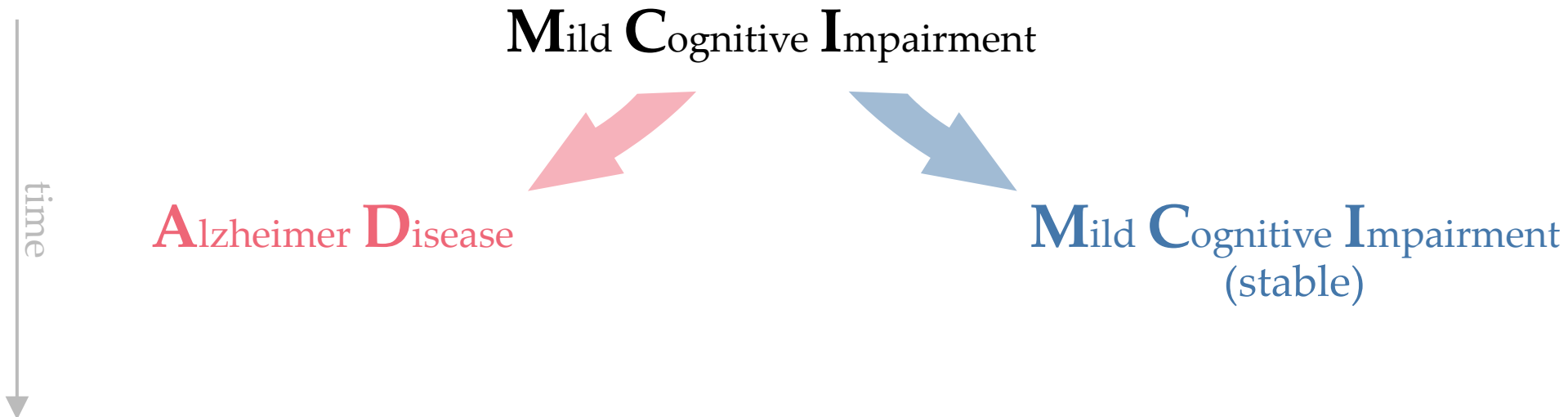
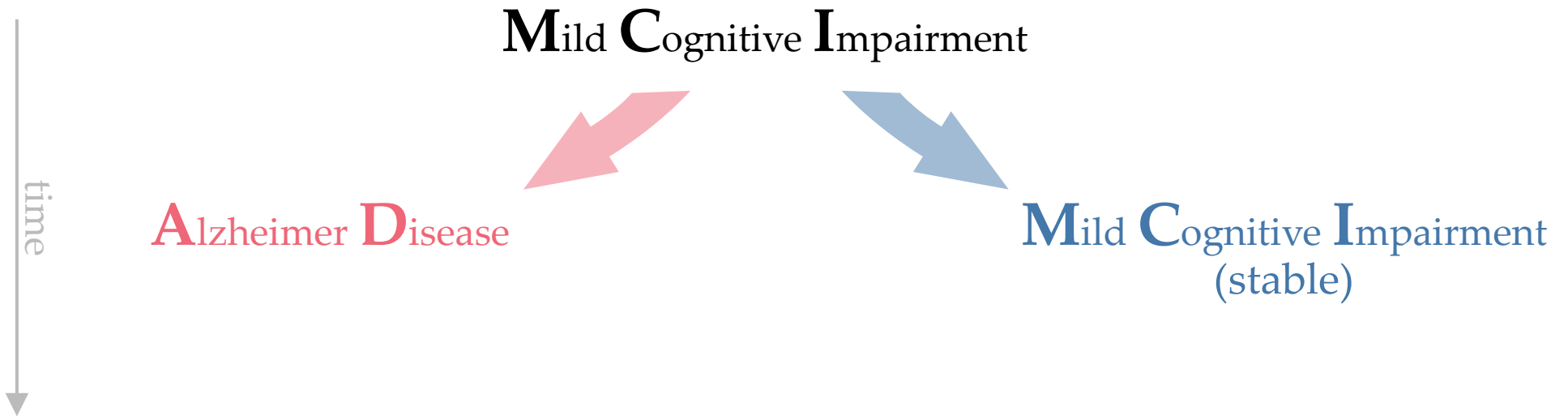


Mild Cognitive Impairment





♂ Gender

♂ AGE

🕒 ANARTERR

🕒 GDTOTAL

🕒 RAVLT

🕒 CATANIMSC

🕒 TRABSCOR

🕒 AVDELTOT

🕒 TRAASCOR

🕒 AVDEL30MIN

🧠 LRHHC

🧠 LRLV

🧬 Apoe4

👤 FAQ

♂ Gender

♂ AGE

🕒 ANARTERR

🕒 GDTOTAL

🕒 RAVLT

🕒 CATANIMSC

🕒 TRABSCOR

🕒 AVDELTOT

🕒 TRAASCOR

🕒 AVDEL30MIN

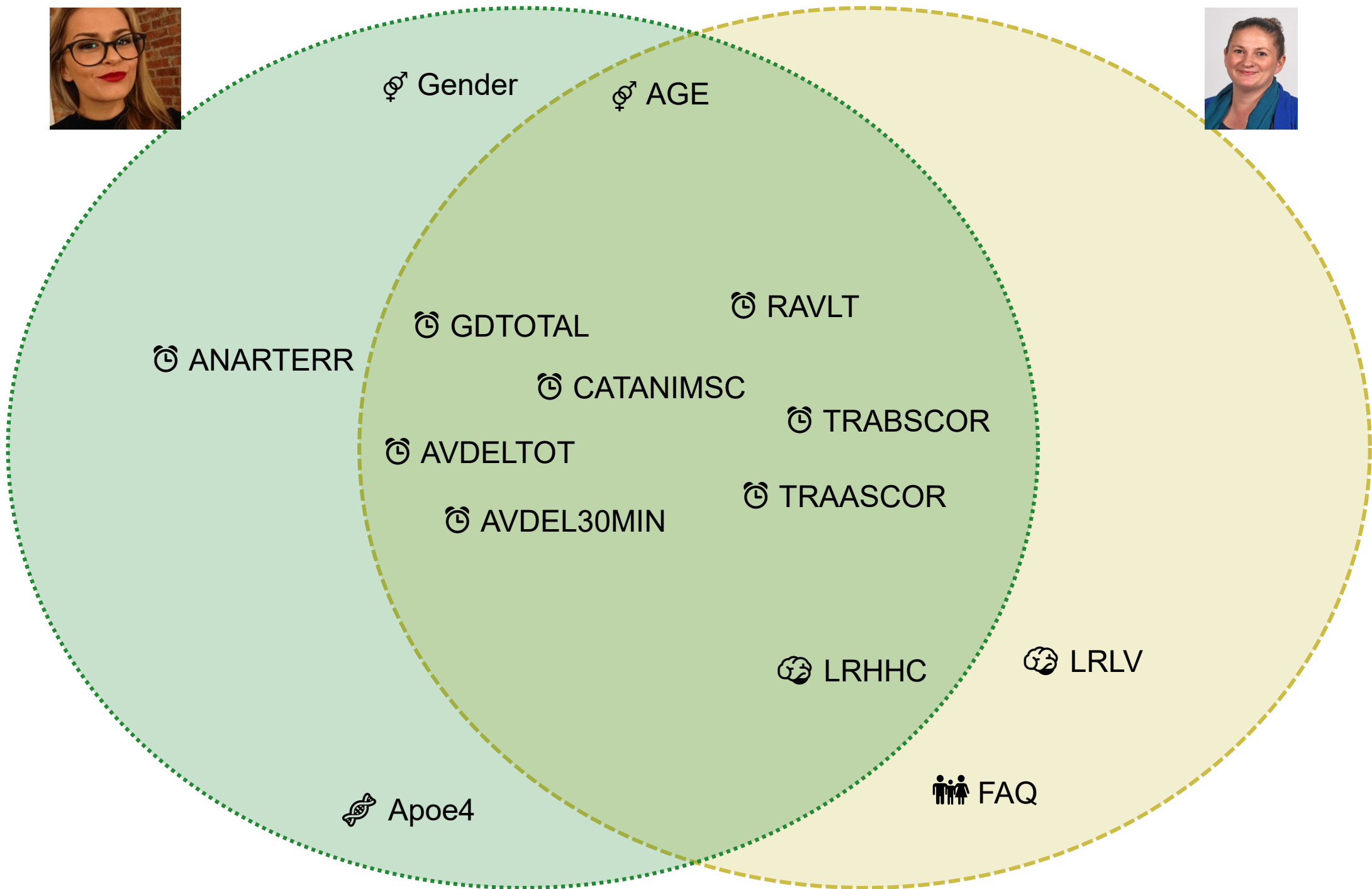
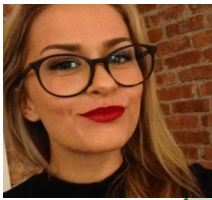
🧠 LRHHC

🧠 LRLV

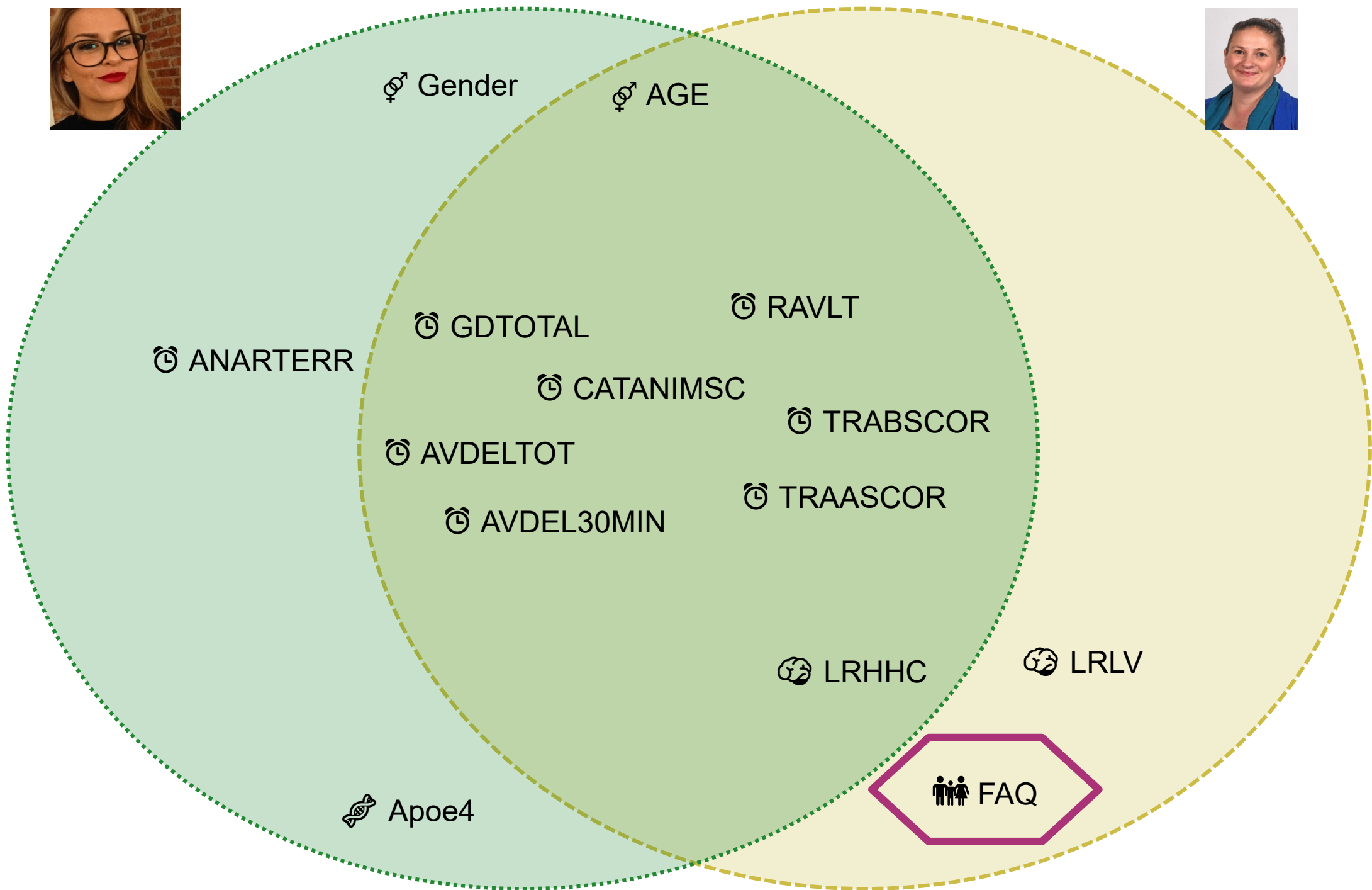
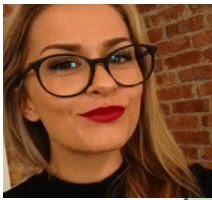
🧬 Apoe4

👤 FAQ

How 'good' are these features at prognosing the later onset of Alzheimer?

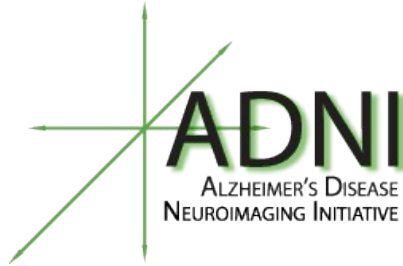


How 'good' are these features at prognosing the later onset of Alzheimer?



How 'good' are these features at prognosing the later onset of Alzheimer?

Data source:

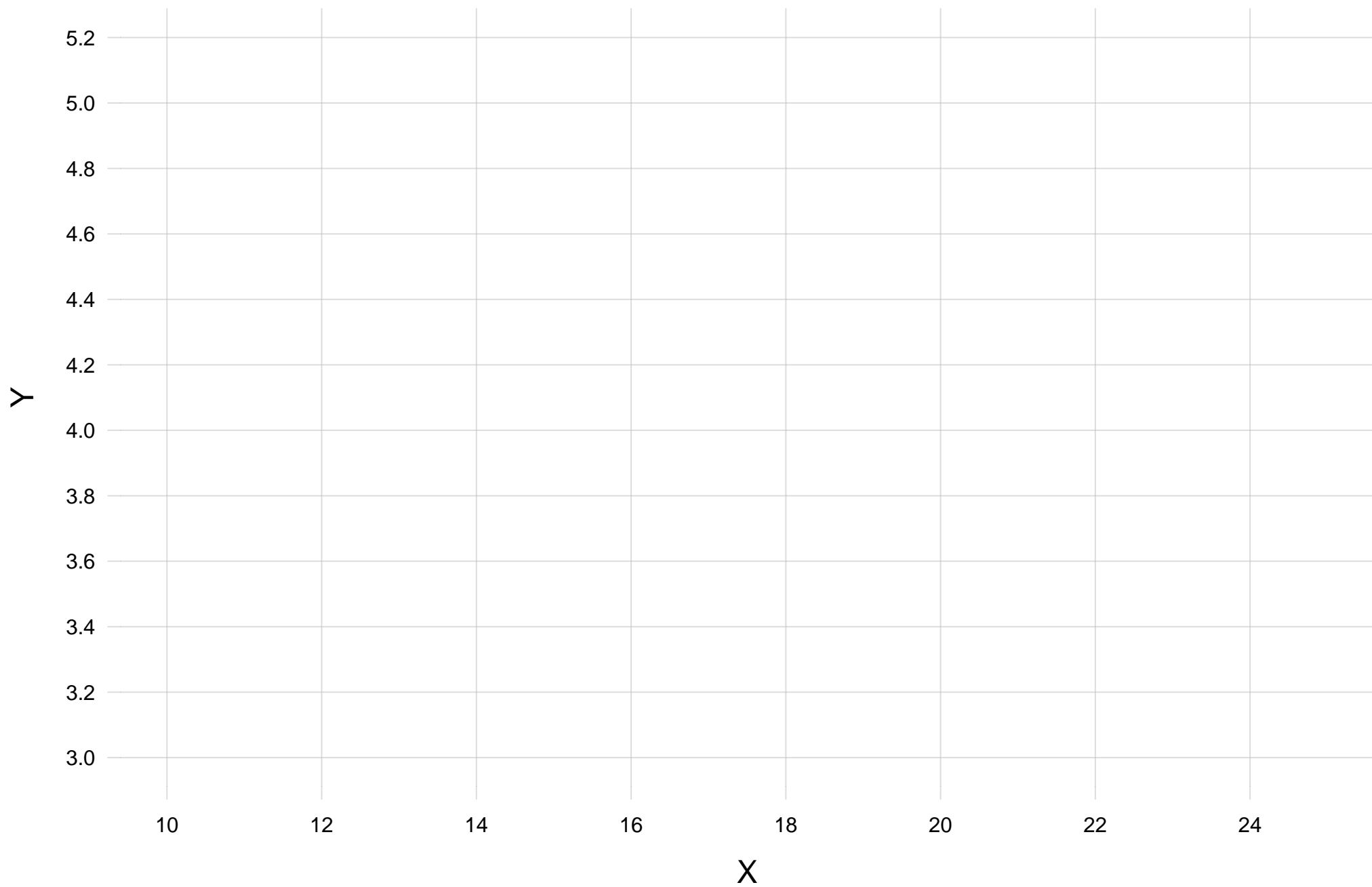


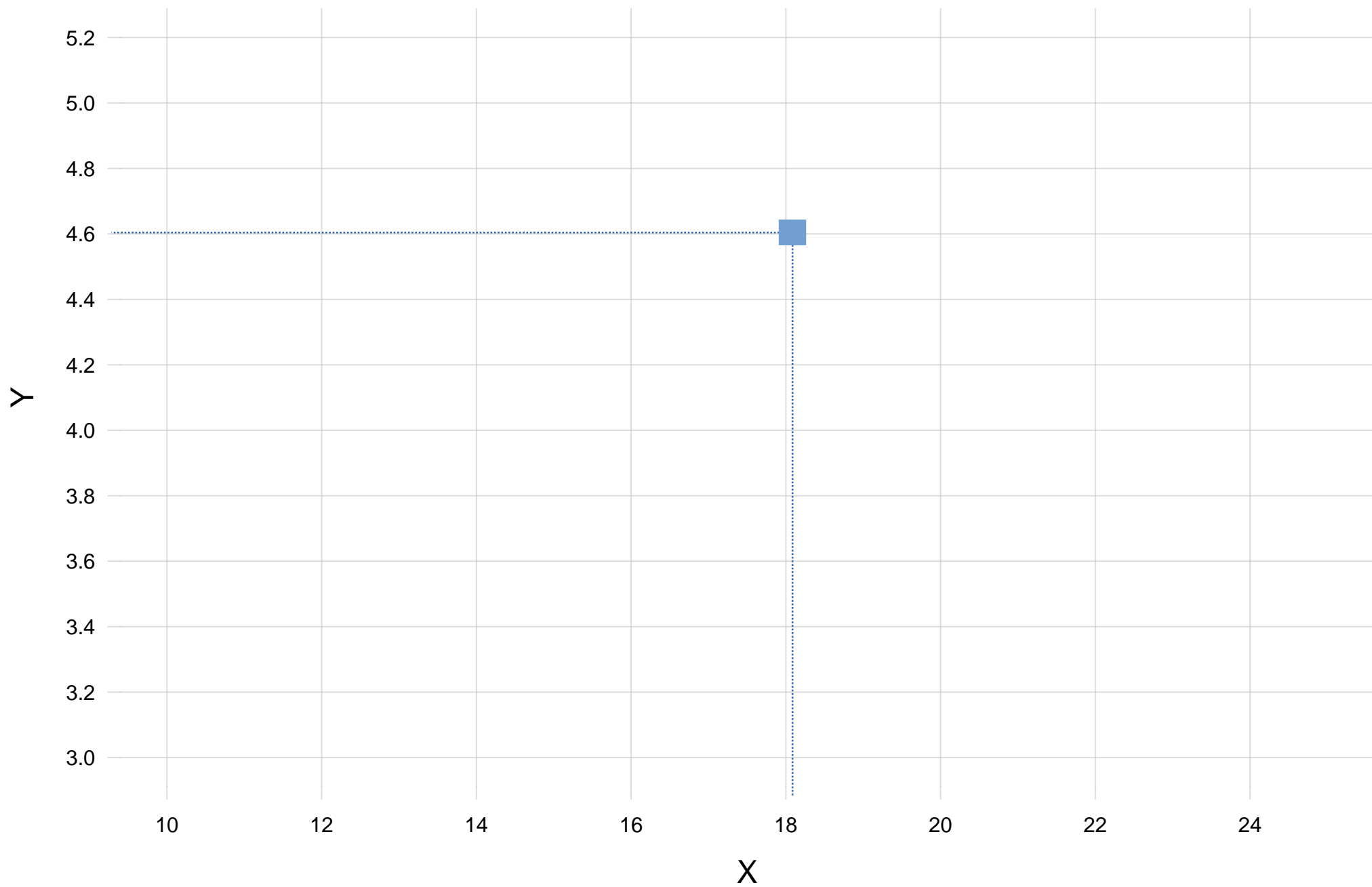
Ingrid's study: 12 + 1 variates, 678 datapoints

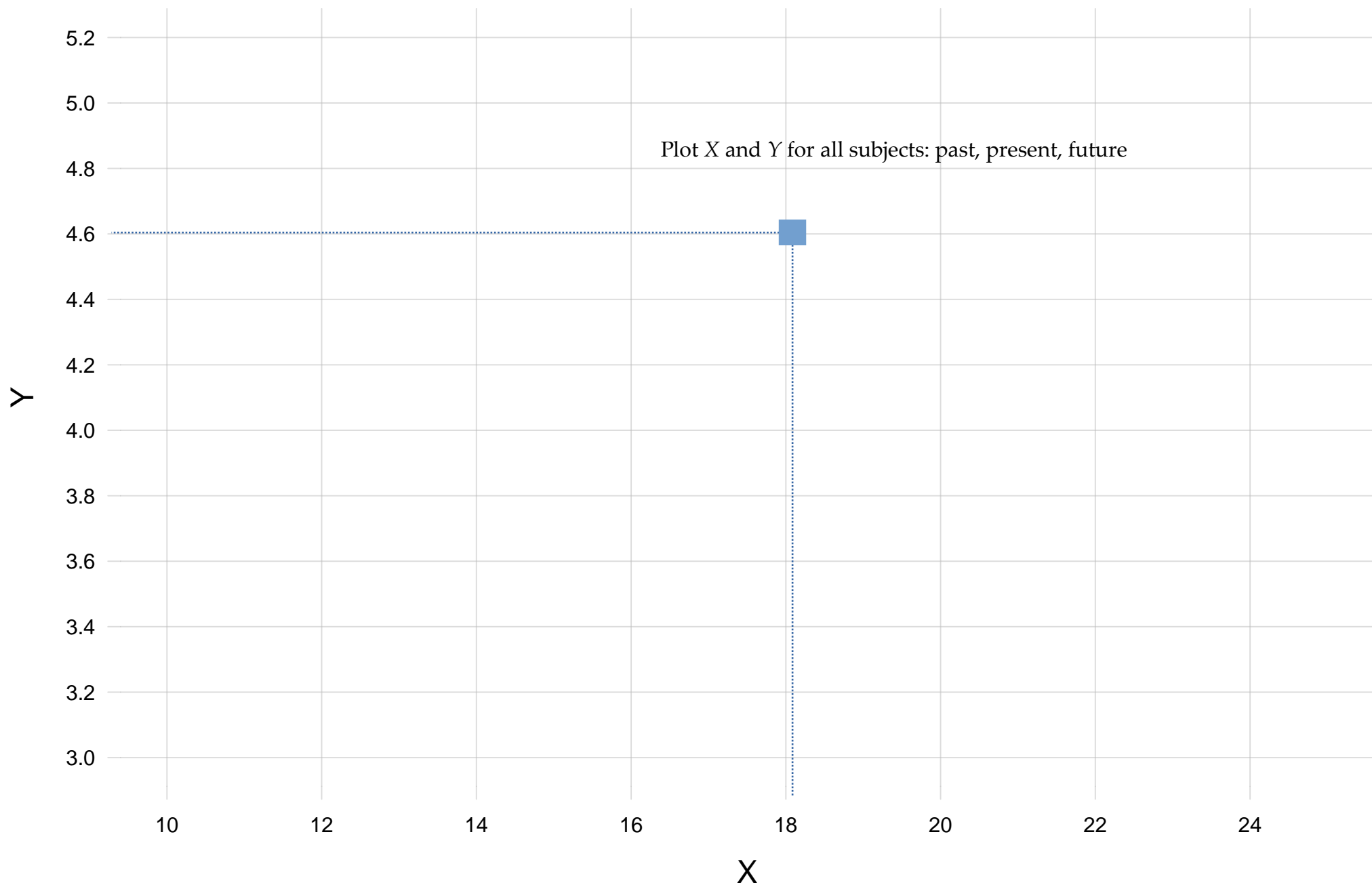
Alexandra's study: 11 + 1 variates, 708 datapoints (43 missing values)

Computation time: ~65 h/study (3 parallel sessions to assess numeric convergence)

HPC **UNINETT** jigma2

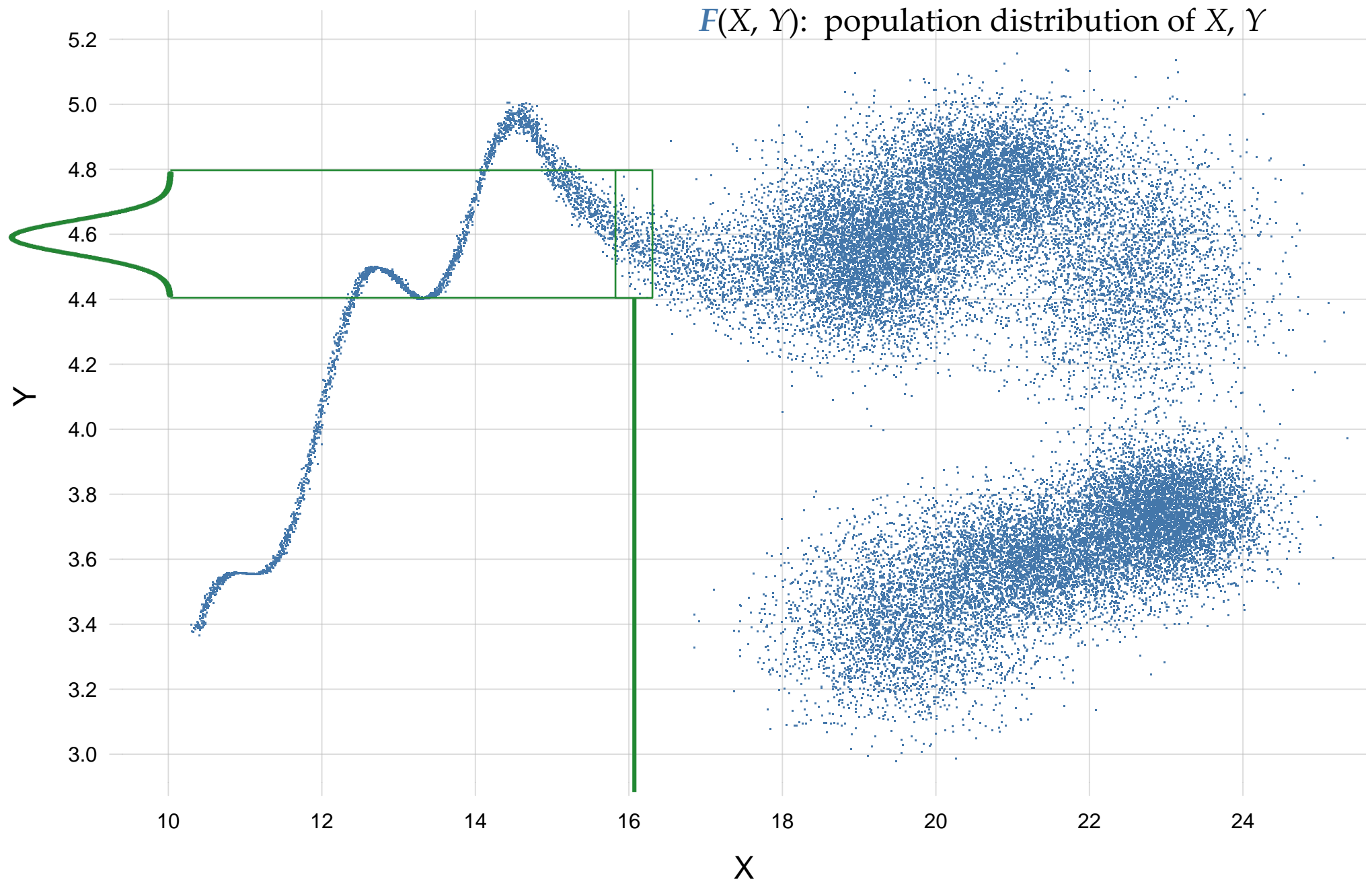


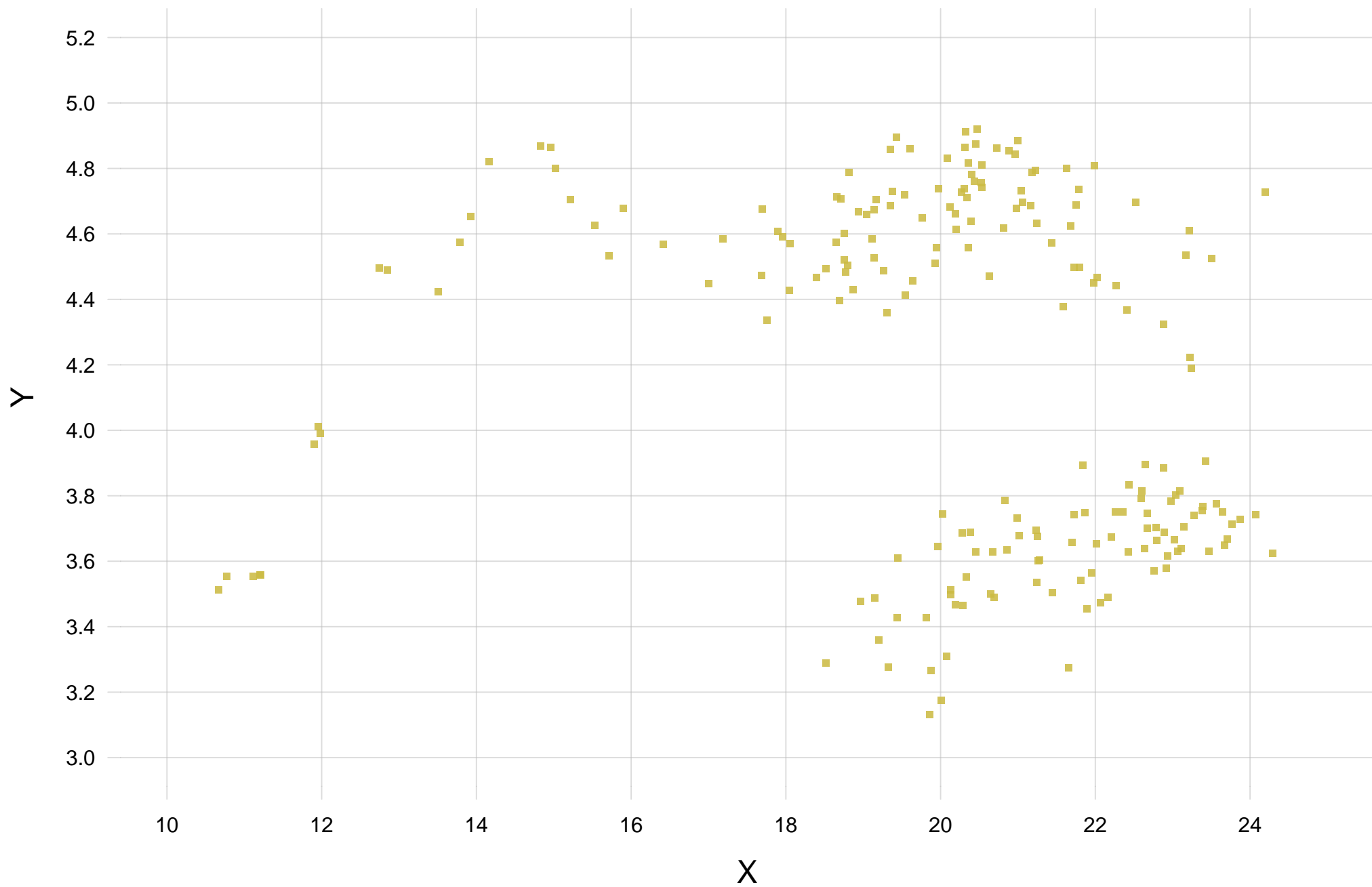




New patient: $X = 16$

$\Rightarrow Y \approx 4.5-4.7$





$$P(y \mid x) = \textcolor{blue}{F}(y \mid x)$$

$$P(y \mid x) = \int F(y \mid x) \, p(F \mid \text{data}) \, dF$$

probability = average over all possible population distributions

$$P(y \mid x) = \int F(y \mid x) \, p(F \mid \text{data}) \, dF$$

probability = average over all possible population distributions

$$p(F \mid \text{data})$$

$$P(y \mid x) = \int F(y \mid x) \, p(F \mid \text{data}) \, dF$$

probability = average over all possible population distributions

$$p(F \mid \text{data}) \propto \underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \dots}_{\text{how well the candidate distribution fits the data}}$$

