ECS766P Data Mining

Week 11: Data Mining Applications & Data Ethics

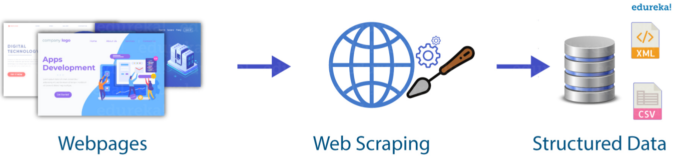
Emmanouil Benetos
emmanouil.benetos@qmul.ac.uk

December 2021

School of EECS, Queen Mary University of London

Last week: Web Mining

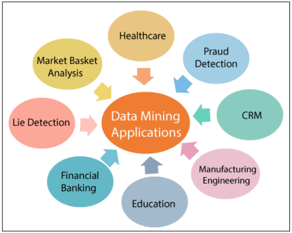
- Six paradigms for today's Internet
- Technology review
- Internet Mining Applications
- Ingesting Internet data
- Search Engine Indexing & Ranking



Webpages → Web Scraping → Structured Data

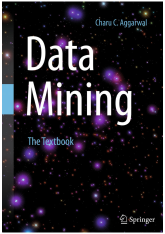
This week's contents

1. Mining Text Data
2. Mining Timeseries Data
3. Data Ethics



Reading

- Chapters 13, 14, 16, and 20 of C. C. Aggarwal, “Data Mining: The Textbook”, Springer, 2015 [non-essential reading]



Data Ethics content adapted from material by Dr Usman Naeem and the Institute of Coding (IoC)

`http://eecs.qmul.ac.uk/ioc/`

Mining Text Data

Mining Text Data: Introduction

Mining Text Data

The *text domain* is sometimes challenging for mining purposes because of its *sparse* and *high-dimensional* nature. Therefore, specialised algorithms need to be designed. The first step is the construction of a bag-of-words representation for text data.

Several *preprocessing steps* need to be applied, such as *stop-word removal*, *stemming*, and the *removal of digits* from the representation.

Algorithms for problems such as clustering and classification need to be modified as well. The *k-means method*, *hierarchical methods*, and *probabilistic methods* can be suitably modified to work for text data.

Mining Text Data: Introduction

Text data are found in many domains:

- Digital libraries
- Web and Web-enabled applications
- News services

Modeling of Text:

- A sequence (*string*)
- A multidimensional record

Mining Text Data: Multidimensional Representations

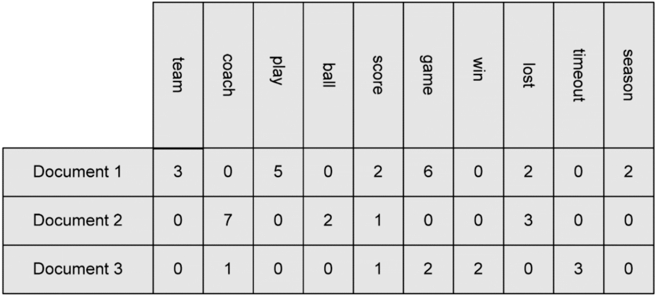
Some terminology:

- Data point: document
- Data set: corpus
- Feature: word/term
- The set of features: lexicon

Vector Space Representation:

- Common words are removed
- Variations of the same word are consolidated
- Displays frequencies of individual words

Mining Text Data: Vector Space Representation



	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Figure: vector space representation for a collection of documents. This particular representation is also called a document-term matrix.

Mining Text Data: Specific Characteristics of Text

Number of “Zero” Attributes (*Sparsity*):

- Most attribute values in a document are 0. This phenomenon is referred to as high-dimensional sparsity.
- Affects many fundamental aspects of text mining, such as distance computation.

Nonnegativity:

- Frequencies are nonnegative.
- The presence of a word is statistically more significant than its absence.

Side Information:

- Hyperlinks or other metadata associated with a document.

Mining Text Data: Data Preprocessing

Stop Word Removal:

- Words in a language that are not very discriminative for mining
- Articles, prepositions, and conjunctions

Stemming:

- Consolidate variations of the same word
- Singular and plural representations, different tenses, common root extraction

Punctuation Marks:

- Commas, semicolons, digits, hyphens

Mining Text Data: tf–idf representation

Inverse Document Frequency:

$$idf(w) = log_{10}(|D|/|D_w|)$$

where $|D_w|$ is the number of documents in which the word w occurs, and $|D|$ is the total number of documents.

