LECTURER: TAI LE QUY

INTRODUCTION TO DATA SCIENCE

Who am I?

- Name: Tai Le Quy
- PhD at L3S Research Center – Leibniz University Hannover
  - Topic: Fairness-aware machine learning in educational data mining
- MSc in Information Technology at National University of Vietnam
- Profile: tailequy.github.io
- Email: tai.le-quy@iu.org
- Materials: https://github.com/tailequy/IU-IntroDS

INTRODUCTORY ROUND

Who are you?

- Name
- Employer
- Position/responsibilities
- Fun Fact
- Previous knowledge? Expectations?

INTRODUCTION TO DATA SCIENCE
TOPIC OUTLINE

INTRODUCTION TO DATA SCIENCE	1
DATA	2
DATA SCIENCE IN BUSINESS	3
STATISTICS	4
MACHINE LEARNING	5
SUMMARY SESSION	6

UNIT 1

INTRODUCTION TO DATA SCIENCE

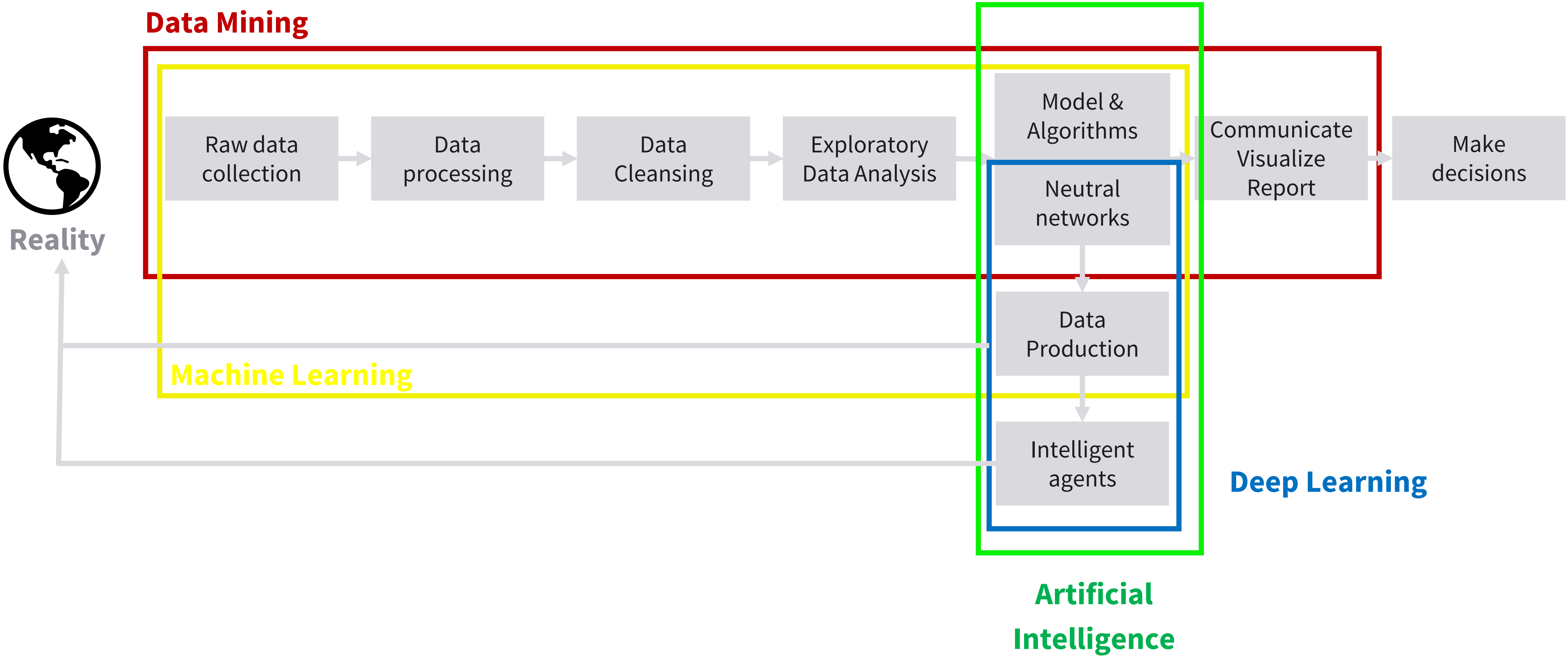
STUDY GOALS

- Understand what is meant by data science and why we need data science.
- Understand the main terms and definitions relating to data science.
- Explain the role of a data scientist.
- Describe the typical activities carried out within the field of data science.

