MACHINE LEARNING

Feature Extraction

Muhammad Afif Hendrawan, S.Kom., M.T.

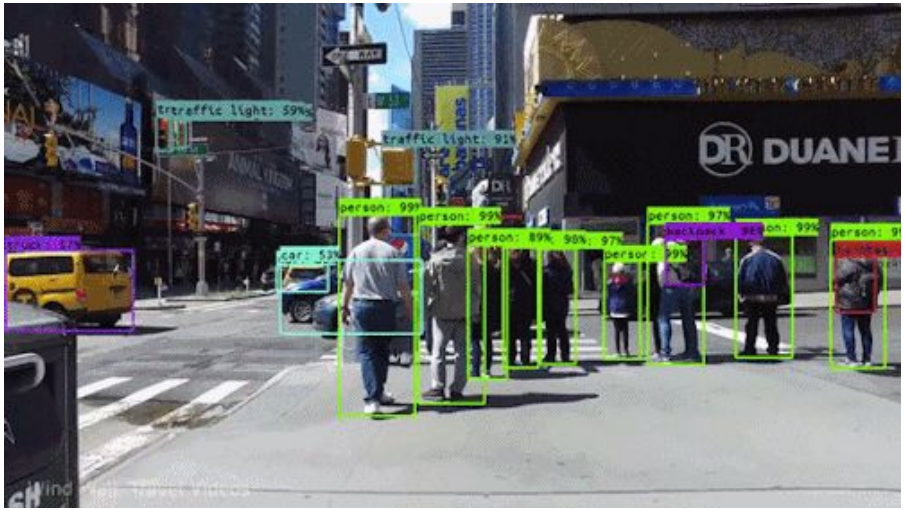
Outlines

- The Relation Between Recognition, Pattern, Feature, and Feature Extraction
- Feature Extraction on Various Types of Data
- Introduction to Preprocessing
  - Data Imputation
  - Normalization
  - Standardization
  - Encoding
- Splitting Data Strategy

The Relation Between Recognition, Pattern, Feature, and Feature Extraction

What is Recognition and Pattern?

- **Recognition** is regarded as a **basic attribute** of human beings, as well as other living organisms.
- Recognition is carried out by **looking for pattern**, namely, the description of an object



