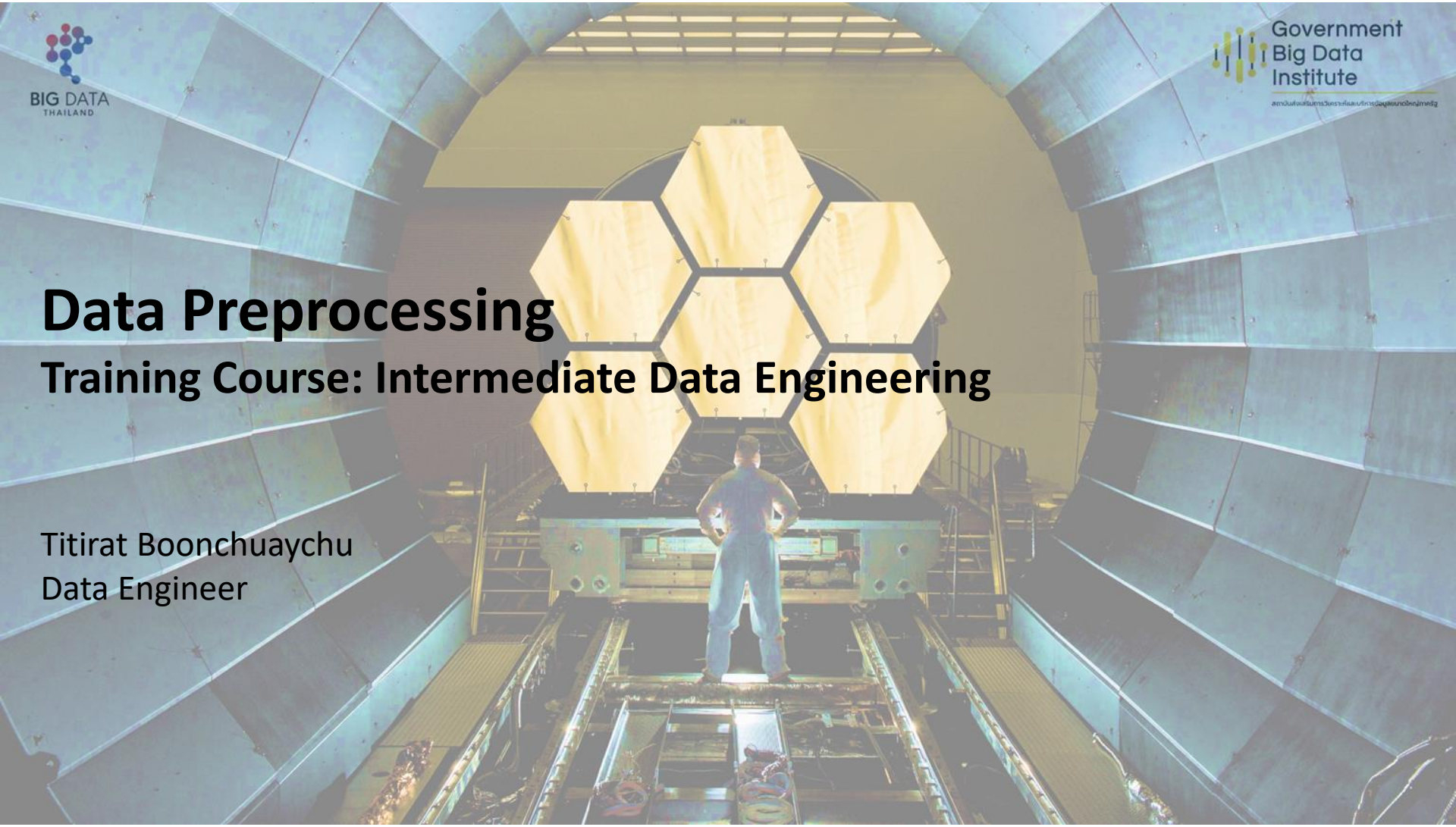


Data Preprocessing

Training Course: Intermediate Data Engineering

Titirat Boonchuaychu
Data Engineer

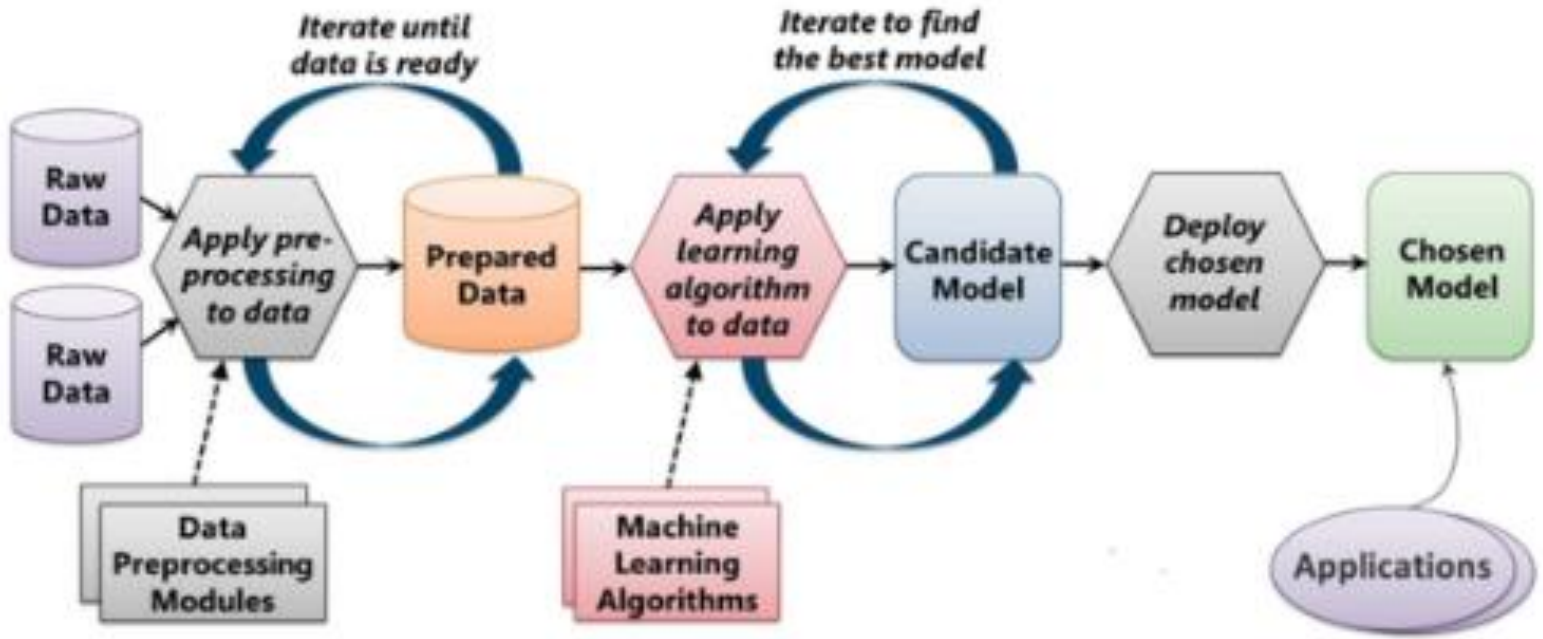
Agenda

- ❑ What Is Data Preprocessing?
- ❑ Data Preprocessing Importance
- ❑ Data Preprocessing Steps
  - Data Quality Assessment
  - Data cleaning
  - Data integration
  - Data reduction
  - Data transformation

What Is Data Preprocessing?

- The process of **transforming raw data** into an **understandable format**.
- Data preprocessing transforms the data into a format that is **more easily** and **effectively processed** in **data mining**, **machine learning** and other data science tasks.

