Bayesian nonparametric density regression
&
information analysis

example
from a hypothetical Alzheimer study

Luca

CDA 2022

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

ASA Statement on Statistical Significance and *P*-Values

1. Introduction

Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions drawn from research data. The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with an index called the *p*-value. While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of *p*-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since *p*-values were first introduced.

In this context, the American Statistical Association (ASA) believes that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. This statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies. Rather, the statement articulates in nontechnical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

2. What is a *p*-Value?

Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

3. Principles

1. ***P*-values can indicate how incompatible the data are with a specified statistical model.**
   A *p*-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the *p*-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.
2. ***P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**
   Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.
3. **Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.**
   Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that *p*-values alone can ensure that a decision is correct or incorrect. The widespread use of "statistical significance" (generally interpreted as "$p \le 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.
4. **Proper inference requires full reporting and transparency**
   P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913

Taylor & Francis
Taylor & Francis Group

EDITORIAL

Moving to a World Beyond "$p < 0.05$"

Some of you exploring this special issue of *The American Statistician* might be wondering if it's a scolding from pedantic statisticians lecturing you about what *not* to do with *p*-values, without offering any real ideas of what *to do* about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.

1. "Don't" Is Not Enough

There's not much we can say here about the perils of *p*-values and significance testing that hasn't been said already for decades (Ziliak and McCloskey 2008; Hubbard 2016). If you're just arriving to the debate, here's a sampling of what not to do:

- Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the *p*-value passed some arbitrary threshold such as $p < 0.05$).
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your *p*-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

Don't. Don't. Just…don't. Yes, we talk a lot about don'ts. The *ASA Statement on p-Values and Statistical Significance* (Wasserstein and Lazar 2016) was developed primarily because after decades, warnings about the don'ts had gone mostly unheeded. The statement was about what not to do, because there is widespread agreement about the don'ts.

Knowing what not to do with *p*-values is indeed necessary, but it does not suffice. It is as though statisticians were asking users of statistics to tear out the beams and struts holding up the edifice of modern scientific research without offering solid construction materials to replace them. Pointing out old, rotting timbers was a good start, but now we need more.

Recognizing this, in October 2017, the American Statistical Association (ASA) held the Symposium on Statistical Inference, a two-day gathering that laid the foundations for this special issue of *The American Statistician*. Authors were explicitly instructed to develop papers for the variety of audiences interested in these topics. If you use statistics in research, business, or policymaking but are not a statistician, these articles were indeed written with YOU in mind. And if you are a statistician, there is still much here for you as well.

The papers in this issue propose many new ideas, ideas that in our determination as editors merited publication to enable broader consideration and debate. The ideas in this editorial are likewise open to debate. They are our own attempt to distill the wisdom of the many voices in this issue into an essence of good statistical practice as we currently see it: some do's for teaching, doing research, and informing decisions.

Yet the voices in the 43 papers in this issue do not sing as one. At times in this editorial and the papers you'll hear deep dissonance, the echoes of "statistics wars" still simmering today (Mayo 2018). At other times you'll hear melodies wrapping in a rich counterpoint that may herald an increasingly harmonious new era of statistics. To us, these are all the sounds of statistical inference in the 21st century, the sounds of a world learning to venture beyond "$p < 0.05$."

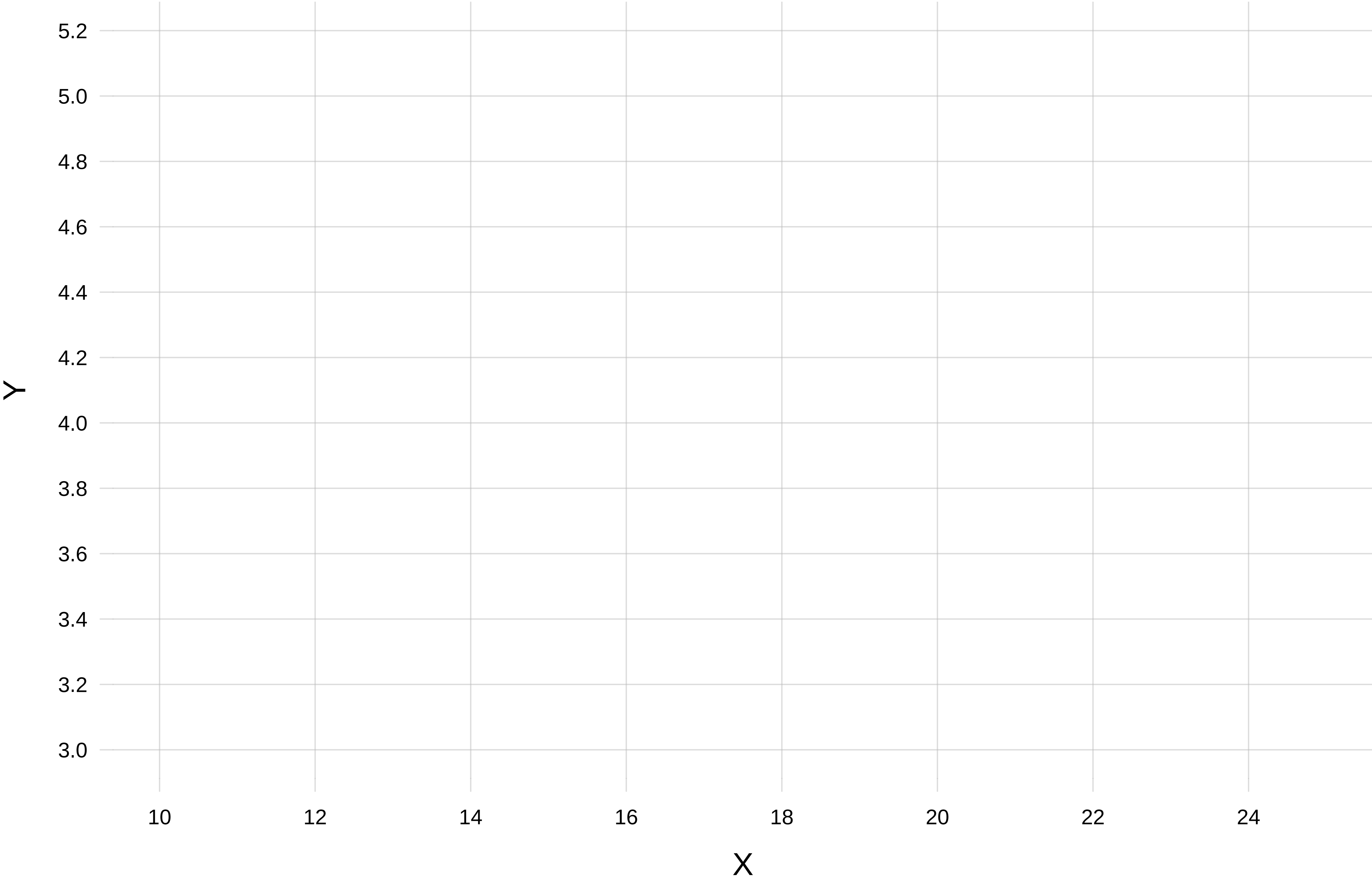
This is a world where researchers are free to treat "$p = 0.051$" and "$p = 0.049$" as not being categorically different, where authors no longer find themselves constrained to selectively publish their results based on a single magic number. In this world, where studies with "$p < 0.05$" and studies with "$p > 0.05$" are not automatically in conflict, researchers will see their results more easily replicated—and, even when not, they will better understand *why*. As we venture down this path, we will begin to see fewer false alarms, fewer overlooked discoveries, and the development of more customized statistical strategies. Researchers will be free to communicate all their findings in all their glorious uncertainty, knowing their work is to be judged by the quality and effective communication of their science, and not by their *p*-values. As "statistical significance" is used less, statistical thinking will be used more.

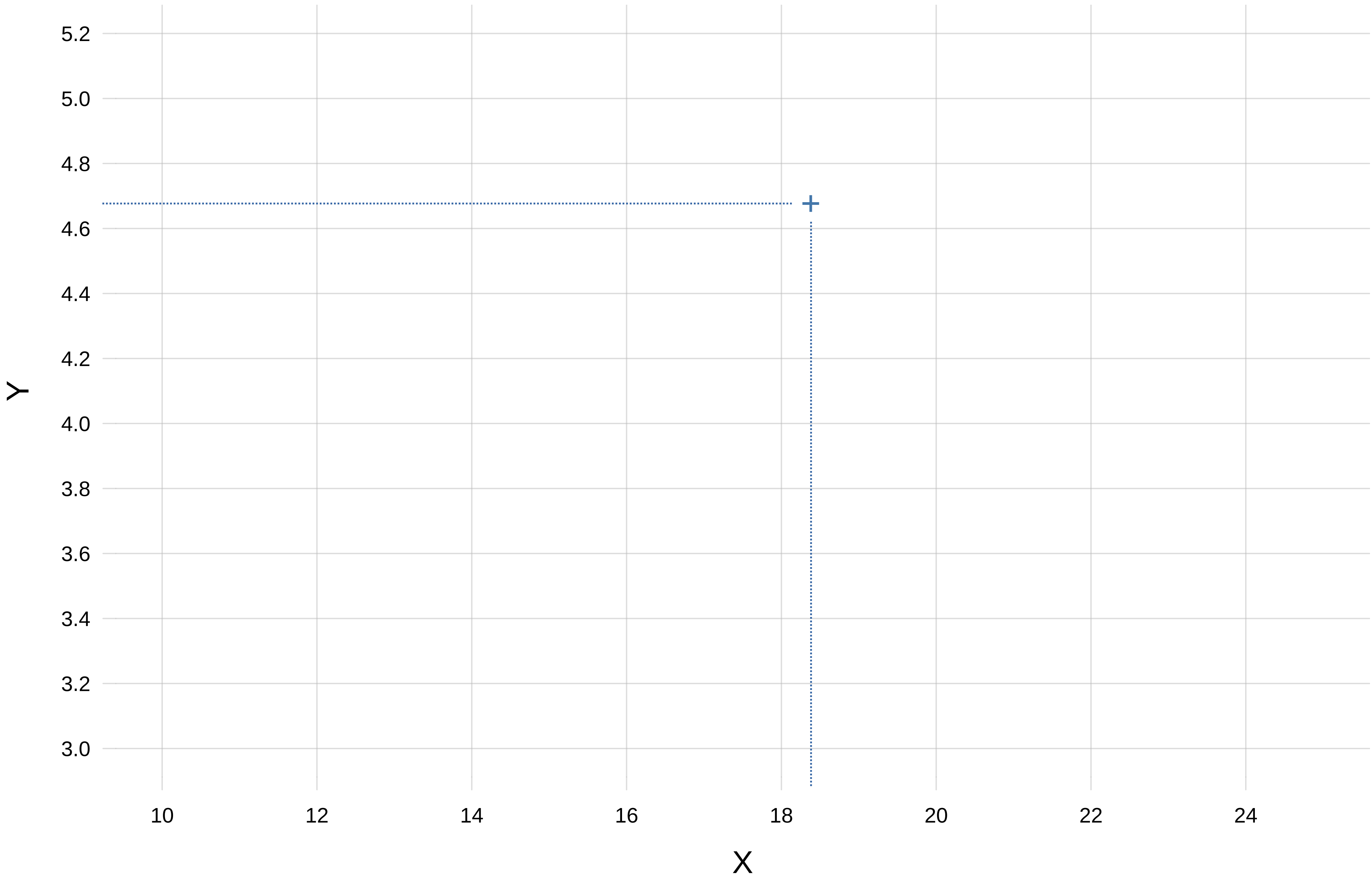
The *ASA Statement on P-Values and Statistical Significance* started moving us toward this world. As of the date of publication of this special issue, the statement has been viewed over 294,000 times and cited over 1700 times—an average of about 11 citations per week since its release. Now we must go further. That's what this special issue of *The American Statistician* sets out to do.

To get to the do's, though, we must begin with one more don't.

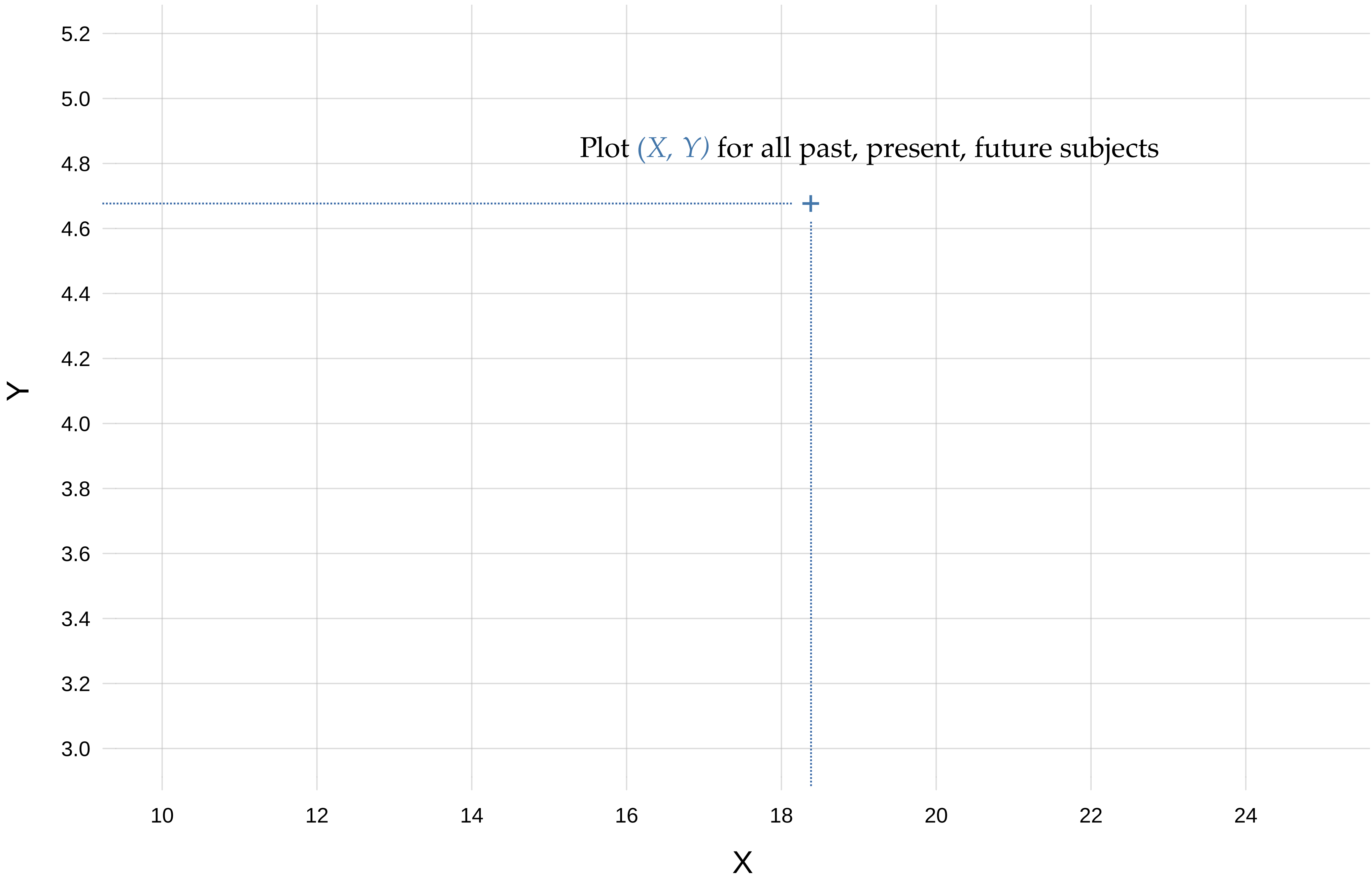
- Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the p-value passed some arbitrary threshold such as $p < 0.05$).
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

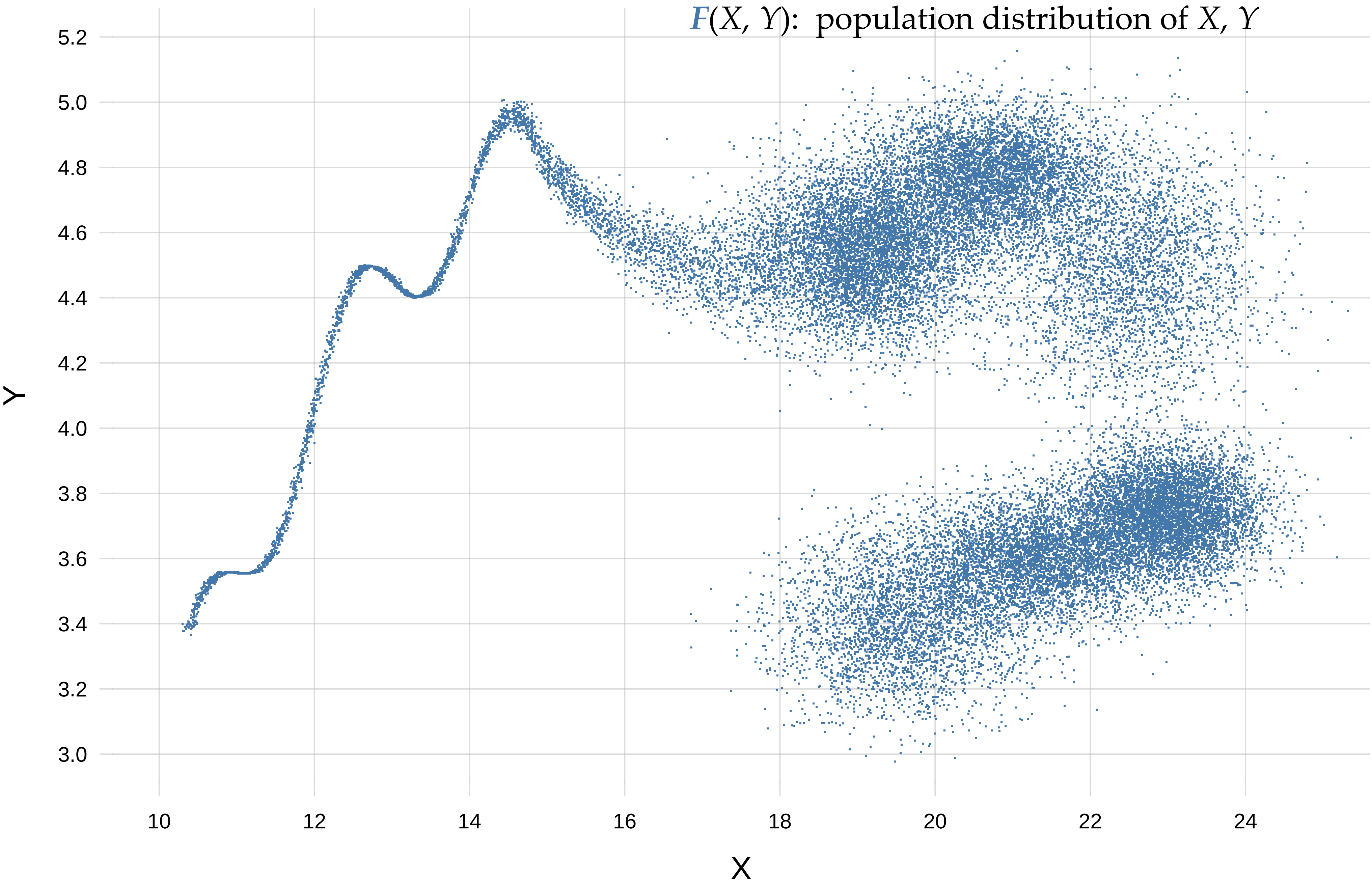
We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.


