Huawei HCIA - IA

Machine Learning - Basics Concepts

Lucas Sousa | Madson Dias - Feb 2021

Agenda

Main Topics

- Machine Learning Types
- Machine Learning Process

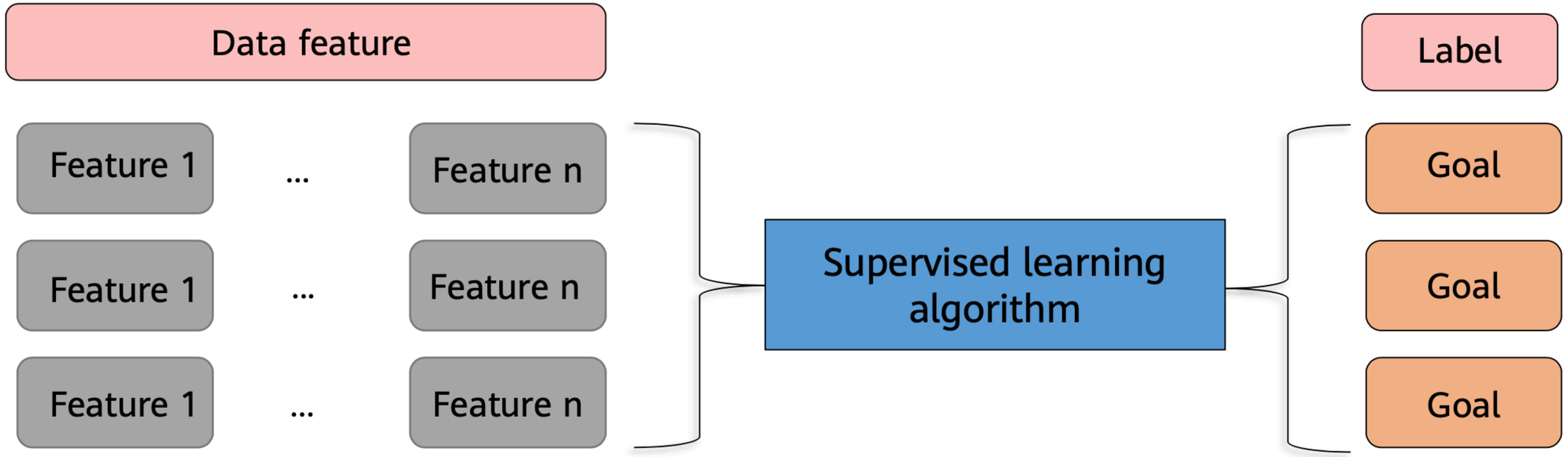
Machine Learning Types

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

Supervised Learning

- Obtain an optimal model with required performance through training and learning based on the samples of known categories. Then, use the model to map all inputs to outputs and check the output for the purpose of classifying unknown data.

Supervised Learning



Weather	Temperature	Wind Speed
Sunny	Warm	Strong
Rainy	Cold	Fair
Sunny	Cold	Weak

Enjoy Sports
Yes
No
Yes

Supervised Learning

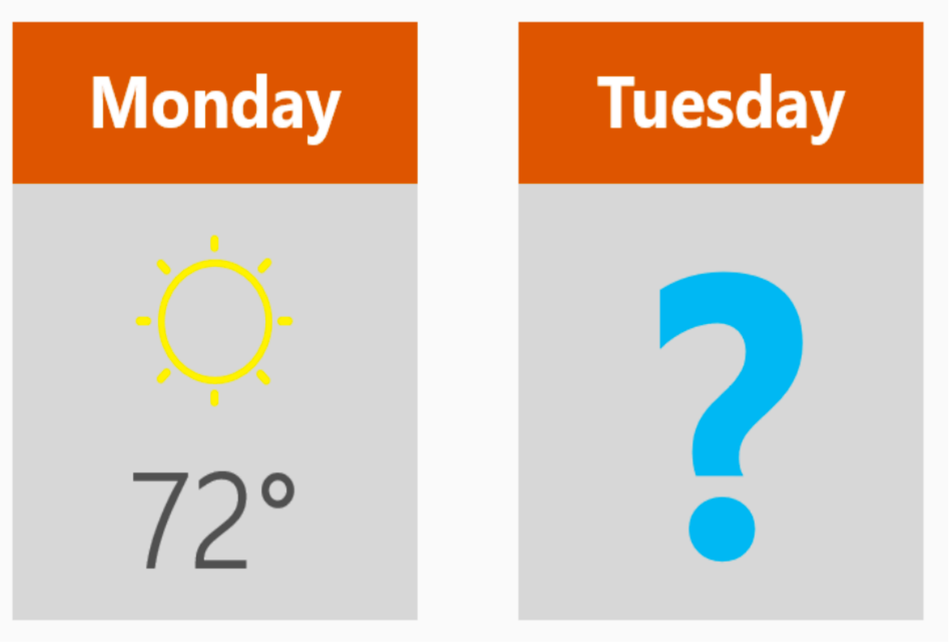
Regression Questions

- Reflects the features of attribute values of samples in a sample dataset. The dependency between attribute values is discovered by expressing the relationship of samples mapping through functions.

Regression Questions

How much will I benefit from the stock next week?