EXPLAIN SIMPLY

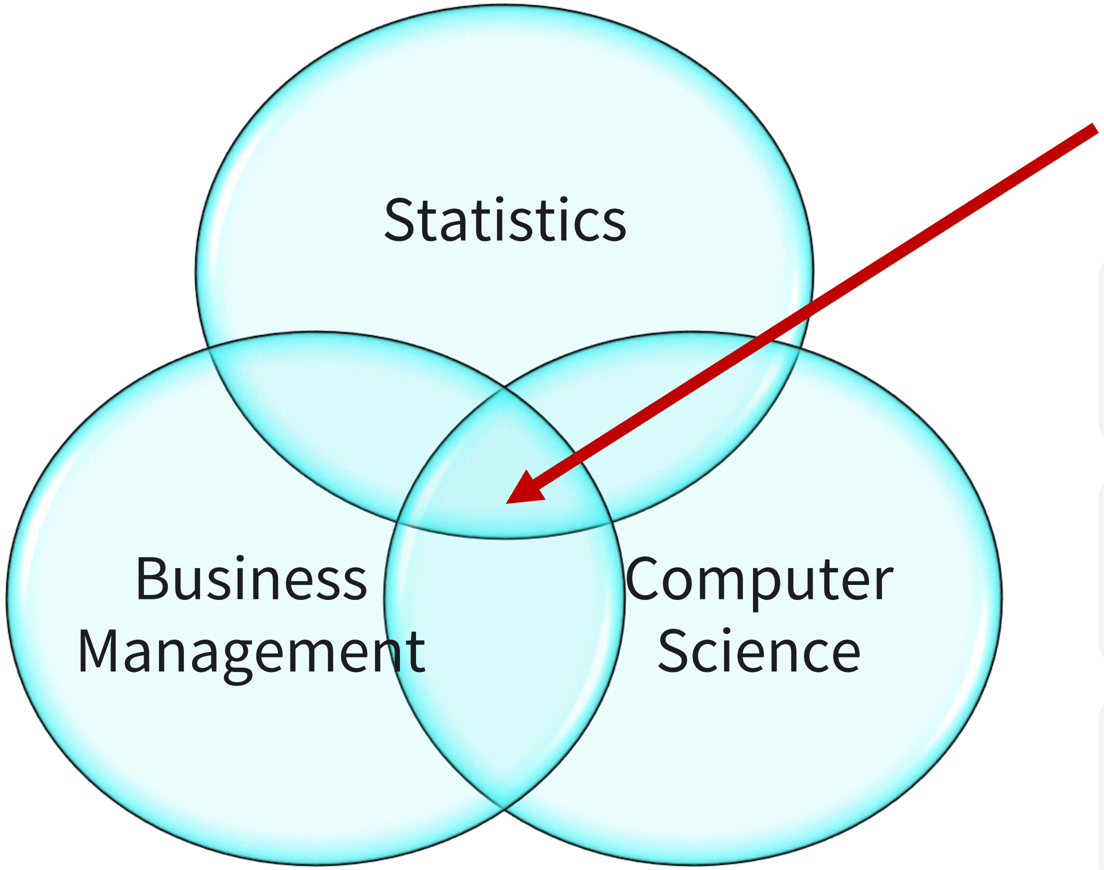
- What is data science?
- What are the benefits of data science?
- What fields are related to data science and how?

DATA SCIENCE



Source of the image: Own creation based on Altexsoft, 2021.

DATA SCIENCE’S RELATED FIELDS
THE DATA SCIENCE VENN DIAGRAM



DATA SCIENCE

Extracts meaningful **insights** from **raw data**.

