LOGO

BigData Engineering

9주차: Machine Learning Basic

강의 : 신경섭

We covered…

- Python Basics
  - Basic functions, Types, List, Dictionary
- Numpy library
- Pandas Library
  - Series data structure
  - Dataframe data structure
  - Indexing/Merge/Groupby
- Matplotlib
  - Scatter/Line/Bar chart
  - Subplot,histogram
- Data acquisition

Today’s Subjects

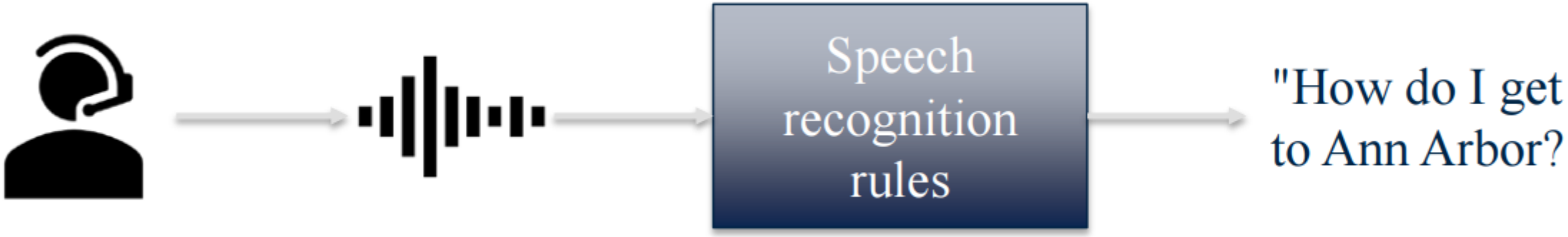
- Applied Machine Learning for data analysis
- Machine learning workflow with data

What is Machine Learning (ML)?

- The study of computer programs (algorithms) that can learn by example
- ML algorithms can **generalize** from existing examples of a task

– *e.g. after seeing a training set of labeled images, an image classifier can figure out how to apply labels accurately to new, previously unseen images*

Speech Recognition



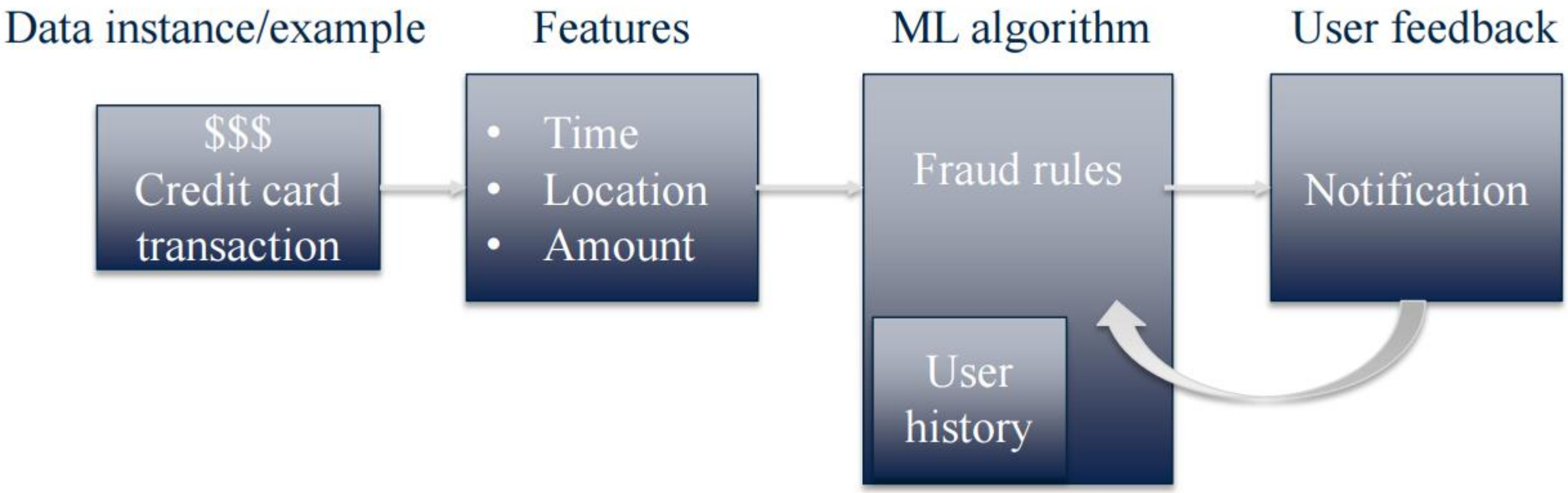
Machine Learning models learn from experience

