Event Started: 3/6/2012 2:00:00 PM

Please stand by for realtime relay captions. .

>> Good afternoon and thank you for standing by. I at this time all participants are on listen only. After the presentation, we will conduct a question-and-answer session. At that time if you'd like to ask a question, please press star one. I'd like to inform participants that today's call is being recorded. If anyone has any objections, you may disconnect at this time and I would now like the turn the call over to your host conference today. You may begin.

>> Thank you for coming and it's a pleasure to introduce David, who is our fifth distinguished lecture series speaker and we have one more, we have six this year so this is the fifth one. I'm really delighted to have David here, I have heard him before and I really liked his talk and he said it would be similar and actually that's really good. So I'm happy it's similar. Principal investigator of the Watson jeopardy project. That work began in 2007 trying to explore the feasibility of designing a system that can rival human champions at the game of "Jeopardy" and in January, February 2011 it actually did, beating Brad Rutter and Ken Jennings and I think he will tell us some of the way -- how it was done, but not too much. So thank you and join in.

>> Thank you. First I'll apologize. I'm coming down from a cold so I may cough or sound a little nasally as I go through this. It's a tremendous pleasure actually to have an opportunity and really an honor to speak here at the national science foundation. I mean, I've always, you know, science is in my life really. I have two young daughters who both want to be scientists too, so it's exciting to have come, you know, this far with my interest, my passion and career. It feels good to be here. Yes, I want to talk to you a little bit about Watson, the computer that won in jeopardy and what it's all about, why IBM took this challenge on and what it really means, I think, for both science and the business. Why would IBM take something like this on at all? And we had this tradition now with these grand challenges and there's a few things that, you know, the company looks for, one is the drive for scientific advances, to define a problem that we know if we were able to accomplish it, we would really push the scientific field forward but we are also a business and it's important to be able to do something that, you know, has significance to the IBM customers and it's really exciting and helps us to communicate to the broader audience if we can do something that captures the broader imagination of the general public. In many ways the computer was one of those kinds of challenges, got people to think a little bit about what can computers do, what are the limits of computation? How does computer intelligence compare to human intelligence and helped us advance ideas in parallel -- in computer, shows relevance to the customers and it has important advances in science as well. It kind of captured those three things and I think you'll that "Jeopardy" was an idea that also, you know, hit on these three points as well but of course a very different -- it represents a very different kind of problem. If we compared the task of playing chess, the task of really understanding, you know, questions and dealing with human language, we see two very different kinds of problems and it's interesting to reflect on them, because we think about chess, it's well defined, you know, mathematical problems, there's a limited number of moves, a limited number of pieces, a finite board and some sense a very mathematical problem. Everything is unambiguously defined and the computer, albeit has to be powerful and smart and has to navigate this long space but nonetheless, when you think about that and you step back a bit, it's more of a mystery to me that humans can play chess than computers can play chess. You know, I mean, I don't think that we

traditionally thought that way. We thought, wow, you know, humans are really smart, you know, and they are and they have a particular kind of intelligence, but, you know, once you kind of really start to understand the problem, like, wow. Language is really on the other end of the spectrum. Here we have, you know, humans are prolific at this. I mean, we are like language machines. You know, we communicate with each other in language, we write in language, we don't even think about it, how -- we are always doing it and we create languages within sub languages within sub languages, we are always creating languages and this is so natural to the human cognition. It's something like, you know, mathematical problem. It's very ambiguous. It's extremely contextual. What any word means depends on its context and usage that its meaning is actually changing. Speaking about my daughters, I mean, now Laura is 10 and Nina is eight. A couple of years ago interesting scenario where I would spend a lot of time at my house trying to engage the kids particularly in science so very common experience in my household guys, come down here, you're going to see this really, really interesting, doing some experiment in the kitchen sink or trying to program a robot to do something interesting, trying to get them excited about these projects. Come down here, it's really, really interesting and they would run down and look at this and this was a common experience and one time, you know Nina stopped at the top of the stairs, daddy. Yes. I think interesting things are boring. clapping [LAUGHTER]

>> So she still wants to be a scientist so I'm not giving up on her, but, you know, what was interesting about that was she was really making a joke. She was being very serious. This is what the word for her came to mean. She heard it enough times and enough times I bored the -- out of her that she says gee, interesting means boring. This was her experience with the word, this was the meaning that it took on. And that's, you know, what language is like. It's all about, you know, the usage, what does it mean, and for humans, we are connecting those, the meaning comes from associating that word and that phrase with the experience in which that word or phrase occurred, we are connecting it to our cognitive experiences. It's not like it's well designed in some completely circumscribed mathematical definition, not like that at all. It's a very human thing. How do you get computers to deal with this? They can't ground those words, can't ground the meaning of the words in human experience. Contrary to popular belief, computers are not human. They don't have human experience, right? It's just a bunch of zeros and ones. But this wasn't as we started to talk more and more about the "Jeopardy" problem to a broader and broader audience, one of the things we learned is that, you know, it seems like people's understanding of computers, the general public is very opaque, it's quite surprising for someone who has been in computer science by whole life and always been sort of an AI guy and always thinking about this. It was shocking to find out that people really clueless. They use, you know, iPhones and computers all the time but they don't really know what's going on inside. And so we are talking about what's hard and what's easy for a computer because we were getting this almost this bipolar reaction where some people -- you know, we would talk about the whole "Jeopardy" and some people would say, oh, wait a second. What's the big deal? Don't computers know everything? And then on the other end of the spectrum people would come and say, goes to take over like skynet, Watson is going to eat our children. Do people really kind of get it? What's easy, what's hard for a computer? I had a radio show host, a radio show out in San Francisco, California interview me, a live -- this was before the game and so starts off the interview and says, so, computers can play "Jeopardy," how is this going to work? You're going to put all the clues into like a spreadsheet and then with all the answers and then when it shows up on the screen, the computer is going to look up the clue and speak out the answer, right? I'm horrified. You know, luckily it was like radio, not television because I would have been rolling my eyes, you know. So no, I mean the computer doesn't have the -- doesn't -- I mean, it doesn't even know what questions it's going to be asked. The guy goes, well then how does it do it? Exactly. Now you're getting it. Now we are synching it. We started talking about what's easy and hard for a computer. Here's the natural -- times pie Q divided by 647,000? Anybody. Greater than or less than one so 50/50. Human's not so good. The

