NANYANG
THE INNOVATIVE POLYTECHNIC

Introduction to AI and Machine Learning

School of Engineering
Nanyang Polytechnic

Topics

1. Overview of AI, ML, and DL
2. Machine Learning Types and Techniques
3. Machine Learning Modeling Process
4. Practical: Regression Models in ML

3. Machine Learning Modelling Process

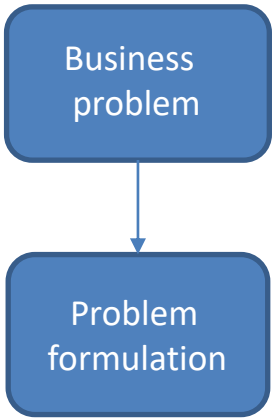
3.1. The Machine Learning Pipeline

3.2. Use Case and Applications

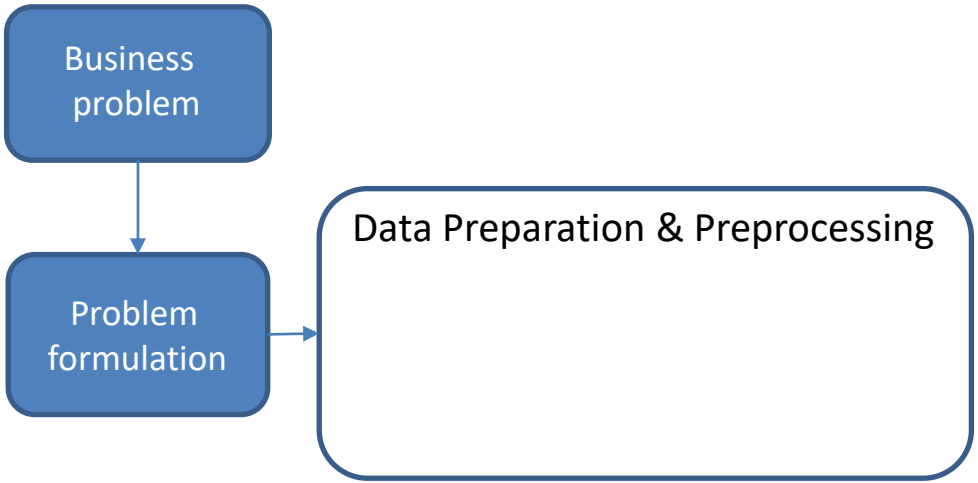
3.3. Activity: Business Problem and Formulation

