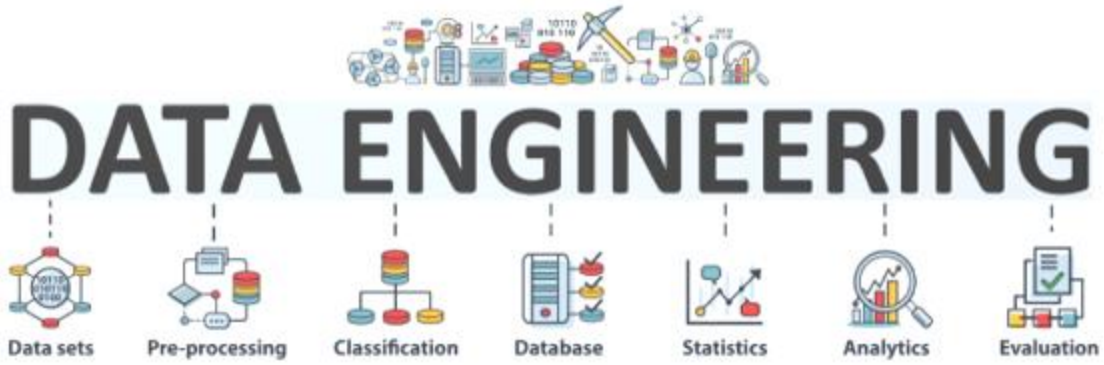


香港城市大學
City University of Hong Kong

Lecture 1: Introduction & Data Ecosystem

CS5481 Data Engineering

Instructor: Linqi Song

WHAT IS Data Engineering?

- Big data is changing the way we do business and creating a need for data engineers who can collect and manage large quantities of data.
- Data engineering is the practice of designing and building systems for collecting, storing, and analyzing data at scale.

Outline

- 1. Course organization
- 2. Overview of the data engineering ecosystem
- 3. Types of data
- 4. Languages and tools for data professionals

Bird’s-eye view of this course

- Data pipeline
  - Data acquisition, data processing, and data storage
  - Data management
- Data processing techniques for data driven applications
  - Information retrieval, recommendations, social network analysis, anomaly detection

Instruction pattern

- 3-hour in-class learning
  - 2-hour lecture, Wednesdays, 20:00-21:50, Yeung LT-1
    - Cover main topics of the course
  - 1-hour tutorial
    - In-class hands-on ability, discussions, presentations, etc.
    - T61, Wednesday 17:00 - 17:50, Yeung G7510;
    - T62, Wednesday 18:00 - 18:50, Yeung G7510;
    - T63, Wednesday 19:00 - 19:50, Yeung B7520/B7510.
- After class
  - Homework assignments and projects
  - Reading recent advances of related fields, implementation details,
  - Papers, technical blogs, GitHub, etc.
- QA
  - Canvas -> Discussions is most preferred
  - Email for more personalized questions
  - TAs for hands-on issues

Teaching team

- Prof. Linqi Song
  - Yeung Y6415, linqi.song@cityu.edu.hk

- TAs

  Ms. Yuxuan Yao (lead TA, yuxuanyao3-c@my.cityu.edu.hk),
  Mr. Mingyan Liu (lead TA, mingyaliu8-c@my.cityu.edu.hk),
  Mr. Guanzhi Deng (lead TA, guanzdeng2-c@my.cityu.edu.hk),
  Mr. Yunhe Li (yunheli4-c@my.cityu.edu.hk),
  Mr. Zengyan Liu (zengyaliu2-c@my.cityu.edu.hk),
  Mr. Jilin Cao (jilincao2-c@my.cityu.edu.hk),
  Mr. Zhuo Han (zhuohan3-c@my.cityu.edu.hk).

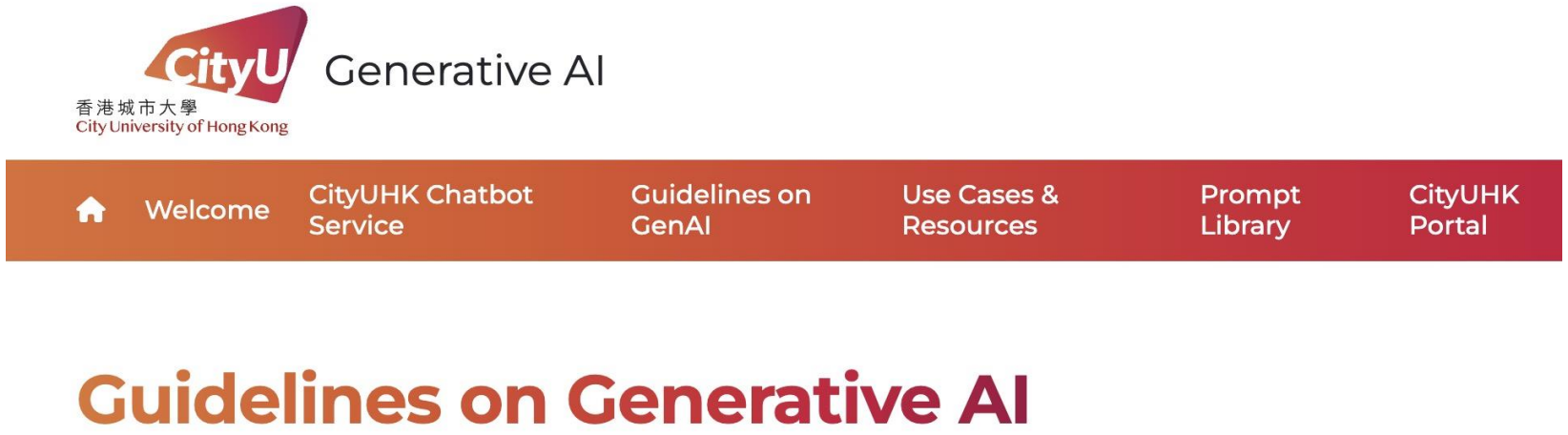
Assessment

- Continuous assessment (60%)
  - 2 individual homework assignments (each 15%)
    - Answer questions and/or programming to implement simple data engineering tasks
  - 1 group project with presentations (30%)
    - To enhance student engagement and peer learning through student-led tutorials, teamwork, and more interactions among peers. Students form groups and collaborate on tiny group projects, including pre-class preparation and in-class group discussions and presentations.
    - Form a group of up to 6 students (before Week 5)
    - Select one topic among several given topics
    - Do experiments and show your innovation and novelty
    - Reports + codes + others (datasets, proofs, figures, etc.)
    - Presentation (Weeks 13)
- Final exam (40%)

Proper use of Large Language Models (e.g., ChatGPT)

- We will follow the university's policy on Generative AI usage in homework/projects.

https://www.cityu.edu.hk/GenAI/guidelines.htm



Schedule

Week	Date	Topic	HW
1	Sep. 3	Introduction (data ecosystem)	
2	Sep. 10	Data acquisition	Assign HW1.
3	Sep. 17	Data preprocessing	
4	Sep. 24	Data visualization	Assign group project.
5	Oct. 1	No class (Public holiday)	
6	Oct. 8	Advanced topics and recent trends: Large Language Models for Data Engineering	HW 1 due. Assign HW2.
7	Oct. 15	Data indexing	
8	Oct. 22	Data querying	
9	Oct. 29	No class (Public holiday)	
10	Nov. 5	Data driven applications (1): Information retrieval and recommendations	HW2 due.
11	Nov. 12	Data driven applications (2): Social network analysis and anomaly detection	
12	Nov. 19	Data management	
13	Nov. 26	Project presentation	Group project due.

