FACULTY OF INFORMATION TECHNOLOGY

Machine Learning

(Học Máy)

Semester 2, 2023/2024

Situation …

- We are drowning in data, but starving for knowledge!

"Looks like you've got all the data – what's the holdup?"

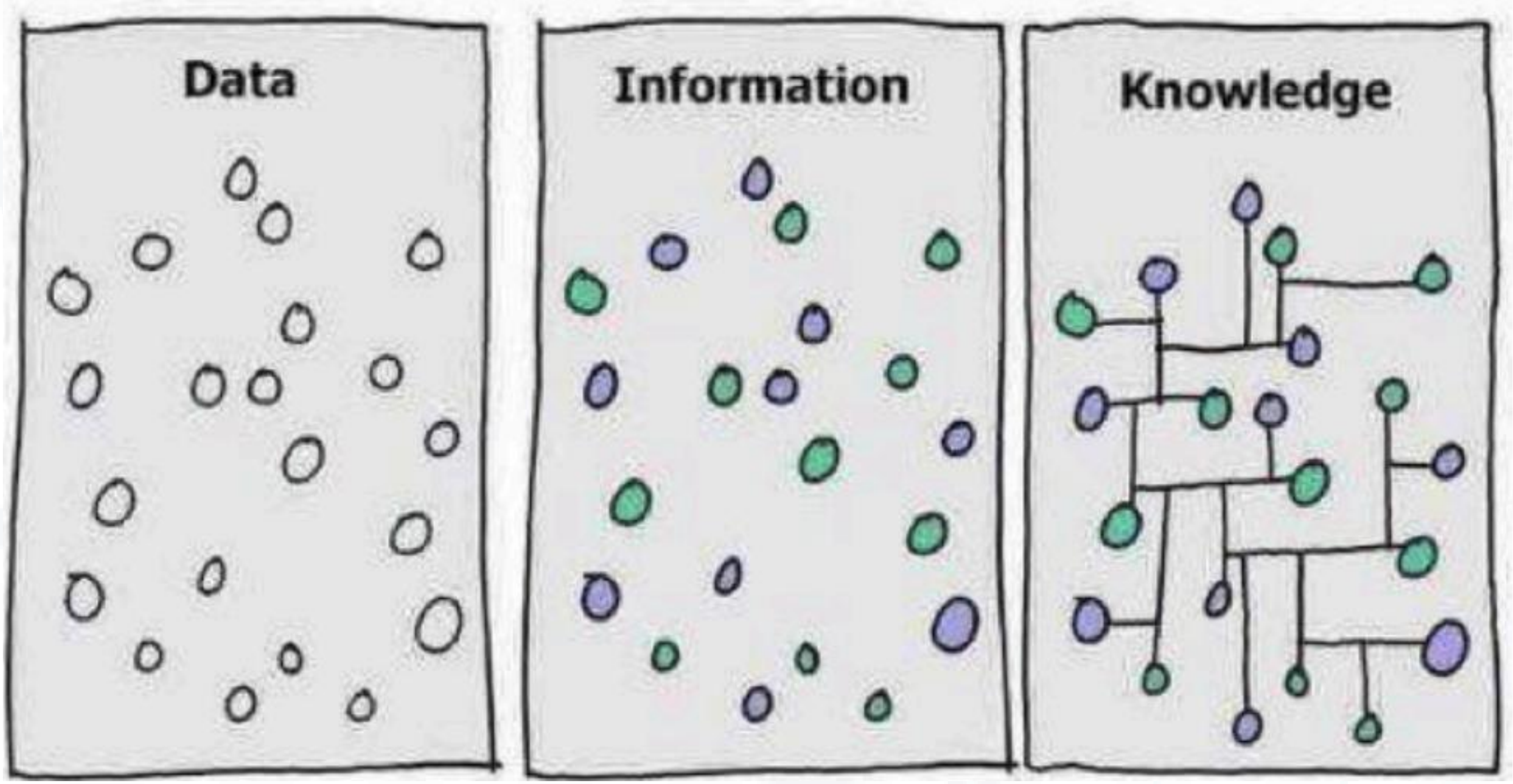
What are Data?

- It can be any unprocessed fact, value, text, sound, or picture that is not being interpreted and analyzed.

- Data are the most important part of all Data Analytics, Machine Learning, Artificial Intelligence.

- Without data, we can’t train any model and all modern research and automation will go in vain

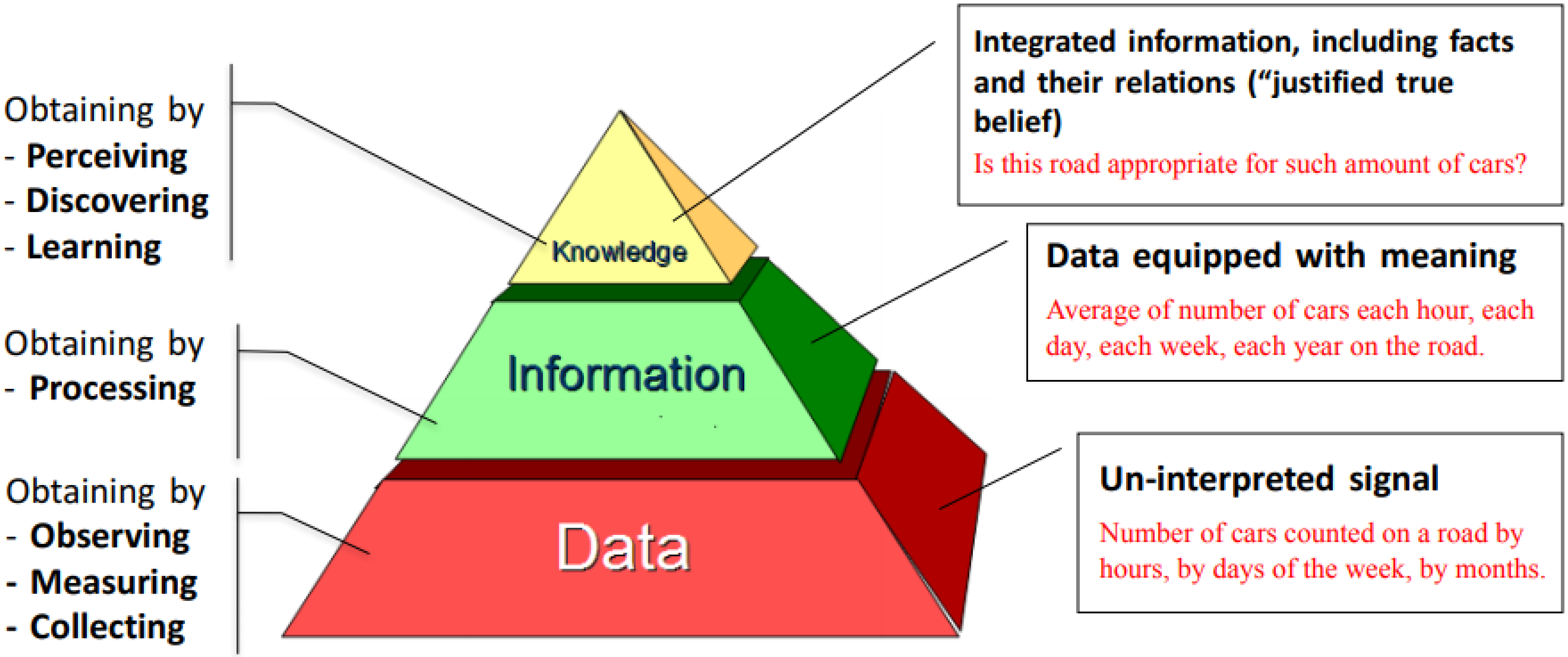
Data, Information, Knowledge



From Julien Blin

Data, Information, Knowledge (cont.)

- Knowledge can be considered data at a high level of abstraction and generalization.



Obtaining by
- **Perceiving**
- **Discovering**
- **Learning**

Obtaining by
- **Processing**

Obtaining by
- **Observing**
- **Measuring**
- **Collecting**

Integrated information, including facts and their relations ("justified true belief)

Is this road appropriate for such amount of cars?