5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X



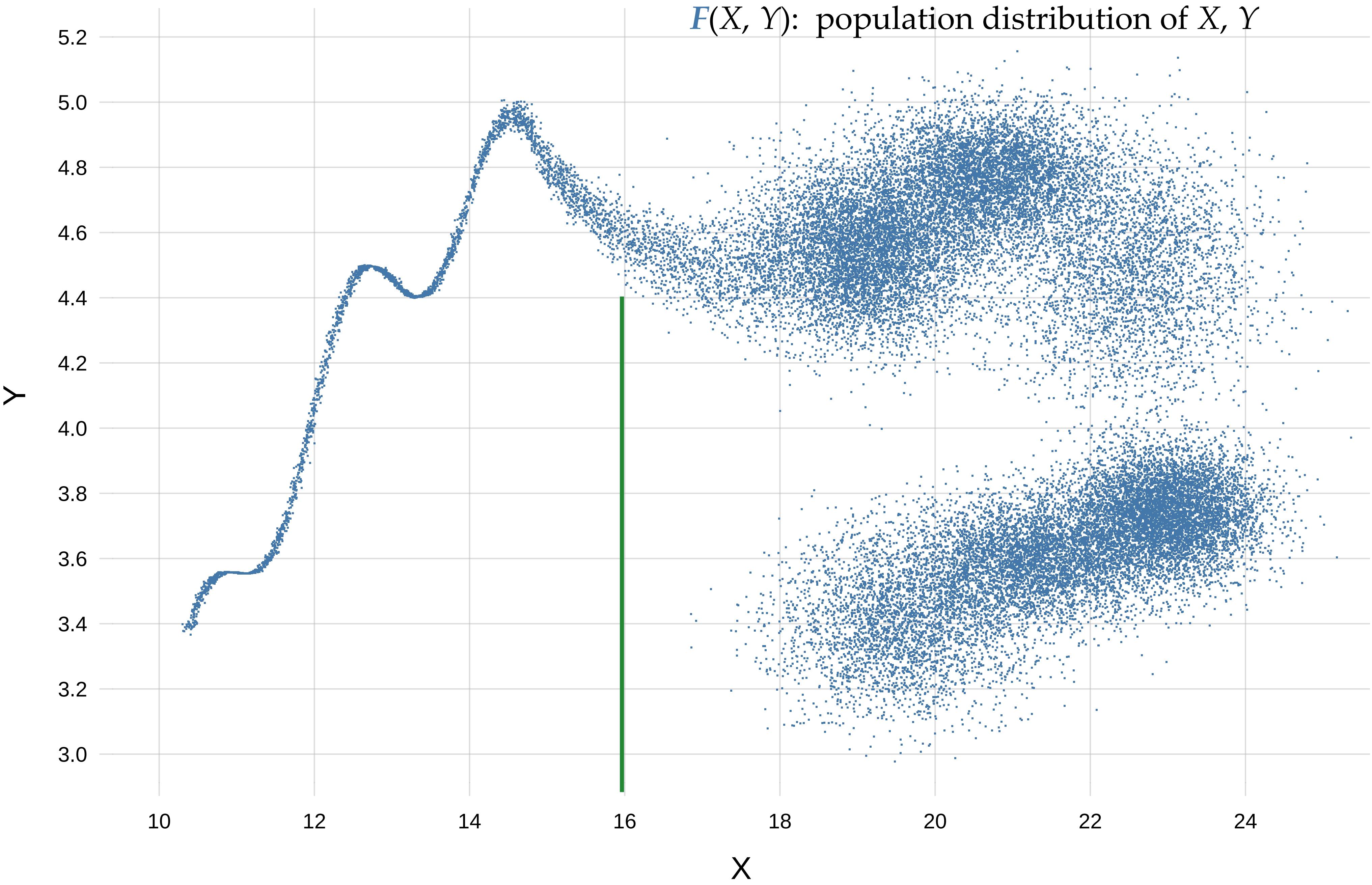
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X


Plot (X, Y) for all past, present, future subjects
Y
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
10
12
14
16
18
20
22
24
X



$F(X, Y)$: population distribution of X, Y
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X

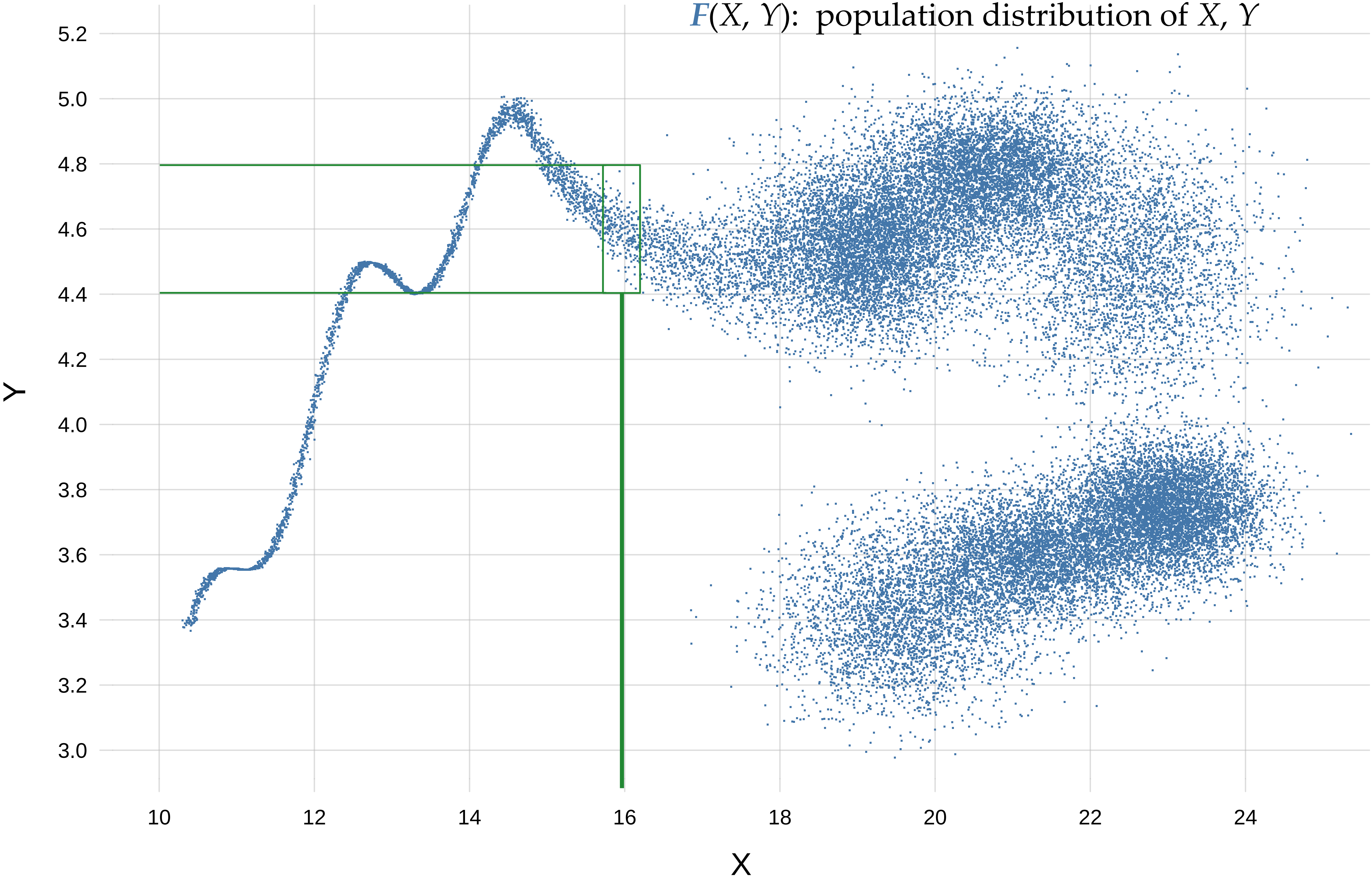
patient: $X = 16$



$F(X, Y)$: population distribution of X, Y
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X

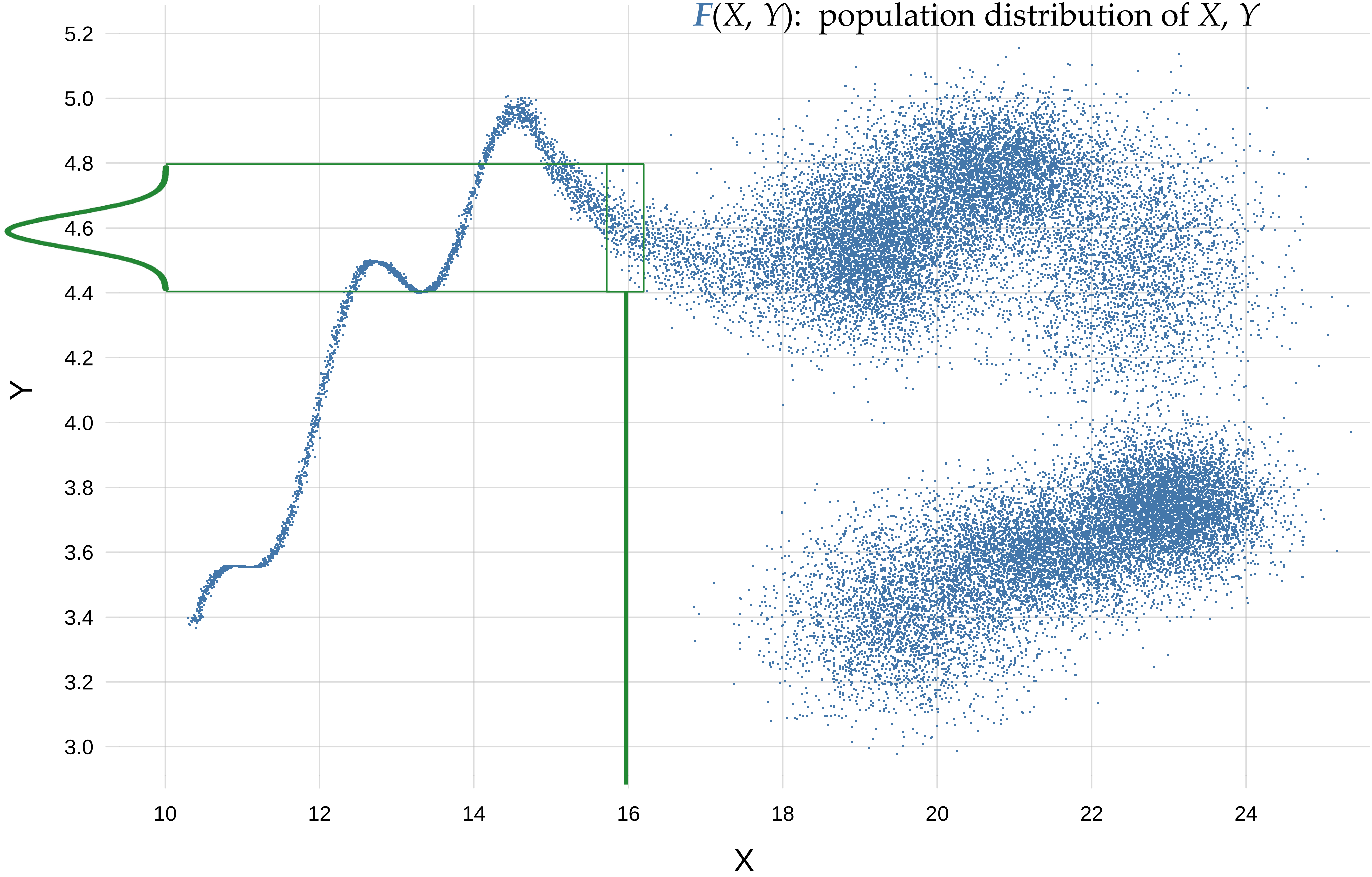
patient: $X = 16$

$\Rightarrow \quad Y \approx 4.5\text{–}4.7$

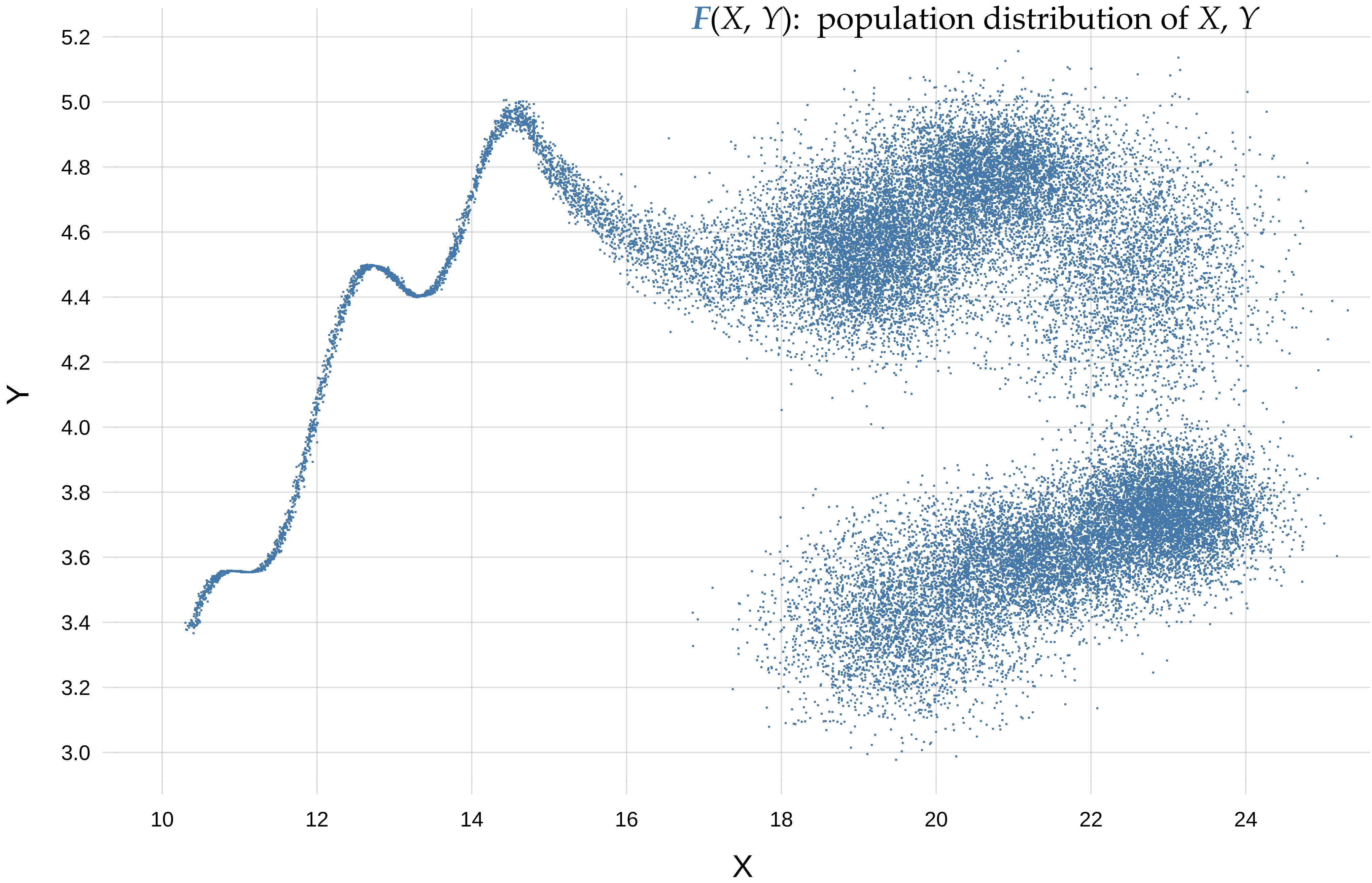


patient: $X = 16$

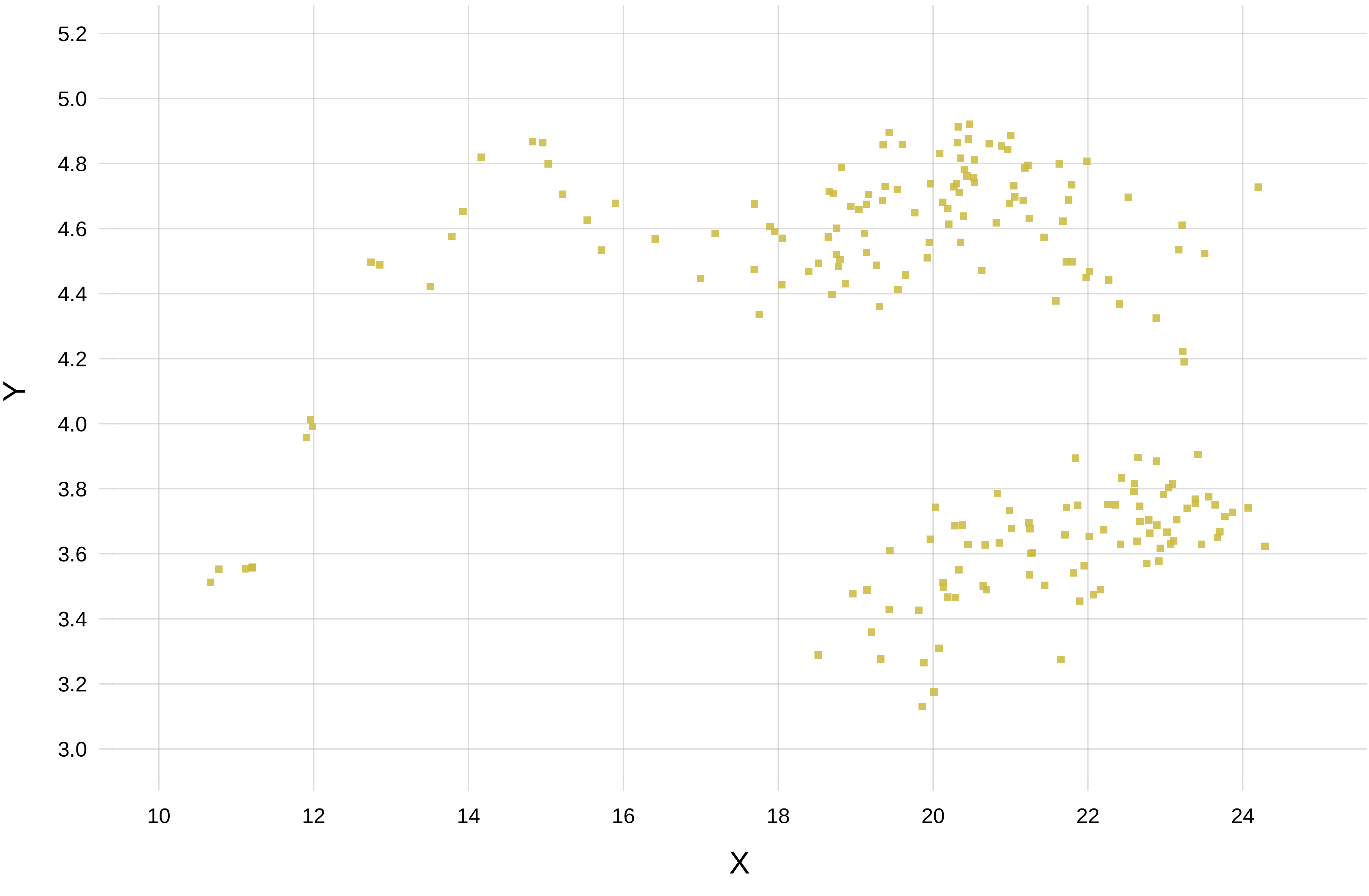
$\Rightarrow \quad Y \approx 4.5\text{–}4.7$



$$\mathrm{P}(y \mid x) = \boldsymbol{F}(y \mid x)$$



F(X, Y): population distribution of X, Y
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X



5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X

$$\mathrm{P}(y \mid x) = \boldsymbol{F}(y \mid x)$$

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

probability = average over all possible population distributions

$$\mathrm{P}(y \mid x) = \int \boldsymbol{F}(y \mid x)\ \mathrm{p}(\boldsymbol{F} \mid \text{data})\ \mathrm{d}\boldsymbol{F}$$

probability = average over all possible population distributions

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

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

probability = average over all possible population distributions

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

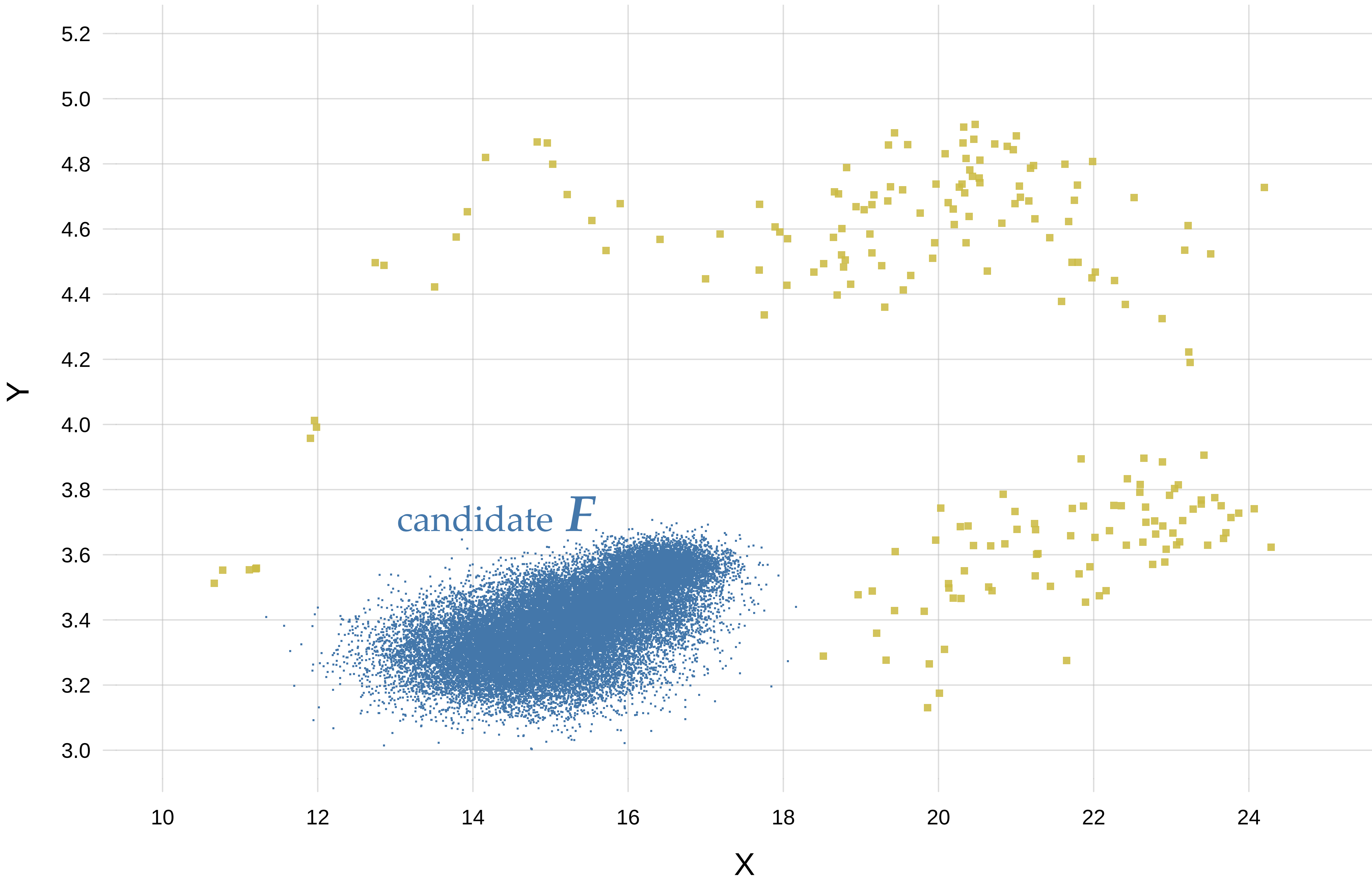
$$P(y \mid x) = \int F(y \mid x)\ p(F \mid \text{data})\ dF$$

