Data Science: What It Is and How It Can Help Your Company

Table of Contents

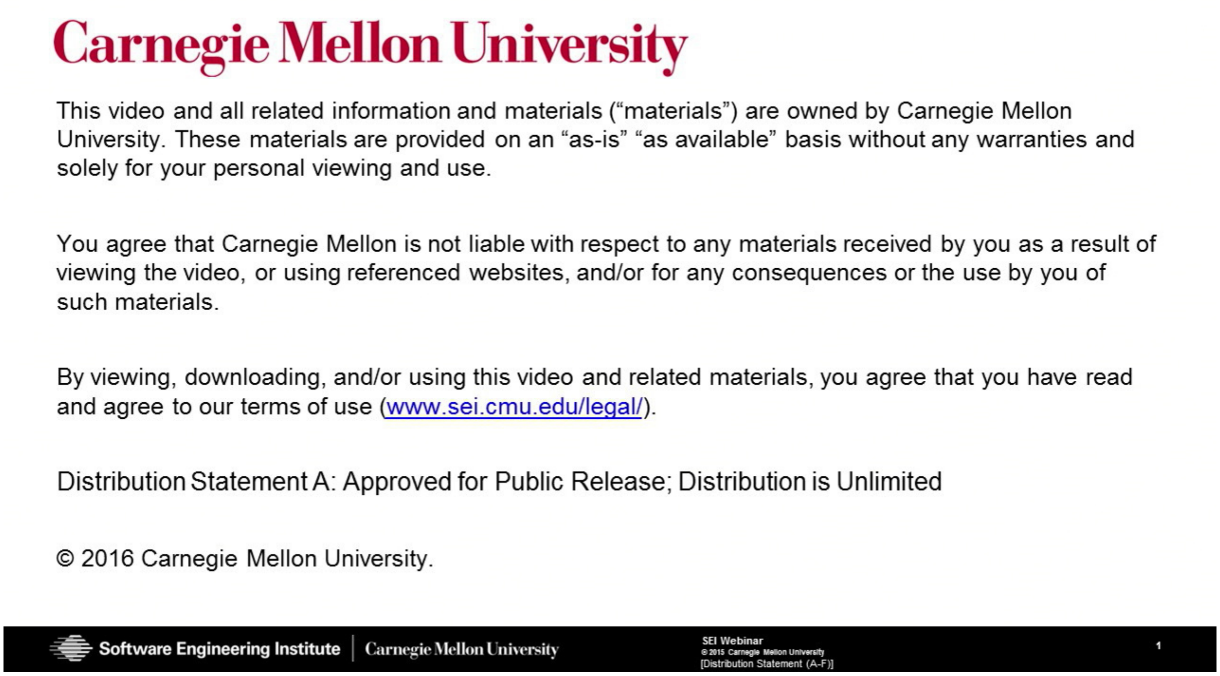
SEI WEBINAR SERIES | Keeping you informed of the latest solutions

SEI WEBINAR SERIES | Keeping you informed of the latest solutions

#SEIwebinar

Software Engineering Institute | Carnegie Mellon University

Carnegie Mellon University



Carnegie Mellon University

This video and all related information and materials ("materials") are owned by Carnegie Mellon University. These materials are provided on an "as-is" "as available" basis without any warranties and solely for your personal viewing and use.

You agree that Carnegie Mellon is not liable with respect to any materials received by you as a result of viewing the video, or using referenced websites, and/or for any consequences or the use by you of such materials.

By viewing, downloading, and/or using this video and related materials, you agree that you have read and agree to our terms of use (www.sei.cmu.edu/legal/).

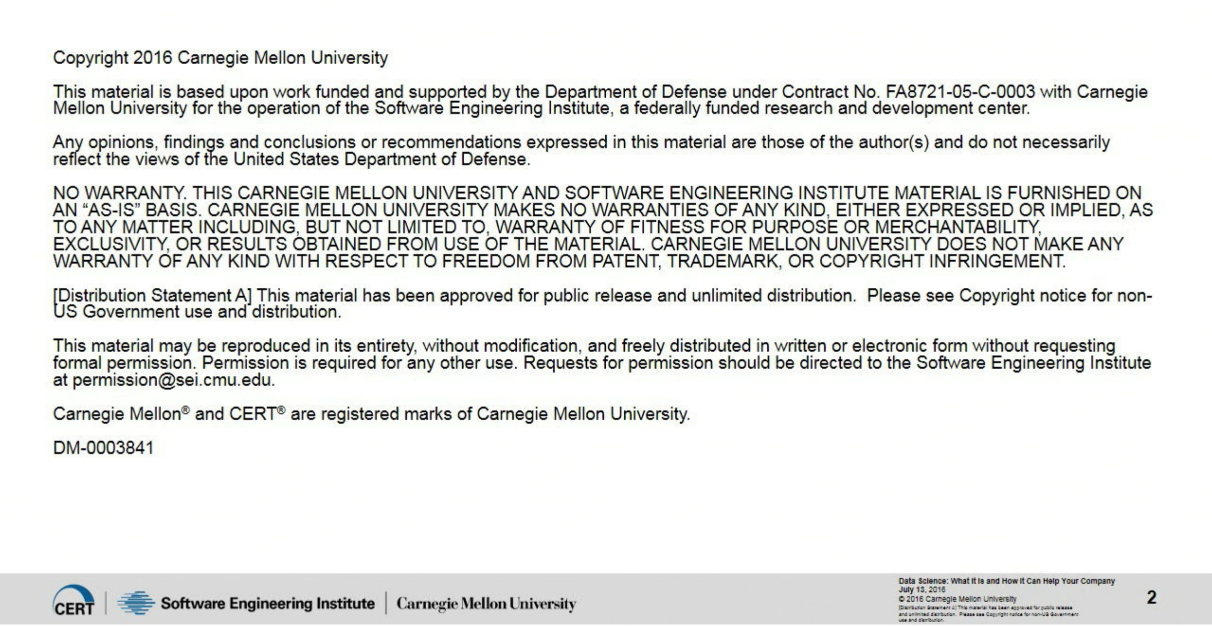
Distribution Statement A: Approved for Public Release; Distribution is Unlimited

Software Engineering Institute | Carnegie Mellon University

SEI Webinar
© 2015 Carnegie Mellon University
[Distribution Statement (A-F)]

1

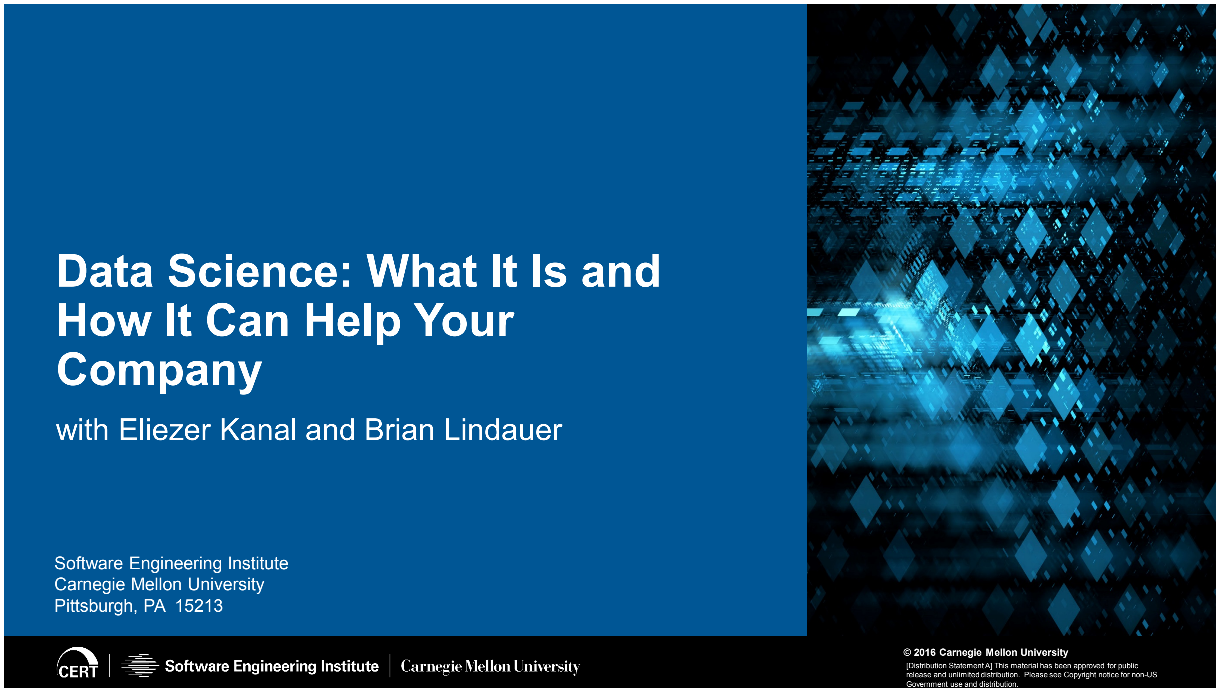
Copyright 2016 Carnegie Mellon University



Data Science: What It Is and How It Can Help Your Company
July 13, 2016
© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

2

Data Science: What It Is and How It Can Help Your Company



Data Science: What It Is and How It Can Help Your Company

with Eliezer Kanal and Brian Lindauer

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

**004 Presenter: And hello from the campus of Carnegie Mellon University in Pittsburgh, Pennsylvania. We welcome you to the Software Engineering Institute's webinar series. Our presentation today is "Data Science: What It Is and How It Can Help Your Company".

Depending on your location, we wish you a good morning, a good afternoon, or good evening. My name is Shane McGraw. I'll be your moderator for today's presentation. And I'd like to thank you for attending. We want to make today as interactive as possible. So, we will address questions throughout the presentation and again at the end of the presentation. You can submit your questions to our event staff at any time through the ask a question or chat tabs on your control panel. We will also take a few polling questions throughout the presentation. And they will appear as a pop-up window on your screen. The first question we'd like to ask is how you heard of today's event.

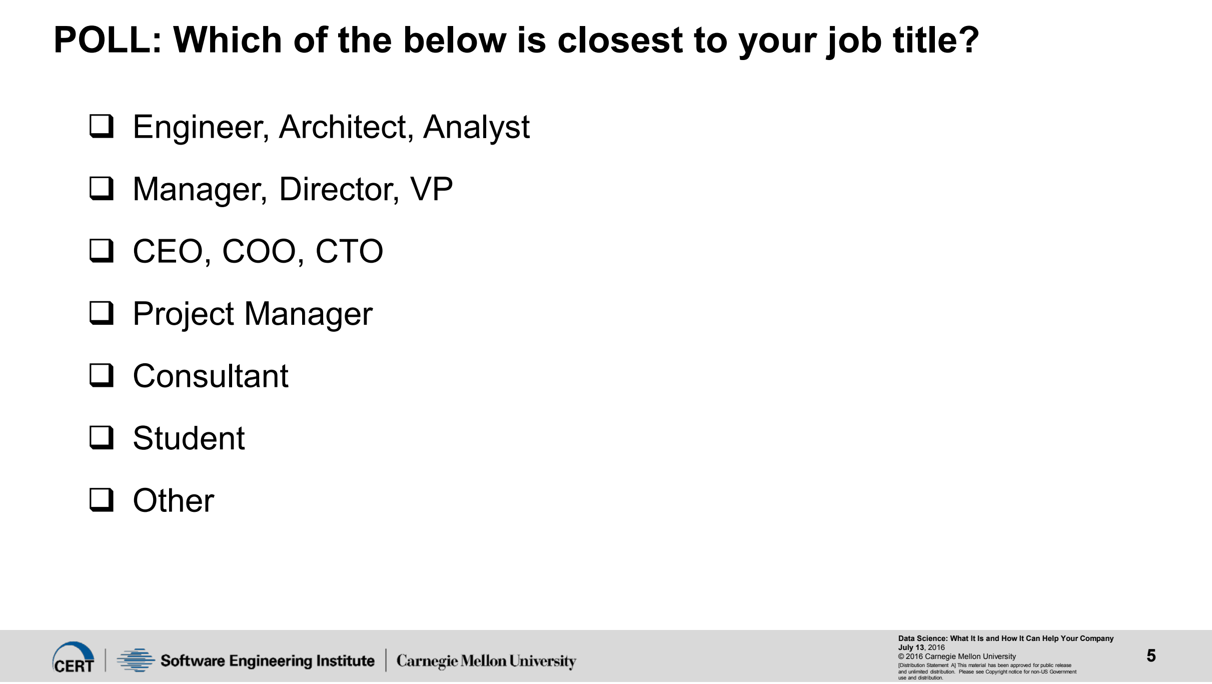
Another three tabs I'd like to point out are the download materials, Twitter, and survey tabs. The download materials tab has a PDF copy of the presentation slides there now along with other data science related work and resources from the SEI. For those of you using Twitter, be sure to follow CERT_division and use the hashtag SEIwebinar.

Now, I'd like to introduce our presenters for today. Elli Kanal is a technical manager in CERT's science of cybersecurity group. He started his management career at Highmark Inc. leading a team of analysts within the health plan operations division. Brian Lindauer is a research scientist in CERT's Science of Cyber Security group. His research focus is on applications of machine learning and data mining to cybersecurity problems.

And now, I'm going to turn over to Elli Kanal and Brian. Elli, Brian, welcome.

Presenter: Thank you very much, Shane. I appreciate that. And today's topic, as we said earlier, is data science, what it is and how it can help your company. To get started, quickly I just want to put out a couple poll questions regarding your background just to see who we have today in the audience.

POLL: Which of the below is closest to your job title?



POLL: Which of the below is closest to your job title?