Data equipped with meaning

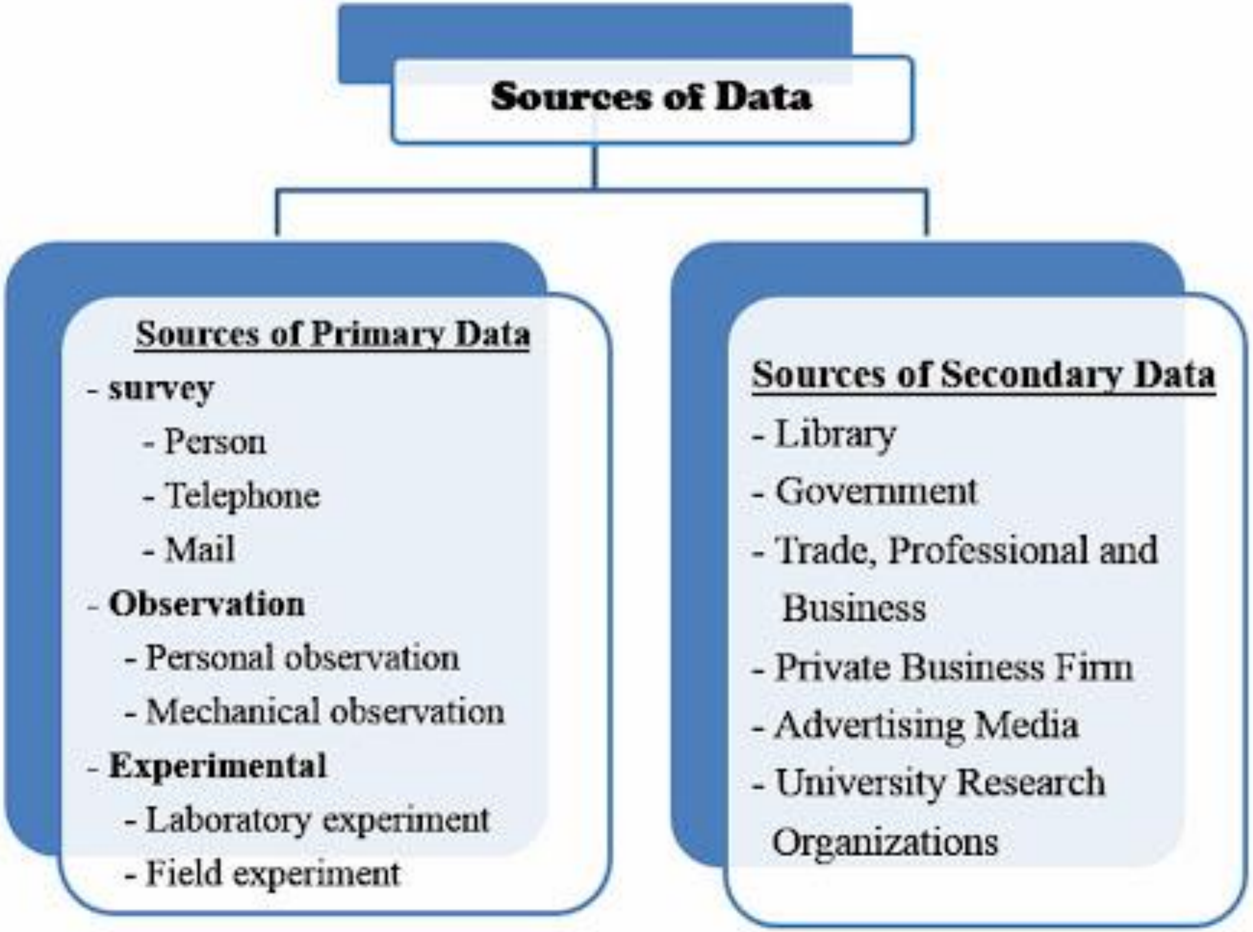
Average of number of cars each hour, each day, each week, each year on the road.

Un-interpreted signal

Number of cars counted on a road by hours, by days of the week, by months.

Where do data come from?

Where do data come from?



Sources of Data

Sources of Primary Data

- **survey**
  - Person
  - Telephone
  - Mail
- **Observation**
  - Personal observation
  - Mechanical observation
- **Experimental**
  - Laboratory experiment
  - Field experiment

Sources of Secondary Data

- Library
- Government
- Trade, Professional and Business
- Private Business Firm
- Advertising Media
- University Research Organizations

What are Structured Data?

- Collection of data *objects* and their *attributes*
- An *attribute* is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an *object*
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Types of Attributes

- Nominal Attributes:
  - The values of a nominal attribute are symbols or names of things (referred to as categorical)
  - Examples: eye color (brown, blue, …), zip codes, strings
- Ordinal Attributes:
  - An attribute with possible values that have a meaningful order or ranking among them
  - Examples: grade (e.g., A+, A, A−, B+, B, B−, and so on), size, ...

Types of Attributes (cont.)

- Binary Attributes:
  - A nominal attribute with only 2 categories or states: 0 (absent) or 1 (present)
    - Symmetric binary: both outcomes are equally important
      - Example: gender
    - Asymmetric binary: outcomes are not equally important
      - Example: medical test (positive vs. negative),
      - Convention: assign 1 to most important outcome (e.g., HIV positive)

Types of Attributes (cont.)

- Numeric Attributes:
  - a measurable quantity (integer or real values)
  - Examples: dates, temperature, time, length, value, count.
  - Special case: Binary/Boolean attributes (yes/no, exists/not exists)

- Discrete (counts) vs Continuous (temperature)

Numeric Relational Data

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points/vectors in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an n-by-d data matrix, where there are n rows, one for each object, and d columns, one for each attribute

Temperature	Humidity	Pressure
30	0.8	90
32	0.5	80
24	0.3	95

Categorical Relational Data

- Data that consists of a collection of records, each of which consists of a fixed set of categorical attributes

ID Number	Zip Code	Marital Status	Income Bracket
1129842	45221	Single	High
2342345	45223	Married	Low
1234542	45221	Divorced	High
1243535	45224	Single	Medium