probability = average over all possible population distributions

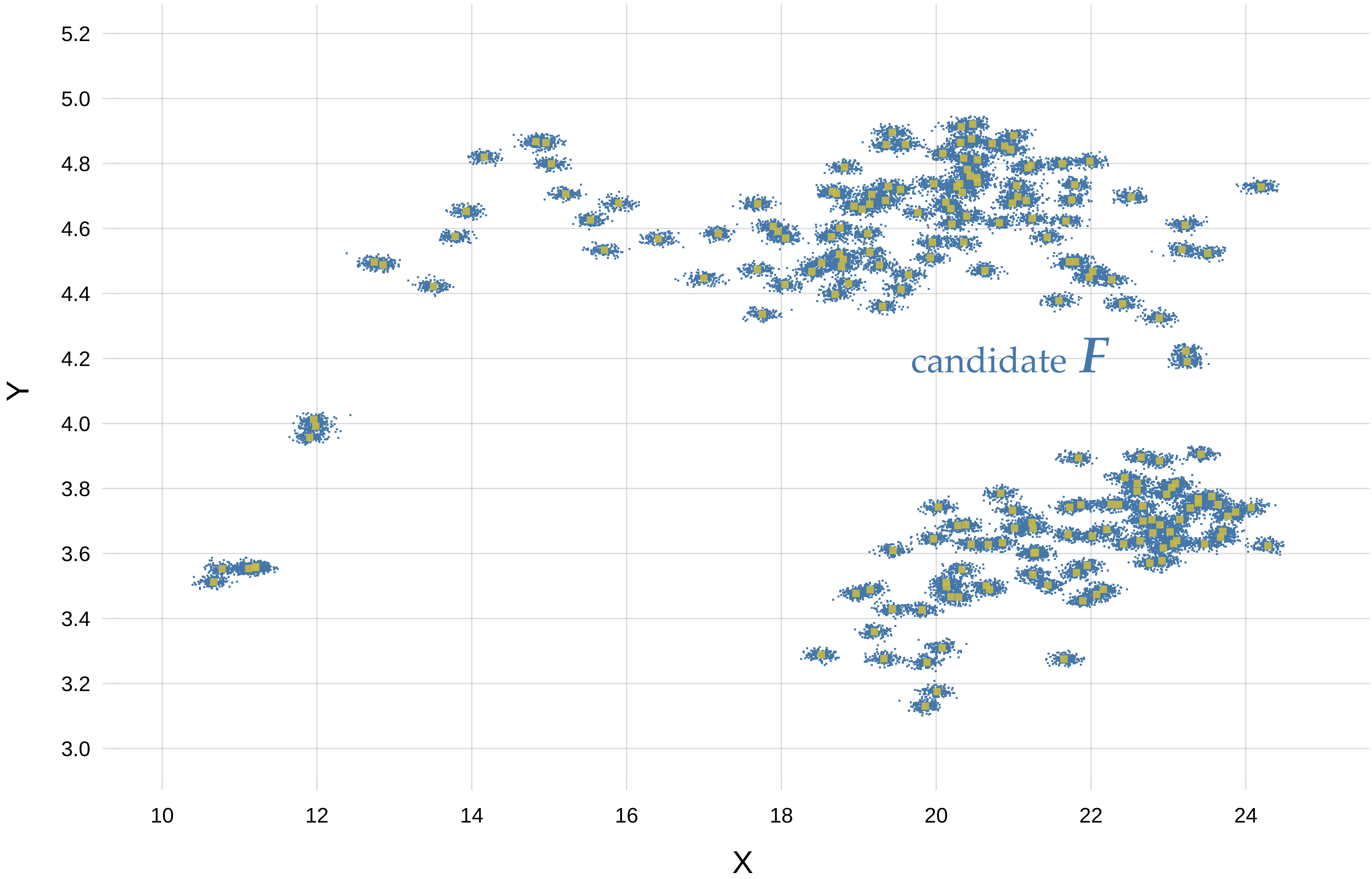
$$p(F \mid \text{data}) \propto \underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \ldots}_{\text{how well the 'candidate' distribution fits the data}} \times \underbrace{p(F \mid \text{prior info})}_{\text{extra-data knowledge}}$$

poor candidate: doesn't fit the data

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

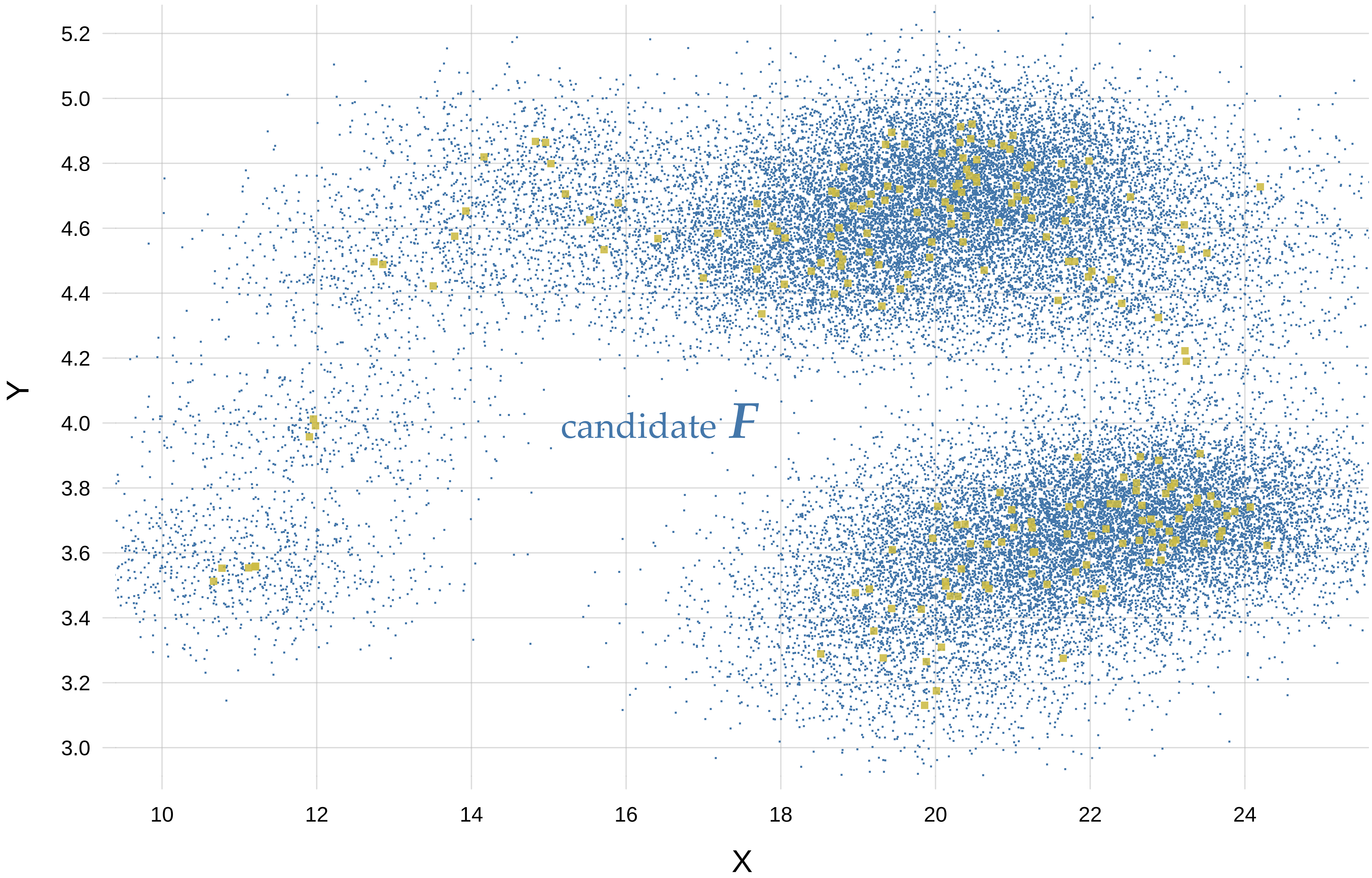


poor candidate: biologically implausible $\underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \ldots}_{\text{high}} \times \text{p}\underbrace{(F \mid \text{prior info})}_{\text{low}}$



reasonable candidate

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

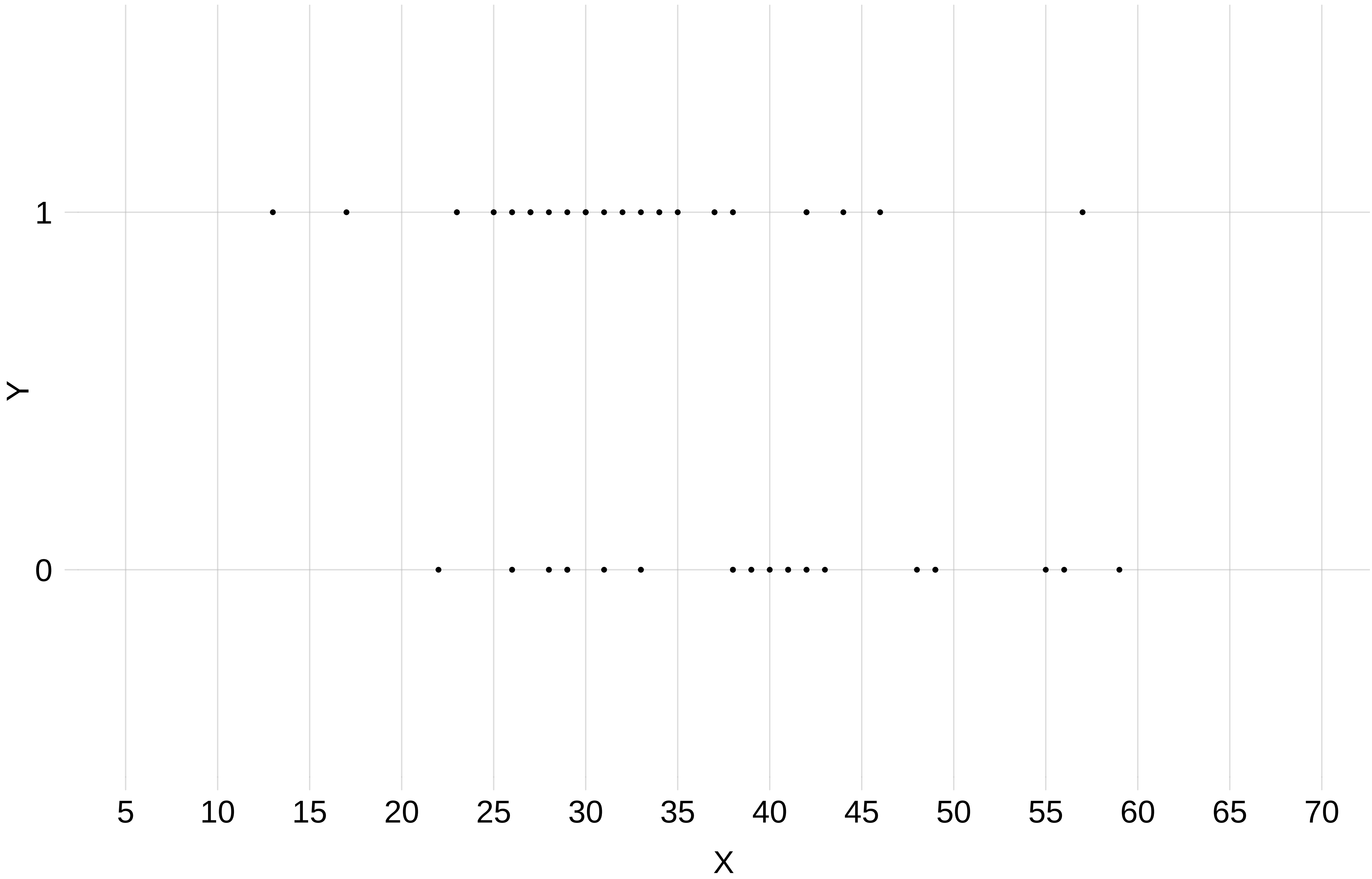


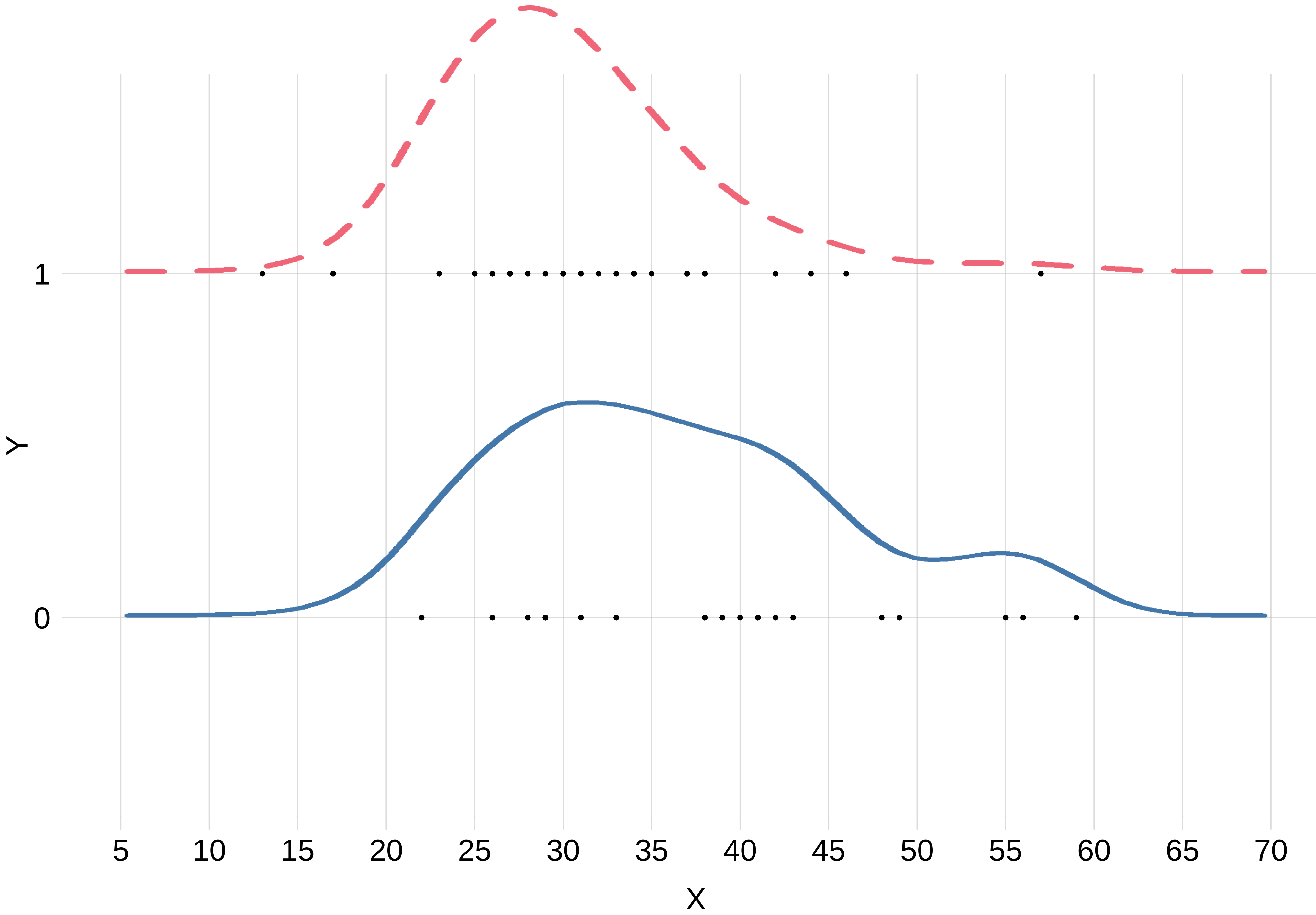
intuition $\rightarrow$ *mathematics*

$$\cancel{\textit{intuition} \longrightarrow \textit{mathematics}}$$

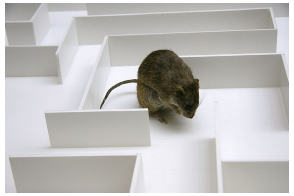
first principles → *mathematics* → *intuition*

('Bayesian')

