FIT5196 DATA WRANGLING

Week 5

Data Discovery and Collection

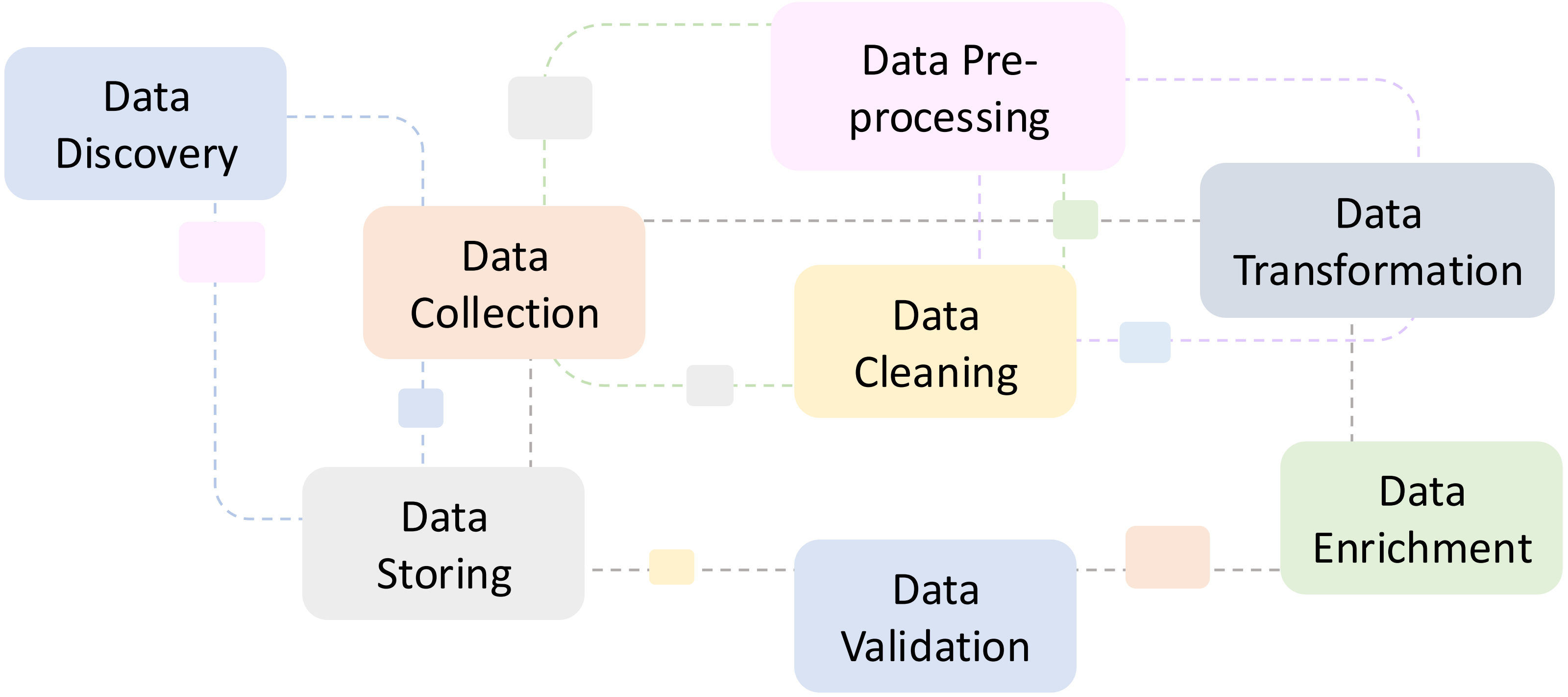
By Jackie Rong

Faculty of Information Technology

Monash University

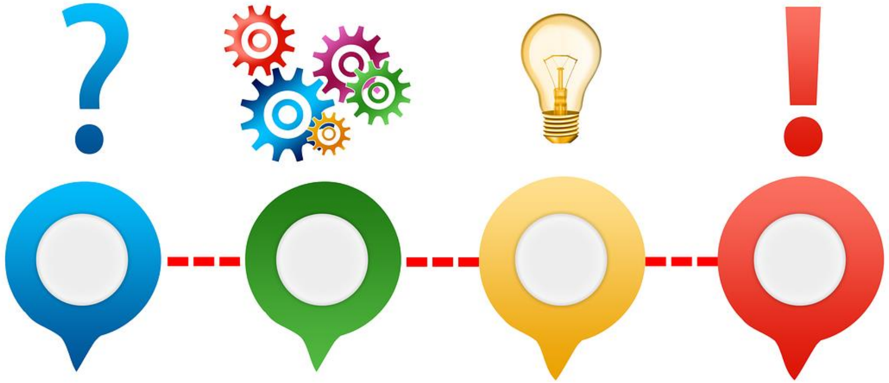
Data Wrangling Tasks (Recap)

In the **Data Pre-processing** stage, preliminary data preparation tasks are performed to make raw data more suitable for analysis.



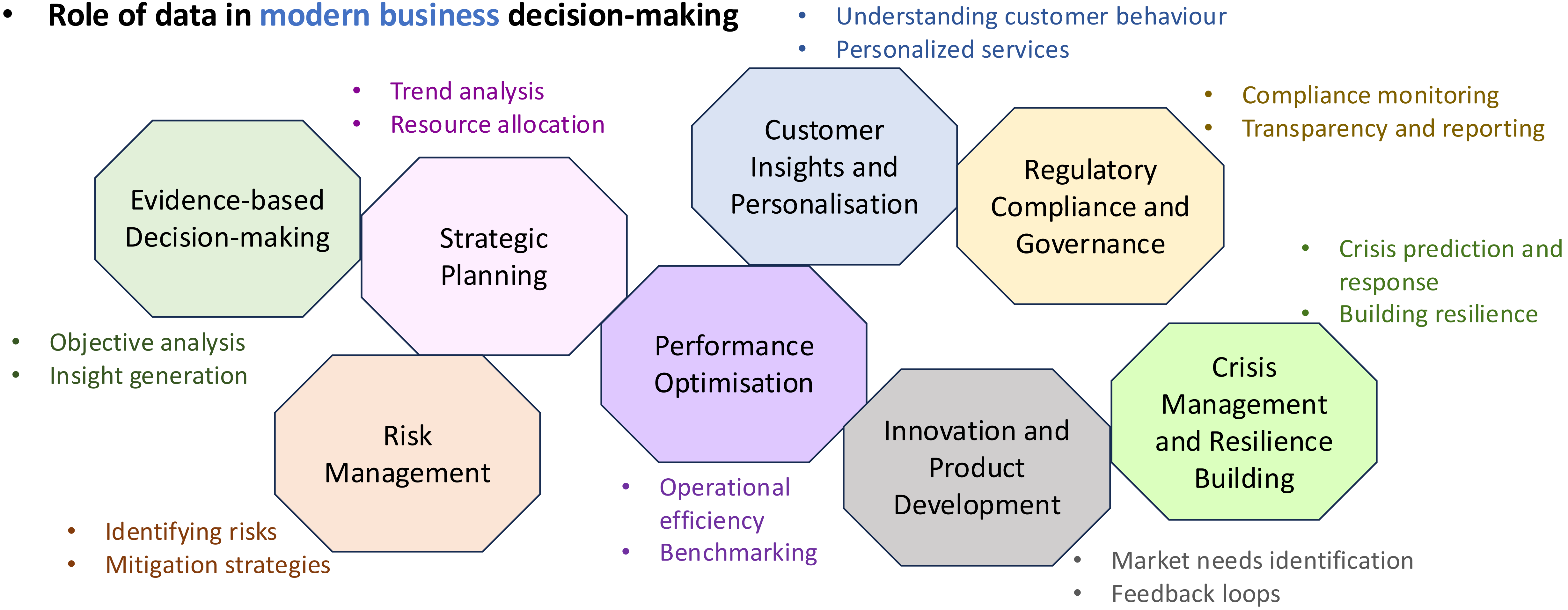
Outline

- What is Data Discovery?
- Data Discovery Process
- What is Data Collection?
- Data Collection Methods



