Week 1 - Overview

- Data science
    - History
    - Definition

Narrow: machine learning on big data
Broad: extraction of knowledge/value from data through the complete data lifecycle process

    - Machine Learning

Machine Learning is concerned with the development of algorithms and techniques that allow computers to learn.

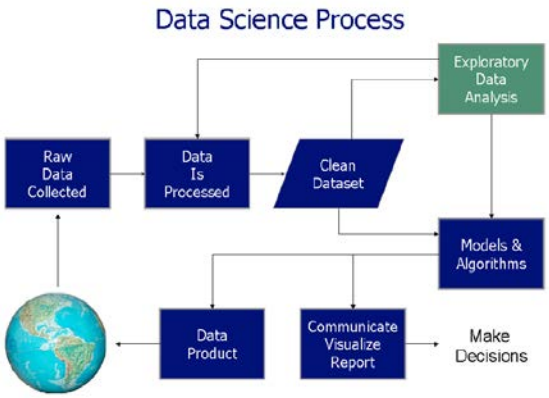
- Data scientist
    - Skills

Business: product development, business
Machine learning/Big data: unstructured data, structured data, machine learning, big and distributed data
Mathematics/Operations research: optimisation, mathematics, graphical models, Bayesian and Monte Carlo statistics, algorithms, simulation
Programming: systems administration, back end programming, front end programming
Statistics: visualisation, temporal statistics, surveys and marketing, spatial statistics, science, data manipulation

    - Roles
    - Job descriptions & requirements

Data developers: were relatively strong in all areas except Statistics
Data researchers: had little Programming
Data businesspersons: had little Mathematics/Operations research and even less Programming
Data creatives: were relatively strong in all areas except Mathematics/Operations research

- Data science process & value chain



Collection: getting the data
Engineering: storage and computational resources across the full lifecycle
Governance: overall management of data across the full lifecycle
Wrangling: data preprocessing, cleaning
Analysis: discovery (learning, visualisation, etc.)
Presentation: arguing the case that the results are significant and useful
Operationalisation: putting the results to work, so as to gain benefits or value

- What is data science?

addresses the data science process to extract meaning/value from data

Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others

- What is a data scientist?

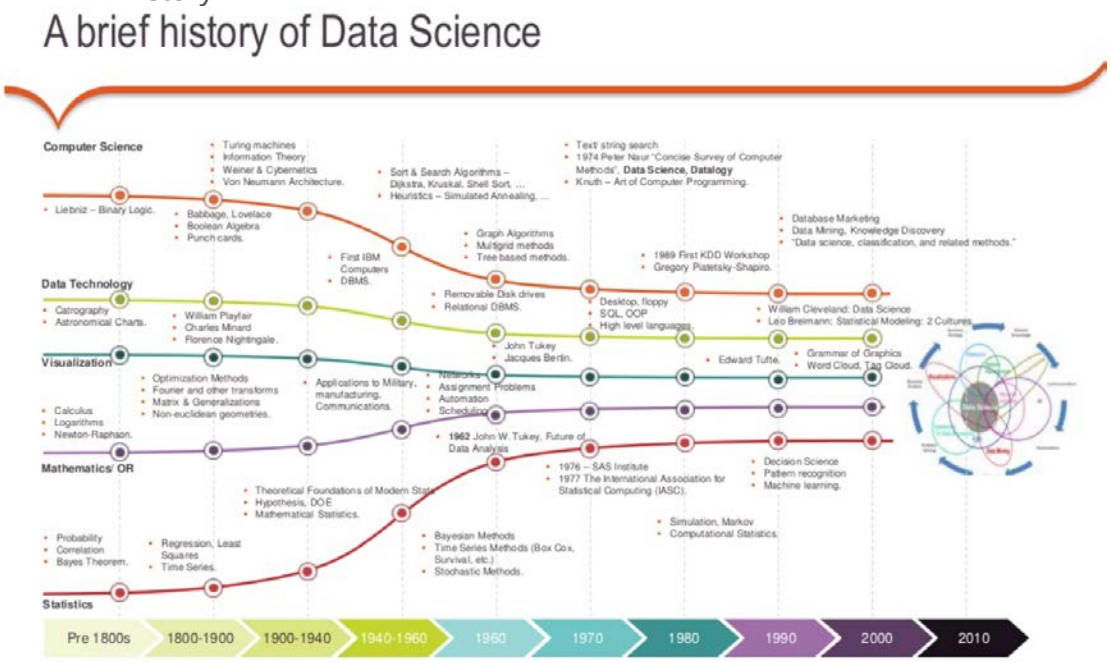
A **data scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes through each stage in the data lifecycle. Moreover, **Data science** is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process.

