# 2019 S2 FIT5145 Final Review

| | |
|---|---|
| Created | @Oct 22, 2019 11:54 AM |
| Tags | FINAL REVIEW FIT5145 |

## Table of Contents

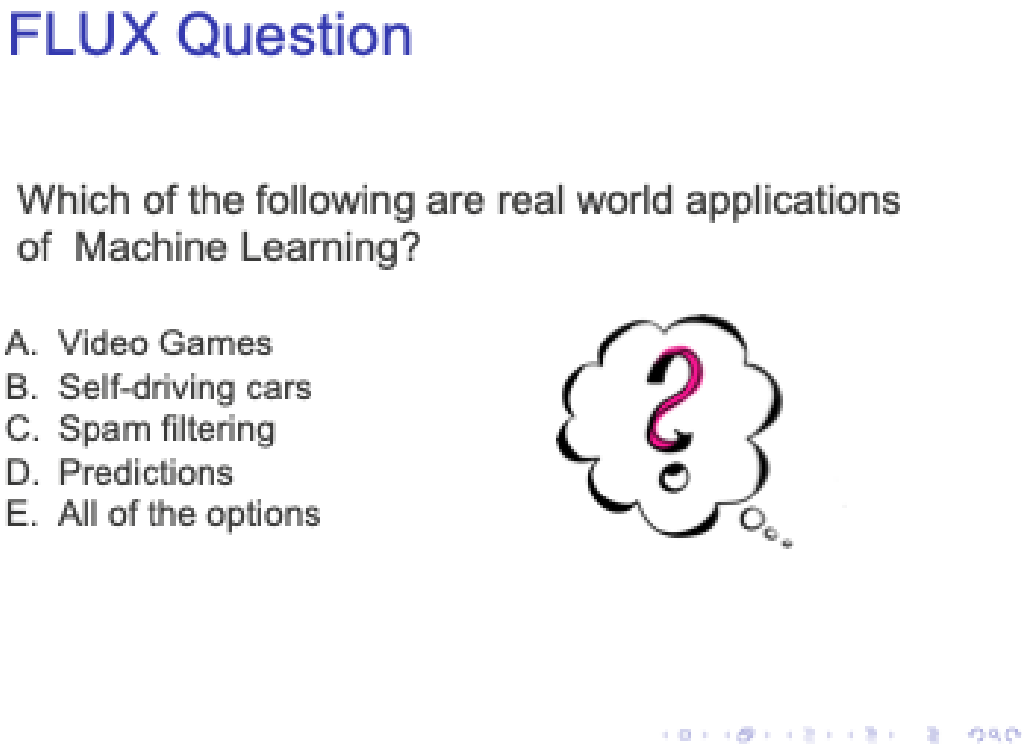
# Lecture 1 - Introduction to Data Science

- What is Machine Learning (p23)
  - Machine Learning is concerned with the development of algorithms and techniques that allow computers to learn.
  - The underlying theory is `statistics`

▼ Why use Machine Learning (p24)

- Human expertise is not available (Martian exploration)
- Humans cannot explain their expertise as a set of rules or their explanation is incomplete and needs tuning (speech recognition)
- Many solutions need to be adapted automatically (user personalisation)
- Situation changes overtime (junk email)
- There are large amounts of data (discover astronomical objects)
- Humans are expensive to use for the work (handwritten zipcode recognition)

▼ FLUX Question (p32) E

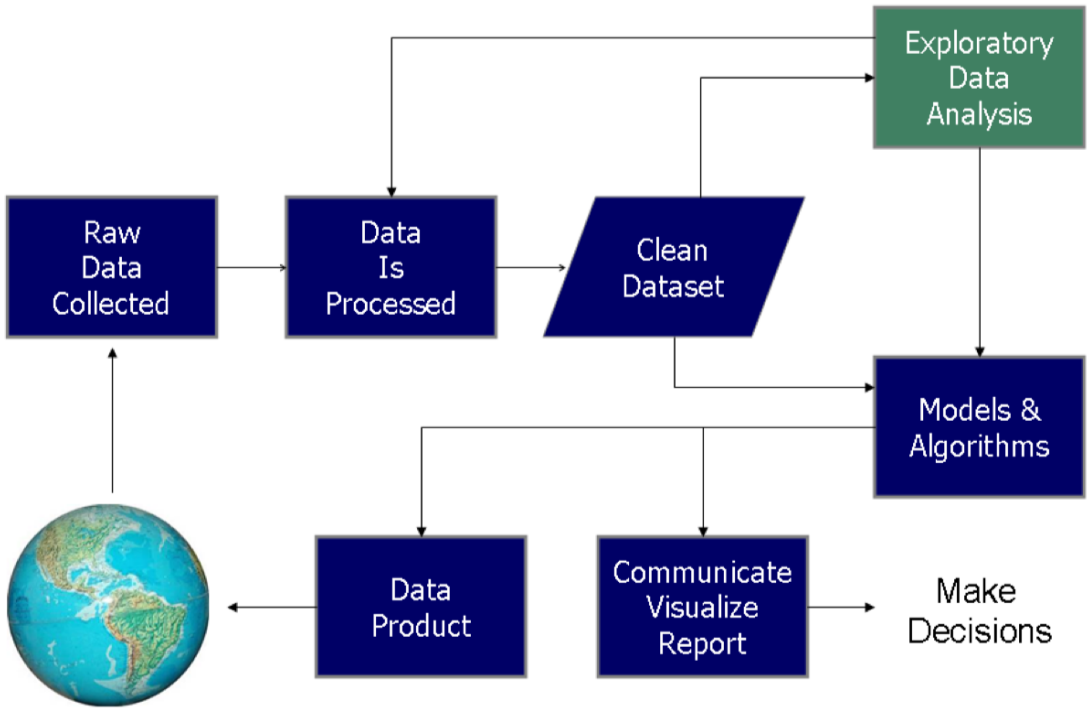



▼ The Data Science Process: (p38-54)

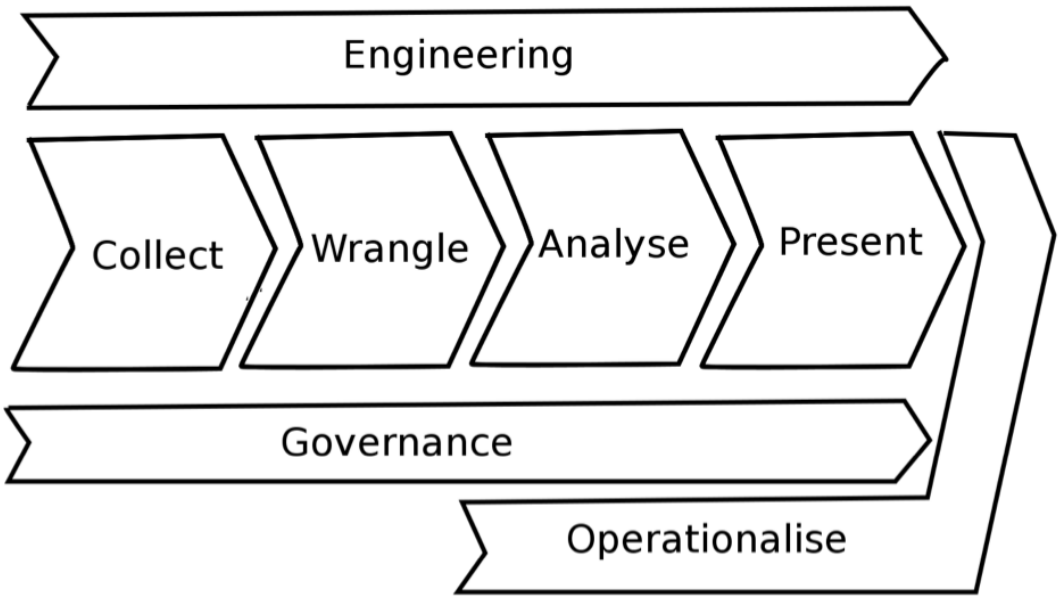
1. Pitching ideas
2. Collecting data
3. Monitoring
4. Integration
5. Interpretation

6. Governance
7. Engineering: back-end work
8. Wrangling: inspecting and cleaning the data
9. Modelling: proposing a conceptual/mathematical/functional model
10. Visualization
11. Operationalization: putting the results to work

- Data Science Process Flowchart (p55)

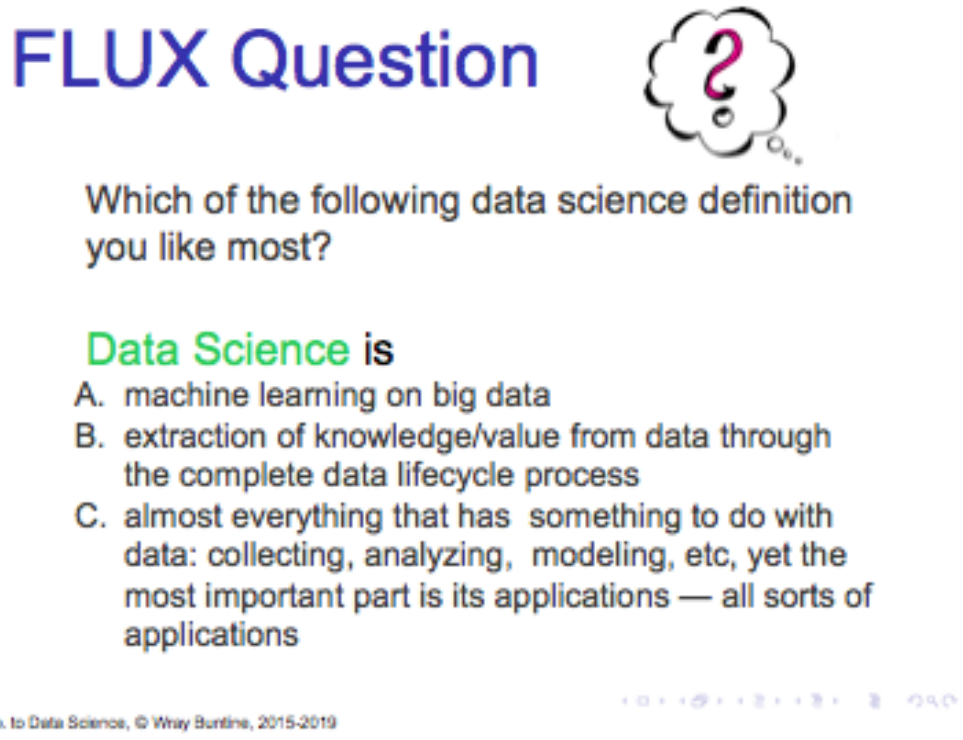



- Standard Value Chain (p58)
  - Collection → Engineering → Governance → Wrangling → Analysis → Presentation → Operationalisation




