FIT5197M1: Introduction to Modelling for Data Science

Asef Nazari

FIT5197M1: Introduction to Modelling for Data Science

Asef Nazari

Generated by Alexandria (https://www.alexandriarepository.org) on March 11, 2017 at 10:52 am AEDT

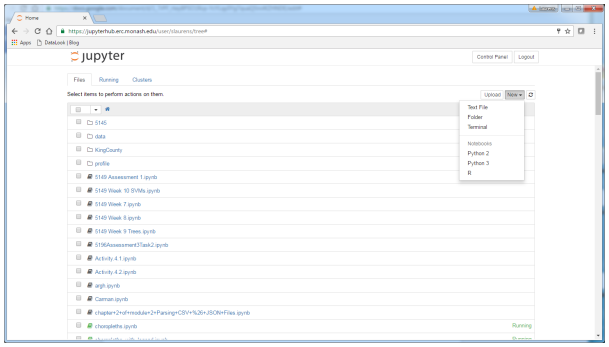
Contents

1
R Programming 1

Monash Online JupyterHub - How To Access and Get Setup

Monash Online's JHub Server is a multi-user online JuPyteR notebook server, setup and hosted by Monash University. Access is with your standard Monash login (authcate), i.e. username (not email) & password. You can access the server through this link **https://jupyterhub.erc.monash.edu/**

JuPyteR Notebooks allow you to create and share documents that contain code, equations (latex), visualisations (typically, as output of running a piece of code), HTML and explanatory text (they used to be called iPython Notebooks, hence the file extension, .ipynb, which has remained). The name JuPyteR is from the programming languages Julia, Python and R but there are many more language options now; JuPyteR Notebooks are 'language agnostic'.



To the above is an example of JuPyterHub's 'home' page. Yours won't have any notebooks until you create them or upload them (see menus 'Upload' and 'New' at right. Current language options are Python 2, Python 3 and 'R'). You can also upload data sets and create folders (you may find this useful for creating different folders for different units as you study).

JuPyterHub

- authenticates users and launches separate Jupyter notebook servers for each user.
- is cross-platform as it runs in your web browser.
- provides the same environment for all users as it will have the same language versions & libraries.
- can be updated easily (i.e. software versions and libraries/packages)

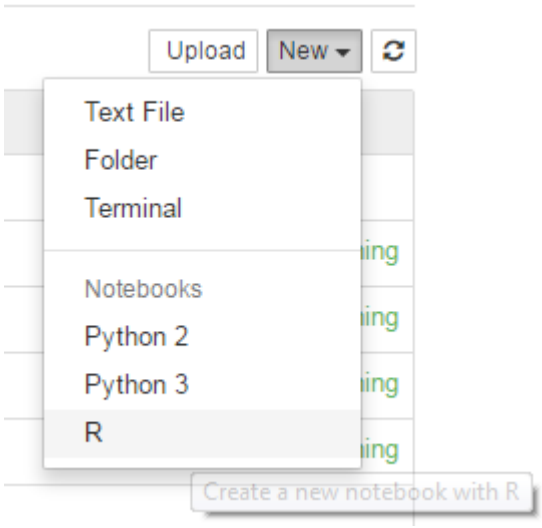
Currently available Software

- Python2
- Python3
- R (and SPARK)
- as well as a range of standard libraries & packages.

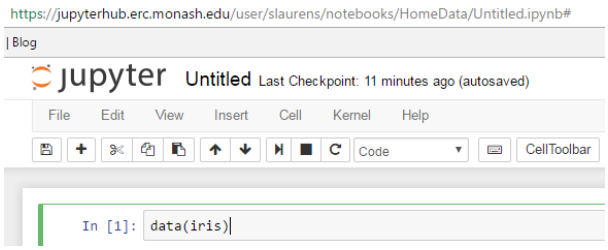
Your first JuPyteR notebook

One. Login to **https://jupyterhub.erc.monash.edu/**

Two. Create a new notebook ('R' in this case) using the 'New' menu then 'R' (as below)



Three. To see:



Four. The highlighted green cell, 'In [ ]:' is ready for some R code, let's load some data, inspect it, then plot it. As shown above, type:

```
data(iris)
```

Then Shift-ENTER to run the code (or use the menus), to see... nothing!!
Actually, no error is good, the data has been loaded.

Five. Now let's look at the data; Type:

```
head(iris)
```

Then run the code using Shift-ENTER to see a table of data:

```
In [1]: data(iris)
```

```
In [2]: head(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

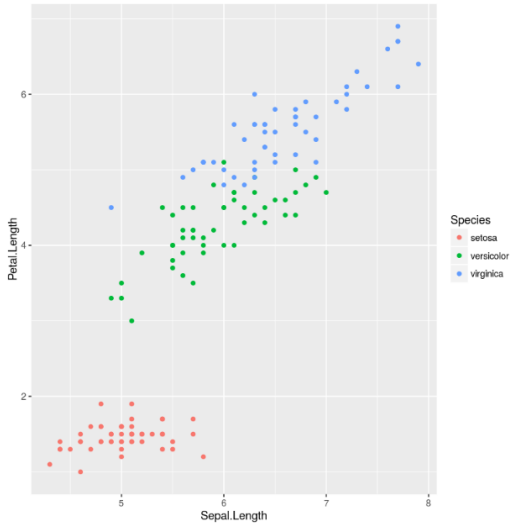
It's flower data, Irises.

Six. Load a plotting library (ggplot2), and plot the data:

```
require(ggplot2)
qplot(Sepal.Length, Petal.Length, data = iris,
                    color = Species)
```

Although it is generally possible to use library(ggplot2), but here just use require(ggplot2).

