GENERAL ASSEMBLY

INTRO TO DATA SCIENCE AND ANALYTICS

John Sandall

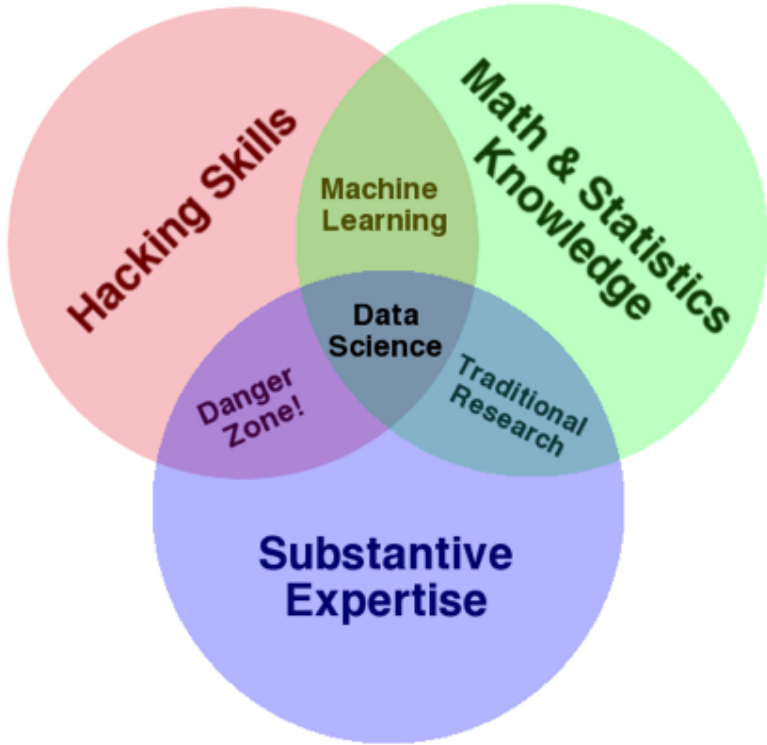
Data Science Consultant

@john_sandall

WELCOME!

I. WHAT IS DATA SCIENCE?
II. DOING DATA SCIENCE
III. TRAITS OF A SUCCESSFUL DATA SCIENTIST
IV. TOOLS
V. OPPORTUNITIES

I. WHAT IS DATA SCIENCE?



ONE MORE THING!

Communication skills

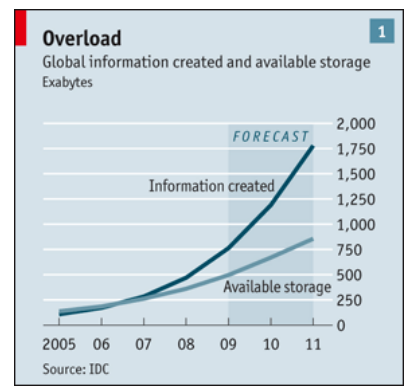
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

- 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.


NETFLIX
amazon
facebook
Google™
LinkedIn
FiveThirtyEight
Nate Silver's Political Calculus
foursquare
Etsy
2012
BARACKOBAMA.COM
Microsoft

- Every two days we create more information than we did up until 2003, around two exabytes

- Creation of data outstrips current capabilities to store it



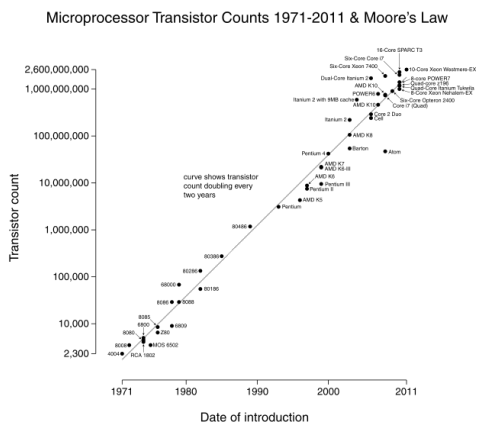
From: http://www.economist.com/node/15557443

- There has never been so much data available to us
- This data differs from much of the data available so far
- Several factors have made this data accessible to us:
  - low hardware costs
  - high hardware performance
  - free tools

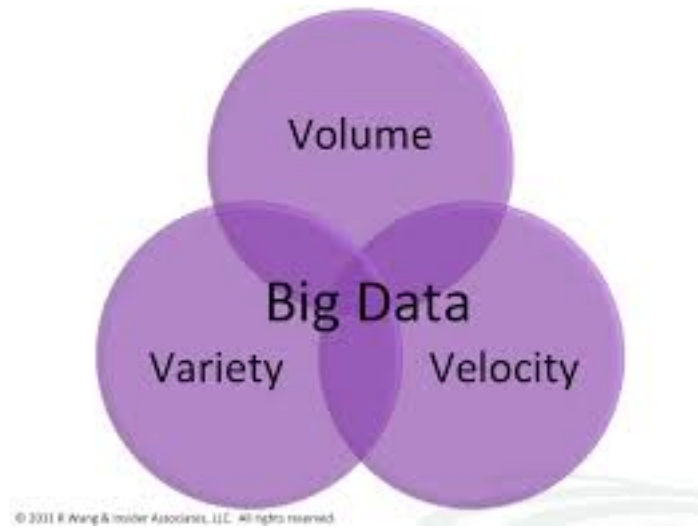
- Computing hardware has come down in cost

- Cloud computing technology has made it possible for everyone to use IT infrastructure

- Moore’s law holds and computers are more powerful



- It describes data that:
  - doesn't fit in memory
  - doesn't fit on a machine



From: http://blog.softwareinsider.org/2012/02/27/mondays-musings-beyond-the-three-vs-of-big-data-viscosity-and-virality/screen-shot-2012-02-27-at-11-18-56-am/

- The increase of computing power has made older techniques practical

- New advances in statistics and machine learning make it possible to analyze large amounts of data

- Not only hardware costs are dropping

- Open Source – Linux, Apache, Hadoop, MySQL

- In addition to infrastructure there are also analysis tools

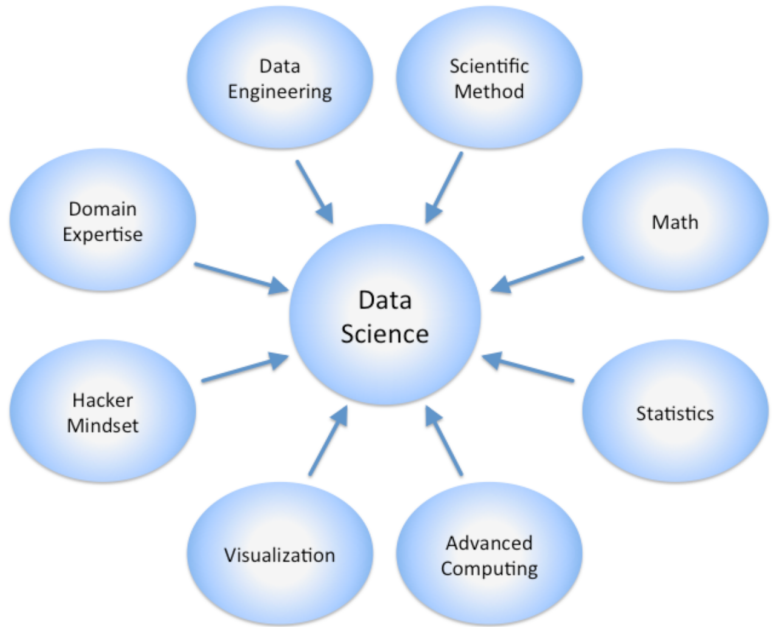
- Who knows R or Python?

