Data mining
Skillnaden mellan observationella och experimentella data

Data mining

- Metoder för att **automatiskt** upptäcka **icke-trivial användbar** information i **stora datamängder**

Data mining: (Mot-)exempel

Data mining:

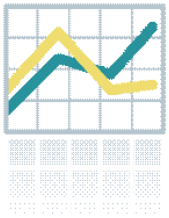
- Upptäcka att vissa namn är vanligare i vissa regioner i Sverige
- Gruppera webbsidor som returneras från sökning på ordet ”Amazon” beroende på vem som besöker dem

INTE data mining:

- Slå upp ett nummer i telefonkatalogen
- Googla på ordet ”Amazon”

Data mining: 2 användningsområden

- **Förutsäga framtiden**
  - Klassificering
  - Regression
- **Beskrivande**
  - Associationsregler
  - Klustring
  - Upptäcka konstigheter
  - Visualisering



Varför data mining?

Vetenskapligt svar:

- Enorma mängder data samlas in hela tiden (GB/h)
  - sensorer på satelliter
  - teleskop som kollar av himlen
  - simuleringsdata
  - DNA-experiment

- Traditionella statistiska metoder omöjliga
- Data mining kan hjälpa vetenskapare att
  - klassificera data
  - formulera hypoteser

Varför data mining?

Kommersiellt svar:

- Mycket data samlas om:
  - köpbeteenden
  - surfning och sökning på Internet
  - bank- och finanstransaktioner

- Datorer allt billigare och kraftfullare
- Starkt tryck att tillhandahålla **bättre**, **skräddarsydda** tjänster

Data mining?

Övning:

Tänk på **fyra saker** du har gjort idag eller igår som har **registrerats** och **resulterat i data** som kan användas för data mining.

Klassificering: Hur hittar vi skattefuskare?

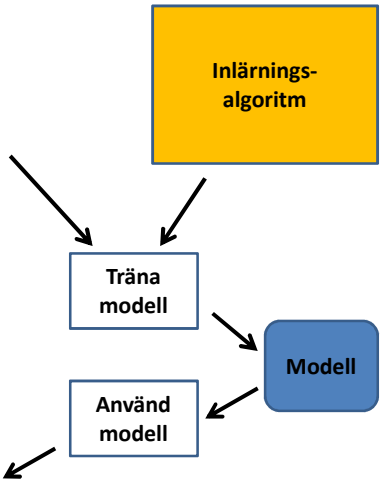
#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja

Sven: Ingen återbäring, skild, tjänar 120K?

Klassificering

#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	260K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	360K	Ja

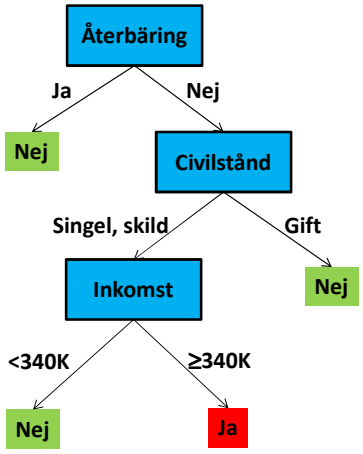
11	Nej	Gift	250K	?



Beslutsträd

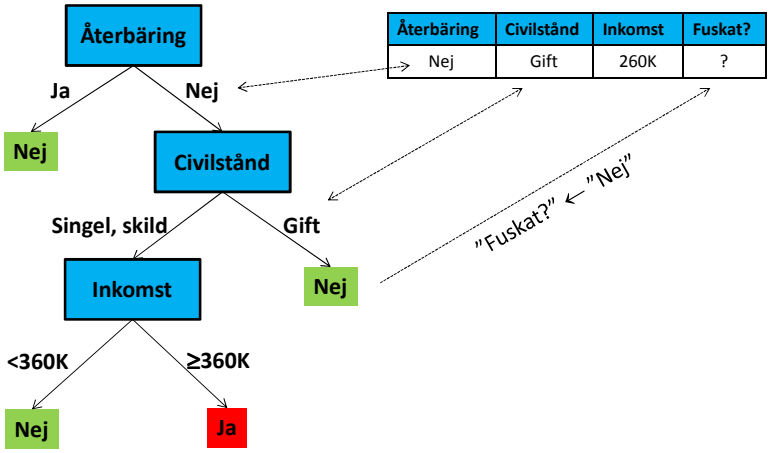
#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja

Träningsdata



Modell: beslutsträd

Använda beslutsträd för klassificering



Observationella vs experimentella data

- Data mining ger oss **observationella** data
- Observationella data kan hjälpa oss att dra slutsatser om **korrelationer** mellan variabler
- Experimentella data kan hjälpa oss dra slutsatser om **kausalitet** (orsaker → effekter)

Experiment vs. data mining

Experiment:

- Vi **vet** vad vi letar efter
  - Formulera noll-hypotes
  - Sampling
  - Förkasta eller acceptera hypotesen
- Variera de ingående variablerna och studera effekten på resultat-variabeln.