- Definition of Data Science and Big Data (p65-69)
  - narrow: machine learning on big data
  - broad: extraction of knowledge/value from data through the complete data lifecycle process

▼ FLUX Question (p70) C



# Lecture 2 - Jobs, Roles and the Impact

- Difference between `Data analysts`, `Data Scientists`, `Data engineers` (p14)
  - Data analysts: develop insights with data
  - Data scientist: develop data models and products
  - Data engineers: manage data infrastructure, automate data processing and deploy models at scale
- Standard Value Chain (p19)
- Skills of Data Scientists (p24)
  - 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

▼ Which of the following has not been proposed as a definition of Data Science?

a. Extraction of actionable knowledge from data through the complete lifecycle process.

b. The use of learning on big data in a distributed computing environment.

c. Organisation, storage, modelling, reporting on data and the delivery of value.

d. Almost everything concerned with data, especially applications.

e. Data-driven processes in science or business.

▼ Impact of Data Science (p33)

- Cloud Service: datafication of you
- Science: scientific method
- Social Good: p47 examples
- Futurology: healthcare and automobiles
- Others: crowd-sourced competitive sport, infographics

▼ FLUX Question (p37) D

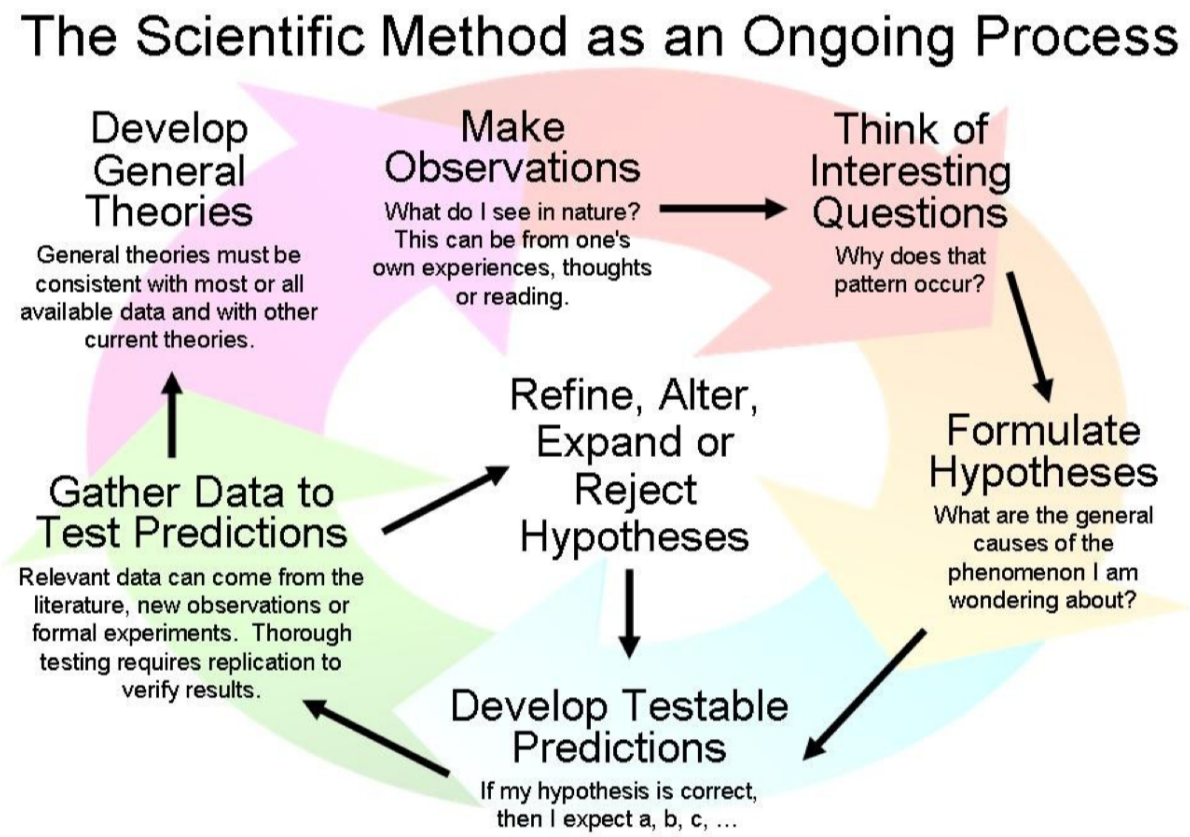
## FLUX Question

What role has the internet had in the development of data science?

**A.** the first big users of data science were internet giants
**B.** source of data for use
**C.** avenue for data science tools
**D.** all of the options



- Scientific method and data science (p40-41)




▼ The Scientific Method can be impacted by Data Science because:

a. It shows correlation is adequate for providing deep understanding

b. The technology to adopt data-driven science exists, and its importance is more widely understood.

c. It offers new statistical theories to use.

d. Non-scientists can contribute via data science competitions.

▼ Which of the following is not a historical precursor to Data Science?

a. Google and Amazon's data-driven operations.

b. IBM's cognitive computing.

c. Statistical platforms such as R, SAS or Weka.

d. Cheap computing hardware.

▼ When did Data Science begin?

a. 1962, with the article "The Future of Data Analysis".

b. Data Science has been happening for a long time, it just called something else.

c. 1974, when Peter Naur used the term 'data science' freely in 'Concise Survey of Computer Methods'.

d. 1996, IFCS, "Data science, classification, and related methods".

> We have seen several views of the history of Data Science, most differ, some go
> back to the dawn of civilisation (e.g. Wolfram). You can trace the history of the term "Data
> Science" but that's not the same.

▼ According to Cukier and Mayer-Schoenberger:

a. Google's clever use of high-quality translations allowed them to develop a better automated translation system.

b. Datafication is the digitization of formerly analogue content.

c. Observing and making use of correlations from big data is a move away from trying to develop deep causal models, traditionally done in science.

d. A key technology for using big data is the use of statistical sampling, being able to infer information from small, carefully curated samples.

> This is presented in the video and the article in Overview of Data Science. The idea is that correlations are easier to detect and can still be used even if the underlying mechanisms are not understood.

▼ Which of the following is not a characteristic of a data scientist according to Mike Loukides of O'Reilly?

a. Be able to program computation on data.

b. Be interdisciplinary.

c. Be versatile.

d. Be able to read a statistical journal article.

e. Be entrepreneurial.

> Some data scientists might need to read a book on machine learning, and all need to be able to use some standard statistical or machine learning software.

---

# Lecture 3 - Data Business Models

▼ Advantages & Disadvantages of Motion Chart (p2)

A:

- time dimension allows deeper insights & observing trends
- good for exploratory work
- motion allows identification for this out of common “rhythm”
- “appeal to the brain at a more instinctual intuitive level”

D:

- not suited for static media
- display can be overwhelming, and controls are complex
- not suited for representing all types of data, e.g. other graphics might be suitable for business data
- “data scientists who branch into visualization must be aware of the limitations of uses”

▼ How many dimensions of data can Motion Charts display?

a. 4+ dimensions

b. 1 dimension

c. 2 dimensions (e.g. x and y)

d. 3 dimensions

> Motion Charts require a minimum of 3 dimensions (category, datetime, numerical
> data), but can handle more.