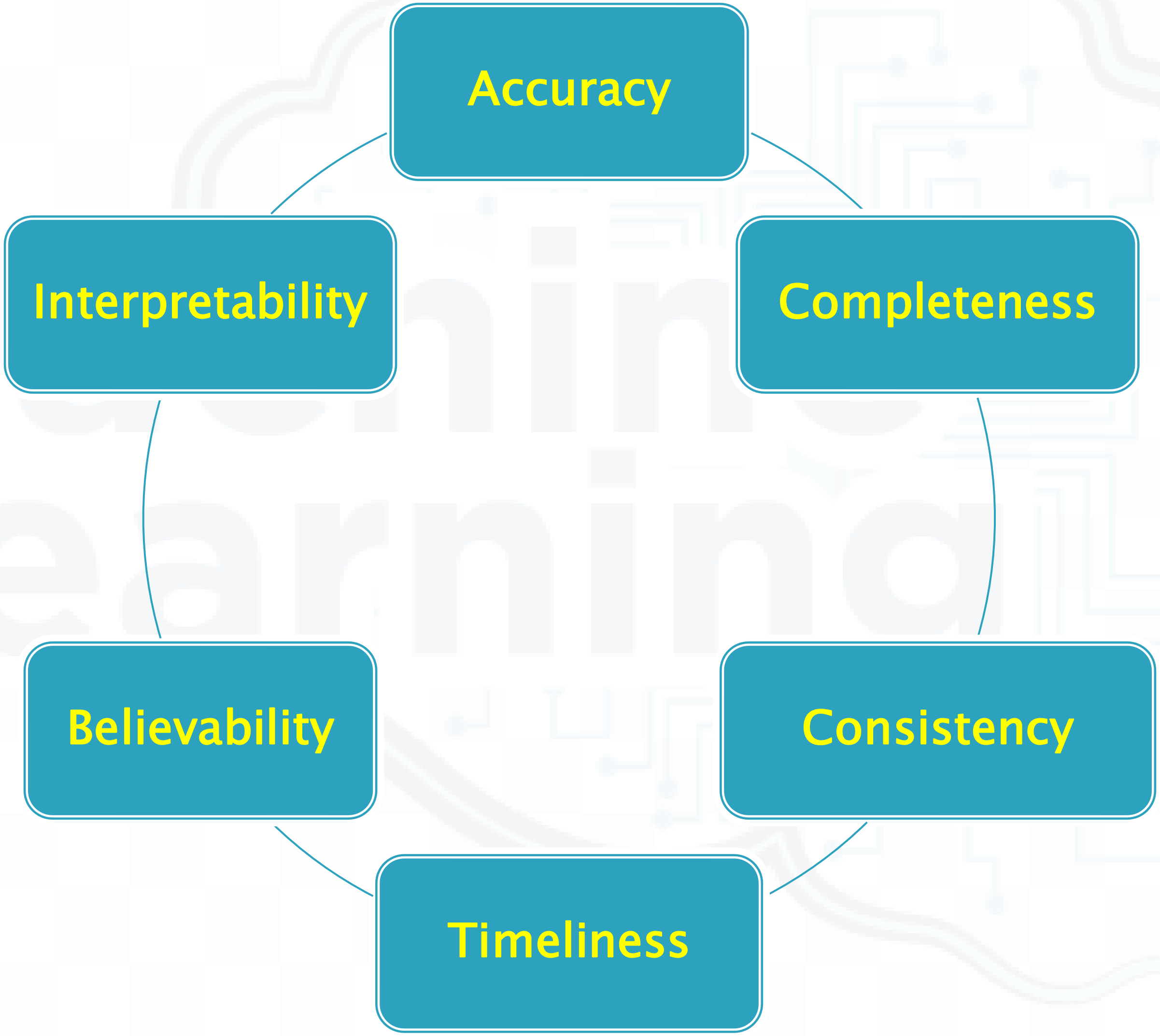
Mixed Relational Data

- Data that consists of a collection of records, each of which consists of a fixed set of both numeric and categorical attributes

ID Number	Zip Code	Age	Marital Status	Income	Income Bracket	Refund
1129842	45221	55	Single	25000	High	0
2342345	45223	25	Married	3000	Low	1
1234542	45221	45	Divorced	200000	High	0
1243535	45224	43	Single	150000	Medium	0

Boolean attributes can be thought as both numeric and categorical

Data Quality



Accuracy

Completeness

Consistency

Timeliness

Believability

Interpretability

Data Quality: Why Preprocess the Data?

- Accuracy:
  - correct or wrong, accurate or inaccurate
- Completeness:
  - not recorded, unavailable, …
- Consistency:
  - Whether the same data kept at different places do or do not match? some modified but some not, dangling, …
- Timeliness:
  - timely update?
- Believability:
  - how trustable the data are correct?
- Interpretability:
  - how easily the data can be understood?

Data Quality Issues – Examples

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., Occupation = " " (missing data)
  - Noisy: containing noise, errors, or outliers
    - e.g., Salary = "−10" (an error)
  - Inconsistent: containing discrepancies in codes or names, e.g.,
    - Age = "42", Birthday = "03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., disguised missing data)
    - Jan. 1 as everyone's birthday?

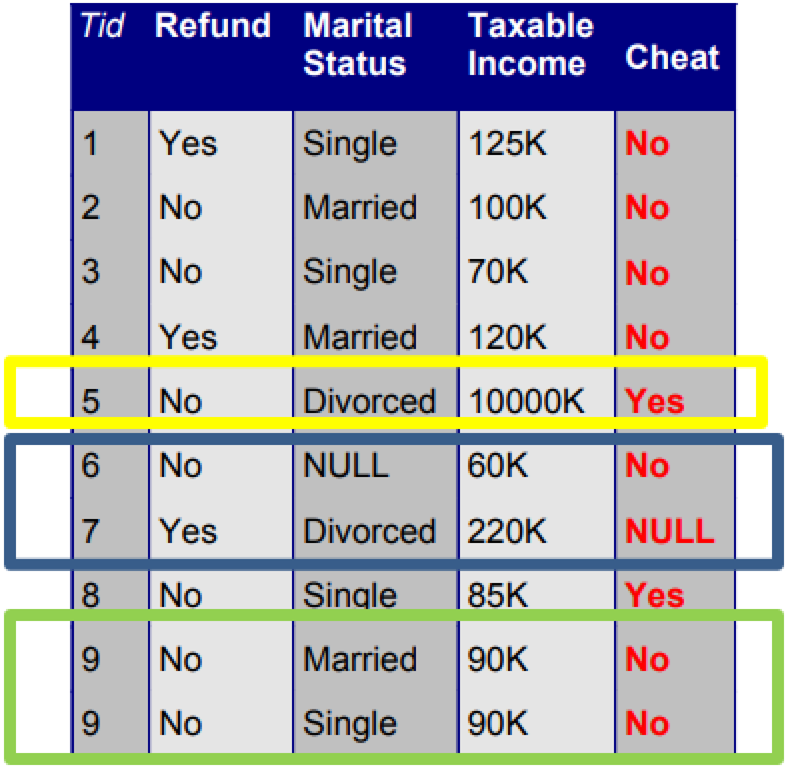
Examples of data quality problems

- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

A mistake or a millionaire?

Missing values

Inconsistent duplicate entries



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

Descriptive statistics

- Central tendency:
  - Mean:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

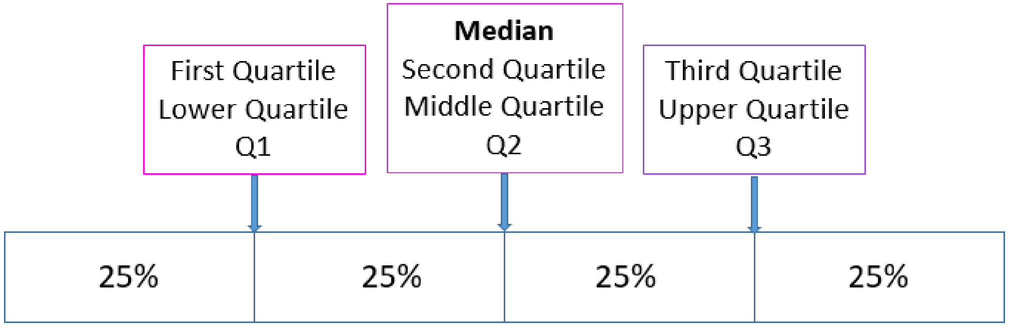
  - Median:

$$Median = \begin{cases} x_{[(n+1)/2]} & \text{if } n\, odd \\ \dfrac{x_{[n/2]} + x_{[(n/2)+1]}}{2} & \text{if } n\, even \end{cases}$$

  - Mode: the value that occurs most often in the dataset
  - Midrange: (Max + Min)/2

Descriptive statistics (cont.)

- Quartiles – tứ phân vị:
  - The first quartile (Q1): the 25th percentile
  - The second quartile (Q2): the 50th percentile (median)
  - The third quartile (Q3): the 75th percentile