- What is R?
- What is R Markdown?

To maintain a reproducible workflow, you need to record what steps you take in a process. R Markdown is an authoring format (.Rmd files) that enables us to combine embedded R code

Week 2 – Impact of Data Science

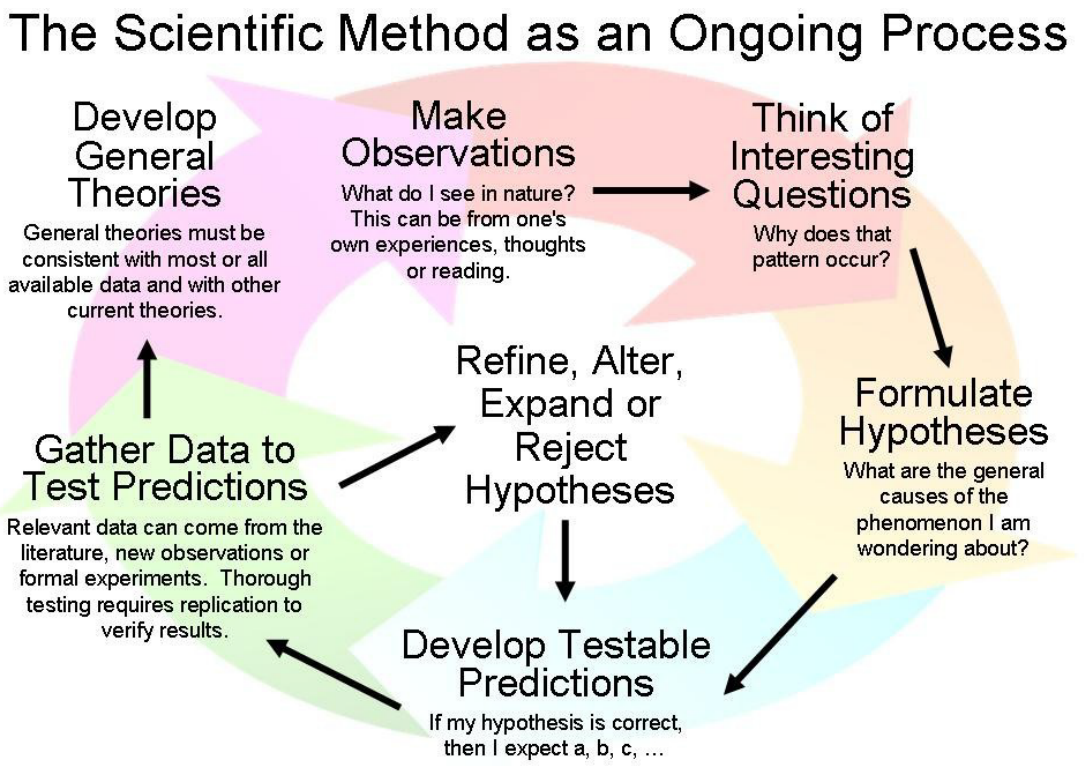
- Data science
    - History



    - Impact
        - On other disciplines
        - On society

Data, Predictions, and Decisions in Support of People and Society
Data Science for Social Good movement trainingdata scientists to support community and charity.

        - Scientific method



12

- Futurology

Your stomach can be instrumented to assess contents, nutrients, etc.
Your bloodstream can be instrumented too assess insulin levels, etc.
Your "health" dashboard can be online and shared by your GP

- R
  - R Markdown

R Markdown is an authoring format (.Rmd files) that enables us to combine embedded R code
with formatted text, so we can

  - ggplot2: visualising and aesthetics
    - Graphs & facets

will tell it to create a geom, a geometrical shape.
The facet creates the subplots for each category.

  - Data wrangling
    - wrangling verbs

Process of transforming "raw" data into data that can be analysed to generate valid actionable results and insights

    - Tidy data

Tidy data is an approach to structuring how you store and use data

- How has data science impacted on society?
- What is the basic syntax of R?
- How can you make data tidy?

