

Een speld in een hooiberg:  
zoeken naar de genetische eigenschappen  
die een ziekte veroorzaken

**CWI**

Centrum Wiskunde & Informatica



**Utrecht University**



A

C

G

T



zus?

ik



zus?





G Disease

# De analyse

## DNA profiel



ACGCCGATATTTCCAGAGTCCCGATTTCGAAGGTA



ACCCCGGTACTTCCAGAATCCCGATACGAAGGTA



ACGCCGGTACTTCCCGAGTCCGGATTTCGAGGGTA



ACGCCGGTATTACCCGAGTCCGGATTTCGAAGGTA



ACGCCGATATTTCCAGAATCCCGATACGAGGGTA



ACCCCGATATTACCCGAGTCCCGATACGAGGGTA

# De analyse

## DNA profiel



ACGCCGATATTCCAGAGTCCCGATTCGAAGGTA



ACCCCGGTACTTCCAGAATCCCGATACGAAGGTA



ACGCCGGTACTTCCCGAGTCCGGATTCGAGGGTA



ACGCCGGTATTACCCGAGTCCGGATTCGAAGGTA



ACGCCGATATTCCAGAATCCCGATACGAGGGTA



ACCCCGATATTACCCGAGTCCCGATACGAGGGTA

# De analyse

## DNA profiel



GATTAGCTA



CGCTAACAA



GGCTCGGTG



GGTACGGTA

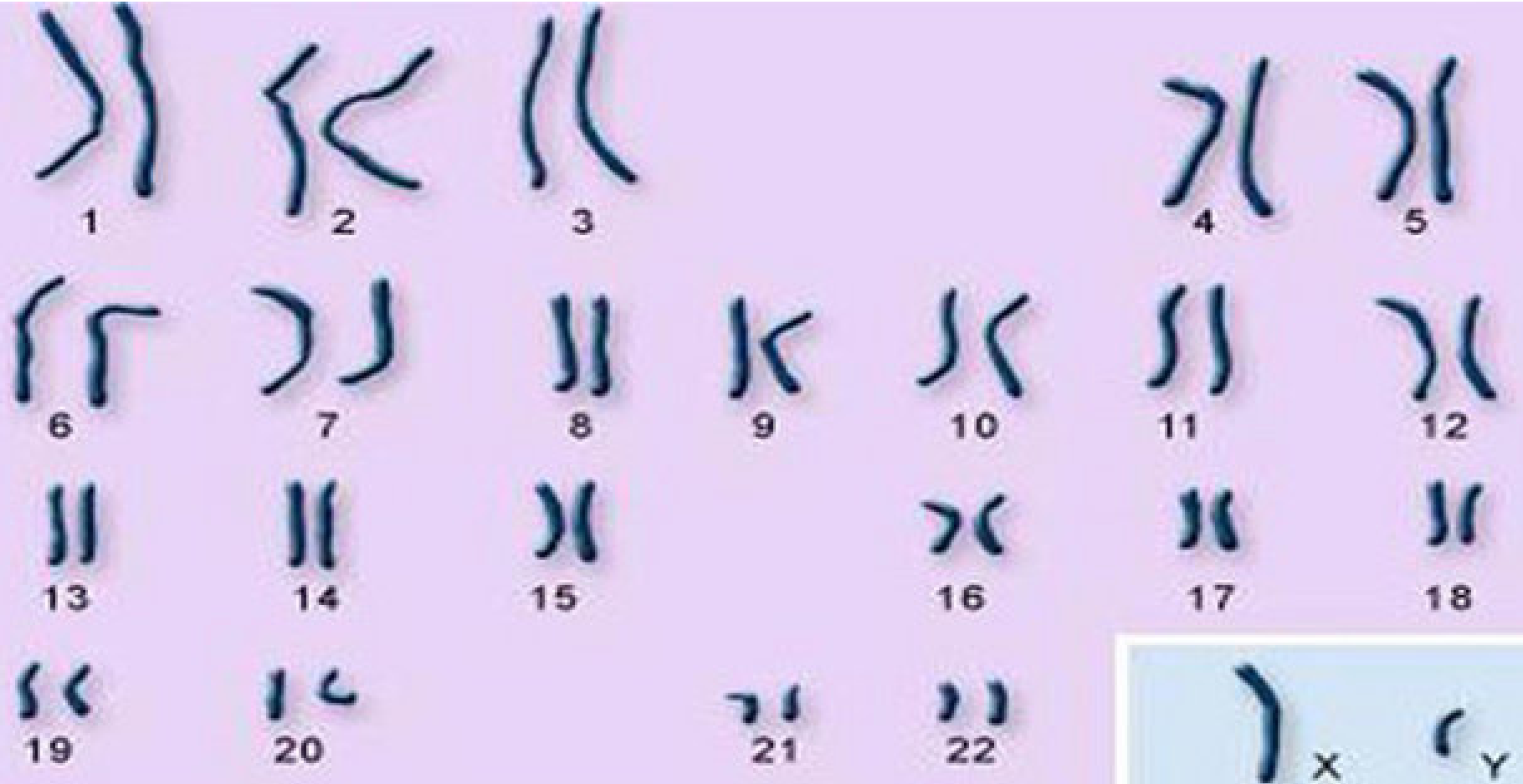


GATTAACAG



CATACGCAG

# Dit was maar de helft...



# Volledig DNA profiel

DNA profiel



GATTAGCTA  
CATTAACAA

# Volledig DNA profiel

DNA profiel



GATTAGCTA



CGCTAACAA



GGCTCGGTG



GGTACGGTA



GATTAACAG



CATACGCAG

G-C A-G T-C T-A A-C G-A C-G T-A A-G

# Volledig DNA profiel

DNA profiel



GATTAGCTA  
CATTAACAA

G-C   A-G   T-C   T-A   A-C   G-A   C-G   T-A   A-G



1/0   1/1   0/0   0/0   1/1   1/0   1/1   1/0   0/0

# Volledig DNA profiel

OS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	dark_13	dark_23	dark_21	yellow_5	yellow_6	dark_1	dark_7
462	chr5:462	T	G	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0	0/1
638	chr5:638	T	G	.	.	PR	GT	1/1	0/1	1/1	0/0	0/0	0/1	0/0
662	chr5:662	C	T	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/1	0/0
911	chr5:911	C	T	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0	0/0
923	chr5:923	A	G	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0	0/0
942	chr5:942	G	A	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0	0/0
948	chr5:948	A	T	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0	0/0
961	chr5:961	T	C	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0	0/0
1378	chr5:1378	C	T	.	.	PR	GT	0/0	./.	0/0	0/0	0/0	0/0	0/1
1941	chr5:1941	T	A	.	.	PR	GT	0/1	0/0	0/0	0/0	0/0	0/0	0/0
1977	chr5:1977	G	A	.	.	PR	GT	0/1	0/1	0/0	0/1	0/0	0/1	0/1
2001	chr5:2001	T	G	.	.	PR	GT	0/1	0/1	0/0	0/1	0/0	0/1	0/1
2003	chr5:2003	G	A	.	.	PR	GT	0/1	0/1	0/0	0/1	0/0	0/1	0/1