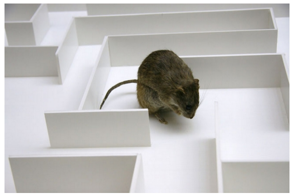

1
0
Y
5
10
15
20
25
30
35
40
45
50
55
60
65
70
X



1
0
Y
5
10
15
20
25
30
35
40
45
50
55
60
65
70
X



Healthy

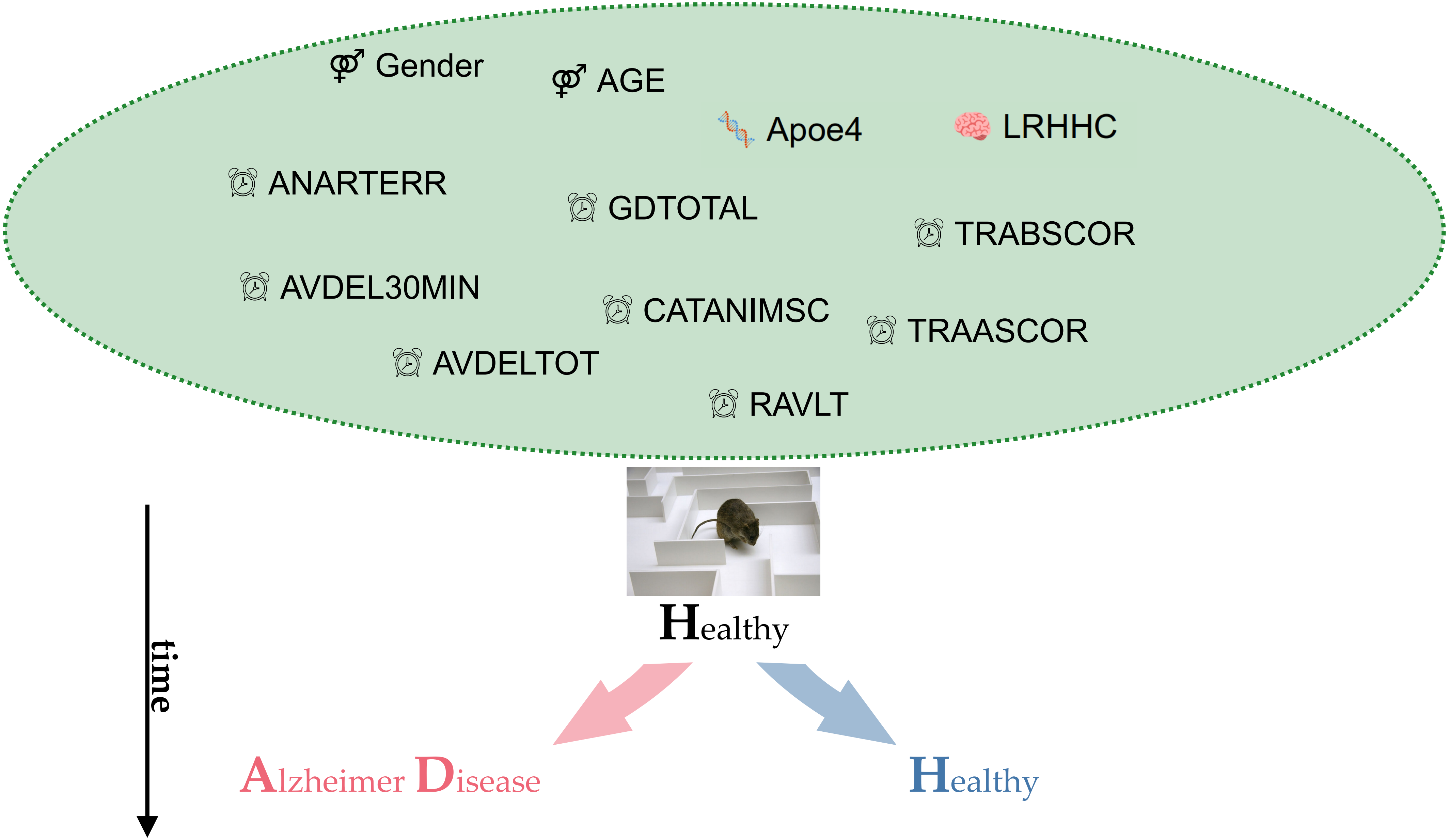


Healthy

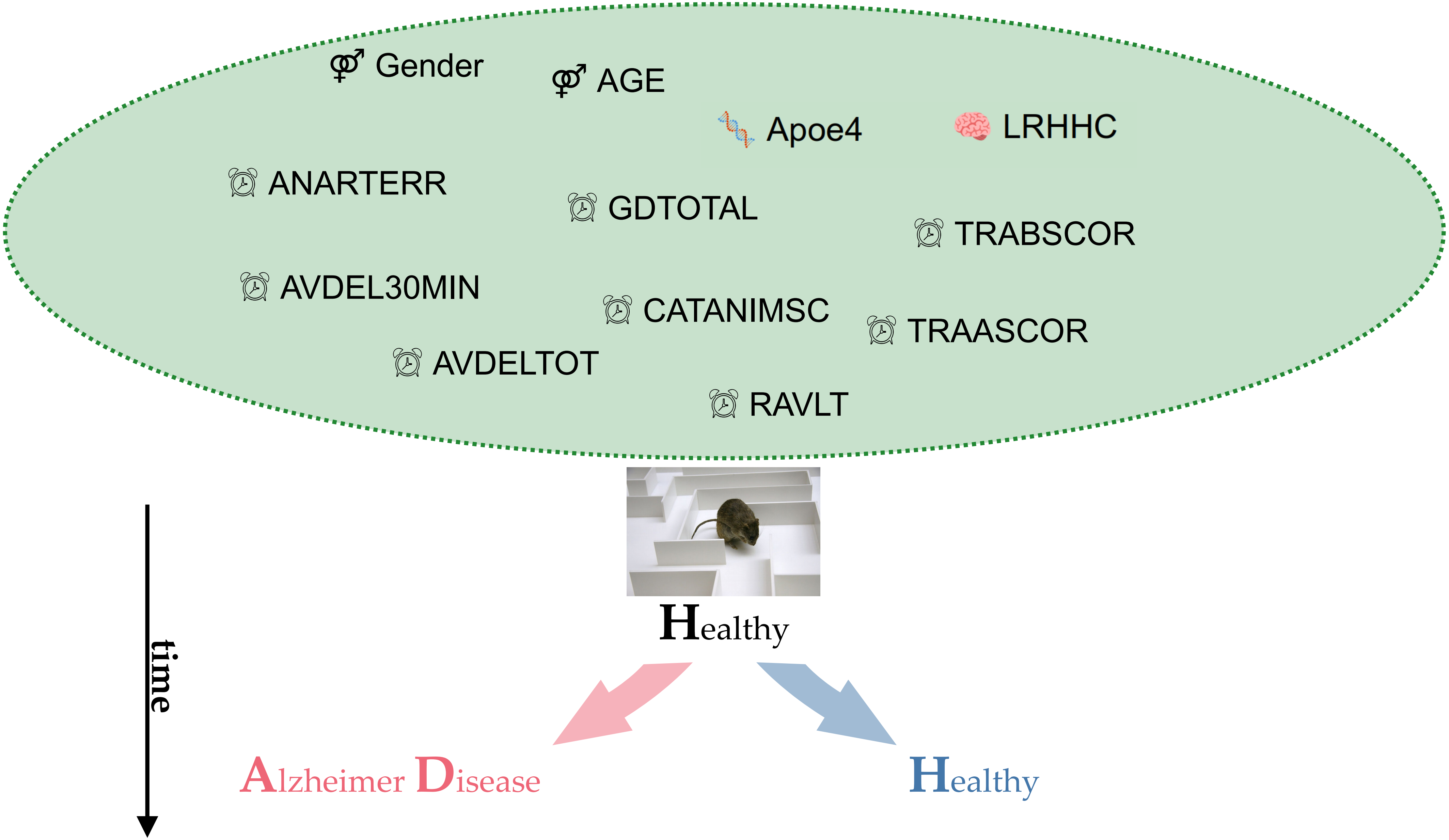
Alzheimer Disease

Healthy

time


Gender
AGE
Apoe4
LRHHC
ANARTERR
GDTOTAL
TRABSCOR
AVDEL30MIN
CATANIMSC
TRAASCOR
AVDELTOT
RAVLT
Healthy
time
Alzheimer Disease
Healthy

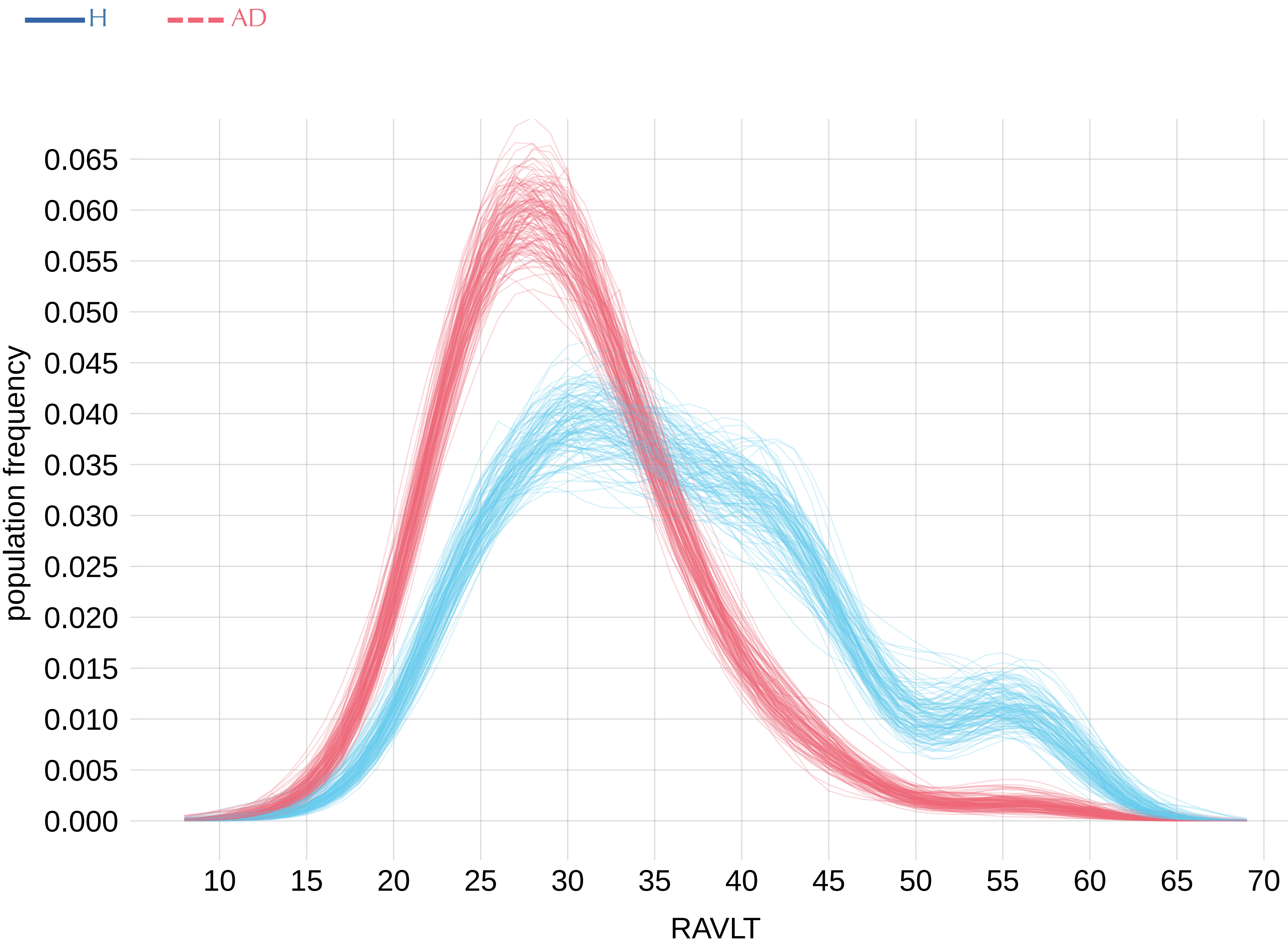
How ‘good’ are these features at prognosing the later onset of Alzheimer?

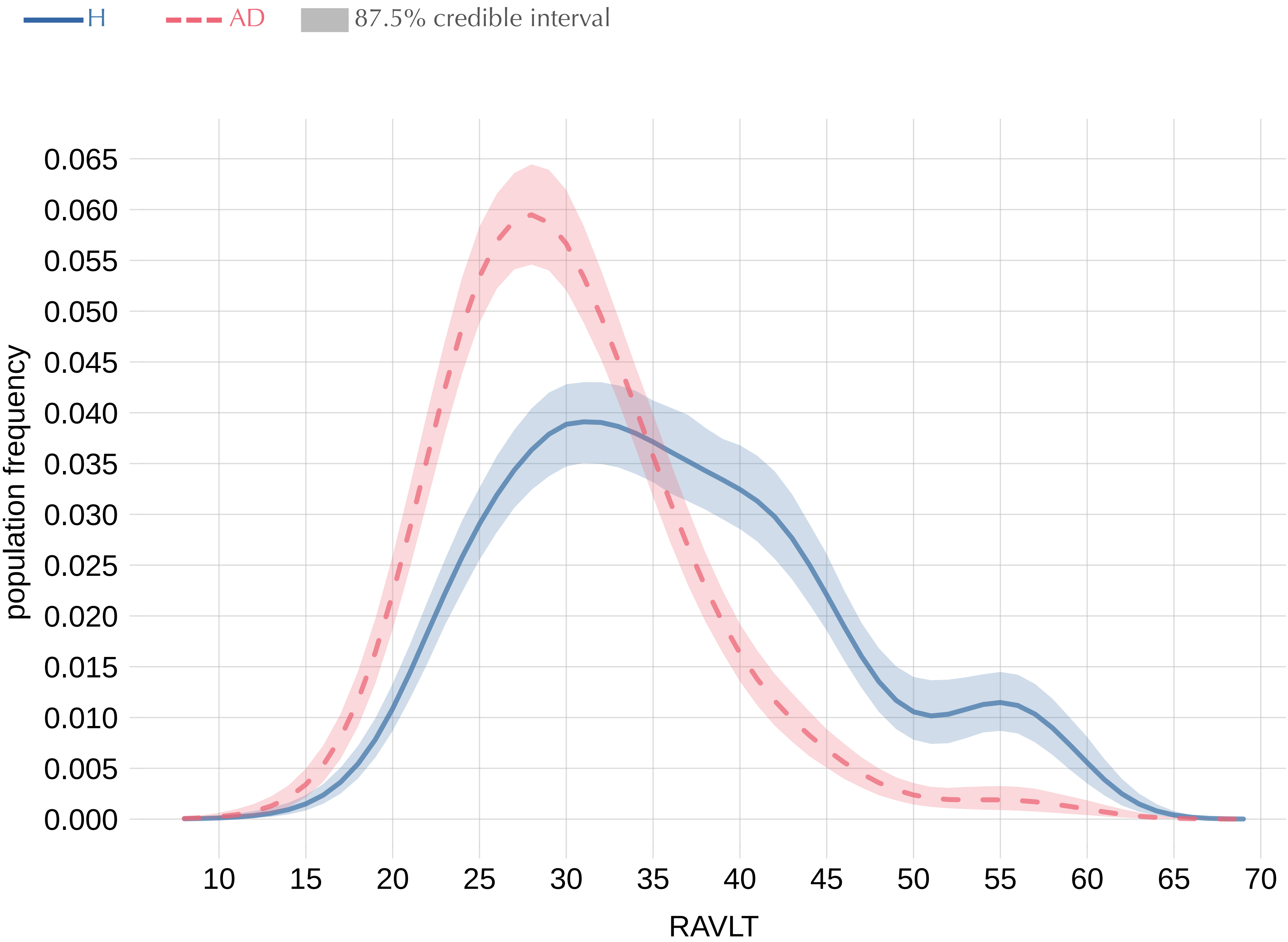


12+1 variates, 678 datapoints

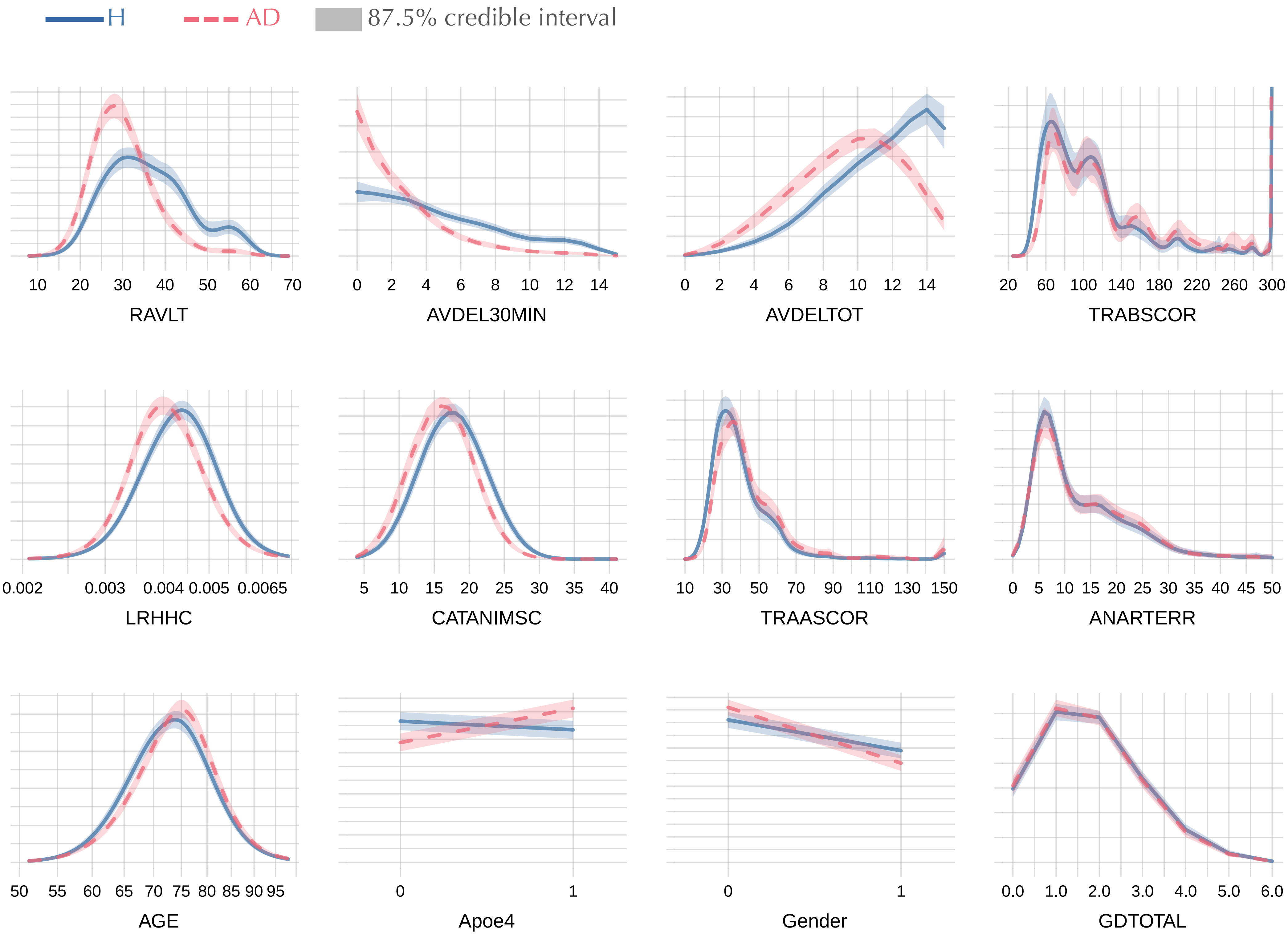
Computation time: ~65 h (3 parallel sessions to assess numeric convergence)

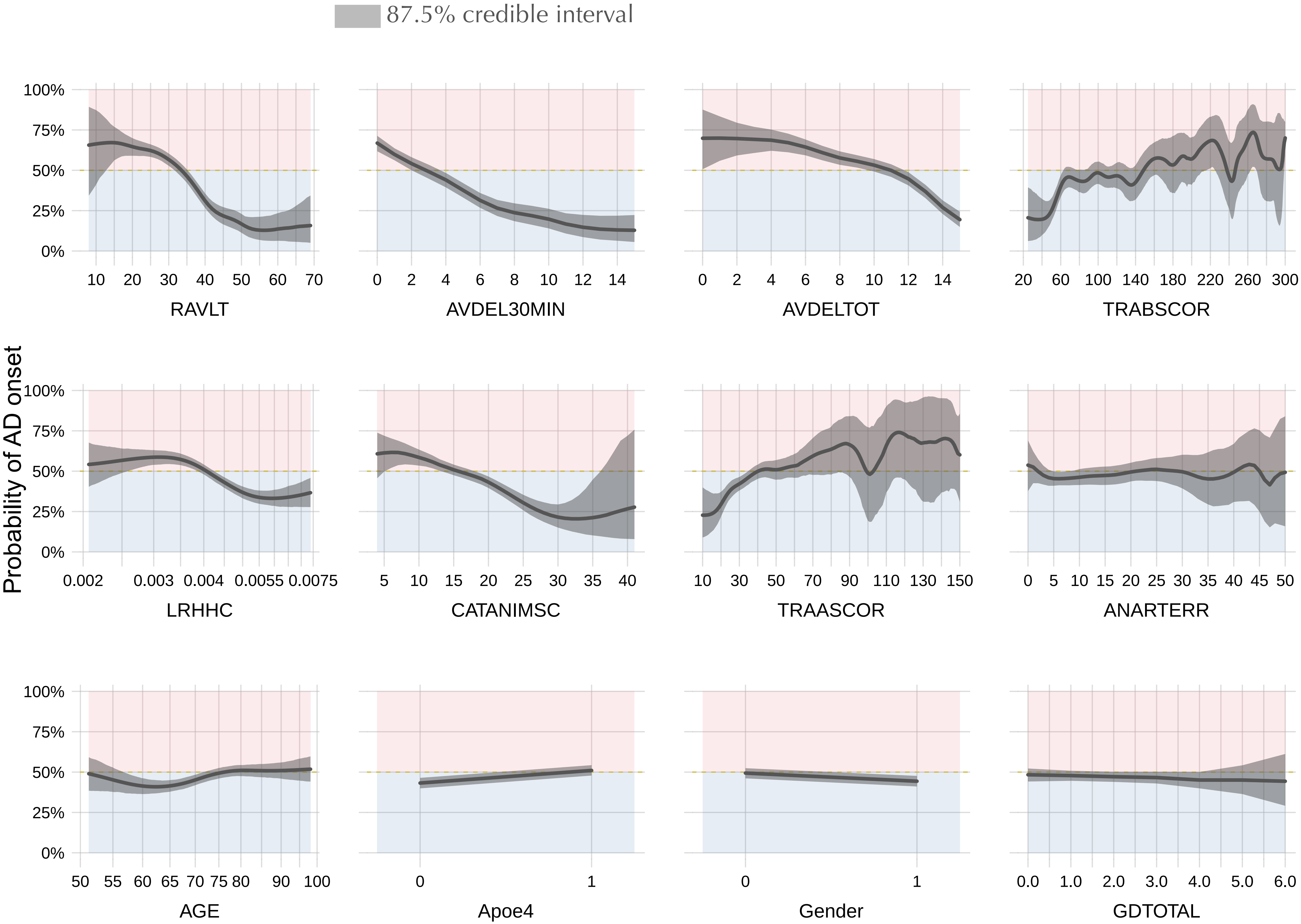
HPC


H
AD
population frequency
0.065
0.060
0.055
0.050
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
15
20
25
30
35
40
45
50
55
60
65
70
RAVLT



H
AD
87.5% credible interval
population frequency
0.065
0.060
0.055
0.050
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
15
20
25
30
35
40
45
50
55
60
65
70
RAVLT


H
AD
87.5% credible interval
RAVLT
AVDEL30MIN
AVDELTOT
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
ANARTERR
AGE
Apoe4
Gender
GDTOTAL

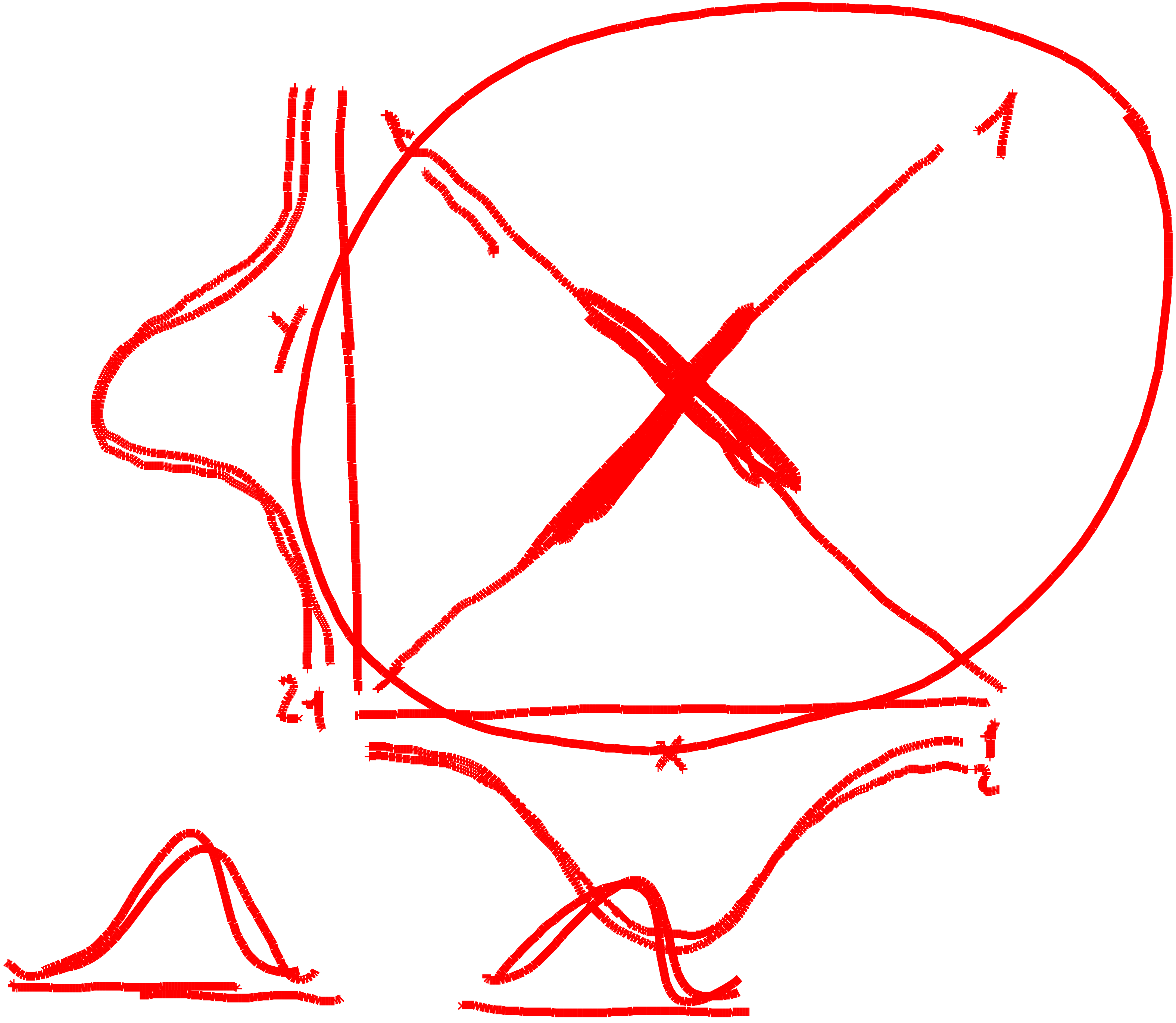

87.5% credible interval
Probability of AD onset
100%
75%
50%
25%
0%
RAVLT
AVDEL30MIN
AVDELTOT
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
ANARTERR
AGE
Apoe4
Gender
GDTOTAL

