CHULA ENGINEERING
Foundation toward Innovation

COMPUTER

Chula Big Data and IoT
Center of Excellence
(CUBIC)

Data Mining Group
Machine Intelligence and Knowledge Discovery Lab
Chulalongkorn University

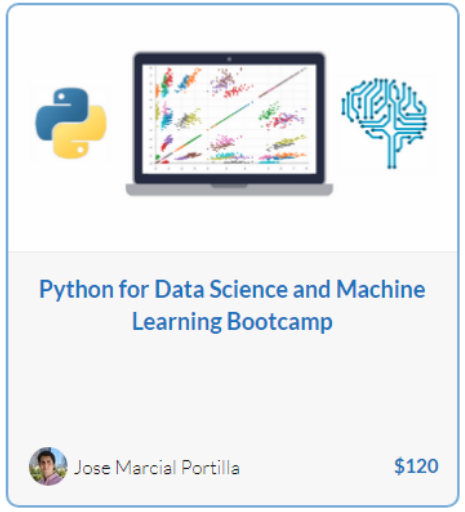
Classification & Clustering

Python for Data Analytics

Peerapon Vateekul, Ph.D.
Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University
Peerapon.v@chula.ac.th

Outlines

- Fundamental terminology
- Data analytics tasks
- Scikit-learn: Machine learning library in Python



PIERIAN DATA

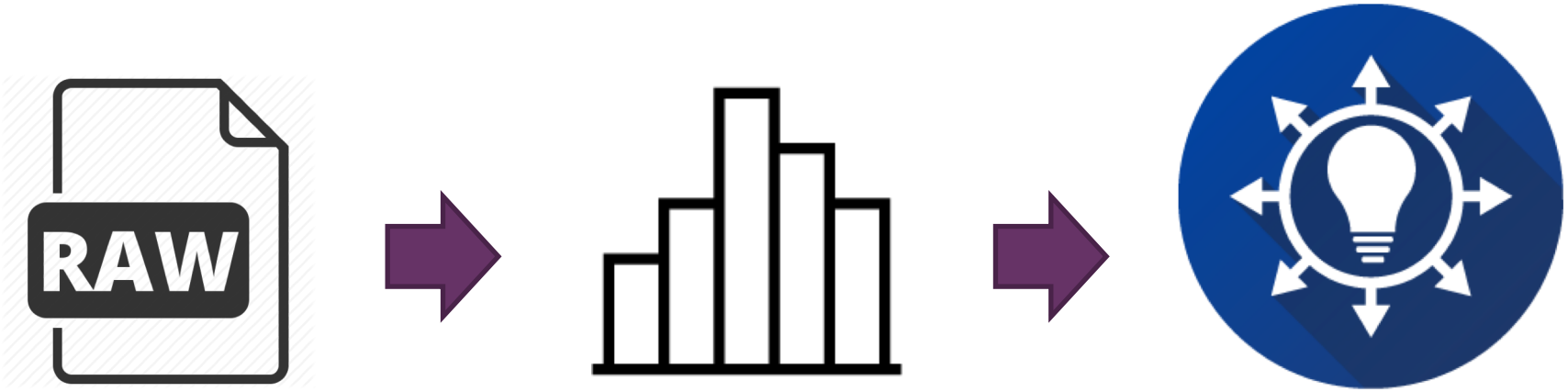
https://www.pieriandata.com/

- Understand data analytics **tasks**
- Be able to identify tasks and **tools** (technique) from a given problem

Data Analytics Tasks

Data analytics

- **Data analytics** refers to the science of examining and exploring data in order to (i) understand it and (ii) discover useful information.
- This can lead to **better decision** and also **data product**.



Data analytics process

4 • Data Visualization

3 • Data Analytics

2 • Data Storage

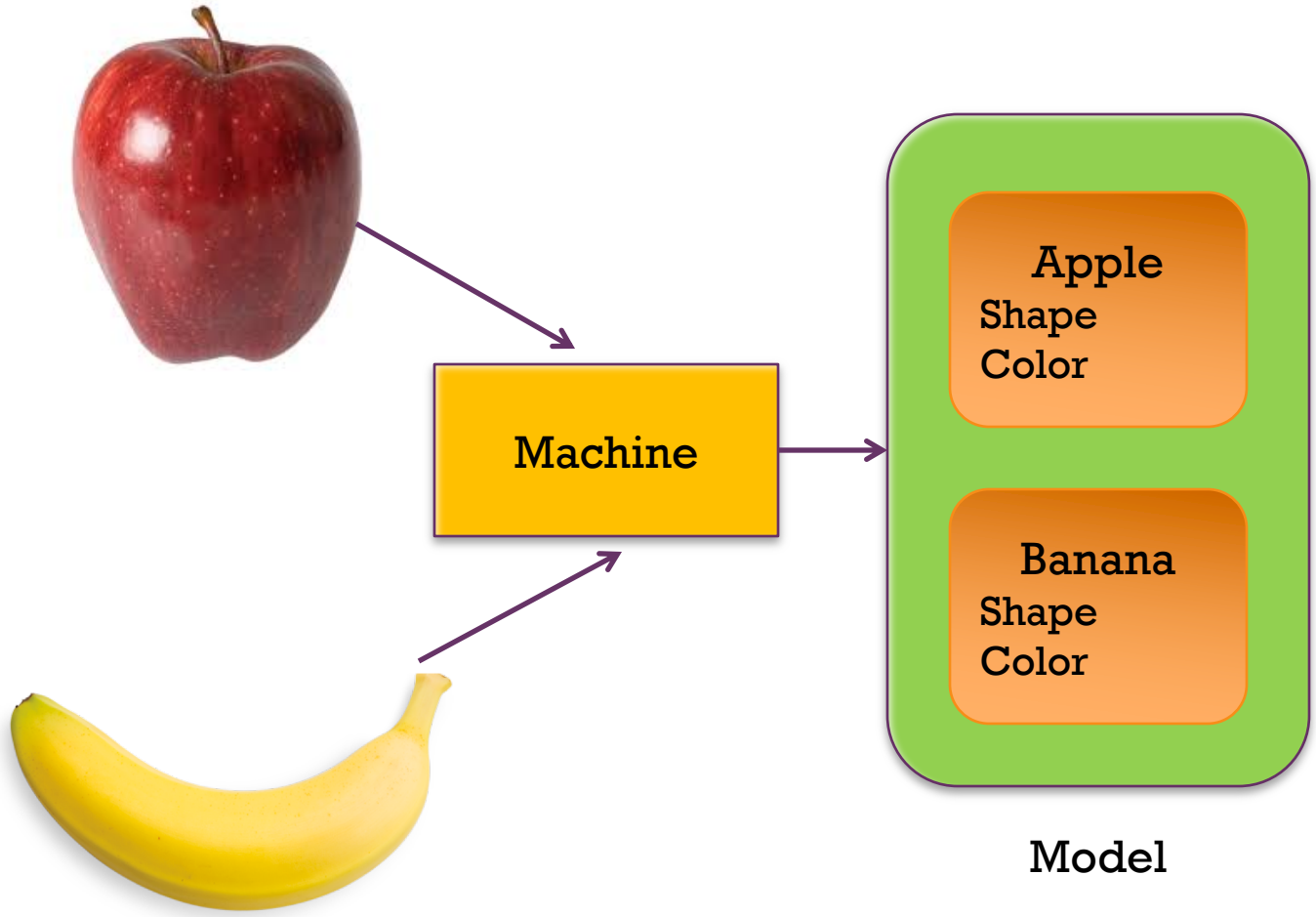
1 • System Infrastructure

Data Mining

- An automatic process of
- discovering useful information
- in large data repositories
- with sophisticated algorithm

- Supervised Learning (predictive task)
- Unsupervised Learning (descriptive task)
- Special task

Task1: Supervised learning (predictive task)

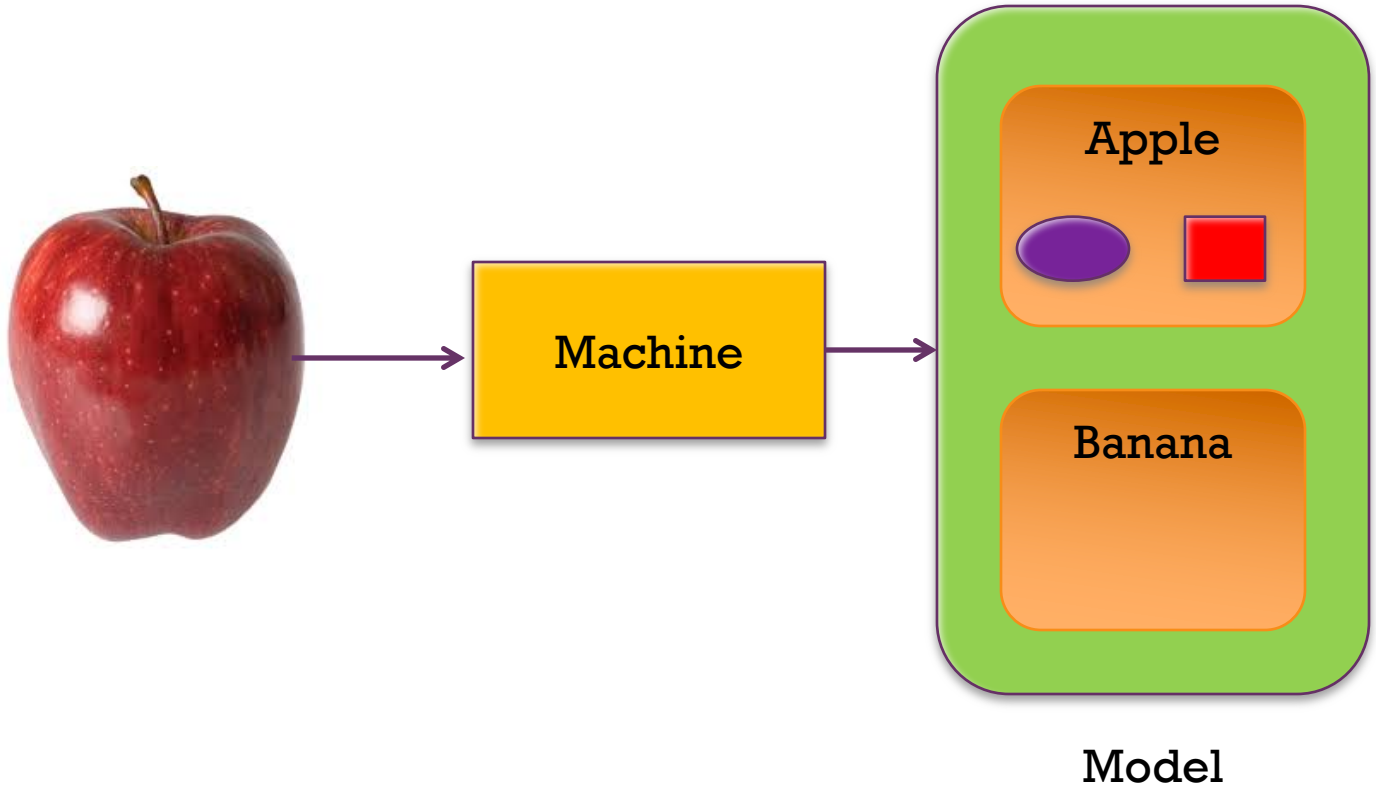


Supervised Learning

- Learn from labeled examples
- Example: Learn to classify apples from bananas
- Model: Criteria for classification

1. Training Phase
2. Testing Phase

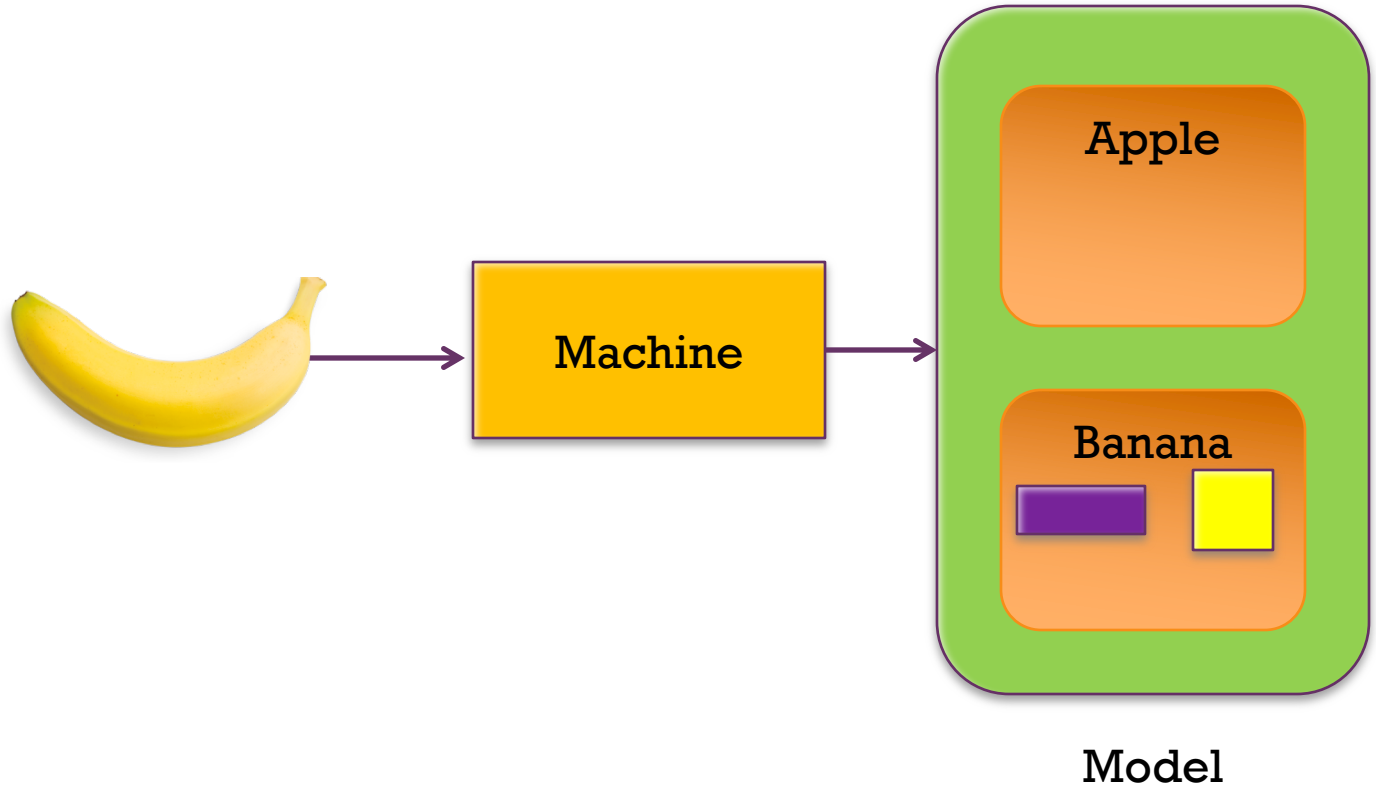
Supervised learning (cont.): Training Phase



Supervised Learning

- Learn from labeled examples
- Example: Learn to classify apples from bananas
- Model: Criteria for classification

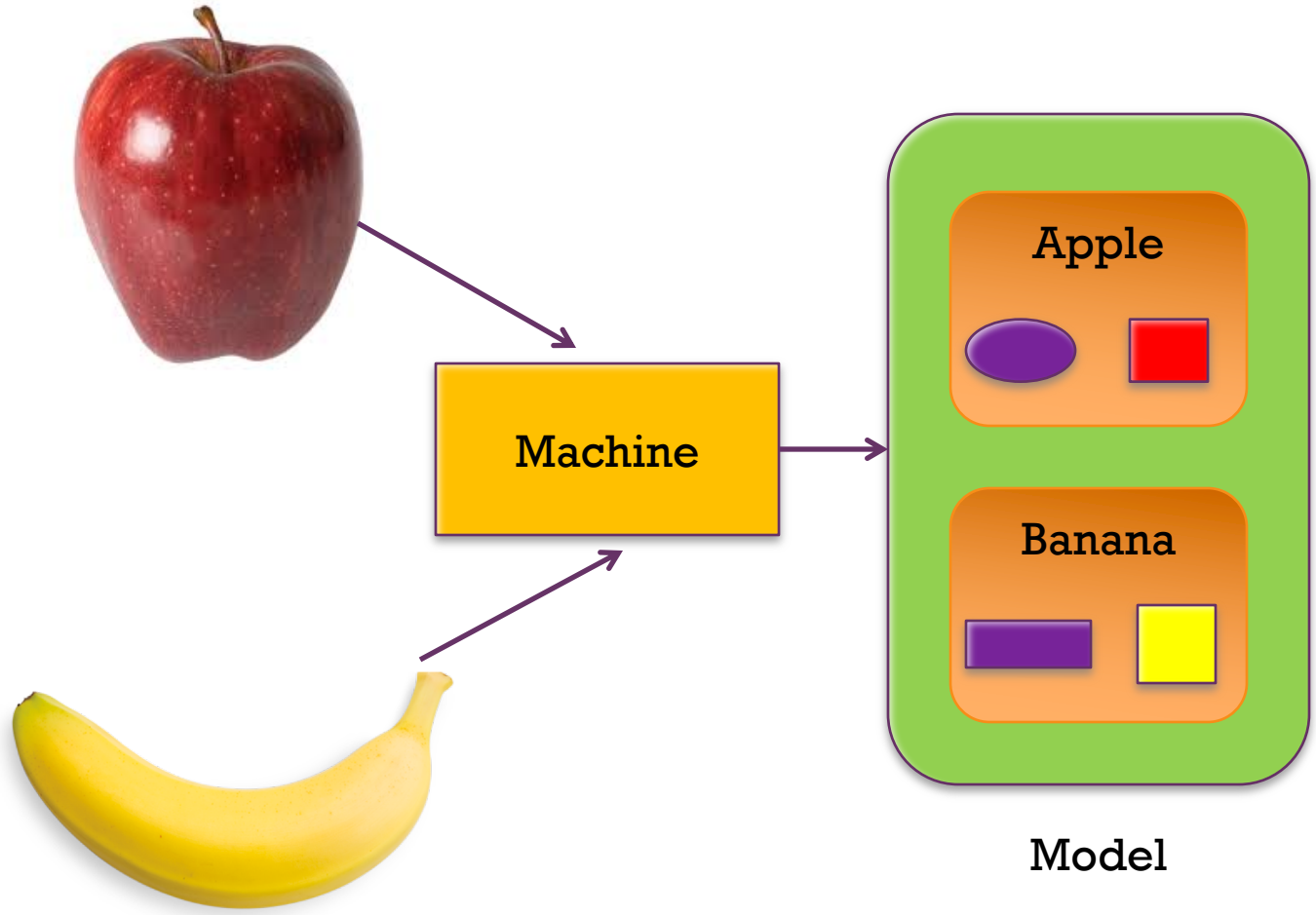
Supervised learning (cont.)
Training Phase