# Volledig DNA profiel

DNA profiel



GATTAGCTA  
CATTAACAA

G-C   A-G   T-C   T-A   A-C   G-A   C-G   T-A   A-G







 1/0   1/1   0/0   0/0   1/1   1/0   1/1   1/0   0/0



 1   2   0   0   2   1   2   1   0

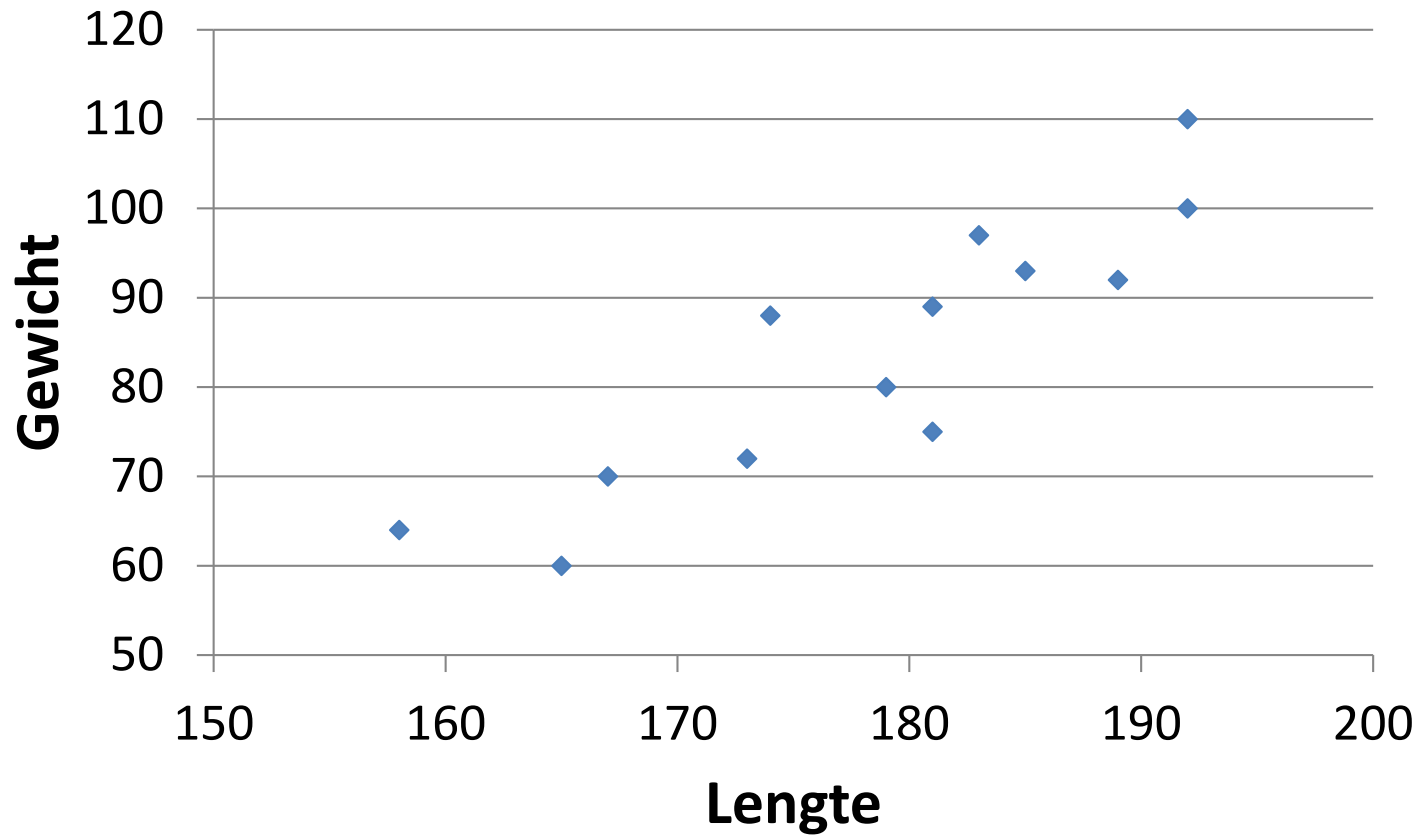
# Volledig DNA profiel

## DNA profiel

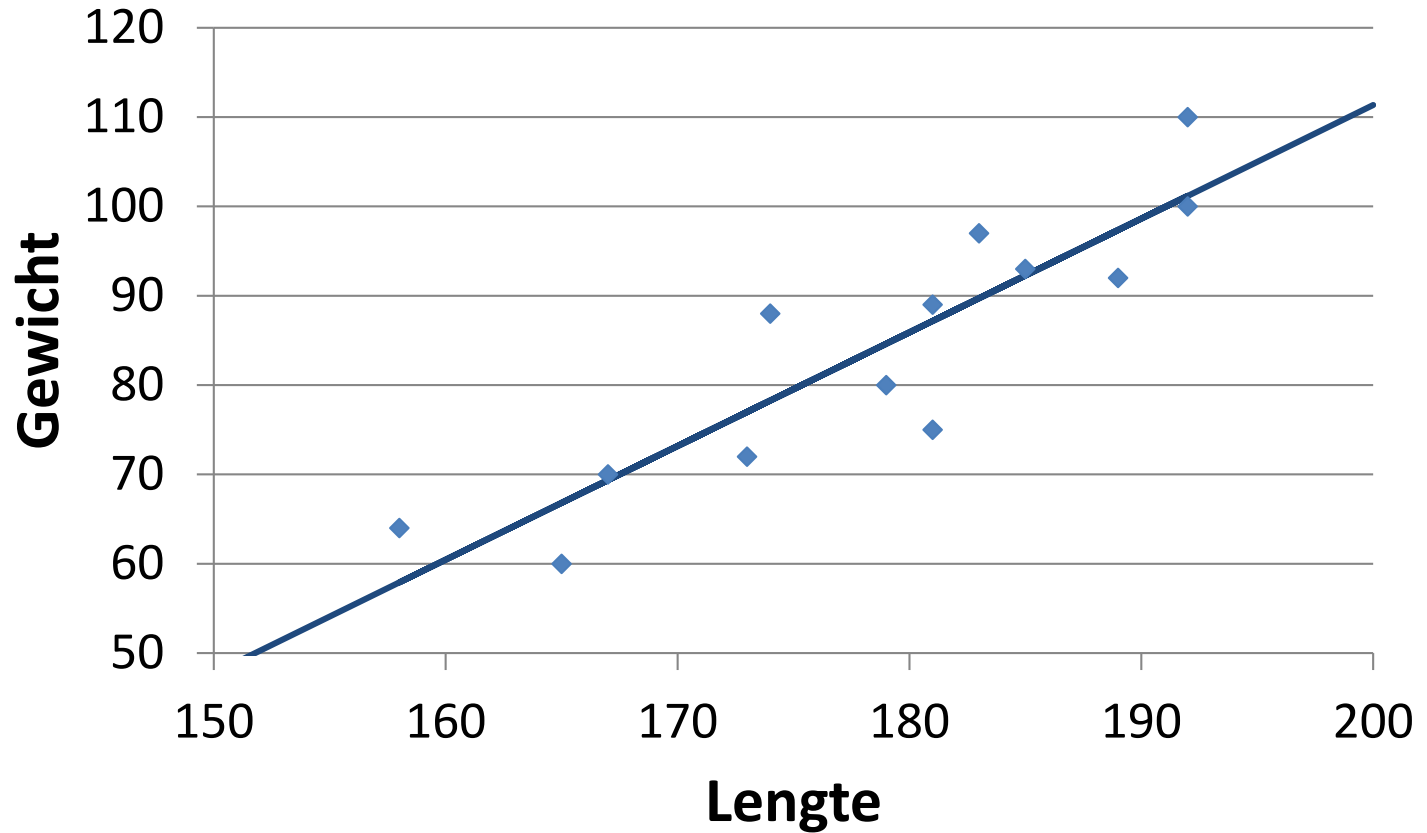
	1	2	0	0	2	1	2	1	0
	0	2	1	0	1	0	0	0	0
	1	1	0	0	2	1	2	1	0
	1	0	0	0	1	0	0	1	0
	2	0	0	1	0	1	0	0	1
	1	1	0	1	2	2	2	0	2

