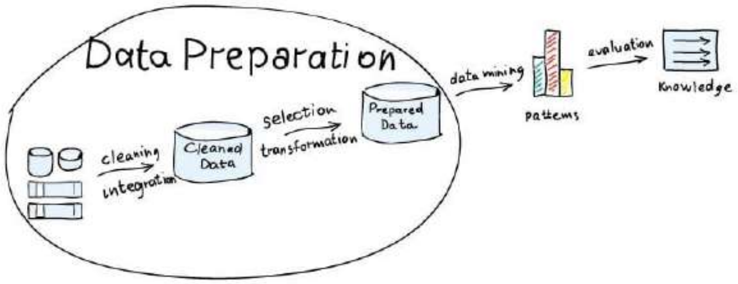
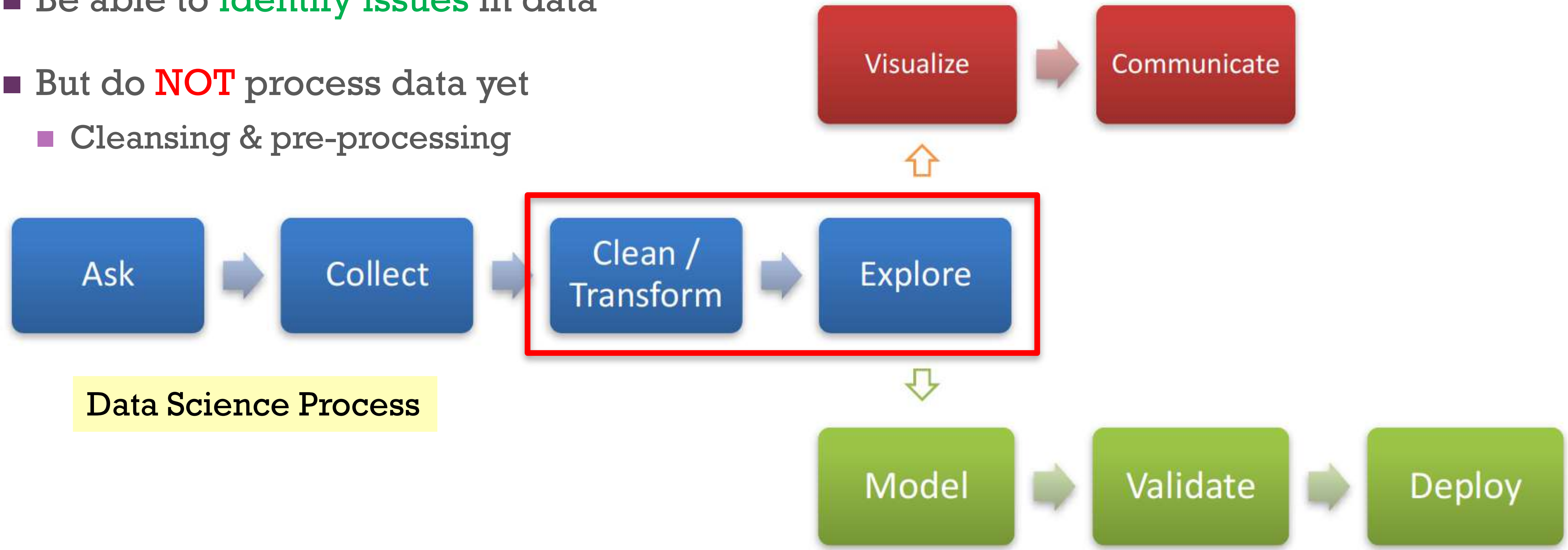


Data Preparation

Peerapon Vateekul, Ph.D.
peerapon.v@4amconsult.com

Data Science Process

- Be able to explore data
- Be able to identify issues in data
- But do NOT process data yet
  - Cleansing & pre-processing



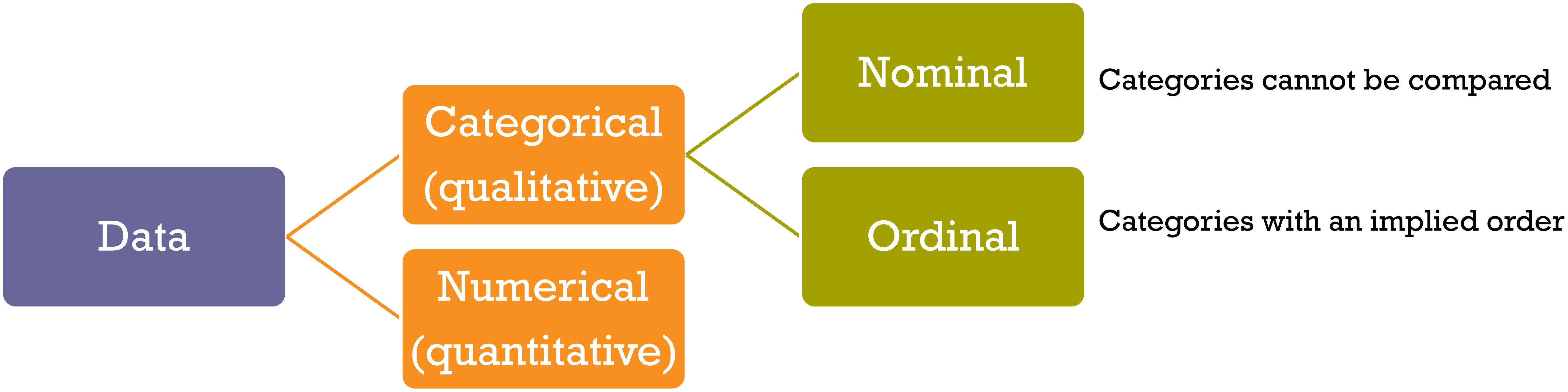
Data Science Process

Terminology: Data table

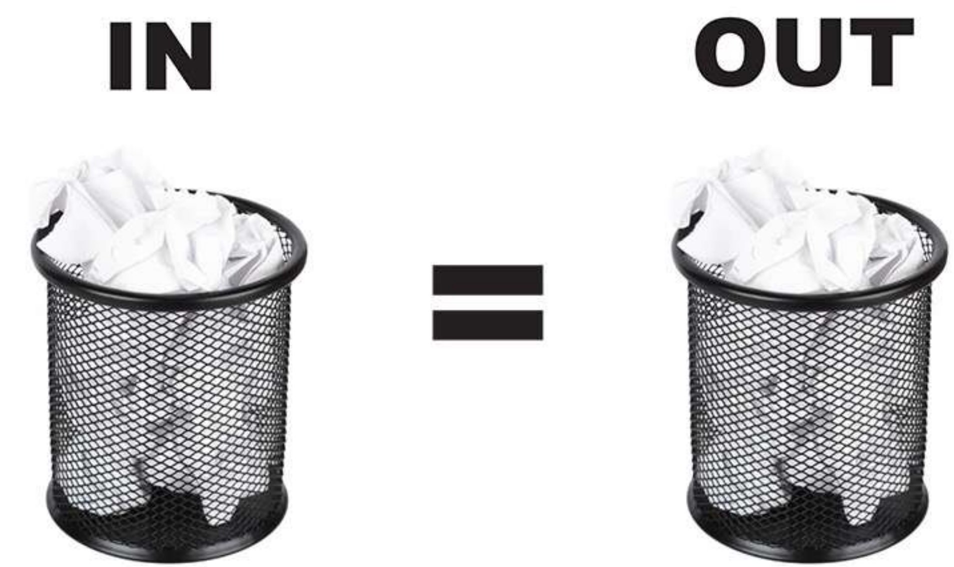
	inputs			target
Age	**Income**	**Gender**	**Province**	**Purchase**
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes
32	35,000	Male	Bangkok	No

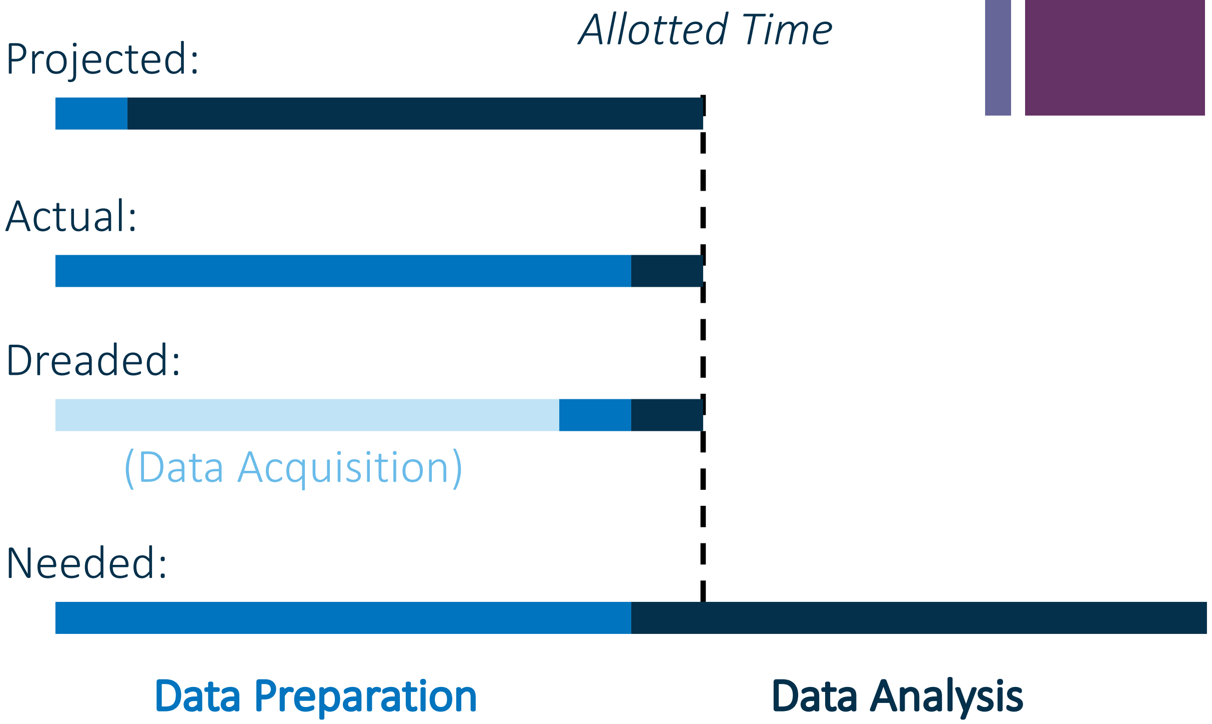
- Row
  - Example, instance, case, observation, subject
- Column
  - Feature, variable, attribute
- Input
  - Predictor, independent, explanatory variable
- Target
  - Output, outcome, response, dependent variable

