GENERAL ASSEMBLY

DATA SCIENCE

CLASS 1: INTRO TO DATA SCIENCE

AGENDA

0. WHAT IS A DATA SCIENTIST?

I. HOW DATA SCIENTISTS ADD VALUE

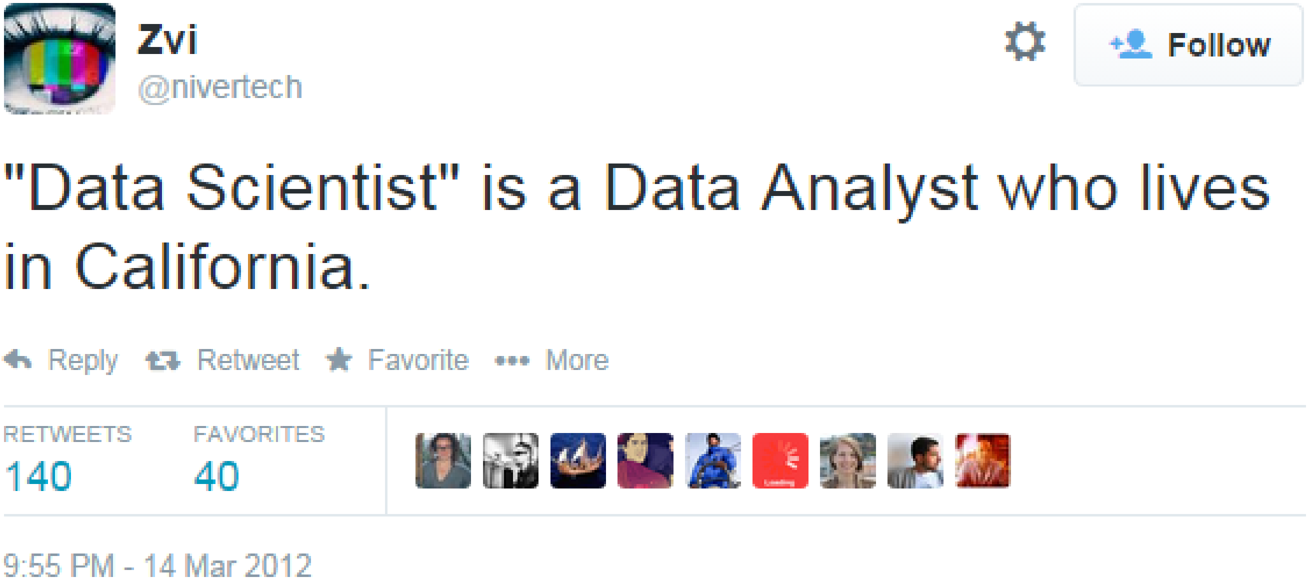
II. THE DATA SCIENCE WORKFLOW

III. QUALITIES OF A GOOD DATA SCIENTIST

0. WHAT IS A DATA SCIENTIST?

WHAT IS YOUR DEFINITION?

WHAT IS A DATA SCIENTIST?