Term Frequency:

$$tf(w, d) = \frac{f_{w,d}}{\sum_{w' \in d} f_{w',d}}$$

The ratio of the number of appearances $f_{w,d}$ of word w in document d, divided with the total number of words in document d.

Term frequency–Inverse document frequency (tf–idf):

$$tfidf(w, d) = tf(w, d) \cdot idf(w)$$

Mining Text Data: Representative-Based Algorithms

Most clustering algorithms can be extended to text data, following some modifications.

Representative-Based Algorithms: Since the vector space representation of text is also a multidimensional data point, algorithms such as *k*-means can be used for text data.

Modifications:

- Choice of Similarity Function: Cosine similarity
- Computation of the cluster centroids:
    - Low-frequency words in the cluster are not retained
    - A representative set of words are retained for each cluster (200 to 400 words)
    - Have significant effectiveness advantages

Mining Text Data: Scatter/Gather Approach

The scatter/gather approach is effective because of its ability to combine hierarchical and k-means algorithms.

- While the k-means algorithm scales as $O(k \cdot n)$, it is sensitive to initialisation.
- While hierarchical partitioning algorithms are very robust, they typically do not scale well.
- A Two-phase Approach:
    1. Apply a procedure to create a robust set of initial seeds (**buckshot** or **fractionation** procedure)
    2. Apply a k-means approach on the resulting set of seeds

Mining Text Data: Scatter/Gather Approach

Buckshot

- Select a seed (sample of documents) of size $\sqrt{k \cdot n}$
    - k is the number of clusters
    - n is the number of documents
- Apply agglomerative hierarchical clustering to this initial sample of seeds
    - The time complexity is $O(k \cdot n)$
- Agglomerative clustering methods
    - The individual data points are successively merged into higher-level clusters.

Mining Text Data: Scatter/Gather Approach

Fractionation

- Break up the corpus into n/m buckets, each of size m
- An agglomerative algorithm is applied to each bucket to reduce them by a factor $\nu \in \{0, 1\}$
- Then, we obtain νn agglomerated documents over all buckets
    - An "agglomerated document" is defined as the concatenation of the documents in a cluster.
- Repeat the above process until k agglomerated documents

Mining Text Data: Scatter/Gather Approach

Fractionation

- Types of Partition
  - Random partitioning
  - Sort the documents by the index of the j^{th} most common word in the document. Contiguous groups of m documents in this sort order are mapped to clusters.
- Time Complexity
  - $O(nm(1 + \nu + \nu^2 + ...)) = O(nm)$

Mining Text Data: Scatter/Gather Approach

k-means algorithm

When the initial cluster centers have been determined with the use of the buckshot or fractionation algorithms, one can apply the k-means algorithm with the seeds obtained in the first step.

- Each document is assigned to the nearest of the *k* cluster centers
- The centroid of each such cluster is determined as the concatenation of the documents in that cluster
- Furthermore, the less frequent words of each centroid are removed

Mining Timeseries Data

Mining Timeseries Data: Introduction

Mining Timeseries Data

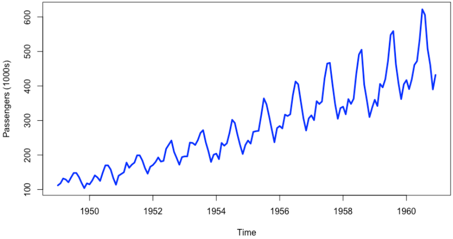
Timeseries data is common in many domains, such as *sensor networking*, *healthcare*, and *financial markets*.

Typically, timeseries data needs to be *normalised*, and *missing values* need to be imputed for *effective processing*. Numerous data reduction techniques such as *Fourier* and *wavelet transforms* are used in timeseries analysis. The choice of *similarity function* is the most crucial aspect of time series analysis.

Forecasting is an important problem in timeseries analysis because it can be used to make *predictions* about data points in the future. Most timeseries applications use either *point-wise* or *shape-wise analysis*.

Mining Timeseries Data: Introduction

- Temporal data may be either *discrete* or *continuous*:
  - Continuous temporal data sets are **timeseries**
  - Discrete temporal data sets are **sequences**
- Time series data are viewed as contextual data representations, with contextual and behavioural attrributes.
- Two types of models:
  - Real-time analysis
  - Retrospective analysis



Mining Timeseries Data: Data Preparation

Multivariate Time Series Data

A time series of length n and dimensionality d contains d numeric features at each of n timestamps $t_1, ..., t_n$. Each timestamp contains a component for each of the d series. Therefore, the set of values received at timestamp t_i is $\overline{Y}_i = (y_i^1, ..., y_i^d)$. The value of the j^{th} series at timestamp t_i is y_i^j.