Terminology: Kinds of data

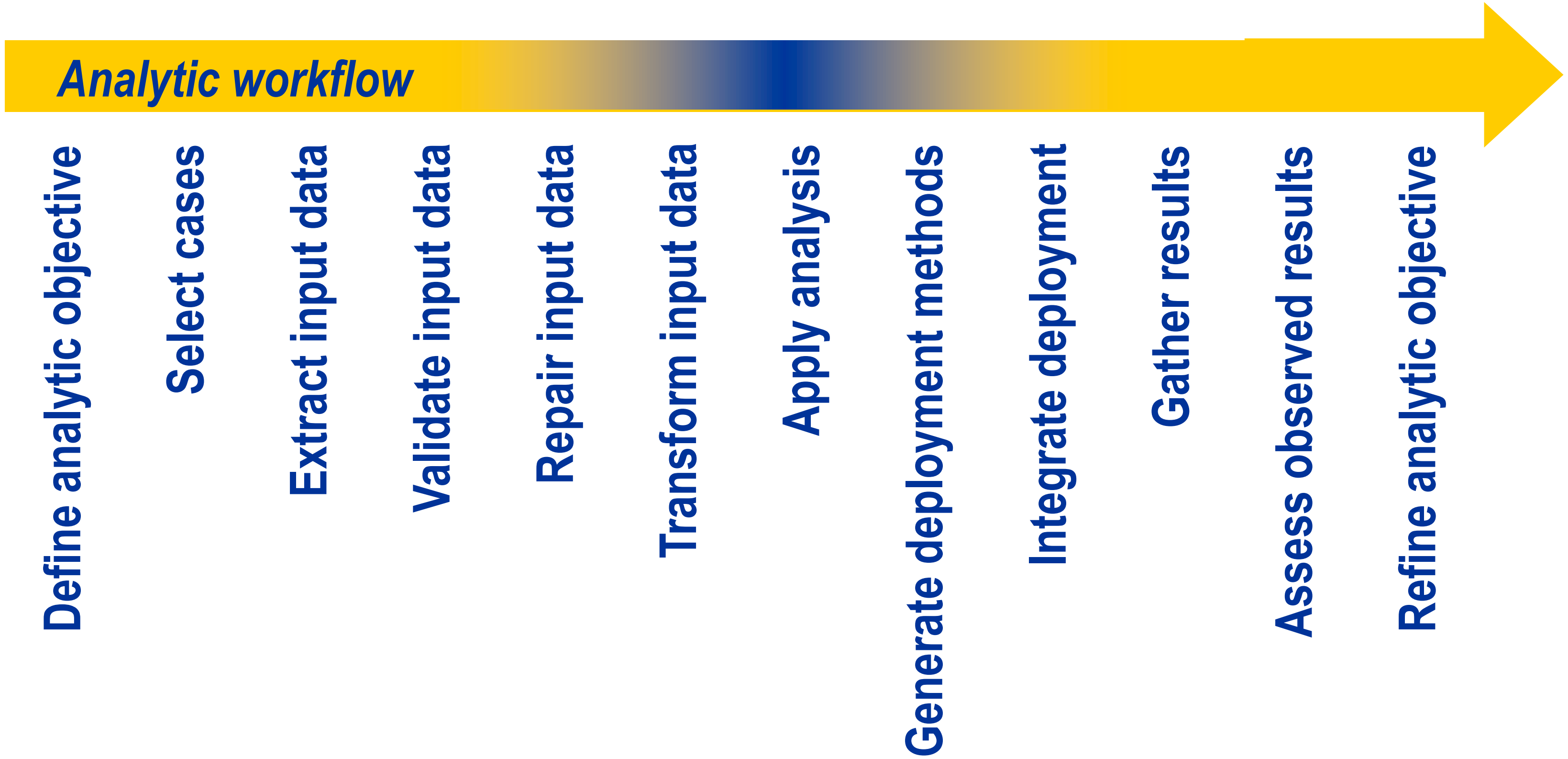


Data preparation is very important!

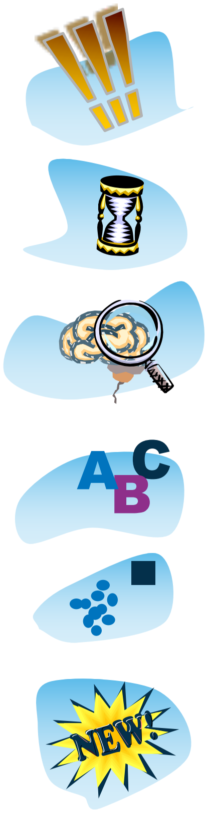




Analytics workflow



Data preparation challenges



- Massive data sets
- Temporal infidelity
- Transaction and event data
- Non-numeric data
- Exceptional, extreme, and missing values
- Stationarity

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

$$Spend = 500 + 2 \times Age + 3 \times Province$$


Practice is everything.
Periander
quotefancy

Data Preparation

1. Business Understanding
2. Examining Data Set
3. Narrowing Down Features
4. Data Preparation
5. Additional Feature Extraction
6. Feature Selection (Optional)
7. Normalization (Optional)
8. Adjusting Imbalanced Data
9. Splitting Data

+

1) Business Understanding

- Clearly defined business problem
- Understand what we want to know
  - Break business problems to data science problems
  - Identify machine learning problem categories
- Set success criteria

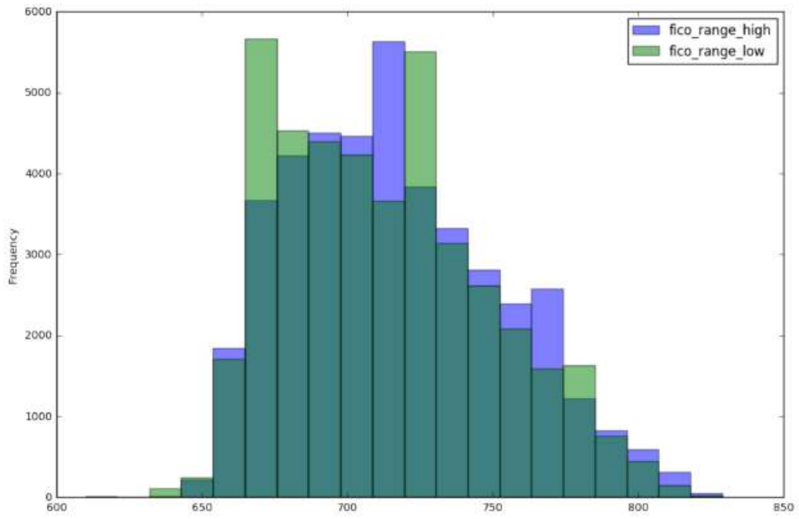
+

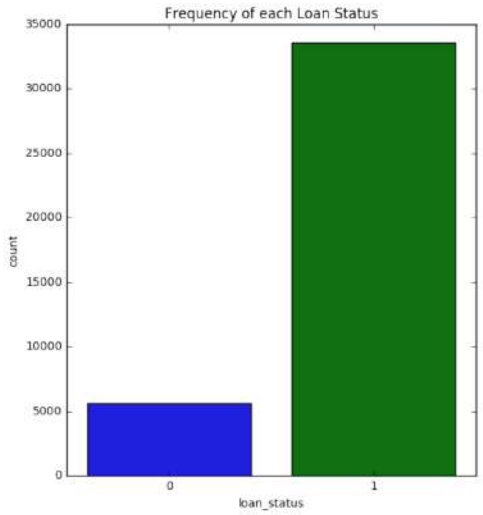
2) Examining Data Set

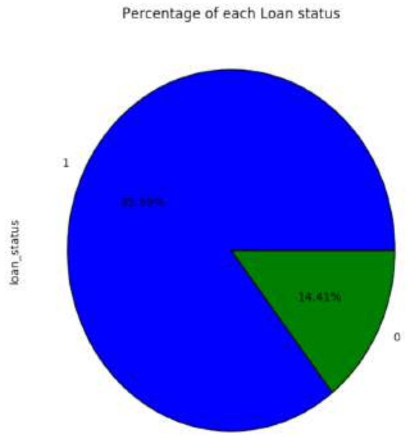
- Head, Tail
- Count row, column
- Data Redundancy : age, birthdate
- Impossible Data : -1 in age column
- Data Type Mismatch : 'john@gmail.com' in phone column
- NULL Value
  - Count Missing value (%)
  - None, NaN, [Unidentify], space -> None

2) Examining Data Set (cont.)

