Jan Meskens / Onderzoek

Wat is **"Predictive Analytics"**?

Historische
en/of huidige
data

Voorspellingen
over de toekomst

Toepassing: **fraudebestrijding**

Opsporen fraude met aanrijdingsformulieren [SAS]

Toepassing: **marketing**

Gepersonaliseerde reclamefolder

[Colruyt]

Toepassing: **eHealth – intensieve zorgen**

Nierfalen binnen de 3 dagen?

Kans op ontsteking?

Hoe lang op de afdeling blijven?

UZ Gasthuisberg [Ramon et al.]

Toepassing: **landbouw**

Slacht de dieren die volgend jaar **het minste melk** zullen geven

[Witten et al.]

In deze **infosessie…**

Deel 1	Wat is **predictive analytics**? Courante **predictieve modellen** **Toepassingen** van deze modellen
Deel 2	Predictive analytics **tools** Predictive analytics **cases** Wat hebben we **geleerd**?

In deze **infosessie…**

Deel 1	Wat is **predictive analytics**? Courante **predictieve modellen** **Toepassingen** van deze modellen
Deel 2	Predictive analytics **tools** Predictive analytics **cases** Wat hebben we **geleerd**?

Data warehouse
Predictive analytics
Database
Data mining
Descriptive analytics
Machine learning
Statistics
AI

"He said his first buzz word today!"

Data

Data Mining

Descriptive analytics

Predictive analytics

Statistiek, Database & DW, Machine learning & AI

Data

Data Mining

Descriptive analytics

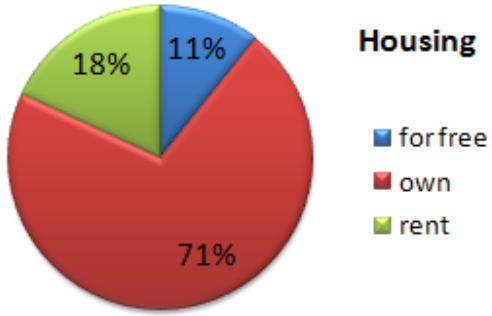
Predictive analytics

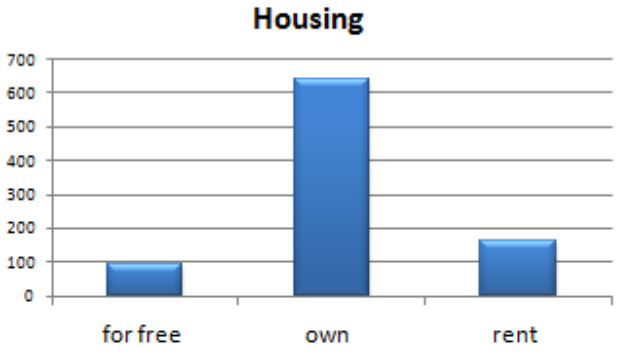
Statistiek, Database & DW, Machine learning & AI

Descriptive analytics: verklaar het **verleden**

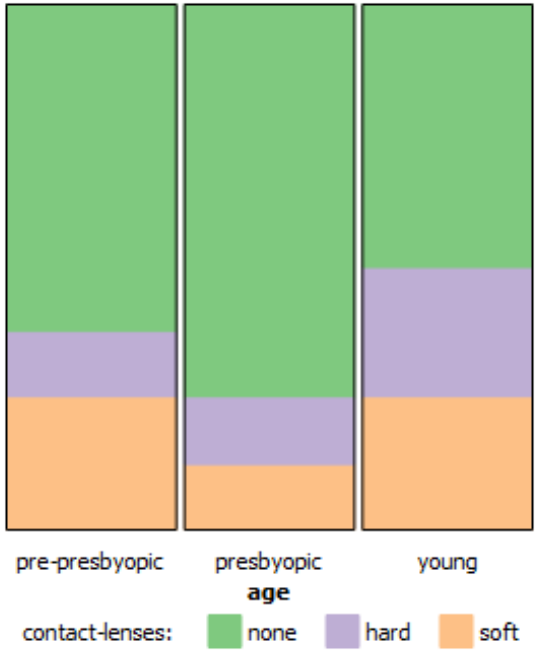
Frequency Table

Housing	Count	Count%
for free	96	10.67%
own	641	71.22%
rent	163	18.11%



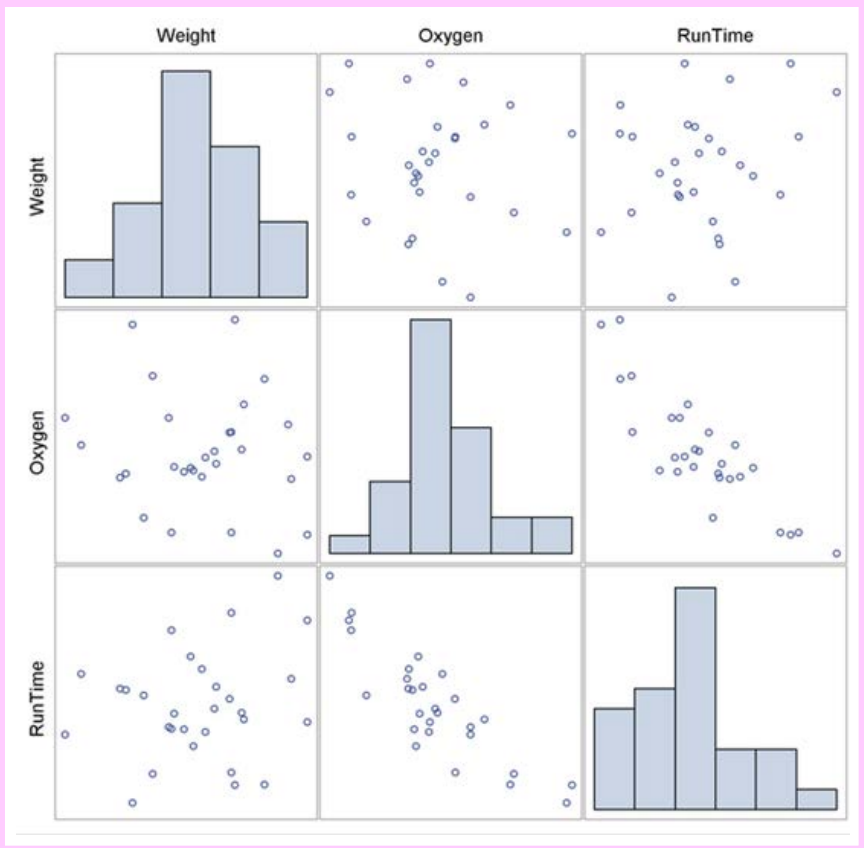


Taart diagram & histogram

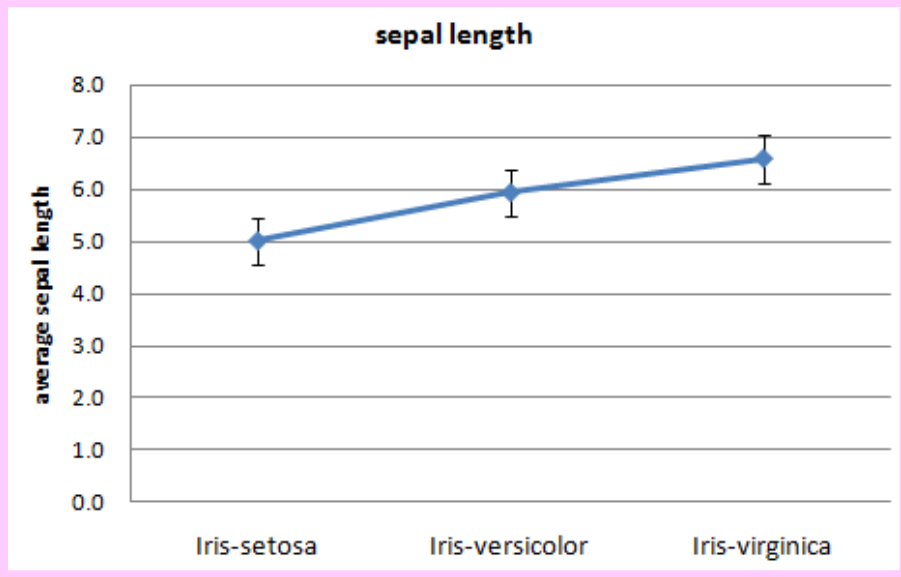


Kolom diagram

Descriptive analytics: verklaar het **verleden**



Scatter plot matrix



Line chart

Gemiddelde, ANOVA,
t-test, standaard afwijking, ...

Beschrijvende statistiek

Data

Data Mining

Descriptive analytics

Predictive analytics

Statistiek, Database & DW, Machine learning & AI

Predictive analytics: voorspel de **toekomst**

Typische dataset:

X_1	X_2	X_3	X_4	X_5	Y
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

Onafhankelijke variabele(n)

(**Predictor** variabelen(n))

Afhankelijke variabele(n)

(**Target** variabele(n))

Predictive analytics: voorspel de **toekomst**

Fraude met aanrijdingsformulieren:

Datum	Schade	Plaats	...	Fraude?
...	...	...	...	YES
...	...	...	...	YES
...	...	...	...	NO

Onafhankelijke variabele(n)

(**Predictor** variabelen(n))

Afhankelijke variabele(n)

(**Target** variabele(n))

Predictive analytics: voorspel de **toekomst**

X_1	X_2	X_3	X_4	X_5
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

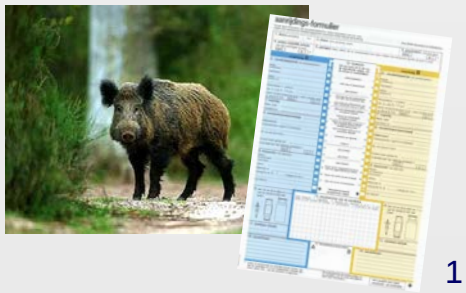
Predictief Model

f(x1,…,x5) = y

Y
…
…
…

Bv. model bij fraude met aanrijdingsformulieren:

f(datum,schade,plaats,…) = YES/NO



Predictive analytics vs. expert knowlegde

Predictief model

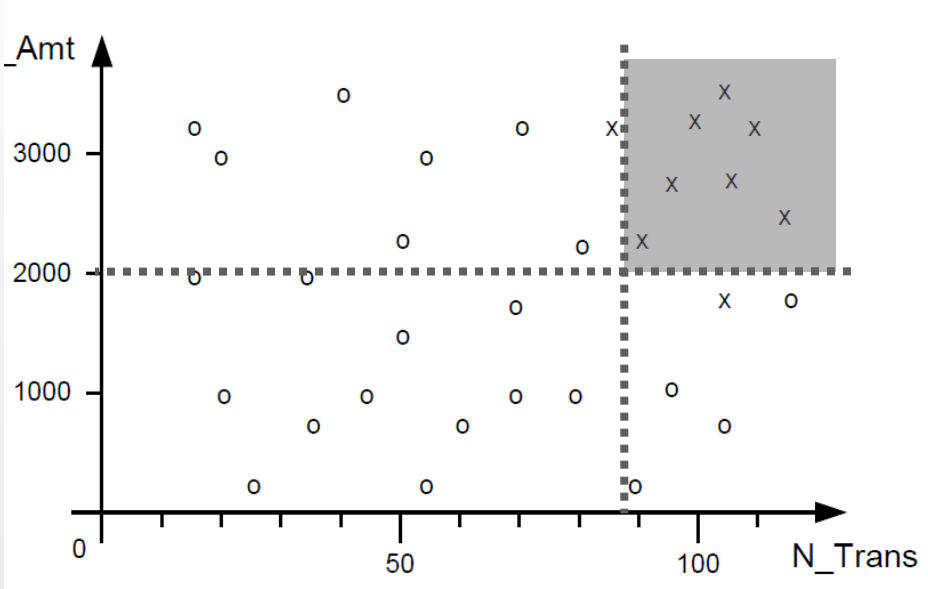
Model wordt uit de data gehaald

Business rules

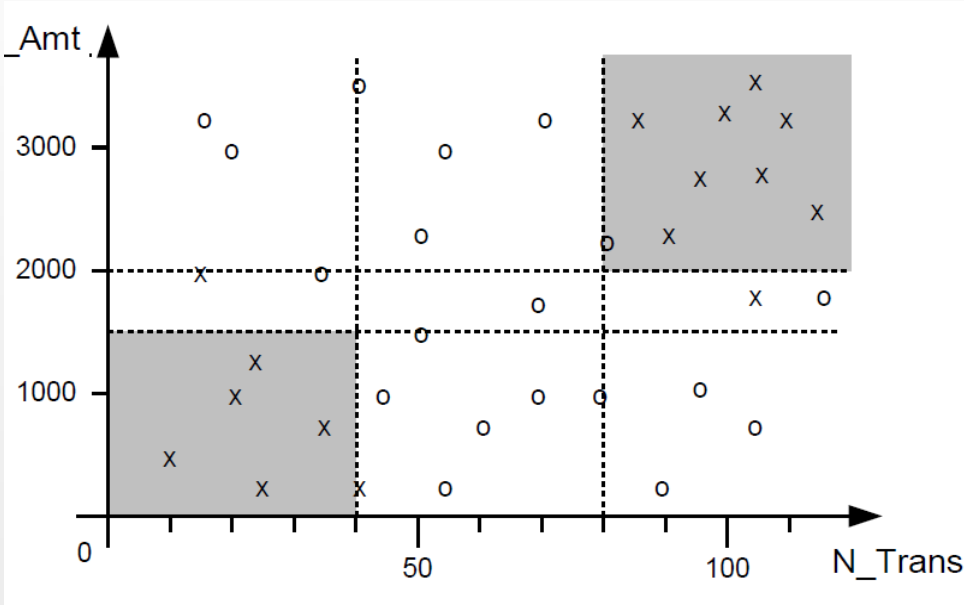
Model wordt door expert gemaakt

Expert driven business rules

Eenvoudige regel

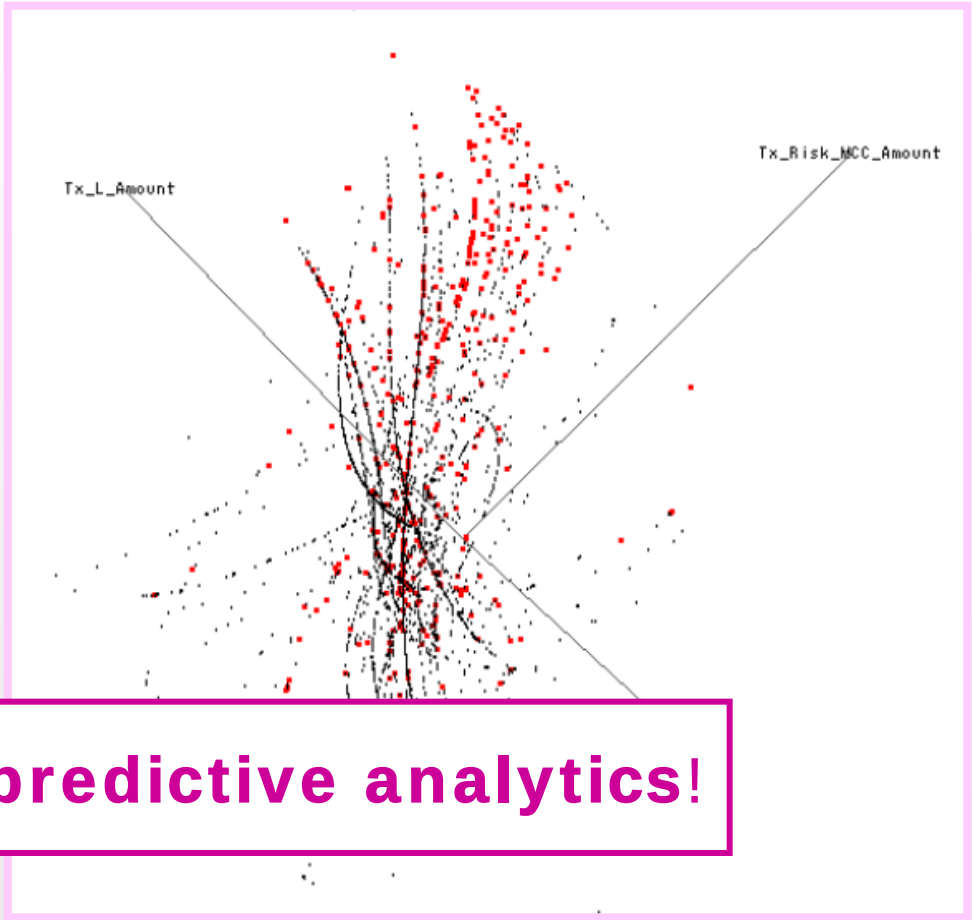


Complexere regel



Expert driven rules opstellen vaak **onmogelijk**

- Data zeer complex
- Meer dan twee dimensies
- Moeilijk om patronen manueel te definieren



Wordt opgelost met **predictive analytics**!

Predictive analytics **project aanpak**

4. Feedback

3. Interpretatie en evaluatie

2. Predictive modeling

1. Voorbereiden data

Data

D1	D2	D3	D4	D5

Dataset

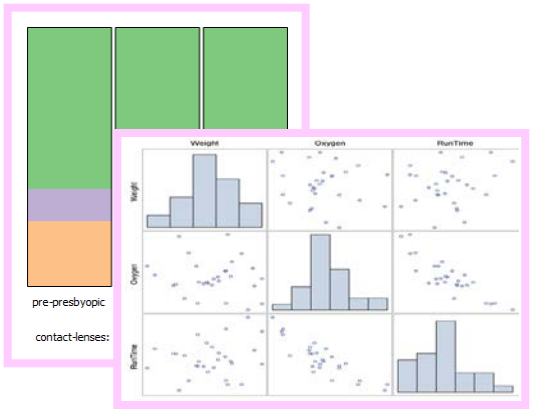
Predictief model

Kennis over de toekomst

Data **voorbereiding**

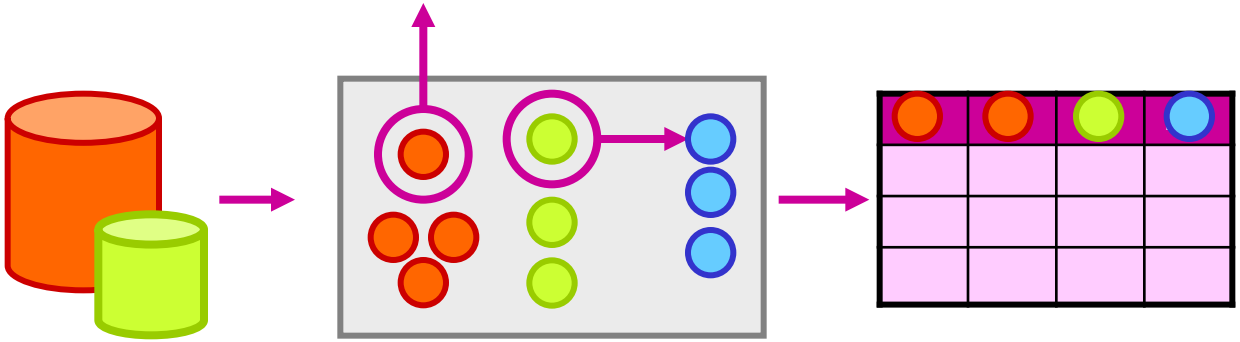
Data

Analyseren



Welke variabelen?
Verdeling data?
Betekenis data?
Data kwaliteit?

Extraheren en omvormen



Data

Dataset

Predictive analytics **project aanpak**

4. Feedback

3. Interpretatie en evaluatie

2. Predictive modeling

1. Voorbereiden data

Data

D1	D2	D3	D4	D5

Dataset

Predictief model

Kennis over de toekomst

Predictive model **opbouwen (=training)**

Input **Output**

X_1	X_2	X_3	X_4	X_5		Y
A	0,1	3	Yes	200	→	2
B	0,4	5	Yes	150	→	4

Trainingsdataset → **Trainingsalgoritme** vh predictief model → **Predictief model**

f(input) = output
Accuraatheid %

Accuraatheid van het predictief model

Confusion matrix:

	Predicted TRUE	Predicted FALSE
Actually TRUE	% True Positive (TP)	% False Negative (FN)
Actually FALSE	% False Positive (FP)	% True Negative (TN)

Accuraatheid = TP + TN

Interpretatie confusion matrix is **belangrijk!**

	Predicted CANCER	Predicted NO CANCER
Actually CANCER	% True Positive (TP)	% False Negative (FN)
Actually NO CANCER	% False Positive (FP)	% True Negative (TN)

Moet zo **laag** mogelijk zijn!

Is een predictief model **generaliseerbaar**?

Hoe gedraagt het model zich bij

data ≠ trainingsdata

≠ model accuraatheid

- Gaat enkel over **trainingsdata**
- **Te optimistisch** over performantie van het model

Naïve oplossing: trainingsdata opsplitsen

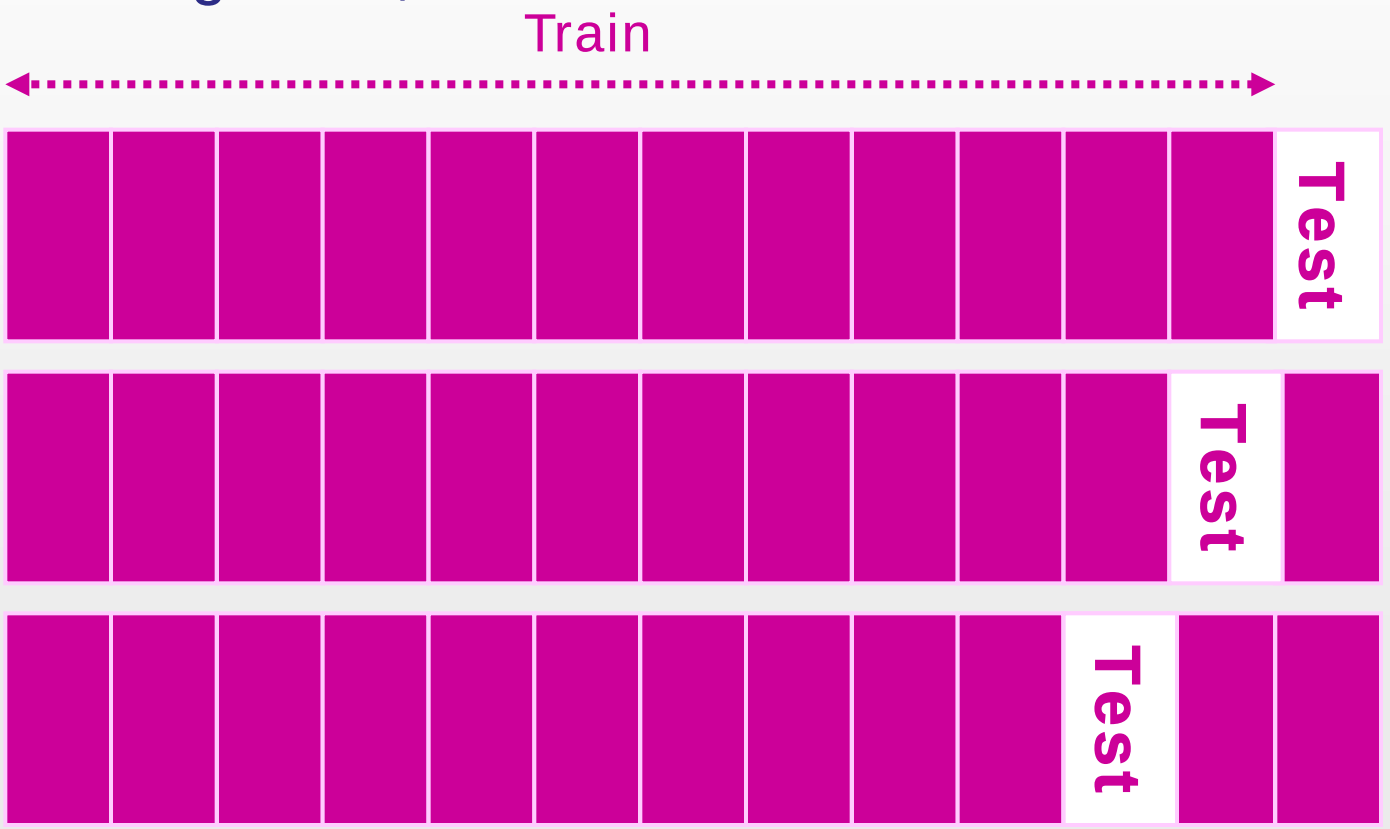
Data

Train: model leren	**Test:** generaliseerbaarheid berekenen

Probleem:
50% training, 50% test reduceert training set enorm

Oplossing: **"cross validation"**

- Data opslitsen in k delen (folds)
- k-1 trainingssets, 1 testset



Predictive analytics **project aanpak**

4. Feedback

3. Interpretatie en evaluatie

2. Predictive modeling

1. Voorbereiden data

Data

D1	D2	D3	D4	D5

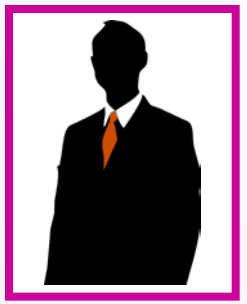
Dataset

Predictief model

Kennis over de toekomst

Interpretatie en evaluatie model output

X_1	X_2	X_3	X_4	X_5	Y
A	0,1	2	Yes	200	?
B	0,7	6	Yes	150	?
A	0,2	8	No	300	?