Resources

- Computing resources
  - Jupyterhub for tutorials, home assignments and group projects.
  - CS department: https://mljh.cs.cityu.edu.hk/
  - Department subscribed Huawei cloud or Google cloud
  - Google Colab: https://colab.research.google.com/
  - Other resources: Kaggle (kaggle.com/notebooks), other online resources, CS Lab MMW 2462
- Other online courses

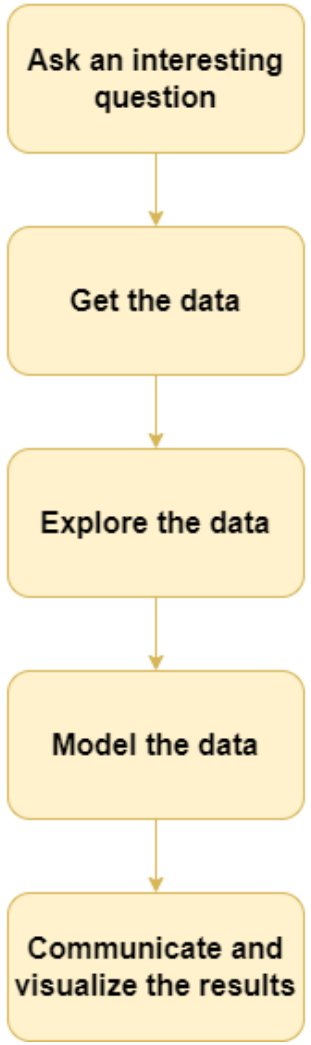
How to learn this course well?

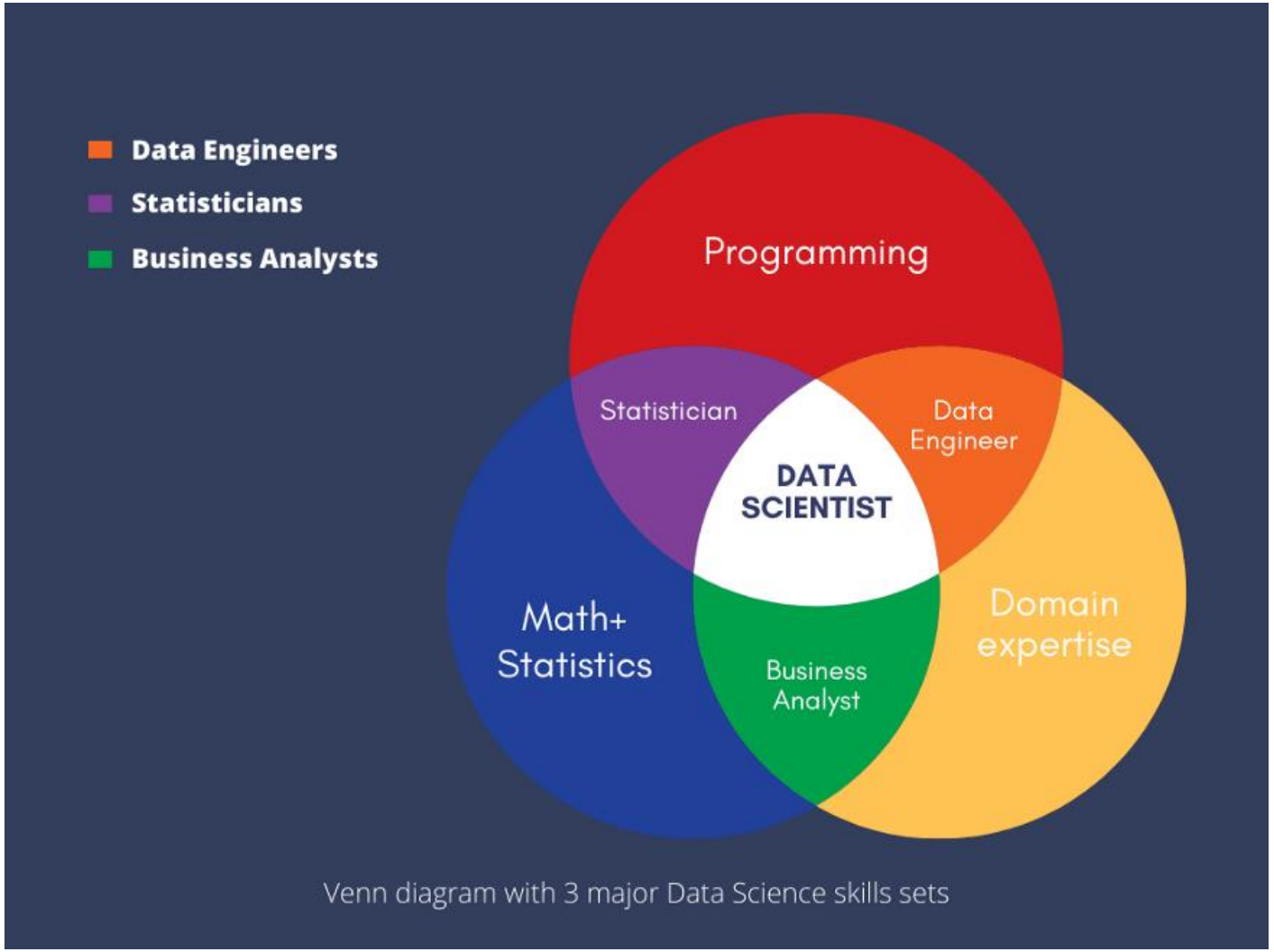
- Higher-level postgraduate courses
  - What problems to solve?
  - How to approach the problem?
  - Systematic ideas instead of details
- This data engineering course
  - Get your hands dirty, as it is mainly about how to process data and implement systems for domain applications
  - Follow recent academic and industrial advances
  - Discuss with others

Outline

- 1. Course organization
- 2. Overview of the data engineering ecosystem
- 3. Types of data
- 4. Languages and tools for data professionals

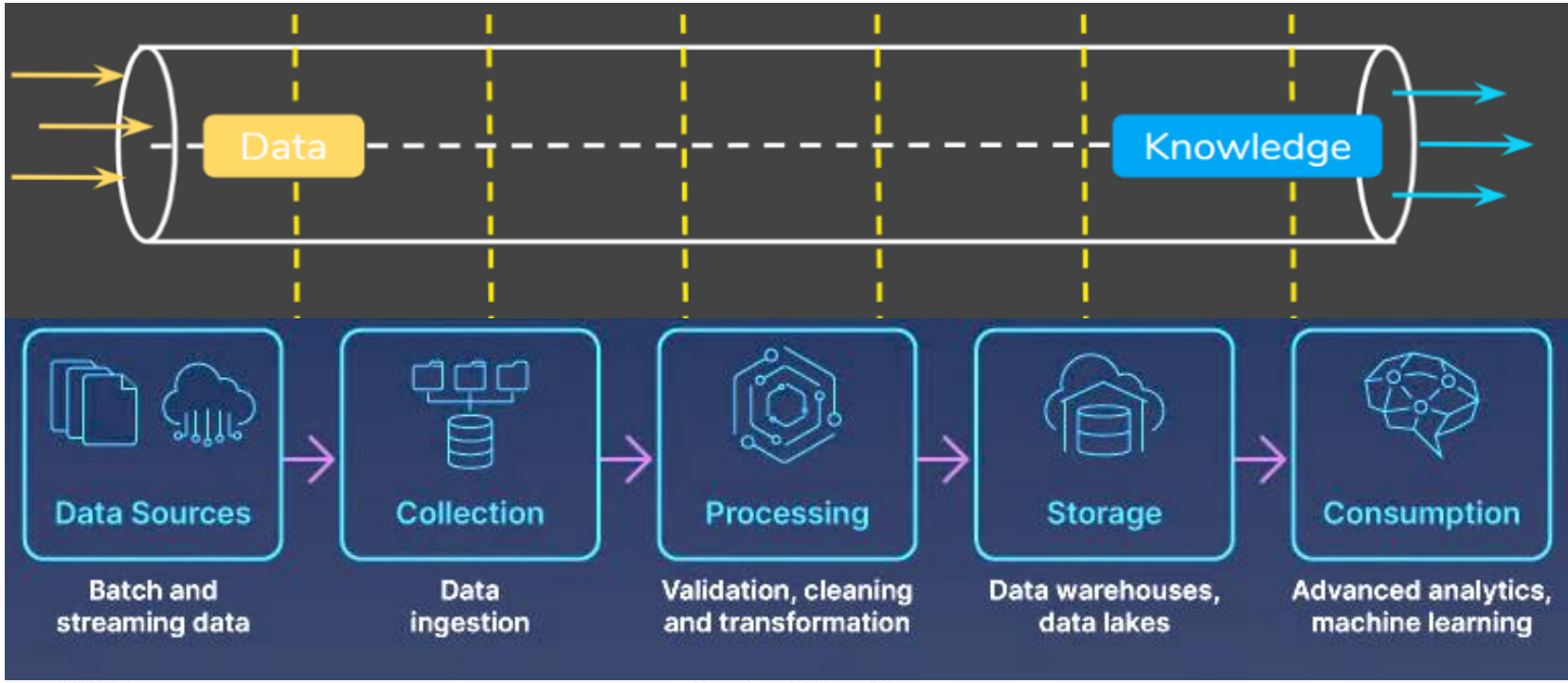
Venn diagram with three dimensional skills for ‘data related work’: coding + maths + domain knowledge