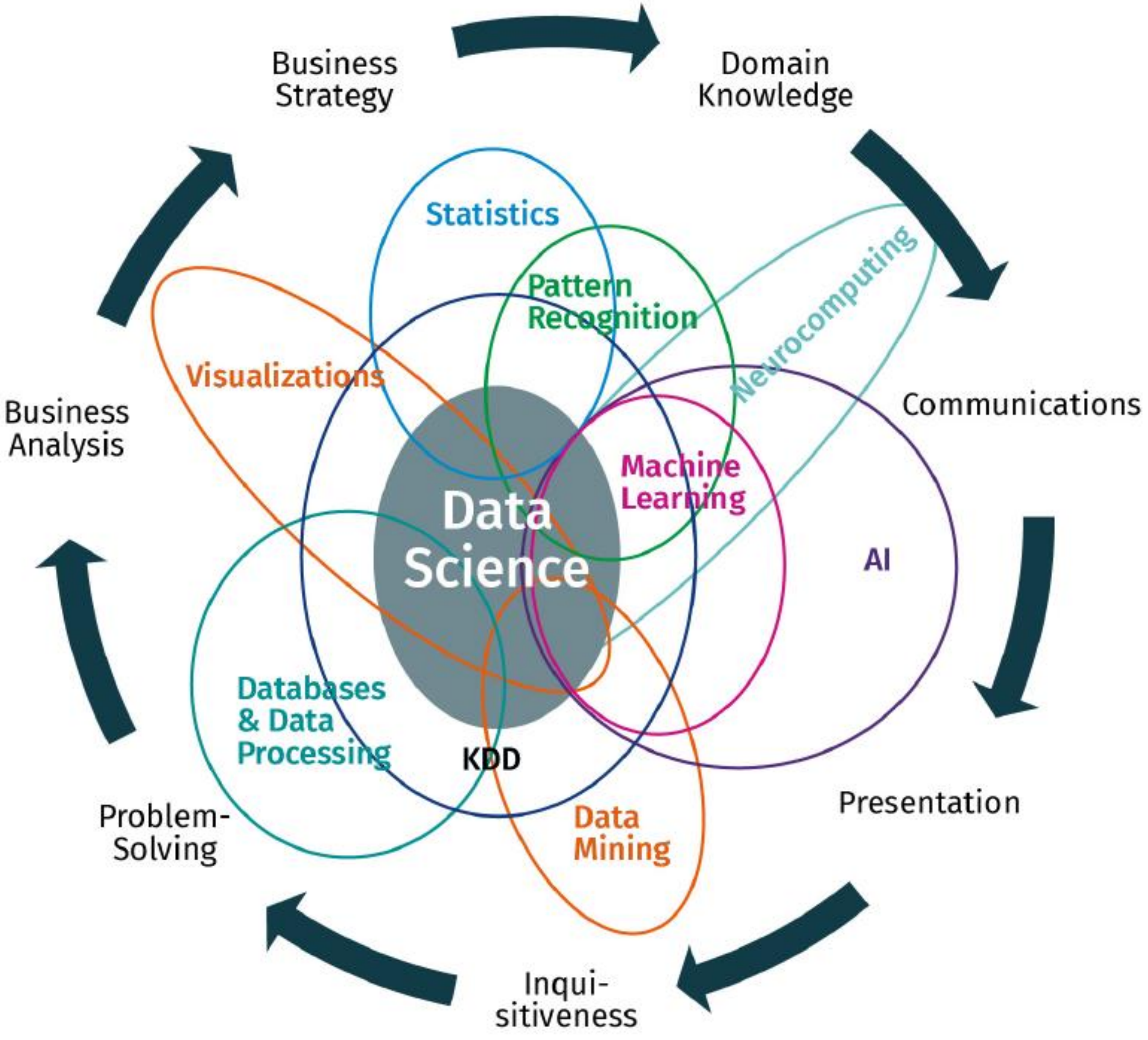
Unlocking the **real values** and **insights** of the data

Focused on the **ways** that people can **understand** and **use** data.

Enable companies to make **smarter business decisions.**

Source of the text: Pollock, 2018; Saleh, 2014.
Source of the image: Zöller, 2020, p.15.