In a univariate time series, the value of d is 1. In such cases, a series of length n is represented as a set of scalar behavioral values $y_1, ..., y_n$, associated with the timestamps $t_1, ..., t_n$.

Mining Timeseries Data: Data Preprocessing

Handling Missing Values

The most common methodology used for handling missing, unequally spaced, or unsynchronised values is linear interpolation.

Let y_i and y_j be values of the timeseries at times t_i and t_j, respectively, where $i < j$. Let t be a time drawn from the interval (t_i, t_j). Then, the interpolated value of the series is given by:

$$y = y_i + \left(\frac{t - t_i}{t_j - t_i}\right) \cdot (y_j - y_i)$$

Polynomial interpolation or spline interpolation are also possible.

Mining Timeseries Data: Data Preprocessing

Noise Removal

- Binning
  - Grouping data into time intervals of size k
  - Averaging value of data points in each interval
  - Let $y_{i \cdot k+1} \ldots y_{i \cdot k+k}$ be the values at timestamps $t_{i \cdot k+1} \ldots t_{i \cdot k+k}$. The new binned value is:

$$y'_{i+1} = \frac{\sum_{r=1}^{k} y_{i \cdot k+r}}{k}$$

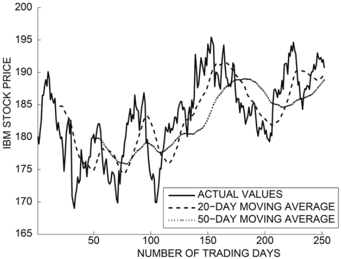
- Moving-Average Smoothing: Moving-average (*rolling averages*) methods reduce the loss in binning by using overlapping bins, over which the averages are computed. Here a bin is constructed starting at each timestamp in the series.

Mining Timeseries Data: Data Preprocessing

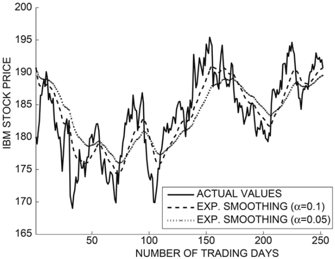
- Exponential Smoothing

  The smoothed value $y_i^{'}$ is defined as a linear combination of the current value y_i, and the previously smoothed value $y_{i-1}^{'}$. Parameter $\alpha \in (0, 1)$ controls the smoothing:

$$y_i^{'} = \alpha \cdot y_i + (1 - \alpha) \cdot y'_{i-1}$$



(a) Moving average smoothing



(b) Exponential smoothing

Mining Timeseries Data: Data Preprocessing

Normalisation

- Minmax normalisation to (0,1)
  Let the minimum and maximum value of the time series be *min* and *max*, respectively. Then, the time series value y_i is mapped to the new value $y_i^{'}$ in the range $(0, 1)$ as:

$$y_i^{'} = \frac{y_i - min}{max - min}$$

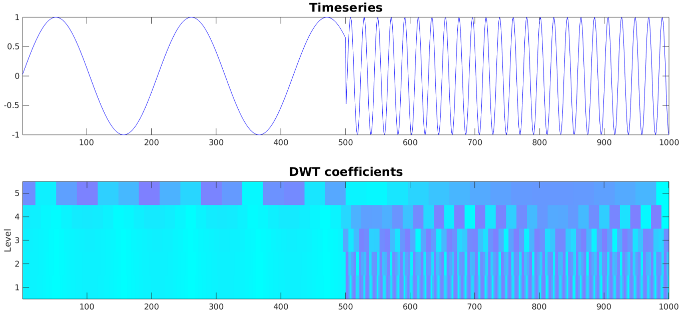
- z-score normalisation
  Let μ and σ represent the mean and standard deviation of the values in the timeseries. Then, the timeseries value y_i is mapped to a new value z_i as:

$$z_i = \frac{y_i - \mu}{\sigma}$$

Mining Timeseries Data: Data Transformation

Discrete Wavelet Transform (DWT)

