LECTURER: TAI LE QUY

DATA SCIENCE

TOPIC OUTLINE

UNIT 2

USE CASES AND PERFORMANCE EVALUATION

STUDY GOALS

On completion of this unit, you will have learned …

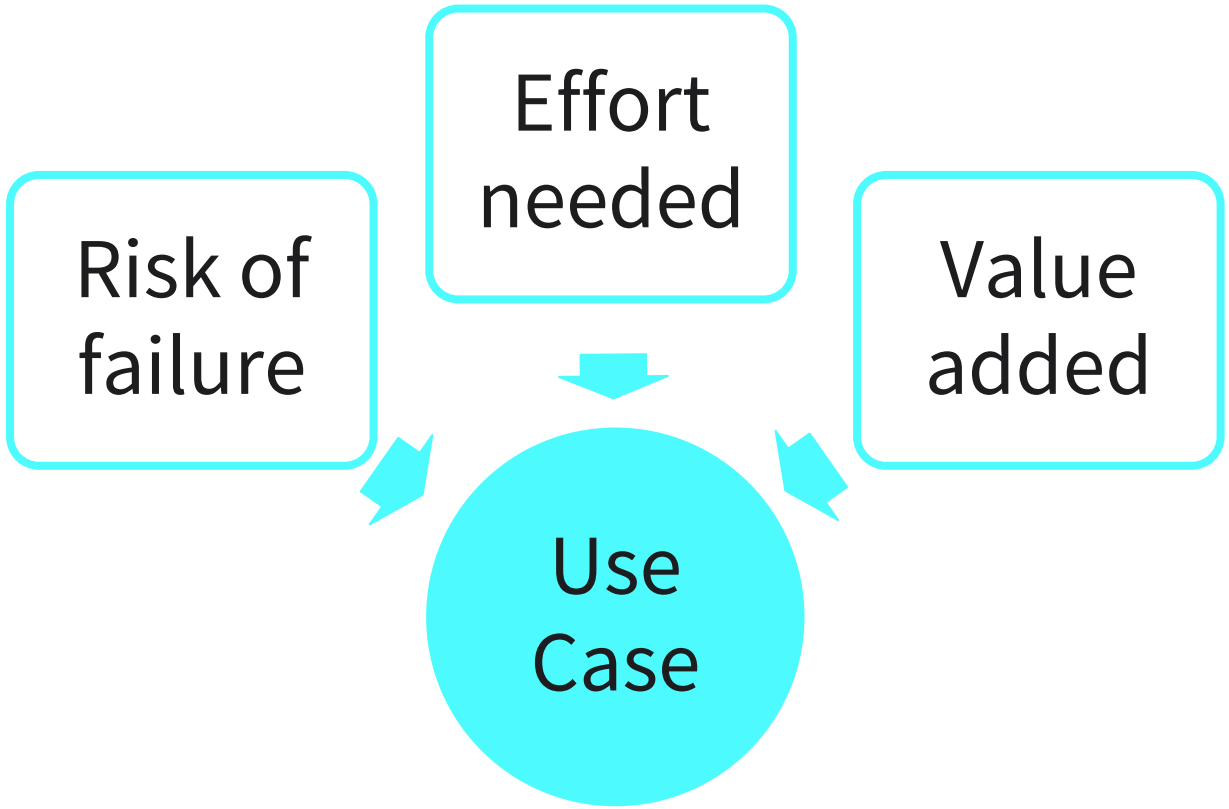
- the importance of a use case for business.
- how to identify use cases.
- the steps to develop a predictive model for a specific use case.
- the metrics to evaluate the performance of a predictive model.
- the role of KPIs in business-centric evaluation.
- the different cognitive biases which influence the decision-making process.