- ❑ Engineer, Architect, Analyst
- ❑ Manager, Director, VP
- ❑ CEO, COO, CTO
- ❑ Project Manager
- ❑ Consultant
- ❑ Student
- ❑ Other

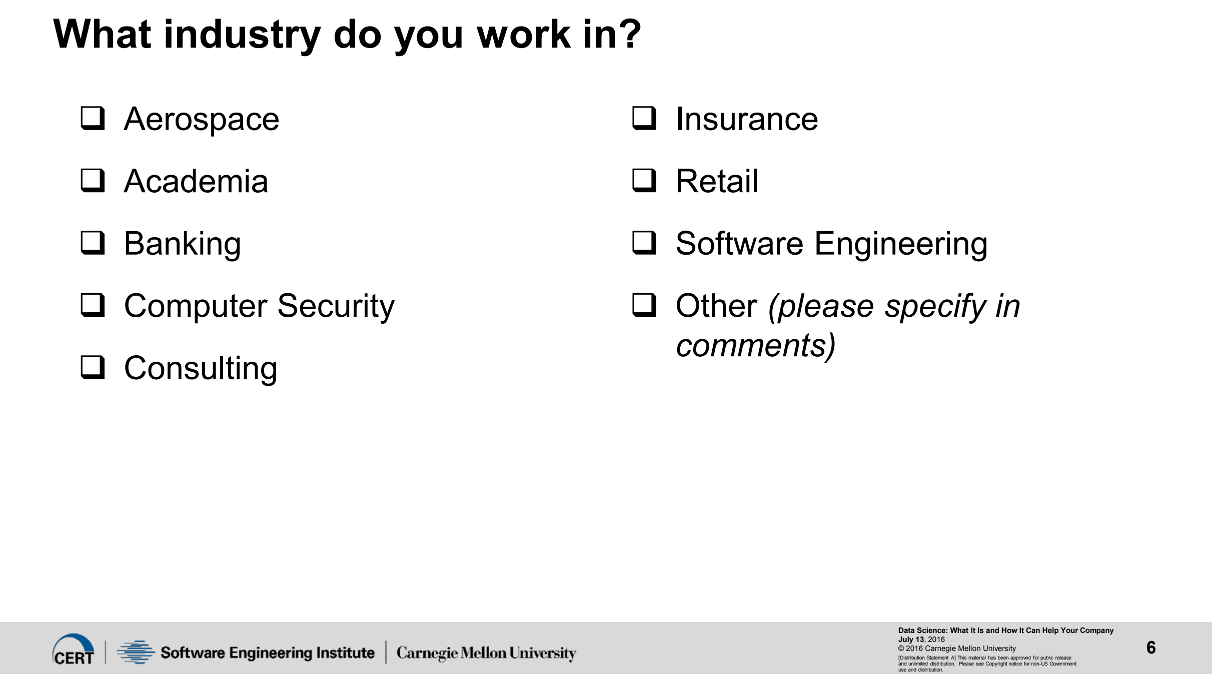
CERT | Software Engineering Institute | Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016
© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

5

**005 I quickly want to know basically what your job title is, get a sense as to where you guys stand in your different organizations,

What industry do you work in?



What industry do you work in?

- ❑ Aerospace
- ❑ Academia
- ❑ Banking
- ❑ Computer Security
- ❑ Consulting
- ❑ Insurance
- ❑ Retail
- ❑ Software Engineering
- ❑ Other *(please specify in comments)*

CERT
Software Engineering Institute
Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016

6

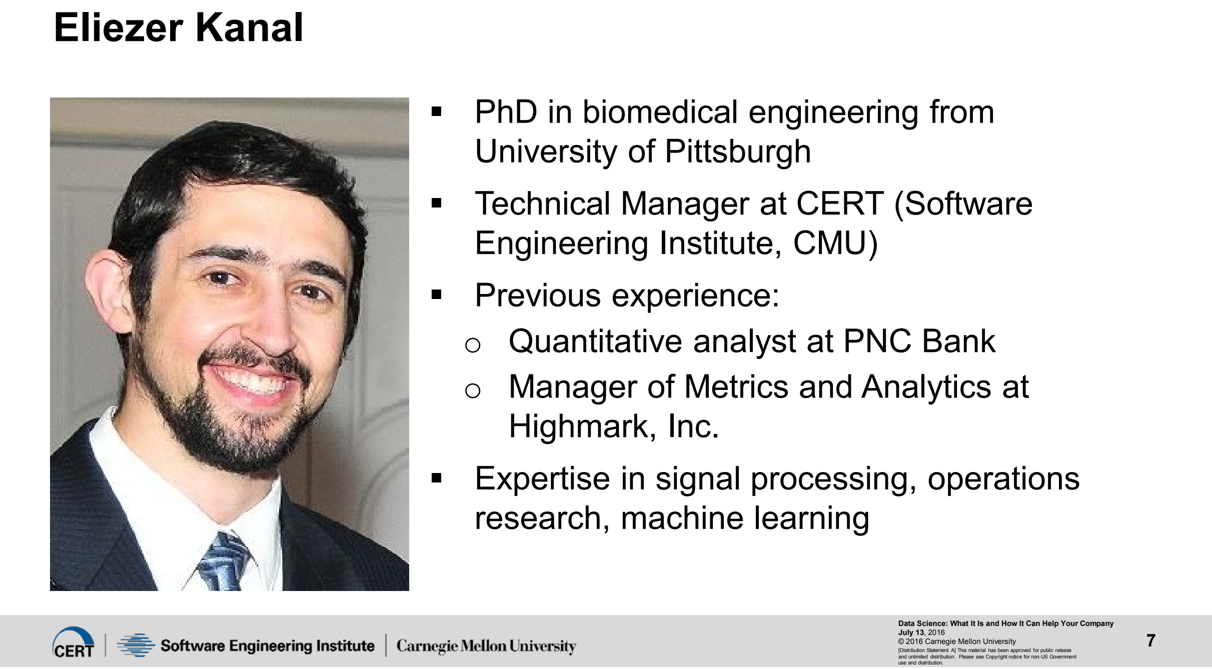
**006 As well as what industry you're coming from.

The overall presentation that we want to go through today follows the following general outline. We want to go through what data science is broadly, how you define it, what you expect someone who has a data science title, go through a couple different examples of just what machine learning is, really kind of a brief treatment of that, and then go through as many case studies as we can just detailing how does data science really apply in the modern world, how can it benefit your organization. And based on that, I want to give you guys a good sense as to what data science is and how it helps you.

I want this to be as applicable as possible to all of you guys. So, please, let us know what you-- where you're from, what your background is. And we'll try to make this as tailored as we can to you guys specifically.

More on that, we have a number of case studies presented from a number of different industries. But if there's something you guys want to hear specifically or something that's interesting to you, please put it in the comments and pass it on to Shane. And he'll pass it on to me. And we'll try to address that as much as we can.

Eliezer Kanal



Eliezer Kanal

- PhD in biomedical engineering from University of Pittsburgh
- Technical Manager at CERT (Software Engineering Institute, CMU)
- Previous experience:
  - Quantitative analyst at PNC Bank
  - Manager of Metrics and Analytics at Highmark, Inc.
- Expertise in signal processing, operations research, machine learning

CERT | Software Engineering Institute | Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016
© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

7

**007 So, as those come in to Shane, fantastic. Cool, thank you very much. So, it looks like we have actually quite a bunch of engineers here today, a couple managers, project management, and some other people as well, very nice, cool. So, thank you very much. And again, if you could also fill out the--

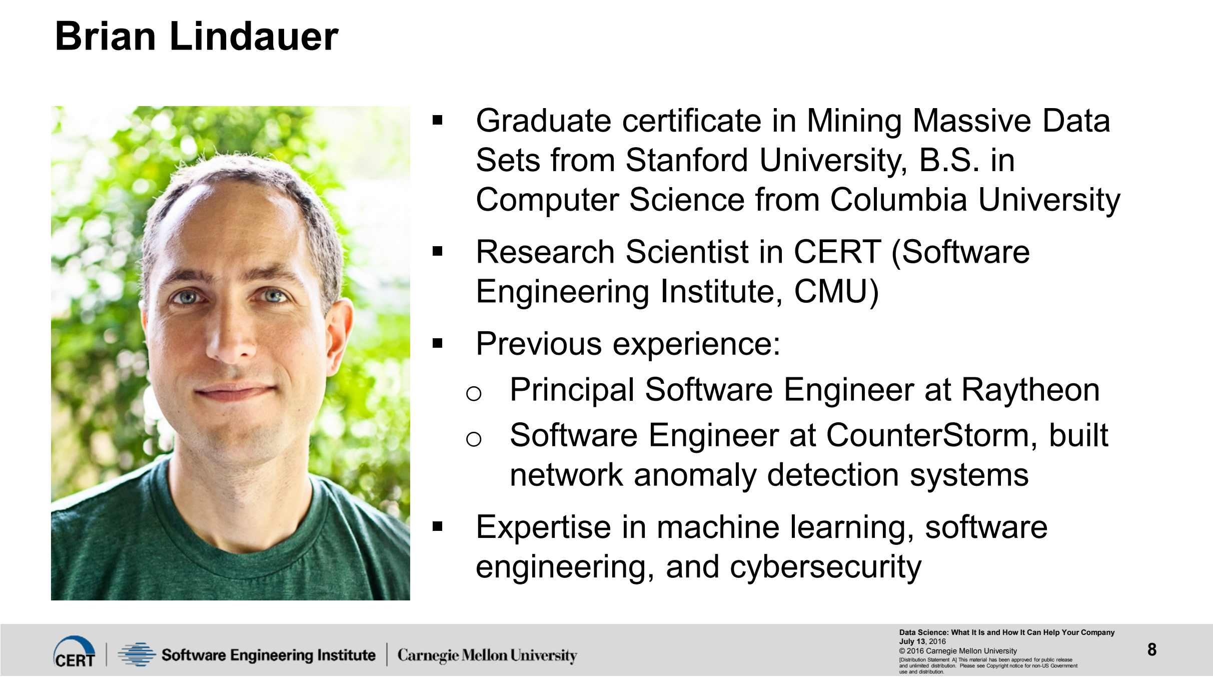
Presenter: We'll do the industry one next.

Presenter: The industry one coming down soon.

So, I'll start talking just a bit about my own background. So, I'm currently-- excuse me, to start out, I got my PhD from biomedical engineering from the University of Pittsburgh. I was doing a lot of work related to neuroscience and the brain, understanding how the brain responds to reward conditions. I then recently joined CERT here as a technical manager. We have quite a few interesting projects related to data science and some work on malware, some work going on insider threat, which we'll talk about a little bit later, some other work going on related to understanding coding, which we'll talk about later. But prior to that, my background actually varied a bit from the work that I'm doing here. I was doing some work at Highmark doing data science as it applies to insurance, Highmark Inc., as well as PNC Bank doing some work there for statistical modeling as it relates to corporate credit risk. Quite a lot of different skills were in use in each of those positions. But I ended up over here.

Brian, you want to introduce yourself?
Presenter: Sure, so my background is a little different from Elli's.

Brian Lindauer



Brian Lindauer

- Graduate certificate in Mining Massive Data Sets from Stanford University, B.S. in Computer Science from Columbia University
- Research Scientist in CERT (Software Engineering Institute, CMU)
- Previous experience:
  - Principal Software Engineer at Raytheon
  - Software Engineer at CounterStorm, built network anomaly detection systems
- Expertise in machine learning, software engineering, and cybersecurity

CERT | Software Engineering Institute | Carnegie Mellon University

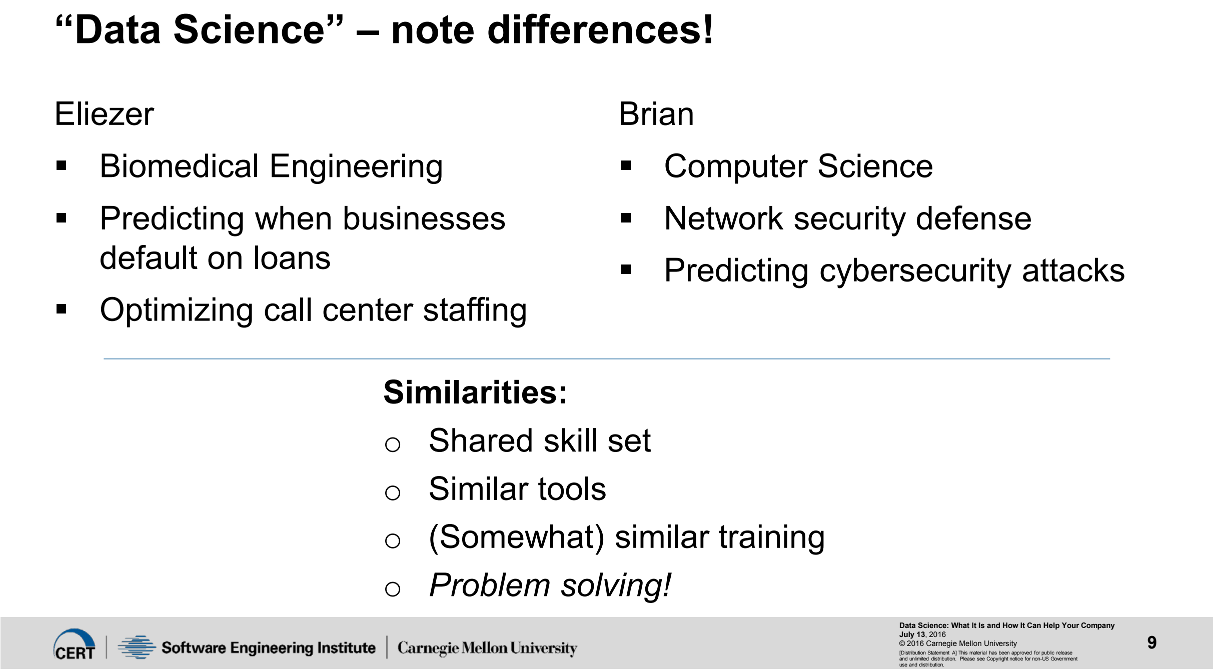
Data Science: What It Is and How It Can Help Your Company
July 13, 2016

8

**008 My degree is in computer science with some graduate work in data mining and machine learning. And really, for most of my career, I was doing software engineering and specifically was working on security problems and network security before I was kind of transitioning into a more heavily data science world. So, it's a little bit different from your background, Elli.

