MONASH University

FIT1043 Introduction to Data Science

Week 11: Data Management and Data Governance

Ts. Dr. Sicily Ting

School of Information Technology Monash University Malaysia

With materials from Wray Buntine, Mahsa Salehi

Week 10 Coverage
Big Data Processing

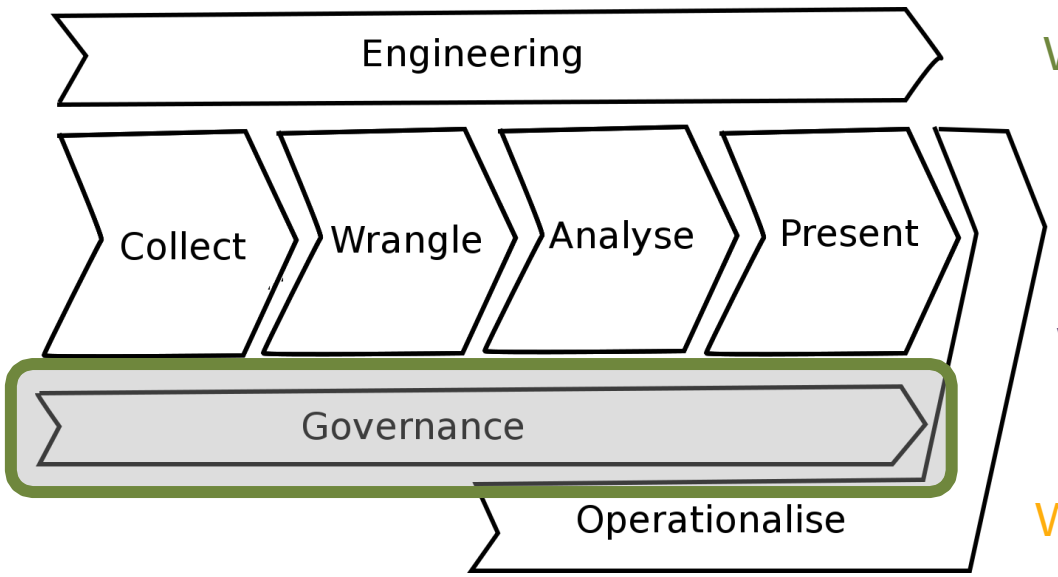
MONASH University

Week	Activities	Assignments
1	Overview of data science	
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	Assignment 2
9	Characterising data and "big" data	
10	Big data processing	
11	Data Management and Data Governance	
12	Industry guest lecture (tentative)	Assignment 3

MONASH University

Week 1

Overview of data science



Weeks 9-10

Week 4

Week 3

Weeks 5-7

Week 11

Weeks 2 & 8

Tools for data science

Data Governance

Week 11 Outline

- Data Management and Data Governance?
- Data Access and Protection
- Regulatory requirements - Compliance

MONASH University

Learning Outcomes

Week 11

By the end of this week you should be able to:

- Differentiate the data management requirements from an internal (data lifecycle) and external (data value chain) perspectives
- Understand the conflicting business and legal objectives, and
- The relationship between ethics, privacy, storage, security and analysis.

MONASH University

Data Management and Data Governance

MONASH University

Why Manage Data?

In short …

The data is very **valuable**, data collection is usually time consuming and hard

Large amount of data and documents are being generated with high growth rate

Multiple sources of data (general business documents, ERP systems, etc.)

Enterprise resource planning (ERP) refers to a type of software that organizations use to manage day-to-day business activities such as accounting, procurement, project management, risk management and compliance, and supply chain operations.

http://www.blue-pencil.ca/what-is-data-management-and-why-it-is-important/

Data Management

Needed to manage data when the volume becomes large and a strategy is required.

- Important in order to maximize business requirements of balancing the cost and the need.
- Strategies such as the retention period, access mechanism, the archive storage, or the format of various types of data during the various phases of the data lifecycle.

There is also the management of data security, where assessment on the risk and minimizing or mitigating the risks are required.

The approaches to the data management is a continuous monitoring as it will change over time. New technology and services will impact data management.