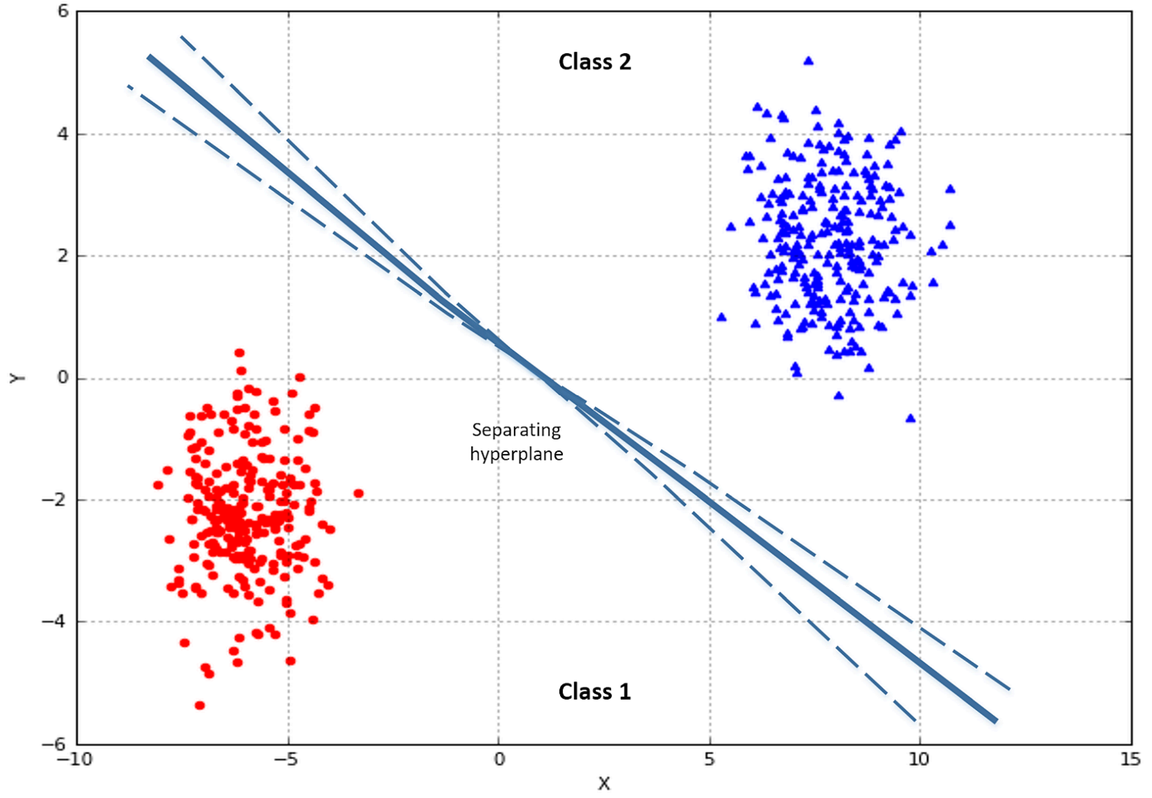
What’s the temperature on Tuesday?



Supervised Learning

Classification Questions

- Maps samples in a sample dataset to a specified category by using a classification model.



Classification Questions

Will there be a traffic jam on XX road during the morning rush hour tomorrow?

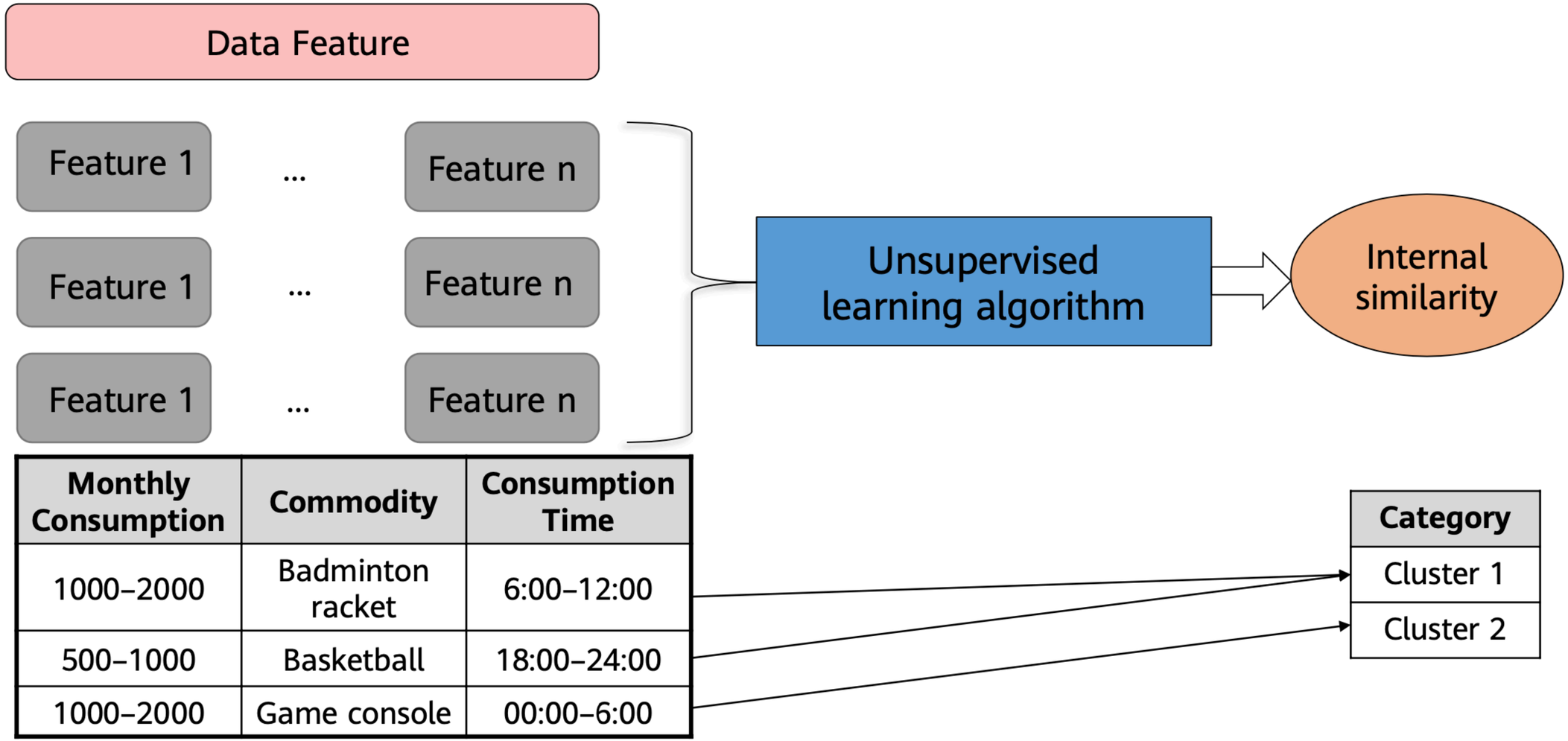
Classification Questions

Is this email a spam?

The main difference between classification and regression problems are their output.

