Data Mining

Lecture 2

IN THE PREVIOUS EPISODE...

In the previous lecture…

- **What is Data Mining?**
  - Information extraction
  - Data excavation
  - Data intellectual analysis
  - Search for regularities
  - Knowledge extraction
  - Pattern analysis
  - Knowledge Discovery in Databases, KDD
  - Statistics and ML
- **Data**
  - Facts
  - Sources
  - Metadata
- **Methods and stages of Data Mining**
  - Discovery
  - Forecasting
  - Exception analysis

Lecture outline

- Data Mining problems:
  - Information and knowledge.
  - Classification and clustering.
  - Forecasting and visualization

INFORMATION AND KNOWLEDGE

Information and knowledge

Information and knowledge

- Data mining tasks:
  - Classification
  - Clusterization
  - Association
  - Forecasting
  - Visualization

Information and knowledge

- **Classification**
  - Detecting features characterizing group of items in the given dataset – classes. Thus new object can be attributed to a predefined class.
  - Methods:
    - Nearest Neighbor
    - K-Nearest Neighbors
    - Bayesian Networks
    - Decision Tree classifier
    - Neural networks

Information and knowledge

- **Clusterization**
  - Dividing objects into groups undefined beforehand according to the newly discovered common charachteristics.
  - Methods:
    - K-means
    - Agglomerative Clusterization
    - Mean shift
    - Affinity propagation
    - Kochonnen cards

Information and knowledge

- **Association**
  - Uncovering associative rules of the linked objects or events.
  - Methods:
    - Apriori algorithm

Information and knowledge

- **Forecasting**
  - On the basis of analysis of historical data missing or future values are predicted.
  - Methods:
    - Mathematical statistics (regression analysis)
    - Neural networks

Information and knowledge

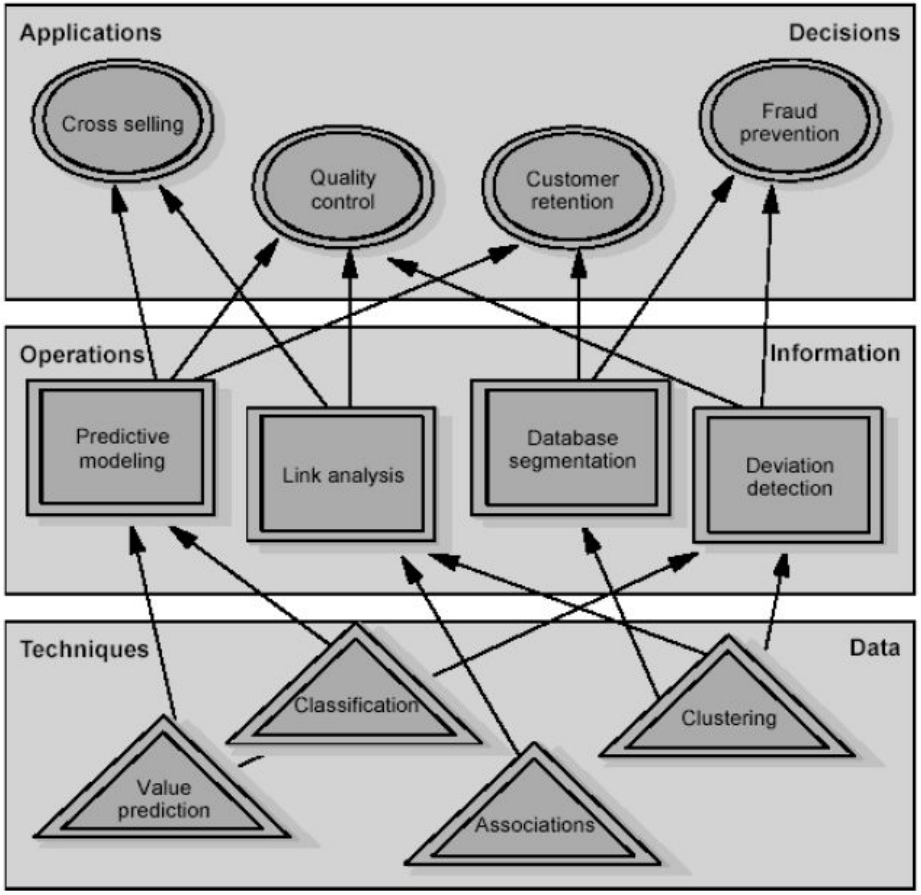
- **Visualization**
  - Creating graphical representation of the analyzed data.
  - Methods:
    - 2-D and 3-D visualizations
    - Graph representations
    - Dendrogramme

