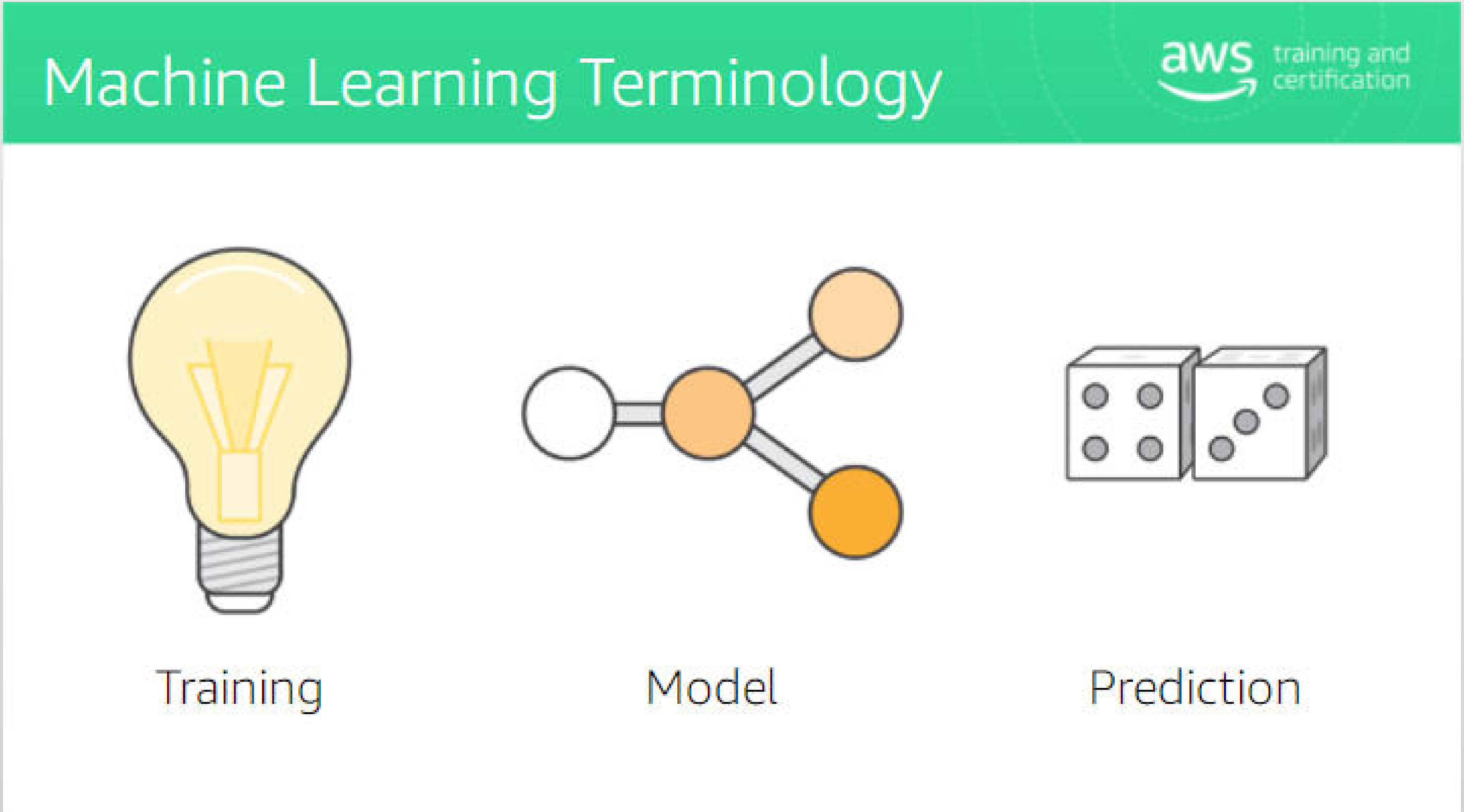
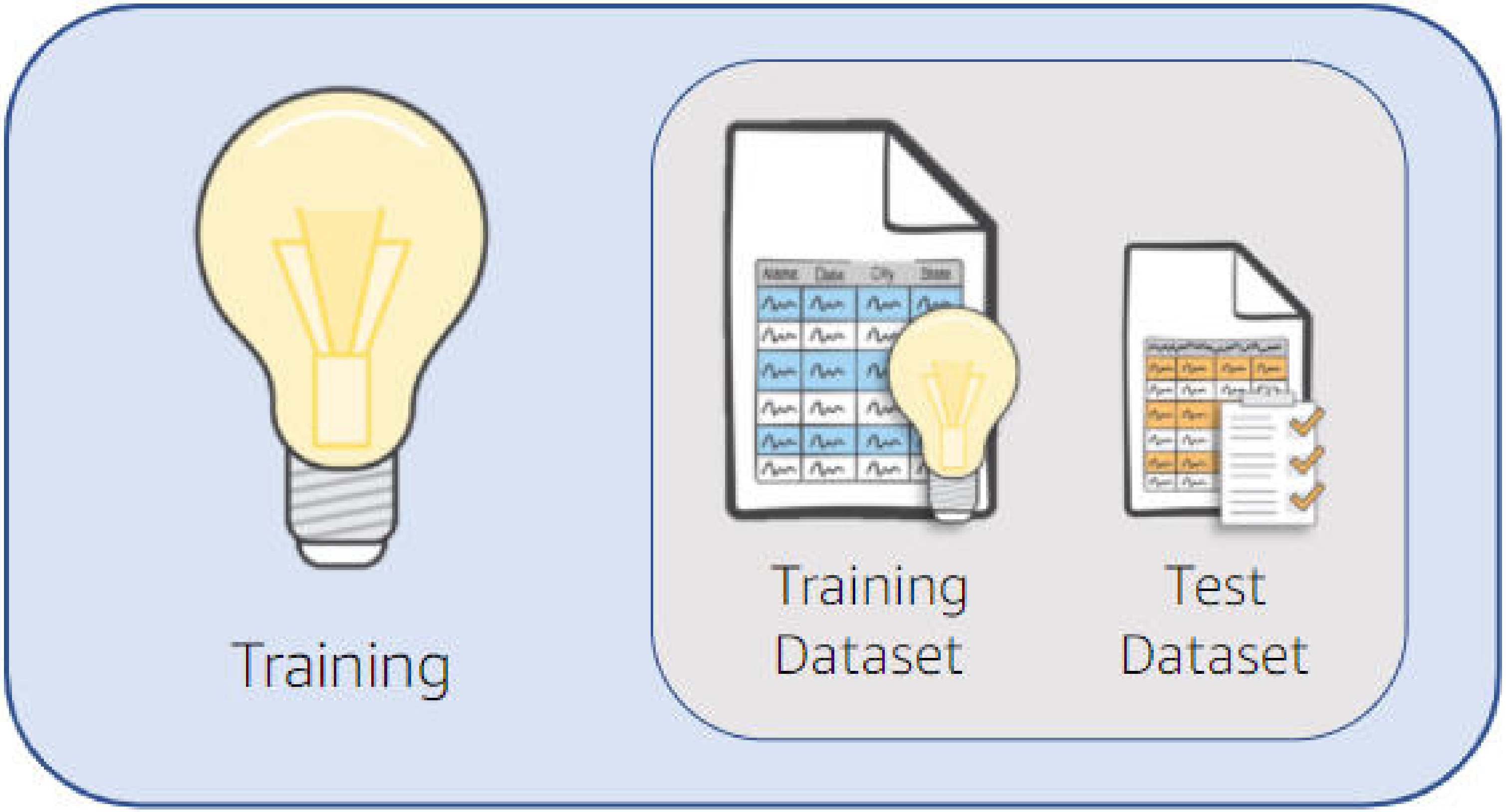
In This Course

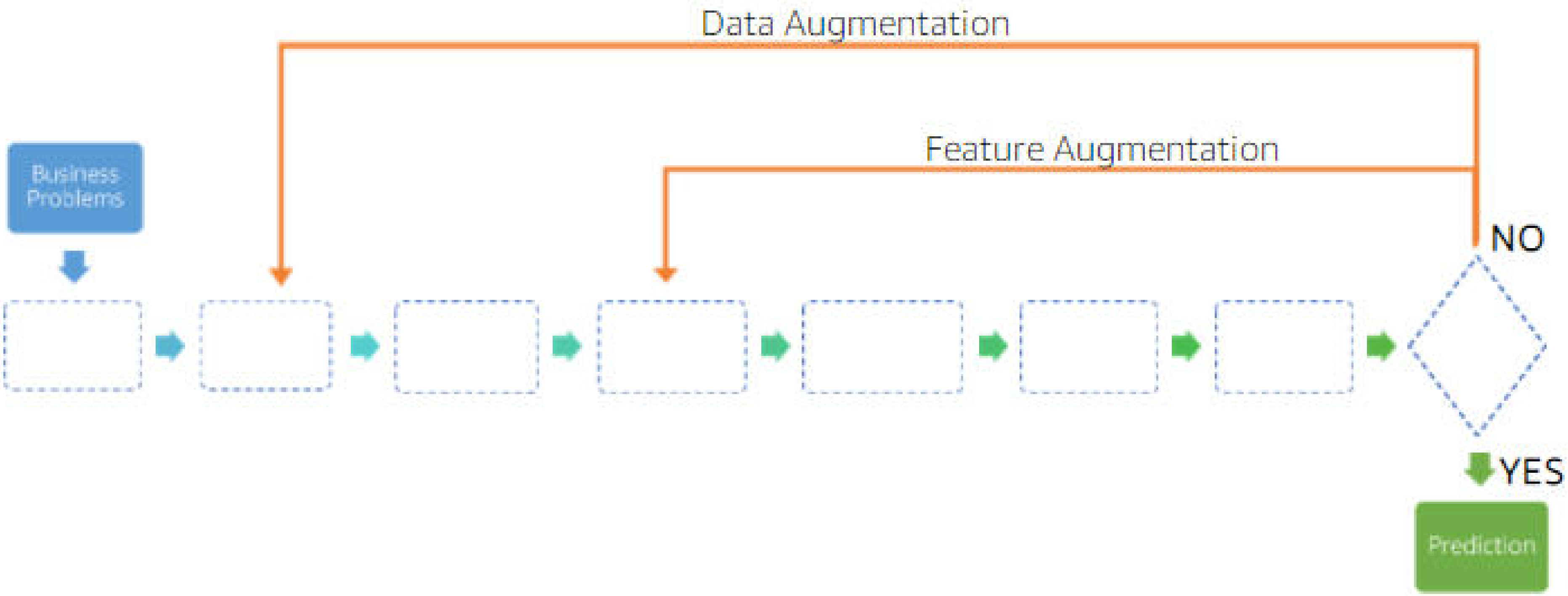
- Common Machine Learning Terminology
- The Machine Learning Process


Machine Learning Terminology
aws training and certification
Training
Model
Prediction