- **In regression problems the outputs are continuous values.**
- **In classification problems the outputs are discrete values.**

Unsupervised Learning



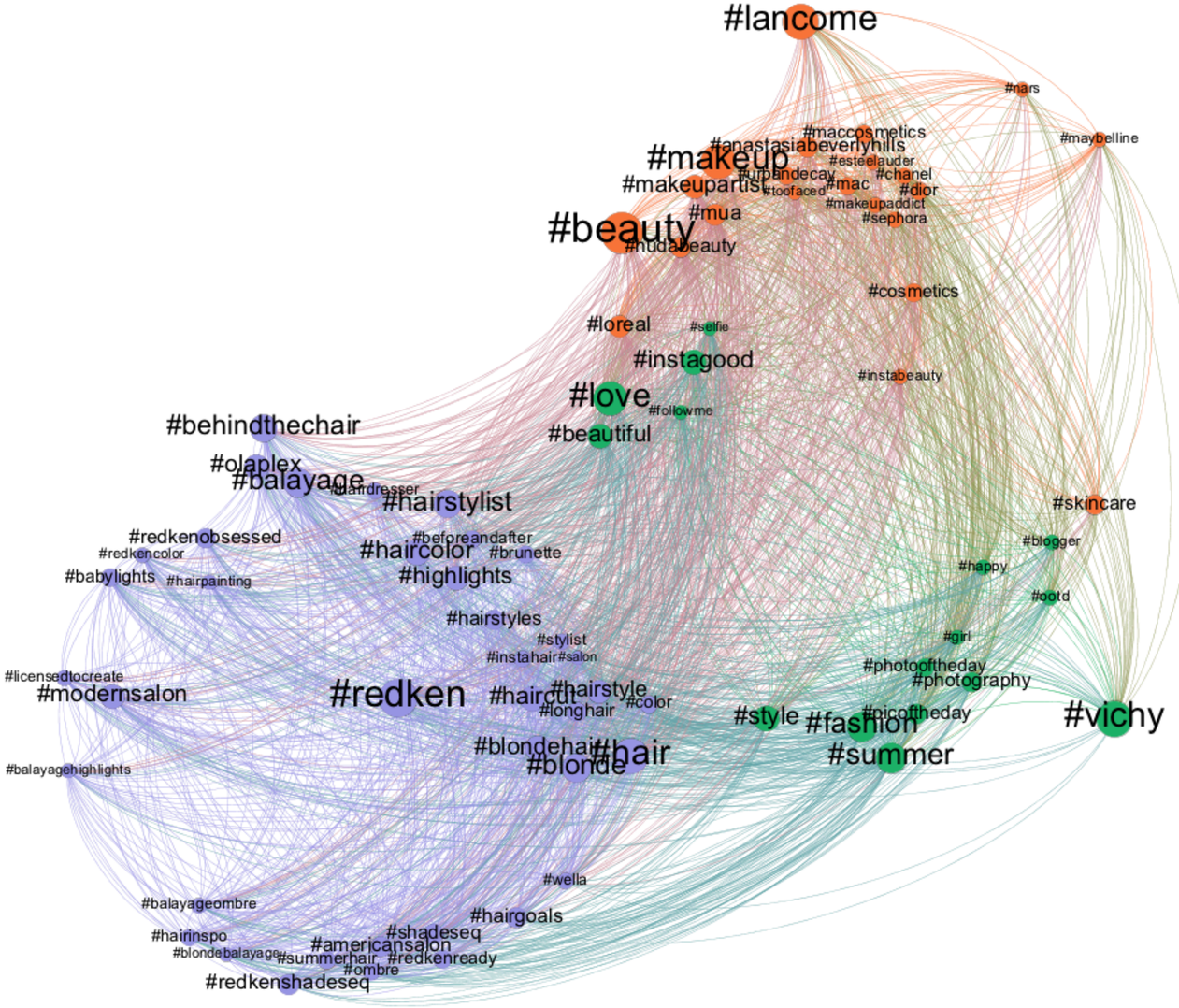
Unsupervised Learning

Clustering Questions

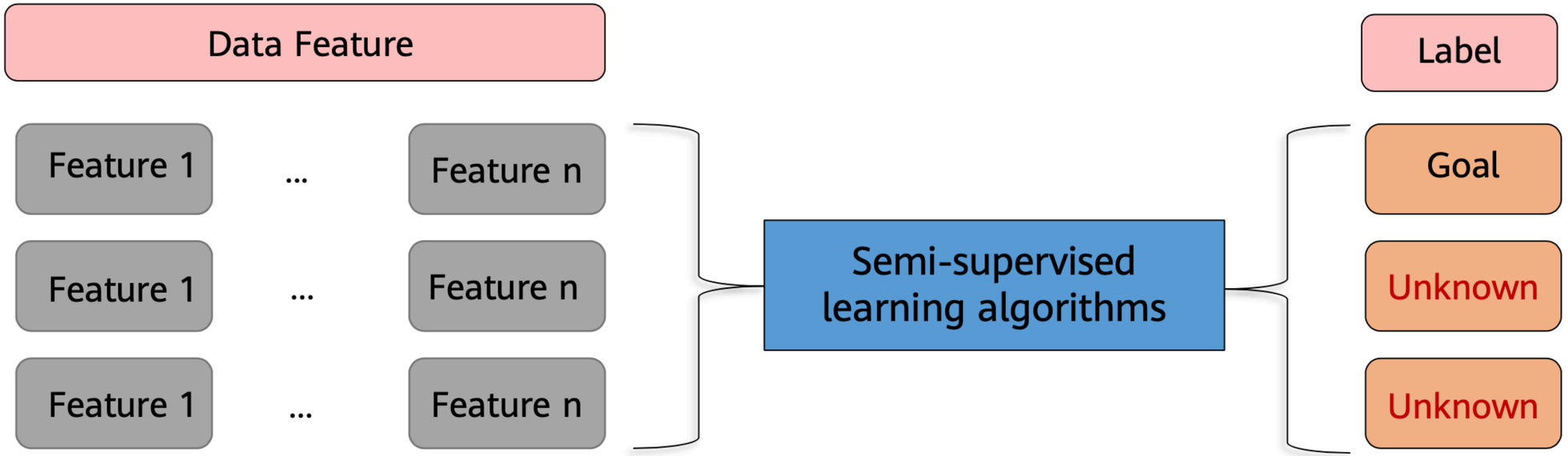
- Classifies samples in a sample dataset into several categories based on the clustering model. The similarity of samples belonging to the same category is high.

Clustering Questions

What subjects are similar?