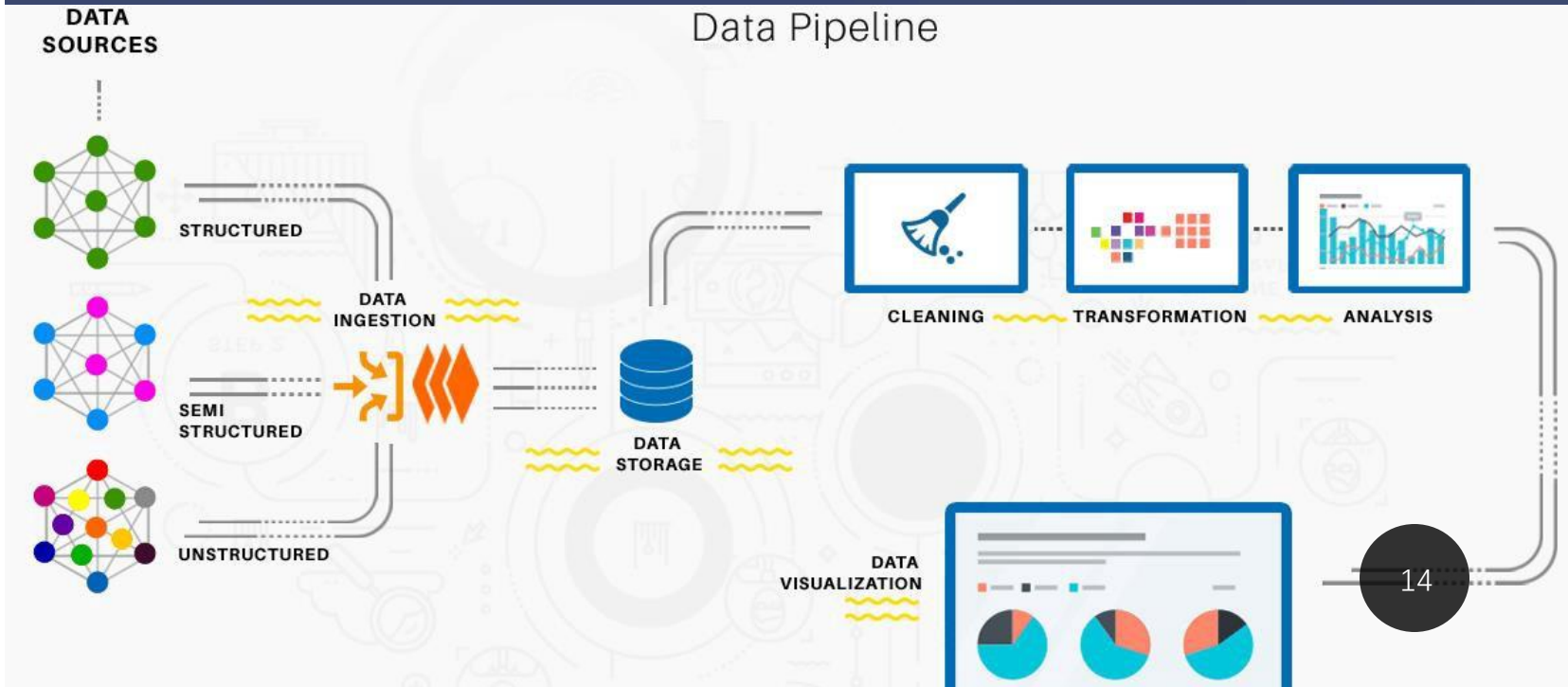




Data engineering pipeline

- **The Goal of Data Engineering** is to provide organized, standard data flow to enable data-driven models such as machine learning models, data analysis.





Data engineering ecosystem

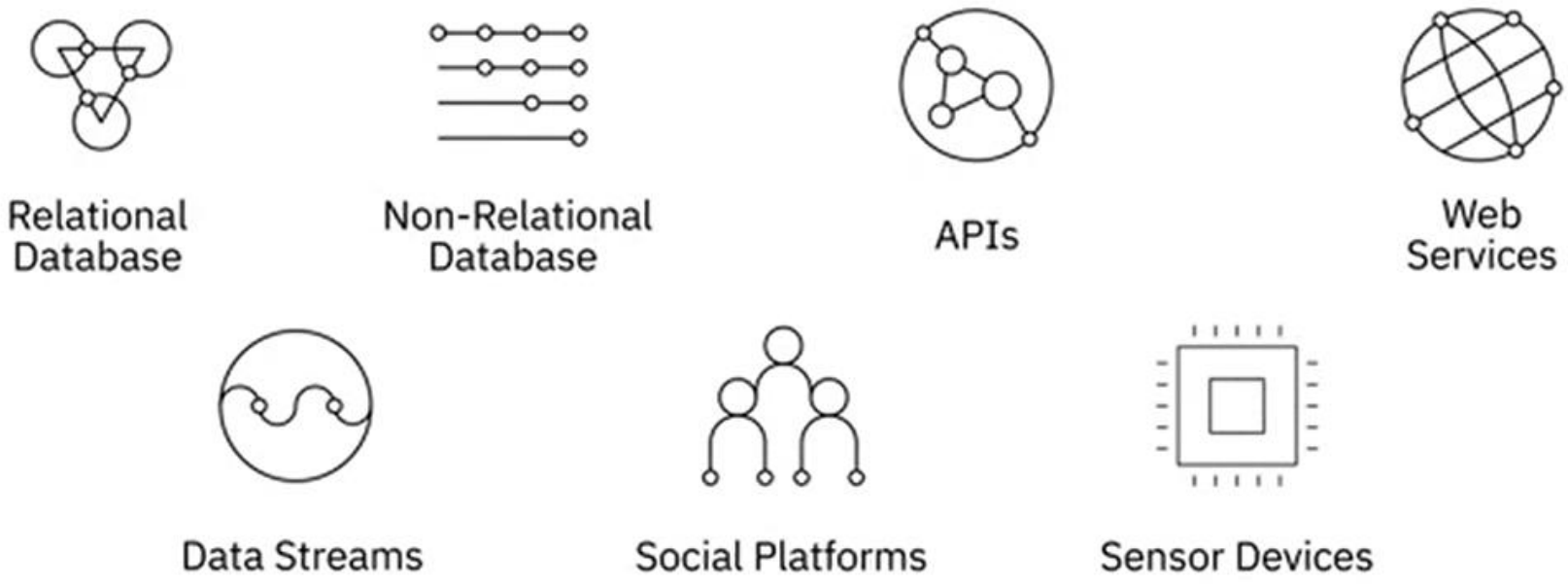
A Data Engineer's ecosystem includes the infrastructure, tools, frameworks, and processes for:

- Extracting data from disparate sources
- Architecting and managing data pipelines for transformation, integration, and storage of data
- Architecting and managing data repositories
- Automating and optimizing workflows and flow of data between systems
- Developing applications needed through the data engineering flow

It is a diverse, rich , and challenging ecosystem.