Information and knowledge

- **Data Mining tasks classification**
  - **By strategy**
    - Supervised learning
      - Classification
      - Forecasting
    - Unsupervised learning
      - Clusterization
  - By model type
    - Descriptive
      - Informative, summarizing, differentiating data charachteristics
      - Characteristics and comparison
    - Predictive
      - Trend analysis

Information and knowledge

- **From task to application**



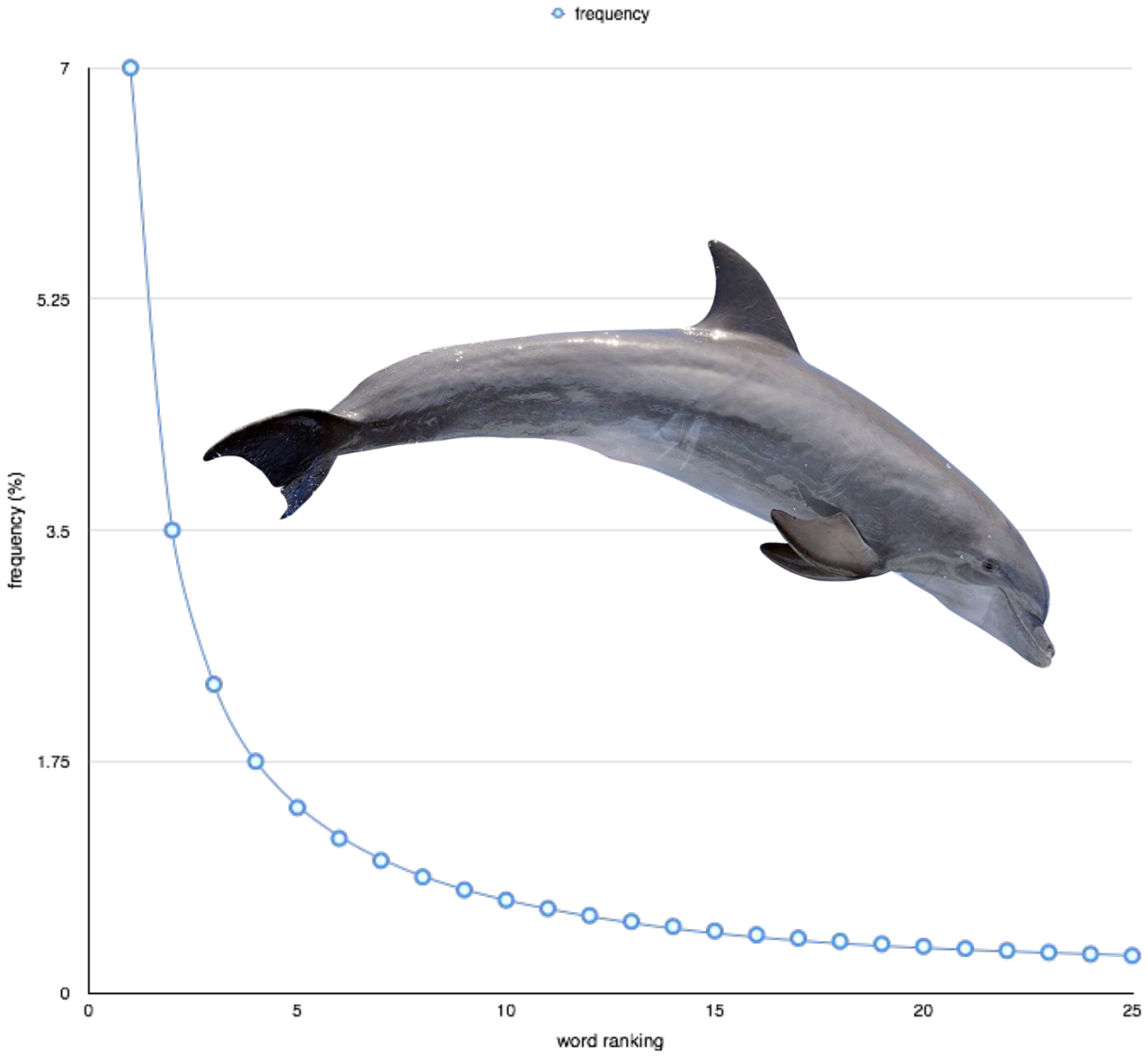
Information and knowledge

- **Information**
  - Any message about anything
  - Intelligence as the object of storage, processing and transfer
  - Quantitative measure of entropy detraction, system organization. Information theory.

https://getpocket.com/explore/item/listening-for-extraterrestrial-blah-blah

Can we tell if aliens are speaking to us?

