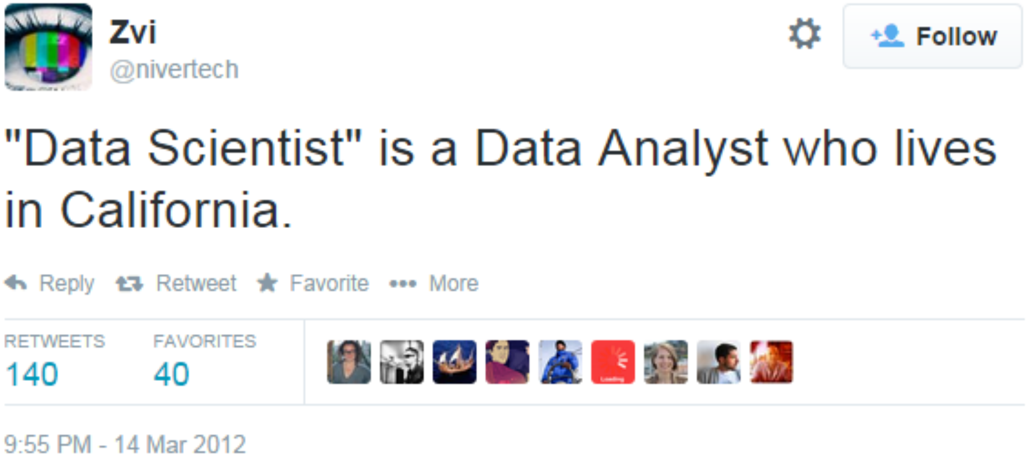
AGENDA

I. WHAT IS A DATA SCIENTIST?

II. DATA SCIENCE WORKFLOW

I. WHAT IS A DATA SCIENTIST?

WHAT IS A DATA SCIENTIST?



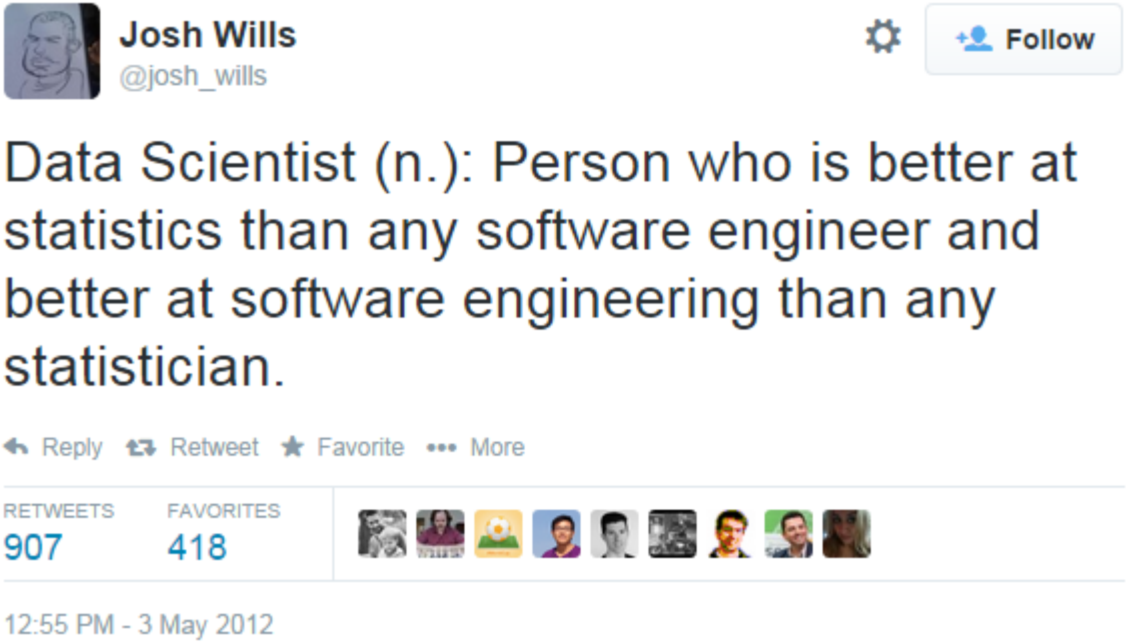
Zvi
@nivertech

"Data Scientist" is a Data Analyst who lives in California.

