Data Mining 2024
Introduction

Ad Feelders

Universiteit Utrecht

The Course

- Literature: Lecture Notes, Book Chapters, Articles, Slides (the slides appear in the schedule on the course web site).
- Course Form:
  - Lectures (Tuesday, Thursday)
  - Lab session (Thursday after the lecture).
- Grading: two practical assignments (50%), a digital exam in Remindo (50%), and 4 homework exercise sets (5% bonus).
- Web Site: https://ics-websites.science.uu.nl/docs/vakken/mdm/
- MS Teams: mainly for finding team mates, and questions about practical assignments. Videos of lectures from previous years will be uploaded.

Personnel

Lecturer: Ad Feelders

Teaching Assistants:

- Panagiotis Andrikopoulos
- Ziv Hochman
- Stylianos Psara

Practical Assignments

Two practical assignments: one assignment with emphasis on programming and one with emphasis on data analysis.

1. Write your own classification tree and random forest algorithm in Python or R, and apply the algorithm to a bug prediction problem (30%).
2. Text Mining: predict whether hotel reviews are genuine or fake (20%).

Assignments should be completed by teams of 3 students.

We will not teach you how to program in Python or R. If you don't know either of these languages yet, you will have to invest some time.
Of course we will try to help you if you have questions about them.

Homework exercise sets

There are four homework exercise sets:

1. Classification trees, bagging and random forests.
2. Undirected graphical models (Markov random fields).
3. Frequent pattern mining.
4. Directed graphical models (Bayesian networks).

We will use Remindo for handing in the exercise sets.

What is Data Mining?

Selected definitions:

- (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 (!))

不是簡單的事

Valid: 模式必須是正確的，不是隨機出現的

Novel: 模式要有新意，而不是大家都知道的

別人收集好的資料

從資料中推導規則或模式

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, heterogeneous, ...) data bases.

有關的

由不同成分形成的

Models

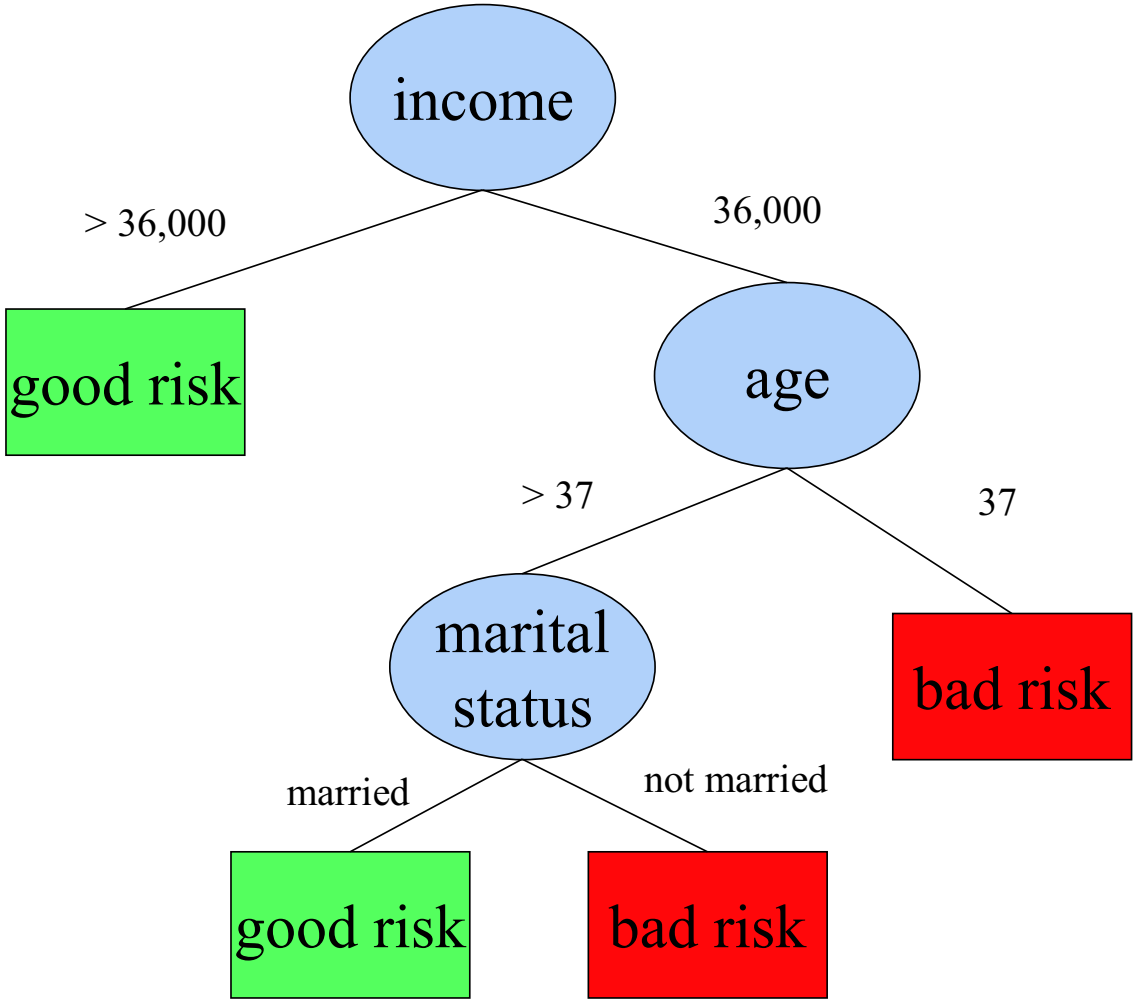
A model is an abstraction of a part of reality (the application domain).

抽象化 → 現實的一個簡化版本

In our case, models describe relationships among:

- attributes (variables, features),
  資料的欄
- tuples (records, cases),
  資料的列
- or both.

Example Model: Classification Tree



Patterns

Patterns are **local models**, that is, models that describe only part of the database.