Presenter: Yeah, that's actually one of the first things that I kind of wanted to highlight.

“Data Science” – note differences!



“Data Science” – note differences!

Eliezer

- Biomedical Engineering
- Predicting when businesses default on loans
- Optimizing call center staffing

Brian

- Computer Science
- Network security defense
- Predicting cybersecurity attacks

Similarities:

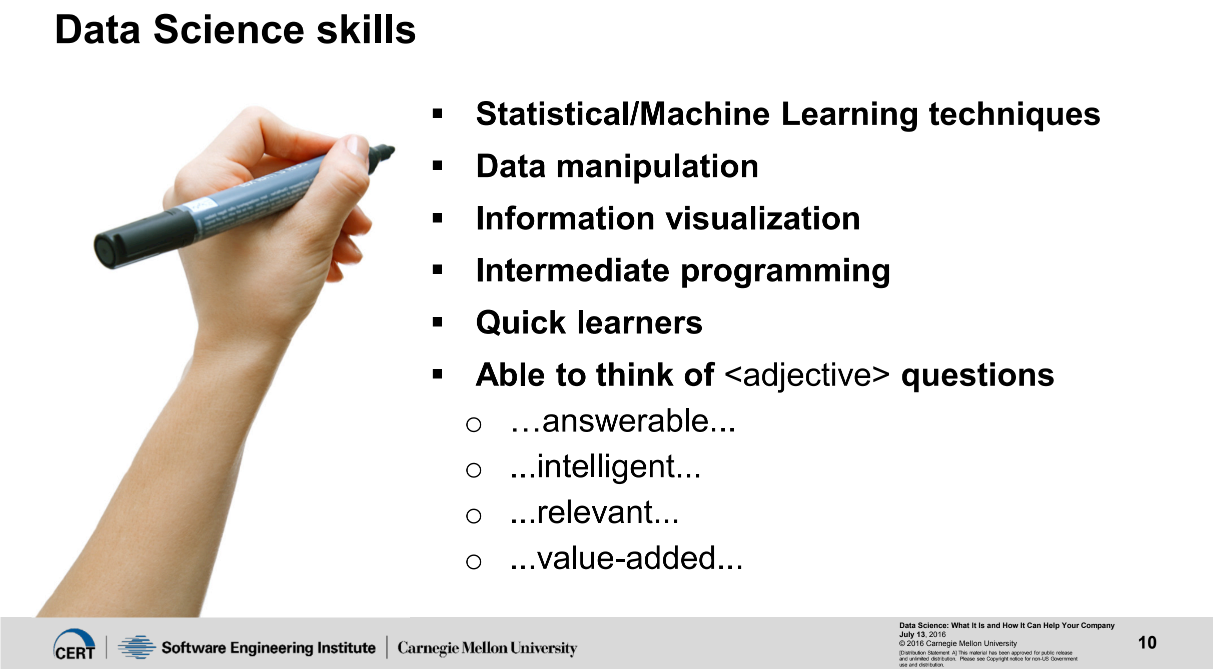
- Shared skill set
- Similar tools
- (Somewhat) similar training
- *Problem solving!*

**009 So, one of the most interesting things that I think you've probably seen, and I've definitely seen in the data science field is that the backgrounds that people come from kind of varies tremendously. So, you can see my background is much more biomedical engineering. I had some work in banking. My training was-- I mean my formal training kind of was all over the place, some math, some sciences, some engineering, some neuroscience, biology. And then I went to the bank. And I did some work in insurance, some of the topics are listed on this slide. That varies kind of a lot from your background, much more traditional computer science, directly into cybersecurity, doing some work with cybersecurity attacks, network analysis and whatnot. And it really does look pretty different. But, I think even to that--

Presenter: Right, I mean even given that, when you start looking closely, we have a lot of similarities. A lot of the coursework that we had was on the same topics. A lot of the methods that we use, the tools that we use are the same. And you know this, it's not unusual to have data science people from very different backgrounds that come together and work on the same problems.

Presenter: Right, right. I mean I have, at this point, I'm working as a manager here. I worked as a manager at Highmark. And I've hired people with backgrounds from physics, statistics, public policy. I mean it's really-- the training can kind of come in from anywhere. And there's a kind of a core set of skills I think that any really good data scientist would have.

Data Science skills



Data Science skills

- Statistical/Machine Learning techniques
- Data manipulation
- Information visualization
- Intermediate programming
- Quick learners
- Able to think of <adjective> questions
  - …answerable...
  - ...intelligent...
  - ...relevant...
  - ...value-added...

CERT | Software Engineering Institute | Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016
© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

10

**010 We can go through a couple of those right here. The first skill I would argue that any data scientist really has to bring to the table is a solid understanding of statistical machine learning techniques. There's kind of the bread and butter of a data scientist is to understand how to take data and do something useful with it. And in order to do that useful thing, there's a lot of statistics and machine learning that has to go into that just understanding how to build models, how to make predictors and stuff like that.

Presenter: Right, I mean that's the core of our toolset, and that's kind of the magic that people see really comes from-- a lot from that.

Presenter: Right, when you look in the public press, and you're talking about all these cool projects that these different companies are putting out, what's impressing people is the data science part of it, the machine learning that you can actually see in a product.

But once you get past that, I think there's a number of other skills that are very important. One of the first ones which I think people don't recognize how critical it is or how much time gets spent on it is just manipulating data. I mean I've gotten data in the form of PDFs. I've gotten data in the form of hand-- literally printed documents that I had to figure out how to get off the document into a database that I could do something with. And just taking the data off that and putting in a database, that's half the work.

Presenter: Right, I had a colleague who was doing his first sort of project in this area. And he came to me and said, "I'm spending eight percent of my time just kind of cleaning up the data and getting it into the right format. Am I doing something wrong?" And I said, "No, this is exactly what's expected." I mean that's where most of your time is spent.

Presenter: It would be nice if all that data came across cleanly packaged up like a little bow. But that really very rarely happens. And once you have this data, and you're putting it all together, and you have something useful with it, that doesn't really give too much unless you have a really good visualization to go along with it. So, I may have a nifty dataset that I've cleaned up. And it's really perfect. I may have some really useful statistical tools. But unless I can show the output that I want to give to my end user, to the business, to the customers, there's really not much use in it. So, the ability to take that data and visualize it is just crucial.

Presenter: Yeah, and that's actually an area that I've noticed is very different coming from software engineering into sort of a data science role is that communicating your results is a much bigger part of this. Of course, in software engineering, you also have that. But here, it's much more prominent that-- a much more prominent part of your job to not only do the analysis, but to communicate it to people.

Presenter: Right, and at this point, I think there are certain companies that actually employ people whose sole job is visualization. If you look at certain media companies, the whole point of these people's job is to take some very cool data that they have that supports an article or something like that and to make the visualization as smooth as possible, as easy to understand as possible, as easy to manipulate. And often the data can be very complicated. You're looking over maps. You're looking over time. You're looking over multiple dimensions. Making that so that anyone can go to a webpage, click a few buttons, and get a real intuitive sense, that's a skill.

Presenter: Right, it's really become its own sort of subfield of journalism.

Presenter: Exactly, exactly. And journalism's just one area. And you can do that in any of the industries that anyone here is a part of. That's just as critical.

I think that a lot of people get the sense that data scientists are programmers. It's computer science all the way around. Everyone here has this really solid understanding of everything about a computer. And I think that it's worth highlighting that they're probably not. I mean the vast majority of people who have a data science title-- I don't know if vast majority. Many of them are really just mediocre programmers. And I would definitely argue that's probably fine because their skill set is not programming. Their skill set is working with data and generating value from the data. And if they have to work with the programmers the company already has, that's okay.

Presenter: Right, a lot of the code that you write in doing this kind of work is sort of a one off thing, I know we have a lot of engineers in the audience, and the sorts of things that a lot of engineers might consider sort of a throwaway thing because what you're looking for is the output of this one program or script that you've written. And then you want to kind of present that. Once you get to the point where you want to sort of turn this into a real production system that can run at scale, that can be maintained over time, then you'll often see people bring in individuals with more of a software engineering background and maybe with some-- sort of enough knowledge of the machine learning and the data science side to work with the statisticians.

Presenter: You start bringing traditional IT. You start bringing in the rest of the architects and saying, "All right, now you've made something useful. Let's-- you've made a useful prototype, let's turn that into a useful product." And there's definitely a good marriage that happens at that point.

Presenter: Right.

Presenter: I think, so those are kind of-- that's the bread and butter of what goes together. I think in order to have someone really be able to do this in any good company over a period of time, I think the ability to pick up a topic or pick up an expertise pretty quickly turns out to be absolutely critical. And the ability to say here is-- welcome to where you are. Welcome to your company. We're going to be focusing on this for the next X number of months or whatnot or years. It's time for you to pick it up and pick it up quickly.

Presenter: Right, it's a very, very fast moving field. I say that coming from software engineering, which is already a really fast moving field. And this is sort of one level faster moving from my point of view. People are developing new methods all the time. And really the ability to pick them up and keep up with the latest and greatest algorithms is paramount.

Presenter: Right, right, and again, I was talking about before I think, the field of data science is generally full of people who know their way around a computer. And one of the things that they like to do is share their tools. And to that extent, kind of like you can go to some big box store and buy a hammer, you can go out on the Internet. And you can actually find quite a few tools that are useful for data science. And staying on top of what's available and how to use it and how to make that you can get value out of it is-- I mean that's a full time job by itself let alone actually doing the work.

Presenter: Right, and it's great that that hammer is available. But it's also important to know that having the hammer doesn't make you a carpenter. And it's still really important to learn how it works and how to use it.

Presenter: Right, right, and I think the last thing I want to emphasize here is that all of this is just within the data science part. But as all the people here in the audience are well aware, you're always working with the subject matter expert. And being able to get a good understanding of the subject matter, that itself is a skill. Any good data scientist recognizes that they are not going to become the expert. They're learning from the experts, taking what they gain, and then trying to use that to build something. But it's always a give and take. Every time you're building something, you're working with the people who know it better than you to try to develop a good product.

Presenter: Right.

Presenter: Can we chime in real quick just with an audience member asking, "Can you explain the difference between data mining, data warehousing, and data science, and big data?" So, all these buzzwords, is there a quick recap for some of the folks that are brand new to this.

Presenter: Can you state that just real quick?

Presenter: So, we've got data mining, data warehousing, data science, obviously, and big data.

Presenter: Okay, I'll take a first stab. And we'll see where we can go with this. That's actually a very good question because a lot of the-- this field is rife with buzzwords. There's definitely no shortage of that. So, what we're trying to talk about today, we're going to call broadly data science. The topic of data science I would argue is what you see on the board. There's one more point which I'm going to get to, but quite frankly, these are the critical elements of what would make a person a data scientist. So, data science is extracting value from data, actionable value from data.

Data mining is the art of working with the data itself. So, if data science encompasses all the things that I have here showing on the slide, data mining is specifically looking through the data and trying to extract that. That's going to incorporate things like we were talking about with BTLs, actually extracting data, figuring out how to get the data into a database, taking the data back out of it, creating some sort of a model. All that stuff would go into mining. Warehousing, and this is going to sound silly, there's an art to how to store data. Actually, you probably know more about that than I do.

Presenter: Sure, so databases are a whole field in themselves. And when I think of data warehousing, I think of designing data storage in such a way that the data can be stored efficiently and also queried efficiently to answer questions that you might want to answer in the process of doing data mining. I would also add that I don't know that there's sort of a consensus about what all of these terms mean, especially data science. Having used it in the title of our webinar notwithstanding, I think it is a bit of an ambiguous term. And I even heard somebody once remark is there any other kind of science besides data science? So--

Presenter: Really?

Presenter: Yeah, and so data warehousing I think has more to do with the database and the storage of it in a way that it can be analyzed efficiently. That last one I guess was big data?

Presenter: Yeah.

Presenter: And so, I would say that's the subset of all of this that has to do with data that is very large, again a very subjective term. What is large? And that changes from year to year as our computational abilities change. But when data gets to a certain size, it calls for a certain treatment, parallel processing and different kinds of algorithms. Excuse me. So, it does kind of have its own set of requirements.

Presenter: Right, and I think one point to make is that the distinction between data and big data is really a practical one in that it's small data, if I can call it that, you can handle on your desktop, or on a slightly larger server, or on a server that a medium sized company can afford, or go up to there. Big data is data that is absolutely enormous that only the largest companies can possibly house it, or data that is slightly smaller than that, but it takes a very highly trained team to handle it, or it's slightly smaller than that, but it's still bigger than you can actually easily work with. And it's a gradient all the way through. But the distinction is how specialized does your tool have to be. And as the data gets more and more difficult to work with from a practical standpoint, you need to develop special tools to handle it. And that led to the creation of this concept of big data.

Presenter: So, we had a fifth term thrown in, applied statistics. Is there a difference there or another way to differentiate that?

Presenter: I mean to me, applied statistics and data science are potentially synonymous. There are some parts of-- I guess the two terms that people puzzle over more frequently are statistics and machine learning. And there's sort of a maybe ninety-five percent overlap in those fields. But there are a few things that people with a machine learning background do that people with a statistics background might not and vice versa. But I would say, generally speaking, you can think of an applied statistician and a data scientist as being one and the same.

