LECTURER: Nghia Duong-Trung

DATA UTILIZATION

WHO I AM

- Name: Dr. Nghia Duong-Trung
- Senior Researcher / Project Lead @ The German Research Center for Artificial Intelligence (DFKI GmbH)
- Academic Teacher @ IU
  - Teaching courses: Machine Learning, Deep Learning, Artificial Intelligence, Data Science, Data Utilization, Neural Nets and Deep Learning, Introduction to Computer Science
  - Thesis Supervision.

- Profile: https://sites.google.com/ismll.de/duongtrungnghia/

INTRODUCTION TO DATA UTILIZATION_DLMBBD01

- Course book: Data Utilization_DLMBBD01, provided by IU, myCampus.
- Basic Reading DLMBBD01, provided by IU, myCampus.
- Video gallery: Sprint Session by Sheikh Radiah Rahim Rivu, myCampus.
- Practical exam example, provided by IU, myCampus.
- Online tests and evaluation, provided by IU, myCampus.
- Additional teaching materials:

https://github.com/duongtrung/IU-Data-Utilization-DLMBBD01

The GitHub repository is the additional teaching materials, consisting of extra slides regarding the current information on the learning domain. It helps students in discussion and group work. It does not necessarily reflect new questions in examination, and it does not present the IU on those extra content.

TOPIC OUTLINE

TOPIC OUTLINE

Problem-Solving	5.2
Decision Support	6.1
Data Security and Data Protection	6.2

UNIT 1

INTRODUCTION TO DATA UTILIZATION

On completion of this unit, you will have learned…

- … definitions of data, information, and knowledge.
- … the main characteristics of big data.
- … the different categories of data.
- … general frameworks for data utilization and business analytics.
- … different types of pattern recognition problems and their applications.

EXPLAIN SIMPLY

1. What are the stages from raw data to understanding?
2. What distinguishes *big data* from conventional datasets?
3. Describe the general framework of data utilization.

- Industry 4.0 (fourth industrial revolution) has turned **everything *smart***.
  - based on data exchange automation
  - smart manufacturing
  - intelligent robots
  - AI
  - IoT
- Progress made in tech. results in high volumes of data generation.

DATA, INFORMATION, AND KNOWLEDGE

DATA
Those elements that are taken: extracted through observation, computation, experiments and record-keeping; partial, selective and representative—does not speak for itself

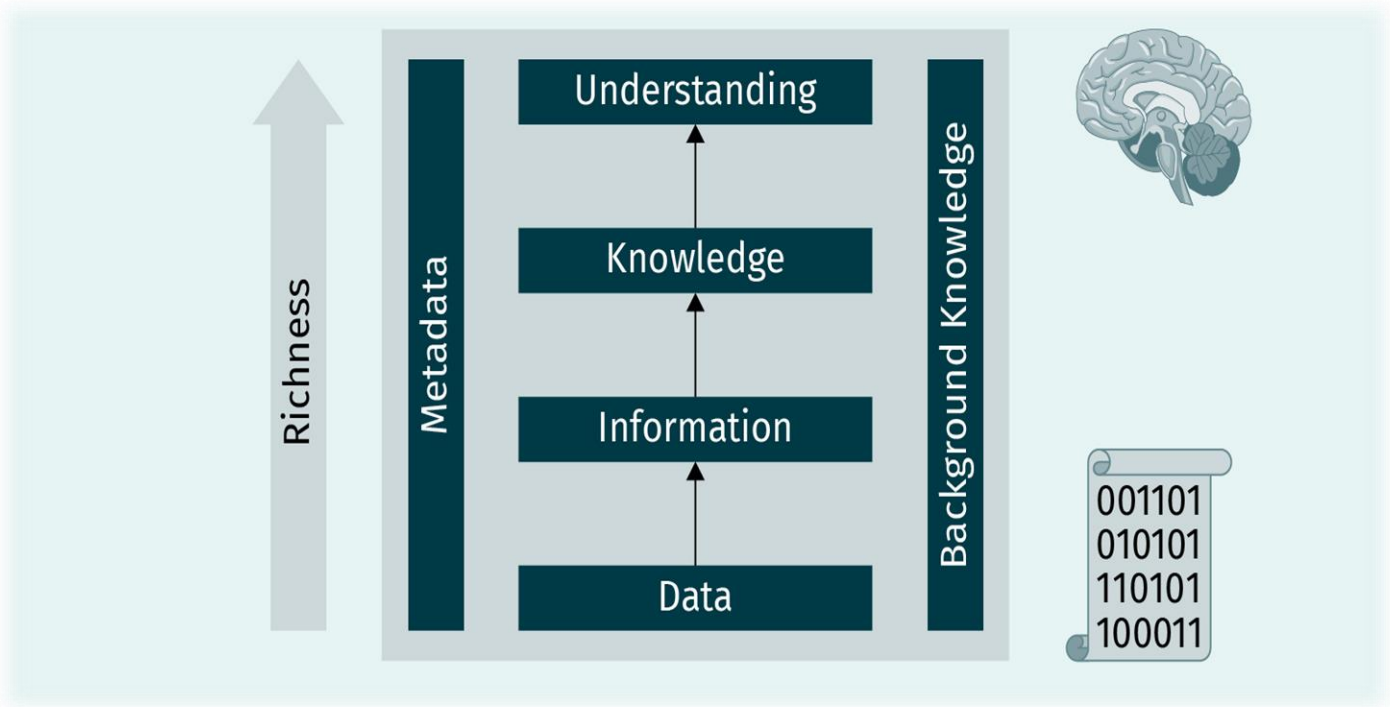
INFORMATION
Data in context: a set of data along with its explanations, reasons and interpretations

