GENERAL ASSEMBLY

INTRO TO DATA SCIENCE & ANALYTICS

Dr J Rogel-Salazar

j.rogel.datascience@gmail.com
@quantum_tunnel / @ dt_science

WELCOME!

LOGISTICS 3

Contact:

Email: j.rogel.datascience@gmail.com

Twitter: @quantum_tunnel / @dt_science

AGENDA 4

I. WHAT IS DATA SCIENCE?
II. DOING DATA SCIENCE
III. WHAT DOES IT TAKE TO BE A (SUCCESSFUL) DATA SCIENTIST
IV. TOOLS AND GEAR
V. CHALLENGES AND OPPORTUNITIES

DATA...

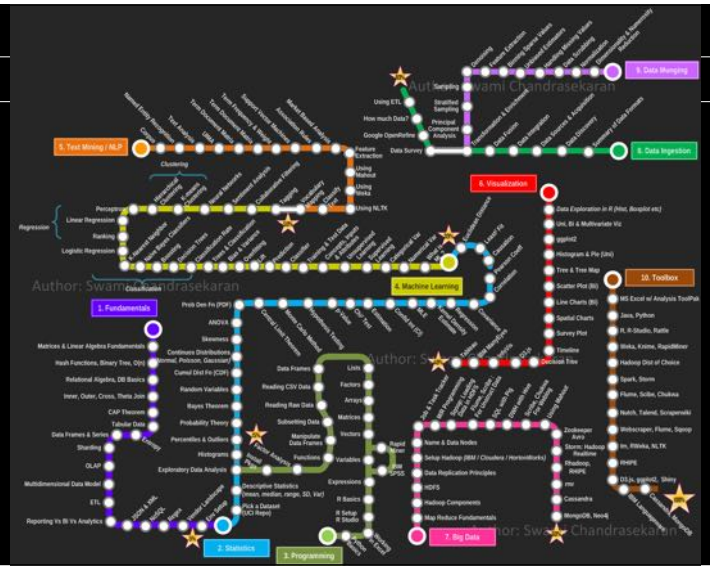
ABOUT YOU

- Who are you?
- Why are you interested in Data Science?
- Any experience with Python (or R)?
- What is your exposure to Data Science so far?
- Do you have any applications in mind?

I. WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

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

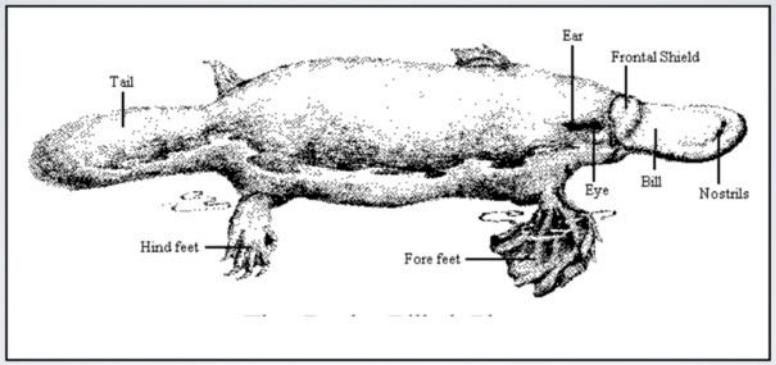


WHAT IS DATA SCIENCE?

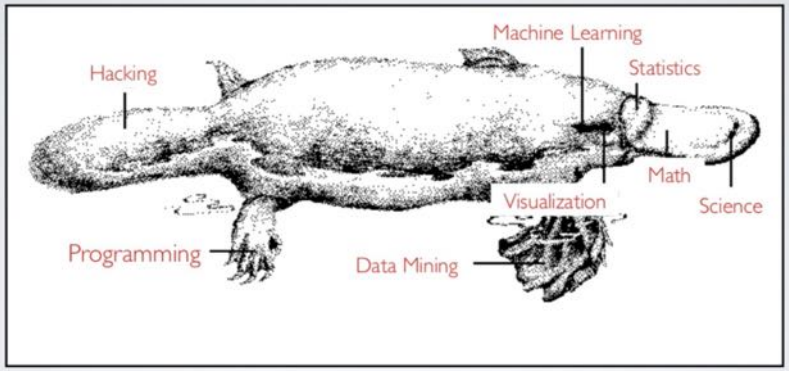
A set of tools and techniques used to extract useful information from data.

An interdisciplinary, problem-oriented subject.