- Digital
- Sensors
- Transactions
- Open data

- Innovation
    - Unmet needs
    - Niche segments

- Optimisation
    - e.g. sending taxis to where the demand *will* be

- Given that the data is there
- It aims to:
    - Make sense of data
    - Use appropriate tools



II. DOING DATA SCIENCE

- The availability of data and tools makes data science possible

- Data scientists try to *find new patterns*

- Make *predictions* and *classify*

- When exploring data, it might be interesting to find out what patterns there are in the data

- Techniques to find patterns are called *unsupervised*

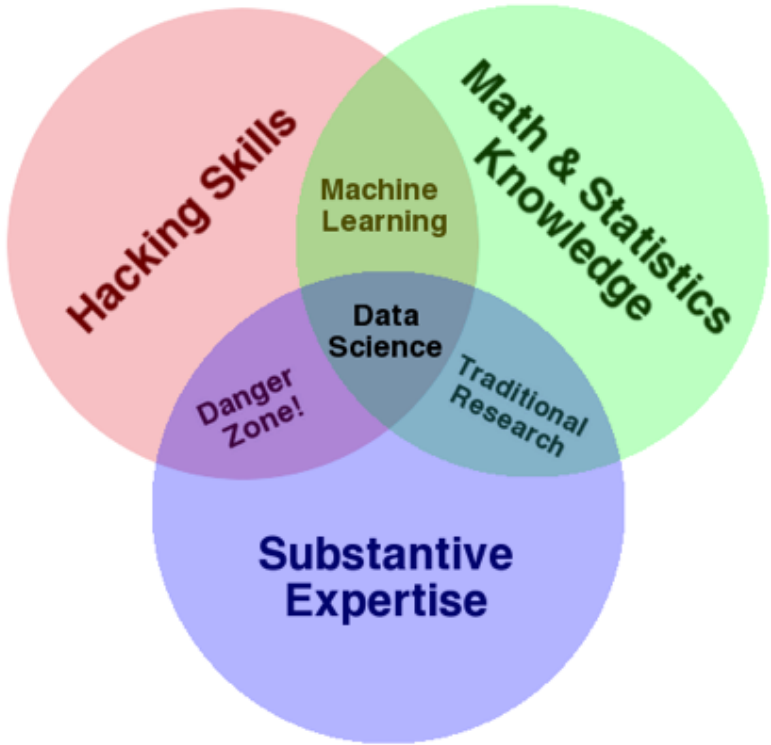
- To determine which category an observation belongs to we use a *classifier* – a *supervised* method

- Another important tasks is prediction

- Given these observations what will happen next

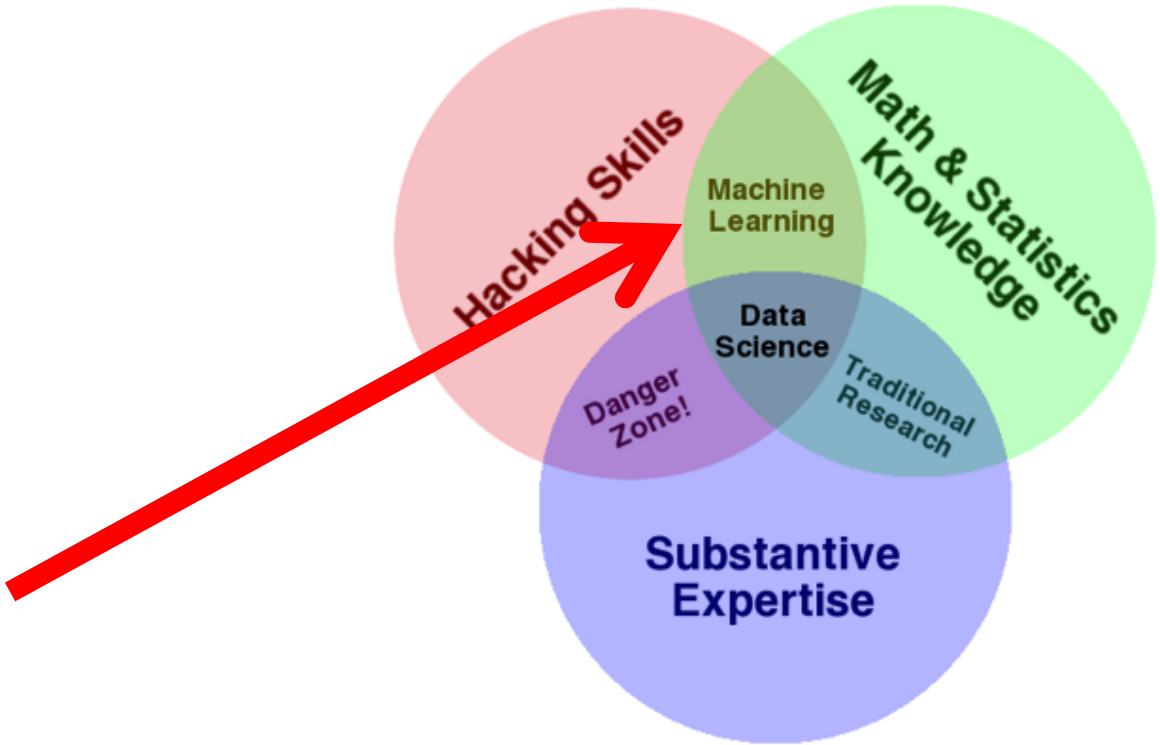
- Techniques that need to data to be used are called *supervised* and they will produce continuous predictions

REMEMBER THIS?



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

YOU ARE HERE



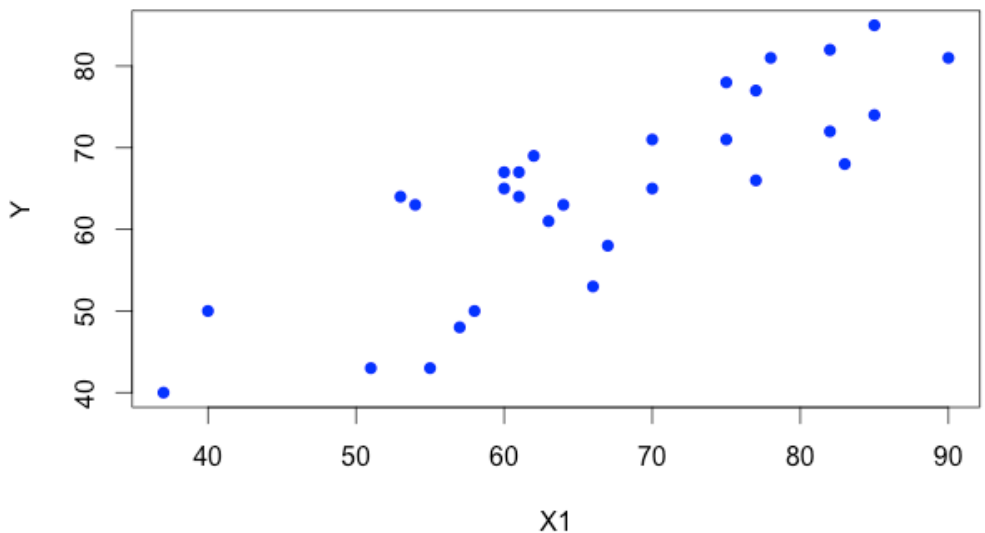
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

supervised	*making predictions*
unsupervised	*extracting structure*

supervised

making predictions

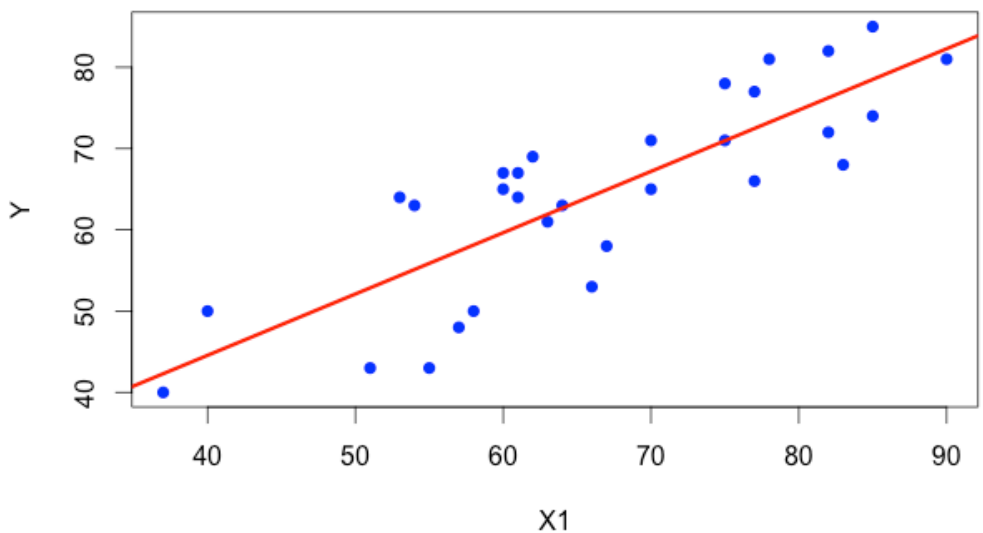
Y	X1
43	51
63	64
71	70
61	63
81	78
43	55