EXPLAIN SIMPLY

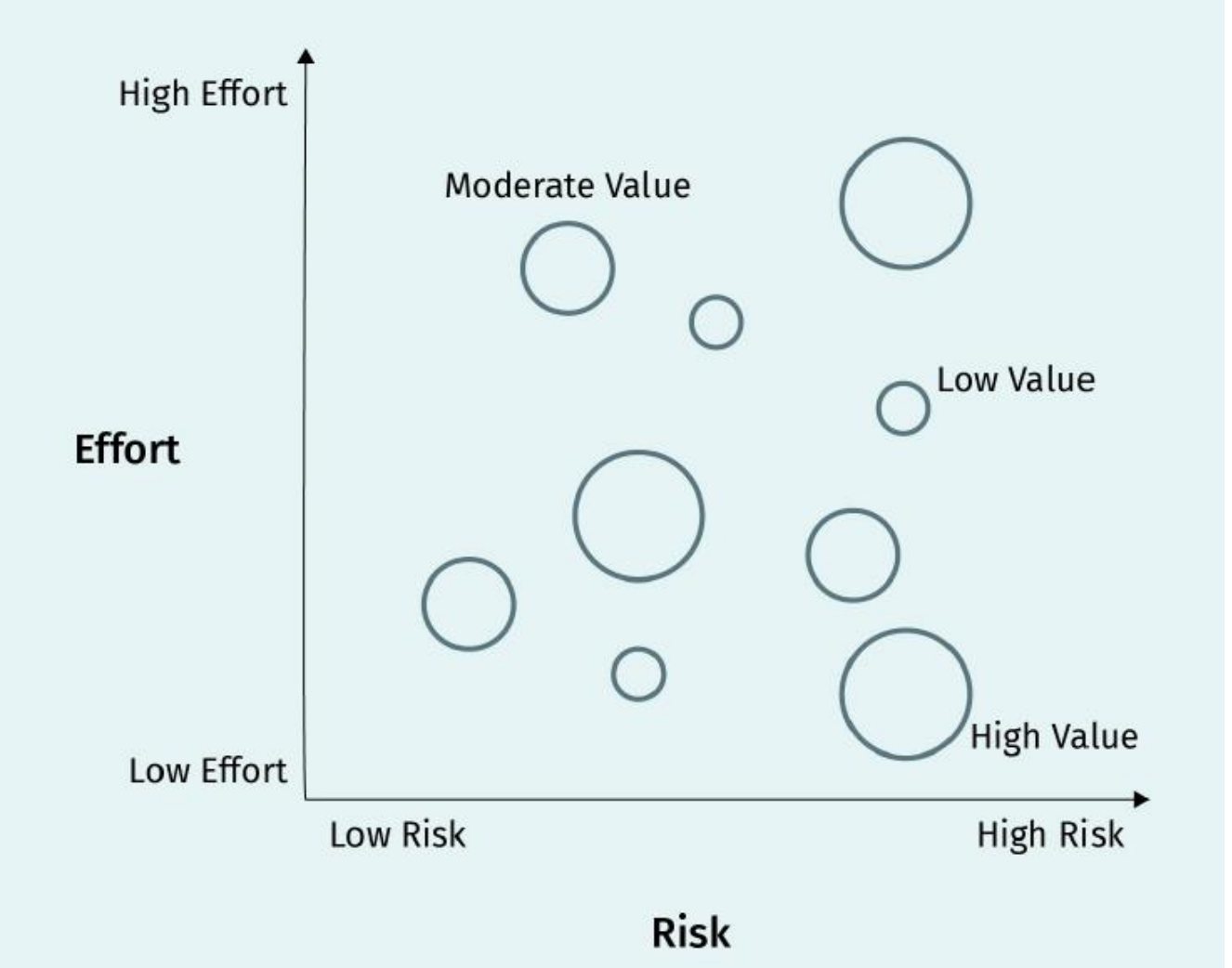
1. Identify a potential model evaluation metrics for a classification use case.
2. Explain why bias is a challenge in data science and mention one de-biasing technique.
3. Name three characteristics of effective business KPIs.

