Veel van weinig

Hoe vinden we diagnostische markers voor kanker op basis van een beperkt aantal genetische profielen?

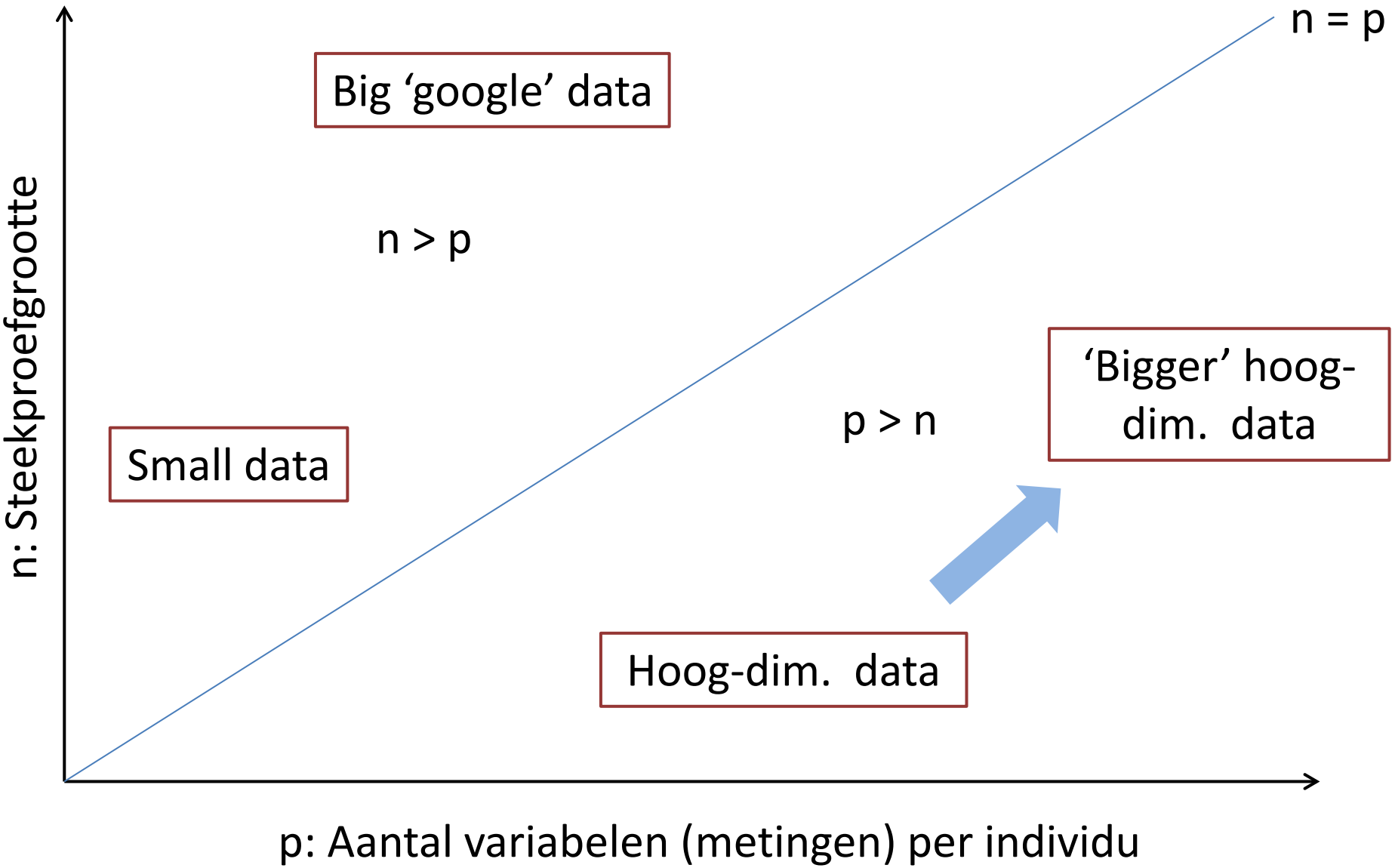
Mark van de Wiel
mark.vdwiel@vumc.nl

Afdeling Epidemiologie & Biostatistiek
Afdeling Wiskunde
VUmc/VU, Amsterdam

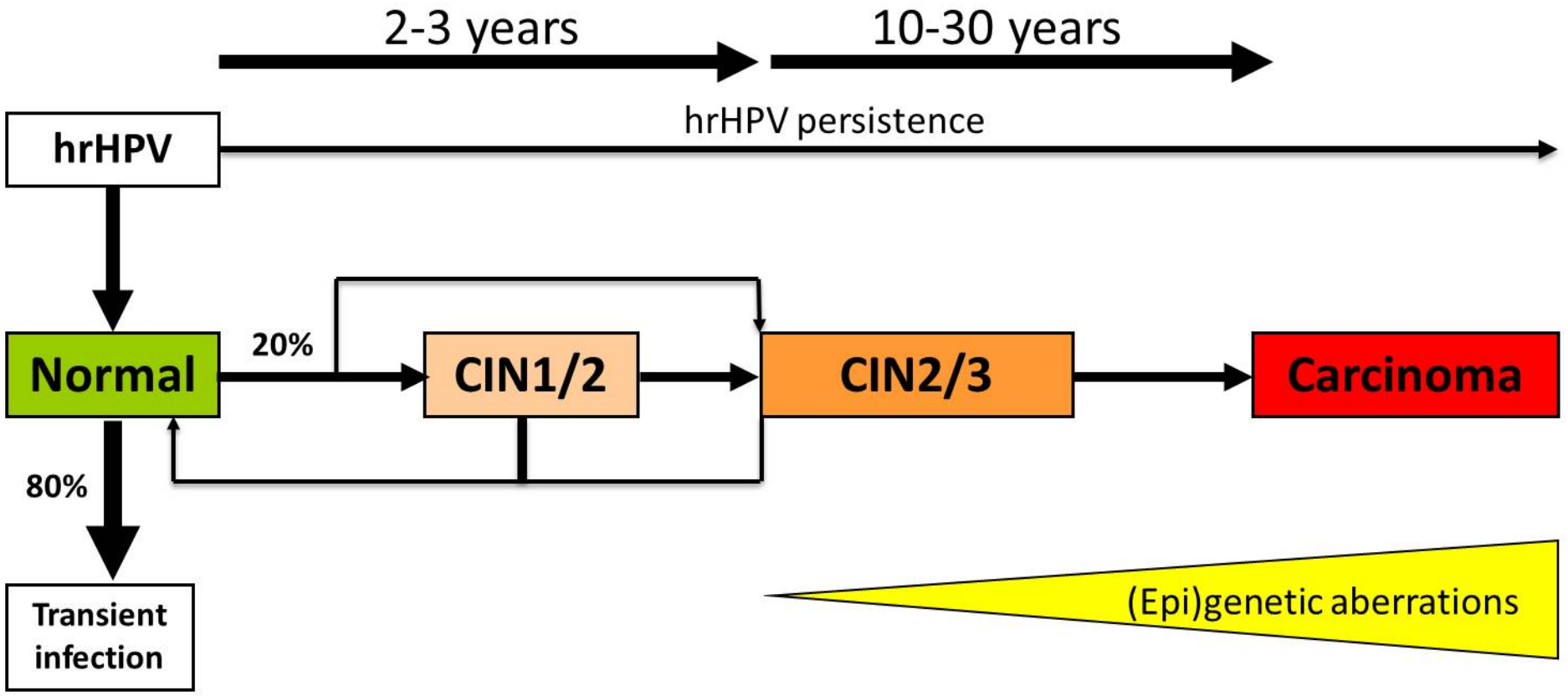
vrije Universiteit

VU medisch centrum

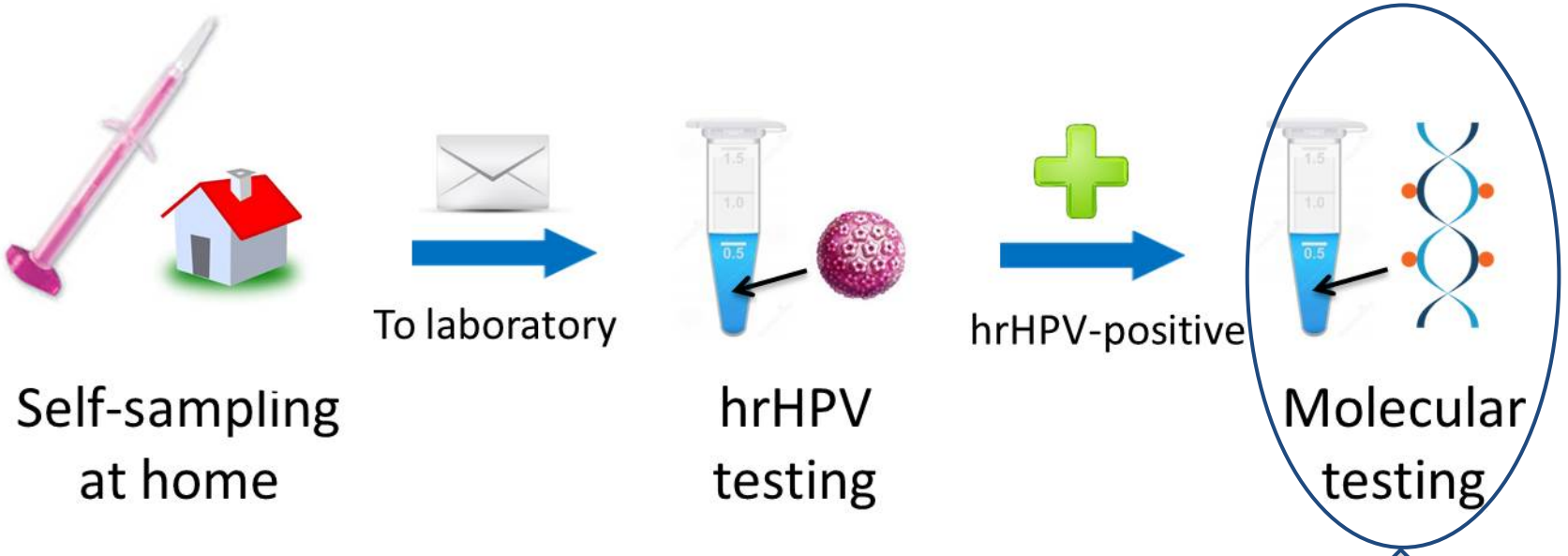
Eerst een ‘big’ misverstand…



Ontwikkeling van baarmoederhalskanker

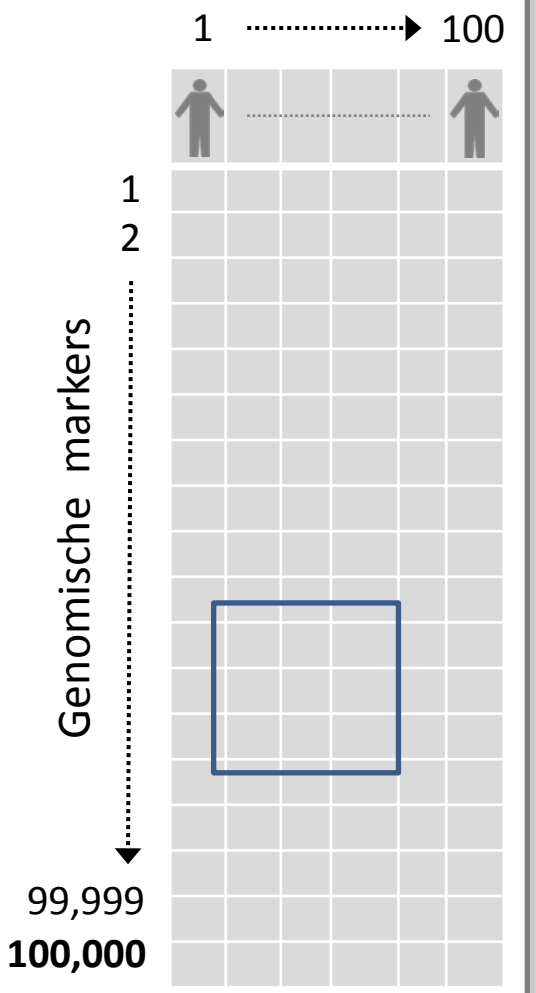


