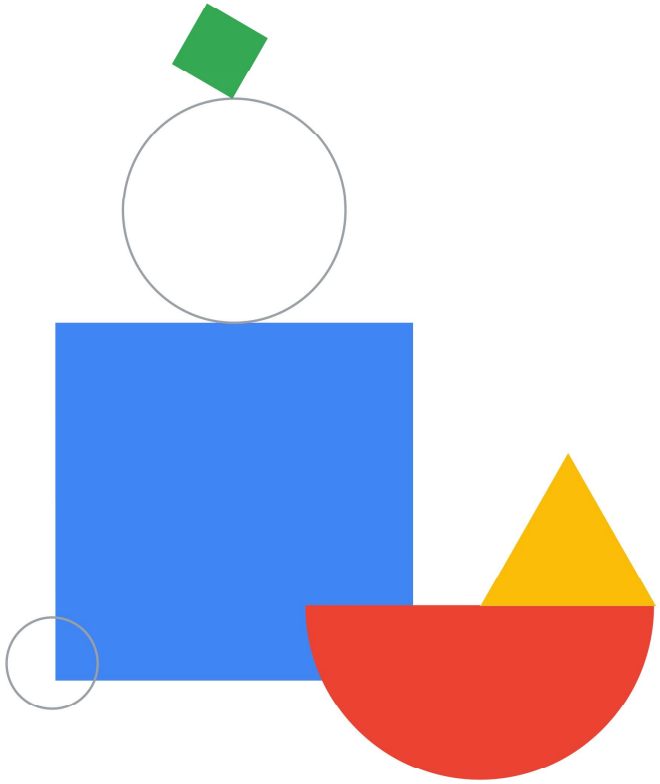
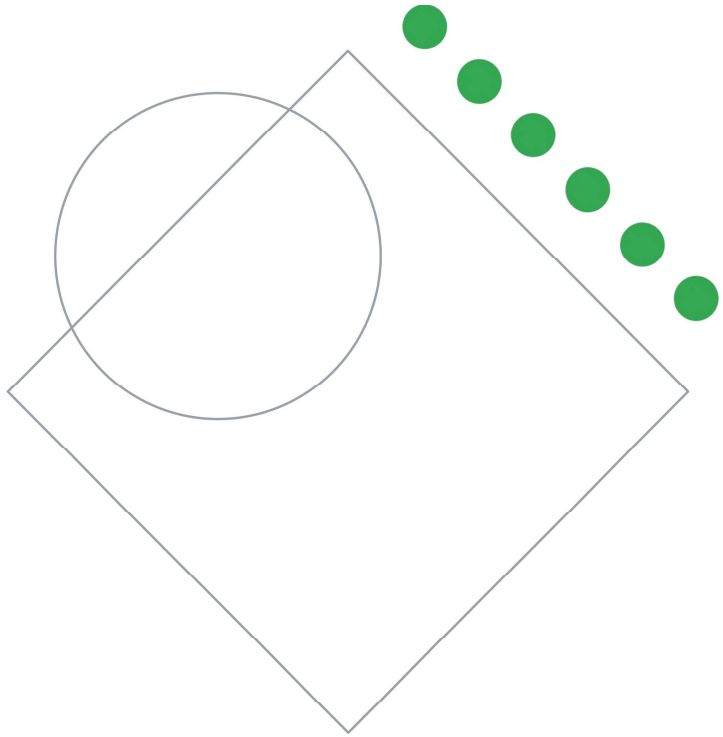
Google Cloud

