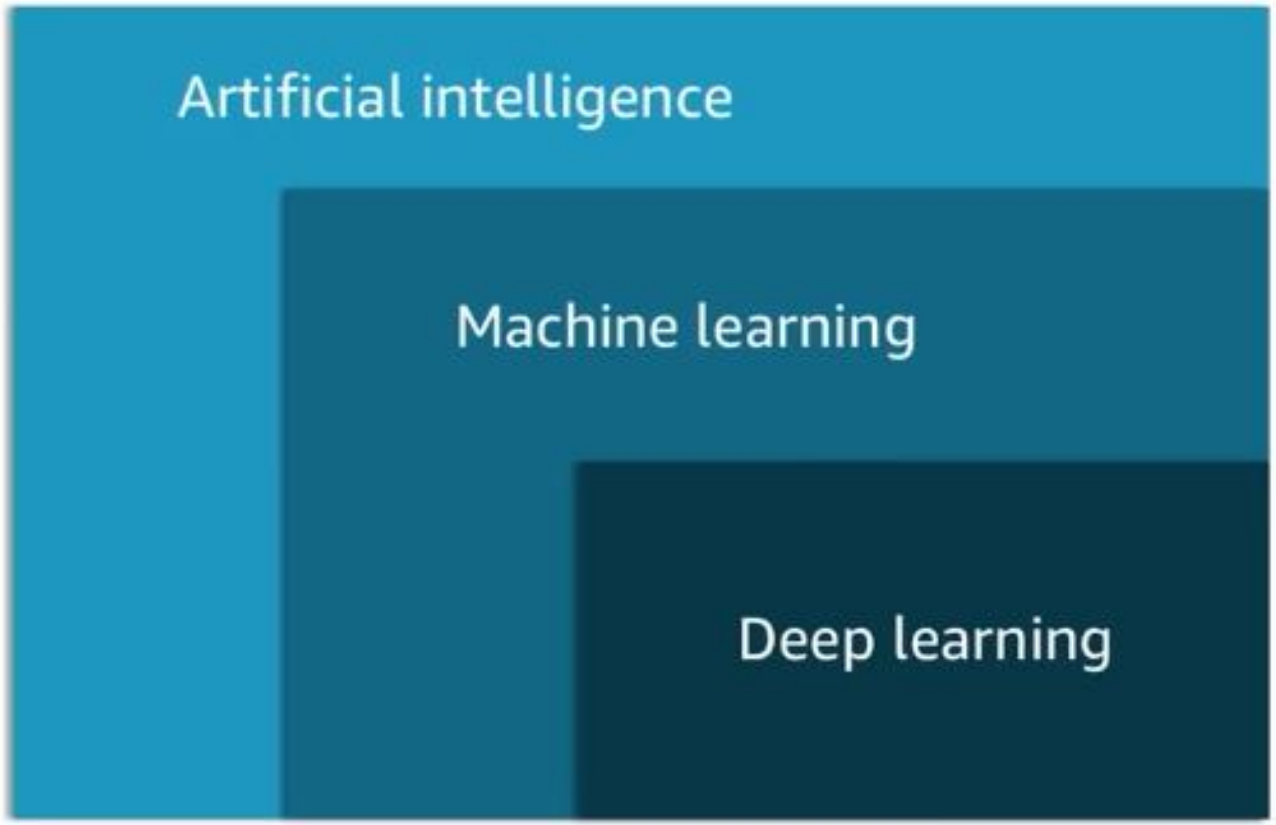
Module 2: Introduction to Machine Learning

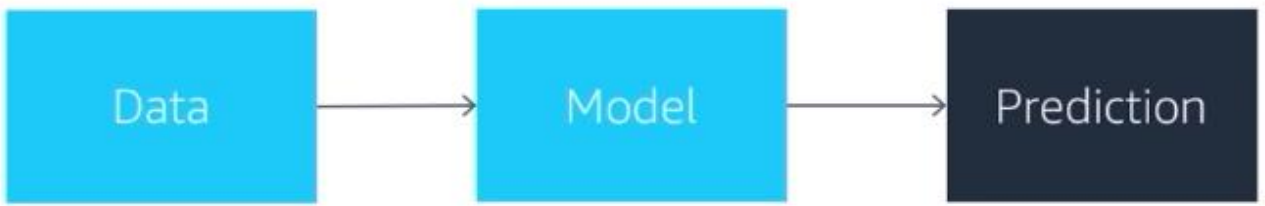
Section 1: What is machine learning ?

Artificial intelligence, machine learning, and deep learning



Machine learning

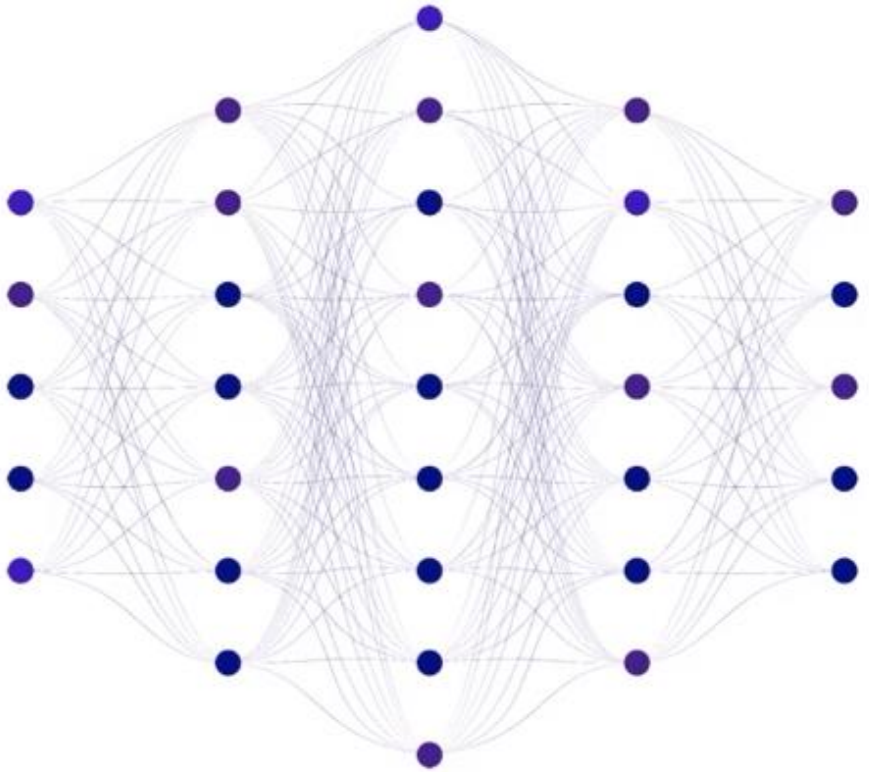
Machine learning is the scientific study of algorithms and statistical models to perform a task using inference rather than instructions.



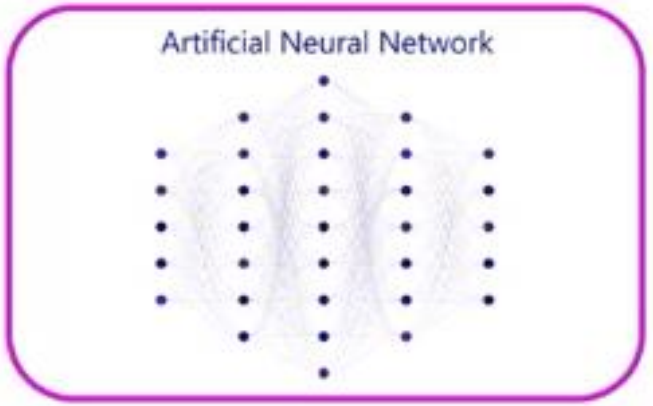
Machine learning flow

Deep learning

Artificial Neural Network



ML and technology advancements



Traditional computing

Cloud computing and Moore's law

Modern machine learning

Section 1 key takeaways

- Artificial intelligence
  - Machines performing human tasks
- Machine learning
  - Training models to make predictions
- Deep learning
  - Neural Networks
- Technology and economic advancements have made machine learning more accessible to individuals and organizations

Module 2: Introduction to Machine Learning

Section 2: Business problems solved with machine learning

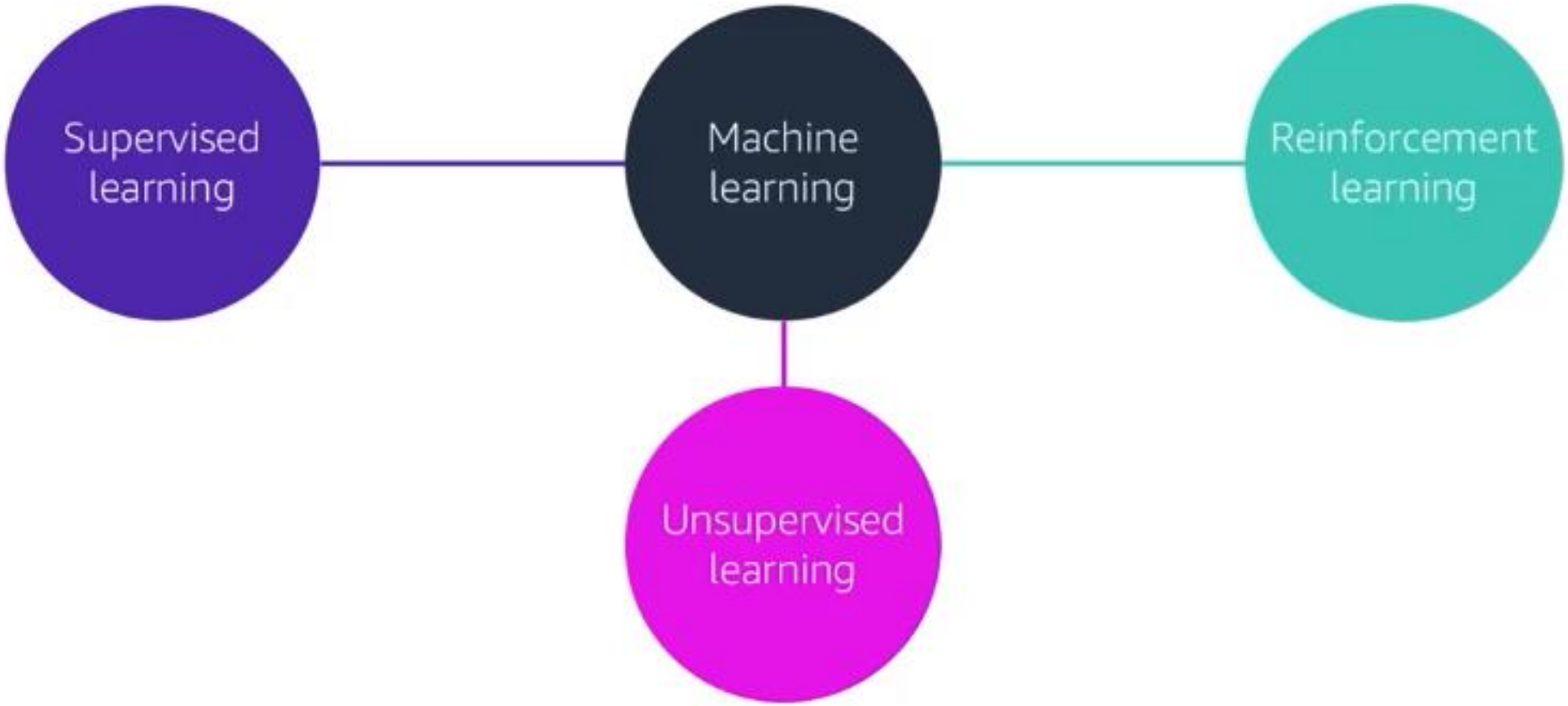
Common business use cases

Spam versus regular email

Recommendations

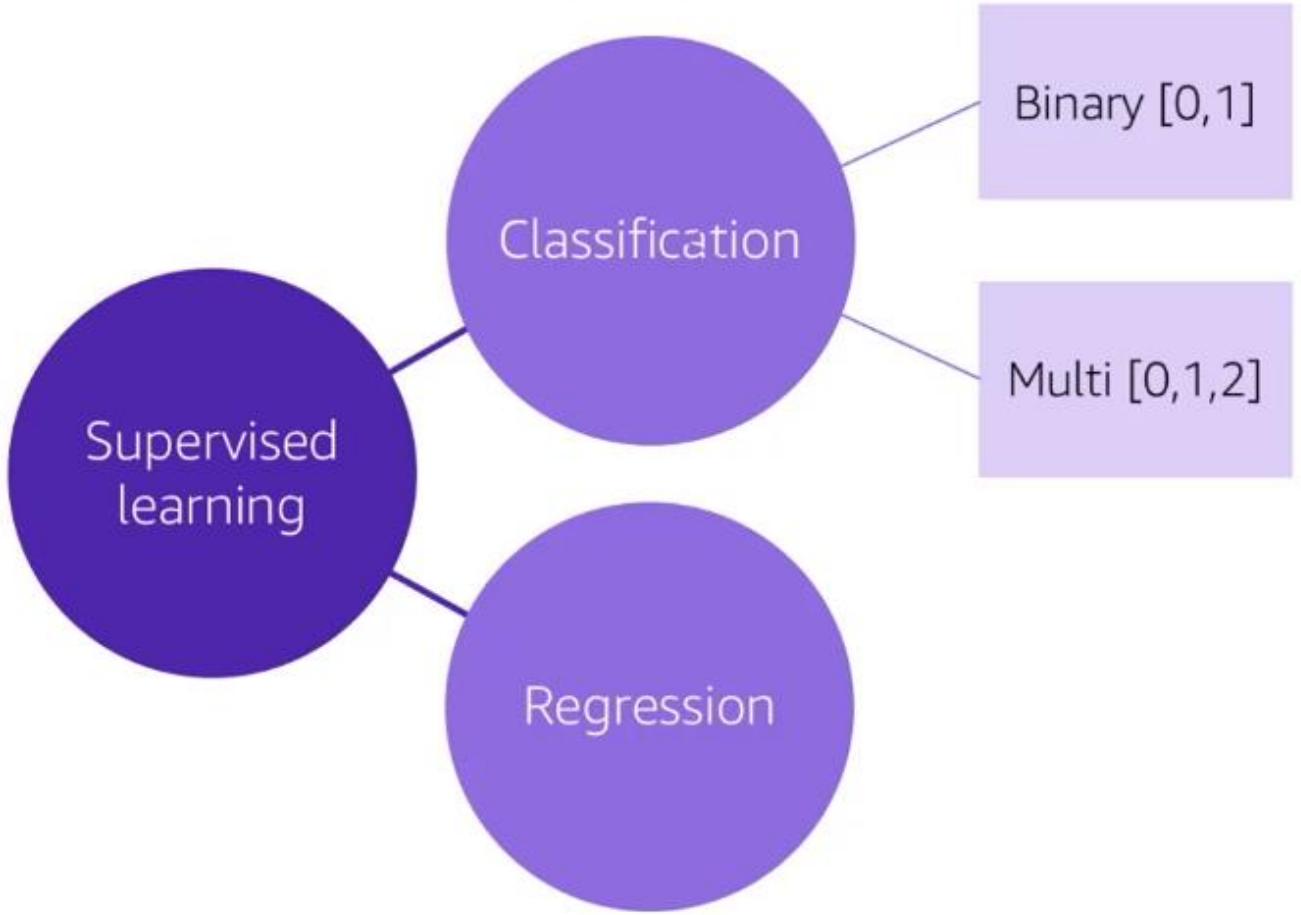
Fraud

Types of machine learning



Supervised learning

Learn by identifying patterns in data that is **already labeled**.

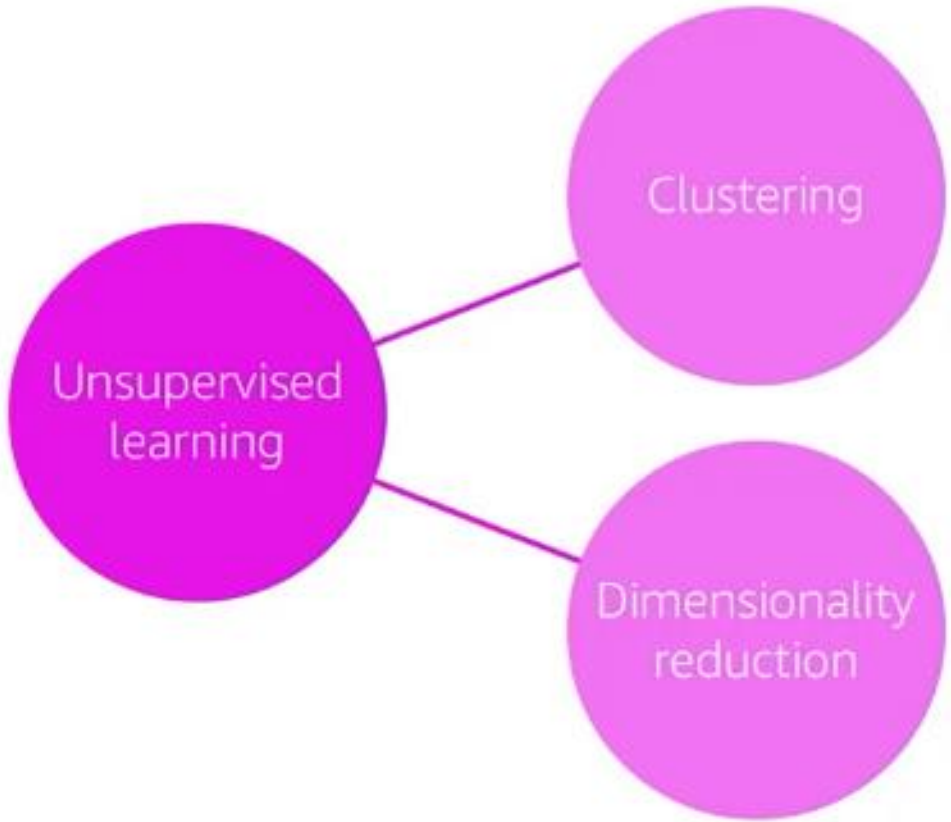


- Fraud detection
- Image recognition
- Customer retention
- Medical diagnostics
- Personalized advertising

- Product sales prediction
- Weather forecasting
- Market forecasting
- Population growth prediction

Unsupervised learning

The machine must uncover and create the labels itself.