Wat wil men testen?



Welke moleculaire markers (bijv. genen) moeten getest worden?

Data



Matrix $\mathbf{X} = (X_{ij})_{i=1,\ldots,p;j=1,\ldots,n}$

n = 100 = steekproefgrootte

p = 100,000 = # variabelen

=

$X_{10,2}$	$X_{10,3}$	$X_{10,4}$
$X_{11,2}$	$X_{11,3}$	$X_{11,4}$
$X_{12,2}$	$X_{12,3}$	$X_{12,4}$

=

10	0	15
122	5	67
34	0	1209

Response vector $\mathbf{Y} = (Y_1,\ldots,Y_n)$

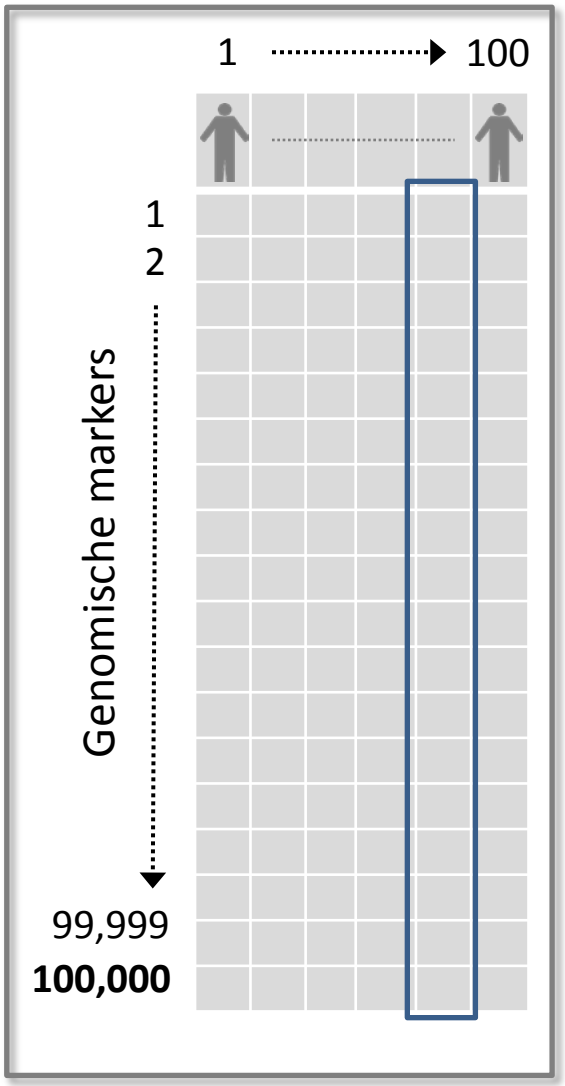
Binair (ziek/niet ziek):

Y = 1 1 0 0 1 ...