Data Management

In short …

Data management is the development, execution, and supervision of

- plans,
- policies,
- programs, and
- practices

that ***control***, ***protect***, ***deliver*** and ***enhance*** the **value** of data and information assets.

The Data Management Association or DAMA, defines data management as "the development of architectures, policies, practices, and procedures to manage the data lifecycle."

Data Management

Data management includes the following topics:

- data security
- data sharing
- data destruction
- data reference and master data management
- data warehousing and business intelligence management
- document and record storage
- records management
- data governance
- data architecture
- database management
- data quality management
- contact data systems

http://www.blue-pencil.ca/what-is-data-management-and-why-it-is-important/

MONASH University

Importance of Data Management

This is related to our Data Science pipeline.

See “How to avoid a data management nightmare”, a video created by NYU Health Sciences Library

Data Management and Data Science

Medical informatics: for predicting fungal infections from nursing notes, the team needs to abide by confidentiality and security

Internet advertising: what implicit and explicit data is stored about a user

Data Governance

Data is an asset to businesses and must be protected and managed.

- **Handling and usage of the data**.

Issues handled by data governance:

- How data is organized, *protected*, and accessed in the various environments,
- Who are allowed to ***access*** which portion of the data (privacy and confidentiality),
- Who will manage the data and be accountable for it, and
- Policies to maintain ***compliance*** with various laws & regulations, such as GDPR and PDPA.

Data Governance

In short …

Data governance focuses more narrowly on the issues surrounding *access*, *protection*, *compliance* and *usage* with the goal of bringing maximum benefits to a business or organization.

Data Management is the management of the internal data lifecycle, while data governance is related to the usage to provide the value of the data but with considerations on various aspects.

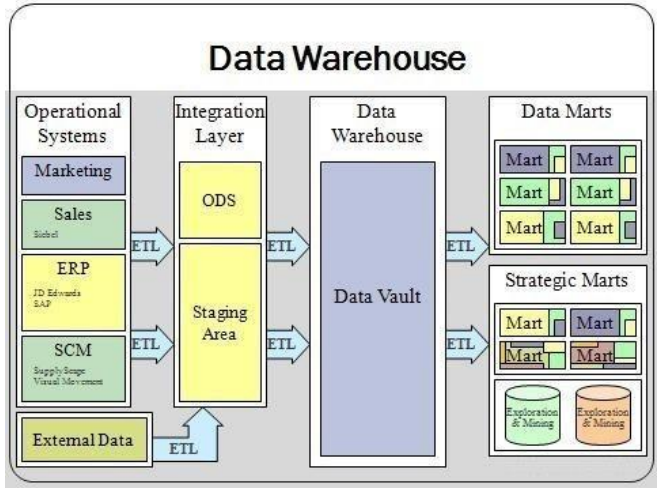
Data Access and Protection

MONASH University

Organizational Data Management Plan

Deals with issues:

- Integration and data warehousing
  - See Data Warehousing- an Overview (Youtube, 2:44 - 4:45- 5:33)



- Replication and persistence
- Standardising the vocabulary used across the organisation, e.g., job titles
- Security!

MONASH University

Privacy versus Confidentiality

Privacy is (for our purposes) having control over how one shares oneself with others, e.g. closing the blinds in your living room

Confidentiality is information privacy, how information about an individual is treated and shared, e.g. excluding others from viewing your search terms or browse history

Security as the protection of data, preventing it from being improperly used, e.g. preventing hackers from stealing credit card data

Social Media and Loss of Confidentiality

See: "The curly fry conundrum: Why social media 'likes' say more than you might think" by Jennifer Golbeck (TED)

- Target is predicting which women is pregnant from their purchases
- Many things can be predicted from Facebook "likes"

Implicit data: Data that are not explicitly stored but inferred with reasonable precision from available data

Explicit data is any information a consumer actively provides on their own. Common examples include their name, gender, email address and home address, and they're typically offered up during a customer's first transaction, or when signing up to receive a newsletter or other information through a site.