Median
Second Quartile
Middle Quartile
Q2

First Quartile
Lower Quartile
Q1

Third Quartile
Upper Quartile
Q3

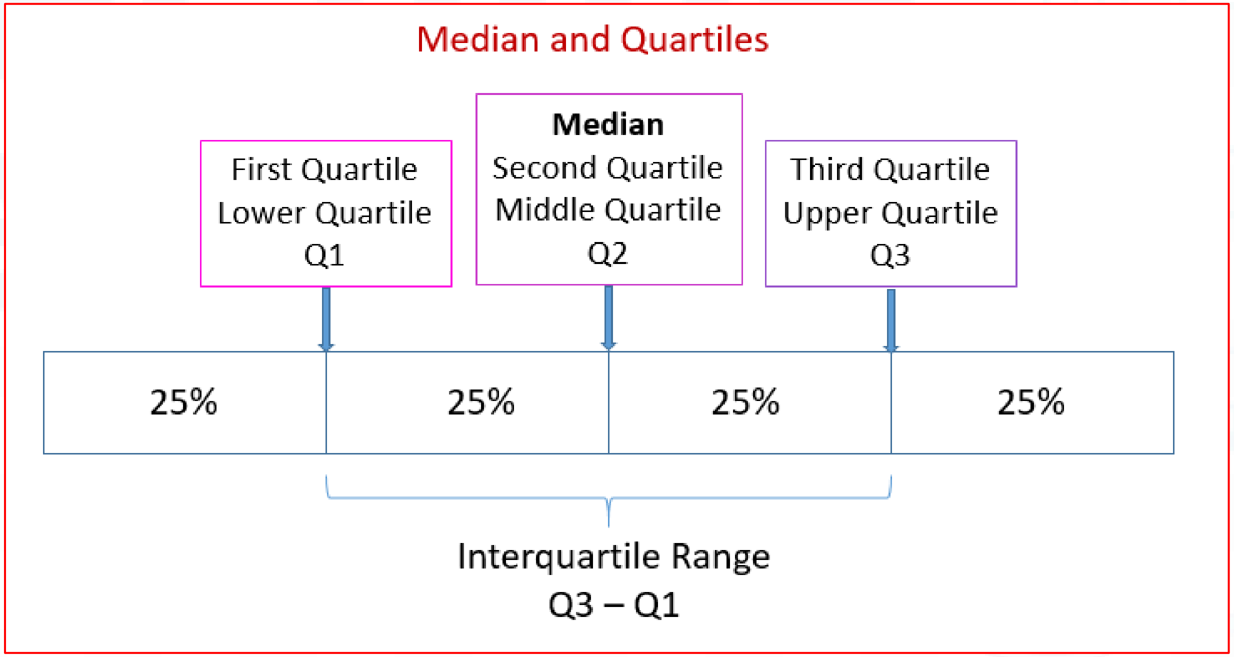
25%	25%	25%	25%

- Variance = Standard deviation²

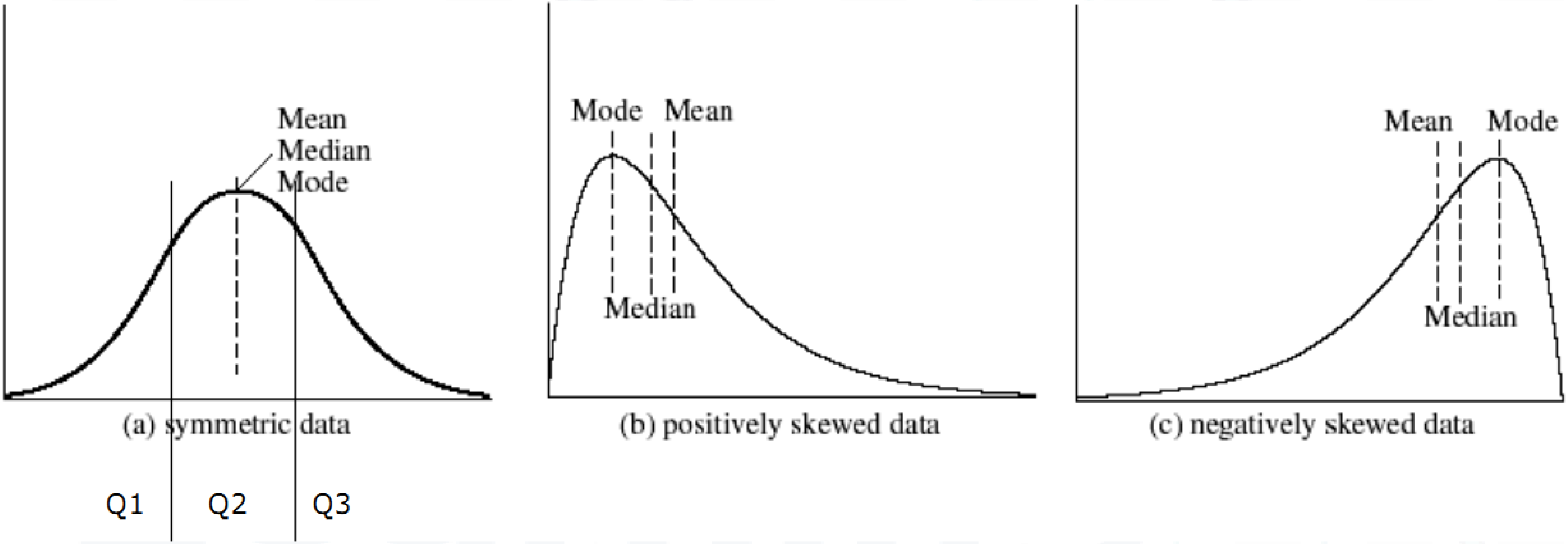
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Descriptive statistics (cont.)

- Outliers (the most extreme observations): values lying above Q3 or below Q1 about 1.5 x IQR (Interquartile Range)



Descriptive statistics (cont.)



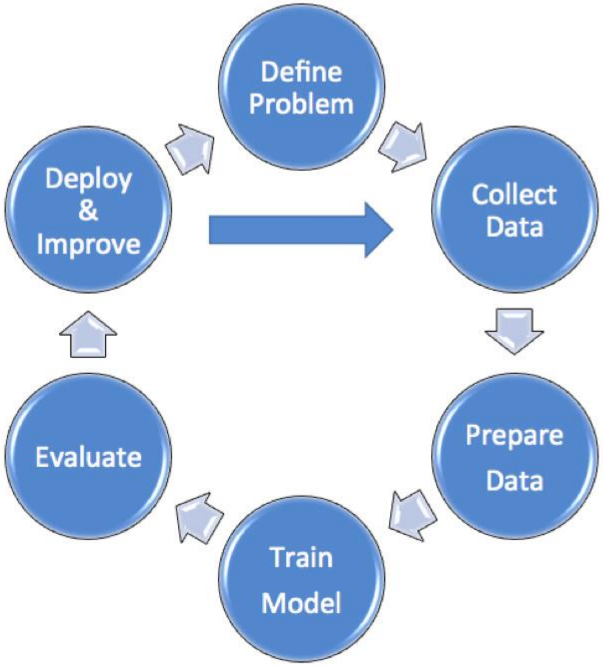
(a) symmetric data

(b) positively skewed data

(c) negatively skewed data

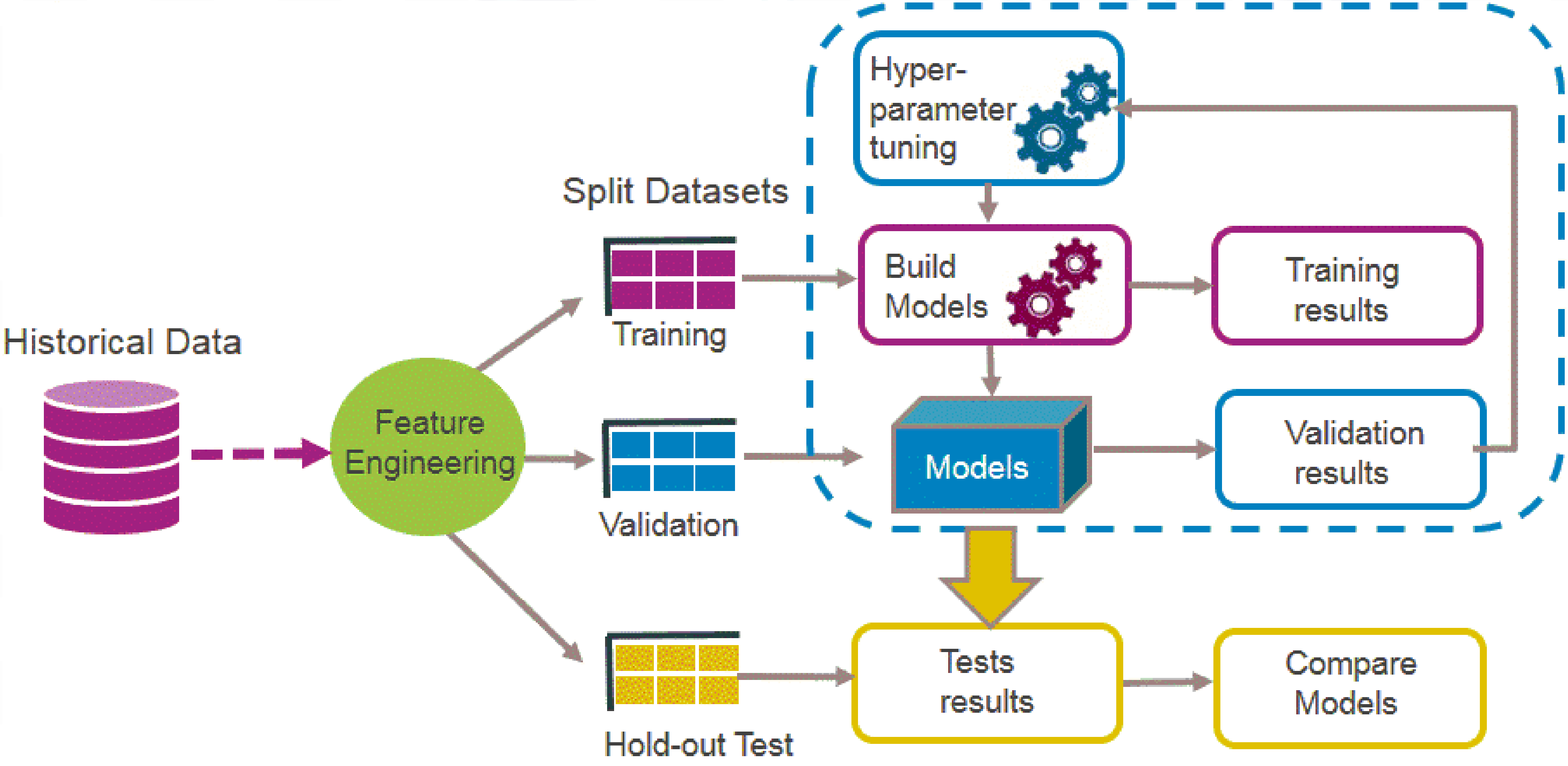
- **Important measures**:
  - median, Q1, Q3, Maximum, Minimum
  - Minimum → Q1 → Median → Q3 → Maximum

