NPA Data Science

4: Data Types and Formats, Open Data and Ethics

Topic Outcomes

- Describe common data types and data formats including structured and unstructured data
- Explain types & sources of large datasets, the philosophy and sources of open data
- Explain the concept of data ethics, including data bias, with reference to national and international standards and frameworks

Data Types and Formats

A key way to separate data into types is through the categories of qualitative and quantitative data.

Qualitative data is defined as information that describes or categorises something. It answers the broad question of “What qualities does this have?”. It cannot be easily measured or counted and therefore often doesn’t contain numbers. For example, you might interview customers to determine which social media platform they use most. You would then categorise the responses by platforms such as Facebook, Twitter, Quora, Snapchat, etc. Or an ecommerce retailer may poll shoppers to see which colour - teal, gray, or white - is preferable for a specific item. (Note: if you combine all the results from the poll - e.g. 45 teal, 70 gray, and 52 white, this becomes quantitative data.)

In some instances, a number or code may be assigned to qualitative descriptions or categories. For example, a company may assign numbers 1-5 to a satisfaction survey: Very satisfied (5), Satisfied (4), Somewhat satisfied (3), Somewhat dissatisfied (2), and Dissatisfied (1).

There are three types of qualitative data: binomial data, nominal data, and ordinal data.

- **Binomial data** (or binary data): this divides information into two mutually exclusive groups. Examples of binary data are true/false, right/wrong, accept/reject, etc.
- **Nominal data (**or unordered data): this groups information into categories that do not have implicit ranking. Nominal data examples include colors, genres, occupations, geographic location, etc.

- **Ordinal data** (or ordered data): as the name implies, information is categorised with an implied order. Examples of ordinal data are small/medium/large, unsatisfied/neutral/satisfied, etc.

Quantitative data is anything that can be measured or counted. This is also called numeric data because it deals with numbers. There's a wide range of quantitative data examples in statistics such as monthly revenue, distance of a race and time of the winner, calories in a meal, temperature, salary, etc.

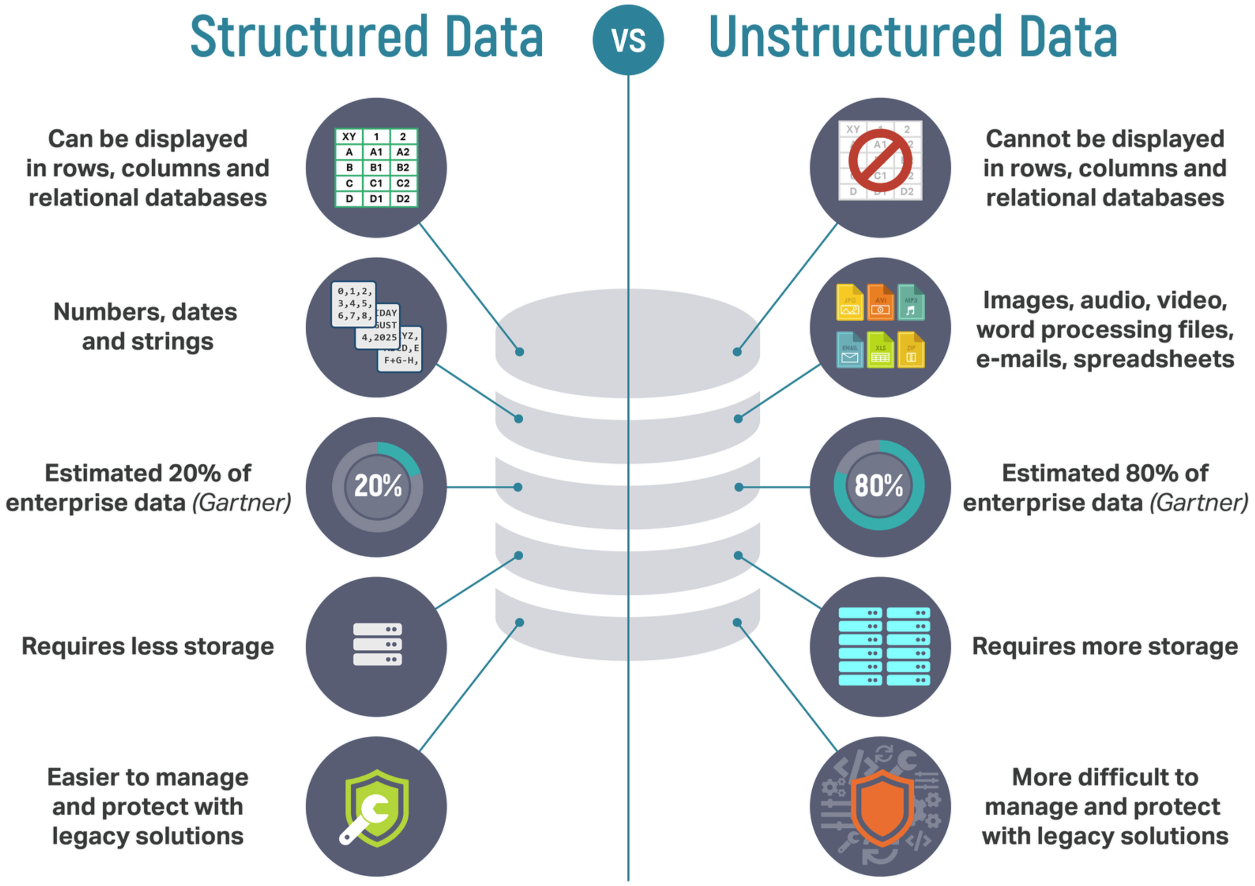
There are two types of quantitative data: continuous data and discrete data.