RETWEETS 140 FAVORITES 40

9:55 PM - 14 Mar 2012

WHAT IS A DATA SCIENTIST?



Josh Wills
@josh_wills

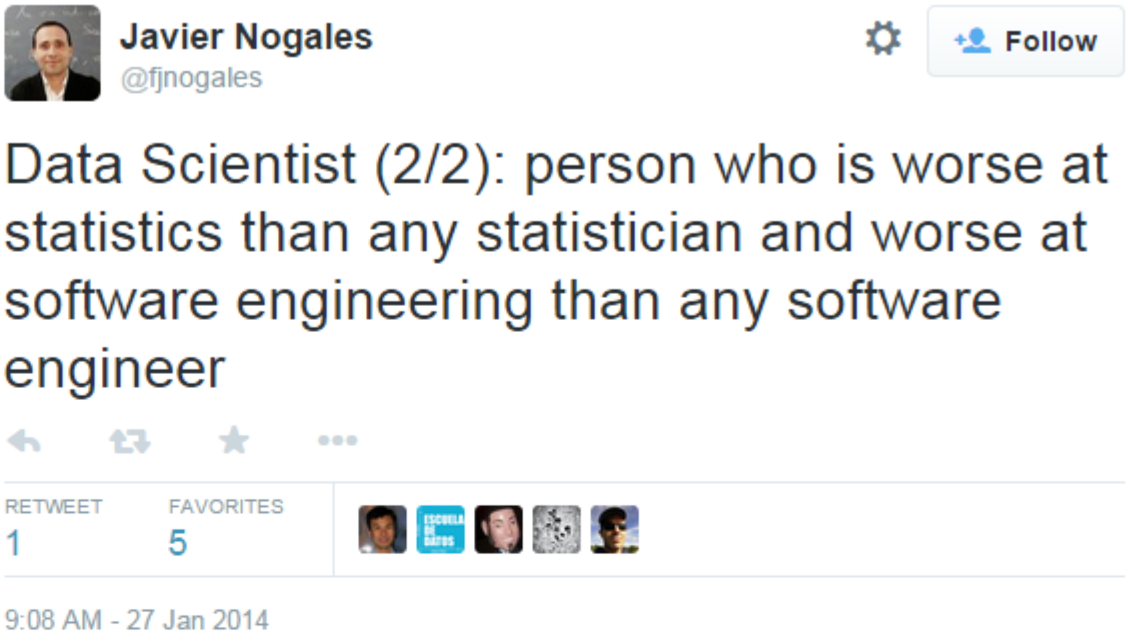
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

RETWEETS 907

FAVORITES 418

12:55 PM - 3 May 2012

WHAT IS A DATA SCIENTIST?



Javier Nogales
@fjnogales

Data Scientist (2/2): person who is worse at statistics than any statistician and worse at software engineering than any software engineer

RETWEET 1 FAVORITES 5

9:08 AM - 27 Jan 2014

WHAT IS YOUR DEFINITION?

WHAT IS A DATA SCIENTIST?

“Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways.”

Data Scientist Type A (for Analysis):

- Primarily concerned with **making sense of data** or working with it in a fairly **static** way.
- Similar to a statistician, but knows all the **practical details of working with data** that aren't taught in statistics: data cleaning, dealing with large data sets, visualization, domain knowledge, etc.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

WHAT IS A DATA SCIENTIST?

“Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways.”

