GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 1: DATA SCIENCE OVERVIEW

Rob Hall

DAT13 SF // March 9, 2015

TODAY'S AGENDA

1. Course producer introduction
2. Student introductions
3. Instructor introductions
4. Lecture: Introduction to Data Science
5. Tutorial: Introduction to iPython
6. Q&A

Introductions

- Your name
- A brief summary of your background (e.g. work, school, etc.)
- What you hope to get out of the class
- One interesting / surprising / random factoid about yourself

Meet Your Instructional team

Rob Hall

Rob Hall is a product leader and data scientist who creates business value from data.

Rob leads the Product Management team at Jut, a big data analytics startup backed by Accel, Lightspeed Venture Partners, and Wing. Prior to Jut, Rob led software product strategy at Stem, an energy startup at the nexus of real-time data, predictive analytics, and energy storage.

In addition to enterprise software, Rob has deep consumer Internet experience. At Overture and Yahoo!, Rob drove significant increases in monetization, relevance, and user engagement by applying machine learning algorithms. He has also launched search and social media products used by tens of millions of people every month. Rob graduated from Cornell University with a BS in Engineering and earned his MBA in Finance from the Wharton School of the University of Pennsylvania.

GENERAL ASSEMBLY

overture a YAHOO! company

YAHOO!

JUT

Matthew Ghent

Matthew Ghent graduated from Penn State with a degree in Life Science. He earned a MS degree from USC in Cell and Neurobiology. While at USC he worked on determining molecular signatures of drug response in childhood cancers.

He works as a bioinformatic data analyst / data scientist at Invitae, a genetic diagnostic company. If he's lucky he spends the day writing python, exploring data with pandas and making visualizations.

He is a graduate of DAT5

GENERAL ASSEMBLY

INVITAE

Meet Your Instructional team

Ankit Jain

Ankit Jain is a data scientist who is passionate about building technology products leveraging data science.

Ankit currently works as a data scientist at Clearslide, a leading sales engagement platform. He uses maths,machine learning and programming skills to develop predictive models which eventually make their way into the product. Prior to Clearslide, he worked as a quant at Bank of America and interned at Facebook as a data scientist.

Ankit graduated with a Masters in Financial Engineering from Haas School of UC Berkeley. Prior to that, he did his BS in Electrical Engineering from IIT Bombay, India.

clearslide Bank of America

Meet Your Instructional team *

Chetan Nandakumar

Chetan is a UC Berkeley PhD graduate, where his graduate research focused on the intersection of computer vision and human visual perception. He conducted targeted experiments on humans where data analysis techniques were key in his ability to assess learning strategies employed by the brain. He then used these insights in the crafting of computer vision algorithms.

Outside of the PhD, he has had development experience at HP and IBM Research Labs and has had product experience while at Leap Motion.

GENERAL ASSEMBLY

IBM

LEAP MOTION

Berkeley UNIVERSITY OF CALIFORNIA

RESOURCES

Instructor: Rob Hall (robhall.ga@gmail.com)

Experts-in-Residence:

Matt Ghent (**MATTGHENT@GMAIL.COM**)

Ankit Jain (**ASJ.ANKIT@GMAIL.COM**)

Chetan Nandakumar (**CHETAN.NANDAKUMAR@GMAIL.COM**)

Course Producer: Vanessa Ohta

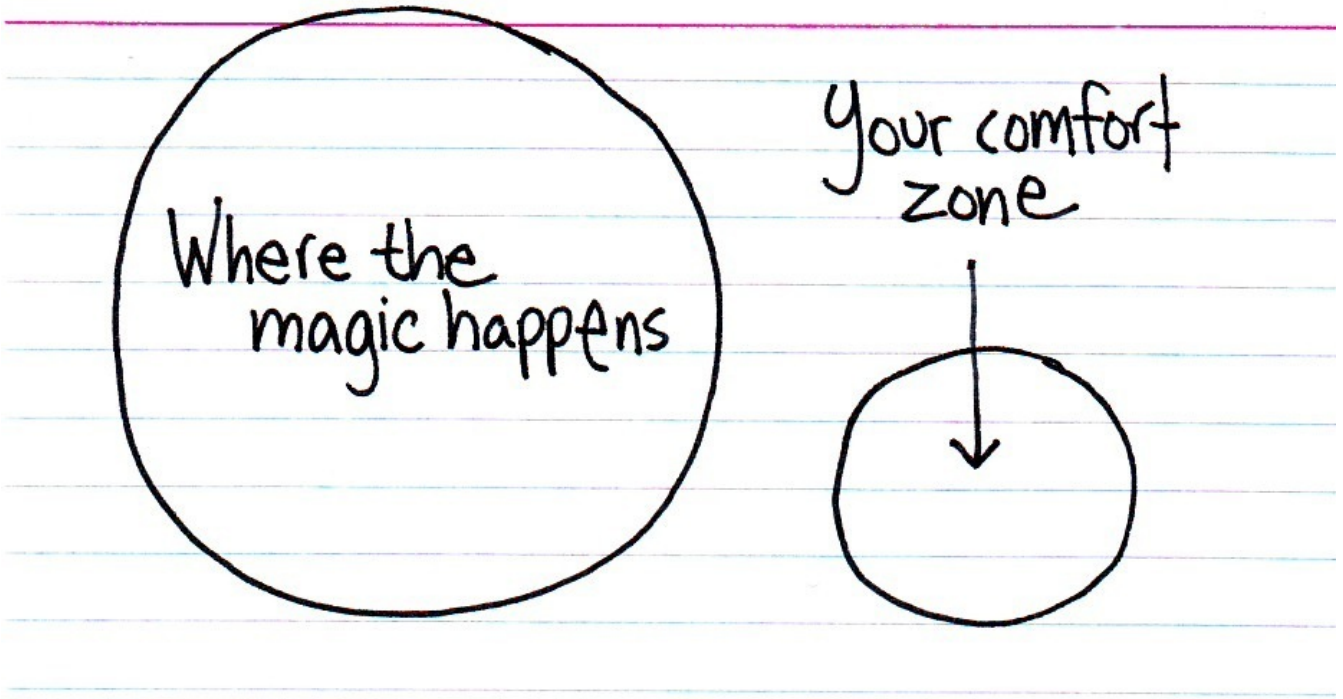
Course Times: 6:30pm-9:30pm, Mondays & Wednesdays (225 Bush St, 3rd Floor, Classroom 7)

Office Hours: See syllabus

Class Dates: March 9, 2015 - May 20, 2015

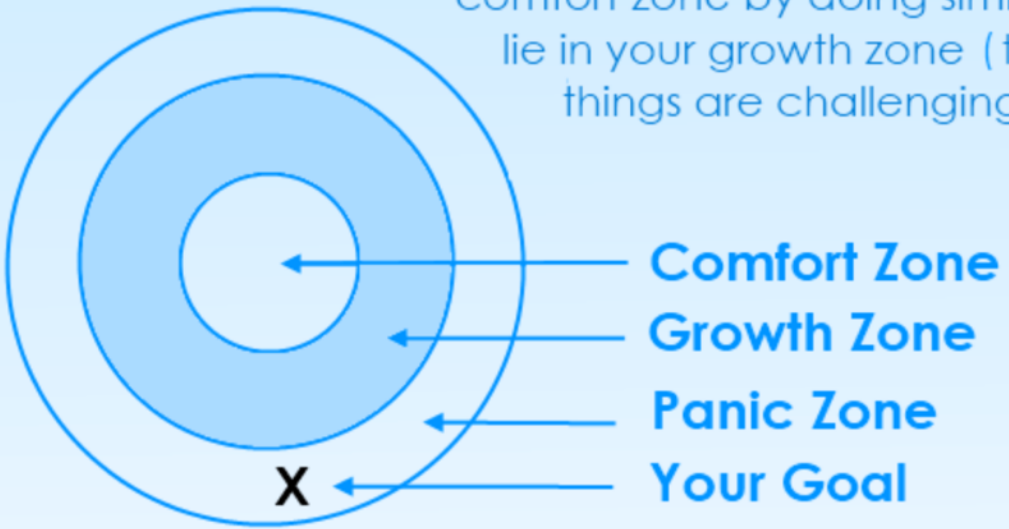
CONGRATULATIONS...