THE EXTENDED DATA SCIENCE VENN DIAGRAM



Source: Tierney (2012)

BENEFITS OF DATA SCIENCE

Improves the **decision-making** of the company.

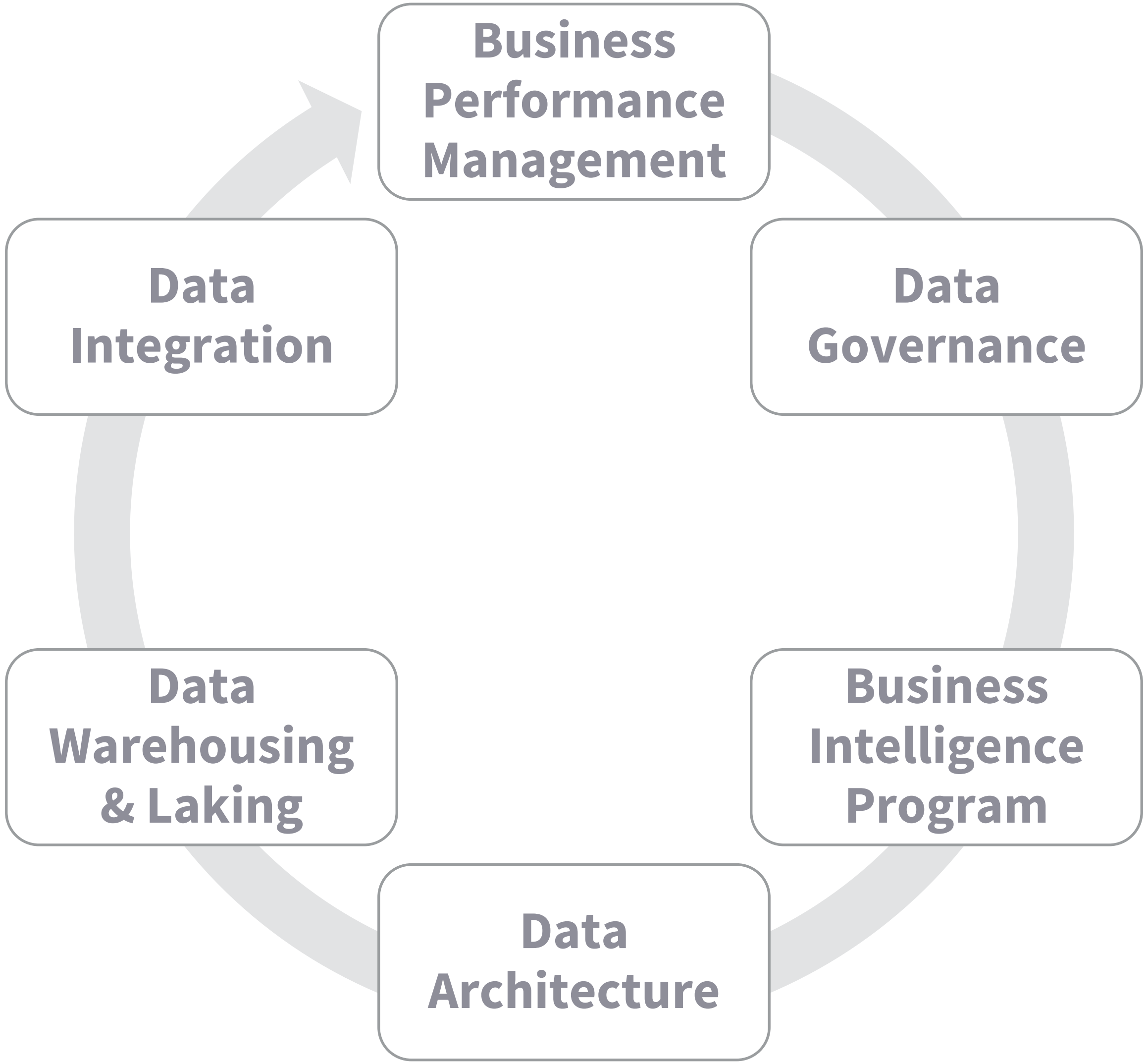
Enhances **operational efficiency**, business routine, and workflows.

Recognizes and informs companies of their **target audiences**.