MONASH University

Confidentiality

Read “Empower consumers to control their privacy in the Internet of Everything” by Carla Rudder (blog)

For many apps or services, you must either accept their data sharing policies or you can’t use their services fully

There could be an agent to interact in a narrative form with individual consumers,

- For instance the app might ask: ‘Are you willing to share your health data with company X?’

Regulatory Requirements Compliance

MONASH University

Compliances and Regulations

Ethics: the moral handling of data

There should be regulations in place to ensure that confidentiality is protected

The process of ensuring you meet regulations is called compliance

Compliances and Regulations

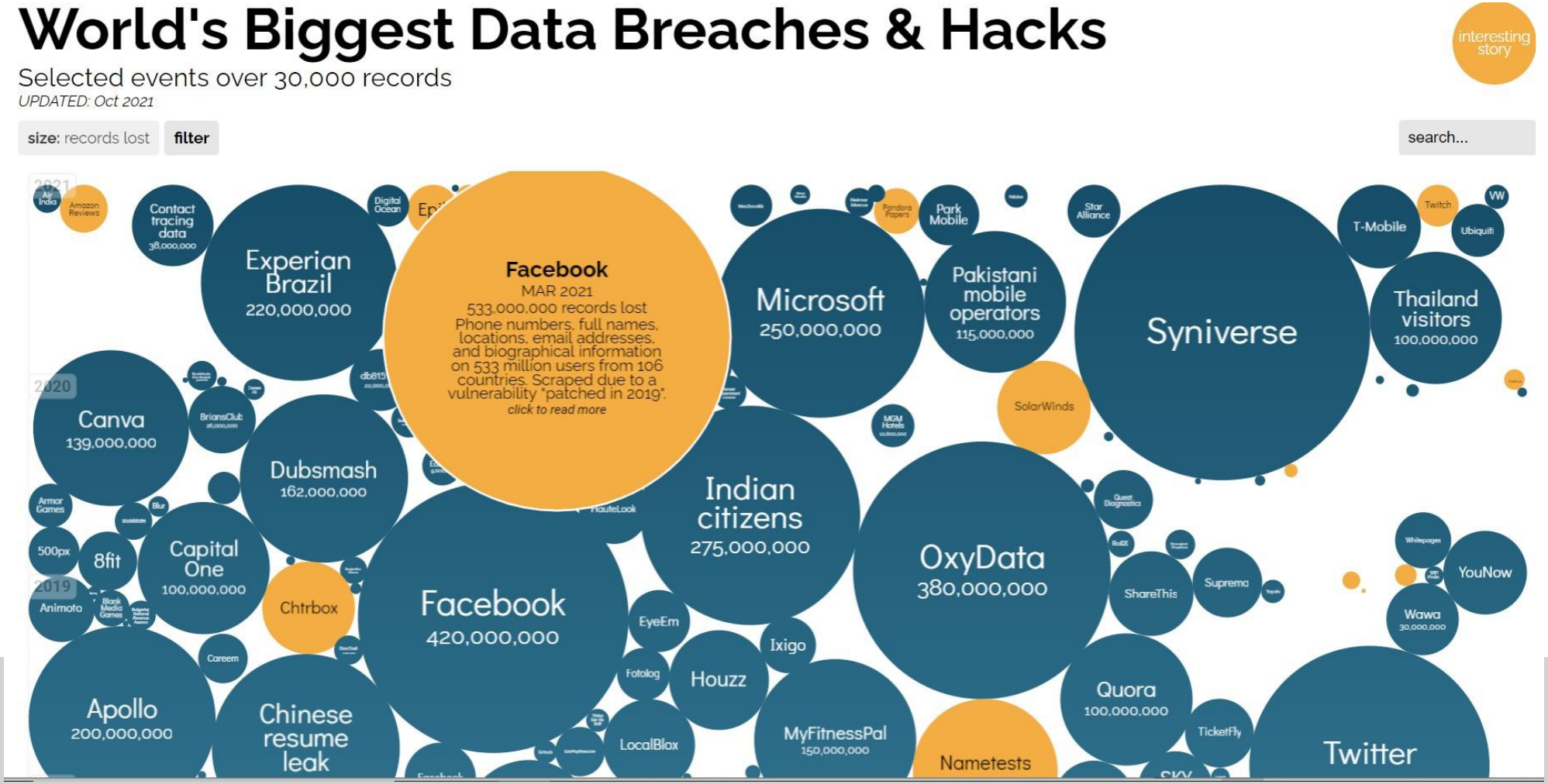
Example

PCI (Payment Card Industry) standard

- Aims to reduce credit card fraud
- By placing specific regulations, e.g., credit card information should be stored only in encrypted format.
- Companies who handle credit card have to comply with PCI standards
- Audit (validation of compliance) is done annually

Some Food for Thought

https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/



PDPA (9 Obligations)

Personal Data Protection Act

Consent: needed before personal data is collected, used, or disclosed

Purpose limitation: an organisation must inform an individual of its purpose for collecting, using, or disclosing personal data; also, the collected data must not be used for anything other than the initial intended purpose.

Notification: individuals must be notified of the purpose before they may give their consent to have their personal data collected, used, or disclosed.

Access and correction: individuals have the right to request access to their personal data in an organisation's possession or control, and be allowed to correct any error to his/her personal data.

Accuracy: an organisation should make reasonable effort to collect accurate and complete personal data, especially if any decisions made using the personal data affects the individual, and if the personal data will be disclosed to another organisation.

PDPA (9 Obligations)

Personal Data Protection Act

Protection: reasonable security arrangements must be made to preventunauthorised access, use, disclosure, copying, modification, and disposal of personal data in an organisation's possession or control

Retention limitation:An organisation may only keep personal data until a certain period, after which it must remove or delete documents containing such permanently.

Transfer limitation: personal data may not be given outside of the nation unless the recipient country has data protection standards commensurate to that of the PDPA

