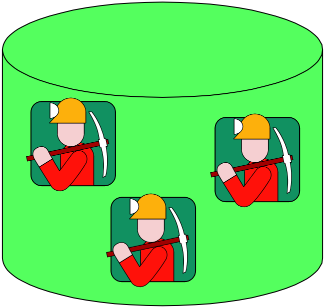
Data Mining 2013
Introduction

Ad Feelders

Universiteit Utrecht



The Course

- Literature: Lecture Notes, Articles, Slides (appear on the course web site).
- Course Form: Lectures and computer lab sessions with R (weeks 38-44).
- Grading: Two practical assignments and a written exam.
- Web Site: http://www.cs.uu.nl/docs/vakken/mdm/

What is Data Mining?

- (Knowledge discovery in databases) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al)
- Analysis of *secondary* data (Hand)
- The induction of understandable models and patterns from databases (Siebes)
- The data-dependent process of selecting a statistical model (Leamer, 1978 (!))

What is Data Mining?

Data Mining as a subdiscipline of computer science:

is concerned with the development and analysis of algorithms for the (efficient) extraction of patterns and models from (large) data bases.

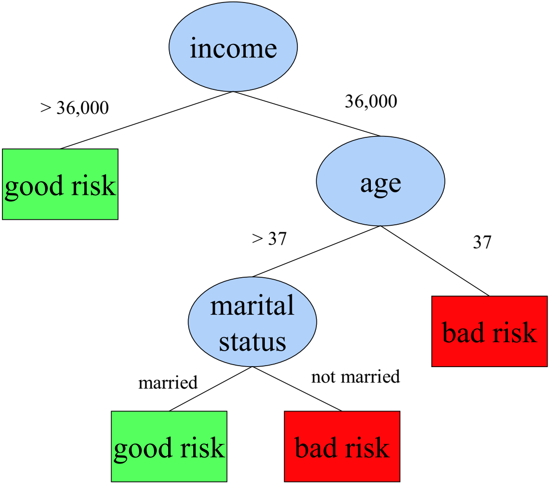
Models

A model is an abstraction of (part of) reality (the application domain).

In our case, models describe relationships among:

- attributes (variables, features),
- tuples (records, cases),
- or both.

Example Model: Classification Tree



Patterns

Patterns are **local models**, i.e., models that describe only part of the database.
For example, association rules:

$$Diapers \rightarrow Beer, \; support = 20\%, \, confidence = 85\%$$

Although patterns are clearly different from models, we will use *model* as the generic term.

Diapers → Beer

Reasons to Model

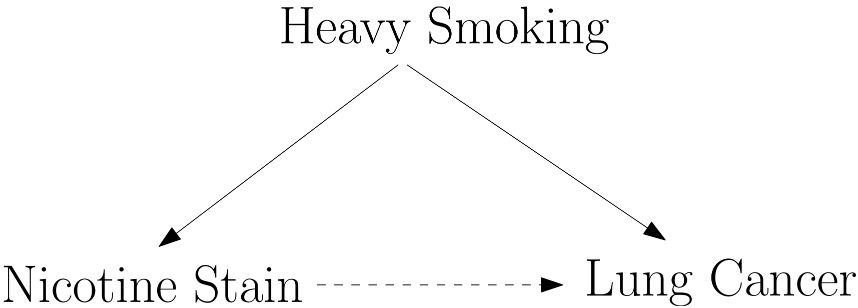
A model

- helps to gain *insight* into the application domain
- can be used to make *predictions*
- can be used for *manipulating/controlling* a system (causality!)

A model that predicts well does not always provide understanding.

Correlation $\neq$ Causation
Can causal relations be found from data alone?

Causality and Correlation



Washing your hands won't help!