For example, association rules:

Diapers → Beer → 如果買尿布，顧客很可能也會買啤酒
support = 20% → 這個規則在整個資料中出現的比例
confidence = 85% → 當買尿布時，有 85% 的機率也買啤酒

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

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

Diapers $\rightarrow$ Beer

Reasons to Model

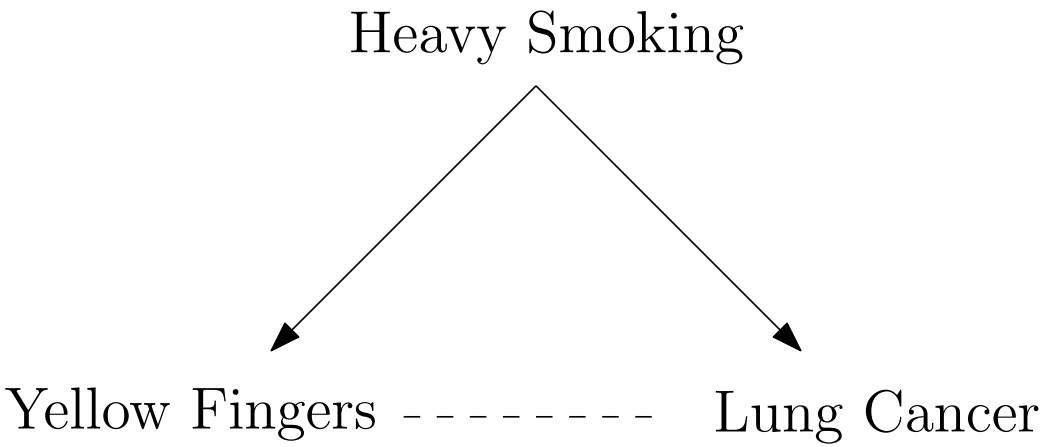
A model

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

A model that predicts well does not always provide understanding.

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

Causality and Correlation



Washing your hands doesn't help to prevent lung cancer.

Induction vs Deduction

演繹 歸納

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 有缺陷的 or otherwise) are defective

Inductive vs Deductive: Acceptance Testing Example

100,000 products; d is proportion of defective products

sample 1000; $\hat{d}$ is proportion of defective products in the sample

Suppose 10 of the sampled products turn out to be defective (1% of the sample; $\hat{d} = 0.01$)

Deductive: $d \in [0.0001, 0.9901]$ 範圍極廣，幾乎涵蓋了所有可能值 → 結論非常保守

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

95% confidence interval:

$$\hat{d} \pm se(\hat{d}) \times z_{0.975} = 0.01 \pm \underbrace{\sqrt{\frac{0.01 \times 0.99}{1000}} \times 1.96}_{\approx 0.006}$$

Experimental data

The experimental method:

- Formulate a hypothesis of interest.

  For example: "This fertilizer increases crop yield"
- Design an experiment that will yield data to test this hypothesis.

  For example: apply different levels of fertilizer to different plots of land and compare crop yield of the different plots.
- 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 yield is somewhat higher under trees where birds roost.
- Conclusion: bird droppings increase yield;
- ... or do moderate amounts of shade increase yield?

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

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

每加侖英里數（燃油效率） 氣缸數 引擎排氣量 馬力 車重

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.

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

吸入

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"
...
```

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
```

