Motivation of data mining

● There is often information “hidden” in the data that is not readily evident

● Human analysts may take a long time to discover useful information

● Much of the data is never analyzed at all

* The Data Gap: Data growth against the number of analysts. In spite of the large increase in data, the number of FTE (Full Time Equivalent) involved in data analysis remains in practice constant.

orange line: data growth

blue line: number of analysts

**Classification**

* Given a collection of records
  + Each record contains a set of *attributes*, one of the attributes is the *class*.
* Find a *model* for class attribute as a function of the values of other attributes.
* Goal: previously unseen records should be assigned a class as accurately as possible.

**Decision** **tree**

* Find the best split
* Measure the node impurity
  + Gini index
  + Entropy
  + Misclassification Error

**Association Rule Discovery**

Supermarket shelf management.

– Goal: To identify items that are bought together by sufficiently many customers.

**Open Source Data Mining Tools**

KNIME: Written in Java and based on Eclipse, KNIME is easy to extend and to add plugins

RapidMiner: Written in the Java Programming language, this tool offers advanced analytics through template-based frameworks.

R-Programming: It’s primarily written in C and Fortran. And a lot of its modules are written in R itself.

Orange: a Python-based tools with components for machine learning