LECTURER: TAI LE QUY

INTRODUCTION TO DATA SCIENCE

UNITS 1-5

REVIEW OF UNITS 1-5

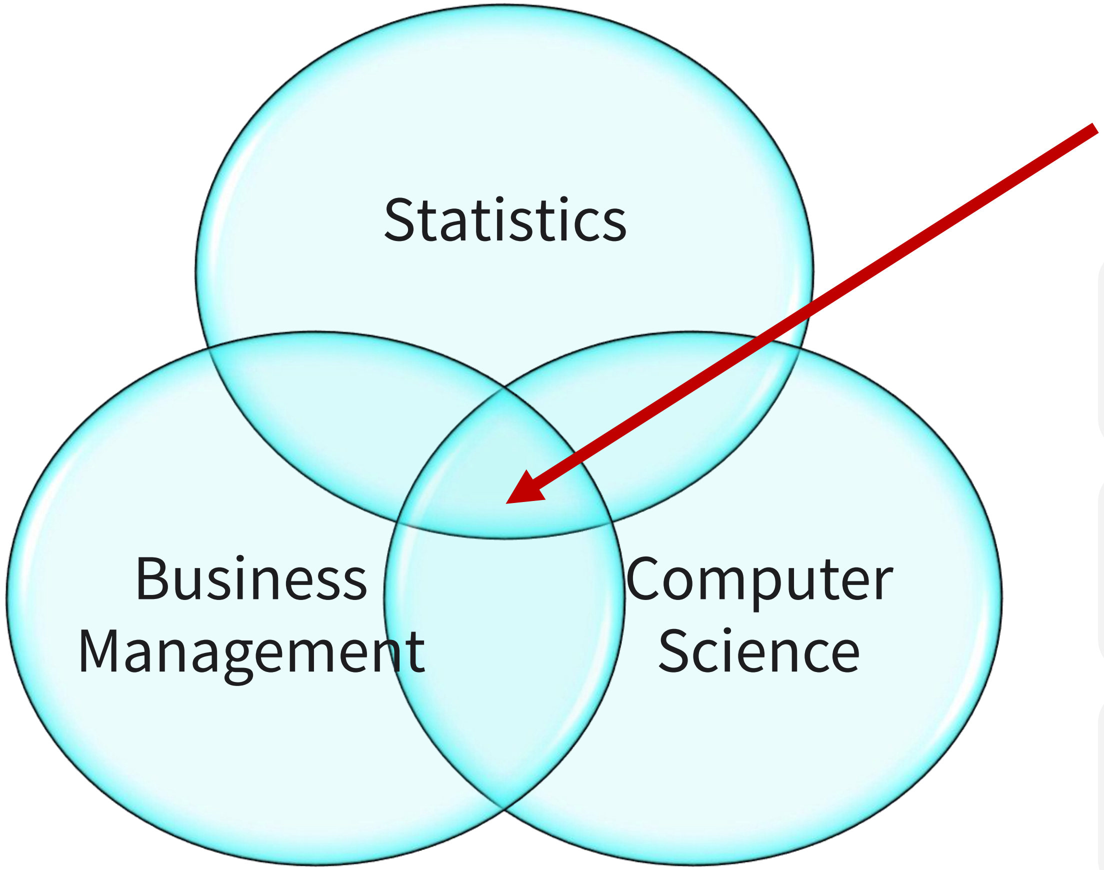
STUDY GOALS

- What is meant by data science?
- Why we need data science?
- Understand the main terms and definitions relating to data science.
- Learn about the 5 Vs of big data.
- Understand the issues concerning data quality.
- Understand what a data science use case is.
- Learn about the machine learning canvas.
- Identify the importance of statistics in data science.
- Know about probability and its relation to the prediction model's outputs.
- Understand the concept of machine learning and how it can be applied.

EXPLAIN SIMPLY

- How is the term “data science” defined?
- How is the term “data” defined?
- What is a Machine Learning Canvas?
- What is the importance of statistical parameters?
- What is machine learning?
- What are the most useful applications of machine learning?

DATA SCIENCE'S RELATED FIELDS



DATA SCIENCE

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

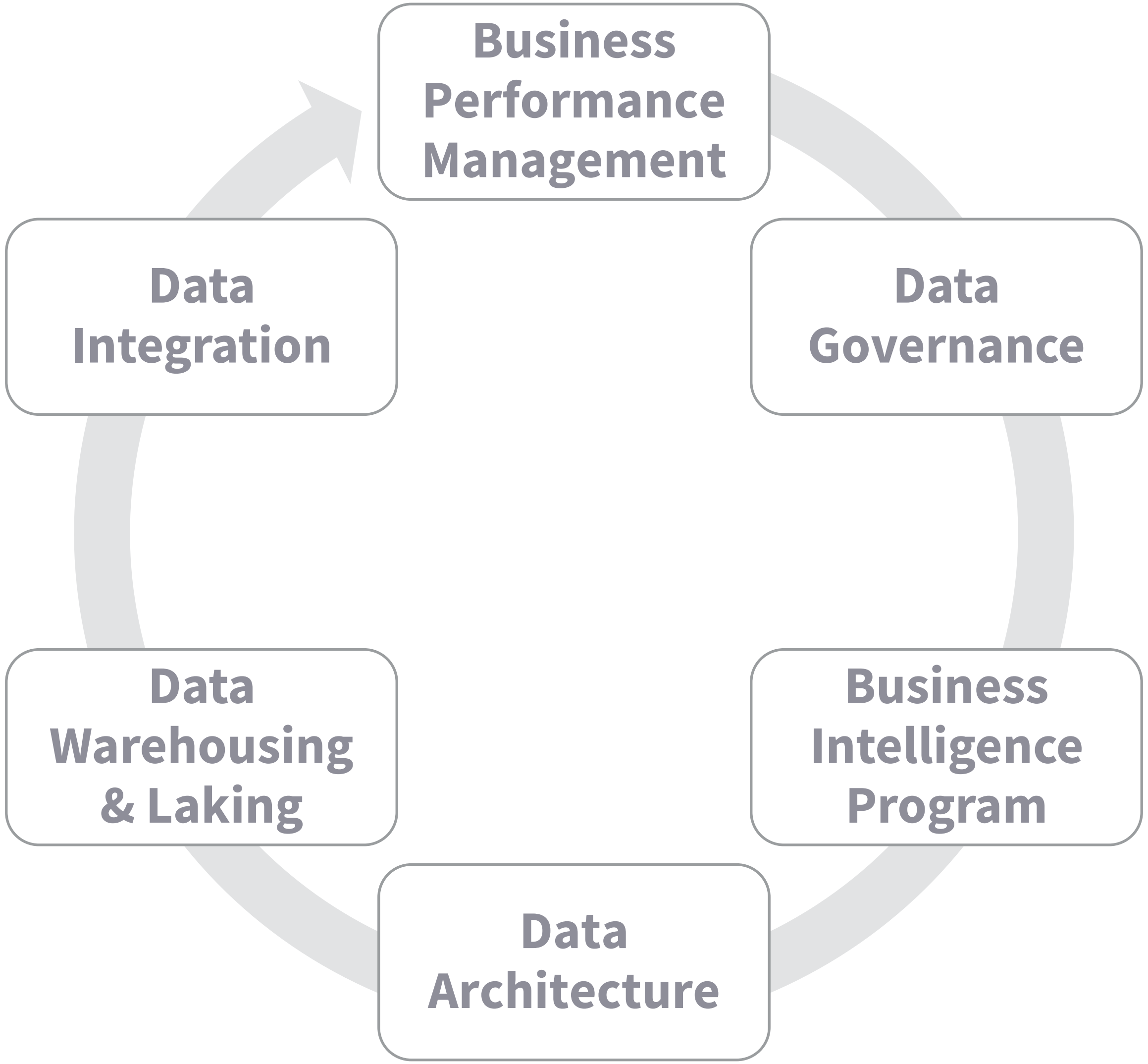
Unlocking the **real values** and **insights** of the data

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.