- Numerical features
  - Out of ranges
  - Distribution:
    - histogram
- Categorical features
  - Miscodes
  - Unique count
  - Distribution:
    - Frequency table
    - Bar chart
- Target feature
  - Understand proportion of each class/value







3) Narrowing Down Features

- Narrowing Down Features
  - Understanding each Features
  - Removing Irrelevant Features
  - Removing Temporal Infidelity Features
  - Removing Unqualified Features

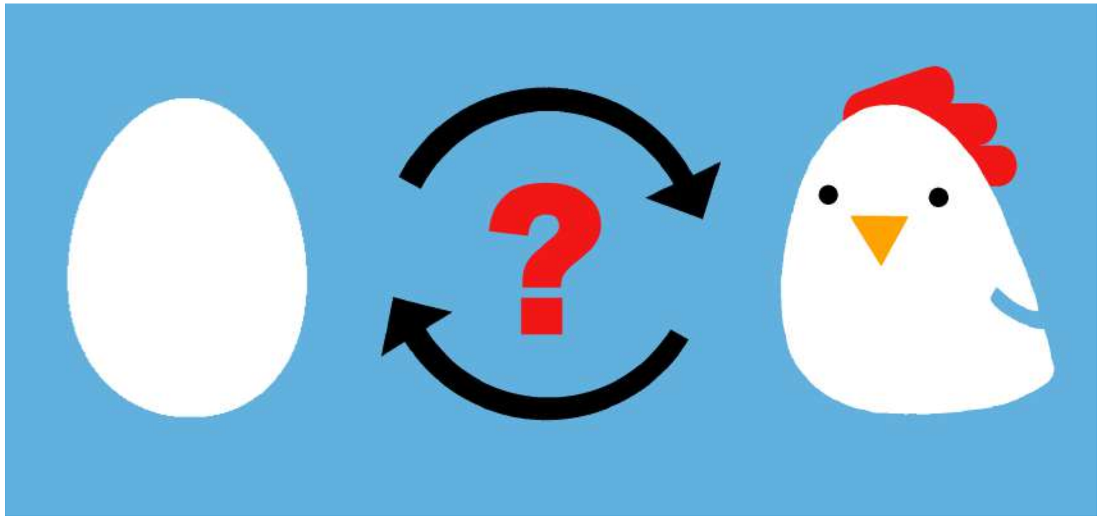
3) Narrowing Down Features (cont.) - Removing Irrelevant Features

Remove irrelevant features <u>manually</u>

Domain expert

Inputs

Target



3) Narrowing Down Features (cont.) - Removing Temporal Infidelity Features

- Occurs when the input variables contain information that will be unavailable at the time that the prediction model is deployed.
- Assume that the model will be deployed in July-2017
  - Should we include a variable called "FICO2017", which is calculated at the end of the year?

3) Narrowing Down Features (cont.) - Removing Unqualified Features

- Id's (lack of generalization; overfit)
- Variables with missing values >

  50%

- Categorical variables
  - Too many unique values (treat as Id's)
  - Flat values (underfit)

- Special ways to treat these data
  - Zip code
    - Distance to closet branch
  - Date/time
    - Recency
  - Categorical
    - Recode, consolidation (grouping)

4) Data Preparation

- Data Preparation
  - 4.1) Imputing Missing Values
  - 4.2) Data Type Conversion
    - Numeric to Categorical
    - Categorical to Numeric
  - 4.3) Truncate Outliers
  - 4.4) Feature Transformation

4.1) Data Preparation (cont.) - Imputing Missing Values

? != 0

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

- Numerical variables:
  - Mean
  - Median
- Categorical variables:
  - Mode

4.2) Data Preparation (cont.) - Data Type Conversion

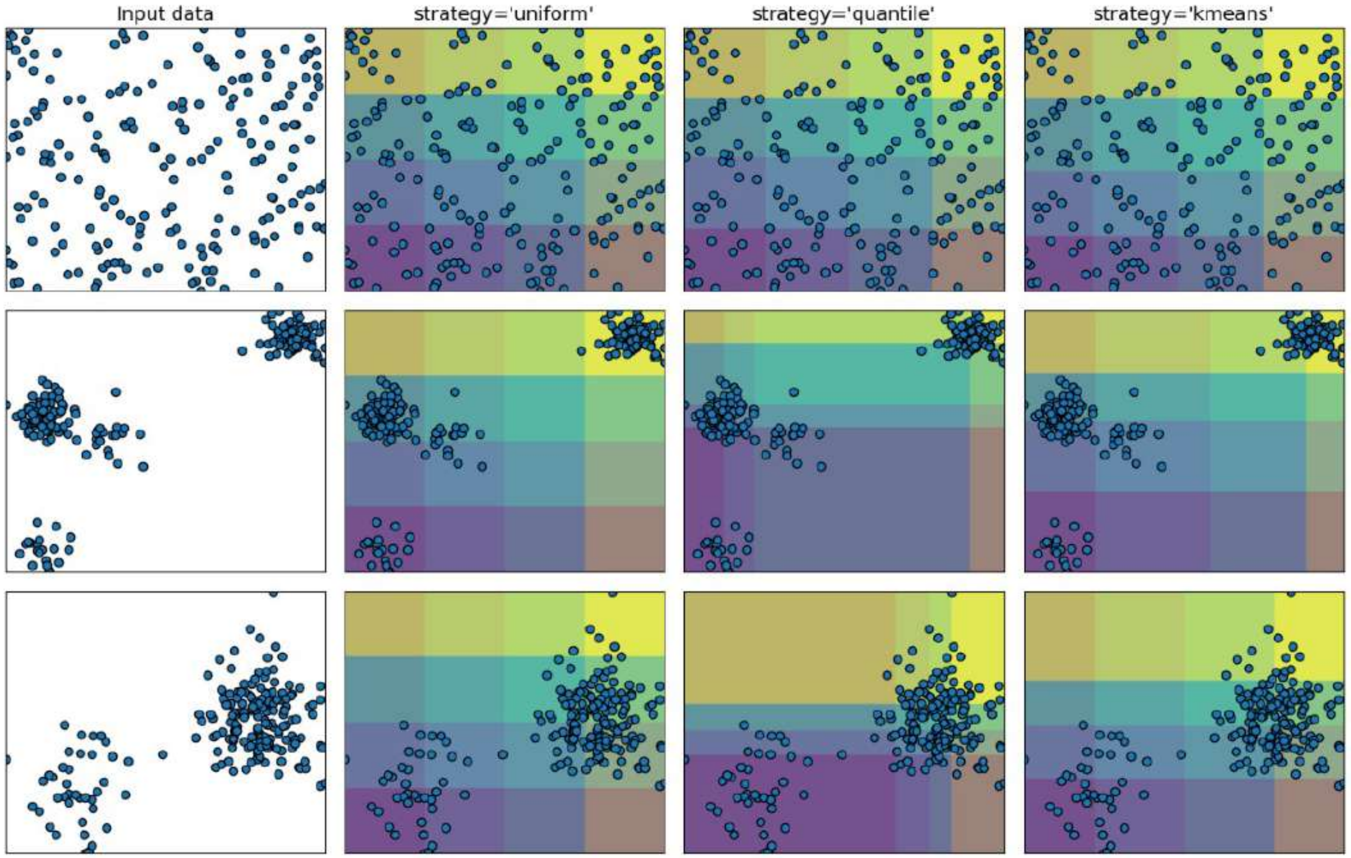
- Some models accept any kind of data
  - While some model accept categorical data
  - While some model accept numerical data
  - Some model accept any kind of data, but affects accuracy and performance
- Need to prepare data for each model

$$\hat{y} = \widehat{w}_0 + \widehat{w}_1 x_1 + \widehat{w}_2 x_2$$

$$Spend = 500 + 2 \times Age + 3 \times Province$$

4) Data Preparation (cont.) - Data Type Conversion (Numeric → Categorical)

- Binning
  - Uniform
    - All bins in each feature have identical widths.
  - Quantile
    - All bins in each feature have the same number of points.
  - K-means (clustering algorithm)
    - Values in each bin have the same nearest center of a 1D k-means cluster.
  - Domain Expert



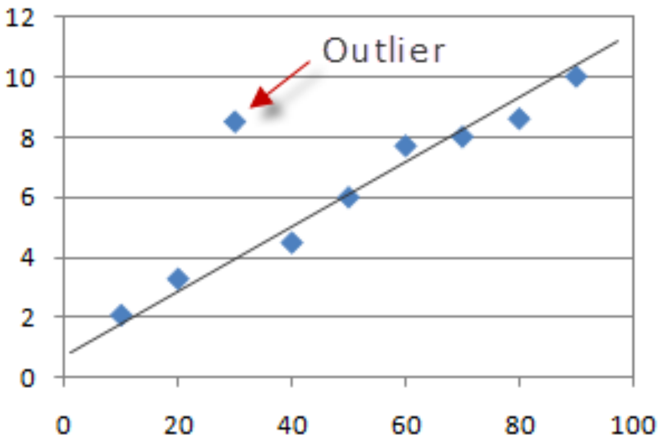
Reference :
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.cut.html
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.KBinsDiscretizer.html

4) Data Preparation (cont.)
- Data Type Conversion (Categorical → Numeric)

