



PyData  
Cardiff



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

John Sandall  
11th April 2018

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

# MY BACKGROUND



Imperial College  
London

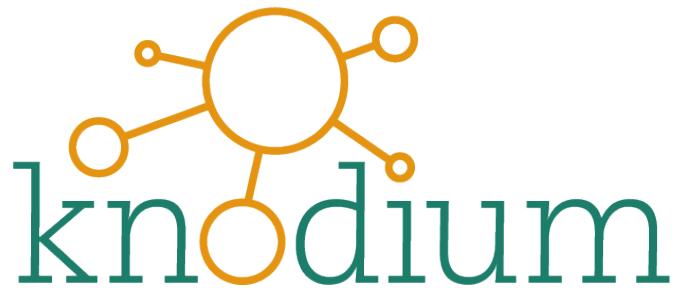
2010

# MY BACKGROUND



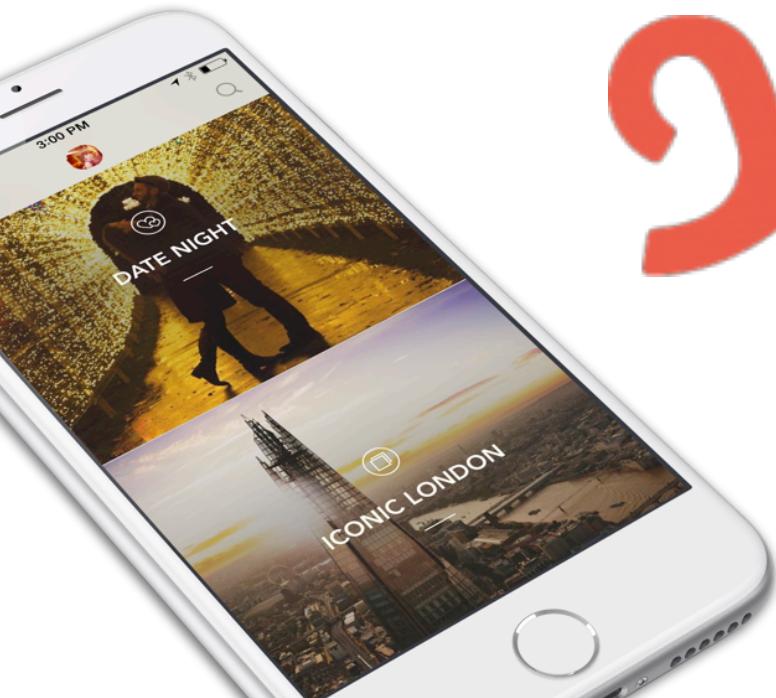
2011

# MY BACKGROUND



2012

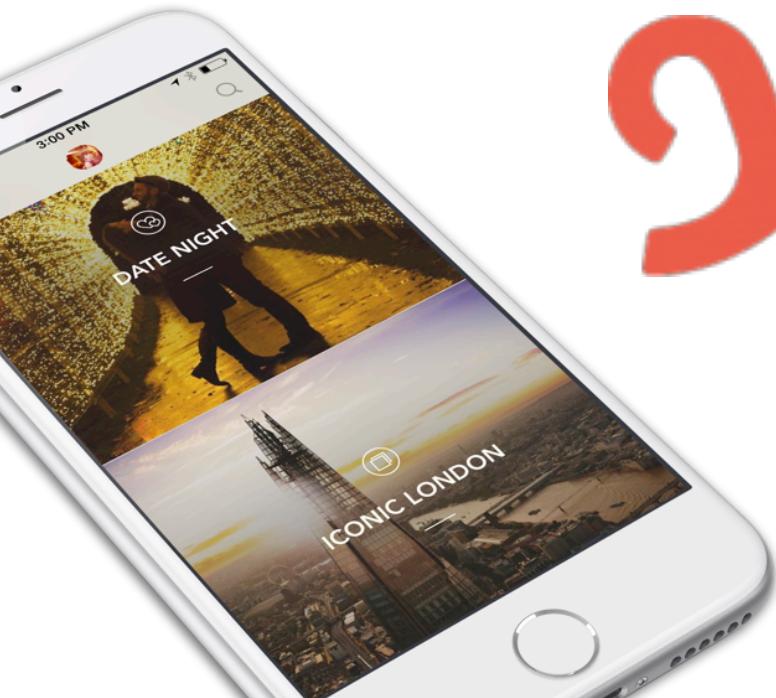
# MY BACKGROUND



# yPlan

2013

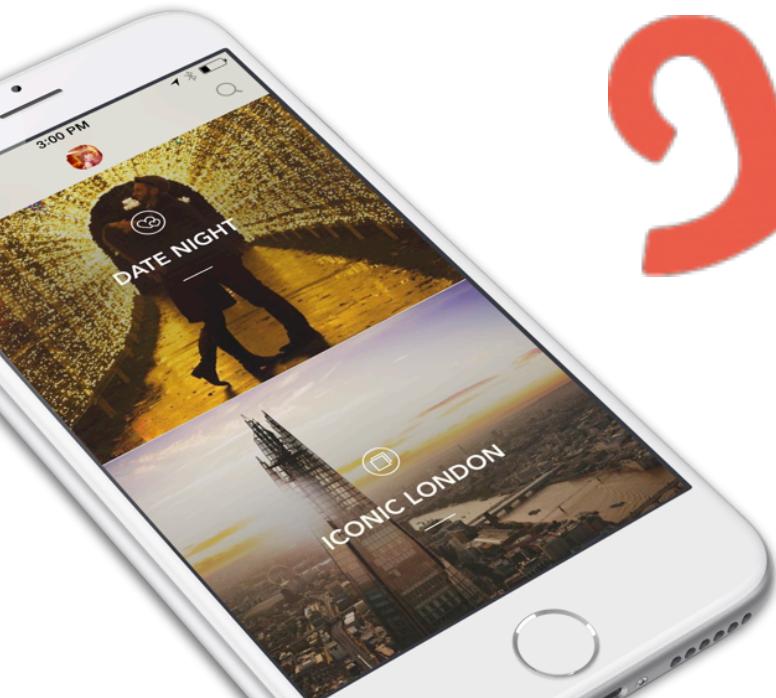
# MY BACKGROUND



# yPlan

2014

# MY BACKGROUND



# yPlan

2015

# MY BACKGROUND



BNP PARIBAS



ACCEL<sup>®</sup>  
PARTNERS



Felix



OmnicomGroup

jbi training  
THE IT PROFESSIONALS CHOICE



RACING POST



HealthUnlocked



Kingston Smith

2016-2018

# AGENDA

- I. What is Data Science?
- II. The Last 10 Years
- III. Machine Learning 101
- IV. The PyData Ecosystem
- V. Tips For Success

# 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.

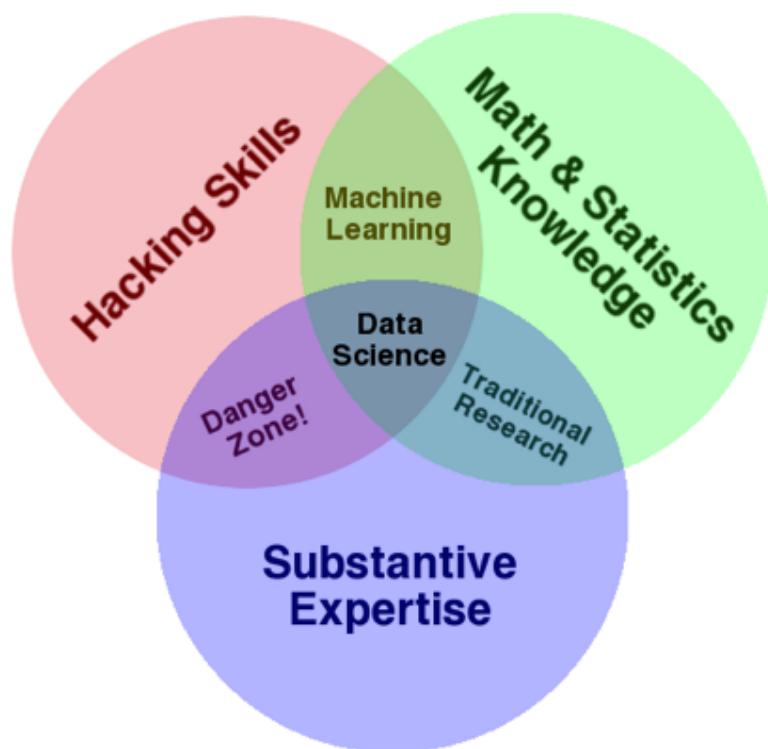
# WHAT IS DATA SCIENCE?

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

# 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.

# THE QUALITIES OF A DATA SCIENTIST



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

# 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.

# 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

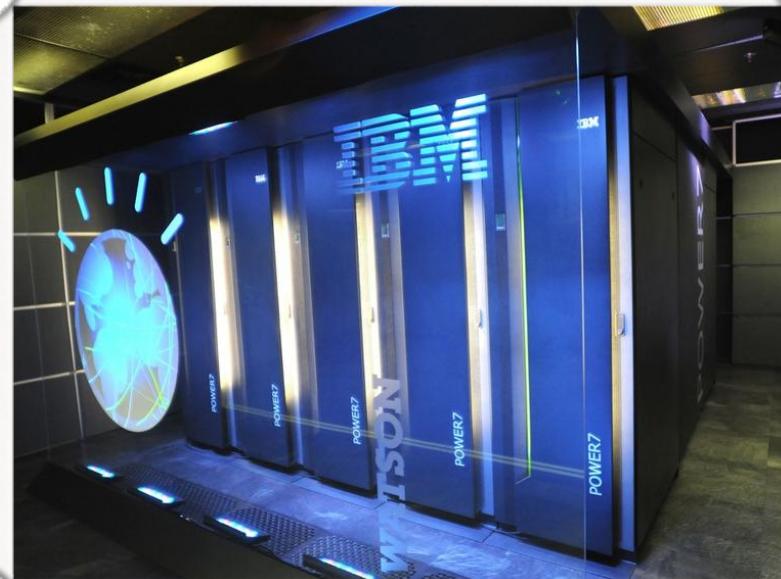
 **caroliniine** just did a great call with Lloyd Alter for my green marketing story! [about 2 hours ago](#) from [twitterific](#)

