

LECTURER: TAI LE QUY

INTRODUCTION TO DATA SCIENCE

Introduction to Data Science

1

Data

2

Data Science in Business

3

Statistics

4

Machine Learning

5

Summary Session

6

UNIT 3

DATA SCIENCE IN BUSINESS



On completion of this unit, you will have learned ...

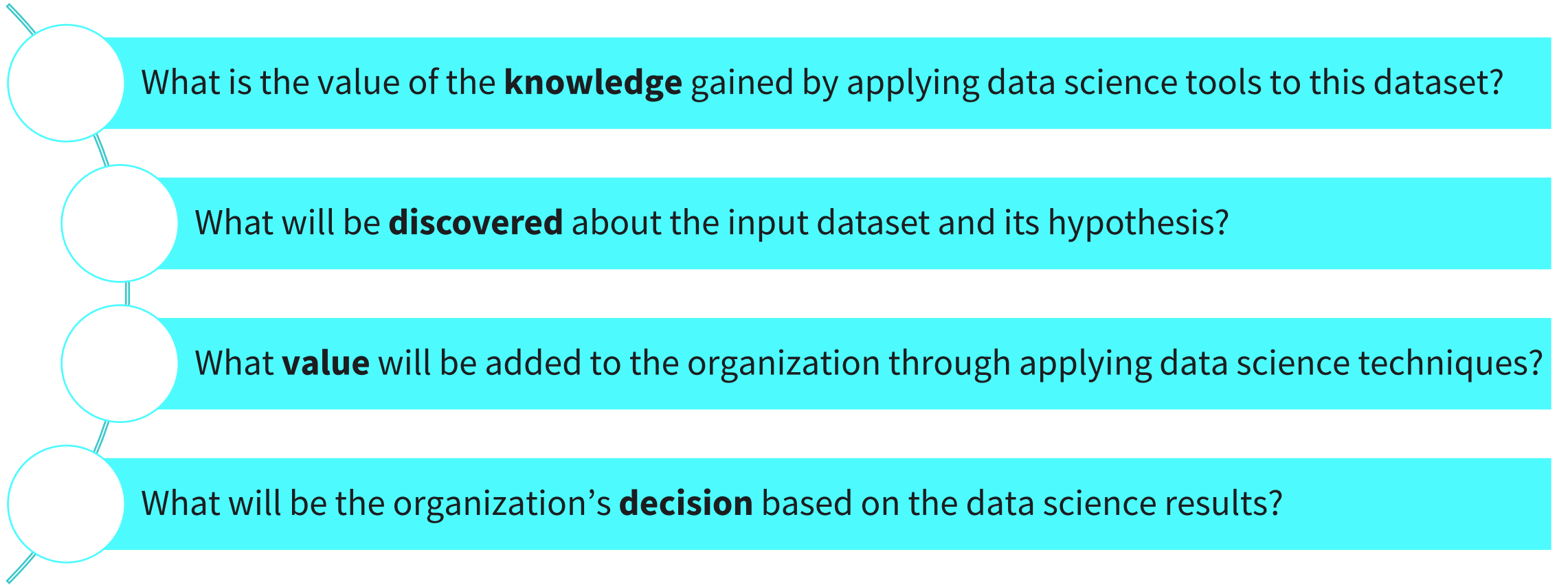
- ... what a data science use case is.
- ... about the machine learning canvas.
- ... about the model-centric performance evaluation.
- ... about the role played by KPIs in operational decisions.
- ... to identify the influence of the cognitive biases.



- What is a DSUC?
- What is a Machine Learning Canvas?
- How can the performance of models be evaluated?
- What are the characteristics of KPIs in operational decisions?
- What is cognitive bias? And how can we solve this problem?

IDENTIFICATION OF USE CASES

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?