- **Continuous data:** this is information that can be measured. It refers to one point within a range (or continuum). Technically, continuous data can be infinitely more precise. For example, if you use a scale at home, your dog may weigh 35 pounds. But the veterinarian's scale might show more precisely that the dog weighs 35 pounds and 7.63 ounces. Other examples of continuous data include the speed of a car, the weight of a toddler, the time a train departs, and the rate of revenue growth.
- **Discrete data**: this is information that can be counted. Generally, discrete data contains integers (i.e. finite values) and cannot be more precise. For example, the number of goldfish in an aquarium is discrete since they can be physically counted and it's impossible to have 3.7 goldfish. Other examples of discrete data include number of customers, number of languages a person speaks, and number of apps on your phone.

Beyond the qualitative and quantitative categories, it is also important to be aware of how data can be **structured, semi-structured or unstructured.**

In computer science, a data structure is a particular way of organising and storing data in a computer such that it can be accessed and modified efficiently. More precisely, a data structure is a collection of data values, the relationships among them, and the functions or operations that can be applied to the data.

- **unstructured:** no predefined structure; 80-90% of data in this format; difficult to analyse
- **semi-structured:** self-describing structure; allows for flexibility; i.e. CSV, JSON and XML formats
- **structured:** ordered in rows/columns; defined by a data model; easily queried


Structured Data
VS
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
Estimated 20% of enterprise data (Gartner)
20%
Requires less storage
Easier to manage and protect with legacy solutions
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
80%
Requires more storage
More difficult to manage and protect with legacy solutions

Structured data usually resides in relational database management systems (RDBMS). Fields store length-delimited data phone numbers, Social Security numbers, or postcodes. Even text strings of variable length like names are contained in records, making it a simple matter to search. Data may be human- or machine-generated as long as the data is created within an RDBMS structure. This format is eminently searchable both with human generated queries and via algorithms using type of data and field names, such as alphabetical or numeric, currency or date.

Common relational database applications with structured data include airline reservation systems, inventory control, sales transactions, and ATM activity. Structured Query Language (SQL) enables queries on this type of structured data within relational databases. Some relational databases do store or point to unstructured data such as customer relationship management (CRM) applications. The integration can be awkward at best since memo fields do not lend themselves to traditional database queries. Still, most of the CRM data is structured.

Semi-structured data do not obey the formal structure of data models associated with relational databases or other table forms. They contain tags or markers to discrete semantic elements. The advancement in the use of the Internet increased the presence of semi-structured data where full text data is not the only type of data available for exchange of information. Semi-structured data are found in object oriented databases.

Unstructured data is essentially everything else. Unstructured data does not have a recognizable structure. It is an unorganised collection of huge data with a variety of objects that have no importance until identified in a structured form. Unstructured data has internal structure but is not structured via pre-defined data models or schema. It may be textual or non-textual, and human or machine-generated. It may also be stored within a non-relational database like NoSQL.

Typical human-generated unstructured data includes:

- **Text files:** Word processing, spreadsheets, presentations, email, logs.
- **Email:** Email has some internal structure thanks to its metadata, and we sometimes refer to it as semi-structured. However, its message field is unstructured and traditional analytics tools cannot parse it.
- **Social Media**: Data from Facebook, Twitter, LinkedIn.
- **Website:** YouTube, Instagram, photo sharing sites.
- **Mobile data:** Text messages, locations.
- **Communications:** Chat, IM, phone recordings, collaboration software.
- **Media:** MP3, digital photos, audio and video files.
- **Business applications:** MS Office documents, productivity applications.