Induction vs Deduction

Deductive reasoning is *truth-preserving*:

1. All horses are mammals
2. All mammals have lungs
3. Therefore, all horses have lungs

Inductive reasoning *adds information*:

1. All horses observed so far have lungs.
2. Therefore, all horses have lungs.

Induction (Statistical)

1. 4% of the products we tested are defective
2. Therefore, 4% of all products (tested or otherwise) are defective

Inductive vs Deductive: Example

100,000 products, sample 1000
Suppose 10 are defective (1% of the sample)

Deductive: $d \in [0.0001, 0.9901]$

Inductive: $d \in [0.004, 0.016]$ with 95% confidence.

$$0.01 \pm \underbrace{\sqrt{\frac{0.01 \times 0.99}{1000}} \times 1.96}_{\approx 0.006}$$

Experimental data

The experimental method:

- Formulate a hypothesis of interest.
- Design an experiment that will yield data to test this hypothesis.
- Accept or reject hypothesis depending on the outcome.

Experimental vs Observational Data

Experimental Scientist:

- Assign level of fertilizer randomly to plot of land.
- Control for other factors that might influence yield: quality of soil, amount of sunlight,...
- Compare mean yield of fertilized and unfertilized plots.

Data Miner:

- Notices that the yield is somewhat higher under trees where birds roost.
- Conclusion: droppings increase yield.
- Alternative conclusion: moderate amount of shade increases yield.("Identification Problem")

Observational Data

- In observational data, many variables may move together in systematic ways.
- In this case, there is no guarantee that the data will be “rich in information”, nor that it will be possible to isolate the relationship or parameter of interest.
- Prediction quality may still be good!

Example: linear regression

$$\widehat{\text{mpg}} = a + b \times \text{cyl} + c \times \text{eng} + d \times \text{hp} + e \times \text{wgt}$$

Estimate a, b, c, d, e from data. Choose values so that sum of squared errors

$$\sum_{i=1}^{N} (\text{mpg}_i - \widehat{\text{mpg}}_i)^2$$

is minimized.

$$\frac{\partial \widehat{\text{mpg}}}{\partial \text{eng}} = c$$

Expected change in mpg when (all else equal) engine displacement increases by one unit.

The Data

```
> cars.dat[1:10,]
   mpg cyl eng  hp  wgt
1   18   8 307 130 3504 "chevrolet chevelle malibu"
2   15   8 350 165 3693 "buick skylark 320"
3   18   8 318 150 3436 "plymouth satellite"
4   16   8 304 150 3433 "amc rebel sst"
5   17   8 302 140 3449 "ford torino"
6   15   8 429 198 4341 "ford galaxie 500"
7   14   8 454 220 4354 "chevrolet impala"
8   14   8 440 215 4312 "plymouth fury iii"
9   14   8 455 225 4425 "pontiac catalina"
10  15   8 390 190 3850 "amc ambassador dpl"
```

Engine displacement is defined as the total volume of air/fuel mixture an engine can draw in during one complete engine cycle.

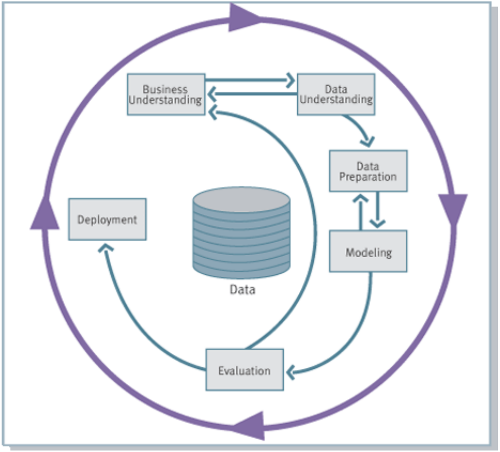
Fitted Model

```
Coefficients:
              Estimate Pr(>|t|)
(Intercept) 45.7567705  < 2e-16 ***
cyl         -0.3932854 0.337513
eng          0.0001389 0.987709
hp          -0.0428125 0.000963 ***
wgt         -0.0052772 1.08e-12 ***
---

Multiple R-Squared: 0.7077

> cor(cars.dat)
           mpg        cyl        eng         hp        wgt
mpg  1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442
cyl -0.7776175  1.0000000  0.9508233  0.8429834  0.8975273
eng -0.8051269  0.9508233  1.0000000  0.8972570  0.9329944
hp  -0.7784268  0.8429834  0.8972570  1.0000000  0.8645377
wgt -0.8322442  0.8975273  0.9329944  0.8645377  1.0000000
```

KDD Process: CRISP-DM



This course is primarily concerned with methods and algorithms for the modeling phase.