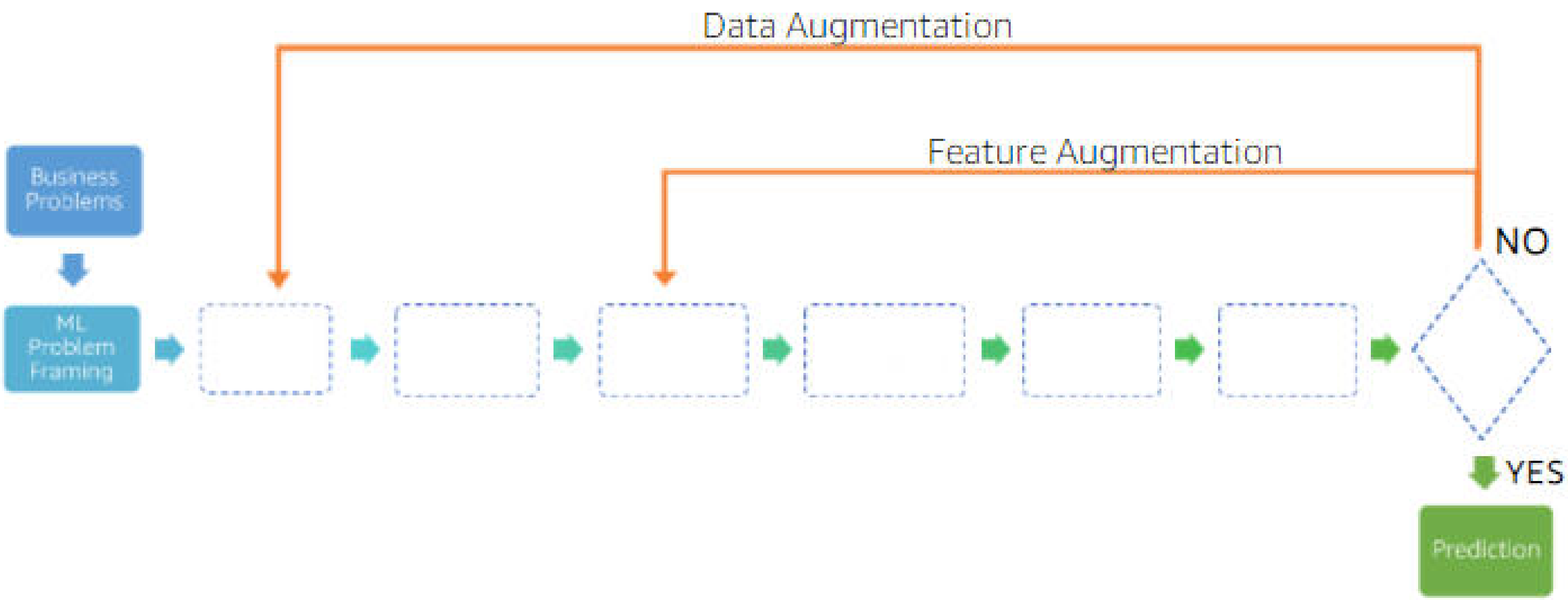
Machine Learning Terminology



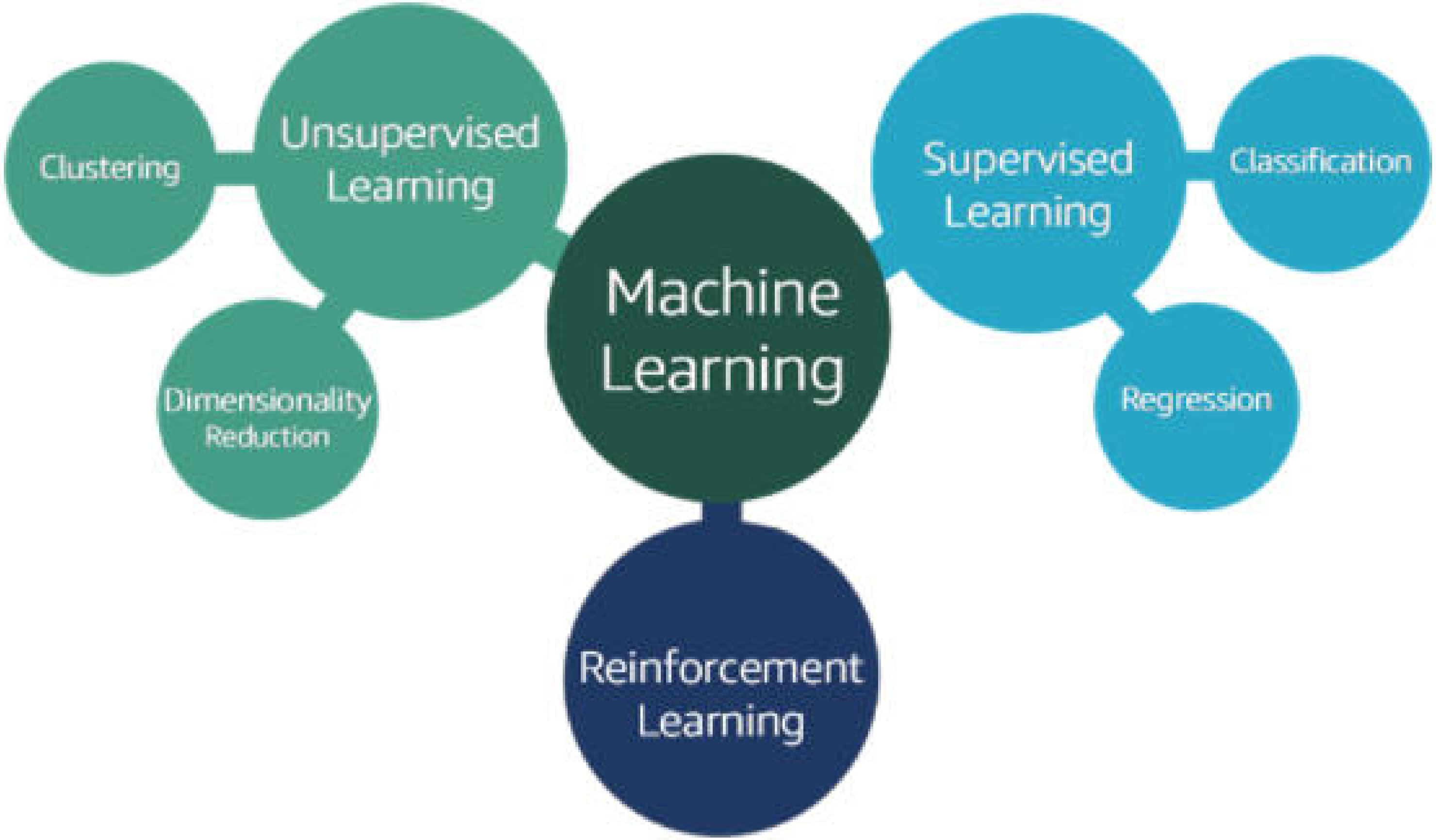
Step 1: The Business Problem



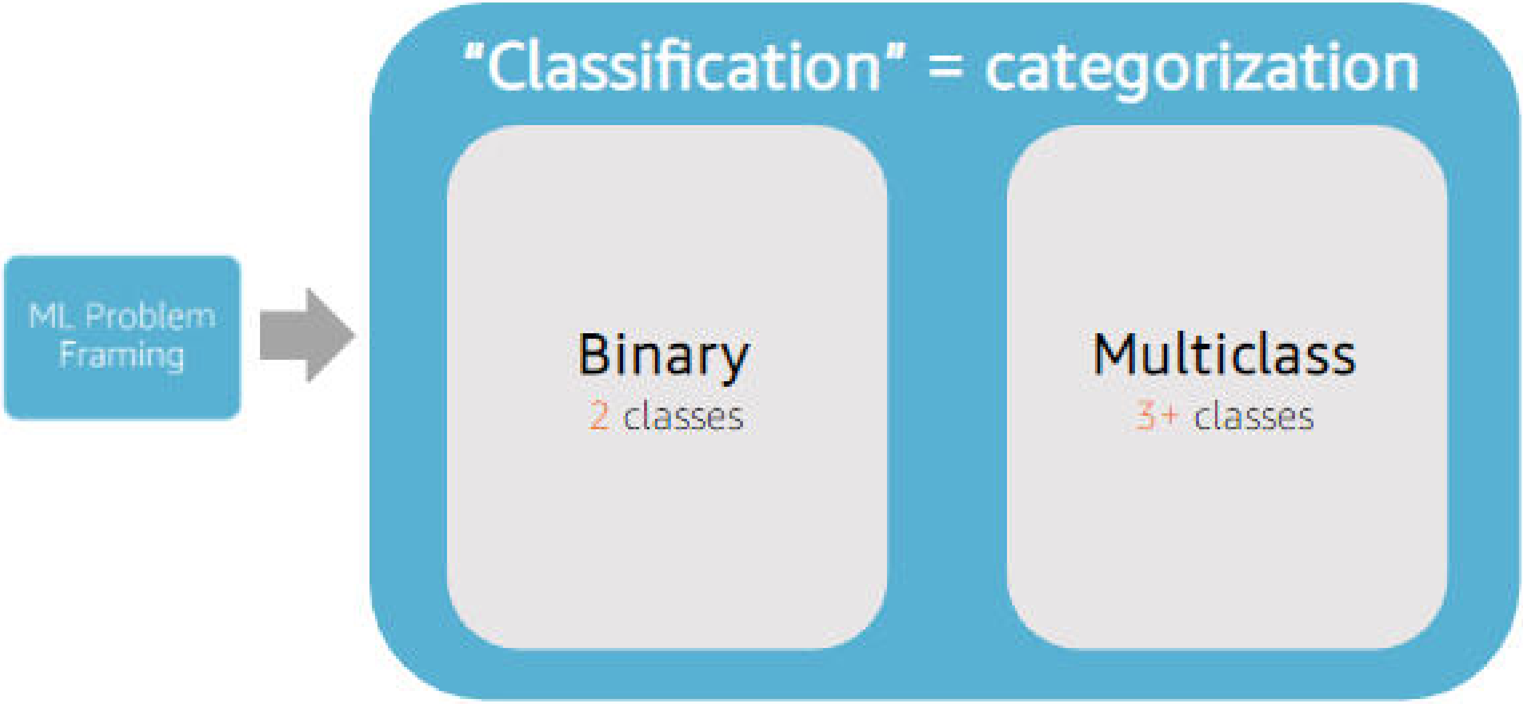
Step 2: The Machine Learning Problem



Questions to Ask



Machine Learning Problems



Machine Learning Problem Definition

aws training and certification

Key elements

- Observations
- Labels
- Features

Example: Income classification problem

- Predict if a person makes more than $50K

Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bachelors	14	Single	Adm-clerical	Male	<50K (-1)
31	Masters	18	Married	Engineering	Female	>=50K (+1)

Machine Learning Problem Definition

Key elements

- Observations
- Labels
- Features

Example: Income classification problem

- Predict if a person makes more than $50K

Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bachelors	14	Single	Adm-clerical	Male	<50K (-1)
31	Masters	18	Married	Engineering	Female	>=50K (+1)

Numeric

Machine Learning Problem Definition

aws training and certification

Key elements

- Observations
- Labels
- Features

Example: Income classification problem

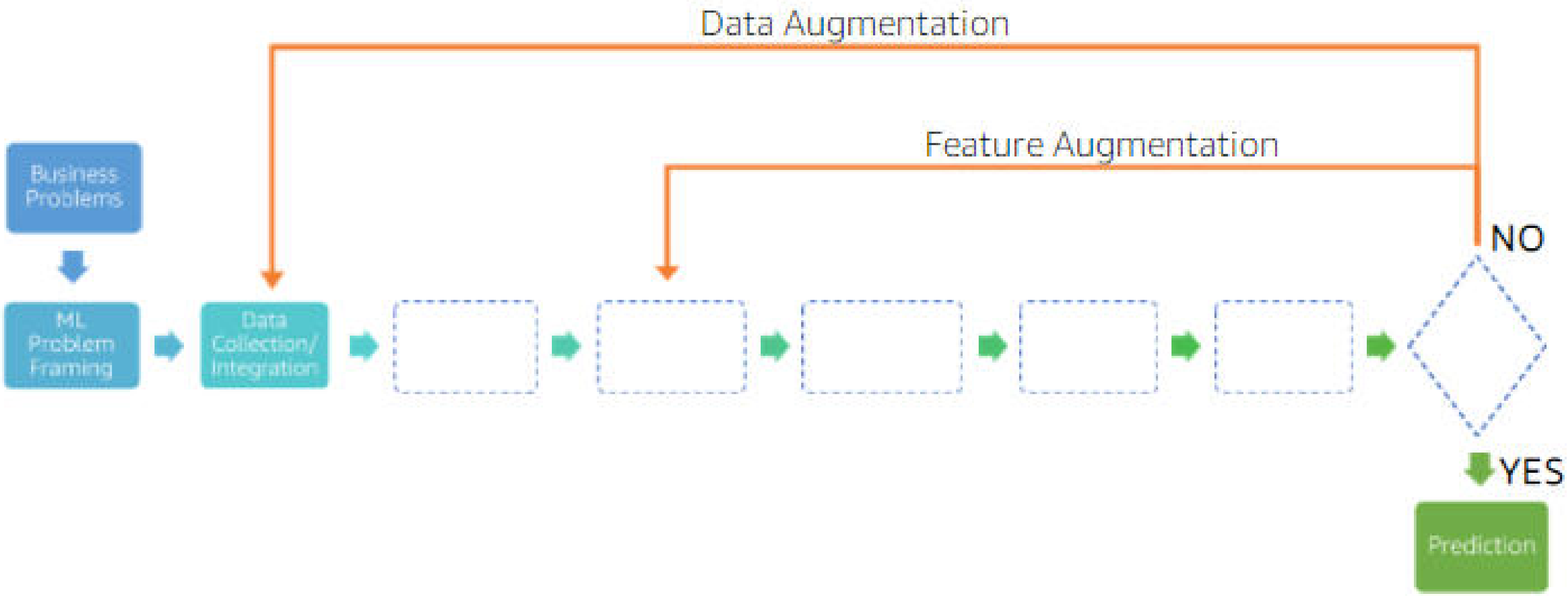
- Predict if a person makes more than $50K

Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bachelors	14	Single	Adm-clerical	Male	<50K (-1)
31	Masters	18	Married	Engineering	Female	>=50K (+1)

