

DATA MINING

Outline

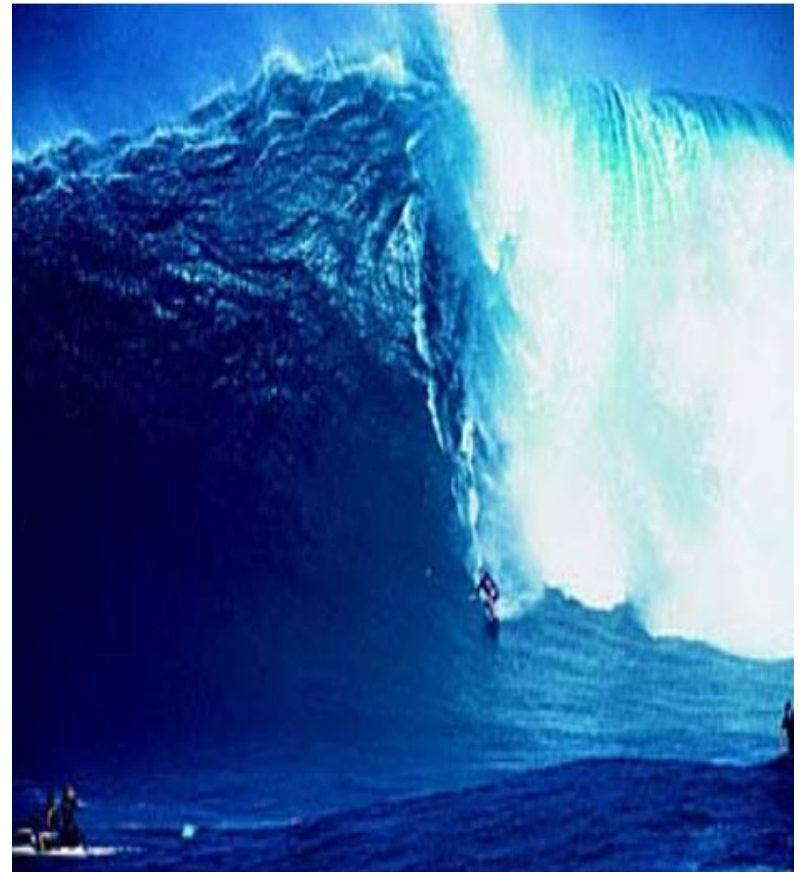
2

- What is Data Mining?
- The Data Mining Process
- Tasks and Applications
- Data Mining Software

What is Data Mining?

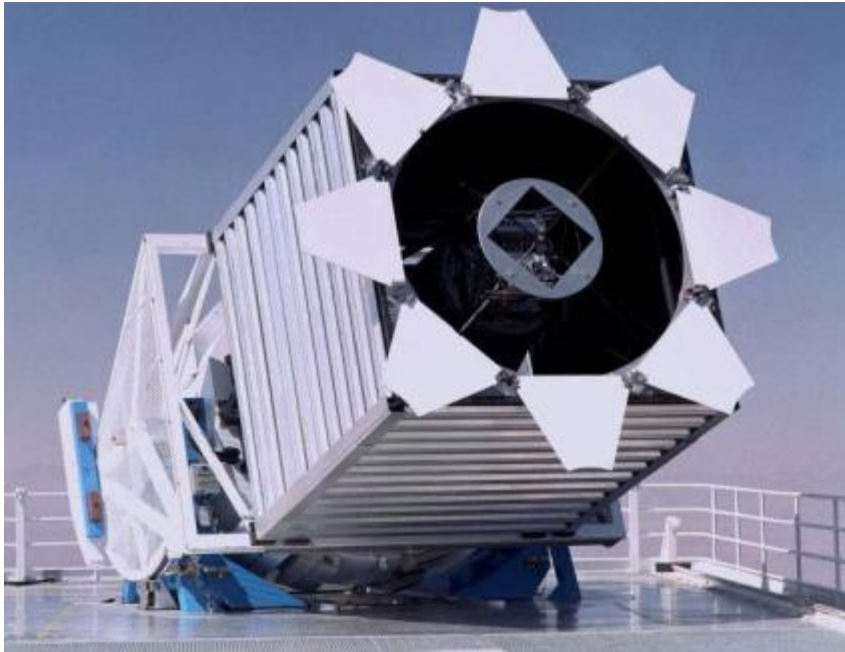
3

- **Large quantities** of data are collected about all aspects of our lives
- This data contains **interesting patterns**
- Data Mining helps us to
 1. **discover these patterns** and
 2. **use them for decision making** across all areas of society, including
 - Business and industry
 - Science and engineering
 - Medicine and biotech
 - Government
 - Individuals



“We are Drowning in Data...”

4



Sloan Digital Sky Survey

$\approx 200 \text{ GB/day}$

$\approx 73 \text{ TB/year}$

Predict

- Type of sky object:
Star or galaxy?

“We are Drowning in Data...”

5



US Library of Congress

≈ 235 TB archived

≈ 40 Wikipedias

Discover

- Topic distributions
- Historic trends
- Citation networks

“We are Drowning in Data...”

6



Facebook

≈ 10 TB/day added

\approx over 300 Petabyte in

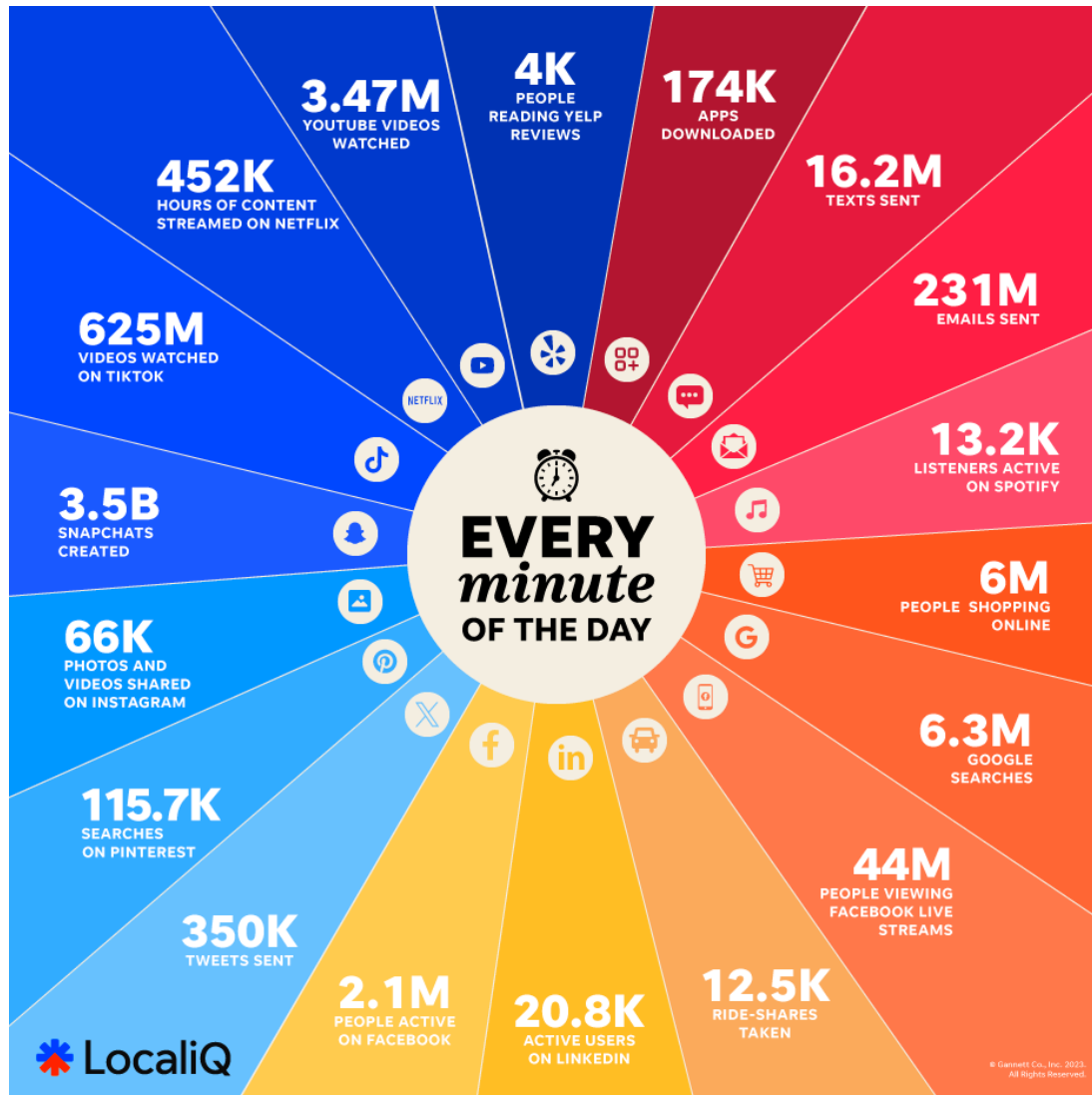
Facebook's data warehouse

Predict

- Interests and behavior of over one billion people

“We are Drowning in Data...”

7



Predict

- Interests and behavior of mankind

“We are Drowning in Data...”

8

Law enforcement agencies collect unknown amounts of data from various sources

- Cell phone calls
- Location data
- Web browsing behavior
- Credit card transactions
- Online profiles (Facebook)
- ...