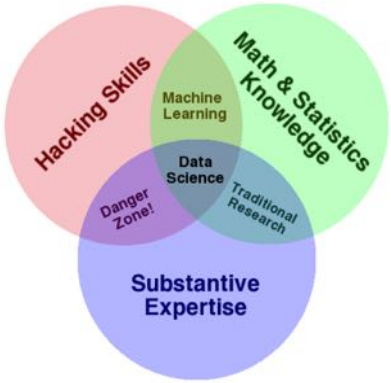
The Duck – Billed Platypus



The Platypus – Billed Data Scientist

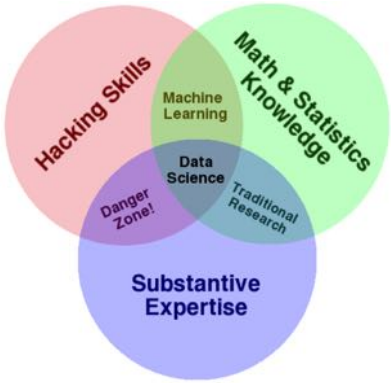


THE QUALITIES OF A DATA SCIENTIST 13



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

THE QUALITIES OF A DATA SCIENTIST 14



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

ONE MORE THING!

Communication skills

15

Class *DataScientist* {

Is skeptical, curious. Has inquisitive mind
Knows Machine Learning, Statistics, Probability
Applies Scientific Method. Runs Experiments
Is good at Coding & Hacking
Able to deal with IT Data Engineering
Knows how to build data products
Able to find answers to *known unknowns*
Tells relevant business stories from data
Has Domain Knowledge

}

WHAT IS DATA SCIENCE? 16

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

An interdisciplinary, problem-solving oriented subject.

The application of scientific techniques to practical problems.

WHAT IS DATA SCIENCE? 17

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.

WHO USES DATA SCIENCE? 18

NETFLIX amazon facebook Google LinkedIn FiveThirtyEight foursquare Etsy 2012 BARACKOBAMA.COM Microsoft

DATA EVERYWHERE 19

- By 2010, according to Eric Schmidt, every two days we created as much information as we did from the dawn of civilisation up to 2003
- Creation of data outstrips current capabilities to store it

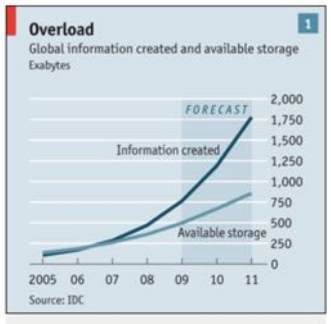


Image Source: http://www.economist.com/node/15557443

DATA EVERYWHERE 20

- 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

WHAT ABOUT COMPUTING? 21

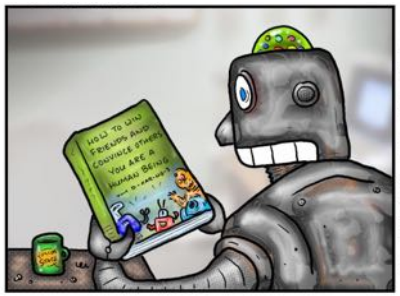
- Computing hardware has come down in cost
- Cloud computing tech has made it possible for everyone to use IT infrastructure
- Moore's law holds and computers are more powerful

AND WHAT OF THAT "BIG DATA" THING? 22

- Generic term to describe data that:
  - Doesn't fit in memory
  - Doesn't fit on a machine

NOW WHAT? 23

- The increase in computing power has made older techniques impractical
- New advances in stats and machine learning make it possible to analyse large amounts of data



Unit Bob crams for his Turing Test.

OPEN SOURCE TOOLS 24

- Not only are there drops in hardware costs, but also…
- Availability of open source tools: Linux, Apache, Hadoop, MySQL
- Apart from infrastructure there are also analysis tools
- Heard of R or Python?

ALL THIS DATA! WHERE IS IT COMING FROM? 25

- Digital
- Sensors
- Transactions
- Documents
- Open data

WHY SHOULD WE CARE ABOUT DATA 26

- Innovation

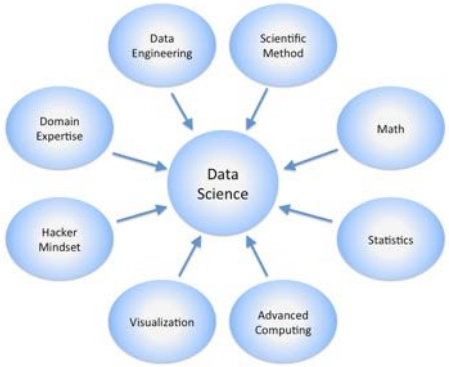
  - Unmet needs
  - Niche segments

- Optimisation

  - e.g meetings demand when/ where required

SO... WHAT IS DATA SCIENCE? 27

- Given that data is already here, data science aims to:
  - Make sense of it
  - Use appropriate tools
  - Transform data into information
  - Information into knowledge
  - Knowledge into action



