Data Management
Review

Dr Emma Murphy

Week 11, April 2021

OLLSCOIL TEICNEOLAÍOCHTA
BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL
UNIVERSITY DUBLIN

Overview

- Module Revision

- CA2 Q&A

Module Revision

Data Management Topics
Ethical Tools and Frameworks
Data Law and Regulation
Data Protection and Data Privacy
Data Governance
Data Management and the data lifecycle
Data Quality
Data Security
Ethics and Bias in Data

What is this module about?

The aim of this module is to **analyse** and **evaluate** the role of data management in an organisation or sector and the various **roles, processes, tools, techniques** and **requirements** that are involved in the data management function.

Data is an asset

- “In today's digital infrastructure, data has become a company asset. It has a status similar to that of oil, steel and railways during the Industrial Revolution”. (Hasselbach and Tranberg, 2017)

Data Management Definitions

- ”The professional discipline of data management addresses the challenges of managing data and information as an enterprise asset, to better deliver value to an organization and its stakeholders. As with other asset management disciplines, this requires considering managing the data asset throughout the life cycle and considering its proper handling, from planning for acquiring or creation of the data through its maintenance and use and into its disposition once its purpose is concluded.”

- (O'Keefe and O’Brien, 2018)

Data impacts people

- Information that allows us to identify a person and make a determination about their eligibility for a loan,
- Or information that trains an artificial intelligence system that provides sentencing recommendations to judges,
- Or whether it is information about the performance of a car's engine in environmental impact tests,
- (O'Keefe and O'Brien, 2018)

Ethics

- At its simplest, ethics is a system of moral principles. They affect how people make decisions and lead their lives.
- Ethics is concerned with what is good for individuals and society and is also described as moral philosophy.
- The term is derived from the Greek word *ethos* which can mean custom, habit, character or disposition.

http://www.bbc.co.uk/ethics/introduction/intro_1.shtml

Ethical Frameworks

- Perhaps you believe that that individuals only ever act in their own self-interest (egoism)
- You believe it sensible to to take a decision that has the best possible outcome for all concerned (utilitarianism).
- Or you may believe that there are some universal rights which all humans have (deontology).
- Or perhaps you do things each day because you think it makes you a good person and not because of any duty or consequences that this action might have (virtue ethics).

(SOAS, 2015)

First Principles Ethical Test

- Does the outcome of your design/algorithm/process outcome contribute positively to ‘the good’, or positive preservation of human rights?

- (O'Keefe and O’Byrne, 2018)

First Principles Ethical Test

What is the outcome?

