PyData

Bristol 2018

A Brief Introduction to Data Science, Machine Learning and the PyData Ecosystem

John Sandall
Data Science & Engineering Consultant
@john_sandall

15th March 2018

AGENDA

I. What is Data Science?

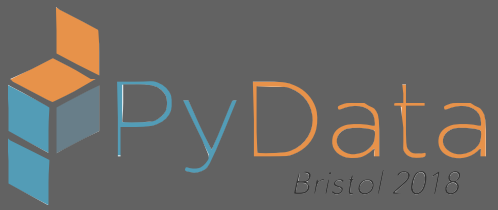
II. The Last 10 Years

III. Machine Learning 101

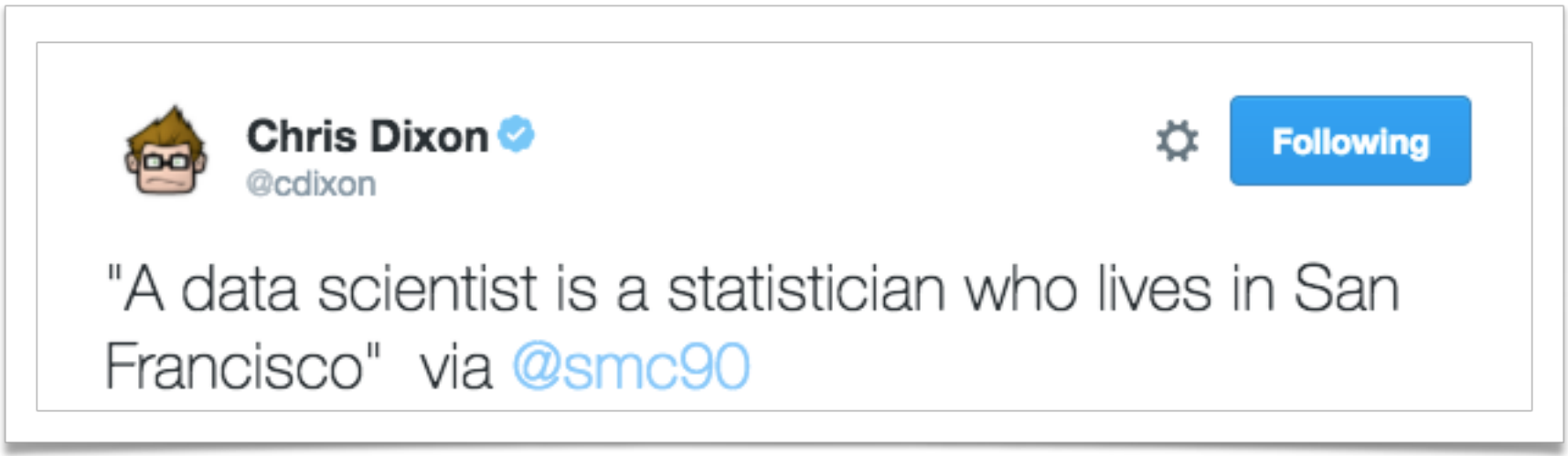
IV. Traits of a Successful Data Scientist

V. Opportunities

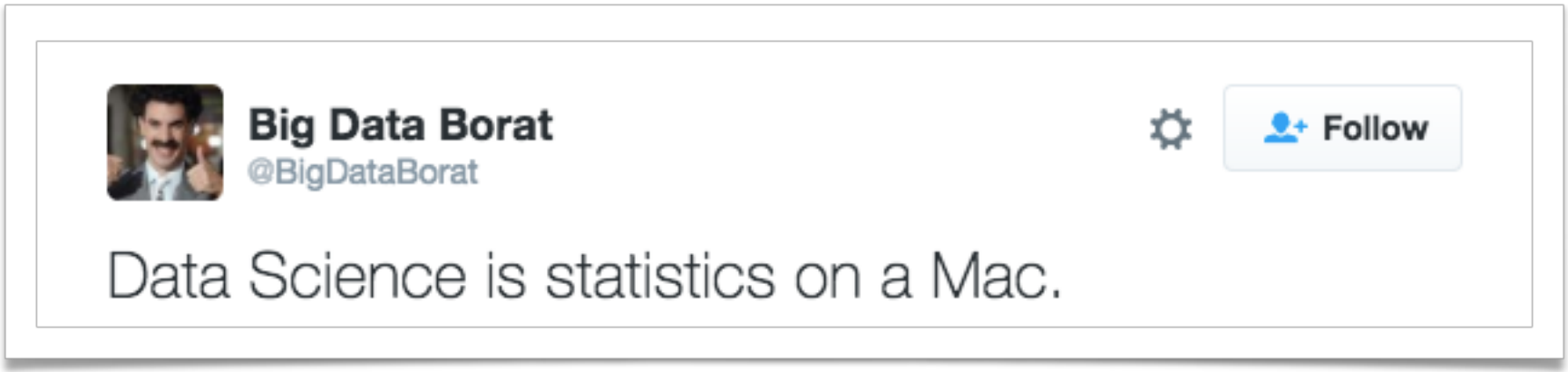
PART I.
WHAT IS DATA SCIENCE?



WHAT IS DATA SCIENCE?



WHAT IS DATA SCIENCE?



WHAT IS DATA SCIENCE?

(((Josh Wills)))
@josh_wills

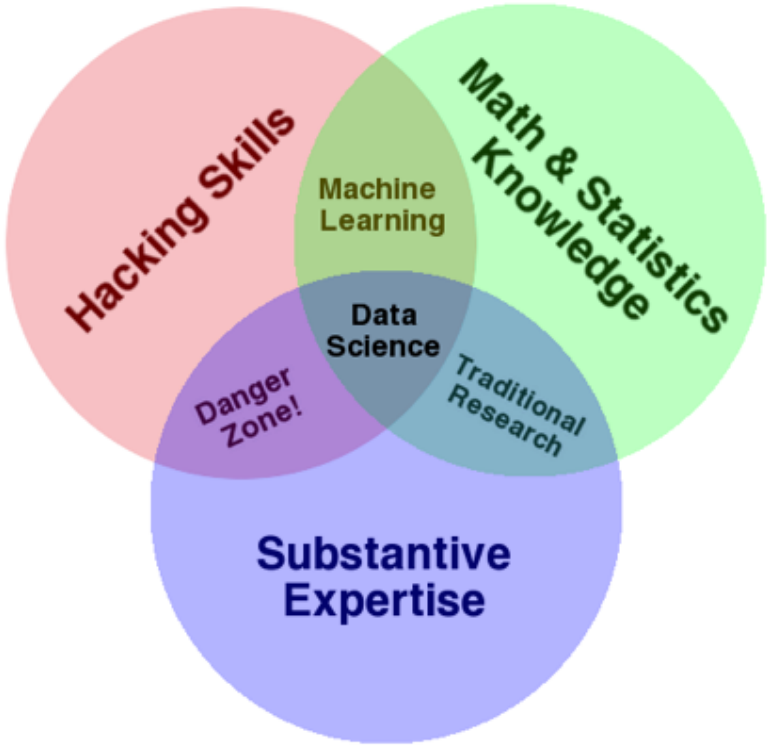
Follow

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

WHAT IS DATA SCIENCE?

- A set of **tools & 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.

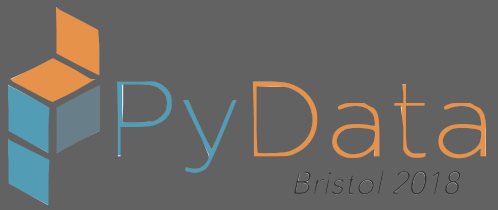
THE QUALITIES OF A DATA SCIENTIST



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

EARLY ADOPTERS OF DATA SCIENCE & ENGINEERING

PART II.
THE LAST 10 YEARS

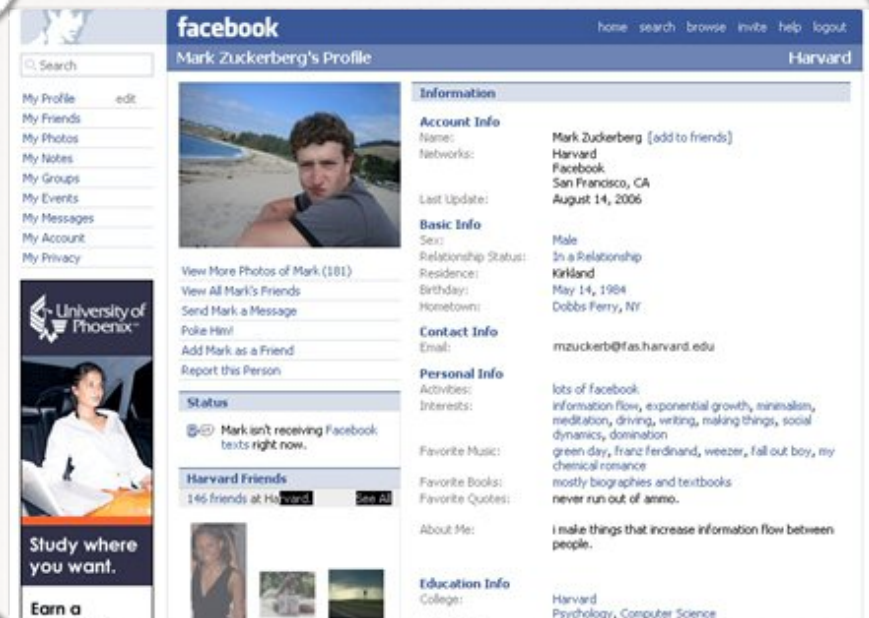


2007: A PIVOTAL YEAR

iPhone released

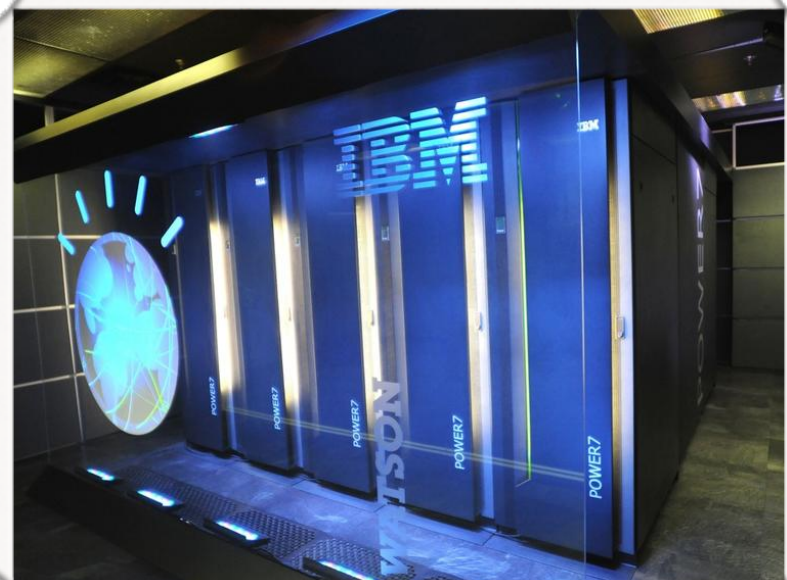
Android launches

2007: FACEBOOK & TWITTER BOTH GO GLOBAL



2007: INFORMATION REVOLUTION

Kindle launches



IBM Watson created

2007: THE OPEN SOURCE ECOSYSTEM ACCELERATES

- Not only hardware costs are dropping

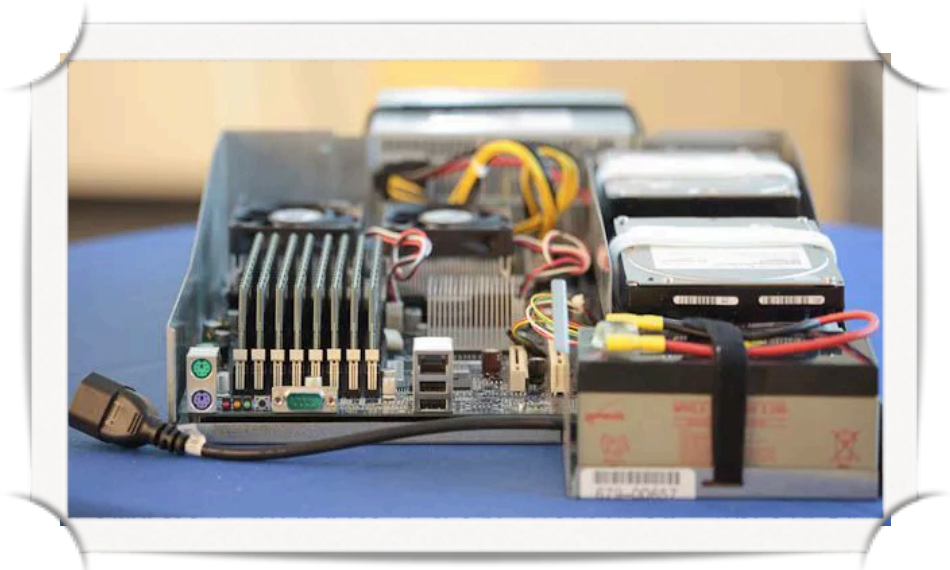
- Open Source – Linux, Apache, Hadoop, MySQL

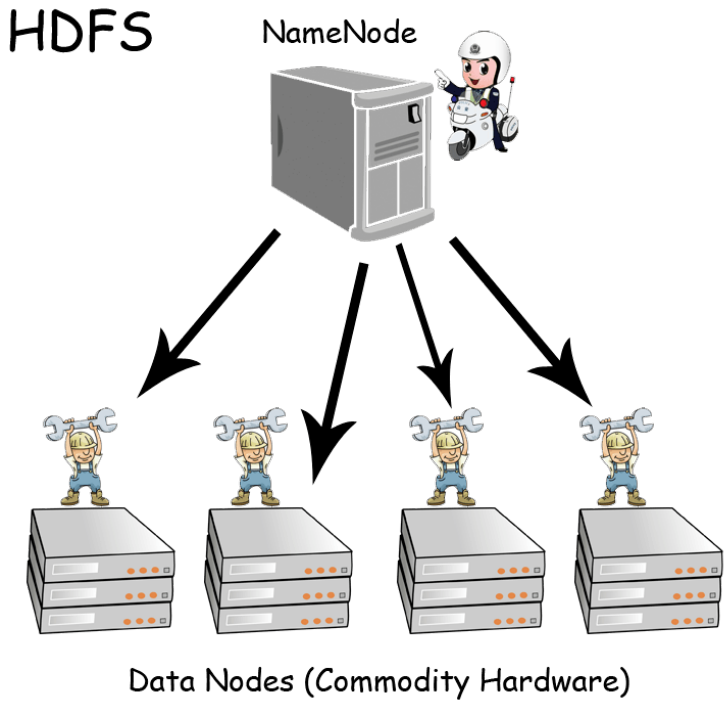
- In addition to infrastructure there are also analysis tools

- Who knows R or Python?