supervised

Y	X1
43	51
63	64
71	70
61	63
81	78
43	55

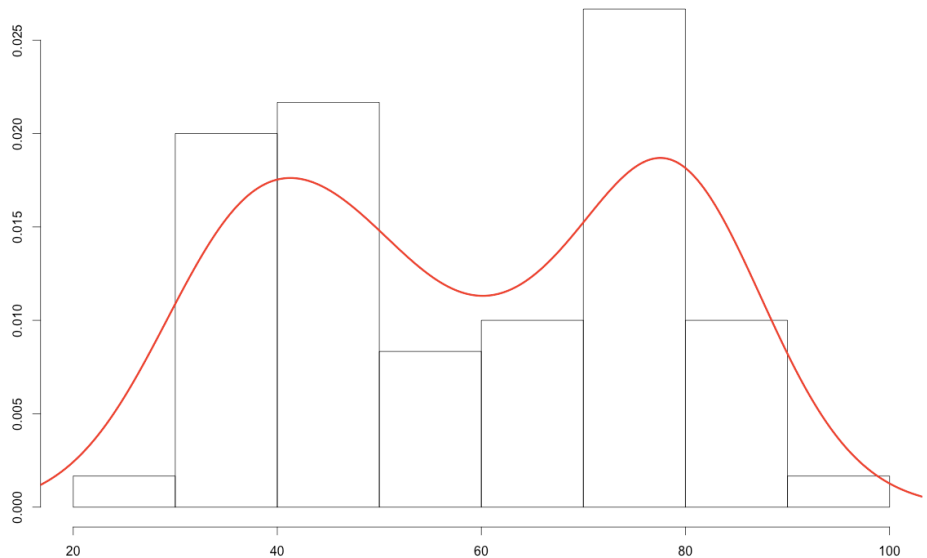
making predictions



unsupervised

extracting structure

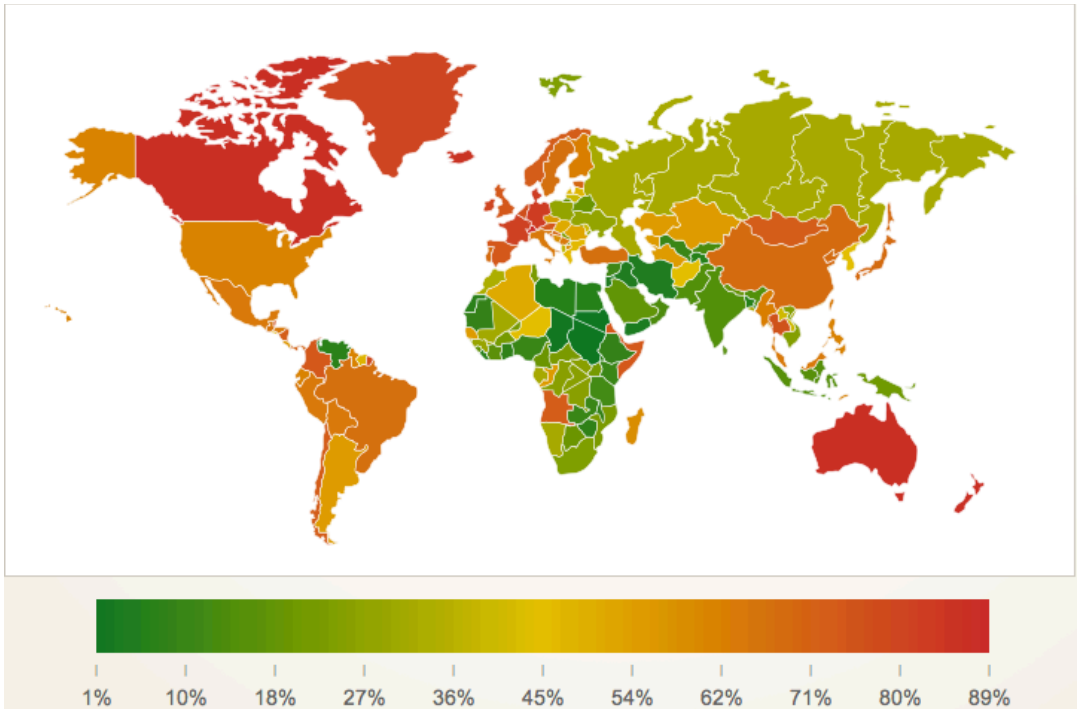
92
73
86
84
83
49
68
66
83
80
67
74
63



	continuous	*categorical*
	quantitative	*qualitative*

	continuous	***categorical***
supervised	*regression*	*classification*
unsupervised	*dimensional reduction*	*clustering*

REGRESSION EXAMPLE: PREDICTING PHONE SALES



GDP

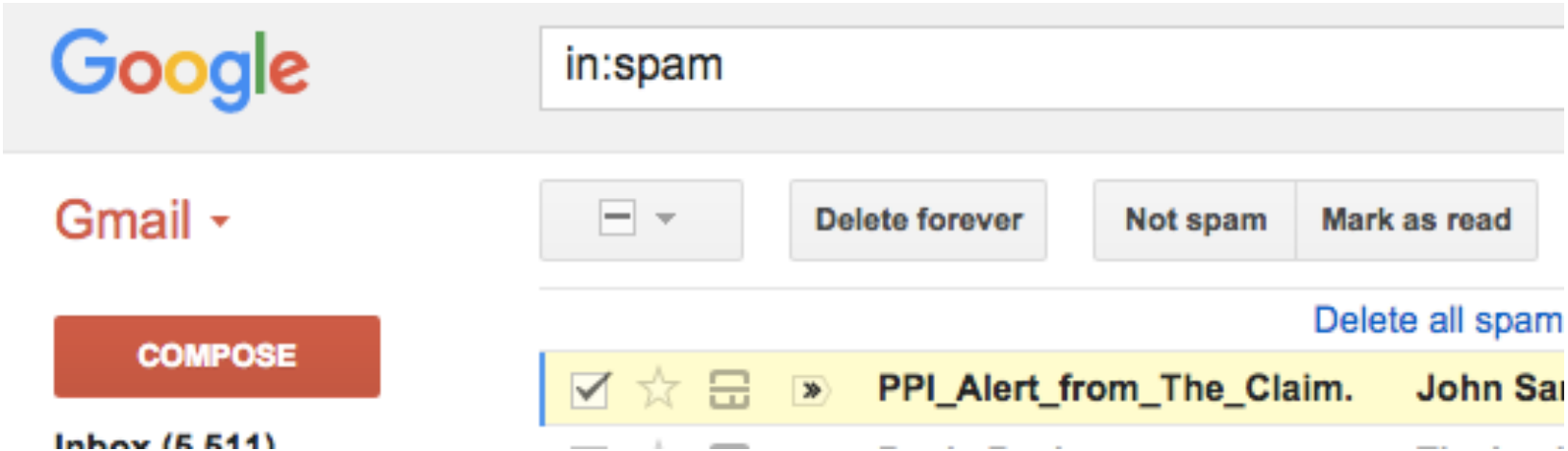
population

Gini

phone penetration %

GDP growth rate

	continuous	***categorical***
supervised	*regression*	*classification*
unsupervised	*dimensional reduction*	*clustering*

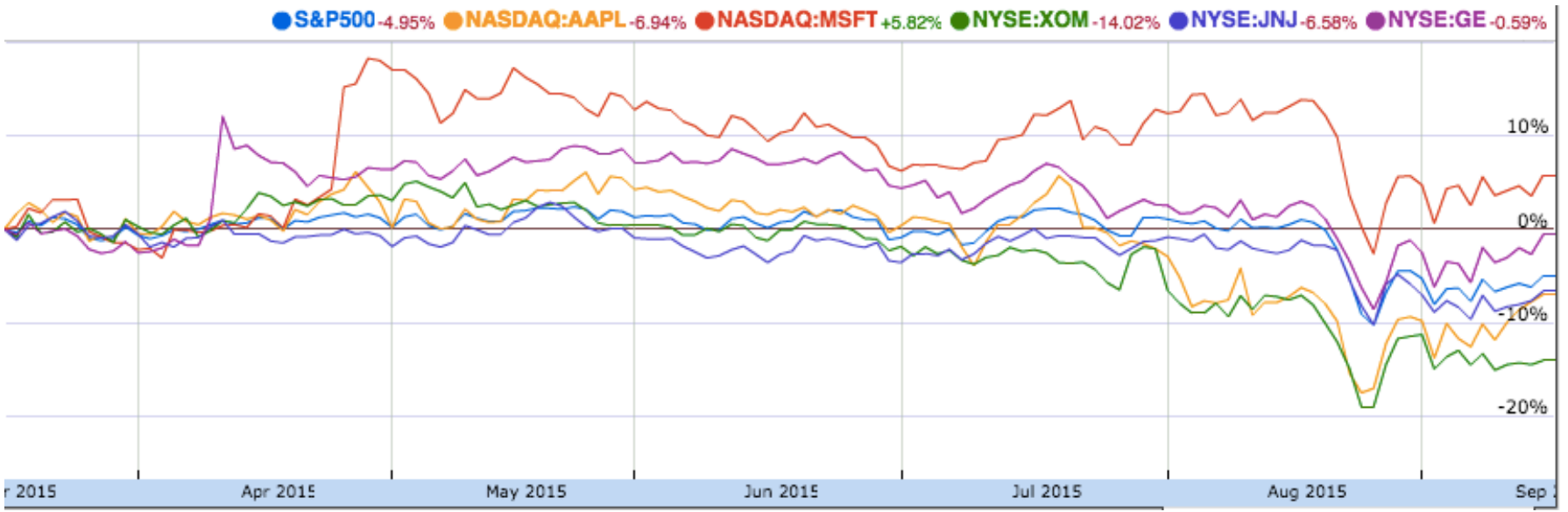


$$$ *Act now!* *As seen on* *Satisfaction guaranteed*

100% free *All natural* *Bargain* *!!!*

	continuous	***categorical***
supervised	*regression*	*classification*
unsupervised	*dimensional reduction*	*clustering*