Typical machine-generated unstructured data includes:

- **Satellite imagery:** Weather data, land forms, military movements.
- **Scientific data:** Oil and gas exploration, space exploration, seismic imagery, atmospheric data.
- **Digital surveillance:** Surveillance photos and video.
- **Sensor data**: Traffic, weather, oceanographic sensors.

When extracting knowledge from unstructured data, there is a need to convert it to a structured form, which helps in the analysis of data. For the conversion of unstructured data to structured data, a machine learning algorithm can be used, including logistic regression, linear regression and decision trees, as shown in the example of working with the data generated by IoT devices (Verma et al. 2020).

The data captured by IoT devices is produced in a mix of data formats, including structured, semi-structured, and unstructured data. This data might include analog signals, discrete sensor readings, device health metadata, or large files for images or video. Because IoT data is not uniform, no one-size-fits-all approach exists for storing IoT data.

In the context of IoT and Big Data, a type of data to be aware of is **metadata.** Technically, this is not a separate data structure, but is one of the most important elements for analysis of big data. Metadata, is data, about data. It offers additional information about a specific set of data. For example with photographs, the metadata describes when and where photos were taken. This metadata then provides fields for dates and locations which themselves can be considered structured data.

Storing and using data

In thinking about data types we also think about how data is stored internally in the computer. So for example as:

- Integers: whole numbers with no decimal or fractional parts
- Floating point: numbers that can contain a decimal or fractional part
- Character: a single text character which can be a letter, number or symbol
- Boolean: can take two possible values such as true/false or yes/no. Often stored as 0 and 1
- Date and time: the number of days or seconds passed since the 'epoch' date, normally 1/1/1970.

Changing the data type affects the precision and value of the data stored.

Data structures as, an organised collection of data types include:

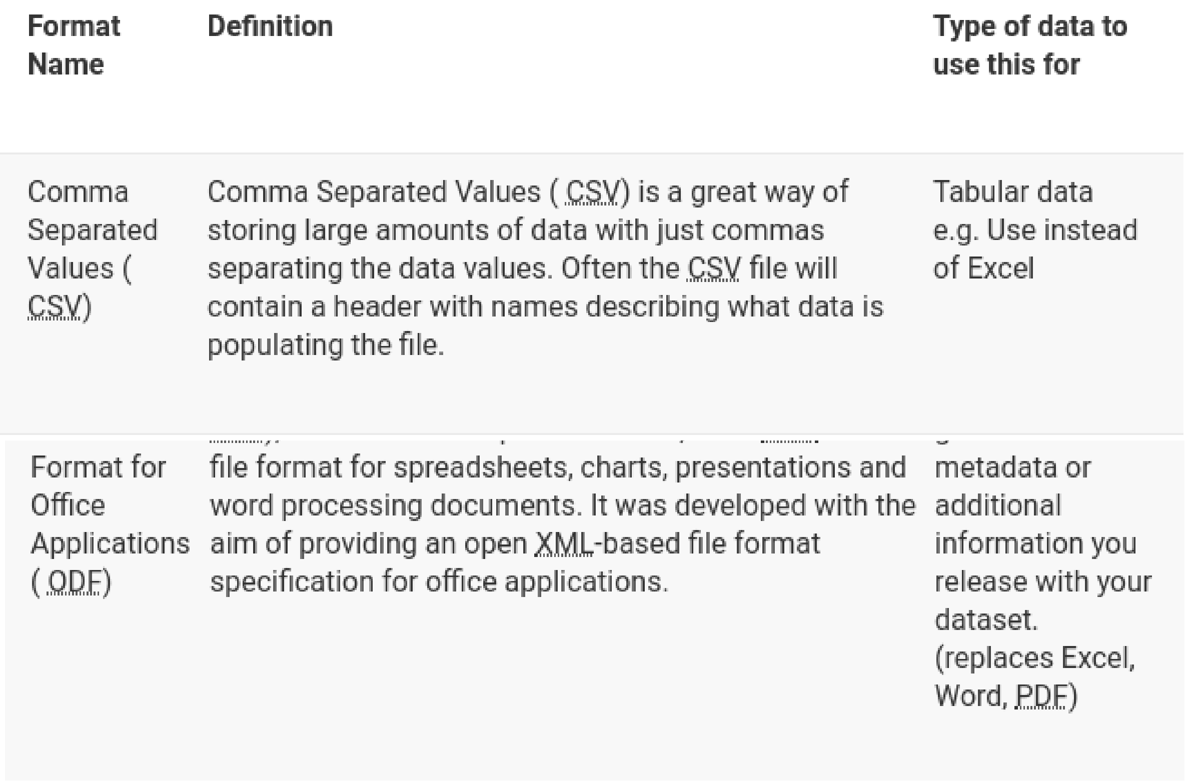
- Strings: a collection of characters combined to create alphanumeric text
- Array: a structure of a fixed size which can hold items of the same data type
- Vector: a one-dimensional array
- List: a dynamically sized structure which can contain different data types
- Data frame: a two-dimensional structure designed for holding datasets. Each column can hold different data types, but all must contain the same number of items.