Importance of Data in Decision-Making

- **Role of data in modern business decision-making**



Evidence-based Decision-making

- Objective analysis
- Insight generation

Strategic Planning

- Trend analysis
- Resource allocation

Risk Management

- Identifying risks
- Mitigation strategies

Customer Insights and Personalisation

- Understanding customer behaviour
- Personalized services

Performance Optimisation

- Operational efficiency
- Benchmarking

Regulatory Compliance and Governance

- Compliance monitoring
- Transparency and reporting

Innovation and Product Development

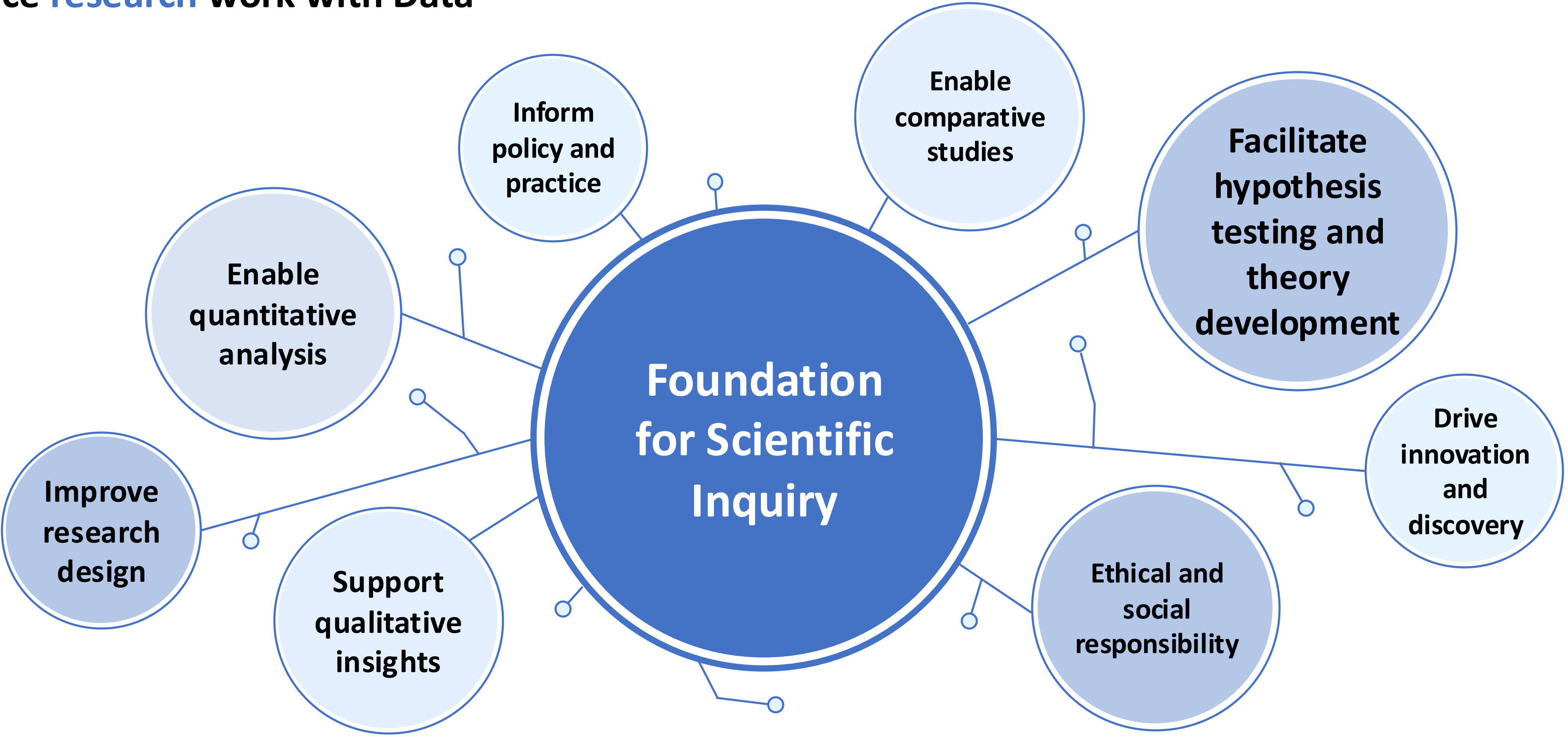
- Market needs identification
- Feedback loops

Crisis Management and Resilience Building

- Crisis prediction and response
- Building resilience

Importance of Data in Decision-Making (cont.)

- Enhance research work with Data



Foundation for Scientific Inquiry

Inform policy and practice

Enable comparative studies

Facilitate hypothesis testing and theory development

Drive innovation and discovery

Ethical and social responsibility

Support qualitative insights

Improve research design

Enable quantitative analysis

Data Discovery

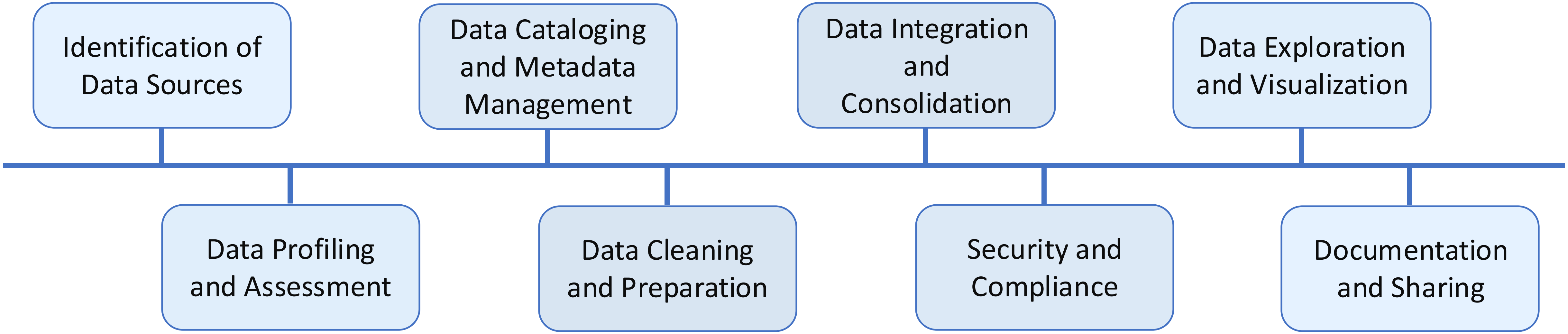
- **Data discovery** is the process of identifying and understanding data sources that can be used for analytical purposes.
- The **primary purpose** of data discovery is to
  - Gain actionable insights into the available data,
  - Understand its potential for analysis,
  - Determine how it can be used to support decision-making and research objectives.