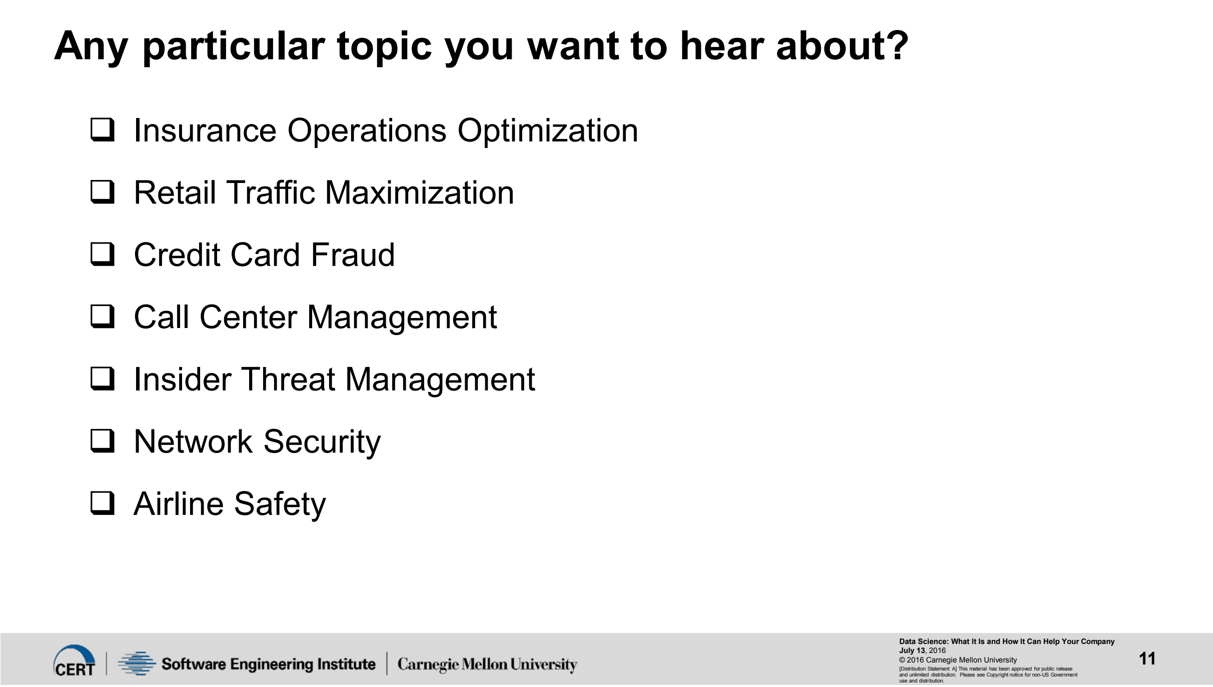
Presenter: Great.

Presenter: Right. And one of the points that I want to convey with this whole discussion is if your company's looking to hire someone in any of these roles, what are you looking for in that individual. You're trying to find certain criteria when you're looking to hire. So, what are you looking for?
So, the goal of this particular slide that I have up, and I'll get to the last one in just a minute, is really to convey here are the sorts of things you're going to want to make sure the person you're hiring has. I think that an applied statistician would be a fairly similar list to what we have here.

Presenter: Right, they would be stronger on the statistical part and probably, generically speaking, a little weaker on say the software engineering part. And so, when you're hiring, it is useful to think about, for yourself and for the specific role that you're hiring for, what kind of data scientist you want. Hopefully, they have some capabilities in all these areas. But there's going to be a broad array of people who are all-- you legitimately call them all data scientists, and you want to know where do you need them to be most strong.

Presenter: Right, and that actually is a perfect segue, thank you, for our last point here on the bullet list, which is that a strong data scientist will be able to think of fill-in-the-blank types of questions that will really add value to your organization. And this, I think, is where all of them should be hopefully as similar to each other as possible. So, you know an applied statistician may have stronger math background than a computer scientist who would have stronger database background or programming background or what have you. But someone who is coming to apply to a data science position should be able to think of questions for your organization. Once they've talked to the subject matter experts, and once they've worked with you on this, the questions should be answerable questions. They'll be able to assist you in determining whether this question has an answer, whether the question is actually relevant to the work that's going on, and whether the question actually adds value. It's very easy sometimes, it's obviously easier than you'd think to get a great question, which doesn't actually help anything. And you really want to make sure that what you're asking provides value to the company and that it drives sales, or it drives revenue, or it drives eyeballs, or whatever it is you're trying to drive. So, with that--

Any particular topic you want to hear about?



Any particular topic you want to hear about?

- ❑ Insurance Operations Optimization
- ❑ Retail Traffic Maximization
- ❑ Credit Card Fraud
- ❑ Call Center Management
- ❑ Insider Threat Management
- ❑ Network Security
- ❑ Airline Safety

CERT | Software Engineering Institute | Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016

11

**011 I wanted to move to our fir-- oh, excuse me, I forgot about the poll.

Presenter: So, I'll launch the poll in a second. I just wanted to put a bow on the industries of the people who are here real quick. Just--

Presenter: Okay, cool.

Presenter: Do that real quickly. As I pose the next-- let me-- I'll pose the next polling question first so they have time to vote.

Presenter: Oh, fantastic, yes, looking over that so you--

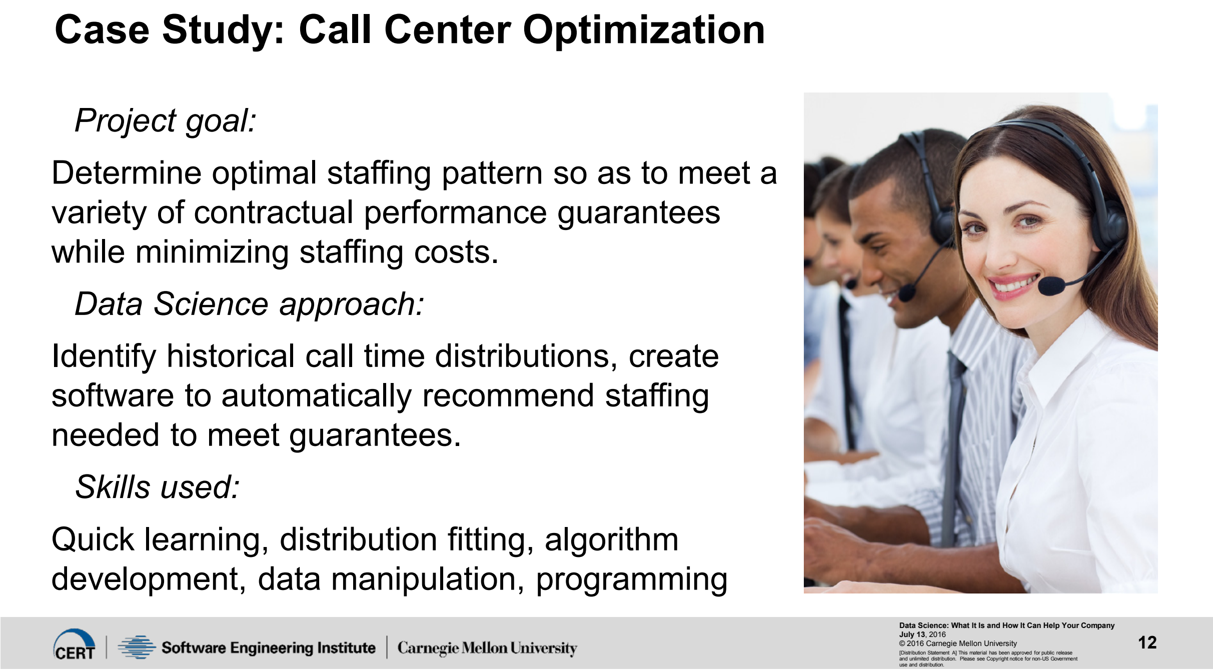
Presenter: Yeah, and then we'll all just--

Presenter: So, looking at this here, it looks like we have quite a few people from the computer security field, some software engineers, and a mix of other domains.

Presenter: Yeah, and some of the other domains that people were chiming in with were DoD analysts, energy industry, federal contractor, healthcare space, telecommunications, manufacturing, ERP. So, a wide range of--

Presenter: Okay, fantastic. And this is actually nice because we have quite a few different case studies I think that's going to actually touch on some of these. And as I said before, if there's a particular question that you want to bring up or an area where you think this could be applied that we don't get to, please feel free to just bring it up and ask about it.

Case Study: Call Center Optimization



Case Study: Call Center Optimization

Project goal:

Determine optimal staffing pattern so as to meet a variety of contractual performance guarantees while minimizing staffing costs.

Data Science approach:

Identify historical call time distributions, create software to automatically recommend staffing needed to meet guarantees.

Skills used:

Quick learning, distribution fitting, algorithm development, data manipulation, programming

CERT | Software Engineering Institute | Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016
© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

12

**012 So, moving on to our-- the first case study which we have to discuss, that's call center optimization. So, the field of answering phones is honestly a lot more big than you'd expect. Basically, a call center is somewhere where people who just sit on the telephone all day. I think people are mostly aware of that. But what they don't recognize is quite how difficult it is to run a call center smoothly. You have many, many people there. Scheduling shifts is exceedingly difficult.

You can possibly have thousands of clients going through a single call center. They could be handling people in different time zones. Each of the clients may have their own performance guarantee out of that call center. They want answer my phone calls within thirty seconds, answer my phone calls within a minute, make sure that this percentage of people are happy when they're done. So, trying to ensure that a call center is run smoothly is really a pretty mean trick.

So, in order to do that, one of the-- there's a lot of data science that you can apply. So, one of the techniques that we've applied-- actually, I had to apply this when I was working at Highmark, you can look at the historical call distributions. Actually, before I get into that, this is something that you can try to do without a data science approach. You can try to do this by hand. Pull some call records. Ty to see exactly what are the calls like, what kind of people are calling in, what times of day. But that turns out to be very unwieldy when you have such large call centers.

So, you can use a data science approach. Using the computer, you can compare all these different-- not compare, you can look at all these different types of phone calls, all these different shifts, all the different things you have coming in, put it all into one large model. And you can generate an optimal schedule that will say all right, using the scheduling that we have here, we're going to have the best number of-- the fastest calls answered. We're going to be able to meet all of our performance guarantees but simultaneously have the fewest people on staff because every extra person who's in that call center is pretty difficult to-- you have to pay them. They're all pretty expensive. So, making sure to manage that is honestly kind of difficult.

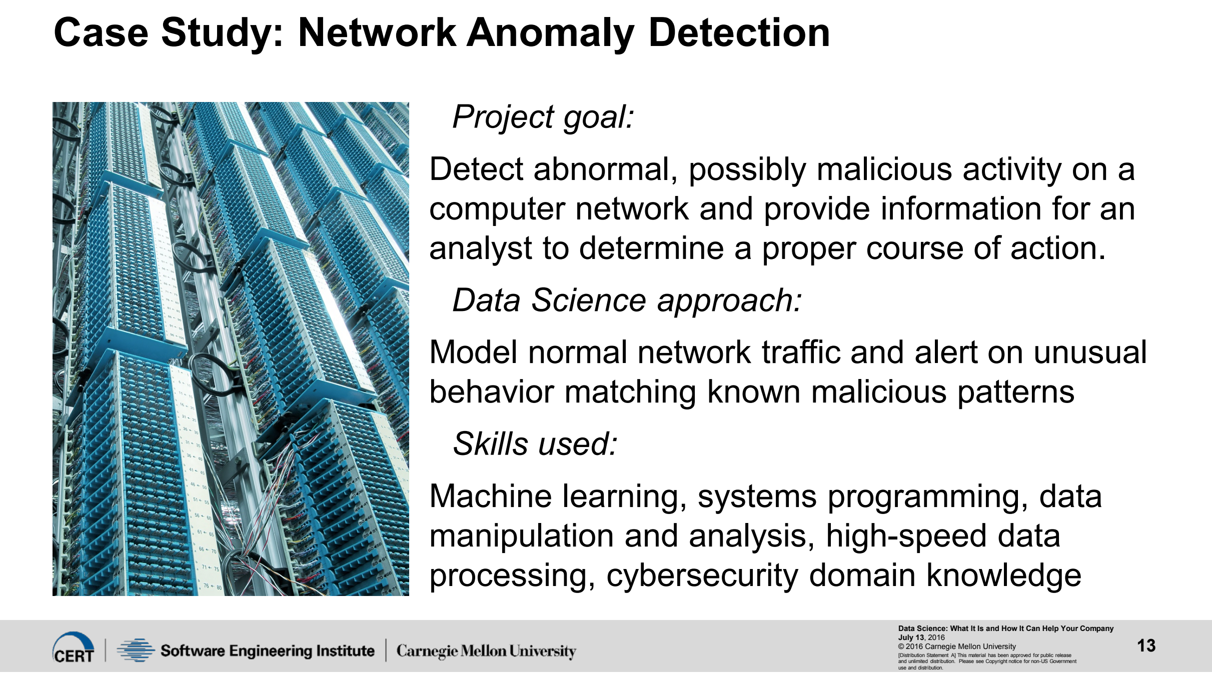
Presenter: What kind of data would you be using as inputs to do that optimization?

Presenter: Right, yeah. One thing which I learned quickly is that I didn't realize quite how much data can be recorded about a phone call. And it turns out quite a lot. Every single step of the phone tree, when you're pushing those buttons to say press one, press three, we can store all that. Time of day, length of call, we can-- after the fact people say, "I was happy. I was not happy." We do look at that. That's stuff that people take into account. And based on all that information, we can say all right, you called. You were upset. Were you upset during an outage? Was there too many people trying to call in? Was there some event going on? And by looking at all this different stuff together, we can really make a pretty robust model. And it's pretty nifty all the things you can create.

The tools that we have are kind of the same stuff we mentioned before. You're handling big data. So, you have to extract all this stuff. You have to store all this stuff. There's huge amounts of data being generated, gigabytes, terabytes depending on how large it is. You need to be able to visualize the call schedule. And at the end of the day, you're making an algorithm to actually do the planning for you. So, all this stuff has to be created.

Presenter: Right.