Focus on:

- increasing knowledge gain from data (e.g., better customer understanding)
- reducing business risk (e.g., predict machine outage upfront)
- decreasing effort (e.g., automate processes)



USE CASE IDENTIFICATION AND VALUE PROPOSITION



Source of figure: : IU International University, Course Book DLMBDSA01, Dorard, 2017.

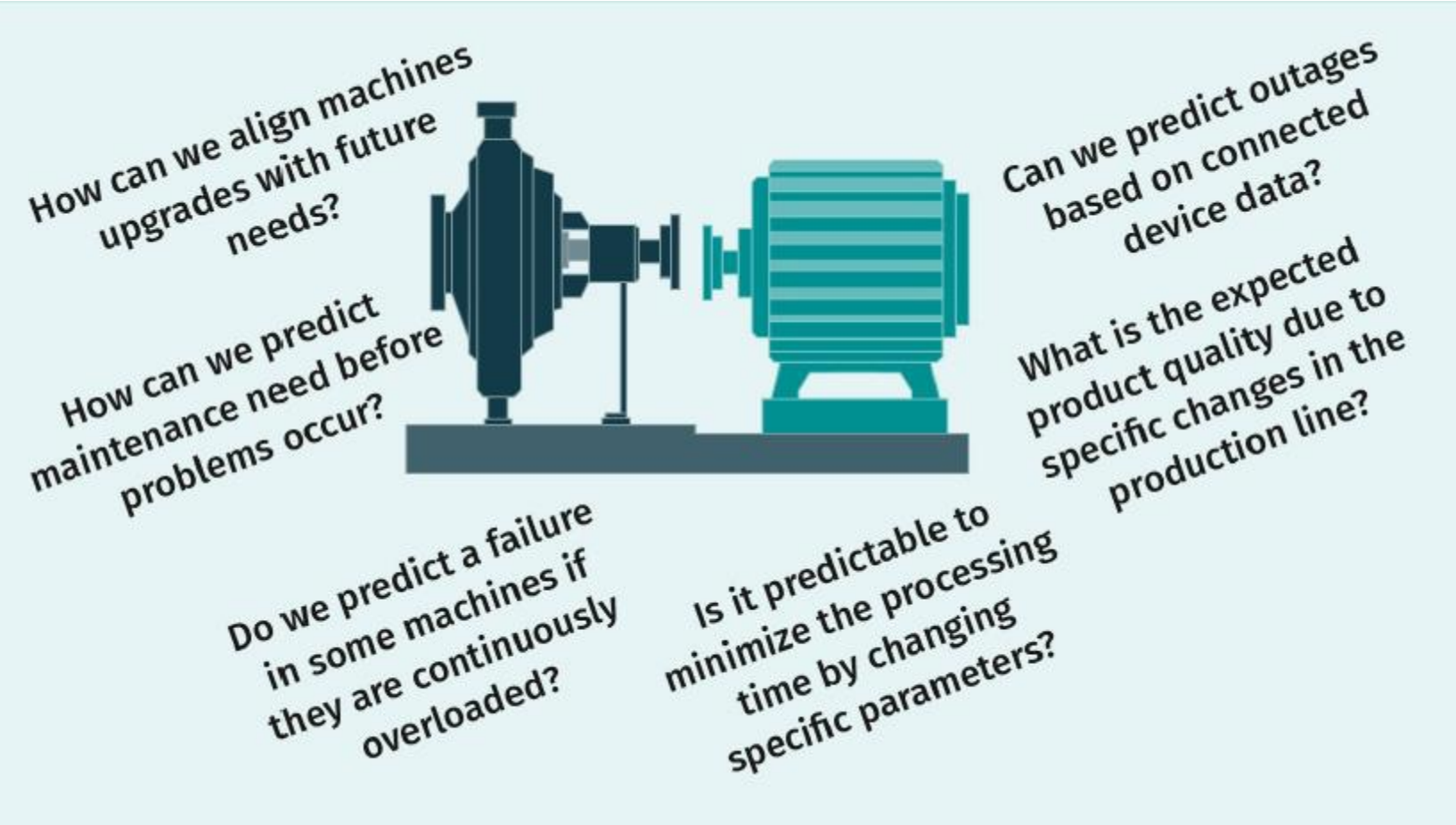
USE CASE IDENTIFICATION AND VALUE PROPOSITION

