

2014 Thought Leader Forum

Proceedings

June 10 – 11, 2014
Greenwich, CT

Welcome

TO THE

THOUGHT
LEADER
FORUM

change
inquire
interaction

CHALLENGE

ENGAGE

Learn
agility
Luck & Skill

connect

WISDOM
{MADNESS}
OF
CROWDS

Michael Mauboussin

Introduction

We are all in the business of forecasting. Getting the future right has significant benefits, and getting it wrong can be very costly. Our natural tendency is to turn to experts when the stakes are high. Yet the ability of experts to predict the future varies widely across a range of realms. For example, experts consistently outperform the average person in dealing with problems in stable and linear systems. Think of an engineer designing a bridge or a novice at chess playing a grandmaster.

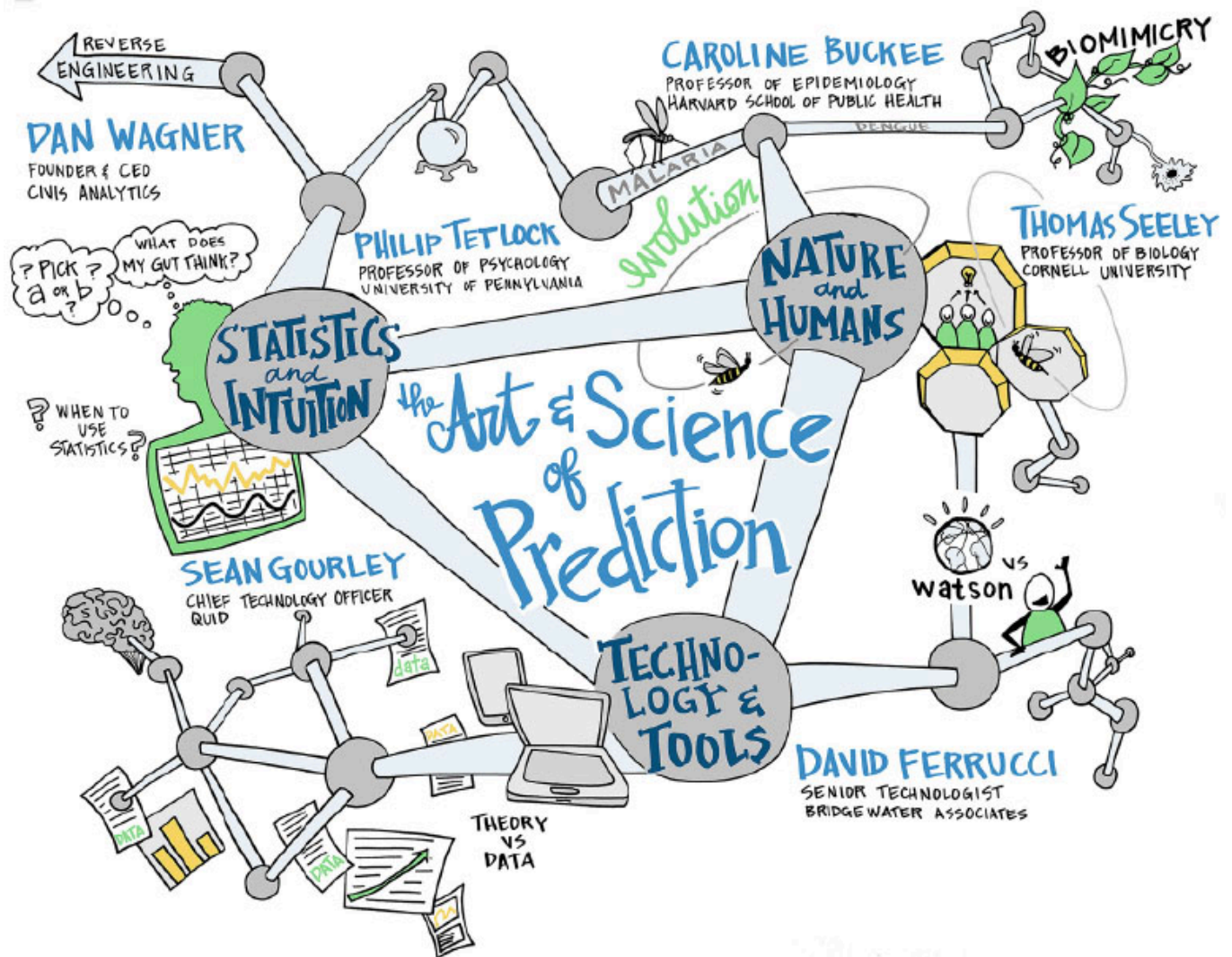
In contrast, experts are poor at predicting the future in unstable and non-linear fields, including economics, politics, and other social systems. These are precisely the types of problems the financial services industry faces. That's why we at Credit Suisse continue to look for original and creative ways to improve our ability to predict. It is also why "The Art and Science of Prediction" was the theme of the Thought Leader Forum in 2014.

The theme in 2014 unites three inter-related concepts that are at the center of some of the most exciting research today: how to blend the ability of humans and computers, when to rely on instincts versus statistics, and how to train to make better decisions. The forum brought together leading thinkers from a range of disciplines, each of whom discussed one or more of these concepts.

The first concept explores the interaction between humans and computers. The challenge is to combine the best of human and computer capabilities while avoiding the limitations of each. Freestyle chess, where individuals play one another but also have the option to get input from computers, is a good illustration. In 1997, Deep Blue beat the world champion, Garry Kasparov, a decisive victory for machine over man. But the top freestyle chess players are better than either man or machine. This raises a host of interesting questions: When should we defer to a computer and when should we override it? How do we make sense of massive amounts of data without getting bogged down by it? And how can we combine data from seemingly disparate sources to help us make better decisions?

The second concept is when to trust our intuition and when to rely on statistical analysis. The tension between the two is rising in many fields today. In baseball, it's the scouts versus the statistics nerds. But it's also fundamental versus quantitative analysts in investing, the old-school political campaigns versus a new breed of statisticians in politics, and doctors versus algorithms for how best to approach treatment in medicine. This tension is closely related to the question of when experts are useful. A key takeaway is experts are poor in non-linear and unstable domains. One method to combine these viewpoints is to consider how much weight to place on the base rate, which is basically what's happened before, and how much weight to place on the circumstances of an individual case. The relative contribution of luck or skill to the final outcome informs that weighting.

The final concept is about how to prepare to make better decisions. Airplane pilots commonly use flight simulators. What is the equivalent of a flight simulator in other realms? Such training would allow us to prepare more effectively for the different environments that will arise. Preparation is also relevant when considering the choices we make in our organizations, including the structures we set up and the people we hire. We also emphasize the importance of cognitive diversity in ensuring that the crowd is wise and in preventing it from acting mad.





Michael Mauboussin **Credit Suisse**

Michael Mauboussin is a Managing Director of Credit Suisse in the Investment Banking division, based in New York. He is the Head of Global Financial Strategies, providing thought leadership and strategy guidance to external clients and internally to Credit Suisse professionals based on his expertise, research and writing in the areas of valuation and portfolio positioning, capital markets theory, competitive strategy analysis, and decision making.

Prior to rejoining Credit Suisse in 2013, he was Chief Investment Strategist at Legg Mason Capital Management. Michael originally joined Credit Suisse in 1992 as a packaged food industry analyst and was named Chief US Investment Strategist in 1999. He is a former president of the Consumer Analyst Group of New York and was repeatedly named to *Institutional Investor's* All-America Research Team and *The Wall Street Journal* All-Star survey in the food industry group.

Michael is the author of *The Success Equation: Untangling Skill and Luck in Business, Sports, and Investing*, *Think Twice: Harnessing the Power of Counterintuition*, and *More Than You Know: Finding Financial Wisdom in Unconventional Places*. He is also co-author, with Alfred Rappaport, of *Expectations Investing: Reading Stock Prices for Better Returns*.

Michael has been an adjunct professor of finance at Columbia Business School since 1993 and is on the faculty of the Heilbrunn Center for Graham and Dodd Investing. He is also chairman of the board of trustees of the Santa Fe Institute, a leading center for multi-disciplinary research in complex systems theory. Michael earned an AB from Georgetown University.

Michael Mauboussin

Welcome Remarks

Good morning, everybody. For those I haven't met, my name is Michael Mauboussin, and I am head of Global Financial Strategies at Credit Suisse.

On behalf of all my colleagues at Credit Suisse, I want to wish you a warm welcome to the 2014 Thought Leader Forum. For those of you who joined us last night, I hope you had a wonderful evening. We certainly have a very exciting lineup for today.

I want to do a couple of things this morning before I hand it off to our speakers. First, I want to highlight some of the themes associated with the art and science of prediction. I then want to discuss the forum itself and what you all can do to contribute to its success.

Let me start with some high level thoughts on the art and science of prediction. There are three thoughts in particular that I'd like to make. I think you'll hear each of these themes throughout the day.

Here are the three themes. They're somewhat interrelated, as you'll see.

The first relates to the interaction between humans and computers. The question is, "How do we combine the best of what humans and computers can do and avoid the worst of what humans and computers can do?"

The second is the classic distinction about whether you should trust your gut or go with statistical analysis. It's important to know when your intuition is likely to work well and when it's going to fail you.

A related issue, which I think is relevant to all of your organizations, is whether there are cultural changes that must occur within an organization as it shifts its thinking about approaches to decision making. As better methods or sets of methods come along, how likely is it that an organization will change?

The final theme is about how we prepare to make better decisions. This could be how we think about choices in our organization or how we hire people.

We study all sorts of topics in school, but there are very few courses on how to make better decisions. How do we close that gap?

So let me start at the top. I have to tell you this is a concept that I have been thinking about a lot the last six to 12 months. I think this was really kicked off by a discussion I heard from Sean Gourley about eight months ago.

The best way to explain this is through chess. We can say 1997 is when machine beat man. That's when Garry Kasparov lost to Deep Blue. Now, today, the best computer programs can beat the best humans.

In the last 15 years or so, there's been the advent of "freestyle chess," where individuals play one another but they also can avail themselves of computers to help guide their moves.

It turns out that those freestyle chess players are better than either man or machine. So, man plus machine beats man or machine.

Further, a number of these great freestyle chess players are not great chess players. They're not grandmasters who've been augmented. They're decent players who are augmented.

So the question becomes, what is the skill that these freestyle chess players have? How do they know when to defer to the computer, and how do they know when to override the computer?

Michael Mauboussin

Welcome Remarks (Continued)

It's almost like this equivalent between quantitative and fundamental strategies. When do you know when to go with a quant? When do you know when to go with a fundamental to override that? Can we define that skill and how we think about that in a corporation or investing context?

The next idea has to do with the interface. Computers are obviously amazing at ones and zeros. Humans are great at seeing patterns.

How do we feed data to the computer so that it can leverage its strength, and how do we compute output from those computers so that we can use the data most effectively as humans?

Naturally, techniques such as visualization are essential, but the key question is how we humans will be able to ingest large amounts of data in a way that we can make sense of it—in the world of finance, business, science, and politics.

A related idea, which also I think is fascinating, is that of mash-ups. How can we bring together data from what are seemingly disparate sources to help us make better decisions? We'll hear about that today, as well, in the instance of tracking disease.

The final point is that we have to acknowledge that as we develop technologies we're going to create systems that are more complex than we can understand. This leads to what Charles Perrow, who is a retired professor of sociology at Yale, called "normal accidents."

One example of a normal accident is the Flash Crash in May of 2010. A single, seemingly innocuous, trade appeared to trigger a series of unanticipated responses. Since many of those responses were a function of algorithms, they were very difficult to stop once they started to cascade.

Same description could be used for the Northeast power blackout in 2003. Again, a really innocuous failure in Ohio led to this cascade of failures.

The question for all of us becomes, as we use technology to improve our world and improve our efficiency, are we injecting a degree of fragility into what we're doing as well that will lead to these large scale effects?

The next issue is when we can rely on our intuition or even our analysis, and when we should rely on statistics. So, differently, where are experts helpful? This is obviously a crucial question as to where we should rely on intuition versus statistics.

There are a couple very well-known psychologists, Danny Kahneman and Gary Klein, who co-wrote a paper on this topic. Kahneman takes a much dimmer view of experts than Klein, who celebrates the role of experts. The title of the paper pretty much said it all. It's called "A Failure to Disagree."

Their point was that one of the best ways to think about this problem is to think about domains in which experts are most likely to be very effective and domains where they are unlikely to be effective.

Where domains are stable and linear, you can develop expertise and intuition that is very powerful. In domains that are non-linear and unstable, you see a lot of failure of experts. Again, this is another theme I'm sure we'll hear a great deal about throughout the day.

A related point, an idea that's very close to my heart, is the consideration of when you should rely on a base rate and when you should rely on the specific circumstances.

After Kahneman won the Nobel Prize, a colleague asked him, "What is your favorite paper you ever wrote?" He mentioned the 1973 paper called, "On the Psychology of Prediction."

Michael Mauboussin

Welcome Remarks (Continued)

In that paper, which he co-wrote with Amos Tversky, they said that there are three elements you need to make a prediction.

One is the base rate, which is basically a history of what's happened before. Second is the facts about an individual case. And the third is a mechanism to weight both of those in your actual forecast.

If you think about a continuum of all luck / no skill to all skill / no luck, one of the simple rules is if an outcome is mostly determined by luck, you should place most of your emphasis on base rates. By contrast, when it's mostly skill, you can focus much more on the specific circumstances.

Just knowing where an activity is on that continuum can give you an informed way to think about how to weight those two things and make a more thoughtful statistical prediction.

The last point is not analytical but it's in every way as important. That is, how can we have organizations think about the world in a different way and incorporate statistical analysis in what they do? This in large part was the drama that was created by Michael Lewis in *Moneyball* a decade ago, when he depicted this squaring off between the scouts and the statistical nerds.

The drama doesn't end with baseball, of course. It's everywhere you look. It's quantitative versus fundamental analysis in investing. It's old-school political campaigns versus a new breed of statistical people. It's the seat-of-the-pants Mad Men versus scientists using A/B testing in advertising. It's doctors versus algorithms for treatments in the hospital.

So the degree to which moving from one set of methods to another set of methods creates an old and a new guard – how will organizations adapt and change to those things?

Let me wrap up with a final section of my three points, and that is how do we prepare to make better decisions going forward?

Whenever I think about this, I often think about how pilots train to improve what they are doing. Of course, the way they typically do that is the use of flight simulators. A flight simulator allows a pilot to hone his or her skills in order to be prepared for different environments in the real world.

The question is how do we assemble our own flight simulators? How do we train ourselves to think more effectively about what's going on in the world?

There are a couple of things that have been pretty well established. One of them we saw beautifully demonstrated last night with Tom Seeley's talk: collectives can solve very difficult problems very effectively.

That is something we now commonly call the wisdom of crowds. Here's the key to the wisdom of crowds – and this needs to be underscored every single time you talk about it – it only works under certain conditions. You have to have the proper conditions in place for the wisdom of crowds to be effective.

Among the most important conditions is diversity. You can achieve diversity within a group as the honeybees did as we saw last night, or you can actually have diversity in your own head, which is much more difficult to do, but also can be done and help you with your better outcomes.

Now, I was with a client not long ago, and he made a statement that completely resonated with me. He said, "You know, I grew up in the investment business hearing not about the wisdom of crowds, but rather hearing about the madness of crowds." That's true for me too. The first book I read was about the madness of crowds not the wisdom of crowds.

Michael Mauboussin

Welcome Remarks (Continued)

That's really my final thought, which is, how do we avoid the madness of crowds? How do we avoid being part of a group that especially is doing something that's suboptimal? In our organizations, how do we create structures that take advantage of that when it arises, or on the other side, avoid it in our organizations?

Those are just a couple of my thoughts going into the day. I look forward to exploring these and other topics with all of you as we go through this.

A couple of things before we start. First, I want to mention we have with us today the very talented folks from Ink Factory. Either Dusty or Sue today will be graphically recording all of our presenters today. For those who were here last night, you saw Dusty hard at work. They are synthesizing the words of our speakers into images and text in order to visually represent the key concepts of each of the themes and each of the talks that you'll hear today.

Their slogan – which I love – is, "You talk, we draw, it's awesome." I really do think it is awesome. I think you'll agree with all of that. Please feel free to take pictures of those boards. You'll see after every presentation the boards will be outside. Last night's board is out there now.

Let me just finish with what we're trying to do today. First, we want to provide you access to speakers who you're unlikely to see in your day-to-day interactions but who are nonetheless capable of provoking useful thought and dialogue. We understand your time is very valuable. And we appreciate that you're sharing some of that time with us in order to expand your horizons.

Second, we really want to encourage a free exchange of ideas. Notice that our speaking slots are a little bit longer than normal. We have breaks between all the speakers allowing you to interact. What we're trying to do is really encourage a lot of back and forth.

We purposefully call this a "forum" instead of a "conference" because a forum is a dialogue, an interaction. We would encourage that environment. Inquire, challenge, exchange are really the key themes.

Finally, we at Credit Suisse would like this to be a great experience for you. So please don't hesitate to ask me or anybody else on our team for anything. We will certainly do our very best to accommodate you.



Philip Tetlock **University of Pennsylvania**

Philip Tetlock is the Annenberg University Professor at the Wharton School and the School of Arts and Sciences at the University of Pennsylvania. He received his PhD in 1979 from Yale University and, with occasional appointments elsewhere, worked mainly at the University of California Berkeley between 1979 and 2010. He has received numerous awards from professional and scientific organizations, including the American Academy of Arts and Sciences, the National Academy of Sciences, the American Psychological Association, the American Political Science Association, and the American Association for the Advancement of Science. He has published more than 200 articles in peer-refereed journals and edited or written nine books.

Phil has a long-standing interest in the enormous difficulty that human beings have in “thinking in time” in drawing logically consistent causal inferences from history and in making empirically accurate conditional forecasts. He has conducted numerous studies of how historians think about “possible pasts” (things that might or would or could have happened if...) and of how political and economic experts think about “possible futures” (things that might or could yet happen). His 2005 book, *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton University Press), explores these issues in depth.

He is currently a co-principal investigator (with his wife Barb Mellers) in the Good Judgment Project, a large-scale forecasting tournament sponsored by the Intelligence Advanced Research Projects Agency (2011-2015). Readers who are interested in participating as volunteer forecasters in the scientific research project are encouraged to visit www.goodjudgmentproject.com. In February 2014, Phil took on the role of Chief Scientific Adviser to Good Judgment LLC, a firm dedicated to translating the scientific discoveries of the Good Judgment Project into implementable advice for private- as well as public-sector organizations.

EXPERT JUDGMENT

20 YEARS DARK

LAST 5 YEARS
LIGHTING
CANDLES

LIGHT

PROBABILITY

IARPA FORECASTING

TOURNAMENTS

400??

MAYBE

VAGUE
VERBIAGE

ACCURACY?

NO TRACK RECORD

HARD TO LEARN

CALIBRATION
&
RESOLUTION

MEASURE
OUTCOMES

SCORECARD

NO
0%

YES
100%

NEED
BENCHMARKS

CONTINUUM

of PREDICTABILITY

MORE

LESS

TOP 2%

WALL ST.
SILICON VALLEY

SUPER FORECASTER!

*FALSE
POSITIVE
RATE

FLUID
INTELLIGENCE



SUPERIOR
PERFORMANCE

- DRIVERS
- RIGHT PEOPLE
- INTERACTION
- TRAINING
- ALGORITHMS

WISDOM

EXTREMIZING



of **CROWDS**

DIVERSITY POWS
ABILITY



DID I
FOLLOW the
PROCESS

VISION

- RIGHT MINDSET
- TEAM SUPPORT

INCREASE KNOWLEDGE

BETTER ?

TAIL
RISK
SENSITIVITY

MINI INDICATORS
will
BRIDGE
the

GAP

OPTOMETRY
TRUMPS
NOSTRADAMUS

E
P F Q Z E
A L F B H
C D M Q R F

IMPROVE our
VISION

= **PHILIP TETLOCK**

Philip Tetlock

The Good Judgment Project

Michael Mauboussin: I'm very pleased to introduce our first speaker today, Professor Philip Tetlock. Phil is the Annenberg University Professor in Psychology at both the Wharton School and the School of Arts and Sciences at the University of Pennsylvania.

I first heard of Phil's work following a meeting with Danny Kahneman in 2005. After that meeting, someone asked Danny what he was reading. He said that one of the most important books he'd read in a very long time was a book written by Phil called *Expert Political Judgment*. By the way, for those who don't own that book it's an essential read for anybody.

Now, most of you probably know the book, or know of the book, which shows that experts in certain fields struggle to make accurate predictions. But Phil didn't let it go at that. He became very determined to figure out how to cultivate good judgment, and so the Good Judgment Project [GJP] was born. Since this is a project that's very much a work in progress, we are all going to be very privileged to hear some of Phil's thoughts that are hot off the press.

I'll add on a personal note that in doing research on my books I've had to do a lot of reading about decision making. Pretty much everywhere I'd look I'd run into Phil's outstanding research. He's contributed a great amount over the years, including topics that we won't be able to cover today. I've learned an enormous amount from him for which I'm very grateful. Please join me in welcoming Professor Phil Tetlock.

[applause]

Phil Tetlock: Thank you, Michael, for that gracious introduction.

I've been studying expert judgment for about 30 years now. You could divide my 30-year plus career on expert judgment into two phases, one of which lasted for about 25 years, the other which has lasted about five years now.

For the first 25 years, it was somewhat born of exasperation with expert judgment. Frankly, the focus was on how bad expert judgment is in many ways. How miscalibrated experts are. How prone to belief perseverance experts are. A variety of errors and biases, somewhat in the Kahneman spirit.

The last five years have been different. We've been focusing less on cursing the darkness and more on lighting candles. The focus has been on how much we can improve probability judgments of possible futures.

There are four key messages I'm going to try to convey in this presentation. The first is that business-as-usual vague-verbiage forecasts slow learning cycles.

And the second message is that you can accelerate learning cycles by using forecasting tournaments, which require making explicit odds estimates on topics you don't normally make odds estimates on.

What do I mean by learning cycles? Well, I mean you'll more quickly learn what you're good and bad at. You'll more quickly learn which of your advisors are better or worse at particular things in particular domains.

Evidence suggests that you'll also become more skilled at using probabilities in a granular fashion, in distinguishing many degrees of uncertainty. I will unpack exactly what that means later. Accelerating learning cycles I think we can all agree, sounds like a pretty good thing.

My third message: I will turn to the key lessons from the tournaments we've been running to-date.

Philip Tetlock

The Good Judgment Project (Continued)

Due to an unusually innovative government agency – the Intelligence Advanced Research Projects Activity, which is the R&D [research and development] branch of the Office of the Director of National Intelligence – we have had funding for four years now for running forecasting tournaments. Sometimes in competition with other universities, sometimes in competition with our own prediction market and sometimes in a shadow competition with predictions emerging from analysts working inside the intelligence community itself. We've learned a lot about how to cultivate good judgment, and it is more or less hot off the press.

Fourth and final message I will be delivering: I want to talk about what the next challenge is in advancing knowledge, and what the next generation of tournaments should focus on, because we are not only learning things about how rapidly people can learn when they make explicit probability judgments, we're also learning something about how to design better tournaments that can improve organizational functioning in even more fundamental ways.

So, that's the agenda.

Now, part one, "vague verbiage forecasts slow learning cycles." I think we all know that pundits and consultants tend to prefer vague verbiage. "A Russian invasion of the Eastern Ukraine is *possible*." "The new Chinese oil rig *could*, or *might*, or *may* signal, in the South China Sea," or the East China Sea, or wherever it might be. "The recent Greek election has the *potential* to..." You've all heard things of that sort. Vague verbiage is very vague, and researchers have quantified just exactly how vague it is.

They've estimated things called quantitative equivalence ranges. When you ask people to translate some of these phrases into probabilities, you often get an enormous range of answers.

"It might happen" takes on meanings ranging from 0.09 to about 0.64. That's how readers decode it. "It could happen," a range from 0.02 to 0.56. "It's a possibility..." "It's a real possibility..." There you've got a real range.

[laughter]

Phil: "It's probable..." "Maybe," even innocent old "maybe." "Distinct possibility," that's a phrase that's caused a great deal of mischief in the intelligence community – both the Bay of Pigs and in a couple of other situations. "Risky." "Some chance." And then there are a few phrases that there isn't that much wiggle room.

As George Tenet, a former director of the Central Intelligence Agency, discovered, "Slam dunk" people really do interpret it to mean something very close to 100 percent. And when you get it wrong, and you're that extremely confident, you take a huge reputational hit, and arguably you should.

These properties of vague verbiage make it impossible to gauge the track records of famous pundits and lucrative consulting shops, some of whose services you may have purchased. Makes it impossible to know how accurate are the thought leaders you read in the *Financial Times* or the *Wall Street Journal* or whatever your favorite publications might be. These are all really high IQ guys. They can give great talks, they can give a better talk than I can. They're wonderful, they're personable, they're charismatic, but nobody has the faintest idea what their track records are. They prefer to keep their own track records.

Vague verbiage also makes it impossible to learn to make the subtlest possible probabilistic distinctions within a given environment. A famous poker player once said that the way to tell the difference between a really serious poker player and a talented amateur is the serious players can tell the difference between a 60/40 proposition and a 40/60 proposition. We talked to some other serious poker players who are among the "superforecasters" in our current research, and they say, "No, no, no. That's not aggressive enough. It's the ability to tell the difference between 55/45 and 45/55."

Philip Tetlock

The Good Judgment Project (Continued)

It's an interesting question how differentiated probability judgments can become. The great psychologist Amos Tversky once said facetiously in a talk that he thought that human beings deep down could really only distinguish three levels of probability: "It's gonna happen," "it's not gonna happen," and "maybe." Those of you who really follow that kind of work, when you see probability weighting functions and so forth, they imply that people give a lot of weight to probabilities at the extremes. When probabilities move from 0.3 to 0.7, there's not that much impact. But when they move from 0.1 to 0, or from 0.9 to 1, boy that has decision impact. In these forecasting tournaments, people do learn gradually. It's a slow process, but they learn to make increasingly differentiated assessments of uncertainty. In domains where you don't normally have such assessments.

Now, there's a question that all this raises. Obviously, pundits and consultants don't like to make precise forecasts because they worry about taking a reputational hit. It raises the question why anyone who has very high status within a particular social system would ever lower him or herself to participate in a forecasting tournament in which the best possible outcome is probably breaking even.

Imagine, for example, that I'm a senior China analyst inside the intelligence community, and I'm the go-to guy for the presidential daily briefing on China and the national intelligence estimates on China. People value me, and now I see these upstarts start scoring probability judgments inside the intelligence community. And the upstarts say to me: "By the way, we want you to compete against these superforecasters who don't have access to any classified information."

How am I going to react to all of that? It doesn't take a lot of imagination to suppose I'm not going to react very favorably. In fact, there is a lot of pushback within most organizations to the idea of measuring the accuracy of probability judgments. The more senior people are, the more likely they are to be resistant. That actually makes good microeconomic sense.

So, all this implies that you should never expect high status Goliaths to help low-status Davids to kill them. You just shouldn't expect things like that to happen. But it did. It did happen in the U.S. government.

The Intelligence Advanced Research Projects Activity [IARPA], which is the R&D branch of a way-over-\$5-billion plus per-year bureaucracy, has funded a set of upstarts at about \$5 million per year, and has engaged these upstarts, in a level playing field competition to see who could assign better probability estimates to events of interest to policymakers, who of course are the customers of intelligence agencies.

So, what happened? It's remarkable. I'm talking about a project that according to the most basic laws of Bureaucracy 101 with rational political actors should never have occurred. The story of why it occurred is one that will probably never be told. It must suffice to say this project would have been impossible but for the efforts of a small group of very dedicated, even visionary, civil servants, whose names I cannot even mention.

That's part one: vague verbiage makes it hard to learn, makes it hard to evaluate thought leaders, pundits, miscellaneous consultants, and the like.

Part two is that forecasting tournaments can accelerate learning. Forecasting tournaments are, in this sense, superior to business as usual. I see forecasting tournaments as a disruptive technology that destabilizes stale status hierarchies dominated by smooth talkers who lack foresight but who are adept at transforming the unforeseen into the retrospectively inevitable.

[laughter]

Phil: That comes from one of our annual reports to IARPA. But you can see why the hypothetical senior China analyst might not like a line like that. It's a bit in your face.

Philip Tetlock

The Good Judgment Project (Continued)

The IARPA tournaments, what do they look like? In the beginning, there were five university-based research programs which won the competition to enter into the competition. The Berkeley...I was a professor for many, many years at Berkeley, and then I moved recently to Wharton at Penn...There was a Berkeley/Penn, a Penn/Berkeley team, there was an MIT team, a Michigan team, and so forth. We were competing with each other to generate better probability estimates.

Each day the tournament was alive, our statisticians and our forecasters would generate forecast streams which would be submitted at 9 AM Eastern Time to the U.S. Government (and its consultancy proxies) using whatever methods the academics thought would work best. If you believed in astrology you could use astrology. You could use anything you wanted. Obviously we weren't using that. But it was designed to incentivize methodological innovation. I think it did exactly that.

IARPA, over the last three years, has now posed almost 400 questions. Some of them actually do fall quite squarely in the financial domain, but most of them are more centrally relevant to national security. Questions like, Will North Korea test a nuclear weapon by "blup"? Will Mario Monti vacate the office of prime minister by date X? Will there be a lethal confrontation in the Crimea before date Y? Will there be a lethal Sino-Japanese confrontation in the East China Sea? Will there be a sovereign debt downgrade of country X? So, a wide range of questions.