computer is very good at this. Actually gave a talk once or a -- where a guy actually got it right. I mean, within like a decibel point or two so if you're really good at logs and magnitudes and numbers you can actually approximate this rapidly but still it was scary. Now, that same person had never heard of the game "Jeopardy," so you know what that tells you. How many people saw the game of Watson? Wow. Oh, yeah, right. How many people rooted for Watson? The rest of you can leave. So there's some humans here. What else are computers good at? So select the payment where the -- Davy Jones and the type of product equals the laptop, that's like a pseudo -- query and so computers are really good at navigating tables where you put all this information in a table, and then you can look up Davy Jones in one column and then map that to another table and get invoices and navigate that over to the payment and the computer knows that David Jones is David Jones because, well, you know, -- computers are good at numbers and they compare, the numbers representing the D, the A so -- of course you can confuse the computer quickly. You put Dave Jones in the query and David Jones in the table and it's gone, lost. Doesn't know anything anymore because a D is not a space. So the computer really understands -- now you have to make the computer smarter, maybe you have to get it -- tell what nicknames are and then maybe things that could match. You don't do that, right? You just say let's use unique identifiers, let's use social security numbers, let's use numbers that uniquely identify the individual so we don't have this complicated problem of ambiguity, not knowing what things really mean. So computers are really good at this, you know, billions of rows really quickly, this is what drives so many computer systems, business systems. So if I had this question where was X born and I told the computer, you know what? Here's how to represent this query and you can put in your SQLs and code, the answer is going to be in a table like that where if you find Albert Einstein, you look up his birth place, the name in the first column, the answer in the second column, we called that structured because the semantics, actual meaning is defined by the table. Ahead of time I ran a query that can be answered with that table so what unstructured means, I don't know what the query is going to be. I didn't anticipate it. I didn't write -- I didn't even know what words are going to be used in it and I just got this where was I born and now I read a bunch of stuff and somewhere I read one day from among an interview [inaudible] The answers there but to get it I have to know why the answer there, who are the people, who are the places? The birth place, why is this answer a "Where" question? So, you know, if you start saying I'm going to ask you any kind of question and you have to read a bunch of stuff and get the answer, this isn't looking things up in a database. This gets a lot more camp indicated -- complicated, I have to do more analysis of the text and the question and understand where the answer might be and interpret that. So here's another one, X ran this, and I have -- I looked it up. If I knew what X ran this, what I meant by that, I could look up jack Welch and I can get GE but imagine getting this string. I don't know what that means and I read. If [inaudible] Tenure at GE, I get with some level of confidence that jack Walsh is a painter at GE. But you start to see where things kind of get complicated with natural language. Just for the heck of it, you know, pulled out, because one of the places we are taking Watson is in medicine and, you know, medicine is fraught with this kind of problem. There's lots of information in textbooks and reference material and doctors' notes and it's expressed in all kinds of different ways. So how could I match what's going on in a patient's case to all the literature and what might be going wrong or right with them if I can't really understand how to work out all the language? So food would get stuck when she was swallowing can mean the same thing can cause food to move slow in the efoss Gus -- -- esophagus. When things can change, meaning terminology can affect meaning, the terms are not the same, do they really mean the same thing, causation changes meaning. You know, it can cause food to move slowly is the same thing as a difficulty, chronology, is a sudden onset of chill the same things as chills? Subtle differences in the meaning depending on the context. Fever, temperature, high temperature, well, everyone has got a temperature. Not everybody has fever. Fever after acute symptoms subside can mean something very different of what's happening in that particular patient's case. So magnitude chronology, location, flank pain the same thing as low back pain, same thing as kidney pain, same thing of pain between the upper abdomen and the back. So location can change, different ways to express this. This kind of becomes a nontrivial problem when you think about all the different ways, different things might be expressed. So along comes "Jeopardy." "Jeopardy" is an interesting problem because what it does is sets us up to say if you can do well at this, you're pushing the science along critical dimensions if the space of question answering and better language processing. So for one, you know, broad open domain, they ask about all kinds of things so I can't sort of anticipate all the things that are going to be asked. If you're standing in this direction you should look to check out the wainscoting.

>> Down.

>> Down, that's right. So interior design. What's interesting about this -- okay I saw that question now. I'm going to be able to answer every question about direction, directions. Let me think. I'm going to build a database and put in all the different directions. There's north, south, east-west. Let's do all the degrees around a compass. And then you could group things to north, south, east. It gets complicated quickly but then I got to capture the up down, you know, right, left, got to capture that as well and then maybe in front of me, you know, behind me, you know, there's all of that. So then there's another "Jeopardy" question that says this is the direction of fabric, this was grain. How many people were going to put grain in their directions? How many people? Not one. How do you anticipate the meaning of words in different contexts so interior design. Splits the nucleus and -- anybody? [inaudible]

>> The first person mentioned by name in the man in the iron mask a previous book by the same author. Anybody for 1000? 1000 peanuts. And of the four countries in the world, the one that U.S. does not have relations with, the one that's farthest north, North Korea. One weird thing about information, it's changing. It's changing. I hate that. So -- it makes life difficult. But so it happens a lot, though. So interior design, biology, you know, classic literature, geopolitics, conflict language. You have to be good. You have to be able to get a lot right to compete with the best of this game. And you have to know what you know which was really a fascinating and important challenge scientifically. You just can't buzz in Willie nilly, yeah, I got something. I found something. The search will always find something. In fact search will always find about a million things. You have to produce an accurate probability that your answer is right because if you buzz in and you're wrong, you lose that with the clue and an advantage to your two competitors who now can take the extra time to come up with a better answer and also benefit from the wrong answer that you provided. Do you think I've said that before, once or twice? Confidence to be able to predict that. You know, you really need to be competitive, you have to get that confidence in that answer in three seconds. So this actor Audrey's husband from 19- -- direct her as [inaudible] What's the question asking for? Asking for the director of green mansions really. It's asking for the director of the film. It's an actor but nobody really knows this person to be an actor. The answer is Mel Farrar. We don't really know. I mean -- so I'm on the guy who looks up the "Jeopardy" question and after it's answered and they have moved onto the next one, I'm still going, what did it say? What was the clue again? I'm that guy. That's how good I am at this. So, you know, this is -- so we have to -- so this is a grammatical parse, the main verb is directed and you'll see that, you know, it's directed her, the acronym [inaudible] On these actor, Audrey, and this expression -- modifies, you know, the husband, modifies director, the actor, so forth. Director of Rima, who is a girl. Who is a bird. Now you have to apply some other analysis to assign some -- is Audrey a person? Is director, is that like directs, like directs a movie or guides or leans? Is this a time of year, a duration. Is Rime a person, a bird, a character. Green mansion is a play, a movie, a house? What does it mean in this context? There's lots of questions both in terms of how to parse the sentence and apply semantics to it and if you make any commitment at this stage of the game and you shut down all the other possibilities like you're lost. You have to really start to say, well, I don't really completely understand it. So in other words I can make mistakes in any