- SETI project
- Zipf law



Information and knowledge

- **Information properties**
  - Completeness for decision making
  - Trustworthiness
  - Value
  - Adequacy
  - Actuality
  - Clarity
  - Accessibility
  - Subjectivity

Information and knowledge

- **Knowledge**
  - Complex of facts, regularities and heuristic rules helping to solve problems
  - Knowledge evolves on the interconnection of information of different origin
  - Denham Gray “ is the absolute usage of information and data, together with the practical experience potential, abilities, ideas, intuition and beliefs of people.

Information and knowledge

- **Knowledge properties**
  - Structure
  - Easiness of access and digestion
  - Laconicism
  - Non-controversy
  - Processing procedures

CLASSIFICATION AND CLUSTERING

Classification and clustering

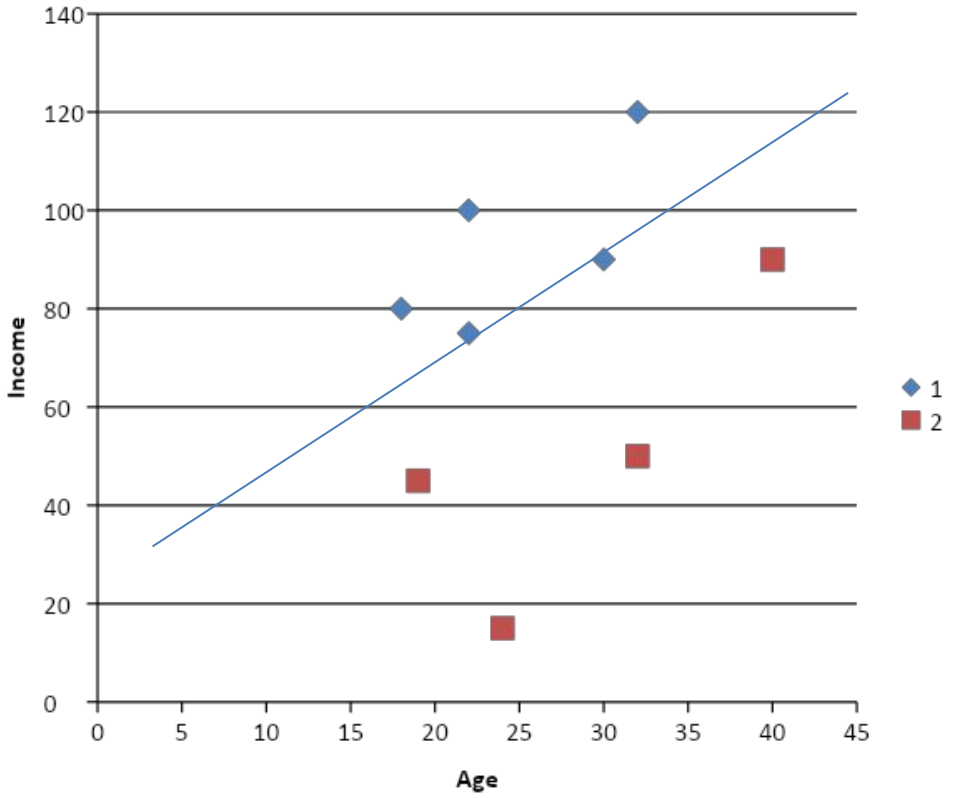
Classification - is a division or category in a system which divides things into groups or types.

- Supervised learning
- Predicting class based on feature vector consisting of continuous and categorical value

Classification and clustering

Classification example

ID	Age	Income	Class
1	18	80	1
2	22	100	1
3	30	90	1
4	32	120	1
5	24	15	2
6	25	22	2
7	32	50	2
8	19	45	2
9	22	75	1
10	40	90	2



Classification and clustering

Classification process

Data

- Preprocess (clean, feat eng)

Train/test split

Training

- Classification models

Testing

- Metrics:
  - Accuracy
  - Precision
  - Recall
  - F1

Application

Classification and clustering

Classification applications

- Face recognition (image)
- OCR (text)
- Text genre detection (text)
- Speaker recognition (sound)