Continu (bijv. BMI):

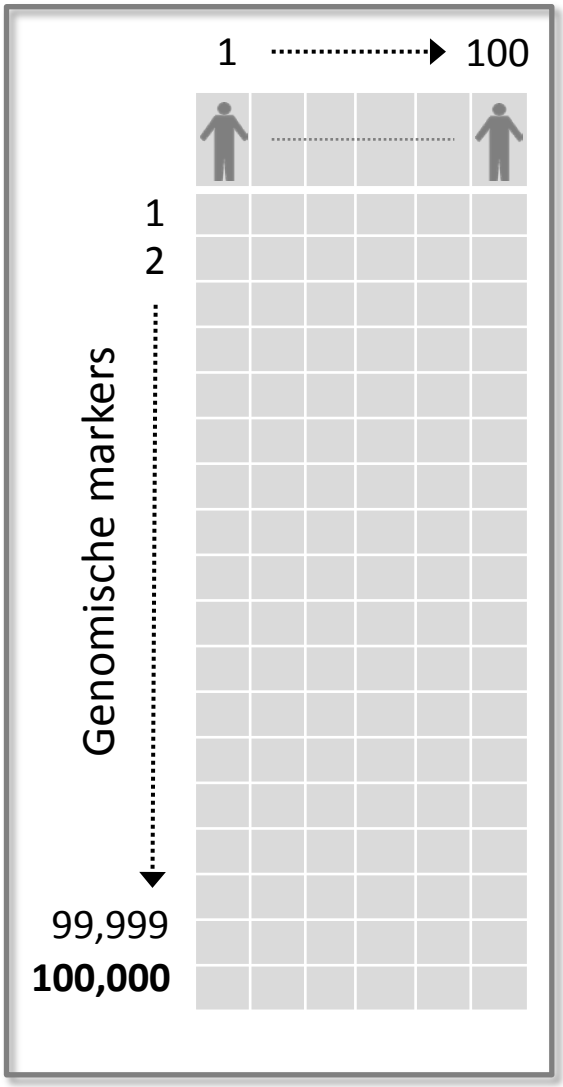
Y = 24.6 17.2 21.3 29.4 19.1 ...

Genomisch profiel

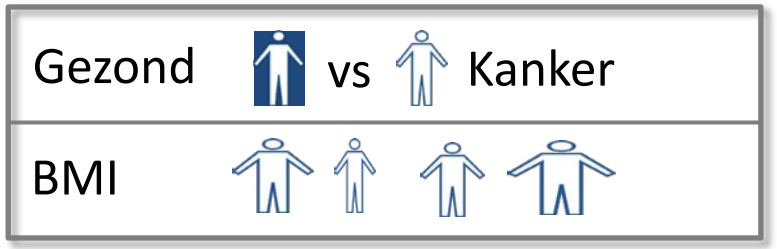


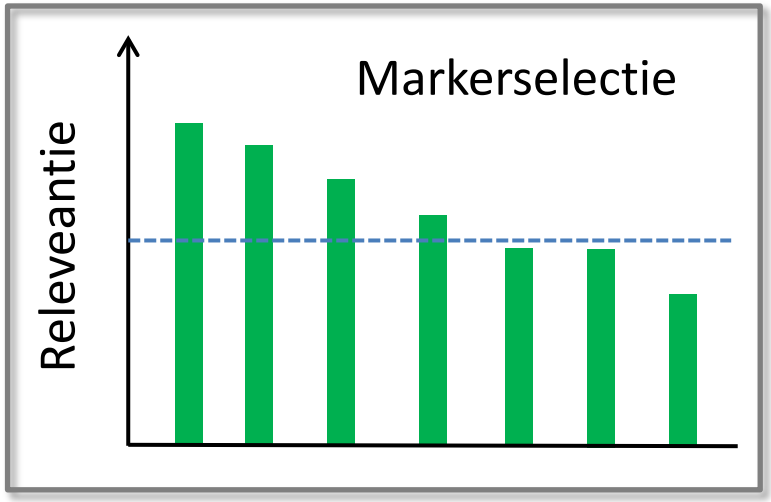
= Genomisch profiel = X_j

Predictie



Predictie Model





Wat willen we?

- (**X,Y**): training data. Waarom training?

➡ Omdat we voor deze individuen de response al hebben!

- X_{nieuw} = genprofiel nieuw individu.

- Doel 1: Een predictor (formule of algoritme) f zodanig dat:

$$\hat{Y}_{nieuw} \approx f_{X,Y}(X_{nieuw})$$

- Doel 2: Verklein genprofiel X tot Z met slechts 5 genen (goedkopere test!)

$$\hat{Y}_{nieuw} \approx f'_{X,Y}(Z_{nieuw})$$

Lineaire Regressie

- Door grote p is het probleem al moeilijk genoeg: laten we f lineair nemen:

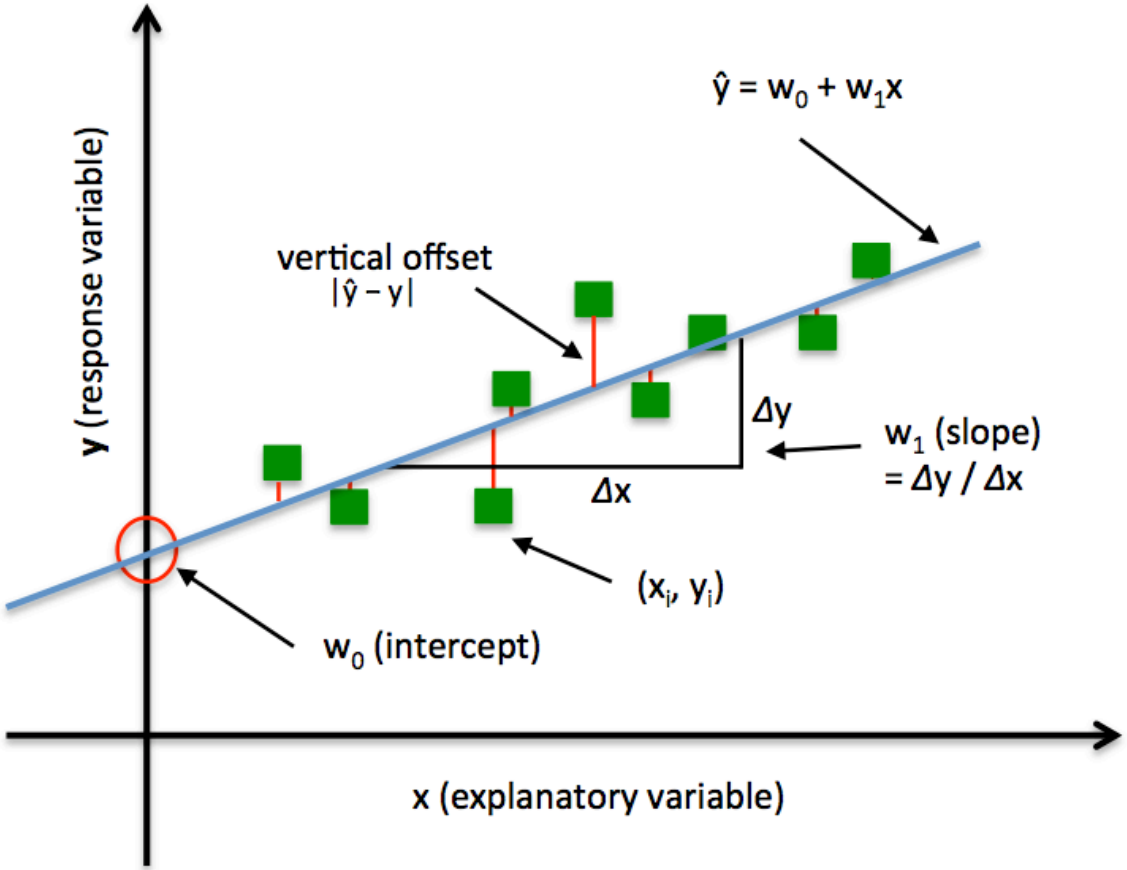
$$Y_j = w_0 + w_1 X_{1j} + w_2 X_{2j} + \ldots + w_p X_{pj} + \varepsilon_j$$