3.1. The Machine Learning Pipeline

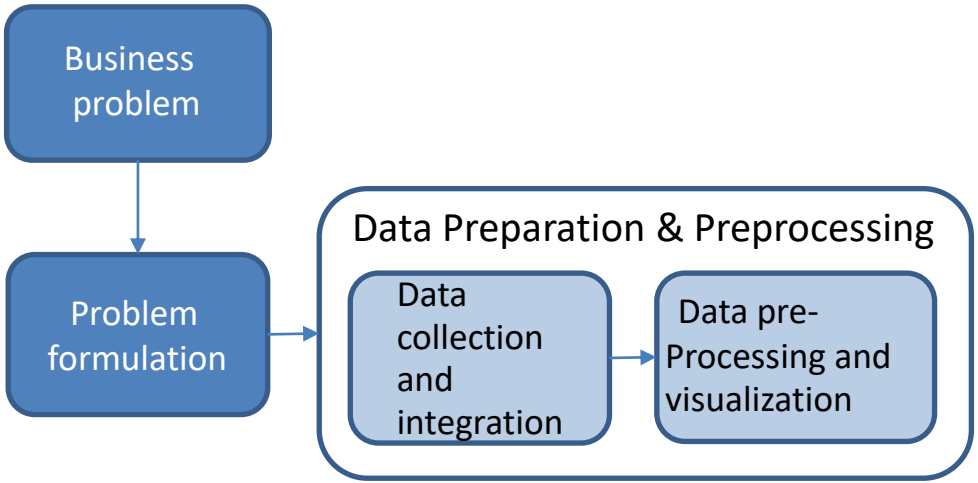
The ML Pipeline



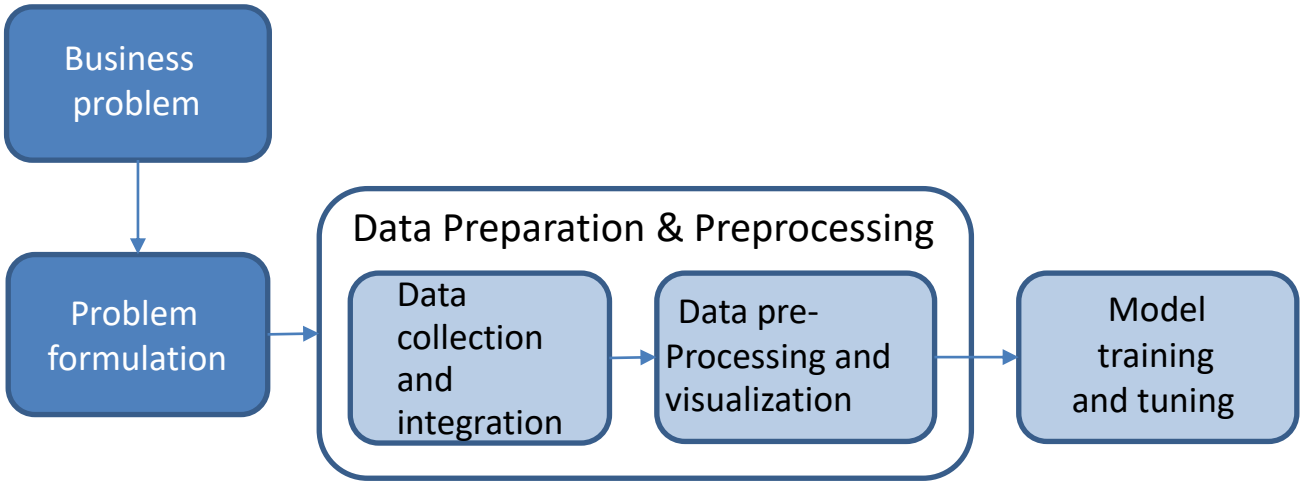
The ML Pipeline



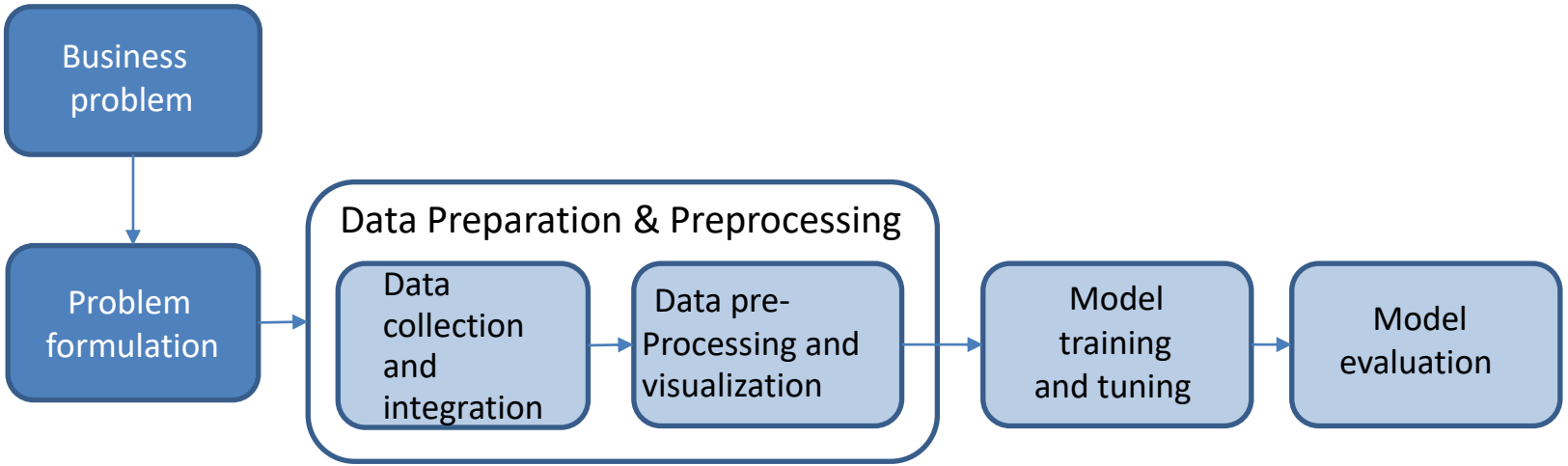
The ML Pipeline



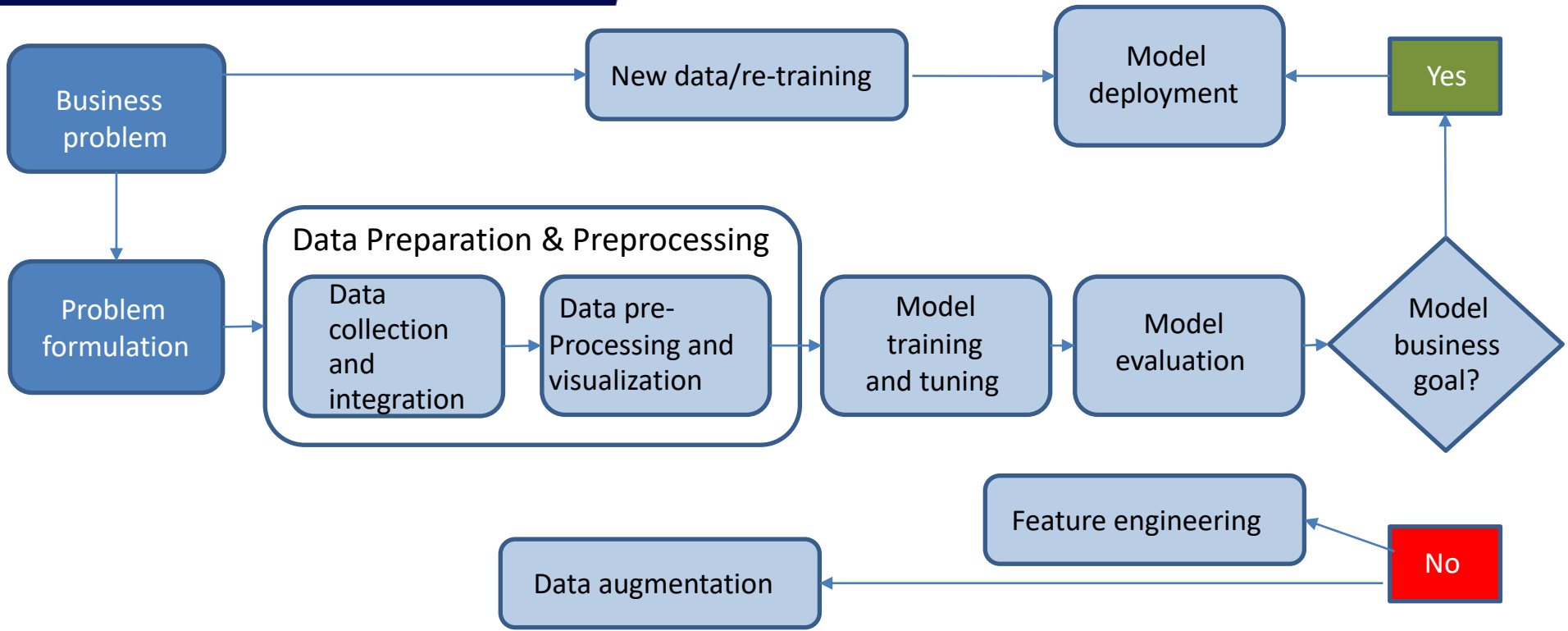
The ML Pipeline



The ML Pipeline