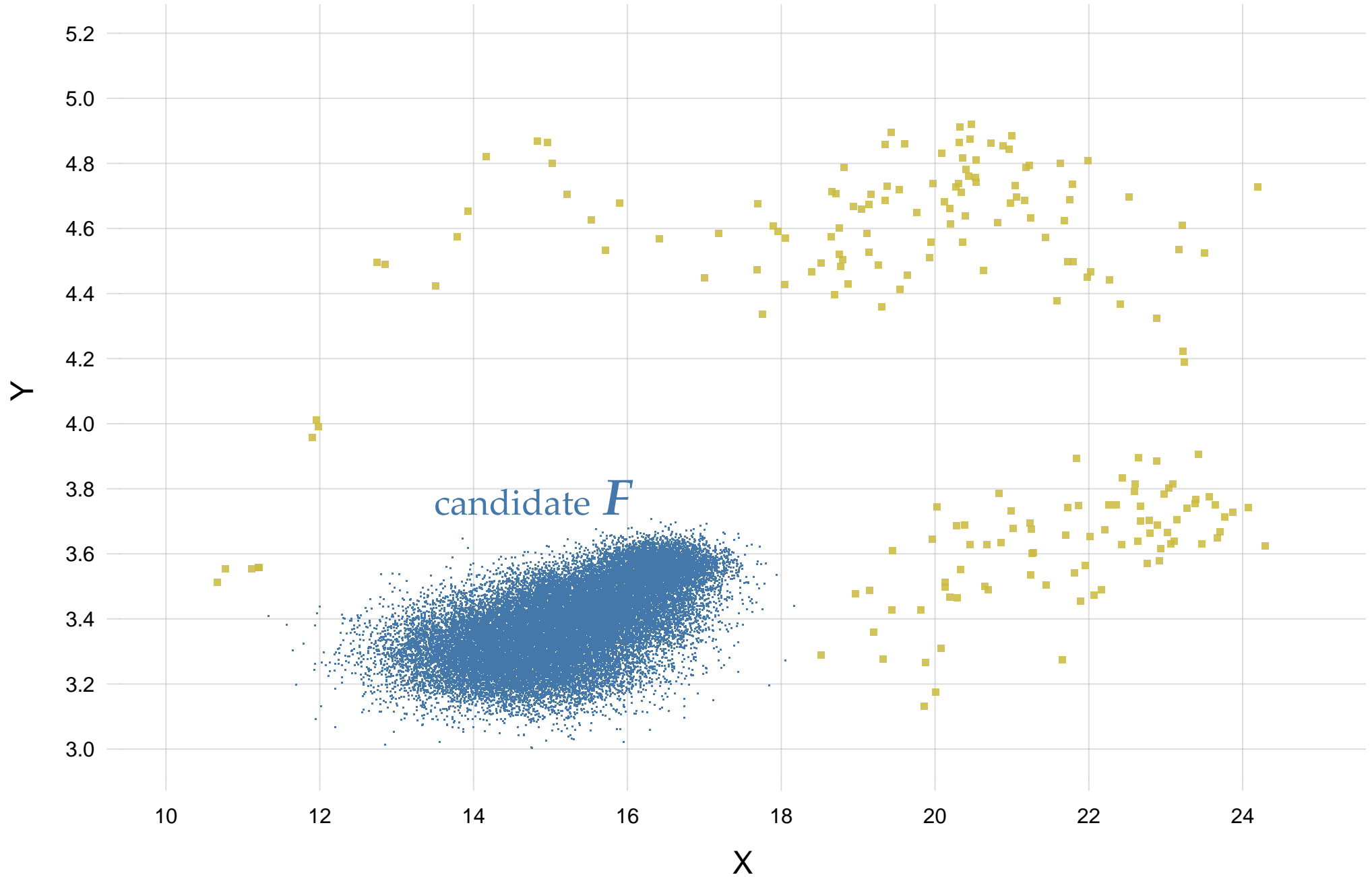
$$P(y \mid x) = \int F(y \mid x) \, p(F \mid \text{data}) \, dF$$

probability = average over all possible population distributions

$$p(F \mid \text{data}) \propto \underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \dots}_{\text{how well the candidate distribution fits the data}} \times \underbrace{p(F \mid \text{prior info})}_{\text{extra-data knowledge}}$$

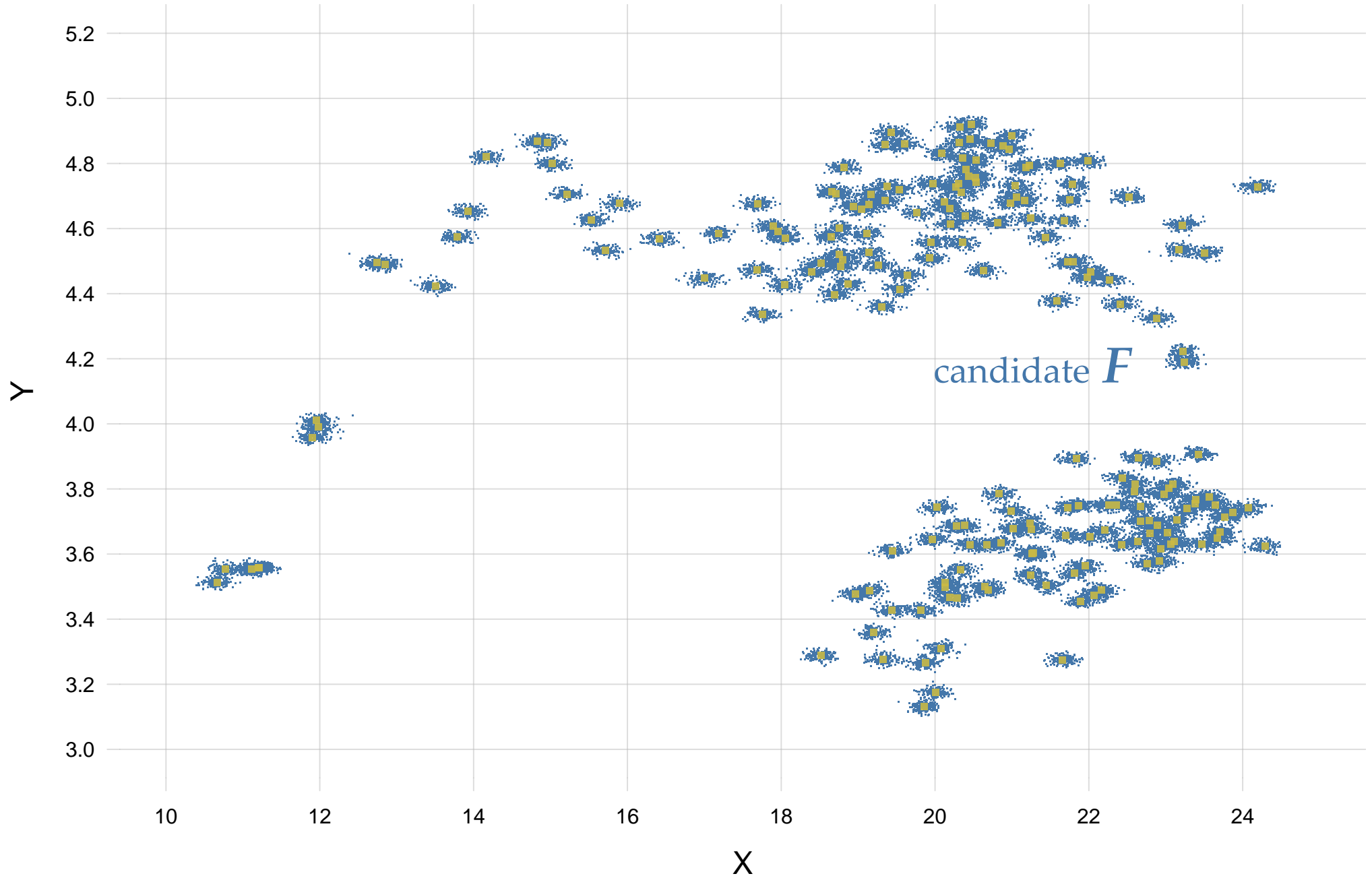
poor candidate: doesn't fit the data

$$\underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \dots}_{\text{low}} \times \underbrace{p(F \mid \text{prior info})}_{\text{high}}$$



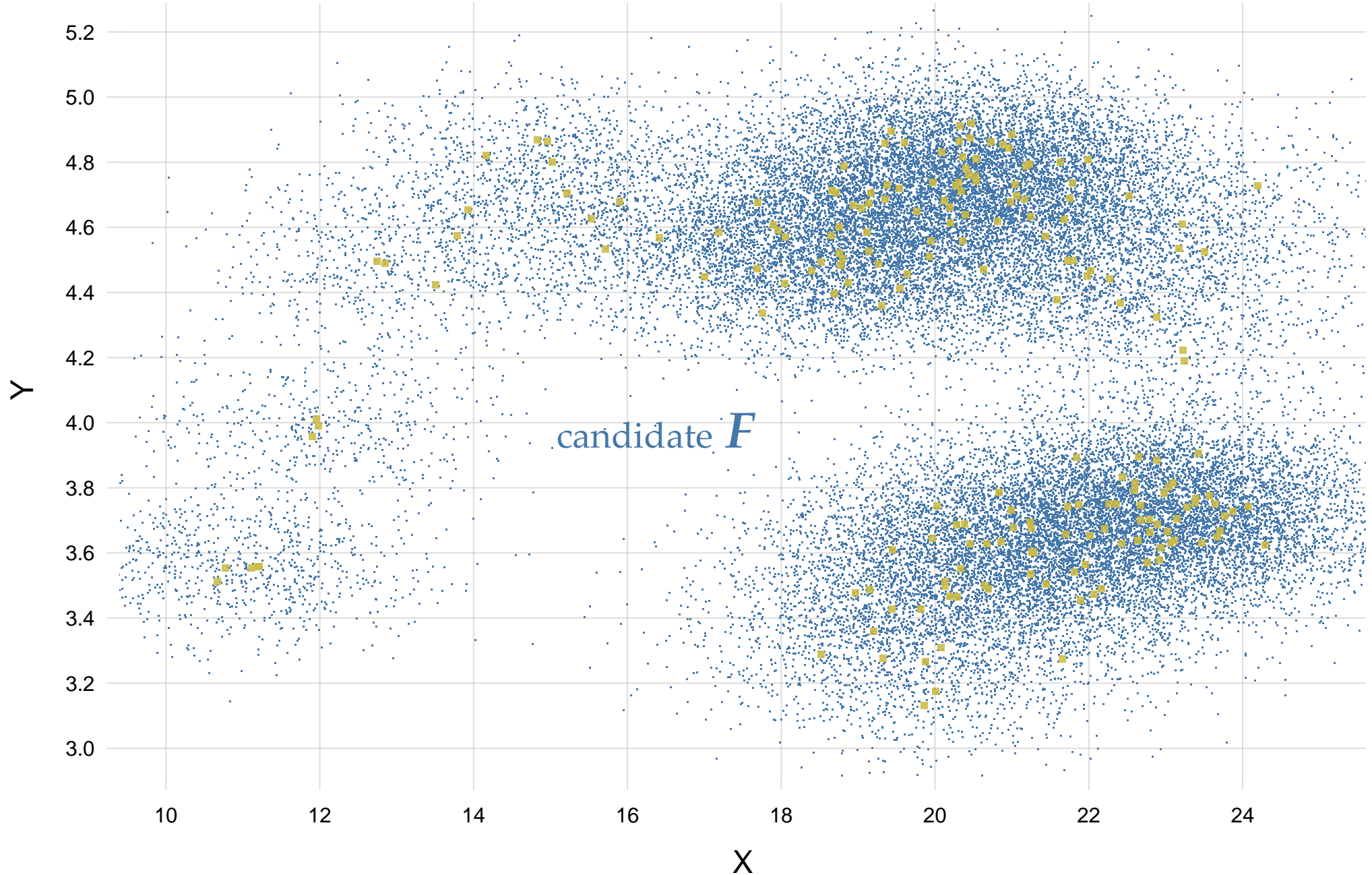
poor candidate: biologically implausible

$$\underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \dots}_{\text{high}} \times \underbrace{p(F \mid \text{prior info})}_{\text{low}}$$



reasonable candidate

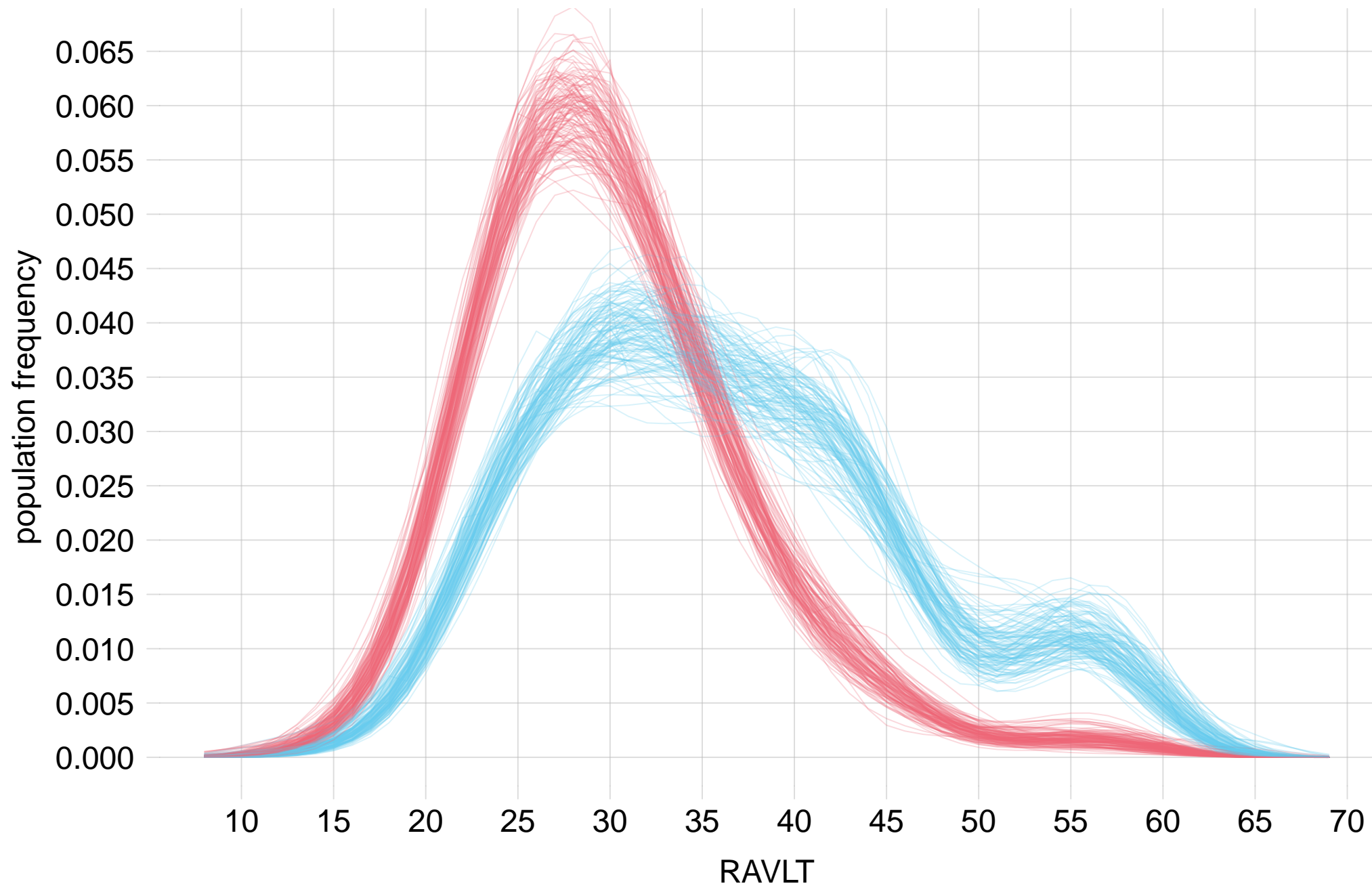
$$\underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \dots}_{\text{high}} \times \underbrace{p(F \mid \text{prior info})}_{\text{high}}$$