Data mining:

- Vi vet **inte** vad vi letar efter
- Data kommer från okontrollerade **observationer**

Observationella data

- Vad kan man dra för slutsatser utifrån följande observationer:
  - Obduktioner visar att avlidna som lidit av Alzheimers sjukdom har en förhöjd halt av aluminium i hjärnans celler.
  - Historiska data visar en förhöjd halt av CO_2 i atmosfären under de epoker då Jorden har en högre medeltemperatur
  - En enkät visar att överviktiga personer tenderar att föredra Coke Light framför vanlig Coke.
  - En fransk konsumentorganisation rapporterade att sannolikheten var högre att ägare till **röda bilar** inte kunde betala sina billån i tid.

Observationella data

- Vad kan man dra för slutsatser utifrån följande observationer:
  - Obduktioner visar att avlidna som lidit av Alzheimers sjukdom har en förhöjd halt av aluminium i hjärnans celler.
  - Historiska data visar en förhöjd halt av CO_2 i atmosfären under de epoker då Jorden har en högre medeltemperatur
  - En enkät visar att överviktiga personer tenderar att föredra Coke Light framför vanlig Coke.
  - En fransk konsumentorganisation rapporterade att sannolikheten var högre att ägare till **röda bilar** inte kunde betala sina billån i tid.

Observationella data

- Vad kan man dra för slutsatser utifrån följande observationer:
  - Obduktioner visar att avlidna som lidit av Alzheimers sjukdom har en förhöjd halt av aluminium i hjärnans celler.
  - Historiska data visar en förhöjd halt av CO_2 i atmosfären under de epoker då Jorden har en högre medeltemperatur
  - En enkät visar att överviktiga personer tenderar att föredra Coke Light framför vanlig Coke.
  - En fransk konsumentorganisation rapporterade att sannolikheten var högre att ägare till **röda bilar** inte kunde betala sina billån i tid.

Observationella data

- Vad kan man dra för slutsatser utifrån följande observationer:
  - Obduktioner visar att avlidna som lidit av Alzheimers sjukdom har en förhöjd halt av aluminium i hjärnans celler.
  - Historiska data visar en förhöjd halt av CO_2 i atmosfären under de epoker då Jorden har en högre medeltemperatur
  - En enkät visar att överviktiga personer tenderar att föredra Coke Light framför vanlig Coke.
  - En fransk konsumentorganisation rapporterade att sannolikheten var högre att ägare till **röda bilar** inte kunde betala sina billån i tid.

Experiment

- Kolla ifall en daglig dos C-vitamin leder till färre förkylningar.
- Hur kan vi designa ett experiment som testar detta?
- **Förslag 1:** Gör en webb-enkät
  - ”C-vitamin gör mig friskare.”
  - ”C-vitamin har ingen påverkan på min hälsa.”
- **Förslag 2:** Samla ett antal försökspersoner och låt dem ta en daglig dos C-vitamin under ett antal månader. Utvärdera sedan om personerna haft färre förkylningsdagar än motsvarande period förra året.

Experiment

- **Förslag 3:** Samla in ett antal försökspersoner. Låt varje person avgöra om de vill ta en daglig dos C-vitamin (grupp A) eller inte (grupp B). Efter testperiodens slut mäter vi om grupp A har haft färre förkylningsdagar än B.

Experiment

- **Förslag 4:** Samla in ett antal försökspersoner. Låt **försöksledaren** avgöra om de kommer att ta en daglig dos C-vitamin (grupp A) eller inte (grupp B). Efter testperiodens slut mäter vi om grupp A har haft färre förkylningsdagar än B.

Experiment

- **Förslag 5:** Samla in ett antal försökspersoner. Låt **slumpen** avgöra om de kommer att ta en daglig dos C-vitamin (grupp A) eller inte (grupp B). Efter testperiodens slut mäter vi om grupp A har haft färre förkylningsdagar än B.