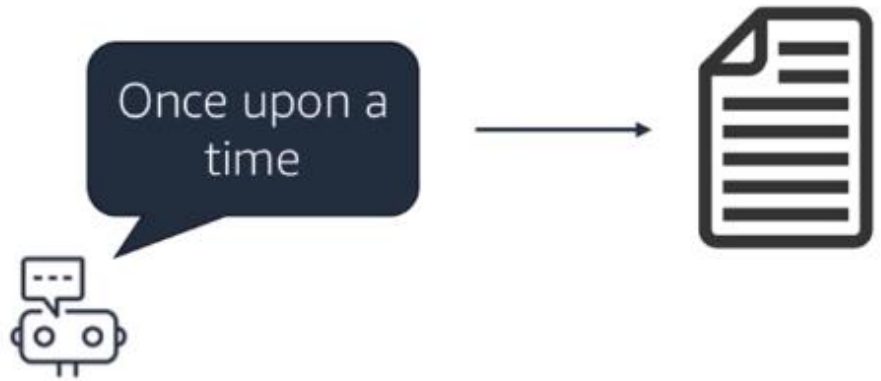


- Product recommendations
- Customer segmentation
- Targeted marketing
- Medical diagnostics

- Visualization
- Natural language processing
- Data structure discovery
- Gene sequencing

Natural language processing

gögn eru lykilatriði ⇌ los datos son clave



Poor story. Little character development. Jumps between scenes like you might get caught stealing. Unexplained bad guys appear with the thinnest of back story. Back to unlimited resources and lets not talk about the mechanics of building something so huge in such an inhospitable place...

Reinforcement learning

aws academy

Learning through **trial and error**

Reinforcement learning

- Game AI
- Self-driving cars
- Robotics
- Customer service routing

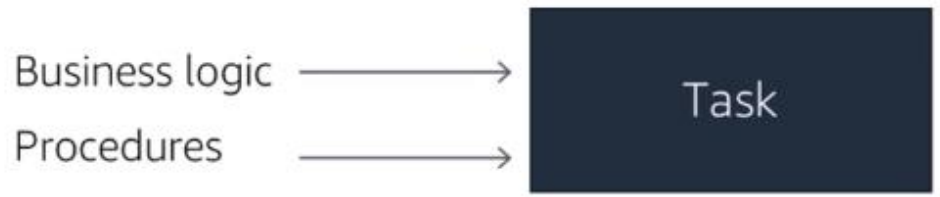
AWS DeepRacer

Best when the desired outcome is known but the exact path to achieving it is not known

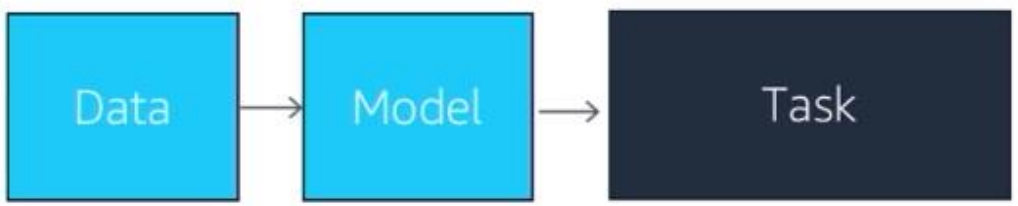
When to use machine learning?

aws academy

Classical programming approach



Machine learning approach



Use machine learning when you have:

- ✓ Large datasets, large number of variables
- ✓ Lack of clear procedures to obtain the solution
- ✓ Existing machine learning expertise
- ✓ Infrastructure already in place to support ML
- ✓ Management support for ML

Section # key takeaways

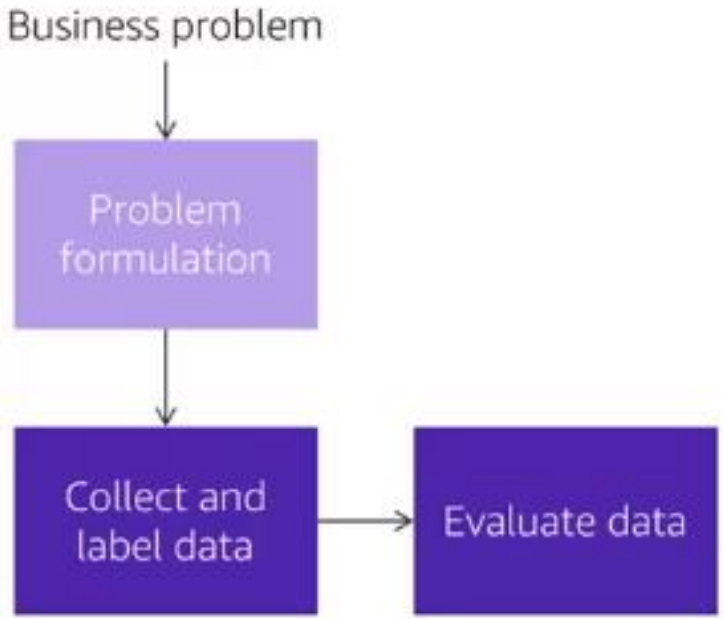
aws academy

- Machine learning applications affect everyday life
- Machine learning can be grouped into –
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Most problems are supervised learning

Module 2: Introduction to Machine Learning

Section 3: Machine learning process

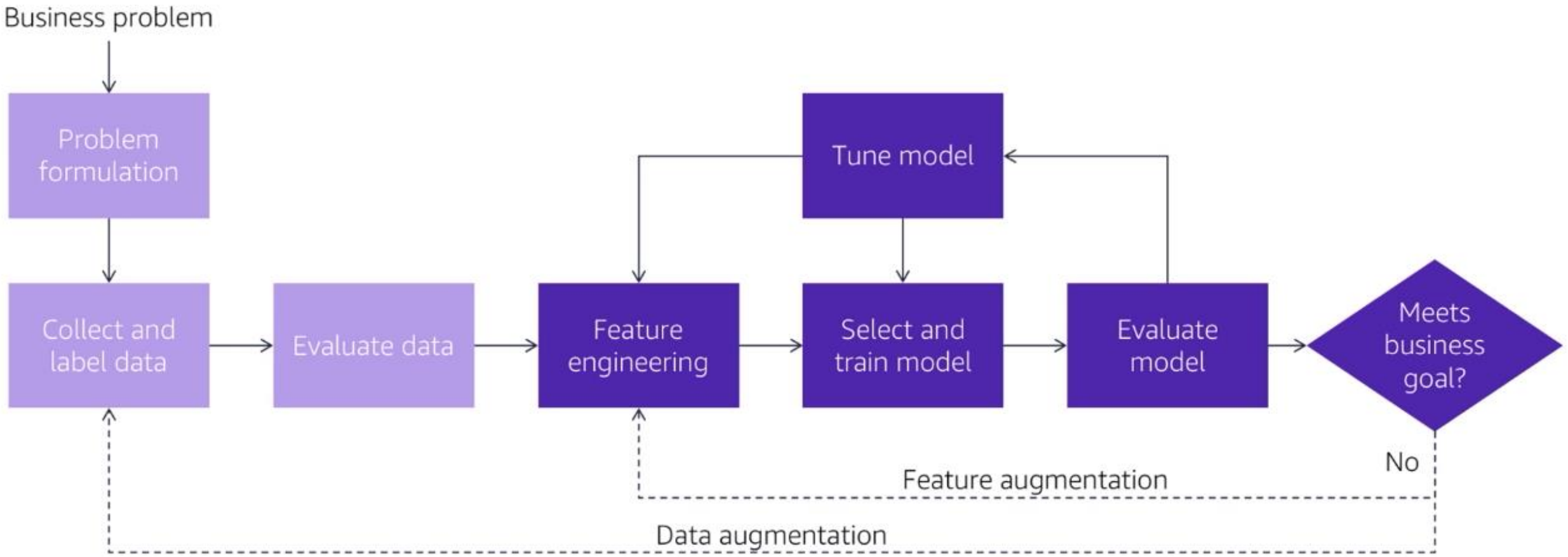
ML pipeline: Data preparation



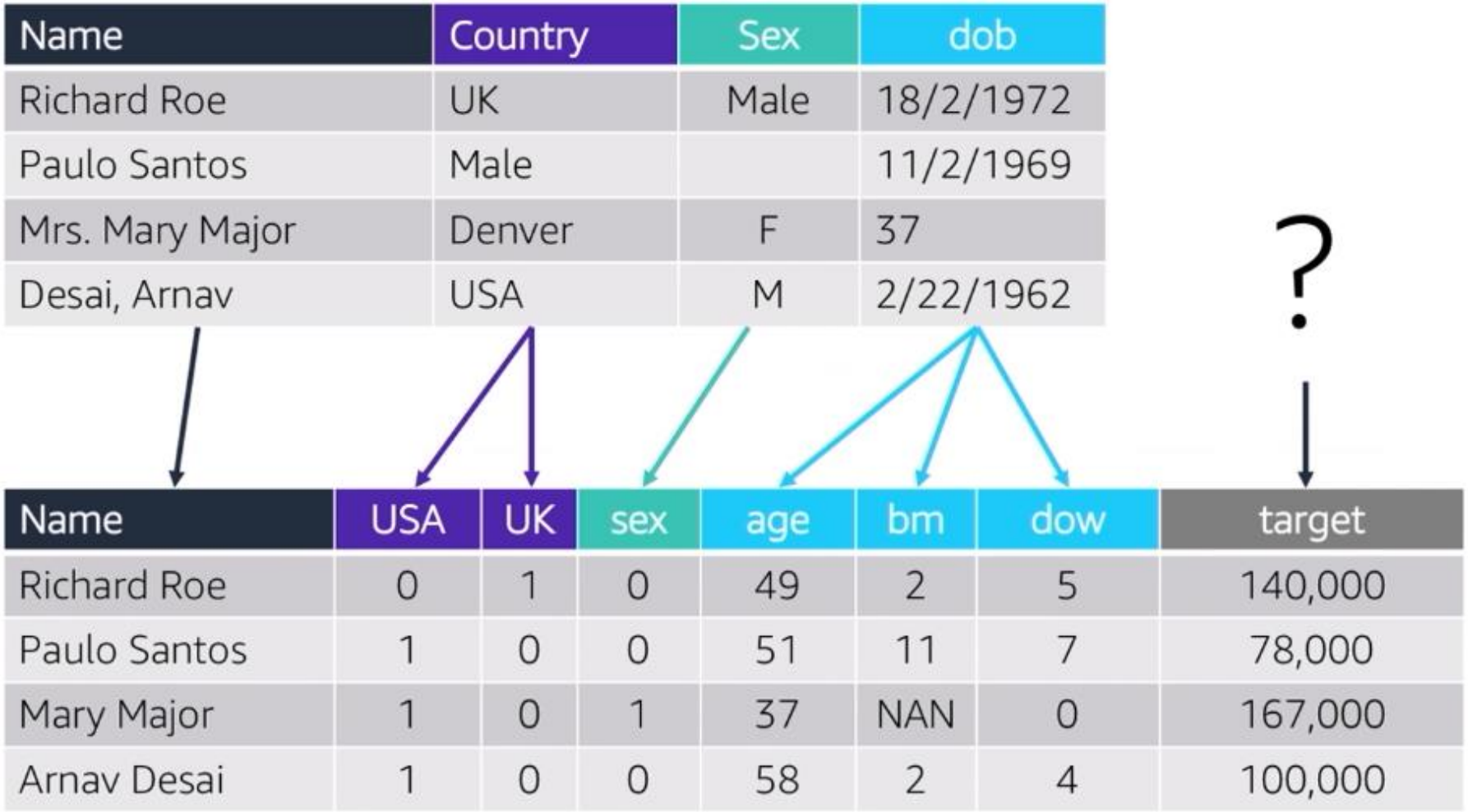
Data handling and cleaning

Name	Country	Sex	dob
Richard Roe	UK	Male	18/2/1972
Paulo Santos	Male		11/2/1969
Mrs. Mary Major	Denver	F	37
Desai, Arnav	USA	M	2/22/1962

ML pipeline: Iterative model training



ML pipeline: Feature engineering

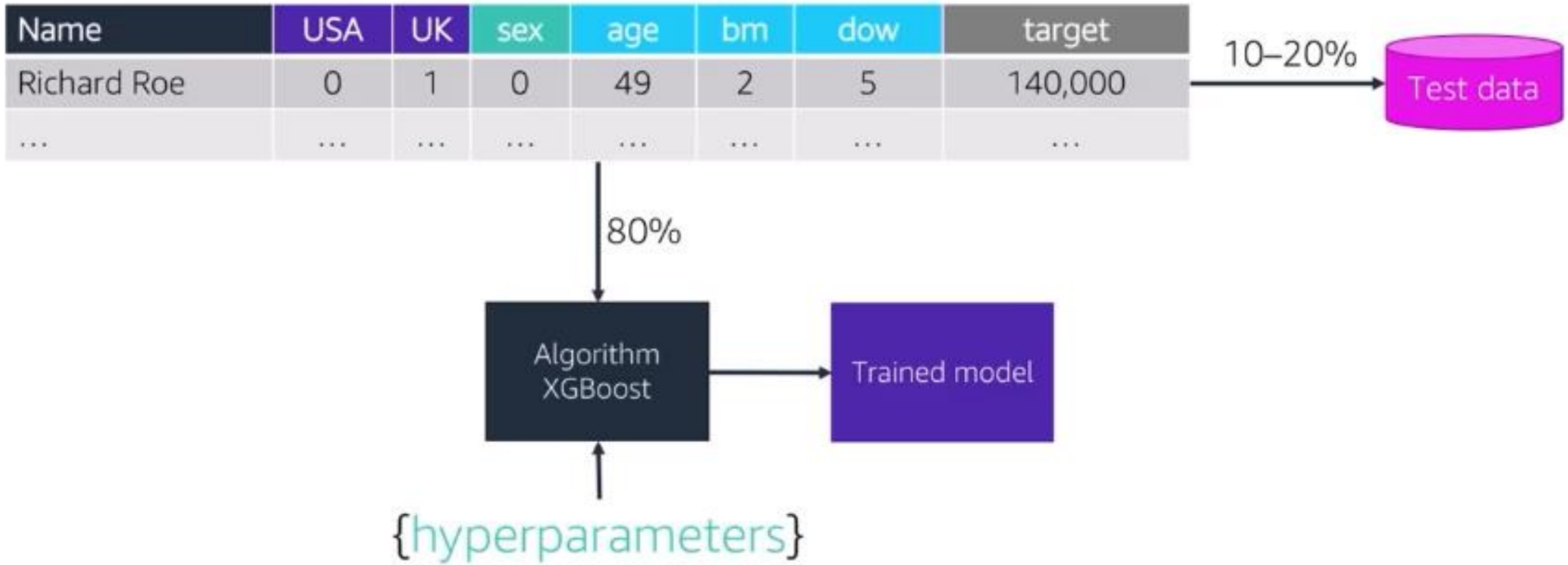


Name	Country	Sex	dob
Richard Roe	UK	Male	18/2/1972
Paulo Santos	Male		11/2/1969
Mrs. Mary Major	Denver	F	37
Desai, Arnav	USA	M	2/22/1962

