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

STUDY GOALS

On completion of this unit, you will have learned …

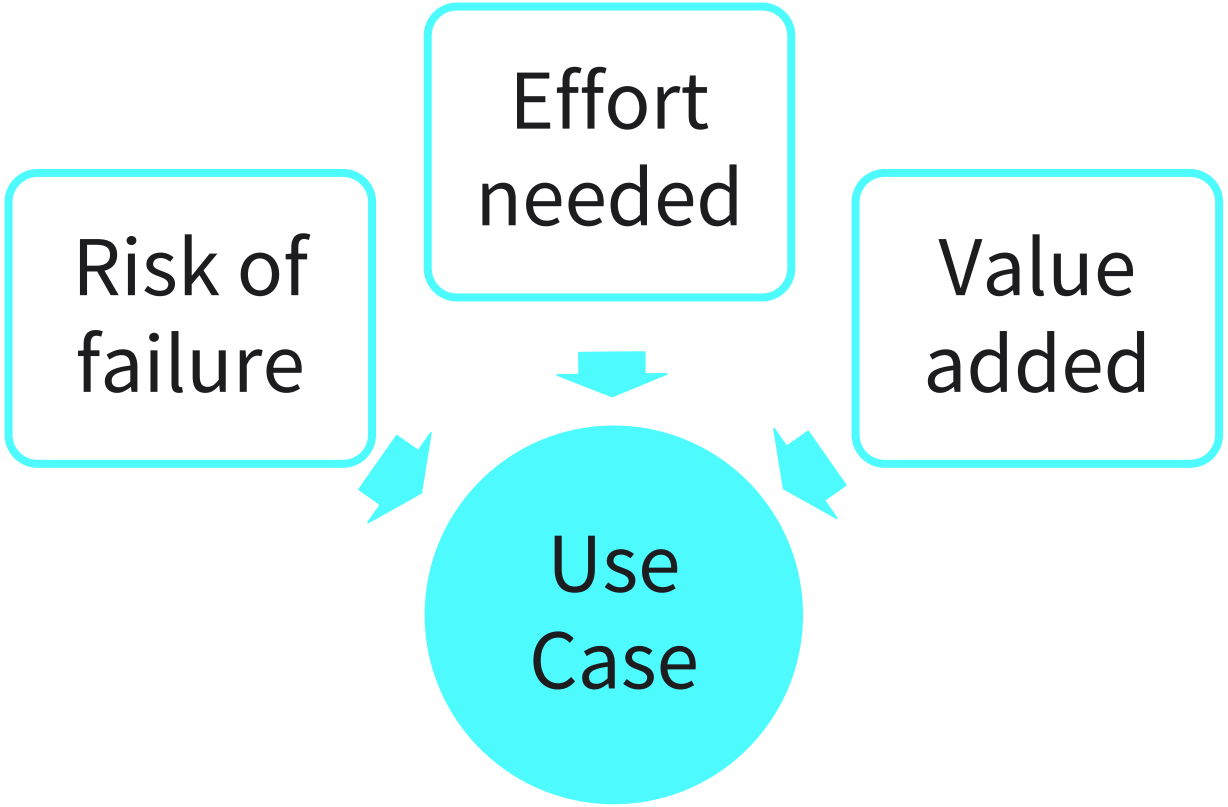
- the importance of a use case for business.
- how to identify use cases.
- the steps to develop a predictive model for a specific use case.
- the metrics to evaluate the performance of a predictive model.
- the role of KPIs in business-centric evaluation.
- the different cognitive biases which influence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

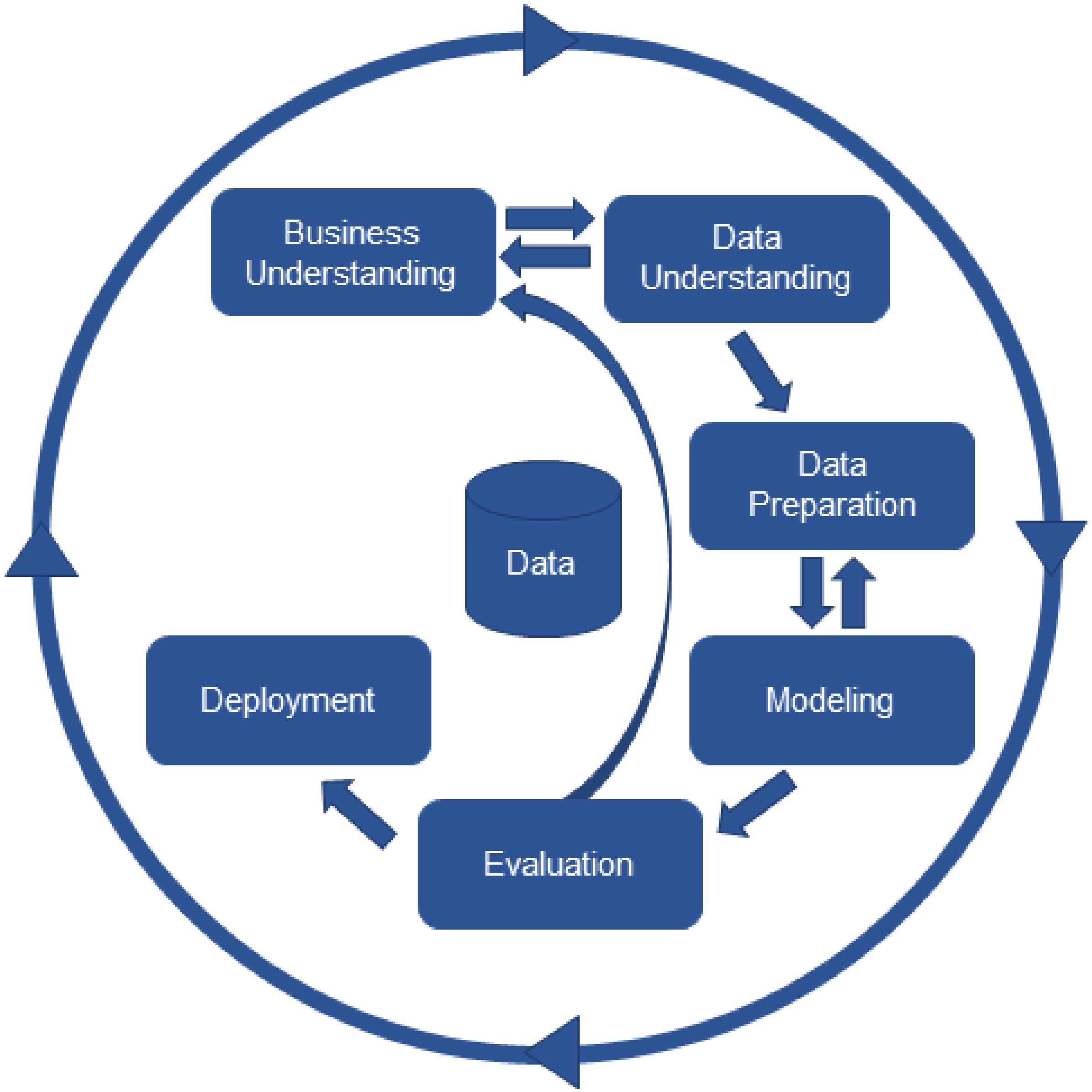
- increasing knowledge gain from data (e.g., better customer understanding)
- reducing business risk (e.g., predict machine outage upfront)
- decreasing effort (e.g., automate processes)