part of this so I have to be able to pursue a computation down many different possibilities and see what happens down the line. A natural resolution what does her refer to? You know, so any things like that. One of the things you might be thinking about is wow, parallel processing, going to have to pursue many different threads at once. We looked at about 20 -- early on we looked at about 20,000 "Jeopardy" questions and we tried to say how broad is this domain? Can I really build a database and anticipate all the questions that are going to be asked and every way they are going to be phrased. What kinds of things they ask about like this direction, this book, this actor, blah, blah, blah, and we got this frequency chart and you see this really long phenomenon so even if there are 20,000 questions, even if you try to tackle the head of the tail, you cover less than 10% of the data so things like lady, product, providence, maker, object, disease are basically all equivalent in terms of their weight. They are significant to the problem. 13% of the questions didn't refer to anything at all. It just said like this or it so this is already down to 87%. So this is a very proud problem. There problem -- broad problem. There wasn't going to be any easy way to anticipate the question. We had to use textual resources. The bible, screenplays, books or whatever and try to find potentially supporting, you know, answers with their supporting evidence and then score the evidence. Doesn't really answer the question. Can I get some other evidence to support whether or not this is the right answer kind of thing. So here's a question. So Lincoln blocks. Treasury secretary chase -- [inaudible] Anybody? Resignation is the right answer. Now, did you actually know the historical fact or did you just say plausible inference? Yeah. So me too. Didn't know the history. So what -- how do you use a plausible inference? What kinds of things do you submit and what kinds of things would you submit in the context of a president and what kinds of things you would submit that are even worth asking about, because you use that kind of computation too. I'm going to ask for resignation because that's more significant or relevant. All these things come up to coming up with a plausible inference. We gave this presentation to a 6th grade and they didn't know the history either so don't feel bad. But -- and they used plausible inference and they came up with a different answer. [LAUGHTER]

>> Because in their context, interesting is boring, you submit a request, you don't submit a resignation. What is that? So contacts meaning -- and, you know, with context it allows us to be predictive, which is important to try and answer things because you've got to come up with   hypotheses, you have to take the prediction and take the next step and context helps you to predict. In this case the 6th graders got it wrong. So one of the things Watson does is analyze large volumes of text and we do that -- get that, you know, grammatical structure if you will like that kind of diagramming thing, subject, verb, object, prepositional phrase, and then we do Symantec analysis, so a person, place, thing, you know, inventor, we do that different ways and then we generalize so we generalize by doing, you know, statistics, it can be statistics. So we say if you ask me what is an inventor likely to patent it would be an invention. You may gee, that's pretty obvious. Don't computers already know that? This computer didn't know that. We didn't write down all of these billions of action yomzaxioms that might be true. Watson would answer, you know, a state. Didn't know you wanted a school. Why would it know that? That's common sense. If you have computed on these statistics over large volume of context, this person is probably asking for a school because that's what people talking about mostly. Fluid is a liquid, liquid is a fluid. Which one is true?

>> Liquid.

>> Liquid is a type of fluid. Plasma to gas is a type of fluid. When people use language, they don't care that much. No offense to physicists, but liquid is fluid, fluid is liquid, whatever, same thing and that's what we have learned from analyzing this. What's most likely to synch? When we extract this information and then generalize, we will also extract the reputation of the context. So that's what's likely to sink in the ocean but if you're talking about the game of pool it's an eightball or a cue ball is

most likely to sink and so we are learn this stuff, the computer will learn this stuff on its own. The right answer is cytoplasm, but Watson will look at this, interpret different ways, do many, many searches and generate many hostilities, what we call hypotheses or possible answers [inaudible] Might be 100 of these and then each one becomes the root of a new computation where now Watson goes out and says I want to try to now get proof. I want to see which of these will -- can I gather more believable evidence and one is right over the other and it looks for a different dimension of evidence. Is that answer, you know, like cytoplasm is it what the question is asking for? The question is asking for a liquid. So in fact the green cytoplasm is a fluid surrounding the nucleus. That's an interesting bit of evidence. If I only knew that fluid was a type of liquid, I can start to make that connection and I can take that passage, that piece of evidence and I could use it to increase my beliefs, increase my confidence so I can find the right answer. Is fluid a liquid and one of the things that uses many different ath rich --algorithms -- represented by words and it says is fluid a liquid and word net has the true text in there, it knows the answer is "no" to that, or at least as far as it knows it's not true. Comes back and says can't help you. Then it goes to that knowledge base and automatically builds. Yeah, I have some evidence that fluid can be a liquid and so boost its confidence that cytoplasm could be classified as a liquid. Why would it believe that knowledge is extracted from that naturally occurring text over something like wordnet and the answer is training, learn it. The answer is, it works in the context of the problem I have defined. That's really the answer. We get that way -- we get that through learning. In other words, we will run lots in our lots of questions and we will use these algorithms, they compete and what happens is when you're answering "Jeopardy" questions, this algorithm and its associated resource gains weight and the system learns that this is trust worthy, this is predictive of the answer as much as that is in some cases, you know, maybe more. If you train on physics tests, you could expect this would not earn a lot of weight. This would not -- it would not learn to trust this, it wouldn't be predictive of the answer as much as wordnet would because this was a more reliable resource for that task. So when you have hundreds of algorithms like this, you have to learn the weights, what works in the context of your training data. Your training data in this case was "Jeopardy" questions so I'm giving you an insight into how Watson works. In May 1998 [inaudible] Explores a rival in India. Anybody? Excellent. So I want to talk about different kinds of evidence so here's -- you look at this as this question is a bunch of key words and we all know key words are pretty cool, pretty powerful. In fact, they would light up this passage. In may Gary arrived in India after he celebrated his anniversary in Portugal. Fantastic, got May, got India, celebrated, Portugal. Who is to say Gary is not an explorer? We are all explorers in our own way. The next sentence might read Gary returned home and explored his attic looking for a photo album. So now you have evidence that Gary -- well Gary is the subject of the verb to explore, what better definition do you want for an explorer? Right, the person explores, they are an explorer. We will use that to judge things. But of course it's wait a second. I have an entire country celebrating the anniversary of this explorer. This has got to be one of the three people I remember from high school history, Columbus, Magellan or Degama. This is a form of evidence, we don't want to throw this out. But we want better evidence. On 927th of May, 1498, Vasco da Gamma landed in Kappad beach. I have to look far and wide and consider a lot of content and once I get it, I can't just throw it out. I have to say well, is there any set of algorithms or ideas I have that are going to allow me to understand relative to this question, allow me to match it if you will. Allows me to say well something happens -- [inaudible] That could refer to 1498. If I -- I didn't write down all of these axioms because I didn't predict all these questions and how the question may relate to what's being said in the context. I'm not going to get there anytime soon. So but what I can do is by analyzing the language, I can find out there's -- these words are used in similar context so some reason to believe they might mean the same thing. I have to be willing to deal with all kinds of uncertainty. I'm getting some signal. Here's where I actually use the database and I think Kappad beach is in India. This is matching so maybe -- can I get any evidence of Vasco da Gamma is an explorer? There's a lot. I'm I 100% Vasco da Gama is right? No, not sure. Do I have more confidence? Yeah, I do. So I'm liking Vasco da gama better than Gary.