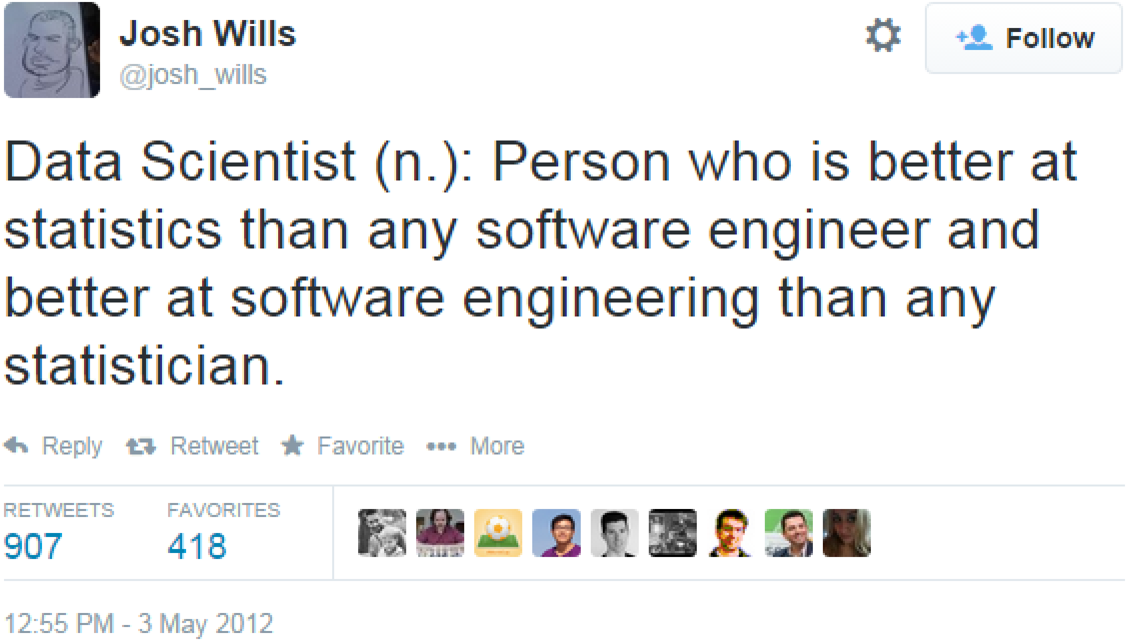


"Data Scientist" is a Data Analyst who lives in California.

RETWEETS	FAVORITES
140	40

9:55 PM - 14 Mar 2012

WHAT IS A DATA SCIENTIST?



Josh Wills
@josh_wills

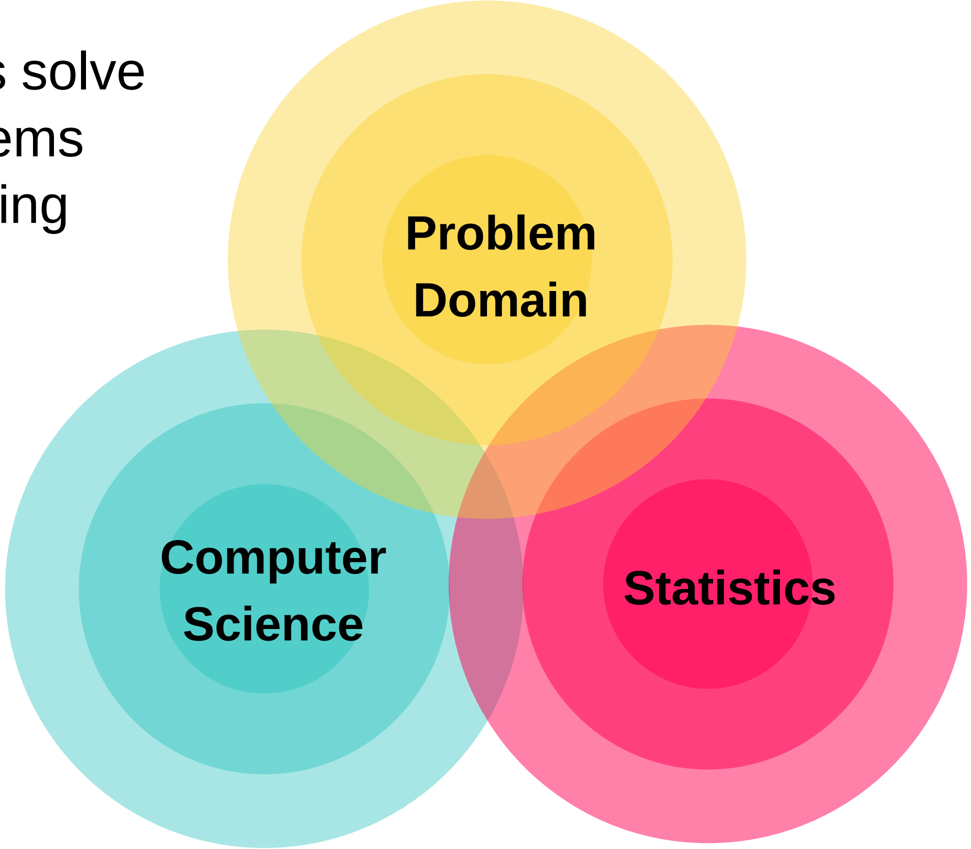
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