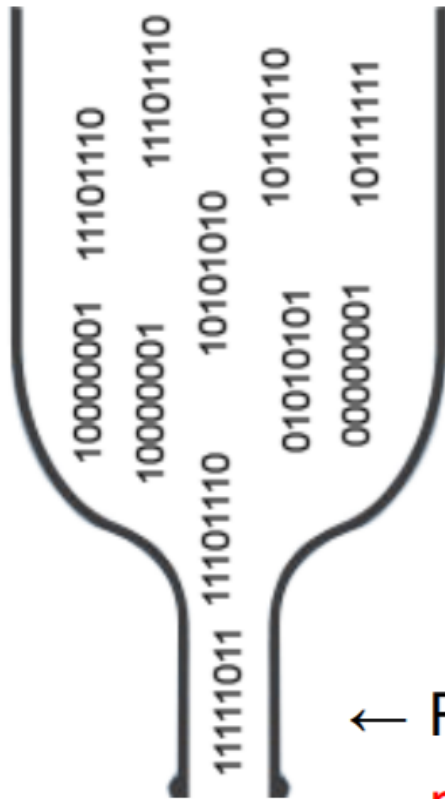
Predict

- Terrorist or not?
- Trustworthiness



“...but starving for knowledge!”

9



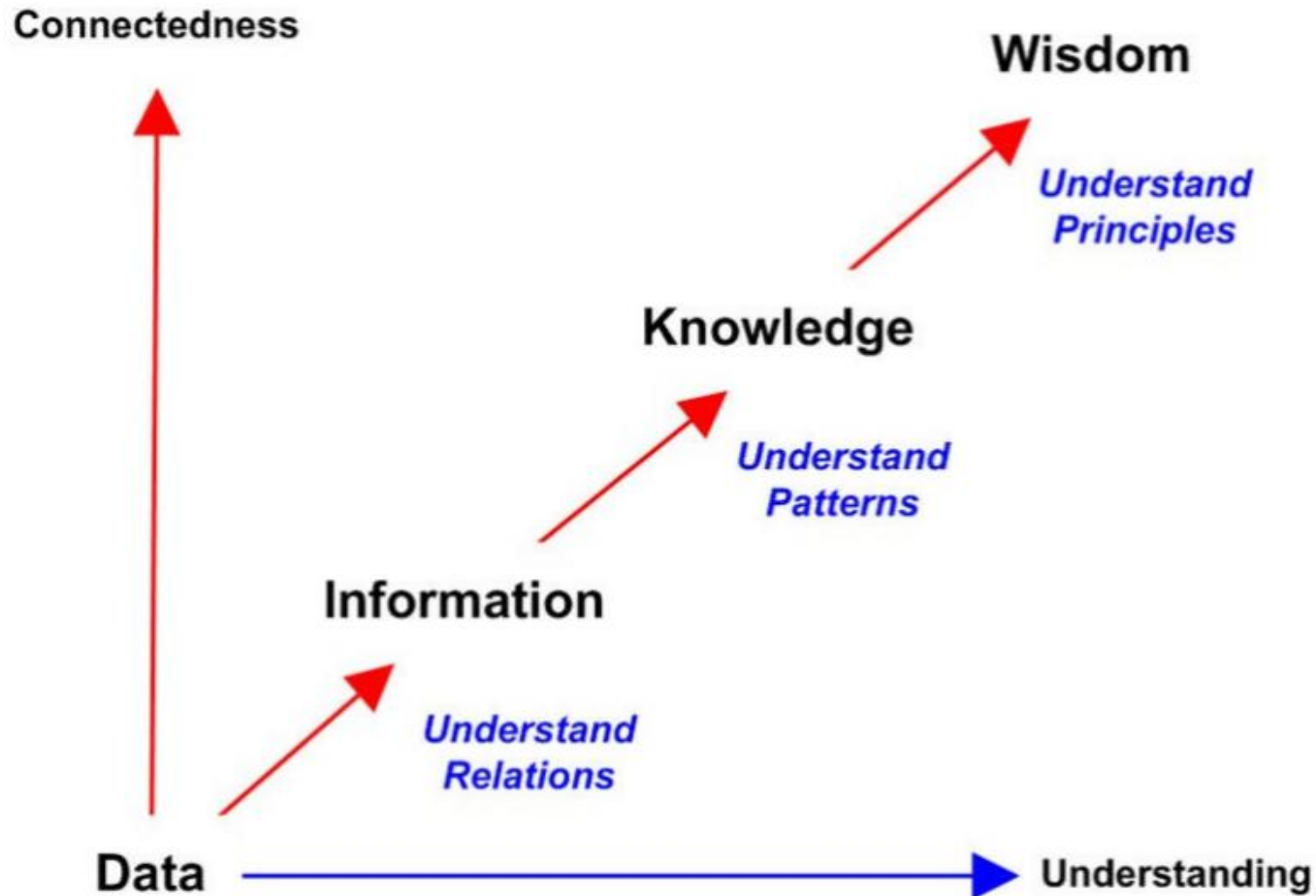
← Rate at which data are produced

← Rate at which data can be understood

manual interpretation is hardly feasible!

Data, Information, Knowledge, and Wisdom

10



Gene Bellinger, Durval Castro and Anthony Mills. "Transforming Data to Wisdom."

Definitions of Data Mining

11

“Exploration & analysis, of large quantities of data in order to discover meaningful patterns.”

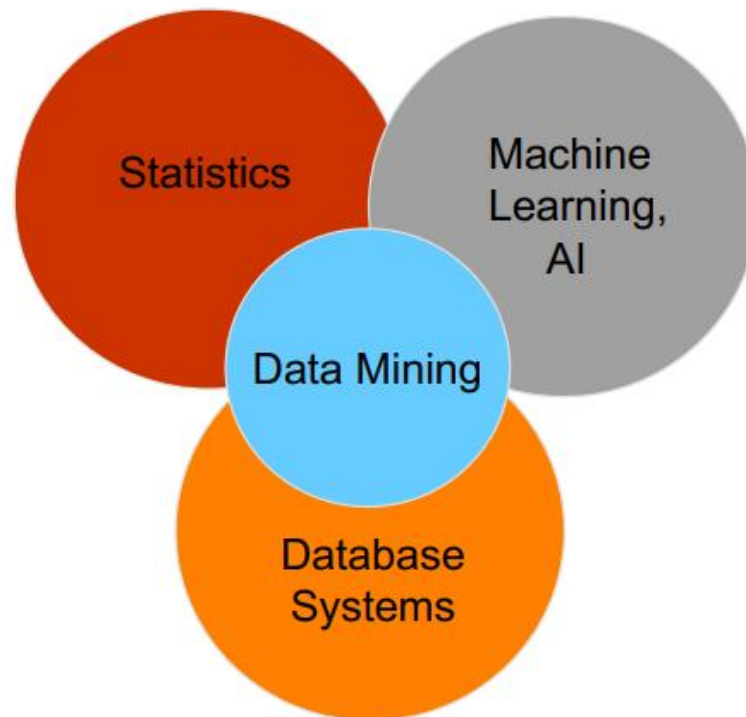
“Data mining is nothing else than torturing the data until it confesses ...and if you torture it enough, you can get it to confess to anything.”

(Fred Menger, year unknown)

Origins of Data Mining

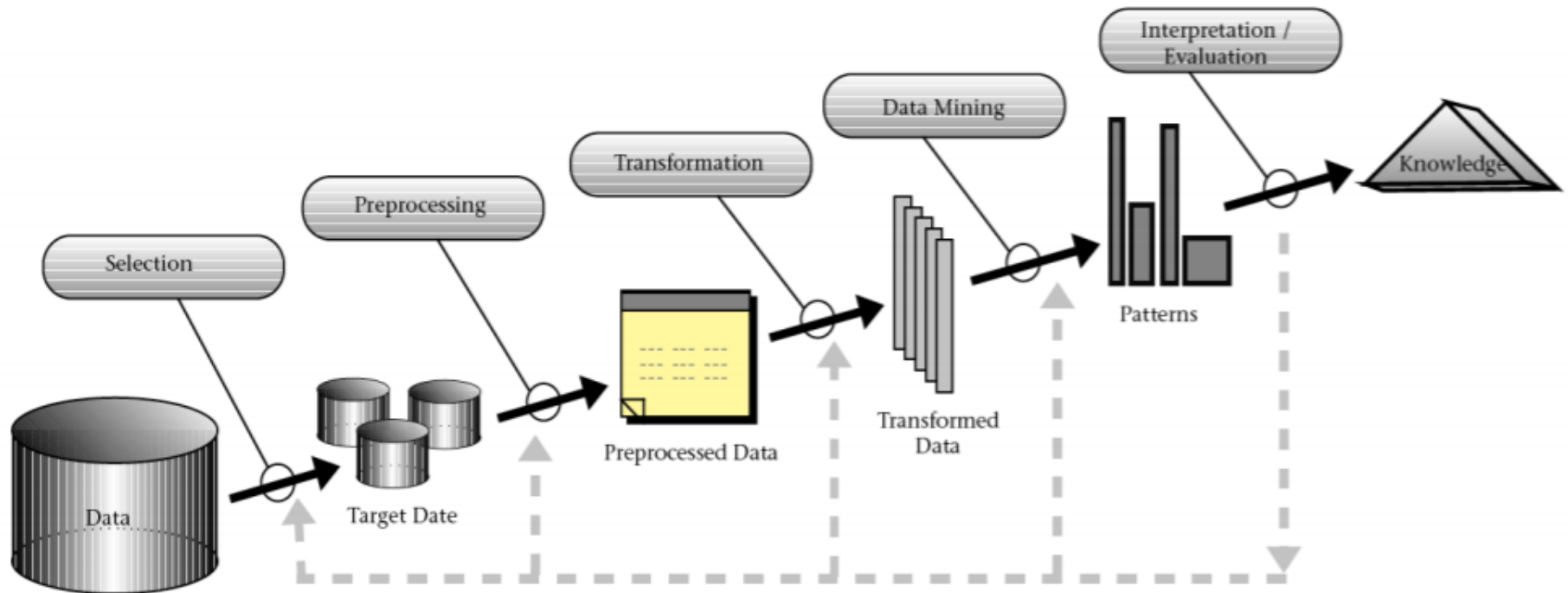
12

Data Mining combines ideas from statistics, machine learning, Artificial intelligence, and database systems