▼ Motion Charts is given data from 1900 to 2000, but only intermediary years are displayed. What process is Motion Charts using to display this?

a. Extrapolation

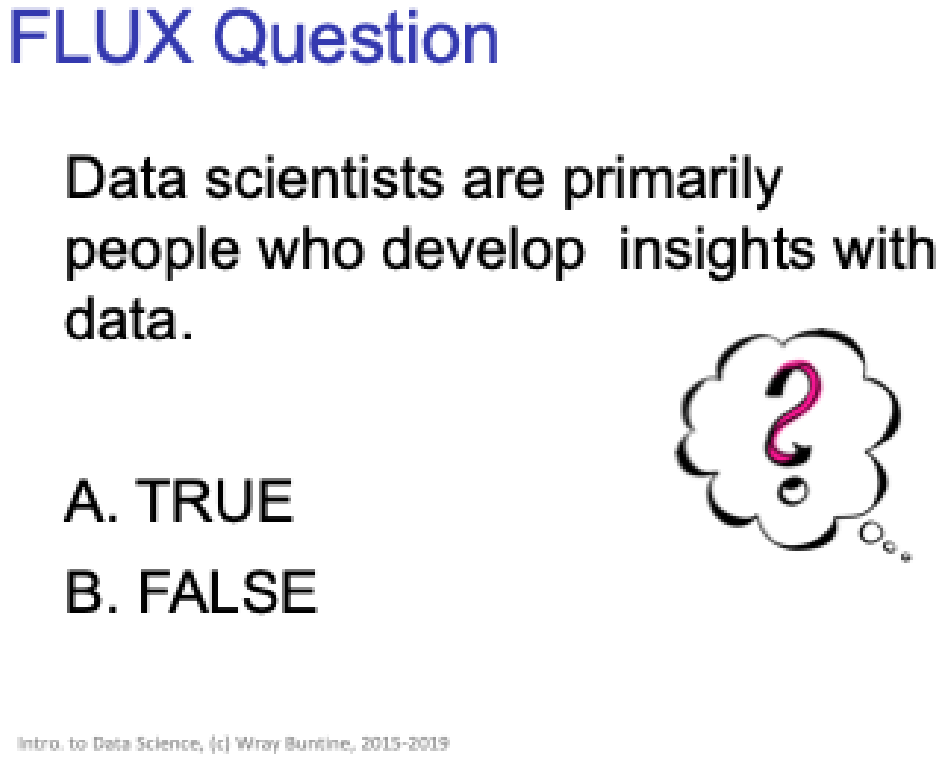
b. Data munging

c. Hyperpolation

d. Interpolation

> Motion Charts is ‘filling in the gaps’. Interpolation is ‘between the points’ of a
> sequence of values.

▼ FLUX Question (p5) B




▼ Standard Value Chain (p11)

- Collection → Engineering → Governance → Wrangling → Analysis → Presentation → Operationalisation

▼ Pivotal's data value chain takes a broader view of the use of data, from capture and storage through to changing business practices. How does their notion of a data scientist differ from that given by our Standard Value Chain?

a. Their DS doesn't identify candidate business tasks to apply data science too.

b. Their DS doesn't propose changes to applications.

c. Their DS doesn't do visualisation.

d. Their DS doesn't do the major data extraction task.

> That's what Data engineers do.

▼ Which of the following would not be regarded as a step in the data value chain?

a. Security.

b. Transform and prepare.

c. Engineering.

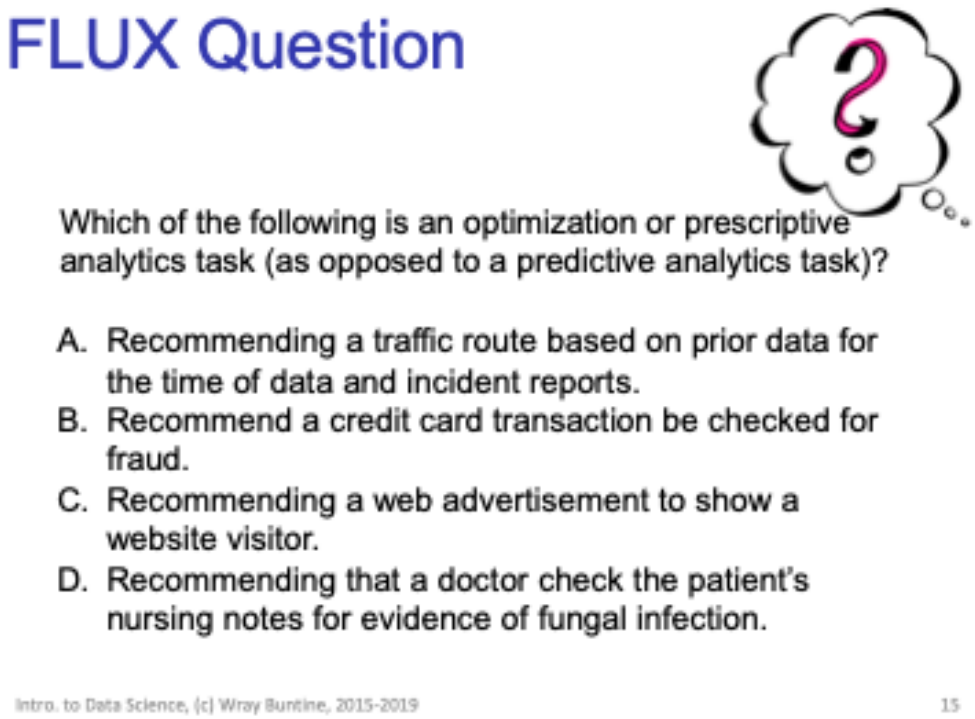
d. Presentation.

> This is but a small part of the management or governance step.

▼ Analytic Levels (p14)

- Descriptive analytics: gain insight
- Predictive analytics: make prediction
- Prescriptive analytics: recommend decisions

▼ FLUX Question (p15) A

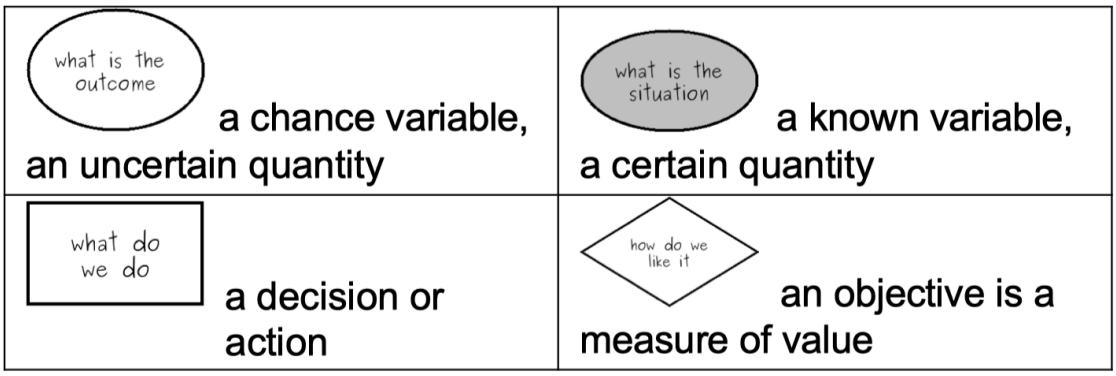



> The point is it has to have "optimization"

▼ When working with SAS Visual Analytics, what is the meaning of the term “drill down”?

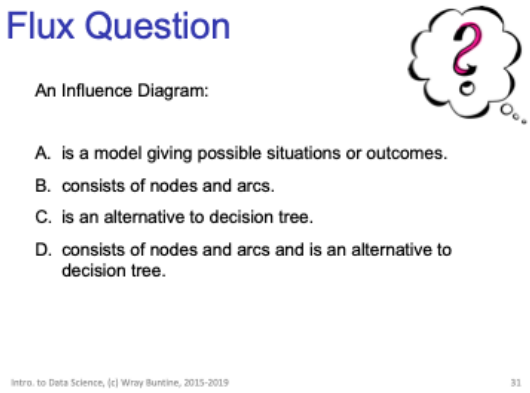
a. To build the database.
b. To explore lower levels of data in a hierarchy.
c. To enter data.
d. To put your tools down and have a break.

- What is Influence Diagrams (p20)
  - directed graphical model
  - ▼ four types of nodes
    - changce nodes
    - known variable nodes
    - action/decision nodes
    - objective/utility nodes




- When do we connect an arc to a node & Four types of nodes (p21-30)
  - `influence`
  - cause, time sequence

▼ FLUX Question (p31) D




> Influence Diagram and Decision Trees are complementary views

- Business models: A business model describes the rationale of how an organisation creates, delivers, and captures value, in economic, social, cultural or other contexts

▼ Data Business Models (p45)

  - 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

▼ FLUX Question (p49) B

## FLUX Question

What is .ipynb?

A. An illegal file extension.
B. Interactive Python NoteBook.
C. Intelligent Python Nota Bene.
D. Typo, it should be 'pinyin'



▼ FLUX Question (p50) D

## FLUX Question

What is a dataframe?

A. An array.
B. A list.
C. A theory about data.
D. A structure that stores tabular data



# Lecture 4 - Application Areas and Case Studies

▼ NIST Analysis (p8)

- data sources: where does the data comes from
- data volume: how much there is
- data velocity: how does the data change over time
- data variety: what different kinds of data is there

- data veracity: is the data correct
- software: what software needed to de 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

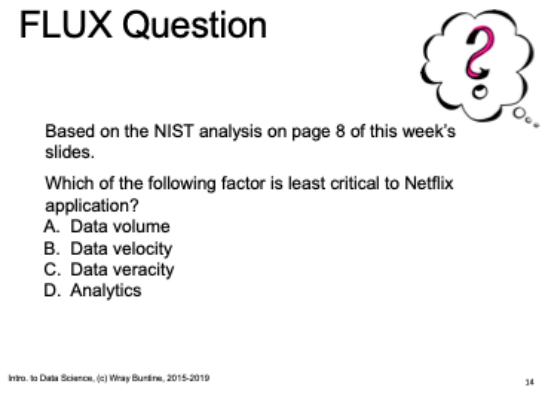
▼ Netflix case study

- data sources: user movie ratings, user clicks, user profiles
- data volume: in 2012: 25 million users, 4 million ratings/day, 3 million searches/day, video cloud storage of 2 petabytes
- data velocity: video titles change daily, rankings/ratings updated
- data variety: user rankings, user profiles, media properties
- software: Hadoop, Pig, Cassandra, Teradata
- analytics: personalised recommender system
- processing: analytic processing, streaming video
- capabilities: ratings and search per day, content delivery
- security/privacy: protect user data; digital rights
- lifecycle: continued ranking and updating
- other: mobile interface