Machine Learning process

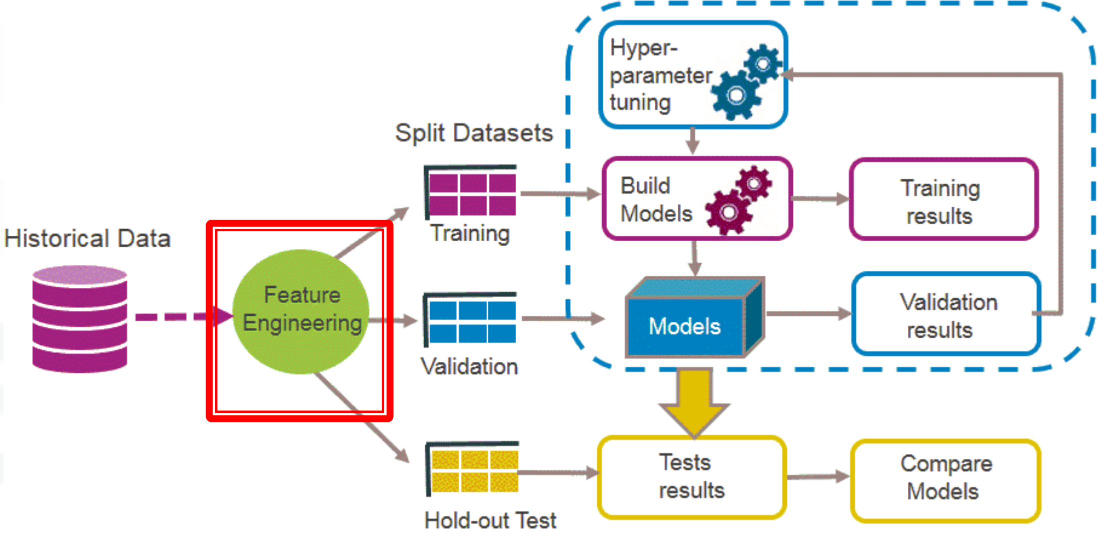


Machine Learning process

- A typical Machine Learning process:

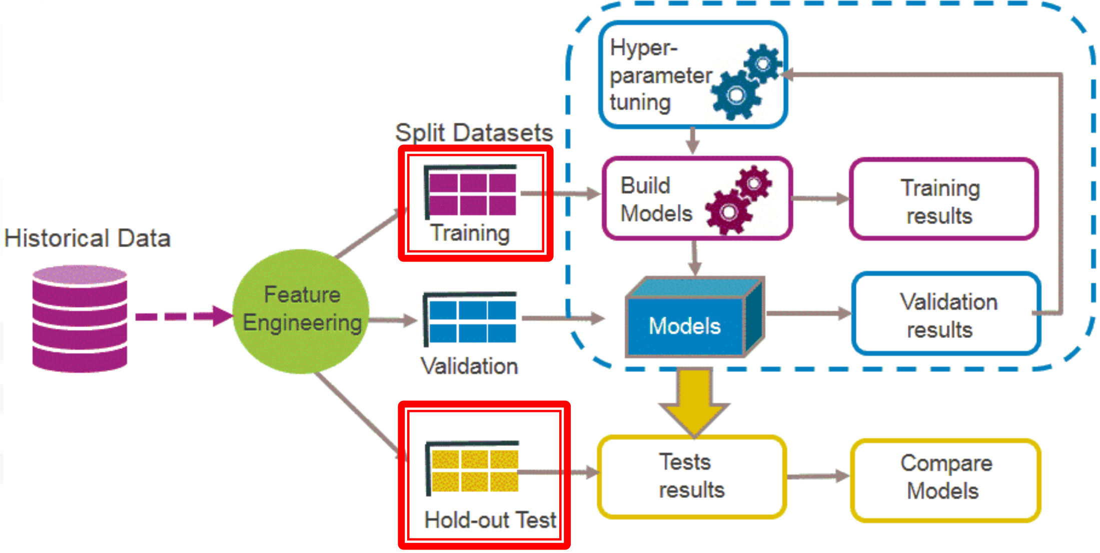


Machine Learning process (cont.)



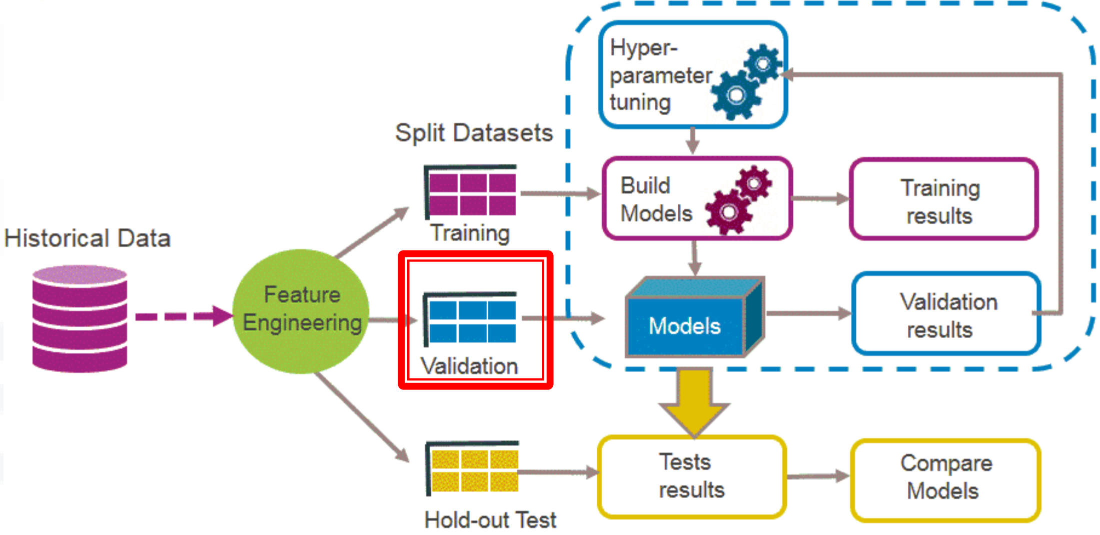
- Feature engineering:
  - the process of selecting, manipulating, and transforming **raw data** into features that can be used for building models.

Machine Learning process (cont.)



