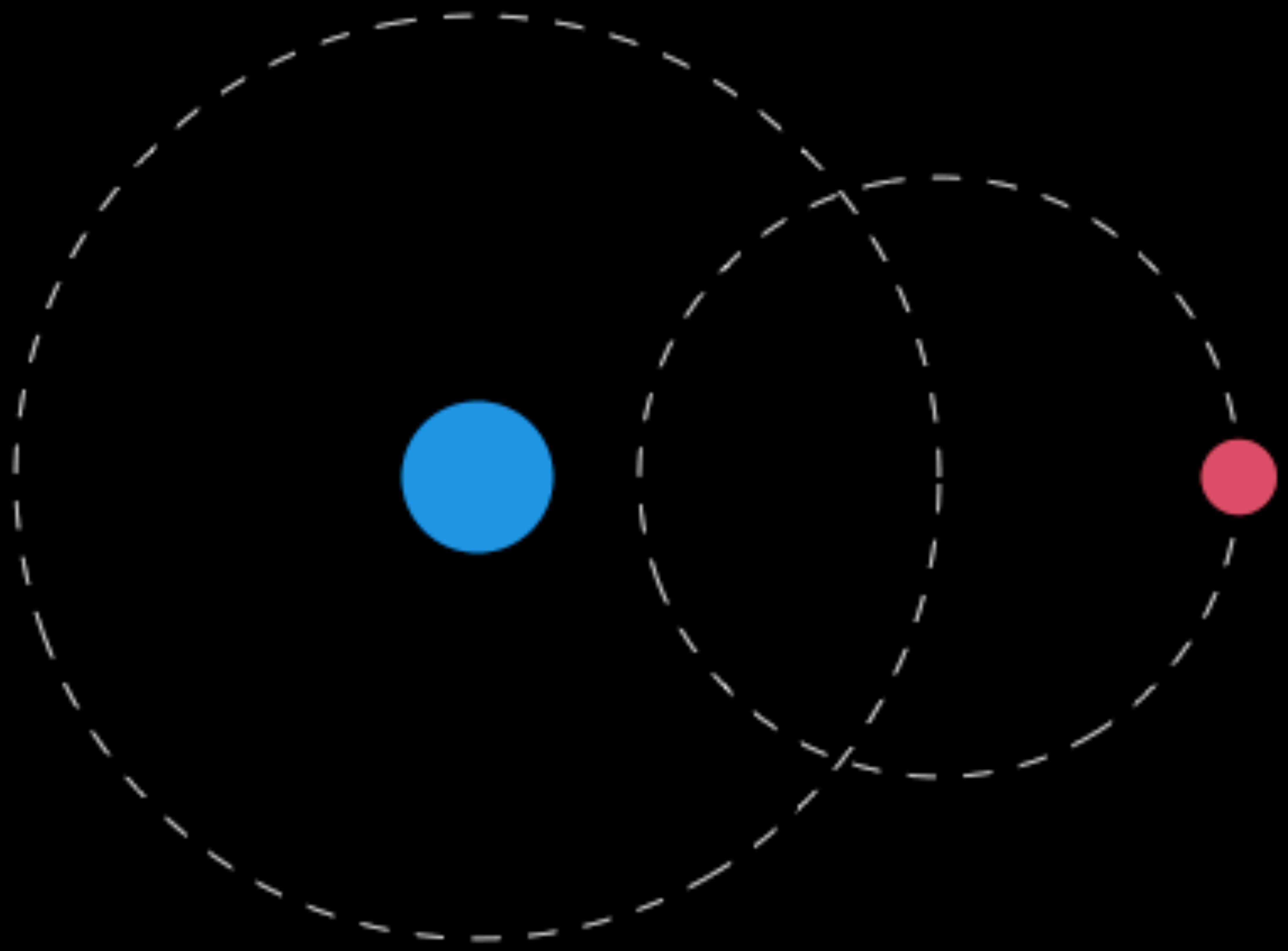
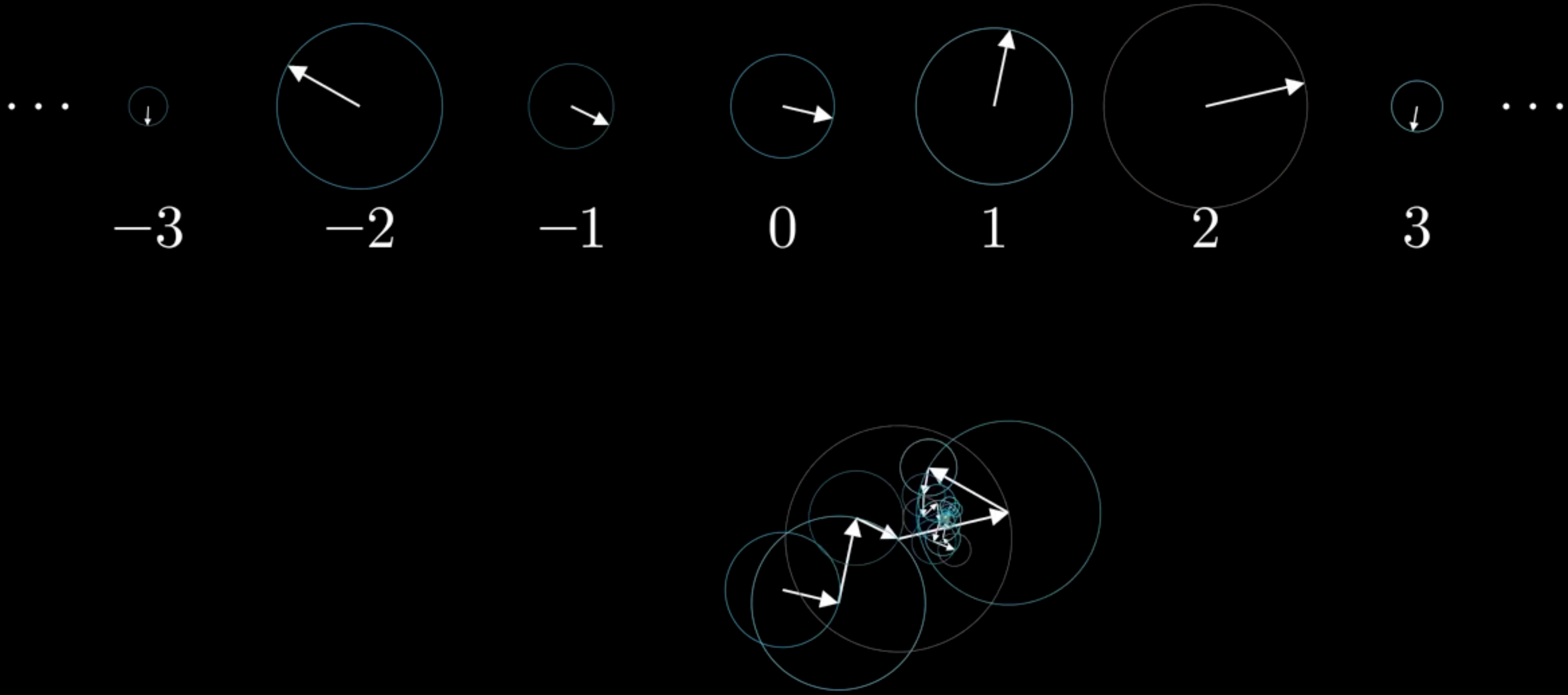
Statistical Rethinking

2023

4. Categories & Curves





www.3blue1brown.com

Drawing Inferences

How to use statistical models to get at scientific estimands?

Need to incorporate causal thinking into

(1) How we draw the statistical models
(2) How we process the results

Drawing Inferences

(1) How we draw the statistical models

Generative model + multiple estimands
=> multiple estimators

(2) How we process the results

Only very simple causal estimates show up in summary tables

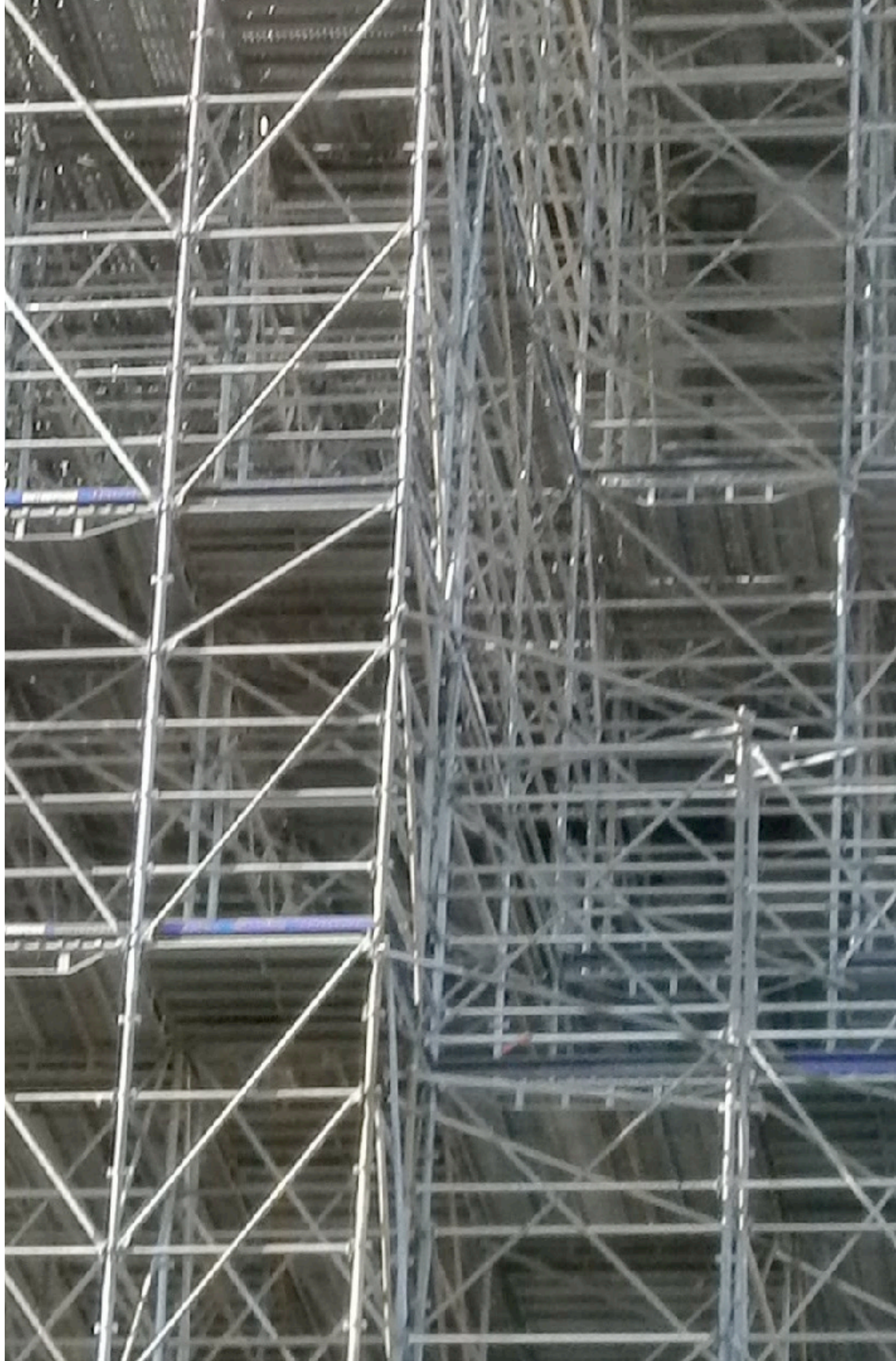
Categories & Curves

Linear models can do extra-linear things

Categories (indicator & index variables)

Splines and other additive structures

We require these tools to build causal estimators



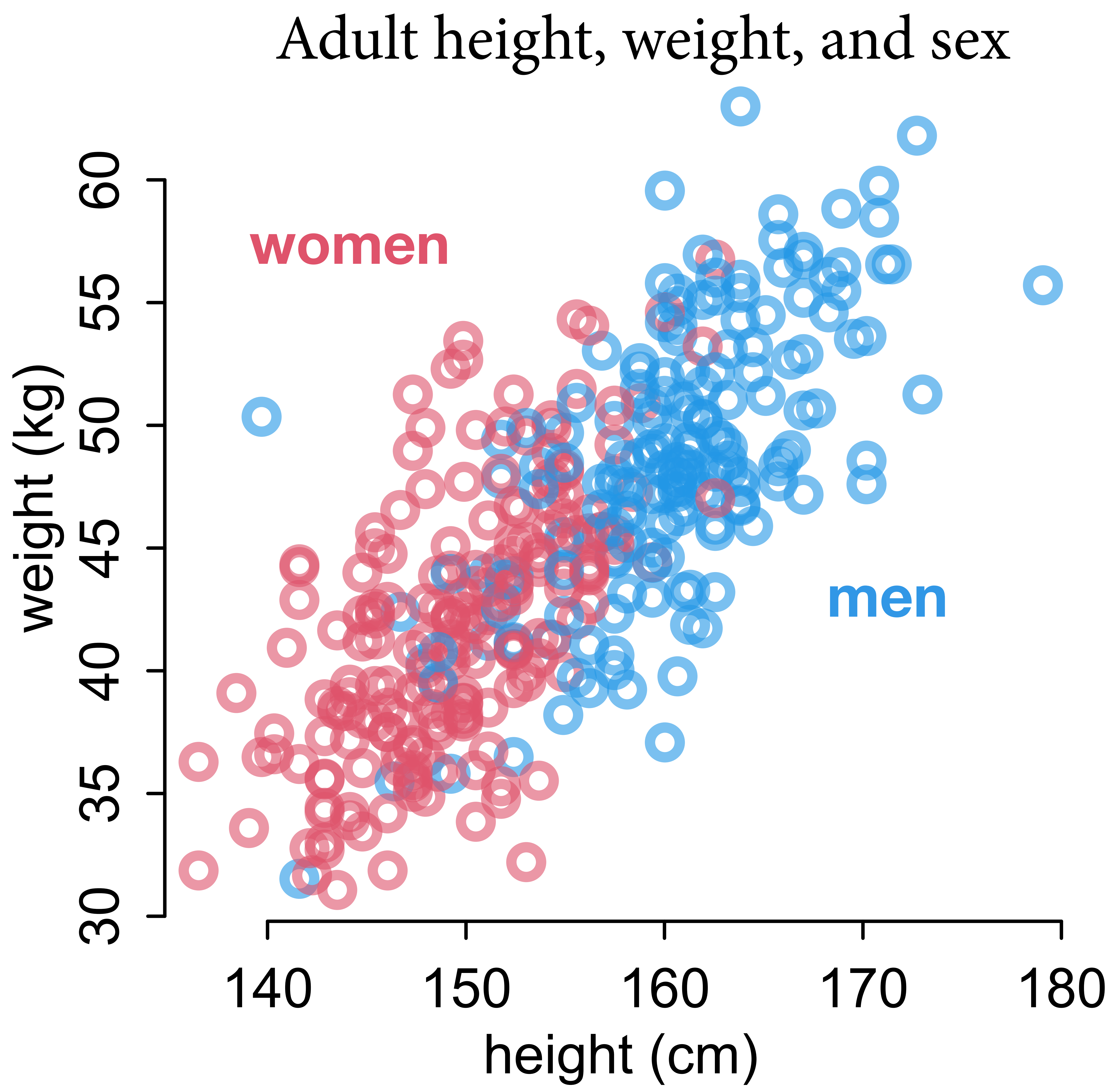
Categories

How to cope with causes that are not continuous?

Categories: discrete, unordered types

Want to **stratify** by category: Fit a separate line for each

```
> head(Howell1)
   height   weight age male
1 151.765 47.82561  63    1
2 139.700 36.48581  63    0
3 136.525 31.86484  65    0
4 156.845 53.04191  41    1
5 145.415 41.27687  51    0
6 163.830 62.99259  35    1
>
```



Think scientifically first

How are height, weight, and sex **causally** related?

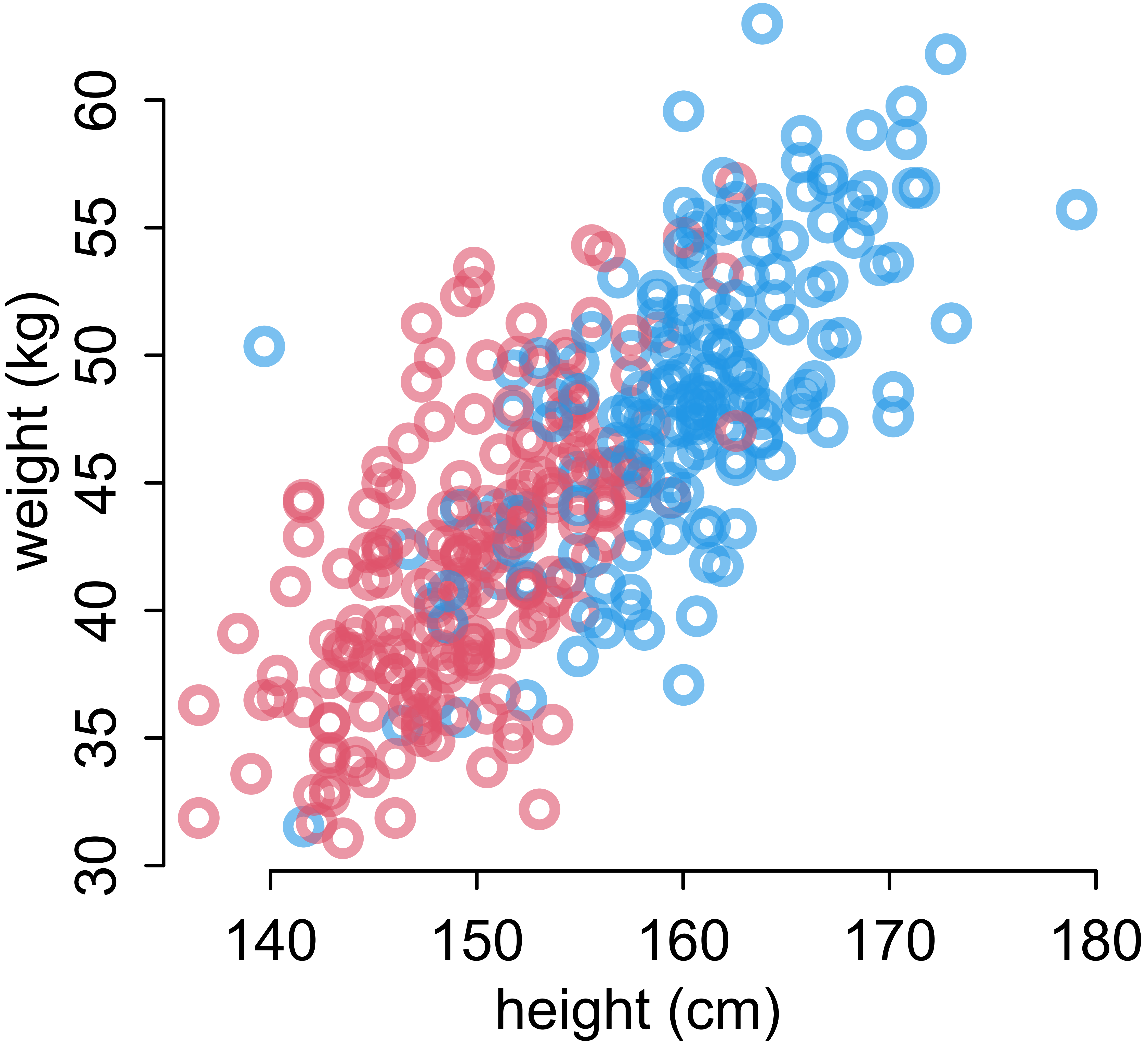
How are height, weight, and sex **statistically** related?