Assists the automated aspect of **HR recruitment** to perform more accurately.

Source of the image: Own creation.

BUSINESS INTELLIGENCE



Business Performance Management

Data Governance

Business Intelligence Program

Data Architecture

Data Warehousing & Laking

Data Integration

Source of the image: Microsoft Archive.

BUSINESS INTELLIGENCE

Benefits

- Data driven business decisions
- Increased efficiency
- Boost ROI
- Increased competitive advantage
- Improved customer experience

Source of the image: Microsoft Archive.

DATA SCIENCE TERMS

Data Handling

Training Set	The **dataset** used to learn the desired task.
Testing Set	Assesses the **performance** of machine learning model.
Outlier	A **data record**
Data Cleansing	The **process** of removing redundant data, etc.

Data Features

Feature	**Measure** of the data; height, etc.
Dimensionality Reduction	The process of **reducing the dataset.**
Feature Selection	The process of **selecting relevant features.**

Source of the image: Zöller, 2020, p.17.

Learning Paradigms

Machine Learning

- **Algorithms** or **mathematical** models
- Uses information to achieve a **desired task or function**.

Supervised Learning

- The subset of Machine Learning, based on **labeled data**.
- Distinguished in **regression** and **classification**.

Unsupervised Learning

- The subset of Machine Learning, based on **unlabeled data**.
- **Clustering** and **dimensionality reduction**.

Deep Learning

- The application of **networks** of computational units.
- Used to **learn** through tasks.

Source of the image: Zöller, 2020, pp.17-18.

DATA SCIENCE TERMS

Model Development

Decision Model

- Assesses the data to **recommend a decision**.

Regression

- Estimates the **dependence** between variables.

Cluster Analysis

- A set of **data records** into **clusters**.

Classification

- Categorizes entities into **predefined classes**.

Model Performance

Probability

- How **likely** it is that a certain **event occurs**.

Standard Deviation

- How spread out the **data values** are.

Type I Error

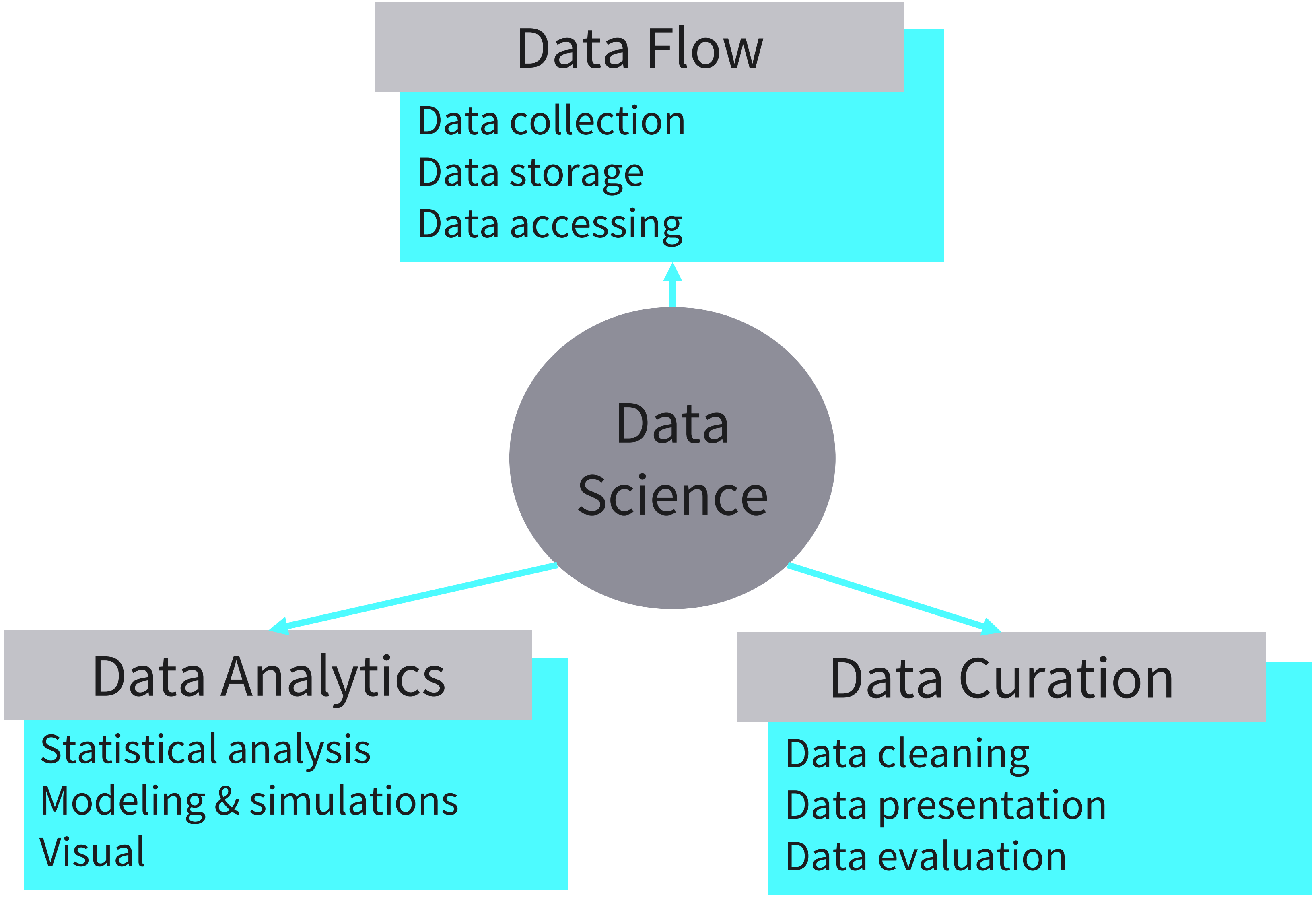
- False **positive** output.