- Every organization has to identify the kind of use cases to be tackled and ensure that the relevant datasets are available.
- They need to answer the following questions:
  - What is the value of the knowledge gained from applying data science tools to the dataset?
  - What will be learned about the dataset?
  - What will be learned about the hypothesis the data science tools will test?
  - What will be the value of that knowledge if the prediction model developed shows good business performance? If it shows a negative business outcome?

Value Propositions in Customer-Related DSUCs

Source of figure: : IU International University, Course Book DLMBDSA01, Dorard, 2017.

USE CASE IDENTIFICATION AND VALUE PROPOSITION



Value Propositions in Operational-Related DSUCs

Source of figure: : IU International University, Course Book DLMBDSA01, Dorard, 2017.

Value Propositions in Fraud-Related DSUCs

Source of figure: : IU International University, Course Book DLMBDSA01, Dorard, 2017.

- Finding the function which relates selected features of the data (inputs) to the objective value of the DSUC (output)
- The output of a prediction model:
  - A probability (for a **classification** model) or
  - A probability density distribution and/or a number with a degree of uncertainty (for a **regression** model)
- The DSUC value is presented to the end user (e.g., a manager) so they can determine the corresponding action

MACHINE LEARNING CANVAS

The Machine Learning Canvas (v0.4)

Designed for | Designed by | Date | Iteration

Decisions

How are predictions used to make decisions that provide the proposed value to the end-user?

ML Task

Input, output to predict, type of problem

Value Propositions

What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?

Data Sources

Which raw data sources can we use (internal and external)?

Collecting Data

How do we get new data to learn from (inputs and outputs)?

Making Predictions

When do we make predictions on new inputs? How long do we have to featurize a new input and make a prediction?