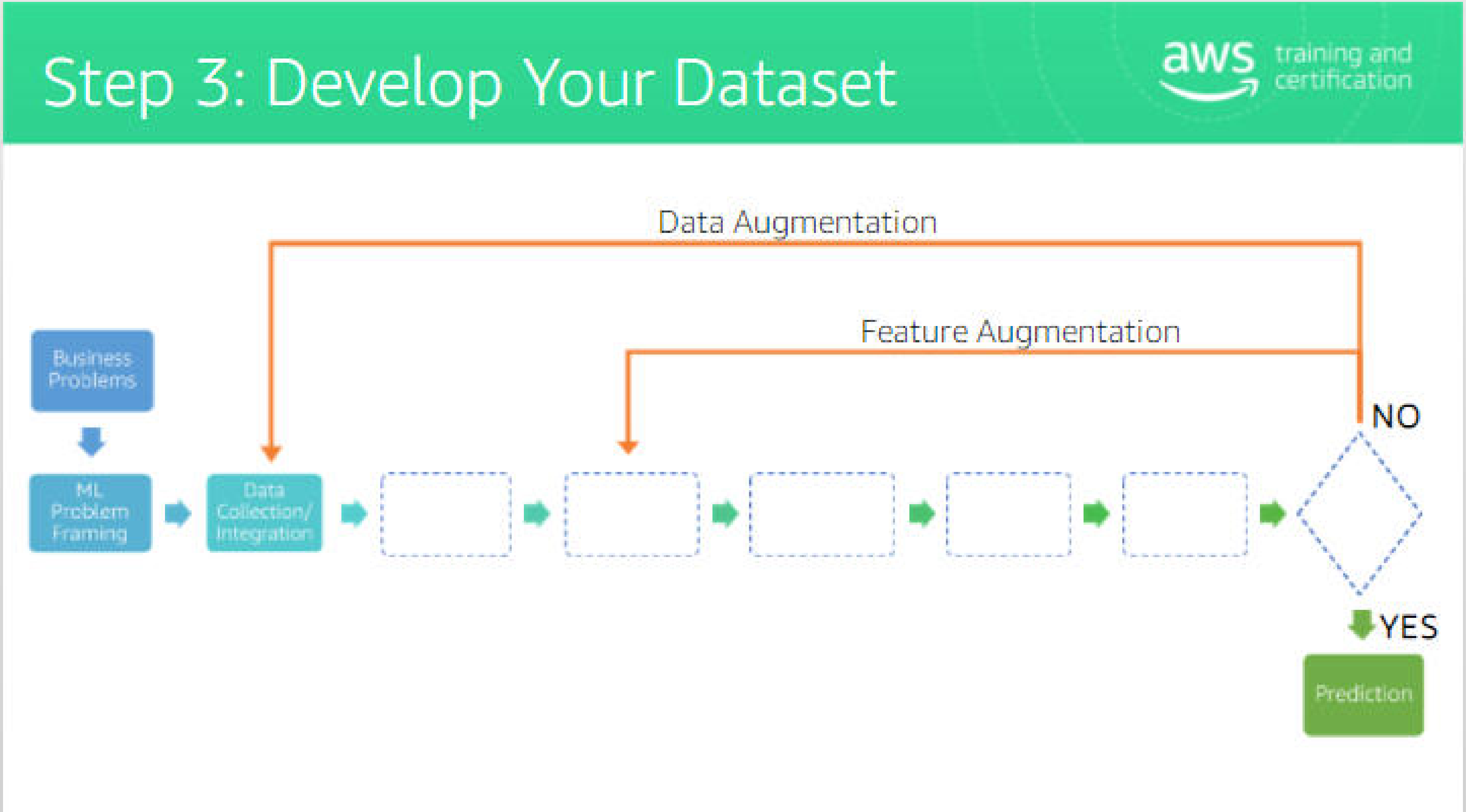
Numeric

Categorical

Step 3: Develop Your Dataset



Machine Learning Terminology and Process



Data Collection & Integration

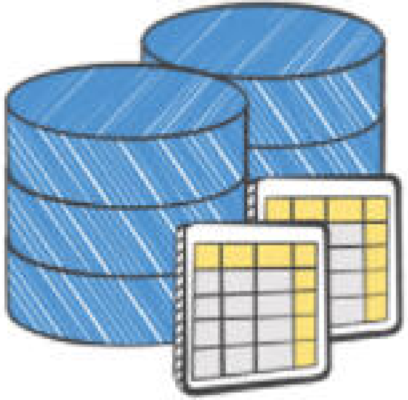
Amazon S3

Amazon DynamoDB

Amazon Redshift

Web pages

Data Collection & Integration



Structured

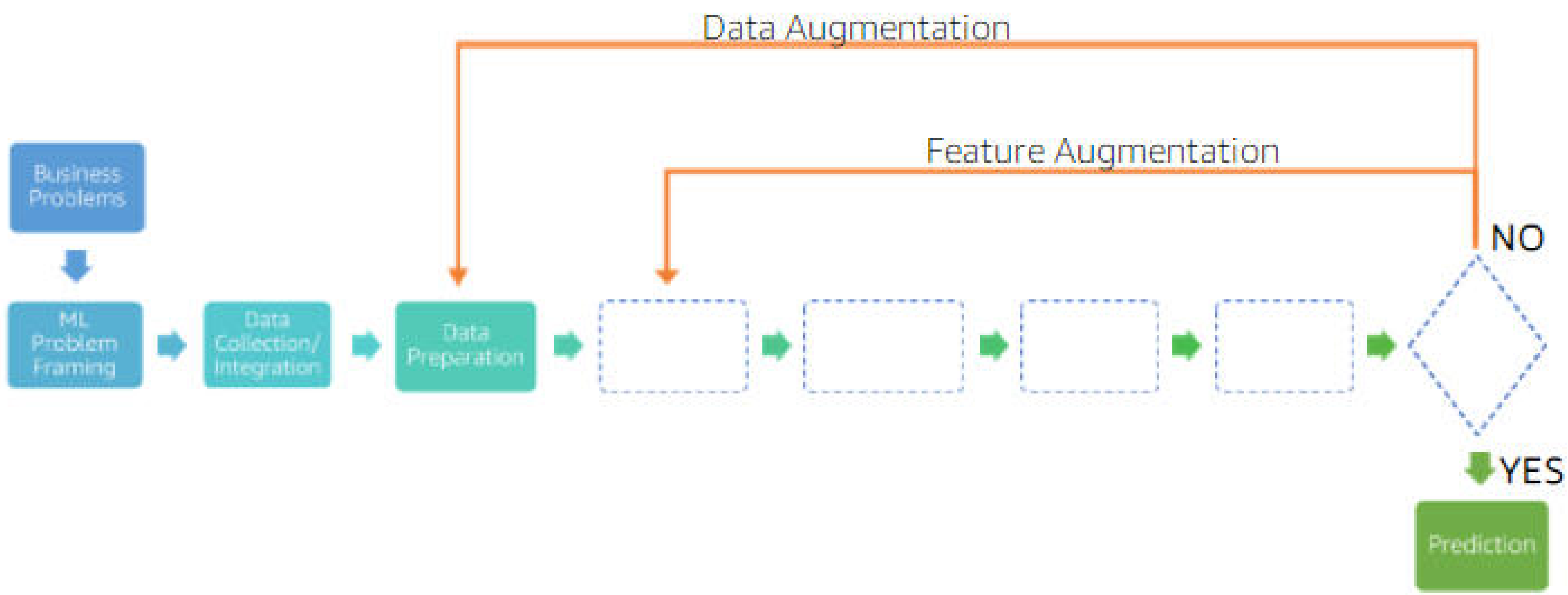
```
{
  {
    "first_name": "John",
    "last_name": "Doe"
  },
  {
    "first_name": "Jane",
    "last_name": "Doe"
  }
}
```

Semi-structured

```
111.22.33.444 - -
[19/Nov/2017:05:44:17 -0700] "GET
/images/imagename.png HTTP/1.1"
200 124
123.45.67.89 - -
[19/Nov/2017:05:44:18 -0700] "GET
/javascript/config.js
 HTTP/1.1" 200 239
```

Unstructured

Step 4: Data Preparation



Data Cleaning

Data Preparation: Handling missing feature values and outliers.

Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bachelors	14	Single	Adm-clerical	Male	0
31	Masters	18	Married	Engineer	Female	1
44	Bachelors			Accounting	Male	0
150	Bachelors	14	Married	Engineer	Female	0

Data Cleaning

Data Preparation: Handling missing feature values and outliers.

- Introduce new indicator variable to represent missing value
- Remove the rows with missing values
- Imputation

Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bachelors	14	Single	Adm-clerical	Male	0
31	Masters	18	Married	Engineer	Female	1
44	Bachelors			Accounting	Male	0
150	Bachelors	14	Married	Engineer	Female	0

Outlier

Missing values

Data Cleaning

Data Preparation: Handling missing feature values and outliers.

- Introduce new indicator variable to represent missing value
- Remove the rows with missing values
- Imputation

Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bachelors	14	Single	Adm-clerical	Male	0
31	Masters	18	Married	Engineer	Female	1
44	Bachelors			Accounting	Male	0
150	Bachelors	14	Married	Engineer	Female	0

Outlier

Missing values

Data Cleaning

aws training and certification

Data Preparation: Handling missing feature values and outliers.

- Introduce new indicator variable to represent missing value
- Remove the rows with missing values
- Imputation

Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bachelors	14	Single	Adm-clerical	Male	0
31	Masters	18	Married	Engineer	Female	1
44	Bachelors			Accounting	Male	0
150	Bachelors	14	Married	Engineer	Female	0

Outlier

Missing values

Impute Missing Values

...A technique for handling missing values or outliers.

If the missing attribute is numerical:

- Mean
- Median

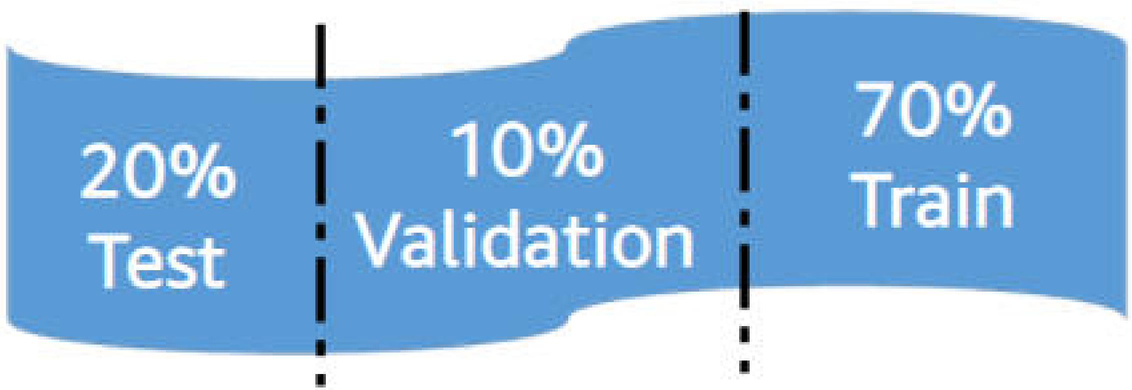
Shuffle Training Data

- Shuffling results in better model performance for certain algorithms
- Minimizes the risk of cross validation data under representing the model data AND model data not learning from all type of data

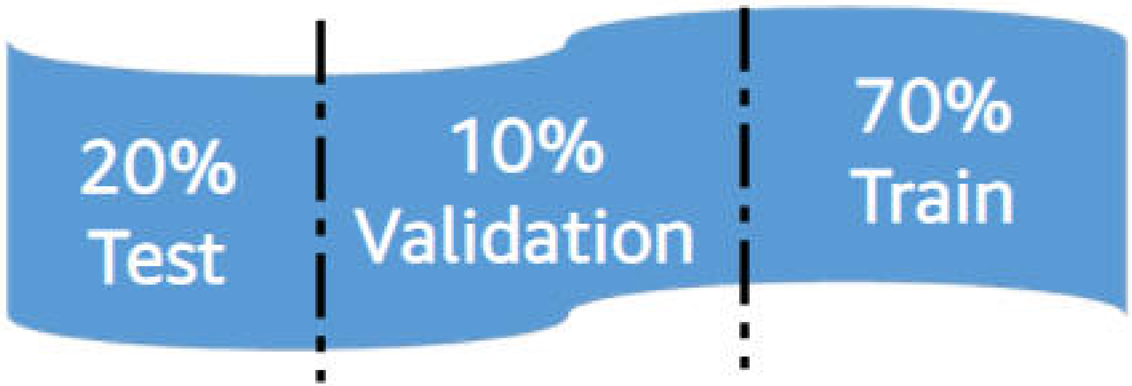
```
In [22]: train_data = train_data.sample(frac = 1)
```