The Data Mining Process

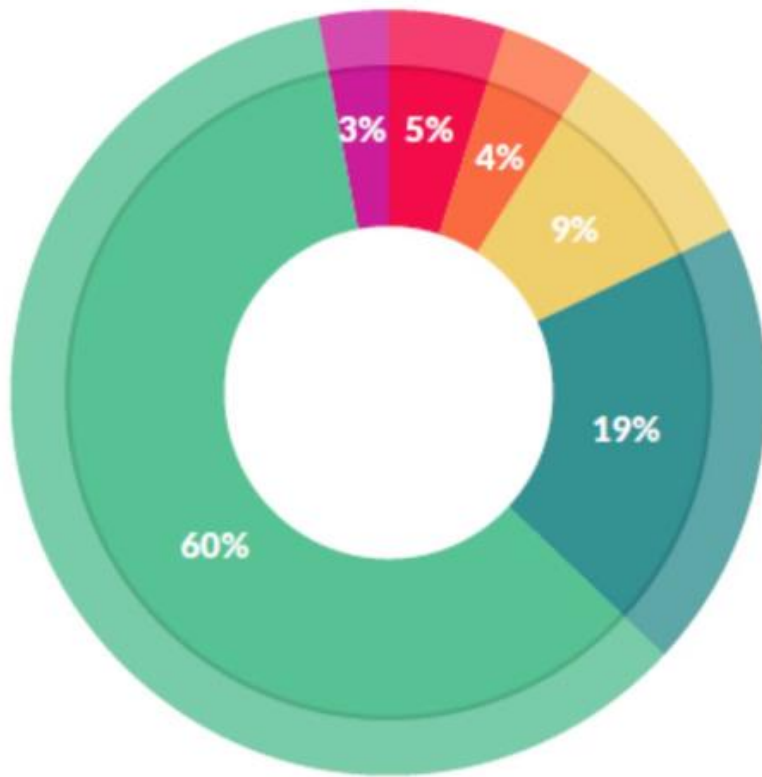
13



Source: Fayyad et al. (1996)

How Do Data Scientists Spend Their Days?

14



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Tasks and Applications

15

Descriptive Tasks

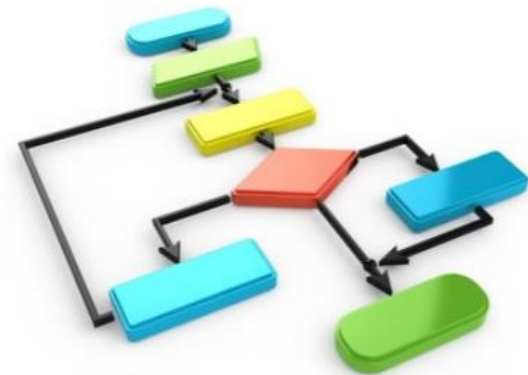
- Goal: Find patterns in the data.
- Example: Which products are often bought together?

Predictive Tasks

- Goal: Predict unknown values of a variable
 - given observations (e.g., from the past)
- Example: Will a person click a online advertisement?
 - given her browsing history

Machine Learning Terminology

- descriptive = unsupervised
- predictive = supervised



Data Mining Tasks

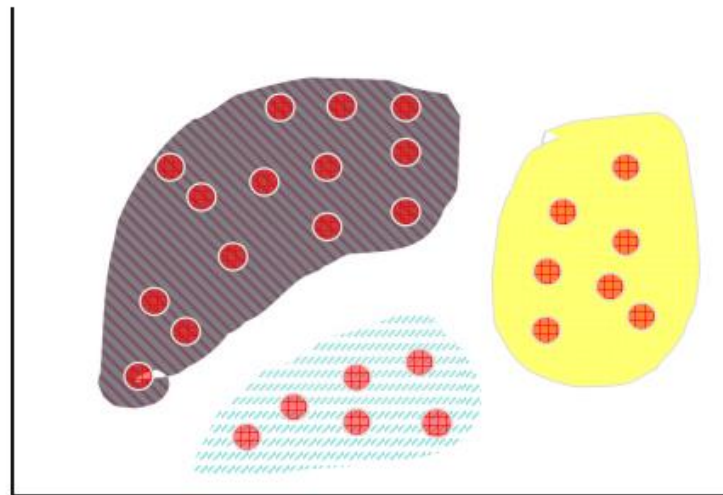
16

- Clustering (descriptive)
- Classification (predictive)
- Regression (predictive)
- Association Rule Mining (descriptive)

Clustering

17

- Given a set of data points, and a similarity measure among them, find clusters such that
 - Data points in one cluster are similar to one another
 - Data points in separate clusters are different from each other
- Result
 - a descriptive grouping of data points



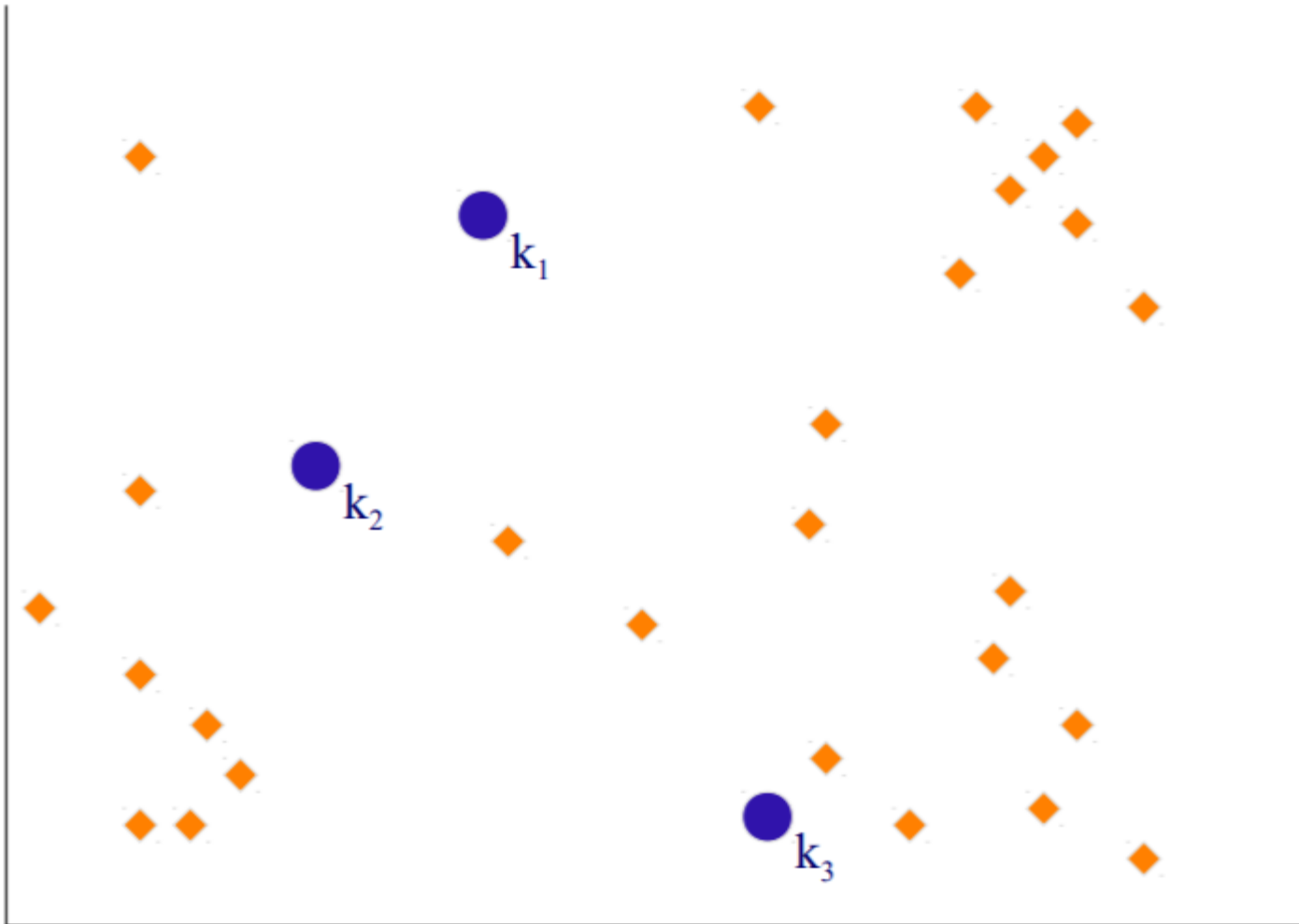
K-Means Clustering

18

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K , must be specified manually