Running a tournament requires clearing a certain amount of scientific underbrush. There were three big questions we had to clear away to make the tournament work. One is we needed to demonstrate we could reliably measure the accuracy of human probability judgments. Second, again, we need to decide against which benchmarks is it reasonable to measure human beings. Should we be measuring them against time series models? Should we be measuring them against *Bloomberg* consensus panels? Exactly what should the benchmarks be? Should we be measuring them just against the dart-throwing chimpanzee?

Phil: The third question is, when we fall short of benchmarks, what can we do to augment the judgment skills of both individuals, teams, and organizations.

Measuring the accuracy of probability judgments. This was a technique that was originally developed by meteorologists and statisticians. It is a very straightforward technique. There are many other techniques we could talk about for measuring probability judgments.

The basic idea is you've got a meteorologist who is making predictions, the probability of rain. He's at 90 percent here, 50 percent here, 50, 80. There's an outcome. It rains, rains, doesn't rain, rains.

Then you code reality as either zero, if it doesn't rain, or as one, if it does rain. And you use a simple quadratic scoring rule. You deviate the probabilities against one or zero.

If you say, for example, there's 90 percent probability of rain and it rains, you get a great Brier score. Low Brier score, it's like golf, low scores are great, 0.02. Here, you're with the dart-throwing chimpanzee, right at maybe maximum uncertainty.

Here, again, same thing. Here, you say 80 percent yes, and you get again a pretty good Brier score. You average it all out. It comes to about to 0.28.

When you do this over a large numbers of questions, you can get fairly stable statistical estimates of judgment skill that are correlated with a variety of things that are meaningful, psychologically and organizationally.

If you had a perfectly accurate theory of a deterministic system, your Brier score could approach zero. If you were just guessing, your Brier score would hover in the vicinity of 0.5. And if you were an inverse

Philip Tetlock

The Good Judgment Project (Continued)

clairvoyant, and the opposite of everything you said happened, then you would have a ridiculously high Brier score of 2.0.

You can break these Brier scores down into two key metrics: calibration and resolution.

Calibration is your ability to assign probabilities that over the long term correspond to the objective frequencies with which those events occur.

For those events you assign 70 percent probability to, they tend to occur about 70 percent of the time. For those events you assign 30 percent probability to, they occur about 30 percent of the time. That's calibration. That's a nuanced sense for the limitations of your knowledge.

But, you want more than just calibration. You also want resolution. You want the ability to assign much higher probabilities to things that occur than to things that don't occur. I'll show you what I mean by that.

This is an example of a forecaster who has really good calibration, because when the forecaster says there's a 40 percent likelihood of things happening, things happen 40 percent of the time. When the forecaster says there's a 60 percent chance of things happening, things happen 60 percent of the time. 50 percent likelihood, things happen 50 percent of the time.

The only problem is this forecaster never says anything very different from minor shades of "maybe." Your forecaster never moves outside the bracketed range between 0.4 and 0.6.

This is an example of a forecaster who has wonderful calibration and very good discrimination. You can see all of the points are along the diagonal there.

And, finally, this is what perfection would look like. This is omniscience or God. This would be someone, who as soon as the question is asked, can either tell you zero or one. "It's not going to occur, it is going to occur," and do it with infallible accuracy. Needless to say, nobody ever even remotely approximates this.

What's the performance range on the Brier scale? It's zero omniscience, 0.5 the chimpanzee, and lots of variation in between.

What kinds of benchmarks should we use for judging the performance of humans? That was another question we had to wrestle with before the tournaments.

There were minimalist benchmarks, like the dart-throwing chimpanzee, simple extrapolation in time series models. More of the same. Can you out-predict a "more of the same" algorithm.

There are moderately aggressive benchmarks for assessing performance, like the unweighted mean or median of the wisdom of the crowd which is quite difficult to beat.

This was, by the way, the metric we decided to use initially in the IARPA tournament. And then expert consensus panels, which you guys are familiar with, like the Philadelphia Fed or the Economist Intelligence Unit, or *Bloomberg*.

Finally, maximalist benchmarks: the most advanced statistical techniques and Big Data models, and beating, finally, the ultimate litmus test, I guess, for people in this room, is beating deep, liquid markets.

How well can our best forecasters and our best algorithms do against various benchmarks? There's no across-the-board answer to that question. It depends on the type of problem with which we're dealing.

It's useful to think of a predictability continuum. At the most predictable end, Foucault's pendulum. If you know Newtonian physics and you know a little bit about the planet earth, you're going to be able to predict, with virtually perfect accuracy, the motion of Foucault's pendulum. Your Brier score will be zero.

Philip Tetlock

The Good Judgment Project (Continued)

At the other end of the continuum, there are no experts at the roulette wheel. It doesn't matter how much you train, it's not gonna help.

Finally, there's the interesting case of predicting the trajectories of hurricanes. Nobody can predict when and where exactly a hurricane will form. There are butterfly wings and all that kind of stuff intruding and producing chaos and turbulence. But once a hurricane has reached a certain level of maturity, it is possible to predict.

Forty years ago, meteorologists weren't very good at it. Now they are pretty good at it. It's a success story. Meteorologists have become better at it, partly because the underlying science has become better, but also because meteorologists are in the regular habit of scoring their calibration and resolution – the components of the Brier score. They're getting feedback and getting better at it.

There's a reason why meteorologists are among the best-calibrated professional groups studied in the literature on expertise. They're up there with expert poker players, expert bridge players. It's a small, rarified crowd of people who are really super well-calibrated.

How much can we improve foresight? When you look at the results of the IARPA tournament, we were able to do about 50 to 70 percent better than the unweighted average of the crowd in a control group.

One of the things you should appreciate about what goes on in the IARPA tournament is that we run experiments. Forecasters are randomly assigned to experimental conditions. There's a random assignment to a control group. We compute the average forecast or prediction in that control group. That's the benchmark we have to beat.

If we're going to say that we have a training system that improves probability judgment, it has to improve probability judgment above and beyond that control group. It's rather like an FDA [Food and Drug Administration] drug trial, you've got to beat the control group.

We were pleased that we were able to replicate that performance now in three years, in randomized trials in which thousands of forecasters make predictions on hundreds of questions.

We have a lot of data—and our margin of victory was sufficient to knock the other academic competitors out of the competition after the second year, and it puts us in direct competition now with our own prediction market, run by Inking, and in indirect competition with predictions emerging from the intelligence community itself.

We now know that we were even able to do better than experts with access to classified information because David Ignatius wrote a story in the *Washington Post* in 2013 in which he revealed that, according to his sources, our best forecasters and best algorithms were able to do about 30 percent better than the forecasts being generated within the intelligence community.

I think that's pretty amazing. *NPR* [National Public Radio] ran a story on it, with a terribly tacky title “Are you smarter than a CIA agent?”—and that triggered a veritable avalanche of new volunteers. They featured one of our very best forecasters, a pharmacist in Delaware, who was tracking Syrian refugee flows. The story went viral.

From my point of view, what we've learned over the last three years is there are four categories of drivers of superior performance. We know this with unusual confidence because of the experimental nature of the design. These are rigorous, controlled experiments. These aren't just correlational designs.

Philip Tetlock

The Good Judgment Project (Continued)

We know it's important to get the right people on the bus, because we can compare our control group to other control groups. We know we had a little bit of success in getting smarter, more open-minded forecasters. We had better, raw human material.

We know about the benefits of interaction, experimentally again, because we can assess it against a control group.

We get about a 10 to 20 percent boost. Forecasters do better when they're working either collaboratively in teams that are structured in certain ways we could talk about, or competitively in prediction markets.

We've been able to design training modules that improve judgment in the range of 10 percent. These are cognitive de-biasing exercises.

Finally, our statisticians have done a wonderful job in developing algorithms, weighted averaging algorithms that give more weight to our better forecasters, and then "extremize." I'll explain the logic of extremizing in a minute.

Turning to the algorithms, the classic wisdom of the crowd generalization is that the average forecast tends to be more accurate than most of the individuals from whom the average was derived.

That's, of course, the famous Galton ox story that was popularized by James Surowiecki, and there's a lot of truth to that. We replicate that finding over and over in various forms.

The Good Judgment Project does not use unweighted averaging. Our statisticians have found it's much more effective to use weighted averaging. We give more weight to higher IQ, more open-minded forecasters who have better track records. Plus we add something that is somewhat controversial, we called an "extremizing parameter." I'm going to explain what that is.

What the transformation parameter does is, at 0.5, it's equal to the actual expressed probabilities, but it extremizes, so it turns 0.6 into about 0.72, and 0.4 into about 0.27. It extremizes.

It looks like this, mathematically. It's a log-odds with shrinkage plus noise. There's a key parameter here. It's a , parameter a . That determines the amount of the transformation.

The amount of the transformation, according to our statisticians – I think they've made a very compelling case for this – depends on the sophistication and diversity of the forecaster pool. The more diverse the forecasting pool, the more extremely you can extremize.

I'll give you a thought-experiment example. Many of you have probably seen the movie *Zero Dark Thirty*, in which James Gandolfini, the late actor, played Leon Panetta when he was Director of the CIA [Central Intelligence Agency] at the time of the operation to target Osama Bin Laden.

There was a debate inside the intelligence community about whether Osama Bin Laden was living in a particular compound in the Pakistani town of Abbottabad. In the movie...and Hollywood manages to get things deeply right and deeply wrong, simultaneously, it's interesting how they manage to do that, but that's art, that's the art part of the art and science...

[laughter]

Phil: They're around the table and the director says, "What's your best probability estimate of this?" Each of the advisors gives an estimate. Just assume – the movie doesn't unfold like this – but assume that Panetta's advisors, each of them says to him, "0.7, 0.7, 0.7." Sounds like the answer is 0.7.

If the advisors were clones of each other, the answer would be 0.7.

Philip Tetlock

The Good Judgment Project (Continued)

But if one of the advisors is drawing on human intelligence, another on satellite intelligence, another on some other form of intelligence, if they're drawing on very different forms of intelligence, and each of them is reaching a 0.7 conclusion, what's the right answer?

We don't know exactly. There's no mathematical answer to that, but you have a pretty strong intuition now that it's more than 0.7.

That's the intuition the extremizing parameter captures.

The art and science, what should this extremizing parameter look like?

You could create an extremizing parameter that's so extreme that as soon as you go up to 51 percent, you go to one. As soon as you go to 49 percent, you go down to zero. That actually works to some degree, but not as well as this. The more moderate extremizing parameter is a better way to go.

Let me say a few words. I mentioned superforecasters. In the IARPA tournament, we take the top two percent of our forecasters and we assign them to elite teams. They work with each other in teams of 12.

Last year, we had about 10 teams of 12 superforecasters working together. They have performed phenomenally.

The superforecasters are the only method of generating probabilities that comes close to beating the log-odds transformation I just described to you.

Human beings cannot beat this transformation. But the supers give it a run for its money, which I think is quite remarkable.

The supers are also better than all untransformed sources of data in the IARPA tournament. They're better than the unweighted average of the wisdom of the crowd, better than average teams, and better than the prediction markets. They're better than the aggregated top teams. They are a force of nature.

Of course, the question you're all asking is "When are they going to fall down to the ground? When is regression toward the mean going to kick in?" I don't know the answer to that question, but I think there is going to be some regression.

Individual supers do regress to some degree toward the mean. Somehow the superforecaster teams have been able to resist that thus far. But I'm fully aware that there is a large chance component to the events that we're trying to predict in the IARPA tournament, and that regression toward the mean is a real process here.

But there is a skill component as well. And as Michael [Mauboussin's] recent book underscores, the amount of regression toward the mean you should expect is going to be a function of your intuition about the ratio of importance of luck and skill.

The more you think it's just luck, the more you should expect very rapid regression toward the mean. The more you think it's skill – the less you should expect—and right now, we're finding a substantial amount of skill.

Here's another illustration of how super the supers are. These are called, in engineering, receiver operating characteristic curves. They refer to the ability of a perceiver or observer to achieve a really high hit rate at a low cost in false positives.

Again, the more godlike you are, the more capable you are of predicting all economic recessions, up to one, up to perfection, at no cost in false positives. It would be a perfect ROC, omniscient.

Philip Tetlock

The Good Judgment Project (Continued)

The supers are way, way better than top teams, top individuals, and others. It's quite remarkable how well they manage to do on the ROC curves.

One of the big key takeaway lessons from tournaments: forecasting is a skill grounded in cognitive styles and abilities. This has implications for personnel selection. There are tests you can use to select people.

Forecasting is a skill that can be cultivated across a broad swath of your workforce. It can be done learning by doing, simply participating in tournaments, becoming increasingly granular, learning how to distinguish 40/60 from 60/40 propositions with repeated practice. It can be done through cognitive de-biasing exercises.

It can be done by assigning forecasters to teams that have learning cultures. There's another talk that could be had about how exactly you organize teams to make them work as effectively as they have in the IARPA tournament.

Finally, and very importantly, well-designed algorithms for distilling the wisdom of the crowd can produce substantial boosts in accuracy. They are, statistically, our biggest effect.

But tournaments cannot deliver miracles. You all know that. You might be thinking already, "This is too good to be true."

I'm not claiming to be a modern day Nostradamus. I think we want to steer clear of the looking-for-forecasting-superstars fallacy. I think the biggest mistake that consumers of forecasting products make is looking for visionaries.

If someone promises you a 90 percent hit rate on geopolitical, geoeconomic questions, the right question to ask is, "What's your false positive rate and who can independently verify the claims?" I think those are very reasonable questions to ask.

In a stochastic world, high hit rates always come at some price in false positives. That's true for our supers, that's going to be true for anybody else. That's like the old joke about economists having predicted 11 of the last 5 recessions or whatever.

Can we improve your vision? How relevant is this to the people in this room? There, the devil is going to lurk in the details.

First question, how close are you already to the optimal forecasting frontier?

If you're convinced that you're already close to the optimal frontier, there's really not much reason for you to be interested in trying to improve foresight. The cost benefit equation, there just isn't much payoff.

You might think that you're at the optimal forecasting frontier even though you're only doing slightly better than the dart-throwing chimpanzee, but that's because you're working in such an extremely different environment that nobody can do much better than a dart-throwing chimpanzee.

You might think that you're close to the optimal frontier because you're considerably above that. You don't know where the optimal forecasting frontier is however, until you run forecasting tournaments and run experiments and assess whether or not it is possible to get some incremental accuracy. Otherwise, you're simply taking it on faith.

It also hinges on whether your forecasters have the right mindset and the right types of team support. Both are crucial.

Philip Tetlock

The Good Judgment Project (Continued)

What is the right mindset? If you as a forecaster subscribe to the theory that it's predominantly luck or subscribe to the theory that it's predominantly innate ability – it's like *g*, it's hard wired into your genes, general intelligence – if you subscribe to either of those theories, there's very little incentive for you to invest in improving foresight.

The people who improve are the people who believe that it's possible to improve. There is an element of self-confirming prophecy here. It's real and it's worth thinking about.

The team support is also a crucial factor because this is hard work and having support from teammates is valuable.

Another factor is how does management balance process versus outcome accountability in evaluating employees. This is a big issue in management literature.

In the intelligence community, it is gospel that you don't hold the analyst, the lower level analyst accountable for errors, getting it wrong, if they went through the right processes.

If they have a good process score in constructing their reports and they're wrong, they don't get nailed. The Director of the CIA might get fired, but the lower level analysts are insulated in a process accountability system. They think it's fair.

It's like a contract between the employees and the organization. You follow these best process practices and we will indemnify you against the stochastic nature of the environment. That's the solution they have arrived at.

Unfortunately, that solution can ossify into bureaucratic ritualism and box checking in which analysts can anticipate what managers want to hear and the managers want to get things done fast, and the process accountability can degenerate quite rapidly.

Our research suggests that the best way to get the best out of your employees is to experiment with hybrid process outcome systems where you want to hold people accountable for best processes, but you also want to hold them accountable for innovating and finding solutions beyond the existing process recommendations of the organization. Exactly how you balance that, I call it a nasty trade-off. It's a tough one.

Then, also whether your statisticians know how to design algorithms and tailor them to the sophistication and diversity of your forecasting pool. High-ability, high-diversity pools, you have some substantial warrant for extremizing.

In my view, this is unglamorous, but optometry trumps prophecy. It's a bit like this. It's a bit like looking at the Snellen eye chart. You gradually become a little bit clearer, and how much clearer you can become is obviously going to hinge on the nature of the forecasting environment within which you're working.

GJP's methods can improve foresight using experimentally tested tools, personnel selection, training, teaming, incentives, and algorithms. Those are the things that work.

The limits? GJP's methods are not perfect. We can't get anywhere close to zero and one and the god-like omniscience I showed earlier. On average, GJP's best methods assign probabilities between about 24 and 28 percent to things that don't happen and probabilities of 72 to 76 percent to things that do. It's a lot better than 50/50. It's much short of god or omniscience.

OK, last section, Part Four, the next steps in advancing knowledge. I think we've learned some things about how to do forecasting tournaments better. I think the second generation of tournaments that we've

Philip Tetlock

The Good Judgment Project (Continued)

been talking about running with IARPA and the intelligence community, and maybe with private-sector entities as well, these are the things I think they should focus on.

The quality of questions as well as the quality of answers – what does it mean to generate a good question for a forecasting tournament? We talked about that a bit.

The accuracy and judgments of tail risks as well as your judgments of mainstream risks, the Nassim Taleb critique of forecasting tournaments.

And finally the performance of humans versus models versus hybrid systems, on which I hope that IARPA will decide to launch a new tournament. I don't know that for sure.

Populating tournaments with better questions. Tournaments require rigorous and resolvable questions like, "There will be a Sino-Japanese clash in the East China Sea with 10 or more dead." There needs to be something very, very specific, a specific downgrade, a downgrade from this to this. You don't want room for after-the-fact wiggle room about who was right and who was wrong.

Forecasting tournaments ideally have that kind of clarity, but rigorous questions aren't often the questions that top policymakers most care about. Top policymakers are less interested in whether there's an accidental clash in the East China Sea that kills some people than they are in what are Chinese geopolitical intentions? How far do they intend to push on various fronts?

It's very hard to ask the question, "Have Chinese geopolitical intentions become more confrontational?" and then come back in a year and say, "Ah, you were right, and you were wrong." It doesn't work very well.

So, we need to bridge the gap by developing what we call multiple mini-indicators of big issues and triangulating. We call these things Bayesian question clusters. We have many, many questions, each targeting a big policy issue. While trying to become more aggressive, we had a set of questions this year about whether the Chinese would interdict a U.S. naval vessel, a big issue. It actually happened. It didn't make the news though.

A Sino-Japanese conflict, South China Sea fishing conflicts with Vietnam and the Philippines, the Japanese prime minister visiting a World War II shrine. Will China sell a UAV [unmanned aerial vehicle] to another country, and the conflict between China and the Philippines over the Second Thomas Shoal. Each of these is a manifestation...if all these questions resolved in the confrontational direction, that would tip the policy debate somewhat toward the view that China is becoming more confrontational.

That's one way of going, improving the quality of the questions. The other way of going is creating more tail-risk sensitivity and addressing what I think of as the "Nassim Taleb question."

I consider Nassim to be a friendly critic. Actually he and Michael [Mauboussin] once jointly visited a class I was doing out at Columbia many years ago. Nassim has some very good grounds I think for being skeptical of all this.

Imagine the superforecasters tell us that there's a four percent likelihood of a Sino-Japanese clash claiming 10 or more lives in the next 16 months or some such time period. That's interesting, but it's not all that reassuring. It doesn't tell us what's the likelihood of not 10 dead but how about 1,000 dead or a million dead, World War III, extreme tail risks.

We need to develop forecasting tournaments that are more tail-risk sensitive. We need to develop indicators that are sensitive to skill in tail-risk domains, which is extremely difficult because assessing the

Philip Tetlock

The Good Judgment Project (Continued)

accuracy of probabilities in the 0.00001 range arguably would require forecasting tournaments that would last tens of hundreds of thousands of years, which may be even beyond Ray Kurzweil's lifespan.

[laughter]

Phil: More tail-risk sensitivity. Here are the two workarounds we have for that right now: by helping forecasters learn to make subtler distinctions, and by helping managers to pose more probative questions.

Learning to make subtler distinctions, I'm going to revisit now the theme on training.

We've been developing techniques for training people to become more logically consistent in their judgments of tiny probabilities because we know from the cognitive psychological literature that people are really bad at dealing with very small probabilities. Estimates of low-probability events are, to put it very kindly, inconsistent. They are often wildly inconsistent.

You ask people a question like, "Chance of someone who drives one hour per day having a crash above the insurance deductible in one day versus 10 years?" What they'll typically do is they'll underestimate the probability in 10 years. Then they'll overestimate the probability in one day.

They'll often give probabilities in one day that would imply they would have many accidents in the 10-year period. This is a hard thing to learn how to do, and people need to be trained how to do it.

We obviously can't pin down the probability of World War III or whatever other extreme tail-risk events, an epidemic or whatever it might be. We can't do that, but we can design exercises that make forecasters more logically consistent and less prone to biases such as partition dependence.

One of the things about tail risk is that people underestimate it until the tail risk is called to their attention, at which point they grossly overestimate it. So it's very, very difficult to get it right, between oblivion and hypersensitivity. How do you strike that balance between oblivious and hypersensitive?

Learning to pose more probative questions: that's the other way of making tournaments more tail risk sensitive. Probative means here shedding light on unobservable tail risks. One way to do it is developing early warning indicators of scenario trajectories. So the sort of thing with China, if China really were becoming extremely confrontational, what sorts of things do subject matter experts tell us we should be observing in the next 12 to 18 months. Developing questions of that sort.

Another very interesting and subtle way of finding out who's better and who's worse with respect to tail risks, is by asking people not just to make predictions about events, but to make predictions about the predictions other people are making. This is a technique I owe to one of our collaborators in GJP, Drazen Prelec at MIT developed a technique called Bayesian truth serum.

It turns out the people who are better at predicting other people's predictions are also better at predicting the event of interest, and therefore you have a good basis ex ante, for knowing who's going to be a better predictor, because you know immediately who's correctly predicting other people's predictions. It's also helpful to be making predictions not only about other people's predictions now but other people's predictions in the future. So someone who said a year ago that consensus opinion was that Russia would be more confrontational a year later, like now, would obviously have hit a bullseye.

Once we know which forecasters are better on these logical consistency tests, and the forecasters who are better at predicting early warning indicators and better at predicting others' predictions, once you've got that information then I think you have a statistical basis for giving greater weight to certain types of people in tail risk-style tournaments. You don't have to wait 10,000 years necessarily.

Philip Tetlock

The Good Judgment Project (Continued)

But how much uncertainty are you reducing? The amount of vision, the amount of clarity you can get on tail risk is always going to be extremely limited. You might be moving from totally blurry to just a slightly less than totally blurry. That's the right way to think about it on the eye chart continuum.

I'm going to close by dispelling some false dichotomies. How do you think of accuracy? Most people think you're either right or wrong.

The right way to think of it, according to the Good Judgment Project, is that you should think about it as a continuum. How do you explain accuracy? Most people think luck or skill, and as Michael [Mauboussin] pointed out, you should think about it as a continuum. There's a luck-skill continuum. What are the underlying sources of skill? Nature versus nurture? Again, that's not the right way to think about it. It's nature and nurture in various combinations. Fluid intelligence does have a genetic component to it, that's true, and it's important for forecasting accuracy.

But there are all sorts of things that are under your control that can also produce greater accuracy. So it's not all or none.

So how are we different from some of these other people who are in the public limelight? First, Daniel Kahneman, who's an old colleague of mine from Berkeley, as well as a mega-famous guy now, humans have cognitive shortcomings. That's the essential Kahneman argument. In fact, these are virtually universal cognitive shortcomings.

They're very close to hardwired into our perceptual cognitive apparatus. We agree with Danny that there's a lot of bias, but we're more optimistic about the feasibility of improving judgment. If there's daylight between our positions, it's there. We're more cautiously optimistic about the potential for improvement.

Nassim Taleb says the world is radically unpredictable, so tournaments that lead people to think it might not be radically unpredictable are a disservice rather than a service. We agree that there's a lot of radical unpredictability out there. But we also see a key role for probabilistic foresight in helping institutions prioritize.

To use Nassim's terms about anti-fragilizing institutions, anti-fragilizing banks or other institutions is extremely expensive. So you have to set priorities, and when you set priorities, you need probabilities.

Then finally, someone else you may have heard of, Bruce Bueno de Mesquita, who's at NYU, wrote the book *The Predictioneer's Game*, and has a moderate size consulting firm: Analytical tools, game theory tools, can predict the future.

We think there's some truth to that. But we think that you don't need to construct really detailed game theoretic models for each problem, that you could deliver levels of accuracy comparable to what he can deliver much more cheaply using tournament methodologies.

That is where I position myself in this space. There are obviously other people we could have compared ourselves to, but I've talked a long time, so...we should open things up for questions. Yes?

Question: How would you apply this in an investment setting? So you have a team of analysts, portfolio managers, and you set up a tournament some way to have them make varying forecasts, and then how would you implement that in our world?

Phil: I would say that exactly how you implement it is going to be a function of what top management considers to be the decisions that are sufficiently probability-sensitive that they're worth including in a tournament. If there are domains in which managers say, "Boy, if we had a better bead on these probabilities, it would influence this investment strategy, or if we knew that Janet Yellen was really

Philip Tetlock

The Good Judgment Project (Continued)

extremely dovish rather than just moderately dovish, that would influence what we would do with bup, bup, bup."

I'm not the finance guy. I would defer to you on how you would do that, but I think tournaments work best if senior management believes that there are categories of questions that are sufficiently probability-sensitive, that it's worth the investment to set up a tournament. Again, the need for better questions—and for setting up competitions to generate better questions as well as better answers.

Yes?

Question: How are the superforecasters judged up front? In other words, how do you identify them? And second, what were their characteristics – in thinking of hiring these types of people, what characteristics do you look for?

Phil: Some of them are Wall Street analysts. There are two domains that are overrepresented among superforecasters, geographically and professionally. I would say they are Wall Street financial professionals and Silicon Valley programmers. There are lots of other people who are supers too. There's the Delaware pharmacist, and there are all sorts of eccentric people scattered across the country who are interesting and very talented.

But what do supers look like? They were the top two percent in a given forecasting year. They got the best Brier scores, a very simple rule for choosing them. Then they get to work with each other, and they love to work with each other. One of the supers said upon being assigned to a super team he said, "Oh, I thought I was the only intelligent person in the tournament!" These supers are elitist.

[laughter]

Phil: They are unapologetic elitists, and this is a pretty elitist strategy we're using here. They score better on fluid intelligence, they score better on crystallized intelligence, they score better on cognitive style. They don't treat their beliefs as precious possessions. They treat their beliefs as testable hypotheses. They're intellectually agile. They've got quite a bit of crunching power.

And they believe this is a skill that can be cultivated. You could have all the intelligence and cognitive style prerequisites a superforecaster has, but if you believe that it's essentially luck or it's essentially hardwired ability, there would be very little reason for you to invest the effort required to cultivate this skill.

Unfortunately, I didn't bring the graphic that Michael [Mauboussin] has seen before. It was a test that was developed in the 1930s. The militaries in World War II latched onto it as a basis for distinguishing farm boys who have officer potential from farm boys who just go directly into the infantry. The Soviet red Army also apparently used it—or so I have heard.

It was an officer potential measure that doesn't depend on literacy. You don't have to be able to read, really. All you have to do is look at the patterns, see how the patterns are evolving, and make a prediction about what the next pattern will look like based on the generative rule that was causing the evolution of the previous patterns. This is called "Raven's Progressive Matrices." It's a classic measure in the psychological literature. It works.

I think it works as well as it does in our tournament because our questions are so extremely heterogeneous. Look at the enormous variety of questions we're asking. General intelligence does play a bigger role.

If you were running a tournament in a more concentrated domain, general fluid intelligence would become less important and crystallized would become more important.

Philip Tetlock

The Good Judgment Project (Continued)

Question: How old is the forecasters group, on average?

Phil: That's interesting. There is a significant group who are early retirees who are looking for things to do. Relatively affluent early retirees are overrepresented. But I would say the average age is probably in the late 30s—and the median a bit lower.

Question: What makes them continue to operate well and not become dysfunctional and disagreeable?

Phil: We give them some guidance on how to interact together in teams, the old Andy Grove constructive confrontation mantra – “How to disagree without being disagreeable.” Some of the superforecasters, their social skills are not that great. One superforecaster team did crash and burn this year, actually. But, overall, they really did well.

We give them guidance on how to do this. We give them guidance on a technique called “precision questioning.” We give them guidance on how to get along with each other, how to be supportive of each other and not to be destructive, how to be constructively critical rather than destructively critical. I think that training does help.

They're really attuned to the dangers of groupthink. If there's any danger in the super teams, it's not really groupthink as much as it is destructive competition. But, for the most part, we've been able to keep that under control.

Let's see. Yes?

Question: What explains the performance of the intelligence community, which has better information? Four average forecasters with better data to work with should do better than a superforecaster.

Phil: That's a great question. If you were to ask some people in the intelligence community, they would say, “That's because we haven't begun to fight. We weren't taking you seriously. And if we ever did take you really seriously, we would kick your butt.” That's one reaction.