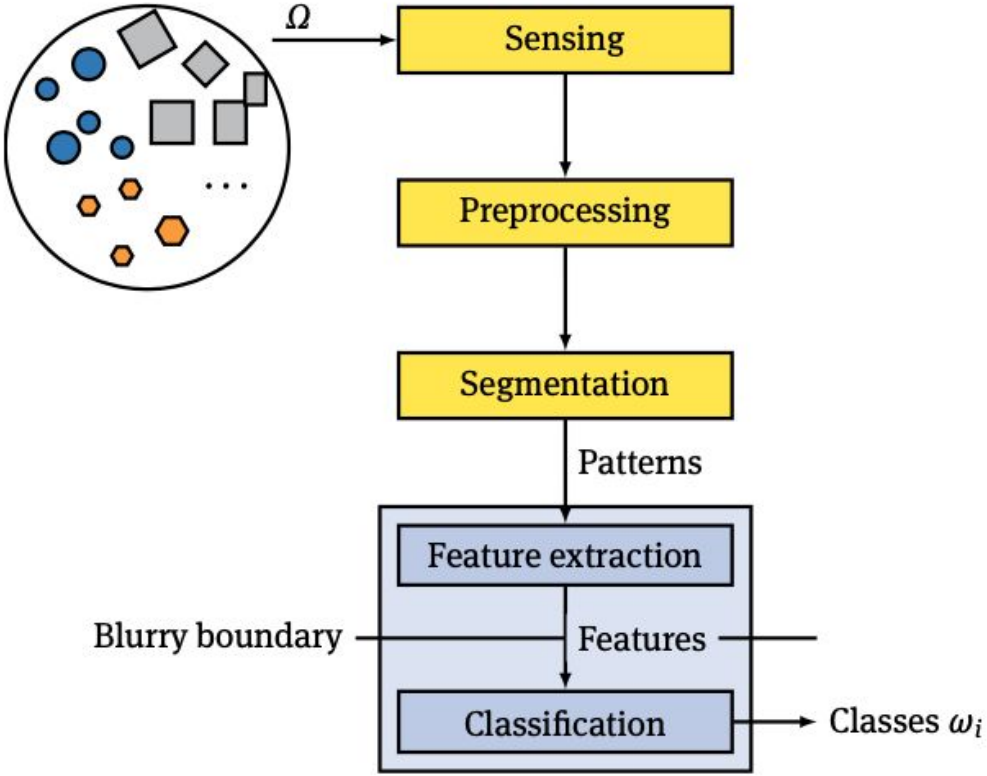
Types of Recognition

- The recognition of concrete items
  - Recognizing characters, pictures, musics, or objects around us.
  - This is called as sensory recognition (visual and aural pattern recognition).
- The recognition of abstract items
  - Recognizing old argument, solution of problem, while eyes and ear closes.
  - This is called as conceptual recognition.

Pattern Recognition

- The **categorization** of input data into identifiable class via the **extraction of significant features or attributes** of the data from a background of irrelevant detail.
- **Recognition** is to **distinguish and classify object**.
- **The problem** is **the discrimination** of the input data between population via the search of **features or invariant attributes** among members of population.

A Pattern Recognition System



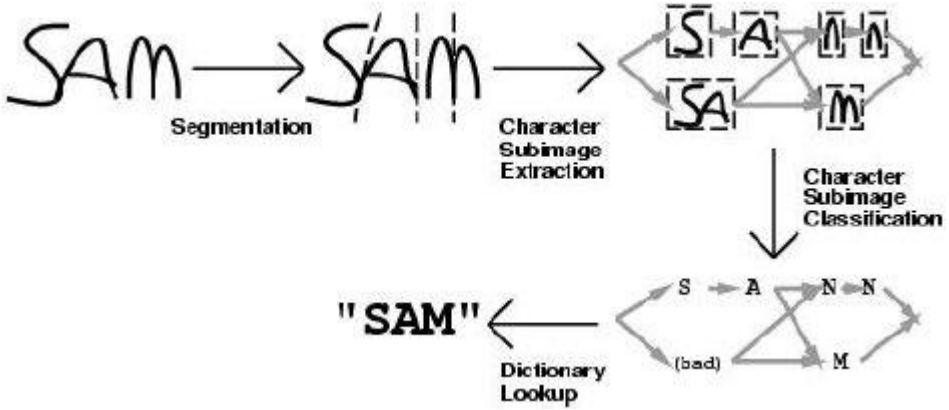
(Beyerer, J., et al, 2017)

Pattern Recognition on Classification Task

Task of Classification	Input Data	Output Response
Character recognition	Optical signals or strokes	Name of character
Speech recognition	Acoustic waveforms	Name of word
Speaker recognition	Voice	Name of speaker
Weather prediction	Weather maps	Weather forecast
Medical diagnosis	Symptoms	Disease
Stock market prediction	Financial news and charts	Predicted market ups and downs

What is feature?

- **Feature** is **numeric representation of raw data**.
- Feature is attribute, characteristic, traits, marks, or **something special about an object**.



(Breuel, 2001)

What is Feature Extraction?

A process of **dimensionality reduction** by which an initial set of raw data is reduced to more manageable groups of processing

These are dimensions

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Feature Extraction on Various Data