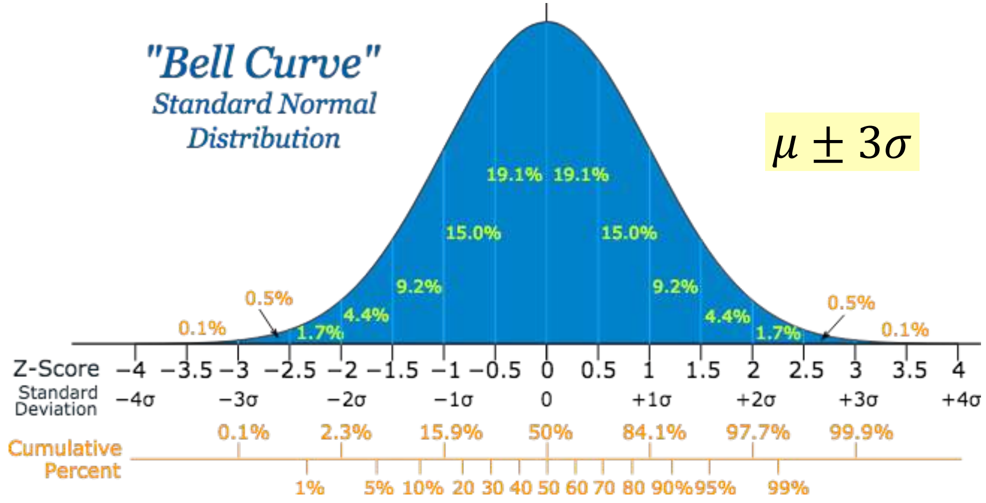
One-hot vector (dummy codes)

Level	D_A	D_B	D_C	D_D	D_E	D_F	D_G	D_H	D_I
A	1	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0
E	0	0	0	0	1	0	0	0	0
F	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	1	0	0
H	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	1

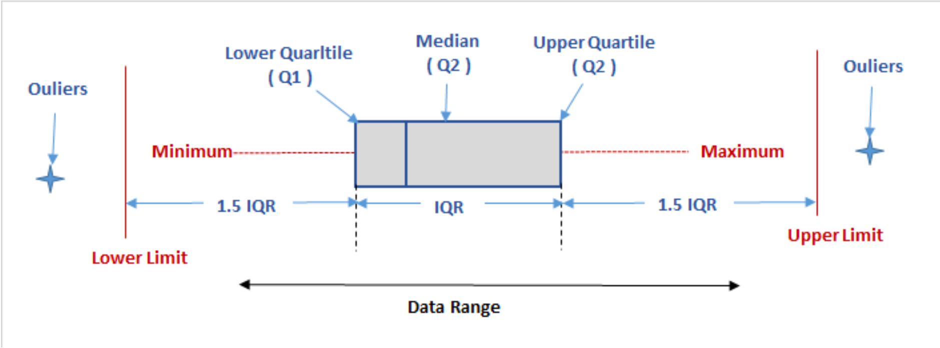
4.3) Data Preparation - Truncate Outliers



- Outlier, leverage points, extreme values



$\mu \pm 3\sigma$



Percentile
1st
2.5th
5th
10th
25th
50th
75th
90th
95th
97.5th
99th

4.3) Data Preparation - Truncate Outliers

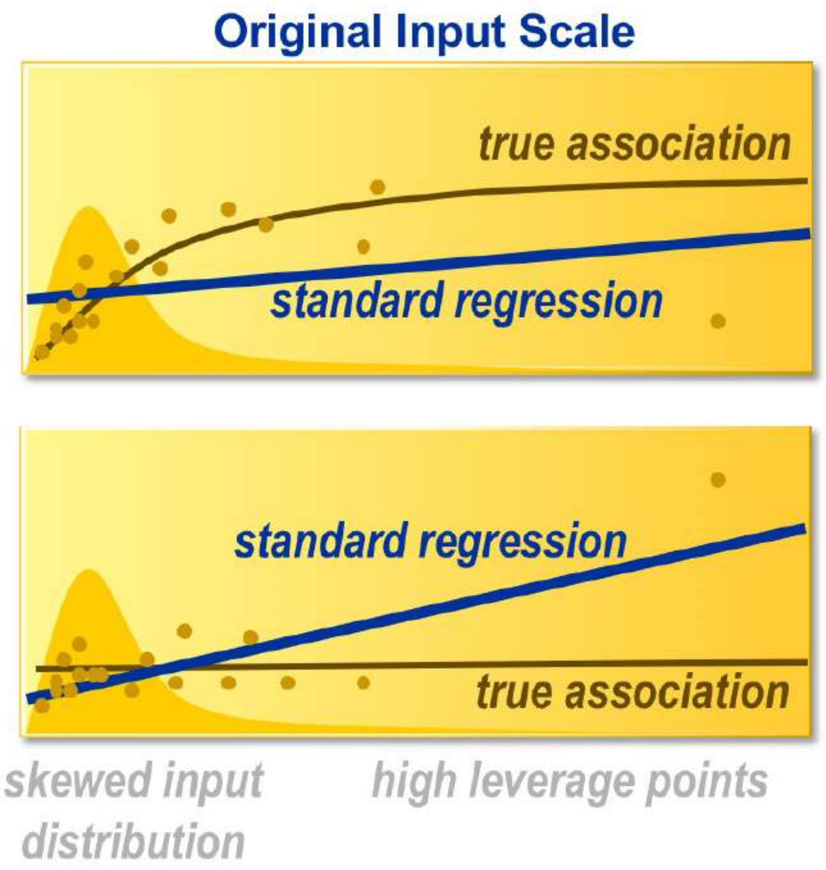
	Spending
	3,000
	3,200
	4,000
	4,500
	5,000
	1,000,000
Mean	169,950
Stddev	406,640.5

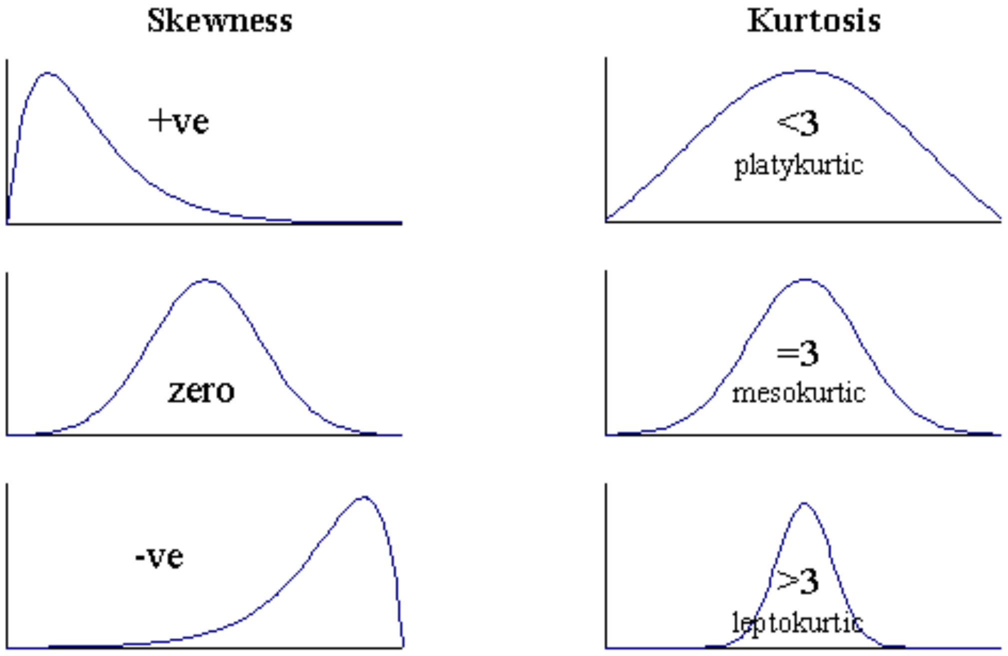
	Spending
	3,000.00
	3,200.00
	4,000.00
	4,500.00
	5,000.00
Mean	3,940.00
Stddev	847.35

4.4) Data Preparation - Feature Transformation

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

- Skewness
- Example: Salary, Balance in bank account
- **Solutions: Log, Binning**





4.4) Data Preparation - Feature Transformation

	Spending	Log(Spending)
	3,000	3.48
	3,200	3.51
	4,000	3.60
	4,500	3.65
	5,000	3.70
	1,000,000	6.00
Mean	169,950	3.99
Stddev	406,640.5	0.99