In R, use the tidyverse collection of packages and the verbs for tidying
- **gather:** take a data set from wide to long
- **spread:** go from long to wide
- **separate:** split variables in one column to multiple columns.

- What is data wrangling?

Process of transforming "raw" data into data that can be analysed to generate valid actionable results and insights

Week 3 – Visualising statistics

• Types of analysis

Descriptive Analytics: gain insight from historical data

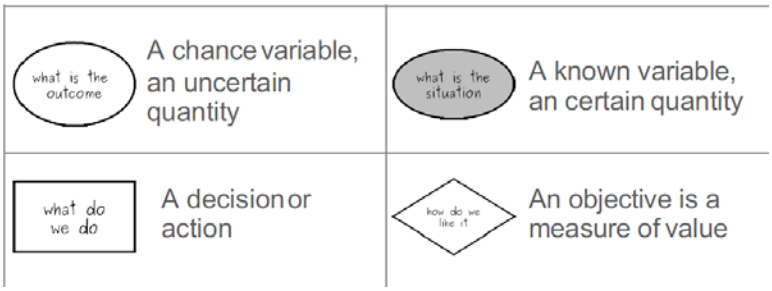
Predictive analytics: make prediction using statistical and machine learning techniques

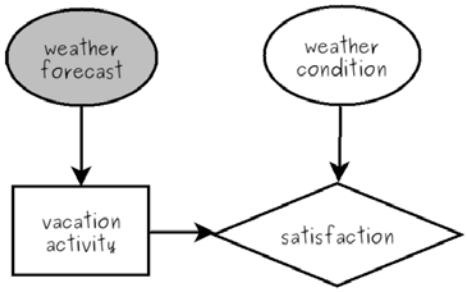
Prescriptive analytics: recommend decisions using optimization, simulation

• Modelling

    • Influence diagrams

consists of nodes and arcs and is an alternative to a decision tree.





• Growth laws

Moore's Law – capability and size of IT - Number of transistors per chip doubles every 2 years (starting from 1975)

Koomey's Law – capability and size of IT - Amount of battery needed will fall by a factor of 100 every decade

Bell's Law – purpose of IT - Roughly every decade a new, lower priced computer class forms based on a new programming platform, network, and interface resulting in new usage and the establishment of a new industry

Zimmerman's Law – relationship between - Zimmerman is creator of Pretty Good Privacy (PGP), an early encryption system

Surveillance is constantly increasing" and Privacy constantly decreasing

privacy and IT

• Business models

A business model describes the rationale of how an organization creates, delivers, and captures value, in economic, social, cultural or other contexts.

    • SaaS

software as a service(SaaS)

• Basic statistics

The practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative samples

    • Mean, variance

Variance is the average of how much values tend to differ from the mean.

    • Variable types

Categorical, qualitative
Groups or categories; **Nominal** – no natural ordering; **Ordinal** – ordered
Quantitative• Numerical; **Discrete** – specific values, like counts;**Continuous** – like temporal data

- Outliers and box plots

Outliers are values outside of the expected parameters for the data
Outliers need to be identified and decided on before the analysis is completed
Below Q1 - 1.5 IQR -- Above Q3 + 1.5 IQR

- Motion charts

Allow us to focus on the relationship between multiple variables over time.
Advantages: Time dimension allows deeper insights & observing trends
Disadvantages: Not suited for staticmedia

- Choosing visualisations

Variable type	How to map	Common errors
Categorical, qualitative	Display Category + Count/Proportion, often as an area plot or with a small number of categories mapped to colour or symbol.	Not including 0 on the Count/Proportion axis. Not ordering categories.
Quantitative	Position along an axis	Displaying as a bar, especially when showing mean values. Mapping to colour.
Date/Time	Time-ordered axis, different temporal resolutions to study long term trend, or seasonal patterns. Lines typically connect measurements to indicate temporal dependence.	Time order corrupted.
Space	Conventional projections of the sphere, map aspect ratio	Wrong aspect ratio

- How has data science changed business models?
- Information brokering service: buys and sells data/information for others.
- Information-based differentiation: satisfies customers by providing a differentiated service built on the data/information.
- Information-based delivery network: deliver data/ information for others.
- Information provider: business selling the data/ information it collects.
- How do the growth laws relate to data science?
- What statistics are commonly used in analysis?
- How is data commonly visualised?
- How can you visualise data in R?