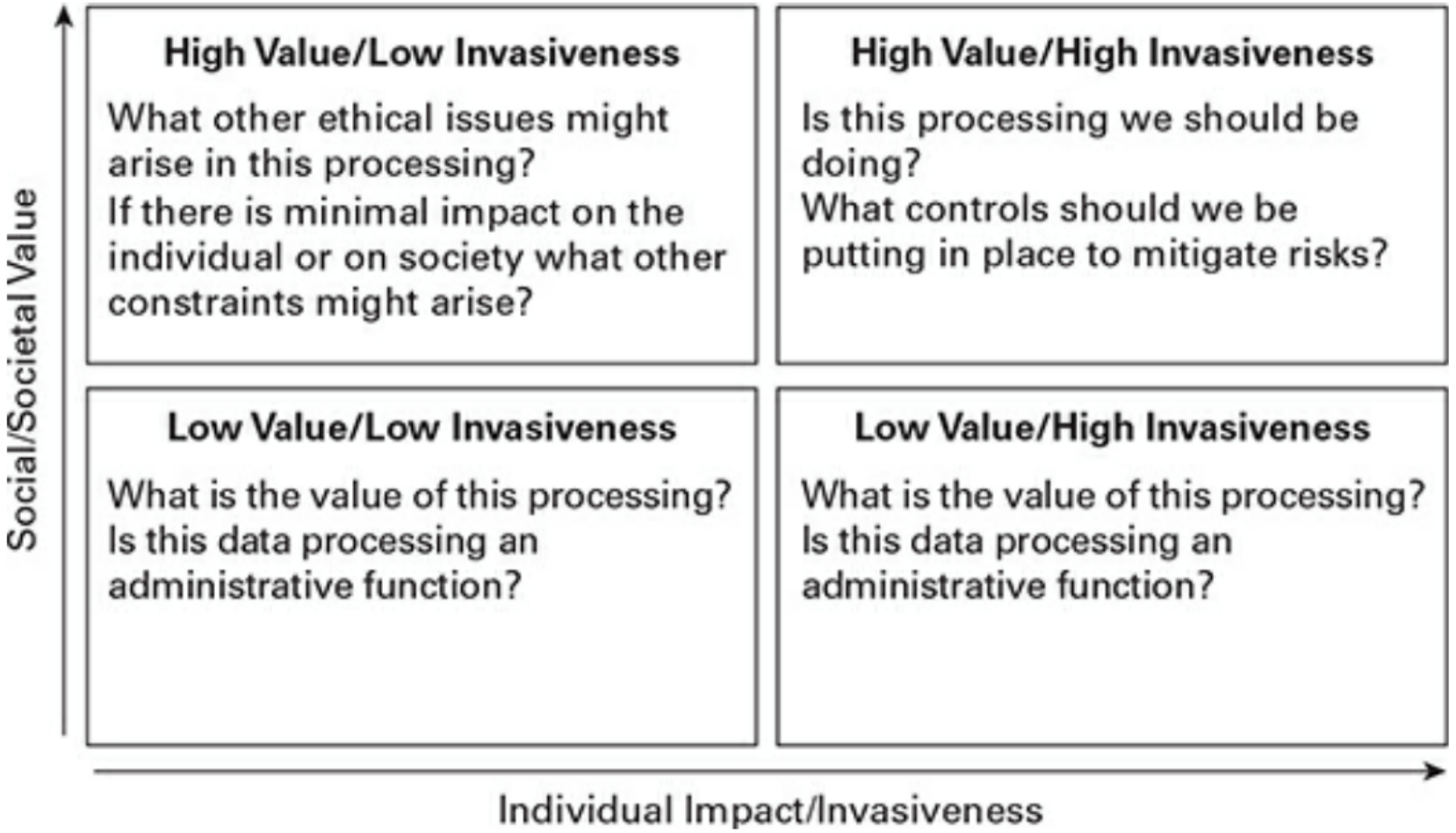
1. Does it preserve or enhance human dignity?
2. Does it preserve the autonomy of the human?
3. Is the processing necessary and proportionate?
4. Does it uphold the common good?

- does the outcome violate any of these four points?

Ethical Dilemmas in Data management

- If processing is significantly invasive or harmful to the rights or freedoms of individuals, but delivers an equally significant social benefit, it may well be that there is a valid ethical trade-off to be made.
- If, however, the processing is invasive to the individual but of limited value to society then the trade-off is less compelling. Also, even in scenarios where the trade-off is skewed against the individual, organizations might be able to take some action to redress that balance through education, communication or other mechanisms.

Social vs Individual



Law and Technology

- “Law likes what is physical and present within boundaries of states, regions, and nations. It likes less what is intangible, such as that which is digital” (Lasprogata, 2019)

Law and Technology

- “Intensifying this misalignment is the fact that technology evolves at a rapid pace. Law, on the other hand, moves as slow as a snail. It is always catching up to technology” (Lasprogata, 2019)

Data Privacy

- Privacy law is not new. Privacy and information privacy as concepts are firmly linked to the ethical imperative to respect human rights.

(DAMA, 2017)

OECD

- In 1980, the Organization for Economic Co-operation and Development (OECD) established Guidelines and Principles for Fair Information Processing that became the basis for the European Union's data protection laws.