2007: R & PYTHON BECOME ENTERPRISE FRIENDLY

2007: THE BIG DATA REVOLUTION BEGINS

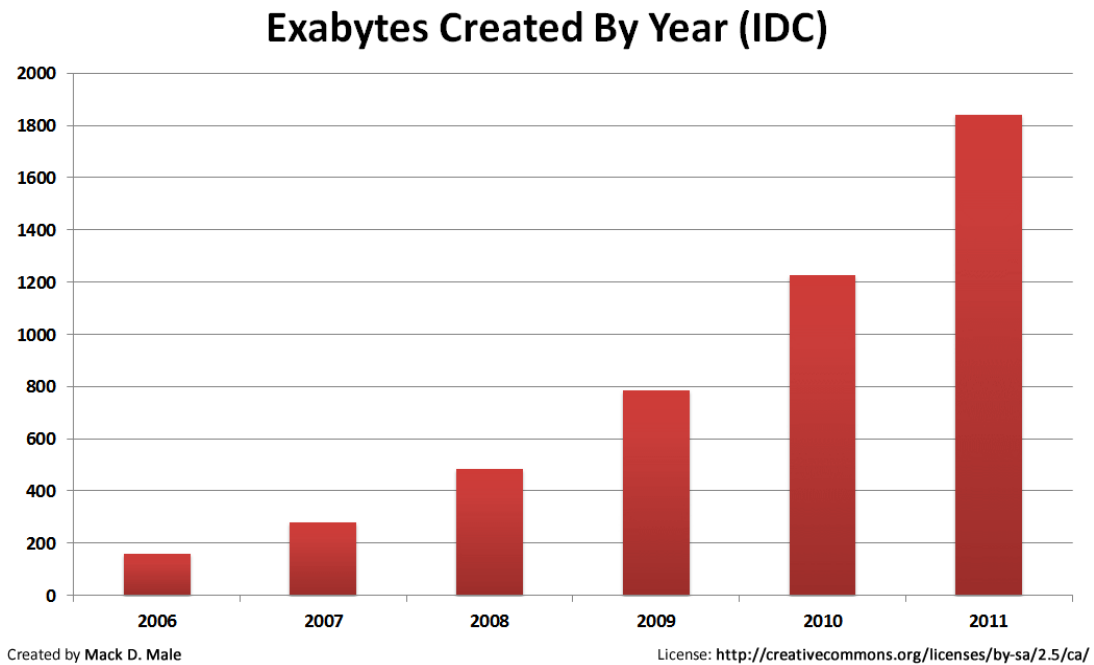




EVOLUTION OF DATA CREATION

2011:

Every two days we create more information than we did up until 2003 (around two exabytes).

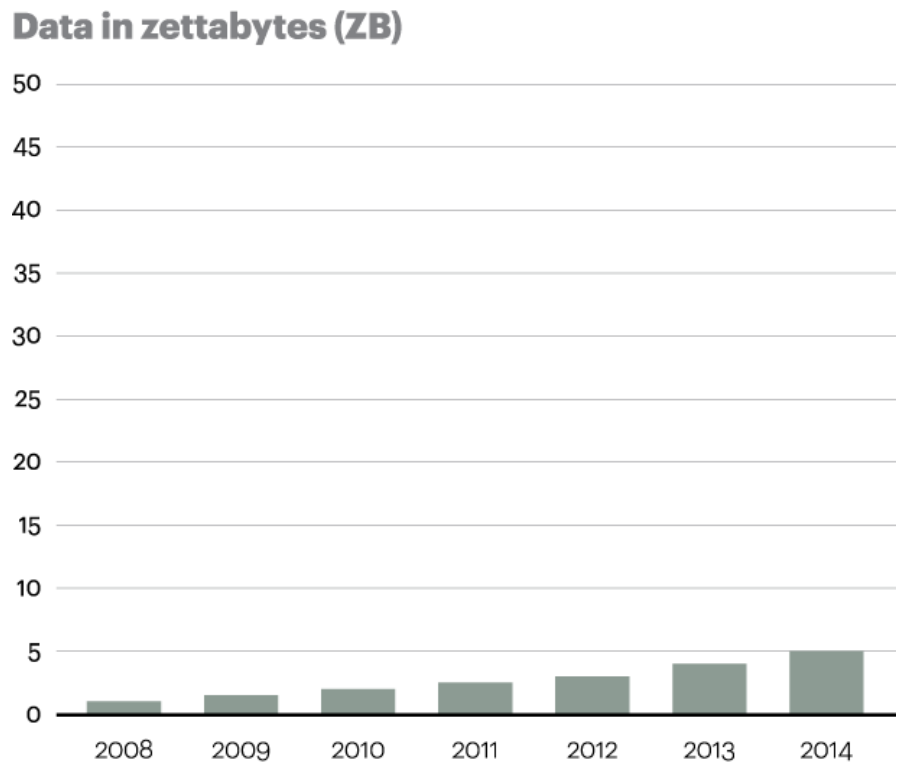


1 exabyte (EB) = 1000 petabytes (PB) = 1 billion gigabytes (GB)

EVOLUTION OF DATA CREATION

2014:

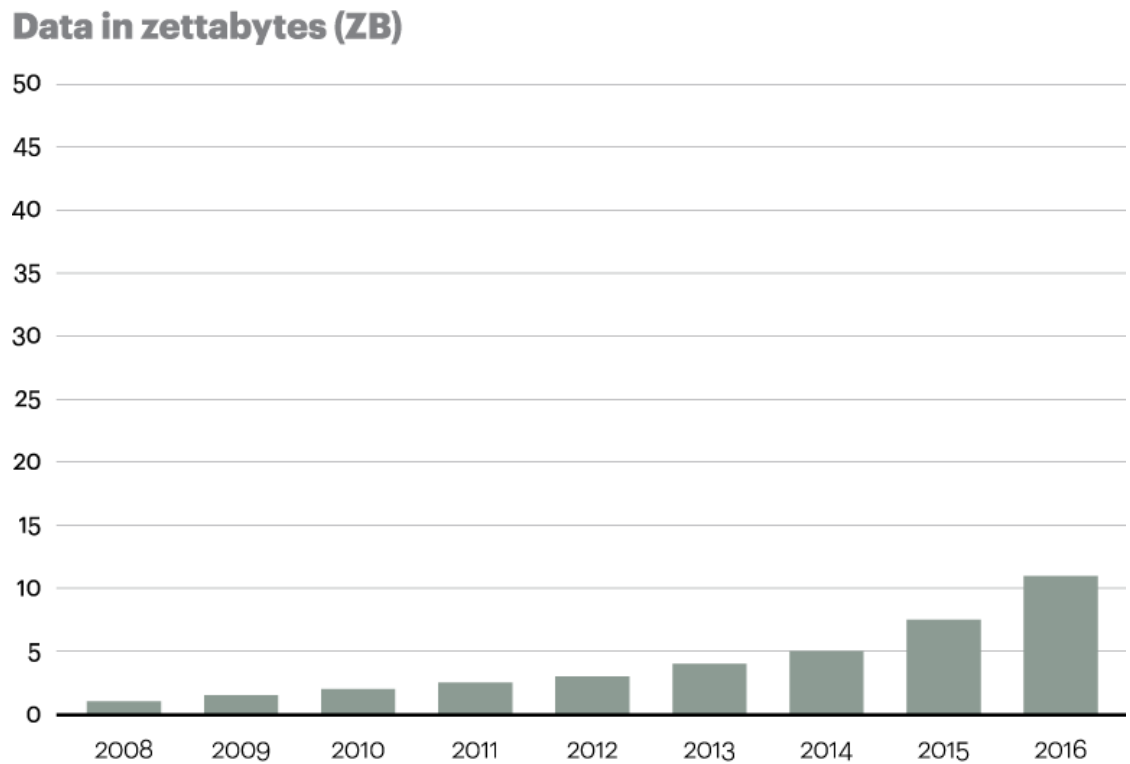
Oracle estimates total data created annually now surpasses five Zettabytes



EVOLUTION OF DATA CREATION

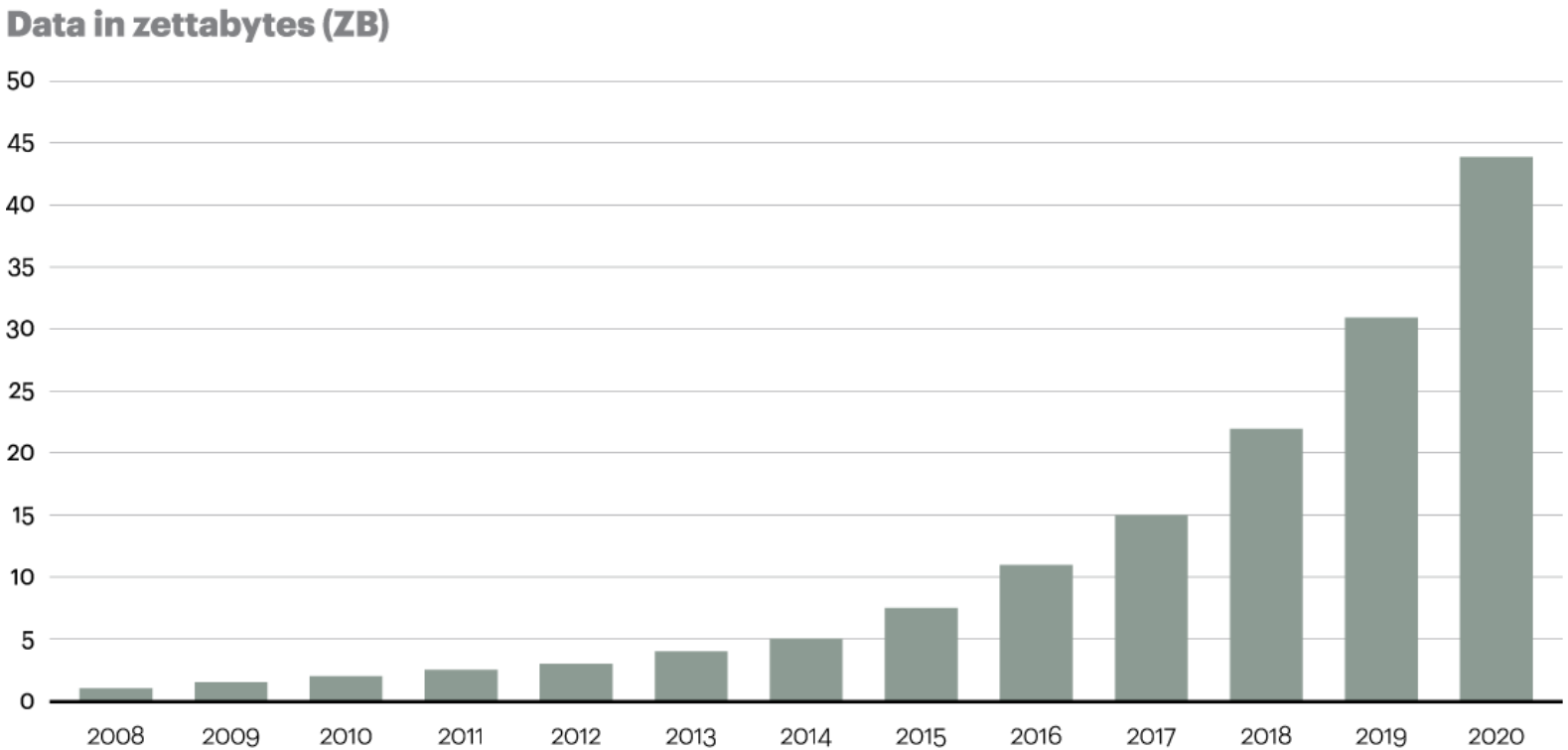
2016:

Data is growing at 40 percent compound annual rate, now hitting over 10ZB annually