$$\begin{tikzcd} H \arrow[r] & W \\ S \arrow[u] \arrow[ur] & \end{tikzcd}$$

The causes aren't in the data

$$H \longrightarrow W$$

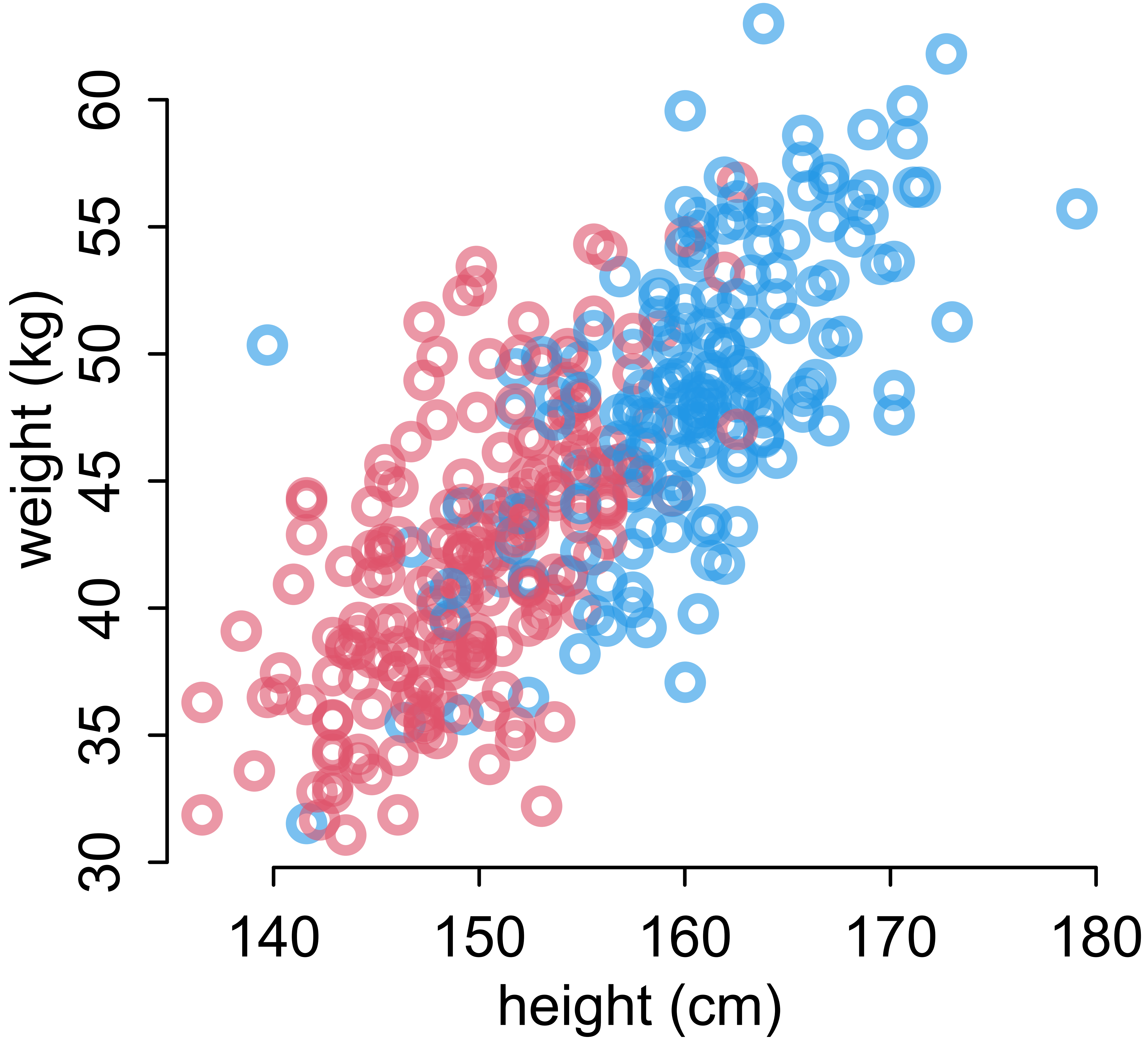
$$H \longleftarrow W$$



The causes aren't in the data

$H \longrightarrow W$

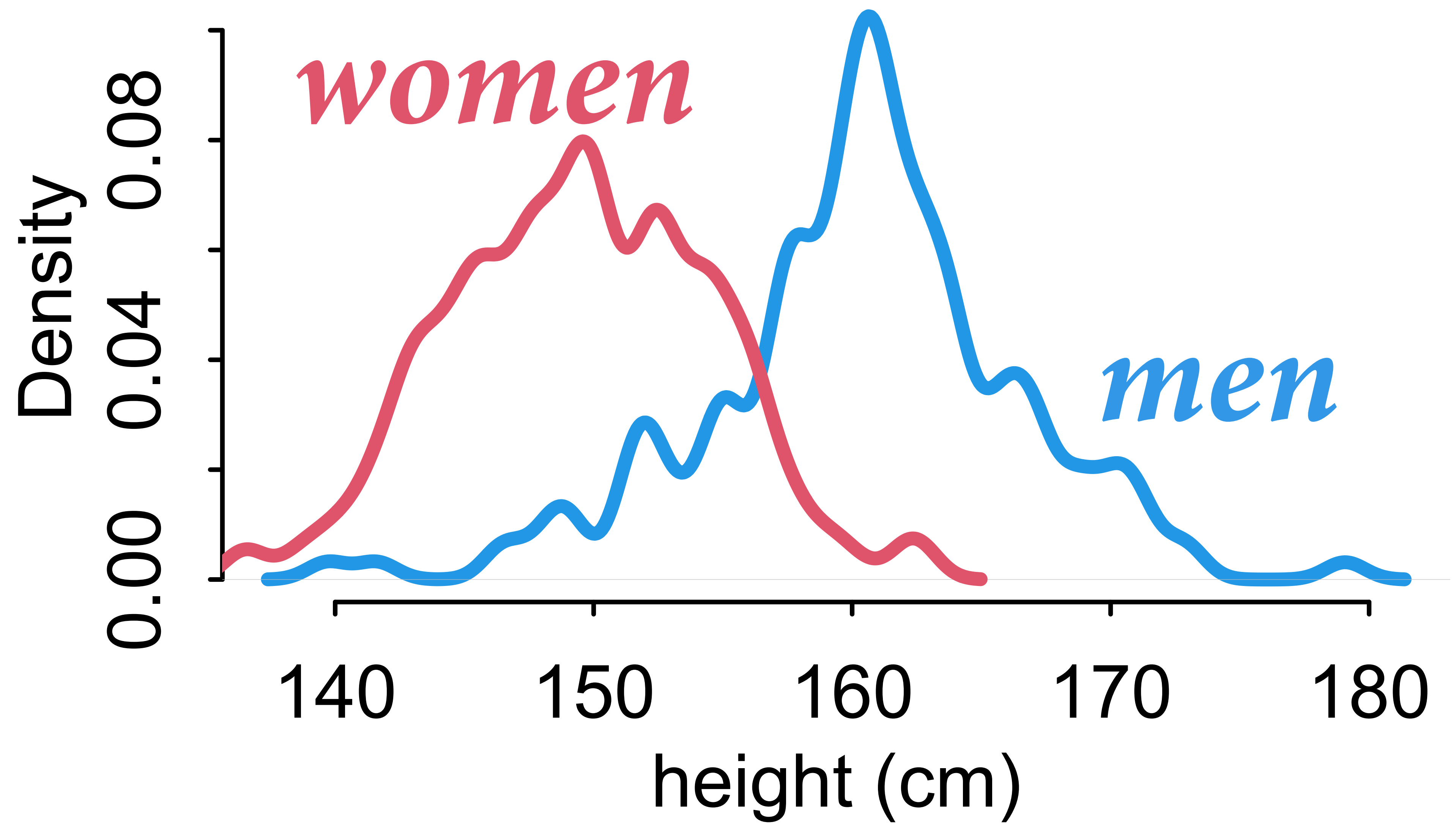
$H \longleftarrow W$



The causes aren't in the data

$$H \longrightarrow S$$

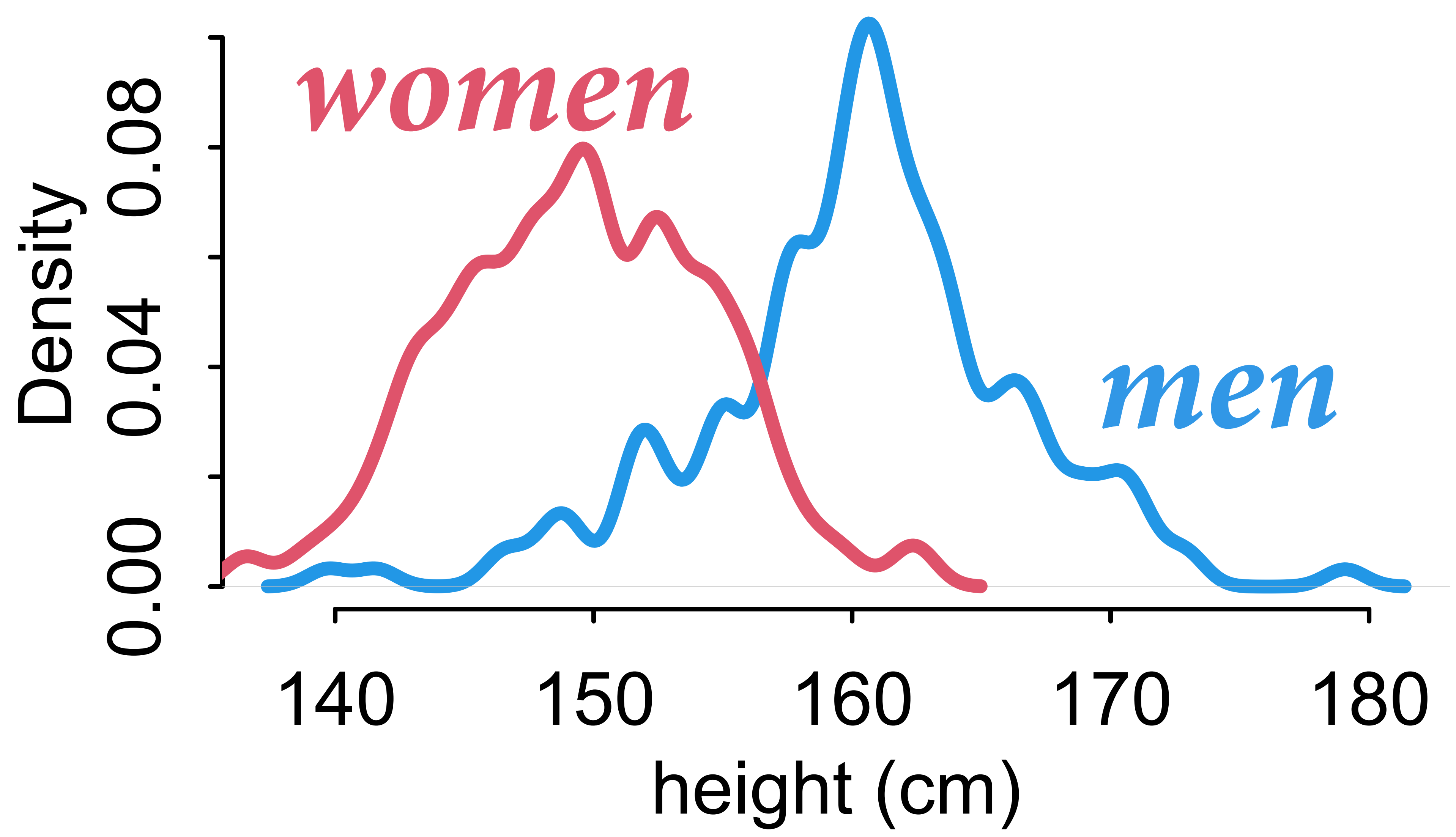
$$H \longleftarrow S$$



The causes aren't in the data

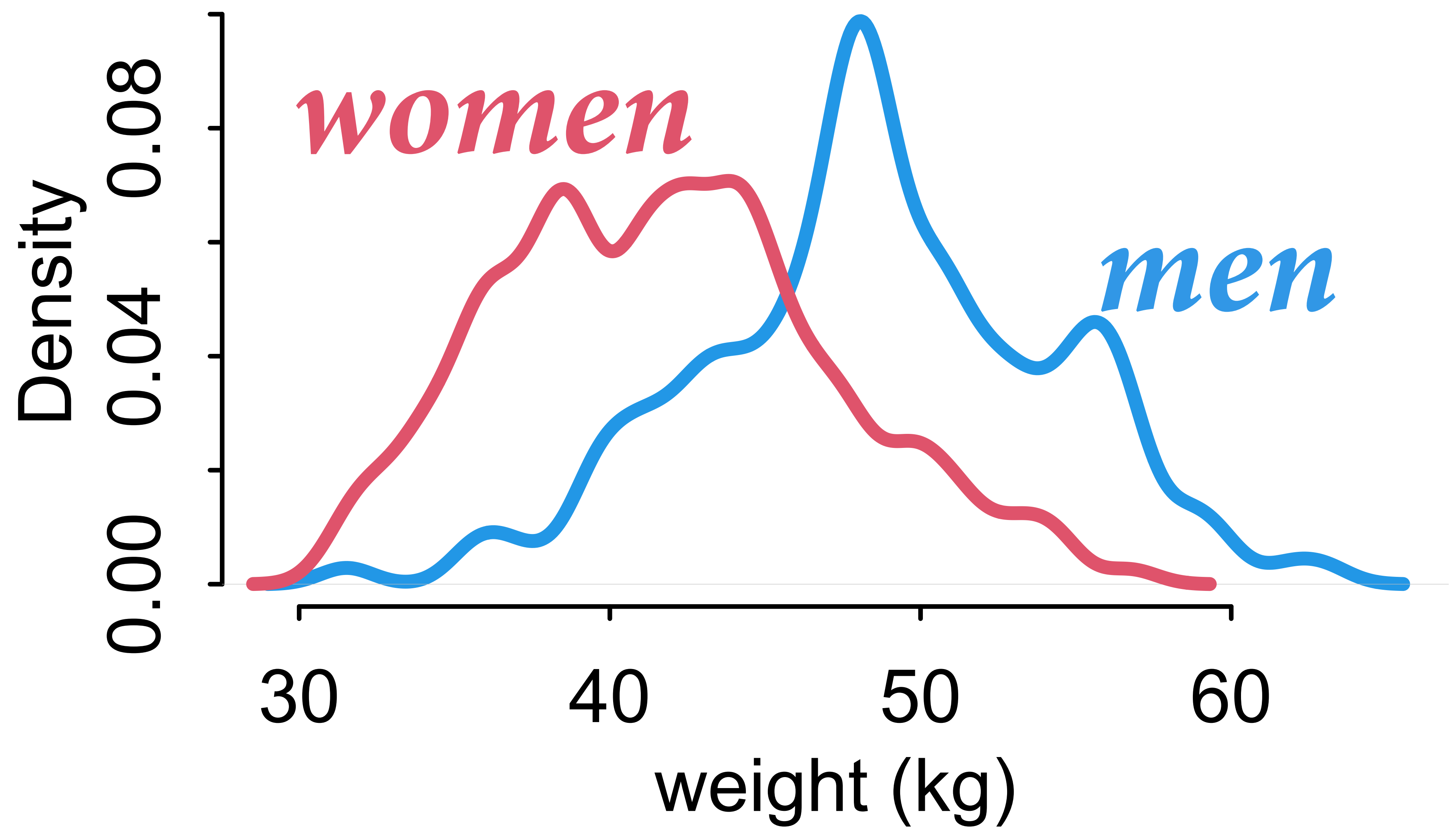
$$H \longrightarrow S$$

$$H \longleftarrow S$$



$$W \longrightarrow S$$

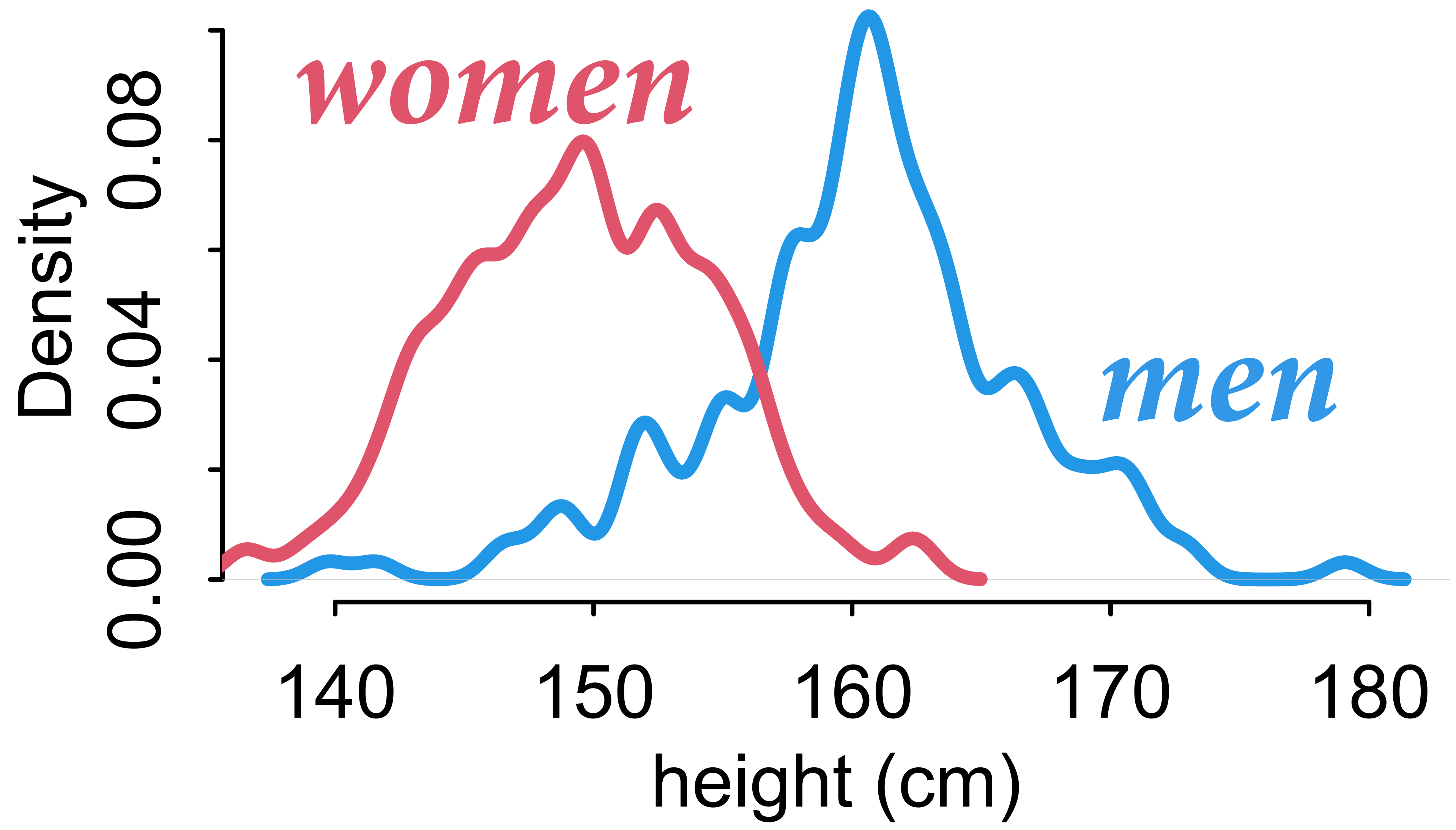
$$W \longleftarrow S$$



The causes aren't in the data

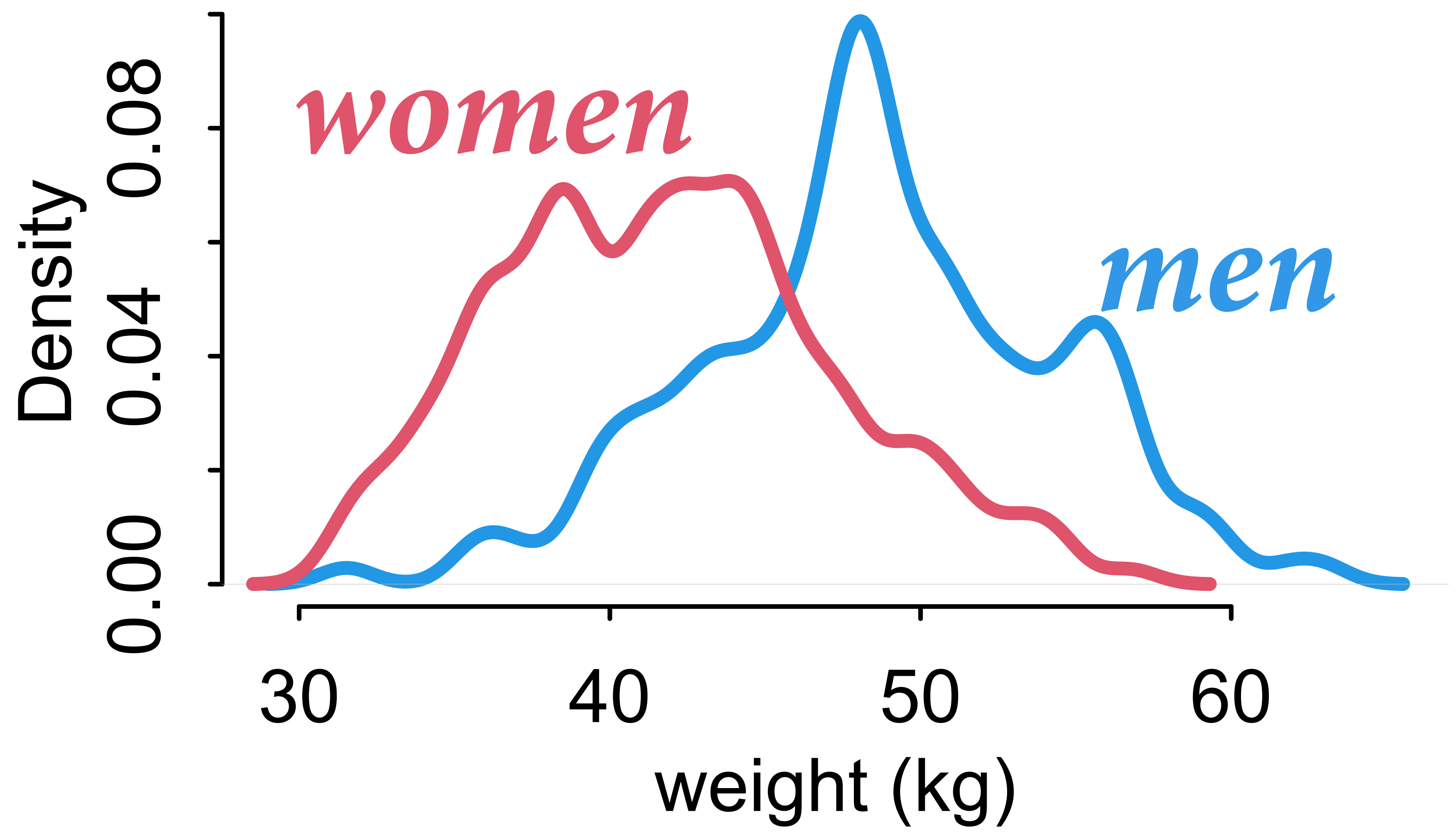
$$H \longrightarrow S$$

$$H \longleftarrow S$$



$$W \longrightarrow S$$

$$W \longleftarrow S$$



height influences weight

$$\begin{tikzcd} H \arrow[r] & W \\ S \arrow[u] \arrow[ur] & \end{tikzcd}$$

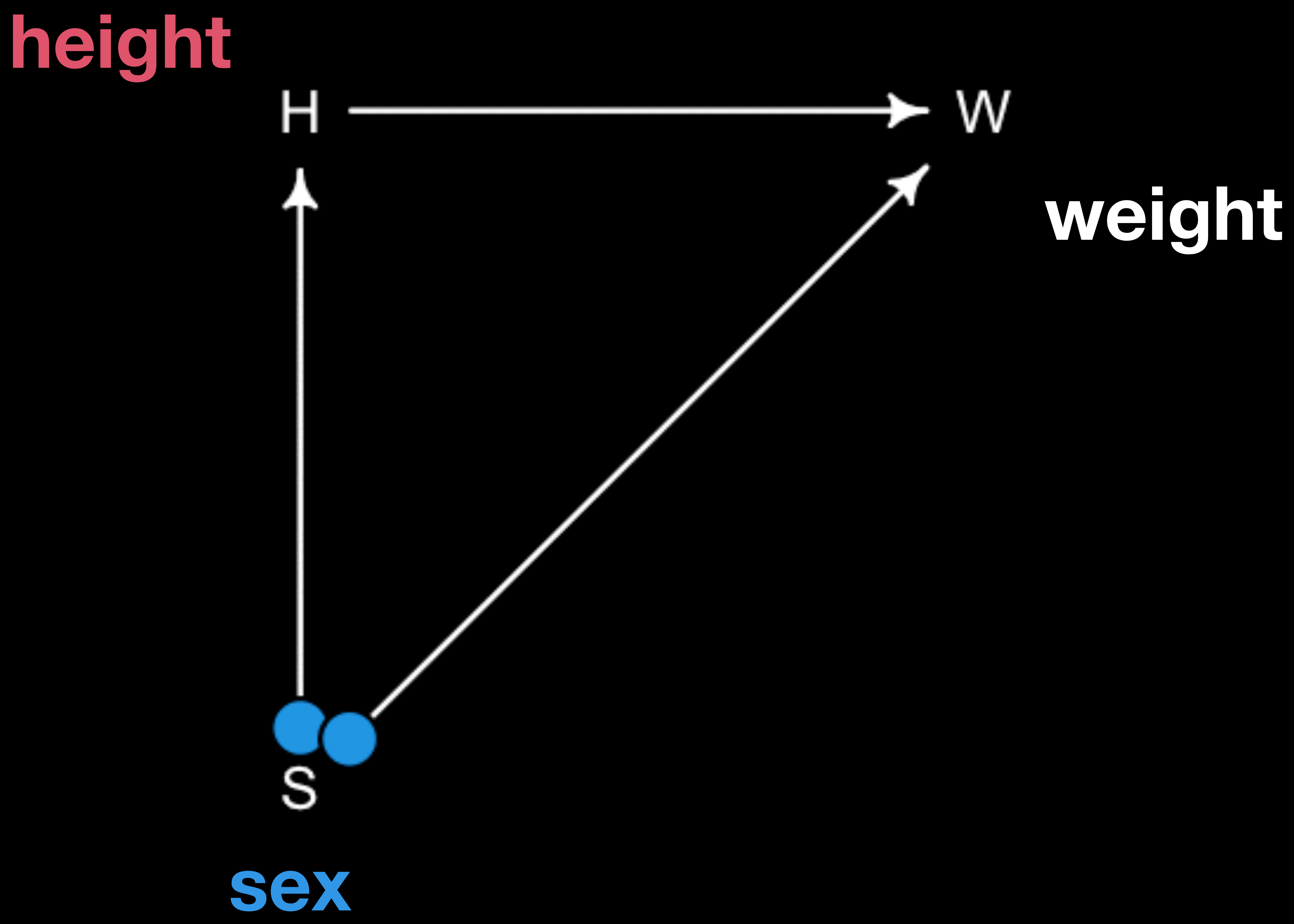
weight is influenced by both height & sex

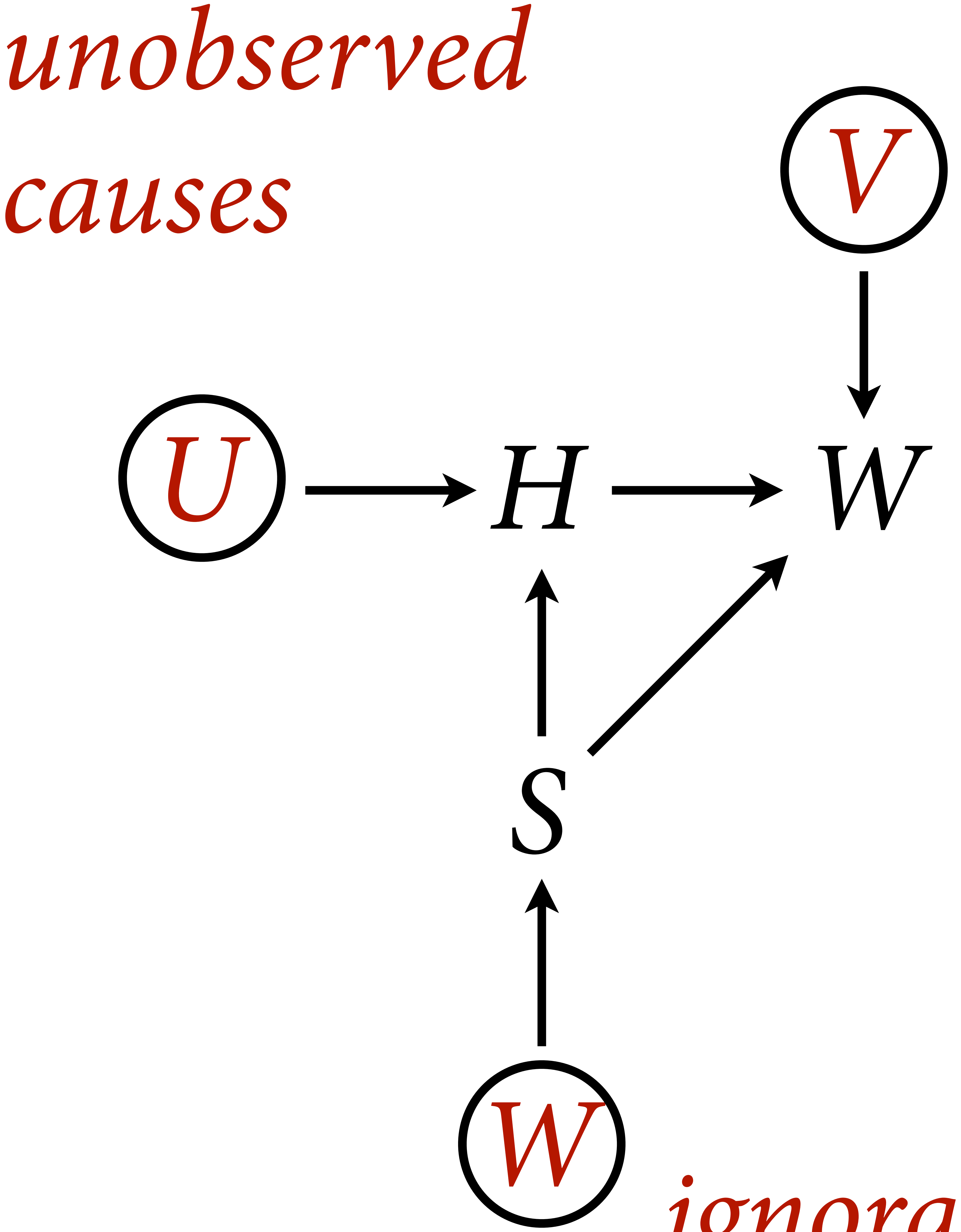
sex influences both height & weight

$$
\begin{tikzcd}
H \arrow[r] & W \\
S \arrow[u] \arrow[ur] &
\end{tikzcd}
$$

$$H = f_H(S)$$

$$W = f_W(H, S)$$


height
H
W
weight
S
sex

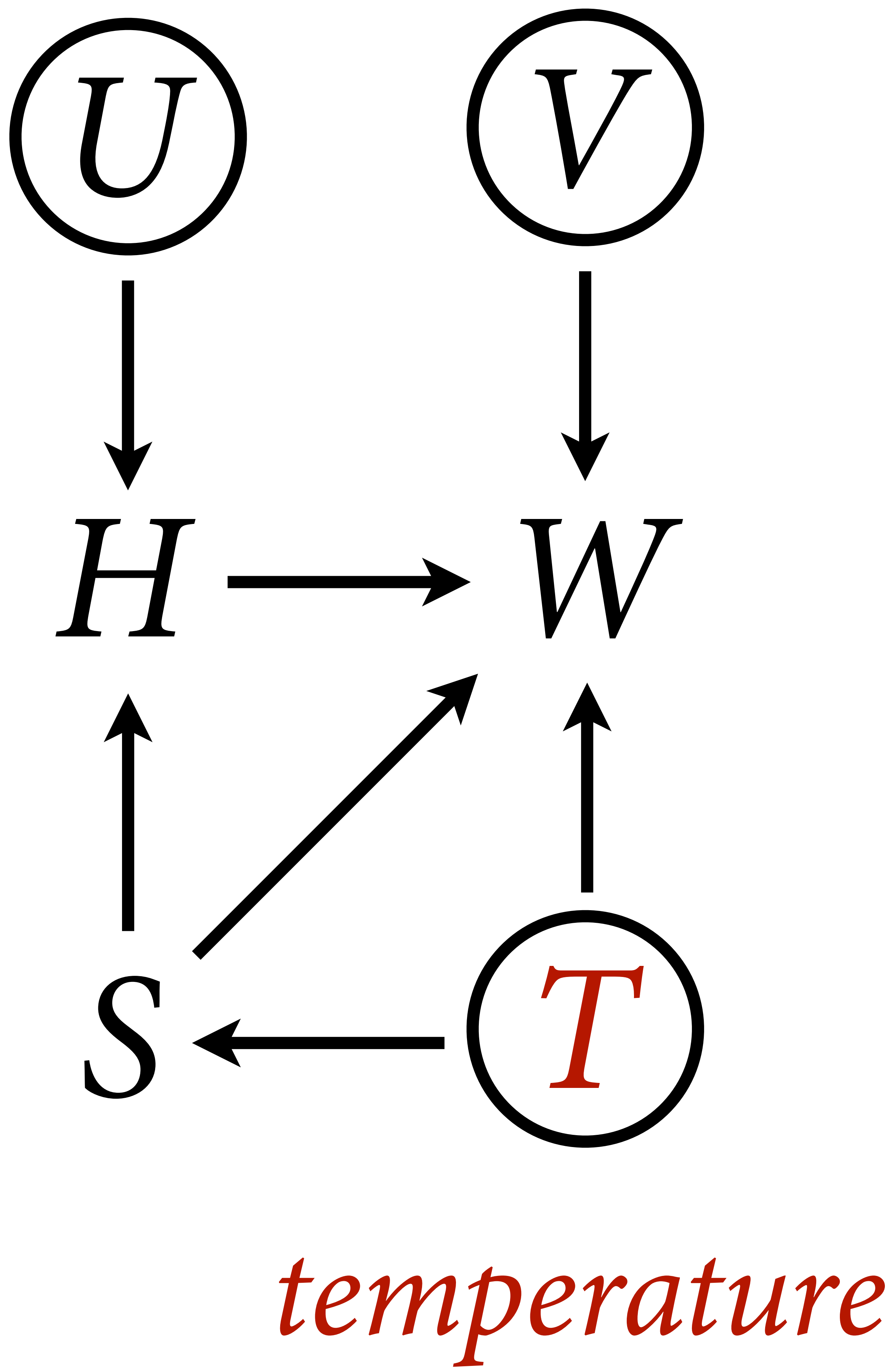


ignorable unless shared

$$H = f_H(S, U)$$

$$W = f_W(H, S, V)$$

$$S = f_S(W)$$



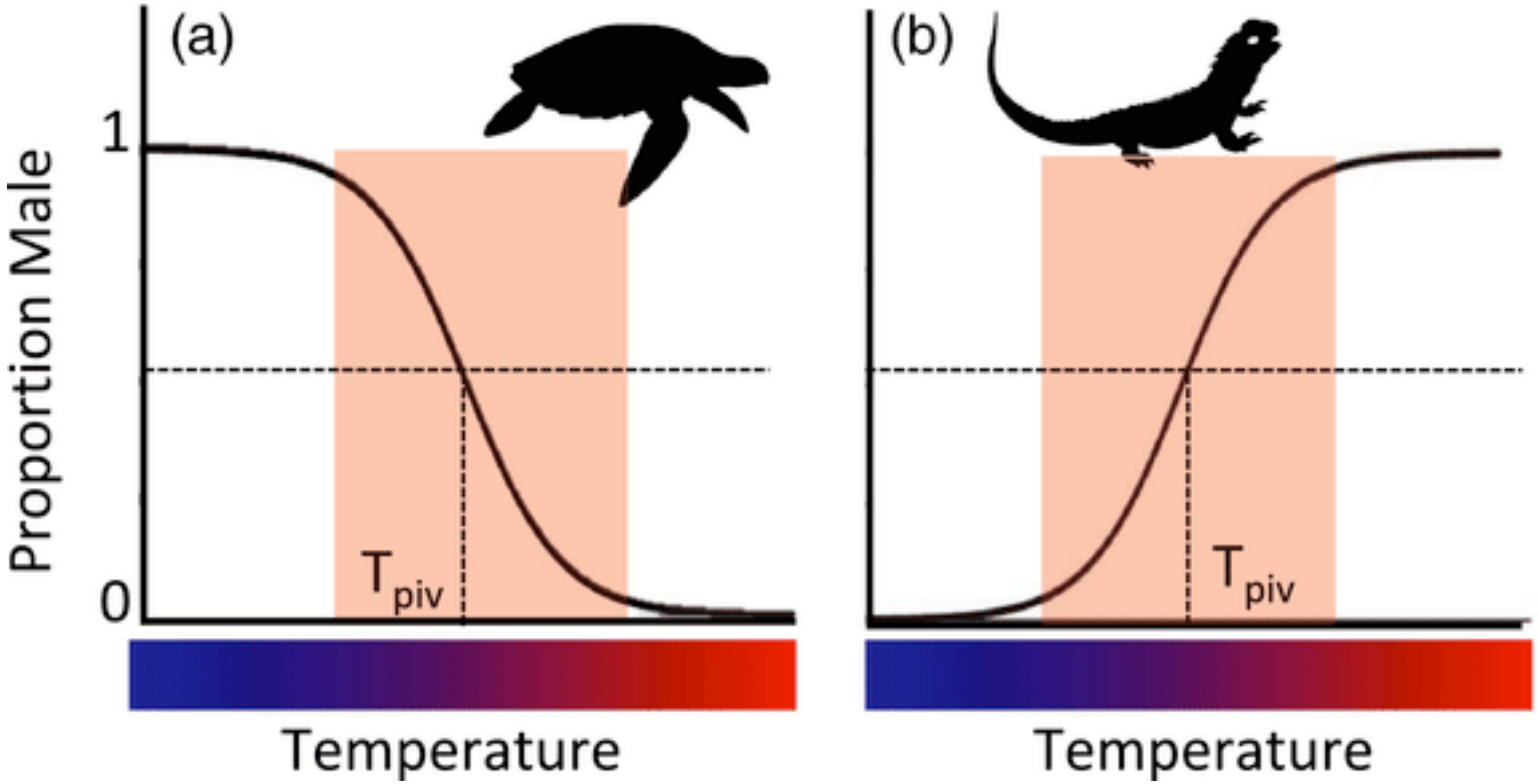
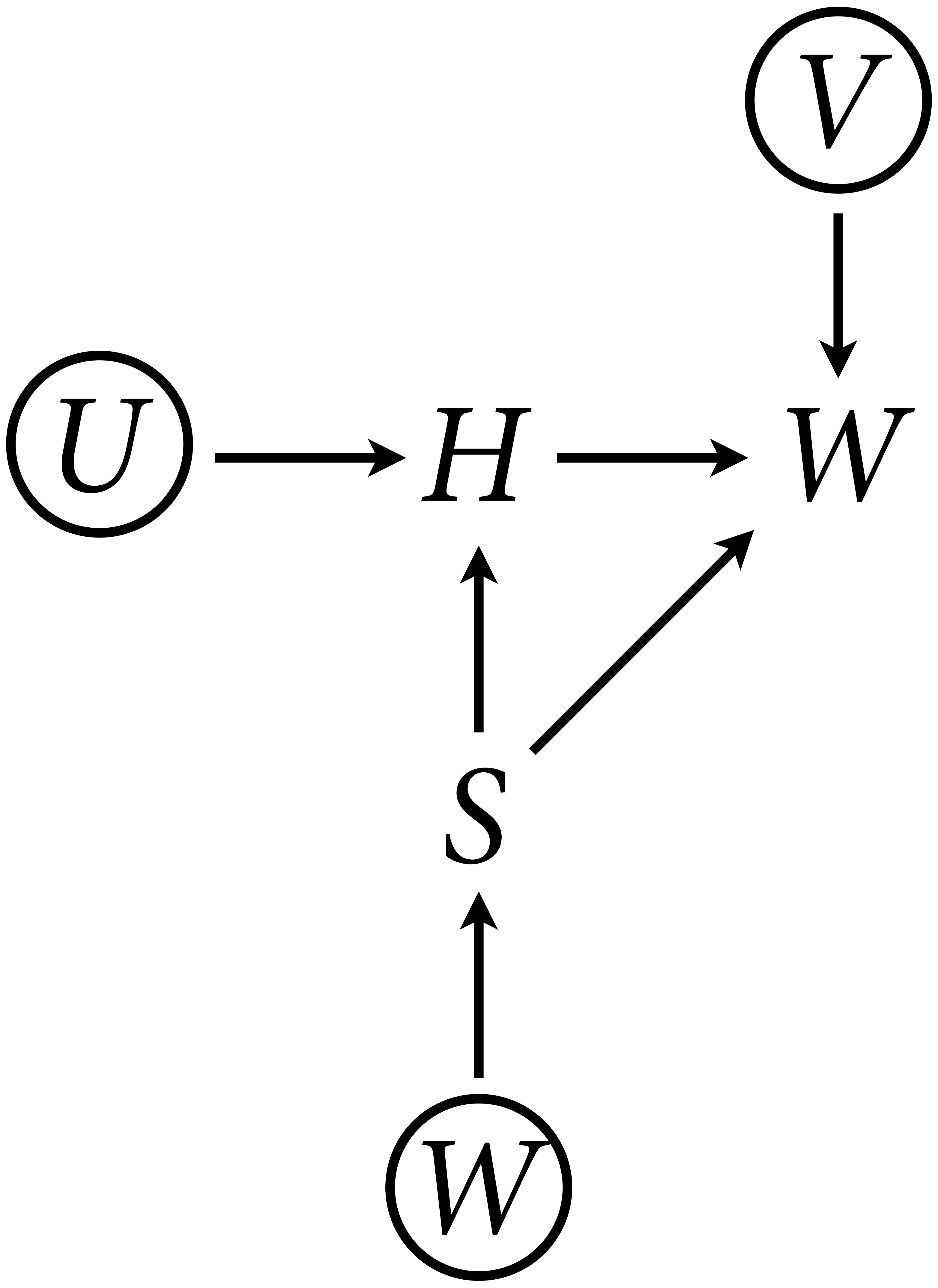


Figure: Lockley & Eizaguirre 2021



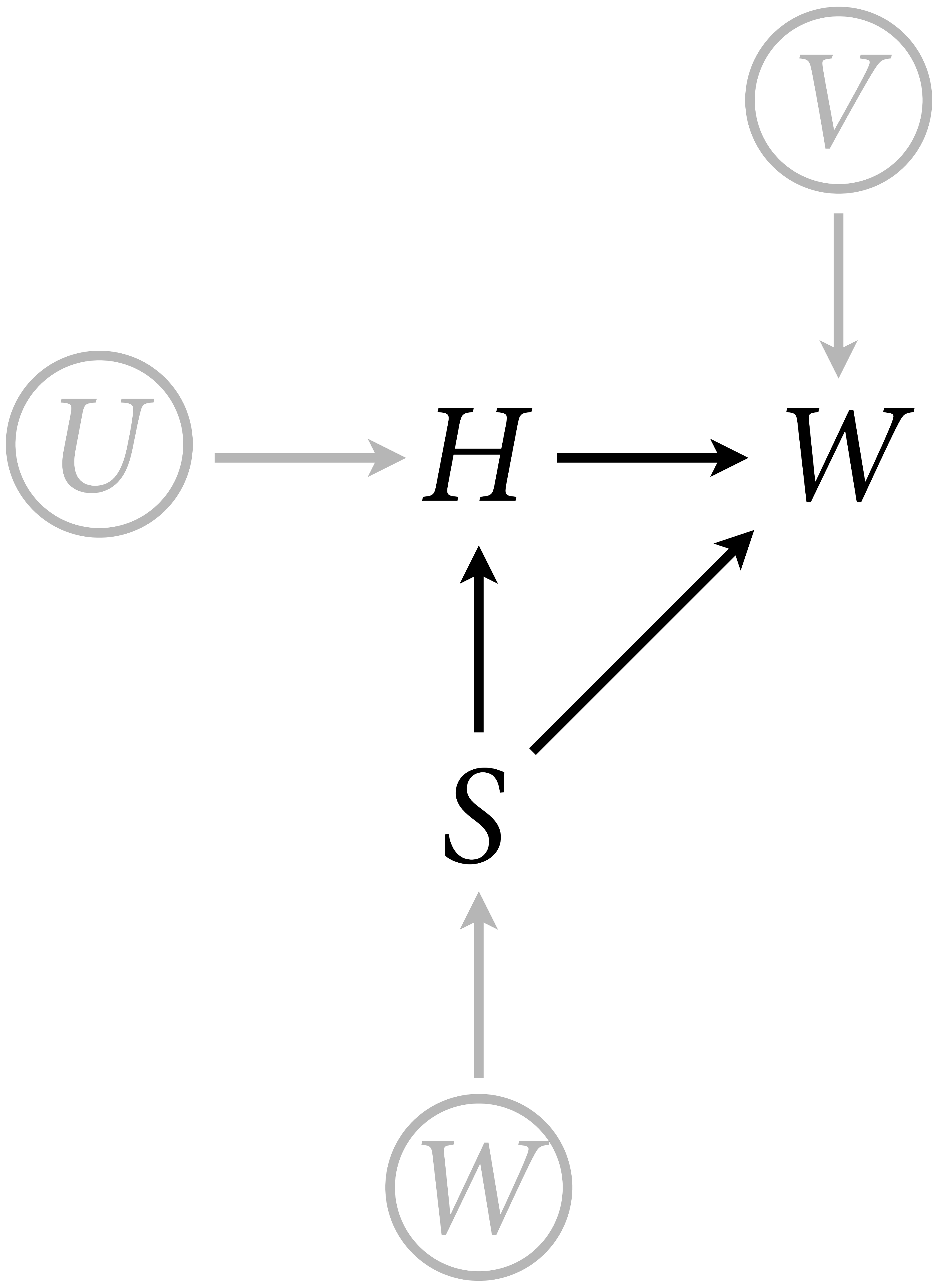
$$H = f_H(S, U)$$

$$W = f_W(H, S, V)$$

$$S = f_S(W)$$

Synthetic people

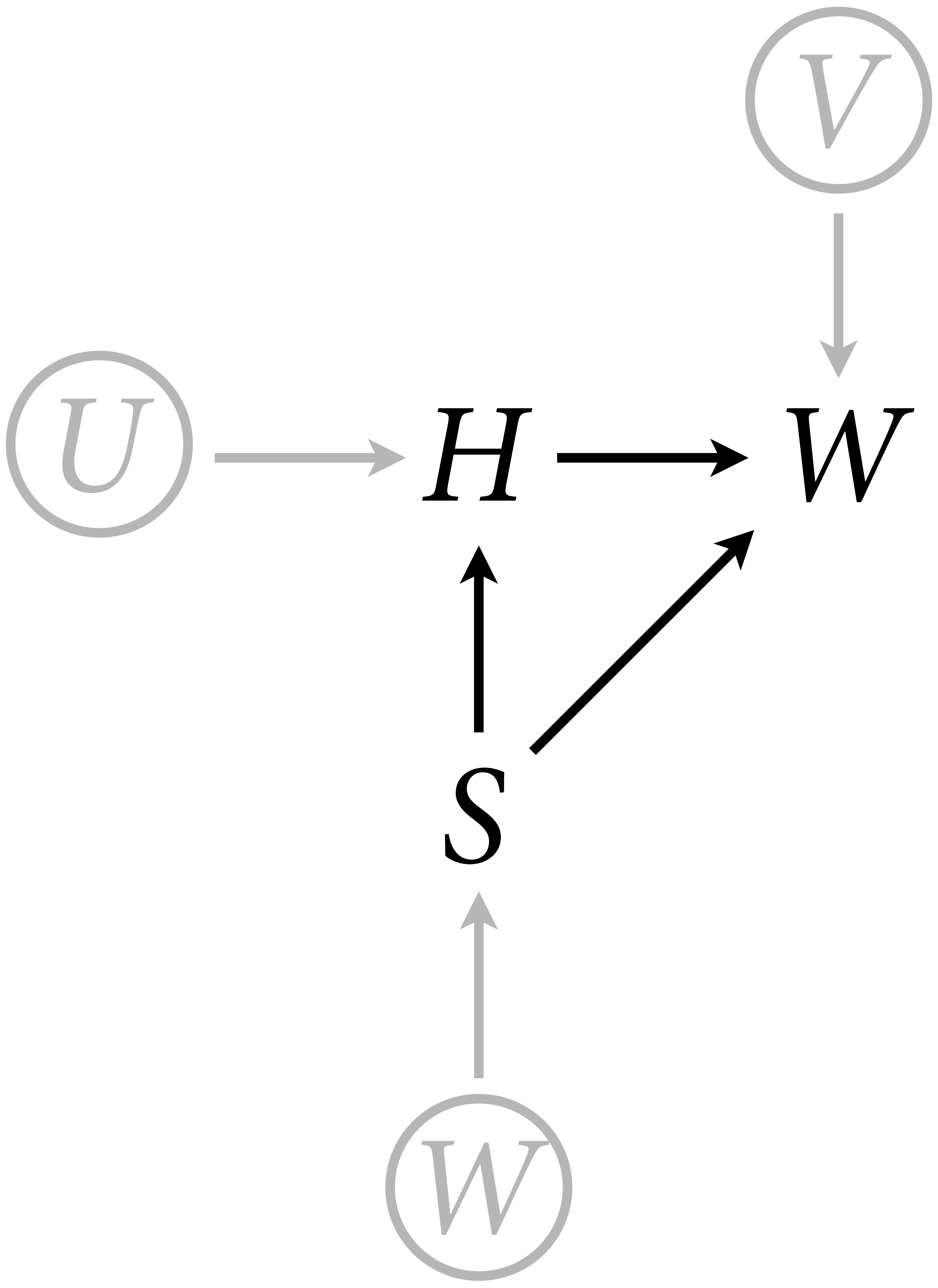
```
# S=1 female; S=2 male
sim_HW <- function(S,b,a) {
    N <- length(S)
    H <- ifelse(S==1,150,160) + rnorm(N,0,5)
    W <- a[S] + b[S]*H + rnorm(N,0,5)
    data.frame(S,H,W)
}
```



Synthetic people

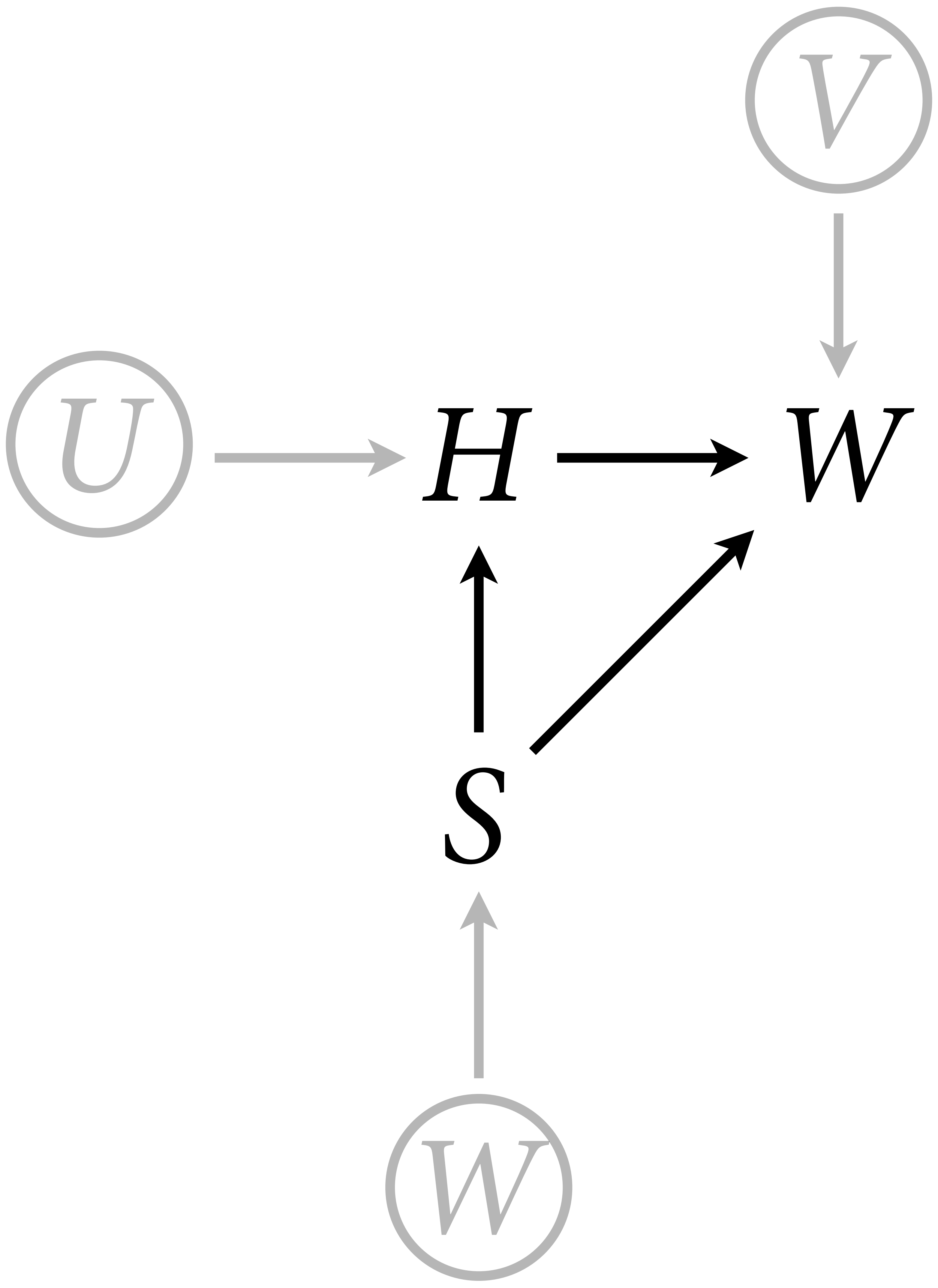
```
# S=1 female; S=2 male
sim_HW <- function(S,b,a) {
    N <- length(S)
    H <- ifelse(S==1,150,160) + rnorm(N,0,5)
    W <- a[S] + b[S]*H + rnorm(N,0,5)
    data.frame(S,H,W)
}
```

for each S, different height-weight line



```r
# S=1 female; S=2 male
sim_HW <- function(S,b,a) {
    N <- length(S)
    H <- ifelse(S==1,150,160) + rnorm(N,0,5)
    W <- a[S] + b[S]*H + rnorm(N,0,5)
    data.frame(S,H,W)
}

S <- rbern(100)+1
dat <- sim_HW(S,b=c(0.5,0.6),a=c(0,0))
head(dat)
```

```
  S        H        W
1 1 149.2428 80.07163
2 2 164.8629 94.79647
3 2 158.1800 90.59944
4 1 148.2153 70.68285
5 2 155.5794 91.43798
6 2 154.0733 92.18237
```



Think scientifically first

Different causal questions may need different statistical models

Q: Causal effect of H on W?

Q: Causal effect of S on W?

Q: Direct causal effect of S on W?

$$\begin{tikzcd} H \arrow[r] & W \\ S \arrow[u] \arrow[ur] & \end{tikzcd}$$

Think scientifically first

Different causal questions need different statistical models

Q: Causal effect of H on W?

Q: Causal effect of S on W?

Q: Direct causal effect of S on W?

$$\begin{tikzcd} H \arrow[r] & W \\ S \arrow[u] \arrow[ur] & \end{tikzcd}$$

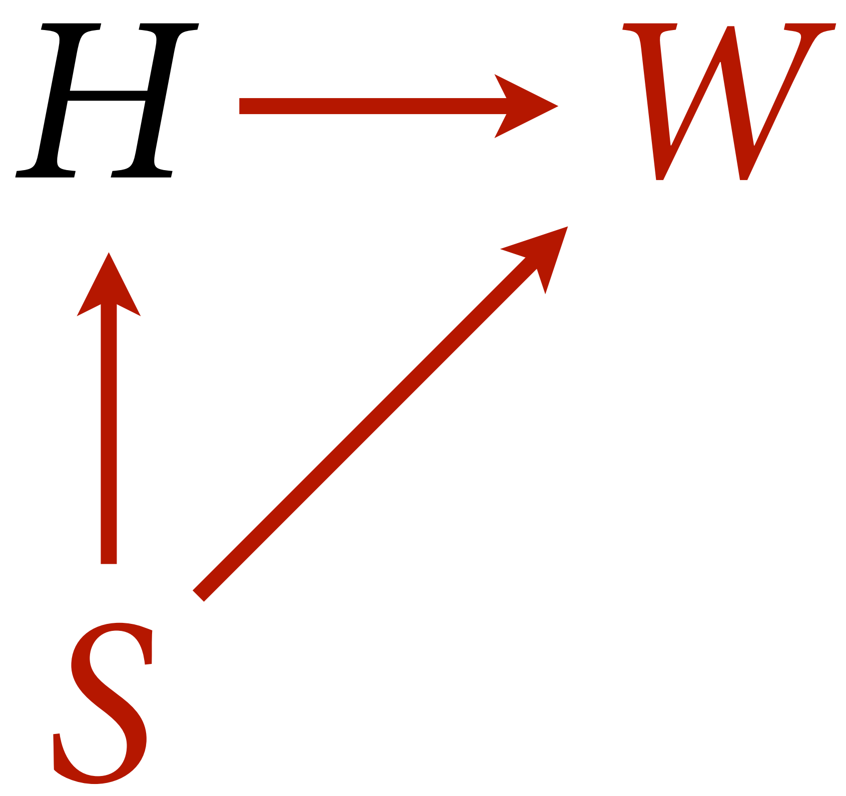
Think scientifically first

Different causal questions need different statistical models

Q: Causal effect of H on W?