 **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: INFORMATION REVOLUTION

IBM Watson wins Jeopardy TV gameshow in 2011



Kindle launches

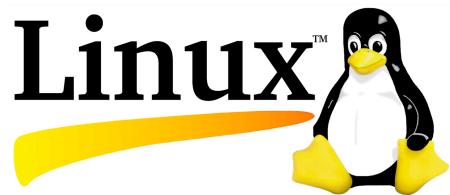


IBM Watson created

# 2007: THE OPEN SOURCE ECOSYSTEM ACCELERATES



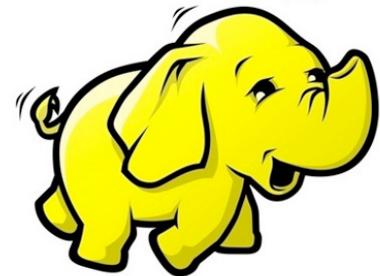
# 2007: THE OPEN SOURCE ECOSYSTEM ACCELERATES



# 2007: THE OPEN SOURCE ECOSYSTEM ACCELERATES



*hadoop*



python

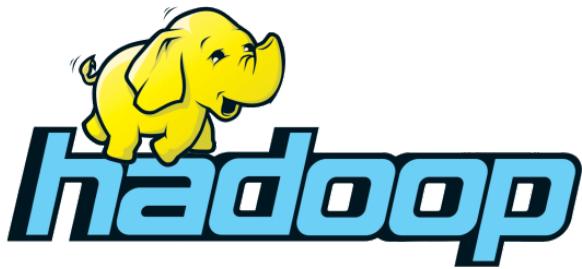


## 2007: R & PYTHON START TO BE ENTERPRISE FRIENDLY

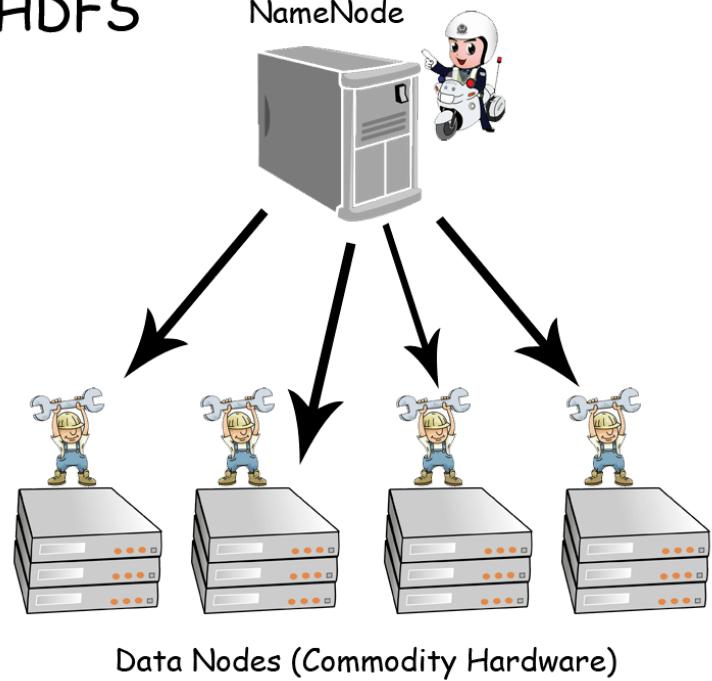
The screenshot shows a news article from Computerworld. At the top left is the Computerworld logo with three horizontal bars followed by the word "COMPUTERWORLD". To the right is a green button labeled "INSIDER" with a white arrow pointing right, and "Sign In |". Below the logo, the word "NEWS" is written in red. The main headline in large black font reads "Microsoft unwraps a big-data analytics platform based on R".

The screenshot shows a news article from PCWorld. At the top left is the PCWorld logo with three horizontal bars followed by "PCWorld" and "FROM IDG" below it. Below the logo, the word "NEWS" is written in red. The main headline in large black font reads "Anaconda's Python-based analytics hit the enterprise with new subscription plans". Below the headline, a smaller text line reads "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).

# **EVOLUTION OF DATA CREATION**

**2011:**

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

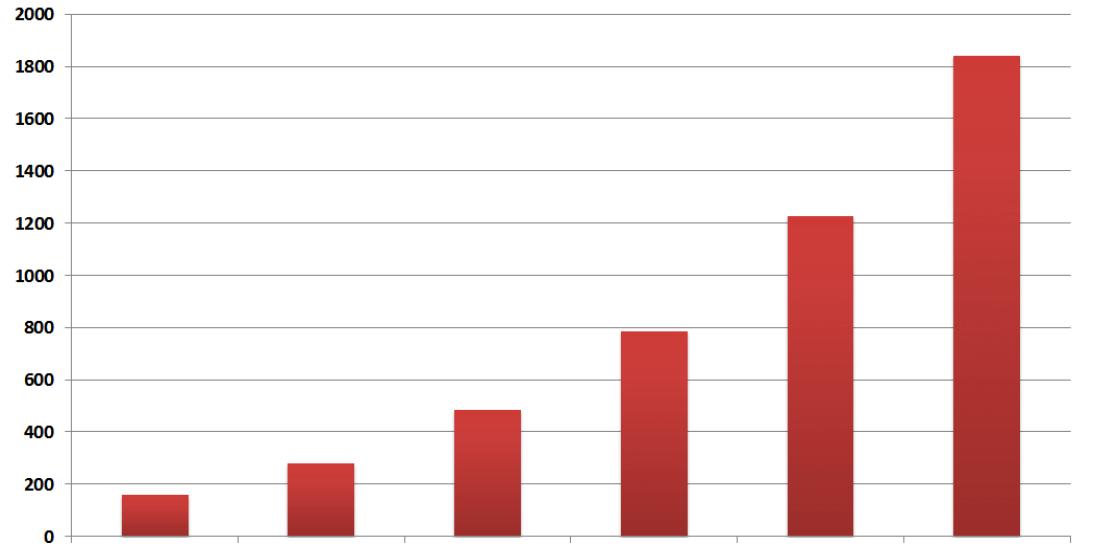
**1 exabyte (EB) = 1000 petabytes (PB) = 1 billion gigabytes (GB)**

# EVOLUTION OF DATA CREATION

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

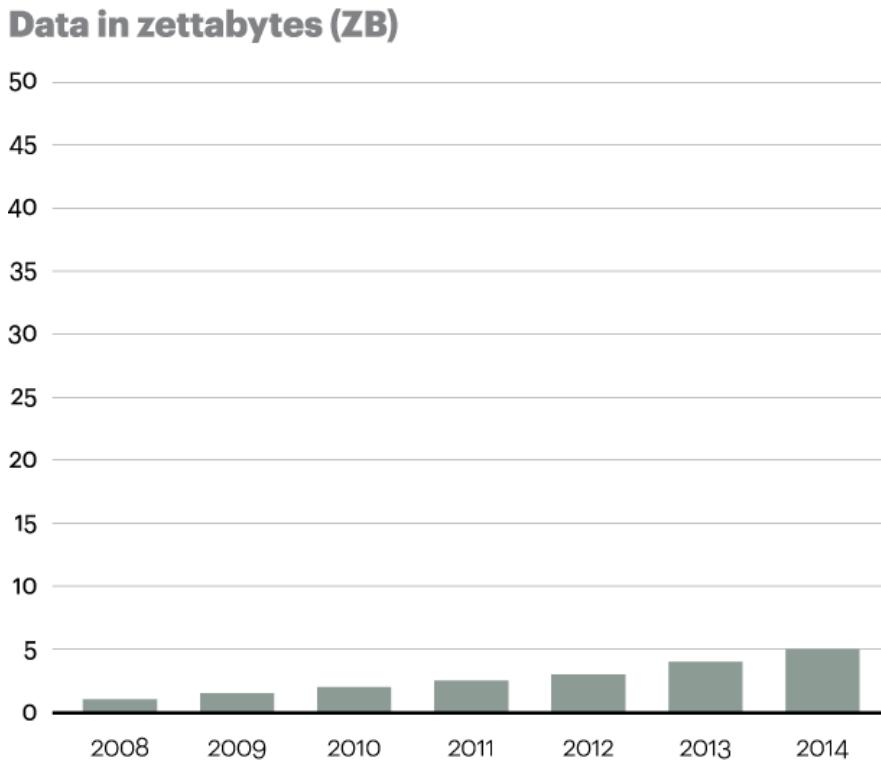
License: <http://creativecommons.org/licenses/by-sa/2.5/ca/>

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

# EVOLUTION OF DATA CREATION

2014:

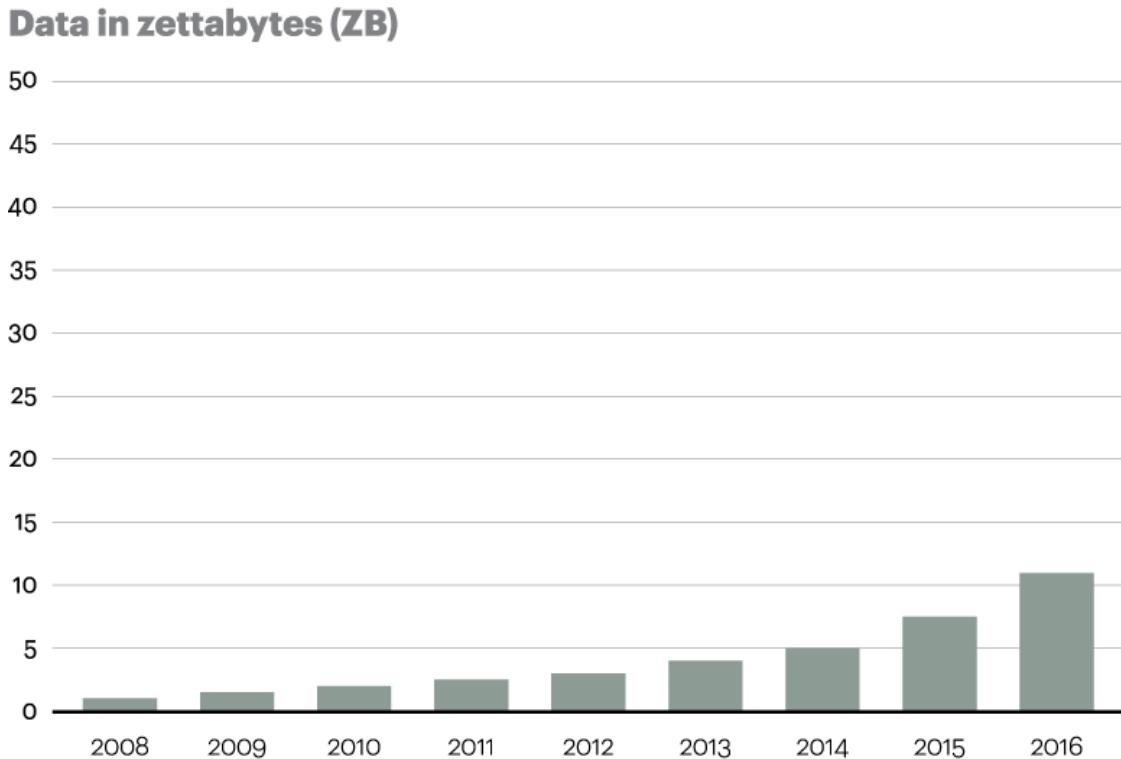
Oracle estimates total  
data created annually  
now surpasses five  
Zettabytes



