

LECTURER: NGHIA DUONG-TRUNG

DATA SCIENCE

TOPIC OUTLINE

Introduction to Data Science

1

Use Cases and Performance Evaluation

2

Data Preprocessing

3

Processing of Data

4

Selected Mathematical Techniques

5

Selected Artificial Intelligence Techniques

6

UNIT 2

USE CASES AND PERFORMANCE EVALUATION



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

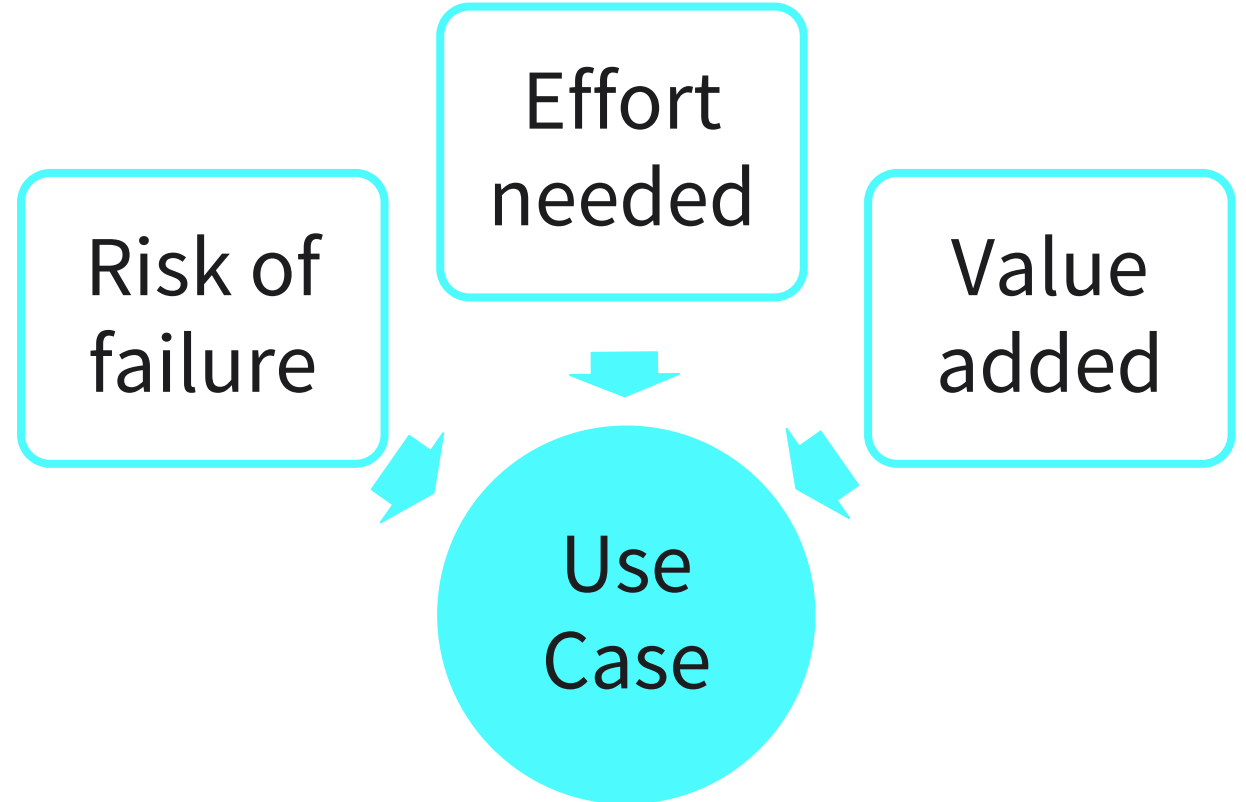
- the importance of a use case for business.
- how to identify use cases.
- the steps to develop a predictive model for a specific use case.
- the metrics to evaluate the performance of a predictive model.
- the role of KPIs in business-centric evaluation.
- the different cognitive biases which influence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



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:

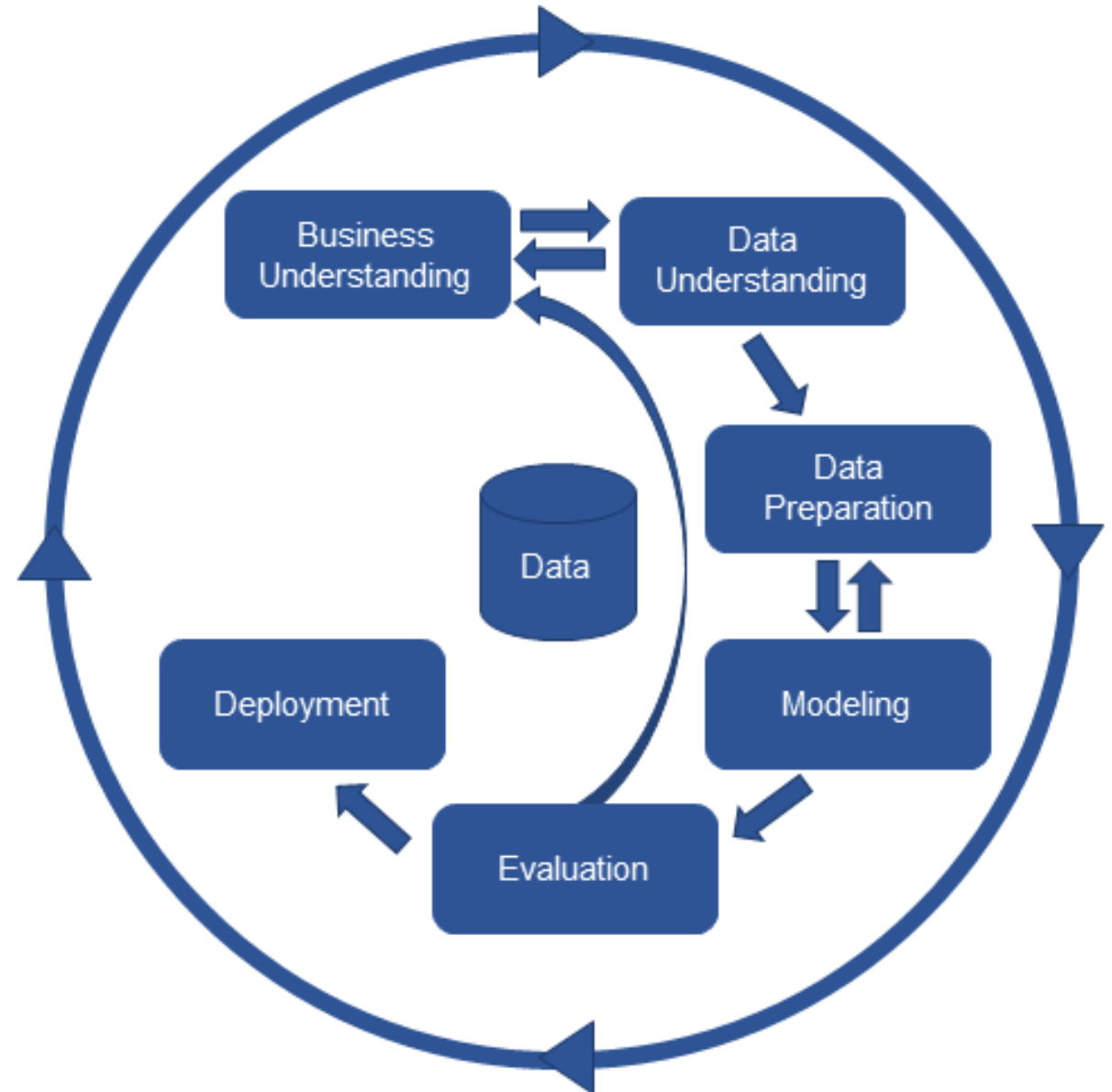
- increasing knowledge gain from data (e.g., better customer understanding)
- reducing business risk (e.g., predict machine outage upfront)
- decreasing effort (e.g., automate processes)



CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

Wirth, R & Hipp, Jochen. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

- business goals remain at the centre of the project
- iterative approach
- both technology and problem-neutral



DATA SCIENCE USE CASES (DSUCS)

<https://data-flair.training/blogs/data-science-use-cases/>

- Data sources
- Type of ML



ONLINE COURSES

<https://www.udemy.com/course/deep-learning-machine-learning-practical/>

Machine Learning Practical Workout | 8 Real-World Projects

- Project 1: ANN – car sales prediction
- Project 2: Deep NN – CIFAR10 classification
- Project 3: Prophet time series – Chicago crime rate
- Project 4: Prophet time series – Avocado market
- Project 5: LE-NET Deep Network – Traffic sign classification
- Project 6: NLP – Email spam filter
- Project 7: NLP – YELP reviews
- Project 8: User-based collaborative filtering – Movie recommender system

ONLINE COURSES

<https://www.superdatascience.com/courses/data-science-for-business-case-studies>

Data Science for Business | 6 Real-world Case Studies

1.Task #1 @Human Resources Department: Develop an AI model to Reduce hiring and training costs of employees by predicting which employees might leave the company.

2.Task #2 @Marketing Department: Optimize marketing strategy by performing customer segmentation

3.Task #3 @Sales Department: Develop time series forecasting models to predict future product prices.