Type II Error

- False **negative** output.

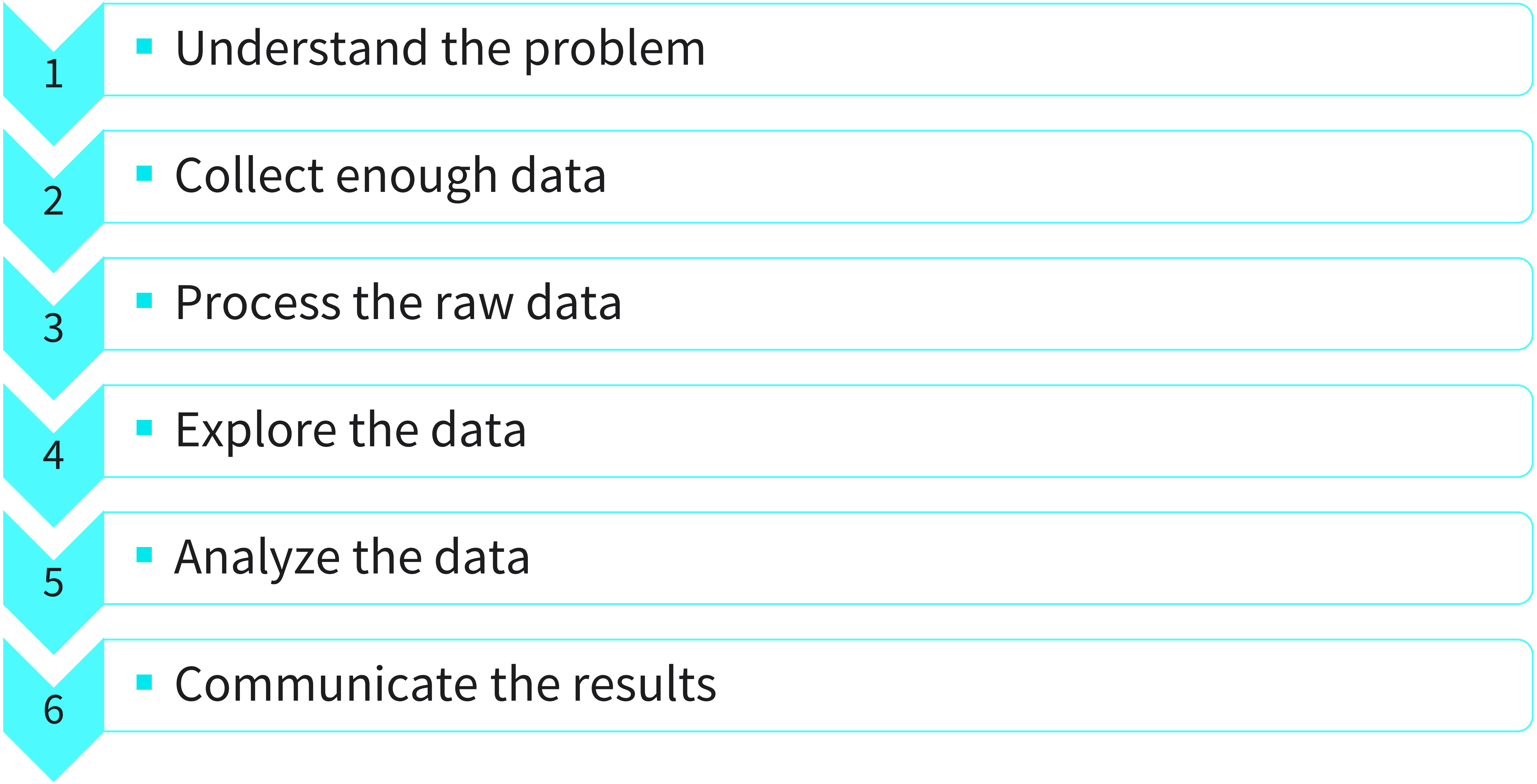
Source of the image: Zöller, 2020, pp.18-19.