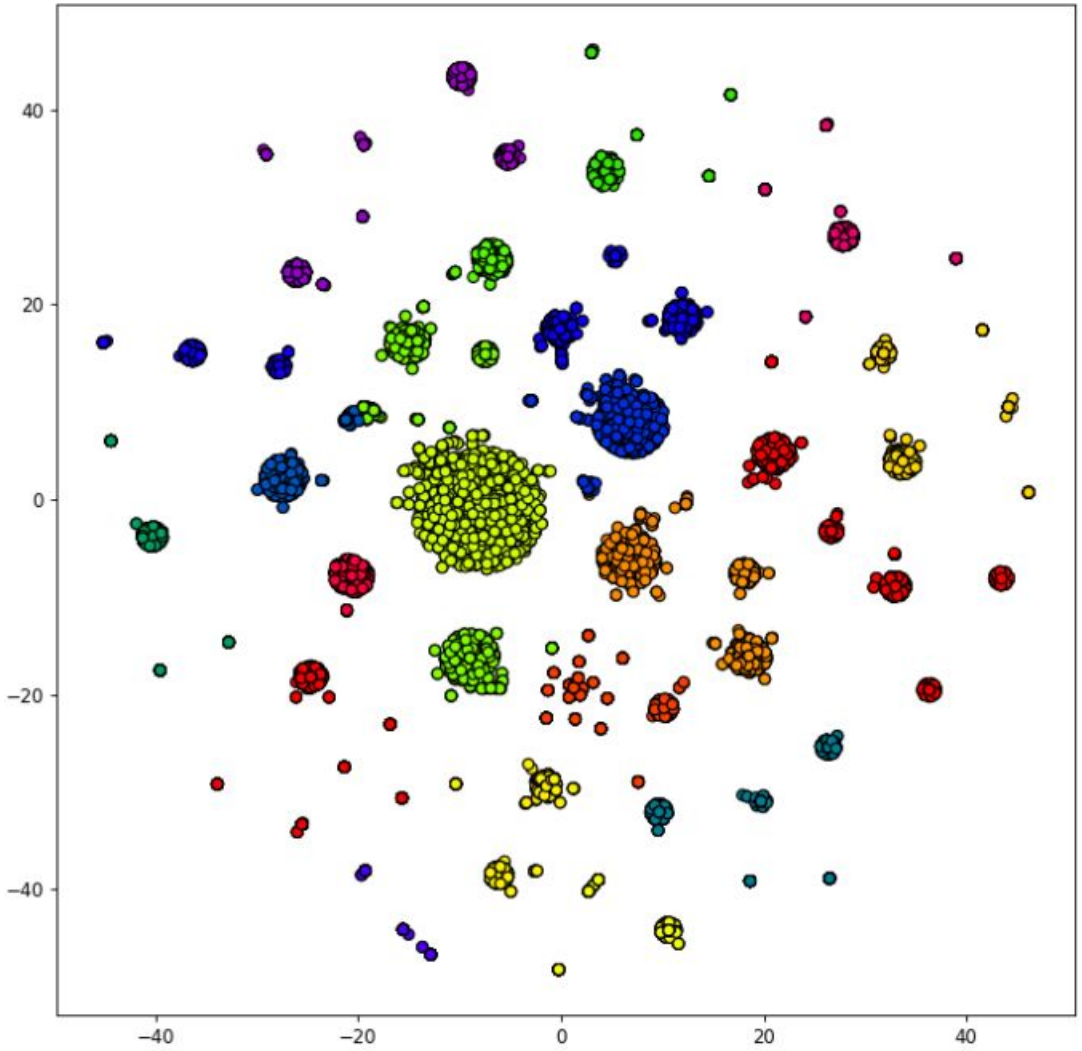
Classification and clustering

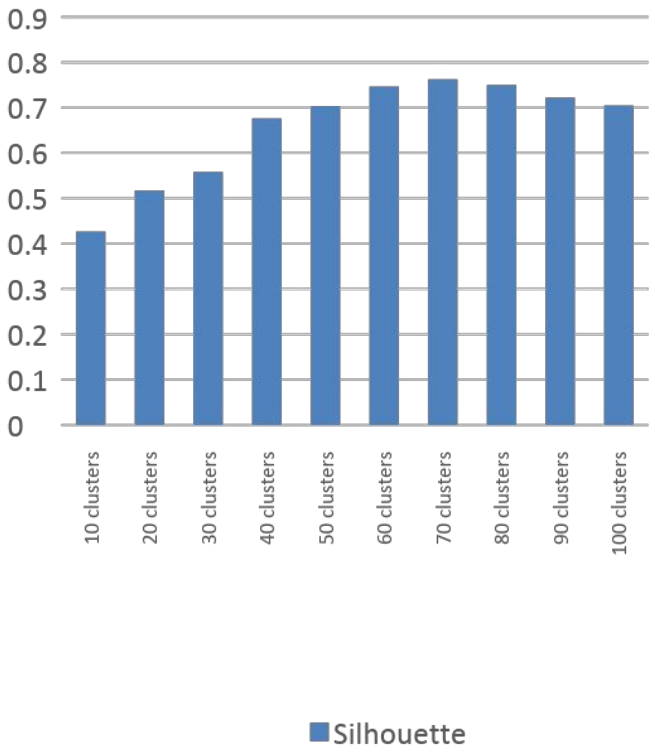
Clustering - is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

