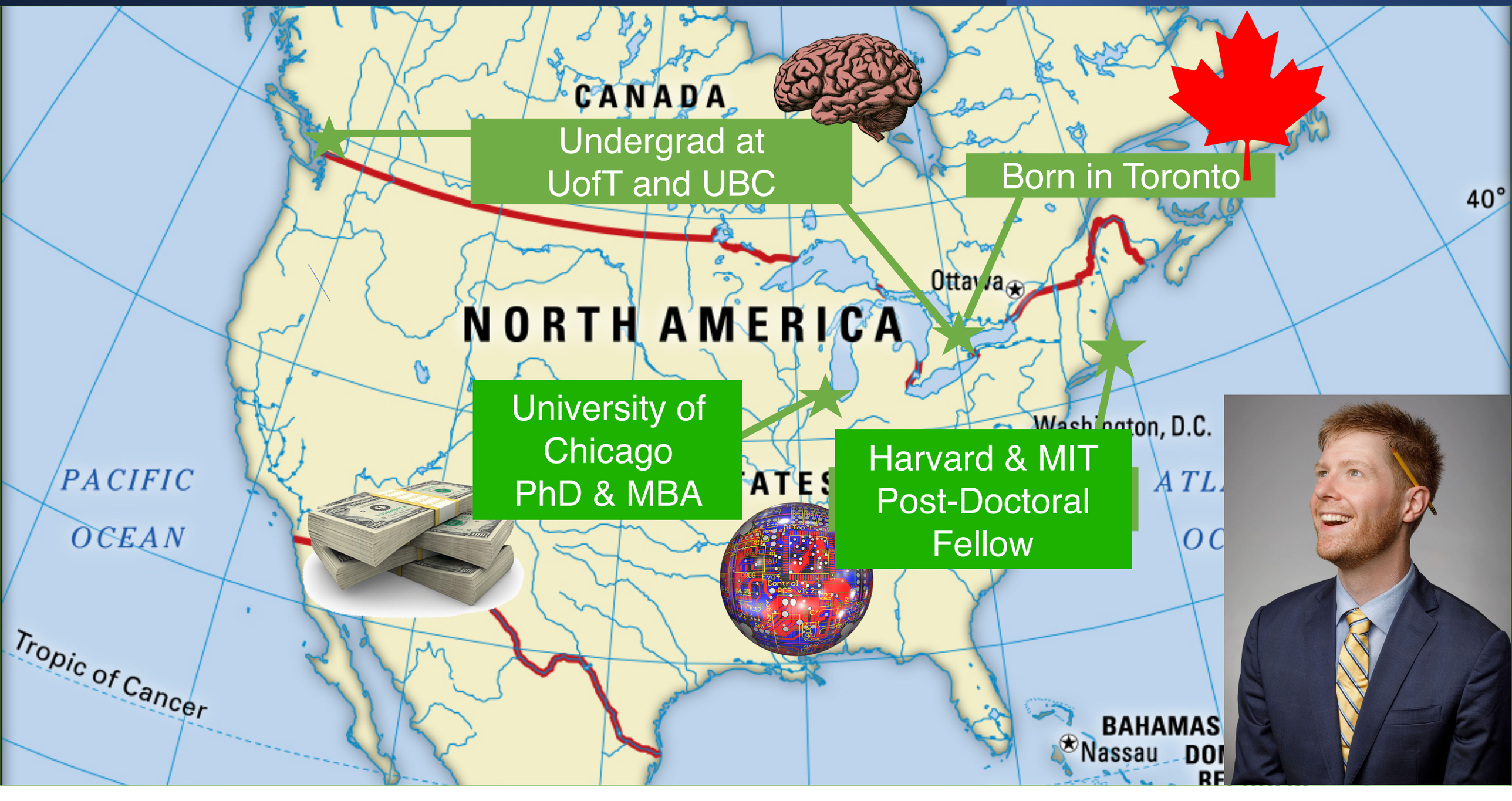
Text Analysis for Social Scientists and Leaders

Class 1: Introduction & Humans

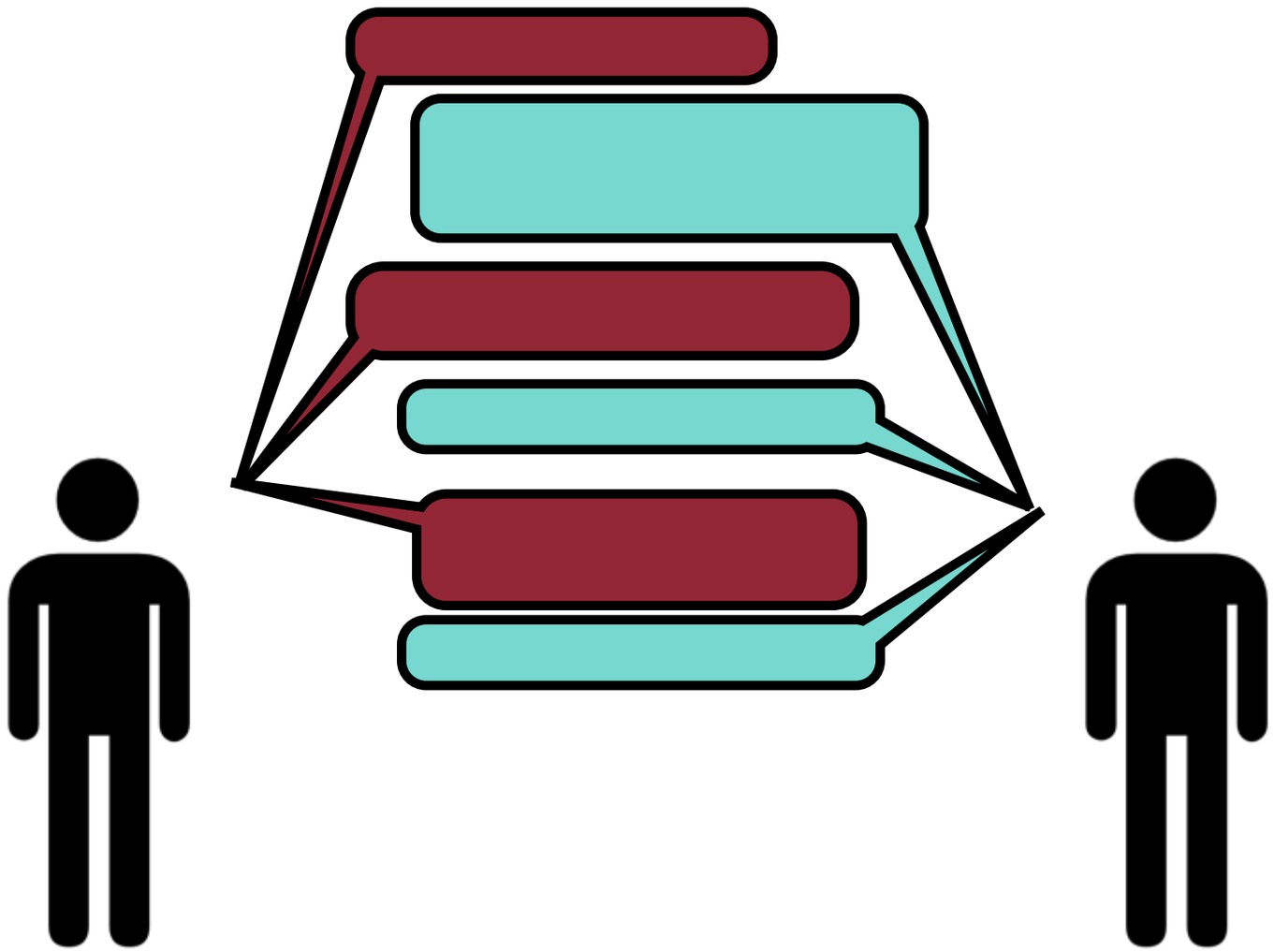
Prof. Michael Yeomans

Who am I?

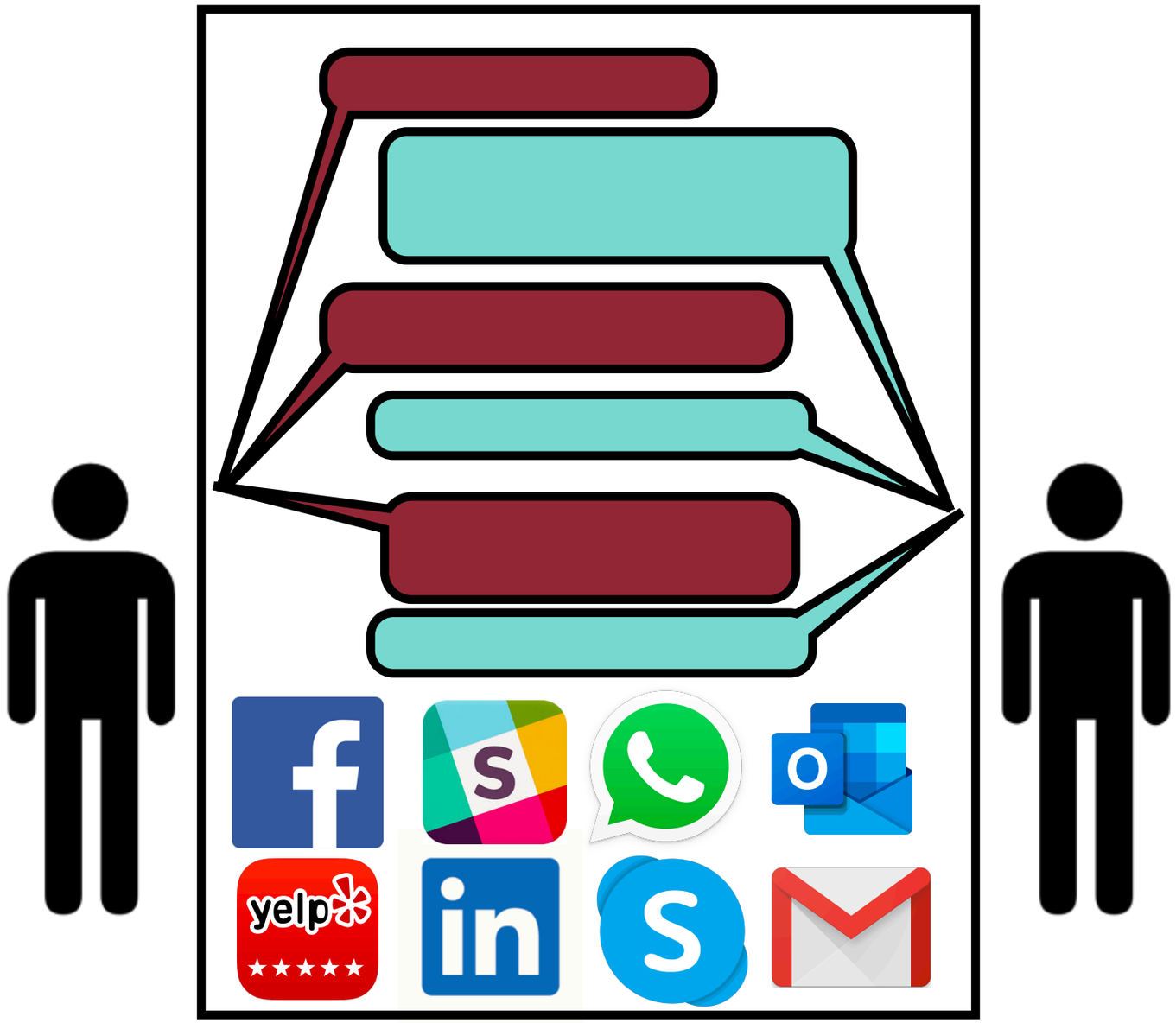

CANADA
Undergrad at
UofT and UBC
Born in Toronto
40°
Ottawa
NORTH AMERICA
University of
Chicago
PhD & MBA
Harvard & MIT
Post-Doctoral
Fellow
Washington, D.C.
PACIFIC
OCEAN
Tropic of Cancer
BAHAMAS
Nassau

Conversations are Everywhere

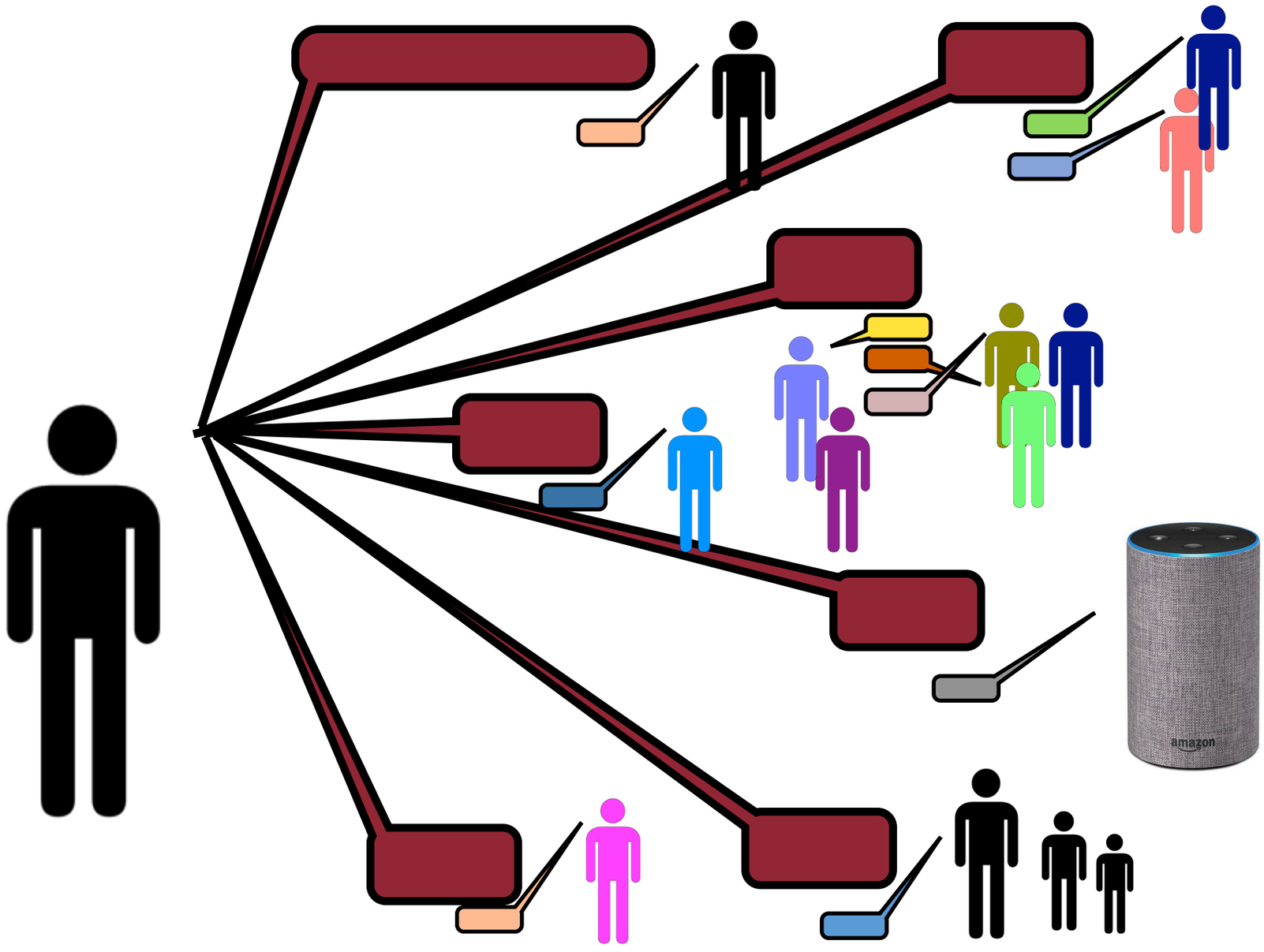
New Modes of Conversation



New Modes of Conversation



New Kinds of Conversation



Your Marks

Your Marks

None!