Challenges in Data Discovery

- Volume and complexity of data
- Data quality and silos
- Dynamic and evolving data
- Data privacy and security concerns
- Lack of metadata and documentation
- Interoperability and integration issues
- Resource constraints
- Finding actionable insights

Data Discovery Process

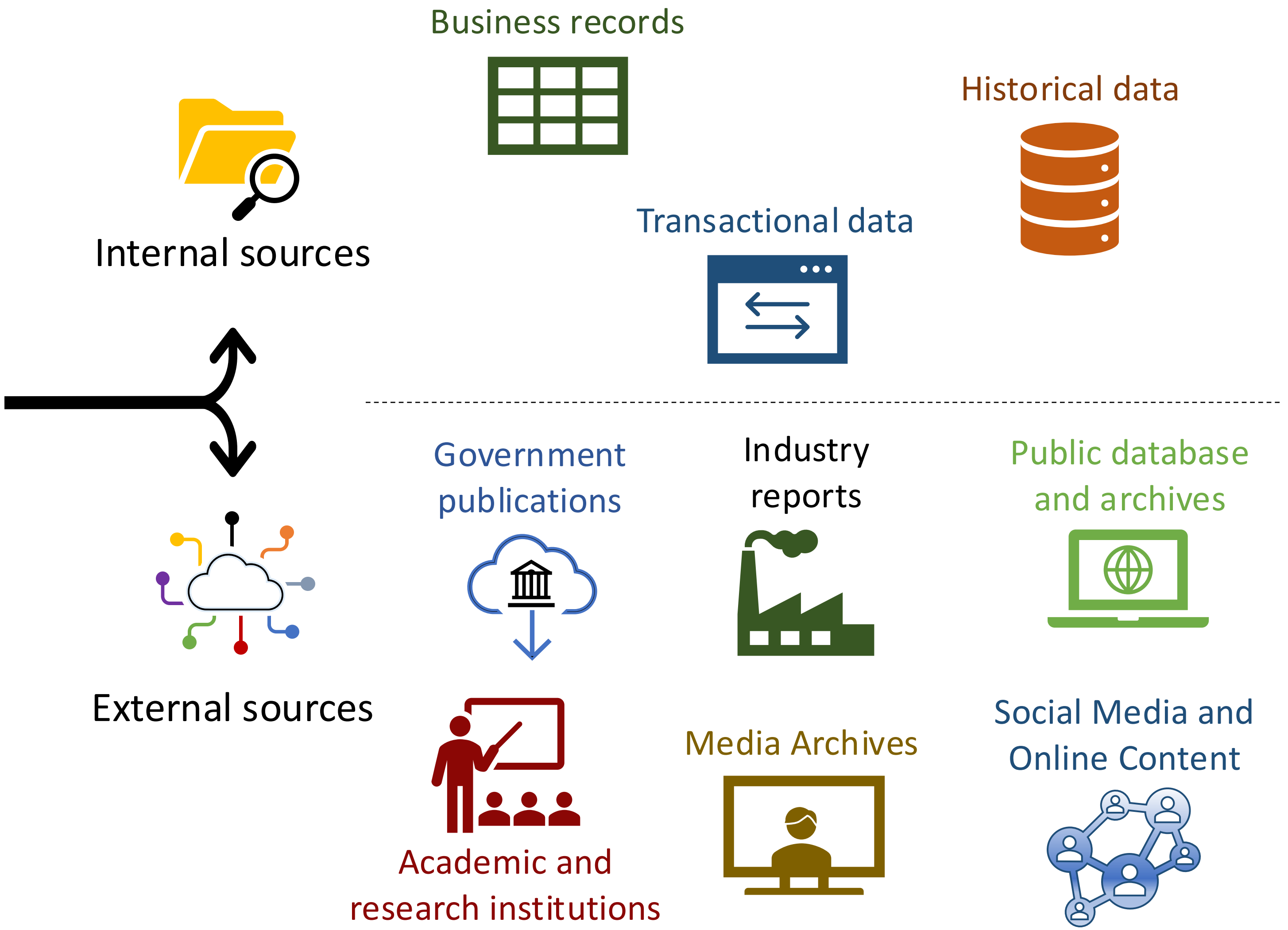
- Data discovery process involves a series of tasks aimed at identifying, understanding, and preparing data for analysis.


Identification of Data Sources
Data Profiling and Assessment
Data Cataloging and Metadata Management
Data Cleaning and Preparation
Data Integration and Consolidation
Security and Compliance
Data Exploration and Visualization
Documentation and Sharing

Data Discovery Process (cont.)

- **Identification of Data Sources**
  - Inventory Existing Data

> **Existing data**, often referred to as *secondary data*, encompasses information that has already been collected for purposes other than the specific research or analysis at hand.



Internal sources

Business records

Historical data

Transactional data

External sources

Government publications

Industry reports

Public database and archives

Academic and research institutions

Media Archives

Social Media and Online Content

Data Discovery Process (cont.)

- **Identification of Data Sources**
  - Inventory Existing Data

Existing data, often referred to as *secondary data*, encompasses information that has already been collected for purposes other than the specific research or analysis at hand.

Advantages of Using Existing Data

- **Cost and time efficiency**
  - Collecting new data can be expensive and time-consuming. Utilizing existing data can significantly reduce both costs and time to insight.
- **Access to broad and diverse data**
  - Existing data can provide access to a wide range of information across different geographies, time periods, and populations.
- **Benchmarking and trends analysis**
  - Allows for the comparison of internal data against industry benchmarks or historical data, facilitating trend analysis and strategic planning.

Data Discovery Process (cont.)

- **Identification of Data Sources**
  - Inventory Existing Data

Existing data, often referred to as *secondary data*, encompasses information that has already been collected for purposes other than the specific research or analysis at hand.

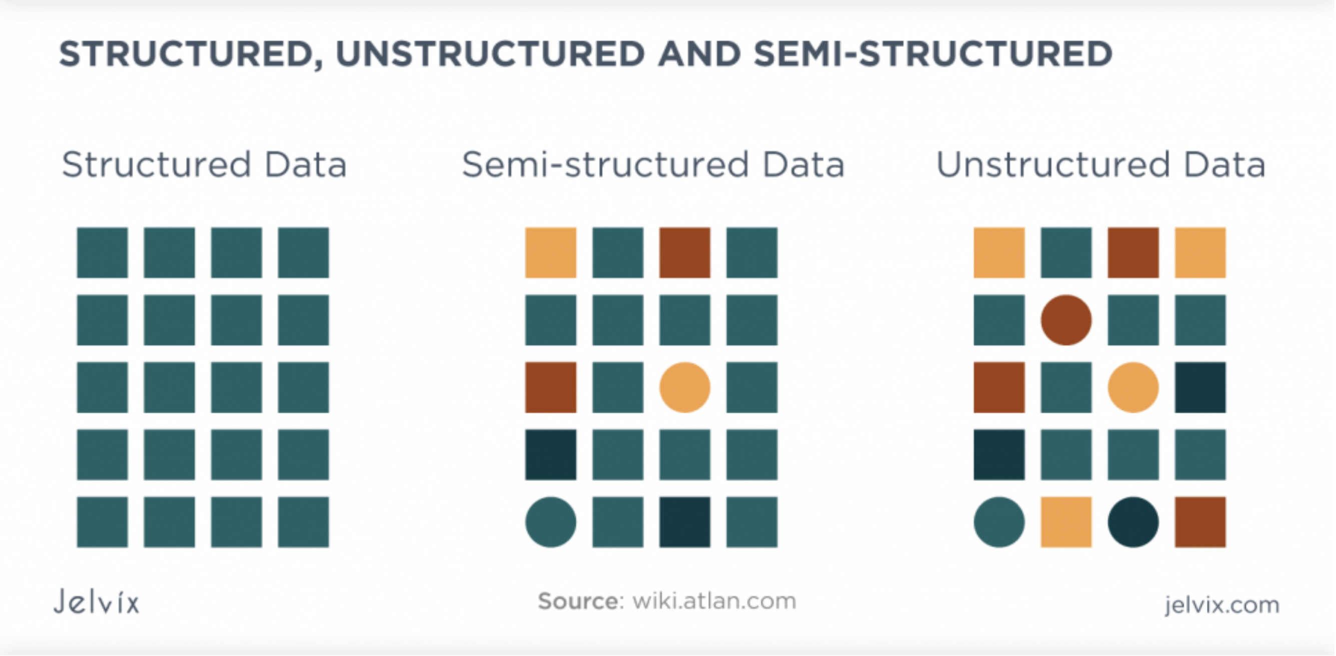
⚠ **Evaluate Data Relevance**
Assess the relevance of each data source to the business questions or analytical projects at hand.