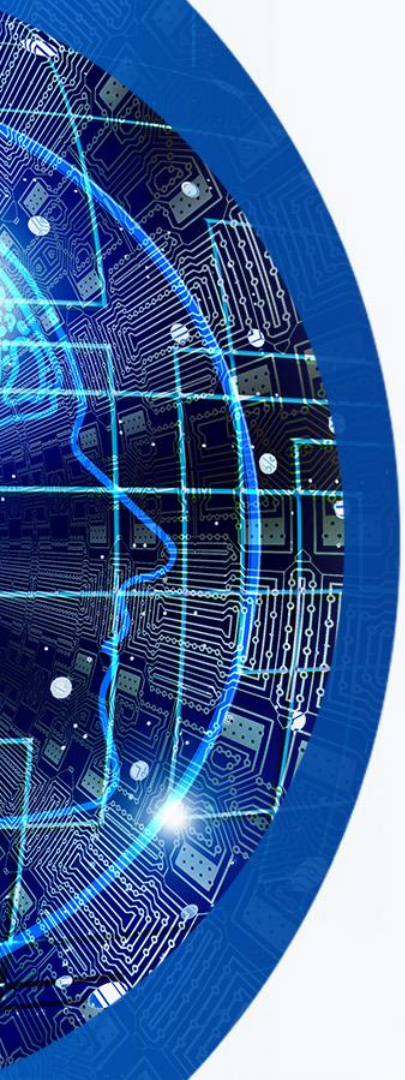
The ML Pipeline



ML is not a solution for every type of problem

You can solve it with simple rule or computations

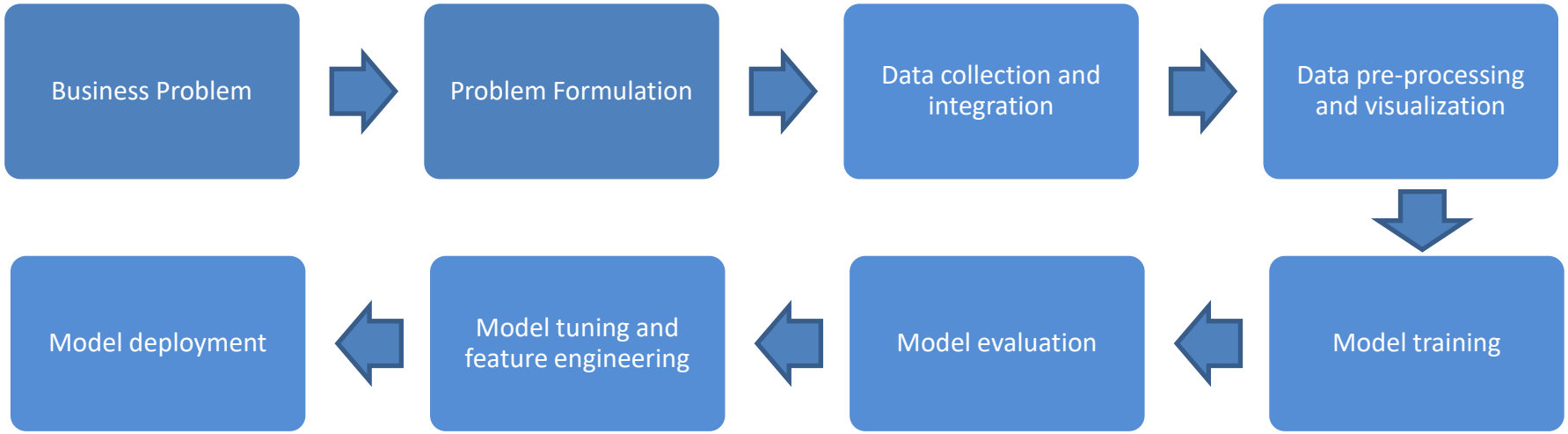
- You cannot code the rule to make a prediction
- You cannot scale predictions