By using data structures, particularly data frames, it allows faster and easier processing of properly structured datasets.

One way to change the data without changing its type is by changing the format of the display. For example, with a floating point, the stored value might be 0.4893, and you can display this value as a percentage of 48.9%. In addition to display formats, there are file formats that you should be aware of. Data is stored in different digital file formats for sharing and transporting. The most common format for tabular data is a .csv (comma separated value) file. There are many different ways of storing data, depending on its contents. Examples include databases (xml, csv, tab), geospatial shapefiles (shp, dbf), image (png, jpg), audio (mp3, wav), video (mp4, mov). The chosen format should ensure long-term access and preservation of the data.

Open data formats (ODF) are the preferred option, such as odt for documents and odp for presentation files. Open Formats are non-proprietary and platform independent. They can be accessed by anyone and do not require access to licensed software. E.g. Microsoft formats are not open as they use proprietary software.

The most appropriate format will depend on the type of data. Any type of data can be stored in an open format, but it is likely you will have to transform the data from its original format.



Format Name	Definition	Type of data to use this for
Comma Separated Values (CSV)	Comma Separated Values (CSV) is a great way of storing large amounts of data with just commas separating the data values. Often the CSV file will contain a header with names describing what data is populating the file.	Tabular data e.g. Use instead of Excel
Format for Office Applications (ODF)	file format for spreadsheets, charts, presentations and word processing documents. It was developed with the aim of providing an open XML-based file format specification for office applications.	metadata or additional information you release with your dataset. (replaces Excel, Word, PDF)

Open Data

Open data is data that can be freely used, shared and built-on by anyone, anywhere for any purpose. If you've ever checked an online weather forecast, used your smartphone's GPS to find a 24hr supermarket or calculated how much your local council paid for road repairs, you've used Open Data. For a long time, however, accessing this government data was very difficult, if available at all. The concept of Open Data is very new. It originated with the belief that the enormous amount of information routinely collected by government entities should be available to all citizens. Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. To be open, data must be technically open (available in non-proprietary formats) and legally open (in the public domain).

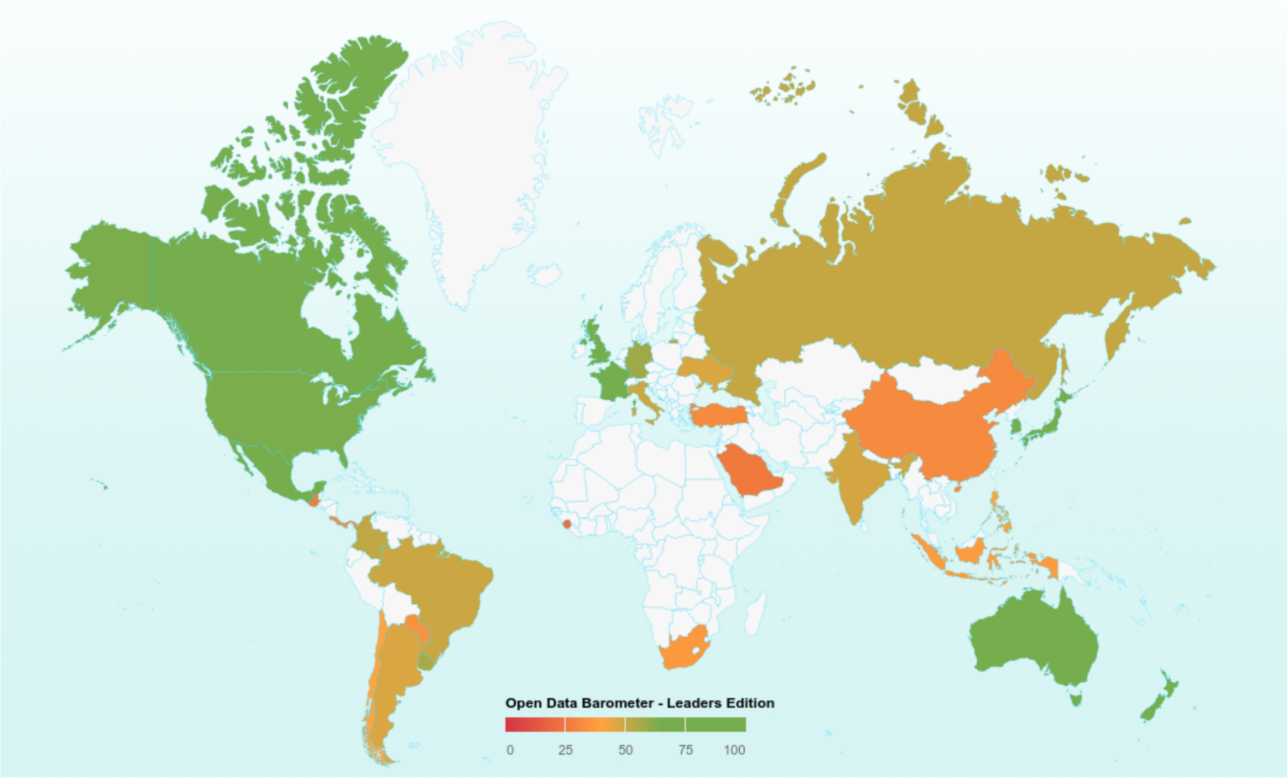
The principles behind open data make it very powerful:

- Availability and access: everyone can access the data
- Re-use and redistribution: everyone can share and reuse the data
- Universal participation: anyone can use the data

The characteristics of what makes data 'open' includes: data must be (1) complete (2) primary (3) timely (4) accessible (5) accurate and (6) machine processable and made online in persistent archives. Furthermore, (7) access should be non-discriminatory (8) data formats should be non-proprietary and (9) data licenses should be unrestricted and bear no usage cost. Data is only truly open if most of these criteria are met.

Openness is considered a good governance principle, and open government data initiatives have emerged all over the world. Open government data (OGD) are non-privacy-restricted and non-confidential data, produced with public money and made available without any restrictions on their usage or distribution. The release of OGD is stimulated by initiatives such as the international Open Government Partnership, in which more than 75 countries are participating (Open Government Partnership, 2017).

The Open Data Barometer is a global measure of how governments are publishing and using open data for accountability, innovation and social impact.



The UK is ranked top of 86 countries by the Open Data Barometer, which measures a country's readiness to secure benefits from open data, its publication of key datasets and evidence of emerging impacts from open government data.

The UK's central repository of public sector open data, data.gov.uk, contains nearly 15,000 datasets published with an Open Government License. Substantial open data resources are also published by non-government sources, such as nonprofits and community groups. (ODI, 2016).

UK companies are using government and non-government open data from a wide range of sectors. For example:

- Arup uses open data to help plan smart cities and mitigate against risk and natural disasters in the built environment.
- Doorda uses open data to help citizens and businesses discover and understand what is happening on their streets and in their local communities.
- FoodTrade maps the food supply chain system to help people buy and sell fresh produce, contributing to the creation of a fair, sustainable and local food system.
- GeoLytix combines geospatial data with domain expertise to help people make better decisions about the location of their businesses.

The benefits of open data

When data are made widely available and easy to use, the benefits can be significant: They can help streamline government services, stimulate economic opportunities, improve public safety and reduce poverty. Some of the main benefits of open data include:

- Improved public services
- Transparency and democratic control
- Ability to spot global patterns
- Measurement of government policies
- Increased government effectiveness
- Facilitates innovation

As the benefits of Open Data impact broader populations and additional useful options are discovered, governments and institutions worldwide are eager to launch new or expand existing Open Data programs. It will take time to fully understand the complexity and broad potential of Open Data, which is derived from the “open” environment of licensing. As Open Data is still in its early stages, best practices and communities are just beginning to emerge. The majority of data remains private and not accessible to the public.