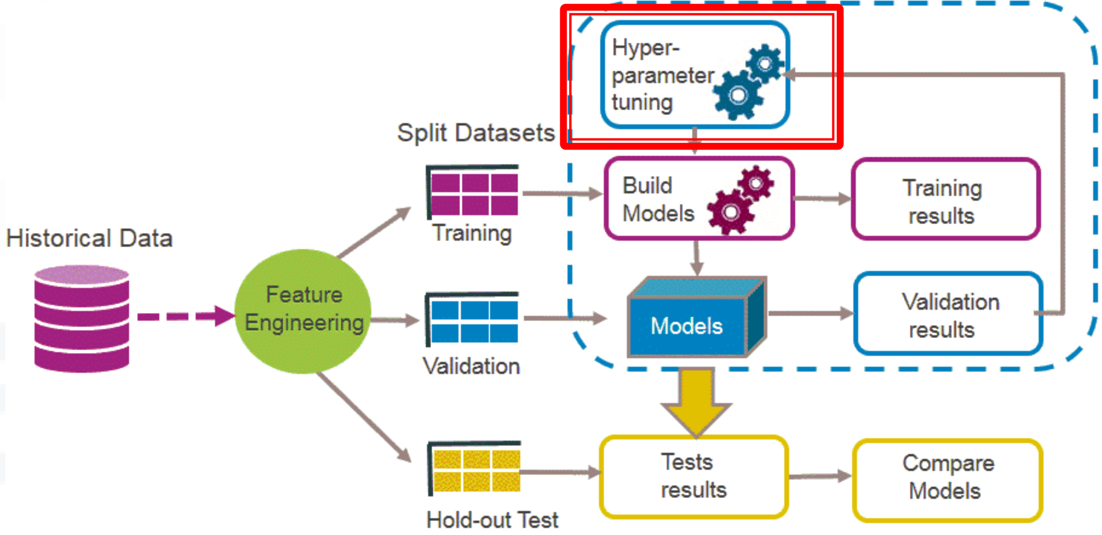
- **Training set**:
  - The sample of data used to fit the model.
- **Testing set**:
  - The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Machine Learning process (cont.)



- Validation set:
  - The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while **tuning model hyperparameters**.

Machine Learning process (cont.)

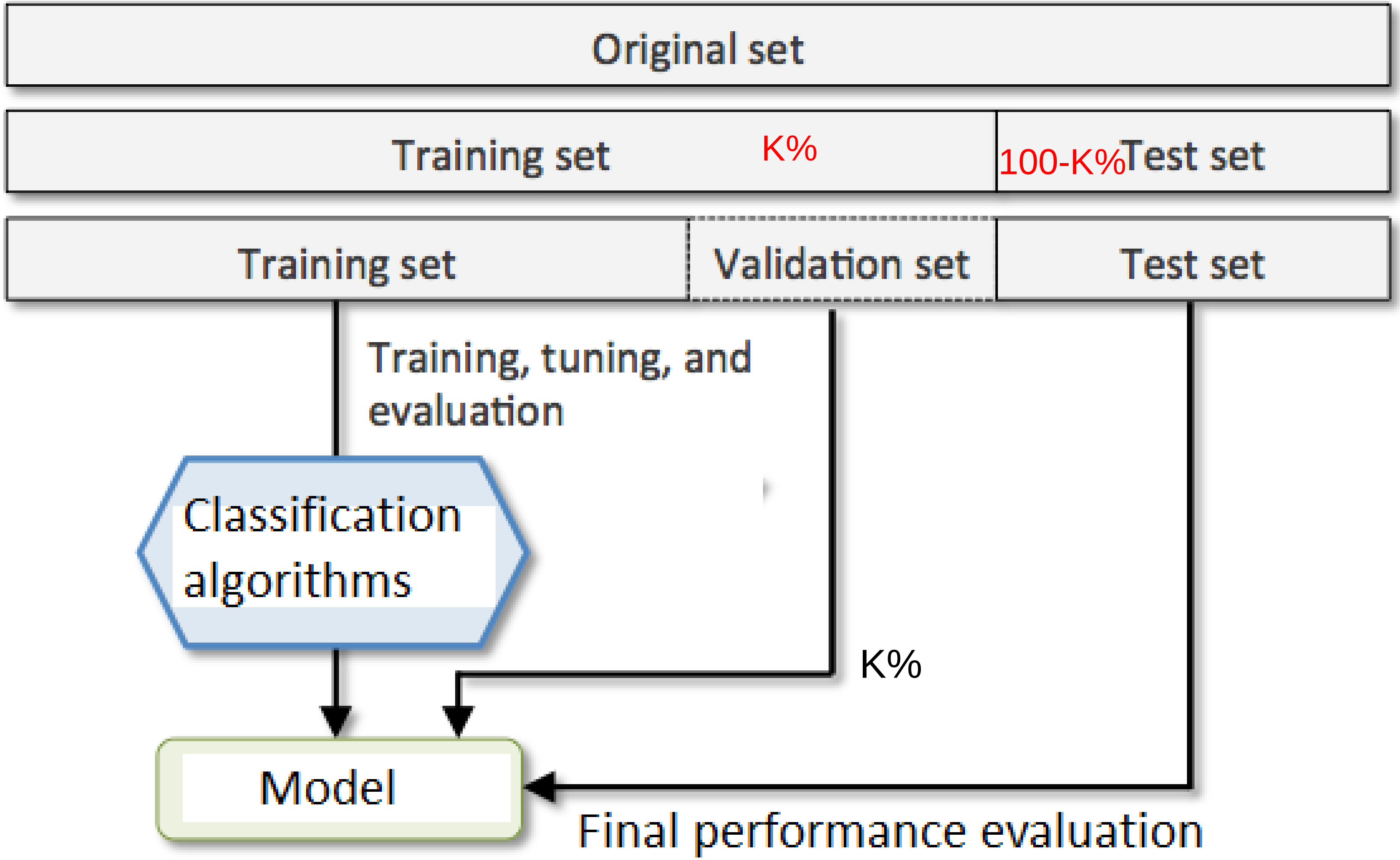


- Hyperparameters:
  - Parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning.
  - The prefix 'hyper_' suggests that they are 'top-level' parameters that control the learning process and the model parameters that result from it.

About the dataset split ratio

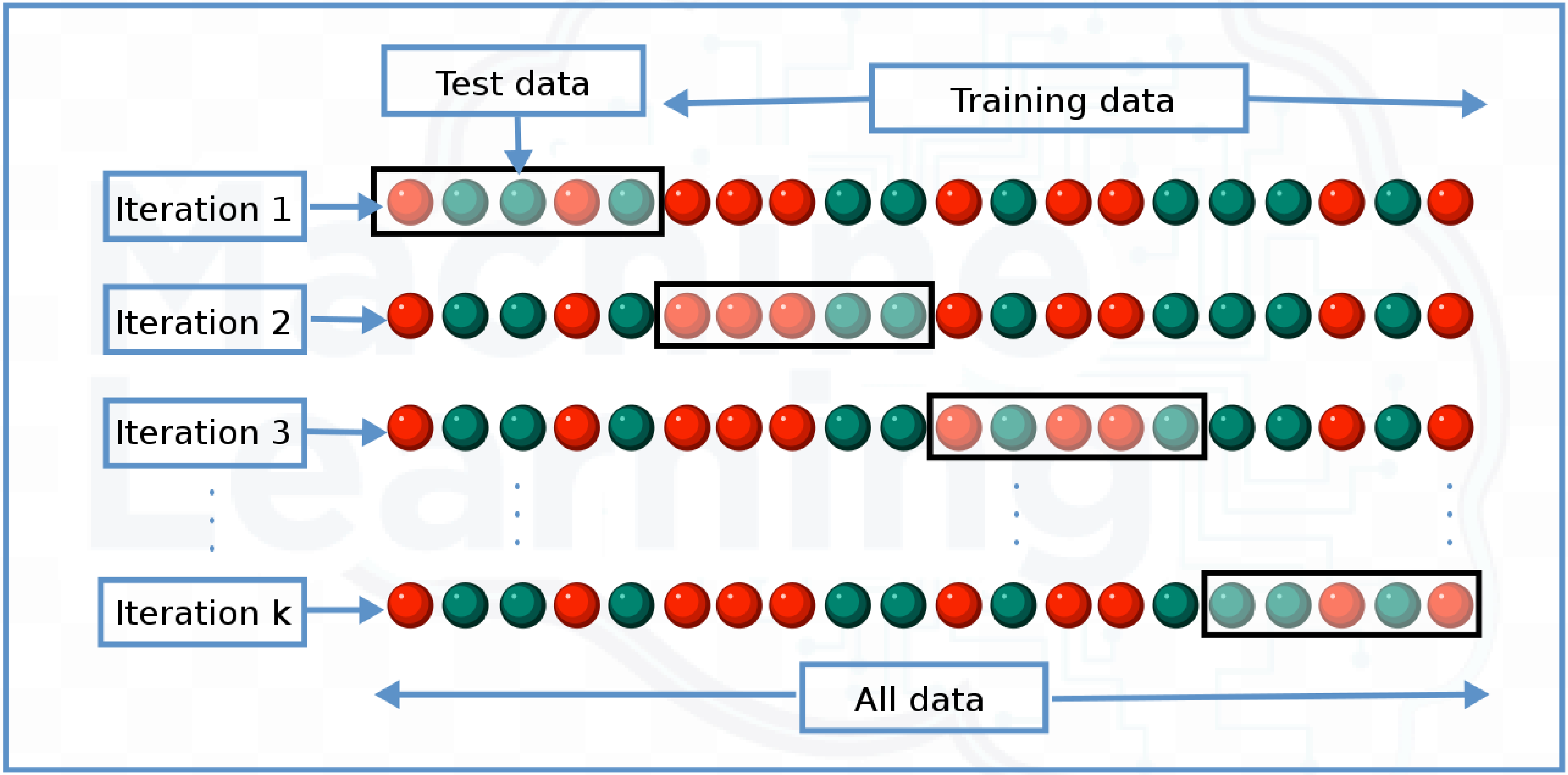
- Split ratio depends on two things:
  - First, the total number of samples in your data
  - Second, the actual model we are training.
- **Example**:
  - Models with very few hyperparameters will be easy to validate and tune ➔ a small validation set
  - Models with many hyperparameters➔ a large validation set

About the dataset split ratio (cont.)