Pr(>|t|) → p 值，表示係數是否顯著。 $p < 0.05$ → 顯著

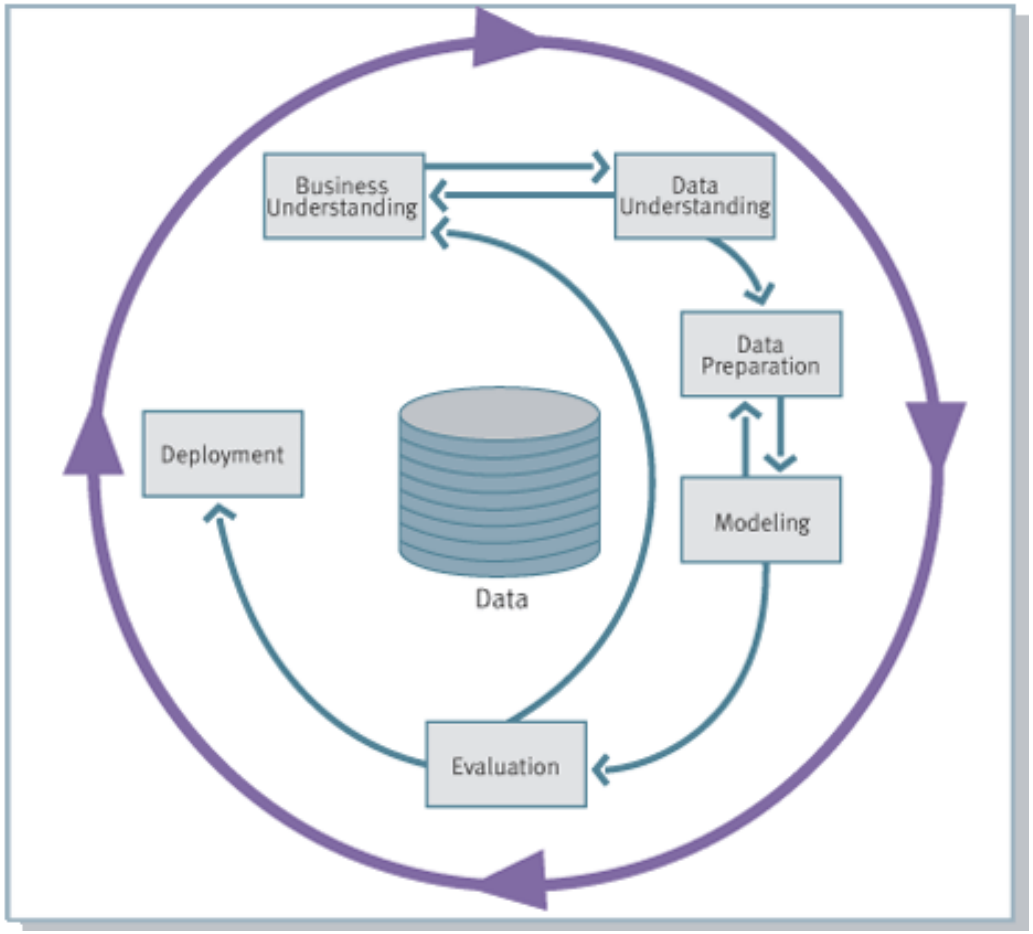
R-squared 越接近 1 → 模型對資料解釋力越高

```
> 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

Cross-Industry Standard Process for Data Mining

KDD: Knowledge Discovery in Databases



This course is mainly concerned with the modeling phase.

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 the record:

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

consistent by changing attribute values.

前後一致的

What change(s) would you make? (Wooclap)

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

What to do with missing values?

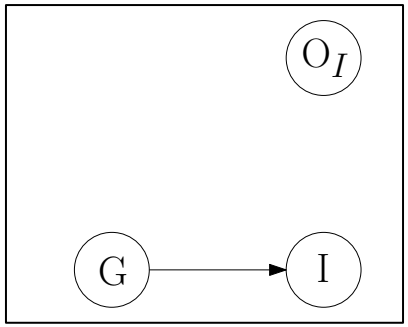
- One can remove a tuple if one or more attribute values are missing.
  Danger: how representative is the remaining sample?
  Also, you may have to ignore a large part of the data!
- One can remove attributes for which values are missing.
  Danger: this attribute may be important.
- You do *imputation*: you fill in a value.
  「填補」缺失值
  Note: but not just any value!

Missing Data Mechanisms: Not Data Dependent (NDD)

機制；成因

Suppose we have data on gender and income.
Gender (G) is fully observed, income (I) is sometimes missing.
O_I indicates whether income is observed ($O_I = 1$) or not ($O_I = 0$).

For example, missingness is determined by the roll of a die.



- There will be no bias if we remove tuples with missing income.
- If we do imputation, what values should we fill in? (Wooclap)

Missing Data Mechanisms: NDD

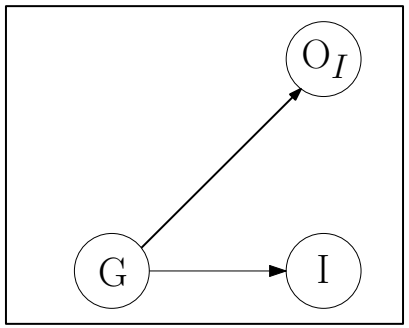
