Event ID: 2862460
Event Started: 3/3/2016 12:19:35 PM ET

Please stand by for real-time captions.

Thank you very much. Good morning. This is a series about -- the data science seminar series we have at NSF. This is a great pleasure and with enthusiasm I'd like to introduce your speaker today Dr. RV Guha. Will talk about modeling of complex systems. I would take a few minutes to introduce him. His undergraduate at technology -- and PhD at Stanford. He's working on a number of different projects and after which he had the one of his many startup companies. After that he moved [Inaudible - static] then he joined Netscape. With all of the A-list companies he's been involved with he created [Indiscernible] they also contributed to a Netscape program and also worked on RSF. Real-time simple syndication. Then he worked on one of his other side companies -- [Indiscernible] after she went to IBM. And then for the last 10 years or so he's been [Indiscernible] research at Google until last month and now he's a free bird and giving talks.. Giving a talk. [Applause]

Thank you for having me here. My talk today is about a bunch of different things. Feel free to stop me and ask questions. The first part I give this talk about [Indiscernible]. What the hell is data science? I can of the concrete definition to quote empirical modeling and then one of the deep dives into a few research areas that I'm personally involved with. One of structure data and the other is what I call teachable learning systems. And finally a close by making the case for a data comments.

At the core of engineering is the task of modeling. Creating models is essential for building, predicting and controlling systems. A model is a set of variables in a situation and its constrained which tells you how the different configurations can occur. It's essential for capturing behavior of the system.

There was engineering before models. But we needed to go to the -- you need a model. When you to Kitty Hawk to the you needed a model.

An interesting part -- models are why [Indiscernible] interesting -- this kind of modeling has the reason why engineering has taken off so much in the past few years. The kind of modeling you do is something I call analytic models at its core you have a relatively small set of fundamental equations having to do with concealed mechanics materials heat transfer and fluids. And so on which capture the properties of the system. The system is modeled with these equations.

When the -- about 315 years ago in the context of calculating Bridge dimensions it became much more complicated to do by hand. We start building more and more complex systems and have computers and we used the fundamentally the same approach. When you build a dream letter its wings can flex by 26 feet. Needed model how 26 feet is okay and 27 is not. At the end of the day the core models and we use them to scale up to much larger structures. This approach has been tremendously successful over the last 30 or 40 years.

The need to have these court analytic equations which tell you the phenomenon -- the causal properties and the limitations. But these incredibly complicated phenomena is systems that involve people or economics and so on and there are [Indiscernible] how these people behave are why a bunch of -- are why these people are doing the things we do are why this seen him him by is actually traffic. We don't understand these things. Over the last few years there's been a different approach that's been used which is basically lots of data and lots of data and simply fit a curve. The difference between correlation and causation. It massively successful in the last 10 years. One of the most high profile initial victories for was Google spell correction. There's an attempt by many for a year to have to use dictionaries and such to do correction. It just did not work out because much the time -- [Indiscernible]. Since then extensive use of the -- so much of Google's [Indiscernible] as a function -- function of a user and vendor and user and all these kinds of things. There driven by critical mass. Facebook newsfeeds -- it look more recently features like Google photos with a recognize faces and this picture was taken next to the Statue of Liberty. [Indiscernible] most of these belong to a web ecosystem. Products designed by [Indiscernible] but computer people the kind of products are built. The opportunity is much larger. It's already nice to certain part of California from a certain time it's probably an ad for organic kale chips will get a response. But I believe there's more important projects we can do. The whole goal is we can take these public systems and use this kind of information. My favorite challenge problem is always discussion going on about economic policies of good government or that government. [Indiscernible] different people believe different things and because they say it's more of a doesn't mean it's real. [Indiscernible] imagine if you had data -- Internet data and not individuals and enough data to characterize their behavior and so on. And a new policy you could run simulations and realize that the same policies have one effect on the German population and a different effect on the Greek population. It up and you can have an expectation of what would happen under these different circumstances. This is a huge problem and huge impact with a juicy problem.