Interpretatie

Nieuwe data

Predictief model

Waarde van variabele **Y**

Predictive analytics **project aanpak**

4. Feedback

3. Interpretatie en evaluatie

2. Predictive modeling

1. Voorbereiden data

Data

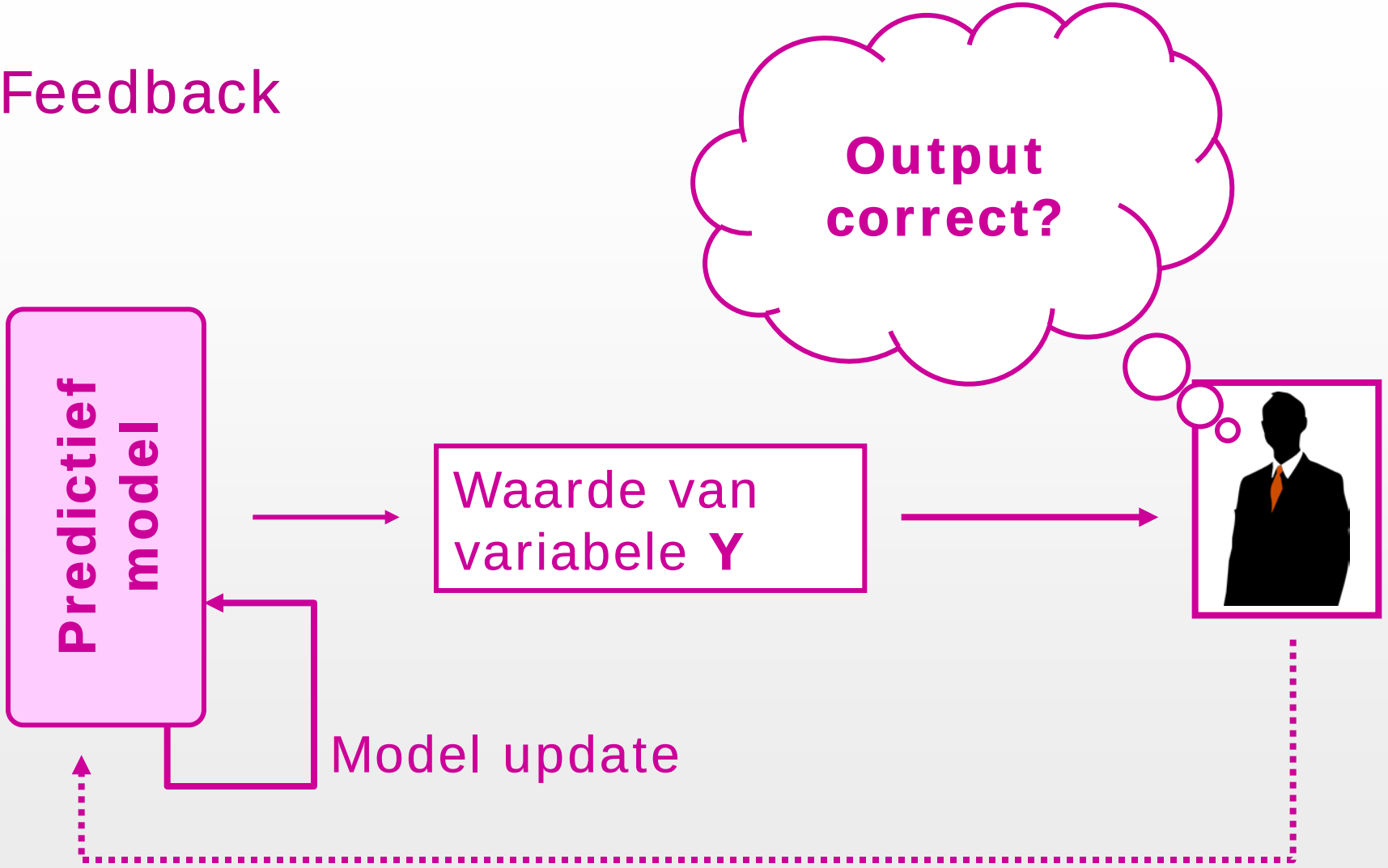
D1	D2	D3	D4	D5

Dataset

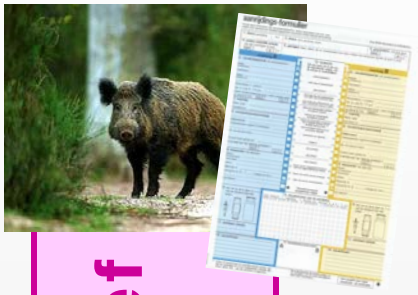
Predictief model

Kennis over de toekomst

Feedback


Output correct?
Predictief model
Waarde van variabele **Y**
Model update
Correct: JA / NEE

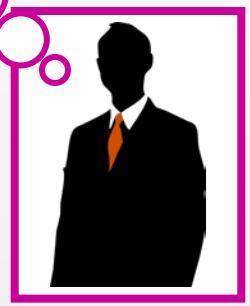
Feedback



Predictief model

Fraude

Fraude onderzoek



Model update

Fraude: Ja / Nee

Predictive analytics **project aanpak**

4. Feedback

3. Interpretatie en evaluatie

2. Predictive modeling

1. Voorbereiden data

Data

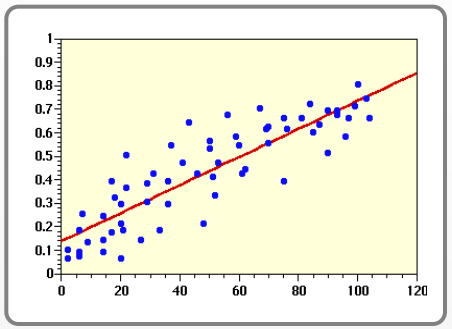
D1	D2	D3	D4	D5

Dataset

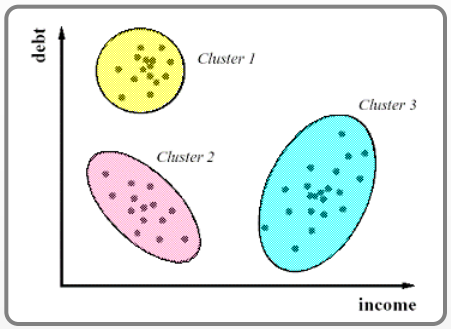
Predictief model

Kennis over de toekomst

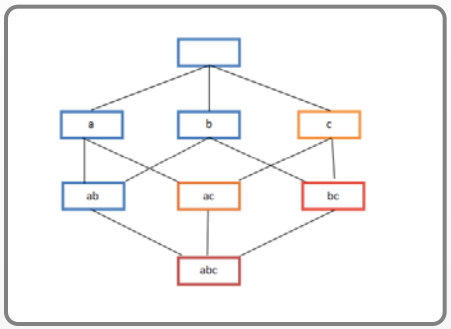
5 predictieve modellen



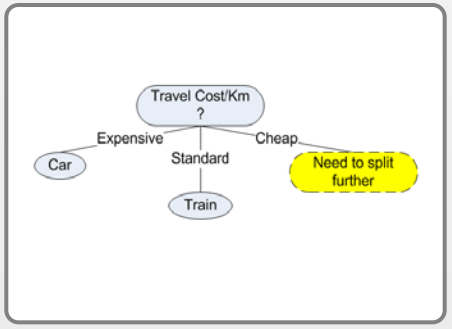
Statistics



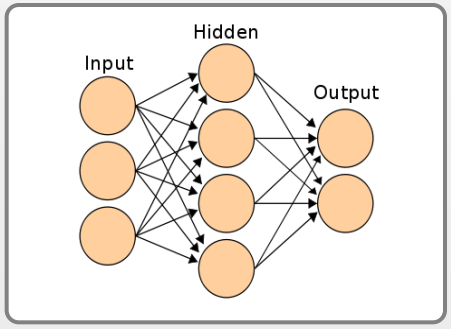
Clustering



Association rules

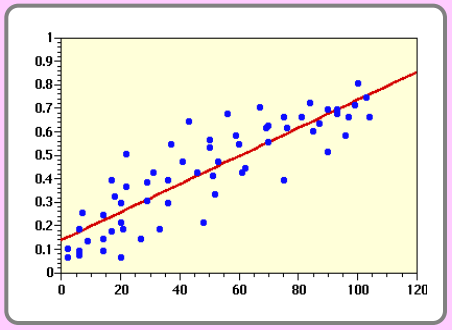


Decision trees

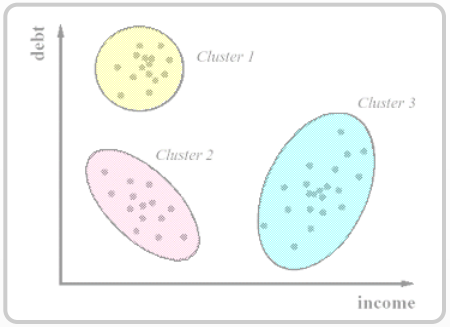


Neural networks

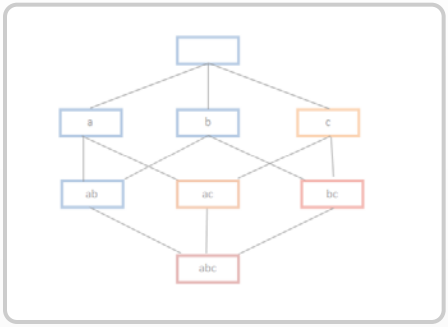
5 predictieve modellen



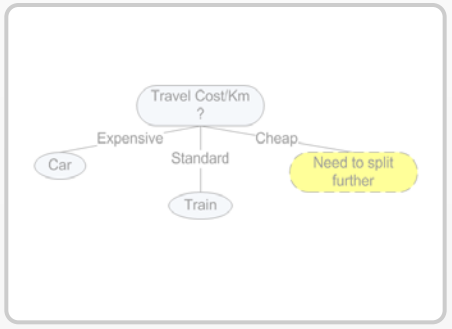
Statistics



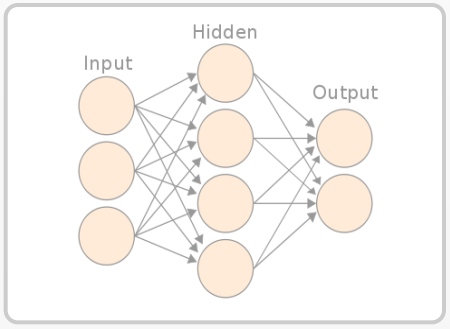
Clustering



Association rules



Decision trees



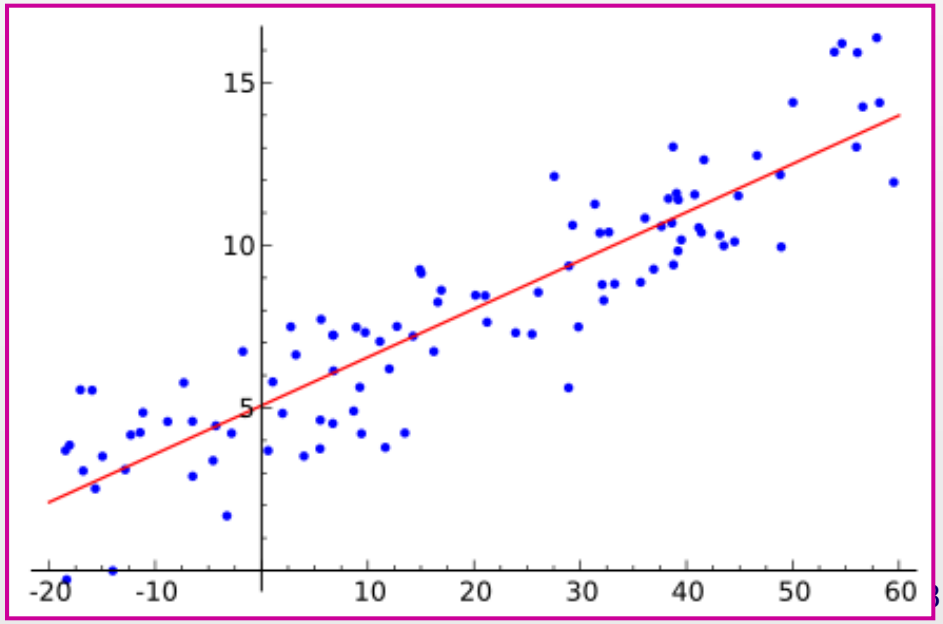
Neural networks

Statistische modellen

Geen beschrijvende statistiek zoals gemiddelde, standaard afwijking, histogrammen, ...

Wel **voorspellende statistiek:**

Regressie analyse



Regressie analyse

Simpele lineare regressie:

$$y_i = a + b \cdot x_i + r$$

Multiple lineaire regressie:

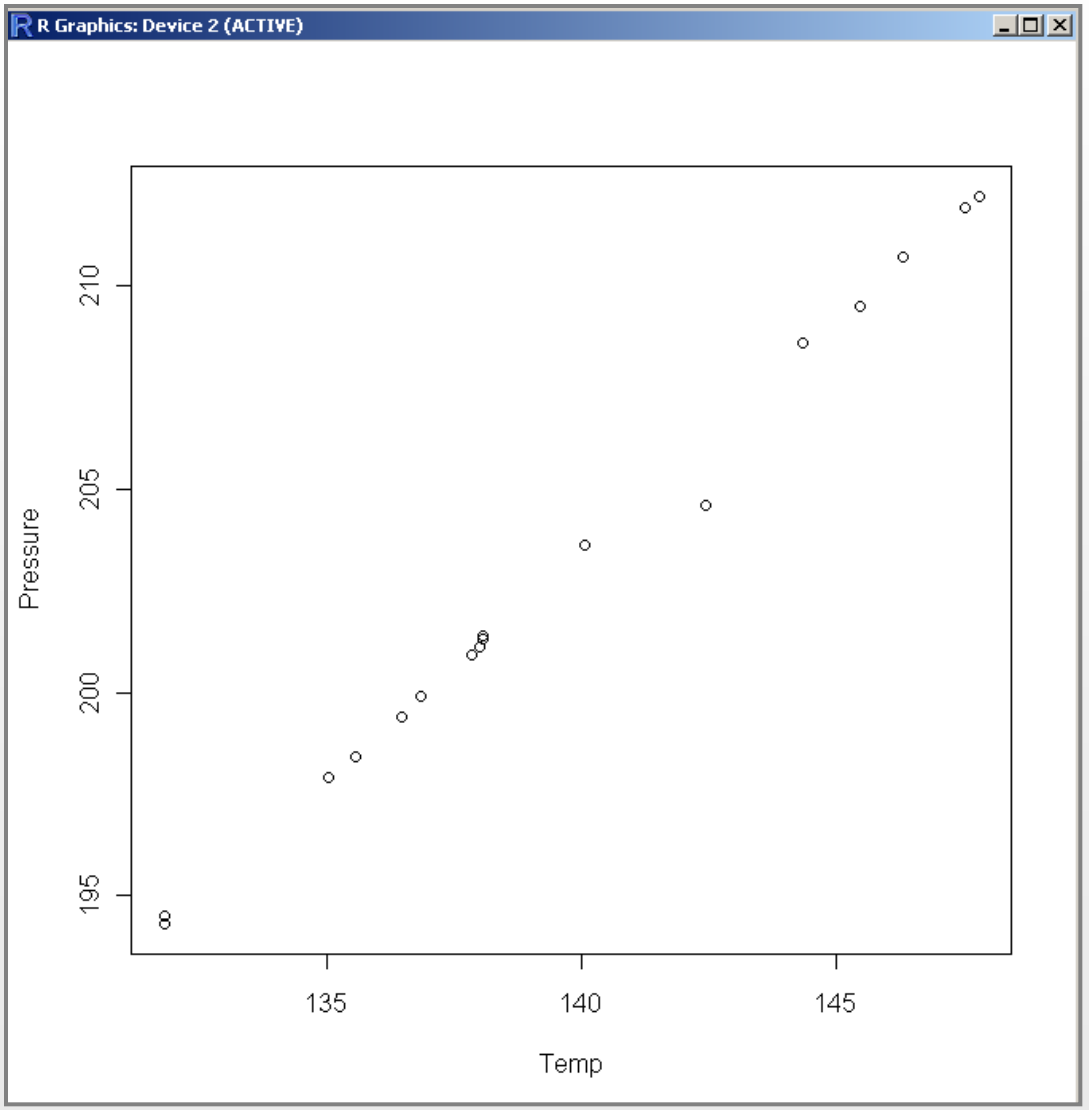
$$y_i = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_m \cdot x_m \ldots + r_i$$

Non-lineaire regressie:

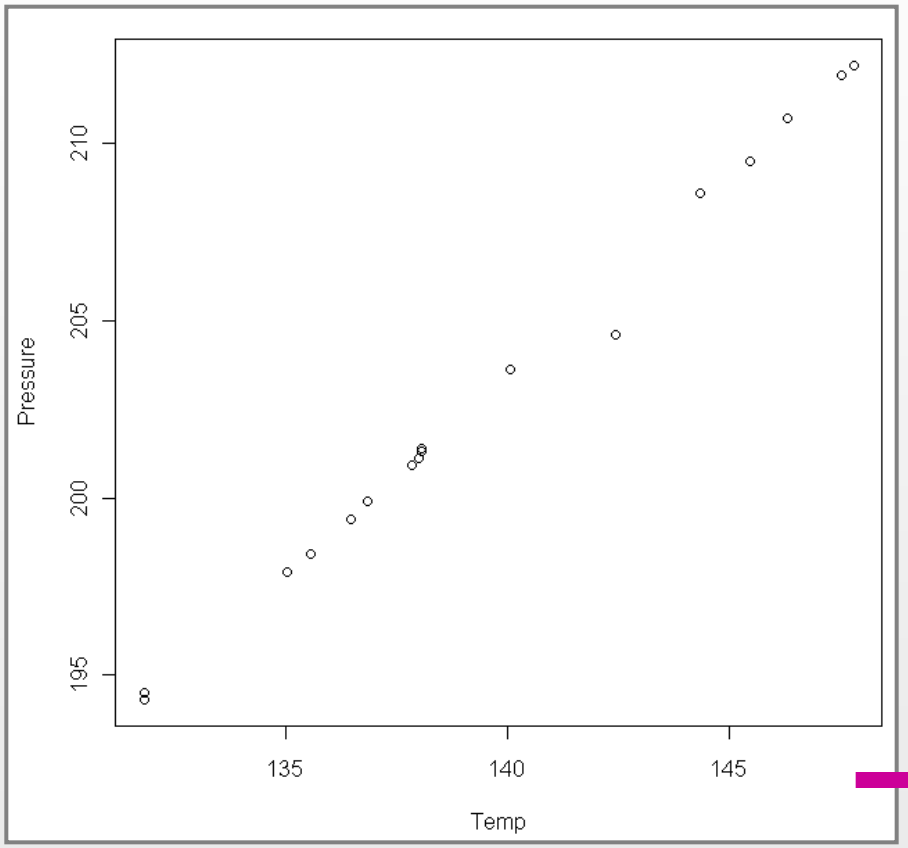
polynomiale regressie, logistische regressie

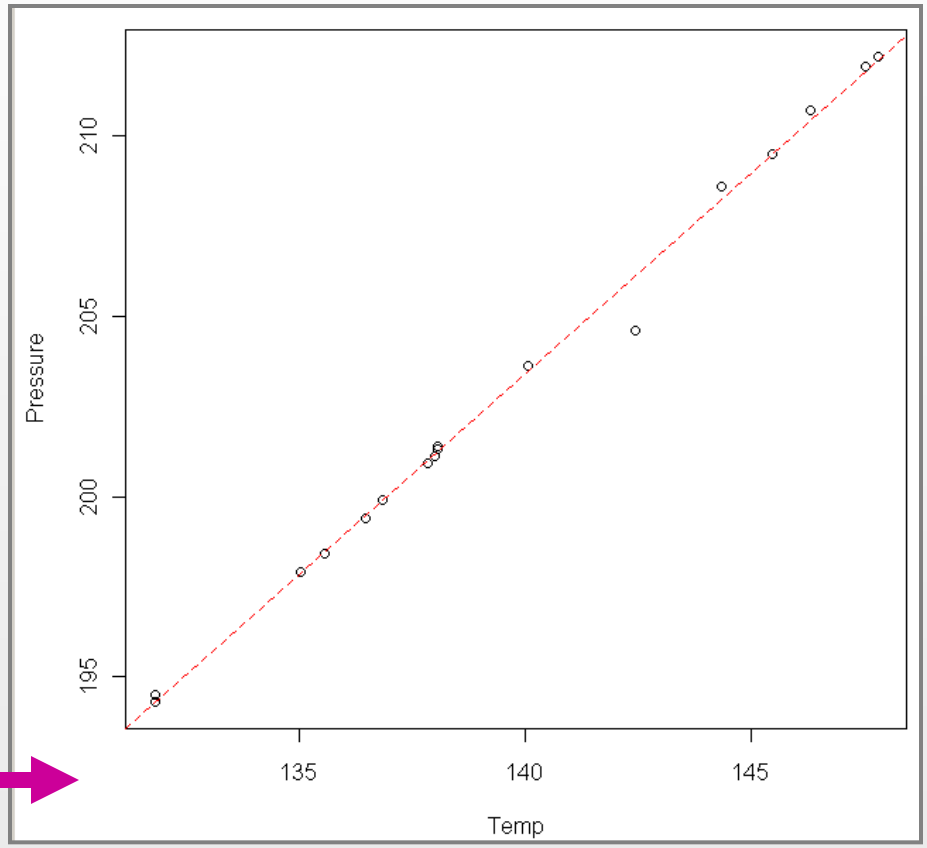
Simpele lineare regressie met **R**

Temp	Pres
194.5	131.79
194.3	131.79
197.9	135.02
198.4	135.55
199.4	136.46
200.9	136.83
201.1	137.82



Simpele lineare regressie met **R**





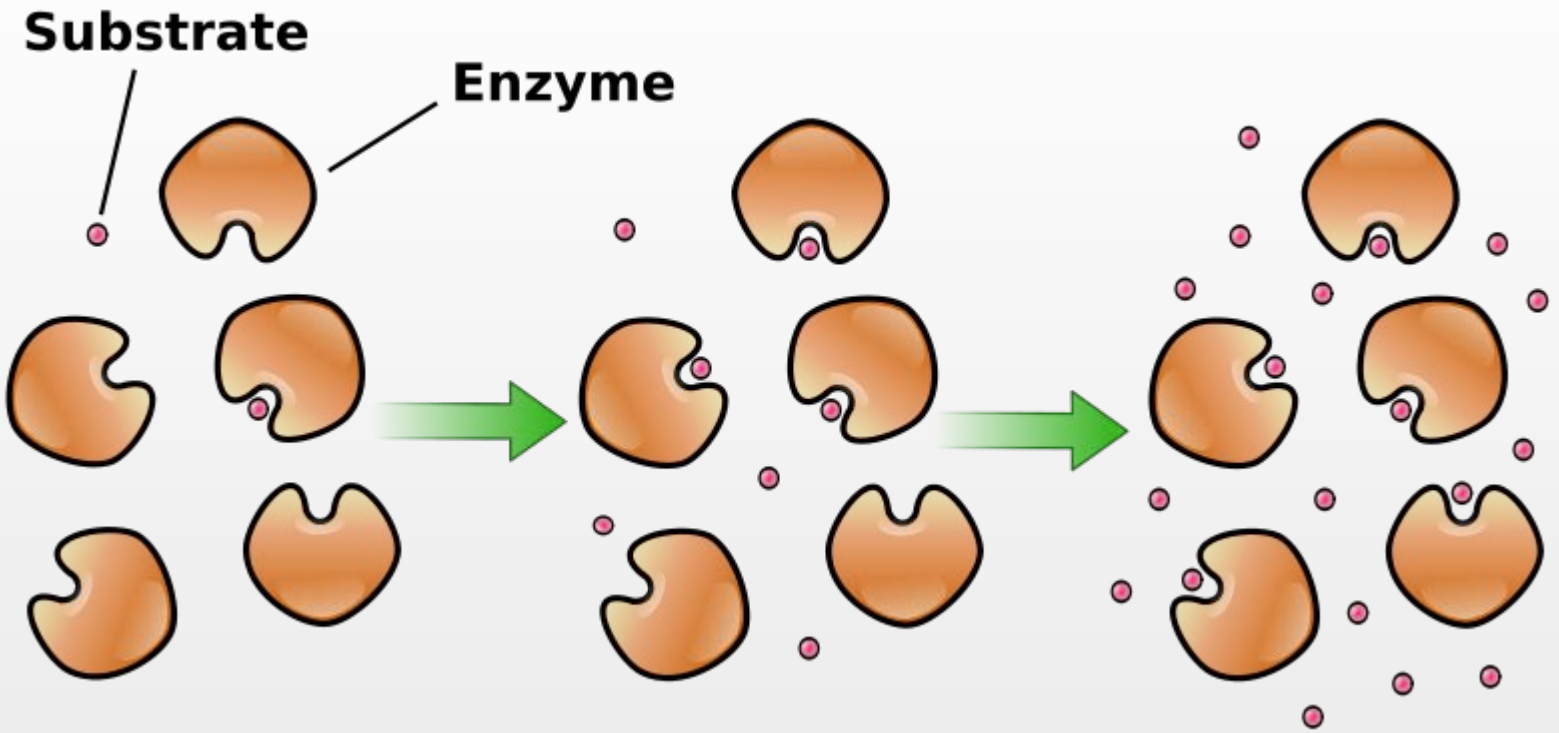
Simpele lineare regressie met **R**

R Console

```
> m1 <- lm(Temp~Lpres,data=forbes)
> plot(Temp~Lpres, data=forbes,xlab="Temp",ylab="Pressure")
> abline(coef(m1),col="red",lty="dashed")
>
```

