



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

John Sandall
15th March 2018

Data Science & Engineering Consultant
[@john_sandall](https://twitter.com/john_sandall)

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?



Chris Dixon

@cdixon



Following

"A data scientist is a statistician who lives in San Francisco" via [@smc90](#)

WHAT IS DATA SCIENCE?



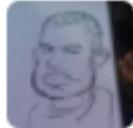
Big Data Borat
@BigDataBorat



 Follow

Data Science is statistics on a Mac.

WHAT IS DATA SCIENCE?



((((Josh Wills))))

@josh_wills



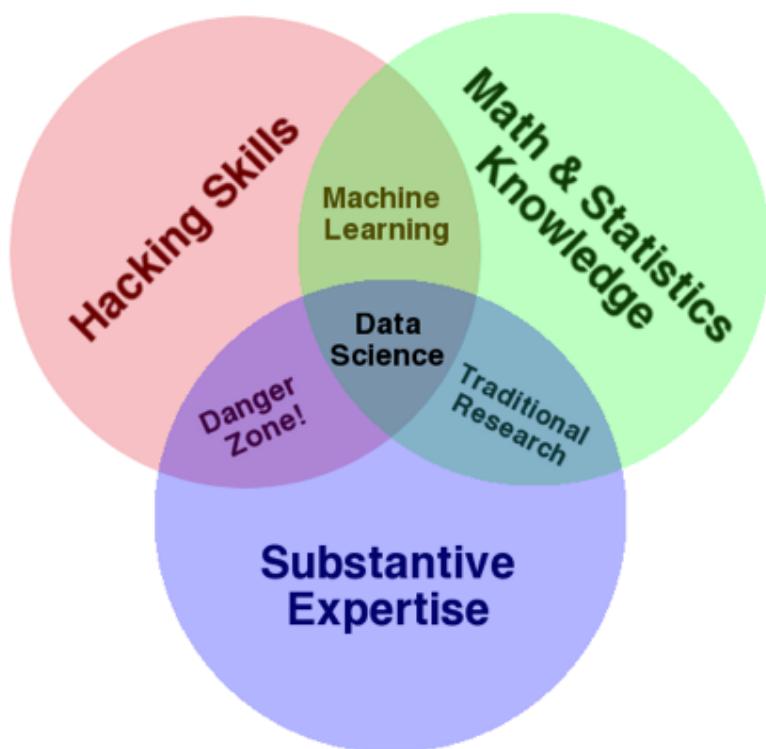
Follow

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

facebook

Mark Zuckerberg's Profile

Harvard

Information

Account Info

Name: Mark Zuckerberg [add to friends]
Networks: Harvard
Facebook
San Francisco, CA
Last Update: August 14, 2006

Basic Info

Sex: Male
Relationship Status: In a Relationship
Residence: Kirkland
Birthday: May 14, 1984
Hometown: Dobbs Ferry, NY

Contact Info

Email: mzuckerb@fas.harvard.edu

Personal Info

Activities: lots of facebook
Interests: information flow, exponential growth, minimalism, meditation, driving, writing, making things, social dynamics, domination
Favorite Music: green day, franz ferdinand, weezer, fall out boy, my chemical romance
Favorite Books:
Favorite Quotes:
About Me: I make things that increase information flow between people.

Education Info

College: Harvard
Psychology, Computer Science

Status

Mark isn't receiving Facebook texts right now.

Harvard Friends

146 friends at Harvard See All

Study where you want.

Earn a













Twitter

Home | Your profile | Invite | Public timeline | [Ba...](#)

What are you doing? Characters available: 140

IM is down at the moment. We're working on restoring it. Thanks for your patience!

Update

What You And Your Friends Are Doing

 **kierstenster** isn't sure she wants to move in with Liz and Alana. She will miss Babar, Niki, and some more Babar. [39 minutes ago](#) from web

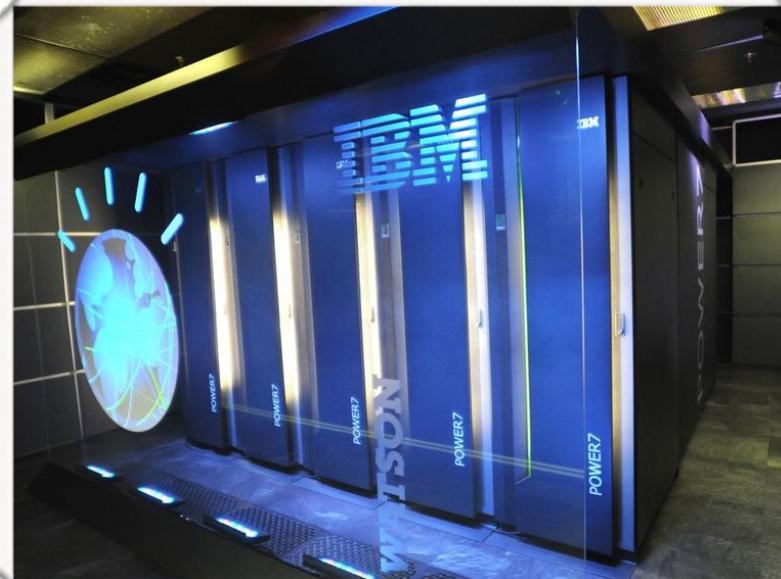
 **caroliniine** just did a great call with Lloyd Alter for my green marketing story! [about 2 hours ago](#) from [twitterific](#)

 **aprilini** I'm back at the office. Yeah, that's right. You heard me. Working. What an idea. I didn't say I was happy about it. [about 2](#)

2007: INFORMATION REVOLUTION



Kindle launches



IBM Watson created

2007: THE OPEN SOURCE ECOSYSTEM ACCELERATES



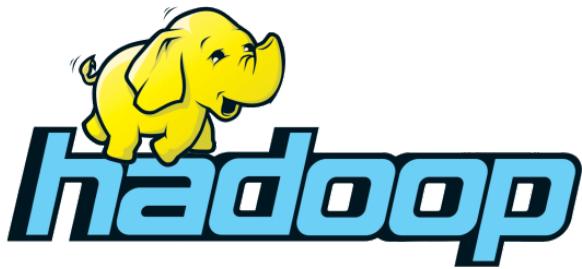
- Not only hardware costs are dropping
- Open Source - Linux, Apache, Hadoop, MySQL
- In addition to infrastructure there are also analysis tools
- Who knows R or Python?

2007: R & PYTHON BECOME ENTERPRISE FRIENDLY

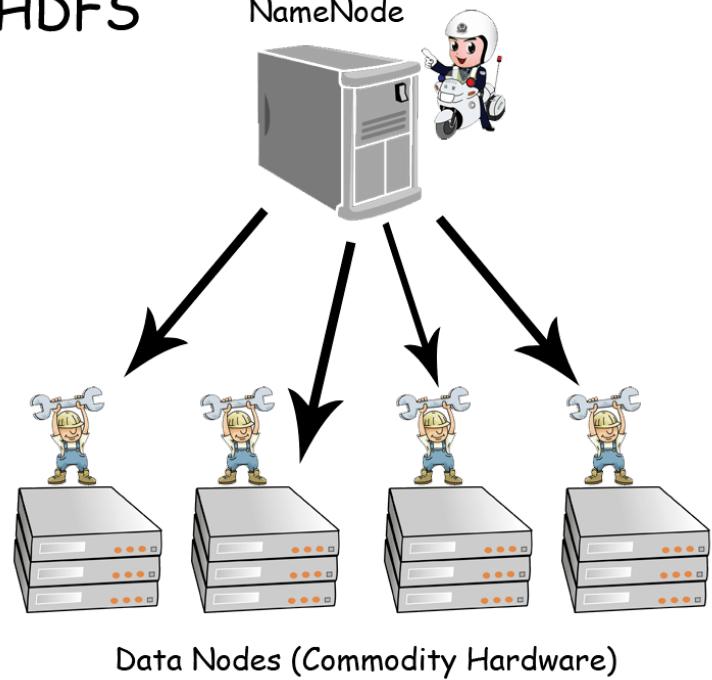
The screenshot shows a news article from Computerworld. At the top, there is a black header bar with the Computerworld logo and navigation icons. To the right of the logo, there are buttons for "INSIDER" and "Sign In". Below the header, the word "NEWS" is written in red capital letters. The main title of the article is "Microsoft unwraps a big-data analytics platform based on R".

The screenshot shows a news article from PCWorld. At the top, there is a dark red header bar with the PCWorld logo and the text "FROM IDG". Below the header, the word "NEWS" is written in red capital letters. The main title of the article is "Anaconda's Python-based analytics hit the enterprise with new subscription plans". A subtext at the bottom states "Also on the Python front, Teradata targets DevOps with a new module of its own".

2007: THE BIG DATA REVOLUTION BEGINS



HDFS

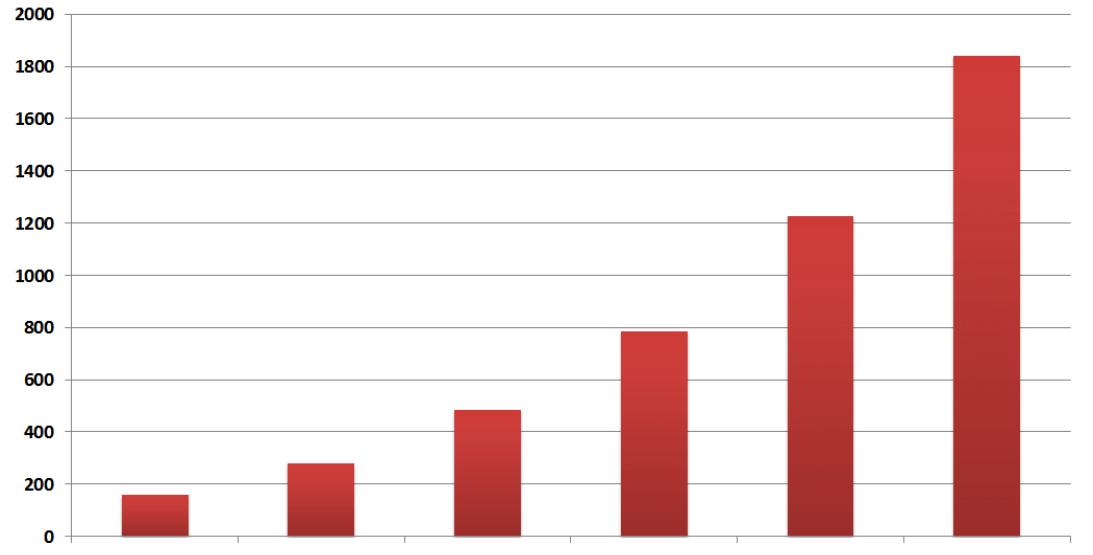


EVOLUTION OF DATA CREATION

2011:

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

Exabytes Created By Year (IDC)