Feature Extraction on Structured Data

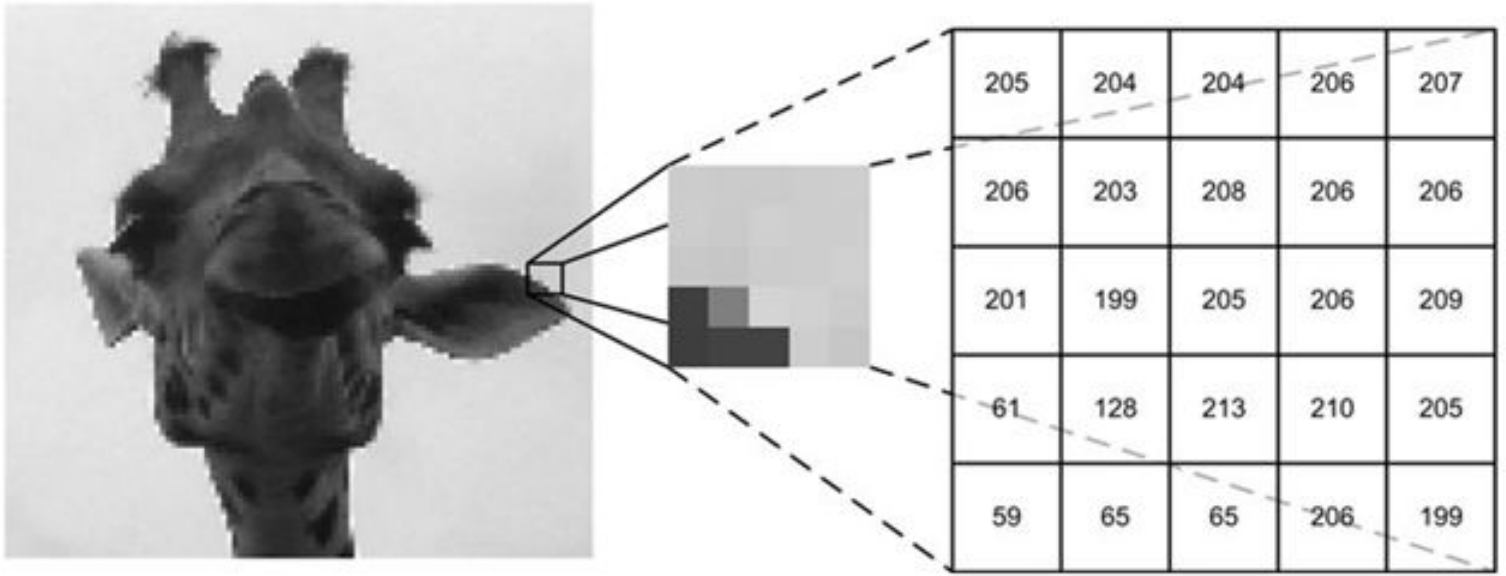
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

	Survived	Pclass	Age	Sex	Cabin
0	0	3	22.0	male	DECK
1	1	1	38.0	female	C85
2	1	3	26.0	female	DECK
3	1	1	35.0	female	C123
4	0	3	35.0	male	DECK

Feature Extraction on Semi-Structured Data

```
<University>
 <Student ID="1">
  <Name>John</Name>
  <Age>18</Age>
  <Degree>B.Sc.</Degree>
 </Student>
 <Student ID="2">
  <Name>David</Name>
  <Age>31</Age>
  <Degree>Ph.D. </Degree>
 </Student>
....
</University>
```

Feature Extraction on Unstructured Data



Introduction to Preprocessing

What is Preprocessing?

- The process of **manipulation or dropping of data before it is feeded or used** to another process in order **to ensure or enhance the performance of machine learning algorithms**
- There are several process that can be done in preprocessing, namely,
  - Data imputation
  - Normalization
  - Standardization
  - Resizing
  - Encoding
  - More . .

Data Imputation

- Filling the missing data with another value.
- Data imputation strategies,
  - Specific value
  - Mean
  - Median
  - Mode
  - KNN
  - Interpolation
  - More . .

Fare	Cabin	Embarked
7.2500	NaN	S
71.2833	C85	C
7.9250	NaN	S
53.1000	C123	S
8.0500	NaN	S

	Survived	Pclass	Age	Sex	Cabin
0	0	3	22.0	male	DECK
1	1	1	38.0	female	C85
2	1	3	26.0	female	DECK
3	1	1	35.0	female	C123
4	0	3	35.0	male	DECK