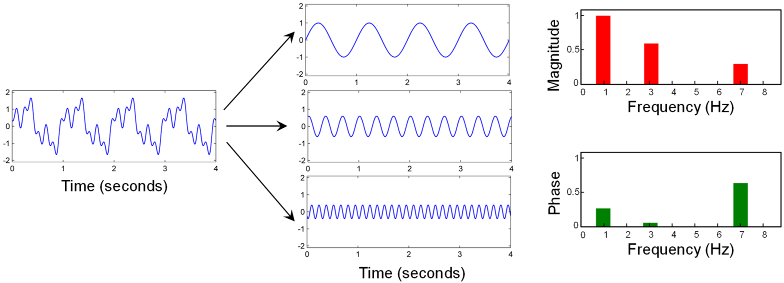
- DWT converts a timeseries to multidimensional data.
- A key advantage is that the DWT can capture both frequency and temporal information.



Mining Timeseries Data: Data Transformation

Discrete Fourier Transform (DFT)

Idea: Decompose a given signal into a superposition of sinusoids (elementary signals).



- The magnitude reflects the intensity at which the sinusoid of a specific frequency appears in the signal.
- The phase reflects how the sinusoid has to be shifted to best correlate with the signal.

Mining Timeseries Data: Data Transformation

Discrete Fourier Transform (DFT)

Any series of length n can be expressed as a linear combination of smooth periodic sinusoidal series. Consider a time series $x_0 \ldots x_{n-1}$. Each coefficient X_k of the Fourier transform is a *complex* value which is defined as follows:

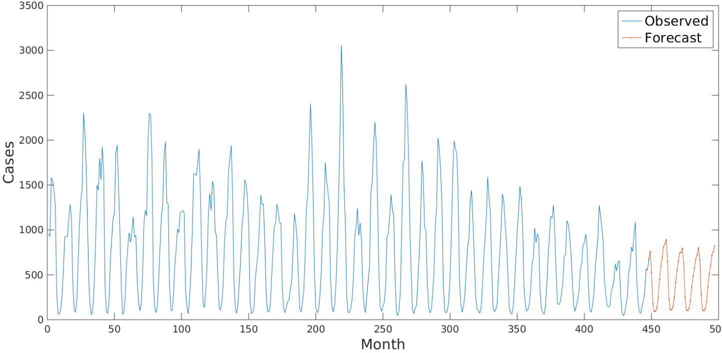
$$X_k = \sum_{r=0}^{n-1} x_r \cdot e^{-ir\omega k} = \sum_{r=0}^{n-1} x_r \cdot \cos(r\omega k) - i \sum_{r=0}^{n-1} x_r \cdot \sin(r\omega k) \quad \forall k \in \{0 \ldots n-1\}$$

Where ω is set to $2\pi/n$ radians, and the notation i denotes the imaginary number $\sqrt{-1}$.

Mining Timeseries Data: Forecasting

The **prediction** of future trends has applications in:

- Retail sales
- Stock markets
- Weather forecasting
- Medicine and health



Mining Timeseries Data: Forecasting

Timeseries can be either **stationary** or **nonstationary**:

- A stationary stochastic process is one whose parameters, such as the mean and variance, do not change with time.
- A nonstationary process is one whose parameters change with time.

In forecasting, we typically convert or assume timeseries to be stationary and use statistical parameters for forecasting.

Statistical methods for timeseries forecasing:

- Autoregressive (AR) models
- Moving average (MA) models
- Autoregressive Moving Average (ARMA) models
- Autoregressive Integrated Moving Average (ARIMA) models

Data Ethics

Is our underlying data fit for purpose?

The objective of this section is to provide students with an understanding of the key ethical and legal issues as well as challenges that they might face when working on data mining. The lecture will also provide insights on how to address these issues based on the UK's Data Ethics Framework.

Fundamental questions:

- *"Does my analysis of the dataset infringe on a user's privacy?"*
- *"Does the use of a particular dataset lead to ethical issues?"*
- *"Is the dataset accurate and fit for purpose?"*

What is Data Ethics?

In simple terms, ethics can be considered as conducting an activity in a ‘good’, ‘acceptable’ or ‘right’ way. But, how can we determine what is ‘good’, ‘acceptable’ or ‘right’? This can be subjective, as this is something that is normally based on the values that are the norm within different groups of people, which is normally influenced by factors such as culture. The moral philosophy discipline categorises ethics into the following two perspectives:

- **Kantian:** *The ethical action is driven by moral values and principles of the individual.* This perspective is not concerned about the consequence of an individual’s actions.
- **Utilitarian:** *The action is ethical if the intention is to maximise positive outcomes for a larger population of individuals.* This perspective is concerned about the consequence of an individual’s actions.