The Machine Learning Process



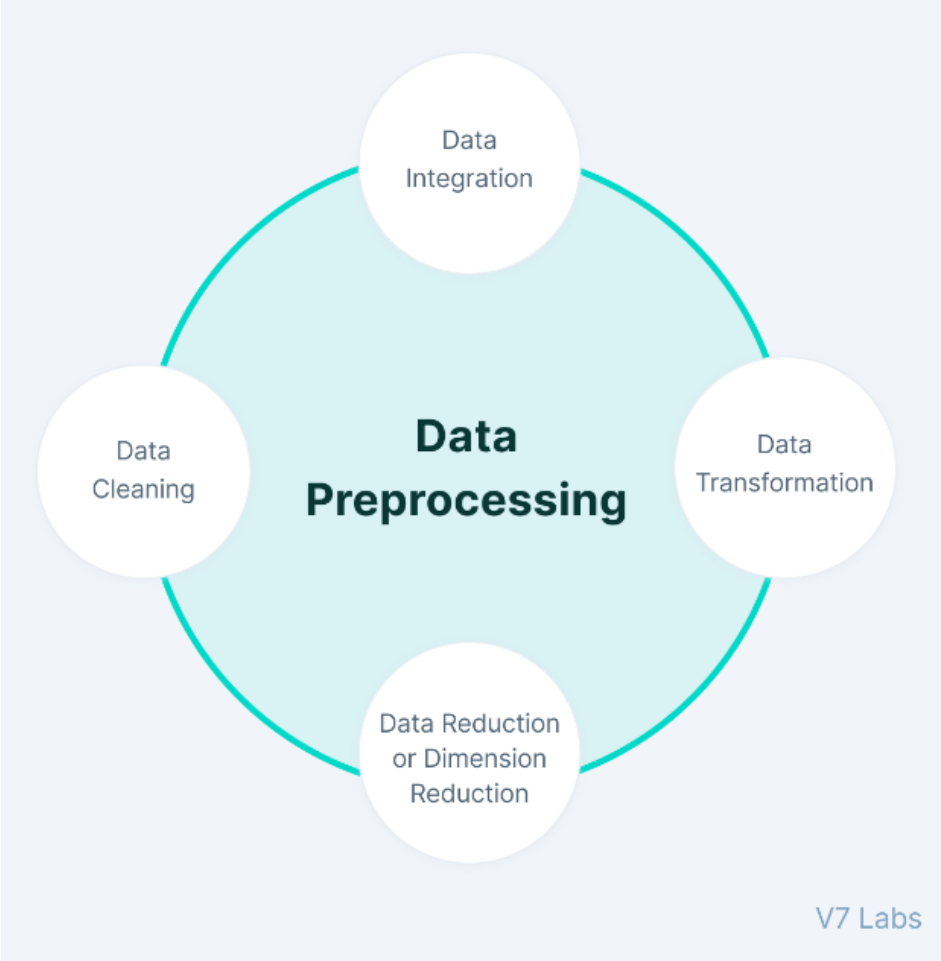
From "Introduction to Microsoft Azure" by David Chappell