Building the systems is an absolute black art. Your stories of companies which do data mining to look for different things for the government but you talk to anyone in the company like this and [Indiscernible]. 80% of the work is in data mining and getting [Indiscernible]. It's a brutal and unexplainable unpredictable and there are many things -- to look at complex system or any of the other systems it's layers and layers on top of the core system to catch all the [Indiscernible]. At some point the layers become as big [Indiscernible] I'm going to do a deep dive on a couple of topics. The first and going to talk about web scale structure data. We had this infatuation I'm creating extremely large structures of structure data. And it's really have any 195 and I turned my attention to doing it on a large scale. I'm going to talk about what I call schema.org and then also get into some interesting math called reference by description. We usually started targeting the structure data was not far from the surface. Starting as early as 95 there's number of websites that basically had relation databases on the backend and used scripts to put it into a human computer form. There were many attempts to make structure data [Indiscernible]. Is the more data you got there had this thing which to the structure data and present into a human form is be some to function of [Indiscernible] the structure data is the first class think I go to my favorite technologist for an example. IMDb which has a bunch of information about Chuck Norris and similar information. The idea that both of these sites would give us this information in a platform that had -- [Indiscernible] due to the graph data model the second one is a common vocabulary. If we have every site giving us this data in a different format would be in trouble. In a different schema we be in trouble. There are many attempts at making this happen MCF and RDF and so on. Most of these in the only one successful or things like RSS and the card but they are successful but very narrow in scope. The goal is to have a wide variety of data and widely used [Indiscernible] Google and Yahoo and other started using the structure data in research results. It was a [Indiscernible] there can be many different schemas and have chaos. So may different ways [Indiscernible]. It was started in August 2010 initially was Google Microsoft and Yahoo answers and it was others have joined. It provides core vocabulary for people places events offers et cetera. And you provided data and it will be in the search engines. The structure data in the search was initially a driving application. An example of this is this is an event taking place not too far from here. Next weekend. Something happening with food. The way it looks to a human being it turns out that underneath this is a markup which shows this information. There's all this structure data underneath the page. In addition to the human data there's the structure data onto the page. How does this work? There is a Pugsley 700 terms of vocabulary is initially by 50 million sites. If you look at Google search index about 30% of pages have search index have this markup and any page which has it there's roughly 25 of these triples. Over the last 12 months or so it's gone by 30%. It spread from webpages to emails so 50% of US and EU emails that have sales confirmation and things like that have this markup. These are all the main sites now. There are few things that happened. It started out companies and now it is open to everyone to join. So an informal structure. Other entities that wanted have extensions for their particular niche thing. Is just a core and GS one is organization that the standards for [Indiscernible] the financial transaction people wanted an extension and real estate. There are new interesting applications. You might get a reservation on open table today but that page email confirmation has this markup which tells you yes your reservation at such a place at such and such a time and someone. This gets picked up by Gmail we send it to your phone is not just for Google to make it happen. And then what happens is the Google knows that you would appointment at 6:30 that takes one hour to get there so excited 30 it will tell you it's time for you to leave. It's working across -- is not specific to open table or Gmail. It's touching many hundreds of millions of information

---

[Indiscernible] every airline ticket information you get it's across the board. [Indiscernible] here's another application. Wise is doing better than so many of the other options? You have to deliver value. In the simplicity value trade-off is there which was in there before. [Indiscernible] and the most important is the we drop some key semantic principles. In particular [Indiscernible]. There's about a few thousand terms. Any particular site you probably 10 or 20 of these things and be talking about the concept of [Indiscernible]. We look at things like Chuck Norris and find Oklahoma there are millions of these things. Just in the world of movies there is half 1 million movies that have been produced and see cannot expect all these to synchronize to have any ID for all these things and synchronize. It's not going to work. We is a thing called reference by description and applications like search to do that. What is reference by description? Most things we communicate about don't have me. But they should. If I say Lincoln I can the president or the town or the car and so on. So I give a reference by describing it. I'm associating a description and this phenomenon is ubiquitous. A bunch of articles in the New York Times and when you find every entity that is referenced has additional facts mentioned about the entity. And so with [Indiscernible]. A problem and communication is not reproducing at one point exactly or proximally a message selected another point. = Frequency packages -- messages have meaning. There refer to or are correlated accorded some system with certain physical or conceptual entities. Some of the message -- the semantic aspects of communication are relevant to the engineering problem. That makes sense. He was try to solve the problem of the two ends of the phone. The signal across but now our value that the somatic [Indiscernible].

There are so many feeds and so many things coming in and most observations you could combine. It's basically an in-store problem. The hope is this could be the basis for reference by description. It means that other than just saying James, neurosurgeon in Gilroy so the receiving program can put it -- quote connected. How big is going to be? It's not enough to just say I'm going attach the following fields. Because in some cases it's not enough. There's an complexity. [Indiscernible] we have a new model -- why is like in the president? It works because we share common knowledge about the world. If I was start talking to a Klingon that would work. We also noted Lincoln -- [Indiscernible] the sender has some entity represented in the message with augmented with descriptions is headed across to the receiver. That's the communication model. It has to do with questions like had we shared knowledge and shared language? How big this description need to be? What's the minimum required for sharing? For contact they had the situation where the message from Italy. There's no idea of shared but there's underlying knowledge of the world in this case they make use of the fact that can we bootstrap what's the commutation overhead? How much can I reveal about something when I'm 20 identified? I don't give you a name I to say a person who shops at Walmart) so much who likes dog food and to -- how much can I believe for real the identity -- before I reveal the identity?

