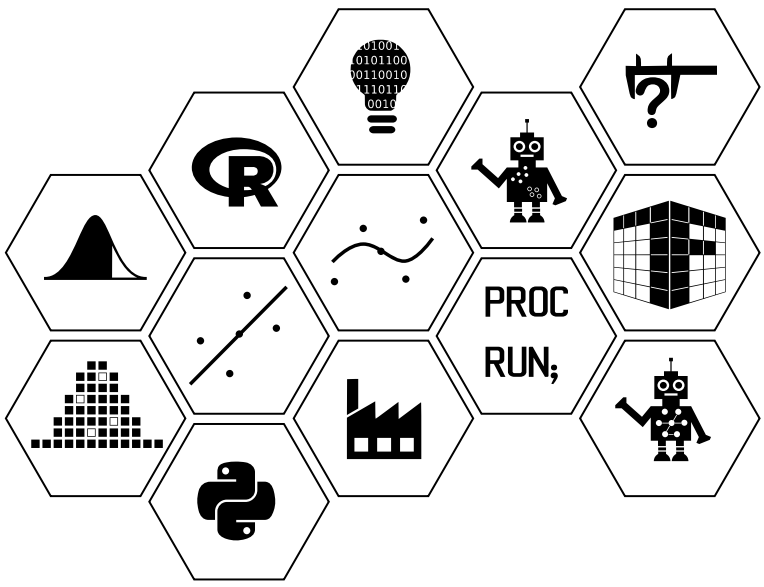
Learning from Data/Data Science Foundations

Week 1: What is Learning from Data?



DATA ANALYTICS
GLASGOW

Learning from data

Week 1 learning material aims

The material in week 1 covers:

- what is learning from data?
- an introduction to different approaches to learning from data;
- data sources, structures and data types;
- collecting and collating data;
- an introduction to approaches for extracting/summarising data.

What is learning from data?

Traditionally, in order to answer specific questions of interest, for example, about demographics, opinions or attitudes of the world's population, business or industry performance, environmental quality or medical advances (to name a few), it was necessary to design a study through which data on the specific question of interest could be collected.

However, these days data are everywhere! We are continuously surrounded by data. Some of which is immediately recognisable as a source of data, and some of which is more subtle e.g. information that can be extracted from tweets, emails or images.

This course will start by introducing different data sources, structures and types, and methods of collecting and collating data, identifying the challenges associated in each case. The course will then introduce tools and techniques to summarise, visualise and describe the data, and to enable us to sensibly extract information, in order to investigate specific questions of interest. While summaries and visualisations of data are important, and essential, tools in *Learning from Data*, we can go much further than this by using advanced and technical analytics approaches to extract information, and hence form conclusions in situations ranging from where we have only a small amount of data available to a big data problem.

The welcome video below provides an introduction to **Learning from Data** and to the course tutors [1].

Video

Learning from Data

Duration 4:16

There are two main advanced analytics approaches to *Learning from Data*, and these are illustrated in the diagram in Figure 1. We can develop a model to describe our data, or we can use algorithms to search our data for structure and pattern. In order to 'learn from data', we can either use ideas from probability theory to develop a model enabling us to describe patterns and relationships and hence draw conclusions (accounting appropriately for any uncertainty that may exist in our estimated results), or we can use algorithmic approaches to mine and search the data for pattern and structure. In this course we will focus on the first of these, and we will use what you have learned from probability theory to introduce and develop, in particular, statistical models to describe our data. We will introduce how we can gain insights from data by assuming that the data arise from a specific probability distribution. These tools and techniques will relate specifically to the following courses (which are also available as individual courses or as part of the full MSc in Data Analytics and MDataGov ODL programmes):

- predictive modelling
- advanced predictive models
- uncertainty assessment & Bayesian computation
- data mining and machine learning I/II

The algorithmic approaches involving machine learning, in particular, are covered in the courses:

- data mining and machine learning I/II
- large scale computing.

The diagram in Figure 1 illustrates the different approaches and provides an indication of where some of these approaches initially appear within our Data Analytics courses:

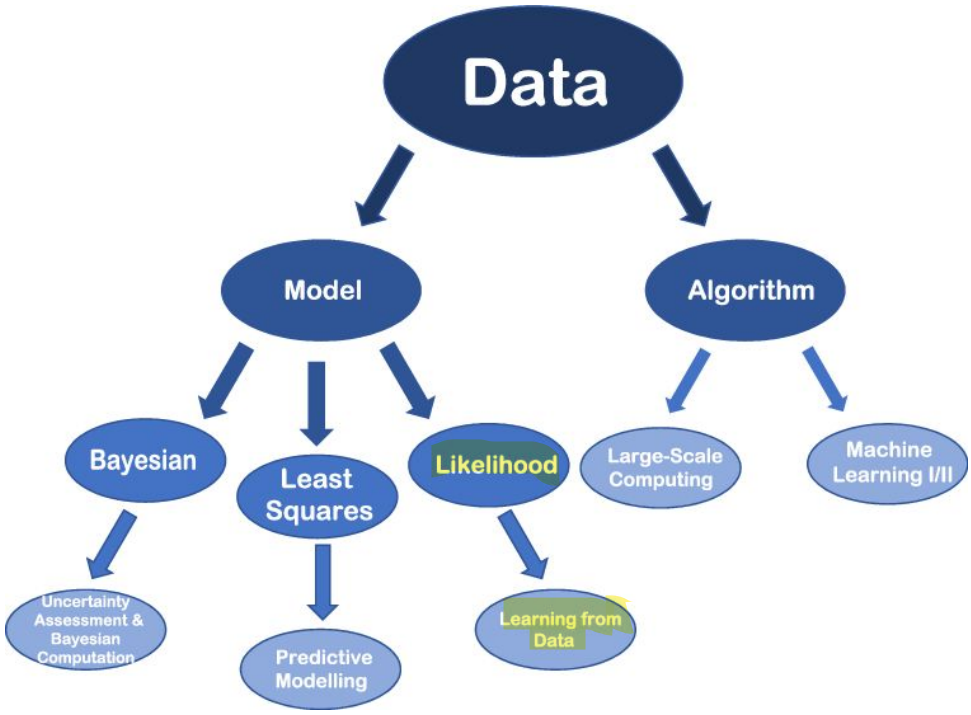


Figure 1: *Approaches to learning from data*

and more background is provided in the second video for this week.

Data sources, structures and terminology

We'll start this course by thinking about what we mean by data. We can regard data as anything that provides us with information. We are usually interested in information that is available, for example, over many different attributes of interest, or on members of the population, or information that is recorded over time or over space, or public opinions on particular topics. Advances in technology have meant that we are surrounded by data, and there are millions of data elements being collected on all aspects of daily life every day. Satellite data, automatic monitoring sensors, computer data analytics, DNA sequencing, citizen science, to name but a few. Alternatively, or additionally, we might design a study in order to collect specific information in order to answer key questions of interest. Regardless of the source of the data, key issues concern: the relevance of the data, the data quality and the associated uncertainty in the data products. Reliable and robust analysis of such data requires both computational approaches for the analysis of such data and key knowledge to extract scientific truth in the presence of uncertainty. These are some of the aspects that we'll consider within this course.

Data sources

Data could appear in a fairly routine or obvious form of, for example:

- collections of numbers about a particular attribute obtained by monitoring or measuring;
- demographic information on individuals;
- opinion polls;

or it could be retained in (and hence obtained from) a source e.g.:

- emails
- tweets
- images
- maps
- newspaper articles.

Task 1

What are the advantages and challenges with using the information sources above to obtain data?

Data types and structures

As illustrated by some of the sources in the previous section, data can be *quantitative* or *qualitative*, i.e. numerical based (quantitative) or textual/image based information (qualitative), and can be available in a *structured* or *unstructured* way. We usually think of structured data as data that are already in a form that are ready for us to directly summarise and hence investigate. For example, data that are recorded systematically in a database or tabular form. Unstructured data may involve data in its raw collection form e.g. data contained in images or information contained in documents that require to be processed before we are in a position to analyse the information and investigate specific questions of interest. It's helpful to define both the data structure and type since this information is important to enable us to identify an appropriate follow-up analysis.

Definition 1

Structured data

Data are highly-organized, stored in tabular, spreadsheet or database form and are formatted in such a way that makes them easily searchable.

Definition 2

Unstructured data

Data that have no pre-determined format or linkage, making it much more difficult to collect or collate, process, and analyze the data.

Examples of unstructured data could be information stored as text, video, audio, mobile activity, social media activity, or facial recognition. Extracting useful information from unstructured data can be complex and highly technical and requires advanced analytical tools.

There is also a sub-category here of **semi-structured** data.

Definition 3

Semi-structured data

This can include data that have been transported and retrieved over the web, for example, emails, tweets, webpages, sensor or satellite data. JavaScript Object Notation (JSON) and eXtensible Mark-up Language (XML) are examples of formats used for storing and transporting data which makes it easily readable. Both can be used to retrieve data from a webserver, are human readable and can be used by various programming langages. Data that have been stored and transported using such formats are easier to visualise, summarise and analyse than in the case of completely unstructured data.

Definition 4

Metadata

This form of data provides information to describe collected or sourced data of interest. For example, an associated spreadsheet, or layer of a data array that contains factual information about the data objects, units of measurement, information on how the data were collected, collated etc.

Definition 5

Synthetic/simulated data

This form of data is artificially created. Synthetic or simulated data usually reflects the statistical properties of another dataset of interest. It can be of interest to help investigate the performance of data science approaches when we know the true characteristics of our data and/or to replace the original data in situations where anonymity is a concern.

The scale and volume of unstructured data is increasing and the articles below give some examples of this, with claims of up to 80% of data available being unstructured.

- Webpage example 1: **Data Management - Solutions Review article**
- Webpage example 2: **IBM article**

Task 2

The following description was taken from the paper 'Exploring methods for identifying related patient safety events using structured and unstructured data', Journal of Biomedical Informatics, (2015). Identify the different data sources within this description and state whether each are structured or unstructured, quantitative or qualitative.