METADATA
Data about data/information: data providing information about one or more aspects of the data (e.g., descriptive summaries, high-level categorization of data and information

KNOWLEDGE
Organized, synthesized, or summarized information to enhance comprehension, awareness, or understanding: metadata + awareness of the context

BACKGROUND KNOWLEDGE
Prior knowledge; abstract notions deep inside the subconscious mind

Different people with varying background knowledge will understand **different** things from the **same information**; essentially, the **cognition process** depends on **background knowledge.**



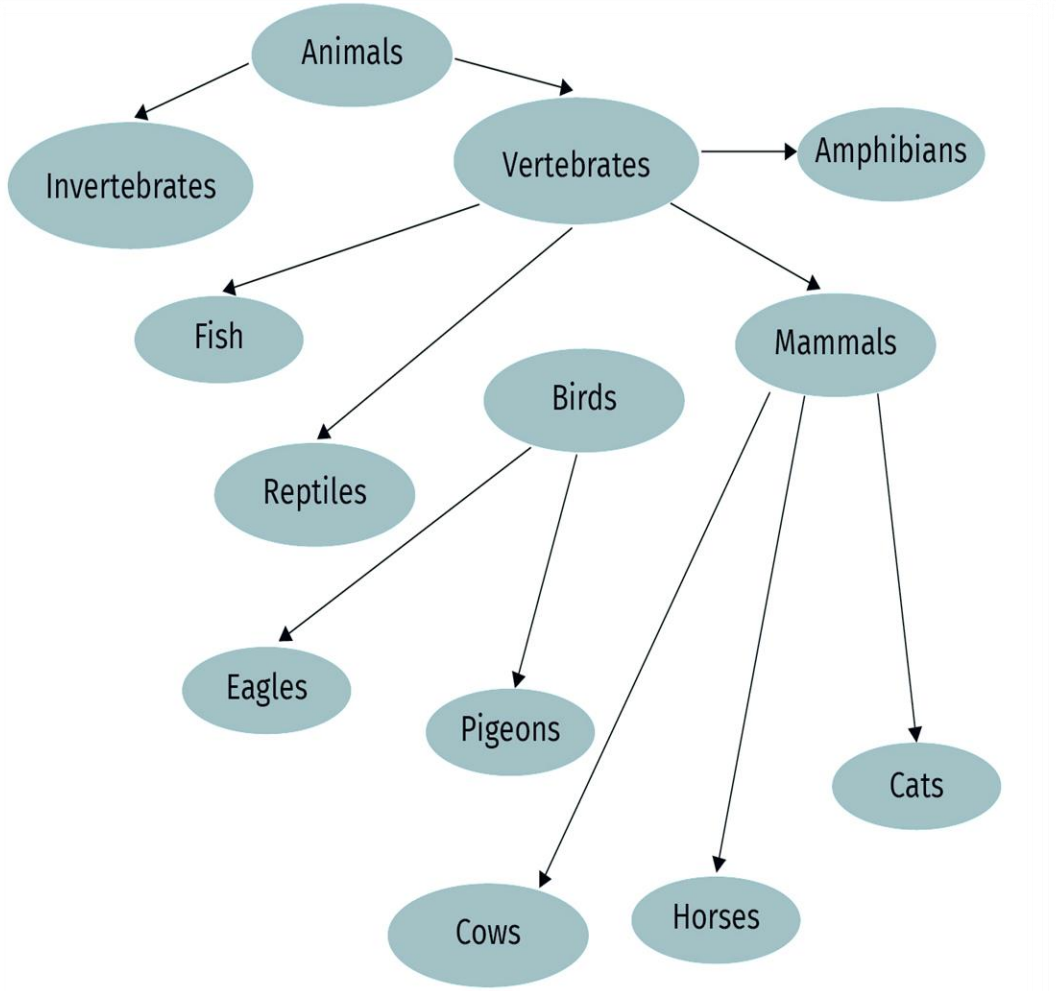
Source of the graphic: Course Book DLMBBD01, p. 15