- Unsupervised learning
- Attributing a data point to a cluster based on its similarity to other data points with respect to a set of characteristics

Classification and clustering

Clustering example





Classification and clustering

Clustering process

Data

- Preprocess (clean, feat eng)

Train/test split

Training

- Clustering models

Testing

- Metrics:
  - Silhouette
  - Jackard measure

Application

Classification and clustering

Clustering applications

- Topic modeling (texts)
- Text to speech (sounds)
- Client base clustering (business)

FORECASTING AND VISUALIZATION

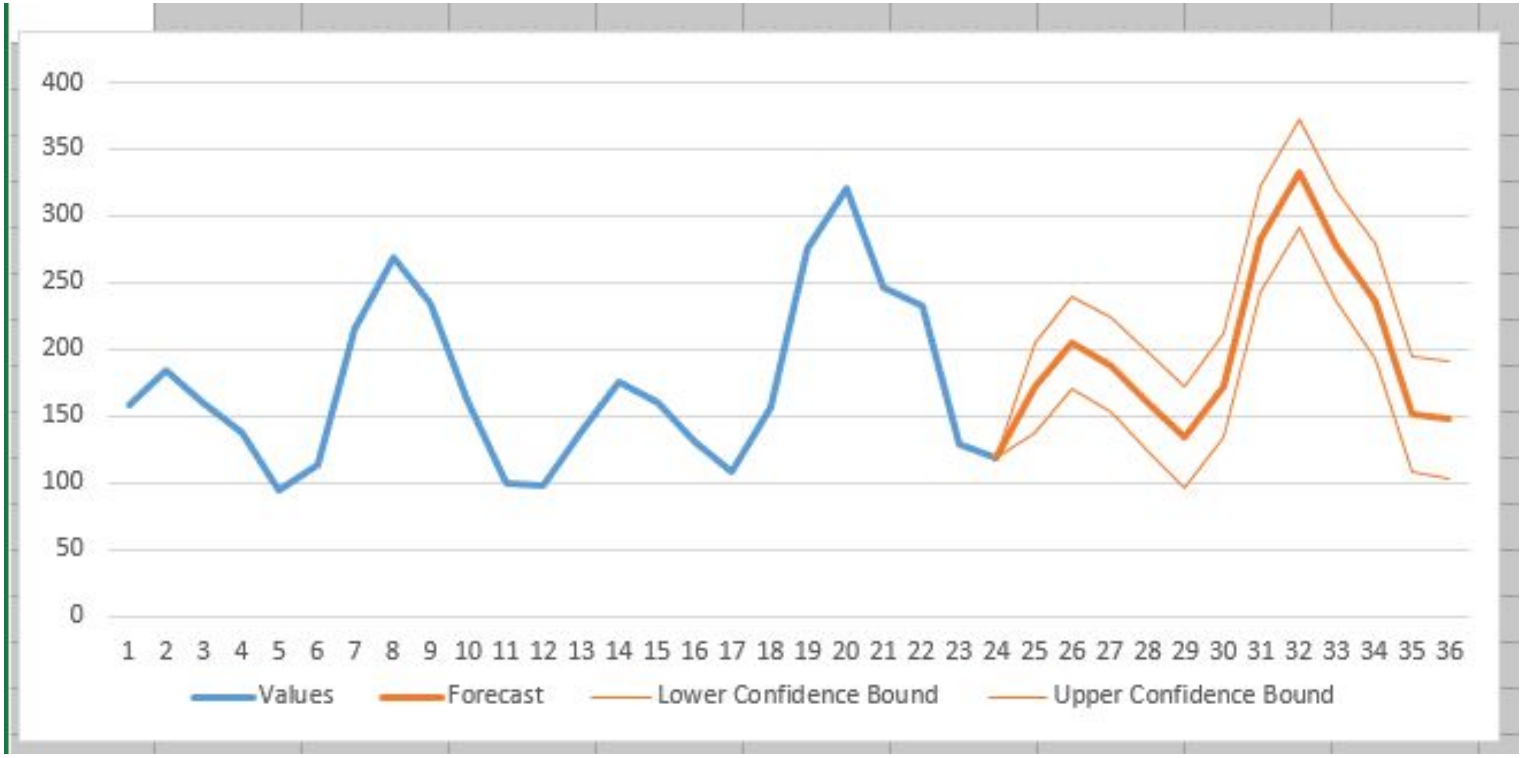
Forecasting and visualization

Forecasting - is the process of making predictions of the future based on past and present data and most commonly by analysis of trends. A commonplace example might be estimation of some variable of interest at some specified future date. Prediction is a similar, but more general term.

- Supervised learning

Forecasting and visualization

Forecasting example



Forecasting and visualization

Forecasting process

Data

- Preprocess (clean, feat eng)