RETWEETS 907
FAVORITES 418

12:55 PM - 3 May 2012

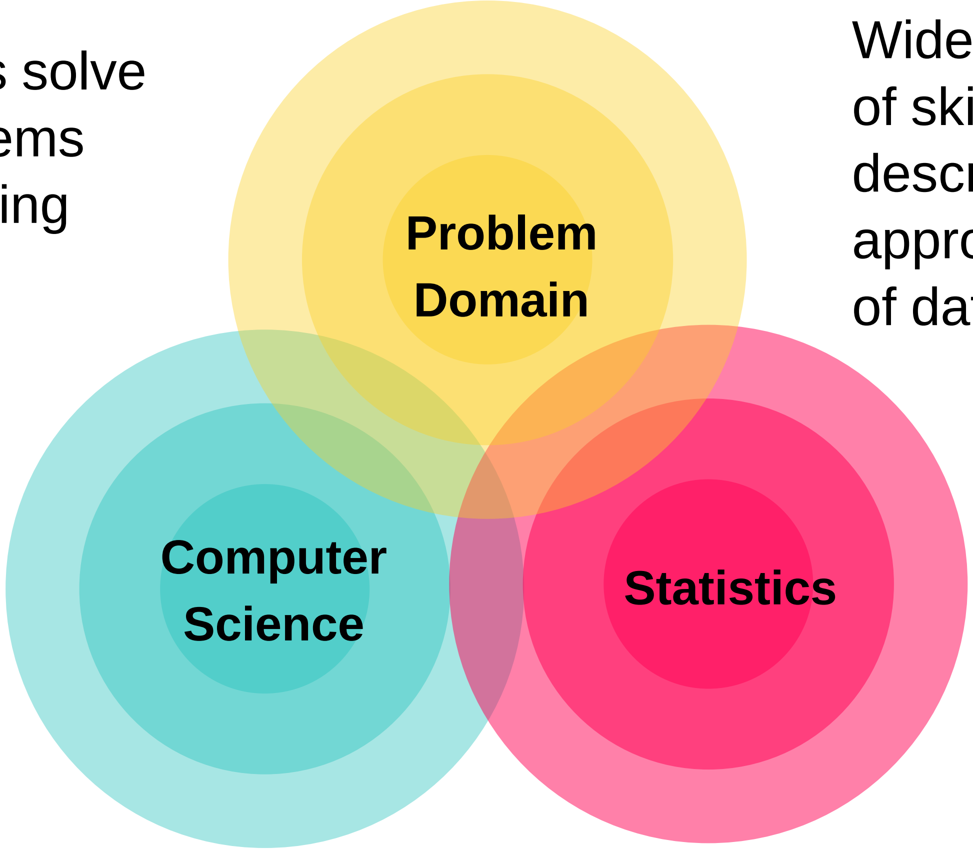
WHAT IS A DATA SCIENTIST?

Data Scientists solve complex problems using data mining techniques



WHAT IS A DATA SCIENTIST?

Data Scientists solve complex problems using data mining techniques



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

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

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

Source: https://www.youtube.com/watch?v=fPzmnRj671Y

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

II. THE DATA SCIENCE WORKFLOW

THE DATA SCIENCE 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

Vital Areas: Heart Rate, Blood Pressure, etc…

Want to collect all data on the claim form (mostly free text)

II. EXPLORE AND PREPARE DATA

Aggregate data at the minute level

Cluster like words

III. BUILD AND EVALUATE MODELS

Compare Decision Tree with Logistic Regression

Start with Naïve Bayes Classifier

IV. COMMUNICATE RESULTS

Create custom dashboard for doctors and nurses

Create report and dashboard proof of concept

III. QUALITIES OF A GOOD DATA SCIENTIST

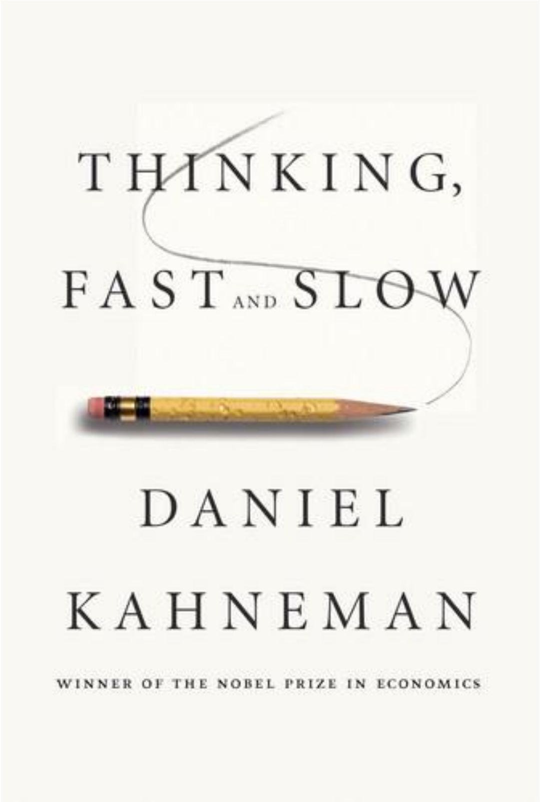
ASKS RATIONAL QUESTIONS

As demonstrated by Nobel Laureate Daniel Kahneman, it can be difficult to think rationally.