- ε_j is de statistische voorspelfout
- Doel 1: vind $\mathbf{w} = (w_0, w_1, \ldots, w_p)$ zo dat ε_j klein, oftewel

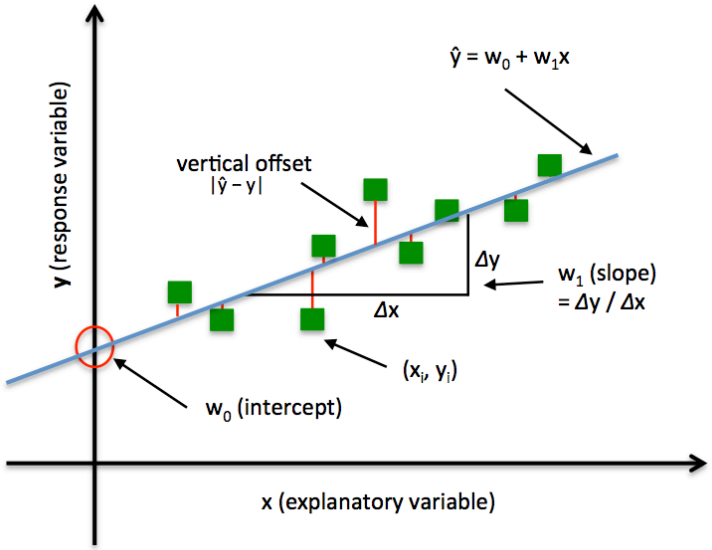
$$Y_j \approx w_0 + \sum_{i=1}^{p} w_i X_{ij}$$

Regressie, 1 gen…

$$Y_j = w_0 + w_1 X_{1j} + \varepsilon_j$$



Regressie, kleinste kwadraten



Minimaliseer $|\ |^2 + |\ |^2 + |\ |^2 + |\ |^2 + \ldots$

$$\Updownarrow$$

$$\min_{w_0,w_1} \sum_{j=1}^{n} \left(Y_j - (w_0 + w_1 X_j)\right)^2$$

$$\Updownarrow$$

$$\frac{d}{dw_0} \sum_{j=1}^{n} \left(Y_j - (w_0 + w_1 X_j)\right)^2 = -2 \sum_{j=1}^{n} \left(Y_j - (w_0 + w_1 X_j)\right) =^{\text{stel}} 0$$

$$\frac{d}{dw_1} \sum_{j=1}^{n} \left(Y_j - (w_0 + w_1 X_j)\right)^2 = -2 \sum_{j=1}^{n} X_j \left(Y_j - (w_0 + w_1 X_j)\right) =^{\text{stel}} 0$$

$$\rightarrow \hat{w}_1 = \frac{\sum_{j=1}^{n} (X_j Y_j - X_j \bar{Y})}{\sum_{j=1}^{n} (X_j^2 - X_j \bar{X})}, \quad \hat{w}_0 = \frac{1}{n} \sum_{j=1}^{n} \left(Y_j - \hat{w}_1 X_j\right)$$

Lineaire Regressie

$$Y_j = w_0 + w_1 X_{1j} + w_2 X_{2j} + \ldots + w_p X_{pj} + \varepsilon_j$$

$$(\hat{w}_0, \ldots, \hat{w}_p) = \min_{w_1,\ldots,w_p} \sum_{j=1}^{n} \left(Y_j - (w_0 + w_1 X_{1j} + \ldots + w_p X_{pj}) \right)^2$$

$$\hat{\mathbf{w}} = (\hat{w}_0, \ldots, \hat{w}_p) = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}$$

(waarbij we **X** hebben uitgebreid met $X_0 = (1,\ldots,1)$, lengte n)

Voorbeeld, p=2

$$Y_j = w_0 + w_1 X_{1j} + w_2 X_{2j} + \varepsilon_j$$

$$\hat{\mathbf{w}} = (\hat{w}_0, \hat{w}_1, \hat{w}_2) = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}$$

Bijvoorbeeld:

Y_j = BMI van individu j

X_{1j} = Hoeveelheid vet die individu j per maand eet

X_{2j} = Hoeveelheid suiker die individu j eet per maand

Simulatie voorbeeld, p=2

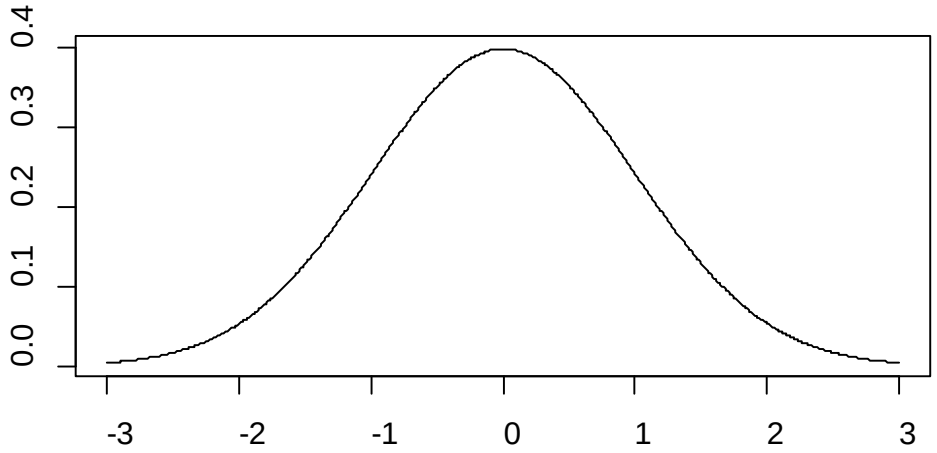
Simulatiemodel:

$$Y_j = w_0 + w_1 X_{1j} + w_2 X_{2j} + \varepsilon_j$$

$$w_0 = 10, w_1 = 1, w_2 = 2$$

$$\varepsilon_j \sim N(m, sd)$$

$$m = 0, sd = 1$$



Standaardnormale verdeling

Voorbeeld in R, p=2

R: Statistische software

- **Gratis**
- Lichte installatie met basis functionaliteit
- Bijdragen van wiskundigen uit heel de wereld (packages)
- Download: https://cran.r-project.org/
- Erg populair voor analyse Big data/hoog-dimensionale data