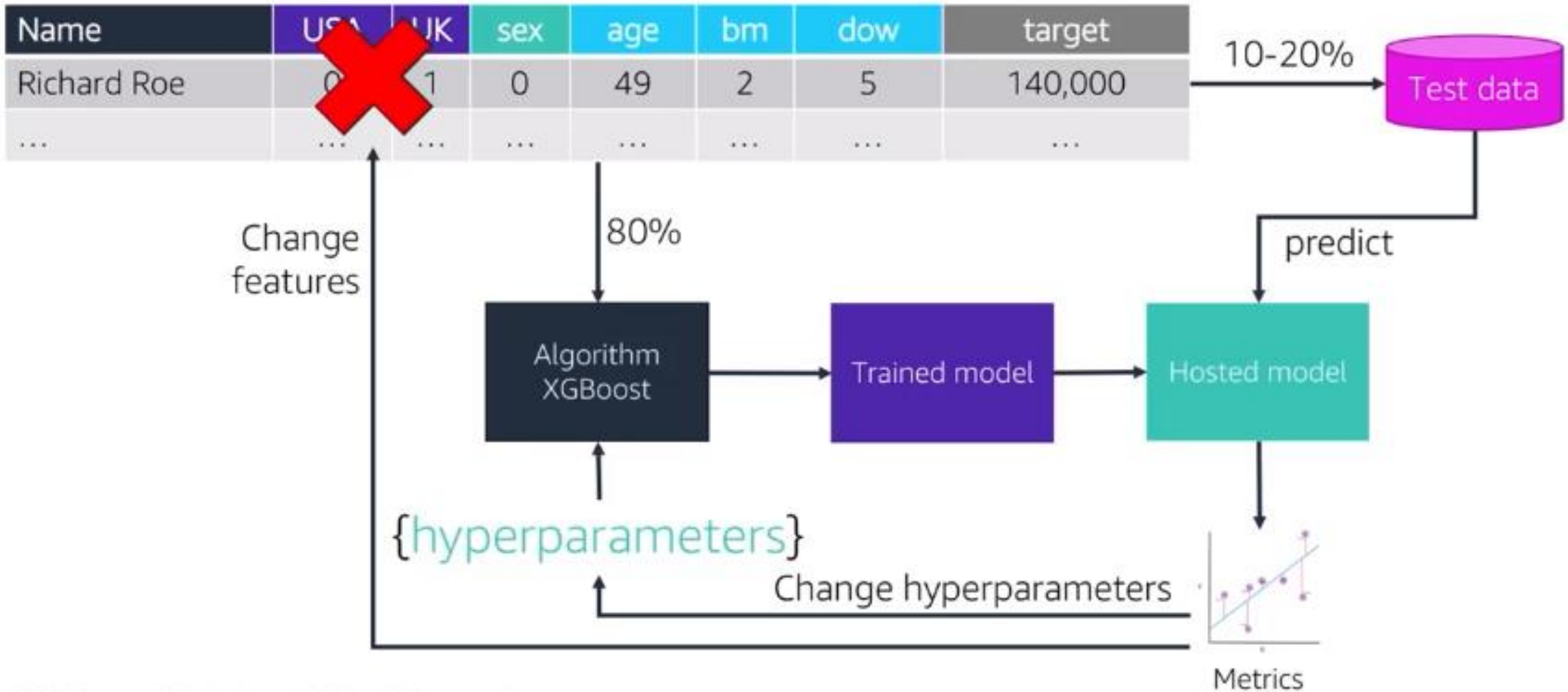
?

Name	USA	UK	sex	age	bm	dow	target
Richard Roe	0	1	0	49	2	5	140,000
Paulo Santos	1	0	0	51	11	7	78,000
Mary Major	1	0	1	37	NAN	0	167,000
Arnav Desai	1	0	0	58	2	4	100,000

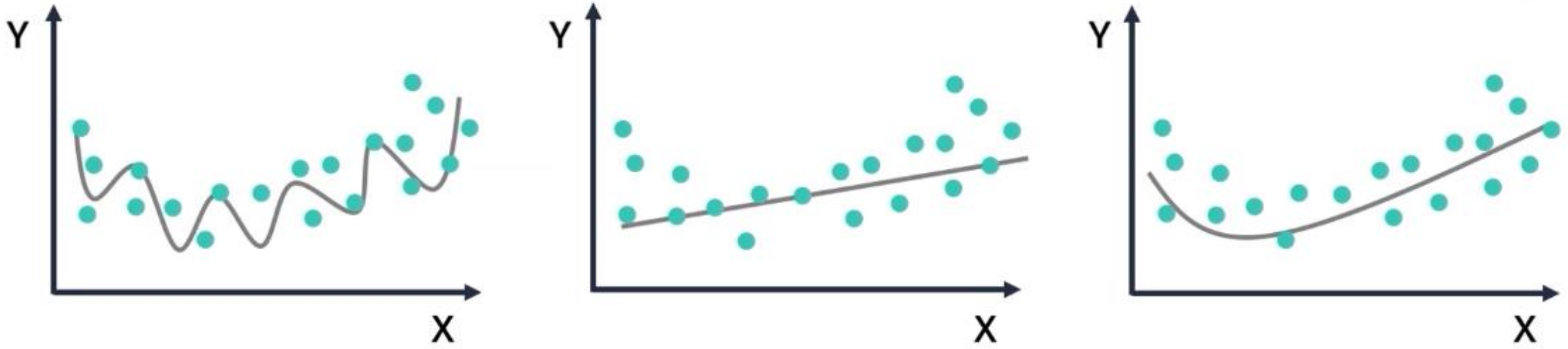
ML pipeline: Model training



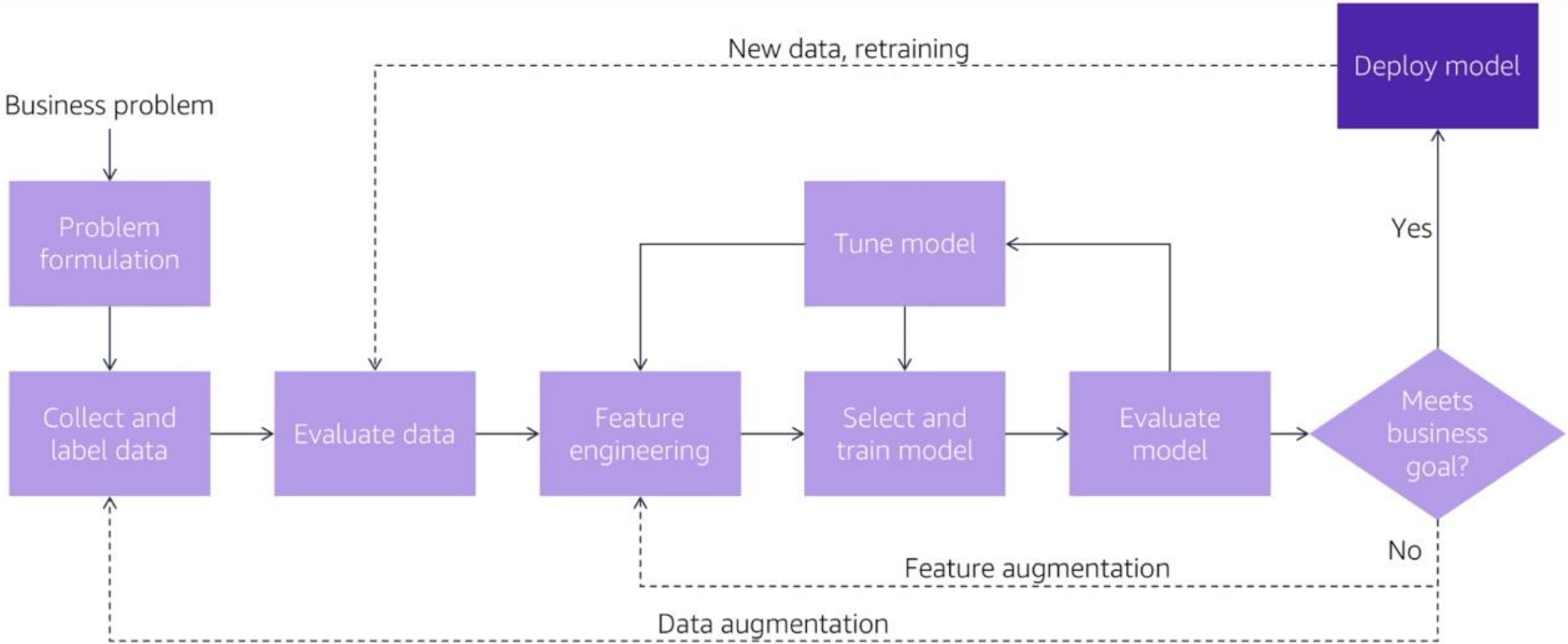
ML pipeline: Evaluating and tuning the model



Overfitting and underfitting



ML pipeline: Deployment



Machine learning challenges

Data

- Poor quality
- Non-representative
- Insufficient
- Overfitting and underfitting

Users

- Lack of data science expertise
- Cost of staffing with data scientists
- Lack of management support

Business

- Complexity in formulating questions
- Explaining models to the business
- Cost of building systems

Technology

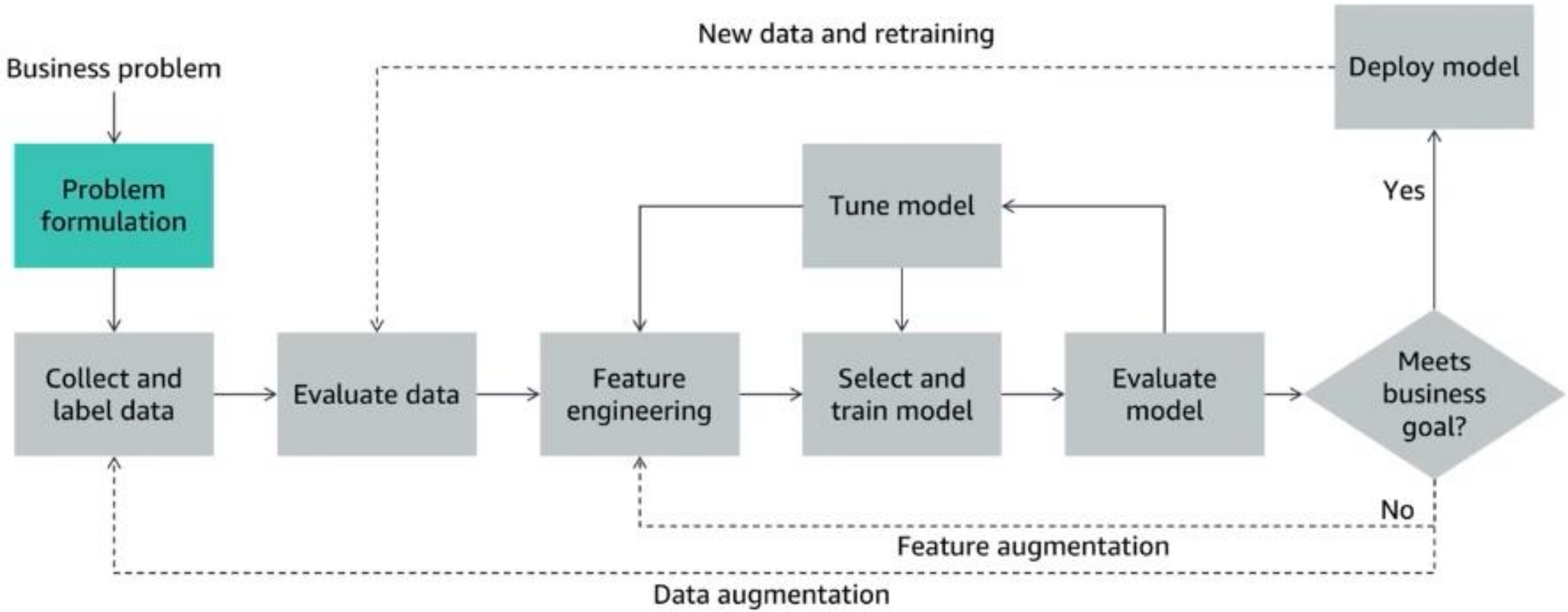
- Data privacy issues
- Tool selection can be complicated
- Integration with other systems

Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 1: Formulating machine learning problems

aws academy

Machine learning pipeline



Define business objective

What is the business goal?

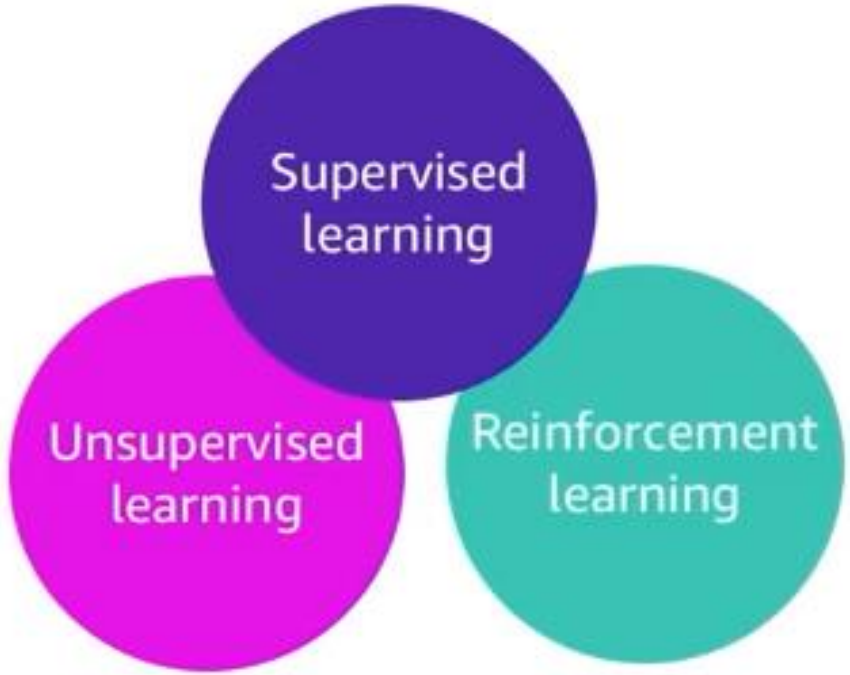
Questions to ask:

- How is this task done today?
- How will the business measure success?
- How will the solution be used?
- Do similar solutions exist, which you might learn from?
- What assumptions have been made?
- Who are the domain experts?

How should you frame this problem?

aws academy

- Is the problem a machine learning problem?
- Is the problem supervised or unsupervised?
- What is the target to predict?
- Do you have access to the data?
- What is the minimum performance?
- How would you solve this problem manually?
- What's the simplest solution?



Example: Problem formulation

You want to identify fraudulent credit card transactions so that you can stop the transaction before it processes.

Why?

Reduce the number of customers who end their membership because of fraud.

10%

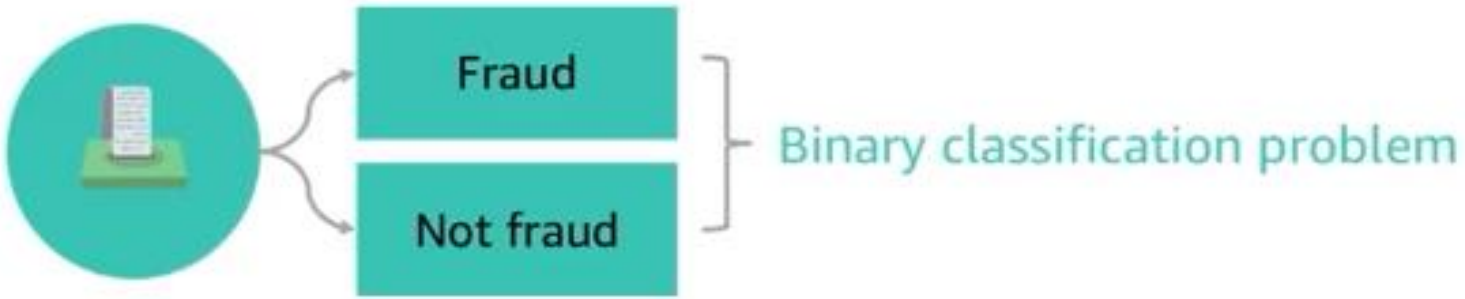
reduction in fraud claims in retail

Can you measure it?

Move from qualitative statements to quantitative statements that can be measured.

Make it an ML model

Credit card transaction is either ***fraudulent*** or ***not fraudulent***.