ONTOLOGY

- A way to express knowledge by means of defining **concepts,** their **properties**, and **interrelationships.**
- Organizes information and knowledge.
- Describes individuals (instances), classes (concepts), class attributes, and the relationships between concepts.
- **Diversity features:** differentiate members within a group (identifying feature)
- **Similarity features:** features coming from an upper set or superset

HIERARCHICAL STRUCTURE OF ONTOLOGY

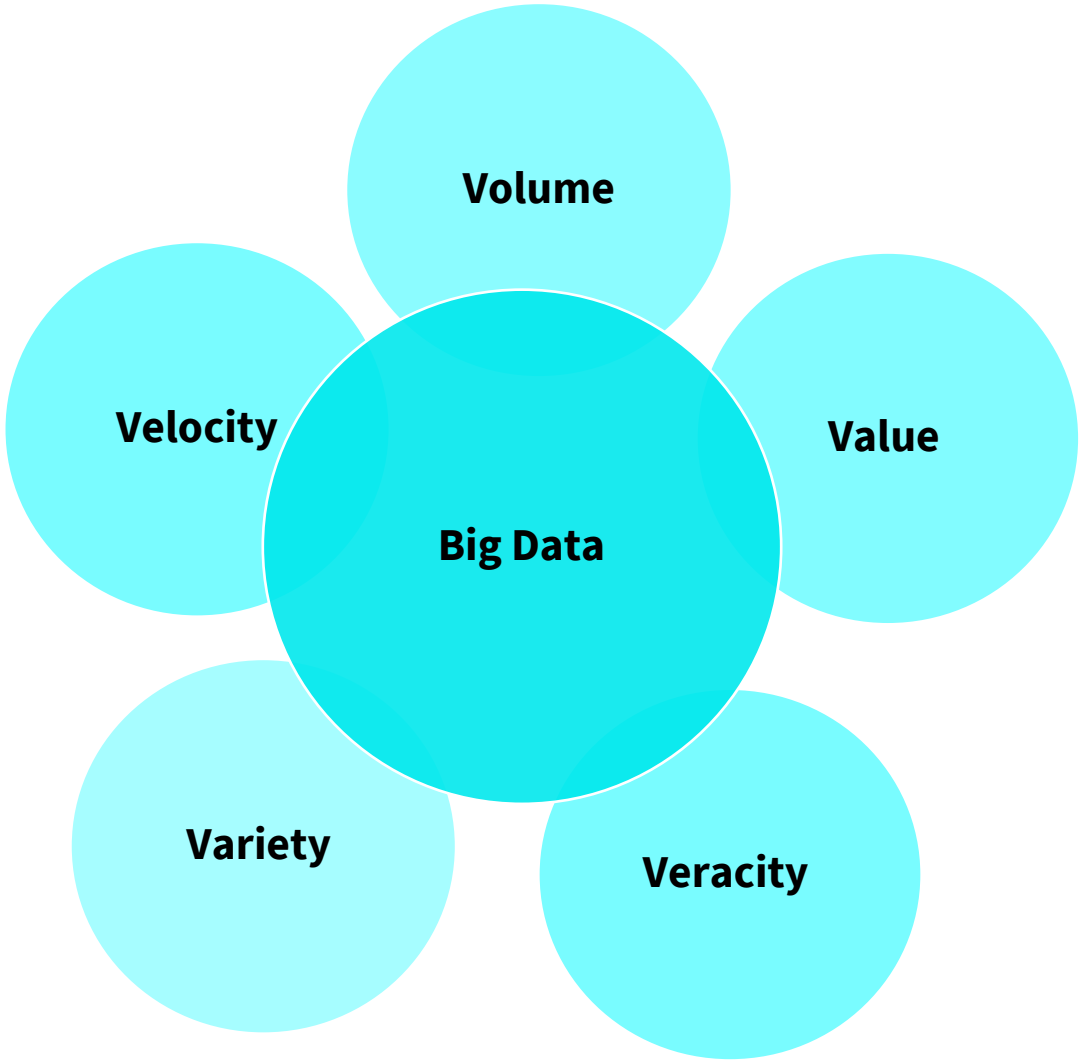
- Ontologies are hierarchical taxonomies used for representing knowledge in **generalized** and **specialized classes**
- Advantages:
  - Separating entities and operations
  - Reusing knowledge
  - Avoiding data repetition