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

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

Machine Learning Practical Workout | 8 Real-World Projects

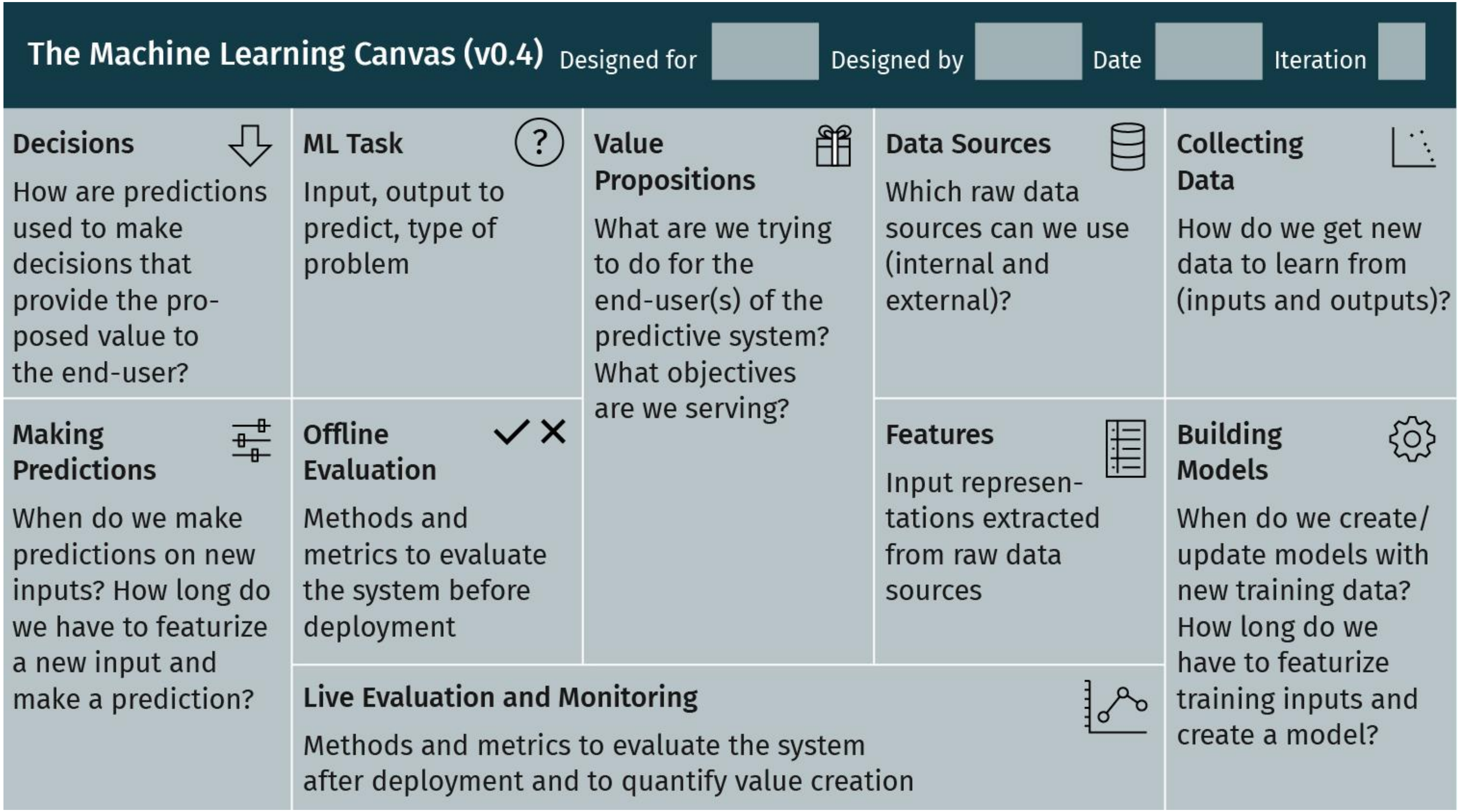
- Project 1: ANN – car sales prediction
- Project 2: Deep NN – CIFAR10 classification
- Project 3: Prophet time series – Chicago crime rate
- Project 4: Prophet time series – Avocado market
- Project 5: LE-NET Deep Network – Traffic sign classification
- Project 6: NLP – Email spam filter
- Project 7: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

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

Data Science for Business | 6 Real-world Case Studies

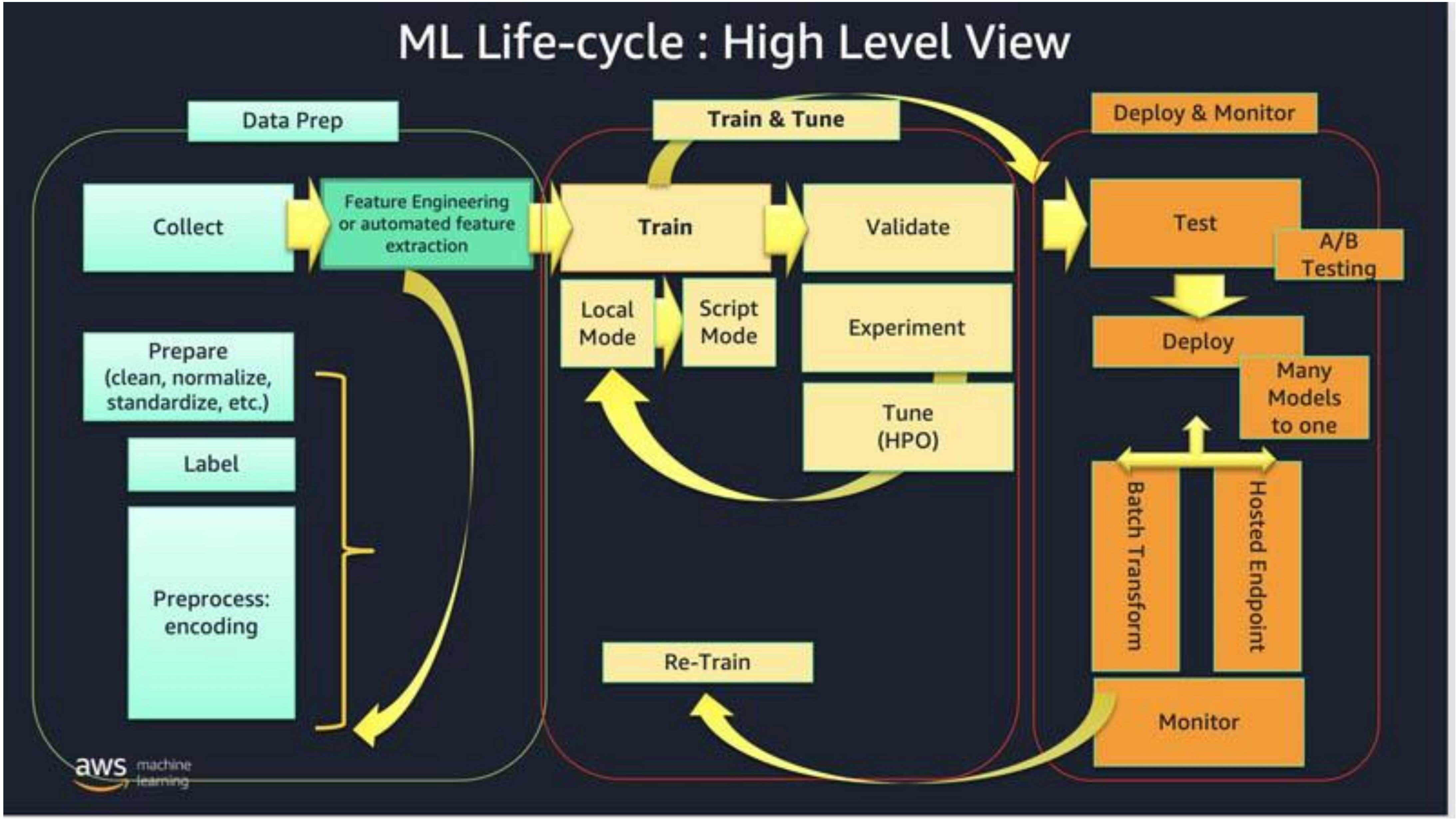
1. **Task #1 @Human Resources Department:** Develop an AI model to Reduce hiring and training costs of employees by predicting which employees might leave the company.
2. **Task #2 @Marketing Department:** Optimize marketing strategy by performing customer segmentation
3. **Task #3 @Sales Department:** 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.