Data Preparation

Once we know the data available to solve the data mining problem, we have to prepare the mining database (often one table):

- Select data
- Clean data
- Construct data
- Integrate data

Select Data: Why Reject Data?

- attributes with too many errors
- attributes with too many missing values
- attributes that have no relevance from a domain experts point of view
- a sample may be sufficient

Data Cleaning

Cleaning data is a complete topic in itself, we mention two problems:

1. data editing: what to do when records contain *impossible* combinations of values?
2. incomplete data: what to do with missing values?

Data Editing: Example

We have the following edits (impossible combinations):

$E_1 = \{$*Driver's Licence=yes, Age* $<$ *18*$\}$
$E_2 = \{$*Married=yes, Age* $<$ *18*$\}$

Make record

Driver's Licence=yes, Married=yes, Age=15

consistent with minimal number of changes.

Of course it's better to *prevent* such inconsistencies in the data!

Seminal Paper: I.P. Fellegi, D. Holt: A systematic approach to automatic edit and imputation, Journal of the American Statistical Association 71(353), 1976, pp. 17-35.

What to do with missing values?

- One can remove a tuple if one or more attribute values are missing. Danger: how representative and random is the remaining sample? Also, you may have to throw away a large part of the data!
- One can remove attributes for which values are missing. Danger: this attribute may play an important role in the model you want to induce.
- You do *imputation*, i.e., you fill in a value. Danger: the values you guess may have a large influence on the resulting model.

Missing Data Mechanisms: MCAR

Suppose we have data on gender and income.

Gender is fully observed, income is sometimes missing.

MCAR: Income is missing completely at random.

$$\Pr(\text{income} = ?) = 0.1$$

- No big deal.
- Remove tuples with income missing.
- No bias.

Missing Data Mechanisms: MAR

Probability that income is missing depends on gender.

$$\begin{aligned} \Pr(\text{income} = ?|\text{gender=male}) &= 0.2 \\ \Pr(\text{income} = ?|\text{gender=female}) &= 0.05 \end{aligned}$$

- Not so nice, but no disaster.

Missing Data Mechanisms: MAR

Probability that income is missing depends on gender.

$$
\begin{aligned}
\Pr(\text{income} = ?|\text{gender=male}) &= 0.2 \\
\Pr(\text{income} = ?|\text{gender=female}) &= 0.05
\end{aligned}
$$

- Not so nice, but no disaster.
- For example: use observed data to estimate mean income for each gender and fill in this conditional mean.

Missing Data Mechanisms: MNAR

Probability that income is missing depends on value of income itself.

$$\begin{aligned}
\Pr(\text{income} = ?|\text{income} > 8000) &= 0.4 \\
\Pr(\text{income} = ?|2000 < \text{income} < 8000) &= 0.01 \\
\Pr(\text{income} = ?|\text{income} < 2000) &= 0.2
\end{aligned}$$

- We can't "repair" this unless we have knowledge about the missing data mechanism.

Missing Data

- Unfortunately, we cannot infer *from the observed data alone* whether or not the missing data mechanism is MAR.
- We might have knowledge about the nature of the missing data mechanism however ...
- Example: Credit Scoring.

Construct Data

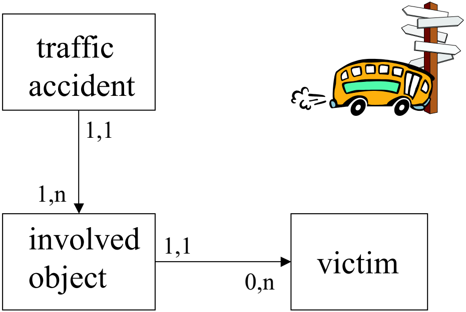
Quite often, the data you have is not the data you want to analyze, for example:

- you have dates of birth and you guess age plays a role.
- you have data on income and fixed expenses and you think disposable income is important.
- you have to analyze text data, for example the body of e-mail messages to develop a spam filter.