Semi-supervised Learning

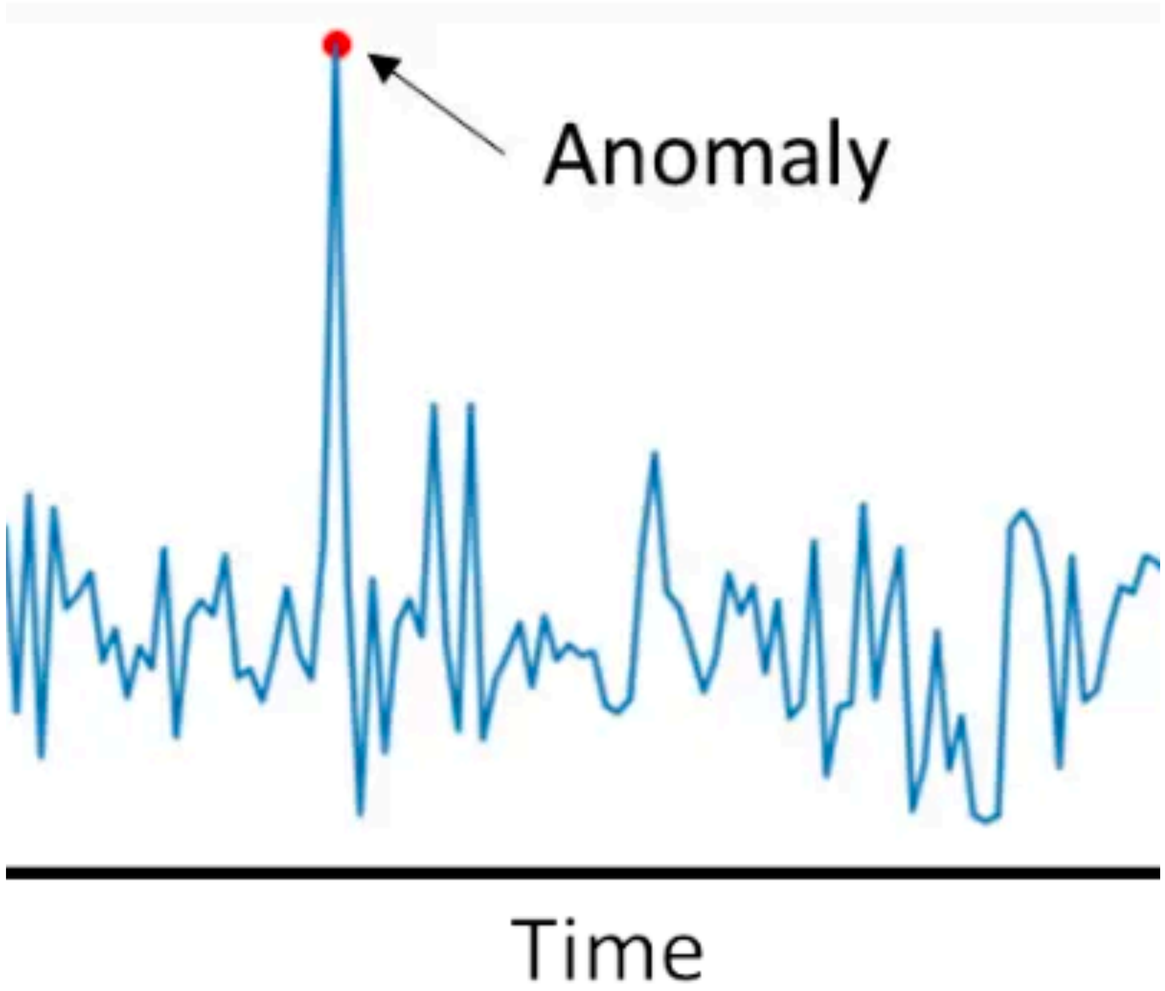


Weather	Temperature	Wind Speed
Sunny	Warm	Strong
Rainy	Cold	Fair
Sunny	Cold	Weak

Enjoy Sports
Yes
/
/

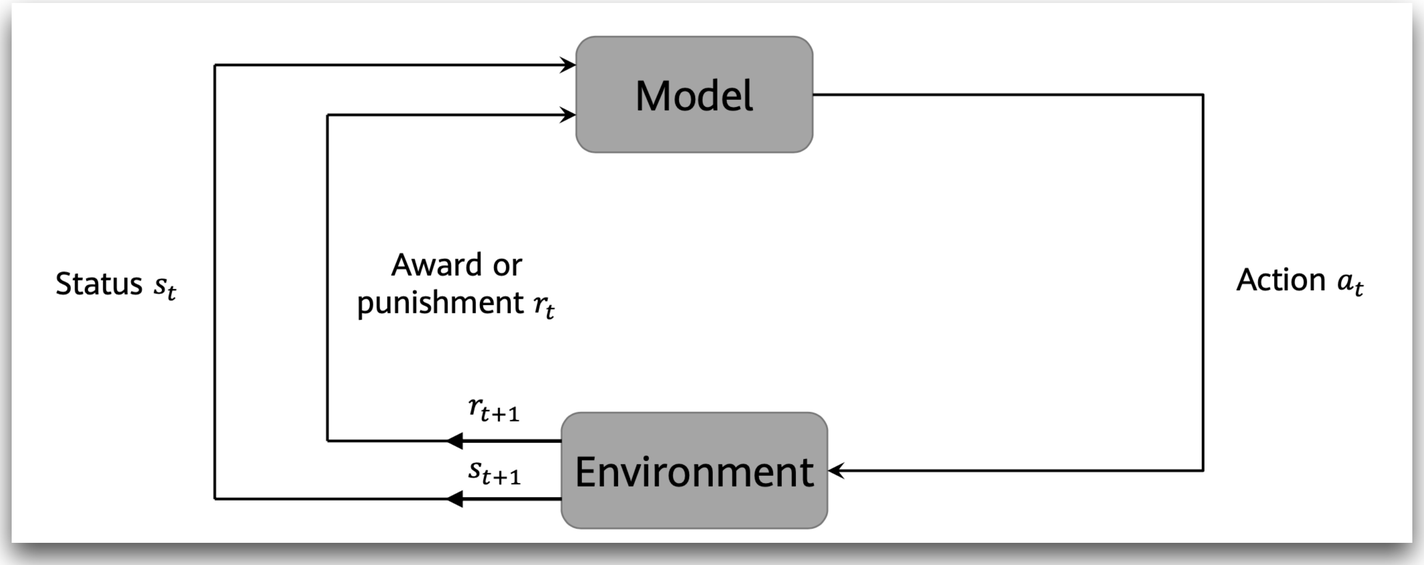
Semi-supervised Questions

Is this sample an anomaly?



Reinforcement Learning

- The model perceives the environment, takes actions, and makes adjustments and choices based on the status and award or punishment.