MACHINE LEARNING CANVAS: HTTPS://WWW.OWNML.CO/MACHINE-LEARNING-CANVAS

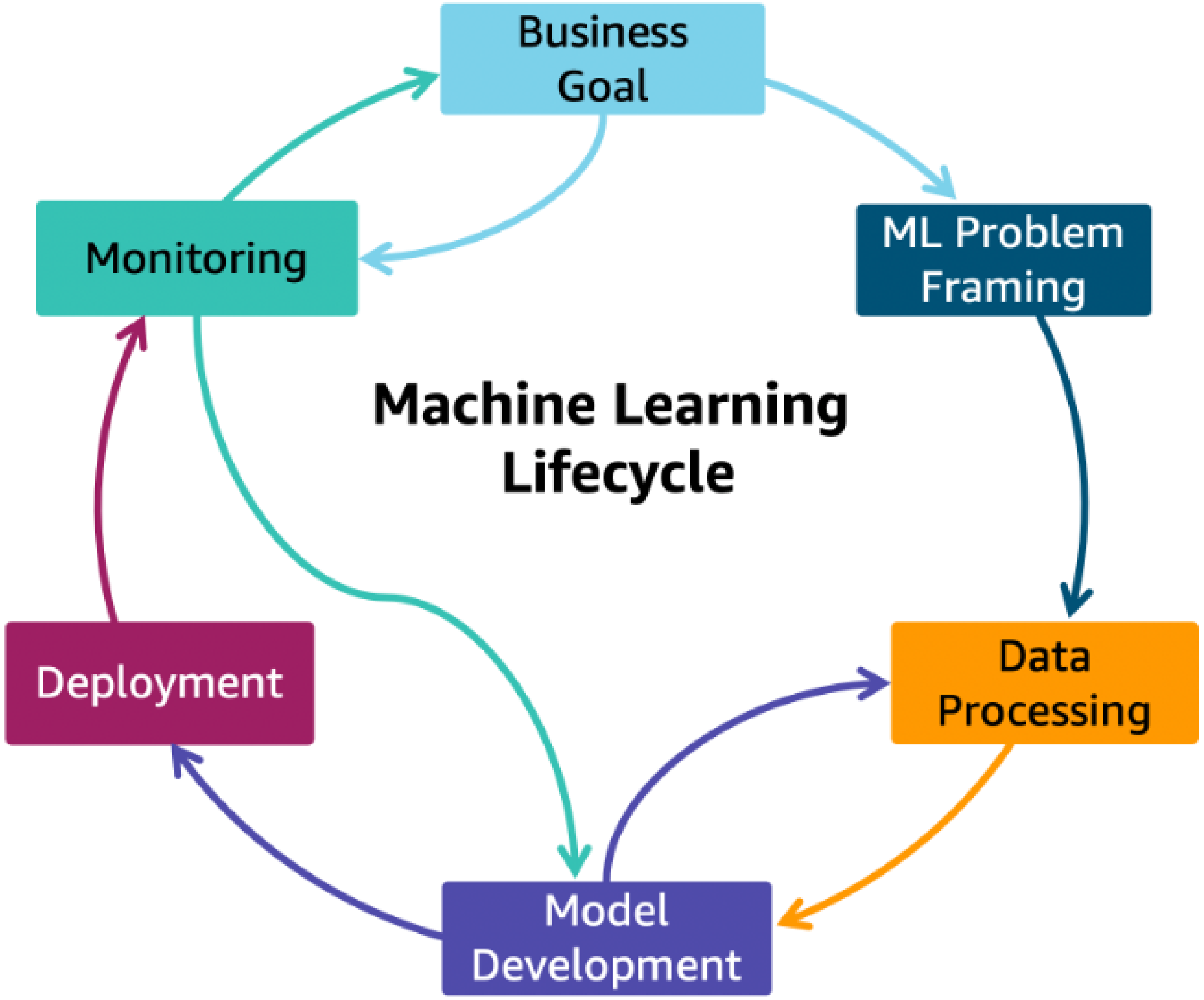


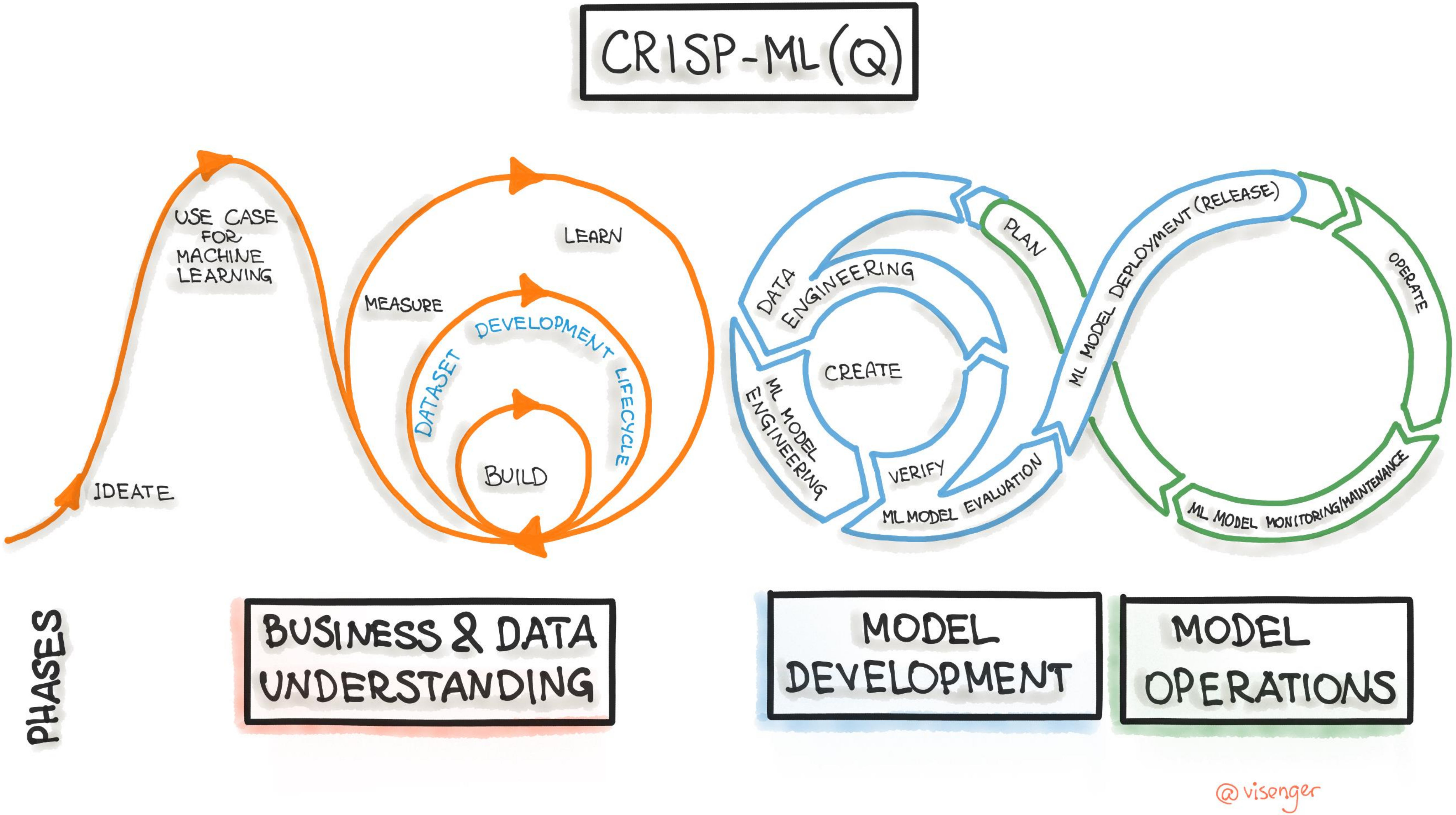
Source of the graphic: IU International University, Course Book DLMBDSA01, Dorard, 2017.