Test-Validation-Train Split

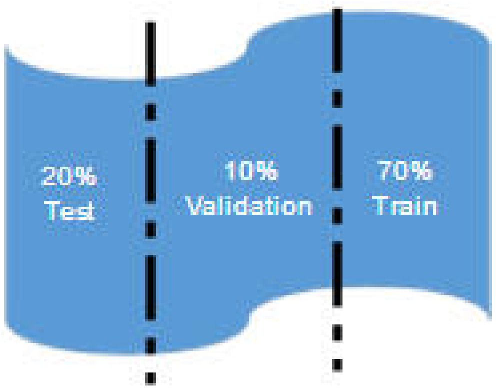
aws training and certification



Test-Validation-Train Split

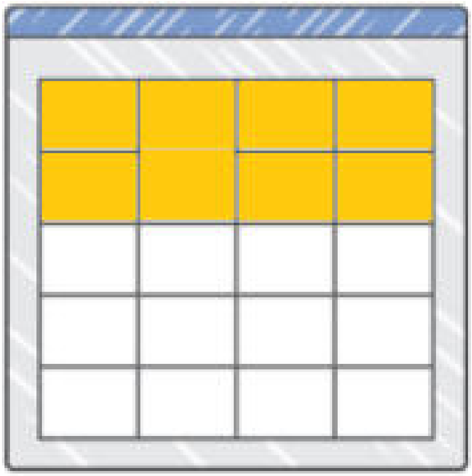


Cross Validation



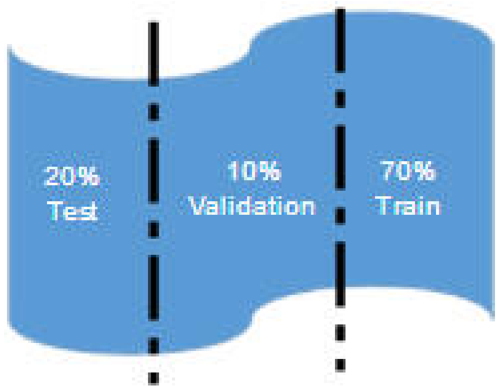
Validation

Leave-one-out



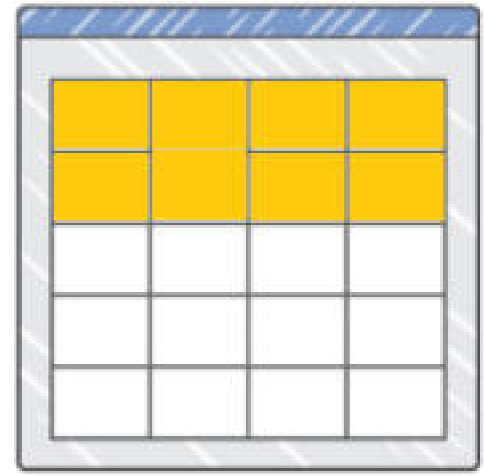
K-fold

Cross Validation



Validation

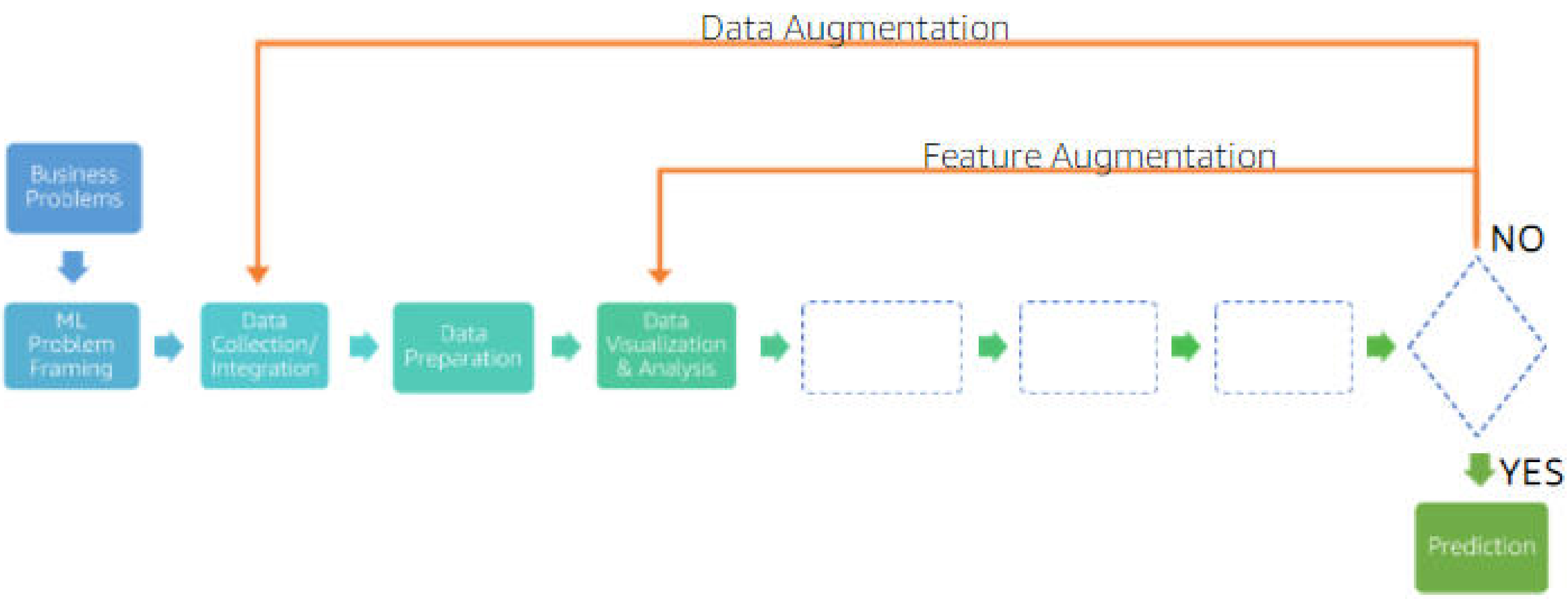
Leave-one-out



K-fold

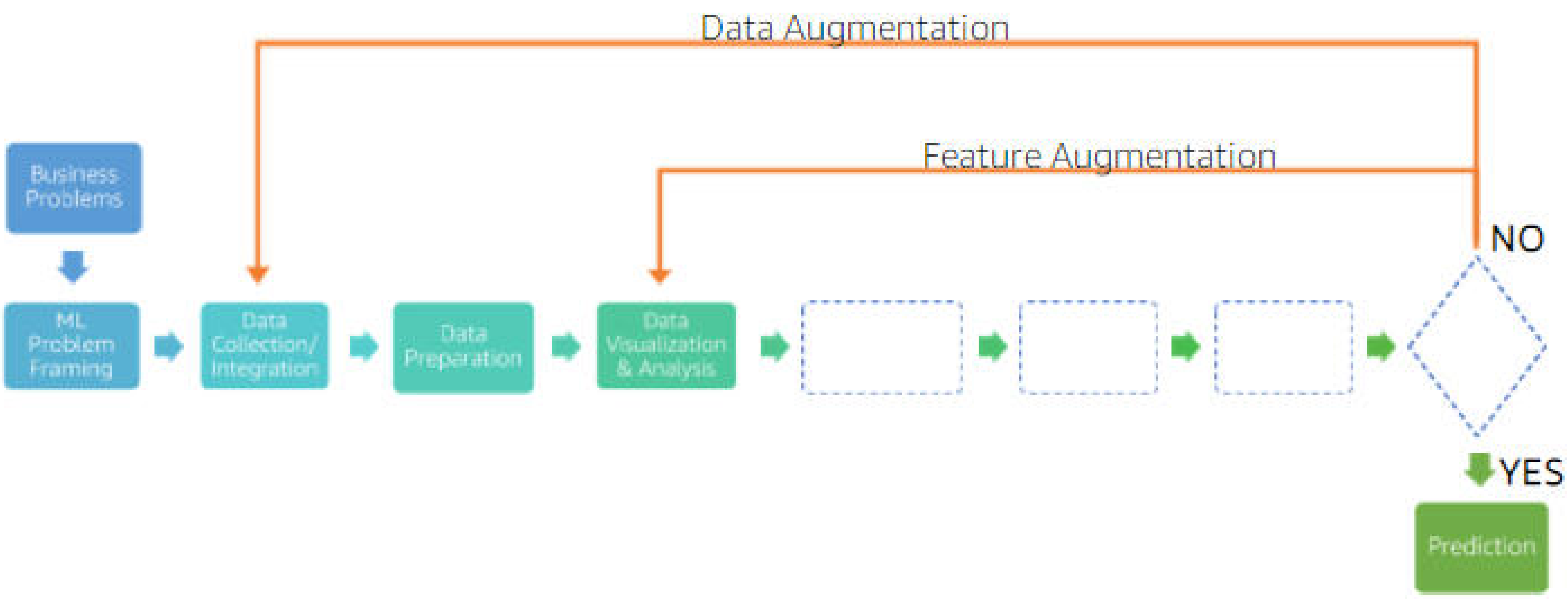
Step 5: Data Visualization & Analysis

aws training and certification



Step 5: Data Visualization & Analysis

aws training and certification



Data Visualization & Analysis

aws training and certification

Feature: An attribute in your training dataset.

Features

NOT a feature

Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bachelors	14	Single	Adm-clerical	Male	0
31	Masters	18	Married	Engineer	Female	1

Data Visualization & Analysis

aws training and certification

Types of Visualization & Analysis:

- Statistics
- Scatter-plots
- Histograms

Feature & Target Summary

aws training and certification

Numerical

In [24]:
```
train_data.describe()
```

Out[24]:

	age	capital-gain	capital-loss	hours-per-week
count	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	1077.648844	87.303830	40.437456
std	13.640433	7385.292085	402.960219	12.347429
min	17.000000	0.000000	0.000000	1.000000
25%	28.000000	0.000000	0.000000	40.000000
50%	37.000000	0.000000	0.000000	40.000000
75%	48.000000	0.000000	0.000000	45.000000
max	90.000000	99999.000000	4356.000000	99.000000

Categorical

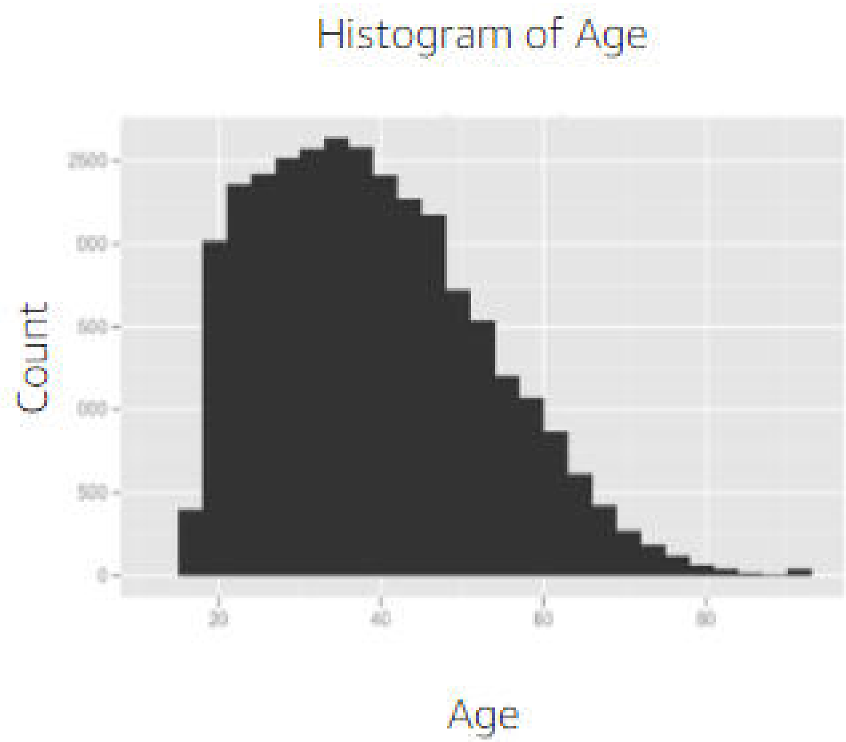
In [25]:
```
for variable in categorical_variables:
  print ("-------------------------------")
  print ("Histogram for " + variable)
  print ("-------------------------------")
  print (train_data[variable].value_counts())
  print ("")
```

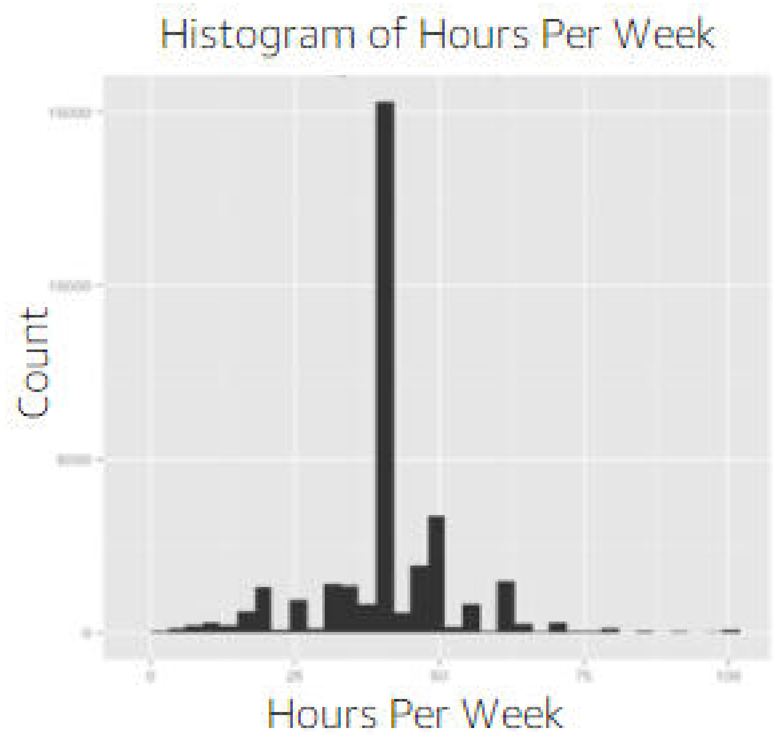
```
-------------------------------
Histogram for workclass
-------------------------------
Private              24532
Self-emp-not-inc      2541
Local-gov             2093
State-gov             1298
Self-emp-inc          1116
Federal-gov            960
Without-pay             14
Never-worked             7
Name: workclass, dtype: int64

-------------------------------
Histogram for education
-------------------------------
HS-grad         10501
Some-college     7291
```

Feature & Target Histograms

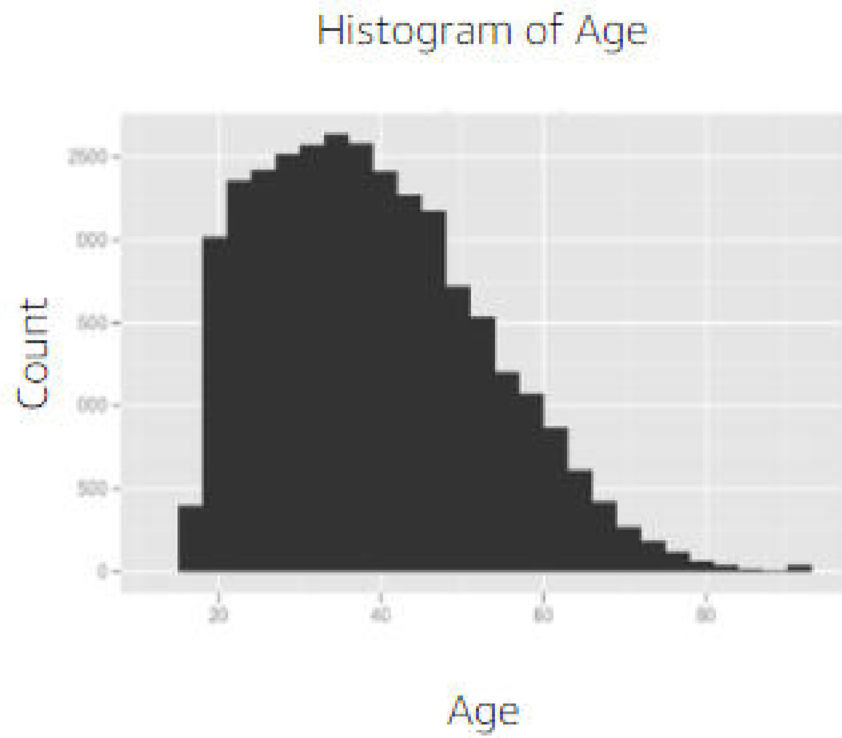
Histograms can help detect skews.