Week 4 – Data quality

- Big data

- The Vs

Volume: The quantity of data to be stored

Velocity: The speed at which data enters the system and must be processed
Variety: Variations in the structure of the data to be stored
Veracity: correctness, truth, i.e., the lack of ...
Variability: change in meaning over time

- Growth laws

Moore's Law: Velocity, Volume
Koomey's Law: Variety
Bell's Law: Variety, Veracity
Zimmerman's Law: all of them

• NIST Case studies

- Analysis framework

Data sources: where does the data comes from? data volume: how much there is?
Data velocity: how does the data change over time?
Data volume: how much there is
Data variety: what different kinds of data is there?
Data veracity: is the data correct? what problems might it have?
Software: what software needed to do the work?
Analytics: what statistical analysis & visualisation is needed?
Processing: what are the computational requirements?
Capabilities: what are key requirements of the operational system?
Security/privacy: what security/privacy requirements are there?
Lifecycle: what ongoing requirements are there?
Other: are there other notable factors?

Data sources: clinical and claims data
Data volume: 1000 centres, 12 million patients, 4 billion clinical events
Data velocity: approx. 1 million clinical events/day
Data variety: free text, lab results, pathology, outpatient, etc.
Data veracity: different standards in different places
Software: Hadoop, Hive, Teradata, PostgreSQL, MongoDB
Analytics: visualisation for data checking; standardisation of incoming data; general data analysis
Processing: analytic processing, handling the volume
Capabilities: models to support subsequent cohort studies
Security/privacy: privacy and confidentiality required
Lifecycle: full data management required

• Data quality

- Wrangling

Data needs to be cleaned, so it can be (re)used
Volume – With a lot of data, irregularities creep in
Velocity – Data can be out-of-date very quickly
Variety – Data can be in a different formats and types that don't work well together
Veracity – The accuracy or consistency of data from different sources or sets or circumstances

- Missing data & strategies

Need to find where data is missing. Visualise the invisible!
Need to decide what to do with what we don't have
Sometimes we actually need to wrangle values for the missing data!

- Imputation

Simple parametric: Use the mean or median of the complete cases for each variable.
Simple non-parametric: Find the *k* nearest neighbours with a complete value and then average these.
Multiple imputation: Use a statistical distribution
Designing the imputation should take into account dependencies that you have seen between **missingness** and **existing** variables.

- NaN and NA

NaN – Not a Number
Value is not empty
Value is not a string or character
Value is not a number

NA - NA is not the same as NaN!
Value is not empty
Value is not the expected type
Value is Not Available

- Shadow matrix in R

arks the location of missings in the original data table -- `bind_shadow()`
We can then use the shadow matrix to see how the missing values relate to other variables in the table

- How do the Vs relate to Big Data?
- How can you describe the aspects of a data science project?
- What is data quality?
- How can you handle missing data?

If a small fraction of cases have several missings, drop the cases.
If a variable or two, out of many, have a lot of missings, drop the variables.
If missings are small in number, but located in many cases and variables, you need to impute these values (replace with substituted values) to do most analyses.

Week 5, Data sources

- Sharing data
  - Open data

Data that is "freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control

  - Data sources
  - Complexities of using shared data
  - Getting data
    - API, SaaS, etc.

Download files: from website, from archives
Via hardware
Send queries: from website interfaces, databases
Application Programmer Interface (API)
Software-as-a-Service (SaaS)

- Data standards
  - Formats: machine-readable, containers,markups

Machine-readable data: data (or metadata) which is in a format that can be understood by a computer, e.g., XML, JSON
Markup language: system for annotating a document in a way that is syntactically distinguishable from the text
Digital container: file format whose specification describes how different elements of data and metadata coexist in a computer file

  - Metadata

structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource
data about data
structured so that a computer can process & interpret it

Descriptive: describes content for identification and retrieval
Structural: documents relationships and links
Administrative: helps to manage information

  - Semi-structured data: XML, JSON
  - Recording models: PMML

PMML: Predictive Model Markup Language
PMML provides a standard language for describing a (predictive) model that can be passed between analytic software (e.g. from R to SAS).