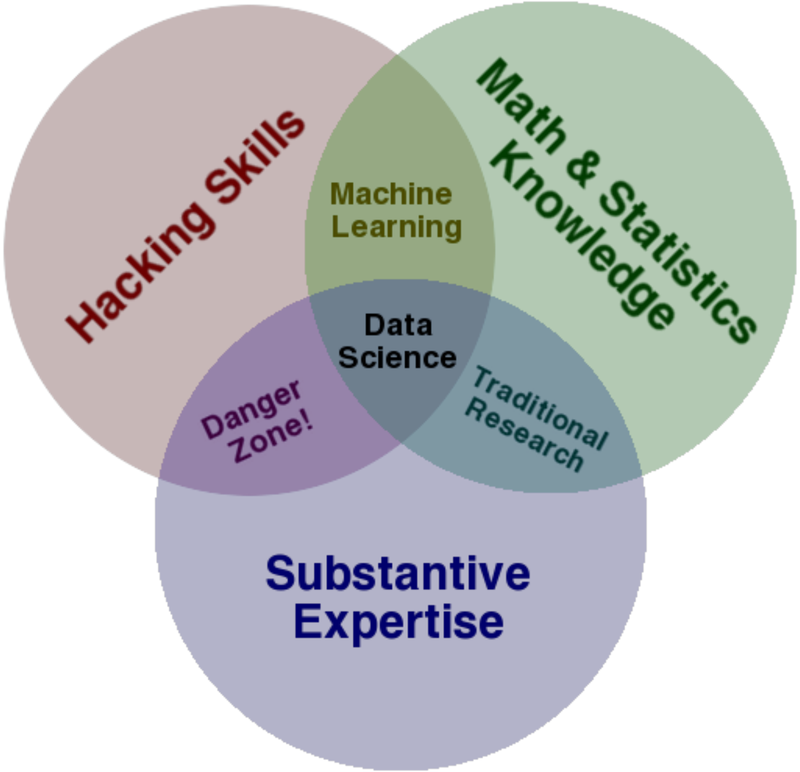
Data Scientist Type B (for Building):

- Some statistical background, but **strong coder or software engineer**.
- Primarily concerned with **using data “in production”**: building models which interact with users (by giving recommendations, for example).

Our course is focused primarily on **Type A**.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

WHAT IS A DATA SCIENTIST?



Wide variance in terms of skillsets: many job descriptions are more appropriate for a **team of data scientists!**

Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

WHAT IS A DATA SCIENTIST?

Hadley Wickham’s advice for becoming a data scientist:

Statistical knowledge
“I think you need some knowledge of specific statistical/machine learning techniques, but a deep theoretical understanding is not that important. You need to understand the strengths and weaknesses of each technique… The vast majority of data science problems can be solved by a creative assembly of off-the-shelf techniques, and don't require new theory.”

Source: https://gist.github.com/hadley/820f09ded347c62c2864

WHAT IS A DATA SCIENTIST?

Hadley Wickham’s advice for becoming a data scientist:

Programming skills

“You need to be fluent with either R or Python. There are other options, but none of them have the community that R and Python have, which means you'll need to spend a lot of time reinventing tools that already exist elsewhere.”

Source: https://gist.github.com/hadley/820f09ded347c62c2864

WHAT IS A DATA SCIENTIST?

Hadley Wickham’s advice for becoming a data scientist:

Domain knowledge

“…A data scientist should be able to contribute meaningfully to any project, even if you're not intimately familiar with the specifics. I think this means you should be generally well read… and an able communicator. A good data scientist will help the real domain experts refine and frame their questions in a helpful way. Unfortunately I don't know of any good resources for learning how to ask questions.”

Source: https://gist.github.com/hadley/820f09ded347c62c2864

WHAT IS A DATA SCIENTIST?