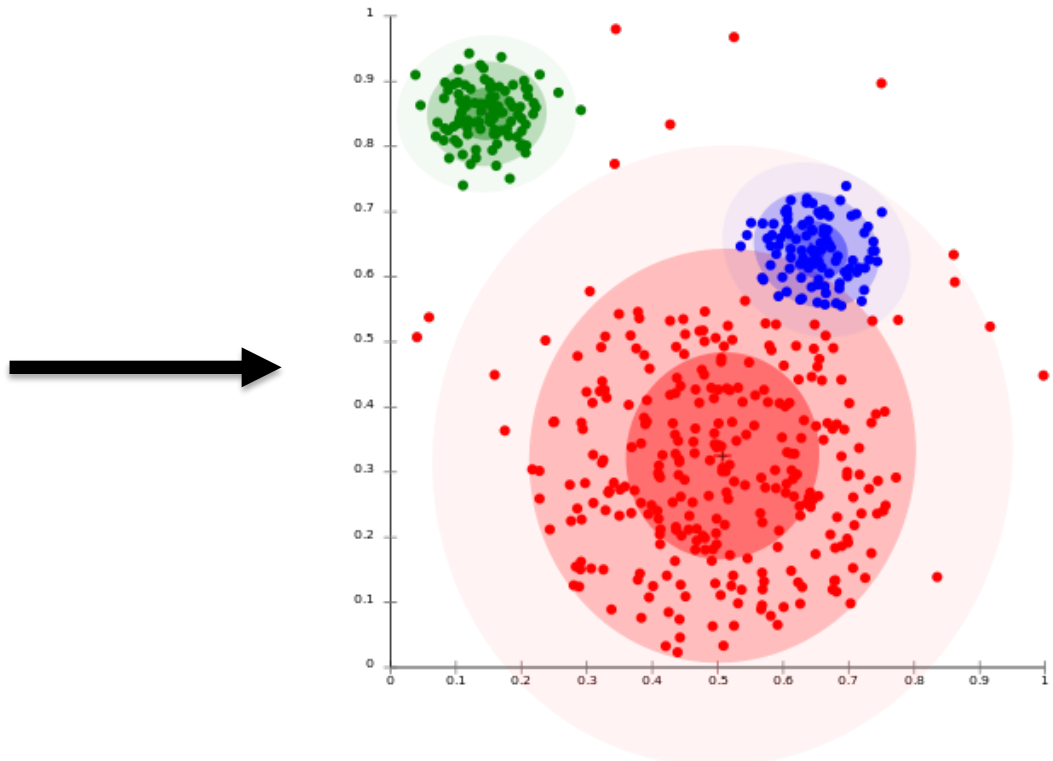
DIMENSIONAL REDUCTION EXAMPLE: A STOCK INDEX



DIMENSIONAL REDUCTION EXAMPLE: A STOCK INDEX

	continuous	***categorical***
supervised	*regression*	*classification*
unsupervised	*dimensional reduction*	*clustering*

coordinates
(continuous data)



	continuous	***categorical***
supervised	*regression*	*classification*
unsupervised	*dimensional reduction*	*clustering*

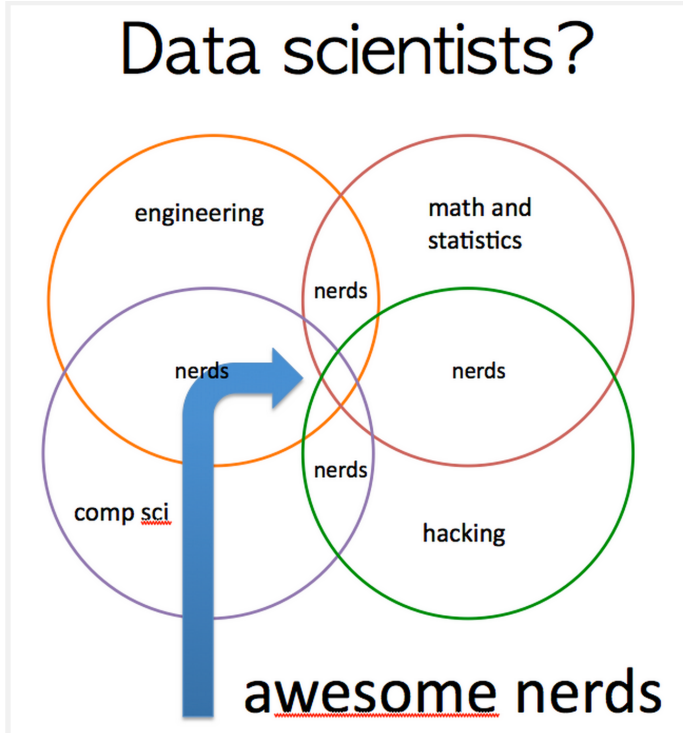
III. TRAITS OF A DATA SCIENTIST

- There are many definitions of data science
- We'll go with this one:

"Data science is the combination of analytics and the development of new algorithms," says Mason. "You may have to invent something, but it's okay if you can answer a question just by counting. The key is making the effort to ask the questions."

- What does hacking mean in this context?

- Is there anything else missing?



- Besides technical skills, attitude is also important:
    - Curiosity
    - Rigor
    - Communication skills
    - Business acumen
    - Playing well with others

- Patterns don't just present themselves

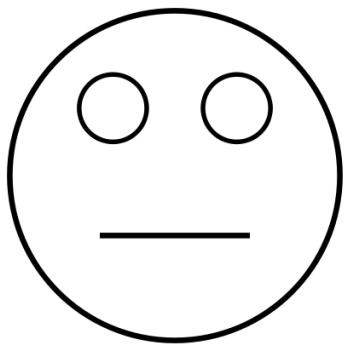
- An outlier could start an interesting line of enquiry

- Staying up to date with developments in the field

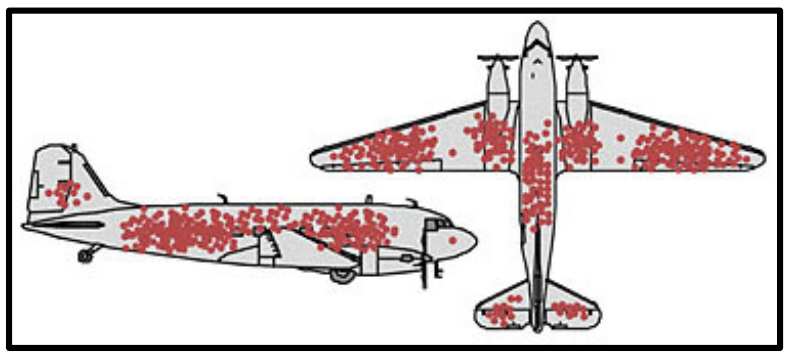
I have no special talent. I am only passionately curious. A. Einstein

- Solving a business problem using data requires
    - Knowledge of technology stack
    - Programming knowledge
    - Understanding how systems are implemented
    - Math/Stats

- Humans are hardwired to see patterns
- People give more weight to information that confirms their believes
- When sifting through large amounts of data we are bound to find patters
- It is important to tell signal from noise



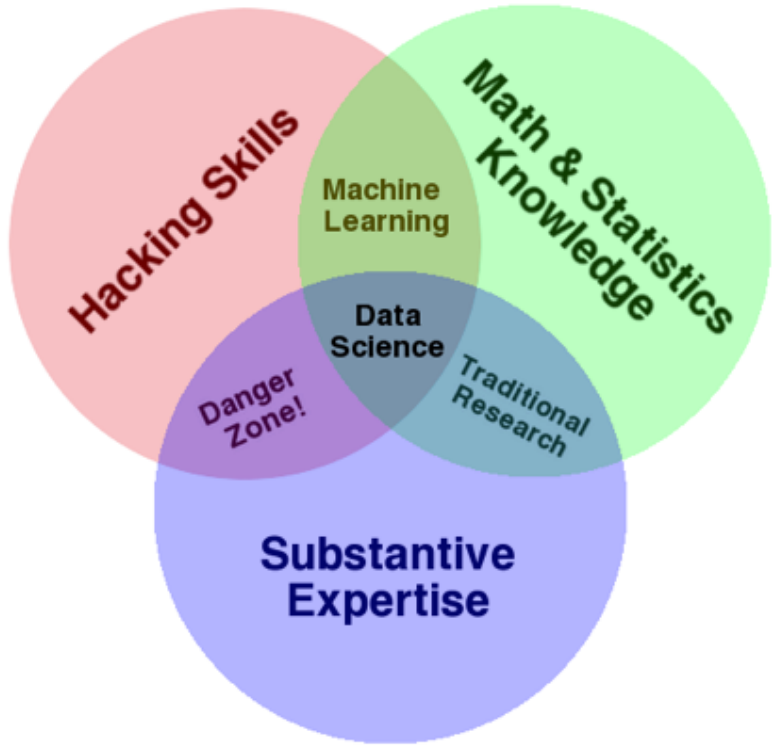
Is this a face?