# Intermezzo: Logistische regressie

# Lineaire regressie

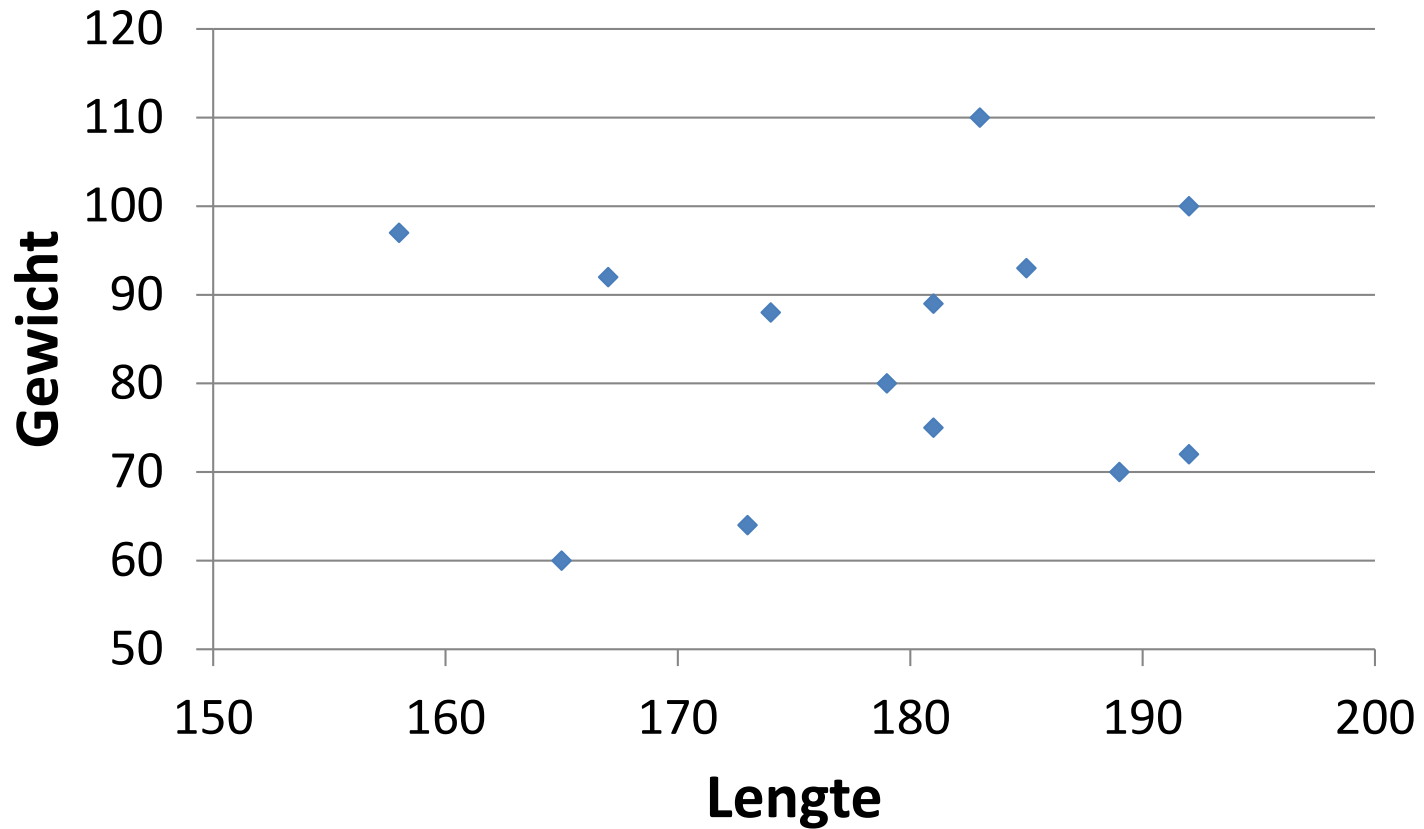


# Lineaire regressie

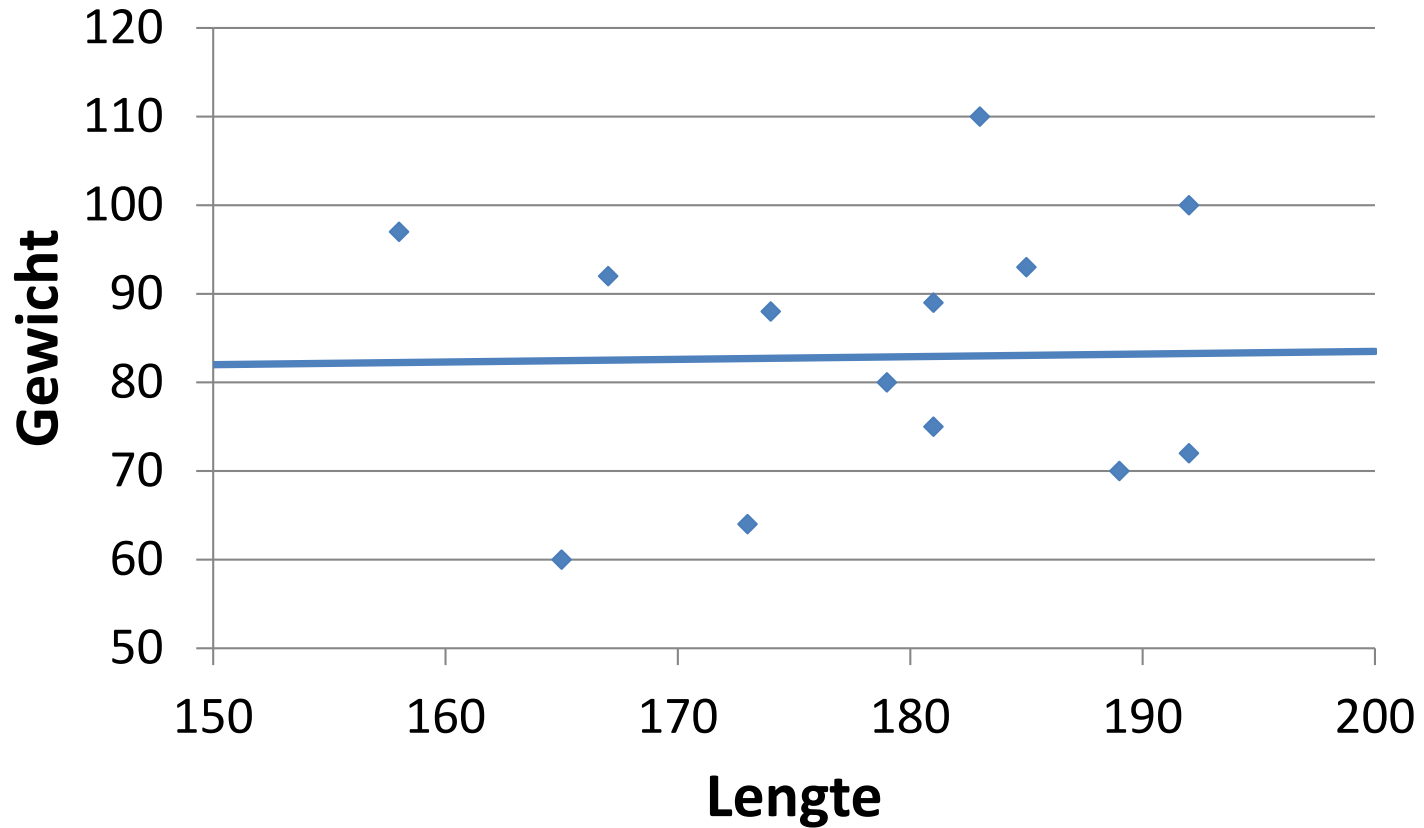


$$\hat{y} = \beta_0 + \beta_1 x$$

# Lineaire regressie



# Lineaire regressie



$$\hat{y} = \beta_0 + \beta_1 x$$

$\neq 0?$

# Hypothese test

$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0$$

$H_0$ : geen verband    versus     $H_1$ : een verband