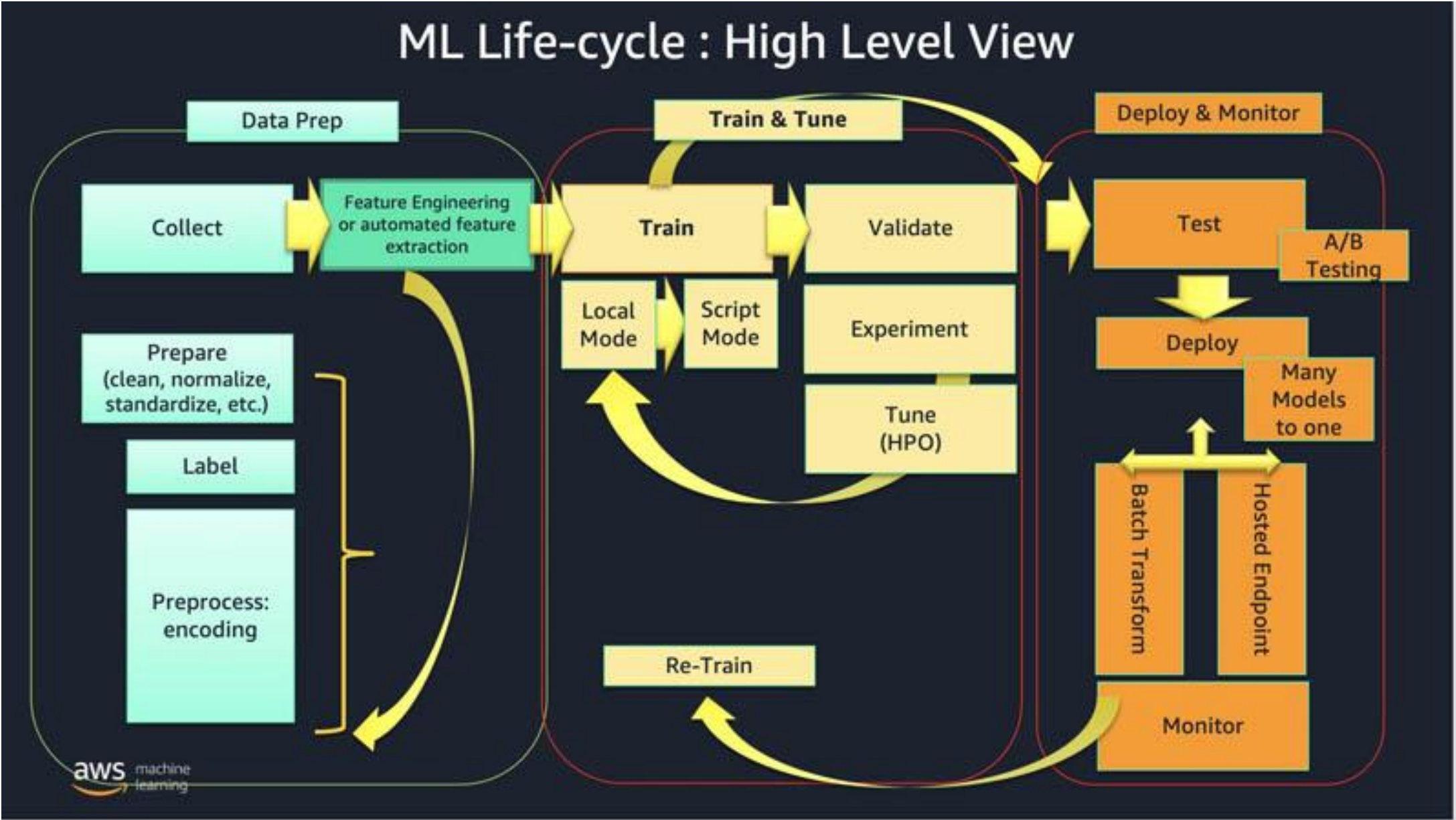
4.Task #4 @Operations Department: Develop Deep Learning model to automate and optimize the disease detection processes at a hospital.

5.Task #5 @Public Relations Department: Develop Natural Language Processing Models to analyze customer reviews on social media and identify customers sentiment.