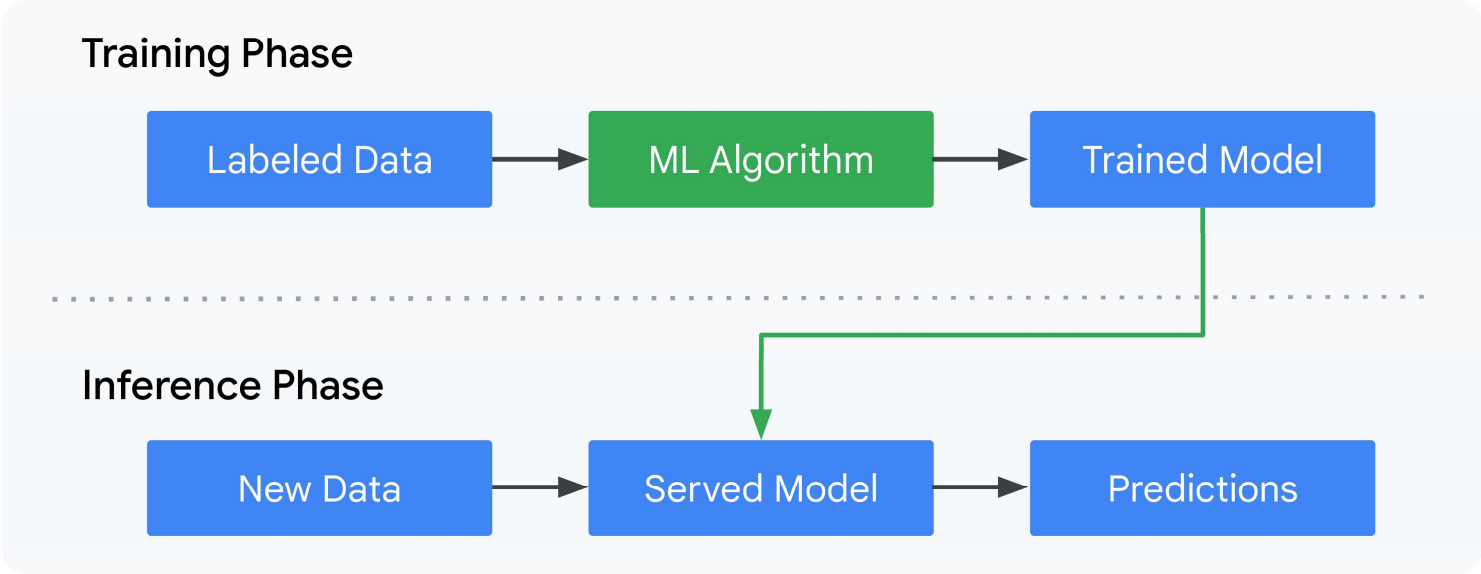
Launching into Machine Learning



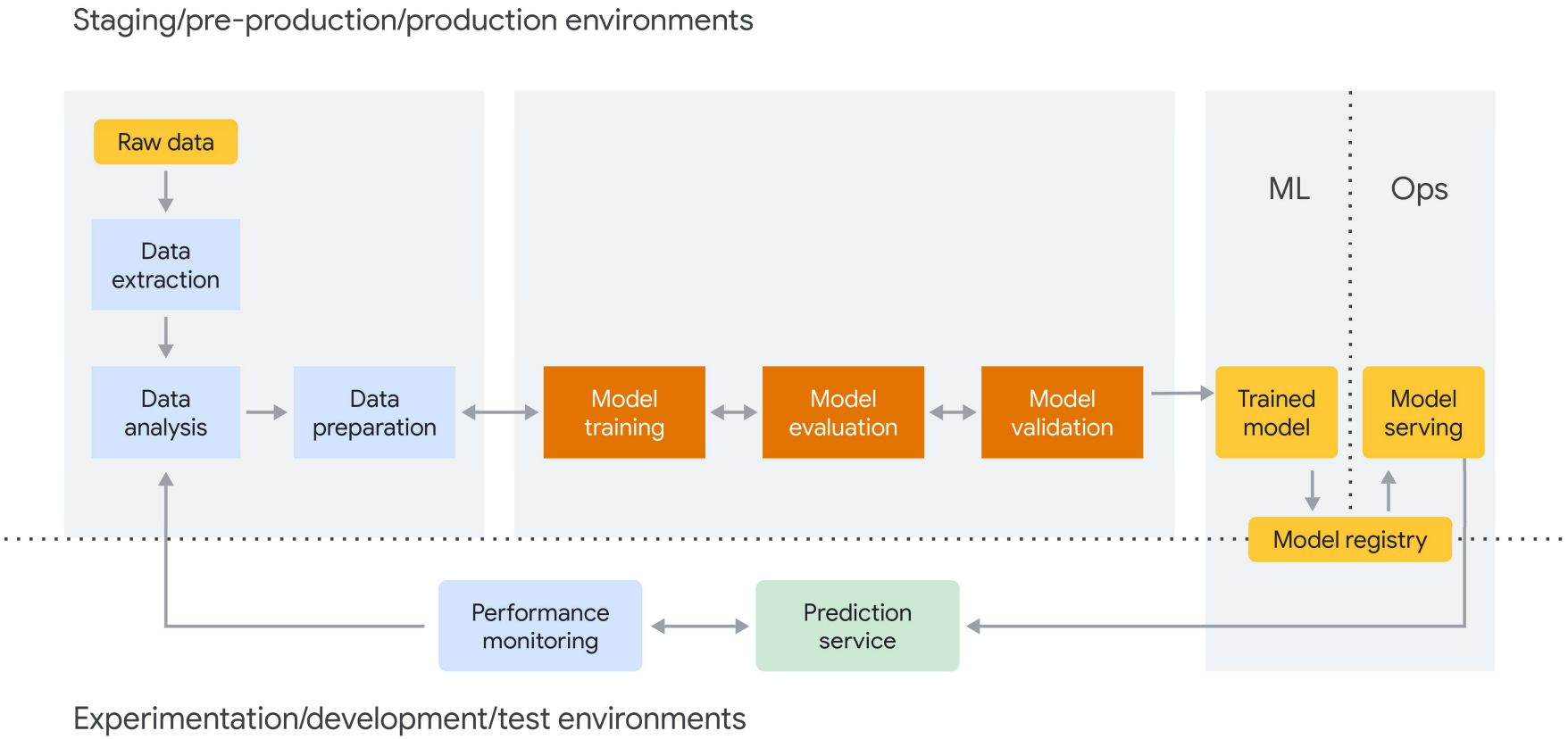
Get to know your data: Improve data through Exploratory Data Analysis