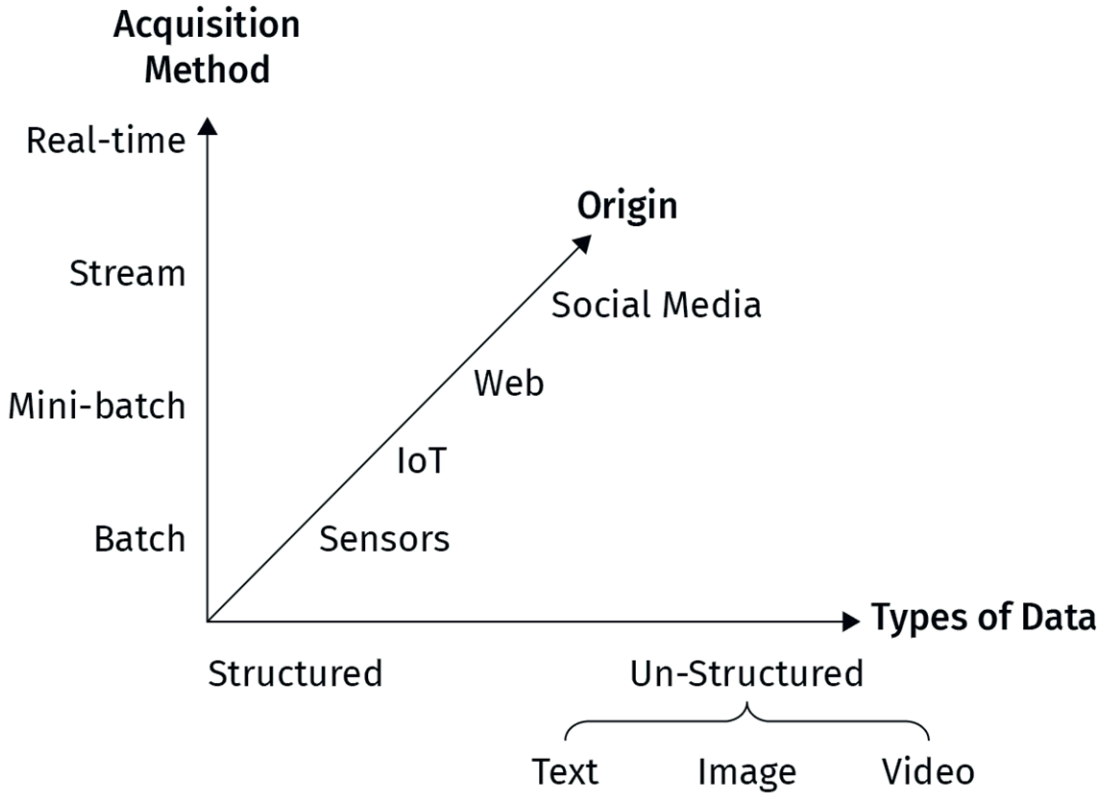
Source of the graphic: Course Book DLMBBD01, p. 18