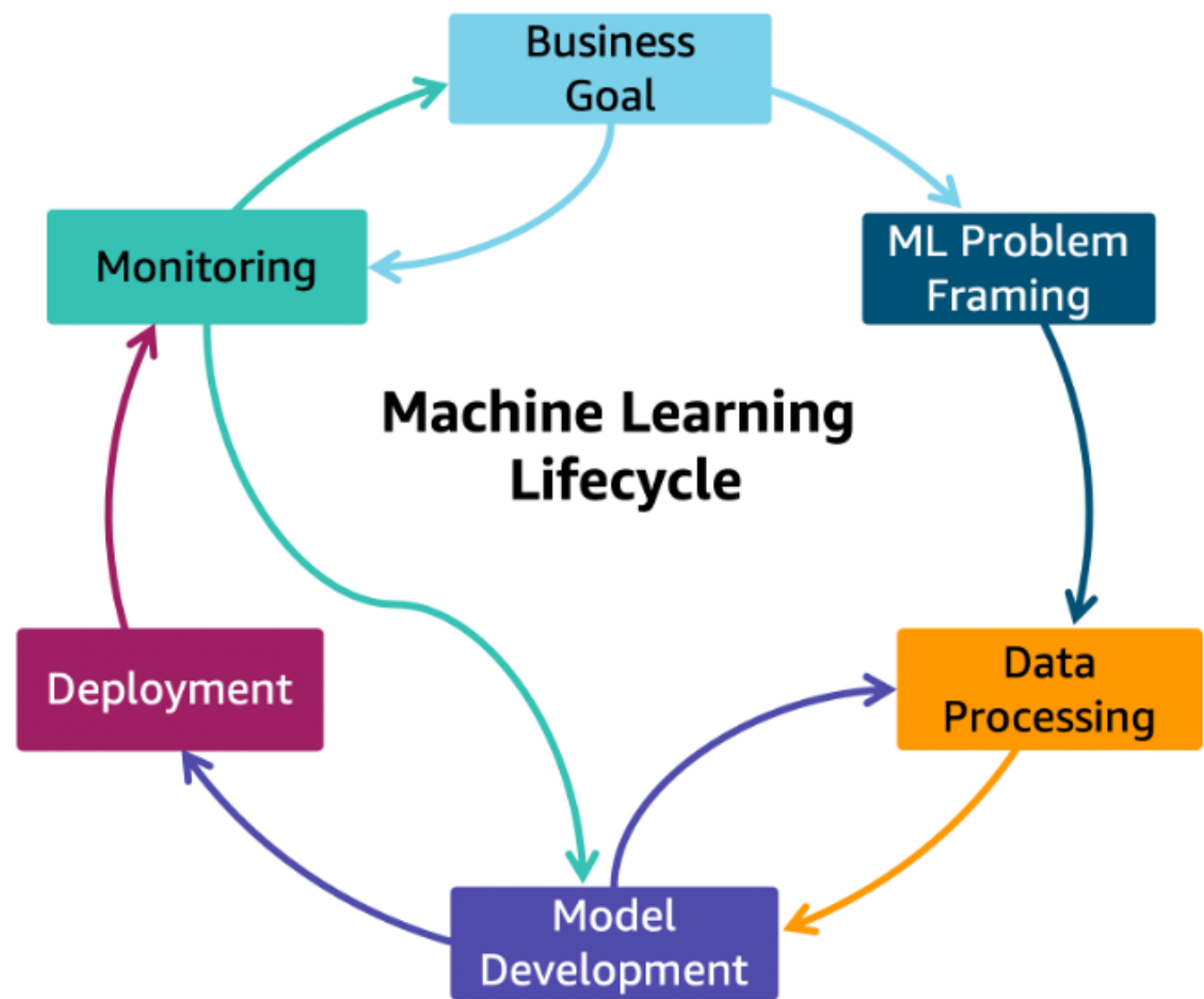
6.Task #6 @Production/Maintenance Departments: Develop defect detection, classification and localization models.