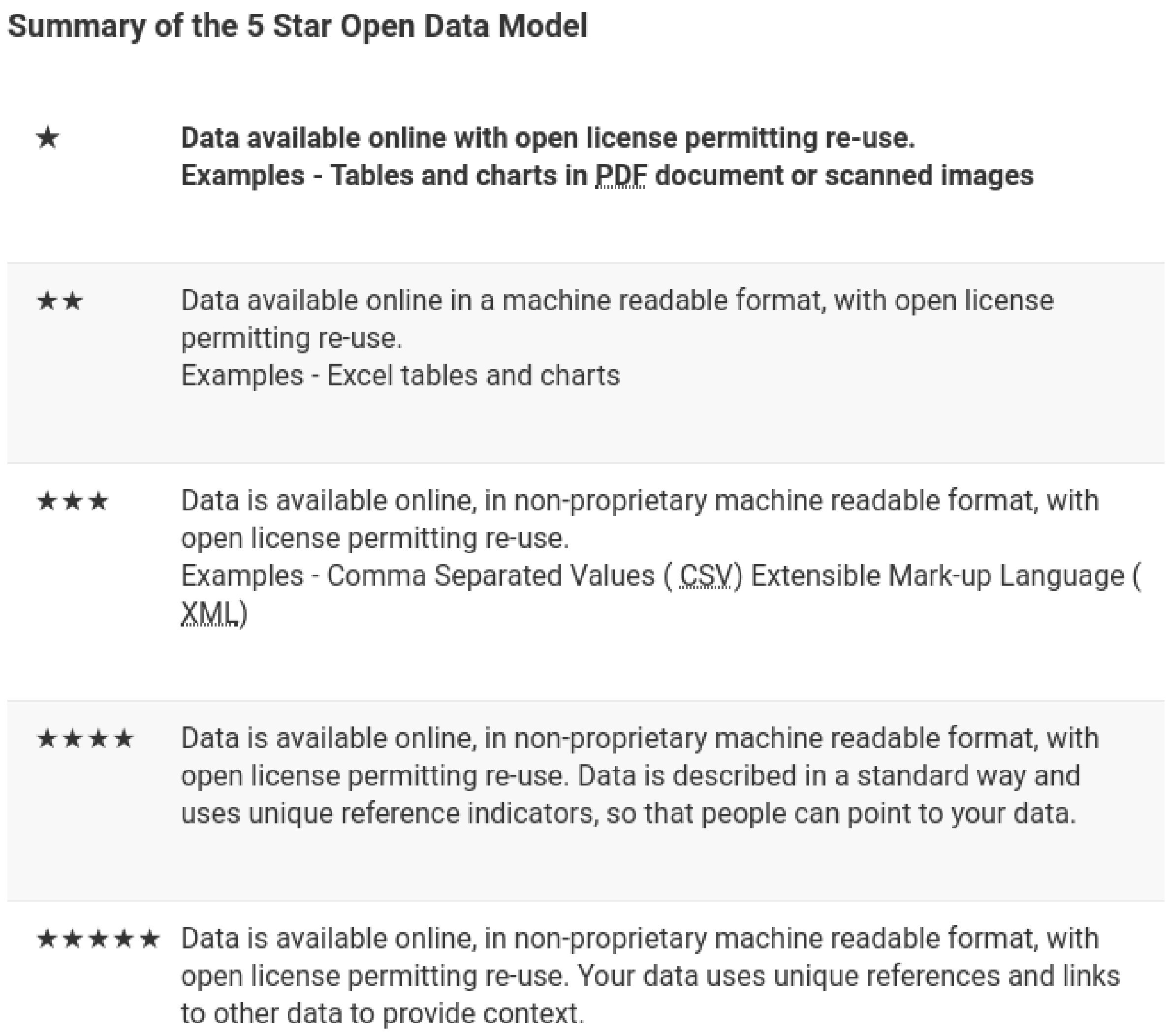
Open Data in Scotland

The Scottish Government has an open data strategy and has adopted the G8 Principles of Open Data which are:

- **Open Data by Default -** Those holding public data should make it open and available for others to re-use. Those collecting new data should make sure that releasing data for re-use is built into the process. Over time releasing data openly should become the default business practice.
- **Quality and Quantity -** The amount of public data available is huge but the data quality varies. Published data must be supported by metadata. Metadata provides information about the data itself. Good metadata allows re-users to understand the data and its limitations.

- **Usable by All -** Data should be published in a manner which supports both easy discovery and easy re-use of the data. This includes making sure it is in a format which supports re-use and it has an open licence. Data will be made available free, with defined exceptions.
- **Releasing Data for Improved Governance -** Public authorities will release data which supports delivery of better public services. They will use the data to improve the services and policies they deliver. Public authorities should aim to engage and inform the public through the release of open data.
- **Releasing Data for Innovation -** Release of data will create wider economic and societal benefits. Others will be encouraged to make use of the data and develop new products or services for non-commercial and commercial use.

The Scottish Government aspires to the 5 star model suggested by the founder of the world wide web Tim Berners-Lee and stated that by 2017, all public authorities in Scotland should be publishing their data in a format of 3 star or above.



Summary of the 5 Star Open Data Model

★	**Data available online with open license permitting re-use.** **Examples - Tables and charts in PDF document or scanned images**
★★	Data available online in a machine readable format, with open license permitting re-use. Examples - Excel tables and charts
★★★	Data is available online, in non-proprietary machine readable format, with open license permitting re-use. Examples - Comma Separated Values (CSV) Extensible Mark-up Language (XML)
★★★★	Data is available online, in non-proprietary machine readable format, with open license permitting re-use. Data is described in a standard way and uses unique reference indicators, so that people can point to your data.
★★★★★	Data is available online, in non-proprietary machine readable format, with open license permitting re-use. Your data uses unique references and links to other data to provide context.