Offline Evaluation

Methods and metrics to evaluate the system before deployment

Features

Input representations extracted from raw data sources

Building Models

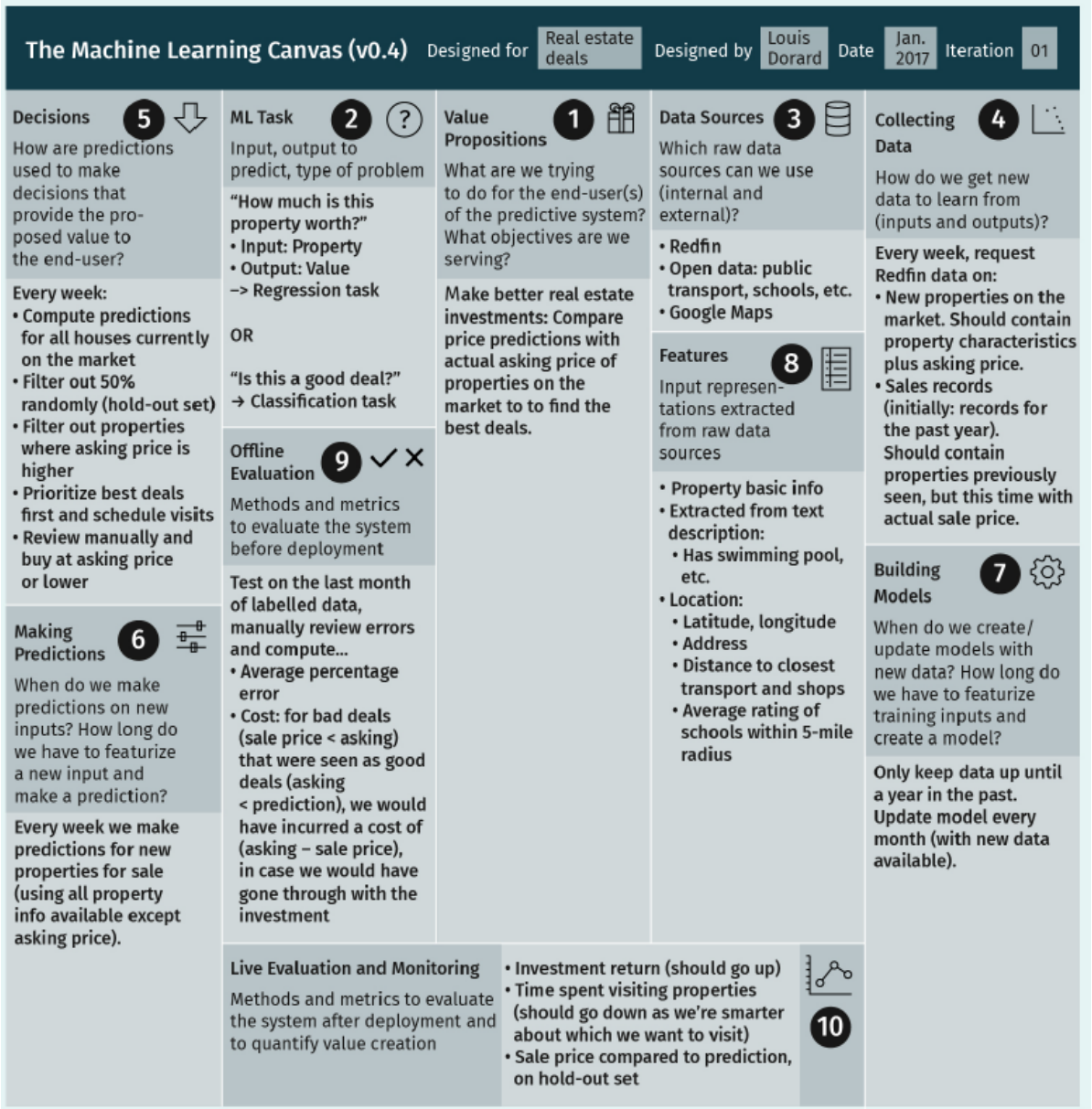
When do we create/update models with new training data? How long do we have to featurize training inputs and create a model?

Live Evaluation and Monitoring

Methods and metrics to evaluate the system after deployment and to quantify value creation

Source of the graphic: IU International University, Course Book DLMBDSA01, Dorard, 2017.

MACHINE LEARNING CANVAS



Source of the graphic: IU International University, Course Book DLMBDSA01, Dorard, 2017.

CLASSIFICATION MODEL EVALUATION METRICS

- **Accuracy** → Fraction of correct predictions (TP+TN) of all predictions
- **Precision** → Cost of False Positive (FP) is high (*Spam Detection*)
- **Recall** → Cost of False Negative (FN) is high (*Cancer Detection*)