Another possibility is that the parts of the intelligence community we're dealing with is not accustomed to using quantitative probabilities. It's just taking time for them to get accustomed to doing it.

Question: Is this an overconfidence in the data? Do you think overconfidence is doing that?

Phil: That's another explanation that's been offered—by David Ignatius. It's that the intelligence analysts are overwhelmed by classified information of dubious usefulness. It's a signal/noise problem. It could be that our forecasters have the benefit of *not* having classified information, rather than the liability.

Yes?

Question: What other biases lead to bad forecasts? Think about the first of January – people think all these things are going to happen, a lot more than probably do in a typical 12-month period.

Phil: Yes. There could be a very simple heuristic that spares our better forecasters from that error. It is a base rate heuristic that change is a lower probability outcome, on average, across thousands of forecasting problems.

There is a tendency to exaggerate the probability of change. It's also a more interesting forecast. And forecasters want to be interesting. There are incentives operating that encourage probabilistic inaccuracy.

Question: When there are no base rates at all, when you're dealing with a pure sui generis, unique event, what can the forecasters do?

Philip Tetlock

The Good Judgment Project (Continued)

Phil: Well, as Michael [Mauboussin] said about the classic Kahneman and Tversky paper, the only thing that's left is inside view. You're going to have to draw on causal principles. You say, "There's nobody who resembles this person in history."

It's very hard for me to imagine, by the way, that something is totally unique. Even for extremely idiosyncratic events, it's typically possible for our best forecasters to cobble together a reasonable base rate.

But in the cases where no base rate comes to mind at all, you will be required to rely on the inside view and causal generalizations that plug into the inside view. Maybe causal generalizations from game theory or from of some other body of theory in social science. Then the question is, "What kind of forecast does that lead to?" But you would have to rely on pure causal reasoning in that scenario.

My experience is that's a little dangerous. But it is often essential. Certainly, it is true that our forecasters weave together statistical and causal reasoning all the time.

Question: You talked a lot about the value of diversity in bringing together the forecasters and then the weighting schemes. Can you talk a little bit about the savvy metrics or attributes you look for in finding diversity across the forecaster groups?

Phil: When I use the term "diversity," I'm not using the term "diversity" in the way human resources in your organizations use diversity, which is more of a legal construct having to do with protected groups and discrimination. I'm using diversity in the sense of functional diversity or substantive diversity.

How many points of view on China are represented around the table? What are they? Are there people around the table who know Chinese well, that are following the Chinese media in depth, who have lived in China? Are there people who have known people on the Chinese Central Committee? Things like that.

Question: You get these people who want to become forecasters. Is it like a trial and error to figure out how diverse they are, or are there questions you can ask?

Phil: This is a great question. Diversity makes our statisticians tear their hair out because they know it's important in determining how extreme the extremization algorithm should be: what value should A take in that formula. They know that's really important logically and mathematically. They haven't been able to pin down the specific attributes of forecasters that are essential for diversity in different domains.

I think that's because the questions are so heterogeneous. As we move toward a model where we have question clusters, a big, big question cluster on China, we'll be able to define diversity more meaningfully. But they flit from China to Myanmar, to India, to Greece. It's very difficult. It's a work in progress.

Question: Can you discuss how Nassim Taleb's views on path dependency fit in here?

Phil: Part of Nassim's argument is deeply philosophical. Take the argument you see nowadays between people who think there's an analogy between 1914 and 2014, the Anglo-German rivalry that preceded World War I and the Sino-American tensions today. You have rising power in Germany and China. You have an entrenched power in England or the United States.

The claim is that there's a causal law operating here. You look through history all the way back to Sparta and Athens in the Peloponnesian War. You see a pattern in which a rising power coming up against a hegemon, a dominant power, that the likelihood of war goes boom. Therefore, the likelihood of something really nasty happening in Asia is way up there.

Philip Tetlock

The Good Judgment Project (Continued)

Now, the path dependency argument that you mentioned here is a very interesting one. Imagine I think that World War I was not inevitable. I think World War I was a fluke that occurred because some crazy guy shot the archduke in Sarajevo. And if you were to rerun early 20th century history a thousand times, World War I would emerge at most once.

If I'm an extreme path-dependency person like that, I'm not going to be too impressed by the analogy between 1914 and 2014. I think 2014 is just very different from 1914.

The assumptions you make about counterfactuals and path dependency play a pivotal role in how you use historical analogies, which play a pivotal role in how you generate predictions on China. Does that make sense?

There's an assertion-to-evidence ratio problem here. Nassim does make a lot of assertions, but the amount of evidence for some of the assertions is not as great as you might hope. Some of the claims he's making are of rather sweeping, philosophy of history character.

Question: Path dependency could be very important to the financial markets, for example. You might be approaching an event, but you could go in one direction or the other. All of a sudden, all the versions come into play, from this direction and another one from that direction. That's the sort of problem that we often deal with in the financial markets.

Phil: I see a lot of argument among my economist friends about the importance of path dependency. The economists, for example, say the Flash Crash, [snaps] it's gone. What's the lasting legacy of the Flash Crash? They see markets as very self-correcting, as in general-equilibrium models.

Whereas others, like Nassim and many other people who think about history, see it as much more path-dependent. I don't know what the distribution of opinion in this room would be about markets and path dependency.

Question: Can you talk about this idea that more narrow domains...Is this a transferable skill? Somebody who does well at predicting sports can also be the better predictor in markets, can be the better predictor in weather? Is it agnostic to the domain, or not?

Phil: My earlier discussion about general intelligence, the importance of general intelligence as a predictor of performance in an extremely heterogeneous question tournament of the IARPA sort, implies that there is a foresight g , to some degree, yes.

But the evidence that we have that is cultivatable--through practice, training, and teaming--suggests it's not so much a g as it is something that's going to be domain-specific. The things you really care about, you're going to get pretty good at. I think there is some generalizability, but I think there are also pockets of extreme domain-specificity.

Question: Can you talk about some de-biasing techniques, in the context of tail risks? Are there de-biasing techniques that are applied more generally you can tell us about?

Phil: The most basic de-biasing technique is to find the right reference and comparison classes before you make a prediction. What kinds of situations are similar to this situation, from the past? When multiple comparison classes come to mind, then you probably do some kind of averaging or weighted averaging of those comparison classes. That then becomes your initial probability estimate, and then you go into the details of the inside view of the case. That works pretty well as a rough, first approximation for predicting, say US GDP [gross domestic product] growth.

Philip Tetlock

The Good Judgment Project (Continued)

It may work well in what we call the 10/90 domain. For a question to make it into the IARPA tournament, subject matter experts have to consider that it's at least 10 percent likely, but not more than 90 percent likely. They want them to be somewhat difficult. Our tournament has deliberately excluded tail risks. That's one of the reasons I emphasize the importance of a tail-risk focus for second generation tournaments.

Question: Is there any correlation between two predictors and what people refer to as the autistic spectrum?

Phil: You mean more Asperger-ish than autistic?

Question: Yes.

Phil: My guess-timate: Not very high. It may be slight, but not very high. By slight, perhaps a correlation coefficient of 0.10, 0.15. That's just a guess. My hunch is that you're onto something, but it's small.

Audience Member: I was a participant in the most recent Good Judgment Project.

Phil: Oh, thank you.

Question: There were basically two elements to making a good prediction – your initial prediction and updating your view. Could you talk about the relative importance of these two things?

Phil: Lots of complicated issues having to do with timing, updating, and updating through time. Here's what I can say. I could say that when you look at the correlates of accuracy for judgments made when questions were first asked, in the middle phase of the questions, and in the final phase of the questions, they're more or less identical.

It doesn't make a big difference. Accuracy correlates are remarkably consistent, across periods of the forecasting window.

Question: If we look at the process of assessing probabilities, one of the factors you emphasize is open-mindedness, a willingness to reassess one's views. Do you see a different pattern in this process for superpredictors compared to other predictors?

Phil: They are disciplined belief updaters. Supers actually show less volatility in their probability judgments, over time, than regular forecasters do. They update more often, but they update by smaller amounts, which I think is consistent with this idea of greater granularity of superforecaster judgments.

I told you about the old Amos Tversky joke, that people only distinguish three levels of probability. When you look at the national intelligence estimates, the National Intelligence Council, which does the top estimates for the government, they distinguish seven levels of probability.

Our supers are distinguishing about 20 levels of probability. People on the NIC, the National Intelligence Council, are perfectly capable of doing this as well, but they're not using the right measurement tools. When I say our supers are distinguishing about 20 levels of probability, that's the equivalent to about 0.05, 0.10, and 0.15.

How do I know that? We know it because when you round off their forecasts, there's degradation of accuracy. It's not that the superforecasters are making silly distinctions when they're using intervening numbers. They're actually doing a reasonable job of improving their accuracy.

It's a very interesting measurement question of how many levels of uncertainty people are capable of distinguishing in different problem domains. Everybody in this room can obviously do better than three.

Philip Tetlock

The Good Judgment Project (Continued)

In a game like poker, where you get very rapid feedback, and you have a well-defined sampling universe, you can probably get poker experts who can go to 0.5013. Certainly the poker computers can do that. Probably you'd have to be very Asperger-ish to be able to go that far.

In some domains, I think people can become super-granular. In others, granularity is harder to achieve. It's a cost-benefit question for you guys. How valuable is it to you to invest in becoming more granular in a particular content domain? But granularity is definitely a strong correlate of accuracy, much stronger than Asperger's.



Caroline Buckee
Harvard School of Public Health

Caroline Buckee joined Harvard School of Public Health in the summer of 2010 as an Assistant Professor of Epidemiology. Her focus is on elucidating the mechanisms driving the dynamics and evolution of the malaria parasite and other genetically diverse pathogens.

After receiving a PhD from the University of Oxford, Caroline worked at the Kenya Medical Research Institute to analyze clinical and epidemiological aspects of malaria as a Sir Henry Wellcome Postdoctoral Fellow. Her work led to an Omidyar Fellowship at the Santa Fe Institute, where she developed theoretical approaches to understanding malaria parasite evolution and ecology. Her work at Harvard extends these approaches using mathematical models to bridge the biological scales underlying malaria epidemiology; she works with experimental researchers to understand the molecular mechanisms within the host that underlie disease and infection, and uses genomic and mobile phone data to link these individual-level processes to understand population-level patterns of transmission.

Her work has appeared in high profile scientific journals such as Science and PNAS, as well as being featured in the popular press, including CNN, New Scientist, Voice of America, NPR, and ABC.



Caroline Buckee

Novel Ways to Understand How Disease Spreads

Michael Mauboussin: Welcome back, everybody. Certainly one of the fascinating questions of our time is understanding how things, including fads or fashions, investment ideas, and of course, diseases, propagate across a network of interconnected agents.

Our next speaker, Professor Caroline Buckee, is doing some fascinating work to gain insight into this question. As she does so, she has the potential to save hundreds of thousands of lives in the process. Caroline Buckee is an assistant professor of epidemiology at the Harvard School of Public Health. Her work focuses primarily on understanding the dynamics and evolution of malaria and other diverse pathogens.

I first met Caroline at the Santa Fe Institute where she was a postdoc and an Omidyar fellow. She had the opportunity to explore some theoretical approaches to this problem in a very multidisciplinary setting. There are a couple specific things about Caroline's work that really excite me.

First is that by transcending disciplinary boundaries and utilizing novel data sets, Caroline and her colleagues are developing deep insights into the spread of disease. It's the interaction between models, computing power, and data that are improving our understanding of this tricky set of problems.

Second is that this research, as you'll see, has the potential to save lots of lives. The work is, of course, inherently interesting as it'll illuminate the processes by which these things spread, but at the same time, it can truly inform in a very important way, public policy.

This is a cool melding of theory and practice. Please join me in welcoming Professor Caroline Buckee.

[applause]

Caroline Buckee: I'm at the Harvard School of Public Health in the Department of Epidemiology and in the Center for Communicable Disease Dynamics. Actually, we recently participated in the flu forecast tournament for this last season, and members of my group won that, where you have to predict the timing and size of the flu epidemic each year. So we're starting to do these kinds of forecasting exercises as well.

What I want to talk about are the populations that I'm most interested in, which are among the poorest communities on the planet. Specifically, children. And so this handbook [refers to "Where There is No Doctor"] has been translated into tens of different languages, and it's invaluable in the field, where essentially if there's no medical care available, this has the answers to everything you might want to know.

As an epidemiologist, we think on a population level and we try and understand the spread of disease through populations. And so what we're concerned about is how we try and fill the knowledge gaps where there's no data.

In most places in low-income settings, when we talk about epidemiological data, we're really thinking about people getting into Land Rovers and driving through the mud and going and asking people surveys and taking blood and these kinds of things. I think that this is the gap where big data and digital data can really have a massive impact for public health. The danger is that there's too much hype and there's no research question.

I think that the analogy has been made that big data is like searching for a needle in a haystack, and my colleague suggests that maybe it's like searching for one needle among a giant pile of needles, and you need to understand which needle you want and why you want it. Without that, without good methods, you're lost.

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

I'm going to talk about some of the projects that we have to try and understand where big data fits in infectious disease epidemiology. Just to place this in context, I think one of the most important and interesting places that epidemiology is changing is in understanding human behavior patterns and specifically migration and mobility.

The extent of travel that each individual undertakes in their lifetime has changed enormously over the last couple of generations. This study just tracks the lifetime extent of movements in four men from one family for four generations.

So here's the great grandfather, and he went on the order of tens of kilometers. The grandfather went to a couple of counties in the U.K. The father went to regions within Europe. And of course, we're all familiar with these travel patterns – we quickly migrate globally now. What that means is that we are a globally connected world, and infectious diseases that emerge over here can quickly spread.

We're all familiar with this. In 2009, there was the swine flu outbreak which started in Mexico. In the course of a few months, it spread globally. Now, we were fortunate that it wasn't that transmissible and it didn't cause that severe a disease, but the potential for global pandemics is real, and I think that we need to start taking these threats seriously.

Certainly, this is a tail risk type of event, a global pandemic, but, given the frequency of emergence of these types of events in the last decade, I think it's fair to say that this is something that concerns us a lot.

This is happening right now, in the Caribbean. If any of you are going on vacation in the Caribbean, we now have a new virus called Chikungunya that's emerging. We don't know much about this virus. It's spread by the same mosquito that spreads West Nile, so we are at risk of it here in the United States. We had our first reported case in Puerto Rico last week. So, if you're going to the Caribbean, wear bug spray.

This is a real and emerging kind of a problem that affects everyone, but the reality is that, if you look at global mortality due to infectious disease, it's almost all concentrated in low-income settings in Sub-Saharan Africa and Asia and a little bit in South America. Infectious diseases still account for a quarter of all deaths, and, in children, 65 percent.

They really represent a huge impact in terms of the burden of mortality and economics in low-income settings. I show this slide to my students to argue for dynamical models in addition to statistics. When we think about chronic diseases, like diabetes or cancer, you can imagine many people playing their own pinball machines. The expected outcome is the average of everybody's individual game.

There can be stochasticities and different things can happen, but the fact is that your outcome isn't necessarily related to somebody else's outcome. When you're talking about communicable infections, like any contagion process, the outcome of everybody depends on everyone else, and there's feedback in the system. So you need dynamic models. And ideally, you need to understand something about the mechanism of contagion.

I'm going to just walk through the workhorse of infectious disease, the susceptible-infectious-recovered model, the SIR model. As a mathematical model, as we imagine everything as circles, this is a population of people. They're all the same. There's no age structure, and it's a closed population, so we'll imagine that there's no births, deaths, and nobody's coming in and out.

Everybody can be classified in this one compartment with respect to the disease. Everyone is susceptible. We add one infectious person to the mix, and, suddenly we have two categories of infected people. Most people are in this compartment. Somebody has now become infectious for whatever reason, and that person can now transmit the disease to others.

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

This arrow represents the flow between these two compartments, and we can start to build up a dynamic model that describes the change in the fraction of the number of people in each of these compartments. Here, the first person gave the disease to some people, and then these people passed it on, and so on.

Eventually, the original infections, those people become immune in this example, and the dynamic continues. We usually model this as some deterministic differential equations, although there are innumerable variations on that theme. If you write down those equations and simulate the model, you get an epidemic curve that looks a bit like this.

At the beginning of the epidemic, everybody is susceptible. There's an exponential growth in the number of infected people that peaks somewhere. As your susceptible population runs out, the disease dies out, and your recovered population increases and then plateaus. Although this is an incredibly simple framework, and the assumptions are incredibly simple, it actually describes epidemic outbreaks pretty well in a lot of cases.

This is data from a dengue project in Pakistan that I'll talk more about in a minute, but this occurred between August and November in a population that had not experienced dengue before, and it shows this kind of...The colors are just different districts.

You can see this classic epidemic curve. You get exponential growth. You get a peak and then presumably, your susceptible fraction is now being used up and the epidemic dies out. What we observed is the infected class. We don't really have a measure of the susceptible people and you can measure the recovered people, but it's difficult.

Generally, this is hospitals reporting cases. That's what we have to work with. We can fit parameters and then we can understand some of the mechanisms behind the infectious epidemic.

I just want to point out here that there's this perpetual trade-off that mathematical modelers face, in any field really. So both of these are models of cars. They're both toy cars, but this one is very complicated and there's a lot of detail. This one just has the very bare bones features of what a car is. It has some wheels that go around and it has a frame.

The simple model that I showed you is like this. It's transparent. Your assumptions are very clearly encoded, and you know what the mechanisms are driving the epidemic. We can add detail by increasing the simulation detail. We can put people in different compartments, have births in there, have different age structure, have people moving around.

There are simulations for flu in the States where people are literally encoded walking their dogs in the morning, going to work, what do they eat, all of this kind of thing. The question is, do you care about all of that extra detail? Do you want to give up the transparency and generality by having all of that stuff in there?

Generally, the epidemic models that are simplest often provide the most insight. The simple models can lead us to some interesting threshold parameters. The basic reproduction number is this idea that's very general, that if, on average, the first infected person infects more than one other person, then your epidemic will spread. And if they don't, then your epidemic will die out.

There's this threshold parameter R_0 , that's one of the stock parameters that we try and measure for an epidemic to figure out the potential for onward spread. Of course, this has analogies in Internet meme spread or product adoption and many other situations.

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

We want to try and think about what are the components of R -naught? How can we measure them and what are they that make something epidemic or not? How can we use that to predict how something's going to spread?

This first parameter is, OK, if somebody coughs on me and they have a virus, what's the probability that I become infected? That's actually very difficult to measure and you have to know something about the inherent biology of the pathogen.

We have the duration of infectiousness, so the probability of transmission given contact is going to increase this rate of flow. The duration of infectiousness determines how rapidly this compartment is diminished. We can measure this for most pathogens pretty well, once we have a reasonable sample size and we know what we're looking for.

Then, we have this critical parameter, the probability of infectious contact. So that data relies on both infected people and susceptible people and how often they mix.

Generally, what we are asked by policymakers and what we think about, is how can we reduce R -naught below one, stop an epidemic? How can we intervene? We can intervene on the infectious contact rate by vaccinating people. This is of course, being one of the classic, public health measures that's made a huge difference to our life spans.

This is modeled and the simplest model is just by moving people from the susceptible compartment straight to recovered. Now, sometimes, if it's an imperfect vaccine, it doesn't necessarily move them to the recovered compartment, but it can reduce the probability of transmission given contact. That's also useful.

We can also quarantine people and this happens naturally sometimes if the symptoms are bad enough that you stay home. Your contact rate with other people is going to reduce anyway. We can use education programs and other ways to reduce that contact between our susceptible and infected populations. Then, we can treat people.

If we can identify the infectious reservoir and treat them, we can reduce the number of people in that category and we can therefore reduce the spread of the epidemic.

One of the biggest problems that we have in using these types of basic models to understand the spread of infection is that we are lacking data on how people move around and therefore, how these contact rates vary spatially.

So we look here at this study showing the spatial and temporal determinants of infection. Here we have spatial scale from the neighborhood scale up to within a city, regional, and then all the way up to international. And here we have temporal scale. These are daily movements, periodic movements, seasonal movements, very important in sub-Saharan Africa, long term and migration.

The impact that they have on the spread of infection ranges from just exposing people within their local communities, all the way up to global spread and pandemics. But in between here, we have regional spread and all of the meta-population dynamics that underlie epidemic spread through a country and a region.

So, where's the data gap? We have data from small-scale survey studies where you go out to a village and you say, "Where did you go yesterday? Where did you go last week? How often do you go there?" and these kinds of things. That can tell us something about the daily exposure people have in their communities.

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

We can also strap little GPSs to people and then see where they go every day. People do that, but it's limited in scope, so you can't really strap GPSs to millions of people, unfortunately. I would love to.

[laughter]

We have census data and we have micro census data, and these tell us about long-term movements and migration. Generally, the question is, "Where did you live this time last year?" That's fine, but really the temporal resolution is not on the order of magnitude that we need to understand disease.

And then we have airline and shipping networks. We know kind of global connectivity through airline routes, but of course, those populations are going to be somewhat different than the populations driving dynamics at the regional level.

In comes big data and we're all really excited. You've probably seen some variation on this graph. This is the number of mobile phone subscriptions over the last decade, in the billions, and of course the biggest increases have been in the developing countries.

Where are we? Suddenly, we have billions of sensors on a lot of people in regions where previously, we would only be able to go and find out about those people if we got into a Land Rover and we drove there and we had teams of community health workers and so forth.

What we're seeing from our own research is that the penetration of these subscribers in rural places and low-income settings is astonishing. I think that soon, we will have a really amazing representative sample among mobile phone data sets.

Every time you make a call or a text, the mobile phone operator logs your phone number, the person you called, and a cell tower ID. If you know the latitude and longitude for that cell tower, then you have a location for that person at that time. More specifically, you have a location for their SIM card. That's not always the same thing.

But let's imagine now we have these records over time for each person. Here's our person, making some calls at Tower A – we know where Tower A is. Then that person moves to Tower B, makes calls there, we can infer movement between locations. Finally, they move to Tower C.

We have a longitudinal record of that person's location over time. These logs are kept routinely by mobile operators for churn analysis and other things. They provide one of the most amazing data sets for the dynamics and location of populations that we've had to work with ever.

If you know something about the populations in those places – we use mainly satellite data, other types of remote sensing information – we can delineate population distributions and densities from them, and then we can build up models of the flow and dynamics of human movement between each sub-population.

That's incredibly powerful, especially given the fact that...People always say to me, "That's good, but you don't really know how biased it is." Yes, it's probably better than zero, which is what we had before.

We're working hard, actually, to figure out how to adjust for biases and ownership and think about ways that we can figure out who it is that we're capturing in these types of data and especially, the movement of children is obviously going to be important.

It's a pretty remarkable data set. Suddenly, we have this whole area. We have an amazing source of information for this whole area. We can get daily or even hourly locations for people. In cities, at least, we can get resolution on the order of a city block. We know where everyone is, we know where our susceptibles are. If we have good clinical data, we can start building really good models.

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

This is really exciting. Most of my community is really excited about this opportunity. What have we been doing with this? Here's Kenya. This is a map that shows the road networks in Kenya. You can see they follow population density patterns pretty well. Kenya is sparsely populated in the North and pretty densely populated in the Central and Western regions and on the coast.

What I've mapped on top, the color on top of the road network here is an average measure of mobility. High, the red, is people that travel a lot and they travel far. The blue is people that don't travel very much. What you can see is that people in these poor regions have to travel pretty far. There's been a long debate in the development world about the burden of travel in rural areas and the impact of remoteness on people's ability to access markets and healthcare and things like that.

Now, we can actually observe these people. We can see whether it's true. Is it true that people in rural communities travel further? It seems like it is, especially if you can couple it with other types of information.

What we did is we wanted to see whether these kind of average aggregate measures had any bearing on healthcare. We took the standard state of the art measures where you map onto the roads, the average time to a health clinic. Here, this is this area in Western Kenya. You can see that there are centers where the travel time to health clinic is short, and then there's areas around them where it's a little bit harder to get access to healthcare.

We wanted to ask the question, are our estimates better than this for predicting health outcomes? Here are our estimates from the mobile phone data and you can see that even in places that have quite similar remoteness in terms of their travel times to health clinics, we have quite heterogeneous average behavior in terms of mobility.

In some sense, this is capturing more than simply a travel time, a static measure. We're looking at dynamics, and we're looking at something mobility probably captures more about the fundamental behavior of people.

What's really interesting is that our measures are a significantly better predictor of which households are missing antenatal care and childhood immunizations. We can layer these kinds of maps onto current policy guidelines to figure out where you need to target your healthcare access initiatives. I think that's pretty exciting.

We can also look at seasonality. People obviously travel very differently in low income settings compared to here. We have a ton of data here about commuting and travel around holidays. There's been this kind of dogma that school closures and school terms basically define our flu seasonality and/or measles seasonality. In low income settings, it's much less clear, because there's a lot of agricultural migration and there are different reasons for people moving around.

This, again, is Kenya. I'm just showing a measure of flux, population movement. This is Christmas, and it's a weird scale because of our data set. Lots of movement around Christmas and then two more peaks. Working together with the researchers at Princeton who are trying to figure out whether Kenya should institute rubella vaccination, we're talking to them and they were like, "We can't figure out these three peaks of rubella that happen in Kenya. We don't know why."

Here's rubella, and you can see these three peaks that keep happening. It turns out that our mobility estimates are better at predicting those rubella peaks than either school terms or rainfall and agriculture. Because we have this actual direct measure rather than proxies like school terms or travel times, we can start actually understanding the mechanism and understanding the dynamics that drive these types of diseases.

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

Those are large scale flux in looking at aggregate mobility measures. We can now bring our SIR-type frame works to bear on this kind of data set as well. This is a picture of a meta-population model. Again, it's just circles, because that's how everything looks to me. It's meant to be these linked populations with different risk.

This is a high-risk population, and these are low-risk populations. We have movements between them that we can measure. Then what we want to ask is, how much imported infection is there going to be from this place to our low-risk settlement? And when our low-risk settlement guys go on holiday or they visit this high risk place, how often are they going to get something and bring it back?

Within each of these populations, we want to be able to parameterize a little SIR model at its most basic and add onto our infected term a constant risk of importation from the other one. So we can build up these networks connected by mobility and use the mobile phone data to parameterize the movements.

We have a high risk and a low risk population, and we want to know what's going to happen. Whereas before we had very crude estimates based on standard hospital reporting requirements and estimates from census data and things like that, now we have the opportunity to collect real-time data on infection rates using syndromic surveillance – that's basically Flu Near You and dengue, these types of new mobile apps where you can engage with people and ask them about their symptoms. We have great climate data now. If anything that's environmentally forced, we have huge amounts of data that we can integrate to adjust our risk.

Then we have some remote sensing technologies that allow us to get really good estimates of the boundaries of settlements and their sizes and densities. We want to ask the rate of spread. Is it going to come to this place? How quickly is it going to come to this place? What can we do about it? How can we target our resources? How much is it going to cost?

The way to parameterize those models now is through these huge data sets that we have on where people are going and how they're mixing. Another kind of big data thing that's really transforming the field is also parasite genomic data, pathogen genomic data. We also have sequences from all of these viruses and things. We can say how related are the sequences here to these ones. Can we measure importation that way as well? All of these new data sets need to be integrated in a rigorous way to come up with some of these estimates and to make better predictions.

What's interesting is that our mobile phone data seems to also be slightly better at predicting population densities and distributions. It correlates very well with our standard estimates of where everyone is. And especially in places like Pakistan, it seems to do a really good job of counting people that are known to be missing from the current data sources.

I'm going to now talk about two examples. My first love is malaria.

[laughter]

It's a very interesting eukaryotic parasite. It's spread by mosquitoes. It's been with us for many, many years. They think that King Tutankhamun died of malaria. Lord Byron died of malaria. It's been with us for a long time, and it still kills a child every minute.

The sad fact is that it's completely treatable and it's completely preventable. Every single one of those deaths is unnecessary. And we know how to fix this. We do understand malaria control. We have some tools, they're not perfect. We still have this burden, and I think it's increasingly unacceptable.

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

If you have malaria, it's pretty horrible. You'll get fever and chills, very severe headaches. Your muscles and bones ache. Children that we see in the clinic die of severe malarial anemia, often coupled with malnutrition. They die of respiratory distress, and they die of coma. This little girl had a coma. What happens is the parasites get sequestered away, stuck in capillaries, and then they go into metabolic distress and die.

The communities that malaria affects are the poorest, most remote communities on the planet. They have poor access to healthcare. Malnutrition is a big problem. Co-infections are a big problem. It's been shown in many cases that if you can eliminate malaria from a community, you instantly improve overall health, education, and the economic opportunities.

I just think it's a really urgent problem to be worked on. We have made some progress over the last 100 years. This is the map of the extent of malaria in 1900, sub-Saharan Africa, but all the way up throughout the U.S., we had quite a lot of malaria here, and into Europe, Mediterranean, and all across into Northern Australia.

In 2007, we've now shrunk that map considerably. Malaria now remains a problem of the tropics and really lower sub-tropics. Sub-Saharan Africa is still a big problem and South Asia, Southeast Asia, and South America, but of course, we eliminated it from the U.S. and from Australia. That's in large part due to a global malarial eradication program in the 1950s.