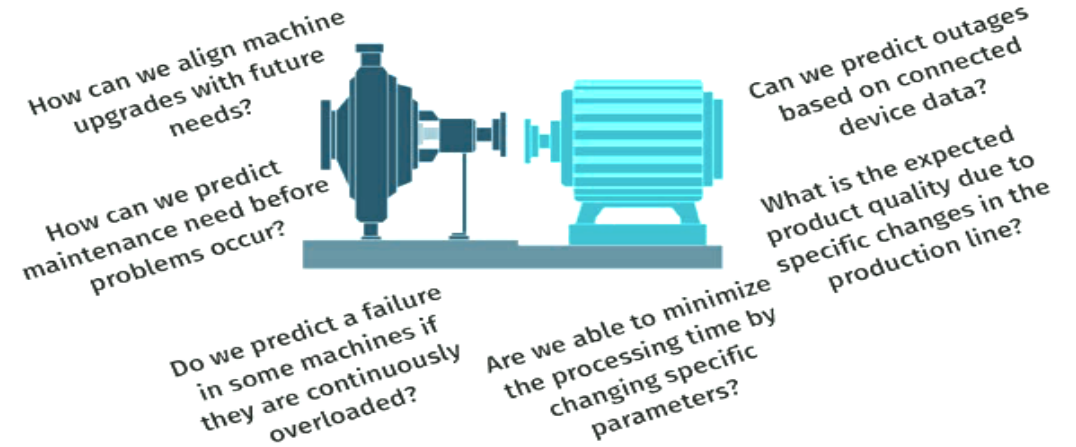
IDENTIFICATION OF USE CASES

EXAMPLES OF DSUC

Achieved Value by Data Science in "Customer"-Related Use Cases



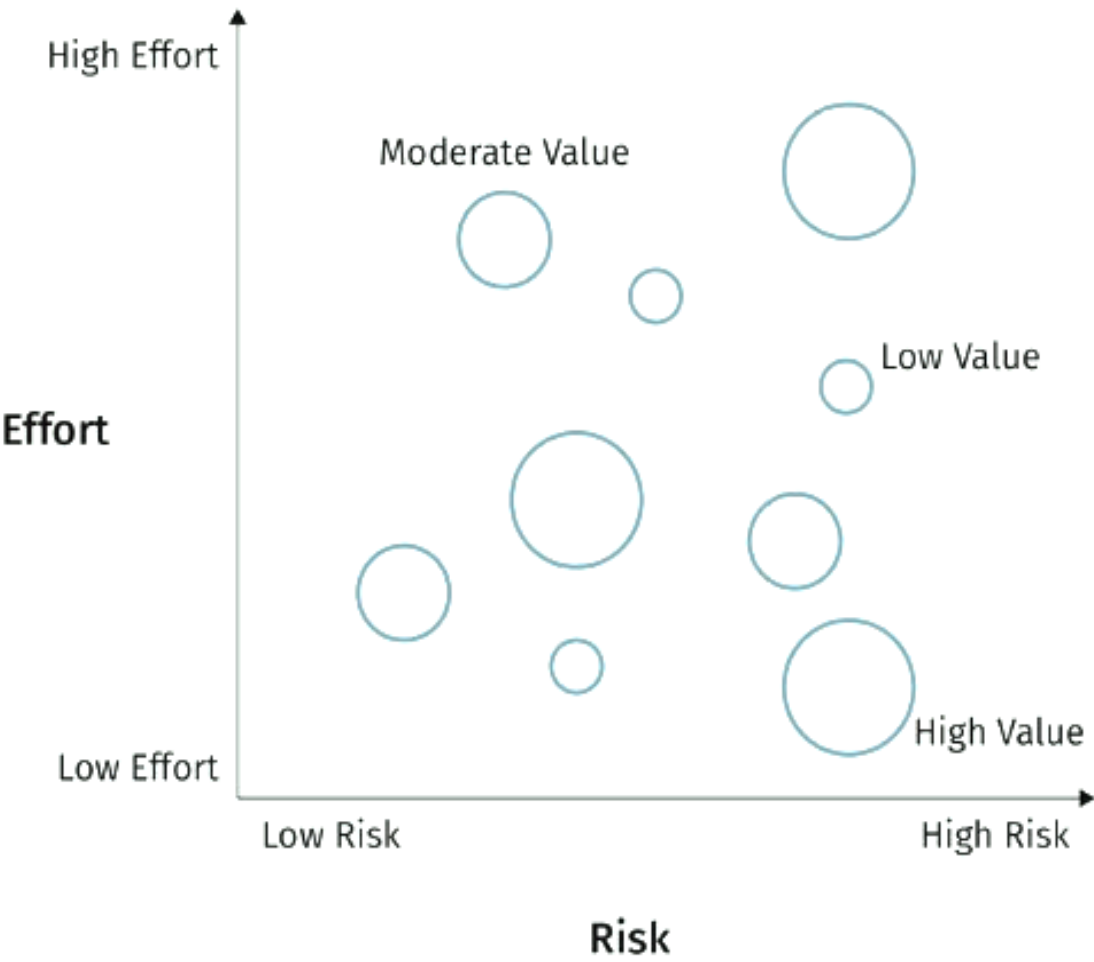
Achieved Value by Data Science in "Operational"-Related Use Cases