Some Broad Questions

Why R?

Some Broad Questions

Why R?

You know it

Some Broad Questions

Why R?

You know it

I know it

Some Broad Questions

Why R?

You know it

I know it

Why English?

Some Broad Questions

Why R?

You know it

I know it

Why English?

You know it

Some Broad Questions

Why R?

You know it

I know it

Why English?

You know it

I know it

Some Broad Questions

Why R?

You know it

I know it

Why English?

You know it

I know it

Je ne connais pas assez de français

Some Broad Questions

Why R?

You know it

I know it

Why English?

You know it

I know it

Je ne connais pas assez de français

Why Text?

Ubiquitous

Some Broad Questions

Why R?

You know it

I know it

Why English?

You know it

I know it

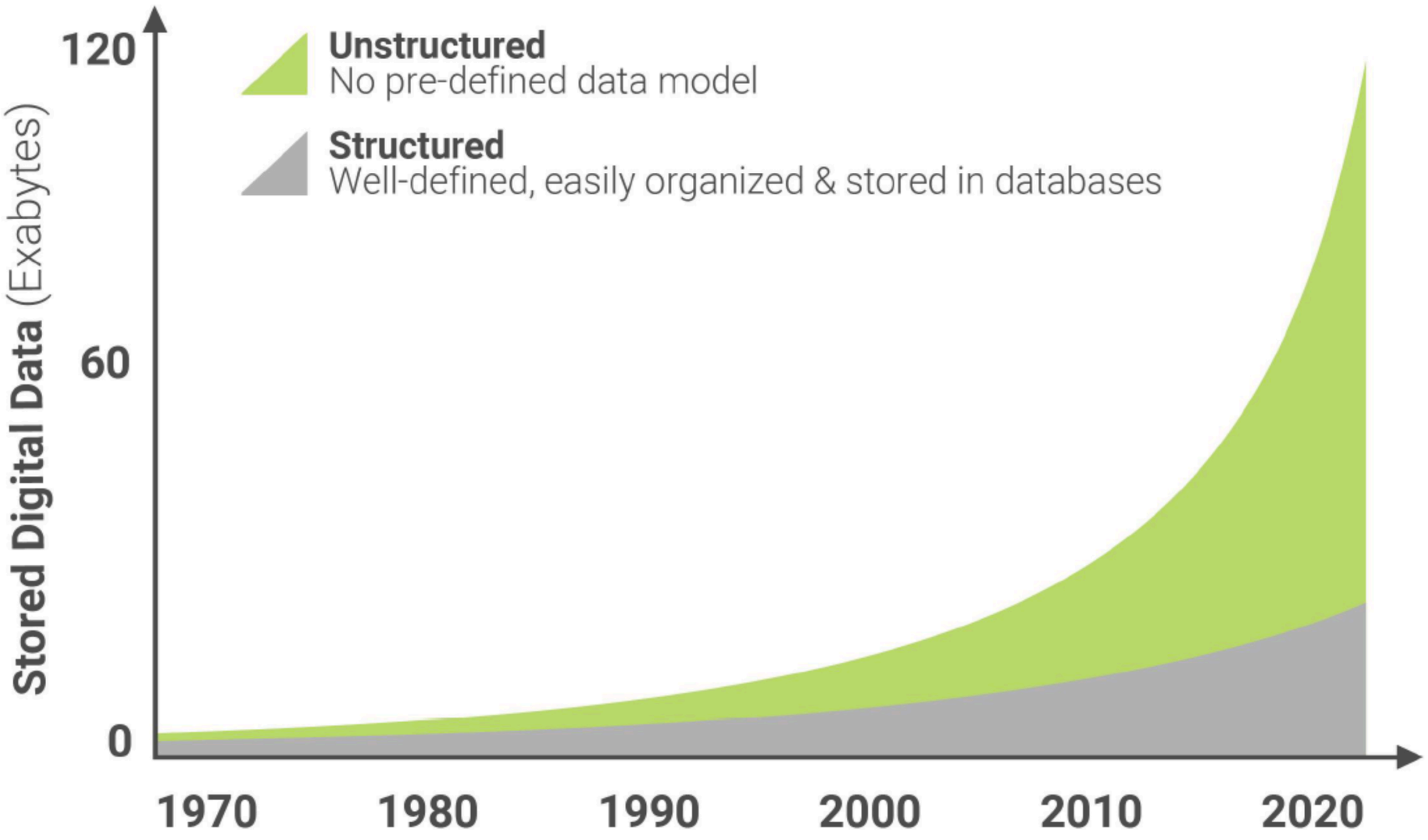
Je ne connais pas assez de français

Why Text?

Ubiquitous

Increasingly easy to measure

Unstructured Opportunities



One Question for this Class

How do we turn words into numbers?

One Question for this Class

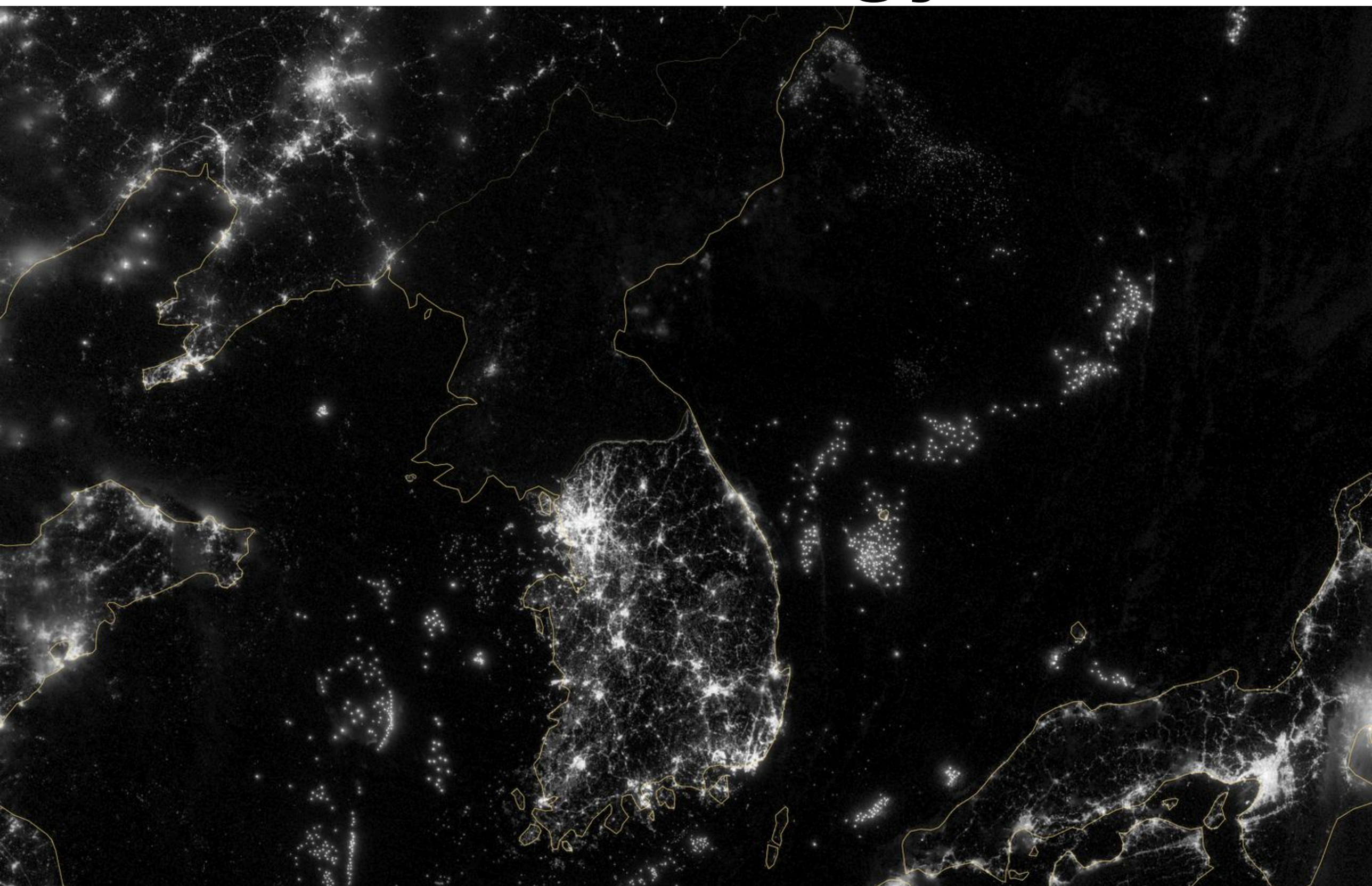
How do we turn words into numbers?

Measurement Validity

How well does a number represent the behaviour itself?

(Cronbach & Meehl, 1955; John & Benet- Martinez, 2000;
Flake, Pek & Hehman, 2017; Fried & Flake, 2018;
Mullainathan & Speiss, 2017)