>> When "60 Minutes" premiered this man was U.S. president. What's interesting about this question is who writes this? So "Jeopardy" writers write these questions but there's something more fundamental here which is that when you write something down, when you write an essay, an article, reference book, whatever, you don't anticipate all the questions that someone else is going to ask and how they are going to phrase them. You want -- a user comes to a problem in a way that's different than the way someone else wrote down the knowledge that they could potentially use. So this person coming at it in an interesting way, but not the way anybody wrote it, so it's not just a simple finding, you know, here's a possible answer, does this passage say the right thing? Do I understand the language? I really have to start breaking this down because when I try to answer this question, I can't get a good answer. I have to know I'm not getting a good answer. I'm not getting good confidence. Then he have to decompose the question saying maybe I can get this -- pose a new question, in 1968 this man was U.S. president. Break the question down into subparts, it will solve them independently and put the pieces back together and there are other strategies to putting the pieces back together. Does this mean it gets all the questions right? No. But this is one of the strategies it uses. Shirts, TV remote controls, telephones, anybody?

>> Button.

>> Button, very good. I find that answer uniquely human because we did, you know, spreading accusation and we found there are oils, plastics, things under my couch, things made by man, lots of things that are common, right? Didn't matter how much I complained to the "Jeopardy" producers about this, they insist the answer was really just buttons.

>> Don't forget, human questions written for humans. We have a computer in the mix here. So we call out the missing link, you know, the thing that these three concepts have in common but sometimes it's not obvious like that's a common bond question in "Jeopardy" where you splitly know they are -- explicitly know. He told reporters he still thinks he was first.

>> Hillary.

>> Hillary is the right answer. To get that you had to make that leap from George Mallory to Mount Everest and then who was first at Mount Everest? So we think of Mount Everest as a missing link. You have to start putting together kind of a chain of inferences, if you will, or chain of association and of course it's complicated because lots of other things are related to George Mallory. Why am I leaping off of Mount Everest so how the different things that are related to George Mallory, how they relate to the rest of the question, what kinds of things can you be first at. There's different ways to slice and dice it. A lot of "Jeopardy" questions were like this. When you're working medicine, it's very much like this. You have to find a chain of connection, if you will, to get to the answer and you have to know that the missing link is not the answer. For a long time, you know, the project was about four years but there was sort of a period where we were struggling with Watson answered with a missing link so you have to kind of be able to say wait a second, that's not good enough. I have to get smarter about what the answer is and what needs to get to the answers and actually this is kind of interesting because, you know, this work has been done by Dr. Swanson who used text analysis to find connective paths to generate hypotheses, so looking at texts associated with [inaudible] Found that connecting things being fish oils use that to generate the hypothesis, maybe fish oil can help with Raynoud's disease God proven but it was -- got proven but it was used to find the connections. It gives you a sense of the classes of things that the Watson system was going to have to deal with in terms of processing language and