# EVOLUTION OF DATA CREATION

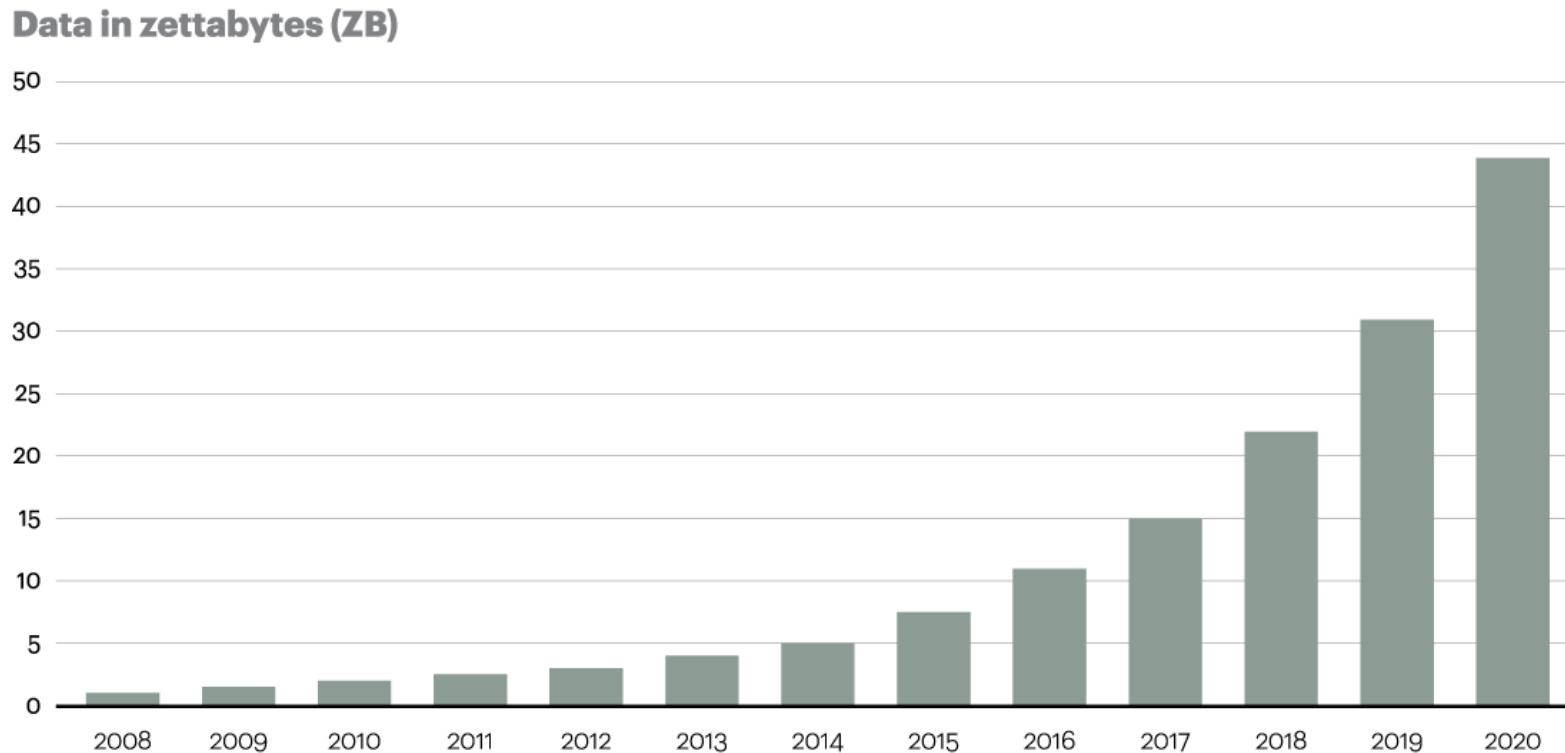
**2016:**

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



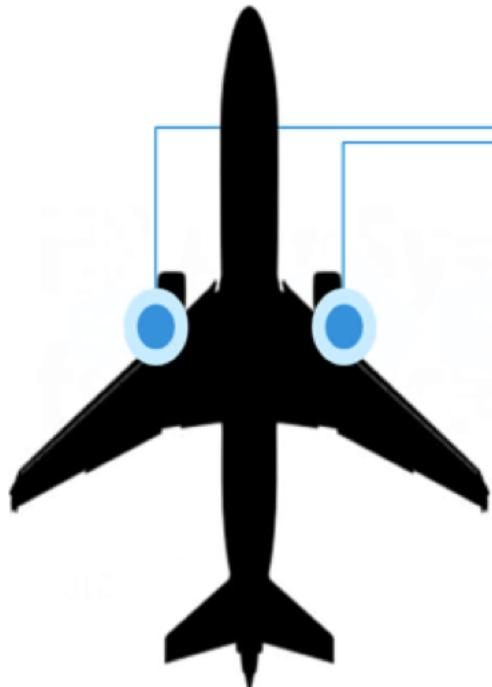
# EVOLUTION OF DATA CREATION

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

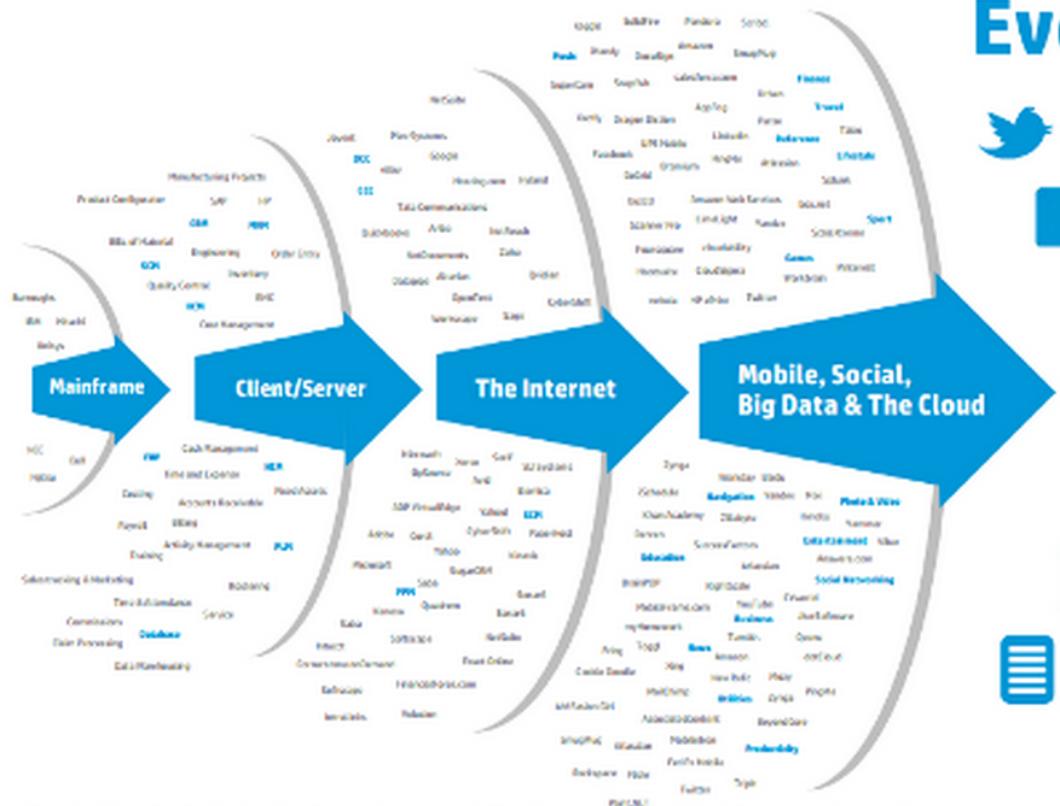
six-hour, cross-  
country flight from  
New York to Los  
Angeles

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

days in a year

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

# WELCOME TO DATA OBESITY!



# Every 60 seconds

 98,000+ tweets

**f** 695,000 status updates

 11million instant messages

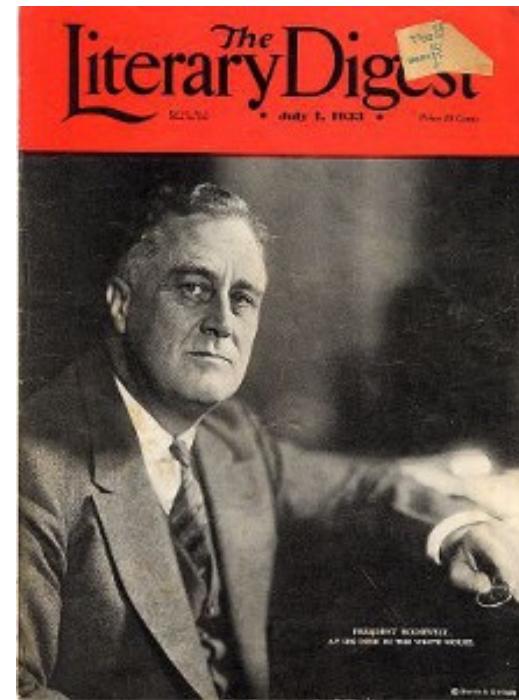
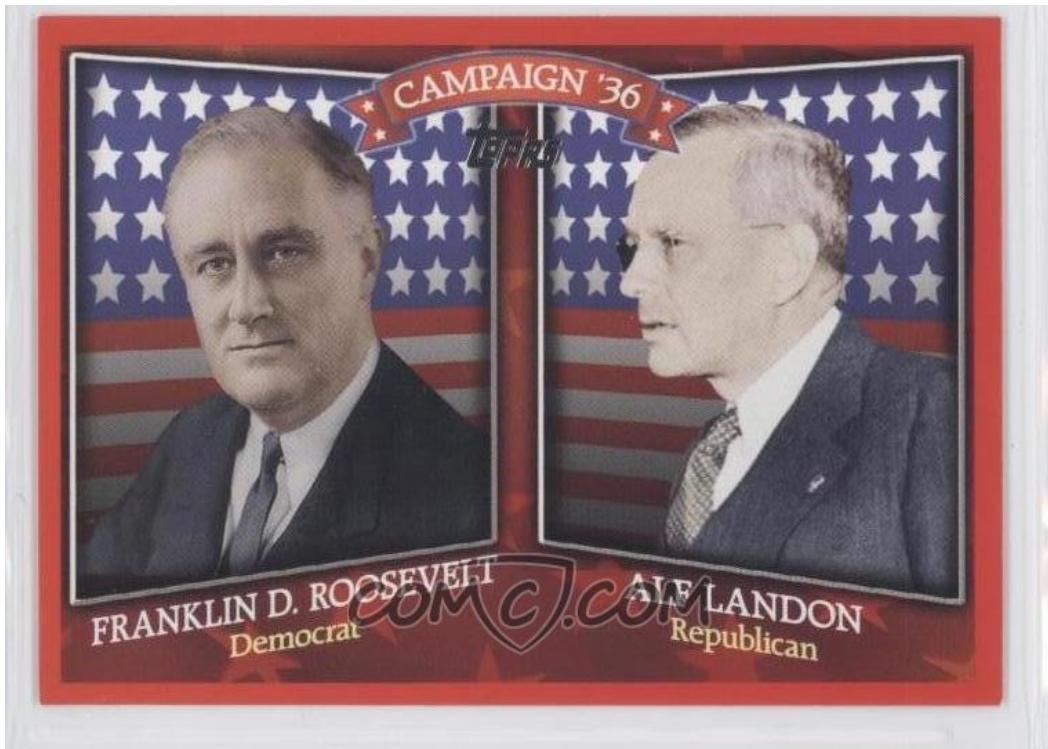
 698,445 Google searches