An analogy

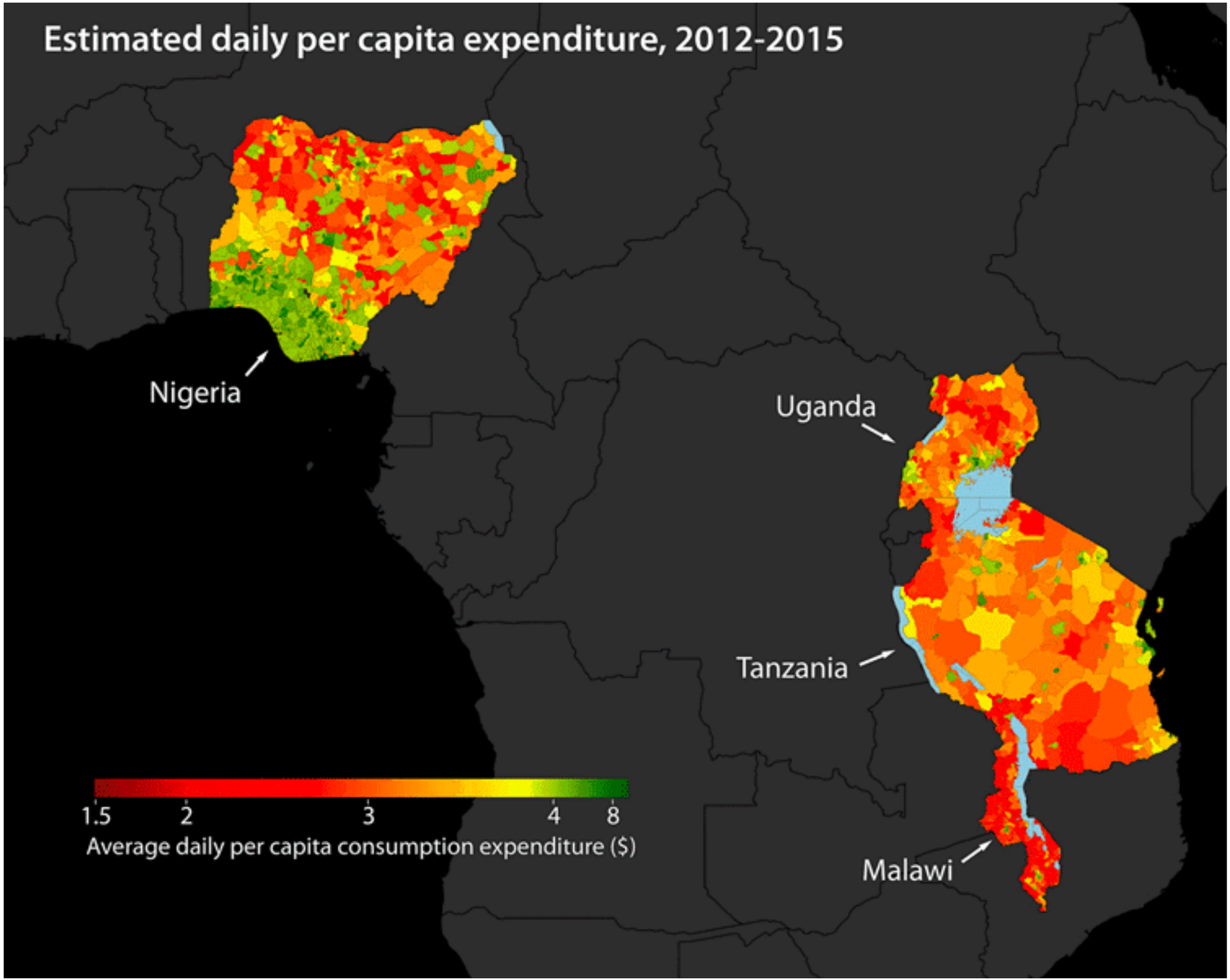


An analogy

Jean et al., 2017

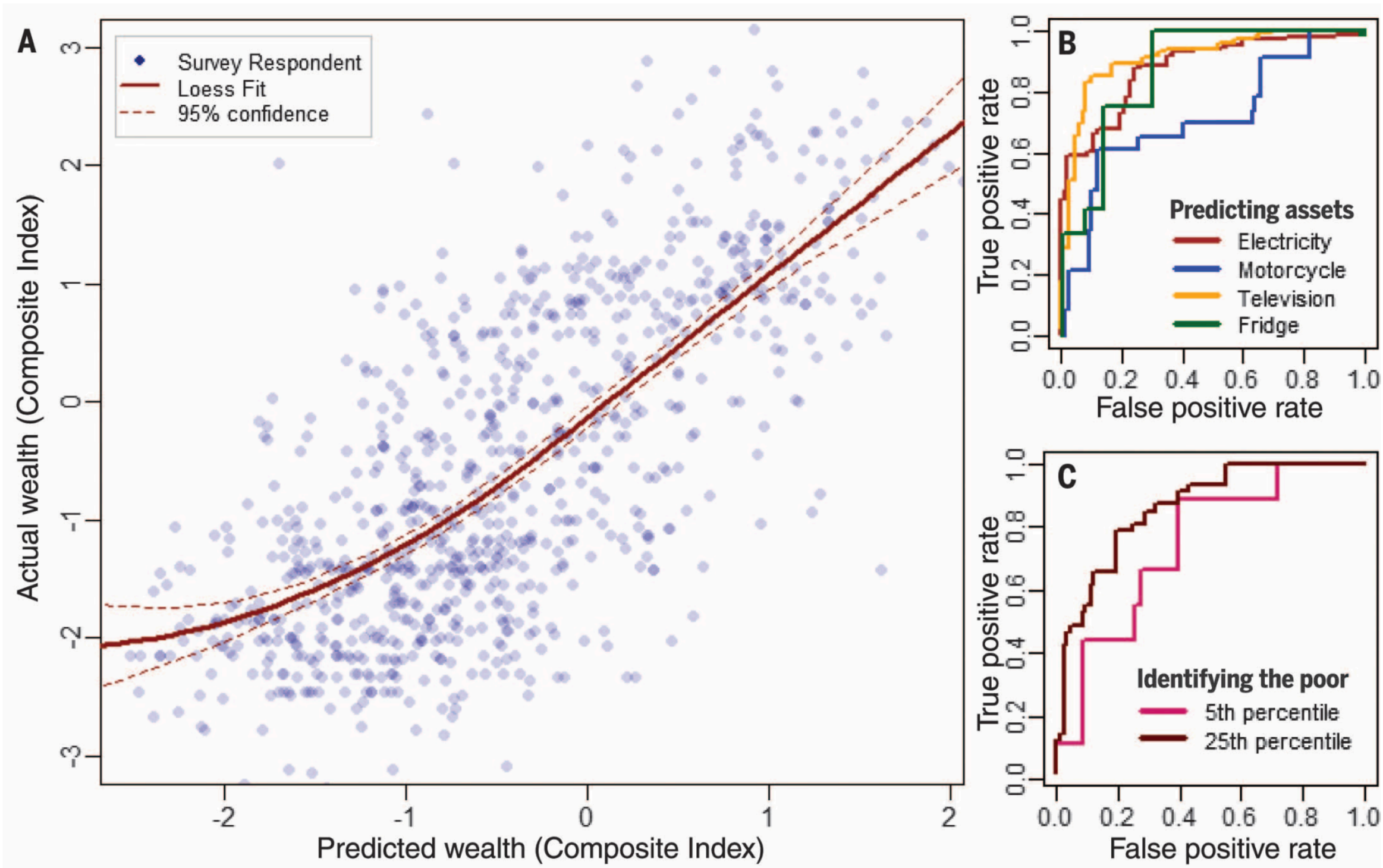
An analogy

Jean et al., 2017



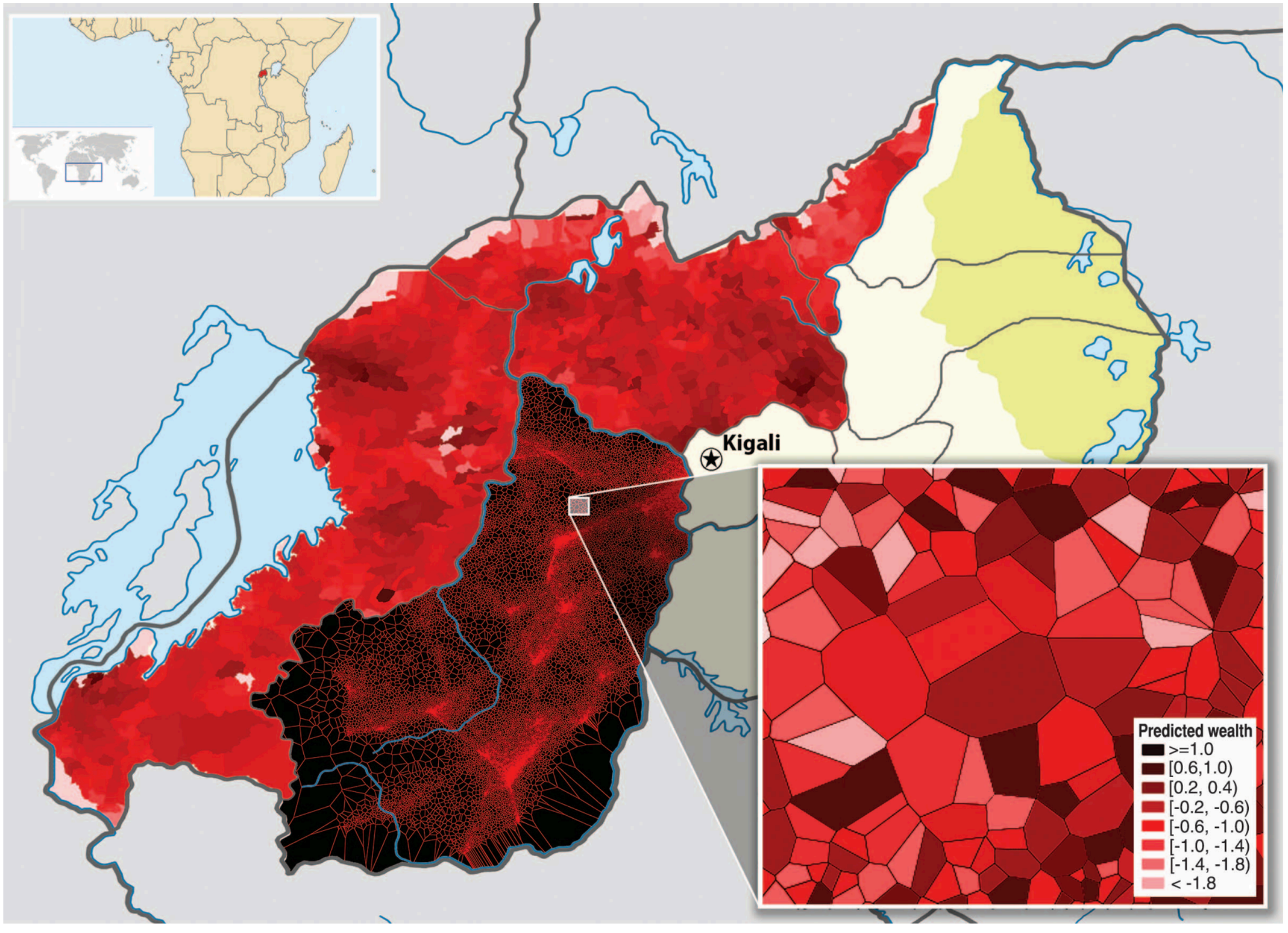
An analogy

Blumenstock et al., 2015



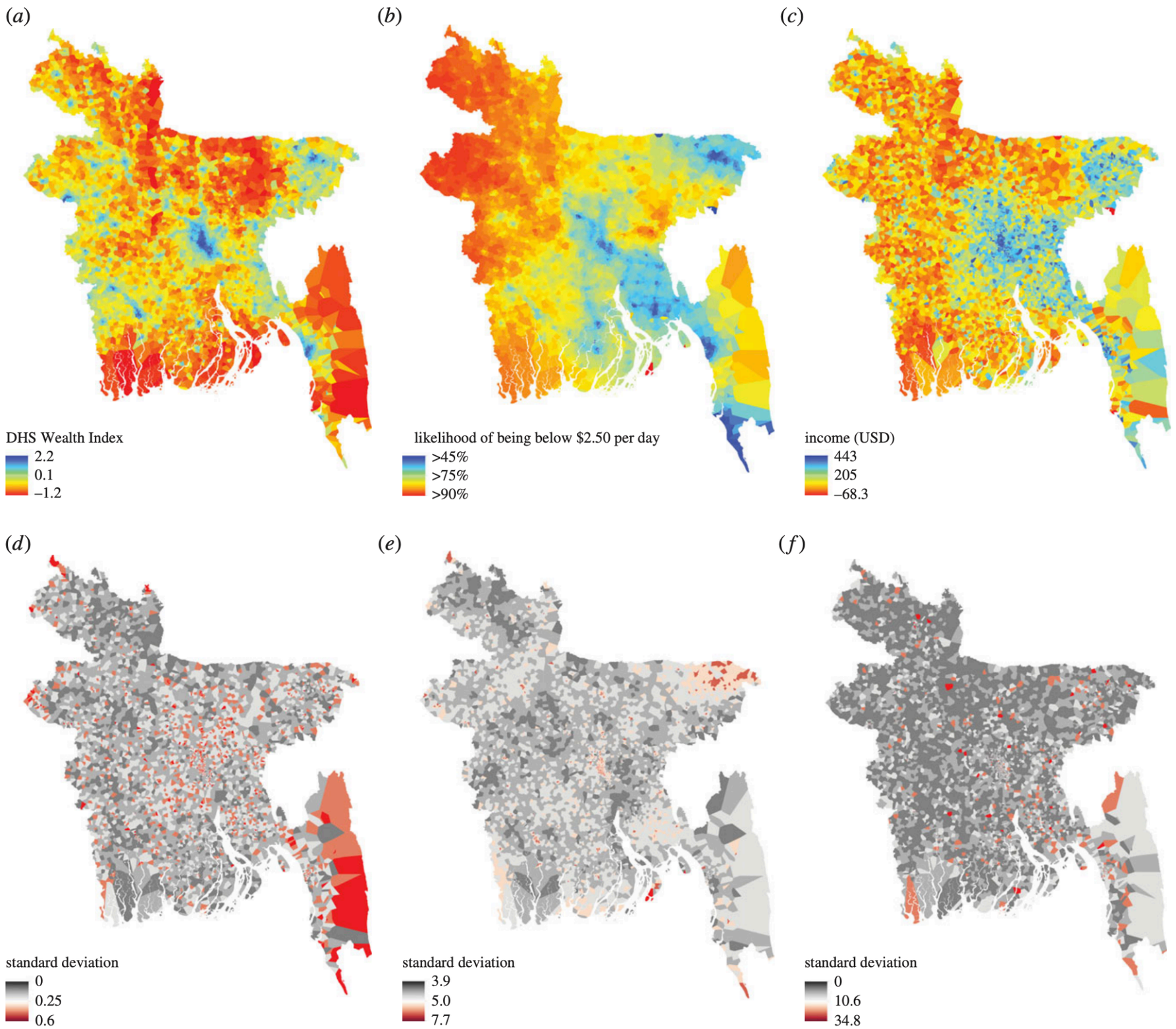
An analogy

Blumenstock et al., 2015



An analogy

Steele et al., 2017



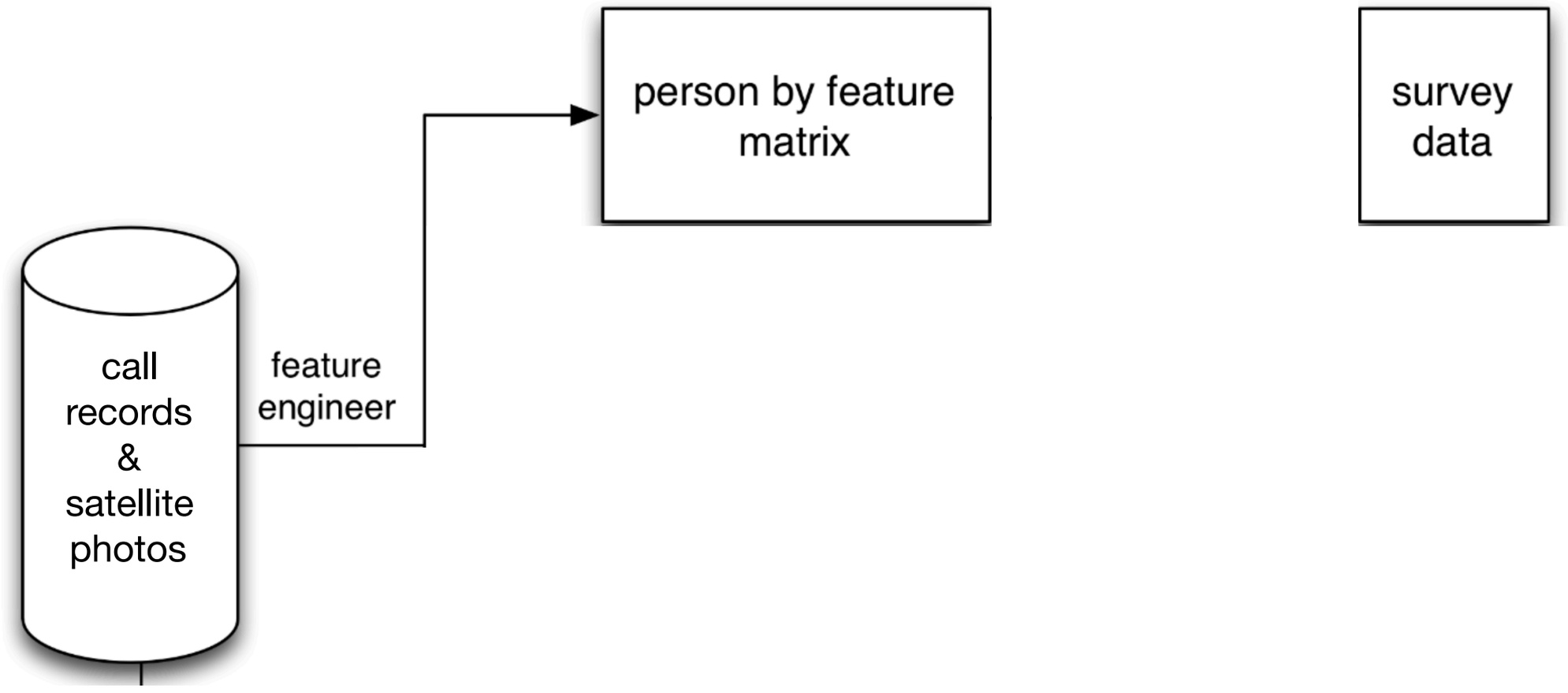