BUSINESS INTELLIGENCE



Source of the image: Microsoft Archive.

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.

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.

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

DEFINITION

DATA

Facts, observations, assumptions, or incidences

Data describe quantity, quality, statistics, symbols, or other units of meaning

Data are processed to return information

Source of the text: Zöller, 2022

TYPES OF DATA

Qualitative Data	Quantitative Data
Describe qualities or characteristics	Can be quantified or expressed as a number
Cannot be counted	Can be counted
Words, objects, pictures, observations, and symbols	Numbers and statistics
Answer to questions: What characteristics? What property?	Answer to questions: How much? How often?
Identify conceptual framework in an area of study	Test hypotheses, analyses the connection between cause and effect.

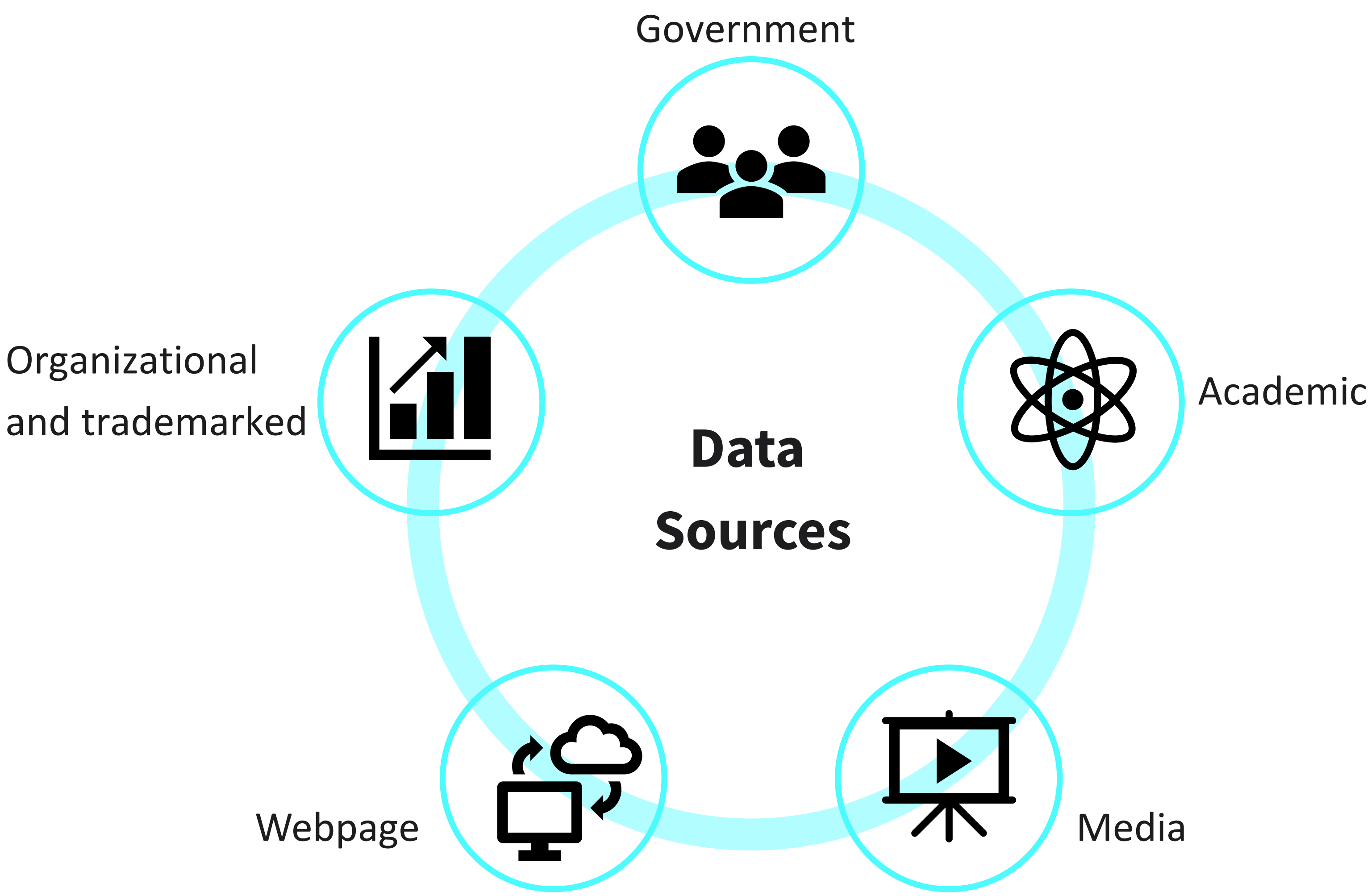
Source of the text: Zöller, 2022

SHAPES OF DATA

	Structured Data	Unstructured Data	Streaming Data
Characteristics	• predefined data models • usually, text or numerical • easy to search	• no predefined data models • can be text, images, audio, or other formats • difficult to search	• continuously generated • by sensors • large amount • processed incrementally
Applications	• inventory control • airline reservation systems	• word processing • tools for editing media	• state monitoring • process control
Examples	• phone numbers • customer names • transaction information	• reports • imagery • email/messages	• real-time sensor values • stock updates

