DATA MINING

Outline

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

What is Data Mining?

- Large quantities of data are collected about all aspects of our lives

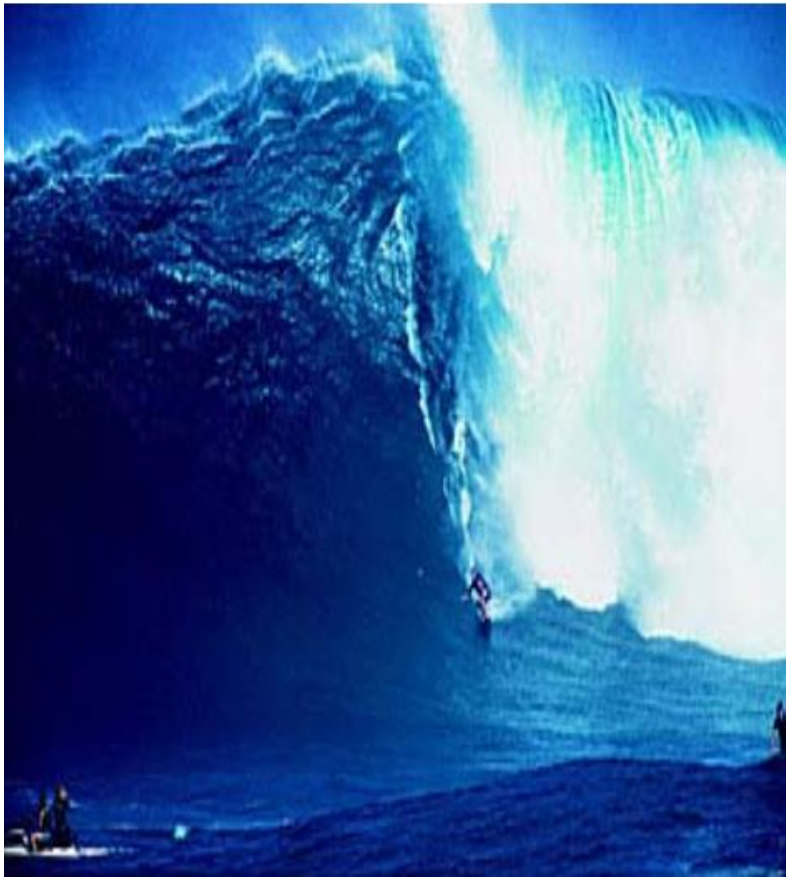
- This data contains interesting patterns

- Data Mining helps us to

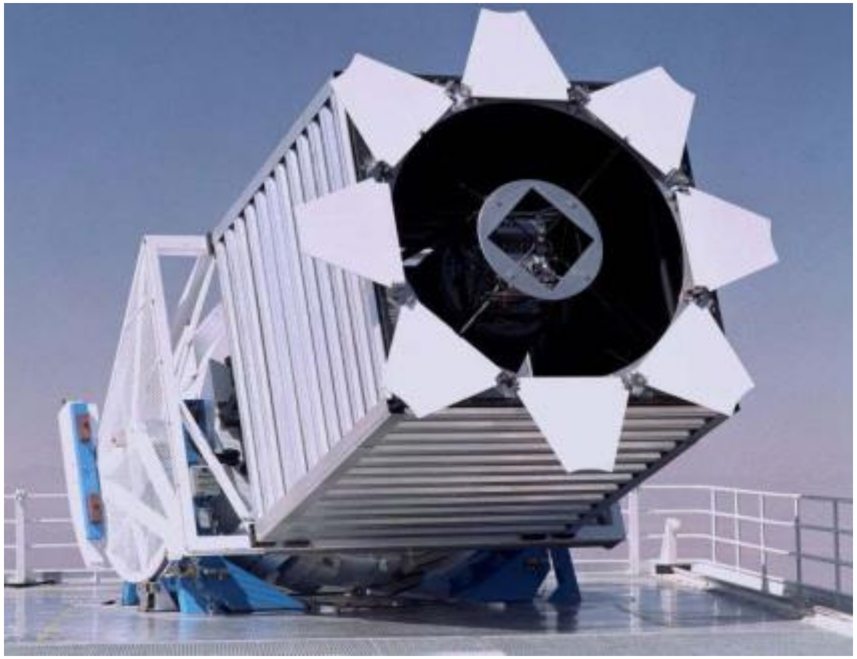
1. discover these patterns and

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...”



Sloan Digital Sky Survey

≈ 200 GB/day

≈ 73 TB/year

Predict

- Type of sky object:

Star or galaxy?

“We are Drowning in Data...”

US Library of Congress

≈ 235 TB archived

≈ 40 Wikipedias

Discover

- Topic distributions
- Historic trends
- Citation networks

“We are Drowning in Data…”

Facebook

≈ 10 TB/day added

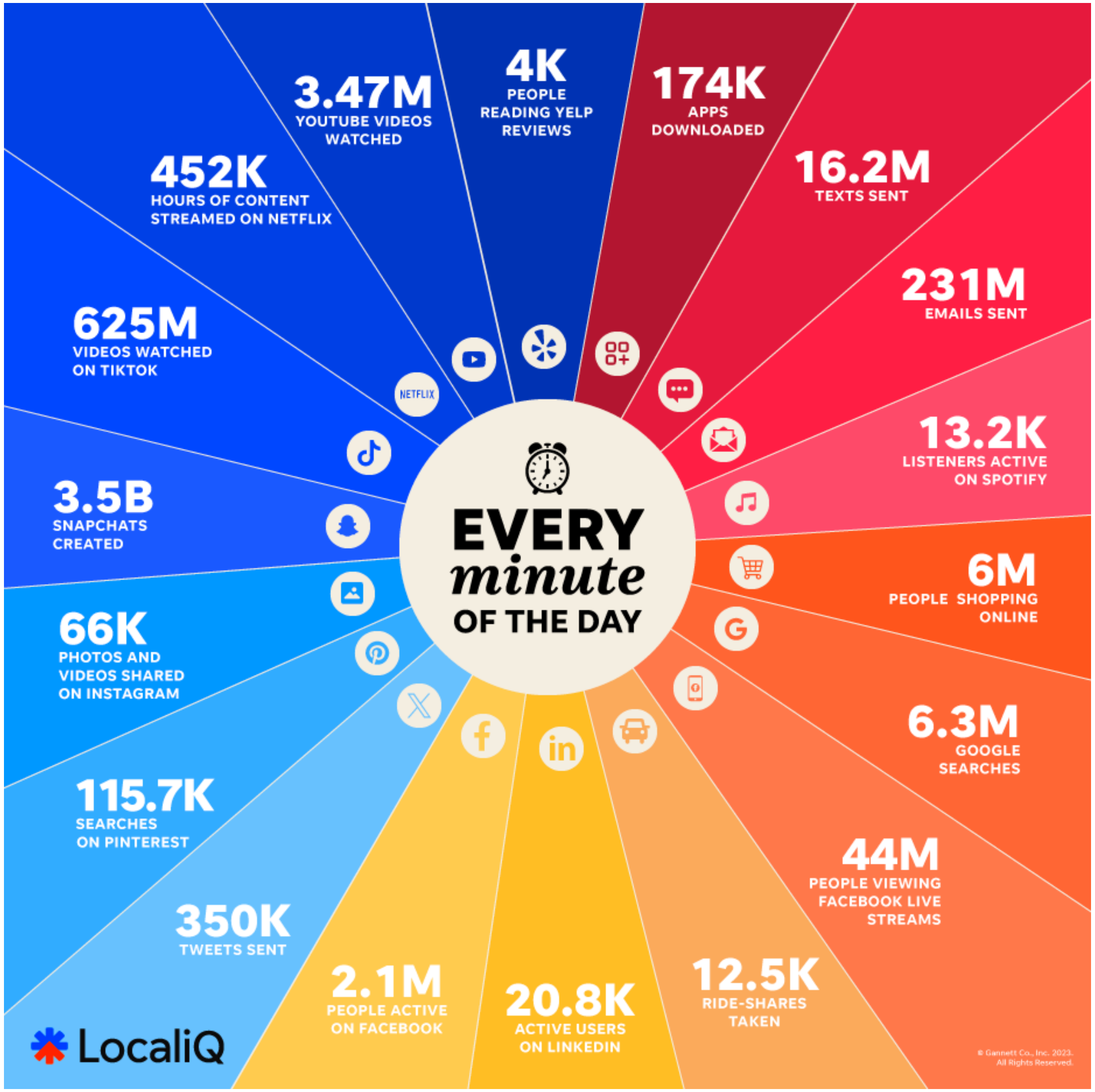
≈ over 300 Petabyte in

Facebook‘s data warehouse

Predict

• Interests and behavior of over one billion people

“We are Drowning in Data…”



Predict

- Interests and

behavior of mankind

“We are Drowning in Data...”

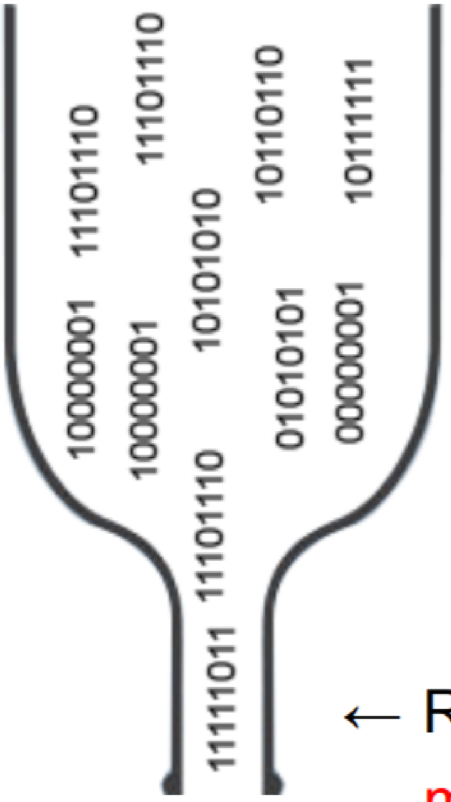
Law enforcement agencies collect unknown amounts of data from various sources

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

Predict

- Terrorist or not?
- Trustworthiness

“…but starving for knowledge!”

