Data Science

Interest in data science is at an all time high, and really has exploded in popularity in the last couple of years.
A fun way to see this is to hit up the Google Trends website.
Google Trends shows search keyword information over time.
We could see that the term ‘data science’ is massively popular, really all across the globe.

We can also see that related queries include topics like python.
And we can even see certain tool kits that we’ll be teaching in this course.
For instance, here's the trend for python pandas, which is the focus of module two and three in this course.

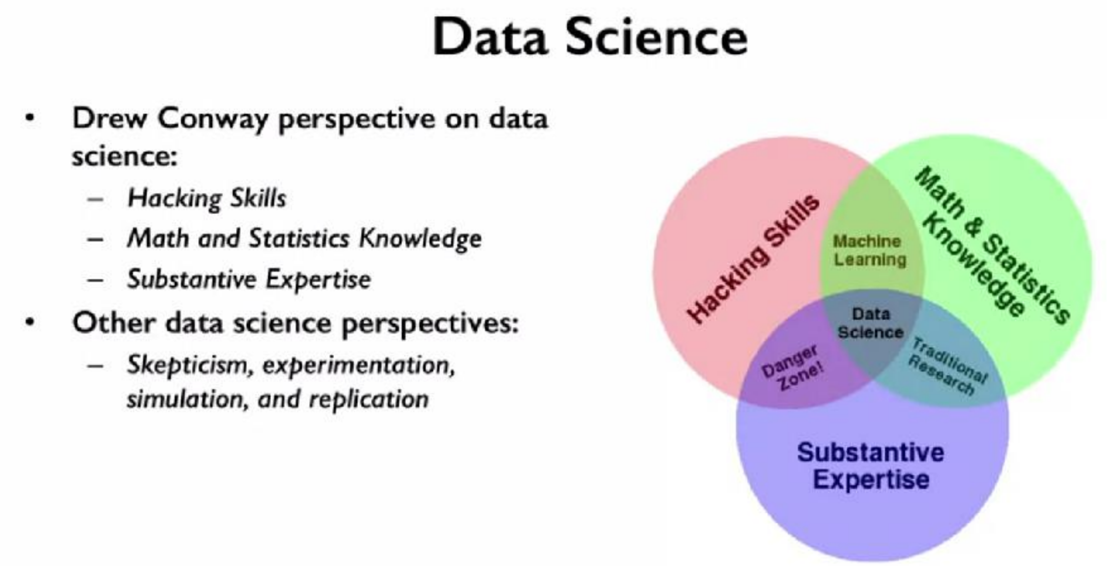
Before we jump in to a discussion of data science, I'd like to take a moment and have you reflect on what you think data science is, and why you came to this course to explore.
If someone you ran into asked you what data science was all about, what would you tell them?

The history of data science goes back a little further than 2004, which is where the Google search term history begins.
But this, at least, gives a sense for how popular the area is now.
I think the popularity of interest in the area comes from the network and data driven society we find ourselves living in.
When people think of the term data scientist, they tend to think of Google or Amazon or Facebook, places with big artificial intelligence research teams, and certainly these are some amazing companies who are doing great things with data science.

But data scientist aren’t just limited to careers with tech companies.
Almost every company is turning to data science to better understand how to build products, serve customers and leverage new opportunities.
And companies aren't the only one.

The space is used by a team which includes programmers, behavioral scientists and data scientists who aim to build technologies to support next generation teaching and learning.
The background of individuals here is very broad and includes folks with computer science and information degrees, psychology degrees, and law and business degrees.
And here they apply data driven methodologies to aid in their discovery, from statistical analysis, machine learning and text mining, to information visualization.
And this need for data driven intelligence and skills is growing in companies and organizations throughout the world.
Let's start with a look at the roots of the field of data science.

The techniques and methodologies of data science stem from the fields of computer science and statistics.
One of the most well cited diagrams describing the field comes from Drew Conway where he suggested data science is the intersection of hacking skills, math and stats knowledge, and substantial expertise.
This diagram might be a bit of an oversimplification, but I think it's a great start.