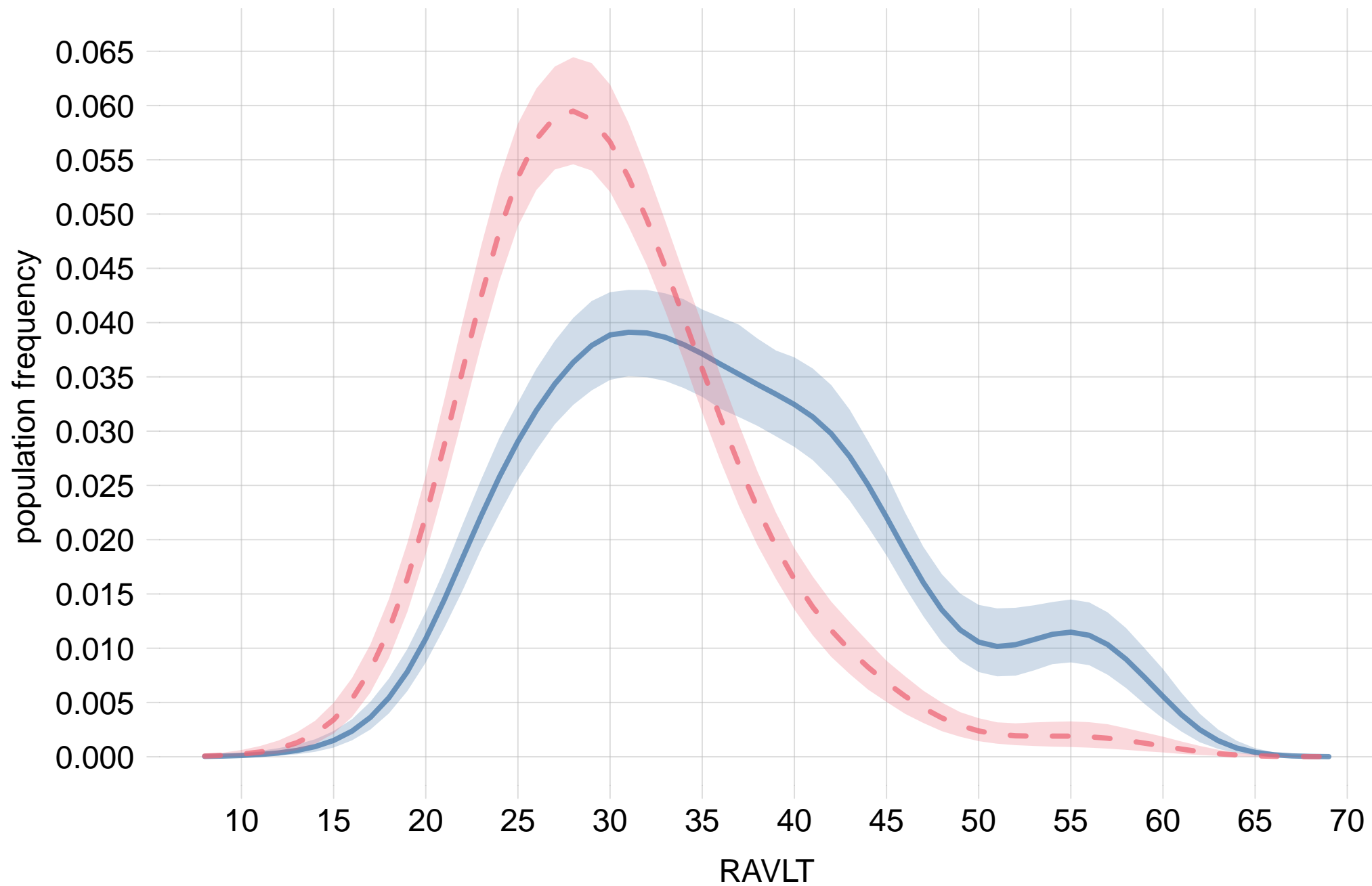


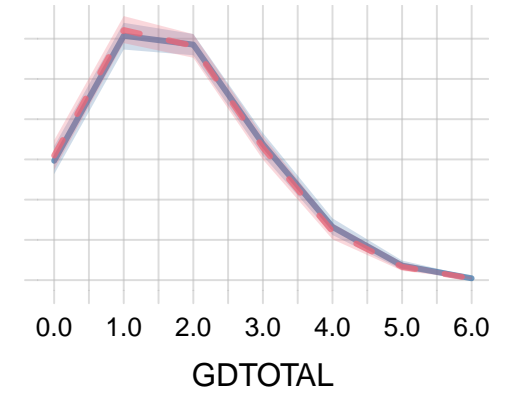
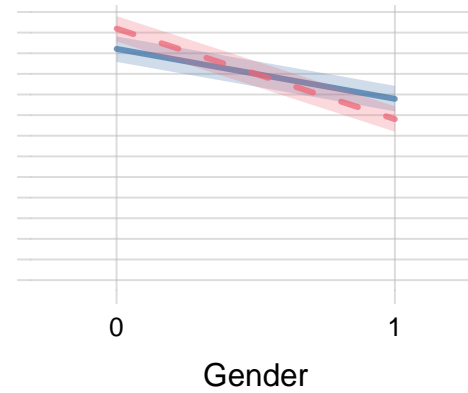
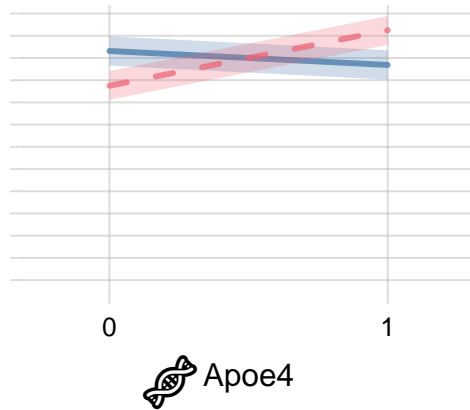
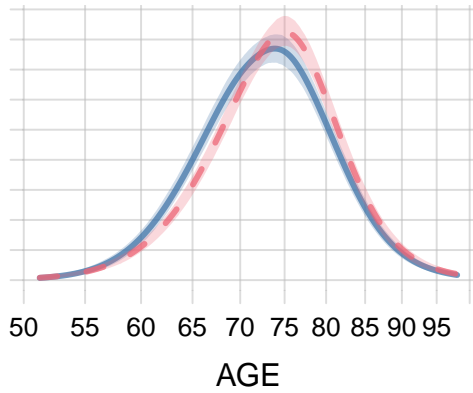
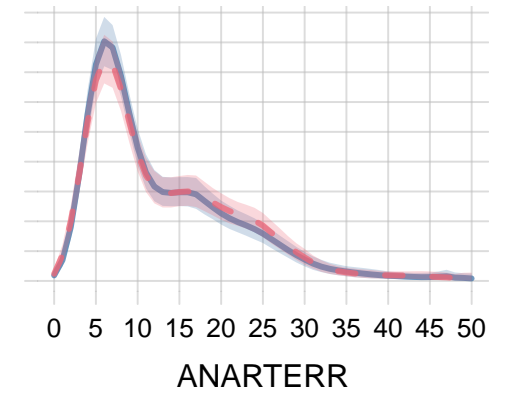
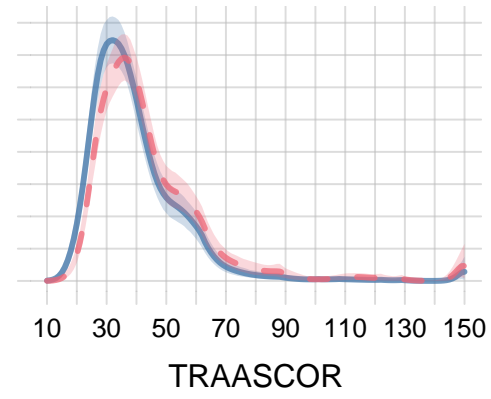
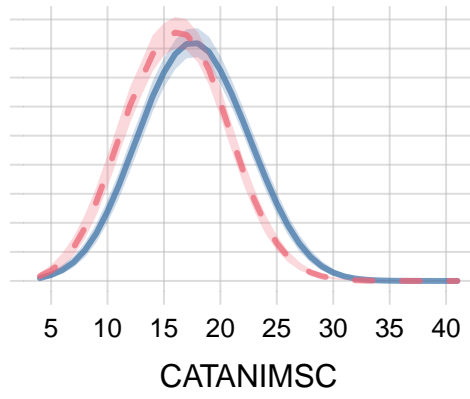
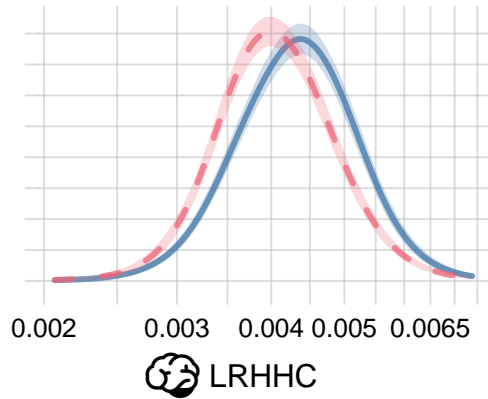
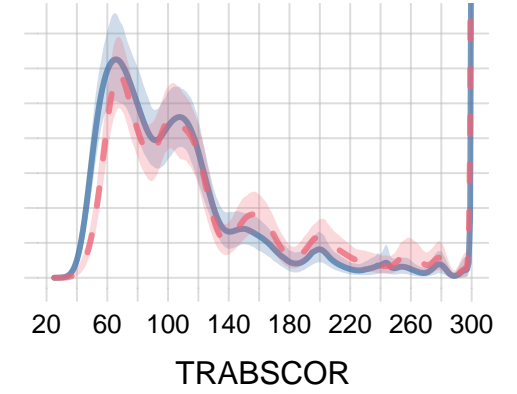
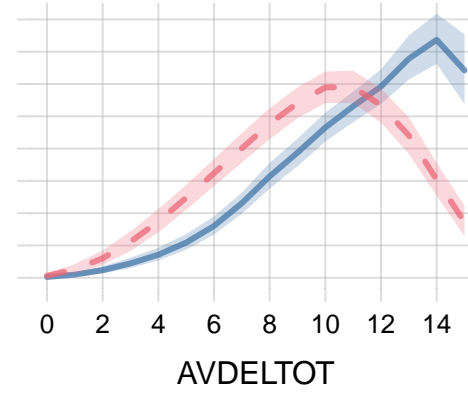
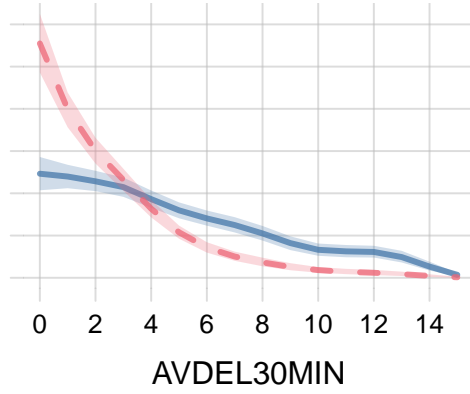
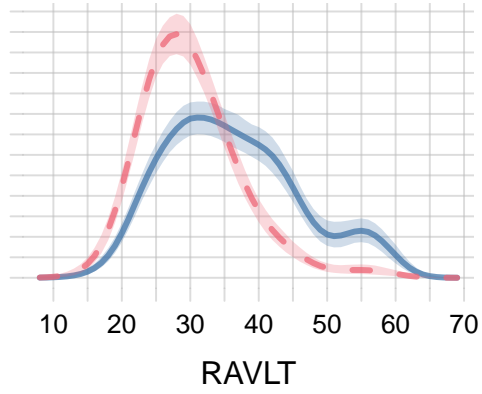
intuition → *mathematics*