The paper describes patient safety event reports which contain information such as the time and site of occurrence, role of the participants (physician, nurse, technician, etc.), patient demographic and clinical attributes, a classification of the severity and type of event and a free-text field in which the reporter can provide a narrative describing the patient safety event in greater detail.

Data terminology

We will just give a very brief introduction here to some of the terminology used for data 'processes' in data science:

Terminology	Description
data lakes	A central repository that allows you to store and process all of your data, structured or unstructured.
data wrangling	Transforming/manipulating data from one 'raw' data format to another format to aid analysis.
data harvesting	Collecting data from various different (typically online) sources into a repository, 'data lake'. This could involve extracting important data from a source. Also called e.g. web scraping or data extraction. It's the process of extracting data for further analysis, which might include statistics or machine learning.
data warehouse	A central repository for storing structured/cleaned/transformed data from a variety of sources.

Collecting/collating data

You might be in the situation that you already possess data and you are interested in the information that you can extract from the data. However, as we'll see in week 2, you need to be very careful here. Searching for meaning in data without a particular focus can be misleading and could, at its most extreme, result in wrong and dangerous conclusions being made. For any study of interest, we have to consider carefully the question of interest that we wish to address, whether or not we already possess appopriate data and, if not, how we will obtain appropriate data in order to investigate our questions of interest. There are two possibilities here, **primary** data collection and **secondary** data collection.

Primary data

The first is that we need approaches or methods in order to collect and/or collate data of interest. In this scenario we generally obtain data directly from individuals, objects or processes and we refer to this as **primary data**. Such data are usually collected for the purpose of answering a particular research question and can be collected in a way that the data are usable in the form it is collected in. These data are also usually more reliable since we control how the data are collected and can monitor its quality. There are many possible ways to collect primary data which are summarised below.

Surveys

Suppose that we are interested in gathering information and hence data about the **population** of a country. Logistically and financially, it is not possible to contact or speak to every member of the population. Therefore, a survey can be a useful mechanism, and can be implemented in a variety of forms.

A *survey* is a data collection method where a **sample** of respondents are selected from a large population in order to gather information about that population. The process of identifying individuals from the population to contact is known as **sampling**. To gather data through a survey, a questionnaire is generally constructed to prompt information from the sample of respondents. Surveys are frequently used when the subjects are people, and questions are asked of them. There are many methods to administer the questionnaire such as:

- Personal interview: An interviewer asks questions face-to-face with each respondent.
- Telephone interview: An interviewer asks questions to respondents over the phone.
- Online interview: An email is sent inviting respondents to participate in an online survey.
- Mailed questionnaire: A printed questionnaire is sent to the postal address of the respondent.
- Focus groups: A small group of people are identified, and are guided in discussion to identify attitudes and experiences of the group.

Task 3

What are the advantages and disadvantages with each of these methods to administer a questionnaire?

One major shortcoming of surveys occurs if we fail to **sample** from our **population** correctly; if this is the case our sample may not be **representative** of the population and as such may give inaccurate results. For example, if we were interested in average income and our sample of respondents only consisted of retired people.

Another disadvantage with many of these techniques is non-response and inaccurate response which can bias (i.e. our results are not representative of all views of the population, and inclined or prejudiced towards particular groups or opinions) the results by introducing non-sampling errors. This bias can occur regardless of the sampling method used.

We've introduced terminology informally here which will be useful throughout the course of: **sample**, **population**, **bias**. We'll define these terms formally in the weeks to follow but for now it's useful to be aware of the associated ideas.

Supplement 1

For more information on different methods of sampling please see the reading material folder for week 1 on Moodle.

Note: Anyone that has taken the course Probability and Sampling Fundamentals will already be aware of this material.

Direct observation

Another method of data collection is **direct observation** where we measure and observe the attributes (**variables**) of interest ourselves, without changing existing conditions. This could involve collecting data on subjects only once or over a period of time. For example, suppose we are interested in the biodiversity of a forest. One way of collecting data would be to identify and count each tree and therefore directly observe and record the data ourselves. Again, collecting data in this way would require some appropriate sampling scheme - otherwise you would find yourself counting every single tree in the forest.

Experiment

We can also collect data through an **experiment**, where a researcher assigns a *treatment* and observes the *response*. Sometimes, a *control group* (a group receiving no treatment or a *placebo*) may be used to compare the effectiveness of a treatment. Experiments are very popular in the medical field, for example in testing the effect of different drugs on a particular disease. One of the biggest advantages of using an experiment is that you can explore **causal** relationships that the previous methods cannot. They are however expensive, time-consuming and can be unethical.

Secondary data

Another rich source of data are data that are collected/collated after another researcher or agency that initially gathered the data has made it available, this is known as **secondary data**. Examples include census data published by the US Census bureau or stock prices data published by CNN. In this case, the data already exists and may be publicly available for download from a database or it may need to be enriched with data from other sources before it meets your research needs. The use of publicly available data is a far cheaper method of data collecting/collating and can be used to answer many research questions. However, as well as the data potentially not being complete it is also more difficult to verify the accuracy of secondary data.