$$\text{Accuracy} = \frac{\Sigma TP+TN}{\Sigma\ TP+FP+TN+FN}$$

$$\text{Recall} = \frac{\Sigma TP}{\Sigma\ TP+FN}$$

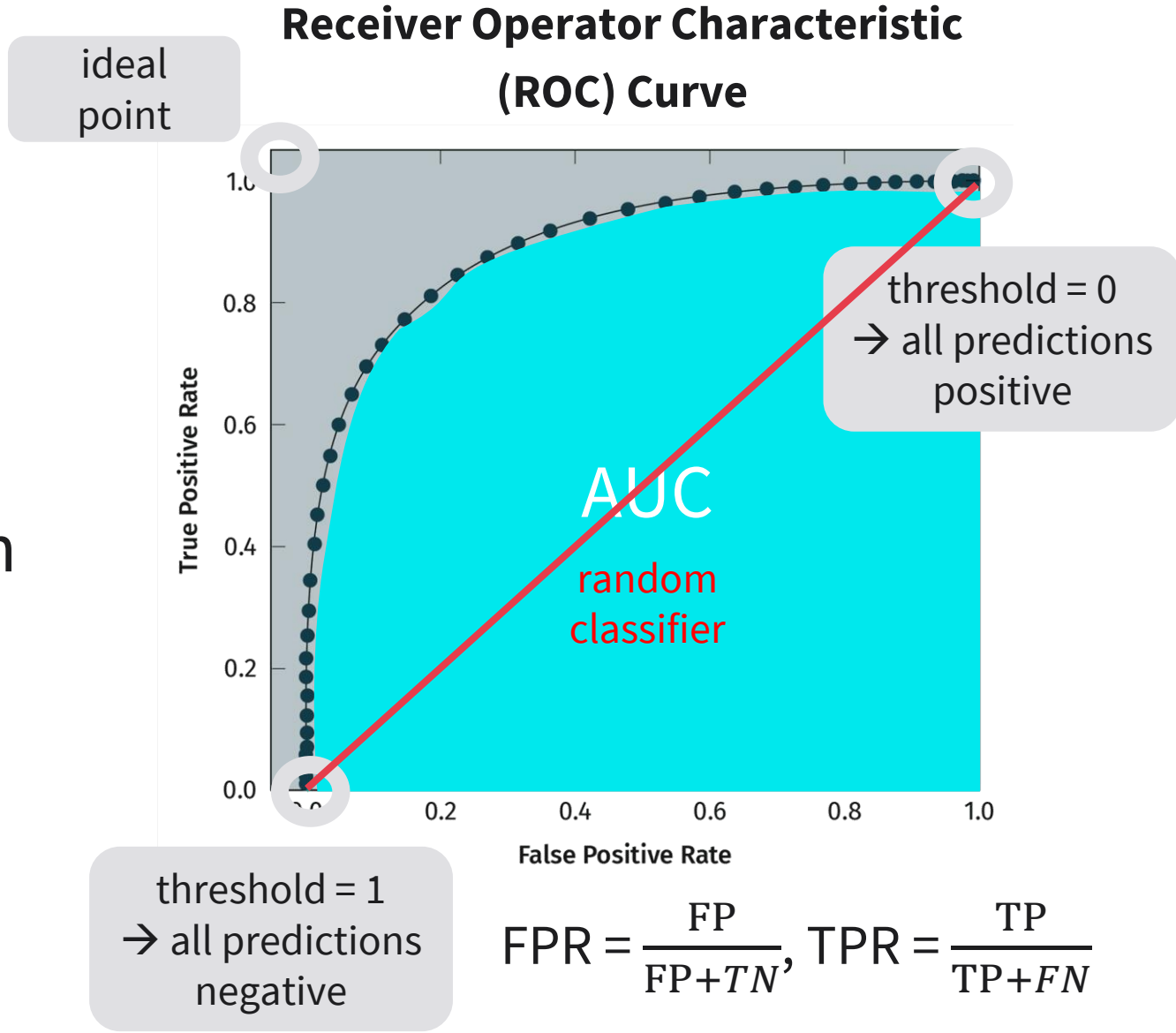
$$\text{Precision} = \frac{\Sigma TP}{\Sigma\ TP+FP}$$

Confusion Matrix

		Actual Class	
		YES	NO
Predicted Class	YES	TP	FP
	NO	FN	TN

Source of the graphic: IU International University, Course Book DLMBDSA01.

- classification models output **probabilities**
- ROC = visualization of model performance with different **thresholds** for a probability to be positive/negative prediction
- best model performance:
  - curve close to upper left corner
  - higher Area under the Curve (AUC)



Source of the graphic: IU International University, Course Book DLMBDSA01.