Created by Mack D. Male

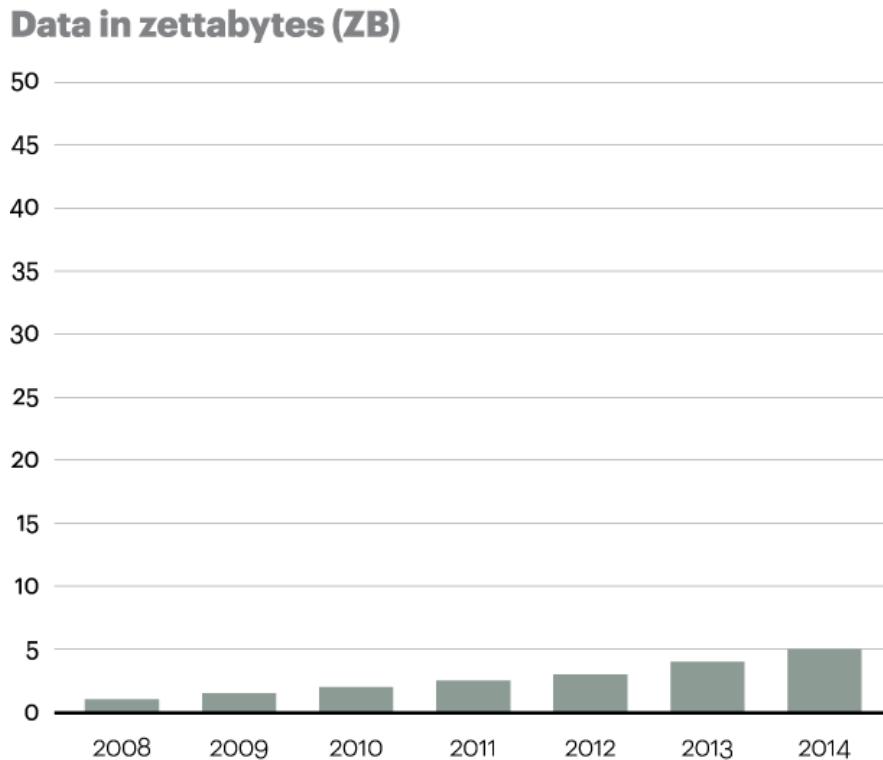
License: <http://creativecommons.org/licenses/by-sa/2.5/ca/>

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

EVOLUTION OF DATA CREATION

2014:

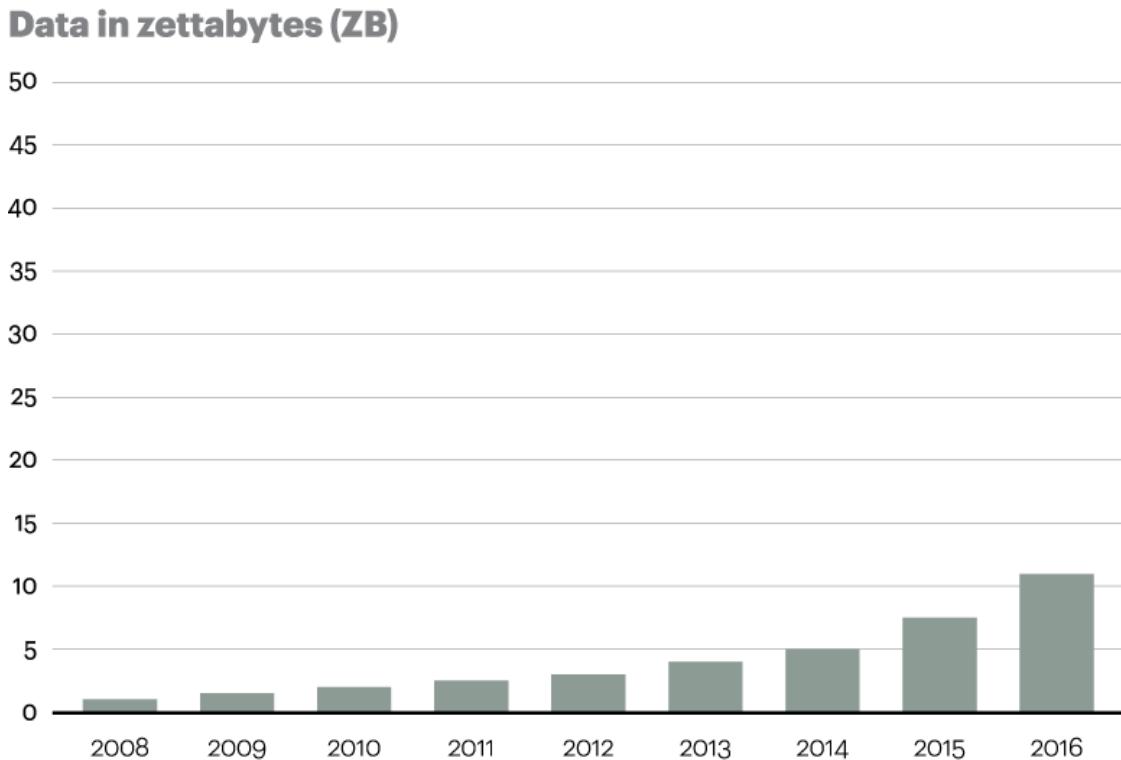
Oracle estimates total
data created annually
now surpasses five
Zettabytes



EVOLUTION OF DATA CREATION

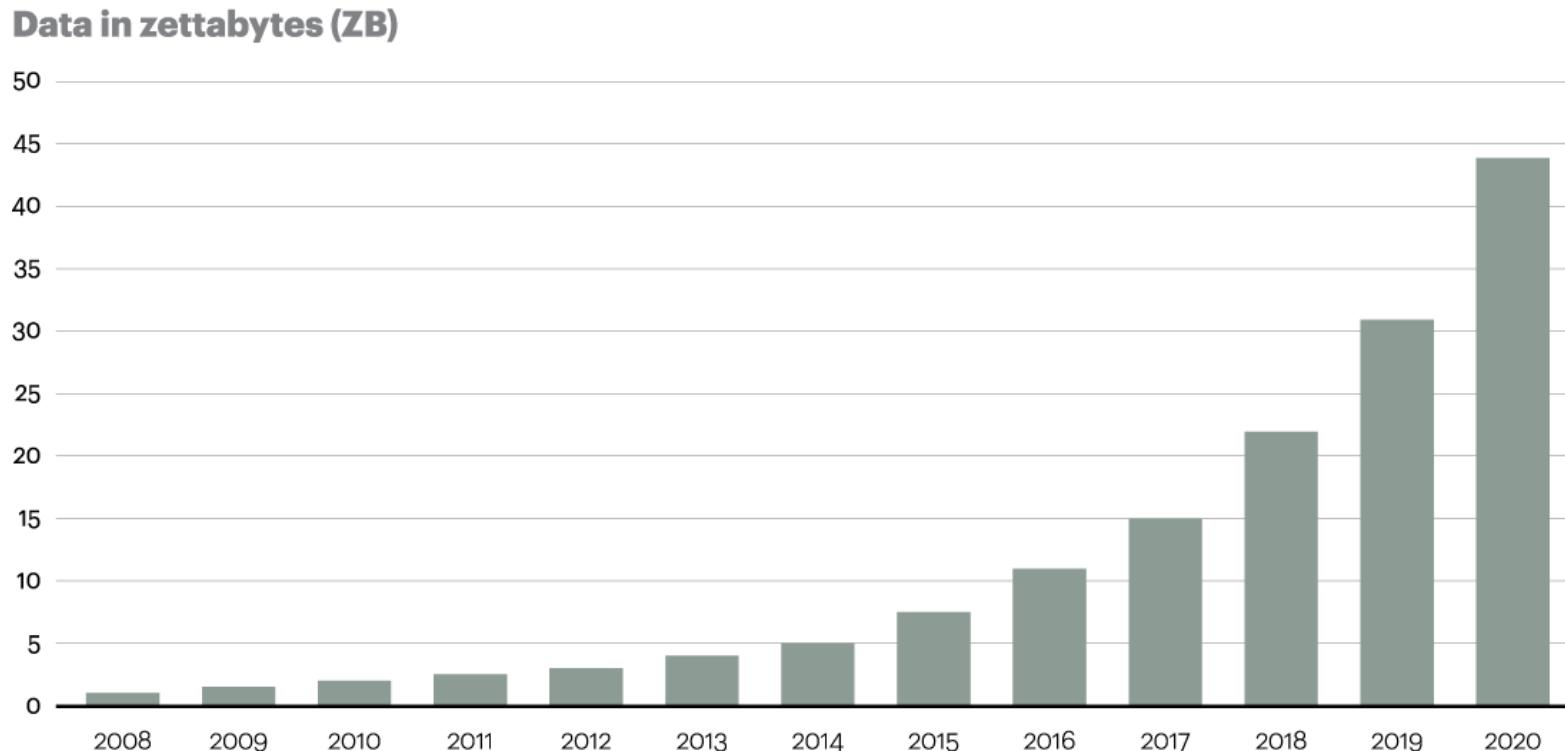
2016:

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



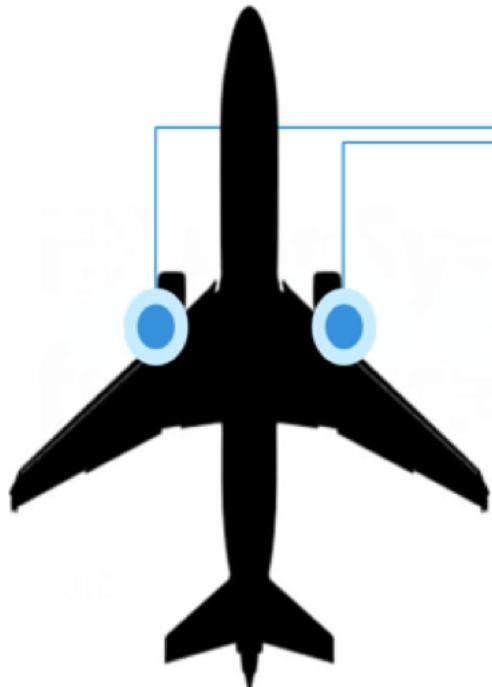
EVOLUTION OF DATA CREATION

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



WHERE IS DATA COMING FROM?