2048 - Mozilla Firefox

2048

https://gabrielecirulli.github.io/2048/

2048

SCORE 2754

BEST 5396

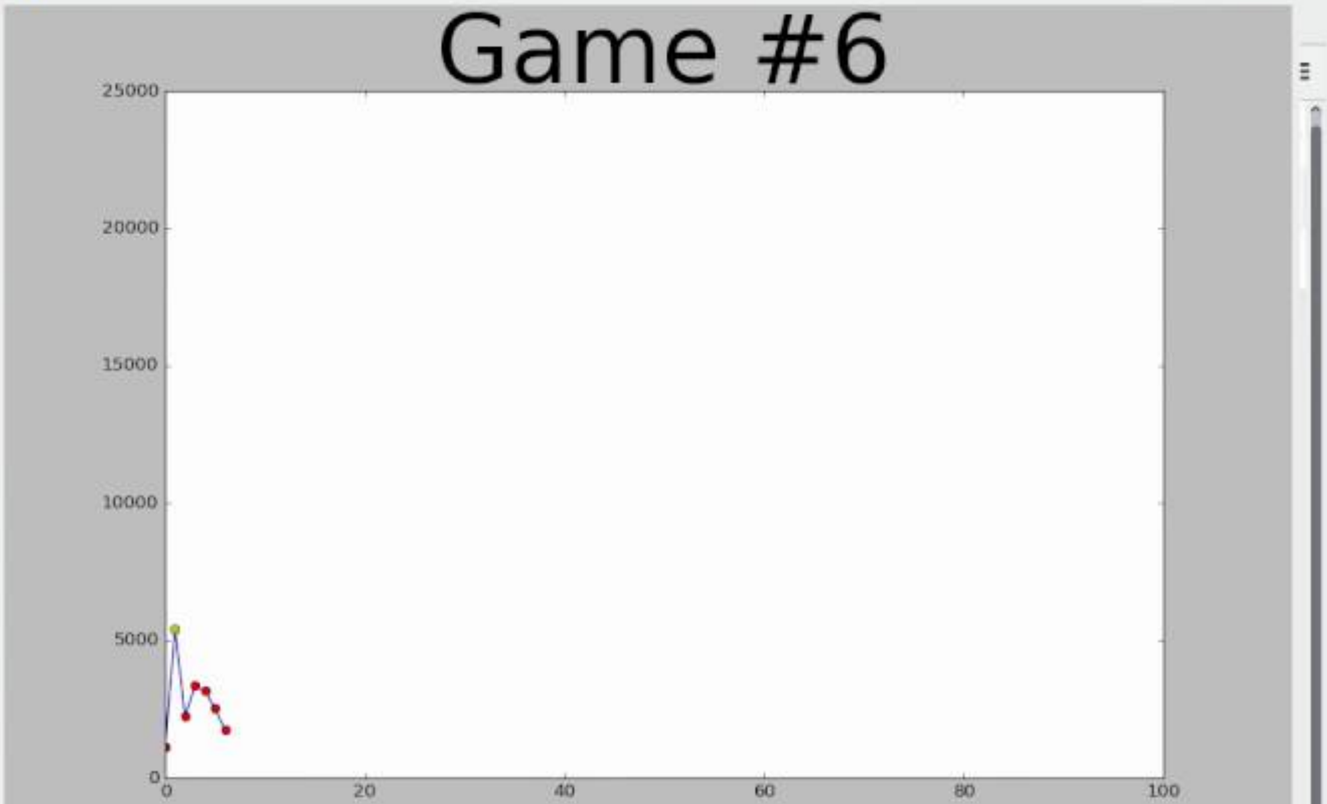
Join the numbers and get to the **2048 tile!**

New Game

NEW: Get the new 2048 app for iOS and Android! x

2	2	4	16
	8	32	64
2	16	256	8
	8	64	32

Figure 1




```
break
```

```
/usr/lib/python3/dist-packages/matplotlib/backend_bases.py:2437: MatplotlibDeprecationWarning: Using default event loop until function specific to this GUI is implemented
  warnings.warn(str, mplDeprecation)
```

```
0 (136, 64, 1136)
1 (385, 512, 5396)
2 (212, 128, 2240)
3 (275, 256, 3352)
4 (262, 256, 3160)
5 (209, 256, 2528)
6 (177, 128, 1752)
```

my2048-rl-theano-n-tuple-Copy | 2048 - Mozilla Firefox | 2048-NN : bash — Konsole | Figure 1 | us | 9:37 上午

Reinforcement Learning

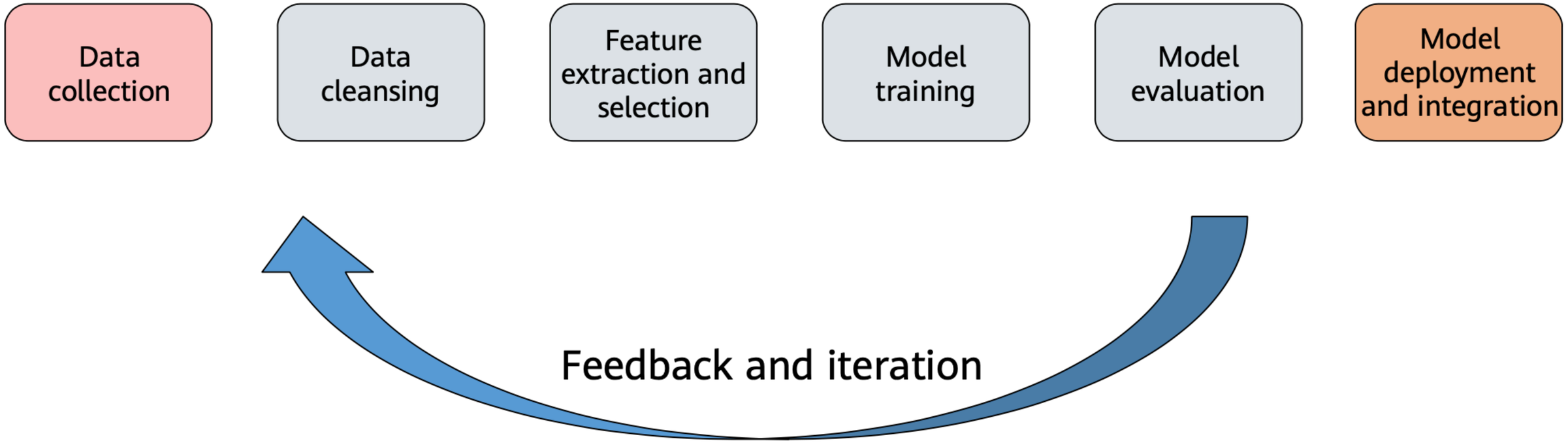
Best Behavior

Always looks for best behaviors. Reinforcement learning is targeted at machines or robots.