We could perform imputation as follows:

「填補」缺失値

- If person is male, pick a random male with income observed and fill in his value.
- If person is female, pick a random female with income observed and fill in her value.

Missing Data Mechanisms: Seen Data Dependent (SDD)

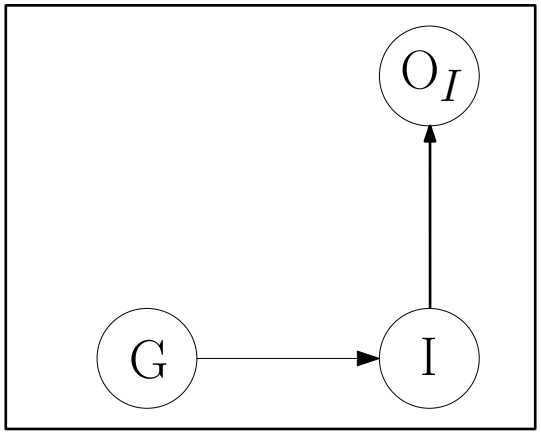
For example: men are less likely to report their income than women.



- This time there *will* be bias if we remove tuples with missing income. What bias? (Wooclap)
- Imputation: same as before, still works.

Missing Data Mechanisms: Unseen Data Dependent (UDD)

For example: people with high income are less likely to report their income.



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

Missing Data

- NDD (Not Data Dependent) is a necessary condition for the validity of complete case analysis.
- SDD (Seen Data Dependent) provides a minimal condition on which valid statistical analysis can be performed without modeling the missing data mechanism.
- Unfortunately, we cannot infer (推斷) *from the observed data alone* whether the missing data mechanism is SDD or UDD (Unseen Data Dependent).
- We might have knowledge about the nature of the missing data mechanism however ...
- Practice: if you don't know, assume SDD and hope for the best.

Construct Features

把「原始資料」轉換成「有用的變數」，方便分析或建模。

Quite often, the raw data is not in the proper format for analysis, for example:

- You have data on income and fixed expenses and you think disposable income is important.
  可自由使用的