Sensor data from a cross-country flight



$$20 \text{ TB} \times 2 \times 6 \times 28,537 \times 365$$

20 terabytes of
information per
engine every hour

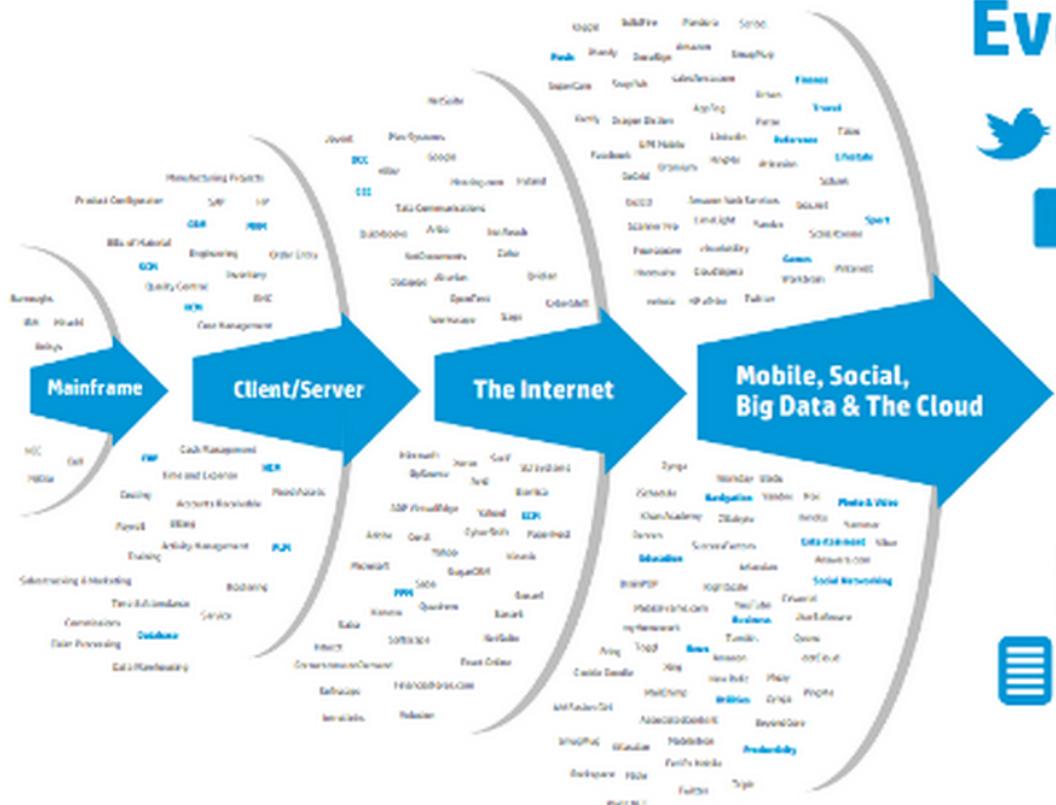
twin-engine
Boeing 737

six-hour, cross-
country flight from
New York to Los
Angeles

of commercial
flights in the sky in
the United States on
any given day.

days in a year

$$= 2,499,841,200 \text{ TB}$$



Every 60 seconds

98,000+ tweets

695,000 status updates

11million instant messages

698,445 Google searches

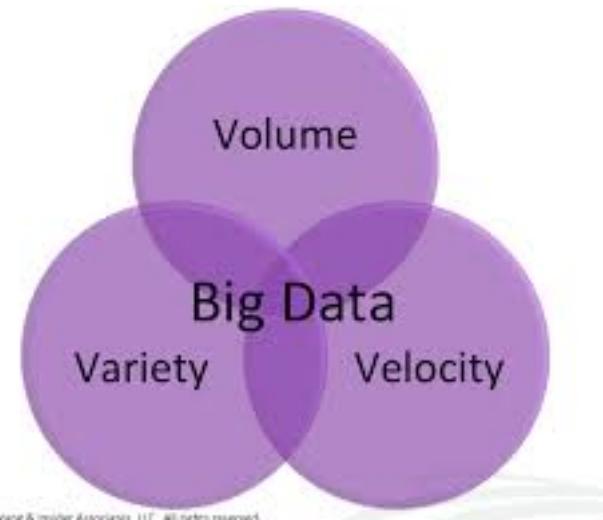
168 million+ emails sent

1,820TB of data created

217 new mobile web users

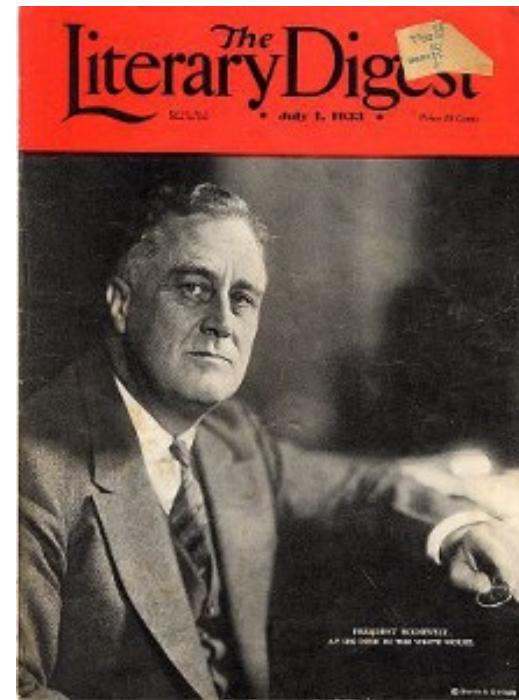
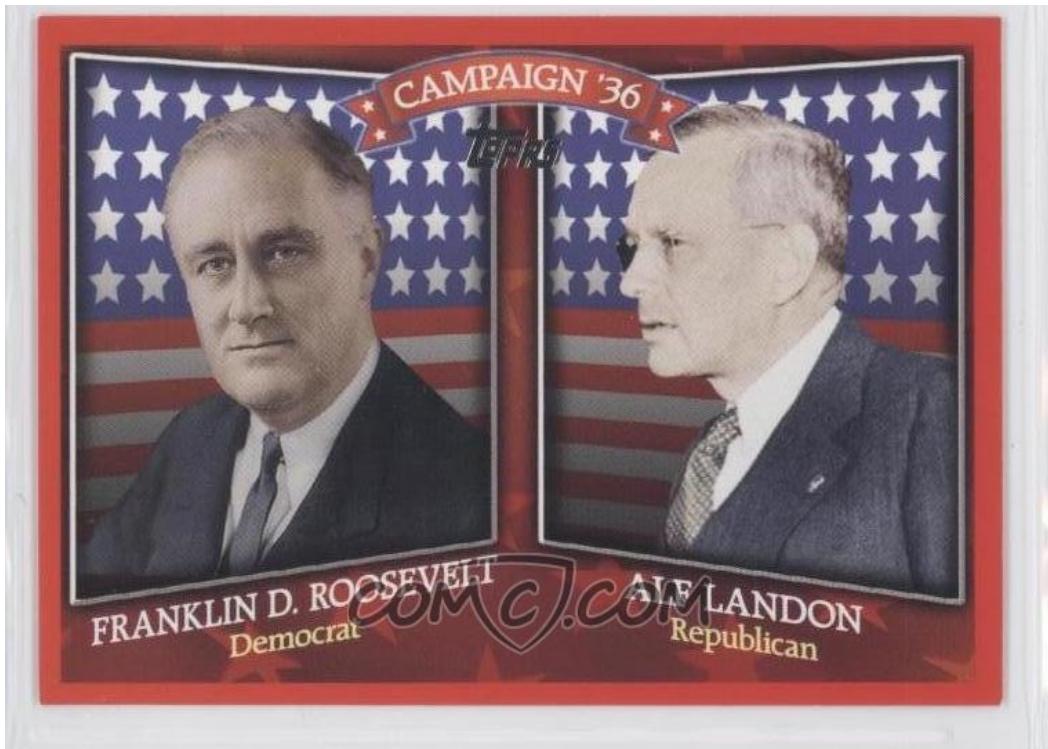
BIG DATA

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

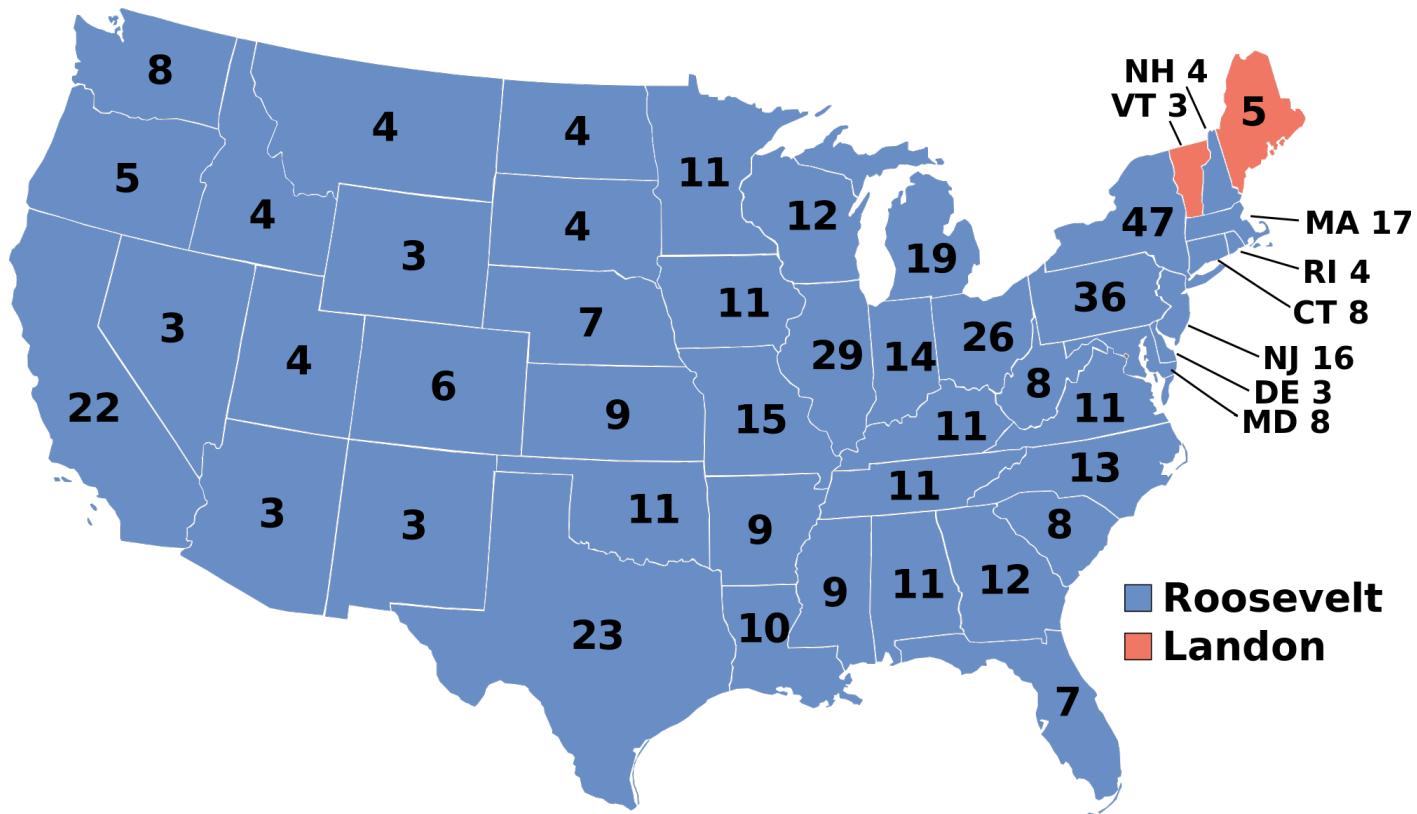


© 2009 R. Wang & Insider Associates, LLC. All rights reserved.