Sources of Open Data

Open data can be found in many forms in different sectors of society. For example including the cultural sector (libraries and archives), sciences (research outputs), finance (government accounts and financial markets), weather and environmental and the charity sector.

Explore the UK Government's available open data: https://data.gov.uk

And/or the case studies available on open data in Scotland: https://www.gov.scot/publications/open-data-resource-pack/pages/12/

A wider range of sources of open data are available for viewing and downloading, and cover such areas as:

- Government and global data
- Financial and economic data
- Crime and drug data
- Health and scientific data
- Academic data
- Environmental data
- Business directory data
- Media and journalism data
- Marketing and social media data

Data Ethics

Data science provides huge opportunities to improve private and public life, as well as our environment (consider the development of smart cities or the problems caused by carbon emissions). Unfortunately, such opportunities are also coupled to significant ethical challenges. The extensive use of increasingly more data—often personal, if not sensitive (big data)—and the growing reliance on algorithms to analyse them in order to shape choices and to make decisions (including machine learning, artificial intelligence and robotics), as well as the gradual reduction of human involvement or even oversight over many automatic processes, pose pressing issues of fairness, responsibility and respect of human rights, among others. These ethical challenges can be addressed successfully.

However, it is questionable that there is a universal code of data ethics. There may be too few commonalities across the specific uses of data science to pull together a single code. Principles of data ethics that hold in medicine may not hold in finance because the social roles occupied by medical professionals and financiers differ significantly. They have meaningfully different obligations to their clients and society, and so it is reasonable to expect that their uses of big data for good and ill will similarly vary. Depending on whom you ask, ethics can have many different meanings. Some believe it has to do with their feelings, and knowing the difference between what is right and what is wrong. Others will tell you it is doing what the law requires of them. Ethics are, and mean, different things to different people depending on the situation.

Data ethics is an emerging branch of applied ethics which describes the value judgments and approaches we make when generating, analysing and disseminating data. This includes a sound knowledge of data protection law and other relevant legislation, and the appropriate use of new technologies. It requires a holistic approach incorporating good practice in computing techniques, ethics and information assurance.

The Data Ethics Framework guides the design of appropriate data use in government and the wider public sector. The Data Ethics Workbook gives an idea of the questions to ask when considering the ethical issues involved in a project in line with the principles of the Data Ethics Framework.

The Open Data Institute also has something called the Data Ethics Canvas. Which is a tool for anyone who collects, shares or uses data. It helps identify and manage ethical issues – at the start of a project that uses data, and throughout. It encourages you to ask important questions about projects that use data, and reflect on the responses. These might be:

- What is your primary purpose for using data in this project?
- Who could be negatively affected by this project?

The Data Ethics Canvas provides a framework to develop ethical guidance that suits any context, whatever the project's size or scope.

Data Ethics in business

With the emergence of new information technologies, personal data can be collected on an unprecedented scale. E-mails, instant messages, social media posts, real time geolocation data, etc, can all build up a picture of who we are as individuals. The individual data in isolation might not, but putting it all together can. Is it right that some large organisations can build up a picture of our life based on these pieces of data? By piecing together data from different sources, it is possible, but the question remains, should they do that?

Google and various other large internet-based companies use what is called data analytics. This is the practice of collecting, analysing and interpreting customer data to detect trends and patterns. From a business point of view, this information is invaluable. It allows the business to understand its customers better, and be able to change their business model accordingly. Advertisements can be more geared towards individuals and, in turn, profits can go up.

Optional: Watch this TED talk on The Future of Your Personal Data — Privacy vs Monetisation https://www.youtube.com/watch?v=JIo-V0beaBw.

As well as the legal obligations companies and organisations have regarding personal data, there are also the ethical considerations they must keep in mind when handling personal, and sometimes sensitive, data. Data sometimes needs to be shared with other people, for example in research. Many companies carry out research when they are developing new products, or gathering their customers' views. However, in the UK we have a 'Duty of confidentiality' that is based on common law, which means that when confidential information comes to the knowledge of an individual, or a group, but it would be unfair if that knowledge were disclosed to third parties, then the information cannot be shared.

Businesses and organisations should have, at the very least, some sort of data sharing policy that employees can take guidance from. Very large companies will have ethics committees that guide their principles and make decisions about what is the right thing to do with data in each situation.