Train/test split

Training

- Forecasting models
  - Regression
  - ARIMA

Testing

- Metrics:
  - R2
  - MAE
  - MSE

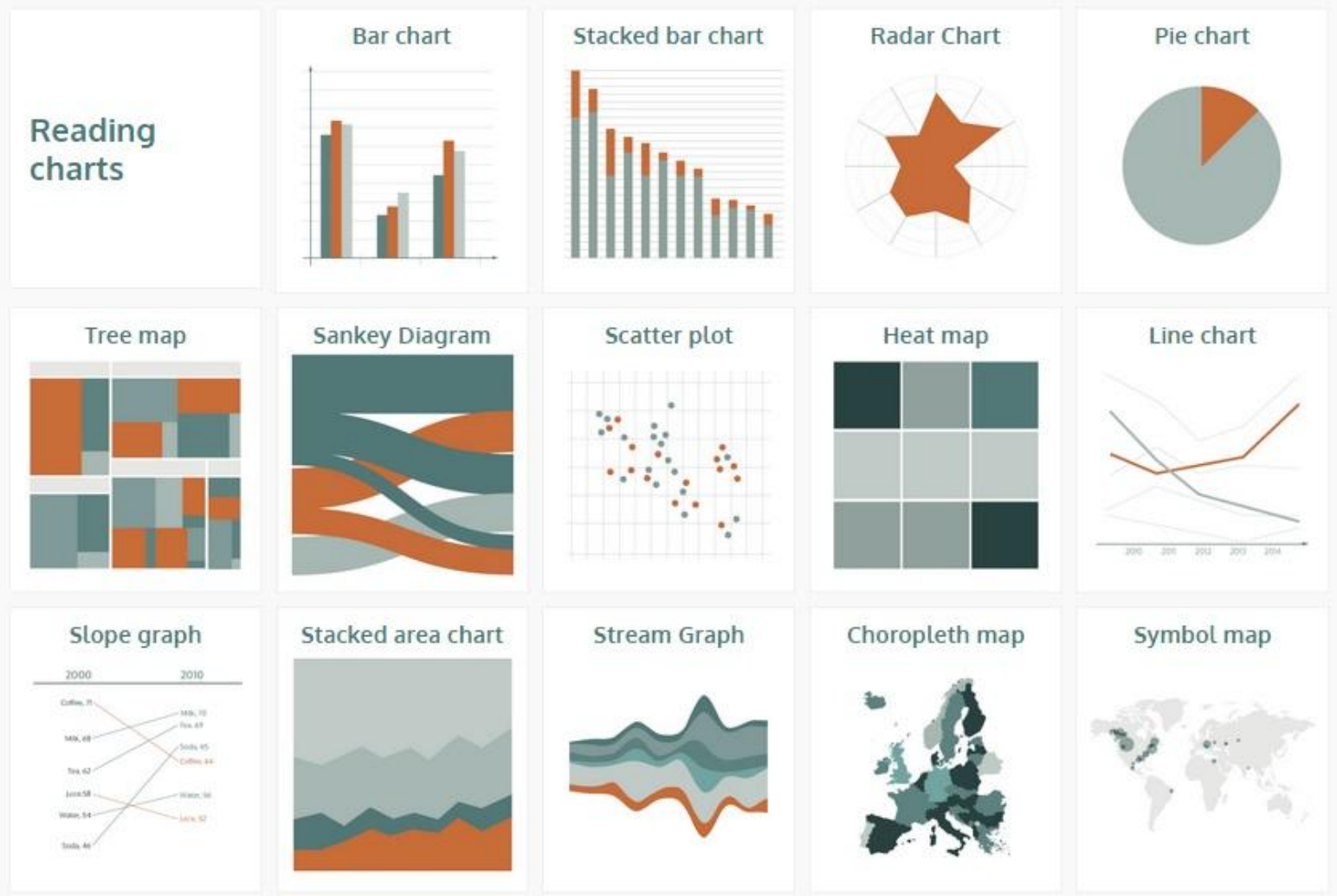
Application

Forecasting and visualization