BIG DATA: A CAUTIONARY TALE



BIG DATA: A CAUTIONARY TALE



OTHER APPLICATIONS OF DATA SCIENCE & MACHINE LEARNING

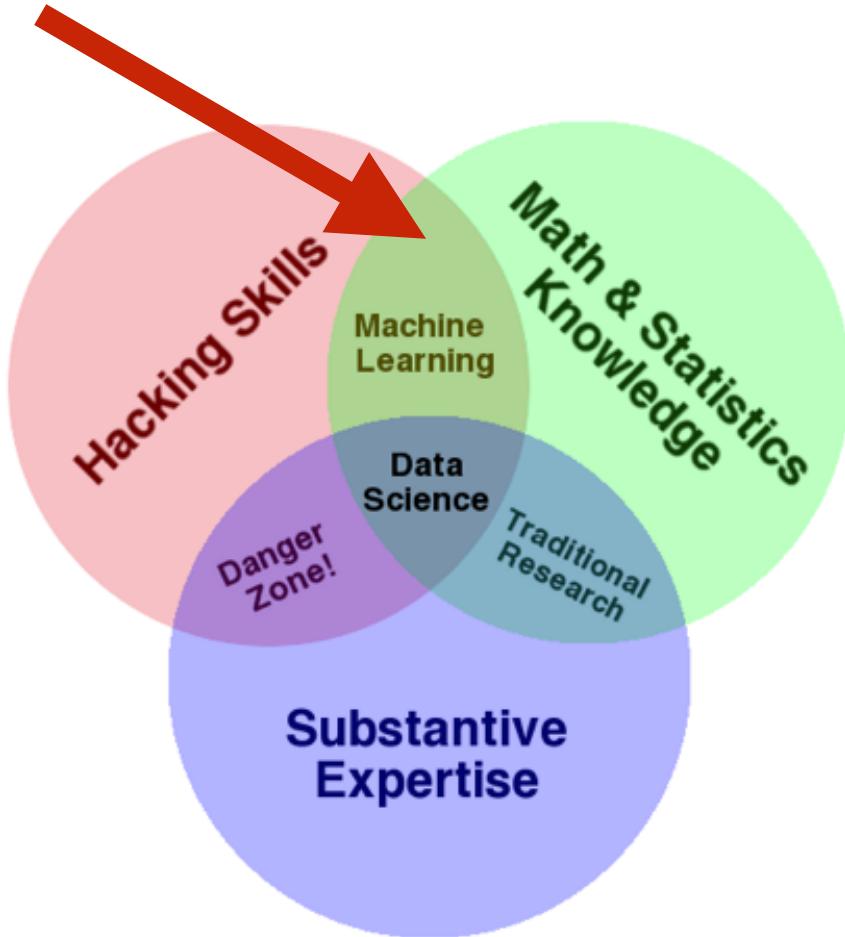
1. Search engines
2. Recommendation systems
3. Image recognition
4. Speech recognition
5. Gaming
6. Price comparison/optimisation
7. Route planning (driving, airlines, social network virality!)
8. Fraud / risk detection
9. Logistics (deliveries of goods, of people, of data)
10. Self-driving cars
11.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



PART III.

MACHINE LEARNING

YOU ARE HERE!