- **Autopilot:** Should it brake or accelerate when the yellow light starts to flash?
- **Cleaning robot:** Should it keep working or go back for charging?

Machine Learning Process

Machine Learning Process



Basic Machine Learning Concept

Dataset

- A collection of data used in machine learning tasks. Each data record is called a sample. Events or attributes that reflect the performance or nature of a sample in a particular aspect are called features.

Basic Machine Learning Concept

Training Set

- A dataset used in the training process, where each sample is referred to as a training sample. The process of creating a model from data is called learning (training).

Basic Machine Learning Concept

Test set

- Testing refers to the process of using the model obtained after learning for prediction. The dataset used is called a test set, and each sample is called a test sample.

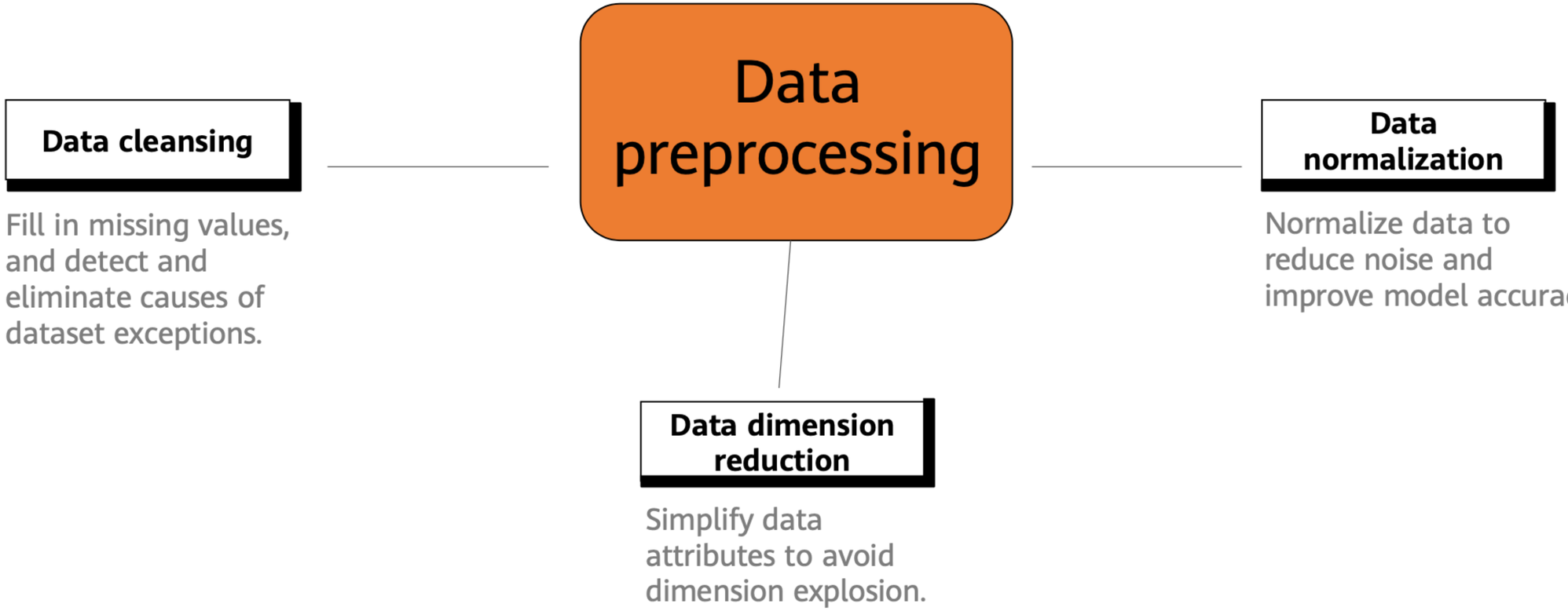
Checking Data Overview

Typical dataset form (Hold-out)

		Feature 1	Feature 2	Feature 3	Label
	No.	Area	School Districts	Direction	House Price
Training set	1	100	8	South	1000
	2	120	9	Southwest	1300
	3	60	6	North	700
	4	80	9	Southeast	1100
Test set	5	95	3	South	850

