GENERAL ASSEMBLY

DATA SCIENCE

CLASS 1: INTRO TO DATA SCIENCE

AGENDA

I. WHAT IS A DATA SCIENTIST?

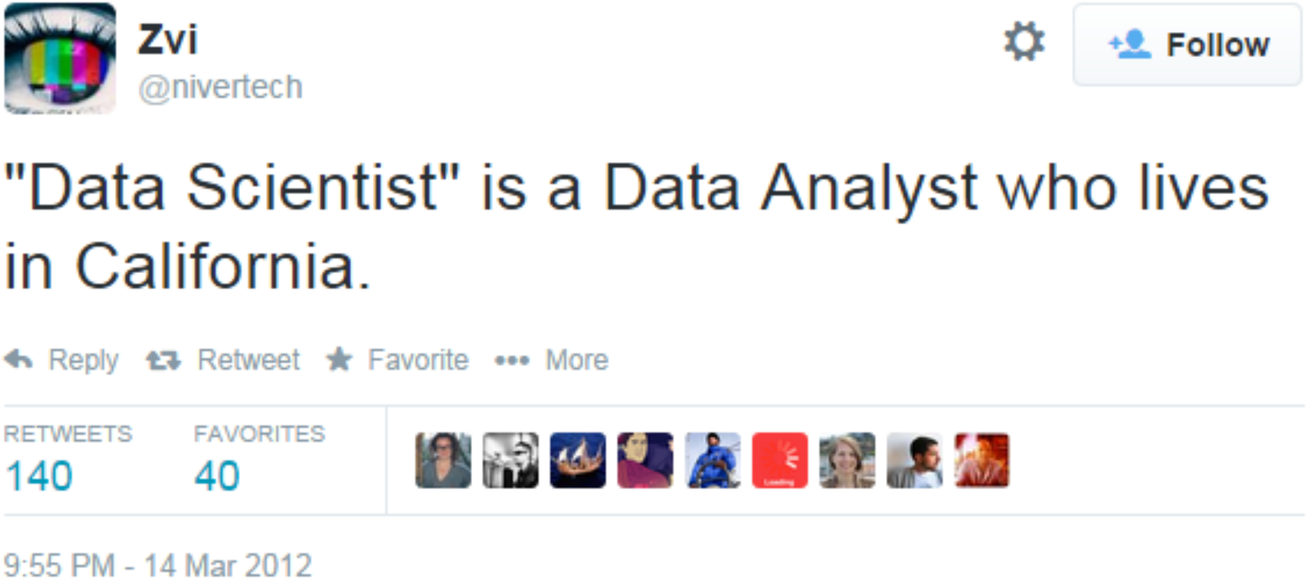
II. HOW DATA SCIENTISTS ADD VALUE

III. THE DATA MINING WORKFLOW

IV. QUALITIES OF A GOOD DATA SCIENTIST

I. WHAT IS A DATA SCIENTIST?

WHAT IS A DATA SCIENTIST?