BIG DATA

Large datasets that can *only* be analyzed by powerful computers to reveal **patterns.**



Data can be categorized by **type**, **origin**, and **method of acquisition.**



Source of the graphic: Course Book DLMBBD01, p. 20

TYPES OF DATA

- Structured data: tubular form and typically stored in relational databases (SQL, Oracle, Microsoft Excel)
- Spatial data: showing location of features, geographical in nature, discover which subsets are spatially autocorrelated or associated
- Image data: "picture is worth a thousand words ", image analysis extracts useful information (barcode reading, facial recognition, camera surveillance, and movement detection)
- Text: contains a depth of buried information

ORIGINS OF DATA

DATA SOURCES
(the subject is narrowed down to organizations)

Internal	External
– a primary source of info in organizations; operational information sys. – internal information sys.: ERP, CRM – structured data stored in relational data bases or kept in spreadsheets – unstructured data such as recorded meetings, photos and videos which may need to be analyzed in some cases	– publicly available and free (e.g., statistics issued by a government) or private and available for a fee(e.g., marketing reports) – world wide web and social media contain a huge amount of data to be analyzed – Web mining: learning about the distribution of info; classifying web pages – Mining comments in social media, demographic information, and social networks

ACQUISITION METHODS

Data Streams

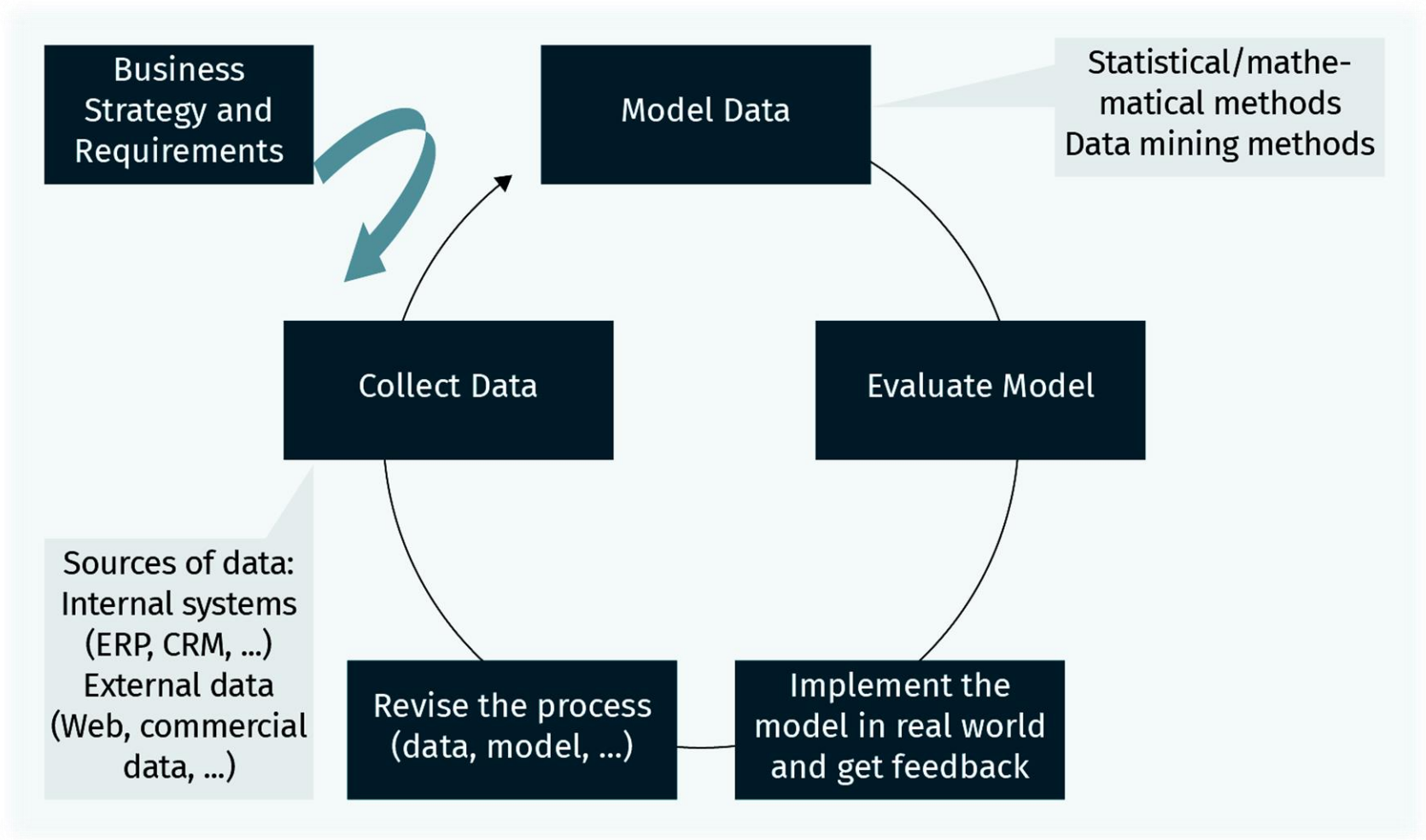
- Generated in a continuous manner and transmitted by different sources, such as IoT devices, sensors, and CCTV cameras.
- Processed incrementally as they are received.
- **Concept drift:** Statistical properties of the variables may change over time so predictions based on the data may be inaccurate.