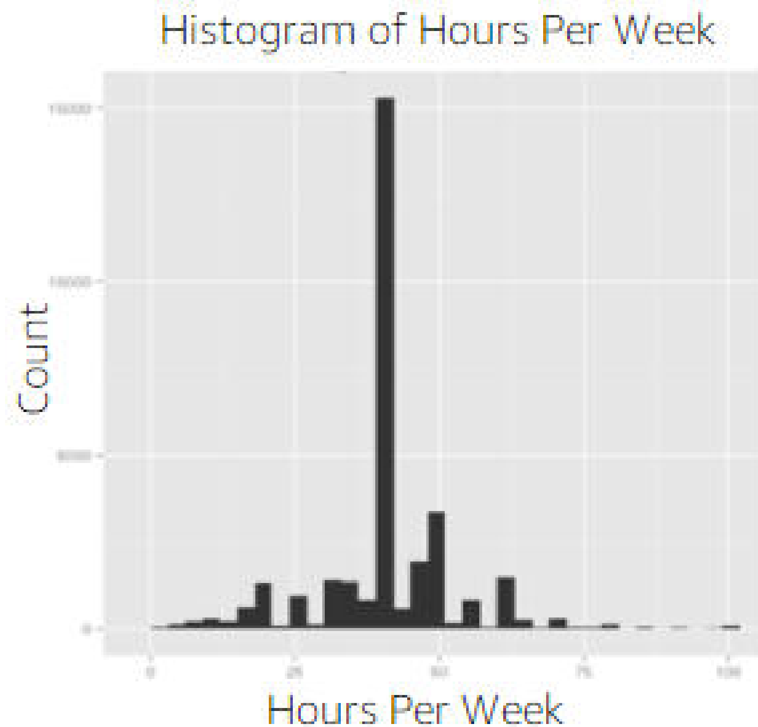




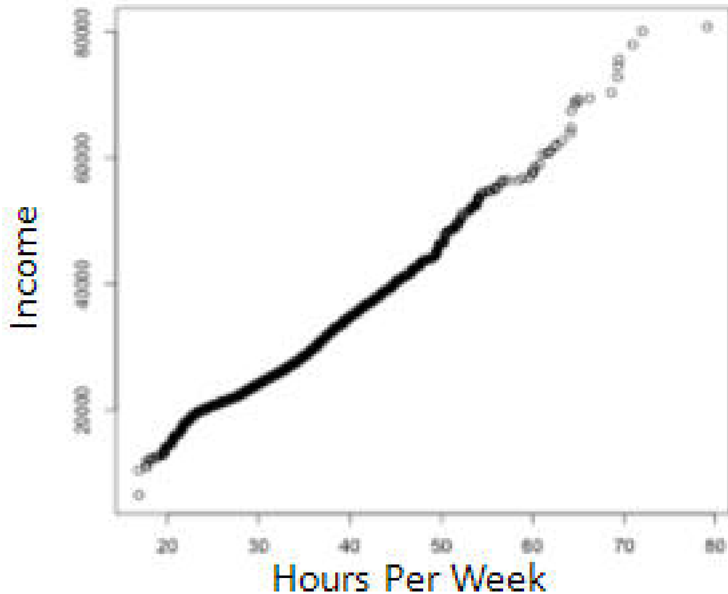
Feature & Target Histograms

Histograms can help detect skews.

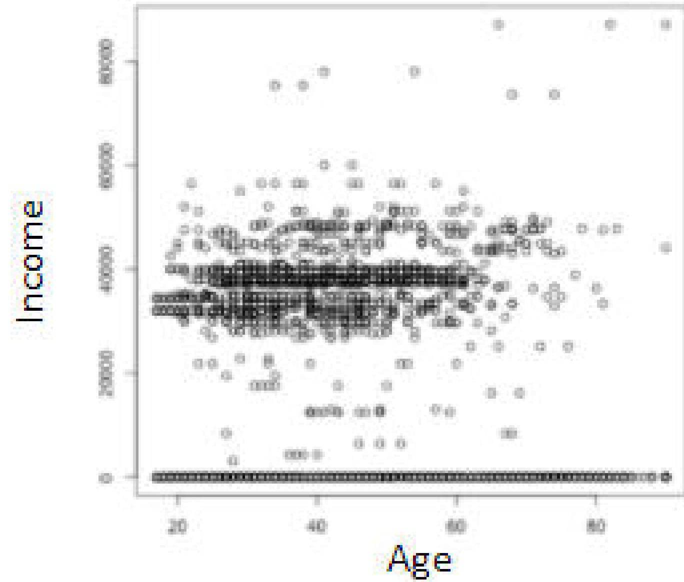




Feature-Target Correlation: Scatter Plots

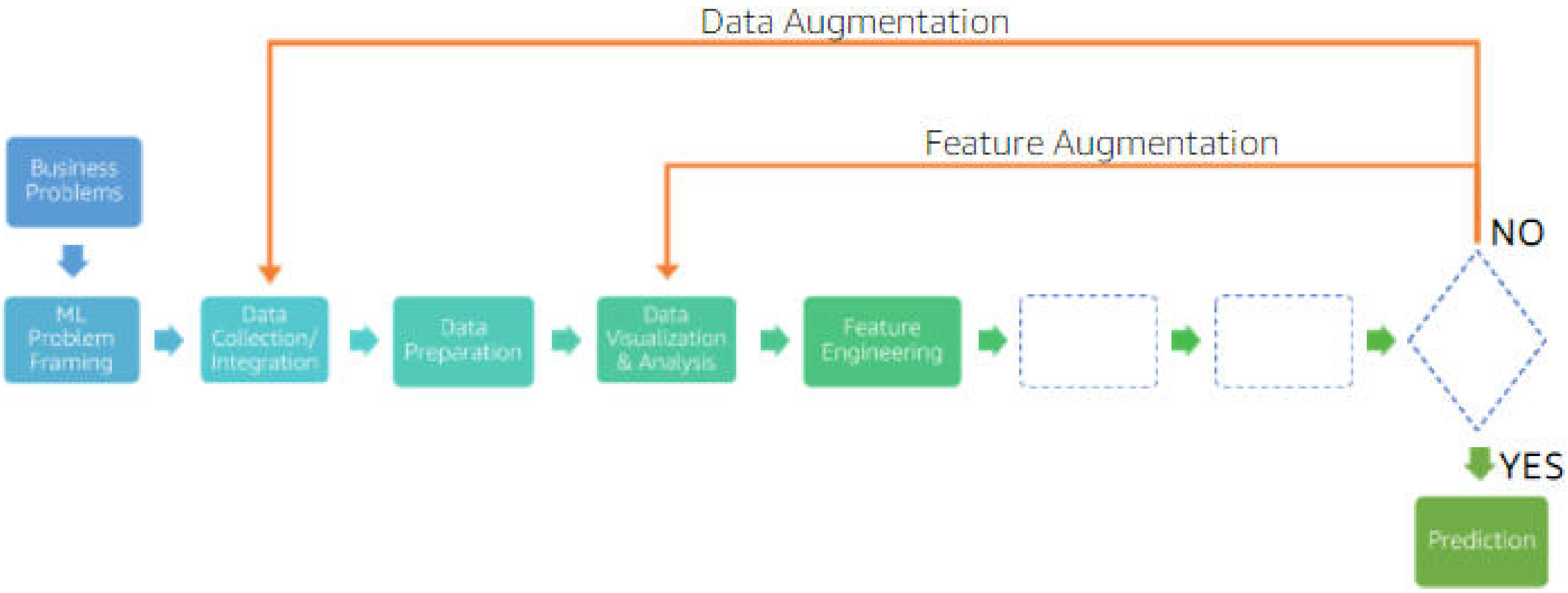


Hours per week is **strongly** correlated with income!



Age is **weakly** correlated with income!

Step 6: Feature Engineering



Feature Engineering

Converts raw data into a higher representation

Feature Engineering

aws training and certification

Converts raw data into a higher representation

Numeric Value Binning

aws training and certification

To introduce non-linearity into linear models, intelligently break up continuous values using **binning**.

Age	Binned Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bin1	Bachelors	14	Single	Adm-clerical	Male	-1
31	Bin2	Masters	18	Married	Engineer	Female	+1
44	Bin3	Bachelors	16	Married	Accounting	Male	-1
62	Bin4	Bachelors	14	Married	Engineer	Female	-1

Numeric Value Binning

To introduce non-linearity into linear models, intelligently break up continuous values using **binning**.

Age	Binned Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bin1	Bachelors	14	Single	Adm-clerical	Male	-1
31	Bin2	Masters	18	Married	Engineer	Female	+1
44	Bin3	Bachelors	16	Married	Accounting	Male	-1
62	Bin4	Bachelors	14	Married	Engineer	Female	-1

20 40 60

Binned Age: Bin1 Bin2 Bin3 Bin4

Quadratic Features

Derive new non-linear features by combining feature pairs.

Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bachelors	14	Single	Business	Male	-1
31	Masters	18	Married	Business	Female	+1
44	Bachelors	16	Married	Accounting	Male	-1
62	Masters	14	Married	Engineer	Female	-1

Quadratic Features

Derive new non-linear features by combining feature pairs.

Age	Education	Years of education	Marital status	Occupation	Sex	Label
19	Bachelors	14	Single	Business	Male	-1
31	Masters	18	Married	Business	Female	+1
44	Bachelors	16	Married	Accounting	Male	-1
62	Masters	14	Married	Engineer	Female	-1

Quadratic Features

aws training and certification

Derive new non-linear features by combining feature pairs.

Age	Education	Years of education	Marital status	Occupation	Sex	Education + Occupation	Label
39	Bachelors	16	Single	Business	Male	Bachelors_Business	-1
31	Masters	18	Married	Business	Female	Masters_Business	+1
44	Bachelors	16	Married	Accounting	Male	Bachelors_Accounting	-1
62	Masters	14	Married	Engineer	Female	Masters_Engineer	-1

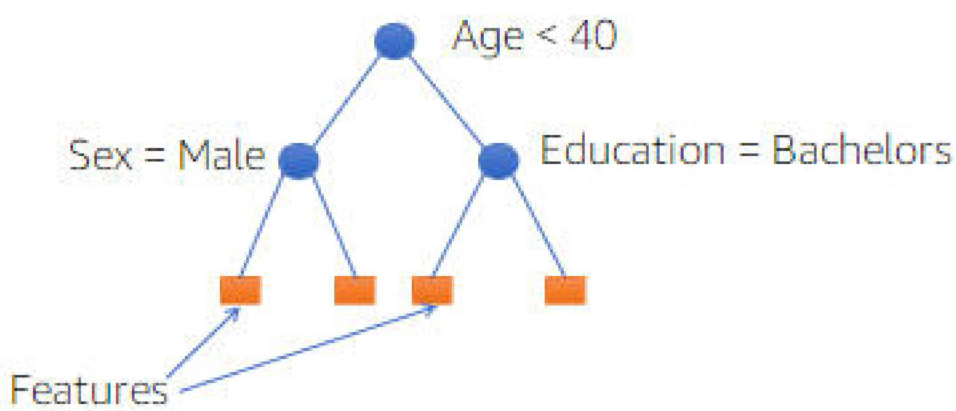
Quadratic feature over Education and Occupation

Non-Linear Feature Transformations

For numeric features:

- Log, polynomial power of target variable, feature values - may ensure a more "linear dependence" with output variable
- Product/ratio of feature values

Tree path features: use leaves of decision tree as features:

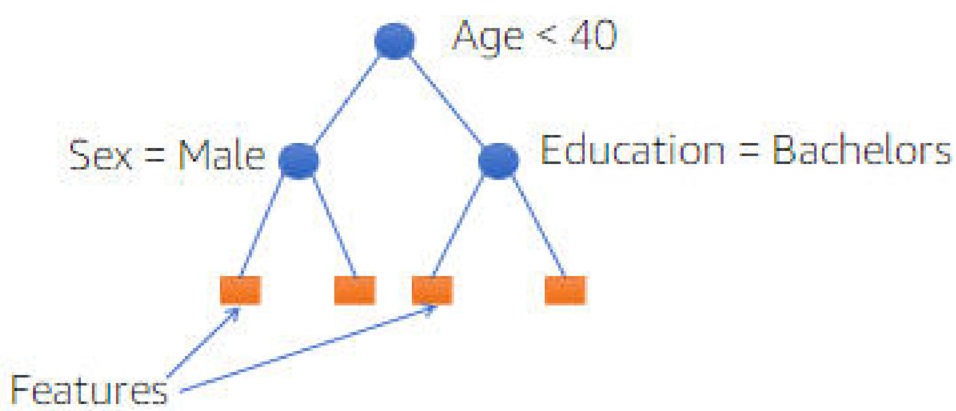


Non-Linear Feature Transformations

For numeric features:

- Log, polynomial power of target variable, feature values - may ensure a more "linear dependence" with output variable
- Product/ratio of feature values

Tree path features: use leaves of decision tree as features:



Domain-Specific Transformations

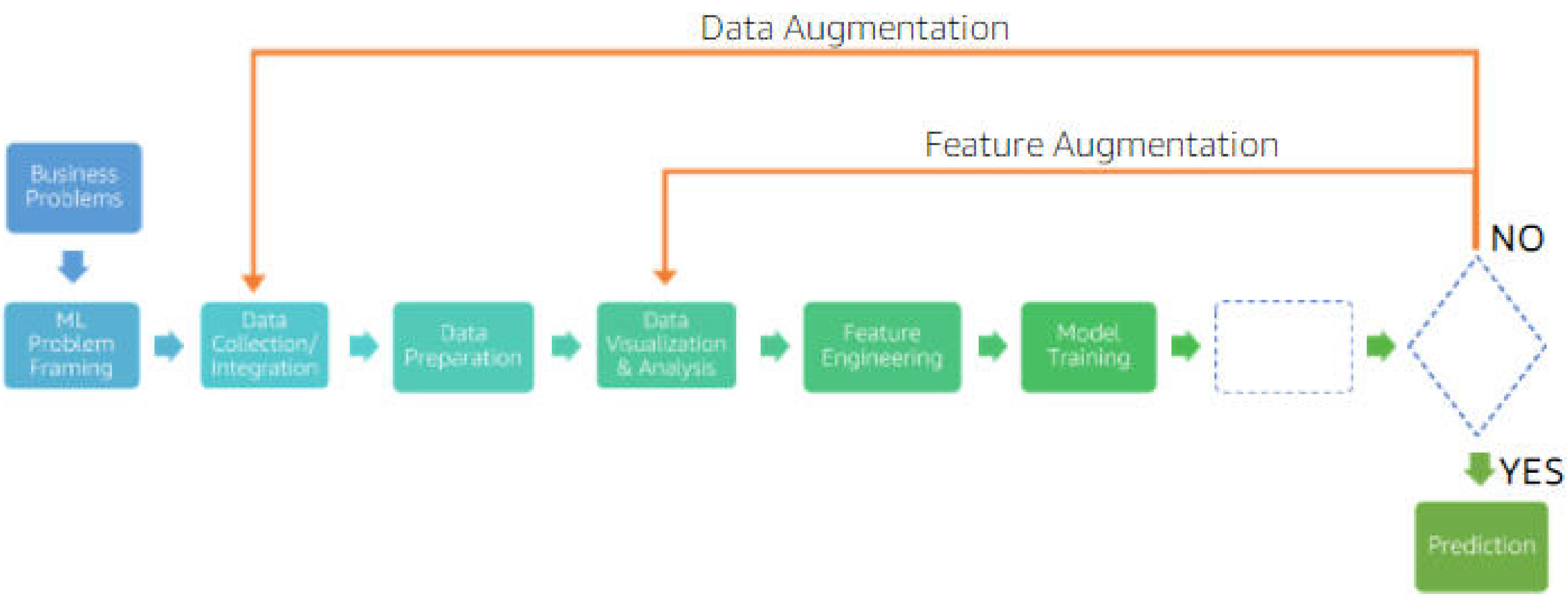
Text Features:

- Stop-words removal/Stemming
- Lowercasing, punctuation removal
- Cutting off very high/low percentiles
- TF-IDF normalization

Web-page features:

- Multiple fields of text: URL, in/out anchor text, title, frames, body, presence of certain HTML elements (tables/images)
- Relative style (italics/bold, font-size) & positioning

Step 7: Model Training



Parameter Tuning

aws training and certification

Loss Function

- Square: regression, classification
- Hinge: classification only, more robust to outliers
- Logistic: classification only, better for skewed class distributions

Regularization

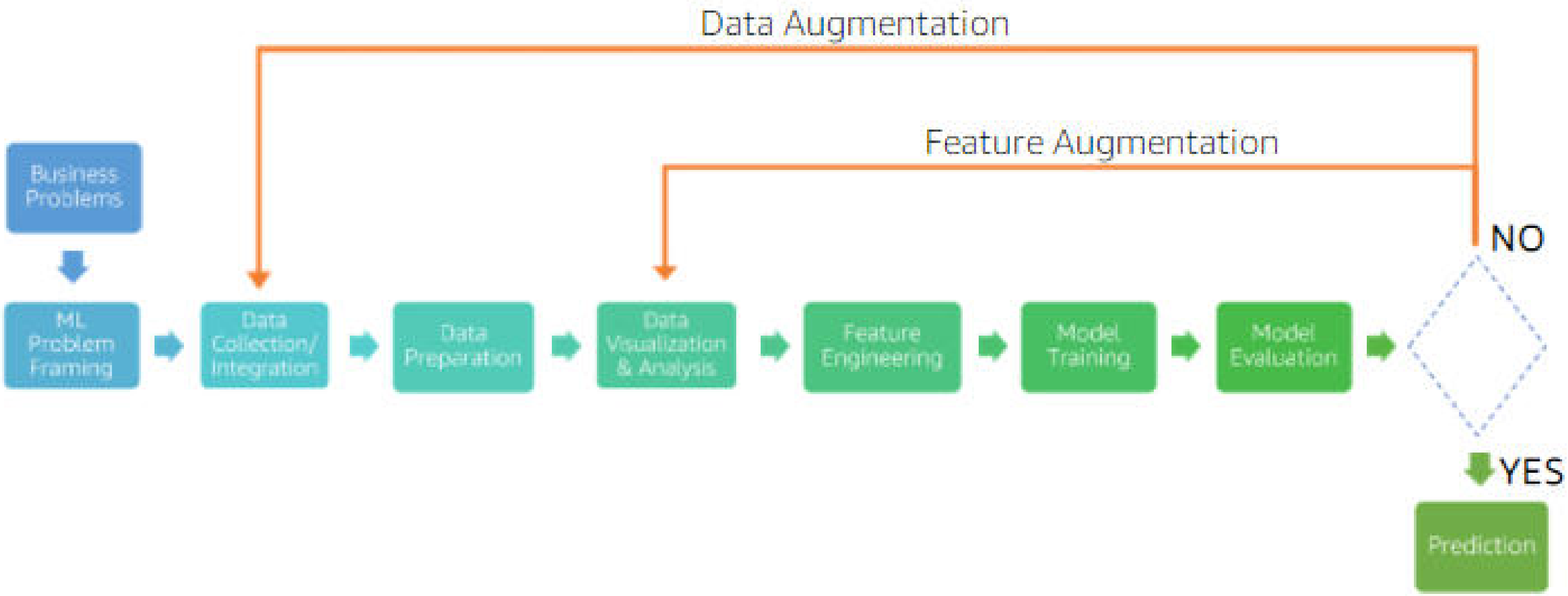
- Prevent overfitting by constraining weights to be small

Learning Parameters (e.g. decay rate)

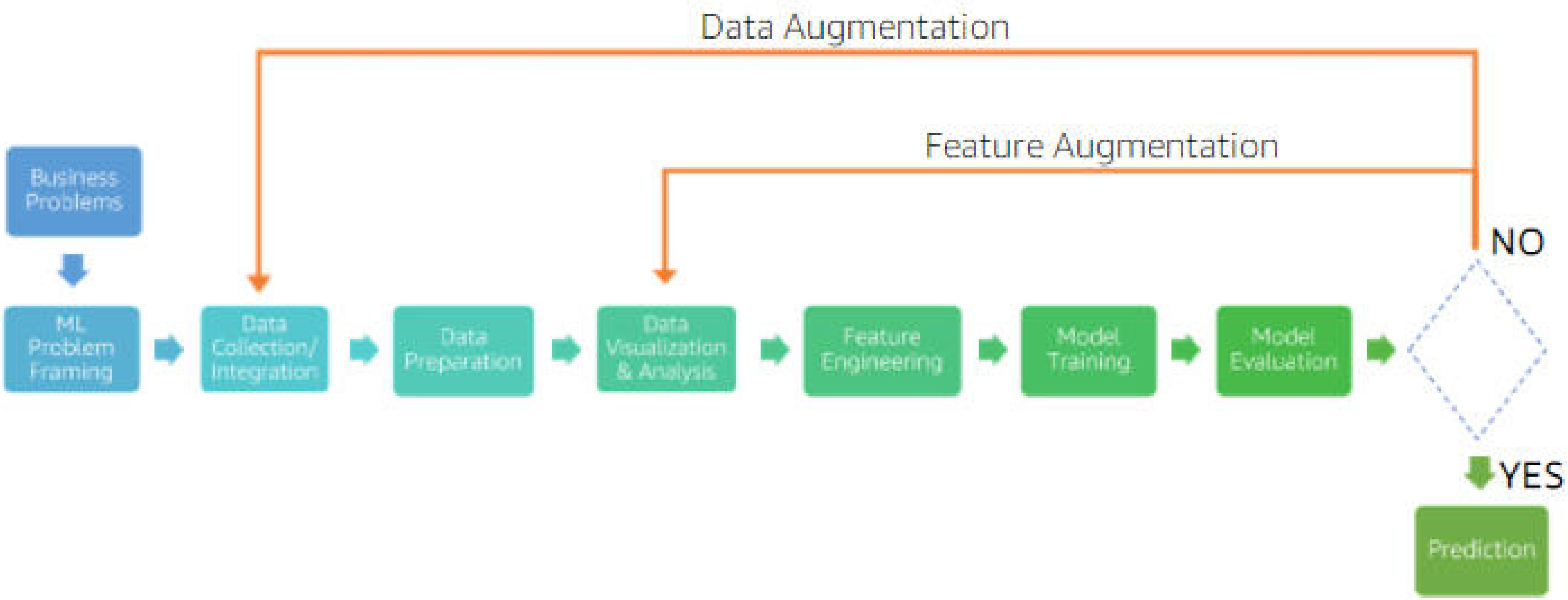
- Decaying too aggressively – algorithm never reaches optimum
- Decaying too slowly – algorithm bounces around, never converges to optimim

Step 8: Model Evaluation

aws training and certification



Step 8: Model Evaluation

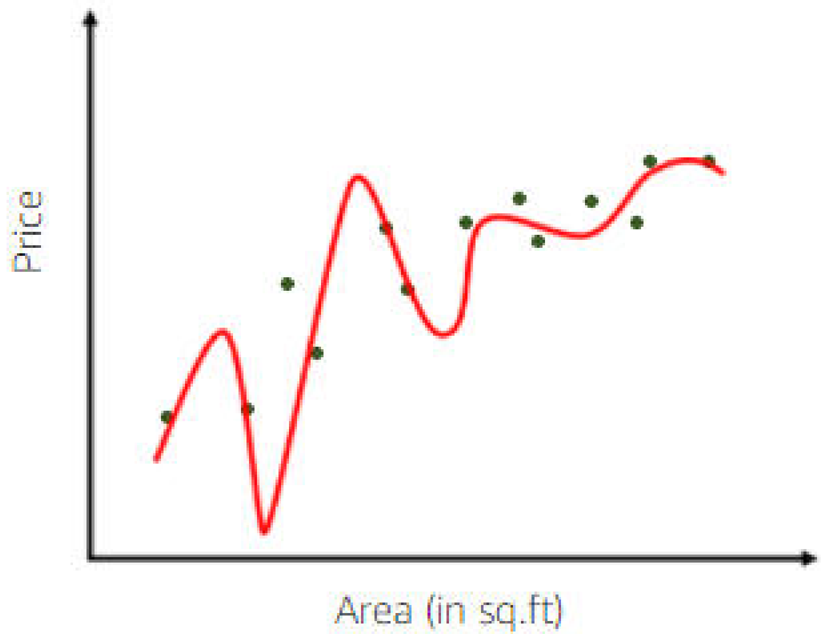


Overfitting & Underfitting

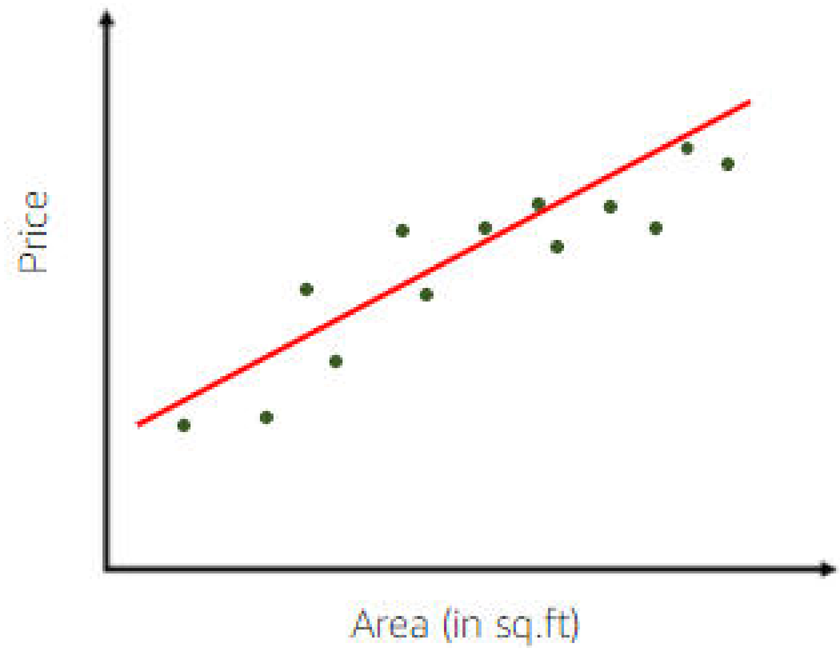
Don't fit your training data to obtain maximum accuracy.