Machine learning may help address a variety of business needs

- Categorization
- Predictive routing
- Fraud detection
- Personalized Advertising
- Voice assistants
- Dynamic pricing
- Email filtering
- Self-driving cars
- Customer churn prediction

Applying the ML Process



3.2. Use Case and Applications

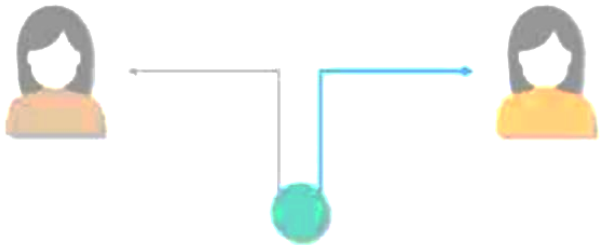
Case study: The Amazon call center problem

Business problem and problem formulation phase

Business Problem → Problem Formulation → Data collection and integration → Data pre-processing and visualization → Model training → Model evaluation → Model tuning and feature engineering → Model deployment

Business Problem

How do you route customers to the correct agent?



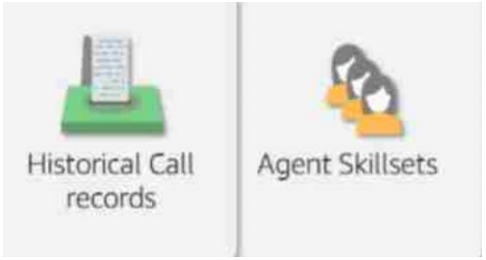
ML Problem

Identifying patterns in customer data that we could use to predict accurate customer routing.

Data collection and integration phase

Business Problem → Problem Formulation → **Data collection and integration** → Data pre-processing and visualization → Model training → Model evaluation → Model tuning and feature engineering → Model deployment

Since we wanted to base our prediction on past data from customer service calls we were dealing with supervised learning



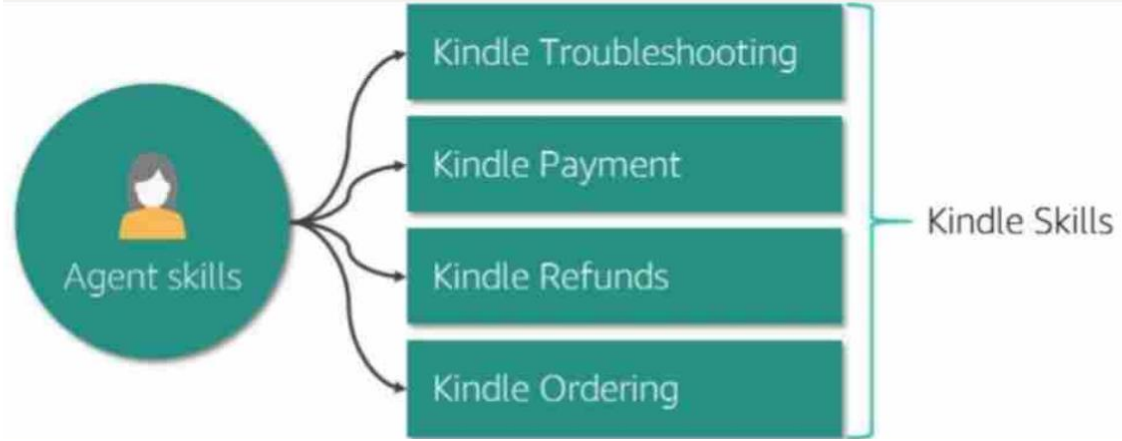
?

- What were customer recent order?
- Did customer own a kindle?
- Were they Prime member?

Data pre-processing and visualization phase

Business Problem → Problem Formulation → Data collection and integration → **Data pre-processing and visualization** → Model training → Model evaluation → Model tuning and feature engineering → Model deployment

This phase includes data cleaning and exploratory data analysis

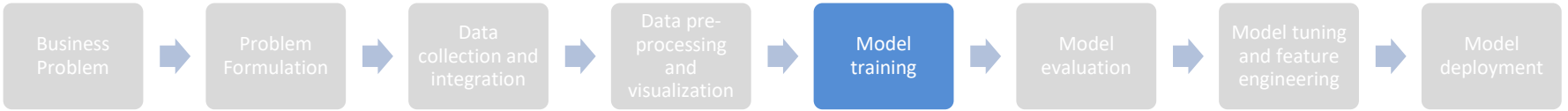


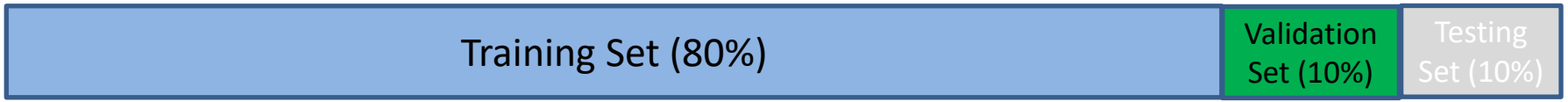
Data pre-processing and visualization phase (Cont'd)