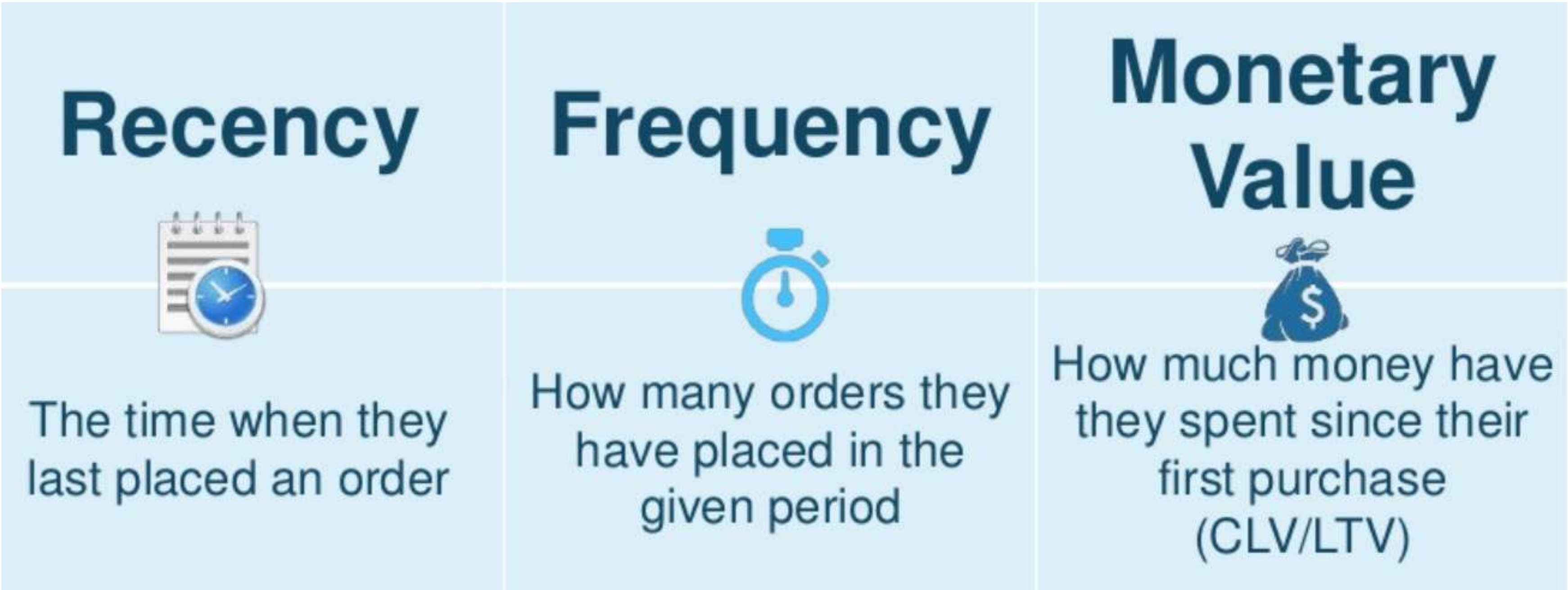
	Spending	Log(Spending)
	3,000.00	3.48
	3,200.00	3.51
	4,000.00	3.60
	4,500.00	3.65
	5,000.00	3.70
Mean	3,940.00	3.59
Stddev	847.35	0.09

5) Additional Feature Extraction

- Additional Feature Extraction
  - Domain Expert
  - RFM
  - Trend
- *** Repeat Data Preparation Steps on Created Features (If Needed)

5) Additional Feature Extraction - RFM

- Additional Feature Extraction
  - Calculated variables
  - Behavior from transactional data (RFM/RFA)



5) Additional Feature Extraction - RFM

- Example : Online Shopping Data

- Monetary
  - Sum spending
- Frequency
  - Count spending/visiting
- Recency
  - How much time has elapsed since a customer's last spending/visiting

Last 12 months

Account	Spending Sum	Frequency Visit	Frequency Spending	Recency Visit	Recency Spending
A	20,000	5	2	7	10
B	100,000	30	27	2	2
C	80,000	7	7	5	5

5) Additional Feature Extraction - Trend

- Trend
  - Comparison between short term behavior and long-term behavior
  - Example :
    - Comparison between "Spending in last 3 months" and "Spending in last 12 months"

Account	Spending Last 3 months	Spending Last 12 months
A	5,000	20,000
B	**1,000** Warning !!!	100,000
C	**60,000**	80,000

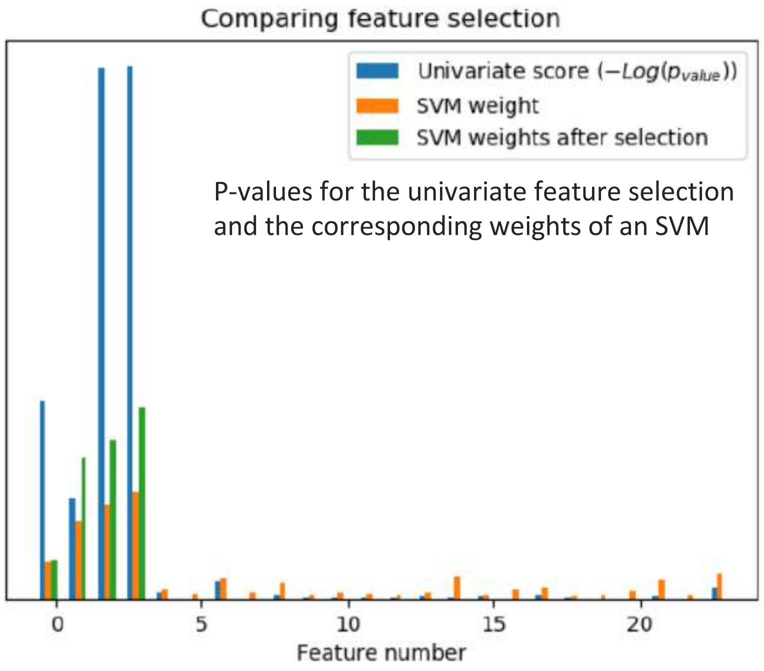
6) Feature Selection (Optional)

- **Feature Selection is one of the core concepts in machine learning**
- **Hugely impacts the performance of your model**
  - **Reduces Overfitting**:
    - Less redundant data means less opportunity to make decisions based on noise.
  - **Improves Accuracy**:
    - Less misleading data means modeling accuracy improves.
  - **Reduces Training Time**:
    - Fewer data points reduce algorithm complexity and algorithms train faster.

Reference : https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

6) Feature Selection (Optional) - Univariate Selection

- Univariate Selection
  - Statistical tests can be used to select those features that have the strongest relationship with the output variable.
  - Example : chi-squared (chi^2) statistical test
  - **SelectKBest in Sklearn Library**



P-values for the univariate feature selection and the corresponding weights of an SVM

only the 4 first ones are significant

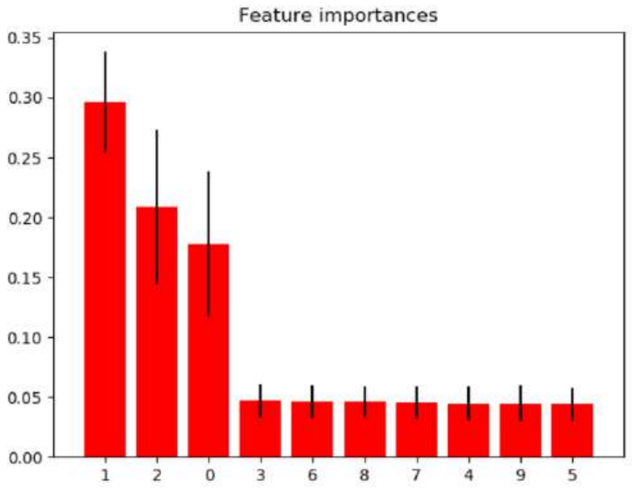
We can see that univariate feature selection selects the informative features and that these have larger SVM weights.

Out:
```
Classification accuracy without selecting features: 0.789
Classification accuracy after univariate feature selection: 0.868
```

Reference : https://scikit-learn.org/stable/auto_examples/feature_selection/plot_feature_selection.html#sphx-glr-auto-examples-feature-selection-plot-feature-selection-py

6) Feature Selection (Optional) - Feature Importance

- Feature Importance
  - Feature importance property of the model.
  - Feature importance is an inbuilt class that comes with Tree Based Classifiers



3 features are informative while the remaining are not.

```
import numpy as np
import matplotlib.pyplot as plt

from sklearn.datasets import make_classification
from sklearn.ensemble import ExtraTreesClassifier

# Build a classification task using 3 informative features
X, y = make_classification(n_samples=1000,
                           n_features=10,
                           n_informative=3,
                           n_redundant=0,
                           n_repeated=0,
                           n_classes=2,
                           random_state=0,
                           shuffle=False)

# Build a forest and compute the impurity-based feature importances
forest = ExtraTreesClassifier(n_estimators=250,
                              random_state=0)

forest.fit(X, y)
importances = forest.feature_importances_
```

Out:

```
Feature ranking:
1. feature 1 (0.295902)
2. feature 2 (0.208351)
3. feature 0 (0.177632)
4. feature 3 (0.047121)
5. feature 6 (0.046303)
6. feature 8 (0.046013)
7. feature 7 (0.045575)
8. feature 4 (0.044614)
9. feature 9 (0.044577)
10. feature 5 (0.043912)
```

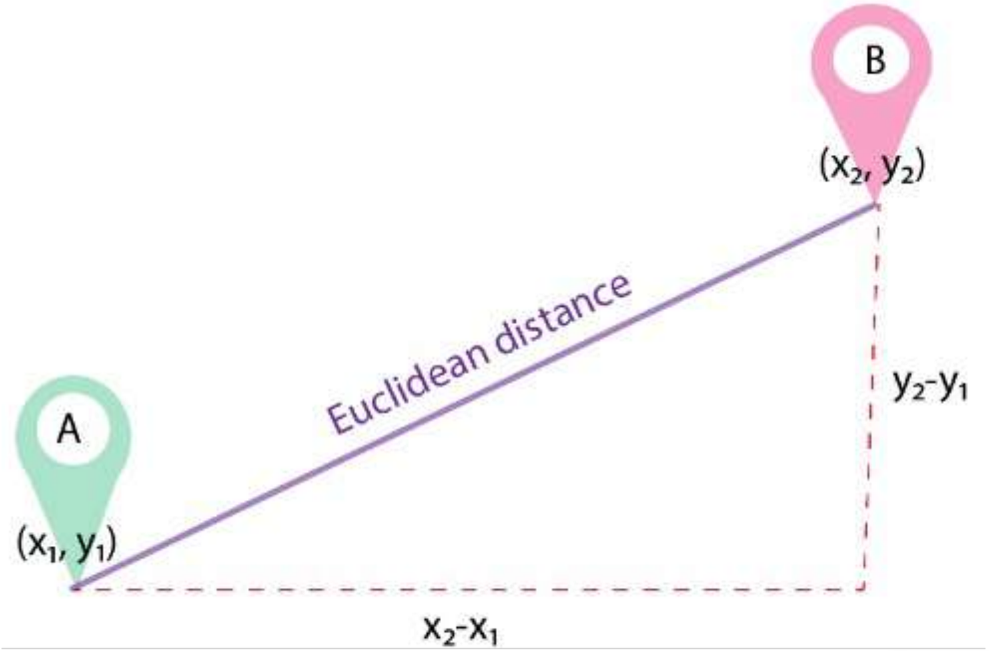
Reference : https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