Chi square-test

$p$ -waarde: de kans dat je  $H_0$  onterecht verwerpt

Verwerp  $H_0$  als  $p < \alpha = 0.05$









G Disease

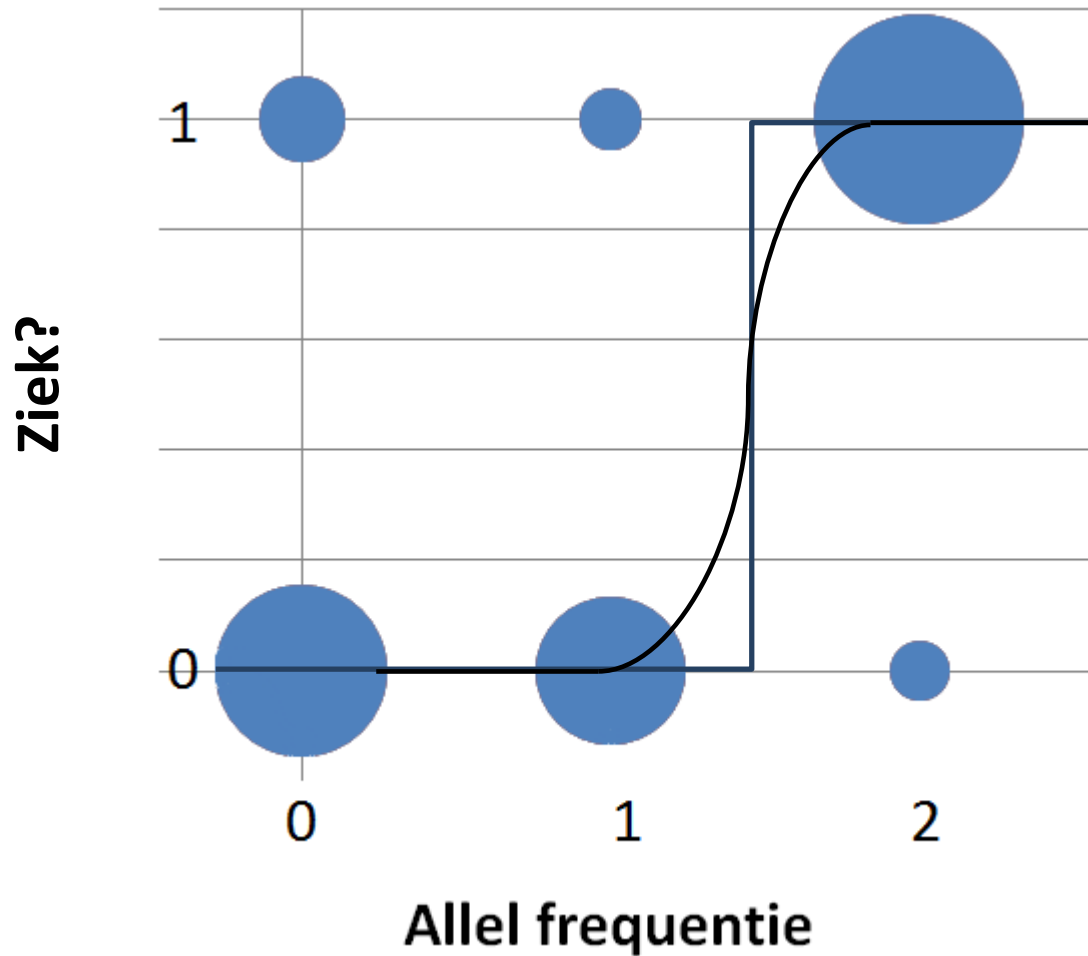
Terug naar  
genetica

# Weet u nog?

## DNA profiel

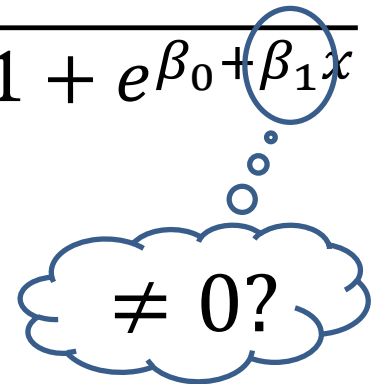
	1	2	0	0	2	1	2	1	0
	0	2	1	0	1	0	0	0	0
	1	1	0	0	2	1	2	1	0
	1	0	0	0	1	0	0	1	0
	2	0	0	1	0	1	0	0	1
	1	1	0	1	2	2	2	0	2