Time-Related Data

- time-related and sequential in nature
- Stock exchange data can be mined to uncover trends that can help investment strategies.
- Mine computer network data streams to find intrusions. Such intrusions are discovered by analyzing the sequence of data or by comparing the current frequent pattern with those in the past.

- **Three-phase cycle** for data utilization:
  - collection
  - modelling
  - usage (experience delivery)
- **Six-phase cycle:**
  - Add business strategies and requirements (search questions concerning the subject at hand) to the beginning of the cycle.

GENERAL FRAMEWORK OF DATA UTILIZATION



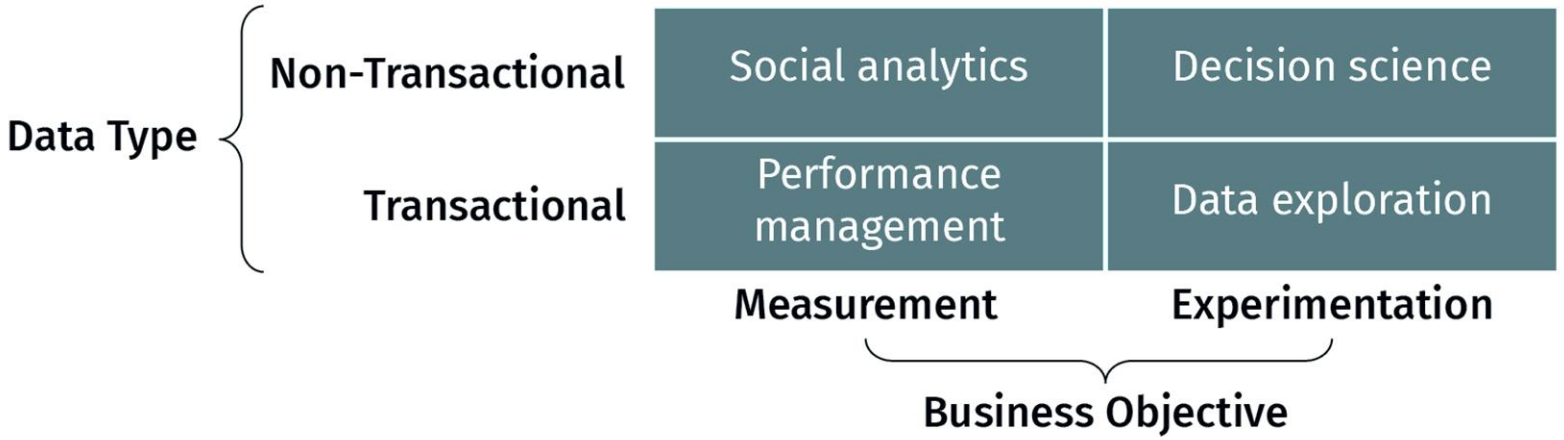
Source of the graphic: Course Book DLMBBD01, p. 23

Several models are proposed according to the type of application:

Analysis and description	Answers what question
Explanation	Concerned with understanding and answering *Why*-questions and providing insight into the behavior of individuals.
Prediction	Extracts knowledge from past events and exploits this knowledge to forecast future events with an acceptable degree of reliability.
Design and Action	Seeks to understand how people do certain activities.