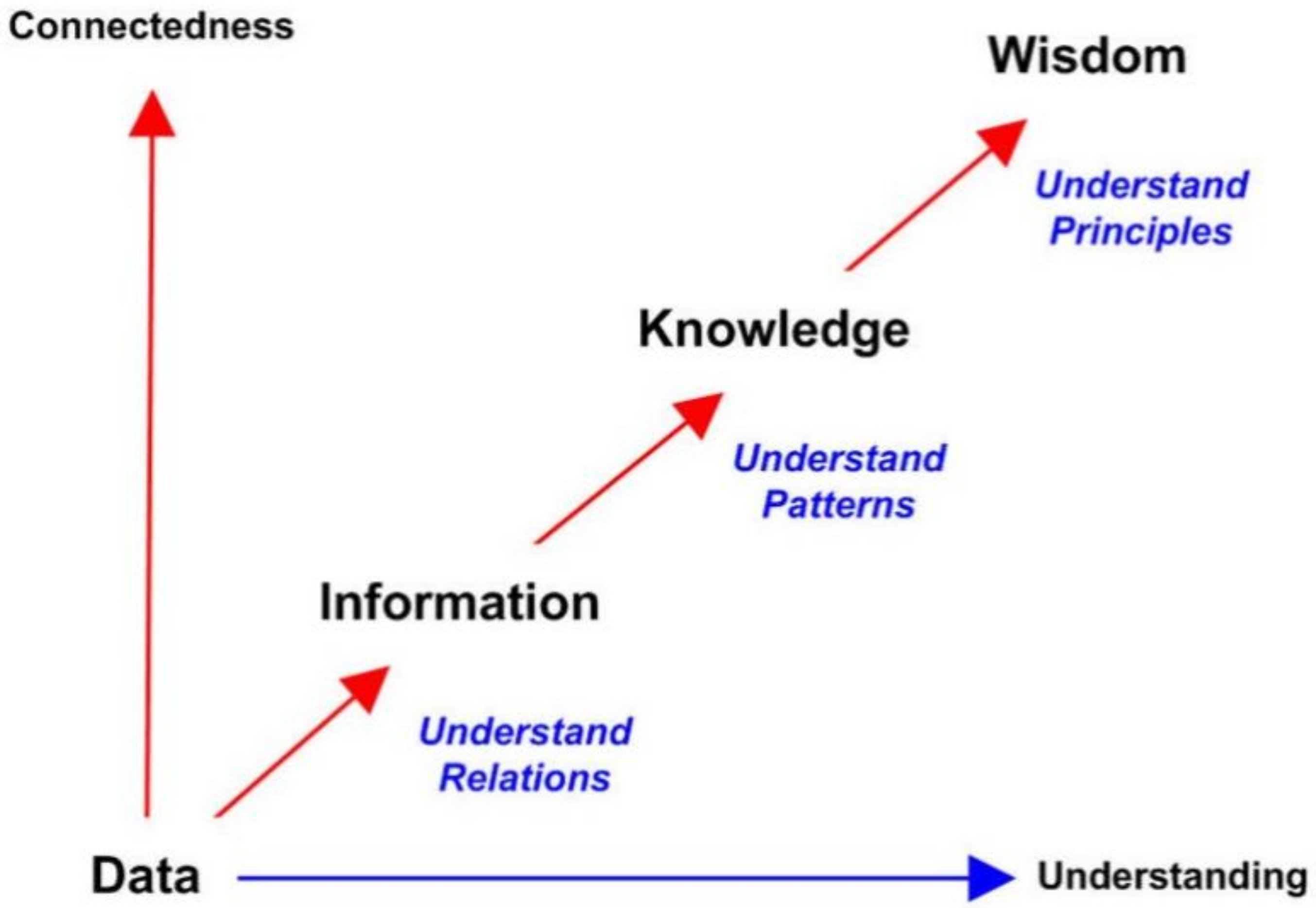


← Rate at which data are produced

← Rate at which data can be understood

manual interpretation is hardly feasible!

Data, Information, Knowledge, and Wisdom



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

Definitions of Data Mining

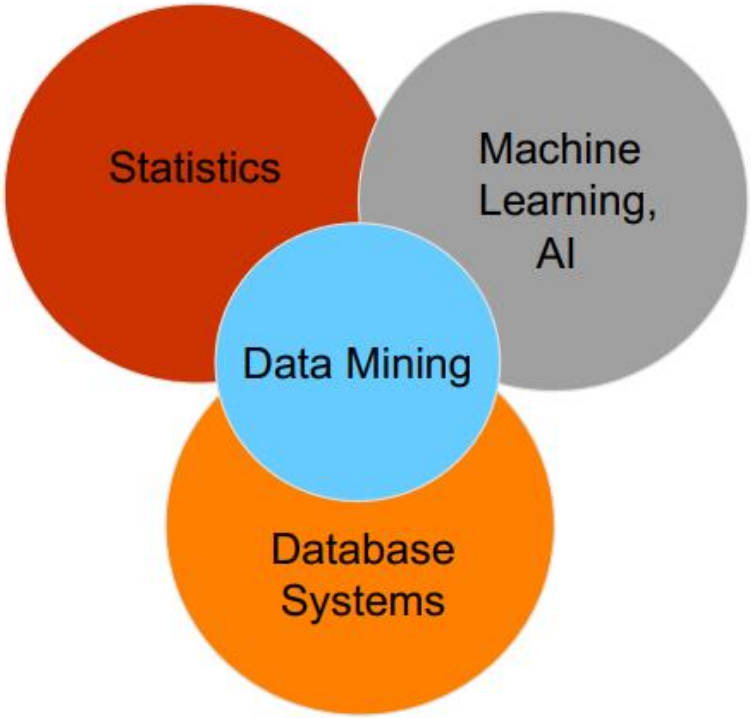
“Exploration & analysis, of large quantities of data in order to discover meaningful patterns.”

“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.”

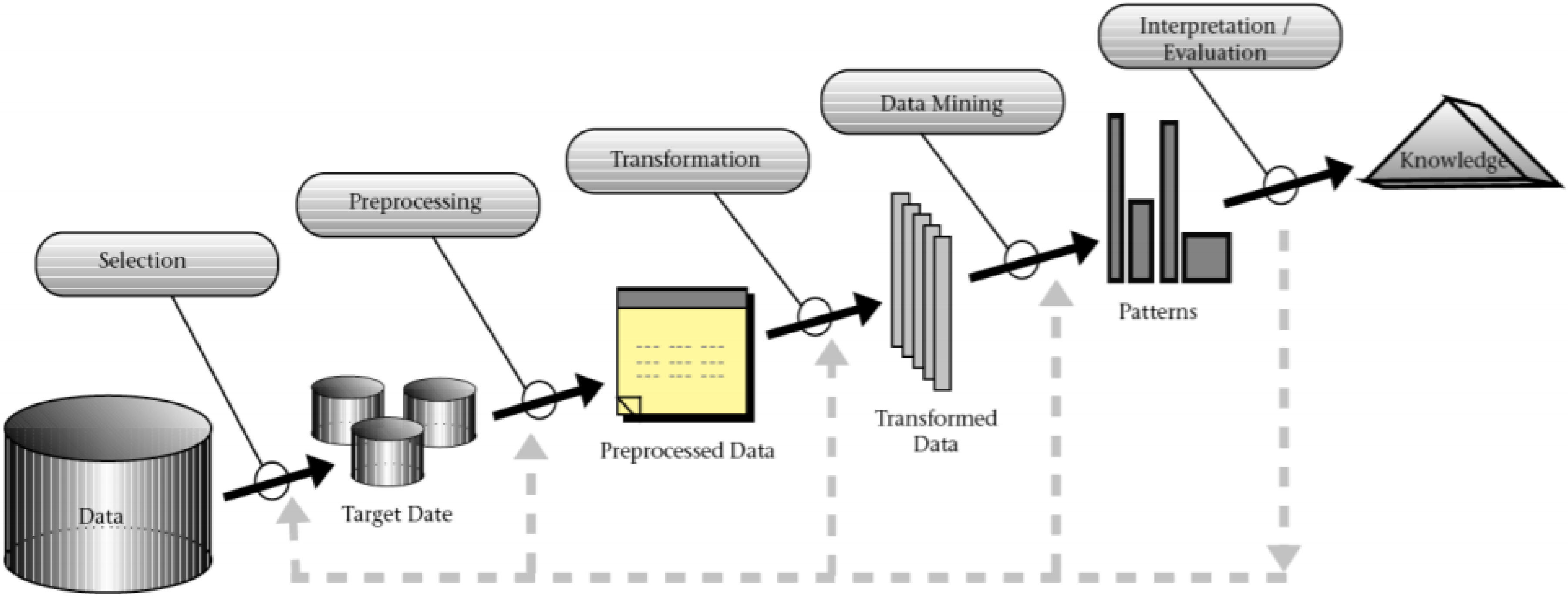
(Fred Menger, year unknown)

Origins of Data Mining

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

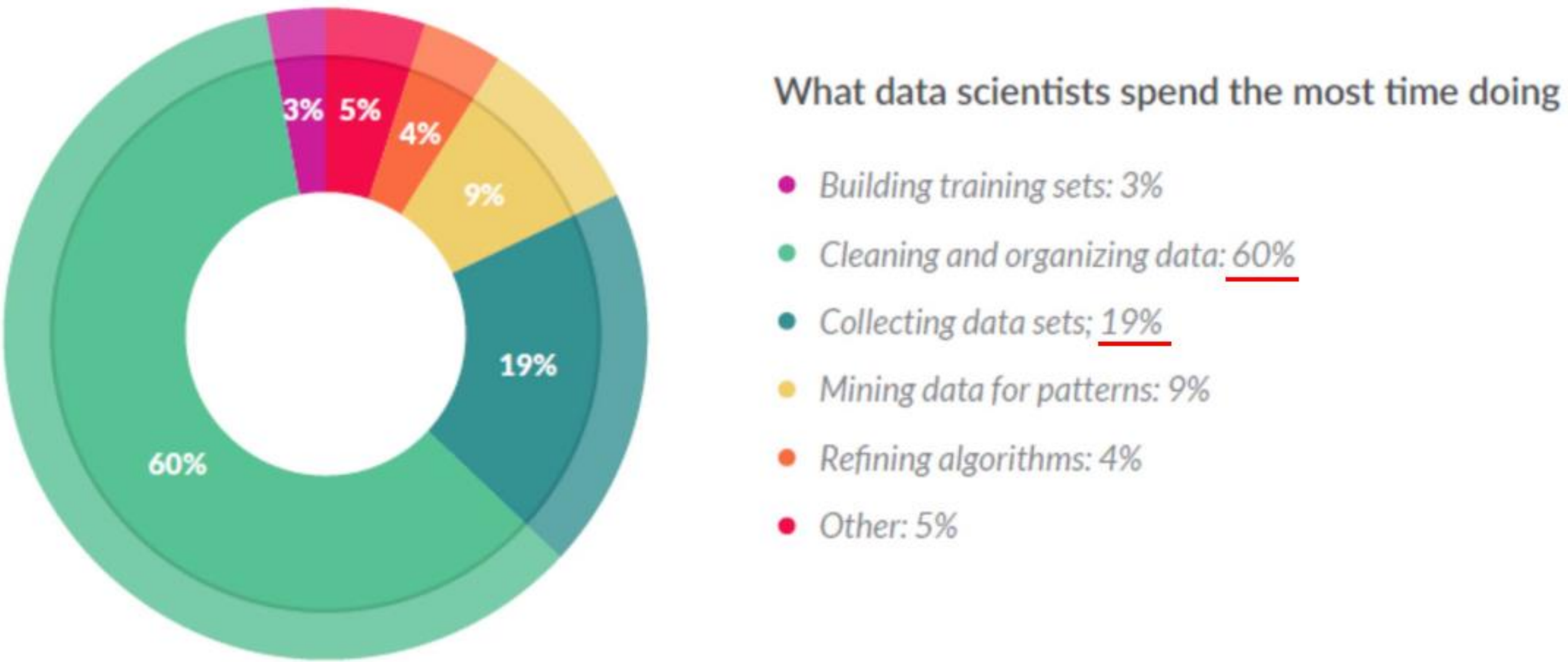


The Data Mining Process



Source: Fayyad et al. (1996)

How Do Data Scientists Spend Their Days?