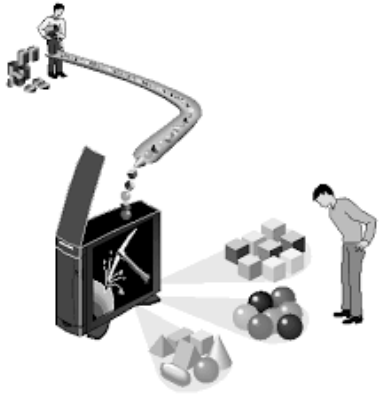
II. DOING DATA SCIENCE

DOING DATA SCIENCE? 29

- The availability of data and tools makes data science possible
- Data scientists try to find new patterns
- Make predictions and classify

FINDING AND CLASSIFYING OBSERVATIONS 30

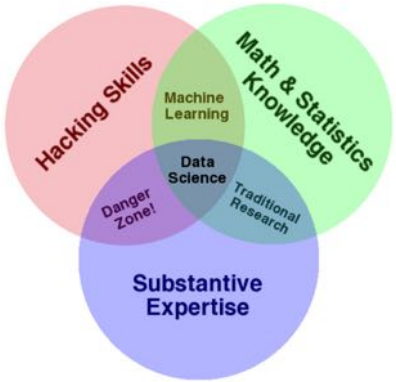
- When exploring data, it is important to find out what patterns are there
- Techniques to find patterns are called **unsupervised**
- To determine which category an observation belongs to, we use a **classifier** - a **supervised** method



PREDICTION 31

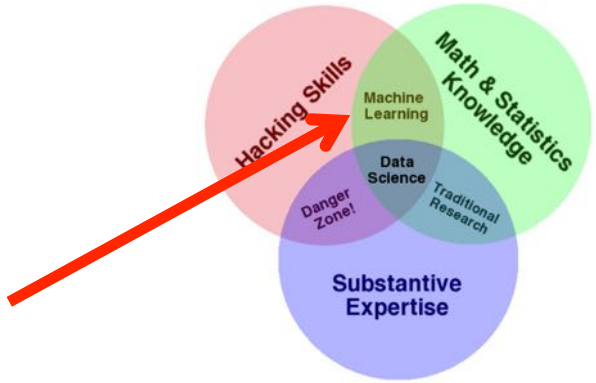
- Another important task is prediction
- Given a set of observations, what will happen next?
- Techniques that need the data to be used are supervised, and will produce continuous predictions

REMEMBER THIS? 32



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

YOU ARE HERE 33



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

TYPES OF LEARNING PROBLEMS 34

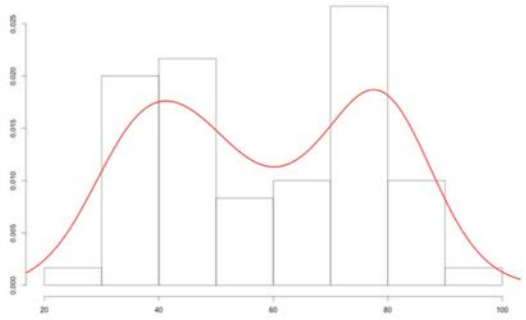
supervised	making predictions
unsupervised	extracting structure