1. **Fitten** lineaire functie
2. **Plotten** Temp~Lpres
3. **Lijn weergeven** van functie m1

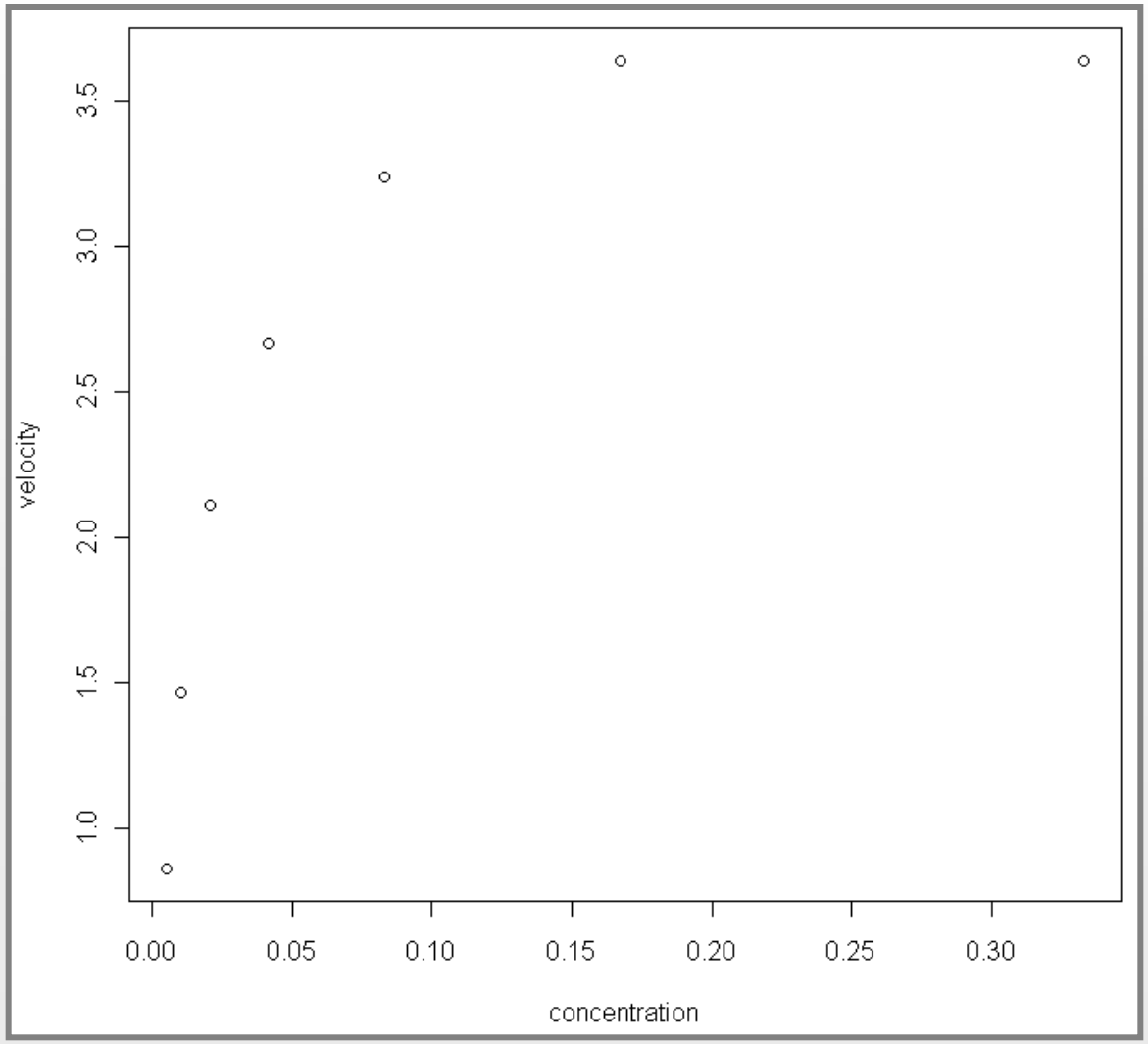
Simpele lineare regressie met **R**



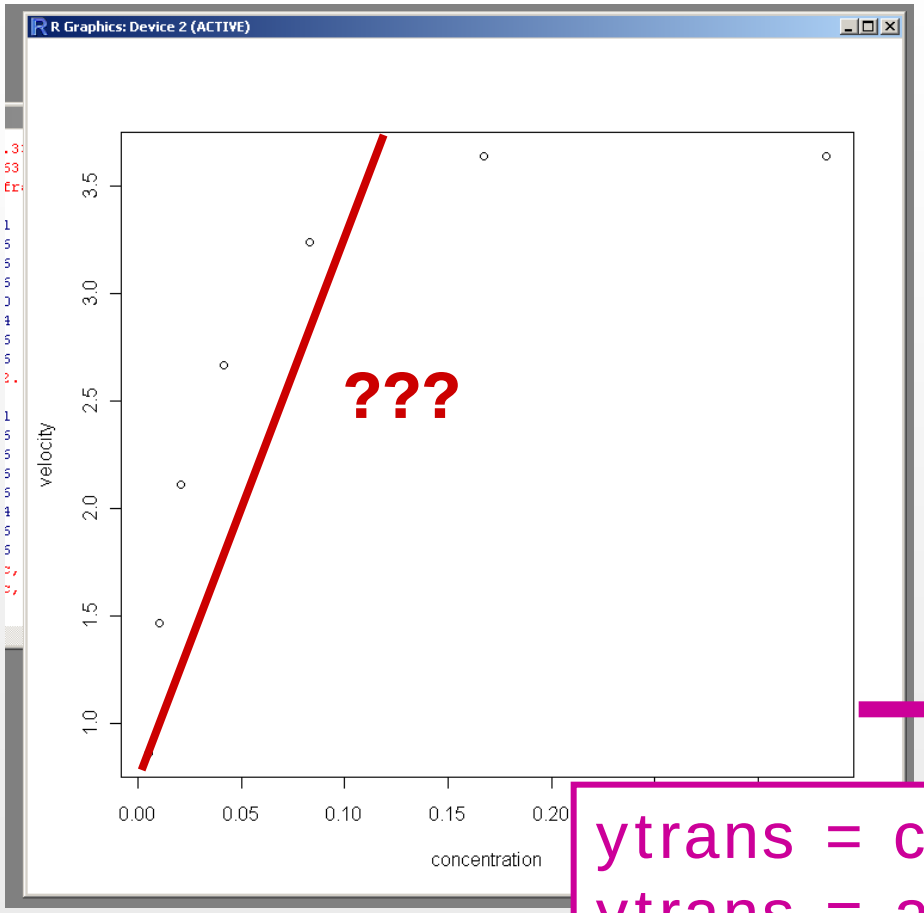
Lage concentratie **Hoge concentratie**

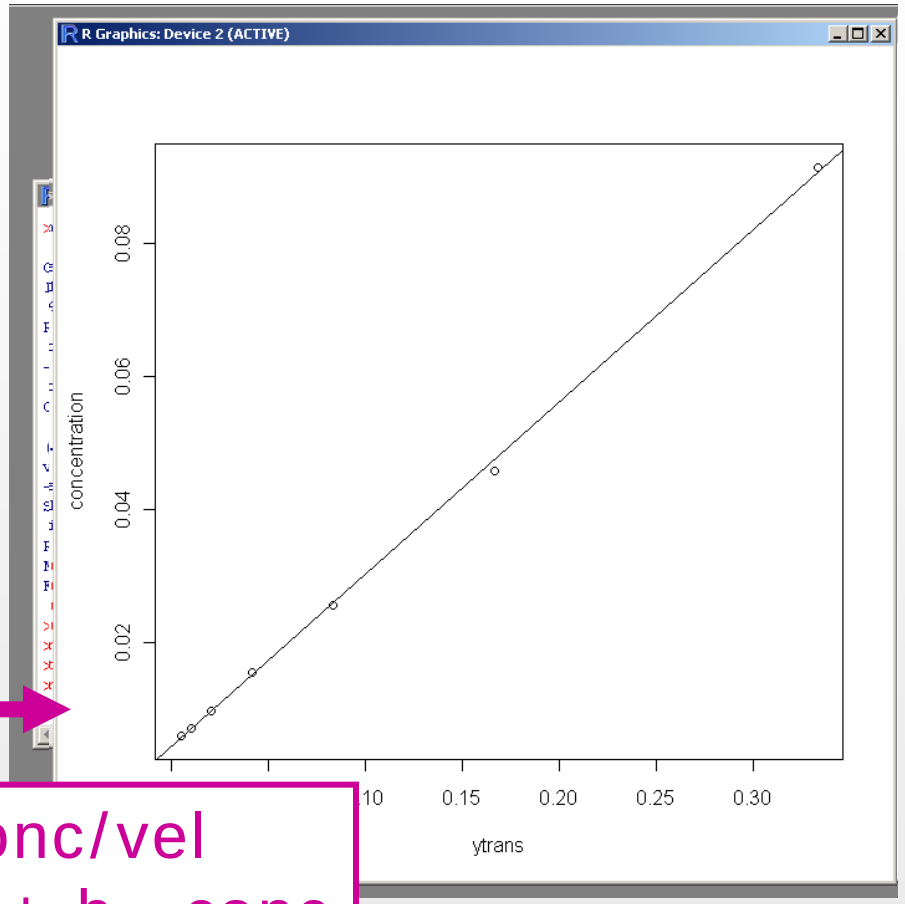
Simpele lineare regressie met **R**

Concentration	Velocity
0.3330	3.636
0.1670	3.636
0.0833	3.236
0.0416	2.666
0.0208	2.114
0.0104	1.466
0.0052	0.866



Simpele lineare regressie met **R**

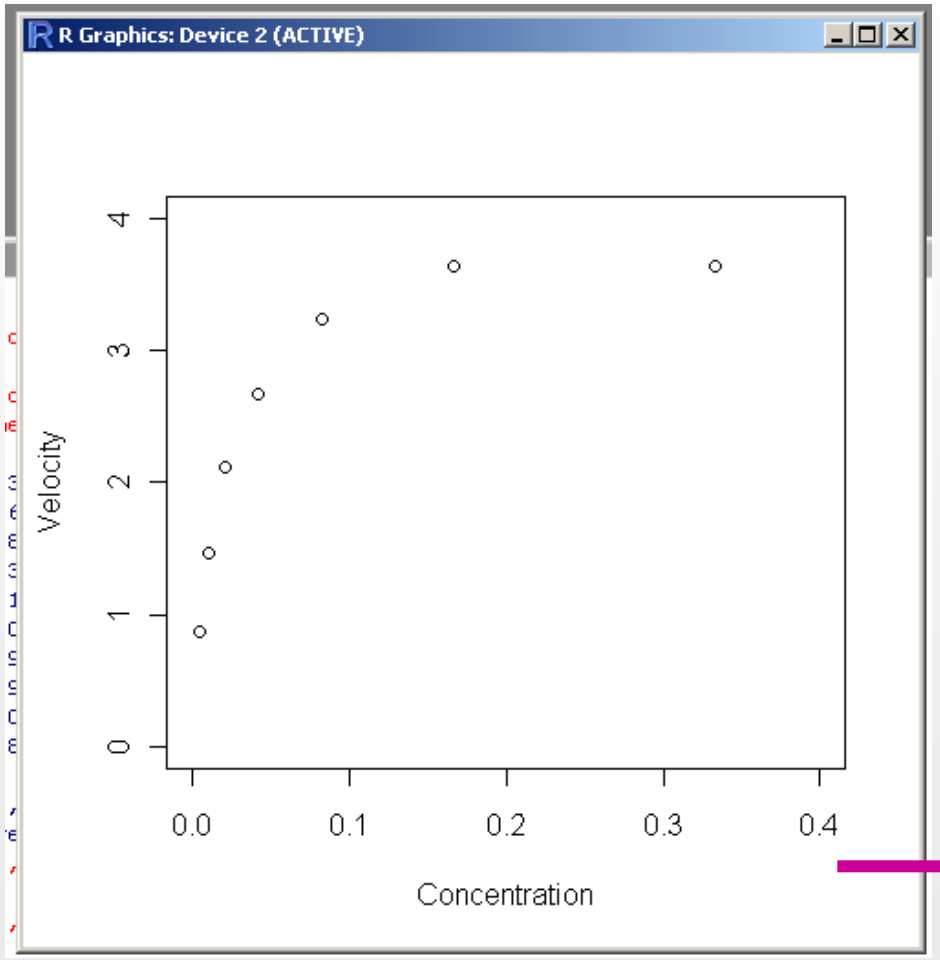


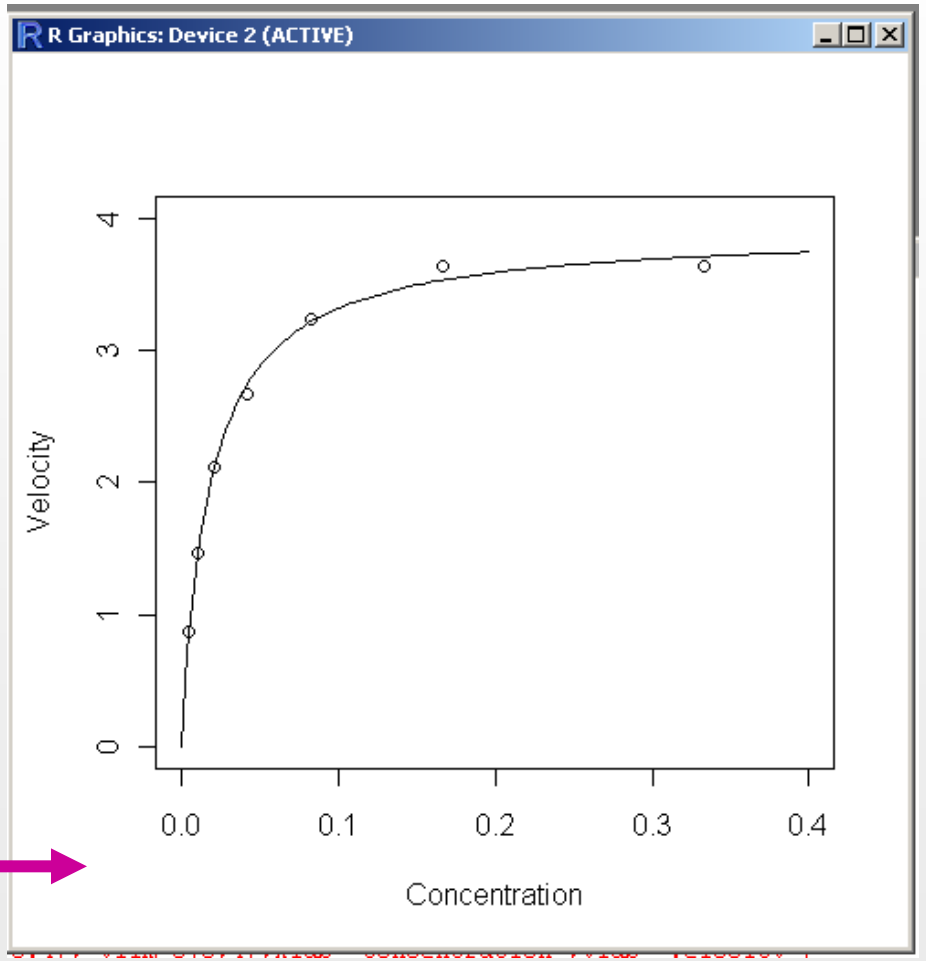


ytrans = conc/vel
ytrans = a + b . conc

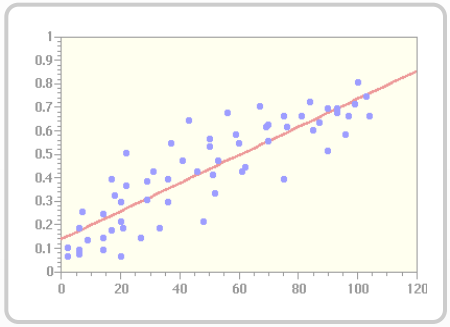
Transformatie

Non-Lineare Regressie **in R**

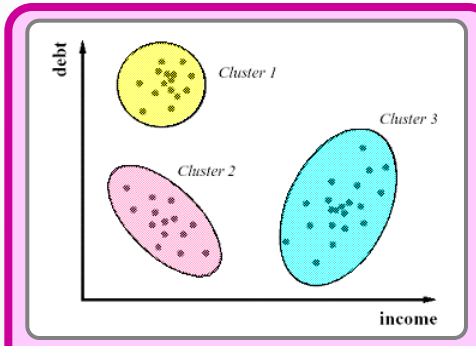




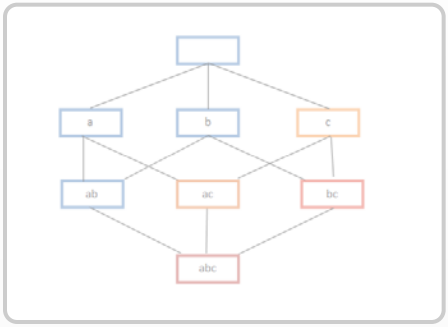
5 predictieve modellen



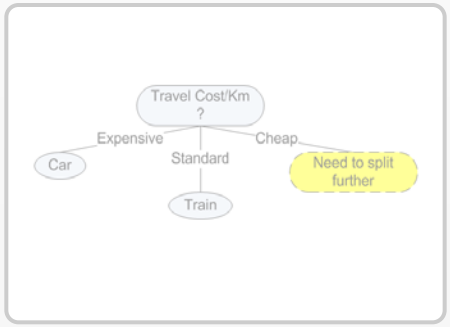
Statistics



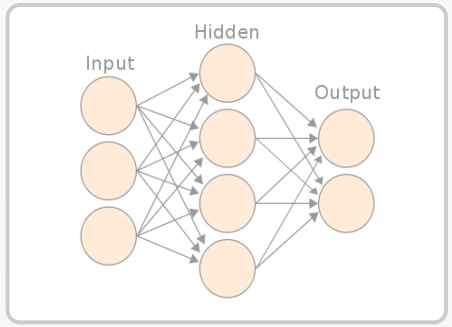
Clustering



Association rules



Decision trees

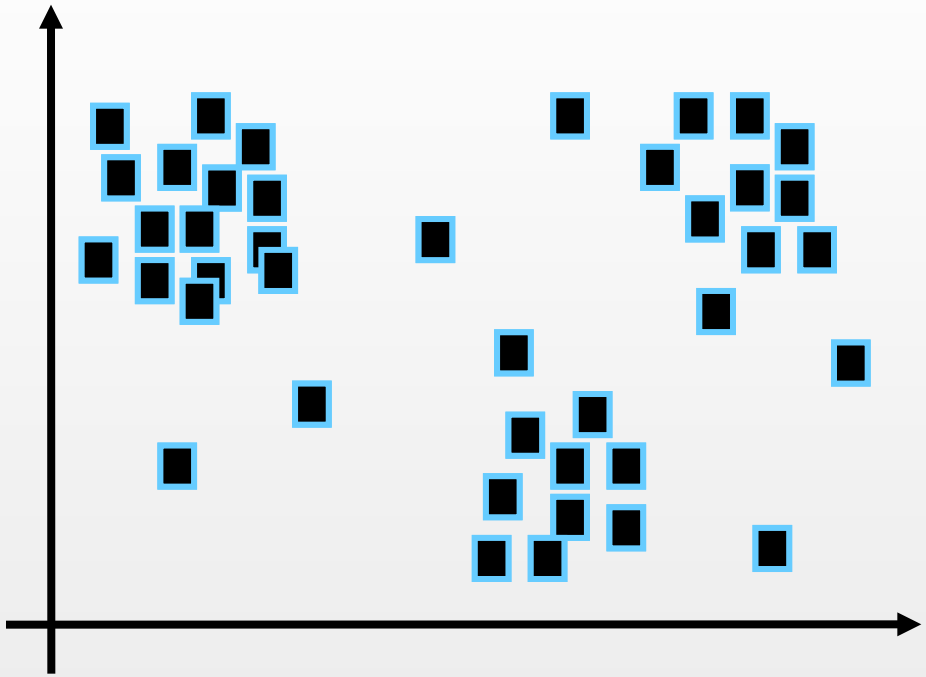


Neural networks

Clustering

Clustering: **K-Means algoritme**

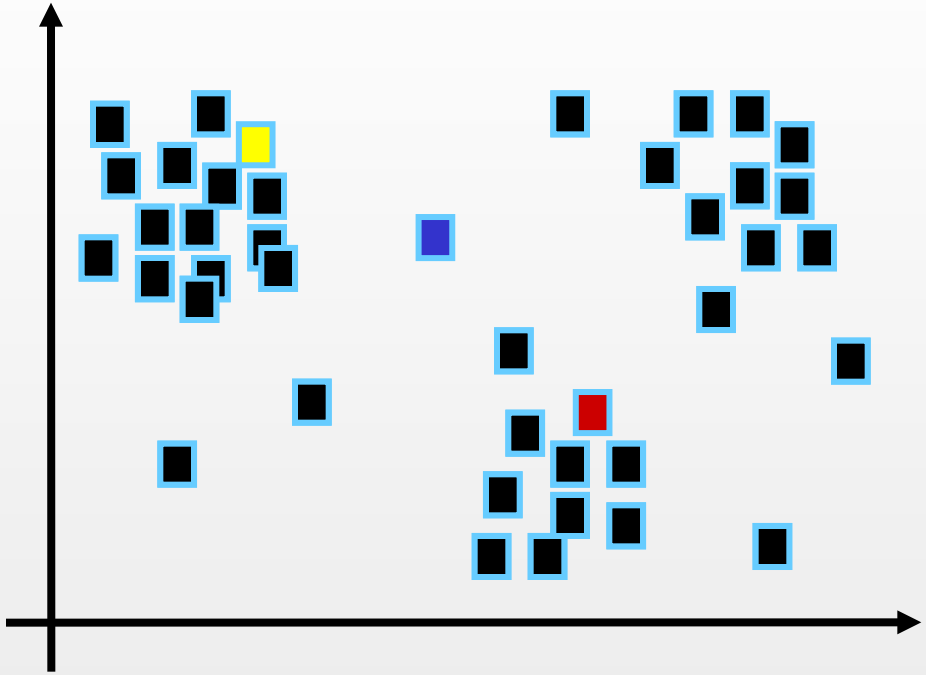
Input: aantal clusters **N**



Clustering: **K-Means algoritme**

Input: aantal clusters **N**

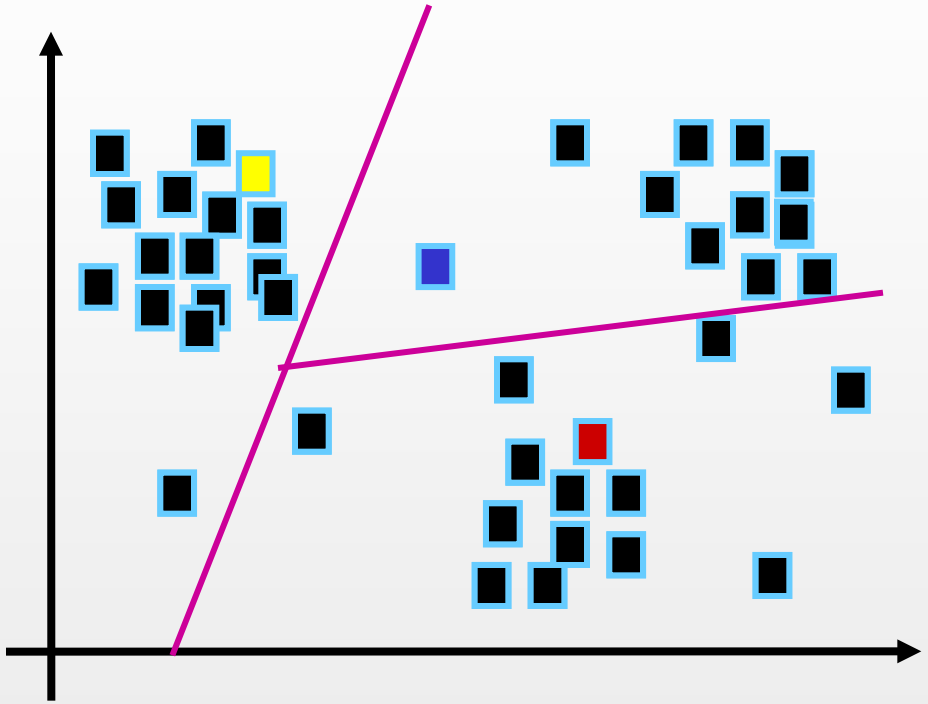
1. Kies N random cluster centra



Clustering: **K-Means algoritme**

Input: aantal clusters **N**

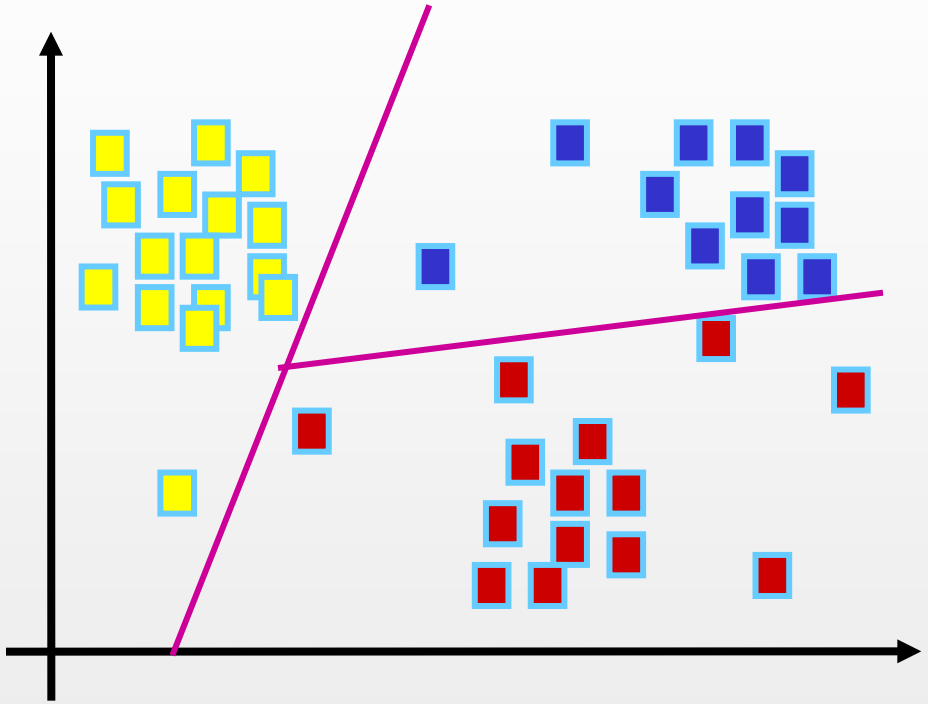
1. Kies N random cluster centra
2. Verdeel in clusters