# Logistische regressie









Sigmoïde functie:

$$y = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$



# Weet u nog?

## DNA profiel

	1	2	0	0	2	1	2	1	0
	0	2	1	0	1	0	0	0	0
	1	1	0	0	2	1	2	1	0
	1	0	0	0	1	0	0	1	0
	2	0	0	1	0	1	0	0	1
	1	1	0	1	2	2	2	0	2





ANNO MDCCXL

mm  
ll  
kk  
ii

ll  
mm  
kk  
ii  
hh  
gg  
ff  
ee  
dd  
cc  
bb  
aa

mm  
ll  
kk  
ii  
hh  
gg  
ff  
ee  
dd  
cc  
bb  
aa

mm  
ll  
kk  
ii

# Hypothese test

$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0$$

$H_0$ : geen verband    versus     $H_1$ : een verband

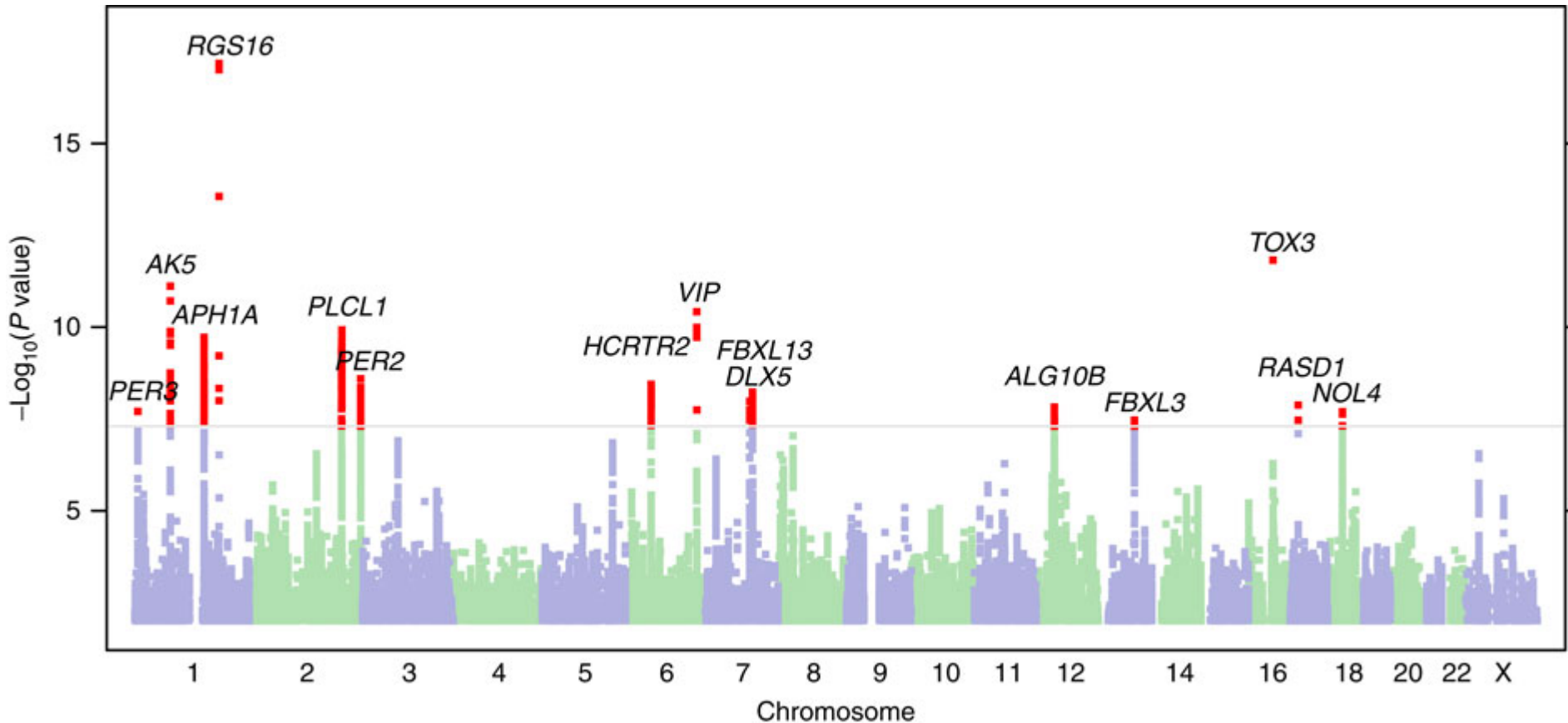
Chi-square-test

$p$ -waarde: de kans dat je  $H_0$  onterecht verwerpt

Verwerp  $H_0$  als  $p < \alpha = 0.05$

# Genoom-wijde associatie studies

$p < 10^{-7}$   
 $\Leftrightarrow \text{Log}_{10}(p) < -7$   
 $\Leftrightarrow -\text{Log}_{10}(p) > 7$



# GWAS successen

**GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer**

nature  
genetics

Multiple loci identified in a genome-wide association study of prostate cancer

Science News

Gene variant that protects against Alzheimer's disease identified

**NIH** **NATIONAL CANCER INSTITUTE**

**A Story of Discovery: HER2's Genetic Link to Breast Cancer Spurs Development of New Treatments**

Neurology®

**Analysis of GWAS-linked loci in Parkinson disease reaffirms PARK16 as a susceptibility locus**



Disease

Maar...