The Hindsight Illusion

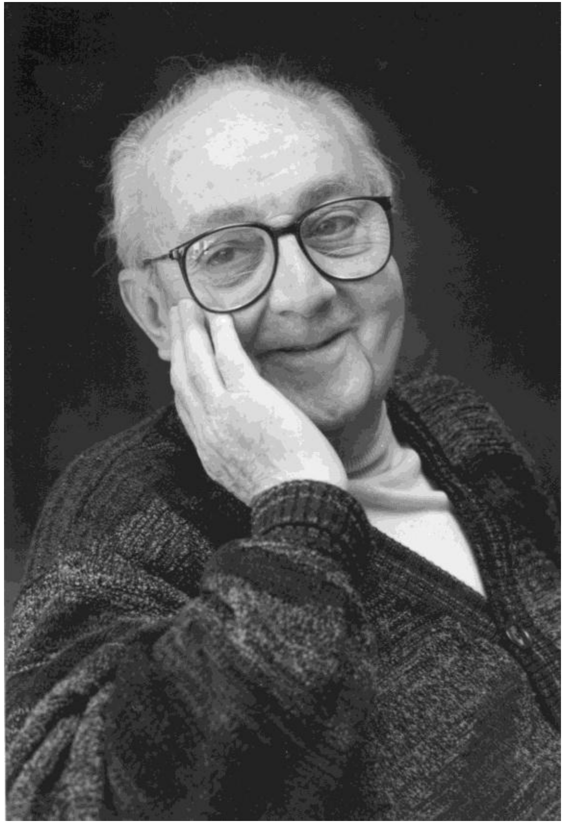
Law of Small Numbers

The Anchoring Effect



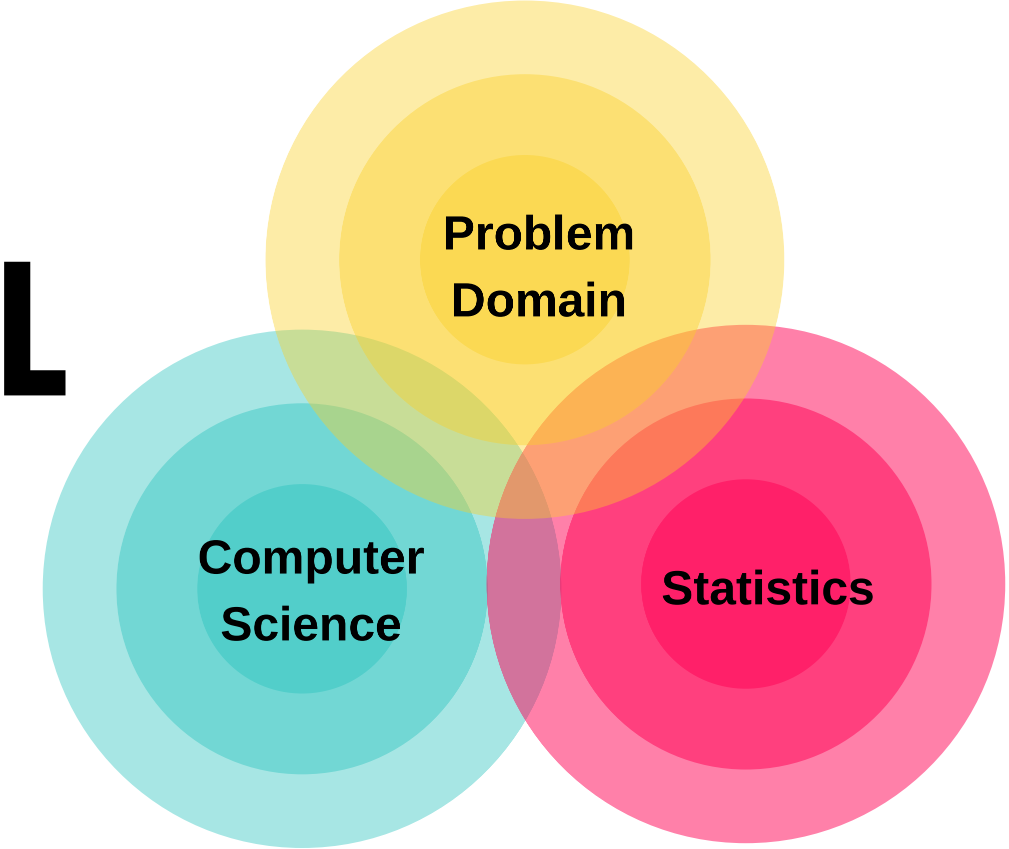
UNDERSTANDS THE PROS & CONS OF DIFFERENT TECHNIQUES

STATISTICIANS, LIKE ARTISTS, HAVE THE BAD HABIT OF FALLING IN LOVE WITH THEIR MODELS – GEORGE BOX



COMMUNICATES CLEARLY

RETAINS INTELLECTUAL HUMILITY


Problem Domain
Computer Science
Statistics