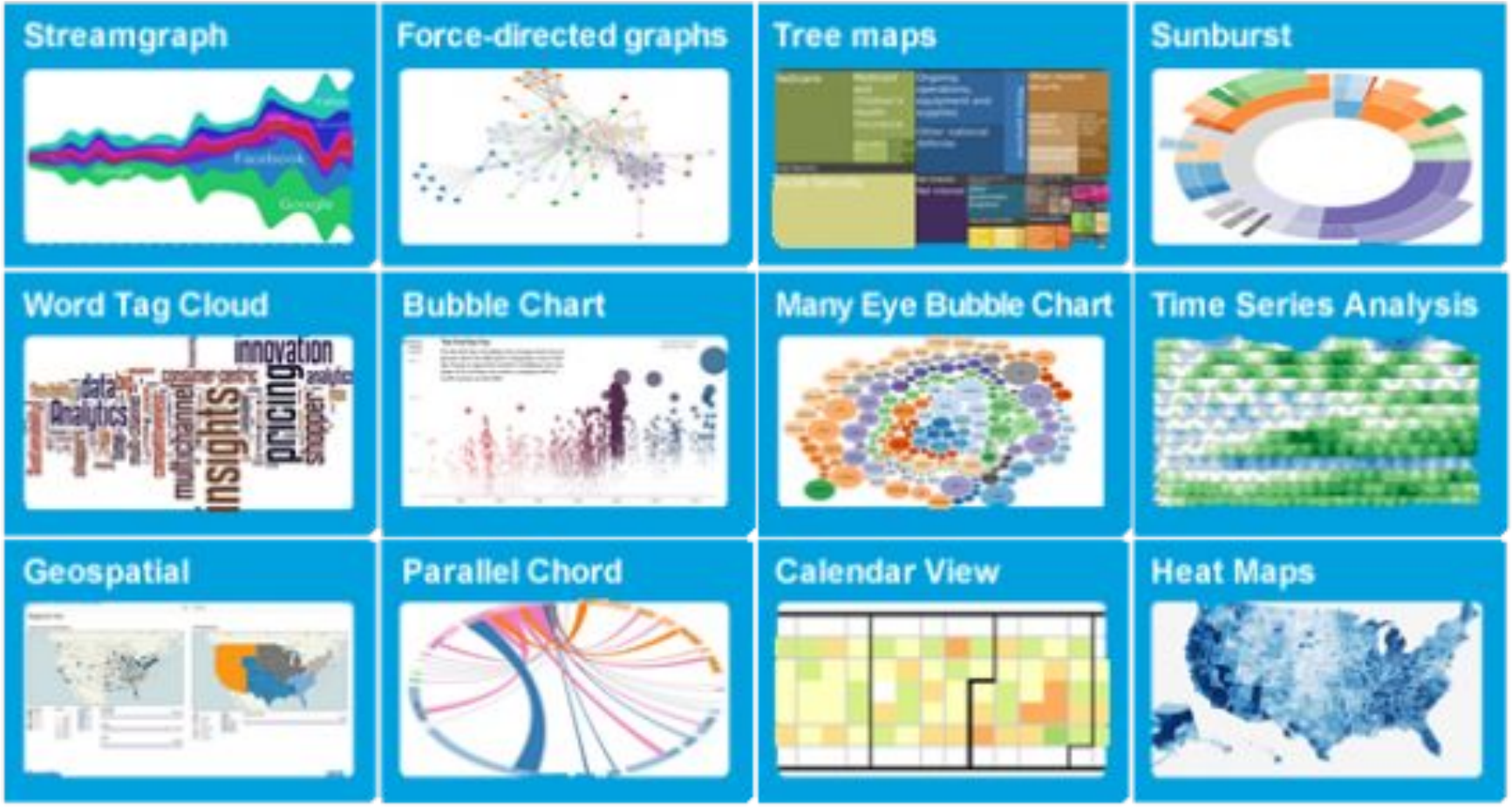
Forecasting application

- Pricing (cars, real estate)
- Price movements (time series)
- Missing values and interpolation
- Revenue predicts (business)