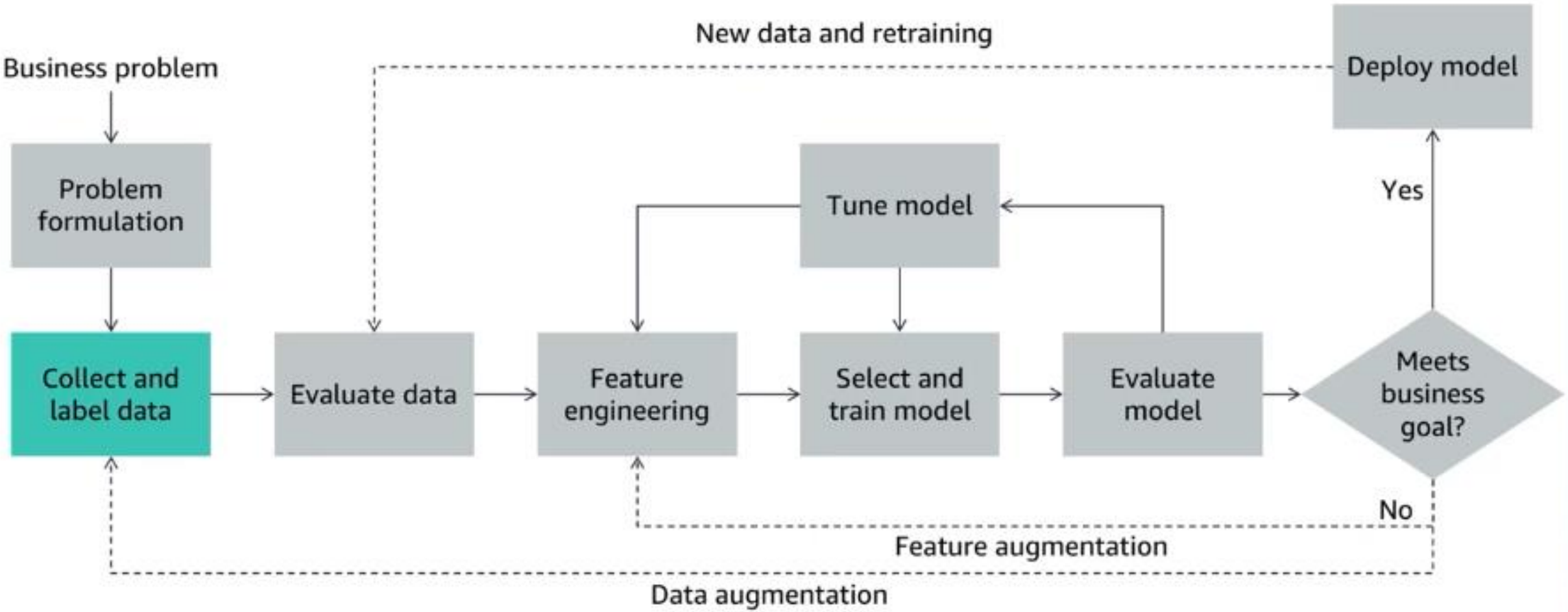
Use historical data of fraud reports to help define your model.

Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 2: Collecting and securing data

aws academy

Machine learning pipeline



What data do you need?

- How much data do you have, and where is it?
- Do you have access to that data?
- What solution can you use to bring all of this data into one centralized repository?

Data sources

- **Private data**: Data that customers create
- **Commercial data**: AWS Data Exchange, AWS Marketplace, and other external providers
- **Open-source data**: Data that is publicly available (check for limits on usage)
  - Kaggle
  - World Health Organization
  - U.S. Census Bureau
  - National Oceanic and Atmospheric Administration (U.S.)
  - UC Irvine Machine Learning Repository
  - AWS

Observations

ML problems need a lot of data—also called observations—where the target answer or prediction is already known.

Customer	Date of transaction	Vendor	Charge amount	Was this fraud?
ABC	10/5	Store 1	10.99	No
DEF	10/5	Store 2	99.99	Yes
GHI	10/5	Store 2	15.00	No
JKL	10/6	Store 2	99.99	?
MNO	10/6	Store 1	99.99	Yes

Feature

Target

Get a domain expert

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

Storing data in AWS

aws academy

Amazon Simple Storage Service (Amazon S3)

Amazon FSx

Amazon Elastic File System (Amazon EFS)

Amazon Relational Database Service (Amazon RDS)

Amazon Redshift

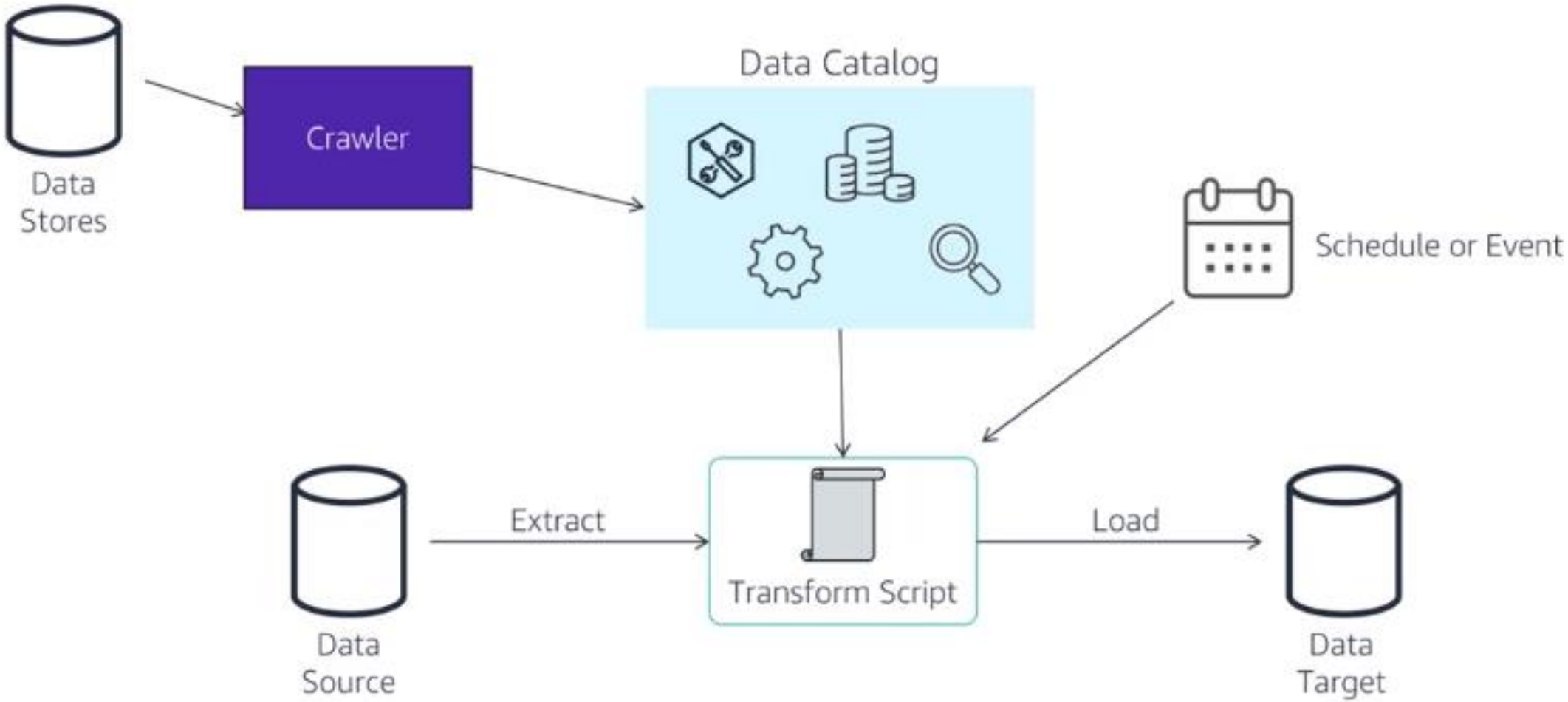
Amazon Timestream

Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 2a: Extracting, transforming and loading data

aws academy

Extract, transform, load (ETL)



ETL with AWS Glue

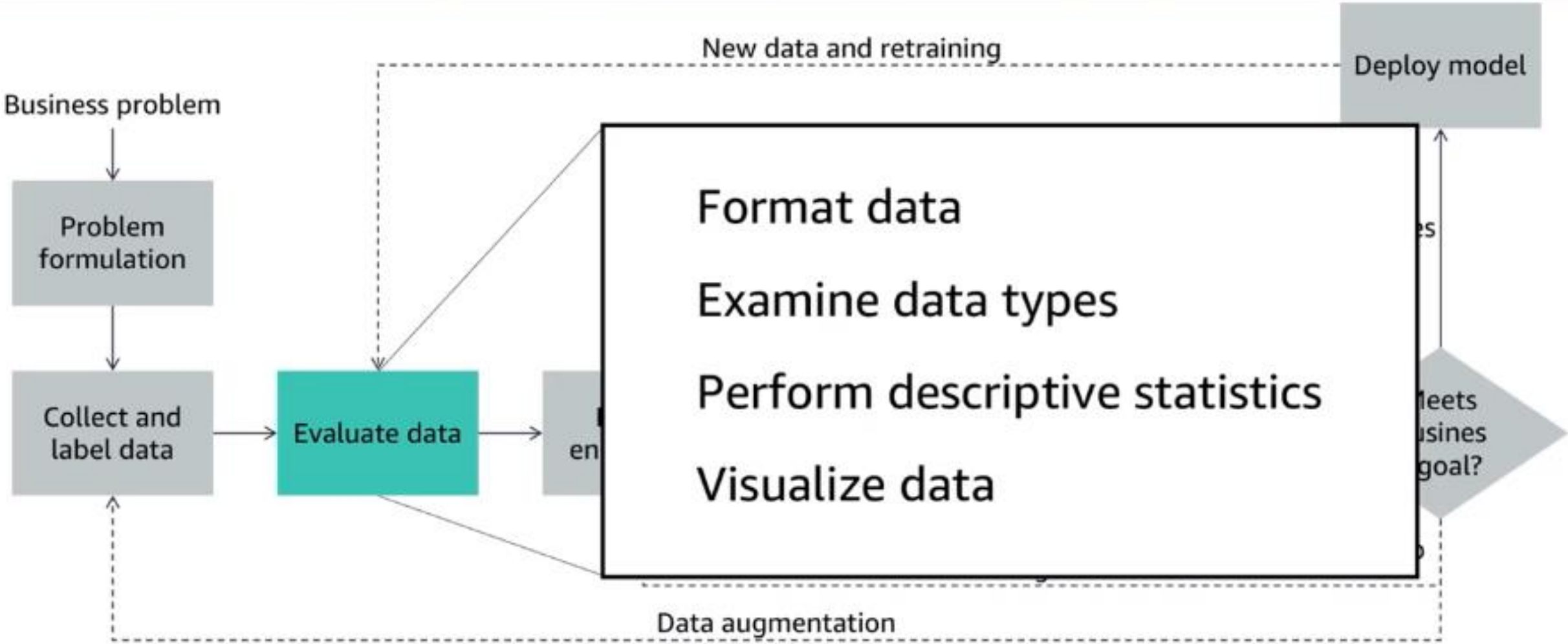
- Runs the ETL process
- Crawls data sources to create catalogs that other systems can query
- ML functionality

AWS Glue can *glue together* different datasets and emit a single endpoint that can queried.

Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 3: Evaluating your data

Machine learning pipeline



You must understand your data

Before you can run statistics on your data, you must ensure that it's in the right format for analysis.

Customer	Date of Transaction	Vendor	Charge Amount	Was This Fraud?
ABC	10/5	Store 1	10.99	No
DEF	10/5	Store 2	99.99	Yes
GHI	10/5	Store 2	15.00	No
JKL	10/6	Store 2	99.99	?
MNO	10/6	Store 1	99.99	Yes

Loading data into pandas

- Reformats data into tabular representation (DataFrame)
- Converts common formats like comma-separated values (CSV), JavaScript Object Notation (JSON), Excel, Pickle, and others

```
import pandas as pd
url = "https://somewhere.com/winequality-red.csv"
df_wine = pd.read_csv(url,';')
```

pandas DataFrame

aws academy

Number of instances

```
df_wine.shape
```

```
(1599, 12)
```

Number of attributes

```
df_wine.head(5)
```

Columns/Attributes

Rows/Instances

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 3a: Describing your data

aws academy

Descriptive statistics

aws academy

Use descriptive statistics to gain insights into your data before you clean the data:

Overall statistics

Multivariate statistics

Attribute statistics

Statistical characteristics

aws academy

```
df_wine.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	pH	sulphates	alcohol	quality
count	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00
mean	8.32	0.53	0.27	2.54	0.09	15.87	46.47	3.31	0.66	10.42	5.64
std	1.74	0.18	0.19	1.41	0.05	10.46	32.90	0.15	0.17	1.07	0.81
min	4.60	0.12	0.00	0.90	0.01	1.00	6.00	2.74	0.33	8.40	3.00
25%	7.10	0.39	0.09	1.90	0.07	7.00	22.00	3.21	0.55	9.50	5.00
50%	7.90	0.52	0.26	2.20	0.08	14.00	38.00	3.31	0.62	10.20	6.00
75%	9.20	0.64	0.42	2.60	0.09	21.00	62.00	3.40	0.73	11.10	6.00
max	15.90	1.58	1.00	15.50	0.61	72.00	289.00	4.01	2.00	14.90	8.00

Categorical statistics identify frequency of values and class imbalances

aws academy

```
df_car.head(5)
```

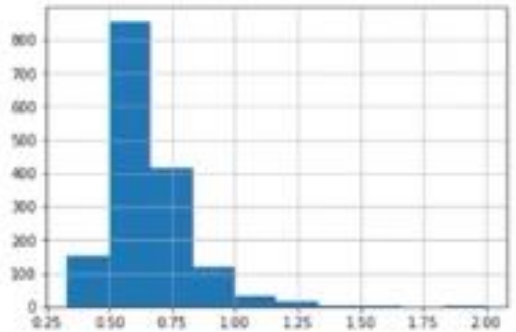
	buying	maint	doors	persons	lug_boot	safety	class
0	vhigh	vhigh	2	2	small	low	unacc
1	vhigh	vhigh	2	2	small	med	unacc
2	vhigh	vhigh	2	2	small	high	unacc
3	vhigh	vhigh	2	2	med	low	unacc
4	vhigh	vhigh	2	2	med	med	unacc

```
df_car.describe()
```

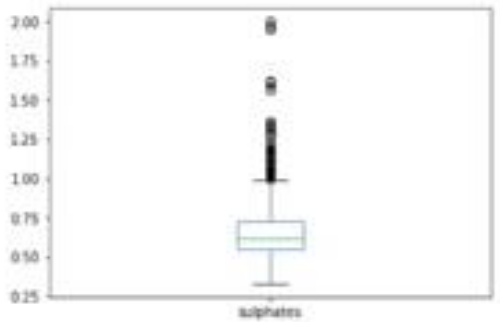
	buying	maint	doors	persons	lug_boot	safety	class
count	1728	1728	1728	1728	1728	1728	1728
unique	4	4	4	3	3	3	4
top	low	low	2	2	big	low	unacc
freq	432	432	432	576	576	576	1210

Plotting attribute statistics

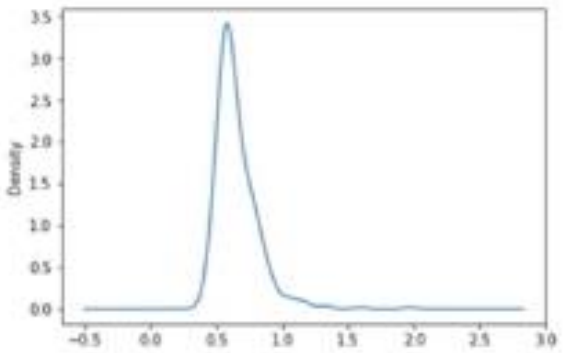
```
df_wine['sulphates'].hist(bins=10)
```




```
df_wine['sulphates'].plot.box()
```

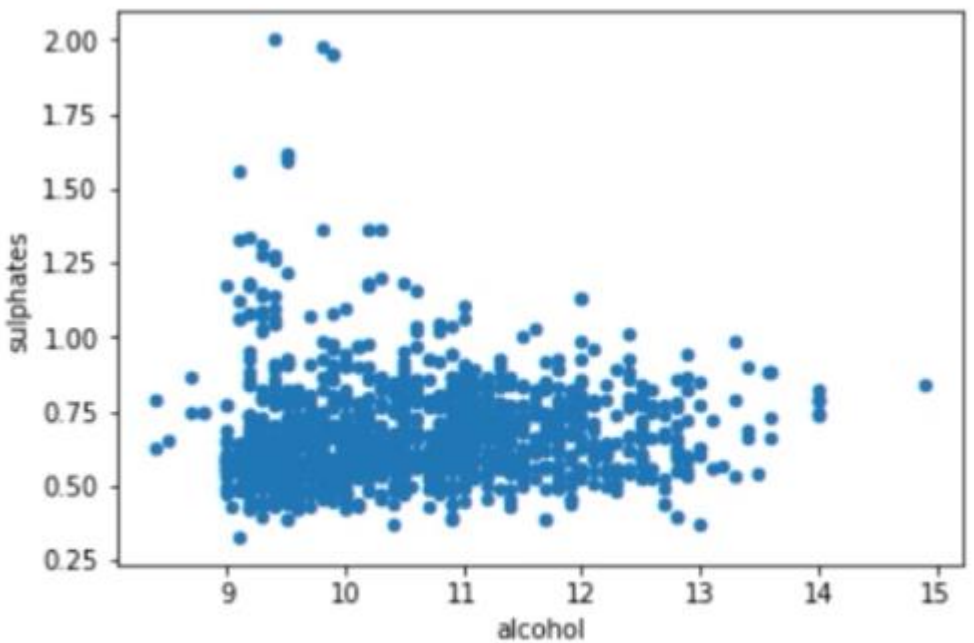