Clustering: **K-Means algoritme**

Input: aantal clusters **N**

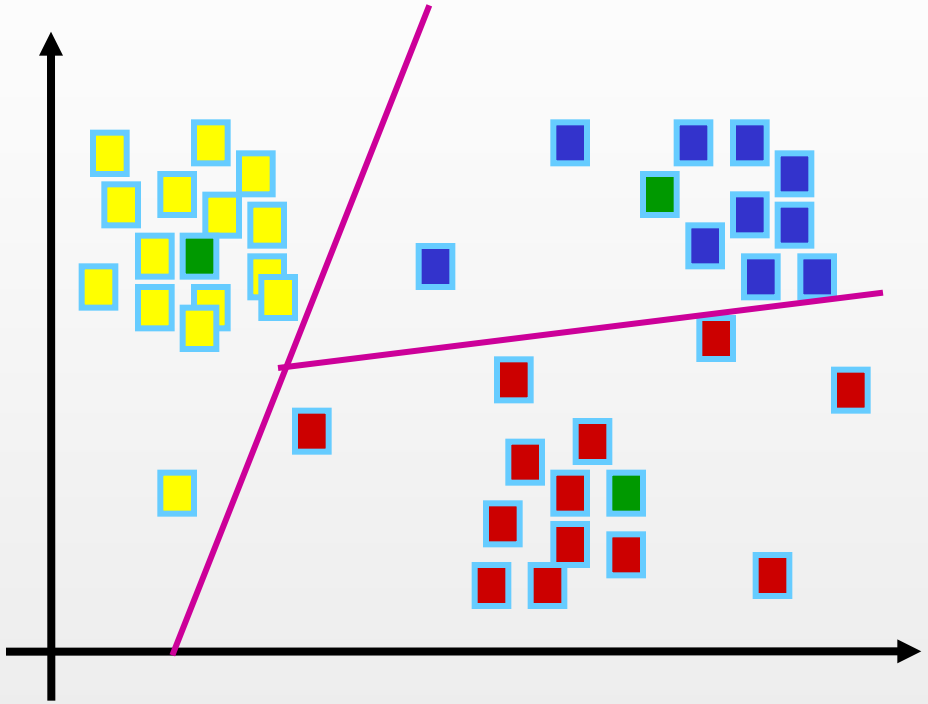
1. Kies N random cluster centra
2. Verdeel in clusters



Clustering: **K-Means algoritme**

Input: aantal clusters **N**

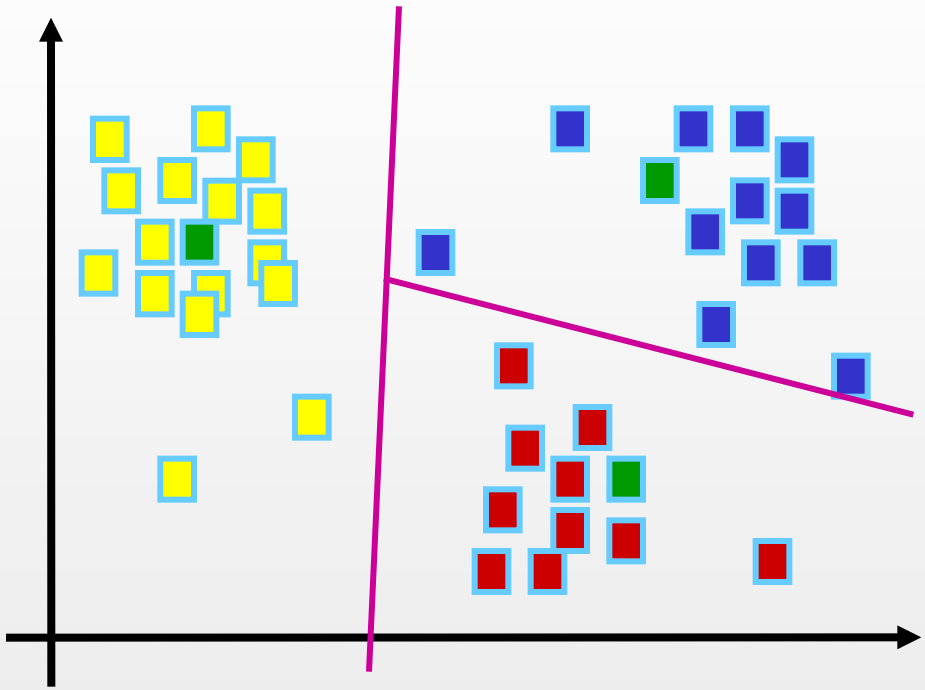
1. Kies N random cluster centra
2. Verdeel in clusters
3. Bepaal N nieuwe cluster centra



Clustering: **K-Means algoritme**

Input: aantal clusters **N**

1. Kies N random cluster centra
2. Verdeel in clusters
3. Bepaal N nieuwe cluster centra
4. N clusters gevonden



Clustering voorbeeld: IRIS dataset

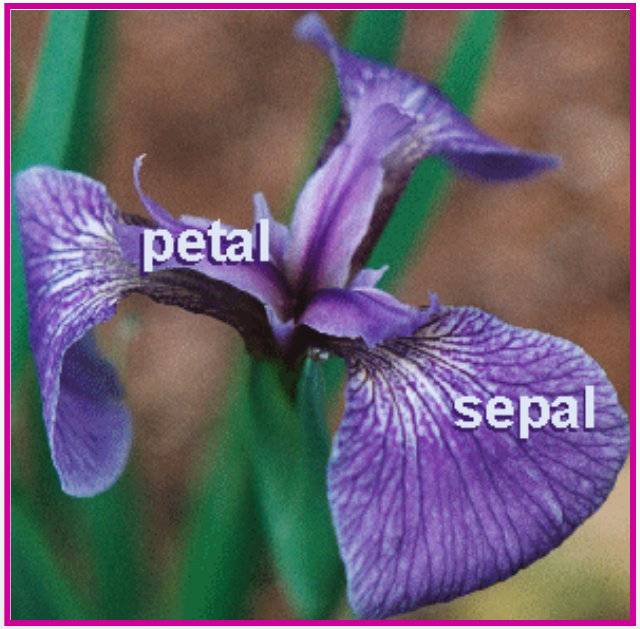
Iris setosa

Iris versicolor

Iris virginica

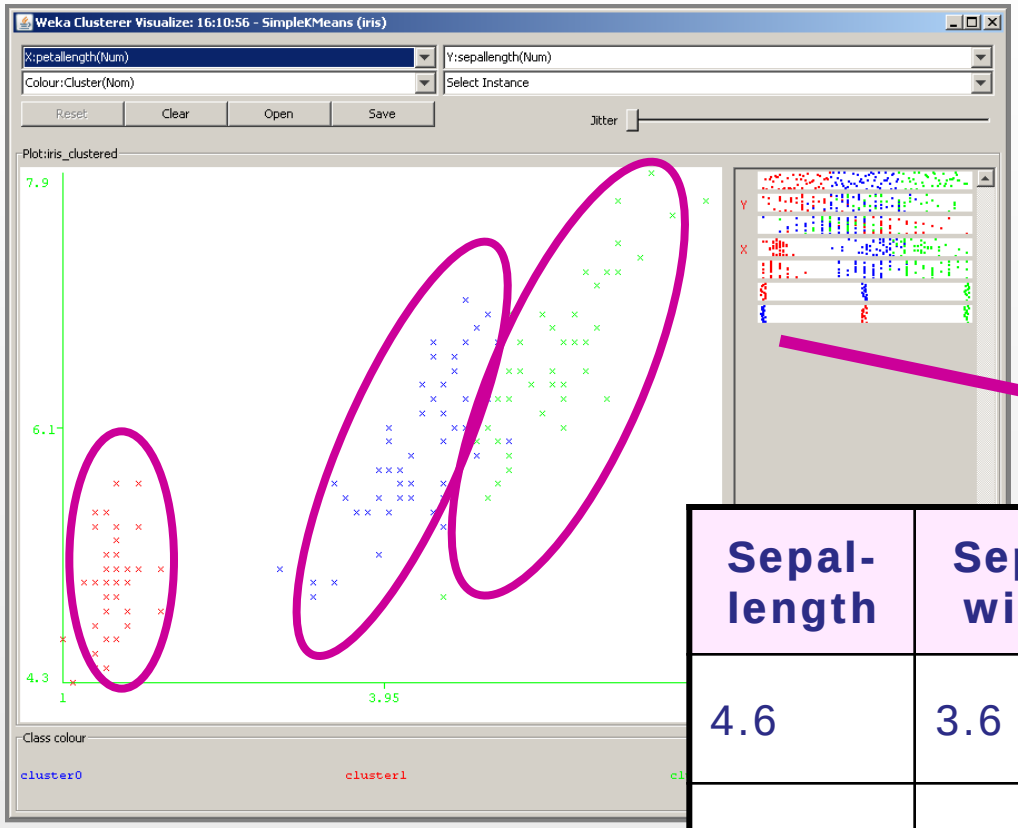
Clustering voorbeeld: IRIS dataset

Sepal-length	Sepal-width	Petal-length	Petal-width	Class
4.6	3.6	1.0	0.2	Setosa
4.3	3.0	1.1	0.1	Setosa
5.1	2.5	3.0	1.1	Versicolor
5.7	2.6	3.5	1.0	Versicolor
5.8	2.8	5.1	2.4	Viriginica
6.3	2.8	5.1	2.4	Virginica
6.9	3.1	5.1	2.3	Virginica
...	...	...	...	...

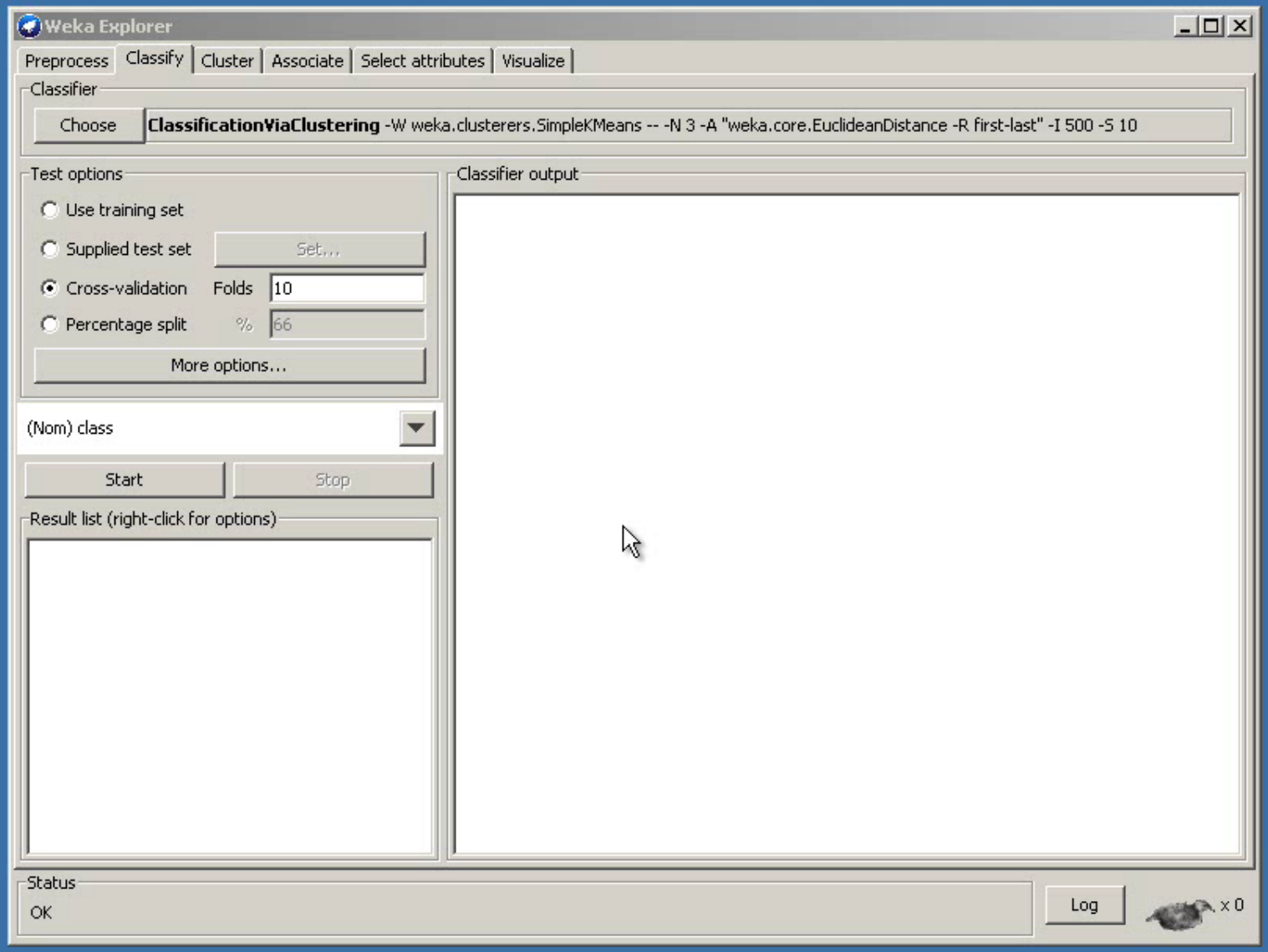


Clustering in WEKA

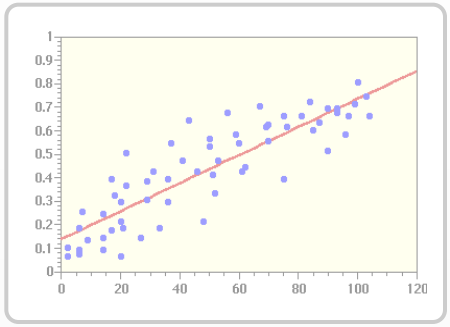
Voorspellen met Clustering



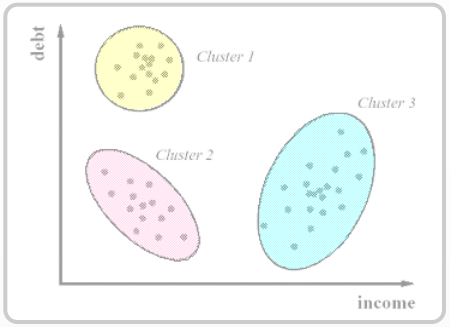
Sepal-length	Sepal-width	Petal-length	Petal-width	Class
4.6	3.6	1.0	0.2	?
4.3	3.0	1.1	0.1	?
5.1	2.5	3.0	1.1	?


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ClassificationViaClustering -W weka.clusterers.SimpleKMeans -- -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) class
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

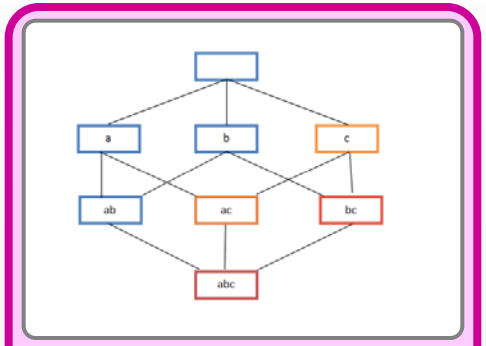
5 predictieve modellen



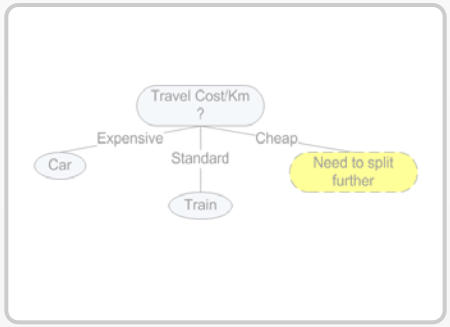
Statistics



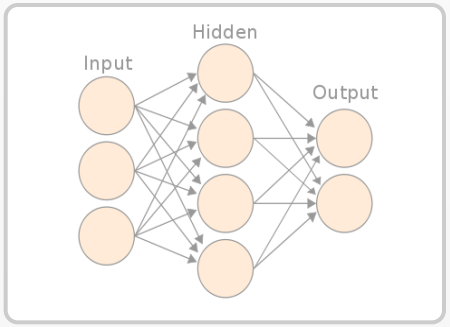
Clustering



Association rules



Decision trees



Neural networks

Association Rules

ID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Voorbeeld:
Associaties tussen aankopen

{Butter, Bread} => {Milk}
{Milk} => {Bread}

IF (Butter AND Bread) THEN Milk
IF Milk THEN Bread

Association Rules

{Butter, Bread} => **{Milk}**

Antecedent

Consequent

Support:
Hoeveel keer (in %) komt het antecedent voor?

Confidence:
Hoeveel keer (in %) volgt de consequent op de antecedent ?

{Bread} => {Milk}

ID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Support: 3/5 = 60%

Confidence: 2/3 = 66,7%

Associaties met het **Apriori Algoritme**

Input: minimum support N

1. Bepaal **support** van elk element

ID	Support
{Milk}	0,4
{Bread}	0,6
{Butter}	0,4
{Beer}	0,2

2. Schrap elementen met **support** < N

Associaties met het **Apriori Algoritme**

Input: minimum support N

ID	Support
{Milk}	0,4
{Bread}	0,6
{Butter}	0,4
{Beer}	0,2

3 Breid associaties uit met 1 element en bepaal **support**

ID	Support
{Milk,Bread}	0,4
{Milk,Butter}	0,2
{Bread,Butter}	0,2

4 Schrap elementen met **support < N**

5 Herhaal tot er geen associaties met **support >= N** gevonden worden...

Association rules in bankgegevens met **SPSS**

Kinderen **Auto** **Spaar rekening** **Lopende rekening** **Hypotheek** **PEP**

Leeftijd

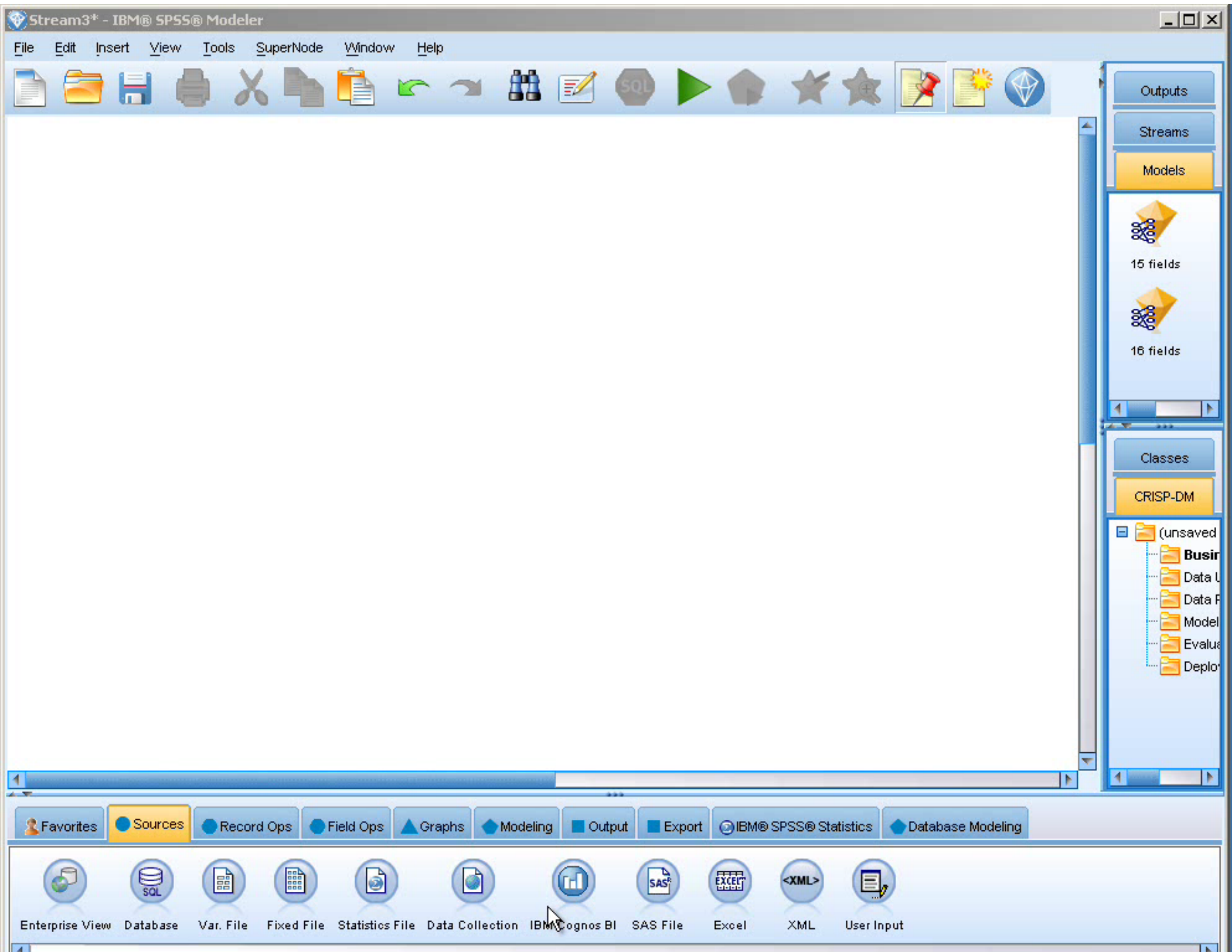
Geslacht

Regio

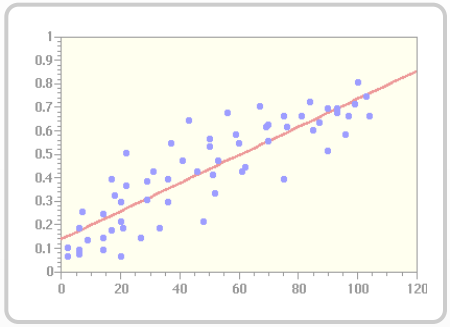
Inkomen

Getrouwd

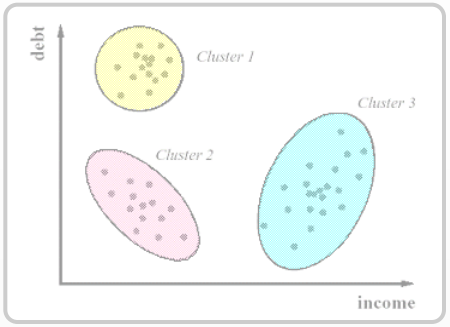
age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
48	FEMALE	INNER_CITY	17546.000	NO	1	NO	NO	NO	NO	YES
40	MALE	TOWN	30085.100	YES	3	YES	NO	YES	YES	NO
51	FEMALE	INNER_CITY	16575.400	YES	0	YES	YES	YES	NO	NO
23	FEMALE	TOWN	20375.400	YES	3	NO	NO	YES	NO	NO
57	FEMALE	RURAL	50576.300	YES	0	NO	YES	NO	NO	NO
57	FEMALE	TOWN	37869.600	YES	2	NO	YES	YES	NO	YES
22	MALE	RURAL	8877.070	NO	0	NO	NO	YES	NO	YES
58	MALE	TOWN	24946.600	YES	0	YES	YES	YES	NO	NO
37	FEMALE	SUBURBAN	25304.300	YES	2	YES	NO	NO	NO	NO
54	MALE	TOWN	24212.100	YES	2	YES	YES	YES	NO	NO
66	FEMALE	TOWN	59803.900	YES	0	NO	YES	YES	NO	NO
52	FEMALE	INNER_CITY	26658.800	NO	0	YES	YES	YES	YES	NO
44	FEMALE	TOWN	15735.800	YES	1	NO	YES	YES	YES	YES
66	FEMALE	TOWN	55204.700	YES	1	YES	YES	YES	YES	YES
36	MALE	RURAL	19474.600	YES	0	NO	YES	YES	YES	NO
38	FEMALE	INNER_CITY	22342.100	YES	0	YES	YES	YES	YES	NO
37	FEMALE	TOWN	17729.800	YES	2	NO	NO	NO	YES	NO
46	FEMALE	SUBURBAN	41016.000	YES	0	NO	YES	NO	YES	NO