Case Study: Network Anomaly Detection



Case Study: Network Anomaly Detection

Project goal:

Detect abnormal, possibly malicious activity on a computer network and provide information for an analyst to determine a proper course of action.

Data Science approach:

Model normal network traffic and alert on unusual behavior matching known malicious patterns

Skills used:

Machine learning, systems programming, data manipulation and analysis, high-speed data processing, cybersecurity domain knowledge

CERT | Software Engineering Institute | Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016

13

**013 Presenter: Moving on to--

Presenter: Yeah, so we've got a bunch of computer security people in the audience. So, I guess this will be a good one.

Presenter: Just to chime in real quick. That was the winner of our poll question, was the network security one.

Presenter: Fantastic.

Presenter: Okay, great. So, kind of traditionally the way we do security on a network or on computers is through what are called signatures, which is where somebody writes down a pattern of the bad behavior that they want the system to look for and then alerts when that pattern is encountered. And that's-- everybody's familiar with their antivirus saying I recognize that this piece of software is malicious. There's the same sort of system that runs on most organizations' networks called an intrusion detection system. One of the problems with this approach is that somebody has to actually sit down and write these signatures for what they're looking for. And of course, this is an adversarial environment. And there's always somebody trying to come up with the new thing that's going to get around the signatures in your system.

So, the idea of anomaly detection is instead of writing down everything that's bad, let's watch the network, learn everything that's good and normal and then produce an alert when there's something that doesn't match our model of what's normal. So, this is, in general, a very, very hard problem, some people think maybe not solvable in the general sense. But things that we, as an industry, have learned from trying to do anomaly detection on networks, we've been able to apply that elsewhere in computer security to host-based systems, to specific kinds of networks. So, it's a very interesting area.

And it's a different kind of problem really than what you were just talking about because you're dealing with

streaming data. Network traffic can
be very high-volume. And you have
to be very concerned with processing
it in real time. So, there's definitely
elements to the problem that are
different from the call center situation
where you're kind of doing something
in one big batch and then producing
the answer and taking action on that
answer. It's something that's kind of
updating constantly.

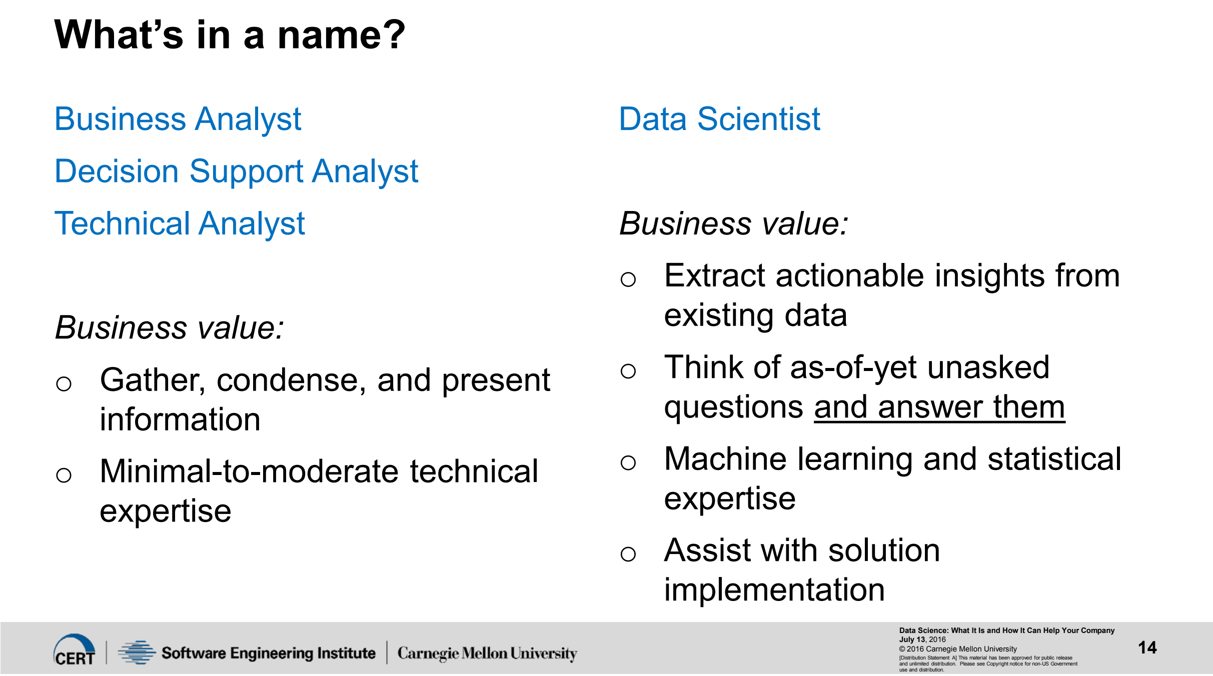
Presenter: Right, I imagine there's
actually quite a lot of data you have
in the network already within your
own internal network. Is there other
stuff you can marry that with outside
the work?

Presenter: Right. Yeah, I mean so
one of the ways to deal with the
problem of anomaly detection
producing a lot of kind of false alarms
is to bring in more context. So, there
are other kinds of records besides
network data that you might look at
like file access records, logins, to
understand who's doing this activity
and what their job role is and things
like that. So, there are definitely lots
of other kinds of data on a network
that can be used to do machine
learning for security on a network.
And using them kind of in an ensemble
fashion is definitely sort of a best practice.

Presenter: Right. Right, very cool.
And we'll get-- I see also that insider
threat was something that you guys
were interested in talking about later.
So, we'll definitely get to that as well.
If you have any more questions
about either insider threat or just

network detection activity, detection network monitoring, please feel free to send them in. And we'll definitely get to them as we get through. So, one-- Kevin covered a little bit of those case studies here.

What’s in a name?



What’s in a name?

Business Analyst
Decision Support Analyst
Technical Analyst

Business value:

- Gather, condense, and present information
- Minimal-to-moderate technical expertise

Data Scientist

Business value:

- Extract actionable insights from existing data
- Think of as-of-yet unasked questions and answer them
- Machine learning and statistical expertise
- Assist with solution implementation

CERT | Software Engineering Institute | Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016
© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

14

**014 I wanted to kind of try to differentiate in the spirit of what we had before what data science is from some roles you may be familiar with already. There's a traditional analyst role in many companies. They think of it as a business analyst, decision support analyst, technical analyst. Depending on where you are, the names may change. But especially when I started in this position, people would say what's the distinction between you and these folks. What makes a data science person different than-- I say this actually in a loving sense because this is something which I consider myself, what's the difference between you and an Excel jockey, someone who knows how to really, really manage a spreadsheet, create some pretty complex models. What's the difference between you and that guy?

And I would argue there's actually quite a big difference, mostly in the ability to expand past the Excel tool. When you have your traditional business analyst, that person's going to be typically tasked with taking data from a variety of sources, condensing it together. I'm going to say that that data's usually fairly small. We're talking maybe a million rows of data, not much past that. Condense that into an understandable format, possibly some modeling on it. And then they're going to convey that to management, whatever their findings are. Minimal to moderate technical expertise, ability to use Excel definitely is understood. Possibly you'll be able to--know some smaller programming languages, maybe some solver, depending on where you are. If you get into the financial world, it definitely gets more complicated. And the models you have in the financial field can get quite complicated. But even within that, they're typically still within tools such as Excel and such.

I would argue that if you have a data scientist as opposed to the business analyst, what you're getting is the ability to handle data far beyond what a business analyst could handle. And even more so, you would expect more value from that individual. So, you're going to have-- you're going to be expecting them to not only take the data and understand how to manipulate it and find something from it but on their own start coming up with value added ideas. Rather than saying find me this report, or get me this particular bit of information, they should be coming up with their own ideas, think of as of yet unasked questions.

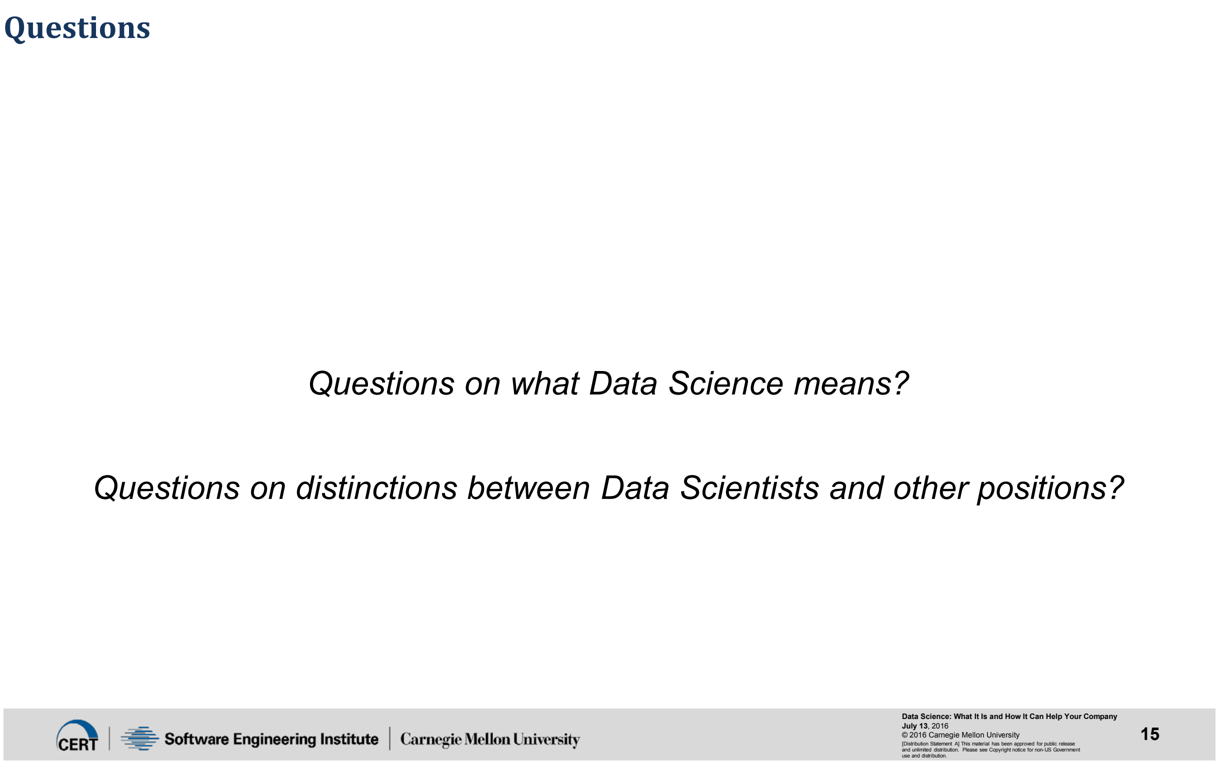
This goes without saying all the machine learning and statistical expertise is probably going to be part of a data scientist's training, whereas a business analyst may not have that nearly as much. Business analysts and data scientists both would probably assist in implementation of solutions that are being identified. But a data scientist may be able to more so just because they have more expertise, more familiarity with the topic at hand, with the statistics, with machine learning, hopefully with whatever it is they're bringing to the table, even with the data handling, than you would expect from a traditional business analyst.

Presenter: So, I would just add to that that these are the things that are implied in those titles. But that's not to say that you couldn't have somebody who went to school to be a business analyst and has all of these skills. It's just sort of generically what's implied by those titles, I think.

Presenter: Definitely. Definitely. And I have to admit, I've seen people who have the title business analyst who have skills far beyond what I would expect. I see people with data science who definitely don't have some skills that I would consider required for the field. So, it goes both ways. Obviously, it's a continuum. But just to give you a sense to frame it, when you're trying to look at a resume, you're trying to look at a position you're posting, here's what you may expect from the two different positions.

Presenter: Right.

Presenter: Cool.



Questions

Questions on what Data Science means?

Questions on distinctions between Data Scientists and other positions?

CERT
Software Engineering Institute
Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016

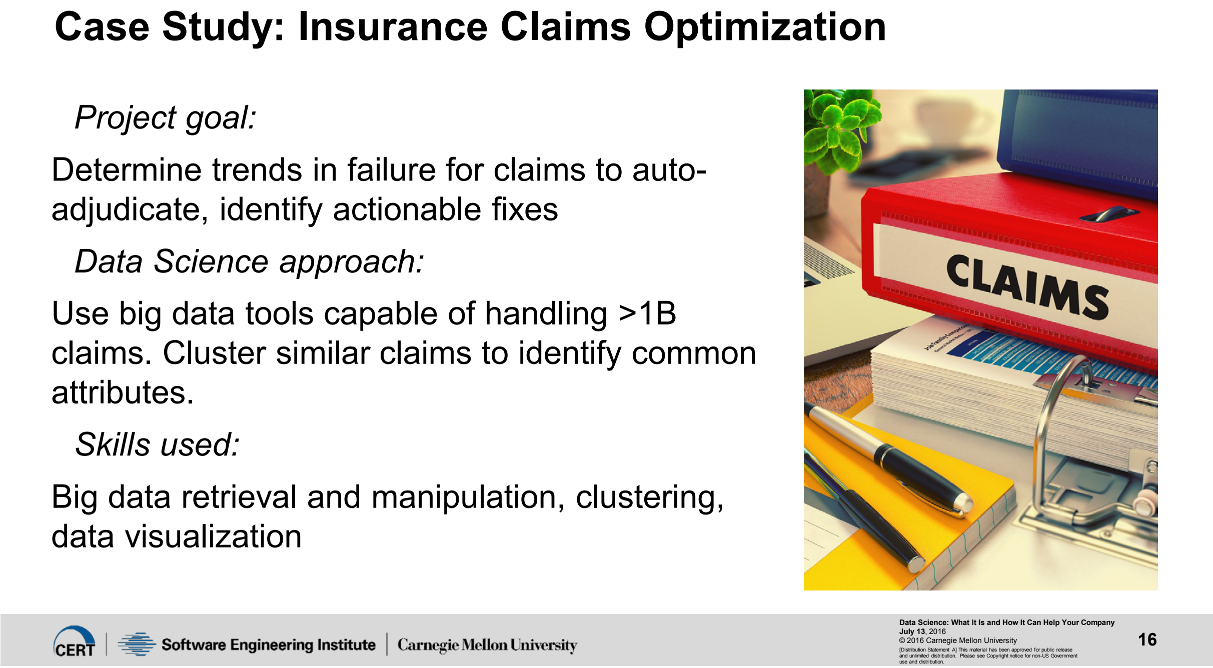
15

**015 I did want to put up here just quickly, if there's any more questions on what data science means, questions on distinctions between data scientists, other types of positions, part of my goal here is that if you're going to be looking for a data scientist, you want to bring on data science, bring it into your organization. what would you be looking for, what qualities are integral to that individual or to the field. So, I want to make sure you have a solid understanding of that.