This is from a U.S. textbook. They sprayed DDT everywhere. It worked really well. But in the 60's, basically what happened was money ran out. They didn't really consider it viable in Africa. There was resistance to DDT and there are all kinds of logistical problems and the program was totally stalled and they gave up.

For decades after that, basically, all of the emphasis was on control – control morbidity, control mortality, do what we can to treat and provide healthcare for people to stop children dying. But forget about transmission, we can't eliminate it.

Then, this happened in 2007, Bill and Melinda Gates had the audacity to say the word “eradication” and everyone was up in arms. But by putting it back on the agenda, it's really shifted how the malaria community thinks about this problem. If we go back to our SIR problem, you'll see that the basic model is still almost exactly the same with a few modifications.

In fact, this model was used to great success in the first eradication program. Now, again, we have our people, they can be susceptible and they can be infected. We have no recovered class because with malaria, you never really become immune to infection even though you become immune to disease symptoms. So you just shuttle back and forth between susceptible and infected in the most simple model.

Then, we have to model mosquitoes also. Mosquitoes can also be susceptible. They can have a latent period which turns out to be very important, and then they're infectious. This latent period is close to the lifespan of a mosquito which is why vector control works so well because it kills them quickly.

Then, of course, these two sets of equations are linked by the contact between people and infected mosquitoes. So, feeding rates and how inherently infectious we are in the relative density of mosquitoes and humans.

How can we intervene? There's no vaccine, and the vaccine that you may have read about that's been rolled out recently is a pretty crappy vaccine from the perspective of transmission blocking. It seems to have an impact on severe disease, which is good. With insecticide-treated nets, we not only block

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

mosquitoes from feeding on people, thus reducing the contact rate, we also increase the death rate because there's insecticide in there. That's a good thing to do.

If we treat people, we shove people back into this group. It may not have a huge impact on transmission, but at least people get better. And at very low prevalence of infected people, it's probably helping overall transmission efforts. Then, with vector control, the other goal is to just reduce the density of vectors. That's straightforward.

We have this kind of suite of tools. They're not perfect, but in combination, they work in non-linear ways to actually be quite effective. The problem is that nobody is ever immune to infection and the vast majority of cases have no symptoms at all, so you don't know if they're infected. Plus, the diagnostic tools that we have right now aren't sensitive enough to actually measure those very low density infections that are nonetheless infectious to mosquitoes.

If you want to eradicate malaria, if you actually want to get rid of malaria forever, you have to think about these people because they are the ones that are traveling around spreading the infection and undermining control programs without even knowing it, and we can't target them very easily.

What are we doing about this? This is Kenya again. In the background, you can see the malaria risk. The red is high risk around Lake Victoria. We have high risk on the coast. And then we have fairly low risk in the middle of Kenya, although some transmission.

These gray blobs in the background represent our satellite estimates of the distribution and location of populations. Then, the black and blue dots are the location of cell towers. You can see again that the cell towers follow population density. Our location estimates are best in places like Nairobi where we have a very high density of cell towers, and they're less good up here in the north where there aren't that many people.

We have ways for adjusting for this, but that is a concern. In any case, we have all of our mobile phone data and what we want to ask is, where do we focus interventions? For a start...Kenya doesn't have a lot of money, but they want to eliminate and so, how do we do that efficiently? How long will it take? That's critical to the cost and whether it's even feasible to consider elimination.

If drug resistance reaches the coast, how quickly will it spread? I'll show you a slide in a minute, but all drug resistance to malaria drugs has emerged in Southeast Asia, spread to East Africa, and then spread across the continent through human travel. At the moment, there's one drug left to which there is no resistance and it's the first-line therapy for severe malaria.

There's drug resistance now emerging in Southeast Asia. If it reaches Africa, we're in real trouble, and there's going to be a huge amount of mortality among children. This is a real concern and we want to be able to predict how that's going to happen and which regions are at the highest risk for resurgence of disease if we do fail at elimination. That's an important problem, too, because in the last eradication program, when it failed, a bunch of places had massive epidemics afterwards.

These are the kinds of questions that we want to be able to answer. What we've done is we've built up our mathematical models. We've taken all of our layers of data and then we've modeled the spread of parasites in people traveling around. We've done a kind of a source-sink analysis: Where are people coming from and where are they going to and where are parasites coming from and where they are going to? Those are the different questions. But if you look here on the left, you can see that the source of human travel is kind of scattered around these high density places. Nairobi remains a sink for travel, so the sheer volume of people traveling into the capital is amazing. It's a sink from all across the country.

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

If you look at the parasites – so this is kind of the mathematical model imposed on our mobility matrices – what you can see is the vast majority of malaria in Kenya is coming from one region and spreading this way. Nairobi is a huge sink, but also this highland region and these areas that have low transmission, but they still have mosquitoes.

What this suggests is that instead of going for low-hanging fruit, where you're continually mopping up these imported infections, you've got to address this problem, put all your interventions there, and this may well just go away. It kind of changes how we think about resource allocation, and it provides direct predictions about how you can impact transmission. It provides some kind of specific policy guidelines.

In Nairobi, because the estimates are so finely grained, we can go a step further and we can ask, where are the likely hot spots of local transmission. So in Nairobi, it's been historically thought that there's no local transmission at all because it's too cold, although we know that the mosquitoes are there and Nairobi used to have malaria back in the early 1900s.

Using our model, what we did is we took these blobs, which represent hospitals around the city, and we compared their clinical cases to our predicted cases from our modeling exercise.

The color of the blob represents the ratio of observed-to-predicted imported infections. And the idea here is that if you predict basically no infections, but you're seeing tons of clinical cases, then when you want to find out where local transmission is happening, those are the places you should go.

If you actually look at where those places are, it's in the slums surrounding the city, which is maybe not surprising, but it gives some evidence to back up teams of people going out there and looking for vectors. We're working with the Kenyan CDC and other researchers in the region to try to figure out how to integrate some of these types of approaches into their policy guidelines.

I think the biggest problem for malaria research is this, as I said. Cambodia now, we have confirmed drug resistance to artemisinin-based drugs. That's kind of a disaster. So there's a lot of emergency action going on. One of the biggest questions is, how quickly is it spreading? What is the impact of Myanmar opening up on the emergence of drug resistance, and where will it go next? Where should we be looking for it?

We know it's got as far as Bangladesh. And because of the high population densities in Bangladesh and India and the frequency of travel to the Middle East and Africa of migrant laborers, this is a huge problem. When we talk about prediction, they should probably have a forecasting tournament for predicting the spread of drug resistance, but they're scrambling to do anything right now. This I think is a very urgent problem.

Here, this is just a model that my collaborator Andy Tatem put together to look at where we think the parasites are going out of that region. It's basically everywhere.

Now, I'm going to end by shifting from a problem that's been endemic for thousands of years in human populations to this issue of emerging epidemics. I'm going to illustrate it with dengue because the underlying model is essentially the same as for malaria with different parameters. We don't know as much about dengue as we do about malaria and there are some key differences that I think underlie how worried people are about this, specifically in Washington.

Over the last 50 years, we've seen the emergence of dengue cases, both spatially and in the number of cases. It's a virus that's spread by a different vector than the malaria parasite. It's spread by the West Nile virus-carrying mosquito, *aedes aegypti* and *aedes albopictus*, actually. This mosquito bites in the day and that means bed nets are a no-go, which is problematic.

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

There is no treatment for dengue. There's a vaccine that's in clinical trials. There are some troubling aspects to it, like it might cause more disease, which is bad. But essentially this vector is an urban vector. This is Singapore and they have problems with dengue, increasing problems with dengue.

Because it's urban, the massive increases in population densities in these mega cities across Asia are really problematic in terms of the future of this virus. Whereas malaria is a rural disease, dengue is an urban disease and it's very, very hard to get rid of the vector because it basically breeds in puddles and tires and trash, so it's almost impossible to get rid of it.

The other reason they're worried is because of course, we have these vectors in the States. Half the world is at risk of...If dengue spread, it would be half the world would be affected. There are many places locally that have high suitability for dengue transmission. And so, I think the possibility for a major global problem with regards to dengue is very real.

This is an example of how infectious disease epidemiology can be challenging. So we are working with Telenor Group, it's a large mobile operator, Telenor Pakistan. We talked to them because in 2011, like I said at the beginning, there was this big dengue outbreak in the northeast of the country where they've never seen dengue before, and they think that basically the tire industry transported tires across the country and carried the mosquito with it.

So now the mosquito is everywhere in Pakistan. There was this huge outbreak that caused over 20,000 clinical cases, so probably many more mild and asymptomatic cases, with a focus in Lahore. And then in 2013 there was another outbreak. And so we were talking to Telenor Pakistan and trying to make the case that CDR [call detail record] analytics could really help.

The idea here is not so much like in the case of malaria, where we'd be working with policymakers to try and like shift policy incrementally, but rather the idea here is, can we use CDRs in relatively real time to make forecasts that can be acted upon? And can we work in partnership with a mobile operator who also has the ability to engage with those populations and reach them through education messaging, which is really one of the only tools we have to address these types of epidemics.

My colleagues at the University of Peshawar helped us with accessing the data and giving us the local expertise that we need. This took a year of organizing, signing NDAs, getting people on board, and writing to Ministries of Health. Then, we had to go to Islamabad and spend ten days there, getting data, and not looking at it because of the privacy implications, and the data couldn't leave the country which is fine. These projects are very bulky, but they have the potential to be really transformative.

Here's the 2013 outbreak. The colors again represent different places in Pakistan. And so there was this small, mini-outbreak that happened at the beginning of August in the northwest of the country. Then, it spread again to Lahore. Then, we have this other outbreak, but it's much later than the 2011 outbreak, it's October-November.

The question is, what can CDRs add to our predictive capacity about dengue in Pakistan in the future? This is preliminary stuff, but just to give you an idea, here's a picture of Lahore. The red dots are cases of dengue, and we have them geo-coded pretty well. And the black dots are cell towers, so we have really great resolution on these and we have three hour chunks of where everybody is. It's like eight million people.

This is the largest data set ever to be interrogated in this way, and I think the first time that anyone's tried to do this for an epidemic. The green dots are the hospitals where the cases were reported, and there's

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

been anecdotes that you're at higher risk near the hospital. We want to check out that. So this is the data that we're working with. You can see these clusters of towers that correspond to villages.

What can we see? All right, so we also got all of the weather data for all of the weather stations across Pakistan. Here's the temperature one for Lahore. You can see it's hot and then tails off here. At some point, it's going to get cold enough that the vectors are no longer able to survive.

Here's rain. There was a massive flood – rainfall and a flood associated with it – in August. We think that that kicked off that first peak, and then it goes down a little bit. Here's dengue cases. Here we have this peak, and then we have this very pronounced second peak. And the question is, what's driving this? Can we measure it with human behavior measures?

Now, here's the first pass CDR analysis. So we're looking at movements in and out of Lahore here and what you can see, July to August, everyone leaves and then comes back. This is Ramadan and it's amazing. Everybody does something completely different during Ramadan, which we saw in Kenya to some extent over Christmas, but this was much more pronounced.

A working hypothesis is that everyone moves around and there's flooding here. There's a time of like a month between one case and the next case, so we think that this flooding probably caused an outbreak here, Ramadan spread it around, and then we see this increase in the aggregation, basically. The number of susceptible people in Lahore is growing. We've adjusted this for increasing numbers of subscribers and stuff like that.

These two factors, people moving around for Ramadan then coming together, then aggregating, have made the conditions perfect for this kind of outbreak. We're working with Telenor and researchers in Pakistan to try to figure out what these types of approaches can add. Can we do a really good job of prediction in real time?

If there was an epidemic – and they've got a polio epidemic right now, actually – and we had CDRs every two weeks, we could say people moved here, here, and here. Live times are like this. This is where the cases are. You need to put surveillance in these five places and you need to send educational messages to all of these people about these specific things.

It's not simply do more vector control, it's when you travel, make sure you do this. If you are going to the tire market, wear long clothes. Clean this particular type of water container. These are very practical things, but they are the tools that we have and I think that this is really a chance to impact these epidemics and stop them potentially.

I'll end by basically saying that malaria kills a child a minute. It's I think pretty unacceptable given that it's treatable and preventable. That's among the poorest communities. But these pandemic threats, they're very low probability events, but if they happen they're going to affect everyone in this room. They're real, and so I think that now we have the technology, the methods, and the data to start actually having impact on these places.

And just as I kind of pitched, this work isn't funded – it's incredibly frustrating – and so if anyone would like to support us in this, I really think we have a chance to make a big difference to a lot of people. So I'd love to hear any questions, and thank you for your time. Here's my details.

[applause]

Question: Can you talk more about how the patterns change around Ramadan?

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

Caroline: Well, it's interesting. Ramadan happens at a different time every year, and so it's not necessarily that it's consistent. What we're doing is we're trying to look back and say, "Is it predictive?" and the specifics, not just that flux causes disease, but rather, does it cause disease in a specific direction and can we make a mechanism that we can use to make predictions when we see particular types of movements in the future?

We have one year's worth of data. I don't know how much inter-annual variability there is, and so it might be very regular, it might not. Continuing data analysis and continuing methods development is going to be required to figure out, do we really need monthly CDRs, is this enough? Can we just now predict dengue?

Again, the nice thing about this data is it's low cost. We're not going out and trying to do surveys every year, which is incredibly expensive. If we can come up with a plausible mechanism that turns out to be predictive, then that's great. My working hypothesis is that it's going to be the aggregation of movement, climate variables, and the specifics of how the outbreak happens each year.

Question: It seemed like when you looked at that slide you said, "OK, this is the source." As you said, this is where you're actually concentrating. What's stopping the Kenyan Government from going to the source and wiping out the mosquitoes, or whatever it is that they could do to...?

Caroline: Because it's kind of the reverse of what people tend to want to do. That's not an easy win. You'll notice, actually it's on the border with Uganda, so what I didn't talk about is cross-border migration, which is a huge issue, especially in the Greater Mekong region. I think that we have ways to deal with that, because quite often, operators log handset numbers. Then we can actually, regardless of SIM card, we can start getting at that.

I think more than that, there's this issue that it's just huge inertia and a difficulty getting things done. I think that's why these partnerships with the operators is great, because we can say to them, "Let's do a randomized control trial where we try sending text messages to people that we identify as high risk, and asking them to use a bed net when they travel."

We can do that. We can do it quickly and scalably and get results and see if it works. I think ministries of health, they're very slow, cumbersome monsters.

Question: What is your opinion on DDT?

Caroline: Oh, DDT, yeah, so I think the consensus is that it has its place. It's incredibly useful in some situations. DDT resistance is a big problem. So, at the end of the eradication program, there were some populations where like it didn't work at all, the mosquitoes were just resistant, so there's a big push to actually come up with novel types of insecticide and different ways of doing things.

I worked with a colleague who, she's found this hormone, if you put it in a bed net, it basically stops the females from laying eggs, and so you could imagine all these new, different ways of manipulating the vectors. There's GM [genetically modified] mosquitoes of course.

DDT has its place. Resistance is the biggest issue I think, at least in sub-Saharan Africa. But I definitely think it's not a kind of all-or-nothing kind of thing.

Question: Can you expand on infectiousness and how it relates to different epidemics? What is the probability of transmission given a contact event?

Caroline: OK, it's a difficult thing, right? I think that's...The infectiousness is probably less important than the mode of transmission and the way it affects people. So, with SARS [severe acute respiratory syndrome] for example, we were kind of lucky because you get symptoms and then you're infectious. If

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

there was something that happened where you're infectious for ages and then suddenly you die – kind of like HIV, actually, you're infectious and you're transmitting and you may not know it – that's really problematic, and that's how you spread something really fast. So, it's not necessarily the per contact transmission rate that's the most important thing. It's mode of transmission and the kind of epidemiology of that particular bug.

The real worry for these epidemics, dengue is one, but the zoonotic reservoir is a big issue. There's a map that a lot of people I collaborate with put up, where if you draw a four hour flight around Ho Chi Minh City, you capture a quarter of the world's people and 75 percent of the world's chickens.

Like flu and other viruses that live in birds, ducks, chickens, poultry that people live in their house with, there's a huge reservoir of viral pathogens and live animal markets in these types of places. And so the uncertainty lies in when those kind of zoonotic events are going to happen. But the inherent transmissibility itself is not the biggest issue so much as the epidemiological characteristics, I think.

Question: How concerned are you about severe health risks?

Caroline: That's difficult to answer. A severe epidemic...I think the other issue here is who it affects. These tail-end events we're talking about like "all of you guys are dead" which people tend to care about more than like "African babies are dead." The problem is that, there are horrible diseases that happen all the time and have been happening forever that are really bad, but because they're not epidemic and because they don't affect us, they're perceived as less bad.

In terms of severity, you can get the whole spectrum. In evolutionary terms, there was this old hypothesis that all pathogens should evolve to be asymptomatic. They shouldn't cause symptoms because the most successful one wouldn't kill anyone, because you're basically cutting off chances for transmission if you kill people. We know now that that's not really true. There's no kind of global pathogen agreement that that's a good idea.

Individual pathogens are simply evolving by natural selection, so if they work, they work, and if they kill some people, then they'll balance that with how good they are at transmitting. There's this idea that there's an intermediate level of virulence that has evolved to be stable, if that answers your question.

That's part of the reason that an emergent thing is problematic because it hasn't adapted with humans for a long time. It can just rip through the population and kill everyone. That's what happened with the Bubonic plague and Spanish flu and these kinds of things. There was no immunity.

Question: How worried should we be about the Ebola virus?

Caroline: Well, I would if I lived in West Africa. Yes, so, when you say we, do you mean...?

Question: Well, if it gets on a plane, for example.

Caroline: It's not clear how transmissible it is either. So what we're worried about is this combination of highly transmissible and highly virulent. And that's these kind of classic examples of epidemics. That's what they were. Measles is highly virulent and highly transmissible, but we're kind of used to that.

Question: You mentioned that Gates used the word "eradicate" and how that was kind of controversial. Do they have a different view of the dynamics of the disease or does it have implications for the kind of research...Were they funded or what's the story around it?

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

Caroline: The story is that, the whole community, there was basically no research into transmission, and we were just going to live with this burden forever, but we'd try and manage symptoms. That was the kind of general framework of where we were.

They, Bill and Melinda Gates basically said, "No, that's not acceptable. We should be able to eradicate this. It's the kind of burden where the economic impact over a long time is just enormous, so let's put in effort now, eradicate it, and no longer have those costs."

It was controversial because the entire community was like, "Well, it's not possible." They sort of challenged that idea like, "It is possible." The classic thought experiment is: It's treatable, right, so if we had like World Take Your Malaria Pill Day, and everyone took it at the same time, maybe once a month for three months, there would be no malaria. It's possible.

They kind of put that idea back on the table and because, frankly, they have so much money and there wasn't much money in malaria, they can dictate kind of how science goes. It's like, "Great, we'll work on transmission."

Question: Less on pills.

Caroline: Well, there's been such an emphasis on severe disease and pathogenesis and those kinds of things, that there's been some really understudied biology, there's been a bit of a shift back, and towards basic tool development, like field tools that scientists tend to prefer fancy molecular stuff. Yes?

Question: Were the vaccines difficult to develop or is that they immune to change?

Caroline: Both, so we don't really understand immunity to disease, like why some kids have this...Two kids will have the same number of parasites in their blood and one will be playing soccer and one will be in a coma. We don't understand why.

The parasite is enormously genetically diverse, so that's part of the reason why you never get immune because you get a different strain every time. One issue is that your body never sees the same malaria parasite twice.

The other issue is that we just don't seem to be very good at making immune responses. So, the parasite manipulates the host immune system and we don't make very effective responses against the things that are shared by all malaria parasites. So that would be the ideal. You have a vaccine where the target is something that all malaria parasites have, but we don't make good antibody responses to those things. Yes?

Question: Do you have good access to hospital records?

Caroline: Clinical data is actually the biggest source of the big data gap, ever. We can't get good clinical data. We actually...For Kenya, there's been so much research there for a long time and so we work with the Malaria Atlas Project.

They've been building high resolution maps using quite sophisticated kind of geo-spatial techniques to take a bunch of cross-sectional and longitudinal studies and smooth the estimates to come up with a one by one kilometer grid of the estimate of the fraction of infected people.

We use that to basically assign our probabilities of infection for our travelers, and then we kind of run it that way. Similarly, you can think about, "Well, what's the kind of disease rate there? Do we know anything about the demographics?" We can put all of those things into the model. What was the other part?

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

Question: Do you have clinical reports over time for infected people?

Caroline: Oh yeah, no. Definitely not at this stage. In Nigeria, we're working with Malaria No More. And in Nigeria, they've just made a law that all drugs have to be authenticated through one's mobile device. There's a thing you scratch off, you text it in to the authentication service, and they text back if it's counterfeit or real.

That means that we should be getting...We should start getting individual data on like how many episodes did you get treated for at least.

We've talked to them about sinking back-end, like trying to preserve anonymity while also sending intervention, educational messages to the correct people. It's really hard and I think at this stage, it's still very much in its infancy, and nobody's figured it out. There's been so many little things tried and nothing's really worked. Nothing's really scaled. I think it will, but it takes time. Yes?

Question: Can you talk more about using insecticides? And, if you could create a genetically-modified mosquito which presumably could wipe out other mosquitoes, what are the issues associated with that? How plausible is that?

Caroline: I think, yes they would. It is quite political, because getting rid of a whole species of mosquito, I would think it would be great, but there are lots of really interesting technologies here that are still in their infancy.

One really interesting one for dengue control is that you infect them with wolbachia, which is a bacteria, and the wolbachia effectively makes a mosquito sterile and it passes it on. So you're spreading sterility through the population...There are lots of ways to make a lot of males, and then it's obviously useless, so...

[laughter]

Then, there are all these different techniques. I would say that they're exciting, but they're in their infancy still. I do think that they would be used if the trials and the field trials seem to be productive. In fact, there are currently field trials in Northeast Australia of the wolbachia infected dengue vectors, and they released them in the same place that they released the cane toad and the rabbits. I mean, it's really terrible, but anyway those are exciting and I think they're real. I think they're interesting.

Question: Is there anything that really surprised you when you sorted the data?

Caroline: Like in the Pakistan data, I was really surprised how often people were calling. Like a lot and so that's great because we have finer resolution estimates, but it's kind of weird to me. I think the sheer volume of travel in and out of big cities in these low income settings is amazing.

From standard models in the kind of economics, people often use gravity models, which approximate the number of people that travel between places. They're just off by so much. So the standard models that we've been using to think about the spatial dissemination of infections are not doing well, and this was the first time that we could actually see it and see how.

These centers, these sort of megacity hubs, are incredibly important for the spread of infection and all of the social and economic repercussions. And so that was one of the interesting things.

Question: What other major bodies are looking at research in contagious diseases? Are you working together?

Caroline Buckee

Novel Ways to Understand How Disease Spreads (Continued)

Caroline: There are very few groups that actually have data, because it's incredibly difficult to get. Actually, Flowminder is a kind of a loose network of academics, and we kind of try and pool our resources and share data so that we can leverage...We have data from Namibia, Rwanda, and Kenya. We're trying to build up our data sets.

The problem is that it's really hard to get buy-in from operators, and so very few people actually, they're not interested in giving you the data, and it's very hard to access it. There's really very few people using CDR analytics in particular.

There are groups across the [United] States and in the U.K., mainly, who do a lot of kind of theoretical work on this and look at spatial transmission using mathematics. It's quite a small, close-knit community, so we work with them a lot, like the rubella group, that's the Princeton group. I work with people at Oxford a lot.

The nice thing there is that they have far-reaching networks with clinicians and field sites across these places, and so it's really about bringing together everyone, figuring out how to get the data to the right people and do local capacity building. I don't know if that answered your question, but I would say it's a pretty small group at this point.

Question: What do you think about the enthusiasm surrounding big data and other tools?

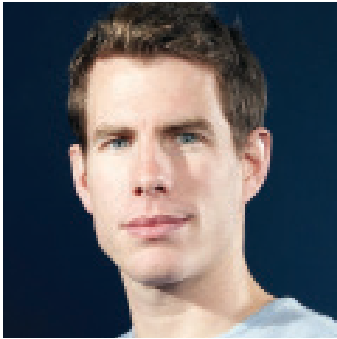
Caroline: There's obviously been a massive amount hype around big data, especially big data for development. My problem with it is that it's all been top-down, so it's like, "Yay, big data! Let's solve the world. Send in the data scientists."

When in reality, it should be we need to figure out where drug resistance is going, this is what we need to do that, how can we find the right data in a rigorous way to make it happen.

I think there's been a kind of over-enthusiasm for participatory surveillance, crowd-sourcing everything, big data, and even this idea that you can do hypothesis-free science now.

I disagree with that concept, and I think what has to happen is this kind of, once the froth has all died down, you'll be left with people who are like, "Yeah, I still want to cure malaria." The Google Flu Trends thing, I think, is a harbinger of things to come. I do think that participatory surveillance can be used very effectively.

We need to understand it better. We need to do a better job of rigorously linking it to real data. We need to do validation and testing, and we need to be a bit more honest about the capabilities and how it can impact the actual control.



Sean Gourley

Quid

Sean Gourley is a physicist, decathlete, political advisor, and TED fellow. He is originally from New Zealand, where he ran for national elected office and helped start New Zealand's first nanotech company.

Sean studied at Oxford as a Rhodes Scholar where he received a PhD for his research on the mathematical patterns that underlie modern war. This research has taken him all over the world, from the Pentagon to the United Nations and Iraq.

Previously, Sean worked at NASA on self-repairing nano-circuits and is a two-time New Zealand track and field champion. Sean is now based in San Francisco where he is the co-founder and CTO of Quid, an augmented intelligence company.



Sean Gourley

Augmented Intelligence

Michael Mauboussin: I'm pleased to introduce our next speaker, Sean Gourley, who splits his time between doing mathematical research and working as a co-founder and chief technology officer for Quid, which helps decision makers digest and understand public information.

Sean has done work in two areas I believe are fascinating and highly relevant for this group. The first is using mathematical and statistical techniques to understand patterns and regularities in the social sciences. His work is focused on modern war, but we've seen similar work on initiatives, for example, in the realm of cities. This work addresses fundamental questions about how we can model social systems at a macro level.

The second area, which will be much more of the focus of today's discussion, is around the concept of augmented intelligence. We do know that humans are good at some things, such as pattern recognition, and bad at other things, such as rapid, massive calculations. We also know that computers are good at some things and bad at others.

The challenge of augmented intelligence, or the opportunity of augmented intelligence, is to figure out how to combine the best of what humans do with the best of what computers do in order to make thoughtful predictions, while of course avoiding the worst of what both humans and computers do.

One final note about Sean is that he was New Zealand's national champion in the decathlon twice; 2000 and 2002. As you know, the winner of the Olympic decathlon is generally defined as the world's greatest athlete, so I think we can say that Sean held the title of New Zealand's greatest athlete at least twice!

Please join me in welcoming Sean Gourley.

[applause]

Sean Gourley: Great to be out here and I think I've got a full 75 minutes, which is probably the longest talk that I've had to do for a while! I think, from what I've seen, with your questions, we're going to have some fun at the end of it as we kind of go backwards and forwards.

But before we get to all that, I want to take you back to my childhood; and probably, like a lot of yours, mine was spent in front of a computer screen. I spent so much time playing computer games, Galaga in particular, and I think some of you of a certain age will remember Galaga.

I actually got pretty good at playing this game—so good that I could get in all sorts of high scores. But no matter how good I got, I'd always get killed by the aliens and dodging the bullets and missiles. My reactions were not quick enough. I simply couldn't move fast enough. The computers would win.

I'd later come to learn that the human reaction time is about 120 milliseconds and that's about as far as we go, and so computers, of course, can think faster than that. That has ramifications in the world of high frequency trading.

But we're better than that, right? We can think strategically. We can think about bigger, more conceptual frameworks. We can think about problems that computers can't think about.

In 1997, a lot of that changed when many of us watched Kasparov take on, and get defeated 2-1 by, IBM's Deep Blue supercomputer. Now, for many of us, this was a story of artificial intelligence triumphing over the human ability to think. That through sheer computational horsepower, a computer could think up to eight steps ahead, whereas a grandmaster was really only about four.

This story resonated and became sort of the dominant story that many of us think about today regarding artificial intelligence—that it simply trumps the human brain.

Sean Gourley Augmented Intelligence (Continued)

Of course, for Kasparov, this wasn't so simple. Kasparov, being the smart guy that he is, realized that computers were going to beat humans at chess, but that didn't have to be the be all and end all. What if humans teamed up with computers? What if humans and computers started to come together?

He formulated a new type of chess, which he called freestyle chess. Freestyle chess was an interesting concept. You could team up with other grandmasters or you could use a computer by itself if you felt that was the best solution. You could use a human and a computer together. You could have any combination of that. You could put yourself on psychedelic drugs if you thought you'd get more clarity. Really, there were no basic rules. You played chess to the best of your ability.

I really like this kind of philosophy. How good can we get at playing chess?

The world championships online in 2006, 48 teams entered from around the world. They were broadly split into two categories. One of the groups, represented here by Hal, you can think of as the headless computers. Pure computer—no humans. On the other side, you've got what they call centaurs—humans and computers working together to kind of create, hopefully, better chess.

As the games went on and the competition started to evolve, one thing became very clear. The centaurs were dominating. They were dominating so much that the final four teams were all centaurs.