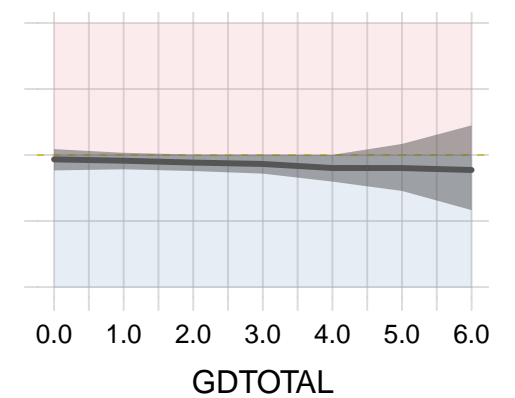
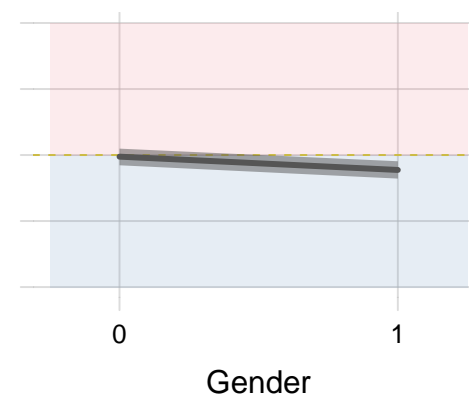
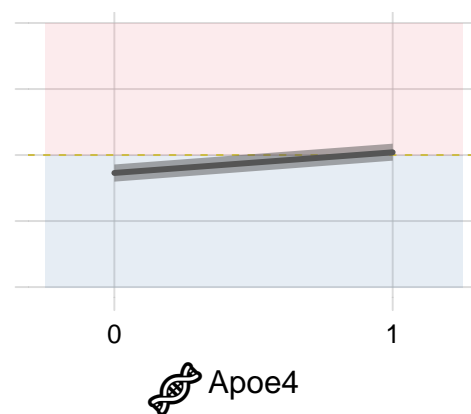
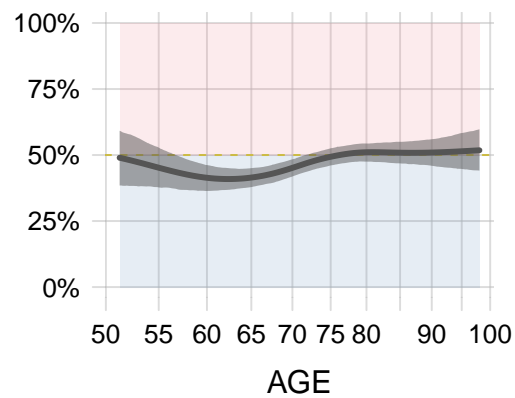
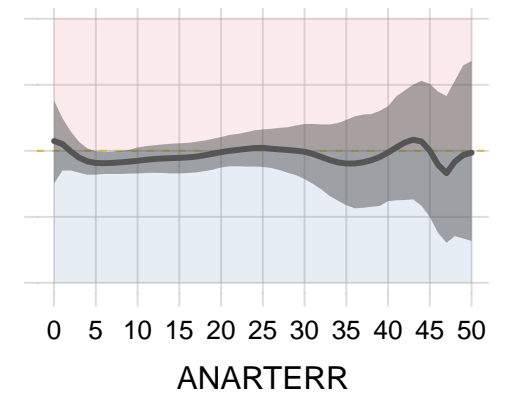
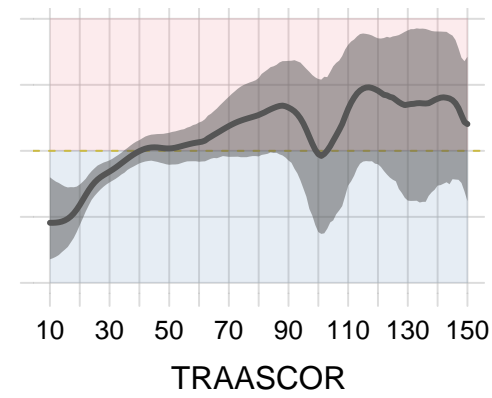
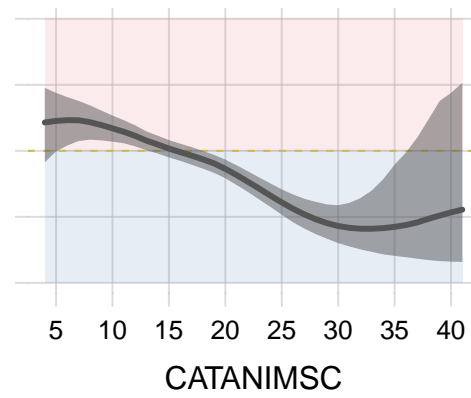
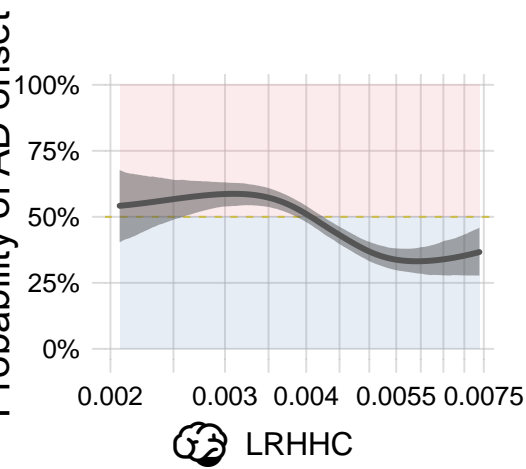
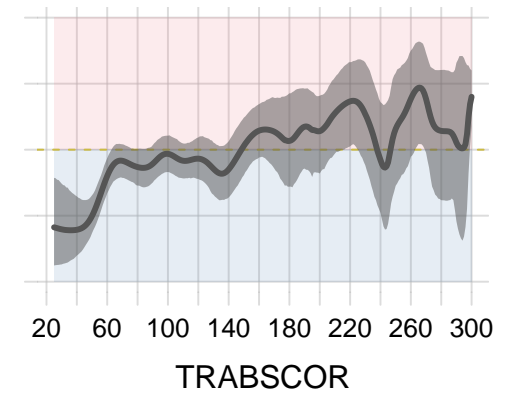
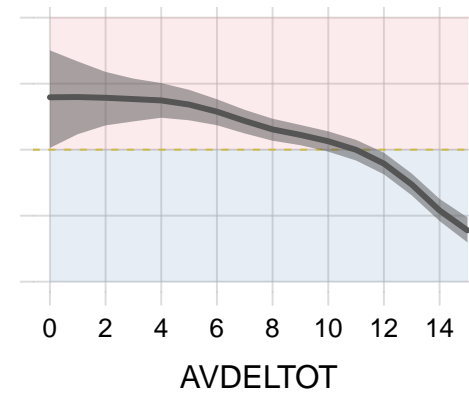
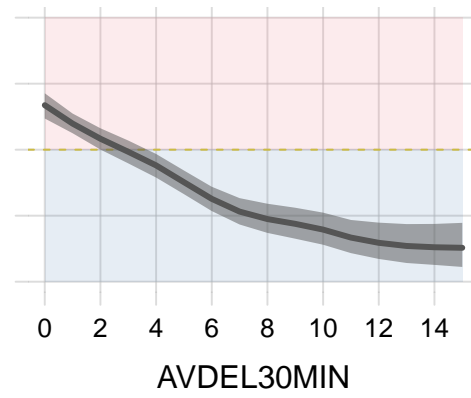
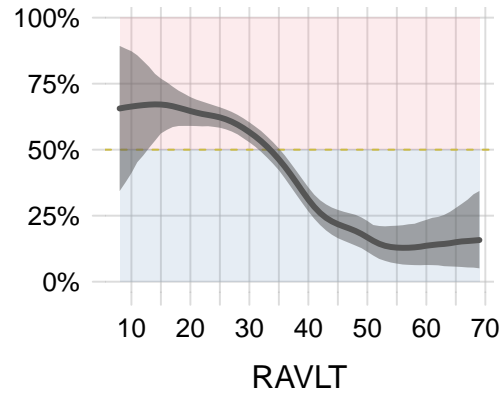
first principles \rightarrow *mathematics* \rightarrow *intuition*

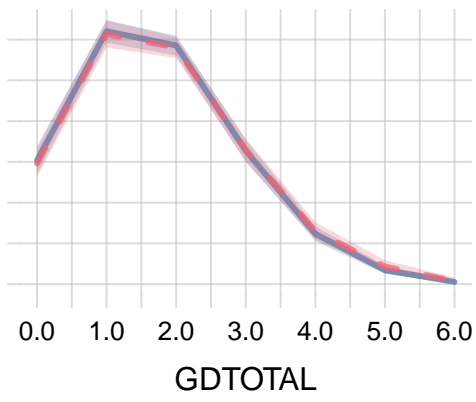
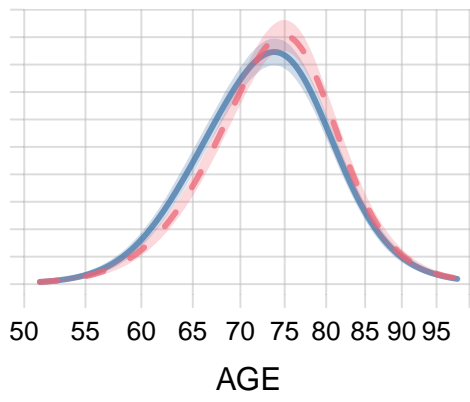
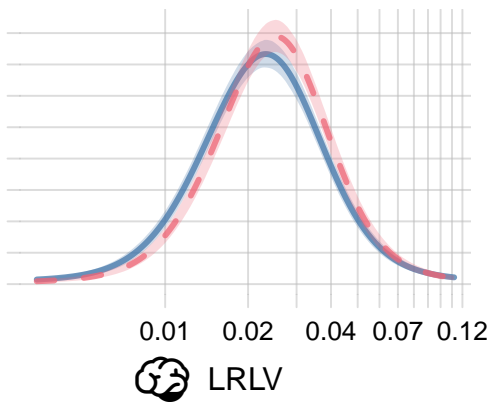
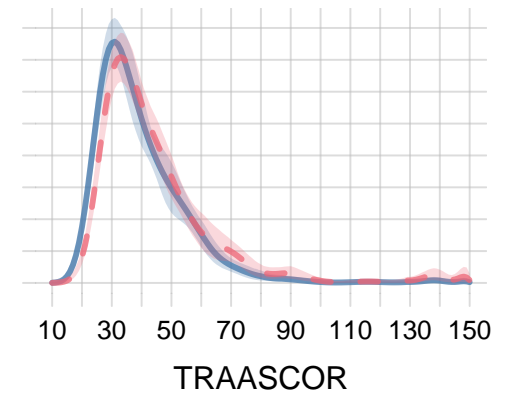
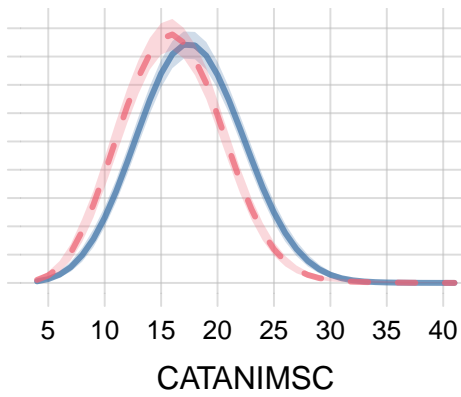
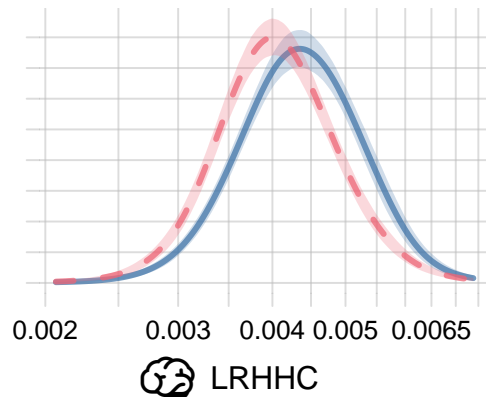
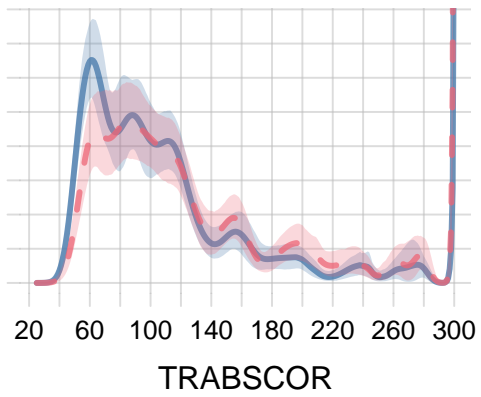
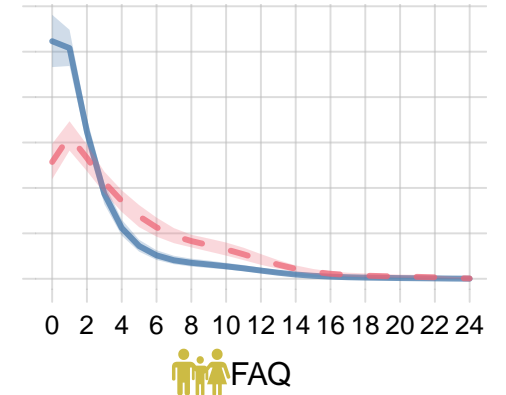
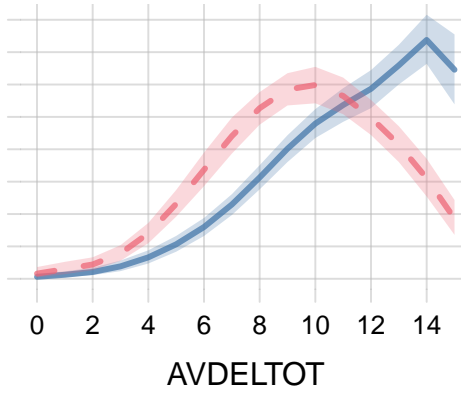
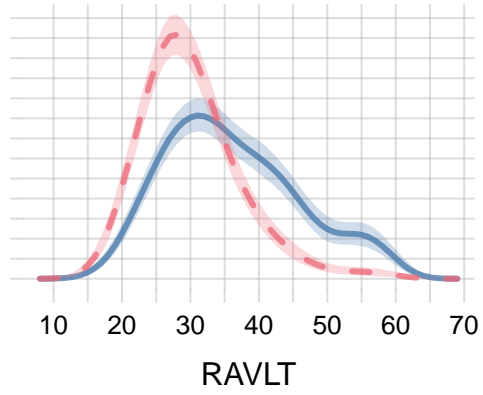
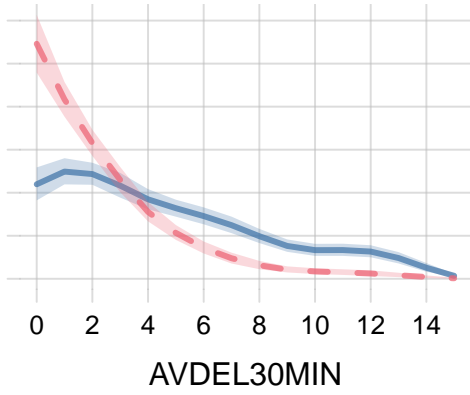
(‘Bayesian’)