```
df_wine['sulphates'].plot.kde()
```

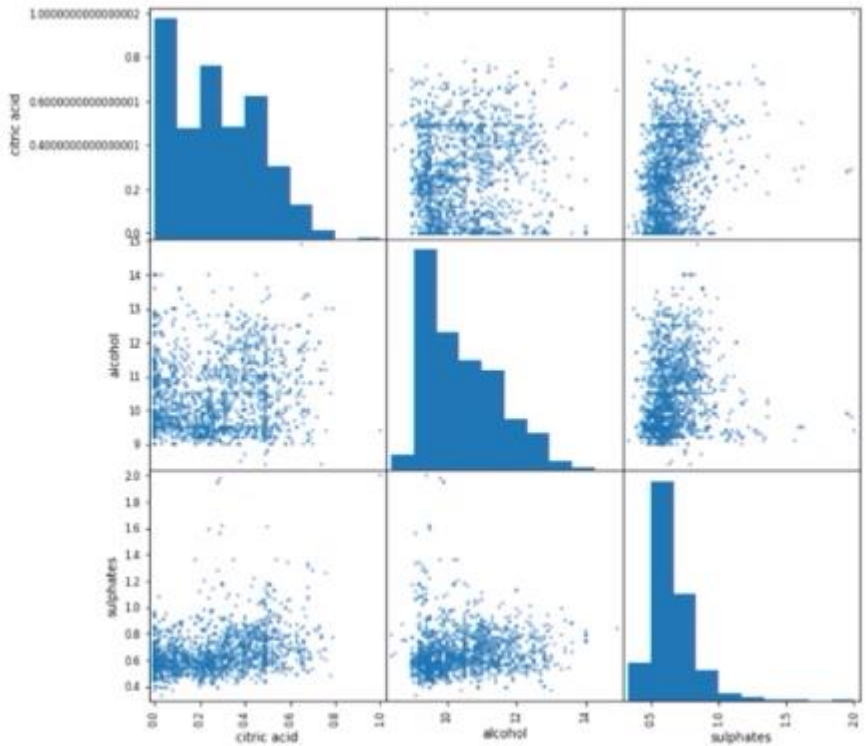


Plotting multivariate statistics

```
df_wine.plot.scatter(
    x='alcohol',
    y='sulphates')
```




```
pd.plotting.scatter_matrix(
    df_wine[['citric acid',
        'alcohol',
        'sulphates']])
```



Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 3b: Finding correlations

Correlation matrix

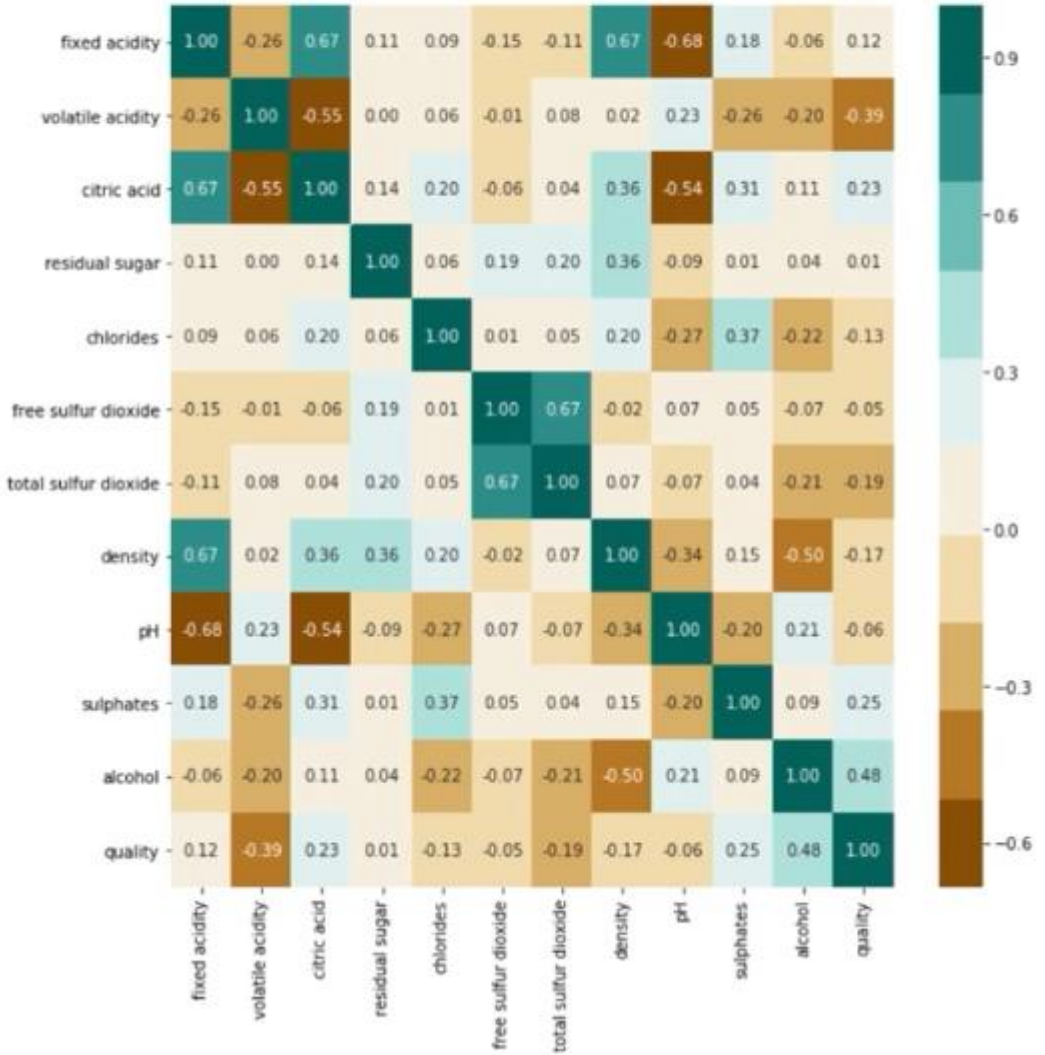
```
corr_matrix = df_wine.corr()
corr_matrix["quality"].sort_values(ascending=False)
```

```
quality                1.000000
alcohol                0.476166
sulphates              0.251397
citric acid            0.226373
fixed acidity          0.124052
residual sugar         0.013732
free sulfur dioxide   -0.050656
pH                    -0.057731
chlorides             -0.128907
density               -0.174919
total sulfur dioxide  -0.185100
volatile acidity      -0.390558
Name: quality, dtype: float64
```

Do alcohol and sulphates correlate to wine quality?

Correlation matrix heat map

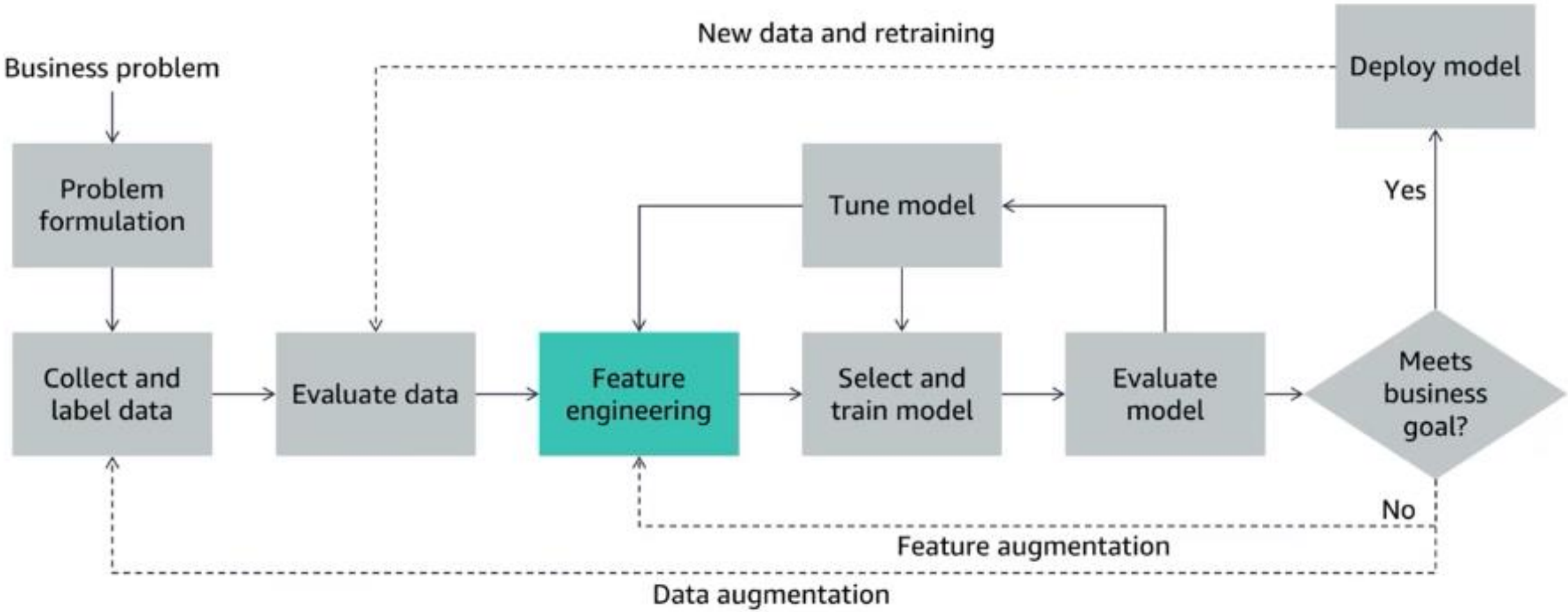
```
import seaborn as sns

correlations = df_wine.corr()
fig, ax = plt.subplots(figsize=(10, 10))

colormap =
    sns.color_palette("BrBG", 10)

sns.heatmap(correlations,
    cmap=colormap,
    annot=True,
    fmt=".2f")

ax.set_yticklabels(colum_names);

plt.show()
```



Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 4: Feature engineering

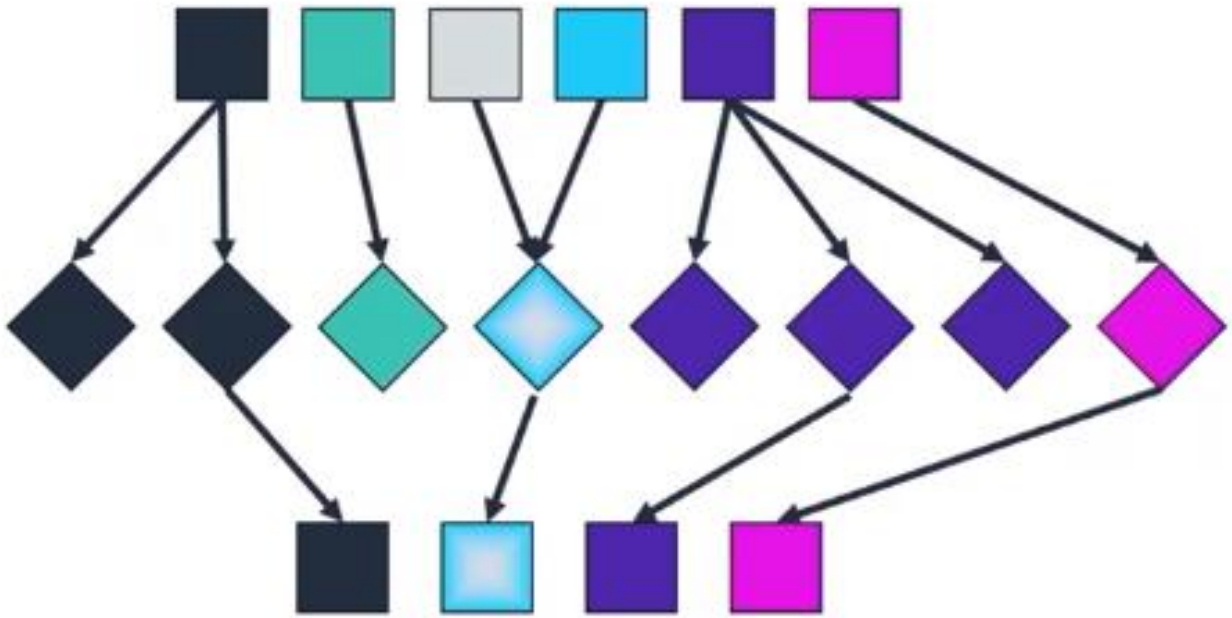
aws academy



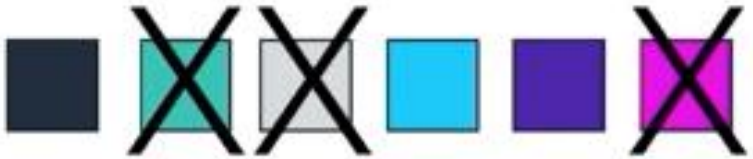
Feature selection and extraction

aws academy

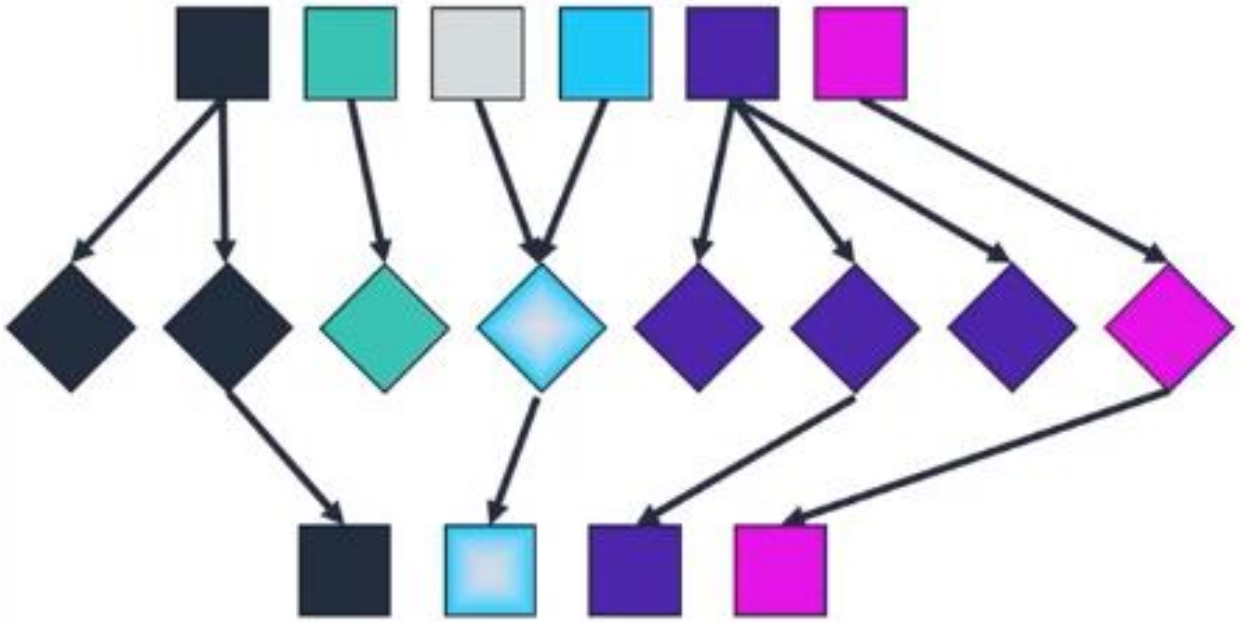
Feature Extraction



Feature Selection



Feature extraction



- Invalid values
- Wrong formats
- Misspelling
- Duplicates
- Consistency
- Rescale
- Encode categories
- Remove outliers
- Reassign outliers
- Bucketing
- Decomposition
- Aggregation
- Combination
- Transformation
- Normalization
- Dimensionality reduction

Encoding ordinal data

Categorical data is non-numeric data.

Categorical data must be converted (encoded) to a numeric scale.

Maintenance Costs	Encoding
Low	1
Medium	2
High	3
Very High	4

Encoding non-ordinal data

If data is non-ordinal, the encoded values also must be non-ordinal.

Non-ordinal data might need to be broken into multiple categories.

...	Color
...	Red
...	Blue
...	Green
...	Blue
	Green

...	Red	Blue	Green
...	1	0	0
...	0	1	0
...	0	0	1
...	0	1	0
...	0	0	1

Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 4a: Cleaning your data

Cleaning data

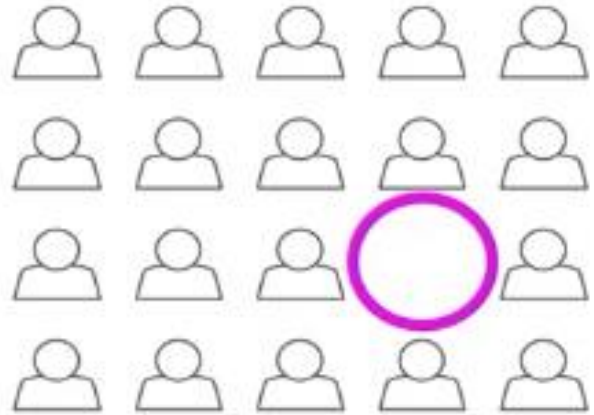
aws academy

Types of data to clean:

Type	Example	Action
Variations in strings	Med. vs Medium	Convert to standard text
Variations in scale	Number of doors vs. number purchased	Normalize to a common scale
Columns with multiple data items	Safe high maintenance	Parse into multiple columns
Missing data	Missing columns of data	Delete rows or impute data
Outliers	Various	

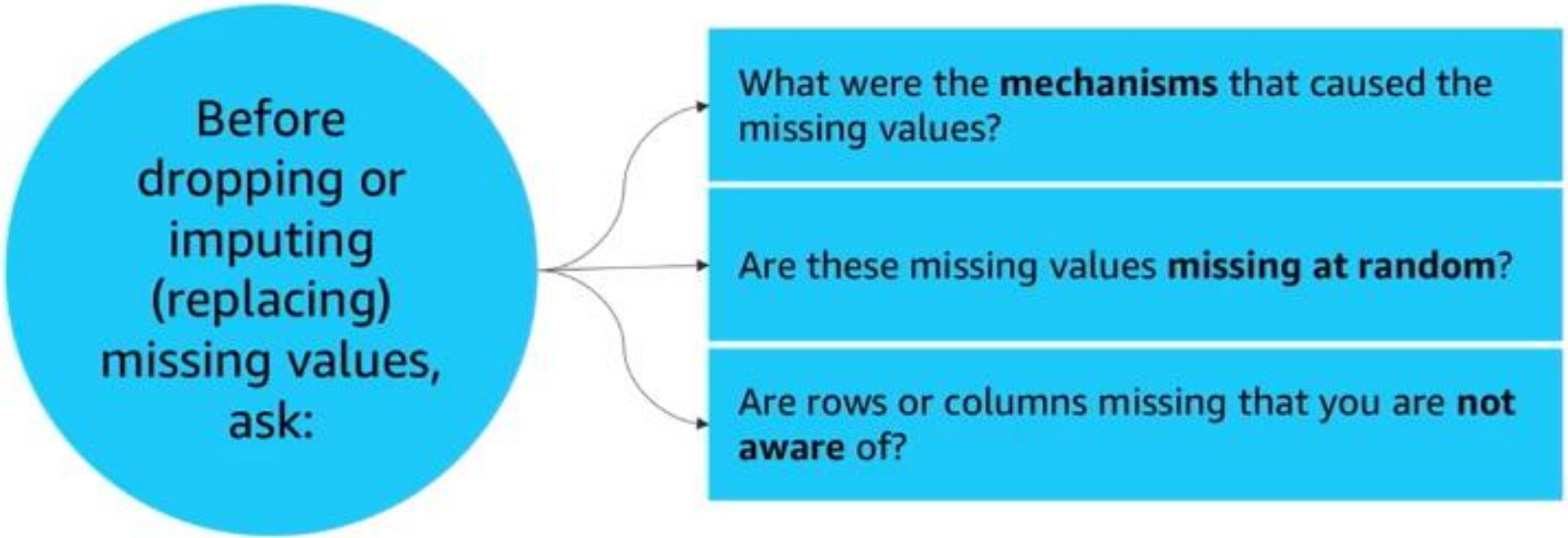
Finding missing data

- Missing data makes it difficult to interpret relationships
- Causes of missing data –
  - Undefined values
  - Data collection errors
  - Data cleaning errors
- Example pandas code to find missing data –