Three of them we knew: they were Russian grandmasters and big military-grade type supercomputers. But the fourth team kept its identity hidden. It was called ZackS. ZackS played some really interesting chess. It was at times kind of bold, at times kind of unorthodox; but it actually started to win, and so much so that it won the whole competition.

Now, at this point, everyone sort of expected Kasparov to walk on stage, collect the prize and claim victory. That would be a nice story—except the truth is much more interesting.

The truth is there were two human players, most definitely not grandmasters. One was a database administrator and the other a soccer coach. They were playing on three consumer grade computers, each running a different version of an AI system, with one being a computer they borrowed from their parents.

Think about that. The best chess playing team in the world was running consumer grade laptops and weren't grandmasters. They beat the best chess players and the best computers.

The reason they did this was not because any one of those pieces were the best; it was because they knew how to drive the machines. They knew when to listen to AI-1 and not AI-3. They knew when to listen to themselves. They knew when to go back and say, "We're going to make this a little more aggressive." They played the player they were playing.

This I think is really, really key. You have to learn how the algorithms work. You have to know when to obey one algorithm and not another. When to listen to yourself, and when to defer to the machine that's in front of you. They did this the best and they ended up on top.

I think the story for some of the most difficult problems that we are facing is not one of artificial intelligence, but instead it's one of augmented intelligence: humans and machines interfacing to solve the most difficult problems we face. Chess is one example of that, of course, but there are others.

Think of a big complex world with high dimensionality on one side, and we can use mathematical techniques to reduce the complexity of that world. On the other side, we have the human brain that we were born with. We can use visualization techniques to enhance that. Somewhere in the middle, we find a space of augmented intelligence, where we hopefully can get enough grasp on the world to actually navigate it.

Sean Gourley Augmented Intelligence (Continued)

One of the things that it's actually used for is weather prediction. Weather prediction is a fascinating problem. I don't know if there are any meteorologists in the room, but that's one of the best prediction games out there.

Meteorologists, unlike most people that deal with data, actually build models of the world. They build models of the interaction of air particles, an entire model of the United States and Canada to predict the weather.

It's not simply linear regression that's being done. It's a chaotic system, it has feedback effects. It's highly non-linear. So much so that back in the 1950s, a three day weather forecast was kind of considered untenable.

Also, a couple nice things about meteorologists are: one, they're stubborn – that didn't stop them; and two, they actually keep records of their predictions. When they say it's going to rain with 83 percent probability, they measure if it rained or not, which I think is a beautiful thing that we should do more of. The nice thing about that is we can track the success of those predictions over time.

You can see that the percentage accuracy, for both 36- and 72-hour forecasts, rose steadily. In the 1950s, the 36-hour accuracy is about 25 percent. By 2005, it's about 78 percent.

Interestingly, you see the same kind of thing with the improvement on the 72-hour forecast. Again, these are very, very difficult non-linear problems to solve.

Now, of course, it's being powered by massive changes in computational power. You're dealing with something like an IBM 701 back in the day. Today, they've now got IBM supercomputers running at many times greater than that—about a 15 billion-fold increase in computational power.

We've got a 15 billion-fold increase in computational power to drive a roughly linear increase in a non-linear problem. That's kind of a nice thing to think about. But where are humans in all of this? If we've got 15 billion times the computational power being thrown at this problem, it feels like humans probably don't have a place. The nice thing about the weather forecasters is they keep track of that as well.

In the blue, we've got the machines only, and in the red we've got humans and machines. We're getting about a 16 percent jump. That stays relatively constant through the time period that was measured. There were big changes in computational power there.

Much like ZackS on the chess example, they knew when to listen to the machine and when not to. They knew when to say, "We're going to use the extreme algorithm and not the drought algorithm," "We're going to use the rainy algorithm," or, "I know that this algorithm always gets it wrong with the westerly winds, so I'm going to correct for that." That's what the meteorologist is sitting and doing. They get about a 16 percent improvement.

However, much like the archetypal factory being run by a man and a dog, the man is there to feed the dog. The dog is there to keep the man away from the machines.

[laughter]

If you are not a meteorologist, you probably shouldn't mess with this model. You're probably not going to improve it. If you're not a chess grandmaster—or at least if you don't know the algorithms—you shouldn't mess with it. For most of you, what the machine tells you to do will probably be right.

There's a small percentage of us around that can actually interact and add value on top. That's fascinating because they are the hardest problems, and 16 percent on some of the hardest problems is a massive

Sean Gourley Augmented Intelligence (Continued)

jump. Keep that in mind that we see that across a range of different systems. Human plus machine, around 10 to 15 percent improvement.

To do that, we need to understand the interface, because whilst machines are very good at reading eigenvectors and linear algebra and so on, most of us don't work very well in that kind of interface. We're much more visual, so we need to understand that interface.

To comprehend that interface, we need to understand this computer that we all have up in our heads. We want to think about intuition.

I think intuition is something that we deem an expert to have. They know what to do when it seems like there's no way they could have known it.

Now, an expert has intuition. It's mystical except we know now that it's actually not. We know that from Wan et al. in *Science* 2011, who put together a really fascinating paper on this looking at the neural circuitry for intuition and what goes on in the brain when people have it [Wan et al., "The Neural Basis of Intuitive Best Next-Move Generation in Board Game Experts," *Science*, January 21, 2011, 341-346.].

Going back to chess, it is like the toy model for every kind of system like this. They went back to chess, and they projected a chessboard up onto an fMRI machine. They put the people into their fMRI machine and measured the parts of the brain that lit up. They did this with two groups: expert chess players and amateur chess players.

They showed the image for just one second because they didn't want the conscious brain to kick in. What happened was the precuneus part of the brain, which is up and to the back, lit up. It lit up twice as strong in the experts as it did the amateurs.

For those that aren't neuroscientists, the precuneus is the part of the brain associated with pattern recognition. It's the thing that sees a signal where everyone else sees noise. In one second with the chess board shown the experts saw a signal and the amateurs didn't see anything. The next thing they did was flash up something for just two seconds . . .

[laughter]

. . . and said, "Choose your best move." This was really fascinating because now a different part of the brain lit up. This was the caudate nucleus. The caudate nucleus is a much more primitive part of the brain down and to the bottom just about at the basal ganglia. It's the part of the brain associated with learned response functions.

It lit up for the experts. It didn't really light up for the amateurs, and it lit up in a really, really interesting way.

Here you can see the strength of activity on the vertical and the correct response percentage on the horizontal. What this is actually saying is that the stronger the signal from the caudate nucleus when that image was flashed up for two seconds, the more likely they were to be correct.

It's as if the experts at a subconscious level had an internal hot-and-cold switch, that they would get a feeling that a thing was right or wrong.

Breaking that down, the neural circuitry for intuition images come through the visual cortex. They're abstracted to a simpler dimension with the precuneus to identify a signal.

Sean Gourley Augmented Intelligence (Continued)

That signal is compared back against all the other signals that we have with the caudate nucleus kicking in with the hot-and-cold response function to pick the best one. It's really expensive to build this. If you're not a chess grandmaster today, you're probably out of luck.

It takes a huge amount of time to train these neural circuits to identify in less than a second what the pattern is, what the best move is, what they should do about it. Probably one of the most computationally expensive tasks in all of computer programming is to build neural circuitry for intuition.

It's valuable because as an expert they're able to offload the computational part of their thinking to the subconscious, and that frees them up with all the time and the consciousness to think about the implications of that, whilst the amateurs are still trying to figure out if there's some sort of signal here or not.

What if we could build software to do this first part? What if we could build software to identify patterns, to make connections, to abstract structure so that you could put any data set in front of it and get that same feeling of intuition that an expert has? This is exactly what we set out to do about four years ago.

Four years ago, we set out to build a company to do just that, and it's called Quid. We're based in San Francisco. There are 65 of us working there. I'm going to take you back to San Francisco in a little bit, but before we get there I just want to talk about how we got there.

I was lucky enough to get a chance to go up to Oxford and study to do my Ph.D. work. I originally went over there really to study in a physics lab to do biomolecular motors. I spent the first couple of weeks in that lab, and soon realized Oxford was a much more interesting place outside of it.

[laughter]

I thought to myself, "There's got to be a better way to do this whole physics Ph.D. I've got something to do outside of the physics lab."

I was sitting down at one of these high-table dinners they do at Oxford early on, and I remember sitting next to the former head of the CIA. We got into a debate after a few glasses of wine about the upcoming invasion or however you want to frame that of Iraq. Suffice to say we disagreed about the likely efficacy of that decision, but I could see where he was coming from. I could absolutely see his points.

In many ways I disagreed with him, but I didn't have anything to back that up. It was incredibly fascinating for me. We're making a massive decision here to go into a country and not really know exactly how this thing's going to unfold, and there's no real framework to do so. For me being a physicist on this, I think it set off a lot of the pattern recognitions for me.

I said, "Well, jeez. What if we could get data? What if we could start to analyze this? What if we could come up with a theoretical framework for understanding how insurgencies start to unfold?"

That's exactly what I set out to do, and that became my Ph.D. work, studying the insurgencies in places like Iraq and to study the chaos and uncertainty that occurs all around the world. These are very difficult systems to wrap our heads around and things that shouldn't really lend itself to this kind of analysis.

Of course, the first thing to do that, as a physicist, is to acquire data. You need data streams as an empirical perspective, but when you have an accent like mine, they say, "Well, I'm sorry. You're not going to get our data."

The U.S., rightly or wrongly on that, had controls on the information that was coming in. That almost stopped me dead in my tracks at the start. I consoled myself by watching Fox News.

Sean Gourley Augmented Intelligence (Continued)

[laughter]

When I was watching TV late one night at Oxford, and I remember listening to and seeing the talking heads. But I remember seeing something else, this little ticker down at the bottom.

This little ticker of news that came across, scrolling little events that happened with bombs and attacks that were going off. I remember thinking, "What if we were able to extract that? What if we were able to start training computers to read that? What if we could have a system that monitored that, and pulled out that data."? That would get us past the whole classified thing, right?

That's exactly what we did. We built out a system that would monitor news reports and the blogs that were starting to emerge from hundreds of different sources, and train them to find attacks.

We trained them to find people that had been killed, when it happened, where it happened, how many people were killed and so on.

We got quite successful at it, and when we did that we had a database of all of the attacks that were there. We could then look at that to see if there was actually a signal contained inside it.

We didn't know how good our data was until [Julian] Assange released the WikiLeaks information.

It was quite interesting. The open data had 81 percent coverage of what the U.S. military had, and that's a pretty good result. But they only had 70 percent of what we had. Now, let that sink in for a second.

Three people on laptops in Oxford training computers to read ambient news sources pulled in more significant event information than the entire U.S. military on the ground, with an information landscape of 2006.

The information landscape has massively changed. However much information you have inside your organization, there is much more outside of it.

Of course, this has implications for security as well, but that's the kind of first step – the massive amount of information that can be claimed from ambient sources. The second is that there are statistical signatures that exist inside this.

If you look at the frequency of attacks versus the attack size for Iraq on a log-log plot across the entire conflict, you see that the first order of a straight line approximates a power law distribution. The exponent here is the slope of the line of 2.3. Out of all the chaos unfolding across all of the different groups that we're attacking and evolving, we see a strong statistical signature like this.

The interesting thing is this is repeated across many different conflicts around the world. You see the same signature in Afghanistan, Columbia, and Sierra Leone. You even see it in Northern Ireland.

It's as though when people get together to kill each other they do so in mathematically precise ways, which doesn't really make sense until you think about this. It's incredibly hard as an insurgent force to take on a strong opposition.

It's so hard, in fact, there are very few ways to do this. And if you evolve and you don't find the solution, you don't exist. There's not enough data to be measured. For any system where there's enough data to be measured, it means that the insurgents have found an organizational structure that is typified by these statistical signatures.

The Pentagon was quite interesting, and of course, the first thing they start to say is, "Can you predict?"

Sean Gourley Augmented Intelligence (Continued)

You don't get these kinds of statistical signatures without predictive ability, that's the nature of it, but it's also a very strange question. Why would I want to predict that three people are going to be killed in the northern suburbs of Baghdad without wanting to sit down and actually change it?

As we sat around the Pentagon table and thought about this surge going into Iraq and whether to send 30,000, 60,000, or no troops, it wasn't really a question of "Can we predict the attacks?" It was, "How are we going to change the system so that this signature doesn't exist at all?" To do that, you need more than a simple statistical model. You need to start to think about how the insurgents are organizing. You need to create models of how the insurgents are allocating resources and create models of how they're making decisions.

This is what we did. We decided to create models inside of our computers to simulate the insurgents, and the output of the models had to come back and agree with the statistical signatures that we're observing.

You could start to change these. You could say, "Well, what happens if we change the number of troops on the ground? What happens if attack the large groups and not the small groups?"

What happens if we tried to break communication? What happens if we did nothing at all?" We published this in 2009 on the cover of *Nature* [Juan Camilo Bohorquez, Sean Gourley, Alexander R. Dixon, Michael Spaga and Neil F. Johnson, "Common Ecology Quantifies Human Insurgency," *Nature*, December 17, 2009.].

For me, it was a great thing to get this research out. To finally, after seven years at that point since that conversation I'd had at that dinner table with the CIA, sit down and say, "Well, actually here's a framework. Here's something that we can start to use to make some very significant decisions in our world."

As I was in Iraq, kind of wandering the hills thinking about what this all meant, I was doing some math in my head.

We had about 300,000 dollars in European Union funding, six people working on the problem, and I was thinking, "We want to build a system that ambiently tracks all the information in the world to start to find these statistical signatures, these patterns, that allows us to make some very important decisions, and it's probably going to take a little more money to do that."

I figured \$100 million and 1,000 people should be a good start, and of course at that point I kept walking and decided that that might not happen, until I realized there was a magical place called Silicon Valley where people with crazy ideas do these kinds of things. They're actually given money to do them. It's just all very magical.

I actually met up with Peter Thiel, who is [a] founder of PayPal, but also an early investor in Facebook, and worth more than God at the moment.

He gave me the first \$2.5 million to get this started, so we actually began at the end of 2009, and started to build a team that would build the software.

With 65 of us out in Silicon Valley, downtown San Francisco, we train our data scientists very young in our office and we're about two-thirds technical. There are about 10 percent of us that are pure data scientists, and the rest are kind of sales and operations.

We're shipping software today to many different groups around the world, a software that you can plug into inside of your web browser to tap into these different data streams to make use of the artificial intelligence to find the information.

Sean Gourley Augmented Intelligence (Continued)

There's a range of clients: everyone from the big tech companies of Microsoft and IBM, all the way though to some of the more interesting consumer companies like Red Bull and Lego, and sort of everyone in between. A huge range of customers from hedge funds to advertising agencies through to government decision-making.

The reason for that is because once you try to compute it to read and make connections between objects like a human does, it actually turns out there are quite a lot of applications for that across a whole range of industries.

They ask questions kind of like this: "What are the dominant narratives about climate change in India, and how does it vary within age groups," or "What are my competitors doing with advanced flexible display technology, and should I partner or compete with them?"

These are the kinds of questions that experts would generally ask, but would take a team of four analysts six weeks of reading to get this information back to you. If you happen to be so lucky to work with McKinsey, that will cost you \$1.2 million to do that.

Again, humans are expensive to read this information. One that we did here with NASA looked at the structure of the commercial space seeker system.

The commercial space industry is no longer a government monopoly. They're likely to be launching Space X and taking humans capabilities now up to the space station, but this changes the landscape quite dramatically.

If you want to understand the commercial space industry, one place you might look is Google. We might do a search on the news stream and in the last few months we see 4,200 articles on the space industry, and that's kind of great.

It returns it back in 0.17 seconds, and thanks to Google, you have your reading for the next six weeks. Again, what's the one article that I should read?

That seems like kind of a strange question, "Is there one article I should read about this week?" You start to see the issues with AI saying, "Here's the best thing for you," when the reality is I want to read everything. I want to see all of it.

Can we turn those 4,200 articles into a dynamic interactive user interface to let humans explore what's going on? We can do that. What we do is we take the text from the article.

We do this for every single news article, blog, Twitter post that's published, extract that out, and start to look for key concepts within that.

We have algorithms that will read through these things to extract out the key concepts to create what we thought of as a digital signature—a sort of fingerprint of every article.

Now, with that, we can then start to make comparisons to every other article. If we find one that's similar, we can connect it and project it onto a two-dimensional space, something that we can finally interact with.

That's how it works in theory, but we can have a look now to see how it works in practice. Here we're looking at the Quid software now inside the web browser.

You can see "No" represents a story here about the global space industry, but stories aren't just by themselves, they have connections. They have connections to similar stories, and we can actually start to see that there.

Sean Gourley Augmented Intelligence (Continued)

Now, there's no space in this at the moment and there's no importance, so we can size by degree. We can also play a force directed algorithm to start to bring some spatial resolution into what we're looking at.

As we set some parameters and apply that, we can now see the similar stories cluster together, and the different stories kind of fly apart, and a landscape starts to emerge of the different topics that are now being discussed in the global space industry.

We can start to apply a clustering algorithm. The colors now represent similar communities at a high level, and we can zoom in and take a little look around, much like we do on a Google map, except this now is not a map of geographic places.

This is a map of all the different ideas and concepts that are being discussed around the global space industry. You move across there and in the yellow you see some of the European Space Agency explorations to far out planets.

We can come down through the middle, everything from cybersecurity to down at the bottom here, the GMD kill vehicles and the interceptors from Boeing and Raytheon. A bunch of stories written about what's going on there.

Within the space of a few minutes you can orientate yourself to all those thousands of articles to see the different topics and concepts that are being discussed. The same thing that you'd hope that your set of analysts would return to you after six weeks, but maybe they wouldn't quite get it right. Applying the cluster, we can see it will start to name the different elements as well.

What's interesting here is—you can see up into the left—you've got NASA in the budget. For many of us, the space industry is NASA, but you can see that the conversation is actually a very small part of actually what's going on.

Much larger is the other topics that are going there. For example, the U.S. military, connected next to Earth observation, military defense, commercial satellite imagery and also connecting up into the middle there, cyber war and space debris.

By far the largest cluster, or group of things, is what comes around from a satellite broadband, mobile broadband, launch vehicles.

What you start to realize is the commercial space industry is not really about space as we generally imagine it with rockets and astronauts. It's about communications systems. It's about sending data backwards and forwards between satellites to power your mobile lifestyle – sort of the first thing that pops up.

Now, if you were to ask an expert to sketch out what was going on on the back of a napkin over drinks, they might put something like this together to draw up the different elements and explain and kind of walk you through that.

This is their cognitive model. Well, we can actually have a pretty similar kind of cognitive model from the machine, the same kinds of concepts aggregated and clustered together, their connections representing similarities, and we can go and explore that. We can interact with it.

It's got a tech top field to it. In essence, we're sort of uploading that information into our brain.

Question: Quick question, does the software select category names or is it an analyst?

Sean: It takes three guesses and then an analyst can edit it. It might be like observation/earth/imaging. Then, the editing will come through and will go off. That's earth observation.

Sean Gourley

Augmented Intelligence (Continued)

Yeah, it's good at guessing, but it's not quite as good as a human would be. Again, there's a sort of editing component.

With that we can again apply the algorithm. This is a very high level clustering algorithm. We say it splits into broadly two groups. You can see the structure of that network does lend itself with different bits of the information across either side.

You can kind of think of this as United States versus the rest of the world. What's interesting there is the different conversation when we're talking about stuff in the United States compared to when we're talking about the rest of the world. That's interesting.

What's really interesting is the stuff that happens right at the boundaries—these two little places. We go and zoom and take a look at what's happening between the United States in the digital world. You see a bunch of articles here about the overhaul of the export licensing system.

Of course, there's the ITAR [International Traffic in Arms Regulations] regulations that forbid any space technologies to be exported to the rest of the world for fear that the Soviets might start to use them. The algorithms have correctly identified that that's something that's happening at the interface of these two clusters. Quite a powerful nuanced insight.

We go up to the other side and we see here what's happening in this interface, which is more around NASA, as we see [Charles] Bolden, the chief administrator of NASA, preparing for a trip to China. Again, as they start to make overtures to connect with other countries that have space capabilities.

These, as you start to navigate it, can be quite nuanced insights that you can observe, not just the major clusters, but how they're interacting with each other.

This is the High Sierra. If anyone has been there you know how beautiful it is. I spent a few weeks there one summer running around with an ecologist, looking at these high alpine meadows to see the food systems that were contained within them.

We're introducing the brown trout and looking at the food web that would emerge from putting a new species into that system.

The species that thrived were in color. The ones that were decimated were in gray. You can think of the same thing of putting SpaceX into an ecosystem and how the ecosystem of every other player would start to evolve. We can start to see the connections between that when we read the text.

We can read a text like this and say that there was an event. It was a \$160 million contract, Orbital Sciences and SpaceX. We know when it happened. We know the company Thaicom was involved. We know that there's an event that occurred between objects. We can represent that visually, but much better to do that in three dimensions.

On the red here, we've got SpaceX at the top and you've got its first order connections through different financial events with other entities. In the green are the second order ones. This is the food web of the ecosystem of the global commercial space industry.

Then we can go and navigate it. But we can do this for any system. We can do this for cars. We can do it for ideas. We can do it for patterns and scientific papers. All of a sudden, this space, these kinds of connections that experts hold becomes accessible to any of us.

Sean Gourley Augmented Intelligence (Continued)

At this point, the expert can't actually hold all these objects in the head. I think this tool is quite revolutionary insofar as what it enables us to do to explore, upload, find patterns and make decisions in a very, very complex world.

One of the interesting uses of this is narrative analysis. This is really interesting for a couple of places. One is the industry of advertising and the publicists groups of the world. It's also interesting for narrative analysis for things like headline risk in publicly traded companies.

What are the stories that we're telling about the companies and are they the right stories? How do these stories propagate through? Also interesting of course from a political perspective.

We can take something here like Occupy Wall Street. Of course, thousands of stories written about an amorphous agglomeration of people. What does it look like?

We can take a projection of that again and we can see here the projection of all the stories. We can see here the clusters. We can see here, in yellow, all the stories that mentioned inequality as it starts to radiate out from the political discussion.

We can also see the separation between Occupy New York and Occupy Oakland, as different conversations were had representing a schism between those two parts. We see here no clear distinction between the Republicans and the Democrats as they struggle to come up with a political response to what was actually going on.

This is looking at a post hoc, but of course, you can do all the stuff in real time as the stories come in, the landscape evolves, the topology changes. We can see the ideas radiating as they're published. This, of course, is very powerful; whether you're an ad agency, a political group, or you're trying to monitor things within the financial market. You can see the stories, and even do more than that—seed the stories yourself to change that...

Question: Sean, did you go back in history and then watch it?

Sean: Yes, you can go through a bunch of the whole *New York Times* archive. You can rewind and see the stories evolve. You can do that with a range of different sources. Everything is time stamped, so really fascinating.

A lot of the hedge funds that are using that are looking at what the launch looked like in 2007 when Apple put out its iPhone and how does that compare to the launch today from the Microsoft Surface?

Our world is complex. We do need to make decisions within it. We don't have all the capabilities within our human brain, but nor is there a magical button that we can push from a machine. We have to reach towards this augmented intelligence, I believe, to allow us to navigate that and to take the best of the machines and the best of the humans and put them together.

I want to contextualize this. We have, of course. I believe 2013 was the biggest year for data that we've ever had. It got to crazy points of coffee table books being produced about data scientists and all these kind of weird things.

It was so much so that I think my favorite paper that came out was an article on the physics archive called... 'I wrote a paper about Twitter predicting x and all I got was this lousy paper'. It seemed that Twitter could predict everything and yet it was only producing papers.

Sean Gourley

Augmented Intelligence (Continued)

It got so big, of course, that it got to the point where David Brooks was compelled to write a column titled "What Data Can't Do," because it seems it can do everything. He had some pretty good points. He outlined four things that data struggles with.

The first was that it struggles to capture the nuance of human emotion. It can see that a person sent an email from A to B. But it really struggles to capture the sense of love between a father and his son.

The second one was that it holds itself up as a badge of authenticity, as a badge of objectivity. It's numbers, whilst oftentimes brushing away where the data came from, any of the models that underlay it, or any of the kind of the biases that were contained. We looked at the number.

The third thing was, very much so, as we get more and more data, we tend to lean on the crutch of correlation at the expense of looking for causation. The more variables we have, the more opportunities we have for correlation to occur. If we don't start looking for causation, then we maybe lead ourselves down the path that's wrong.

But perhaps his biggest critique was that big data can't solve big problems. That we're sitting here with an amazing technology and we're using it to solve trivialities.

The hypothetical data scientist sitting in their lab in Silicon Valley optimizing children's breakfast cereal and advertising packaging on top of it to take the color, the size, the shape, the form, the position on the store shelf, to get the eye tracking software to see where people were looking.

Also, that we can get three percent better yield on the product without sitting back and saying, "Well, jeez, what about obesity? What about diabetes?" Things that are a lot harder to actually understand with a simple kind of structure of information streams that we seek to optimize.

I think the reason for that, and we need to rewind the clock – why are we here with this amazing technology optimizing archetypal children's breakfast cereals at three percent? I think we have to appreciate that over the last six years as a society we've undertaken one of the biggest experiments in quantification that perhaps we'll ever do.

We've all taken our relationships, our personal information, and we've uploaded it into a structural database from the likes of Facebook, so that they can mine that information and serve us ads. This is a massive experiment.

We've all run it. We've run it on the scale of billions. I don't know what exactly it's going to do to us, but we're running it.

Out of this experiment really came data science as we know it. It came out of that with the two people that coined the phrase, Jeff Hammerbacher and D.J. Patil, and in the very instrumentally early stages of data science in defining what we have today with it.

Jeff was the head of the data analytics team at Facebook. D.J. was the head of the data analytics team at LinkedIn. Data science that's been captured by these people reflects their experience. That's fair enough. It's going to reflect your experience.

Your experience is millions of data points, structured information streams, A/B testing, so you can know whether something's better or worse with the ultimate goal of serving an ad.

To their credit, they're really good at doing this stuff. So good that just through the public likes we can predict if you're single or in a relationship, whether you use drugs, your race, your religion, even your gender.

Sean Gourley

Augmented Intelligence (Continued)

We can predict this information with startling accuracy even if you don't choose to reveal it. Even if your friends don't reveal it either. What we found out of all of this was that liking curly fries is a top predictor of intelligence.

[laughter]

That's an actual result. That was interesting. But we laugh at that because it says nothing about intelligence. We use this data to find correlations without understanding what's actually going on underneath it.

The point of that is if you're a data scientist at LinkedIn, you don't care. All you care is that if someone says I want to advertise to smart people and if you've got curly fries on your home page, you're probably going to get an ad for whatever they're selling. We haven't strived to understand what's happening.

I think data science is quite limiting in the way it's being cast and in the way it starts to talk about what data can do. I think data intelligence is a much more interesting way to think about this. Data science versus data intelligence can sort of be summed up like this.

In data science you're looking for 10 percent improvements, whereas in data intelligence you're looking to change that game. The goal of data science is to predict and optimize. The goal of data intelligence is to create and change.

The decisions are primarily made by algorithms in data science, whether it's the ads being served on Facebook or high-frequency trades. Algorithms are making those decisions, whereas in data intelligence it's very much the humans.

The data is often big and clean, whereas in data intelligence it is small, messy, and it wasn't necessarily designed for the problem in front here. The communication in data science is done through equations, whilst in data intelligence it's stories. You could think, broadly speaking, of data science as being tactical, whilst data intelligence is being strategic in nature.

Some of us have sort of gone through that over the last decade of experience working with data. I want to share with you five heuristics that I use to think about data and data intelligence and particularly how to use data to solve big problems.

The first of these is data needs to be designed for human interaction. A human will read this very well on the left and interact with and explore it. A machine will read the one on the right. We need to remember what we're actually dealing with. We need to design human-centered user interfaces for our data.

The second is we need to understand the limits of human processing. We need to know that we can't make a strategic decision faster than 650 milliseconds, that we hold about five to seven objects inside of our head and that we're comfortable with 150 relationships. That's the kind of biological evolution that brings us to that point.

We should also know that beyond the human capability, there will be open game for the algorithms. We get the high-frequency trading environment, where 90 percent of trades in US equity markets are made by algorithms.

They'll exploit that because they think faster. So appreciate the human biases, appreciate the human limitations, but also appreciate that when we can't go past it, the machines will dominate.

Sean Gourley Augmented Intelligence (Continued)

Data is messy. It's incomplete and it's biased. We need to know what that data is. One of the things I always say is if you've got data and you haven't opened it up inside of a text editor, if you opened it up inside of Excel and read at least one roll of it, you shouldn't be using any models that are derived from it.

There's a kind of data that we get out of places like Afghanistan—handwritten notes from al Qaeda operatives that don't lend themselves to a lot, but they're very important. You need to use them, but you're not going to necessarily be able to look that up inside of a database and run a regression analysis on top of it. We need to deal with that.

Data needs a theory, because if it doesn't, you have things like this: The number of people who drown by falling into a swimming pool is highly correlated with the number of films that Nicolas Cage appeared in.

[laughter]

I knew Nicolas Cage was a bad man. That's correlation at its worst. As we get more variables, more variables recorded, you can find this for pretty much anything. This is at the heart.

Chris Anderson of *Wired* magazine said, "Theory is dead." This is the world you've created, Chris. We need theories because we need to understand what's going on. We need theories because we also need to make predictions. Predictions are great if the future's going to look like the past.