You can measure ambiguity nicely using Shannon's -- effectively each of these things as unique reference there's no entropy. [Indiscernible] the entropy of the possible message is a good measure of the shared language. This is a classic way of talking about information shared. And it doesn't directly work here so I'm 20 description of some if they start saying things like the person has two eyes and one heart and two lungs is not useful. It's not discriminating. We need to do is look for information content of the most discriminating description. And say what are the things that are salient points. And that actually gets a nice mathematical equation. And [Indiscernible] Raven, with a description of checking to see this unique I might not be able to look at multiple category description so measure how many disruptions and willing to look at? The final thing is [Indiscernible] describes of the in terms of other things that are less ambiguous. Lincoln the president -- this image with the ambiguity of the descriptor [Indiscernible]. This trade-off between shared language and shared commutation. Humans use a form of discrepancies are very flat and easy to decode. You can a member of shared knowledge you can bootstrap from no shared language. But at the cost of initial computation and initial descriptions. You also get description size for not identifying description and there's a phase transition when a boundary where you go nonidentifying to mostly identifying.

The talk about teachable learning systems. These are more than [Indiscernible] their unpredictable and this is supposed to be a school bus. You have no idea why. The other thing is there very bad for data predicting the because you need so much data. And contrast with the human -- [Indiscernible] a can tell you this has a trunk. If you look at classical AI systems knowledge base systems you tell them things. But you to tell them everything. They're extremely expressive in terms of what you can tell them but they're very predictable and typically cannot learn new things. Most learning base systems learned lots of simple generalizations mostly propositional representations and public things they cannot. Why is it that a kick in recognize [Indiscernible] the gross anatomic characteristics [Indiscernible]. The goal is to create a systems that combine the two. There's a way of combining them but it's very crude. [Indiscernible] what we want to do is actually the general knowledge in a form and the into the learning system so that you can improve the results. Not only improve performance but also use that to generate information. The key problem here is that most interesting statements like the ones I just mentioned are impossible to fit with feature vectors. [Indiscernible] the problem is if you go for knowledge-based representation with the learning system all of the learning systems depend on differentiation. There's some reason [Indiscernible] and map these into a set of points and vectors and then give them bootstrap from that and you get adjusting semi normal [Indiscernible]. But you can do the problems that I mentioned.

I think empirical modeling is great but work remains. It's a black art. Employment data together to learning components and so on. The only way this will make progress is by actually having 100 X model system. The way you get one times more research is data. Data research sets [Indiscernible]. This loan sky survey and more recently image net. [Indiscernible]. This difference between these. Using a data set -- here's a link to it download and have fun. [Indiscernible] work they don't have access to mapping computers and so on. They do it [Indiscernible] you've lost the 9% of information. It's just too hard. Proposal is to burn data Commons. Google and Microsoft are going to do this but they think of a data set [Indiscernible] if you want to take a petabyte and move it once the cost is about 20 5K. Those storage prices going down much faster [Indiscernible]. It makes no sense to move data around all the time. It's much more sense to move the code. Videocassette interesting observations? And the data [Indiscernible] imagine students and faculty with accounts [Indiscernible] you upload your code. The goal is it you have an idea and you have some code with mining algorithm or whatever and you run it should be less than 30 minutes. Will be interesting ecosystem of shared data sets and tools and applications. It's derivative data that makes interesting. [Indiscernible] I have the whole rest of the web to link it to to make my page more interesting. I have the rest of the web with images I can pull in into my page to make it more interesting. [Indiscernible]. The largest training systems in the data -- student -- [Indiscernible]. Something like 150 million events like that. A large data set. That page has markup in English and a structured language and others had to check the same thing across different languages. Huge problem. Here are five papers of the last four years. Why? On the data set. It's a data share but it's billions of data across millions of sites. If you look at IMDb [Indiscernible]. If someone wants to use all of this has to go [Indiscernible] then you have to put this together. What we need is an program called stitch which takes these data and present for someone to come along and cinches it together and in some of his I can do data mining and look at this. Them have layers of abstraction [Indiscernible]. You have things like browser history and privacy.

A GPS tracking that people try to figure out what are you doing and when you know that equipment simulations. In Palo Alto is a school called gun school. And there's a [Indiscernible]. Can we draw this abstraction and make the data available to someone else can come and figure out [Indiscernible]. There's lots of data in the cloud is there a nationally [Indiscernible]. What's happened in 1993 [Indiscernible] it could have been different. You had America Online and Bill Gates bought a book called the vote ahead with its but the information superhighway and that was not even mentioned once. -- Internet was not mentioned once. And 95 Emison was shut down. They had so much money and so many ideas [Indiscernible]. In the fall of 94 [Indiscernible] the reasons his web thing would not succeed and he was right on every one of the reasons. Does not have identity or security. On the other side you have putting up pictures of coffee cans. What happened? Anyone could contribute. It is not optimized for performance. Increment talking about why the Internet would never work [Indiscernible]. We look at the document -- [Indiscernible]. It easy to start contributing.