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%*

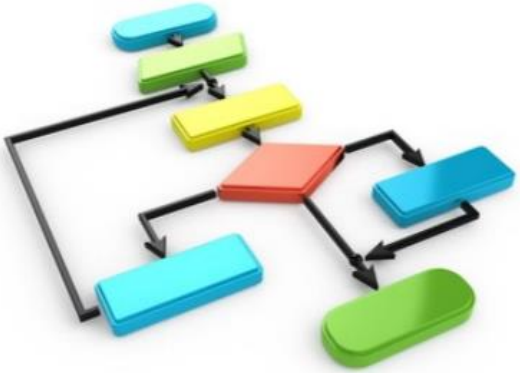
Tasks and Applications

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

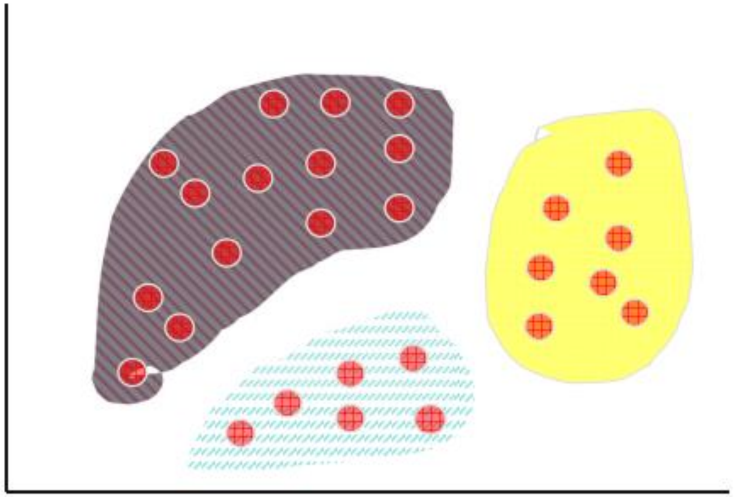
- descriptive = unsupervised
- predictive = supervised

Data Mining Tasks

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

Clustering

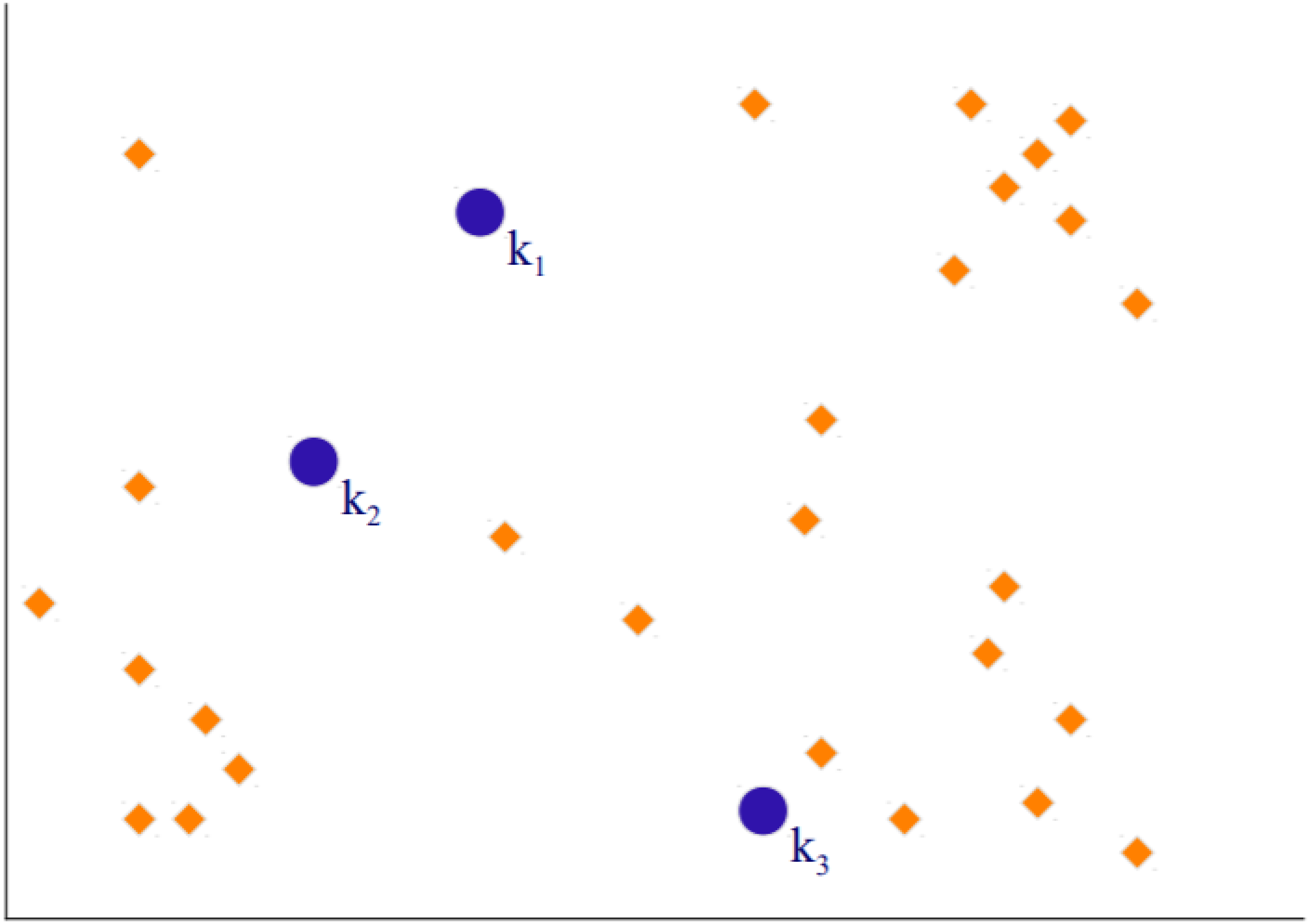
- Given a set of data points, and a similarity measure among them, find clusters such that
  - Data points in one cluster are similar to one another
  - Data points in separate clusters are different from each other
- Result
  - a descriptive grouping of data points