An analogy

An analogy

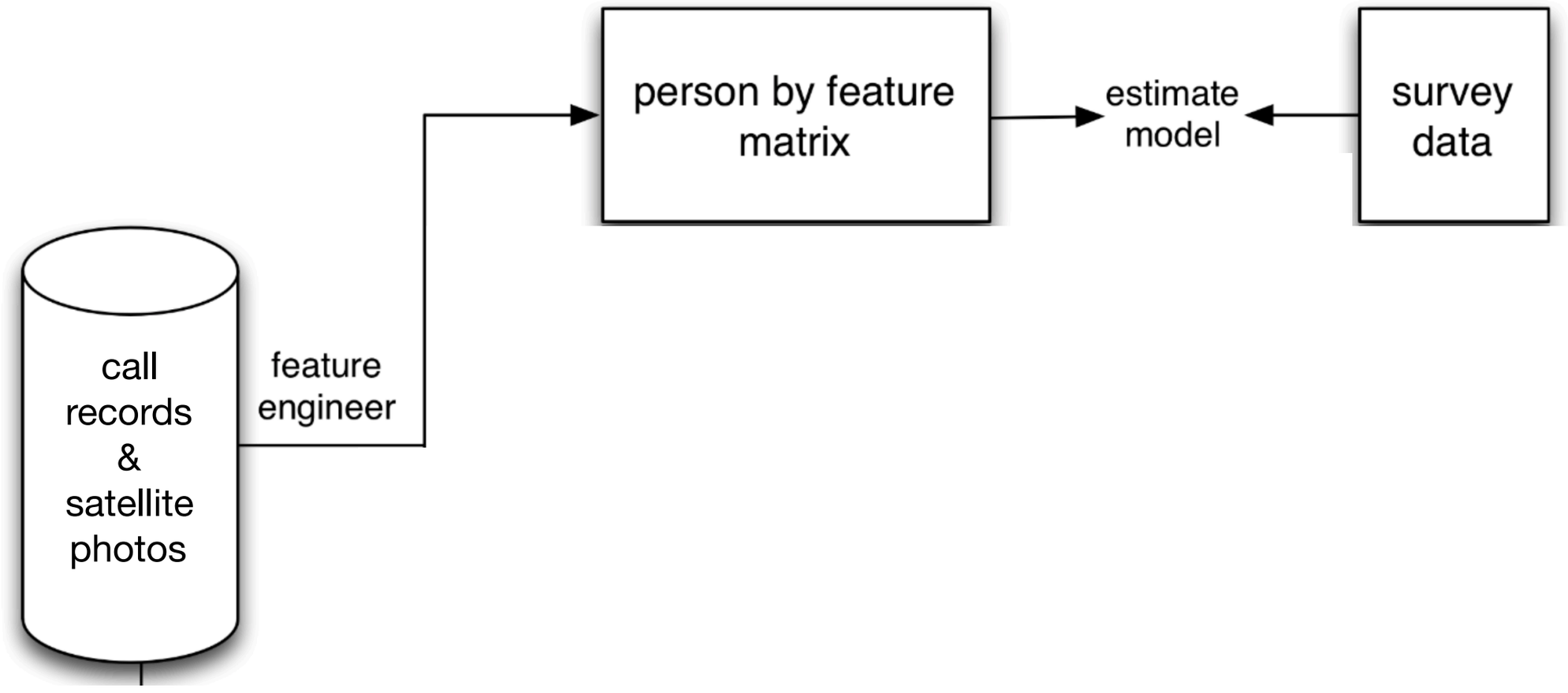
survey data

call records & satellite photos

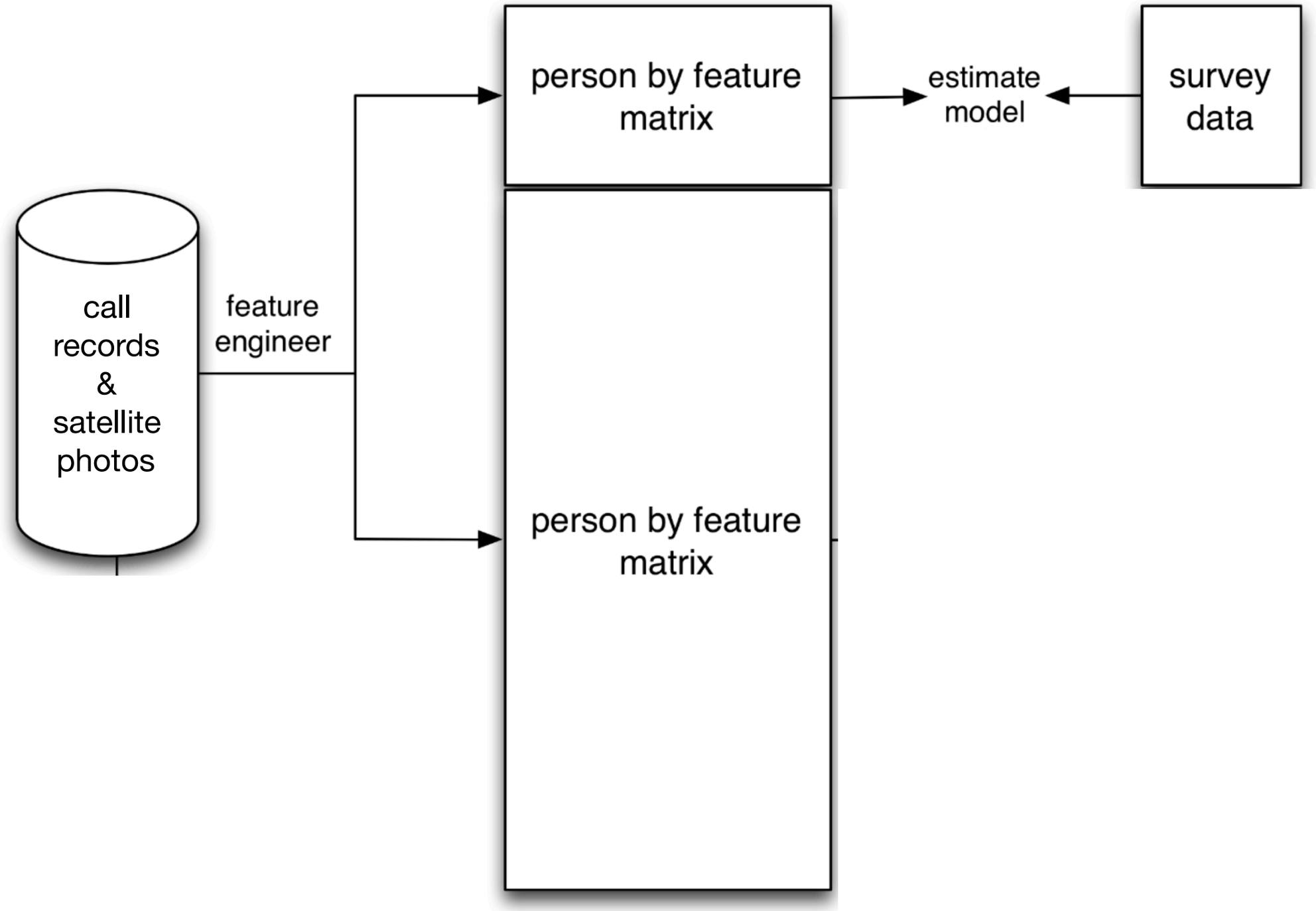
An analogy


person by feature matrix
survey data
call records & satellite photos
feature engineer

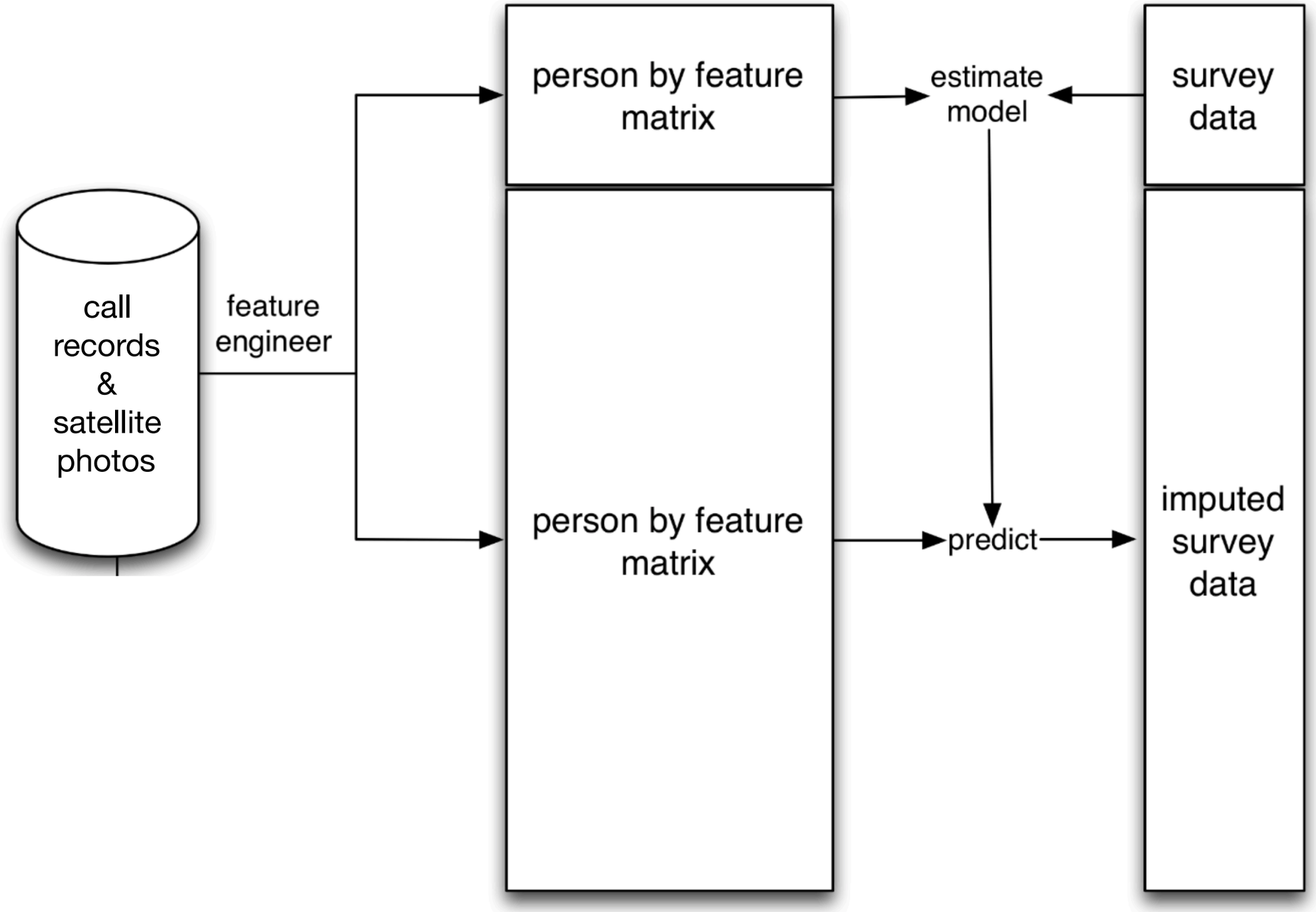
An analogy


person by feature matrix
estimate model
survey data
call records & satellite photos
feature engineer