Collating/extracting data

As well as using recorded information which has been made publicly available, we are sometimes in the situation where we have access to a data resource, for example, data automatically collected on members of the public every time they wish access to a particular wifi network. In such a situation, people often realise that they have a potentially rich data resource, and they want to extract information from the data to see what they can learn. However, a word of caution here, before analysing data it's very important to consider the questions that you would like to answer, as mentioned above, and we'll return to this in week 2.

Other possible sources for gathering data include **data scraping** where a computer programme extracts data from an output (e.g. a website) into a spreadsheet or local file. This is one of the most efficient ways to extract data from the web. Some common uses for data scraping include price comparison websites or conducting market research by trawling public data sources (e.g. twitter).

Data linkage

With data coming from many different sources, being of different types and different structures it can be a challenge to link data together appropriately in order to analyse patterns, look for relationships and answer questions of interest. A variety of approaches are available, including:

- data assimilation (to link collected/collated data with data from mathematical models);
- data fusion to link together data of many different types;
- in geographical data, a geospatial identifier can be used to link data ;
- upscaling/downscaling approaches are used to link data on different scales;
- unique identifiers are used for record linkage in administrative and health data.

Supplement 2

A fuller coverage of the latter is provided, as an example, at this link: **Linking Data for Health Services Research: A Framework and Instructional Guide**

The video below provides an introduction to approaches to *learning from data* and examples of identifying data sources/structures, types and methods of collecting data for the following examples:

- What is the percentage support for the Labour party in the UK population (at a given point in time)?
- What effect will changes in climate have on average surface water quality of a lake?
- What is the probability that tweets have come from the same source?

Video

Approaches and Data Stuctures

Duration 3:59

Summarising data

Once we have identified our questions of interest that we wish to answer, sourced, collected and collated our data, we then need appropriate methods to summarise, visualise and describe the data in such a way as to investigate questions of interest. This will be the main focus of week 2 for *structured data*. However, there are a couple of additional questions that we need to consider initially. For *unstructured data*, how do we extract information from sources to obtain data that we can analyse? For *structured data*, what types of data are we working with?

Unstructured data

The topic of how to extract information from unstructured data, could be a full course (or multiple courses) in itself. Here, we simply provide a very brief introduction to highlight this area and the challenges and complexities involved. A standard approach is to create structured data from unstructured data by extracting information from unstructured data that is then in a form to be visualised and analysed. A couple of examples are:

Example 1

Tweets

Suppose you are interested in analysing tweets to try to identify who sent them. The following website extracts information from tweets and analyses the language in order to try to identify the sender. Analysing the tweets of different users

Example 2

semi-structured data

RESTful API's (Application Programming Interface) send and receive information through a URL interface, allowing communication using HTTP methods. API's allow us to remotely access datasets, and specifically request elements of that data. This is extremely useful for cases when data are updated frequently. Access to API's is often authenticated, requiring users to specify an authentication key. Data can be accessed in R by using the httr package and making direct HTTP requests.

Data obtained from API calls are usually structured in XML or JSON file formats, allowing the data to be structured in a hierarchical format. Packages such as jsonlite in R can parse JSON files for use in R.

The OMDb API is an open web service that hosts movie information. To make a call to this service, the user must specify an API key and the name of the movie of interest.

This example looks to pull information from an API from an online database of movies for a specified movie. The example requires an API key, but the one in the example can be used (it can support 1000 calls a day).

```
# Require the following packages
library(httr)
library(jsonlite)

# Access movie information data from OMDb API

# To obtain access, must obtain API key
# (can be obtained from http://www.omdbapi.com/)
api_key <- "aeb77595"

# Specify the title of which movie to search for (replace
spaces with +)
movie <- "Indiana+Jones+and+the+temple+of+doom"

# Create web path
path <- paste("http://www.omdbapi.com/?
t=",movie,"&apikey=",api_key,sep="")

# Access HTTP data using GET command
call <- GET(url=path)
```

```
# To see if a HTTP request is successful, check the reply
status code (successful code is 200)
status_code(call)
```

R Console

```
[1] 200
```

```
# To view the content of the call, the number of characters is
truncated here for display only.
str(content(call), nchar.max=20) # remove ``nchar.max'' for all
info
```