+ Normalization (Optional)

- Normalization (Optional)
  - MinMax Normalization
  - Z-Score Normalization

Normalization (Optional)

ID	Age (x)	Salary (y)
x1	26	20000.00
x2	30	45000.00
x3	60	70000.00

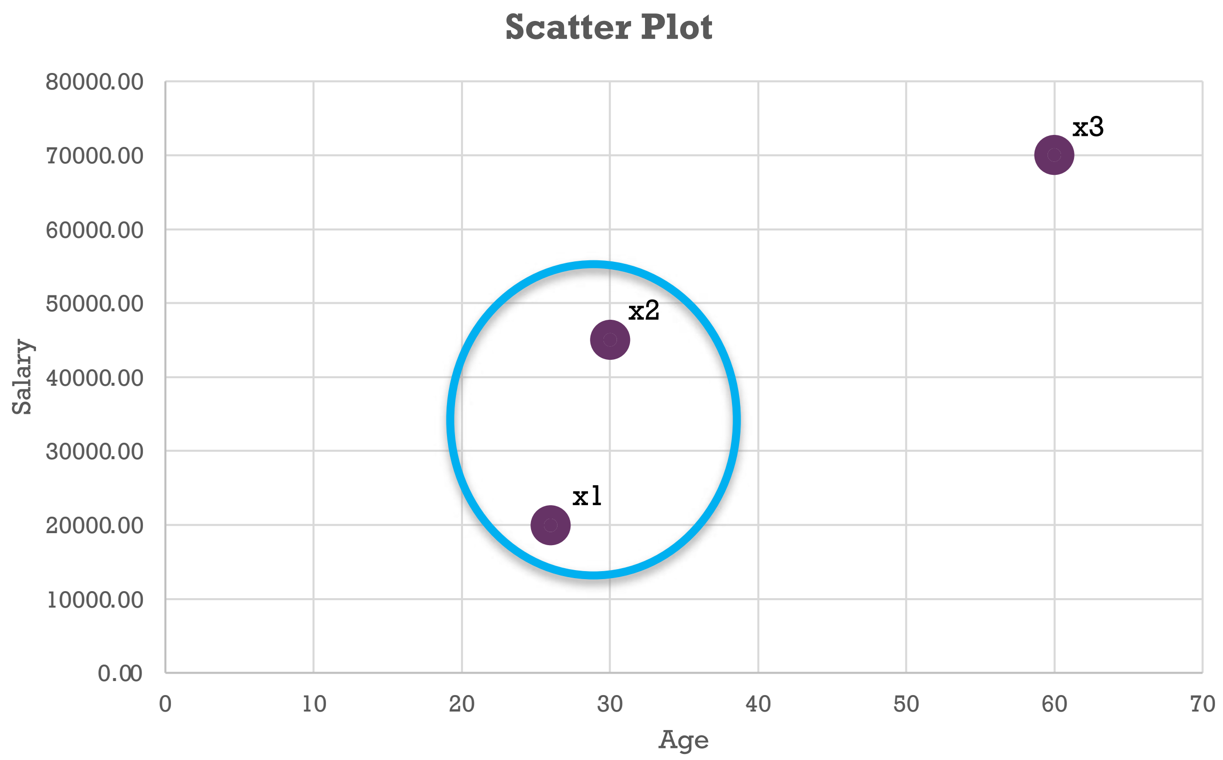


ID1	ID2	Diff_age	Diff_age^2	Diff_salary	Diff_salary^2	Sum	Sqrt
x1	x2	4	16	25000	625,000,000	625,000,016	25,000
x1	x3	34	1156	50000	2,500,000,000	250,0,001,156	50,000.01
x2	x3	30	900	25000	625,000,000	625,000,900	25,000.02

Reference : https://medium.com/@sarkarchandrima12/different-types-of-distances-used-in-machine-learning-7d6a757a2e50

Normalization (Optional)

ID	Age (x)	Salary (y)
x1	26	20000.00
x2	30	45000.00
x3	60	70000.00



Normalization (Optional)

ID	Age (x)	Salary (y)
x1	26	20000.00
x2	30	45000.00
x3	60	70000.00

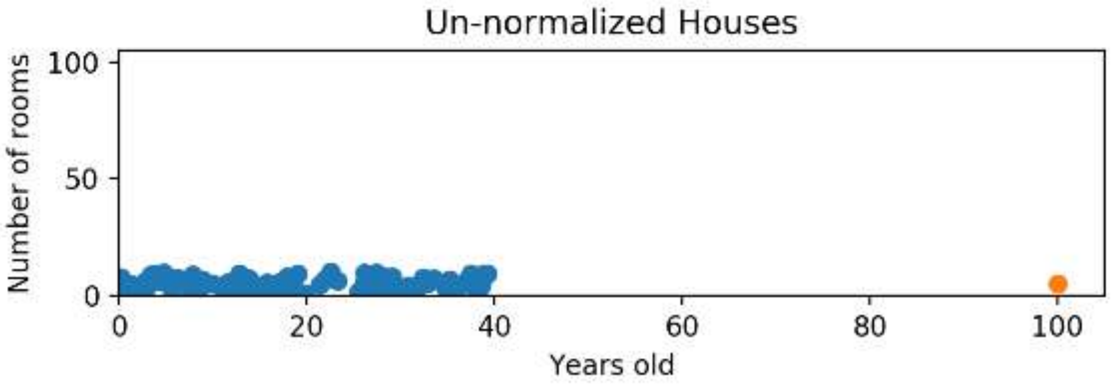
ID	Norm Age (x)	Norm Salary (y)
x1	0	0.29
x2	0.12	0.64
x3	1.00	1.00

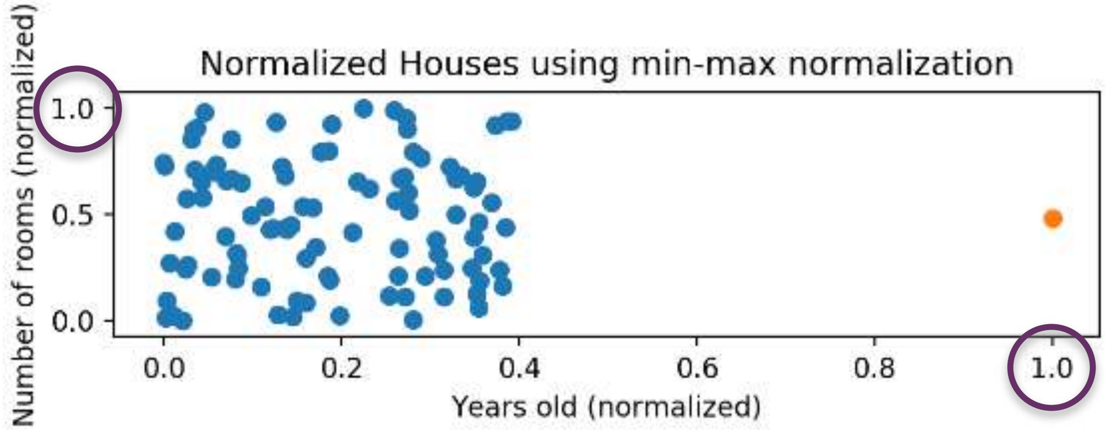
ID1	ID2	Diff_age	Diff_age^2	Diff_salary	Diff_salary^2	Sum	Sqrt
x1	x2	0.12	0.01	0.36	0.13	0.47	0.69
x1	x3	1.00	1.00	0.71	0.51	1.71	1.31
x2	x3	0.88	0.78	0.36	0.13	1.24	1.11

Normalization (Optional) - MinMax Normalization

- Most common ways to normalize data
- For every feature
  - The minimum value transformed into a 0
  - The maximum value transformed into a 1
  - Every other value gets transformed into a decimal between 0 and 1

$$\frac{value - min}{max - min}$$





It does not handle outliers

Reference : https://www.codecademy.com/articles/normalization

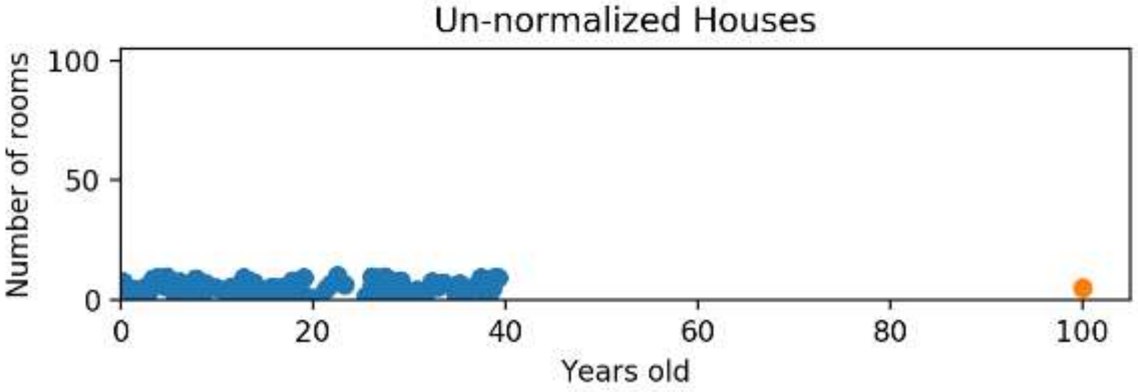
7) Normalization (Optional) - Z-Score Normalization