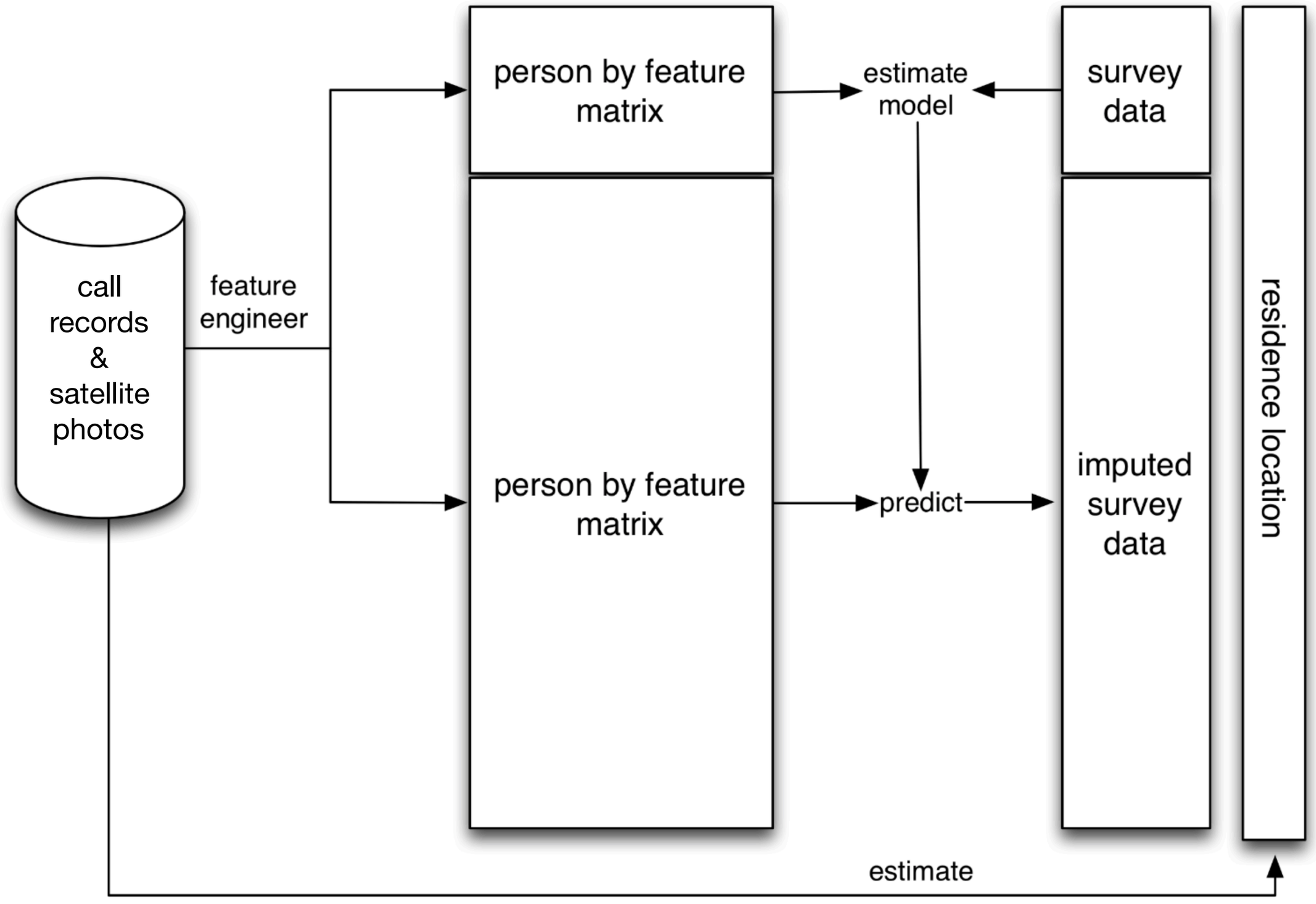
An analogy


person by feature matrix
estimate model
survey data
call records & satellite photos
feature engineer
person by feature matrix

An analogy


call records & satellite photos
feature engineer
person by feature matrix
estimate model
survey data
person by feature matrix
predict
imputed survey data

An analogy


call records & satellite photos
feature engineer
person by feature matrix
person by feature matrix
estimate model
predict
survey data
imputed survey data
residence location
estimate

Applied Measurement Validity

A model of the world

$$y = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Mullainathan & Speiss, 2017

Applied Measurement Validity

A model of the world

$$y = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

$$\hat{\beta}$$

Intervention

Mullainathan & Speiss, 2017

Applied Measurement Validity

A model of the world

$$y = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

$$\hat{\beta}$$

Intervention

If I change x, what will y do?

Mullainathan & Speiss, 2017

Applied Measurement Validity

A model of the world

$$y = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

$\hat{y}$

Prediction

What y should I expect?

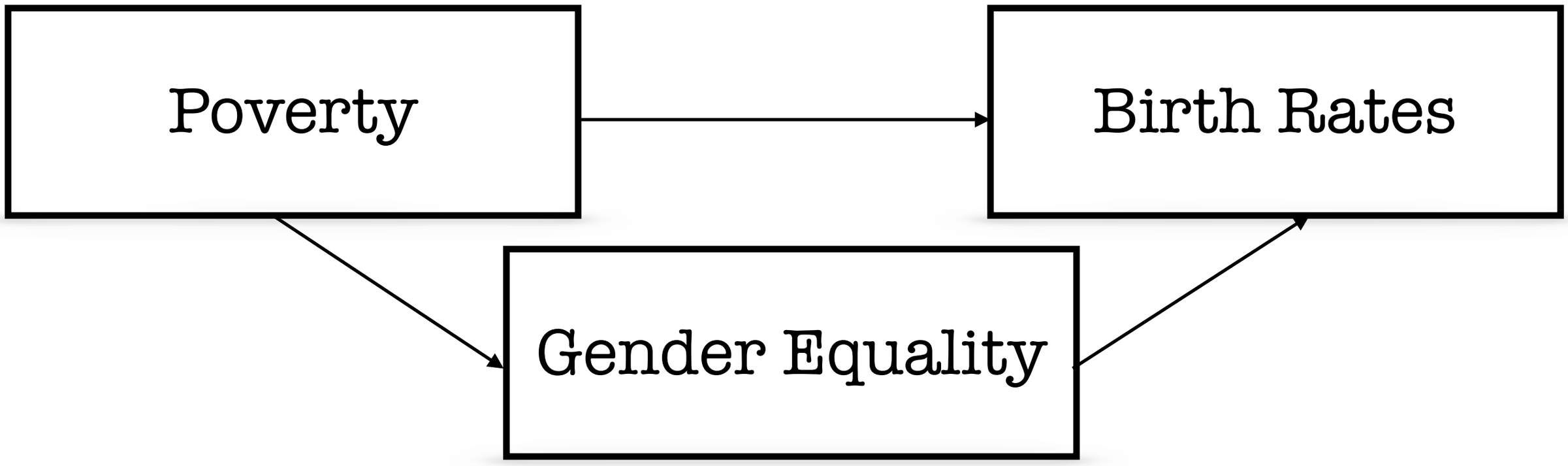
$\hat{\beta}$

Intervention

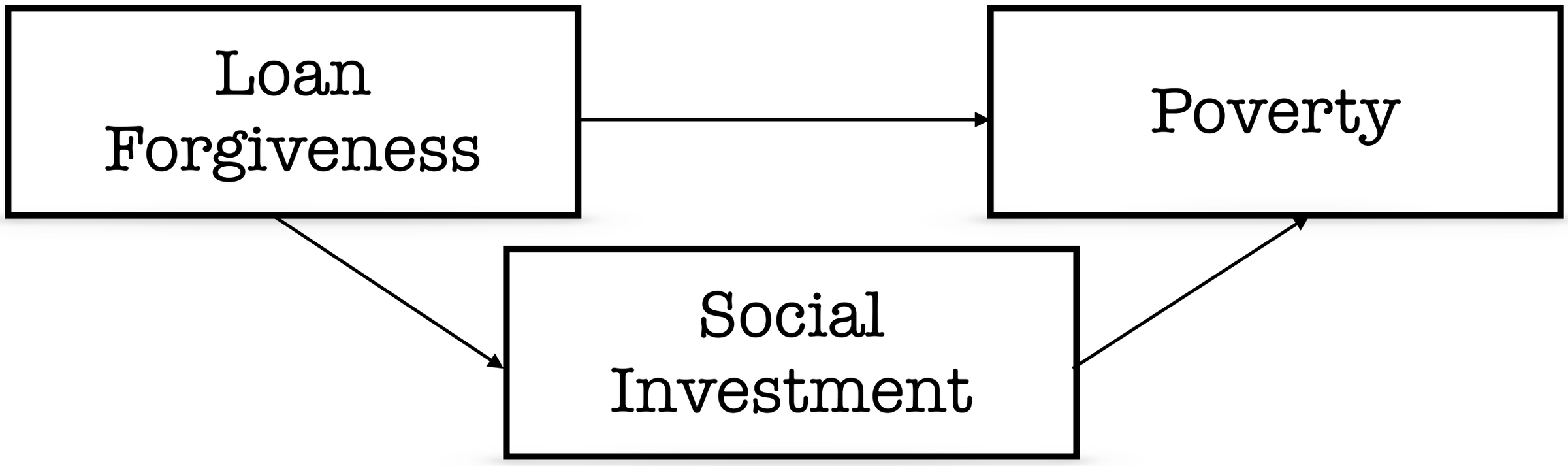
If I change x, what will y do?