Source of the text: Zöller, 2022

SOURCES OF DATA

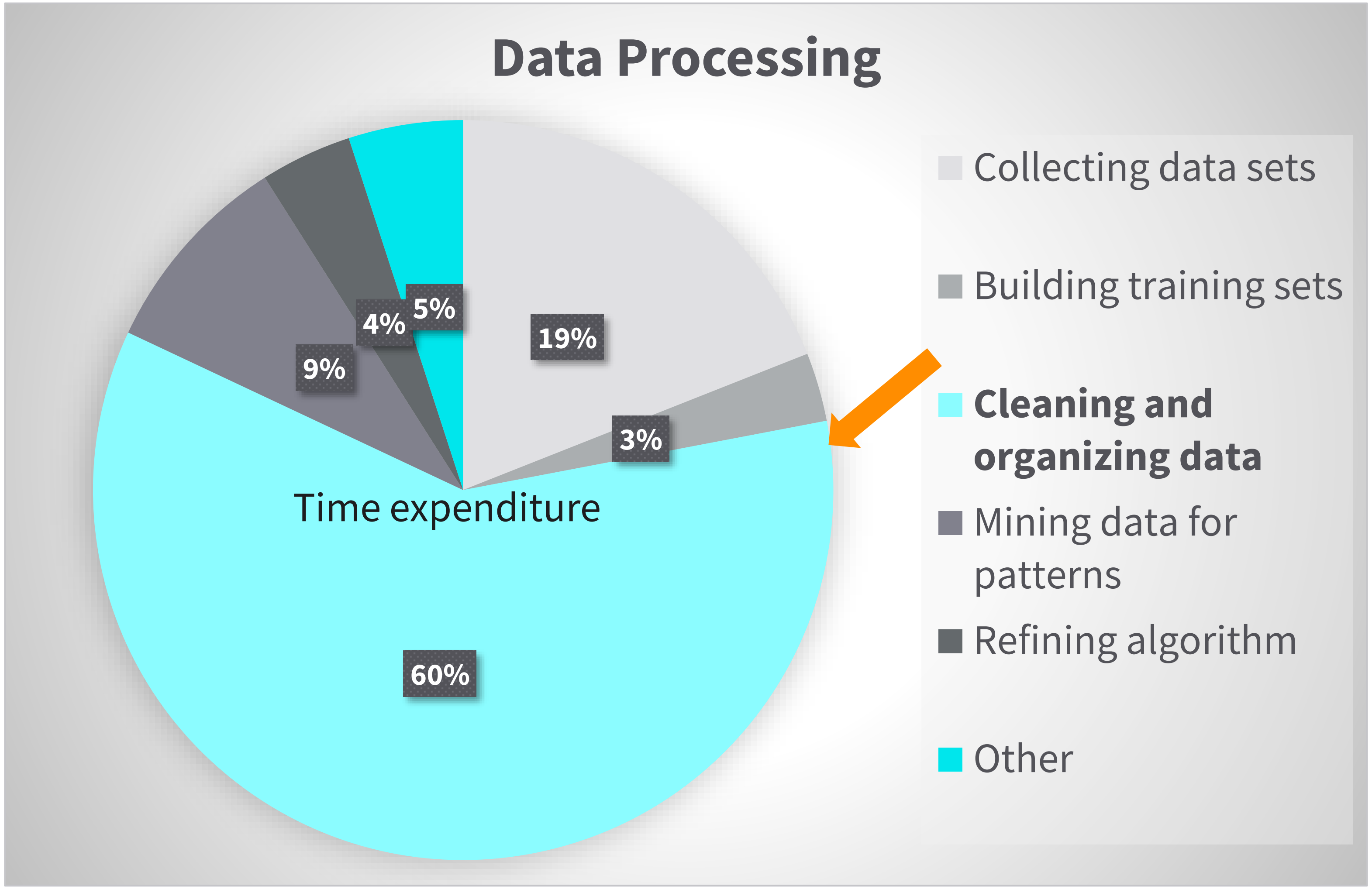


Source of the text: Zöller, 2022

THE 5VS OF BIG DATA

1. **Volume**: amount, scale
2. **Variety**: structured/unstructured
3. **Velocity**: frequency
4. **Veracity**: quality
5. **Validity**: suitability

Source of the text: Zöller, 2022



Source of the image: Zöller, 2022, p.24

Data Transformation Methods

Variable scaling

- Scaled values: {0,1} or {-1,1}
- To ensure equal weighting

Variable decomposition

- One variable → multiple variables
- To improve data representation

Variable aggregation

- Some variables → one variable
- To reduce data volume

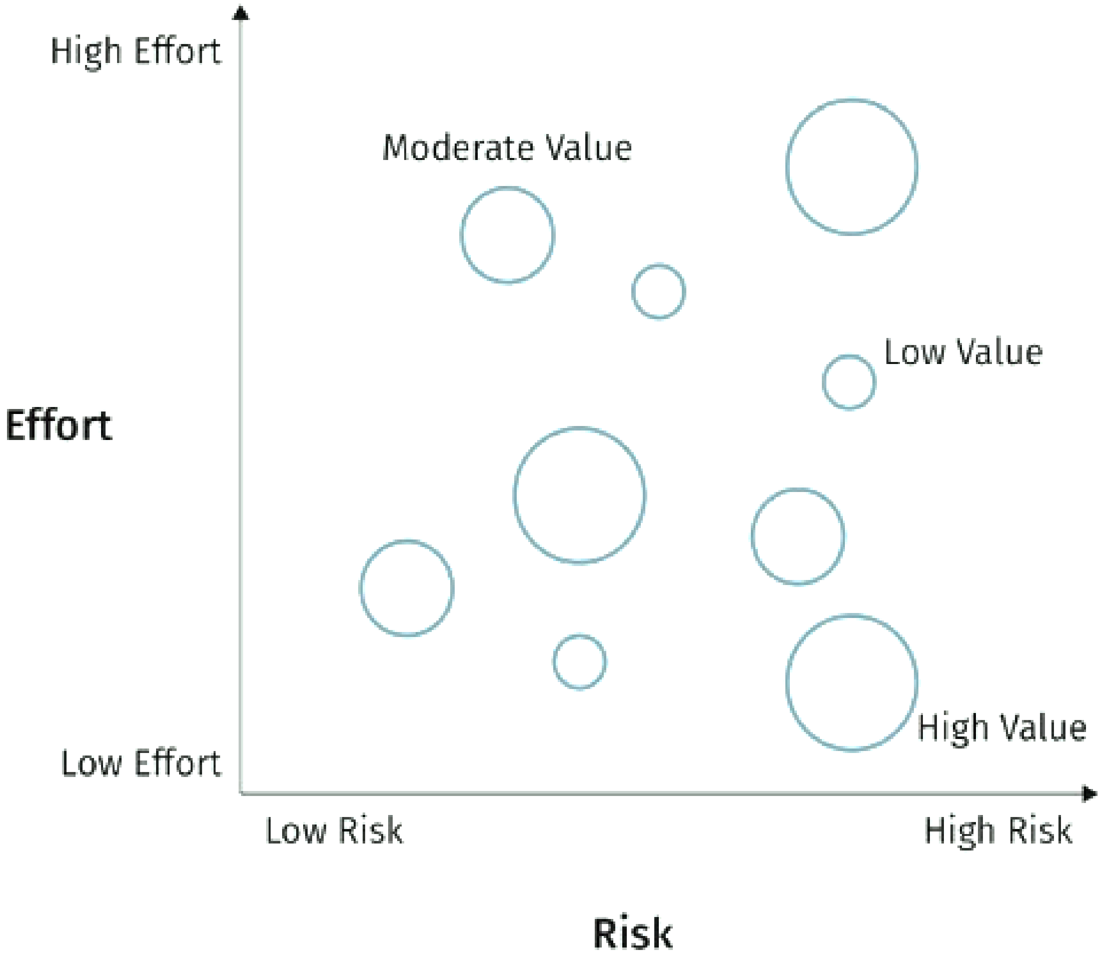
Source of the text: Zöller, 2022

Important questions have to be answered to identify the suitable data science use cases (**DSUC**) for the business objectives:

- What is the value of the **knowledge** gained by applying data science tools to this dataset?
- What will be **discovered** about the input dataset and its hypothesis?
- What **value** will be added to the organization through applying data science techniques?
- What will be the organization's **decision** based on the data science results?

Source of the text: Zöller , 2022.

DSUC Portfolio:



Source of the image: Zöller, 2022, p. 39

DSUC

- Define important questions for the business objectives
- Identify the suitable DSUC

Dataset

- Collect data
- Generate data if neccessary
- Label data
- Add comments
- Observe anomalies

Pre-processing techniques

- Correct errors/noises
- Scan redundant or missing data
- Select relevant features

Machine learning methods

- Establish mathematical functions
- Create training & testing set
- Train & test the model

Model Implementation

- Predict unseen data
- Update the model

Source of the text: Zöller , 2022.

MACHINE LEARNING CANVAS

<table>
<tr>
<td>5. Decision
How are predictions used to make decisions that provide the proposed value to the end-user(s)?</td>
<td>2. ML Task
Input, output to predict, type of problem</td>
<td rowspan="2">1. Value Propositions
What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?</td>
<td>3. Data Sources
Which raw data sources can we use (internal and external)?</td>
<td>4. Collecting Data
How do we get new data to learn from (inputs and outputs)?</td>
</tr>
<tr>
<td rowspan="2">6. Making Prediction
When do we make predictions on new inputs?
How long do we have to featurize a new input and make a prediction?</td>
<td>9. Offline Evaluation
Methods and metrics to evaluate the system before deployment</td>
<td>8. Features
Input representations extracted from raw data sources</td>
<td rowspan="2">7. Building Models
When do we create/update models with new training data?
How long do we have to featurize training inputs and create a model?</td>
</tr>
<tr>
<td colspan="3">10. Live Evaluation and Monitoring
Methods and metrics to evaluate the system after deployment and to quantify value creation</td>
</tr>
</table>

Source of the text: Dorard, 2017.