Presenter: So, we got into that a little bit earlier. So, but there are lots of other questions on other topics, which we'll let you get your next case. Out of your next pause, we'll address some of those questions I know we did address.

Presenter: Fantastic. So, we're going to move on from here actually, to another--

Case Study: Insurance Claims Optimization



Case Study: Insurance Claims Optimization

Project goal:

Determine trends in failure for claims to auto-adjudicate, identify actionable fixes

Data Science approach:

Use big data tools capable of handling >1B claims. Cluster similar claims to identify common attributes.

Skills used:

Big data retrieval and manipulation, clustering, data visualization

**016 Click-- to another case study. Again, this is insurance claims optimization. And again, the goal is just to try to emphasize how this can be used in a number of different settings. So, one of the-- we were talking in the context before of big data. We said it's kind of a continuum. So, my first exposure to what I would consider to be big data was working with claims volume. So, it's not unusual for large insurance companies to have billions of claims come through. And if you're looking at the specialty, each individual line item on a claim, that number can go even larger because one hospital stay can have easily two hundred items that are associated with it. So, part of the field of claims processing is making sure that all goes through the system really smoothly.

It took me some time to explain this or to even try to convince people that this is true. When you have such a large complicated system where you're trying to process literally billions of claims, and all of that processing is going through an automated system, you're trying to keep as few people as you can involved as possible because every time someone touches it A, it's slowing it down. And B, it's expensive because a human's looking at it rather than the automated system. There's going to be places where things get stuck. There's going to be areas of difficulty, hot spots.

And people always want to try to find where those are. Data science is a really useful tool to try to understand where are the problems, and based on where those problems are, how can we fix it. What types of fixes are needed? Do we need system fixes? Do we need people? What's causing the issue? So, again in this area, we're able to look at a whole bunch of these different claims. We're able to cluster them together to see these claims have similar parameters. And they're all getting stuck for some reason at this phase of processing. These ones don't. These are going through smoothly. And just try to understand what does the entire system look like. I think looking at that as an individual, it's way too much data to try to understand. But looking at it with the computer, you can kind of visualize the whole system.

Presenter: So, I don't really have any experience in this industry. If you build one of these models that works for some particular insurance company, could you pick up that model and deploy at another insurance company to tell you which claims are going to be problematic? Or do you need a data scientist to build a completely new model?

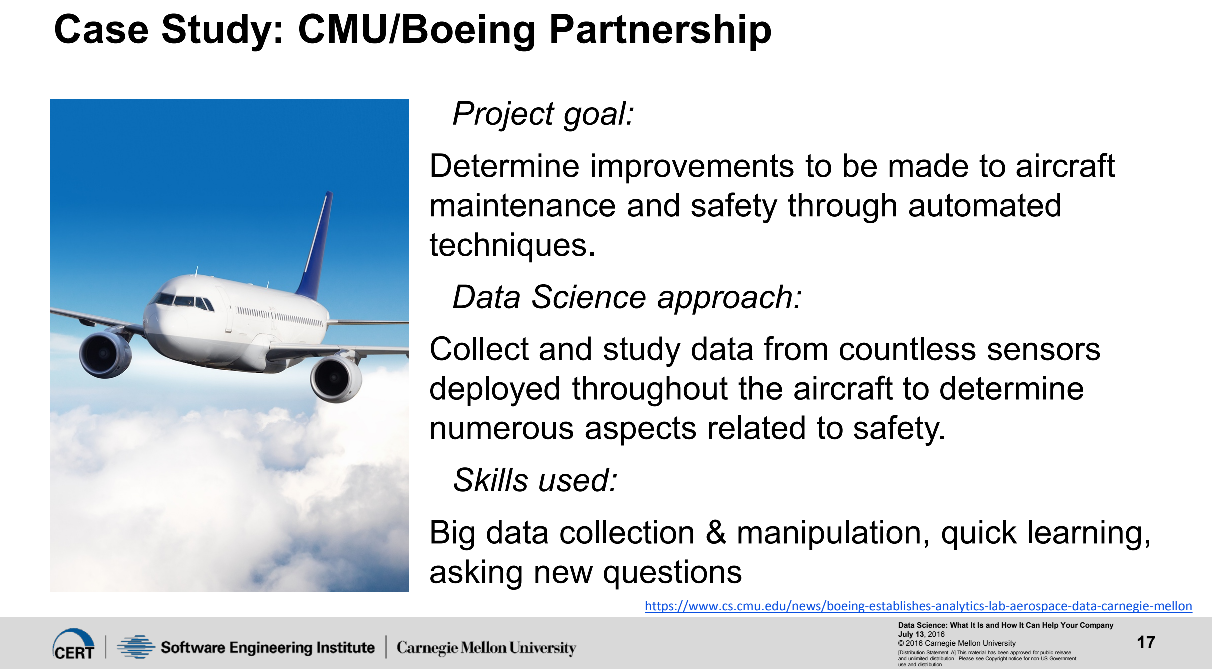
Presenter: Right. And that's actually-- the question you just asked can be broadly applied basically to any field. So, if I have a data science concept, and I apply it here, can I apply it anywhere else in the same domain? And the answer typically is yes with modifications. So, I think it's not uncommon to see companies that will release products that will say, "This is the data science solution. And we're going to help you manage all of your X." I think network analysis is a very fine area. There are many tools that will purport to analyze your network for you. And they can do all sorts of statistical work. But it just so happens that every company has their own processes. Every company has their own nuances. It could be that organizational structure is different. It could be who reports to whom. But based on all that, every solution you get is going to have to require some customization.

Presenter: So, you might have the same model, but you need to train it on that company's data just like in any other domain.

Presenter: Exactly, or even it could be that the model might even need some tweaks itself. Depending-- and many companies they're going to follow a broadly a very similar process. But it could be that some-- do something totally different. And without getting into specifics, this company processes it's claims by doing this particular step first. And this one doesn't do that check. Or this one has a totally different set of checks or does things out of order. All that stuff will affect the way the model works. And these things are not-- we'll get a little bit later as to what machine learning works and what the computer's learning. But suffice to say, it's not exactly you can build it once, and you've solved all the problems. It usually needs a little bit of customization.

So, actually we can use that as a segue to get into--

Case Study: CMU/Boeing Partnership



Case Study: CMU/Boeing Partnership

Project goal:

Determine improvements to be made to aircraft maintenance and safety through automated techniques.

Data Science approach:

Collect and study data from countless sensors deployed throughout the aircraft to determine numerous aspects related to safety.

Skills used:

Big data collection & manipulation, quick learning, asking new questions

https://www.cs.cmu.edu/news/boeing-establishes-analytics-lab-aerospace-data-carnegie-mellon

CERT | Software Engineering Institute | Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016
© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

17

**017 A little bit about-- never mind. First, we're going to get into a case study on Boeing. Go ahead.

Presenter: Okay, so Carnegie Mellon and Boeing have a partnership to do data analysis on aircraft data. So, one of the things that you'd like to do if you're an airline or an aircraft manufacturer is minimize the amount of downtime for your aircraft. And the one way to do that is to know ahead of time that a part is going to fail because then you can address that issue at your own sort of leisure as opposed to right before the flight is going to depart and you've got a bunch of angry passengers.

So, something that people might not realize is that aircraft have a lot of sensors on them. And there's a lot of data that's collected. And so, in this body of work, what Carnegie Mellon is helping Boeing do is sort of develop methods and techniques to take all of that sensor data and make predictions about when certain parts are going to fail. And that way, airlines can use that information to keep their planes flying better.

Presenter: Right, so I could be a little naive. I kind of picture an airline as just an enormous flying bus.

Presenter: No, no it's really like--

Presenter: Is there really that much data?

Presenter: Yeah, I mean modern-- it's really surprising. Modern aircraft, I saw a statistic that the Airbus 8350 I think has-- the current version has sixty-five hundred sensors. And this is surprising, they generate two and a half terabytes of data a day, just one aircraft. And that, I think this article said that that's going to triple in the next version of that aircraft. So, it really does become a big data problem.

Presenter: So, what kind of stuff are we storing here?

Presenter: So, it's like temperature, humidity, stress on parts. They have fiber optic sensors that can record information about stress, so a lot of things like that, environmental factors.

Presenter: Okay, and I guess that'll all get used eventually to figure out what part's going to wear out, or which parts are under too much stress, which parts have-- I don't even--

Presenter: Right, and I guess there's also-- there's data about sort of internal network of the aircraft that all of the stuff going over the wire internally.

Presenter: Right. That's pretty-- and I guess we can even use this-- there's similar studies-- again, this is stuff which is going in the field of data science because people see this in pop press regarding all the moving cars that are out there and automated car driving and stuff like that. I don't think that's quite in the same field. But again, there's enormous amounts of sensor data being collected to feed these machine learning algorithms.

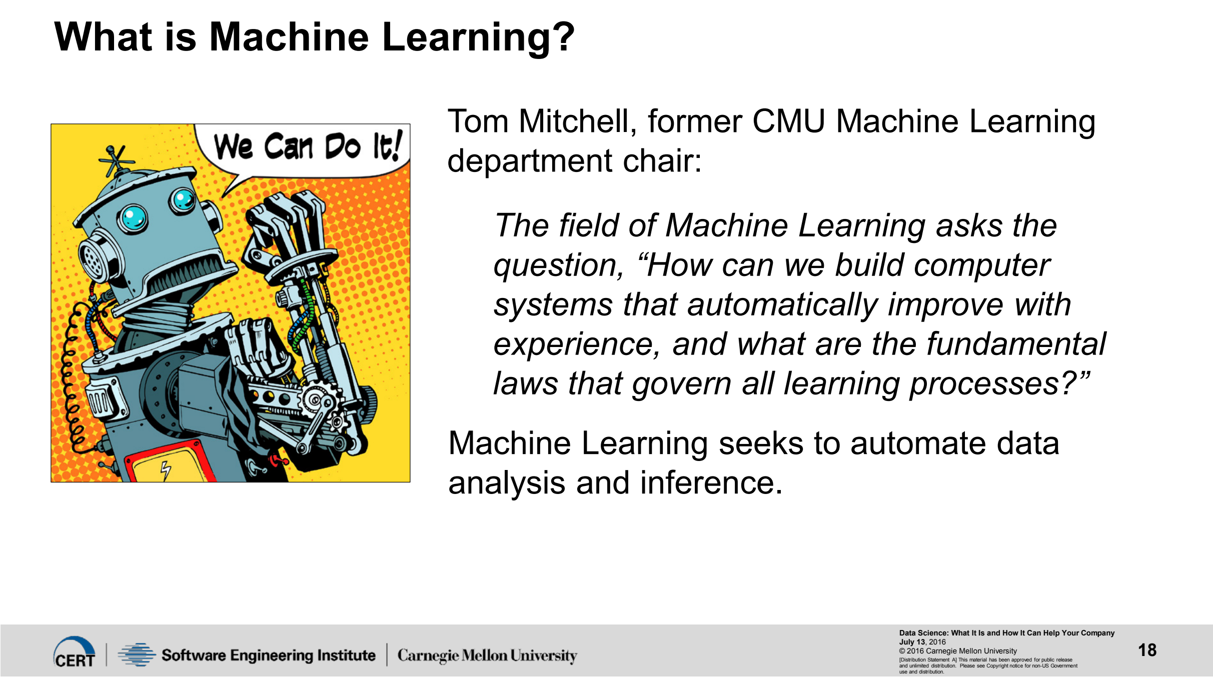
Presenter: Sure. Bill Scherlis who's a professor in the school of computer science at CMU gives this-- I've heard him give this talk where one of the things that he said was Ford is now a software company. I think it was a quote from a Time magazine article or something that there's so much software and so many electronics and sensors in all of these systems, cars, aircraft. And that's all producing data that we can be doing things with.

Presenter: It's fun-- yeah, and then I remember when I was working back at PNC Bank, one of the things which I was fighting a battle to try to convince people is that the bank is actually a data company because all they're doing is taking data, taking customer information trying to figure out who should we lend to, who's likely to have more risk, who's likely to have more fraud, and treating-- I think many more companies these days have much more value in their data than they probably expect.

Presenter: Yeah.

Presenter: Cool.

What is Machine Learning?



What is Machine Learning?

Tom Mitchell, former CMU Machine Learning department chair:

> *The field of Machine Learning asks the question, "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"*

Machine Learning seeks to automate data analysis and inference.

CERT | Software Engineering Institute | Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016
© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

18

**018 So, moving on from there, I think it's worth just doing a brief intro into what is machine learning. I think this is probably a little bit too difficult of a topic to do anything more than a cursory coverage of. But I figured I would use one of these quotes here from one of the former machine learning chairs here at CMU, Dr. Tom

Mitchell. And in his words, "The field of machine learning asks the question, 'How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?'" And there's two parts to that. And I think he intentionally separated that out. You really want to take the-- you're trying to build a computer system that understands how to learn. And in order to do that, there's almost a philosophical question you have to ask about what does it mean to learn. And trying to answer that can get a little bit hairy.