understanding questions and repeating evidence. How good did it have to be to win? And this is a graph that captured that. We used it as a way to measure Watson's performance over time. And the way to think about this is the dots in the center of the graph, we call this the winner's graph. These are actual games, "Jeopardy" games and what I'm doing is plotting the performance of the winner of those games. So, you know, take the center of the green cloud, but then you see -- and draw that down to the X axis and the X axis is the percent answer to the number of questions, the percent of questions in a game that the winning player was confident enough and fast enough to buzz in for and get to answer that question. So about, you know, again, you're nearly 50% so that means the winning player is confident and fast and winning questions away from the two competitors. They are getting way more than their share. If it's random, each one would get about 1/3 and then only 33%. And then they are doing the between 85 and 95% precision. That's really good. They are getting between 85 and 95% of those questions that they were fast enough to buzz in for and confident enough to get right. On average was confident enough and fast enough to buzz in for about 62%, acquire 62% of the entire game on average, so from 70, 81%, 81%, that means you're just watching Ken Jennings answer questions. You don't even know who else is playing. And so he's really -- you know, and how many of you do that? You just have to assume you know all things. Unfortunately, he was right about that. He did about 92%, you know of precision on average but you'll see there are games, you know, he won with only answering 40%. There are games he won with, you know, of the questions he answered, and he got, you know, just 81% right and, you know, only answered 51%. So it depends on the competition. It depends on what was in the game, you know, some games he was better at, some games not as good at and of course there are daily doubles. A big luck element in "Jeopardy." Who were you playing against? Did you get the daily double, not get the daily doubles? Who in their right mind would bet their career on getting a computer to play this game given the luck element? So, you know, this one -- we had a kind of state-of-the-art techniques in it, had been a small project, though, you know, typical kind of research project, three to five people and we took that system and we turned it, you know, on "Jeopardy" and got this confidence curve and what this says is that the way to think about this, a little bit tricky is you're -- the way we have the confidence, imagine you had all the questions and the system determines a confidence, you know, those how short ones was getting it right and then delivered his answer and what it did is ranked all answers based on the confidence so from the highest confidence to the lowest confidence. So if you look right here, 5% of the questions that had the highest confidence in his answer, it got 47% right. As it had less and less confidence, in other words, more and more questions but now my confidence is going down, it was doing worse and about 15%, less than 15%. That left of course a really big gap between that and the winners cloud and that's when, you know, most of the team turned around and left. Some people were really frightened but for the most part we had a fantastic team, very talented people and dedicated people and this was even more. This is a challenge we took on in 2007 and it's a move that state-of-the-art capability from performing at that level to slicing across that winner's cloud. So what was some principles? Some principles were we weren't going to get this done by building some big database of questions and answers, too slow, too narrow. We really had to do the analysis of content. There wasn't going to be one algorithm that solved the problem. There's no silver bullet. One guy came to me in the beginning of the project and said for us -- we need maximum equations on -- we will have this equation that can answer any question. I said if that's true, we are screwed. This really needs to be built on, you know, the decades of natural language process and the information retrieval, we need to figure out how to put these capabilities together in an architectural that would allow us to advance them rapidly and solve this problem and that's ultimately what we did. Many different methods, many different algorithms combining together them continually giving them the data. We are going to have to look at different interpretations of the question, different possible answers, different paths toward answers, different collections of evidence. How do we do all of this and fast enough to compete in "Jeopardy" we are going to have to do a lot of it in parallel. We ended up with DQA. This is a high level, the software architecture of the system so you have the question, the topic of the category, analyze it, you may have to break the question up into different parts and pursue each power in parallel. Like that, you know, a premiere thing, other questions like that. But, you know, the first step you will do is you will analyze the question and the topic and you'll generate many, many searches, come back with hundreds of possible answers or candid answers, could be 100, could be 75. This was a dial we would turn to manage the system is how many candidates you would come up with because every time you came up with another candidate, you were supporting a whole nother thread to the system. Let's say you had 100. And then for each candidate, now let's say you had 10 pieces of evidence to see if you could support that candidate's right answer so now you have a thousand evidence answer pairs, and then for each evidence answer pair we have 100 algorithms, you know, each algorithm scoring that evidence answer pair so you could potentially end up with millions of scores, combinable. So -- and then if I have each evidence answer pair, you know, Christopher Columbus, some passage, the evidence and then algorithm that scores it, so now I have this other score, how do I weigh one algorithm's contribution to that answer over another algorithm's? There's hundreds, how do I get those weighed? We would run the entire system on a whole bunch of questions and the machine learning would look for those weights that -- how to balance all those various algorithms, whether the coefficient so -- that collection of weights is called a model. We come up with many models. Different models is because different classes of questions like if you're solving an anogram you are going to balance the algorithms very differently than if you were solving a history question or something like that. So we came up with different models and different fundamental classes of questions. With that architecture, with that approach, we now had to take those -- we had to now develop new algorithms and advance our base capabilities and continue to IT rate on that to make Watson smart enough to play "Jeopardy." It wasn't like we came out with that architecture, we were done, the hard work began, the research of advancing those algorithms. Just to give you a taste of where we were as we evolved. This was the early version of Watson and some of the questions answered it did. So actually "Jeopardy" questions, the American dream, decades before Lincoln, Daniel Webster spoke of government made for, made by and answerable to them. Anyone?No one. This was an actual Watson answer early on in the development but --

>> Exactly.

>> Exclamation point was warranted for the end of this in 1918. Very good, Watson. Watson, there you go. Exclamation point that ends a sentence, what year? I really don't get the problem. The answer was World War I. Also in 1994, 25 years after this event, one participate said "For one crowning moment, we were creatures of the cosmic ocean." Watson, the big gang. Not a lot of people around. To comment on the big gang: You have to understand now, you know, as, you know, scientists and the team working on this, at that stage of the game was the fact that we answered with an event, I was like wow, we answered with an event. Good thing. The queen's English, give a Britt a tinkle when you get into town and you've done this. Again I thought that was right. And the last one, not -- all of these you can kind of see each one of these kind of represents this and you say yeah, that's hard. This one, I'm not sure what this one explains but show it to you anyway. This was the father of back tierology -- bacteriology. Watson, how tasty was my little Frenchman. [LAUGHTER]

>> You can imagine on that one we were just like....

>> So what what's really a big part of what we did was develop a methodology on how to take a lot of different researchers working on many different algorithms and advance them rapidly, combined so it was really a methodology for innovating rapidly, you know, very goal oriented. We had this end to end system metric, we needed to get better at our precision and confidence, we needed to move that line