R Console

```
List of 25
$ Title     : chr "Ind"| __truncated__
$ Year      : chr "1984"
$ Rated     : chr "PG"
$ Released  : chr "23 May 1984"
$ Runtime   : chr "118 min"
$ Genre     : chr "Action, Adventure"
$ Director  : chr "Steven Spielberg"
$ Writer    : chr "Wil"| __truncated__
$ Actors    : chr "Har"| __truncated__
$ Plot      : chr "A s"| __truncated__
$ Language  : chr "Eng"| __truncated__
$ Country   : chr "United States"
$ Awards    : chr "Won"| __truncated__
$ Poster    : chr "htt"| __truncated__
$ Ratings   :List of 3
..$ :List of 2
.. ..$ Source: chr "Int"| __truncated__
.. ..$ Value : chr "7.5/10"
..$ :List of 2
.. ..$ Source: chr "Rotten Tomatoes"
.. ..$ Value : chr "83%"
..$ :List of 2
.. ..$ Source: chr "Metacritic"
.. ..$ Value : chr "57/100"
$ Metascore : chr "57"
$ imdbRating: chr "7.5"
```

```
$ imdbVotes : chr "491,007"
$ imdbID    : chr "tt0087469"
$ Type      : chr "movie"
$ DVD       : chr "13 May 2008"
$ BoxOffice : chr "$179,870,271"
$ Production: chr "N/A"
$ Website   : chr "N/A"
$ Response  : chr "True"
```

```
# Data are currently in JSON format, needs to be converted to
text
text_call <- content(call,as="text",encoding="UTF-8")
df <- fromJSON(text_call,flatten=T)
head(df) # the first few items are shown for display only, run
``df'' for all info.
```

R Console

```
$Title
[1] "Indiana Jones and the Temple of Doom"

$Year
[1] "1984"

$Rated
[1] "PG"

$Released
[1] "23 May 1984"

$Runtime
[1] "118 min"

$Genre
[1] "Action, Adventure"
```

Example 3

Fingerprint matching and image processing

A fingerprint image is totally unstructured. To analyze a fingerprint, key points on the print are automatically identified and then mapped. The map, which is structured data, is what is actually matched. Unstructured prints aren't analyzed; the extracted information is.

The `fingerprint` package in `R` can be used to analyse similarities and differences between fingerprints.

More generally, the `magick` package in `R` provides tools for cutting, editting and filtering of images, working with and combining image layers, exporting images to rasters and overlaying images on graphics files.

The following list provides some of the topics of interest for extracting information from unstructured data when we are interested in text mining e.g. analyzing natural language text.

Topic	Description
word frequency	the frequency that certain words appear
collocation	do words appear together more often than by chance
text classification	assigning tags or labels to text depending on content
word stemming	reducing words to unify across documents e.g. 'learns', 'learner', 'learning' could all be reduced to 'learn'
stem completion	the reduced word isn't a real word and hence e.g. 'comp' may become 'computer' or 'computable'
sentiment analysis	analysing sentiment in text e.g. identifying opinions
keyword extraction	automatically extracting the most important words

The following table provides some examples of `R` packages (it is by no means exhaustive), and their use for extracting and manipulating information from unstructured data.

Package	Short description
`tm`	text mining in R
	importing data, handling a collection of texts (corpus), metadata management, preprocessing methods
`wordcloud`	word clouds - visualise differences and similarity between documents
`SnowballC`	implements stemming - collapsing words to a common root for compression
`NLP` `OpenNLP`	machine learning for Natural Language Processing
`languageR`	statistical analysis on text data, e.g. vocabulary richness, vocabulary growth
`koRpus`	text mining, e.g. text readability
`RKEA`	keyword extraction from texts
`Tidytext`	text mining for e.g. sentiment analysis, determining feelings a writer is expressing in text
`textrank`	finding keywords and most relevant sentences in text

Example 4

To illustrate a simple example of using a couple of these packages. The following `R` code creates a word cloud using some of the introductory text from this document (this is only one example way to do this). The data are contained in the RData object for week 1, which can be found at: RData

```
# We need to install various packages....
install.packages("wordcloud")
install.packages("RColorBrewer")
```

```
install.packages("tm")
install.packages("dplyr")
```

```
library(wordcloud)
library(RColorBrewer)
library(tm)
library(dplyr)

# Create a vector containing only the text
# (note: this is a simple example with all text in a single
vector, the first step here, of reading the csv file, is not
required if you use the 'worddata' object from within the Rdata
file)
worddata <- read.csv("worddata.csv")
Text <- worddata$Text
# Create a corpus, which is just a collection of documents
# (Although, here it's simple and we only have one)
mytext  <- Corpus(VectorSource(Text))

# Use the dplyr and tm packages to tidy up the text a little.
# Specifically we want to remove numbers, punctuation, white
space
# Transform everything to lower case and remove common stop
words and frequent words such as 'and', 'of'.

mytext <- mytext %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)

mytext <- tm_map(mytext, content_transformer(tolower))
mytext <- tm_map(mytext, removeWords, stopwords("english"))

# create a dataframe containing each word in your first column
# and their frequency in the second column.

mytextdata <- TermDocumentMatrix(mytext)
mtdmatrix <- as.matrix(mytextdata)
mywords <- sort(rowSums(mtdmatrix),decreasing=TRUE)
mywordsdf <- data.frame(word = names(mywords),freq=mywords)

set.seed(412) # for reproducibility
```

```
# this command ensures that the same word cloud is generated
each time,
# the number in brackets is arbitrary, and can be any number
that you would like to choose.

# Produce the wordcloud
# The arguments here use the 'words', the frequencies of the
words, the minimum frequency of words to be included, the
maximum number of words to be included, a non-random order
indicates that words should be decreasing in size relative to
frequency, the proportion of words to be rotated 90 degrees
(rot.per), and a colour chart for the words

wordcloud(words = mywordsdf$word, freq = mywordsdf$freq,
min.freq = 1, max.words=50, random.order=FALSE,
rot.per=0.25,colors=brewer.pal(8, "Dark2"))
```