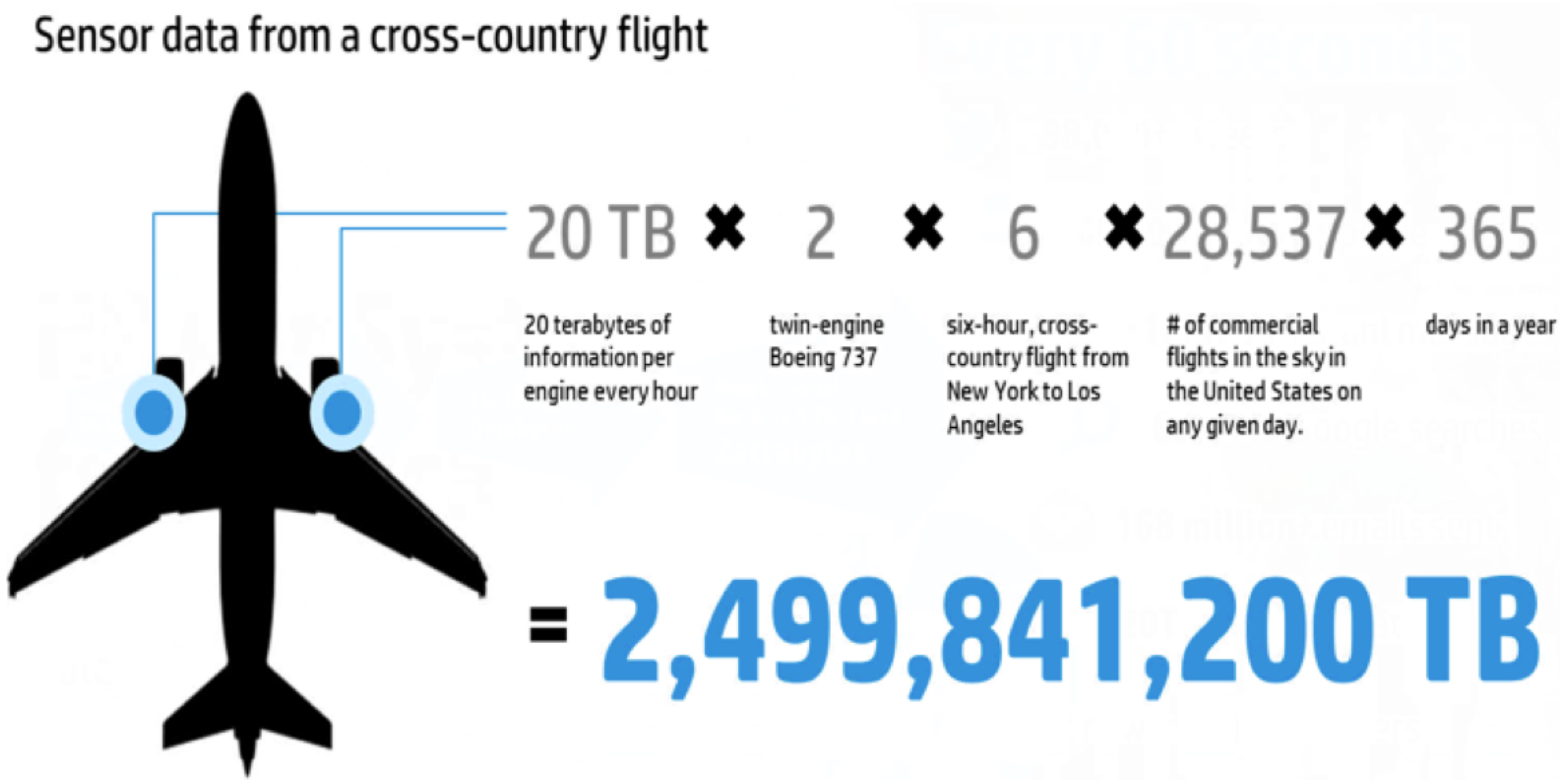


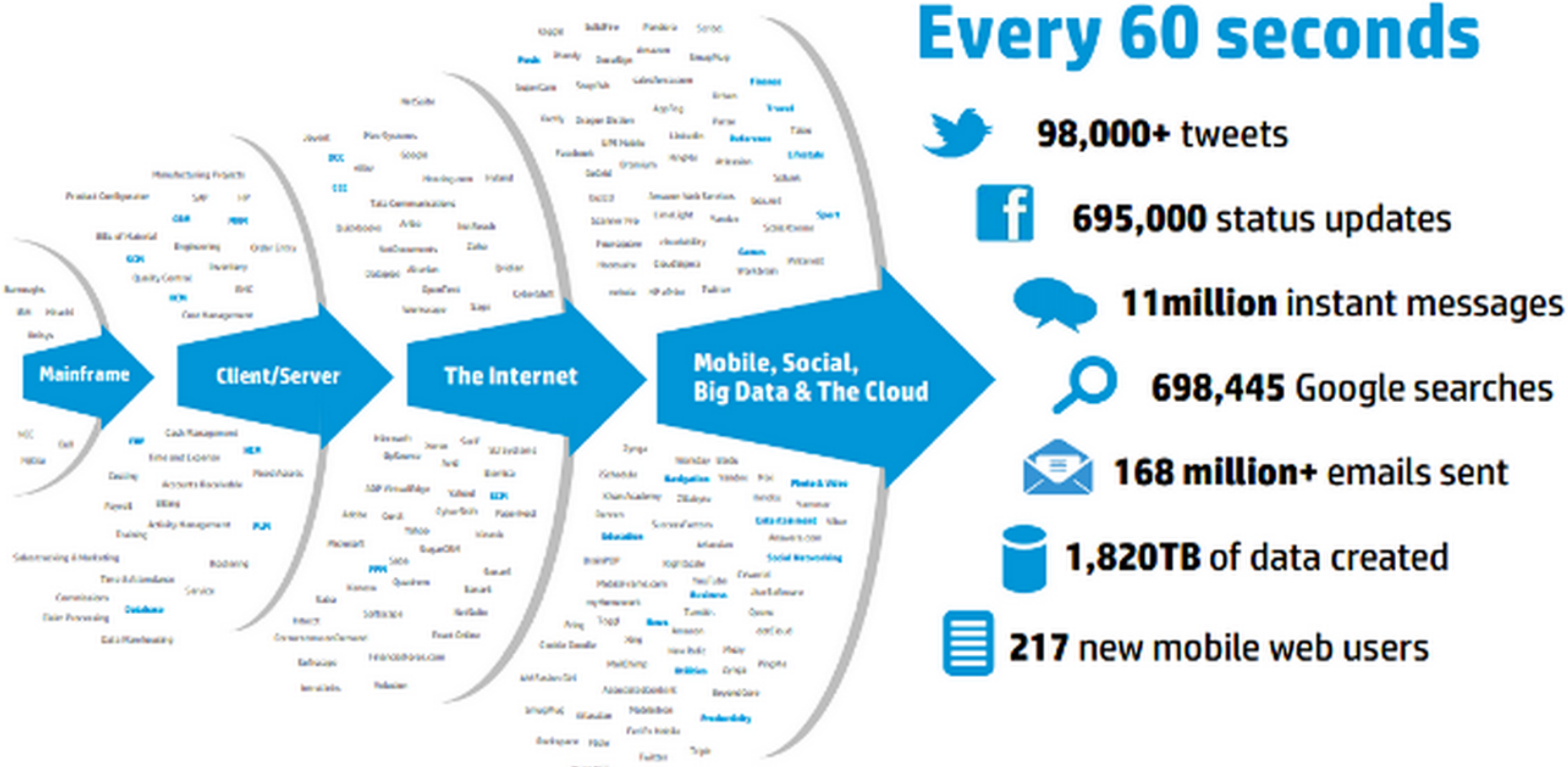
EVOLUTION OF DATA CREATION

Forecasts suggest annual data creation will hit nearly 45ZB by 2020



WHERE IS DATA COMING FROM?





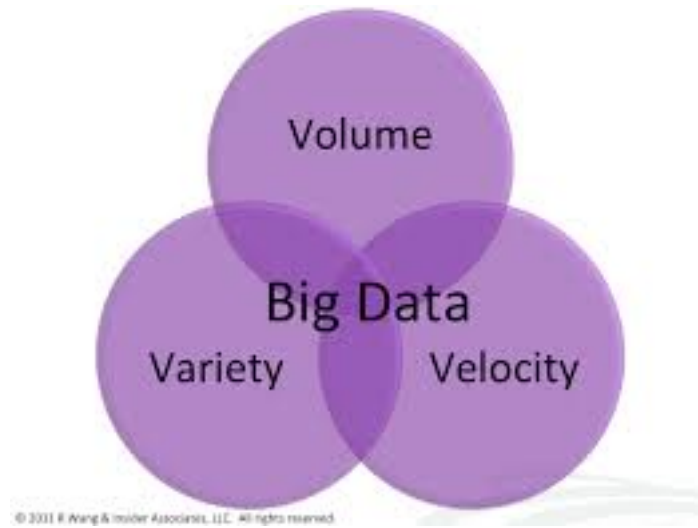
Every 60 seconds

- **98,000+** tweets
- **695,000** status updates
- **11million** instant messages
- **698,445** Google searches
- **168 million+** emails sent
- **1,820TB** of data created
- **217** new mobile web users

http://www.datasciencecentral.com/profiles/blogs/basic-understanding-of-big-data-what-is-this-and-how-it-is-going

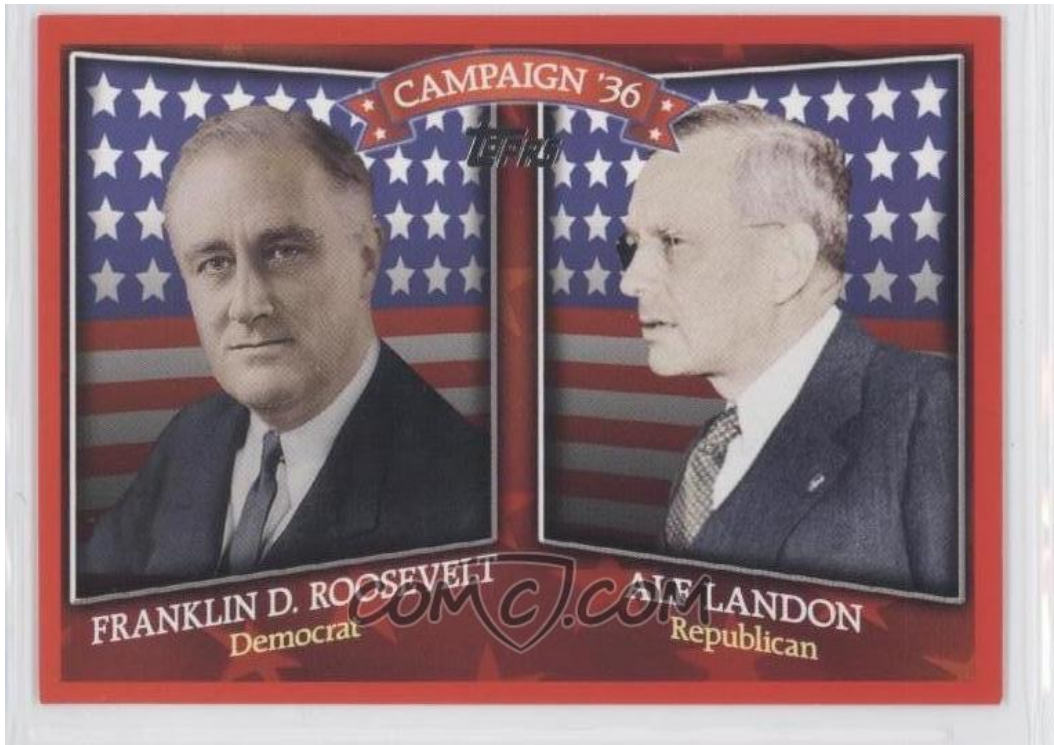
BIG DATA

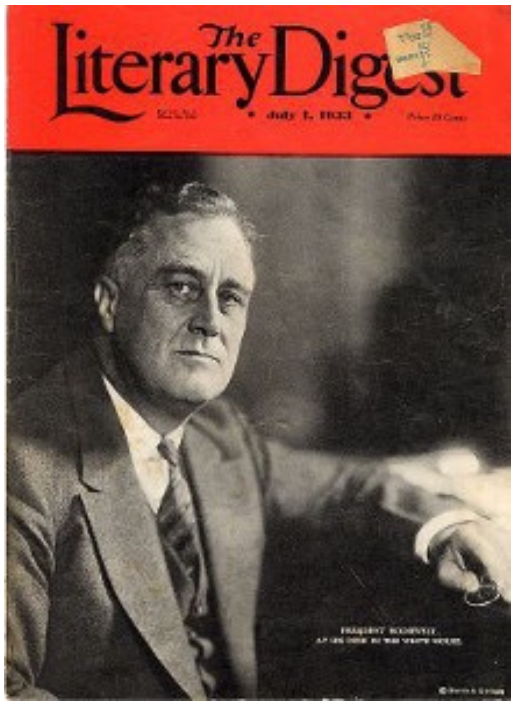
- It describes data that:
  - doesn't fit in memory
  - doesn't fit on a machine



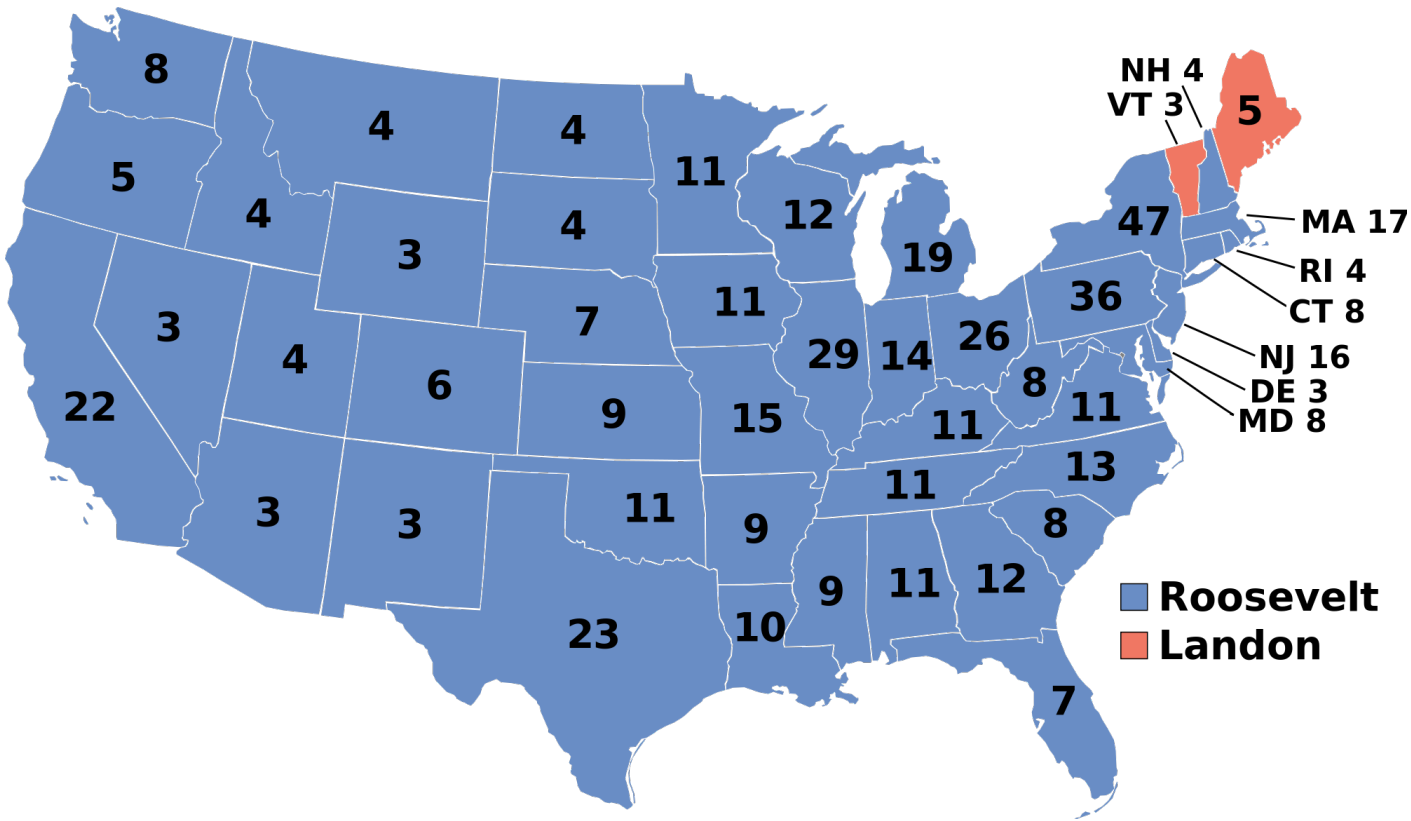
From: http://blog.softwareinsider.org/2012/02/27/mondays-musings-beyond-the-three-vs-of-big-data-viscosity-and-virality/screen-shot-2012-02-27-at-11-18-56-am/