```
In [3]: require(ggplot2)
Loading required package: ggplot2
In [4]: qplot(Sepal.Length, Petal.Length, data = iris, color = Species)
```



Seven. Save your notebook (you might like to name it too, click on 'Untitled' at top)

If you are not familiar with JuPyteR notebooks the following online tutorial is a good place to get started with JuPyteR notebooks:

- Jupyter Notebook Users Manual (https://athena.brynmawr.edu/jupyter/hub/dblank/public/Jupyter%20Notebook%20Users%20Manual.ipynb) which shows how to view and execute computer programs and to create executable documents or documents with visualisation.

We recommend using JuPyteRHub for GDDS units that require coding in Python or R (also SPARK). You can also setup JuPyteR notebooks on your local (home) machine, see:

https://www.alexandriarepository.org/module/appendix-b-setting-up-your-programming-environment/

Programming in R

In this section, we provide a very introductory programming in R, which will help you in accomplishing all the assessments in this unit. However, it is really important to have your R programming skill highly developed. I suggest using this guide as a start point. Search for each topic on the Internet and find more materials and examples. Never limit yourself to the materials here. For each section, you can download related iPython notebook (it is just the name, we have nothing to do with Python programming in this unit) and read the instructions and do the commands. Also, there are R files of each section, which contains only the R commands. It does not have explanations. That's why the zeroth section does not have an R file. So, download the notebook, upload it to your jupyterhub directory, open it, read it and do it.

For each section, there are three files. The notebook contains a zipped ipython notebook. You need to download it, unzip it, and then upload it into your jupyterhub home. The second one, r-file, is R commands only. There are not any explanations that were provided in the notebooks. This is for people who are interested in learning through Rstudio. To have access to the explanation, I uploaded a pdf version of the notebooks to read the explanations.

- **Section 0: How to Use Markdown [notebook (https://www.alexandriarepository.org/wp-content/uploads/20170227091110/R0000-MarkDownPractice.ipynb.zip), pdf (https://www.alexandriarepository.org/wp-content/uploads/20170227182217/R0000-MarkDownPractice.pdf)]**
- **Section 1: The First Touch of R [notebook (https://www.alexandriarepository.org/wp-content/uploads/20170227091208/R0001.ipynb.zip), r-file (https://www.alexandriarepository.org/wp-content/uploads/20170227091254/R0001.r), pdf (https://www.alexandriarepository.org/wp-content/uploads/20170227182302/R0001.pdf)]**

1. Using R as a Calculator
2. Assignments
3. Managing Variables
    1. Finding variables ls()
    2. Removing variables rm()

- **Section 2: Data Types [notebook (https://www.alexandriarepository.org/wp-content/uploads/20170227091439/R0002.ipynb.zip), r-file (https://www.alexandriarepository.org/wp-content/uploads/20170227091521/R00021.r), pdf (https://www.alexandriarepository.org/wp-content/uploads/20170227182336/R0002.pdf)]**
    1. Main data Classes
    2. Vectors
    3. Lists
    4. Numbers
    5. Changing class of variables
    6. Factors
    7. Missing Values
    8. Sub-setting
    9. Vectorised Operations
- **Section 3: Data Tables [notebook (https://www.alexandriarepository.org/wp-content/uploads/20170227091624/R0003.ipynb.zip), r-file (https://www.alexandriarepository.org/wp-content/uploads/20170227091644/R0003.r), pdf (https://www.alexandriarepository.org/wp-content/uploads/20170227182526/R0003.pdf)]**
    - Matrices
    - Dataframes

- Reading and writing data in R
- Managing your files
- Built-in datasets
- Packages
- Frequently used functions

- **Section 4: Controlling the Execution flow [notebook (https://www.alexandriarepository.org/wp-content/uploads/20170227091754/R0004.ipynb.zip), r-file (https://www.alexandriarepository.org/wp-content/uploads/20170227091811/R0004.r), pdf (https://www.alexandriarepository.org/wp-content/uploads/20170227182611/R0004.pdf)]**
    1. Logical expressions
    2. Control structure loops
        1. If structure
        2. For structure
        3. While structure

2
Data Science and Models

This subject (and the Data Science course in general) is primarily about training you up in the technology, at the centre of the circle below, although if you're specialising in one of the applied sciences, it will be of interest as well.

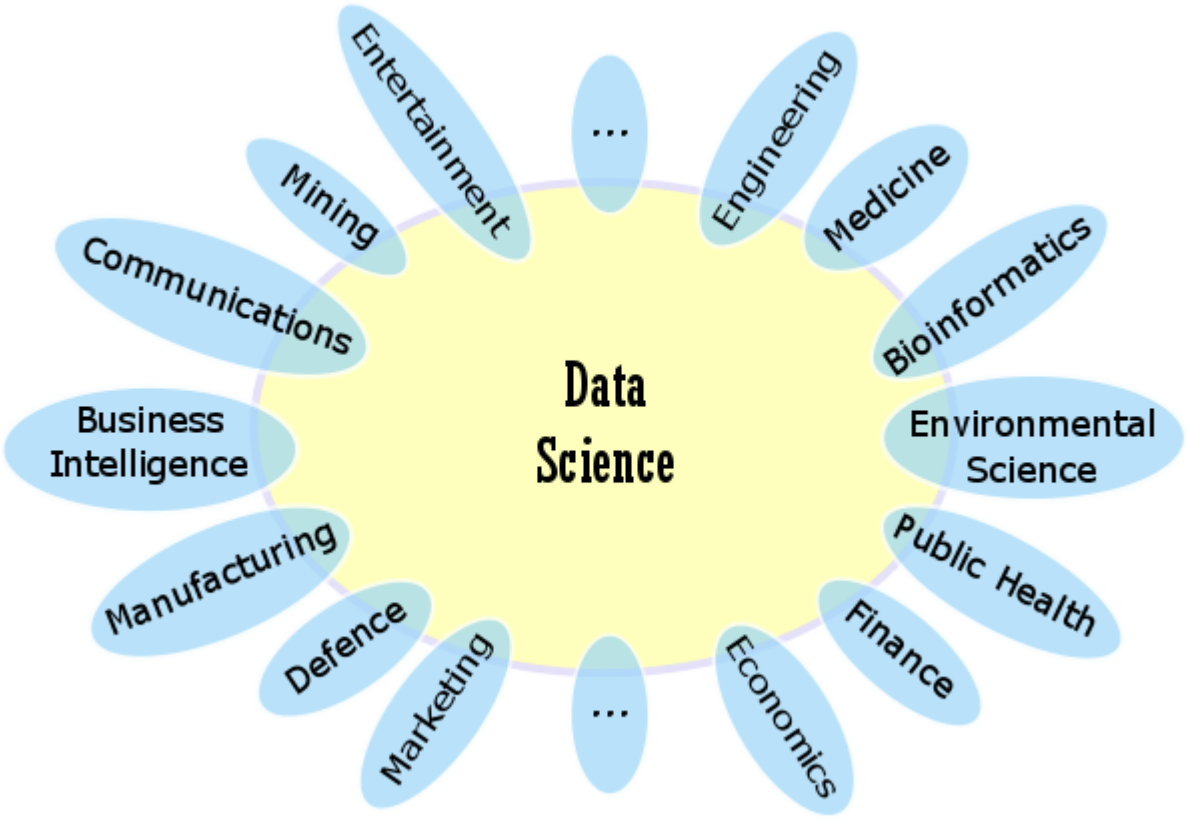