- **Labeled examples**
  (Email spam detection)

- **User feedback**
  (Clicks on a search page)

- **Surrounding environment**
  (self-driving cars)

Machine Learning for fraud detection



Feature Representation

Feature representation

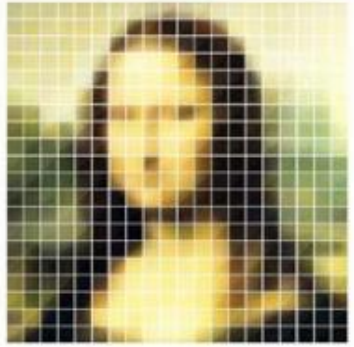
Email

```
To:  Chris Brooks
From:  Daniel Romero
Subject:  Next course offering
Hi Daniel,
Could you please send the outline for the
next course offering?  Thanks! -- Chris
```

Feature	Count
to	1
chris	2
brooks	1
from	1
daniel	2
romero	1
the	2
...	

A list of words with their frequency counts

Picture



A matrix of color values (pixels)

Sea Creatures

Feature	Value
DorsalFin	Yes
MainColor	Orange
Stripes	Yes
StripeColor1	White
StripeColor2	Black
Length	4.3 cm

A set of attribute values

What is Applied Machine Learning?

- Understand basic ML concepts and workflow
- How to properly apply 'black-box' machine learning components and features
- Learn how to apply machine learning algorithms in Python using the scikit-learn package

Supervised Machine Learning

Training set

X Sample		Y Target Value (Label)	
	x_1	Apple	y_1
	x_2	Lemon	y_2
	x_3	Apple	y_3
	x_4	Orange	y_4

Classifier

$f : X \rightarrow Y$

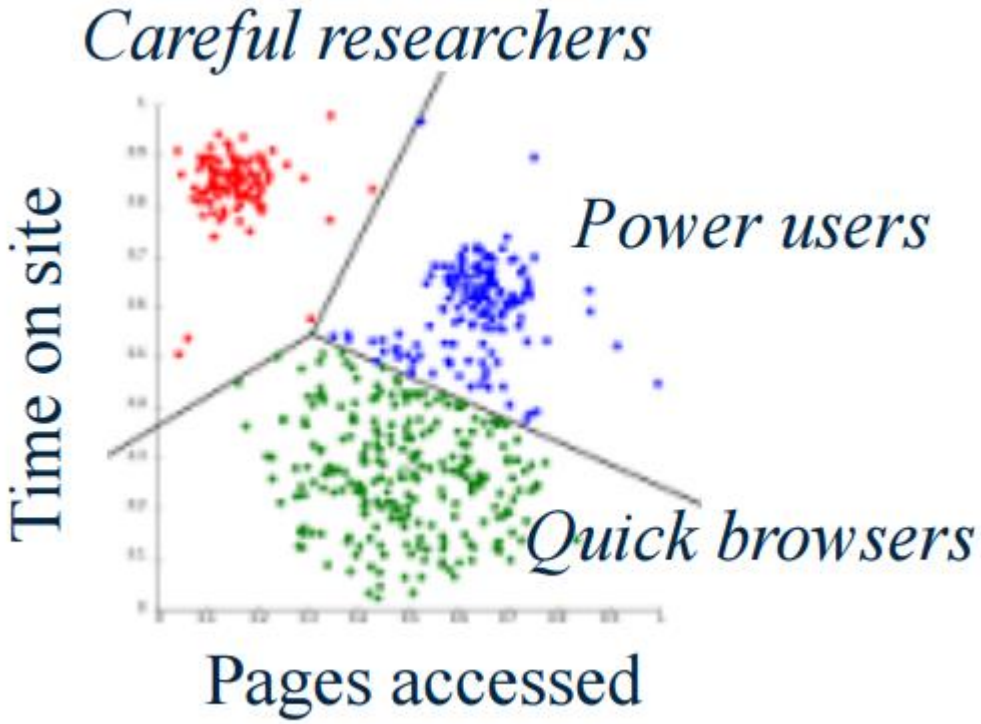
At training time, the classifier uses labelled examples to learn rules for recognizing each fruit type.