Data Preprocessing Importance

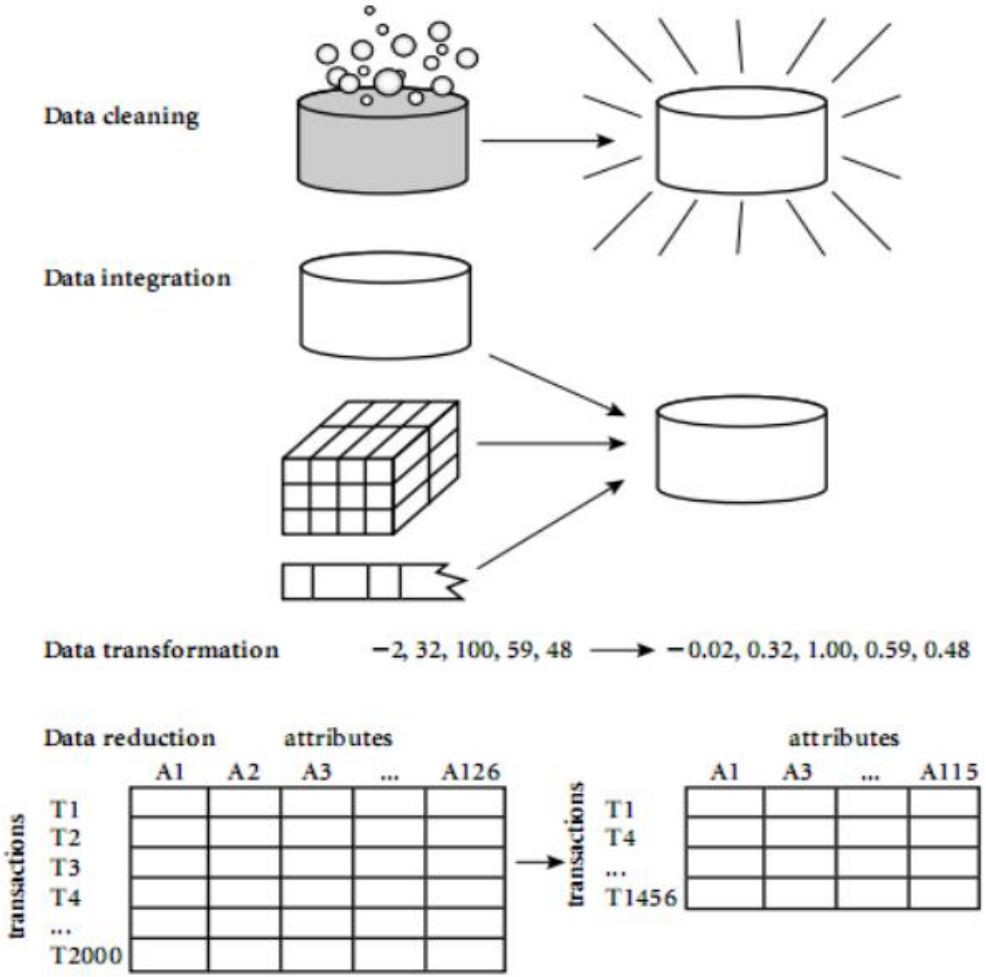
- Applying data mining algorithms on this **noisy data** would not give quality results as they would **fail** to **identify patterns effectively**.
- **Quality decisions** must be **based on quality data**. *Data Preprocessing is important to get this quality data*, without which it would just be a Garbage In, Garbage Out scenario.

Source: Garbage in/ out scenario

Data Preprocessing Steps



Data Preprocessing Steps



Source: [Data preprocessing]

Data Preprocessing Steps

1. Data Quality Assessment

- ❑ The completeness with **no missing** attribute values
- ❑ **Accuracy** and **reliability** in terms of information

Data Preprocessing Steps

1. Data Quality Assessment

- ❑ **Consistency** in all features
  - Mismatched data types

id	monthly_income	currency	country
1	20000	Thai baht	Thailand
2	860	United States Dollar	United States

id	monthly_income	currency	country
1	20000	Thai baht	Thailand
2	30100	Thai baht	United States

Convert to a single currency

Data Preprocessing Steps

1. Data Quality Assessment

- ❑ Consistency in all features
  - Mixed data values

Source 1

id	firstname	lastname	gender
1	John	Leno	man

Source 2

id	firstname	lastname	gender
1	Matt	Json	male

Convert gender to *male*

Data Preprocessing Steps

1. Data Quality Assessment

- ❑ Maintain data **validity**
  - Data outliers
- ❑ It does **not** contain any **redundancy**