- Combining data
  - joins

For data sets to be joined, they must have something in common

- Scripting languages: R, python, Unix shell code
    - Wildcard

* （0 or many）；?(one)

    - Piping

Piping: output from one command streams as input to another

    - Directing input/output

> and < to indicate the input and output

    - Moving files and directories
    - Analysing file contents: `grep, awk`

grep: output any lines in FILENAME that match PATTERN
awk: process text files in various ways

    - Handling big data

Piping shells commands buffers their execution

- Standardisation
    - software

Ideally, the software used for data should be standardised
Rapid prototyping is what scripting languages are ideal for

    - workflow

Need to standardise how data is accessed
Need to be able to reproduce

    - processes

The context of working with data also needs to be recorded

- What information can you store about data?
- What is the significance of open data?

open data provides new opportunities for business, new products and services, and can raise productivity
open data supports public understanding and citizen engagement
scientists need to better publicise their data (with help from universities, etc.)
industry sectors should work with regulators and coordinate industry collaboration
collaboration across sectors in both public and private settings

- Why are standards needed in the data science process?
- What are some basic shell script commands?
- How can you join data?

Week 6 – Modelling data

- Temporal data
    - Temporal elements

Date, Time

    - Extraction and conversion

Decompose/convert once, use often
Numbers are easier to work with than words

    - Visualisation

Line plots: Highlight temporal continuity within and between time periods
Calendar plots: Helps visually identify irregularities

- Statistical modelling

Statistical models represent the relationships between variables

    - Variables: dependent, independent

A model can be used to predict about the dependent variable, given information about the independent variables

    - Causation vs correlation

Causation indicates that one event is the result of the occurrence of the other event
Identifying causation requires controlled experiments

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

- Regression modelling
  - Model family

A **model family** is the proposed function form used to describe that relationship

  - Learning parameters/fitting a model

Parameters are unknown values that need to be estimated/learnt/trained from the data.

  - Simple linear regression model

one independent variable, straight-line, representing the strength of the relationship

- How are temporal elements included in data?

The temporal aspect of the data can be of different types. Specific and Relative.

- How are correlation and causation different?
- How can you model data?

Week 7 – Fitted modelling

- Truth of data
  - Error

Error measures the distance between the prediction & the actual value.

  - Correlation coefficient

measure the **strength and direction** of the linear relationship of two variables

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Simple linear regression
  - Residuals: Mean Square Error

the distances between the observed values and the predicted values

  - Goodness of fit: fitting variation
    - R-squared

value for a fitted model is a key goodness of fit statistic.
R2 is between 0 and 1
1 is good, variability in y is fully explained by the model
0 is bad, no variability in y is explained by the model

- Polynomial regression
  - Degree

Bayesian information criterion (BIC) includes a penalty for using more variables

  - Underfitting, overfitting

Poor fit due to high bias called under-fitting -- cannot fit the data well
Poor fit due to low bias called overfitting -- fits the data too well

  - Testing and training

Split up the data we have into two nonoverlapping parts, a training set and a test set

  - Bias-Variance tradeoff

Bias: measures how much the prediction differs from the desired regression function

Variance: measures how much the predictions for individual data sets vary around their average

- No Free Lunch Theorem

If a [learning] algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.

- Multiple models & ensembles

An ensemble is a collection of possible/ reasonable models

- How do you know if a model fits the data?
- How can multiple models sometimes help?

Week 8 – Grouping data

• Segmentation

Sometimes the segmenting of data is because of the context of the data
Sometimes we don't have pre-determined segments, but we want segmentation
- Some of the data may be similar
- Better decision-making if we consider each segment independently

latent means the variable is never observed in the data.

• Regression trees

A regression tree is a supervised machine learning algorithm that predicts a **continuous-valued response variable** by learning decision rules from the predictors (or independent variables)

Rather than using a single function to represent the data, divide the data into similar segments, then make predictions in each segment

- ANOVA

  - As in for the population inside that node, this pair of predictor/value **improves** the chosen criteria (e.g., ANOVA) the most.

- ANOVA criterion = $SST - (SSG1 + SSG2)$
- $SST = \sum(y_i - \bar{y})^2$, total variation of the dependent variable.
- $SSG1$ & $SSG2$ use the SST formula but with the values for the two subgroups created by the partition.