Overfitting & Underfitting

Don't fit your training data to obtain maximum accuracy.

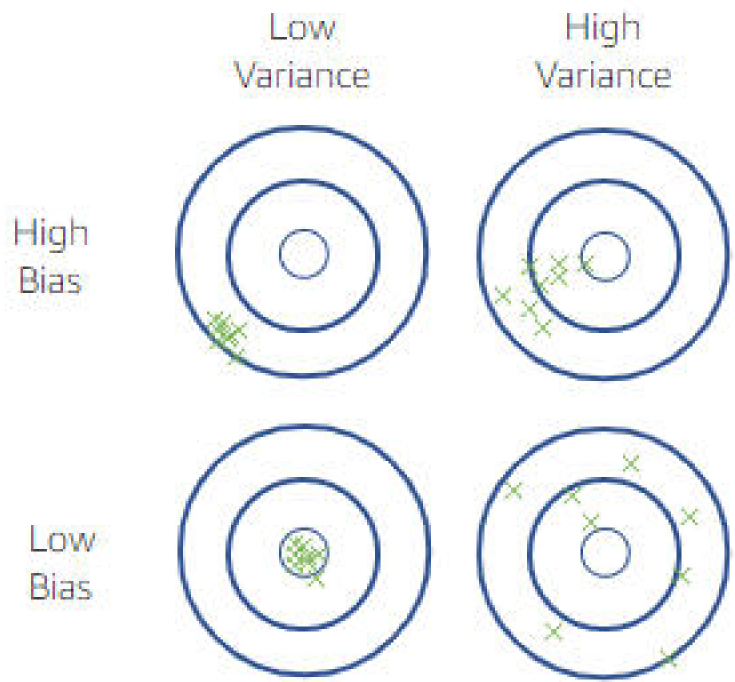


Overfitting



Underfitting

Bias-Variance Tradeoff



Evaluation Metrics

aws training and certification

Metrics when regression is used for predicting target values:

- Root Mean Square Error (RMSE)
- MAPE (Mean Absolute Percent Error)
- R^2: How much better is the model compared to just picking the best constant?

$$R^2 = 1 - (\text{Model Mean Squared Error} / \text{Variance})$$

Evaluation Metrics

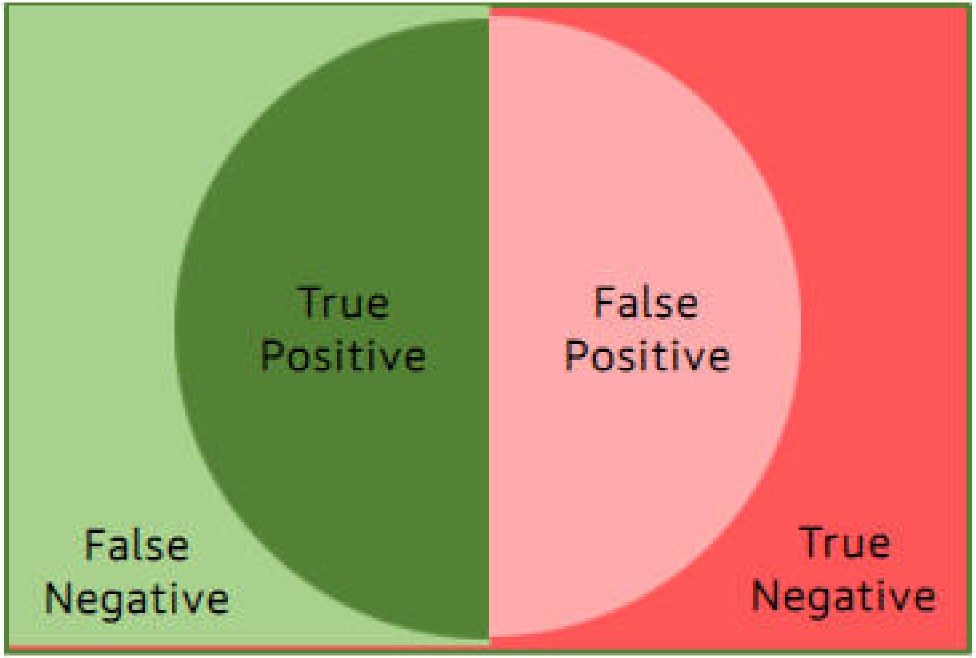
Metrics when classification is used for predicting target classes:

- Confusion Matrix
- ROC Curve
- Precision-Recall

	Actual +1	Actual -1
Predicted +1	True Positive	False Positive
Predicted -1	False Negative	True Negative

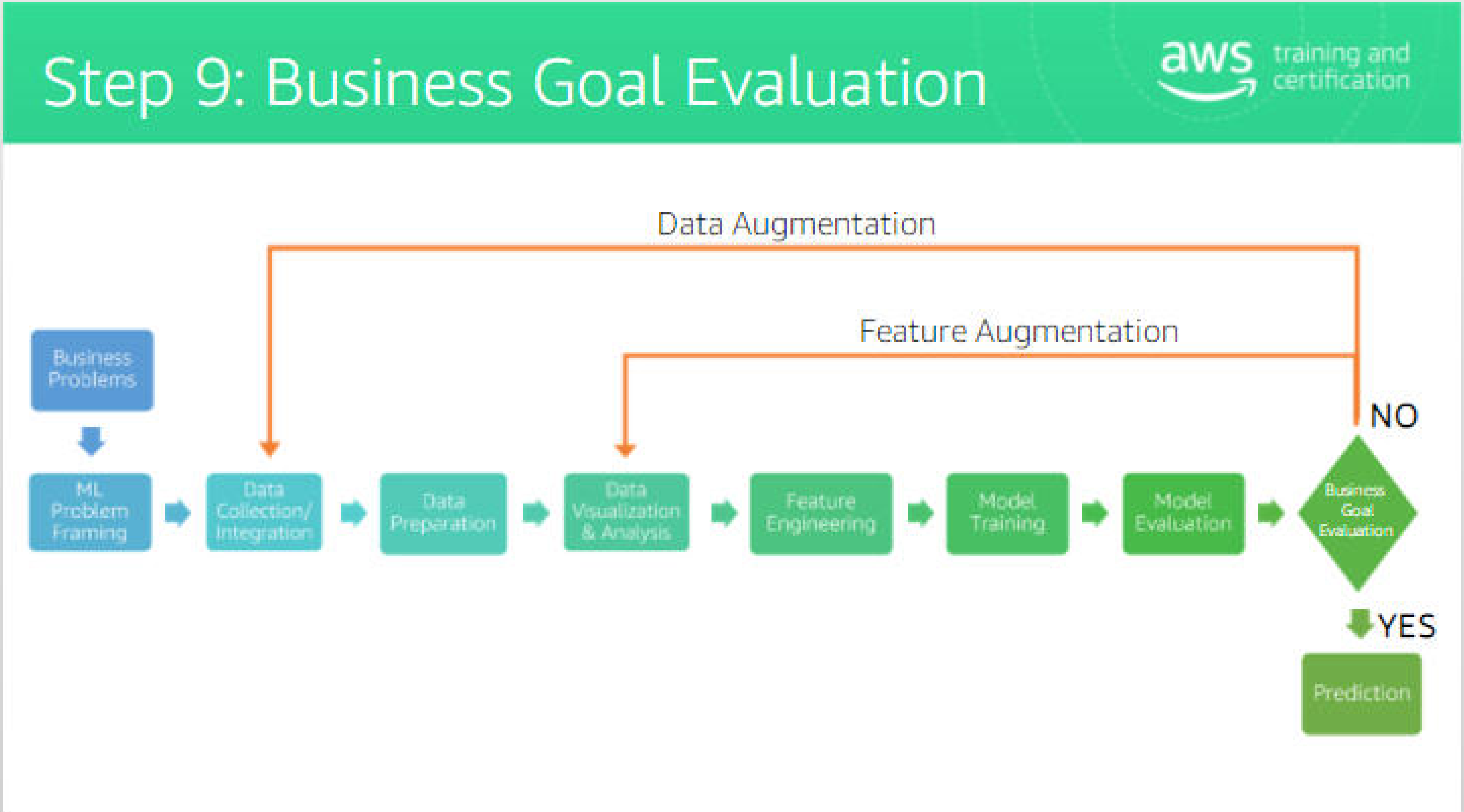
Precision – Recall

$$Precision = \frac{TP}{(TP + FP)}$$



$$Recall = \frac{TP}{(TP + FN)}$$

Machine Learning Terminology and Process

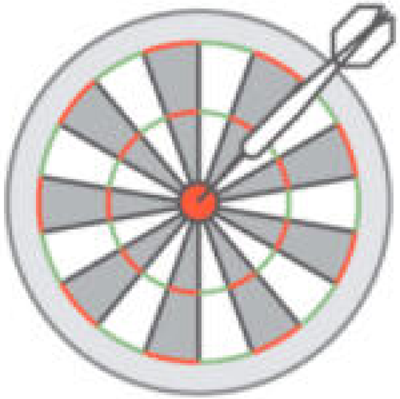


Business Goal Evaluations

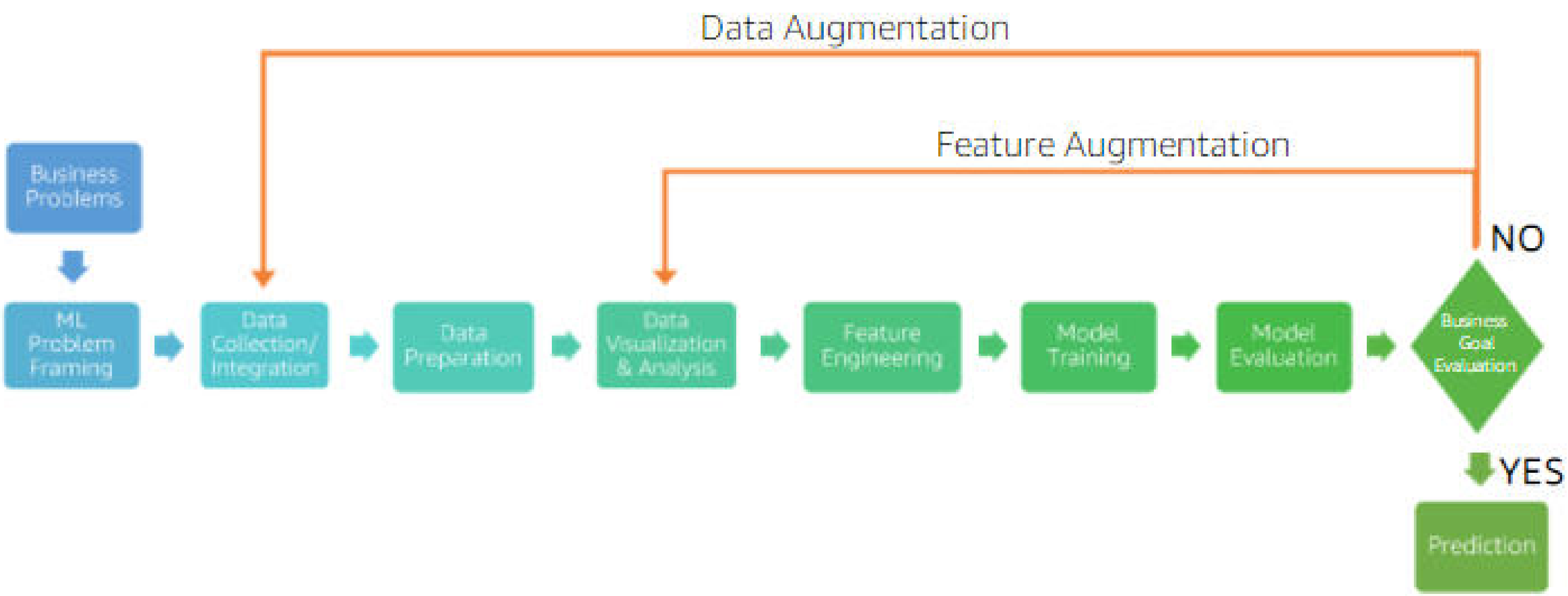
1. Evaluate how the model is performing related to business goals.
2. Make the final decision to deploy or not.