Companies should take into consideration ethics when dealing with people's data, such as:

- Inform users how their data will be stored, preserved and used
- Inform users how their confidentiality will be maintained, for example by anonymising data
- Obtain consent, either written or verbal, before sharing data

Accenture (2016) outlined a list of principles for data ethics for data scientists and practitioners which can be read in full here.

1. The highest priority is to respect the persons behind the data.
2. Attend the downstream uses of datasets.
3. Provenance of the data and analytical tools shapes the consequences of their use.
4. Strive to match privacy and security safeguards with privacy and security expectations.
5. Always follow the law, but understand that the law is often a minimum bar.
6. Be wary of collecting data just for the sake of more data.
7. Data can be a tool of inclusion and exclusion.
8. As much as possible, explain methods for analysis and marketing to data disclosures.

9. Data scientists and practitioners should accurately represent their qualifications, limits to their expertise, adhere to professional standards, and strive for peer accountability.

10. Aspire to design practices that incorporate transparency, configurability, accountability, and auditability.

11. Products and research practices should be subject to internal, and potentially external ethical review.

12. Governance practices should be robust, known to all team members and reviewed regularly.

Ethical Dilemmas

Many of the ethical issues that organisations and their employees face have to do with privacy. Different scenarios will have different answers. Debating these answers can open more ethical dilemmas, and each one of you may have different answers depending on what you think the right thing to do is. Most people will be aware of the The Ethical Dilemma of the Facebook Privacy ‘Breach’, relating to the use of data by the company Cambridge Analytica. You can find some other articles relating to Facebook and data ethics here:

https://www.theguardian.com/technology/2016/jun/17/facebook-ethics-but-is-it-ethical

https://theconversation.com/uk/topics/facebook-ethics-40179

Beyond the more obvious examples of illegal activity and the use of data in unethical ways, an ethical dilemma can also arise where there is no real ‘harm’ done. Consider the following example: A waiter in a Dubai hotel overheard a man musing with his wife, who was in a wheelchair, that it was a shame he could not get her down to the beach. The waiter told the management, who passed this information on, and the next afternoon there was a wooden walkway down the beach to a tent that was set up for them to have dinner in. In that situation, the customer experience was the only thing that mattered to the hotel, but the fact that the information had been obtained by overhearing a private conversation was put to the side. This is where the question of ethics comes in. Is it okay to do something for the benefit of a customer (or even the benefit of the organisation) without the customer being aware that you are using their data to do so?

References

Accenture. 2016. Universal Principles of Data Ethics: 12 guidelines for developing ethics codes. Available at: https://www.accenture.com/_acnmedia/pdf-24/accenture-universal-principles-data-ethics.pdf

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. 2007. Dbpedia: A nucleus for a web of open data. In The semantic web (pp. 722-735). Berlin: Springer.

Charalabidis, Y. et al. 2018. The Open Data Landscape. In: The World of Open Data: Concepts, Methods, Tools and Experiences. (pp.1-9). Berlin: Springer.

Enterprise Big Data Framework. 2019. Available at: https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/

Floridi, L. and Torredo, M. 2016. What is Data Ethics? Available at: https://royalsocietypublishing.org/doi/full/10.1098/rsta.2016.0360

Open Data Institute. 2015. Open data means business: UK innovation across sectors and regions. London, UK. Available at: https://theodi.org/article/research-open-data-means-business-pg5/

Open Data Institute. 2019. The Data Ethics Canvas Available at: https://theodi.org/article/data-ethics-canvas/

Pickell, D. 2019. 50 Best Open Data Sources, G2 Learning Hub. Available at: https://learn.g2.com/open-data-sources

Ruijer, E., Grimmelikhuijsen, S., van den Berg, J. and Meijer, A. 2018. Open data work: understanding open data usage from a practice lens. International Review of Administrative Sciences. 86(1): 3-19.

Scottish Government. 2015. Open Data Resource Pack. Available at: https://www.gov.scot/publications/open-data-resource-pack/pages/7/

UK Government, Data Ethics Framework, Available at: https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework

Uria-Recio, F. 2018. 5 Principles for Big Data Ethics: Towards Data Science. Available at: https://towardsdatascience.com/5-principles-for-big-data-ethics-b5df1d

Taylor, C. 2018. Structured vs. Unstructured Data. Datamation. Available at: https://www.datamation.com/big-data/structured-vs-unstructured-data.html

Verma, S. et al. 2020. An Unstructured to Structured Data Conversion using Machine Learning Algorithm in Internet of Things (IoT). 3rd International Conference on Innovative Computing And Communication. Available at SSRN: https://ssrn.com/abstract=3563389 or http://dx.doi.org/10.2139/ssrn.3563389