...for getting out of your comfort zone!



How to Grow Your Comfort Zone

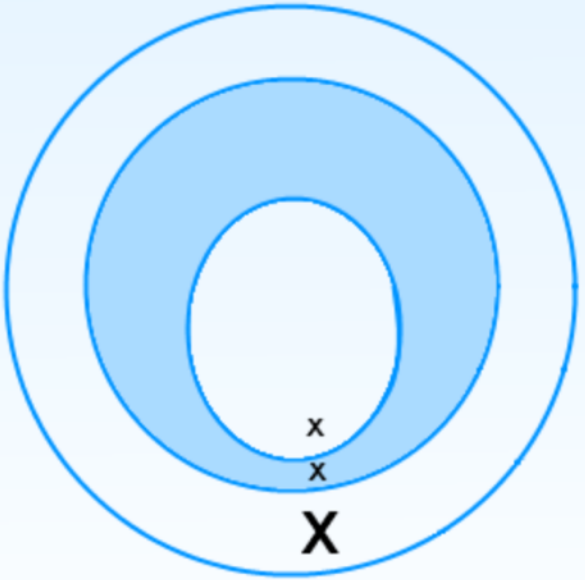
Any goal or challenge may fall into one of three zones - your comfort zone, growth zone, or panic zone. If your goal is currently in your panic zone, i.e. it would be too scary to do now, you will need to grow your comfort zone by doing similar challenges that lie in your growth zone (the zone in which things are challenging or scary, but do-able).



How to Grow Your Comfort Zone

As you pursue challenges in your growth zone, those challenges become easier and your comfort zone expands.

Eventually, challenges that were previously in your panic zone begin to fall into your growth zone, and ultimately within your comfort zone.



AGENDA

I. WHAT IS DATA SCIENCE?
II. THE DATA SCIENCE WORKFLOW
III. VISUALIZATIONS AS A MEDIUM

LAB:
IV. WORKING AT THE UNIX COMMAND LINE
V. INTRO TO I-PYTHON

I. WHAT IS DATA SCIENCE?

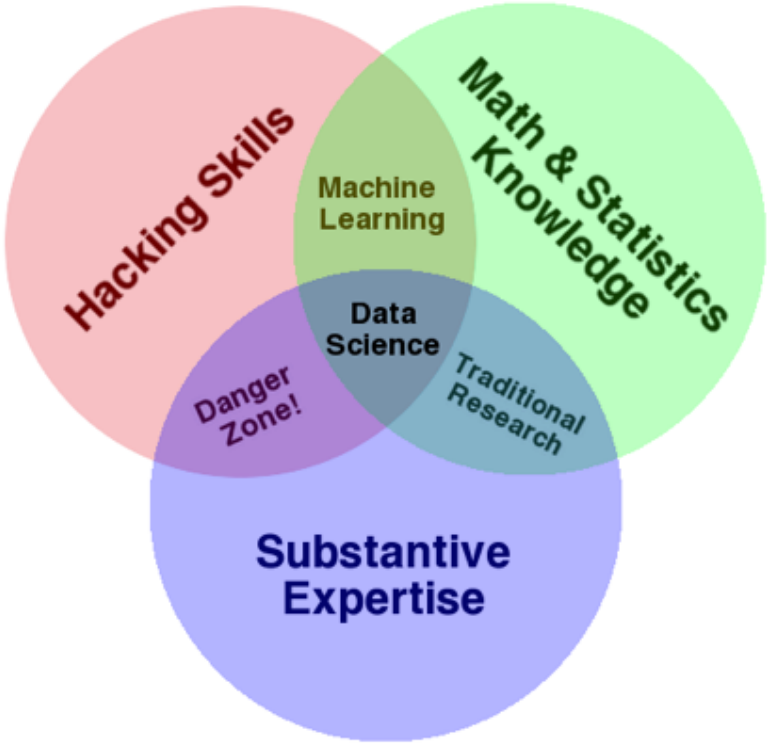
WHAT IS DATA SCIENCE?

- A set of tools and techniques used to extract useful information from data.

WHAT IS DATA SCIENCE?

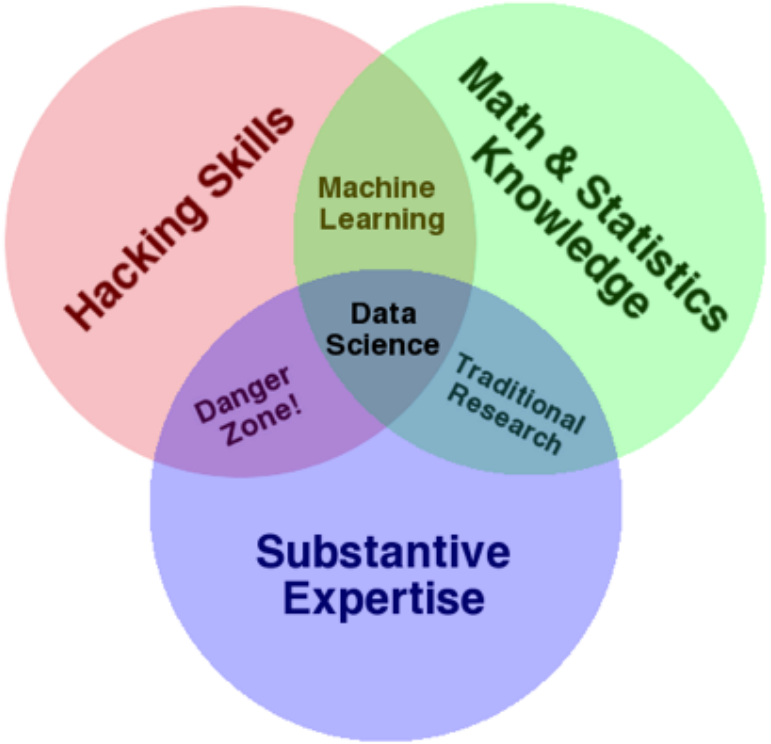
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject.

THE QUALITIES OF A DATA SCIENTIST



source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

THE QUALITIES OF A DATA SCIENTIST



ONE MORE THING!