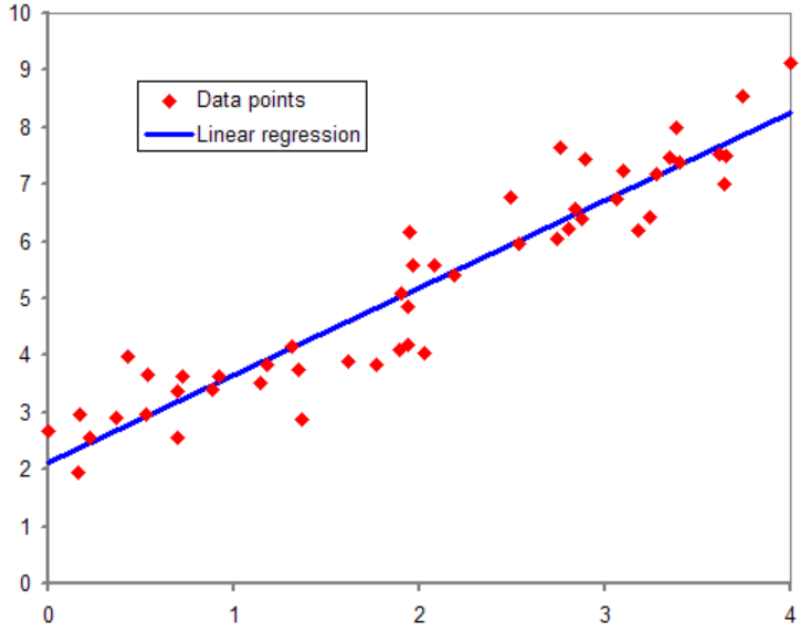
Data Preprocessing Steps

2. Data cleaning

- **Missing** data
  - Ignore the tuples
  - Filling in the values manually, predicting the missing values

predicting the missing values (examples)