 168 million+ emails sent

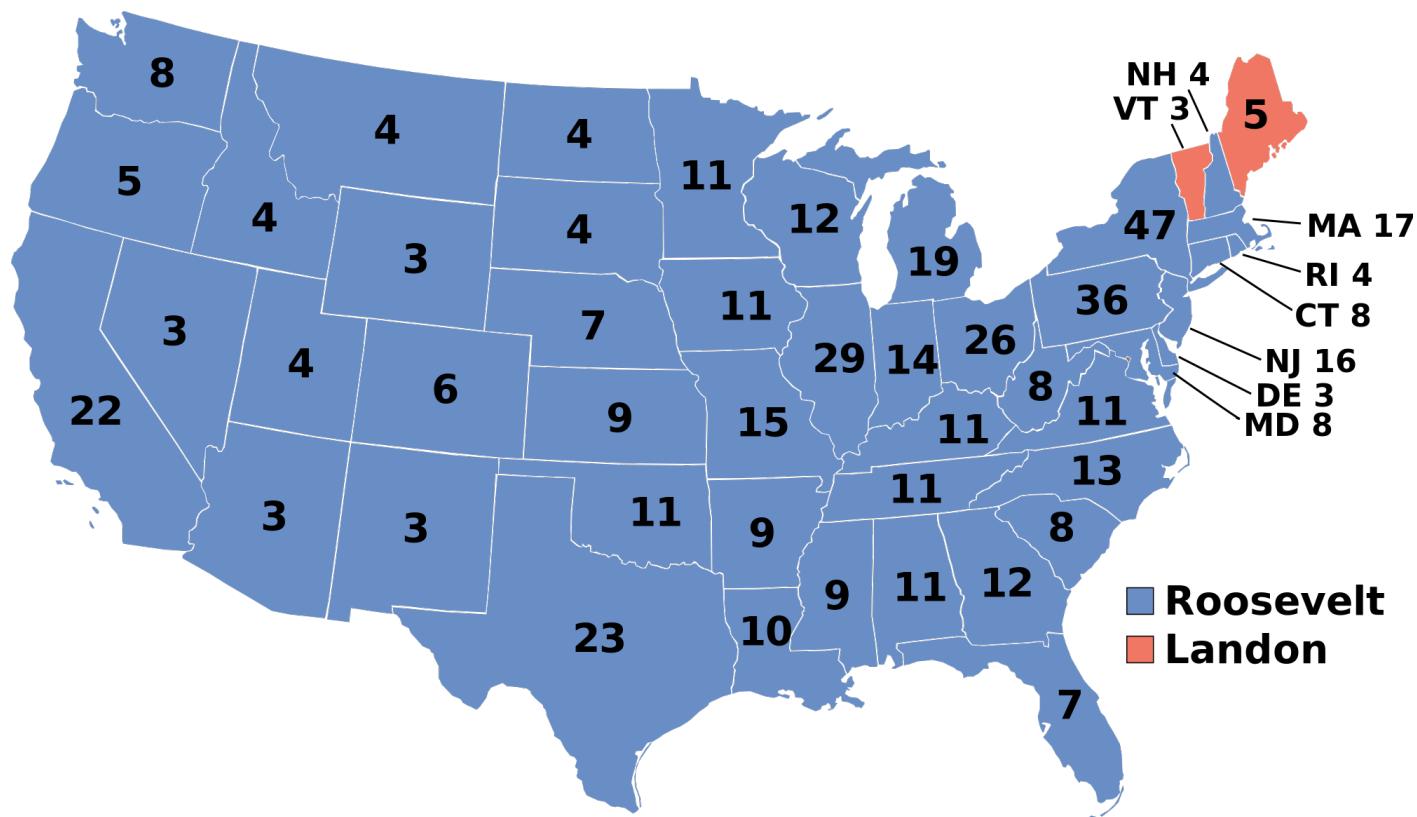
 **1.820TB** of data created

## 217 new mobile web users

# BIG DATA: A CAUTIONARY TALE



# **BIG DATA: A CAUTIONARY TALE**



# **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. **...**

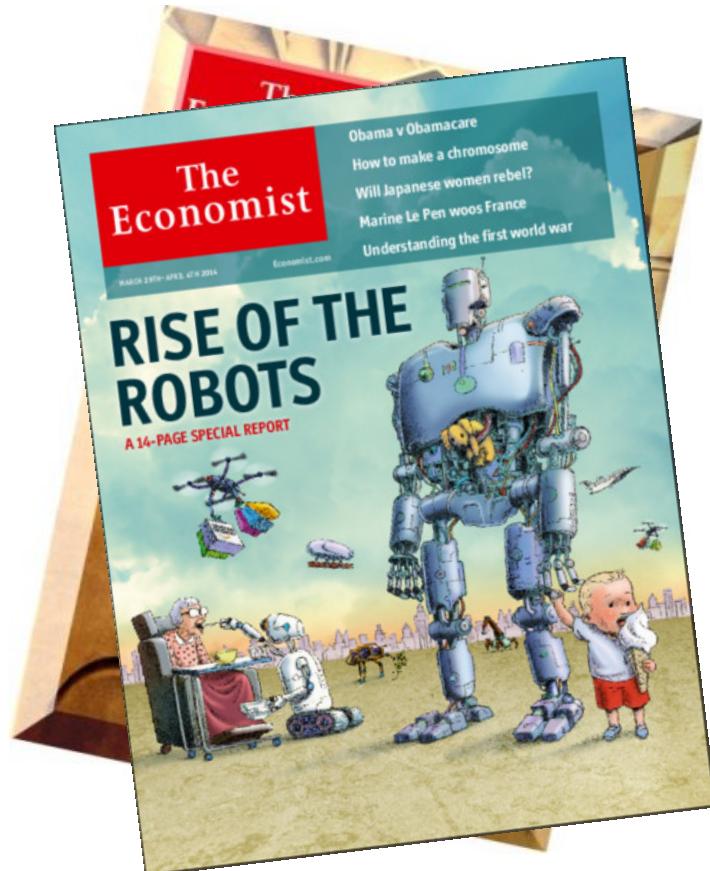
# APPLICATIONS OF DATA SCIENCE & MACHINE LEARNING

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. ...



# OTHER APPLICATIONS OF DATA SCIENCE & MACHINE LEARNING

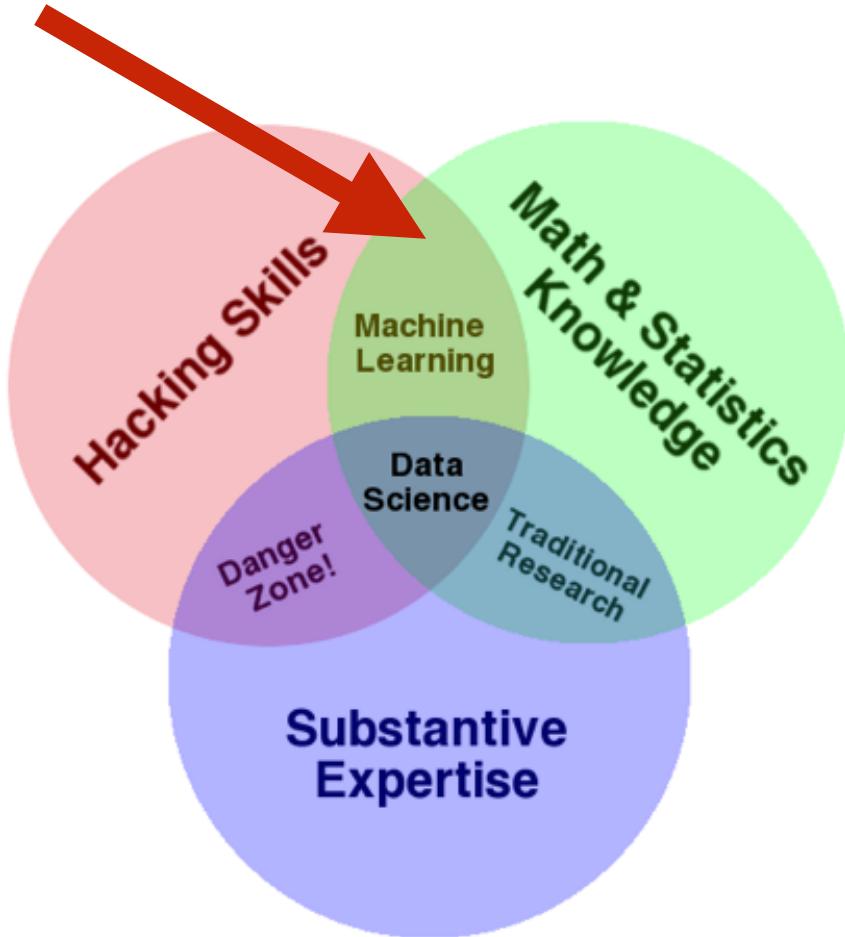
1. Search engines
2. Recommendation systems
3. Image recognition
4. Speech recognition
5. Gaming
6. Price comparison/optimisation
7. Route planning (driving, airlines, social network virality!)
8. Fraud / risk detection
9. Logistics (deliveries of goods, of people, of data)
10. Self-driving cars
11.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