TYPES OF LEARNING PROBLEMS 35

Unsupervised

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

Extracting Structure

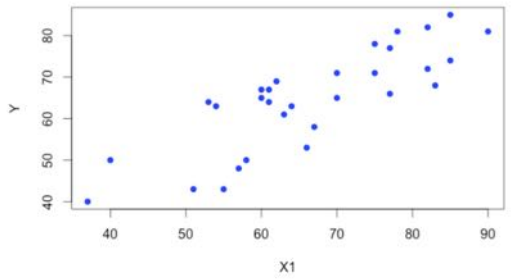


TYPES OF LEARNING PROBLEMS 36

Supervised

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

Making Predictions

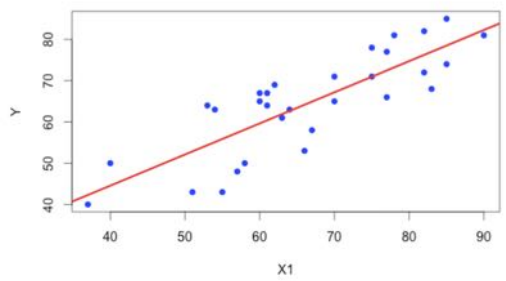


TYPES OF LEARNING PROBLEMS 37

Supervised

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

Making Predictions



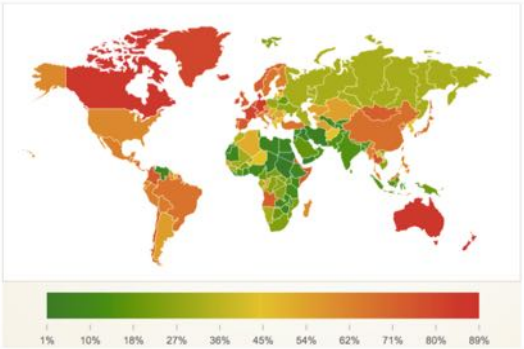
TYPES OF DATA 38

	continuous	categorical
quantitative		
qualitative		

TYPES OF ML SOLUTIONS 39

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

REGRESSION EXAMPLE: PREDICTING PHONE SALES 40



GDP

Population

Gini

Demographics

Phone penetration %

Growth rate

TYPES OF ML SOLUTIONS 41

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

DIMENSIONALITY REDUCTION EXAMPLE: STOCK INDEX 42

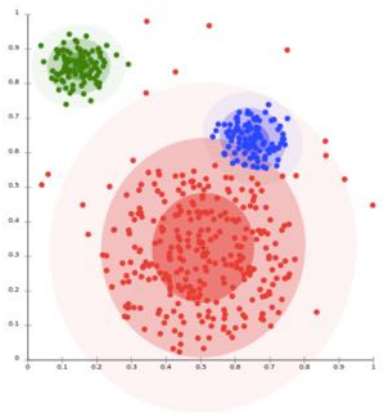
DIMENSIONALITY REDUCTION EXAMPLE: STOCK INDEX 43

TYPES OF ML SOLUTIONS 44

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

CLUSTERING EXAMPLE: USER LOCATIONS 45

Coordinates
(continuous data) →



TYPES OF ML SOLUTIONS 46

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

CLASSIFICATION EXAMPLE: SPAM FILTERING 47

$$$ Bargain

100% FREEE

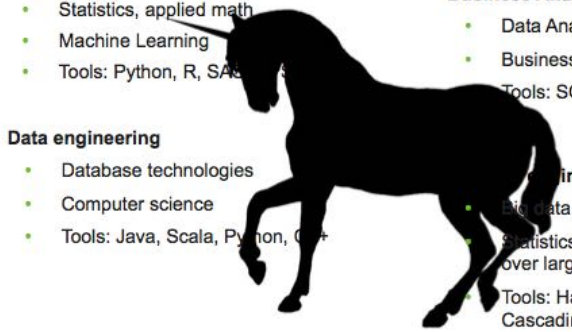
ACT NOW!!!!

££££

Satisfaction “Guaranteed”

III. WHAT DOES IT TAKE TO BE A (SUCCESSFUL) DATA SCIENTIST

Which makes them hard to find…



Applied Science
- Statistics, applied math
- Machine Learning
- Tools: Python, R, SA

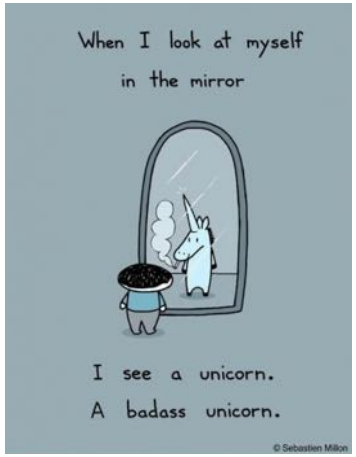
Business Analysis
- Data Analysis, BI
- Business/domain expertise
- Tools: SQL, Excel, EDW

Data engineering
- Database technologies
- Computer science
- Tools: Java, Scala, Python, C++

ineering
- ata pipeline engineering
- tistics and machine learning
  ver large datasets
- Tools: Hadoop, PIG, HIVE, Cascading, SOLR, etc

Hortonworks

WHAT DOES A DATA SCIENTIST DO? 50



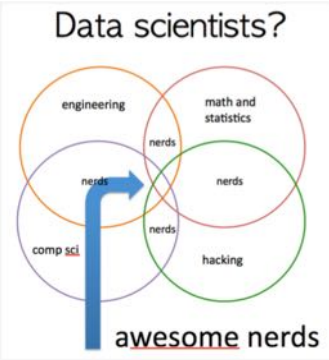
DATA.. SCIENTISTS… 51

- **There are many definitions….**
- "Data science is the combination of analytics and the development of new algorithms… 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."

Hillary Mason

COMBINATION OF SKILLS 52

- What does "hacking" mean in this context?
- Is there anything else missing?