Machine learning phases



An ML pipeline recap



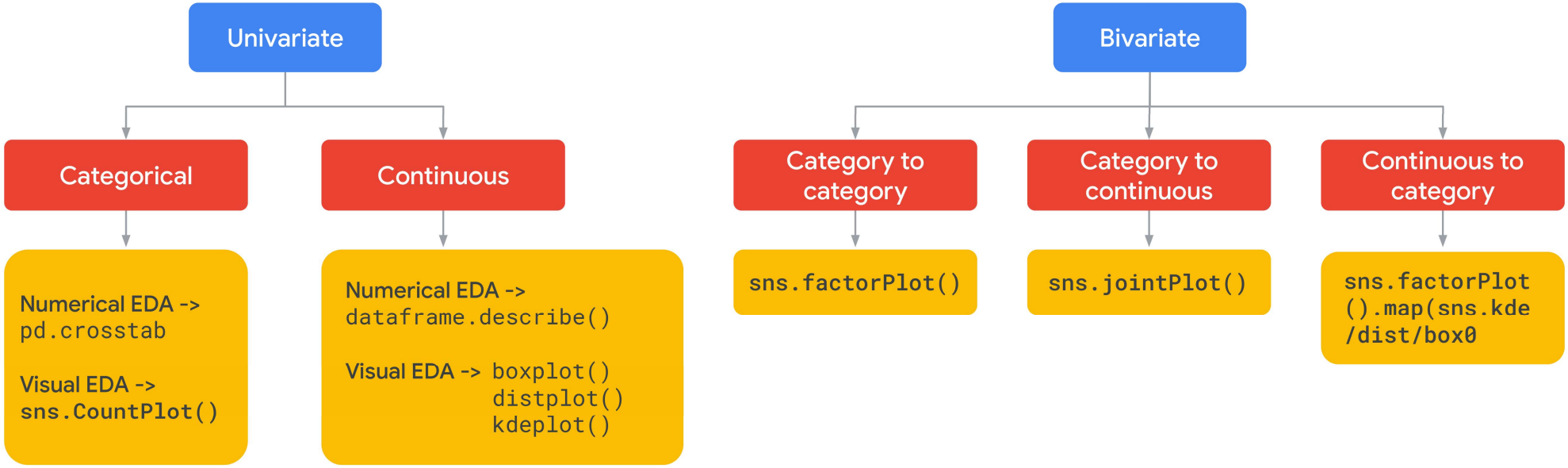
Attributes related to the data quality

1. Accuracy of Data
2. Consistency of Data
3. Timeliness of Data
4. Completeness of Data

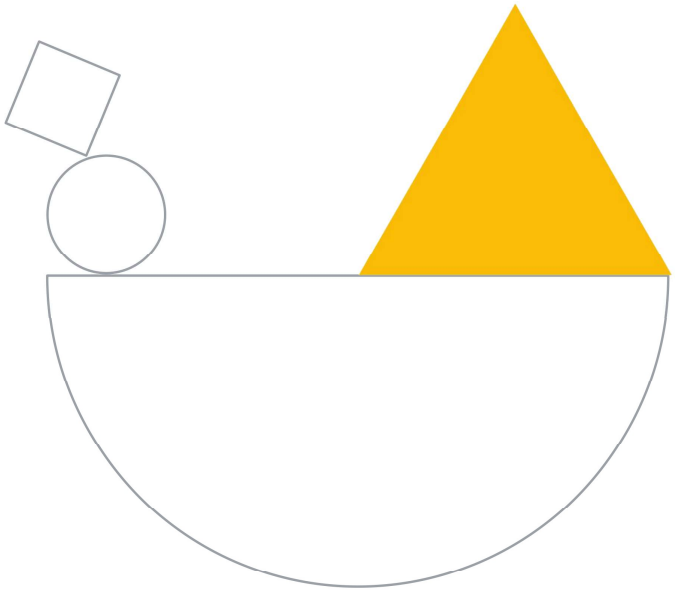
Ways to improve data quality

1. Resolve Missing Values
2. Convert the Date feature column to Datetime Format
3. Parse date/time features
4. Remove unwanted values
5. Convert categorical columns to “one-hot encodings”