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				

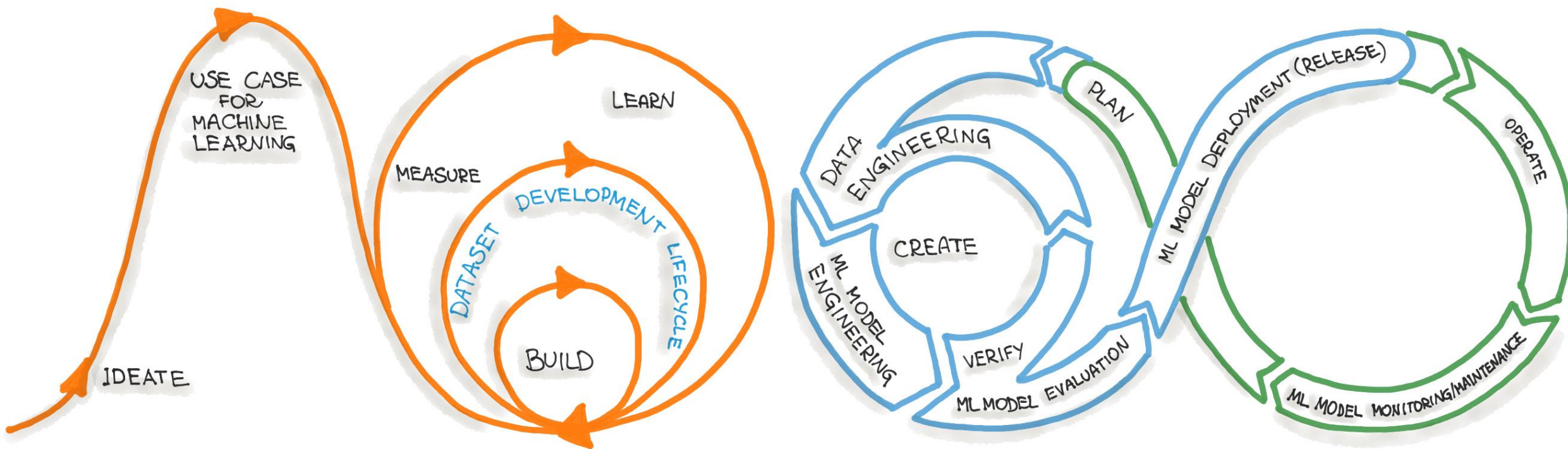
MACHINE LEARNING LIFE CYCLE



MACHINE LEARNING LIFE CYCLE



CRISP-ML(Q)



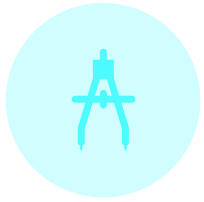
PHASES

BUSINESS & DATA
UNDERSTANDING

MODEL
DEVELOPMENT

MODEL
OPERATIONS

CHARACTERISTICS OF EFFECTIVE BUSINESS KPIS



easy to comprehend
and simple to measure
*(reduce number of
customer complaints)*



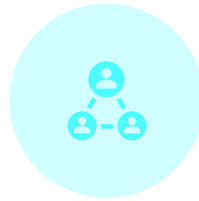
comprised of small,
measurable elements
*(amount of daily
production, employee
workload)*