DATA SCIENCE’S ACTIVITIES



Source of the image: Zöller, 2020, p.20.

DATA SCIENCE’S ACTIVITIES



1. Understand the problem
2. Collect enough data
3. Process the raw data
4. Explore the data
5. Analyze the data
6. Communicate the results

simplilearn

TOP 10
DATA SCIENCE
PROJECTS FOR 2024

REVIEW STUDY GOALS

- Understand what is meant by data science and why we need data science.
- Understand the main terms and definitions relating to data science.
- Explain the role of a data scientist.
- Describe the typical activities carried out within the field of data science.

SESSION 1

TRANSFER TASK

TRANSFER TASKS
CASE STUDY

Scenario

John is a data scientist working in a team of Business Intelligence. He is going to start a new data science project to improve the **marketing process** of the company.

Questions

1. Which tasks may John be responsible for?
2. Which benefits can the company achieve from John’s data science activities?

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

LEARNING CONTROL QUESTIONS

1. Machine learning is a set of algorithms or mathematical models that use information extracted from data in order to achieve a desired task or function.

   a) Correct

   b) Incorrect

LEARNING CONTROL QUESTIONS

2. Cluster analysis is a type of supervised learning used to partition a set of data records into clusters.

   a) Correct

   b) Incorrect

LEARNING CONTROL QUESTIONS

3. Data scientists follow a group of actions that encompasses all possible elements of the process that need to be addressed. Put the various steps in the correct order.
   - Understand the problem
   - Explore the data
   - Communicate the results
   - Process the raw data
   - Collect enough data
   - Analyze the data

LIST OF SOURCES

Text

Zöller, T. (2020). *Introduction to Data Science*. IU International University of Applied Science.

Pollock, N. J., Healey, G. K., Jong, M., Valcour, J. E., & Mulay, S. (2018). Tracking progress in suicide prevention in Indigenous communities: A challenge for public health surveillance in Canada. *BMC Public Health*, 18(1320). Retrieved from https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-018-6224-9

Saleh, B., Abe, K., Arora, R. S., & Elgammal, A. (2014). Toward automated discovery of artistic influence. *Multimedia Tools and Applications*, *75*, 3565—3591.

Images

Altexsoft. (2021). *Data science vs machine learning vs AI vs deep learning vs data mining: Know the differences*. https://www.altexsoft.com/blog/data-science-artificial-intelligence-machine-learning-deep-learning-data-mining/

Zöller, 2020, p.15.

Zöller, 2020, p.17.

Zöller, 2020, pp.17-18.

Zöller, 2020, pp.18-19.

Zöller, 2020, p.20.