(OECD, 2011)

2018 GDPR

- Originally proposed by the European Commission in January 2012, the GDPR (Regulation (EU) 2016/679) was adopted by the European Parliament in April 2016 and published in the Official Journal of the European Union on 4 May 2016. Following a two-year transition period, it was enforced in all 28 EU member states on 25 May 2018.

https://www.itgovernance.eu/en-ie/data-protection-ie

GDPR – Personal data

- Under the GDPR, *personal data* is data that relates to or can identify a living person, either by itself or together with other available information. Examples of personal data include a person's name, phone number, bank details and medical history.
- A *data subject* is the individual to whom the personal data relates.
- Organisations that collect or use personal data are known as *data controllers* and *data processors*.

https://www.citizensinformation.ie/en/government_in_ireland/data_protection/overview_of_general_data_protection_regulation.html

GDPR Penalties

Serious infringements

- For the most serious infringements (for example, not having sufficient customer consent to process data or violating the core of privacy by design concepts) organisations can be fined up to 4% of their annual global turnover or €20 million, whichever is greater.

https://www.citizensinformation.ie/en/government_in_ireland/data_protection/overview_of_general_data_protection_regulation.html

ePR

- In January 2017, the European Commission proposed a new Regulation on Privacy and Electronic Communications (ePR) as part of its digital single market strategy.
- The ePR will replace the 2002 ePrivacy Directive (the 'cookies law') and all member state laws that implemented it – including Ireland's ePrivacy Regulations 2011.

https://www.itgovernance.eu/en-ie/eprivacy-regulation-epr-ie

So what is personal data?

- According to GDPR:
- ‘Personal data’ means any information relating to an identified or identifiable natural person (‘data subject’).
- “An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.”

Personal Data - Identifiers

- An individual is 'identified' or 'identifiable' if you can distinguish them from other individuals.
- A name is perhaps the most common means of identifying someone. However whether any potential identifier actually identifies an individual depends on the context.
- A combination of identifiers may be needed to identify an individual.
- The GDPR provides a non-exhaustive list of identifiers, including:
    - name;
    - identification number;
    - location data; and
    - an online identifier. ('Online identifiers' includes IP addresses and cookie identifiers which may be personal data.)
- .

https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data/

Anonymisation

- In order to be truly anonymised under the GDPR, you must strip personal data of sufficient elements that mean the individual can no longer be identified.
- However, if you could at any point use any reasonably available means to re-identify the individuals to which the data refers, that data will not have been effectively anonymised but will have merely been pseudonymised. This means that despite your attempt at anonymisation you will continue to be processing personal data.

https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/

Pseudonymisation

- Pseudonymisation may involve replacing names or other identifiers which are easily attributed to individuals with, for example, a reference number.
- Whilst you can tie that reference number back to the individual if you have access to the relevant information, you put technical and organisational measures in place to ensure that this additional information is held separately.

Pseudonymisation

- Pseudonymising personal data can reduce the risks to the data subjects and help you meet your data protection obligations.
- However, pseudonymisation is effectively only a security measure. It does not change the status of the data as personal data. GDPR makes it clear that pseudonymised personal data remains personal data:
- "…Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person…"

What are the lawful bases for processing?

Basis	Description	Example
Consent	Freely given, specific and informed, unambiguous, provides a clear indication of wishes.	Informed Consent
Contract	Covers processing of personal data of employees.	Employee Contracts
Legal Obligation	This must be enshrined in EU or member state law.	Delivery of water, health care.
Vital Interest	This has to be on a common sense basis, prioritising the best interest of the data subject.	Life or death situation, i.e. accessing someone's health data to give lifesaving aid.
Public Interest	This covers official authorities.	Revenue Commissioners.
Legitimate Interest	Processing is necessary to the controller or processor, balanced against the data subject rights.	A bank processing customer details.

Controllers and Processors

- Controllers are the main decision-makers – they exercise overall control over the purposes and means of the processing of personal data.
- If two or more controllers jointly determine the purposes and means of the processing of the same personal data, they are joint controllers. However, they are not joint controllers if they are processing the same data for different purposes.
- Processors act on behalf of, and only on the instructions of, the relevant controller.

https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/

Who is the controller?

- To determine whether you are a controller or processor, you will need to consider your role and responsibilities in relation to your data processing activities.
- If you exercise overall control of the purpose and means of the processing of personal data – ie, you decide what data to process and why – you are a controller.
- If you don't have any purpose of your own for processing the data and you only act on a client's instructions, you are likely to be a processor – even if you make some technical decisions about how you process the data.

https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/

Activity

- Identify who is the data controller and the data processor in this example and explain why.