- mean
- regression method

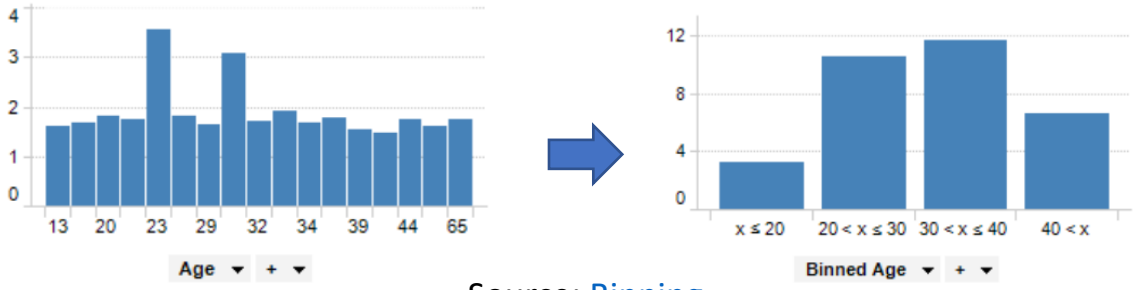


Source: Regression

Data Preprocessing Steps

2. Data cleaning

- **Noisy** data
  - Binning



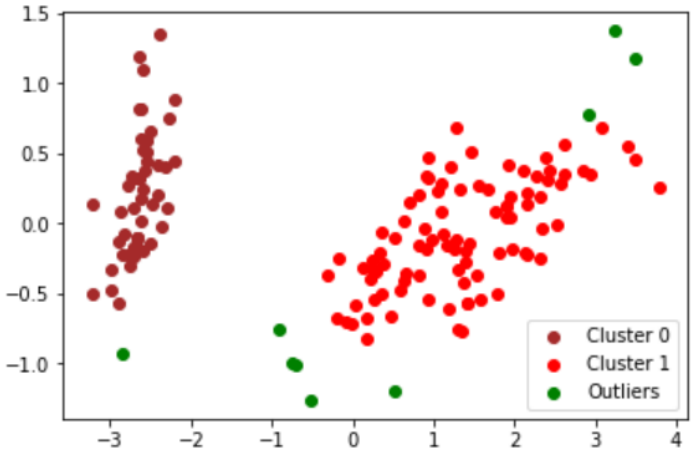
Source: Binning

  - Regression

Data Preprocessing Steps

2. Data cleaning

- ❑ Noisy data
  - Clustering



Source: Clustering

Data Preprocessing Steps

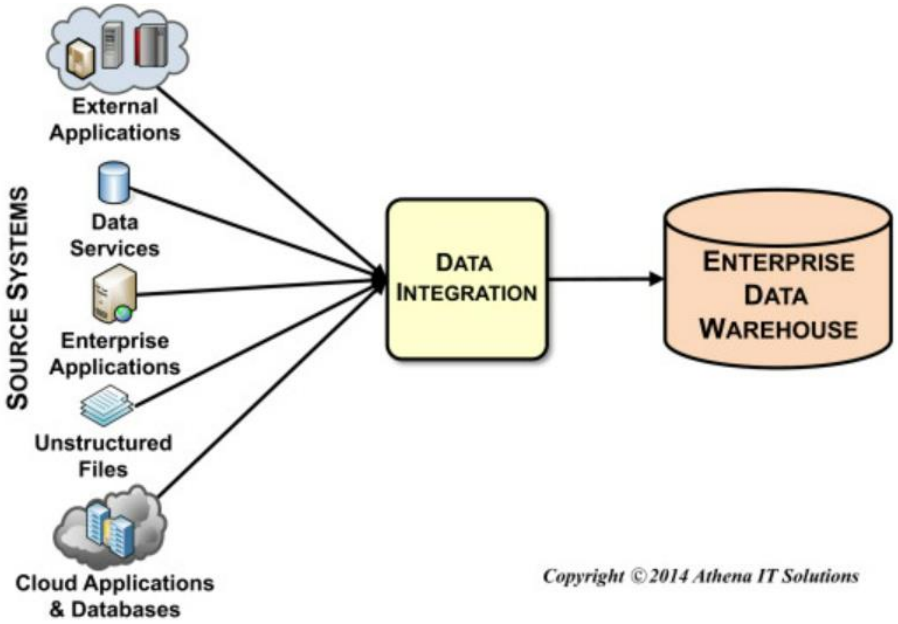
2. Data cleaning

- ❑ Removing **outliers**
  - Clustering

Data Preprocessing Steps

3. Data integration

- Schema integration and object matching



Data Preprocessing Steps

3. Data integration

- Removing **redundant** attributes from all data sources
- Detection and resolution of data value **conflicts**

Data Preprocessing Steps

4. Data reduction

- **Attribute selection**, also known as feature selection

  The feature selection methods

  - Filter method
  - Wrapper method
  - Embedded method

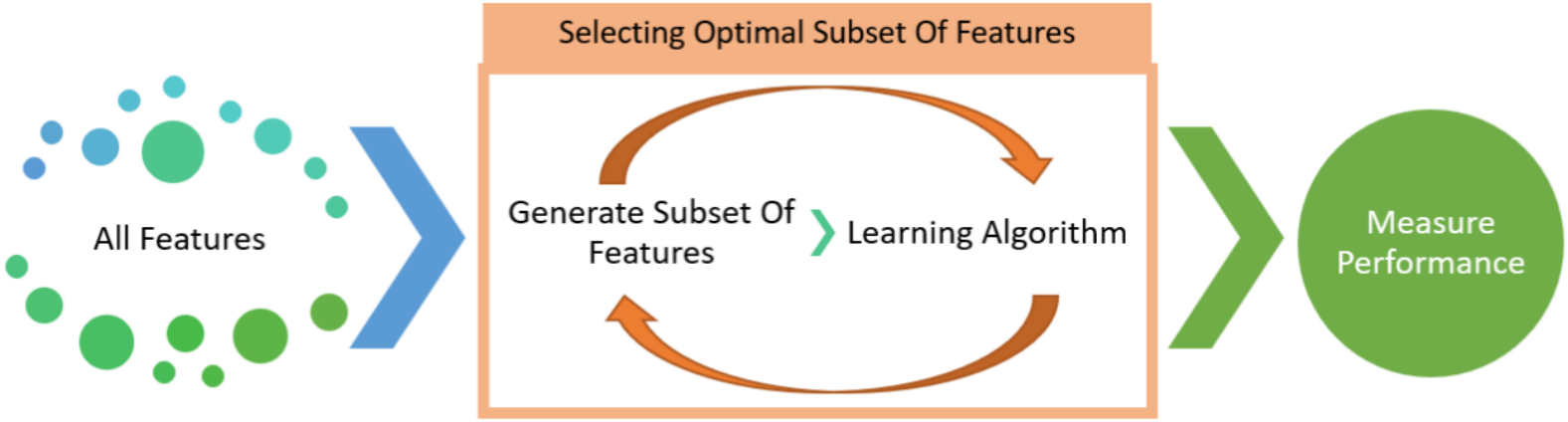
Filter method

- **Based on attributes** of the feature.
- Performance does not depend on model used.
- Effective in computation time and **robust to overfitting**.
- This method should be used for preliminary screening.
- **Advantage**: Filter methods are much **faster** compared to wrapper methods.
- **Disadvantage**: Usually ***not*** *the* ***best performance*** in terms of reducing features.