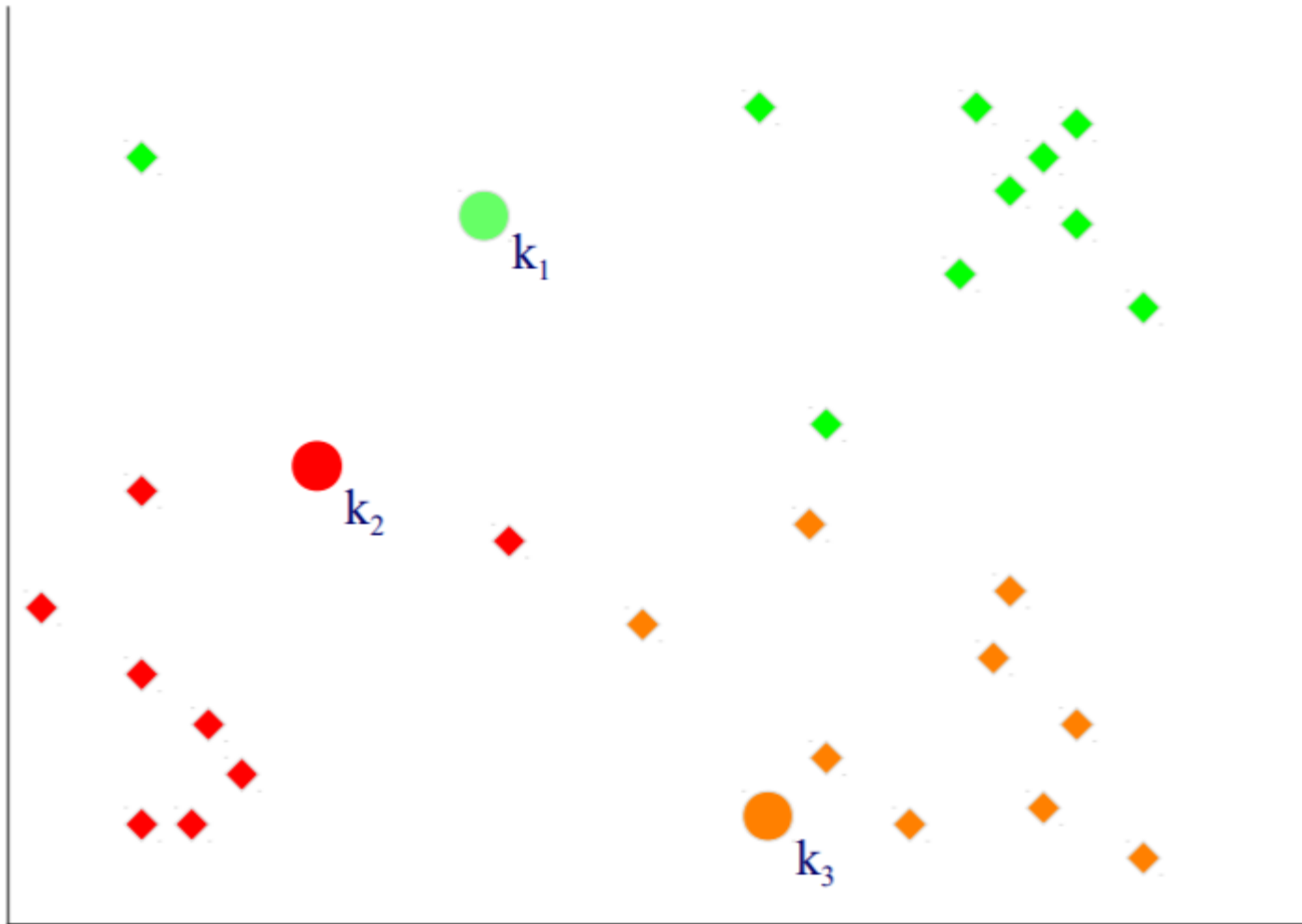
K-Means Example, Step 1

19



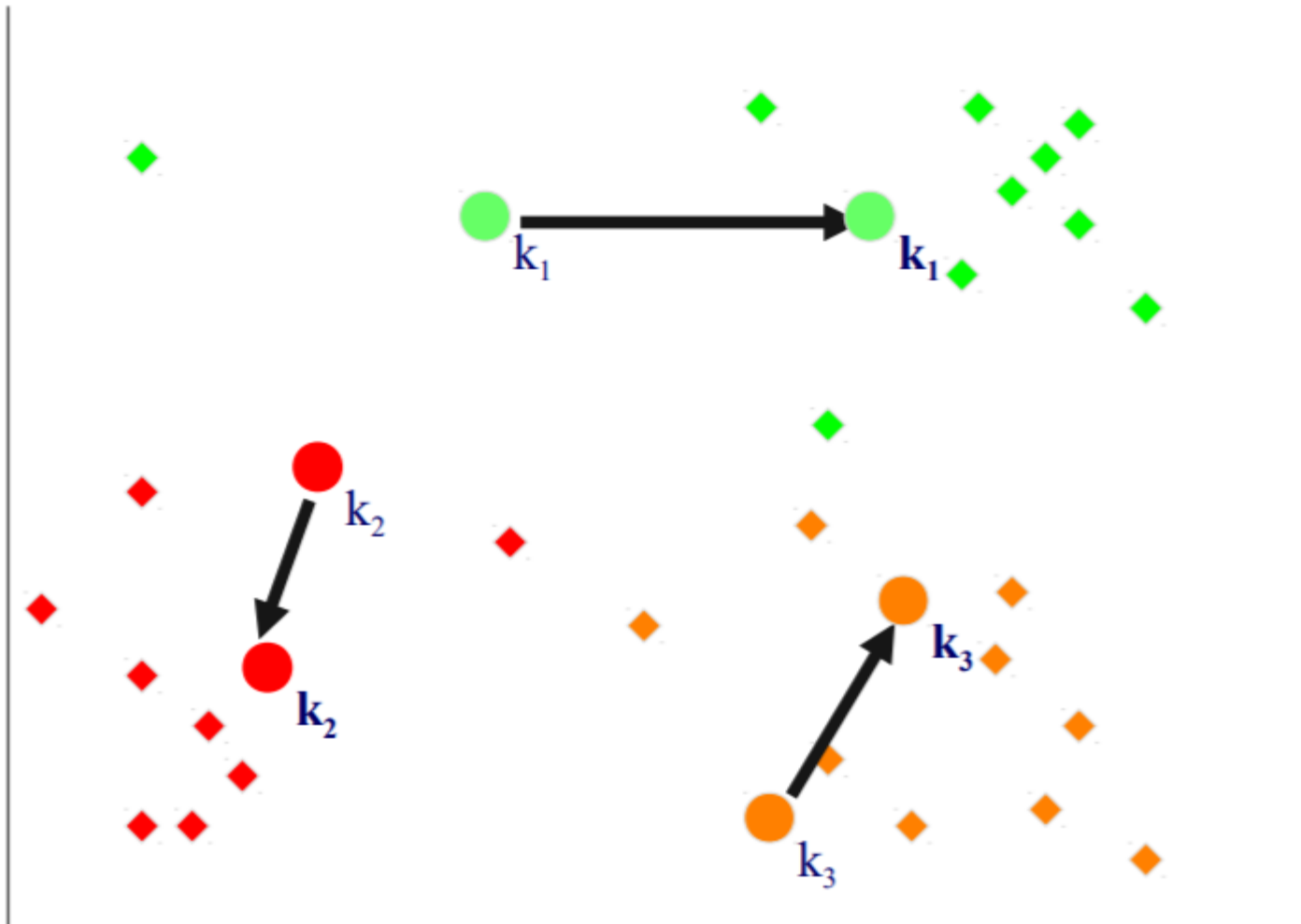
K-Means Example, Step 2

20



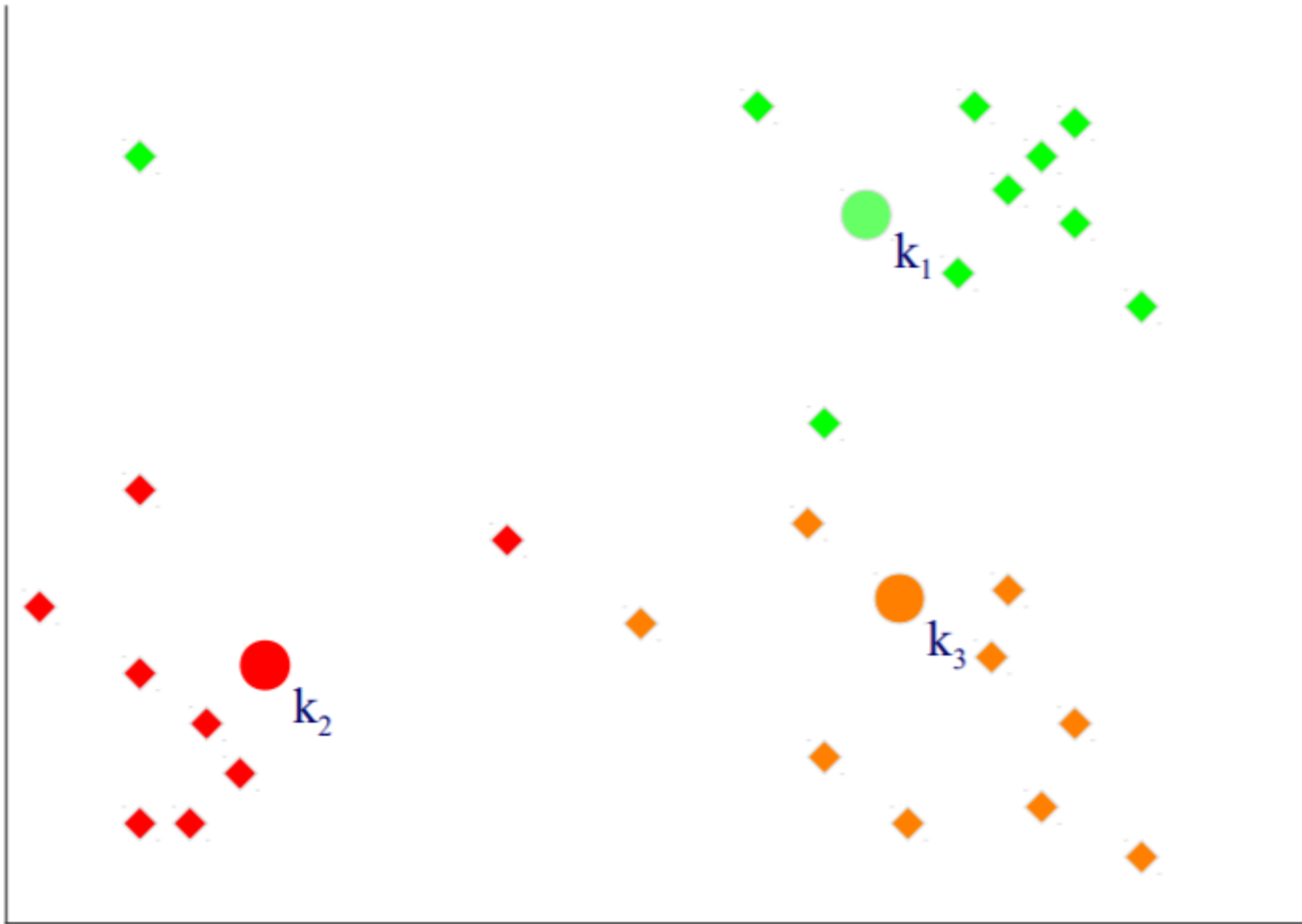
K-Means Example, Step 3

21



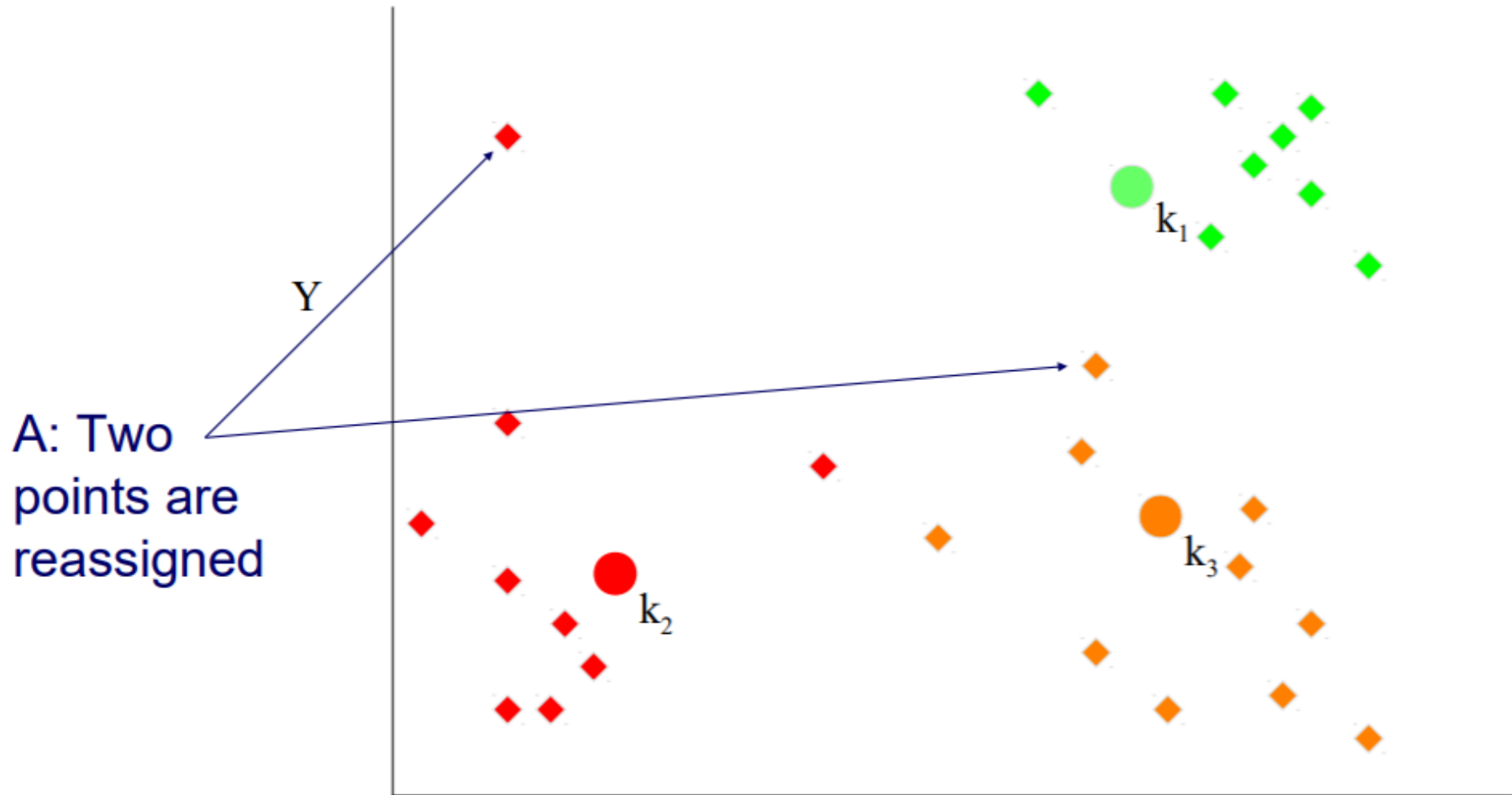
K-Means Example, Step 4

22



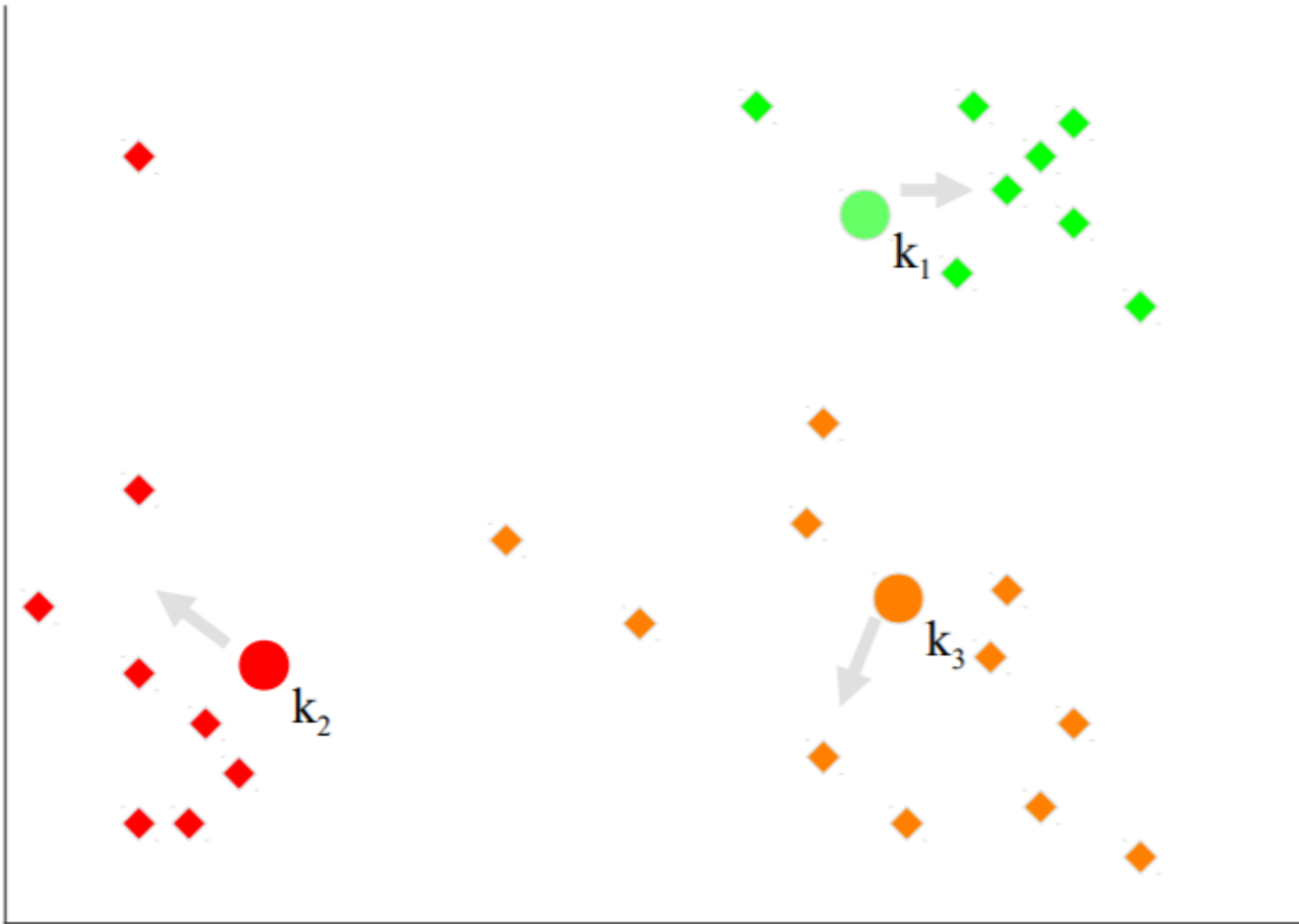
K-Means Example, Step 5

23



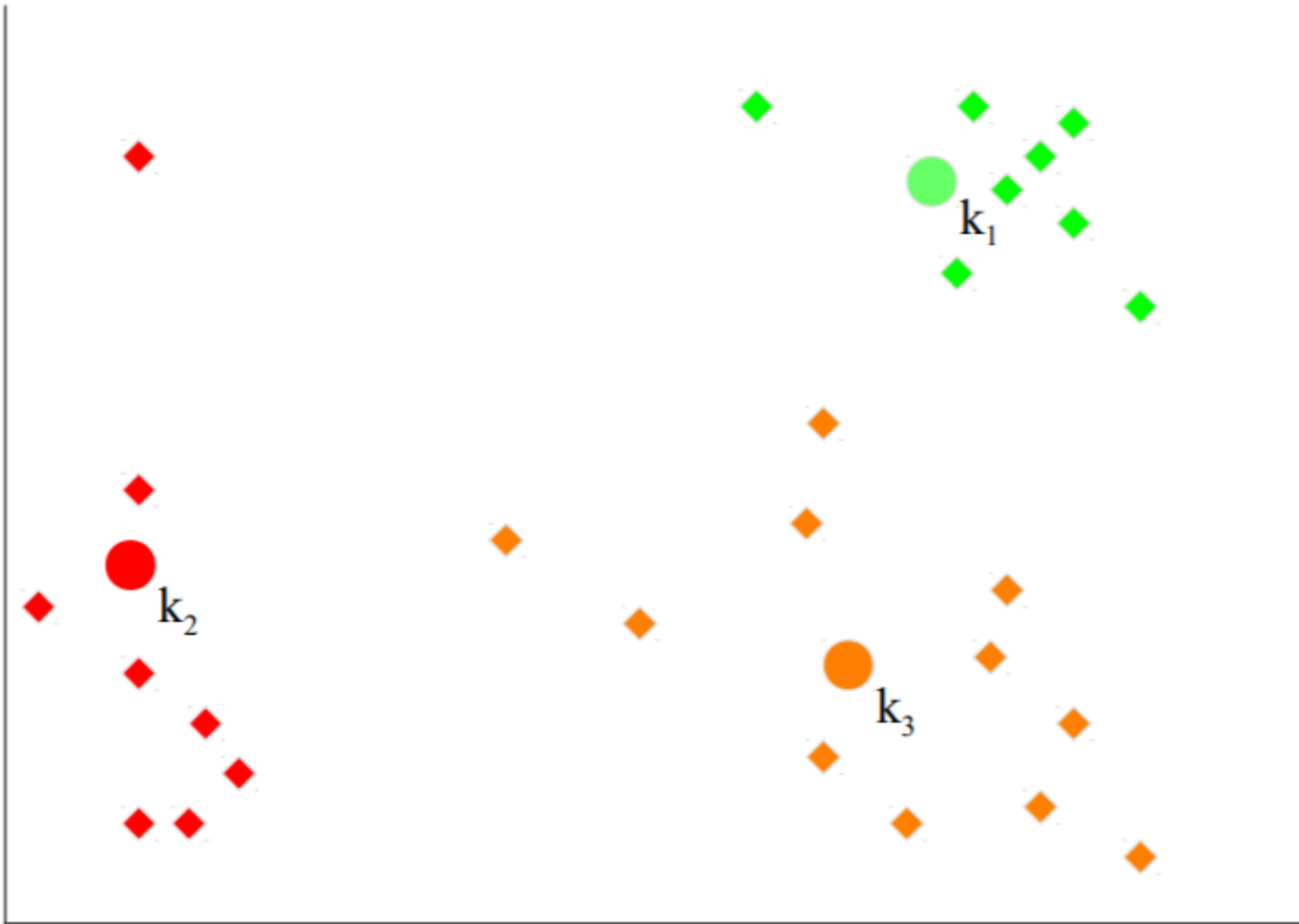
K-Means Example, Step 6

24



K-Means Example, Step 7

25



Clustering: Applications

26

- Application area: Market segmentation
- Goal: Subdivide a market into distinct subsets of customers
- Approach:
 - Collect information about customers
 - Find clusters of similar customers
 - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters



Clustering: Applications

27

Application area: Document Clustering

- Goal: Find groups of documents that are similar to each other based on the important terms appearing in them
- Approach
 - Identify frequently occurring terms in each document
 - Define a similarity measure based on the frequencies of different terms
- Application Example: Grouping of stories in Google News



Classification

28

- Goal: **Previously unseen records** should be assigned a class from a **given set of classes** as accurately as possible.

Approach:

- Given a collection of records (**training set**)
 - each record contains a set of **attributes**
 - one attribute is the **class attribute (label)** that should be predicted
- Find a **model** for predicting the class attribute as a function of the values of other attributes



Classification: Example

29

- **Training set:**



"tree"



"tree"



"tree"



"not a tree"



"not a tree"



"not a tree"

- **Learned model:** "Trees are big, green plants without wheels."

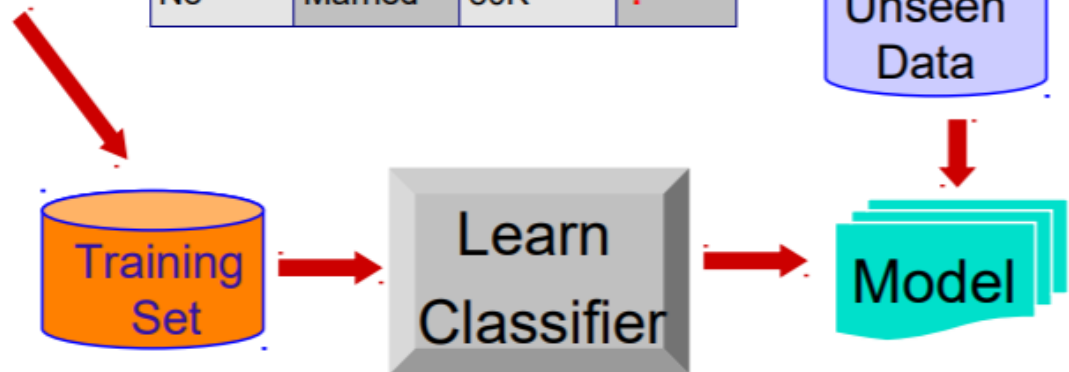
Classification: Workflow

30

Class/Label Attribute

| <i>Tid</i> | 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 |

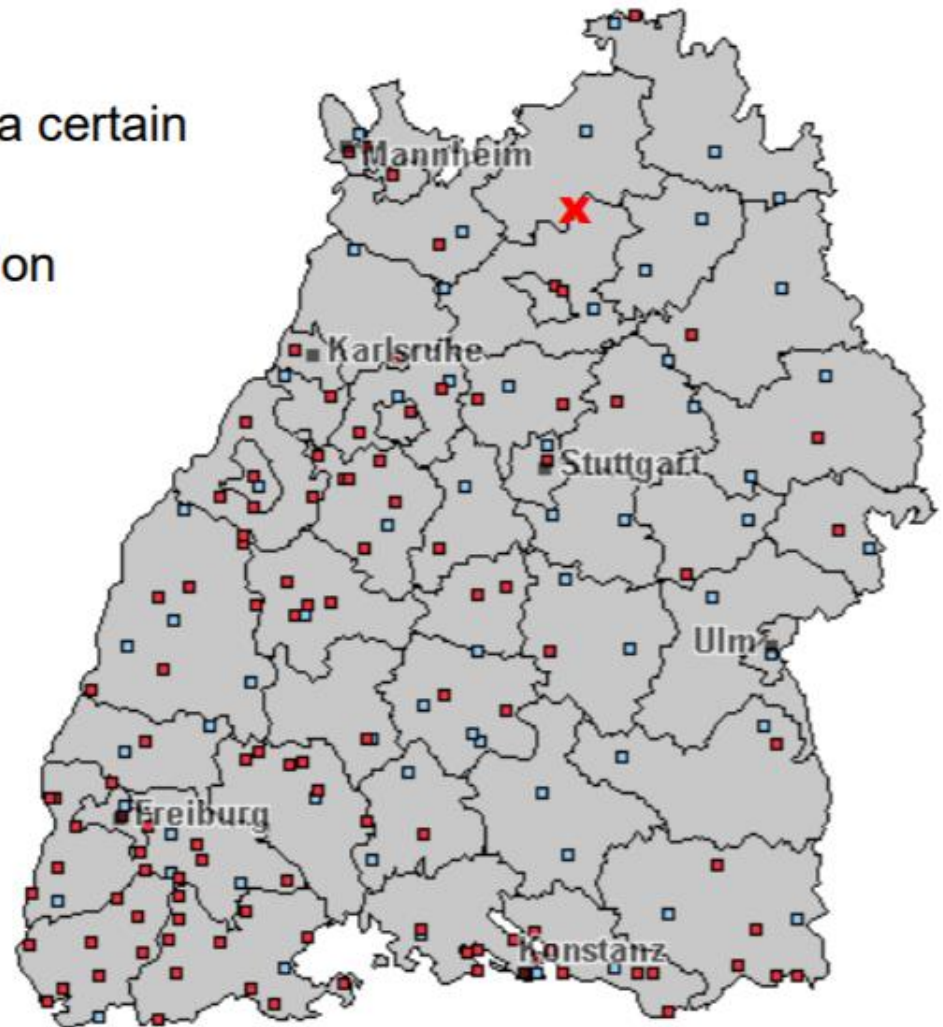
| Refund | Marital Status | Taxable Income | Cheat |
|--------|----------------|----------------|-------|
| No | Single | 75K | ? |
| Yes | Married | 50K | ? |
| No | Married | 150K | ? |
| Yes | Divorced | 90K | ? |
| No | Single | 40K | ? |
| No | Married | 80K | ? |