Evaluation metrics for a classification model

- Potential outcomes of classification: True positive, true negative, false positive, and false negative.
- Evaluation metrics to measure the quality: Precision, accuracy, recall, and F-Score.

Confusion matrix		**Prediction**	
		Positive	Negative
Ground truth	Positive	True positive	False negative
	Negative	False positive	True negative

EVALUATION METRICS

PRECISION

$$\frac{TP}{TP + FP}$$

RECALL

$$\frac{TP}{TP + FN}$$

ACCURACY

$$\frac{TP + TN}{TP + TN + FP + FN}$$

F-SCORE

$$2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Source of the text: Zöller , 2022.

Evaluation metrics for a regression model

Absolute error	$\varepsilon = \lvert d - y \rvert$
Relative error	$\varepsilon^* = \left\lvert \frac{d - y}{d} \right\rvert \cdot 100\%$
Mean absolute percentage error	$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left\lvert \frac{d_i - y_i}{d_i} \right\rvert \cdot 100\%$
Square error	$\varepsilon^2 = (d - y)^2$
Mean square error	$MSE = \frac{1}{n} \sum_{i=1}^{n} (d_i - y_i)^2$
Mean absolute error	$MAE = \frac{1}{n} \sum_{i=1}^{n} \lvert d_i - y_i \rvert$
Root mean square error	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (d_i - y_i)^2}$

Source of the text: Zöller , 2022.

There are several factors that impact judgments and decisions, and **biases** are an essential influence that might lead to inaccuracy.
The following table represents common **cognitive biases** and their proposed **de-biasing** techniques

Cognitive Bias	Description	De-Biasing Technique
Anchoring	Occurs when the estimation of a numerical value is based on an **initial value** (anchor), which is then **insufficiently adjusted**.	**Remove** anchors, have **numerous** and counter anchors, use **various experts** using specific anchors.
Confirmation	Occurs when there is a desire to **confirm** one's **belief**, leading to **unconscious selectivity** in the acquisition and use of evidence.	Use **multiple experts** for assumptions, counterfactual challenging probability assessments, use **sample evidence** for alternative assumptions.
Desirability	Favoring alternative options due to a bias that leads to **underestimating** or **overestimating** consequences.	Use **multi-stakeholder studies** of different perspectives, use **multiple experts** with different views, use appropriate **transparency** rates.
Insensitivity	Sample sizes are **ignored,** and extremes are considered **equally** in small and large samples.	Use **statistics** to determine the likelihood of extreme results in different samples, use the **sample data** to prove the logical reason behind extreme statistics.

Source of the text: Zöller, 2022.

Important statistical parameters

- **Maximum** – Greatest value
- **Minimum** – Smallest value
- **Mean** - Arithmetic average of values
- **Median** - Located in the middle
- **Standard deviation** - Shows the distribution of values

Source of the text: Zöller, 2022.

- **Probability** – The likelihood that an event will happen
  - $0 \leq P \leq 1$
  - $P = 0$: It is impossible for the event to occur.
  - $P = 1$: The event will definitely occur.

- **Probability theory** – Core theory for many Data Science techniques

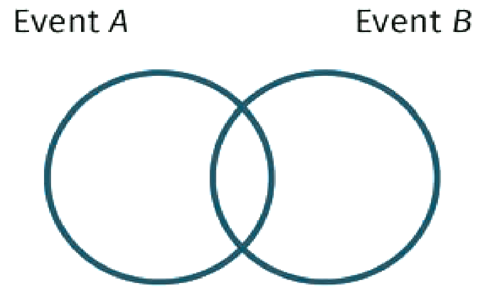
Source of the text: Zöller, 2020.

Mutually exclusive events:

events cannot occur at the same time

Multi independent events:

events can occur simultaneously without affecting each other



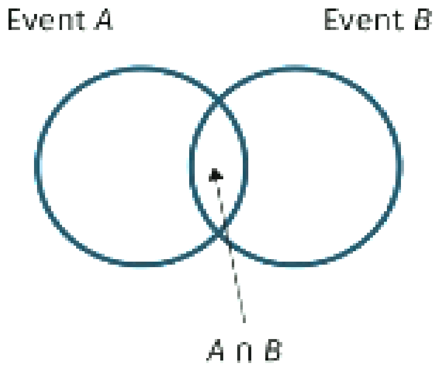
$P(A \text{ and } B) = P(A \cap B) = P(A) \cdot P(B)$;

$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Multi conditional probability:

events are correlated

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Source of the text: Zöller, 2020.

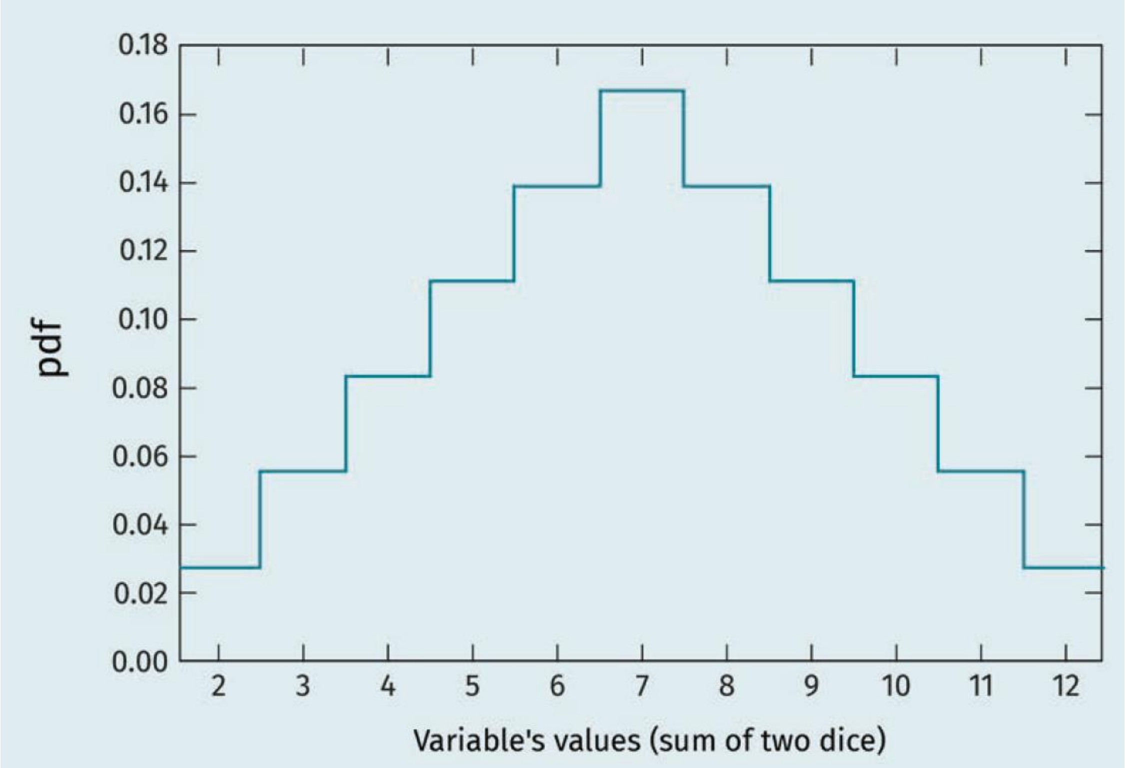
Probability distribution:

- A random variable can take on a given set of values.
- The **occurrence** of each of these values has a certain **probability**.