Normalization

- Scaled the data into range of 0 - 1.
- Using this following formula

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization

- Standardized the data by removing the mean (mean=0) the scaling the standard deviation to 1
- Using this following formula

$$x'' = \frac{x - \mu}{\sigma}$$

Encoding

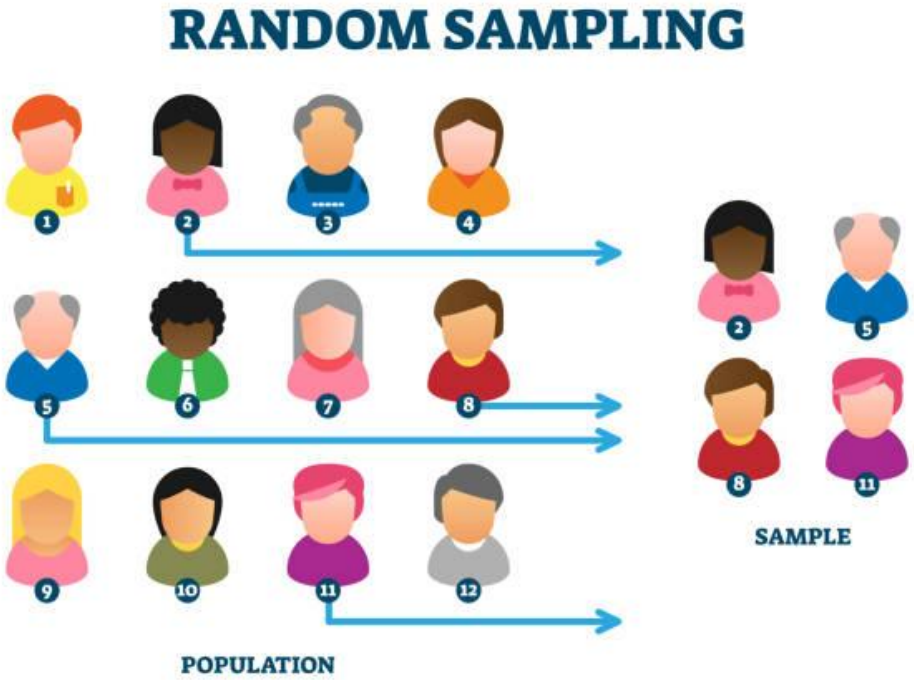
- The process of converting the data into specified format
- Several encoding strategy,
  - Ordinal Encoding
  - Label Encoding
  - One-Hot Encoding
  - Dummy Encoding

	Survived	Pclass	Age	Sex	Cabin
0	0	3	22.0	male	DECK
1	1	1	38.0	female	C85
2	1	3	26.0	female	DECK
3	1	1	35.0	female	C123
4	0	3	35.0	male	DECK

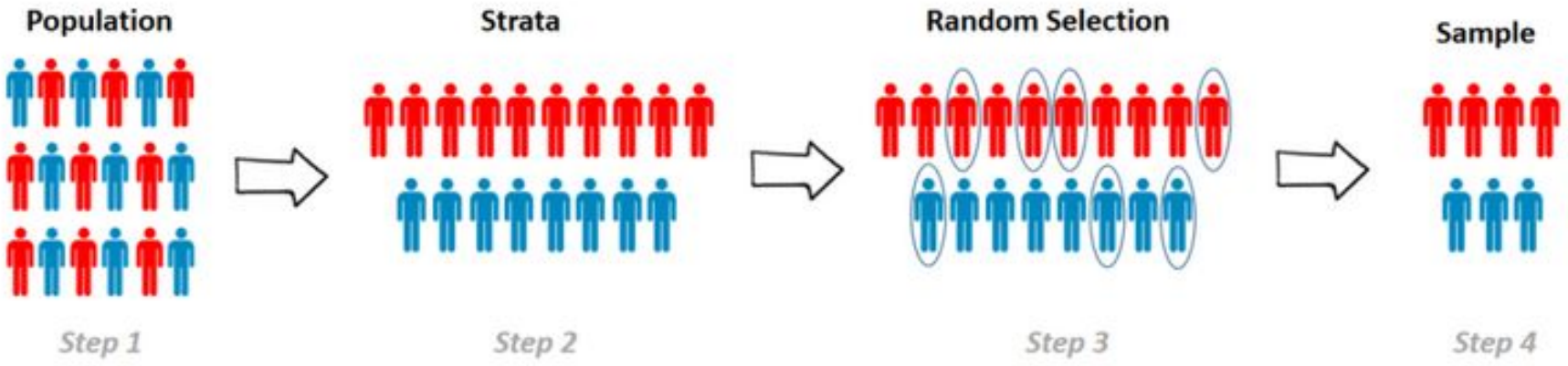
	Survived	Pclass	Age	Sex	Cabin
0	0	3	22.0	1	115
1	1	1	38.0	0	81
2	1	3	26.0	0	115
3	1	1	35.0	0	55
4	0	3	35.0	1	115