- *A brewery has many employees. It signs a contract with a payroll company to pay the wages. The brewery tells the payroll company when the wages should be paid, when an employee leaves or has a pay rise, and provides all other details for the salary slip and payment. The payroll company provides the IT system and stores the employees' data.*

Activity

- Identify who is the data controller and the data processor in this example and explain why.
- *Your company/organisation offers babysitting services via an online platform. At the same time your company/organisation has a contract with another company allowing you to offer value-added services. Those services include the possibility for parents not only to choose the babysitter but also to rent games and DVDs that the babysitter can bring. Both companies are involved in the technical set-up of the website. In that case, the two companies have decided to use the platform for both purposes (babysitting services and DVD/games rental) and will very often share clients' names.*

Going beyond the legislation

- “Ethical companies in today's big data era are doing more than just complying with data protection legislation. They also follow the spirit and vision of the legislation by listening closely to their customers.” (Hasselbach and Tranberg, 2017)

Ethics and government

- At a civil or national level, government and laws facilitate societal enforcement of the agreed ethical norms of a nation or civil body.
- Laws and the legal system do not determine ethics but they codify decisions made as to what behaviour should be enforced, and provide a path for mediation and escalation in disputes.
- We take decisions based on our individual ethical views, the codified laws, and any alleged infringement is decided on by an independent judiciary.

- (O'Keefe and O'Brien, 2018)