Evaluation depends on:

- Accuracy
- Model generalization on unseen/unknown data
- Business success criteria

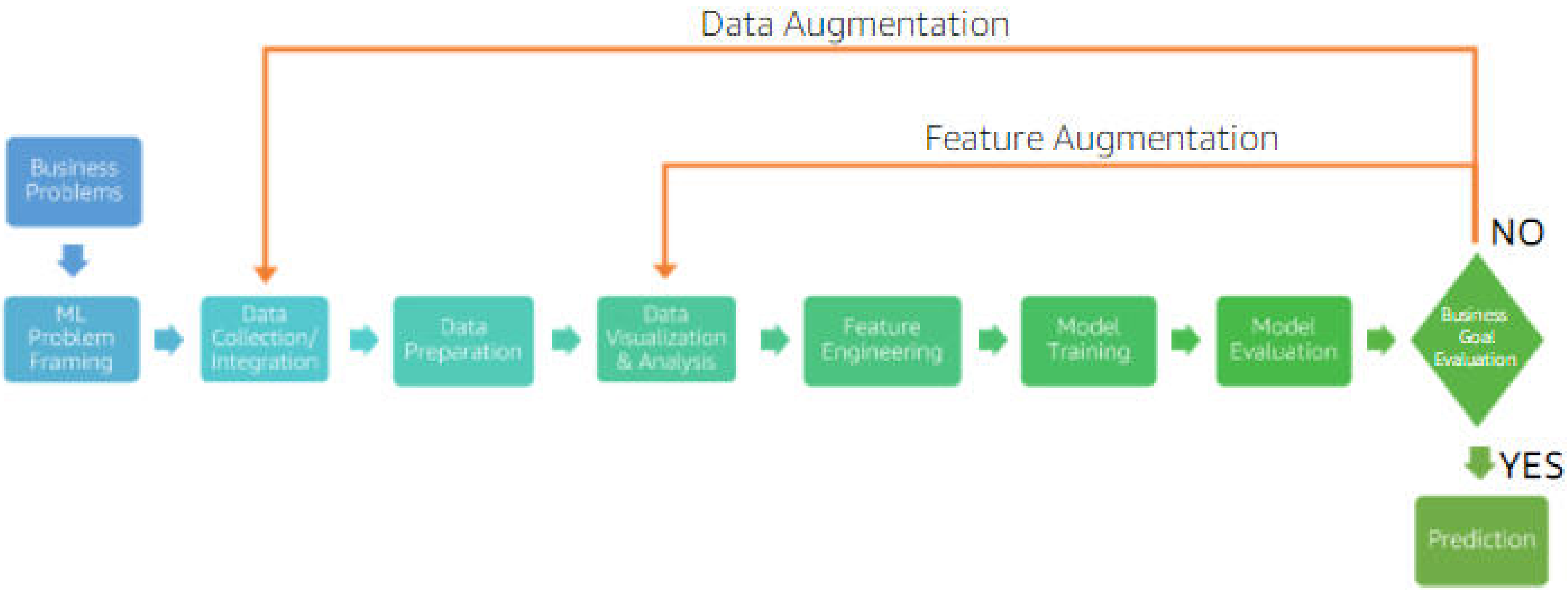


Augmenting Your Data

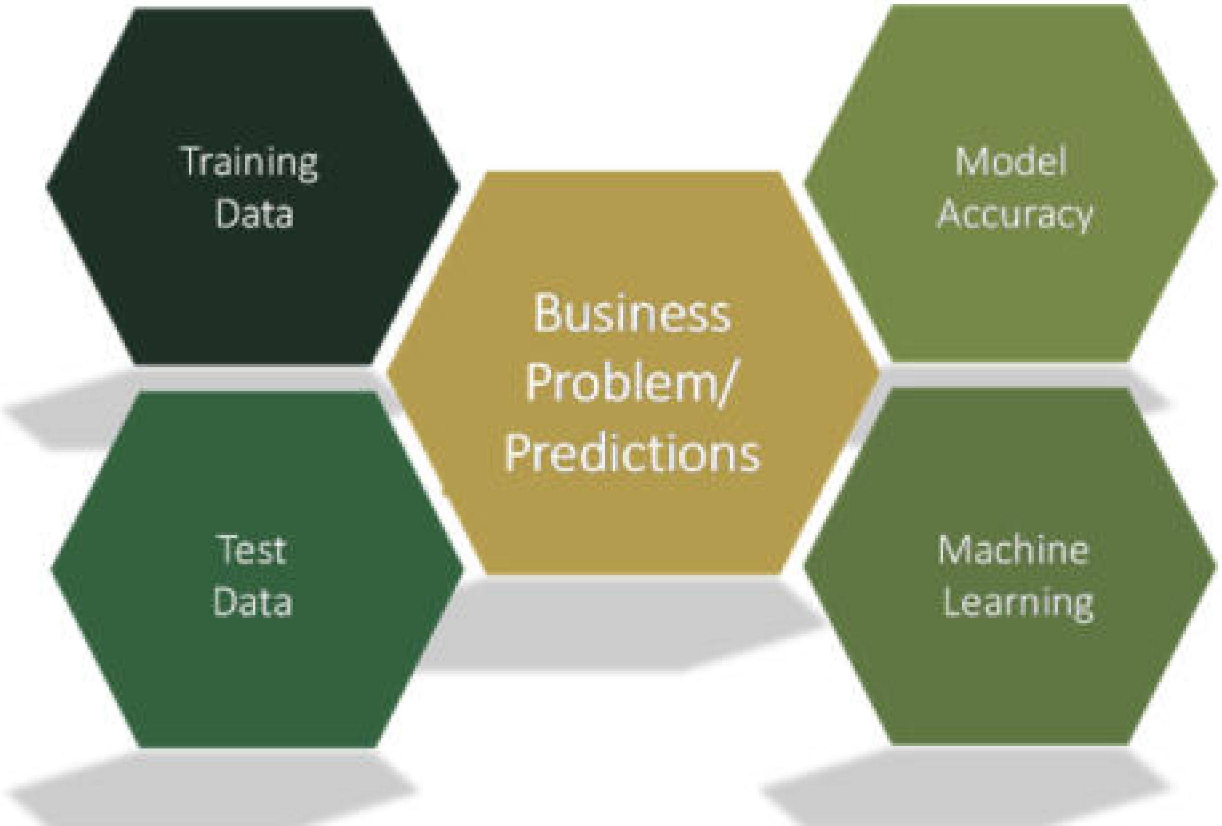


Machine Learning Terminology and Process

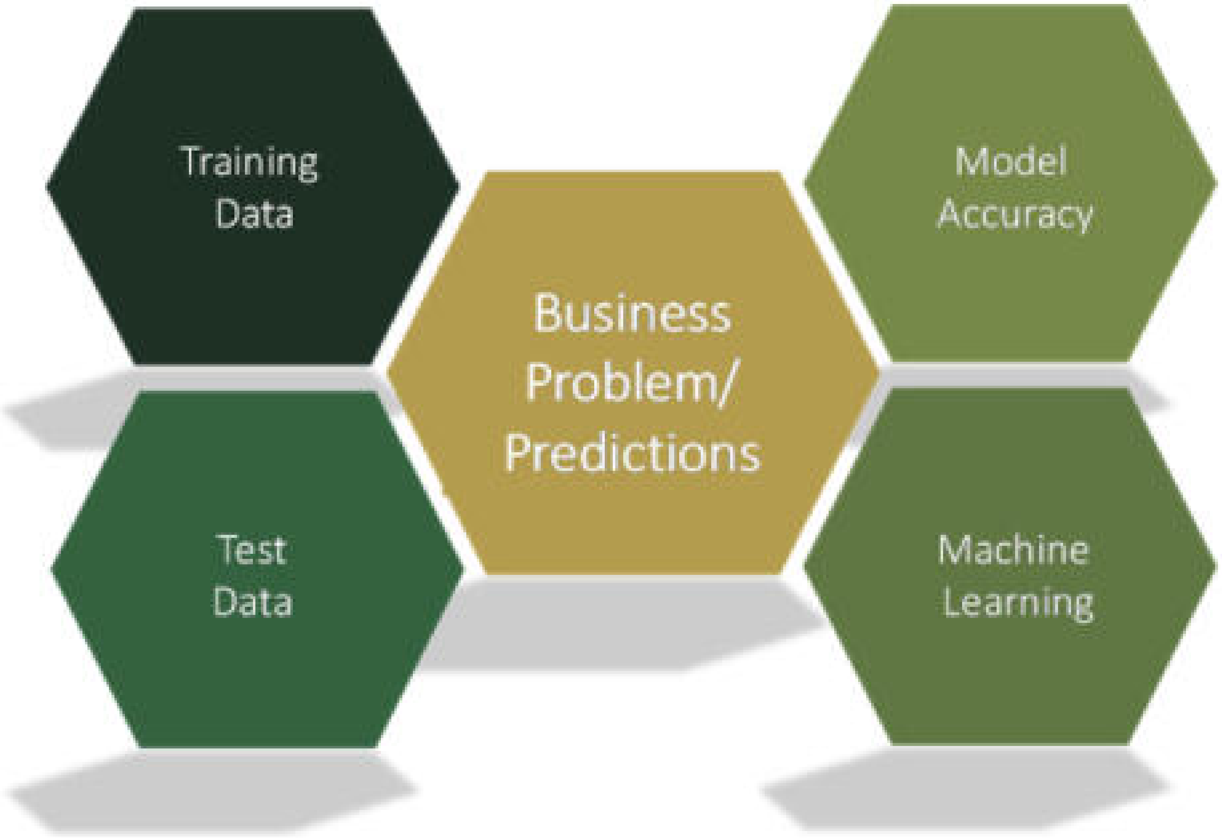
Prediction



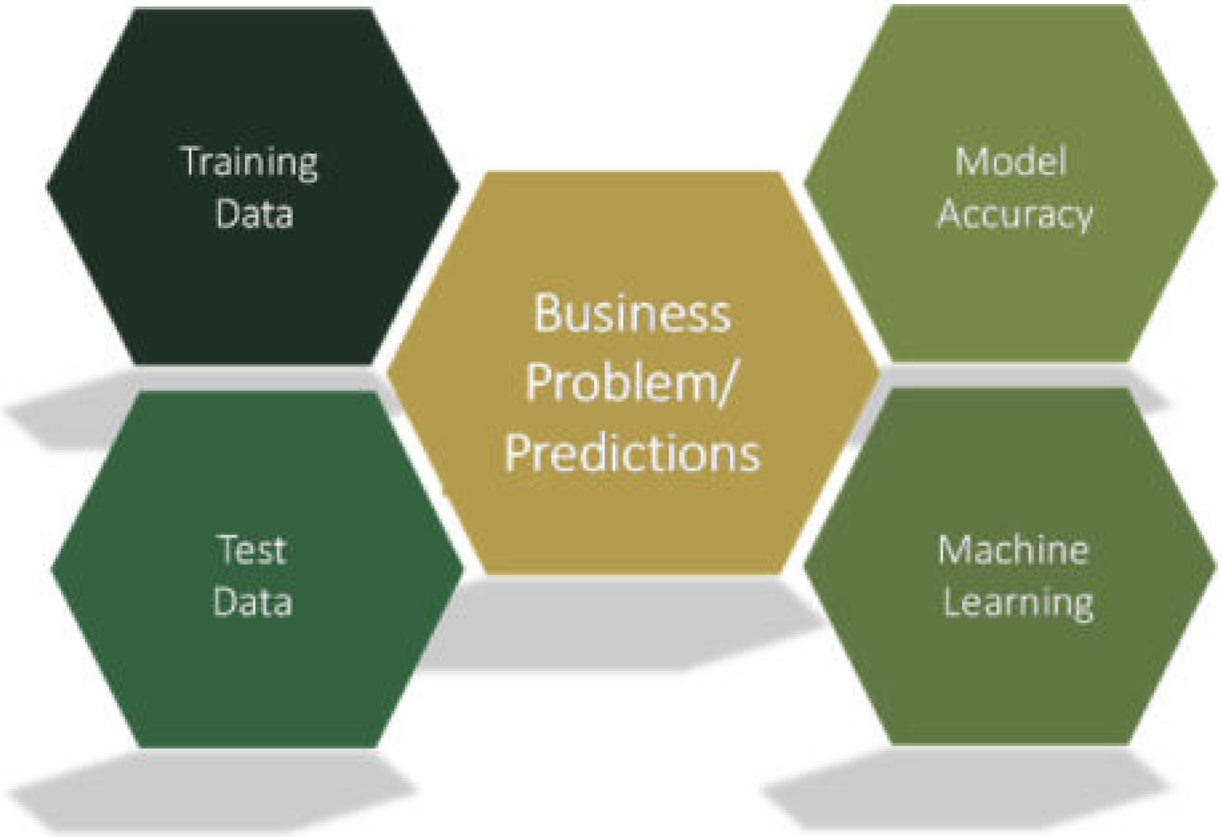
Summary


Training Data
Model Accuracy
Business Problem/ Predictions
Test Data
Machine Learning

Summary


Training
Data
Business
Problem/
Predictions
Model
Accuracy
Test
Data
Machine
Learning

Summary


Training Data
Business Problem/ Predictions
Model Accuracy
Test Data
Machine Learning

Thanks for watching!

aws training and certification

aws training and certification

Certificate of Completion

Hem Bahadur Gurung

Has successfully completed

Machine Learning Terminology and Process

Director, Training and Certification	1 hour **Duration**	10 September, 2021 **Completion Date**

Director, Training and Certification