Introduction to Data Science and Analytics

Gianluca Campanella

Today’s seminar

What is Data Science?

Who is a Data Scientist?

What’s it like to be a Data Scientist?

Planning your Data Science career

What is Data Science?

What is Data Science?

Mathematics and
Statistics / Operational Research

Computing and
Software Engineering

Visualisation and
Communication Skills

Domain expertise

What is Data Science?

A problem-solving approach
based on the scientific method

What does Data Science deal with?

Problems!

Can we improve…

- The quality of offers we send to our customers?
- Road safety?
- How we identify people at high risk of cancer?

What does Data Science deal with?

Predictions?

How likely…

- Is a customer to respond to some offer?
- Are traffic accidents to occur in a certain area?
- Is a person to develop cancer in the next 10 years?

What does Data Science deal with?

Mechanisms?

Why…

- Does a customer decide to respond to some offer?
- Do traffic accidents occur regularly in certain areas?
- Do people develop cancer?

What is Data Science?

Statistics

- Predates computers
- People can understand why something happens in the face of uncertainty

Machine Learning

- ‘Algorithmic modelling’ (L. Breiman)
- Computers can learn rules without explicit programming

Deep Learning

- Less structured inputs
- Computers can learn structure without explicit programming

What is Data Science?

	Predictions	Mechanisms
Analysis	**Descriptive** What’s happening?	**Diagnostic** Why is it happening?
Building	**Predictive** What’s likely to happen?	**Prescriptive** What do I need to do?

Recap

Data Science

- Evidence-based problem solving and decision-making
- Multidisciplinary but domain-driven
- Analysis-focused or building-focused

Who is a Data Scientist?

Who is a Data Scientist?

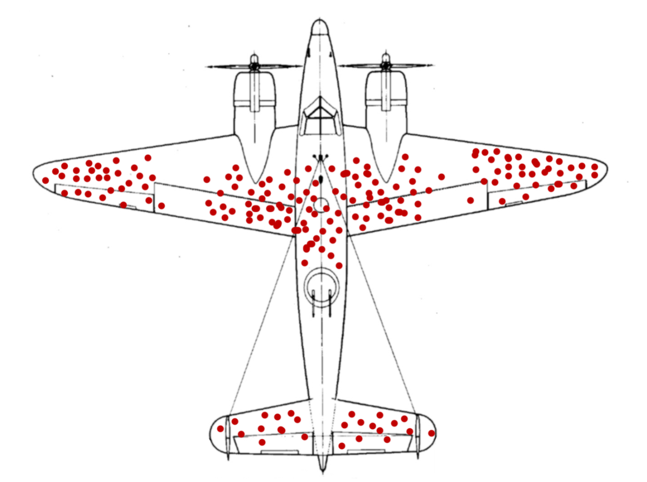
Someone who can…

- Get a ‘feel’ for the data
- Communicate effectively
- Work well in a team

What’s this ‘feel’ for the data?

- Passion for the domain
- Curiosity about the data
- Intuition and creativity
- Common sense
- Rigour and accuracy
- Relevance

What’s this ‘feel’ for the data?

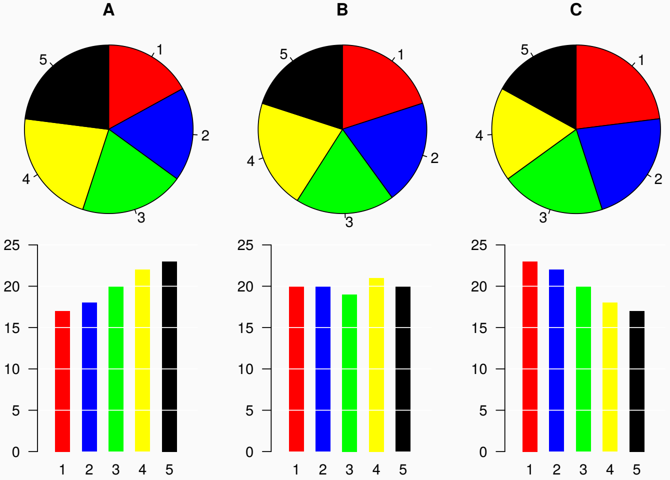


Via *Wikimedia Commons*

How do I communicate effectively?