Presenter: Yeah, yeah definitely.
And the kind of learning that we usually talk about in machine learning, I think once you learn-- once you kind of understand how it works, it loses a little bit of its magic. But you understand how it's practical. I mean a lot of it is really just building statistical models, and using data to fit those models, and then making decisions based on the data that you fit to that statistical model.

So, the model itself, the thing that you're-- that's learning is specific. It's for a specific task. It's not kind of a general learning in the kind of science fiction AI sense or the Terminator Skynet thing. It's-- there's a specific task even for very complex systems like a self-driving car. So, I think sometimes, it can be a little surprising how non-general this kind of learning is.

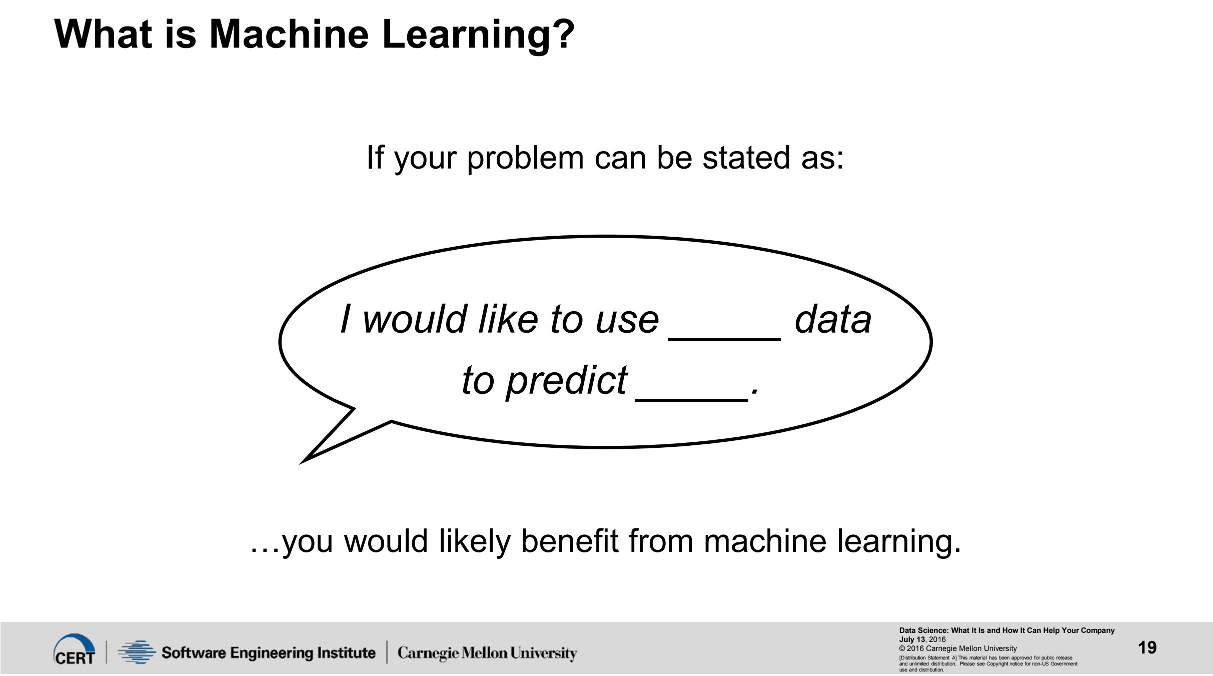
Presenter: Right, I think a lot of people here are probably familiar with the fact that you can go to Google and do a search for bird, and it'll give you birds, images of birds. But if you ask that same system, and it looks to you the same because they have one interface. But if you had asked the bird system, or the image recognition system, to recommend you how to make a sandwich, it couldn't do that because each system is trained independently. They know their own trick, so to speak.

Presenter: Right, but if I'd given it a bunch of examples of pictures of tigers instead of pictures of birds, it could learn how to identify tigers. It's specific to identifying things in photos.

Presenter: Exactly, they're general within their specific area, but you're not-- it's not like a-- one of the things I've heard someone say a while ago is trying to preach the intelligence of a two-year-old human. And even that is very difficult. You can ask a human child quite a few questions that would stump the most advanced AI that we have nowadays. So, definitely, I think that's an interesting point.

One thing which may of interest to you guys listening is understanding so now that we have this kind of a definition, so to speak, of machine learning--

What is Machine Learning?



What is Machine Learning?

If your problem can be stated as:

I would like to use _____ data to predict _____.

…you would likely benefit from machine learning.

CERT | Software Engineering Institute | Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016
© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

19

**019 When do you think that-- when would you be able to actually apply that to problems you have. And I think this is well stated. Brian, I think you can--

Presenter: Yeah, so I think one good heuristic for knowing when you have a machine learning problem, and it's not true one hundred percent of the time, but it's often true that if you can state your problem in the form of I would like to use some specific piece of data to predict some specific outcomes, there's a pretty good chance that's a problem that you could address with machine learning. So, I'd like to predict-- I'd like to take historical stock values and predict tomorrow's returns. Or I'd like to take customers' shopping history and predict whether they'll buy specific items next time they visit my website. These are kind of-- fit very well into the mold of machine learning.

Presenter: Right, and I think that some people in the audience might-- you have similar questions, but you can phrase in this context. So, you may be thinking how do I secure my network. Well, if instead of asking how do I secure my network, you say how do I use the net flow data that I have right now to secure my network or to predict attacks on my network, all of a sudden, you've taken the problem that you have and framed it as data science.

Presenter: Right, or I'd like to take a binary piece of software and predict whether it's malicious. There's a machine learning problem.

Presenter: Exactly. And by framing the questions in this sense, you're able to take what seems to be an almost impossible question, but then suddenly frame it in a way that not only is it very possible, but it can even be automated using an intelligent machine learning algorithm.

Presenter: Right, we should point out that this isn't always true. You can't say I'd like to take the ambient temperature and predict whether this piece of software is malicious.

Presenter: Right, and again, that's part of where the value of having a data scientist comes in. I think I remember, there's a joke that I heard quite a while back that if I could take a-- if I want to ask a computer the question nowadays of what's-- is this picture a picture of a bird or so, that's totally possible. But I want to tell it to predict who will be the next president, that's just not even-- there's no way to do it. It's not possible whatsoever. So, there's always questions you can and can't ask. And based on that, it's worth being able to have someone who knows how to answer those questions.

Presenter: Right. Although, even that one, if you wanted to say I want to take all polling data from the last six months and predict who will be president, that might be possible.

Presenter: You might have something going there.

Presenter: Nate Silverman might have something to say about that.

Presenter: You want to break for questions?

Presenter: We do have lots of questions coming in. So, one from Ravi asking, "What are the different forecasting which can be done on production support/application incidents based on historical data?"

Presenter: I'm assuming this is in the context of network security?

Presenter: Yes. What
are the different forecasting which can be done on production

support/application incidents based on historical data?

Presenter: Forecasting that can be done on production/support incidents. So, in other words, I'm interpreting this-- and feel free to put a comment in if it's not right. But in other words, based on support request that we're getting, what kind of stuff can we expect to get in the future.

Presenter: Sure, okay so, definitely one thing you can do is if you've got systems, and you've got support requests coming in, you could predict whether the volume is going to continue to increase for those events. You could kind of I guess make predictions perhaps about what parts of the system are the root cause of this problem by training your system on prior data. So, you might learn that when certain problems arise, they most often apply to some-- arise from some root cause. Those are a couple things that come to mind immediately. Hopefully, we interpreted the question correctly.

Presenter: Yeah, I'm sure somebody will chime in if not. What about next one? As we know, correlation does not imply causation. How does data science hold these as unrelated, yet forecast trends or identify root causes of past performance?

Presenter: I can take that. So, there's actually a very interesting field I'm becoming familiar with more because some products we're doing, that is actually detected to causal matters. So, correlation does not imply causation. That is definitely true. And the fact that we call ourselves data scientists doesn't mean we get to change that.

But there are actually statistical techniques, and I'll call them statistical techniques even though some might argue on that, that will allow you to construct causal maps. So, when this happens, this other thing might happen. And I don't think it's too much of a stretch to say that. We know this is true, right? Before it rains, there's usually going to be a rain cloud. And there's not really much argument about that. There may be cases where it's not. But that's-- you can get other causation maps for that.

So, when someone's trying to predict a causal model-- yeah, when someone's trying to predict something as causing something else, you can frequently construct a causal model for it.

Presenter: Kind of based on the temporal relationship of the events, right?

Presenter: True, true. It's also worth stating, correlation does not imply causation is actually-- and I hope this not sound like a cop out. It's actually a philosophical concept. Practically speaking, if correlation works for the purposes of actually selling a product, or actually getting eyeballs on your webpage, or reducing the number of problems you have with your Boeing airplane, that may be actually good enough for a practical purpose.

Presenter: Yeah, I mean there's definitely a big internal debate in machine learning or maybe in AI in general about the relative value of these sort of empirical statistical methods that maybe don't really care about what the underlying mechanism is versus people who would really like to understand to be able to model what is the generative process that causes these things to happen. And with the causal stuff, I guess the sort of traditional physical sciences way of dealing with that, or even social sciences, is through controlled experiments. And we don't, as data scientists, always have the opportunity to run controlled experiments.

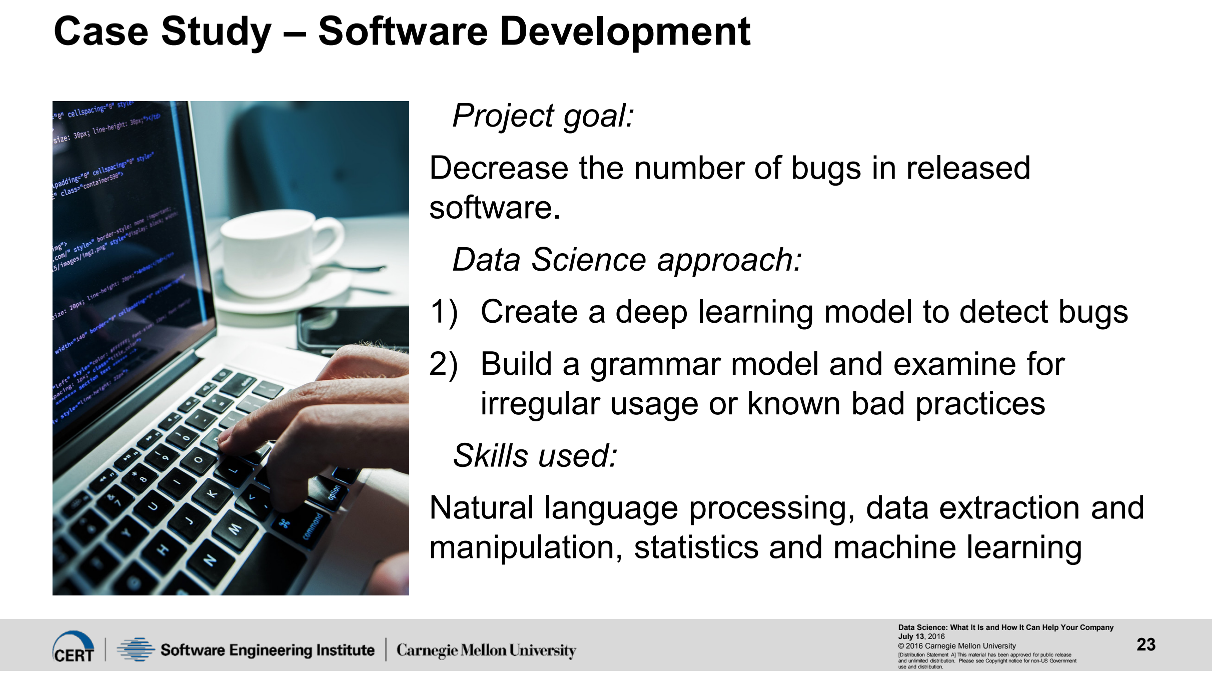
A lot of people do. For example, search engines routinely will show different results to different people as part of a controlled experiment to try to understand what their behaviors are going to be and whether action A causes reaction B. But your point is great that, in general, sometimes correlation is good enough.

Presenter: Right, right, we're pretty happy with that.

Presenter: So, I just wanted to say we know it's 2:31 Eastern time. Due to the late start we got due to our network connection issues, we started about ten minutes late. So, we're going to finish up those ten minutes with the next couple case studies, maybe a couple more questions. But we certainly understand if you have to go at 2:30. So, thank you for attending. Elli, go on to our next case link.

Presenter: Cool. Yeah, thank you very much. I think I'm actually going to jump to the

Case Study – Software Development



Case Study – Software Development

Project goal:

Decrease the number of bugs in released software.

Data Science approach:

1) Create a deep learning model to detect bugs
2) Build a grammar model and examine for irregular usage or known bad practices

Skills used:

Natural language processing, data extraction and manipulation, statistics and machine learning

CERT
Software Engineering Institute
Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016
© 2016 Carnegie Mellon University
[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

23

**023 Software development one given some of the nature of some of the questions that we had before. And I think this one is actually relevant to the first question we had about support tickets. So, you want to take this one, Brian?

Presenter: Yeah, so one thing that you can do with these methods is build models that try to detect bugs in software. And I'm not talking about the sort of-- I know we have a lot of engineers. So, I'm not talking about the sort of syntax errors that a compiler is going to warn you about. But this is more looking at the grammars that make up programming languages. Just like natural language, a program language has a grammar and an order that the terms appear in. And then feeding those, kind of learning what the normal practice is for laying out those grammars and using machine learning models to detect when certain sequences of code indicate a possible bug in the code.

Presenter: Right, and I guess, there's-- kind of with the English language, there's sort of sentence structure where you have nouns, subjects, verbs and such like that. But then once you step a level above that, you actually have the meaning of the sentence. And you can have a sentence which is syntactically correct, but the meaning is either incorrect, or it doesn't make sense, or is probably wrong.