Diagram showing the many applications of data science.

Models are basic to science and so, in particular, to data science. Models typically are used to represent scientific theories: to instantiate them in some form that makes them easier to study or understand. Scale models (and maps, for modelling geography) are perhaps the best-known examples. Mathematical models of scientific theories encode those theories in formulas that can then be manipulated (in proofs) in order to better understand the consequences of those theories. They are often crucial for deriving empirical predictions from a theory and a set of initial conditions, in something like the hypothetico-deductive method [wikipedia] (https://en.wikipedia.org/wiki/Hypothetico-deductive_model) of testing scientific theories.

Scientific method

We can understand science as a version of a general method for learning about the world, described by an OODA loop [wikipedia] (Observe-Orient-Decide-Act) (https://en.wikipedia.org/wiki/OODA_loop). The OODA loop was developed as a model to represent military decision making, where a combatant monitors its environment (observes), figures out what's happening in that environment (orients), decides what to do (decides), and then acts on the decision (acts), resulting in changes to the environment and so new observations. While the OODA loop was recently invented, the very same ideas about rational decision making can be traced all the way back to Aristotle.

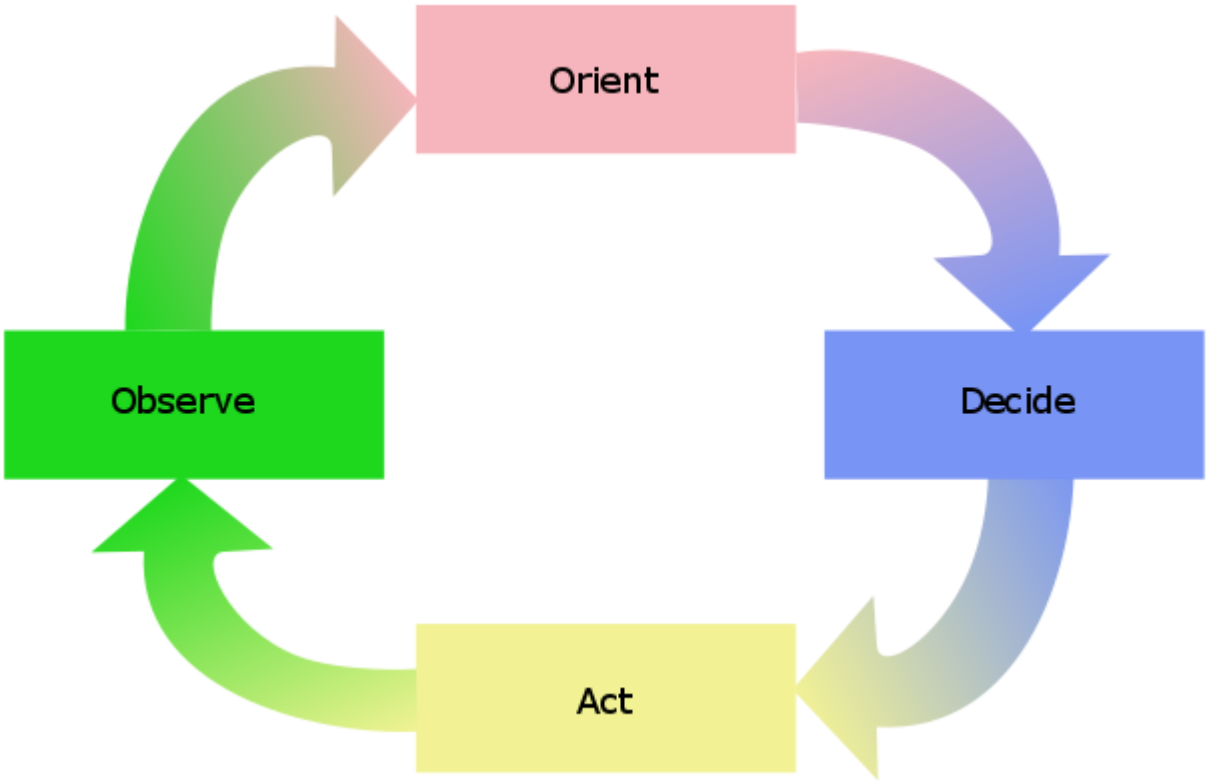
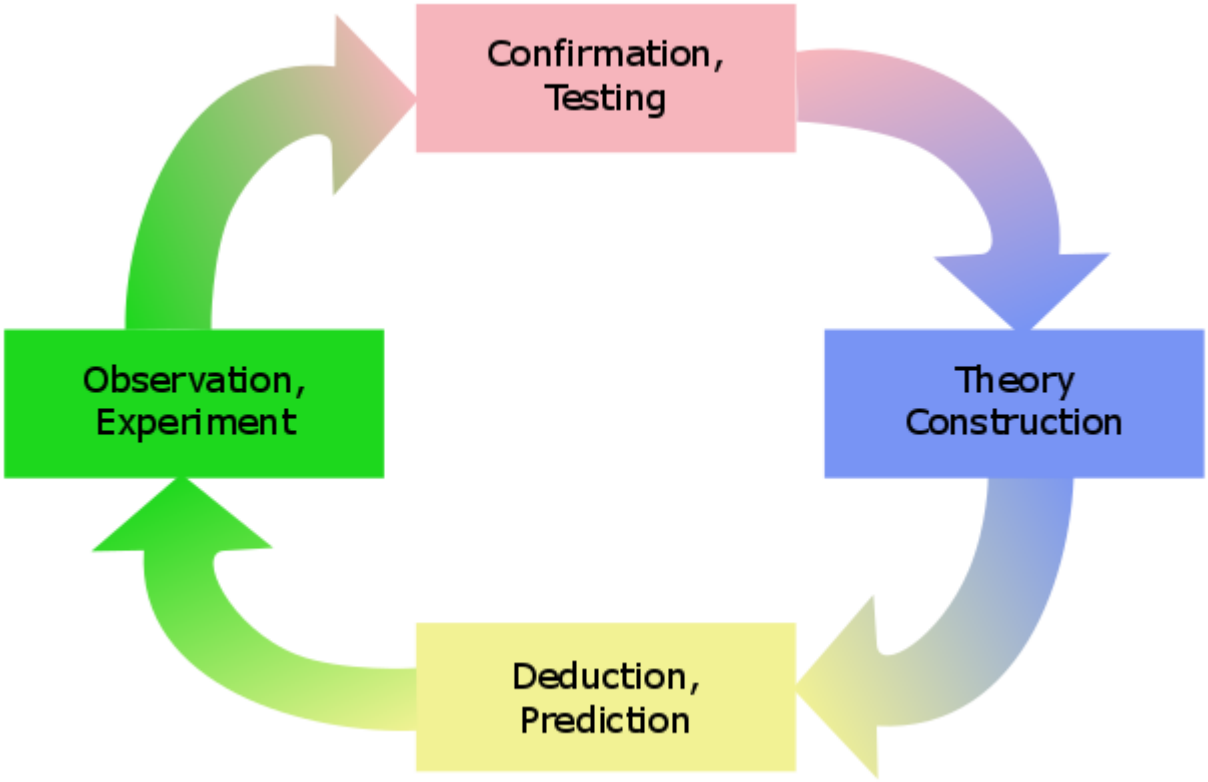


Diagram illustrating the OODA loop.

As a general model for scientific learning we could adapt this to: observing nature and/or the outcome of experiments (observe); interpreting those observations and, in particular, judging whether they confirm or disconfirm our prior beliefs about nature (orient); invent new or alter existing theories to accommodate these interpretations (decide); deduce consequences of the new theory, making predictions and then observing or experimenting with nature (act) to see if those predictions come true (observe).

From the following OODA model, we can see that the role of theories/models is that of representing what we think we have learned from the data and providing the means of directing further testing and data gathering.



Computational models

In this subject, we will be primarily concerned with computational models. These are models which can be implemented as computer programs. They include computer simulations, which are programs that aim to reproduce or mimic processes in the world, such as the formation of galaxies, economic development, the spread of bush fires, or the spread of epidemics. Simulations come in a very large variety of types, including, but hardly limited to, agent-based models, Ising models and other cellular automata, weather and climate models, Artificial Life, discrete-event simulation, and, of special relevance here, stochastic simulation [wikipedia] (https://en.wikipedia.org/wiki/Stochastic_simulation) (i.e., sampling from a probability distribution). These and other kinds of simulation can be studied in computational science, artificial intelligence and other courses. Here, however, we will focus on probabilistic computational models, models which can be automatically learned from sample data and which in turn (via stochastic simulation) can be used to generate similar sample data to that from which they are learned.

Probabilistic models can be thought of as mathematical models since they can be thought of as computer implementations of probability distributions. Usually, however, the mathematics corresponding to any probability distributions we may learn from data mining is far too complex to write down on paper and deal with in a traditional analysis. We shall instead rely on computer programs to do the detailed mathematical work for us, which is, of course, precisely where computers excel beyond human abilities.

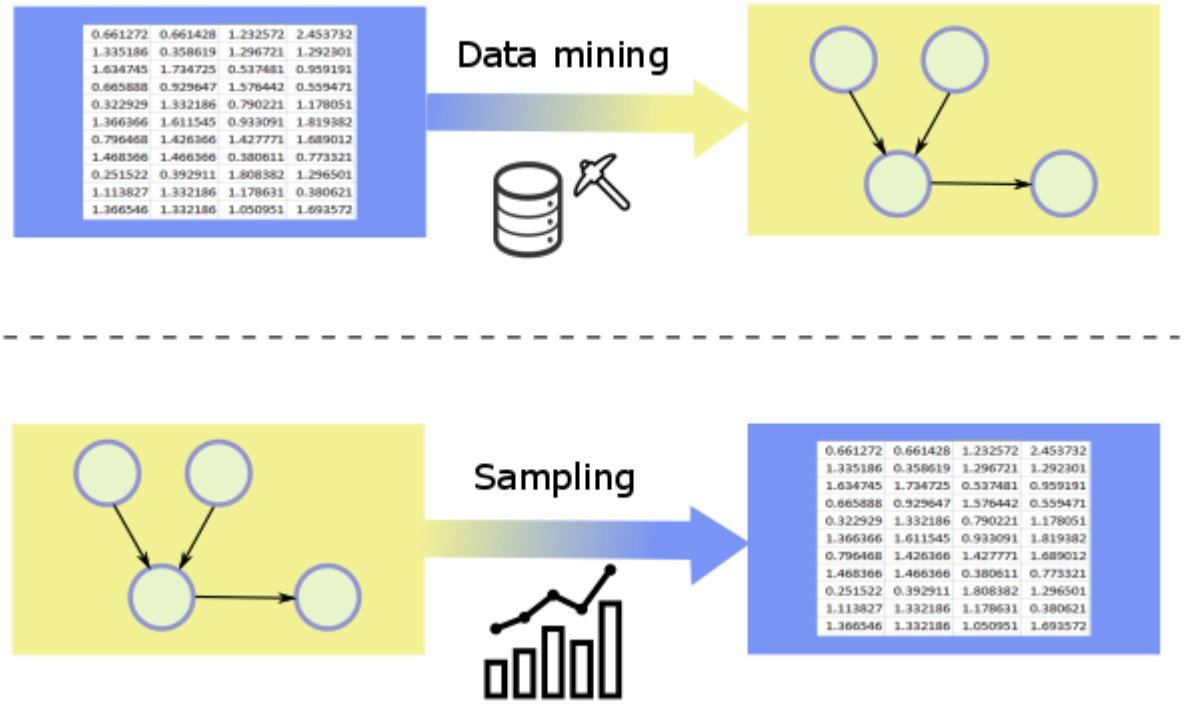


Diagram showing how data mining generates models from data; sampling generates (artificial) data from models.

Key concepts

- **Theory:** A set of statements describing the nature and operation of some system in the world. These statements may or may not be formalised in mathematics.
- **Model:** A representation of some system in the world. These usually correspond to a theory about that system which is either accepted or is under test. A model can often be "operated" so as to simulate the system and its change from an initial state to a subsequent state.