Interesting characteristics of $\boldsymbol{F}(Y, X)$:

- Several high-density regions in the 12D space

- Some features seem more robust if used in a 'discriminative' way: $\mathrm{P}(Y \mid X)$, others in a 'generative' way: $\mathrm{P}(X \mid Y)$

$$\mathrm{P}(Y \mid X_{\mathrm{d}}, X_{\mathrm{g}}) \propto \mathrm{P}(X_{\mathrm{g}} \mid Y)\, \mathrm{P}(Y \mid X_{\mathrm{d}})$$



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

“Language is a product of, and reflects, thinking.
Sloppy usage reflects sloppy thinking, a kind of thinking
incompatible with good scientific habits of mind”

(D. J. Helfand)

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**: ???✓✓✓

increasing importance

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 ⇒ **A** has 'importance = **0**'
- Discard **B**: still 100% win ⇒ **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**: ???✔✔✔

increasing importance ↓

discard **A**: still 100% win

discard **B**: still 100% win

discard **C**: 1/9 winning chance

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

increasing importance

Scenario 1:
choose one clue

C

A
B

Scenario 3:
discard one clue

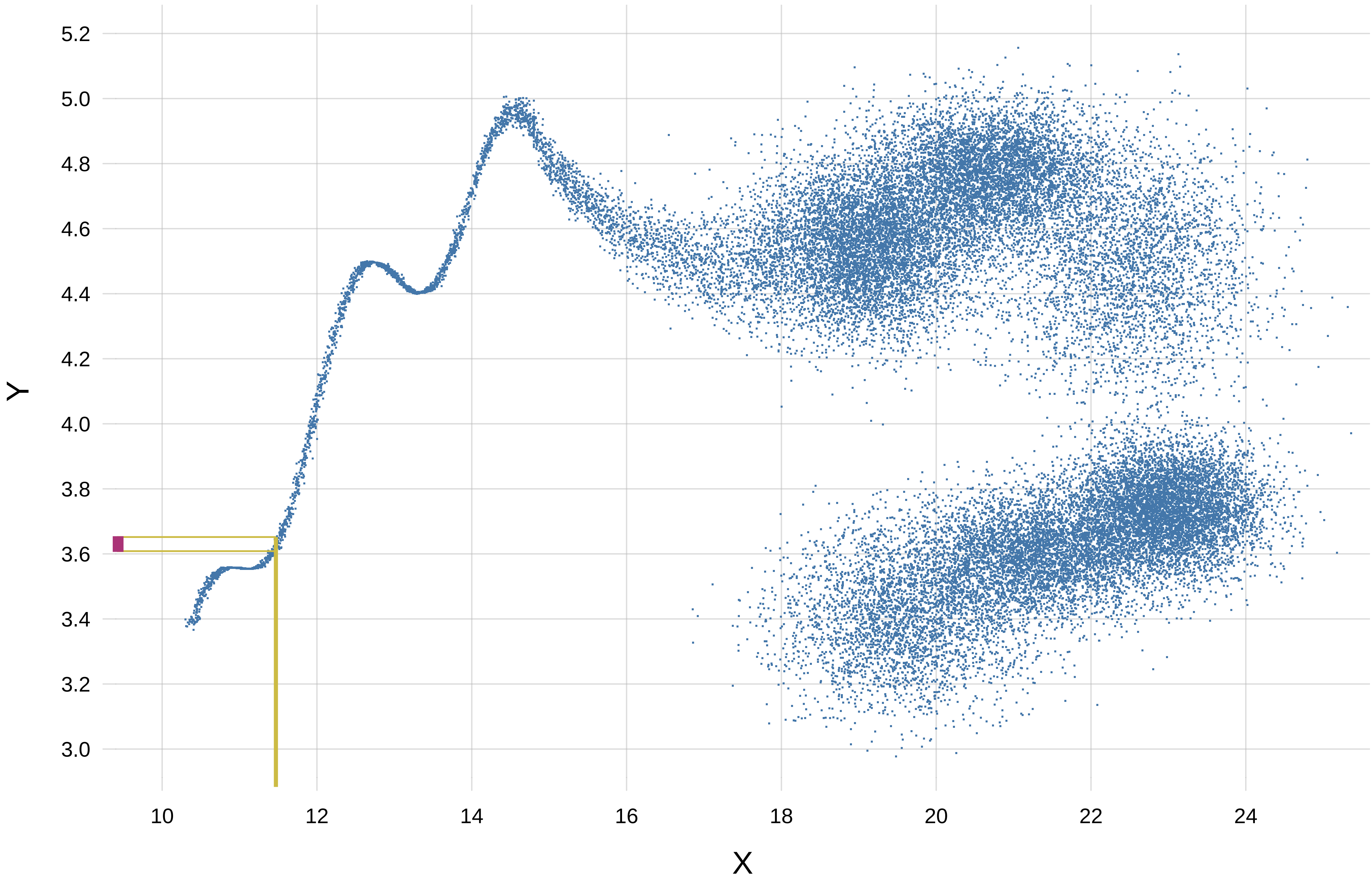
A
B

C

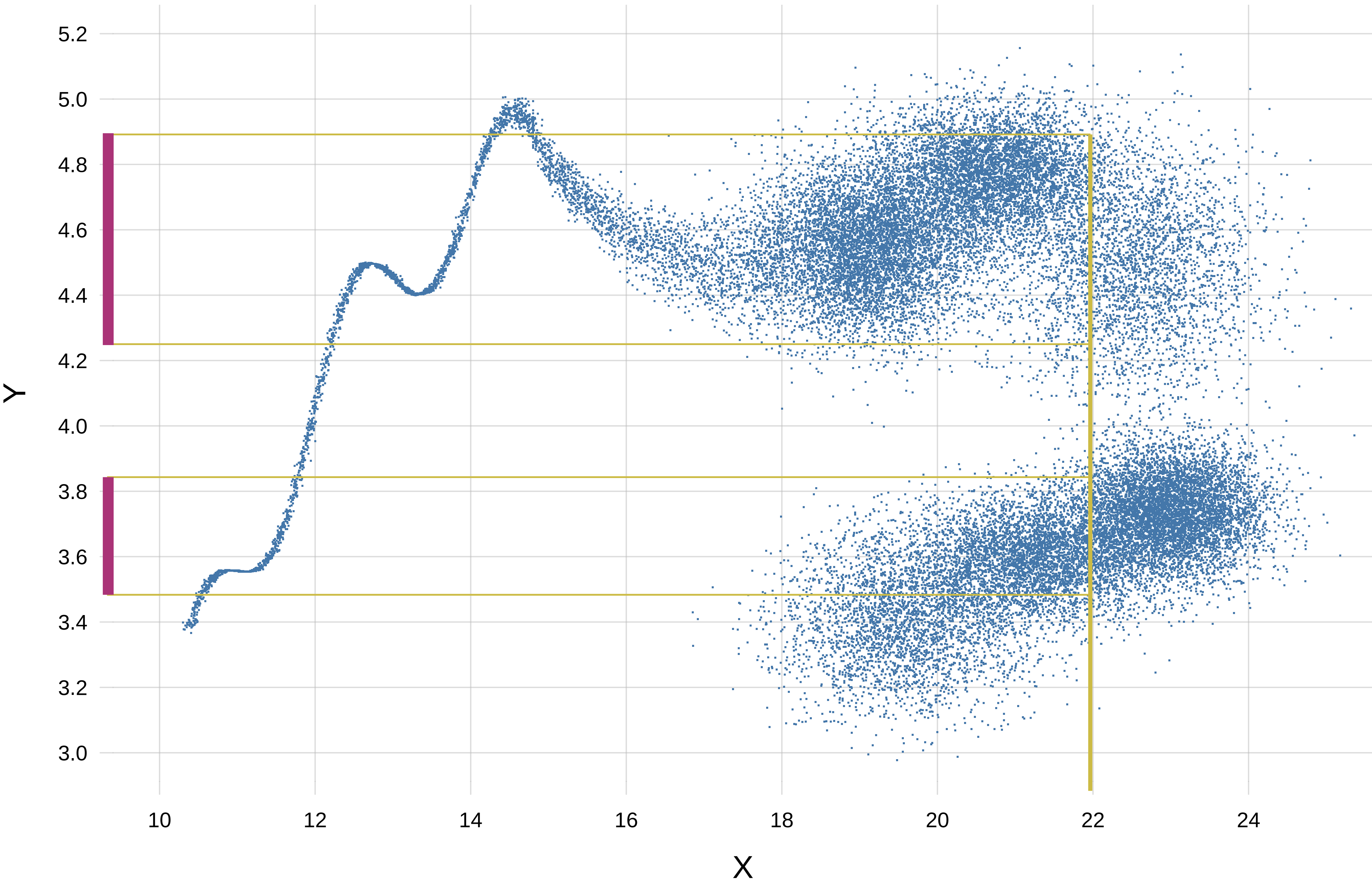
increasing importance

‘Importance’ or ‘predictive power’ of X is *context-dependent*
(which other features are we considering?)

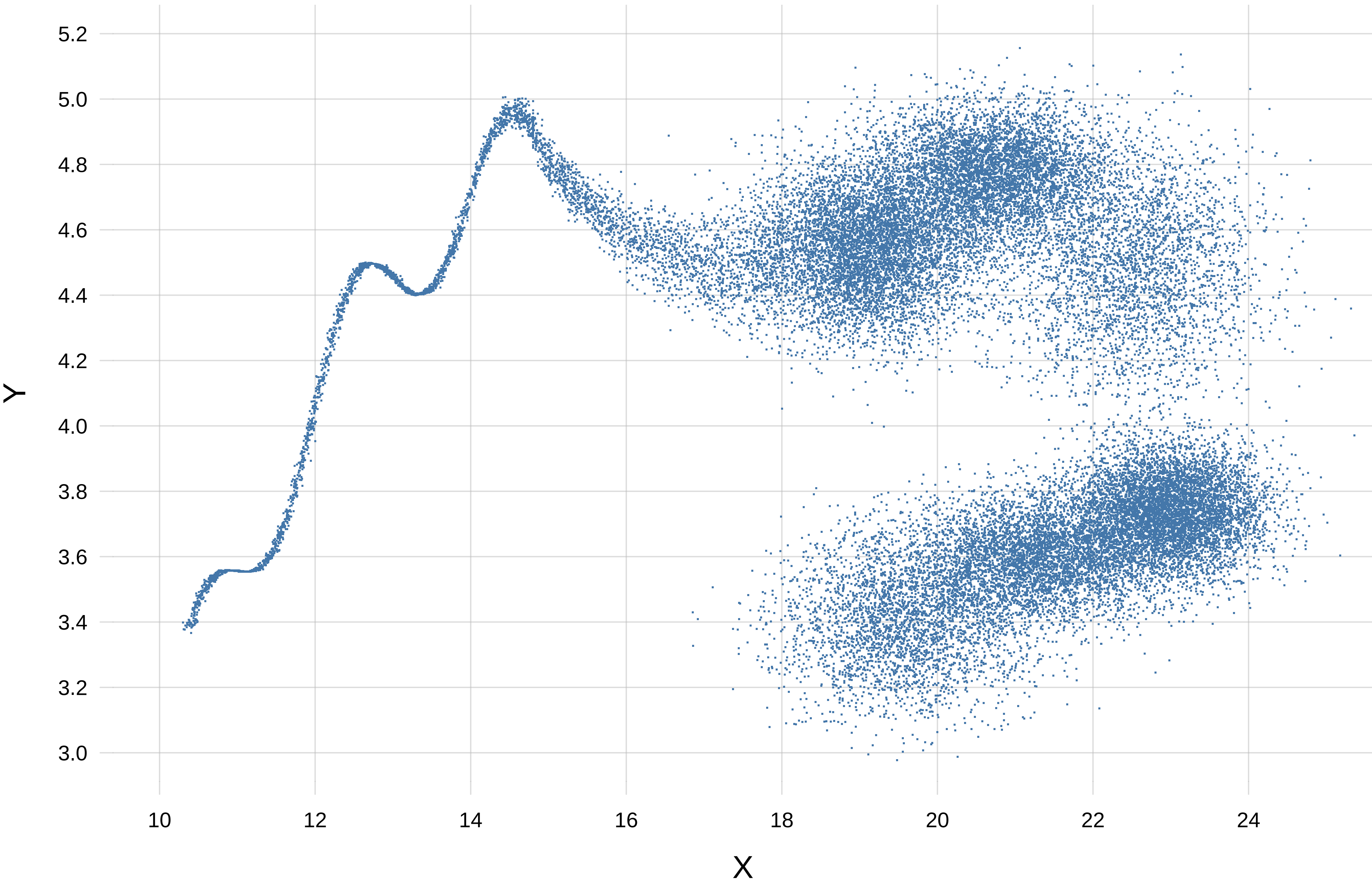
$x = 11.5 \;\Rightarrow\; y \approx 3.60\text{–}3.65$

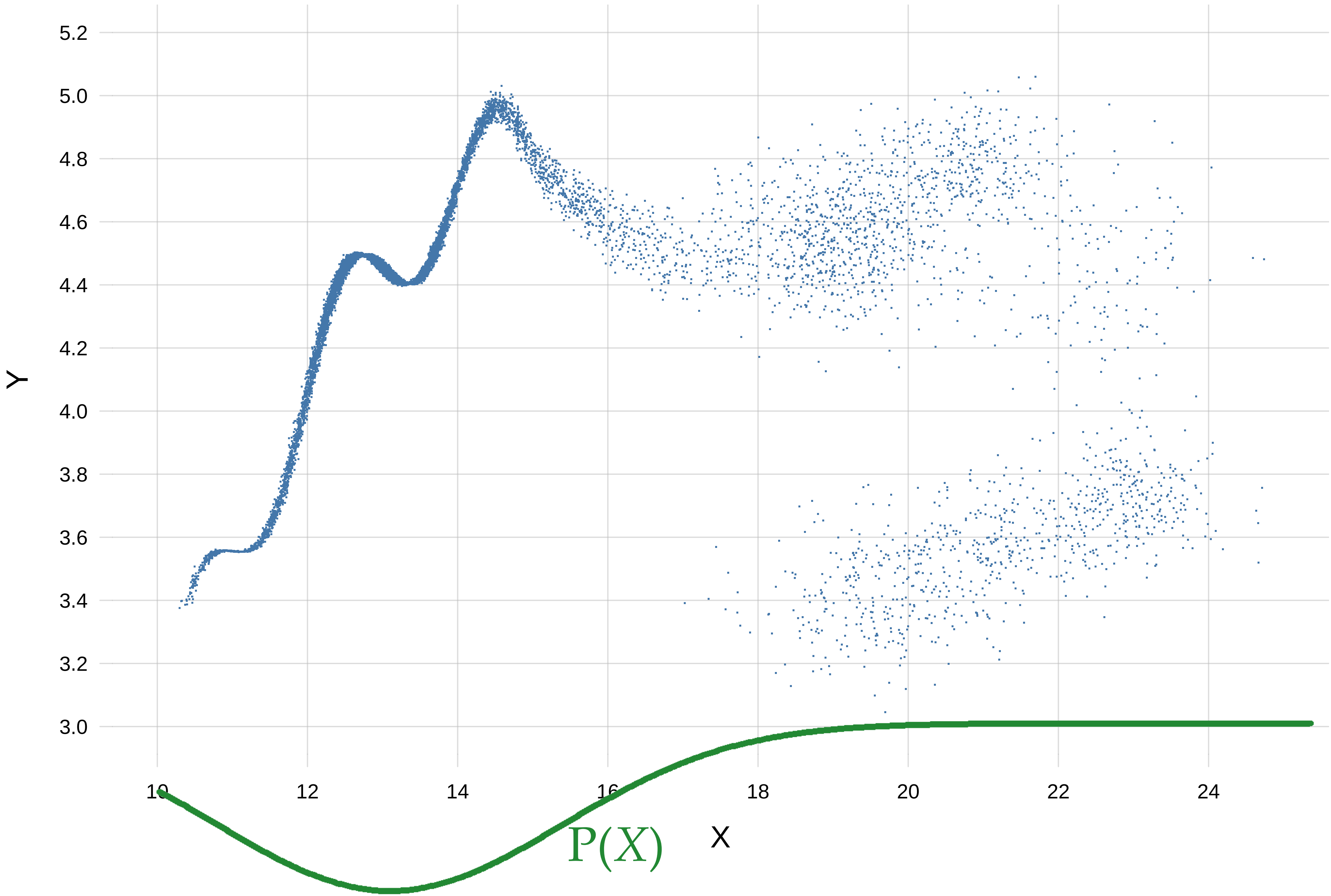


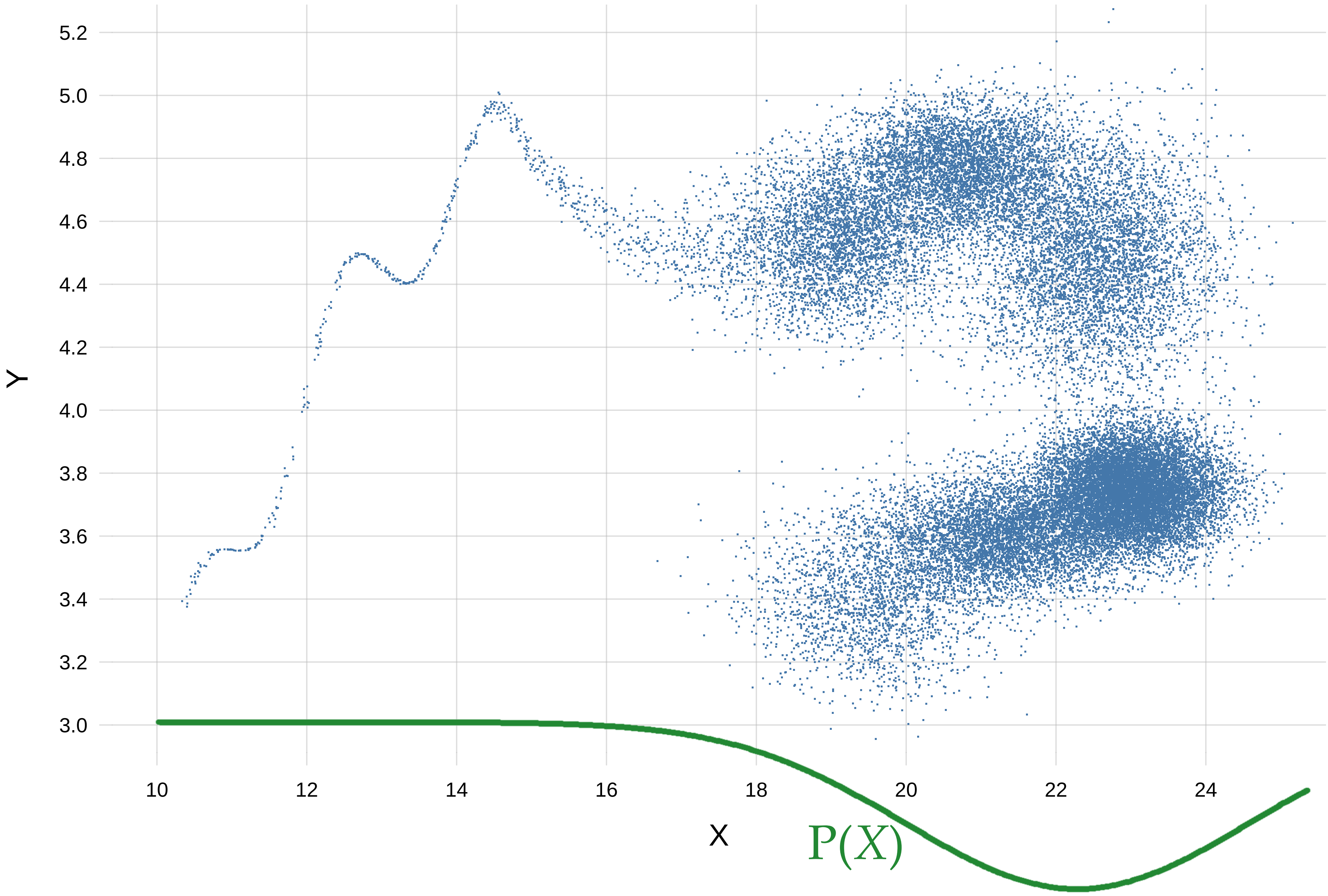
$x = 22 \;\Rightarrow\; y \approx 3.50\text{–}3.85$ or $4.25\text{–}4.90$



What is the ‘overall predictive power’ of *X*?