# Gen-gen interacties



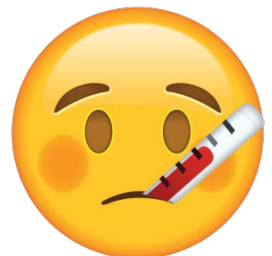
AND



# Gen-gen interacties



XOR



# Gen-gen interacties - effect

Aantal interacties	Aantal tests	Rekentijd	Grens $p$ -waarde
1	$n$	Week	$0.05/n$



**Wiskundige optimalisatie**

**Expert systemen**

**Machine learning**

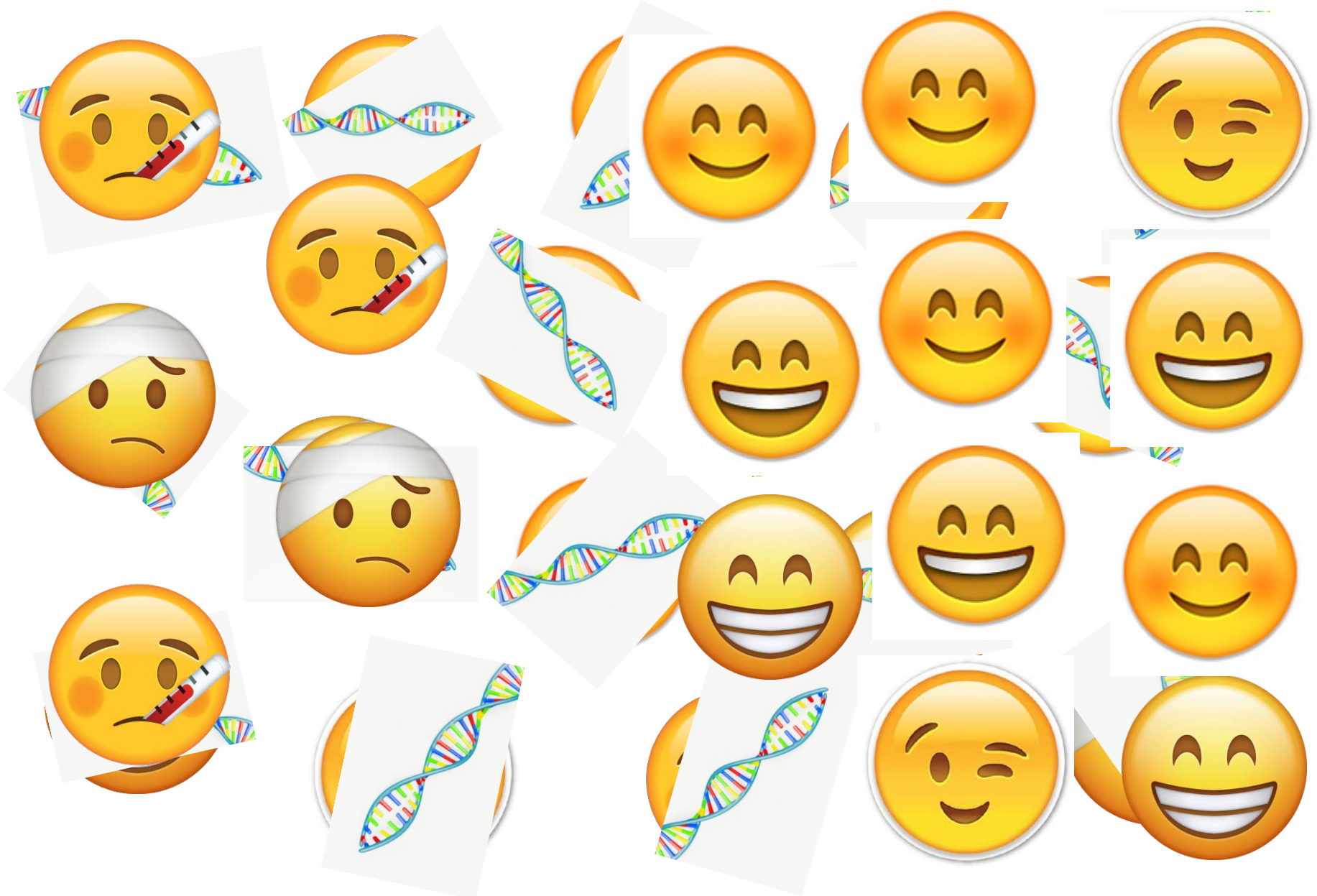
**Kunstmatige intelligentie**

**Big data analytics**

**Deep learning**

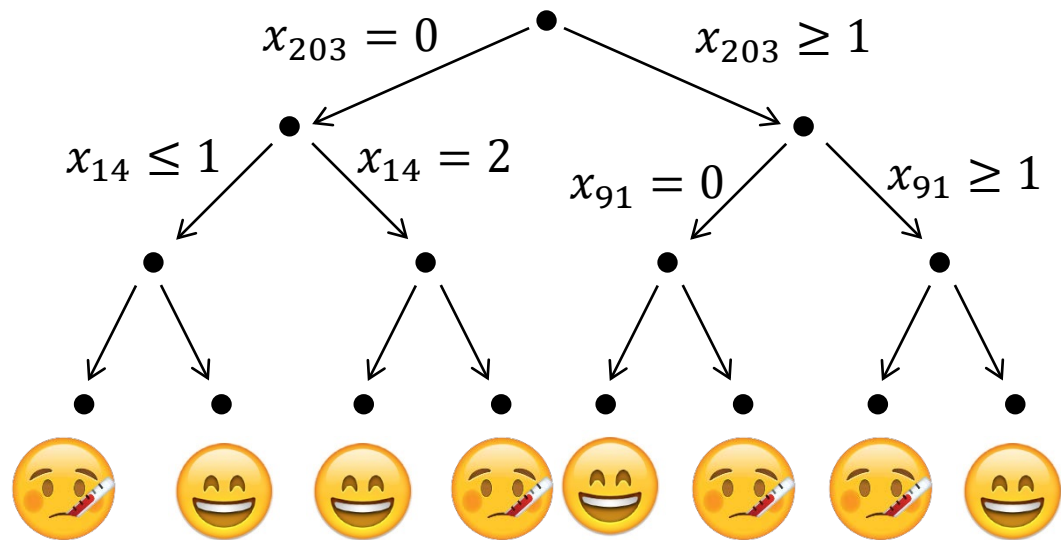
**Statistiek**

# Gezonde en zieke mensen



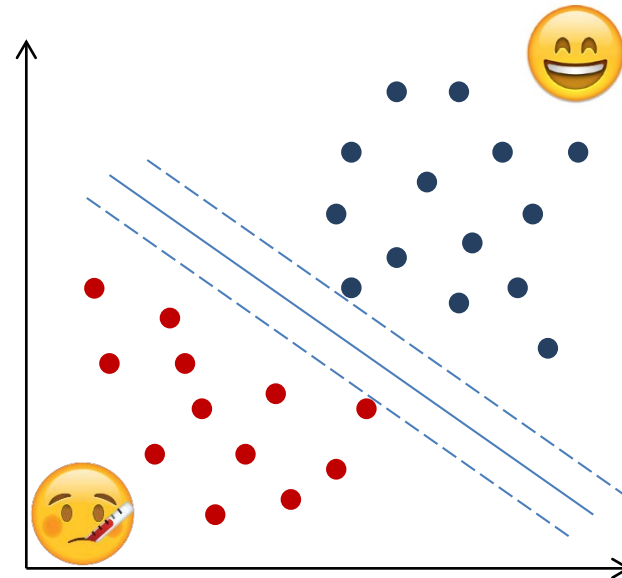