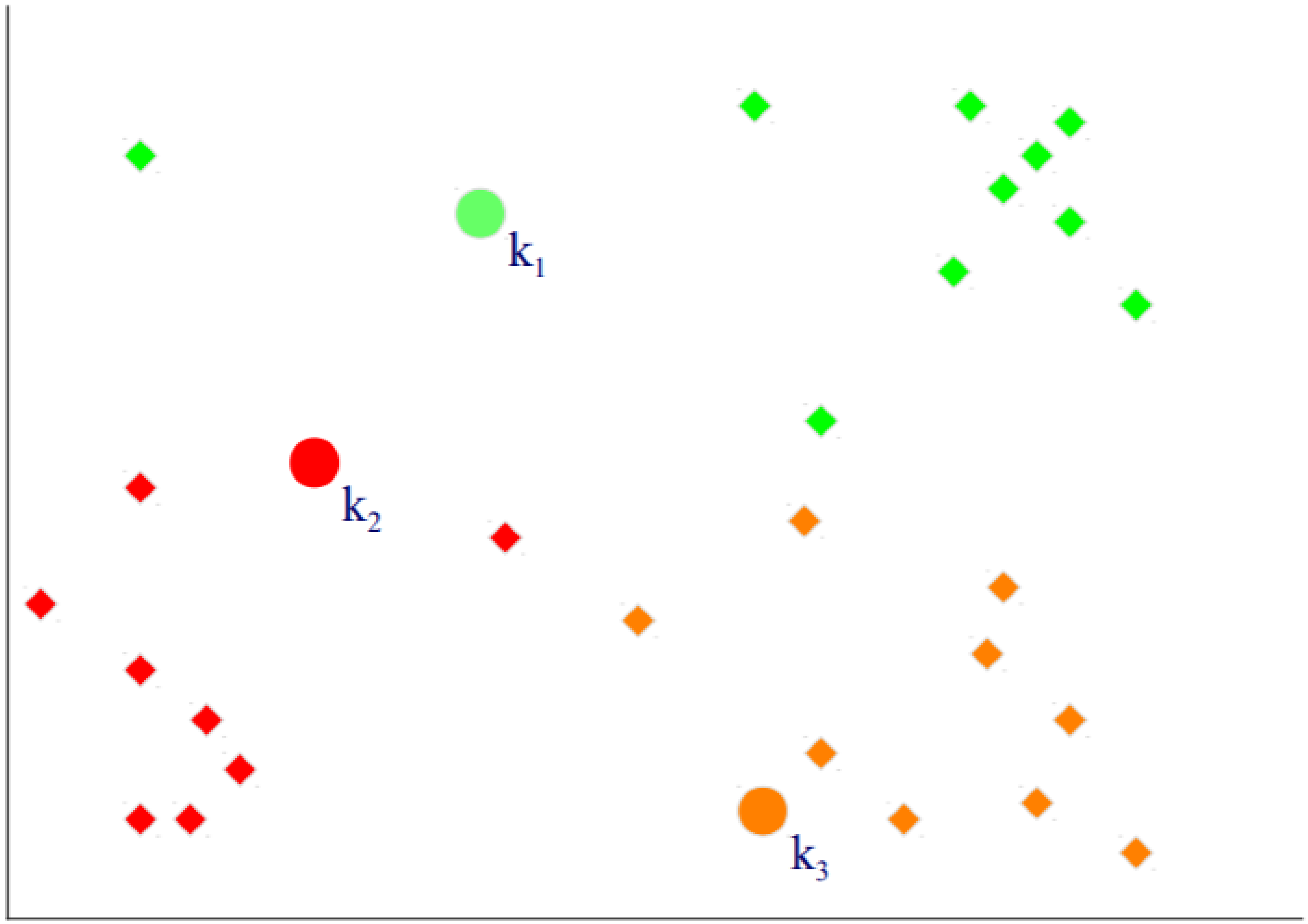
K-Means Clustering

- 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

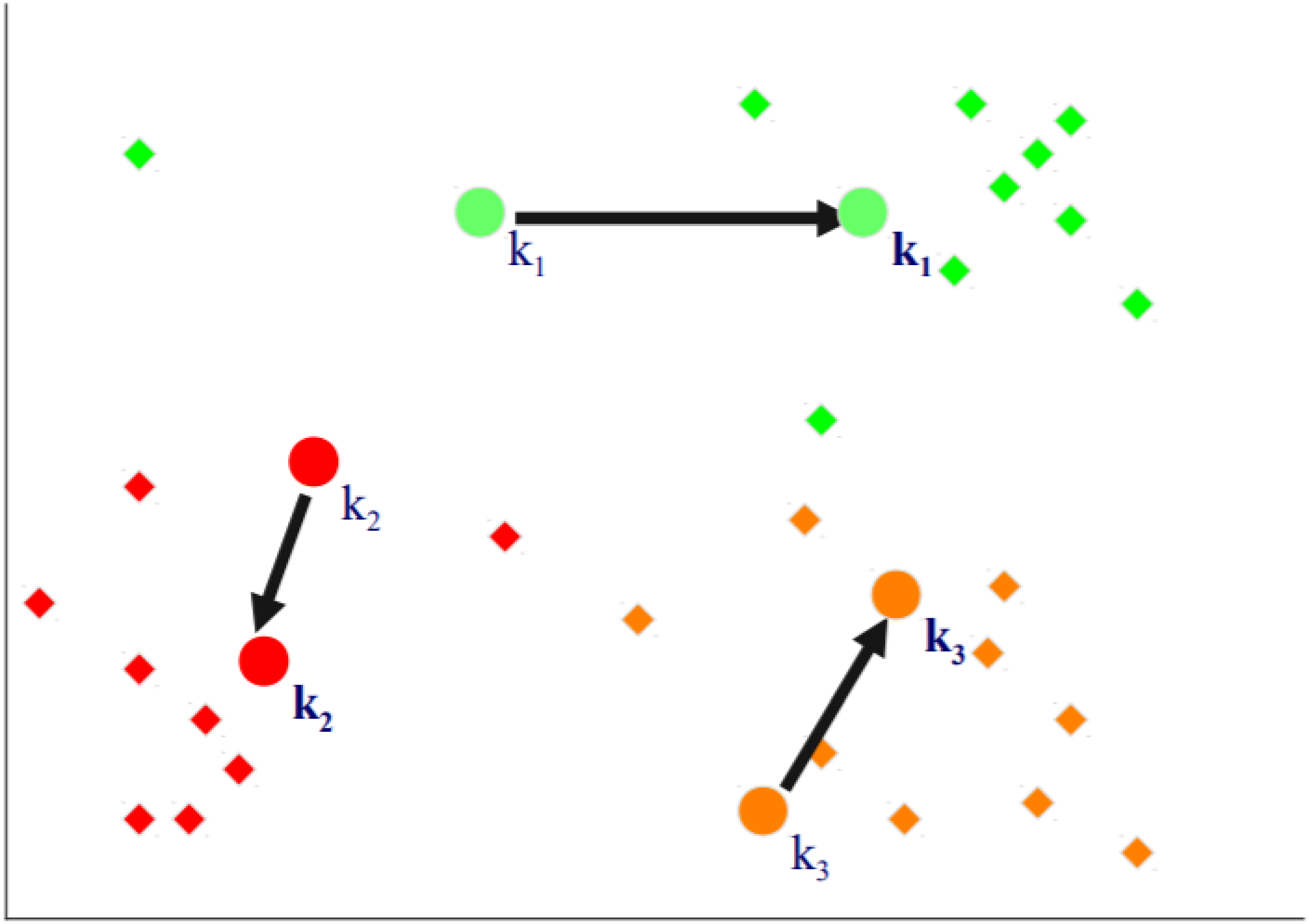
K-Means Example, Step 1



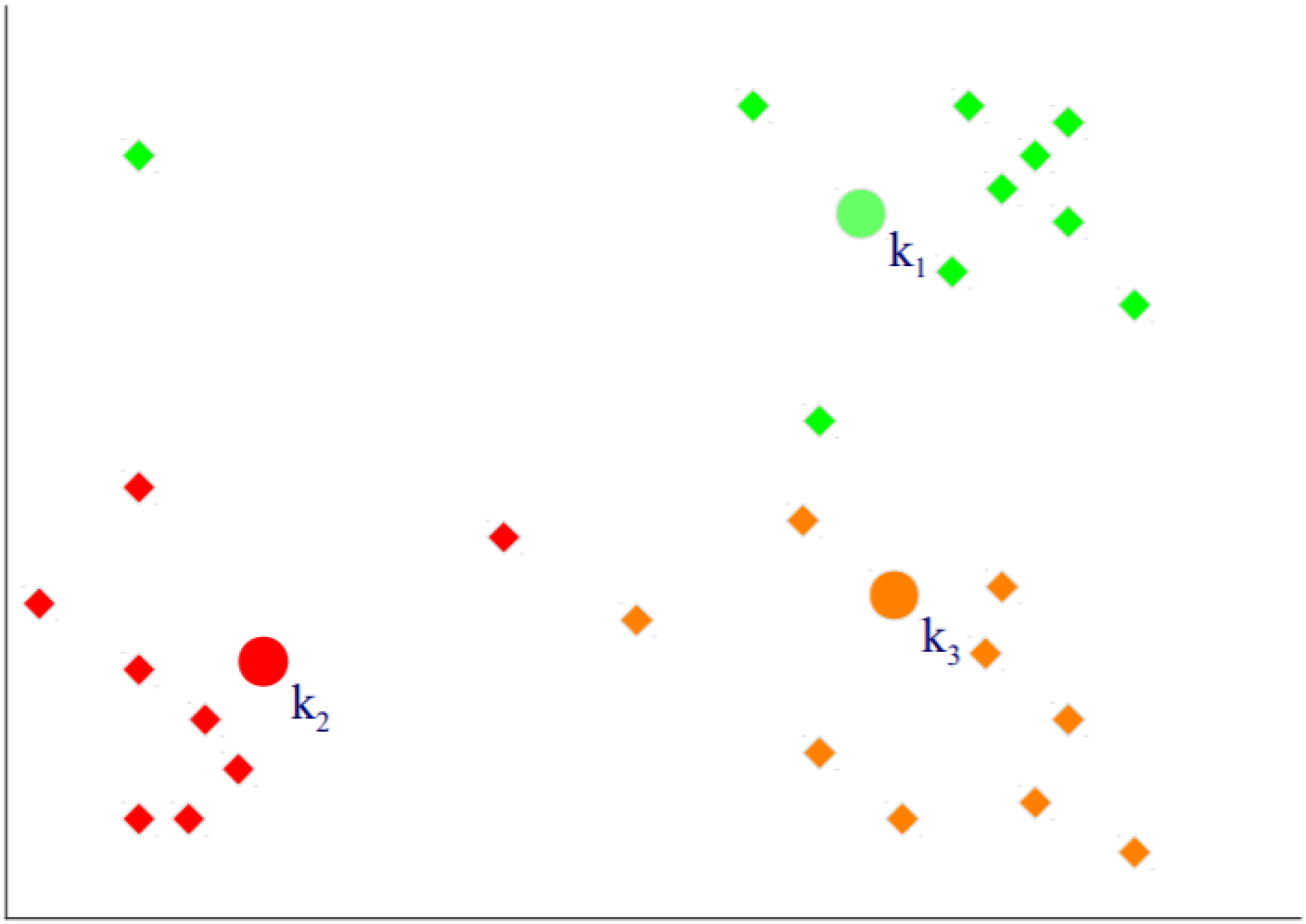
K-Means Example, Step 2



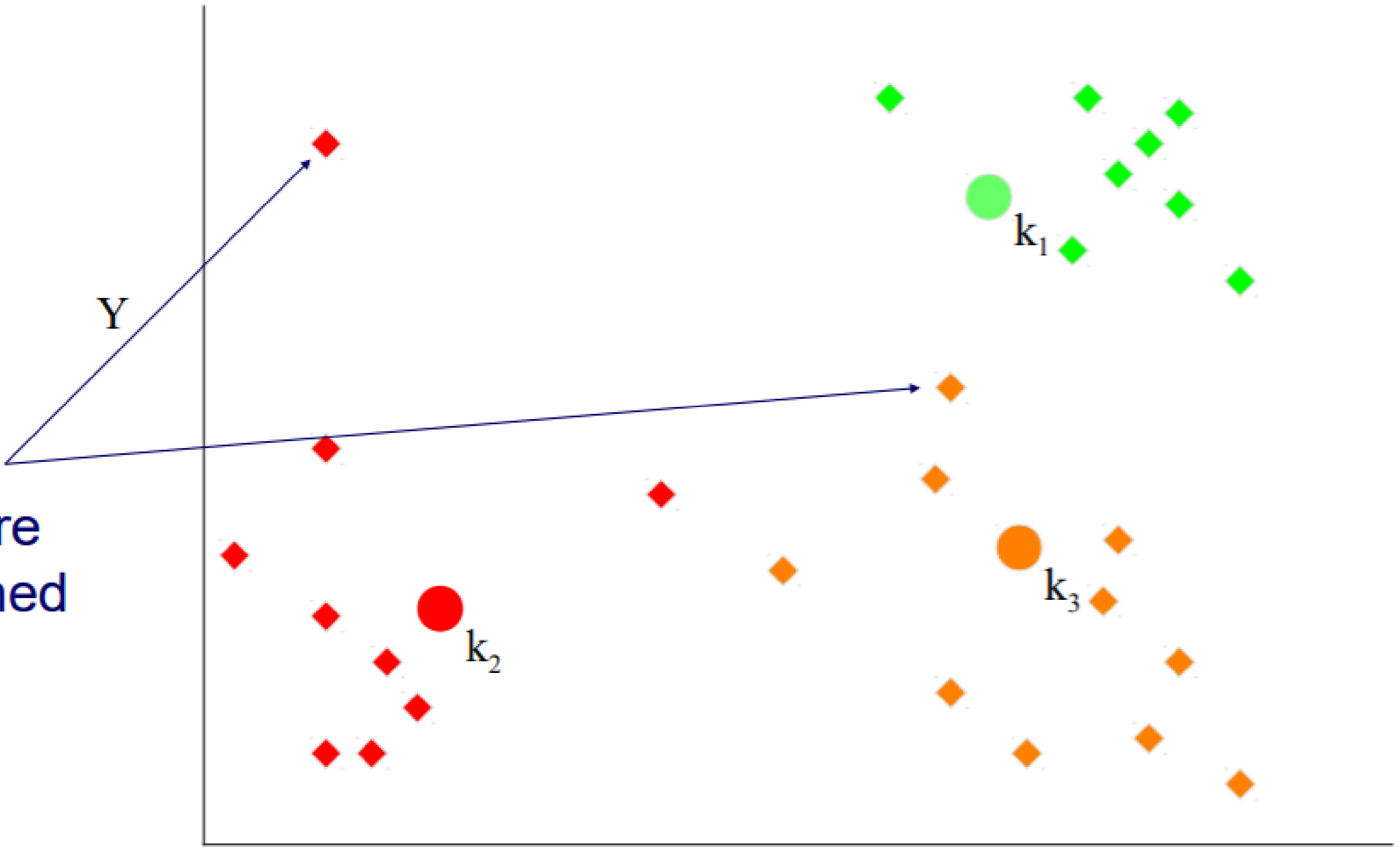
K-Means Example, Step 3



K-Means Example, Step 4

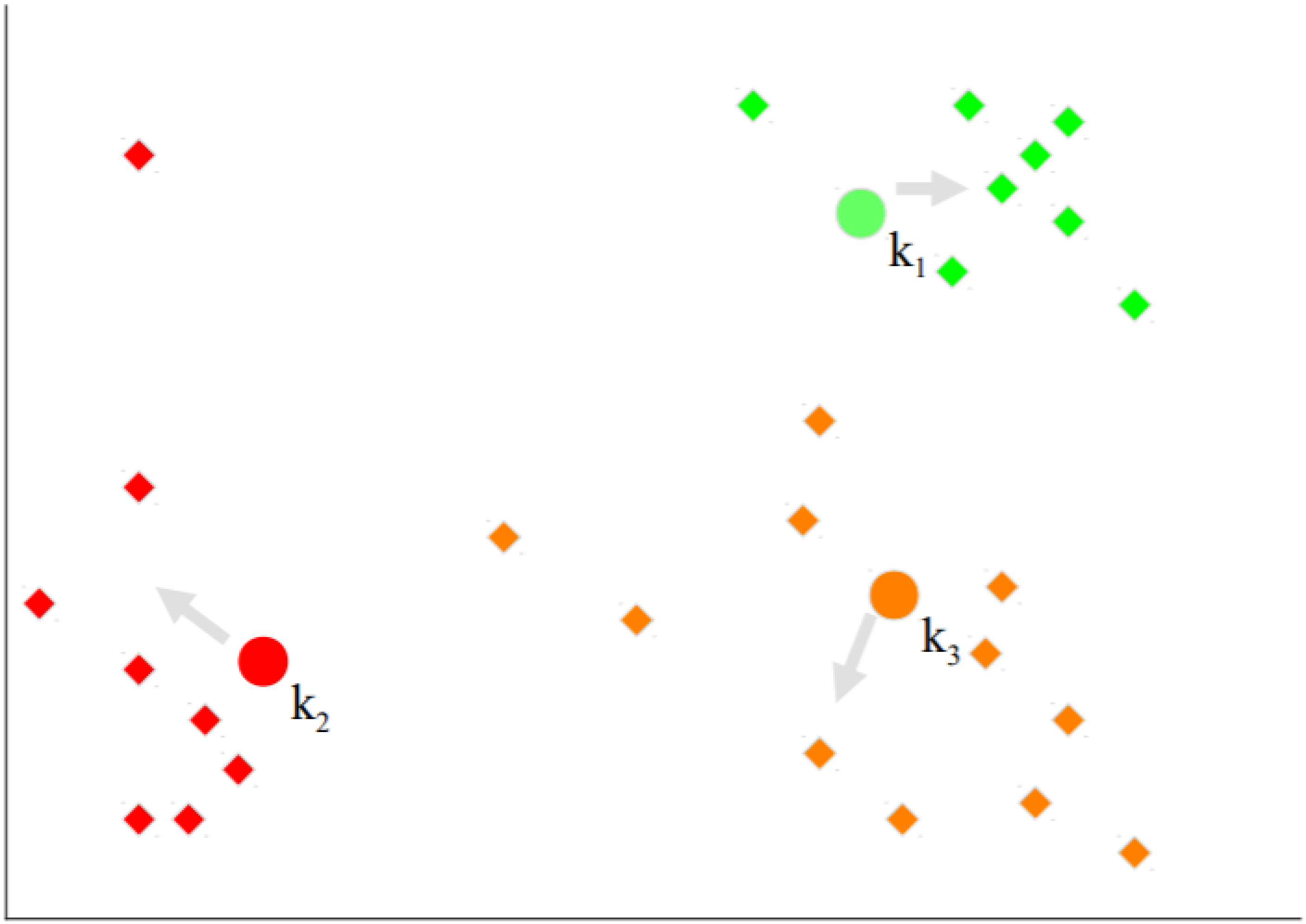


K-Means Example, Step 5

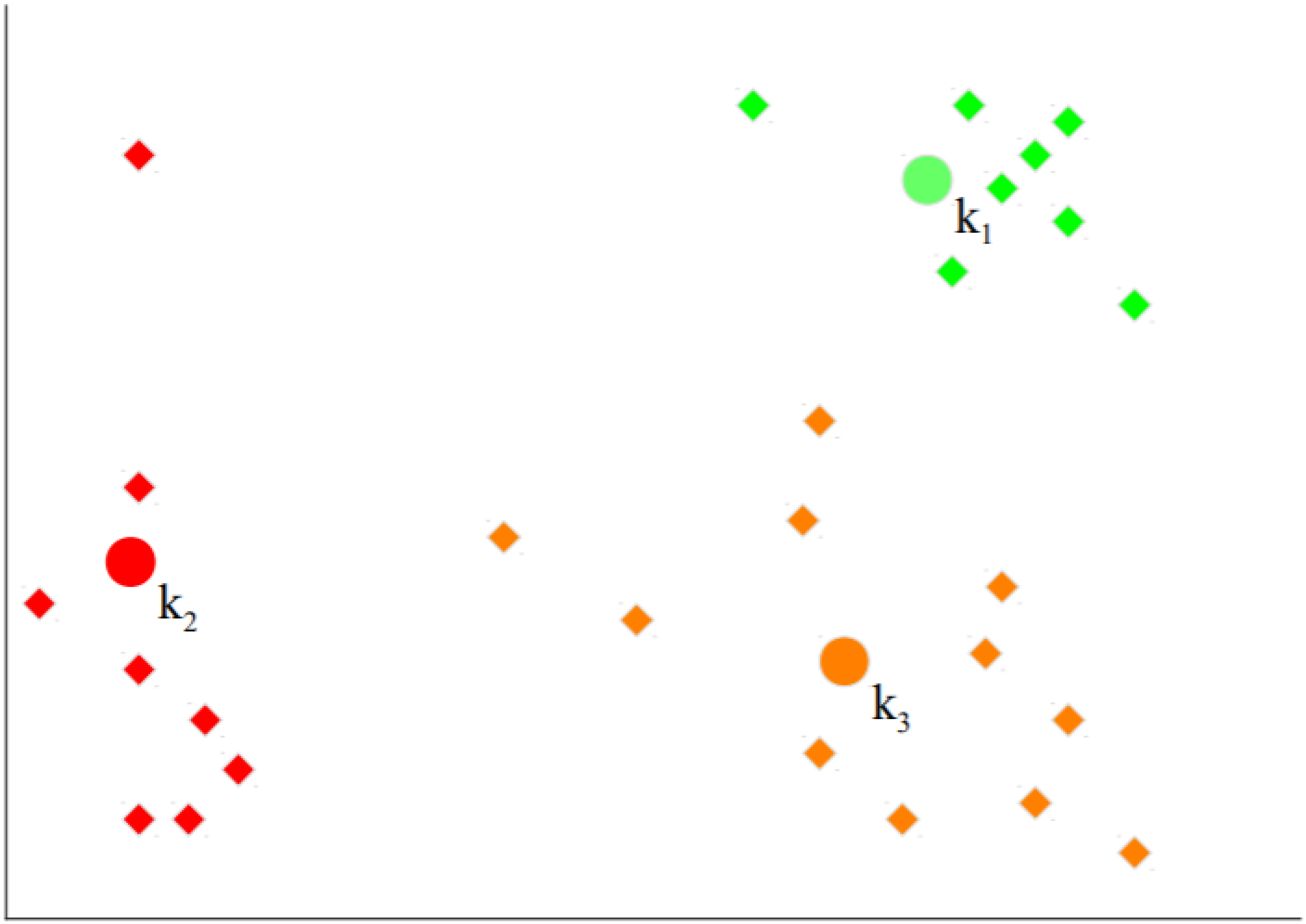


A: Two points are reassigned

K-Means Example, Step 6



K-Means Example, Step 7



Clustering: Applications

- Application area: Market segmentation
- Goal: Subdivide a market into distinct subsets of customers

- Approach:

– Collect information about customers

– Find clusters of similar customers

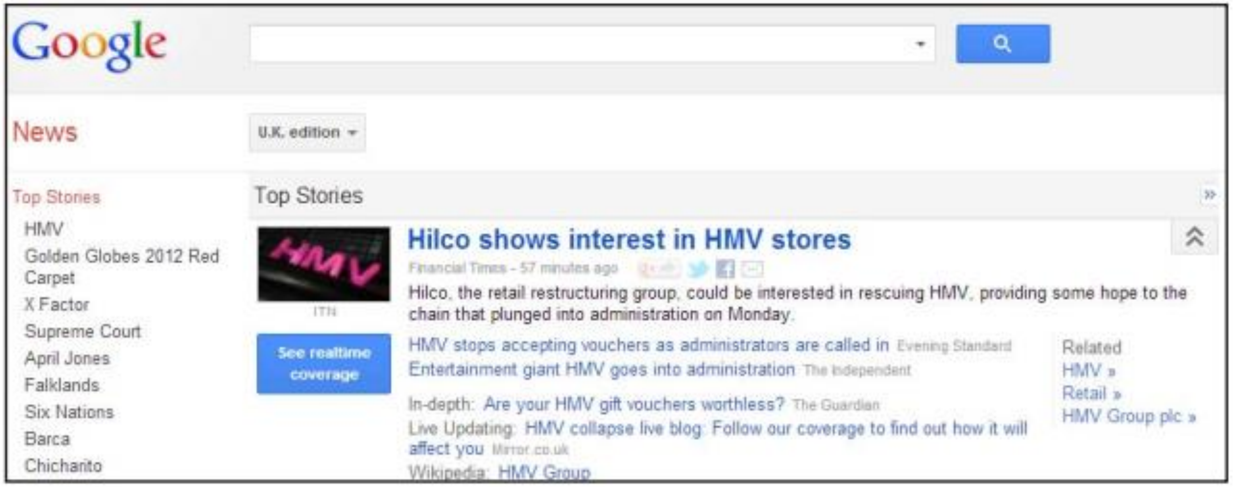
– Measure the clustering quality by observing buying patterns