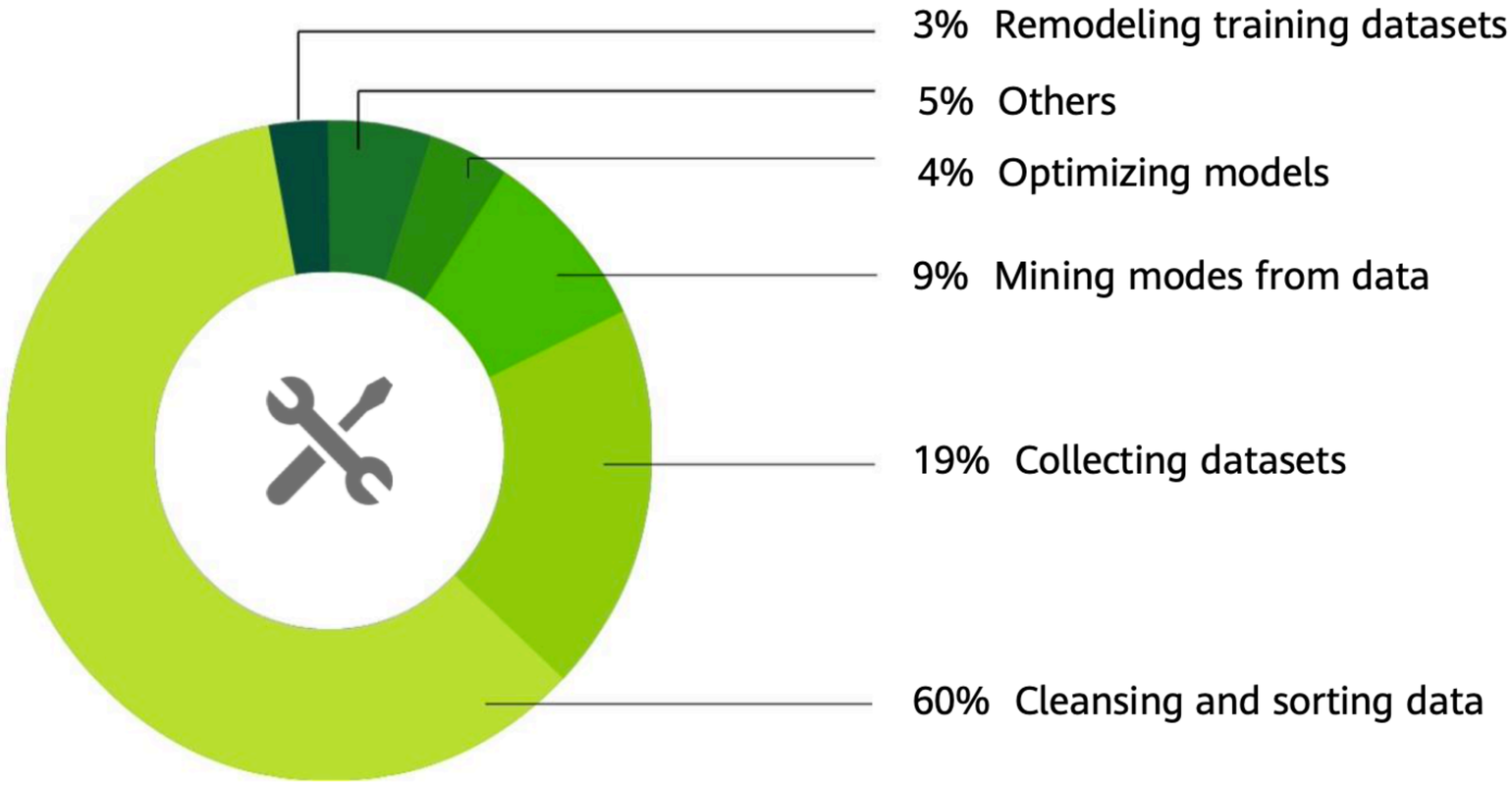
Importance of data processing

- Data is crucial to models. It is the ceiling of model capabilities. Without good data, there is no good model.



Workload of Data Cleansing

Statistics on data scientists' work in machine learning



CrowdFlower Data Science Report 2016

Data Cleansing

- Most machine learning models process features, which are usually numeric representations of input variables that can be used in the model.

Data Cleansing

- In most cases, the collected data can be used by algorithms only after being preprocessed. The preprocessing operations include the following:
  - Data filtering
  - Processing of lost data
  - Processing of possible exceptions, errors, or abnormal values
  - Combination of data from multiple data sources
  - Data consolidation

Dirty Data

Generally, real data may have some quality problems.

- **Incompleteness:** contains missing values or the data that lacks attributes.
- **Noise:** contains incorrect records or exceptions.
- **Inconsistency:** contains inconsistent records.

Dirty Data

Generally, real data may have some quality problems.

#	Id	Name	Birthday	Gender	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Missing value

Invalid value

Value that should be in another column

Misspelling

Invalid duplicate item

Incorrect format

Attribute dependency

Data Conversion

- After being preprocessed, the data needs to be converted into a representation form suitable for the machine learning model. Common data conversion forms include the following:

  - With respect to classification, category data is encoded into a corresponding numerical representation.

Data Conversion

- After being preprocessed, the data needs to be converted into a representation form suitable for the machine learning model. Common data conversion forms include the following:

  - Value data is converted to category data to reduce the value of variables (for age segmentation).

Data Conversion

- After being preprocessed, the data needs to be converted into a representation form suitable for the machine learning model. Common data conversion forms include the following:

  - In the text, the word is converted into a word vector through word embedding (generally using the word2vec model, BERT model, etc).
  - Process image data (color space, grayscale, geometric change, Haar feature, and image enhancement).

Data Conversion

- Feature engineering
  - Normalize features to ensure the same value ranges for input variables of the same model.
  - Feature expansion: Combine or convert existing variables to generate new features, such as the average.

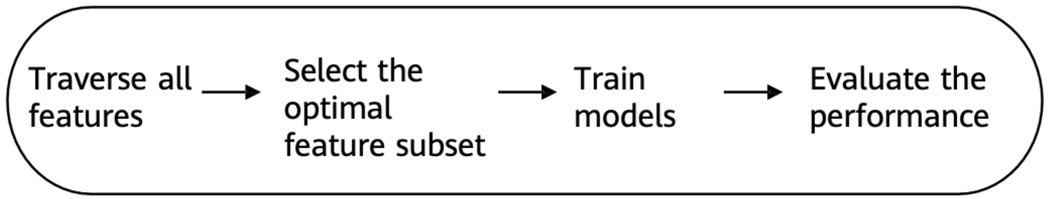
Necessity of Feature Selection

Generally, a dataset has many features, some of which may be redundant or irrelevant to the value to be predicted.

Feature selection is necessary in the following aspects:

Feature Selection Methods - Filter

Filter methods are independent of the model during feature selection.



Procedure of a filter method

By evaluating the correlation between each feature and the target attribute, these methods use a statistical measure to assign a value to each feature. Features are then sorted by score, which is helpful for preserving or eliminating specific features.

Common methods

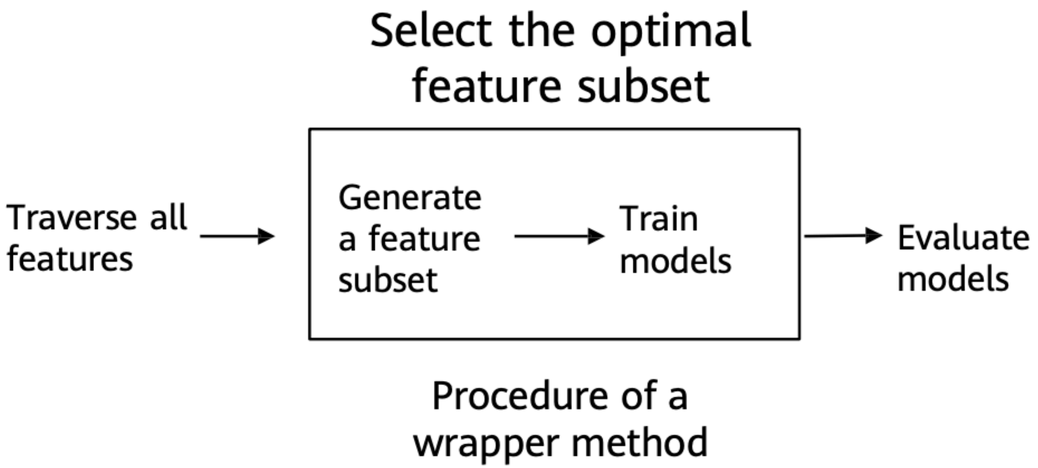
- Pearson correlation coefficient
- Chi-square coefficient
- Mutual information

Limitations

- The filter method tends to select redundant variables as the relationship between features is not considered.