# Machine learning methoden

- Lineaire/logistische regressie
- Random forest



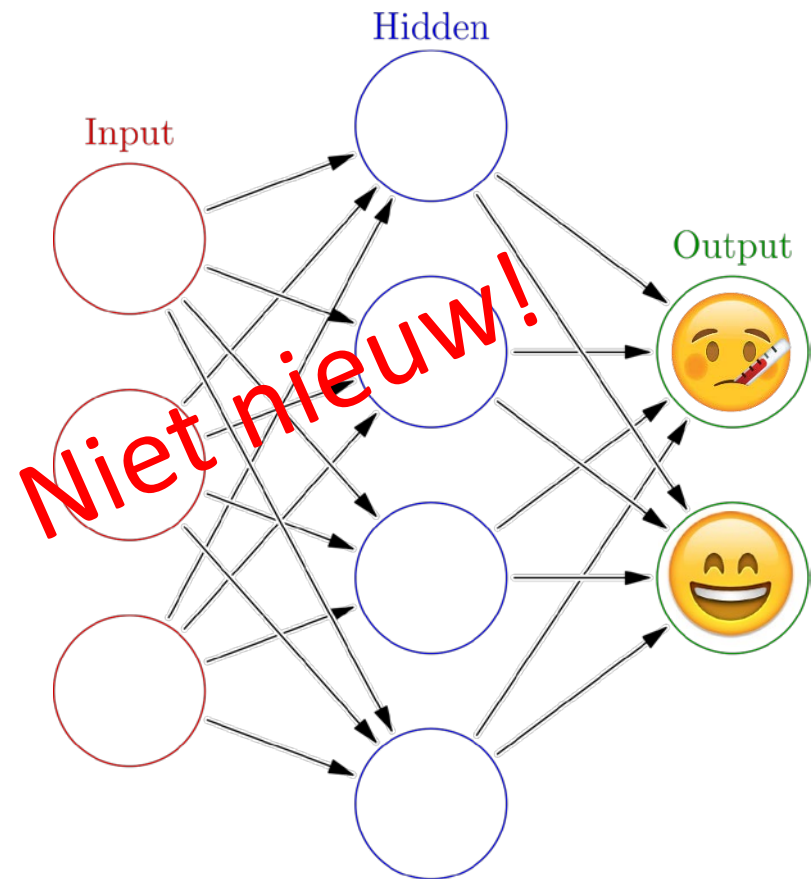
# Machine learning methoden

- Lineaire/logistische regressie
- Random forest
- Support vector machine



# Machine learning methoden

- Lineaire/logistische regressie
- Random forest
- Support vector machine
- Neuraal netwerk
- Bayesiaanse netwerken
- Genetische algoritmen
- ...





# Huidig onderzoek

## Kansen:

- Grote datasets
- Sterke computers (GPUs)

## Uitdagingen:

- Grote datasets
- Interpreteerbaarheid



DNA Disease

# Regressie, maar nu geheeltallig...