PERSONALITY 53

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

CURIOSITY 54

- Patterns don't just present themselves
- An outlier could start an interesting line of enquiry
- Staying up-to-date with developments in the field

TECHNICAL ABILITY 55

- Solving business problems using data requires
  - Knowledge of technology stack
  - Programming knowledge
  - Understanding how systems are implemented
  - Maths/Stats

RIGOUR 56

- 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 something
- It is important to tell signal from noise

COMMUNICATION 57

- 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 organisations
- This is not technical and requires persuasion skills

BUSINESS ACUMEN 58

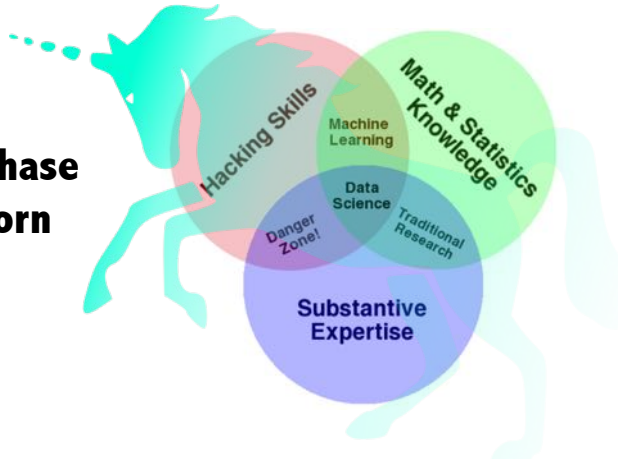
- Data science is about finding new things
- Of all the things we can do, which one is the most important?
- There might be something unexpected in the data, but does it matter?
- The best solution to a problem may not be practical

DATA SCIENCE IS A TEAM SPORT 59

- Very few people are strong in all areas
- Each discipline is vast enough to require a life time too become a master
- Data scientist often work ion teams that include: data engineering, reporting, operations, etc.
- Any other roles?

HOW TO BUILD A DATA SCIENCE AND ANALYTICS TEAM 60

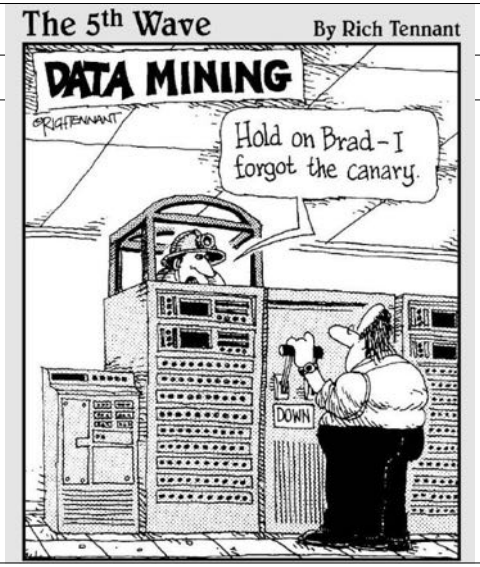
You may want to chase the infamous Unicorn



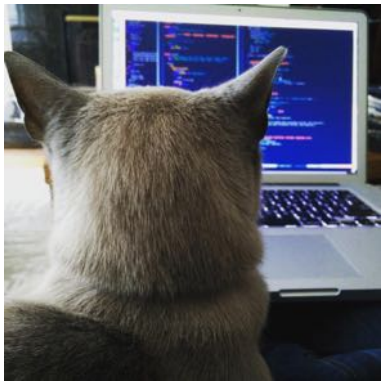
Or… you can be more realistic and become a Jackalope data scientist

This can help you build an effective team

IV. TOOLS AND GEAR



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



R

65

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

PYTHON

66

- A very useful general-purpose language
- Rapidly growing since 2000s
- Libraries to access DBs, APIs, machine learning, plotting, network analysis, NLP, web dev, etc
- For reference: it is interpreted and dynamically typed

DATABASES - SQL AND NOSQL

67

- Relational databases
- SQL - Structured query database
- The main DB tech for many a year
- What has changed?

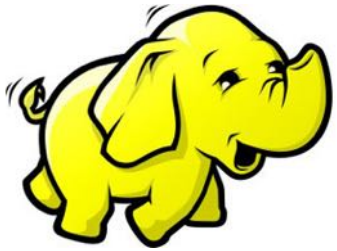
NOSQL

68

- Emergence of web scale data
- Distributed, large scale, non structured data

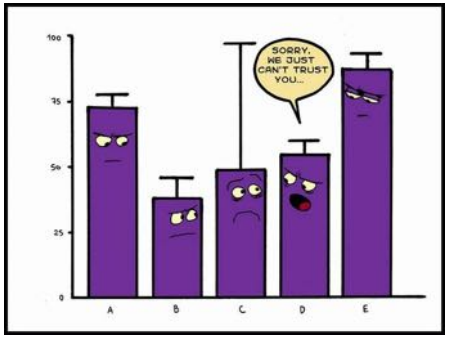
HADOOP 69

- Comes from 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 and manipulated via MapReduce



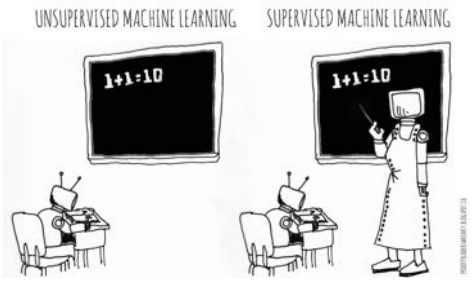
STATISTICS 70

- A science of data
- Predates computers
- Emphasises formal statistical inference (in low dimensionality)
  - Confidence intervals
  - Hypothesis tests