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

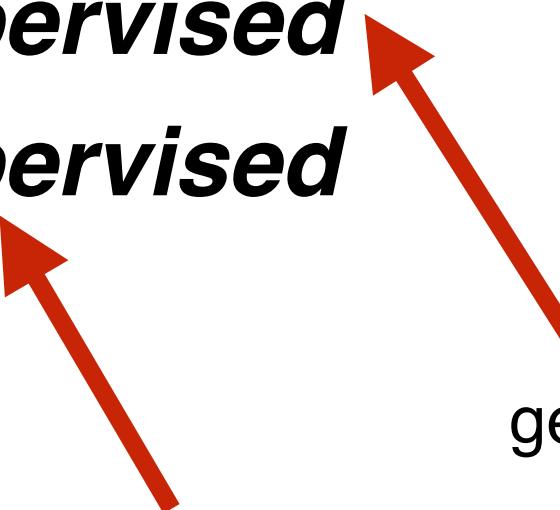
TYPES OF MACHINE LEARNING PROBLEM

supervised
unsupervised

making predictions
extracting structure

representation

generalisation



TYPES OF MACHINE LEARNING PROBLEM

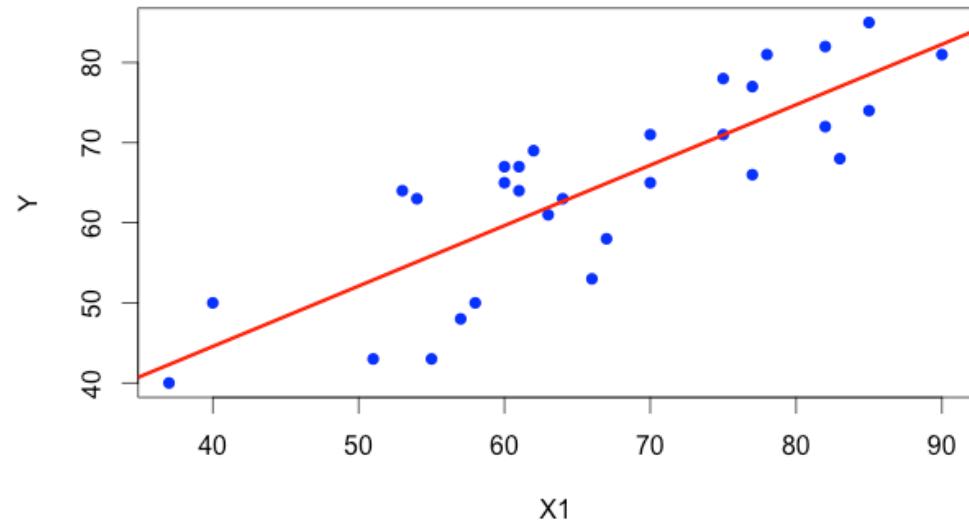
supervised

unsupervised

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

making predictions

extracting structure



TYPES OF MACHINE LEARNING PROBLEM

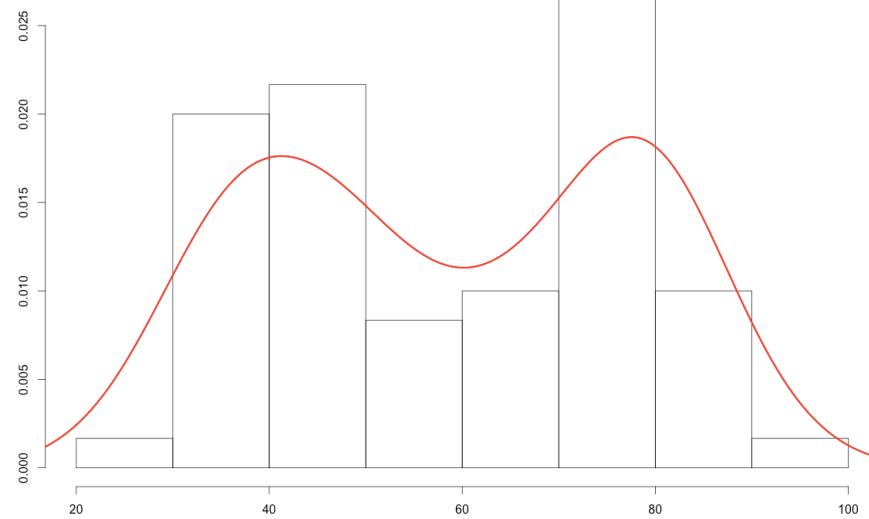
supervised

unsupervised

making predictions

extracting structure

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



TYPES OF DATA

continuous

quantitative

e.g. height

categorical

qualitative

e.g. eye colour

TYPES OF ML PROBLEMS

supervised

unsupervised

continuous

regression

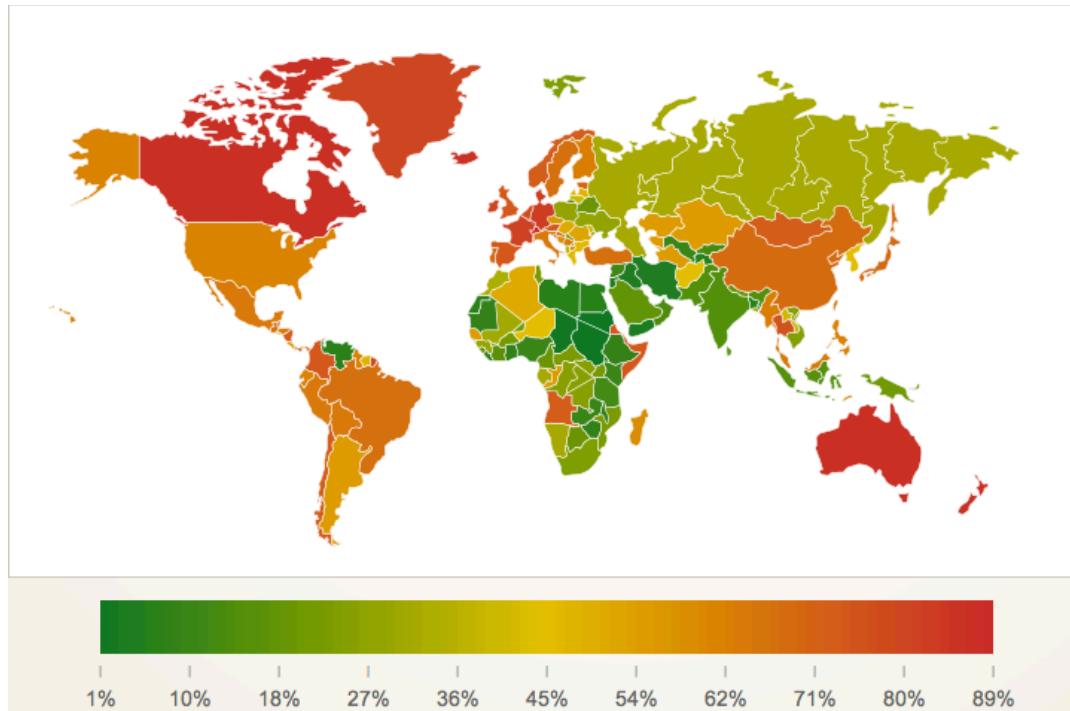
*dimensional
reduction*

categorical

classification

clustering

REGRESSION EXAMPLE: PREDICTING IPHONE SALES



GDP

population

Gini

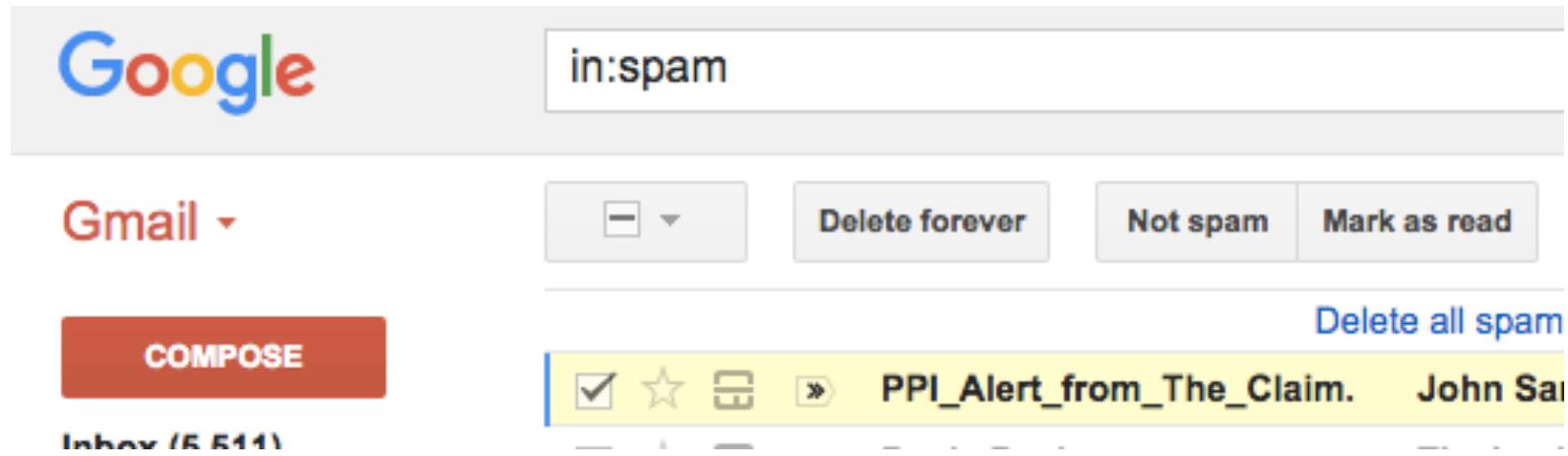
phone penetration %

GDP growth rate

TYPES OF ML PROBLEMS

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimensional reduction</i>	<i>clustering</i>

CLASSIFICATION EXAMPLE: SPAM FILTERING



\$\$\$

Act now!

As seen on

Satisfaction guaranteed

100% free

All natural

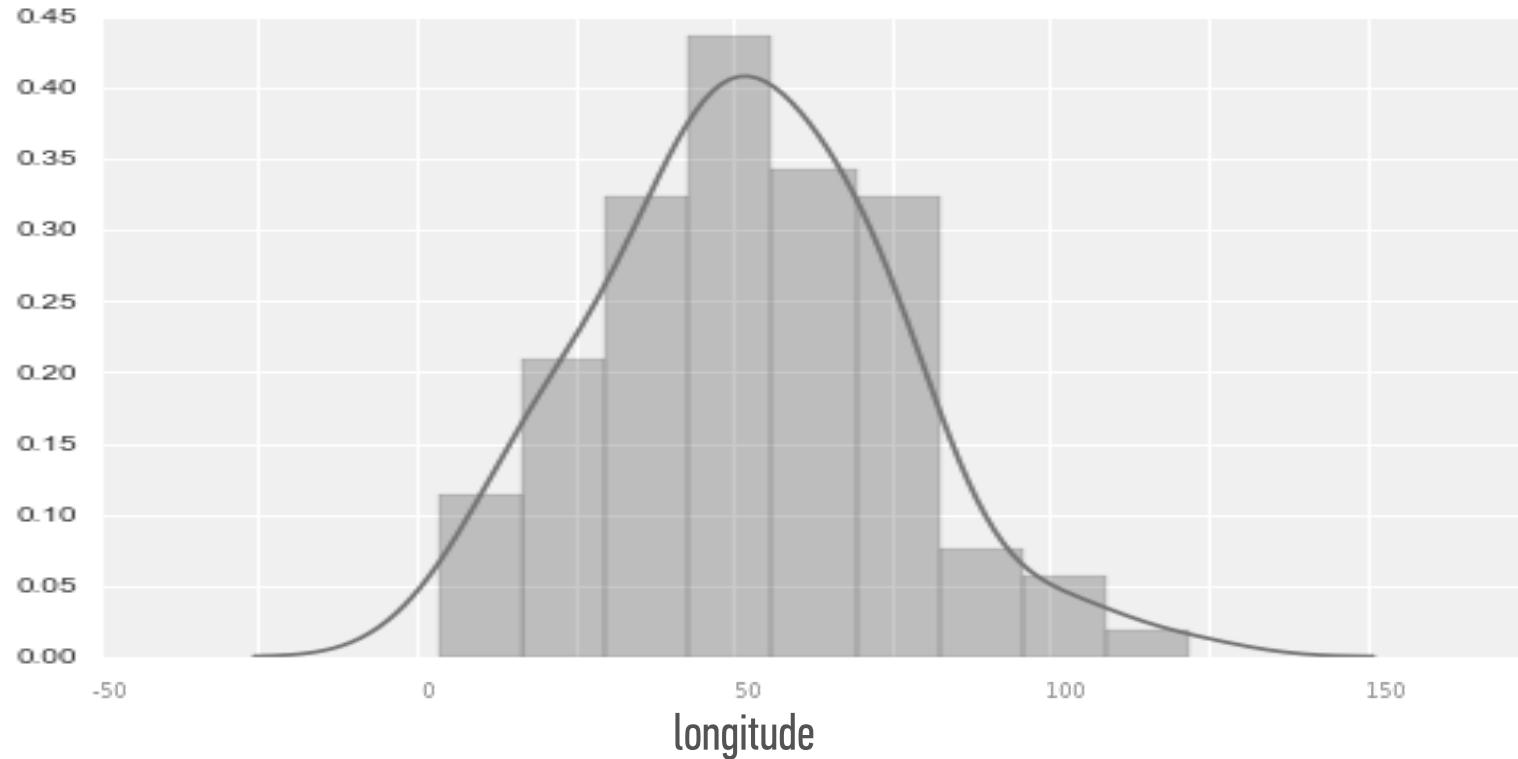
Bargain

!!!

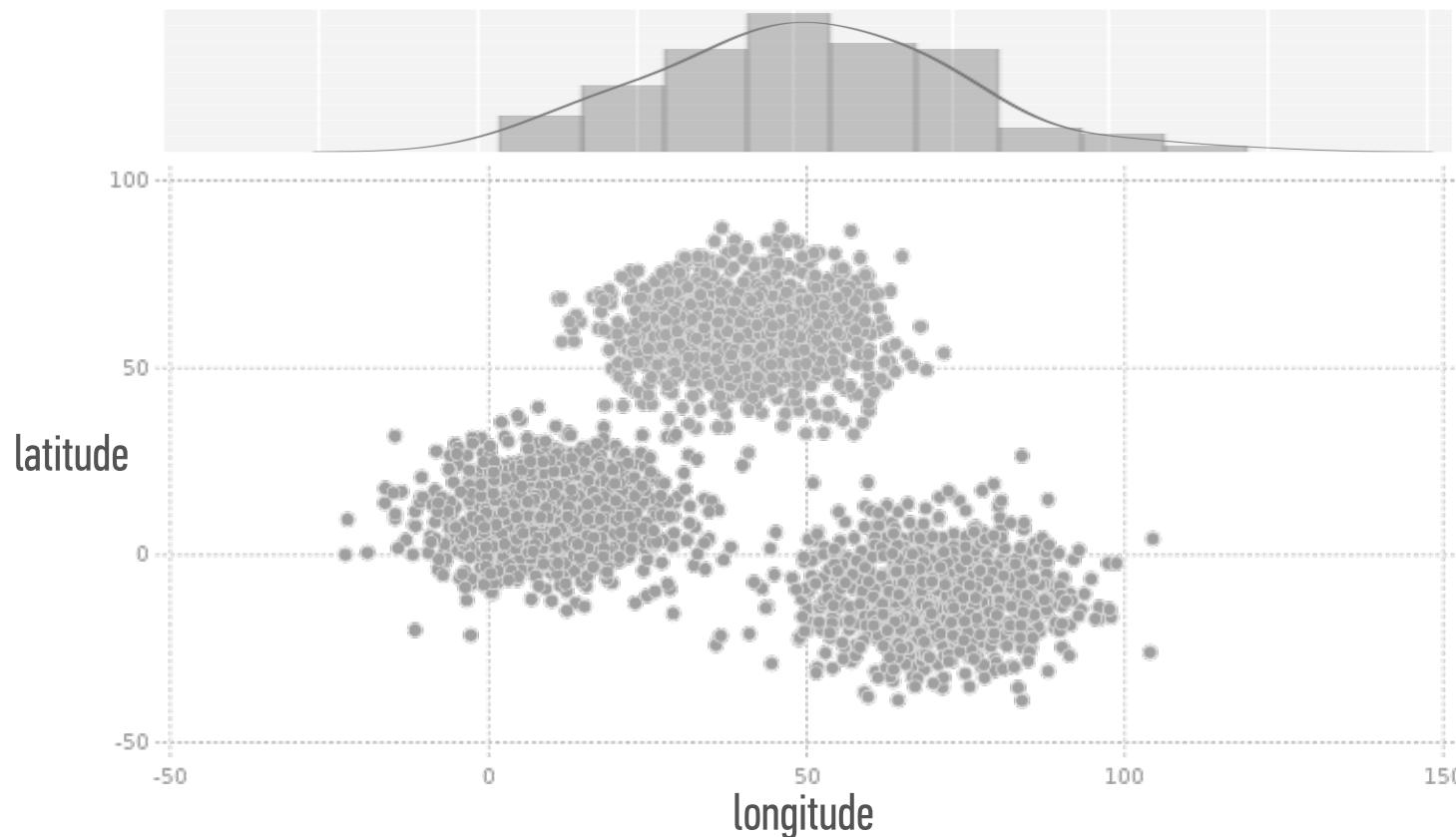
TYPES OF ML PROBLEMS

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimensional reduction</i>	<i>clustering</i>