Q: Causal effect of S on W?

Q: Direct causal effect of S on W?



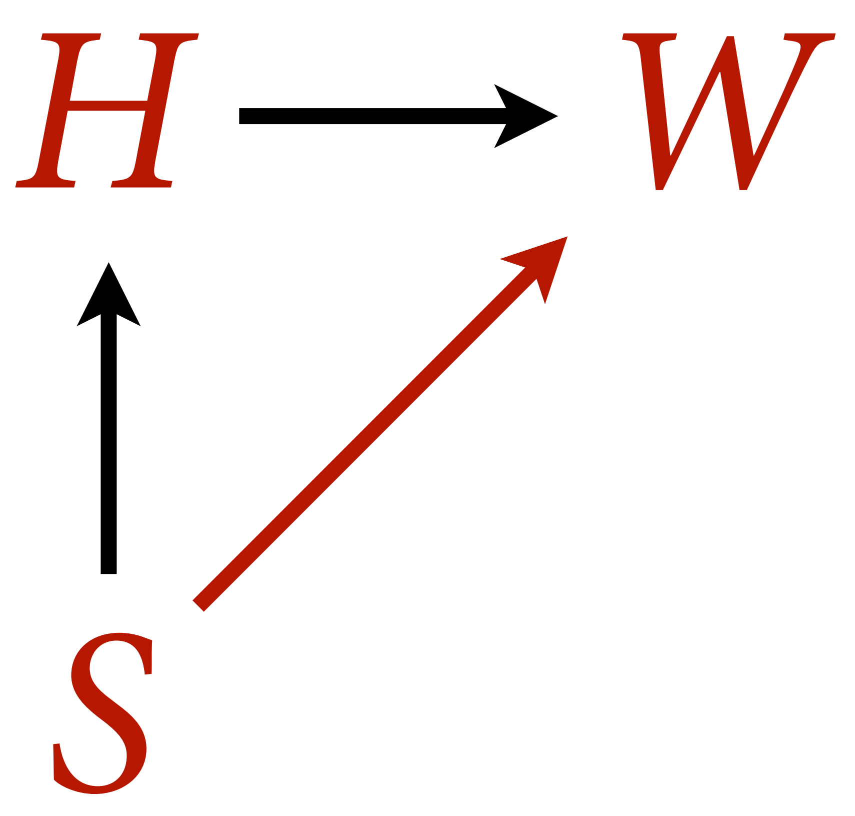
Think scientifically first

Different causal questions need different statistical models

Q: Causal effect of H on W?

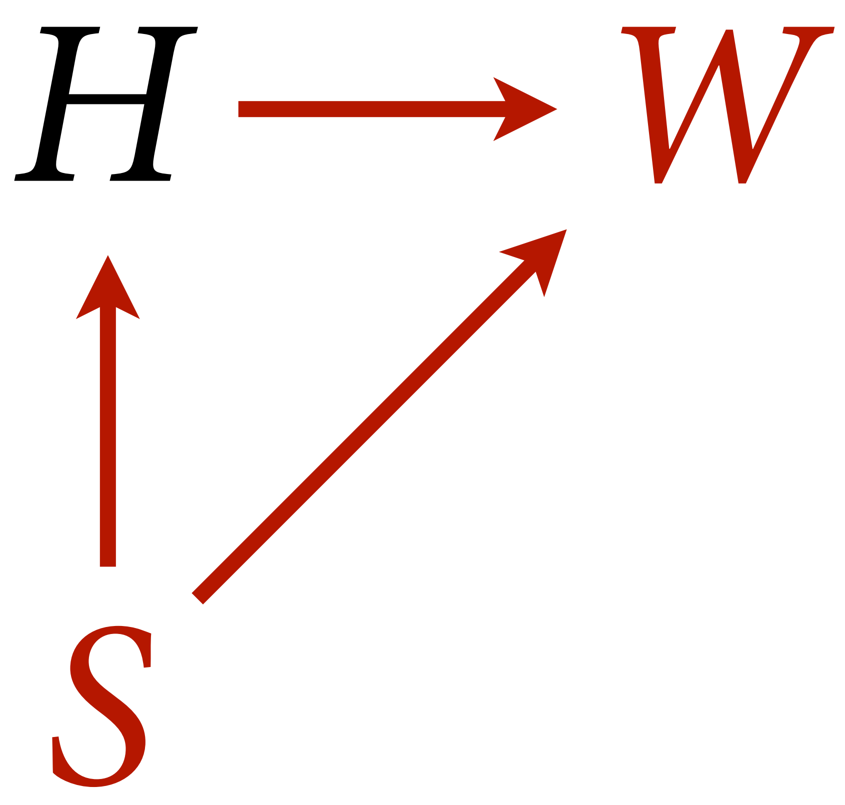
Q: Causal effect of S on W?

Q: Direct causal effect of S on W?

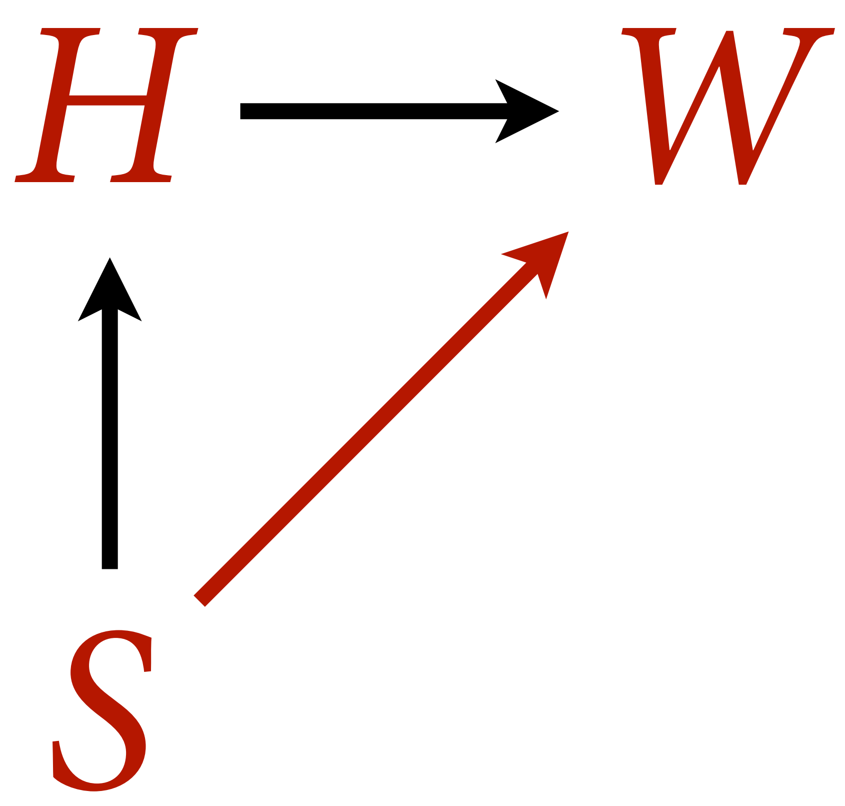


From estimand to estimate

Q: Causal effect of S on W?



Q: Direct causal effect of S on W?



Stratify by S: different estimate for each value S can take

Drawing Categorical Owls

Several ways to code categorical variables

(1) indicator (0/1) variables

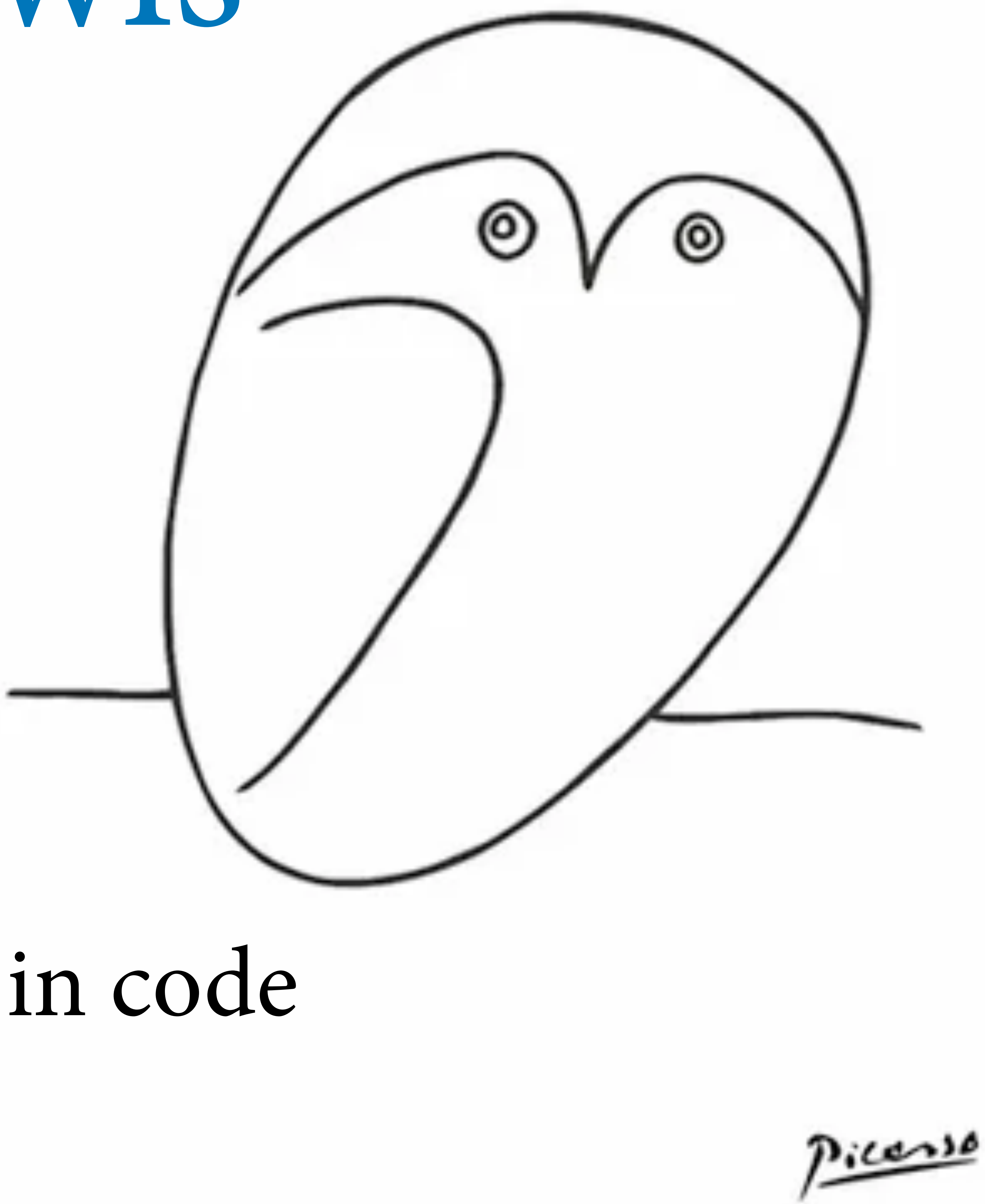
(2) index variables: 1,2,3,4,…

We will use index variables:

Extend to many categories with no change in code

Better for specifying priors

Extend effortlessly to multi-level models



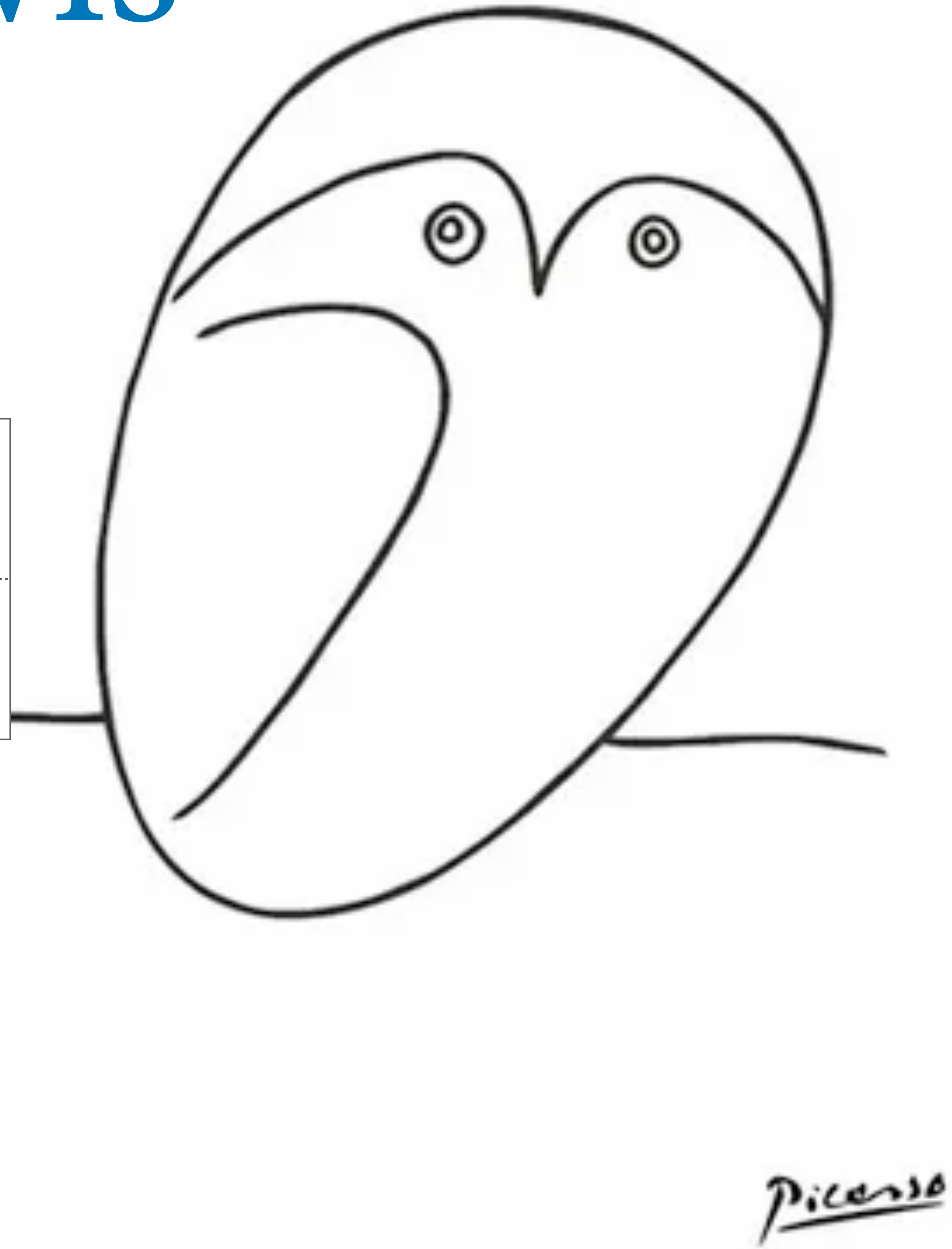
Drawing Categorical Owls

Example:

Category	Cyan	Magenta	Yellow	Black
Index Value	1	2	3	4

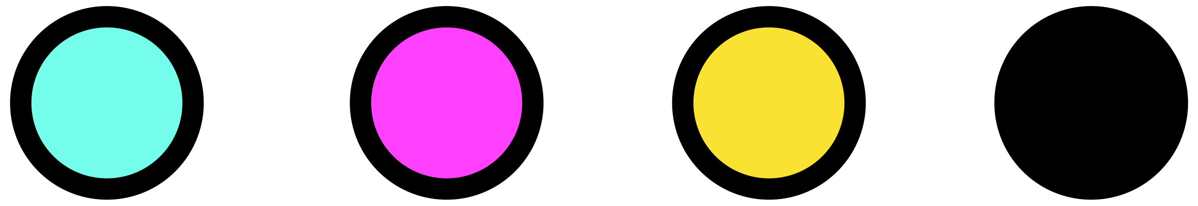
Influence of color coded by:

$$\alpha = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]$$



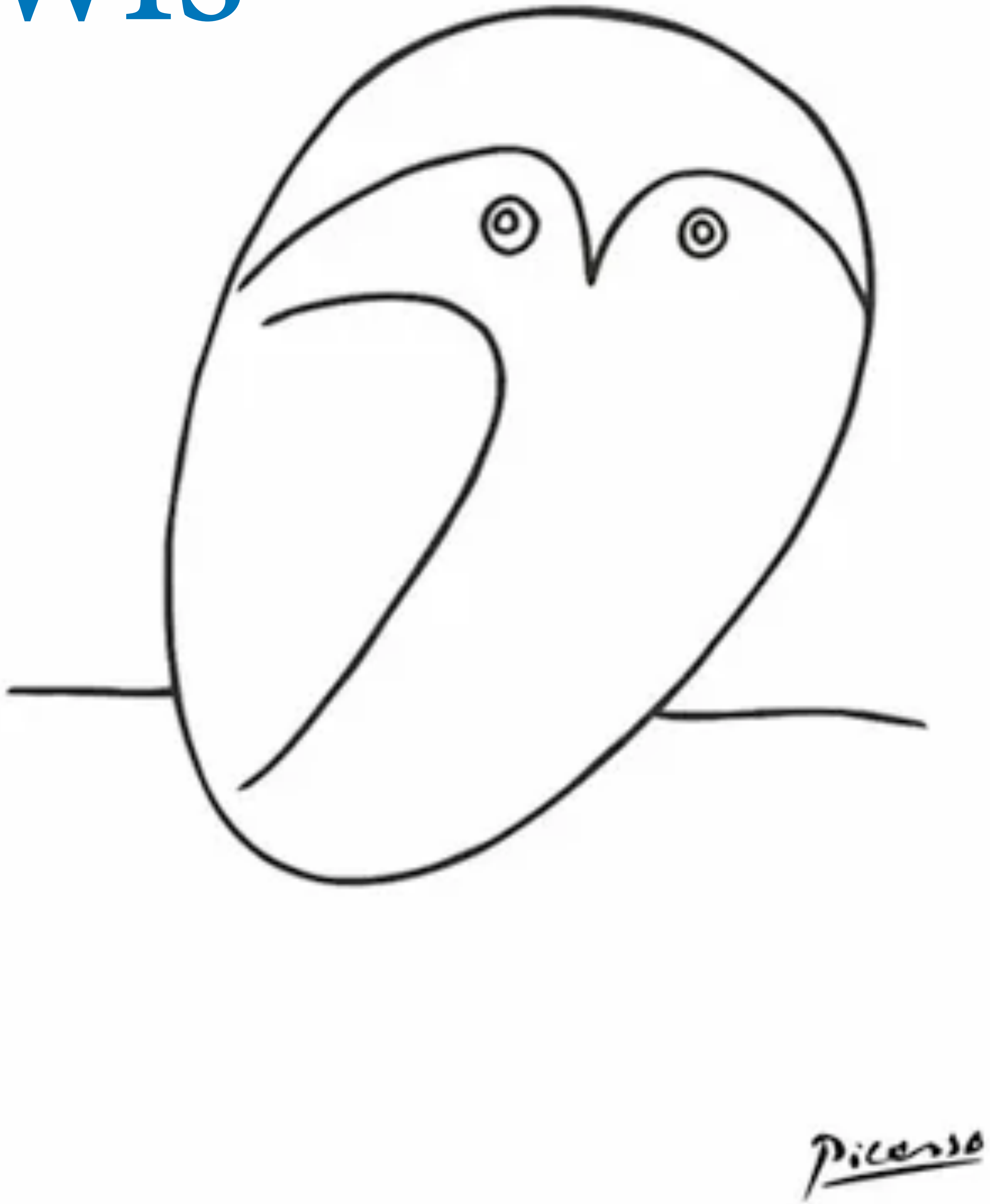
Drawing Categorical Owls

$$\alpha = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]$$



$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{\text{COLOR}[i]}$$



Using Index Variables

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha$$

intercept

Using Index Variables

	H	W	S
1	152	48	2
2	140	36	1
3	137	32	1
4	157	53	2
5	145	41	1
6	164	63	2
7	149	38	1
8	169	55	2
9	148	35	1
10	165	54	2
11	154	50	1
12	151	41	2

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{S[i]}$$

sex of i-th person

S = 1 indicates female
S = 2 indicates male

$$\alpha = [\alpha_1, \alpha_2]$$

Two intercepts, one for each value in S

Using Index Variables

	H	W	S
1	152	48	2
2	140	36	1
3	137	32	1
4	157	53	2
5	145	41	1
6	164	63	2
7	149	38	1
8	169	55	2
9	148	35	1
10	165	54	2
11	154	50	1
12	151	41	2

i = *1*

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{\boxed{S[i]}}$$

S[1]=2

$$\alpha = [\alpha_1, \boxed{\alpha_2}]$$

Using Index Variables

	H	W	S
1	152	48	2
2	140	36	1
3	137	32	1
4	157	53	2
5	145	41	1
6	164	63	2
7	149	38	1
8	169	55	2
9	148	35	1
10	165	54	2
11	154	50	1
12	151	41	2

$i = 2$

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{\boxed{S[i]}}$$

$S[2]=1$

$$\alpha = [\boxed{\alpha_1}, \alpha_2]$$

Using Index Variables

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{S[i]}$$

Priors

$$\alpha = [\alpha_1, \alpha_2] \qquad \alpha_j \sim \text{Normal}(60, 10)$$

Testing

What is the total causal effect of sex? Causal effect of sex is difference made intervening:

```
# female sample
S <- rep(1,100)
simF <- sim_HW(S,b=c(0.5,0.6),a=c(0,0))

# male sample
S <- rep(2,100)
simM <- sim_HW(S,b=c(0.5,0.6),a=c(0,0))

# effect of sex (male-female)
mean(simM$W - simF$W)
```

```
[1] 21.13852
```

$$
\begin{aligned}
W_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha_{S[i]} \\
\alpha_j &\sim \text{Normal}(60, 10) \\
\sigma &\sim \text{Uniform}(0, 10)
\end{aligned}
$$

Now run the estimating model and synthetic sample

```
# observe sample
S <- rbern(100)+1
dat <- sim_HW(S,b=c(0.5,0.6),a=c(0,0))

# estimate posterior
m_SW <- quap(
    alist(
        W ~ dnorm(mu,sigma),
        mu <- a[S],
        a[S] ~ dnorm(60,10),
        sigma ~ dunif(0,10)
    ), data=dat )
precis(m_SW,depth=2)
```

```
a[1]  75.47 0.93 73.98 76.95
a[2]  97.39 0.84 96.05 98.74
sigma  6.26 0.44  5.55  6.96
```

97 – 75 = 22

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha_{S[i]}$$
$$\alpha_j \sim \text{Normal}(60, 10)$$
$$\sigma \sim \text{Uniform}(0, 10)$$

Analyze the real sample

```
data(Howell1)
d <- Howell1
d <- d[ d$age>=18 , ]
dat <- list(
    W = d$weight,
    S = d$male + 1 ) # S=1 female, S=2 male

m_SW <- quap(
    alist(
        W ~ dnorm(mu,sigma),
        mu <- a[S],
        a[S] ~ dnorm(60,10),
        sigma ~ dunif(0,10)
    ), data=dat )
```

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

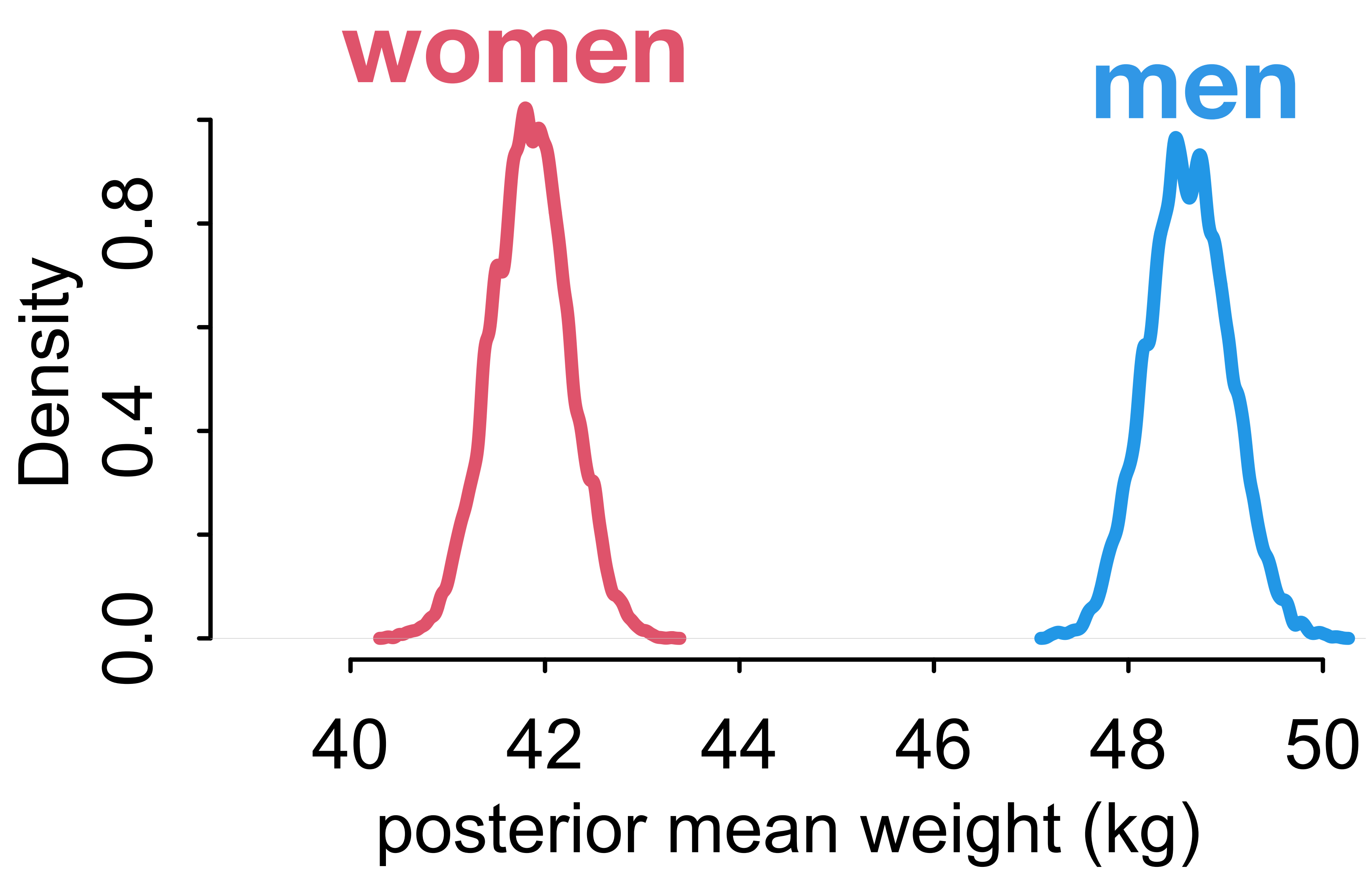
$$\mu_i = \alpha_{S[i]}$$

$$\alpha_j \sim \text{Normal}(60, 10)$$

$$\sigma \sim \text{Uniform}(0, 10)$$

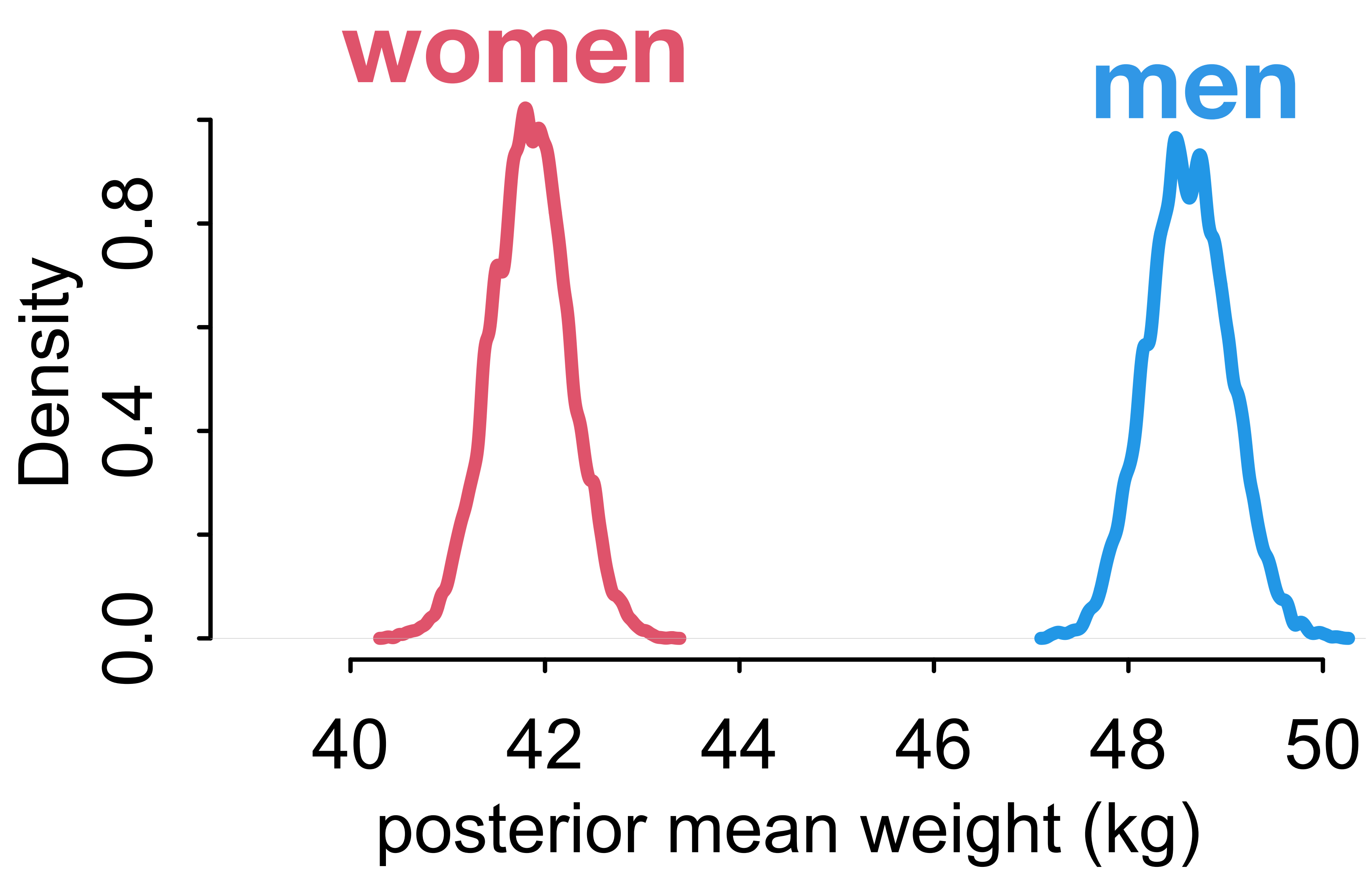
Posterior means & predictions

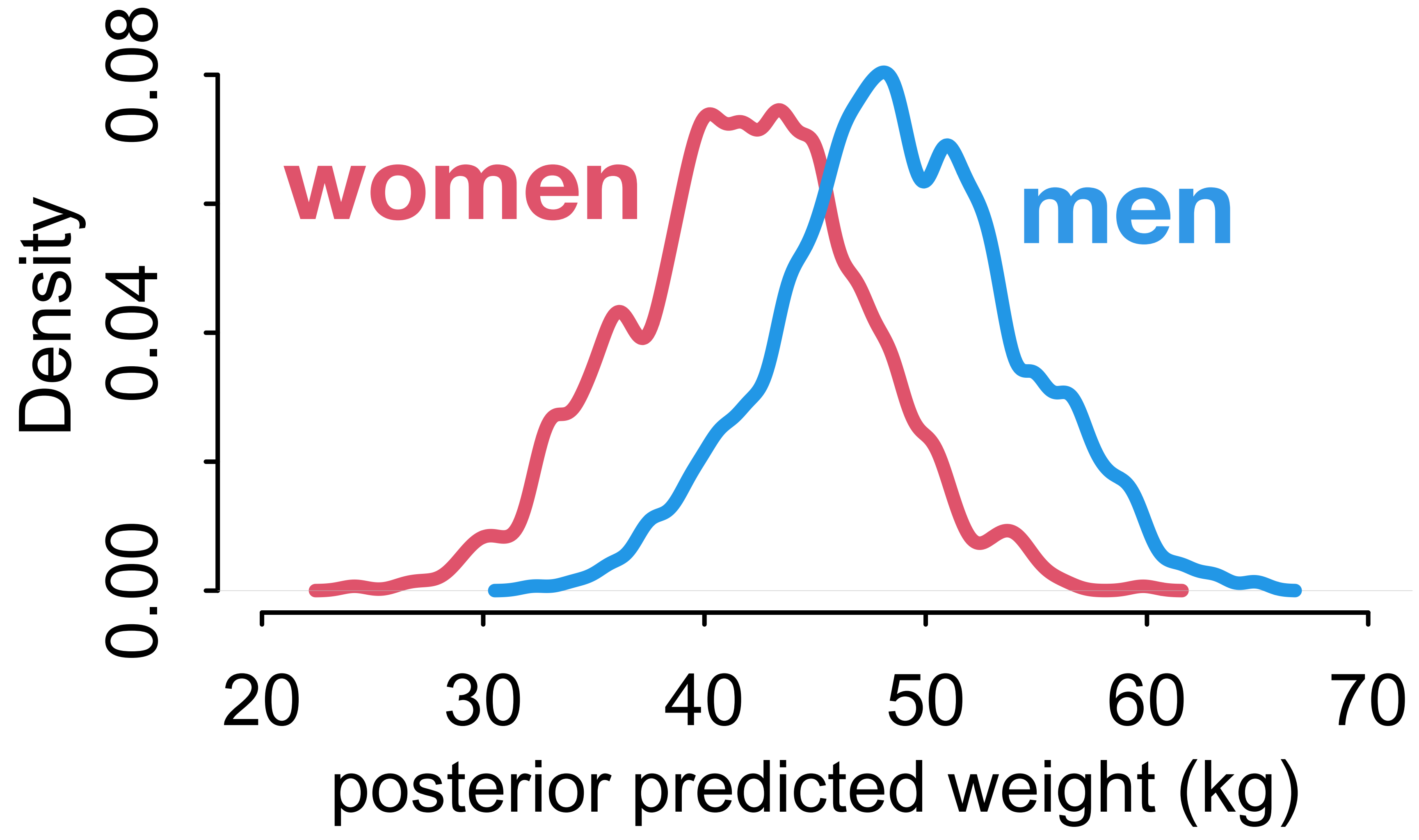
```
# posterior mean W
post <- extract.samples(m_SW)
dens( post$a[,1] , xlim=c(39,50) , lwd=3 ,
col=2 , xlab="posterior mean weight (kg)" )
dens( post$a[,2] , lwd=3 , col=4 , add=TRUE )
```



Posterior means & predictions

```
# posterior mean W
post <- extract.samples(m_SW)
dens( post$a[,1] , xlim=c(39,50) , lwd=3 ,
col=2 , xlab="posterior mean weight (kg)" )
dens( post$a[,2] , lwd=3 , col=4 , add=TRUE )

# posterior W distributions
W1 <- rnorm( 1000 , post$a[,1] , post$sigma )
W2 <- rnorm( 1000 , post$a[,2] , post$sigma )
dens( W1 , xlim=c(20,70) , ylim=c(0,0.085) ,
lwd=3 , col=2 )
dens( W2 , lwd=3 , col=4 , add=TRUE )
```