Supervised Learning

- Learn from labeled examples
- Example: Learn to classify apples from bananas
- Model: Criteria for classification

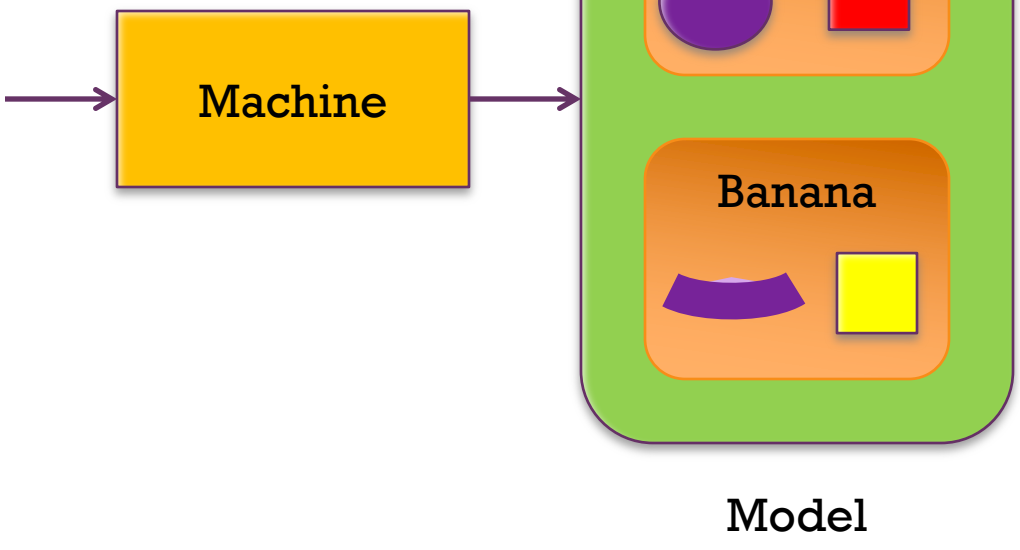
Supervised learning (cont.): Training Phase



Supervised Learning

- Learn from labeled examples
- Example: Learn to classify apples from bananas
- Model: Criteria for classification

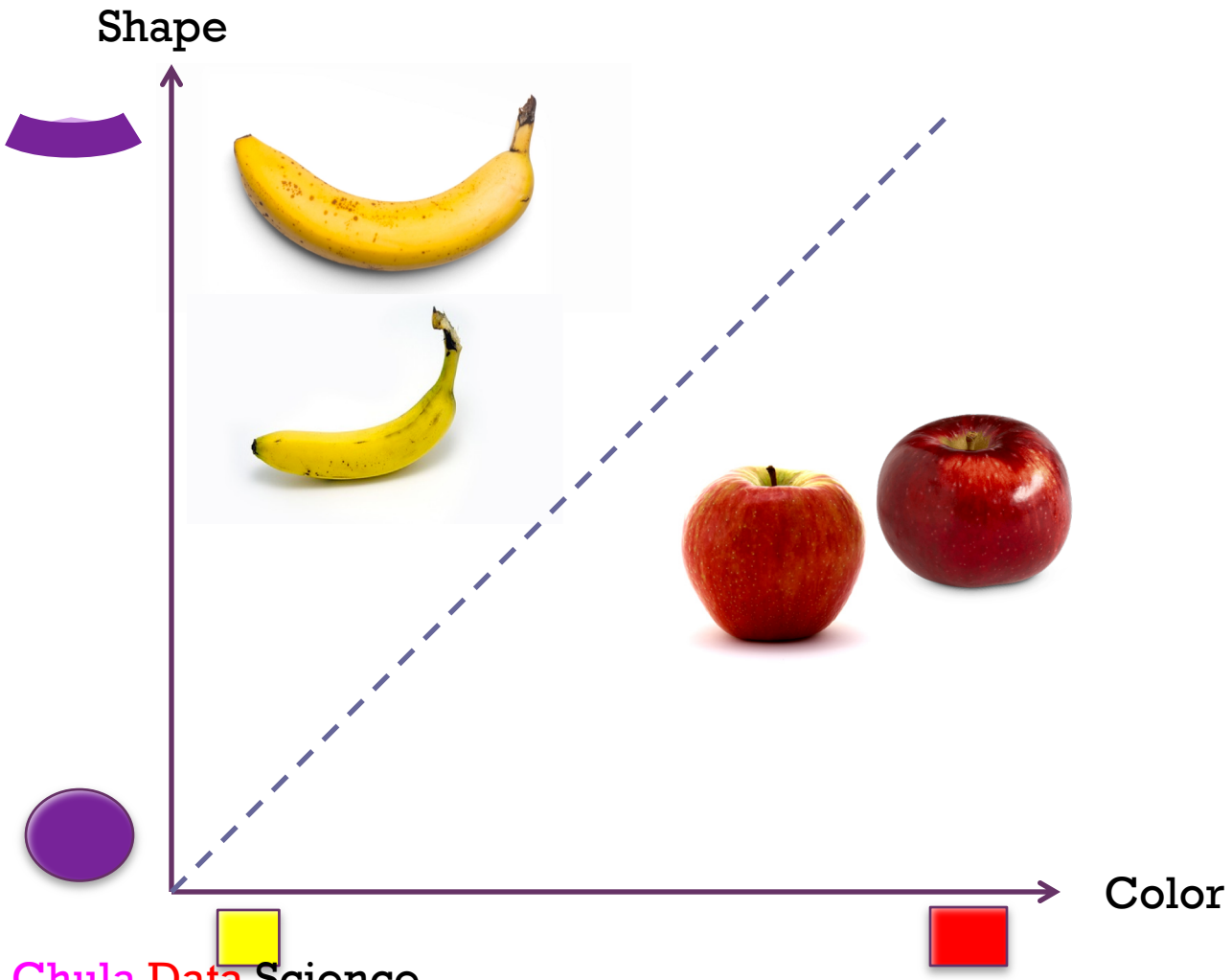
Supervised learning (cont.):
Training Phase → more examples



Supervised Learning

- Learn from labeled examples
- Example: Learn to classify apples from bananas
- Model: Criteria for classification

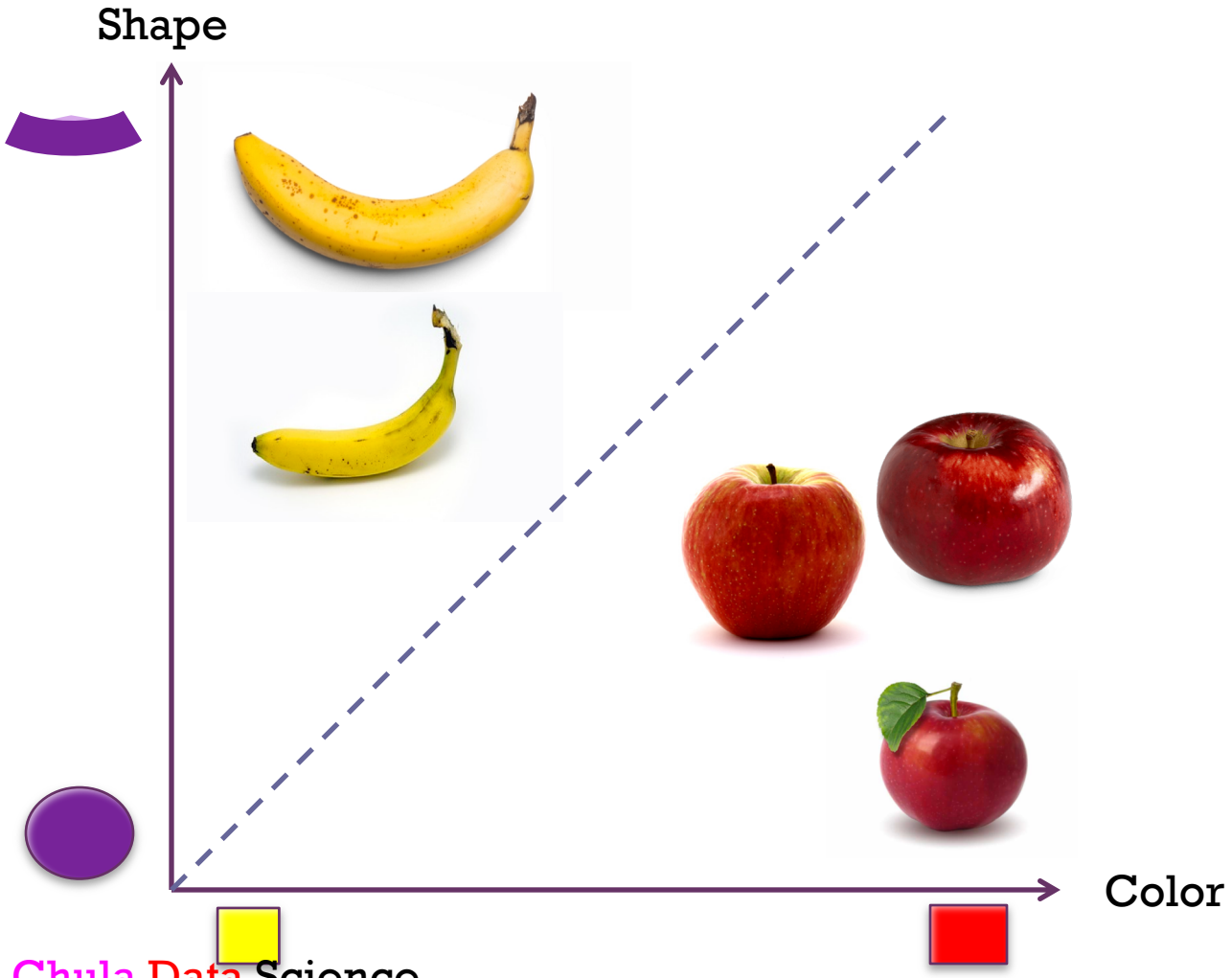
Supervised learning (cont.):
Training Phase → classification model



Supervised Learning

- Learn from labeled examples
- Example: Learn to classify apples from bananas
- Model: Criteria for classification

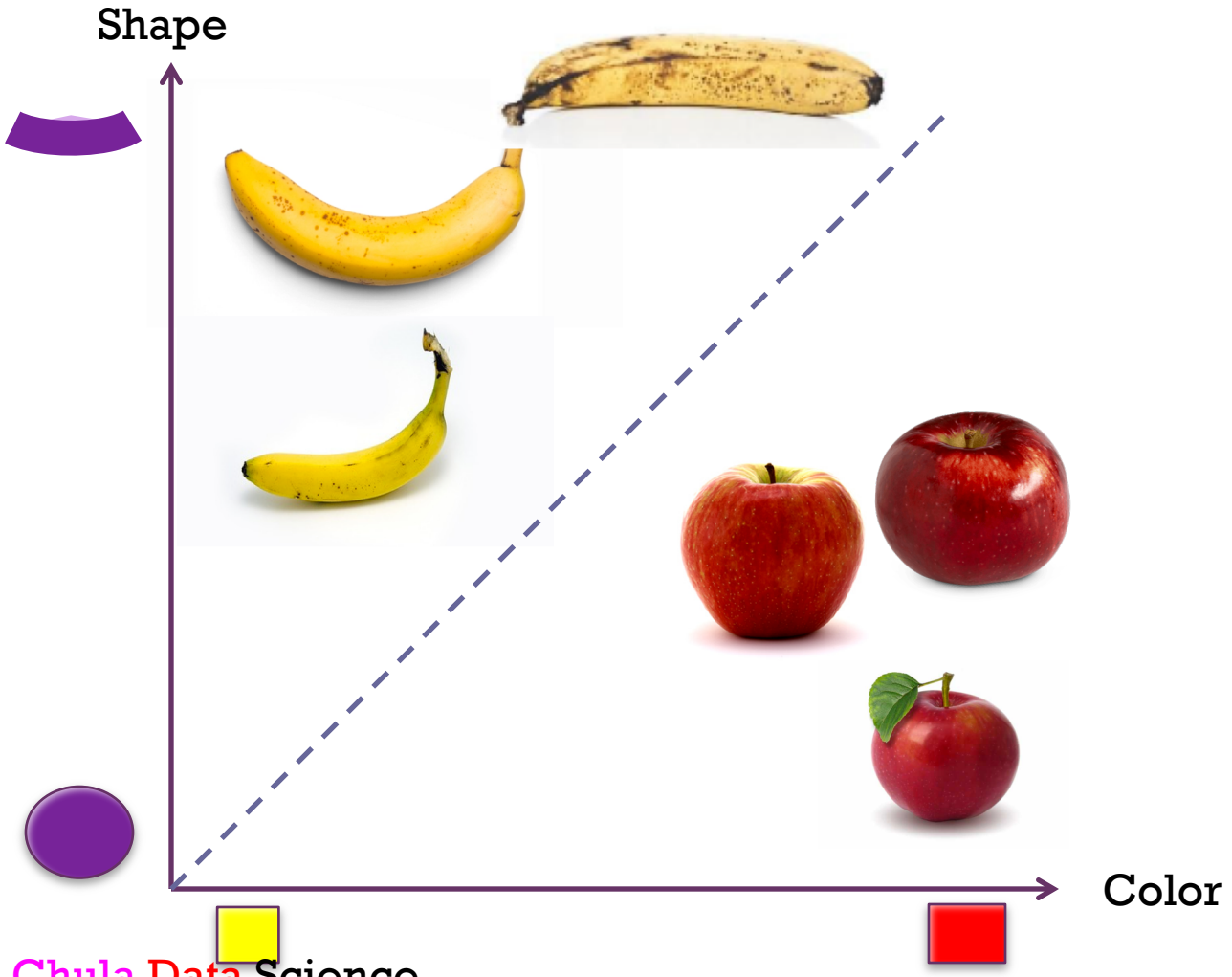
Supervised learning (cont.):
Testing Phase: case 1



Supervised Learning

- Learn from labeled examples
- Example: Learn to classify apples from bananas
- Model: Criteria for classification