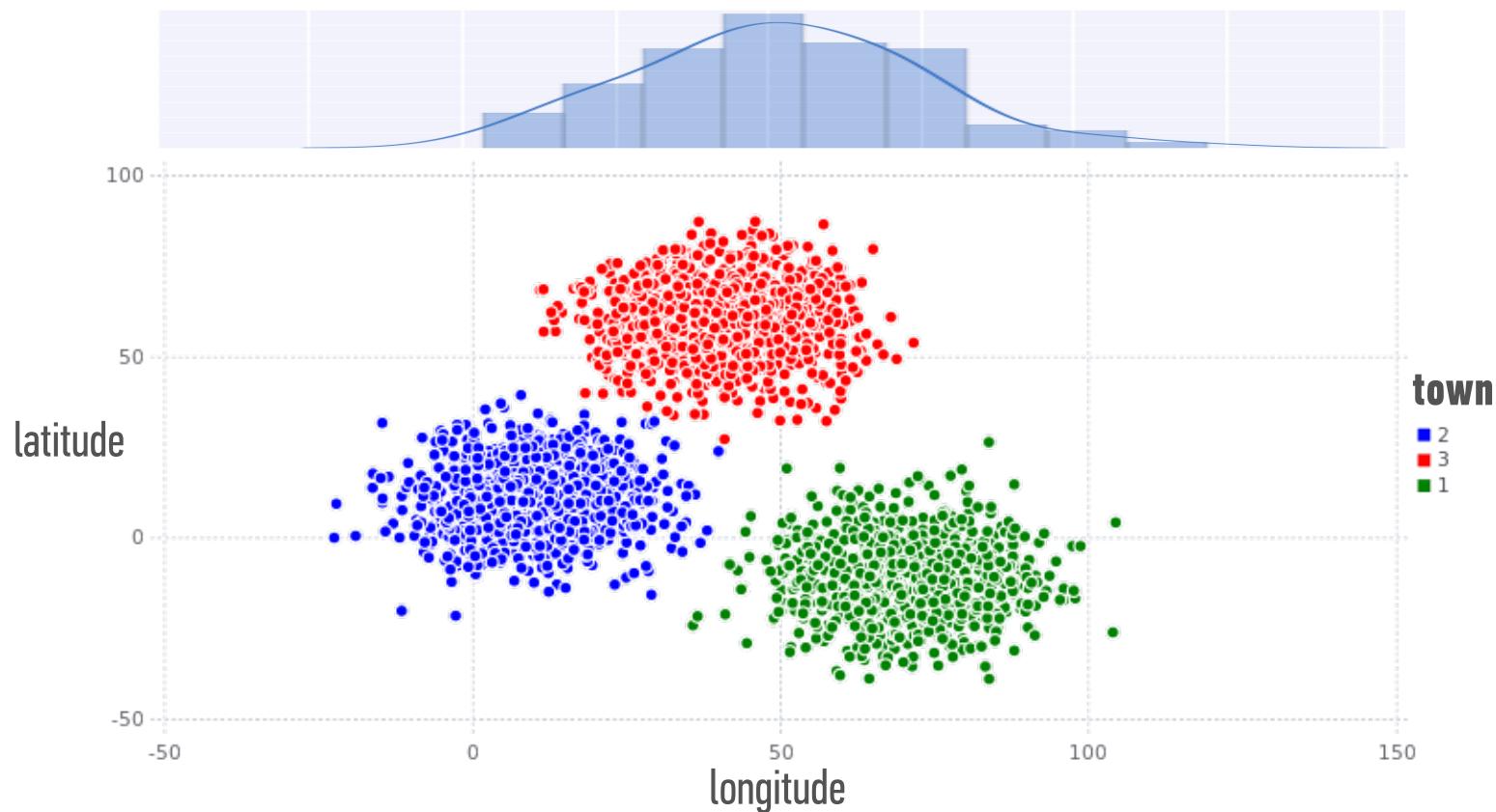
CLUSTERING EXAMPLE: USER LOCATIONS



CLUSTERING EXAMPLE: USER LOCATIONS



CLUSTERING EXAMPLE: USER LOCATIONS



TYPES OF ML PROBLEMS

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimensional reduction</i>	<i>clustering</i>

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

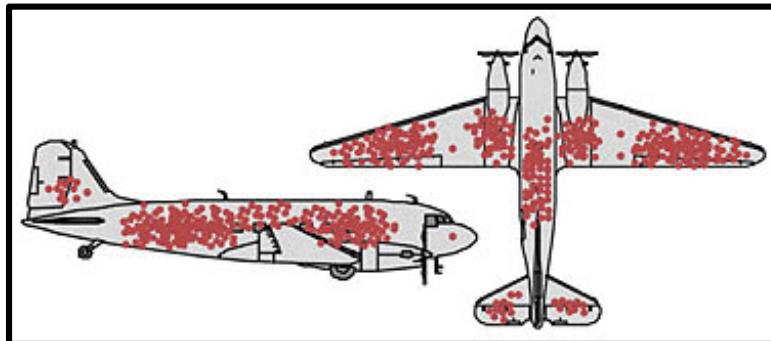
I have no special talent. I am only passionately curious. A. Einstein

RIGOUR

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



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

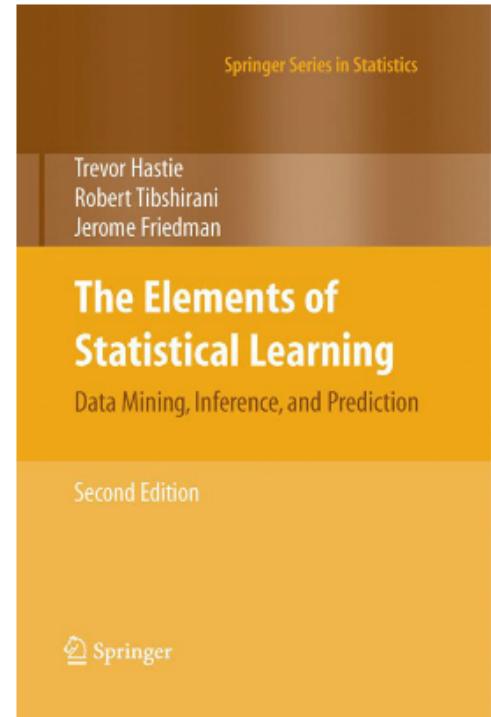
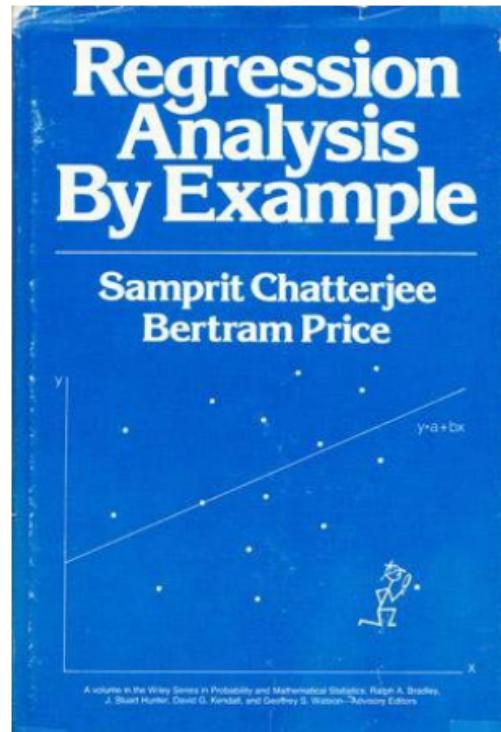
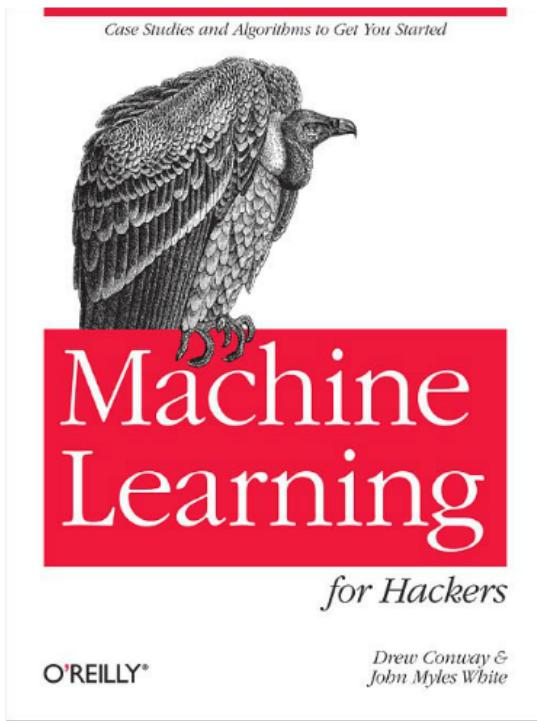
- In April 2012 McKinsey predicted 1.5 million shortage of data scientists
- More and more companies are looking for people to unlock the value in their data
- Rise in available positions

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%	

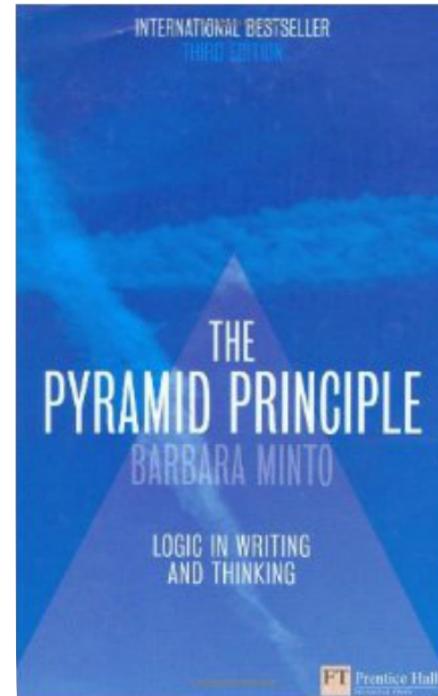
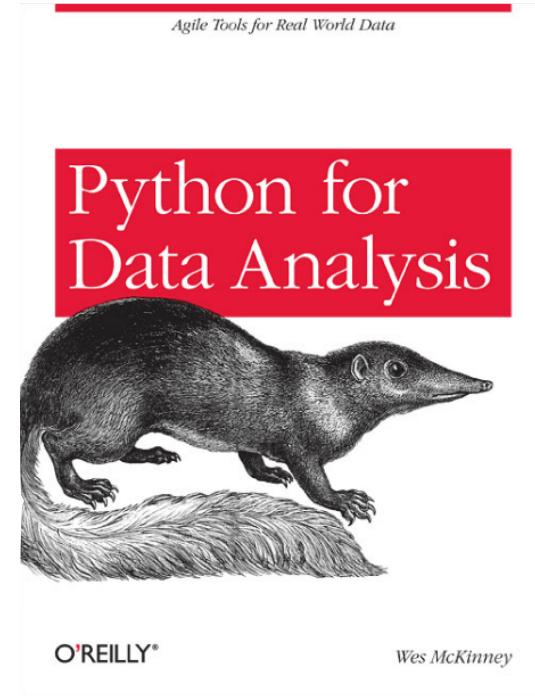
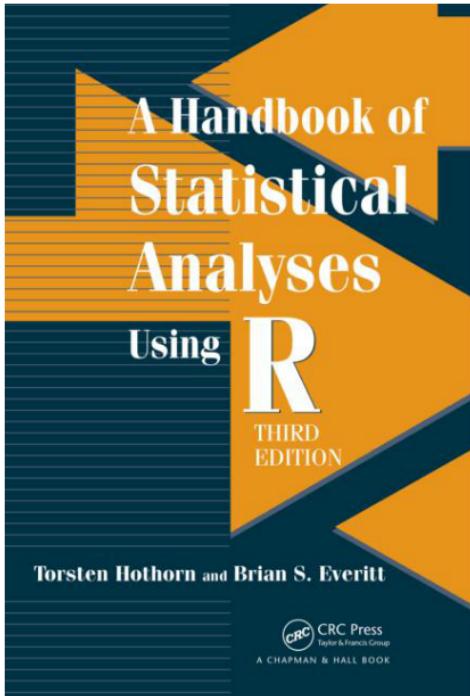
SHORTAGE OF SKILLS