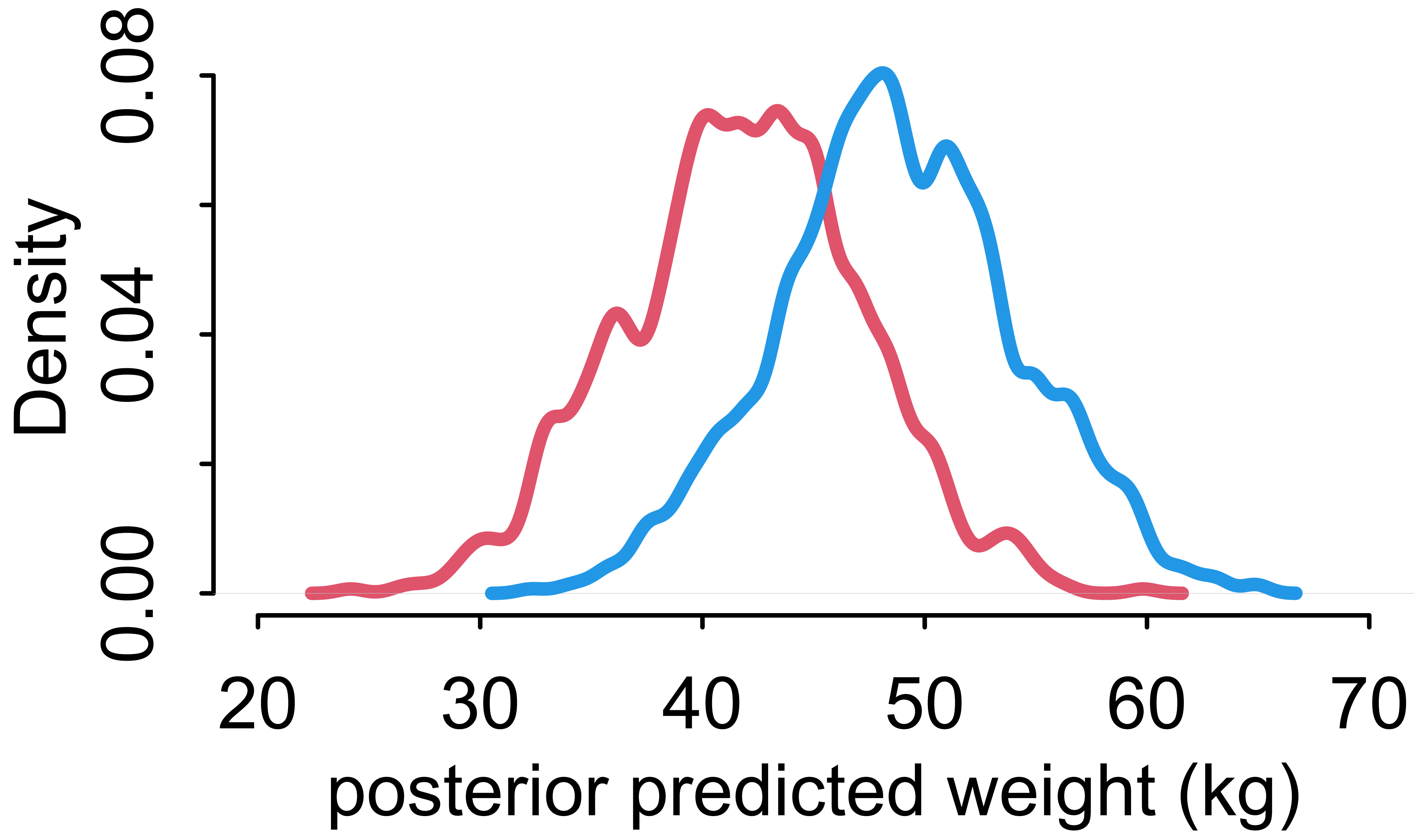


Always Be Contrasting

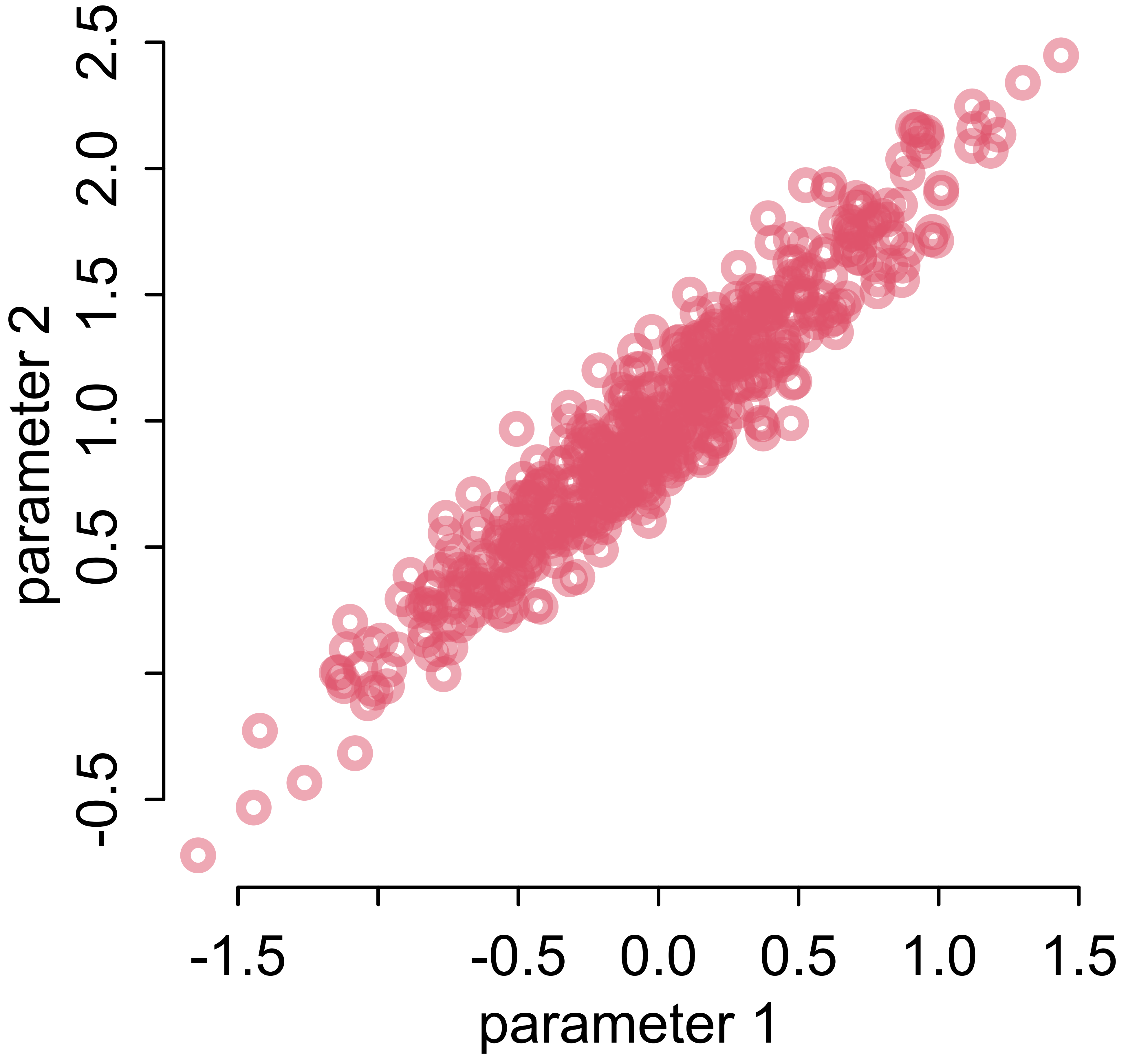
Need to compute **contrast**, the difference between the categories

It is **never** legitimate to compare **overlap** in distributions

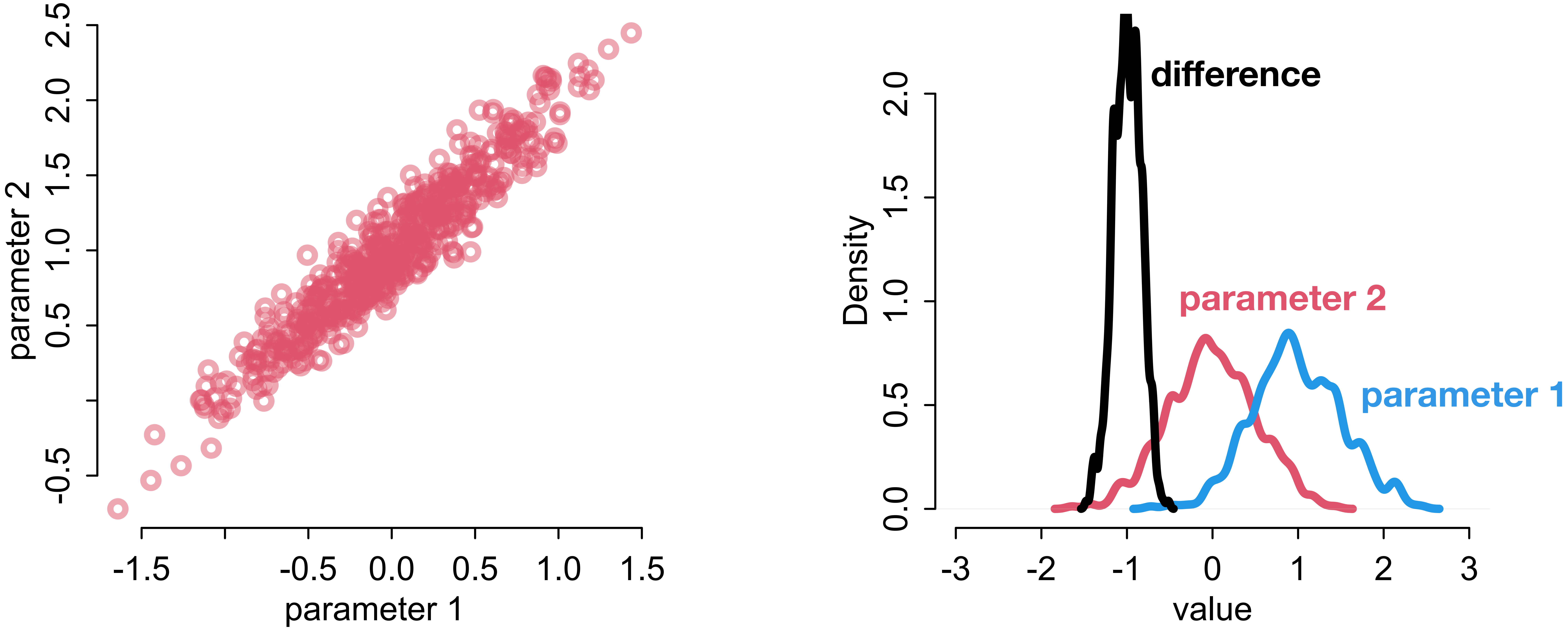
Must compute **contrast distribution**



It is **never** legitimate to compare **overlap** in parameters or lines



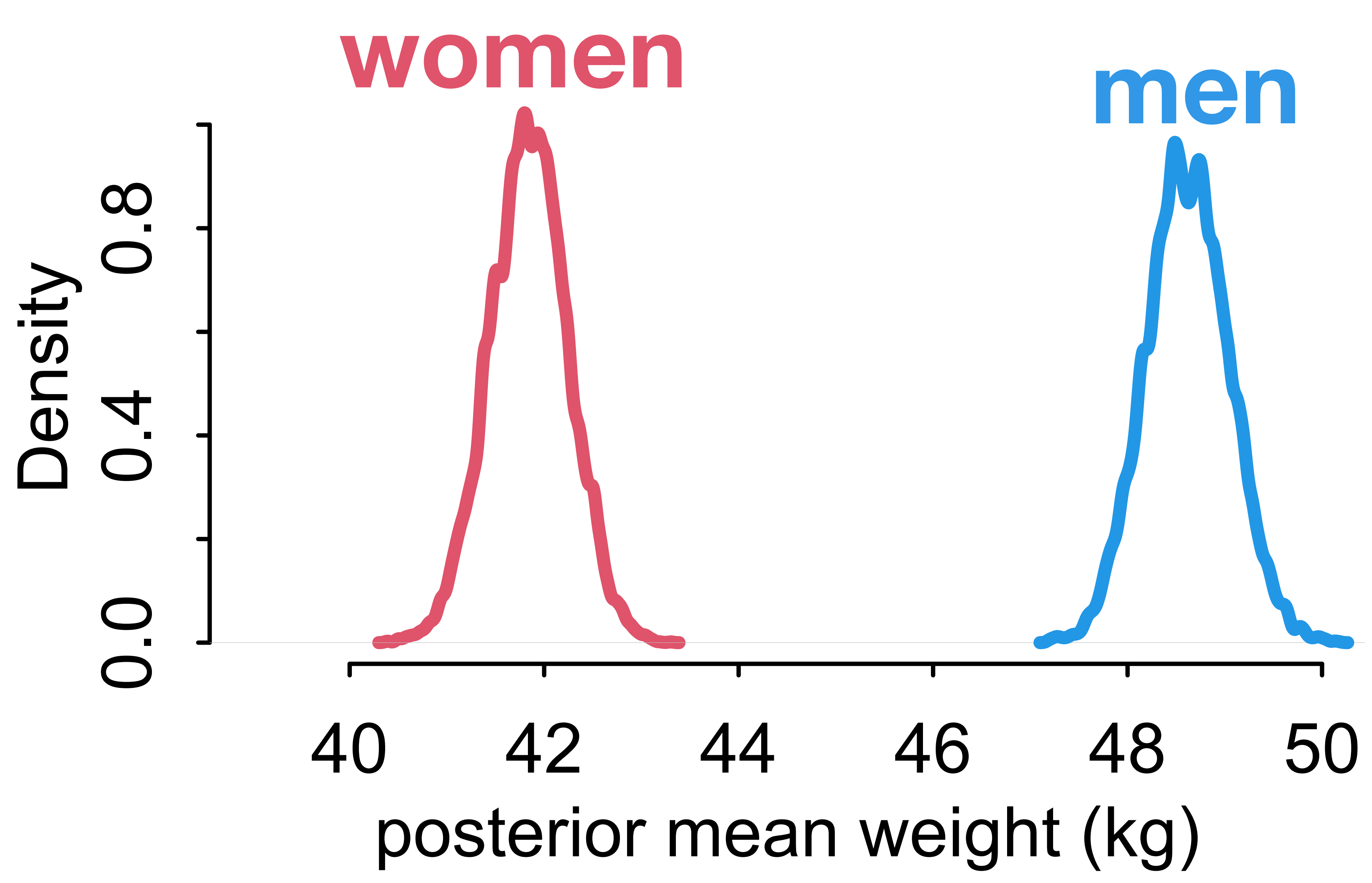
It is **never** legitimate to compare **overlap** in parameters or lines



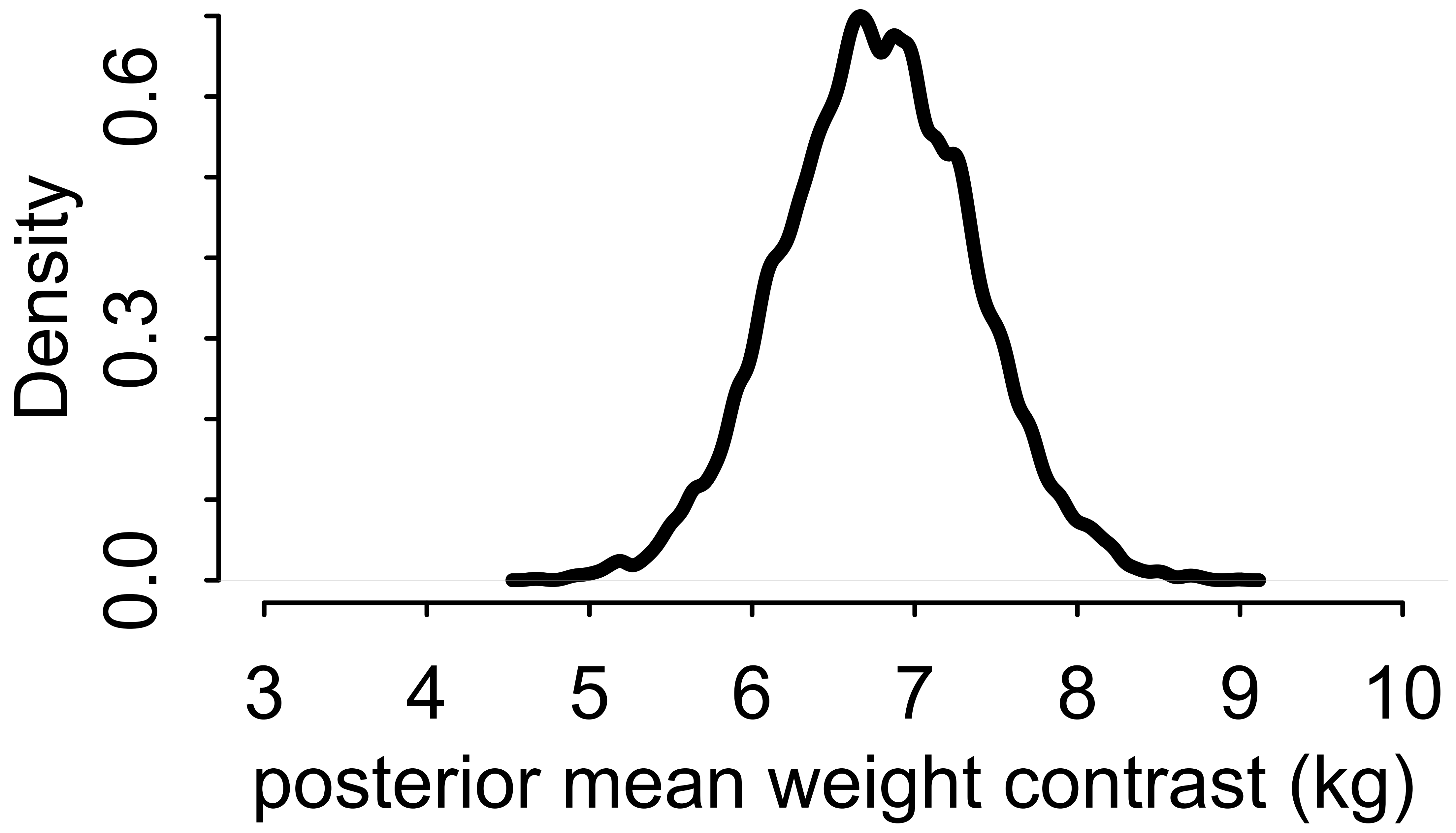
This means no comparing **confidence intervals** or **p-values** either!

Causal contrast

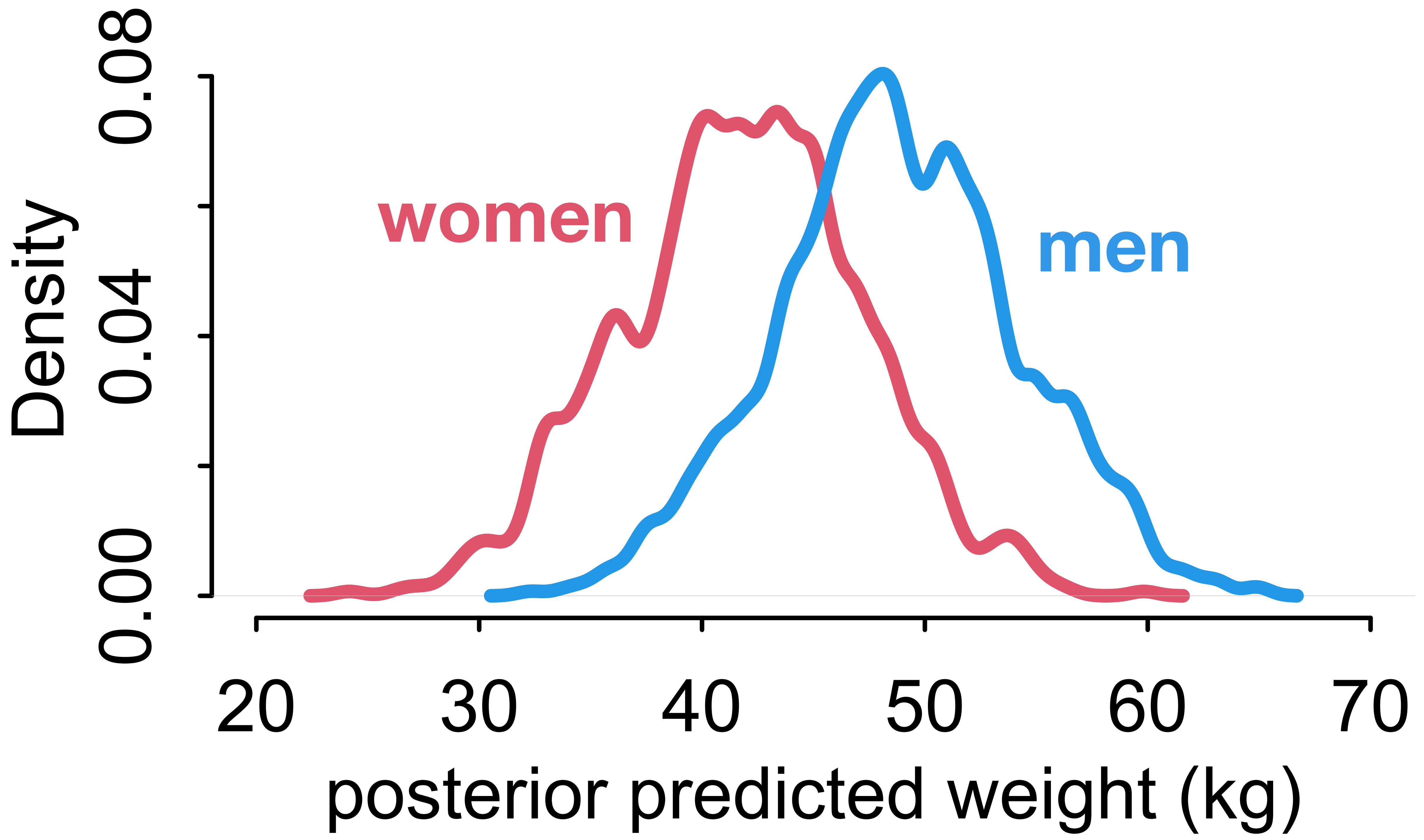
```
> str(post)
List of 2
 $ sigma: num [1:10000] 5.38 5.54 5.45 5.72 5.28 ...
 $ a    : num [1:10000, 1:2] 42.8 41.5 41 41.9 42 ...
```




```
# causal contrast (in means)
mu_contrast <- post$a[,2] - post$a[,1]

dens( mu_contrast , xlim=c(3,10) , lwd=3 ,
col=1 , xlab="posterior mean weight contrast
(kg)" )
```

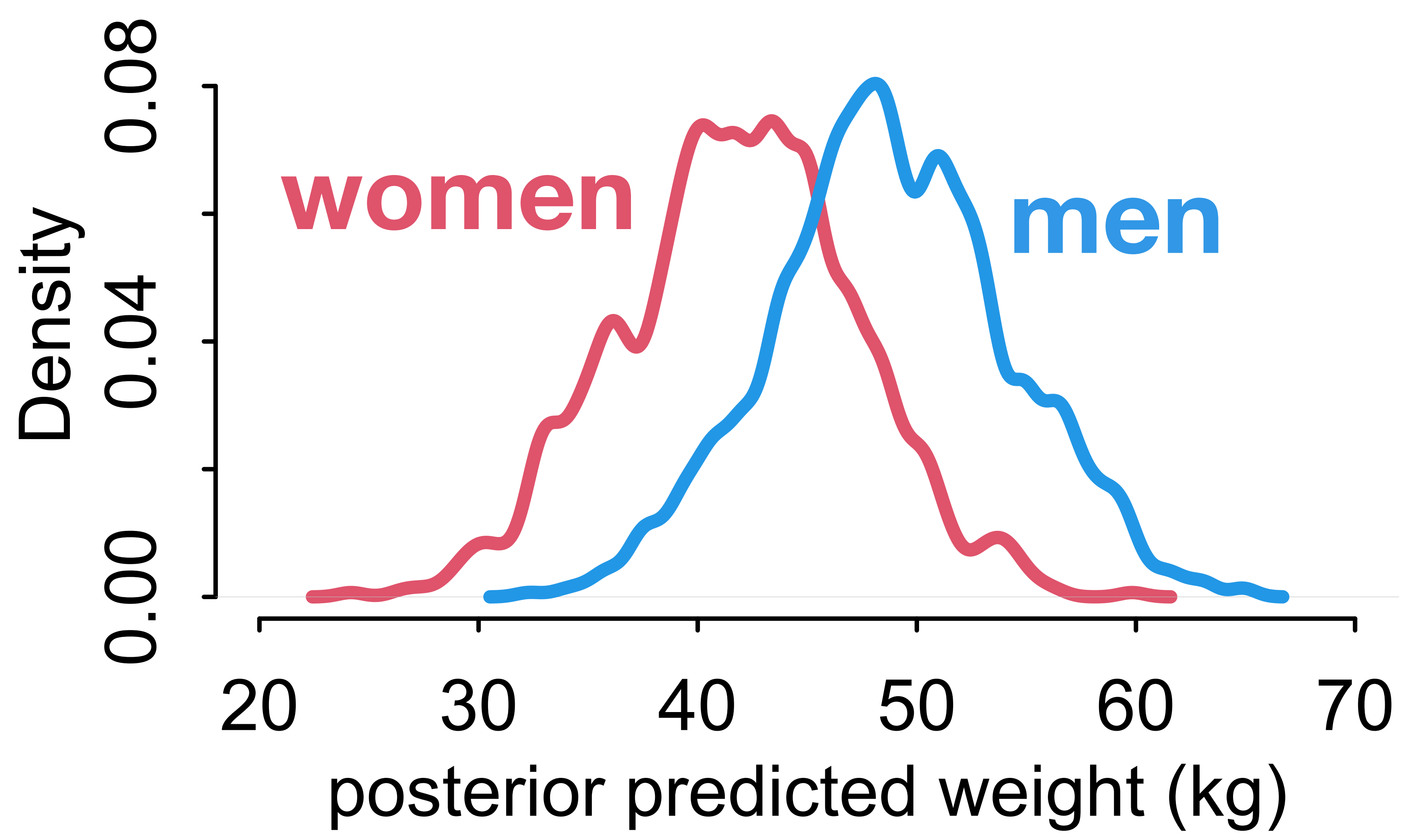


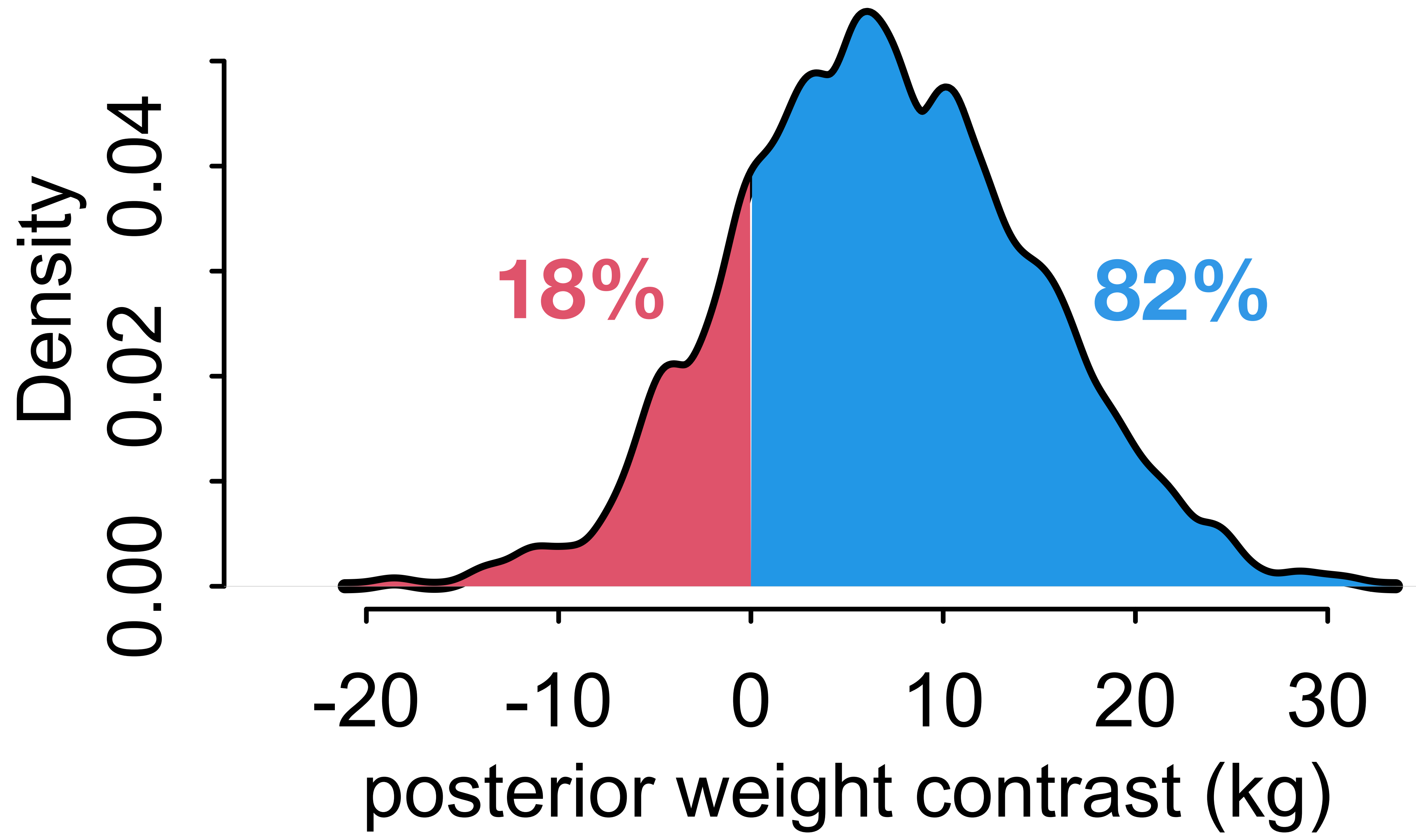
Weight contrast



Weight contrast

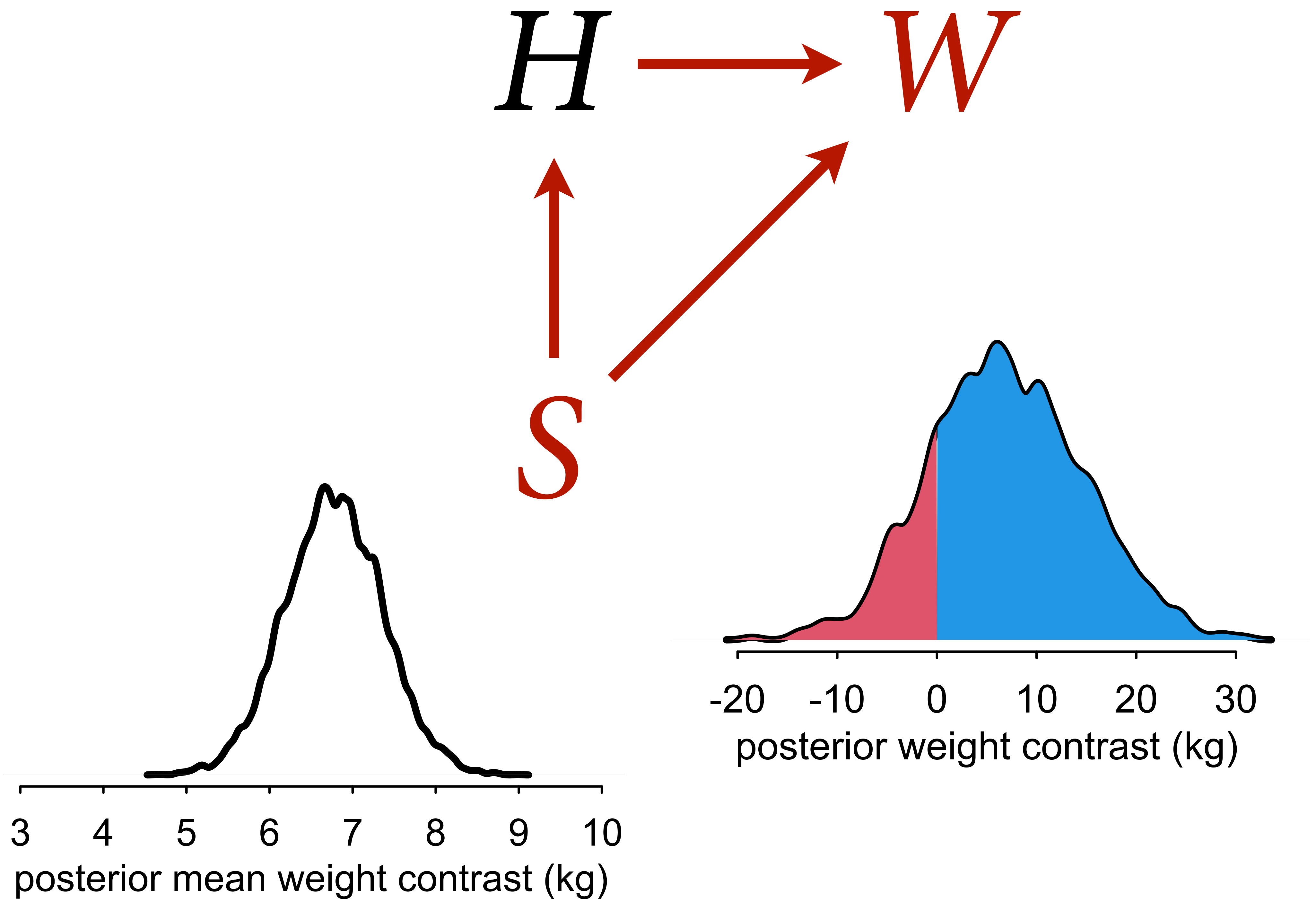
```
# posterior W distributions
W1 <- rnorm( 1000 , post$a[,1] , post$sigma )
W2 <- rnorm( 1000 , post$a[,2] , post$sigma )

# contrast
W_contrast <- W2 - W1
dens( W_contrast , xlim=c(-25,35) , lwd=3 ,
col=1 , xlab="posterior weight contrast (kg)" )


# proportion above zero
sum( W_contrast > 0 ) / 1000
# proportion below zero
sum( W_contrast < 0 ) / 1000
```



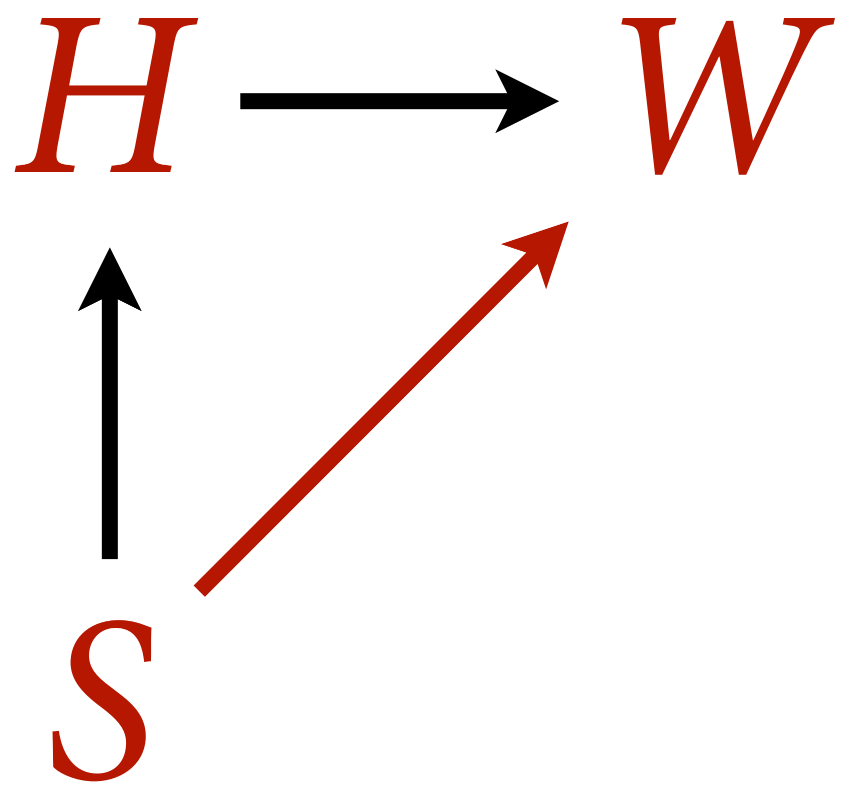


From estimand to estimate

Q: Causal effect of S on W?



Q: Direct causal effect of S on W?

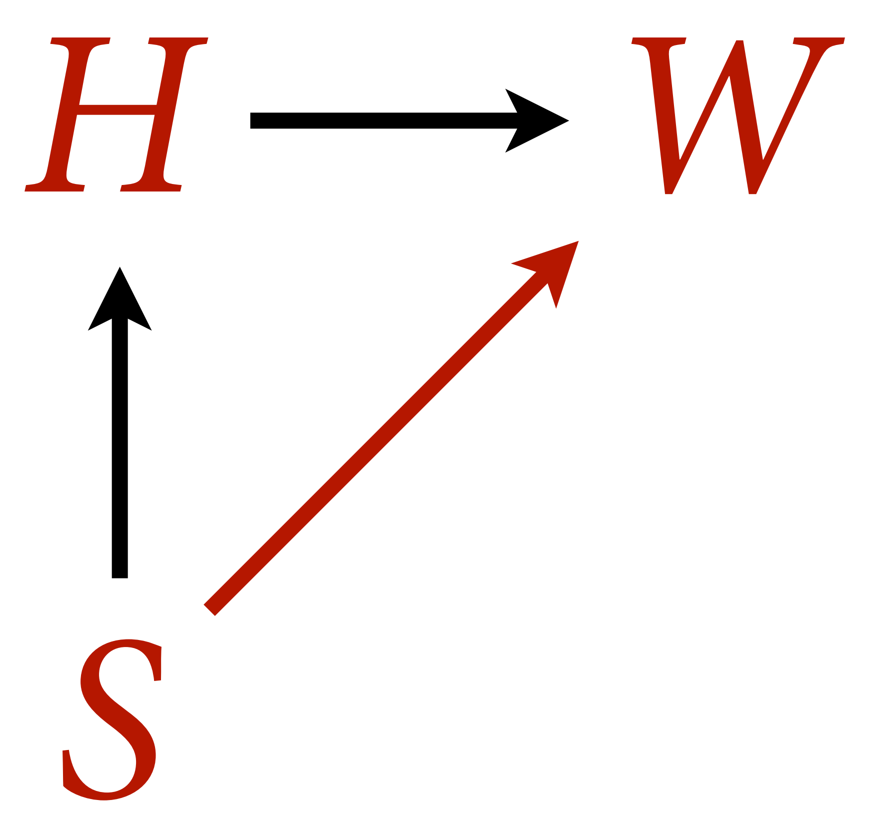


From estimand to estimate

Q: Direct causal effect of S on W?