Future sample

Label: Orange

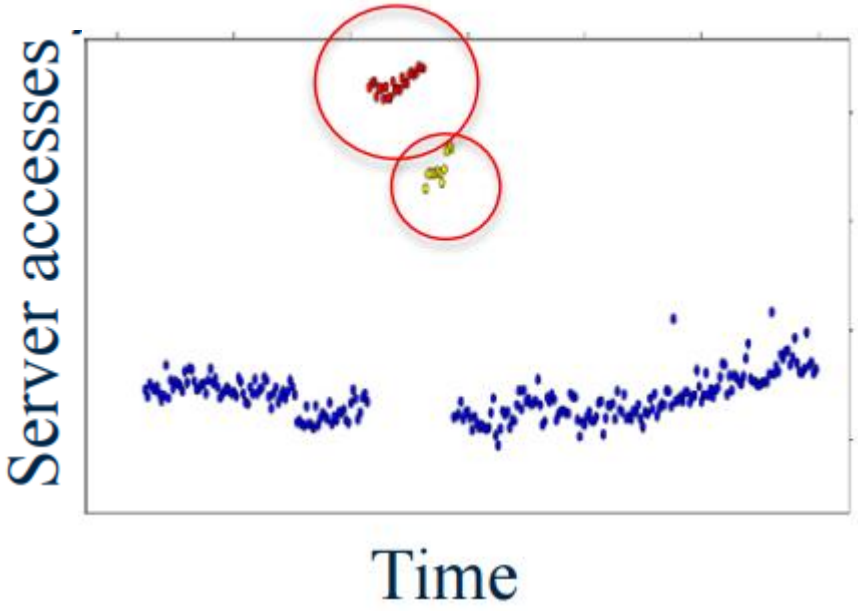
After training, at prediction time, the trained model is used to predict the fruit type for new instances using the learned rules.