forward towards the winner's cloud so there was very much a focus on coming up with algorithms but, you know, there wasn't a single hero or a single technology that was going to make all of this all of a sudden change the world. In fact, algorithms would have 2% impact, half a percent impact. In the early stages you might come up with an idea, you put it into the system and you had 5% alone by yourself. You're like wow, incredible. You know, but this six months later that 5% got eaten away and all of a sudden your algorithm was having a half percent. Other people were coming up with other algorithms and were overlapping. The alternative was to come up with some master design that anticipated all possible variables and interactions between all the falcons. That wasn't going to happen. We wouldn't have written a lot of code for five years. This was a methodology that says yeah, you go in and develop the algorithm where you integrate, we are going to train, try to understand, we are going to try to make the right bets but in the end, the performance is going to tell us, you know, what's working and what isn't working so extreme collaboration. We all got into one room, really tried to work together as a team and to communicate, very high communication, who is doing what, what are you working on to minimize really, you know, obvious overlap, you would sit there and say what are you doing? That's a really good idea. I'm going to take this and interact this direction versus that direction so constant collaboration. We did very rigorous blinding experiment. We would do biweekly integration. People would work on their algorithm on the last version of the system and every two weeks if your algorithm was doing no harm and only if it was doing no harm, you had the privilege of taking your updates and putting them into the next, the integration one. We would integrate and then regress and say okay now let's is see how we are doing. The whole everything all put together? Each of these biweekly months would generate over 10 gigabytes of analysis and data. This was the key, what was going on? How do you take the next iteration of research and you didn't understand what was happening and it's an incredibly complex system? Over the four years we did over 8000 documents and experiments. So we would have a website where all the documents got logged and you could dig into an experiment and see what -- you know, the aggregate were, how is it performing, accuracy, precision, average and things like that and then look at the actual questions in that data set, type of thing the answer was supposed to be was -- the confidence was and then you could drill into a particular question and see what were all its answers and for each one of those answers, it seemed common with a number between zero and one, there were 600 of these columns so you would look at those columns and look at these numbers between zero and one to try to understand how these algorithms were influencing these answers and what would happen. Come on, be honest, what would happen? A human looking at 600 columns between -- what would happen? Nothing, nothing. You would just go blind but nonetheless there were people who would stand up and say, oh, I see what's going on. In fact, no, you don't. And that was the key thing, no, you don't. We need tools to understand what's going on. We need tools that help us plot, you know, how any individual one of those algorithms or features were performing relative to their prediction of getting an answer right. So if you have some algorithm and its confidence score is going from zero to 120, the confidence score going up was it likely the answer would also go up? I hope so. Even though it would go very small, at least be going in the right direction. Actually, only put the opposite direction, that's okay. Going to change the signs. The problem is, well, do you even know what direction it's going? You know, is it bimodal? That's bad, going up and going down. Not so bad. It's bad if you don't know it. If you know it, could you slice it and create two teachers. But you don't know that until you have tools that help you understand how these features are behaving. How about this? Now we can say how do they contribute relative to one another? It's a big chilly today is the category -- here's the top answers over here so arrest Tina and bailiff -- Argentina and Bolivia at the top. Now we have all the hundreds of algorithms here and what you quickly see is a graph that says which algorithms are the right answers and comparing them to the answers. So which algorithms are contributing to the right answer? Look where the green bars are. The yellow bars are saying these are algorithms contributing to the wrong answer. Quickly zoom in. This is with all the model weights and everything so you know exactly what these contributions are. You say oh, that's what I have to pay attention to, see what's going right, what's going wrong. But there's still a lot of them. My brain hurts. So, you know, while -- you know what? We can group them to make them more cognitively more acceptable, a whole bunch of algorithms, geospatial information, popularity, reliability, how reliable is the source. Now all those hundreds of features now, some of them are represented under these groupings which at least have to make some cognitive sense. It actually led to the notion of evidence proceed vial. In the end if you want to apply this technology outside of "Jeopardy," you don't want to just give an answer to a number. Why do you believe that answer? You know, ultimately human I have to make a judgment so you have done a lot of work, you know, you Watson guy. Why are you coming up with this number. I want to see how you broke the evidence down from a location perspective, I like Argentina, popularity perspective, I like Bolivia. Why? Turns out that there's a lot of big dispute between Bolivia and Chile over -- land dispute over a border so a lot of noise out there but wasn't giving you the answer. Reliability, the sources that were giving evidence for Argentina were more reliable than the sources getting evidence for Bolivia. Classification, Watson says they are both countries. Good thing. You could now imagine that you could drill in each one of these bars and actually get the evidence itself. Oh, I see. That's why. Right? So this kind of leads you into where do I want to look and I actually look at the evidence itself. You'll find double columns in -- there's a Bethel college and a seminary in both cities. St. Paul or south bend, Indiana. It's a Minnesota city, right? It turns out that the Bethel college and seminary in both St. Paul Minnesota and south bend, Indiana but you think we don't know the right answer as well. Watson sort of did. It knew that St. Paul was in Minnesota. That was a good thing. Negative evidence for south bend because south bend is in Indiana but a lot more evidence in every other category for south bend, Indiana. Watson should have never believed any of this because in a case like this, the location information should completely overwhelm all the other data. So the evidence profile would teach us something is wrong with how we are weighing the data. So we learn how to make it smarter, how to look at questions differently and understand how to weigh information differently. Another thing, of course is that it's holey. Which city is holier? St. Paul is holier and someone wrote a pun detector, an algorithm that would look for pun relationships between clues and answers. We put the pun detector in, trained it and it earned weight and it referred to St. Paul over south bend. Of course that wasn't the best way to get this but I make a point that you can come in with a new idea, write that idea, integrate it and train in it. It says oh it's predictive of the right answer. Of course what it did, the pun relation detector boosted holy cross, Minnesota well above south bend because after all holy cross is holier than south bend. With the architecture, the algorithms, allowed us to integrate and advance the algorithms we move from that base line all the way to that dark blue line, actually a little bit above that was the system we played with against "Jeopardy," slicing right across the winner's cloud, this means that Watson can identify -- Watson knows how to rank. If I order the questions based on my confidence and I cover 70% of them and this ultimately -- I don't get like 87 or 88% right over here, so -- and anything -- anything with a higher confidence I'm doing even better. We actually mapped the confidence to a probability successfully so in other words, if I'm doing right here, if I'm doing 90% precision, my confidence was 90% so if I'm 90% sure I'm getting it right, I would get 90% of those right. This was good enough to compete against champions. Does it mean you would win every game? No, it doesn't because depending on the question, depending on, you know, the categories, depending on the competition, depending on the speed, how quickly you can answer those questions, depending on the daily doubles, which is a lot of luck involved in, all these things can change the likelihood if you would win a game or not. Knowing that as a scientist it wasn't about that one game. As a professional with a career, it was all about that one game. A scientist, it wasn't -- I kept getting reminded of that by the senior VP. But as a scientist it wasn't about that one game. We played 55 games actually formal games against tournament champion players, people in the same league as Rutter and Jennings and we won about 71% of those games. So that's -- and that was significantly significant but that means we went into that final game with roughly 30%