- Not everyone understands hypothesis tests
- It is important to tell a story
- To do that listen, understand and explain clearly
- Data scientists need to change organizations
- This is not technical and requires persuasion skills

- Data science is about finding new things
- Of all the things we can do, which one is the most important?
- There might be something unexpected in this data, but does it matter?
- The best solution to a problem might not be practical

- There are very few people who are strong in all areas
- Each discipline is vast enough to require a life time to become a master
- Data scientists often work in teams that include: data engineering, reporting, operations, etc…
- What other roles can you think of?



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

IV. TOOLS

- Obtaining data from DBs, APIs
- Processing data
- Reproducible
- Automation

- A statistical programming language
- R started out as an implementation of S with bits of Scheme
- Developed by statisticians for statisticians
- Cutting edge algorithms available
- ggplot2

- A very useful general purpose programming language
- Rapidly growing in popularity since early 2000's
- Libraries to access DBs, APIs, do machine learning, graphing, network analysis, natural language processing, web dev, etc…
- For reference it is interpreted, multi-paradigm and dynamically typed

- Relational databases

- SQL – Structure Query Database

- The main DB tech for decades

- What has changed?

- Emergence of web scale data

- Distributed, large scale, non structured data

- Comes out of an effort to create an open source search engine
- Yahoo! Was an important contributor and a large user
- It is not a database but a data storage and management system
- Data extracted through MapReduce

- A science of data
- Predates computers
- Emphasizes formal statistical inference (confidence intervals, hypothesis tests) in low dimensionality

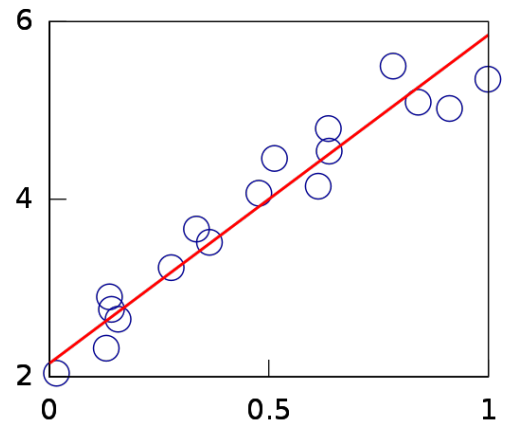
If a statistician presents an estimate to a journalist and says "here is the point estimate of the number of people listening to a given radio station and states that the margin of error is +/- 3% with a 90% confidence interval" there is almost always a follow-up discussion about the margin of error and how the standard error was calculated (simple random, stratified, cluster) why is it a 90% confidence interval rather than a 95% confidence interval. And then someone is bound to ask what a confidence interval is anyway? Then extend this even further and the statistician gives the journalist a p-value? Now there is an argument between statisticians about hypothesis testing and the terms "frequentist" and "Bayesian" start getting thrown around.

- Very close to statistics
- A Computer Sci discipline
- There is an algorithmic component
- Many concepts are similar to stats but have a different name

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant = $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

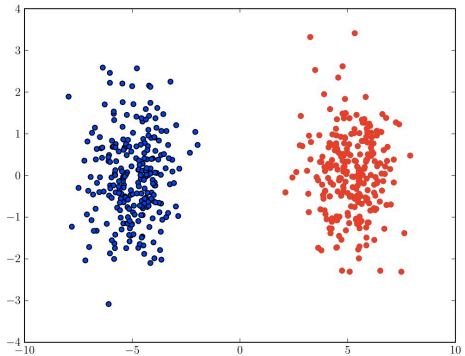
- Has been called the workhorse of data science
- Allows us to characterize the relationship between variables
- Can be used to build a model through which we can predict values of *y* given observations *x*

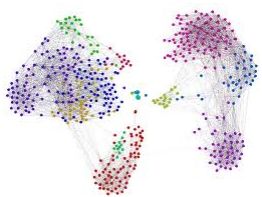


- Are there natural groups in my data?

- Applications include clustering

- Unsupervised learning technique

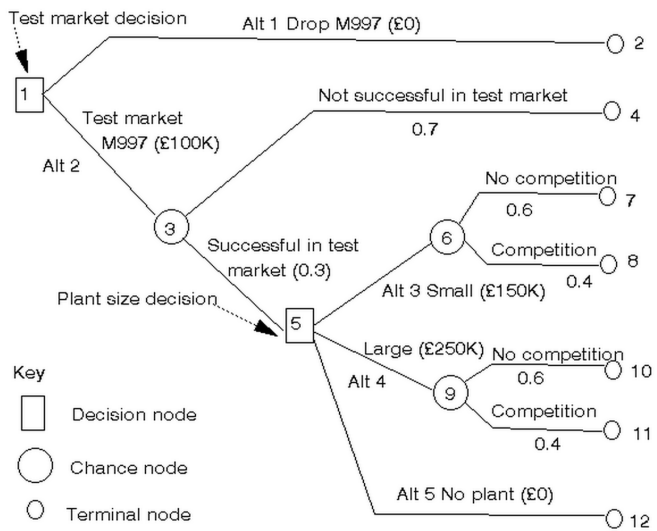




– Builds on decision trees

– What if instead of one tree we had many (ensemble learner)?

– Tells us what category something is in – *classifier*

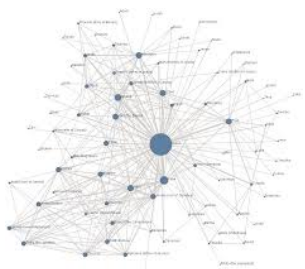


- Data from Twitter, calls among mobile network users

- What is the shortest distance between any two users?

- Are there communities?

- How does product adoption spread among friends?



- Extracting information from text

- What emotion is the customer expressing in this message?

- What are the topics mentioned on these webpages?

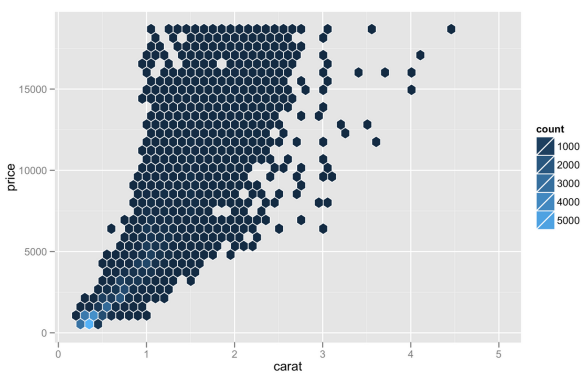
- Which section of this text is of type *x*?