# 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**..."

# 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*

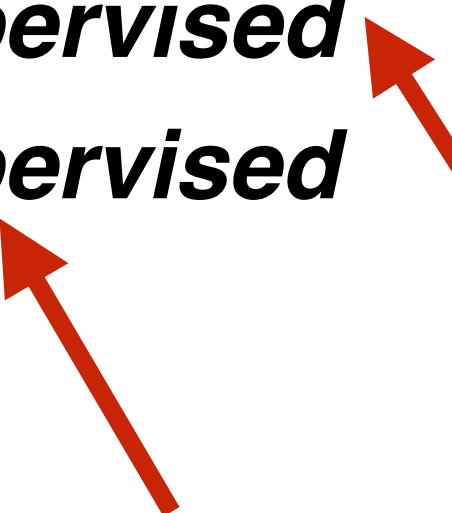
# TYPES OF MACHINE LEARNING PROBLEM

***supervised***  
***unsupervised***

*making predictions*  
*extracting structure*

representation

generalisation



# TYPES OF MACHINE LEARNING PROBLEM

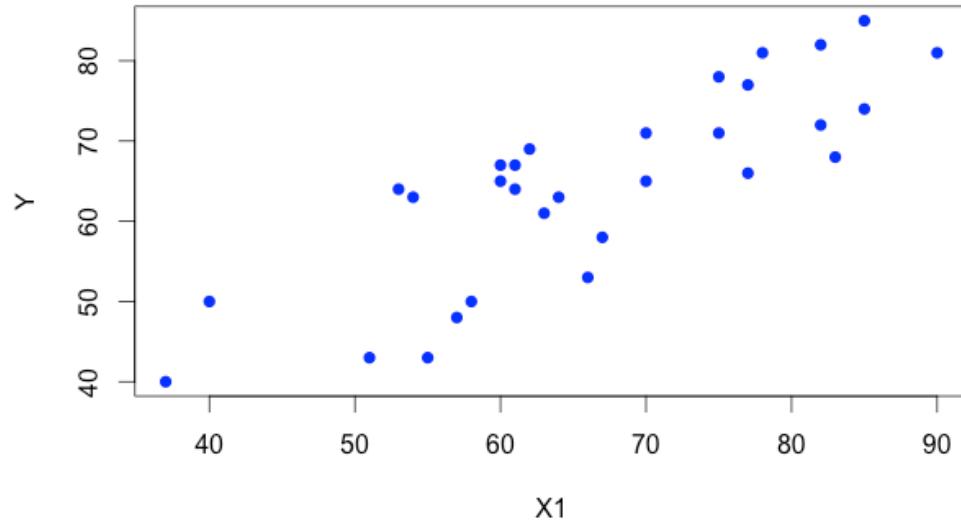
***supervised***

*unsupervised*

*making predictions*

*extracting structure*

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



# TYPES OF MACHINE LEARNING PROBLEM

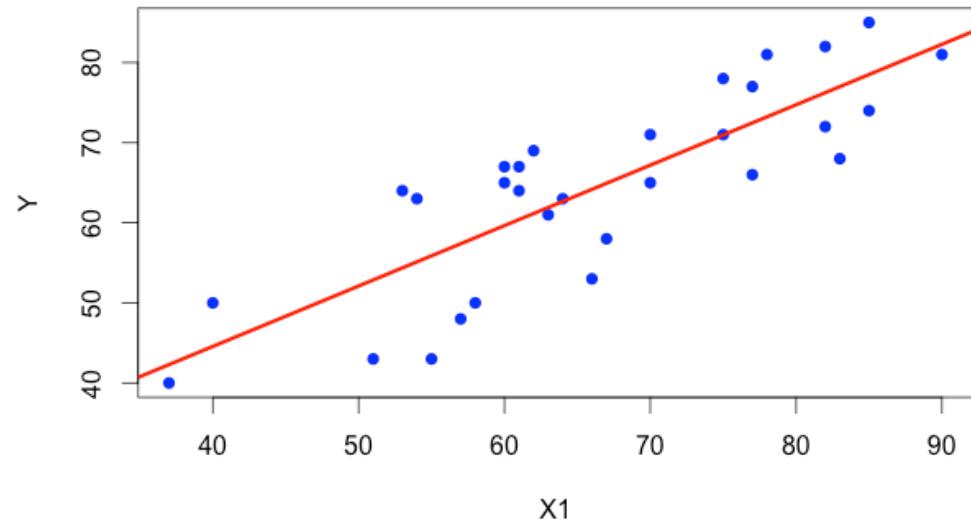
***supervised***

*unsupervised*

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

*making predictions*

*extracting structure*



# TYPES OF MACHINE LEARNING PROBLEM

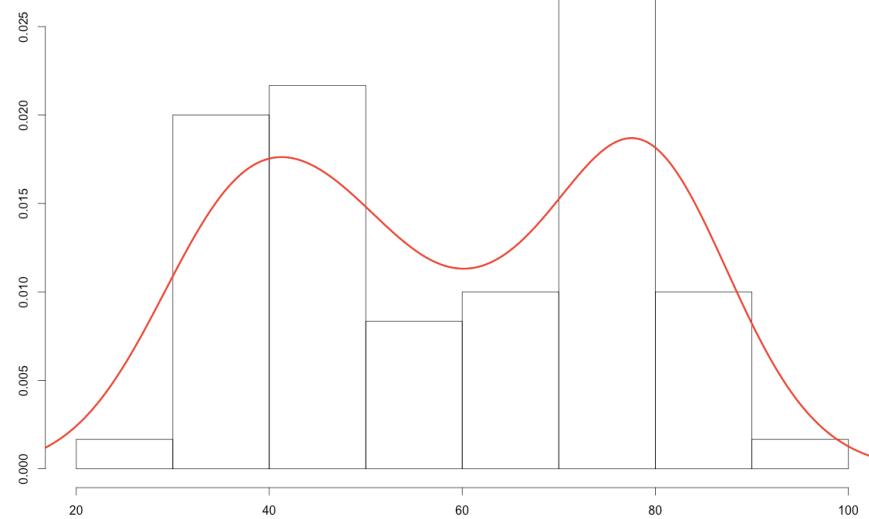
*supervised*