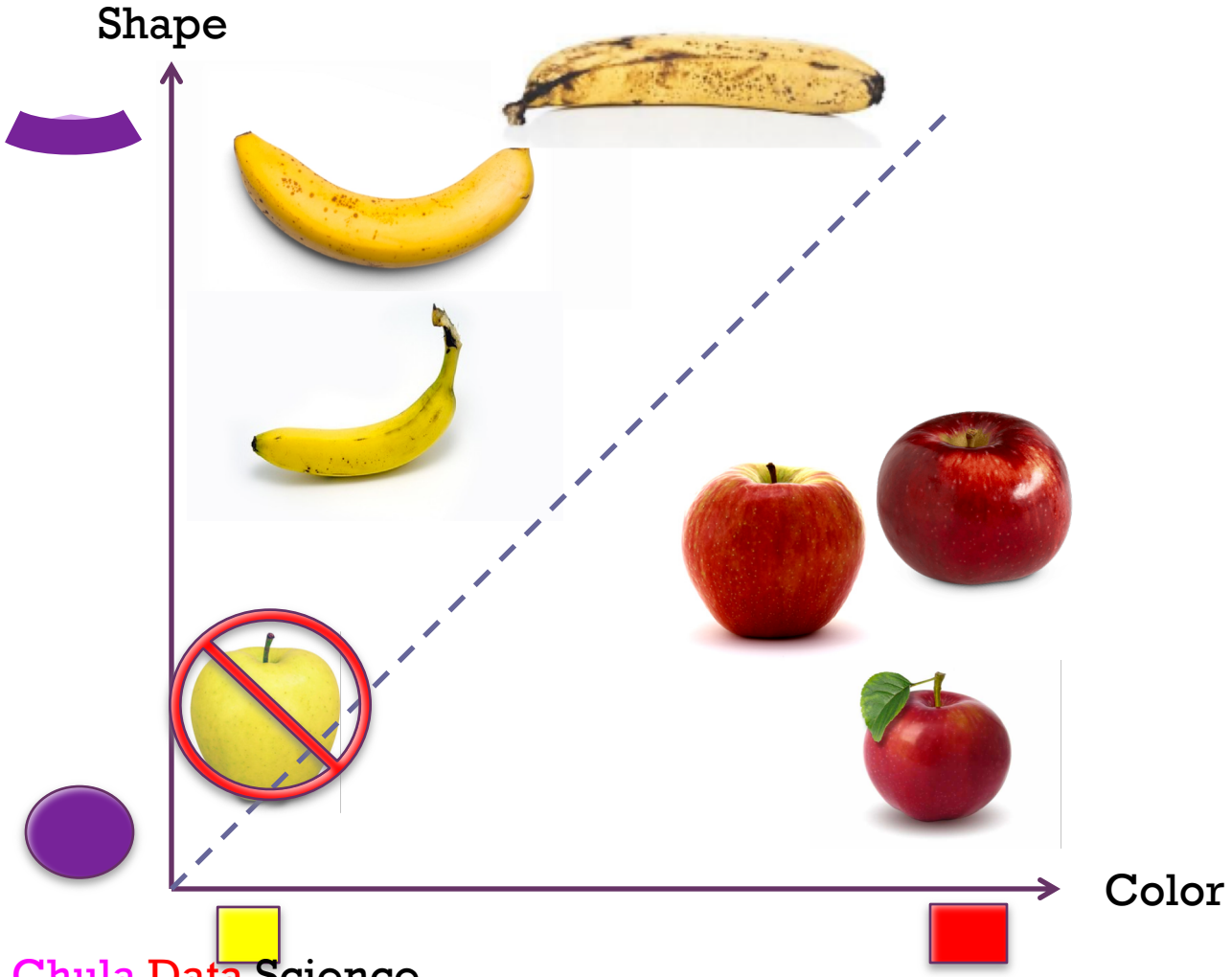
Supervised learning (cont.):
Testing Phase: case2



Supervised Learning

- Learn from labeled examples
- Example: Learn to classify apples from bananas
- Model: Criteria for classification

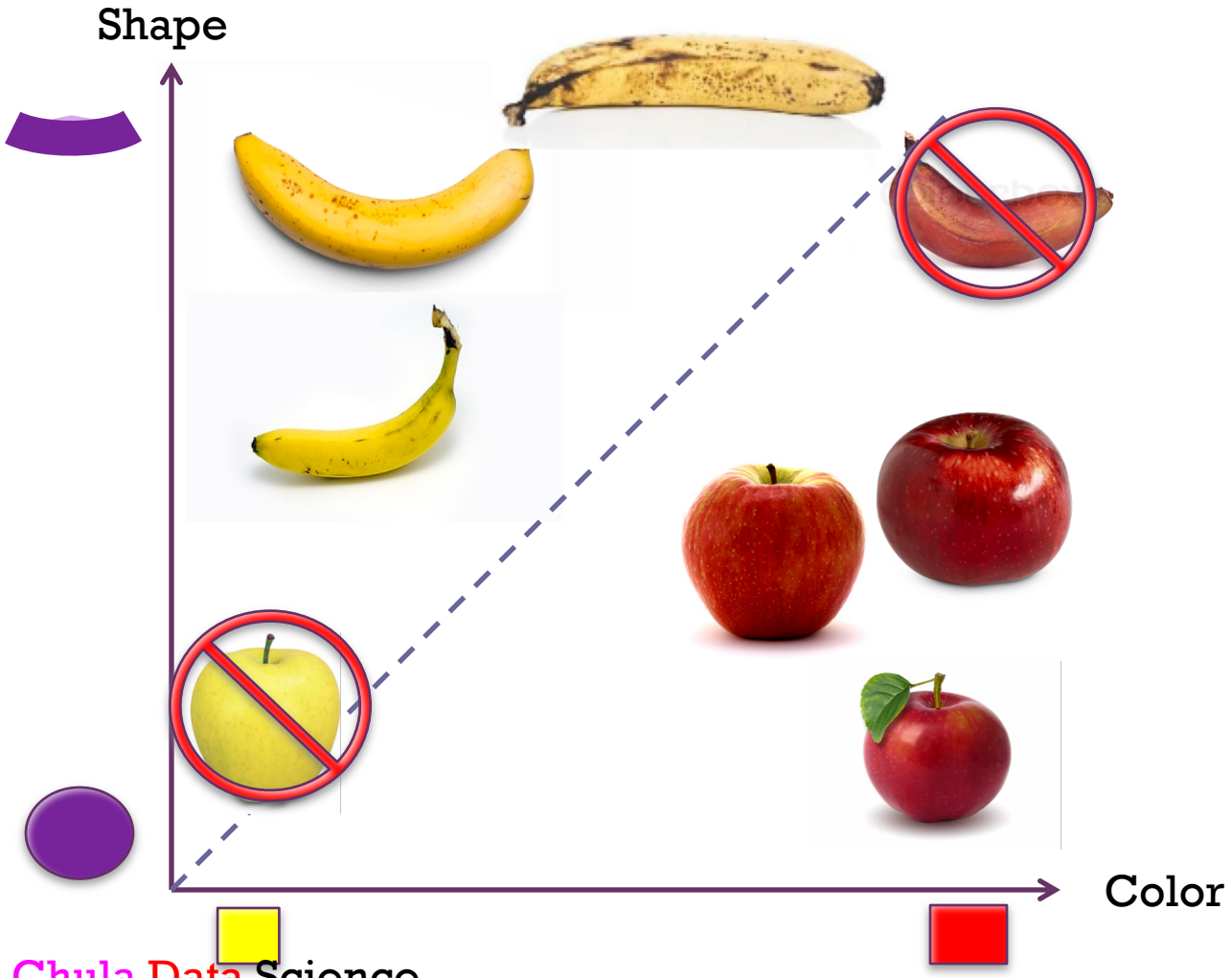
Supervised learning (cont.):
Testing Phase: case3



Supervised Learning

- Learn from labeled examples
- Example: Learn to classify apples from bananas
- Model: Criteria for classification

Supervised learning (cont.): Testing Phase: case4



Supervised Learning

- Learn from labeled examples
- Example: Learn to classify apples from bananas
- Model: Criteria for classification

Supervised learning (recap)

Training Data

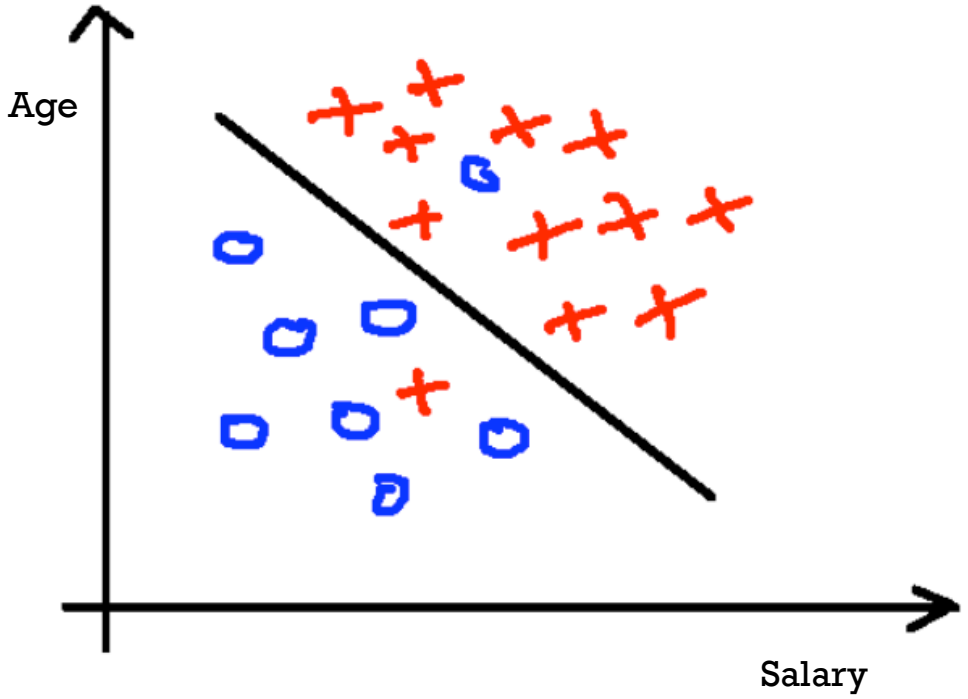
Age	Income	Gender	Province	Purchase
		inputs		target
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes
32	35,000	Male	Bangkok	No

Testing Data

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

Application: Direct Target Customer

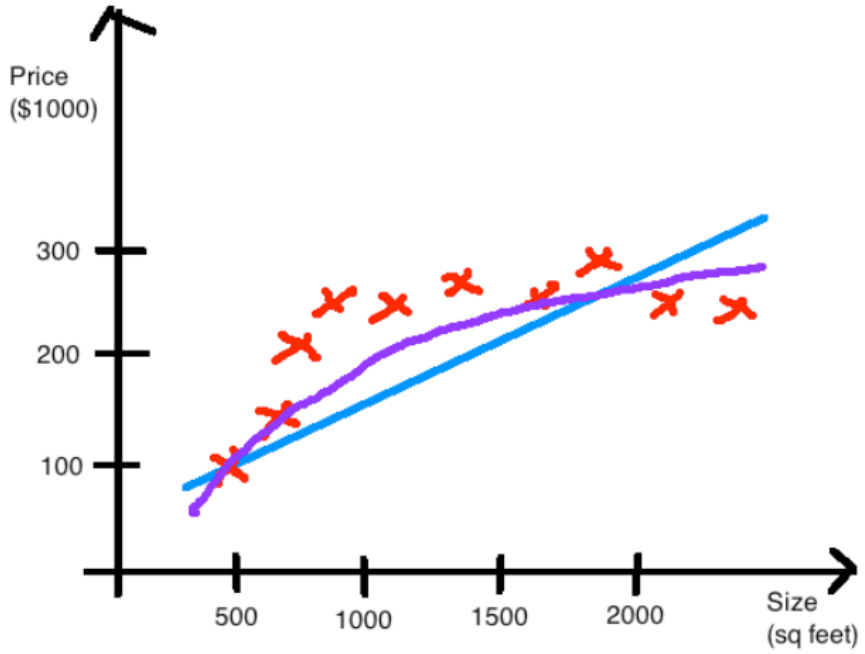
Classification: Predicting a categorical target



Predict targeted customers who tend to buy our product (yes/no)

- **Sample Applications**
  - Database marketing
  - Fraud detection
  - Pattern detection
  - Churn customer detection

Regression: Predicting a continuous target



Predict a sale price of each house

- **Sample Applications**
  - Financial risk management
  - Revenue forecasting
  - Loss reserving

inputs		target
Gender	**Province**	**Amount**
Female	Bangkok	$7,800
Female	Nontaburi	$500
Male	Bangkok	$12,000

Prediction algorithms

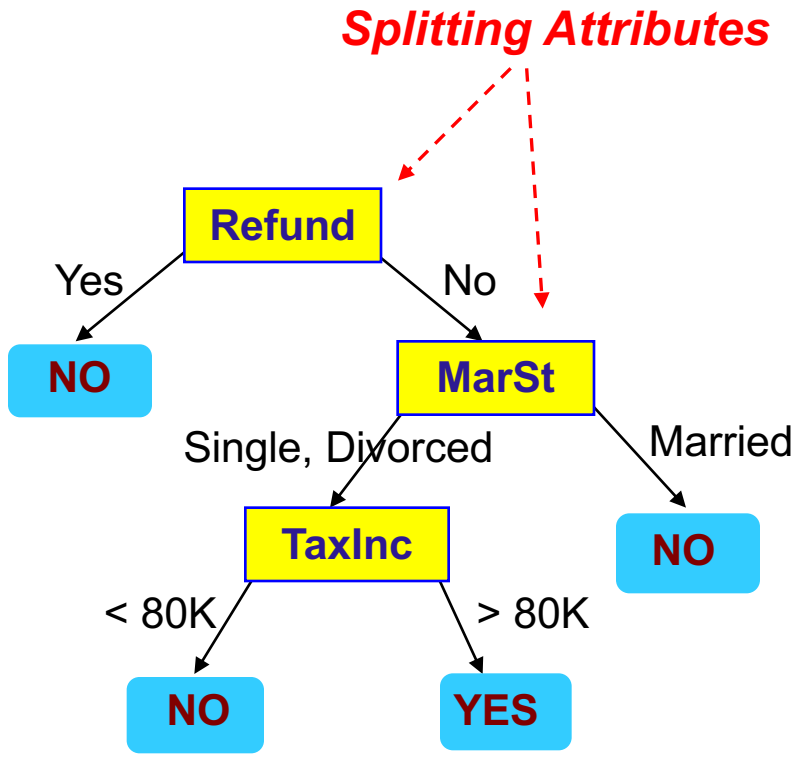
- Decision Tree
- (Logistic) Regression
- kNN
- Support Vector Machine
- Neural Networks (NN)
- Deep Learning

Decision Tree

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

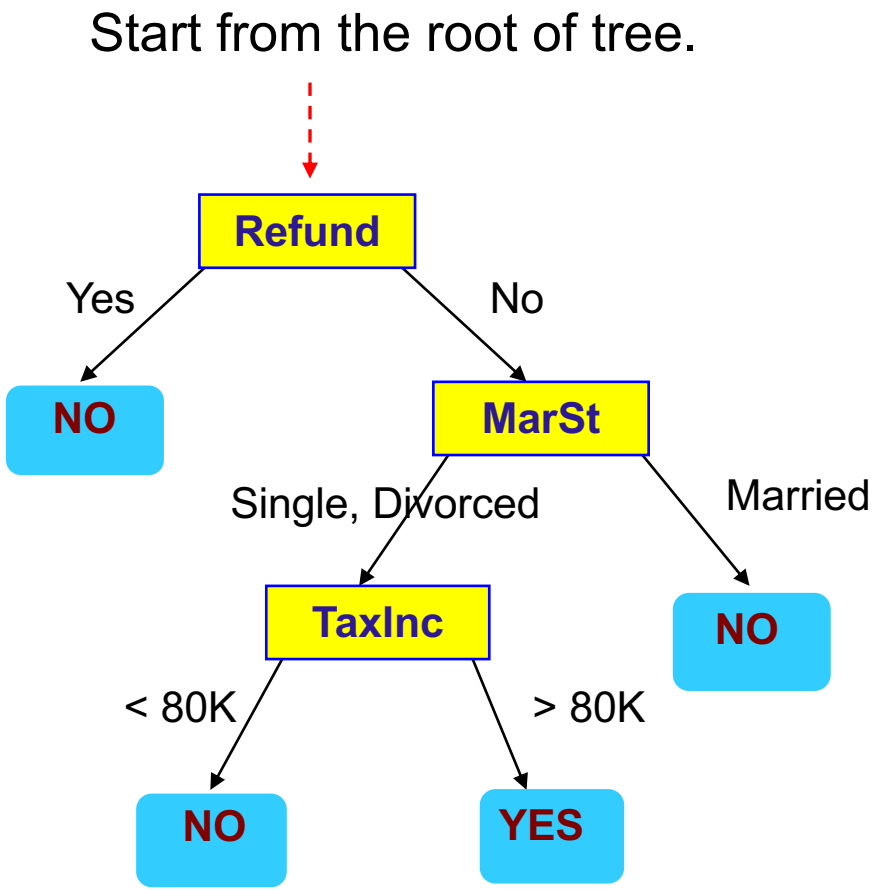
categorical categorical continuous class

Training Data



Model: Decision Tree

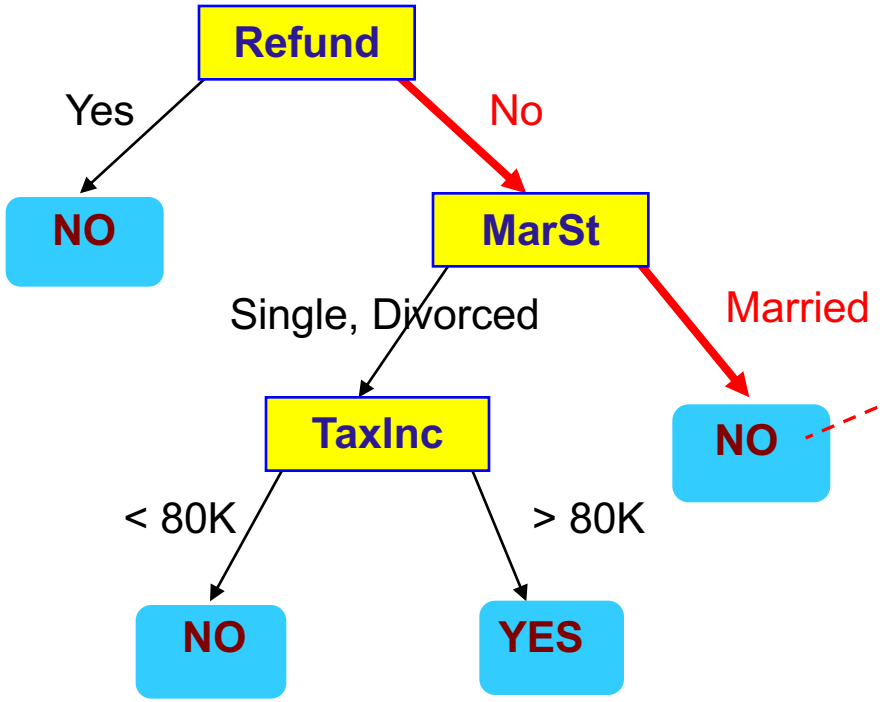
Decision Tree (cont.)