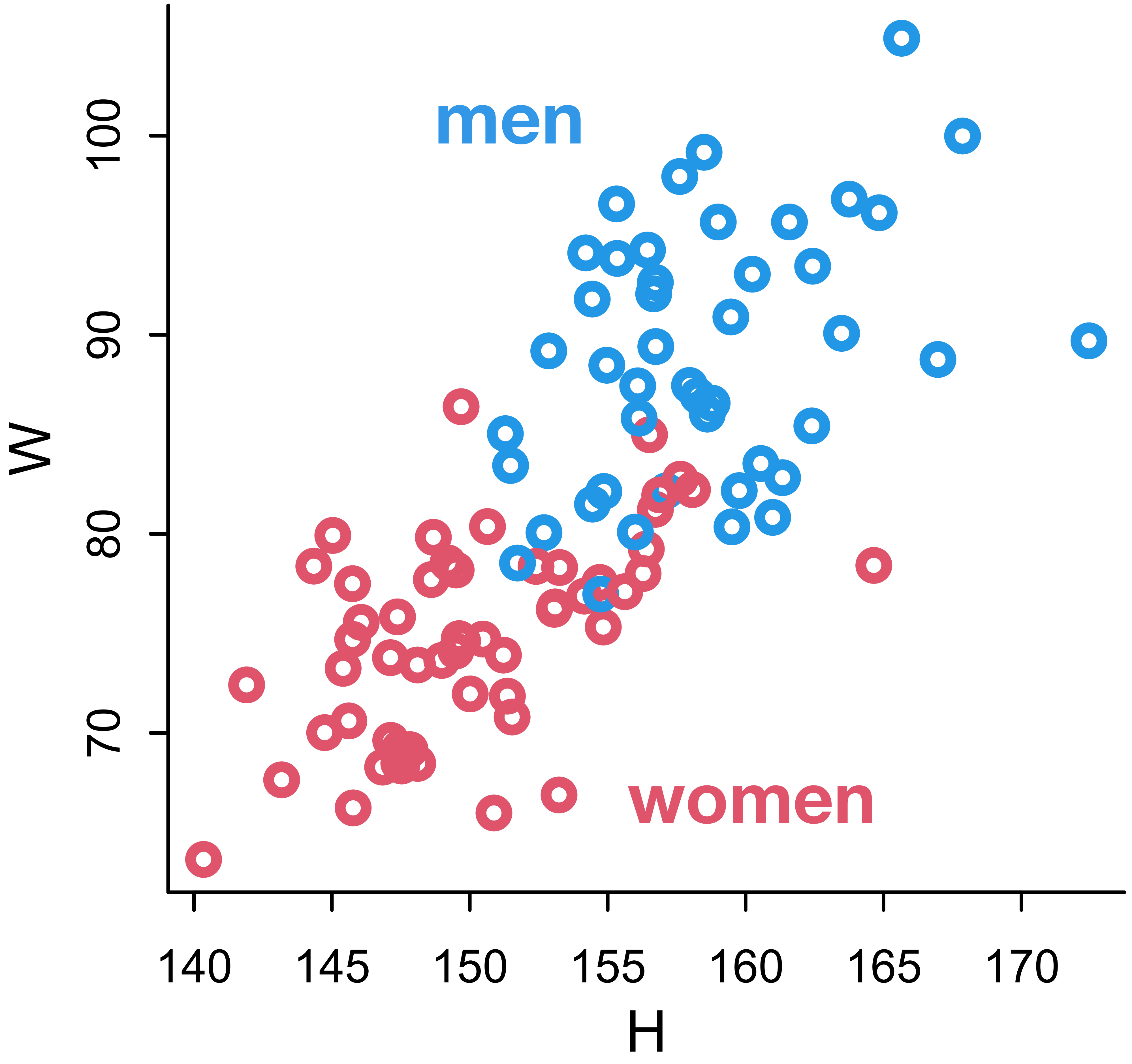
Now we need to “block” association through H

This means stratify by H



```
S <- rbern(100)+1
dat <- sim_HW(S,b=c(0.5,0.5),a=c(0,10))
```

indirect *direct*

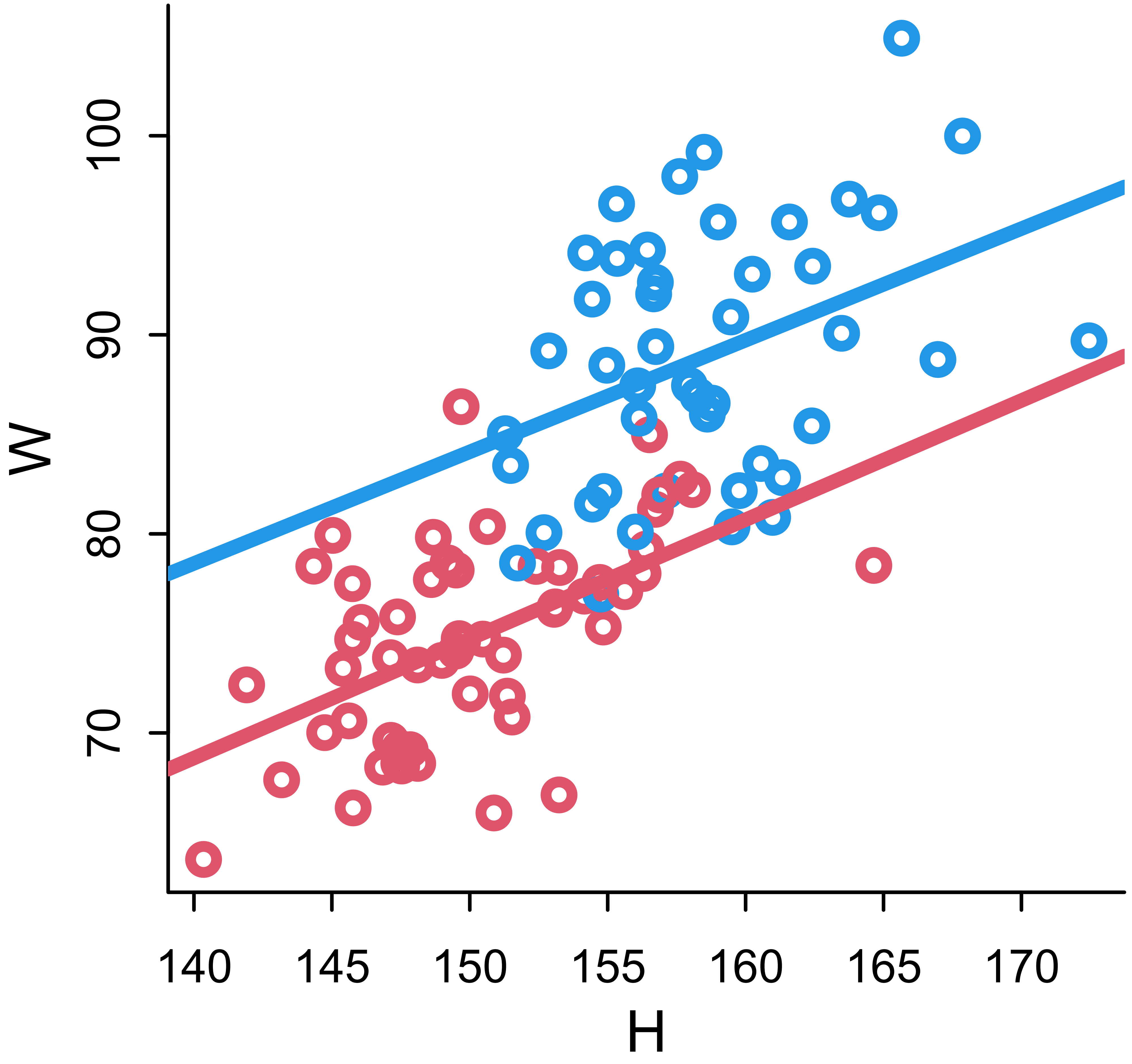


```
S <- rbern(100)+1
dat <- sim_HW(S,b=c(0.5,0.5),a=c(0,10))
```

indirect *direct*

Indirect: H influences W differently (slopes)

Direct: W differs net any slope differences



Index Variables & Lines

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \underbrace{\alpha}_{\textit{intercept}} + \underbrace{\beta}_{\textit{slope}}(H_i - \bar{H})$$

Centering variables

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta(\underbrace{H_i - \underline{\bar{H}}}_{\text{"centered" height}})$$

average height

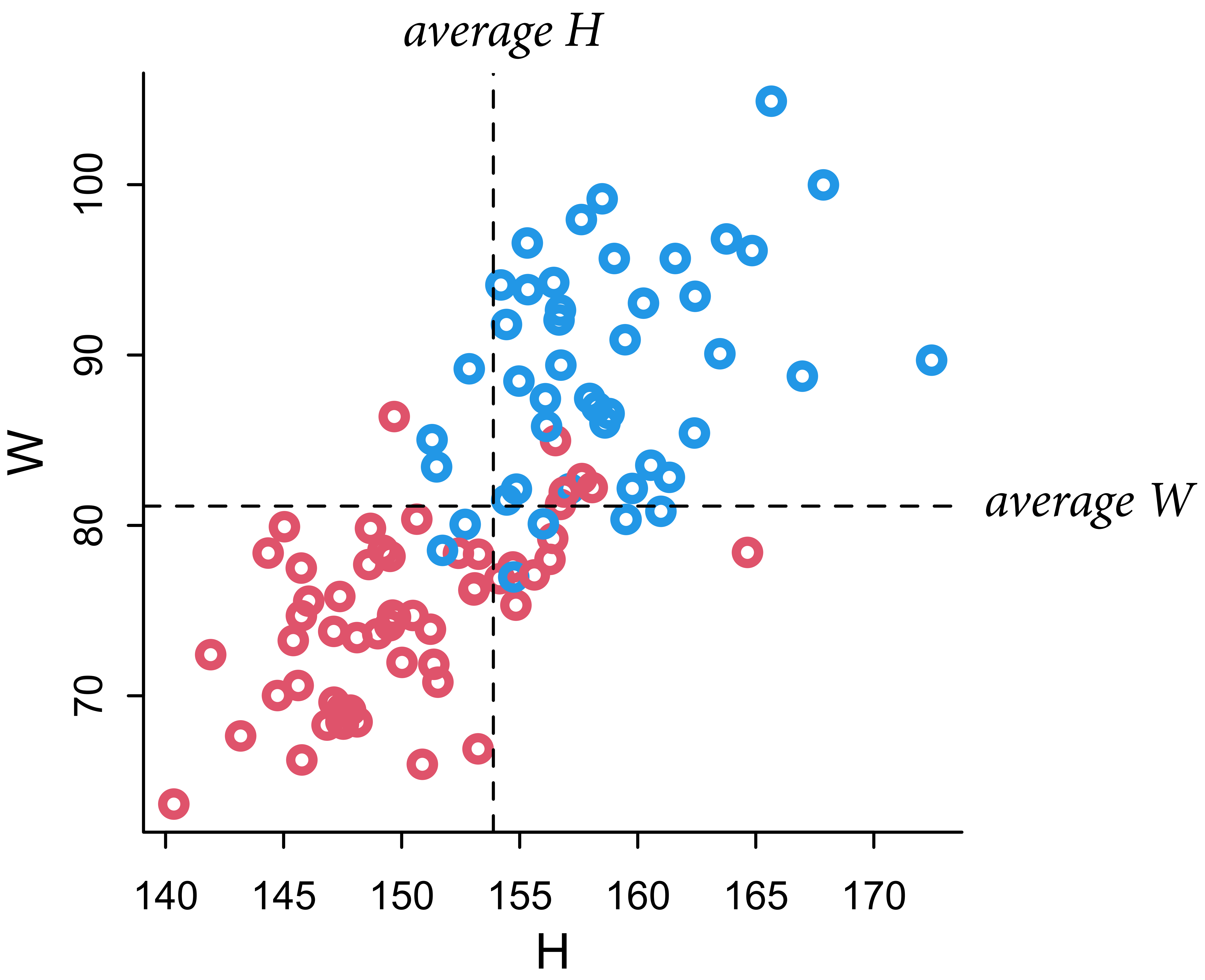
"centered" height

Centering *H* makes it so that *alpha* is the average weight of a person with average height

Centering makes it easier to define scientific priors for *alpha*

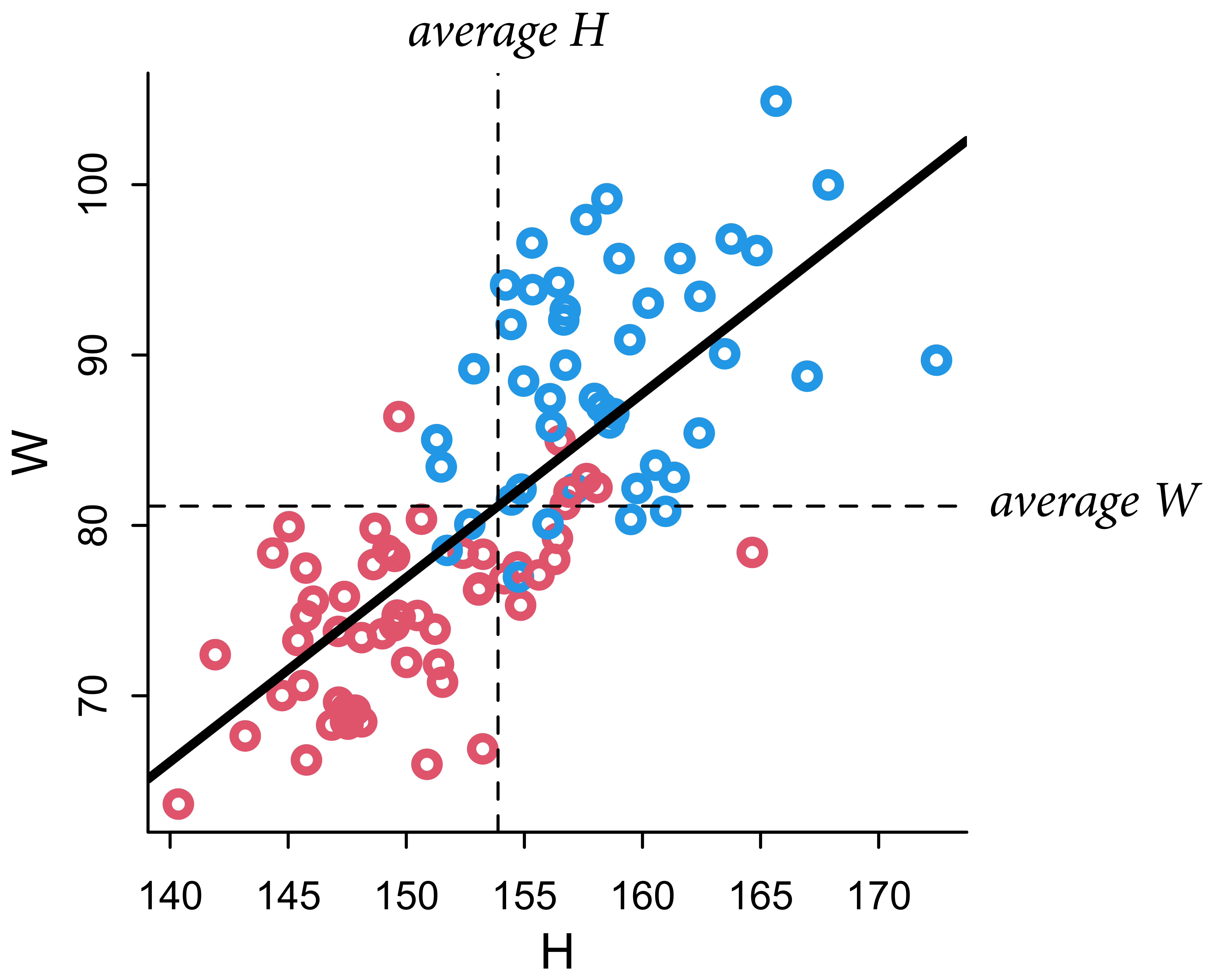
$$W_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta(H_i - \bar{H})$$

Centering H makes it so that *alpha* is the average weight of a person with average height



$$W_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta(H_i - \bar{H})$$

Centering *H* makes it so that *alpha* is the average weight of a person with average height



Index Variables & Lines

	H	W	S
1	152	48	2
2	140	36	1
3	137	32	1
4	157	53	2
5	145	41	1
6	164	63	2
7	149	38	1
8	169	55	2
9	148	35	1
10	165	54	2
11	154	50	1
12	151	41	2

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{S[i]} + \beta_{S[i]}(H_i - \bar{H})$$

sex of i-th person

S = 1 indicates female
S = 2 indicates male

$$\alpha = [\alpha_1, \alpha_2] \qquad \beta = [\beta_1, \beta_2]$$

Two intercepts and two slopes, one for each value in S

Index Variables & Lines

	H	W	S
1	152	48	2
2	140	36	1
3	137	32	1
4	157	53	2
5	145	41	1
6	164	63	2
7	149	38	1
8	169	55	2
9	148	35	1
10	165	54	2
11	154	50	1
12	151	41	2

i = *1*

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{\boxed{S[i]}} + \beta_{\boxed{S[i]}}(H_i - \bar{H})$$

S[1]=2

$$\alpha = [\alpha_1, \boxed{\alpha_2}] \qquad \beta = [\beta_1, \boxed{\beta_2}]$$

Index Variables & Lines

	H	W	S
1	152	48	2
2	140	36	1
3	137	32	1
4	157	53	2
5	145	41	1
6	164	63	2
7	149	38	1
8	169	55	2
9	148	35	1
10	165	54	2
11	154	50	1
12	151	41	2

$i = 2$

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

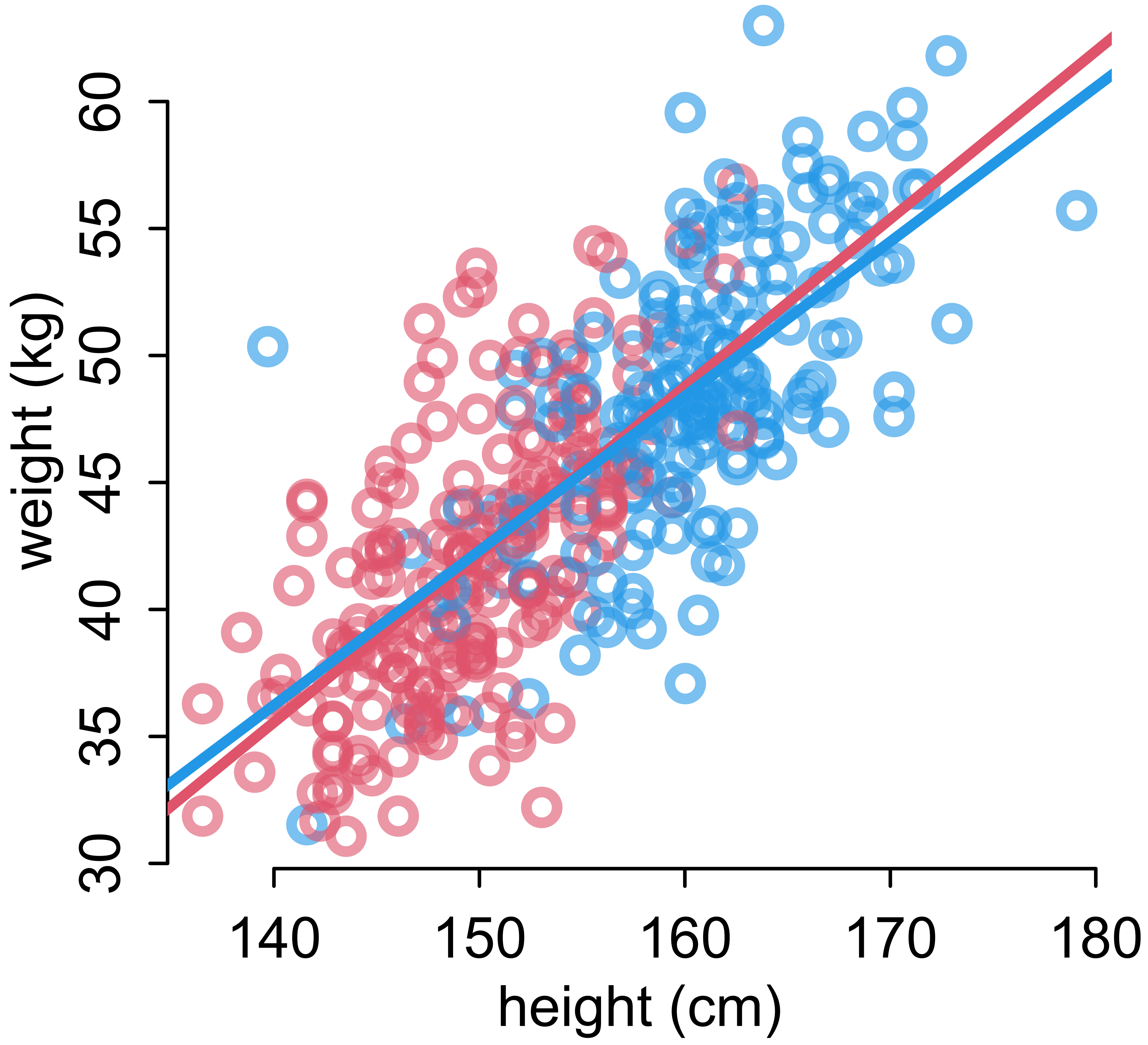
$$\mu_i = \alpha_{\boxed{S[i]}} + \beta_{\boxed{S[i]}}(H_i - \bar{H})$$

$S[2]=1$

$$\alpha = [\boxed{\alpha_1}, \alpha_2] \qquad \beta = [\boxed{\beta_1}, \beta_2]$$

Analyze the sample

```
data(Howell1)
d <- Howell1
d <- d[ d$age>=18 , ]
dat <- list(
    W = d$weight,
    H = d$height,
    Hbar = mean(d$height),
    S = d$male + 1 ) # S=1 female, S=2 male

m_SHW <- quap(
    alist(
        W ~ dnorm(mu,sigma),
        mu <- a[S] + b[S]*(H-Hbar),
        a[S] ~ dnorm(60,10),
        b[S] ~ dunif(0,1),
        sigma ~ dunif(0,10)
    ), data=dat )
```



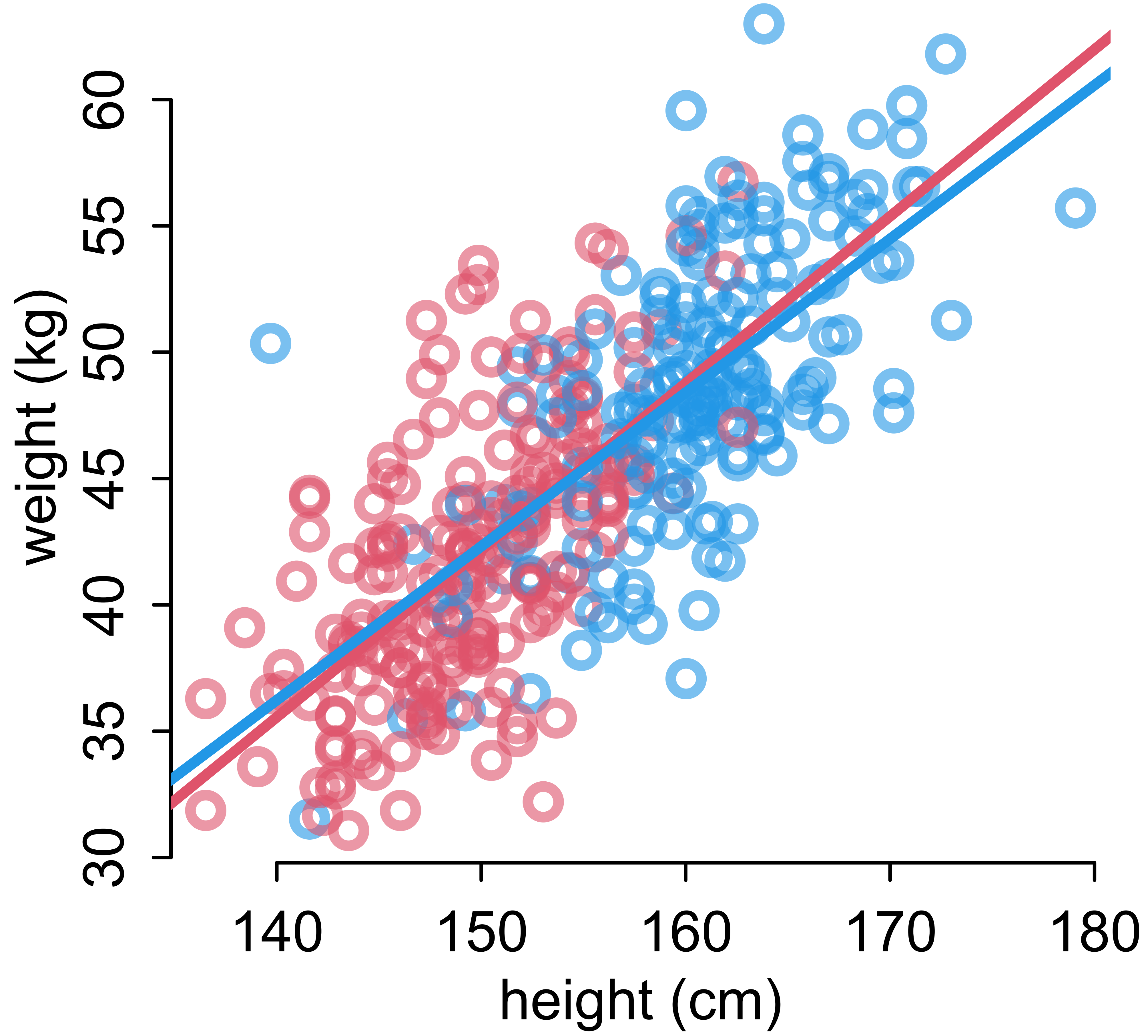
Contrasts at each height

(1) Compute posterior predictive for women

(2) Compute posterior predictive for men

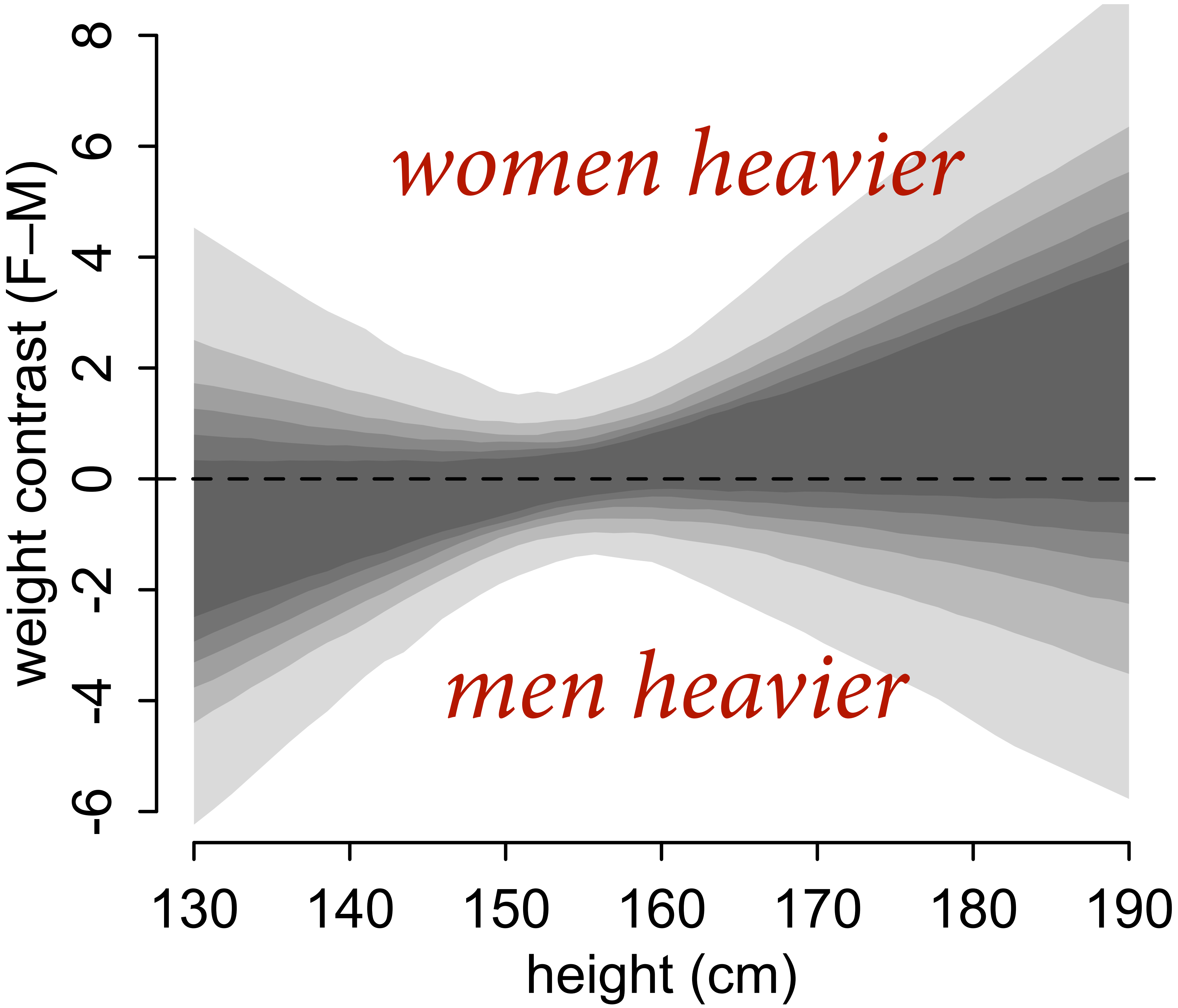
(3) Subtract (2) from (1)

(4) Plot distribution at each height (on right)



Contrasts at each height

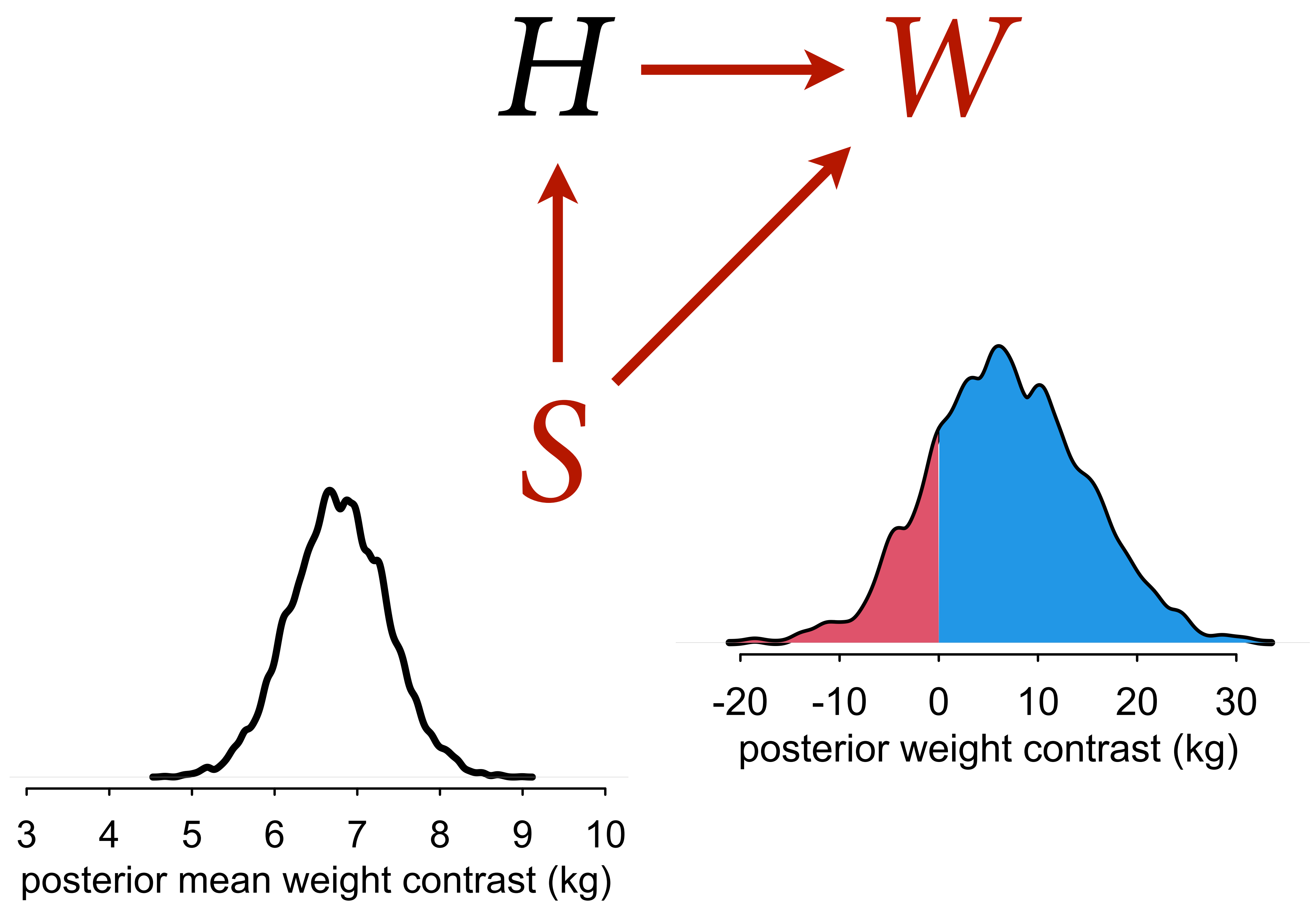
```
xseq <- seq(from=130,to=190,len=50)

muF <-
link(m_adults2,data=list(S=rep(1,50),H=xseq,Hbar=mean(d$height)))
lines( xseq , apply(muF,2,mean) , lwd=3 , col=2 )

muM <-
link(m_adults2,data=list(S=rep(2,50),H=xseq,Hbar=mean(d$height)))
lines( xseq , apply(muM,2,mean) , lwd=3 , col=4 )

mu_contrast <- muF - muM
plot( NULL , xlim=range(xseq) , ylim=c(-6,8) , xlab="height (cm)"
, ylab="weight contrast (F-M)" )
for ( p in c(0.5,0.6,0.7,0.8,0.9,0.99) )
    shade( apply(mu_contrast,2,PI,prob=p) , xseq )
abline(h=0,lty=2)
```



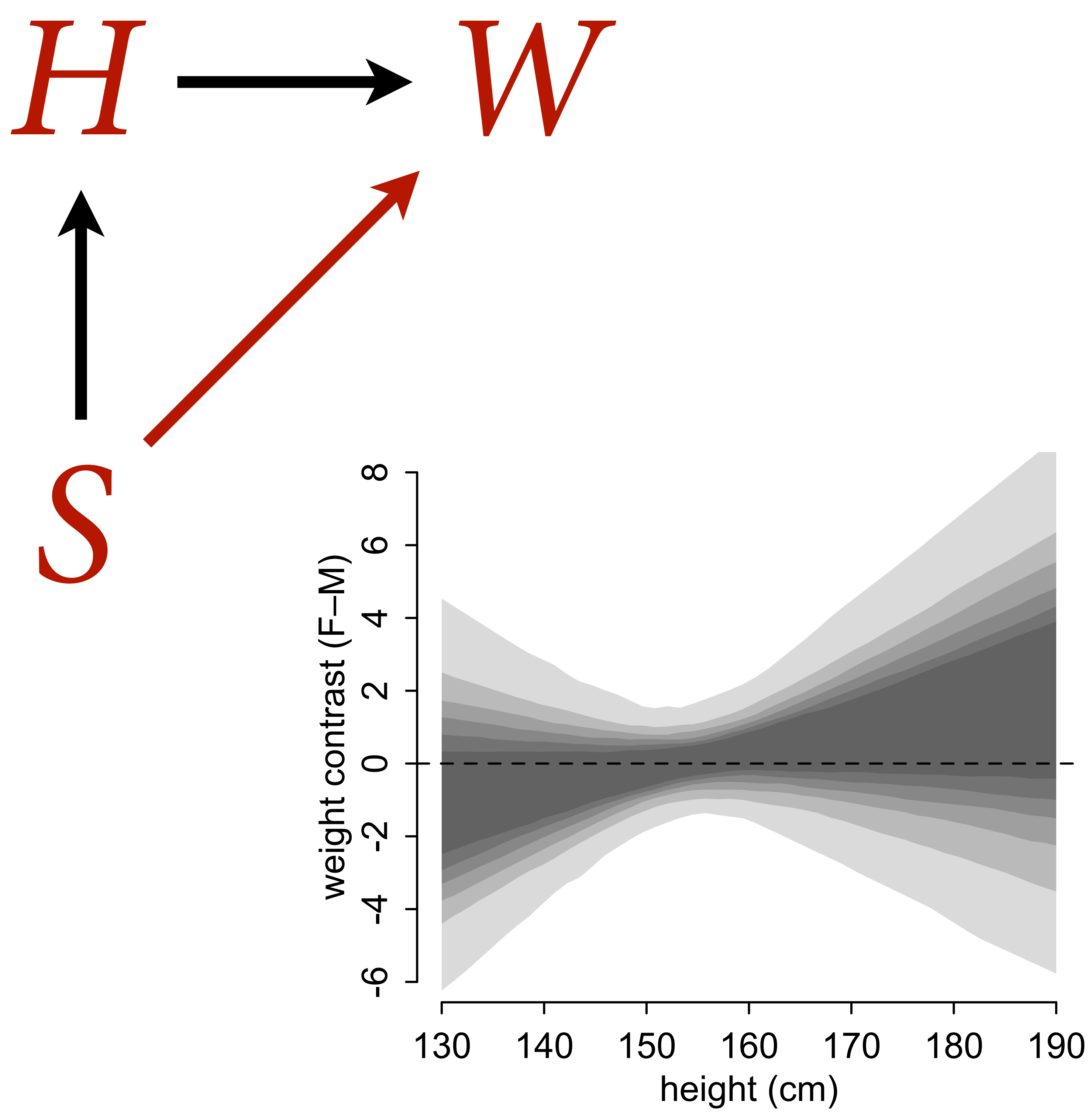
Nearly all of the causal effect of S acts through H

From estimand to estimate

Q: Causal effect of S on W?



Q: Direct causal effect of S on W?