chance of losing. Can you imagine what that felt like? It took about two hours to answer a single question on a 2.6 gigahertz CPU with about 16 gig of ram and took about two hours. Now, "Jeopardy" insisted that would make a boring "Jeopardy" game so -- but again because the architecture was embarrassingly parallel, had these constant splits where you can pursue these answers and this evidence independently, we were able to paralyze the system and scale out of the 2008 -- final system was on 2880 cores, 15 pair bites -- para bites of ramp didn't go to disk and that would average three seconds per question, depending on the question, so it could take eight seconds, could take one second. But about three seconds. This was fast enough to, you know, to play and to compete against champions. If the lab we actually had two systems because one was optimized for development and one was optimized for late tent -- what would happen is every time you answered a question, you know, depending on what point the system was authorized, different points, at one point was taking two hours, we got it taking about an hour per question so every CPU was running the entire system answering a single question taking an hour but if we waited on 1000CPUs, we could answer 1000 questions in an hour. This system was optimized for Lehtinen -- this was for development where we could change things quickly. This year we ended up winning the final -- that wasn't the final question. That was the final score. What happened? The point that I want to make here is was "Jeopardy" a good challenge to take on from a scientific perspective and besides the AI thing, you know, it's interesting to see, you know, this machine that would take 20 -- 80 kilowatts of electricity and 20 tons of air-conditioning computing against a human with a brain that fits in a shoebox with a tuna fish sandwich and a glass of water and it walks. So that was really cool, but from a technology perspective when you talk about natural parsing, disam big waking -- dis ambiguous, extracting knowledge, whether or not one text entails another. These are measurements that are regularly done in the NLP community or all our algorithms got substantially better. In fact we have leading scores in all of these. What's interesting is we didn't know researchers sat there and said I'm going to get better. No researcher said I'm going to get better at this or that. They all worked to make the system win "Jeopardy" and it was a good scientific hypothesis, a good challenge because in fact it did drive all these component technology that have leading scores in the industry which was very nice to see. You know, it's all about applications going forward when we -- what do we do with this? What's beyond "Jeopardy" and I think there are lots of areas where this technology can help but generally areas, real valuable knowledge, natural language resources that we want to get at, understand better and allow that knowledge to help informed decision making, help humans make better decisions. Health care, kind of interesting to see how the way Watson works maps onto doing diagnosis or treat assessment. So what was in these dimensions of evidence can become the kinds of things that you look for like symptoms, personal history, medications findings so we talked about an evidence profile. Here's an evidence profile, what it might look like in medicine and when we combine the scores, final confidence in a particular diagnosis, in this case UTI or urinary tract infection. From a simple perspective we like this diagnosis, you know more than from a -- imagine I can drill in to each of those and look at the information that's taking me there, look at the information that's supporting you. So getting from, you know stuff taken from a patient interview and lab test to the family and personal history indications, test findings, huge volumes of text journals, references and data basis you can imagine this system becoming this continuous process that's analyzing all this data and updating and computing the data profile creating different possibilities. It could be diagnosis, could be treatment options and treatment options, there isn't necessarily a correct answer. We have basically different possibility of treatments, the best matched input data based on the most current background information for different reasons. And you want to kind of know what that is and you want to be able to organize the information according to these evidence profiles so very interesting. We imagine an interface looking like this, we see the actual evidence profile but now drilling in and seeing the actual, from textbook, from website, medical journal, you're seeing the actual evidence itself that contributed to that score, you know for that data so when we think about it, I'll end on this slide but when we think about where we need to go moving from "Jeopardy" and beyond, on the top we see a couple of categories, you know, understanding the input of the question, interacting, explaining the answers and learning, the continuous learning process and then you see kind of the where we were with "Jeopardy" to along here, where we need to go. And with "Jeopardy" we have specific questions like the type of murmur associated with this condition is -- when you get a wealth of questions, there's a lot of data in that question because the person who formulated the question thought hard about, you know, what the answer is and how I formulate the question. We need to go from a single question to just a bunch of data. Even if it's not in medicine, here's everything I know. Tell me what my problem is. The question is always what's my problem and how do I fix it? If you think about it in medicine, what's the problem and how do I fix it? That's always the two main questions but I have lots of data. How do I partition that data and figure out what questions do I have, which data is worth pursuing or not. If you go from a specific question to rich problem scenarios, question and answer out, well, you know, guess what? That's artificial, right and "Jeopardy," how many times Watson wanted to say to Alex, Alex, did you really mean -- or APB, when you expand that, what did you intend for that to be? But you couldn't do that. In the real world you can. You can say I have this evidence but I'm not clear on it and so being able to go from a question and interact with dialogue to help refine the question and to help build better confidence in answers or in what it's meaning makes more sense. We have to go from question or answer out to to something that's more interactive. From the answer panel that you saw on "Jeopardy" to really evidence profiles that we are showing you. It's not just the answer and the confidence. It's how am I weighing the different class of evidence and can I show you the evidence itself and what you decide? It's a teaching process. Really want the teaching process to ultimately not be just that that you train every so often but constantly training and constantly learning from the interaction of the system so you really want it to be collaborative learning process that happens over time. The system providing value and you are working with the system so you're both learning as time goes on. This is kind of where we are going with Watson from a technical perspective and of course the application, one of the application areas is health care. Thank you.

>> Do we have time for questions?

>> Probably a few.

>> I didn't watch the "Jeopardy" show but I did see the video that was linked to the announcement in your seminar here and giving an example you gave of St. Paul and south bend, the question answer about, wrong answer of Chicago -- Toronto, it made a big point in the video that Watson would [inaudible question]

>> How did Watson --

>> Watson didn't want to answer. That was a final "Jeopardy" question and in final "Jeopardy" you have to show your answer. On a normal question, Watson would have never answered because the confidence was way below thresholds so in final jeopardy you have no choice, though and one of the things we did to indicate that was on the answer panel Watson would put like 20 question marks, you know, reminding them I'm answering and what happened was the first airing of the show, they aired it again in September, the first airing of the show in February they did not put the answer panel which caused me incredible frustration for months after that because it showed he had 13% confidence and 11% in Chicago and would have never answered that if it wasn't for final "Jeopardy" where we have to show the final answer. I've been answering that question like, you know, 4000 different times. It didn't want to answer that question. The interesting thing, though, about that was that everyone thought that was such -- it was stupid to even put Toronto in front of Chicago for the same reason that we see there,

was because the category is U.S. cities but the questions didn't say this has to be a U.S. city. It said at the largest airport. So if you look at a lot of the training data in jeopardy just because the categories you would see does not mean the answers you would see. You learn that over and over in "Jeopardy." Categories is overseas and the answer is shuffle board. So Watson knew that because it's Toronto doesn't necessarily mean it's the right answer but nonetheless, very low confidence. Now, if you rephrase the question as this U.S. city, it swaps Chicago and Toronto. It still didn't have enough confidence like 30% when you did that, it has 30% confidence in Chicago.