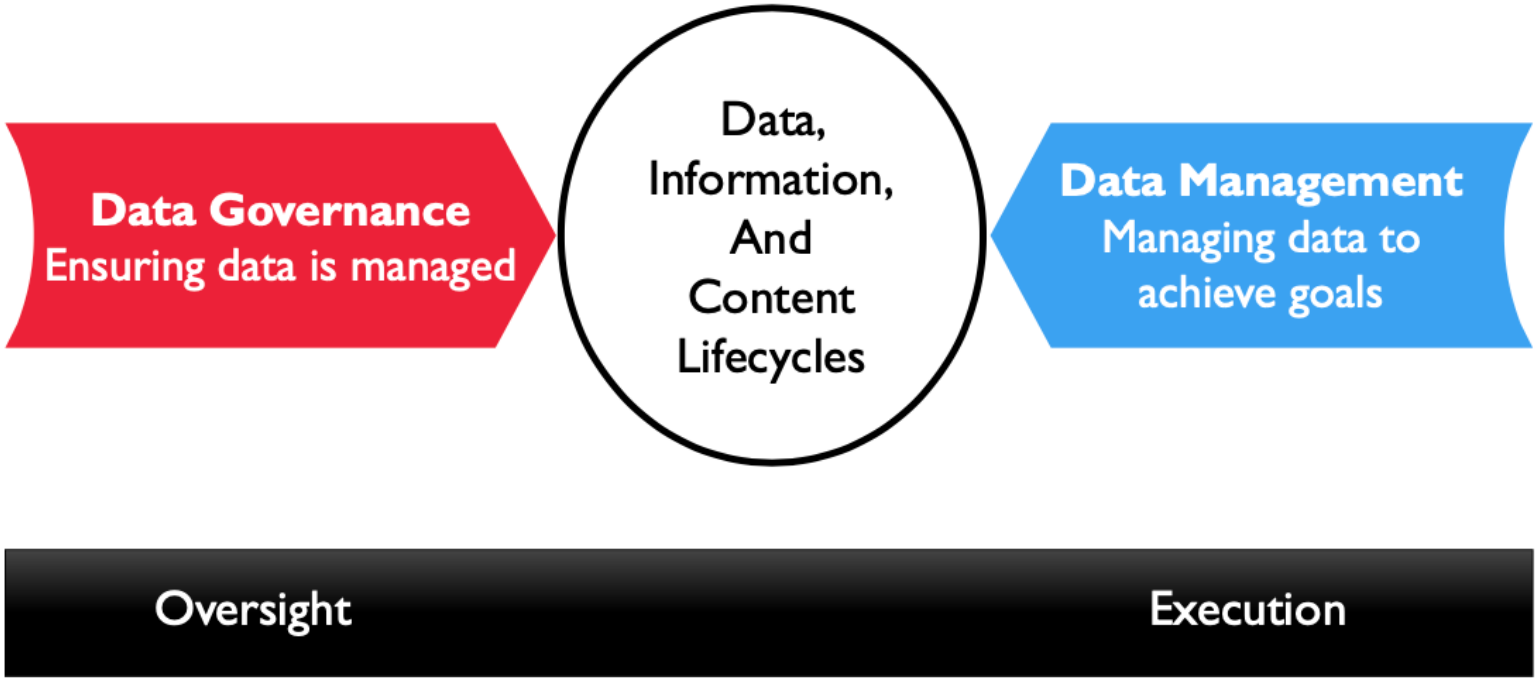
Data Governance and civil government

- Data governance is an analogous system that enables an organization to determine what is considered proper action regarding data and data processes.

- It enables clear definitions, decision-making rights and responsibilities, and provides an escalation path or process of mediation or remediation when people have questions or disputes as to what should be done with what data, by whom and under which circumstances.
- (O'Keefe and O'Brien, 2018)

Separation between oversight and execution

- Just as an auditor controls financial processes but does not actually execute financial management, data governance ensures data is properly managed without directly executing data management. Data governance represents an *inherent separation of duty between oversight and execution.*



(DAMA, 2017)

Data Stewardship

- A steward is a person who has a responsibility to manage the property of another person and may be held to account if that property is lost, stolen, damaged or misused.
- If you have ever rented an apartment or a house, you will have found language in your lease agreement that made you a steward of the property.

Data stewardship

- Data stewardship is the term used to describe accountability and responsibility for the use of data assets in an organization.

- It can be a formal assignment or it can evolve through people organically trying to help an organization better manage its information. At the heart of the data stewardship concept is an ethical value that data is only held on trust and the processing being performed is for the benefit of someone else. This is consistent with the 'customer-centric' ethos of quality management systems.

(DAMA, 2017)

Stewards

- Ideally, stewards are ‘self-selecting’ and the structures you implement in your data governance framework will recognize their efforts and enable them to be more successful and to contribute more.
- Even where they are appointed, the structures that are put in place for data governance in your organization should work to support effective alignment around common standards, both in the context of technical standards and definitions but also in the context of ethical standards and values in the organization. (“Tone from the top”)

(O’Keefe and O’Brien, 2018)

DG Principles

- Leadership and strategy: Successful Data Governance starts with visionary and committed leadership. Data management activities are guided by a data strategy that is itself driven by the enterprise business strategy.
- Business-driven: Data Governance is a business program, and, as such, must govern IT decisions related to data as much as it governs business interaction with data.
- Shared responsibility: Across all Data Management Knowledge Areas, data governance is a shared responsibility between business data stewards and technical data management professionals.

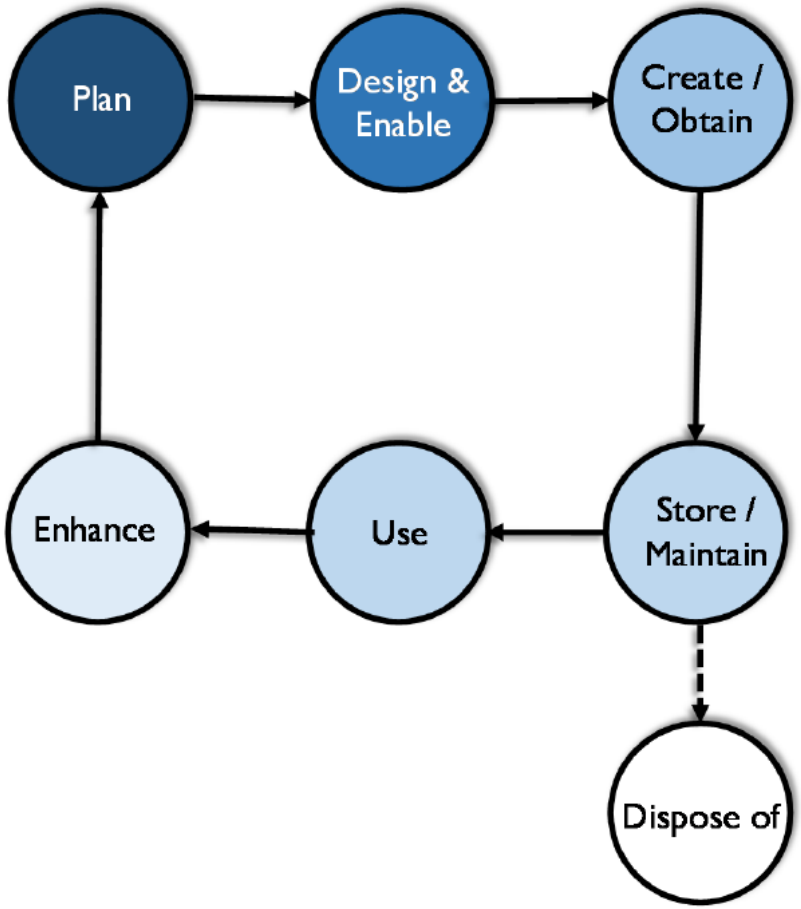
(DAMA, 2017)