Original set
Training set
K%
100-K%
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Classification algorithms
K%
Model
Final performance evaluation

bigdatauni.com

About the dataset split ratio (cont.)



https://en.wikipedia.org/wiki/Cross-validation_(statistics)

Hyperparameters

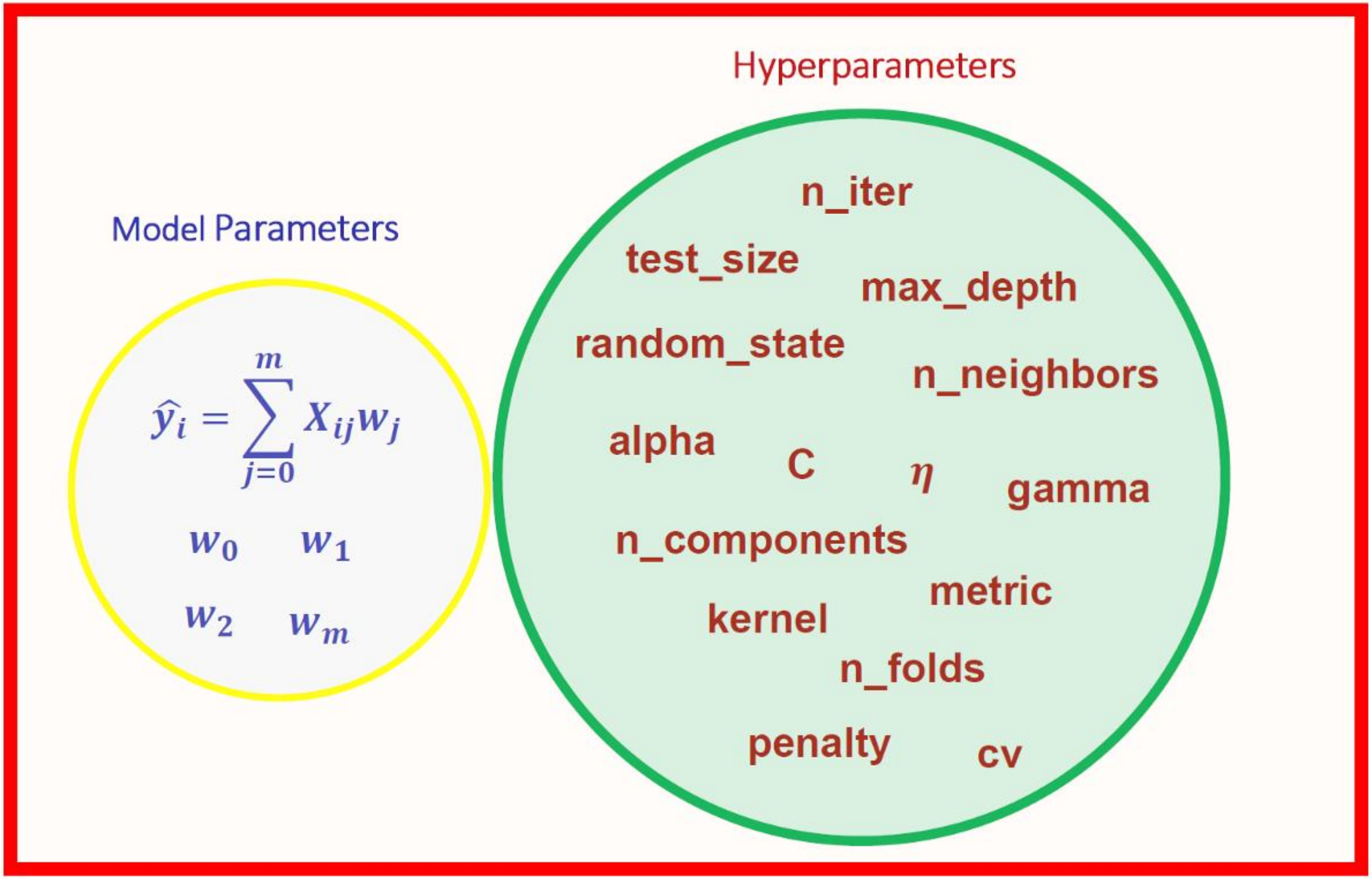
- Examples of hyperparameters:
  - Train-test split ratio
  - Learning rate in optimization algorithms (e.g. gradient descent)
  - Number of hidden layers in a neural network
  - Number of iterations (epochs) in training a neural network
  - Number of clusters in a clustering task
  - Kernel or filter size in convolutional layers
  - Pooling size
  - …

Parameters

- **Parameters**:
  - Are internal to the model
  - Are learned or estimated purely from the data during training

- **Examples of parameters**
  - The coefficients (or weights) of linear and logistic regression models.
  - Weights and biases of a neural network
  - The cluster centroids in clustering