Exploratory data analysis



Machine Learning in Practice

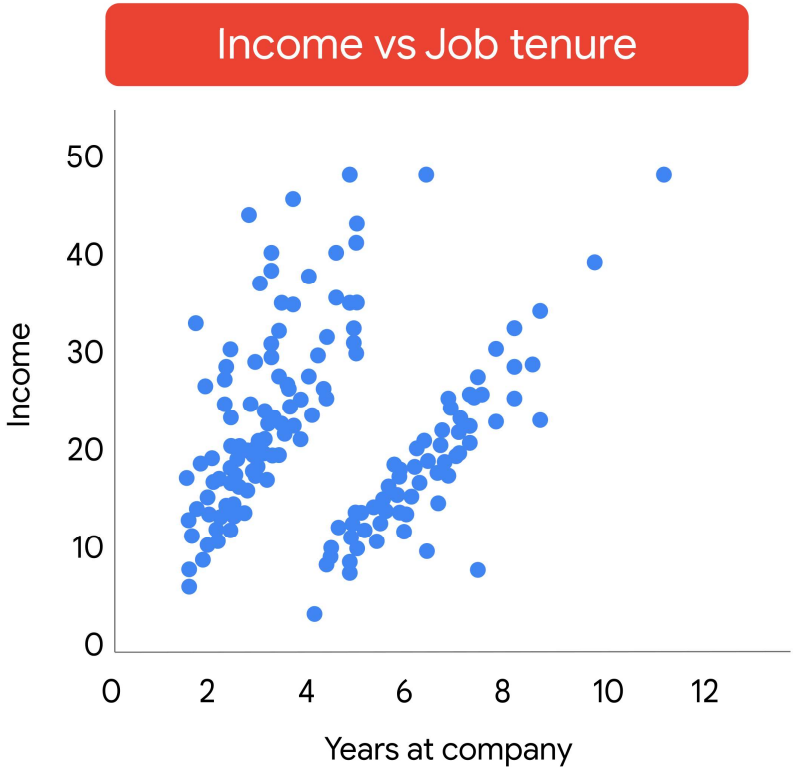


Unsupervised and supervised learning are the two types of ML algorithms

In unsupervised learning, data is not labeled

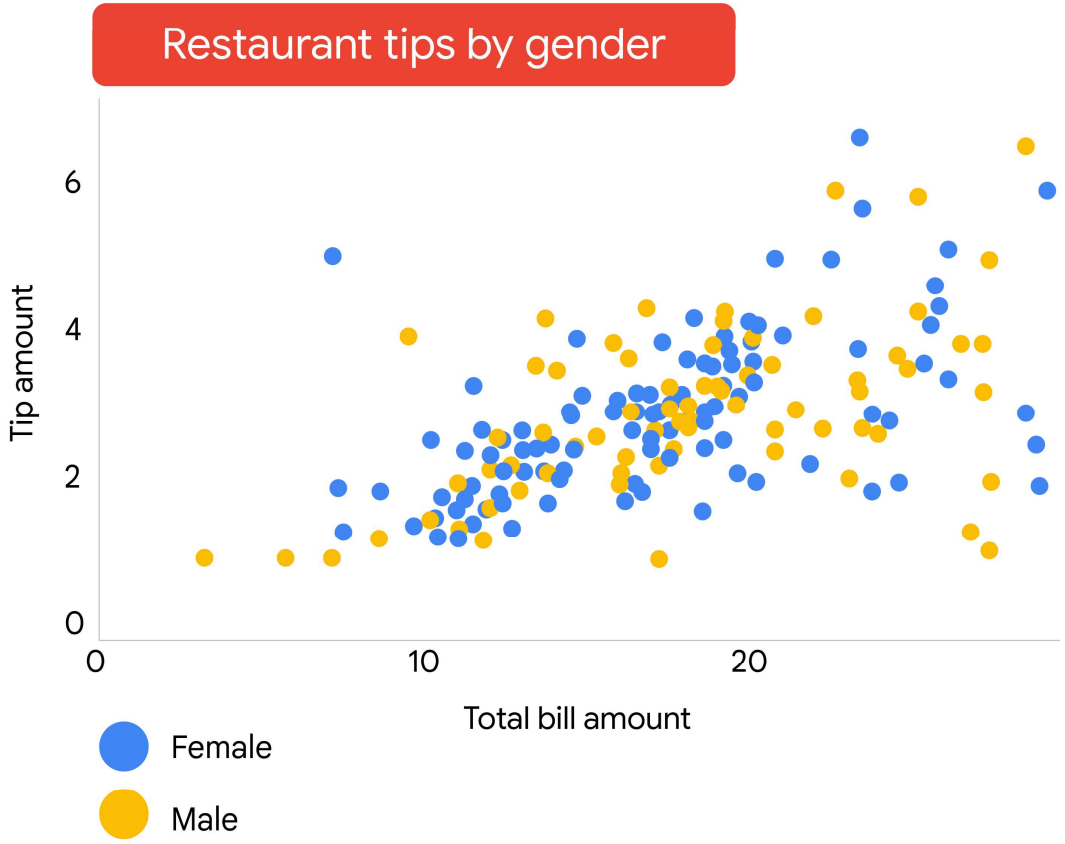
Example Model: Clustering

Is this employee on the “fast-track” or not?



Supervised learning implies the data is already labeled

In supervised learning we are learning from past examples to predict future values



Supervised ML model types:

Regression and classification

	total_bill	tip	sex	smoker	day	time
2	16.99	1.01	Female	No	Sun	Dinner
3	10.34	1.66	Male	No	Sun	Dinner
4	21.01	3.5	Male	No	Sun	Dinner
5	23.68	3.31	Male	No	Sun	Dinner
6	24.59	3.61	Female	No	Sun	Dinner
7	25.29	4.71	Male	No	Sun	Dinner
8	8.77	2	Male	No	Sun	Dinner
9	26.88	3.12	Male	No	Sun	Dinner

Supervised ML model types:

Regression and classification

	total_bill	tip	sex	smoker	day	time
2	16.99	1.01	Female	No	Sun	Dinner
3	10.34	1.66	Male	No	Sun	Dinner
4	21.01	3.5	Male	No	Sun	Dinner
5	23.68	3.31	Male	No	Sun	Dinner
6	24.59	3.61	Female	No	Sun	Dinner
7	25.29	4.71	Male	No	Sun	Dinner
8	8.77	2	Male	No	Sun	Dinner
9	26.88	3.12	Male	No	Sun	Dinner

Option 1
Regression Model
Predict the tip amount

Supervised ML model types:

Regression and classification

	total_bill	tip	sex	smoker	day	time
2	16.99	1.01	Female	No	Sun	Dinner
3	10.34	1.66	Male	No	Sun	Dinner
4	21.01	3.5	Male	No	Sun	Dinner
5	23.68	3.31	Male	No	Sun	Dinner
6	24.59	3.61	Female	No	Sun	Dinner
7	25.29	4.71	Male	No	Sun	Dinner
8	8.77	2	Male	No	Sun	Dinner
9	26.88	3.12	Male	No	Sun	Dinner

Option 1
Regression Model
Predict the tip amount

Option 2
Classification Model
Predict the sex of the customer

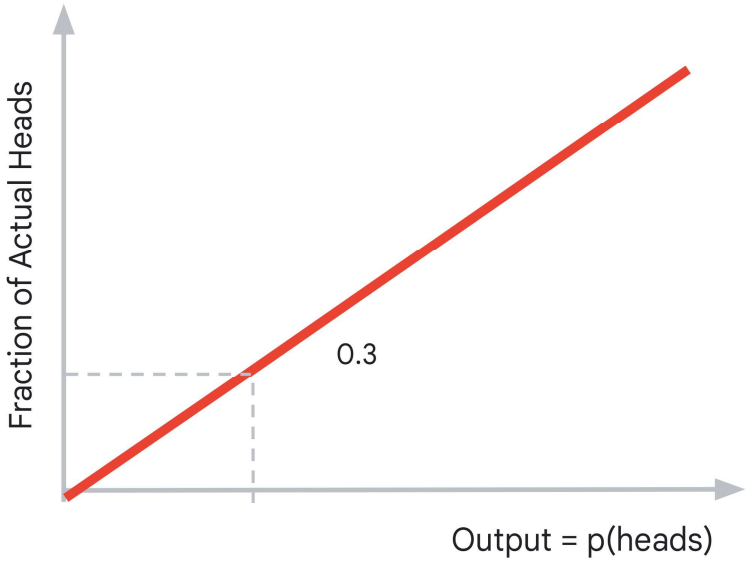
The output of Logistic Regression is a calibrated probability estimate

Useful because we can cast binary classification problems into probabilistic problems:

Will customer buy item?

becomes

Predict the probability that customer buys item



Vertex AI AutoML

Choose a training method

Auto ML

- Create and train a model with minimal technical effort.
- Quickly prototype models or explore datasets before developing in a custom training application.

Custom training

- Create a training application optimized for your targeted outcome.
- Maintain complete control over training application functionality.
    - Target any objective, use any algorithms, develop your own loss functions or metrics, or other customizations.