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

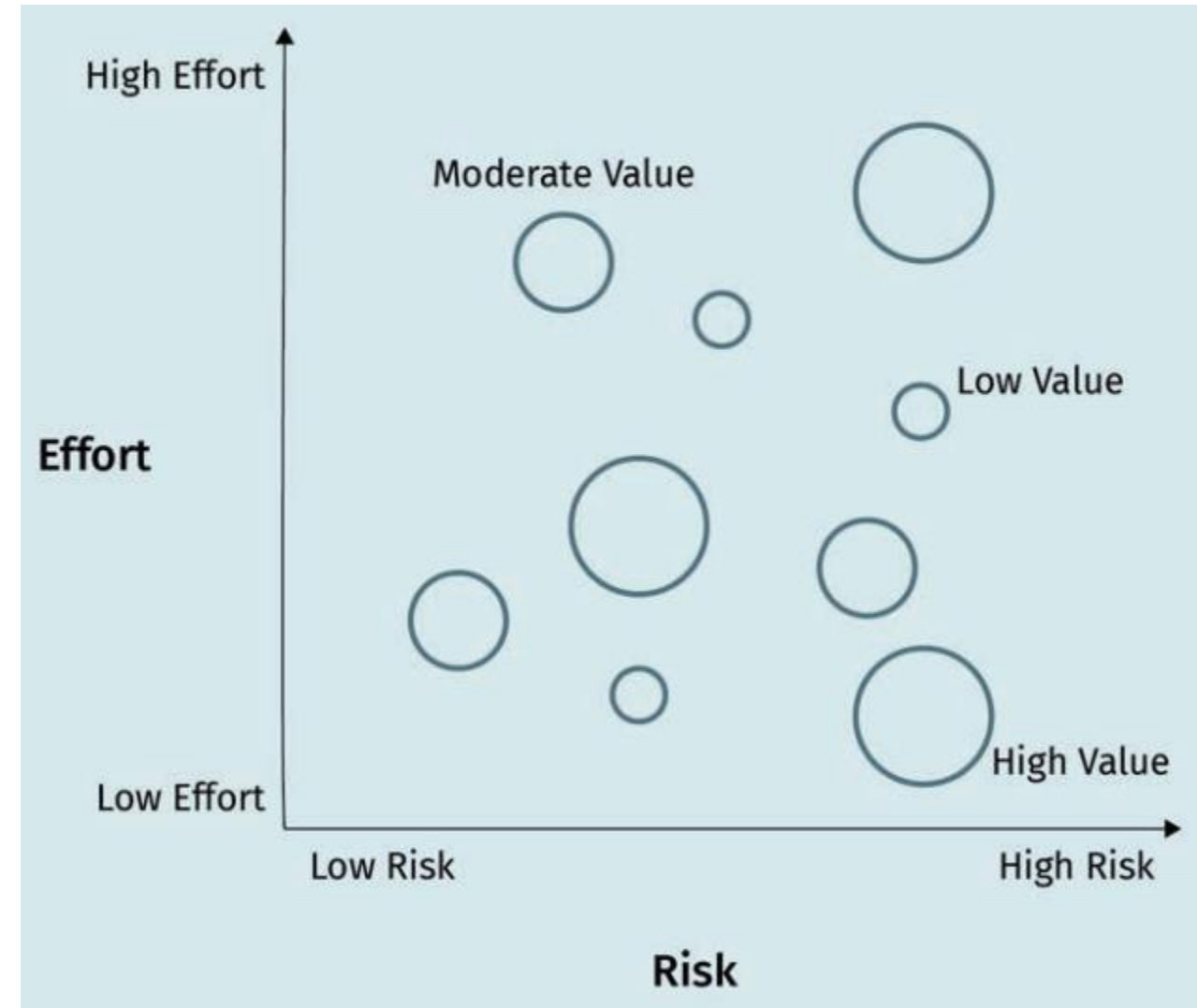


visible across the
entire organization
*(outcome affects
multiple departments)*

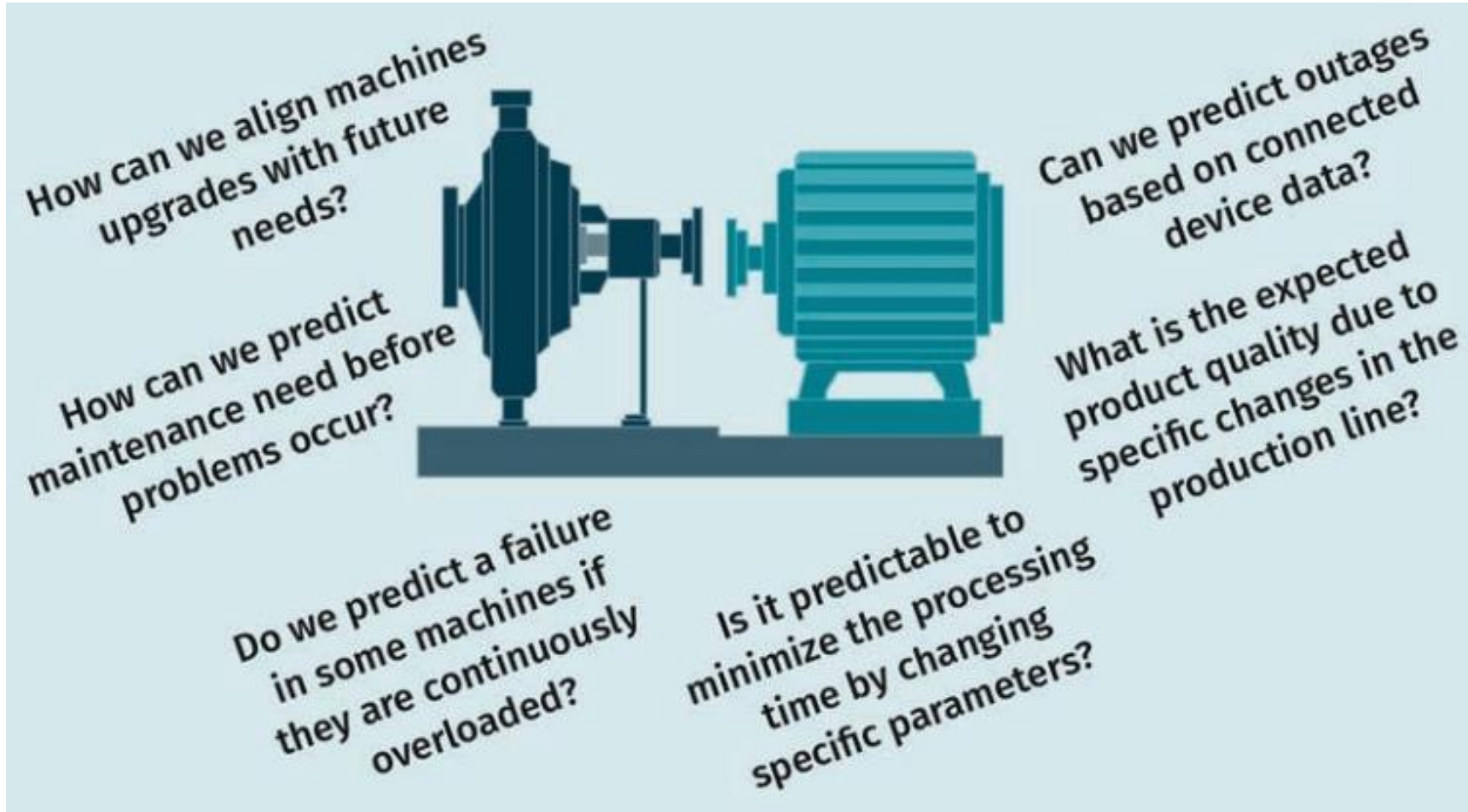
IDENTIFICATION OF AN ORGANIZATION'S USE CASES

Focus on:

- 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 OPERATIONAL-RELATED DSUCS



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*)