- **Prediction:** The use of a theory or model to explore what the system of interest will do under real or hypothetical conditions. Given an initial state of a model, any future state that follows that can be observed in a corresponding real-world system would constitute a prediction. To test whether the prediction is correct, we would need to put some real-world system into the model's initial state and operate it until the future state predicted by the model should occur.
- **Explanation:** Accepted theories or models can be used to explain specific facts, in a similar way in which they predict specific facts. The differences from prediction are that: 1) the theory or model being used is assumed to be true; 2) the fact is known or observed, as are the initial conditions required for the model or theory to lead to the fact in question. Scientific explanations, then, simply show why or how the known fact should have been expected given the theory or model and its initial conditions.
- **Decision:** We decide how to act based in part on our predictions. E.g., if we opt for surgery, what the likely outcomes are. But also we decide based on our values. E.g., what's the value of being cured of a disease, versus encountering side-effects of surgery, versus carrying on without any surgery. In general, we prefer actions that maximise the expected value of the results. Such decision making depends upon using our best predictive models to understand the consequences of our available actions.

3
Data Collection and Sampling

The title of a famous textbook in Statistics which is "Statistics, the art and science of learning from data" clearly express the importance of data. Data is the essence of modelling, and actually, models are prepared to describe data in an elaborated manner. Data are obtained from experiments and are the result of measuring some characteristics or property of objects. These objects are usually human or things. The outcomes of measurements are sometimes called observations. As the characteristic of the interest could be different for different objects under experiment, these characteristics or properties are called variables.

All variables are categorised as *numerical* (quantitative) or *categorical* (qualitative). Numerical variables are those that can be measured on a numerical scale or counted. Categorical data take values that are nonnumerical in nature. It is meaningless to do arithmetic on categorical data. For example, postcodes look numerical, but there is no meaning to add them up or take their average. Therefore, postcodes are treated as categorical data. Categorical data could only be classified into categories, levels or classes. When we decide to use a particular tool to summarise data or graphically represent them, we need to know whether the variables are numerical or categorical.

The set of all the objects of interest is called the *population*. Populations are not generally available to study, or it is very difficult and costly to access them. For example, suppose that you want to measure the height of all Australians to find the average height! The cost issues force the government to have a census every five or ten years in some countries. In a census, we actually trying to capture all the population. Instead of accessing the whole population, we might take a *sample*, a smaller subset of the population, do conduct the study over the sample. Based on the information we get from the sample, we make *inferences* about the population. However, we need to be very careful in choosing a sample. It should be a good representative of the population.

A good sample presents similar characteristics to the population represented. To have a good sample, we need to make sure that every individual in the population has the same chance to be selected in the sample. A sample obtained by this process is called a random sample. Generally, people use random number generators to automatically generate random samples. We will cover this in depth later.

4
Experiment Design and Causality

In the real world, scientists are confronted with several questions, and the first step in answering those questions is to gather information. The questions could cover different aspects of life and business. For example, we would be interested to discover the effect of a newly designed medicine. Or, we want to predict the number of sale for the next year to be able to have enough inventory space for orders. To gather the information, we might perform experiments and surveys. The observations collected through experiments or surveys are called *data*. Before the experiment, we need to find out the best ways of collecting, analysing and making conclusions from data.