Unsupervised Machine Learning

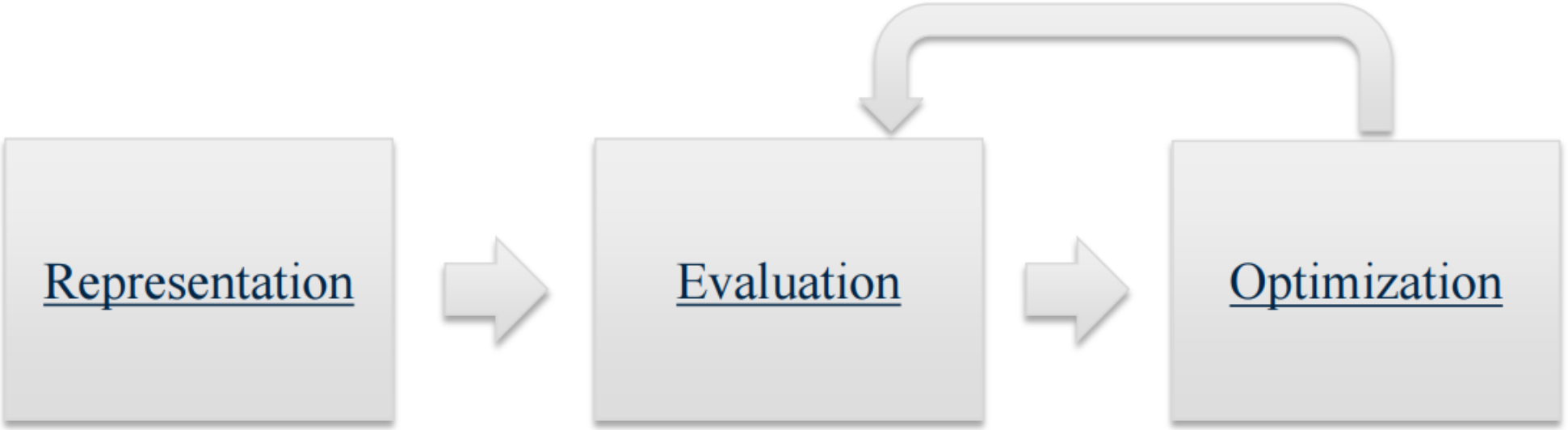


- **Finding clusters of similar users (clustering)**

- **Detecting abnormal server access patterns (unsupervised outlier detection)**



A Basic Machine Learning Workflow



Choose:
- A feature representation
- Type of classifier to use

e.g. image pixels, with
k-nearest neighbor classifier

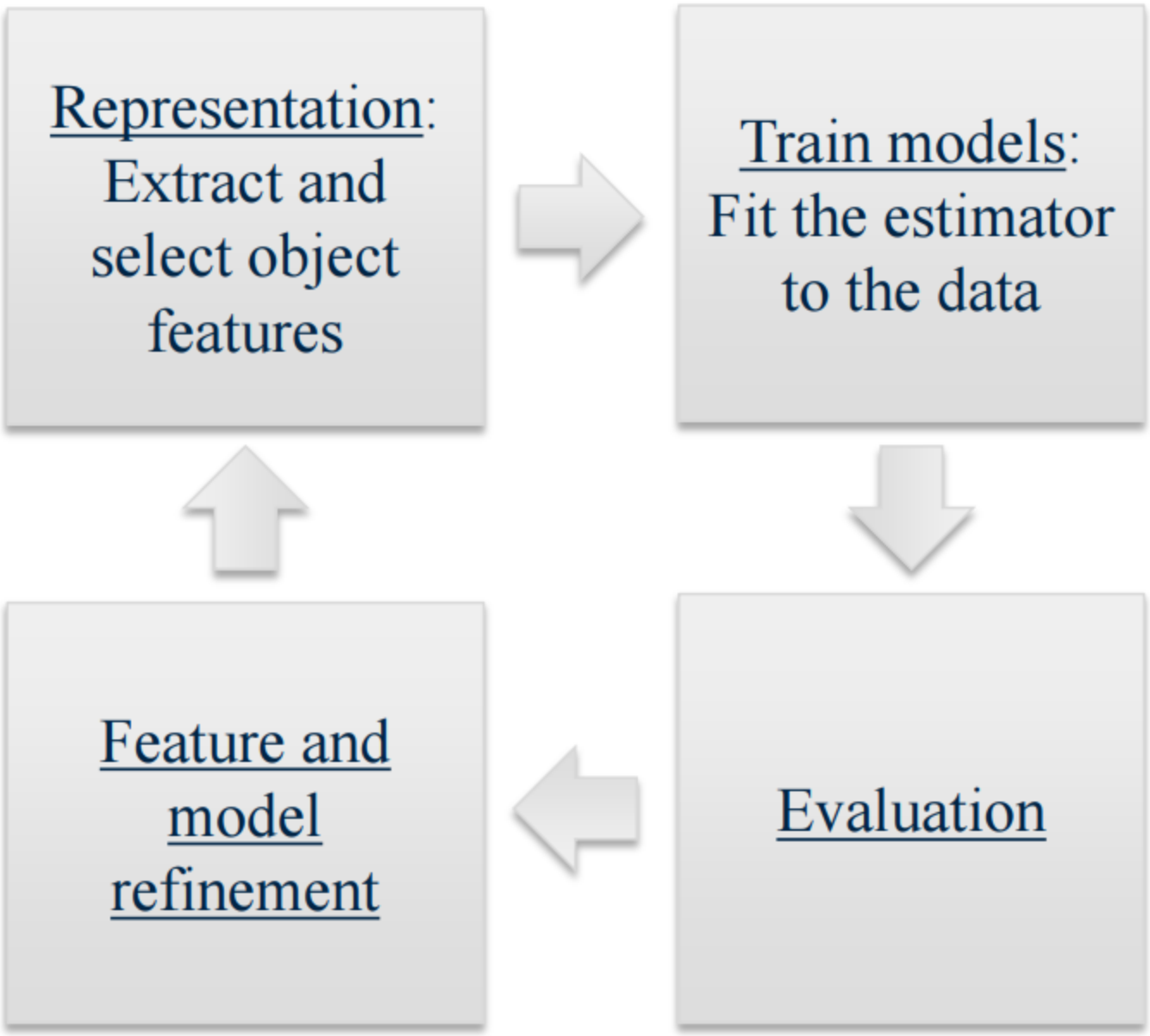
Choose:
- What criterion distinguishes good vs. bad classifiers?

e.g. % correct predictions on test set

Choose:
- How to search for the settings/parameters that give the best classifier for this evaluation criterion

e.g. try a range of values for "k" parameter in k-nearest neighbor classifier

Represent / Train / Evaluate / Refine Cycle



List of skills for data analysis

- Data Visualization
  - Matplotlib, Seaborn, Plotly
  - Data mining
  - Pandas, numpy

- Feature engineering
  - Time series features
  - Categorical features
  - Numerical features
  - Aggregation features
  - Ratio features
  - Product features

- Data preparation
  - Up-sampling
  - Down-sampling
  - SMOTE

- Model development
  - Sklearn : linear, non-linear, tree model
  - Xgboost
  - Lightgbm
  - Catboost
  - LibFFM