Probability distribution function

maps outcomes with their respective probability

- **X-axis**: possible values of the variable
- **Y-axis**: probability of each value

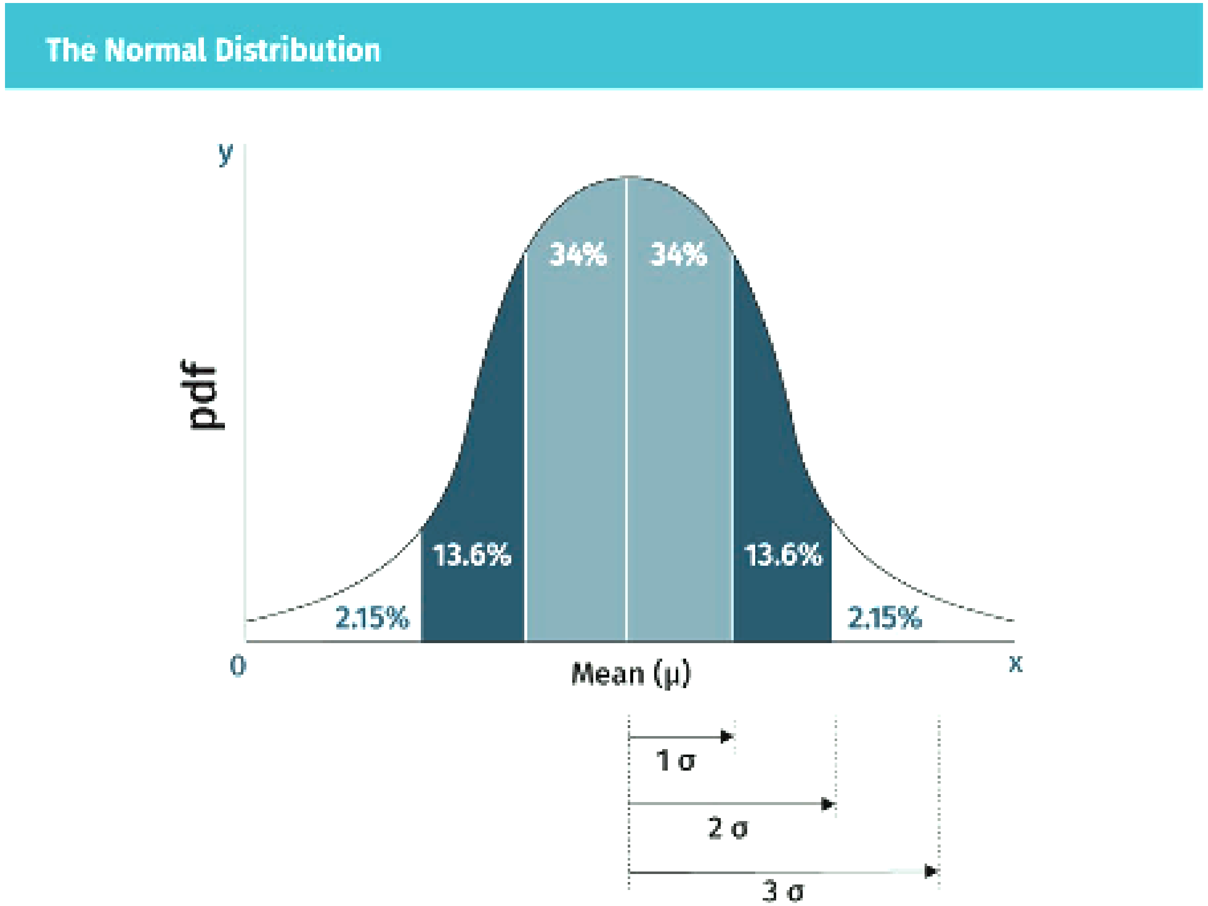


Source of the image: Zöller, 2020.

Normal distribution

- has a bell-shaped curve
- has a symmetrical distribution around the mean value
  - $1\sigma \sim 68\%$
  - $2\sigma \sim 95\%$
  - $1\sigma \sim 99{,}7\%$

Example: Performance assessment of an organization's employees



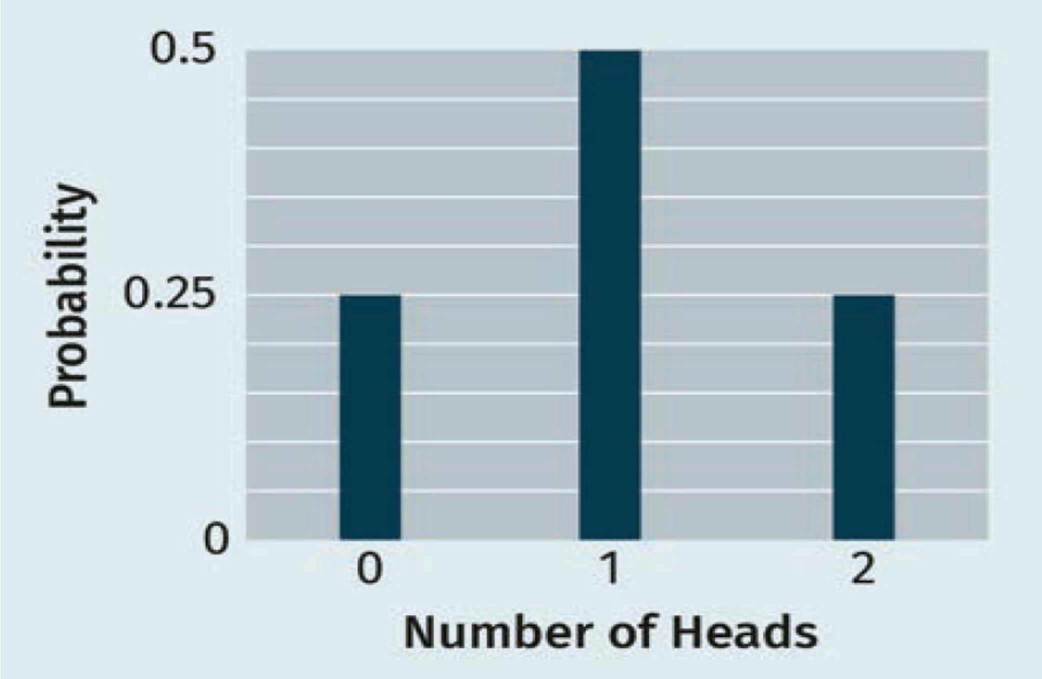
Source of the image: Zöller, 2020.

Binomial distribution

The probability distribution of the **number of successes** in a sequence of independent trials that each can be described by a **binary random** variable.

Example: tossing a coin twice

Possible Outcomes of Tossing a Coin		
Outcome	1st toss	2nd toss
1	Heads	Heads
2	Heads	Tails
3	Tails	Heads
4	Tails	Tails



Source of the image: Zöller, 2020.