Source: Filter method

Wrapper method

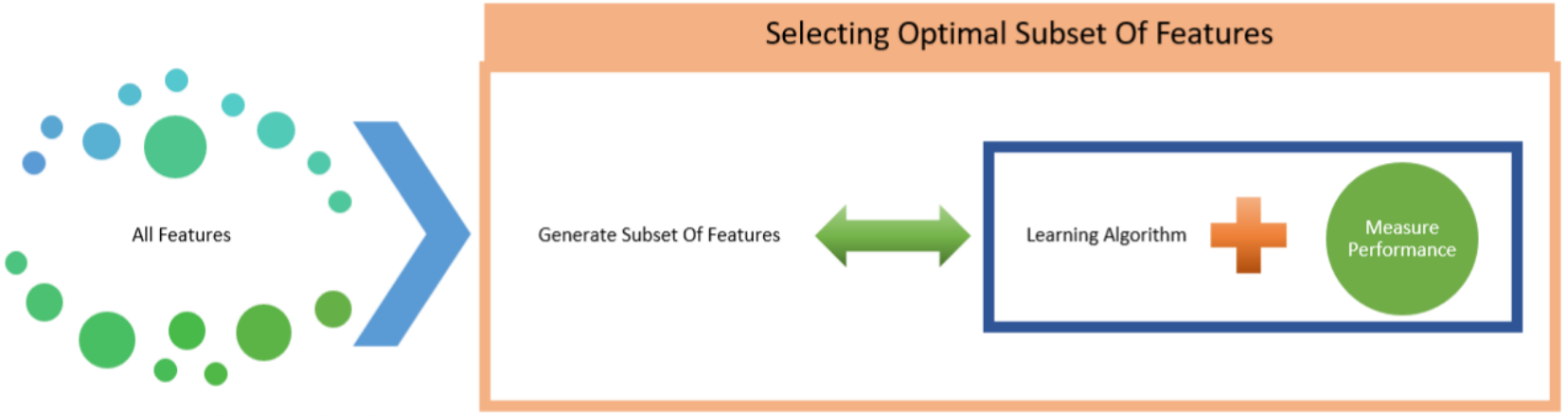
- Performance **depends on model selected** and data underlying.
- Usually can suggests the optimal feature subset.
- Typically very computationally expensive.
- **Advantage**: Possibly ***best performance*** in terms of feature elimination.
- **Disadvantage**: **Terribly slow** when it comes to **large datasets**.



Source: [Wrapper method]

Embedded method

- **Combine** the advantages of both previous methods.
- A learning algorithm takes advantage of its own variable selection process and performs feature selection and classification simultaneously.
- Generally less computationally expensive than Wrapper methods.



Source: Embedded method

Application of feature selection metaheuristics [edit]

This is a survey of the application of feature selection metaheuristics lately used in the literature. This survey was realized by J. Hammon in her 2013 thesis.[47]