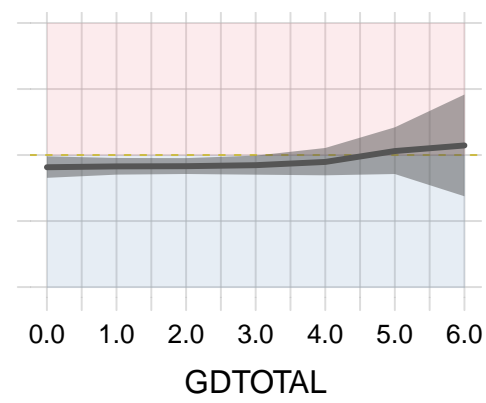
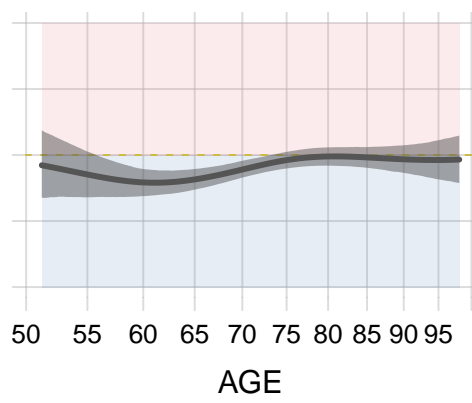
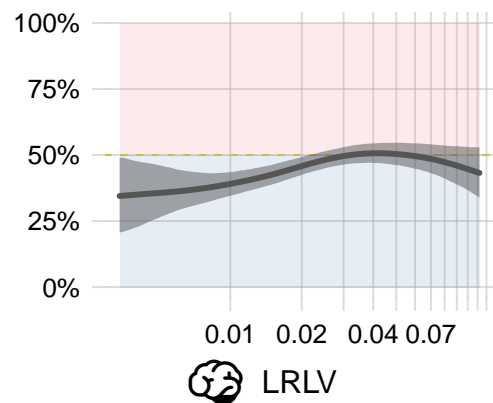
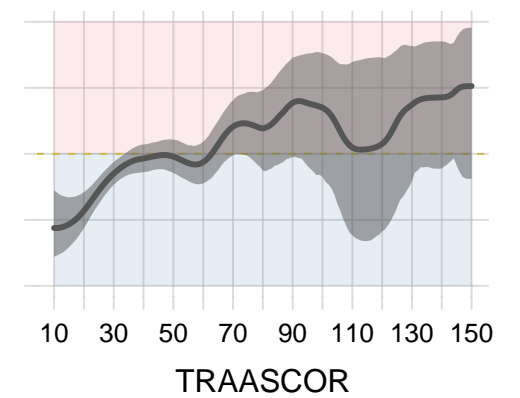
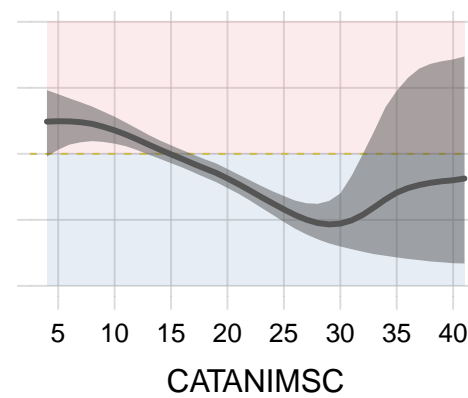
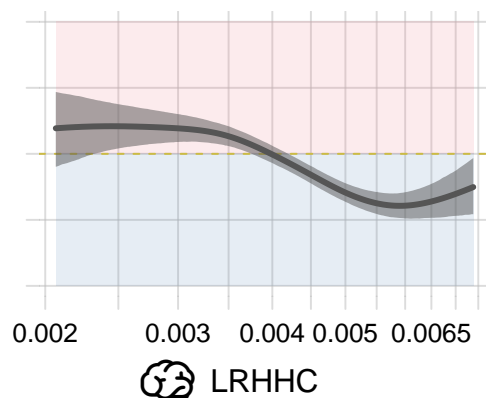
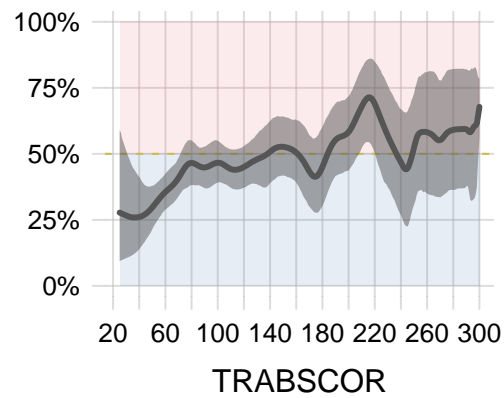
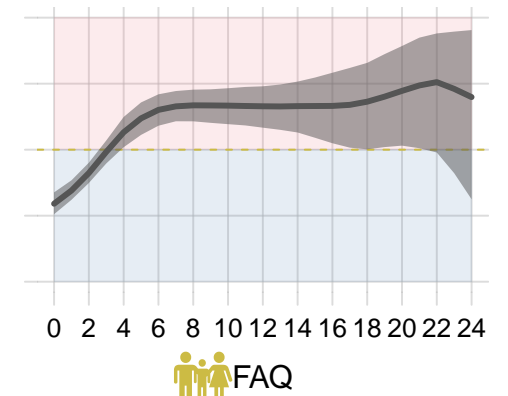
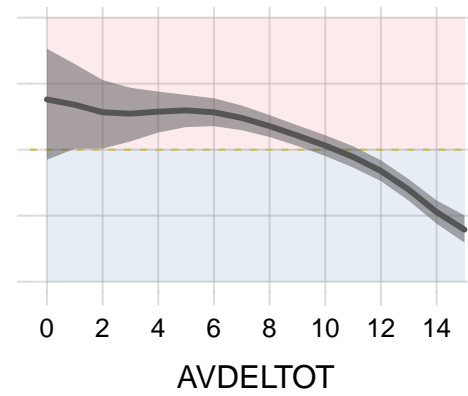
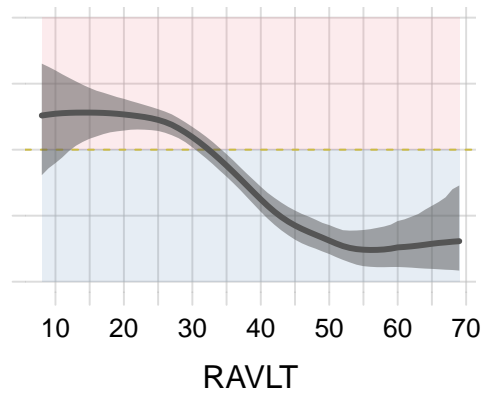
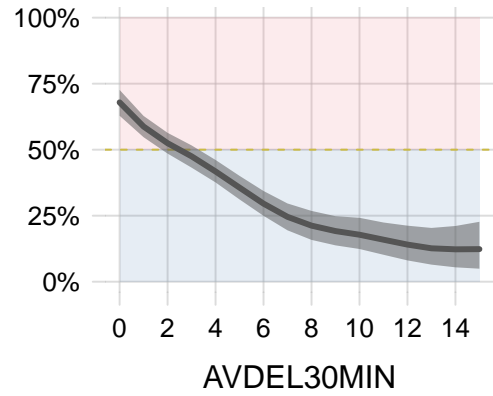












How to quantify the ‘prognostic power’ of a set of features?

Prediction problem:

guess the six digits of the winning lottery ticket ???????

Clue A: ✓✓✓✓??

Clue B: ✓✓✓?✓?

Clue C: ???✓✓✓

What is the ‘importance’ or ‘predictive power’ of each clue?

Scenario 1: we can use **only one** clue

Clue A: ✓✓✓✓??

Clue B: ✓✓✓?✓?

Clue C: ???✓✓✓



Best: **A** or **B** (each gives $1/81$ winning chance)

Worst: **C** (gives $1/729$ winning chance)

Scenario 2: we can use **all** clues

Clue A: ✓✓✓✓??

Clue B: ✓✓✓?✓?

Clue C: ???✓✓✓

→ We fully know the winning number! 💰

Scenario 2: what happens if we **discard** clues?

Clue A: ✓✓✓✓??

Clue B: ✓✓✓?✓?

Clue C: ???✓✓✓

Scenario 2: what happens if we **discard** clues?

Clue A: ✓✓✓✓??

Clue B: ✓✓✓?✓?

Clue C: ???✓✓✓

- Discard A: still 100% win \Rightarrow A has 'importance=0'

Scenario 2: what happens if we **discard** clues?

Clue A: ✓✓✓✓??

Clue B: ✓✓✓?✓?

Clue C: ???✓✓✓

- Discard **A**: still 100% win \Rightarrow **A** has 'importance=0'
- Discard **B**: still 100% win \Rightarrow **B** has 'importance= 0'

Scenario 2: what happens if we **discard** clues?

Clue A: ✓✓✓✓??

Clue B: ✓✓✓?✓?

Clue C: ???✓✓✓

- Discard **A**: still 100% win \Rightarrow **A** has 'importance=0'
- Discard **B**: still 100% win \Rightarrow **B** has 'importance= 0'
- Discard **A and B**: 1/9 winning chance

Scenario 2: what happens if we **discard** clues?

Clue A: ✓✓✓✓??

Clue B: ✓✓✓?✓?

Clue C: ???✓✓✓

- Discard **A**: still 100% win \Rightarrow **A** has 'importance=0'
- Discard **B**: still 100% win \Rightarrow **B** has 'importance=0'
- Discard **A and B**: 1/9 winning chance
 \Rightarrow **A and B** together have 'importance>0'

Scenario 2: what happens if we **discard** clues?

Clue A: ✓✓✓✓??

Clue B: ✓✓✓?✓?

Clue C: ???✓✓✓

- Discard **A**: still 100% win \Rightarrow **A** has 'importance=0'
- Discard **B**: still 100% win \Rightarrow **B** has 'importance=0'
- Discard **A and B**: 1/9 winning chance
 \Rightarrow **A and B** together have 'importance>0'

$$'0 + 0 \neq 0'$$

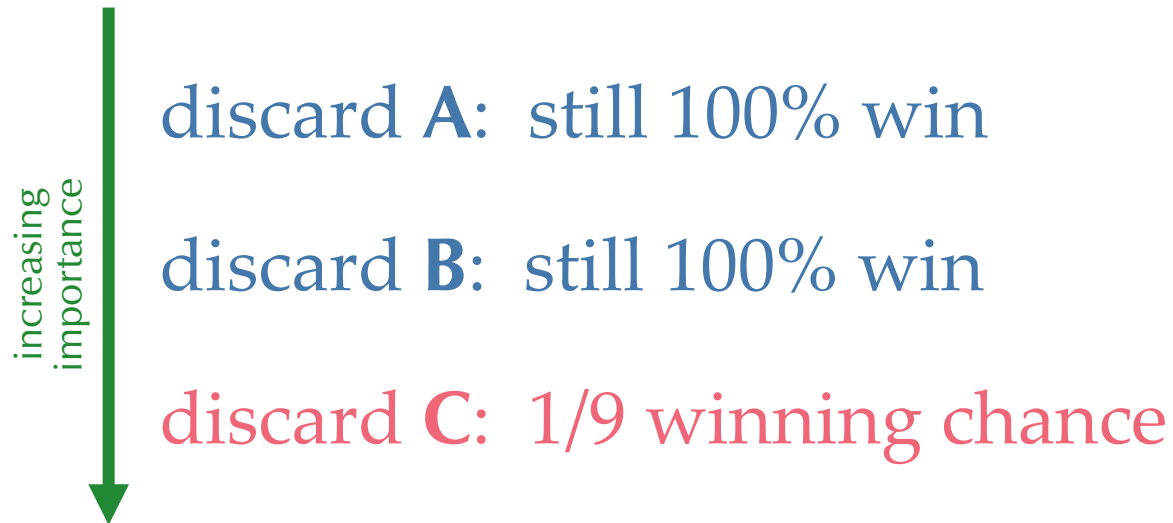
‘Importance’ or ‘predictive power’ is *not* an *additive* property

Scenario 3: we have to **discard one** clue. Which?

Clue A: ✓✓✓✓??

Clue B: ✓✓✓?✓?