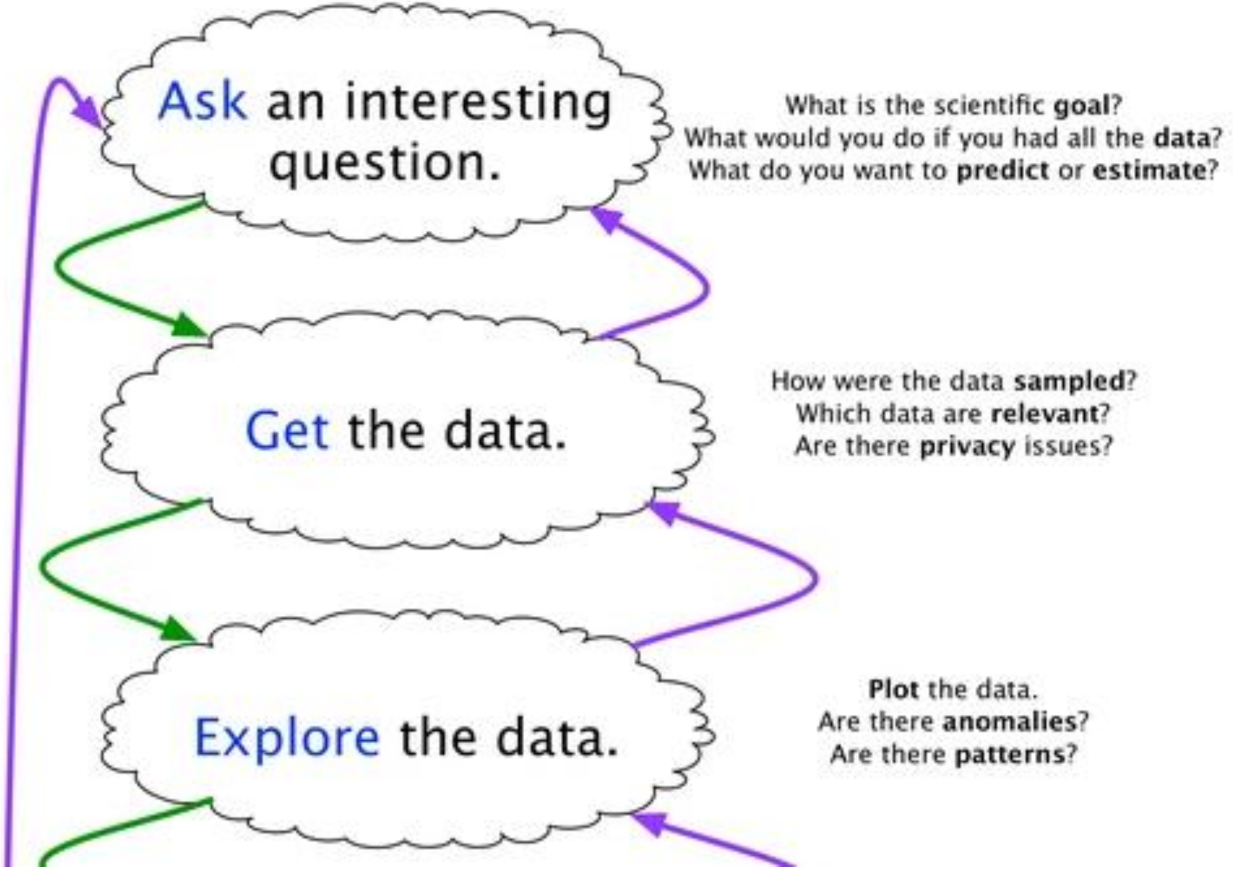
Chris Volinsky (Columbia & AT&T Labs) on “Data Mining vs. Statistics”

- Snark: Data Mining = Statistics + Marketing
- Statistics is known for: **well-defined hypotheses** used to learn about a **specifically chosen population** studied using **carefully collected data** providing inferences with **well-known properties**.
- Data mining isn’t that careful. It is: **data-driven discovery** of **models and patterns** from **massive and observational data sets**.