5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X
P(X)



Y
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
10
12
14
16
18
20
22
24
X
P(X)

The 'importance' or 'predictive power' of X depends on $\mathrm{P}(X)$

The ‘importance’ or ‘predictive power’ of X depends on $\mathrm{P}(X)$

Careful with ‘data balancing’!

Information Theory

The Bell System Technical Journal

Vol. XXVII *July, 1948* *No. 3*

A Mathematical Theory of Communication

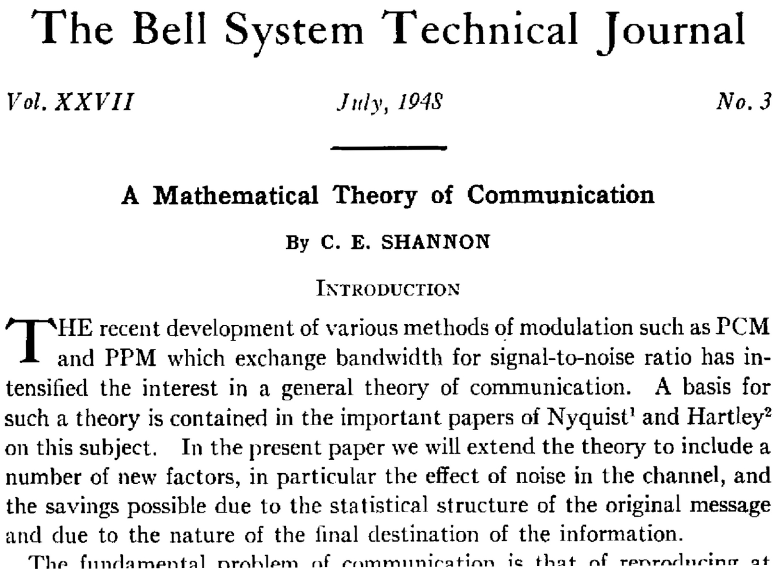
By C. E. SHANNON

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 Nyquist[1] and Hartley[2] 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

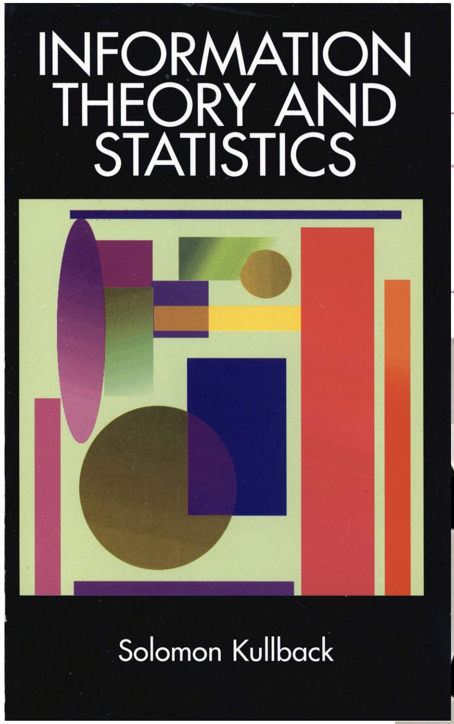
A Mathematical Theory of Communication

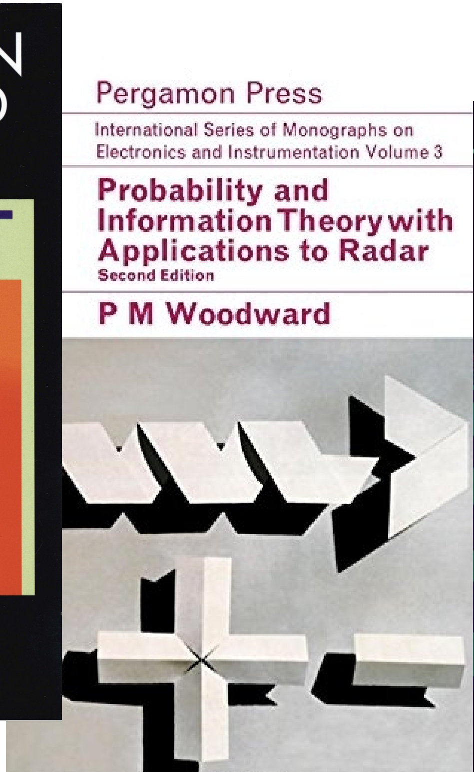
By C. E. SHANNON

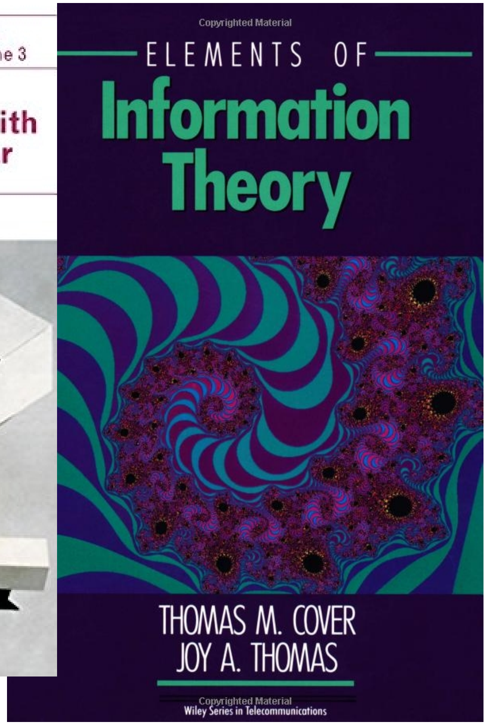
INTRODUCTION

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 Nyquist[1] and Hartley[2] 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.

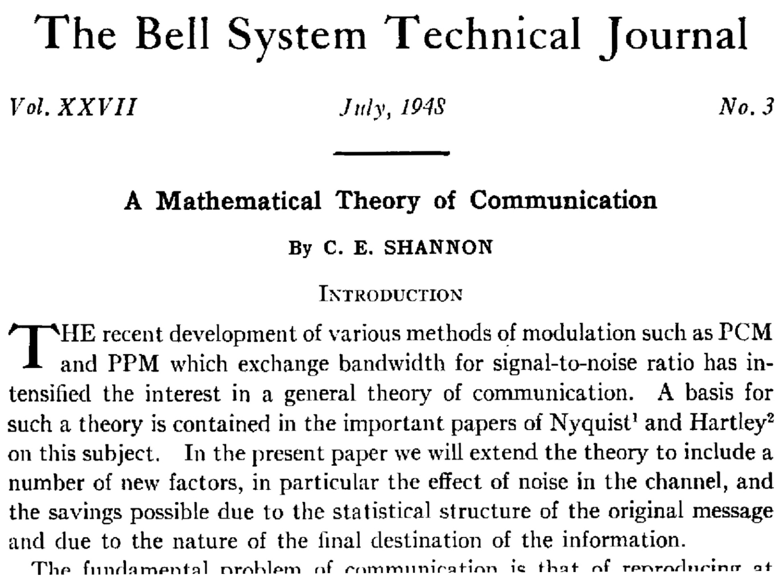
The fundamental problem of communication is that of reproducing at







Information Theory



The Bell System Technical Journal

Vol. XXVII — July, 1948 — No. 3

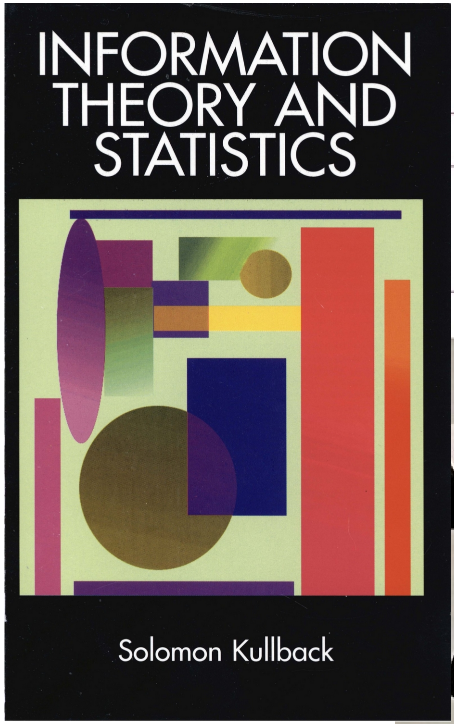
A Mathematical Theory of Communication

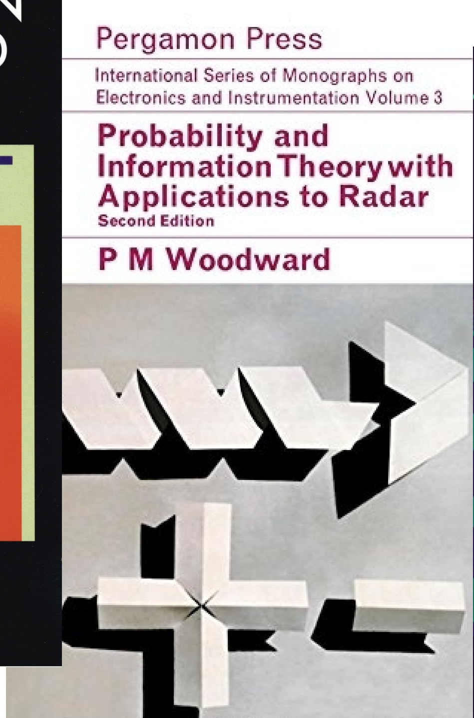
By C. E. SHANNON

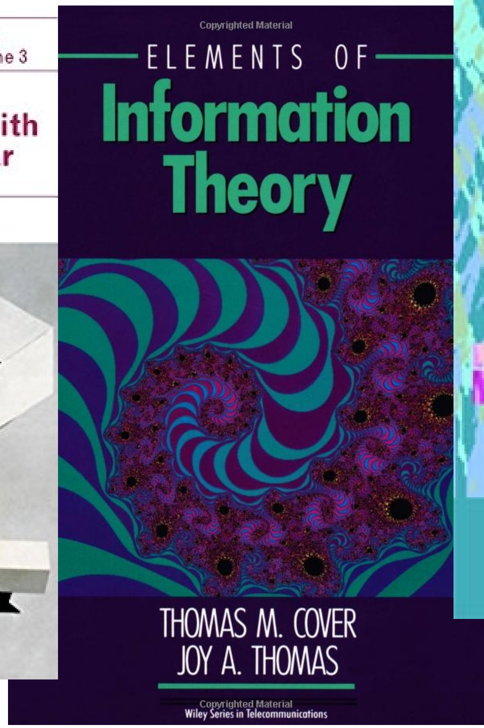
INTRODUCTION

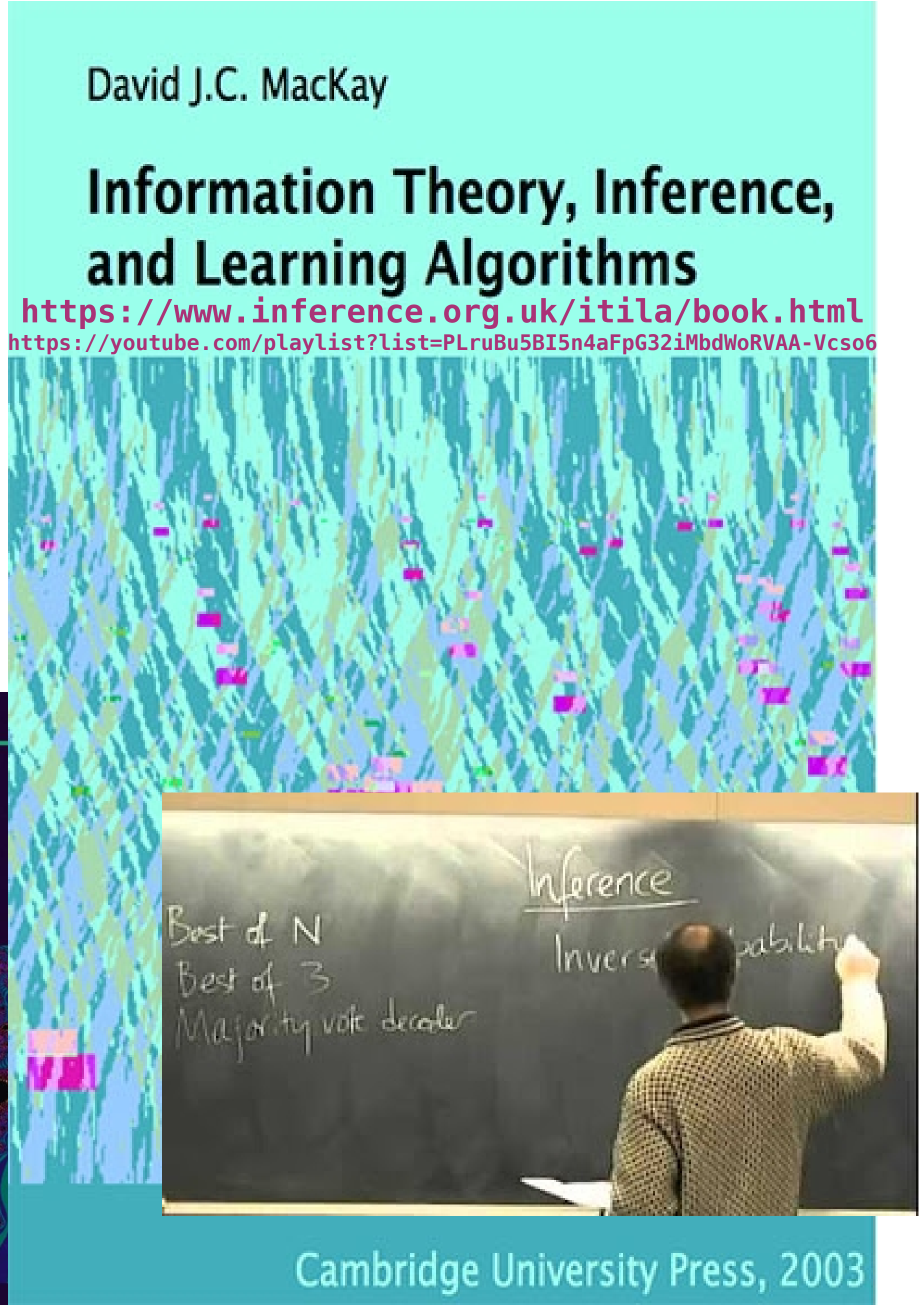
THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. 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









https://www.inference.org.uk/itila/book.html

https://youtube.com/playlist?list=PLruBu5BI5n4aFpG32iMbdWoRVAA-Vcso6

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

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

'predictive power' of X for Y **:=** **Mutual information** between Y and X
(mean transinformation content)

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

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

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

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

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

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

IEC 80000-13

Edition 1.0 2008-03

INTERNATIONAL STANDARD

NORME INTERNATIONALE

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

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

IEC 80000-13

Edition 1.0 2008-03

INTERNATIONAL STANDARD

INFORMATION SCIENCE AND TECHNOLOGY				QUANTITIES
Item No.	**Name**	**Symbol**	**Definition**	**Remarks**
13-24 (*902*)	information content fr *quantité* (f) *d'information*	$I(x)$	$I(x) = \text{lb}\,\frac{1}{p(x)}\ \text{Sh} = \text{lg}\,\frac{1}{p(x)}\ \text{Hart} =$ $\ln \frac{1}{p(x)}\ \text{nat}$ where $p(x)$ is the probability of event x	See ISO/IEC 2382-16, item 16.03.02. See also IEC 60027-3.
13-25 (*903*)	entropy fr *entropie* (f)	H	$H(X) = \sum_{i=1}^{n} p(x_i) I(x_i)$ for the set $X = \{x_1, \ldots, x_n\}$ where $p(x_i)$ is the probability and $I(x_i)$ is the information content of event x_i	See ISO/IEC 2382-16, item 16.03.03.
13-30 (*908*)	joint information content fr *quantité* (f) *d'information conjointe*	$I(x, y)$	$I(x, y) = \text{lb}\,\frac{1}{p(x, y)}\ \text{Sh} = \text{lg}\,\frac{1}{p(x, y)}\ \text{Hart} =$ $\ln \frac{1}{p(x, y)}\ \text{nat}$ where $p(x, y)$ is the joint probability of events x and y	
13-35 (*912*)	transinformation content fr *transinformation* (f)	$T(x, y)$	$T(x, y) = I(x) + I(y) - I(x, y)$ 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)	See ISO/IEC 2382-16, item 16.04.07.
13-36 (*913*)	mean transinformation content fr *transinformation* (f) *moyenne*	T	$T(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) T(x_i, y_j)$ for the sets $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_m\}$, 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)	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 ≈ 0,693 nat ≈ 0,301 Hart
13-24.b	hartley	Hart	value of the quantity when the argument is equal to 10	1 Hart ≈ 3,322 Sh ≈ 2,303 nat
13-24.c	natural unit of information	nat	value of the quantity when the argument is equal to e	1 nat ≈ 1,433 Sh ≈ 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:

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

If Y is binary:

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

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:

$$0\ \mathrm{Sh} \le I(Y;X) \le 1\ \mathrm{Sh}$$

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

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

If Y is binary:

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

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 $22 + 78/2 = \mathbf{61}$ correct prognoses (TP+TN)

If Y is binary:

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

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 $22 + 78/2 = \mathbf{61}$ correct prognoses (TP+TN)

Any prognostic algorithm
which does less than this
is performing poorly

If Y is binary:

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

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

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

In $N=100$ new prognoses:

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

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

Any prognostic algorithm which does less than this is performing poorly

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

If Y is binary:

$$0\ \mathrm{Sh} \le I(Y;X) \le 1\ \mathrm{Sh}$$

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

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

In $N = 100$ new prognoses:

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

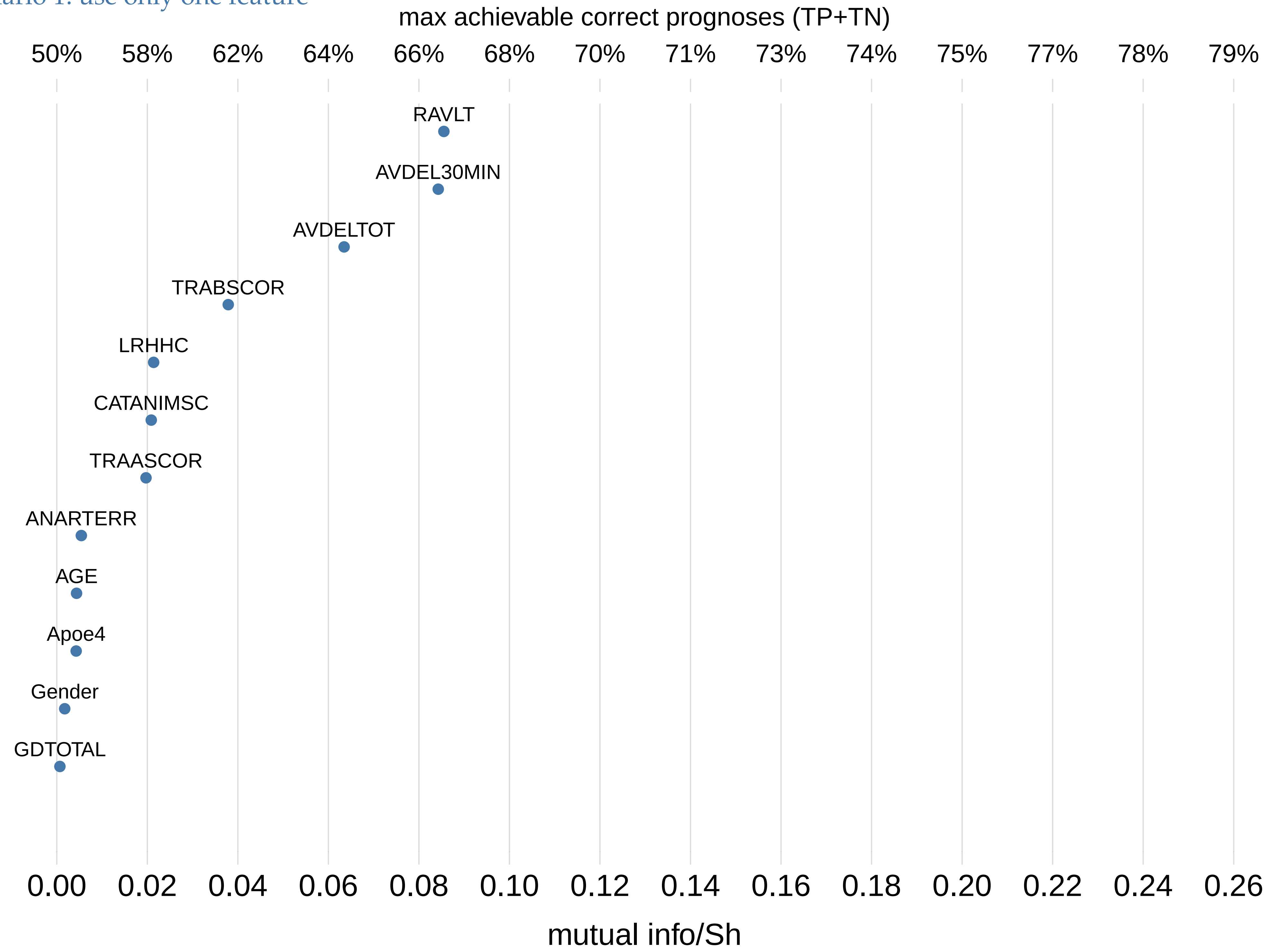
→ approx $22 + 78/2 = \mathbf{61}$ correct prognoses (TP+TN)