Mullainathan & Speiss, 2017

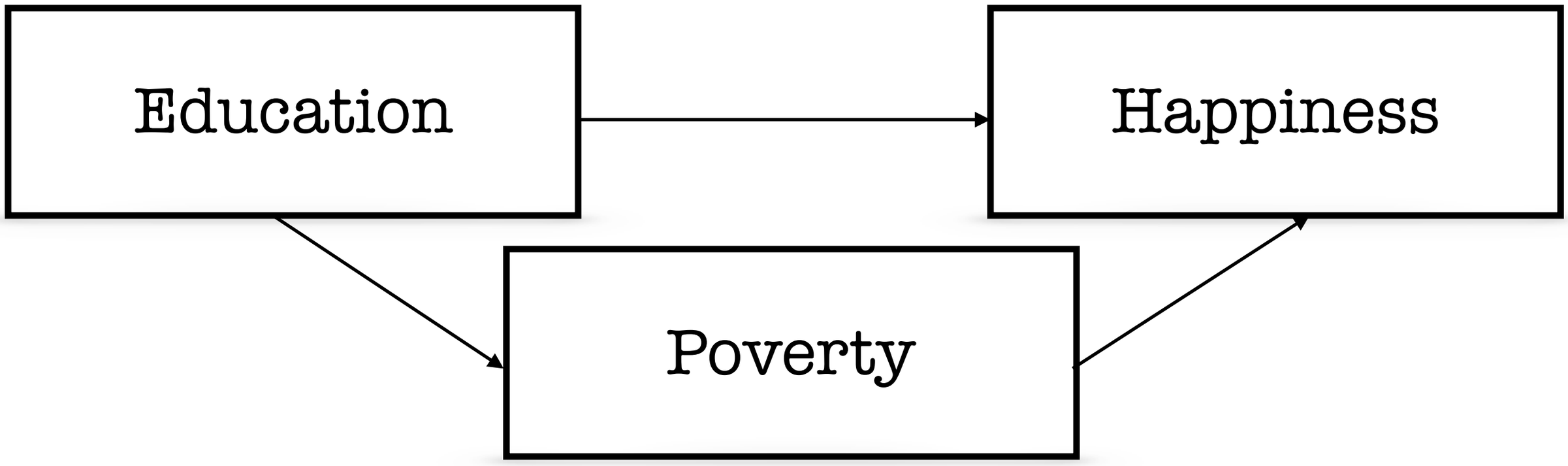
Applied Measurement Validity



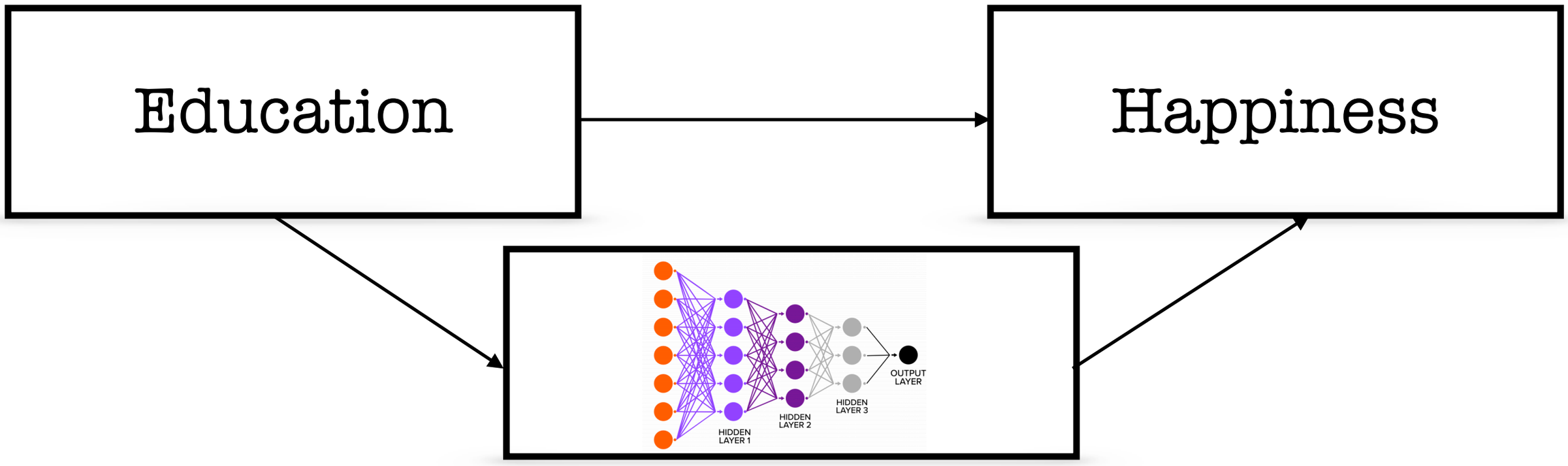
Applied Measurement Validity



Applied Measurement Validity



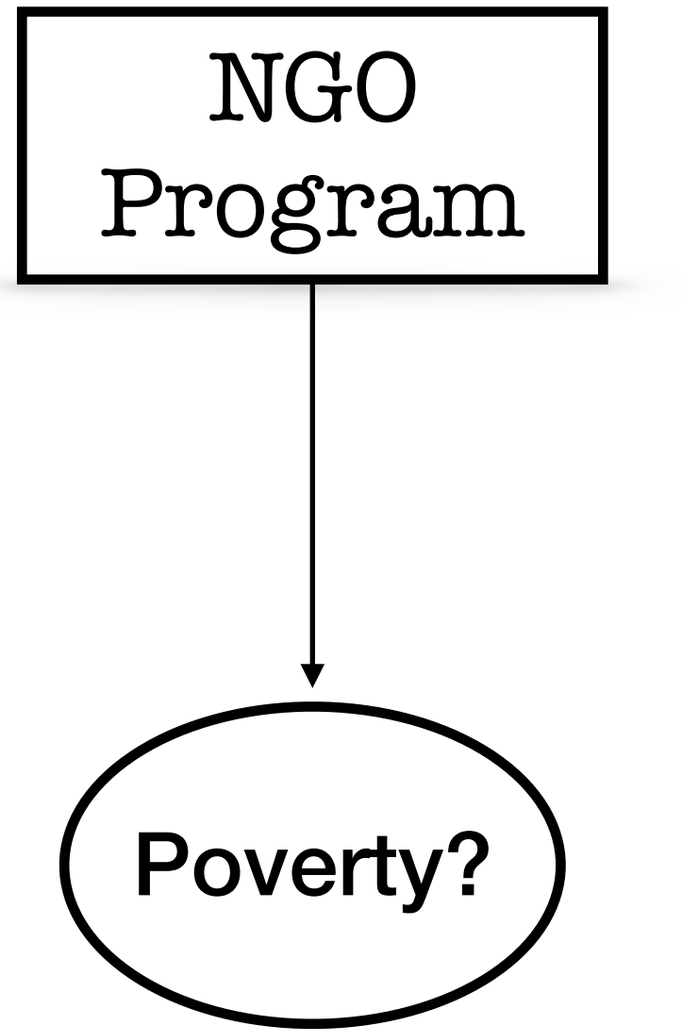
Applied Measurement Validity



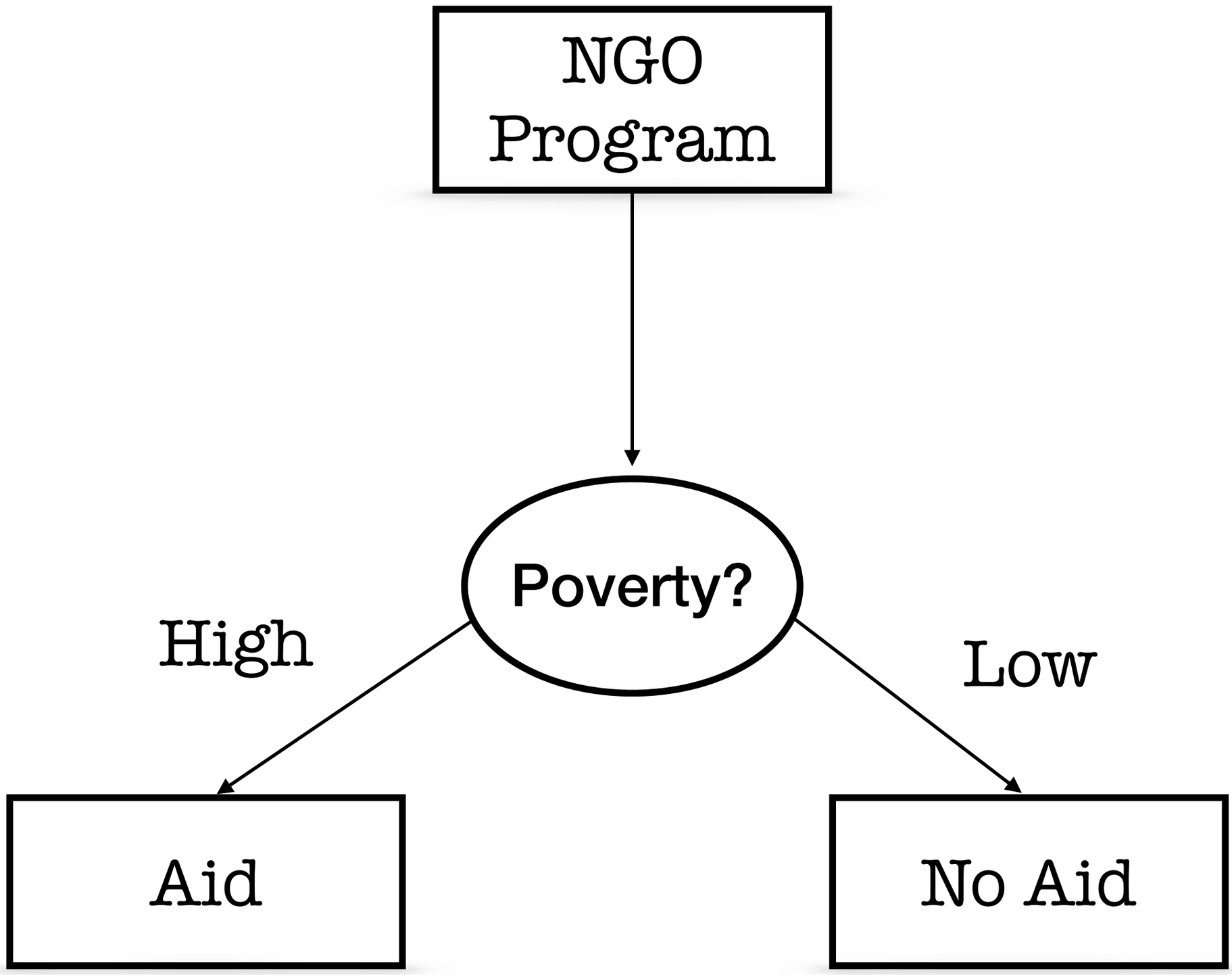
Applied Measurement Validity

NGO
Program

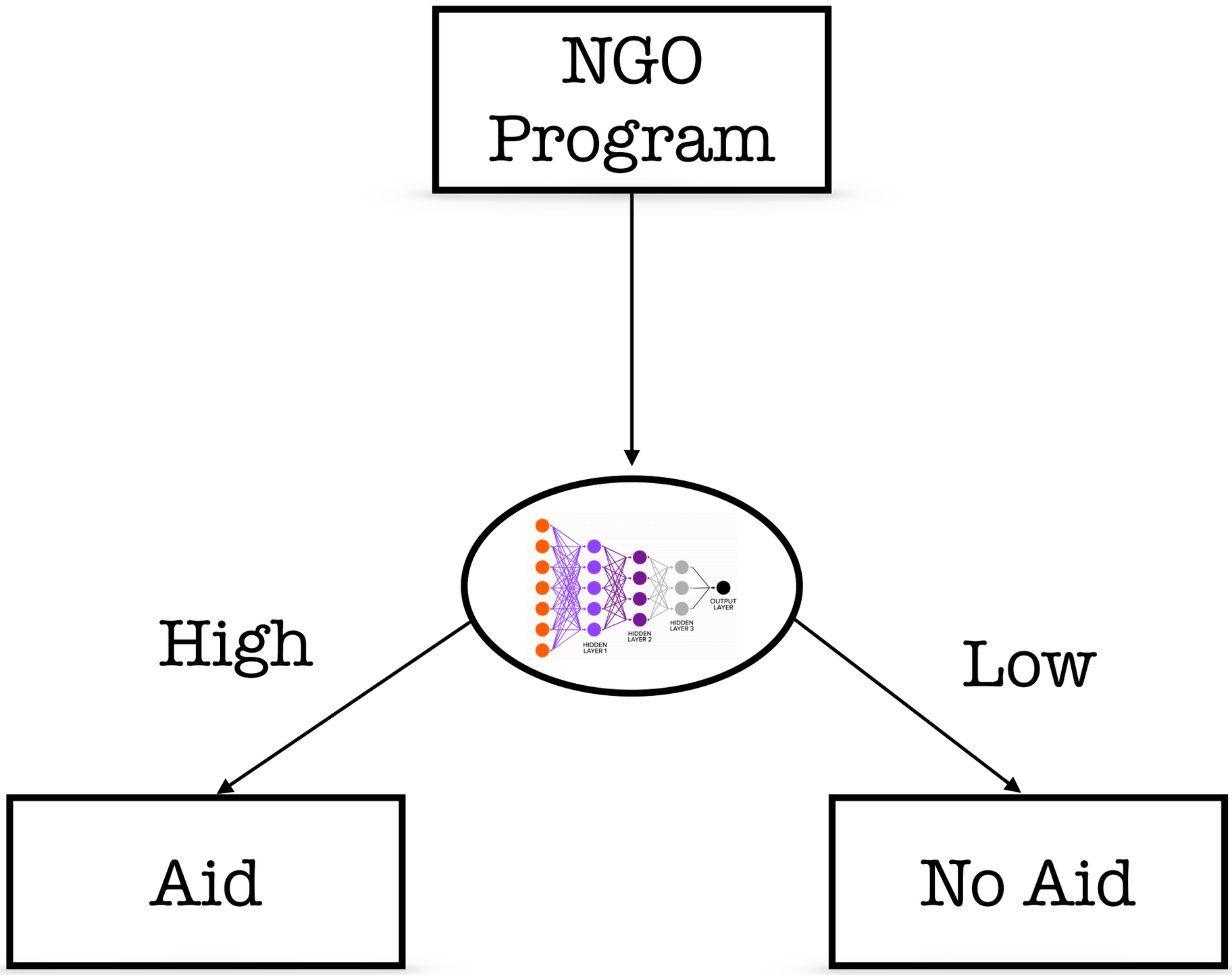
Applied Measurement Validity



Applied Measurement Validity



Applied Measurement Validity




Ceci n'est pas une pipe.

“The treachery of images”


%>%
Ceci n'est pas un pipe.
www.tidyverse.org

magrittR

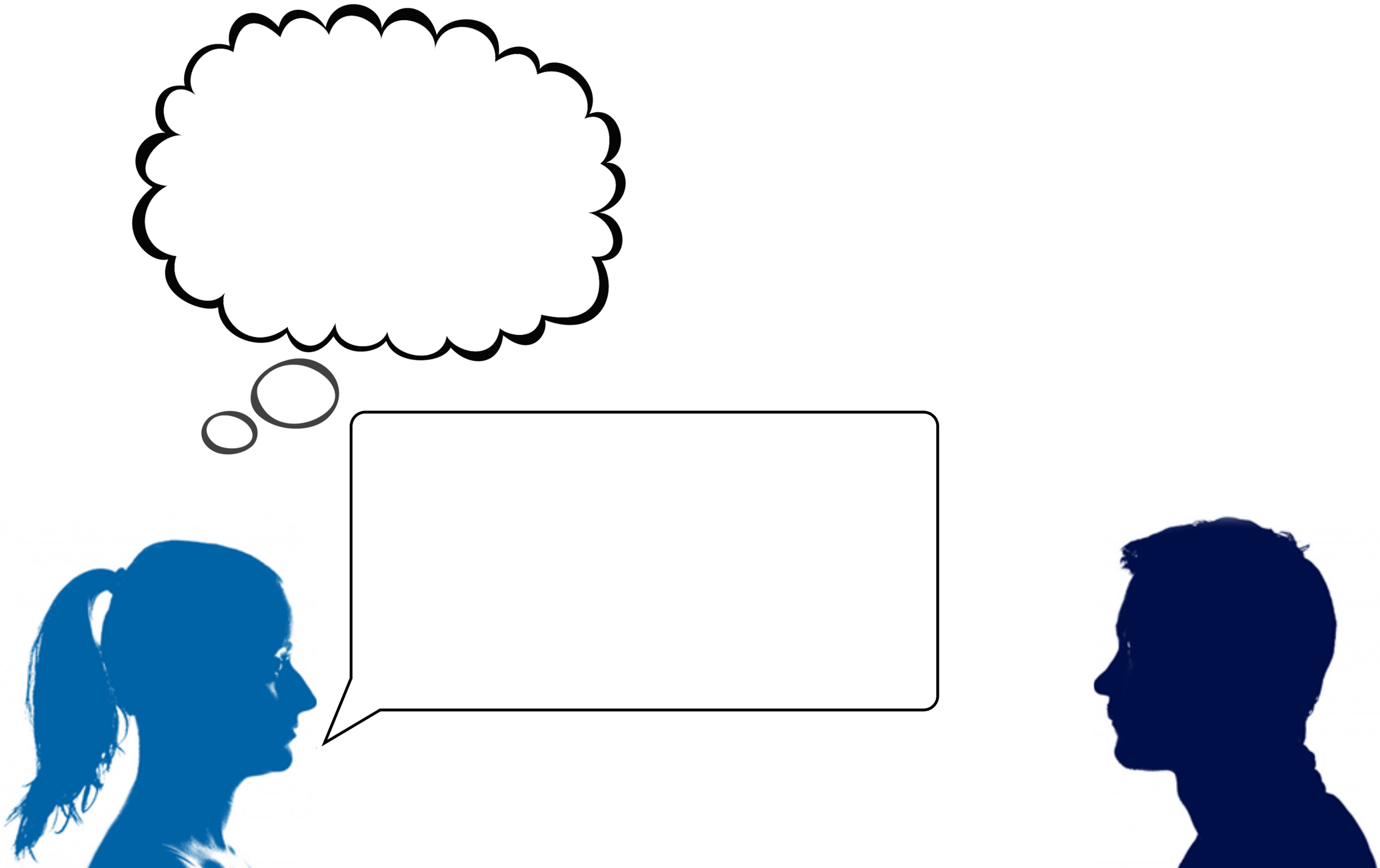
How is an Idea Communicated?

How is an Idea Communicated?