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Decision Tree (cont.)

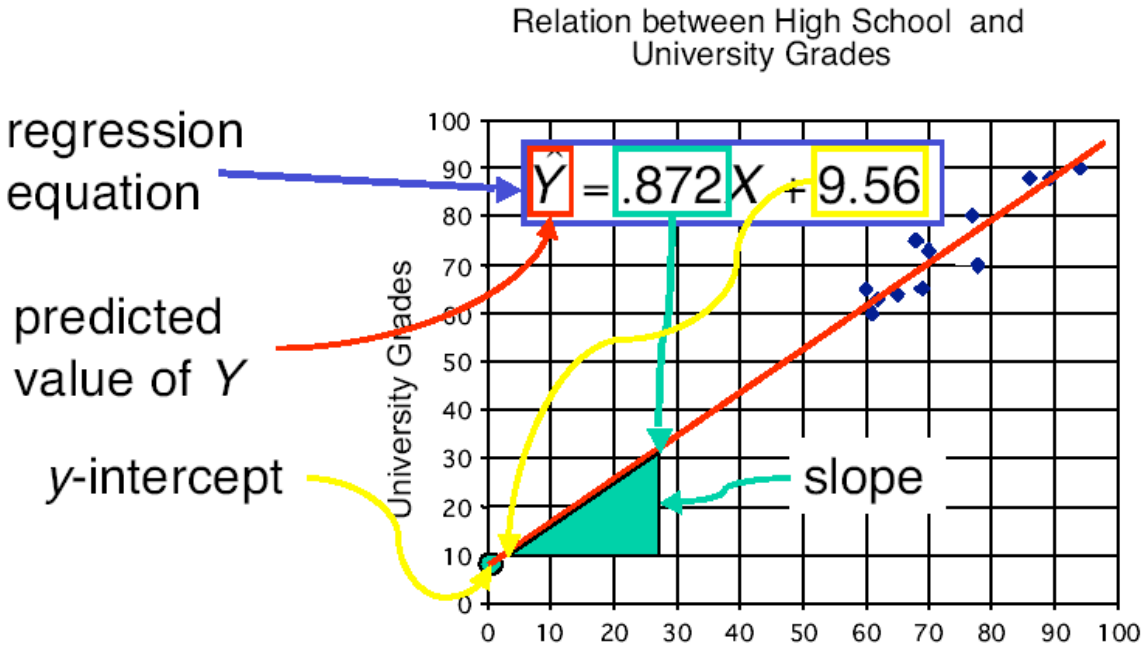


Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Assign Cheat to “No”

Regression



weight, coefficient

$$\underset{\text{target}}{\hat{y}} = \underset{\text{intercept}}{\hat{w}_0} + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

input

- The least square method aims to minimize the following term

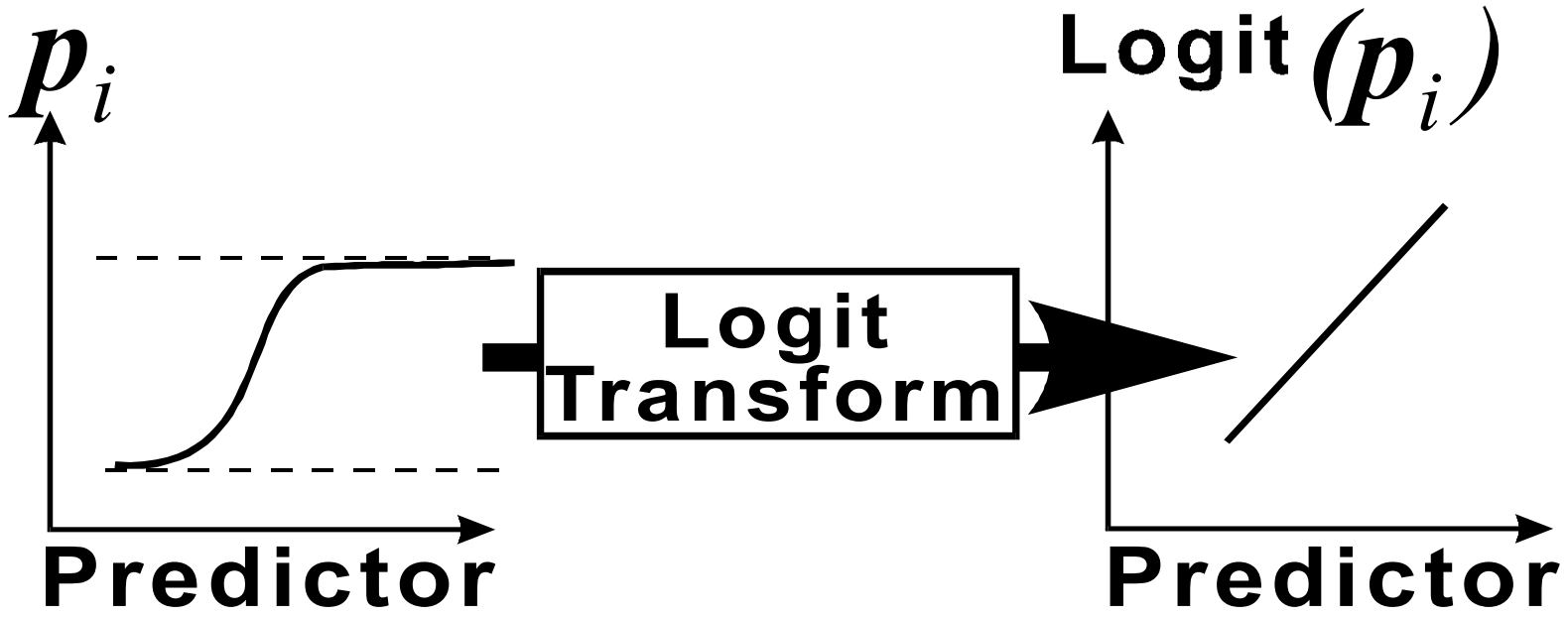
$$\sum_{\substack{training \\ data}} (y_i - \hat{y}_i)^2$$

Logistic regression

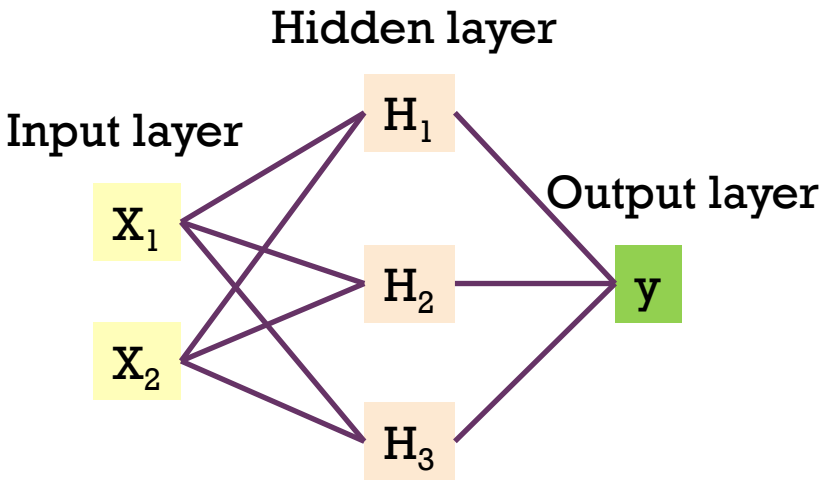
$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 = logit(\hat{p})$$

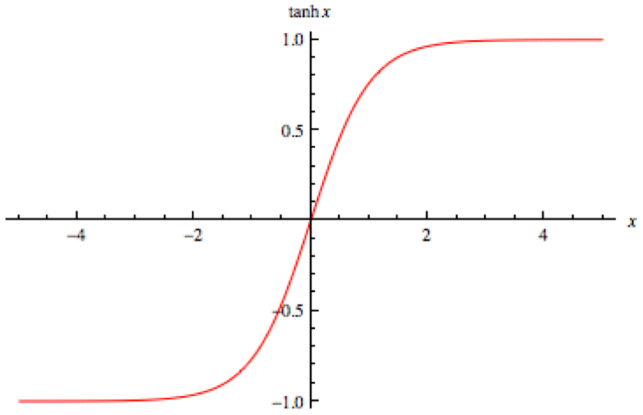
$$\hat{p} = \frac{1}{1 + e^{logit(\hat{p})}}$$

- Maximum likelihood estimates



Neural Networks (universal approximator)





$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 H_1 + \hat{w}_2 H_2 + \hat{w}_3 H_3$$

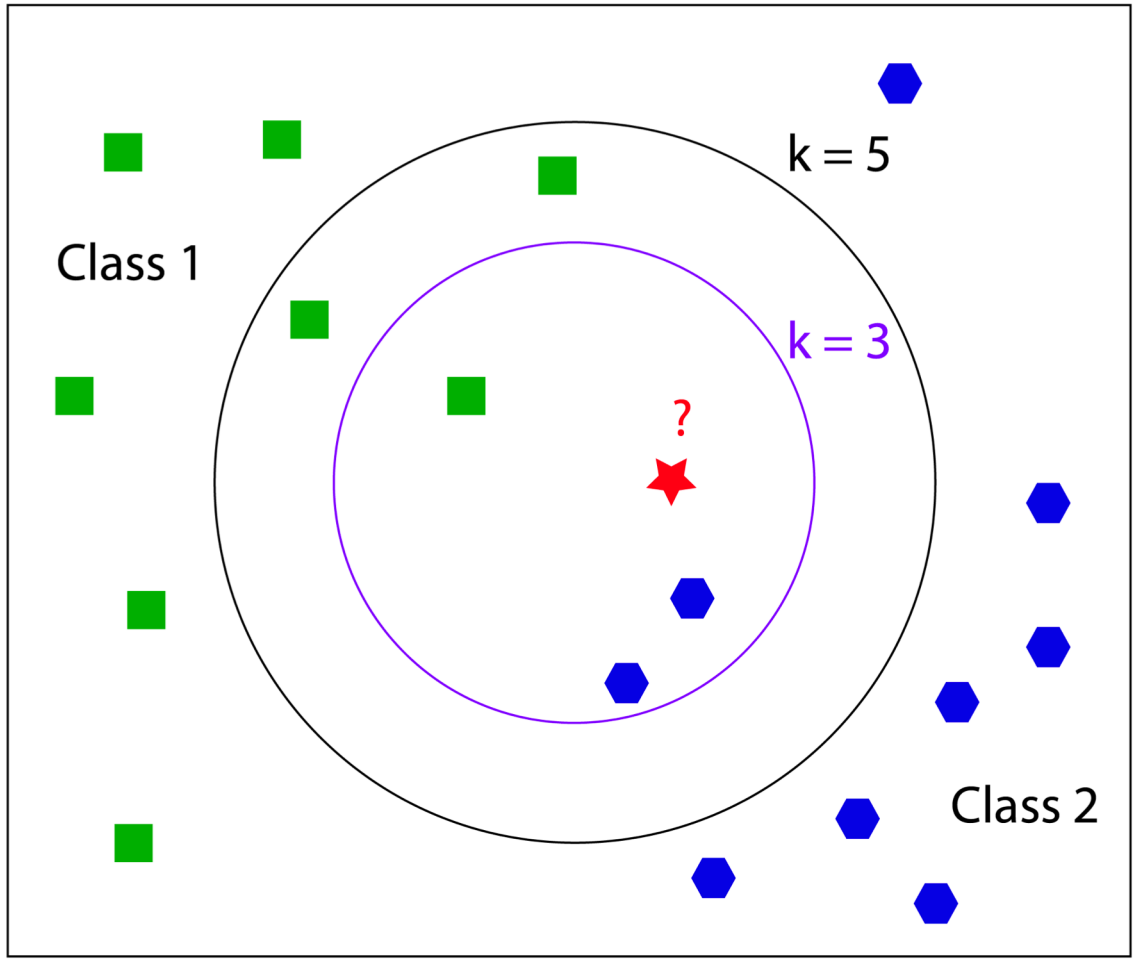
$$H_1 = tanh(\hat{w}_{10} + \hat{w}_{11} x_1 + \hat{w}_{12} x_2)$$

$$H_2 = tanh(\hat{w}_{20} + \hat{w}_{21} x_1 + \hat{w}_{22} x_2)$$

$$H_3 = tanh(\hat{w}_{30} + \hat{w}_{31} x_1 + \hat{w}_{32} x_2)$$

k-Nearest Neighbors (kNN)

- Memory based learning
- Suitable for small data sets
- Merge
  - Voting
  - Average
  - Maximum prob

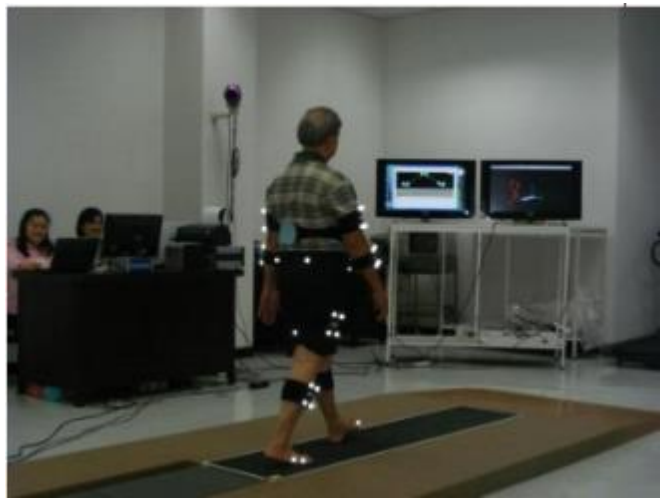


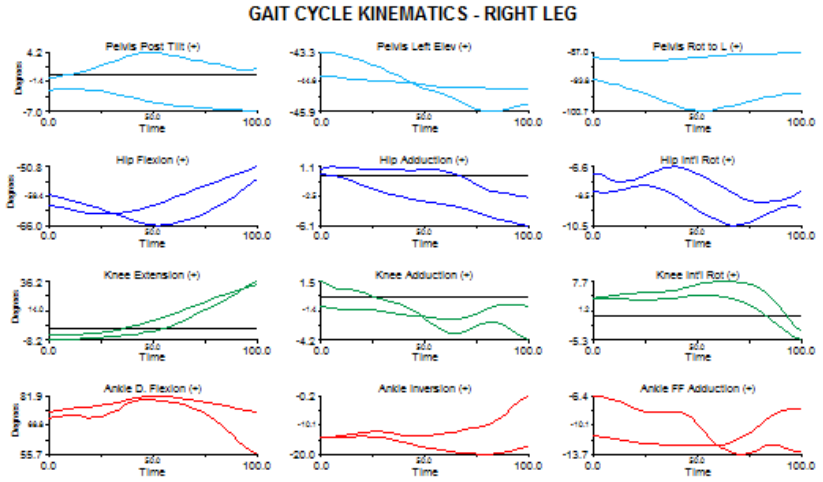
Model Evaluation

- Estimate
  - Sum Squared Error (SSE)
  - Average Squared Error (ASE)
- Decision
  - **Accuracy**
  - Misclassification
  - Precision
  - Recall
  - F1
- Ranking
  - ROC Curve
  - Area Under ROC (c-statistic)
  - **Lift**
  - Gain
  - Response

Case Study: Automated Medical Diagnosis on Movement Disorder Using Gait Data

- This work proposed an automated medical diagnosis in order to classify patients into three classes:
  - Normal, Sick/Knee OA, Sick/Parkinson.





Task2: Unsupervised learning (descriptive task)

1. Clustering
2. Association Rule Mining

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	

Testing Data

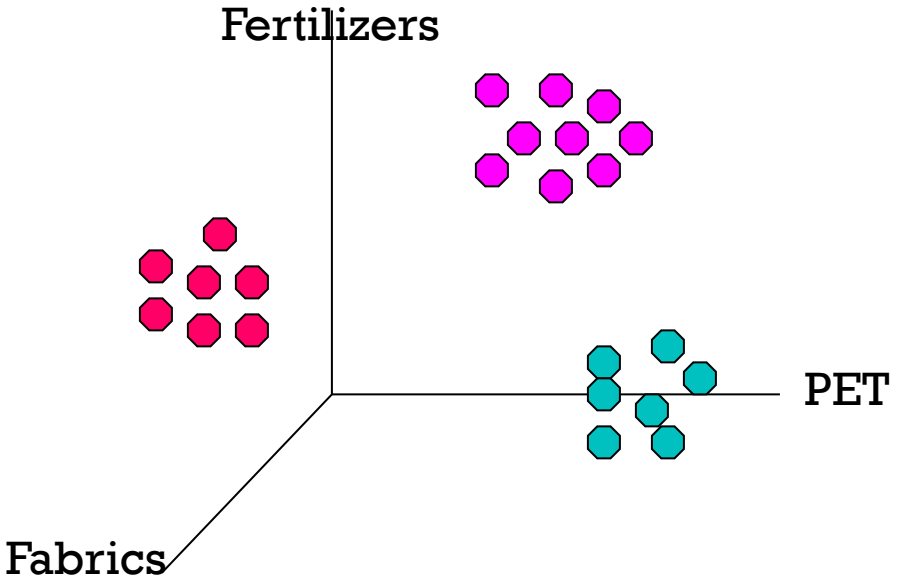
Age	Income		Province	Purchase
25	25,000		Bangkok	?