Business Problem → Problem Formulation → Data collection and integration → **Data pre-processing and visualization** → Model training → Model evaluation → Model tuning and feature engineering → Model deployment

Visual analysis to better understand the data

Model training phase



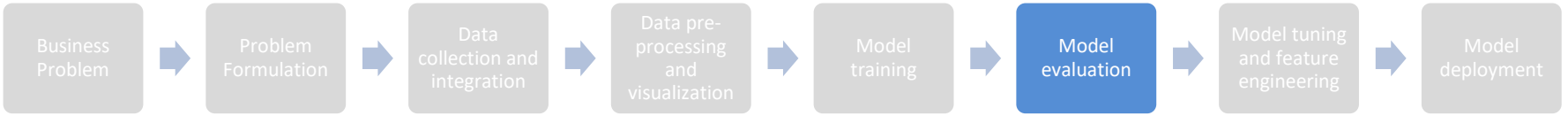


Model Training:

Used 80% of the data to develop (train) the model

Used 10% of the data to improve the model with each training iteration

Model evaluation phase



Model Evaluation:

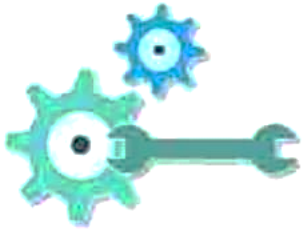
Used 10% of the data to verify that model is performing at or above necessary accuracy

- How often did it route calls correctly on the first try?
- How many times on average did calls have to be rerouted?
- Do these results meet our business needs

Model tuning and feature engineering phase

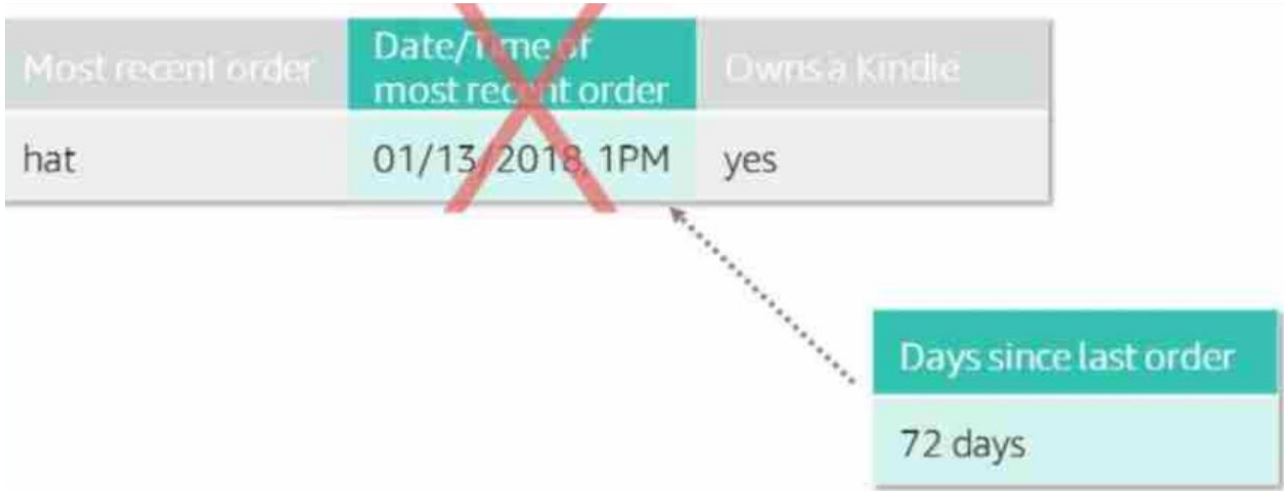
Business Problem → Problem Formulation → Data collection and integration → Data pre-processing and visualization → Model training → Model evaluation → **Model tuning and feature engineering** → Model deployment

After running a training job, we evaluated our model and began a process of iterative tweaks to the model and our data to control how fast or slow our model was learning



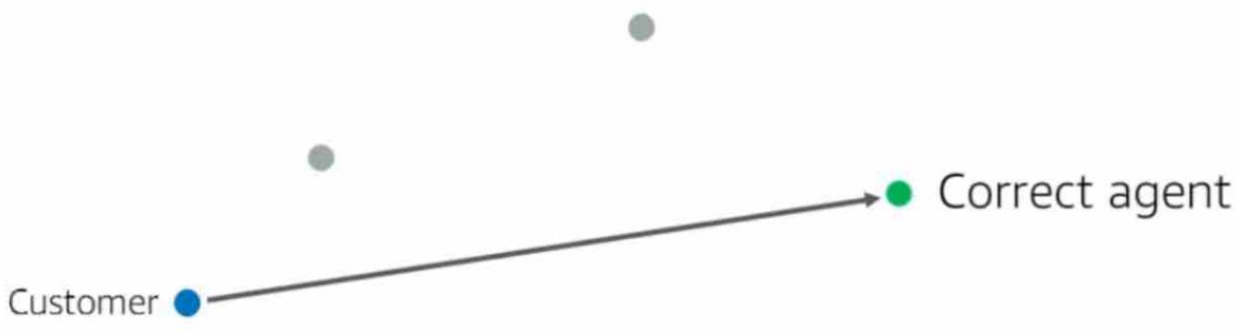
Model tuning and feature engineering phase (Cont'd)