• Classification trees

For classification task, if we want to use a decision tree, the result is a classification tree.
most popular split criteria are **Gini** and **Entropy**.

• Clustering

Clustering tends to be associated with segmentation that allows us to recognize similar combinations of attribute values when we don't have predefined categories.

- Centroids

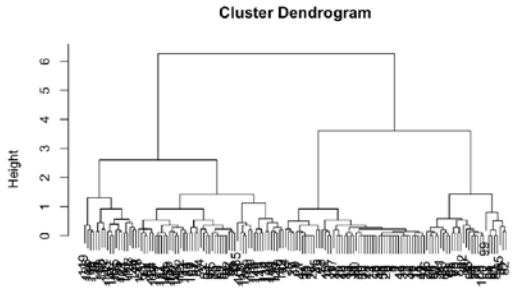
Value of a plot in the centre of the cluster

- K-means

Randomly select centroids for K clusters → Find mean values in each cluster and use that as new centroid

- Hierarchical trees: dendrograms

Clusters within clusters
The results of hierarchical clustering are usually presented in a dendrogram



- Network data: degree, centrality

Degree – How many connections a node has
Degree centrality - most connections
Betweeness centrality - most used for shortest trip between other nodes
Closeness centrality - least 'hops' to travel to other nodes

- How do decision trees relate to data?

divide the data into subsets of similar values
estimate the response within each subset

- How can you group data?

Week 9 – Big data

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

- RDBMS: SQL
    - Unstructured data: NoSQL

Use NoSQL database when: Storing large volume of data with little to no structure and Data changes rapidly
NoSQL databases offer a rich variety beyond traditional relational.

- Distributed systems
    - Hadoop

Hadoop provides an inexpensive and open source platform for parallelising processing:
based on a simple Map-Reduce architecture
not suited to streaming (suitable for offline

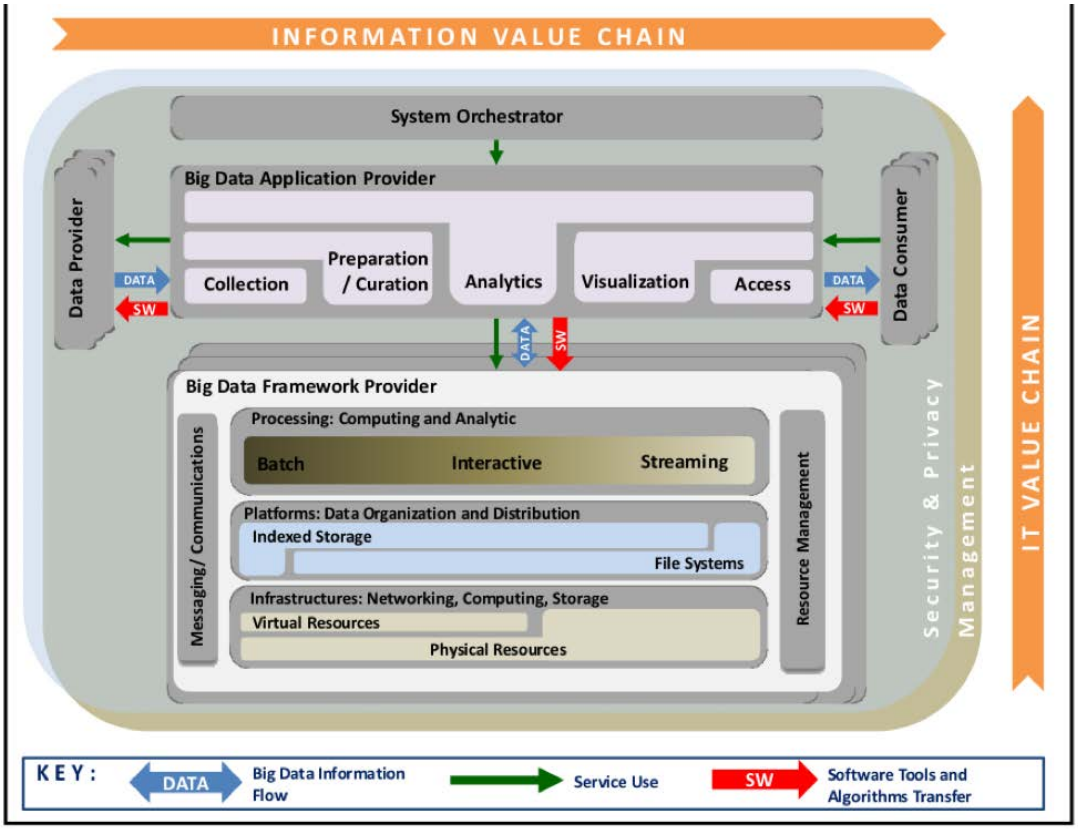
    - Map-Reduce

Simple distributed processing framework developed at Google
intended to run on commodity hardware
requires simple data parallelism followed by some merge ("reduce") process