REGRESSION MODEL EVALUATION METRICS

$$\text{MAE} = \frac{\sum|\hat{Y} - Y|}{n}$$

→ ***robust to outliers***

$$\text{MSE} = \frac{\sum(\hat{Y} - Y)^2}{n}$$

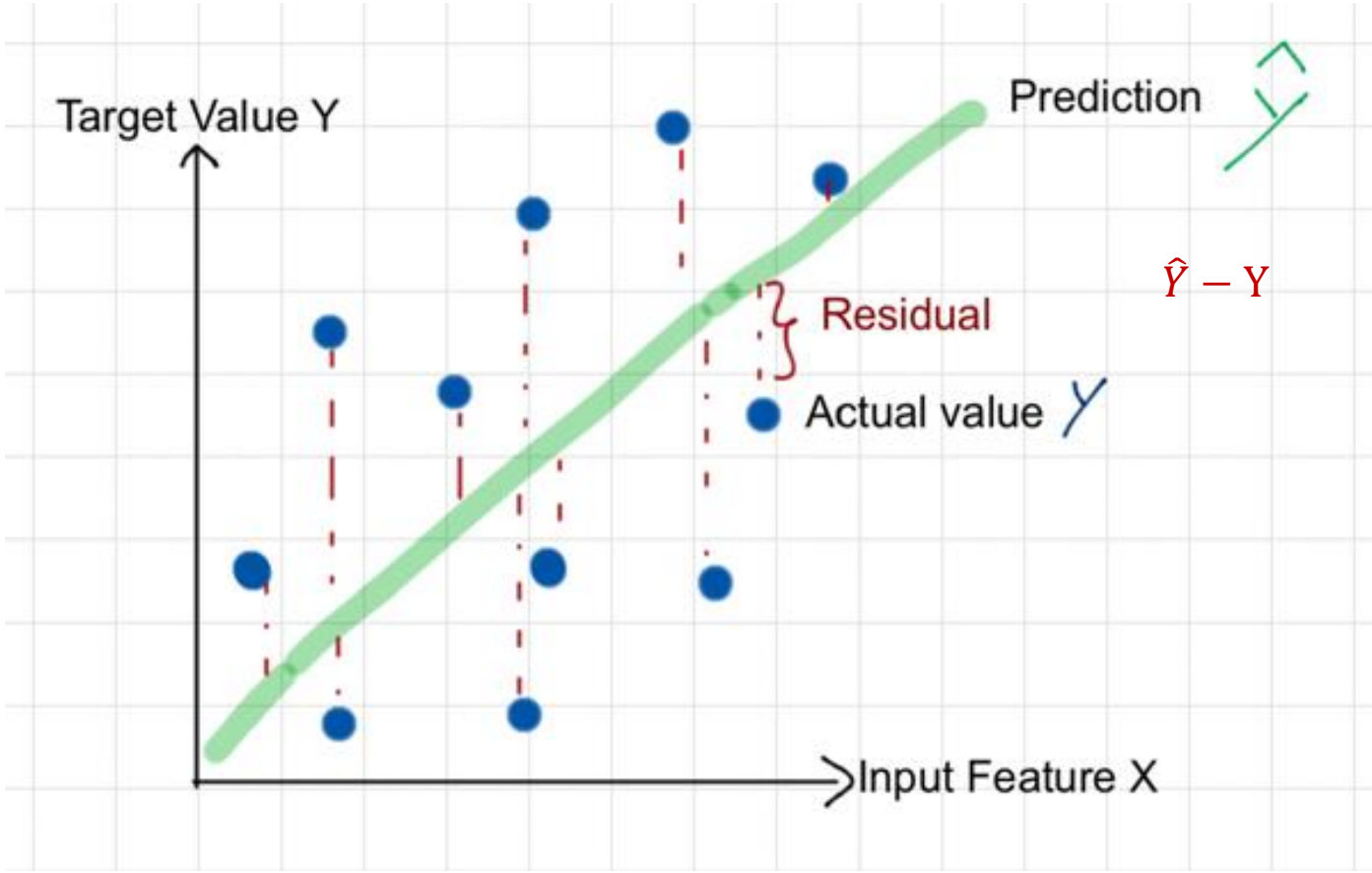
→ ***weights larger errors higher***

$$\text{RMSE} = \sqrt{MSE}$$

→ ***advantage to MSE: original unit***

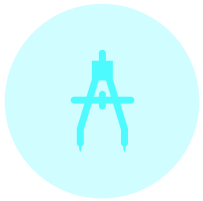
$$\text{MAPE} = \frac{1}{n}\sum\left|\frac{\hat{Y} - Y}{\hat{Y}}\right|$$

→ ***mean of absolute percent differences***



Source of the graphic: Drawing by Antonia Schulze, 2021.

CHARACTERISTICS OF EFFECTIVE BUSINESS KPIS



easy to comprehend and simple to measure

(reduce number of customer complaints)

comprised of small, measurable elements

(amount of daily production, employee workload)

assigned to the relevant task manager

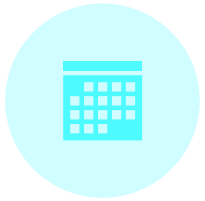
(department head committed)

able to indicate positive/negative variations from the business objective

(increase in products sold)

achievable within the resource constraints

(staff available)



defined with both start and end dates for measuring

visible across the entire organization

(outcome affects multiple departments)

COGNITIVE BIASES AND DE-BIASING TECHNIQUES

Cognitive Bias	Definition	De-biasing techniques
Desirability of options	leads to over- or underestimating probabilities, consequences in a direction that favors a desired alternative	Use incentives and adequate levels of accountability
Confirmation bias	occurs when there is a desire to confirm one's belief, leading to unconscious selectivity in the acquisition and use of evidence	Probe for evidence for alternative hypotheses
Affect influenced	occurs when there is an emotional predisposition for, or against, a specific outcome or option that taints judgments	Involve various stakeholders to get a diverse perspective
Insensitivity to sample size	people tend to ignore sample size and consider extremes equally likely in small and large samples	Use statistics to determine the probability of extreme outcomes in samples of varying sizes