Application	Algorithm	Approach	Classifier	Evaluation Function	Reference
SNPs	Feature Selection using Feature Similarity	Filter		r^2	Phuong 2005[49]
SNPs	Genetic algorithm	Wrapper	Decision Tree	Classification accuracy (10-fold)	Shah 2004[51]
SNPs	Hill climbing	Filter + Wrapper	Naive Bayesian	Predicted residual sum of squares	Long 2007[52]
SNPs	Simulated annealing		Naive bayesian	Classification accuracy (5-fold)	Ustunkar 2011[53]
Segments parole	Ant colony	Wrapper	Artificial Neural Network	MSE	Al-ani 2005[*citation needed*]
Marketing	Simulated annealing	Wrapper	Regression	AIC, r^2	Meiri 2006[54]
Economics	Simulated annealing, genetic algorithm	Wrapper	Regression	BIC	Kapetanios 2007[55]
Spectral Mass	Genetic algorithm	Wrapper	Multiple Linear Regression, Partial Least Squares	root-mean-square error of prediction	Broadhurst et al. 1997[56]
Spam	Binary PSO + Mutation	Wrapper	Decision tree	weighted cost	Zhang 2014[25]
Microarray	Tabu search + PSO	Wrapper	Support Vector Machine, K Nearest Neighbors	Euclidean Distance	Chuang 2009[57]
Microarray	PSO + Genetic algorithm	Wrapper	Support Vector Machine	Classification accuracy (10-fold)	Alba 2007[58]
Microarray	Genetic algorithm + Iterated Local Search	Embedded	Support Vector Machine	Classification accuracy (10-fold)	Duval 2009[59]
Microarray	Iterated local search	Wrapper	Regression	Posterior Probability	Hans 2007[60]
Microarray	Genetic algorithm	Wrapper	K Nearest Neighbors	Classification accuracy (Leave-one-out cross-validation)	Jirapech-Umpai 2005[61]
Microarray	Hybrid genetic algorithm	Wrapper	K Nearest Neighbors	Classification accuracy (Leave-one-out cross-validation)	Oh 2004[62]
Microarray	Genetic algorithm	Wrapper	Support Vector Machine	Sensitivity and specificity	Xuan 2011[63]
Microarray	Genetic algorithm	Wrapper	All paired Support Vector Machine	Classification accuracy (Leave-one-out cross-validation)	Peng 2003[64]
Microarray	Genetic algorithm	Embedded	Support Vector Machine	Classification accuracy (10-fold)	Hernandez 2007[65]
Microarray	Genetic algorithm	Hybrid	Support Vector Machine	Classification accuracy (Leave-one-out cross-validation)	Huerta 2006[66]

Data Preprocessing Steps

4. Data reduction

- **Numerosity reduction**: Replaces the original data by **smaller form** of data **representation**.
  - Parametric Methods
  - Non-Parametric Methods

Parametric Methods (examples)

- Regression

  Linear regression: $y = ax+b$

  x and y are random variable

  a and b are regression coefficients

Non-Parametric Methods (examples)

- **Histograms**: The data representation in terms of **frequency**.
- **Clustering**: The **cluster** representation of the data are used to **replace** the **actual data**.
- **Sampling**: Allows a large data set to be represented by a much **smaller** random **data sample** (or subset).

Data Preprocessing Steps

4. Data reduction

- **Dimensionality** reduction
  - **Feature Selection**: **Removing** the irrelevant or **redundant** features.
  - **Feature Extraction**: Reduce the number of features in a dataset by **creating new features** from the existing ones (and then discarding the original features).

Data Preprocessing Steps

5. Data transformation

- **Aggregation**: **Combines** all of your data together in a **uniform format**.
- **Normalization**: **Scales** your data into a regularized **range** so that you can **compare** it more accurately.
  - Min-Max normalization
  - Z-Score normalization
  - Decimal scaling normalization

Min-Max normalization

- The linear transformation of the original unstructured data.

$$v' = \frac{v - minF}{maxF - minF}(new\_max_F - new\_min_F) + new\_min_F$$

where is the current value of feature F.

Z-Score normalization

- It is also called zero-mean normalization.
- Scale where an average number equals zero and a standard deviation is one.

$$v' = \frac{v - \bar{F}}{\sigma_F}$$

Here is the mean and is the standard deviation of feature F.

Decimal scaling normalization

- Move the decimal point of values of the attribute.
- This movement of decimal points totally depends on the maximum value among all values in the attribute.

$$v' = \frac{v}{10^j}$$

In this formula, j is the lowest integer while Max (|) < 1.

Data Preprocessing Steps

5. Data transformation

- **Feature selection**: New properties of data are created from existing attributes to help in the data mining process.

Data Preprocessing Steps

5. Data transformation

- **Discretization**
  - The **continuous data** here is split into intervals.
  - Discretization pools data into **smaller intervals**.
  - It's somewhat similar to binning, but usually happens **after data** has been **cleaned**.

Data Preprocessing Steps

5. Data transformation

- ❑ Concept hierarchy generation
  - A **hierarchy** within and **between** your **features** that wasn't present in the original data.


BIG DATA
THAILAND
Government
Big Data
Institute
End the section