Feature Selection Methods - Wrapper

Wrapper methods use a prediction model to score feature subsets.



Procedure of a wrapper method

Wrapper methods consider feature selection as a search issue for which different combinations are evaluated and compared. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy.

Common methods

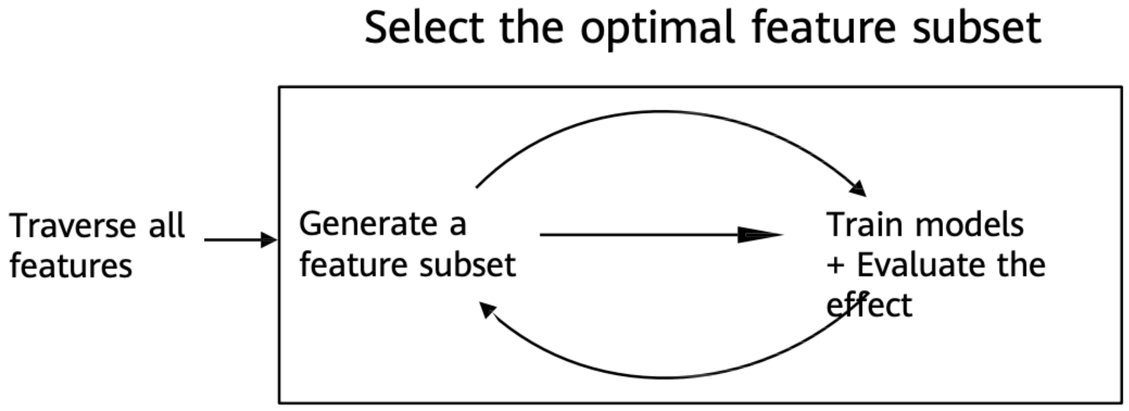
- Recursive feature elimination (RFE)

Limitations

- Wrapper methods train a new model for each subset, resulting in **a huge number of computations**.
- A feature set with the best performance is usually provided for a specific type of model.

Feature Selection Methods - Embedded

Embedded methods consider feature selection as a part of model construction.

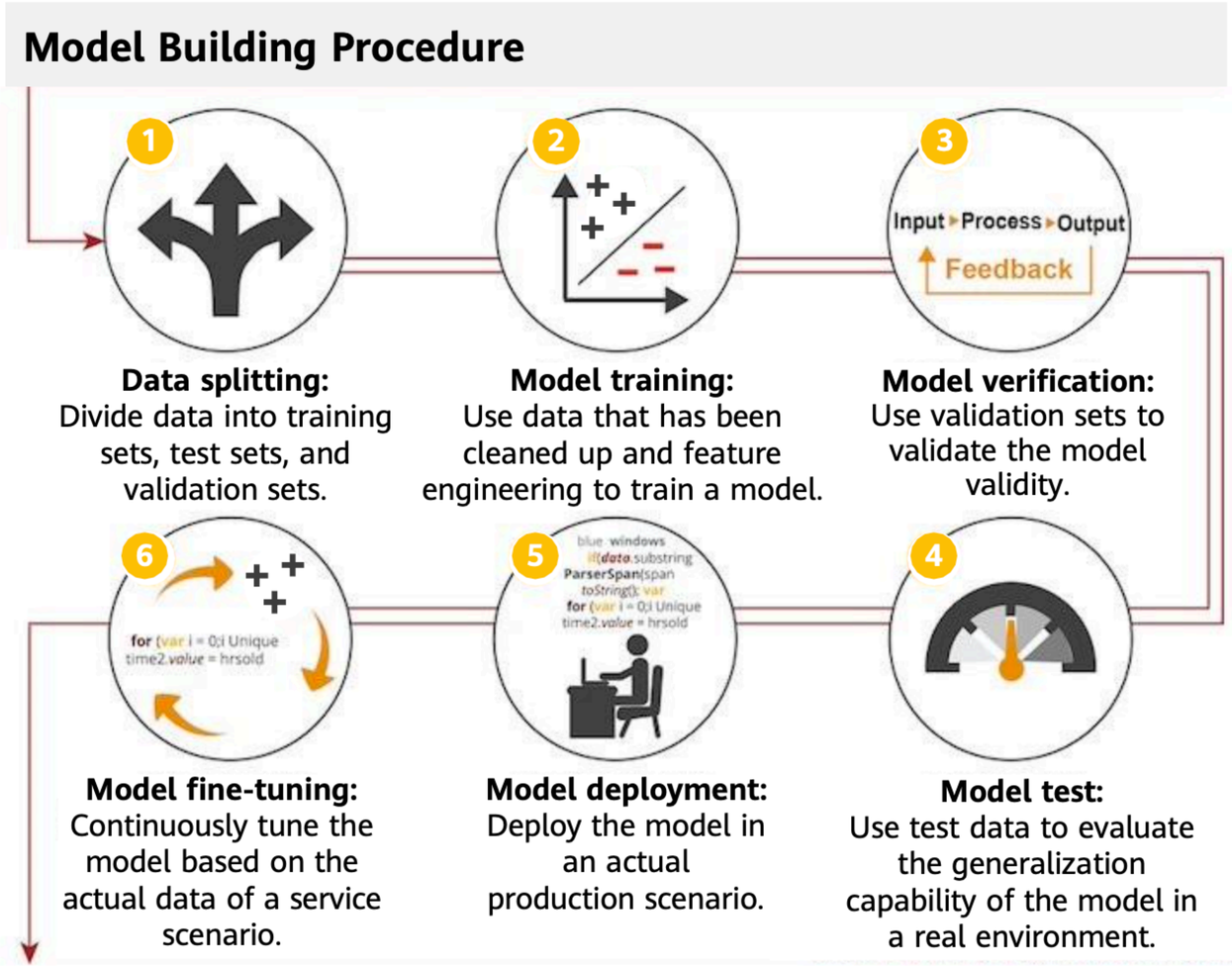


Procedure of an embedded method

The most common type of embedded feature selection method is the **regularization method**. Regularization methods are also called penalization methods that introduce additional constraints into the optimization of a predictive algorithm that bias the model toward lower complexity and reduce the number of features.

Common methods
- Lasso regression
- Ridge regression

Overall Procedure of Building a Model



What Is a Good Model?

- **Generalization capability**
  Can it accurately predict the actual service data?
- **Interpretability**
  Is the prediction result easy to interpret?
- **Prediction speed**
  How long does it take to predict each piece of data?
- **Practicability**
  Is the prediction rate still acceptable when the service volume increases with a huge data volume?