MACHINE LEARNING LIFE CYCLE



MACHINE LEARNING LIFE CYCLE




CRISP-ML(Q)
USE CASE FOR MACHINE LEARNING
IDEATE
MEASURE
LEARN
DATASET DEVELOPMENT LIFECYCLE
BUILD
DATA ENGINEERING
PLAN
CREATE
ML MODEL ENGINEERING
VERIFY
ML MODEL EVALUATION
ML MODEL DEPLOYMENT (RELEASE)
OPERATE
ML MODEL MONITORING/MAINTENANCE
PHASES
BUSINESS & DATA UNDERSTANDING
MODEL DEVELOPMENT
MODEL OPERATIONS
@visenger

CHARACTERISTICS OF EFFECTIVE BUSINESS KPIS

easy to comprehend and simple to measure

(reduce number of customer complaints)

able to indicate positive/negative variations from the business objective

(increase in products sold)

visible across the entire organization

(outcome affects multiple departments)

comprised of small, measurable elements

(amount of daily production, employee workload)

achievable within the resource constraints

(staff available)

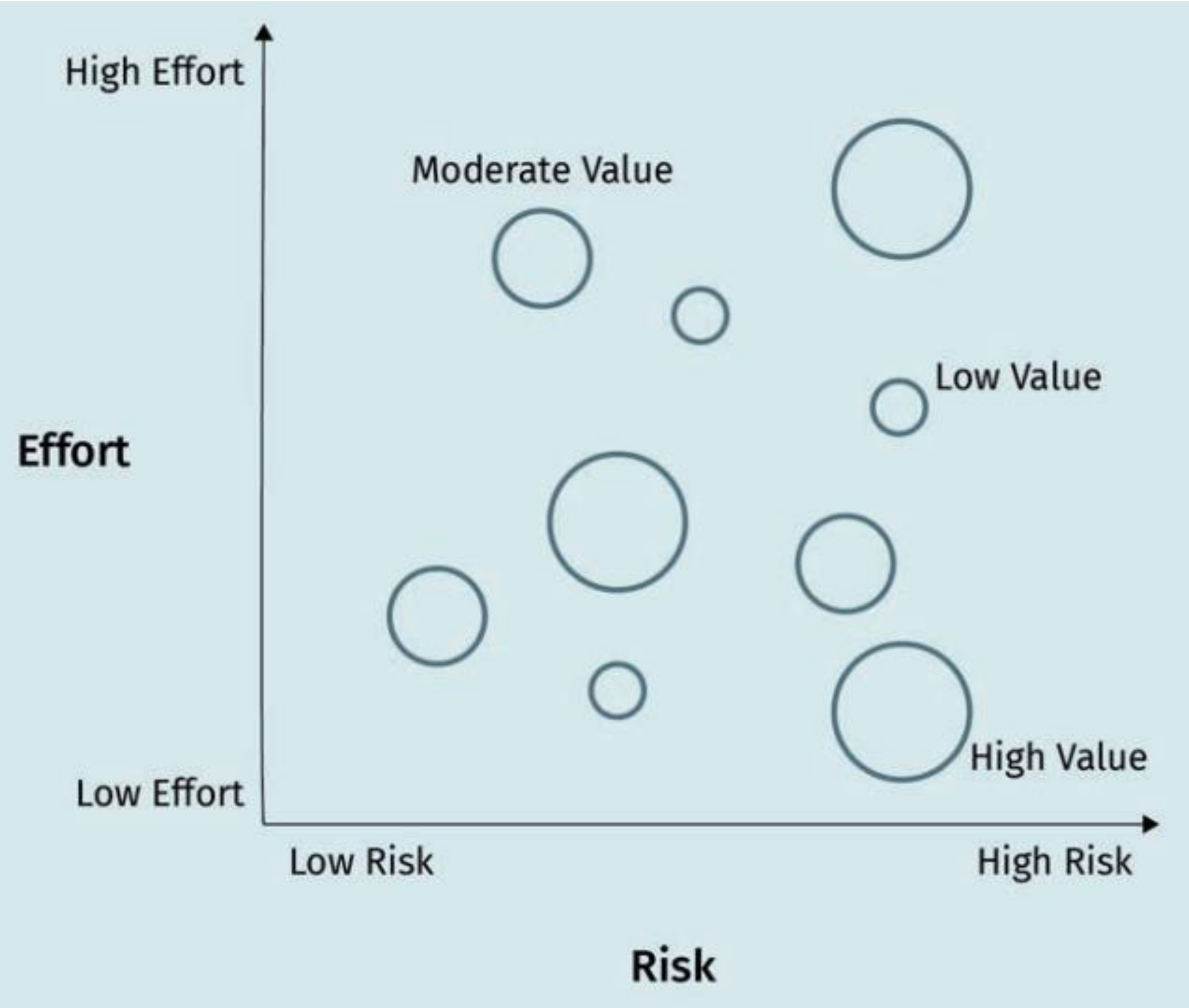
assigned to the relevant task manager

(department head committed)