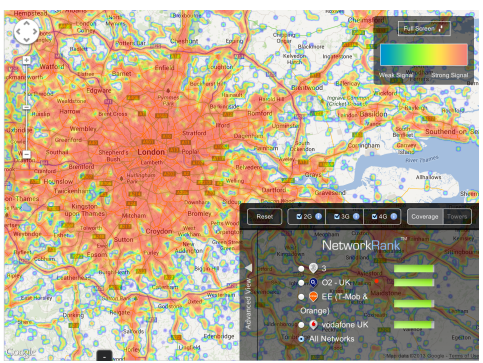
- Machine learning as a service
    - Prediction APIs: wise.io, Google Predictions
    - Entity extraction Open Calais

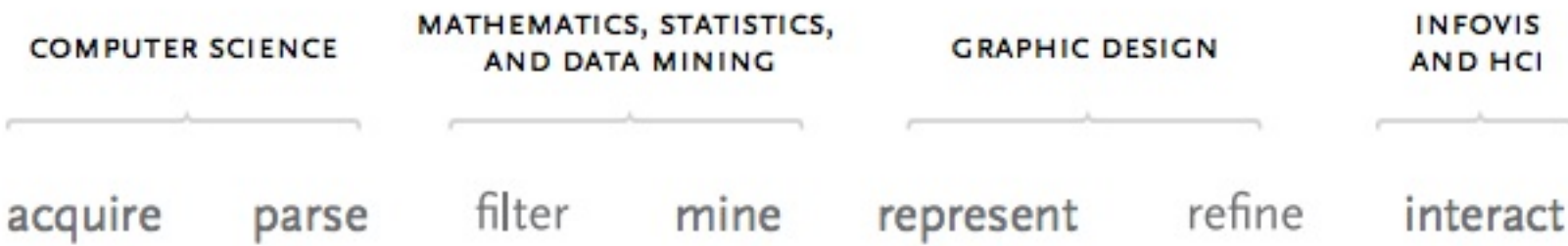
- Cloud Computing
    - Send your code and data run on 100's of machines
    - Spin up powerful servers to run your computations

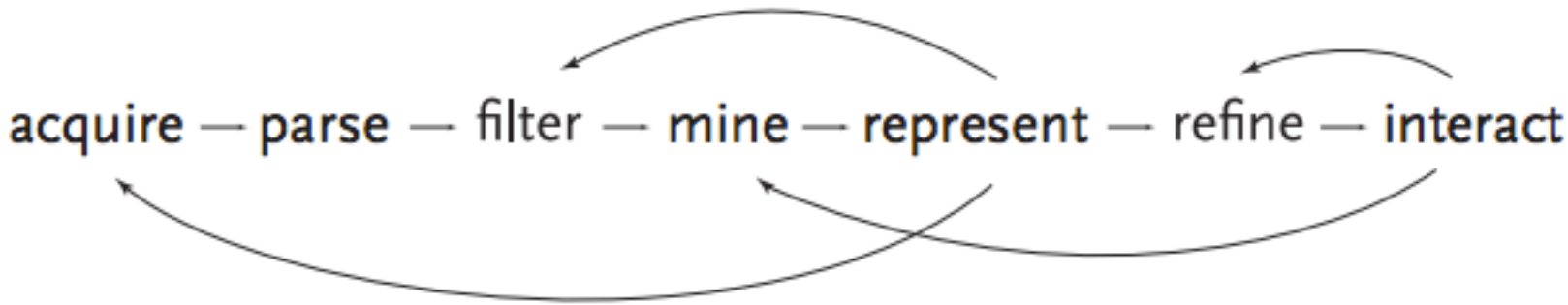
- Exploratory data analysis

- Communicating findings





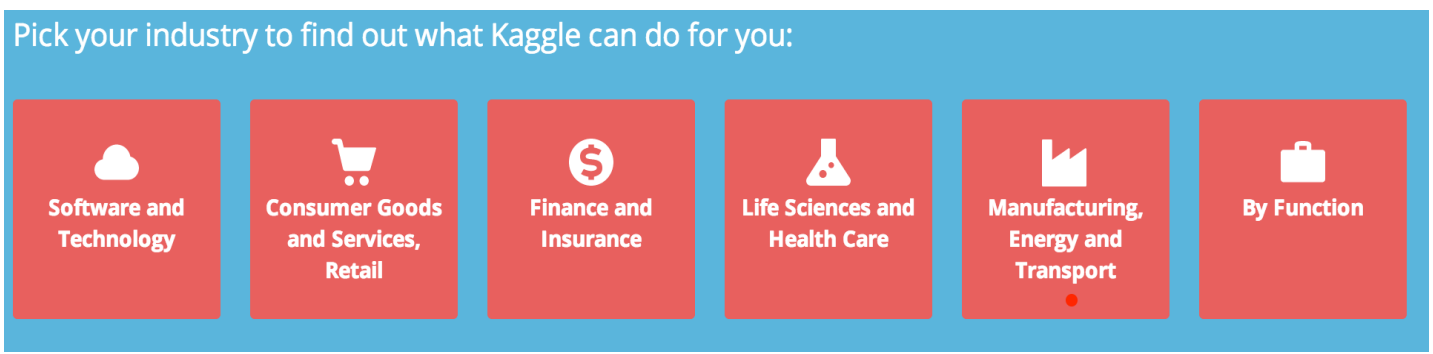

COMPUTER SCIENCE
MATHEMATICS, STATISTICS, AND DATA MINING
GRAPHIC DESIGN
INFOVIS AND HCI
acquire
parse
filter
mine
represent
refine
interact


acquire — parse — filter — mine — represent — refine — interact

V. OPPORTUNITIES

- The organization you work for probably generates more data than it ever has
- Is the value from each data source exhausted?
- Is data from multiple source combined?
- Are users aware of the possibilities offered by social network analysis, natural language processing or clustering?

- Crowdsourcing analytics problems
- Thousands compete by using whatever methods to produce best predictions
- Cash prizes



kaggle

- In April 2012 McKinsey predicted 1.5 million shortage of data scientists
- More and more companies are looking for people to unlock the value in their data
- Rise in available positions

Location London	3 months to 16 Aug 2013	Same period 2012	Same period 2011
Data Scientist			
Rank	566	631	-
Rank change year-on-year	▲ +65	● -	
Permanent jobs requiring a Data Scientist	41	11	0
As % of all permanent IT jobs located in London	0.091%	0.021%	-
As % of the Job Titles category	0.097%	0.023%	-
Number of salaries quoted	31	10	0
Average salary	**£55,000**	£65,000	-
Average salary % change year-on-year	-15.38%	-	
UK excluding London average salary	£60,000	£85,000	£50,000
% change year-on-year	-29.41%	+70.00%	

- Software development becomes commoditized
- Many not very technical ideas only need a WordPress install
- Many new companies are differentiate themselves through their use of data
- Point in case: “EDITD: Using bid data to demystify fashion trends”

- What is now called data science is not new
- The pharma industry has been using similar tools and techniques
- However there broader healthcare industry still has some catching up to do

- What is now called data science is not new
- The pharma industry has been using similar tools and techniques
- However there broader healthcare industry still has some catching up to do
- Marketers have used segmentation and churn for years
- However with the move to digital, marketing is becoming more analytical

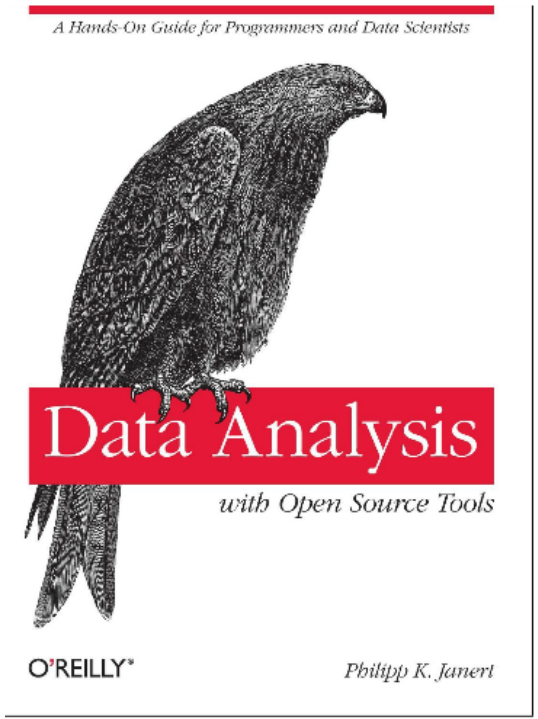
- Government compile statistics about schools

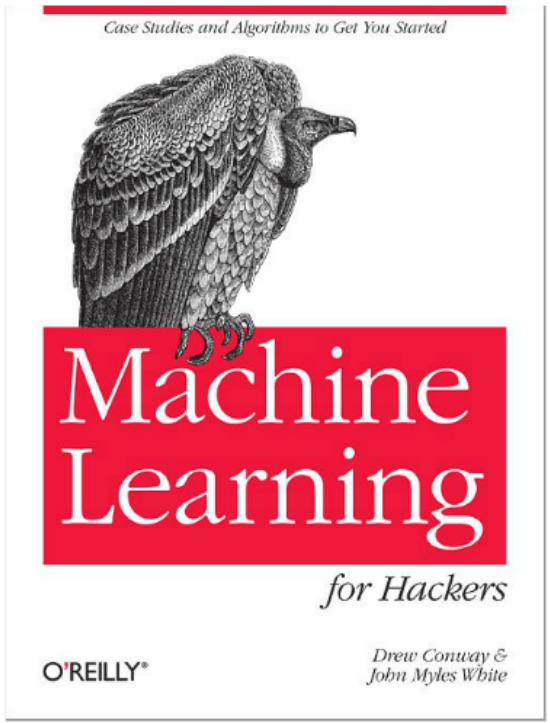
- Online education allows to track each student's progress and tailor the material