When to use AutoML and when to use custom training

	AutoML	Custom training
Data science expertise needed	No.	Yes, to develop the training application and also to do some of the data preparation like feature engineering.
Programming ability needed	No, AutoML is codeless.	Yes, to develop the training application.
Time to trained model	Lower. Less data preparation is required, and no development is needed.	Higher. More data preparation is required, and training application development is needed.
Limits on machine learning objectives	Yes. You must target one of AutoML's predefined objectives.	No.
Can manually optimize model performance with hyperparameter tuning	No. AutoML does some automated hyperparameter tuning, but you can't modify the values used.	Yes. You can tune the model during each training run for experimentation and comparison.

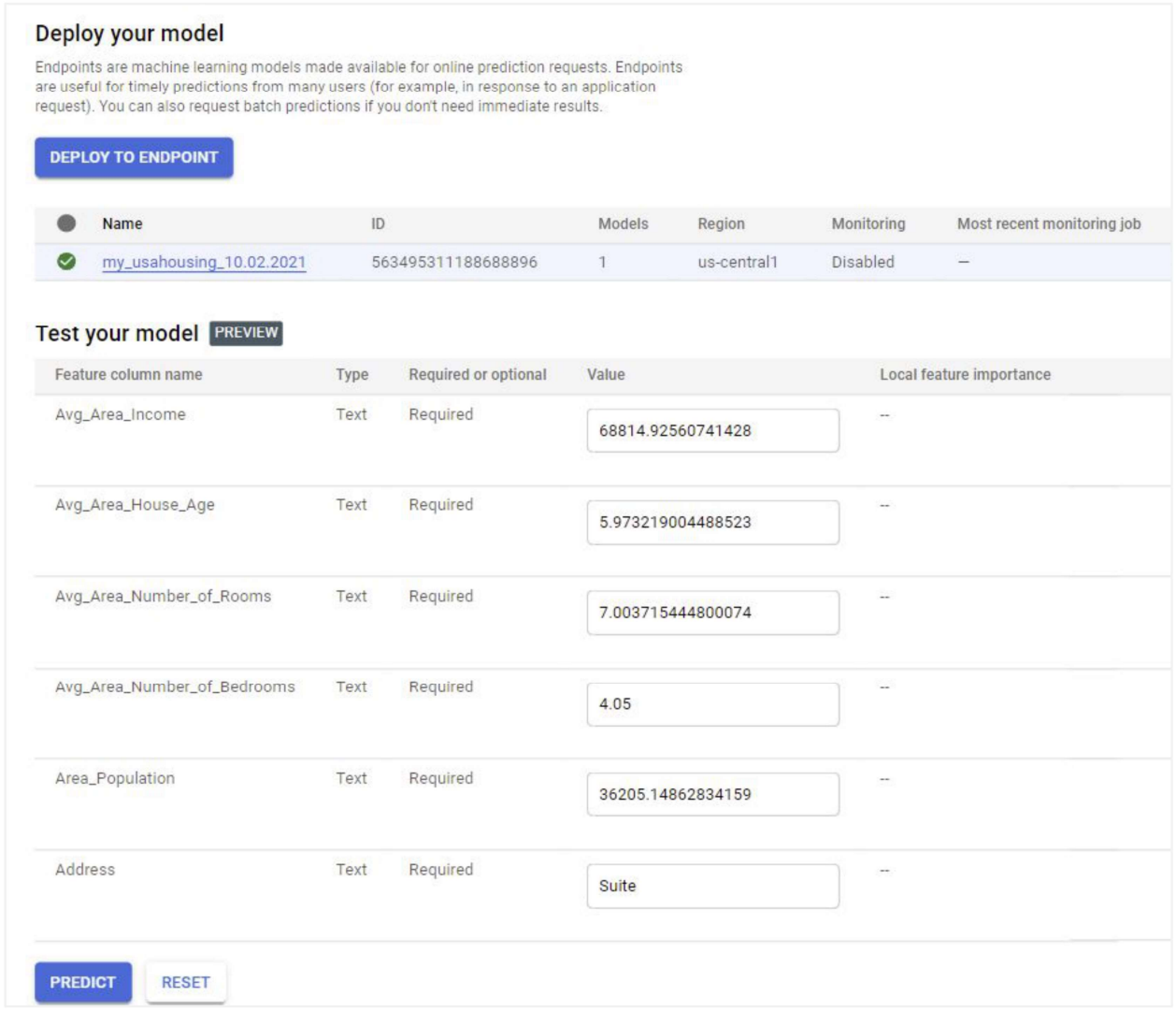
When to use AutoML and when to use custom training

	AutoML	Custom training
Can control aspects of the training environment	Limited. For image and tabular datasets, you can specify the number of node hours to train for, and whether to allow early stopping of training.	Yes. You can specify aspects of the environment such as Compute Engine machine type, disk size, machine learning framework, and number of nodes.
Limits on data size	Yes. AutoML uses managed datasets; data size limitations vary depending on the type of dataset.	For unmanaged datasets, no. Managed datasets have the same limits as Vertex AI datasets that are used to train AutoML models.

Test your model

- After evaluating your model metrics, you can test your model with new data.
- See if the model's predictions match your expectations.
- If not, you may need to continue improving your model's performance.

Deploy your model



Deploy your model and make online predictions

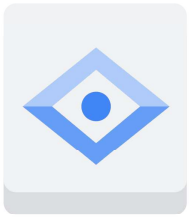
Batch prediction

- Allows you to make many prediction requests at once.
- Is asynchronous (the model won't return a CSV file or BigQUery Table until it processes all prediction requests).

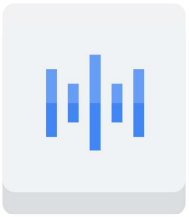
Online prediction

- Deploy your model to make it available for prediction requests using a REST API.
- Is synchronous (the model will quickly return a prediction, but only accepts one prediction request per API call).
- This is useful if parts of your system are dependent on a quick prediction turnaround.