- 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



MY TOP 3 BOOK RECOMMENDATIONS



ONLINE COURSES



Machine Learning
Andrew Ng (Stanford)



Machine Learning
CalTech CS156



DATAQUEST

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

{swirl}

swirlstats.com
"Learn R, in R"



DataCamp

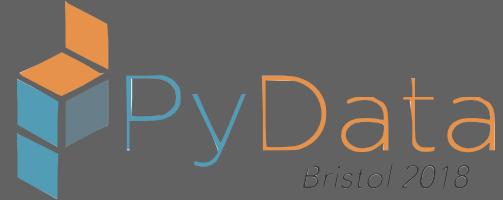
www.datacamp.com
*"Learn data analysis from the
comfort of your browser"
(R, Python, DataViz)*

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



pmc energy

Quora



Dropbox



Spotify®



reddit



YouTube

POWERED BY PYTHON



Gartner

Google

Honeywell

EVERNOTE



INDUSTRIAL
LIGHT & MAGIC
A LUCASFILM COMPANY



The Washington Post

Eventbrite®

the ONION®

splunk>

START AT PYDATA.ORG



ABOUT ▾ EVENTS ▾ DOWNLOADS SPONSOR ▾



A COMMUNITY FOR
DEVELOPERS AND USERS OF
OPEN SOURCE DATA TOOLS

[VIEW UPCOMING EVENTS](#)



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



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



PYDATA AMSTERDAM
MAY 25-27, 2018



PYDATA BERLIN
JULY 6-8, 2018

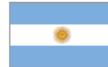


PYDATA EDINBURGH @ EUROPYTHON
JULY 25-29, 2018

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

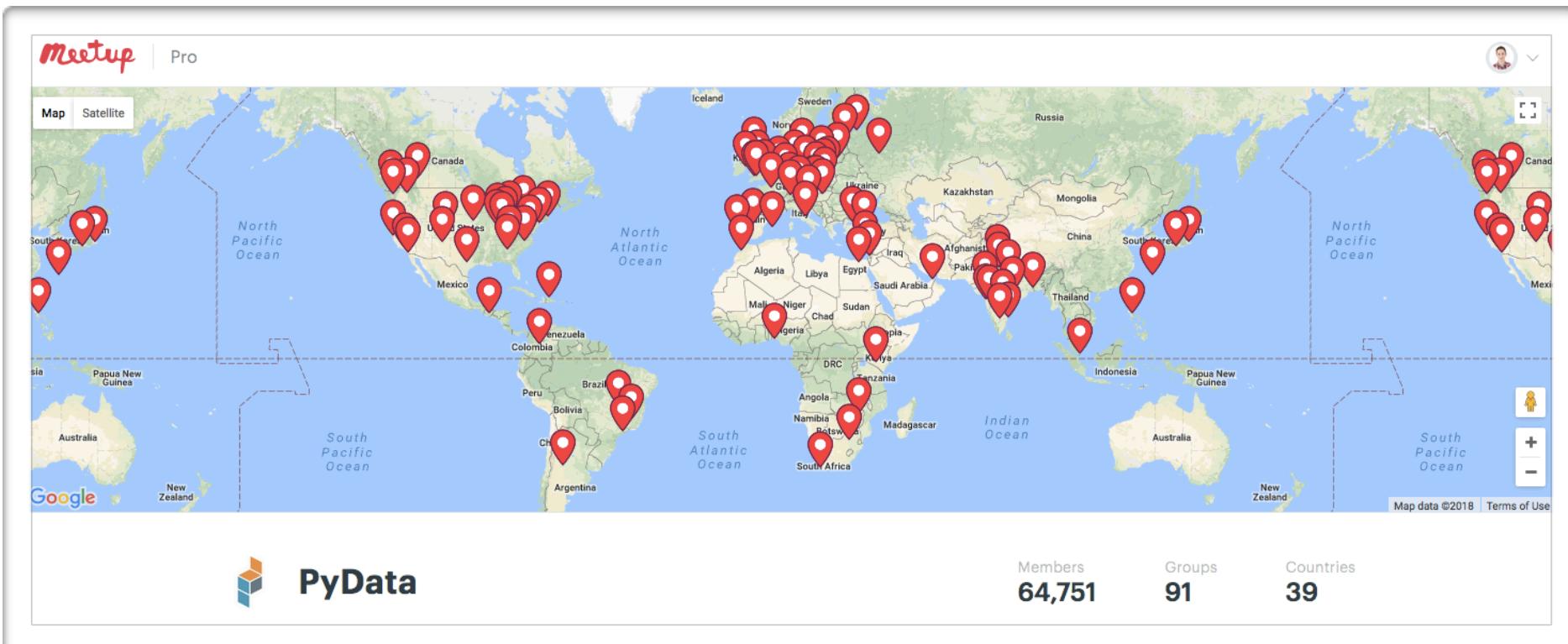


PYDATA WARSAW
NOVEMBER 19-20, 2018



PYDATA NYC
NOVEMBER 2018

MEETUPS



DOWNLOADS & SPONSORED PROJECTS

 PyData

ABOUT ▾ EVENTS ▾ DOWNLOADS SPONSOR ▾

ACCESS THE PYTHON OPEN DATA SCIENCE STACK

Download Cutting Edge Tools in Data Science

DOWNLOADS

Logos with [] around them are NumFOCUS Sponsored Projects



PACKAGES TO START WITH

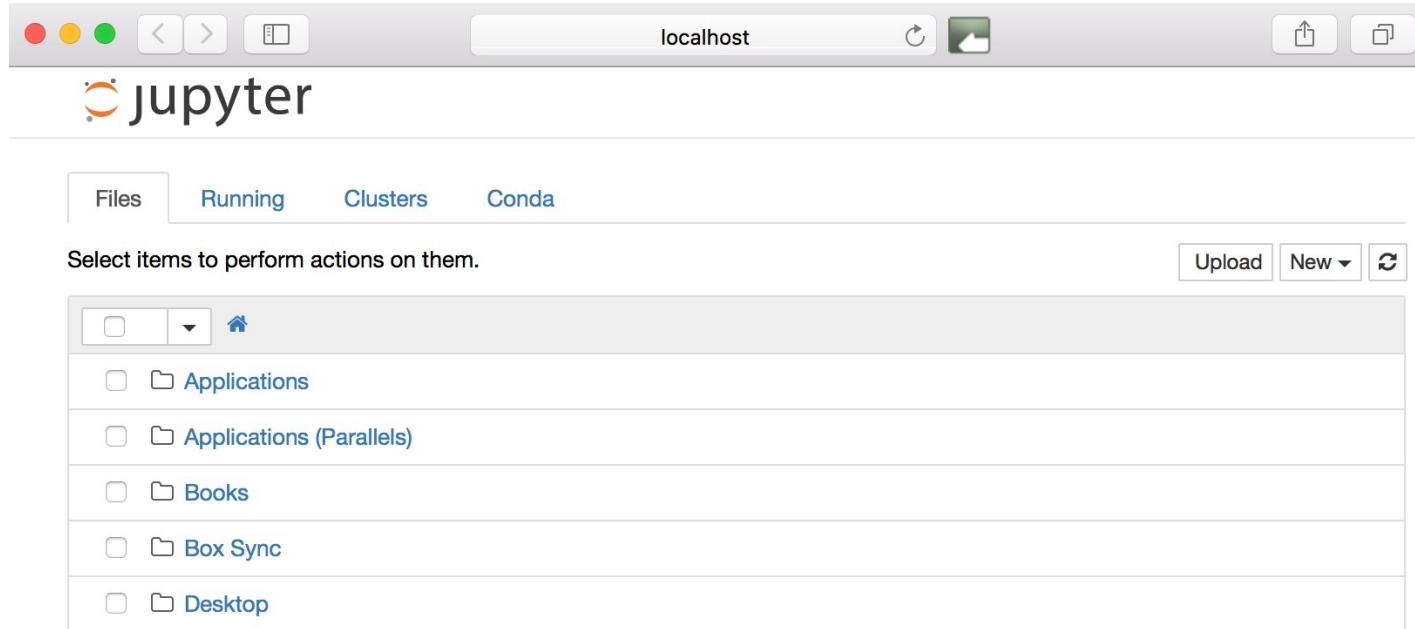
- ▶ **pandas**: manipulate data
- ▶ **SciPy/NumPy**: scientific computing and numerical calculations
- ▶ **Scikit-learn**: machine learning
- ▶ **matplotlib/Seaborn**: data visualisation
- ▶ **spacy/nltk**: natural language processing
- ▶ **statsmodels**: statistical tests
- ▶ **Beautiful Soup**: HTML/XML 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.