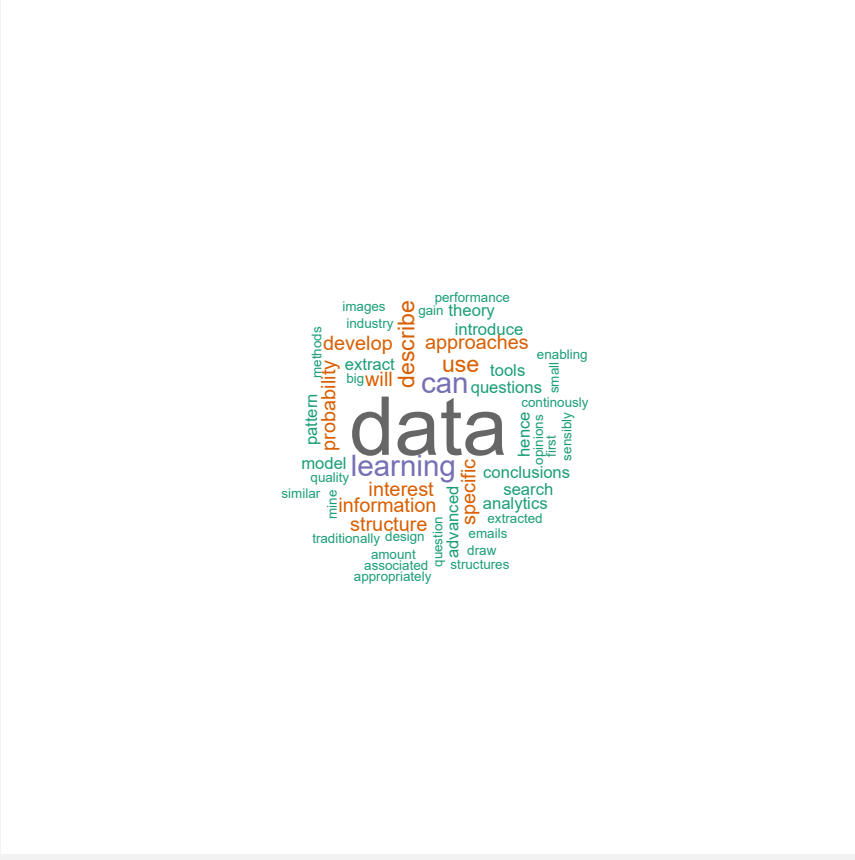


Figure 2

This is as far as we go with our treatment of unstructured/semi-structured data. However, more information can be found in the supplementary references below, and courses to follow. For example, text mining and network data will be covered in particular in the course 'Data Mining and Machine Learning II', and advanced machine learning approaches will be covered in 'Large Scale Computing'.

Structured data

In this course we will focus primarily on structured data (or unstructured data that have been processed to be in a structured form) that are either **quantitative**, and hence **numerical**, or **qualitative**, and in particular, **categorical** in nature.

Four scales of measurement can be defined for numerical and categorical data. We will distinguish among the four scales of measurement as set out in the table below.

Type of Data	**Sub-category**	**Description**
Categorical Scale		The individual observations themselves are categories and not numbers. Numerical summaries of the data are obtained by accumulating observations in categories.
	Nominal Scale	The categories are simply names or labels.
	Ordinal Scale	The categories are naturally ordered.
Numerical Scale		The individual observations are numbers.
	Discrete Scale	There are a finite or countable number of possible values, which we may list in the form: x_1, x_2 (see the definition of a discrete random variable in the *Probability and Stochastic Models* or *Probability and Sampling Fundamentals* courses).
	Continuous Scale	There are uncountably many different possible values (usually all the values in an interval of the real line), so it is not possible to list them all (see the definition of a continuous random variable in the *Probability and Stochastic Models* or *Probability and Sampling Fundamentals* courses).

Supplement 3

There are a variety of different resources that are useful for comparing structured and unstructured data, and to describe methods and approaches to extracting information from unstructured data. The following provides some examples:

An introduction to text analysis in R

A collection of relevant R packages for scraping data from the Web and to interact with Web services can be found at the Web Technologies and Services CRAN Task View

A link to the following e-book is available on the Moodle page:

- Mastering Data Analysis with R by Gergely Daroczi (Sept 2015)

This book might also be of interest:

- Mastering Text Mining with R by Ashish Kumar and Avinash Paul (Dec 2016)

These books are **not** required for this course.