Experiment

- **Förslag 6:** Samla in ett antal försökspersoner. Låt slumpen avgöra om de kommer att ta en daglig dos C-vitamin (grupp A) eller **en tablett som inte innehåller någon verksam ingrediens** (grupp B). Försökspersonerna **vet inte** om de tillhör grupp A eller grupp B. Efter testperiodens slut mäter vi om grupp A har haft färre förkylningsdagar än B.

Experiment

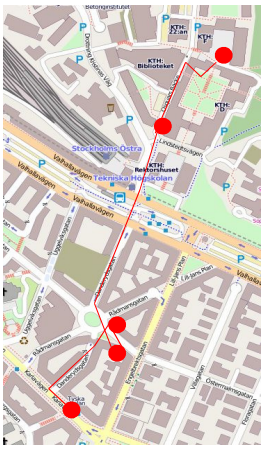
- **Förslag 7:** Samla in ett antal försökspersoner. Låt slumpen avgöra om de kommer att ta en daglig dos C-vitamin (grupp A) eller en tablett som inte innehåller någon verksam ingrediens (grupp B). Varken försökspersonerna **vet inte** om de tillhör grupp A eller grupp B, **och det vet inte heller försöksledaren**. Efter testperiodens slut mäter vi om grupp A har haft färre förkylningsdagar än B.

Några designprinciper för experiment

- Användande av kontrollgrupp
- Randomisering - Slumpen avgör vem som tillhör försöksgruppen och vem som tillhör kontrollgruppen
- Placebo
- Dubbelblinda test

Exempel - Navigeringssystemet

- Vi ville designa ett system som guidar en fotgängare F från A till B
- F har en mobiltelefon med GPS
- Det är känt från tidigare studier att människor föredrar vägbeskrivningar med **landmärken**
- ”Gå mot restaurangen på hörnet” snarare än ”Gå höger” eller ”Gå norrut” eller ”Ta Sveavägen”.



Exempel - Navigeringssystemet

- Vilket landmärke ska vi basera instruktionen på?

Exempel - Navigeringssystemet

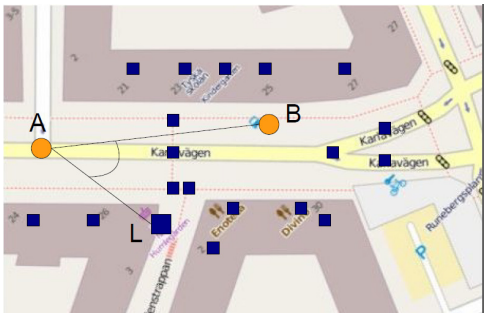
- Vi gjorde följande datainsamling:
- Ett antal personer gick en planerad rutt och beskrev hur de gick (beskrivningen spelades in).
- Vi registrerade vilka landmärken de använde (och vilka de inte använde) i sina beskrivningar.
- Varje landmärke kunde beskrivas med ett antal särdrag (storlek, typ, avstånd, etc.)
- Utifrån denna information kunde vi bygga en matematisk modell som förutsåg vilka landmärken personen skulle använda i nya situationer.

Exempel - Navigeringssystemet

learning instance:

(169822,6.0,4.0,9,1,1,0,0,0,0,0,0,0)
(180196907,0,6.0,109,1,1,0,0,0,0,0,0,0)
(713456474,6.0,0,0,1,0,0,0,0,0,0,0,1)
(928531207,5.0,5.0,40,1,0,0,0,1,0,0,0,0)
(1340895384,6.0,5.0,22,1,0,0,1,0,0,0,0,0)
(1340899520,6.0,4.0,15,1,0,0,0,1,0,0,0,0)
(1340903227,6.0,4.0,13,1,0,0,0,1,0,0,0,0)
(1525463050,6.0,5.0,27,1,0,0,1,0,0,0,0,0)
(1525463077,6.0,5.0,40,0,0,0,0,0,1,0,0,0)
(1755176084,5.0,5.0,25,0,1,0,0,0,0,0,0,0)
(1755176087,5.0,5.0,9,1,1,0,0,0,0,0,0,0)
(1755176089,5.0,5.0,6,0,1,0,0,0,0,0,0,0)
(1755176091,3.0,6.0,82,1,1,0,0,0,0,0,0,0)
(1756229411,5.0,6.0,63,0,0,0,0,0,1,0,0,0)
(1768982670,7.0,5.0,13,1,0,0,0,0,0,0,0,1)
....

L = **928531207**



Exempel - Navigeringssystemet

- Var denna datainsamling ett exempel på data mining eller ett kontrollerat experiment? Motivera!