```
df.isnull().sum() #count missing values for each column
df.isnull().sum(axis=1) #count missing values for each row
```

Plan for missing values



Dropping missing values

aws academy

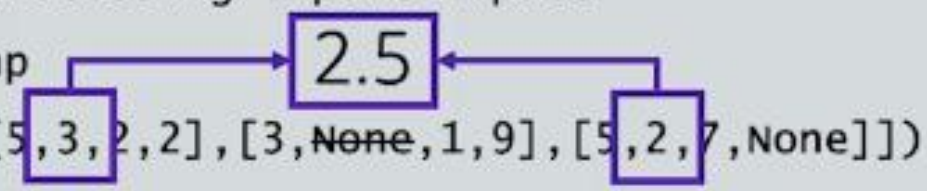
Drop missing data with pandas

- dropna function to drop rows
  - `df.dropna()`
- dropna function to drop columns with null values
  - `df.dropna (axis=1)`
- dropna function to drop a subset
  - `df.dropna(subset=["buying"])`

Imputing missing data

- First, determine why the data is missing
- Two ways to impute missing data –
  - Univariate: Adding data for a single row of missing data
  - Multivariate: Adding data for multiple rows of missing data

```
from sklearn.preprocessing import Imputer
import numpy as np
Arr = np.array([[5,3,2,2],[3,None,1,9],[5,2,7,None]])
imputer = Imputer(strategy='mean')
imp = imputer.fit(arr)
imputer.transform(arr)
```



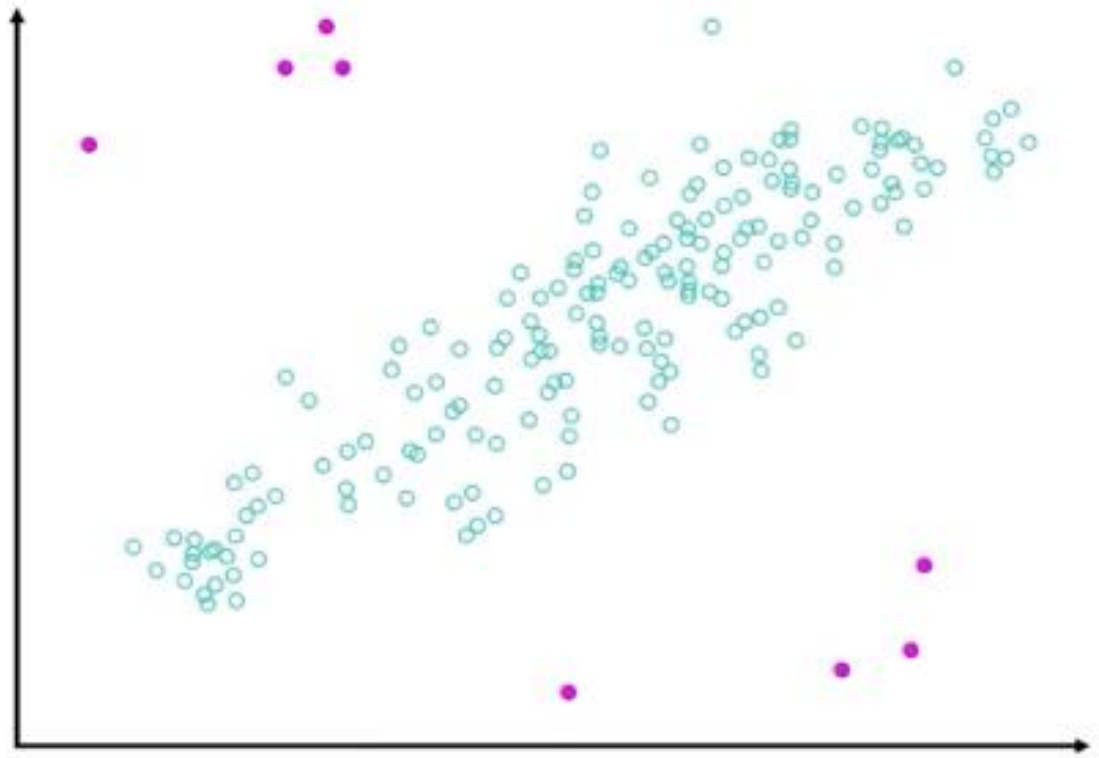
Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 4b Dealing with outliers and selecting features

aws academy

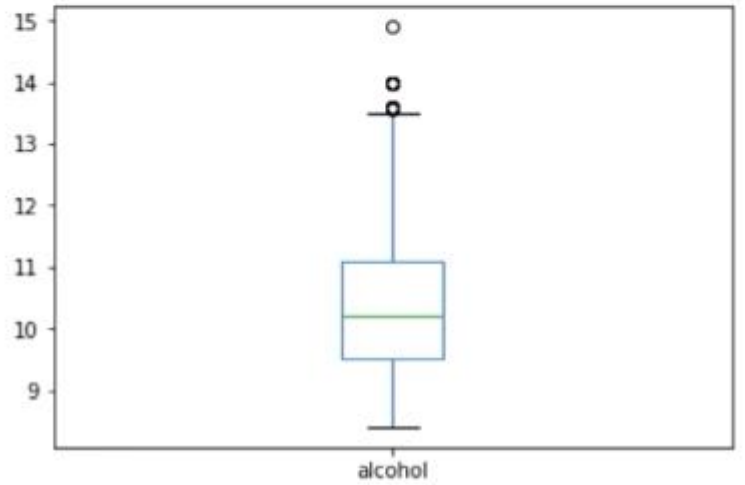
Outliers

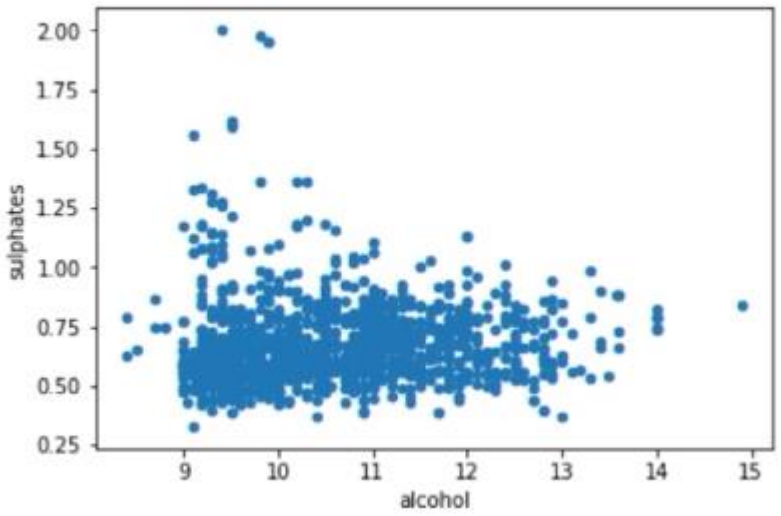
- Outliers can –
  - Provide a broader picture of the data
  - Make accurate predictions difficult
  - Indicate the need for more columns
- Types of outliers –
  - Univariate: Abnormal values for a single variable
  - Multivariate: Abnormal values for a combination of two or more variables



Finding outliers

- Box plots show variation and distance from the mean
  - Example to the right shows a box plot for the amount of alcohol in a collection of wines
- Scatter plots can also show outliers
  - Example to the right shows a scatter plot that shows the relationship between alcohol and sulfites in a collection of wines