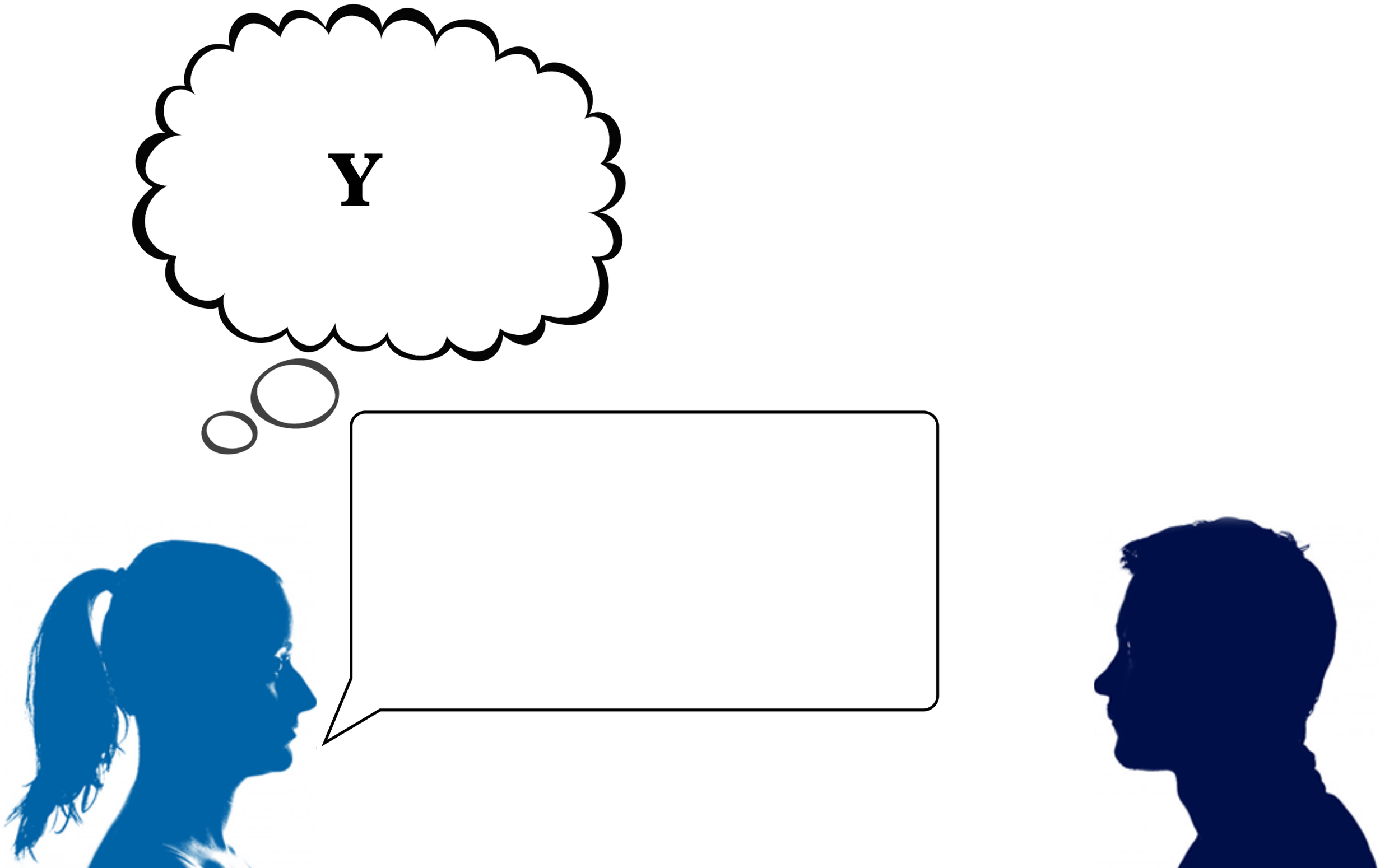
How is an Idea Communicated?

How is an Idea Communicated?

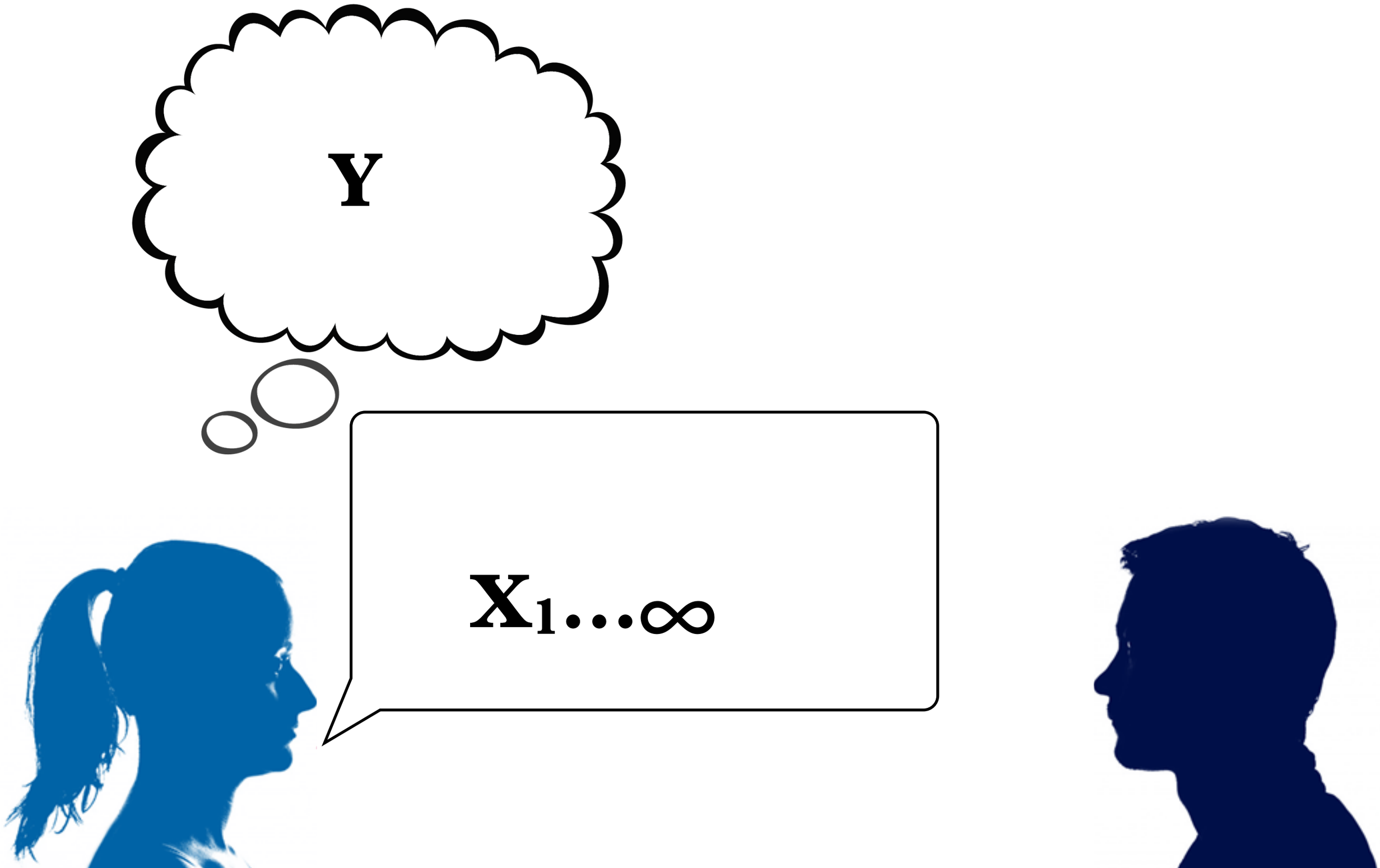
How is an Idea Communicated?



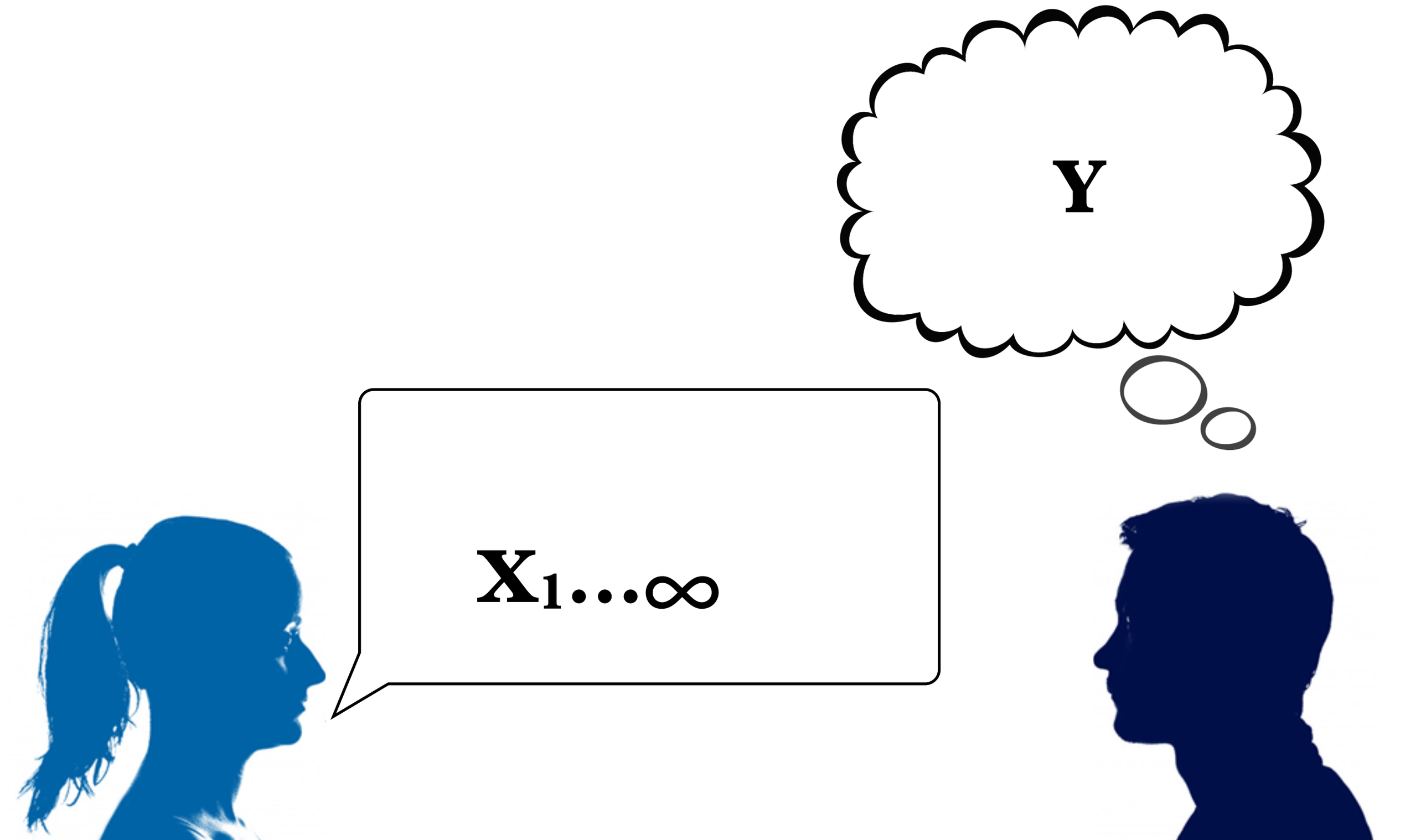
How is an Idea Communicated?



How is an Idea Communicated?



How is an Idea Communicated?



A Linguistic Model

$$\hat{y} = a_0 + e$$

A Linguistic Model

$$\hat{y} = a_0 + x_1 + x_2 + x_3 + \ldots + e$$

Feature Extraction

Select set of observables ***x***

Usually from theory or pre-trained model

Sometimes estimated (e.g. topic model)

A Linguistic Model

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Feature Extraction

Select set of observables ***x***

Usually from theory or pre-trained model

Sometimes estimated (e.g. topic model)

Feature Estimation

Determine **β** weights

A Linguistic Model

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Feature Extraction

Select set of observables ***x***

Usually from theory or pre-trained model

Sometimes estimated (e.g. topic model)

Feature Estimation

Determine **β** weights

Usually estimated empirically

A Linguistic Model

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Feature Extraction

- Select set of observables ***x***
- Usually from theory or pre-trained model
- Sometimes estimated (e.g. topic model)

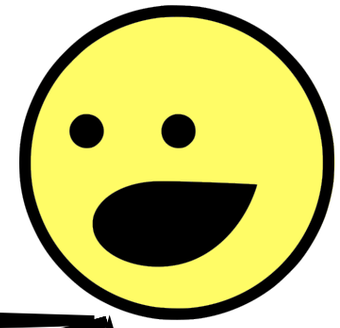
Feature Estimation

- Determine **β** weights
- Usually estimated empirically
- Sometimes guesstimated (e.g. equal weights)

Measurement in Language

Measurement in Language

I think he needs to do a better job of caring about the work environment he is in. If he loses his job which he is on the verge of, he would be in trouble



I am going to give them critical feedback on the work they have done so far. This is a big project and each of us has important work to contribute. I don't feel as if they are doing their best. I want to steer them in the right direction.

I plan to tell him that his failure to politely interact with the customers has resulted in lower sales and is currently putting his job at risk.

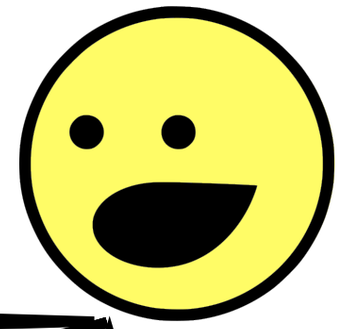
I plan on telling her that she needs to work on her hygiene because she hasn't been smelling very good at the office.

from Blunden, Green & Gino, 2018

Measurement in Language

I think he needs to do a better job of caring about the work environment he is in. If he loses his job which he is on the verge of, he would be in trouble

-1.9



-0.7

I am going to give them critical feedback on the work they have done so far. This is a big project and each of us has important work to contribute. I don't feel as if they are doing their best. I want to steer them in the right direction.

I plan to tell him that his failure to politely interact with the customers has resulted in lower sales and is currently putting his job at risk.