Source of the table: IU International University, Course Book DLMBDSA01 based on Montibeller & Winterfeldt, 2015.

You have learned …

- the importance of a use case for business.
- how to identify use cases.
- the steps to develop a predictive model for a specific use case.
- the metrics to evaluate the performance of a predictive model.
- the role of KPIs in business-centric evaluation.
- the different cognitive biases which influence the decision-making process.

SESSION 2

TRANSFER TASK

TRANSFER TASK

Draft your own data science project checklist. Consider:

- ☑ What are the different steps and aspects to focus on?
- ☑ What are the right questions to ask?
- ☑ Which stakeholders should be involved?
- ☑ Highlight de-biasing techniques in your checklist.

Working in groups

- ☑ Select your domain, e.g., finance, education, e-commerce, insurance, etc.
- ☑ Present in 3-5 minutes

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

1. By increasing the area under the ROC curve we get…

   a) a better performance by the developed classification model.
   b) a worse performance by the developed regression model.
   c) a high false negative rate.
   d) none of the above.

2. The objective of a prediction model is to produce reasonably high accuracy with respect to the…

   a) whole dataset.
   b) cleaned dataset.
   c) testing set.
   d) training set.

3. Cognitive and motivational biases are very important parameters and should be…

   a) included only in the decision-making process.
   b) included only in the pre-processing step.
   c) de-biased and avoided while building the prediction model.
   d) considered when designing the variables of the prediction model variables.

LIST OF SOURCES

Dorard, L. (2017). The machine learning canvas [PDF document]. Retrieved from https://www.louisdorard.com/machine-learning-canvas

Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. O'Reilly Publishers.

Montibeller, G., & Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis. *Risk Analysis, 35*(7), 1230–1251.