Simulatie voorbeeld, grote p

Simulatiemodel:

$$Y_j = w_0 + w_1 X_{1j} + w_2 X_{2j} (+0X_{3j} + \ldots + 0X_{pj}) + \varepsilon_j$$

$$w_0 = 10, w_1 = 1, w_2 = 2$$

Analysemodel:

$$Y_j = w_0 + w_1 X_{1j} + w_2 X_{2j} + \ldots + w_p X_{pj} + \varepsilon_j$$

X_{ij} en X_{kj} onafhankelijk

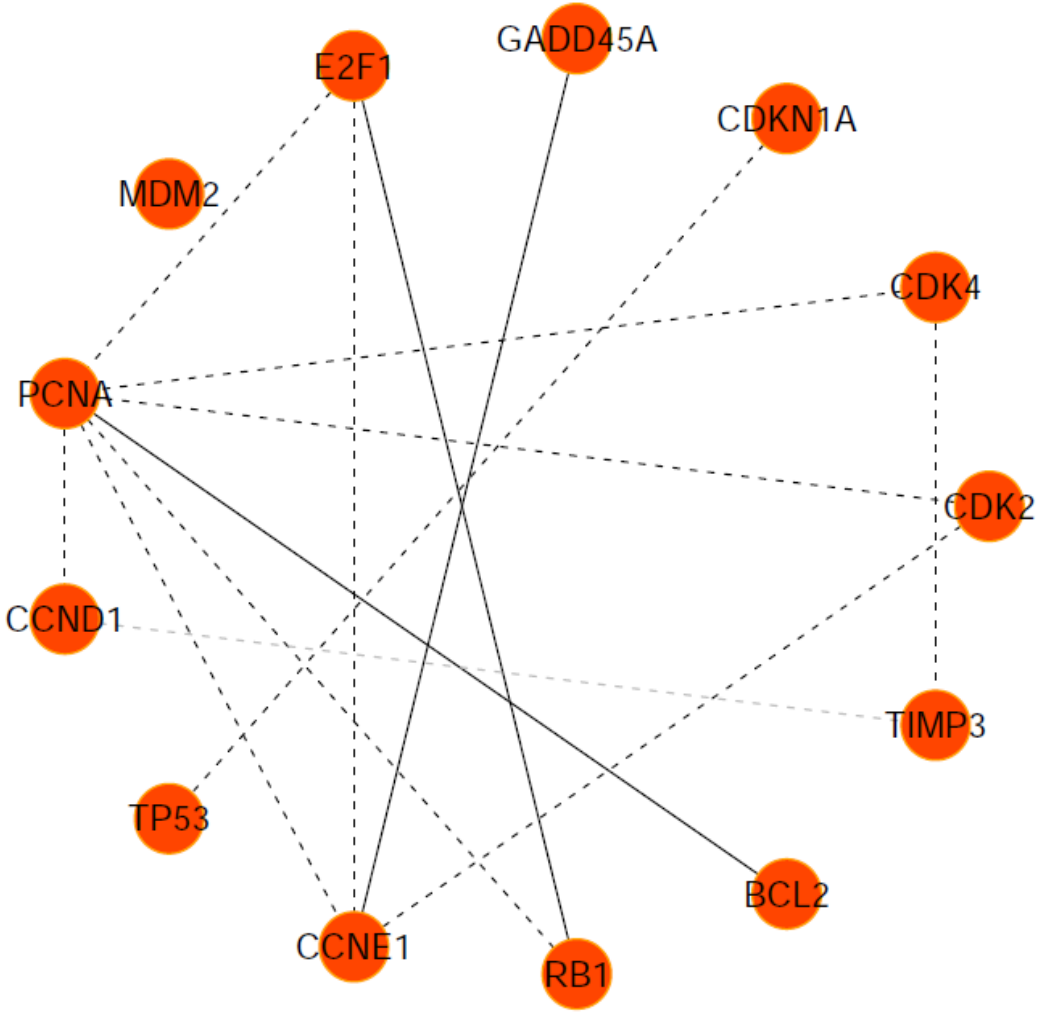
Simulatie voorbeeld, grote p

Analysemodel:

$$Y_j = w_0 + w_1 X_{1j} + w_2 X$$

X_{ij} en X_{kj} (gen i en gen k) onafhankelijk?

NEE!



Regressie voor hoog-dimensionale data

Gewone regressie 'crasht' als p > n!

Wat nu? Regulariseren!

$$(\hat{w}_0,\ldots,\hat{w}_p) = \min_{(w_1,\ldots,w_p)\in C_\lambda} \sum_{j=1}^{n} \left(Y_j - (w_0 + w_1 X_{1j} + \ldots + w_p X_{pj})\right)^2$$

Ridge regressie:

$$B_\lambda = \{\mathbf{w} : (w_1)^2 + (w_2)^2 + \quad \ldots \quad + (w_p)^2 \leq \lambda\}$$

Wat is B_λ voor een gebied?

Regressie voor hoog-dimensionale data

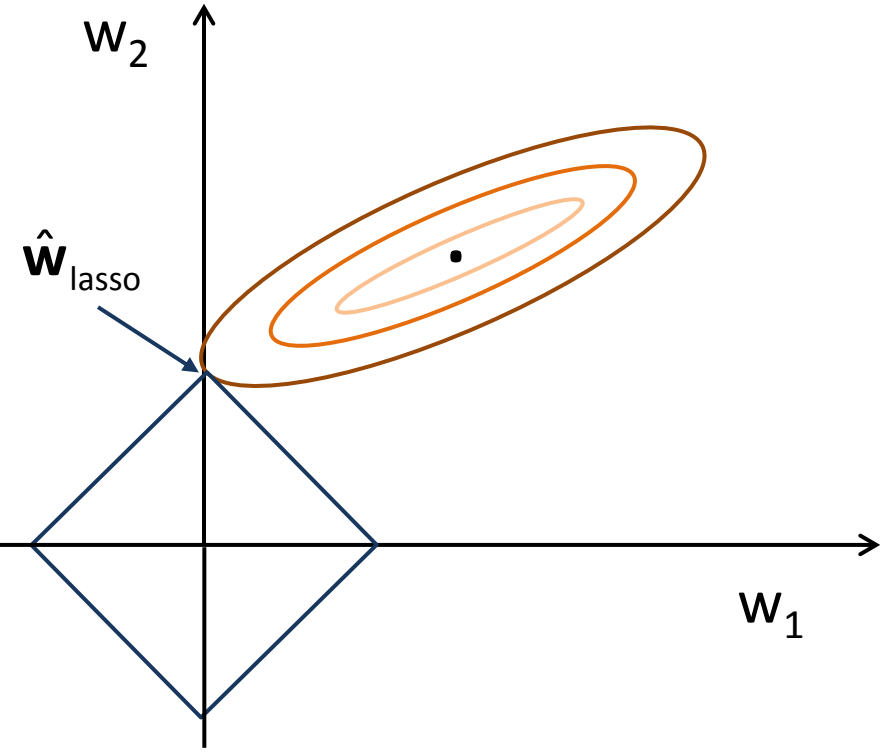
$$(\hat{w}_0,\ldots,\hat{w}_p)=\min_{(w_1,\ldots,w_p)\in V_\lambda}\sum_{j=1}^{n}\left(Y_j-(w_0+w_1X_{1j}+\ldots+w_pX_{pj})\right)^2$$