Limitations of Using Existing Data

- **Relevance**
  - The data may not perfectly match the specific needs of the current analysis or research question.
- **Quality and accuracy**
  - The quality and accuracy of existing data can vary, and it may be outdated or not rigorously collected.
- **Accessibility**
  - Some data, especially from private sources or specific industries, may be difficult to access or require purchase.

Data Discovery Process (cont.)

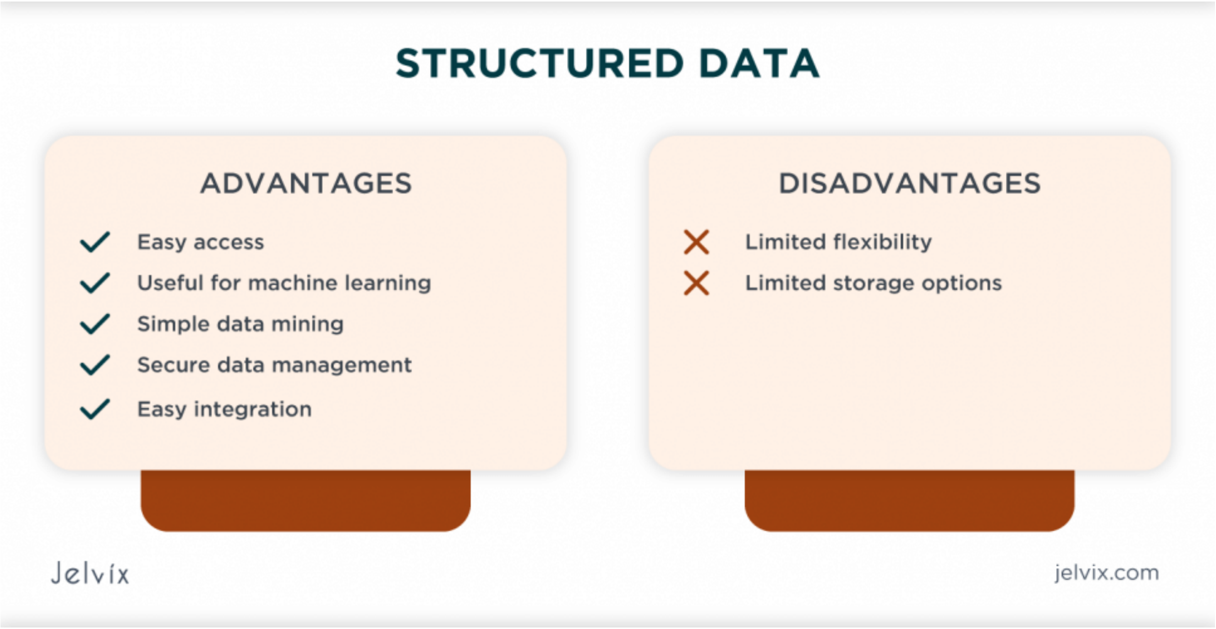
- **Data Profiling and Assessment**
  - Understand data structure
  - Content exploration
  - Quality assessment



Source: wiki.atlan.com

Data Discovery Process (cont.)

- **Data Profiling and Assessment**
  - Understand data structure
    - **Structured data** is highly organized and easily understandable by machine language, typically stored in databases.
      - Relational databases
      - Data warehouses



STRUCTURED DATA

ADVANTAGES

- ✓ Easy access
- ✓ Useful for machine learning
- ✓ Simple data mining
- ✓ Secure data management
- ✓ Easy integration

DISADVANTAGES

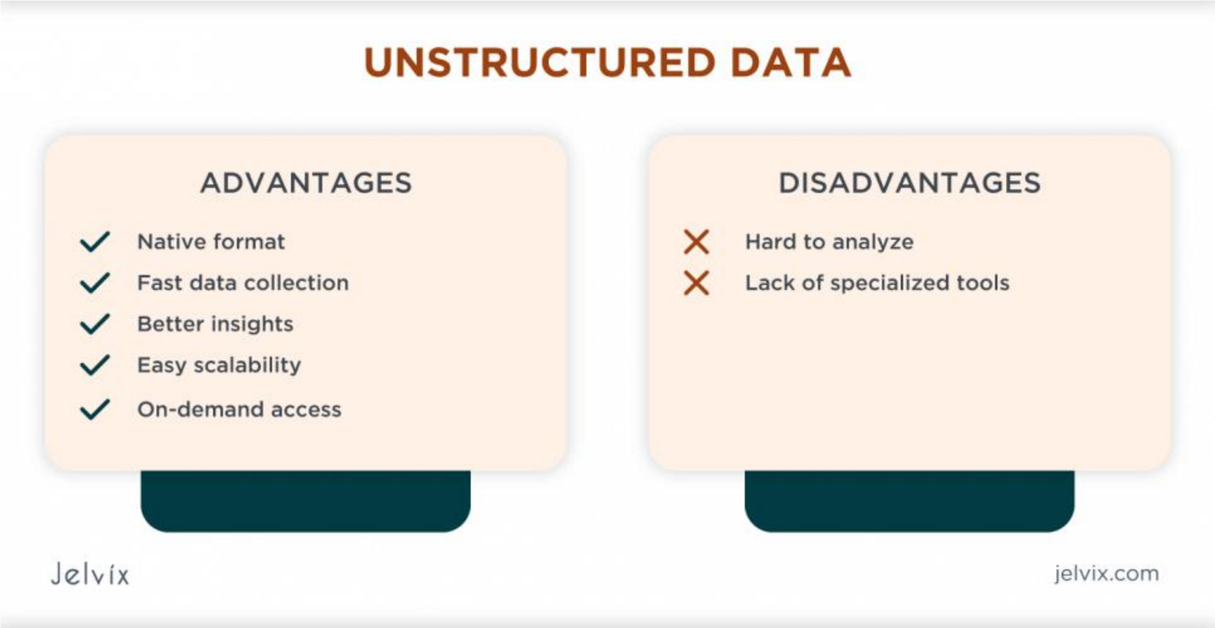
- ✗ Limited flexibility
- ✗ Limited storage options

Jelvix

jelvix.com

Data Discovery Process (cont.)