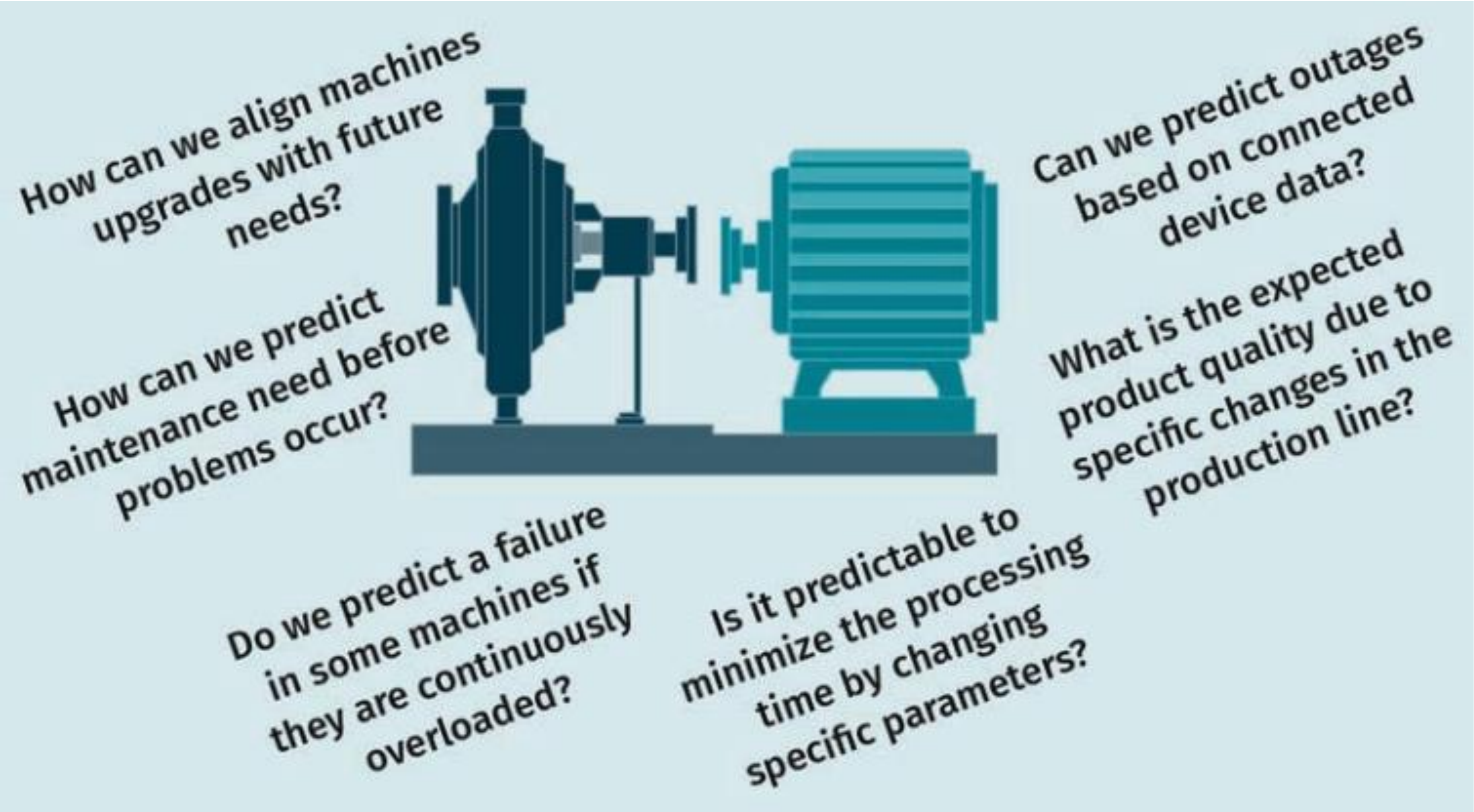
defined with both start and end dates for measuring

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

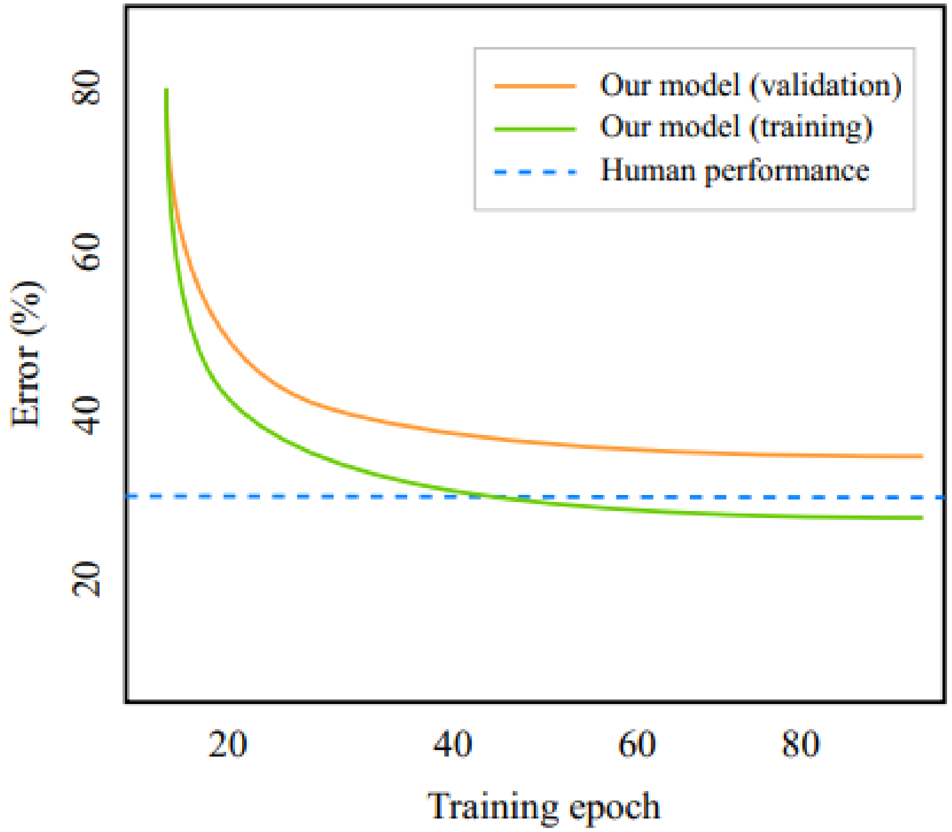


- Goal of ML
  - A model that solves, or helps solve, a business problem. Within a project, the model is often seen as a black box described by inputs, outputs and acceptable level of performance
- Impact of ML
  - ML can replace a complex part in your engineering project or
  - There's a great benefit in getting inexpensive (but probably imperfect) predictions
- Cost of ML
  - The problem difficulty
  - The cost of data, and
  - The need for accuracy
- Simplify the problem
  - Start small for debug, then deploy it bigger