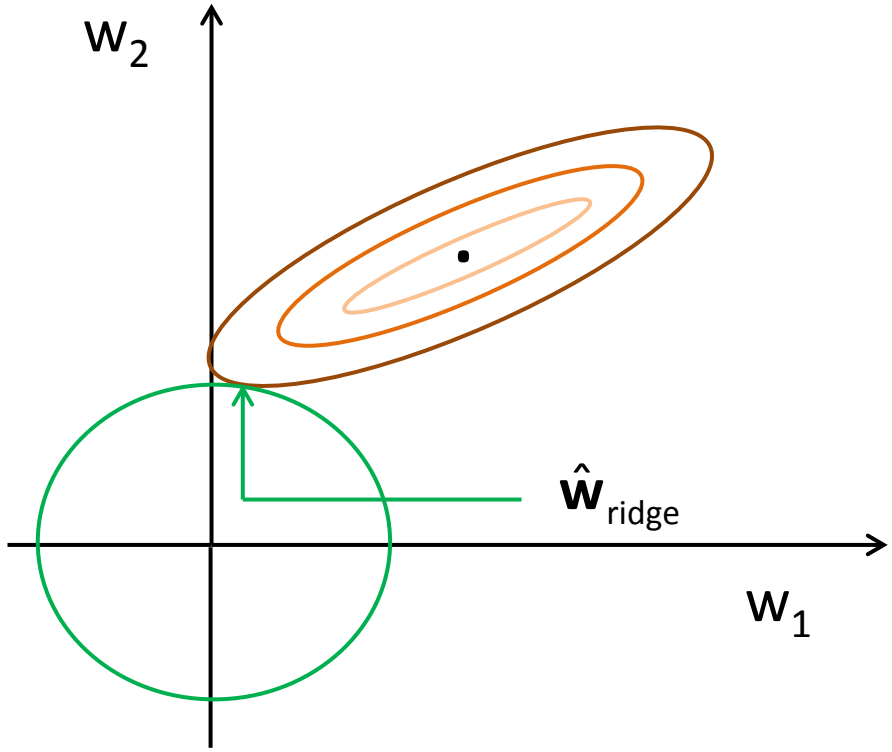
Lasso regressie: $K_\lambda=\{\mathbf{w}:|w_1|+|w_2|+\ \ldots\ +|w_p|\leq\lambda\}$

Wat is K_λ voor een gebied?

lasso vs ridge

$$g(w_1, w_2) = \sum_{j=1}^{n} \left(Y_j - (w_1 X_{1j} + w_1 X_{2j})\right)^2 \qquad [w_0 = 0]$$





lasso: $K_\lambda = \{\mathbf{w} : |w_1| + |w_2| \leq \lambda\}$

ridge: $B_\lambda = \{\mathbf{w} : (w_1)^2 + (w_2)^2 \leq \lambda\}$

lasso en ridge

lasso en ridge regressie:

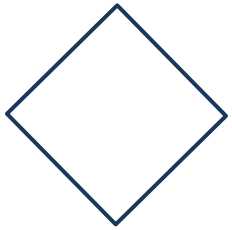
- Geven unieke oplossing (en daarmee predictiemodel)
- lasso zet coëfficiënten op 0 → variabelen-selectie
- Stel: w_0 =18.24, w_{34} = 0.45, w_{3453}= 0.78, w_{7343}= -0.91
- (Z_{1j}, Z_{2j}, Z_{3j}) = (X_{34j}, X_{3453j}, X_{7343j}), dan voorspellingsmodel:

$$Y_{nieuw} = 18.24 + 0.45Z_{1,nieuw} + 0.78Z_{2,nieuw} - 0.91Z_{3,nieuw}$$

lasso en ridge: welke λ?

Predictie en dus ook de kwaliteit ervan hangt af van λ:

Kleine λ 

Of grote λ 

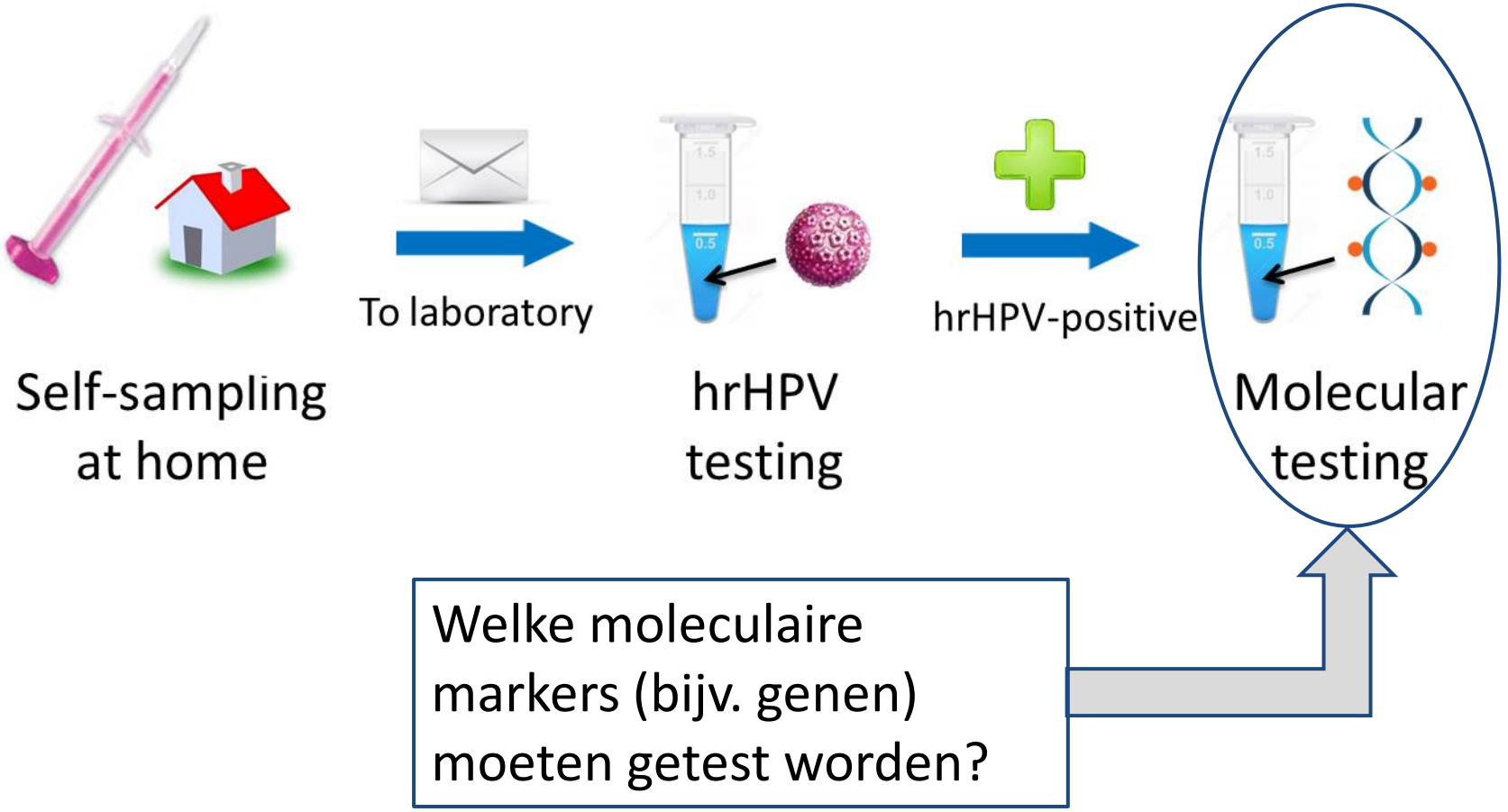
Kleine λ:

- Beperkt erg de invloed van alle genen, ook de belangrijke
+ Stabiele oplossing

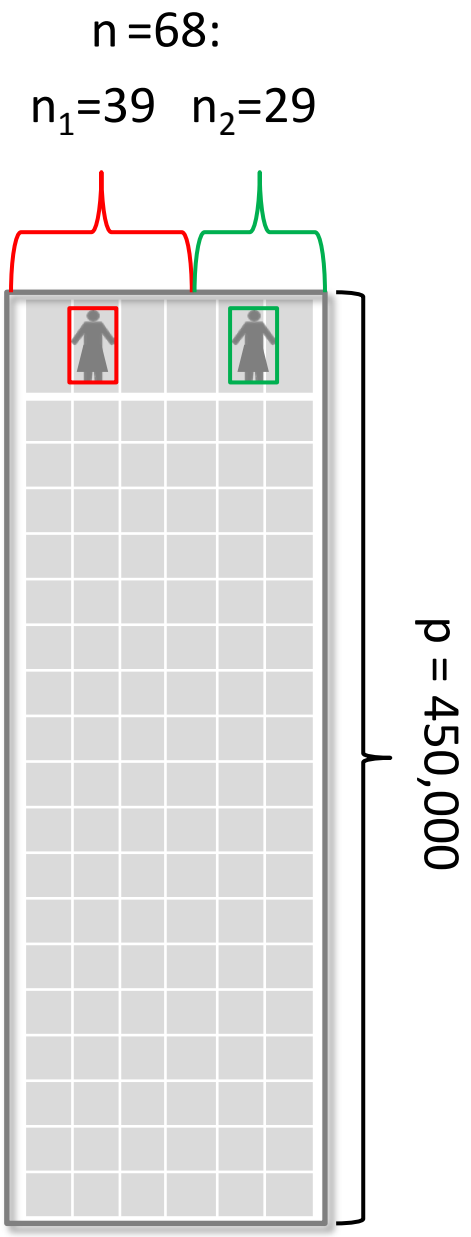
Grote λ:

- Meer vrijheid voor de belangrijke genen
+ Mogelijk instabiele oplossing

Terug naar het probleem…



Terug naar het probleem…



CIN3

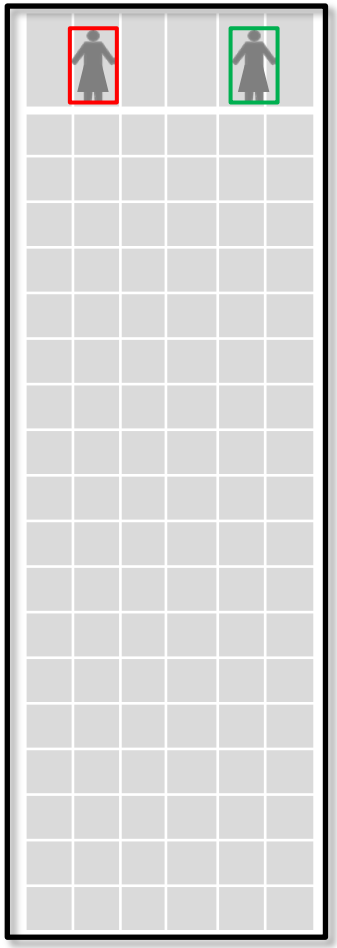
Controle

Doel:

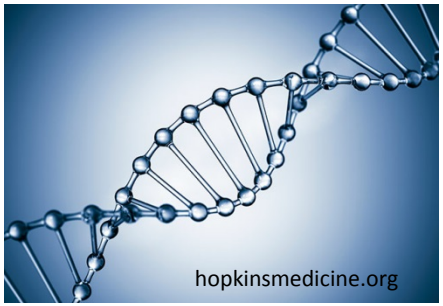
- 10-15 DNA markers die goed CIN3 onderscheiden van controle
- Enorm speld-in-hooiberg probleem
- Big-data aanpak: gebruik andere databronnen om de oplossing te sturen

Co-Data

Databases

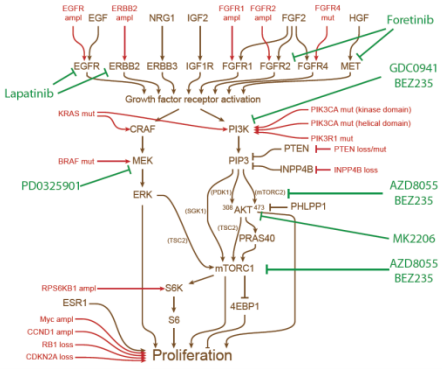


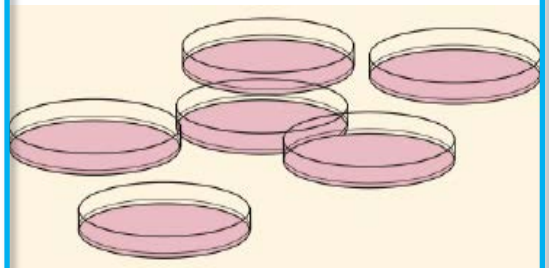
Primaire Data



Andere moleculen

Netwerken





Cell lijnen

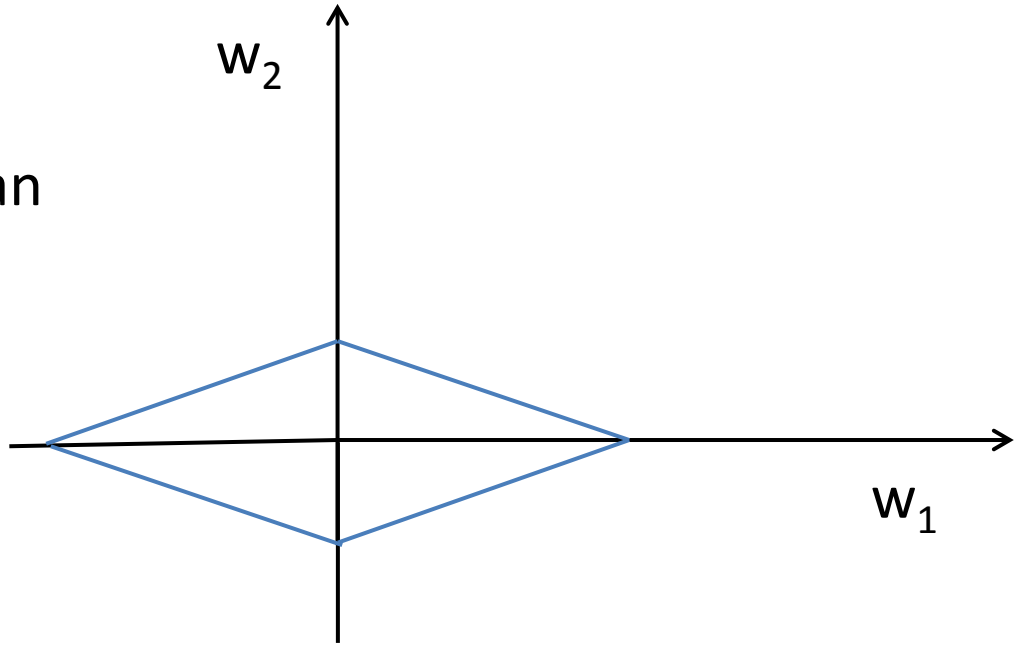
Gebruik co-data

1. Gebruik de co-data om je genomische variabelen in groepjes op te delen.

2. Ieder groepje zijn eigen λ

3. Variabelen uit groepje 1 worden minder beperkt dan die uit groep 2

→ Worden eerder gekozen



Gewogen lasso: $K_\lambda = \{\mathbf{w} : |w_1| \leq \lambda_1, |w_2| \leq \lambda_2\}$

Resultaten

Hoe evalueren we de voorspeller / geselecteerde markers?