SALVATORE MODEL

- Studies the relationship between data type and business objective.
- Sorts data into transactional and non-transactional.
- Categorizes business objectives into measurement and experiment.



Source of the graphic: Course Book DLMBBD01, p. 27

TRANSACTIONAL AND NON-TRANSACTIONAL DATA: AN EXAMPLE

- Transactional:
  - Has a *time* dimension that most businesses rely on to keep track of their day-to-day operations.
  - E.g., invoices, purchase orders, and receipts.
- Non-transactional:
  - information that is not used to complete a transaction. This type of data is typically used for informational or reference purposes only.
  - E.g., Customer name, contract number, addresses.

GARTNER MODEL - UTILIZATION FRAMEWORK

DESCRIPTIVE

- Includes performance analytics, context analytics, and research.
- Completed by research on the development of descriptive statistics around KPIs.
- Answers: What happened?

PREDICTIVE

- Builds a model to make predictions.
- Fulfilled through statistical and data mining methods.
- Answers: What might happen?

FOUR TYPES OF DATA ANALYTICS

DIAGNOSTIC

- It is performed by analyzing probabilities , filtering, etc.
- Answers: Why did it happen?

PERSPECTIVE

- Helps with the creation of recommendations regarding which specific steps to take or the most appropriate decisions to make in order to reach certain goals.
- Answers: What should we do?

Four Main Layers of Data Analytics		
		Tools used
Descriptive	What happened?	Data dashboard, KPI
Diagnostic	Why did it happen?	Probabilities, filtering
Predictive	What might happen?	Statistical and data mining methods (classification, regression, clustering)
Prescriptive	What should we do?	Simulation, optimization models

DATA ANALYSIS AND VISUALIZATION TOOL

- Microsoft Power BI
- Microsoft Excel
- https://dash.plotly.com
- https://github.com/mljar/mercury
- https://towardsdatascience.com/4-python-packages-to-create-interactive-dashboards-d50861d1117e
- https://mode.com/blog/python-data-visualization-libraries

Places objects into categories or classes, depending on the application. These objects/patterns can be images, signals, etc.

Kinds of patterns that can be mined

- Class/concept description: characterization and discrimation
- Classification and regression for predictive analysis (finds common patterns, characteristics, and defines a class to distinguish it from other classes)
- Cluster analysis: groups similar objects together; objects in the same group often have similar attributes, characteristics, and features
- Mining frequent patterns, associations and correlations: identifies patterns that occur often in the data
- Outlier analysis: interest in abnormal data (those that do not comply with general normal behavior)

On completion of this unit, you will have learned…

- … definitions of data, information, and knowledge
- … the main characteristics of big data
- … the different categories of data
- … general frameworks for data utilization and business analytics
- … different types of pattern recognition problems and their applications.

SESSION 1

TRANSFER TASK

TRANSFER TASK

Assume your company wants to analyze and describe its customers' behavior at the end of the year from their purchases and written feedback. Future production is intended to be planned based on this.

1. Define the type of acquired data.
2. Select a proper utilization framework and based on that explain the stages and questions that can be answered using the model .

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

1. “Information that is organized, synthesized, or summarized to enhance comprehension, awareness, or understanding” is the best description of…
   a) … knowledge.
   b) … information.
   c) … ontology.
   d) … metadata.

2. Which of the following types of data is structured?
   a) text
   b) image
   c) ERP system
   d) speech

3. Which of the following is <u>not</u> a benefit of using a hierarchical structure of ontologies for categorizing objects?
   a) sharing of operational knowledge
   b) reuse of knowledge
   c) avoidance of data repetition
   d) prediction of target variables

LIST OF SOURCES

Gregor, Shirley. (2006). The nature of theory in information systems. *MIS Quarterly 30*(3), 611-642.

Margaliot, M. (2008) Pattern Recognition (Theodoridis, S. and Koutroumbas, K.; 2006) [Book reviews]. *IEEE Transactions on Neural Networks 19*(2), 376-376. DOI 10.1109/TNN.2008.929642

Salvatore, P., Iyer, B. & Vesset, D. (2012). Four strategies to capture and create value from big data. *Ivey Business Journal 76*(4). 1-5.