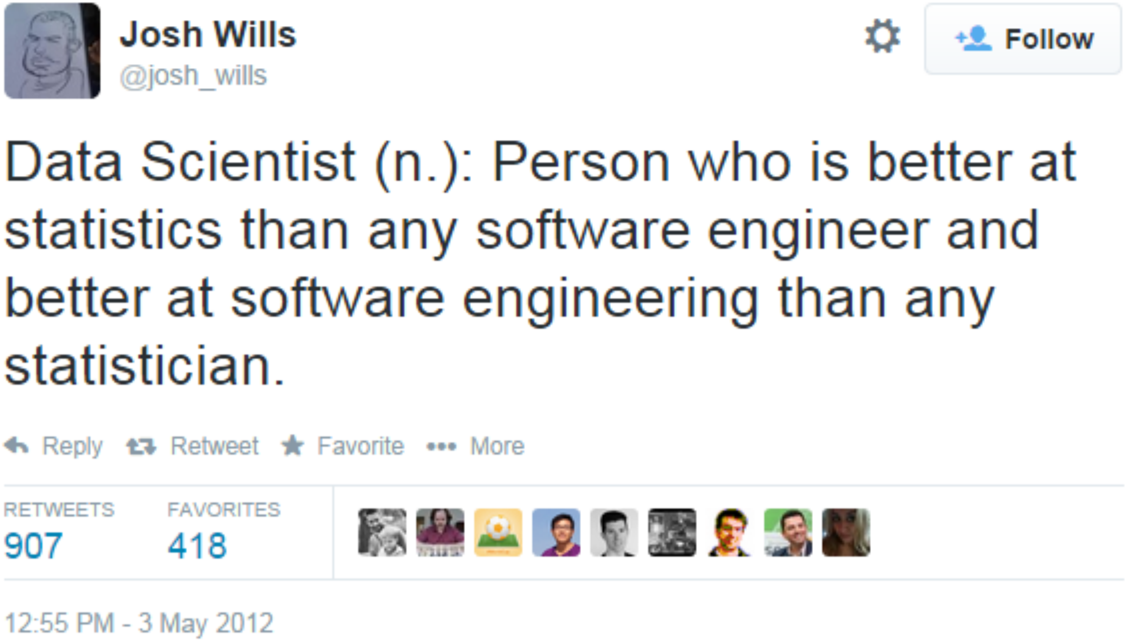
Zvi
@nivertech

"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

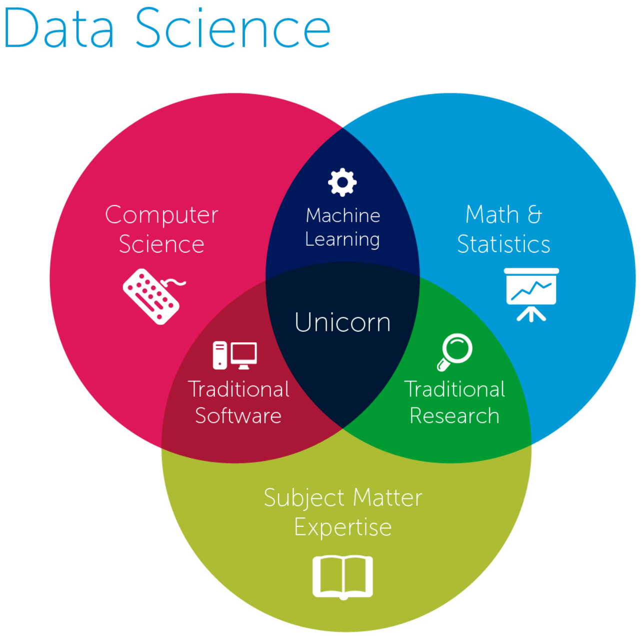
FAVORITES 418

12:55 PM - 3 May 2012

WHAT IS A DATA SCIENTIST?

- Data Scientists are currently defined more by their set of skills than they type of work they do.

- Data Science is a direct byproduct of tech companies' desire to expand the role of engineering to other parts of their business.



WHAT MAKES A DATA SCIENTIST DIFFERENT FROM A DATA ANALYST?

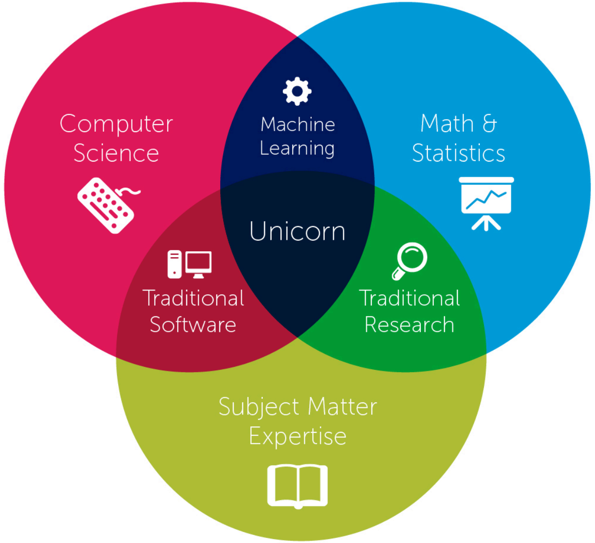
- Data Scientists typically have degrees or training in econometrics, applied math, and statistics.

- Data Scientists develop rigorous, reproducible approaches.

- Data Scientists work with tools and technologies typically reserved for software engineers.

- Data Scientists develop predictive models that automatically take action on new data, without the need for human interaction.

Data Science



I. HOW DATA SCIENTISTS ADD VALUE

HOW DATA SCIENTISTS ADD VALUE

Data mining techniques generally add value by doing one of four things:

1) Predicting the bad
2) Identifying the good
3) Automating existing processes
4) Identifying patterns in data

Data scientists can be found within many fields. let’s look at some additional examples to motivate this course.

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, and blood pressure.

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://datamininglab.com/images/case-studies/ERI_Text_Mining_SSA_Claims_for_Disability_Approval.pdf

II. THE DATA MINING WORKFLOW

THE DATA MINING WORKFLOW

0. **Define the problem / question**

I. **Identify and collect data**

II. **Explore and prepare data**

III. **Build and evaluate model**

IV. **Communicate results**

0. DEFINE THE PROBLEM / QUESTION

Can I predict infection before it occurs?

Can I predict claim approval from the start of the process?

I. IDENTIFY AND COLLECT DATA

Heart Rate, Blood Pressure, weight, etc.

Free form text on the claim form

II. EXPLORE, CLEAN AND PREPARE DATA

Aggregate data at the minute level, delete outliers

Cluster like words, throw out common phrases

III. BUILD AND EVALUATE MODELS

Compare performance of decision tree and logistic regression

Look at error rate of with Naïve Bayes classifier

IV. COMMUNICATE RESULTS

Create custom dashboard for doctors and nurses

Create automated model with output alerts

III. QUALITIES OF A GOOD DATA SCIENTIST

PROACTIVELY FINDS OPPORTUNITIES FOR WORK

CLEARLY DEFINES DATA AND ENGINEERING NEEDS, THE DELIVERED PRODUCT, AND SUCCESS METRICS

FULLY EXPLORES DATA BEFORE BUILDING THE MODEL (AND PROMISING RESULTS)

UNDERSTANDS THE PROS & CONS OF DIFFERENT TECHNIQUES

RETAINS A HEALTHY SKEPTICISM OF THE MODEL’S ACCURACY

COMMUNICATES CLEARLY WITH BOTH BUSINESS OWNERS AND SOFTWARE ENGINEERS

UNDER-PROMISES AND OVER-DELIVERS