- **Data Profiling and Assessment**
  - Understand data structure
    - **Unstructured data** is information that doesn't have a pre-defined data model or is not organized in a pre-defined manner. It's more challenging to collect, process, and analyse.
      - Text data
      - Multimedia data



UNSTRUCTURED DATA

ADVANTAGES

✓ Native format
✓ Fast data collection
✓ Better insights
✓ Easy scalability
✓ On-demand access

DISADVANTAGES

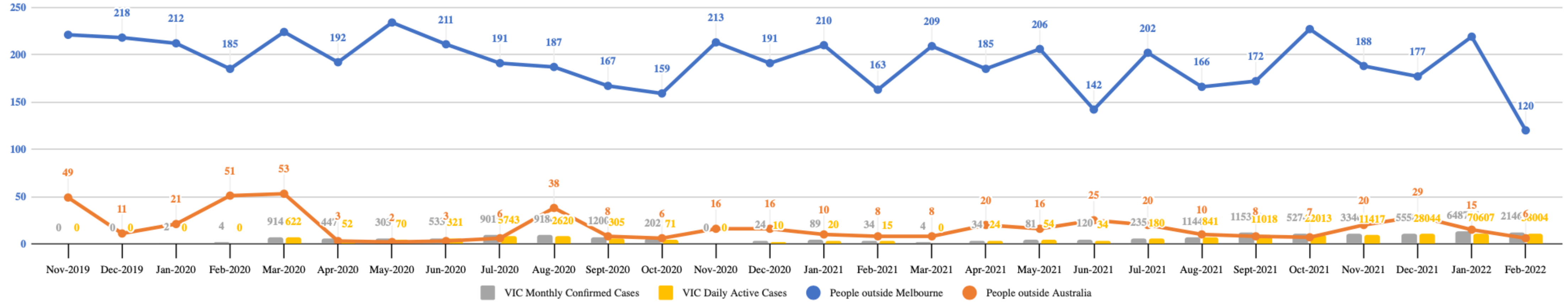
✗ Hard to analyze
✗ Lack of specialized tools

Jelvix

jelvix.com

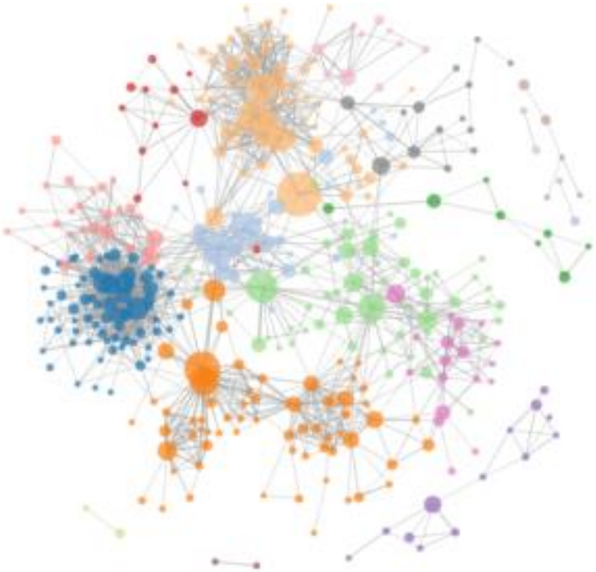
Data Discovery Process (cont.)

- **Data Profiling and Assessment**
  - Understand data structure
    - **Time-series data**
      - Time-series data is data where sequences of values are indexed in time order, often in regular intervals.
      - This is common in financial analysis, sensor data, and application performance monitoring.

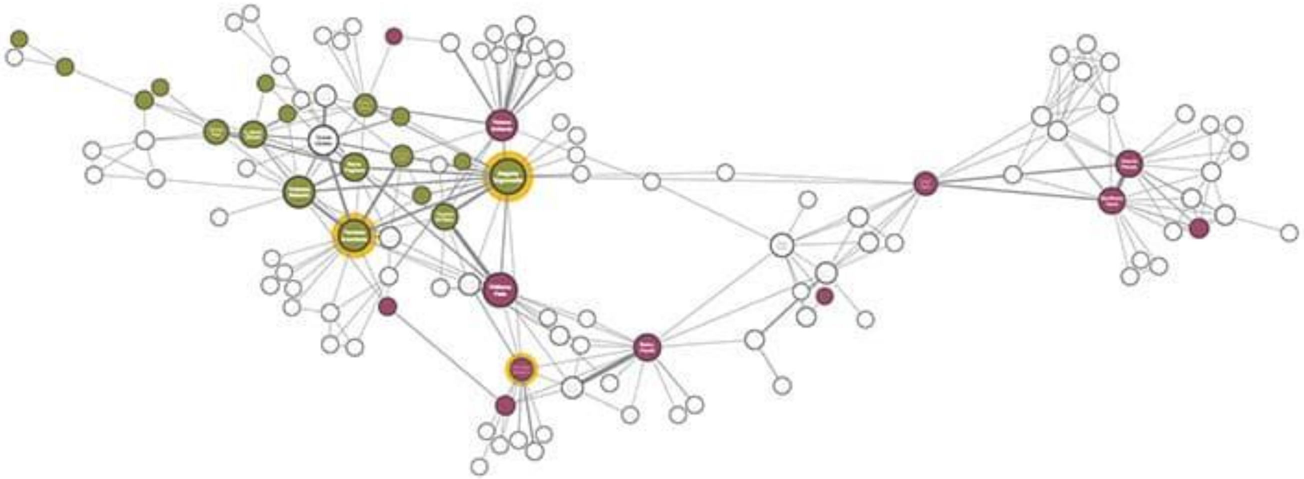


Data Discovery Process (cont.)

- **Data Profiling and Assessment**
  - Understand data structure
    - **Graph data**
      - Graph data models are used to represent relationships between entities in a flexible and intuitive way, making them ideal for social network analysis, recommendation systems, and fraud detection.



Source: Digital Humanities, Network Graph, University of Georgia, https://digi.uga.edu/network-graphs/



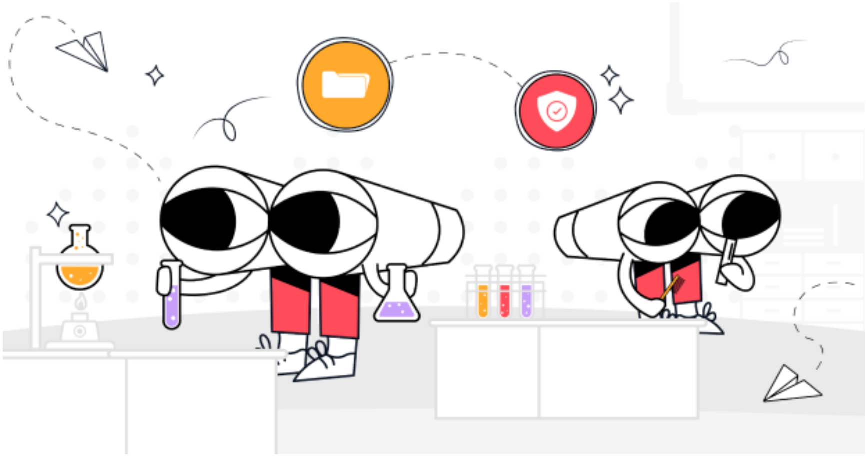
Source: Cambridge Intelligence, what is graph visualization? https://cambridge-intelligence.com/graph-visualization-software/

Data Discovery Process (cont.)

- **Data Profiling and Assessment**
  - Understand data structure
    - **Big data**
      - Refers to data that is so voluminous that traditional data processing software can't manage them.
      - Big data encompasses all the previously mentioned data structures but on a much larger scale and velocity.