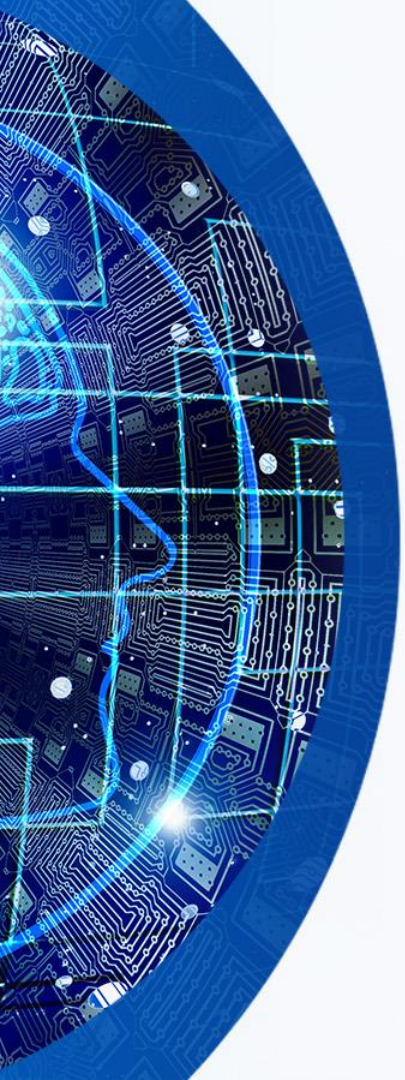
Business Problem → Problem Formulation → Data collection and integration → Data pre-processing and visualization → Model training → Model evaluation → **Model tuning and feature engineering** → Model deployment



Model tuning and feature engineering phase (Cont'd)

Business Problem → Problem Formulation → Data collection and integration → Data pre-processing and visualization → Model training → Model evaluation → Model tuning and feature engineering → **Model deployment**



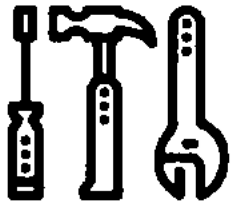


Problem Formulation

- Defining the problem
- Choosing data
- Identifying success
- Summary

Defining the business problem

Example: some products are overstocked and some are understocked, leading to increased overhead costs and missed sales.

Business problem	Goal	Success metric	ML Problem
			
Inaccurate demand prediction is losing the company money	Keep unsold inventory and missed sales low	At the end of each month, have no more than 15% unsold inventory without running out	

What model do we choose?

Use this information to determine the type of machine learning problem you are working with.

Other Problems: Product sales predictions

You want to determine if you should carry a product in stock at all.

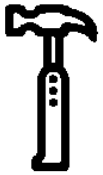
You've decided to rule out any product that will have less than 100 sales

Prediction: **No**

This is a binary classification problem

Example: Product sales predictions

You want to determine the best month to put each product on sales



Prediction: **June**

Prediction: **November**

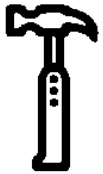
Prediction: **January**

This is a multi-class classification problem

Frame the simplest solution

… but try not to lose important information

Help manage supply and inventory



Demand: 239

> 100 sales ? **Yes**

Simpler, but loses relevant forecasting and sales

Choosing data

Get an understanding of your data

- How much data do you have and where is it?
- Do you have access to that data

Choosing data (Cont'd)

Get a domain expert

- Do you have the data you need to try to address this problem
- Is your data representative?

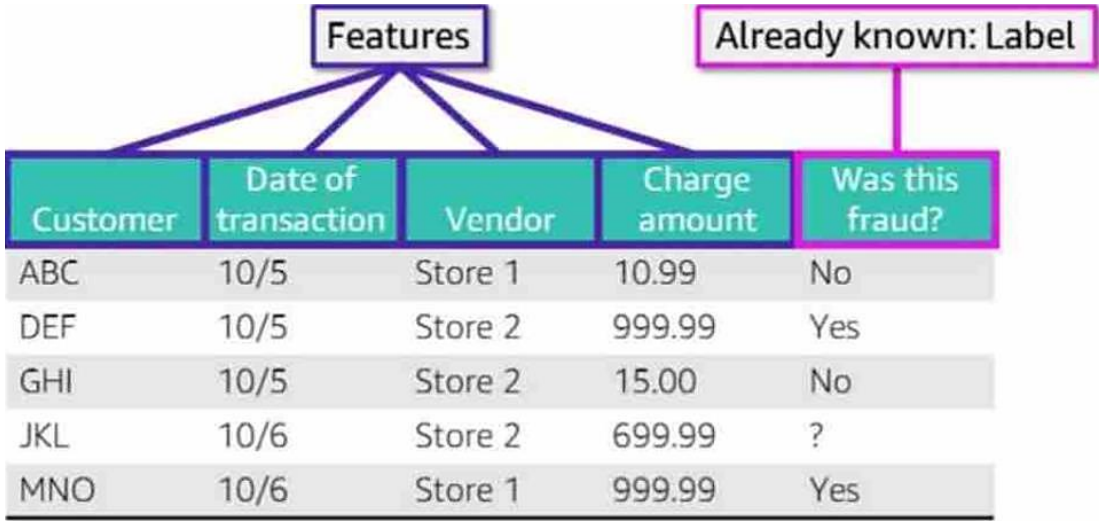
Choosing data (Cont'd)

Evaluate the quality of your data

Product name	Price	Max stock	Current stock	Sales this week
Soap	1.99	20	14	49
Shampoo	6.99	20	2	23
Hair brushes	12.95	30	12	2
Toothpaste	3.50	30	13	40
Toothbrushes	5.00	20	?	?
Lotion	8.75	10	?	?