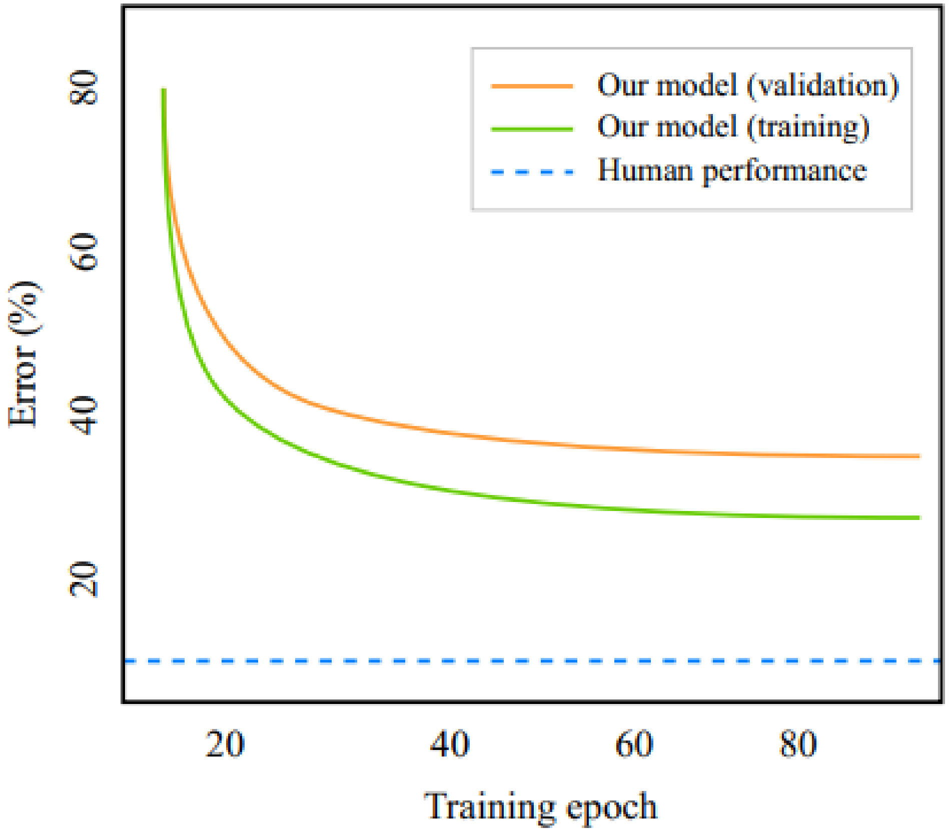
BEFORE THE PROJECT STARTS

- Properties of a successful model
  - It respects the input and output specifications and the performance requirement
  - It benefits the organization (measured via cost reduction, increased sales or profit)
  - It helps the user (measured via productivity, engagement, and sentiment)
  - It is scientifically rigorous
- Team of ML
  - ML skill, software development skill, data engineering skill, data labeling skill, research skill, DevOps
- Why ML projects fail
  - Lack of experienced talent
  - Lack of clearly defined expected deliverables
  - Data infrastructure and labeling challenge
  - Lack of collaboration and alignment
  - Technically infeasible projects

MODEL PERFORMANCE VS HUMAN BASELINE



(a)



(b)

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.

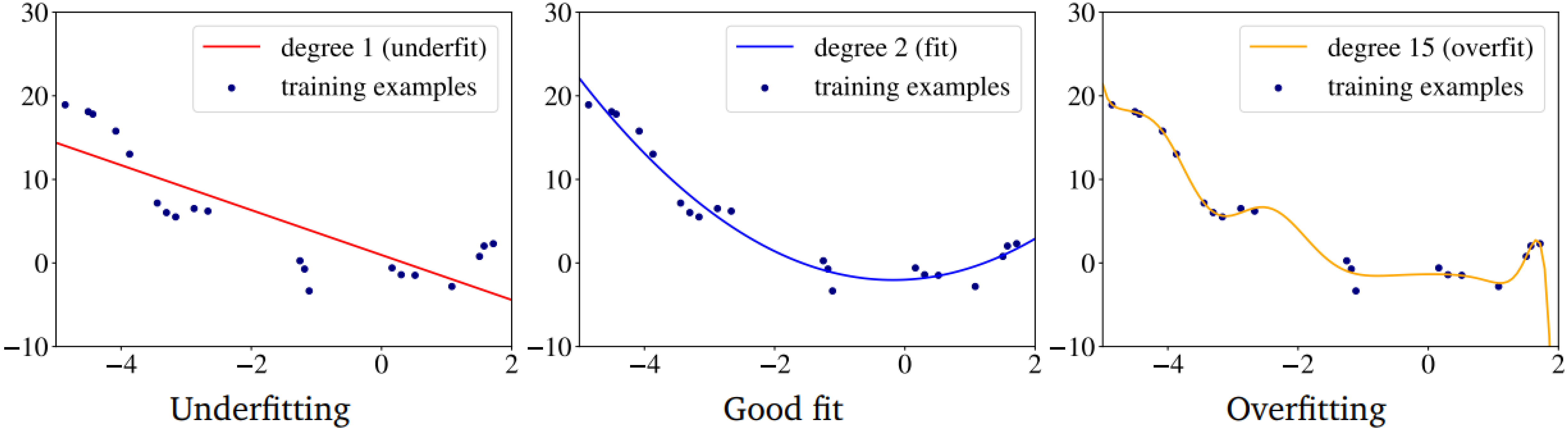
Linear Regression: Single Variable

$$\underset{\text{Predicted output}}{\boxed{\hat{y}}} = \underbrace{\beta_0 + \beta_1}_{\text{Coefficients}} \underset{\text{Input}}{\boxed{x}} + \underset{\text{Error}}{\boxed{\epsilon}}$$

Linear Regression: Multiple Variables

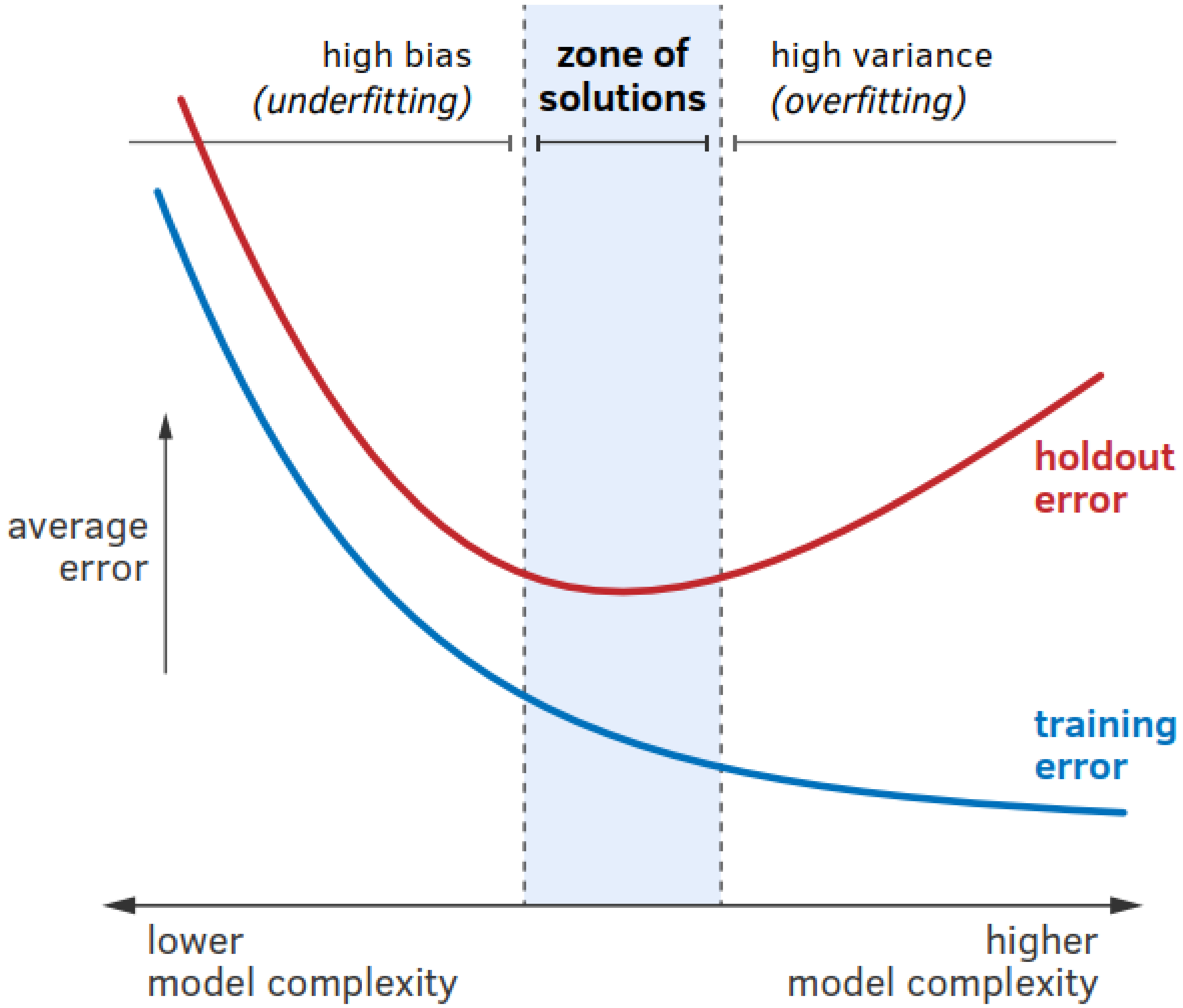
$$\boxed{\hat{y}} = \underbrace{\beta_0 + \beta_1}\boxed{x_1} + \ldots + \underbrace{\beta_p}\boxed{x_p} + \boxed{\epsilon}$$