Data engineering ecosystem: data sources

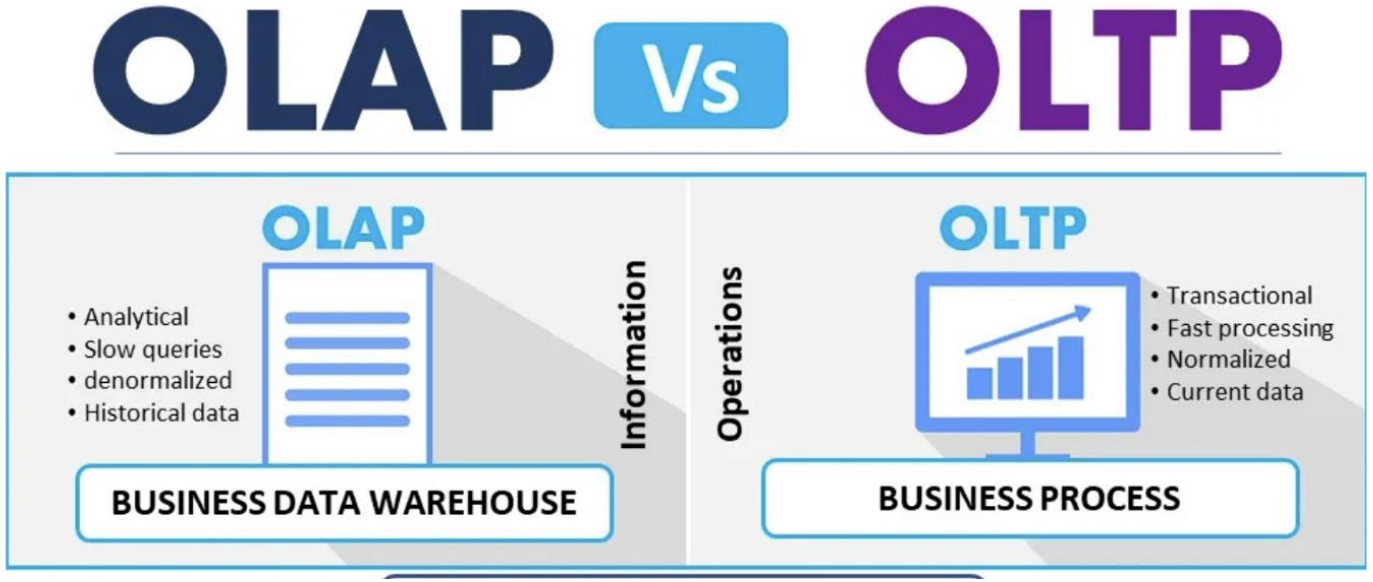
Data comes in a wide-ranging variety of file formats being collected from a variety of data **sources**:



Data engineering ecosystem: data storage

Online Analytical Processing (OLAP) Systems

- Optimized for conducting complex data analytics
- Include relational and non-relational databases, data warehouses, data marts, data lakes, and big data stores

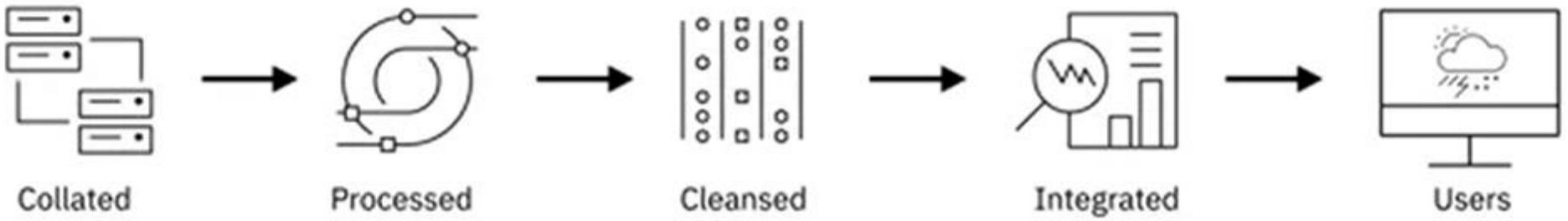


Online Transaction Processing (OLTP) Systems

- Designed to store high volume day-to-day operational data
- Typically relational, but can also be non-relational

Data engineering ecosystem: data integration

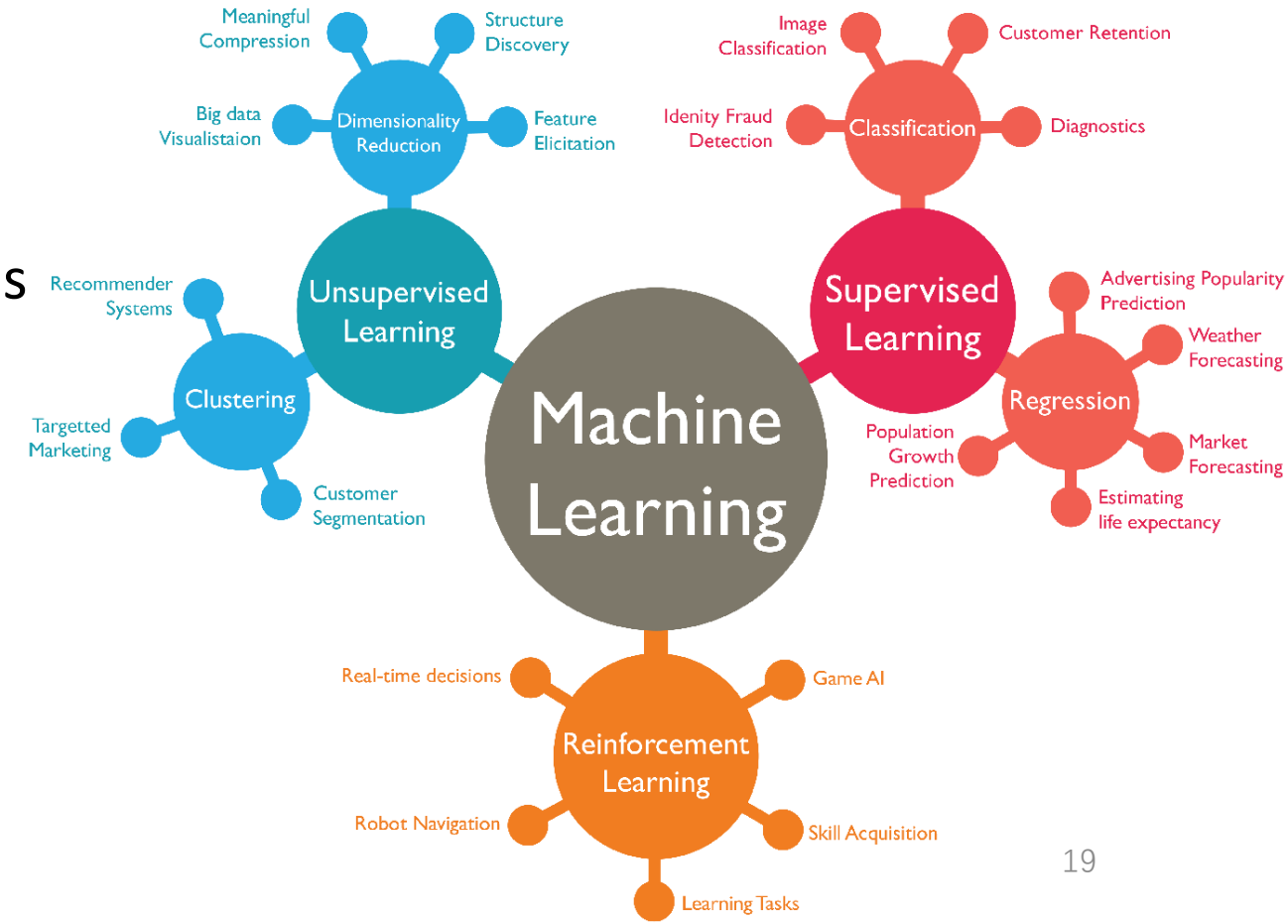
Combine data from disparate sources into a unified view, accessed by users to query and manipulate the data.