Communication skills

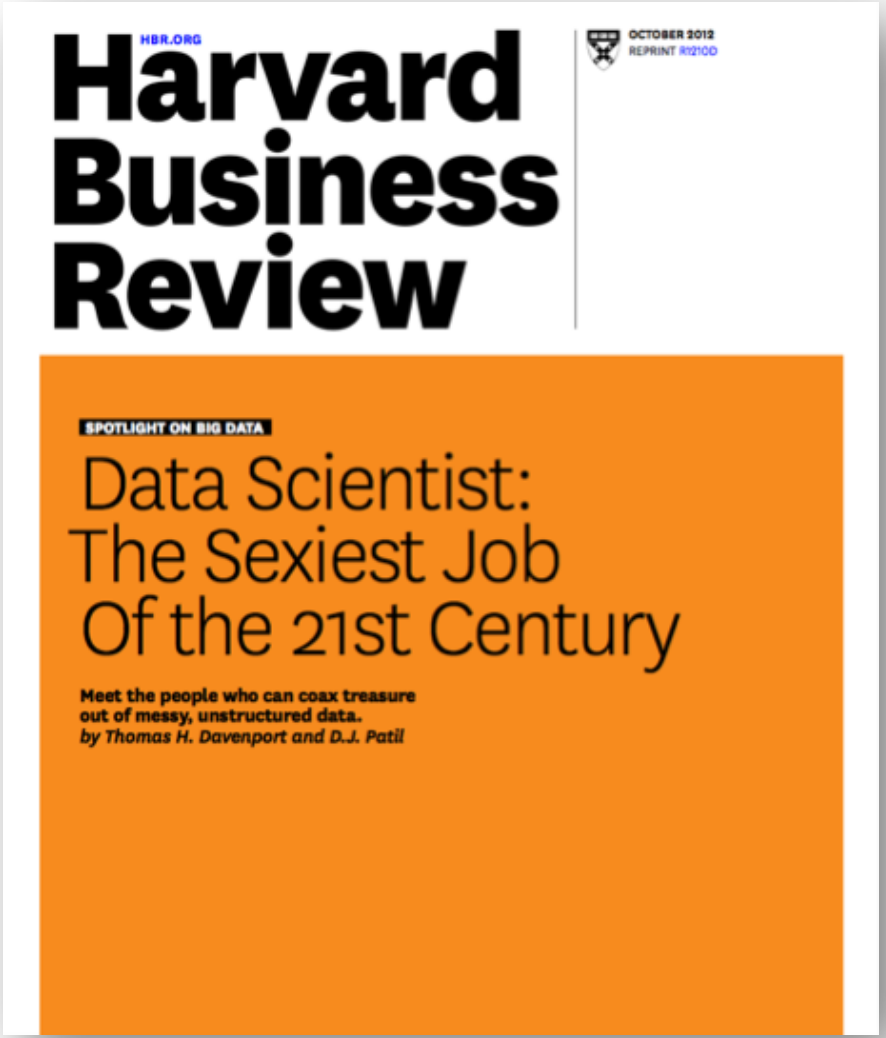
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

WHAT IS DATA SCIENCE?

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.

WHAT IS DATA SCIENCE?

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.
- A rapidly growing field.



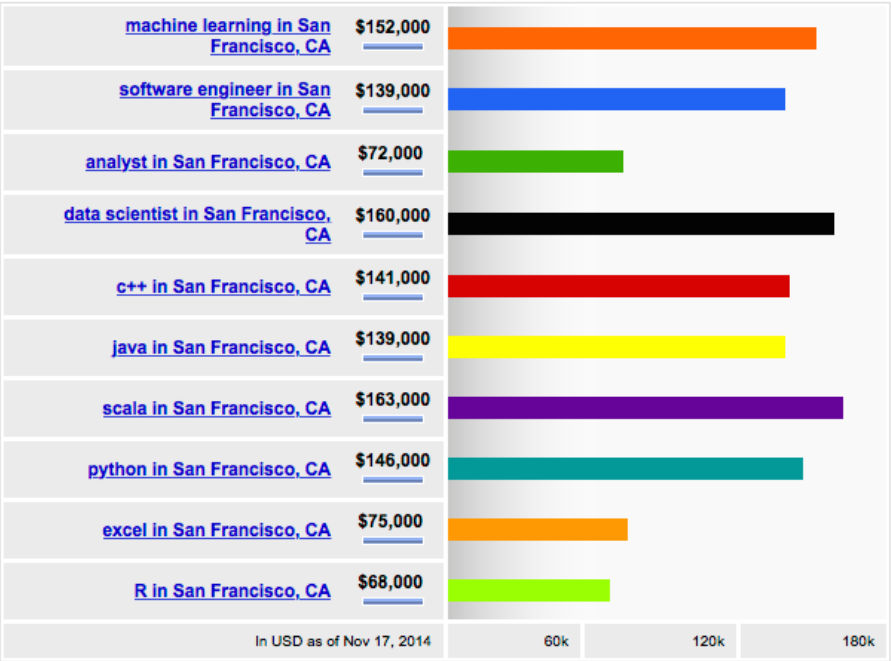
McKinsey estimates 140,000-190,000 shortage by 2018

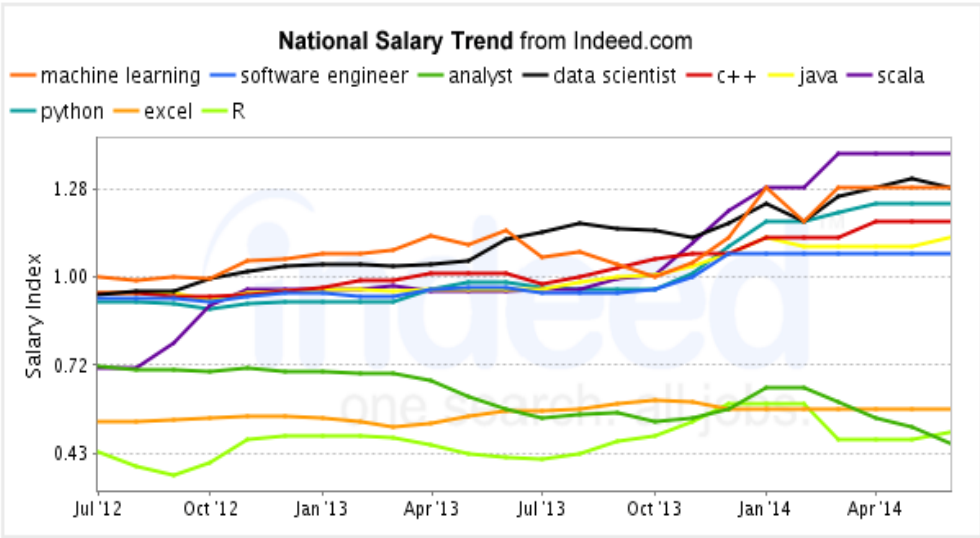
I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?

Hal Varian, Chief Economist at Google, The McKinsey Quarterly, January 2009

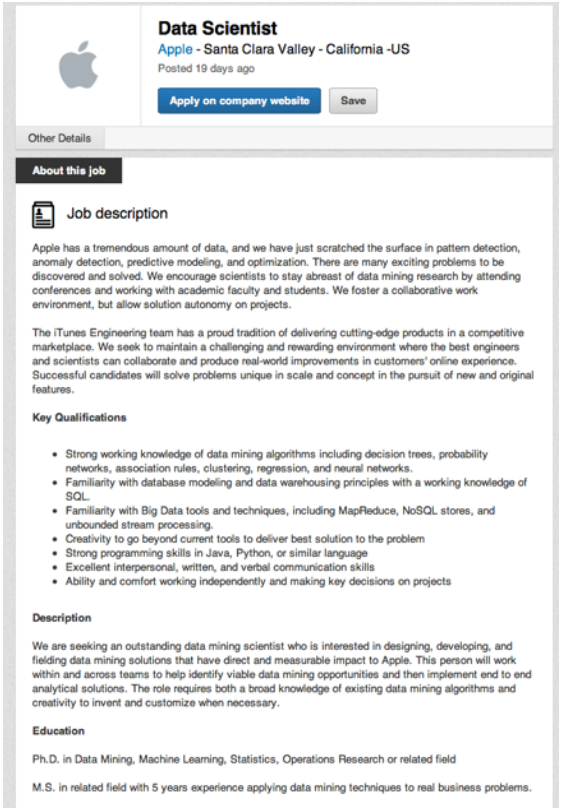
THE MOTIVATOR

Average Salary of Jobs with Titles Matching Your Search





DATA SCIENTISTS WANTED



WHO USES DATA SCIENCE?

WHAT MAKES A GOOD DATA SCIENTIST?