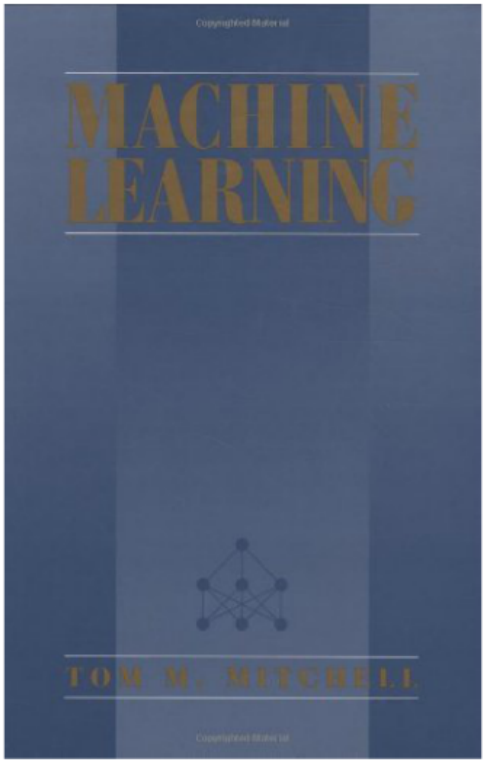
- Online teaching materials supplement normal classes and generate data

- Many companies struggle to recruit in this area
- Traditional analysts too focused on specific tools
- Many programmers don’t have business experience
- Because the field is new there are few people with leadership skills

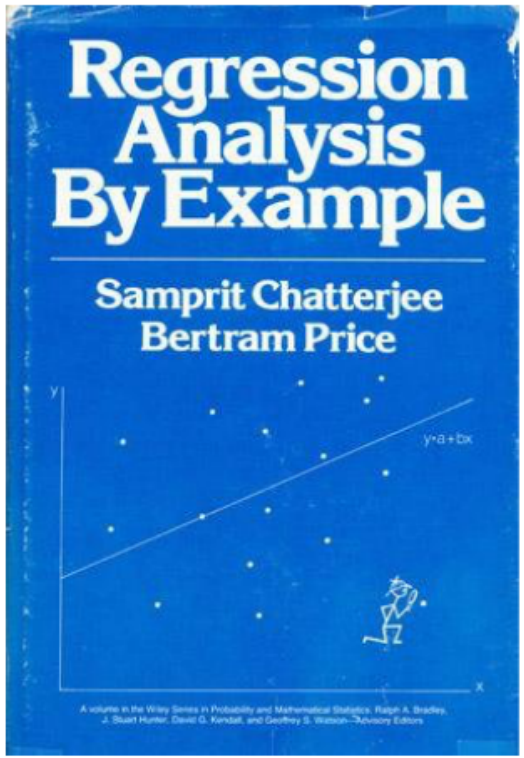
- Mobile devices are generating data which is already being collected - http://opensignal.com/
- Internet of things – all devices will become computerized, constantly connected and generating data
- Quantified self – Tracking almost every aspect of someone's data requires data skills to generate actionable insight.

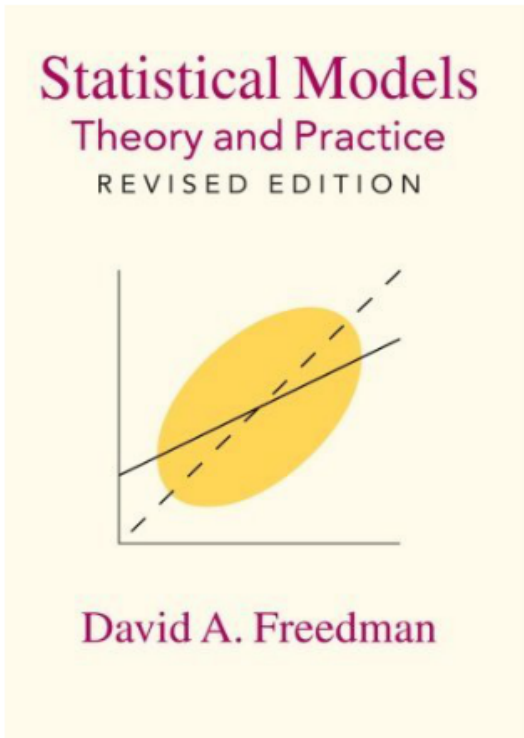

A Hands-On Guide for Programmers and Data Scientists
Data Analysis
with Open Source Tools
O'REILLY
Philipp K. Janert

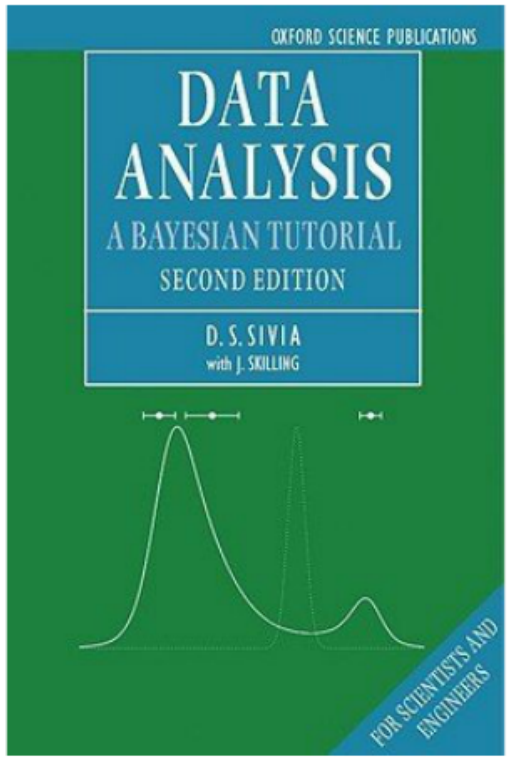

Case Studies and Algorithms to Get You Started
Machine Learning
for Hackers
O'REILLY
Drew Conway &
John Myles White