    - Spark

Spark is a more recent development than Hadoop
includes Map-Reduce capabilities
provides real-time, in-memory processing, much faster than Hadoop processing)

- NIST Big Data Reference Framework



- Data Science Tools & Services
    - Open source software

1. Apache Hadoop Distributed File System(HDFS)
2. Apache Hadoop YARN
3. Apache Spark
4. Apache Cassandra (distributed NoSQL, wide-column store)
5. Apache HBase (distributed NoSQL, wide-column store)
6. Apache Hive (distributed SQL)
7. Apache Mahout (distributed linear algebra with GPU)
8. Apache Pig (data flow and data analysis on top of Hadoop)
9. Apache Storm (distributed real-time computation)
10. Apache Tez (dataflow for Hive and Pig)

    - Case studies
    - APIs
    - SaaS

Examples: Email systems, File sharing systems(Google Drive), Business systems
Pros: Pay as you go, Scale up/down, Low maintenance 软件即服务，就像试用期

- How has big data affected data storage and processing?

Week 10 – Data management

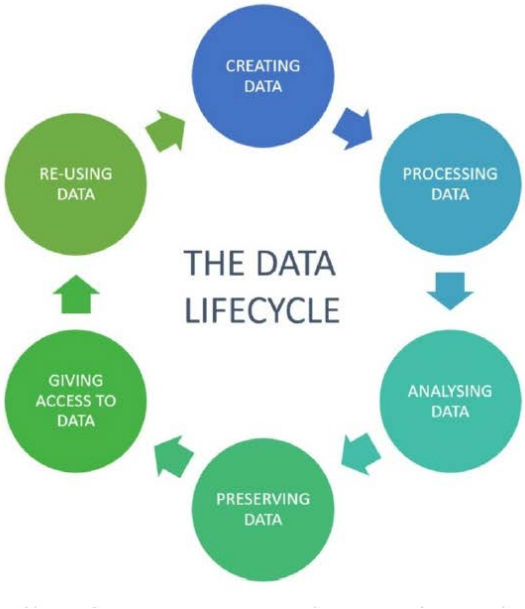
- Data management

Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.

Science: reproducibility and credibility of scientific work, producing artifacts of knowledge, creating scientific data
Business: governance, compliance, information privacy, etc.

Government: a unique legislative environment that regulates them (e.g., “transparency”), archiving, FOIs, support data infrastructure, etc.
Medicine: significant privacy issues, conflicting corporate financial constraints, government regulations and furthering
of medical science

- Data lifecycles



- Data governance
    - Legal requirements: Privacy Act, GDPR, licenses
    - Ethical requirements

Ethics as the moral handling of data
e.g., not selling on other’s private data to scammers44

    - Rights
    - Privacy
    - Confidentiality

- Stakeholders

Stakeholders are any parties that have a relationship with a project/policy/product/data. This includes the data’s source , managers, analysts and users, IT developers, data scientists

- Data management plans

Improvements to efficiency, protection, quality and exposure
Value, Innovation, Data curation
Content: Backups, Survey of existing, Data owners & stakeholders, File formats, Metadata, Access and security, Data organisation, Storage, Data sharing, publishing and archiving

- What is data management?

Data Management is what you do to handle the data
Data usage is no good without good data management
Resources, practises, enacting policies

- What is data governance?

Data Governance is making sure that it is done appropriately
Policies, training, providing resources
Planning and understanding

- What is a data lifecycle?
- How should stakeholders be responsible with data?