k Nearest Neighbors

31

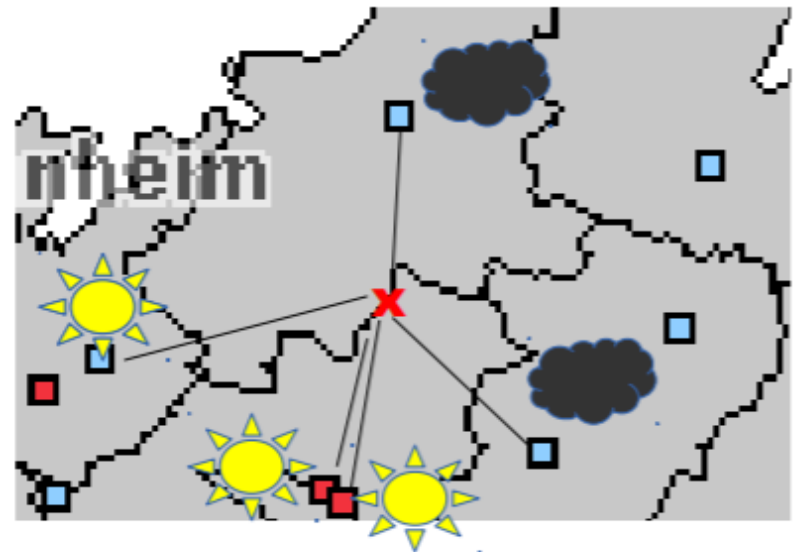
- Problem
 - find out what the weather is in a certain place
 - where there is no weather station
 - how could you do that?



k Nearest Neighbors

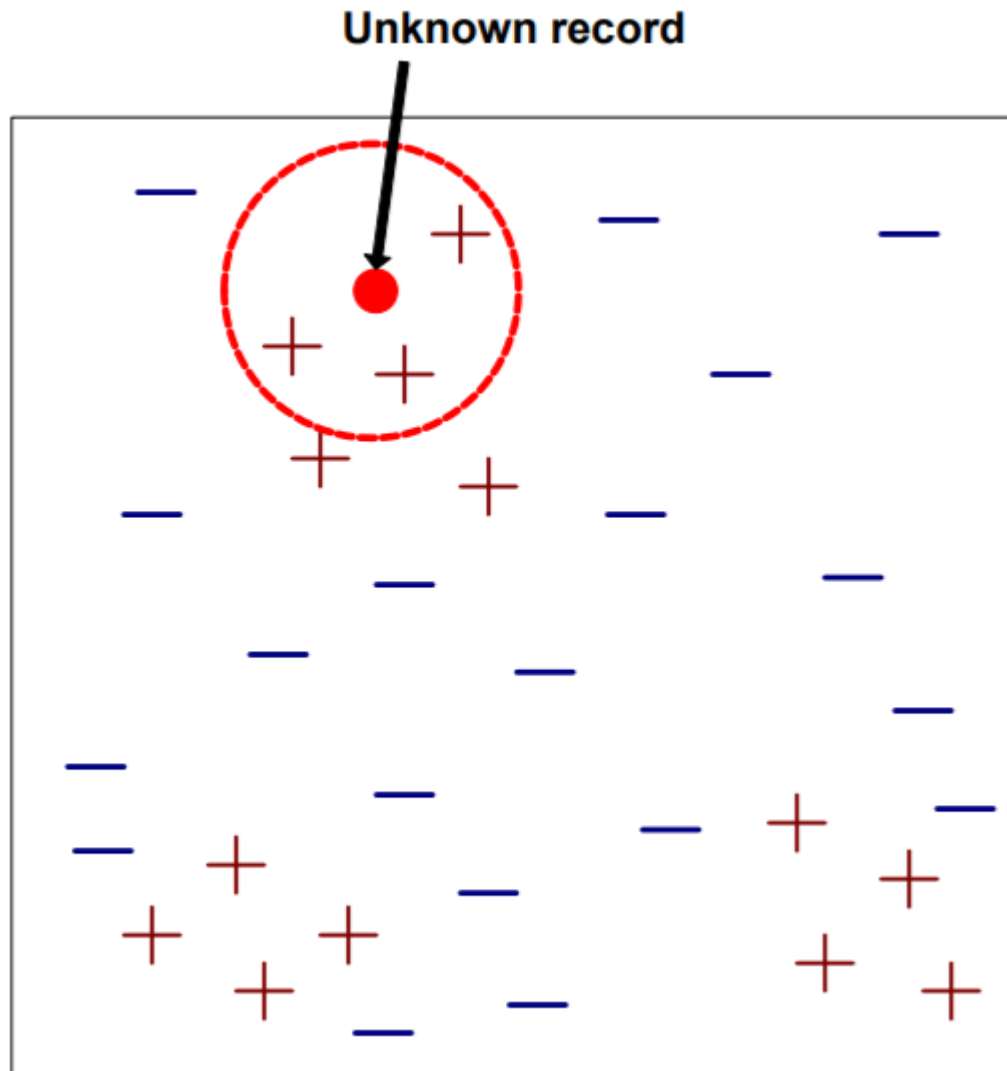
32

- Idea: use the average of the nearest stations
- Example:
 - 3x sunny
 - 2x cloudy
 - result: sunny
- Approach is called
 - “k nearest neighbors”
 - where k is the number of neighbors to consider
 - in the example: $k=5$
 - in the example: “near” denotes geographical proximity



Nearest-Neighbor Classifiers

33



Classification: Application

34



- Application area: Fraud Detection
- Goal: Predict fraudulent cases in credit card transactions.
- Approach:
 1. Use credit card transactions and information about account-holders as attributes
 - When and where does a customer buy? What does he buy?
 - How often he pays on time? etc.
 2. Label past transactions as fraud or fair transactions This forms the class attribute
 3. Learn a model for the class attribute from the transactions
 4. Use this model to detect fraud by observing credit card transactions on an account

Classification: Application



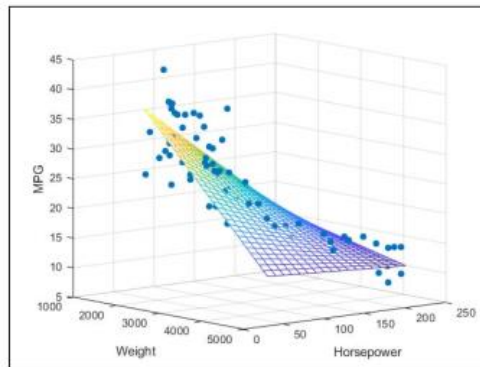
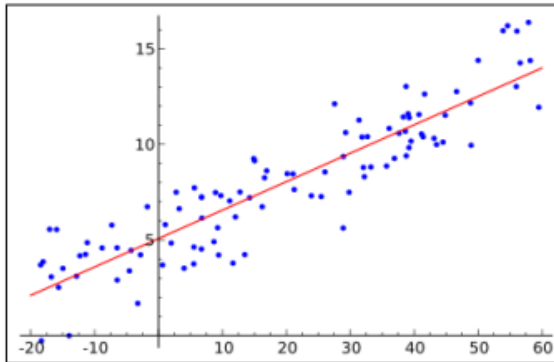
35

- Application area: Direct Marketing
- Goal: Reduce cost of a mailing campaign by targeting only the set of consumers that likely to buy a new product
- Approach:
 1. Use data from a campaign introducing a similar product in the past
 - we know which customers decided to buy and which decided otherwise
 - this {buy, don't buy} decision forms the class attribute
 2. Collect various demographic, lifestyle, and company-interaction related information about the customers
 - age, profession, location, income, marriage status, visits, logins, etc.
 3. Use this information to learn a classification model
 4. Apply model to decide which consumers to target

Regression

36

- Predict a value of a continuous variable based on the values of other variables, assuming a linear or nonlinear model of dependency
- Examples:
 - Predicting sales amounts of new product based on advertising expenditure
 - Predicting the price of a house or car
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.