***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*

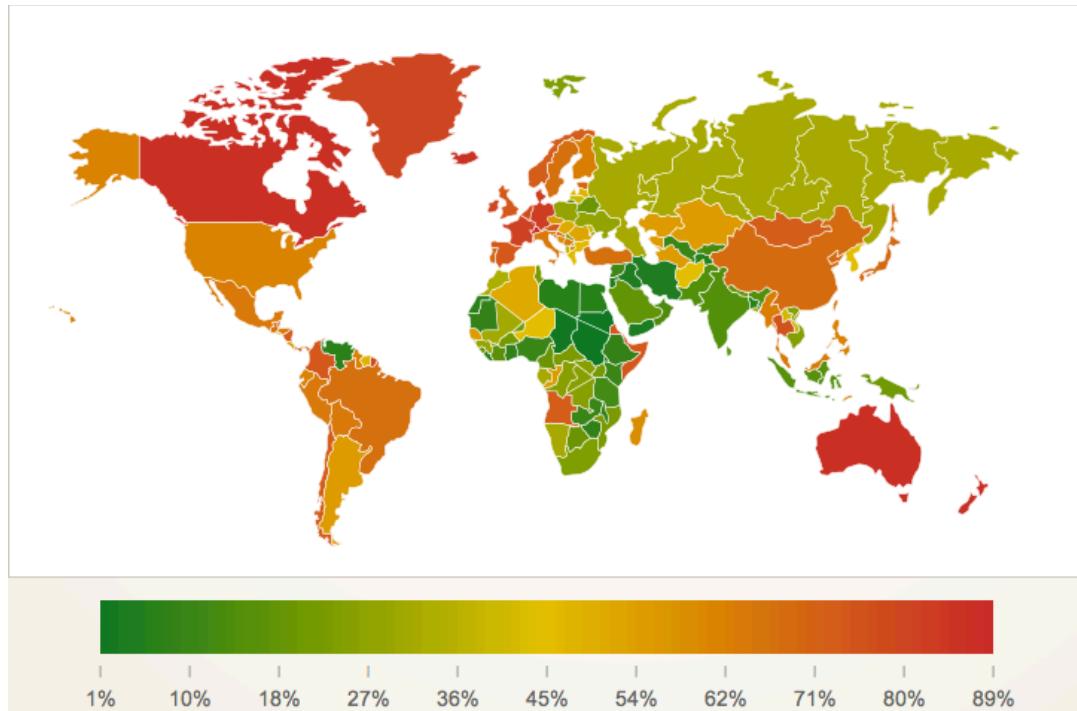
*dimensional  
reduction*

*categorical*

*classification*

*clustering*

# REGRESSION EXAMPLE: PREDICTING IPHONE SALES



*GDP*

*population*

*Gini*

*phone penetration %*

*GDP growth rate*

# TYPES OF ML PROBLEMS

*supervised*

*unsupervised*

*continuous*

*regression*

*dimensional  
reduction*

*categorical*

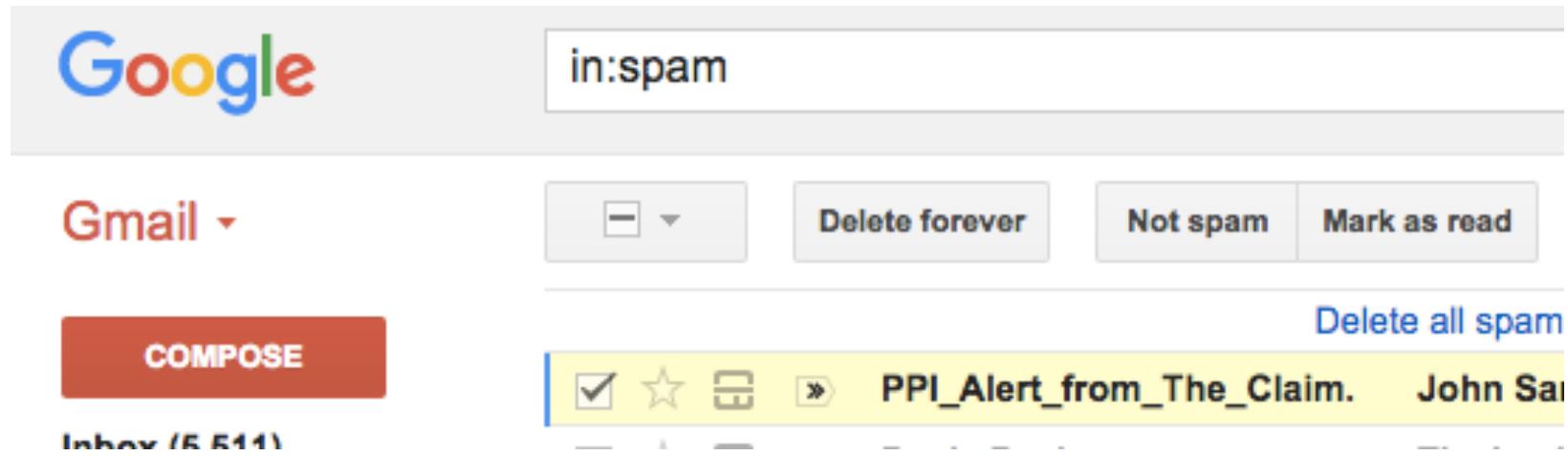
*classification*

*clustering*

# 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*

*!!!*

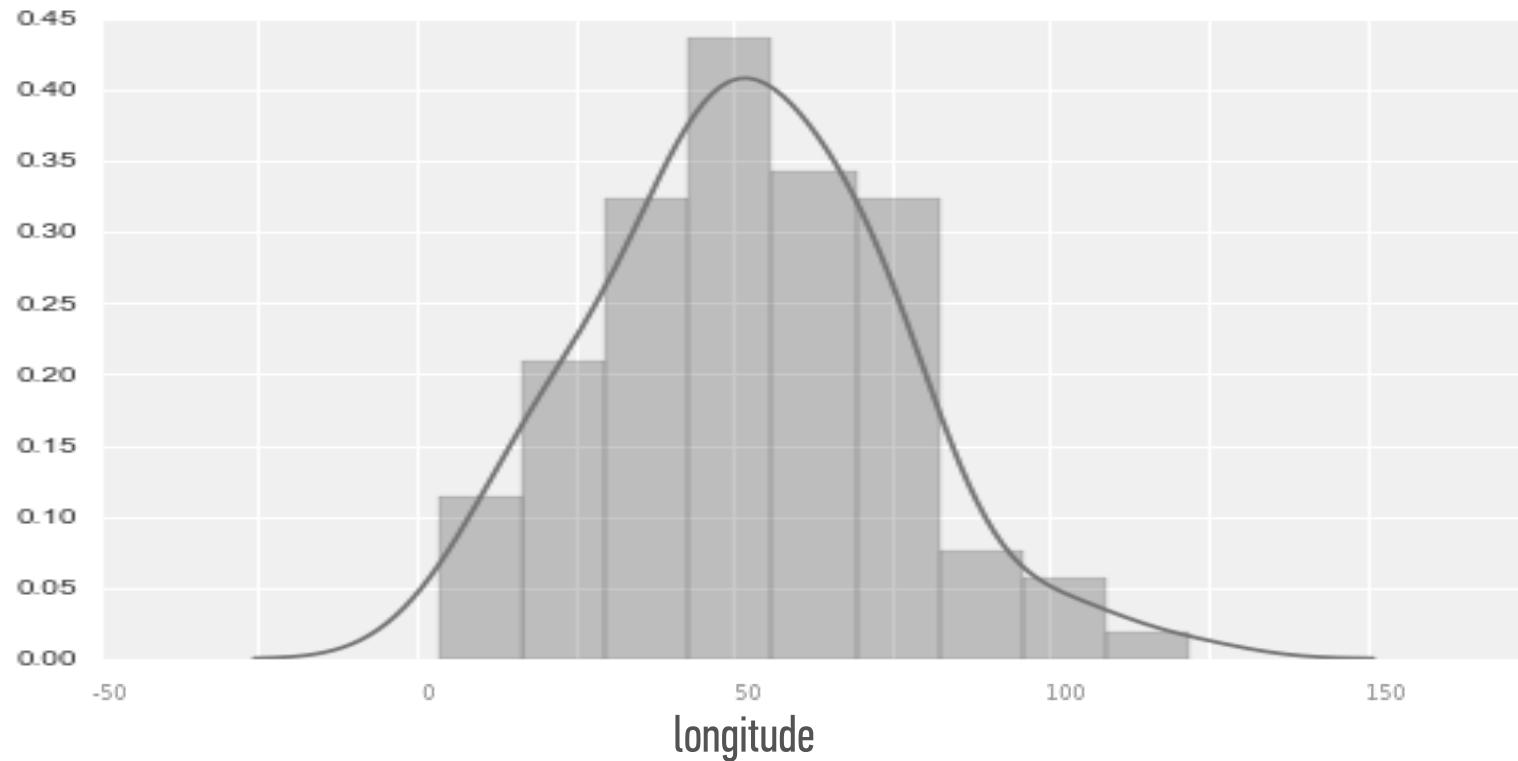
# TYPES OF ML PROBLEMS

	<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>

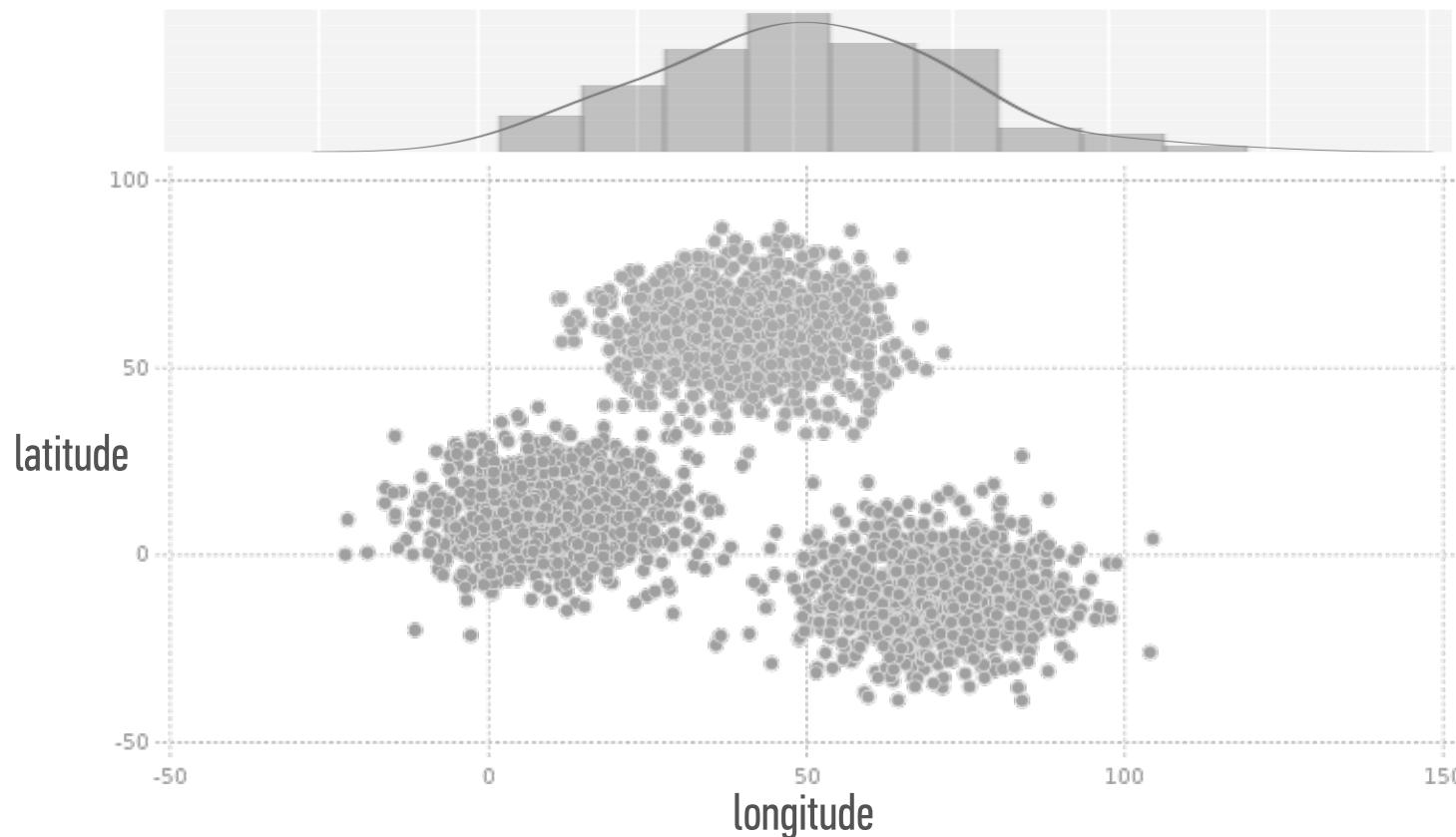
# TYPES OF ML PROBLEMS

	<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>

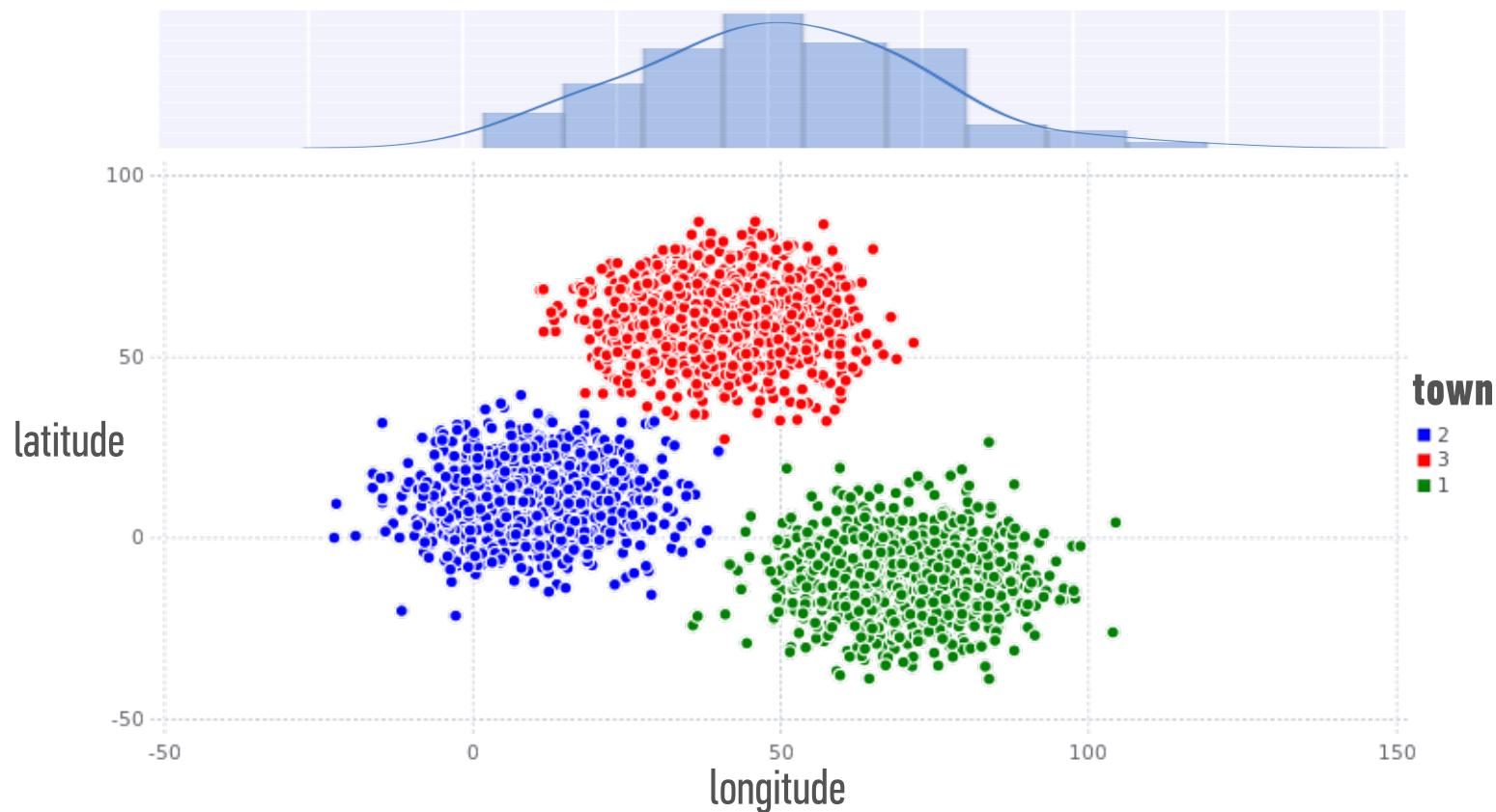