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

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

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

Confusion Matrix

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

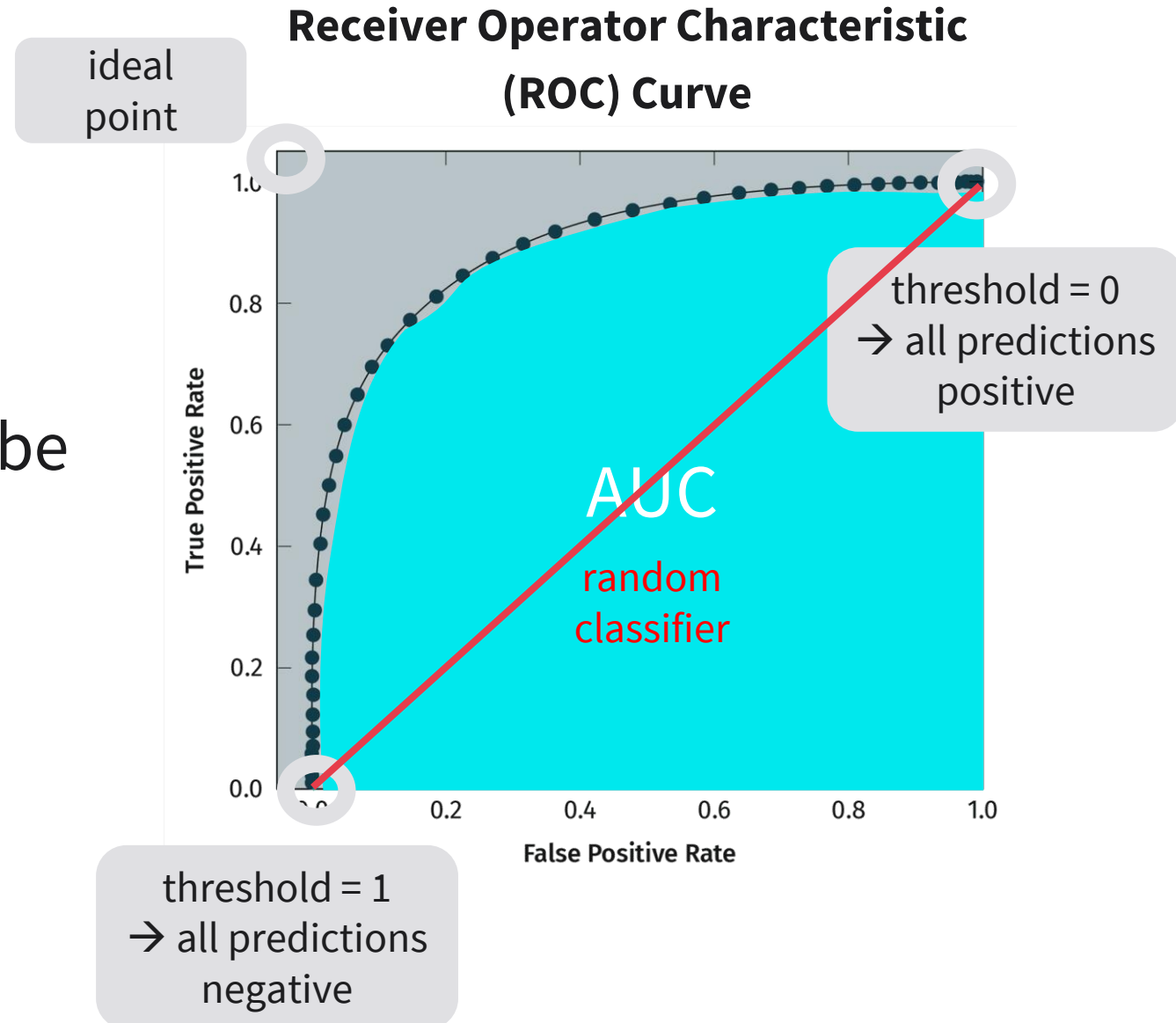
F1-SCORE

- <https://deepai.org/machine-learning-glossary-and-terms/f-score#:~:text=The%20F%2Dscore%2C%20also%20called,positive'%20or%20'negative'.>
- The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

$$\begin{aligned} F_1 &= \frac{2}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \\ &= \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})} \end{aligned}$$

CLASSIFICATION MODEL EVALUATION METRICS

- 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)