The essence of an experiment design is taking a random sample from a population. The aim is to make sure that the sample is a good representative of the population. Hence, we can make inferences about the population based on what we learn from the sample. If there is not enough randomness in choosing sample points, say someone is going to choose for us, it is possible that the sample could be skewed to that person's interest, which may be unintentional. This would introduce bias into a sample. Also, we need to be careful with other issues in sampling including non-response bias (https://en.wikipehttps://en.wikipedia.org/wiki/Non-response_biasdia.org/wiki/Non-response_bias) (answers of non-respondents would be significantly different from those of respondents) and convenience sampling (https://en.wikipedia.org/wiki/Convenience_sampling) (the sample is selected for convenience reasons).

In a statistical activity, we might distinguish between *explanatory variables* and *response variables*. Generally, explanatory variables might affect response variables in a relationship. In some cases, it is not easy to distinguish between them. Generally, you are told which one of the variables is a response variable. However, naming the variables as explanatory or response does not mean that the relationship between them is causal, one causing the other.

Two main types of data collection are *observational studies* and *experiments*. In an observational study, researchers collect data as they appear, and they do not directly interfere how the data is generated. These kinds of studies could provide an association between variables, and they cannot show a causal connection. To show a causal connection, researchers conduct an experiment. For instance, we may suspect the efficacy of a particular medicine on curing a disease. To check if there is a causal connection between the explanatory variable and the response variable, researchers would collect a sample of individuals and split them into two groups. When individuals are randomly assigned into groups, the experiment is called a *randomised experiment*. Then, the first group receive the "real" medicine and are usually called the *treatment group*, and the other group receive a placebo (fake medicine) and are usually called the *control group*. The causal relationship can only be established by a randomised experiment such as this by the process known as *intervention*

In order to identify a causal association, we need to perform comparisons between groups. Through the experiment design, we assign some people to a *treatment group*, who got the treatment, and a control group of others. We then can compare the outcome of these group to establish an association. In establishinintervention causal wikipediag causality, we need to be very careful with *confounding factors*. Confounding factors could have impacts on the all other factors under investigation, and create a fake association. Through the randomisation process, our aim is to remove all confounding factors. Also, we might do an experiment as blind or double blind. In a blind experiment, the individuals do not know whether they are in the treatment group or the control group. In a double-blind experiment, even the people executing the experiment do not know which individual is in the treatment or the control group. There are some evidence that the knowledge of nurses has some impact on the outcome if they are aware of the treatment and the control groups.

5
Data Pre-processing

In this section, you will be introduced to the earliest phases of data analysis: getting to know your data. To be sure, there is an even earlier phase, which is getting to know your domain. For example, if you are working on a data science problem in medicine, then you might want to start out by learning something about the medical domain in question. But exploratory data analysis is the first phase from the point of view of the program in data science. The major components of data exploration are addressed in:

- preprocessing (data wrangling)
- descriptive statistics
- visualization

Data preprocessing

Data preprocessing (also known as "wrangling") is the preparation of data for analysis with data mining or visualisation tools. There are many problems which can interfere with a successful analysis; some of them can be readily addressed with simple preprocessing techniques, which we will explore here. Some of the problems are less readily addressed, either in general or when they are represented in some extreme form in the data.