Data engineering ecosystem: data analysis

Data analysis is to discover useful information, informing conclusions, and supporting decision-making from data, in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.

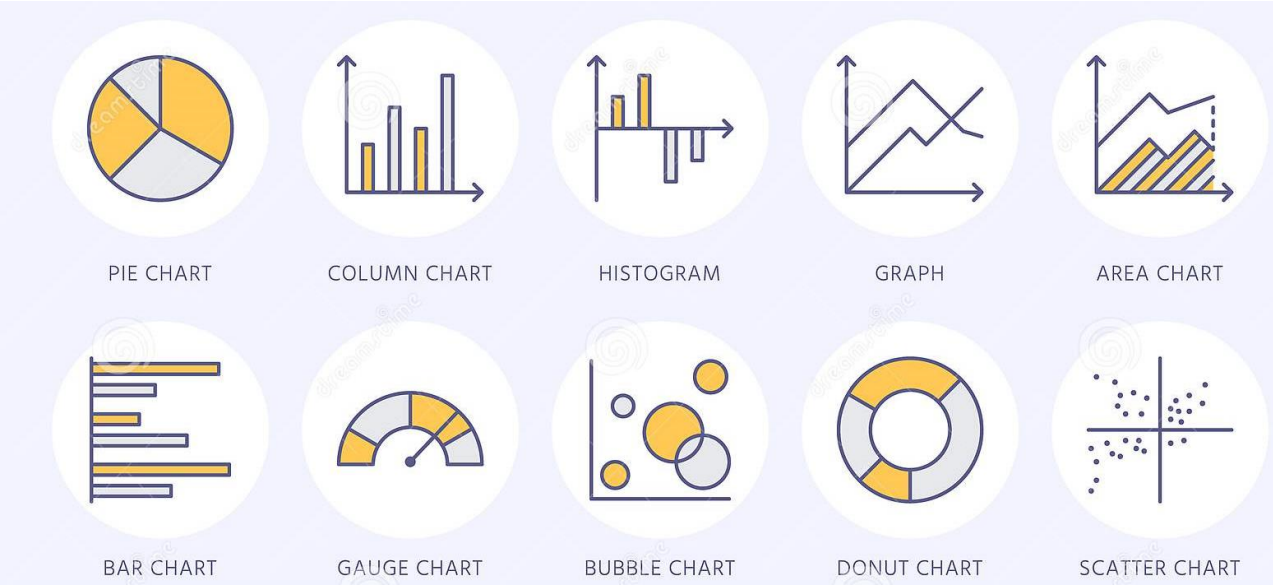
It often involves data mining and machine learning techniques to model and process the data.



Data engineering ecosystem: data visualization

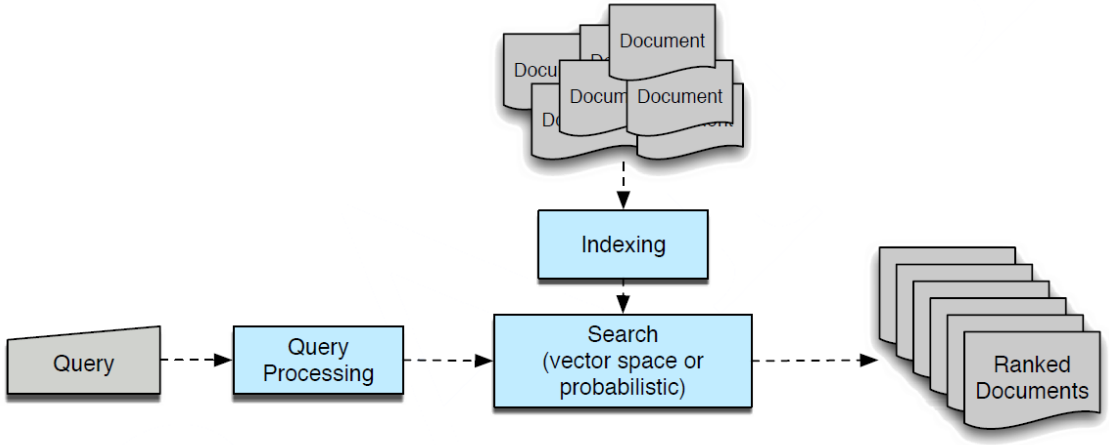
Business Intelligence (BI) and Reporting Tools

- Collect data from multiple data sources and present them in a visual format, such as interactive dashboards.
- Visualize data in real-time and predefined schedule.
- Drag and drop products that do not require knowledge of programming

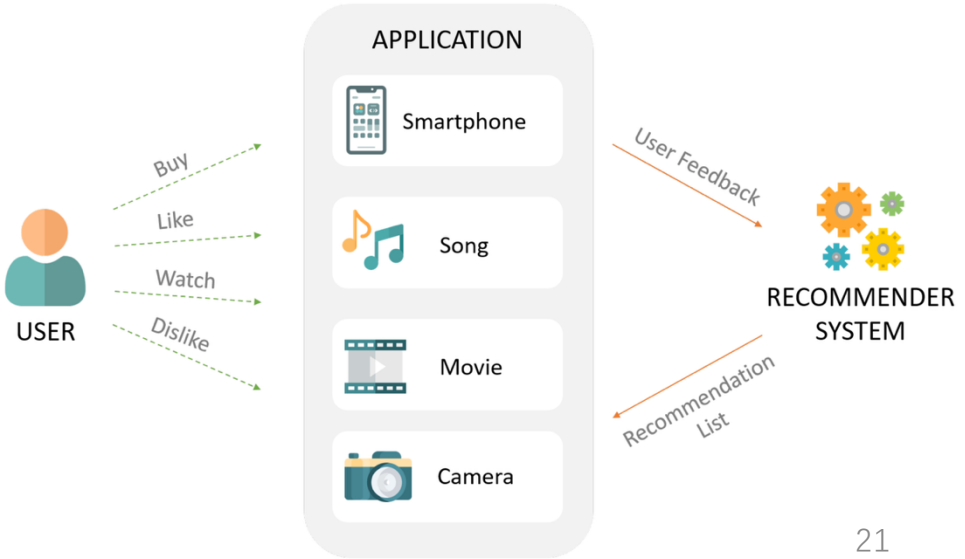


Data engineering ecosystem: data-driven applications (1)

- Information retrieval (IR) is the process of obtaining information system resources that are relevant to an information need from a collection of those resources (texts, images or sounds). Searches can be based on full-text or other content-based indexing.

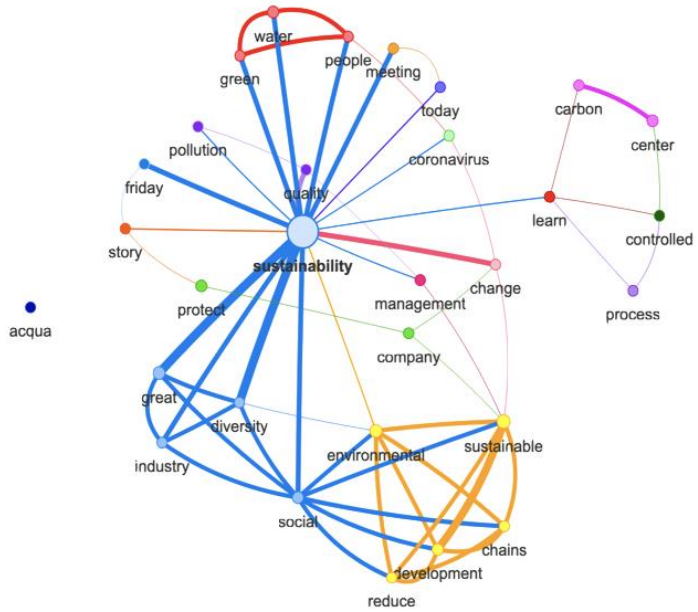


- A recommender system is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item.