MACHINE LEARNING
TOM M. MITCHELL

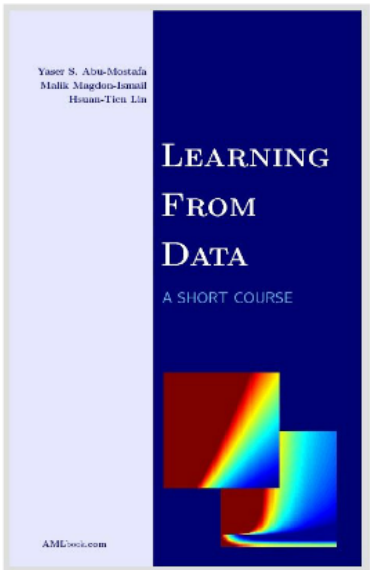
BOOKS – PARAMETRIC METHODS

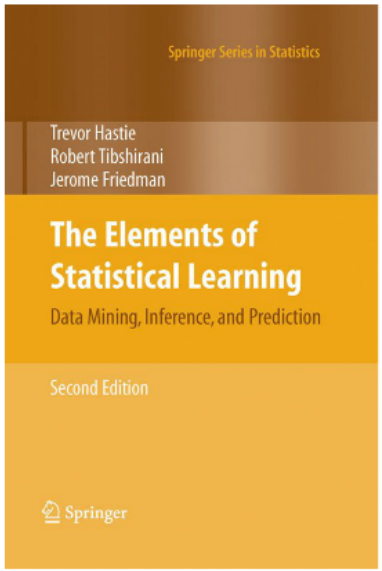


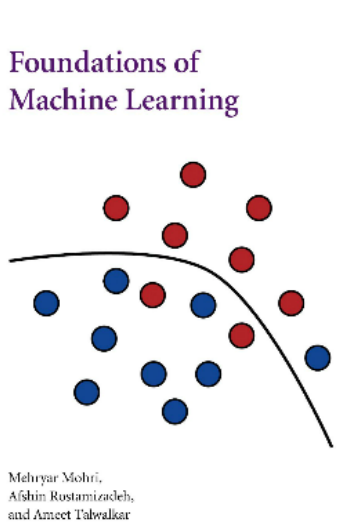


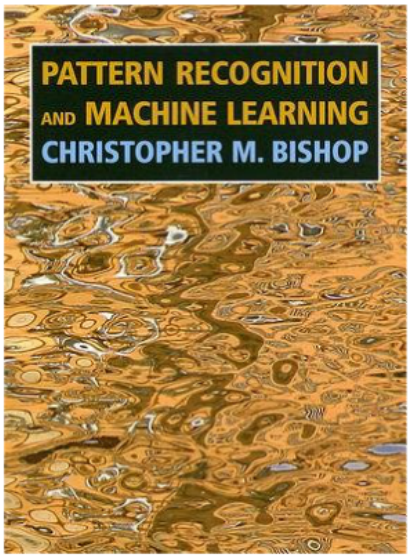


BOOKS – MACHINE LEARNING THEORY

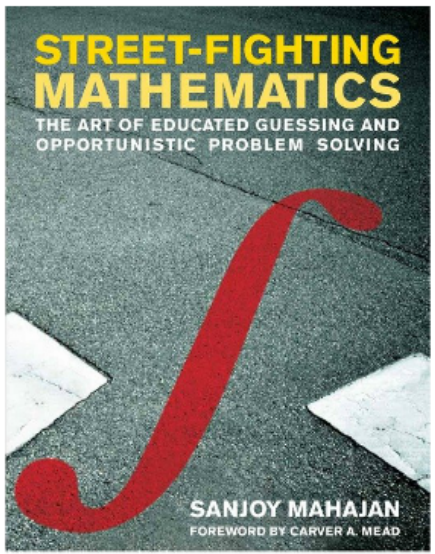


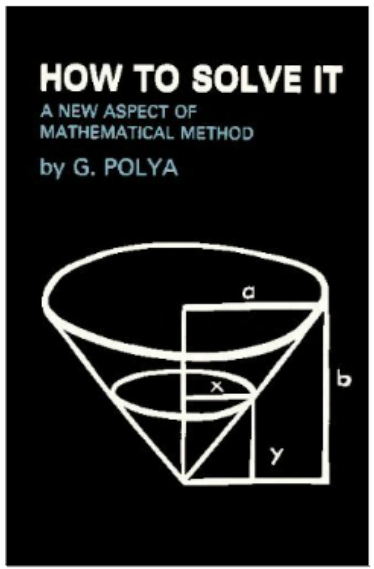


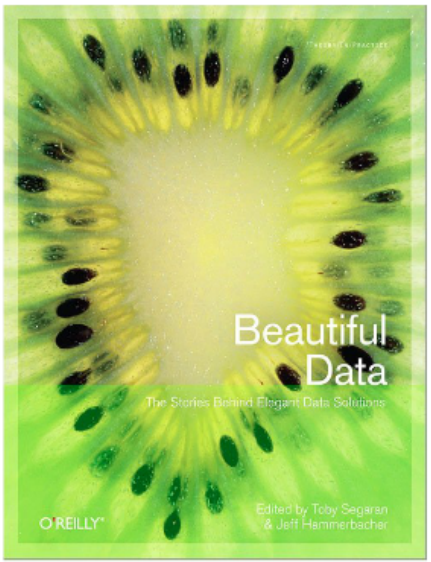


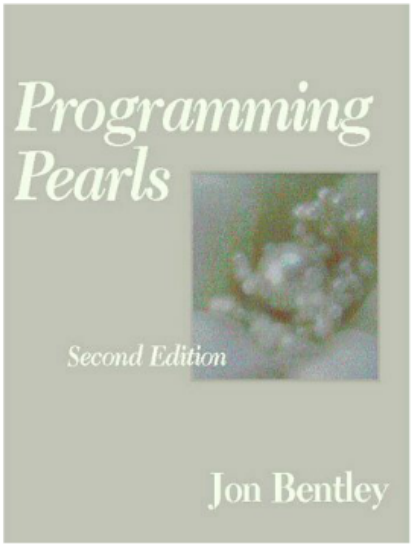


BOOKS - PROBLEM SOLVING

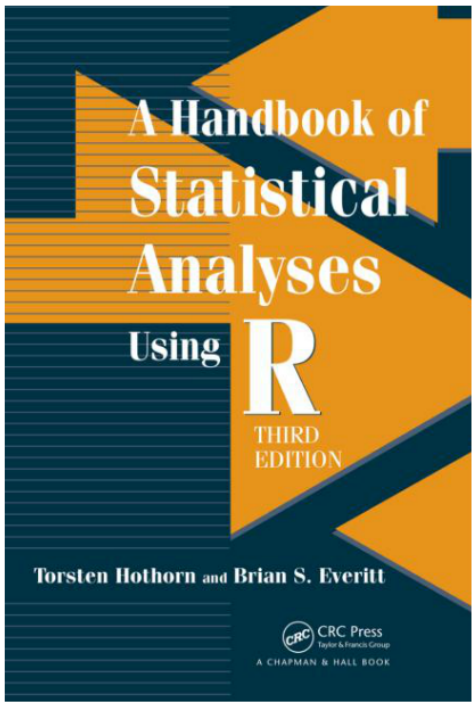


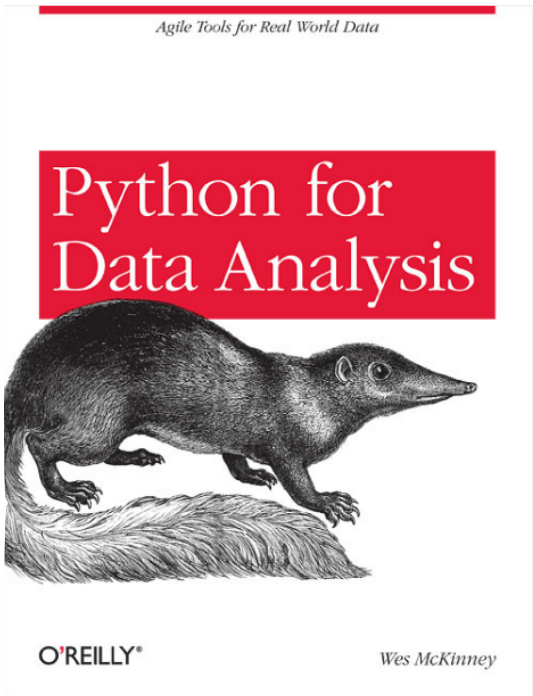


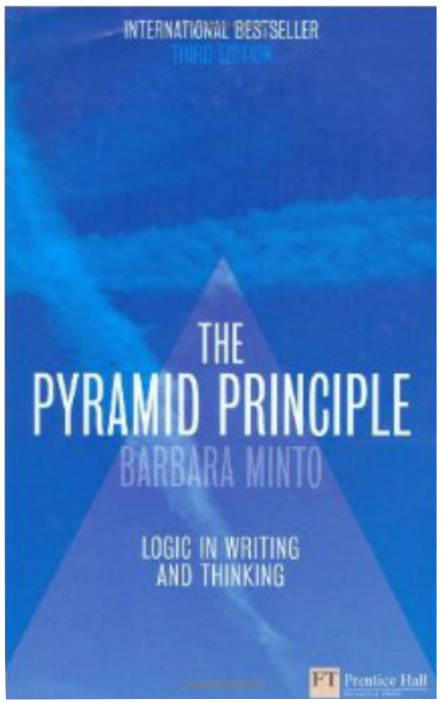




BOOKS – JOHN'S TOP 3 RECOMMENDATIONS







https://news.ycombinator.com/

http://www.r-bloggers.com/

http://www.johnmyleswhite.com/

http://www.hilarymason.com/

http://blog.echen.me/

http://andrewgelman.com/

http://hunch.net/

http://conductrics.com/data-science-resources/

http://fivethirtyeight.blogs.nytimes.com/

https://www.coursera.org/course/ml Stanford ML course

http://www.youtube.com/playlist?list=PLD63A284B7615313A Caltech ML course

http://videolectures.net/Top/Computer_Science/Machine_Learning/

http://www.hilarymason.com/tag/video/

http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/

http://harvarddatascience.com/

http://www.autonlab.org/tutorials/ (Andrew Moore tutorials page)

http://www.columbia.edu/~lah2178/index/Courses.html (Lauren Hannah courses)

- Data Skeptic
- Partially Derivative
- Linear Digressions
- More or Less
- O’Reilly Data Show

- PyData London
- LondonR
- Data Science Meetup London
- Big Data London
- London Machine Learning Meetup
- Quantified Self
- Predictive Analytics London Meetup
- Data Visualization Meetup
- PyLadies London
- Women in Data
- Londata
- Data Science Journal Club

- DataKind
- NHS Hack
- Kaggle
- UK Hackathons & James Meetup
- StartupWeekend
- Code for Good

1. Learn to code

Python. R. Professional software engineering practices.

2. Get statistical

Significance. Inference. Regression. Machine learning.

3. Learn lean

Business skills. Startup methodology. Communication.

4. Experience

Side projects. Github. Kaggle. Hackathons. Stand out.

- Data science is a product of our time

- Being a data scientists requires people and technical skills

- We’re only getting started…

Q&A