Introduction to the Data Science Toolbox

https://dsbristol.github.io/dst/

Daniel Lawson University of Bristol

Lecture 01.1 (v2.0.2)

https://xkcd.com/1838/

Some Data Science Questions. . .

Some Data Science Questions. . .

- What is the point being a mathematical data science?
- How much data in ‘enough’?
- Can there be ‘too much’ data?
- How many CPU hours did training ChatGPT use?
- How big is the Internet?
- How will large models grow into the future?
- Are most data science problems like this?

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
- We might classify tools into four classes:
  - **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

- Every week has a "block of content".
- Everything is at *http://dsbristol.github.io/dst/*
  - two lectures (1hr)
  - one workshop (1hr)
  - **formative worksheet**
  - **summative worksheet** to be entered into Portfolio
- **Week 3, 7, 11**: group project due.
- **Week 12+**: Portfolio due

Expectations

- This unit consists of:
  - around 24 hours of lectures
  - 12 hours of supported workshops
  - around **150 hours** of independent learning
- 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 and Statistics 1
  - Some programming knowledge
- Desirable:
  - Statistics 2 (or equivalent)

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:
  - Use this course to **better understand** concepts in 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 valuable learning.

Some numbers

- In 2023, Wikipedia has 60M pages taking up **22.14 GB** without media, 428.36TB of media on Wikimedia Commons.
- The (text only) internet is about 400TB.
- Ten billion photographs on Flickr. Compressed to 1 MB JPG each, this would be **10 PB**.
- Spotify has 80 million songs. At 3Mb each this would be almost **24 PB**.
- 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**.
- According to OpenAI, the training process of **Chat GPT-3 required 3.2 million USD in computing resources alone**. This cost was incurred from running the model on **285,000 processor cores and 10,000 graphics cards**, equivalent to about **800 petaflops of processing power**.

What is big data?

	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

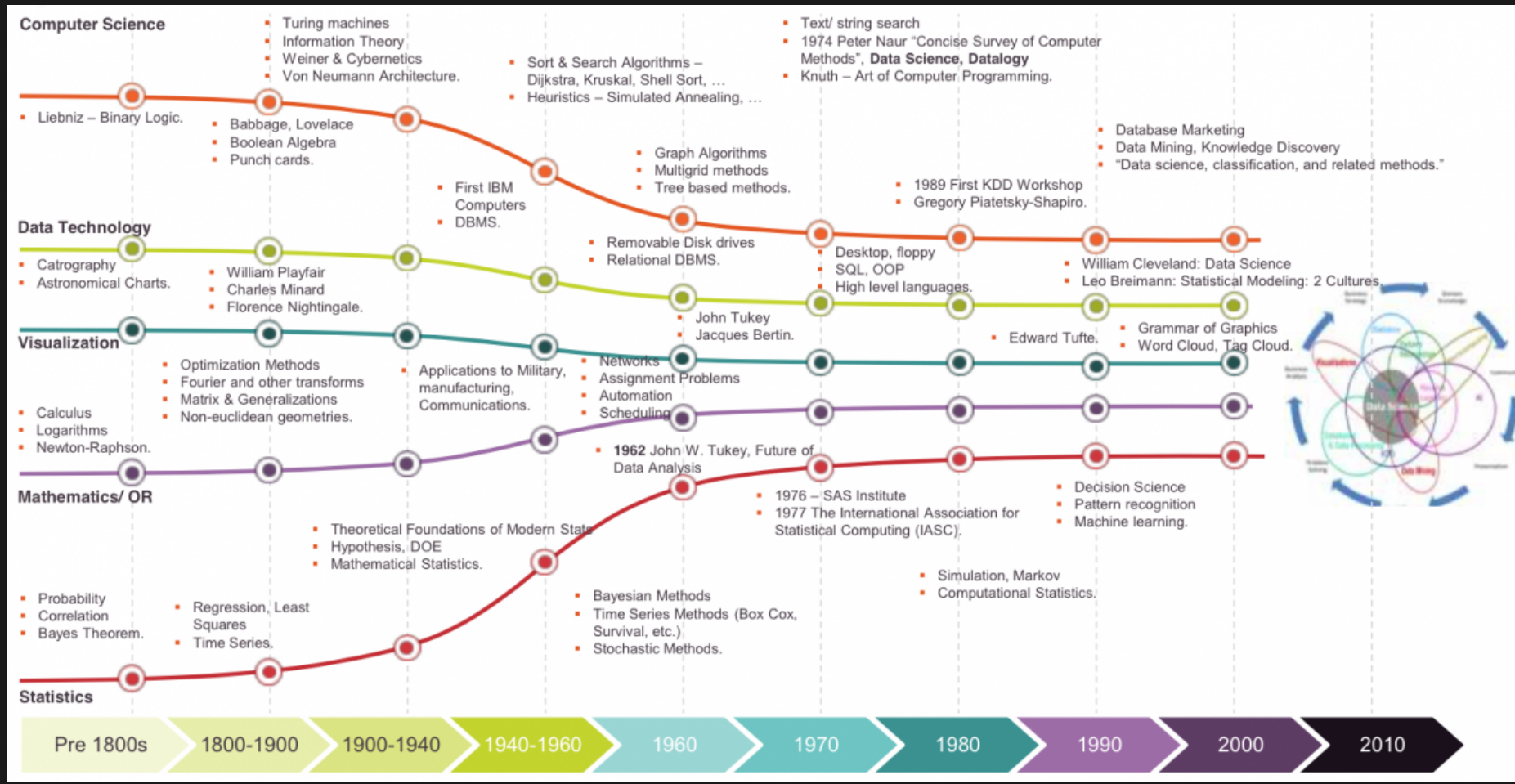
What is data science?

- "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