Stream3* - IBM® SPSS® Modeler
File
Edit
Insert
View
Tools
SuperNode
Window
Help
Outputs
Streams
Models
15 fields
16 fields
Classes
CRISP-DM
(unsaved
Busir
Data U
Data P
Model
Evalua
Deplo
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Database Modeling
Enterprise View
Database
Var. File
Fixed File
Statistics File
Data Collection
IBM Cognos BI
SAS File
Excel
XML
User Input

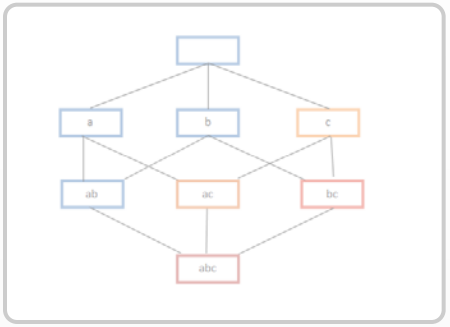
5 predictieve modellen



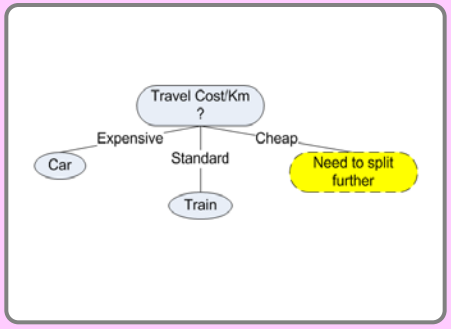
Statistics



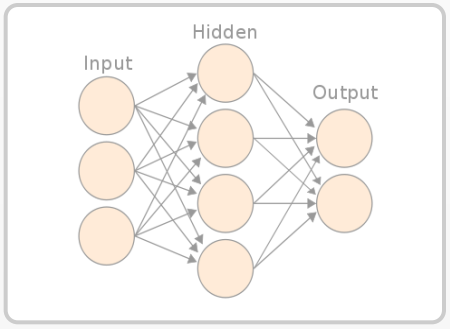
Clustering



Association rules



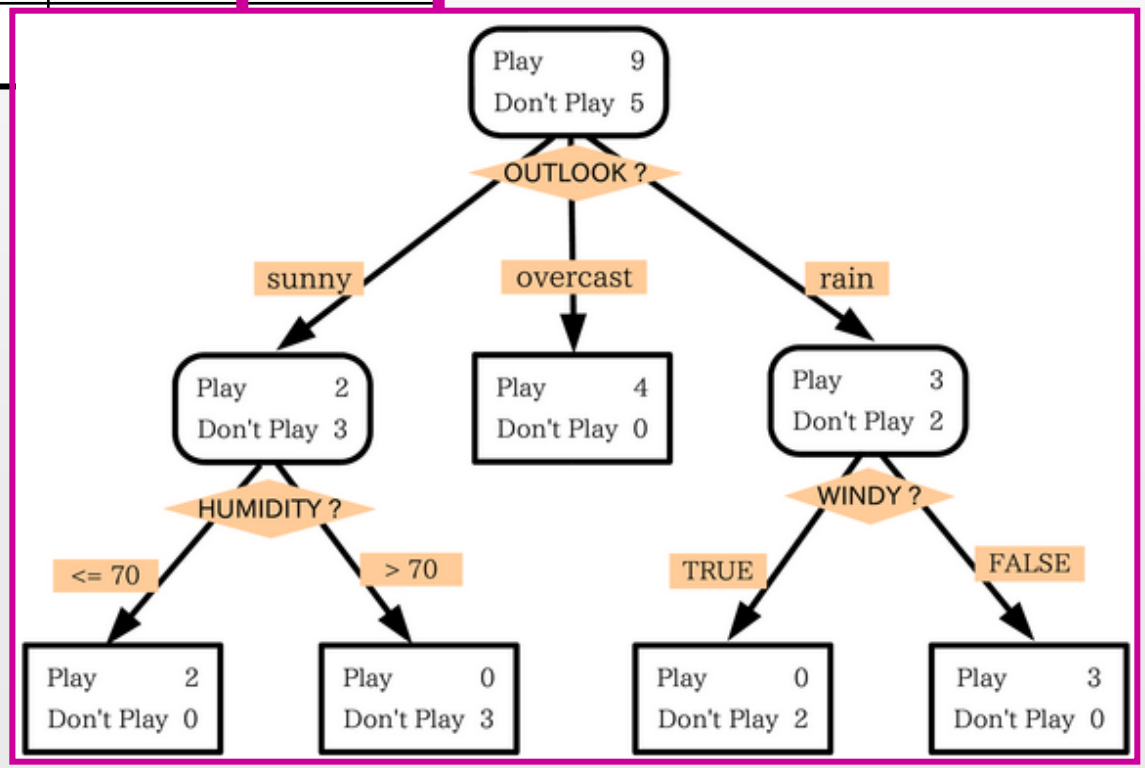
Decision trees



Neural networks

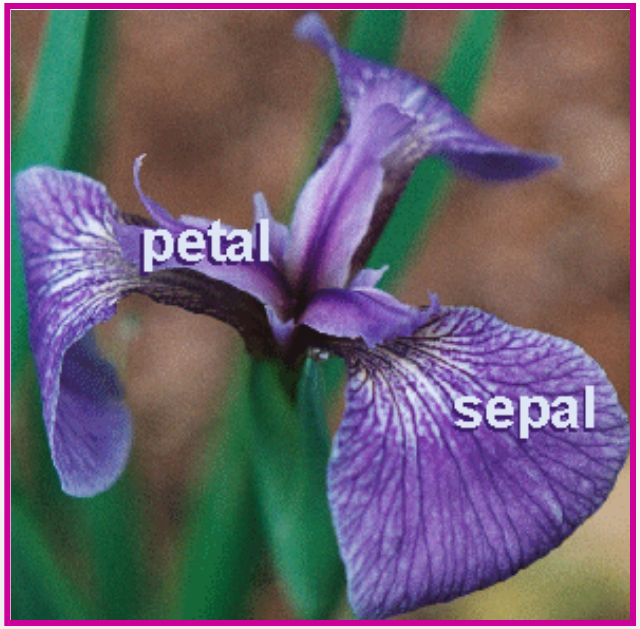
Decision trees

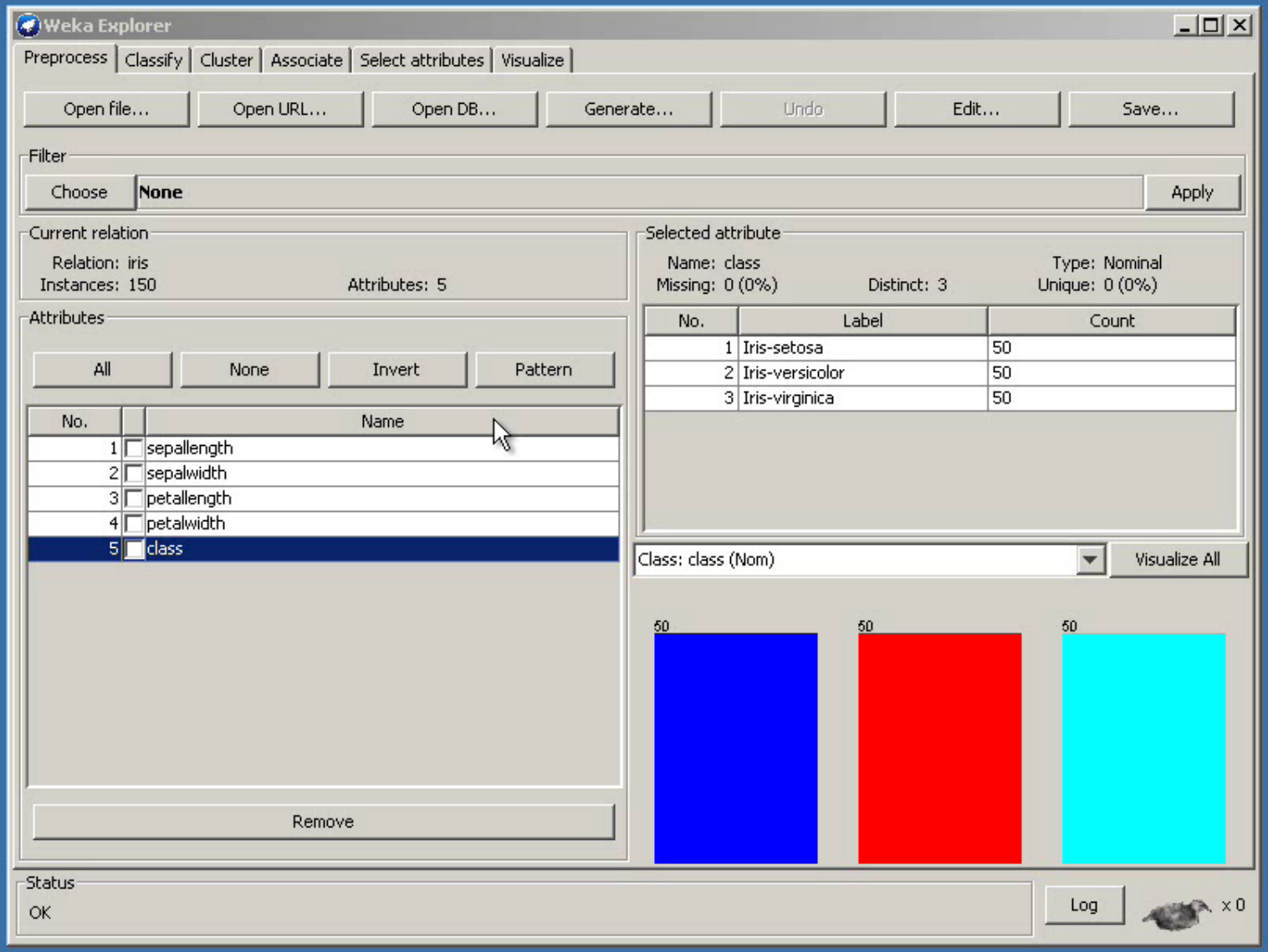
Outlook	Temp.	Hum.	Wind	Play?
Sunny	85	85	F	N
Sunny	80	90	T	N
Overcast	83	86	F	Y
...	...	...		



Decision trees met **Weka**

Sepal-length	Sepal-width	Petal-length	Petal-width	Class
4.6	3.6	1.0	0.2	Setosa
4.3	3.0	1.1	0.1	Setosa
5.1	2.5	3.0	1.1	Versicolor
5.7	2.6	3.5	1.0	Versicolor
5.8	2.8	5.1	2.4	Viriginica
6.3	2.8	5.1	2.4	Virginica
6.9	3.1	5.1	2.3	Virginica
...	...	...	...	...




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
No.
Label
Count
1 Iris-setosa 50
2 Iris-versicolor 50
3 Iris-virginica 50
Class: class (Nom)
Visualize All
50
50
50
Status
OK
Log
x 0

Decision trees in **R**

```
> iris <- read.table("c:/iris.csv",header=T,sep=",")
```

1 Inlezen data

Decision trees in **R**

```
> iris <- read.table("c:/iris.csv",header=T,sep=",")
> fit <- rpart(class~sepallength+sepalwidth+petallength+petalwidth
      , data=iris,method="class")
```