REGRESSION MODEL EVALUATION METRICS

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

→ *robust to outliers*

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

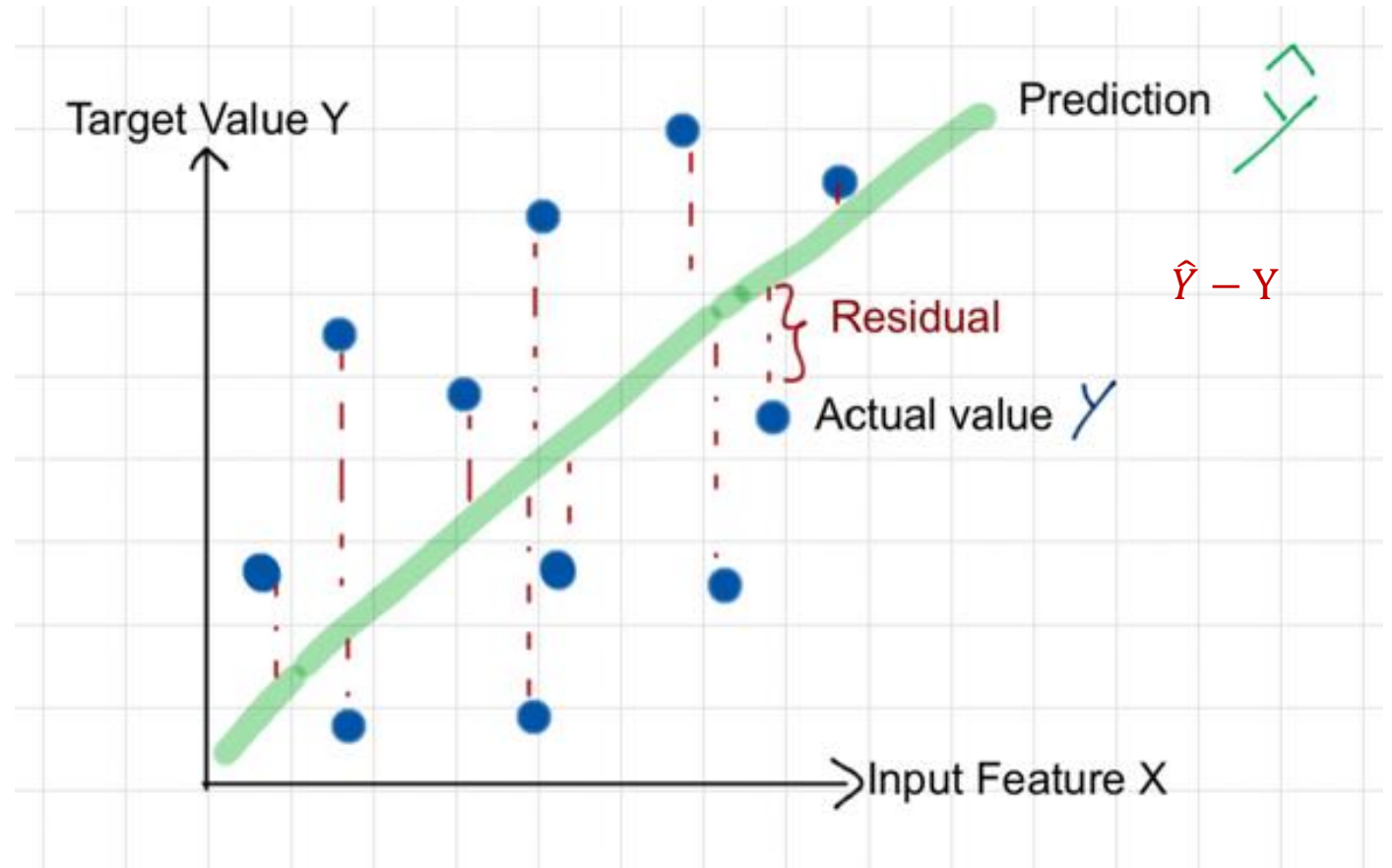
→ *weights larger errors higher*

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

→ *advantage to MSE: original unit*

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

→ *mean of absolute percent differences*



Linear Regression: Single Variable

$$\boxed{\hat{y}} = \beta_0 + \beta_1 \boxed{x} + \boxed{\epsilon}$$

Predicted output

Coefficients

Input

Error

The diagram shows the equation $\hat{y} = \beta_0 + \beta_1 x + \epsilon$. The predicted output \hat{y} is enclosed in a red box, with a red line pointing to the label 'Predicted output' below it. The coefficients β_0 and β_1 are grouped by a green bracket underneath, with a green line pointing to the label 'Coefficients' below the bracket. The input x is enclosed in a blue box, with a blue line pointing to the label 'Input' below it. The error term ϵ is enclosed in an orange box, with an orange line pointing to the label 'Error' below it.

Linear Regression: Multiple Variables

$$\boxed{\hat{y}} = \beta_0 + \beta_1 \boxed{x_1} + \dots + \beta_p \boxed{x_p} + \boxed{\epsilon}$$

The diagram shows the equation $\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$. The predicted output \hat{y} is enclosed in a red box. The coefficients β_0 and β_1 are grouped by a green bracket underneath. Each input variable x_1 and x_p is enclosed in a blue box. The error term ϵ is enclosed in an orange box.

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



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

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.

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.

© 2021 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.