Clue C: ???✓✓✓



→ If we have to discard one clue, it's most important that we keep **C**

Scenario 1:
choose one clue



A B
C

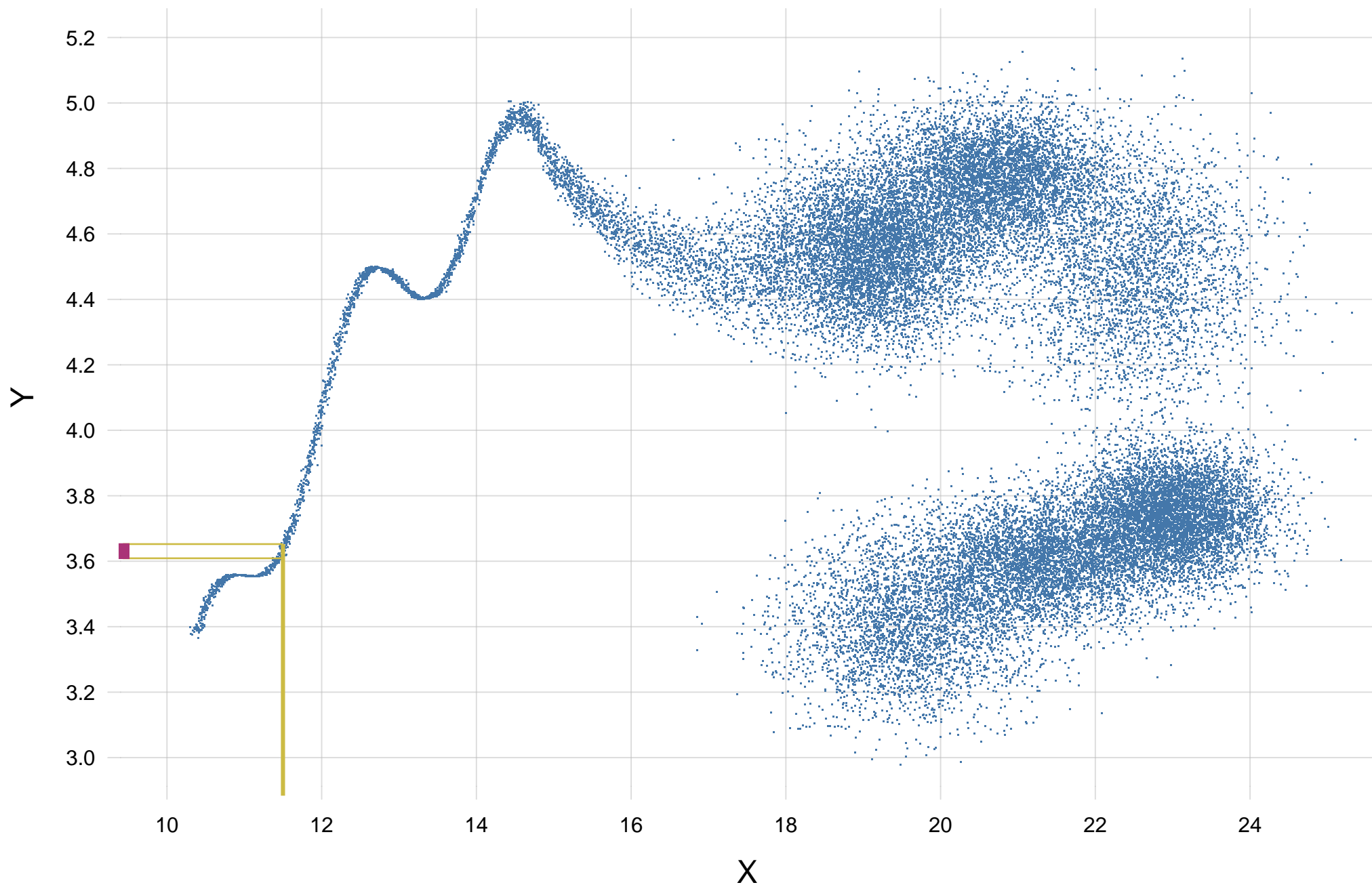
Scenario 3:
discard one clue



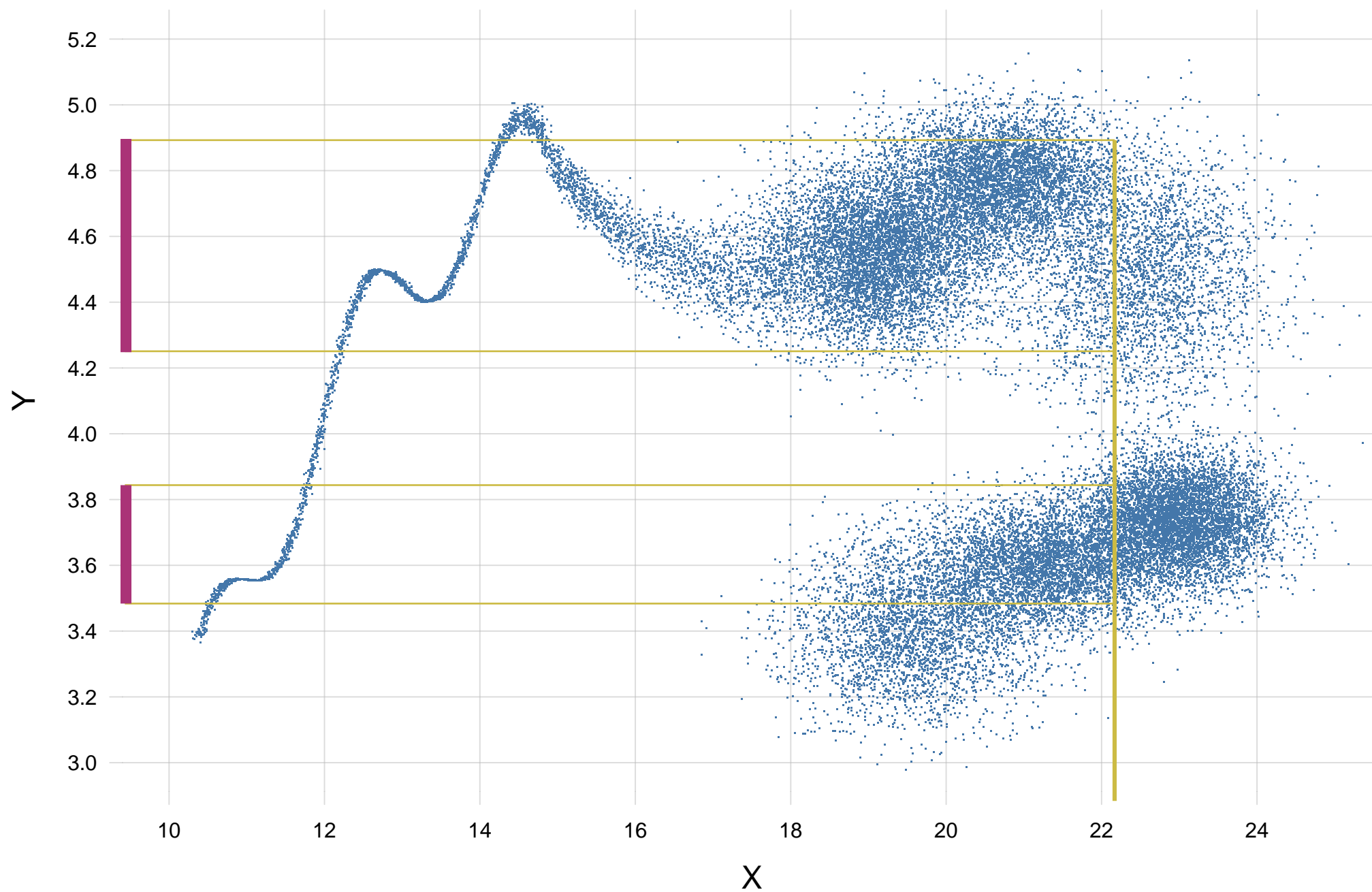
C
A B

‘Importance’ or ‘predictive power’ is *context-dependent*

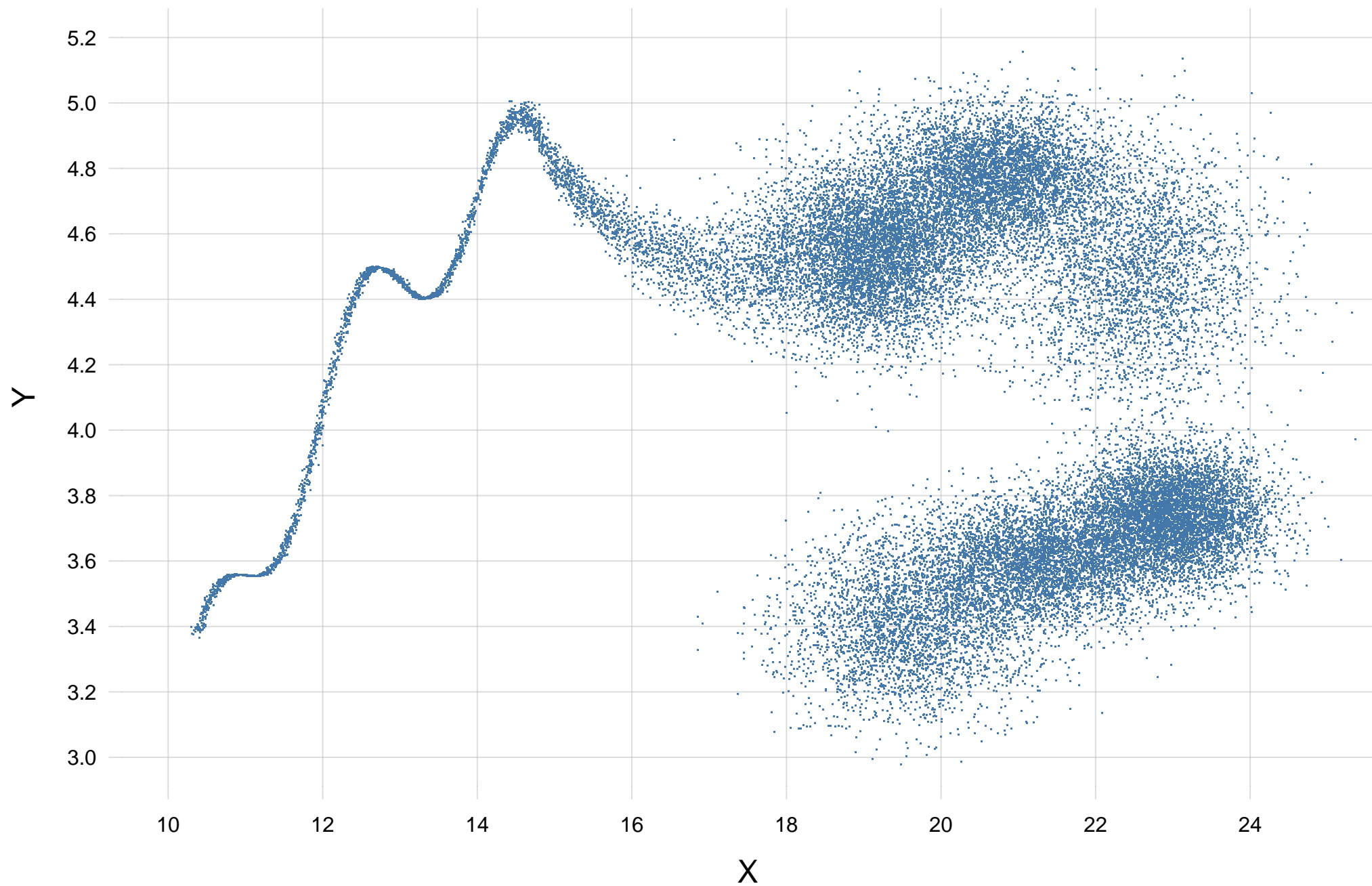
$$x = 11.5 \Rightarrow y \approx 3.60-3.65$$

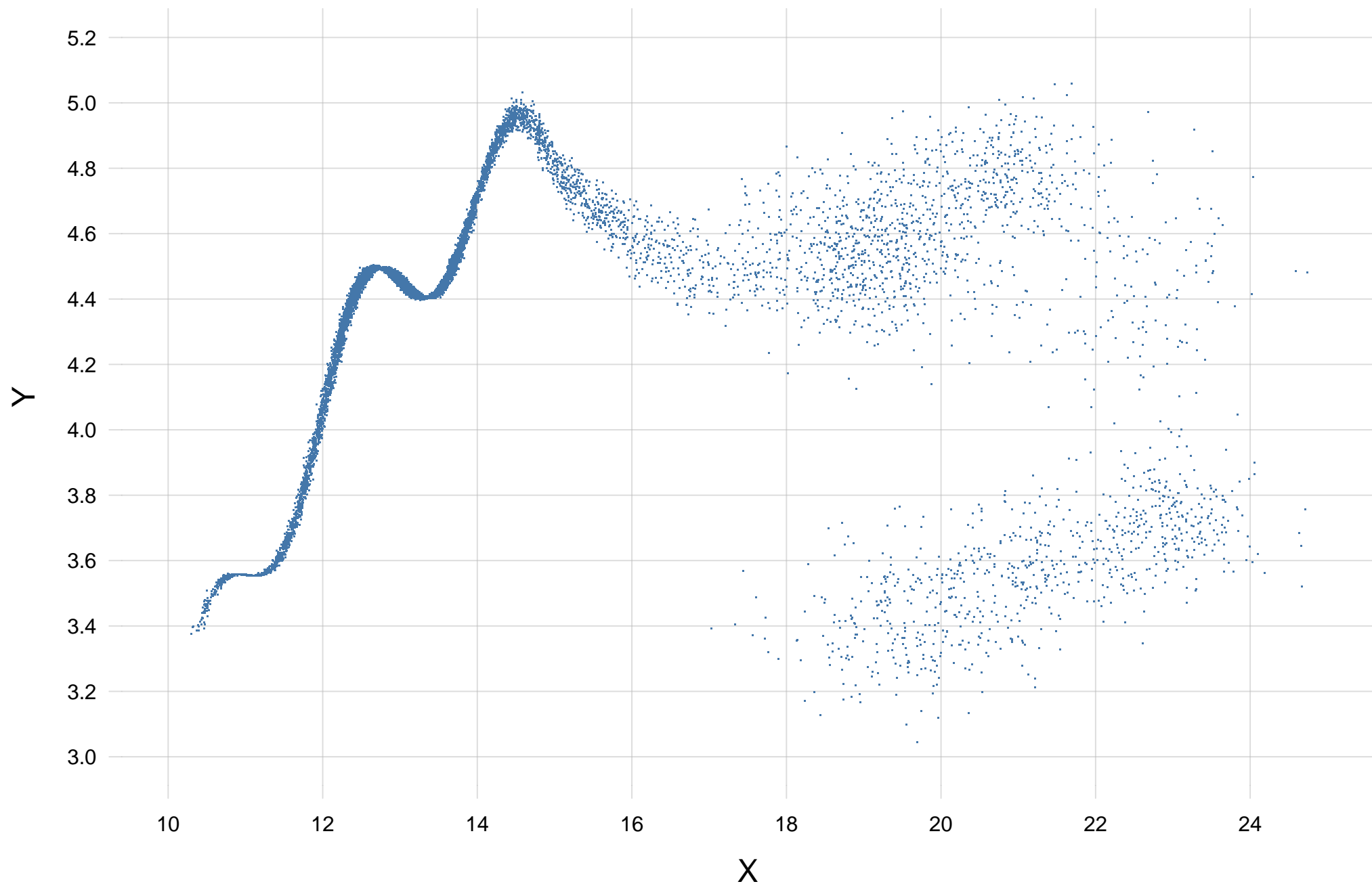


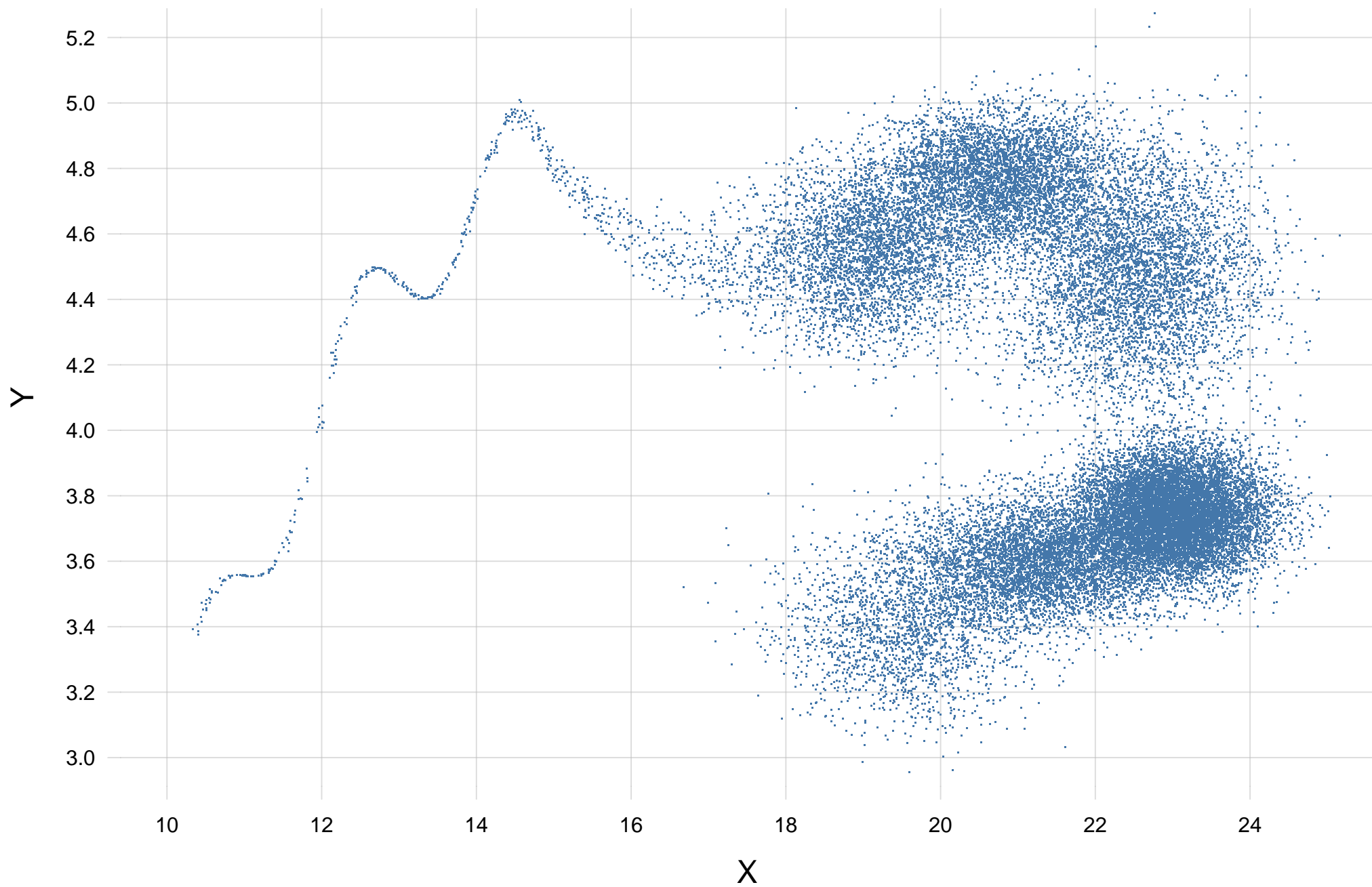
$x = 22 \Rightarrow y \approx 3.50\text{--}3.85 \text{ or } 4.25\text{--}4.90$



What is the 'overall predictive power' of X ?