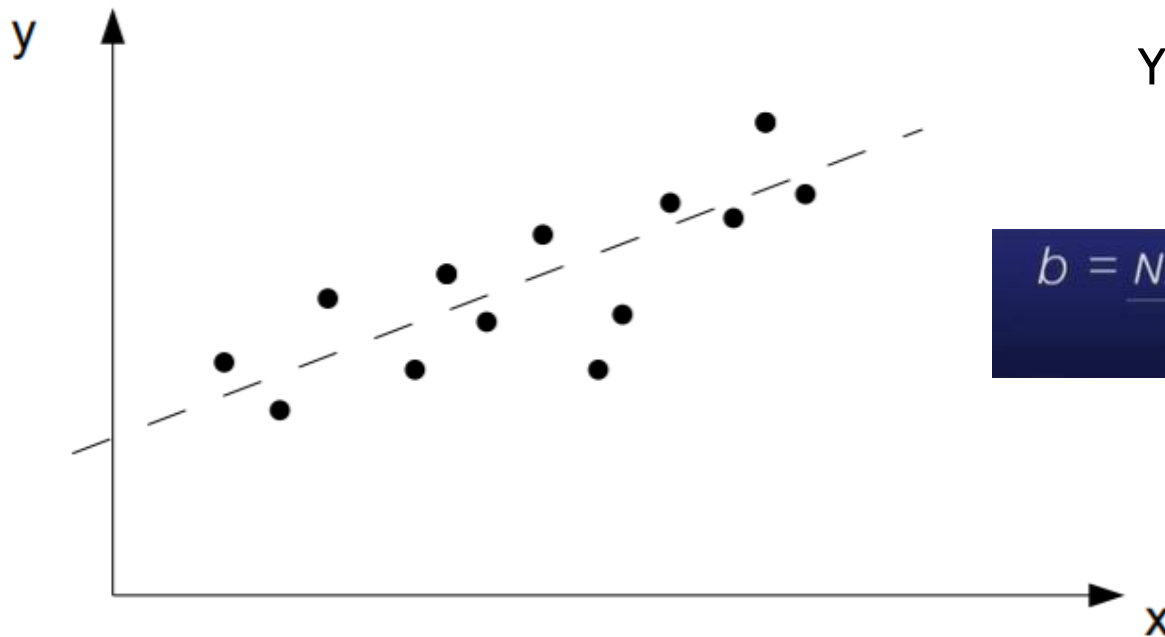


- Difference to classification: The predicted attribute is continuous, while classification is used to predict nominal attributes (e.g. yes/no)

Linear Regression

37

- Assumption: target variable y is (approximately) linearly dependent on attributes
 - for visualization: one attribute x
 - in reality: $x_1 \dots x_n$



$$Y = a + bX$$

$$b = \frac{N \sum XY - (\sum X) (\sum Y)}{N \sum X^2 - (\sum X)^2}$$

Association Rule

38

- Given a set of records each of which contain some number of items from a given collection
- discover **frequent itemsets** and produce **association rules** which will predict occurrence of an item based on occurrences of other items

Market-Basket transactions

| <i>TID</i> | <i>Items</i> |
|------------|---------------------------|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

Examples of Association Rules

$\{\text{Diaper}\} \rightarrow \{\text{Beer}\},$
 $\{\text{Milk, Bread}\} \rightarrow \{\text{Eggs, Coke}\},$
 $\{\text{Beer, Bread}\} \rightarrow \{\text{Milk}\},$

→ denotes co-occurrence,
not causality!

Apriori Algorithm

39

Two-step approach

- First: Frequent Itemset Generation
 - Generate all itemsets whose support \geq minsup
- Second: Rule Generation
 - Generate high confidence rules from each frequent itemset
 - where each rule is a binary partitioning of a frequent itemset

Apriori Algorithm: Frequent Itemset Generation

40

Support

$$s(X \rightarrow Y) := \frac{|X \cup Y|}{|T|}$$

$$s(\{\text{Bread}\}) = 0.8$$

$$s(\{\text{Bread}, \text{Milk}\}) = 0.6$$

$$s(\{\text{Bread}, \text{Milk}, \text{Diaper}\}) = 0.4$$

$$s(\{\text{Milk}\}) = 0.8$$

$$s(\{\text{Milk}, \text{Diaper}\}) = 0.6$$

$$s(\{\text{Milk}, \text{Diaper}, \text{Beer}\}) = 0.4$$

| <i>TID</i> | <i>Items</i> |
|------------|---------------------------|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

support \geq minsup threshold

Apriori Algorithm: Rule Generation

41

Confidence

$$c(X \rightarrow Y) := \frac{s(X \cup Y)}{s(X)}$$

{Milk, Diaper} → {Beer} c=0.67

{Milk} → {Beer} c=0.5

{Diaper} → {Beer} c=0.8

| <i>TID</i> | <i>Items</i> |
|------------|---------------------------|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

confidence ≥ minconf threshold

Association Rule Discovery: Applications

42

- Application area: Marketing and Sales Promotion
- Example rule discovered:
 $\{\text{Bagels, Coke}\} \rightarrow \{\text{Potato Chips}\}$
- Insights:
 - promote bagels to boost potato chips sales
 - if selling bagels is discontinued, this will affect potato chips sales
 - coke should be sold together with bagels to boost potato chips sales

Frequently Bought Together

amazon.com




+



+



Price For All Three: \$87.41

 Add all three to Cart

 Add all three to Wish List

[Show availability and shipping details](#)

Association Rule Discovery: Applications

43

- Customers who bought this product also bought...
 - ...do terrorists order bomb building parts on Amazon?
- Content-based recommendation
 - requirement: much data
 - e.g., Amazon transactions, Spotify logfiles

Frequently bought together



Total price: **\$35.19**

[Add all three to Cart](#)

[Add all three to List](#)

i These items are shipped from and sold by different sellers. [Show details](#)

✓ **This item:** Black Iron Oxide - Fe3O4 - Natural - 5 Pounds **\$18.99**

✓ Elmer's Liquid School Glue, Washable, 1 Gallon, 1 Count - Great For Making Slime **\$10.49**

✓ Purex Sile-Fix Liquid Starch, 64 Ounces **\$5.71** [Add-on Item](#)

 Song

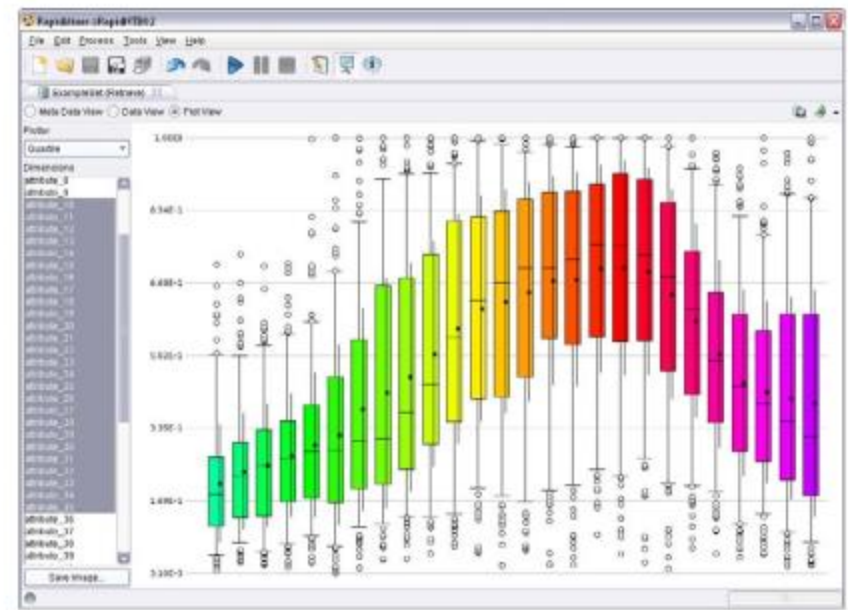
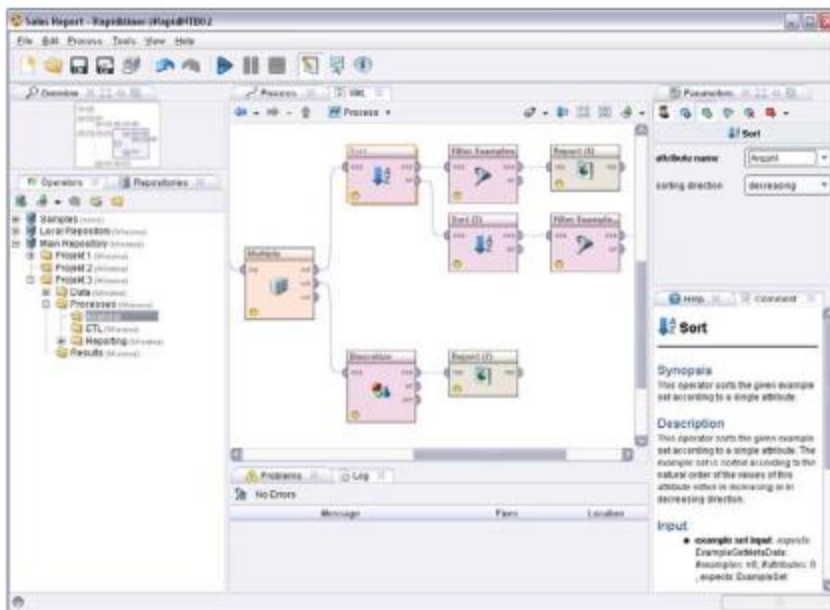
Du hast **The Gathering** und **Tiamat** gehört.
Diesen Song magst du vielleicht auch.



In Memoriam (Intro)
Moonspell

[Als nächstes abspielen](#)

- Powerful data mining suite
- Visual modelling of data mining pipelines
- Commercial tool, offering educational licenses



We use the Anaconda Python distribution

- includes relevant packages, e.g.
 - scikit-learn, pandas
 - NumPy, Matplotlib
- includes Jupyter as development environment



```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import GridSearchCV

knn_estimator = KNeighborsClassifier()
parameters = {
    'n_neighbors': range(2, 9),
    'algorithm': ['ball_tree', 'kd_tree', 'brute']
}
stratified_10_fold_cv = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
grid_search_estimator = GridSearchCV(knn_estimator, parameters, scoring='accuracy',
                                     cv=stratified_10_fold_cv)
grid_search_estimator.fit(iris_data, iris_target)
```