ML/AI solutions



Vision API

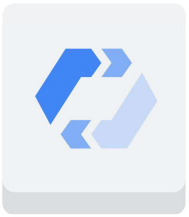


Speech API

Translation API

Natural Language API

Vertex AI



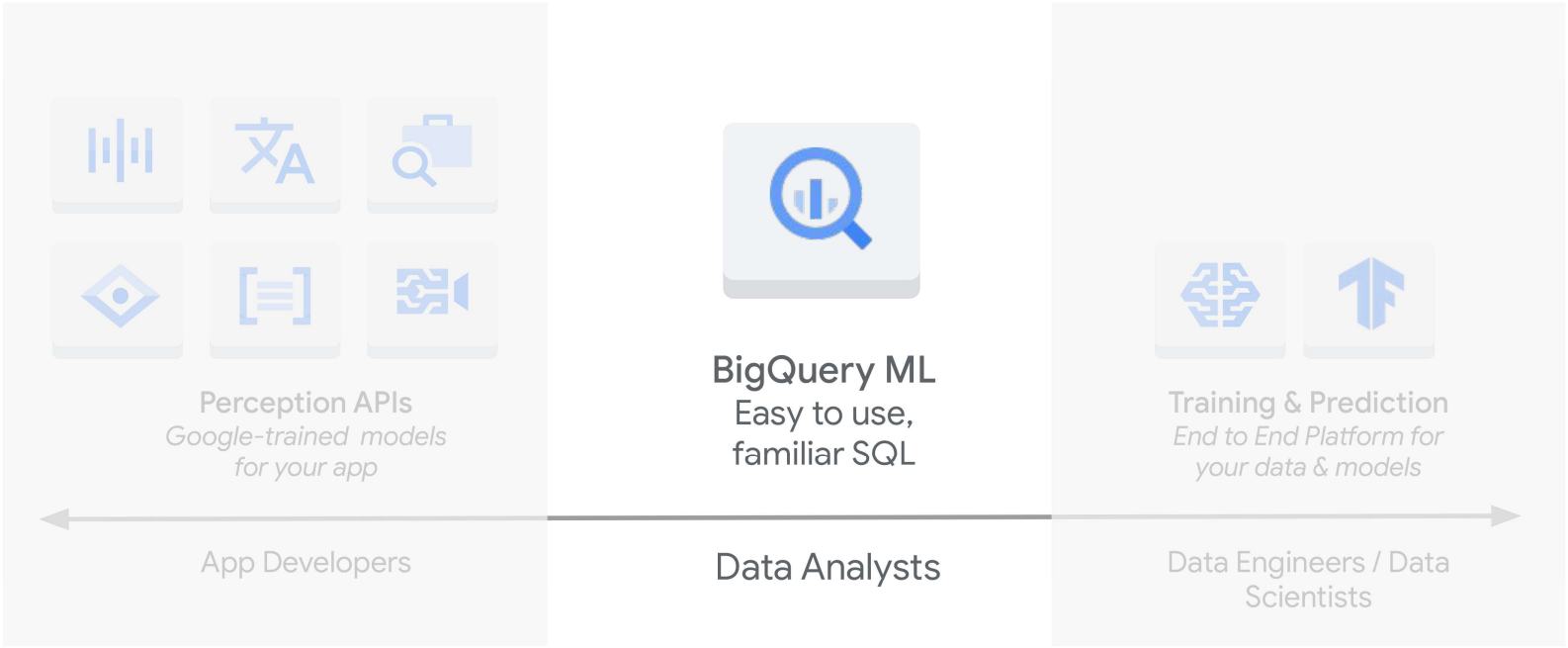
AutoML (Fast)

...and more!

Pre-trained models

It is very common to enrich your data with pre-trained models, to take advantage of unstructured data.

BigQuery ML is a way to easily build machine learning models



Working with BigQuery ML

01

Dataset

02

Create/ train

```
CREATE MODEL `BigQuery
ML_tutorial.sample_model`
OPTIONS(model_type='logistic_reg') AS
SELECT
```

03

Evaluate

```
FROM
  ML.EVALUATE(MODEL `BigQuery
ML_tutorial.sample_model`,
    TABLE eval_table)
```

04

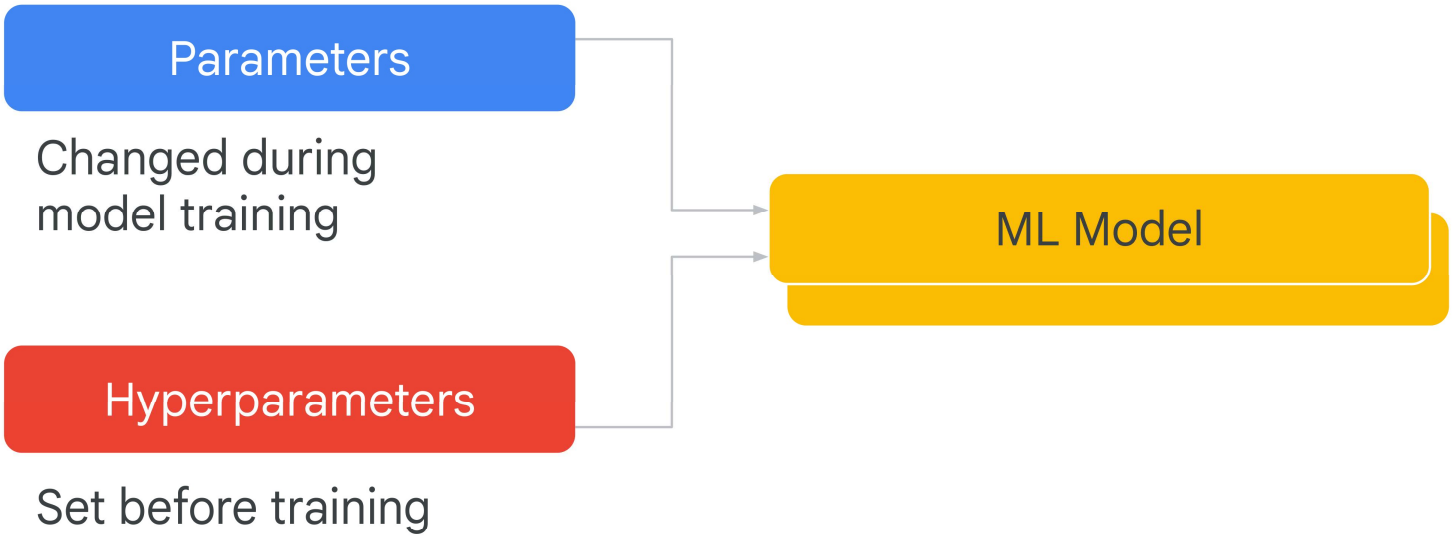
Predict/ classify

```
FROM
    ML.PREDICT(MODEL `BigQuery
ML_tutorial.sample_model`,
      table game_to_predict) ) AS
predict
```