UNDERFITTING AND OVERFITTING



BIAS-VARIANCE TRADEOFF

- Reducing overfitting leads to underfitting, and the other way Around
- Zone of solutions
  - Move to the right by increasing the complexity of the model, and, By so doing, reducing its bias, or
  - Move to the left by regularizing the model to reduce variance by making the model simpler
    - Regularization adds a penalizing term Whose value is higher when the model is more complex



- **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*)

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

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

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

Confusion Matrix

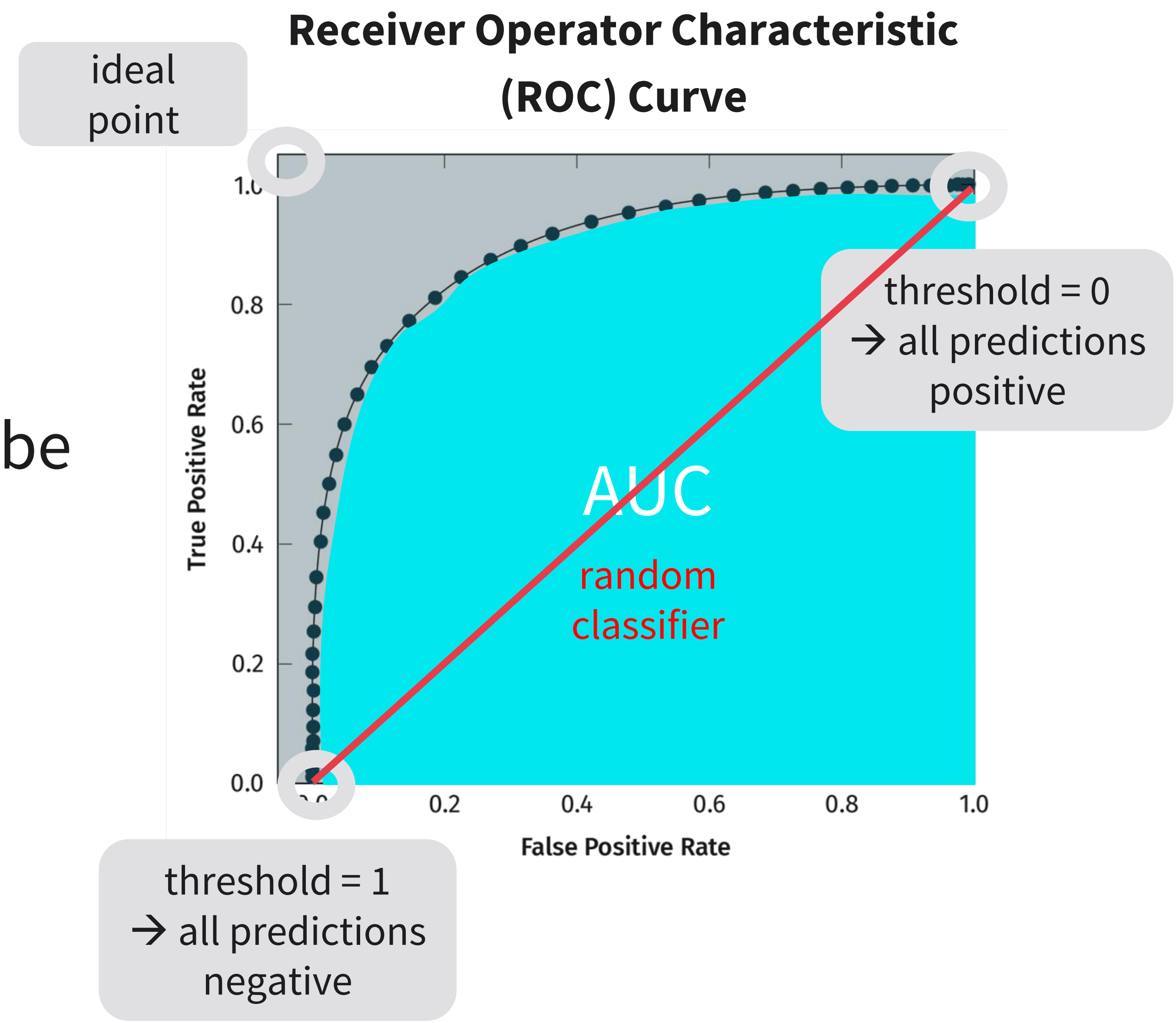
		Actual Class	
		YES	NO
Predicted Class	YES	TP	FP
	NO	FN	TN

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

- 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.