BIG DATA: A CAUTIONARY TALE





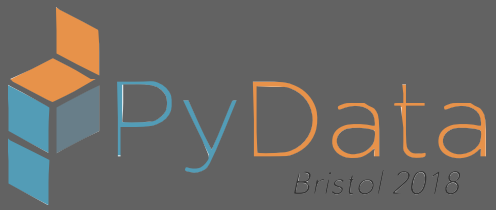
BIG DATA: A CAUTIONARY TALE



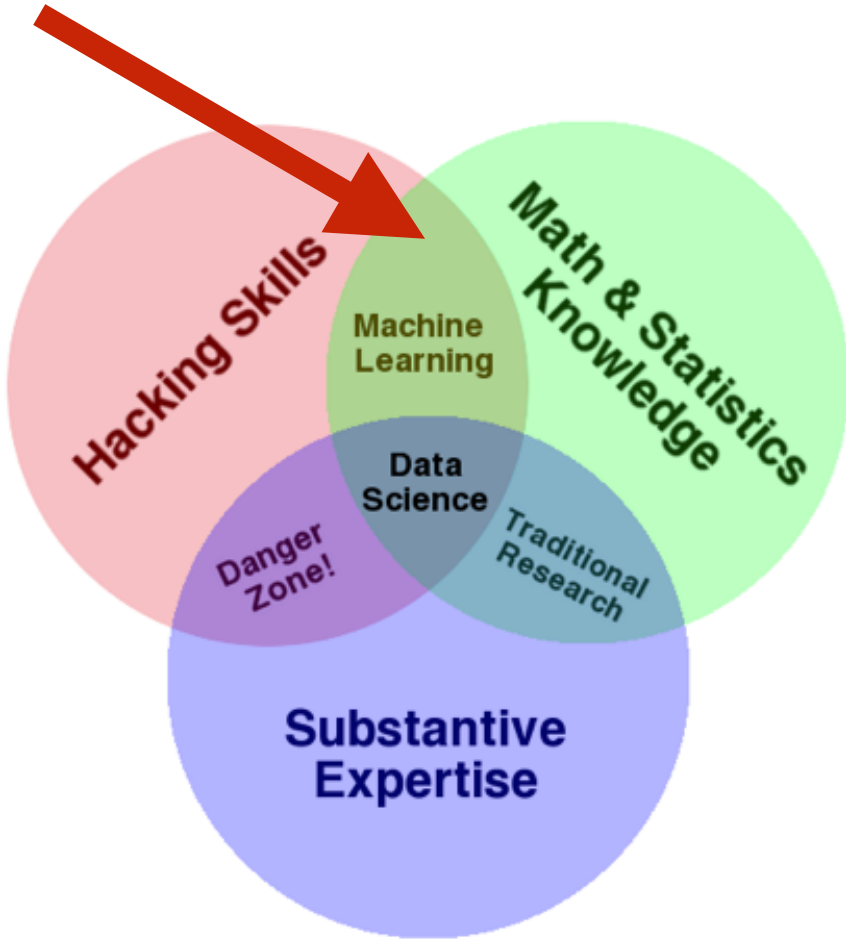
OTHER APPLICATIONS OF DATA SCIENCE & MACHINE LEARNING

1. Search engines
2. Recommendation systems
3. Image recognition
4. Speech recognition
5. Gaming
6. Price comparison/optimisation
7. Route planning (driving, airlines, social network virality!)
8. Fraud / risk detection
9. Logistics (deliveries of goods, of people, of data)
10. Self-driving cars
11. Robots & AI assistants
12. ...

PART III.
MACHINE LEARNING



YOU ARE HERE!


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

WHAT IS MACHINE LEARNING?

From Wikipedia:

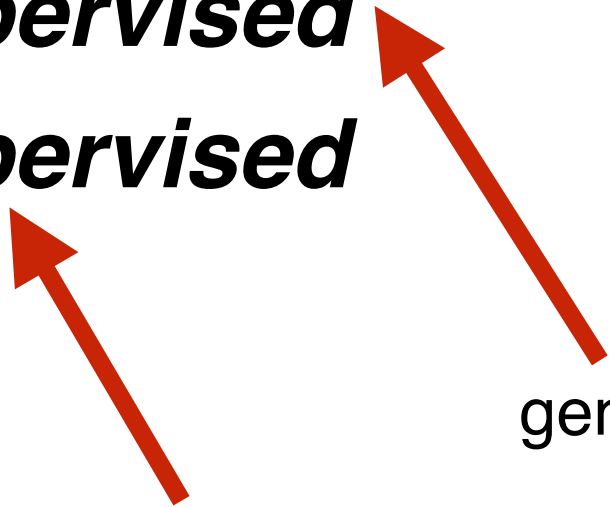
- "Machine learning, a branch of **artificial intelligence**, is about the construction and study of systems that can **learn from data**."

- "The core of machine learning deals with **representation** and **generalisation**..."
  - **representation** – extracting structure from data
  - **generalisation** – making predictions from data

TYPES OF MACHINE LEARNING PROBLEM

supervised *making predictions*

unsupervised *extracting structure*



generalisation

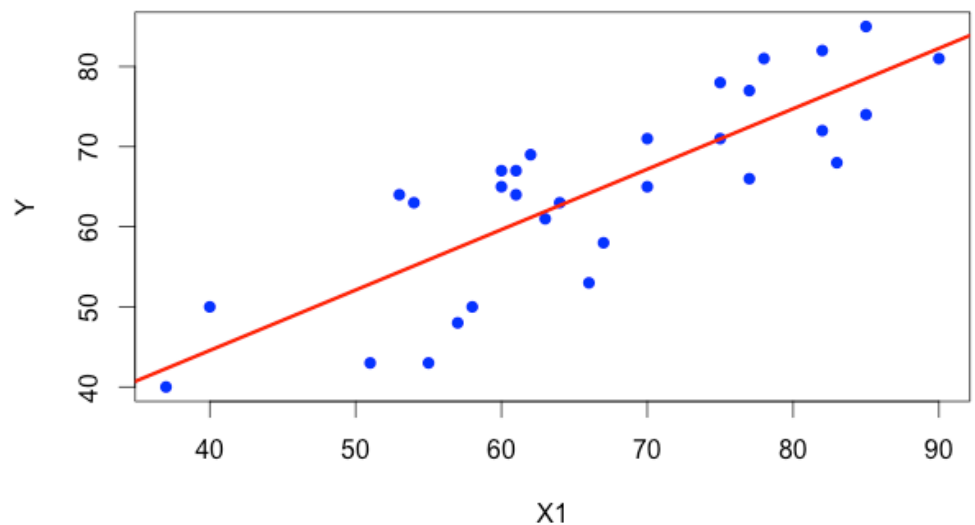
representation

TYPES OF MACHINE LEARNING PROBLEM

supervised *making predictions*

unsupervised *extracting structure*

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



TYPES OF MACHINE LEARNING PROBLEM

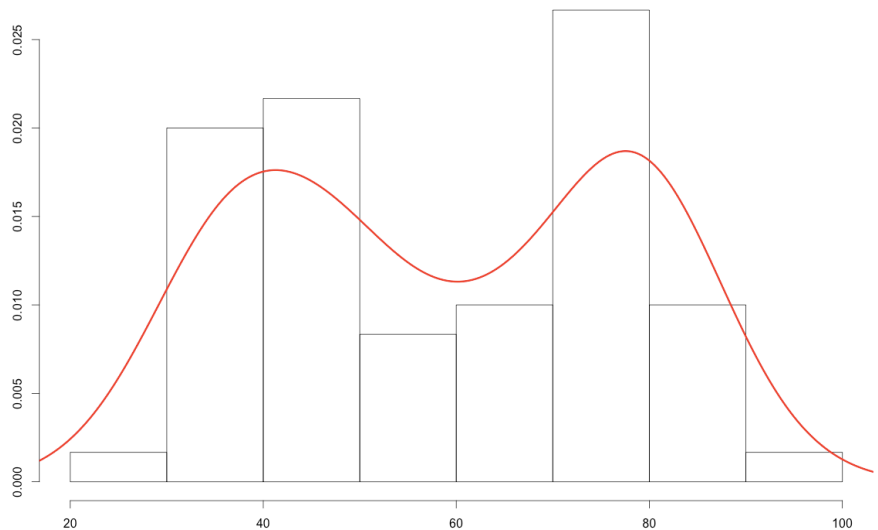
supervised

unsupervised

making predictions

extracting structure

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



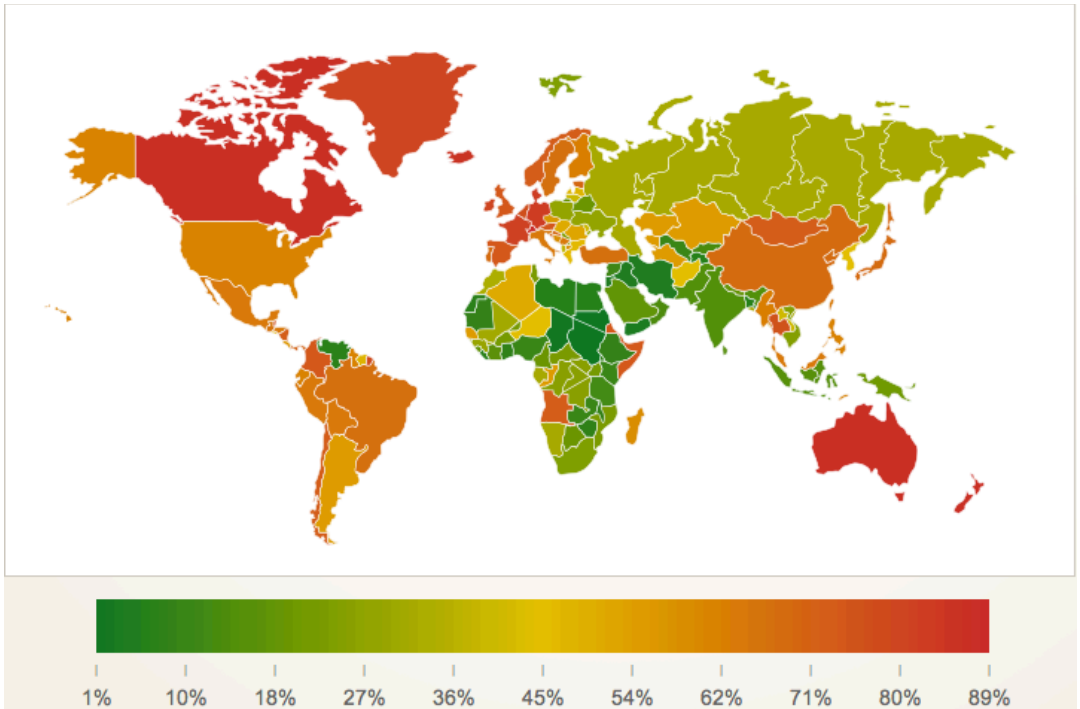
TYPES OF DATA

continuous	***categorical***
quantitative	*qualitative*
e.g. height	e.g. eye colour