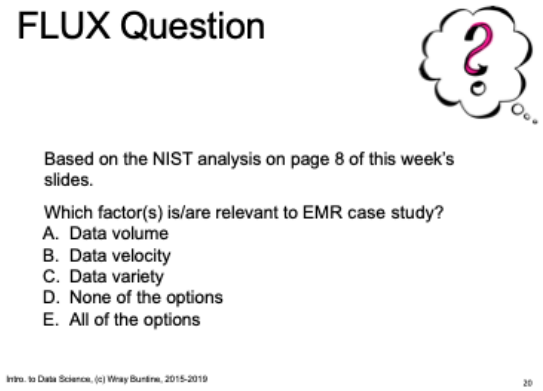
▼ Other cases in Slides

▼ Four Vs of Big Data (p9)

▼ FLUX Question (p14) C




▼ FLUX Question (p20) E



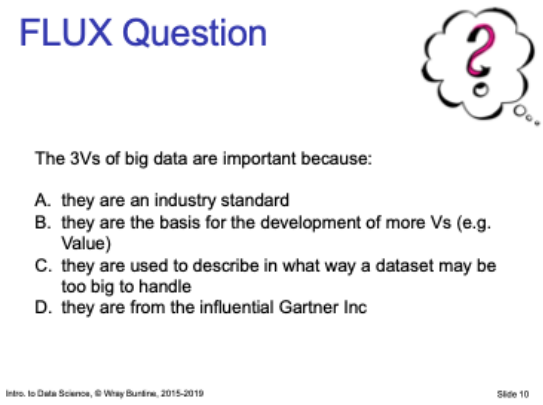

- ▼ Some application area of big data (p32)
  - Health
  - Government
  - Retail
  - Manufacturing
  - Location Technology

---

# Lecture 5 - Characterising data and "big" data

- ▼ Python versus R (p3)
  - both are free
  - R developed by statisticians for statisticians, huge support for analysis
  - Python by computer scientists for general use
  - R is better for stand-alone analysis and exploration
  - Python lets you integrate easier with other systems
  - Python easier to learn and extend than R (better language)
  - R has vectors and arrays as first class objects; similar to Matlab!
  - R currently less scalable
  - The index starts from 1 in R not 0 like in Python
- ▼ Some general characteristics of data sets used to assess a project (p5)
  - the V's (p6-9)
    - 3V & 4V
  - ▼ metadata (p14-16, 24)
    - metadata: structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource → metatdata is data about data
    - ▼ types of metadata:
      - descriptive: describes content for identification and retrieval e.g. title, author of a book

        - structural: documents relationships and links e.g. chapters in a book, elements in XML, containers in MPEG
        - administrative: helps to manage information e.g. version number, archiving date, Digital Rights Management (DRM)
    - ▼ why use metadata:
        - facilitate data discovery
        - help users determine the applicability of the data
        - enable interpretation and reuse
        - clarify ownership and restrictions on reuse
    - ▼ key concepts:
        - machine-readable data: data 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 e.g. Markdown, Javadoc
        - digital container: file format whose specification describes how different elements of data and metatdata coexist in a computer file e.g. MPEG
- dimensions of data
- ▼ growth laws (p37-42)
    - Moore's Law: number of transistors per chip doubles every 2 years
        - ▼ transistor count translates to:
            - more memory
            - bigger CPUs
            - faster memory, CPUs
        - pace currently slowing
    - Koomey's Law: amount of battery needed will fall by a factor of 100 every decade
        - leads to ubiquitous computing
    - Bell's Law: 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. (PC → mobile computing → cloud → IoT)
    - Zimmerman's Law:The natural flow of technology tends to move in the direction of making surveillance easier, and the ability of computers to track us doubles every eighteen months.
- B→K→M→G→T→P→E
- ▼ Which of the following are metadata?
- ▼ FLUX Question (p10) C




> 3 V's charaterise bigness adequately

▼ When a scientist collects data from an explosion, which of the four V's is paramount?

a. Variety

b. Volume

c. Veracity

d. Velocity

▼ Which laws are predated by (and derived from), Moore's Law?

a. Koomey

b. All of them.

c. Zimmermann

d. Bell

▼ Which growth law says power consumption will decrease?

a. Moore's Law.

b. Zimmerman's Law.

c. Bell's Law.

d. Koomey's Law.

---

# Lecture 6 - Data Sources and Case Studies

▼ Unix Shell: Useful for managing and manipulating large files

- without ever loading them fully into memory
- using pipes allow us to process files as a stream
- allows us to deal with files that are too big for applications and/or don't fit into memory

▼ Factors that influence data science

- business needs

- data analysis and general wrangling tools
- the internet
- big business recognition

▼ Database Review (p10,16,17)

- RDBMS: Relational Database Management Systems
- SQL: structured query language

▼ Storing and accessing data

- SQL Database (when data is structured and unchanging)
- No-SQL Database (when storing large volume of data with little to no structure or data changes rapidly)
  - Object DB
  - Doc.DB
  - key-val cache
  - key-val store
  - tabular key-val
  - graph DB
- JSON

▼ Database background concepts

- in-database analytics: the analytics is done within the DB
- in-memory database: the DB content resides memory
- cache: data stored in-memory
- key-value: value accessible by key, e.g. hash table
- information silo: an insular information system incapable of reciprocal operation with other, related information systems

▼ What are the ‘walls’ that the RDBMS/SQL paradigm has hit in business?

a. Businesses function in a continuously changing environment, therefore the DB schema must change.

b. Businesses require data driven decision-making, massive amounts of data.

c. All options are correct.

d. Businesses require insights faster i.e. in real time.

▼ Types of Processing (p19-20)

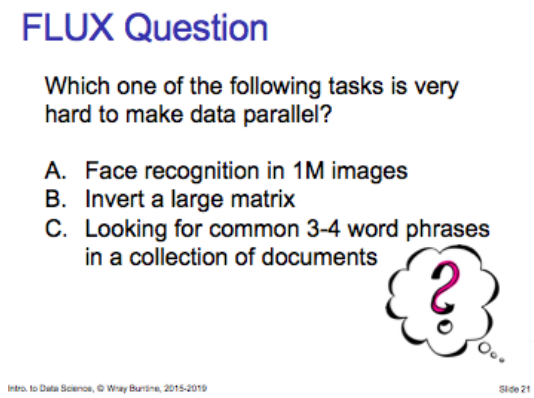
- Interactive: bringing humans into the loop
- Streaming: massive data streaming through system with little storage
- Batch: data stored and analysed in large blocks, "batches", easier to develop and analyse

▼ Processing background concepts

- in-memory: in RAM, i.e., not going to disk
- parallel processing: performing tasks in parallel

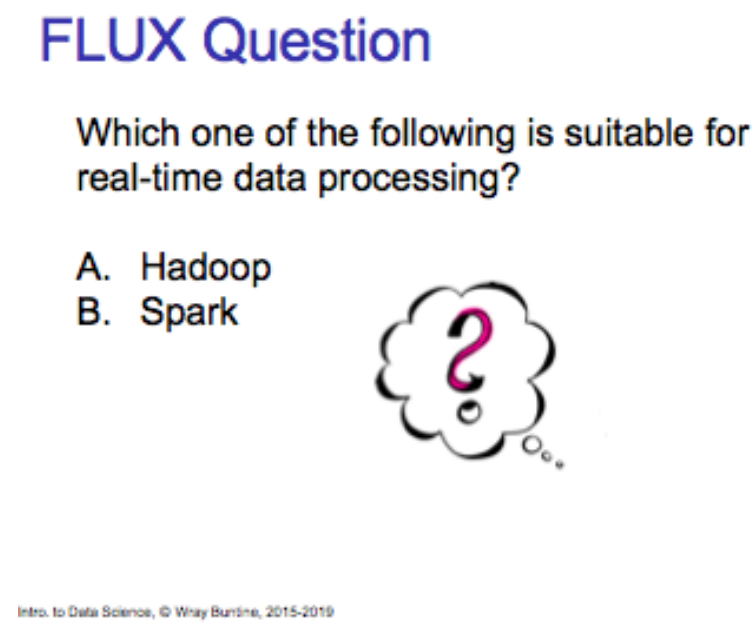
- distributed computing: across multiple machines
- scalability: to handle a growing amount of work; to be enlarged to accommodate growth (not just “big”)
- data parallel: processing can be done independently on separate chunks of data

▼ FLUX Question (p21) B




- Map-Reduce (p23-27)
  - MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks.
  - Stopped using by Google at 2005
- Hadoop: Open-source Java implementation of Map-Reduce
  - based on a simple Map-Reduce architecture
  - not suited to streaming (suitable for offline processing)
- Spark: builds on Hadoop infrastructure
  - includes Map-Reduce capabilities
  - provides real-time, in-memory processing
  - much faster than Hadoop

▼ FLUX Question (p28) B




▼ What is MapReduce?

a. A multi stage process to break up then analyse data.

b. No longer used, Google has found an alternative.

c. Owned by Apache.

d. A way to make maps smaller.

---

# Lecture 7 - Resource and Standards

- Open Data (p21-23)
  - A common format for open data is "Linked Open Data (LOD)"
  - Aim of Linked Open Data (LOD) is to make data accessible, machine readable and self-describing
  - objects given a URI
  - relationships between two objects represented as a triple: (subject, verb, object)
  - relation itself is another URI
  - data has an open license for use

▼ FLUX Question (p24) C

# FLUX Question

Graph database is commonly used to store...?