Data Lifecycle

- As data is used or enhanced, new data is often created, so the lifecycle has internal iterations.
- Data is rarely static.
- Managing data involves a set of interconnected processes aligned with the data lifecycle.

(DAMA, 2017)

DM Lifecycle



Data disposal

- Non-value-added information should be removed from the organization's holdings and disposed of to avoid wasting physical and electronic space, as well as the cost associated with its maintenance.
- There is also risk associated with retaining records past their legally required timeframes. This information remains discoverable for litigation.

(DAMA, 2017)

Data Quality

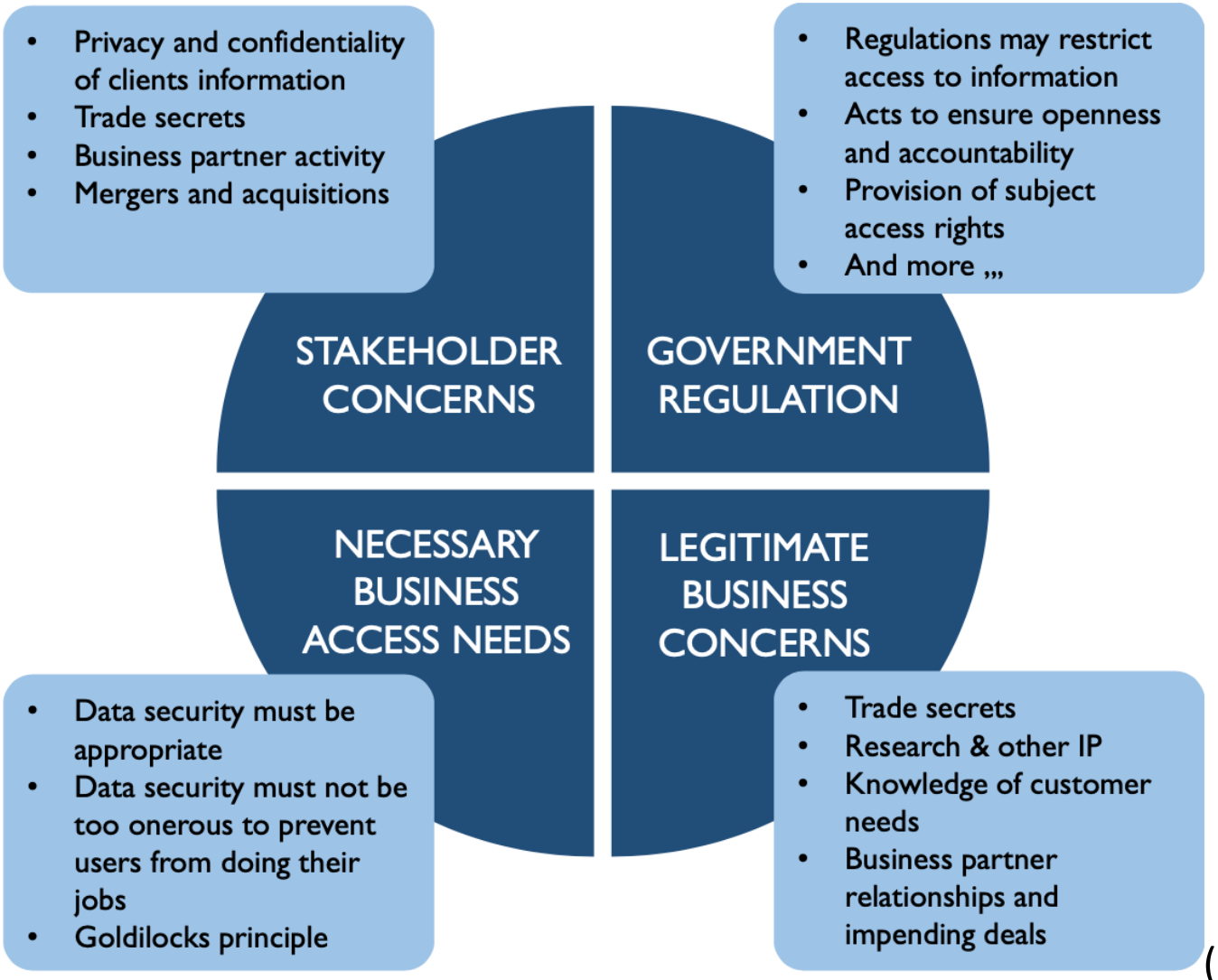
- Just as with ethics, information quality management requires you to have some mechanism for defining and measuring what ‘good’ is.
- It also requires formal processes and methods to plan for, design for, and ensure controls for good-quality information.