2 Fitten decision tree

Decision trees in **R**

```
> iris <- read.table("c:/iris.csv",header=T,sep=",")
> fit <- rpart(class~sepallength+sepalwidth+petallength+petalwidth
      , data=iris,method="class")
> plot(fit, uniform=TRUE, main="Iris tree")
> text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

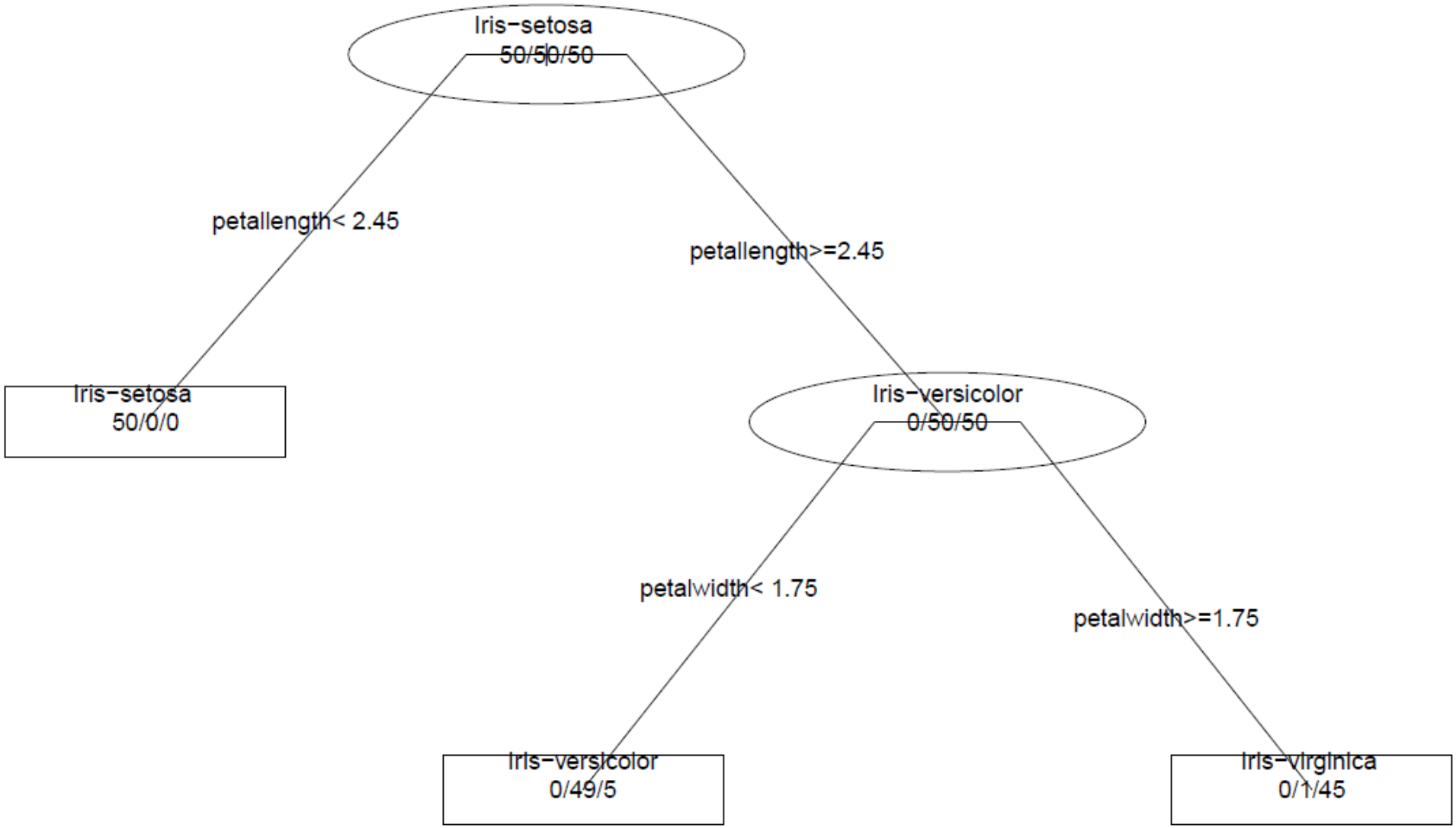
3 Plotten resultaat

Decision trees in **R**

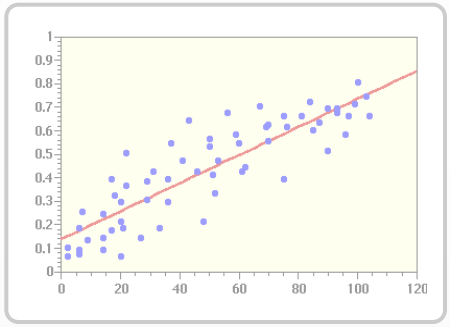
```
> iris <- read.table("c:/iris.csv",header=T,sep=",")
> fit <- rpart(class~sepallength+sepalwidth+petallength+petalwidth
      , data=iris,method="class")
> plot(fit, uniform=TRUE, main="Iris tree")
> text(fit, use.n=TRUE, all=TRUE, cex=.8)
> pred <- predict(fit, newData=data,type="class")
> mc <-table(iris$class,pred)
> print(mc)
                Iris-setosa Iris-versicolor Iris-virginica
 Iris-setosa             50              0             0
 Iris-versicolor          0             49             1
 Iris-virginica           0              5            45
> err <- mc[1,2]+mc[1,3]+mc[2,1]+mc[2,3]+mc[3,1]+mc[3,2]
> 1 - err/length(iris$class)
[1] 0.96
>
```

4 Evalueren resultaat

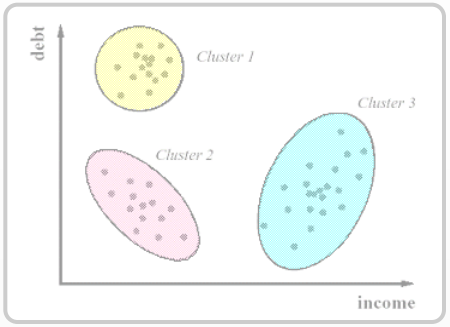
Decision trees in **R**



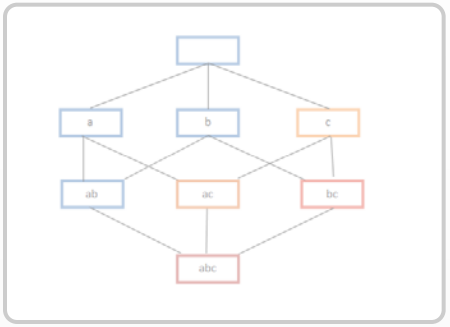
5 predictieve modellen



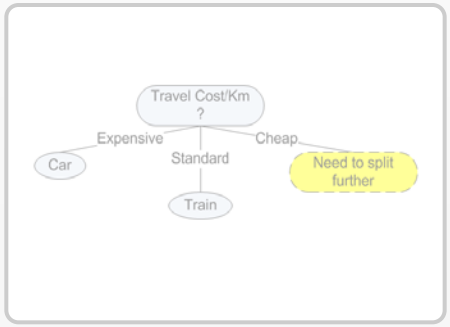
Statistics



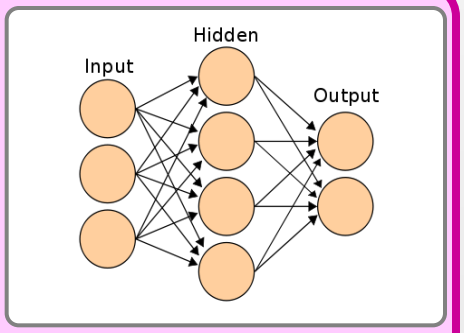
Clustering



Association rules

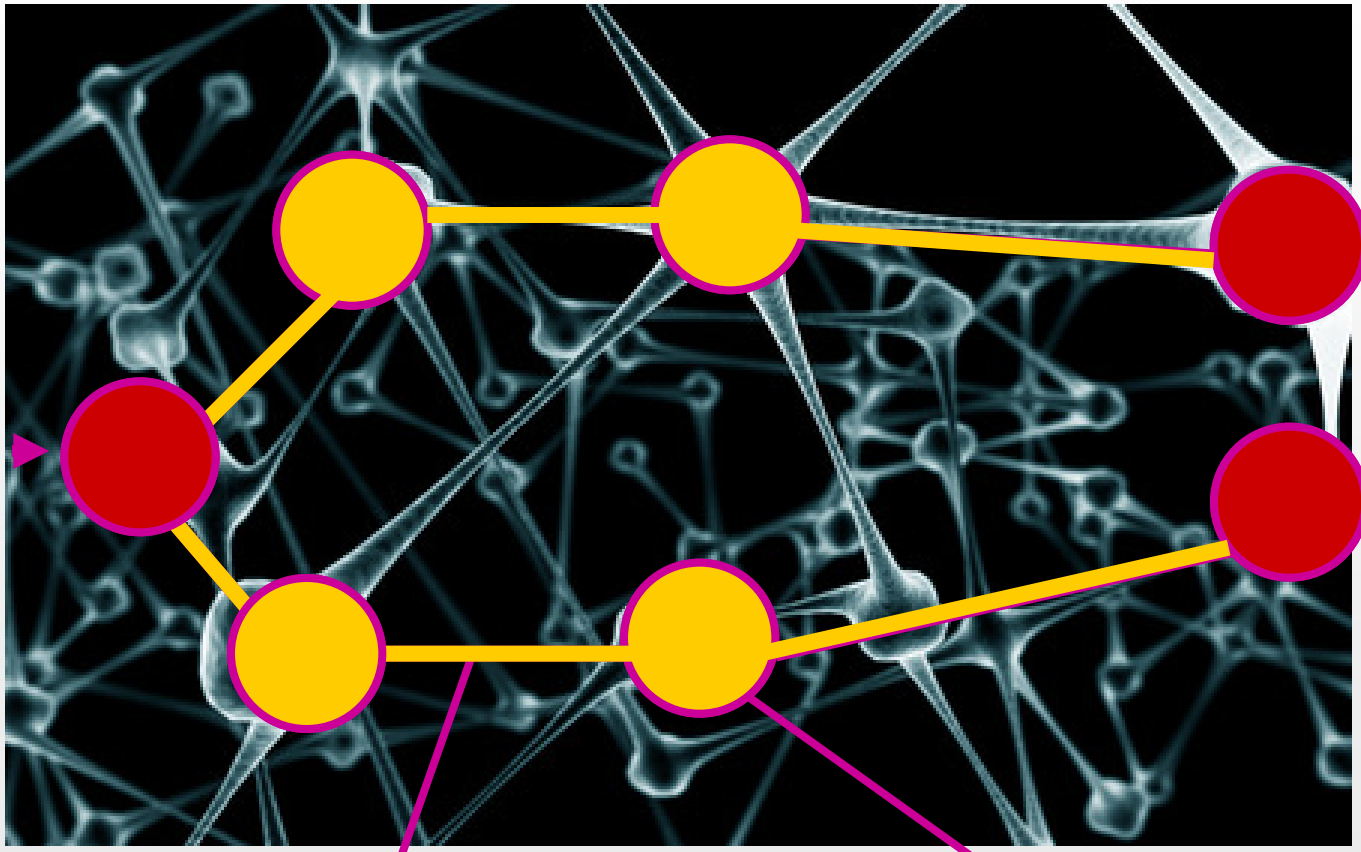


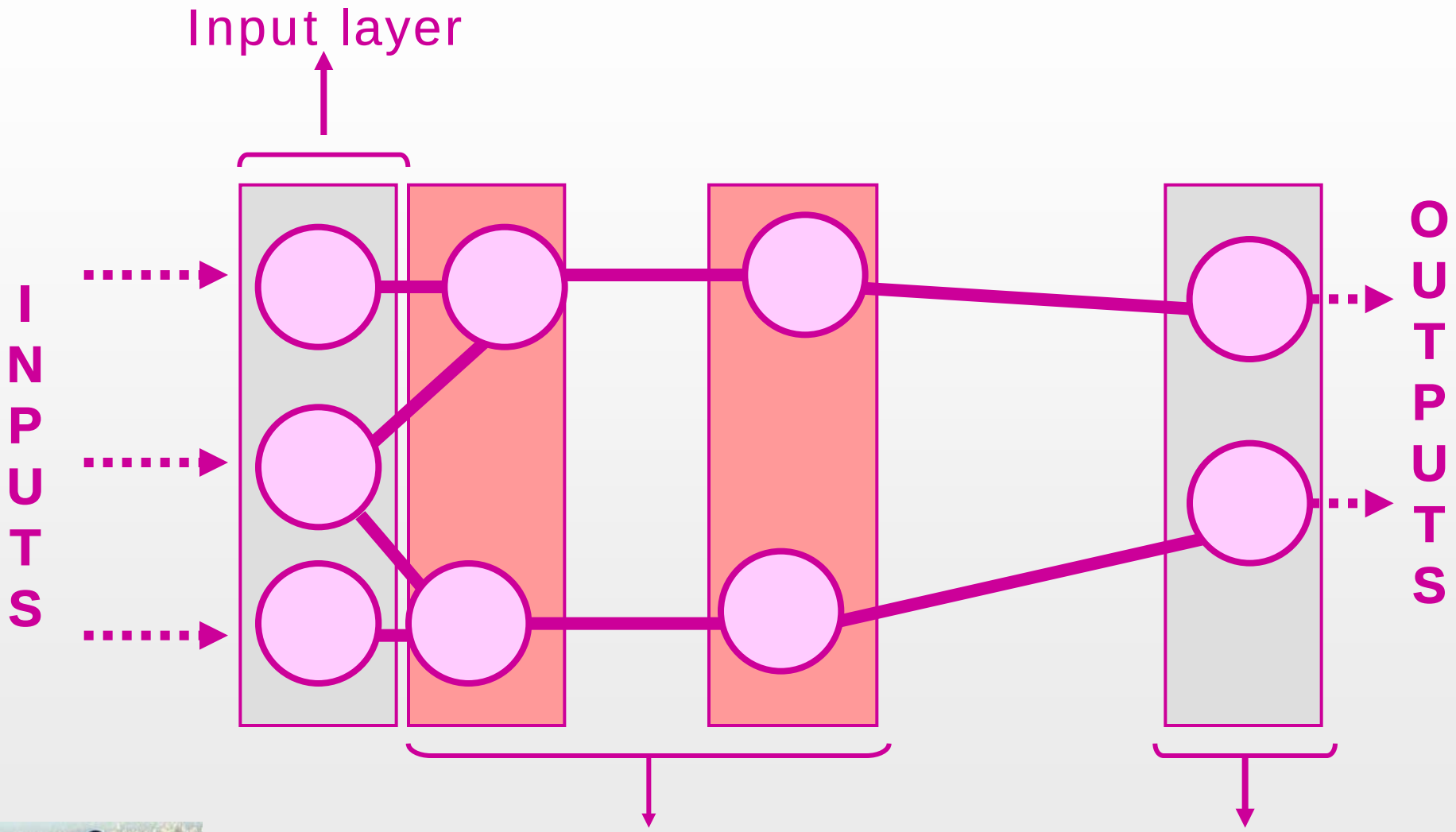
Decision trees



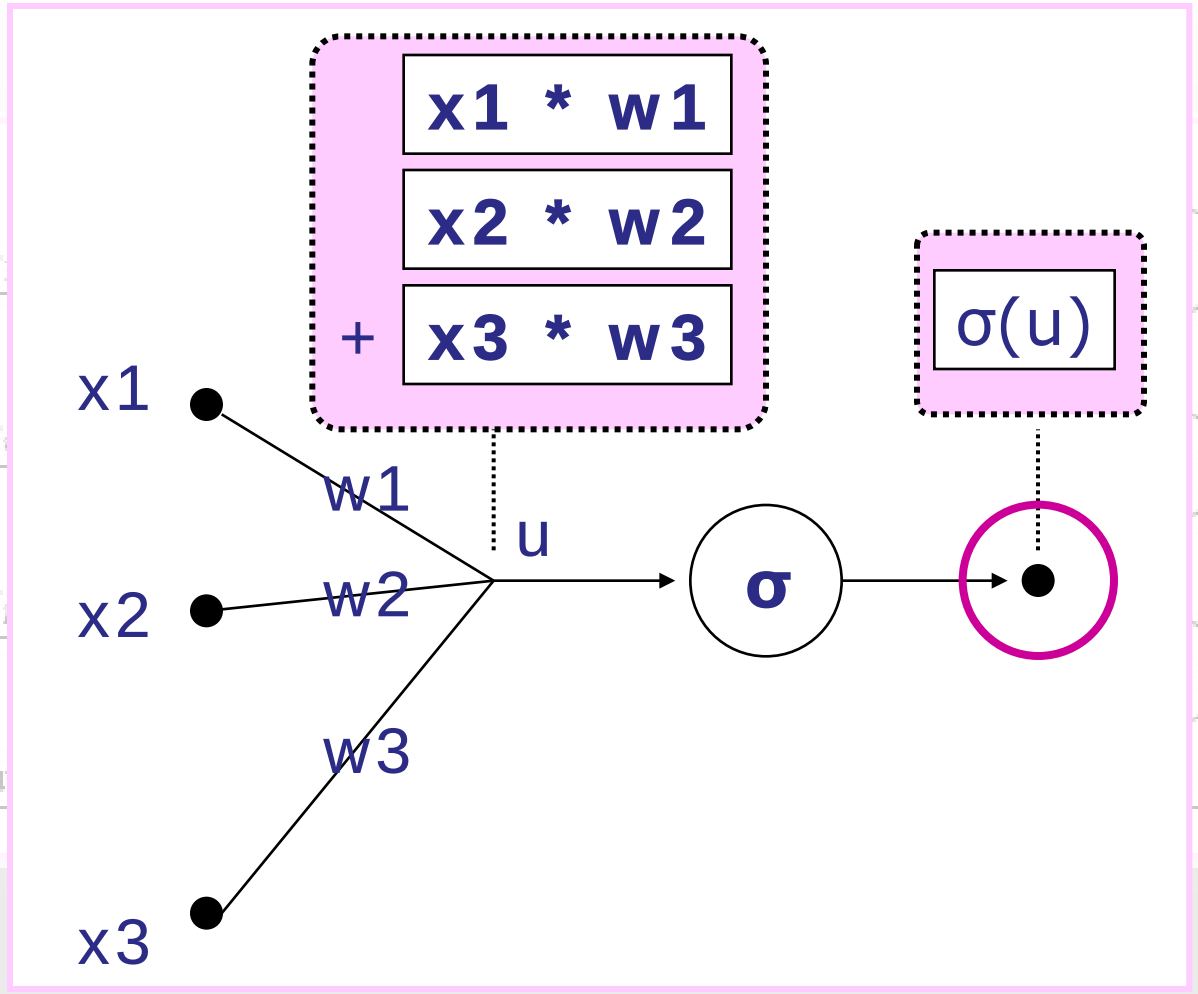
Neural networks

Biologische vs. artificiele **neurale netwerken**



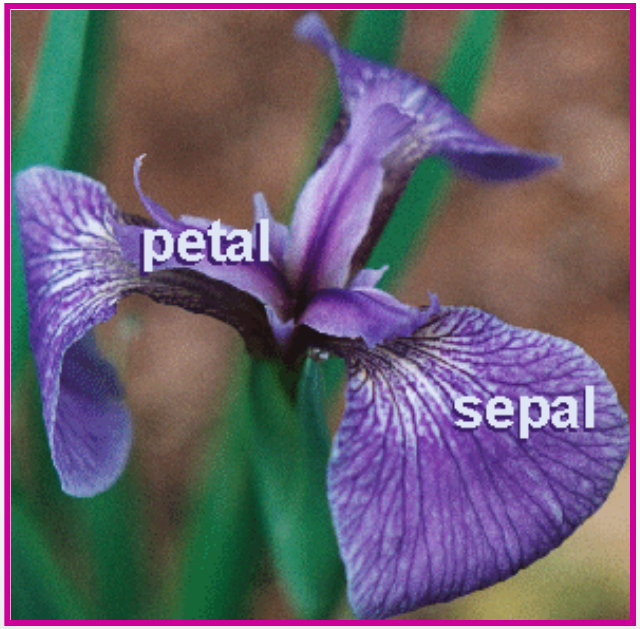

Input layer
INPUTS
OUTPUTS
Hidden layers
Output layer

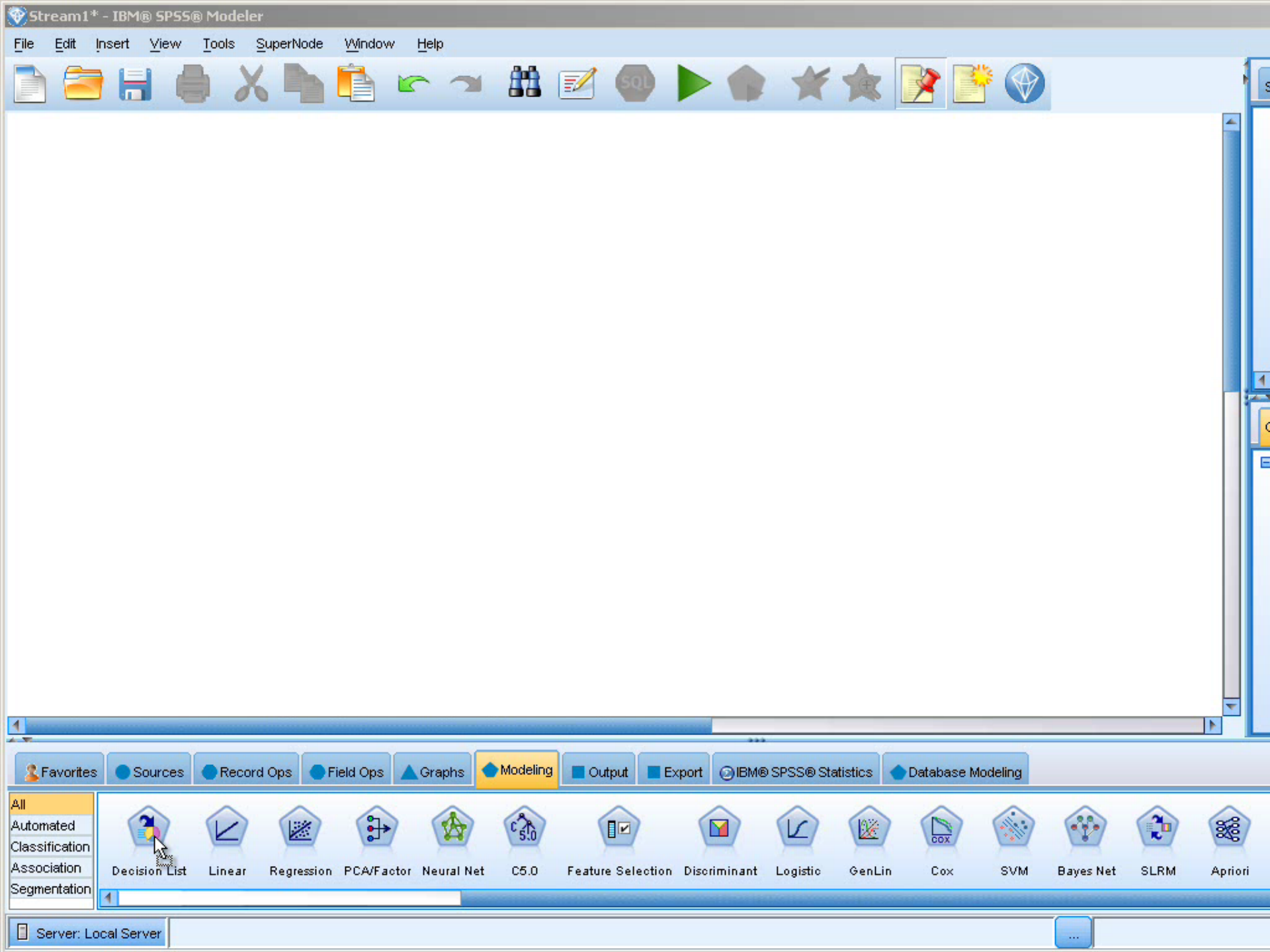
Multi-Layer Perceptron (MLP)

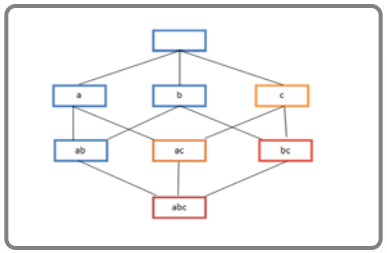


Neurale netwerken in **SPSS**

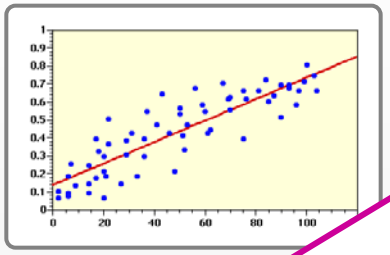
Sepal-length	Sepal-width	Petal-length	Petal-width	Class
4.6	3.6	1.0	0.2	Setosa
4.3	3.0	1.1	0.1	Setosa
5.1	2.5	3.0	1.1	Versicolor
5.7	2.6	3.5	1.0	Versicolor
5.8	2.8	5.1	2.4	Viriginica
6.3	2.8	5.1	2.4	Virginica
6.9	3.1	5.1	2.3	Virginica
...	...	...	...	...



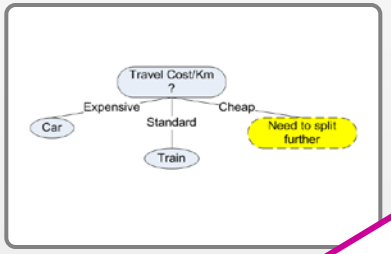

Stream1* - IBM® SPSS® Modeler
File
Edit
Insert
View
Tools
SuperNode
Window
Help
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Database Modeling
All
Automated
Classification
Association
Segmentation
Decision List
Linear
Regression
PCA/Factor
Neural Net
C5.0
Feature Selection
Discriminant
Logistic
GenLin
Cox
SVM
Bayes Net
SLRM
Apriori
Server: Local Server



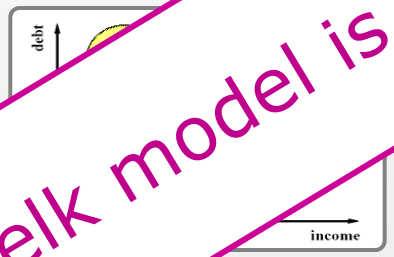
Association rules



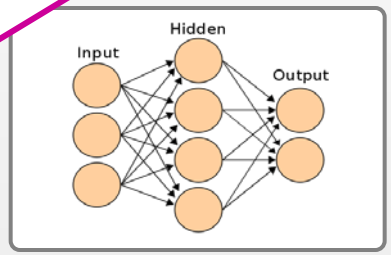
Statisti



Decision



lustering



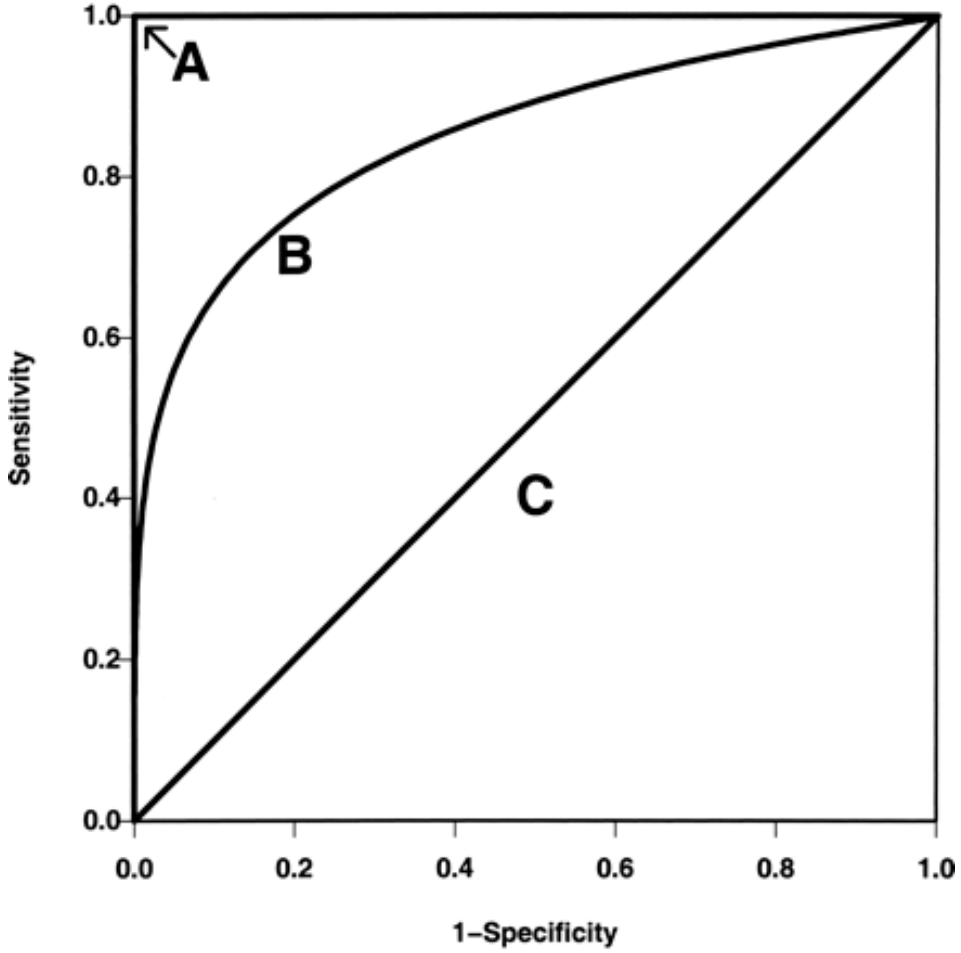
Neural networks

Welk model is het **beste**?

Makkelij erpretatie

Moeilijke interpretatie

Welk model is het beste: **ROC Curve**

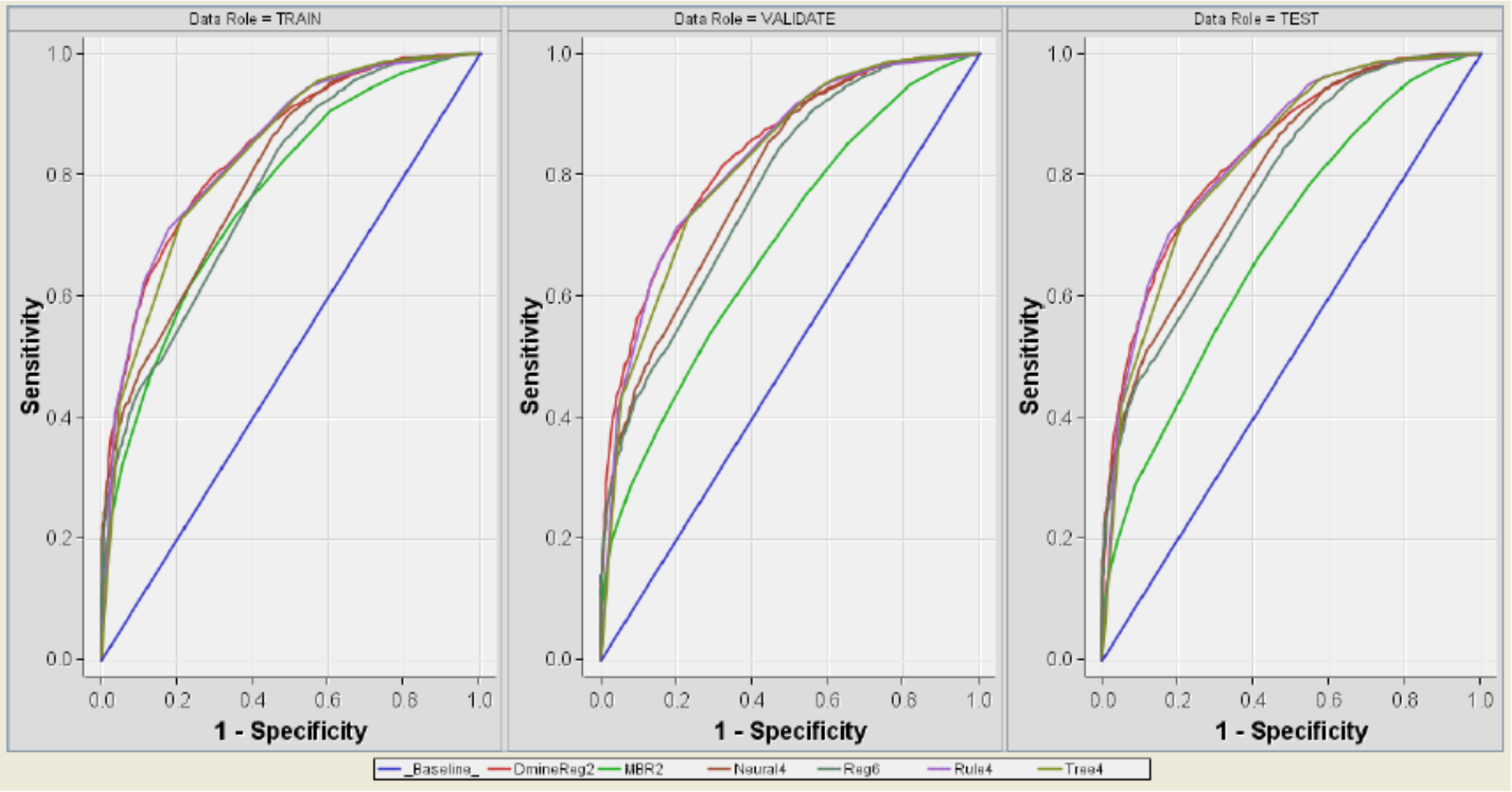


A > B > C

Sensitivity: true positive rate
1-Specifity: false negative rate

	Predicted TRUE	Predicted FALSE
Actually TRUE	% True Positive (TP)	% False Negative (FN)
Actually FALSE	% False Positive (FP)	% True Negative (TN)

Welk model is het beste: **ROC Curve**



In deze **infosessie…**

Deel 1	Wat is **predictive analytics**? Courante **predictieve modellen** **Toepassingen** van deze modellen
Deel 2	Predictive analytics **tools** Predictive analytics **cases** Wat hebben we **geleerd**?

In deze **infosessie…**

Deel 1	Wat is **predictive analytics**? Courante **predictieve modellen** **Toepassingen** van deze modellen
Deel 2	Predictive analytics **tools** Predictive analytics **cases** Wat hebben we **geleerd**?

Tools per analyse methode

Scripting	Form-based	Visual programming
S+	WEKA	SPSS Modeler
R		SAS Enterprise Miner
Matlab		Oracle Data Miner
Octave		TIBCO Spotfire Miner

Open source | Commercial

Tools per analyse methode

Scripting	Form-based	Visual programming
S+	WEKA	SPSS Modeler
R		SAS Enterprise Miner
Matlab		Oracle Data Miner
Octave		TIBCO Spotfire Miner

Open source | Commercial

S+	R
Implementaties van de statistische programmeertaal "S"	
• Commercieel • Command shell • Biedt ook forms aan die de shell afschermen	• Open source • Command shell

Tree fitten in **R vs. S+**

RGui - [Untitled - R Editor]

File Edit Packages Windows Help