What is Data Ethics?

Both Kantian and Utilitarian perspectives have their advantages and disadvantages:

- **Kantian perspective** can be difficult to recognise moral (good) values of an individual.
- **Utilitarian perspective** can overlook minority groups, as this perspective only considers positive outcomes for the larger group of individuals.

Data Ethics is concerned with the values and methods that are adopted when we generate, analyse and disseminate data. Hence, a fundamental objective of data ethics is to ensure that you consider the social and legal implications of how and for what purpose you use the data and algorithms as a data scientist.

Data Ethics - Suggested Reading

- *Mingers, J., & Walsham, G. (2010). Towards ethical information systems: The contribution of discourse ethics. MIS Quarterly, 34(4), 833–854.*
- *Pasquale, Frank & Citron, Danielee Keats (2014) Promoting Innovation While Preventing Discrimination: Policy Goals for the Scored Society. Washington Law Review 89:1413.*
- *Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision making: A call for action on the long-term societal effects of 'datification'. The Journal of Strategic Information Systems.*
- *Vallor, S. (2016). Technology and the virtues: A philosophical guide to a future worth wanting. Oxford University Press.*
- *Gumbus, A., & Grodzinsky, F. (2016). Era of big data: Danger of descrimination. ACM SIGCAS Computers and Society, 45(3), 118–125.*

Case study

“We also should be worried about misdirection of the innovation of scoring in the employment context—particularly if firms can effectively hide misconduct via scores. Existing laws prohibit some discriminatory uses of the data. For example, an employer cannot fire workers simply because they have an illness. But Big Data methods are able to predict diabetes from a totally innocuous data set (including items like eating habits, drugstore visits, magazine subscriptions, and the like). [...] For example, a firm could conclude a worker is likely to be diabetic and that they are likely to be a “high cost worker” given the significant monthly costs of diabetic medical care.

(from Pasquale and Citron, 2014)

What is the Data Ethics Framework?

The Data Ethics Framework has been developed by the UK government that prescribes the design of appropriate data use, which is aimed at statisticians, analysts and data scientists working directly/indirectly within the public sector. The objective of this framework is to encourage ethical data use to build better services, which is based on the following values of the Civil Service Code:

- Integrity
- Honesty
- Objectivity
- Impartiality

Resources:

- Data Ethics Framework `https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework`
- Data Ethics Workbook `https://www.gov.uk/government/publications/data-ethics-workbook`

Which data are we allowed to use?

Quantitative secondary research sources that includes datasets such as census data, birth/death rates, unemployment rates are a type of data normally generated by governments, organisations and charities.

Are we allowed to make use of this data? The answer is ‘yes’, however we need to be aware of legislations related to the usage of data. According to gov.uk, This includes how we:

- Produce statistics
- Protect privacy by design
- Minimise the data needed to achieve our need
- Keep personal and non-personal data secure

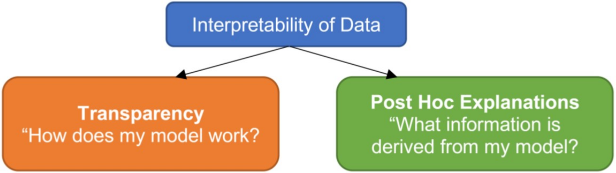
Personal Data Protection

If you intend to use personal data, then you must ensure that you comply with the principles of the General Data Protection Regulation (GDPR) and Data Protection Act 2018 (DPA 2018).

The importance of GDPR cannot be underestimated, as it aims to improve the protection of data subject's rights within Europe. In addition to this, GDPR clearly articulates what companies must do to protect personal data.

Personal Data Protection

Data scientists also need to take into consideration the interpretability of data, as this is also a GDPR requirement.



There are two aspects to the interpretability of data (this legal definition also includes models), which are transparency and post hoc explanations.

Personal Data Protection

Transparency is based on how your model works, while post hoc explanations are based on the information derived from your model. From a GDPR perspective, this is important as a user has the legal right to find out how an algorithmic decision was made about them.