TYPES OF ML PROBLEMS

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

REGRESSION EXAMPLE: PREDICTING IPHONE SALES



GDP

population

Gini

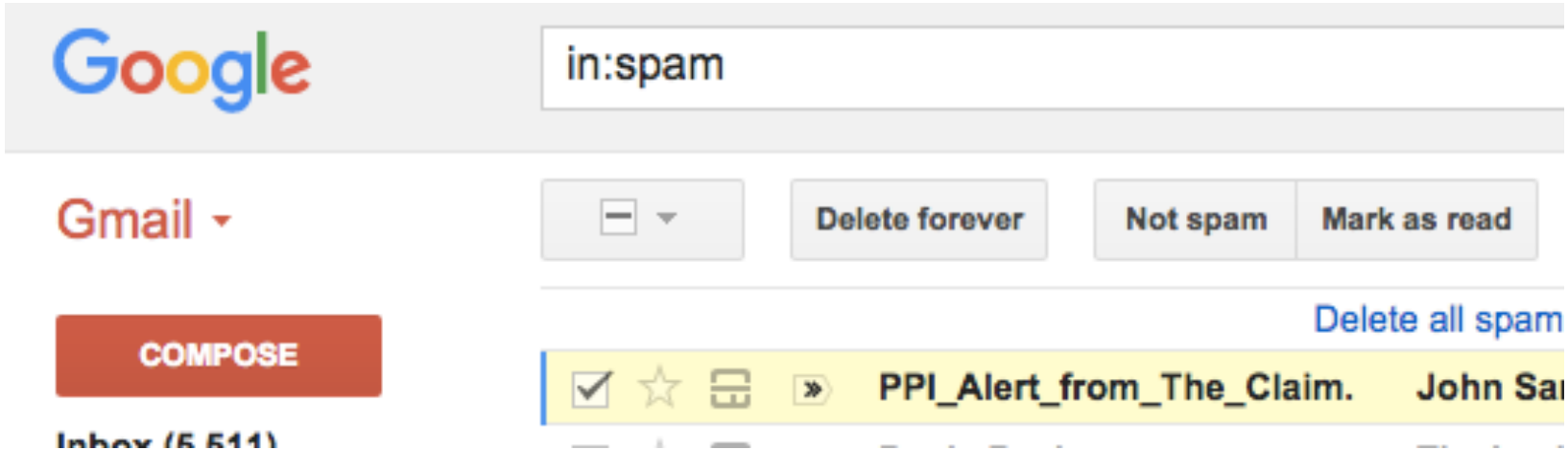
phone penetration %

GDP growth rate

TYPES OF ML PROBLEMS

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

CLASSIFICATION EXAMPLE: SPAM FILTERING



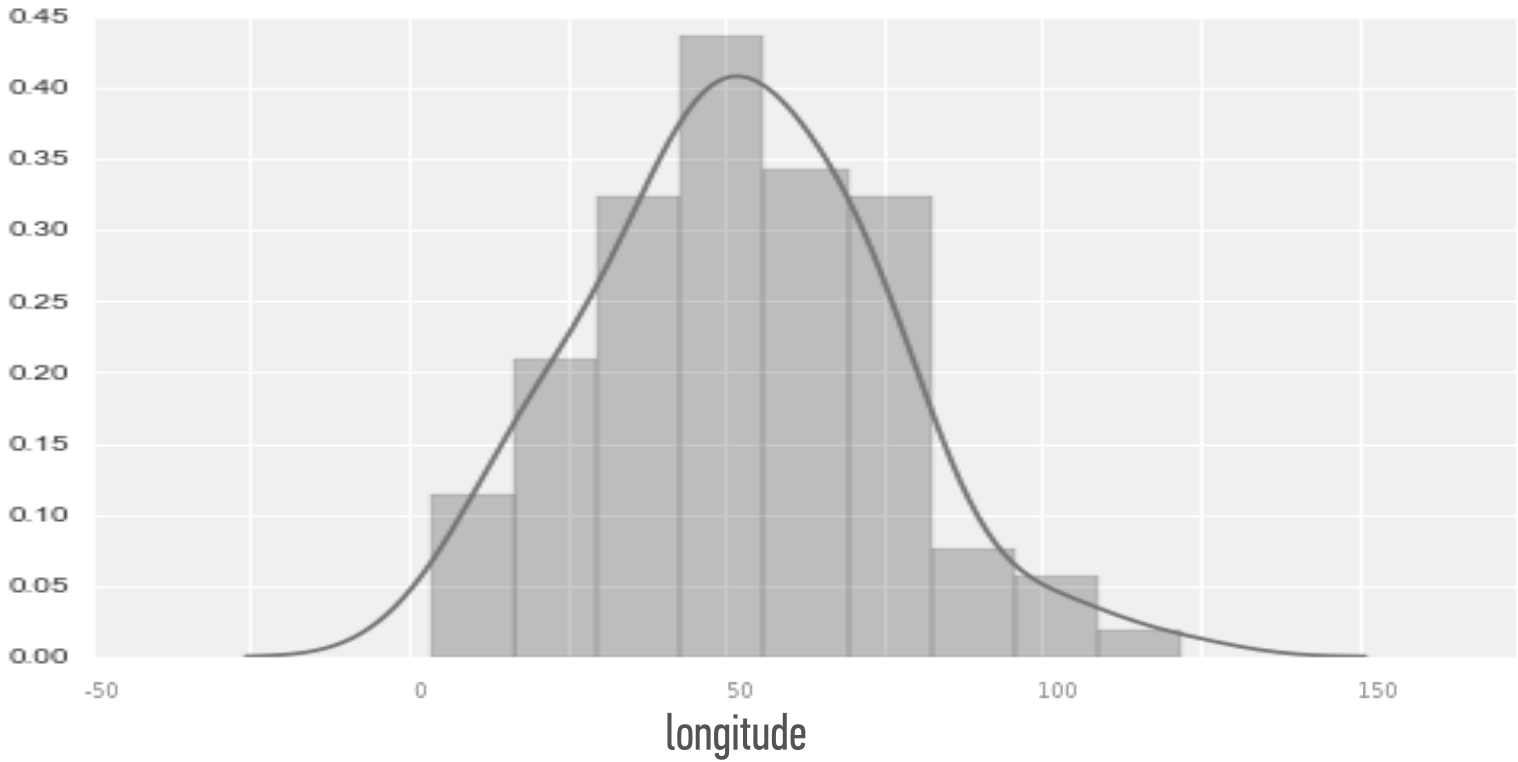
$$$ *Act now!* *As seen on* *Satisfaction guaranteed*

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

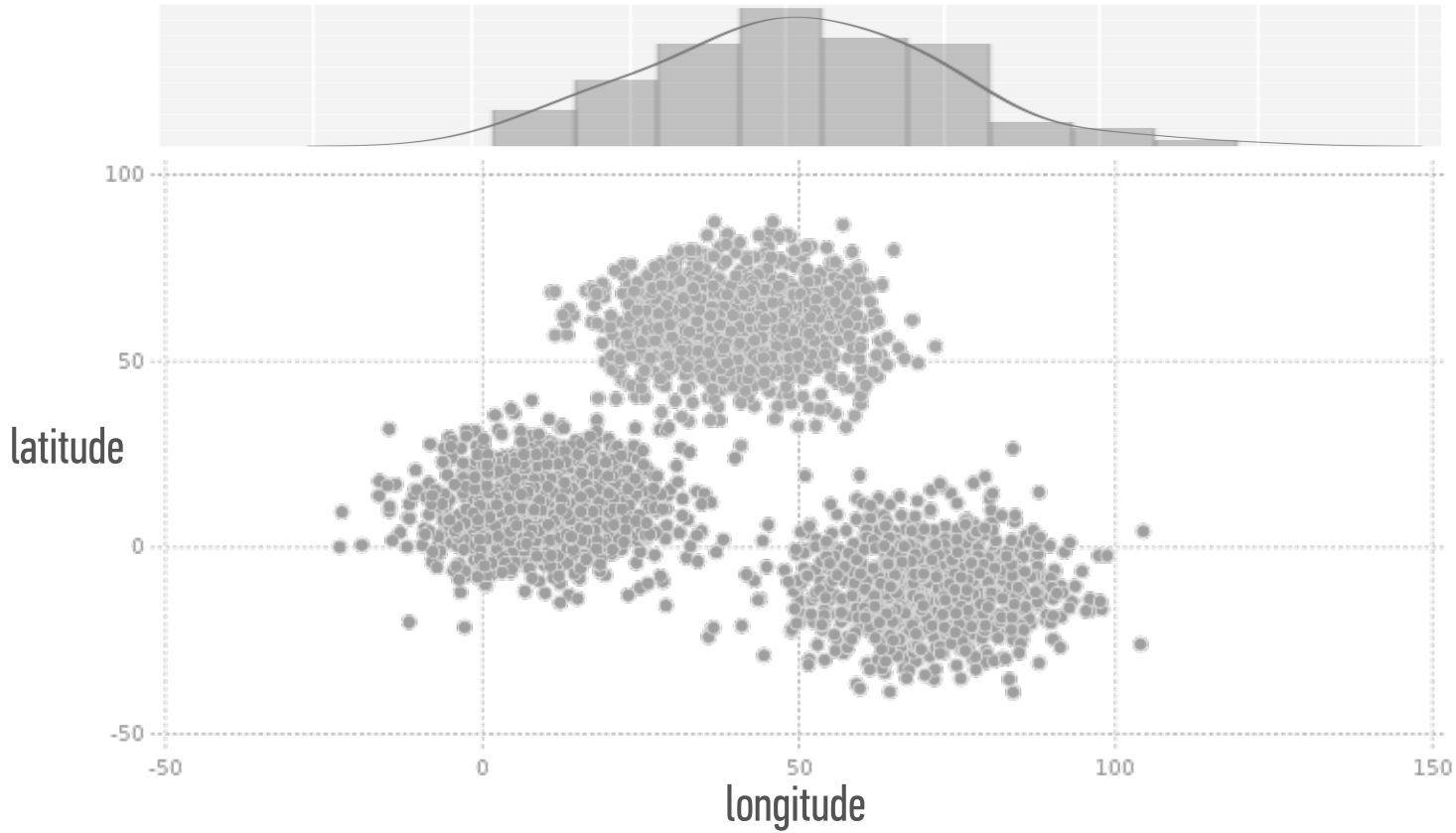
TYPES OF ML PROBLEMS

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

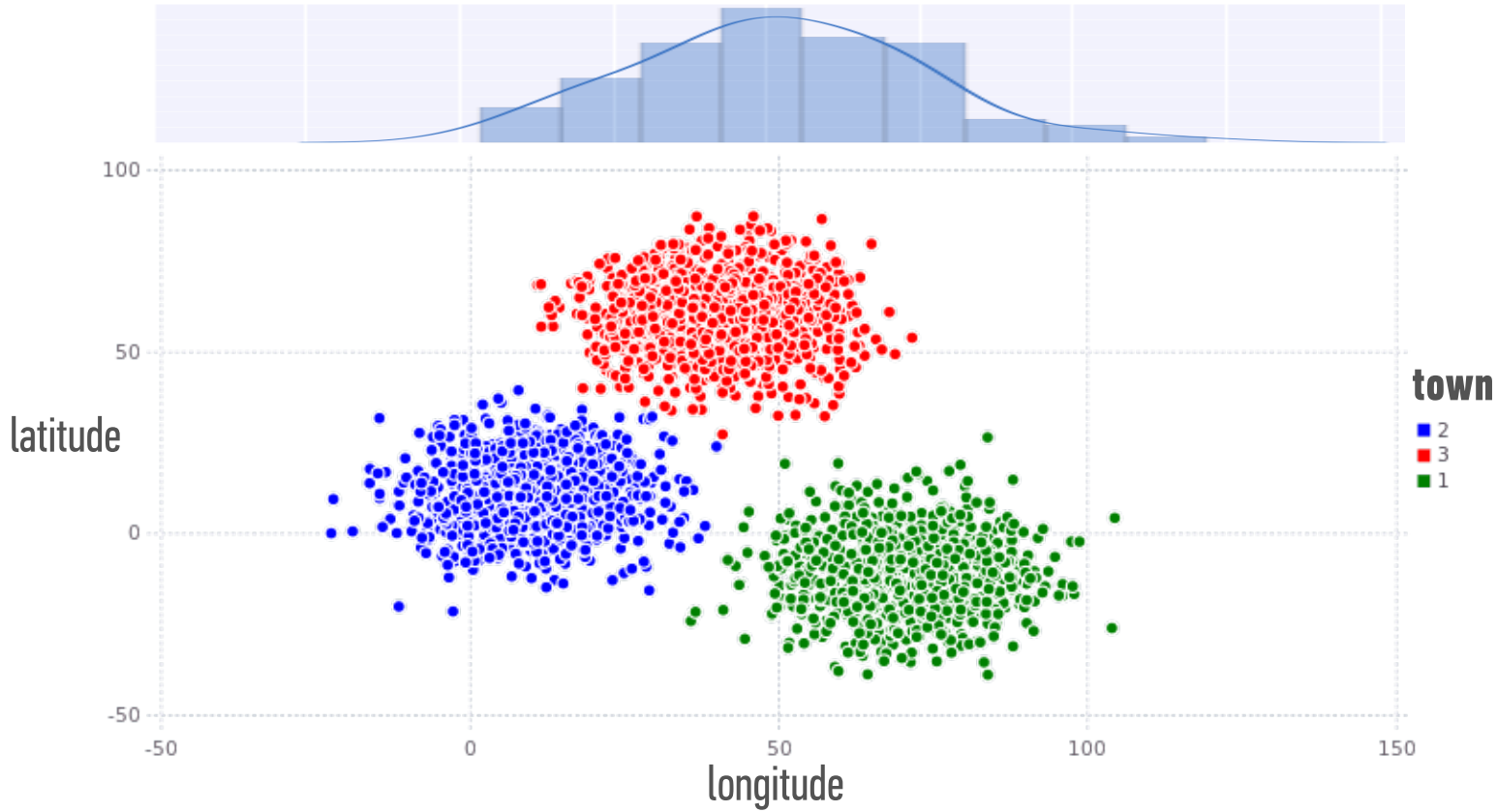
CLUSTERING EXAMPLE: USER LOCATIONS



CLUSTERING EXAMPLE: USER LOCATIONS



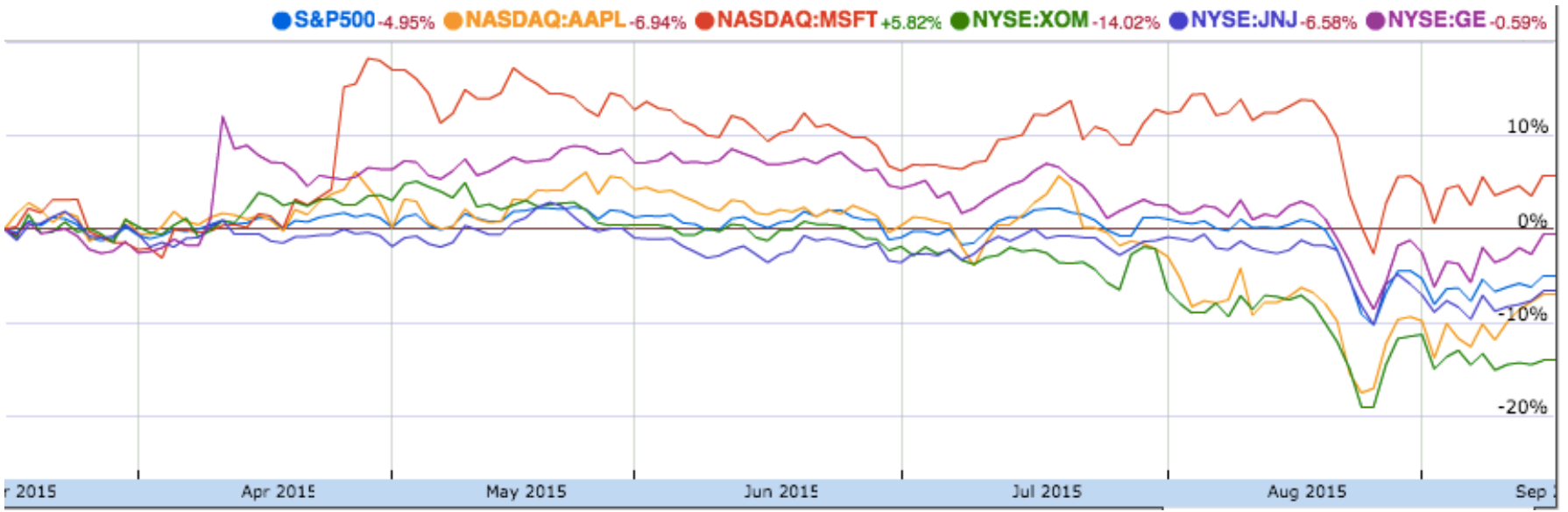
CLUSTERING EXAMPLE: USER LOCATIONS



TYPES OF ML PROBLEMS

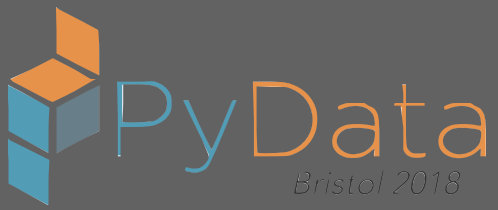
	continuous	***categorical***
supervised	*regression*	*classification*
unsupervised	*dimensional reduction*	*clustering*