Data Science

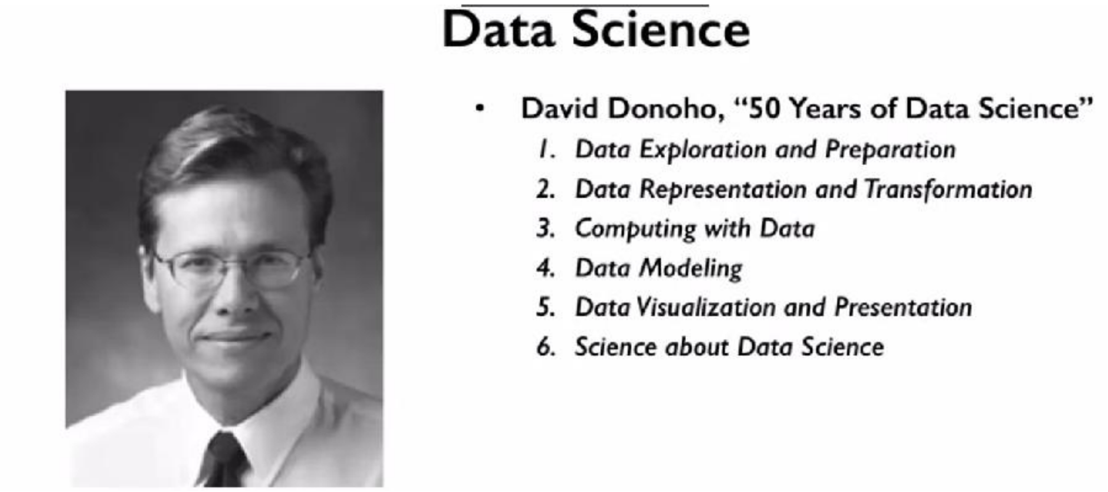
- **Drew Conway perspective on data science:**
  - *Hacking Skills*
  - *Math and Statistics Knowledge*
  - *Substantive Expertise*
- **Other data science perspectives:**
  - *Skepticism, experimentation, simulation, and replication*

Data science is definitely one of those areas where you ask ten people and get ten different answers.

One thing that I think is missing from this diagram is the underlining need for the scientific inquiry.
You don't necessarily get this for having good hacking skills or math and statistic knowledge.
A good data scientist bring skepticism, experimentation, simulation, and replication to bear on understanding a given phenomena.
Whether it be trying to predict sales at the coffee shop, cluster products in an online store, determine who's likely to win in election or mine health data from physiological sensors like Fitbits.

In this specialization we're going to touch on all of these issue broadly but really focus on skills.
In addition to the hacking skills Conway mentions, we're going to develop some communication skills in the form of charting, graphing, and related visualizations.
A good data scientist is one who can communicate their findings clearly to others and having some fundamental tools in the toolkit as useful here.

We can find a more comprehensive and academic discussion in the work of David Donoho, Professor of Statistics in Stanford.
Who's providing excellent commentary of views of the field in a paper entitled 50 years of data science.



Data Science

- David Donoho, "50 Years of Data Science"
  1. *Data Exploration and Preparation*
  2. *Data Representation and Transformation*
  3. *Computing with Data*
  4. *Data Modeling*
  5. *Data Visualization and Presentation*
  6. *Science about Data Science*

There's lots to think about with this paper, but I want to draw your attention to section 8, where he describes the field broadly as being made up of six activities.

- The first of these is Data Exploration and Preparation, this involves cleaning data and manipulating it for further analysis.
- The second is Data Representation and Transformation.

This specialization is going to look at several different forms of representing data.

Tabular structures in the first course, text data in the fourth course, and graph-based data in the last course.

These are just some of the kinds of data you have to deal with as a data scientist.

- The third activity is Computing with Data, and Donoho mentioned specifically the languages of R and Python as being fundamental here.

He goes further and he talks about pipelining and how data scientist need to be able to work with different languages for different parts on an analysis project and I think this is very true.

Unlike enterprise software projects, where you might use one language for implementing all of the functionality you need.

Modern data science projects can span many different languages and computing paradigms.

Knowing when to use the right tool for the job is an important attribute.

- The fourth activity is Data Modeling. I think data modeling is a big space but Donoho speaks specifically about predictive modeling, which we'll talk about in the third MOOC in this specialization and generative modeling.

There's some fundamental differences here, and I think the increased interest in predictive modeling is helping to fuel the current data science push.

In particular, the modern world, with massive data streams and significant computational power, gives us opportunity to rapidly experiment with making predictions.

Allowing us to innovate and evaluate new data science techniques.

- The fifth activity is Data Visualization and Presentation.

Now the word visualization brings with it a number of connotations, such as charting and graphing, as well as 3D visualizations and interactive environments.
In the second course we'll learn the basics of information visualization, but keep in mind that info viz is a whole field of its own.

- Finally, Donoho suggests there's sort of a Meta activity which he calls the science about data science.

That is understanding what works and what doesn't in data science, and building ways to leverage these discoveries.

We see examples of this as new tools and paradigms of computing are invented to change the way data science works.

I think Donoho provides a really nice overview of the area of the data science.
In his paper he also alludes to several investments, research institutions are putting into the field.
Including the University of California in Berkeley, New York University,
MIT and the University in Michigan, here, which just kicked off a $100 million data science initiative.
Data science is becoming a fundamental way of understanding the world around us.
While many institutions, are starting to offer data science masters degrees and certificate.
I think it's important to think of data science as a sort of epistemology, or a way of knowing.
Data science thinking can be useful in a variety of disciplines and careers.
It is a way of approaching problems and critical thinking skills.

But this is a skills based course, so in the next lecture we're going to jump right in and start talking about the Python programming language.