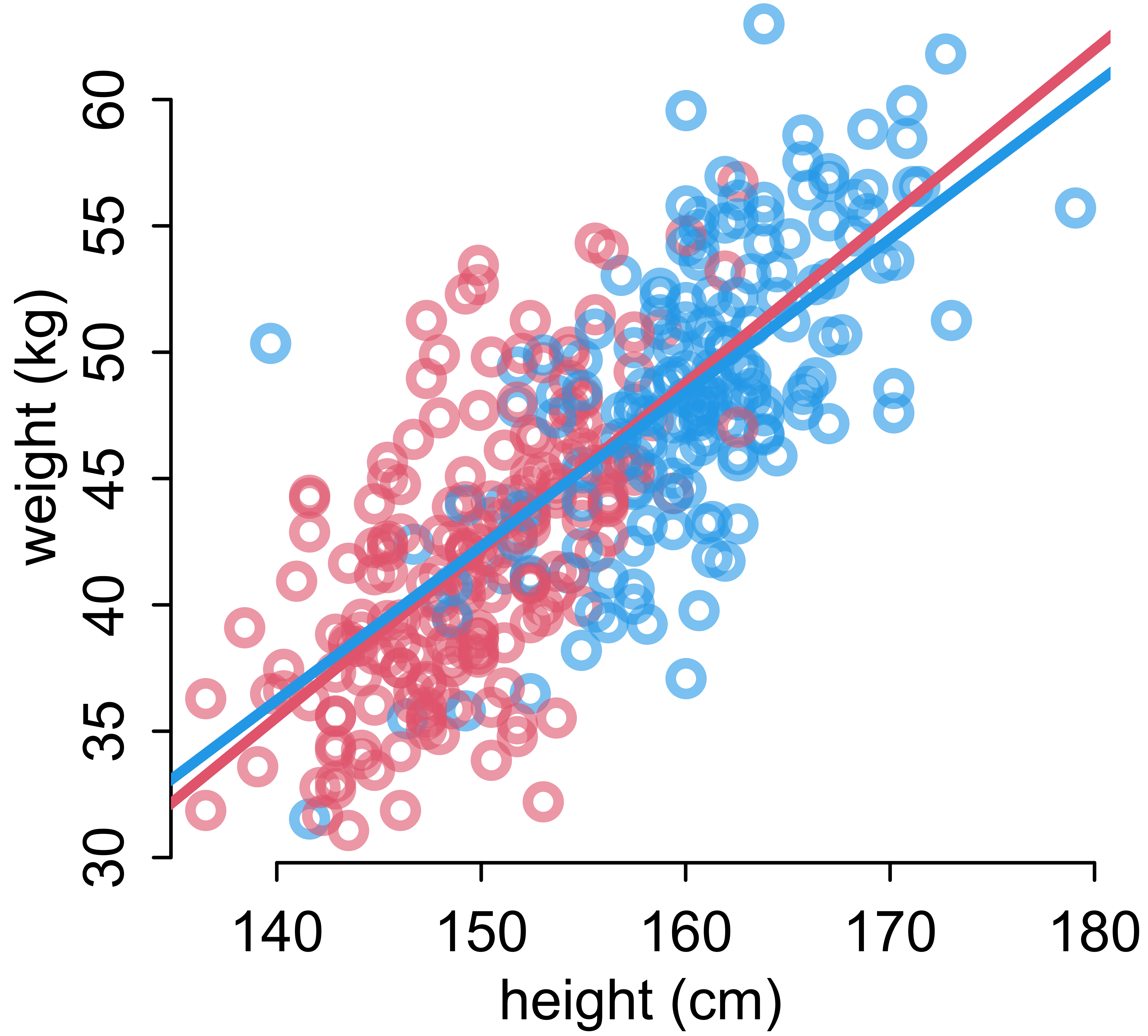
Categorical variables

Easy to use with index coding

Use samples to compute relevant contrasts

Always summarize (mean, interval) as the last step

Want **mean difference** and not **difference of means**



PAUSE

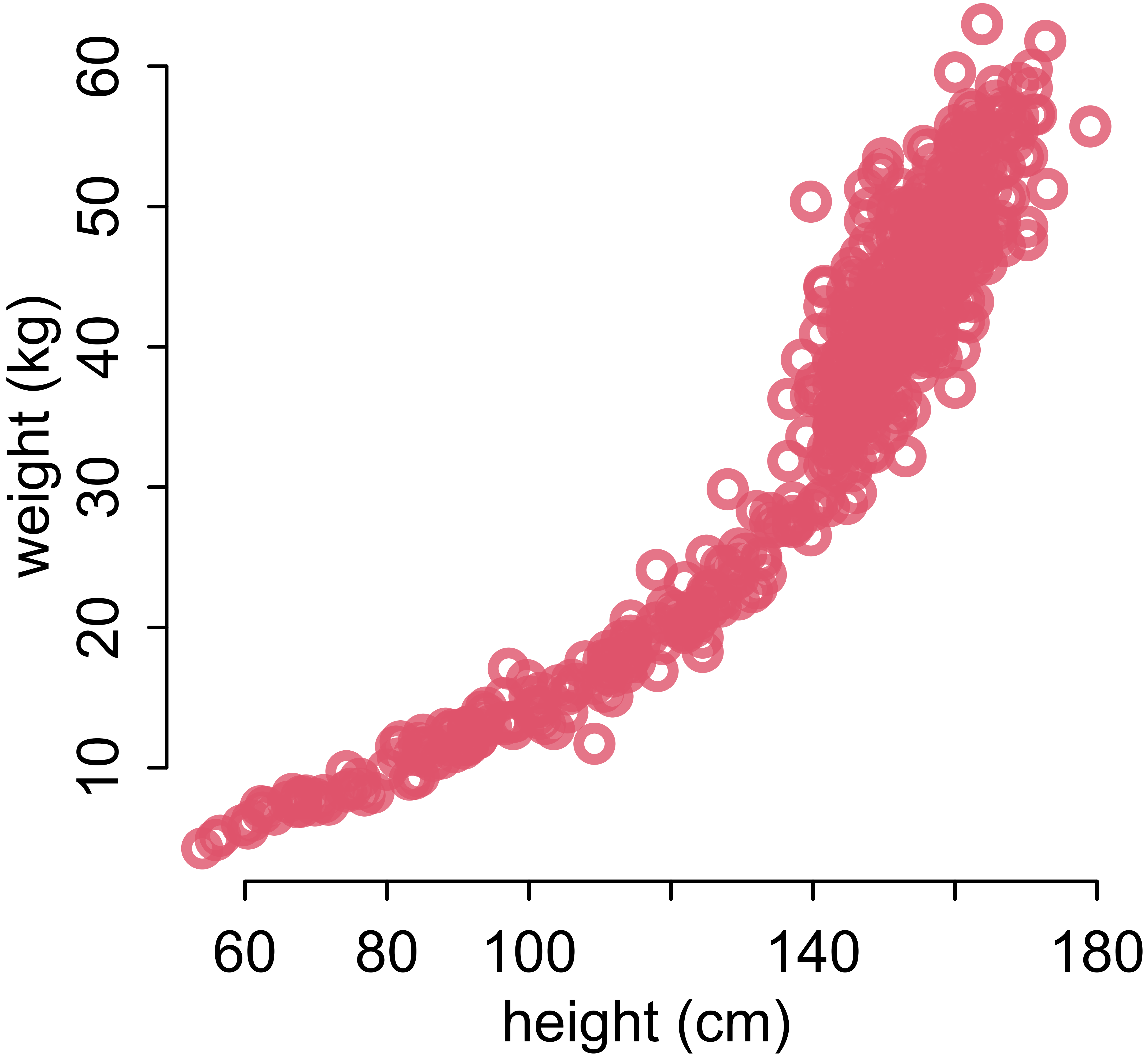
Curves from lines

```
library(rethinking)
data(Howell1)
```

H –> *W* obviously not linear

Linear models can easily fit curves

But this is not **mechanistic**



Curves from lines

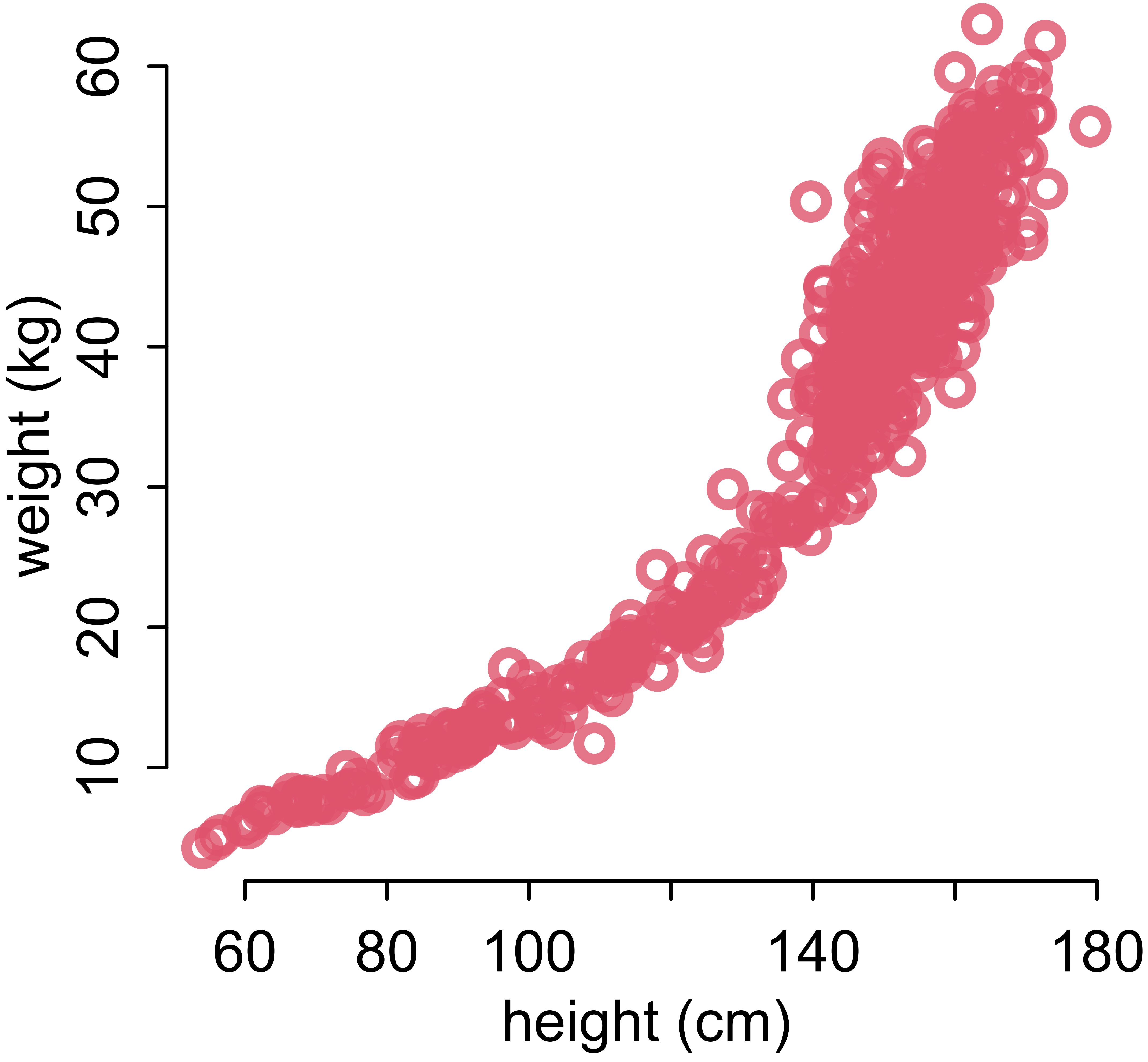
```r
library(rethinking)
data(Howell1)
```

Linear models can easily fit curves

Two popular strategies

(1) polynomials — awful

(2) splines and generalized additive models — less awful



Polynomial linear models

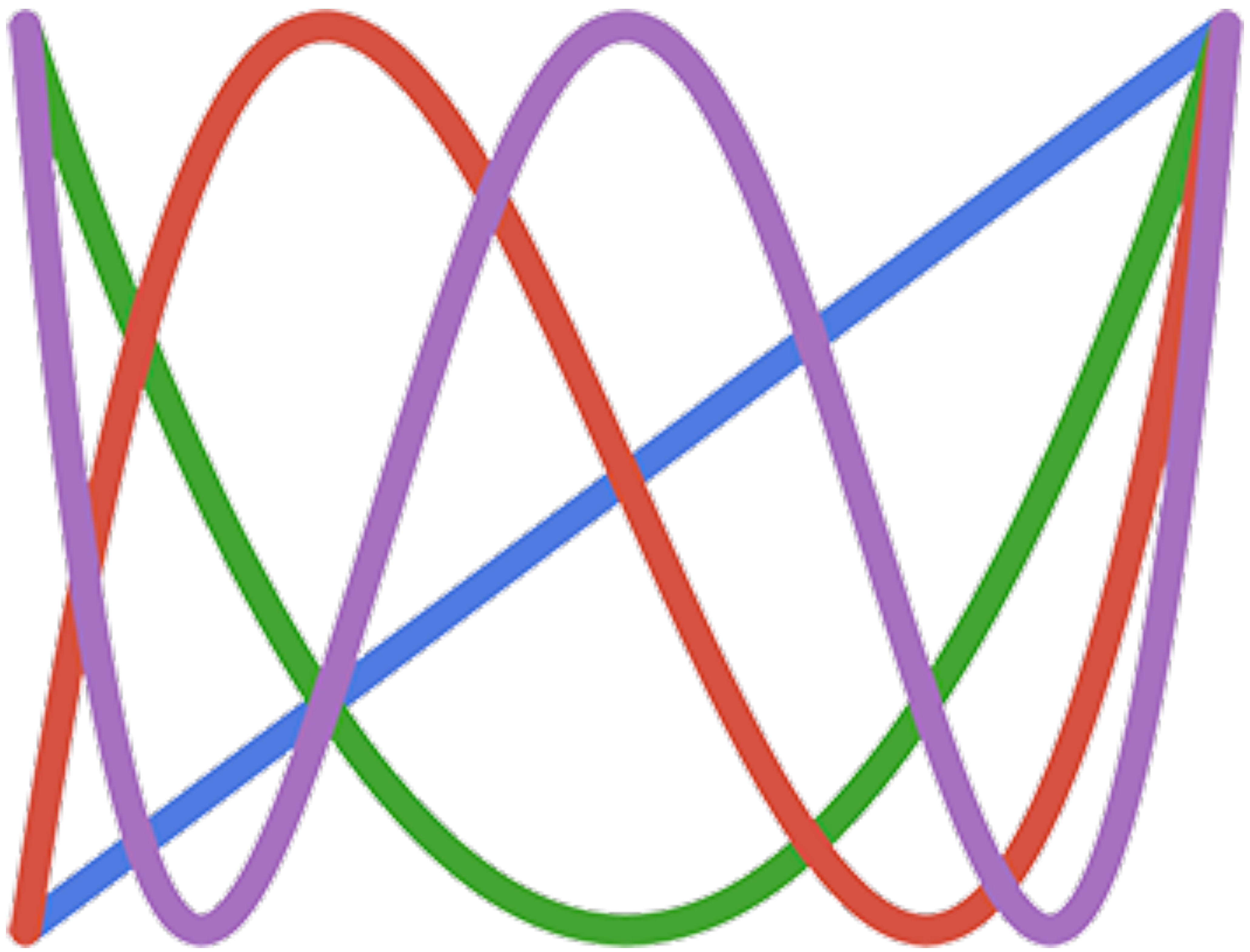
Strategy: polynomial functions

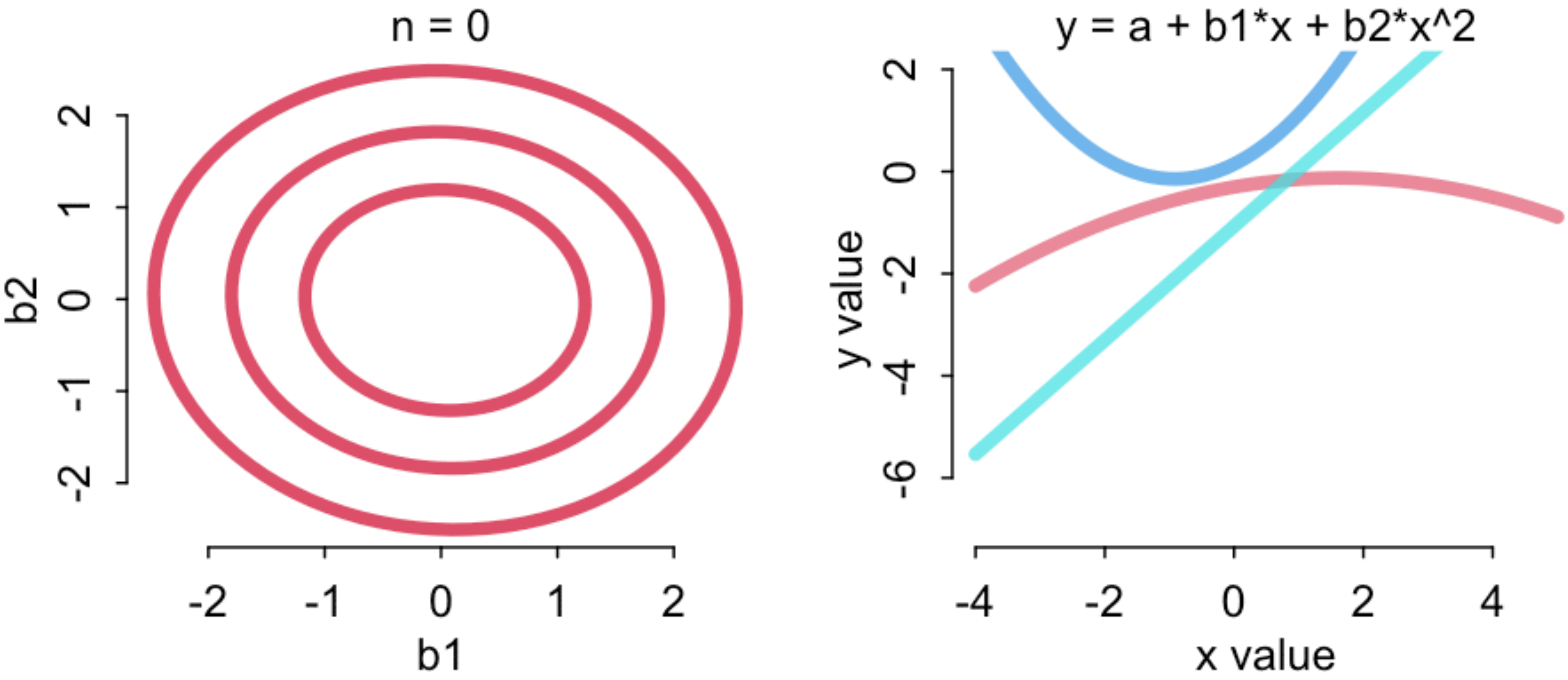
$$\mu_i = \alpha + \beta_1 x_i + \beta_2 x_i^2$$

Problems: strange symmetries, explosive uncertainty at edges

No *local* smoothing; only *global*

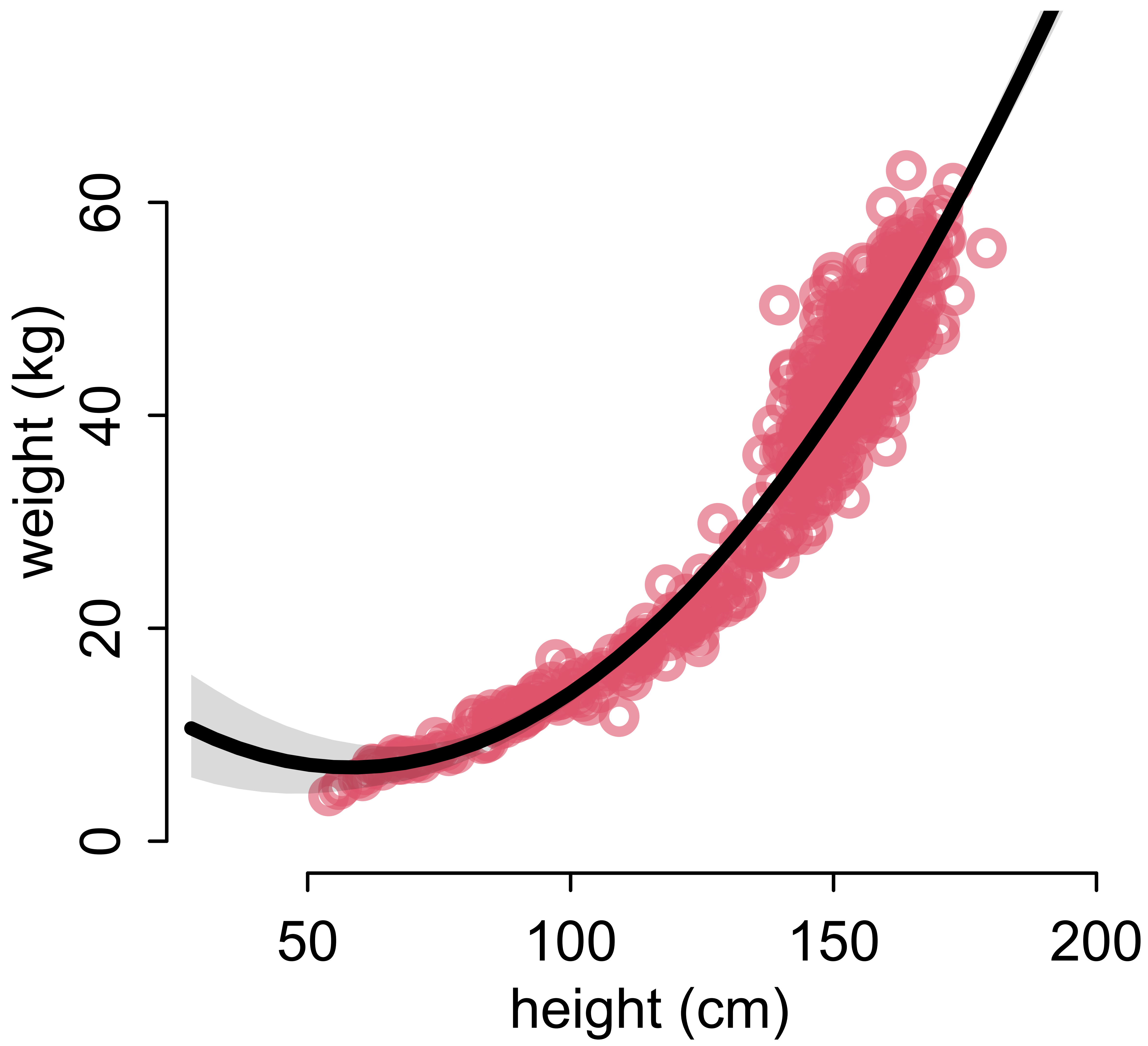
DO NOT USE

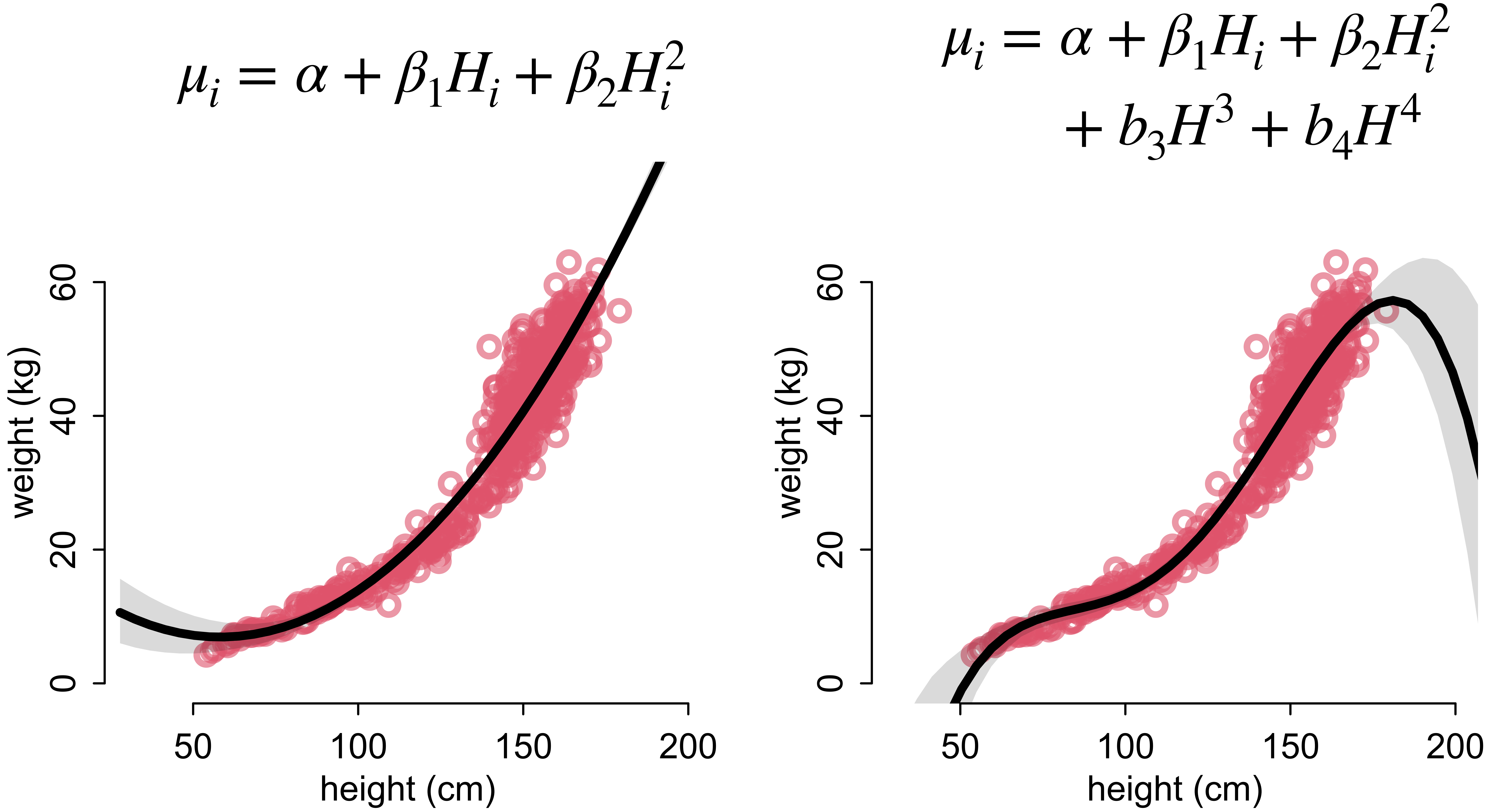




n = 0
b2
2
1
0
-1
-2
-2
-1
0
1
2
b1
y = a + b1*x + b2*x^2
y value
2
0
-2
-4
-6
-4
-2
0
2
4
x value

$$\mu_i = \alpha + \beta_1 H_i + \beta_2 H_i^2$$




$\mu_i = \alpha + \beta_1 H_i + \beta_2 H_i^2$
weight (kg)
0
20
40
60
50
100
150
200
height (cm)
$\mu_i = \alpha + \beta_1 H_i + \beta_2 H_i^2$
$+ b_3 H^3 + b_4 H^4$
weight (kg)
0
20
40
60
50
100
150
200
height (cm)

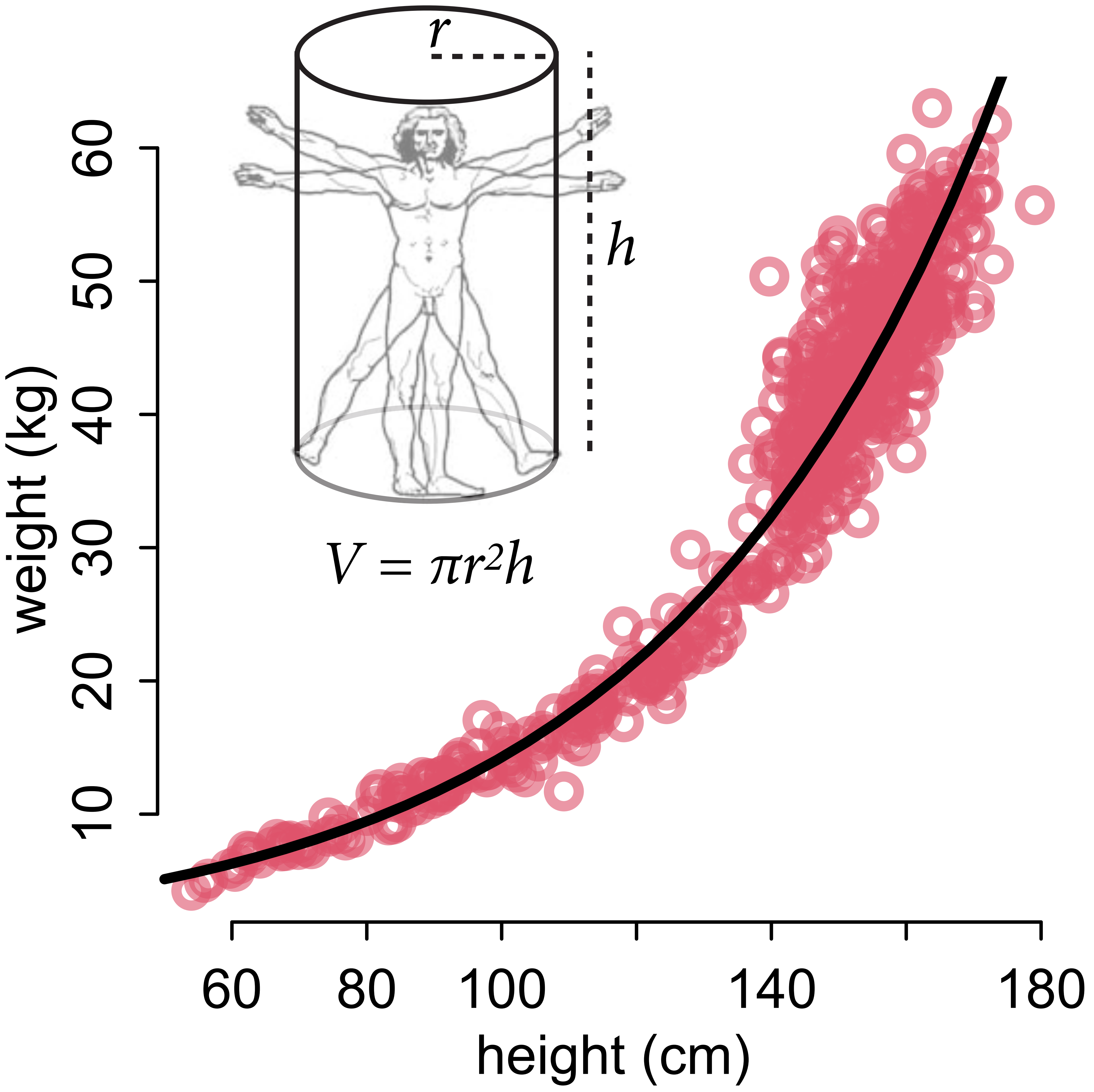
Thinking vs Fitting

Linear models can fit anything (Geocentrism)

Better to think

$$\log W_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta(H_i - \bar{H})$$

Will revisit in lecture 19

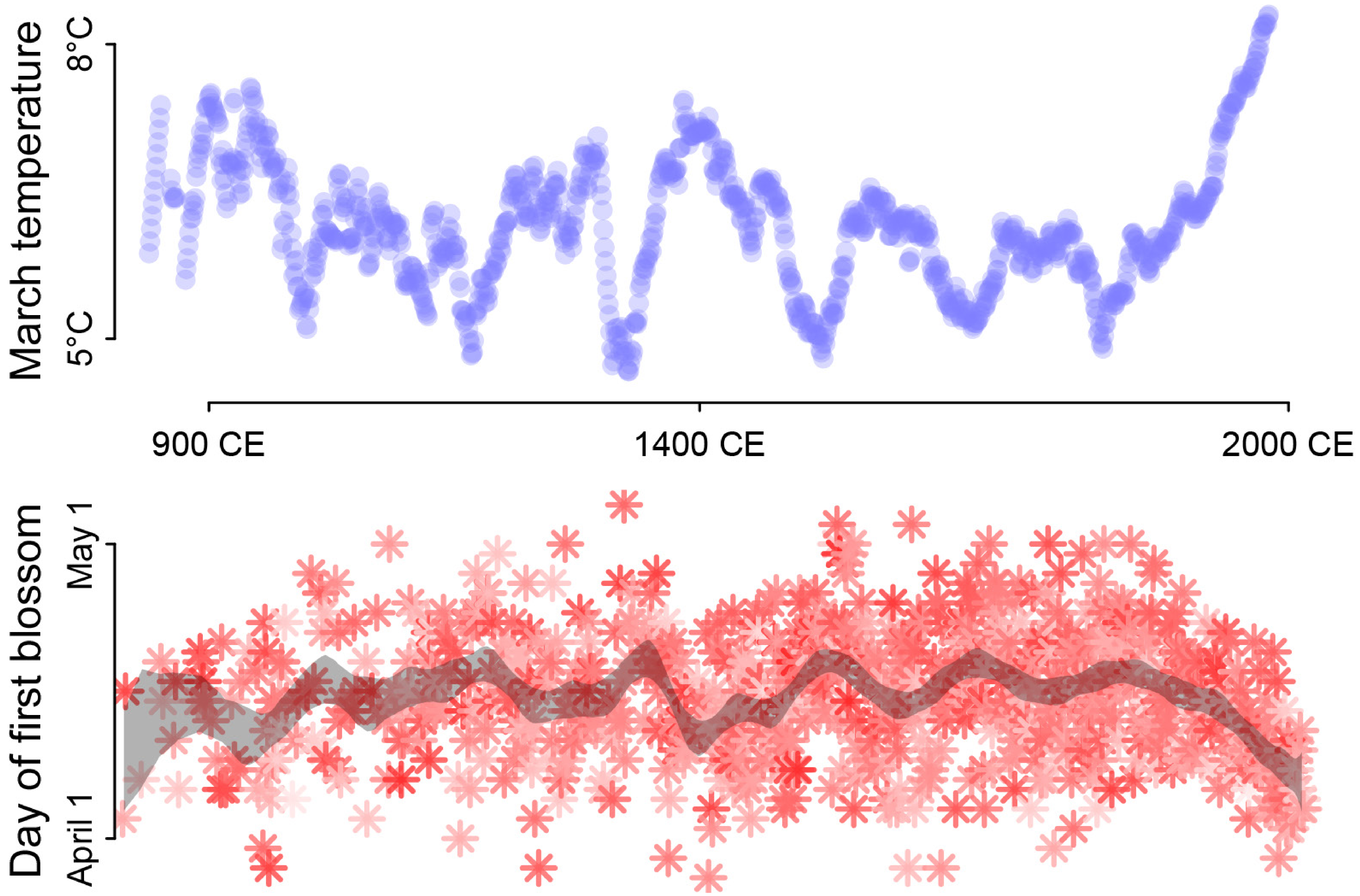


Splines

Basis-Splines: Wiggly function built from many local functions

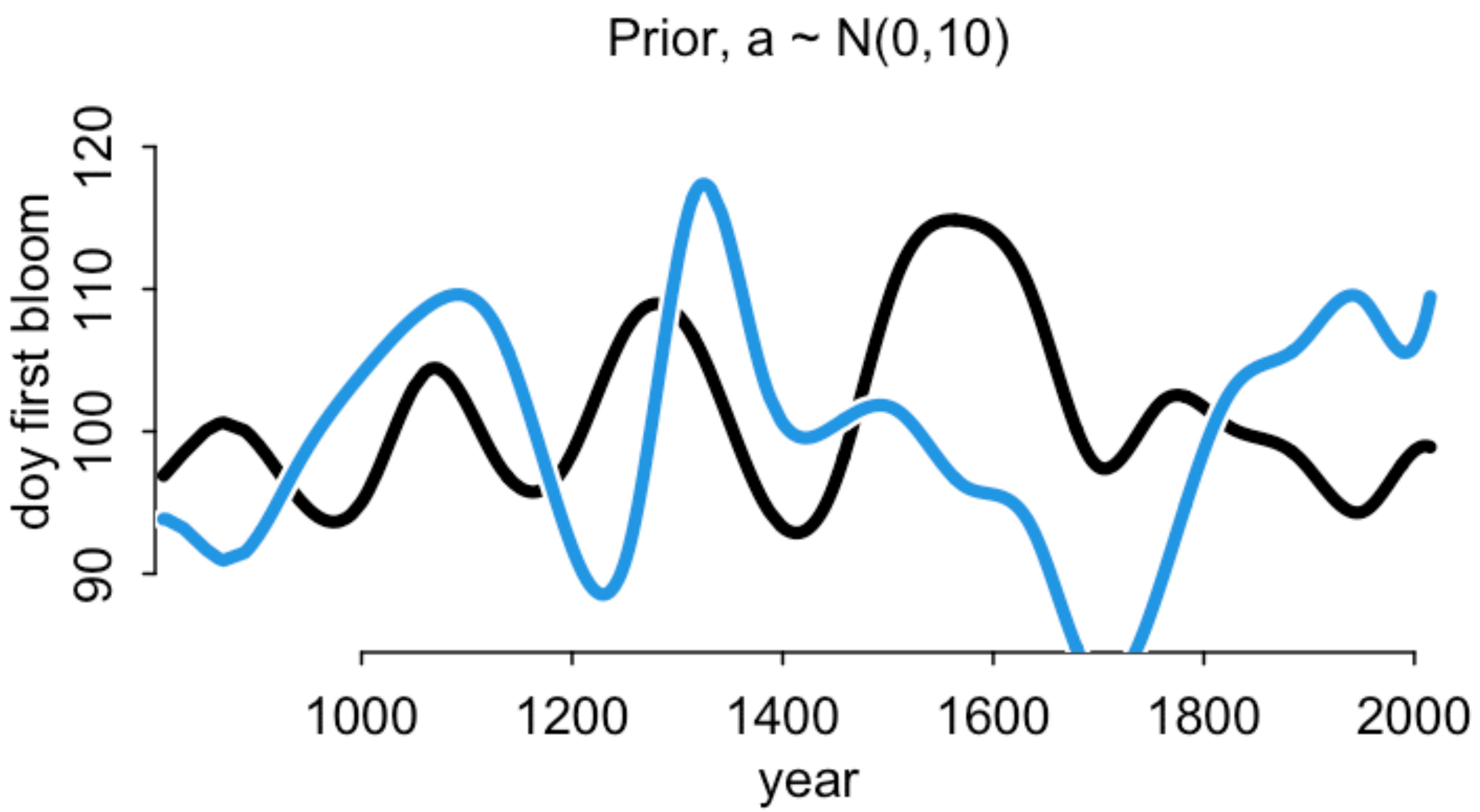
Basis function: A local function

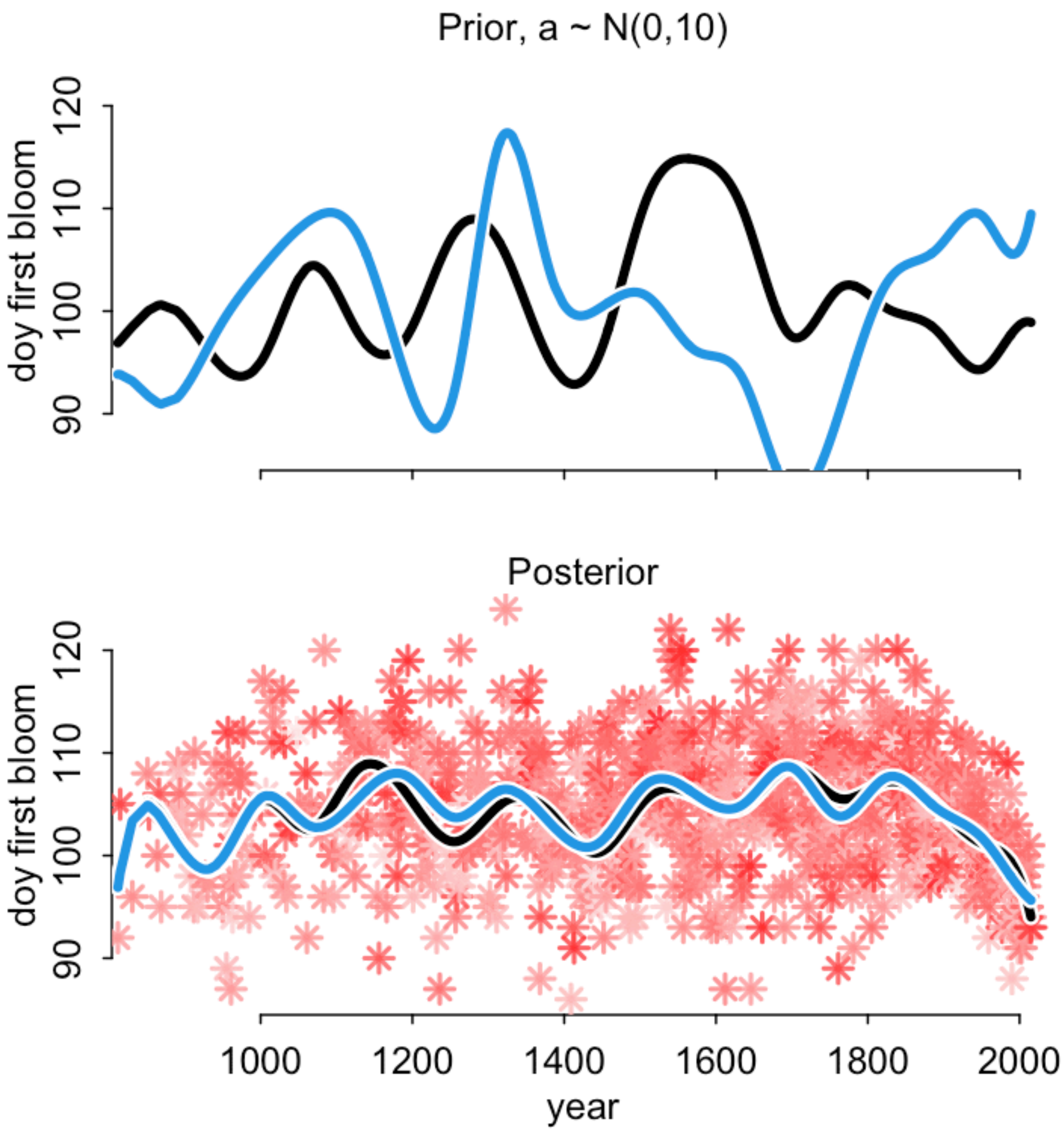
Local functions make splines better than polynomials, but equally geocentric