Many problems can be avoided by an early planning of data collection. If you can anticipate that a study of customer satisfaction will need customer income levels, for example, then in organising a survey you can arrange to ask about income, whereas without anticipating this you may never think to ask and end up data poor. Or, again, if you can anticipate that many people are reluctant to truthfully report their income, you can arrange to check for their accuracy in self-reporting by recording data that can be used to cross check for under-reporting income, such as postcode (comparing with average incomes per postcode), automobile ownership, job, etc. As Benjamin Franklin said, an ounce of prevention is worth a pound of cure. Of course, many analysts have no say in the original collection of data and are simply handed a dataset, warts and all. Then there are only two options: (1) wrangling with the data to reduce or eliminate problems; (2) reporting on the problems and how to avoid them in future data collection.

Watch Ross Gayler (Analytics Specialist at Connected Analytics) share his thoughts on deep learning, credit scoring and the future of data science and data analysis.

(https://youtu.be/nBuhsXPAROU)

Common Data Problems and Some Preprocessing Solutions

- **Format:** Perhaps the most common problem is just that the data available and the analytical tool you'd like to use require different formats. There's nothing special to say here; you just need to reformat your data. An editor with macros or the ability to program in a scripting language will likely be of help. In the case of combining data from distinct databases, there may be specialised tools to help.
- **Noise:** All real data are noisy; that is, there are always sources of inaccuracies in any real data collected. Measurement instruments are never perfect. Aside from that, what is being measured will typically be impacted in many different ways, so that the variable one is attempting to measure is itself fluctuating. The result may be a systematic error, called *bias*, which we look at below. The result may be that a value is simply missing because the measuring device altogether fails on an occasion. For example, in a survey, a respondent may report everything else but simply omit her income. This results in a missing value.

A common problem is simply recording error. It may be that the errors can be readily detected and corrected. For example, if a survey field for gender reads "G" and the first name is "Sally", then it is an easy inference that "G" was mistakenly entered for "F". Whether cleansing the data for all such errors is easy or not will depend on the size of the data set and the nature of the errors. It might be possible to simplify data cleansing by, say, automating checks for simple errors or paying for extra personnel to code data twice, providing an easy source of checks for errors.

Typically some measurement error is added to any measurement. For example, a temperature may be measured as 17.4°C when it is really 17.3°C, simply because the measuring device is accurate only to tenths of degrees (i.e., variations in it and its environment have a standard deviation of 0.05°C or more). This last kind of data problem is best addressed by improving measurement accuracy during data collection. There's little that can be done in data wrangling to improve matters, although you might be able to estimate the amount of variation introduced during measurements. For example, you might find a guaranteed source for data and then measure repeatedly to sample how much noise (and, perhaps, bias) is present. At least that way you'll see how much of a noise problem you have. The final method of dealing with noise is to model it, or eliminate it via modelling

techniques; we'll treat that option in Module 6.