But you need a theory if you want to see what it might look like if it doesn't look like the past, if you're actually going to change it. That's where physics comes in. We start to build theories. Very much to the heart of physics are theories to explain empirical evidence.

Finally, data needs stories, but stories also need data. We tell stories and we tell them very well because they're memorable. They communicate and they percolate through organizations. They're very powerful tools to encode information about complex environments. We're telling them, of course, for as long as we can remember.

Whilst I could show you an equation for the dynamics of insurgency, I could equally well tell you a story about the myth of the Hydra, the multi-headed beast that comes in and cuts off one head only for two more heads to appear. This non-linear function is encoded in that story in the same way as it's encoded within that equation. But this story will be probably told a lot longer than those equations will be.

Big data can solve big problems. Indeed, it must solve big problems, but it's going to be doing it with a framework more akin to data intelligence. To do that, we need to understand that interface between humans and machines.

In short, I think we need to build better centaurs. The centaurs were the mythical creatures that lived in the forest; half man, half horse, the head piles of wisdom and insight that you would go and consult for the big questions that were around.

[laughter]

You could actually type that search query, optimize it against Nicolas Cage films and come up with your geopolitical strategy optimized for social media likes. Plug into a drone that would make decisions.

I was chatting with Michael [Mauboussin] up at [the Santa Fe Institute] about this. He said, "That's got some implications," I said, "Yeah, it does." We talked about what some of those implications were. I want to run through a little bit of that and particularly algorithm design.

Sean Gourley Augmented Intelligence (Continued)

If we design algorithms for online interaction, we could take two really simple vectors. Your history – do I remember all of your history or do I remember none of it? And your identity – can you be anything you want to be? Can you be an avatar? Or do you have to be you?

You can use these two vectors quite simply. I can design a set of social interactions online with two vectors. What do I get? Up onto the right is Facebook—identity and history. Turns out that's a very profitable space to sell ads to.

Up into the left, we've got Reddit, and for those that know Reddit, you can be anyone you want on Reddit, but it does remember who you are and it assigns your karma based on that.

Down to the left, you've got *Fortune*. If you haven't been to *Fortune*, don't go to *Fortune* because you can imagine what happens when you don't let people be an identity and you don't remember anything they do. You get a kind of a weird 12-year-old playground, which is at times genius and most of the means that we ever see come out of *Fortune*, but it denigrates pretty quickly.

On the bottom side, we've got Snapchat, a real identity and no history. We all know what happens when you do that. Or maybe your kids do.

[laughter]

You don't want to know. These four systems occupy these four different quadrants. By just manipulating history and identity, we get four very different types of behaviors. From the kind of behaviors we have where my identity is always going to be remembered for me, versus it's never going to be remembered and we behave very differently.

How do we design these systems and what kind of systems do we want? Because they are being rolled out to billions of people and they do affect how we think and how we interact and socialize.

I think that's really interesting. We think about the stance of how do we design. But I will also say what happens when you have these algorithms that actually allow us to think better? What happens when you have augmented intelligence? What sort of equality or inequality that starts to create?

We can roll that out and think about the different ways the algorithms and humans can interact. What I showed you there with the Facebook example is the dominant way most of us are interacting in the world through Facebook and Google to get information through these filtering algorithms.

These are free systems. We don't pay for Google and they're sort of free. The tradeoff is that we're being treated as a product. We're being bought and sold literally in real time by algorithms, much like the high frequency algorithms that we have on Wall Street, trading backwards and forwards to bid against each other to show you an ad.

The information that you've got there is being mediated by that. That does a couple of things. One is it puts a cap on how much it can cost to run that algorithm. The entire system per user per year for Facebook costs \$1.50. Google costs about \$12. They have to sit under that cap because that's all the money you can get from advertising.

It's very hard to get into the hundreds of dollars from advertising. The complexity of the algorithms is capped, but the goal of the algorithm is not necessarily to give you the best information, to make you think the right way so that you have the best personal relationships, friends, and updates, the right information about space news. It's to sell you ads and we should remember that.

Sean Gourley Augmented Intelligence (Continued)

What we're going to see is the ownership of these algorithms start to switch. Instead of it being owned by Google, and being optimized for Google's advertising needs, we're going to go "You know what, \$12, I'm probably going to spend that. In fact, I might even pay a couple of hundred dollars, but I want to own the algorithm."

I want it optimized to me. I want it to push me to places to think about things I haven't thought of, to make sure to remind me to connect with people I haven't connected with, to kind of constrain me if I'm feeding myself too much junk food type of information. I want it to optimize for me, and I'm prepared to pay for that because information and how we think is incredibly valuable.

Of course, there are going to be people who don't have enough money to be interesting at all for advertising-based algorithms, and they're actually going to be working for the algorithms, the sort of techno serfs of the word. This seems kind of science fiction, but you look at Mechanical Turk, you look at a lot of these micropayments sites, you can't earn enough money. So an algorithm says, "You know what? I'm not smart enough to determine whether or not this picture is work appropriate. Human, is this a safe picture? Is this a picture of someone's genitalia?"

They do that. That's the sort of thing you pay one or two cents to do -jobs that the algorithms can't do. Of course, on the far side of that, you've got augmented intelligence where humans are not just owning the algorithms, but they're actively driving the algorithms. They're interfacing with it. They're not just taking the prediction, but they're actually starting to manipulate and understand.

I think we're going to move from what we've seen; where everyone gets information from these advertising-driven models, to these different things. You can see here that it brings with it a huge amount of inequality as well.

One of the things is, as we build these systems and roll them out, we're going to need to think about what do with a society where some people can own algorithms that make them literally see further and think better, whilst other people are being fed junk food algorithms, serving them information on the likes of Google.

On that, that's probably controversial enough to open up to questions. Thank you, and I'm happy to field any and all.

[applause]

Question: In the interface that you've shown us, you talked a little bit about the ability for analysts, to make the analysts then provide a little bit of supervision. Is there any capacity for all that supervision and all of that to then go and train the total system, so that it could better...you're uploading all that information?

Sean: That's exactly right. As we get more users, they're able to add and delete connections. They're able to change the guesses that we make. The system is learning as it goes through. We have a very nice setup now.

Question: But all users of your software then benefit from those changes?

Sean: That's right.

Question: It's sort of like Wikipedia.

Sean: That's right. They won't see your searches, but the algorithm will say, "I weighted too much the part of speech tagging, I should bring that weight down a little bit. When I'm guessing names, I should be

Sean Gourley Augmented Intelligence (Continued)

a little less happy to use adjectives," and so on and so on. It will change the weights of the guesses that are made for both the naming and the links.

Question: Are humans who are good at collaborating with computers born or made?

Sean: That's a good question. To answer that, I think you need to think about what makes someone good at collaborating with a computer. I think of the two people that won the computer chess were the database administrator and a soccer coach. You can kind of think of two sides of that. You do need to understand the data, you need to have a sense of where the data has come from, where it is, what's going on.

You need to understand how the algorithm is being performed, how it's being programmed. To understand how it's been programmed, to know that it's going to do simulate to climb a local maximum, but it generally misses the global maximums. You need to have that language to know that. But I think you also need to know, in a sense, a strategy and gain internal sense yourself. So you have to have a good awareness yourself.

That's quite rare to have both a knowledge of data, and a knowledge of yourself and the world around you.

I think I'd sort of draw a Venn diagram of people that understand data and algorithms, and people that understand the world. The Venn diagram is maybe the people that are good at artificial intelligence, but that's pretty small. I think we can train it and we should train people to understand algorithms.

I think we can take a lot of the kids that we're pushing into math and science to get out and make sure they understand politics, philosophy, and economics. We probably can help ourselves along that way.

I think we should also appreciate that not everyone is going to be able to add to an algorithm. That's got pretty profound economic implications. If we've got a set of people that can add to algorithms, and a set of people whose jobs are being replaced by algorithms, where does that leave us?

Michael Mauboussin: Phil? Do you have a thought on that?

Phil Tetlock: A really good question. How hard would it be to go back to the computational linguistics of the space industry?

I have done a little bit of work with the intelligence community on Open Source Indicators, and I am curious how close we are to certain benchmarks to others.

One question would be whether your system could answer the question, I know high level NASA executives visited China, without actually mentioning, not using the exactly same terms as those in the newly-released films.

Sean: Could our system answer the question, has a high level NASA executive visited China? There are a couple of points in there, one is we know [Charles] Bolden, we know that he is an executive, we know he's the chairman of NASA, we know that China is an entity, we know the geographic thing to have traveled. We haven't designed the system to answer those questions like that.

For us, we have designed it very much for humans to interface to it, so that you'll search for everything that's going around NASA, and you'll see a cluster there very quickly, like you did. There was cluster of NASA going to China. Now, we can make that quicker and have the computer know that but also, it might lead us down the path of artificial intelligence.

Sean Gourley Augmented Intelligence (Continued)

Question: What is interesting is that it probably ties into the Ferrucci talk – Watson could do that. The other question, and I am not sure whether this is augmented intelligence or this will be used tomorrow, would be: how far are we from being able to pose questions about possible futures, to the occupational district tomorrow?

Is China going to have a man on the moon by 2020, or is it going to test another anti-satellite weapon?

Sean: How far are we away from predictions on top of these models? Is China going to have a man on the moon in 2020, are we going to see anti-satellite weapons?

You could ask a question that to the computer and the computer could spit back an answer. I would contend that you wouldn't be like, "Oh cool, it said that." We'll probably build around geopolitical policy on top of that prediction. You got to want to look a little closer. Even if it could today, would we trust it, and would we really kind of spend hours for just a computer output?

That's if it was there, the second I would say is, that's an intelligence question. We have a bunch of intelligence analysts doing that. They do it at the moment by reading through scores of information, to plug weak signals to create a model about what they think will happen there.

The point of thinking what this thing does is, it takes all of that reading and compresses it right down, and takes the connections that they would make, and starts to make them for them.

They can get the week's information, and we know that there's evidence of both those things happening and then they can start to assimilate that. There's definitely still a role for the analyst. The analyst should take that information. I don't think the computer's at a point where they can do that nor do I think they will be there in the next 10 years.

Question: That was my sense too. We could do augmented intelligence experiments within a large organization that tells you. Some analysts get access to your model, some don't, how much value is there in the ability to perform?

Sean: Absolutely, we would rather start with the three major consultancies, and two of the three. You can guess which two of the three. I did A/B test on that. They actually set up a team of four analysts to do market segmentation for an M&A activity, they said, how long will it take them? How many targets do they find? And what sort of clustering did they do?

It took them about four weeks; a principal and three associates. They found two hundred target companies, and they split it into five different groups. The 10 that are running Quid did the same experiment. They did it in three days and found 1,500 companies, and they split them into 25 groups. These are not cheap people, and you are not only doing it faster, you are doing it with a higher resolution.

That's on one side of things. The second anecdote I'll note is a major search company that runs this for their competitive intelligence process. They ran the A/B tests, since they could do what they did in a week, they could do it in a day.

Question: I was curious about your assessment of the three. Why the exact three? Because part of your description was mainly circumstantially based, knowing when to be aggressive and when not to. Presumably, the grandmasters, that's the way they do it. These guys should be able to be the grandmasters...

Sean: This was not a one-off. It's consistently being done again.

Sean Gourley Augmented Intelligence (Continued)

I think the thing is why people won, of course, there starts to be conjecture on that. The interesting thing was that they wouldn't have beaten any of the other players had they played them by themselves.

We know that. What did they do that the other people didn't do? My guess is that when you become trained as a grandmaster, you are trained in a very different process compared to interacting with a machine. There is no sense that knowing how to play chess as a human is knowing how a set of algorithms play chess.

To know this, you need to really understand how algorithms play chess, which means you need to understand the code behind it. You also need to know how they've been trained, how they've been trained in their endgames, how they've been trained in their opening games. You want to know when they're climbing local maximums and when you're pretty confident it's a global maximum.

I think knowing that, and knowing workings of algorithms, I don't think grandmasters necessarily know all of that. There can be grandmasters that know that. It would be interesting.

One of the interesting things is through computer games, people are playing different chess now. They've all been brought up playing online computer games against each other, against machines. It actually changes the opening styles and we can track the opening styles of chess pre- and post-computer induction.

There is a significant change in how we play. I think that's also interesting. This is pretty early stuff in terms of understanding what it is and to gather the earlier question, what it is that makes a human work well with a machine. No doubt, as that becomes more economically important, we're going to get better at understanding it.

Question: Isn't the key the way grandmasters deal with the chunking of information?

Sean: I think the chunking of information is something that's really interesting, early on the precuneus, the pattern recognition of our brain that abstracts a series of complex things to say- yes, I've seen a signal in this. That's a low dimensional representation of that high dimensional information. Something really innate to us is that chunking of data.

One of the things that humans almost invariably will do. There will almost invariably be five chunks or seven chunks. No one ever comes back and goes you know what, I've seen 123 insurgent groups in Iraq. There's always seven groups.

That's because it's about as many as we can handle. Most people are dealing with three. We chunk information at an abstraction that we can deal with.

Computers do the same as well. They chunk the game. One of the interesting things I think is computers, although they've done a pretty good job of solving chess, they've done a terrible job of solving Go.

Chess has 10 to the power of 40 moves as a sort of state space. Go has 10 to the power of 150 moves. That doesn't seem like a lot, but that's a big jump. They can't even really beat a novice Go player. Go is the game with the stones, if you've played that. It's got a huge state space.

They solve that. Their current best things run on local laptops. They don't run on massive supercomputers because they're basically imploring heuristics. They're playing the game like a human plays it because the state space is so big it would be like going into the ocean with a teacup, even with the biggest computers we have we can't explore that kind of state space.

Sean Gourley Augmented Intelligence (Continued)

It's interesting to see Go develop and how AI will start to solve that system and where we are with that today. I think one of the interesting things is to look for signals of where AI is going. What it can do? What it can't do? Straight after this talk, we'll see some real interesting stuff around natural language processing and question and answers which I think are going to be fascinating.

We should also be cognizant of where it is. We do not have a machine that can accurately determine the right geopolitical move for Iraq. We shouldn't expect to have that given that we haven't solved Go. We'll probably solve Go before we solve that, maybe.

Question: What's this say about the practice of reading?

Sean: Franco Moretti kind of coined the term "distant reading," which I think is a really interesting thing. We tend to think about reading words, but you can actually read at sort of an abstracted level of what's going on. That's important because there are 1.2 million news articles in English published every day.

If it's your full-time job, you may sample 200 of them. Is that the right sample? What are the biases implicit in that sample? How much of that is the financial trade publications versus reading what the consumers are actually thinking? How hard is it for you to unwind that bias about that one article that said that stock was going down even though you're reading at the same time that the consumers love it.

When we're sampling 200 from 1.2 million, we're going to be biased. We're not going to get faster at reading traditionally. If we want to get that encompassing event information, machines will have to read that. They'll be able to answer a lot of things. There's a high level NASA official going to China.

The more complex stuff is still going to need a human to interact with it. That's where that interface stuff become really important.

Question: How might this work for a corporation?

Sean: Imagine if you will there's a large advertising company that is working with a deodorant company. They want to be more manly. You get a conversation of all the stories about manliness. There are different clusters on manliness. There's a cluster of stories around whether Chris Christie is too manly to be the governor. Or some more serious, bullying in the NFL and is that acceptable or not, and facial hair, that's always going to be there.

That's pre-hundred million dollars of advertising spent. However this company moves into that space, you can be sure as hell that network, that space, that projection is going to be different. They can predict how the conversation is going to evolve. But what they choose to do will massively shape what evolves.

I can predict without me making a decision, but it would be kind of strange. A much better thing to say is what part of that message should I take. Should I combine to get two different clusters? Should I take a white space? Should I take one that's owned by a competitor and start to own that myself? You can do the same thing for politics, which side of the fence you're on.

The decisions that we make as we go into that space will change things. That has to be accounted for in any kind of system that you're building.

Question: How would you improve this system?

Sean: What I say to this is get more observations. You read one story about the manliness of facial hair and you don't understand the whole thing, so you get more. What I would say on this here is it is reflecting back everything that's happened in the world. All the stories about X, all of the events about Y.

Sean Gourley

Augmented Intelligence (Continued)

It's not making projections forward. What data is significant? I think that's really important. It's not going forward and saying we think this is going to happen. To the point of how do we make better decisions about the future and what's going to be there, I'd say is make more information accessible to the people that are making the decisions. Given them a better view of the world and higher resolution quicker so they can think more about the implications on top of that.

Humans are making these decisions today. There were a group of people that sat around the Pentagon table that decided to send troops into Iraq. That was biased and they had their own ideologies. They didn't have all the information. They didn't have all that stuff at their fingertips. We *are* doing this.

I think of the question as, "Over the next decade, how can we get a set of tools that can help humans still make those decisions better?"

Question: How do you compare models?

Sean: I think that comes down to the way the forecast is recording your predictions and recording them against what actually happened. If you can make that, then you should be able to do a delta given enough data points of your analyst predictions with this system versus another system.

I think the contention would be if you can give them a better representation of the world as it is today, to make better decisions on top of that about tomorrow.

Question: What are some of the different things being done today?

Sean: One is global macro stuff. I need to understand what's going on in Crimea. There is a lot of stuff going on and how I understand that is going to change the way I price my oil futures or my natural gas.

The way that's done at the moment is reading a lot of data streams and what this does is it starts to make that quicker and more efficient and more accurate. Another one is looking at long-short public equity. You take all the information, particularly around headline risks. If an event happens and it blows up in the social sphere, is that something that is real or is it something that's actually just fabricated?

What are the different statistical patterns that emerge from headline risks around different topics? It's not like just saying, "Is this sentiment positive or is sentiment negative," but actually getting it to say, "What are the heart of the narratives that are being told?"

Once you go down a little bit into private equity, you are starting to look at the selection of all different hardware companies in the world.

There might be 10,000 different hardware companies. Which one should I invest in or which one should I acquire?

I think for me, the most interesting stuff that you can do with this in finance is to look for the stories that we are telling ourselves about the world and then to take an orthogonal data source about stories that are actually happening and start to trade off the delta between them.

That could be very simply saying, I'm going to plug in the Indian news and just look at the stories the Indians are telling about X, and put in the American news, and look for the differences between the two of them. Or, I might take the stories about what we're doing with drones, and take the stories of all the patents that are being filed by Boeing, and any differences between what Boeing is doing that we're not talking about or what we're talking about too much that isn't being done.

Sean Gourley

Augmented Intelligence (Continued)

That in a sense like a trade on Boeing on top of that. Two different data sources that probably no one has read in their entirety is certainly now read across both of them, put the machine on that, and look for deltas and trade against it.

Question: What are some of your personal experiences?

Sean: There are a couple of things. One is if you think you know everything. You think, "I've got it all." That's a really hard thing.

One of the things I'm consistently humbled by is, I'll throw up all of the patents that have been filed in the last two years on natural language processing. I would always find stuff that's being done that I wasn't aware of.

I'll just take the last week of technology news. I live in the Valley. You live, breathe, drink, do the whole thing. I'll take that, I'll do an event extraction algorithm on top of that data feed, I will see things and go, "How the hell did I miss that?"

It will be a big event. It will be a new chip from Intel. How did I miss that? Experts that come to think they know it all struggle with that. I think it's always important to remind, there is stuff going on in the world you don't know about. You can't ever keep up with it all, so don't try. Plug into machines when you're needed to do it.

I think the other one that comes back are people that are looking for that answer. What does a machine tell me to do? It doesn't tell you to do anything. It just shows you the world. What does it tell me to do? You're going to have to make that decision yourself.

That's really hard for people to get past. They are so ingrained that the machine should give us an answer. We're so ingrained that there's a right thing that we don't appreciate that perhaps it's a little more subtle and that maybe we need to still be involved.

Question: What are the different choices available?

Sean: You can have a range of different connections and there's a range of options. Anything from explicit connections like A interacted with B, through implicit connections. An implicit connection could be the concepts of A are similar to the concepts of B. That could also be the emotional language in A is similar to the emotional language in B.

You can have a range of different things. What you choose to project is what you choose to deem to be the valuable connections. We try and give the user options to say, do you want this kind of connection or that type of connection. As you go down, you think of a high dimensional object that's connected in many, many ways.

You've got to project that to two dimensions or three dimensions so that we can navigate it. You're collapsing a dimensional space and trying to guess which collapsing to do. You can do reasonable jobs, but it's much better to give that as an option to the user so that they can have the connections that are meaningful to them.



David Ferrucci
Bridgewater Associates

David Ferrucci is an award-winning Artificial Intelligence ("AI") researcher. He joined Bridgewater in December 2012 as part of the research department to work on new approaches to capture and apply knowledge in markets and management.

Dave has over 25 years of experience in AI and software systems architecture. At IBM Research, he was the principal investigator who built Watson. Watson resulted in a landmark achievement in Artificial Intelligence and natural language processing and was awarded many public accolades including the AAAI Feigenbaum Prize. As a result of Dave's broad impact on science and business, he became an IBM Fellow in 2011.

Dave is published in AI, Natural Language Processing, Text Generation, and Software Architecture and Engineering. He earned a PhD in Computer Science from Rensselaer Polytechnic Institute in 1994.



David Ferrucci

Artificial Intelligence

Michael Mauboussin: It's my real pleasure to introduce our cleanup speaker today, Dave Ferrucci. Dave currently works in the research department at Bridgewater Associates, the well-known hedge fund. Long before I met Dave, I was a big fan of his work and the kind of work he does. Not necessarily directly related to Dave, but I certainly followed closely IBM's efforts to build a computer that could beat Garry Kasparov, which, as you all know, they succeeded in doing in 1997.

But when I heard that IBM was building a machine that would take on the champions in the game of *Jeopardy!*, that was exciting and I was even more hooked on the work. The man behind Watson, which is the champion-beating, *Jeopardy!*-playing computer, was Dave Ferrucci.

Chess is basically a simple game but it does have enormous possible combinations of outcomes. But it's also a game that computers are well designed to be good at. Basically, it's number crunching.

Jeopardy! is a very different ball of wax, as we'll hear. The questions require an understanding of similes, puns, riddles, language processing has to be excellent, and so developing a program to beat humans in that realm where humans are really strong, as well as having all the other information, that is truly extraordinary.

Dave Ferrucci has more than 25 years in artificial intelligence. He's received numerous prestigious prizes in the field. He's going to tell us about this field and what he's learned over the years, and what's to come. So please join me in welcoming Dave Ferrucci. Dave.

[applause]

David Ferrucci: Thank you. It was great to hear Sean's talk because the theme behind that talk is going to be very compatible with the theme that you're going to see in my talk as well. He also did a lot of the heavy lifting, so I'll be able to chat with you about philosophy instead of going into all of that deep stuff. [laughs] It was very compatible.

I will say a couple of things right at the outset, sort of how Sean's talk left me thinking about artificial intelligence versus augmented intelligence. Right now, you and I talking, I'm augmenting your intelligence, you're augmenting mine as we we're dialoguing.

The reason that can happen is because we're both intelligent, we both speak the same language, we have compatible processes and representations. You can think of when two people collaborate, they're augmenting each other's intelligence. But on the other side of that fence, it requires a level of intelligence to do that.

I thought what was really interesting was when he talked about how some people can interact with computers really well and some people can't. I think it's the responsibility of people in AI, right, it's the responsibility of artificial intelligence to build computers that make that interaction easier, that make that interaction possible, for the majority of people to sit down and actually get value out of interacting with that machine.

Anyway, it's an interesting perspective, lots of interesting stuff here. I'll talk a little bit about the evolution of AI from theory driven to data driven systems, and this idea of where the sweet spot is. You'll hear the echoes from Sean's thoughts there, reflections on IBM Watson, a landmark in artificial intelligence and where that fit in, and my view of the future of AI, and how that will evolve.

Just to go from the very beginning, artificial intelligence, who even coined the word? Where does it come from? What does it mean? Two interesting perspectives on this. One, very demanding, which was Turing's perspective, which was computer systems whose interactive behavior is ultimately indistinguishable from

David Ferrucci Artificial Intelligence (Continued)

the humans. How many of you have read that the Turing test, the Turing thing was solved? How many of you believe it? Oh, good.

[laughter]

Has anybody seen the dialogue from the machine that supposedly passed the Turing test? It's really pathetic. How many people have heard of the ELIZA program from the 1960s? Good? It's a little bit better than that I think, not that much.

McCarthy in the 1950s coined the term actually, and he said artificial systems are computer systems that perform tasks that if performed by a human would be associated with intelligence. It's a very different kind of definition. It's sort of ingenious in a way because what it does is it avoids really defining intelligence entirely, and it really puts it on the task.

So, we have Deep Blue and we say "Wow, a computer, Deep Blue, beat a grand chessmaster, and people who can do that, I would associate them...They'd be intelligent if they were able to do that." A computer program beat the best at *Jeopardy!* OK, I associate with that a certain kind of intelligence, so those things are considered artificially intelligent.

But of course, Deep Blue couldn't walk nor can Watson walk off the stage, start conversing with you, talking about economics, making jokes. Neither of those are going to happen. So it's just that task that they did, we say "Gee, if a human did that, we would think that human was intelligent." Sort of a very different kind of definition.

AI has come a long way actually. I don't know how many people are familiar with a lot of the incremental advances that have been going on. One way to divide the world is into the idea of knowing—intelligence as it means to know something, like to interpret, to understand, to reason, to learn, to predict—and then doing—being able to take some action, and do that action in an effective way, walk, see, sense, catch, drive, fly, you know, that kind of stuff.

I'd like to start here. How many people know about the Google car? Everybody's got to know about the Google car. I don't know if you've seen some of the things that are out there. If you haven't, I was going to click on a couple of these things. I have a couple of YouTube videos for those of you not familiar with this kind of stuff. Oh, there's one.

Video: History Grasping consists of throwing and catching, which contrasts with traditional methods that maintain contact state.

David: That's pretty good. Can you do that?

[laughter]

Video: ...In the air, and then move to intercept it. This is an extremely dynamic task with a lot of uncertainty, with very fast dynamics, yet this technique for designing control systems is very robust, and...

David: How many people saw this? This is the big dog. Basically the dog can walk on ice and balance itself. It's learned how to balance itself. It's actually quite...I can't do that. Can you do that? I can't do that. I would just fall.

That's a certain kind of intelligence. You're sensing. You're taking in that sense data, you're responding to it. This you can actually buy.

David: Sound localization, so it's got hearing, it's got pressure sensors, it learns how to walk over different surfaces.

David Ferrucci Artificial Intelligence (Continued)

Dave: Anyway, and then of course the Google car, which is going to change everybody's life. It's amazing. In interpreting understanding, reasoning, and learning, and predicting it's a little bit more challenging in many ways, because what it requires you is to kind of get at meaning, and meaning is ultimately connected with how other humans think and interpret things, because you're taking human data, and you're saying "What does this mean, and how do I understand it, and how do I reason about it in a meaningful way?"

How many people saw the game against Watson, the *Jeopardy!* game? How many people? Not a lot. Those of you who watched it, or didn't watch it, how many people were rooting for the computer?

[laughter]

David: How many were rooting for the humans to win? Are they allowed to stay?

[laughter]

David: It's kind of interesting. Those of you who didn't see it. Here's a little clip.

[video clip begins]

Alex Trebec: Watson?

Watson: What is keratin?

Alex: You are right.

Watson: The Art of the Steal for \$1,200.

Alex: Answer, the other Daily Double.

[applause]

Alex: Watson, what are you going to wager?

Watson: \$1,246, please.

[laughter]

Alex: \$1,246. Here is the clue in The Art of the Steal. The ancient Lion of Nimrod went missing from this city's national museum in 2003, along with a lot of other stuff.

Watson: I'll take a guess. What is Baghdad?

Alex: Even though you are only 32 percent sure of your response, you are correct.

[applause]

[clip ends]

David: A lot going on there. One of the most interesting things about that was that answer panel. In my view...I fought very hard, there's a back story, I won't tell it, I fought very hard to make sure that answer panel ended up on television. Because that answer panel that showed the top three answers with the confidence, I think changed everyone's impression about what was going on.

If you would have just put that computer up there and it answered when it did and you didn't see anything when it didn't answer, I think it would have been completely left up to people's imagination about how they thought computers worked. They would have projected that onto the computer.

David Ferrucci Artificial Intelligence (Continued)

And you know what the majority of people think is going on inside of a computer? It's a rich spreadsheet, a big table. All the questions are already there. The question comes up, you look it up and you look up the answer. That's how people would have imagined it. But as soon as you put that answer panel up there, you say wait a second, it's considering multiple answers.

It's coming up with a probability that one might be right. It must be thinking. The impression of what's going on, whether you call it thinking or not, but it completely changes once you put that answer panel up there. I think that really changed people's perception of what was going on to actually become more aligned with reality.

We talked about chess a lot—finite, mathematically well-defined search space, all responses grounded in precise unambiguous rules, a large but finite set of moves—somebody mentioned 10 to the 40th, clear consequences. It's perfect for a computer. The really surprising thing is that humans can do it.

We learned earlier about how they actually do this somehow differently than just looking over the broad possibilities of moves, but they actually remember these various kinds of patterns and things like that. Human language is a completely different story. Words, they can be images or speech, what I mean by language is just how we communicate with one another.

We can communicate with words. We communicate with images. They lack precise interpretation. There are nearly an infinite number of expressions mapped to a huge variety of meaning. The meaning is only grounded in shared human experience. We tend to think that...when we started with computational linguistics we'd say, "Oh, there are grammars and there's words and words have precise meaning."