Choosing data (Cont'd)

Start identifying features and labels you already have

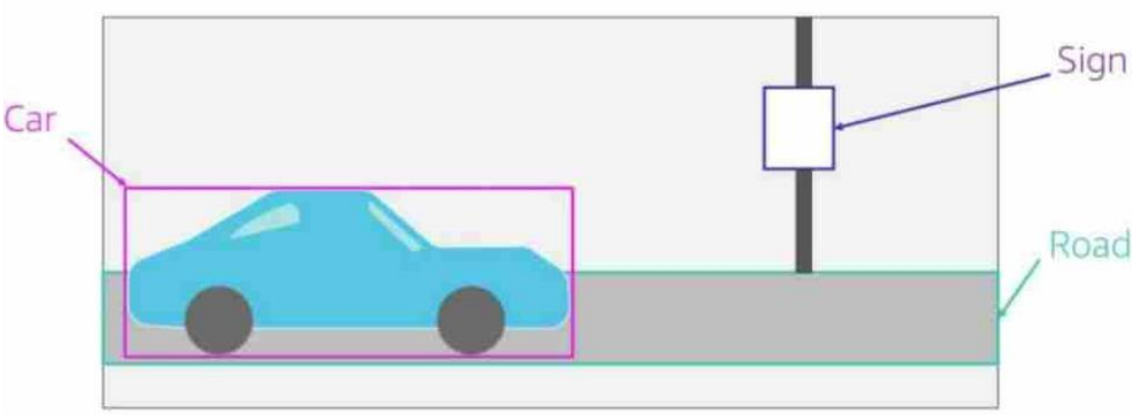


Customer	Date of transaction	Vendor	Charge amount	Was this fraud?
ABC	10/5	Store 1	10.99	No
DEF	10/5	Store 2	999.99	Yes
GHI	10/5	Store 2	15.00	No
JKL	10/6	Store 2	699.99	?
MNO	10/6	Store 1	999.99	Yes

Choosing data (Cont'd)

Do you need a lot of labelled data?

Example: Training data for autonomous driving requires a lot of labels



Ground Truth

You can get help for generating labelled data, e.g. ….

Sends a sample of your data…

… to human who label it and send it…

… back to a service where AI can help with labelling the ground truth

Ground Truth (Cont'd)

The AI sees how the humans labelled the data…

…then uses AI/ML to learn to add label to unseen data where…

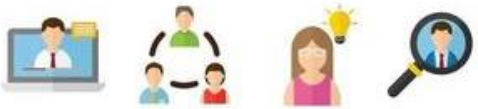
accurate training dataset that's ready to use

Choose the type of workforce

Publicly crowdsourced workforce

You own internal workforce

Pre- approved third- party vendors



Identify Success

How will you know you're doing it right?

Model Performance metrics

- Used during the testing and evaluation sections of ML pipeline
- Typically expressed in terms of accuracy

Business goal metrics

- Used after the model has been deployed
- Measure how well the model is performing in the real world
- Can identify an inappropriate model performance metric

Identify Success (Cont'd)

How will you know you're doing it right?

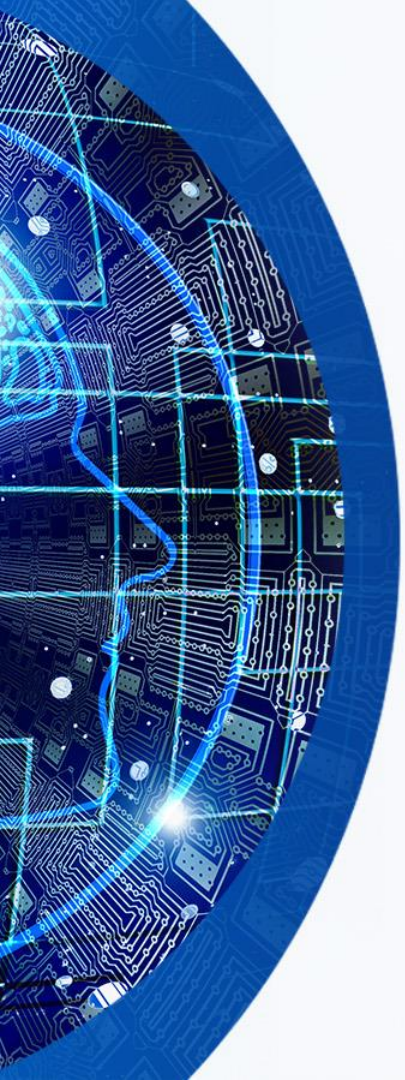
Model Performance metrics

- Example: "The model needs to accurately identify at least 75% of the fraudulent transactions in the test datasets."