Michael E. Driscoll
@medriscoll

Following

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

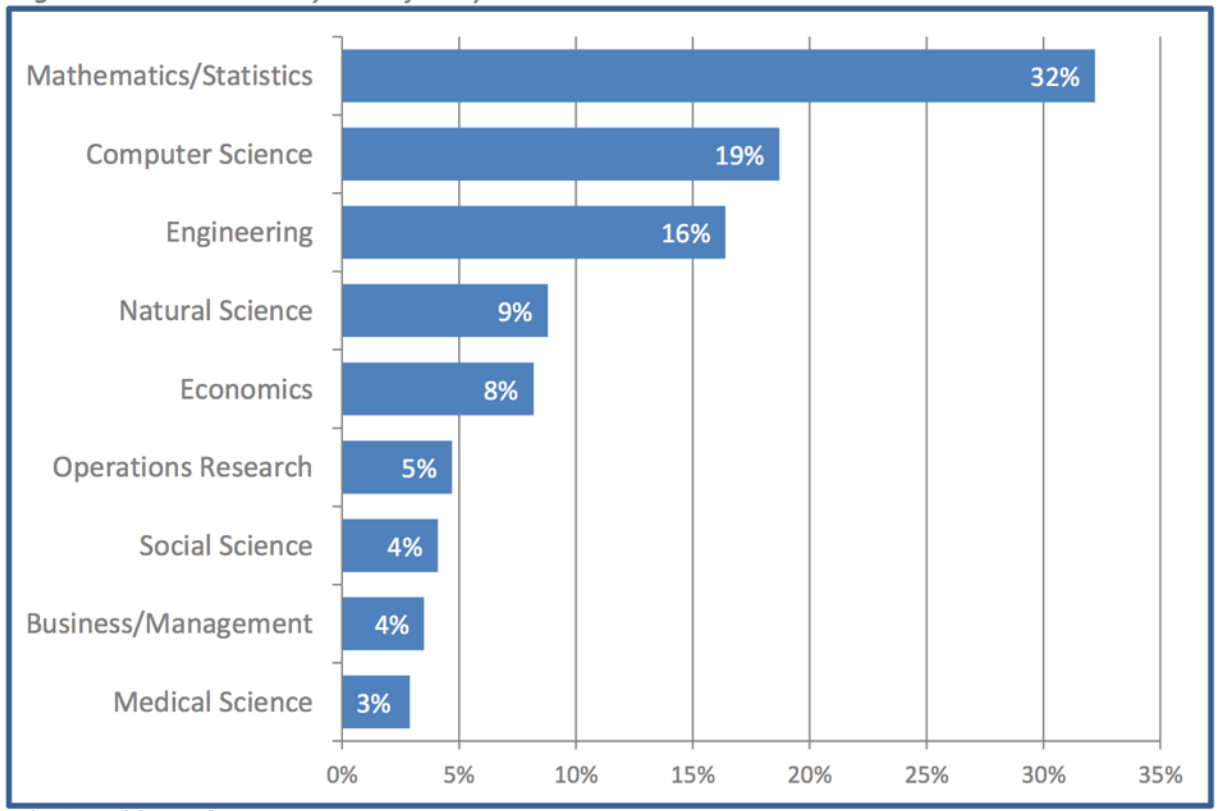
Reply Retweet Favorite More Pocket

WHAT MAKES A GOOD DATA SCIENTIST?

- Statistical and machine learning knowledge
- Engineering experience
- Curiosity
- Product sense
- Storytelling
- Cleverness

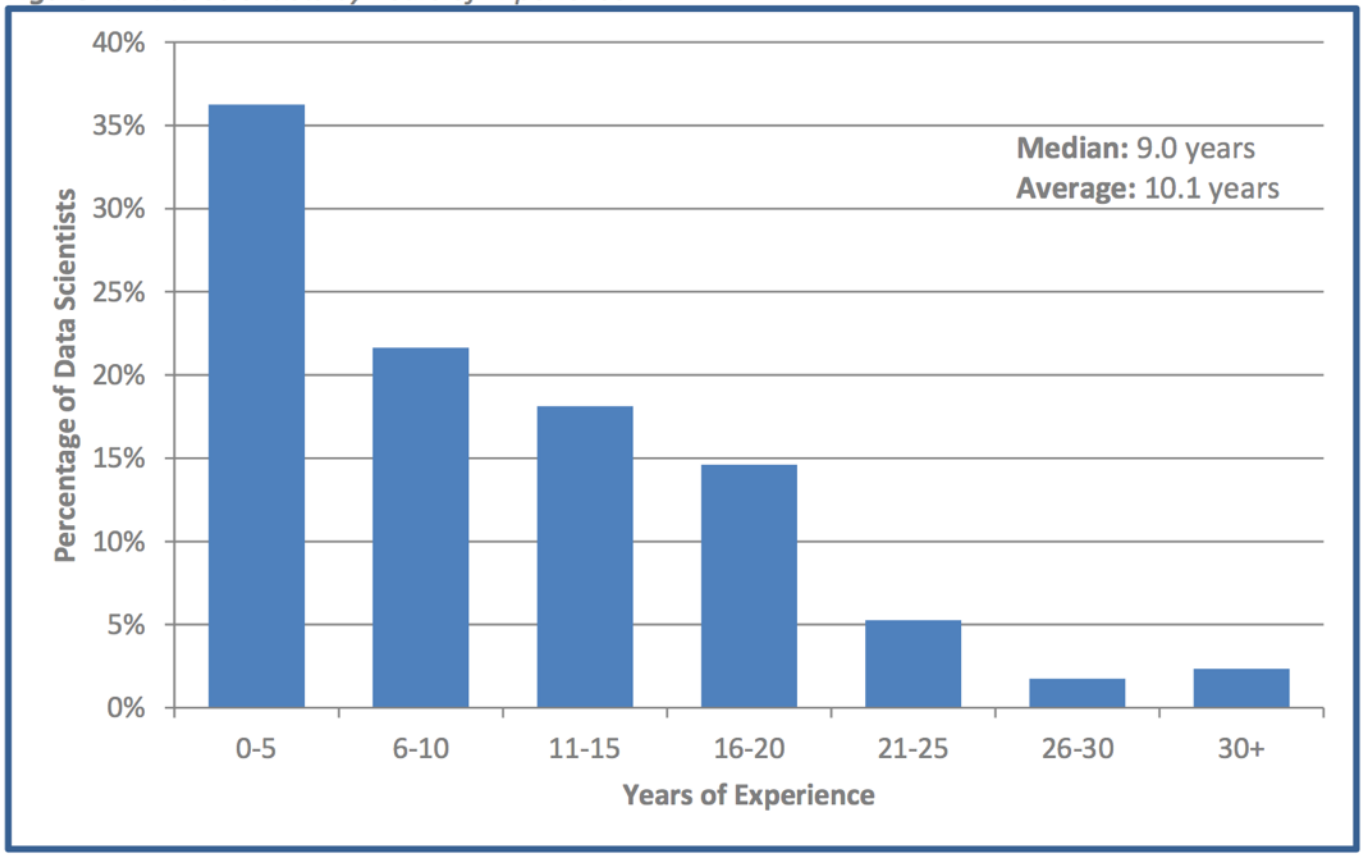
WHO ARE DATA SCIENTISTS?