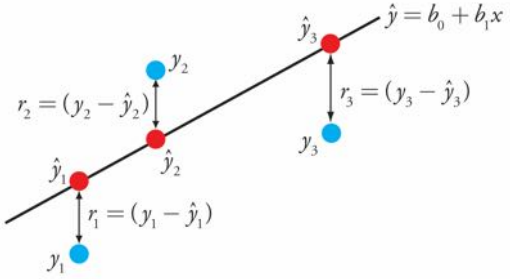
MACHINE LEARNING 71

- Close to stats
- A computer science discipline
- There is an algorithmic component
- Many concepts are similar to stats, but with a different name



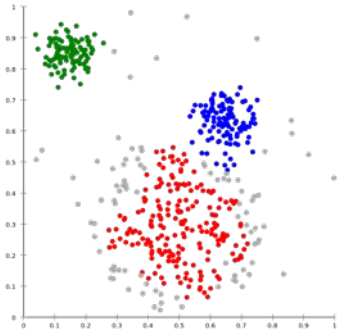
STATISTICS - REGRESSION 72

- The "workhorse of data science"
- Allows us to characterise relationships between variables
- Can be used to build a model to predict values of y given observations of x



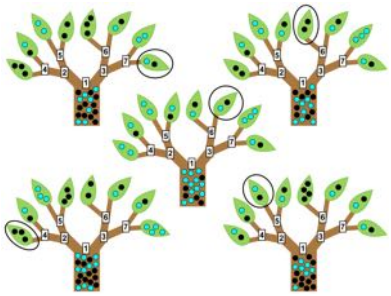
MACHINE LEARNING - CLUSTERING 73

- Are there natural groups in the data?
- Applications include segmentation
- Unsupervised learning technique



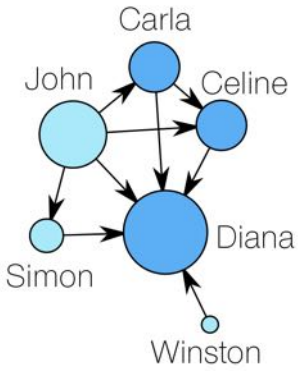
MACHINE LEARNING - RANDOM FORESTS 74

- Builds on decision trees
- What if instead of one tree we had many (ensemble learner)?
- Tells us what category something belongs to - Classification



SOCIAL NETWORK ANALYSIS 75

- Data from Twitter, calls among mobile network users
- What is the shortest distance between to users?
- Are there communities?
- How does product adoption spread among friends?



NATURAL LANGUAGE PROCESSING 76

- Extracting information from text
- What emotion is the customer expressing in a message?
- What are the topics mention on certain webpages?
- Which section of a text is of a certain type?

DISTRIBUTED COMPUTING 77

- Machine learning as a service
  - Prediction APIs: wise.io, Google Predictions
  - Entity extraction Open Calais
- Cloud Computing
  - Send code and data to tun on 100s of machines
  - Spin up powerful servers to run your computation

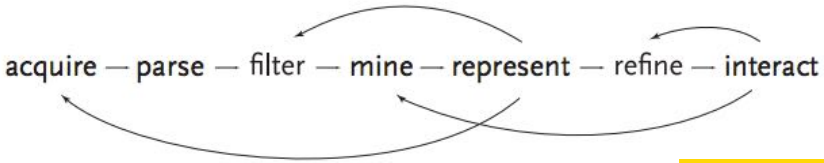
DATA VISUALISATION 78

- Exploratory data analysis
- Communicating findings

THE DATA SCIENCE WORKFLOW 79

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

THE DATA SCIENCE WORKFLOW 80



NOTE

This diagram illustrates the iterative nature of problem solving

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

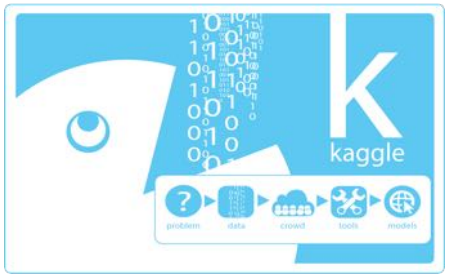
V. CHALLENGES AND OPPORTUNITIES

YOUR JOB

- The organisation you work for probably generates more data than you imagine
- Is the value form each data source exhausted?
- Is data from multiple sources? Combined?
- Are users aware of possibilities offered by SNA, NLP, others?

