Introduction to the Data Science Toolbox

Daniel Lawson University of Bristol

Lecture 01.1 (v1.0.2)

Why Data Science?

- For the first time in history, data is abundant and everywhere
- This is **found data**, that is, it is not gathered for the purpose to which you will put it
- There are four classes of tools that contribute:
  - **Classical statistics**: designed for small, carefully curated data
  - **Machine Learning**: designed for efficient prediction
  - **Algorithms**: the study of what tasks can be efficiently implemented
  - **Infrastructure**: choices of how to structure data and compute resource
- None of these fields alone is enough
- **Data Science** is combining these to solve real-world questions from biased, messy data

What is a Data Scientist?

> Think of him or her as a hybrid of data hacker, analyst, communicator, and trusted adviser. The combination is extremely powerful - and rare.

- Harvard Business Review, via fossbytes[1].

[1] https://fossbytes.com/data-scientist-money-you-can-earn-21st-century/

Course Structure

- **fortnightly** lectures (1hr)
- **weekly double** (2hr) workshop sessions:
  - odd week: **2hrs mixed** lecture/workshop
- **2 semester** run time!
- **every 4 weeks**: assessed mini-project due.

Expectations

- This unit consists of:
  - 12 hours of lectures
  - 12 hours of presentation-led workshops (more lectures)
  - 12 hours of supported workshops
  - **164 hours** of independent learning (including python workshops)
- You may have entered the course on marginal pre-requisites. It is **your responsibility** to catch up any missing knowledge:
  - (Blocks 1-5) Intermediate R, e.g.: http://www.datasciencemadesimple.com/r-tutorial/
  - (Blocks 6-11) Intermediate Python, e.g.: http://chryswoods.com/intermediate_python/index.html
- You are also expected to find **code, methods, documentation and explanations** for yourselves!

Assessment of coursework

- Students will differ in background knowledge of statistics, computer science, and programming. **All three skills are required** to produce high quality coursework.
- However:
  - There is much flexibility in the details of the content that you can choose to present
  - You can emphasise core mathematics, exploitation of library routines, expert knowledge of the data, and brute programming to different degrees
  - **Diligence and brilliance** in any category will be rewarded. You are encouraged to design your coursework content to emphasise your strengths.

Pre-requisites

- All University of Bristol
  - Probability 1
  - Statistics 1
  - Statistics 2 (or equivalent)
  - Some programming knowledge

Intended Learning Outcomes

- ILO1 Be able to **access and process cyber security data** into a format suitable for mathematical reasoning
- ILO2 Be able to **use and apply basic machine learning** tools
- ILO3 Be able to **make and report appropriate inferences** from the results of applying basic tools to data
- ILO4 Be able to use **high throughput computing infrastructure** and understand appropriate **algorithms**
- ILO5 Be able to reason about and **conceptually align problems** involving real data to appropriate theoretical methods and available methodology to correctly make inferences and decisions
- ILO6 Be able to **work as part of a team** to apply mathematical methods to difficult data science problems

Working Individually, Together

- Work together as a **team** to **learn**:
  - You **all** need to get **very good, very fast**, at a diverse set of skills.
  - Use your colleagues to catch up on missing pre-requisites
  - Work together to solve problems, particularly data processing problems common to all students
  - Work together to understand the theory and material
- Work individually to **demonstrate expertise**:
  - Individual Assessments should be written, in entirety, by **you alone**
  - You can receive **acknowledged** help
  - You are allowed to use solutions found elsewhere, as long as you provide evidence that you understand what the code is doing. This **evidence should be unique** to you.

Connection to other courses

- Wider Courses:
  - We'll review content from the pre-requisites
  - We'll use ideas from advanced Statistics, Bayesian Methods, and Probability
  - The CS course on **Machine Learning** covers much more theory and diverse practice
- Connection to other courses in the Mathematics of Cyber Security Unit:
  - Use this course to **better understand** concepts in Cyber Security, Graph Theory, Anomaly detection, etc
  - Bring those concepts into your mini-projects
  - Explore them in detail, with data & methods from this course

Adaptable thinking

- Most courses choose a single notation and stick to it.
- Data Science is a mess in part because **disciplines are not able to talk to each other**. They use different notation and translation is hard.
- To read about how a method developed by a statistician works, you will need to understand how they write. You will need a different statistical "language" to understand a Machine Learning method. And different again to understand a Algorithms method.
- This course will play fast and loose with notation and language style to **normalise a single concept having multiple, analogous, definitions**. It is suggested that, where notations differ confusingly, you keep a crib sheet. Creating this is very valuable learning.