Figure 8. *Data Scientists by Area of Study*

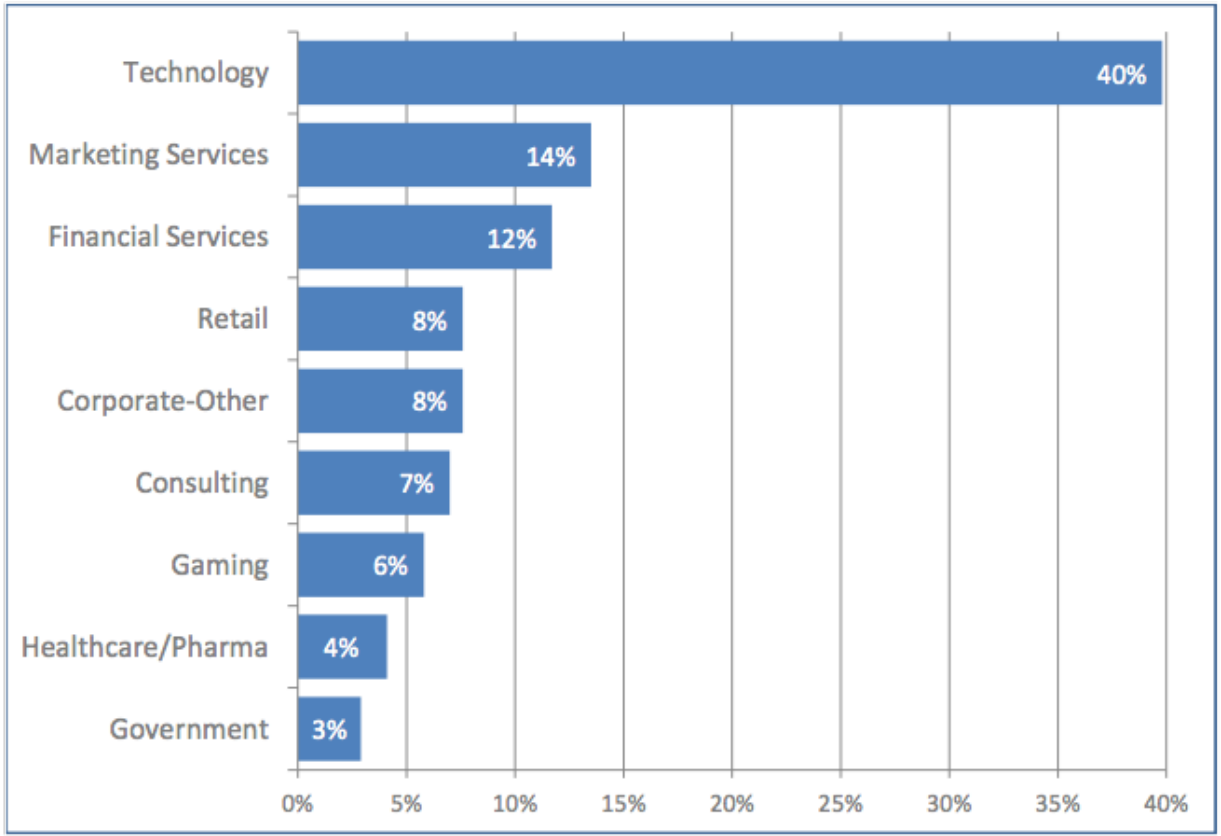


Source: Burtch Works Executive Recruiting study, 2014

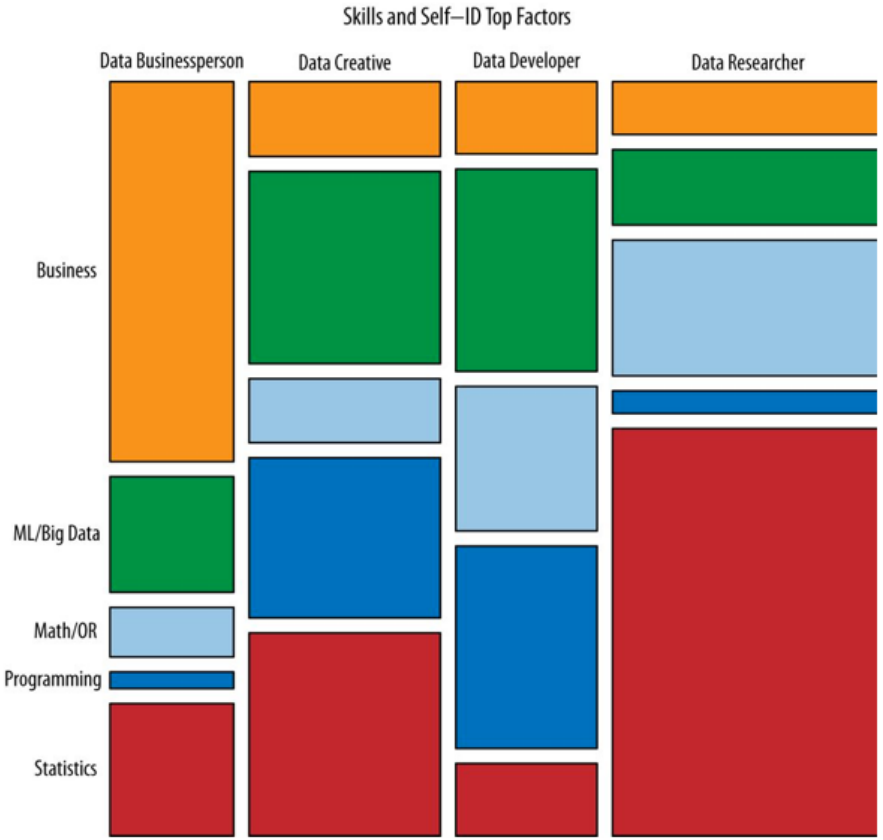
Figure 4. Data Scientists by Years of Experience



WHO ARE DATA SCIENTISTS?



ANALYZING THE ANALYZERS

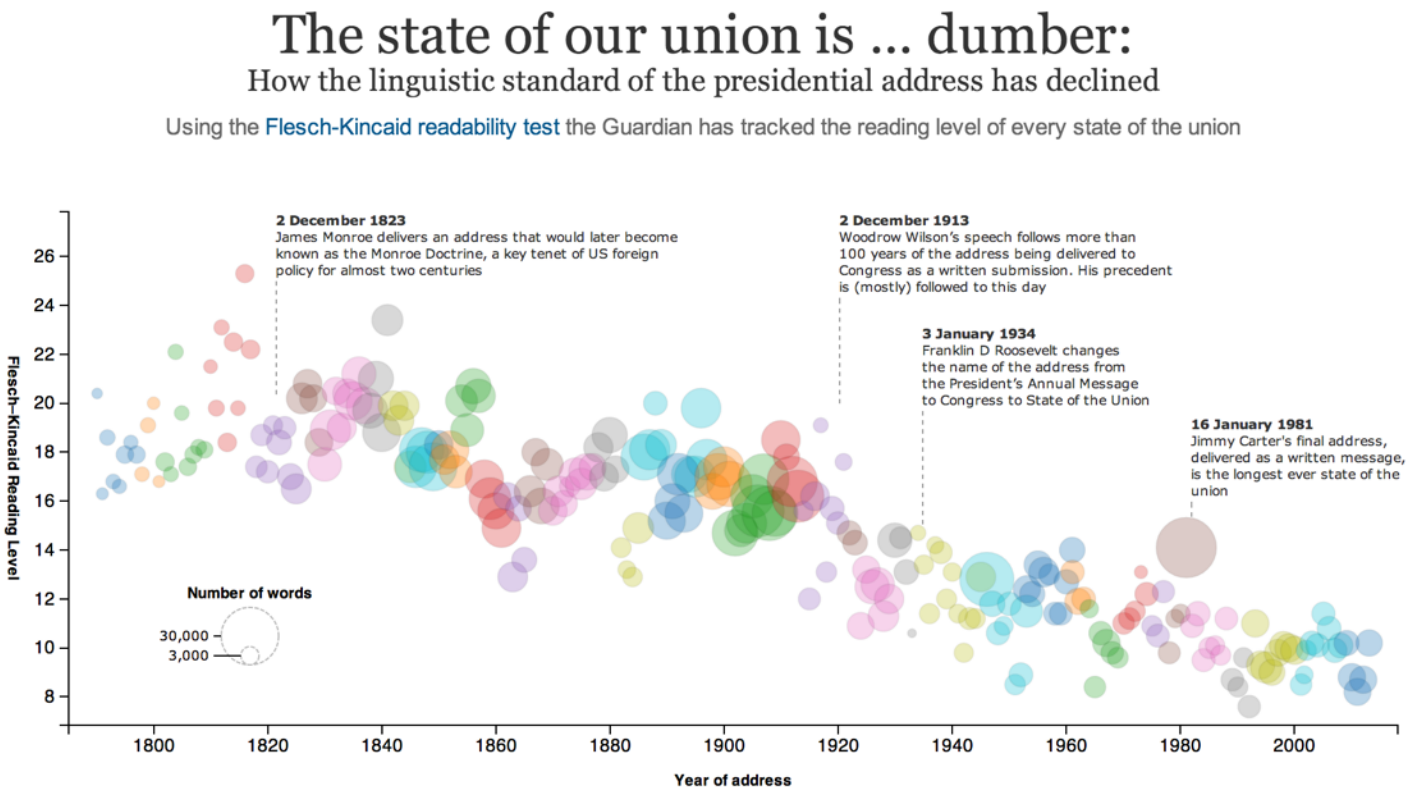


"Analyzing the Analyzers" by Harlan Harris, Sean Murphy, and Marck Vaisman , O'Reilly Strata 2012