Data Commons has to come in the systems do not come from industry. [Indiscernible] they come from academia. -- Academics. They come from people who value simplicity and relevance. Industry can build provide the resources but it has to come from academics. The first version of this Microsoft and Google are both [Indiscernible]. To get this thing bootstrapped and by the end of that. It would move toward [Indiscernible]. You guys are already cutting checks for the professor to buy drones to use for grad students -- it makes no sense to do that because the skills which these guys are operating in economies of scale are better. And time is that is that research.

Have a favorite story about how this [Indiscernible]. It takes about 30 years the technology [Indiscernible] they were all hand done. There were more books but still about that size and make copies of these. [Indiscernible] it should be possible to put a book into a bag -- a saddlebag and to do that you get seven other [Indiscernible]. You could close and come back to where you left with page numbers. You can even a table of contents and so on. And the first book to be done in this form [Indiscernible]. My theory is that is in the computer engineering is a lot like what Gutenberg did with the Bible. More of the same but faster. And we can do is actually do a new kind of modeling see can go and model far more complex things that no one ever talked about. And with that I'll take questions.

[Applause]

We have some time for questions. Are there any questions in the room? And we can queue up online.

[Indiscernible] I'm sure there are websites [Indiscernible] what is the constraint [Indiscernible]

You have to think about [Indiscernible] this piece of information came from this site. A commitment to define -- a user reads [Indiscernible].

[Inaudible - low volume] you already have more shared language. And shared context and likely interpretation for a work.

Could that be built in?

It's built into the theory.

[Indiscernible] point-to-point. He cannot do [Indiscernible]. Talking to a large audience [Indiscernible] I need to be able to put a description [Indiscernible]. There is already an issue of you talking to a complete stranger and you'll like to bootstrap to the language you have eventually. What is the path you follow? [Indiscernible].

I'm education director here. One of the things we are interested in is how to improve the education systems. We had interesting talk a few days ago [Indiscernible]. He's expanded that to schemes of how to improve educational ecology and teachers administrators in the fashion which [Indiscernible]. Your again building a large system for engineers and basically all the systems [Indiscernible]. Isolate the area of engineering [Indiscernible] in critical modeling there's value in that [Indiscernible]. But there's areas and engineering of attempted to go beyond models to achieve that and still have not been able to solve the complex problems. One area is turbulence. [Indiscernible]. There's areas of engineering the week moved in a direction but there still value [Indiscernible] you could be have one or the other in the thing that will have to [Indiscernible]. What happens is that people do these populations [Indiscernible] people are building bridges long before [Indiscernible].

There's no danger of that happening. [Indiscernible]. Most of what I do is what our tree [Indiscernible]. Enough engineering is focusing on the [Indiscernible]. You're looking under the lamppost. That looking at problems you don't want to solve. Maybe they can be solved. The reason that will happen is we put together the data and have a bunch of [Indiscernible]

What you are referring to is showing us take a take these complex problems and challenges and put [Indiscernible] but in a crowd sourced way and get experts [Indiscernible]. It's not one or the other.

[Indiscernible]. I do not mean to say one or the other. But right now [Indiscernible].

On schema.org to have Xs's of the schematic web or what the relationship?

[Inaudible - multiple speakers]

To me is not a standard thing. It is set of ideas. The idea of having -- I don't believe things like [Indiscernible]. The value proposition was wrong. Schema.org [Indiscernible] let's have a large amount of machine crushable data available to people.'s of interesting things would happen. This was schema.org was my intent [Indiscernible] it does make sense for someone else to come along [Indiscernible] unless you have so many different [Indiscernible]. They should use some of the methods but don't expect everyone to coordinate for every study for every patient -- [Indiscernible]. What is the core to standardize on? How do you make that happen? A lot of the things [Indiscernible]. But it's not really [Indiscernible].

[Inaudible - low volume]

What of the economics of uploading your code to the data? Have you considered that? If too many people assigned up --

---

Someone has to pay for it. The hope is that -- the reason for Google and Microsoft [Indiscernible]. At the end of three years of the people who are funding the said I can't live without this. They make the money will show up. [Indiscernible] they're already paying for it in a different way. And hopefully the second or third order would be so large that [Indiscernible]. You can get compute time for the University [Indiscernible]. Eventually there has to be an economic [Indiscernible]. Are there any online questions?

I'm showing no questions on the audio.

We're right on the.. Again thank you very much. [Applause] this concludes today's conference call. You may disconnect at this time.

[Event Concluded]

Actions