A. Structured data
B. Open data
C. Linked open data
D. None of the above options

Intro. to Data Science, © Wray Buntine, 2015-2019 Slide 24

- Data Wrangling (p26-28)
  - What is data wrangling
    - Process of transforming “raw” data into data that can be analysed to generate valid actionable results and insights
  - Why wrangling
    - Data comes in all shapes and sizes
    - Different files have different formatting
    - Mistakes in data entries
  - Examples of wrangling (p28,30,31)
    - extract the core news text, title, and date from a webpage
    - extract the text plus details from a PDF file
    - extract all article titles from an XML file
    - digitise the text from a scanned image
    - extract all the sentences referring to particular individual in an article
    - integrate data sources
    - geocoding
    - convert free text dates to standard format
    - recognise missing values and deal with them
      - removing the row or columns
      - replace with a special "unknown" value
      - replace with an average value
      - or doing nothing
    - deal with outliers or "illegal" values

- discretise the data into a set of values

▼ What does data wrangling mean?

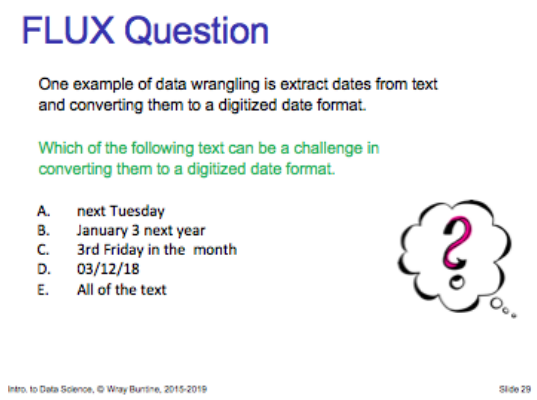
a. To capture data from sensors.

b. To tidy up or clean data.

c. To brand or own data.

d. To steal data.

▼ FLUX Question (p29) C



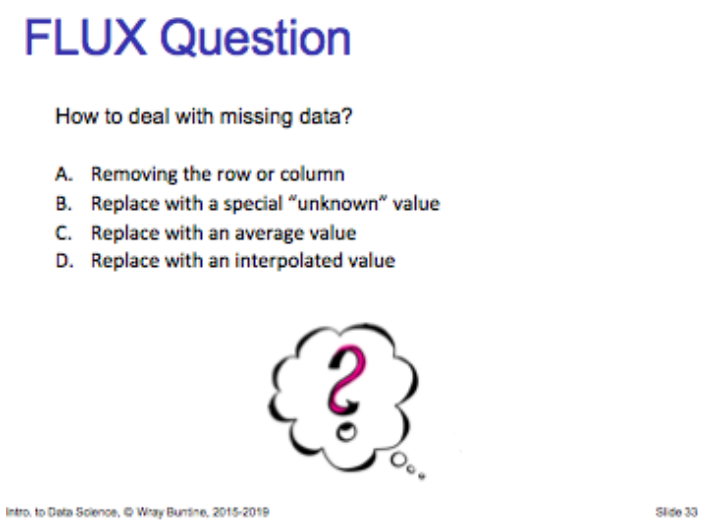

▼ What would be the output if the following patient data was imported, cleaned up and then saved to .csv? 001M11/11/1998 88140 80 10

a. 001M11/11/1998 88140 80 10

b. 001,M,11/11/1998, 88,140, 80, 1,0

c. 001,M,11/11/1998,88,140,80,1,0

▼ FLUX Question (p33) ABCD




▼ Example Standards

- Metadata standards
- XML formats
- Standards for describing the data mining/science process

    - Standard vocabularies for use in Medicine
- Semi-Structured Data (p39)

▼ Which of the following are ‘machine-readable data’?

a. XML

b. All options are correct.

c. JSON

d. RDF

▼ A vector of ages data was saved to file in the following format: {"Age {"0":39,"1":28,"2":44,"3":25,"4":32,"5":33,"6":31,"7":26,"8":22,"9":25,"10":28}} What format is this?

a. RDF

b. JSON

c. XML

d. CSV

- Model Language (p40)
    - 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)

▼ If 5GB is approximately one movie then 500TB is:

a. 500 movies.

b. 1,000 movies.

c. A Gigabyte of movies.

d. 100,000 movies.

▼ What is the correct order, in decreasing size of the following units of digital information?

a. Mega, giga, tera, peta.

b. Exa, zetta, peta, tera, giga, mega.

c. Exa, peta, tera, giga, mega.

d. Exa, peta, zetta, tera, giga, mega.

---

# Lecture 8 - Resources Case Studies

▼ FLUX Question (p6) C

## FLUX Question

Which of the following statement is FALSE?

A. PMML is a standard language for describing a predictive model
B. Semi-structured data is data that is presented in XML and JSON
C. JSON is easier to read than YAML

Intro. to Data Science, © Wray Buntine, 2015-2019 Slide 6

▼ FLUX Question (p7) C

## FLUX Question

A vector of ages data was saved to file in the following format:

{"Age":{"0":39,"1":28,"2":44,"3":25,"4":32,"5":33,"6":31,"7":26,"8":22,"9":25,"10":28}}

What format is this?

A. RDF
B. XML
C. JSON
D. CSV

Intro. to Data Science, © Wray Buntine, 2015-2019 Slide 7

▼ File system vs Database (p11)

- file processing system: a collection of programs that store and manage files in computer hard-disk
  - more data redundancy
  - less flexibility in accessing data
  - doesn't provide data consistency
  - less complex
- database: a collection of programs that enables to create and maintain a database
  - less data redundancy
  - more flexibility in accessing data
  - provides data consistency through normalisation
  - more complex

▼ Popular Open Source Projects (p12)

1. Apache Hadoop Distributed File System (HDFS)
2. Apache Hadoop YARN (resource management system)
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 (data flow for Hive and Pig)

▼ REST API Terminology (p28)

- API: Application Programming Interface
- REST: REpresentational State Transfer
- SaaS: Software as a Service

▼ API and Saas Examples

▼ APIs:

- Facebook API
- Twitter API
- LinkedIn API
- Google Maps API
- ...
-

▼ SaaS:

- Email systems
- File sharing systems
- Business systems

▼ Why Saas and disadvantage (p34)

- Pay as you go
- Scale up/down
- Low maintenance
- Better performance
- `Disadvantage: data privacy`

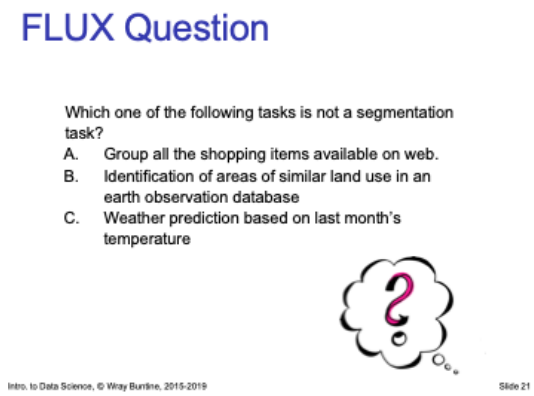
---

# Lecture 9 - Data Analysis Theory

▼ Models of the structural aspects of data analysis problems:

- simple prediction (aka classification/regression) task: Is used to predict an unknown value on the basis of a number of know feature values

    - more complicated prediction task: It contains many variables that link to one another complicated ways, many of variables are unknown, different patients might have different knowns.
    - segmentation (aka clustering) task: Used to identify customer segments, customers are grouped into segments. A segmentation model is a graphical model where the cluster variable is unknown, called latent and the cluster variable identifies the segments.
    - time series forecasting and sequential learning tasks: Used to predict the next value in a series based on the previous value form the same series.
    - causal inference task
- Simple Prediction task (p12-13)
- Complicated Prediction Task (p14-15)
- Segmentation Task (p16,20)
- ▼ FLUX Question (p21) C




- Time Series Forecasting (p22-24)
- Sequential Learning Task (p25)
- Causal Model
- Truth (p28)
- ▼ Quality (p32-33)
    - loss: positive when things are bad, negative (or zero) when they're good
    - gain: positive when things are good, negative when they're not
    - error: measure of "miss", sometimes a distance, but not a measure of quality. Quality is a function of error.
- Linear Regression (p34)

$$\hat{y}(X;\vec{a}) = a_0 + a_1 X$$

$$MSE_{train} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}(X_i;\vec{a}) - y_i)^2$$

- ▼ FLUX Question (p35) D

## FLUX Question

A 'good fit' linear regression line can be achieved by:

A. Drawing a line backwards.
B. Drawing a line through data.
C. Drawing a line through the middle of data.
D. Finding a line through data such that the distance from that line to all of the points is minimised.



▼ What is the purpose of linear regression?

a. To draw a line through data.

b. To fit a model to data.

c. To draw a line through the middle of data.

d. To fit data to a model.

- Polynomial Regression (p36)

$$\hat{y}(X;\vec{a}) = a_0 + a_1 X + a_2 X^2 + ... + a_9 X^9 + a_n X^n = \sum_{i=0}^{n} a_i X^i$$

$$MSE_{train} = \frac{1}{N}\sum_{i=1} N(\hat{y}(x_i;\vec{a}) - y_i)^2$$

▼ FLUX Question (p37) B

## FLUX Question

What is a polynomial regression?

A. Fitting many lines to data.
B. Fitting a curve defined by a polynomial function to data.
C. Fitting a curve to a line.



- More data improves the fit & loss decreases with training data

▼ What rule can you deduce from the following data?

Age, Loan

20, No

21, Yes

22, No

a. Not much, more columns of data may help.

b. Older people don't get loans.

c. Younger people get loans.