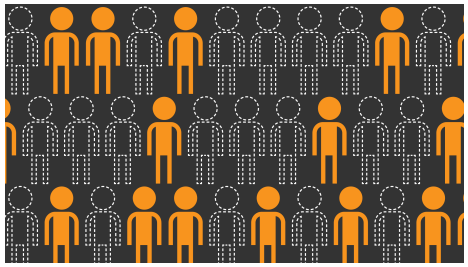
KAGGLE

- Crowd-sourcing analytics problems
- Thousands compete by using whatever methods to produce best prediction
- Cash prizes



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



YOUR STARTUP 85

- Software development becomes commoditised
- Many not very technical ideas only need a WordPress install
- Many new companies differentiate themselves through their use of data

YOUR HEALTHCARE IS CHANGED BY DATA 86

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

HOW INDUSTRIES ARE CHANGED BY DATA 87

- Marketers have used segmentation and churn for years
- However, with the move to digital, marketing is becoming more analytical

HOW EDUCATION ARE CHANGED BY DATA 88

- Government compile stats about schools
- Online education allows to track each student's progress and tailor the material
- Online teaching materials supplement normal classes and generate data

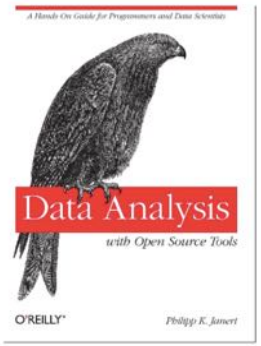
SHORTAGE OF SKILLS 89

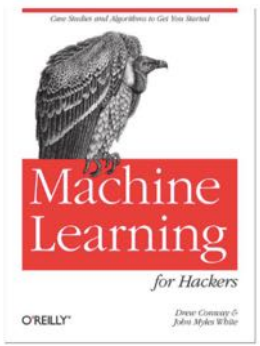
- Many companies struggle to recruit in this area
- Traditional analysts too focused on specific tools
- Many programmers don't have the business experience
- Because the field is new there are few people with leadership skills

THE DATA EXPLOSION CONTINUES 90

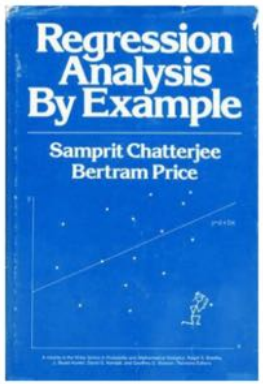
- Mobile devices generate data which is already being collected
- Internet of things - all devices will become computerised, constantly connected and generating data
- Quantified self - Tracking almost every aspect of someone's data requires skills