Source: http://www2.research.att.com/~volinsky/DataMining/Columbia2011/Slides/Topic1-DMIntro.ppt

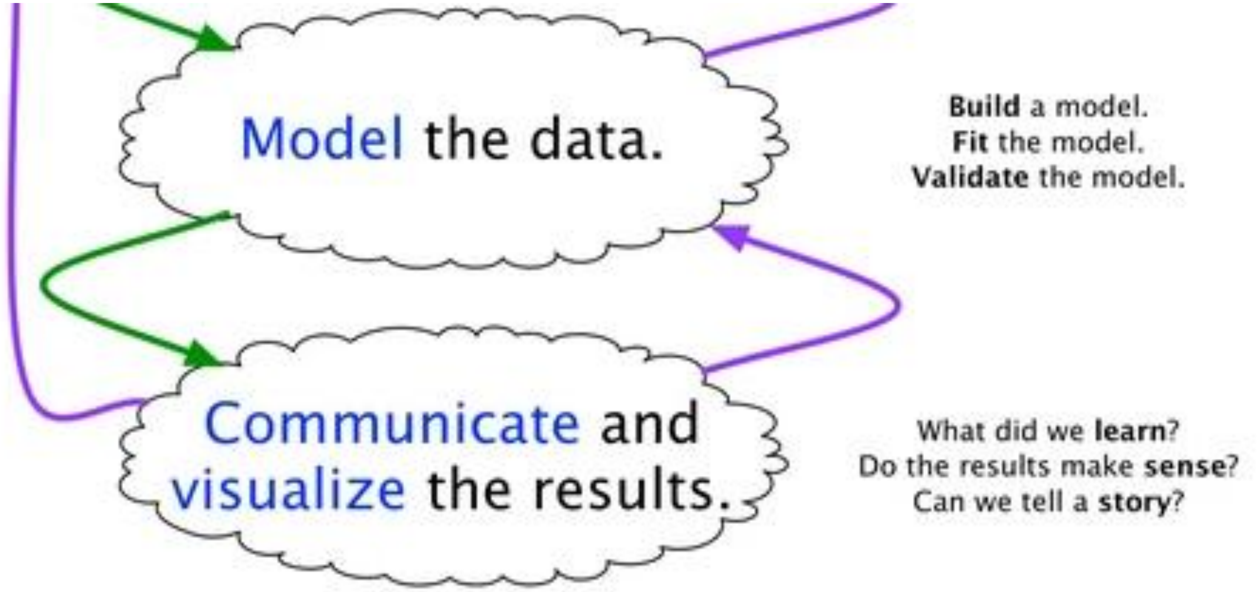
II. DATA SCIENCE WORKFLOW

THE DATA SCIENCE WORKFLOW



Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?

THE DATA SCIENCE WORKFLOW



Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Source: https://www.quora.com/What-is-the-work-flow-or-process-of-a-data-scientist-analyst-and-what-tools-do-you-use-for-this/answer/Ryan-Fox-Squire

EXAMPLE #1: PREDICTING NEONATAL INFECTION

Problem: Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick

Goal: Detect subtle patterns in the data that predicts infection before it occurs

Data: 16 vital signs such as heart rate, respiration rate, blood pressure, etc…

Impact: Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

Image: http://www.babycaretips4u.com/wp-content/uploads/2014/03/premature-baby.jpg
Case Study: http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695

EXAMPLE #2: AUTOMATING GOVERNMENT PAPER-PUSHING

Problem: Processing disability claims at the Social Security Administration is a time-intensive process, with many claims taking over 2 years to adjudicate

Goal: Automate the approval of a subset of the "simplest" disability claims

Data: Free text in the claims form

Impact: Able to fully automate 20% of the simplest claims. Rating accuracy of the algorithm is higher than the average claims examiner.

Case Study: http://dataminninglab.com/images/case-studies/ERI_Text_Mining_SSA_Claims_for_Disability_Approval.pdf