DIMENSIONAL REDUCTION EXAMPLE: A STOCK INDEX



DIMENSIONAL REDUCTION EXAMPLE: A STOCK INDEX

PART IV.
TIPS FOR SUCCESS



TECHNICAL ABILITY

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

PERSONALITY

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

CURIOSITY

- 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

RIGOUR

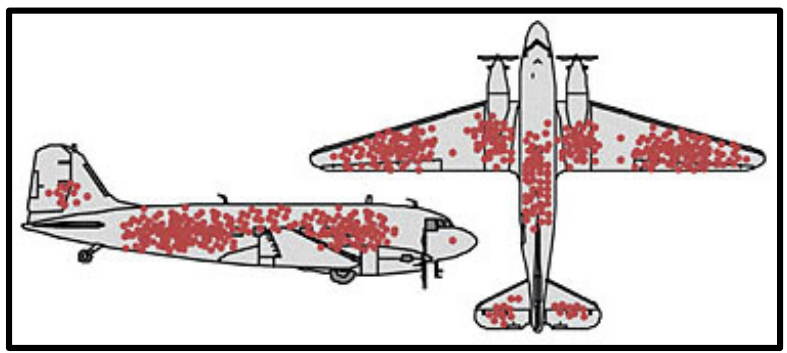
- Humans are hardwired to see patterns

- People give more weight to information that confirms their believes

- It is important to tell signal from noise

- When sifting through large amounts of data we are bound to find patters

RIGOUR IS HARD



COMMUNICATION

- 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

BUSINESS ACUMEN

- 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

NEW JOB

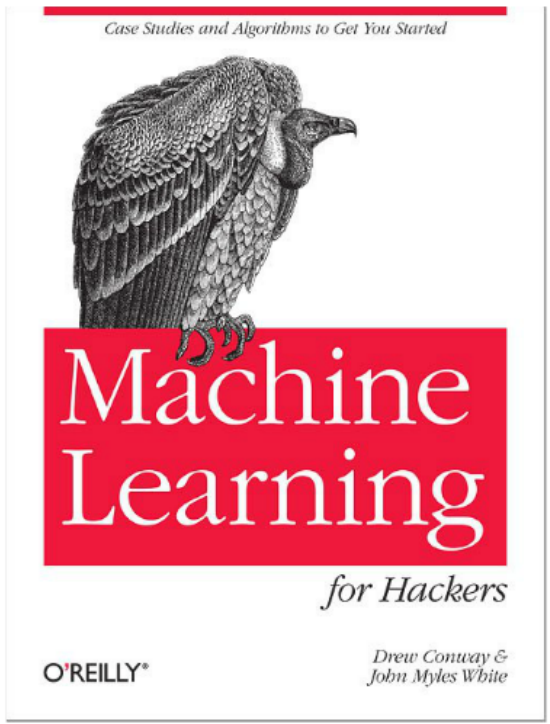
- In April 2012 McKinsey predicted 1.5 million shortage of data scientists
- More and more companies are looking for people to unlock the value in their data
- Rise in available positions

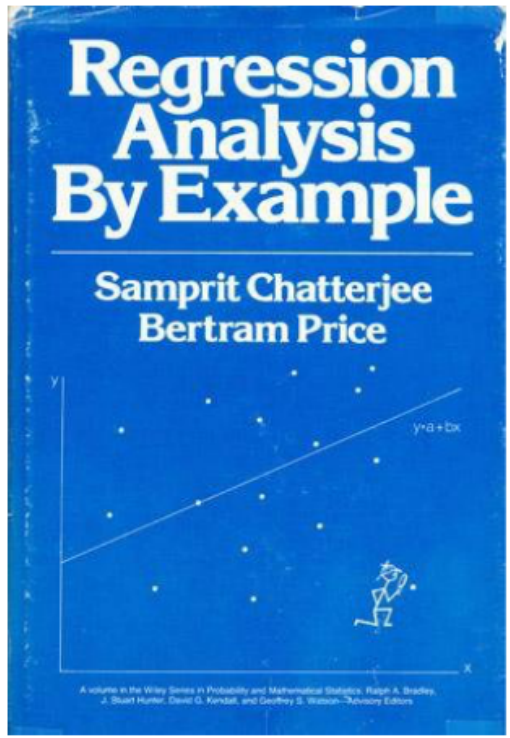
Location: London	3 months to 16 Aug 2013	Same period 2012	Same period 2011
Data Scientist			
Rank	566	631	-
Rank change year-on-year	▲ +65	● -	
Permanent jobs requiring a Data Scientist	41	11	0
As % of all permanent IT jobs located in London	0.091%	0.021%	-
As % of the Job Titles category	0.097%	0.023%	-
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%	

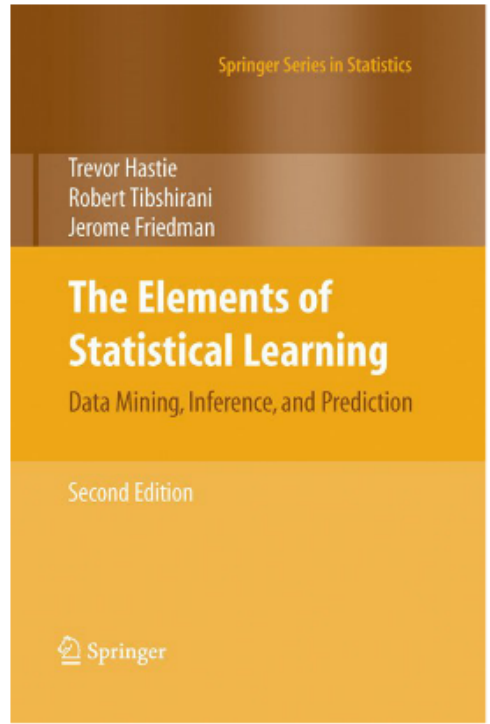
SHORTAGE OF SKILLS

- 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

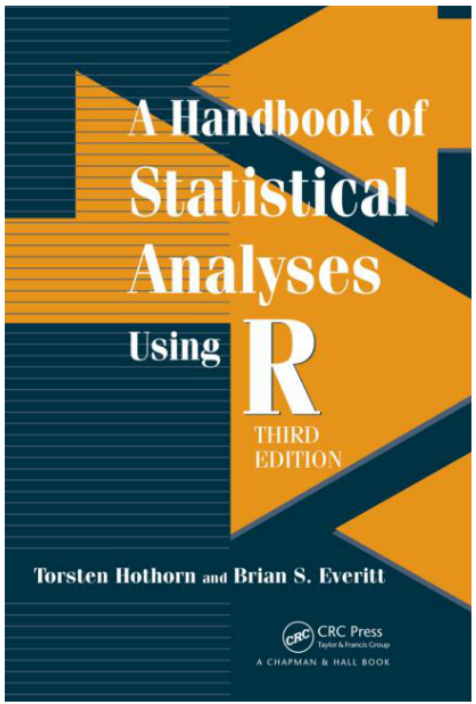
BOOKS

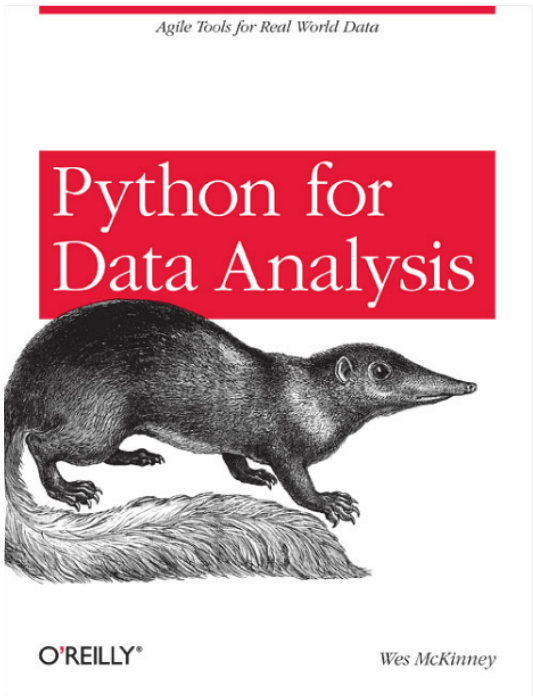

Case Studies and Algorithms to Get You Started
Machine Learning
for Hackers
O'REILLY
Drew Conway &
John Myles White

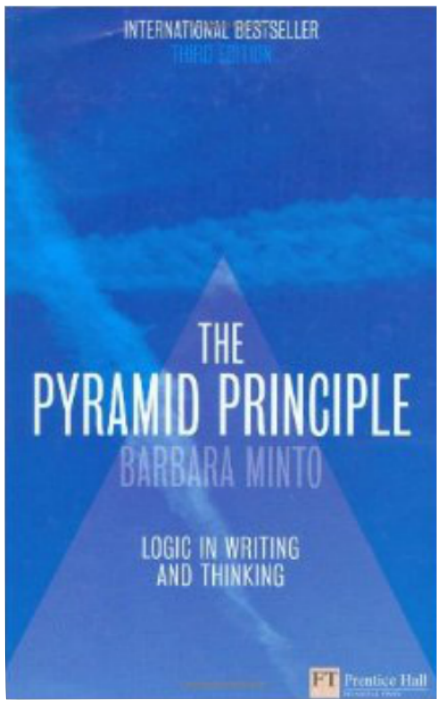

Regression Analysis By Example
Samprit Chatterjee
Bertram Price


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

MY TOP 3 BOOK RECOMMENDATIONS







ONLINE COURSES

Machine Learning
Andrew Ng (Stanford)

Machine Learning
CalTech CS156

www.dataquest.io
Writing code, work with data, build projects in your browser.

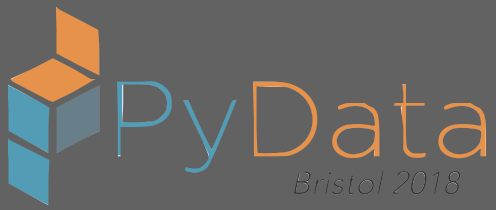
swirlstats.com
"Learn R, in R"

www.datacamp.com
"Learn data analysis from the comfort of your browser" (R, Python, DataViz)

PODCASTS

- **Data Skeptic** (Kyle Polich, I ❤️ the mini-explainer episodes!)
- **Partially Derivative** (light hearted)
- **Linear Digressions** (Udacity)
- **More or Less** (Tim Harford & BBC Radio 4)
- **O'Reilly Data Show** (Ben Lorica, technical with more focus on data engineering)
- **Planet Money** (NPR, economics/data/finance – A/B testing, multiple comparisons)
- **What's The Point** (FiveThirtyEight, how data is changing our lives)
- **Science Vs** (Gimlet Media, new last summer, controversial issues + rigour)

PART V.
THE PYDATA ECOSYSTEM



WHY PYTHON?