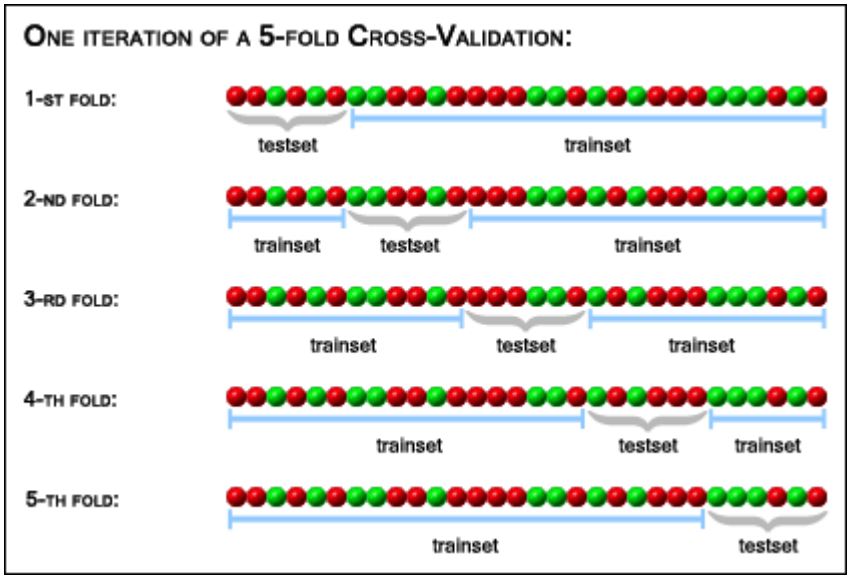
A.

- Als steekproefgrootte groot: Gebruik een *onafhankelijke* test set; vergelijk de predictie met de ware waarde

- Als steekproefgrootte klein: kruisvalidatie

- Predictie-fout:

$$\frac{1}{n_{test}} \sum_{j=1}^{n_{test}} \left| Y_{test,j} - \hat{f}(X_{test,j}) \right|$$



Bron: http://genome.tugraz.at/proclassify/help/pages/XV.html

B. Meet de geselecteerde markers op goedkoop platform voor nieuwe test samples, en evalueer dan de predictiefout

Resultaten

A

1. Voorspeller met co-data 30% accurater dan zonder co-data

2. 12 kandidaat markers

B

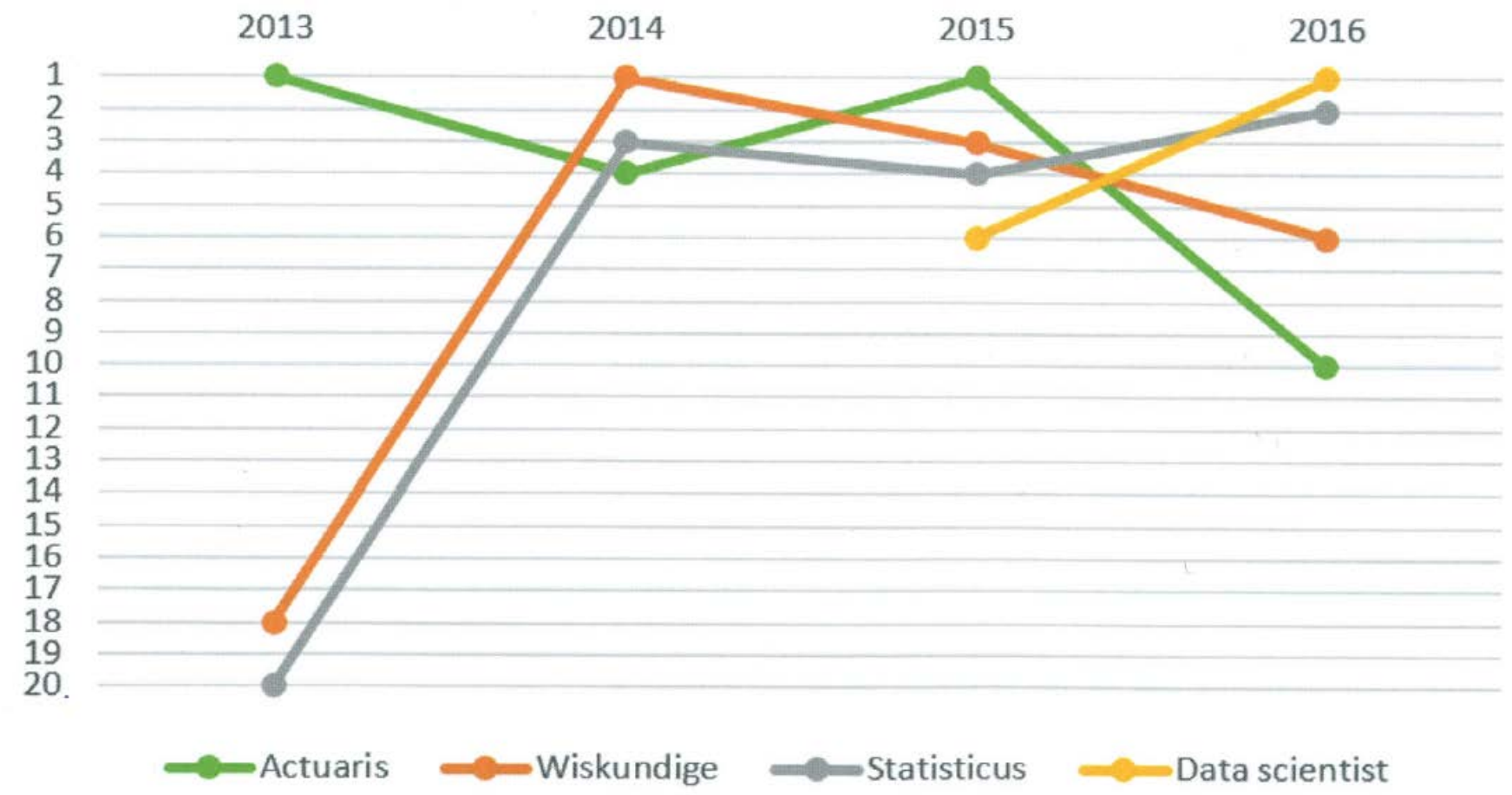
1. Grote groepen met goedkope, diagnostische techniek

2. Verdere selectie: 3 markers

3. 92% accuraat

4. Zeer competitief met huidige test, maar
   - Goedkoper en objectiever
   - Aantrekkelijk voor andere landen

Tot slot



Figuur 1. Trend in beste beroepen volgens Careercast.com