DIAG A
DIAG B
DIAG C


TRANSOM SHAPE
DIAG A
DIAG B
DIAG C
DIAG D
DIAG E
DIAG F
DIAG G
DATE


Prior, a ~ N(0,10)
doy first bloom
90
100
110
120
1000
1200
1400
1600
1800
2000
year

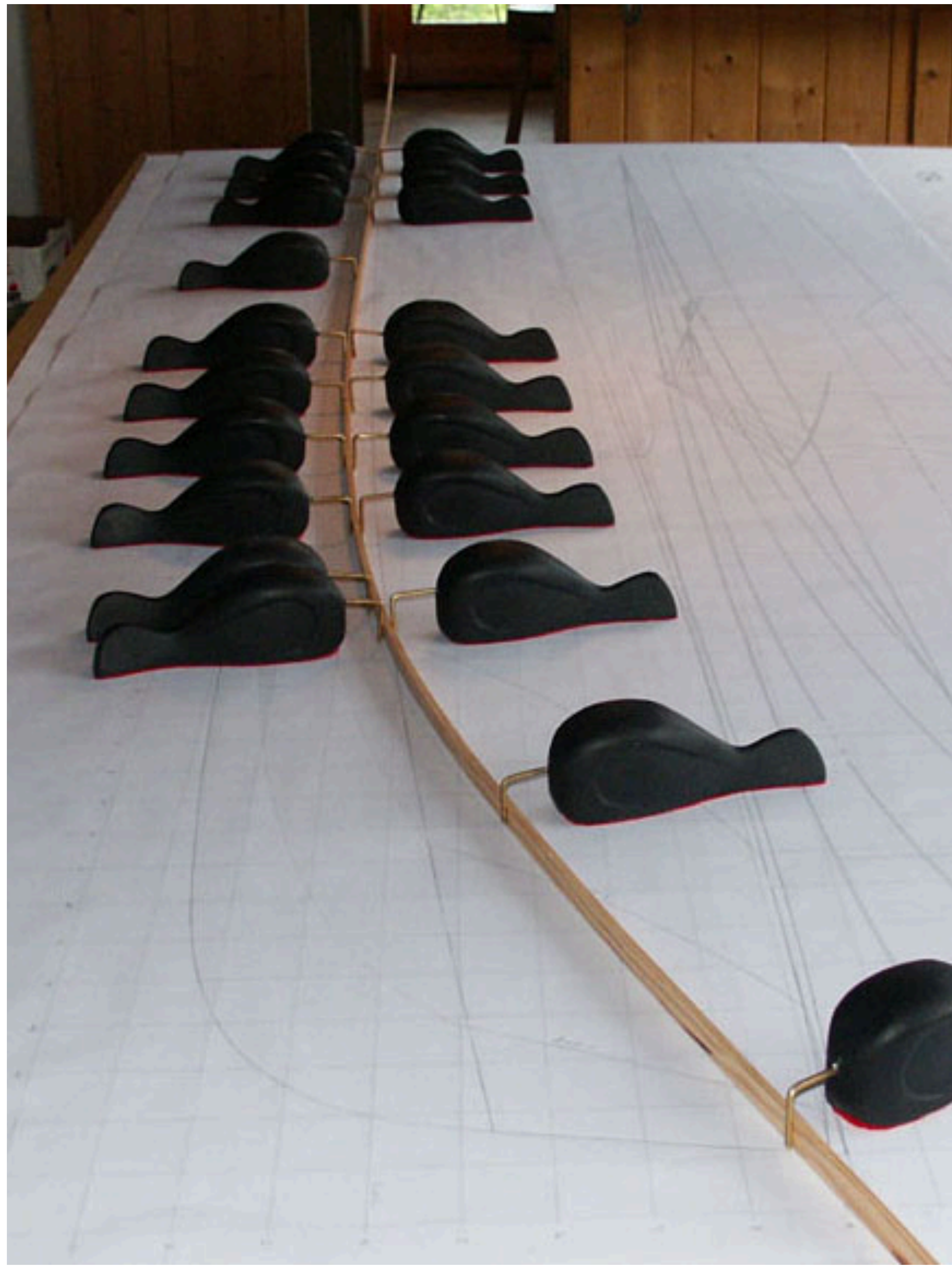


Prior, a ~ N(0,10)
doy first bloom
90
100
110
120
Posterior
doy first bloom
90
100
110
120
1000
1200
1400
1600
1800
2000
year

Going Local — B-Splines

B-Splines are just linear models, but with some weird synthetic variables:

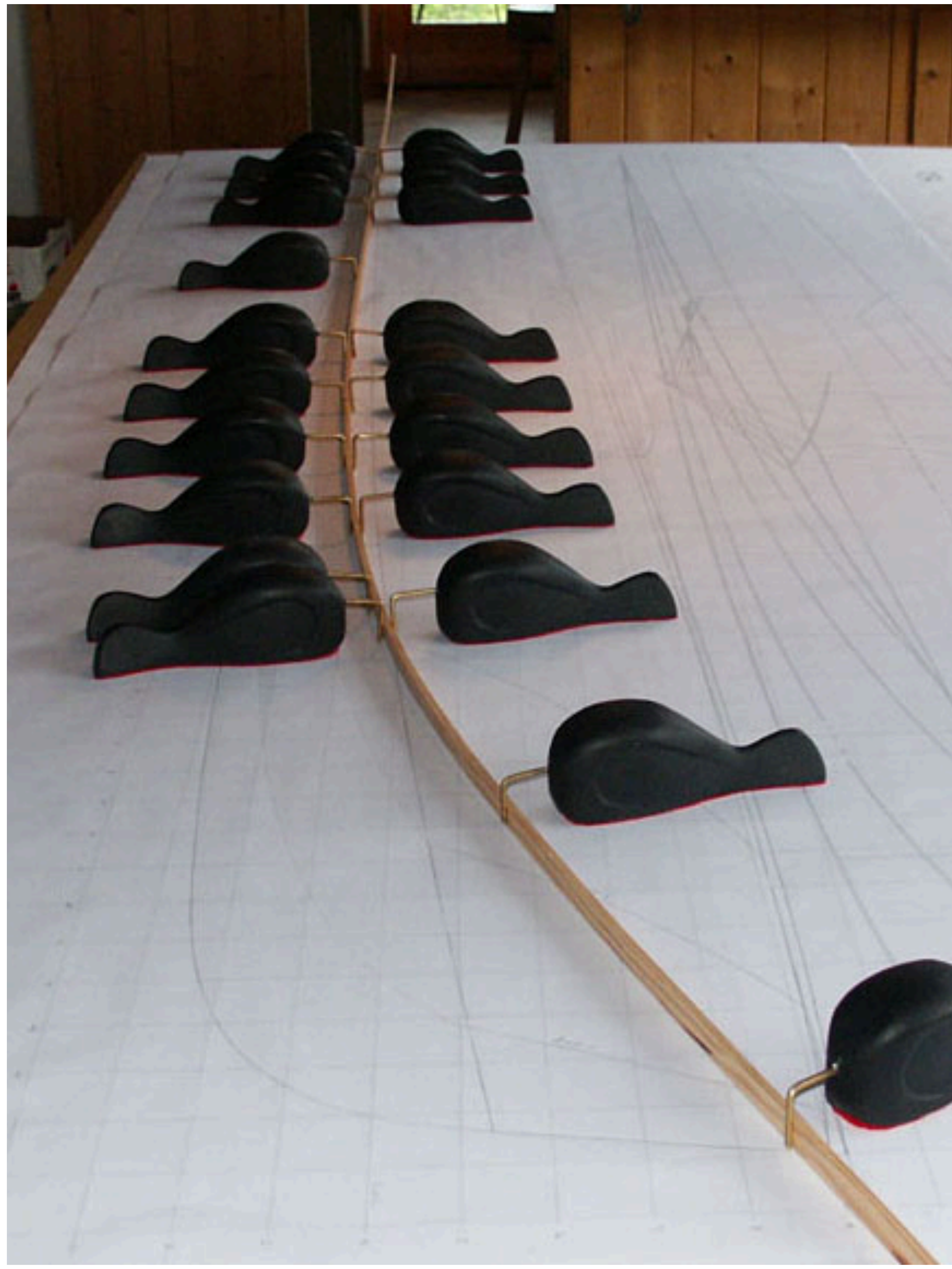
$$\mu_i = \alpha + w_1 B_{i,1} + w_2 B_{i,2} + w_3 B_{i,3} + \ldots$$



Going Local — B-Splines

B-Splines are just linear models, but with some weird synthetic variables:

$$\mu_i = \alpha + w_1 B_{i,1} + w_2 B_{i,2} + w_3 B_{i,3} + \ldots$$



Going Local — B-Splines

B-Splines are just linear models, but with some weird synthetic variables:

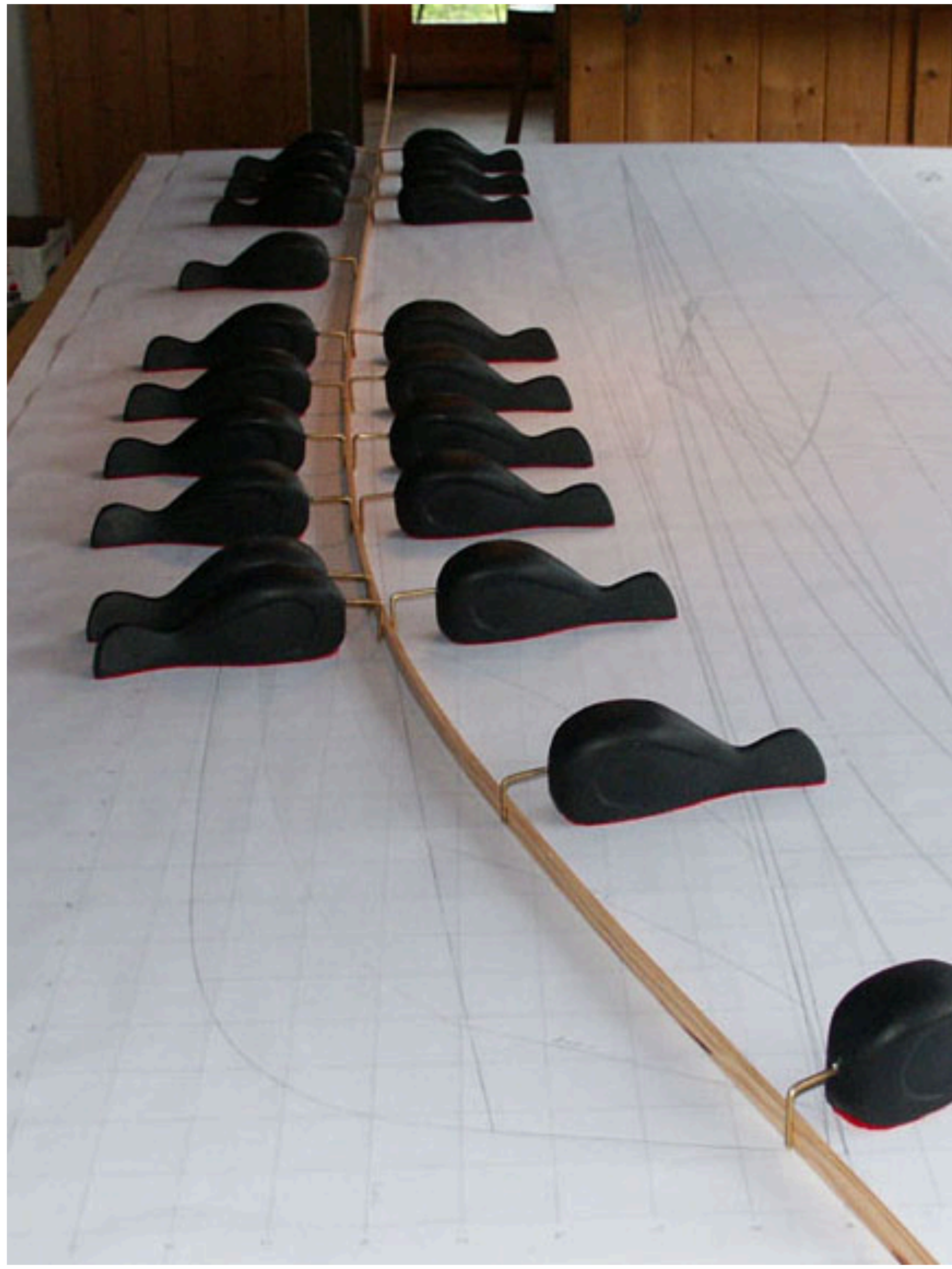
$$\mu_i = \alpha + w_1 B_{i,1} + w_2 B_{i,2} + w_3 B_{i,3} + \ldots$$

Weights w are like slopes

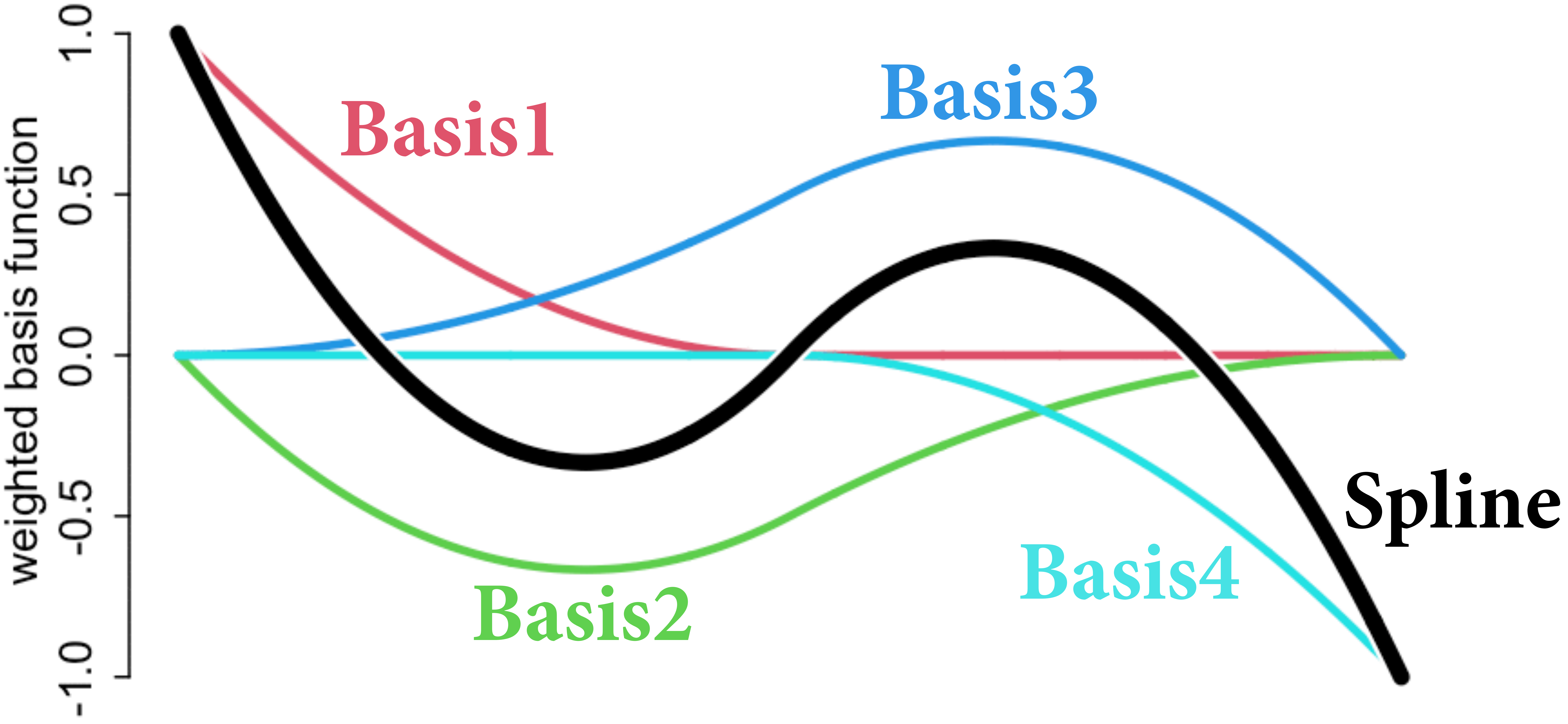
Basis functions B are synthetic variables