Resources:

- General Protection Data Protection Regulation (GDPR)
  `https://gdpr-info.eu`
- Data Protection Act 2018 (DPA 2018)
  `https://www.legislation.gov.uk/ukpga/2018/12/enacted`

Case Study: Autonomous Vehicles

Decision making models are dependent on data that is generated given a particular scenario. One such example is the series of decisions that have to be made given the data captured by the multiple sensors in Autonomous Vehicles (AVs). The questions that we need to think about are:

- *"Who makes these decisions?"*
- *"Are there any legal liabilities for these decisions?"*

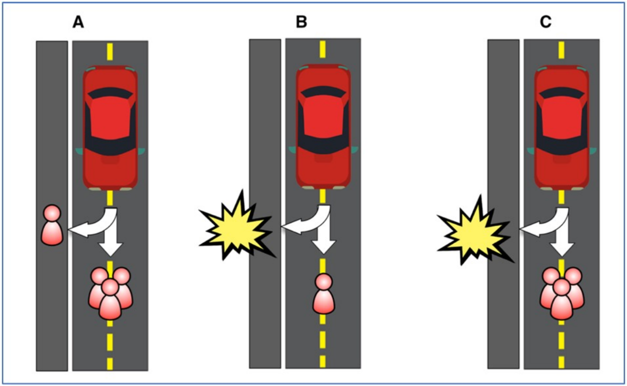
The advent of AVs is seen as a progressive step towards a smart city infrastructure, where the motivation is to provide safe roads by reducing traffic accidents. However, decision-making models will likely make a series of difficult moral decisions if the vehicle is involved in a crash.

Case Study: Autonomous Vehicles

Let us consider the following scenarios:

- Scenario A:
  The vehicle will keep on driving straight on the road and kill a group of pedestrians or
  The vehicle will swerve to the right and kill one person walking on the pavement
- Scenario B:
  The vehicle will keep on driving straight on the road and kill one pedestrian or
  The vehicle will swerve to the right onto the pavement and kill the passenger in the vehicle
- Scenario C:
  The vehicle will keep on driving straight on the road and kill a group of pedestrians or
  The vehicle will swerve to the right onto the pavement and kill the passenger in the vehicle

Case Study: Autonomous Vehicles



These scenarios clearly illustrate why the designing of these models can lead to ethical dilemmas, however will the legal liabilities be the same for a pre-programmed AV and a human-driven car?

Readings:

- *Contissa, G., Lagioia, F., & Sartor, G. (2017). The Ethical Knob: ethically-customisable automated vehicles and the law. Artificial Intelligence and Law, 25(3), 365-378.*
- *Ethics guidelines for trustworthy AI* `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`

Data Reliability

Limitations with datasets can lead to data analysis being misleading and unreliable. Hence, this is seen as an ethical concern.

How do we determine if a dataset is reliable?

We need to take into consideration the lineage of the dataset, as this will allow us to trace back any errors or discrepancies to the beginning of the data analysis process.

Data Reliability

Identification of the data lineage can be done by answering the following set of questions:

- *What is the source of the data?*
- *How was the data collected? Was it by humans? Or automated systems?*
- *Why was the data collected?*
- *Does the data reflect its target population?*
- *Are there any patterns in the data?*
- *Is data likely to change over time?*
- *Are there omissions from the dataset?*
- *What was the sampling method used to collect this data?*

Data Bias

Bias within datasets can be caused by:

- *Datasets that do not accurately represent the cohort that the insights will be based on.*
- *Datasets produced by humans, which can be in the form of curated news articles or social media content, which leads to bias against a group of people.*

There are huge ethical implications of having bias within datasets that can lead to biased models that will be prejudiced and harmful towards people.

An example of this is the case study about the “COMPAS Recidivism Algorithm” that was used to predict a defendant’s likelihood to commit a crime.

Data Bias - Types of Biases

Selection Bias

This type of bias occurs when the dataset does not reflect the population or cohort that the insights or decisions will be based on. This is very common with surveys. Hence, it tends to lead to a situation where you only end up with willing participants who are a small subset of the population and do not reflect the characteristics of an average person. Hence, the existence of this bias is due to the need of working with data that is easily accessible.

Self-Selection Bias

This is a subcategory of the selection bias, where the subject within the analysis selects themselves. For example, if you are running an online poll on how many people in a town can use an email client. The results for this will not represent the entire town as only the participants who had received the poll via email would be the most likely to reply with a response.

Data Bias - Types of Biases

Omitted Variable Bias

This bias occurs when variables or features are omitted from the dataset with the belief that they are not relevant to the output given existing beliefs.