COURSE CALENDAR

Session	Date	Day	Topic	HW	Final Project
Unit 1: Basics of Data Science					
1 U1-1	09-Mar	M	Introduction to Data Science		
2 U1-2	11-Mar	W	Introduction to Python	1	
3 U1-3	16-Mar	M	Data Cleaning & Exploration		
Unit 2: Machine Learning Fundamentals					
4 U2-1	18-Mar	W	Introduction to ML, kNN	2	
5 U2-2	23-Mar	M	Linear Regression & Regularization		
6 U2-3	25-Mar	W	Decision Trees & Random Forests	3	Title Due
7 U2-4	30-Mar	M	Data Acquisition & Web APIs		
8 U2-5	01-Apr	W	Logistic Regression	Midterm	Elevator Pitch In-class
9 U2-6	06-Apr	M	Databases & SQL, More Data Munging		
10 U2-7	08-Apr	W	Clustering with K-Means	4	Midterm Due
Unit 3: Advanced Techniques in Machine Learning					
11 U3-1	13-Apr	M	Dimensionality Reduction		
12 U3-2	15-Apr	W	Naïve Bayes	5	Final Project Proposal
13 U3-3	20-Apr	M	Text Mining & NLP		
14 U3-4	22-Apr	W	Non-linear Methods & SVMs	TBD	Proposal Due
15 U3-5	27-Apr	M	Recommender Systems		
Unit 4: Advanced Topics & Guest Speakers					
16 U4-1	29-Apr	W	Overview of "Big Data" & Distributed Computing Concepts		
17 U4-2	04-May	M	Guest Speaker (TBD)		
18 U4-3	06-May	W	Guest Speaker (TBD)		
19 U4-4	11-May	M	Guest Speaker (TBD)		
Unit 5: Final Project: Working Session & Presentations					
20 U5-1	13-May	W	FP Working Session or Guest Speaker		
21 U5-2	18-May	M	Final Project Presentations Day1		Presentation
22 U5-3	20-May	W	Final Project Presentations Day2		Presentation

WHO USES DATA SCIENCE?



WHO USES DATA SCIENCE?

Music + Data:

http://bit.ly/echonest

II. THE DATA SCIENCE WORKFLOW

THE DATA SCIENCE WORKFLOW

Dataists (Hilary Mason & friends)

- 1. Obtain
- 2. Scrub
- 3. Explore
- 4. Model
- 5. Interpret

source: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

Dataists (Hilary Mason & friends)

- 1. Obtain - pointing and clicking does not scale (APIs, Python, shell scripting)
- 2. Scrub - “Scrubbing data is the least sexy part of the analysis process, but often one that yields the greatest benefits” (Python, sed, awk, grep)
- 3. Explore - look at the data (visualizing, clustering, dimensionality reduction)
- 4. Model - “All models are wrong, but some are useful” / models are built to predict and interpret
- 5. Interpret - “The purpose of computing is insight, not numbers”

source: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

THE DATA SCIENCE WORKFLOW

Jeff Hammerbacher (Facebook, Cloudera)

- 1. Identify problem
- 2. Instrument data sources
- 3. Collect data
- 4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
- 5. Build model
- 6. Evaluate model
- 7. Communicate results

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

Ted Johnson

- 1. Assemble an accurate and relevant data set
- 2. Choose the appropriate algorithm

THE DATA SCIENCE WORKFLOW

Ben Fry

- 1. Acquire
- 2. Parse
- 3. Filter
- 4. Mine
- 5. Represent
- 6. Refine
- 7. Interact

source: http://benfry.com/phd/dissertation-110323c.pdf

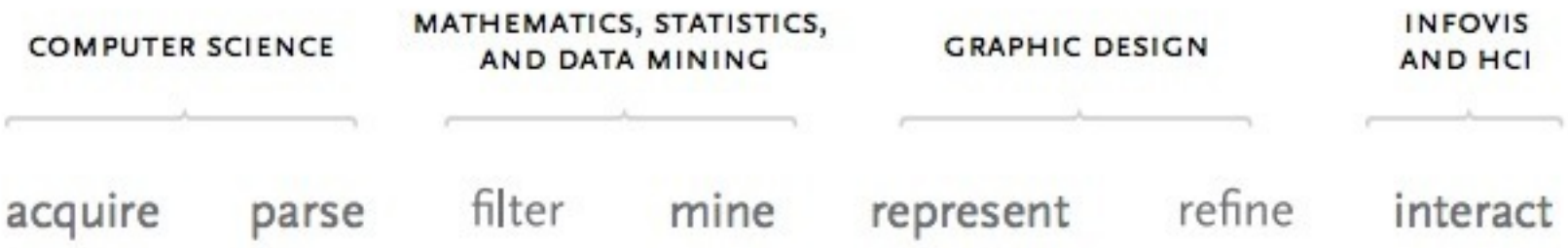
THE DATA SCIENCE WORKFLOW

Ben Fry

- 1. Acquire - the matter of obtaining the data
- 2. Parse - providing some structure around what the data means
- 3. Filter - removing all but the data of interest
- 4. Mine - the application of methods from statistics or data mining, as a way to discern patterns or place the data in mathematical context
- 5. Represent - determination of a simple representation (e.g. graphing)
- 6. Refine - improvements to the basic representation to make it clearer and more visually engaging
- 7. Interact - the addition of methods for manipulating the data or controlling which features are visible

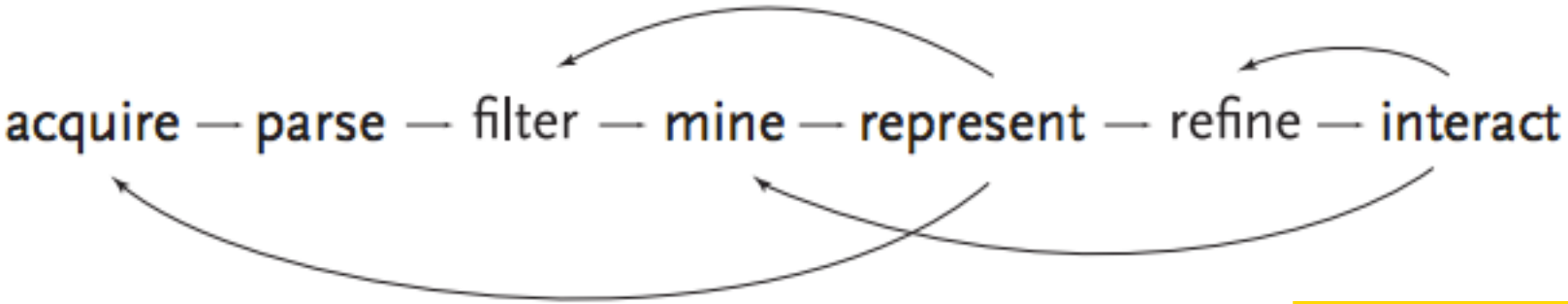
source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW



source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW



NOTE

This diagram illustrates the *iterative* nature of problem solving

source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW

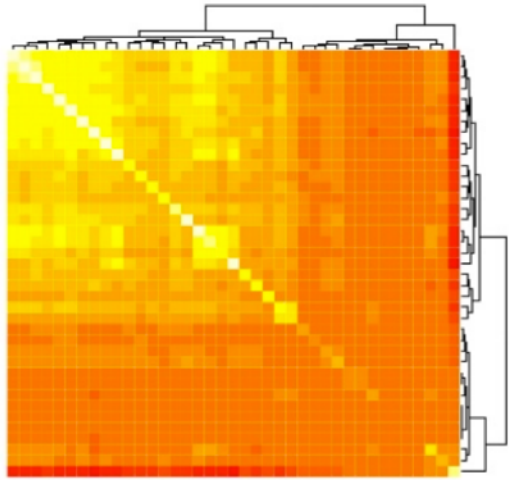
What is needed most?