B values turn on weights in different regions

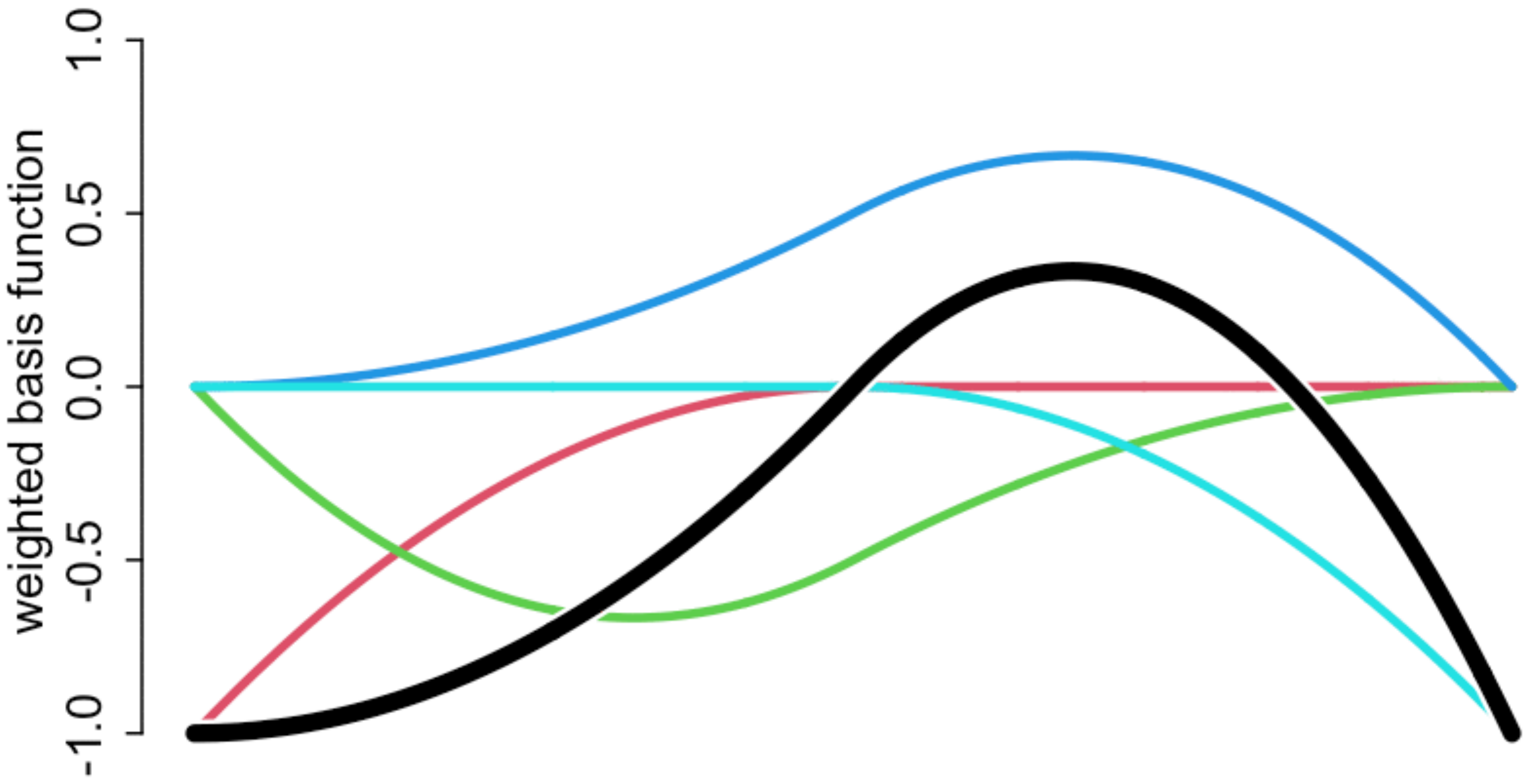
Detailed example starting on page 114



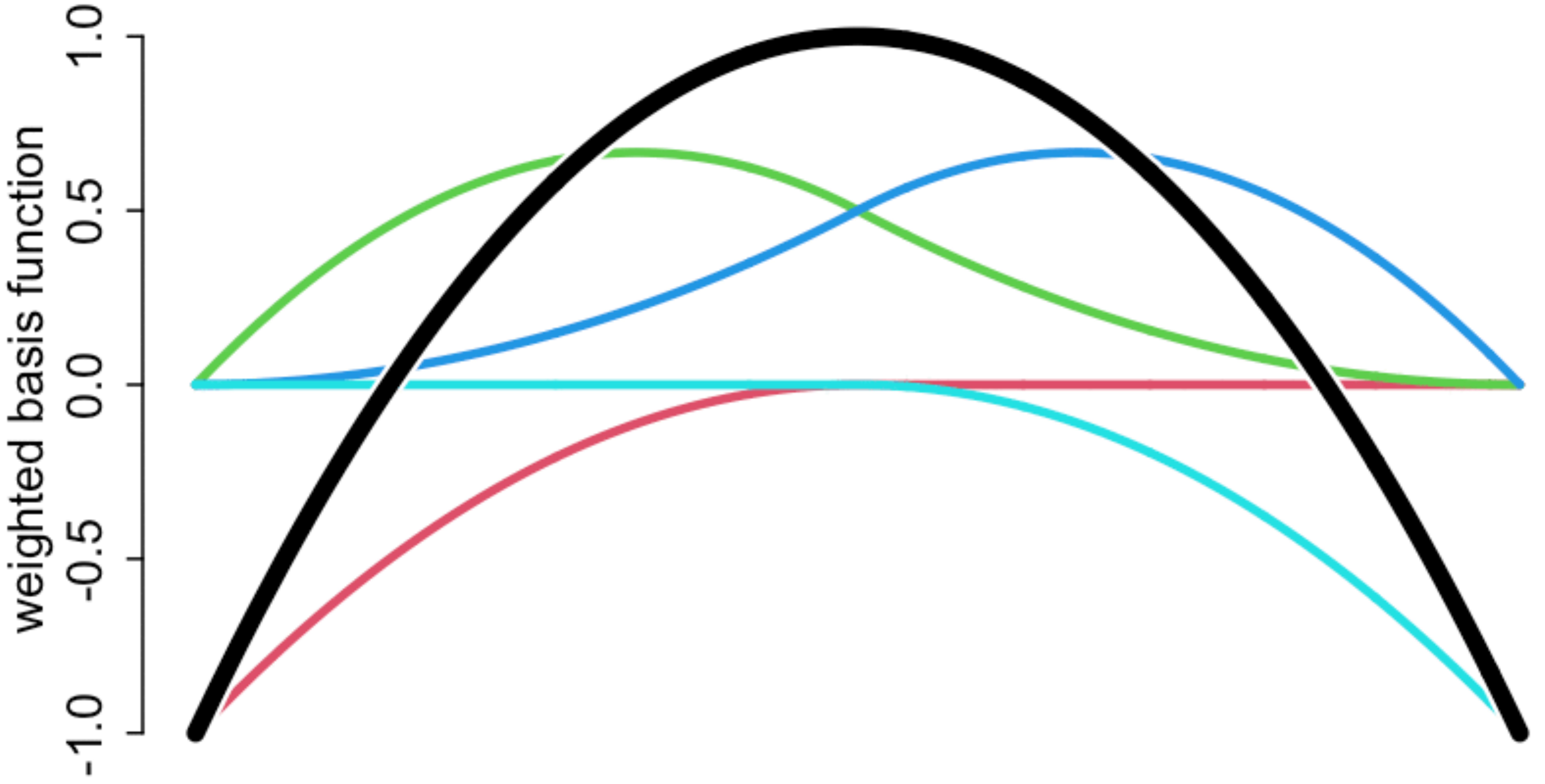
$$w = [1, -1, 1, -1]$$


weighted basis function
1.0
0.5
0.0
-0.5
-1.0
Basis1
Basis2
Basis3
Basis4
Spline

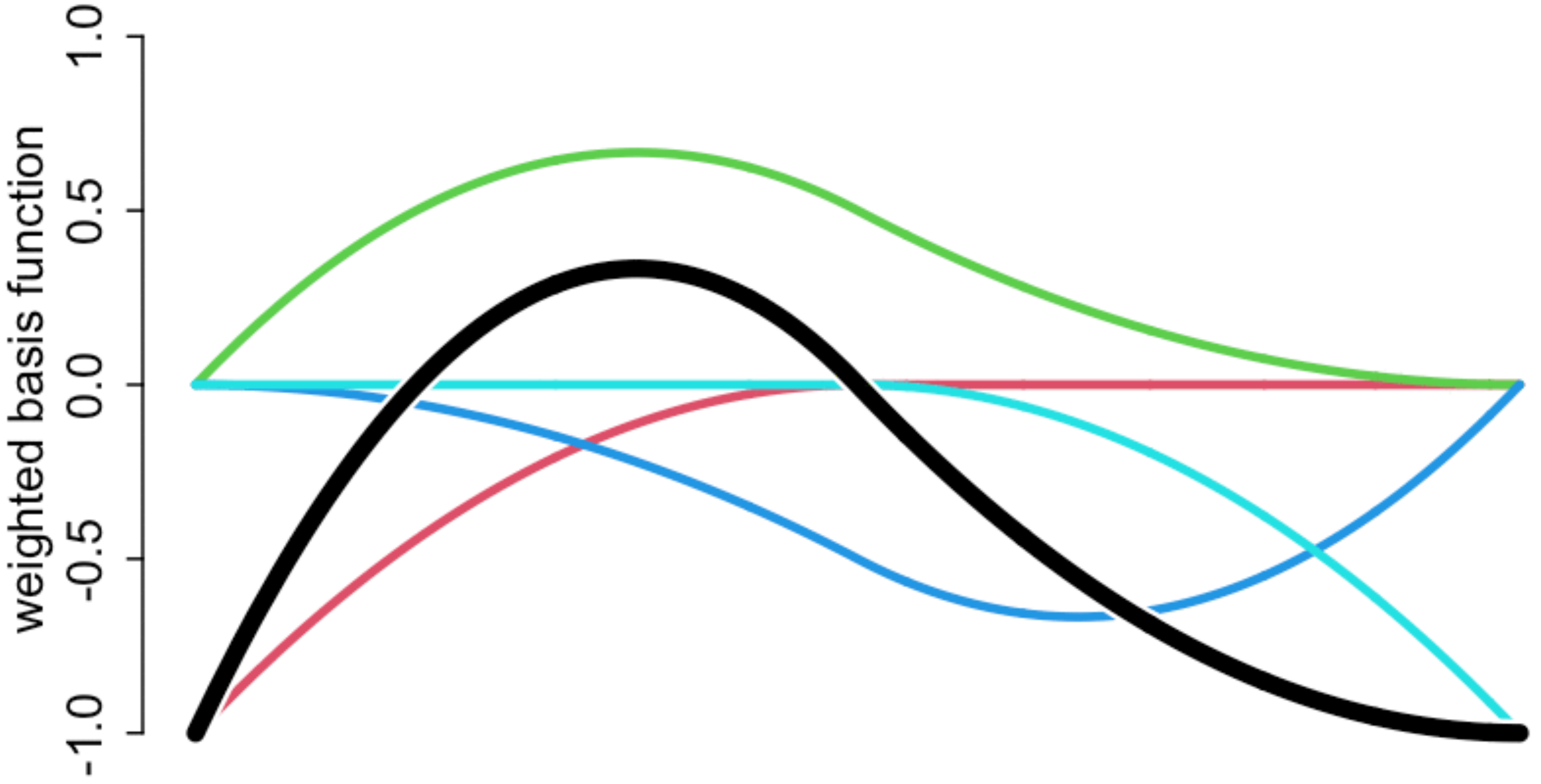
$$w = [-1, -1, 1, -1]$$



$$w = [-1, 1, 1, -1]$$



$w = [-1, 1, -1, -1]$



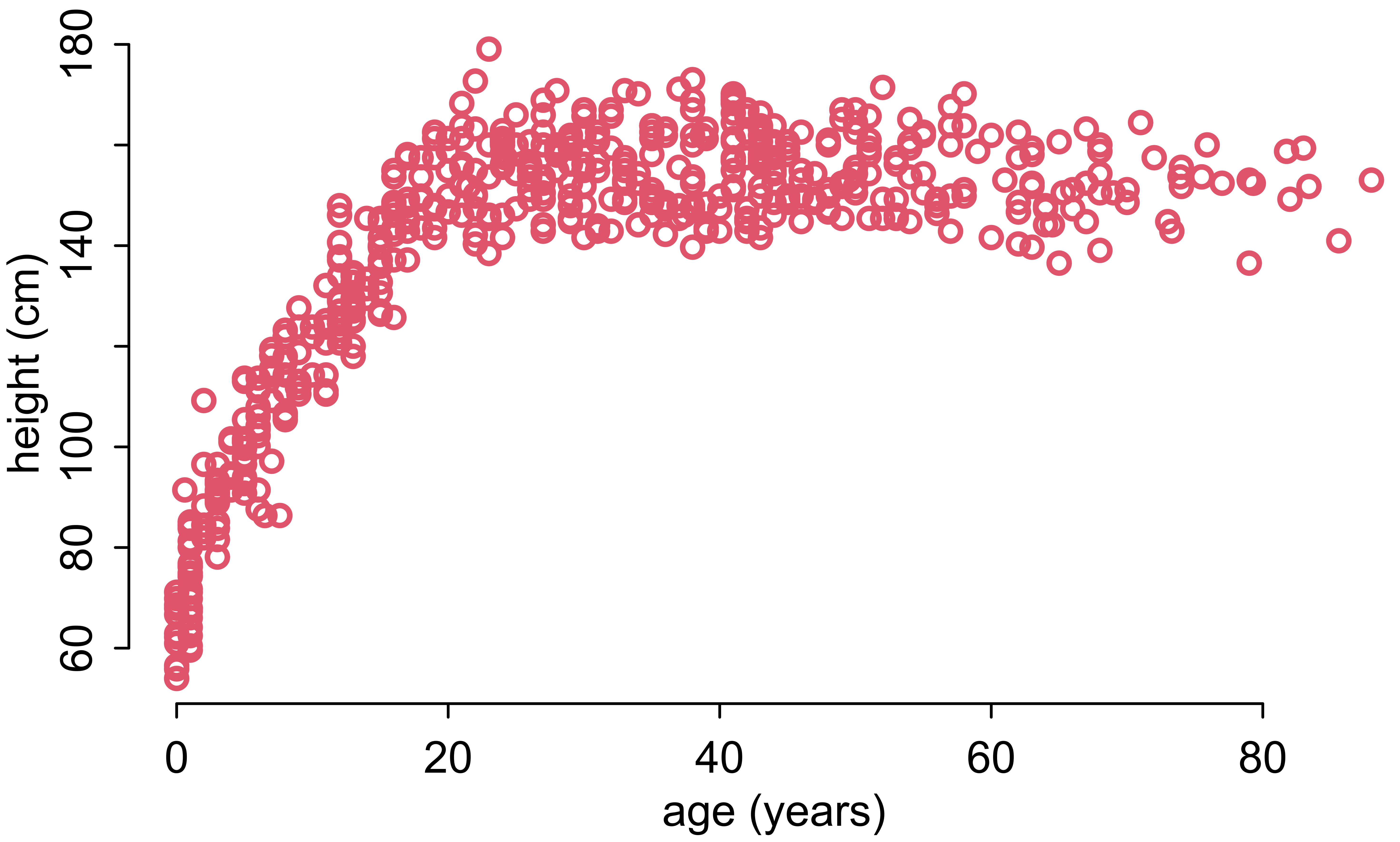
Height as a function of age

Obviously not linear

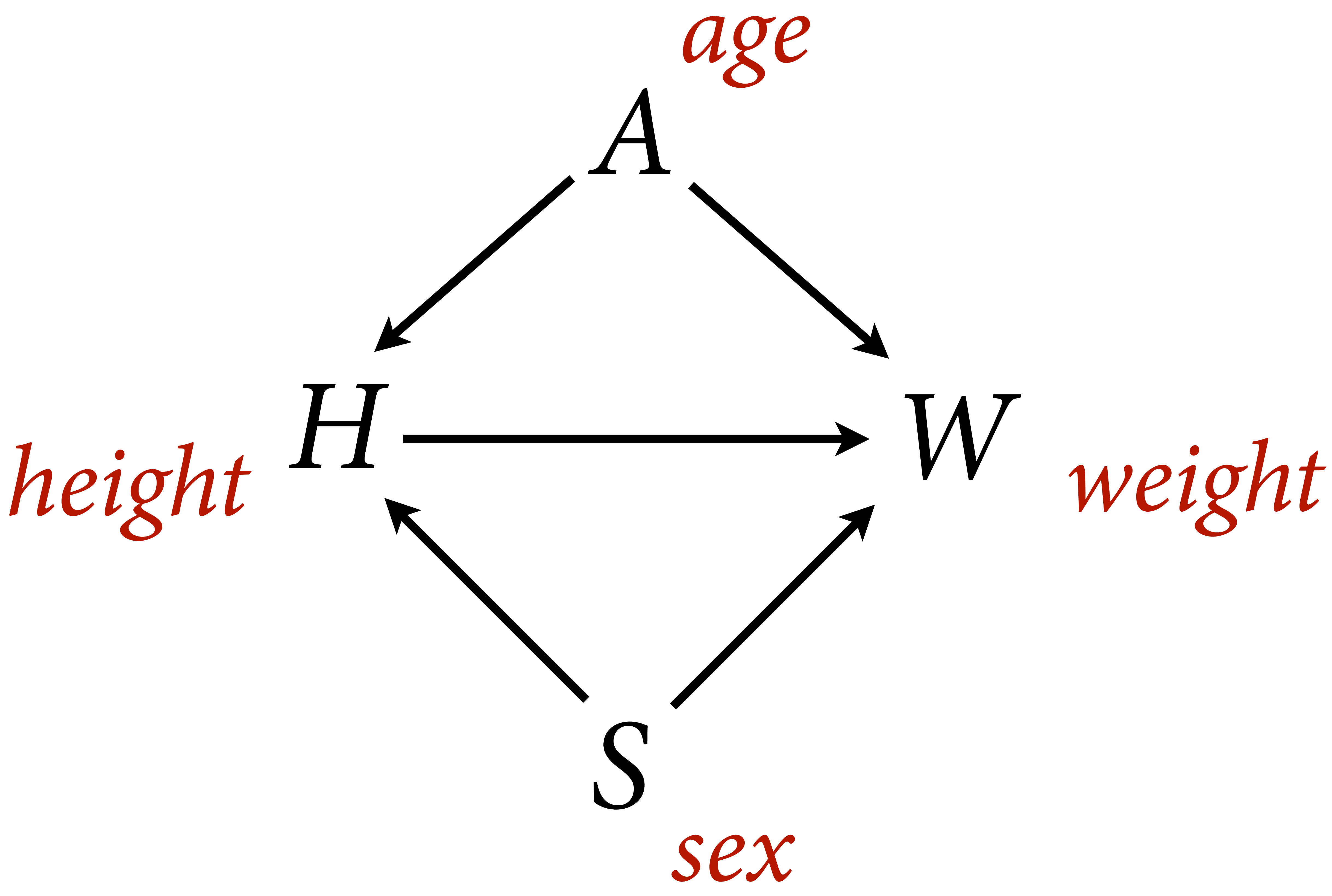
Want function approximately right

Fit spline as example

But biological model would do a lot better



Height as a function of age

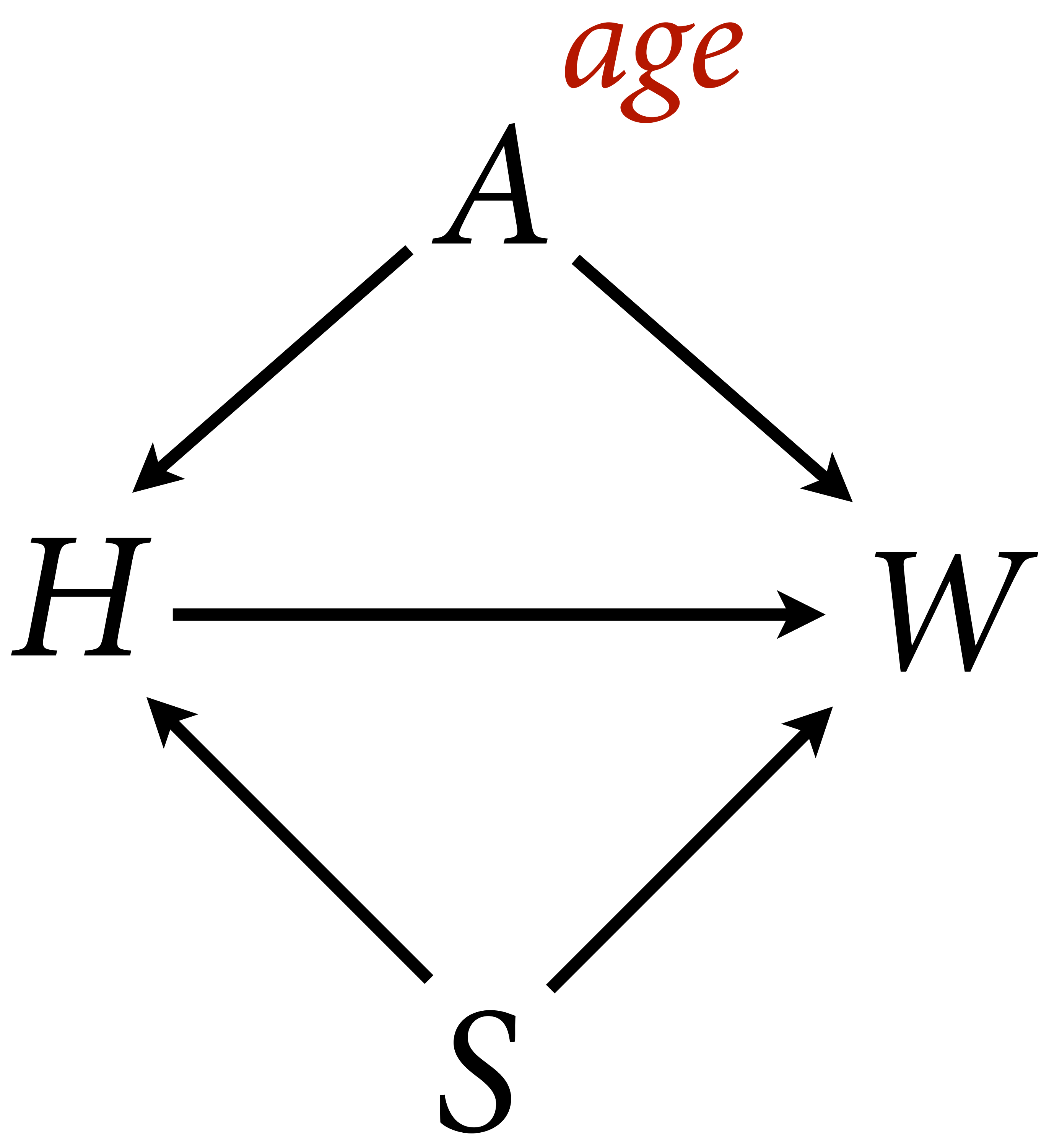


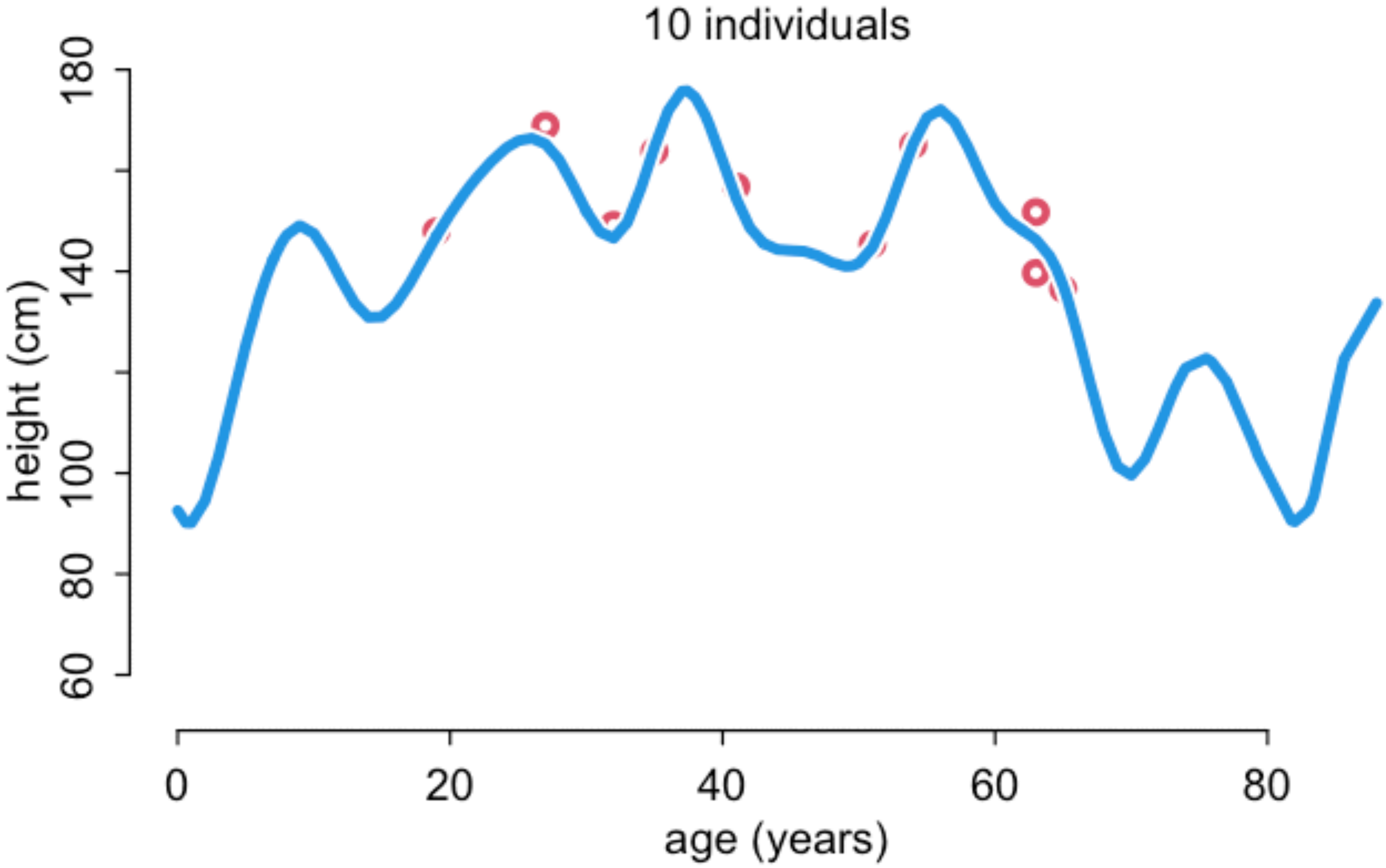
Height as a function of age

How can age cause anything?

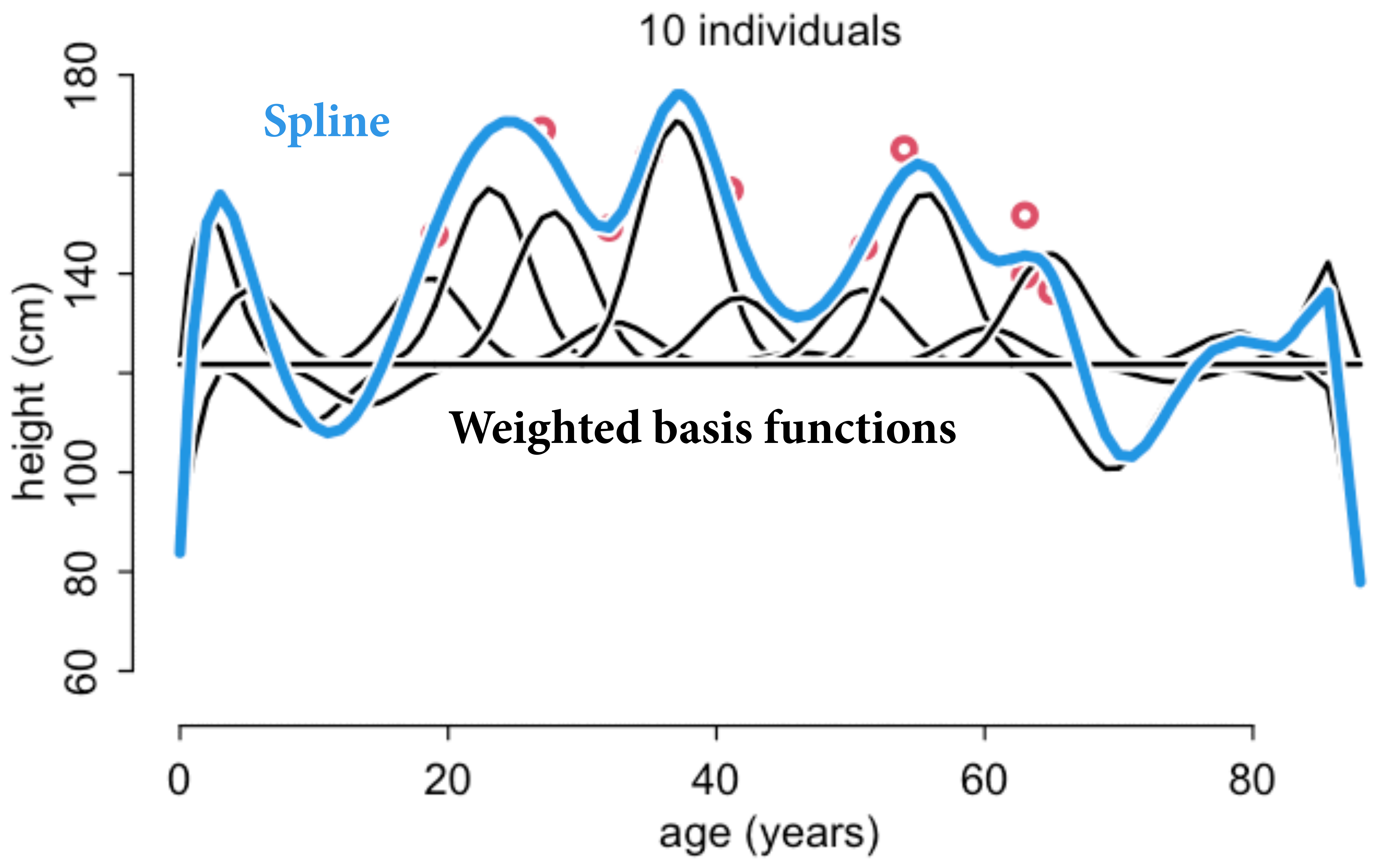
No definable intervention?

Association between age and height due to accumulated growth





10 individuals
height (cm)
180
140
100
80
60
0
20
40
60
80
age (years)



10 individuals
Spline
Weighted basis functions
height (cm)
180
140
100
80
60
0
20
40
60
80
age (years)

Curves and Splines

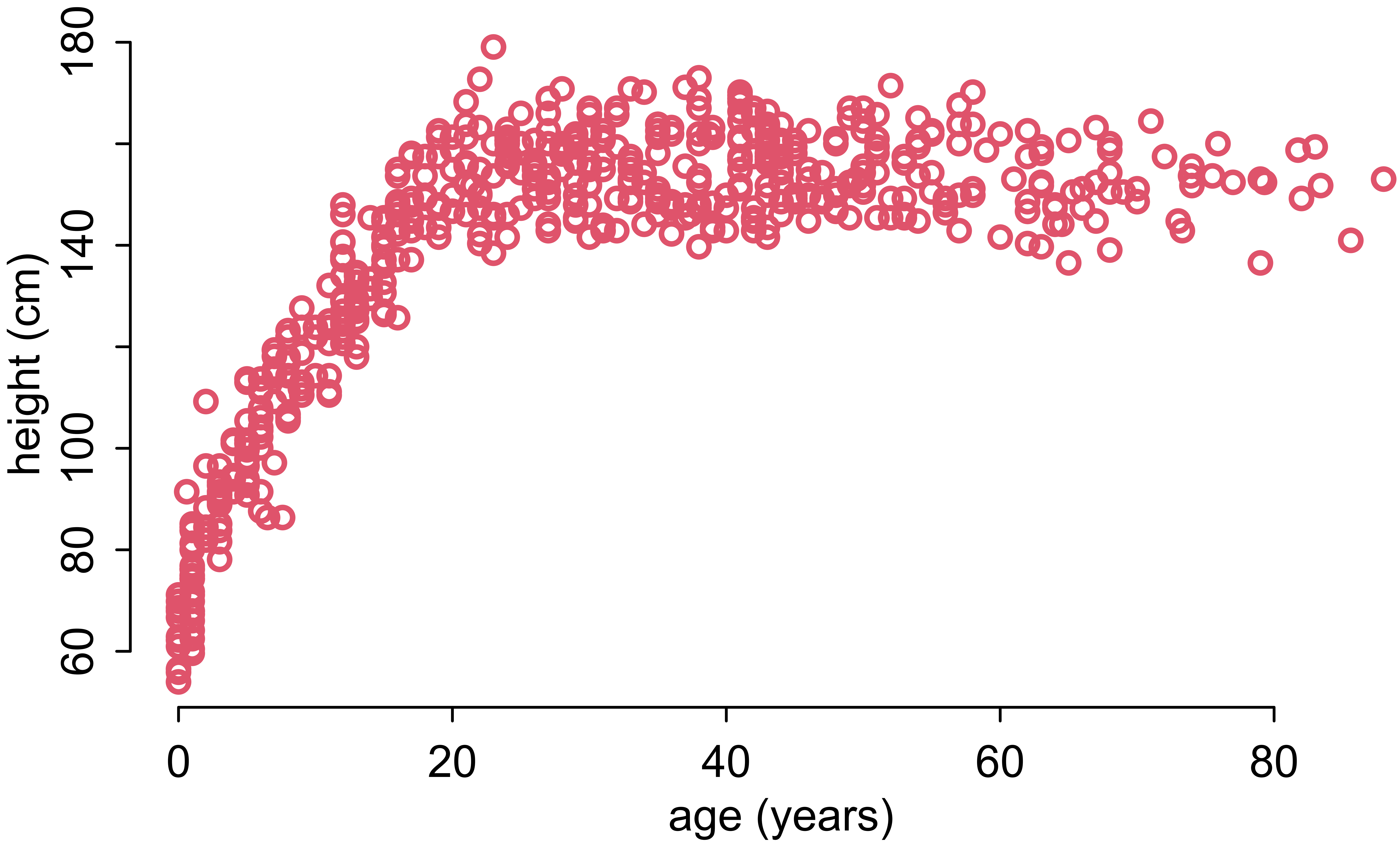
Can build very non-linear functions from linear pieces

Splines are powerful geocentric devices

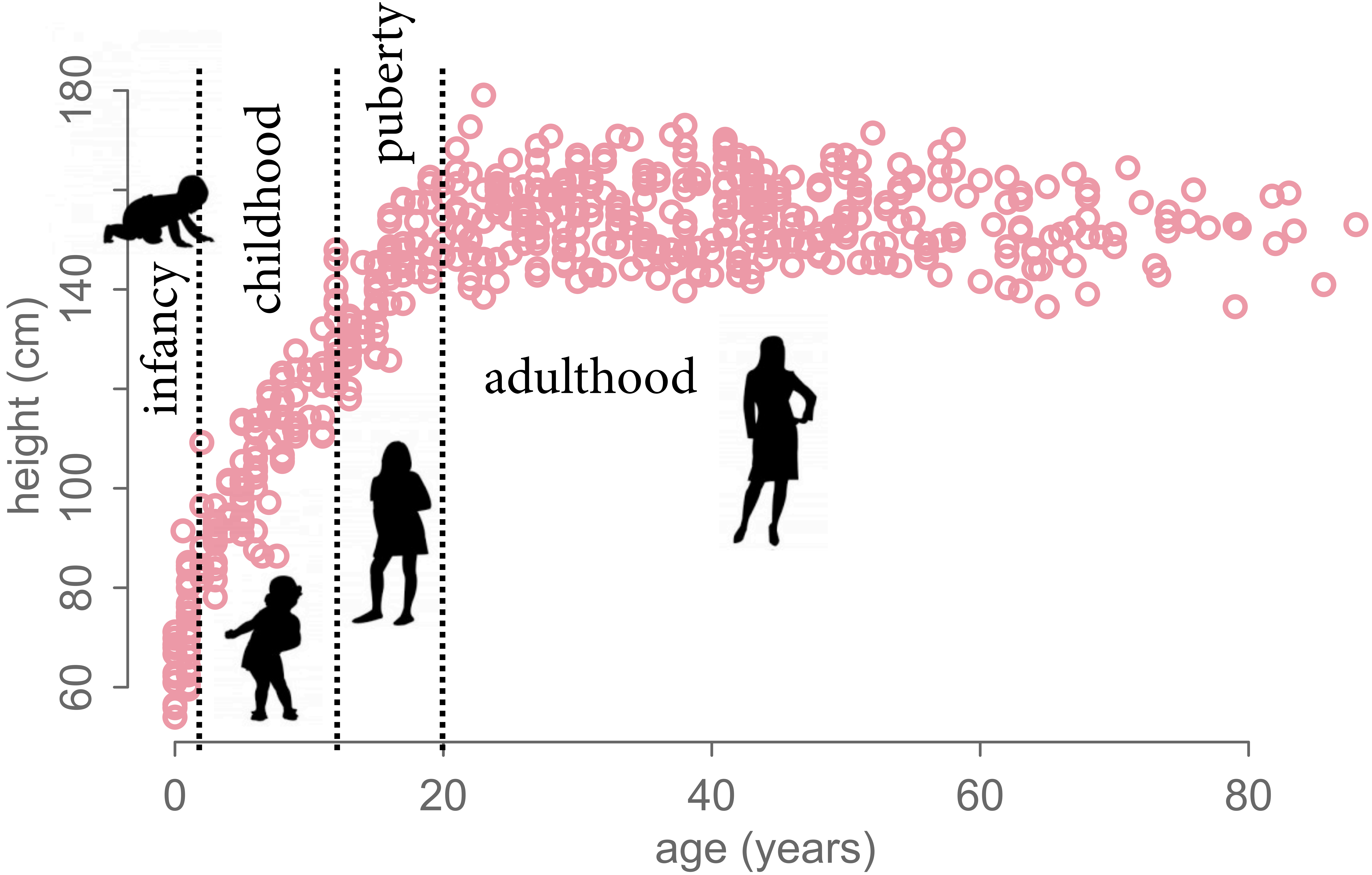
Adding scientific information helps

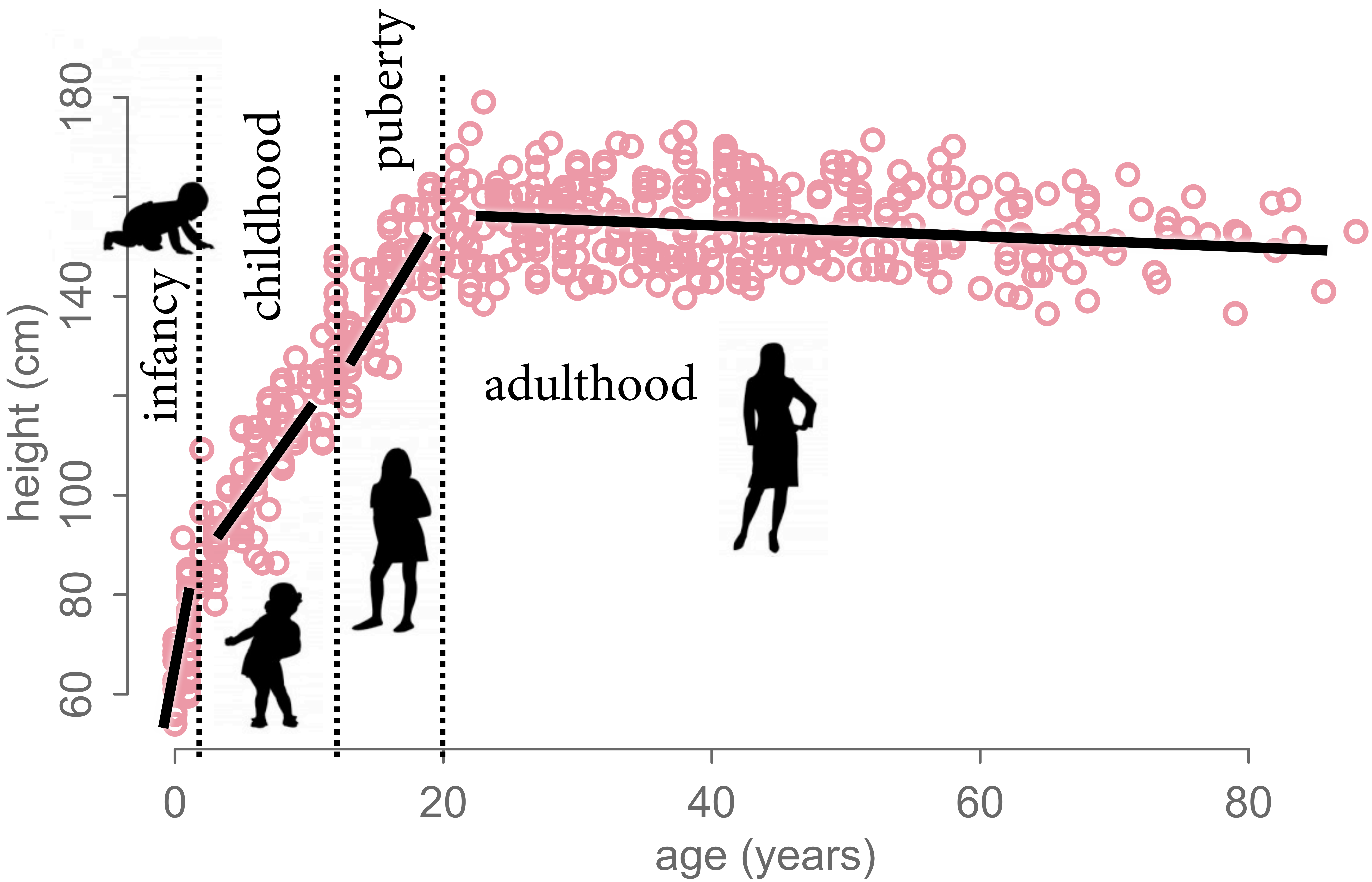
e.g. Average weight only increases with height
e.g. Height increases with age, then levels off (or declines)

Ideally **statistical** model has some form as **scientific** model



height (cm)
180
140
100
80
60
0
20
40
60
80
age (years)


height (cm)
180
140
100
80
60
infancy
childhood
puberty
adulthood
0
20
40
60
80
age (years)



infancy
childhood
puberty
adulthood
height (cm)
60
80
100
140
180
0
20
40
60
80
age (years)

Course Schedule

Week 1	Bayesian inference	Chapters 1, 2, 3
Week 2	Linear models & Causal Inference	Chapter 4
Week 3	Causes, Confounds & Colliders	Chapters 5 & 6
Week 4	Overfitting / Interactions	Chapters 7 & 8
Week 5	MCMC & Generalized Linear Models	Chapters 9, 10, 11
Week 6	Integers & Other Monsters	Chapters 11 & 12
Week 7	Multilevel models I	Chapter 13
Week 8	Multilevel models II	Chapter 14
Week 9	Measurement & Missingness	Chapter 15
Week 10	Generalized Linear Madness	Chapter 16

https://github.com/rmcelreath/stat_rethinking_2023

BONUS

Full Luxury Bayes

We used two models for two estimands

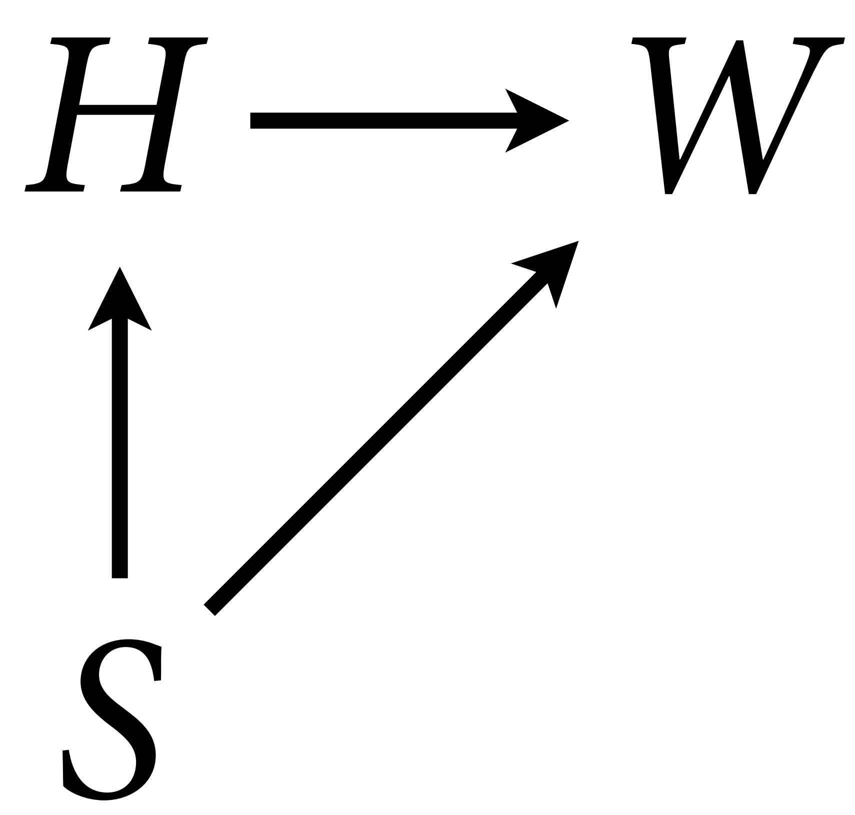
But alternative and equivalent approach is to use one model of entire causal system

Then use joint posterior to compute each estimand

$$\begin{tikzcd} H \arrow[r] & W \\ S \arrow[u] \arrow[ur] & \end{tikzcd}$$

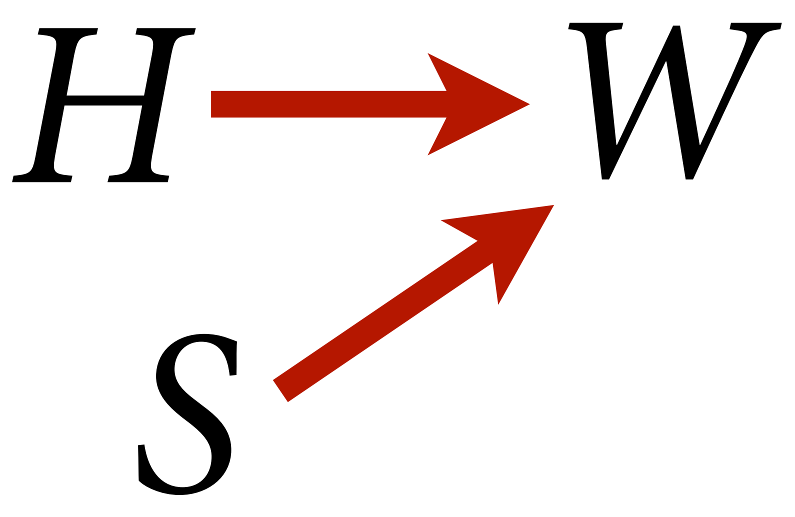
Full Luxury Bayes

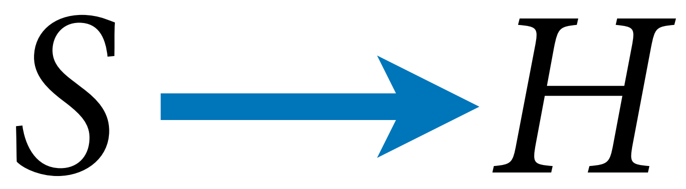
```
m_SHW_full <- quap(
    alist(

        # weight
        W ~ dnorm(mu,sigma),
        mu <- a[S] + b[S]*(H-Hbar),
        a[S] ~ dnorm(60,10),
        b[S] ~ dunif(0,1),
        sigma ~ dunif(0,10),

        # height
        H ~ dnorm(nu,tau),
        nu <- h[S],
        h[S] ~ dnorm(160,10),
        tau ~ dunif(0,10)

    ), data=dat )
```

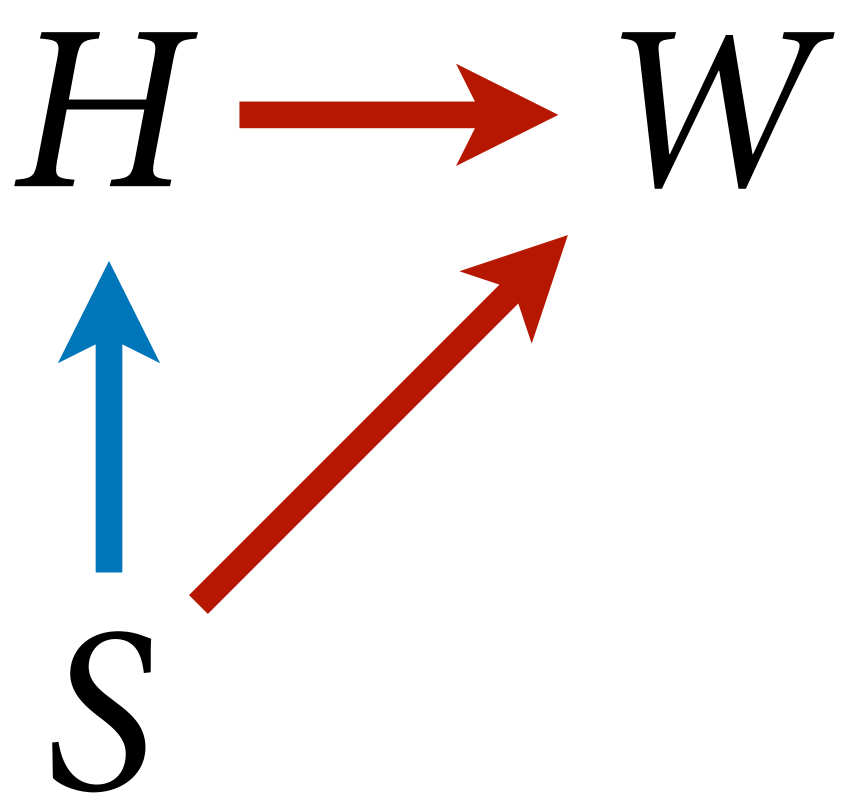


Full Luxury Bayes

```
m_SHW_full <- quap(
    alist(

        # weight
        W ~ dnorm(mu,sigma),
        mu <- a[S] + b[S]*(H-Hbar),
        a[S] ~ dnorm(60,10),
        b[S] ~ dunif(0,1),
        sigma ~ dunif(0,10),

        # height
        H ~ dnorm(nu,tau),
        nu <- h[S],
        h[S] ~ dnorm(160,10),
        tau ~ dunif(0,10)

    ), data=dat )
```







Full Luxury Bayes

```
m_SHW_full <- quap(
    alist(

        # weight
        W ~ dnorm(mu,sigma),
        mu <- a[S] + b[S]*(H-Hbar),
        a[S] ~ dnorm(60,10),
        b[S] ~ dunif(0,1),
        sigma ~ dunif(0,10),

        # height
        H ~ dnorm(nu,tau),
        nu <- h[S],
        h[S] ~ dnorm(160,10),
        tau ~ dunif(0,10)

    ), data=dat )
```

```
> precis(m_SHW_full,depth=2)
        mean   sd   5.5%  94.5%
a[1]   45.17 0.44  44.47  45.87
a[2]   45.09 0.46  44.37  45.82
h[1]  149.53 0.40 148.89 150.18
h[2]  160.36 0.43 159.67 161.04
b[1]    0.66 0.06   0.56   0.75
b[2]    0.61 0.05   0.52   0.70
sigma   4.23 0.16   3.97   4.48
tau     5.52 0.21   5.19   5.85
>
```

Causal effect is consequence of intervention

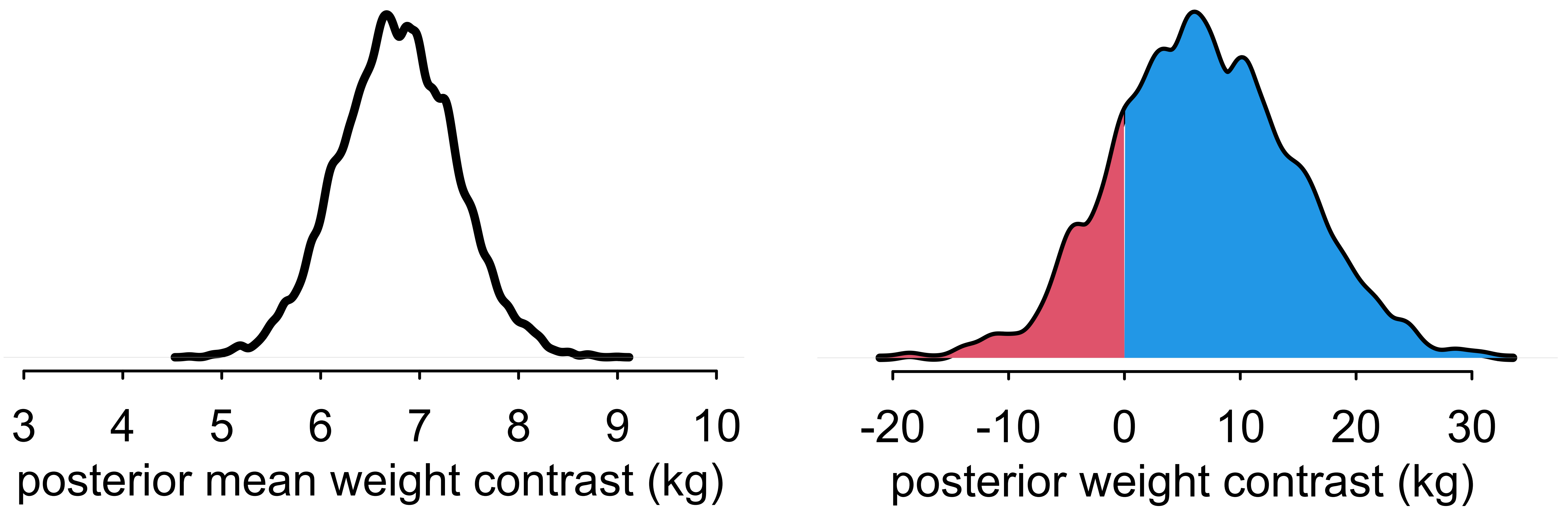
Now simulate each intervention

Simulating Interventions

```
post <- extract.samples(m_SHW_full)
Hbar <- dat$Hbar
n <- 1e4

with( post , {
# simulate W for S=1
H_S1 <- rnorm(n, h[,1] , tau )
W_S1 <- rnorm(n, a[,1] +
  b[,1]*(H_S1-Hbar) , sigma)

# simulate W for S=2
H_S2 <- rnorm(n, h[,2] , tau)
W_S2 <- rnorm(n, a[,2] +
  b[,2]*(H_S2-Hbar) , sigma)

# compute contrast
W_do_S <<- W_S2 - W_S1
})
```

Total causal effect of S on W:
Consequence of changing S at birth

"$p(W|\text{do}(S))$"

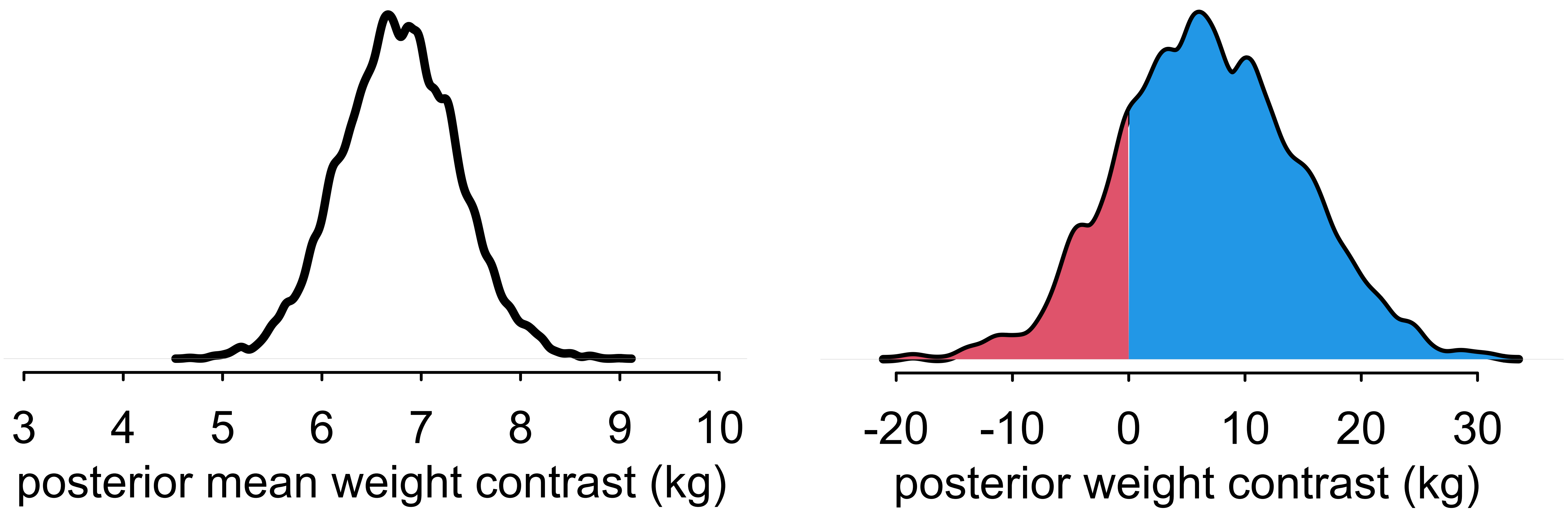


Simulating Interventions

```
post <- extract.samples(m_SHW_full)
Hbar <- dat$Hbar
n <- 1e4

with( post , {
# simulate W for S=1
H_S1 <- rnorm(n, h[,1] , tau )
W_S1 <- rnorm(n, a[,1] +
  b[,1]*(H_S1-Hbar) , sigma)

# simulate W for S=2
H_S2 <- rnorm(n, h[,2] , tau)
W_S2 <- rnorm(n, a[,2] +
  b[,2]*(H_S2-Hbar) , sigma)

# compute contrast
W_do_S <<- W_S2 - W_S1
})
```

```
# automated way
HWsim <- sim(m_SHW_full,
             data=list(S=c(1,2)),
             vars=c("H","W"))
W_do_S_auto <- HWsim$W[,2] - HWsim$W[,1]
```

"$p(W|\mathrm{do}(S))$"



Inference With Linear Models

With more than two variables, scientific (causal) model and statistical model not always same

(1) State each estimand
(2) Design **unique statistical model** for each
(3) **Compute** each estimand

ONE STAT MODEL FOR EACH ESTIMAND

Or -

(1) State each estimand
(2) Compute joint posterior for **causal system**
(3) **Simulate** each estimand as an **intervention**

ONE SIMULATION FOR EACH ESTIMAND