Clustering

- In our class, there are many participants. Should we teach them using the same method?
- May be not! Since they may have different learning behaviors and backgrounds.
- Inputs
  - Education field
  - Level of English communication
  - Level of computer skills
  - Age range
  - Gender

Clustering (cont.)

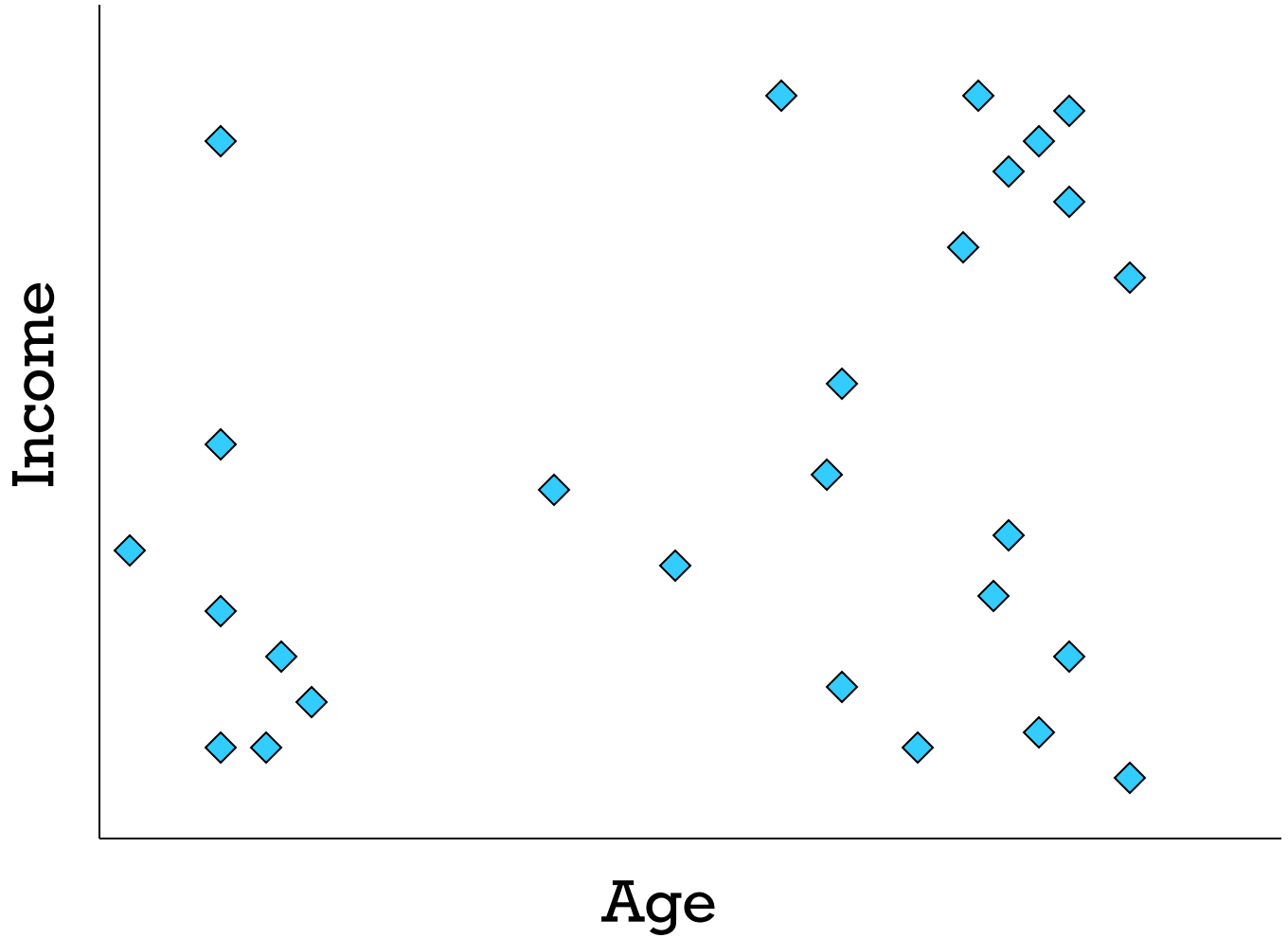
Company	Sedan	Truck	Motorcycles
C1	70M	2M	80M
C2	90M	120M	100M
C3	1M	8M	70M



- **Some techniques:**
  - k-means
  - DB-scan
  - Hierarchical clustering

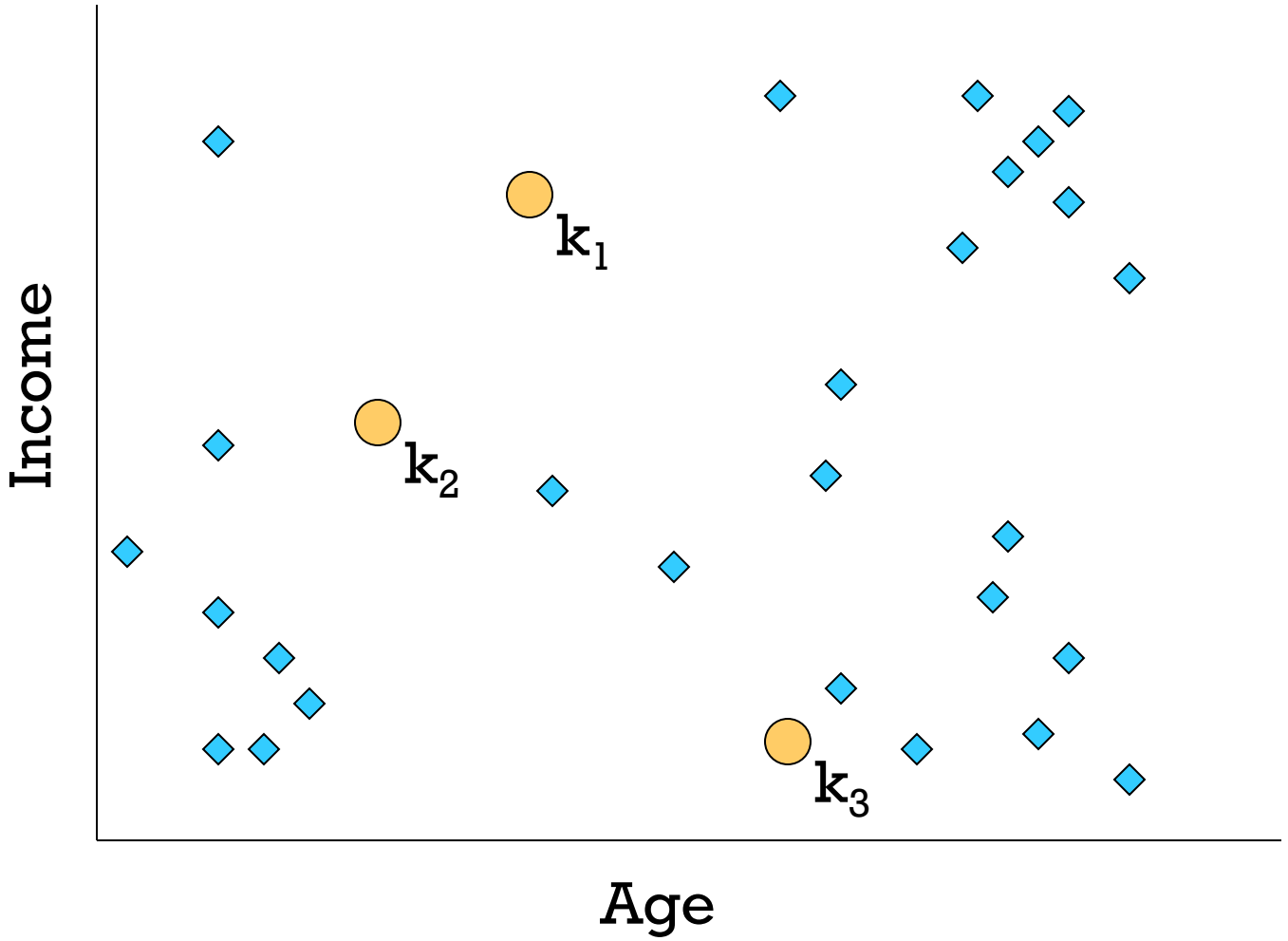
Example: Customer Segmentation

K-means: Step0



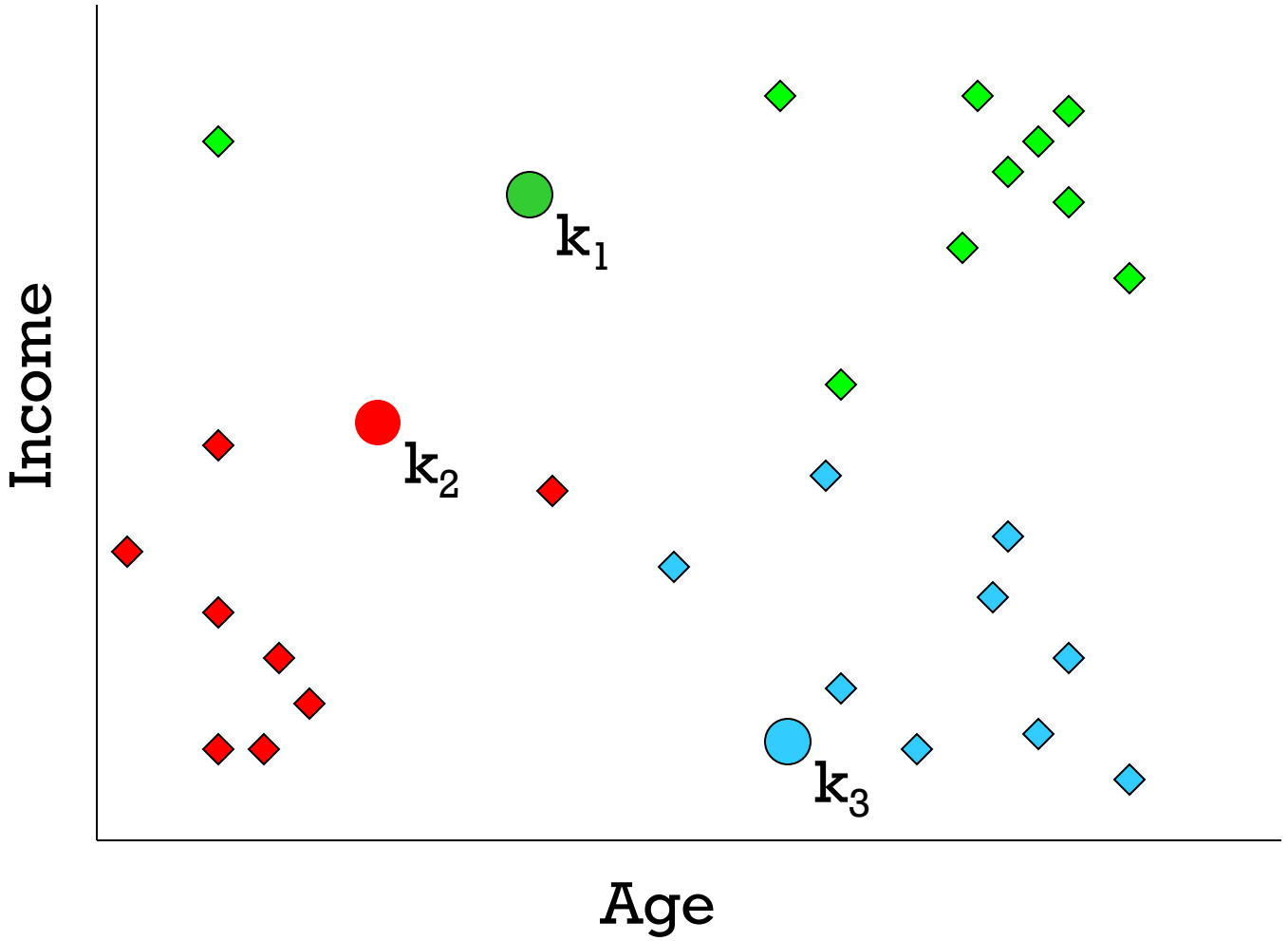
K-means: Step 1

Pick 3 initial cluster centers (randomly)