of customers in same cluster vs. those from different clusters

Clustering: Applications

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



Top Stories

Hilco shows interest in HMV stores

Financial Times - 57 minutes ago

Hilco, the retail restructuring group, could be interested in rescuing HMV, providing some hope to the chain that plunged into administration on Monday.

HMV stops accepting vouchers as administrators are called in Evening Standard

Entertainment giant HMV goes into administration The Independent

In-depth: Are your HMV gift vouchers worthless? The Guardian

Live Updating: HMV collapse live blog: Follow our coverage to find out how it will affect you Mirror.co.uk

Wikipedia: HMV Group

Classification

- 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

- Training set:

"tree"

"tree"

"tree"

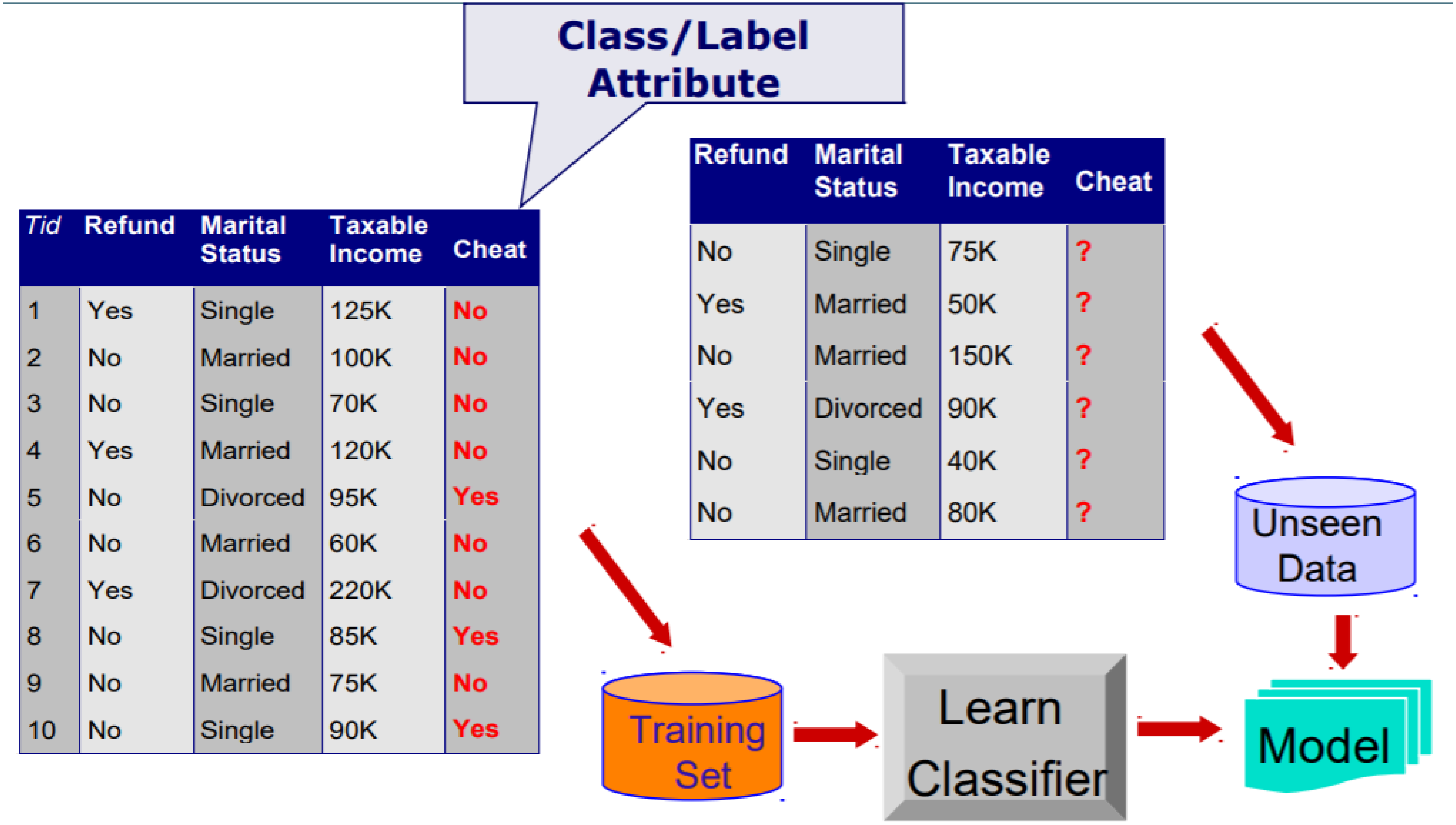
"not a tree"