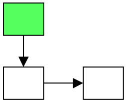
Integrate Data: The Mining Table

- integration of data from different databases, for example:
  - the ‘same’ attribute has two different names
  - two attributes with the same name have a completely different meaning
  - the ‘same’ attribute has two different domains.
- One-one relationships between tables are easy, but what to do with 1-n relationships?

One to Many Relationships: traffic accidents



Traffic accidents: flattening the data



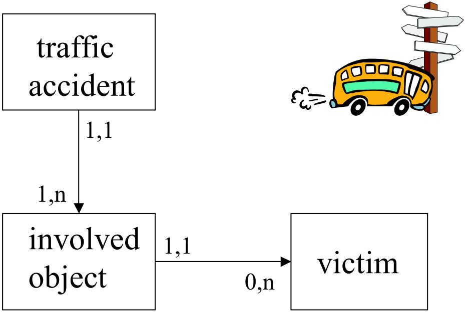
“flatten”

traffic accident

town area?
weather
lighting
maximum speed
involved objects
…
outcome (lethal/victims/ material damage)

Multi-Relational Data Mining

Multi-relational data mining algorithms can handle multiple tables with one-to-many relationships and thus can diminish the required amount of preprocessing.

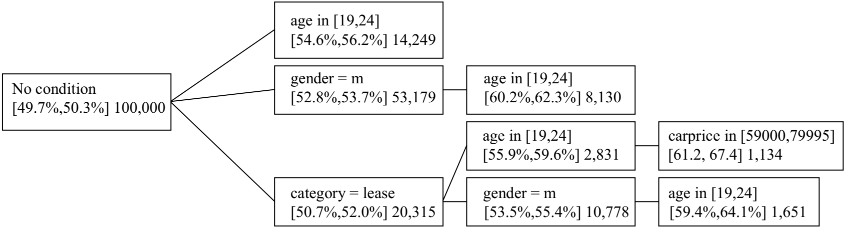


Modeling: Data Mining Tasks

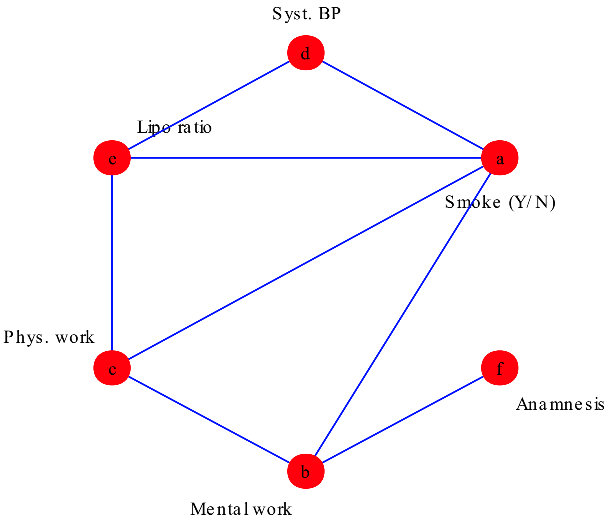
- Classification / Regression
- Dependency Modeling (Graphical Models; Bayesian Networks)
- Frequent Pattern Mining (Association Rules)
- Subgroup Discovery (Rule Induction; *Bump-hunting*)
- Clustering
- Ranking

Subgroup Discovery

Find groups of objects (persons, households, transactions, ...) that score relatively high (low) on a particular *target* attribute.