approximately **80% of the costs** for data-related projects gets spent on data preparation – mostly on **cleaning up** data quality issues: ETL, log files, etc., generally by socializing the problem

unfortunately, data-related budgets tend to go into frameworks that can only be used *after clean up*

most valuable skills:

- *learn to use programmable tools that prepare data*
- *learn to understand the audience and their priorities*
- *learn to socialize the problems, knocking down silos*
- *learn to generate compelling* ***data visualizations***
- *learn to estimate the confidence for reported results*
- *learn to automate work, making process repeatable*



THE DATA SCIENCE WORKFLOW

Modeling

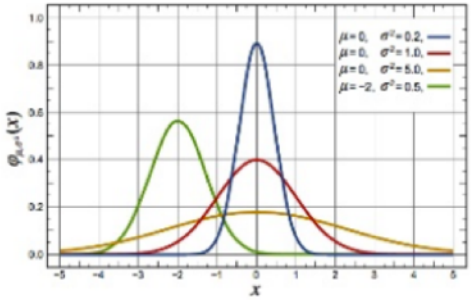
back in the day, we worked with practices based on **data modeling**

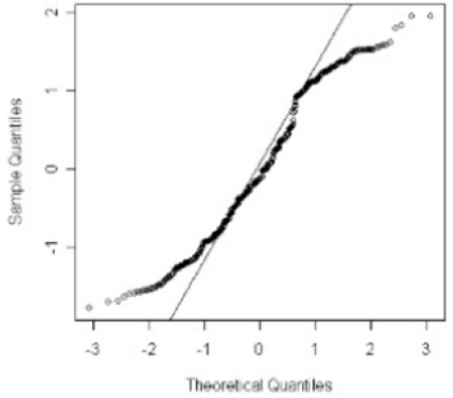
1. *sample the data*
2. *fit the sample to a known distribution*
3. *ignore the rest of the data*
4. *infer, based on that fitted distribution*

that served well with ONE computer, ONE analyst, ONE model… just throw away annoying "extra" data

circa late 1990s: machine data, aggregation, clusters, etc. **algorithmic modeling** displaced the prior practices of data modeling

because the data won't fit on one computer anymore



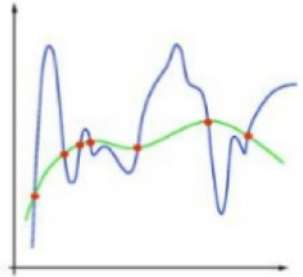


THE DATA SCIENCE WORKFLOW

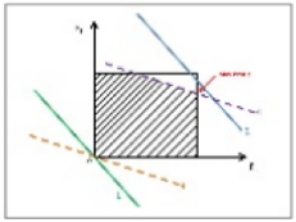
Learning Theory

in general, apps alternate between learning patterns/rules and retrieving similar things…

machine learning – scalable, arguably quite ad-hoc, generally "black box" solutions, enabling you to make billion dollar mistakes, with oh so much commercial emphasis
(i.e. the "heavy lifting")



statistics – rigorous, much slower to evolve, confidence and rationale become transparent, preventing you from making billion dollar mistakes, any good commercial project has ample stats work used in QA
(i.e., "CYA, cover your analysis")



once Big Data projects get beyond merely digesting log files, **optimization** will likely become the next overused buzzword :)

THE DATA SCIENCE WORKFLOW

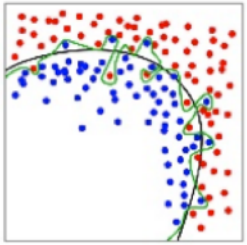
Generalizations about Machine Learning...

great introduction to ML, plus a proposed categorization for comparing different machine learning approaches:

A Few Useful Things to Know about Machine Learning
Pedro Domingos, U Washington
homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

toward a categorization for Machine Learning algorithms:

- **representation**: classifier must be represented in some formal language that computers can handle (algorithms, data structures, etc.)
- **evaluation**: evaluation function (objective function, scoring function) is needed to distinguish good classifiers from bad ones
- **optimization**: method to search among the classifiers in the language for the highest-scoring one

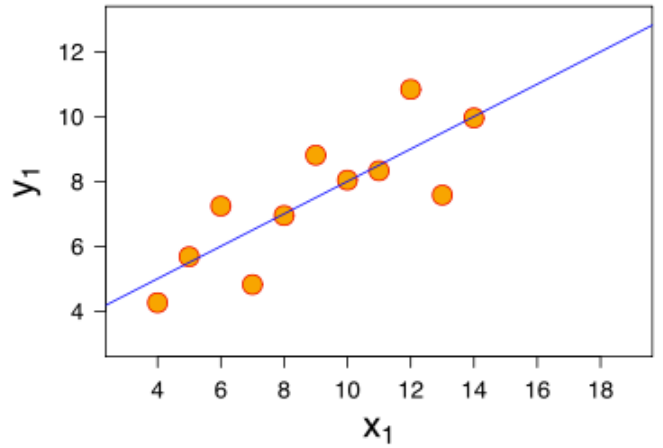


III. VISUALIZATIONS AS A MEDIUM

EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

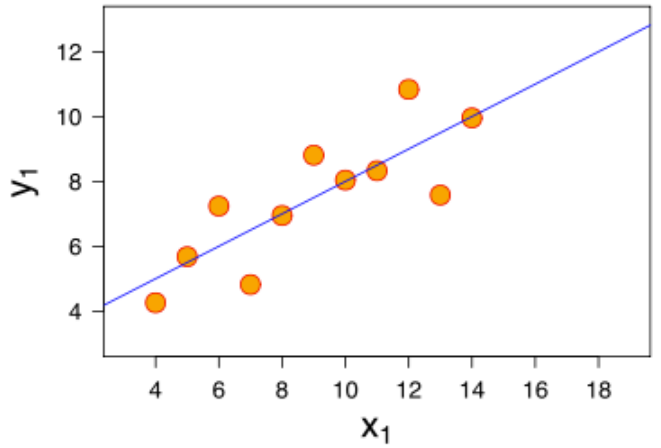
- *eleven (x, y) points*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

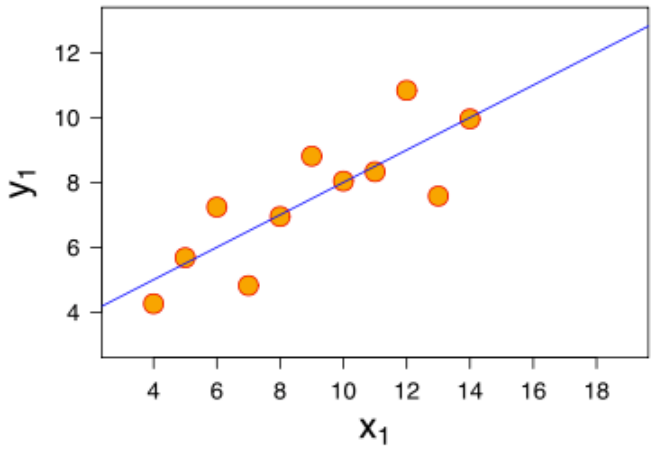
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