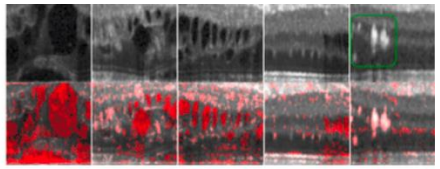
Data engineering ecosystem: data-driven applications (2)

- Social network analysis investigates social structures through the use of networks and graph theory. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them.

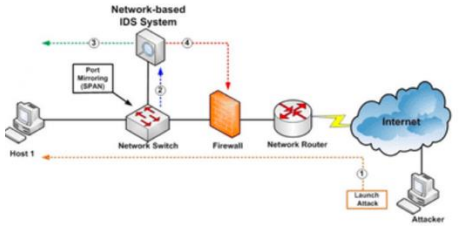


- Anomaly detection is the identification of rare events, items, or observations which are suspicious because they differ significantly from standard behaviors or patterns. Anomalies in data are also called standard deviations, outliers, noise, novelties, and exceptions.

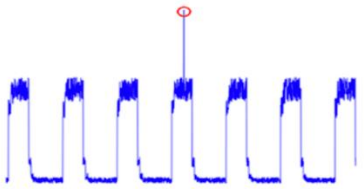
(a) Illegal Traffic Flow detection



(b) Detecting Retinal Damage



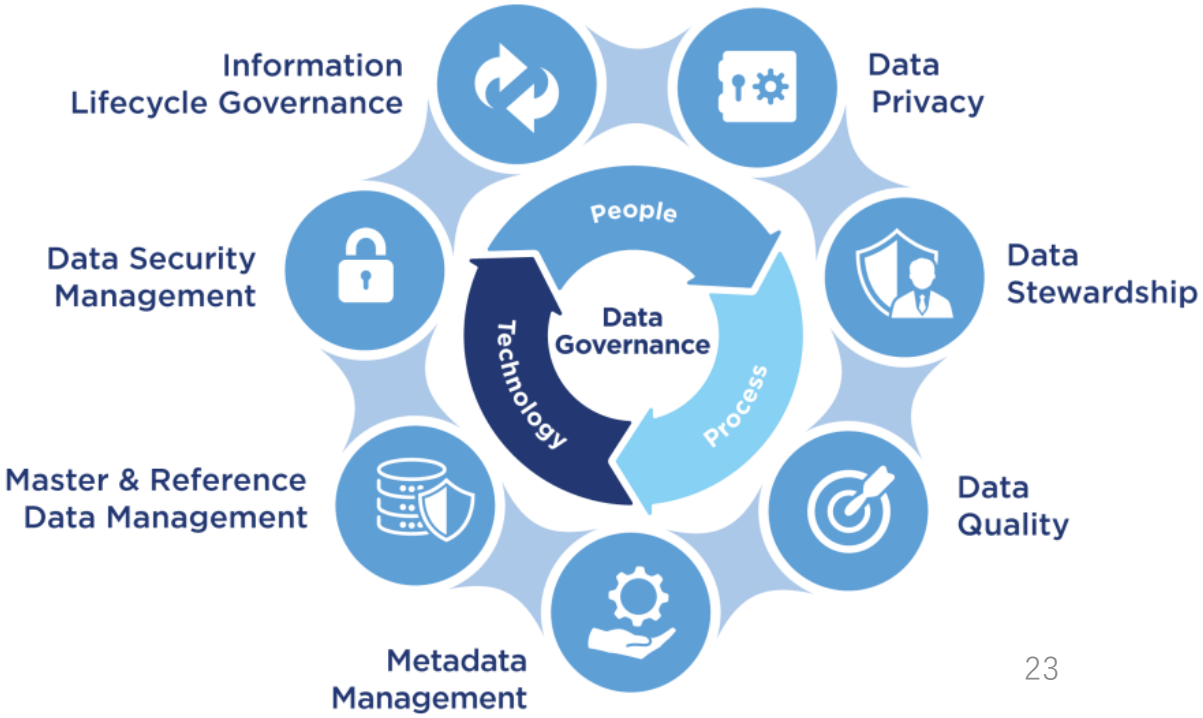
(c) Cyber-Network Intrusion detection



(d) Internet Of Things (IoT) Big-Data Anomaly detection

Data engineering ecosystem: data governance

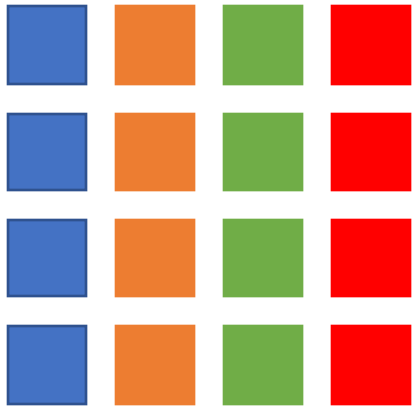
- Data quality: how well suited a data set is to serve its specific purpose, such as accuracy, completeness, consistency, validity, uniqueness, biasness, and timeliness.
- Data security: safeguarding data throughout its entire life cycle to protect it from corruption, theft, or unauthorized access. It covers everything—hardware, software, storage devices, and user devices; access and administrative controls; and organizations' policies and procedures.
- Data privacy: proper handling of sensitive data including personal data and other confidential data, such as certain financial data and intellectual property data, to meet regulatory requirements as well as protecting the confidentiality and immutability of the data.



Outline

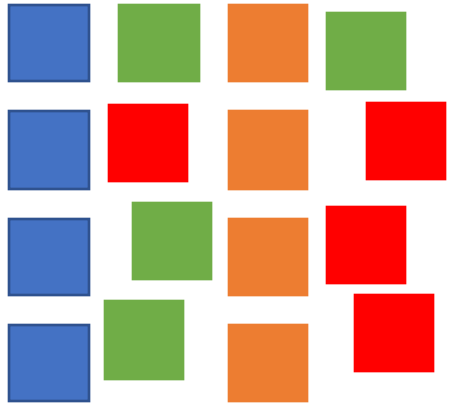
- 1. Course organization
- 2. Overview of the data engineering ecosystem
- 3. Types of data
- 4. Languages and tools for data professionals

Types of data



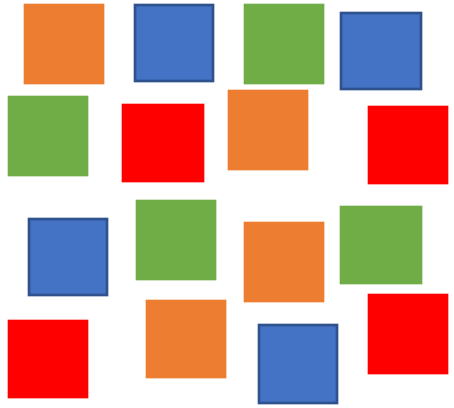
Structured

Data that follows a rigid format and can be organized into rows and columns.



Semi-Structured

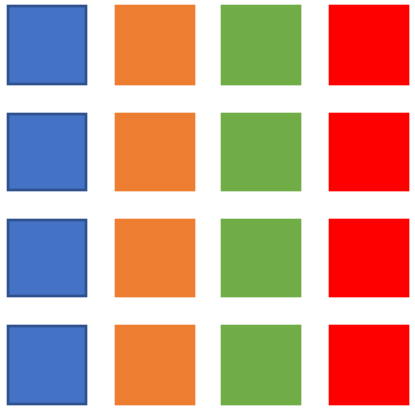
Mix of data that has consistent characteristics and data that does not conform to a rigid structure



Unstructured

Data that is complex and mostly qualitative information that cannot be structured into rows and columns

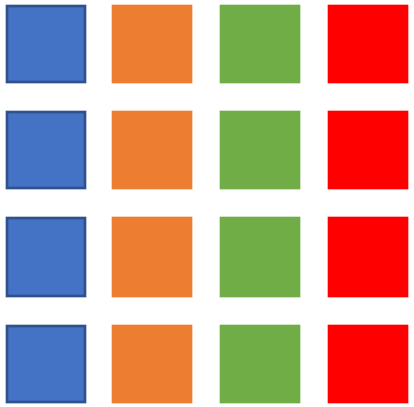
Structured data (1)



Structured data is objective facts and numbers that can be **collected**, **exported**, **stored**, and **organized** in typical databases.