The National Do Not Call Registry: Names registered into the national DNC Registry may not receive unsolicited marketing messages (voice calls, text messages, or fax) from any registered organisation.

PDPA vs GDPR (EU 2018)

The *GDPR has stricter measures than the PDPA for requesting and providing consent. For example

Breach notification: Data controllers must notify supervisory authority, private individuals affected, or the organisation to which it reports of any privacy breaches without undue delay/within the first 72hrs of having become aware of the breach.

Penalties: An organisation that doesn't comply can be fined up to a maximum of 4% of annual global turnover, or €20 million (whichever is greater).

*European Union's General Data Protection Regulation(GDPR)
-Approved by the EU Parliament on 14 April 2016 and came into effect on 25 May 2018.
The GDPR is most significant piece of privacy regulation to emerge in the EU in over 20 years.

PDPA and GDPR

What you need to know about it in Malaysia.

http://www.shearndelamore.com/whatnews/the-impact-of-the-gdpr-on-malaysian-businesses/

Data breach notification

Under the GDPR, notification of a breach of data security to the supervisory authority and the affected individuals is a mandatory requirement in all member states where the breach is likely to "*result in a risk to the rights and freedoms of natural persons*".

The notification has to be made within 72 hours of becoming aware of the data breach and, in the case of delay, the reasons therefor[efn_note]Article 33 of the GDPR.[/efn_note]. At the very least, data controllers have to provide the following information when making the notification to the supervisory authority:-

1. The nature of the personal data breach including, where possible, the categories and approximate number of data subjects concerned and the categories and approximate number of personal data records concerned;
2. the name and contact details of the Data Protection Officer ("DPO") or other contact point where more information can be obtained;
3. description of the likely consequences of the personal data breach; and
4. description of the measures taken or proposed to be taken by the controller to address the personal data breach including, where appropriate, measures to mitigate its possible adverse effects.

Re-Cap Learning Outcomes

Week 11

By the end of this week you should be able to:

- Differentiate the data management requirements from an internal (data lifecycle) and external (data value chain) perspectives
- Understand the conflicting business and legal objectives, and
- The relationship between ethics, privacy, storage, security and analysis.

MONASH University