Poisson distribution

The probability of a given number of independent events occurring in a fixed time interval

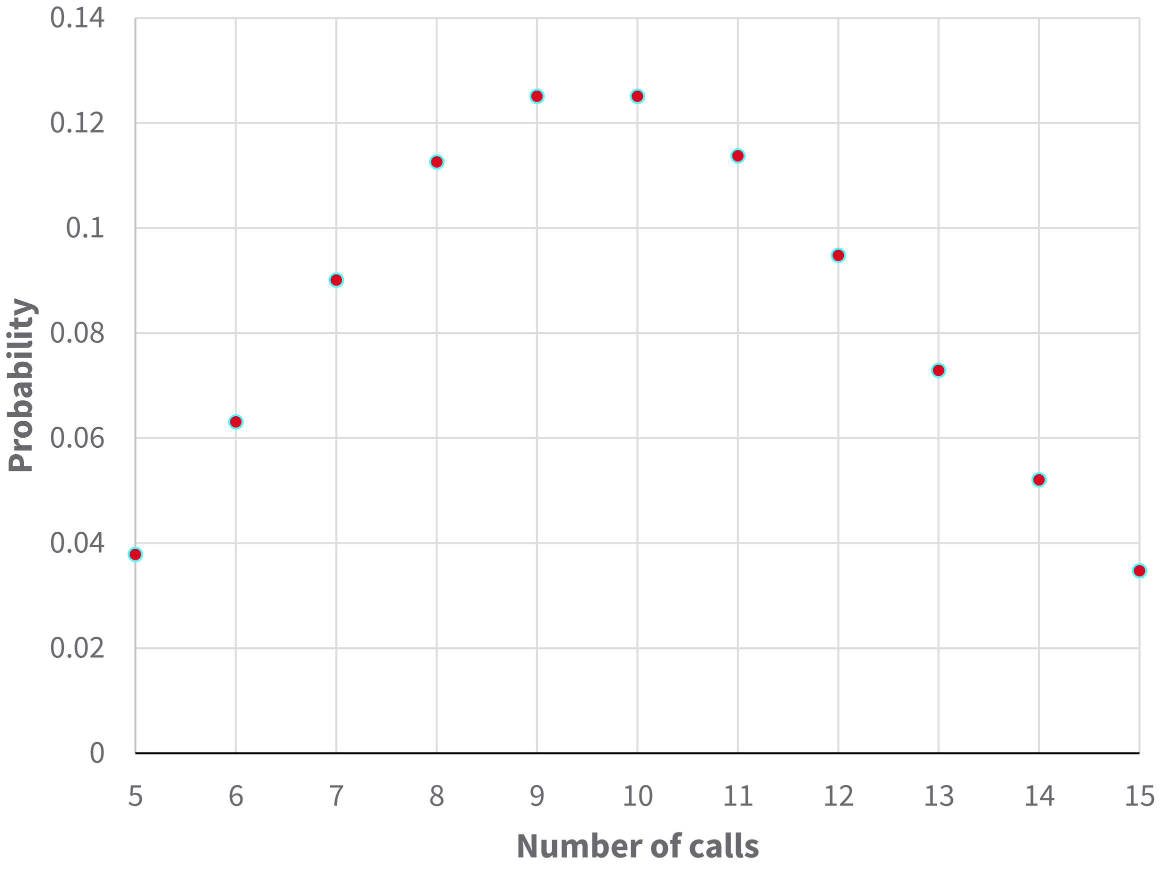
$$P(x) = \frac{e^{-\mu}\mu^x}{x!}$$

Where:

μ – the mean number of occurrences

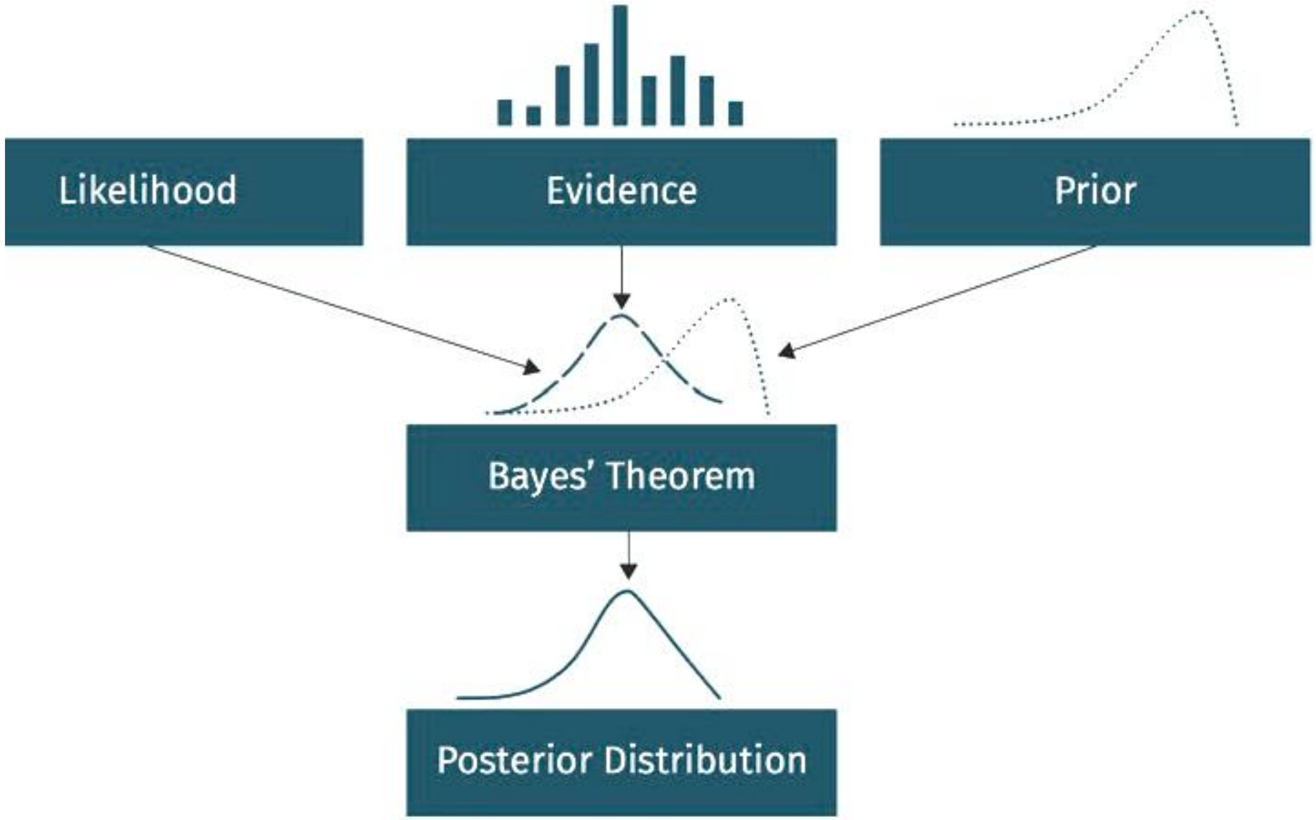
x – the required number of occurrences

Example: The probability that a call center will receive exactly n calls on a given day.



Source of the image: Zöller, 2020.

Bayerian statistics interprets probabilities **as expectation of belief.**



Conditional probability **equation**:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

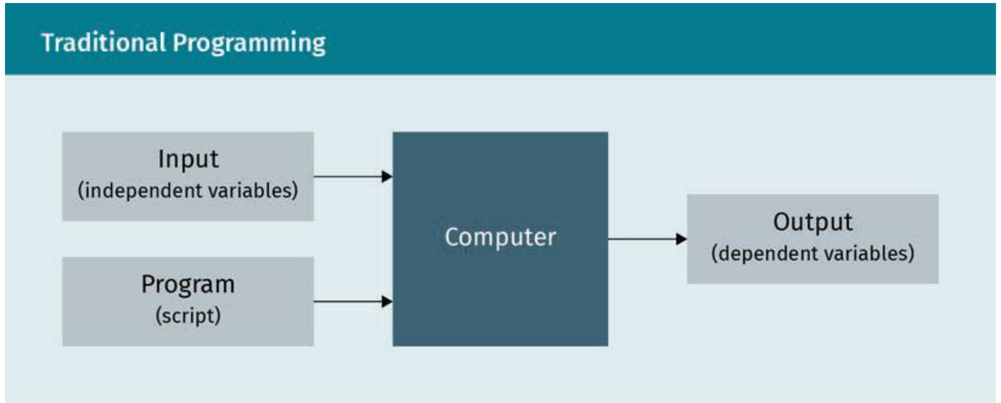
$P(A|B)$ is the **posterior** belief of the event A after observing the **evidence** B.