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

- 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*

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

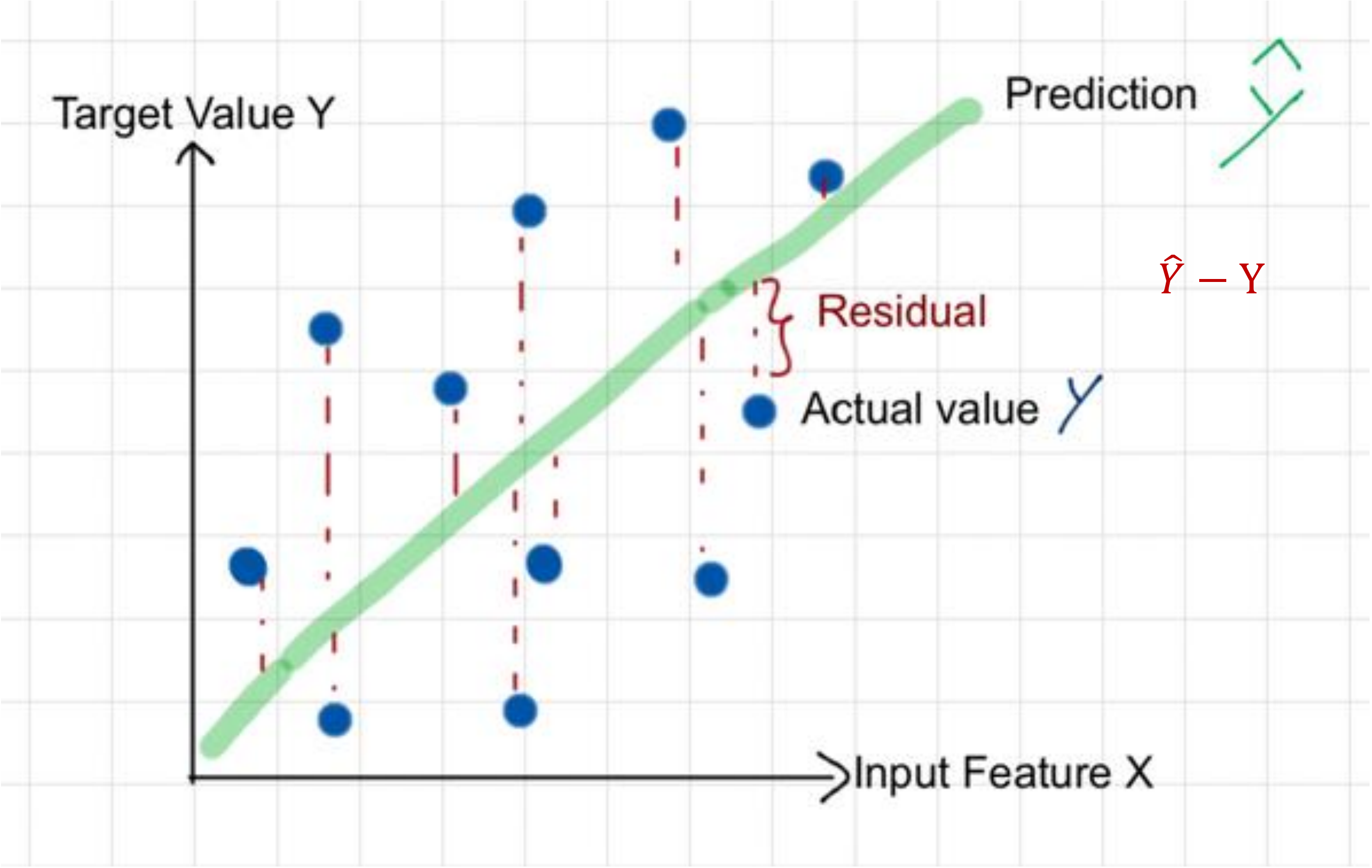
→ *weights larger errors higher*

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

→ *advantage to MSE: original unit*

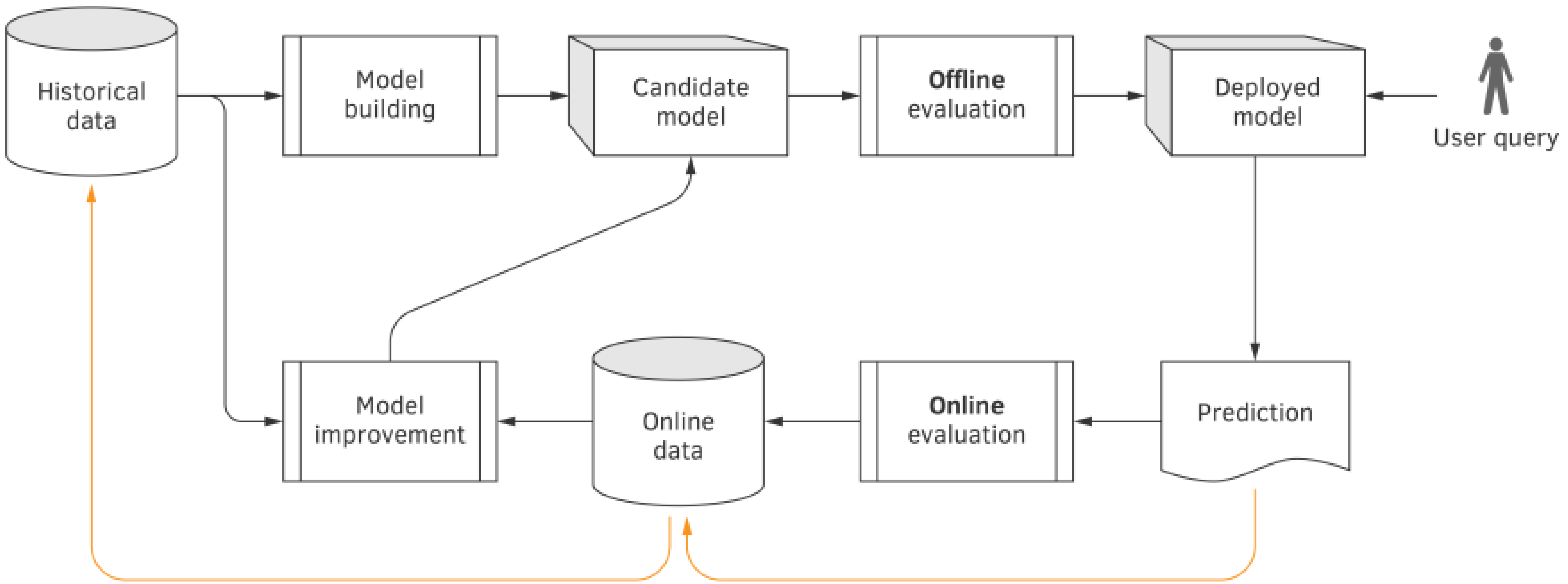
$$\text{MAPE} = \frac{1}{n}\sum\left|\frac{\hat{Y} - Y}{\hat{Y}}\right|$$

→ *mean of absolute percent differences*



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

- An offline model evaluation happens when the model is being trained by the analyst
- The online evaluation happens when the model is being tested in production by using online data



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

THE MACHINE LEARNING CANVAS

Designed for: Designed by: Date: Iteration:

PREDICTION TASK

Type of task? Entity on which predictions are made? Possible outcomes? Wait time before observation?

DECISIONS

How are predictions turned into proposed value for the end-user? Mention parameters of the process / application that does that.

VALUE PROPOSITION

Who is the end-user? What are their objectives? How will they benefit from the ML system? Mention workflow/interfaces.

DATA COLLECTION

Strategy for initial train set & continuous update. Mention collection rate, holdout on production entities, cost/constraints to observe outcomes.

DATA SOURCES

Where can we get (raw) information on entities and observed outcomes? Mention database tables, API methods, websites to scrape, etc.

IMPACT SIMULATION

Can models be deployed? Which test data to assess performance? Cost/gain values for (in)correct decisions? Fairness constraint?

MAKING PREDICTIONS

When do we make real-time / batch pred.? Time available for this + featurization + post-processing? Compute target?

BUILDING MODELS

How many prod models are needed? When would we update? Time available for this (including featurization and analysis)?

FEATURES

Input representations available at prediction time, extracted from raw data sources.

MONITORING

Metrics to quantify value creation and measure the ML system's impact in production (on end-users and business)?

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.

DISCLAIMER

- This is the modified version of the IU slides.
- I used it for my lectures at IU only.