Nicholas Tollervey
@ntoll
Follow
@hynek more anecdote: kids in UK learn Python3 - it's the standard promoted by @Raspberry_Pi & soon #BBCMicroBit via @Micropython. #longterm
RETWEETS
9
LIKES
10
7:47 PM - 17 Feb 2016
9
10

POWERED BY PYTHON

Instagram

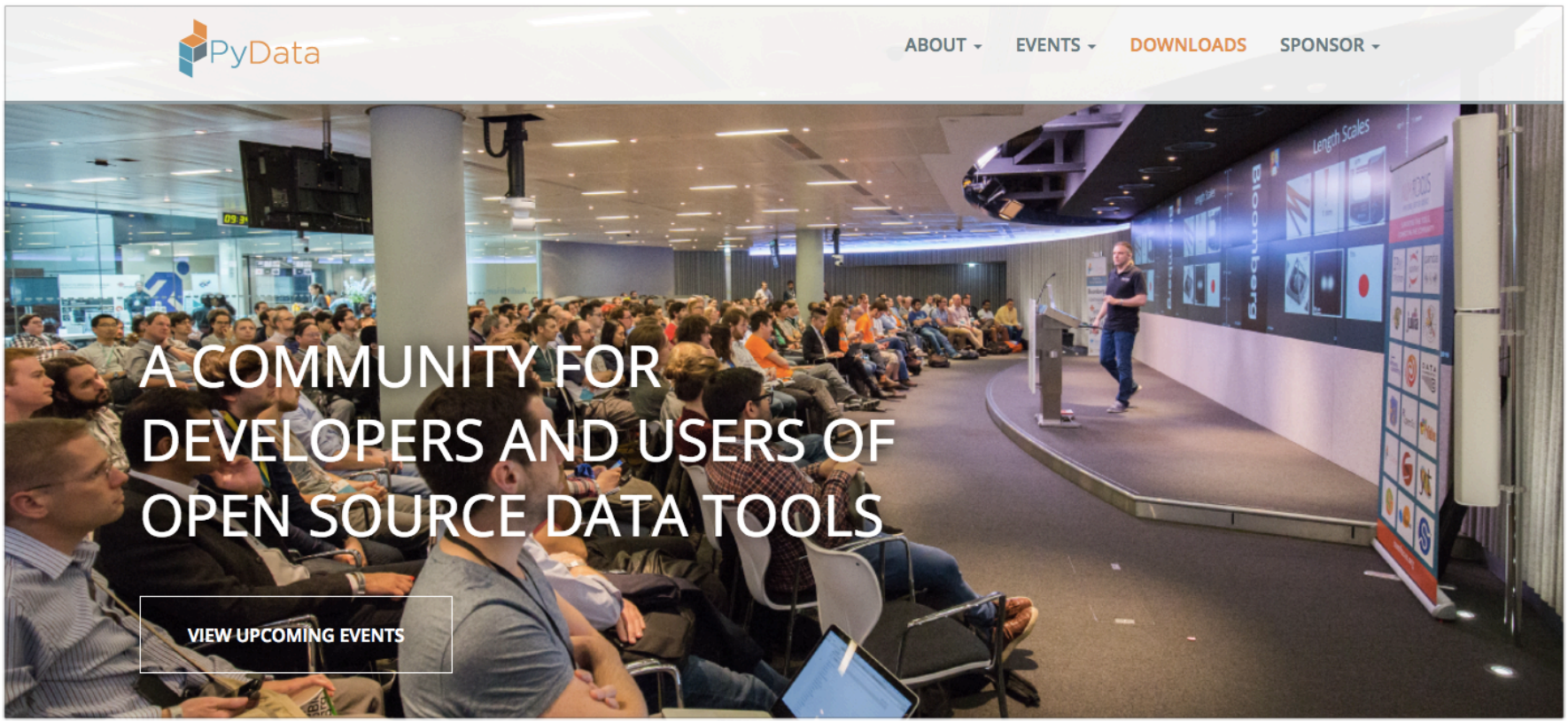
Quora

Dropbox

Spotify

POWERED BY PYTHON

START AT PYDATA.ORG



UPCOMING EVENTS

FENICS'18
MARCH 21-23, 2018
Oxford, UK

PYDATA FLORENCE @ PYCON ITALY
APRIL 20-22, 2018

PYDATA LONDON
APRIL 27-29, 2018

PYTHON IN ASTRONOMY
APRIL 30 - MAY 4, 2018
New York, NY, USA

PYDATA KAUNAS @ PYCON LITHUANIA
MAY 5, 2018

ROPENSCI UNCONF
MAY 21-22, 2018
Seattle, WA, USA

PYDATA AMSTERDAM
MAY 25-27, 2018

PYDATA BERLIN
JULY 6-8, 2018

PYDATA EDINBURGH @ EUROPYTHON
JULY 25-29, 2018

JULIACON
AUGUST 7-11, 2018
London, UK

JUPYTERCON
AUGUST 21-24, 2018
New York, NY, USA

PYDATA CÓRDOBA
OCTOBER 1-2, 2018

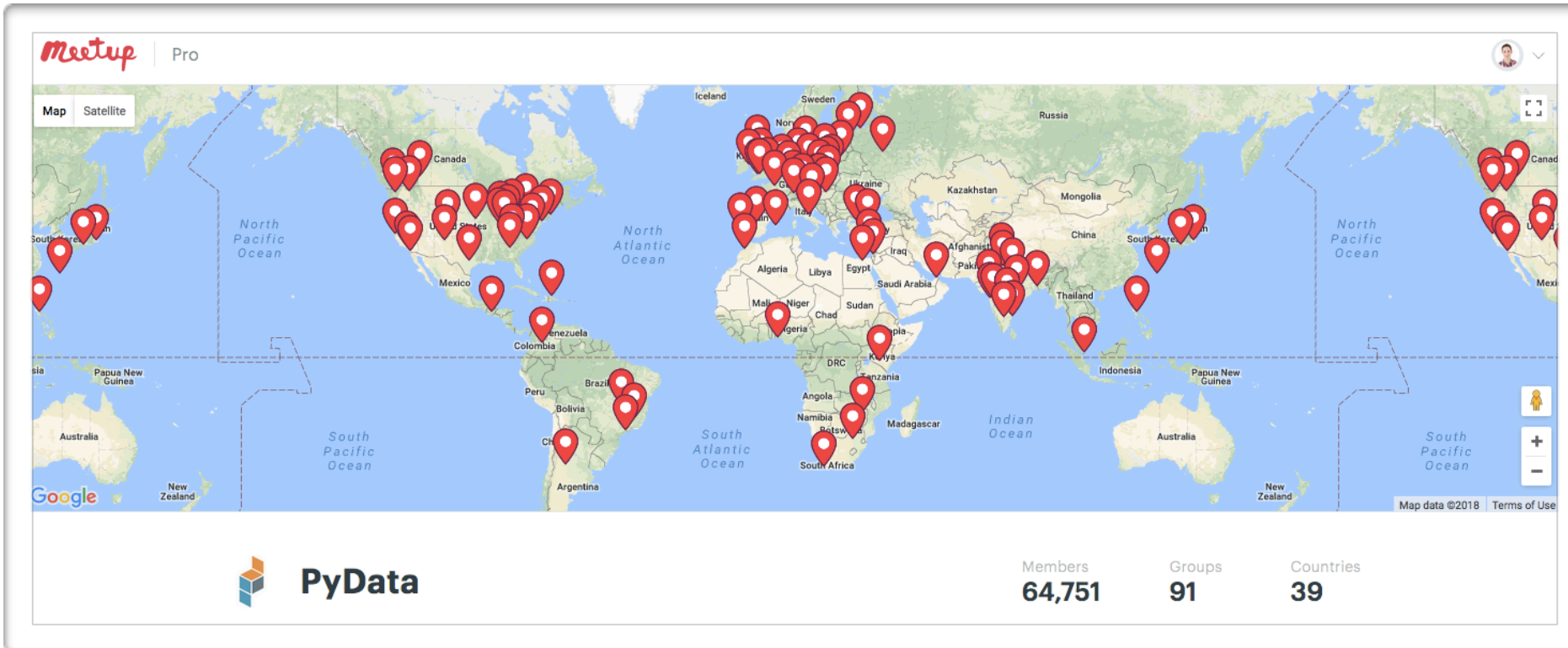
PYDATA LOS ANGELES
OCTOBER 22-24, 2018

PYDATA KARLSRUHE & PYCON DE
OCTOBER 24-28, 2018

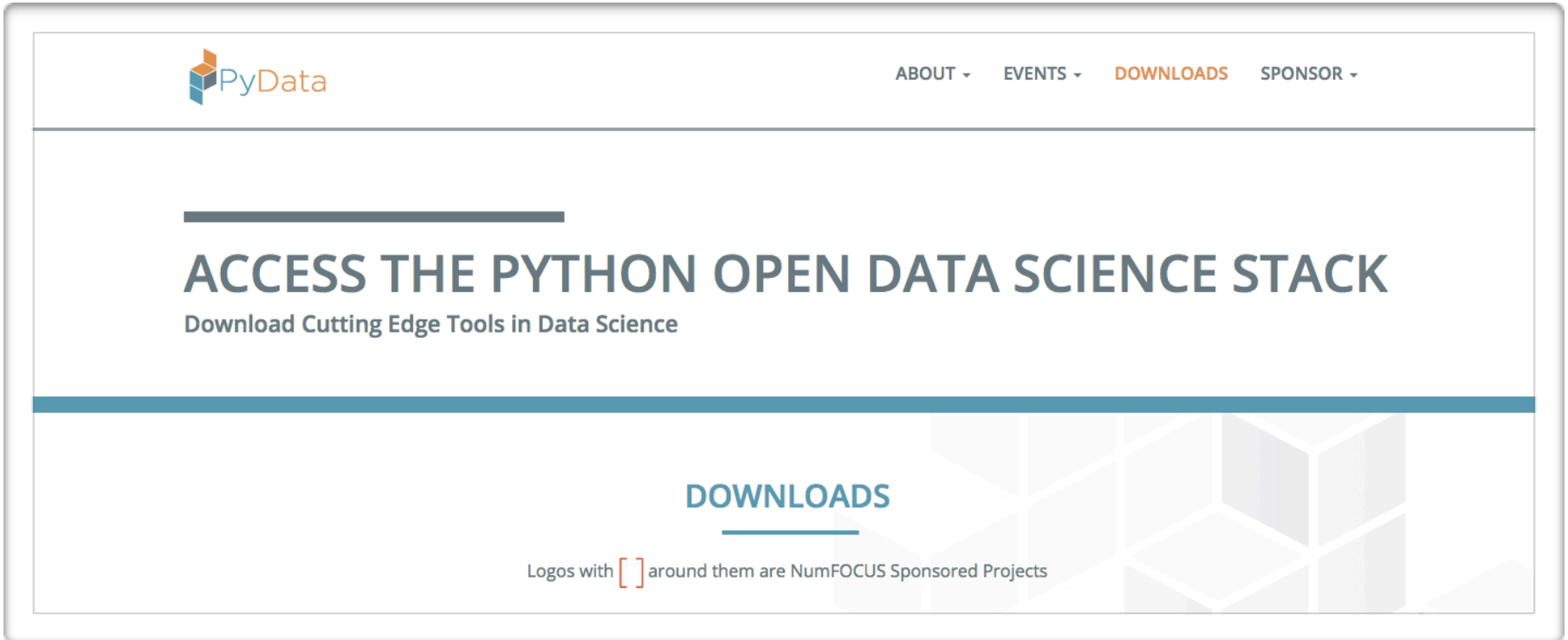
PYDATA WARSAW
NOVEMBER 19-20, 2018

PYDATA NYC
NOVEMBER 2018

MEETUPS


Meetup
Pro
Map
Satellite
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Canada
Mexico
Colombia
Peru
Bolivia
Brazil
Argentina
Iceland
Sweden
Russia
Ukraine
Kazakhstan
Mongolia
China
Iraq
Afghanistan
Saudi Arabia
Egypt
Libya
Algeria
Mali
Niger
Chad
Sudan
DRC
Angola
Namibia
South Africa
Madagascar
Thailand
Indonesia
Papua New Guinea
Australia
New Zealand
Google
Map data ©2018
Terms of Use
PyData
Members
64,751
Groups
91
Countries
39

DOWNLOADS & SPONSORED PROJECTS



PACKAGES TO START WITH

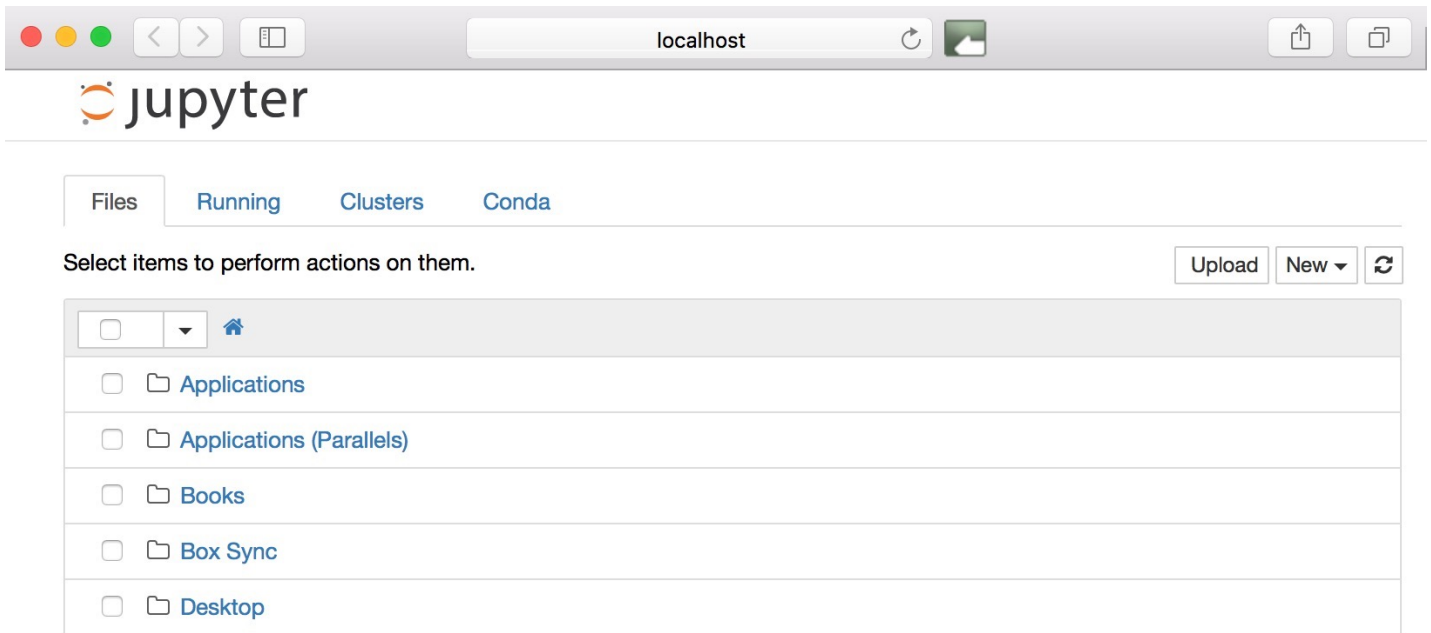
- **pandas:** manipulate data
- **SciPy/NumPy:** scientific computing and numerical calculations
- **Scikit-learn:** machine learning
- **matplotlib/Seaborn:** data visualisation
- **spacy/nltk:** natural language processing
- **statsmodels:** statistical tests
- **Beautiful Soup:** HTML/XML data & web scrapers
- **Jupyter:** interactive programming environment

MY MOST USED PACKAGES

- **pandas:** manipulate data
- **SciPy/NumPy:** scientific computing and numerical calculations
- **Scikit-learn:** machine learning
- **matplotlib/Seaborn:** data visualisation
- **spacy/nltk:** natural language processing
- **statsmodels:** statistical tests
- **Beautiful Soup:** HTML/XML data & web scrapers
- **Jupyter:** interactive programming environment

JUPYTER NOTEBOOK

Jupyter Notebook is a web interface that let's us use formatting along side our code.