Data Discovery Process (cont.)

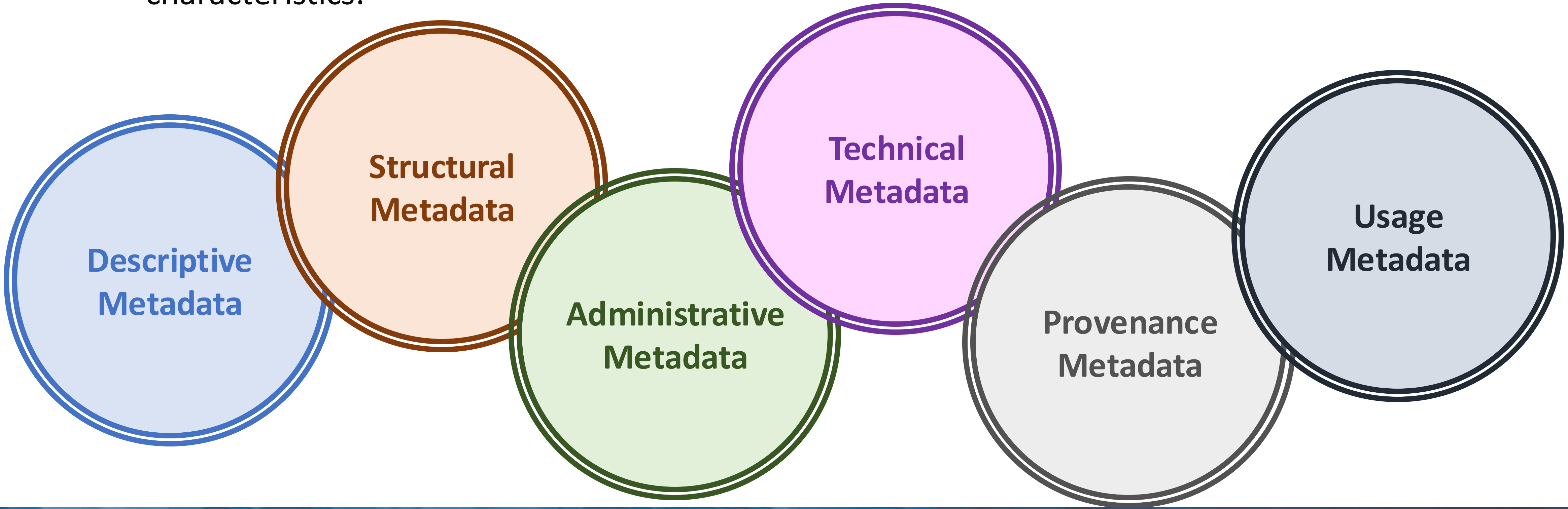
- **Data Profiling and Assessment**
  - **Content Exploration**
    - Delve into the content of the data to understand the type of information it holds, such as categorical, numerical, or textual data.
  - **Quality Assessment**
    - Evaluate the quality of data by identifying issues such as missing values, duplicates, or inconsistencies.



Source: Data Profiling: Definition, Types, Process, & More. https://www.castordoc.com/blog/data-profiling-definition-types-process-more

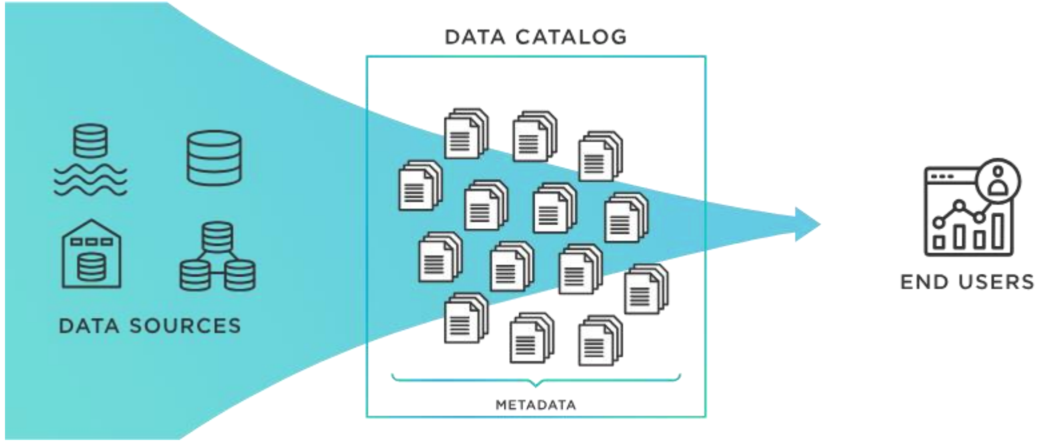
Data Discovery Process (cont.)

- **Data Cataloging and Metadata Management**
  - **Metadata Collection**
    - Gather metadata, which includes information about the data's origin, format, and characteristics.



Data Discovery Process (cont.)

- **Data Cataloging and Metadata Management**
  - **Catalog Creation**
    - Create a searchable catalog of data assets, making it easier for users to find and understand the data they need.
  - **Data Lineage Documentation**
    - Document data lineage, tracing the data from its source through various transformations to its current state, to ensure transparency and trust in the data.



Source: https://www.tibco.com/glossary/what-is-a-data-catalog

Data Discovery Process (cont.)

- **Data Cleaning and Preparation**
  - **Data Cleansing**
    - Address data quality issues identified during profiling, such as correcting errors, filling missing values, or removing duplicates.
  - **Data Transformation**
    - Transform data into a format or structure that is suitable for analysis, which may include normalization, aggregation, or encoding of categorical variables.
- **Data Integration and Consolidation**
  - **Combine Data Sources**
    - Integrate data from multiple sources to create a comprehensive dataset that provides a unified view of the information.
  - **Ensure Consistency**
    - Harmonize data formats, units of measure, and other discrepancies across data sources to ensure consistency.

Data Discovery Process (cont.)

- **Security and Compliance Checks**
  - **Data Privacy**
    - Implement measures to protect sensitive information and personal data in compliance with privacy regulations (e.g., GDPR, HIPAA).
  - **Access Control**
    - Establish data access controls to ensure that only authorized users can access certain data, based on their roles and the data's sensitivity.
- **Data Exploration and Visualization**
  - **Exploratory Data Analysis (EDA)**
    - Conduct an initial exploration of the data to uncover patterns, trends, and anomalies using statistical summaries and visualization tools.
  - **Visualization**
    - Use data visualization techniques to represent data graphically, making it easier to identify relationships, outliers, and patterns.

Data Discovery Process (cont.)

- **Documentation and Sharing**
  - **Document Findings**
    - Document the findings from data exploration, including insights, challenges, and potential uses of the data.
  - **Share Insights**
    - Share the documented findings and data visualizations with stakeholders to facilitate data-driven decision-making.