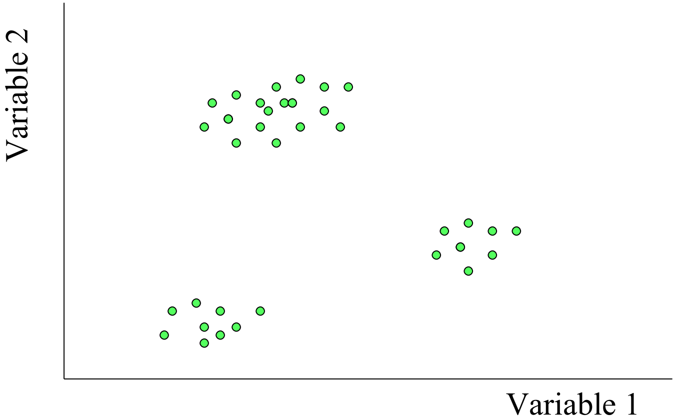


Dependency Modeling: Coronary Heart Disease



Clustering

Put objects (persons, households, transactions, ...) into a number of groups in such a way that the objects within the same group are similar, but the groups are dissimilar.



Ranking

For example:

- Rank web pages with respect to their relevance to a query.
- Rank job applicants with respect to their suitability for the job.
- Rank loan applicants with respect to default risk.
- ...

Has similarities with regression and classification, but in ranking we are often only interested in the *order* of objects.

Components of DM algorithms

Data Mining Algorithms can be regarded as consisting of the following components:

1. A representation language: what models are we looking for?
2. A quality function: when do we consider a model to be good?
3. A search algorithm: how de we find some or all good or best models?

Representation Languages

Representation languages define the set of all possible models, e.g.,

- linear models: $y = b_0 + b_1 x_1 + \cdots + b_n x_n$
- association rules: $X \to Y$
- subgroups: $X_1 \in V_1 \wedge \cdots \wedge X_n \in V_n$
- classification trees
- Bayesian networks (DAGs)

In multi-relational mining these languages are upgraded from "propositional" to "first-order" languages.

Quality Functions

The quality score of a model often contains two elements:

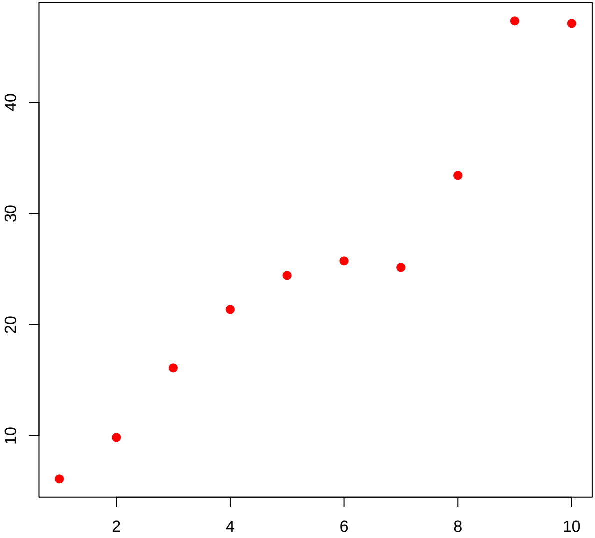
- How well does the model fit the data?
- How complex is the model?

For example (regression)

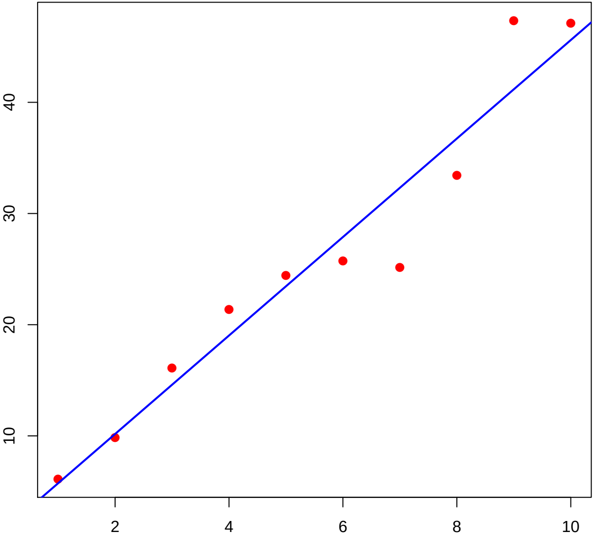
$$\text{score} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + 2 \times \# \text{ parameters}$$

If independent test data is used, the quality score usually only considers the fit on the test data.