Hyperparameter tuning

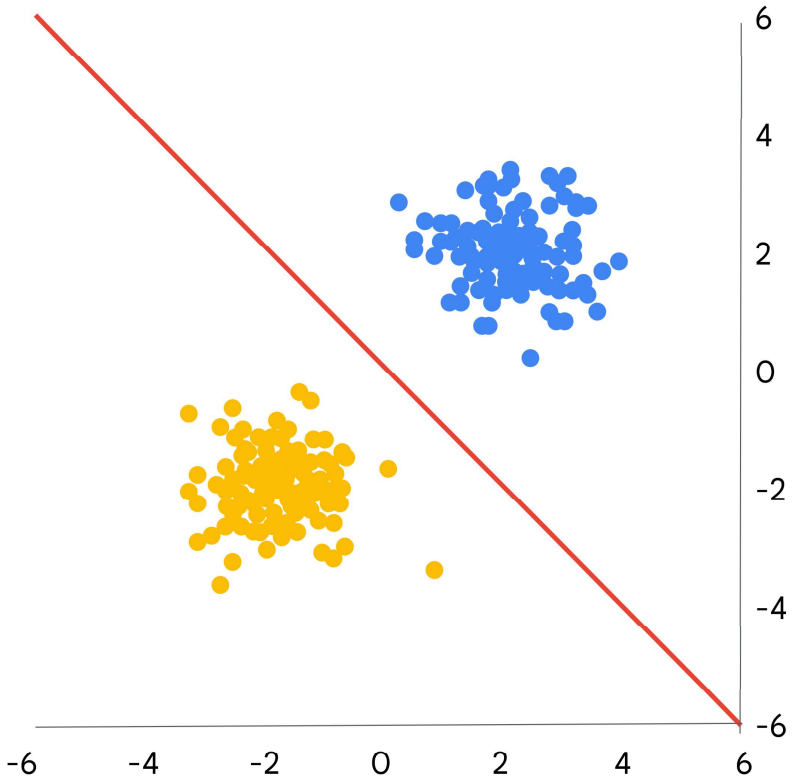
A hyperparameter is a model argument whose value is set before the learning process begins.

ML models are mathematical functions with parameters and hyper-parameters



$$b + m \times x = y$$

How can linear models classify data?



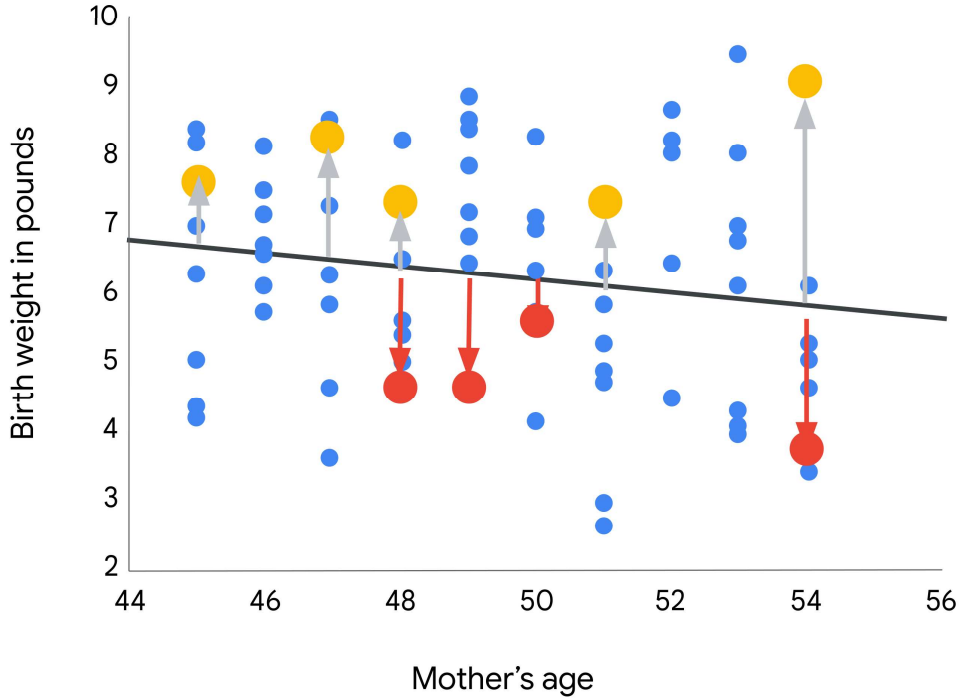
Loss functions

Compose a loss function by calculating errors

Each error makes sense. How about all the errors added together?