Supplement 4

You all might be interested in the following which were new in 2022 from the Royal Statistical Society (RSS):

- The Data Science and AI section
- Advanced Data Science Professional certification
- The Alliance for Data Science professionals
- A website from the RSS to showcase data science in action

Supplement 5

Here are links to some recent articles on data science that might be of interest:

- The 5 biggest data science trends in 2022
- Is data scientist still the sexiest job of the 21st century
- Anaconda - State of Data Science report 2022

Summary - The Basic Approach to Learning from Data

For effectively every part of every context where we wish to *Learn from Data* one can identify the following integral steps in the procedure:

1. Specify clearly each question of interest;
2. Design a suitable means of gathering or identifying data sources from which to collate appropriate data to answer the question posed;
3. Identify data structure and type and extract/manipulate to obtain information required from data sources.

Once we have obtained our data, the next important step is to summarise and visualise the data in an appropriate form as part of quality assuring the data, and to get a subjective answer to our questions of interest. This next step is only the beginning of our analysis, and this will be the main focus of week 2.

Learning outcomes for week 1

By the end of week 1, you should be able to:

- state (and explain briefly) different approaches to learning from data;
- state and recognise different sources of data;
- state and recognise different types of data;
- define the terms *structured*, *semi-structured*, *unstructured* and *metadata*;
- explain (briefly) different approaches to collecting data;
- outline (briefly) methods to extract/summarise information from unstructured data;
- state and explain the different scales of measurement for structured data.

Review exercises and answers to all tasks/review exercises are provided below.

Review exercises

Task 4

The following description was taken from the paper 'Predicting Customer Behavior with Combination of Structured and Unstructured Data', Journal of Physics: Conference Series, (2019). Identify the data sources used and state whether the data sources are structured or unstructured, quantitative or qualitative.

The paper describes a questionnaire that was provided to customers, the questionnaire retrieved the following information from the respondents: age, gender, marital status, occupation, highest academic qualification, states and town the respondents reside in, level of computer literacy, average income per month, hobbies, income range, and frequency of visiting YouTube; Facebook; Twitter; Google. There were also open questions on the customers attraction to online business, their dislike about online business, attraction and dislike about doing business through the mobile phone platforms.

Task 5

For the questions below, think about how you would obtain data to answer the question of interest. Are the data structured or unstructured? Qualitative or quantitative? What are the challenges?

- By how much, on average, will a new drug decrease blood pressure?
- What effect will changes in the UK government have on the share price of large multi-national companies?

Task 6

There are many different data repositories and sources of data available on the web providing lots of valuable information. You might want to look at a few of the websites below to see some examples of the websites available.

UK data service

Satellite data for Lake water temperature

Open UK Government Data

The environmental information data centre

For these websites, what might you want to look for to ensure the data are appropriate to use?

Task 7

Additional supplementary task

Use the code provided earlier within the learning material here (or similar alternative code) to produce a word cloud of a document that describes something about you (e.g. short biography, job description).

Task 8

Additional supplementary task

Use the information in example 2 to download and extract the information for a different movie.

Answer 1

Advantages and challenges of data sources

Collections of numbers, demographic information or information from opinion polls can be summarised, and visualised in quite a straight forward way (and these approaches will be considered in more detail in week 2). However, these data might not be automatically available. A study may need to be planned and conducted in order to obtain the data. For example, through a designed monitoring programme or a survey.

Emails, tweets, images, maps, newspaper articles surround us every day and are readilly available digitially. However, the main challenge is how you collate and extract information from them. For example, how to source and extract key words or numeric information from documents. How do we summarise an image? It might be of interest to extract key information from it e.g. is there a car in an image or a picture of an animal? What are the dominant features of an image?

Answer 2

Data sources, structures and types

Assuming that the safety report information is collated in e.g. a database or spreadsheet:

- Structured data: time, site, role, demographic and clinical attributes, severity classification, type
- Unstructured data: the free text information entered
- Quantitative: time
- Qualitative: It seems likely here that all other aspects are qualitative. The exception is for 'clinical attributes' for which we would need more information to make a judgement.

Answer 3

Personal interview

Advantages include: excellent response rates and enables you to ask follow-up questions to responses that are not clear.

Disadvantages include: method is expensive and time-consuming as it requires interviewer training, transport, and remuneration.

Telephone interview

Advantages include: data can be collected quickly and the telephone interview is cheaper than personal interviews.

Disadvantages include: harder to gain the trust of respondents and so the response rate might not be as good which can introduce bias (i.e. our results are not representative of all views of the population, and inclined or prejudiced towards particular groups or opinions).

Online interview

Advantages include: low-cost way of interviewing many respondents. Another benefit is anonymity; you can get sensitive responses that participants might not feel comfortable providing with personal interviews.

Disadvantages include: poor response rate, might not get a representative sample (i.e. one that contains individuals representing all attributes/characteristics of your population) and it's not possible to seek clarification on responses that are unclear.

Mailed questionnaire

Advantages include: might be able to obtain information that respondents are unwilling to give when interviewing in person.

Disadvantages include: poor response rate, inaccuracy in mailing address, delays or loss of mail could also affect the response rate. Cannot be used to interview people with low literacy and cannot seek clarifications on responses.

Focus groups

Advantages include: require fewer resources and time compared to interviewing individuals. Can request clarification on unclear responses.

Disadvantages include: sample selected may not represent the population accurately and dominant participants can influence the response of others. The data are likely to be more *unstructured* by design.