The reality is you can think of words as having really no intrinsic meaning. You can think of them more as this intricate indexing system into our shared experience. I use a bunch of words and I'm speaking to you and they're lighting up different parts of your brain. They're reminding you of different experiences. And you're sorting out those experiences.

And as I'm speaking to you, you're nodding. And I'm thinking, "Oh, OK. Those words seem to be lighting up the same experiences. That person seems to be understanding." For all I know, you have no idea what I'm talking about. In fact, you might have some idea, but you don't have exactly the same interpretation I have, which might even be worse, because now you're nodding and you're saying, "I understand." And then push comes to shove and I expect you to do something. I predict that you're going to behave a certain way because I saw you understanding my words. And you do something crazy. It turns out you don't understand me at all. Even though I used the words, but they were lighting up different experiences in your head.

This is a great example this from my own life. I do these scientific experiments at home. I have two young daughters. They'll hear me say, "Hey guys, come down here. I'm doing this thing. It's really, really interesting." They would come down, and I'm doing something crazy in the sink or whatever.

I'm doing this over and over again, "Guys, you've got to come down here. This one's really cool. This one's really, really interesting." One time she was seven at the time and I go, "Come on down here." She stopped at the top of the stairs and she said, "Daddy, interesting things are boring."

[laughter]

David: Because of her experience with that word, she assigned "boring," she assigned a completely opposite meaning to that word based on her experience with that word. That's really how language evolves. It evolves through context and through the assignment of that shared experience to those words.

David Ferrucci Artificial Intelligence (Continued)

It's getting at that meaning and that meaning is really sourced in humans and how humans experience the world.

Here, we have this picture. Whatever you see there, don't even tell me. I can interpret this in lots of different ways. A computer stores that image and we can move it around; take it with a camera and then send it to you in a message or an email.

But it's just a bunch of zeros and ones. What does it really mean? When I start looking at a human, because it's optical system can identify colors and say, "There's white, there's black and then red." Now, I can do object detection. I can say, "I know what that thing is. It's a calculator with some dice on top."

I can think, "You know what they meant to tell me? They're calculating the odds. This image was actually labeled by humans with this word here, winning." I wouldn't even have made that connection, but a bunch of humans did.

Another way is human use of tools. Maybe that's what it means. Different levels of interpretation, just a bunch of zeros. Where do those interpretations come from? It comes from our experience with that thing. That thing, that artifact becomes part of our language. We use it to light up those different experiences in your head.

Ultimately, meaning is subjective and we the humans are the subjects. How do we resolve meaning? It's the context that becomes so important in triangulating that meaning given those words. You can think of meaning as this probabilistic of mapping from symbols to a common experience. Richer context narrows those possibilities and improves confidence in that mapping. What are you talking about?

"The bat was flying toward him." A couple of ideas. Those words start lighting up things. There's lots of possibilities. We're just showing two competing possibilities – a bunch of vampire bats or fruit bats or whatever, flying toward a person. We could imagine a baseball bat flying toward me under different conditions; one is my ex-girlfriend, another one is maybe someone hit a home run through the bat. I was in the audience. Different possibilities.

"Billy ran as fast as he could." Now I have two competing possibilities here. One is running away from the bats chasing Billy. The other one is Billy just swung and the bat went flying.

"He made it home safe." Still not really sure. I still got two competing possibilities with more and more context.

Finally, "He scored." I'm pretty sure now it's the baseball game. There's another interpretation I won't mention, but I'm pretty sure it's the baseball game at that point.

As more and more information comes in, I take those words and I start connecting them to that meeting. Really interesting though because subtle differences in the expression can dramatically change the meaning that it happens to light up.

These small changes, it depends on how those words are used, not just the words themselves, but what order and sequence they come in. Here, I just took these words; "safe home" and "at," and I just mixed them up a little bit. I put them into Google image search. Here's what came out. Don't forget, Google search is very much driven by how people annotate things. It's the source of people saying, "This is what this means to me. No, this is what it means to me." Connecting the meaning with the words.

"Safe at home," I get these images, mostly baseball, except for this one right here of some band or something. I don't know what that is. "Home safe." "Safe home." Really different interpretations. It's

David Ferrucci Artificial Intelligence (Continued)

amazing how it makes that point. It's all about how the words are used, not just the words, but even the ordering of the words.

When you look at AI from its very early beginnings, it's, "I want the computer to understand things, answer questions, and make predictions. How am I going to go about that?" The idea was to build theories, to model how the world works. To say, "If $A(x)$ is true, then $B(x)$ is true. If the conditions change and I have $A(x)$ and $C(x)$...and to model all those rules. To identify concepts and relationships between them, usually working on small data.

It was also interesting, Sean talked about the difference between the challenges, that you need more intelligence with small data. More ways to interpret and understand than if you just have big data.

Anyway, scientists worked with small data because they didn't have big data in those days. They interpreted and generalized, came up with an anthology, different concepts, general concepts, more specific concepts, connected them to rules, came up with the theory. Then would ask a question, and they would get these predictions and answers.

They were explicable. Why were they explicable? In other words, it was an open box. In the sense that you could say, "Why'd you come up with that?" The computer can go back and trace all its reasoning, identify the concepts and the rules, and give you a deductive proof.

I just pulled this out of Nate Silver's book when he was talking about the description of why somebody thought there was a recession: "Consumers have extended too much credit to pay for homes that the housing bubble had made unaffordable. Many of them had stopped making their payments to the degree of leveraging the system would compound the problem and so forth and so on. You're demonstrating, 'Here's my understanding of how the world works and why I might predict that there's going to be a recession.'"

Data driven, since things started changing over the last 15 years, huge amounts of data becoming more and more available, tremendous amount of compute power to munge over that data. Lots of different statistical machine learning techniques. We had the advent of big data. A is correlated with B and that may suggest some causation, so I'm good. Show me A and I'll predict B.

Really interesting results in a broad number of areas: health care, economics, e-commerce, Netflix, Amazon, economics, elections. We talked about I think all of them today. What you get back is you get back predictions, but there sort of inexplicable. There's no human compatible theory underlying there.

I can tell you all different features I may have considered. There may be a ton of them, but how are they related and what do they mean? I get a very different kind of prediction or explanation for prediction. The most reliable forward looking indicators are now collectively behaving as if they did on the cusp of a full blown recession. Very different kind of explanation for a similar prediction.

I think this was mentioned, too, I think it's an interesting debate. It's actually been going on for years. It's not that new. People have been talking about induction and deduction forever. It's just kind of made it into the mainstream.

One extreme represented in 2008, Chris Anderson when he was editor-in-chief at *Wired*, "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation whatsoever." It's a fascinating point of view, very extreme.

We had, 2012, Nate Silver saying, "Throwing the theory out is just categorically the wrong attitude. Statistical inferences are much stronger when backed up by theory." And we saw that theme with a couple

David Ferrucci Artificial Intelligence (Continued)

of the speakers today. Yes there's information in data. It gets harder and harder to grab it when the data gets bigger and bigger.

It's an interesting relationship to think about, as the data is growing exponentially, are my computational methods and my algorithms growing fast enough to extract the meaningful data out of there? How much theory do I need to identify whether or not the correlations are meaningful? This becomes a huge opportunity, but also a huge challenge.

What I want to do for the next few minutes is take you through a painfully simple, and I already apologized for all of you who understand this in great depth. I'm just going to take you through this painfully simple example, to fine tune the point, the difference between a data driven approach and a theory driven approach. I have a painfully simple decision to make: should I wear galoshes on my walk to lunch? Because it's been a bit raining, so it's complicated.

First I'm going to start my theory driven approach. I'm going to model the world. These are quasi-versatile logic, it's not perfect because I don't remember all my versatile logic. There exist surfaces. There exist things called paths.

All paths are surfaces. Surfaces can be covered. Surfaces can be wet or dry. There are events. Raining is a type of an event. Something is wet if it's raining and not covered. There are things called people. People wear galoshes. When it's walking and when it's wet, people want to wear galoshes.

I've modeled the world. Here it is. Now, I build the system that does deduction over that. I ask the system, "I will be walking to lunch, should I wear galoshes?"

The system asks, "Is it raining?" It says, "I'm using rules nine and seven. It says, "Yes."

"Is the path covered?" Using again those rules, user, "No." System, "You should wear galoshes." User, "Why?" Because if it's raining, and the path is not covered, and the path is wet, if the path is wet, people wear galoshes. I'm assuming you're a person.

[laughter]

It's fabulous. It's fantastic. I love it. It gives you this theory based description of what you should do.

I want to advance it. What if it's not raining, but the path is still wet? Life gets always more complicated than you'd like to think it is as you build out these theories. But having a theory allows us to engage human thought, intuition, perception, all the things humans are aware of that is very difficult to make the machine aware of, but it can get challenging.

If I want to do this, I have to start thinking, "What's the duration of the raining event? When did it start? When did it end? How long does it take to dry?" I have to worry about ground retention, temperature, and humidity. There might be depressions in the surface. They fill up. They're called puddles.

I have to think about the topology and the retention of water depending on the topology. It takes time for these effects to dissipate and so forth. I end up with an extremely complex problem to try to model in a theory-driven way. I imagine giving you a really robust explanation, but this becomes exceedingly difficult to do and to maintain, and to grow.

It's like the data driven approach. I have my galoshes data. I have all these observations. I have whether it's raining or not. I have people annotating and say whether or not they would be wearing galoshes under those circumstances.

"Should I wear galoshes?"

David Ferrucci Artificial Intelligence (Continued)

"Yes."

"Why?"

"85 percent time it rains, people wear galoshes."

"Great." "Why?"

I can add features to this. I can say, "Well, its prediction qualities are not that great, so I want to throw in more features. I could throw in wet path, I could throw in covered, I could get maybe better predictive results." What if I want to advance my system? It's still not as good, so I want my version 1.1, so I throw in more features. I can just, shoe type, location, tree type, season, temperature, humidity, start time, whether the Red Sox won or not, whether the paths are covered, all kinds of possible things.

I could throw this into the same machine learning algorithm. It should say, "85 percent of the time, that well you should probably just wear galoshes." "Why?" "Just look at this giant table and you do the math."

[laughter]

The relationships and the meaning and how I think about these various features, and how they build out a model of how the world works is just missing from this particular approach even though it may be predictive. If I have to now explain myself or be liable for my decisions, what do I have to point to? How do I now engage in whether or not I believe that prediction in that particular case? I'll give you a great example from my own life.

A little bit serious but it's OK. I'm over it. My dad many years ago went into cardiac arrest in a restaurant. An ambulance took 10 minutes to get there, rushed him to the hospital. Finally he gets there.

The resident comes out and says, "Look, your dad's brain-dead. We want you to sign a "Do-Not-Resuscitate." I said, "How do you know he's brain-dead? He said, "Given where he was, how long the ambulance took, statistically speaking, there's a 98 percent chance he's brain-dead." I said, "OK. There were a hundred of people in the restaurant. Two of them would have survived this. How do you know he's not one of those two?" "OK, you need to speak to the chief cardiologist."

[laughter]

So, I got on the phone with the chief cardiologist. He said, "I know this is very difficult, but we understand these things. We have the statistics. Your dad's brain-dead. The responsible thing to do is to sign a "Do-Not-Resuscitate." I said, "Do you have any deductive evidence?"

He said, "What?"

I said, "How do you know he's brain-dead. Do you have an EEG?"

"No we don't have an EEG."

"Well, how do you know?"

"His pupils are dilated."

I said, "Doing the deduction, are there other reasons why his pupils might be dilated?"

"Well, yes, there's a drug they give at the ambulance that might dilate his pupils."

[laughter]

"What else you got?"

David Ferrucci Artificial Intelligence (Continued)

"I have to come down there and talk to you about this." He came down and we debated this a little bit further. Eventually, he gave up. He was very frustrated with me. Long story short, about 18 hours later, my father was sitting up in bed with zero brain-damage. I would not sign the DNR.

What went on there was while the statistics apply in general, in that particular case, there's very specific evidence that would take you one direction or another. You have to go through and I wanted to go through the deduction. It was a very long process because what the doctor made me do was, because he was against the way I was thinking, he made me make every decision. He said, "What are going to do now?"

I would ask him questions back and forth, then I would have to make a decision. It was interesting how I just want to go with the statistical evidence.

The Holy Grail ultimately is to build systems, when we talk about artificial intelligence, is to acquire, understand, predict, and explain. One thing to observe is how engaged the human is in all of this process. What data is important to even acquire to answer certain kinds of questions? What model do I have to build to represent an understanding? What algorithms do I need to predict? What is worth predicting?

Ultimately, if I make a prediction, can I explain it with regard to the understanding? I'd like to automate this whole thing, and you could think about the human in the source of meaning, coming up with the objective function, coming up with what it means to have a compatible understanding and a meaningful explanation, is engaged in this whole process.

On one hand, I think that to build artificially intelligent systems, we absolutely have to engage the human in every step of the process. The idea is, in what way do we engage them? Can the system incrementally learn from that engagement just like a child interacts over and over again with its environment and with other humans and starts to learn what the meaning of things are and starts to mimic and extend that meaning through those interactions?

The world is changing. We're coming like full circle, we went from theory-driven to power data-driven, which works, of course, exciting, but you're seeing the observations made at this conference as well as you see this other now, 2014 *Wired*, talking about how you need more than just data:

"By claiming that computers will ever organize our data or provide us with a full understanding of the flu or fitness or social connections or anything else, for that matter, they radically reduce what the data and understanding means."

And it talks here about having to engage people and build the model of that understanding."

I think there's a growing and growing awareness that yeah, huge amounts of data out there, huge opportunities, but what does it mean to just find these correlations and not be able to map them to theories and causation?

It's interesting to reflect on what represents difficulty in this space. When you think of very, very low level computer applications like traditional information technology, this one level of difficulty was everything was structured and deterministic. Everything's in databases, I know exactly what all the columns mean. Basically, what I'm doing is I'm selecting information out of the database, maybe I'm doing some math on it, and I'm shipping it somewhere.

It gets a little bit more challenging when you think of predictive analytics, because now what you're looking at, is you're looking for patterns that are...You don't know what the patterns are, so now you're trying to discover the patterns. Here, the data is completely understood and it's almost selecting from it. You're

David Ferrucci Artificial Intelligence (Continued)

actually looking or discovering patterns, but you don't know what the patterns mean. You're forming correlations, but you're wondering about causation.

Then it gets even harder when it's unstructured and probabilistic, meaning, "I'm not even sure how to interpret the data." That's like with language. Language is highly probabilistic as we saw earlier.

These are ways to stratify difficulty in getting the computer to do these kinds of things. Where does Watson fit into this space? Watson's an interesting point in the spectrum. Here we had this grand challenge for advancing AI, but specifically, in the areas of natural language understanding and the challenge went something like this.

I have this broad open domain, meaning you're not going to tell me ahead of time what the questions are. There isn't a spreadsheet where all the questions I'm going to ask you are in there, and you have to look up the answer. I don't know what they're going to be. I don't know what they're going to be about.

Complex language, it's written all sorts of different ways. "If you're standing, it's the direction you should look to check out the wainscoting." Anyone? It's an easy one. \$200. Down, down is the answer, down.

That's interesting, because if you sat there as a theory-driven guy, you're like, "OK, I got it." Don't forget. You'll never see this question again. How much you do invest in understanding that question in building Watson? You're never going to see this question again. He's like, "OK." How do I generalize?

I won't get this question again, but I might get a question about direction. Let me model what a direction is. Let's see. I got up-down. I got left-right. Maybe I should do like 360 degrees around a compass, I could do that. There's relative direction behind me, in front of me. I build this big complex model, probably never see a question about it again.

Oh, a question shows up. It refers to the direction of fabric. The answer is green. Would you have thought to model that, having looked at that question? Probably, you would've missed it. Very challenging to imagine, if the domain is too broad. Let's see. "In cell division, mitosis splits the nucleus and cytokinesis splits this liquid cushioning the nucleus." What are you, finance guys?

[laughter]

Cytoplasm is the answer. 600, you've missed that. You're paying if anybody gets it, right? "Seems this perp was the first murderer in the Bible and to top it all off, he iced his own brother." I'm good. There we go, 800.

Why is this a hard question? It's interesting, because if a computer was doing this, it has to...Hinting to you how Watson works, but it has to start finding connections between words to come up with, what's the common expected meaning of these words. Seems this perp—what's a perp—was the first murderer in the Bible. To top it off, he iced his own brother.

I go to a dictionary, *Webster's* or something, to find what "iced" might connect to, and I don't see murder there. Where do I have to go? It turns out if I went to the *Urban Dictionary*—it's the first definition in the *Urban Dictionary*—"iced" is murder. The problem is, the *Urban Dictionary's* like 50 percent pornographic, bad source to use on national television.

[laughter]

"I tell you, it was so cold today. How cold was it? It was so cold, I wish we were back in '64 when he was emperor. Hot times, if you know what I mean." You are very good, \$2,000.

David Ferrucci

Artificial Intelligence (Continued)

There's so many words in that thing. What's meaningful? What is it asking? You as a human connect to that, because the '64 in Rome, or the emperor lights up for you, but how do I know which should light up and what shouldn't light up if I don't have that shared experience, that I just got a ton of words?

"Of the four countries in the world that the U.S. does not have diplomatic relations with, the one that's farthest north." North Korea, that's right. You have to be able to get these questions. You have to be able to have high precision. To compete with the best, I'll tell you how good you have to be in terms of precision. You also have accurate confidence, meaning you have to know what you know. Because if you buzz in and then get it wrong, you lose that dollar value and your competitor gets it.

So you have to have a good prediction of whether or not you're right, and you have to have a decent probability, an accurate probability associated with that. You have to do it really, really quickly. Within just a couple of seconds; about two to three seconds to compete with the best.

They ask about all kinds of things. Here's a histogram of all the kinds of things, *Jeopardy!* questions, we look at 20,000 randomly-sampled *Jeopardy!* questions, and they ask about a ton of stuff like films and groups and women and singers and persons and founders and diseases and makers and objects, whatever an object is, a heroine, vegetables, months, pets. It's a very long-tail phenomenon.

Even if you tried to model everything you knew about the top few things here, you'd only cover 10 percent of the questions. No way you're going to compete. You have to do something a lot more general than that. You're not going to sit down and build this complex theory of all common sense knowledge.

13 percent of the questions didn't even tell you what the type was. It just said "this" or "it." Everybody thinks, "Oh, the categories tell you what's going on." Not true. Here, the categories were weak indicators of what the answer type was.

Here, if we have U.S. cities but the answer was shuffleboard in this one. The answer was the Erie Canal in this one. Here, we have country clubs. The answer was a mace, a baton; League of Nations was another one there. Author, the answer was the book of *Job*. The answer was Romania.

The categories didn't tell you exactly what the answer type was. It was generally what the answer type might have something to do with. It wasn't as strong of an indicator as most believed. Just understanding the question's a challenge. Here's a question. This actor, Audrey's husband from 1954 to 1968, directed her as Rima the Bird-Girl in *Green Mansions*. What's it asking for? It's asking for the director of the film *Green Mansions*. Did you know that?

You just have to parse the questions, just the syntactic parts. Like back in your grammar school when you wrote out the grammar, but then you got to start interpreting the parts. A husband's a word. It modifies actor but what does it mean? It's not like I can go look it up in a husband database. I don't even know what this means in this case. What about direct? Direct guy, direct like direct a film, lead? What does that mean? Director of...Who's Rima? A person, a bird, a girl, a character. Green Mansion's a place, a movie, a play, a book, a house?

I have to make predictions on everything. Those predictions are going to be probabilistic. I have many of them. I have to pursue all of them and start gathering evidence, building context, narrowing it in to decide, "Which one is more likely? What's the right interpretation here? Which one's more likely the right interpretation over another one?"

If I was Google, I could just throw you a whole bunch of documents that contain those words and you could find the answer. But Watson couldn't hand Alex Trebek 1,500 documents and say, "You find the answer."

David Ferrucci Artificial Intelligence (Continued)

[laughter]

Here's a question. "Treasury Secretary Chase just submitted this to me for the third time—Guess what, pal. This time, I'm accepting it." Resignation, very good.

Did you do that because you know the history or because you use plausible inference? B? Yeah, plausible inference, that's right. Because you're sitting there, you're saying, "I don't know the history," but I sit there, "What kinds of things get submitted?" A bunch of things. I use more context. What types of things get submitted to a president? Probably still a lot of things. Then I use a meta inference, a meta plausible inference. What types of things get submitted to a president that's worth talking about on a television show? You say, "Resignation." Very good, that's the right answer.

Now, you ask this question to a bunch of sixth graders, and they use plausible inference as well, but they have a different shared experience and they come up with a different answer.

["Friend request" shown on screen followed by laughter]

It's reasonable! Watson couldn't always interpret the question such that it could find an exact passage that match, then it would have to basically make predictions. One of the things it did to enhance its intelligence was it would analyze a whole bunch of text, and it would do these syntactic frames, like do these parsing in front of the subject, the verb, the object, and all the modifiers, and then it would analyze them with some semantic knowledge and then generalize.

It would get things like inventors patent inventions, officials submit resignations, people earn degrees at schools. Watson didn't even know that, Watson knew nothing on its own. I didn't write a single rule, or no one in the team wrote rules...

It basically inferred things from analyzing text and looking up patterns. Officials submit resignations, people earn degrees at school. You say, "Why did it have to learn that?" It learned all kinds of things. We didn't dictate what it would learn, we basically let the data drive the inductions that it would make.

That's kind of the interesting thing, because I asked you, "Oh, where did you get your degree?" I might answer with the city. That would be reasonable. But more often than not, people answer with schools. That tells me something. It tells me, if somebody asks a question about where you got a degree, I might more likely answer with a school than something else, because that's what's expected based on the model, the language model that was inferred from all of this human text.

"Fluid is a liquid, liquid is a fluid." Which is true? Any physicists out there? Liquid is a fluid, because there are other fluids like gases and plasmas. When we use language, people use it interchangeably. If you went and you looked at a formal taxonomy, you would see a liquid is a type of fluid, but people use it interchangeably.

Vessels sink, but people sink 8 balls in the game of pool. See those words there? Pool game, that represented a context. That bag of words would represent a context, so that would limit how these inductions were made.

Parsing sentences, trying to find answers, inductive techniques, I'll talk about a few more techniques. But let's step back and think about how hard was this task. This is a way to measure our progress. This was called the winner's cloud graph up here. We have on the x-axis, is the percent of questions answered by the winner of the game. Every dot here is a *Jeopardy!* game. What I'm plotting on the x-axis is the number of questions that the winning player buzzed in for and got a chance to answer.

David Ferrucci Artificial Intelligence (Continued)

If you look at the center of the cloud, I don't know, something around 48 percent. The winning player, on average, was fast enough and confident enough to buzz in for about 48 to 50 percent of the answers.

Of those questions, how many did the winning player get right? Somewhere between 80 and 90 percent, about 85 percent. The black dots are Ken Jennings. Ken Jennings is a phenomenon unto himself. He won 72 games in a row, something like that, 71, 72. His average game, he acquired 62 percent of the board. On *average*! That's crazy. He played a game where he acquired 82 percent of the board. You don't even know who else is playing. He's just sitting there, going like this, answering questions.

82 percent of the game, he acquired...On average, he just clobbered...Look at that cloud. That's unbelievable. In 2007, this was the state of the art, open domain, question-and-answering system. Based on all state of the art techniques, my team at IBM was working on this for a number of years before we did *Jeopardy!* Before, this was called the confidence curve. The way to read this is if it sorted all the questions based on its confidence; this would be the five percent of the questions it was most confident in.

So, that system turned on *Jeopardy!*, this was its performance. 5 percent that was its most confident did about 48 percent accuracy. As it had to answer more and more, it's less and less confident, and it plateaued at about 13 percent. That's a big gap, huh? What was the question?

Question: What was the issue?

David: What happened was its confidence curve stunk. It wasn't predicting very well. What would a really good confidence curve look like? Where it thought it was doing a good job it was. It'd be along the top. As it predicted it was doing badly, it would start to do badly. That would be a more accurate confidence curve, it would be shaped like that. This was not doing a good job in predicting in any way, so lots of problems.

How are we going to solve that problem? I talked about it before when I gave you a hint with the direction question, but we saw this question: "In cell division, mitosis splits the nucleus and cytokinesis splits this liquid cushioning the nucleus." I'm sitting here, I told the executives, "I'll solve *Jeopardy!* in three to five years." Can you imagine me taking each one of these questions and building out a model like this, representing all of biology to get this question I'll never see again? That just wasn't going to happen.

It's an interesting way to solve it, because then I could say, "You know what, let's see, it splits the nucleus cushioning. What liquids cushion the nucleus? What does it mean to cushion? Let me find out what the structure of the cell is." It would be great, because I explain my answers, but this is tough to do.

Let me give you a hint that's closer to how Watson solved the problem in that. "In May 1898, Portugal celebrated the 400th anniversary of this explorer's arrival in India." Anybody?

David: Very good. Now, I can break this down to a bunch of keywords and I can do a search. Because of PageRank on Google and how popular this is because everybody learns it in high school, you're probably going to get Magellan up there and Vasco da Gama up there, but you might get Christopher Columbus, too, because Christopher Columbus hits a lot of these keywords like Portugal and the time frame and India. He thought he was going to India, anyway.

But if you just strictly looked at keywords and no understanding of what's going on, you could pick up this passage. "In May, Gary arrived in India after he celebrated his anniversary in Portugal." It's got all the keywords in common, that's fantastic. Keyword inference is amazing. And Gary's the explorer. Who's to say Gary's not an explorer? We're all explorers in our own way. You can imagine the next sentence.

David Ferrucci Artificial Intelligence (Continued)

The next sentence could legitimately give evidence that Gary's an explorer, because it might say, "Gary returned home to explore his attic, looking for a photo album." Here we are, we have Gary as the subject of the verb "to explore." That's evidence right there. Gary's an explorer. Is it good evidence? Probably not.

Now, we have the same question. Now, we have, "On the 27th of May 1498, Vasco da Gama landed in Kappad Beach." That's interesting, because now, the only keyword I have in common is May. This is not even going to come up, but it's got a lot of interesting stuff in it.

As long as I had a whole bunch of algorithms that can do the matching. Can I match a 400th anniversary in 1898 with 1498? Can I match arriving with landed without having to write a rule for every two words?

With statistical paraphrasing, I could find that gee, "arriving" and "landing in" appear in very similar contexts. Maybe I get some signal that they might mean the same thing. I get some signal here, I get some signal here. And then "Kappad Beach," with geospatial reasoning, I guess I could look this up in a database, here I can say, "Kappad Beach is in India, so if I arrived in Kappad Beach, I've arrived in India." Now, I got Vasco da Gama looking pretty good. Now, I kind of find out evidence that Vasco da Gama is an explorer, probably more evidence than I can for Gary.

If I could weigh that evidence, I'm starting to feel like this is a better answer than Gary. This is more of the way that Watson works. Here is that cytoplasm question again. It doesn't actually build a model of biology, but it starts finding lots of passages and figuring out if there are paths for making the connections from the words and phrases in the question to the different things that it reads.

Can split mean divide? Can reduce mean split? Can myotic imply mitosis? Can liquid imply fluid? If I could find resources, other passages, dictionaries, thesauri, anything that can help me make these connections, I can actually start to build evidence that this passage may support that answer. This is exactly what it did.

Here we have that question. We analyze it, we parse it, we come up with the key terms, the answer type, I'm looking for liquid, some of the relations. We then do lots of searches to come up with lots of documents.

From those documents, we pull out possible answers, not really knowing anything. Well, maybe it's organelle, vacuole, cytoplasm, plasma, mitochondria,...all possibilities. For each one of those possibilities, I assume that's a right answer. Think of it this way. These become competing hypotheses.

I've used search to then use some algorithmic techniques to pull out possible answers from those documents. Each one now becomes a competing hypothesis, and I can go off independently and say, "Can I get evidence for this? Can I get evidence for this? Can I get evidence for this?" For each one of those, I get evidence.

Let's say I have a question, I come up with 100 possible answers. For each answer, I get 100 possible supporting pieces of evidence, so I have 10,000 now answer-evidence pairs. For each piece of evidence, I now want to analyze it with algorithms. I have 100 algorithms that look at each passage and go, "Does this match this other thing in this way or that way or this way or that way?"

Now with 100 algorithms on 10,000 pieces, it's like a million different scores. Now I've got to rank them, and this is where the machine learning comes in. Why should I believe this score? Why should I give this score this way versus that way? I have to do this giant multiple regression, and I use machine learning to do that.

I train on previous *Jeopardy!* answers and questions for which I know the right and wrong answer, and I get a big model, a big set of weights that I can apply to all these different algorithmic techniques that I

David Ferrucci Artificial Intelligence (Continued)

evaluated all these answers with. That allows me to rank the answers and to build a confidence value that I can then map to a probability.

The hard part was coming up with all the algorithms. This is the basic architecture. Do the question and analysis, bring in all these sources, generate the different hypotheses, find evidence for each hypothesis, apply the various models based on training, and then rank the answers.

So, 25 AI scientists, software engineers, at IBM research and university partners, built on decades of groundwork. We used search engines. We used machine learning techniques. We used natural language processing. We advanced, principally, natural language processing techniques and ways to combine them with machine learning.