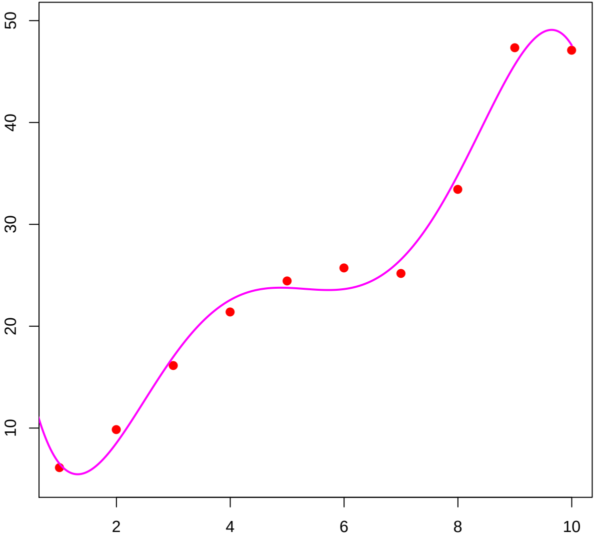
The data



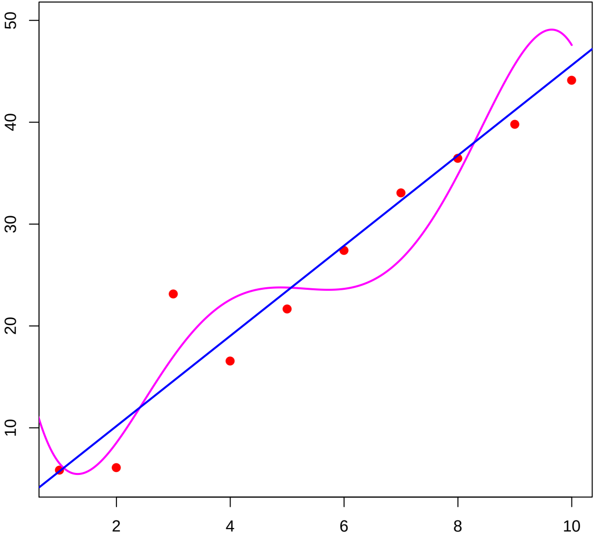
Fitting a linear model



A fifth degree polynomial fits better



Over-fitting: linear model generalizes better



Search for the good models

- Sometimes we can check all possible models, because there are rules with which to prune large parts of the search space; for example, the "a priori principle" in frequent pattern mining.
- Usually we have to employ *heuristics*
  - A general search strategy, such as a hill-climber or a genetic (evolutionary) algorithm.
  - Search operators that implement the search strategy on the representation language. Such as, a neighbour operator for hill climbing and cross-over and mutation operators for genetic search.

Search: example

In linear regression, we want to predict a numeric variable y from a set of predictors $x_1, \ldots, x_n$. We might include any subset of predictors, so the search space contains 2^n models. E.g., if $n = 30$, we have $2^{30} = 1,073,741,824$, i.e. about one billion models in the search space.