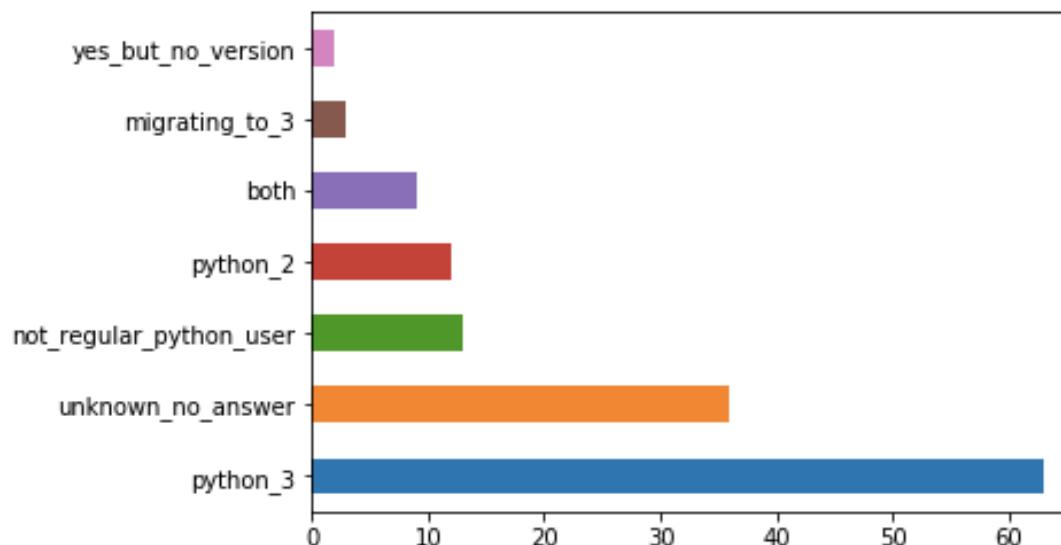


DEMO

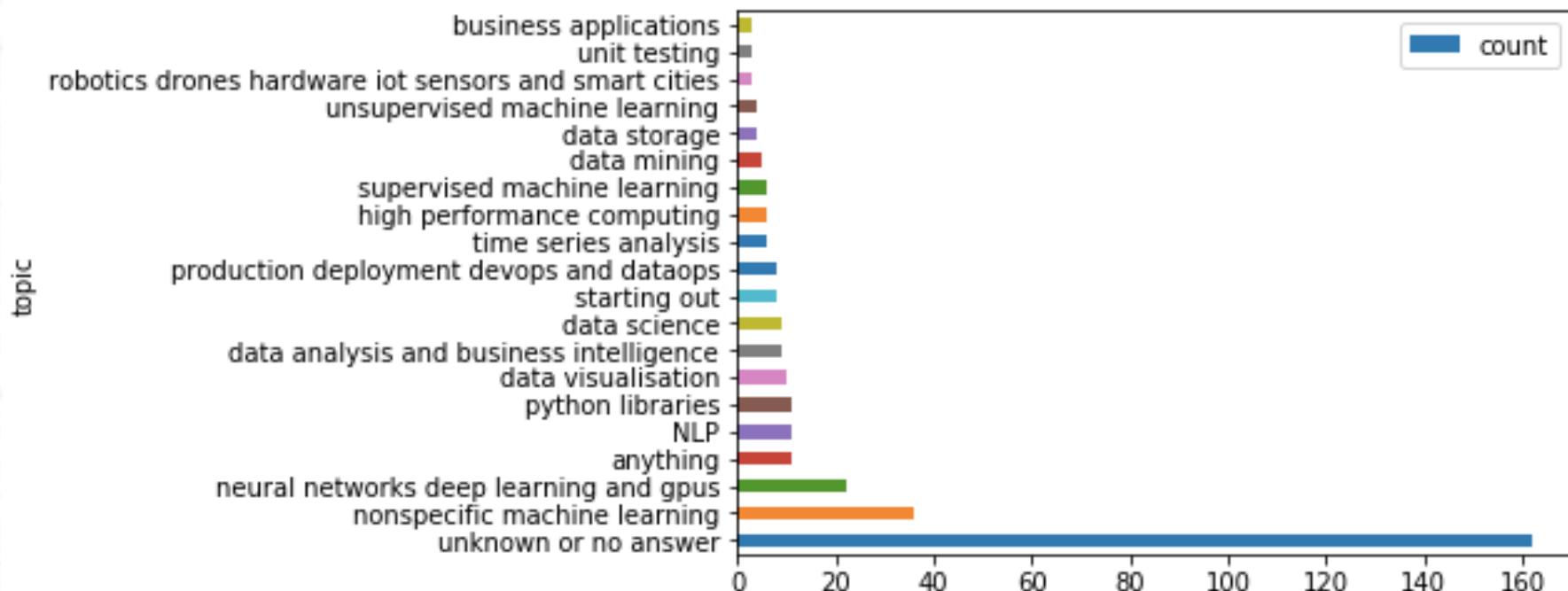
80% of PyData Bristol members use Python 3 vs 20% using Python 2!

** 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"

Historical reporting.

Metrics. KPIs. Segmentation.

Dashboards. BI tools. Pivot tables.

Necessary...keeps the engines running.

Tools: Excel, SQL, Tableau.

"Data Science"

Predictive forecasting.

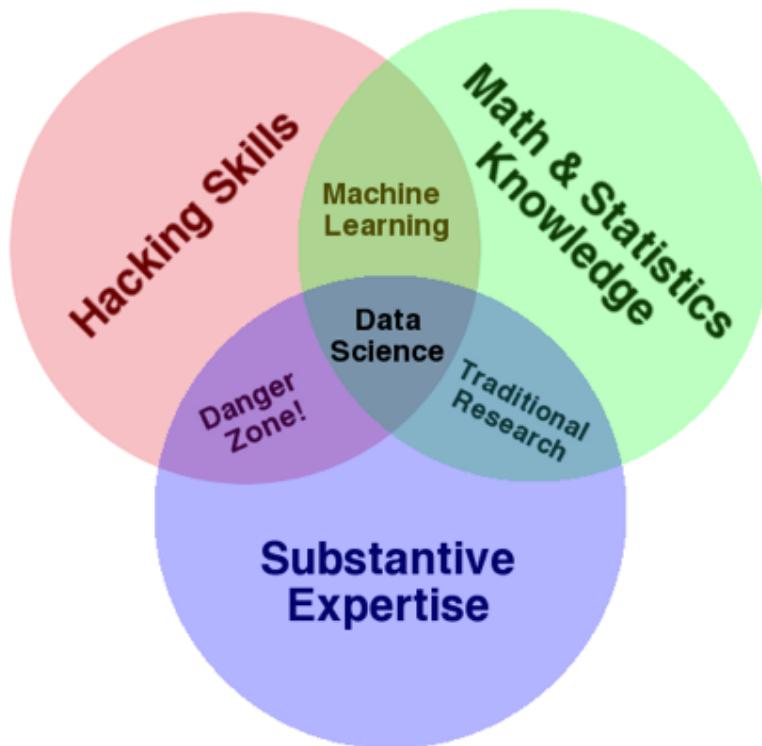
Statistics. Regression. Machine learning.

Coding. Flexibility. Automation.

Exciting...unexpected insights.

Tools: Python, R, scikit-learn.

"DATA 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

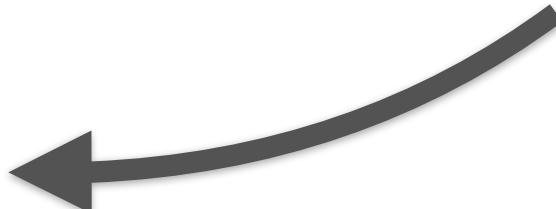
BRISTOL MEETUPS!

- ▶ 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

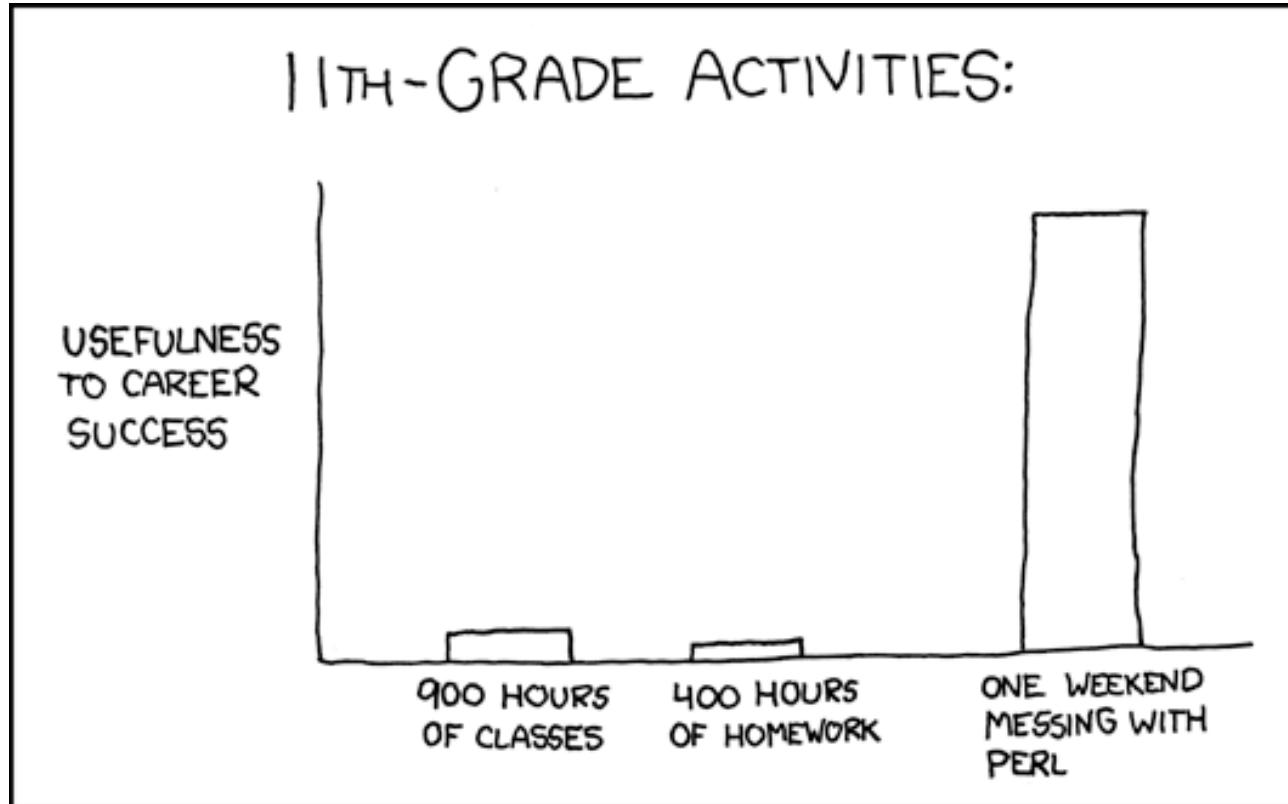
HACKATHONS & DATADIVES

- › 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



MARCH 10, 2017

[MINI] The Perceptron

▶ PLAY 14:46 [!\[\]\(ae75f41d8c940537376a84fb7eef4933_img.jpg\) Download](#)

Today's episode overviews the perceptron algorithm. This rather simple approach is characterized by a few particular features. It updates its weights after seeing every example, rather than as a batch. It uses a step function as an activation function. It's only appropriate for linearly separable data, and it will converge to a solution if the data meets these criteria. Being a fairly simple algorithm, it can run very efficiently. Although we don't discuss it in this episode, multi-layer perceptron networks are what makes this technique most attractive. [View More](#)

dataskeptic.com

THE CORNERSTONE OF A DAUNTING FUTURE?

The screenshot shows the homepage of the website waitbutwhy.com. The header features the site's logo "WAIT BUT WHY" in large orange letters, with "new post every sometimes" in smaller blue text below it. To the right of the logo is a cartoon illustration of a king-like figure with a crown and a sword, surrounded by small stick figures. On the far right of the header is a green icon of a tent and the text "中文". Below the header is a navigation bar with links: home, about, archive, minis, the shed, dinner table, store, and support wbw. To the right of the navigation bar is a search bar and a feed icon. The main content area features a large blue title "The AI Revolution: The Road to Superintelligence" by Tim Urban. Below the title is a note explaining the long writing process. To the right of the title is a yellow sidebar with a sign-up form for email notifications. At the bottom of the page is a footer with social media links and a Patreon button.

The AI Revolution: The Road to Superintelligence

By Tim Urban

Note: The reason this post took three weeks to finish is that as I dug into research on Artificial Intelligence, I could not believe what I was reading. It hit me pretty quickly that what's happening in the world of AI is not just an important topic, but by far THE most important topic for our future. So I wanted to learn as much as I could about it, and once I did that, I wanted to make sure I wrote a post that really explained this whole situation and why it matters so much. Not shockingly, that became outrageously long, so I broke it into two parts. This is Part 1—Part 2 is [here](#).

We are on the edge of change comparable to the rise of human life on Earth. — Vernor Vinge

Join 174,026 other humans and have new posts emailed to you

Email Address

SEND ME NEW POSTS

Follow us using these special men

215,527  34,726  4,089 

Help us exist with Patreon 

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?