Parameters vs Hyperparameters



Model Parameters

$$\hat{y}_i = \sum_{j=0}^{m} X_{ij} w_j$$

w_0 w_1

w_2 w_m

Hyperparameters

n_iter

test_size

max_depth

random_state

n_neighbors

alpha

C

η

gamma

n_components

metric

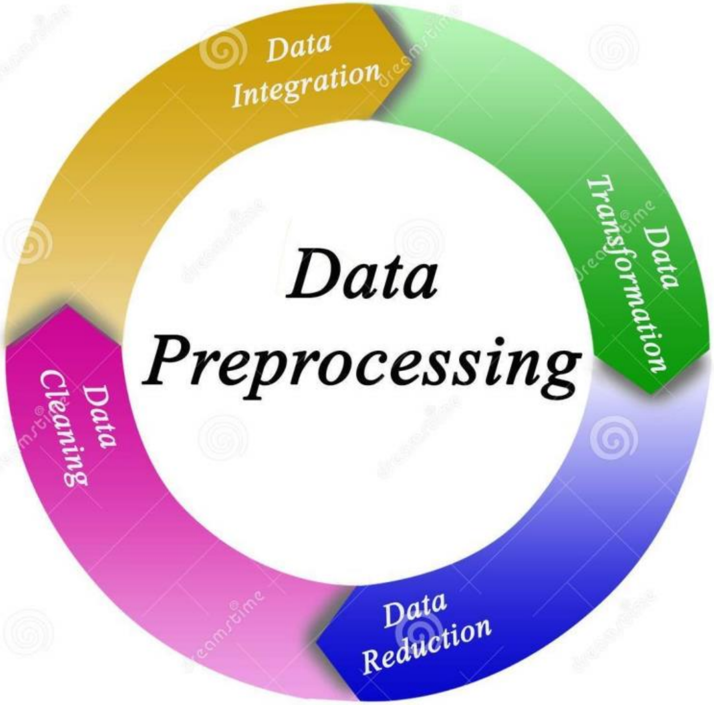
kernel

n_folds

penalty

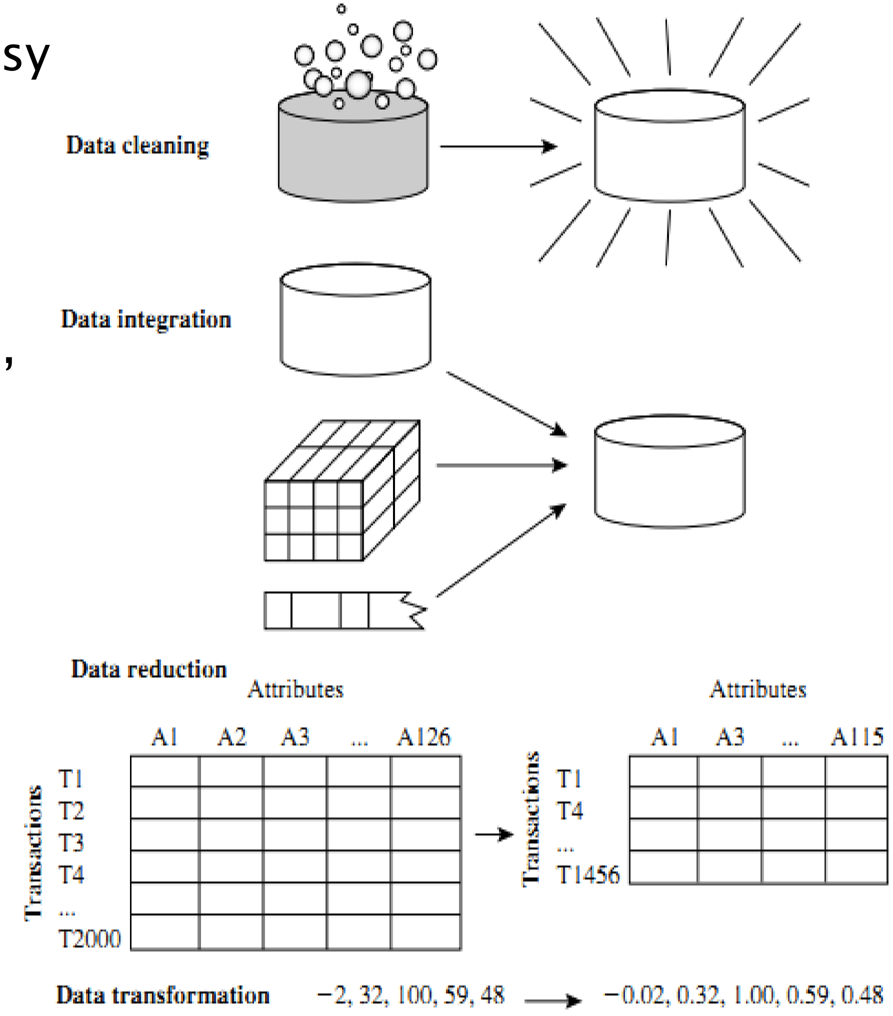
cv

Preprocessing data



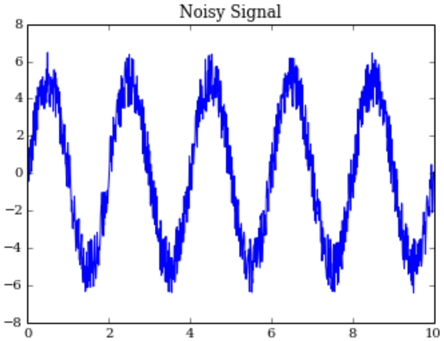
Data preprocessing tasks

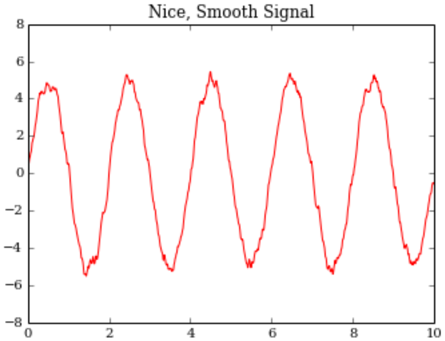
- Data cleaning:
  - Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration:
  - Integration of multiple databases, data cubes, or files
- Data reduction:
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- Data transformation:
  - Normalization
  - Concept hierarchy generation

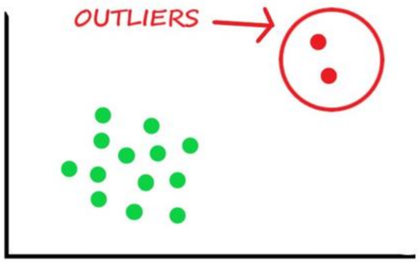


Task 1: Data Cleaning

- Data cleaning:
  - fill in missing values,
  - smooth out noise,
  - identifying outliers,
  - correct inconsistencies.







Task 2: Data integration

- Data integration
  - Combining data from multiple sources into a coherent store
- Schema integration:
  - e.g., A.cust-id ≡ B.cust-number

➔Metadata can be used to help avoid errors in schema integration

  - Metadata: the name, meaning, data type, and range of values permitted for the attribute, and etc.
- Entity identification:
  - Identify real world entities from multiple data sources,
  - e.g., “R & D” in Source 1 and “Research & Development” in source 2. “Male” in Source 1 and “Female” S1, “Nam” and “Nữ” in S2.

Task 3: Data Reduction

- Obtain a reduced representation of the data set
  - much smaller in volume but yet produces almost the same analytical results

- Why data reduction?
  - A database/data warehouse may store terabytes of data
  - Complex analysis may take a very long time to run on the complete data set

Data Reduction (cont.)

- Dimensionality reduction is the process of reducing the number of random variables or attributes under consideration

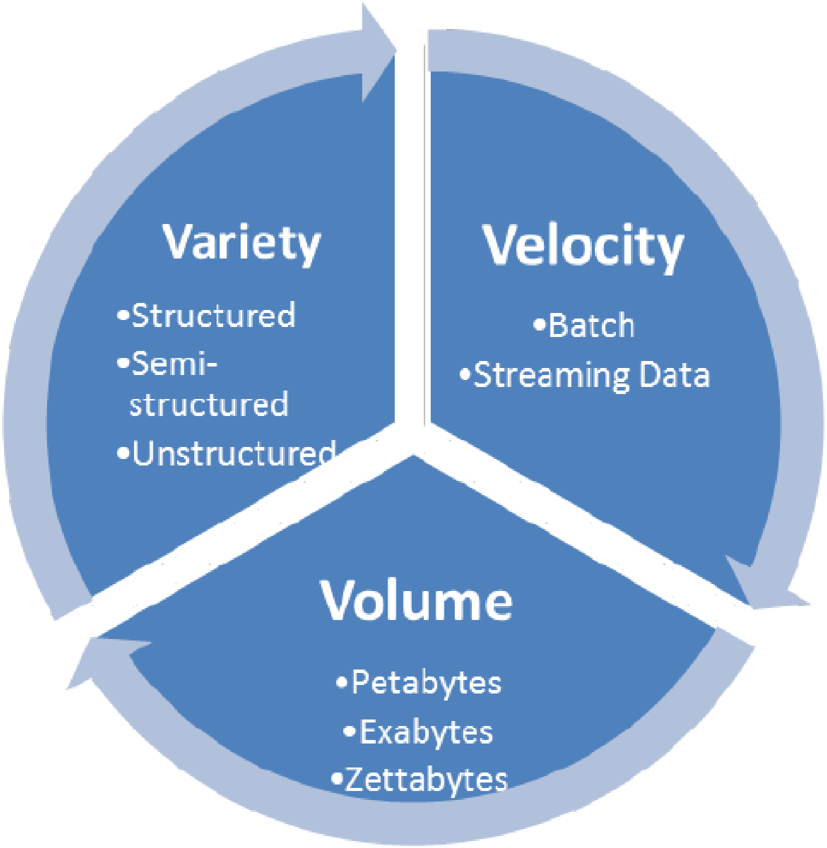
- Numerosity reduction techniques replace the original data volume by alternative, smaller forms of data representation

- Data compression transformations are applied so as to obtain a reduced or “compressed” representation of the original data

Task 4: Data Transformation

- Data are transformed or consolidated into forms appropriate for mining
- Methods:
  - Smoothing: Remove noise from data
  - Attribute/feature construction: New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
  - Discretization: the raw values of a numeric attribute are replaced by interval labels or conceptual labels
  - Concept hierarchy generation for nominal data, where attributes such as street can be generalized to higher-level concepts, like city or country.

Vs of Big data



Variety

- Structured
- Semi-structured
- Unstructured

Velocity

- Batch
- Streaming Data

Volume

- Petabytes
- Exabytes
- Zettabytes

5V'S OF
BIG DATA
1
VOLUME
2
VELOCITY
3
VARIETY
4
VERACITY
5
VALUE

FACULTY OF INFORMATION TECHNOLOGY