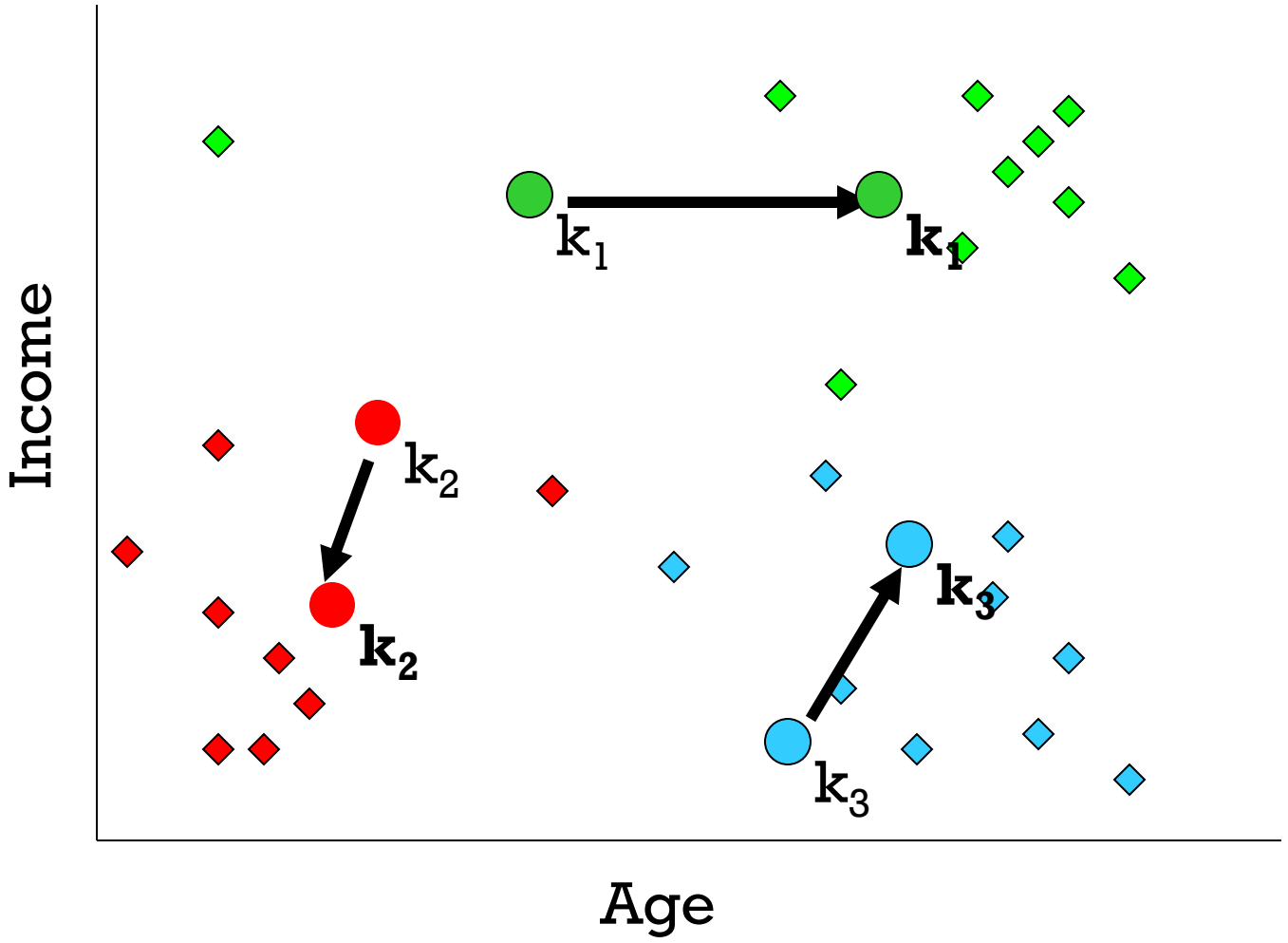
K-means: Step2

Assign each point to the closest cluster center



K-means: Step3

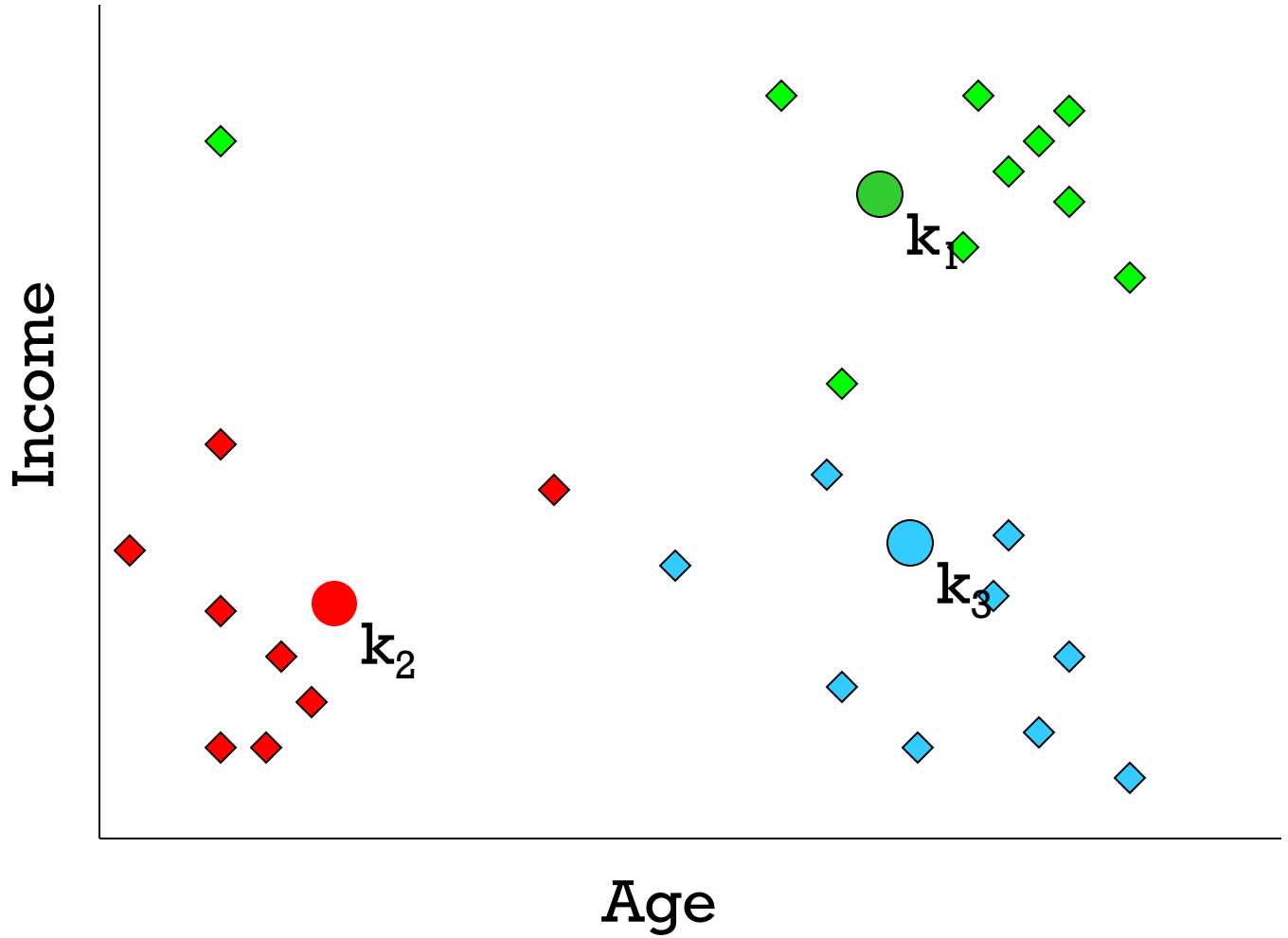
Move
each cluster
center
to the mean
of each cluster



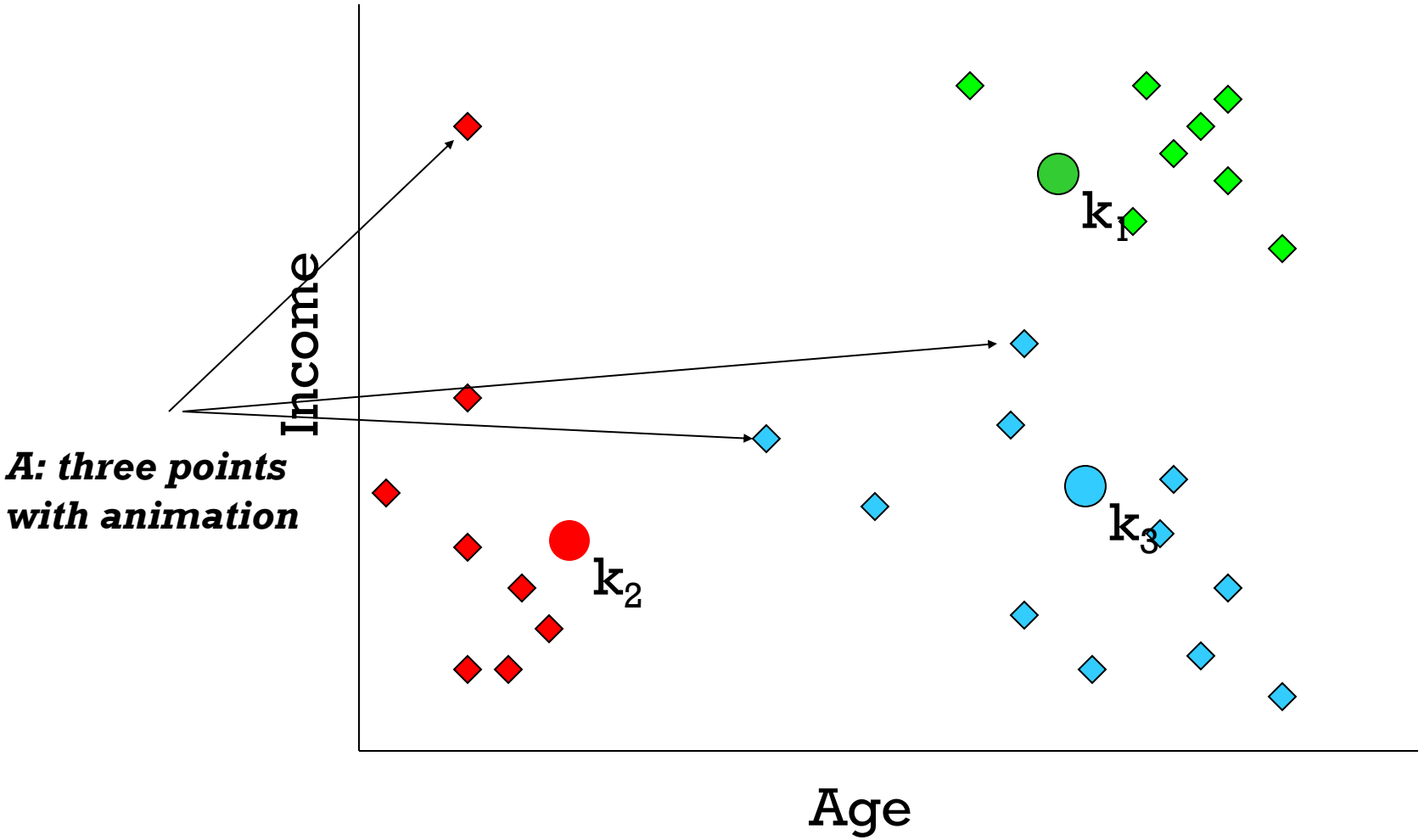
K-means: Step4

Reassign points closest to a different new cluster center

Q: Which points are reassigned?

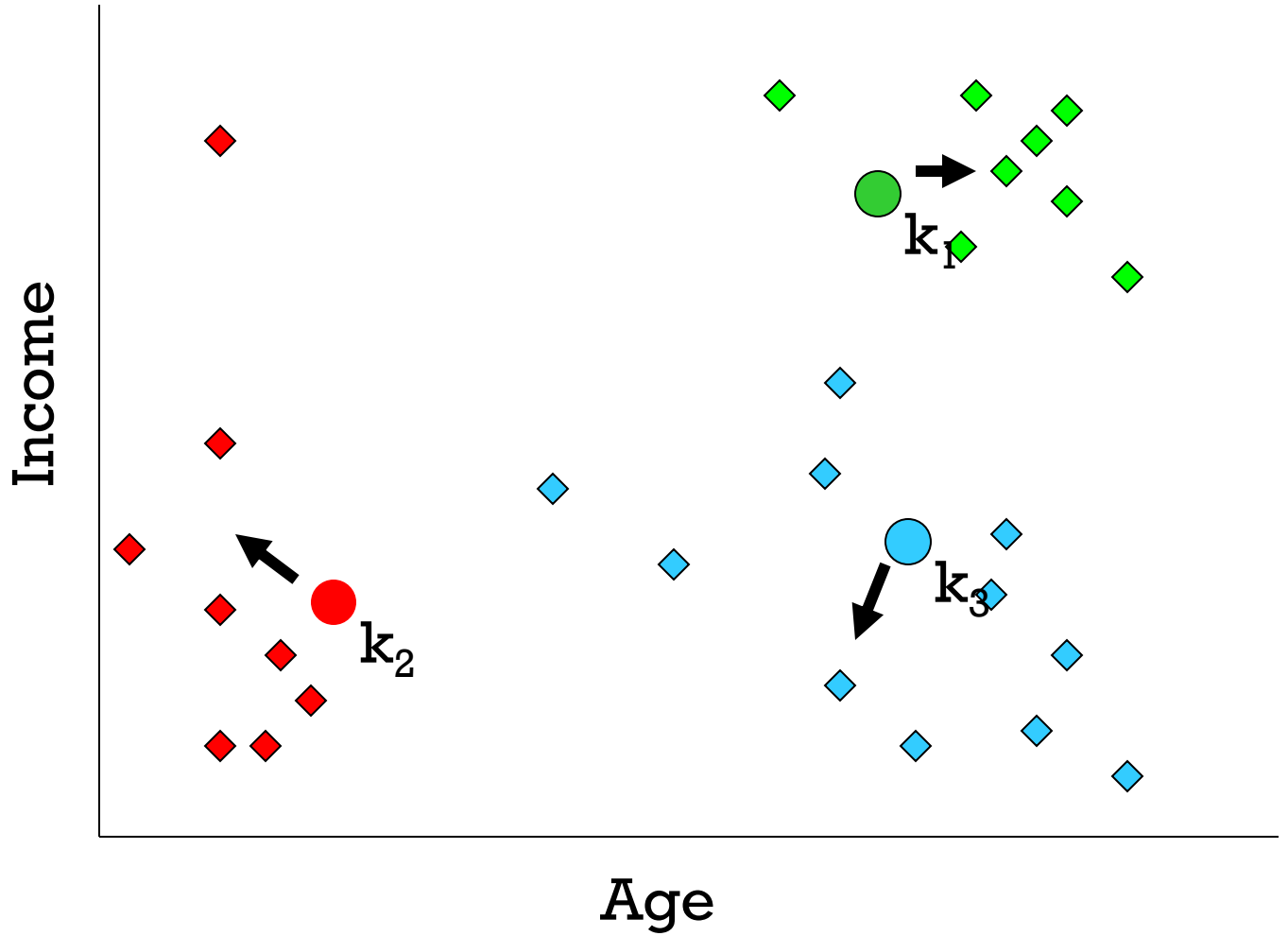


K-means: Step4(a)



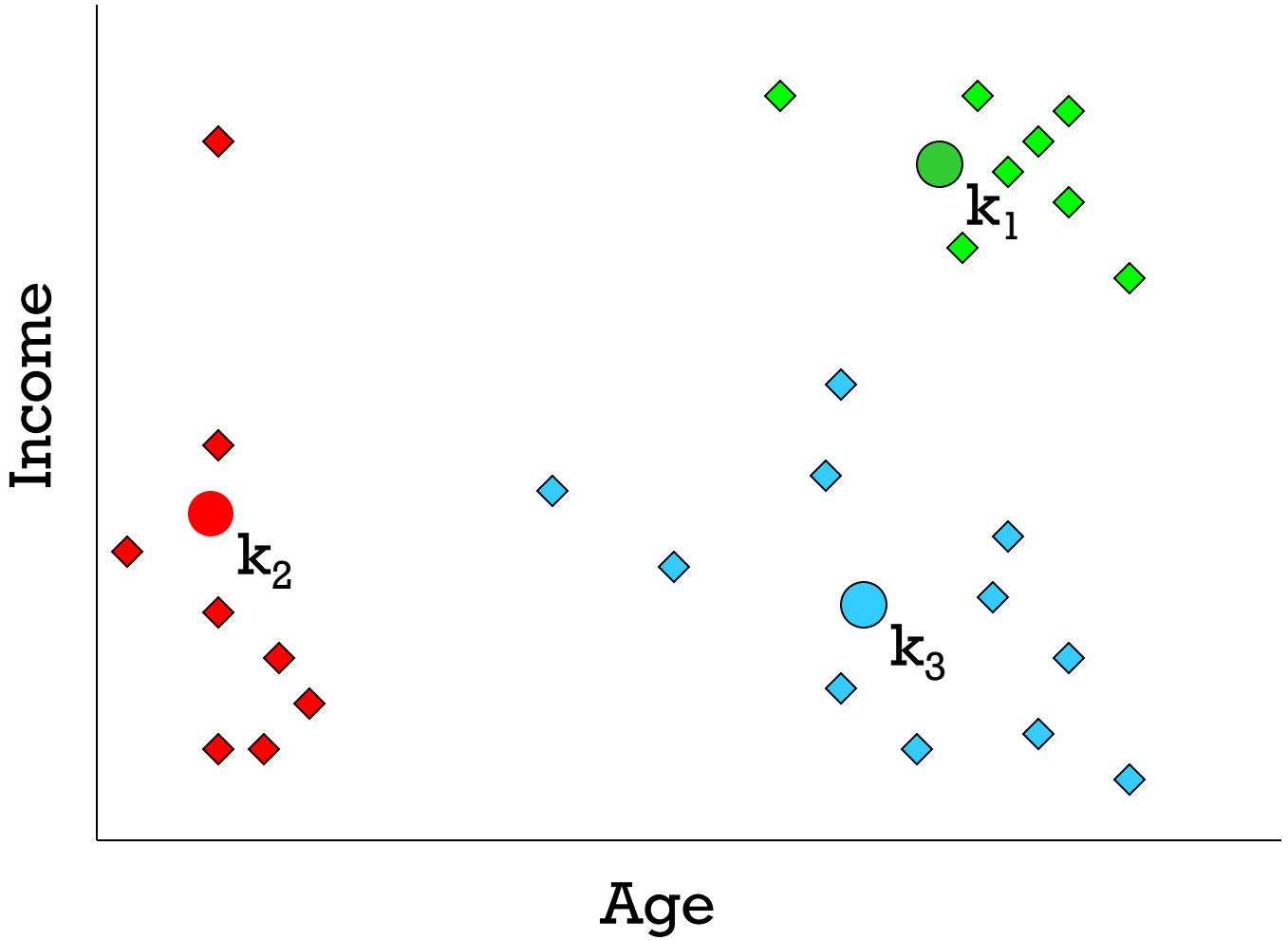
K-means: Step5

re-compute
cluster means



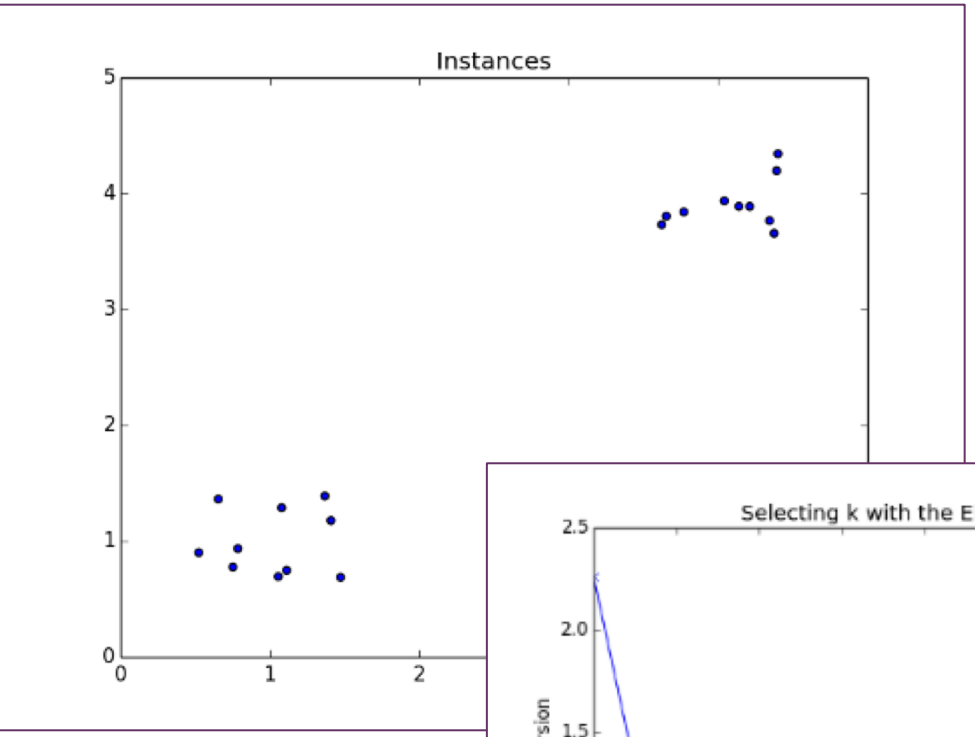
K-means: Step5(a)