Example: Drug test analysis

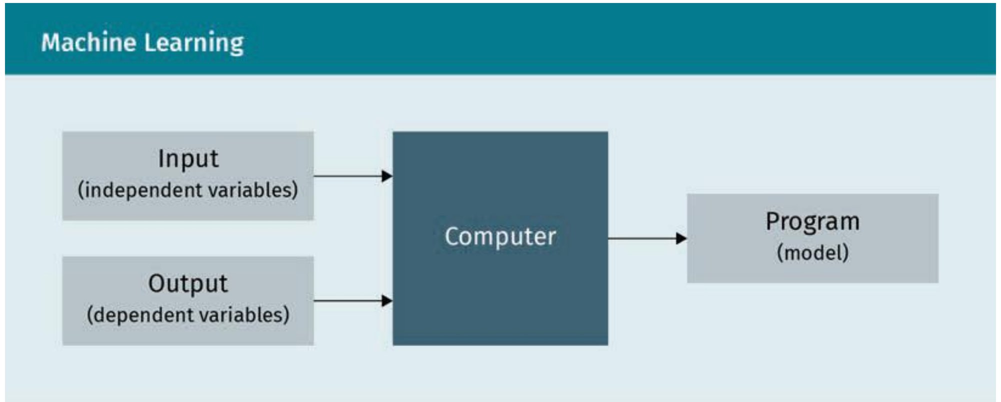
Source of the image: Zöller, 2020.

Machine learning concepts

- **Traditional programming** constructs an **explicit** processing of input variables into desired outputs via a set of **code** instructions.
- **ML** algorithms build **models** based on sample **data**, in order to make **predictions** or **decisions** without being explicitly programmed to do so.



Traditional programming

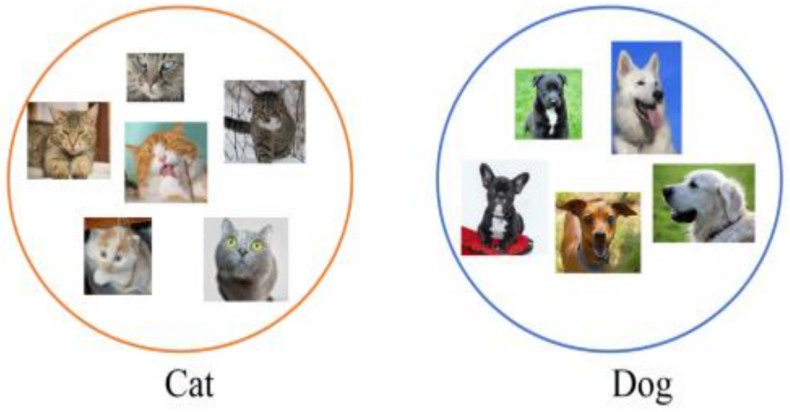


Machine learning

Source of the text: Zöller, 2022.
Source of the image: Zöller, 2022, pp.12-13.

Classification

- **Objective**: Develop a ML model to **map** the inputs to the outputs and **predict** the **classes** of new inputs.
- **Accuracy** can be presented in a confusion matrix.
- Evaluation **metrics**:
  - $Precision = \frac{TP}{TP+FP}$
  - $Recall = \frac{TP}{TP+FN}$
  - $F_{Score} = \frac{2\cdot(Precision\cdot Recall)}{Precision+Recall}$



Dog and Cat classification

The Confusion Matrix

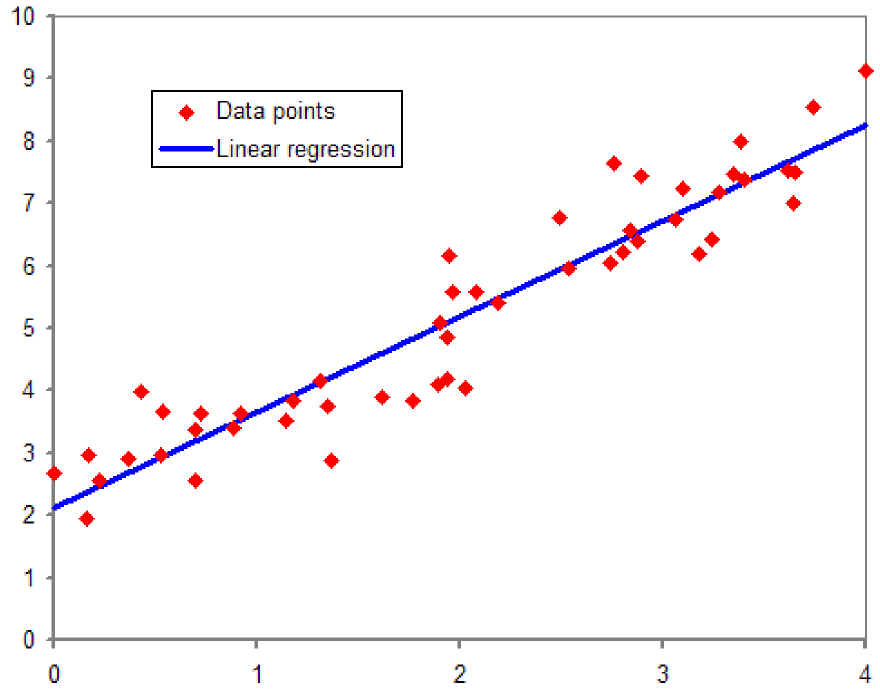
		Model's output	
		Class 1	Class 2
Desired output	Class 1	TP	FN
	Class 2	FP	TN

Confusion matrix

Source of the text: Zöller, 2022.
Source of the image: Sasaki, n.d., Open source.

Regression

- **Objective**: Develop a ML model to **relate** the inputs x to the outputs y and **predict** the output **values** $\hat{y}$ for new inputs
- Evaluation **metrics**:
  - Mean Square Error: $\boldsymbol{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - Root Mean Square Error: $\boldsymbol{RMSE} = \sqrt{MSE}$
  - Mean Absolute Error: $\boldsymbol{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$



An example of Regression

Source of the text: Fernandez, 2020; Jason, 2021; Zöller, 2022.
Source of image: Amatulic, 2007, Public Domain.

Supervised Learning

- **Dataset**: a collection of **labeled** samples, containing both inputs (independent variables) and outputs (dependent variable)
- **Objective**: develop a ML model to **relate** the inputs to the outputs of in the training set and **predict** the outputs for new inputs

Supervised Learning Structure

Source of the text: Zöller, 2022.

Unsupervised Learning

- **Dataset**: a collection of **unlabeled** samples, containing only inputs (independent variables) while outputs (dependent variable) are unknown.
- **Objective**: develop a ML model to **discover** the salient **patterns** and **structures** within the training set.

Unsupervised Learning Structure

Source of the text. Zöller, 2022.