# CLUSTERING EXAMPLE: USER LOCATIONS



# 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>

# 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



# 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>

# PART IV.

# 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

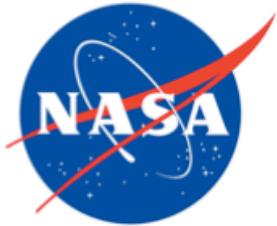


Gartner

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

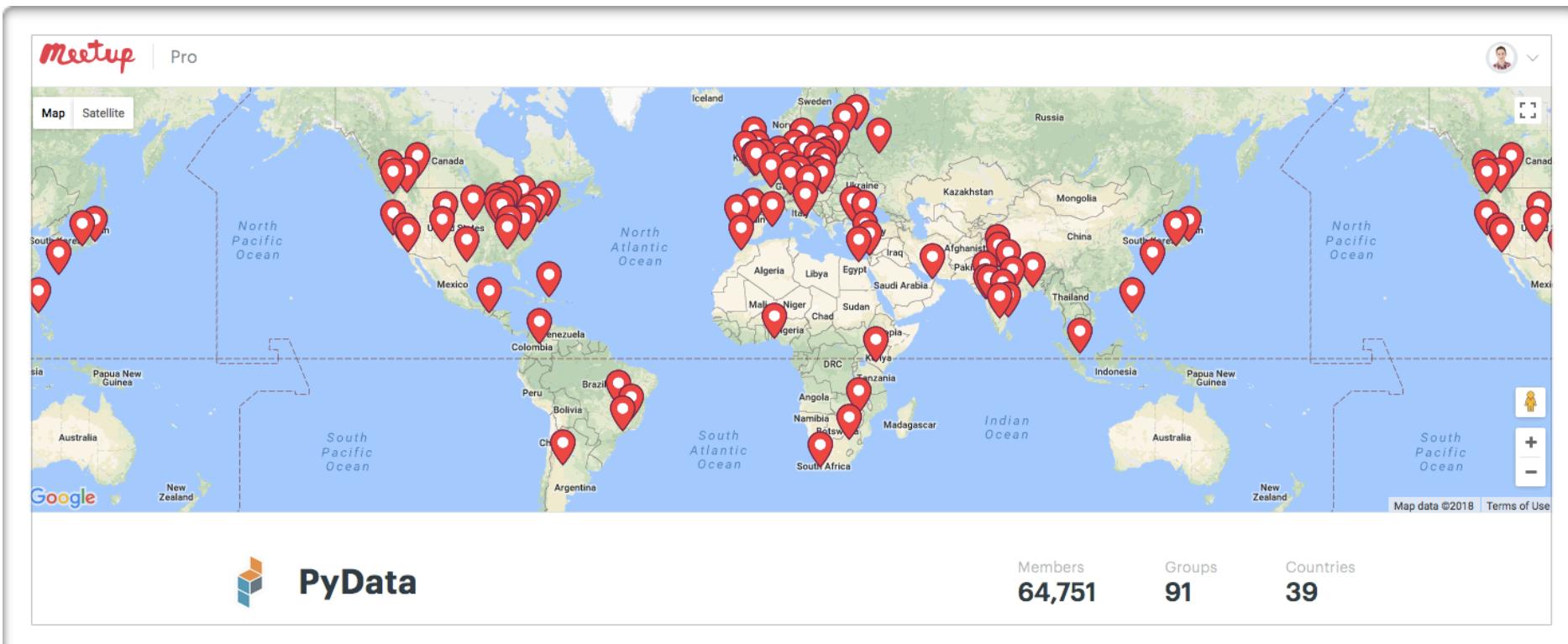


**PYDATA WARSAW**  
NOVEMBER 19-20, 2018



**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

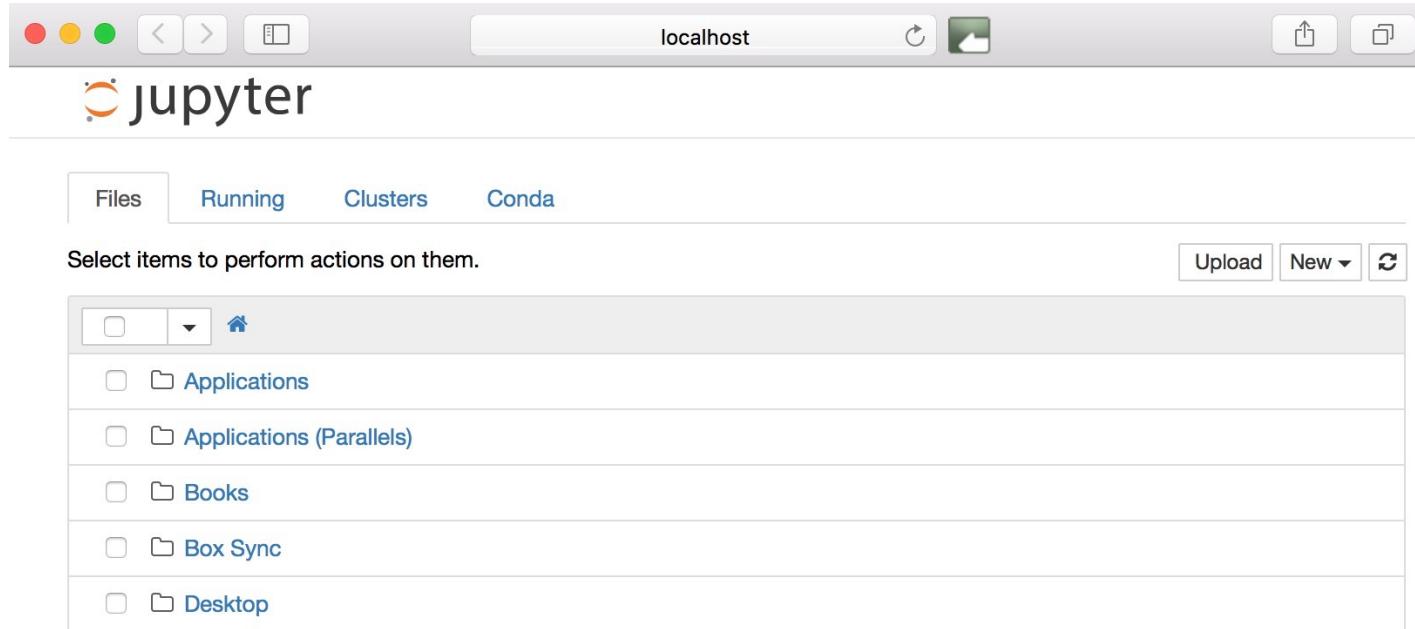
- ▶ **pandas**: manipulate data
- ▶ **SciPy/NumPy**: scientific computing and numerical calculations
- ▶ **Scikit-learn**: machine learning
- ▶ **matplotlib/Seaborn**: data visualisation
- ▶ **spacy/nltk**: natural language processing
- ▶ **statsmodels**: statistical tests
- ▶ **Beautiful Soup**: HTML/XML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