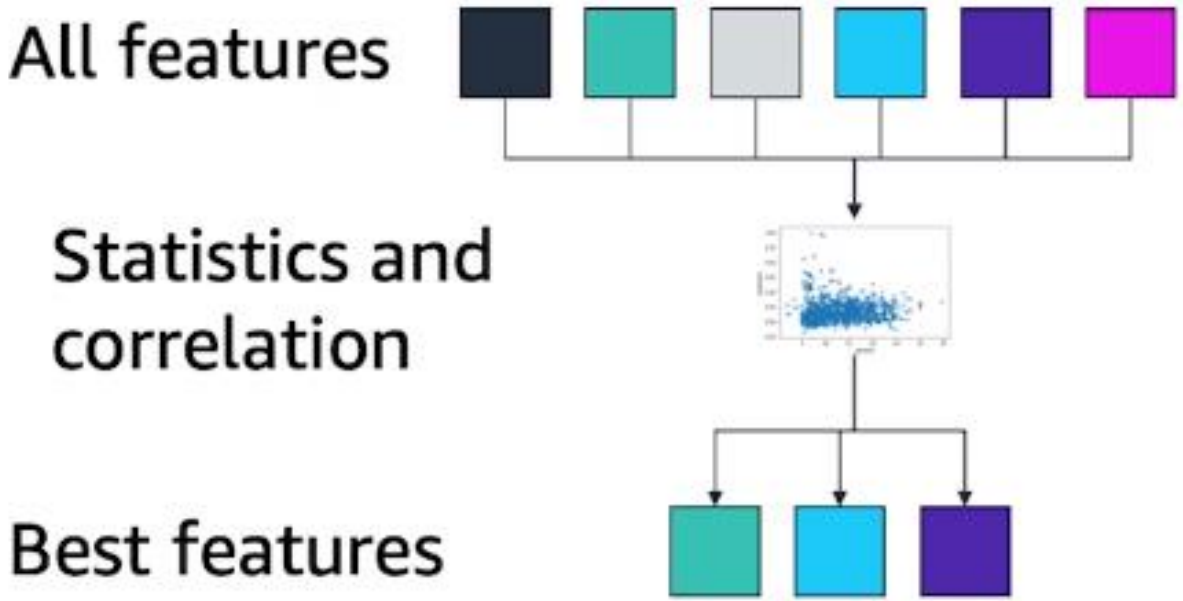
Dealing with outliers

aws academy

Delete the outlier	Transform the outlier	Impute a new value for the outlier
Outlier is based on an artificial error.	Reduces the variation that the extreme outlier value causes and the outlier's influence on the dataset.	You might use the mean of the feature, for instance, and impute that value to replace the outlier value.

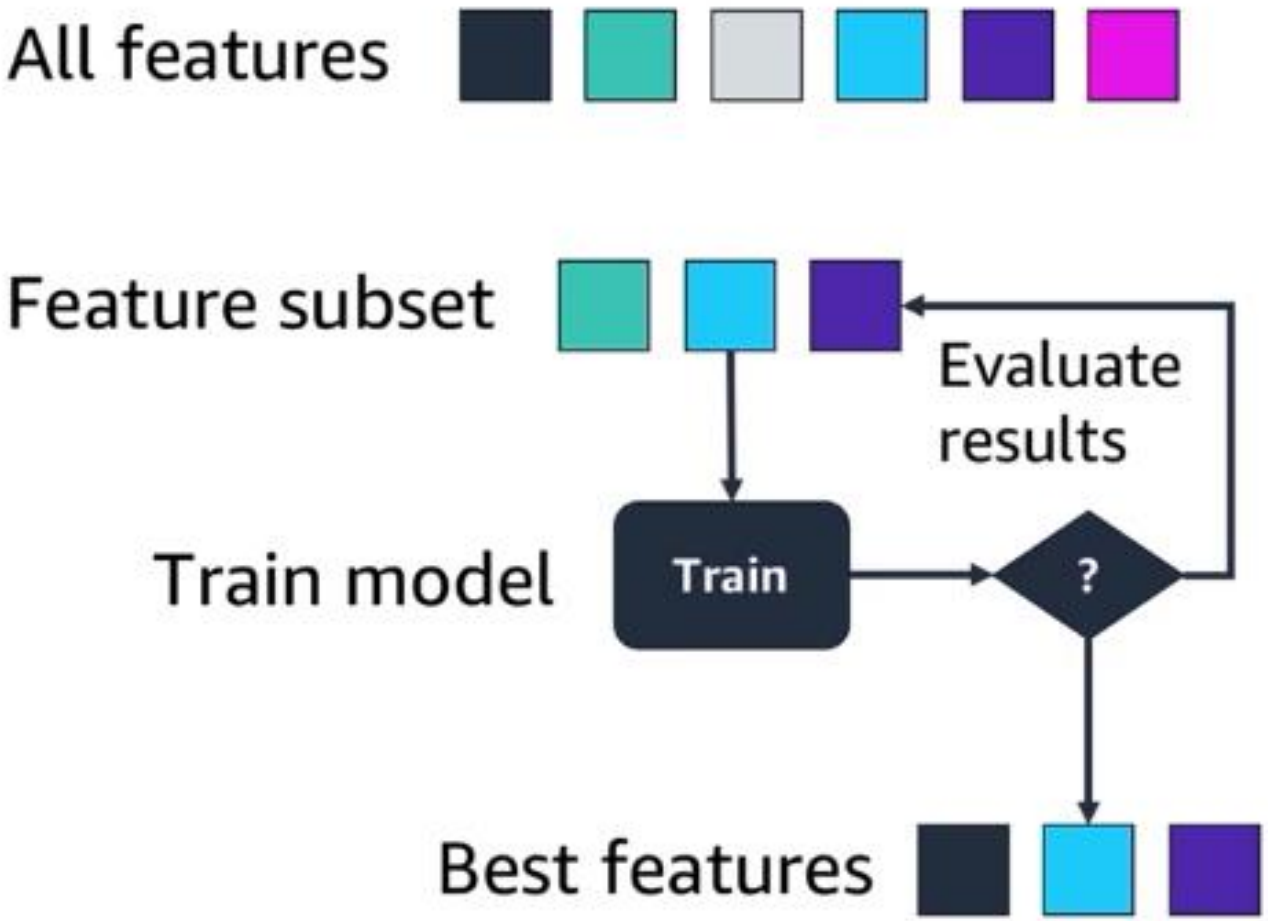
Feature selection: Filter methods

- Measures –
  - Pearson's correlation
  - Linear discriminant analysis (LDA)
  - Analysis of variance (ANOVA)
  - Chi-square



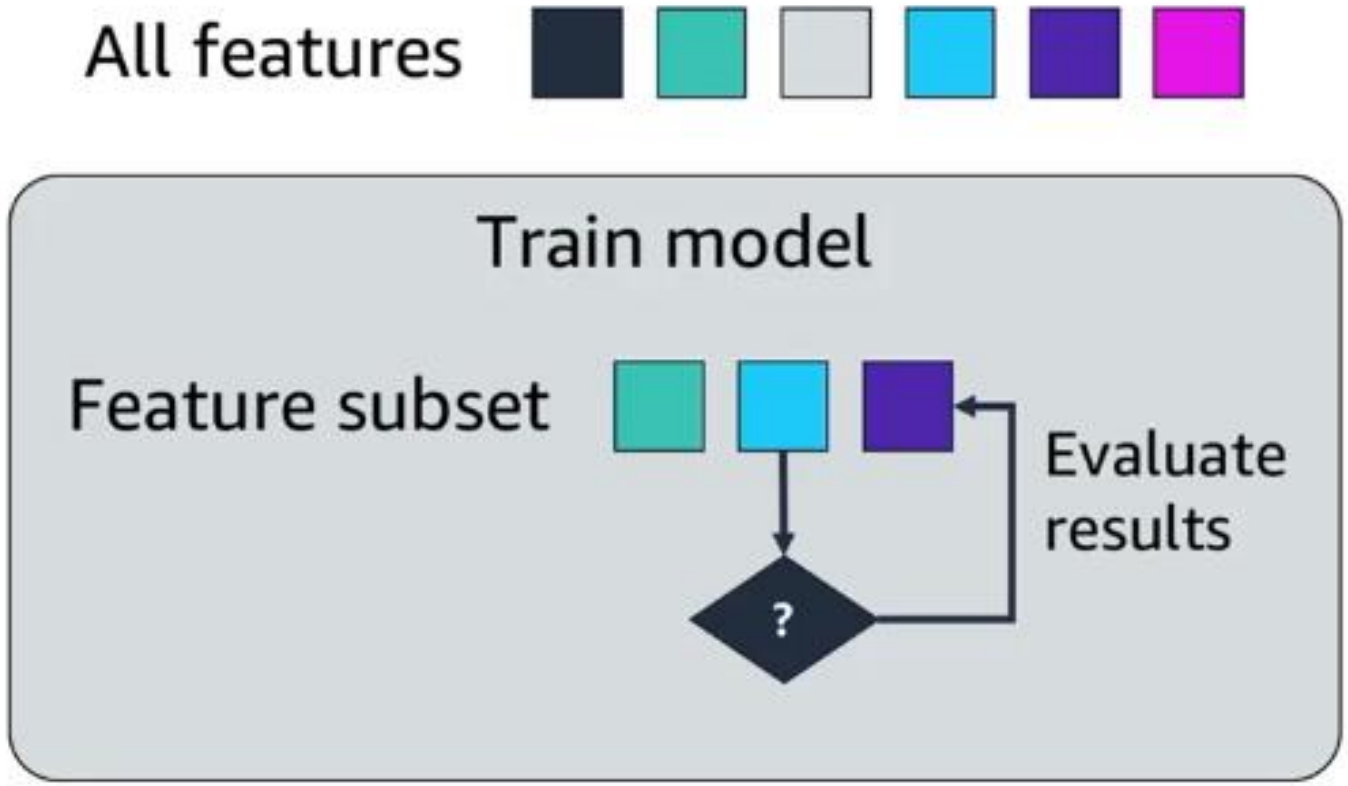
Feature selection: Wrapper

- Methods –
  - Forward selection
  - Backward selection



Feature selection: Embedded methods

- Methods –
  - Decision trees
  - LASSO and RIDGE



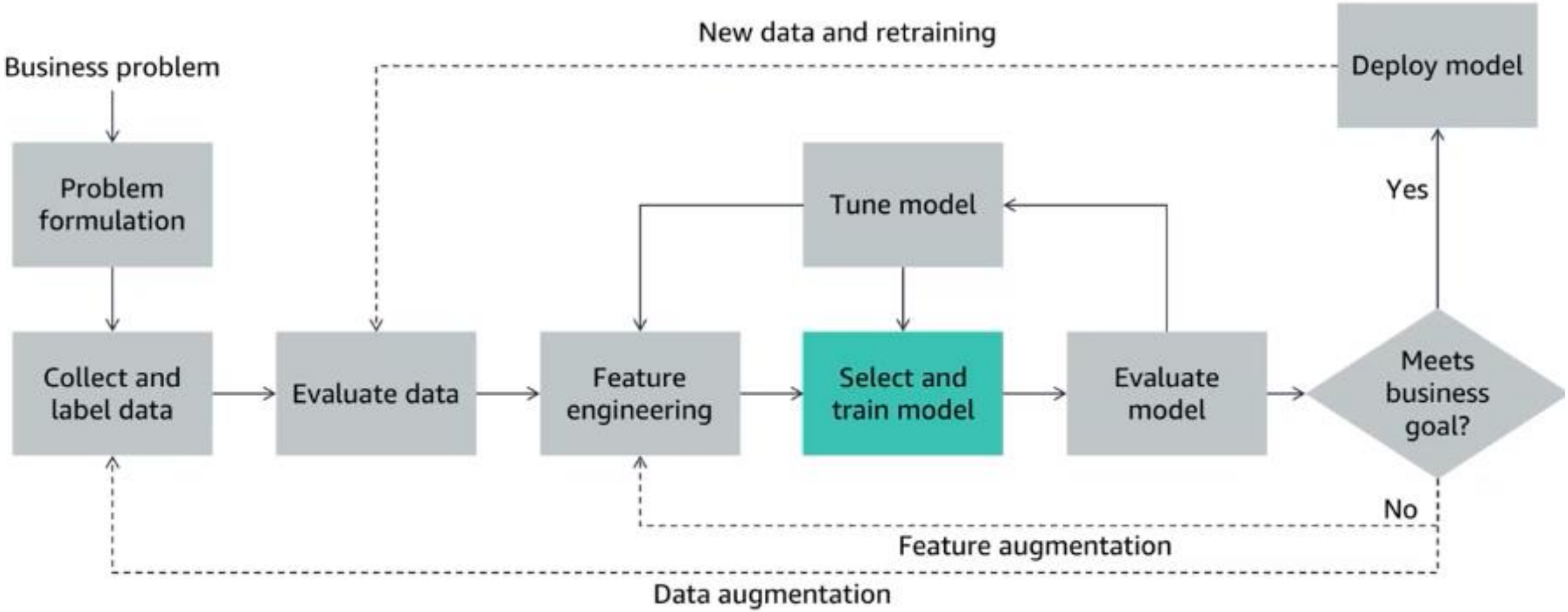
Section 4 key takeaways

- Feature engineering involves –
  - Selection
  - Extraction
- Preprocessing gives you better data
- Two categories for preprocessing –
  - Converting categorical data
  - Cleaning up dirty data
- Use categorical encoding to convert categorical data
- Various types of dirty data –
  - Missing data
  - Outliers
- Develop a strategy for dirty data –
  - Replace or delete rows with missing data
  - Delete, transform, or impute new values for outliers

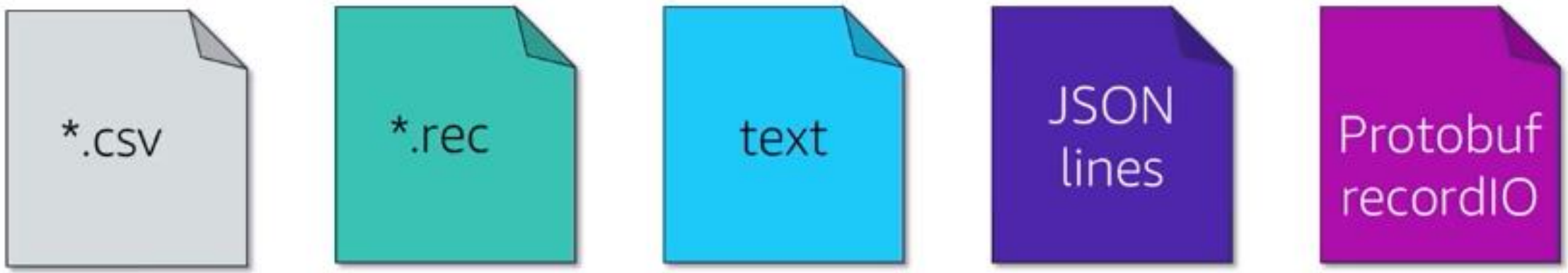
Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 5: Training

Machine learning pipeline

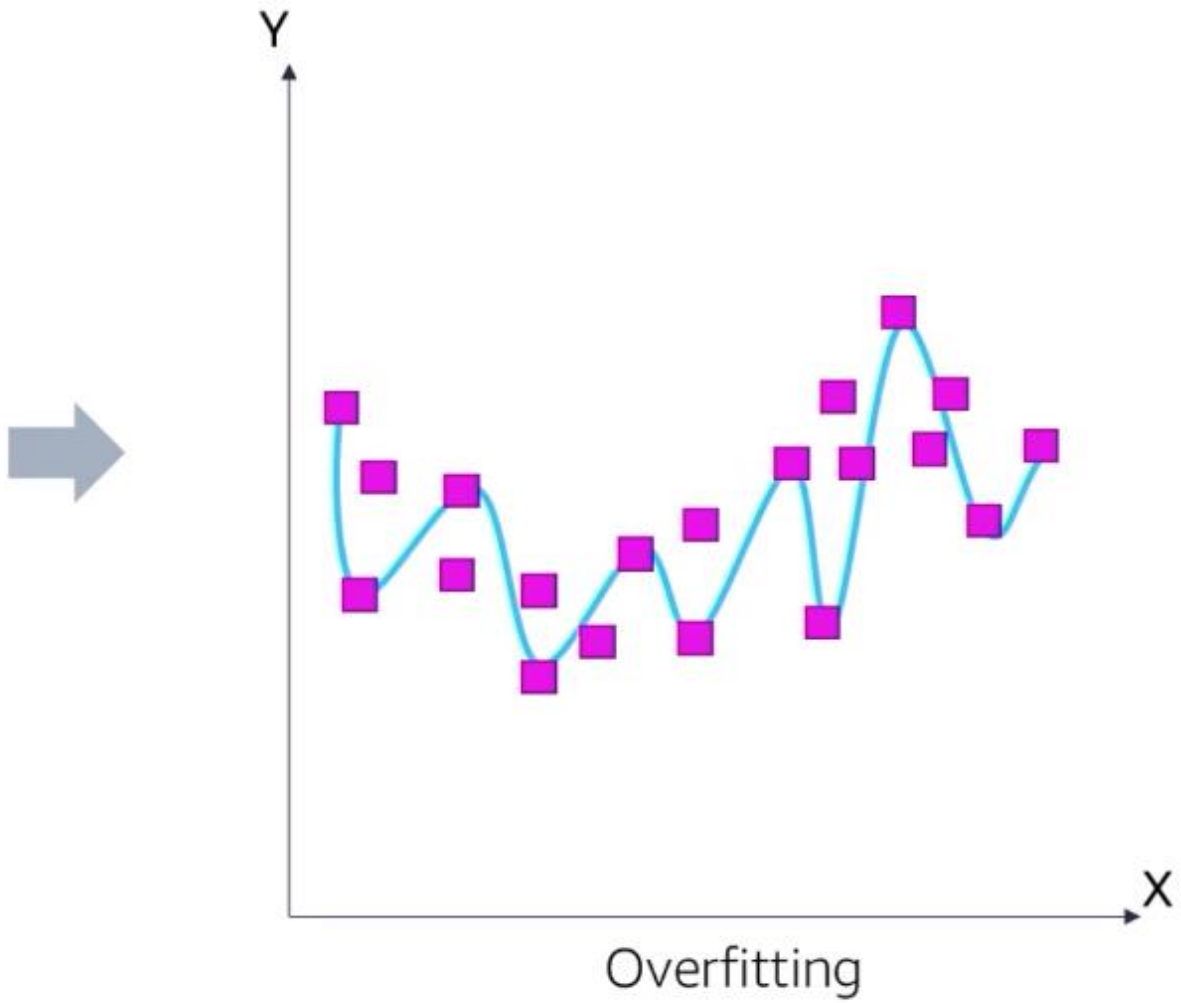


File formats for machine learning

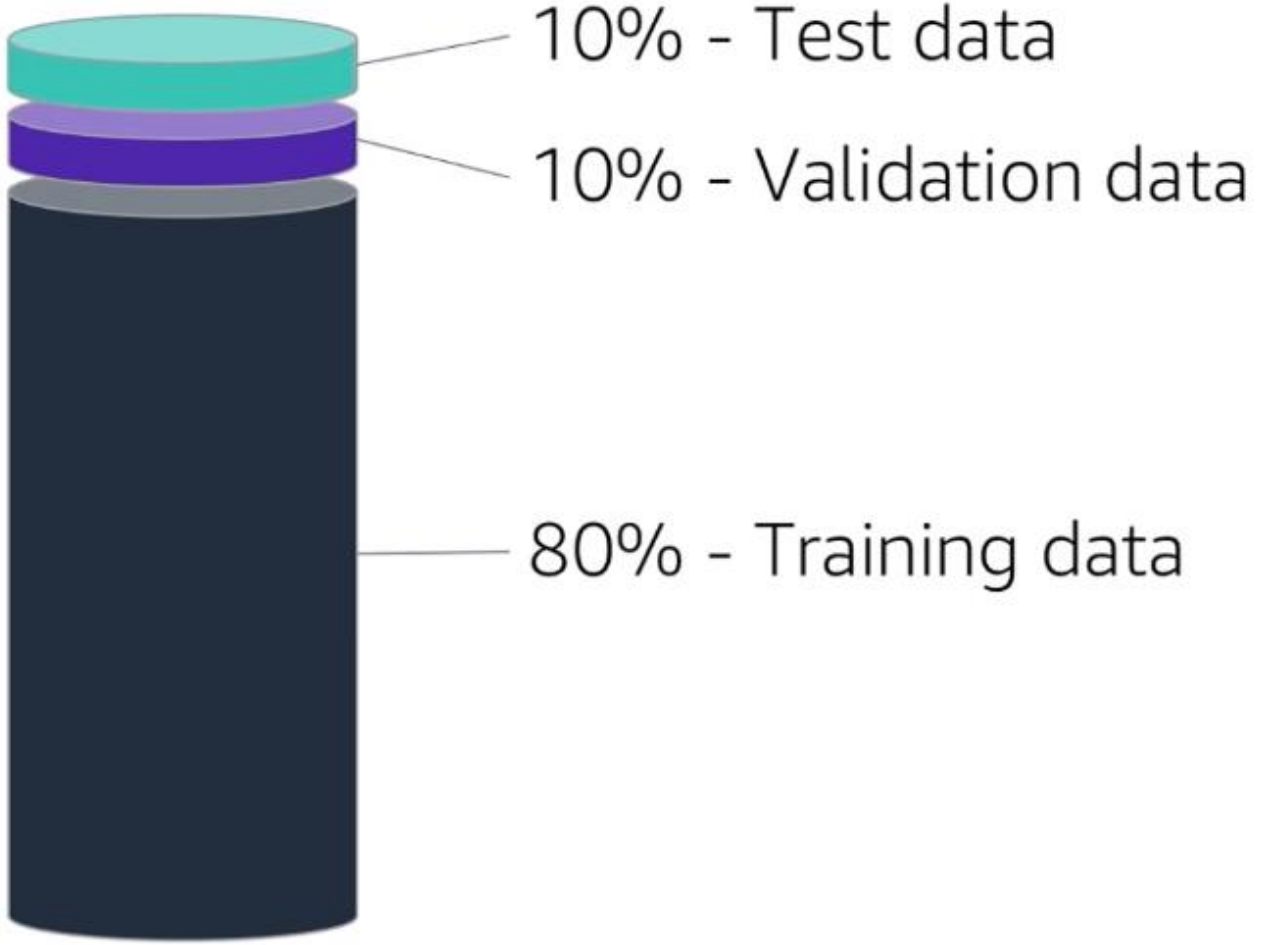


Why split the data?

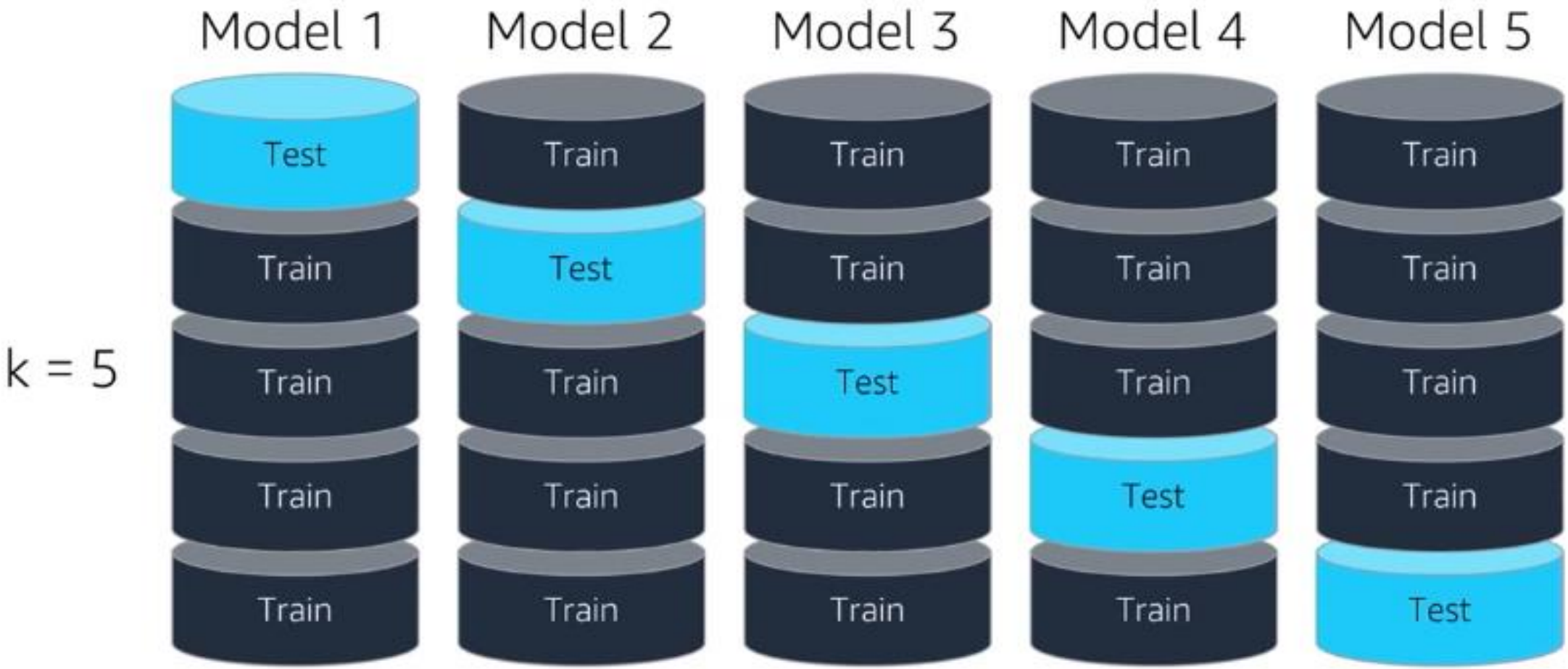
All data that is used for training and evaluation



Holdout method