Business goal metrics

- Example: "Six month after the model has been deployment, we should have at least 50% fewer customer who cancel their cards due to fraudulent transactions

3.3 Activity: Business Problem and Formulation Exercise



Activity

Choose a project and form teams

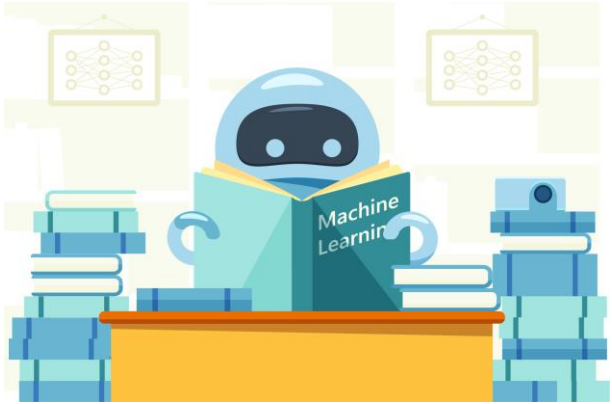
1. Choose the project you would like to work on (refer to the activity sheet).

Activity

Choose a project and form teams

1. Choose the project you would like to work on (refer to the activity sheet).

2. Move to your project designated area of the classroom / breakout rooms



3. Introduce yourself, talk about your background and relevant skills.

A . Break into teams of 2-4 peoples (3 is ideal)

B. Each team should try to have a diverse set of background and skills, to emulate how real world ML teams typically function

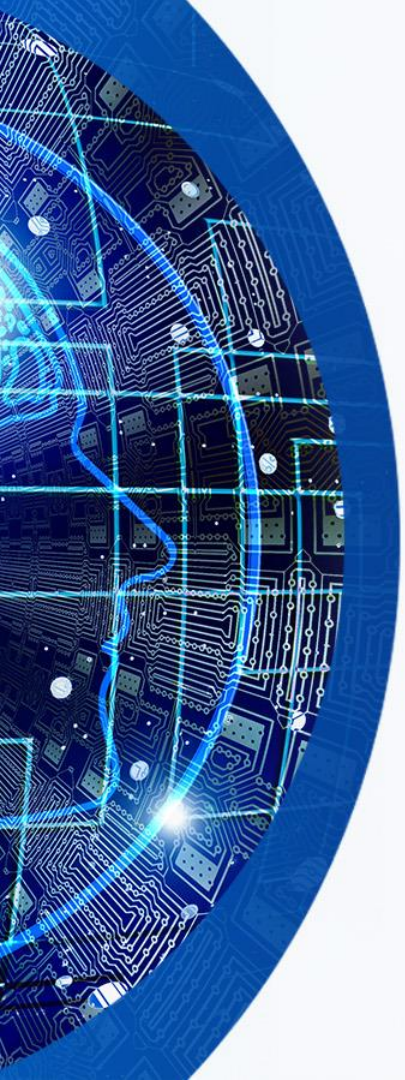
C. Feel free to change project if that makes it easier to form a team.

D. You can work in you own notebooks individually, but consult with each other as a team to develop strategies and troubleshoot problem. Share you expertise with each other, just like you would have to in a real world environment.

Share outs

Periodically, you will asked to shared what you're finding:

- Summarize your finding;
- Talk about any challenges you ran into
- If you'd like, you can use a PowerPoint presentation

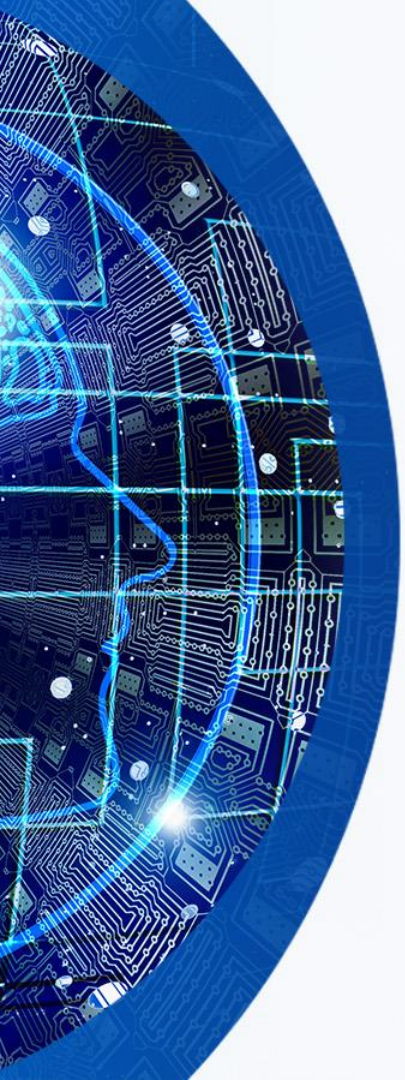


Problem formulation activity

Estimated completion time for each team: 30min

Read through each business scenario and:

1. Determine if and why ML is an appropriate solution to deploy
2. Formulate the business problem, success metrics, and desired ML output
3. Identify the type of ML problem you're dealing with
4. Analyze the appropriateness of data you're working with



Group Sharing

Estimated time for all groups (depending on number of groups): 45 min

Share with the class

- What is the proposed solution from your group?
- Did your group come to a different conclusion? Why?
- What kinds of data would you want to have access to in order to best address the problem?

End of Chapter 3

Q&A