"not a tree"

"not a tree"

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

Classification: Workflow



Class/Label Attribute

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

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

Training Set

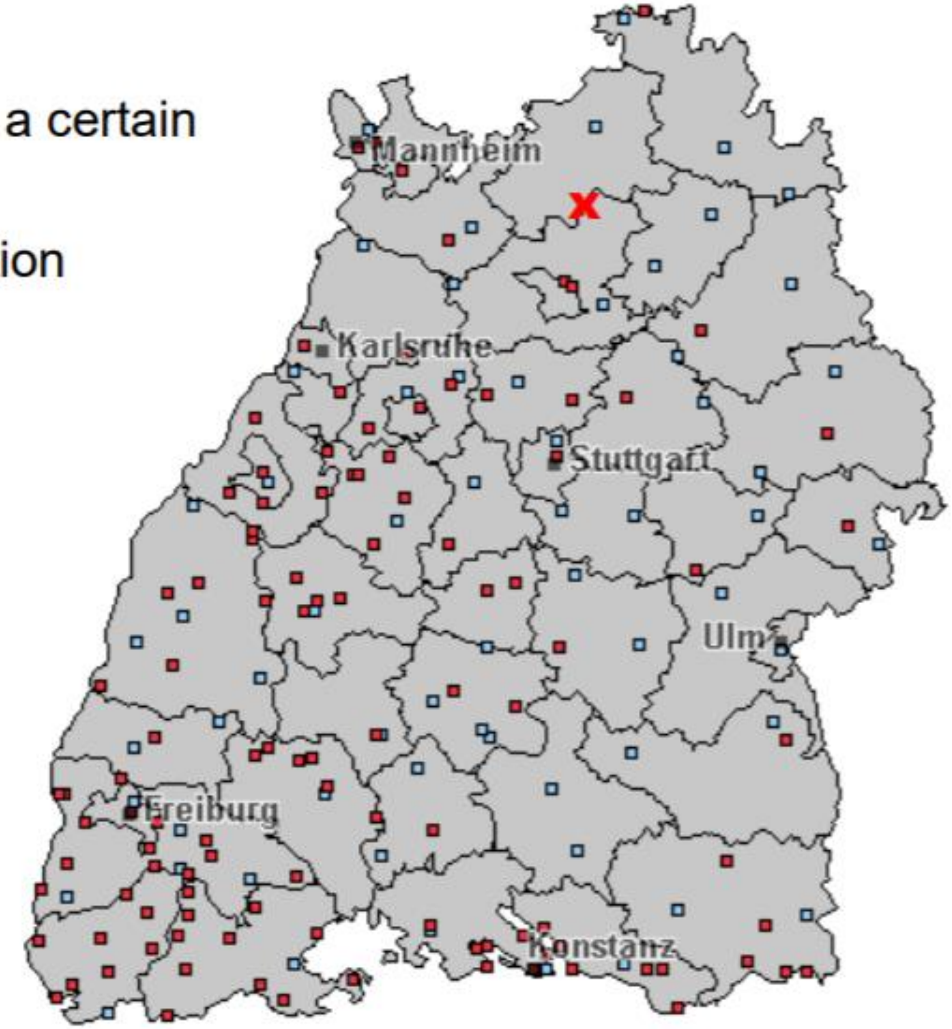
Learn Classifier

Model

Unseen Data

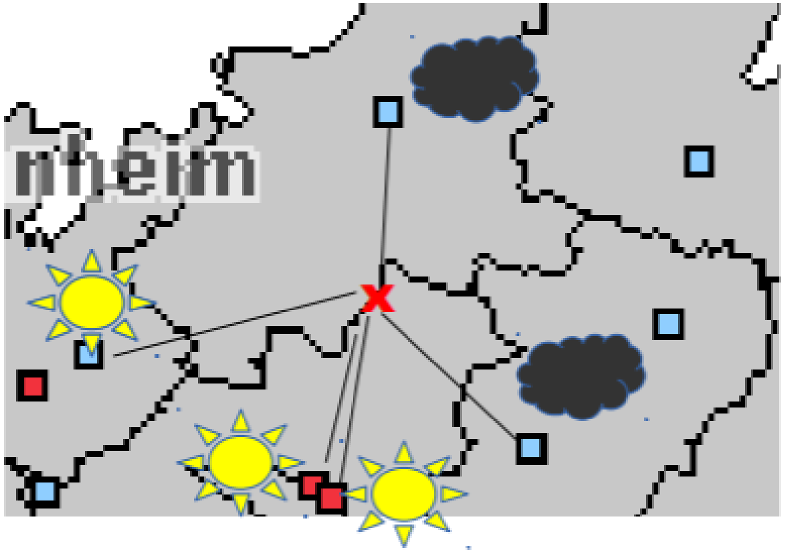
k Nearest Neighbors

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



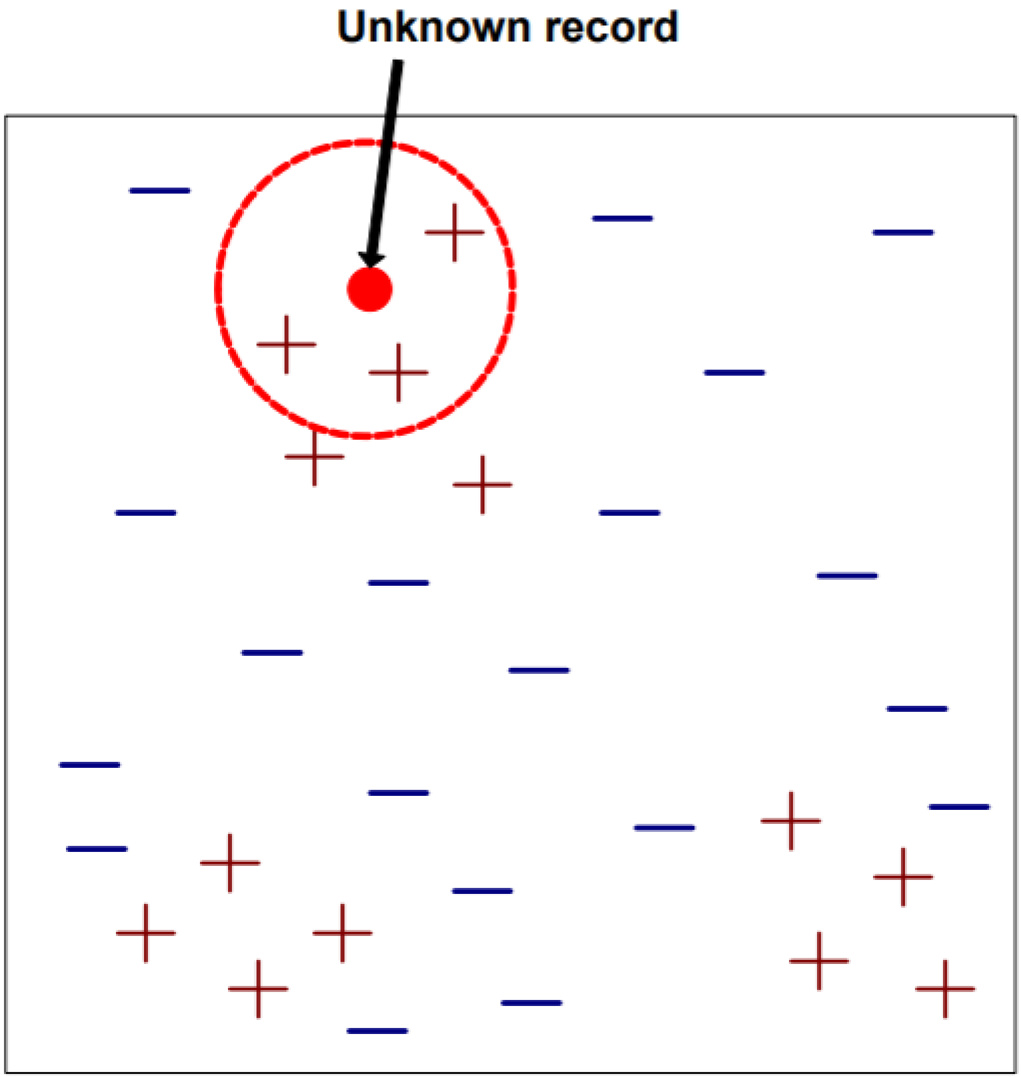
k Nearest Neighbors

- 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



Classification: Application

- 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

- 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:

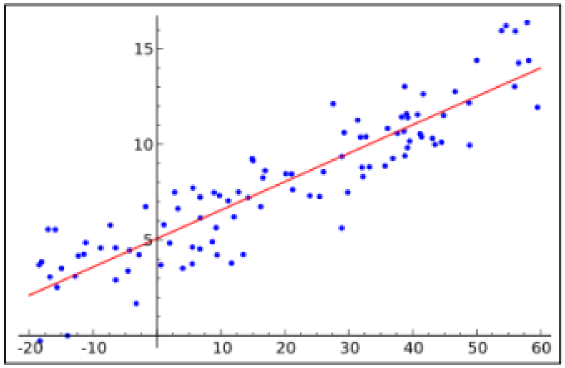
1. Use data from a campaign introducing a similar product in the past
   - we know which customers decided to buy and which decided otherwise
   - this {buy, don't buy} decision forms the class attribute
2. Collect various demographic, lifestyle, and company-interaction related information about the 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

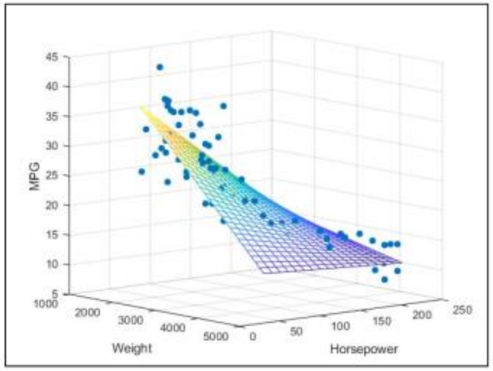
Regression

- 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.