>> [inaudible question]

>> It's one of the areas I'm very interested in, is the whole process we went through and codeifyionying. It's more general than this so I'm very interested in that.

>> I didn't quantify this so I can't establish it are from a statistical standpoint but in the third game it seemed like the third and final game, it seemed like Watson has some hesitation where when it came to popular culture type questions and Ken seemed to see that Watson had a deficiency in that area and he would automatically try to answer the popular culture question. It seemed like he was gaining an advantage. It may be my imagination but just seeing if there's something there.

>> It had no particular weakness in popular culture. It was part of characterizing the weakness from a category perspective. We -- early on the categories didn't tell us a heck of a lot. What told us whether or not Watson might be able to answer the question was the previous questions in that category. So it was hard. I mean, you know, we had one group that came in -- a whole category having to do with knitting or something, something like that. And Watson just, you know, was like a train wreck. Just did really bad. But not for any reason, right? It just didn't understand those questions and, you know what else? It was so variable on whether or not -- you know, why? Because it depends on -- it depends on whether or not it can understand the question well enough and then what it's reading and how it can relate to that question so it depends on what is read too. So you can't really know that. You think you can know it but you can't so -- because everybody would guess and remember we had a group in, and did terrible like a food category and on knitting category. And people said Watson is a chauvinist. Why? Because clearly he doesn't do those categories about women well enough. Really?

>> Input on something like timing and jitters you have accurate response time simulating Humes?

>> We didn't have to simulate a human. We had to use the same interface that humans did. So in other words Watson had to push --

>> I asked that because many of those times you could see the other two had the answer, they chimed in in a slight fraction of a second too late, obviously. I'm wondering did that give an unfair advantage to Watson?

>> So again, the way this was calibrated was there's an interface in the "Jeopardy" system which is you have to push a little plastic button down on a screen. Watson had to push that same plastic button. At that point on it was human versus machine. I can tell you -- what happens is humans listen to the clue, Watson wasn't listening so Watson -- you know, and the way, you know, good expert humans do this is they listen to the Kay dense of the question and -- cadence of the question and time it. They do this remarkably well so where they can come in in less than 10 milliseconds, it took Watson's fastest buzz because Watson had a confidence weighted buzzer scheme, depending on how confidence it was, it it would buzz faster or slower up to a limit. The fastest it can buzz is about 10 milliseconds. Technically speaking by listening and timing, a human can beat the computer to the buzz and did. The human wasn't

as consistently fast as Watson can be if Watson was able to come up with the confidence in the answer. So the human could beat and often did Watson but not consistently, again, you know when we were both kind of ready and ready to go. But humans sometimes are faster at the -- at just computing the answer. In the air game and again we played many games and you saw all kinds of weird things happening. It was interesting because often you would have a human -- by the way, the classic thing to complain about when you're a "Jeopardy" player is that someone else is cheating on the buzzer. This is what you do. This is part of the culture. So what would happen was was it was nice when you had two humans because we had one person there saying it's absolutely impossible to beat Watson to the buzzer. Meanwhile the other human was beating Watson so they couldn't complain. It's frustrating when you know the answer and miss that businesser -- buzzer. When we knew the answer and missed that thing, I was dying. You hate it. In the actual aired game, Watson lost an entire category to the humans because there were short clues like directors who act or something. Watson had the right answer in over 90% confidence and lost every single one. They are like -- sorry.

>> Extending into the medical field and I wonder whether the next show would be, you know, you play against Dr. House, but apart from that, are you really aiming it at the difficult, difficult diagnosis or are you really looking at providing diagnostic tools for every physician?

>> I mean, I don't know that there's a single answer to that. I can tell you what we are really focusing on scientifically but also a commercial effort and I think the commercial effort is being more driven by talking to customers and figuring what their pain points are, how do we adapt this technology to really solve the problem that a hospital, institution, insurance company whatever is facing, where are the costs going, help them manage the health care process better. From a scientific perspective, it's really can I understand the input relative to the context, can I draw, if you will, an inference graph from the input to the most likely diagnosis to the likely treatments and can I annotate that graph with the right evidence from the huge volumes of content that are out there and that's the way we ultimately want to evaluate Watson and it shouldn't -- you know, was it a hard question? Was it an easy question? Same thing with "Jeopardy," very hard to determine how you measure whether it was hard or difficult. You know, we went like as examples, in a "Jeopardy" case, we were analyzing our performance on historical "Jeopardy" games and we would randomly sample from the games, random set and we would test our performance. And we were getting a certain, you know, performance, pretend it was here. And I -- how the game is constantly changing, it's not formulaic evolving and one of the researchers said that's a good point, I have to order the data chronologically and they did that and it was a shock because what happened was the performance went like this and, boom, like that. In 2003 the performance changed by almost 10%. What happened? Were the questions harder? The humans weren't finding them harder? Why was all of a sudden the computer failing on questions post 2003? Something changed. So the best way of characterizeizing what changed -- characterizing what changed was the language used in the question was further removed from the language used in the content. "Jeopardy" got more entertaining and more comical and more this in the way they were formulating the questions. Were they spins sickly -- intrinsically harder? I think when humans think hard, they think lots -- take lots of inference steps. I had to go from here, then to there and then to there and then to there. I had to do five steps, six steps. Another way of thinking about it is not common knowledge, in other words, it's fewer.

>> Yeah.

>> Good obscurity, right? Because they can get all the information that you don't remember. So different ways to measure that.

>> So you talked a lot about algorithms and how you would go through the process of screening. What about knowledge representation and were there any sort of strategic shifts along the way, you know, as you got to higher performance --

>> So, you know, from an organization perspective, we didn't expect there was going to be one -- from jeopardy but rather we had to asim lit different -- assimilate different oncology and use different things for classifying things and did that. We used wordnet, other things. It was loosely formed and formally structured oncologies of every kind as sources of information that would contribute to interpretation of a phrase or how to relate one word to another word so very liberal in our use of that. And in some parts of the system there were kind of formal action mat tick -- axiomatic kind of things. It was more driven by I have an algorithm where you can capture phenomenon that we seem to be missing on and that algorithm is going to depend on more of a formal representation. But if you looked at the bold contributions from the system, it was relatively small coming from more the structure domes representation part but nonetheless it was there and there were clearly opportunities continuing to expand it. We expect there will be more of that actually in medicine.

>> For those of you who have questions, come on up. [event concluded [