Splitting Data Strategy

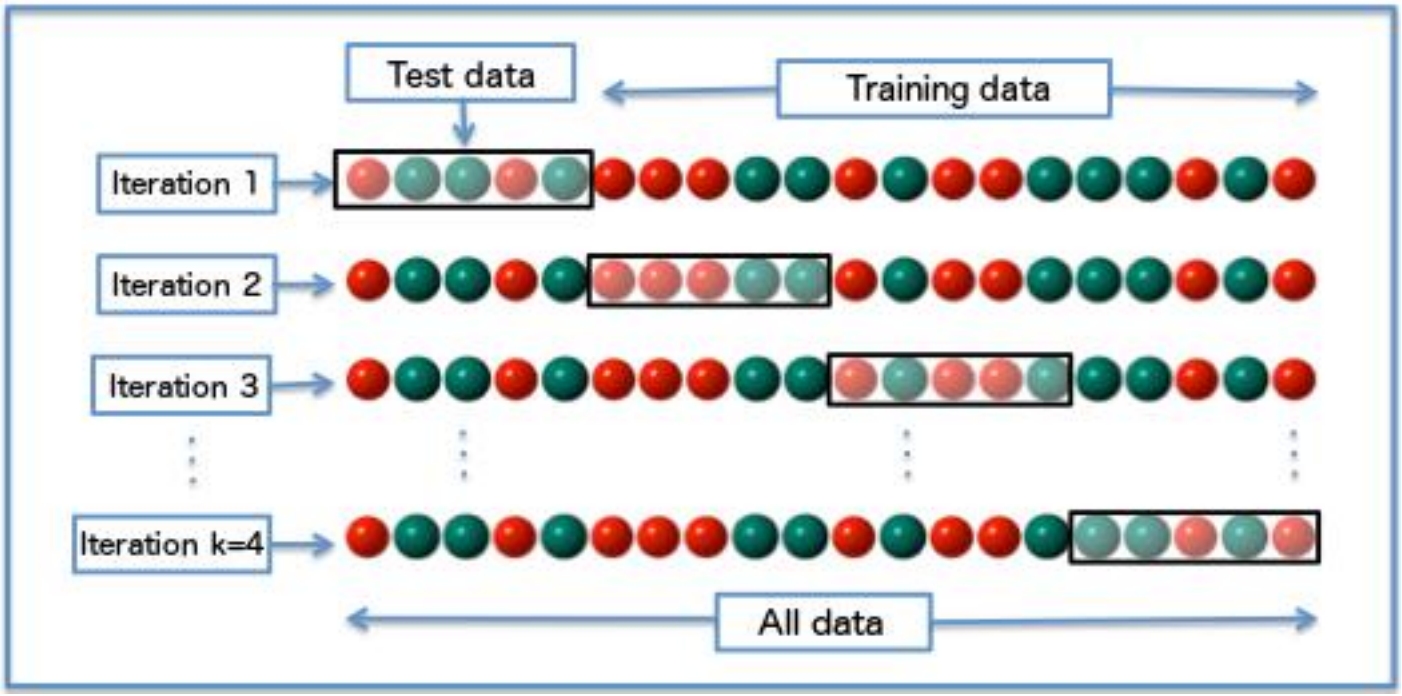
Random Split



Stratified Split



Cross Validation



Let’s get your hands dirty!

Feature Extraction Practice!