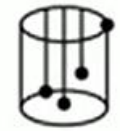
- Has a well defined structures.
- Can be stored in well-defined schemas.
- Can be represented in a tabular manner with rows and columns.

Structured data (2)



	A	B	C	D	E	F	G	H	I
1	Sales Representative	Location	Region	Customer	Order Date	Item	Quantity	Price	Total Sale Amount
2	Sara Snyder	New York	East	Phyllis Johnston	2016-10-30	Things	1	17.83	17.83
3	Sara Snyder	New York	East	Kimberly Little	2016-05-23	Junk	3	12.42	37.26
4	Frances Warren	Massachusetts	East	Justin Dixon	2016-09-27	Widgets	4	53.35	213.4
5	Sara Snyder	Massachusetts	East	Shirley Rivera	2016-02-12	Junk	5	12.42	62.1
6	Diane Gonzalez	Oregon	West	Marilyn Franklin	2016-02-14	Things	8	17.83	142.64
7	Patrick Graham	Washington	West	Henry Sanders	2016-04-11	Widgets	4	53.35	213.4
8	Sara Snyder	Connecticut	East	Benjamin Phillips	2016-09-02	Junk	4	12.42	49.68
9	Frances Warren	New Jersey	East	Theresa Torres	2016-11-26	Junk	4	12.42	49.68
10	Patrick Graham	Oregon	West	Roger Bell	2016-07-13	Junk	10	12.42	124.2
11	Sara Snyder	New Jersey	East	Harold Matthews	2016-06-02	Junk	3	12.42	37.26
12	Frances Warren	New York	East	Roy Young	2016-06-02	Widgets	8	53.35	426.8
13	Sara Snyder	New York	East	Debra Allen	2016-02-20	Things	1	17.83	17.83
14	Randy Watson	Connecticut	East	Alan Dean	2016-06-07	Junk	7	12.42	86.94
15	Randy Watson	Massachusetts	East	Robin Matthews	2016-10-31	Stuff	5	16.32	81.6
16	Randy Watson	New York	East	Randy Burton	2016-03-13	Stuff	4	16.32	65.28
17	Patrick Graham	Washington	West	Terry Nguyen	2016-02-10	Widgets	10	53.35	533.5

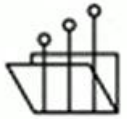
Sources of structured data includes:



SQL Databases

Online Transaction Processing

Spreadsheets

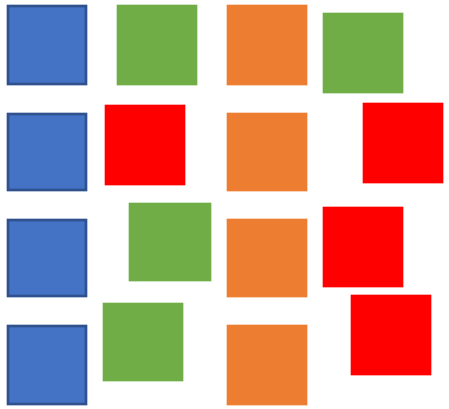


Online forms

Sensors GPS and RFID

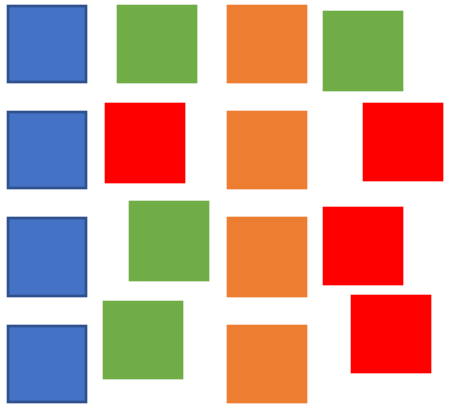
Network and Web server logs

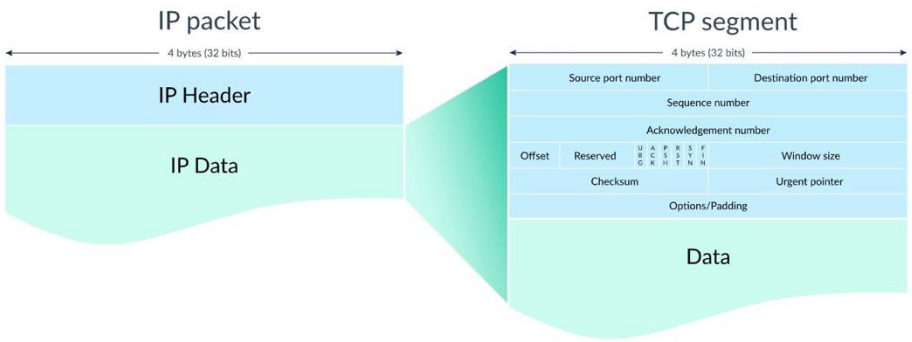
Semi-structured data (1)



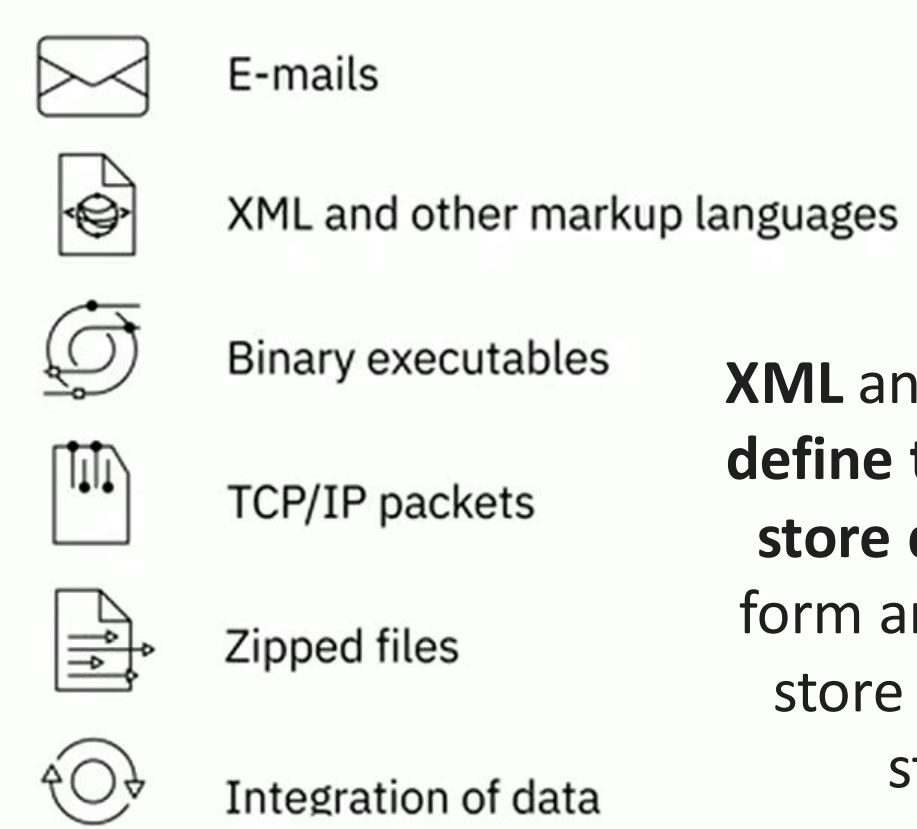
- Has some organizational properties but lacks a fixed or rigid schema.
- Cannot be stored in the form of rows and columns as in databases.
- Contains tags and elements, or metadata, which is used to group data and organize it in a hierarchy.