The ‘predictive power’ of X depends on $P(X)$

⚠ Careful with ‘balancing’! ⚠

The Bell System Technical Journal

Vol. XXVII

July, 1948

No. 3

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley² on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at

Information Theory

The Bell System Technical Journal

Vol. XXVII

July, 1948

No. 3

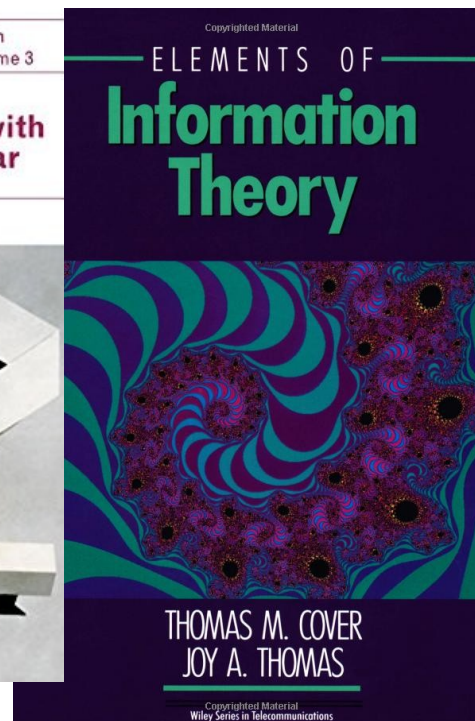
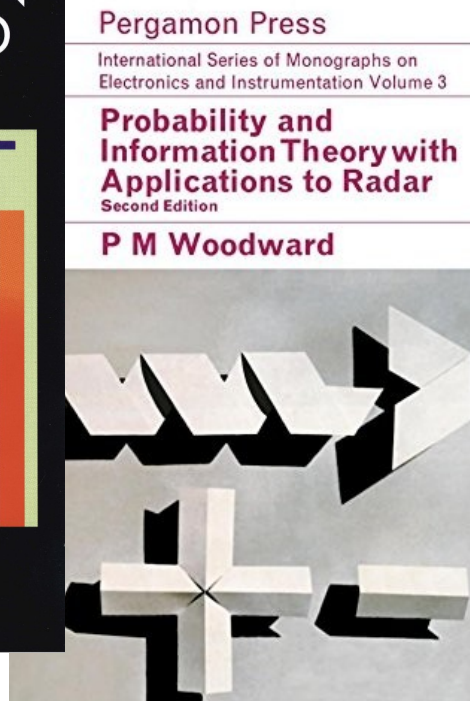
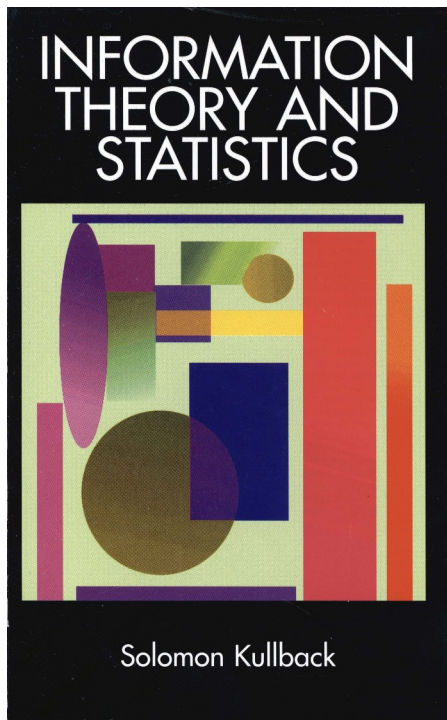
A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley² on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at



Information Theory

The Bell System Technical Journal

Vol. XXVII

July, 1948

No. 3

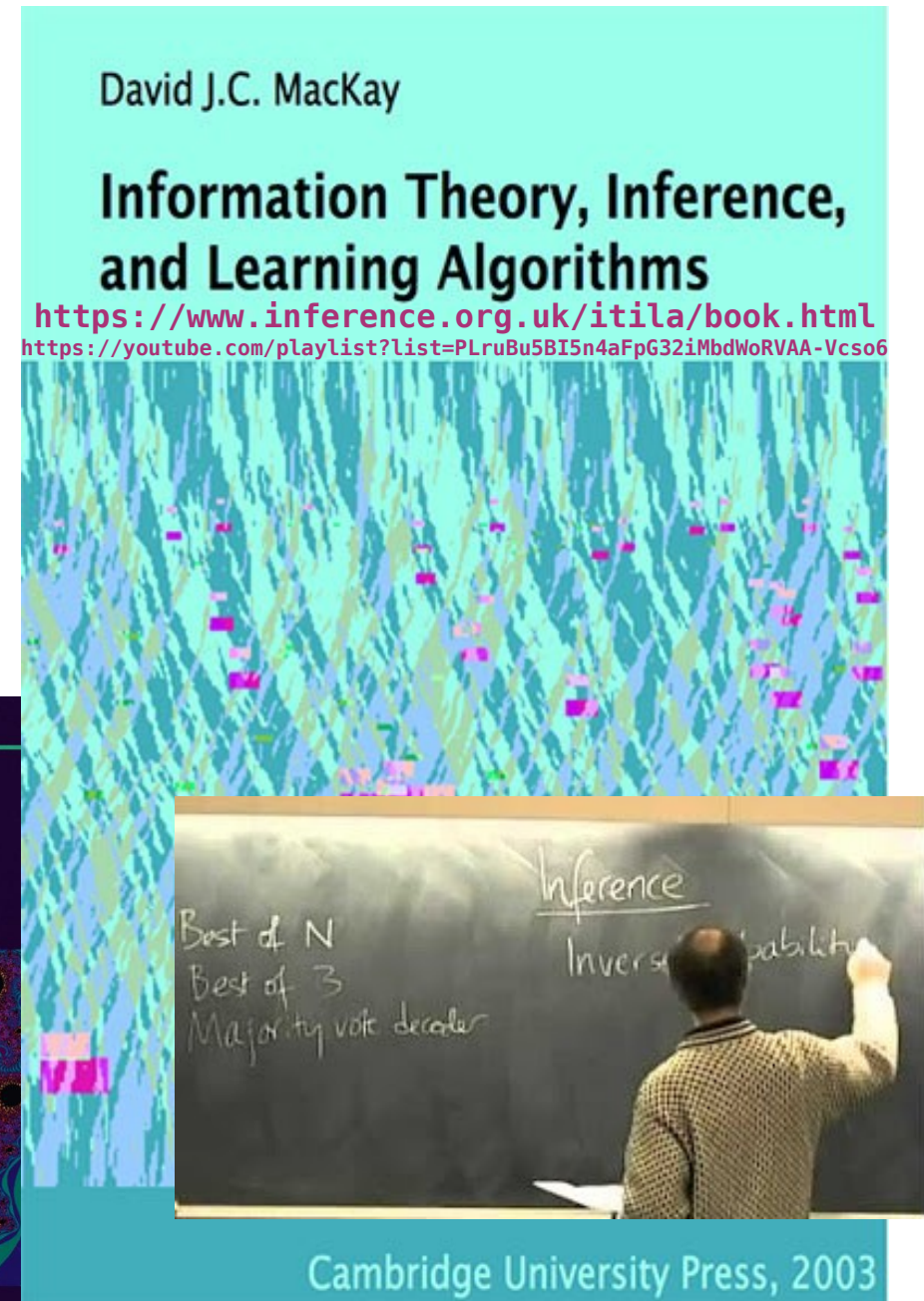
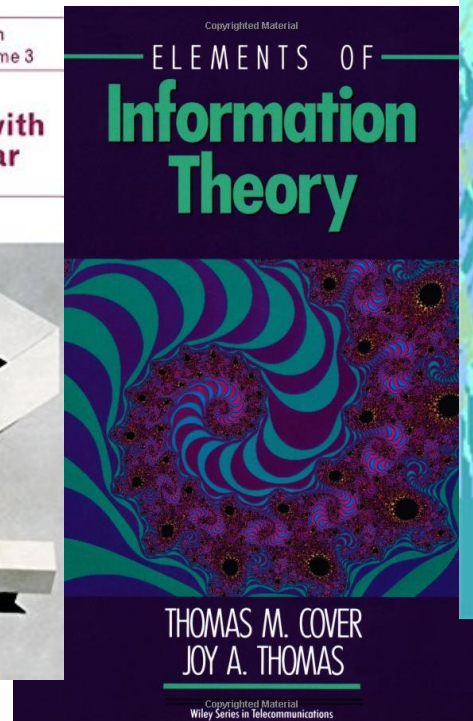
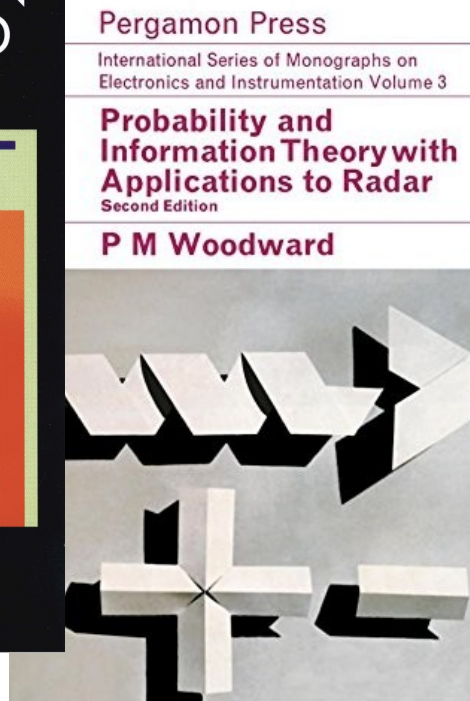
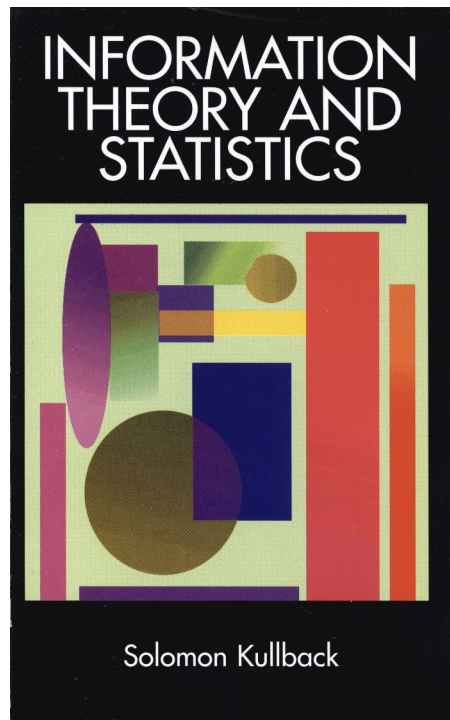
A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley² on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at



‘predictive power’ of X for Y \coloneqq **Mutual information** between Y and X
(mean transinformation content)

$$I(X; Y) := \int p(y|x) p(x) \log \left[\frac{p(y|x)}{p(y)} \right] dy dx$$

$$I(Y; X_1, X_2) \geq I(Y; X_1)$$

$$I(Y; X_1, X_2) \geq I(Y; X_2)$$

$$\text{but } I(Y; X_1, X_2) \neq I(Y; X_1) + I(Y; X_2)$$

INTERNATIONAL STANDARD

NORME INTERNATIONALE

**Quantities and units –
Part 13: Information science and technology**

**Grandeurs et unités –
Partie 13: Science et technologies de l'information**

INTERNATIONAL STANDARD

INFORMATION SCIENCE AND TECHNOLOGY			QUANTITIES	
Item No.	Name	Symbol	Definition	Remarks
13-24 (902)	information content <i>fr</i> <i>quantité (f) d'information</i>	$I(x)$	$I(x) = \lg \frac{1}{p(x)} \text{ Sh} = \lg \frac{1}{p(x)} \text{ Hart} =$ $\ln \frac{1}{p(x)} \text{ nat}$ <p>where $p(x)$ is the probability of event x</p>	See ISO/IEC 2382-16, item 16.03.02. See also IEC 60027-3.
13-25 (903)	entropy <i>fr</i> <i>entropie (f)</i>	H	$H(X) = -\sum_{i=1}^n p(x_i) \lg p(x_i)$ <p>for the set $X = \{x_1, \dots, x_n\}$ where $p(x_i)$ is the probability and $I(x_i)$ is the information content of event x_i</p>	See ISO/IEC 2382-16, item 16.03.03.
13-30 (908)	joint information content <i>fr</i> <i>quantité (f) d'information conjointe</i>	$I(x, y)$	$I(x, y) = \lg \frac{1}{p(x, y)} \text{ Sh} = \lg \frac{1}{p(x, y)} \text{ Hart} =$ $\ln \frac{1}{p(x, y)} \text{ nat}$ <p>where $p(x, y)$ is the joint probability of events x and y</p>	
13-35 (912)	transinformation content <i>fr</i> <i>transinformation (f)</i>	$T(x, y)$	$T(x, y) = I(x) + I(y) - I(x, y)$ <p>where $I(x)$ and $I(y)$ are the information contents (13-24) of events x and y, respectively, and $I(x, y)$ is their joint information content (13-30)</p>	See ISO/IEC 2382-16, item 16.04.07.
13-36 (913)	mean transinformation content <i>fr</i> <i>transinformation (f) moyenne</i>	T	$T(X, Y) = \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) T(x_i, y_j)$ <p>for the sets $X = \{x_1, \dots, x_n\}$, $Y = \{y_1, \dots, y_m\}$, where $p(x_i, y_j)$ is the joint probability of events x_i and y_j, and $T(x_i, y_j)$ is their transinformation content (item 13-35)</p>	See ISO/IEC 2382-16, item 16.04.08.

UNITS INFORMATION SCIENCE AND TECHNOLOGY				
Item No.	Name	Symbol	Definition	Conversion factors and remarks
13-24.a	shannon	Sh	value of the quantity when the argument is equal to 2	1 Sh \approx 0,693 nat \approx 0,301 Hart
13-24.b	hartley	Hart	value of the quantity when the argument is equal to 10	1 Hart \approx 3,322 Sh \approx 2,303 nat
13-24.c	natural unit of information	nat	value of the quantity when the argument is equal to e	1 nat \approx 1,433 Sh \approx 0,434 Hart
13-25.a	shannon	Sh		
13-25.b	hartley	Hart		
13-25.c	natural unit of information	nat		
13-30.a	shannon	Sh		
13-30.b	hartley	Hart		
13-30.c	natural unit of information	nat		
13-35.a	shannon	Sh		
13-35.b	hartley	Hart		
13-35.c	natural unit of information	nat		
13-36.a	shannon	Sh		In practice, the unit "shannon per character" is generally used, and sometimes the units "hartley per character" and "natural unit per character".
13-36.b	hartley	Hart		
13-36.c	natural unit of information	nat		

If Y is binary:

$$0 \text{ Sh} \leq I(Y; X) \leq 1 \text{ Sh}$$

If Y is binary:

$$0 \text{ Sh} \leq I(Y; X) \leq 1 \text{ Sh}$$

X and Y are independent

Using X is no better than flipping a coin

If Y is binary:

$$0 \text{ Sh} \leq I(Y; X) \leq 1 \text{ Sh}$$

X and Y are independent

Using X is no better than flipping a coin

Y is a deterministic function of X

X always yields perfect predictions

If Y is binary:

$$0 \text{ Sh} \leq I(Y; X) \leq 1 \text{ Sh}$$

X and Y are independent

Using X is no better than flipping a coin

Y is a deterministic function of X

X always yields perfect predictions

$$I(Y; X) = 0.22 \text{ Sh}$$

If Y is binary:

$$0 \text{ Sh} \leq I(Y; X) \leq 1 \text{ Sh}$$

X and Y are independent

Using X is no better than flipping a coin

Y is a deterministic function of X

X always yields perfect predictions

$$I(Y; X) = 0.22 \text{ Sh}$$

In $N=100$ new prognoses:

- we are **completely certain** about 22
- we are **completely uncertain** about $100 - 22 = 78$

→ approx $22 + 78/2 = 61$ correct prognoses (TP+TN)

If Y is binary:

$$0 \text{ Sh} \leq I(Y; X) \leq 1 \text{ Sh}$$

X and Y are independent

Using X is no better than flipping a coin

Y is a deterministic function of X

X always yields perfect predictions

$$I(Y; X) = 0.22 \text{ Sh}$$

In $N=100$ new prognoses:

- we are **completely certain** about 22
- we are **completely uncertain** about $100 - 22 = 78$

→ approx $22 + 78/2 = 61$ correct prognoses (TP+TN)

Any prognostic algorithm
which does less than this
is performing poorly

If Y is binary:

$$0 \text{ Sh} \leq I(Y; X) \leq 1 \text{ Sh}$$

X and Y are independent

Using X is no better than flipping a coin

Y is a deterministic function of X

X always yields perfect predictions

$$I(Y; X) = 0.22 \text{ Sh}$$

In $N=100$ new prognoses:

- we are **completely certain** about 22
- we are **completely uncertain** about $100 - 22 = 78$

→ approx $22 + 78/2 = 61$ correct prognoses (TP+TN)

Any prognostic algorithm
which does less than this
is performing poorly

Upper bound \approx $\frac{1}{2} + \frac{1}{2} \sqrt{1 - (1 - 0.22)^{\frac{4}{3}}} \approx 77 \% \pm 0.8\sqrt{N} \% \quad \text{correct prognoses}$

If Y is binary:

$$0 \text{ Sh} \leq I(Y; X) \leq 1 \text{ Sh}$$

X and Y are independent

Using X is no better than flipping a coin

Y is a deterministic function of X

X always yields perfect predictions

$$I(Y; X) = 0.22 \text{ Sh}$$

In $N=100$ new prognoses:

- we are **completely certain** about 22
- we are **completely uncertain** about $100 - 22 = 78$

→ approx $22 + 78/2 = 61$ correct prognoses (TP+TN)

Any prognostic algorithm
which does less than this
is performing poorly

Upper bound \approx $\frac{1}{2} + \frac{1}{2} \sqrt{1 - (1 - 0.22)^{\frac{4}{3}}} \approx 77 \% \pm 0.8\sqrt{N} \%$ correct prognoses

Maximum accuracy attainable
by *any* algorithm which uses only feature set X