It is common to use a hill-climbing approach called stepwise search:

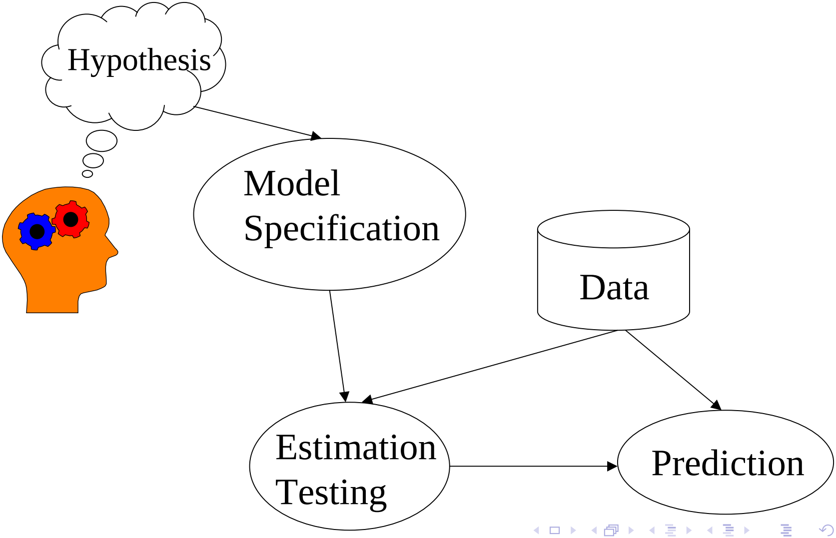
1. Start with some initial model, e.g. $y = b_0$, and compute its quality.
2. Neighbours: add or remove a variable from the right-hand-side.
3. If all neighbours have lower quality, then stop and return the current model; otherwise move to the neighbour with highest quality and return to 2.

Evaluation

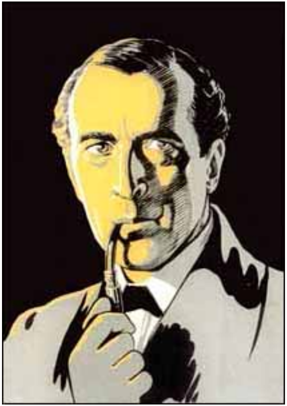
The traditional statistical approach tests one carefully crafted hypothesis. In data mining one generates and tests lots of hypotheses. This means that we have to *evaluate* our results:

- Statistical significance
- Domain significance

Classical Statistics (Theory Driven)



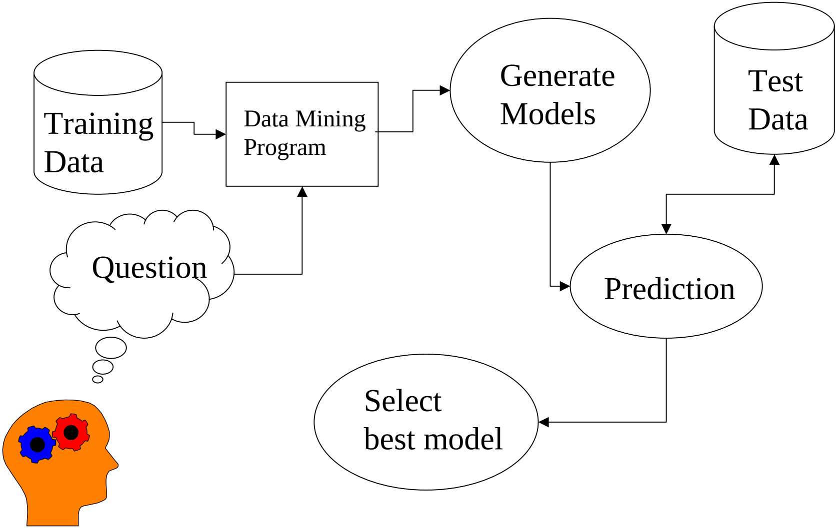
Data Mining: Sherlock Holmes Inference



"No data yet…It is a capital mistake to theorize before you have all the evidence. It biases the judgements."

A study in scarlet

Data Mining (Data Driven)



Statistical Significance

- Train and test: divide the database randomly into two parts
  - the *train set* from which we induce our model
  - the *test set* on which we can test the accuracy of our model
- k-fold cross validation

K-fold cross-validation

- Divide the database in k parts
- For each of the k parts do
  - use this part of the database as the test set for a model induced from the rest of the database.
- the average of the k accuracies is an estimate of the accuracy of the model induced from the complete database.