Semi-structured data (2)



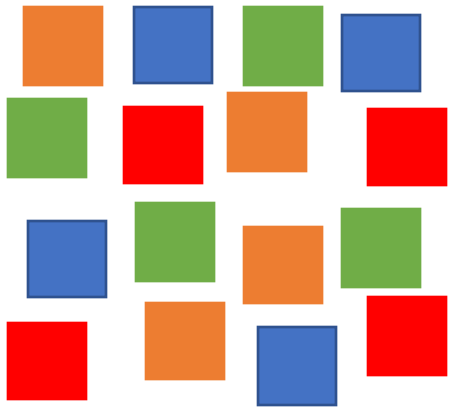


Sources of semi-structured data



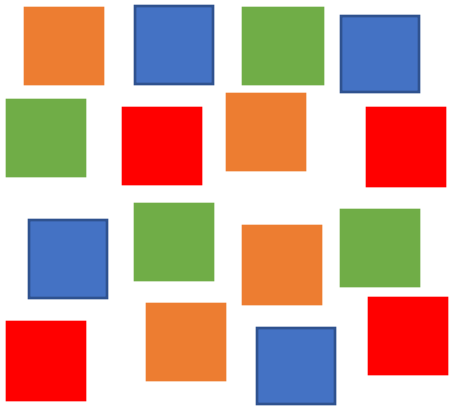
XML and **JSON** allow users to **define tags and attributes to store data** in a hierarchical form and are used widely to store and exchange semi-structured data.

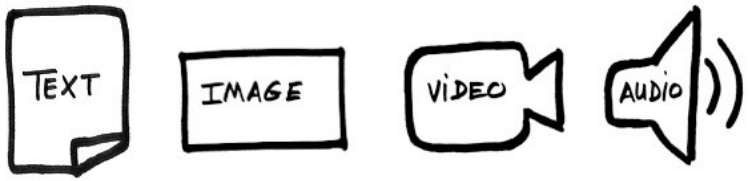
Unstructured data (1)



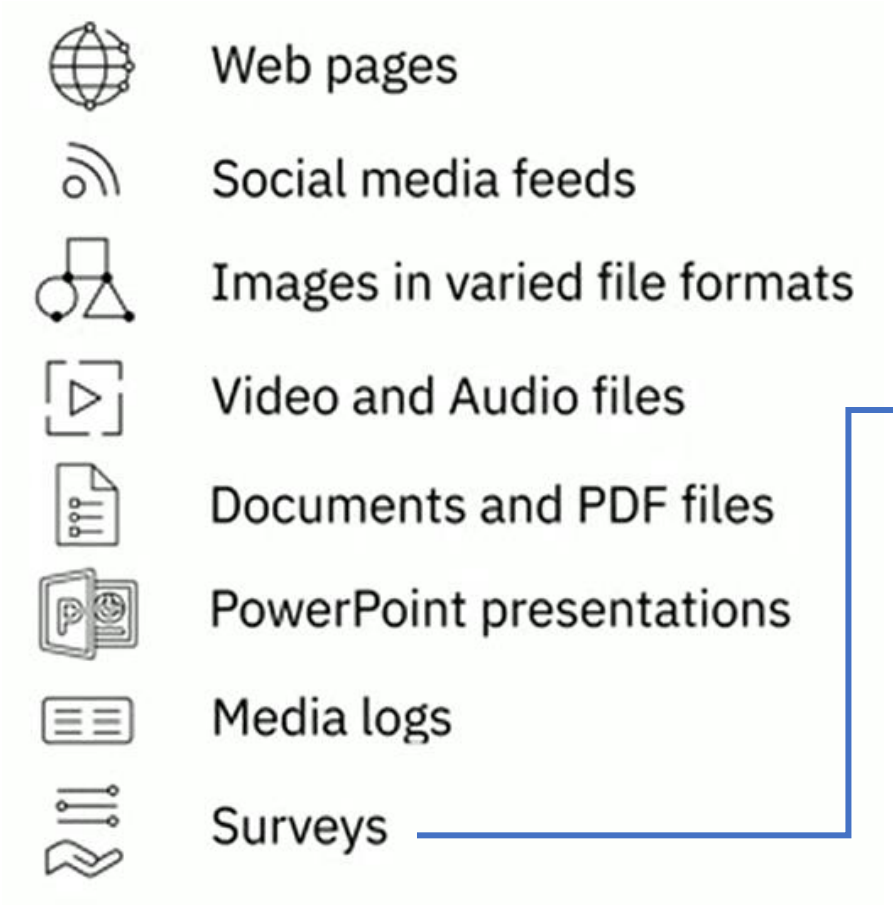
- Does not have an easily identifiable structure.
- Cannot be organized in a mainstream relational database in the form of rows and columns.
- Does not follow any particular format, sequence semantics, or rules.

Unstructured data (2)





Sources of unstructured data



Examples of different types of data

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
 </Student>
....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Outline

- 1. Course organization
- 2. Overview of the data engineering ecosystem
- 3. Types of data
- 4. Languages and tools for data professionals

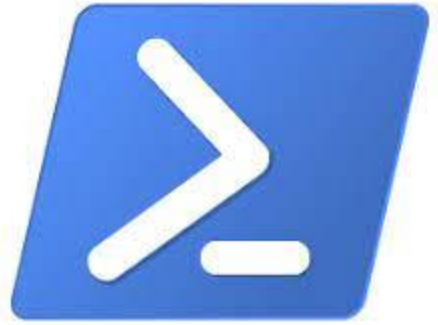
Languages for data professionals

Query Languages

For example, SQL for querying and manipulating data

Programming Languages

For example, Python for developing data applications



Shell and Scripting Languages

For repetitive and time-consuming operational tasks

Query languages

Advantages of using SQL:

- SQL is portable and platform independent.
- Can be used for querying data in a wide variety of databases and data repositories.
- Has a simple syntax that is similar to the English language
- Can retrieve large amount of data quickly and efficiently
- Runs on an interpreter system.

General programming languages – Python

Python is one of the fastest-growing programming languages in the world.

Advantages of using Python:

- Easy to learn and Open-source
- Can be ported to multiple platforms and has widespread community support.
- Provides open-source libraries for data manipulation, data visualization, statistics, mathematics.

Statistical programming languages – R

Advantages of using R-programming:

- Open-source and platform-independent.
- Can be paired with many programming languages and highly extensible.
- Facilitates the handling of structured and unstructured data.
- Can be used for developing statistical tools.

General programming languages – Java

Java is an object-oriented, class based, and platform-independent programming language.

Advantages of using Java:

- One of the top-ranked programming languages used today.
- Used in a number of data analytics processes – cleaning data, importing and exporting data, statistical analysis, data visualization.
- Used in the development of big data frameworks and tools – Hadoop, Hive, Spark
- Well-suited for speed critical projects.

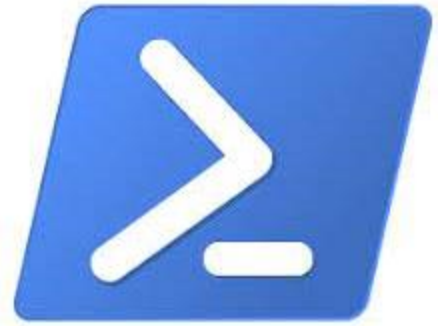
Shell and scripting languages

Typical operation performed by shell scripts include:

- File manipulation
- Program execution
- System administration tasks such as dis backups and evaluating system logs
- Installation scripts for complex programs
- Executing routine backups
- Running batches.

Shell and scripting languages - PowerShell

PowerShell is a cross-platform automation tool and configuration framework by Microsoft that is optimized for working with structured data formats.



- Consists of command-line shell and scripting language
- Object-based and can be used to filter, sort, measure, group, and compare objects as they pass through a data pipeline.
- Used for data mining, building GUIs, creating charts, dashboards, and interactive reports.

香港城市大學
City University of Hong Kong

Thanks for your attention!

References

1. https://www.coursera.org/learn/introduction-to-data-engineering

2. https://macxima.medium.com/data-engineering-572733412d54

3. https://www.analyticsvidhya.com/blog/2021/06/data-engineering-concepts-and-importance/