Achieved Value by Data Science in "Financial Fraud"-Related Use Cases



DSUC Portfolio:



DSUC

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

Dataset

- Collect data
- Generate data if necessary
- 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

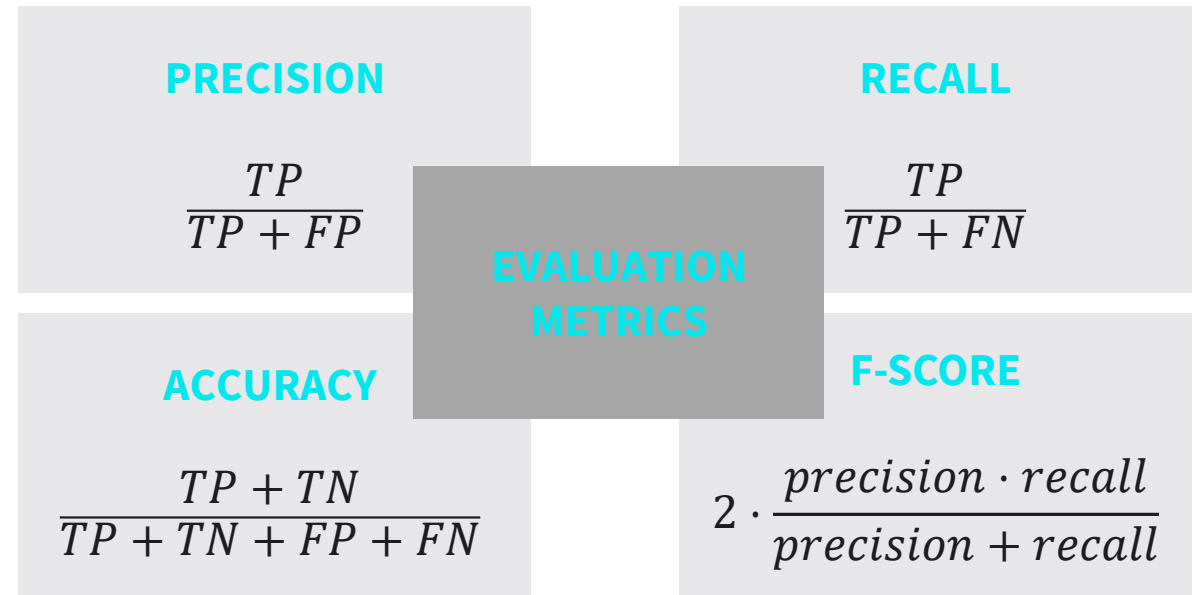
MACHINE LEARNING CANVAS

5. Decision How are predictions used to make decisions that provide the proposed value to the end-user(s)?	2. ML Task Input, output to predict, type of problem	1. Value Propositions What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?	3. Data Sources Which raw data sources can we use (internal and external)?	4. Collecting Data How do we get new data to learn from (inputs and outputs)?
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?	9. Offline Evaluation Methods and metrics to evaluate the system before deployment		8. Features Input representations extracted from raw data sources	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?
	10. Live Evaluation and Monitoring Methods and metrics to evaluate the system after deployment and to quantify value creation			