K-fold cross validation



Shuffle your data

quality	Fixed acidity	sulphates
3	7.2	0.56
3	7.2	0.68
4	7.8	0.65
4	7.8	0.65
4	…	…
5	…	…
5	…	…
5	…	…
6	…	…
6	…	…

Test data

Training data

Training models with Amazon SageMaker

Amazon SageMaker provides ML algorithms that are optimized for speed, scale, and accuracy.

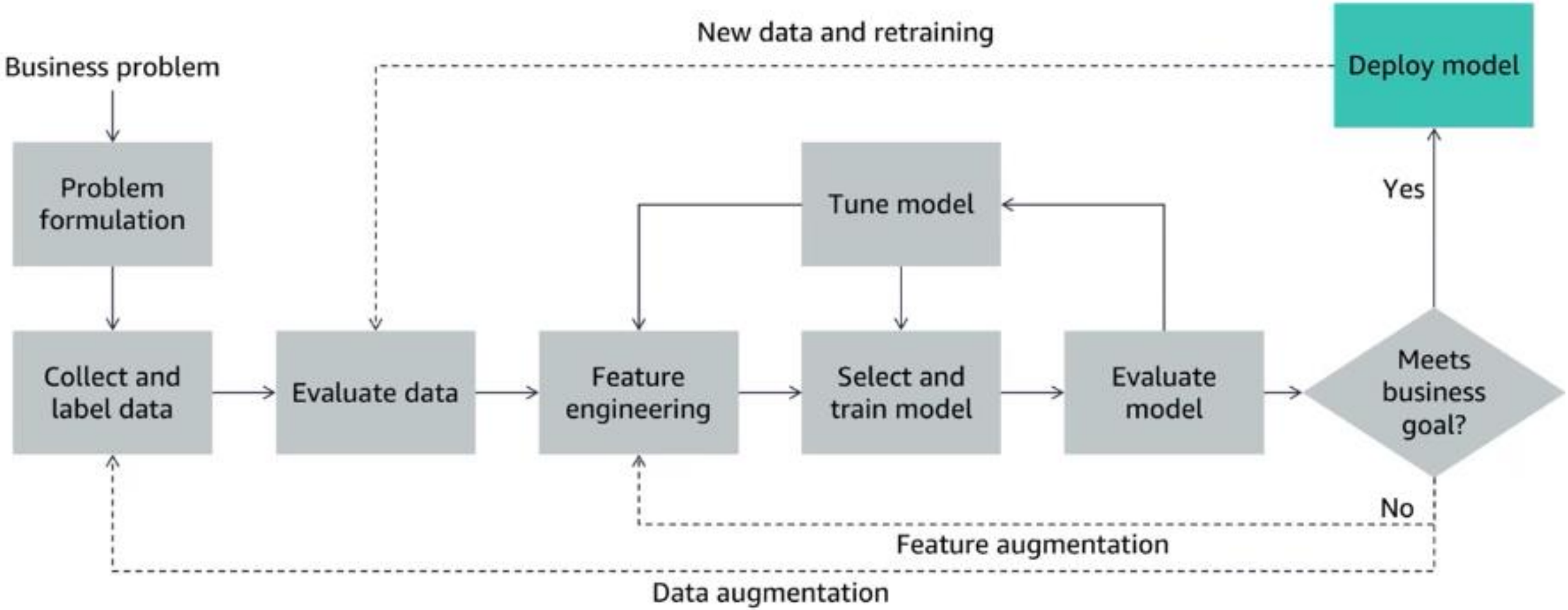
- Amazon SageMaker built-in algorithms
- Amazon SageMaker supported frameworks
- Amazon SageMaker custom frameworks
- AWS Marketplace algorithms

Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

Section 6: Hosting and using the model

aws academy

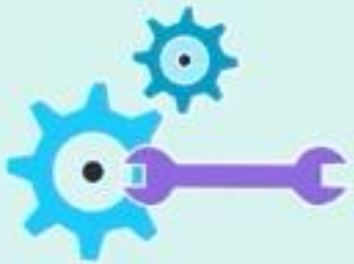
Machine learning pipeline



Is your model ready to deploy?

aws academy

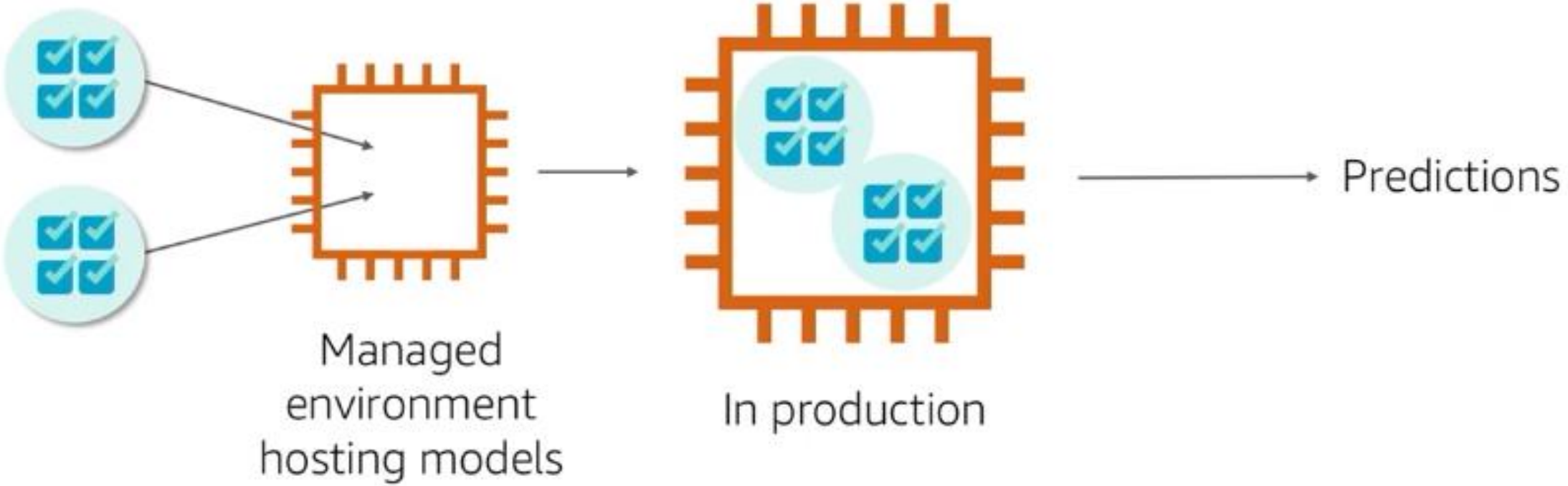
Trained



Tuned

Tested

The goal of deployment



Section 6 key takeaways

aws academy

- Can use two options for deploymen
  - Amazon SageMaker hosting
  - Batch transform
- Deploy only after you have tested your model
  - Goal is to generate predictions for client applications
- Create an endpoint
  - Single-model endpoint for simple use cases
  - Multi-model endpoint to support multiple use cases

Module 3: Implementing a Machine Learning Pipeline with Amazon SageMaker

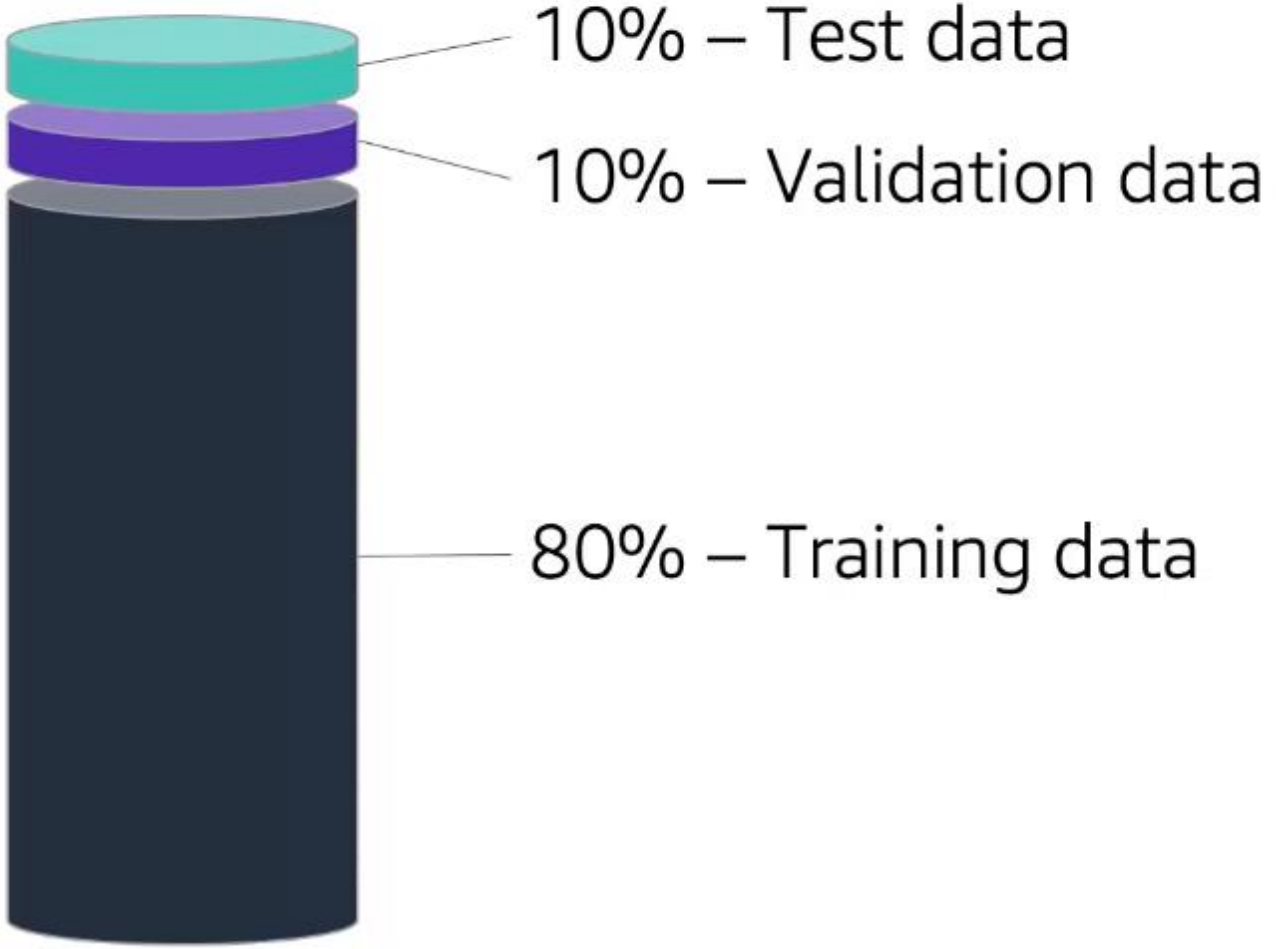
Section 7: Evaluating the accuracy of the model

Evaluation determines how well your model predicts the target on future data

aws academy

Testing and evaluation:

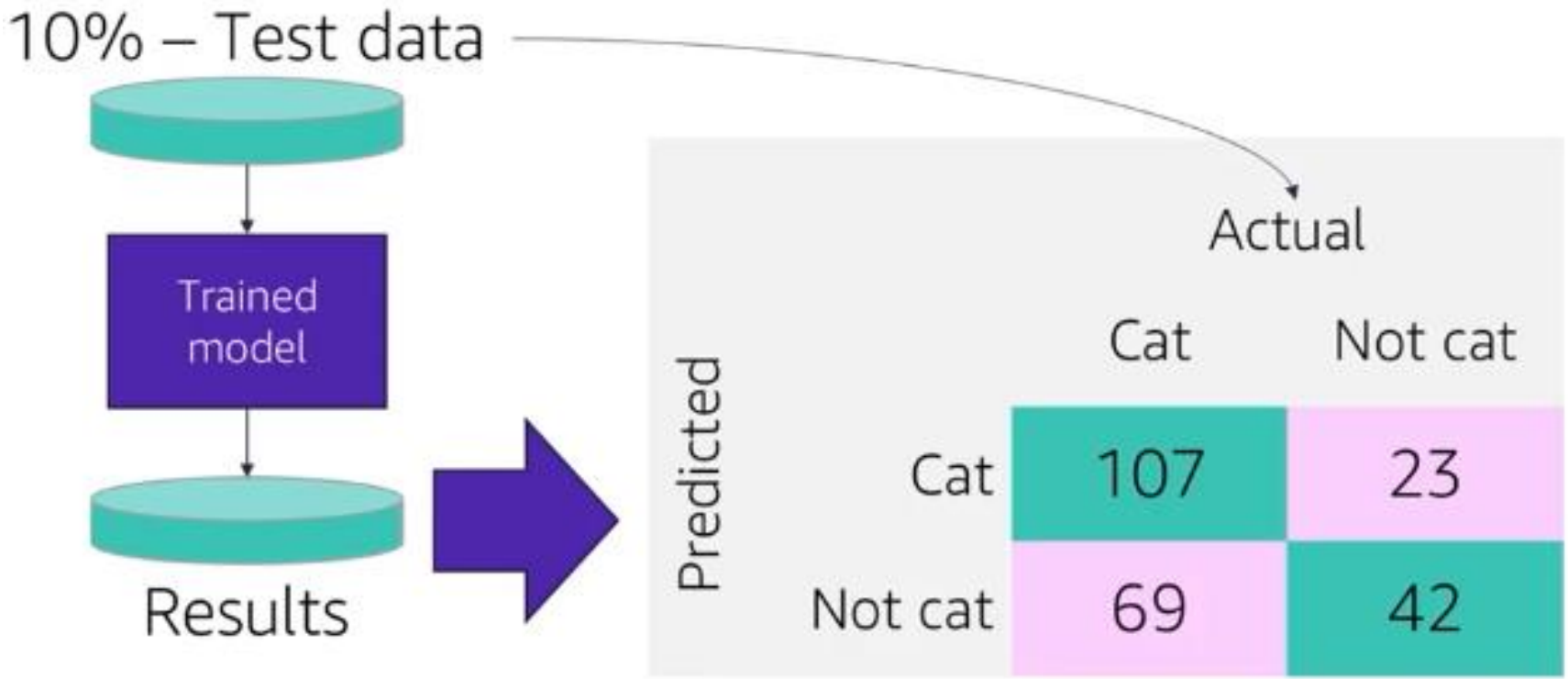
To evaluate the predictive quality of the trained model



Success metric

The model metric must align to both the business problem and the success metric.

Confusion matrix



Credit:
The content is taken from the **AWS Academy Machine Learning Foundations [13258]** course