- You have to analyze text data, for example hotel reviews. You could represent the text as a *bag-of-words*.
  每個詞計數
- Relational data bases: 1:1 relationships between tables are easy, but what to do with 1:n relationships?
  關聯式資料庫

Bag of Words

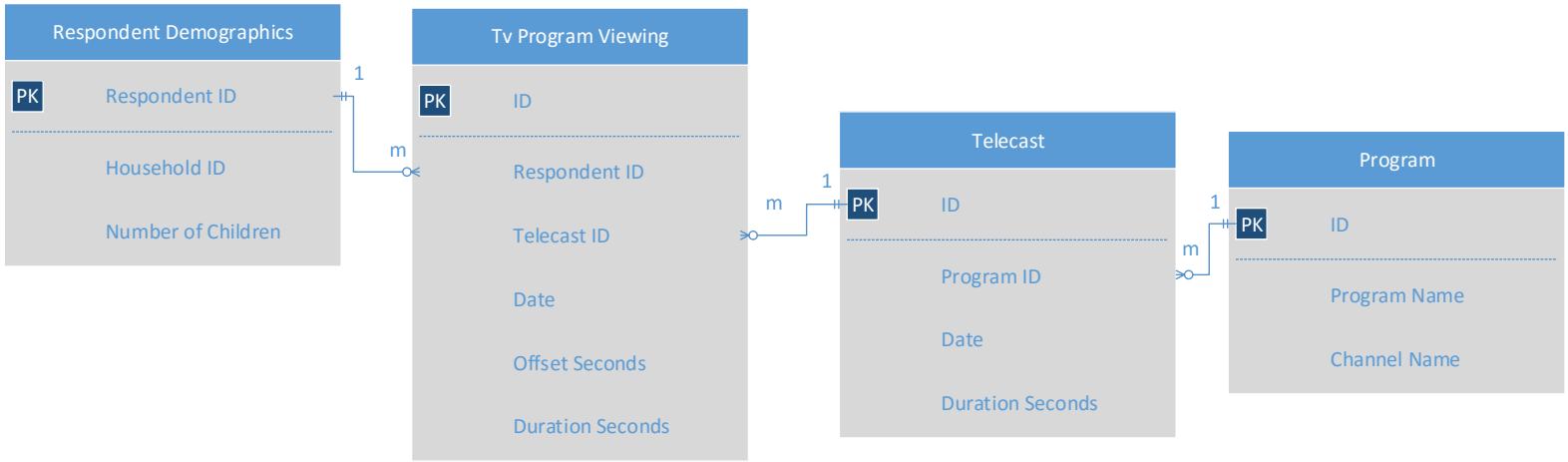
```
Doc 1: a view to a kill
Doc 2: license to kill
```

Bag of words representation:

	a	view	to	kill	license
Doc 1	2	1	1	1	0
Doc 2	0	0	1	1	1

One-to-Many Relationships: TV Viewing

Predict household composition from TV viewing behavior.



Aggregating the data to household level

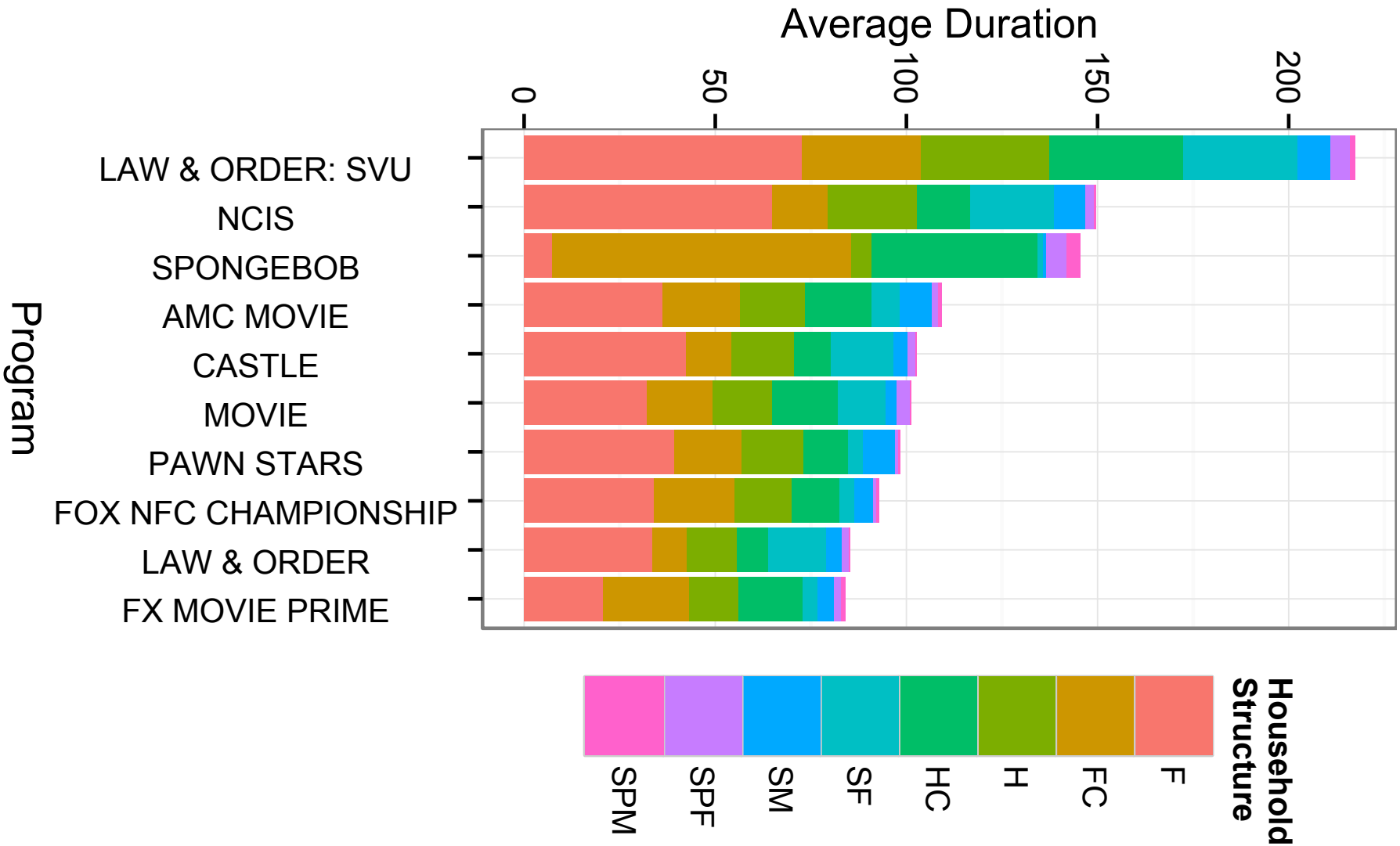
Viewing behaviour has to be aggregated, for example:

- Weekly viewing frequency of different programs.
- Weekly viewing duration of different programs.
- Weekly viewing frequency of different program *categories*.
- etc.

Potentially results in (導致) a huge number of attributes.

Example:
文字資料 (Bag-of-Words)
如果你的資料庫有 10,000 個不同單字 → 每個單字都變成一個欄位 → 就有 10,000 個屬性。

Some descriptive statistics



Modeling: Data Mining Tasks

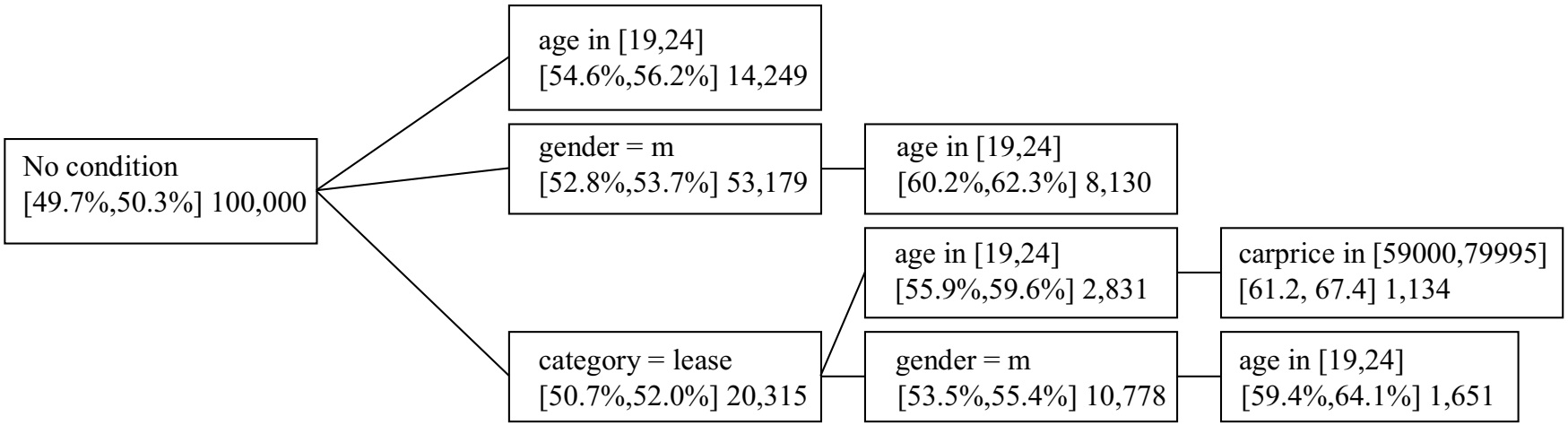
Common data mining tasks:

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

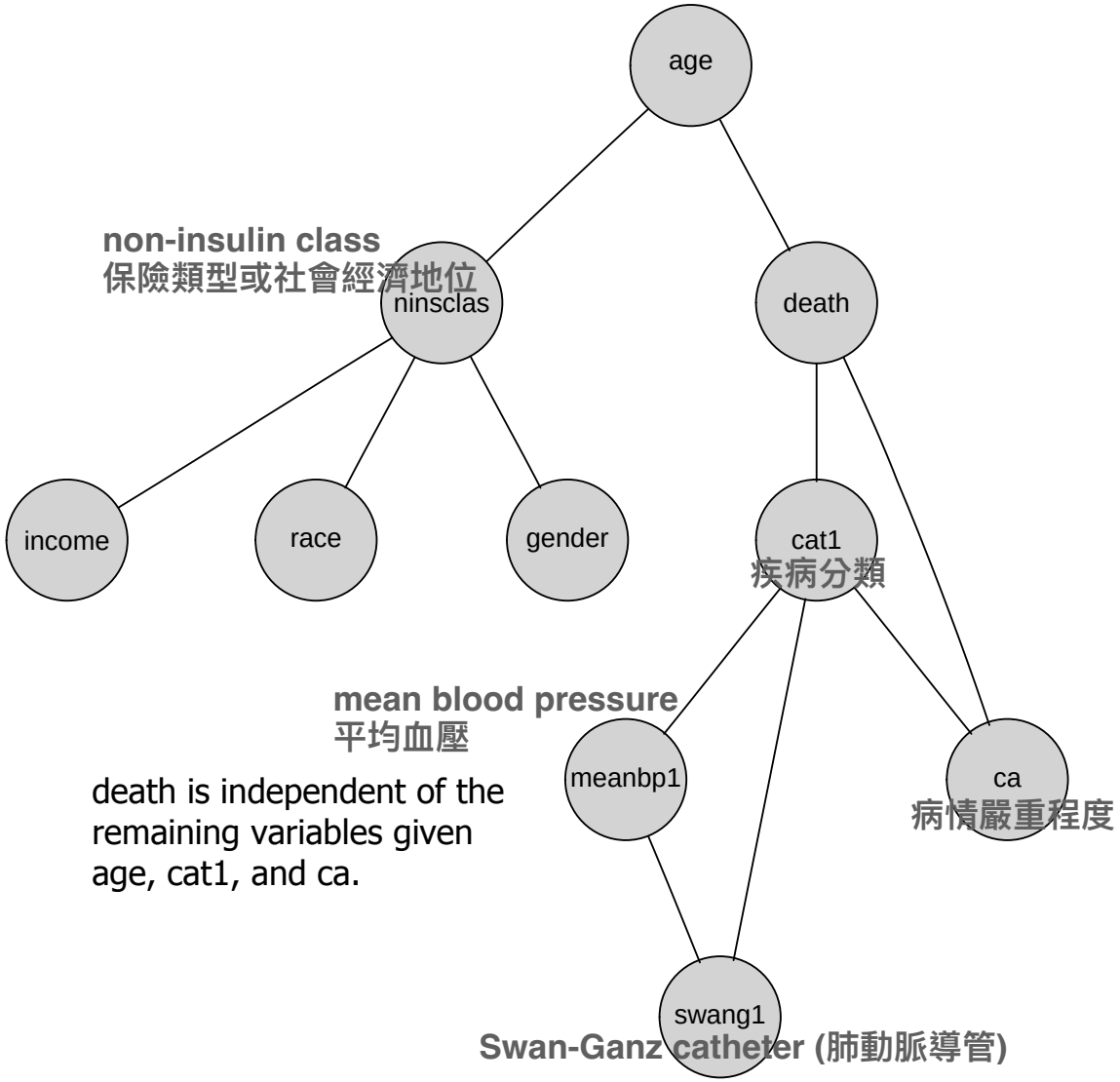
Subgroup Discovery

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

Car insurance example (target: did person claim?):



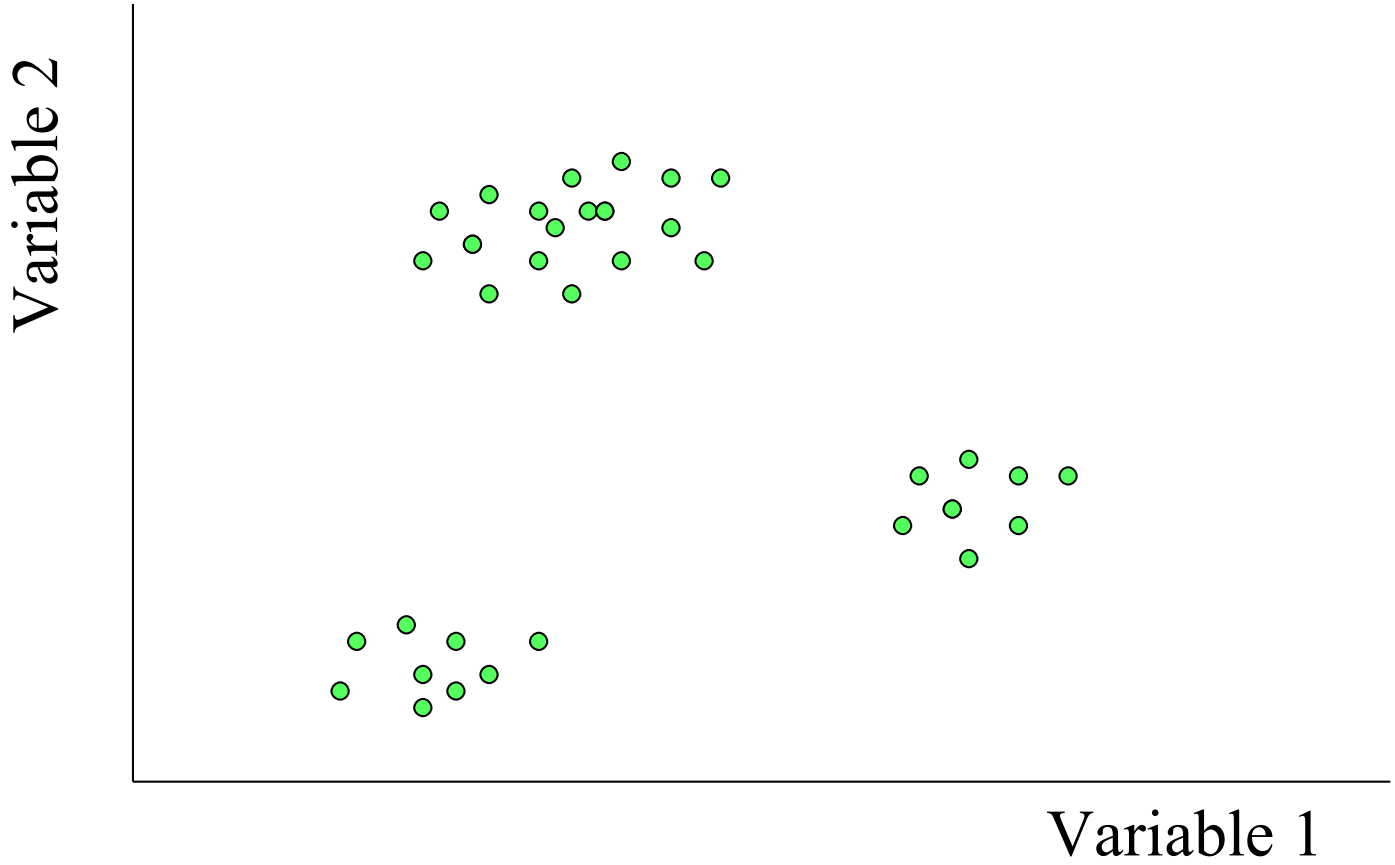
Dependency Modeling: Intensive Care Data



death is independent of the remaining variables given age, cat1, and ca.