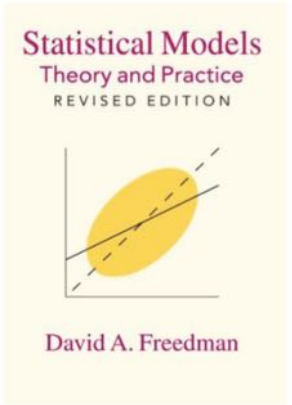
BOOKS - INTRO 91

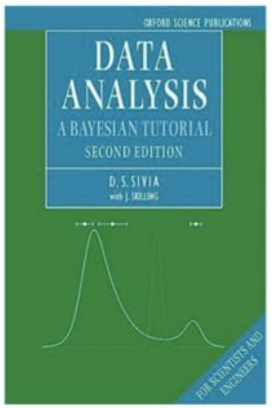




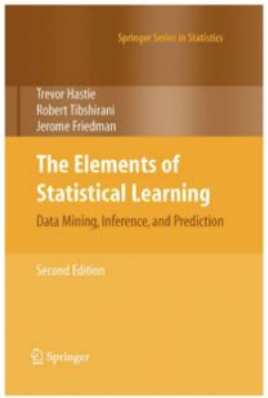
BOOKS - PARAMETRIC METHODS 92

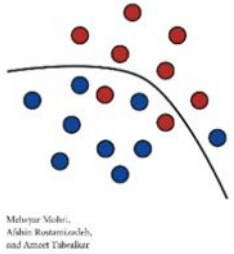


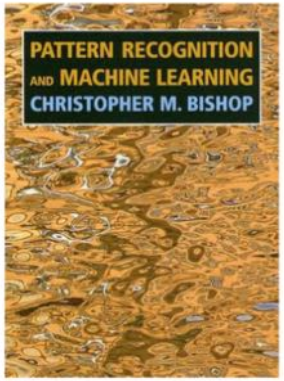




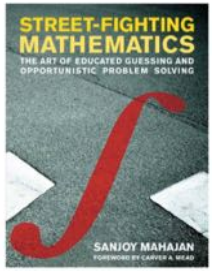
BOOKS - MACHINE LEARNING THEORY 93

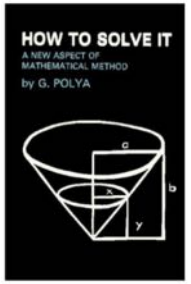


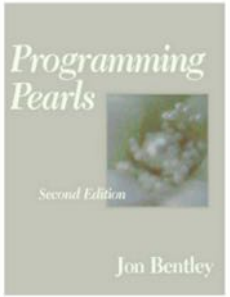




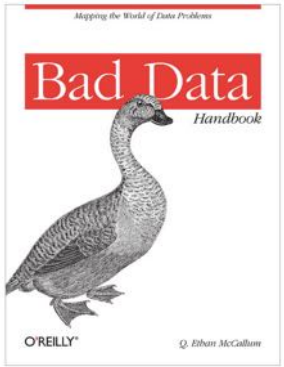
BOOKS - PROBLEM SOLVING 94

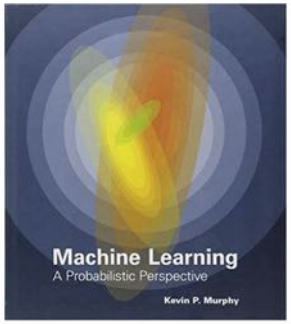


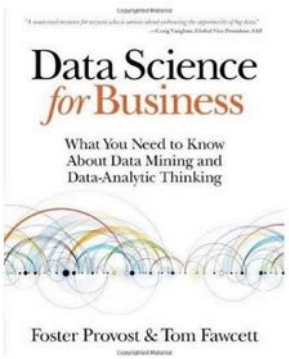




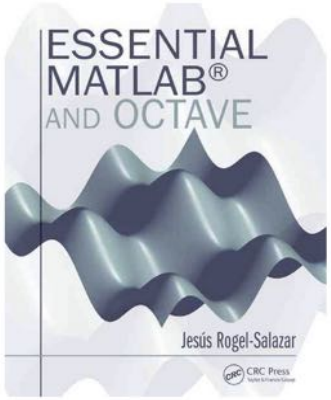
BOOKS - J'S TOP 3 RECOMMENDATIONS 95

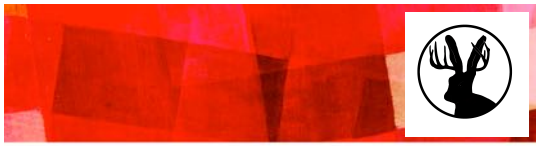






BOOKS - IF YOU ARE FEELING GENEROUS 96





DATA SCIENCE AND ANALYTICS WITH PYTHON

Dr J Rogel-Salazar

DATA SCIENCE BLOGS 97

- news.ycombinator.com
- r-bloggers.com
- hilarymason.com
- blog.echen.me
- hunch.net
- fivethirthyeight.blogs.nytimes.com
- blog.yhat.com
- wesmckinney.com

ONLINE LEARNING 98

- Stanford ML course: coursera.org/course/ML
- hilarymason.com/tag/video
- harvarddatascience.com
- Andrew Moore: autonlab.org
- CalTech ML: youtube.com/playlist?list=PLD63A284B7615313A
- videolectures.net/Top/Computer_Science/Machine_Learning

PODCASTS 99

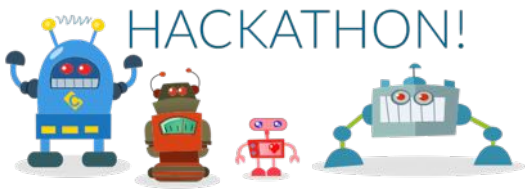
- Data Skeptic
- Partially Derivative
- Linear Digressions
- More of Less
- O'Reilly Data Show
- Podcatst.__init__
- Talk Python to Me
- Data Stories

LONDON MEETUPS 100

- PyData London
- LondonR
- Data Science Meetup London
- Big Data London
- London Machine Learning Meetup
- Visual+Data
- Quantified Self
- PyLadies London
- Women in Data
- Big O Meetup

HACKATHONS AND DATIVES 101

- DataKind
- NHS Hack
- Gaggle
- Hackathons and Jams UK
- StartupWeekend
- Code for Good



HOW TO BECOME A DATA SCIENTIST (OVERSIMPLIFIED) 102

- Learn to code
  Python, R, Spark, etc
- Get statistical
  Significance, inference, regression, ML
- Learn to learn
  Business skills, startup methodology, communication
- Experience
  Side projects, GitHub, Kaggle, Hackathons

CONCLUSION 103

- Data Science is a product of out time

- Being a data scientist requires people and technical skills

- We are only getting started

Q&A

CONTACT

Dr J Rogel-Salazar
j.rogel.datascience@gmail.com
@quantum_tunnel / @dt_science

GENERAL ASSEMBLY WOULD LOVE TO HEAR YOUR HONEST FEEDBACK

You shall receive an email requesting feedback on today's session shortly. We encourage you to complete this as it will allow us to improve the quality and value we provide.

Thank You!!