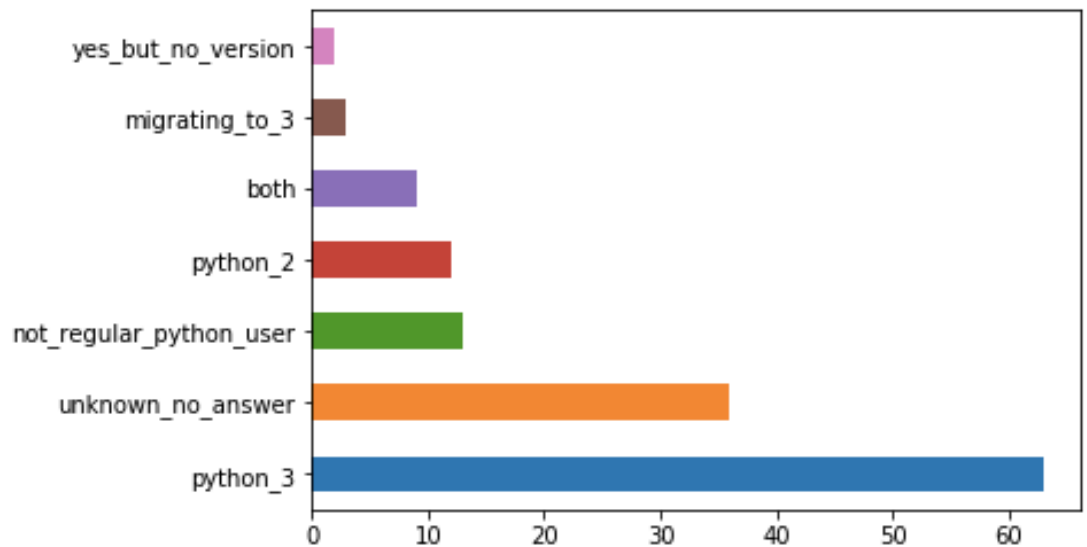


DEMO

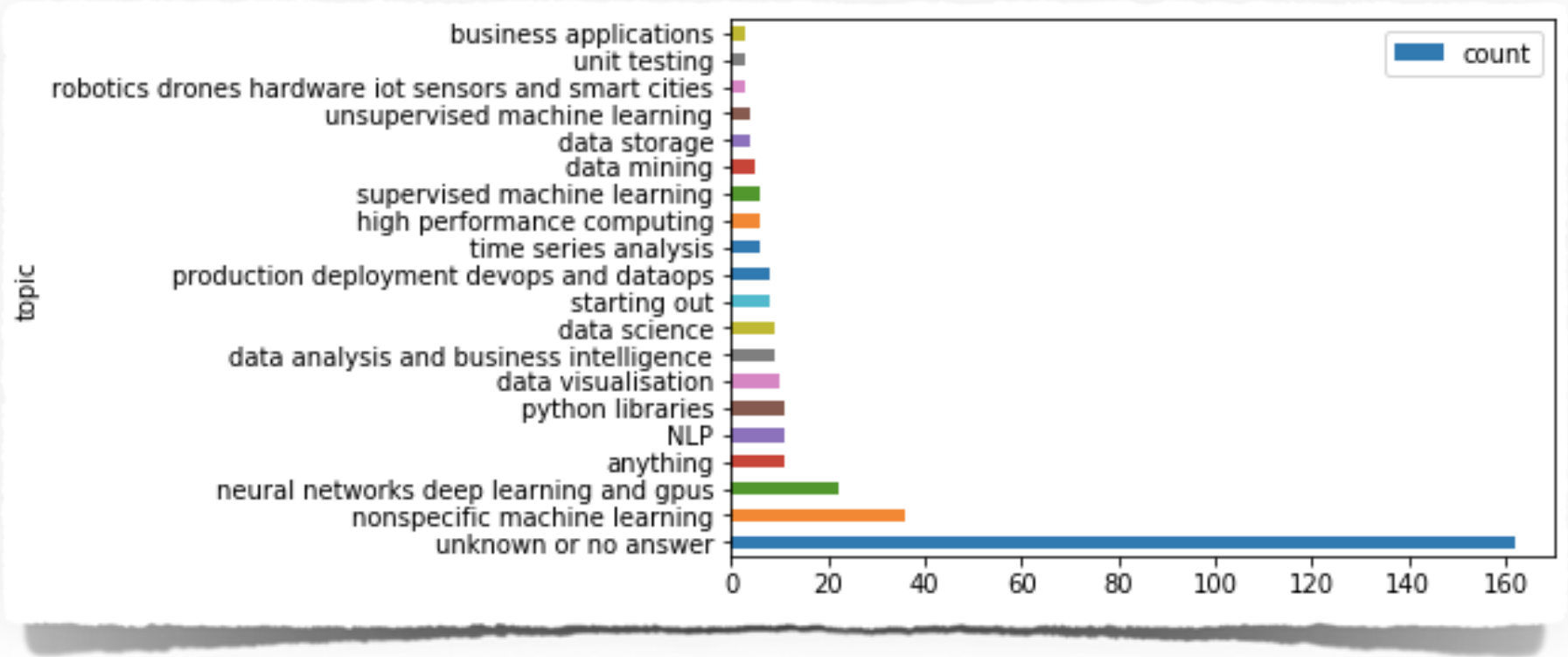
80% of PyData Bristol members use Python 3 vs 20% using Python 2!

** Excluding "don't know" and dividing "both" equally.

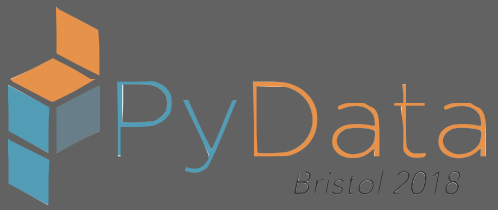
```
Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x1093035c0>
```



TOPICS FOR FUTURE PYDATA BRISTOL MEETUPS



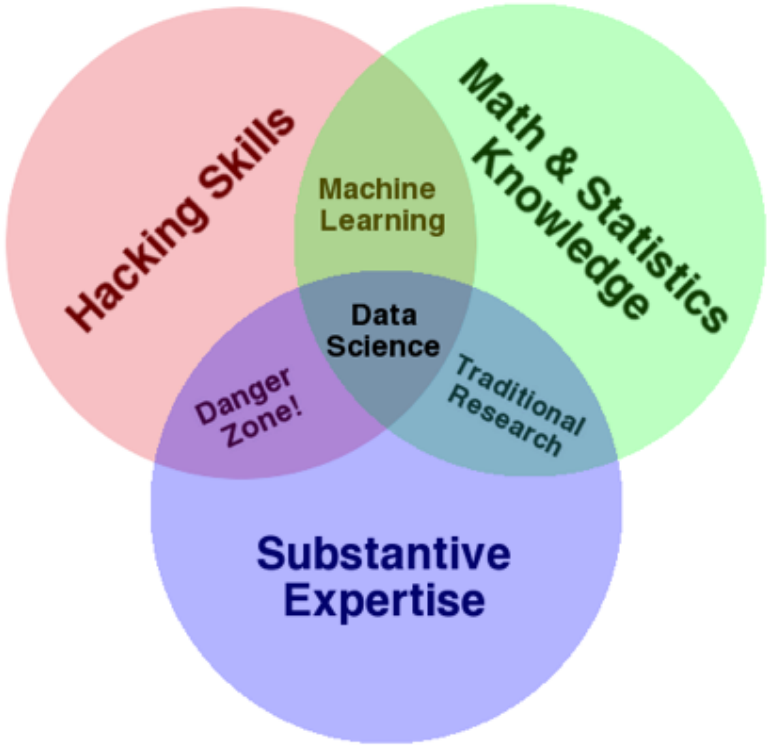
FINAL THOUGHTS



DATA SCIENCE vs DATA ANALYTICS

"Data Analytics"	"Data Science"
Historical reporting.	Predictive forecasting.
Metrics. KPIs. Segmentation.	Statistics. Regression. Machine learning.
Dashboards. BI tools. Pivot tables.	Coding. Flexibility. Automation.
Necessary…keeps the engines running.	Exciting…unexpected insights.
Tools: Excel, SQL, Tableau.	Tools: Python, R, scikit-learn.

"DATA ANALYST"...OR "DATA SCIENTIST"?



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

WHAT IS DATA SCIENCE FOR ME?

"Data Analytics"

Historical reporting.
Metrics. KPIs. Segmentation.
Dashboards. BI tools. Pivot tables.
Necessary...keeps the engines running.
Tools: Excel, SQL, Tableau.

"Data Science"

Predictive forecasting.
Statistics. Regression. Machine learning.
Coding. Flexibility. Automation.
Exciting...unexpected insights.
Tools: Python, R, scikit-learn.

"Data Engineering"

Architecture. Devops. Cloud solutions.
Databases. Data warehouses. Big data.
Integrations (e.g. tracking, channel attribution).
BI tools. Automated reporting. Bespoke solutions.
Version control. Repo management. Code review.

"Strategic Analysis"

Business skills. Startup methodology. Working lean.
Measuring success. KPIs. Data-informed decisions.
Communication. Technical writing. Domain expertise.
Project management. Agile workflows. Problem solving.
Education. Hiring. Mentoring. Advisory.

"BECOME A DATA SCIENTIST WITH THESE 4 WEIRD TIPS"

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.

LONDON MEETUPS

- 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

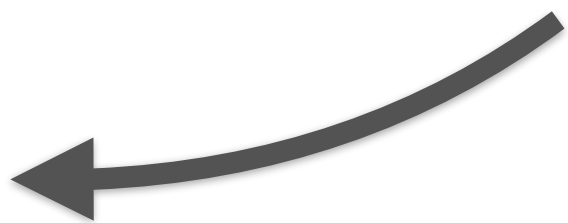
BRISTOL MEETUPS!

- PyData Bristol
- Bristol Data Scientists
- Big Data Bristol
- South West Data Meetup
- Bath Machine Learning Metope
- Bristol Digital Analytics Meetup
- SQL Bristol
- Cardiff R User Group
- Bristech
- South West Futurists
- CodeHub Bristol
- Bath: Hacked

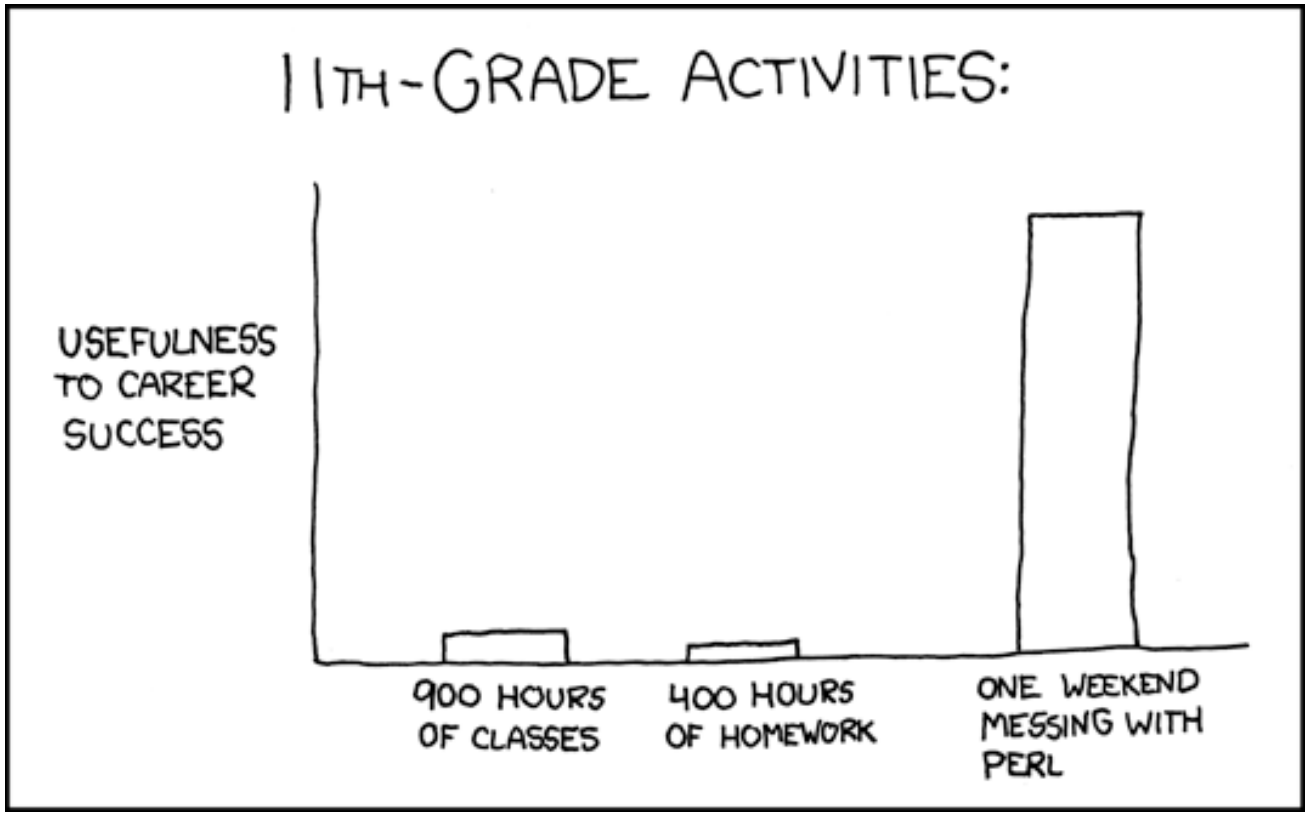
HACKATHONS & DATADIVES

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

"We liberate data, and make useful things"



IF YOU DO NOTHING ELSE...



...GET STARTED TONIGHT!

- **Data Skeptic Podcast**

dataskeptic.com

THE CORNERSTONE OF A DAUNTING FUTURE?

FINAL THOUGHTS

- Data science is a product of our time

- Being a data scientists requires people and technical skills

- We’re only getting started…

THANK YOU

QUESTIONS?