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 (查詢).
- Rank job applicants with respect to their suitability for the job.
- Rank loan applicants with respect to default (違約) risk.
- ...

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

Components of Data Mining algorithms

Data Mining Algorithms can often 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 go about finding good models?

1. 怎麼表示模型
2. 怎麼評估模型
3. 怎麼找到好模型

Representation Languages

Representation languages define the set of all possible models, for example:

- linear models: $y = b_0 + b_1 x_1 + \cdots + b_m x_m$
- association rules: $X \rightarrow Y$
- subgroups: $X_1 \in V_1 \wedge \cdots \wedge X_m \in V_m$
- classification trees
- Bayesian networks (DAGs)

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)

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

誤差平方和 → 擬合程度

2 × 参数数量 → 複雑度懲罰

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

Overfitting on the training data

Slogan:

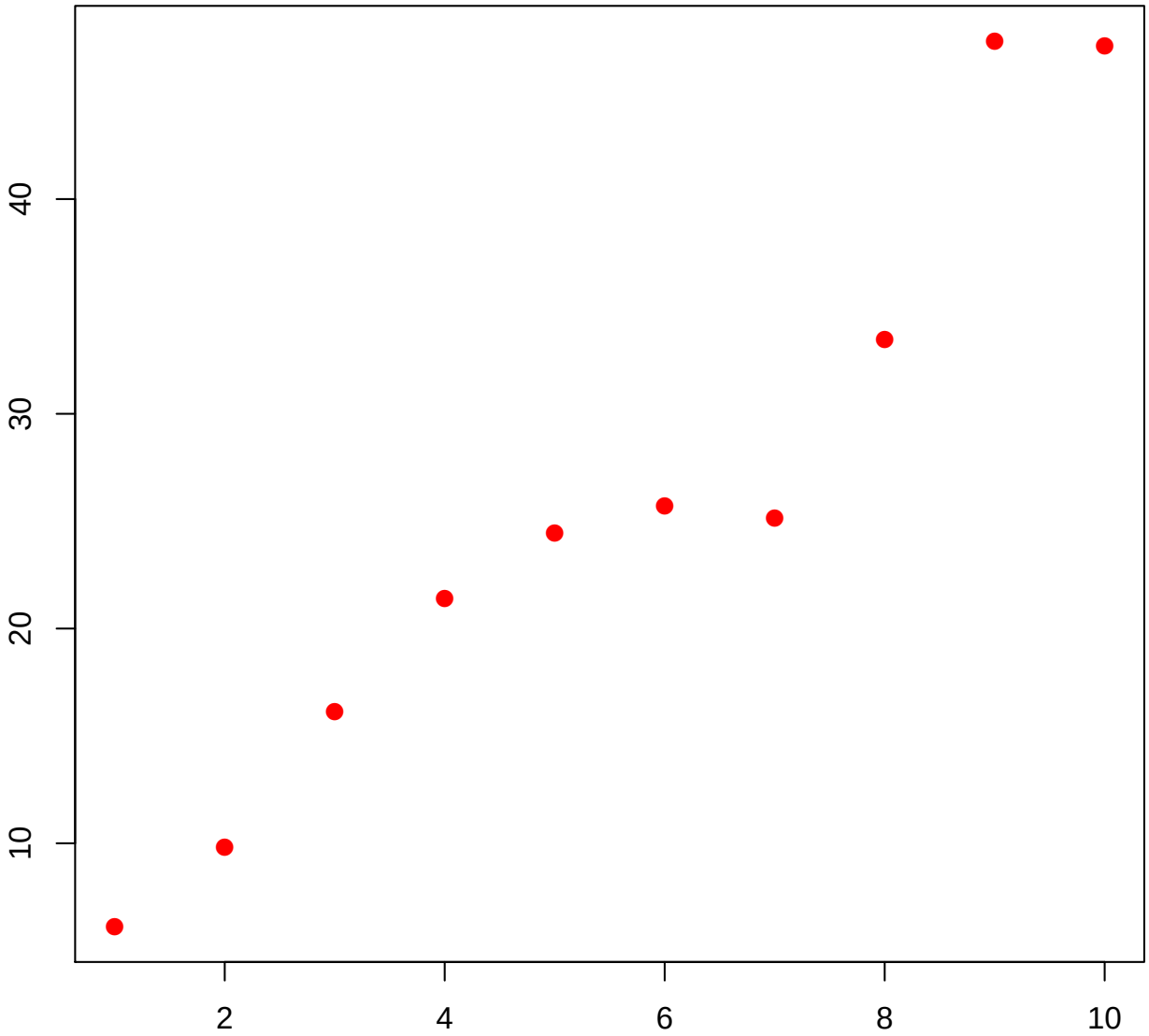
$$\text{DATA} = \text{STRUCTURE} + \text{NOISE}$$

We want to capture the structure, not the noise!

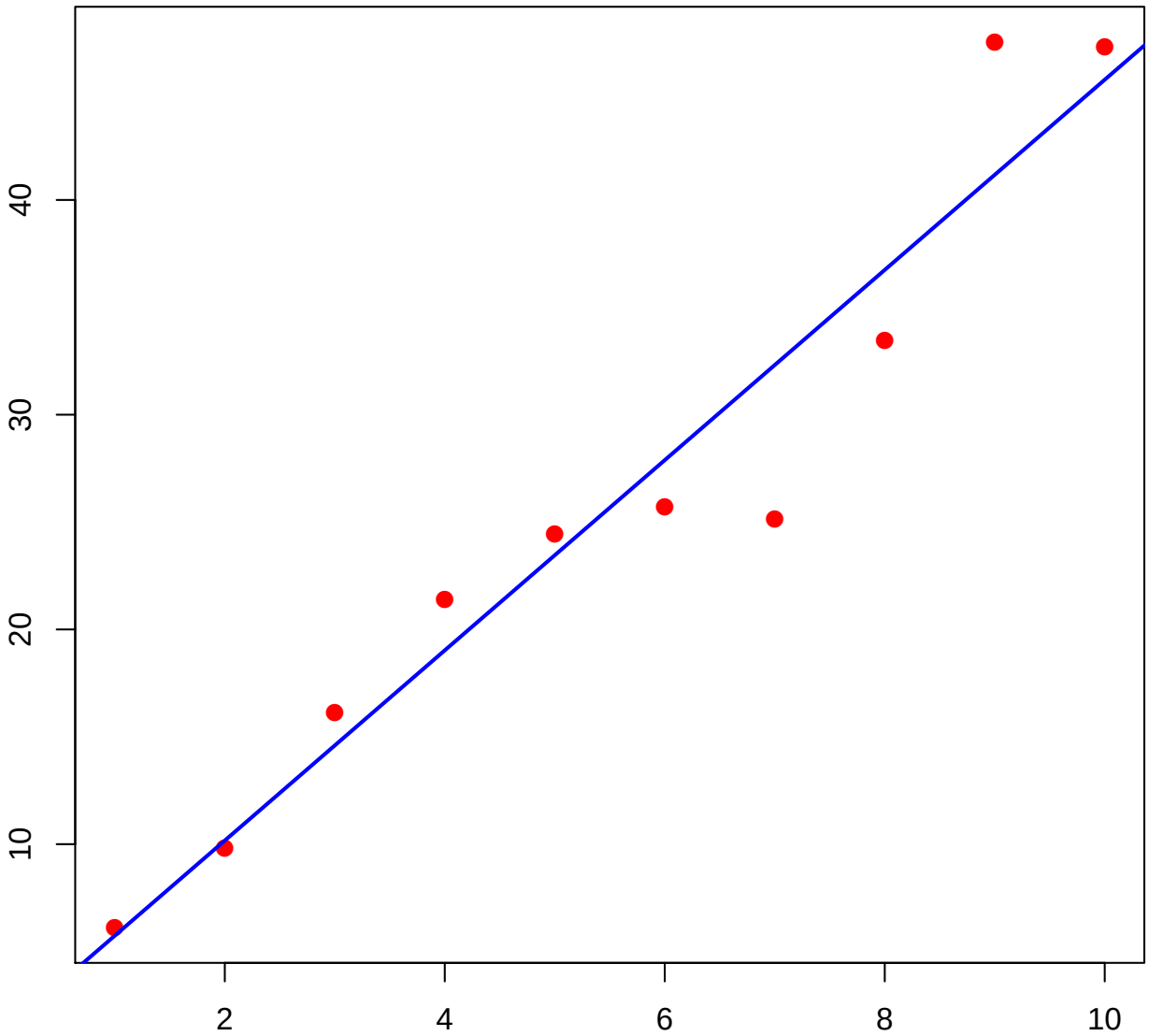
Regression example:

$$y_i = a + bx_i + \varepsilon_i \qquad\qquad i = 1, \ldots, n$$

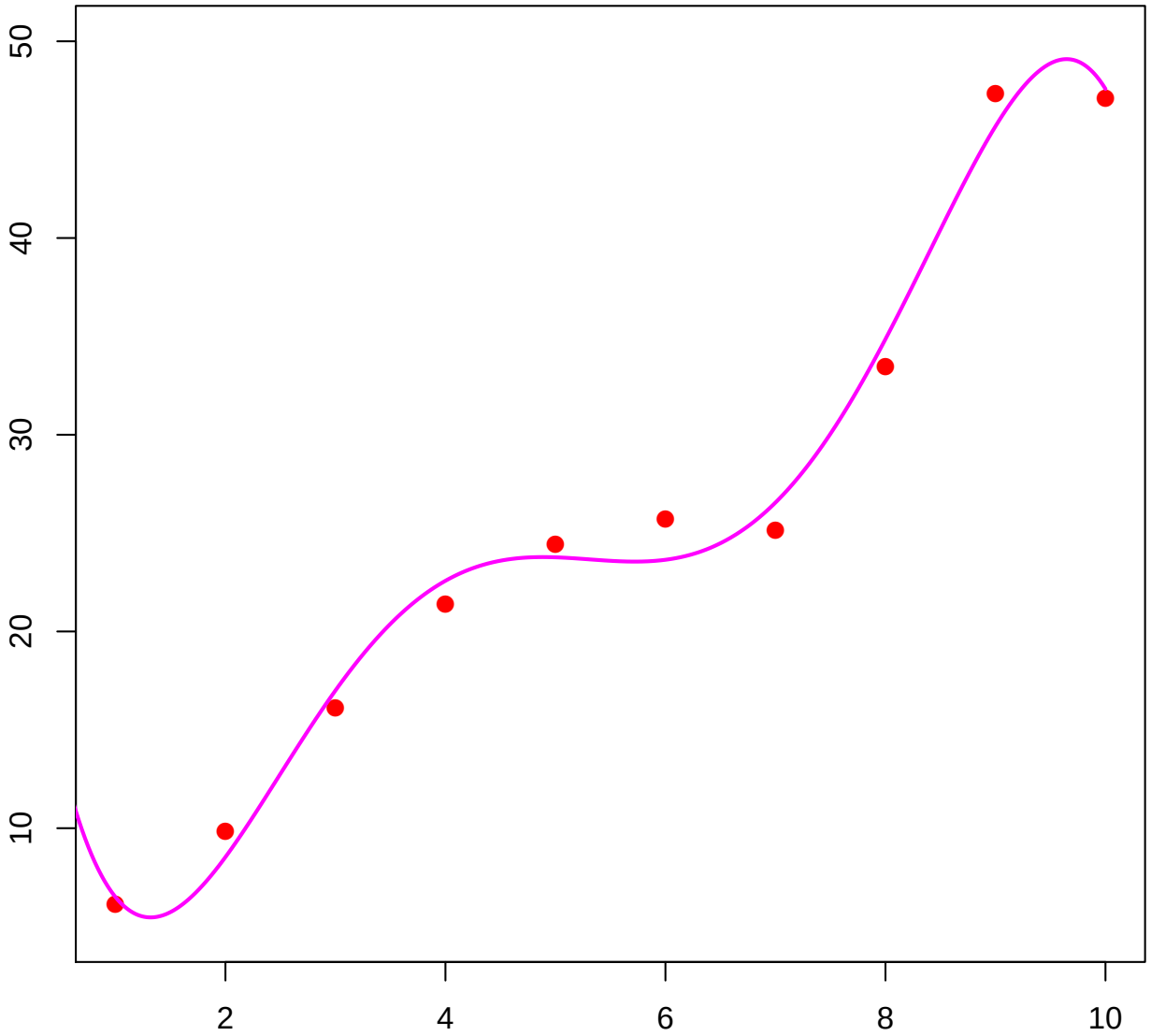
The training data: $y_i = a + bx_i + \varepsilon_i$



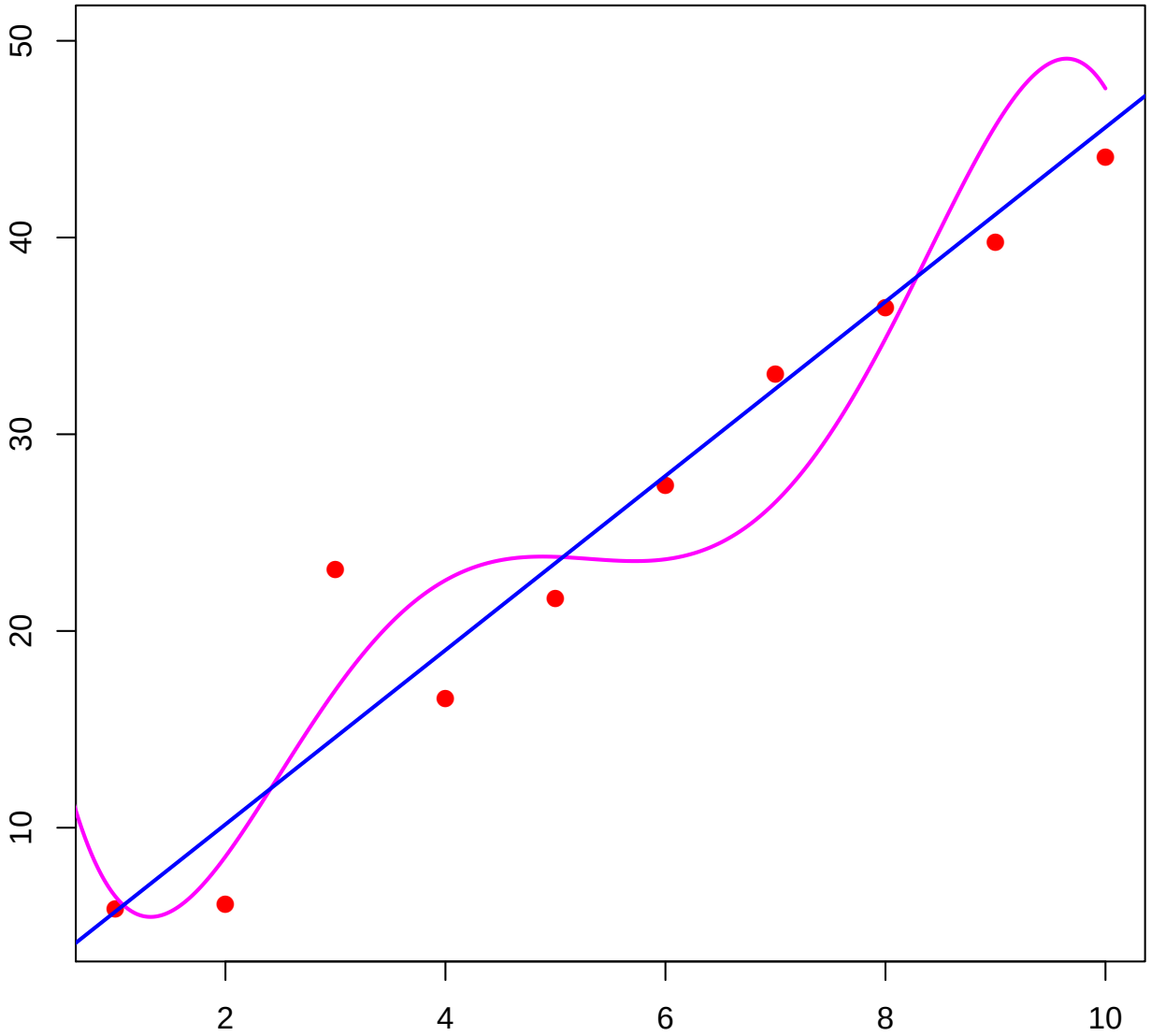
Fitting a linear model to the training data



A degree 5 polynomial fits the training data better!



Overfitting: linear model generalizes better to new data