Answer 4

Data sources, format and type

Assuming that the questionnaire information is collated (or partially collated) already in e.g. a database or spreadsheet

- Structured data: age, gender, marital status, occupation, academic qualification, state and town residing in, level of computer literacy. average income, hobbies, income range, frequency of visits online.
- Unstructured data: open answers on customer attraction to online business, dislikes about online business, and attraction/dislikes to doing business over mobile phones.
- Quantitative: age, average income, income range, frequency data*
- Qualitative: It seems likely here that all other aspects are qualitative.

*Frequency data are sometimes considered as a factor or categorical e.g. if there are only small numbers for each visit.

Answer 5

Data structures and types

By how much, on average, will a new drug decrease blood pressure?

To collect appropriate data, a study would need to be designed and carried out. Particapants would be recruited to the study and likely assigned to one of three groups: a control group receiving a placebo (i.e. 'no effect' treatment), a previous treatment group receiving the current 'gold standard' treatment, and a new treatment group receiving the new drug.

The designed study would produce structured data designed and collected in a particular way to eliminate sources of bias. Careful thought would need to be given to the participants that are recruited to the study e.g. a group of patients all with a history of at least 6 months of high blood pressure but no other (known) underlying medical conditions.

The resulting data would be numerical (quantitative), with the recorded output being blood pressure (say) before and after treatment.

What effect will changes in the UK government have on the share price of large multi-national companies?

Historic data could be sourced from the web providing information on the share price of say the top 100 multi-national companies the day before and the day after UK general elections over a period of time.

Initial data may be unstructured or semi-structured. It may require to be scraped from different web sources.

Challenges here would be around obtaining the data and ensuring the accuracy of the data. It would be important to carefully validate sources of the data. The data would be a mixture of quantitative for the share price (or change in price), and qualitative for the political party elected into government.

Answer 6

Collection and quality

In each case we would be looking, for example, for documentation on the website that contains information on data collection/collation processes, quality checks that have been carried out on the data, metadata to inform how to use and interpret the data appropriately.

Answer 7

R code

The code below needs suitably adjusted for your dataset. The text used here was stored in the first column of a spreadsheet and stored as a .csv file, but other options are possible.

```
library(wordcloud)
library(RColorBrewer)
library(tm)
library(dplyr)

# Create a vector containing only the text
# (the first step here, reading from the csv, is not required
if you use the 'worddata' object from within the Rdata file)
```

```r
worddata <- read.csv("worddata.csv")
Text <- worddata$Text
# Create a corpus, which is just a collection of documents
# (Although, here it's simple and we only have one)
mytext  <- Corpus(VectorSource(Text))

# Use the dplyr and tm packages to tidy up the text a little.
# Specifically we want to remove numbers, punctuation, white
space
# Transform everything to lower case and remove common stop
words and frequent words such as 'and', 'of'.

mytext <- mytext %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)

mytext <- tm_map(mytext, content_transformer(tolower))
mytext <- tm_map(mytext, removeWords, stopwords("english"))

# create a dataframe containing each word in your first column
# and their frequency in the second column.

mytextdata <- TermDocumentMatrix(mytext)
mtdmatrix <- as.matrix(mytextdata)
mywords <- sort(rowSums(mtdmatrix),decreasing=TRUE)
mywordsdf <- data.frame(word = names(mywords),freq=mywords)

set.seed(123) # for reproducibility
# produce the wordcloud
# The arguments here use the 'words', the frequencies of the
words, the minimum frequency of words to be included, the
maximum number of words to be included, a non-random order
indicates that words should be decreasing in size relative to
frequency, the proportion of words to be rotated 90 degrees
(rot.per), and a colour chart for the words.


wordcloud(words = mywordsdf$word, freq = mywordsdf$freq,
min.freq = 1, max.words=50, random.order=FALSE, rot.per=0.25,
colors=brewer.pal(8, "Dark2"))
```

Answer 8

Note: there is sometimes a conflict between the `tm` package and `httr` package. If you have any problems here try detaching the `tm` and `NLP` packages.

```
# Require the following packages
library(httr)
library(jsonlite)


# Access movie information data from OMDb API

# To obtain access, must obtain API key (can be obtained from
http://www.omdbapi.com/)
api_key <- "aeb77595"
# Specify the title of which movie to search for (replace
spaces with +)
movie <- "blade+runner"
# Create web path
path <- paste("http://www.omdbapi.com/?
t=",movie,"&apikey=",api_key,sep="")

# Access HTTP data using GET command
call <- GET(url=path)

# To see if a HTTP request is successful, check the reply
status code (successful code is 200)
status_code(call)

# To view the content of the call, the number of characters is
truncated here for display only.
str(content(call), nchar.max=20) # remove ``nchar.max'' for all
info

# Data are currently in JSON format, convert to text
text_call <- content(call,as="text",encoding="UTF-8")
df <- fromJSON(text_call,flatten=T)
head(df) # the first few items are shown for display only, run
``df'' for all info.
```

Footnotes

1. Many thanks to Suzy Whoriskey for all her contributions to the development of the course material ↩