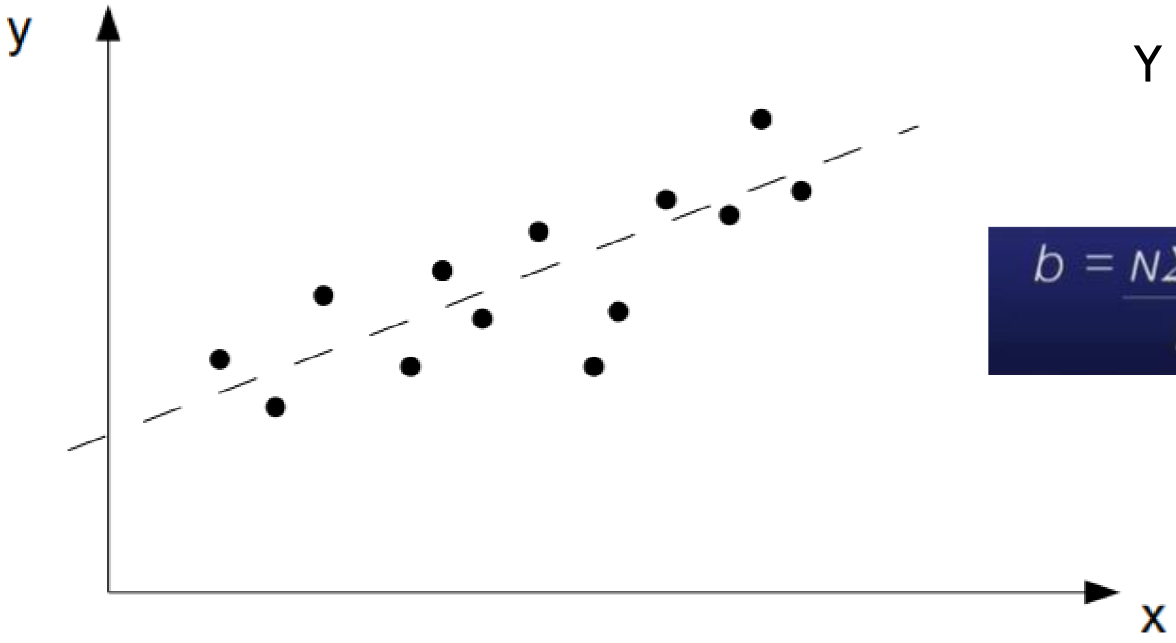




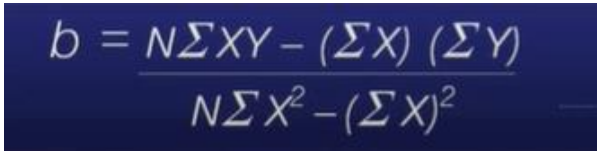
- Difference to classification: The predicted attribute is continuous, while classification is used to predict nominal attributes (e.g. yes/no)

Linear Regression

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



$$Y = a + bX$$



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

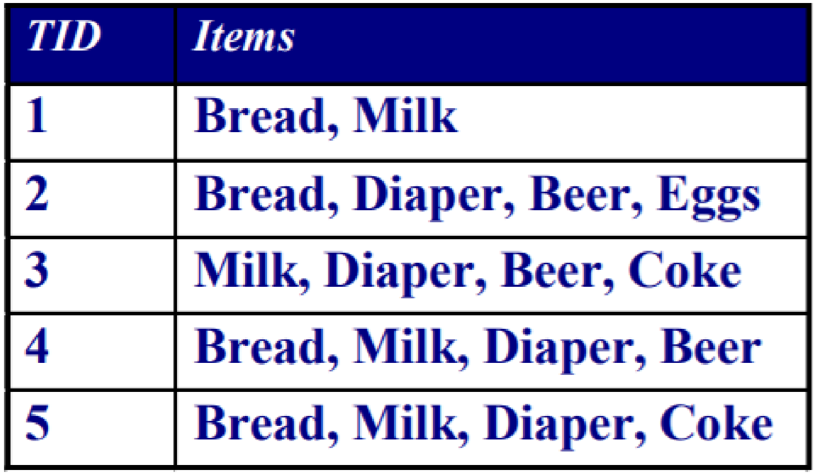
Association Rule

- 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



TID	*Items*
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

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

→ denotes co-occurence, not causality!

Apriori Algorithm

Two-step approach

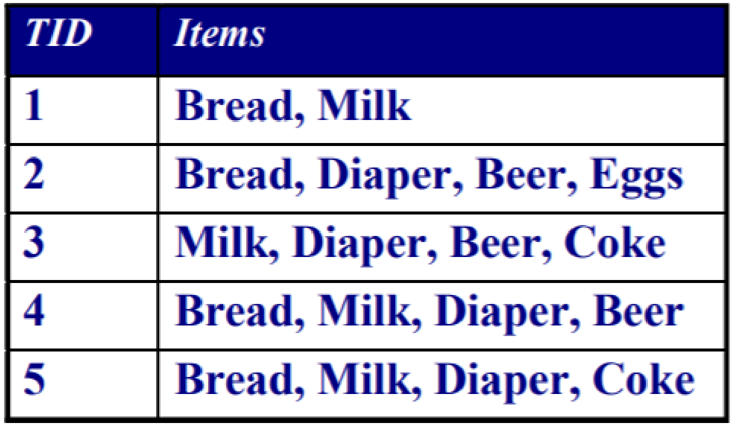
- First: Frequent Itemset Generation
  – Generate all itemsets whose support ≥ minsup

- 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

Support

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



TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

s({Bread}) = 0.8

s({Bread,Milk}) = 0.6

s({Bread,Milk,Diaper}) = 0.4

s({Milk}) = 0.8

s({Milk,Diaper}) = 0.6

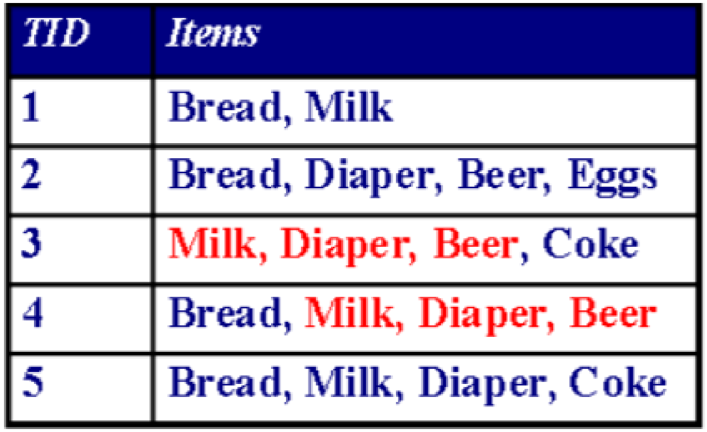
s({Milk,Diaper,Beer}) = 0.4

support ≥ minsup threshold

Apriori Algorithm: Rule Generation

Confidence

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



TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

{Milk,Diaper} → {Beer} c=0.67

{Milk} → {Beer} c=0.5

{Diaper} → {Beer} c=0.8

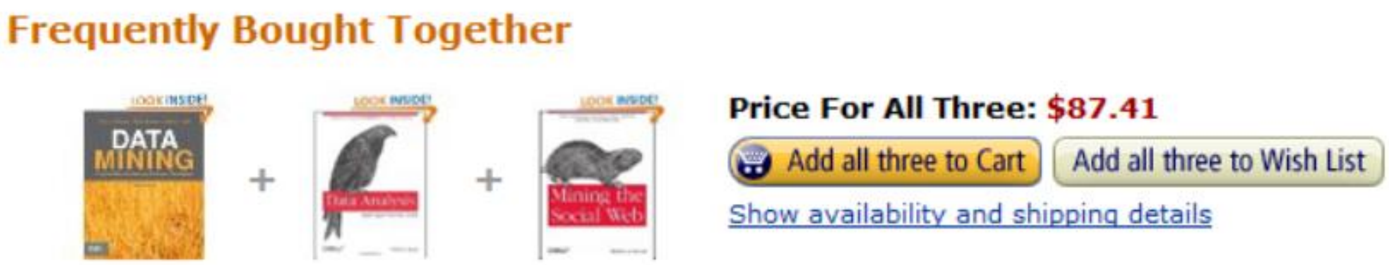
confidence ≥ minconf threshold

Association Rule Discovery: Applications

- Application area: Marketing and Sales Promotion
- Example rule discovered:

  {Bagels, Coke} --> {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

amazon.com



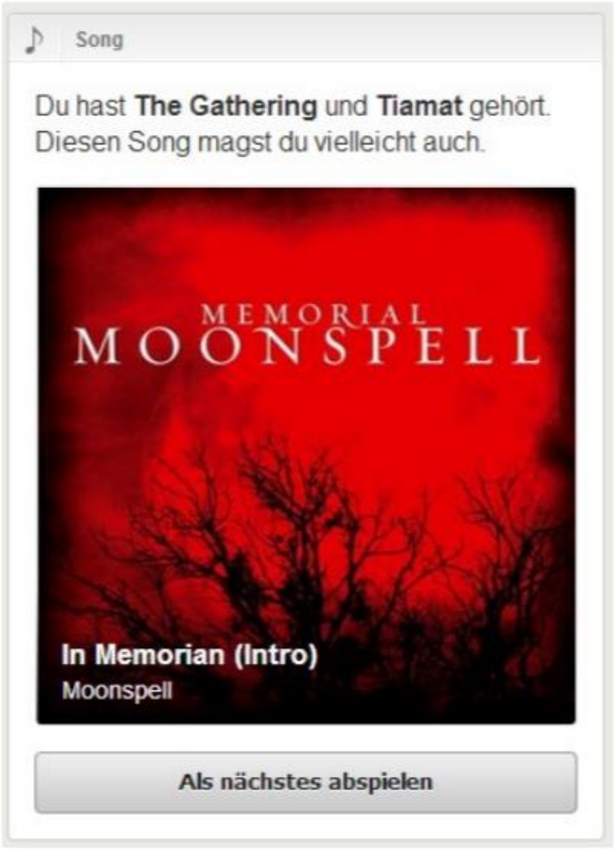
Frequently Bought Together

Price For All Three: $87.41

Association Rule Discovery: Applications

- Customers who bought this product also bought…
  - …do terrorists order bomb building parts on Amazon?