Search for the good models

「a priori principle」是資料探勘裡的一個概念，特別常用在 frequent pattern mining（頻繁模式挖掘）。它的核心意思是：如果一個模式（itemset）是頻繁的，那它的所有子模式也一定是頻繁的；反之，如果一個模式是不頻繁的，那它的超模式一定也不可能頻繁。

補知識：
Frequent pattern mining
Hill-climber
Genetic/evolutionary algorithm

假設我們在挖購物籃資料：
如果 {牛奶, 麵包} 是頻繁的（經常一起被買），
→ 那麼 {牛奶} 和 {麵包} 一定也是頻繁的。
反過來，如果 {牛奶, 蘋果} 不頻繁，
→ 那麼 {牛奶, 蘋果, 麵包} 也不可能頻繁，
→ 可以直接 剪掉這個候選模式，不用再算它的支持度。

Exhaustive search (窮舉搜尋)

- Sometimes we can check all possible models, because there are rules with which to prune (修剪) large parts of the search space; for example, the "a priori principle" in frequent pattern mining.

Heuristic search (啟發式搜尋)

- Usually we have to employ *heuristics*
  - A general search strategy, such as a hill-climber (爬山演算法) or a genetic (evolutionary) (基因演算法 / 演化演算法) 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 (突變) operators for genetic search.

Hill-climbing：鄰居操作（neighbor operator） → 移到更好的鄰居
Genetic algorithm：交叉（crossover）、突變（mutation） → 產生新模型

Search: example

In linear regression, we want to predict a numeric variable y from a set of predictors $x_1, \ldots, x_m$. We might include any subset of predictors, so the search space contains 2^m models. E.g., if $m = 30$, 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 (逐步) search:

1. Start with some initial model, e.g. $y = a$, and compute its quality.
2. Neighbours: add or remove a predictor.
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.

Classical Text Book Approach (Theory Driven)