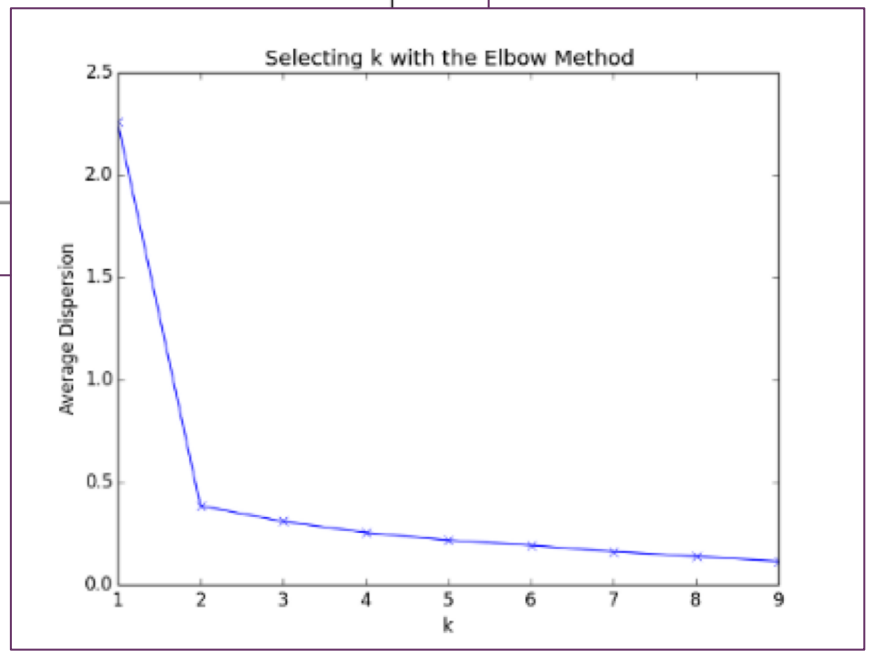
move cluster centers to cluster means

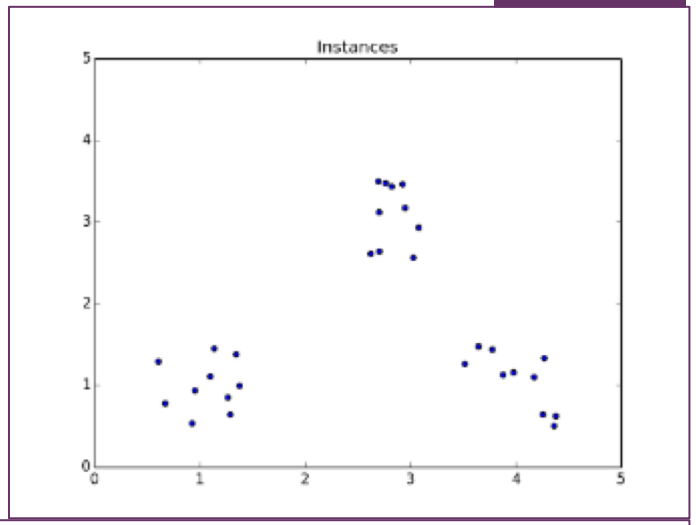


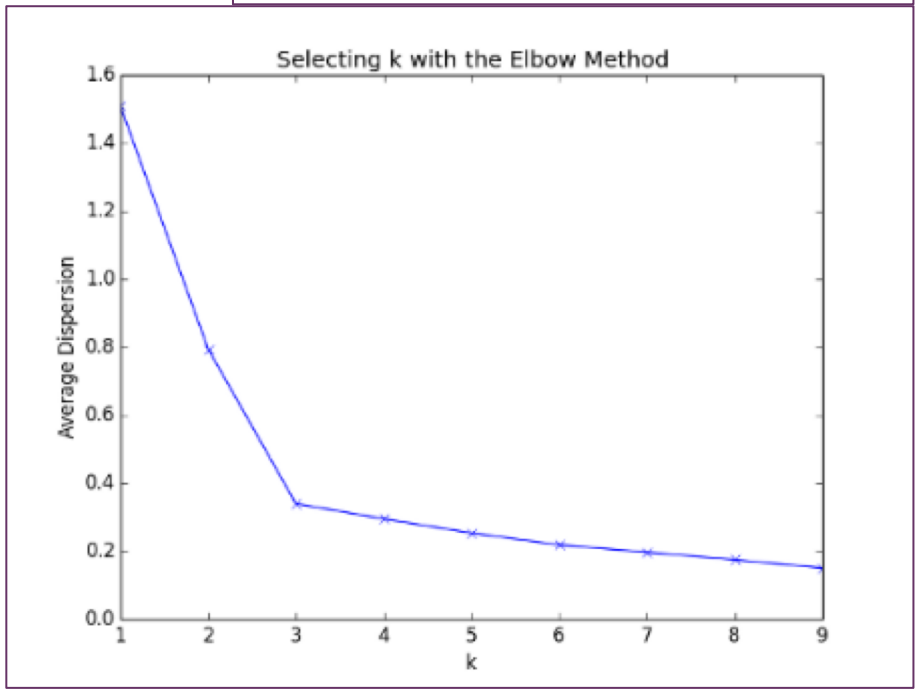
Determine the number of *k*: Elbow Method

https://www.packtpub.com/books/content/clustering-k-means









Market Basket Analysis

Rule	Support	Confidence
Apple => Donut	2/5	2/3
Coconut > Apple	2/5	2/4
Apple => Coconut	2/5	2/3
Banana & Coconut => Donut	1/5	1/3

Implication

	No	Yes	Pasta Total
No	500	3,500	**4,000**
Yes	1,000	5,000	**6,000**
Pizza Total	**1,500**	**8,500**	**10,000**

- Support(Pizza => Pasta) = 50%
- Confidence(Pizza => Pasta) = 83%
- Expected Confidence(Pizza => Pasta) = 85%
- **Lift(Pizza => Pasta) = 83%/85% < 1**

Association Rule Mining (cont.)

- Wal-Mart customers who purchase Barbie dolls have a 60% likelihood of also purchasing one of three types of candy bars [Forbes, Sept 8, 1997]

- Strategies?
  1. Put them closer together in the store.
  2. Put them far apart in the store.
  3. Package candy bars with the dolls.
  4. Package Barbie + candy + poorly selling item.
  5. Raise the price on one, and lower it on the other.
  6. Offer Barbie accessories for proofs of purchase.
  7. Do not advertise candy and Barbie together.
  8. Offer candies in the shape of a Barbie doll.

Caution in Association Rule Mining

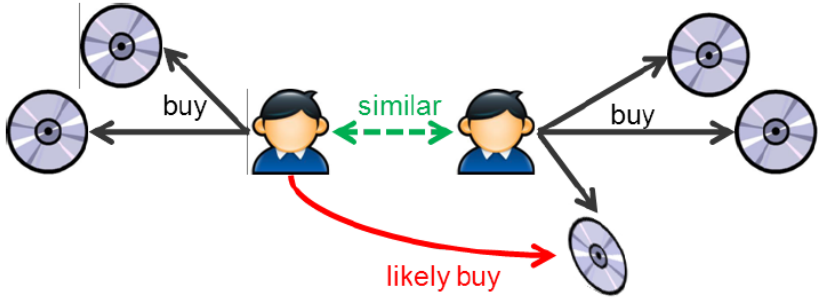
1. Basket size: per bill, customer, day
2. Item level: SKU, product category

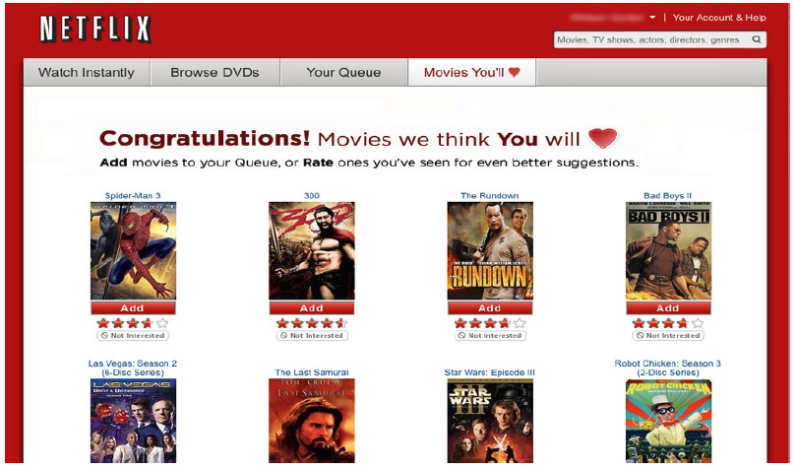
Association Rule Mining (cont.)

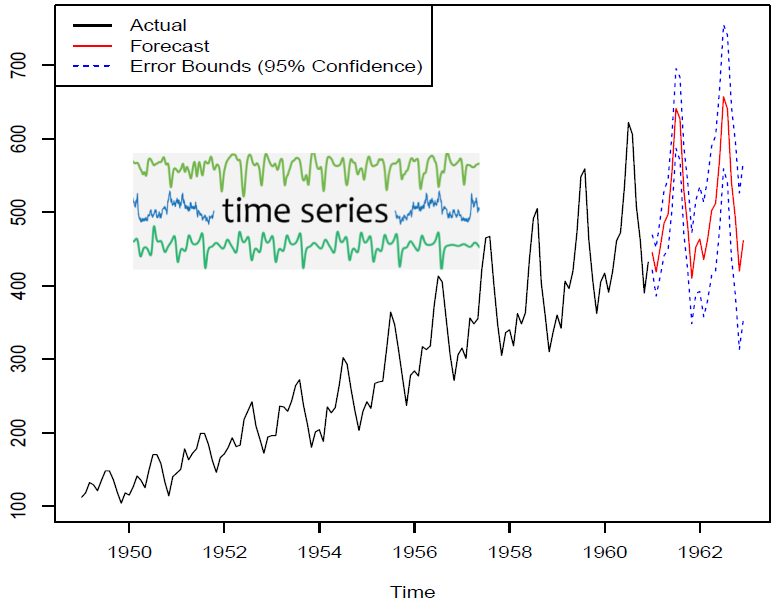
- Wal-Mart customers who purchase Barbie dolls have a 60% likelihood of also purchasing one of three types of candy bars [Forbes, Sept 8, 1997]

- Strategies?
  1. Put them closer together in the store.
  2. Put them far apart in the store.
  3. Package candy bars with the dolls.
  4. Package Barbie + candy + poorly selling item.
  5. Raise the price on one, and lower it on the other.
  6. Offer Barbie accessories for proofs of purchase.
  7. Do not advertise candy and Barbie together.
  8. Offer candies in the shape of a Barbie doll.

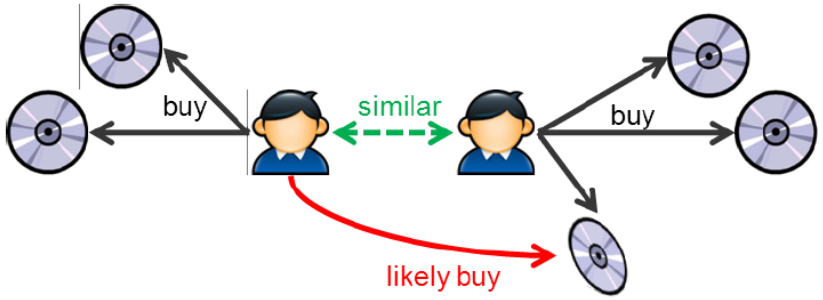
Task3: Special task

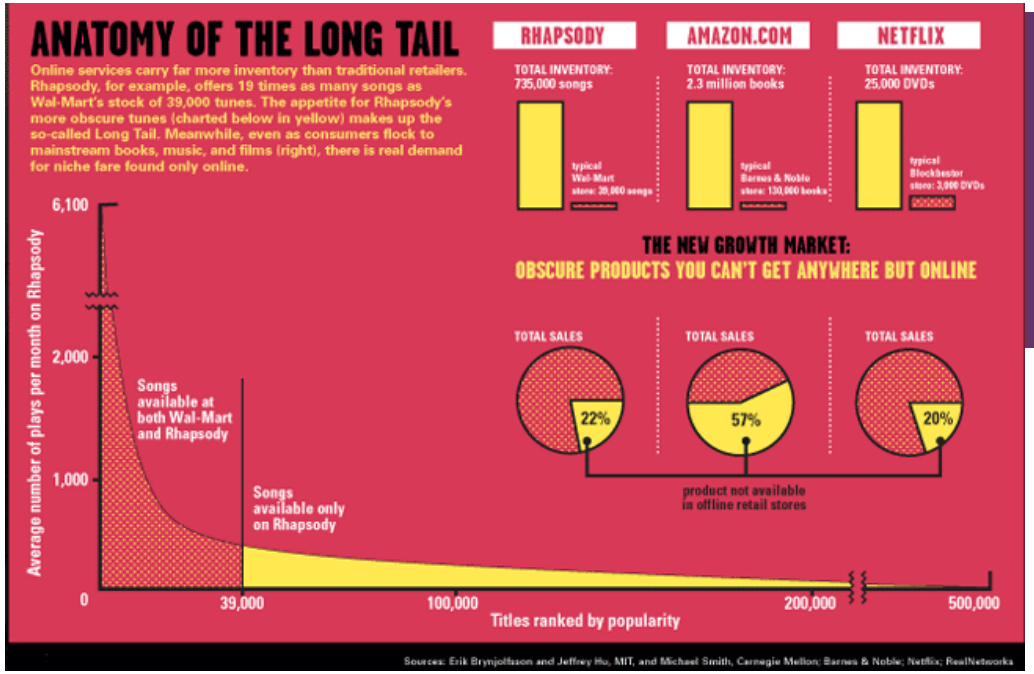


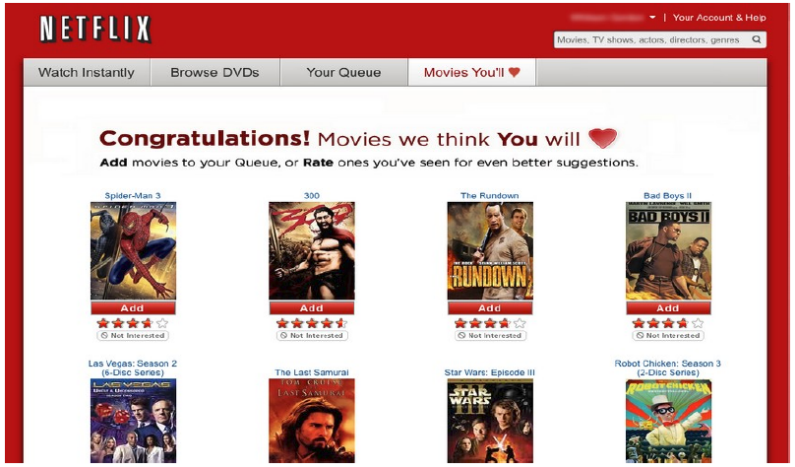




Recommendation system



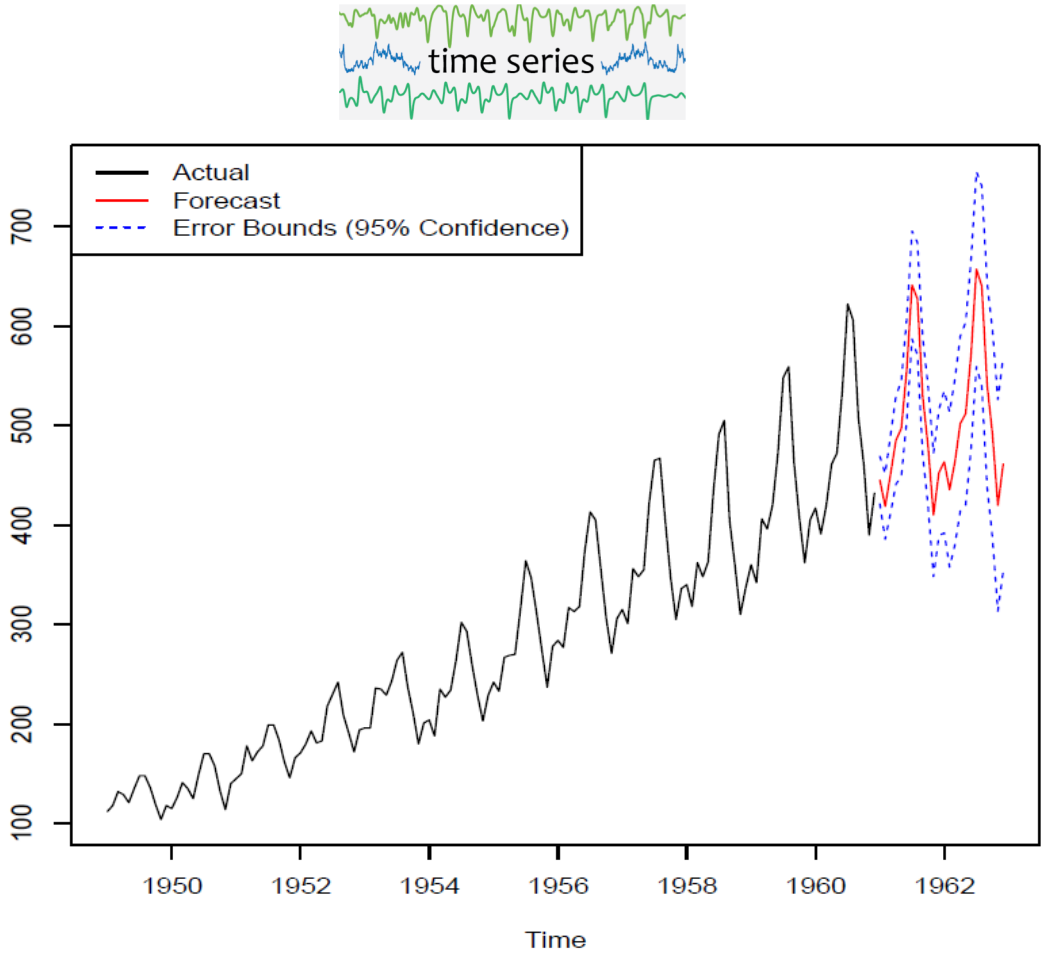




	Harry potter	X-Men	Hobbit	Argo	Pirates
101	5	2	4	?	?
102	?	?	5	2	?
103	1	2	?	?	3
104					
105					
…					