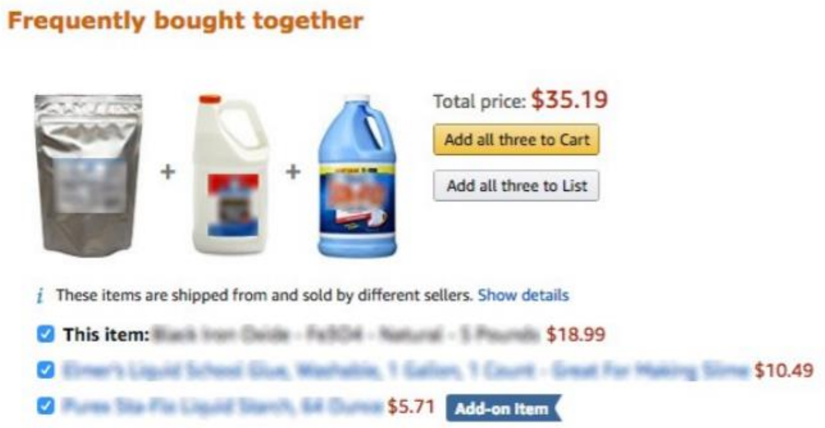
- Content-based recommendation
  - requirement: much data
  - e.g., Amazon transactions, Spotify logfiles



♪ Song

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

In Memorian (Intro)
Moonspell



Frequently bought together

Total price: $35.19

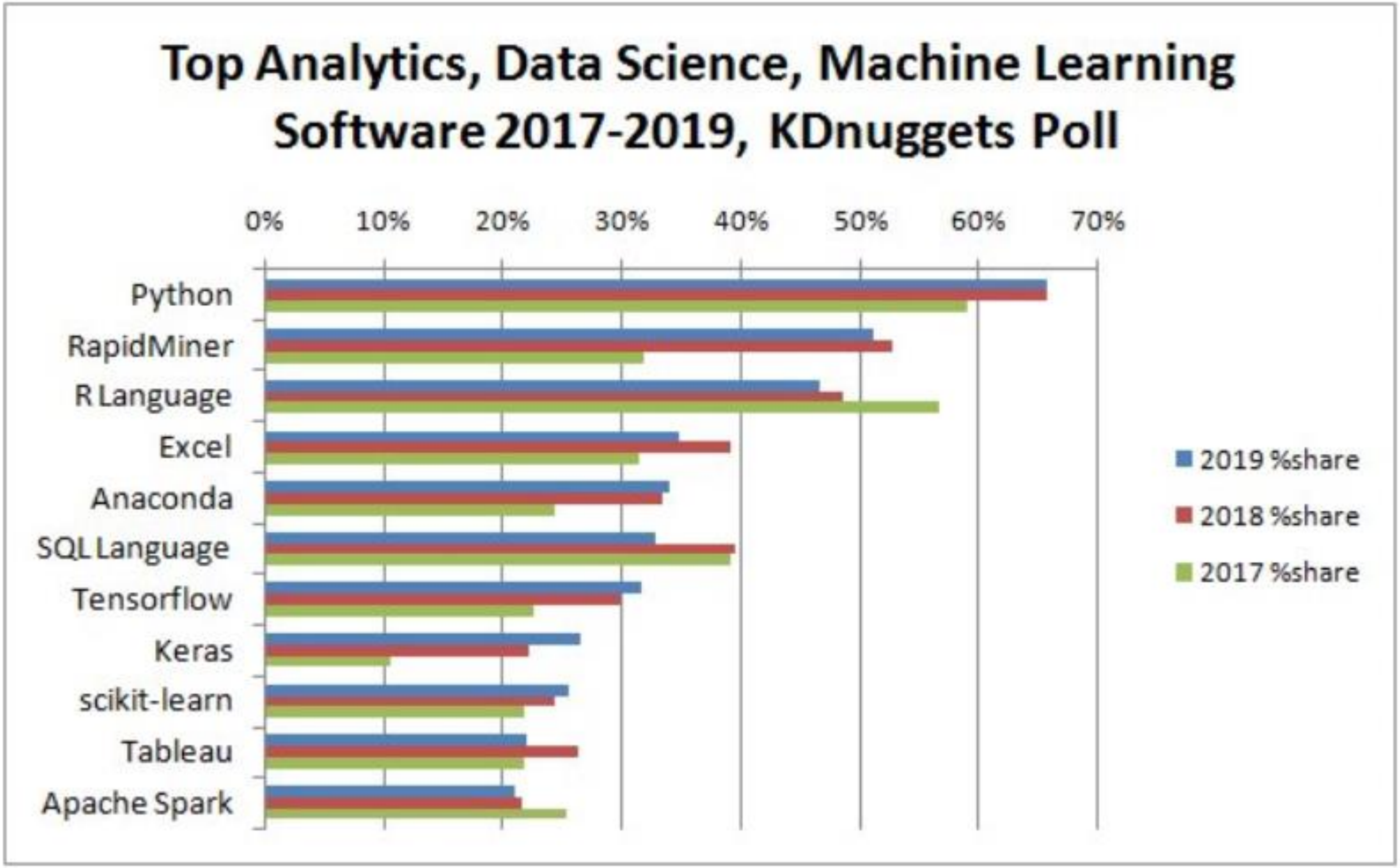
i These items are shipped from and sold by different sellers. Show details

☑ **This item:** [redacted] $18.99

☑ [redacted] $10.49

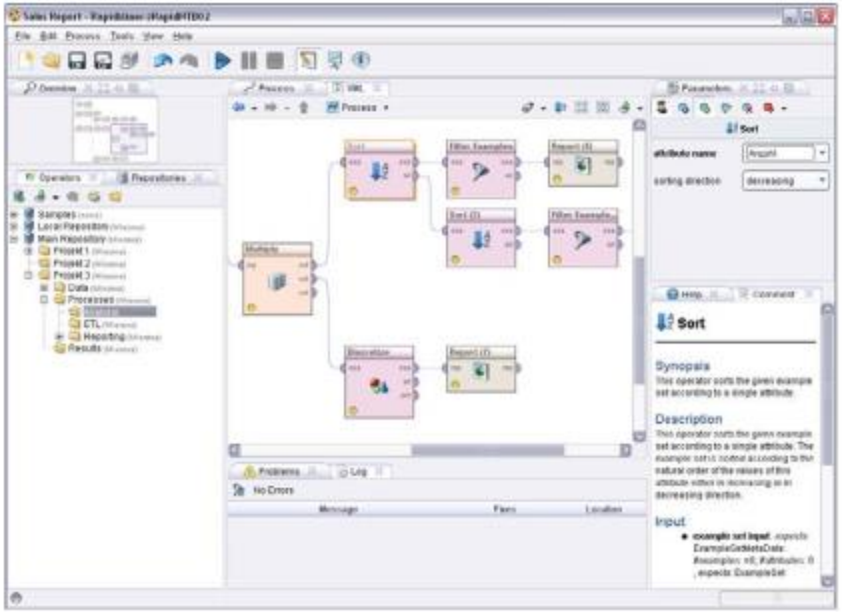
☑ [redacted] $5.71 Add-on Item

Data Mining Software



RapidMiner

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



Sort

attribute name: Amount
sorting direction: decreasing

Sort

Synopsis

This operator sorts the given example set according to a single attribute.

Description

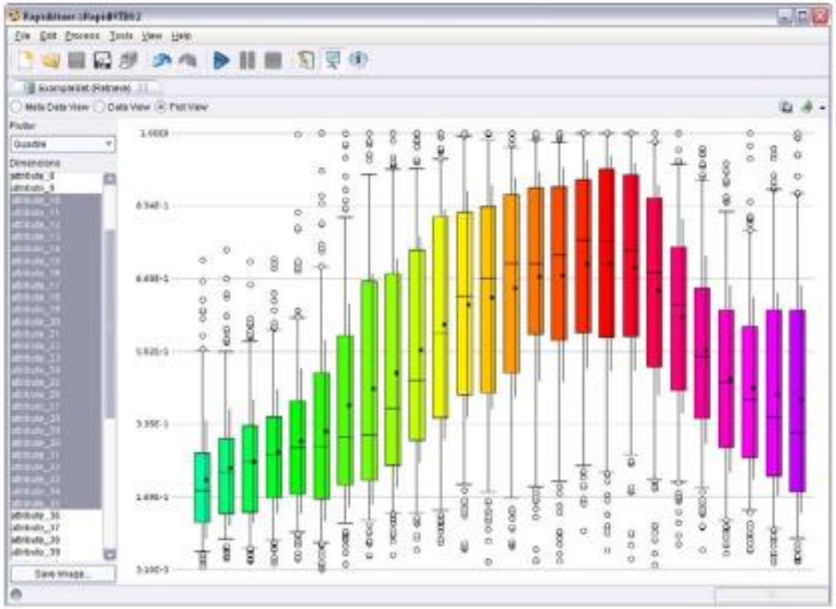
This operator sorts the given example set according to a single attribute. The example set is sorted according to the natural order of the values of this attribute either in increasing or in decreasing direction.

Input

- example set input: expects ExampleSetMetaData: #examples: =0; #attributes: 0, expects ExampleSet

No Errors

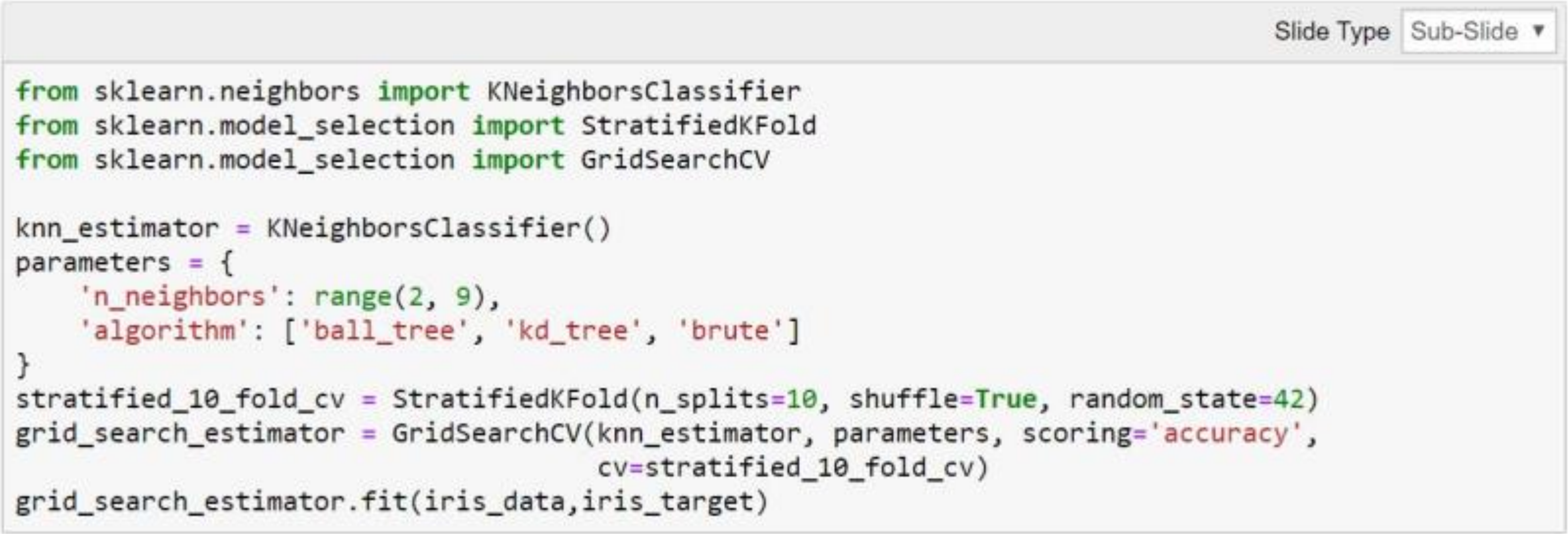
Message	Fixes	Location



Python

We use the Anaconda Python distribution

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



```python
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)
```