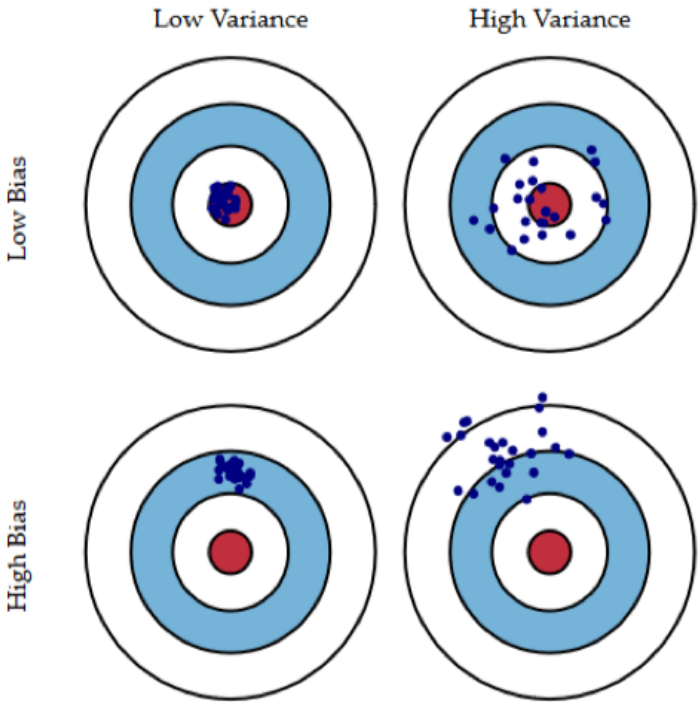
d. Not much, more rows of data may help.

- What is learning curves (p39)
  - sample size against error
- What is overfitting (p40-41)

> the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably – wikipedia

▼ Bias and Variance (p42-44)

- Bias: what is the least error one can get when fitting any possible model to the data (impracticle to achieve).
- 算法的期望预测与真实结果的偏差，刻画了算法本身的拟合能力
- Variance: what is the average error one gets for different data sets over and above the minimum error.
- 训练集变动所导致的学习性能的变化，刻画了数据扰动所造成的影响




- Training Set and Test Set (p45)
- Noise (p46)
- No Free Lunch Theorem: `There is no universally good machine learning algorithm`
- Ensembles (p51-52)

▼ y = ax + b is a function describing a line:

a. Where a is the intercept and b is the incline.

b. Where b is the intercept and a is the incline.

c. Can’t say without a and b.

▼ In a discussion forum, off-topic posts and spam are regarded as an example of:

a. No-free Lunch Theorem.

b. Ensembles

c. Correlation versus causation.

d. Signal versus noise.

---

# Lecture 10 - Data Analysis Process

- Dependence (p6-8)
  - independence: event A and event B are independent if knowing whether A occurred provides no information about whether B will occur or not
- Correlation (p9)
  - dependence in continuous variables
  - statistical notion of correlation usually measures the “linear” dependence between variables

▼ Which of the following are examples of correlation?

a. All options are correct.

b. There is a spill on the floor and a broken bottle.

c. In a supermarket, people who bought chips also bought beer.

d. In a supermarket, people who bought beer also bought chips.

- Causality (p10)

▼ Which of the following are examples of causality?

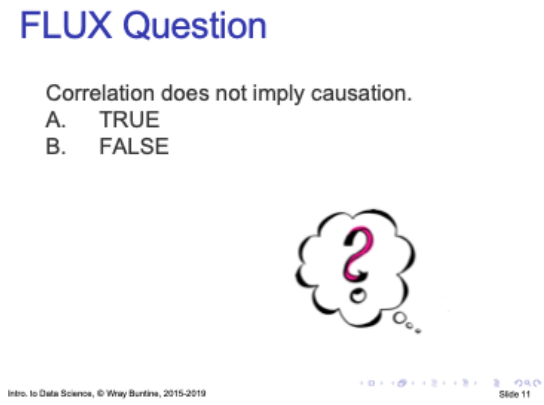
a. A shopper dropped a beer bottle on the floor and it smashed.

b. There is a spill on the floor, a broken bottle of beer and only one shopper near by, they must have caused the mess.

c. In a supermarket, shoppers who bought beer also bought chips.

d. All options are correct.

▼ FLUX Question (p11) A




- Imputation (p15)
- ▼ Characterizing Learning (p17)
  - Prediction: Is the task a simple prediction?
  - Dynamic: Does the task repeat over space or time? (GPS, game playing)
  - Missing Data: Do some of the variables have missing data? (note they cannot be 100% missing)
  - Latent variables: Are there latent variables? e.g., a segmentation task. Note the target variable for a prediction task cannot be latent.
  - Optimisation: Does evaluation/prediction require optimisation after statistical inference (i.e. after prediction)?
- ▼ Types of Data Analysis (p18)
  - Descriptive: quantitatively describe data
  - Exploratory: explore relationships between variables
  - Inferential: infer values of unknown variables
  - Predictive: predict future values
  - Causal: determine if a causal relationship exists
  - Mechanistic: explain causal relationships
- Data Analysis tools (p22-27)
  - ▼ Common Software:
    - access: SQL, Hadoop, MS SQL Server, PIG, Spark
    - wrangling: common scripting languages (Python, Perl)
    - visualisation: Tableau, Matlab, Javascript+D3.js
    - statistical analysis: Weka, SAS, R
    - multi-purpose: Python, R, SAS, KNIME, RapidMiner
    - cloud-based: Azure ML (Microsoft), AWS ML (Amazon)
  - ▼ Machine Learning Platforms:
    - Amazon Web Services ML
    - Google Cloud ML
    - Scikit-Learn

- The R Project
- TensorFlow
- Apache Mahout

▼ Rapid Prototyping

- software development for data science projects is often (almost) one-off ... get the results, but ensure it is reproducible
- not standard software engineering, not “waterfall model”, not “agile”
- little requirements analysis
- the results are tested, not the software and its full capability
- development speed and agility are important
- hence use of scripting languages

▼ examples:

- putting together a processing pipeline
- testing out different alternatives
- trying "cheap hacks" for data cleaning to test ideas before investing more effort
- glueing in custom software that might be difficult or particular to use
- running what is usually GUI or internet API systems in command line mode
- modifying/restarting a processing pipeline

▼ Why was the Google Flu Study was criticised? (p31-33)

a. Google’s search engine and its query logs are constantly changing so it makes the sources unclear.

b. All options are correct.

c. Google apparently did not use standard time series models.

d. Because of the proprietary nature of Google’s search engine’s internals, scientific reproducibility is very difficult.

- cut-off (p48)

▼ confusion matrix (p49-52)

- true positive rate: (aka sensitivity or recall)

$$TP = \frac{true\ positives}{true\ positives + false\ negatives}$$

- false positive rate: (aka fall-out or 1-specificity)

$$FP = \frac{false\ positive}{false\ positive + true\ negative}$$

- sensitivity & specificity
- precision & recall

- ROC curves (p52)

▼ What is hard (p56)

- Model fitting: core statistics/machine learning not usually hard (e.g., many use R as a black box for this)
- Data collection: can be critical sometimes, but often more routine
- Data cleaning: can be a lot of work, but often more routine
- Problem definition: getting into the application and understanding the real problem can be hard
- Evaluation: what is measured? should multiple evaluations be done? can be hard
- Ambiguity and uncertainty: invariably these occur and we need to live with them; can be hard

- Decision / Regression Tree

---

# Lecture 11 - Issues in Data Management

▼ Definition of Terminologies (p7)

- Privacy: having control over how one shares oneself.
- Confidentiality: information privacy, how information about an individual is treated and shared.
- Security: the protection of data, preventing it from being improperly used
- Ethics: the moral handling of data.
- Implicit data: is not explicitly stored but inferred with reasonable precision from available data

▼ Privacy and confidentiality in Data Science:

a. are the same.

b. is about how information about an individual is treated.

c. have different meanings, one is about oneself, the other is about your data.

d. is about having control over how one shares oneself with others.

▼ Four Political camps in the big data world (p10)

1. Corporations: want to use data for business advantages; (opposing consumers)
2. Security conscience: concerned with freedom, liberty, mass surveillance; (opposing intelligence organisations)
3. Open data: want open accessibility, support FOI requests; (opposing security experts concerns with leaks)
4. Big data and civil rights: concerned about big data and citizens; (opposing data brokers selling consumer data)

▼ There are four political camps in data science

a. Actually there are five because one of them can be split into two groups.

b. Is one of the Laws of Data Science.

c. Is just an opinion.

d. Actually there are only two, liberal and conservative.

▼ Data Governance (p16)

- ethics,confidentiality
- security
- consolidation and quality-assurance
- persistence
- regulatory compliance
- organisation policy compliance
- organisation business outcomes

▼ Data Governance:

a. is a made up word, there's a lot of that.

b. means the authority, control and shared decision making over the management of data assets.

c. means the same as Data Management.

d. means who owns the data.