```
fit <- rpart(class~sepallength+sepalwi
, data=iris,method="class")
plot(fit, uniform=TRUE, main="Iris tre
text(fit, use.n=TRUE, all=TRUE, cex=.8
```

R

Tree Models

Model | Results | Plot | Prune/Shrink | Predict

Data
Data Set: iris
Weights:
Subset Rows with:
Omit Rows with Missing Values

Fitting Options
Min. No. of Obs. Before Split: 5
Min. Node Size: 10
Min. Node Deviance: 0.010

Save Model Object
Save As:

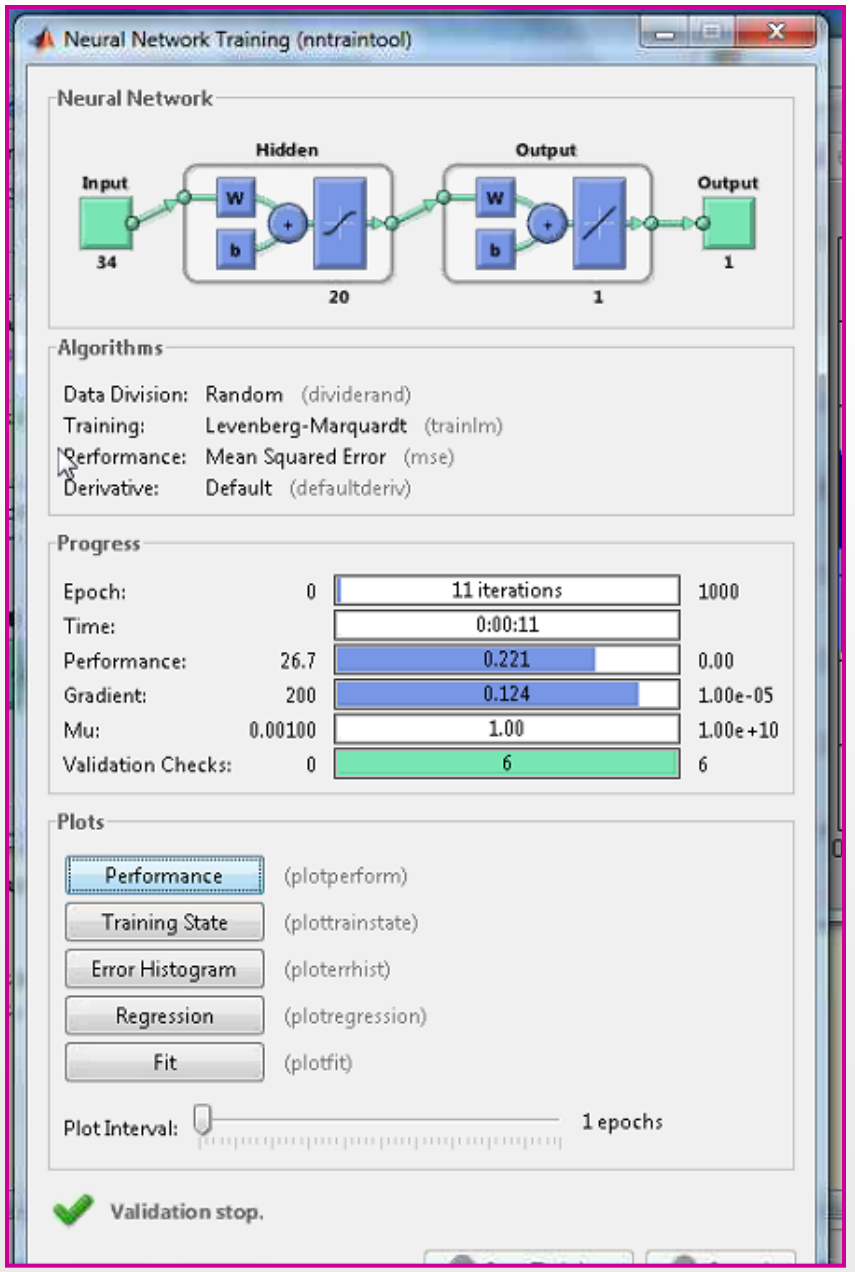
Variables
Dependent: class
Independent: <ALL> sepallength sepalwidth petallength petalwidth
Formula: class~petalwidth+sepallength+sepalwidth+petallength
Create Formula

OK Cancel Apply current Help

S+

Matlab

- Mathworks
- Wiskundige omgeving
- Sterk in **data visualisatie**!
- Functionaliteit verpakt in "toolboxes"
- Toolboxes voor **neurale netwerken**, databases, statistics, ...
- Command line en forms
- **Octave** = OS alternatief



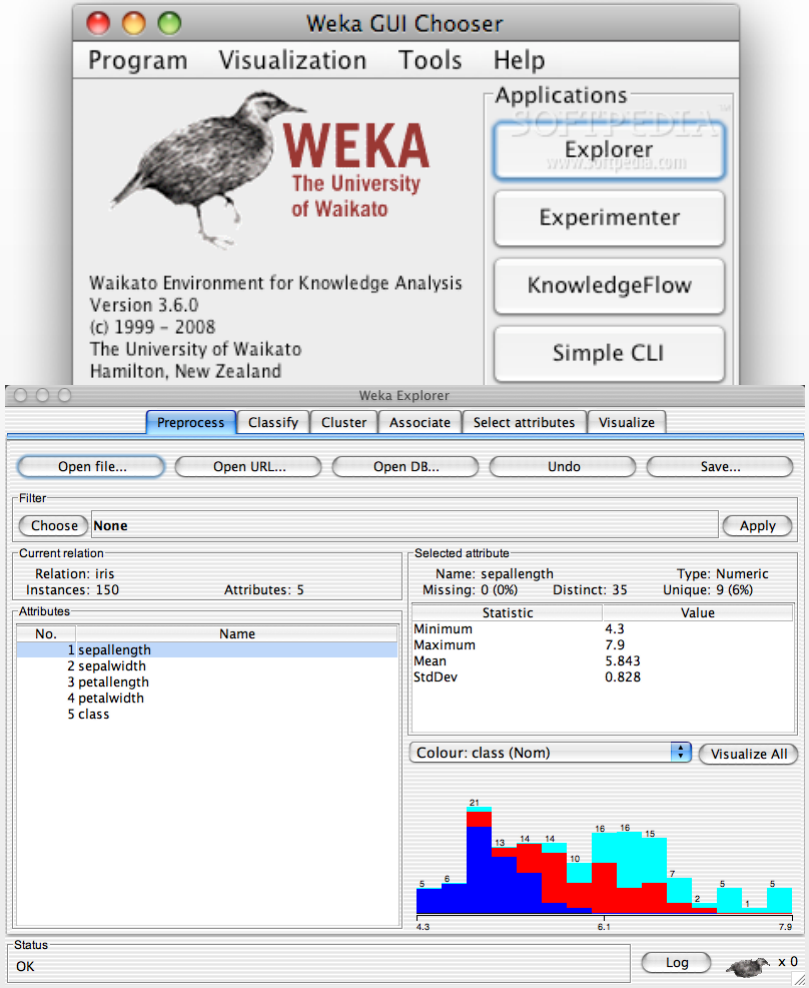
Tools per analyse methode

Scripting	Form-based	Visual programming
S+	WEKA	SPSS Modeler
R		SAS Enterprise Miner
Matlab		Oracle Data Miner
Octave		TIBCO Spotfire Miner

Open source | Commercial

Weka

- **W**aikato **E**nvironment for **K**nowledge **A**nalysis
- Open source tool van de Universiteit van Waikato
- Java
- Zeer veel algoritmes ter beschikking



Tools per analyse methode

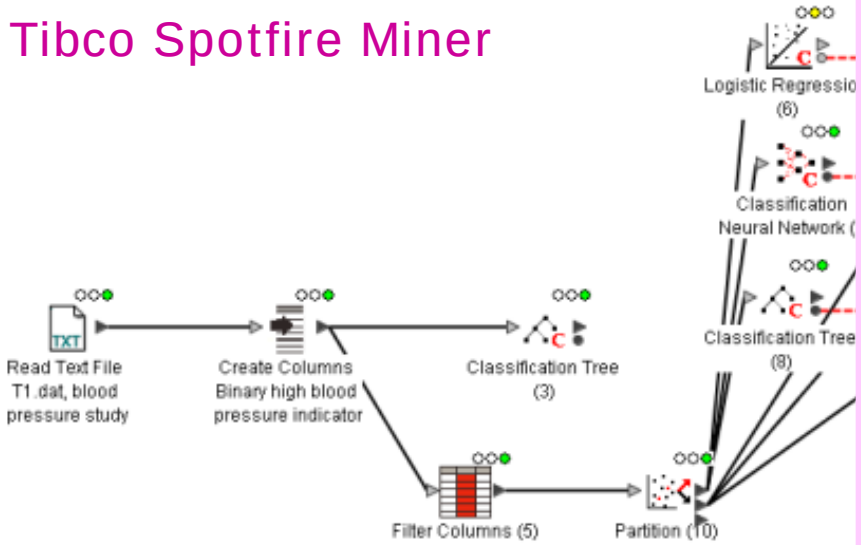
Scripting	Form-based	Visual programming
S+	WEKA	SPSS Modeler
R		SAS Enterprise Miner
Matlab		Oracle Data Miner
Octave		TIBCO Spotfire Miner

Open source | Commercial

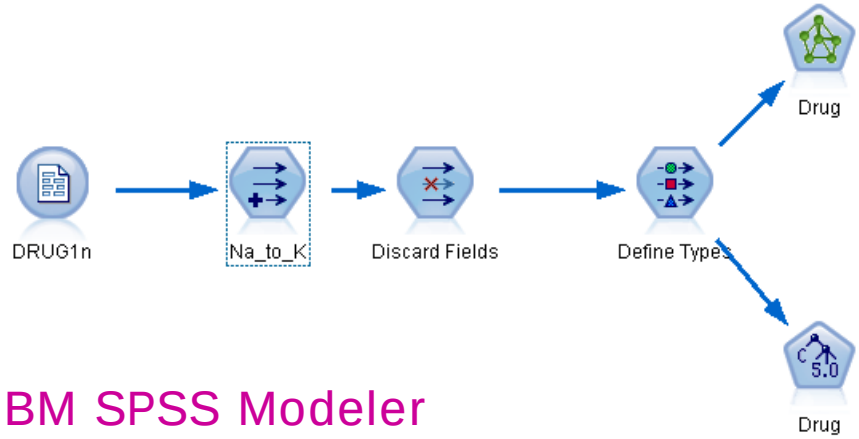
Smals
ICT for society

Visual programming

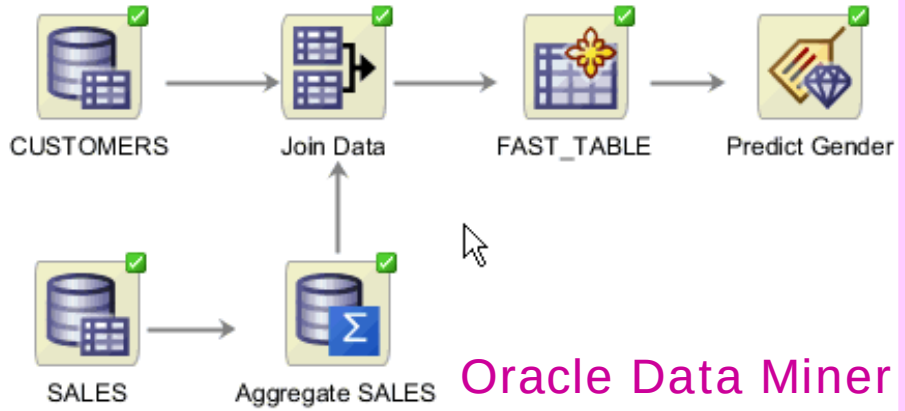
Tibco Spotfire Miner



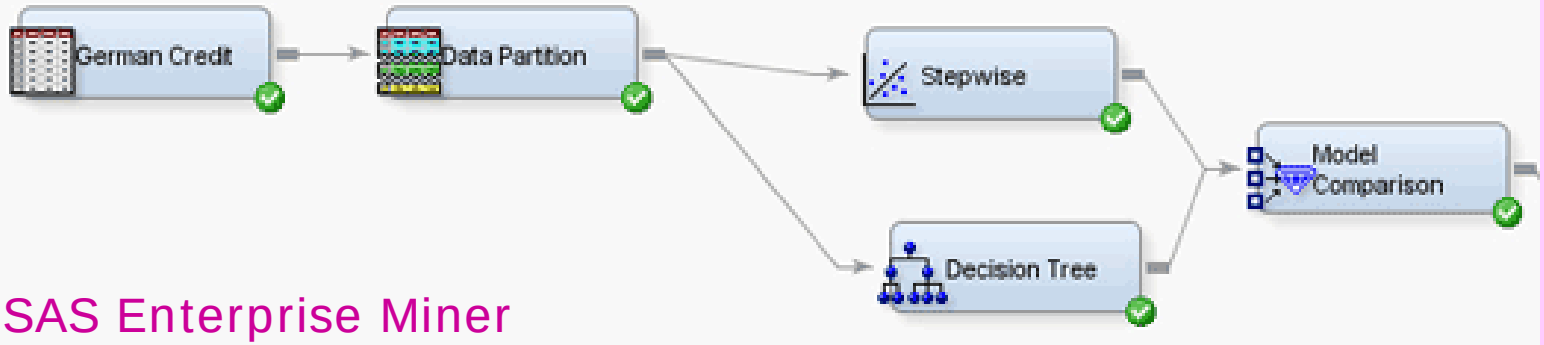
IBM SPSS Modeler



Oracle Data Miner

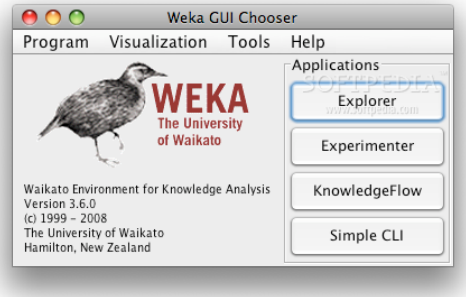


SAS Enterprise Miner

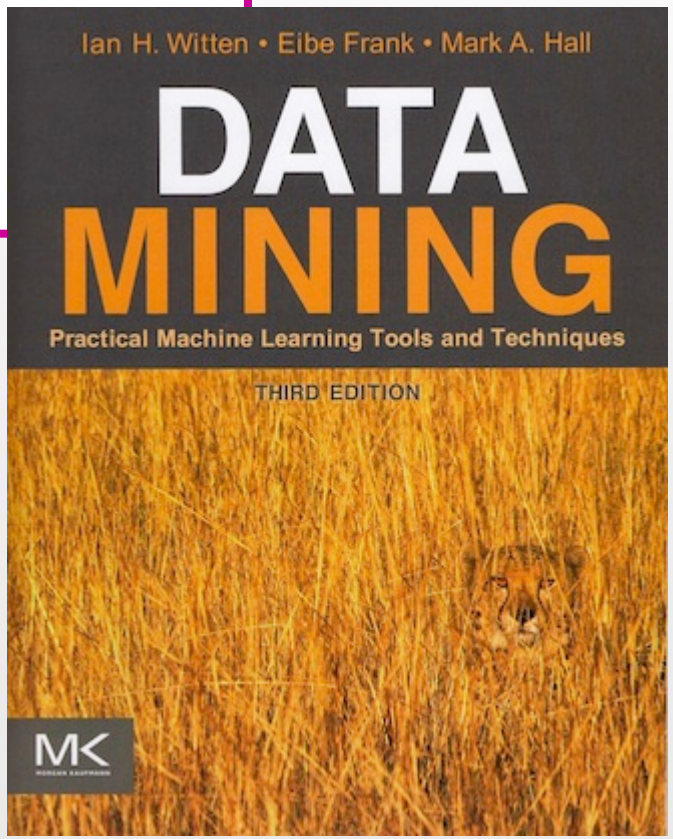


Gevolgd traject met deze tools

Eerste stappen met predictive analytics



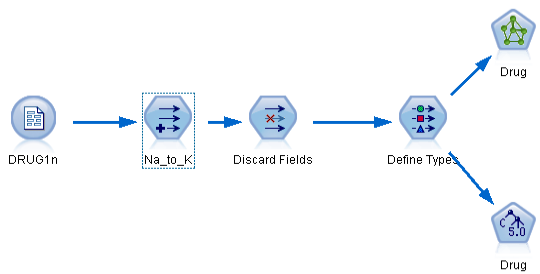
- Algoritmes leren kennen
- Parameters tunen
- Kleine datasets
- Lage instap (free)



Gevolgd traject met deze tools

Eerste stappen met predictive analytics

Eerste projecten met **meer data**



Visual programming

- Methodes om data te filteren, cleanen, combineren,…
- Lage leercurve om algoritmes en data te koppelen
- Kunnen veel data formaten aan

Gevolgd traject met deze tools

Eerste stappen met predictive analytics

Eerste projecten met **meer data**

Geavanceerd gebruik

```
fit <- rpart(class~sepallength+sepalwidth+petallength+petalwidth
, data=iris,method="class")
plot(fit, uniform=TRUE, main="Iris tree")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

Scripting

- Scripts zijn makkelijker uitbreidbaar ivgl met visual programming
- De nodige extra functionaliteit kan geprogrammeerd worden
- Blijft een persoonlijke keuze: programmeren vs. modelleren

Predictive analytics initiatieven SMALS

CASE 1: Strijd tegen de **sociale fraude**

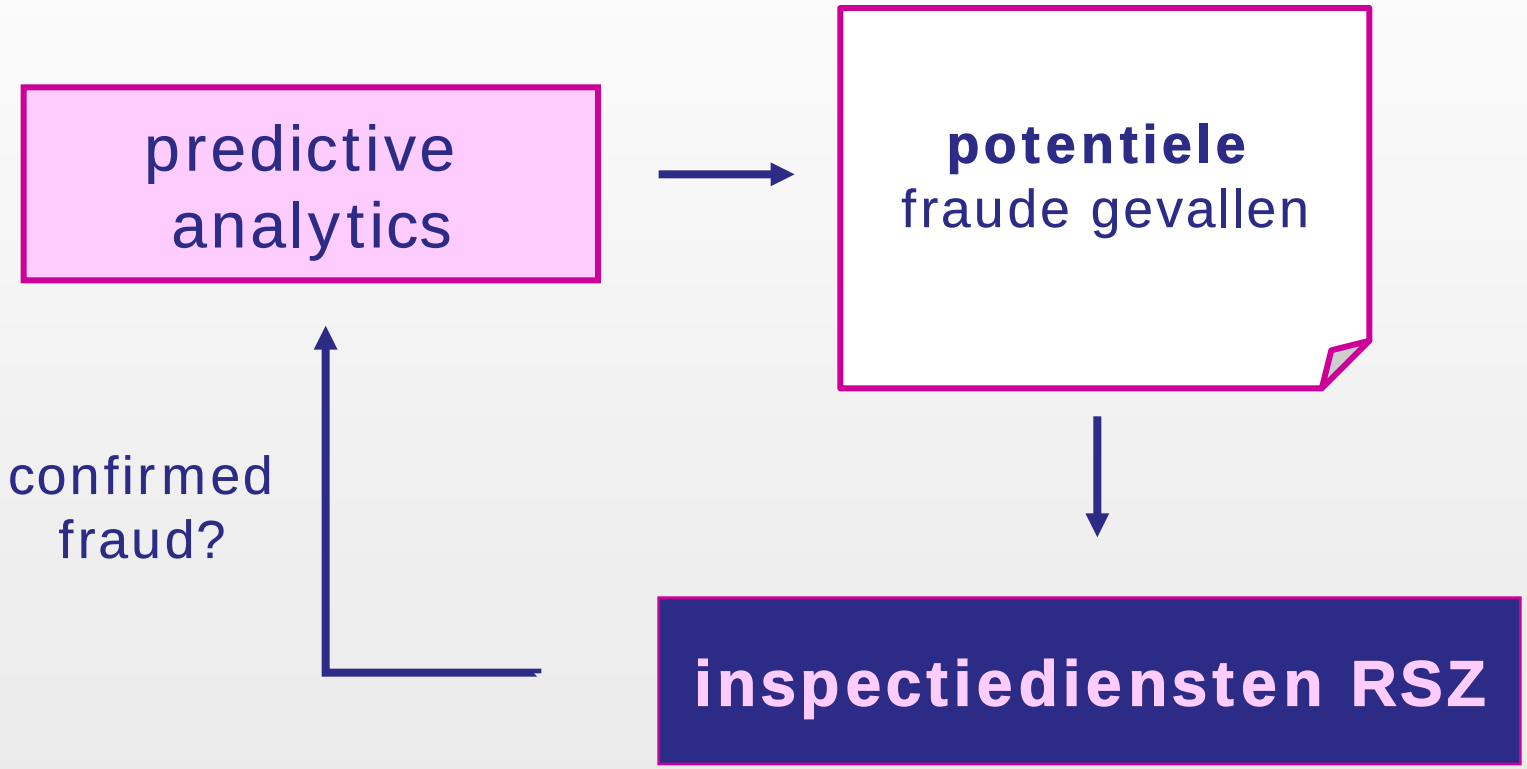
CASE 2: Traceerbaarheid

Predictive analytics initiatieven SMALS

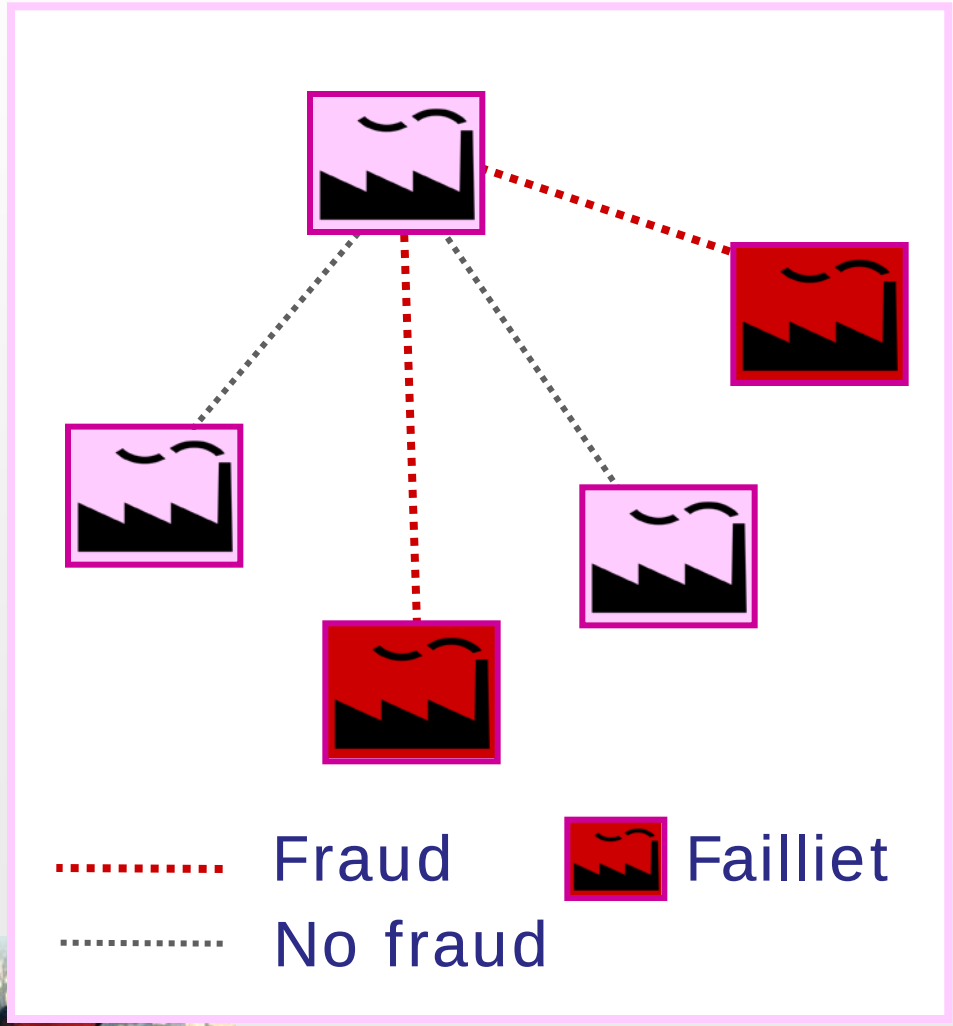
CASE 1: Strijd tegen de **sociale fraude**

CASE 2: Traceerbaarheid

Strijd tegen de **sociale fraude:** verhogen van de **efficientie** van de inspecties!



Predictive modeling bij sociale fraude



Doel:
frauduleuze associaties tussen bedrijven herkennen.

Oplossing:
Association rules met het apriori algoritme

Association rules?