Error = actual (true) - predicted value
Compute the errors:

+0.70
+1.10
+0.65
-1.20
-1.15
+1.10
+3.09
-2.10



One loss function metric is Root Mean Squared Error (RMSE)

01 Get the errors for the training examples.

+0.70
+1.10
+0.65
-1.20
-1.15
+1.10
+3.09
-2.10

02 Compute the squares of the error values.

0.49
1.21
0.42
1.44
1.32
1.21
9.55
4.41

03 Compute the mean of the squared error values.

2.51

04 Take a square root of the mean.

1.58

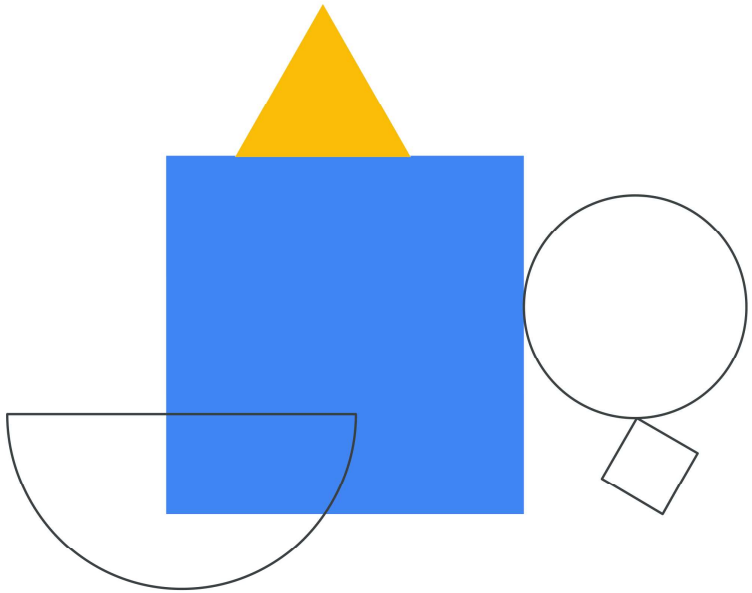
$$\sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}$$

$\hat{Y}_i$ predicted value

$\hat{Y}_i$ labeled value

Lab intro

Develop an intuitive understanding of neural networks using Tensorflow Playground



Activation function

Skewed data can make inappropriate strategies seductive

1000 parking spaces.
990 of them are **taken**.
10 are **available**.

An ML model that always reported that a space was occupied would be right 99/100 times.

Performance metrics allow you to measure what matters

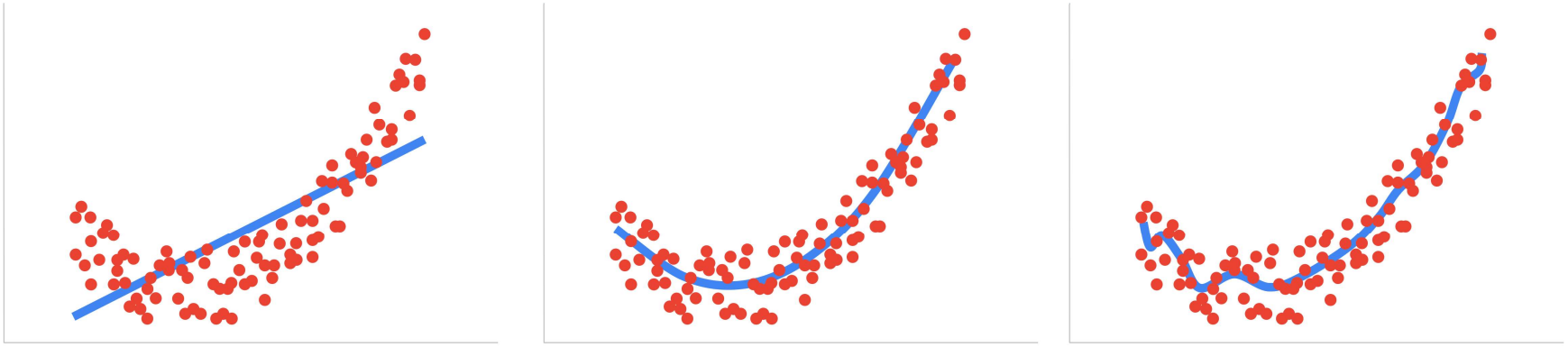
Loss functions

- During training
- Harder to understand
- Indirectly connected to business goals

Performance metrics

- After training
- Easier to understand
- Directly connected to business goals

Beware of overfitting as you increase model complexity

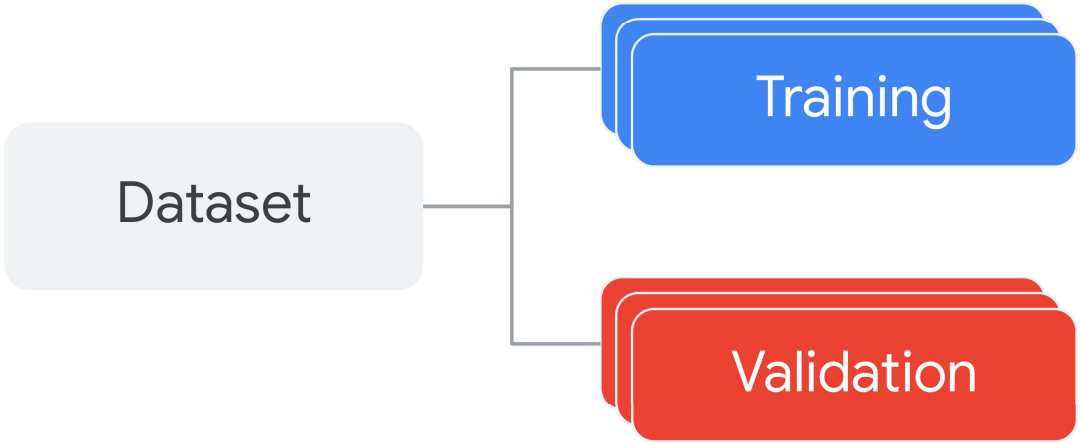


Underfit

Fit

Overfit

Evaluate the final model with cross-validation



Thanks