Semi-Supervised Learning

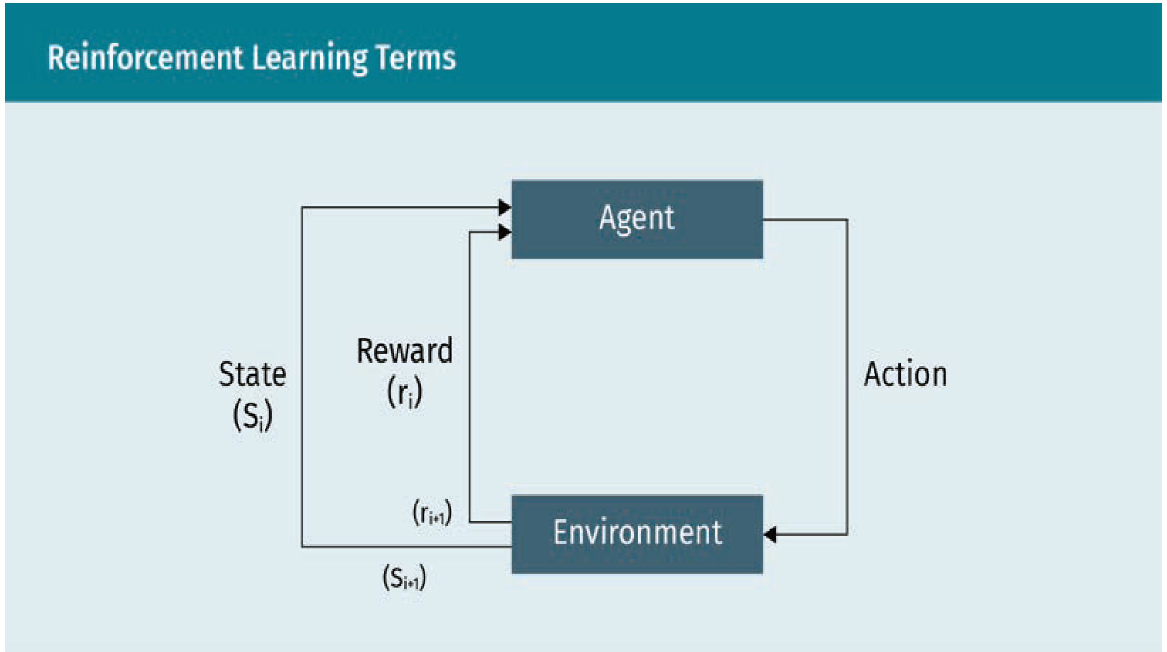
- **Dataset**: a collection of both **labeled** samples (a small portion of data), and **unlabeled** samples (lots of data)
- **Objective**: mix of supervised and unsupervised learning to combine the properties of both.
- **2 steps:**
  - Supervised learning is performed on few labeled data
  - Unsupervised learning is performed on large unlabeled data

Semi-Supervised Learning Structure

Source of the text: Zöller, 2022.

Reinforcement Learning

- **Objective**: To find an **action policy** that achieves a given goal by **trial-and-error** interactions with the environment.
- **“Cause and effect”** method: An action is performed to achieve a maximum reward.
- **Reward function** acts as feedback to the agent.



Reinforcement Learning Structure

Source of the text: Zöller, 2022.
Source of image: Zöller, 2022, p. 24.

EXAM: ORAL ASSIGMENT - ONLINE PRESENTATION

Choice: There are different topic options to choose from for the oral assignment. Please select **only one** to cover in your presentation.

Goal: to determine your ability to present a research topic in the field of Data Science in an understandable way.

Basis: Coursebook.

Further materials: Self-identified literature.

EVALUATION OF THE ORAL ASSIGNMENT

The following criteria are included in the evaluation with the respective percentage indicated:

Criteria	Explanation	Percentage
Introductory remarks	Lead into the topic	5 %
Text	– Structural outline of the presentation – logic of the outline – appropriate emphasis – timed sections	20 %
Argument	Quality of reasoning and research	30 %
Conclusion	– Conclusion – Summary – concise summary of the results	15 %
Rhetoric	– General quality of delivery performance – Comprehensibility – Intonation – use of pause – appropriate use of tone – use of non-verbal effects	15 %
Visual Layout	– General quality of slide presentation – clarity and number of slides – adequate font size – use of graphic effects	15 %

TUTORIAL SUPPORT

Several **options** are available for support with presentations. The **student** is **responsible** for making use of these resources.

Tutors are available for subject **consultation** on the **choice** of topic as well as for specific and **general questions** on academic work.

There is **no provision** for the tutor to confirm acceptable **outlines**, parts of the **content**, or presentation **drafts** since independent preparation is part of the examination.

Hints may be given to facilitate the creation of academic work.

STUDY GOALS

- What is meant by data science?
- Why we need data science?
- Understand the main terms and definitions relating to data science.
- Learn about the 5 Vs of big data.
- Understand the issues concerning data quality.
- Understand what a data science use case is.
- Learn about the machine learning canvas.
- Identify the importance of statistics in data science.
- Know about probability and its relation to the prediction model’s outputs.
- Understand the concept of machine learning and how it can be applied.

SESSION 6

TRANSFER TASK

TRANSFER TASK

Task 1: Choose **one** of the following topics to present: **CRISP-DM** or Microsoft **TDSP**.

Task 2: Present the topic **of Machine Learning Canvas**.

Task 3: Choose **one** of the following topics to present**: Linear Regression**, **Decision Tree Learing**, or **K-Means Clustering.**

Please present your results.
The results will be discussed in plenary.

TRANSFER TASK

GUIDELINES

Task 1: Choose CRISP-DM or Microsoft TDSP to present

1. **Argue** if and **why** process models are useful in the context of data science activities.
2. **Introduce** and **describe** either one of the aforementioned process models. You are free to choose which model you want to **focus** on.
3. **Summarize** how the model you selected for description in 2, addresses the **goals** and **benefits** of process models outlined in 1.

Recommended Literature:

Smart Vision Europe. (2022). *What is the CRISP-DM methodology?* https://www.sv-europe.com/crisp-dm-methodology/

Microsoft. (2022). *What is the Team Data Science Process?* https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/

Task 2: Machine Learning Canvas

1. Give an exposition of the design **goals** underlying the creation of the ML Canvas.
2. Describe and explain the **structure** of the ML Canvas and its constituent **parts**.
3. Summarize the most salient **benefits** of the ML Canvas for setting up and documenting Data Science projects.

Recommended Literature:

Dorard, L. (2019). *Machine learning canvas.* https://www.machinelearningcanvas.com/

Task 3: Machine Learning

Choose Linear Regression, Decision Tree Learing, or K-Means Clustering to present

1. Briefly **introduce** the different machine learning **paradigms** (supervised, unsupervised, semi-supervised) and describe to **which** paradigm your chosen algorithm belongs and **why**.
2. Conceptually describe **how** your method of choice works.
3. Give an **example** of a real-world analytical problem that could be addressed by your chosen method.

Recommended Literature:

Shalev-Shwartz, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press. https://www.cs.huji.ac.il/w~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

LIST OF SOURCES

Text:

Dorard, L. (2017). *The machine learning canvas*. https://www.louisdorard.com/machine-learning-canvas

Fernandez, J. (2020). *Introduction to regression analysis*. https://towardsdatascience.com/introduction-to-regression-analysis-9151d8ac14b3

Jason, B. (2021). *Regression metrics for machine learning*. https://machinelearningmastery.com/regression-metrics-for-machine-learning/

Microsoft. (2022). What is the Team Data Science Process? https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/

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.

Shalev-Shwartz, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press. https://www.cs.huji.ac.il/w~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

Smart Vision Europe. (2022). *What is the CRISP-DM methodology*? https://www.sv-europe.com/crisp-dm-methodology/

Zöller, T. (2020). *Course Book – Introduction to data science*. IU International University of Applied Sciences.

Images

Amatulic. (2007). File:Normdist_regression.png. *Wikimedia Commons.* https://en.wikipedia.org/wiki/Regression_analysis

Sasaki, T. (n.d.). Dog and Cat Classification.png [Open source]. *Kaggle.* https://www.kaggle.com/code/sasakitetsuya/dog-and-cat-classification-by-mobilenet