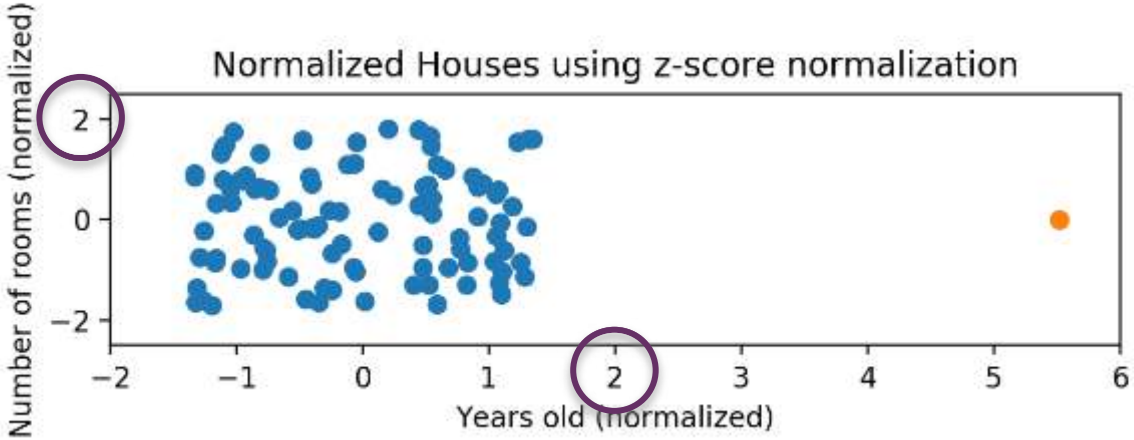
- Strategy of normalizing data that avoids outlier
- For every feature
  - If value is equal to mean, it will be 0
  - If value is below mean, it will be a negative number
  - If value is above mean, it will be a positive number

$$\frac{value - \mu}{\sigma}$$

μ is the mean value of the feature

σ is the standard deviation of the feature





Almost all points are between -2 and 2 on both the x-axis and axis

The only potential downside is that the features aren't on the *exact* same scale

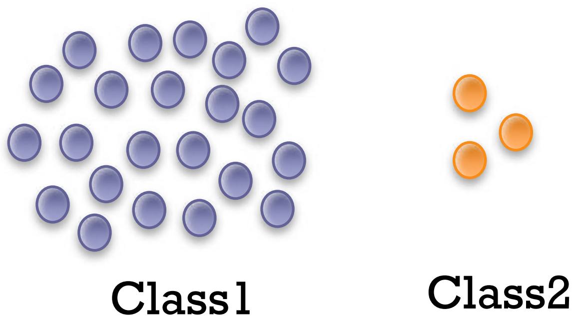
Reference : https://www.codecademy.com/articles/normalization

8) Adjusting Imbalanced Data

Adjusting imbalanced data

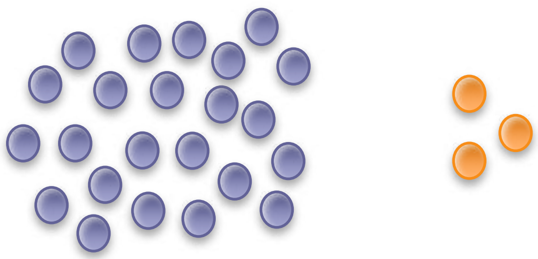
- When number of instance between 2 classes are significantly difference
- Adjust the class distribution of data set

- How?
  - over sampling
  - under sampling



8) Adjusting Imbalanced Data

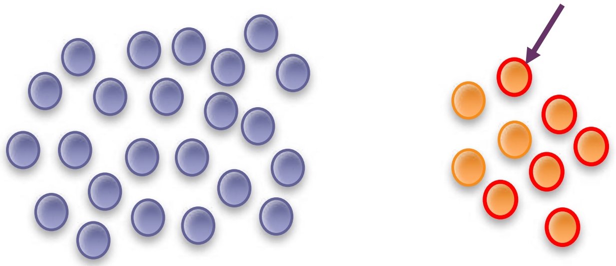
Before



Class1 Class2

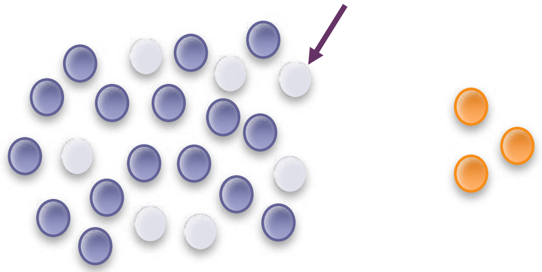
After

- Over Sampling



Class1 Class2

- Under Sampling



Class1 Class2

9) Splitting Data

- Splitting Data
  - Train, Test, Validate
    - Random
    - Stratification
  - K-Fold Cross Validation
  - K-Fold Cross Validation in Time Series Data

9) Splitting Data
- Train, Test, Validate

Training Data

inputs

target

Age	Income	Gender	Province	Purchase
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes
32	35,000	Male	Bangkok	No

Validation Data

Age	Income	Gender	Province	Purchase
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes

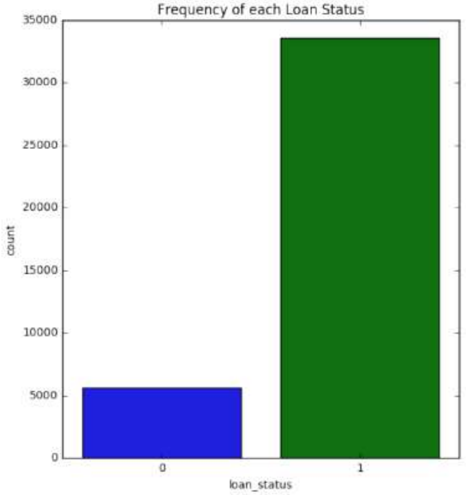
Testing Data

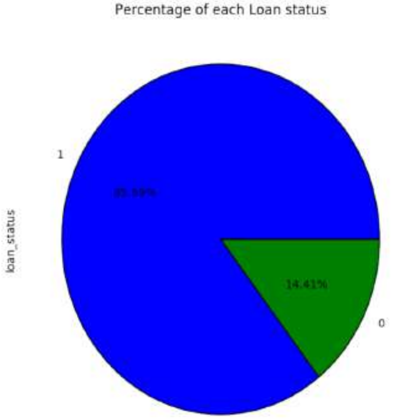
Age	Income	Gender	Province	Purchase
25	25,000	Female	Bangkok	?