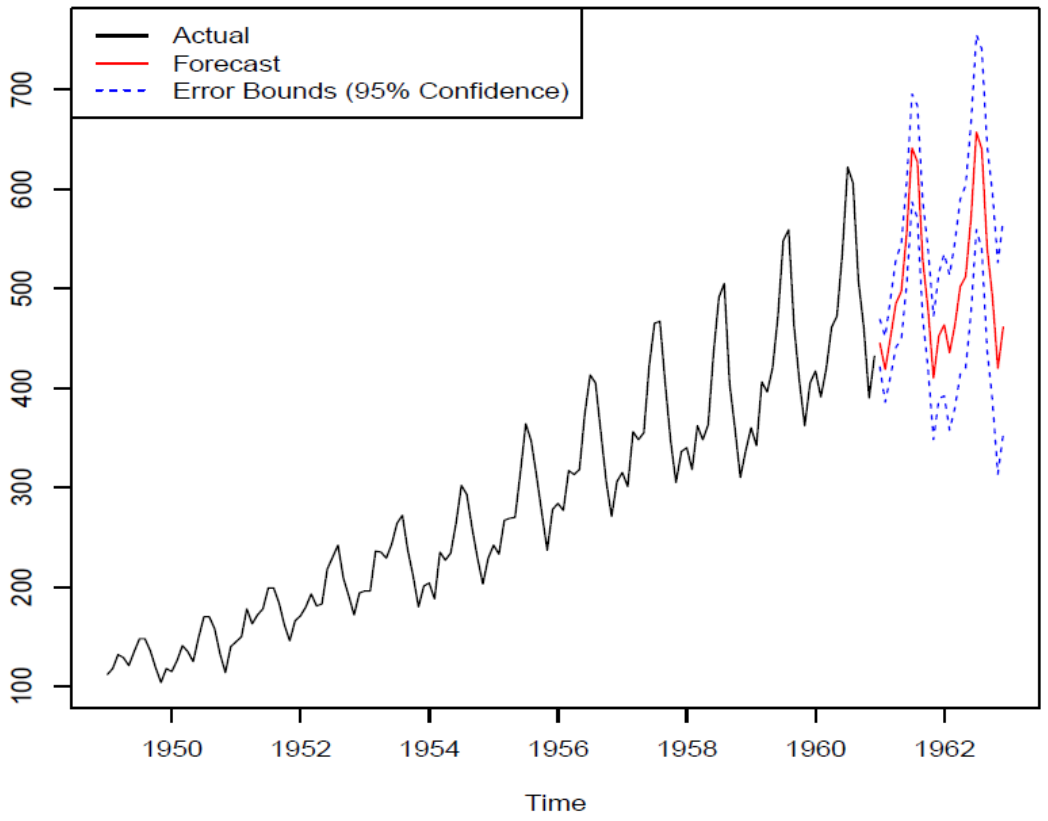
	Harry potte	X-Men	Hobbit	Argo	Pirates
101	5	2	4	1	3
102	4	1	5	2	3
103	1	2	4	1	3
104					
105					
…					

Time Series Analysis (Trend Forecasting)



- Techniques
  - **ARIMA (Autoregressive integrated moving average)**
  - Exponential Smoothing
  - Neural Networks
  - Deep Learning
- Sample Applications
  - Customer trend forecasting
  - Revenue trend forecasting
  - Rainfall forecasting
  - Remaining useful life forecasting (preventive maintenance)

Time Series Analysis (Trend Forecasting)



- Techniques
  - **ARIMA (Autoregressive integrated moving average)**
  - Exponential Smoothing
  - Neural Networks
  - Deep Learning
- Sample Applications
  - Customer trend forecasting
  - Revenue trend forecasting
  - Rainfall forecasting
  - Remaining useful life forecasting (preventive maintenance)

Text Mining

https://ischool.syr.edu/infospace/2013/04/23/what-is-text-mining/

- Text mining, which is sometimes referred to "text analytics" is one way to make qualitative or "unstructured" data **usable by a computer**.
- Convert from unstructured to structured data

Parse → Stop → Stem

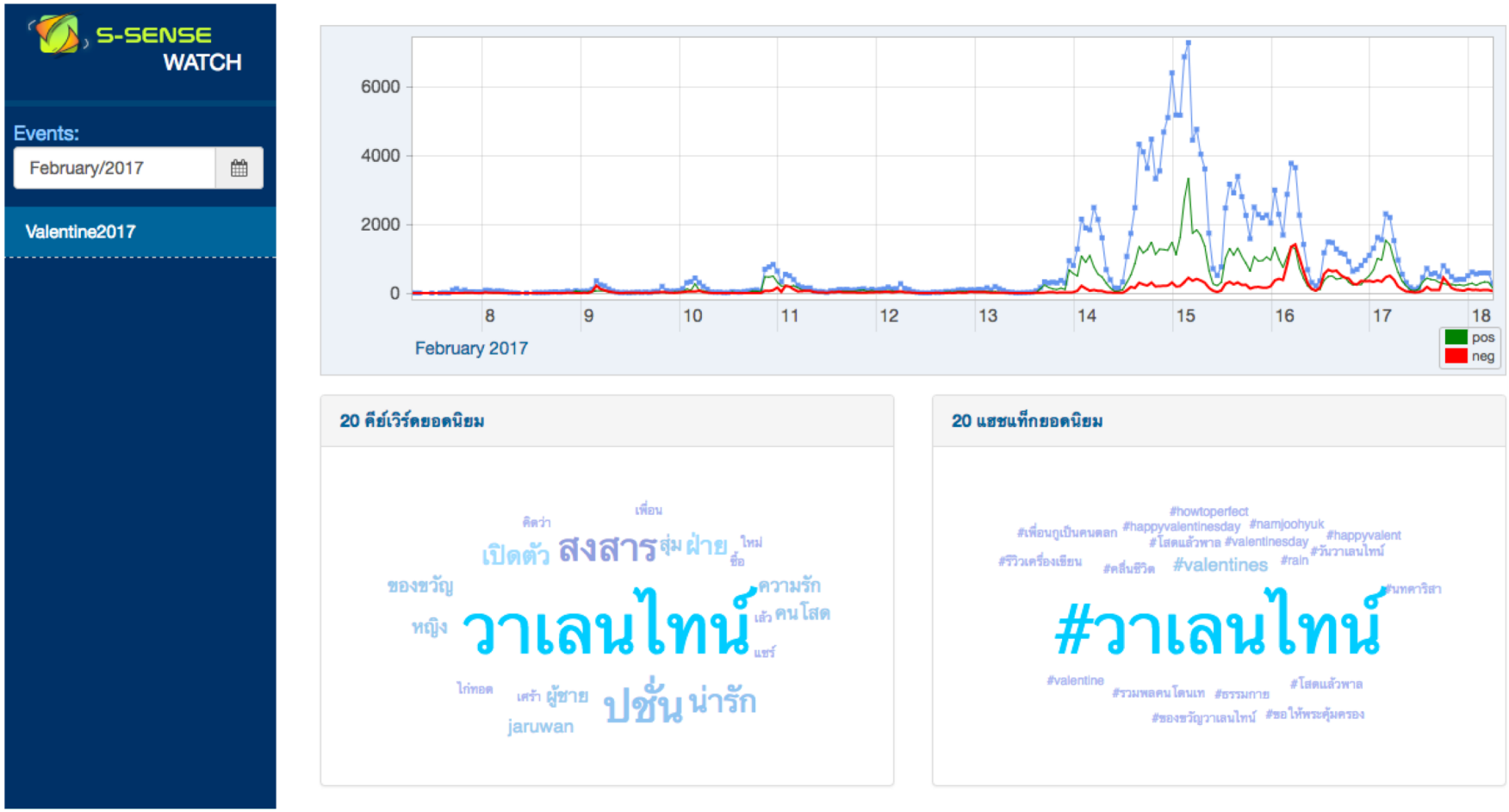
Comments	Good	Like	Hate	Sentiment
Tweet1	7	8	0	😁
Tweet2	1	0	10	😡
Tweet3	2	9	1	😁

Text Mining (cont.): Sentiment Analysis

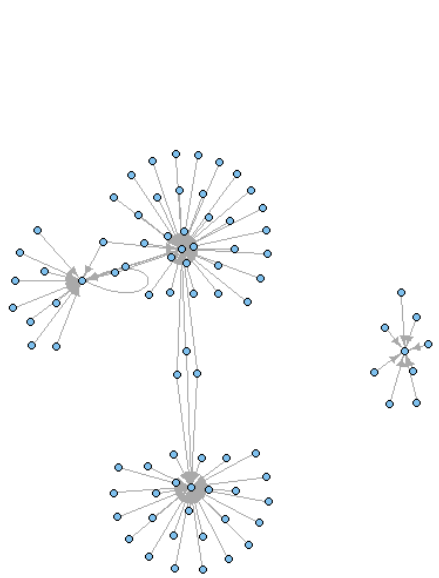
http://pop.ssense.in.th/

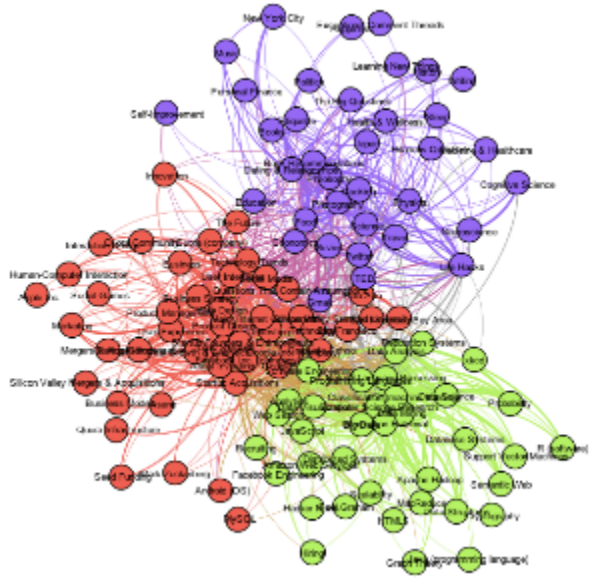
Text Mining (cont.): Emerging Trend Analysis

http://www.ssense.in.th/watch/



Social Network Analysis





- Techniques
  - Centrality: degree, closeness, betweenness, transitivity
  - Community detection
  - Graph Clustering
- Sample Applications
  - Influencer detection
  - Community detection

Scikit-learn:
Machine learning library in Python

Scikit-learn: Machine learning library in Python

- Provides many machine learning tools with a common Estimator interface
- Built in helpers for common ML tasks (e.g., metrics, preprocessing)
- Easily combine algorithms to make a complex pipeline
- Relies heavily on numpy and scipy, often used with pandas

How do you pronounce the project name?

sy-kit learn. sci stands for science!

Why scikit?

There are multiple scikits, which are scientific toolboxes built around SciPy. You can find a list at https://scikits.appspot.com/scikits. Apart from scikit-learn, another popular one is scikit-image.

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, … — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, … — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, … — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction. — Examples

http://scikit-learn.org/stable/index.html

Estimator Interface

Decision Trees

We'll start just by training a single decision tree.

```
In [8]: from sklearn.tree import DecisionTreeClassifier

In [9]: dtree = DecisionTreeClassifier(min_samples_leaf=10, criterion='entropy')

In [10]: dtree.fit(X_train,y_train)

Out[10]: DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=None,
                                max_features=None, max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=10, min_samples_split=2,
                                min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                                splitter='best')
```

Prediction and Evaluation

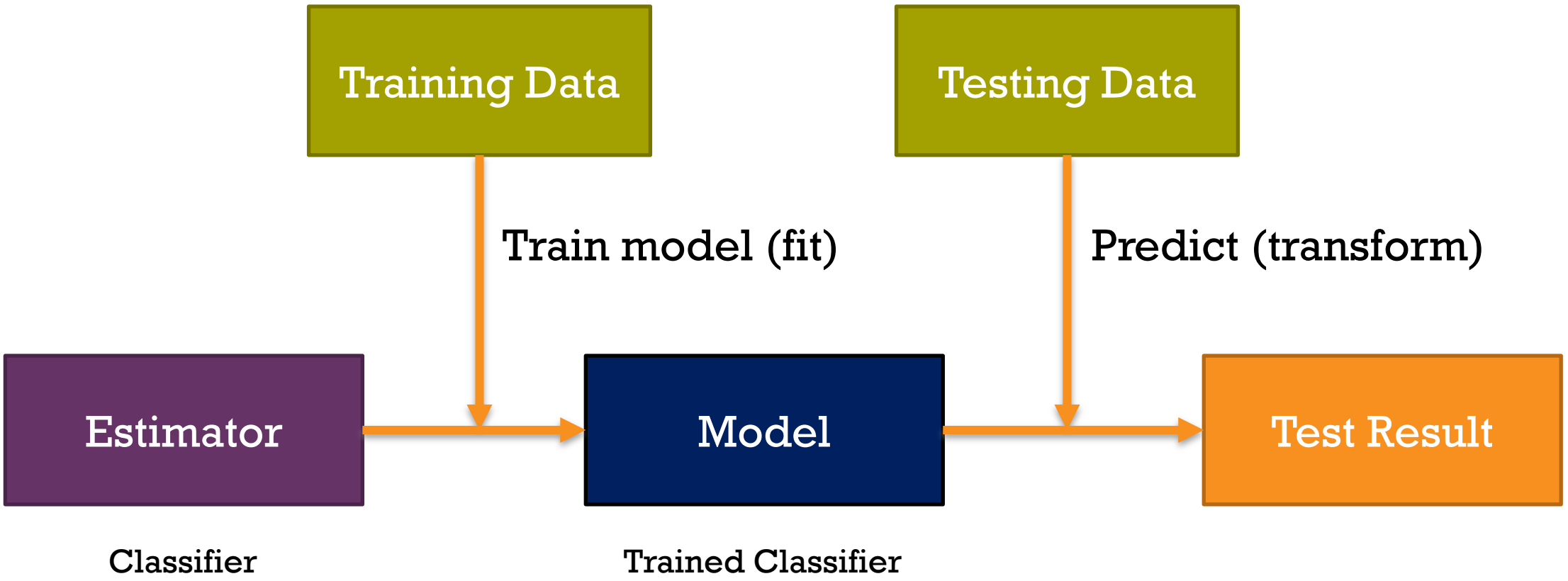
Let's evaluate our decision tree.

```
[11]: predictions = dtree.predict(X_test)

[12]: from sklearn.metrics import classification_report,confusion_matrix

[13]: print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
absent	0.85	0.85	0.85	20
present	0.40	0.40	0.40	5
avg / total	0.76	0.76	0.76	25



Example: Learning to Predict Breast Cancer

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split

cancer = load_breast_cancer()      # Get some data
X_train, X_test, y_train, y_test = train_test_split(
        cancer.data, cancer.target,
        stratify=cancer.target, random_state=1337)

tree = DecisionTreeClassifier(random_state=7331)
tree.fit(X_train, y_train)  # Learn a Decision Function
```

Ref: C. Travis Johnson, Mines Linux Users Group on April 27, 2017

Example (cont.): Evaluating Accuracy of a Model

```
# How well did we do?
train_acc = tree.score(X_train, y_train)
test_acc = tree.score(X_test, y_test)
print("Training Accuracy: {:.3f}".format(train_acc))
print("Testing Accuracy: {:.3f}".format(test_acc))
# Training Accuracy: 1.000
# Testing Accuracy: 0.923
```

Ref: C. Travis Johnson, Mines Linux Users Group on April 27, 2017

Example (cont.): Pipeline

```
from sklearn.pipeline import make_pipeline
from sklearn.svm import SVC
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import GridSearchCV
pipe = make_pipeline(PCA(), StandardScaler(), SVC())
params = dict(pca__n_components=[2, 5, 10],
              svc__C=[0.1, 10, 100])
grid = GridSearchCV(pipe, param_grid=params)
# Next, call grid.fit on some training data
# This will use cross validation to estimation performance using each
# combination of parameters for pipeline in params dict

# With fitted model
print(grid.best_params_)
```

Ref: C. Travis Johnson, Mines Linux Users Group on April 27, 2017