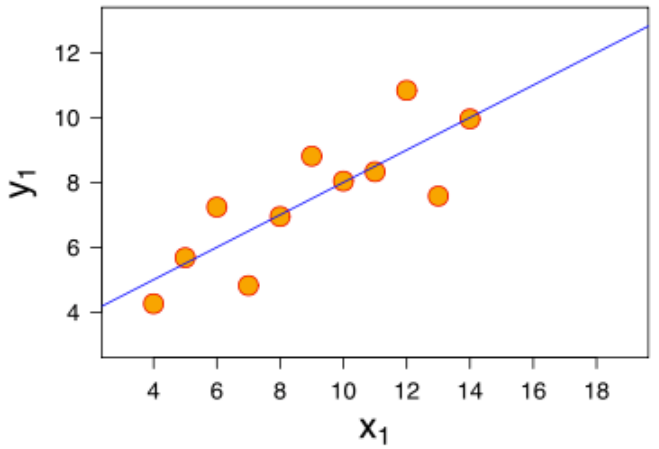
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

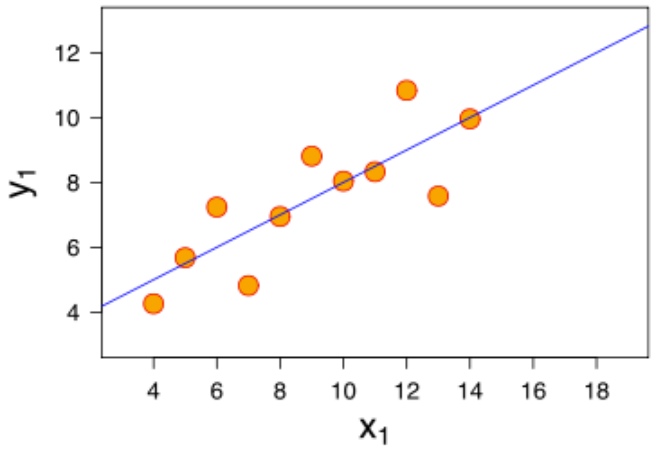
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*
- *correlation of x and y = 0.8*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

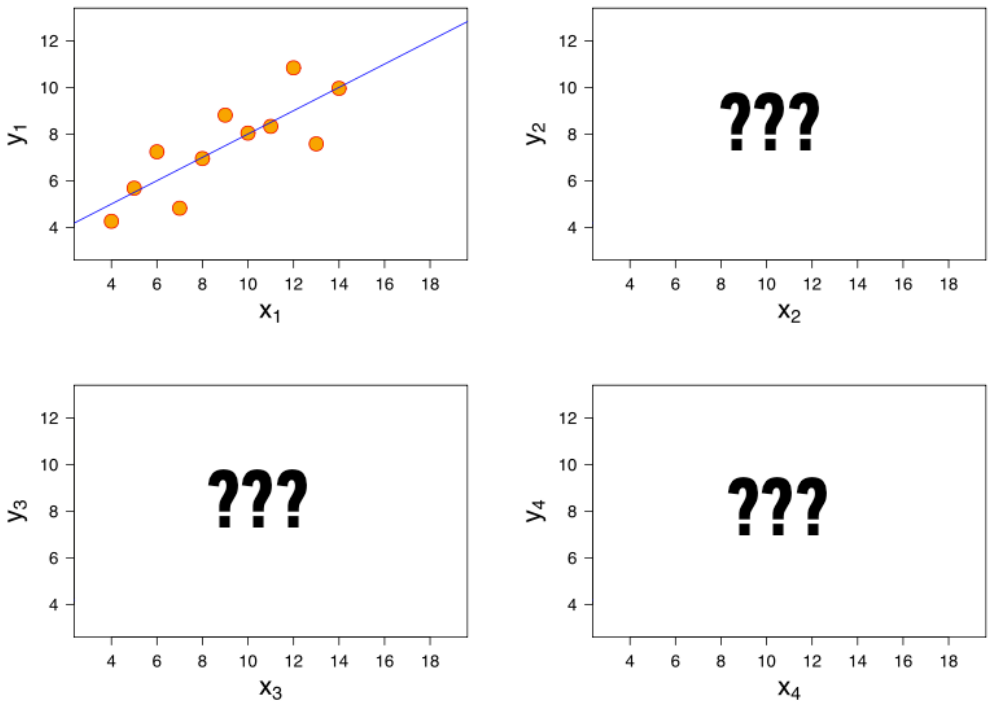
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*
- *correlation of x, y = 0.8*
- *line of best fit: y = 3.00 + 0.500x*



EXERCISE – WHY VISUALIZE DATA?

Now, suppose I give you three more datasets with exactly the same characteristics...

Q: how similar are these datasets?



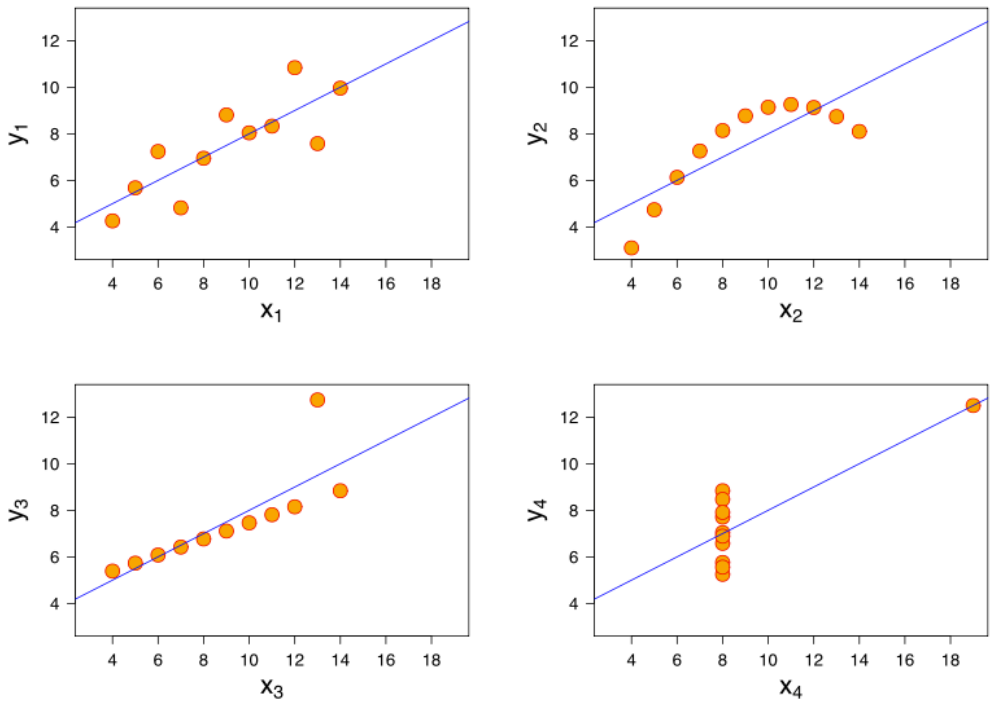
EXERCISE – WHY VISUALIZE DATA?

Now, suppose I give you three more datasets with exactly the same characteristics.

Q: how similar are these datasets?

A: not very!

http://en.wikipedia.org/wiki/Anscombe's_quartet



EXERCISE – WHY VISUALIZE DATA?

Look at your data!

V. INTRO TO I-PYTHON

DISCUSSION