▼ Malicious Use of AI/ML & Mitigation (p21)

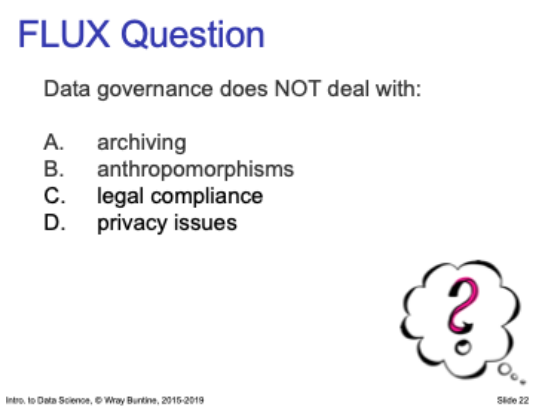
▼ Malicious use:

- faking digital media
- faking interactions (phone calls, teleconferences)
- cyber-attacks taking over autonomous vehicles
- spoofing autonomous weapons systems

▼ Mitigations:

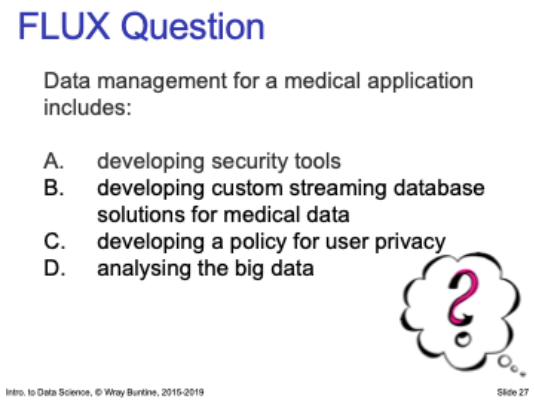
- formal verification, exploring vulnerabilities
- effective licensing of technologies
- education, norms and standards in data science andAI
- policies

▼ FLUX Question (p22) B



- Data Management (p24)
  - 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.

▼ FLUX Question (p27)




▼ Data Management:

a. is the same as data analysis.

b. is the same as data curation.

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

d. is the same as data governance.

▼ Data management project frameworks in Medicine and Health are best based on:

a. the Government framework.

b. are their own unique area.

c. a combination of Government and Science frameworks.

d. a combination of Business/Organisational and Science frameworks.

---

# Introduction to Python for Data Science

▼ Basic Syntax

- Compute mathematical expresions
- Define variables and assign values
- Define a list
  - `B = [1,2,3,4]`
- Access a list (slice)
  - `b = B[1:3] B[1:-1]`

[start:end:step] start包括，end不包括，step默认1，可正可负

start和end都可省略，负数代表倒着数

▼ Load Libraries

- `import matplotlib.pylab as plt`
- `from matplotlib import pylab as plt`

▼ Pandas & DataFrame

- Define a dataframe
- Print a dataframe
- Select a column
  - `df['columnName']`
  - `df.columnName`
  - `df.loc[:,'columnName']`
  - `df[['cn1','cn2']]`
- Select rows
  - `df.loc[2]`
  - `df.loc[[1,2]]`
  - `df.loc[[3:5]]`
  - `df.loc[df['name']=='Amy']`
- Load & Save data
  - `import pandas as pd`
  - `df = pd.read_csv('input.csv',sep=' ')`
  - `df.to_csv('output.csv')`
- Aggregation and GroupBy
  - `df['Mark'].sum()`
  - `df['Mark'].mean()`
  - `df.groupby('Name')['Mark'].mean()`
  - `df.groupby(['Name','ID'])['Mark1', 'Mark2'].mean()`
- Advanced aggregation

▼ Plot Data

- Basic Operations
  - `plt.show()`
  - `plt.plot(df.col_name)`
- Histograms
  - `df.col_name.hist(bins=200)`
- Boxplots
  - `df.boxplot(column='col_name')`
- Bar Charts
  - `plt.bar((1,2,3),df['col_name'])`

▼ Linear Regression

```
from scipy.stats import linregress
slope, intercept, r_value, p_value, std_err = linregress(df['Age'], df['Runs'])
line = [slope*xi + intercept for xi in df['Age']]
```

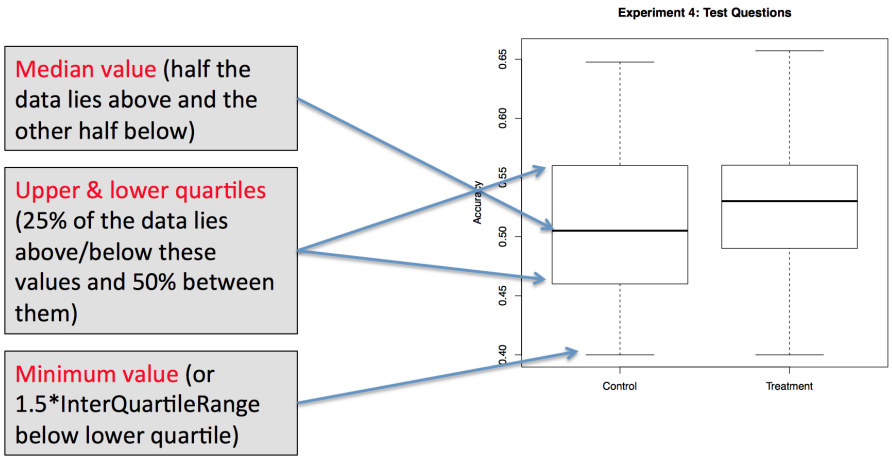
# Introduction to R for Data Science

- Basic Syntax
  - Compute mathematica expressions
  - Define variables and assign values
    - `A ← 10`
    - `A = 10`
  - Define a vector
    - `B ← c(1,2,3)`
  - Concatenate vectors
    - `B ← c(B,c(1,2))`
- Load Libraries
  - `install.packages("ggplot2")`
  - `library(ggplot2)`
- Data Frames
  - Select columns
    - `my_table['names']`
    - `my_table[1]`
    - `my_table[c(1,3)]`
    - `my_table$ages`
  - Select rows
    - `my_table[1,]`
    - `my_table[2:4,]`
  - Load & Save data
    - `my_data ← read.table('my_data.csv')`
    - `my_data ← read.csv('my_data.csv')`
    - `write.csv(my_data,file='my_data.csv')`
  - others
    - `str(my_table)`
    - `head(my_table)`
    - `tail(my_table)`
    - `min(my_table$ages)`
    - `mean(my_table$height)`
    - `sd(my_table$height)`

- `summary(my_table$height)`

▼ Plot Data

- `hist(X)`
- `boxplot(Accuracy~Group,data=my_data)`




- `plot(heights~ages,data=my_table,col='red')`

▼ Fit a linear model

- `fit ← lm(height~age,data=my_data)`
- `summary(fit)`
- `abline(fit,col='red')`

▼ Difference from Python

- start from 1 vs 0
- `A ^ 2` vs `A ** 2`
- `b = B[1:3]` [start,end],end也包括

# Introduction to Shell Command for Data Science

▼ Navigation

- `cd`
- `cd ..`
- `ls`
- `cp`
- `mv`

▼ Reading a Text File

`less myfile.txt`

▼ Some Useful commands

- `wc -l myfile.txt`

- `grep -o -w -c 'elephant' myfile.txt`
- `head myfile.txt`
- `tail myfile.txt`
- `cat myfile.txt`
- `sort -rnk 1,3 -t, myfile.txt`
- pipe & redirection

- awk

---

# Introduction to SAS for Data Science

▼ Load Data

- `data patients;` data类似dataframe
- `infile "/home/abc/def/patient.txt";` 读取

```
input
@1 patno $3. 从第一位开始，类型为string，长度为3
@4 gender $1.
@5 visit mmddyy10. 从第五位开始，类型为mmddyy型日期，长度为10
```

- `format visit ddmmyy10.;` 格式化
- `run;` 运行

▼ Print Data

```
proc print data=patients;
where patno = 123456;
run;
```

▼ Freq

```
proc freq data=patients;
table gender / options;
run;
```

▼ Plotting

- scatterplot `proc sgplot data=patients;run;`

---

# Short Answer Questions

▼ Consider the definition given for data science. Keep in mind that the boundary between data science, data engineering and data analysis is somewhat fluid. What is the importance of the definition?

Definition is like a mission statement (1). So educators have something to focus on (1), and employers know they have the right job descriptions (1), etc. It is fluid though ,... there is no hard and fast answer unless the American Association of Data Scientists becomes a government mandated monopoly. The definitions are gradually coalescing as industry and practitioners find out what works. Importantly, distinctions need to be made

with "near" communities like business analytics and data engineering. There are big overlaps in job roles and functions in practice though.

▼ Briefly review the two slides on the car industry. Note that first they underwent a digitisation process, followed by a datafication process. Give two other non-automotive industries that have had similar developments in recent decades. How do you expect this to change these industries?

***aircraft engine manufacturers***: more effective data on lifetime of engine, better support for maintenance and design, better scheduling of maintenance supermarkets: tracking of products means better knowledge of inventory and lost items; tracking users (around store, as well as purchases) means placement of items, better marketing, better targeted sales.

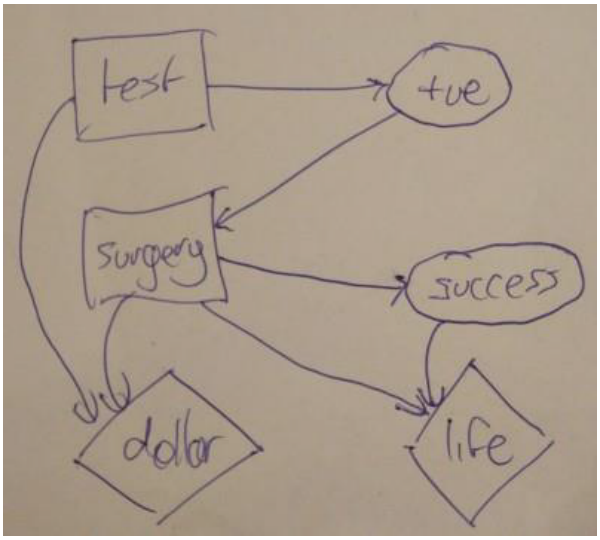
***software vendors*** (Microsoft and many others now "phone home" with data): while we know its used for tracking HCI issues, and maintenance updates, there are probably reasons concerned with modifying/optimising billing practices, going to SaaS etc.