- **Missing values:** Dealing with missing values is a significant part of preprocessing. There are three very common ways of treating a missing value problem. (1) Deleting rows that contain missing values; (2) modal imputation (replacing the missing value by its most common value in the sample); (3) mean imputation (replacing the missing value by the mean value in the sample). The first is simplest, but also the most problematic. Consider the reluctance of wealthy people to report their incomes. If we simply delete those rows, then we are left with data that under-represents wealthy people, resulting in a bias that's likely to pollute any subsequent analysis. Therefore, the latter two options are usually preferred, even if the first remains common just because it's simple. Imputation just means to attribute a value where one is missing based upon whichever average is preferred. Weka [website] (http://www.cs.waikato.ac.nz/ml/weka/) (a free data mining platform), for example, does modal imputation to replace missing values by default. But there can be problems with this as well. In one memorable case of search-and-rescue data, for example, one binary variable had 1% True values, 2% False values and 97% missing values. Modal imputation resulted in a distribution of 1% True values and 99% False. This didn't help us understand the variable much.

Other techniques are more advanced statistical inference methods with names like "expectation maximisation" (EM) and "Gibbs sampling". They seek to estimate a distribution of values for missing values of any particular variable based upon the rest of the data available. These inference methods tend to be better than the simple imputations above, but they typically rely on the simplifying assumption that the "missingness" of a particular variable is unrelated (probabilistically independent) to the actual values of other variables for the sample. This is called missing completely at random [Wikipedia] (https://en.wikipedia.org/wiki/Missing_data#Missing_completely_at_random), and means, when a variable is missing it provides no useful information. This may well be far from the truth. To pick on our standard example in this section, if someone owns a recent model Mercedes Benz and fails to report any income, those two facts are highly unlikely to be independent (*i.e.*, how can they and why would they buy an expensive car if their income is zero?). So treating them as independent will likely result in an underestimated income. As a second example, if a questionnaire asks your religion but only lists 4 common religions as optional answers, then failing to answer probably indicates you are of an alternative religion. That said, given a tool that performs more advanced statistical inference for missing values (and many do just that), then it can be a quick and satisfactory preprocessing repair.

Possibly the most general approach to dealing with missing values (short of going back and getting them recorded!) is to just add a "missing" or "NA" value to the variable and treat it as just another value that the variable takes. The result of doing that is that constructing a model via data mining will now model the "missingness": a predictive model, for example, should be able to tell you under what conditions, say, the income variable goes missing, or doesn't go missing.

- **Missing Variables:** These are extreme forms of missing values, where 100% of the values are missing. It might seem like nothing can be done about this (except measuring the missing variable for the next round of analysis), but there are statistical methods to find candidates for missing variables and to measure their influence on measured variables. In fact, there is a whole subdiscipline in statistics dedicated to that project, called latent variable analysis (latent variable = hidden variable = missing variable). See, for example, *Latent Variable Analysis* [google books] (https://books.google.com.au/books?id=mNMPAQAAMAAJ&q=latent+variables&dq=latent+variables&hl=en&sa=X&ved=0CDEQ6AEwA2oVChMIov2ypOnrxgIVYximCh0YTgTd) by von Eye and Clogg for a standard textbook. Nevertheless, there is nothing cut-and-dry about latent variable analysis, so this is more likely to be considered data analysis proper than data preprocessing. However, I give a very brief overview of some options for it here.

One option that sometimes presents itself is to find an alternative source of values for a known missing variable. For an example, when working with some GPS tracking data in one study, we had time, date, location, direction, etc. What we didn't have, but could use, were meteorological data,

such as wind speed, direction, etc. However, the Bureau of Meteorology makes such historical data available, so we collected it and merged it with our primary data source. (And that was, of course, data preprocessing.)