Data Sources

The course follows cyber-security data from:

- The U.S. National CyberWatch **Mid-Atlantic Collegiate Cyber Defense Competition** (MACCDC 2012). Generated by Johns Hopkins University.
- "Each team is given physically identical computer configurations... the teams have to ensure the systems supply the specified services while under attack from a volunteer **Red Team**... the teams have to satisfy periodic 'injects' that simulate business activities IT staff must deal with in the real world."
- Contains scanning/recon through exploitation as well as some c99 shell traffic. Roughly 22M total connections.
- http://www.netresec.com/?page=MACCDC
- Additional info at www.secrepo.com

Some categories of data

- Numerical, categorical, or binary
- Text: emails, tweets, articles
- Records: user-level data, timestamped event data, log files
- Geo-based location data
- Network data
- Time-series sensor data
- Images and video
- Audio and music

Some numbers

- **48** - The hours of video uploaded to YouTube every minute, resulting in nearly 8 years of content every day.
- **7 Million** - The numbers of DVDs internet traffic information would fill EVERY hour (2017). Side by side, they scale Mount Everest 95 times.
- **3 Billion** - The number of people who were online in 2015, generating 8 zettabytes of data. (One zettabyte equals one sextillion bytes - twenty-one zeros...)
- **30 Billion** - Pieces of content shared on Facebook every day (2017).
- **247 Billion** - The number of e-mail messages sent each day (2017) - up to 80% are spam.
- **90%** - Percentage of the world's data created in the last 2 years.

More numbers

- Library of Congress text database of around **20 TB**.
- Thirteen million photographs, even if compressed to a 1 MB JPG each, would be **13 TB**.
- AT&T **323 TB**, 1.9 trillion phone call records.
- 3.5 million sound recordings, which at one audio CD each, would be almost **2,000 TB**.
- World of Warcraft utilizes **1.3 PB** of storage to maintain its game.
- Avatar movie reported to have taken over **1 PB** of local storage at Weta Digital for the rendering of the 3D CGI effects.
- Google processes **24 PB** of data per day.
- YouTube: More video is uploaded in 60 days than all 3 major US networks created in 60 years. According to cisco, internet video will generate over **18 EB**.

What is big data?

- Large text dataset:
  - 1,000,000 words in 1967
  - 1,000,000,000,000 words in 2006

	Big Data	Small Data
Data Condition	Always unstructured, not ready for analysis, many relational database tables that need merged	Ready for analysis, flat file, no need for merging tables.
Location	Cloud, Offshore, SQL Server, etc.	Database, local PC
Data Size	Over 50K Variables, over 50K individuals, random samples, unstructured	File that is in a spreadsheet, that can be viewed on a few sheets of paper
Data Purpose	No intended purpose	Intended purpose for Data Collection

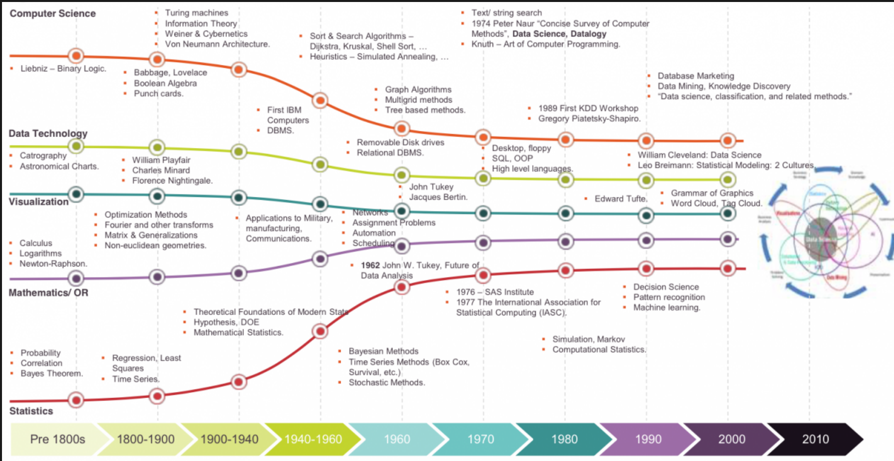
What is data science?

- "Data science, also known as data-driven science, is an **interdisciplinary field about scientific processes** and systems to extract knowledge or insights from data in various forms." (Wikipedia)
- "Data science is an advanced discipline, requiring **proficiency in** parallel processing, map-reduce computing, petabyte-sized noSQL databases, machine learning, advanced statistics and complexity science." (Data Science: An Introduction)
- "Data science is the study of **where information comes from, what it represents and how it can be turned** into a valuable resource in the creation of business and IT strategies." (TechTarget)
- "Data Science: An action plan to **expand the field of statistics** ." (William Cleveland, 2001)

What is data science?

- "Data science, as it's practiced, is a blend of **Red-Bull-fuelled hacking and espresso-inspired statistics**. [...] Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible." (Mike Driscoll)
- "Data science is an **act of interpretation**." (Riley Newman)
- "There is **no such thing as data science**." (Robin Bloor)

History of Data Science



Signposting

Next is **01.2 Exploratory Data Analysis**:

- Types of data
- How to read in data
- How to plot it
- Interpreting what data is, before we use a model