- Specify hypothesis (model) of interest. The model is determined up to a fixed number of unknown parameter values.
  指定
- Collect relevant data.
- Estimate the unknown parameters from the data.
- Perform test, typically whether a certain parameter is zero, using the same data!
  是否顯著

It is allowed to use the same data for fitting the model and testing the model, because we did not use the data to determine the model specification.
規格

Data Mining (Data Driven)

A simple analysis scenario could look like this:

- Formulate question of interest.
- Select potentially relevant data.
- Divide the data into a training and test set.
- Use the training set to fit (many) different models.
- Use the test set to compare how well these models generalize.
- Select the model with the best generalization performance.

In this scenario, we cannot use the training data both to fit models and to test models!

Vacancies in Education Advisory Committee

The master Education Advisory Committee (EAC) of computer science has 3 vacancies:

1. Student member Data Science (DASC)
2. Student member Computing Science (COSC)
3. Student member Game- and Media Technology (GMTE)

The EAC gives solicited and unsolicited advice about individual courses, the curriculum, Education and Exam Regulations (EER), etc.

Typically 5 meetings per year to discuss course evaluations (i.e. Caracal) after each period, and the Education and Exam Regulations (EER).

Financial compensation $\pm$ €600.

Interested? Send e-mail with motivation to `a.j.feelders@uu.nl`.