- Condense findings into recommendations
- Describe assumptions and limitations
- Use storytelling techniques and visual aids
- Interpret sceptically
- Understand limitations and don’t overstate results

How do I communicate effectively?



Via *Wikimedia Commons*

The ‘PR problem’ of Data Science

Inevitably the data are…

- Not quite what you need to solve your problem
- Too limited, too large, too inaccurate, too expensive to obtain…

But (eventually) you…

- End up with a ‘nice’ dataset
- Apply some models

…and it looks incredibly easy from the outside!

What’s it like to be a Data Scientist?

Data Science workflow

1. Define the problem
2. Obtain the data
3. Clean and explore the data
4. Model the data
5. Summarise the results

Time allocation

Which takes longer?

Time allocation

In decreasing order…

1. Defining the problem
2. Obtaining the data
3. Cleaning and exploring the data
4. Managing expectations
5. Summarising the results
6. Learning new things
7. Modelling

Modelling misconceptions

Most well-executed data science projects don’t…

- Use complicated tools
- Fit complicated models

Instead, they do…

- Focus on solving the problem
- Consider whether the data are appropriate — not necessarily big!
- Use relatively standard models
- Interpret results sceptically

The 80—20 rule of modelling

- The first reasonable thing you can do is 80% of the way to the solution
- Everything after that is to get the remaining 20%… often at additional cost!

The 80—20 rule of modelling

- The first reasonable thing you can do is 80% of the way to the solution
- Everything after that is to get the remaining 20%… often at additional cost!

Is it worth it?

Caveat

The Data Science workflow is

non-linear and iterative

Recap

A successful Data Scientist…

- Is insatiably curious… and a bit stubborn!
- Never stops learning
- Is a practical, impact-driven, dependable person who can tell a story
- Can prioritise and manage time effectively
- Knows the limitations of Data Science and how to manage expectations

Planning your Data Science career

Why should you consider it?

Many companies struggle to recruit in this area

- Traditional analysts often focused on specific tools
- Many programmers don’t have business experience
- Few people with leadership skills

The possibilities are endless… and growing!

- Many new companies are built on data
- Most industries are becoming increasingly analytical
- Data as ~~an asset~~ the lifeblood of the organization

What do you need to succeed?

1. Be passionate about your domain
2. Know about research methods and statistics
3. Become a coding ninja

What do you need to succeed?

Be passionate about your domain

- Understand where the data come from
- Know your stakeholders and speak their language
- Communicate to others why a certain question is worth answering

What do you need to succeed?

Know about research methods and statistics

- Remember the 80—20 rule of modelling
- Interpret sceptically
- Understand limitations and don’t overstate results

What do you need to succeed?

Become a coding ninja

- Standardise and automate data collection and analysis
- Standardise and automate everything!
- Share your analyses with others

Taking the first step

You can begin today!

Figure out what you need to learn

- Code comfortably in a programming language (typically Python or R)
- Work with data in that language
- Understand basic statistical concepts

Taking the first step

For example…

1. Get comfortable with Python
   - Complete ‘Intro to Python for Data Science’ on DataCamp
   - Attend ‘Python Programming 101’ at General Assembly
2. Learn data manipulation, analysis, and visualisation with pandas
   - Go through *Python for Data Analysis* by Wes McKinney
   - Try using pandas instead of Excel
3. Brush up on statistics
   - Go through *Think Stats* by Allen B. Downey

… … …

One final word of advice

Find ‘the thing’ that motivates you to practise and learn more…

Then do it!