Presenter: Right, so in this stuff, you're really trying to say can I model not just the syntax, but also the semantics of what's happening here. And then on top of those semantics, try to understand are these semantics problematic. It's a very difficult problem.

Presenter: Right, and to that point, I don't think we're actually seeing too much of this practice right now.

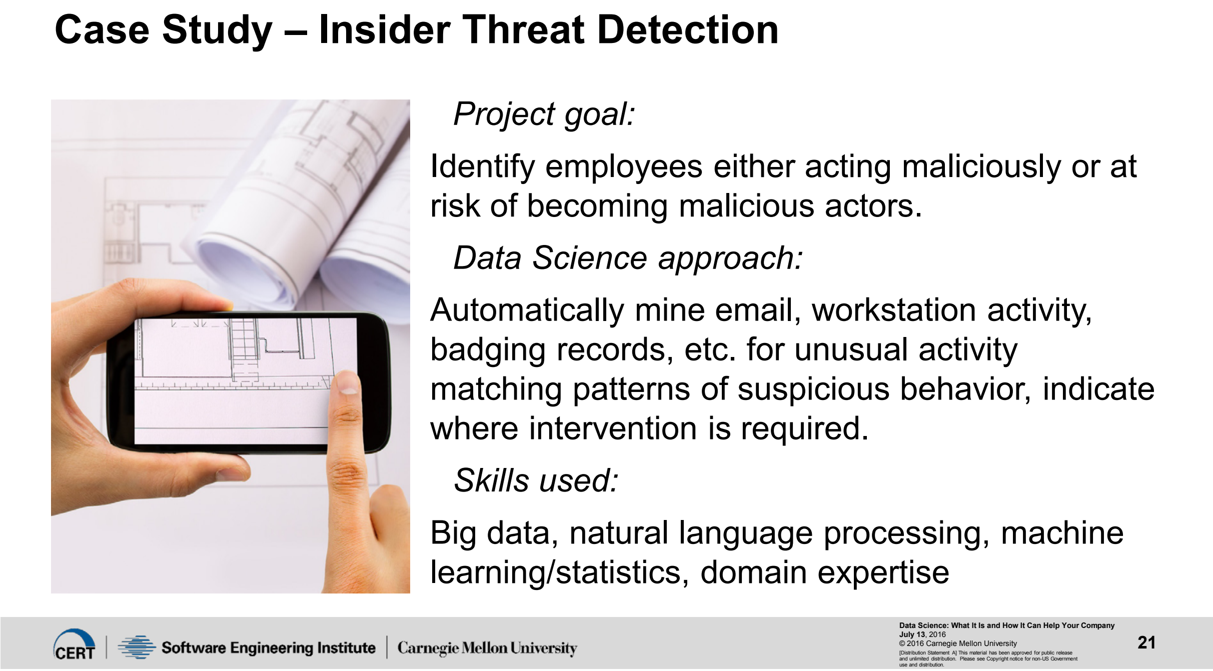
Presenter: Right, I mean this sort of thing is pretty much on the research side. There's kind of-- there are methods that try to do these things that don't involve machine learning necessarily that are used a little more in practice. I mean not in your day to day IDE, but for certain high value systems, you might use code checkers and things like that. Some of this is trickling down I think into end user tools. But for this type of stuff, this is very bleeding edge.

Presenter: Right, right and the field of machine learning is kind of still, relative to many other fields, is pretty young itself. So, this is kind of just an example of where there's definitely still a lot of work to be done to bring it down.

Presenter: Yeah.

Presenter: I think there was a lot of interest before also in the topic of insider threat detection. So, it's definitely worth--

Case Study – Insider Threat Detection



Case Study – Insider Threat Detection

Project goal:

Identify employees either acting maliciously or at risk of becoming malicious actors.

Data Science approach:

Automatically mine email, workstation activity, badging records, etc. for unusual activity matching patterns of suspicious behavior, indicate where intervention is required.

Skills used:

Big data, natural language processing, machine learning/statistics, domain expertise

**021 Trying to cover that before we break.

Presenter: Sure, so we have, as you know, a group here at the CERT that focuses on insider threat detection. And you and I, of course, are not part of that group. But so, I'm just going to talk generically about this problem. But for people in the audience that don't know an insider threat, it basically refers to employees or people with sort of privileged access to a system using that privileged access to act maliciously. So, as opposed to an external sort of the generic idea of an external hacker getting into your network, this is somebody who's already inside of your network because they're an employee or a trusted partner.

And so, there's a lot of interest in many industries in identifying employees or trusted actors that might be at risk of acting maliciously or that are acting maliciously, sabotaging things or stealing intellectual property and taking them off the network. So, there's been a lot of research in this area both in sort of static signature sort of things that you can look for in the same sense that we use those in network security, but also in using methods that are more like the anomaly detection example that we talked about earlier where you monitor all sorts of data on your network, when people are sending email, what websites are getting visited by your company, and things like that and look for anomalies or people who are doing things that are unexpected.

Presenter: Right. So, I guess what kind of happens when you see something?

Presenter: Yeah, I mean that's the million-dollar question. So, that's kind of a policy question for each company and kind of a little above my pay grade. But what I've usually heard from the people who are more expert in this area is what their real goal is is not to catch somebody doing something bad, but to find the people who might be at risk of doing that even accidentally, kind of not doing this on purpose, and intervening. If you've got somebody disgruntled, maybe you address that instead of waiting until they actually do something that's a fire-able offense.

But it does bring up kind of a general question of ethics in data science, I think. For lots of these things, there are ethical considerations.

Presenter: Right, I mean I think we're seeing now, this is based on what's in the news, you see the topics of automated cars, which actually may hit a person. Or one thing which came up recently in a workshop that we were at was if you have a car that's speeding down the road, and a deer jumps out, it's all well and good to say total the car. Kill the deer and save the human. Are there any situations that you'd want to say well, maybe I can save the human and the car and the deer? Do I want to make that risk and say okay, save the wildlife? And there are people who would argue on both sides of that. And understanding all the ethics involved in this, it's definitely sticky.

As machines are actually learning, they're going to make a decision. Or at least they're going to inform a decision. So, they're trying to make sure that that's as clear cut.

Presenter: Right, and as the people building the systems, I think it's really important for all of us in this community to be working with people making policy. If we're going to codify and computer code specific decisions or risk tradeoffs, that those are aligned with where we, as an organization or a society or government, want to be. And that's not up to the individual data scientist or programmer necessarily. We want to go through proper train of thought about that.

Presenter: Most definitely, most definitely. Anything else coming in that we want to touch on?

Presenter: So, just before maybe we get to our final questions, we'll let you do a quick wrap up. I just wanted to let everybody know, before we get to that final Q and A section, a new video series from the SEI. It's called the SEI Cyber Minute. In every Cyber Minute video an SEI expert delivers a quick informative snapshot of our latest research in changing world of all things cyber. And each video is added to the SEI YouTube channel which links from our website. So, look for those.

So, let's get into some final questions from audience, Elli and Brian. A number of questions on can you recommend some good data models and tools for data analysis or science.

Presenter: That really depends on your problem.

Presenter: Yeah, I guess so that's--that kind of brings quite nicely to one of the final points we wanted to make, which was that it really depends, I guess, on what the problem is, what kind of data you have available, and what you're trying to solve. As we said before, when you're teaching your machine something, or when you're doing a machine learning solution, you're really only solving one problem at a time. So, to say what's the broad based tool that can help everyone, it's pretty difficult. And especially given that we have quite a few people in the audience who are coming from the engineering side, if you're trying to implement it, there's a certain set of tools that you want. If you're trying to buy off the shelf, it's a completely different set of tools. And if you're trying to just find talent, that's again a different discussion.

Presenter: Right, this is kind of like asking a software engineer, "What's the best programming language?" It depends on the problem you're trying to solve.

Presenter: Right, I'm afraid that's--

Presenter: But, I mean broadly speaking, some of the things that are popular are deep learning, which is usually first in neural nets, support vector machines, which were more in vogue maybe a decade ago. And as far as tools and toolkits, I mean as far as what you get when you put out a request for resumes for these kinds of things, you'll get a lot of people who do their work in R or Python mostly. But there are lots of other toolkits. There are cloud solutions now. And we probably don't want to be endorsing specific products or anything here, but--

Presenter: Understood.

Presenter: Just one last thing on that, I would definitely say feel free to-- we're going to put up our emails at the end of this, definitely feel free to reach out. And we can continue that discussion if necessary.

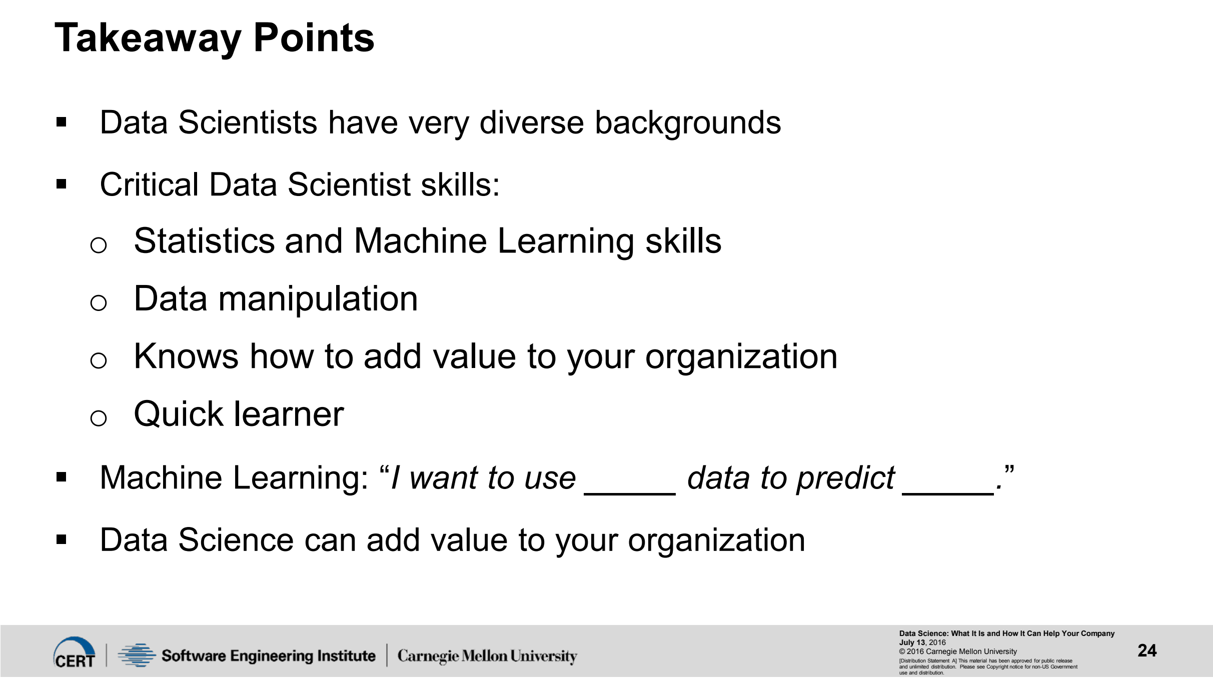
Presenter: Right, so and we just have a lot of comments coming into the chat log, which I'll make sure I get you guys just to see if there's something you may be able to follow up with or write a blog post, something that we can address some of these things that are coming in because there's just not enough time. But another question that came in is, "You mentioned data manipulation less often than other skills. Is this less important? And how close does this come to data management, database engineer?"

Presenter: I would say it's not less important. In fact, it's where you spend the majority of your time. I would say it's just maybe a little less interesting to talk about in this context. It's-- sometimes I feel a little bit like a data janitor. So, but yeah, I mean the ability to, by hook or by crook, get data into the right form-- the form that you need it in and to remove cruft from the data and clean it up and extract the actual information just as a starting point to all of these fancier techniques and models is, like I said earlier, could easily be eighty percent of your time spent.

Presenter: And I'll actually personally up that to ninety. I mean it's an absolutely critical skill. And the fact that we don't talk about it as much is simply because it's just, I'll say, less sexy than the rest of it. It's plumbing. But it's absolutely critical to do.

Presenter: So, we'll let you wrap it up with any last takeaways you want to address. And then we'll close out the--

Takeaway Points



Takeaway Points

- Data Scientists have very diverse backgrounds
- Critical Data Scientist skills:
  - Statistics and Machine Learning skills
  - Data manipulation
  - Knows how to add value to your organization
  - Quick learner
- Machine Learning: "*I want to use _____ data to predict _____.*"
- Data Science can add value to your organization

CERT
Software Engineering Institute
Carnegie Mellon University

Data Science: What It Is and How It Can Help Your Company
July 13, 2016

24

**024 Presenter: Yeah, we'll just very quickly jump to the takeaway points we have here just to highlight a couple bits. It's just data scientists have some very diverse backgrounds. And it's really worth understanding that, when you're looking for a data scientist, you're going to get a lot of diverse backgrounds in your call. There are a couple critical data science skills. And we have some of those listed here. We talked about those throughout the talk.

Machine learning, you're going to try to apply a machine learning technique, you're going to want to put that in the context of I want to use my data X to predict Y. As we said before, there are many other types of machine learning, but if you keep that in mind, you'll be pretty good. And lastly, data science can add value to your organization. So, if this something which sounds interesting to you, please feel free to get in touch with us. And that's that.

Presenter: Great, Brian, Elli, thank you very much for your presentation, excellent job. Folks, again just thank you for taking an hour with us today. We wanted to remind you, our next webinar will be August 17th. And the topic will be building and scaling a malware analysis system by Brent Fry. And lastly, upon exiting today's webinar, we ask that you fill out that survey as your feedback is always greatly appreciated. Have a great day, everyone.

SEI WEBINAR SERIES | Keeping you informed of the latest solutions

SEI WEBINAR SERIES | Keeping you informed of the latest solutions

#SEIwebinar

Software Engineering Institute | Carnegie Mellon University