When working with linear models the most common approach to identifying and estimating the parameters for one or more latent variables is factor analysis. The simplest factor analytic model assumes that the measured variables can be clustered together based upon their probabilistic dependence upon each other (correlations). The interdependencies between variables within a cluster are then assumed to be explained by a common hidden cause, called a factor. This is a common approach, for example, in psychology. It's pretty clear that verbal ability and mathematical ability are correlated with each other. Throw in some puzzle solving and analogical reasoning tasks, followed by a factor analysis, and we have a measure of IQ, the postulated common factor behind all of them. That doesn't mean that the (in)famous factor g [wikipedia] (https://en.wikipedia.org/wiki/G_factor_(psychometrics)) can explain everything about intelligence, or intelligence tests, but that it can (hypothetically) explain what is common to the performance on multiple subtasks within them.

There are many other sophisticated attempts to posit missing variables and their parameters that you'll find if you explore the modelling literature; some of them will come up during further discussions in this unit.

- **Bias:** This refers to an influence in the collection of data which results in systematically inaccurate estimates of population values. An inaccuracy that does not lead to systematically inaccurate estimates is called noise (see above). There are many possible sources of bias, some of the common ones being:
  - *Selection bias*, where members of a population are not selected uniformly randomly. If samples are selected partially for convenience, for example, the result may well be biased with respect to one or more values of interest. On surveys sent out to large classes of people, reporting bias is almost universal: those who respond to surveys are systematically different from those who do not. Another source of selection bias is the exclusion of particular kinds of subjects of a study; e.g., if criminals are excluded from a study, then, while this may be a small percentage, it is likely to be significantly different from the rest of the population in many respects and, so, influencing any socio-economic measurement. A special case of this is self-selection bias; for example, if people who drop out of a medical study (a self-selection process) are more likely to be ill than those who remain, then the medical treatment will likely have its benefits overestimated.
  - *Instrument bias*, when the measurement apparatus being used has some systematic bias. For example, a thermometer may be miscalibrated, so that it always reports one degree C too high.
  - *Statistical bias* refers to problems with estimators being used to making inferences from a sample to the population; it is not a problem with the data itself. Similarly, inductive bias refers to a bias inherent in (or parameterized into) a data mining tool. Such biases will be discussed in connection with specific estimators or tools.

The primary way of dealing with bias is to avoid or correct it. If a selection bias is detected, the sampling procedure is best altered so as to obtain representative samples. This may be impractical, when, say, the data are massive and collected by a procedure outside of your control. In that case, there are alternatives. An attractive one that is often ignored is to estimate the bias and adjust the data for it. If the bias truly is systematic, then it can be estimated accurately, perhaps with a second, smaller representative sample of the population.

- **Outliers:** These are data points that lie well beyond the bulk of samples for a variable on one or more dimensions. If we had no background knowledge, we might decide to delete that sample in

any case. But on occasions, we will be in error in doing so. Many distributions have very long tails (https://en.wikipedia.org/wiki/Long_tail) (that part of a distribution that is way beyond the central hump); the normal distribution has infinitely long tails. So, from time to time, if your data are genuinely normal, it will show an outlier that is a legitimate data point, and clipping it would then be wrong in principle.

Another possibility is that the outlier represents the "same" process, but it is the process itself which is changing. For example, a rapid shift in temperatures might indicate changed climate conditions. Or a rapid change in the size of the ozone hole at the South Pole might indicate the introduction of new chemicals destroying ozone in the upper atmosphere. Simply deleting outliers guarantees a failure to notice such changes. One option that suggests itself is to conduct a dual analysis, one of the data with apparent outliers removed and other with them retained; the comparison of results for the two parallel data sets may suggest sources for the outliers which otherwise might be overlooked.