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 for a regression model

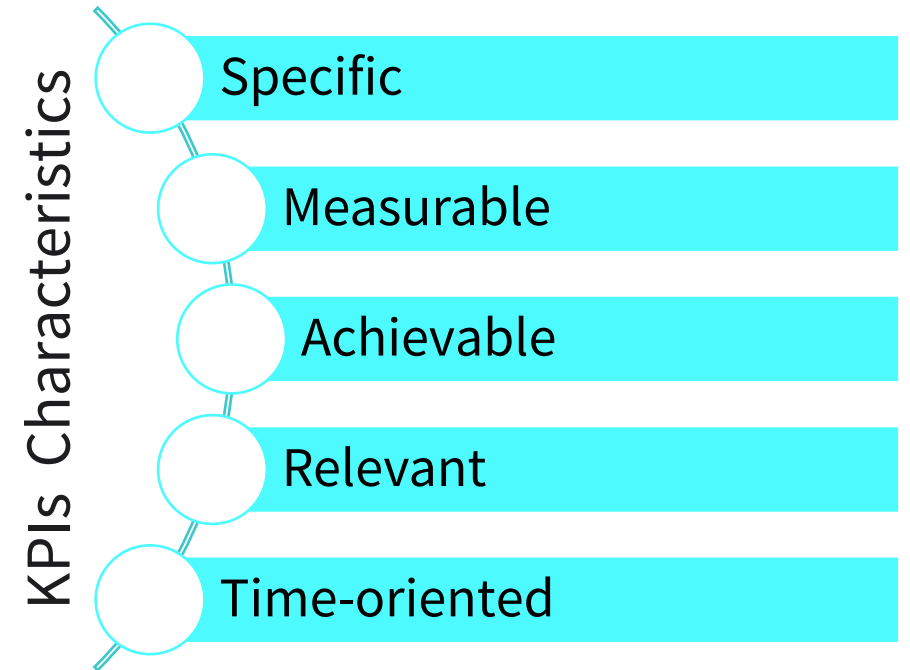
Absolute error	$\varepsilon = d - y $
Relative error	$\varepsilon^* = \left \frac{d - y}{d} \right \cdot 100\%$
Mean absolute percentage error	$MAPE = \frac{1}{n} \sum_{i=1}^n \left \frac{d_i - y_i}{d_i} \right \cdot 100\%$
Square error	$\varepsilon^2 = (d - y)^2$
Mean square error	$MSE = \frac{1}{n}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

KPIs – Key Performance Indicators

- measure the achievement of specific objectives,
- to enhance efficiency, decrease costs, increase profits, increase customer satisfaction, and support company's success.

Some **examples** of KPIs:

- number of complaints
- proportion of tasks executed
- measured time to complete a task
- annual growth



COGNITIVE BIASES

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.



On completion of this unit, you will have learned ...

- ... what a data science use case is.
- ... about the machine learning canvas.
- ... about the model-centric performance evaluation.
- ... about the role played by KPIs in operational decisions.
- ... to identify the influence of the cognitive biases.

SESSION 3

TRANSFER TASK

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your
results.

The results will be
discussed in plenary.



TRANSFER TASK

Create a ML Canvas in the domain of real estate to investigate risky investments and compare the real estate price predictions with the actual prices to determine the best deals.



1. Which of the following is incorrect?

Machine Learning Canvas is ...

- a) a single page user interface
- b) used to identify DSUC
- c) applied to firm's problems
- d) none of the above



2. Data science use cases (DSUC) are identified by:

- a) effort and risk
- b) effort, risk and achieved value
- c) effort and achieved value
- d) risk



3. Which of the following is not one of the common cognitive biases?

- a) anchoring
- b) confirmation
- c) desirability
- d) sensitivity

LIST OF SOURCES

Dorard, L. (2017). *The machine learning canvas*. <https://www.louisdorard.com/machine-learning-canvas>
Zöller, T. (2022). *Course Book – Introduction to Data Science*. IU International University of Applied Science.

© 2022 IU Internationale Hochschule GmbH

This content is protected by copyright. All rights reserved.

This content may not be reproduced and/or electronically edited, duplicated, or distributed in any kind of form without written permission by the IU Internationale Hochschule GmbH.