{Butter, Bread} => {Milk}

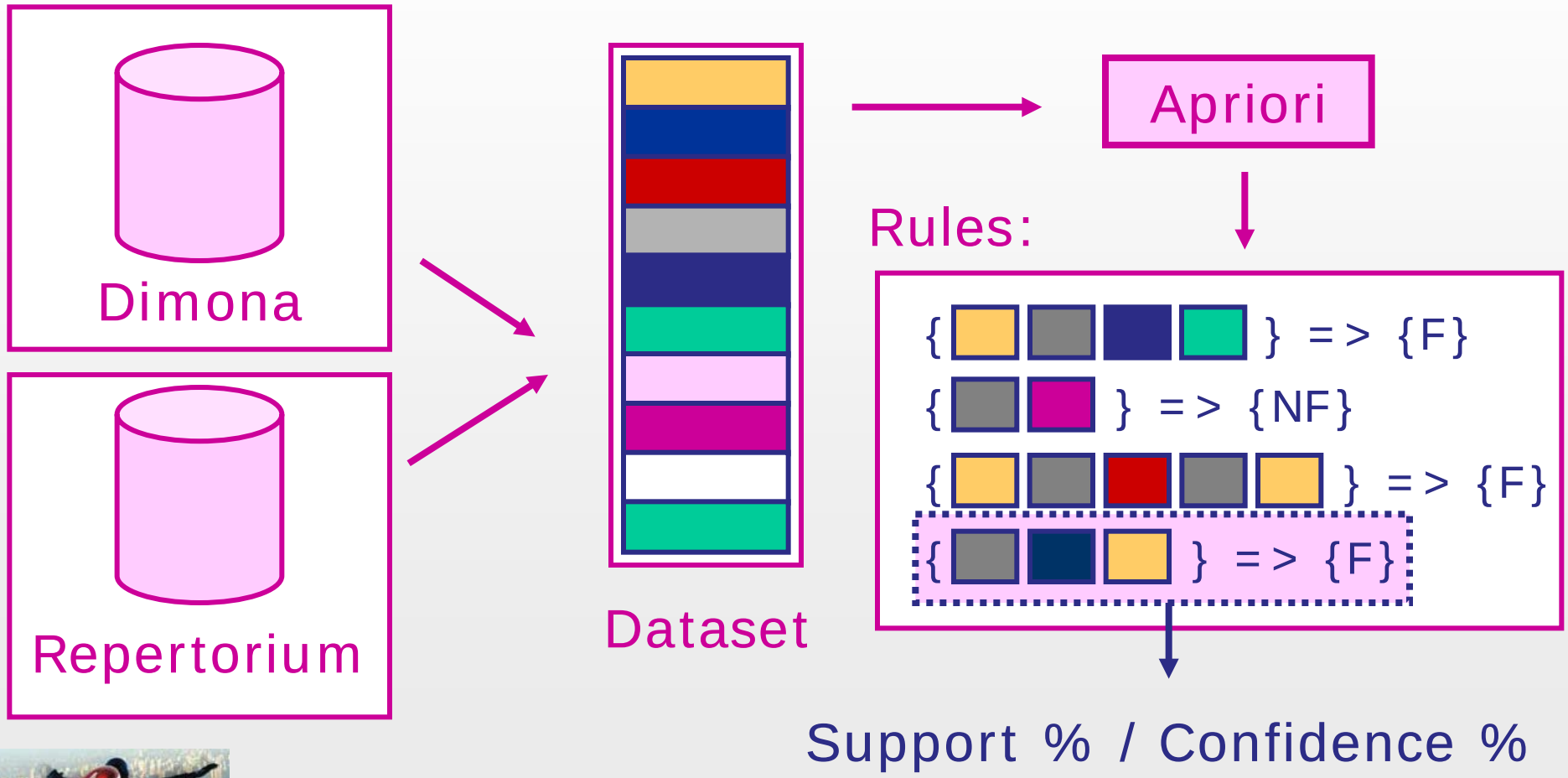
Antecedent

Consequent

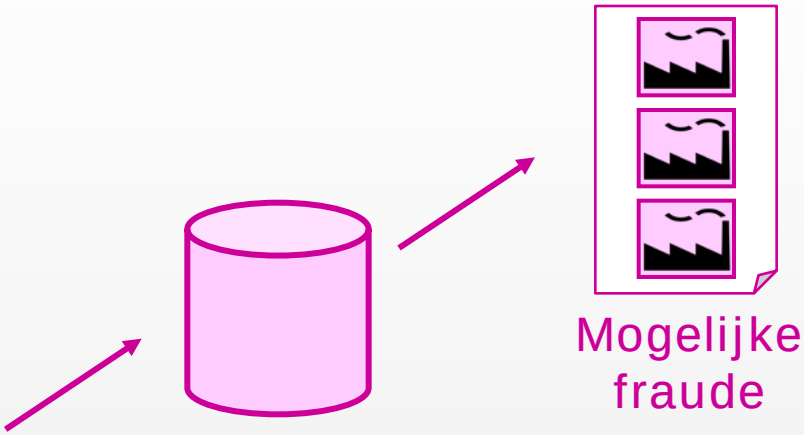
Support:
Hoeveel keer (in %) komt het antecedent voor?

Confidence:
Hoeveel keer (in %) volgt de consequent op de antecedent ?

Association rules met Apriori



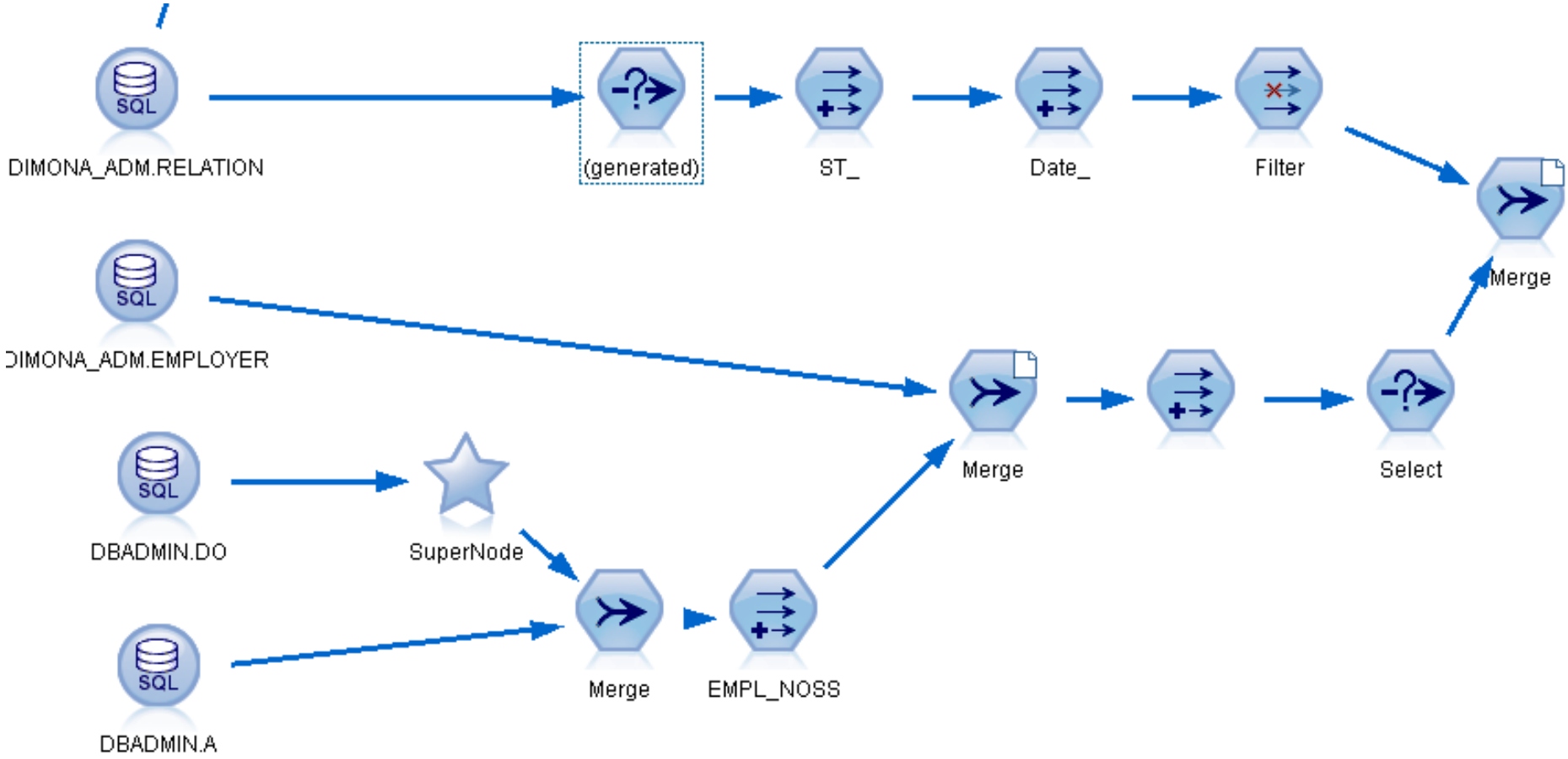
Fraude opsporen met deze association rules



	Confidence
{ } => {F}	100%
{ } => {F}	99,8%
{ } => {F}	95%

Association rules geordend volgens confidence

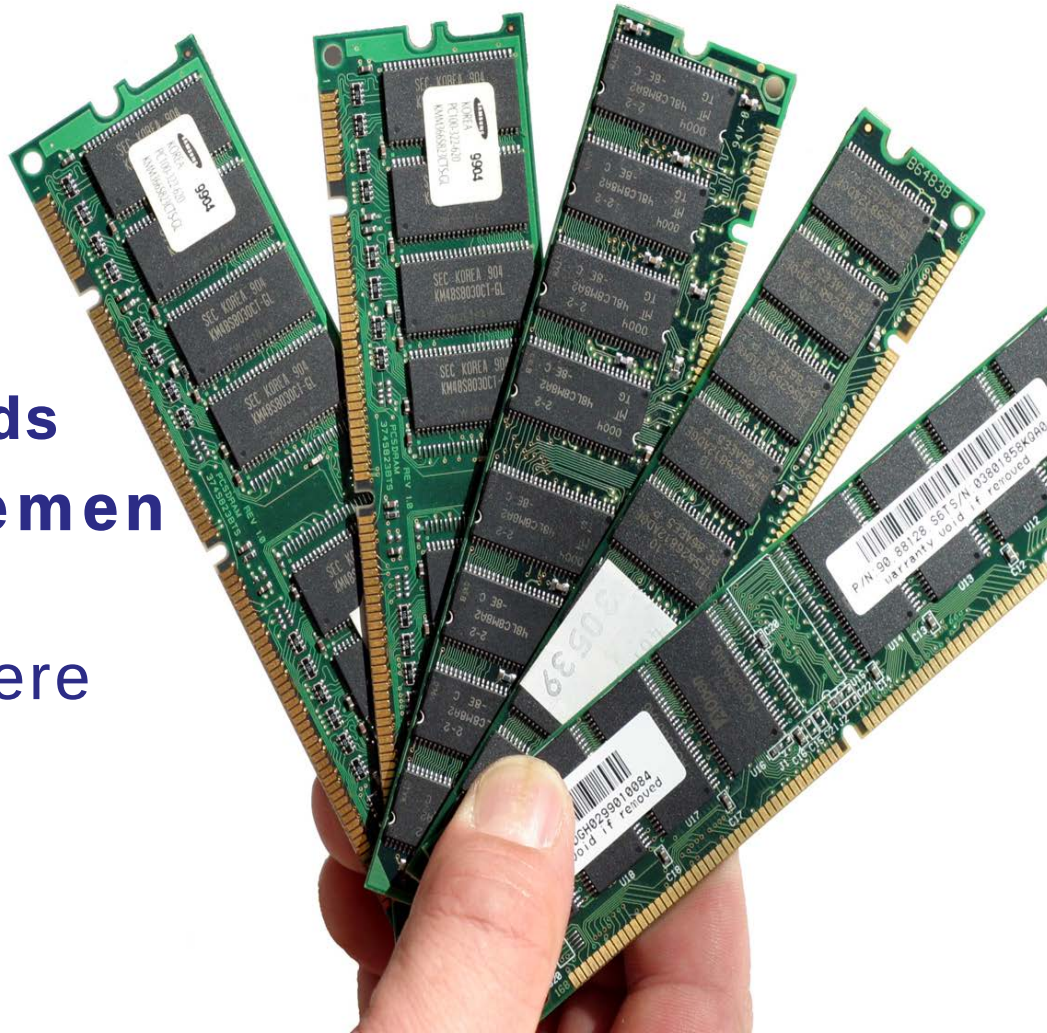
Geimplementeerd in **SPSS**



Documentatie noodzakelijk!

Status case "sociale fraudebestrijding"

- Eerste resultaten veel belovend!
- Fraude = zoeken naar uitzonderingen
  - Alle data is nodig
  - **> 50 miljoen records**
- **Performantie problemen** op "gewone Pc"
- Migreren naar zwaardere servers



Predictive analytics initiatieven SMALS

CASE 1: Strijd tegen de **sociale fraude**

CASE 2: Traceerbaarheid

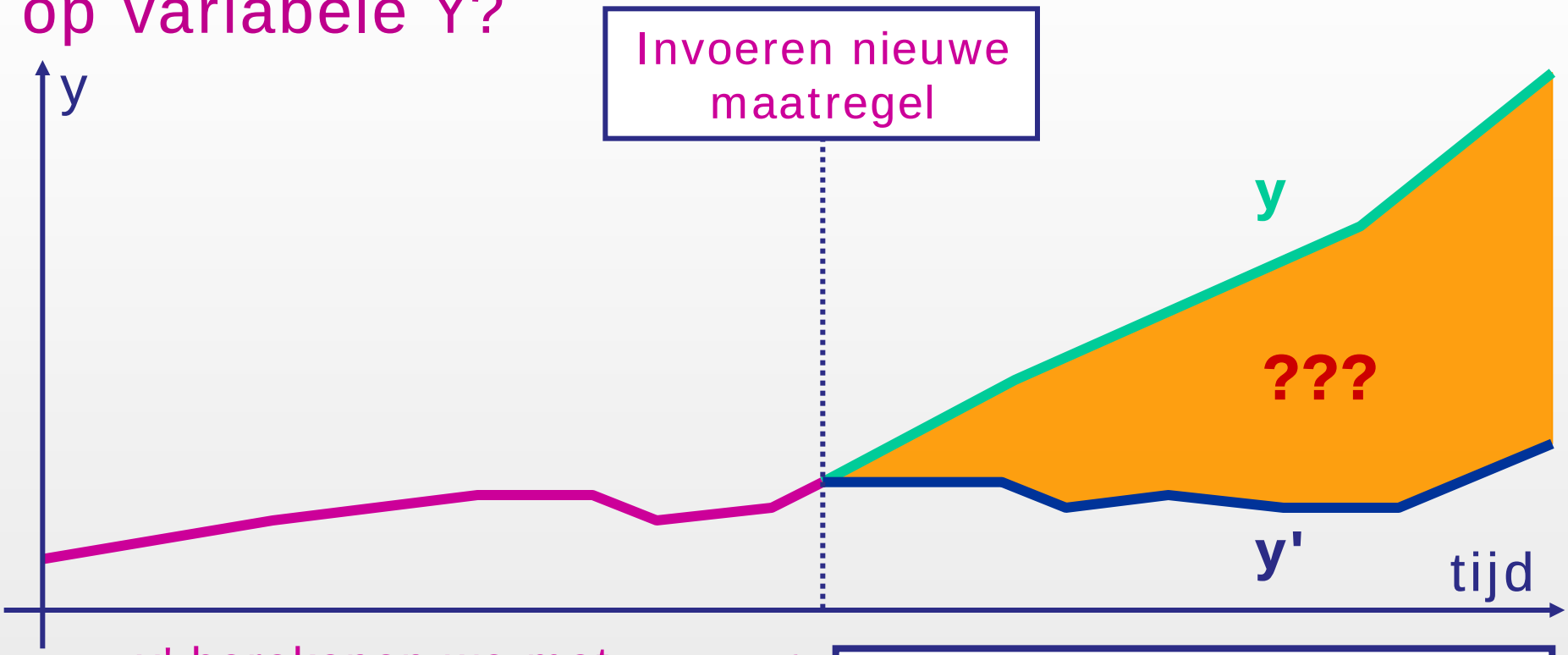
Traceerbaarheid

Wat is de **impact** van een **beleidsmaatregel** op variabele Y?

	Resultaat **na invoering maatregel**	**(y)**
—	Resultaat **als maatregel niet ingevoerd was**	**(y')**
	Impact van een maatregel	**(y - y')**

Traceerbaarheid

Wat is de **impact** van een **beleidsmaatregel** op variabele Y?



y' berekenen we met **Predictive Analytics**

- Zonder invoering maatregel
- Na invoering maatregel
- Impact maatregel

Traceerbaarheid

Wat is de **impact** van ee **el** op variabele Y?

X_1	X_2	X_3	X_4	Y_{old}
a	a	c	d	1
d	e	e	d	2

X_1	X_2	X_3	X_4	Y
a	a	c	e	4
a	a	c	e	5
...	...	...	...	...

Chronologische dataset

Data **voor** maatregel

Invoering nieuwe maatregel !

Data **na** maatregel

Berekenen Y' **zonder invoering maatregel**

1. Bereken predictief model bij data **voor** maatregel

X_1	X_2	X_3	X_4
a	a	c	d
d	e	e	d
...	...	...	...

Predictief model

Y_{old}
1
2
...

2. Pas model toe **op data na de maatregel** om Y' te bekomen

X_1	X_2	X_3	X_4
a	a	c	e
a	a	c	e
...	...	...	...

Predictief model

Y'
1
2
...

Berekenen **impact** maatregel

Resultaat **na invoering maatregel**	**(y)**
Resultaat **als maatregel niet ingevoerd was**	**(y')**
Impact van een maatregel	**(y - y')**

Gevonden dmv **"Predictive Analytics"!**

Wat hebben we **geleerd** van deze predictive analytics cases?

Voorbereiden van data neemt **veel tijd** in beslag

Predictive analytics **project aanpak**

4. Feedback

3. Interpretatie en evaluatie

2. Predictive modeling

1. Voorbereiden data

Data

D1	D2	D3	D4	D5

Dataset

Predictief model

Kennis over de toekomst

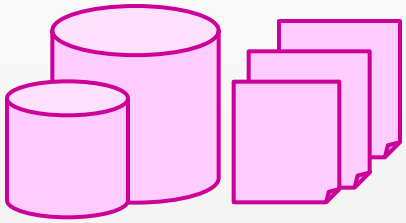
Data voorbereiding is een **interdisciplinair** proces

- Verschillende **rollen** nemen deel
- Business expert(s)
- Data quality expert(s)
- Database expert(s)
- Predictive analytics expert(s)
- Data warehouse expert(s)

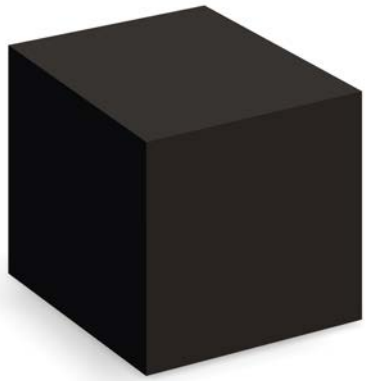
Wat hebben we **geleerd** van deze predictive analytics cases?

Predictive analytics is **geen "black box"**

Predictive Analytics als een **"black box"**



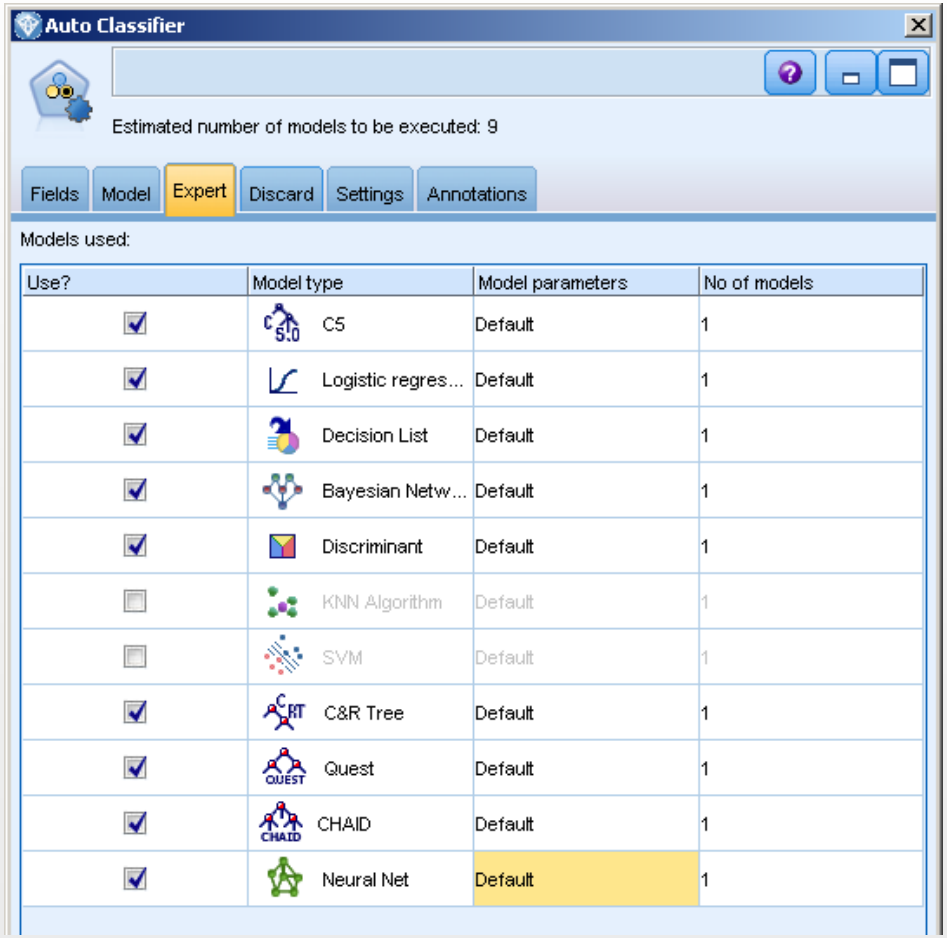
Data



Predictive Analytics tool

Beste predictief model

Mogelijk in tools als SPSS en SAS



- Kan zeer lang duren (meerdere dagen)
- Trage feedback loop
- **Betere optie:**
  - Predictive analytics kennis intern opbouwen
  - Zelf meeredeneren

Wat hebben we **geleerd** van deze predictive analytics cases?

Bouw predictive analytics projecten gradueel uit

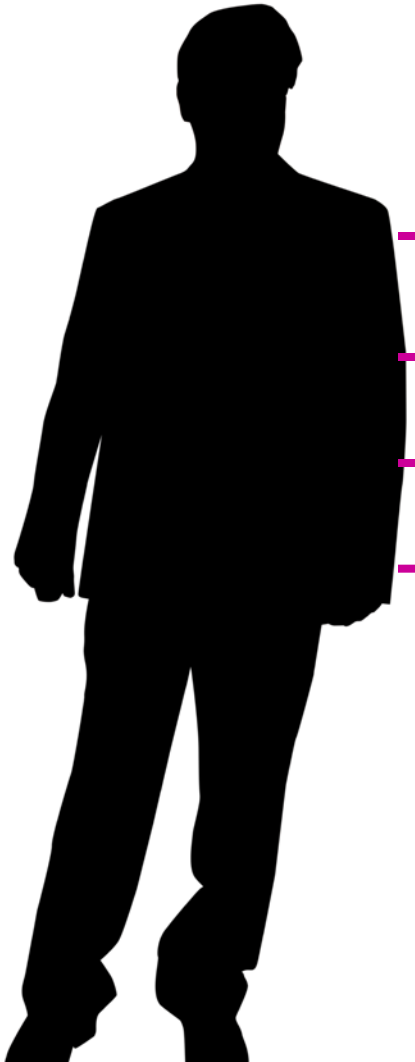
Predictive analytics is **geen magische oplossing**

- Start met **kleinschalige** predictive analytics experimenten
- Bouw groter project uit
- Gradueel interesse wekken van potentiele stakeholders

Wat hebben we **geleerd** van deze predictive analytics cases?

My name is …
Predictive Analyst

Specifieke **rol** nodig voor predictive analytics!

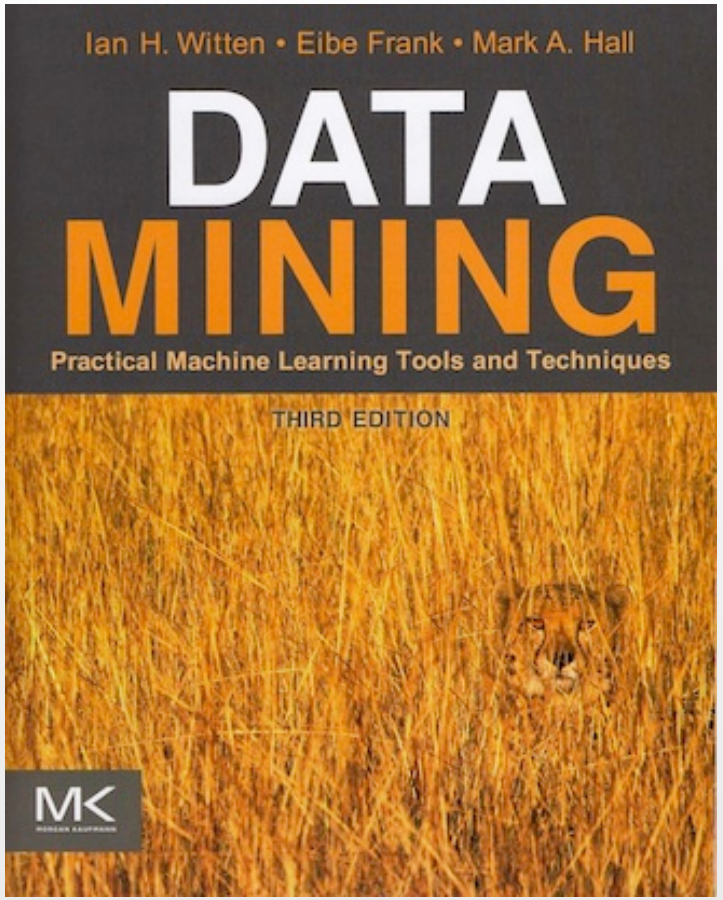


- **Analytics kennis**
- **Business kennis**
- **IT kennis**
- **Communicatie!!**

Aanbevolen literatuur

Predictive Analytics White Paper

Charles Nyce, PhD, CPCU, API
Senior Director of Knowledge Resources
American Institute for CPCU/
Insurance Institute of America



Aanbevolen literatuur

http://chem-eng.utoronto.ca/~datamining/dmc/data_mining_map.htm

Interactief overzicht van meest courante datamining technieken!

Vragen?

Deel 1	Predictive analytics **aanpak** Courante **predictieve modellen** **Toepassingen** van deze modellen
Deel 2	Predictive analytics **tools** Predictive analytics **cases** Wat hebben we **geleerd**?