Observer Bias

The is type of bias occurs when the data scientist subconsciously influences the outcome of their research by:

- Having previous knowledge or subjective feelings about a sample of people being studied.
- Unintentional manipulation of participants during surveys or interviews.
- Cherry picking a group of people who have characteristics that will support the data scientist’s hypothesis.

Data Bias - Types of Biases

Social Bias

Social bias can be positive and negative and refers to being in favor or against individuals or groups based on their social identities. Commonly occurs in data science when using data collected from the web, news, and social media.

The following case study illustrates an example of this, where text features trained on Google News articles exhibited female and male gender stereotypes:

- *T. Bolukbasi et al, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", 30th Conference on Neural Information Processing Systems, 2016.*

Personal vs Sensitive Data

What is the difference between personal and sensitive data?

Personal Data

Personal data is information that can be used to identify an individual. Typical examples are name (first/middle/last name), address, email address, national insurance number, location data, IP address, signature, date of birth and bank account details.

Typically datasets are made up of multiple pieces of personal data which can be combined together to identify an individual.

Personal vs Sensitive Data

Sensitive Data

Sensitive data is a category of personal information that may lead to harm or discrimination if not treated with extra care and security. For example, sensitive information about an individual could be on:

- ethnicity
- religious beliefs
- political views and opinions
- sexual orientation
- trade union membership
- biometric data
- health records
- criminal records

This type of data should be encrypted or pseudonymised and stored separately from other personal data.

Resources:

- *https://www.itgovernance.co.uk/data-protection-dpa-and-eu-data-protection-regulation*
- *https://ico.org.uk/media/for-organisations/documents/1554/determining-what-is-personal-data.pdf*

Quiz

Question 1. Which of the following is considered personal data?

- A Salary/wages
- B Religious beliefs
- C Sexual orientation
- D Philosophical beliefs

Question 2. Which of the following is considered sensitive data?

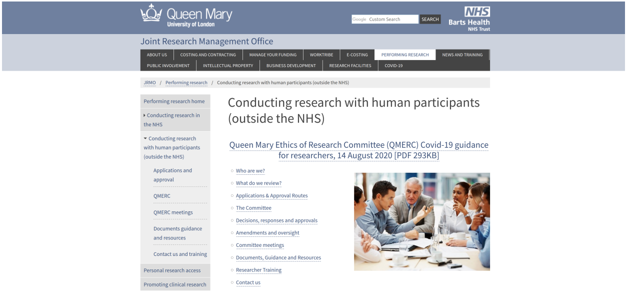
- A Hours of employment
- B Emergency contact person details
- C IP address
- D Religious affiliation

Research Ethics at QMUL

All projects that involve human participants or personal data require ethics approval from the university - including MSc projects!

Most projects which involve surveys/questionnaires can be approved through the 'low risk' ethics approval process which takes approximately 3-4 weeks.

More information: http://www.jrmo.org.uk/performing-research/conducting-research-with-human-participants-outside-the-nhs/



Joint Research Management Office

JRMO / Performing research / Conducting research with human participants (outside the NHS)

Performing research home
▸ Conducting research in the NHS
▾ Conducting research with human participants (outside the NHS)
Applications and approval
QMERC
QMERC meetings
Documents guidance and resources
Contact us and training
Personal research access
Promoting clinical research

Conducting research with human participants (outside the NHS)

Queen Mary Ethics of Research Committee (QMERC) Covid-19 guidance for researchers, 14 August 2020 [PDF 293KB]

- ○ Who are we?
- ○ What do we review?
- ○ Applications & Approval Routes
- ○ The Committee
- ○ Decisions, responses and approvals
- ○ Amendments and oversight
- ○ Committee meetings
- ○ Documents, Guidance and Resources
- ○ Researcher Training
- ○ Contact us

Summary

Mining Text Data It is the process of deriving high-quality information from text-like datasets.

Mining Timeseries Data It comprises methods for analysing timeseries data in order to extract meaningful statistics and other characteristics of the datasets.

Data Ethics evaluates moral problems related to data, algorithms and corresponding practices in order to formulate and support morally good solutions.

Data Reliability refers to the assurance of the accuracy and consistency of datasets.

Data Bias results in skewed outcomes, low accuracy levels, and analytical errors.

Personal Data is information on an individual. Sensitive Data is specific personal information that can cause discrimination.

Questions?

also please use the forum on QM+