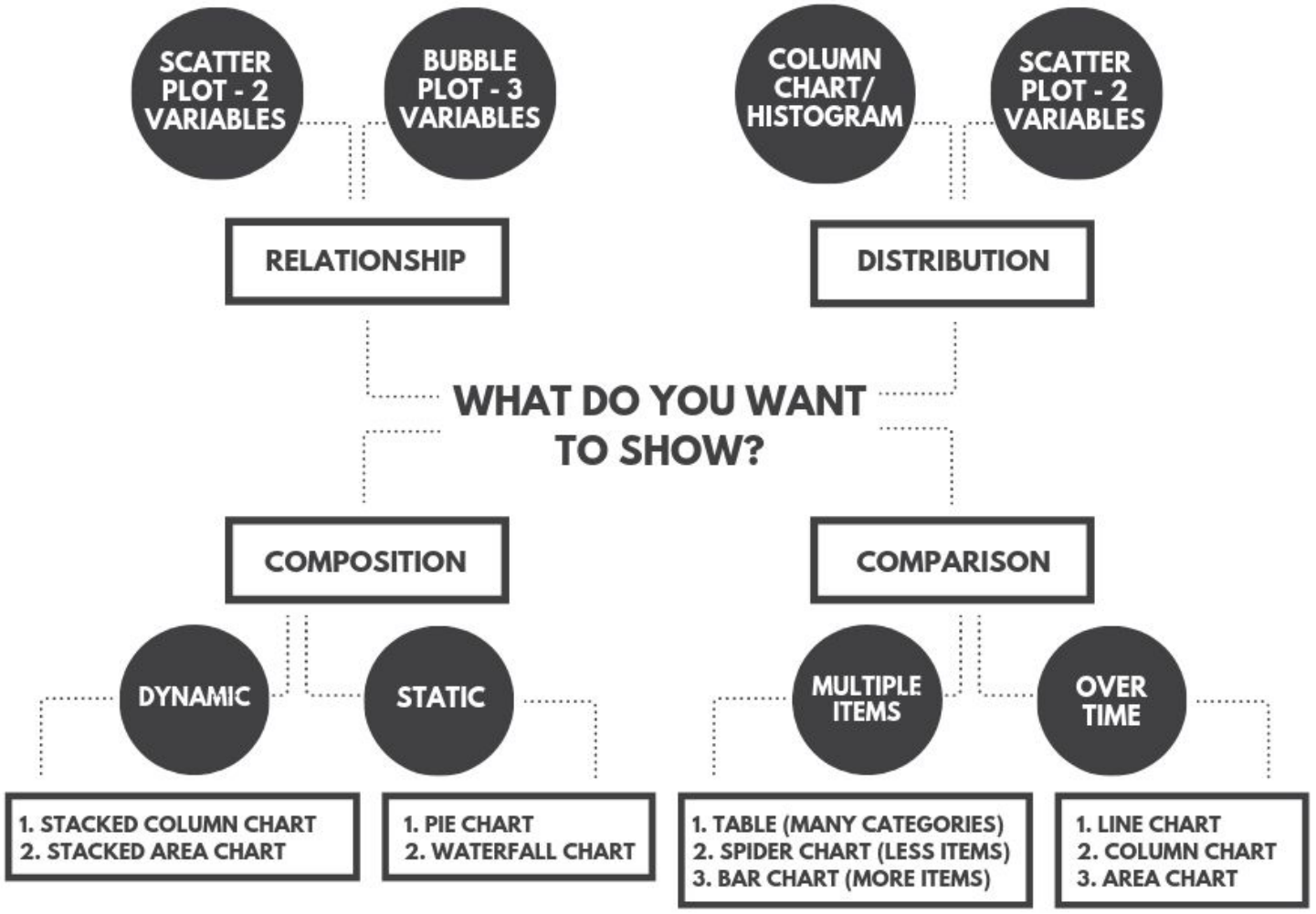
Forecasting and visualization



Forecasting and visualization



Forecasting and visualization



Summary

- Data Mining problems:
  - Information and knowledge.
    - Data-Information-Knowledge
    - Support decision making process
  - Classification and clustering.
  - Forecasting and visualization