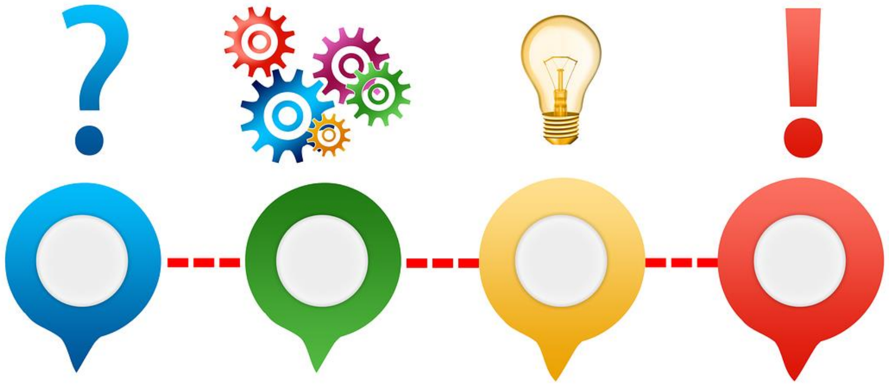
Data Discovery Tools & Platforms

- **Data discovery tools** are essential in today's data-driven world, helping organizations and researchers to uncover insights, trends, and patterns from vast amounts of data.

tableau
splunk>
ALTERYX
Power BI
Qlik Sense
elastic
collibra

Outline

- What is Data Discovery?
- Data Discovery Process
- **What is Data Collection?**
- **Data Collection Methods**



Data Collection

- **Data collection** is the systematic process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.
- The process can vary in methodology.
- Data collection is foundational to the empirical approach in various domains, facilitating a deep understanding of complex issues, guiding strategic planning, and enabling the measurement of outcomes.

Source: https://www.pngegg.com/en/png-nxbqm

Primary vs. Secondary Data

- **Primary Data**
  - **Original data**, is collected firsthand by the researcher for a specific research purpose or project.
  - Primary data is collected directly from the source, allowing the researcher to control the quality, purpose, methodology, and scope of the data.

- **Secondary Data**
  - Information that was collected by someone else for a different purpose but is being used for a new project.
  - Secondary data has already been gathered, complied, and often analysed or interpreted before the current project.

Quantitative vs. Qualitative Data

- **Quantitative Data**
  - Any data that can be quantified or measured numerically.
  - It is data that can be expressed in numbers and involves measurable quantities.
  - The focus is on the quantity of the data rather than its qualitative aspects.
  - It is often used to formulate facts and uncover patterns in research.
  - Often collected using structured research instruments like surveys and experiments.
  - Suitable for statistical analysis to test hypotheses or predict outcomes.
  - Can be displayed through graphs, charts, and tables for interpretation.

- **Qualitative Data**
  - Qualitative data is descriptive and conceptual.
  - It is data that can be observed but not measured with numbers.
  - Often used to understand concepts, thoughts, or experiences and provides insights into the problem or helps to develop ideas or hypotheses for potential quantitative research.
  - Describes qualities or characteristics.
  - Data is usually textual or visual.
  - Analysis can be more subjective and involves interpretation of meanings from the data.

Data Collection Methods – Structured Data

- For **structured data**
  - Surveys and questionnaires
    - Key considerations to ensure the reliability and validity of the data collected.
      - Clearly define objectives
      - Question design
      - Sampling
      - Pilot testing
      - Ethical considerations
      - Distribution method

Source: https://www.pngegg.com/en/png-dberq

Data Collection Methods – Structured Data

- For **structured data**
  - Online forms
    - User experience and design
    - Privacy and security
    - Accessibility
    - Data quality
  - web scraping
    - Legal and ethical considerations
    - Technical challenges
    - Data quality and relevance
    - Efficiency and resource utilisation

SurveyMonkey

Google Forms

qualtrics

Web Scraping Beautiful Soup

Scrapy

Source: https://www.pngegg.com/en/png-xdgxk

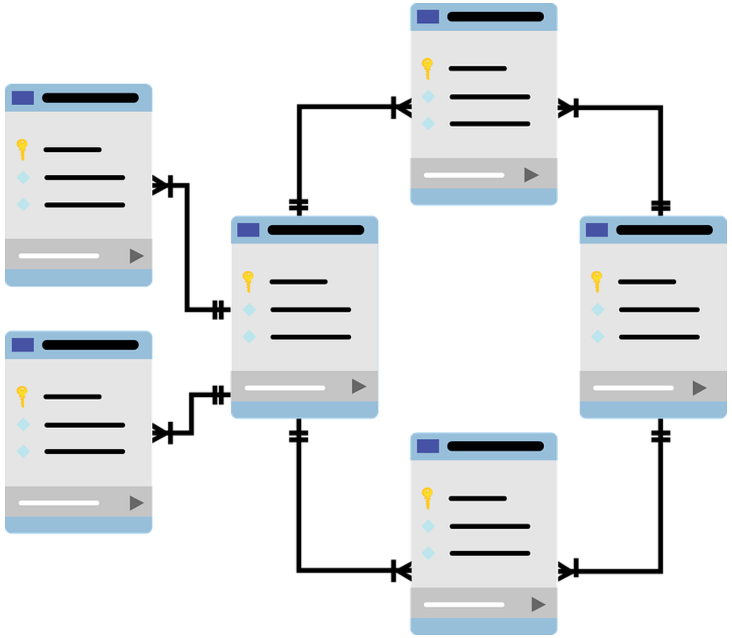
Data Collection Methods – Structured Data

- For **structured data**
  - Relational databases
    - Database design and structure
    - Data quality and integrity
    - Scalability and performance
    - Security measures
    - Backup and recovery
    - Compliance with regulations
    - Data accessibility and documentation
    - Monitoring and maintenance

MySQL

PostgreSQL

SQLite



Source: https://www.pngegg.com/en/png-tqpzo

Data Collection Methods – Structured Data

- For **structured data**
  - API
    - API documentation review
    - Authentication and authorisation
    - Rate limiting and quotas
    - Data pagination
    - Error handling
    - Data efficiency and minimisation
    - Compliance with API terms of service
    - Data storage and management
    - Monitoring and maintenance

Source: https://www.pngegg.com/en/png-iwevt

Data Collection Methods – Unstructured Data

- For **unstructured data**
  - Text mining and natural language processing (NLP)
    - Data quality and volume
    - Data preparation and pre-processing
    - Choose the right NLP techniques and models
    - Understanding context and nuances
    - Ethical considerations and bias
    - Performance evaluation and validation
    - Scalability and computational resources
    - Integration with other data sources

Natural Language
Tool Kit (NLTK)
Basic Text Analytics

python

Natural Language Processing
(NLP) with Python and SpaCy
spaCy

Chat GPT

Source: https://www.pngegg.com/en/png-zaepp

Data Collection Methods – Unstructured Data

- For **unstructured data**
  - Image and video data collection
    - Consistent quality and high-resolution data
    - Diversity and representation
    - Ethical considerations and legal compliance
    - Large file sizes and data organisation
    - Accurate annotations and labelling
    - Conversion and processing for standard formats
    - Ethical use and bias mitigation
    - Bandwidth and transfer
    - Real-time processing