Any prognostic algorithm which does less than this is performing poorly

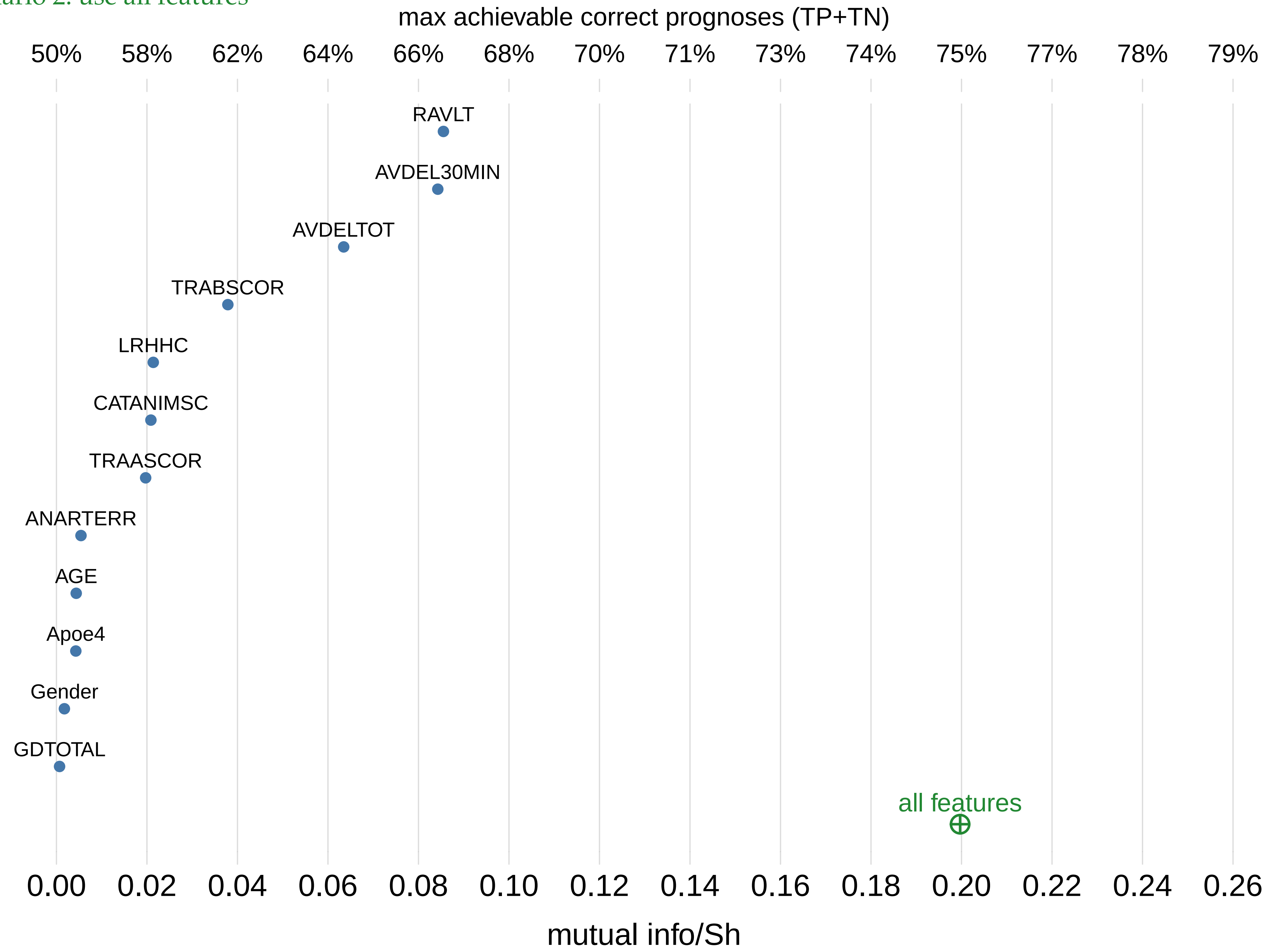
Upper bound ≈ $$\frac{1}{2} + \frac{1}{2}\sqrt{1 - (1 - 0.22)^{\frac{4}{3}}} \approx 77\,\% \quad \pm 0.8\sqrt{N}\,\%$$ correct prognoses

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

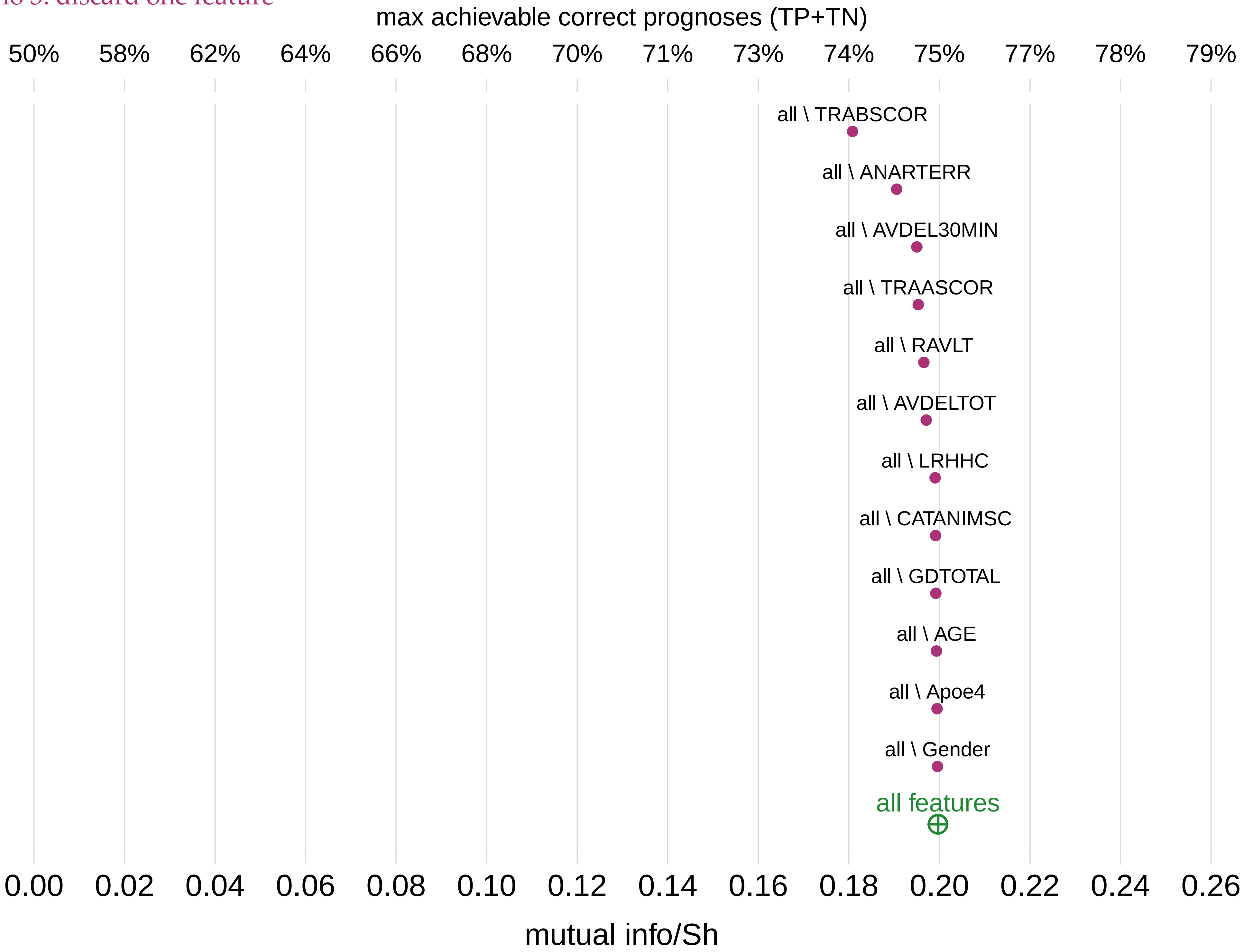
Scenario 1: use only one feature

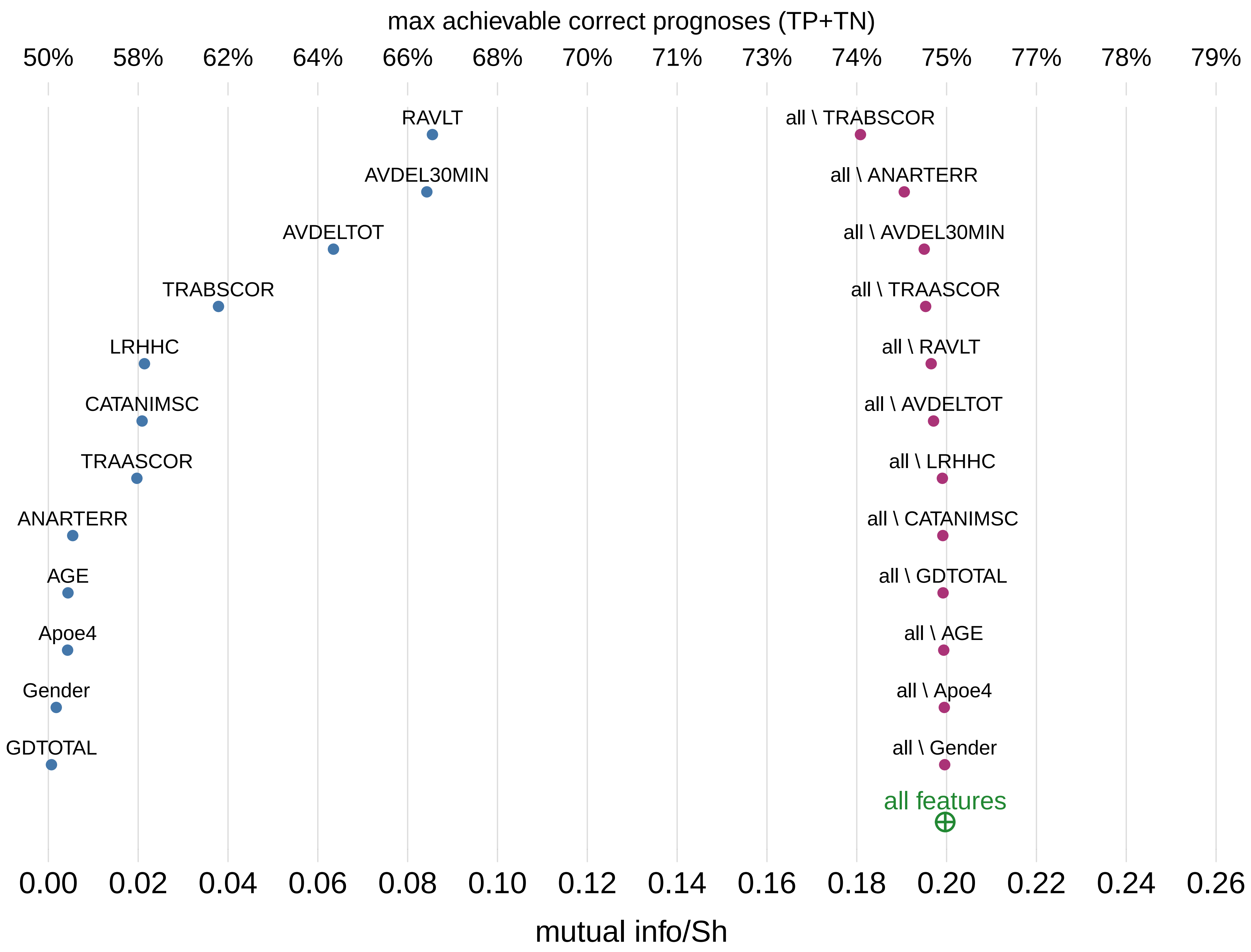


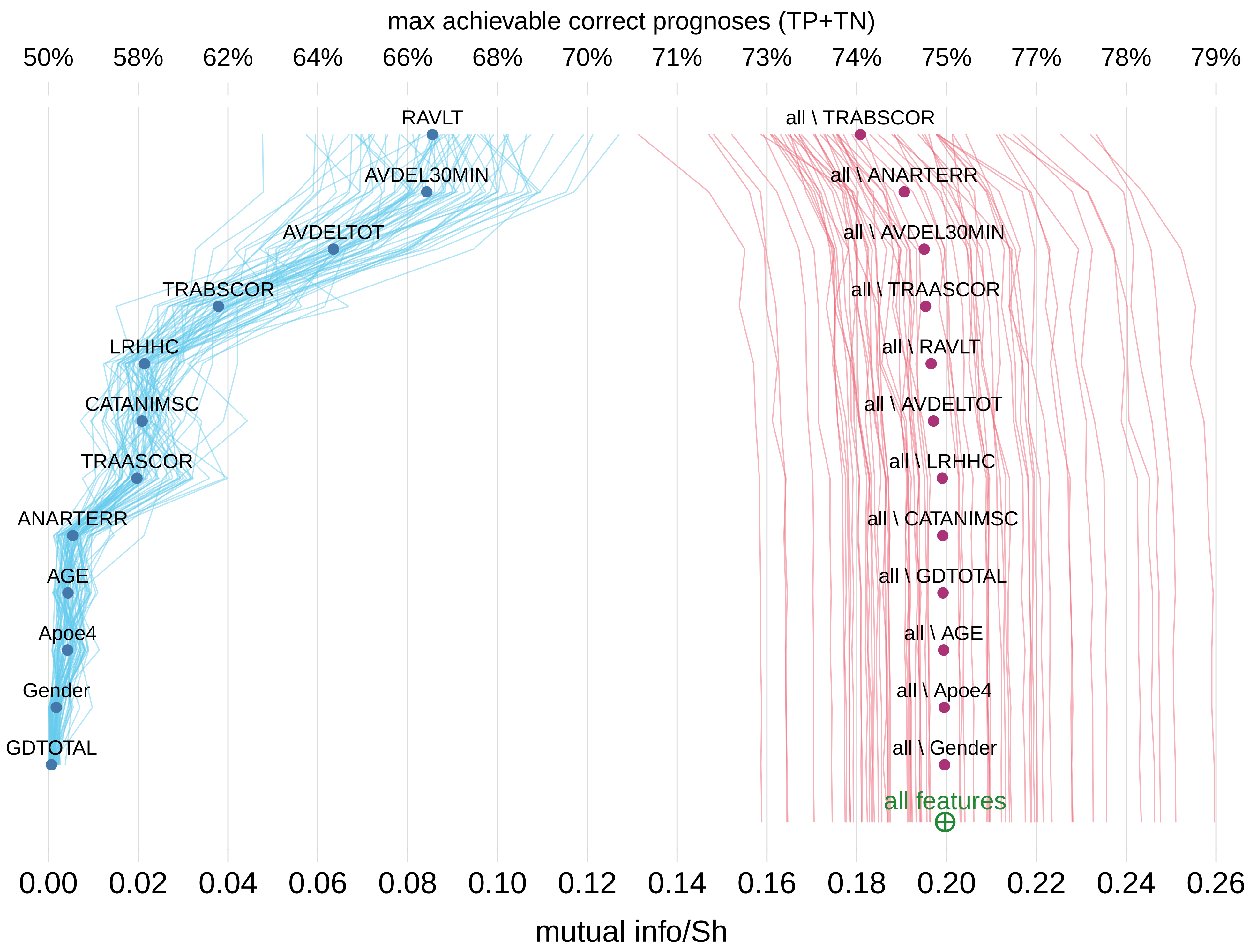
Scenario 2: use all features



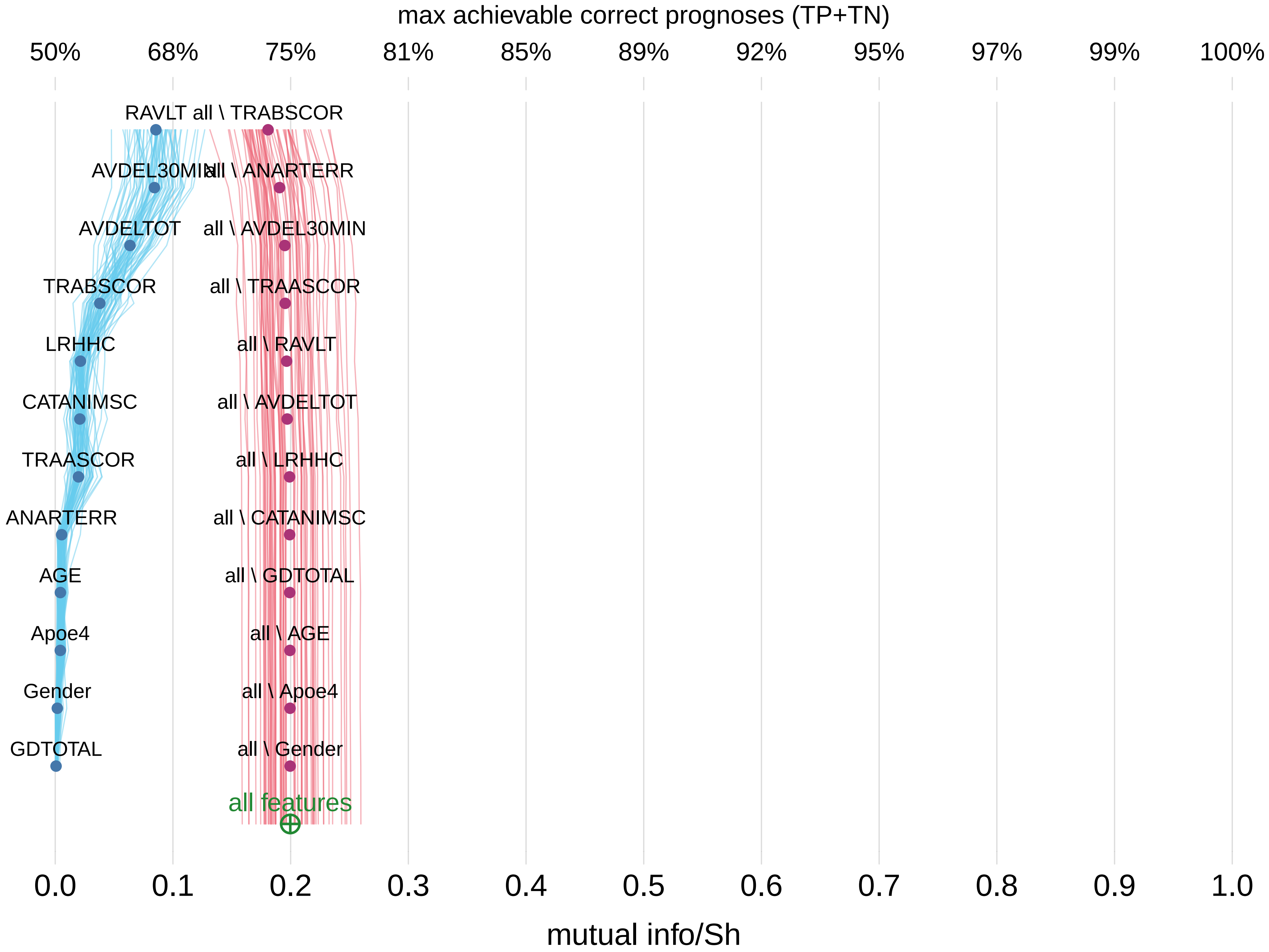
Scenario 3: discard one feature

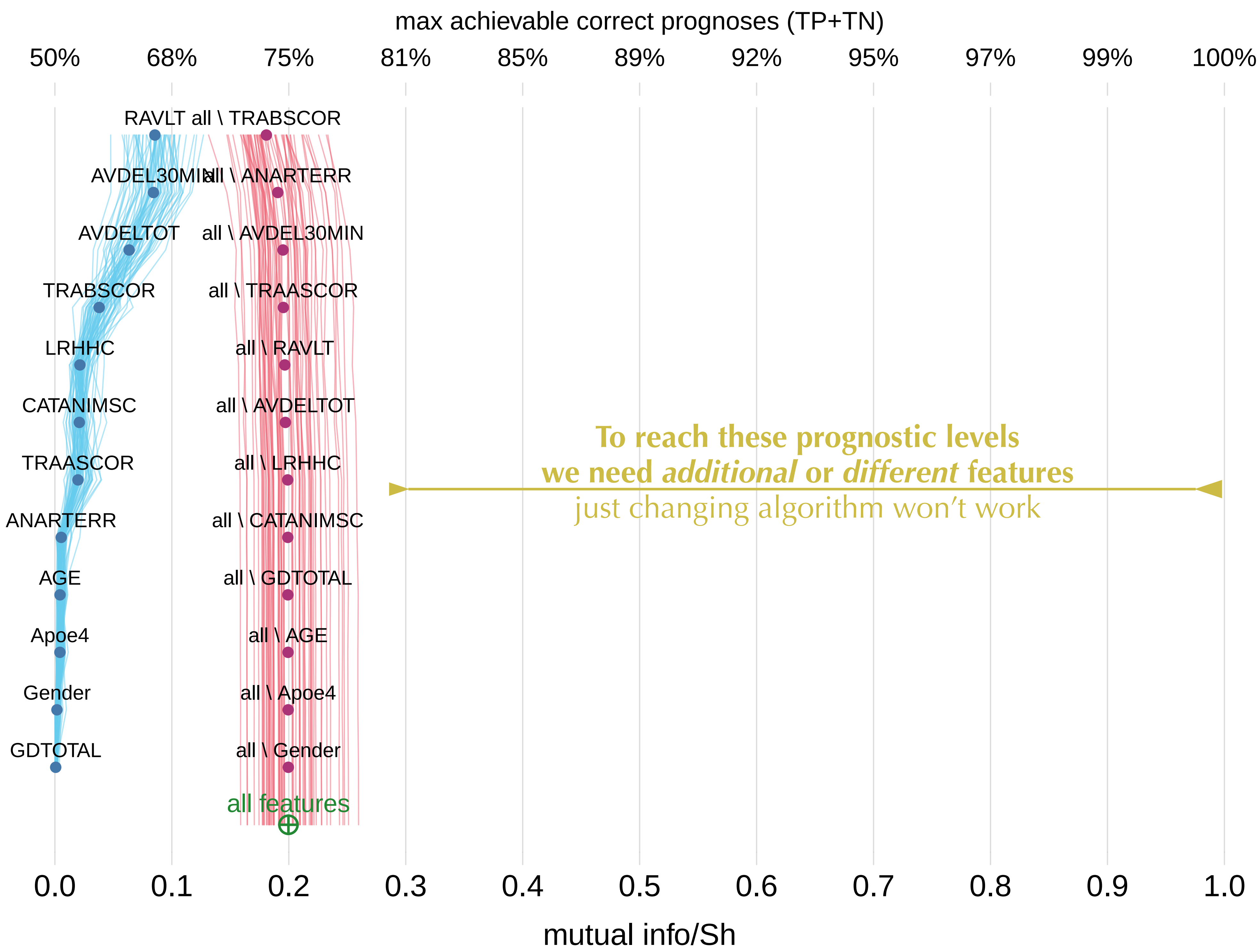



max achievable correct prognoses (TP+TN)
50%
58%
62%
64%
66%
68%
70%
71%
73%
74%
75%
77%
78%
79%
RAVLT
AVDEL30MIN
AVDELTOT
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
ANARTERR
AGE
Apoe4
Gender
GDTOTAL
all \ TRABSCOR
all \ ANARTERR
all \ AVDEL30MIN
all \ TRAASCOR
all \ RAVLT
all \ AVDELTOT
all \ LRHHC
all \ CATANIMSC
all \ GDTOTAL
all \ AGE
all \ Apoe4
all \ Gender
all features
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
0.26
mutual info/Sh