The responsibility of looking after the data lies with all stakeholders – the data scientist is a key to making this happen

Week 11 – Issues

• Data management capability maturity

Data acquisition, processing and quality assurance Goal: Reliably capture and describe scientific data in a way that facilitates preservation and reuse
Data description and representation Goal: Create quality metadata for data discovery, preservation, and provenance functions
Data dissemination Goal: Design and implement interfaces for users to obtain and interact with data
Repository services/preservation Goal: Preserve collected data for long-term use

A mature system manages data all through the data lifecycle and **throughout all projects**.

• Linked data: Semantic web, RDF

Connecting elements within multiple structured data sets
Allows data relating an element to be collected from multiple data sets
Expands the knowledge base of a single dataset
-Linked Open Data (LOD) allows the links and data to be freely shared and accessed

Semantic web :On the web, use the URIs Semantic web when mentioning things
Resource Description Framework (RDF) is another style of language for representing (subject, verb, object) triples, which is used to represent semantics. It is a core representation language for Linked Open Data and the Semantic Web

- Confidentiality

We often don't own or manage corporate/internet/app data about ourselves
The source data is critical for advertisers so we cannot expect companies to be banned/excluded from using it and For many apps/websites, you must accept their privacy data sharing policies to use their services fully
The interface for selecting privacy preferences should move away from individual Internet platforms and be put into the hands of individual consumers

- Privacy

• Surveillance

- Data retention laws

require some telecommunications service providers to retain specific telecommunications data (the dataset) relating to the services they offer for at least 2 years

• Data veracity

It is easy for the modelling to misrepresent what the data is supposed to reflect.Even statistical analysis can be biased!

- Bias

Not all bias is in the numbers, Bias can also be in how you have designed the research
Are the variables appropriate for all situations being modelled?
Are assumptions being made about the context of the data?

- Human-in-the-loop

Need the human perspective in the design, understanding and review of the process, how it is utilised and its results

- Sampling
  - A/B testing

Blind experiments or A/B testing may be used to show if relationship between various variables
The validity of the the hypothesis is based on whether A has a different response to B, where the response is the target variable(那个网站点的人多)

- Significance testing: p-value, k-fold testing

k-fold testing: experiment with k combinations of test and training data

- Scientific method

Scientific method isn't the only valid research methodology!
Still need to make sure any modelling or other research outcomes are valid!

- How can linked data be used?
- Why might data science results not always be appropriate?

Section	Title	Week mentioned
1.1-1.3		1
1.4	What is Data Science	1, 2
1.5	Roles of a Data Scientist	1
1.7	Impact of Data Science	2
2.1	Data & Decision Models	3
2.3	Business Models with Data	3, 9
2.5	Application Areas	4, 5
3.1	Characterising Data	3, 4, 5
3.3	Data Case Studies	4, 5
3.4	Big Data Processing	9
4.1	Introduction to Resources	5
4.5	Standards and Issues	5, 9
4.7	Interviews on Software and Tools	9
4.8	Case Studies of Standards and Issues	5, 9
5.1	Introduction to Data Analysis	7
5.2	Theory of Data Analysis	6, 7
5.4	Tools for the Data Analysis Process	6
5.5	Activity: Decision trees with BigML	8
5.6	Activity: Prediction with BigML	8
5.7	Data Analysis Case Studies	11
6.1	Issues in Data Curation and Management	10, 11
6.2	Frameworks for Data Management	10, 11
6.3	Interviews on Data Management	10, 11

代码关注，以 shell 代码为主

Unix commands

• pwd: path of current directory
• cd DIRPATH: change directory to DIRPATH
• ls DIRPATH: output the filenames of DIRPATH
• cp FILENAME NEWFILENAME: copy FILENAME to NEWFILENAME
• mv FILENAME NEWFILENAME: rename FILENAME to NEWFILENAME
• echo “TEXT”: output TEXT
• cat FILENAME: output the contents of FILENAME
• less FILENAME: output the contents of FILENAME, one screen at a time (can page up and down)

• wc FILENAME: count the number of characters, terms, lines in FILENAME
• grep “PATTERN” FILENAME: output any lines in FILENAME that match PATTERN
e.g. grep “Australia” product*v1.txt
• head FILENAME: output the first lines of FILENAME
• tail FILENAME: output the last lines of FILENAME
• awk: process text files in various ways, including search and replace cf. sed, perl
• man COMMAND: output user manual pages for COMMAND
• COMMAND –?: output shorter help pages for COMMAND
• COMMAND –-help: ditto