# PART V.

# TIPS FOR SUCCESS

# NEW JOB

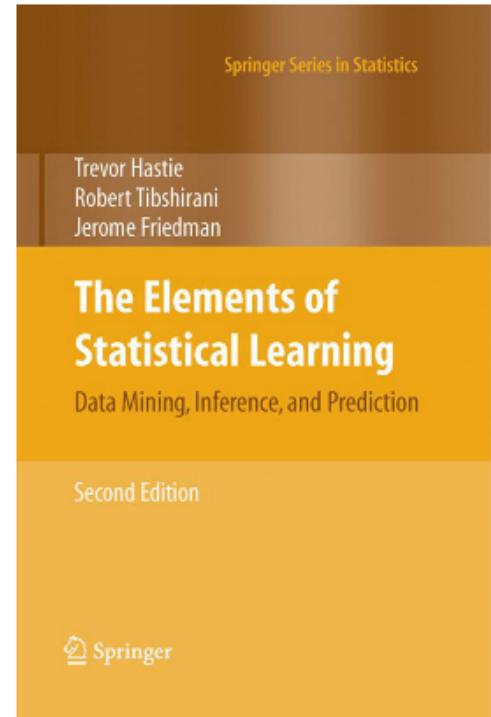
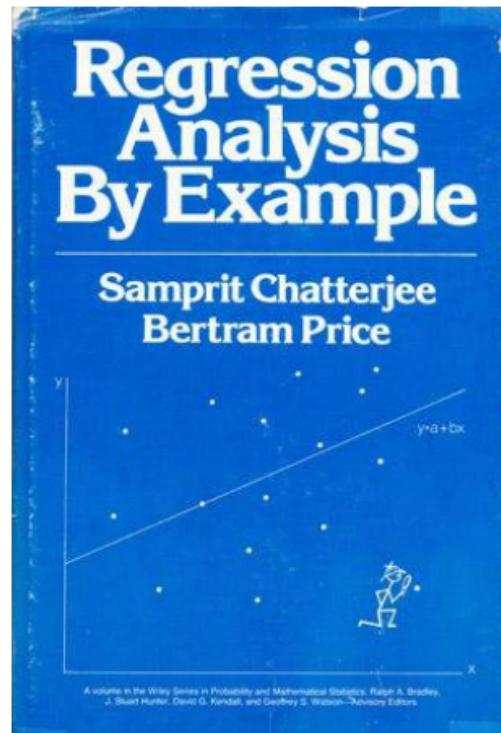
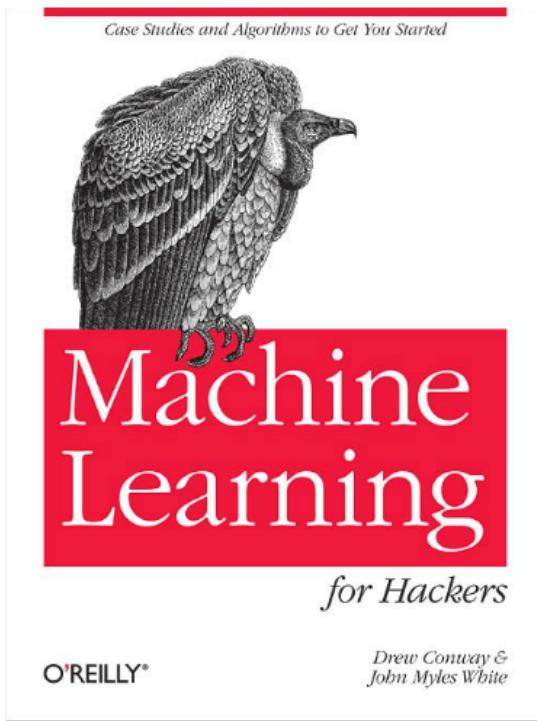
- In April 2012 McKinsey predicted 1.5 million shortage of data scientists
- More and more companies are looking for people to unlock the value in their data
- Rise in available positions

Location	London	3 months to 16 Aug 2013	Same period 2012	Same period 2011
<strong>Data Scientist</strong>				
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	<strong>£55,000</strong>	£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%	

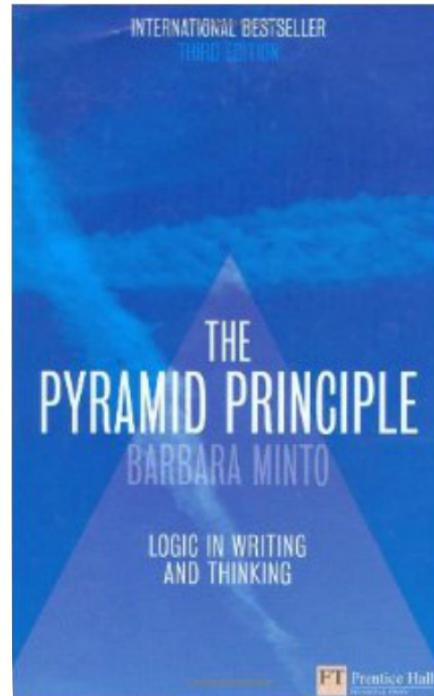
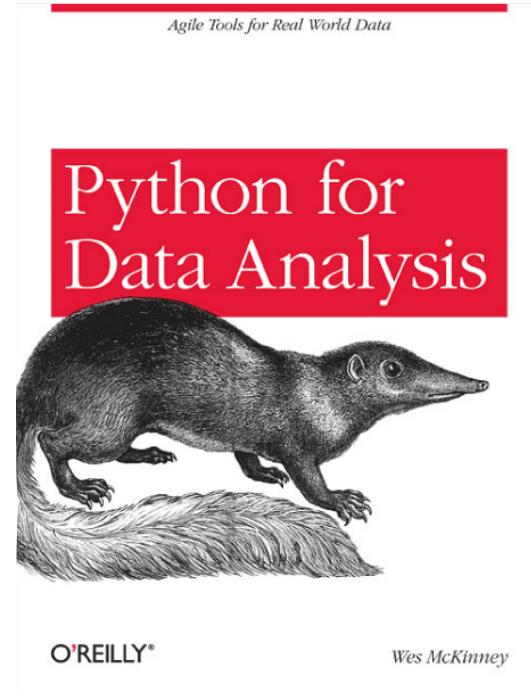
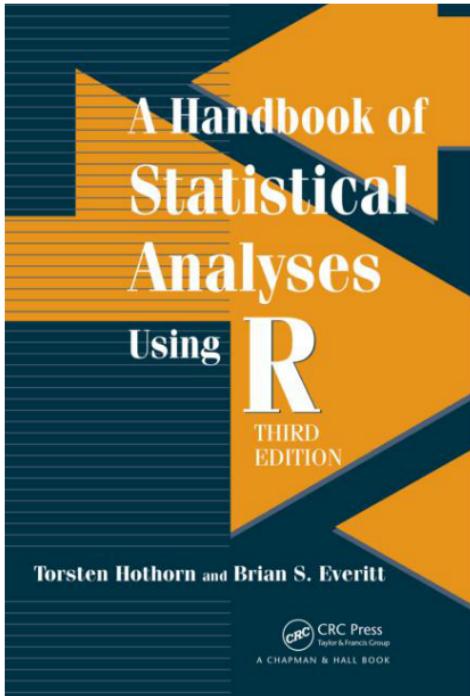
# **SHORTAGE OF SKILLS**

- 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)

# 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

## 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

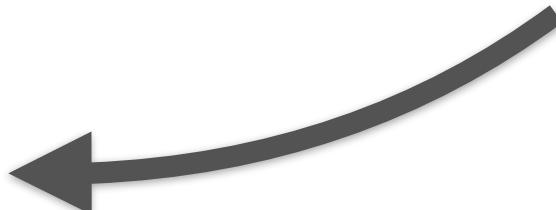
# HACKATHONS & DATADIVES

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

# HACKATHONS & DATADIVES

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

"We liberate data, and make useful things"



# FINAL THOUGHTS

# FOUR STEPS TO SUCCESS

## 1. Learn to code

Python. R. Professional software engineering practices.

## 2. Get statistical

Significance. Inference. Regression. Machine learning.

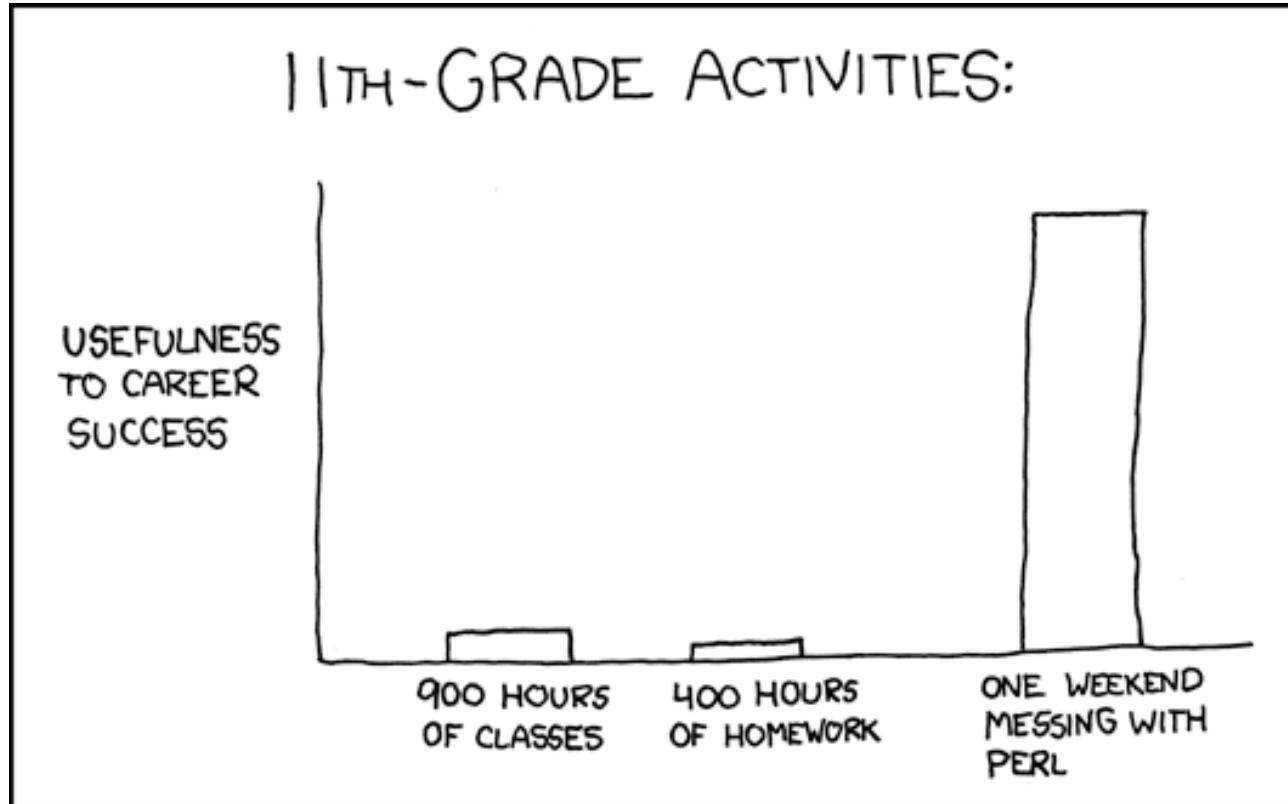
## 3. Learn lean

Business skills. Startup methodology. Communication.

## 4. Experience

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

# IF YOU DO NOTHING ELSE...



# IF YOU DO NOTHING ELSE.....GET STARTED TONIGHT!

## › Data Skeptic Podcast

MARCH 10, 2017

### [MINI] The Perceptron

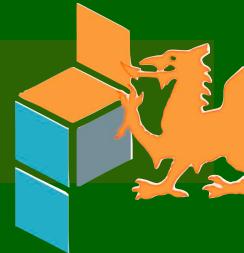
▶ PLAY 14:46    [!\[\]\(f5c60e9ef44a35d4e503f9cee26bcbe0\_img.jpg\) Download](#)

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](http://dataskeptic.com)**

# **FINAL THOUGHTS**

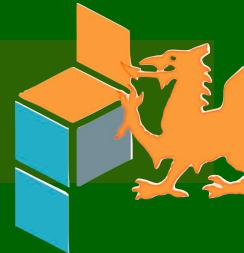
- Data science is a product of our time**
- Being a data scientists requires people and technical skills**
- We're only getting started...**



# THANK YOU

@PyDataCardiff

@john\_sandall



# QUESTIONS?

@PyDataCardiff

@john\_sandall