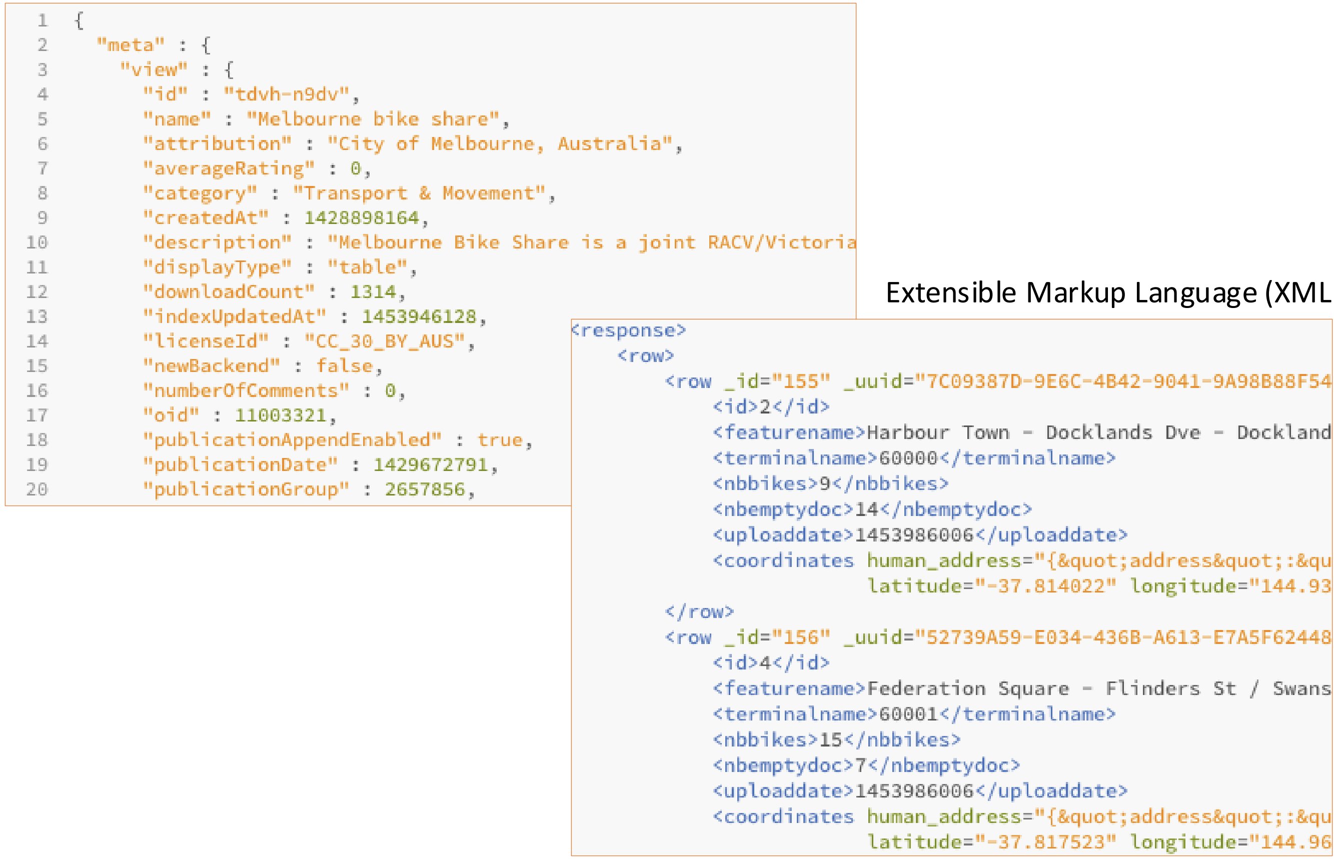
Source: https://www.pngegg.com/en/png-ejjtu

Data Collection Methods – Unstructured Data

- For **unstructured data**
  - Social media and web content
    - Legal and ethical consideration
    - Adherence to APIs terms of use
    - Changes in API access
    - Dynamic content
    - Handling noise
    - Anonymisation and data processing
    - Sampling bias and cultural context
    - Long-term accessibility
    - Archiving and preservation

Data Collection Methods – Semi-structured Data

- For **semi-structured data**
  - JSON and XML data extraction
    - Understand data structure
      - Hierarchical structure
      - schema/schemaless
    - Parsing data using libraries
    - Regular expression
    - Character encoding
    - Handling inconsistencies and errors



JavaScript Object Notation (JSON):

```
{
  "meta" : {
    "view" : {
      "id" : "tdvh-n9dv",
      "name" : "Melbourne bike share",
      "attribution" : "City of Melbourne, Australia",
      "averageRating" : 0,
      "category" : "Transport & Movement",
      "createdAt" : 1428898164,
      "description" : "Melbourne Bike Share is a joint RACV/Victoria
      "displayType" : "table",
      "downloadCount" : 1314,
      "indexUpdatedAt" : 1453946128,
      "licenseId" : "CC_30_BY_AUS",
      "newBackend" : false,
      "numberOfComments" : 0,
      "oid" : 11003321,
      "publicationAppendEnabled" : true,
      "publicationDate" : 1429672791,
      "publicationGroup" : 2657856,
```

Extensible Markup Language (XML)

```
<response>
    <row>
        <row _id="155" _uuid="7C09387D-9E6C-4B42-9041-9A98B88F54
            <id>2</id>
            <featurename>Harbour Town - Docklands Dve - Dockland
            <terminalname>60000</terminalname>
            <nbbikes>9</nbbikes>
            <nbemptydoc>14</nbemptydoc>
            <uploaddate>1453986006</uploaddate>
            <coordinates human_address="{&quot;address&quot;:&qu
                         latitude="-37.814022" longitude="144.93
        </row>
        <row _id="156" _uuid="52739A59-E034-436B-A613-E7A5F62448
            <id>4</id>
            <featurename>Federation Square - Flinders St / Swans
            <terminalname>60001</terminalname>
            <nbbikes>15</nbbikes>
            <nbemptydoc>7</nbemptydoc>
            <uploaddate>1453986006</uploaddate>
            <coordinates human_address="{&quot;address&quot;:&qu
                         latitude="-37.817523" longitude="144.96
```

Data Collection Methods – Semi-structured Data

- For **semi-structured data**
  - Logs and sensor data collection
    - Volume and velocity
      - High throughput
      - Stream processing
    - Variability and structure
      - Diverse formats and standardisation
    - Time-sensitivity
      - Timestamps and time zone awareness
    - Interoperability
    - Automated alerts and actions

Source: https://www.pngegg.com/en/png-mrtrd/download

Data Collection Methods – Semi-structured Data

- For **semi-structured data**
  - Email and communication data collection
    - Privacy and Legal Compliance
      - Consent and Authorization
      - Sensitive Information
    - Data Structuring and Formatting
      - Complex Structures
      - Metadata Extraction
      - Handling Attachments
    - Data Quality and Integrity
      - De-duplication
      - Noise Filtering

Source: https://www.pngegg.com/en/png-hvcvm

Ethical Considerations and Privacy

- **Ethical considerations and privacy** are paramount in the data collection process, guiding how data should be collected, stored, used, and shared.
- These considerations protect individuals' rights and maintain trust between data collectors and subjects.

Informed Consent

- Transparency and openness
- Voluntariness
- Understanding

Privacy and Anonymity

- Protecting personal information
- anonymisation

Compliance with Laws and Regulations

- Legal compliance
- Ethical standards

Data Retention and Disposal

- Retention policy
- Secure disposal

Data Security

- Secure storage and transmission
- Access controls

Minimisation and Necessity

- Data minimisation
- Purpose limitation

Respect for Participants

- Respect for autonomy
- Beneficence and non-maleficence
- Accountability

Equity and Fairness

- Inclusive data collection
- Fair treatment

Ethical Considerations and Privacy

- **Ethical considerations and privacy** are paramount in the data collection process, guiding how data should be collected, stored, used, and shared.
- These considerations protect individuals' rights and maintain trust between data collectors and subjects.

Summary & To-do List

- Please download and read the materials provided on Moodle.
- Review the content learnt from Week 5.

- Assessment 1
  - Continue to work on Assessment 1

- Next week: Data Structuring