max achievable correct prognoses (TP+TN)
50%
58%
62%
64%
66%
68%
70%
71%
73%
74%
75%
77%
78%
79%
RAVLT
AVDEL30MIN
AVDELTOT
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
ANARTERR
AGE
Apoe4
Gender
GDTOTAL
all \ TRABSCOR
all \ ANARTERR
all \ AVDEL30MIN
all \ TRAASCOR
all \ RAVLT
all \ AVDELTOT
all \ LRHHC
all \ CATANIMSC
all \ GDTOTAL
all \ AGE
all \ Apoe4
all \ Gender
all features
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
0.26
mutual info/Sh


max achievable correct prognoses (TP+TN)
50%
68%
75%
81%
85%
89%
92%
95%
97%
99%
100%
RAVLT
AVDEL30MIN
AVDELTOT
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
ANARTERR
AGE
Apoe4
Gender
GDTOTAL
all \ TRABSCOR
all \ ANARTERR
all \ AVDEL30MIN
all \ TRAASCOR
all \ RAVLT
all \ AVDELTOT
all \ LRHHC
all \ CATANIMSC
all \ GDTOTAL
all \ AGE
all \ Apoe4
all \ Gender
all features
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
mutual info/Sh

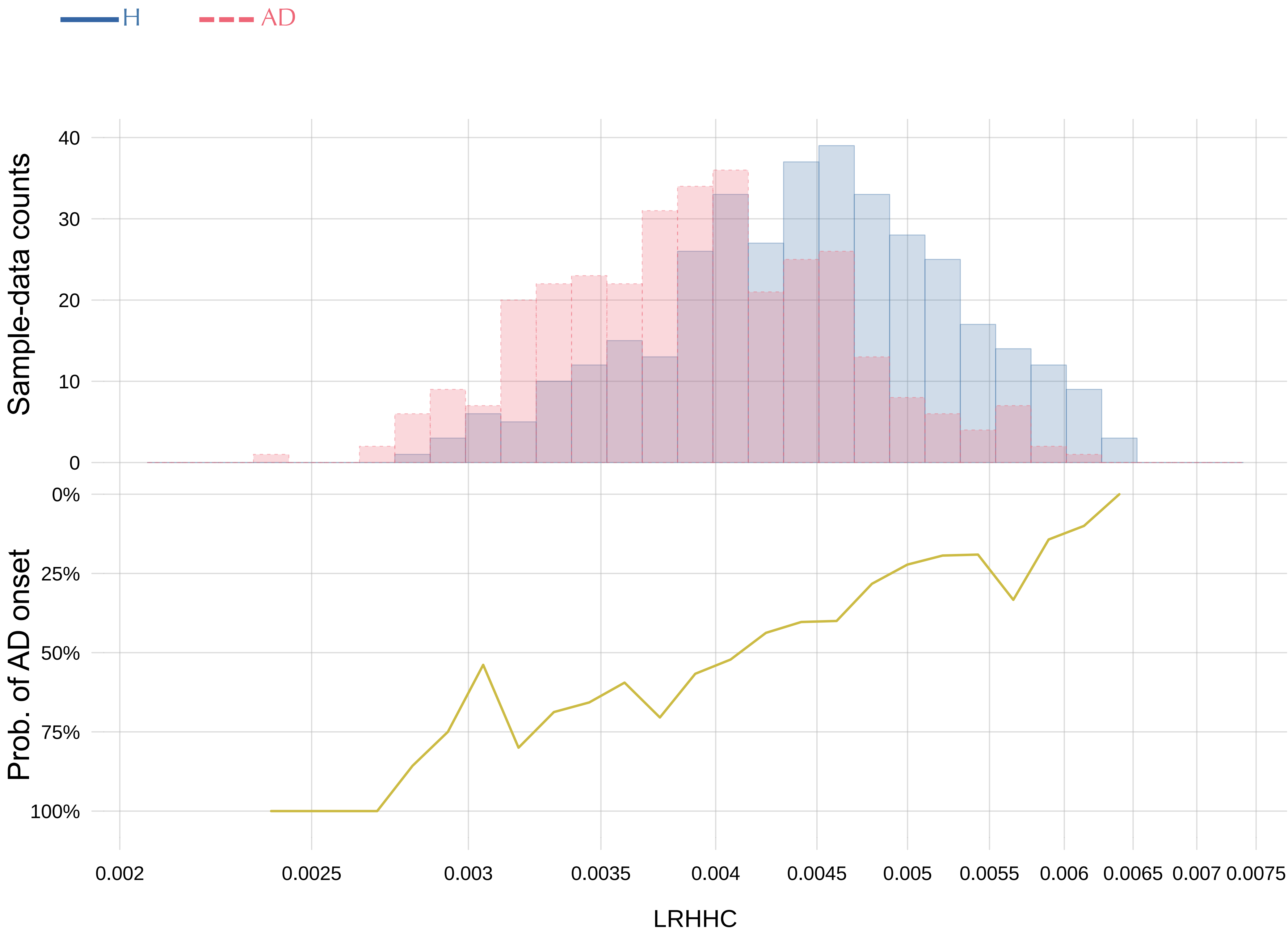


max achievable correct prognoses (TP+TN)
50%
68%
75%
81%
85%
89%
92%
95%
97%
99%
100%
RAVLT
all \ TRABSCOR
AVDEL30MIN
all \ ANARTERR
AVDELTOT
all \ AVDEL30MIN
TRABSCOR
all \ TRAASCOR
LRHHC
all \ RAVLT
CATANIMSC
all \ AVDELTOT
TRAASCOR
all \ LRHHC
ANARTERR
all \ CATANIMSC
AGE
all \ GDTOTAL
Apoe4
all \ AGE
Gender
all \ Apoe4
GDTOTAL
all \ Gender
all features
To reach these prognostic levels
we need *additional* or *different* features
just changing algorithm won't work
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
mutual info/Sh

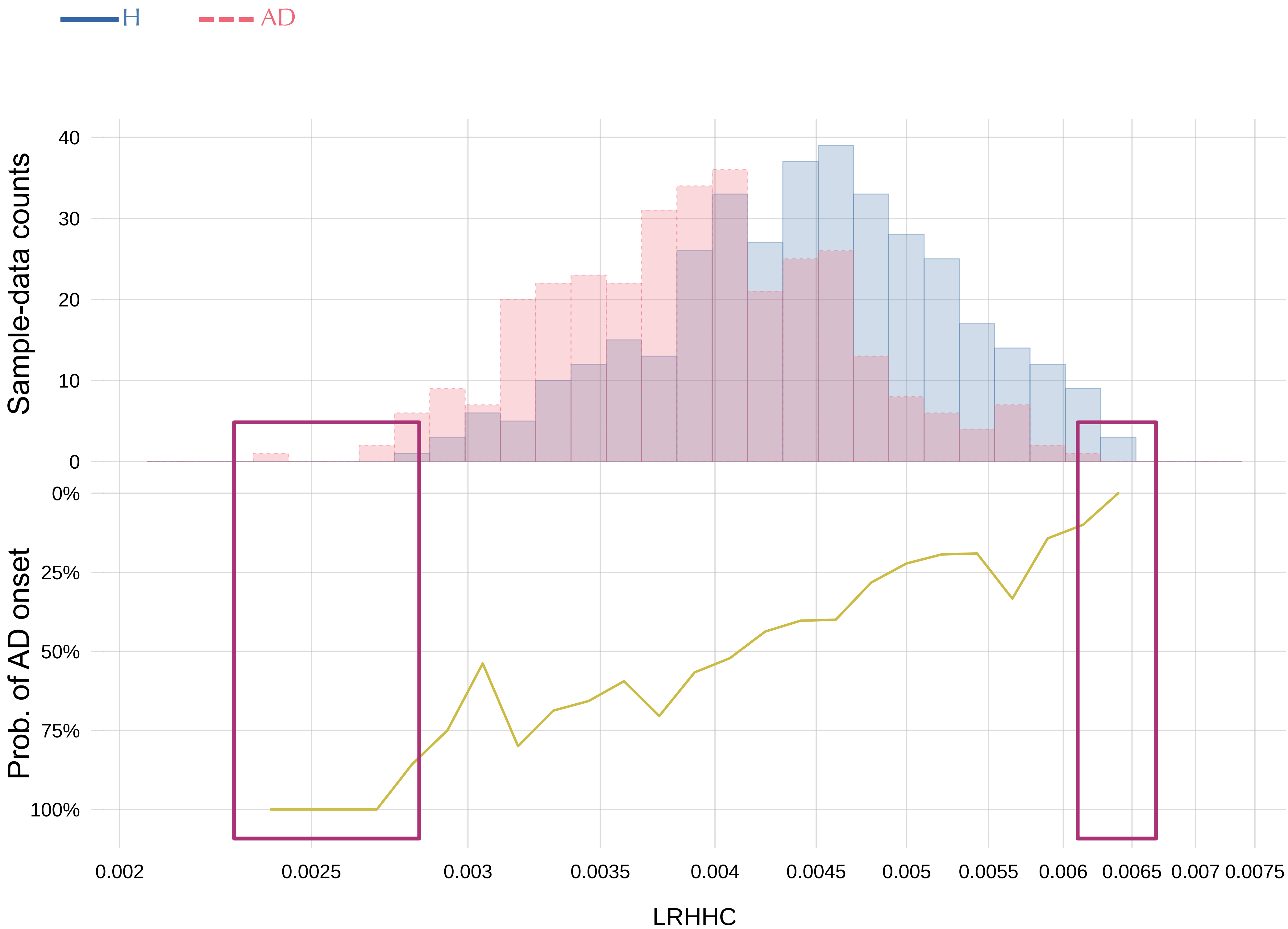
$$\mathrm{P}(Y \mid X)\,\mathrm{P}(X) \equiv \mathrm{P}(X \mid Y)\,\mathrm{P}(Y)$$

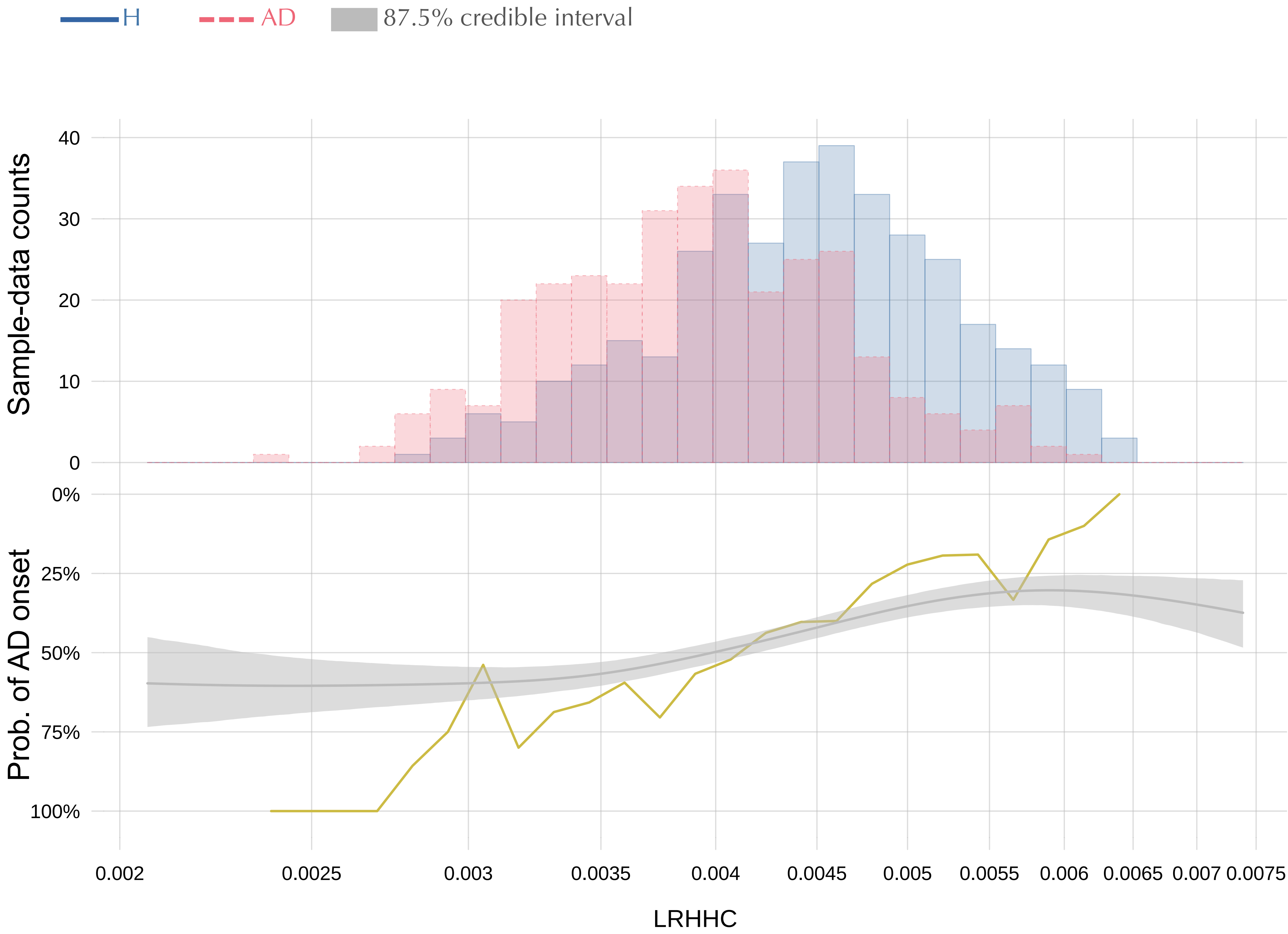
$$\mathrm{P}(Y \mid X)\,\mathrm{P}(X) \overset{!}{\equiv} \mathrm{P}(X \mid Y)\,\mathrm{P}^*(Y)$$

$$\overset{?}{\mathrm{P}}{}^*(Y \mid X)\ \overset{?}{\mathrm{P}}{}^*(X) \equiv \overset{?}{\mathrm{P}}{}^*(X \mid Y)\ \mathrm{P}^*(Y)$$

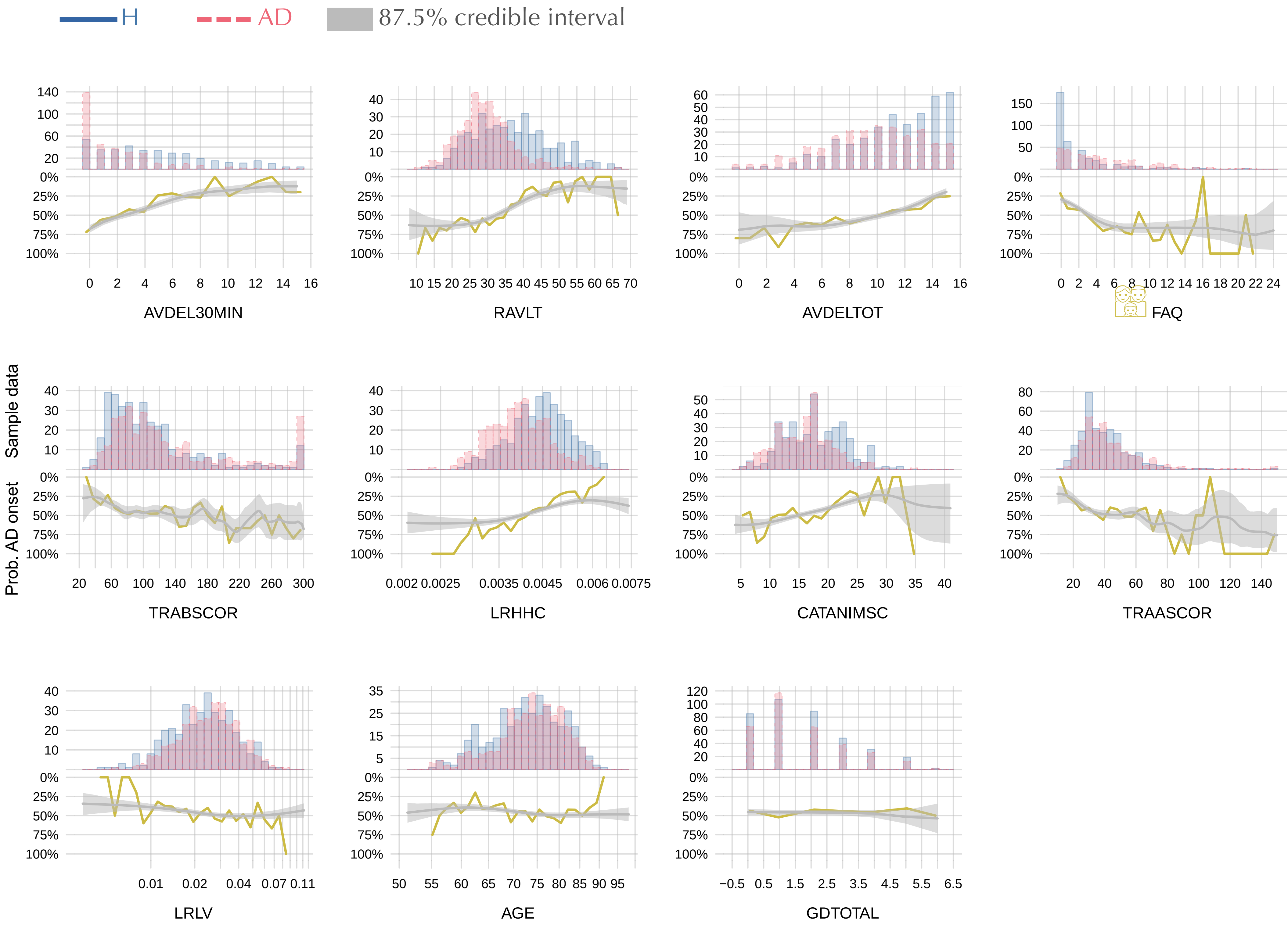


H
AD
Sample-data counts
40
30
20
10
0
0%
25%
50%
75%
100%
Prob. of AD onset
0.002
0.0025
0.003
0.0035
0.004
0.0045
0.005
0.0055
0.006
0.0065
0.007
0.0075
LRHHC


H
AD
Sample-data counts
40
30
20
10
0
0%
25%
50%
75%
100%
Prob. of AD onset
0.002
0.0025
0.003
0.0035
0.004
0.0045
0.005
0.0055
0.006
0.0065
0.007
0.0075
LRHHC


H
AD
87.5% credible interval
Sample-data counts
40
30
20
10
0
0%
25%
50%
75%
100%
Prob. of AD onset
0.002
0.0025
0.003
0.0035
0.004
0.0045
0.005
0.0055
0.006
0.0065
0.007
0.0075
LRHHC


H
AD
Sample data
Prob. AD onset
AVDEL30MIN
RAVLT
AVDELTOT
FAQ
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
LRLV
AGE
GDTOTAL


H
AD
87.5% credible interval
Sample data
Prob. AD onset
AVDEL30MIN
RAVLT
AVDELTOT
FAQ
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
LRLV
AGE
GDTOTAL

Thank you!