▼ What role has the internet had in the development of data science?

This is all over the various initial lectures: internet giants were the first big users of data science (1), making it visible to the broader business community; internet-driven social networks and commerce sites provided rich varieties of data for use (1); open data sources, tools and training all up on the internet; really the internet was the incubator for data science (1).

▼ Your GP says you may have cancer X, so she sends you to an Oncologist. The Oncologist says the options are you can (1) have a test done, which will cost $X, and it has particular proportions of false positives and false negatives. Moreover you can (2) have surgery which costs $Y. The surgery has a particular success rate P but always lead to harm H (e.g., 90% success rate, but you loose part of your lower colon), and on failure the cancer will continue and usually lead to death. Note if you never had cancer, the surgery is always a success, but you incur the harm.

1. Draw the influence diagram where there are two value nodes, dollar cost and life cost, plus other nodes.



Notes: "+ve" in the figure is shorthand for "positive test outcome". No arc from "+ve" to "life" since positive is test outcome and has no relationship to truth of cancer; "success" indicates cancer is gone (though, honestly we don't know, and wont know until maybe one year later!); "surgery"' impacts life because a body part may be chopped out

2. Describe a situation (e.g., P is low, etc.) where, rationally, you would decide to get both the test and optionally the surgery done.

   You always consider both test and surgery if the test has low error rates and little harm (1) and the surgery has a high success rate (1). In this case the test is informative and the surgery is a good bet.

3. Describe a situation where, rationally, you would decide to get neither the test nor the surgery done.

   You take neither test or surgery if the test has high error rates (1) and the surgery has a very high harm (1). In this case you gamble that maybe you don't have cancer.

▼ Consider the video by Foster Provost we saw in Lecture 3. Now look at the NIST analysis on page 8 of the slides for Lecture 4. Name two issues that Foster touched on and why they were relevant to his application.

- data sources: they have account and financial transaction data; note they also used a lot of other features that probably had been extracted from external sources like socioeconomic data
- data variety: bringing in the fine-grained data of credit card transactions allowed the predictions to improve; note it made the prediction more complex;
- data volume: they were getting this for up to 1 million consumers so we can guess data is in the 10-100Gb (i.e., 10-100k per consumer)

▼ Consider the NYT article, "Lord Mayor's Geek Squad" discussed in this week's lecture. Name two issues from the NIST analysis touched on and why they are relevant to this application.

- data sources: the breadth of data available from the city, mostly relational tables
- data volume: most datasets seem to be quite small: Mbs but maybe Gbs for parking data
- analytics: NYC contains no one single application, rather whatever people can do, so many different kinds of analytics could be done

▼ Give two examples of Data Science applications specifically in retail (selling mass produced goods to consumers).

several in marketing: cross-sell for websites; sentiment analysis to ascertain what consumers like; customer segmentation to support advertising campaigns

▼ Give an example of data with a problem with veracity, and discuss the problems it causes. Remember, well understood measurement error (e.g., only recording to 3 decimal places) is not usually considered to be veracity.

Generally systematical or predictable errors are not considered wrongful data. They are those tempered with or containing distorted signals. E.g., ozone hole

example in NASA; in medical, wrong diagnostic codes by human error; diet data is intrinsically errorful; in industries, such as instrument break downs.

▼ Explain how Koomey's Law has affected data science.

Koomey's Law means instruments can be used to gather data in low energy context. In the 2000s that led to the rise of mobile smart phones which helped kick off the social web that creates a lot of the data used in data science. Second, it has more recently led to the rise of the internet of things, where data can be sourced from many different small devices: small cameras around the city, standard household devices giving their status, instrumentation embedded in industrial plants.

▼ Give a recent/future example of Bell's Law, and discuss the kind of new data it provides.

Internet of things is one example, where data can be sourced from many different small devices. This can provide status data on devices as well as monitoring data of devices, temperature in the fridge, engine runtime characteristics, etc.

▼ Present two different kinds of tasks: (a) one which is purely ("embarassingly") data parallel, i.e., very suited for Hadoop, and (b) one which would be very hard to make data parallel. For each, explain why it works or doesn't.

Standard examples for (a) are turning document collection into an inverted index; typical image/doc processing applications (where each is done independently), such as "find faces"; looking for common 3-4 word phrases in a collection; and for (b) are optimisation or global tasks such as compute the top principle component(s) of a large graph, invert a large matrix, design the layout for a computer chip based on a computerised circuit diagram.

▼ Consider a graph database, such as DBpedia. Give an example of a commercial or government application that would use a graph database, and discuss why it is appropriate.

New York Times keeps graphDB linking news articles to people, places and events to support the fact checkers and writers, best way for them to access things; Google has a similar graphDB, much larger, though we are unsure of usage and how it relates to data science

▼ Give an example where two very different data sets needed to be combined in order to make a data science project work.

Many listed in first part of talk: web pharmacovigilance (2nd source is FAERS, combined gives better results); traffic prediction (weather, events, historical traffic)

▼ Define what proxy data is, and give an example of its use.

Term used by paleoclimatologists but means data that is used as a substitute for an unmeasurable variable. We expect it to be highly correlated. Example is (1) search query data about health to indicate the incidence of the terms being queried, (2) crash data on intersections as proxy for "danger to cyclists"

▼ Give an example of a data standard that is in use: name the standard, give its domain, the sort of data it supports, and why it is used.

See week7 slide36

▼ Name a popular Apache open source product for big data (but not Hadoop or HDFS), (A) briefly describe what it does, and (B) give a short use case for a data science project.

See week8 slide12

▼ Name a popular data/information API, (A) briefly describe what is does, and (B) give a short use case for a data science project

See week8 slide30

▼ Figure-Eight provides human-in-the-loop tools and resources for dataset generation and staffing. Find their Human-in-the-loop FAQ and research it briefly. Then: (A) why would one use human-in-the-loop methods to build a training set and (B) give a short use case

NOTE: this could not be in the exam unless we listed the relevant parts of their FAQ somewhere. Now predictive modelling requires high quality data, and the more the better. So if there isn't the data available, one needs to acquire it somehow, and one way to do it is to get labels from available experts. Also, existing data may be poorly labelled, in which case you would like the labels checked. Now one cannot have experts label some things like which customers defaulted on their bank loan, but they usually can label things like images, properties of text data, and so forth. For use cases see their "Success stories" web page.

▼ Peter Norvig talked about the "unreasonable effectiveness of data". What does he mean by that?

This comes from the term "unreasonable effectiveness of mathematics" which is claimed of Physics and Engineering where a smaller number of mathematical principles are the basis for large parts of theory. This is not the case in areas like health, bioinformatics, sociology, economics where Norvig claims instead data can be used to answer many questions.

▼ Give an example of "customer segmentation" from industry. Describe the kind of data used to build the segmentation, and what the segmentation is used for.

(Should be a realistic example ... doesn't have to be confirmed truth). A search engine or internet company (Yahoo, Yandex, ...) might "bucket" their users into 500 different groups, based on their web interactions (search queries, advertisement responses, purchases). This is then used as a "grouping" to support which advertisements and news will be show to the user on their home page or during other activities. So now, each group will have its own advertisement and news predictions, instead of the company having to create predictions personally for every single user.

▼ Describe what bias and variance are.

See week9 slide42

▼ Describe what ensembles are.

See week9 slide51

▼ What is a clinical trial and why are they used? Give an example.

They are experiments ("trials") done in medicine ("clinical research") to answer a specific question about a treatment. Usually they follow the scientific method: subjects are randomly selected for "treatment" or placebo, and the treatment is done blind so that subjects cannot know which they get.

At the end of the trial results are recorded and analysed. Example: have 1000 elderly men take daily aspirin or placebo for a year, and record their blood pressure and cholesterol before and after to answer the question, "does daily aspirin support cardiac health".

▼ Give an example of a Machine Learning Platform and explain why it is used.

(Note we didn't ask you about a specific platform, you could pick the one you knew) TensorFlow is an open source library for ease of computational deployment of numeric algorithms to GPUs and CPUs, originally developed by Google. It is used by machine learning researchers wanting to port their numerical algorithm onto GPUs without them requiring specific knowledge of the hardware and its coding.

▼ What is a scripting language, and what is their relationship to rapid prototyping?

See week10 slide25. Scripting languages are ideal for rapid prototyping, so are often used for it.

▼ What are the problems with Google Flu trends?

See week10 slide31-33. The Google team used specific search queries as proxy data for the existence of flu. These were not disclosed, so the system wasn't open. This also becomes very subject to public perception: news reports could make people go searching for flu even if they didn't have it. Also flu and flu-like diseases are hard to distinguish by the non-expert. Finally, Google didn't use standard time series approaches which would have used the CDCs reported data as well in the prediction.

▼ Describe two kinds of problems that can arise with the application of the scientific method and give examples.

See week10 slide35

---

# Sample Exams

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/89bbf34d-0081-4433-ab07-6b4881cb30d9/FIT5145_Sample_Exam.pdf

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/7d41df1d-a77a-4424-b4f8-8c979b97ed2d/FIT5145_Solutions_to_Sample_Exam.pdf

---

# Alexsandria Complementary

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/89be3ed1-c867-4ce4-9dab-b4c9187296d3/Alexandria_Notes_FIT5145.pdf