9) Splitting Data
- Train, Test, Validate

Simple random sample

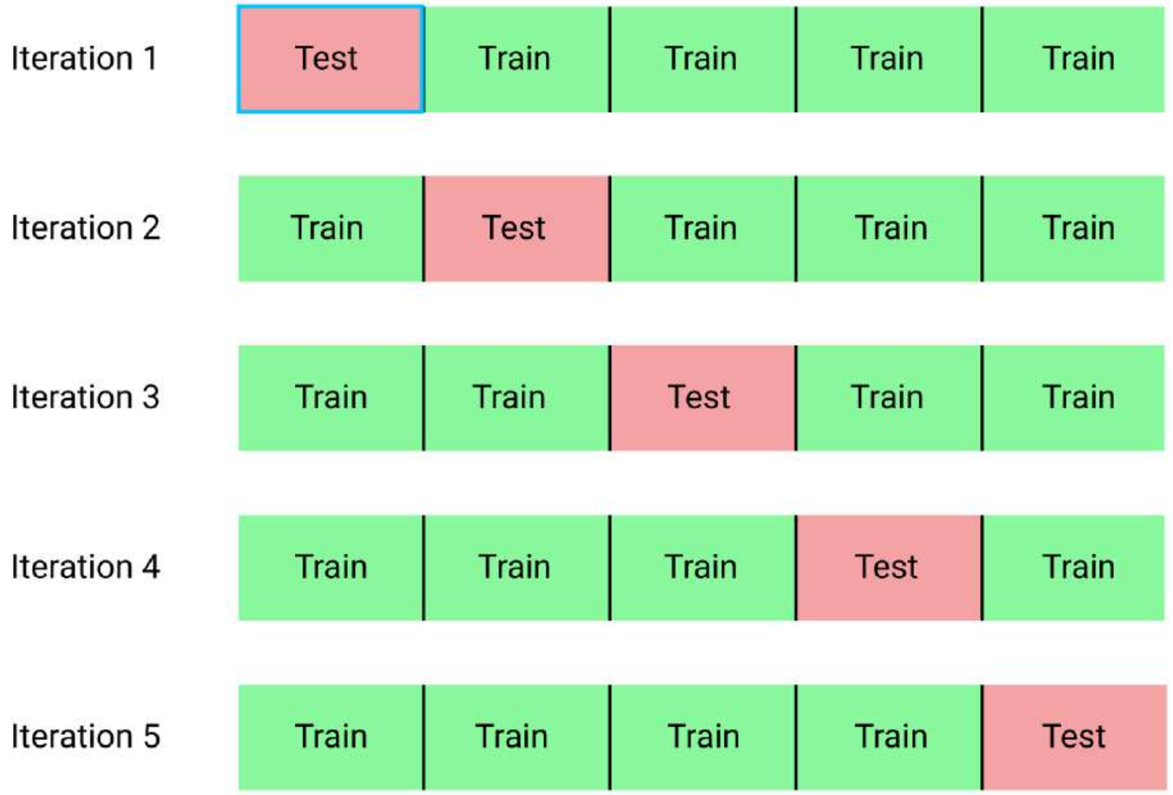
Stratification





9) Splitting Data
- K-Fold Cross Validation

How to fix overfitting issue on test



75%

80%

90%

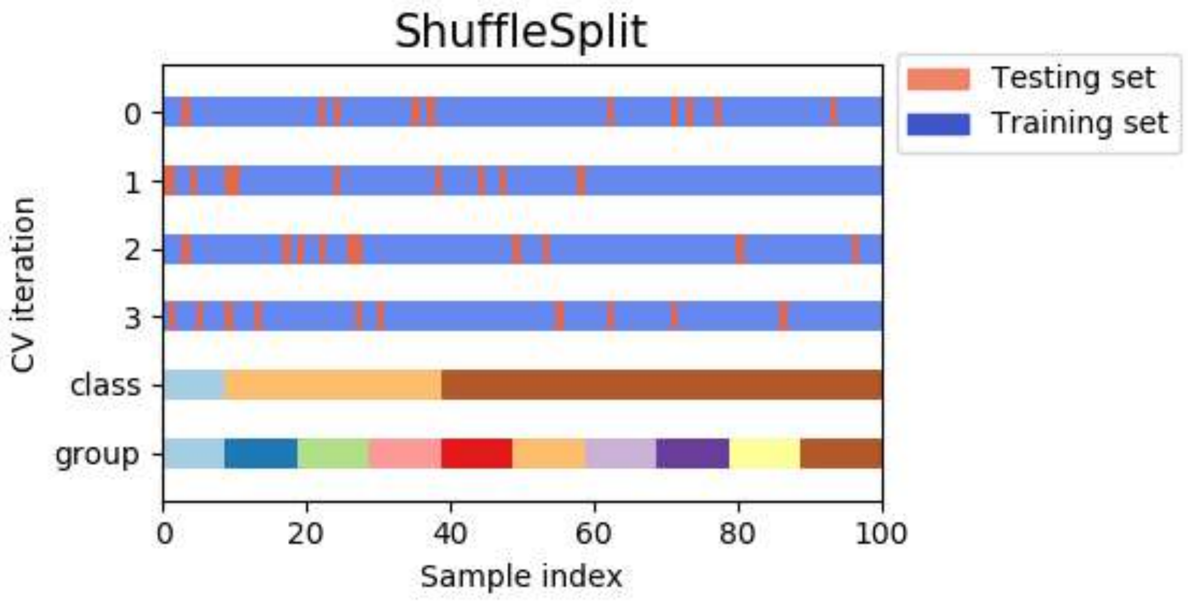
75%

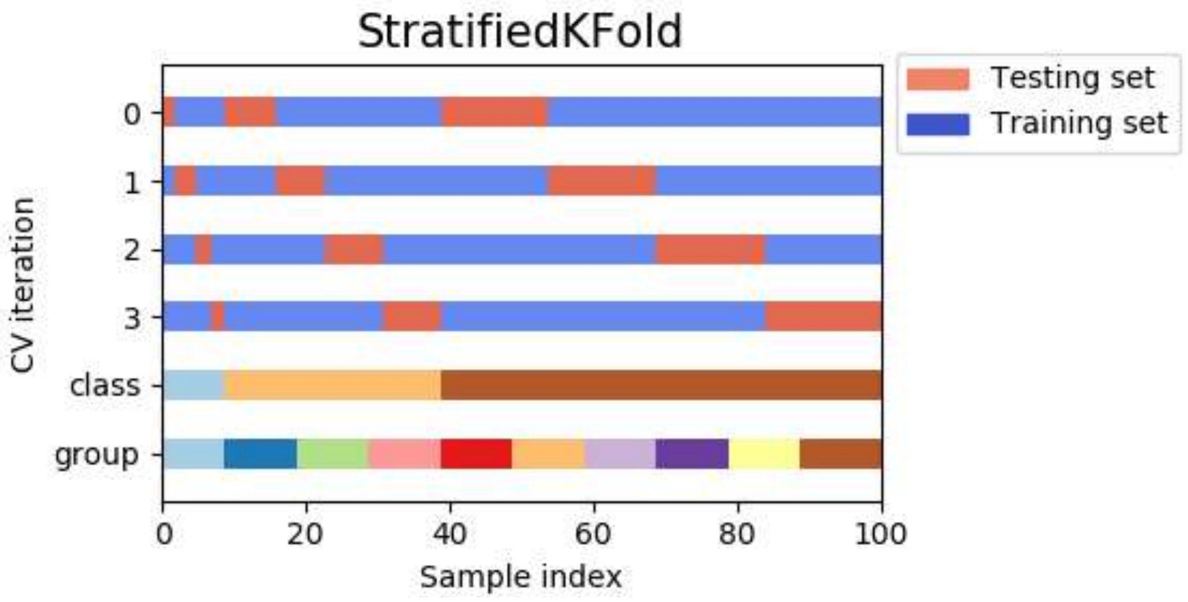
84%

Overall performance = mean(folds) = 80.8%

Reference : https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

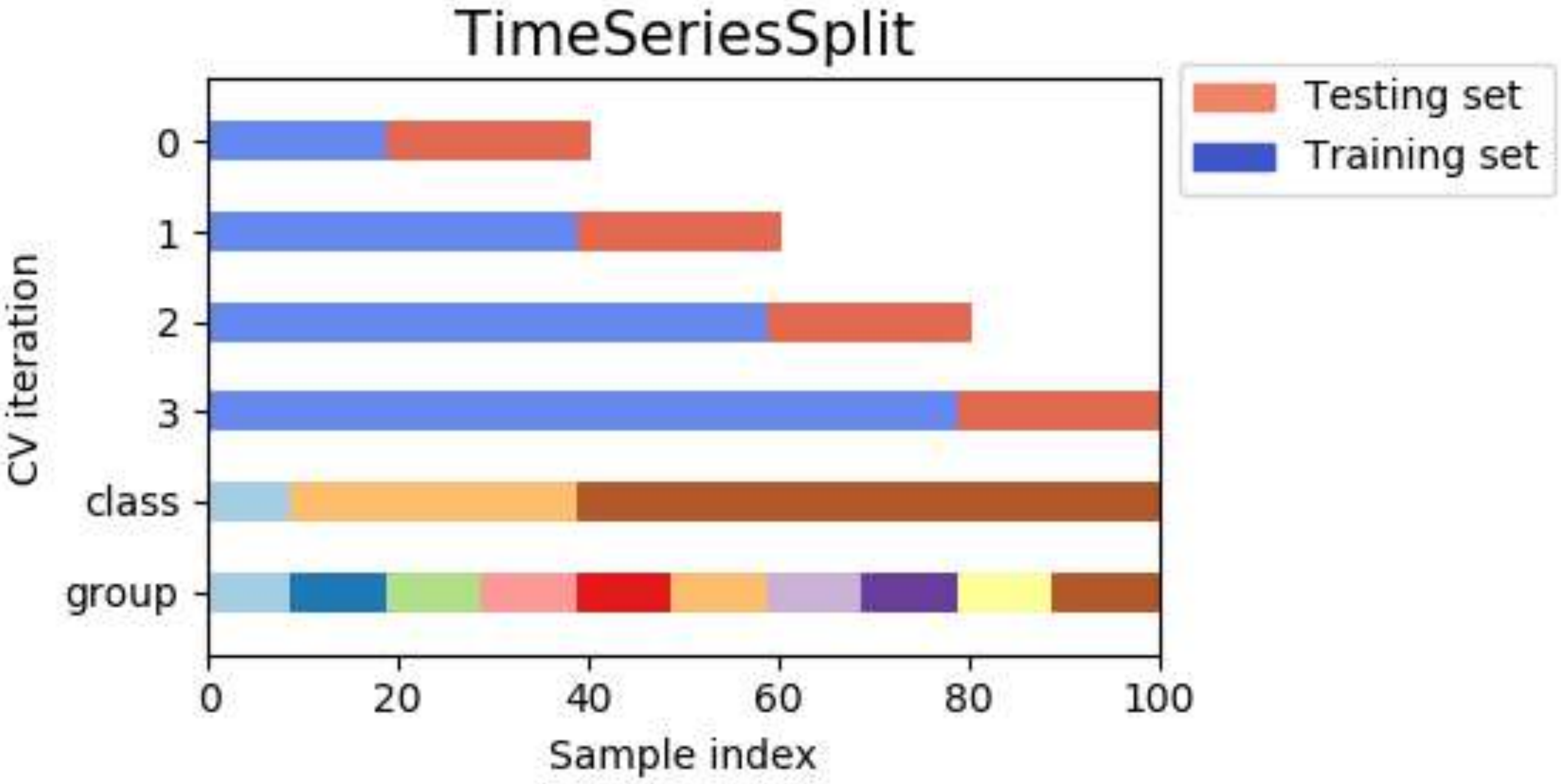
9) Splitting Data
- K-Fold Cross Validation





Reference : https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_indices.html

9) Splitting Data - K-Fold Cross Validation in Time Series Data



Reference : https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_indices.html

Remark: Random Seed

- The experiment must be able to reconstruct (replicate).
- All randoms must be assigned a radom seed.
  - random.seed(12345)
  - random_state option

+

Any Questions?