(O’Keefe and O’Brien, 2018)

Dimensions of Data Quality

- Accuracy
- Completeness
- Consistency
- Integrity
- Reasonability
- Timeliness

Data Security Requirements



(DAMA, 2017)

Encryption

- Hash – Hash encryption uses algorithms to convert data into mathematical representation. The exact algorithms used and order of application must be known in order to reverse the encryption process and reveal the original data.
- Symmetric - Symmetric-key algorithms are algorithms for cryptography that use the same cryptographic keys for both encryption of plaintext and decryption of ciphertext. The keys may be identical or there may be a simple transformation to go between the two keys.
- Private Key – Private Key encryption uses one key to encrypt the data. Both the sender and recipient must have the key to read the original data. Data can be encrypted one character at a time or in blocks. Common private key encryptions include DES, 3DES, AES and IDEA.
- Public Key – In Public Key encryption, the sender and receiver have different keys. The sender uses a public key that is freely available, and the receiver uses a private key to reveal the original data.

Data Security

- Know your obligations
- https://www.dataprotection.ie/en/organisations/know-your-obligations/data-security-guidance

Data Security

- Do we know what types of personal data we hold:
- electronically (including less obvious data such as CCTV images)?
- on paper?
- Can we justify the collection of this information?
- Why do we collect it?
- What it is used for?
- What are the risks?
- How long do we hold it?
- Who has access to it?
- To whom do we disclose it?
- Where will the data be stored?
- Is it held securely?
- How we dispose of the data?

https://www.dataprotection.ie/en/organisations/know-your-obligations/data-security-guidance

Data and discrimination

- Machine learning classifiers train themselves on historical datasets
- Any problems with data become problems for the classifiers
- These problems lead to exclusion or poor representation or poor performance especially when dealing with minorities

Data Bias

- Datasets can exclude underrepresented people from consideration
- Exclusion leads to bad design, poor policies and technology that doesn't work
- Sometimes the exclusion is deliberate albeit for good reasons
- New datasets that are reflective of people with disabilities need to be developed or they will be left out

Preventing Discrimination

Recommendations

1. Active Inclusion
2. Fairness
3. Right To Understanding
4. Access to Redress

(World Economic Forum, 2018)

Remedies

- Fairness initiatives e.g. IBM Fairness 360
- Balancing Techniques e.g. SMOTE
- (Synthetic Minority Oversampling Technique)
- Synthetic Dataset Generation
- Personas

CA2

- Q and A……