We did, during a four-year period, over 8,000 documented experiments, tons and tons of experimental data, dwarfing the actual data we used to drive the machine itself, which was the equivalent of about two million books – not a lot, in the order of 100 gigabytes. After we analyzed 100 gigabytes, we came up with about a terabyte of raw, analyzed data.

This is our progress. Somebody asked about the confidence curve. This is where we started. After four years – this is top line here – look at the shape of that curve. It's beautiful. As we were getting less and less confident, we were actually getting more and more wrong.

If you look at where Ken Jennings was competing, we were doing just about what he was doing, close to 90 percent, so we could play. We could play *Jeopardy!* We can give Ken Jennings a good game. Would we necessarily win? Not necessarily. Went into the game with about a 73 percent chance of winning, so that's a 25 to 30 percent chance of me losing my job.

[laughter]

David: I was told, "You must win."

[laughter]

It was competitive. It was actually more competitive than you might have thought because of the daily doubles, but we ultimately won. This was very, very good performance. It took every bit of those four years to evolve those algorithms to get them to the point that they were doing that well enough.

It's interesting to step back now from an AI perspective and say, "What just happened here?" We just beat a human with a brain. What did we beat it with? We beat it with a machine, about 2,880 cores, running in parallel.

Remember when we generated all those answers, and each answer had to generate a bunch of passages. Every one of those algorithms had to run those passages. I basically divided all that up. They all ran in parallel, because if you ran that computation on a single machine, a single three-gigahertz machine with 64 gigs of RAM, two hours to answer a question. It would have made for a very boring *Jeopardy!* game.

[laughter]

David: With that parallelization, we were able to get that down to two to three seconds. 2,880 cores—one brain. Size of about 10 refrigerators—fits in a shoebox. IBM's made it much smaller now. Not quite the size of a shoebox, though. 80 kilowatts of electricity to power it—tuna fish sandwich and a glass of milk, maybe. 20 tons of cooling—hand fan. 4 years and about 2 million books of content—about 30 years of learning.

David Ferrucci Artificial Intelligence (Continued)

Watson had to have a hand. I don't know if you know that. The idea was that the computer would have equivalent interfaces to the game system, and humans had to push that little plastic button down physically, so the computer had to push that little button down physically. It actually had to be made a hand, a bit of a hand that's on top of that buzzer that would push it down.

The computer was beat to the buzz several times during the game. So, while in general it was faster depending on the questions and how quickly it could answer and how confident it was, it was beat to the buzz. In fact, it was many times beat to the buzz.

I want to talk about a question that was a very interesting type of question. Final *Jeopardy!* questions were like this. Final *Jeopardy!* questions were harder for the computer. It was interesting to notice about these questions. Here's the question: "On hearing of the discovery of George Mallory's body, he told reporters, he still thinks he was first."

Does anybody know the answer to the question, by the way? It's Edmund Hillary, right. We thought of it as a missing link. You had to go from George Mallory and all of the things he's connected to Mount Everest and then think, "Who was first to Mount Everest and get Edmund Hillary?"

There was a missing link in the question. It's unlikely that there was any one passage, no matter how good your natural language scores were, that can sit there and say, "This passage is answering this question, I'm 65 percent sure that this is the answer based on this passage and this one and this one and this one."

It really had to combine information from two passages. It had to make two hops. At the hop from here to George Mallory, Mount Everest, and then make the link from Mount Everest to Edmund Hillary. It was a much harder kind of question to solve. In fact, that led to the next thing, the thing I did right before I left IBM was this WatsonPaths stuff which they've been evolving ever since.

Based on that observation, the idea was, in medicine, when you look at medicine, it's much more complicated. You don't know what the passage is and you may have to consider many of them. The presumption is in that it's really in one place. You have to break the question down and make these connections. That's what WatsonPaths does. It defines paths through the literature.

Here, we have a 63-year-old patient is sent to neurologist. The clinical picture of a resting tumor that began two years ago and so forth. This is actually a United States Medical Licensing Exam question. It gives you a bunch of information and says, "What part of his nervous system is most likely affected?" It was the question.

What WatsonPaths would do is break it up into different segments. Then based on different questions, it would try to find links, very general questions like, "What is this associated with? What does it indicate? What does it contain? Was has it affected? What is it caused by?"

It would weight those connections based on its confidence, and then it would land in different places. It would go to Parkinson's disease, but that's not the answer to the question. Then it would go to basal ganglia, it would go here, and it would go there. Then it would stand out from there.

As it was going from the input in this direction, it would go from the possible answers, which is either a multiple choice question or it would generate a set of possible answers, just like the original Watson did. And it would go from the set of questions backwards and see where it met in the middle.

This now allowed you to answer questions by not assuming that the answer was described in one place, but that you could find paths through the literature that if you read the passages here and the passages here, you'd see an explanation for what the answer was.

David Ferrucci Artificial Intelligence (Continued)

Without a single if-then rule, without a single human writing a domain-specific if-then rule, WatsonPaths would be able to find connections from the input to the answers even though it didn't build a model. Talk about augmented intelligence, which is why I liked Sean's talk. It said, "You human, I don't really understand what went on here, but you, the human, it's likely that if you read the passage here and you read the passage here, you'd be convinced that Substantia Nigra is the answer."

You could imagine this going three or four or five steps deep. This is very different for a person who came from sort of that theory-based approach in AI, building all those rules and learning, "Wow, I just can't get there from here." Having a system like this to augment how humans can make these connections through the literature, sort of a very powerful idea, one of the very promising things that I think came out of Watson.

So, some reflections, huge triumph on building this notion of systems architecture, was hugely important to succeeding at Watson. In other words, not sitting there and saying, "I'm doing something specific for *Jeopardy!*," but coming up with a general architecture, you do question analysis, you generate hypotheses, you let the hypotheses compete, you gather evidence, you score the evidence, use machine learning to weigh the evidence. That general architecture was extremely powerful. It will let us build many, many algorithms to plug into that architecture.

A triumph for combining a diversity of methods. My team, I had a bunch of researchers, and they all worked independently in the following sense: "I'm going to work on an algorithm to generate candidates." "I'm going to work on an algorithm for scoring negation." "I'm going to work on an algorithm for this." "I'm going to work on an algorithm for that." Many, many different ideas.

Some of those algorithms would overlap in their intent, but that was OK. I didn't have to fine tune the relationships between those algorithms. I just put them into this architecture, and then I used machine learning to train on former questions and answers to figure out what those weights should be.

And this is very reminiscent of Marvin Minsky's view, one of the founders of artificial intelligence, at MIT. He wrote a book called *The Society of Mind*, where he argued that intelligence was of this nature. There wasn't one system that knows the answer in your head. There are lots of different calculations that input and get combined.

For example, there wasn't any one component of Watson that knew anything about biology or that knew how to answer the question, or many, many things, that produce signals that those signals got combined.

The other observation was we have yet to make machines learn to understand and explain. Watson did not build a logical model for what it read, that you and I can sit down and start reasoning over that and saying, "Why do you think that? What are the logical connections, and can you do deduction over what you just read?" We have yet to do that. We have yet to build a system that can read anything and build a human compatible, logical model for that and be able to reason over it with deduction.

My view of how things are going to evolve is very similar to Sean's in the following way. I don't think you can take humans out of this equation, because human cognition is the source of meaning.

As I said in the beginning of the talk, "What are words about?" Words are just pointers into a shared human experience.

Every time we take a piece of data, we have to interpret it with regard to what our value system is, what our objective function is, what our intent behind or meaning behind that data is.

David Ferrucci Artificial Intelligence (Continued)

Then if we want to be able to communicate with other humans, we're going to have to map that to a theory. We're going to have to open the box up, not just have all the features. You can put a lot of features in here, and we could have the machine produce inductive hypotheses, basically correlations: "This correlation looks interesting. This correlation looks interesting." But how do I now interpret that data, map it to a theory, so that I could reason about it, and engage the human understanding of the world in that process?

You could say, "The human understanding doesn't matter if I can predict the right answer." But do you really believe that? Because if you want humans to engage in that, to augment their abilities to extend the theory, to explain the theory, to communicate with other humans, how are humans going to extend their understanding of the world? How are they going to extend their value system, their interpretation of the world around them if they're just letting machines process the data? Not to mention, you saw all the spurious correlations that happen, what do they really mean?

My view of the future is that as systems, ultimately, machine intelligence, is going to evolve into this notion of, "How do we use computers to accelerate this virtuous cycle? How do we use computers to rapidly process data to produce and prove inductive hypotheses? How do we engage the human machine interface to understand that and be effective with processing that? How do we support the rapid formulation and testing of theories?"

I look at the role of computers very holistically in trying to accelerate that virtuous cycle. I think I'm done.

[applause]

Question: At one point, you showed that this part of the game, you said it's very important to you. The answer was a probability, and then two covers if you will, with probabilities.

I'm trying to follow through your conversation. It seems to me, it's clear that you go through these steps. I would think at each step, you're deducing what's the most probable answer through these steps, but it's surprising or maybe to add some color on, how does it arrive at that probability estimate at the end, and why is it important? Does it have a whole distribution that there's a second best and third best answer?

It just seems like a lot of things you're compiling....

David: It doesn't compute those probabilities until the end of that pipeline. What it's doing is it's taking the question and it's using the whole process for generating possible answers. So it would become competing hypotheses, like, "I think it's Vasco da Gama, it's Magellan, Columbus, or Gary." I got these competing hypotheses.

For each one of those now, I go and I say, "Let's assume this were the right answer. Can I find evidence for it?" Now you could think of the process literally breaking up into four independent streams, each one going off and, not even thinking about the other one yet, and each one going off and trying to find evidence in support of that.

But now the evidence is independently scored, meaning that if I find the passage supporting Columbus, I'm going to have 100 algorithms evaluating that passage, and it's going to give a score. That score is saying, "Here's the likelihood this passage either supports or refutes Columbus as the answer." I'm going to get all these scores.

Each answer has all these passages and each one has a score. Now, first of all, I have to combine it and say, "I have all these scores for this answer from lots of different passages and lots of different algorithms, how do I combine those? How do I weigh them?"

David Ferrucci

Artificial Intelligence (Continued)

That's where you would train over previous questions and answers to figure out, "I'm going to weigh my typing algorithms higher than my syntactic grammar algorithms, higher than this one, higher than that, however I'm going to weigh them.

I come up with all the weights, and I get this final score for Columbus. I get a final score for Magellan; I get a final score for this. These scores are just confidence scores at that point.

Based on a lot of training data, I could see my confidence scores – how well they work, how they map to a likelihood of getting a question right or wrong. Since I can run it on lots of answers, I can now take that confidence score, and I can map it to a probability. Now each answer has a probability, the likelihood of being right.

Question: Are they typically that tight? It was like 32 percent, 26 percent.

David: Yeah, that's what it would say. It was a pretty good mapping to probability. In other words, if the system said, "60 percent chance I got this right," then when it says 60 percent chance, 60 percent of the time it would get it right.

That's how we did it. We had enough data to do that. You can't always do that, but we had enough data to do that. Yeah?

Question: I thought that was a beautiful presentation. It's extraordinary work. If Watson could handle a question about a job swap between senior Russian leaders—Medvedev and Putin—how much would you have to change for it to handle a question about whether Medvedev and Putin will have another job swap?

David: Let me see if I got this right. What I'm saying is you can read about how there was a job swap between...It doesn't really matter who, two leaders. The question was how would Watson predict whether there was going to be another job swap?

The short answer is that Watson is only going to tell you what has been written. That was the distinction I was making. It's not building an internal logical representation for what's going on. It's not building an understanding that it can then reason over and say, "If this, then that. These are the reasons that likely caused this job swap. Therefore, if I see those reasons again I'm going to do that."

You could build a similar system to say, "I'm going to look for linguistic prior indicators that show up before I then see a job swap." That would be a very separate, narrow application.

Question: There seems to be a lot of application with something like this in the consumer market, yet IBM's done nothing to develop that. What's behind that?

David: IBM's investing a lot in the application of Watson into a bunch of areas. The specifics, I don't work there right now. I couldn't tell you what the business strategy is. Sorry, but you're right. I think it does have applications in that space.

Question: Can you talk a little bit about the factors involved in the algorithm for determining how much to bet in the Daily Double?

[laughter]

David: That's a great question. Humans do the natural language processing so well that they're much more fascinated with the betting strategy. We played over 100 practice games against former *Jeopardy!* contestants, and then we would interview them after, and they were always much more interested in the betting. Because that's the part they found hard. [laughs]

David Ferrucci

Artificial Intelligence (Continued)

I'm going to answer your question...but just to give you a sense, there was like one person working on betting, and there were 24 working on the algorithms. It's interesting. It wasn't even a full person. Nonetheless, they did some interesting things.

A paper was published on it. They did some neural network stuff. They looked at human betting patterns. There was a little bit of a game theory going on.

This is how humans are likely going to bet so what will I do? They used neural network techniques to figure out exactly what's the best way to bet to optimize your chance of winning. They learned interesting things, like humans bet way too low on Daily Doubles than they should. They don't bet aggressively enough, things like that.

Anyway, we used those algorithms. We'd come up with a betting...Daily Double betting was very different than the Final *Jeopardy!* betting.

We would use our priors, our likelihood that we would get a question right or wrong, which was different from Daily Doubles than it was for Final *Jeopardy!* Those priors would change if we learned something about how we were doing in that category.

The final bit that I'll share is that, I remember when the guy working on this, Gerry Tesauro at IBM, came in and said, "Dave, we're coming up with really funky numbers like \$126 and \$967. You want us to round it to the nearest 10, right?"

I was like, "Why would I want you to do that? That's just going to make the math easier for the humans. [laughs] Leave it the way it is."

Question: If you were to calibrate the role of perfect digital recall versus the analog system for the human brain, how important was that in the results?

David: How important was the recall?

Question: The ability to have perfect recall compared to fuzzy recall for a human being. "I know the answer. It's on the tip of my tongue, but I can't get it," which doesn't happen to Watson.

David: Well, it depends what you mean by that. It's interesting that you bring up the point of recall. I want to answer the question two ways.

First, I just want to talk about recall. Precision is whether I get a question right or wrong. Recall is, "Can I even remember...Do I have the answer anywhere in my memory bank," if you will. Watson would consider the top end result, a whole bunch of possible results. Let's say the top 500 answers that it was able to extract from various sources. In those top 500, even 1,000 answers, we experimented with different numbers of competing answers to consider.

Only 85 percent were in the recall. In other words, 15 percent of the questions we couldn't find anywhere. We had nothing to consider. When you see getting, whatever the accuracy was, 85 percent right, that's doing really, really well in that game because our recall total was not that high, that's of the questions you buzzed in for.

Anyways, that's one. The other one – what does it really mean to know or not know or to struggle remembering something? Clearly, Watson's not going to go forward if it didn't have an answer. The real struggle was, "how confident am I in that answer?"

That's why I question what does it mean to have recall. Watson's sitting there going, "I don't know. Might be. Might not be. Should I buzz? Should I not buzz?"

David Ferrucci Artificial Intelligence (Continued)

I got a handle on that. That's true. But how sure is it? I agree there's a difference, but what is exactly the difference? Because humans will sit there and go, "Oh, it's on the tip of my tongue." They're not being able to remember it. Watson struggled with, "Is it even right?" I don't know.

In the end, I wanted to clarify what I thought that meant, but I don't know the answer.

Question: If Ken Jennings were to see the list of probabilities that Watson offered would he have improved on Watson's results?

David: That's a great question. Probably yes. That's a great question. I think probably yes. That's an interesting way to approach your question about the recall thing.

Here's a bunch of possibilities. Given those possibilities that you might not have been able to recall, can you now retrieve the additional information you need to distinguish between those two? That's kind of interesting.

If you look at the way Watson worked where initially it would generate tons of possibilities. Then it would say, "Can I find additional information that will allow me to differentiate them?" If you missed that first part...

Your question?

Question: How did you write the rules for the threshold for when Watson answered a different way?

David: That's a great question. That actually changed. That could change during the game. We could move that threshold around.

What happened is if Watson was way behind, it would actually be more aggressive, and it would lower the threshold. It would have been sitting there thinking, "Dave's screwed. I might as well take a bigger chance."

Then if it was way ahead it would say, "Why look stupid?" And it would actually raise the threshold and answer at a high threshold. It would calibrate during the game depending on where it was.

Question: Did it adjust the algorithms?

David: No, it would not adjust the algorithms. It was just doing weighting. There was a different kind of learning going on when it was in a category. It would do this dynamically.

The really smart way to play *Jeopardy!* is to hunt for Daily Doubles right from the beginning. The advantage of not hunting for Daily Doubles and starting at the top of the category is because at a low cost, you can learn what the category is about.

The first question is a better predictor for what the answer type is than the category itself. That's why Watson would go in low and then it would adjust its likelihood. If it was getting those questions right, it was learning the answer type. When it hit a Daily Double, it would adjust its prior for that.

Question: Does Watson ever take a negative expectation bet if you're behind and it's all or nothing?

David: I don't know. At least, let me put it this way, not that I designed. Now, if that snuck in there [laughs].

Question: There are different types of questions in *Jeopardy!* and some of the ones you showed have different paths to get to the answer. The *Green Mansion* is one you could know who Audrey Hepburn's husband was or you could know that Mel Ferrer directed.

David Ferrucci

Artificial Intelligence (Continued)

David: That's right.

Question: Do you think Watson would do better against humans in the question where there's just one route to the answer or where there are multiple routes to the answer?

David: Watson would likely do better where there were multiple routes provided that the evidence for any given wrote was sufficient enough. That's really what it was looking for. It didn't distinguish between routes in that sense. If some route, which is one portion of the question, got it to a passage that had scored very high, it would just say good.

Question: Can you talk about when Watson would buzz in?

David: So the only thing it did there was it had what I would call the confidence-weighted buzzer scheme.

In other words, it would come up with an answer, and it would compute the confidences with that answer. Then, if its confidence was somewhere in the middle area, it was assuming – and we didn't dynamically adjust this during the game, but we could have, because we were playing the top champions – it was just assuming that these guys were really good.

What happens is its confidence was on the low side. Depending on how it was doing in the game, it would be more timid on the buzzer.

The reason was because the guy next to me is really good. If he's really aggressive and is going to jump in and get it, I'm going to let him have it, because otherwise I'm going to lose it and he's going to get it.

If he grabs it, fine. If he doesn't grab it, I'll take a chance. It actually was doing a confidence-weighted buzzer scheme, which in simulations actually tended to work better.

It also had another effect, which is another thing that we were balancing. It was a secondary objective, which was not to look stupid. We could have taken a very different strategy, which was buzz on every single question, but we would have gotten a lot of things wrong.

Where you set that threshold changed the nature of how you played the game. We were walking that line between looking stupid and winning. We actually had a metric, the "looking stupid metric," that we were trying to minimize.

Michael Mauboussin

Closing Remarks

We're going to call it a day.

I'm going to do a very quick wrap-up. These are my CliffsNotes versions of the last 24 hours. Thank you all, by the way, for your hardiness. I hope I do all the speakers justice in this. Here are my 30-second versions of some of the highlights.

First, a recap from last night from Tom Seeley and the work on the bees. To me, one of the key takeaways for us is the fact that these bees are diverse agents. They're independent and different. They are revealing, or surfacing, options. They're going out and trying to find a lot of alternatives. And they're basically voting on those best options in order to find a home.

One of the essential points is that there *is* an objective function. They know what their ideal home looks like. They also have unified goals and values. They all want to find the best home. Obviously, they're genetically related to one another.

The basic principles all make sense. In real organizations it gets a lot more complicated when we introduce things such as incentives and politics.

All of Phil Tetlock's stuff was great but he had one slide that I thought was extraordinary. And it addressed something that can be applied to organizations right away. He pointed out that he now can find results that are 50 to 70 percent better than the unweighted average of a group. There were four contributing factors to that.

One was the people. That was about a 10 to 15 percent boost. That was a function of this fluid intelligence, this ability to deal with novelty and open-mindedness, people were open-minded.

By the way, fluid intelligence as a form of cognitive intelligence peaks in your early 20s and then drifts lower through life. Most of us are losing our fluid intelligence even though we're picking up crystallized intelligence.

The second is the interaction effects: 10 to 20 percent boost from that. That's "How do we effectively work in teams?" That's obviously a very interesting topic. We have a piece out there where we wrote a little bit about this.

Teams can be very helpful, but they can also be distracting in some ways. It's not just being a team. It's being a team in a constructive way.

Third was the training: about a 10 percent boost from training. That is this idea of cognitive de-biasing. Phil mentioned one or two things on that. That's something that many organizations can probably try to do.

Finally, which I thought was fascinating, were these extremizing algorithms. That was a big chunk of this, 15 to 30 percent.

Let me see if I get this right. If you have high diversity in the group, you want to increase the extremes to get better. Is that right?

There's a quote I want to read from Richards Heuer, who wrote a book called *The Psychology of Intelligence Analysis*. Heuer says, "When faced with a major paradigm shift, analysts who know the most about a subject have the most to unlearn."

We all have a natural tendency to default to experts, an expert in this particular sector or this geography or what have you. But when the world changes, that person is actually the most disadvantaged because he or she has the most to unlearn before he can learn the new stuff.

Michael Mauboussin Closing Remarks (Continued)

I think that's a really interesting thought to also bear in mind – our natural inclination is for our go-to people, but they may be the person or the individuals in the worst situation.

As for Caroline Buckee's talk, it is worth dwelling a moment on the fact that a child a minute dies from malaria. This is remarkable. It seems quite remote to us here, of course, but very important.

What I loved about Caroline's talk is that we have these established models, the SIR model for example, which has been around for some time but can now be informed with new sources of data.

What Caroline and her colleagues are doing is extremely exciting. They are improving on current approaches and are generating new insights. She also made a point, which was echoed by Sean Gourley and Dave Ferrucci, on the role of theory.

I saw an article by Chris Anderson a few years ago called, "The End of Theory." I couldn't believe it. When we think about data, can we now release causality and use only correlations?

For some fields, you may be able to do that, but for the things that we're involved with, causality is essential. So you need good theory to understand large data sets. That said, we will be serving curly fries next year at the lunch time.

[laughter]

Sean Gourley talked about freestyle chess. The question I keep coming back to is, "Can we pin down what is this skill of these freestyle chess champions?"

When do I rely on my computer? When do I override my computer? Which skills are required to know when to override my computer, and does that require me to understand the algorithms and programming?

I can say I certainly am not versed in programming and these algorithms. I think most organizations separate fundamentals and quantitative stuff. Is there a way for us to be more effective at bringing those two things together?

And there's this very open question – there's a reasonable reason to be skeptical about it – but it's a good open question. It's, "Can we use technology to essentially play the role of our subconscious?" That helps us accelerate our sense of intuition and understanding.

You just heard Dave's talk, so there's not much for me to add. I thought that last slide was really, again, just a recap of many of these ideas of how human cognition is really a source of meaning. It's essential to be part of this whole equation. The role of data, again, so essential as a foundation for theory. And then the application of logic, focusing on causality.

I will say this too – this may be a bit of a stretch – but even as he described how Watson makes its choices, it actually returned me back to the discussion about the bees, which is, again, independently unearthing these different possibilities, coming up with an independent voting system, and then going with the best alternative. In some ways I felt like the last couple of comments brought us full circle.

I just want to thank you again for taking a day out of your valuable schedules to join us. Hope you had fun. And I'm hopeful we'll see you again in 2015 for another round. Thanks a lot.

[applause]



Thomas Seeley **Cornell University**

Thomas Seeley, biologist and writer, is the Horace White Professor in Biology at Cornell. He is the chair of the Department of Neurobiology and Behavior. His research focuses on collective intelligence in animal (including human) groups, especially honey bee colonies.

He grew up in Ithaca, New York and began keeping bees while a high school student, when he brought home a swarm of bees in a wooden box. Tom earned his AB in chemistry from Dartmouth College and his PhD in biology from Harvard University. Before working his way home to Ithaca/Cornell in 1986, he joined the faculty at Yale in 1980.

In recognition of his scientific work, he has received the Alexander von Humboldt Distinguished US Scientist Award, been awarded a Guggenheim Fellowship, and elected a Fellow of the American Academy of Arts and Sciences. Tom's debut science book for the general public, *Honeybee Democracy*, was published by Princeton University Press (October 2010) in the US and in six markets worldwide.

Note: No transcript available.





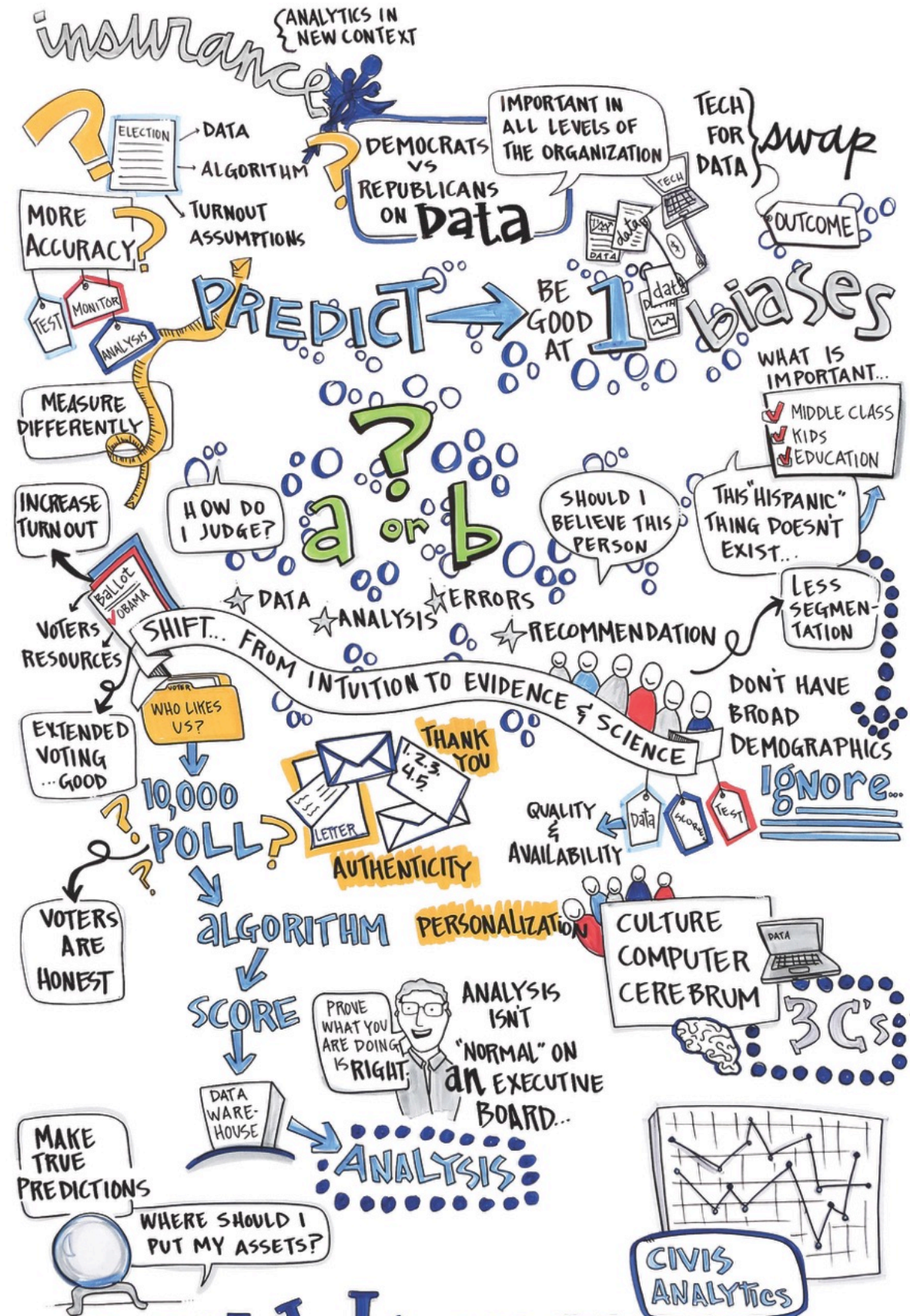
Dan Wagner
Civis Analytics

Dan Wagner is Chief Executive Officer and Founder of Civis Analytics and ambassador for the firm's external partnerships, representing the company's vital, core interests as it engages client business opportunities.

Dan served as the Chief Analytics Officer on the 2012 Obama campaign, overseeing a 54-person team of analysts, engineers and organizers that provided analytics and technologies for voter contact, digital, paid media, fundraising, and communication. His department's work was credited with "reinventing how national campaigns are done" and has been highlighted in Time, MIT Technology Review, The Wall Street Journal, Bloomberg, the Los Angeles Times, and Harper's Magazine.

Dan began his work in politics as the Deputy Voterfile Manager for Barack Obama during the 2008 Iowa caucuses, which defined the relevance of data, and client comprehension of data. Prior to the 2012 cycle, he worked as the National Targeting Director for the Democratic National Committee during the 2010 election cycle. Dan has previously worked with the FTI Consulting Economics Practice. He graduated from the University of Chicago with a degree in Economics and Public Policy.

Note: No transcript available.



Dan Wagner

ink factory
you talk. we draw. it's awesome.



Credit Suisse is committed to a professional and inclusive work environment where all individuals are treated with respect and dignity.
Credit Suisse is an equal opportunity employer.

© 2014 CREDIT SUISSE GROUP AG and/or its affiliates. All rights reserved.