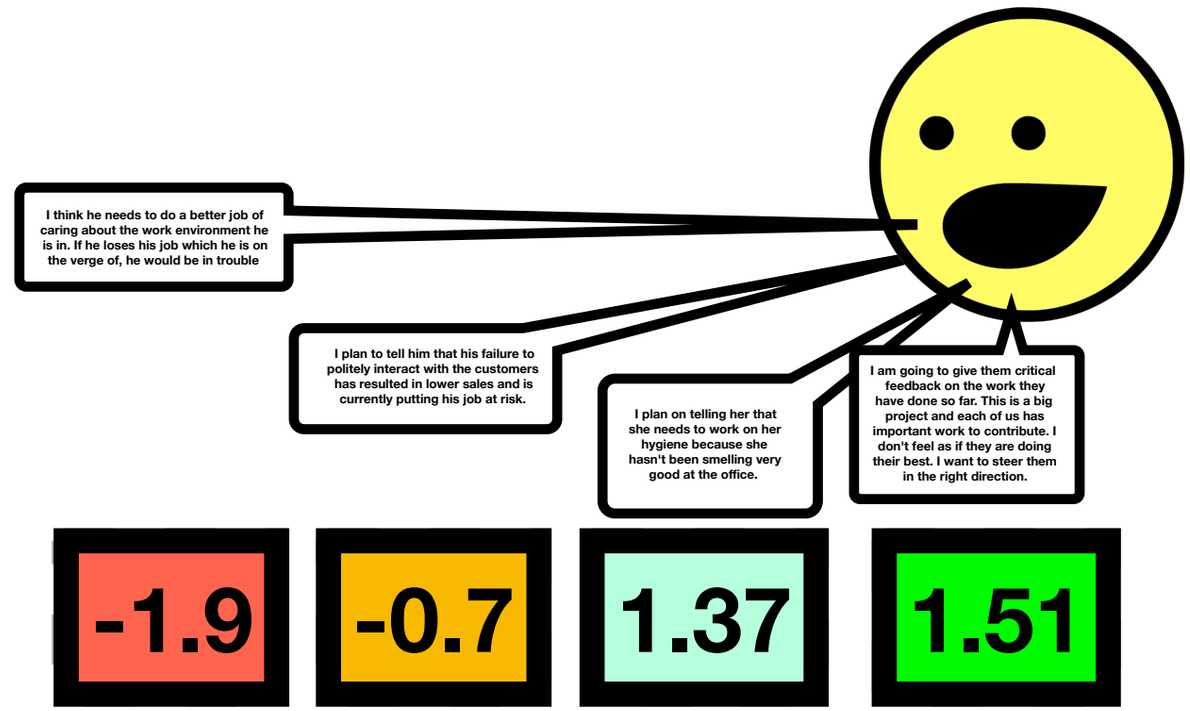
1.37

I plan on telling her that she needs to work on her hygiene because she hasn't been smelling very good at the office.

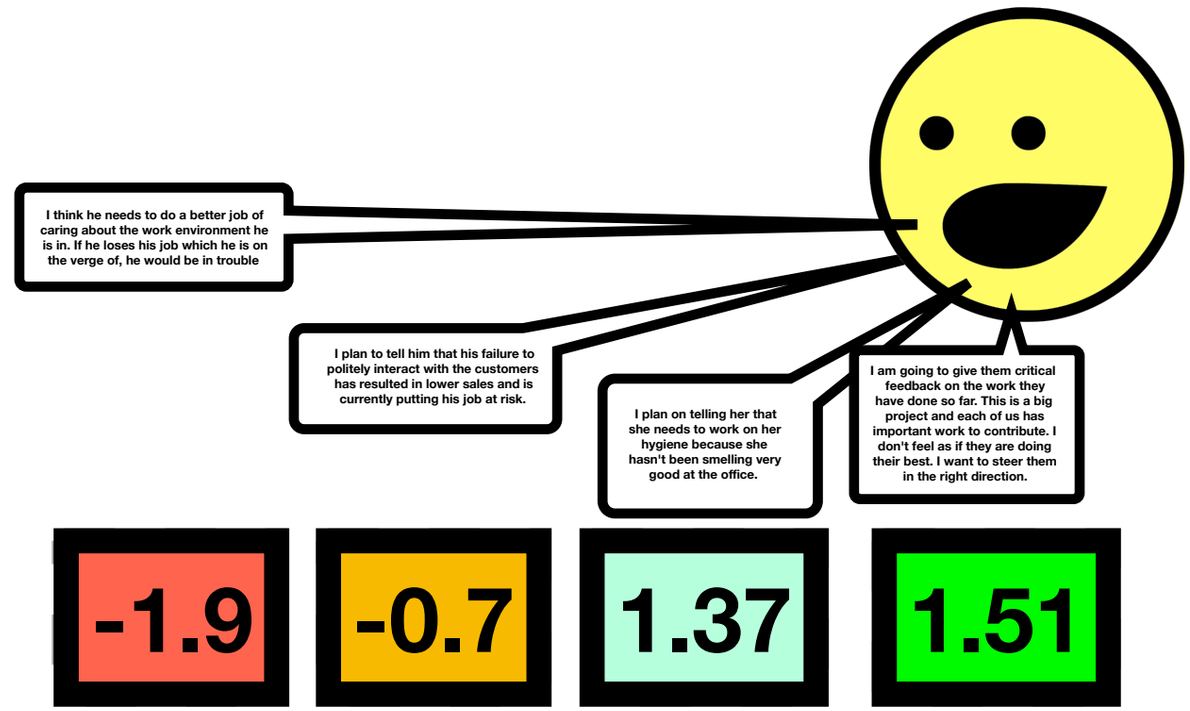
1.51

from Blunden, Green & Gino, 2018

Measurement in Language

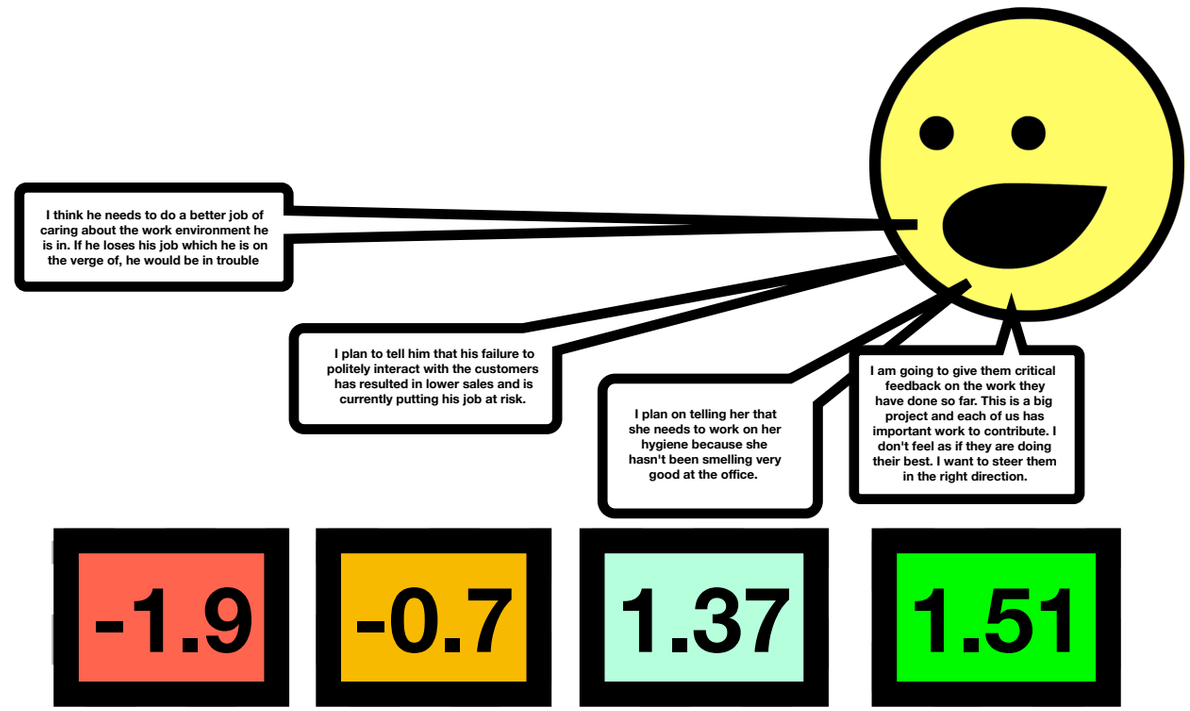


Measurement in Language



"Latent Variables"

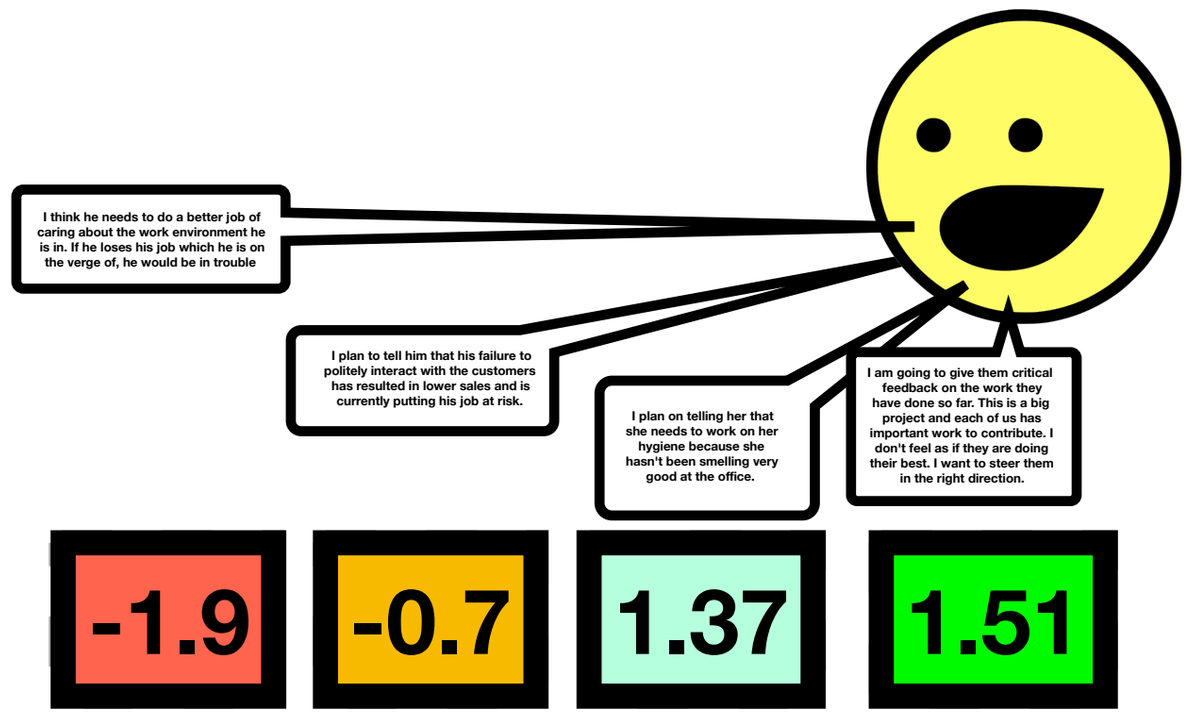
Measurement in Language



"Latent Variables"

Often from theory
Cannot observe directly
Product of author, listener, context, document, etc...

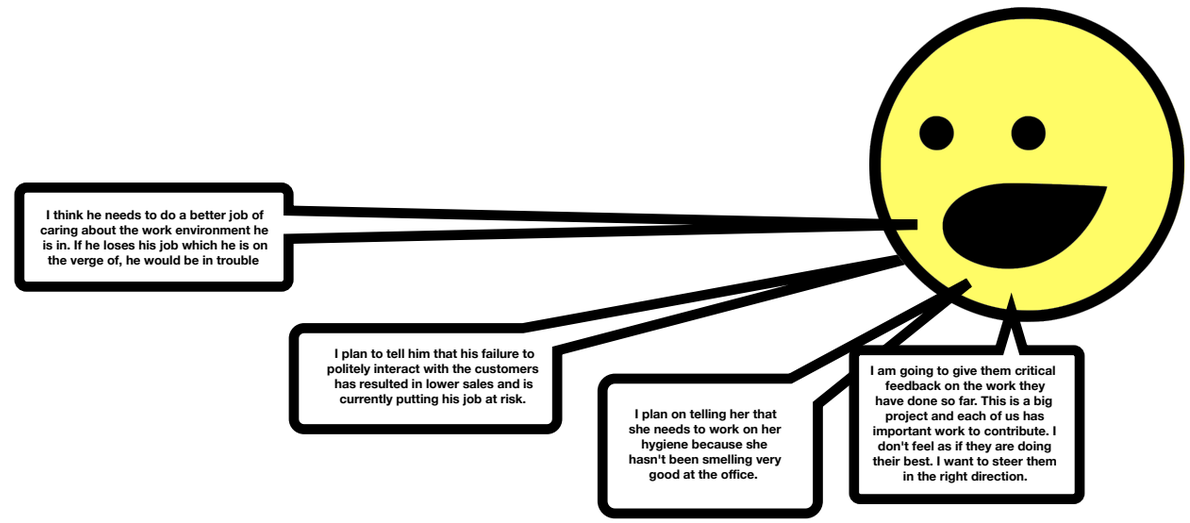
Measurement in Language



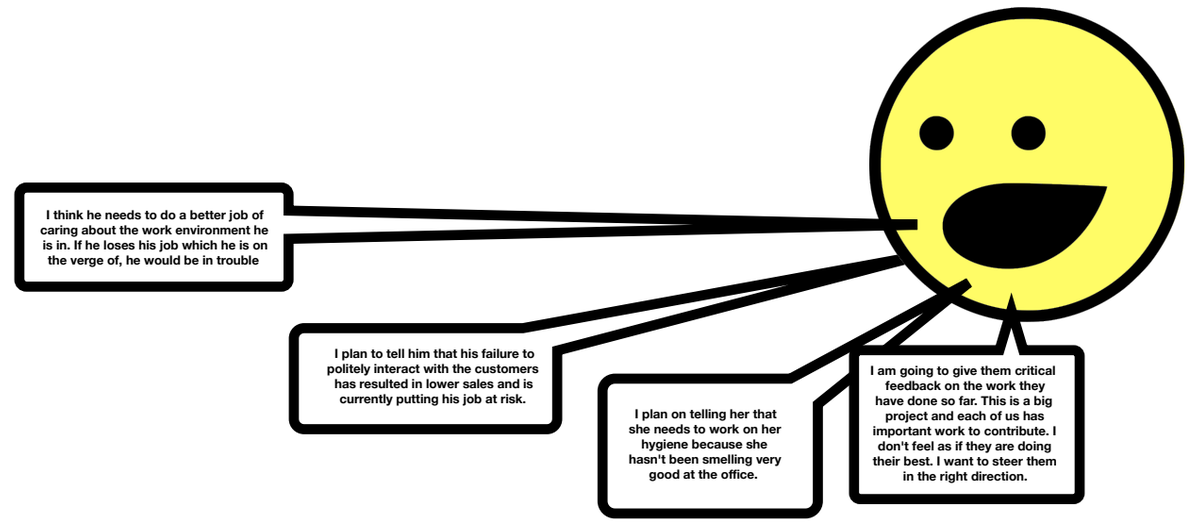
"Latent Variables"

Often from theory
Cannot observe directly
Product of author, listener, context, document, etc...

Measurement in Language



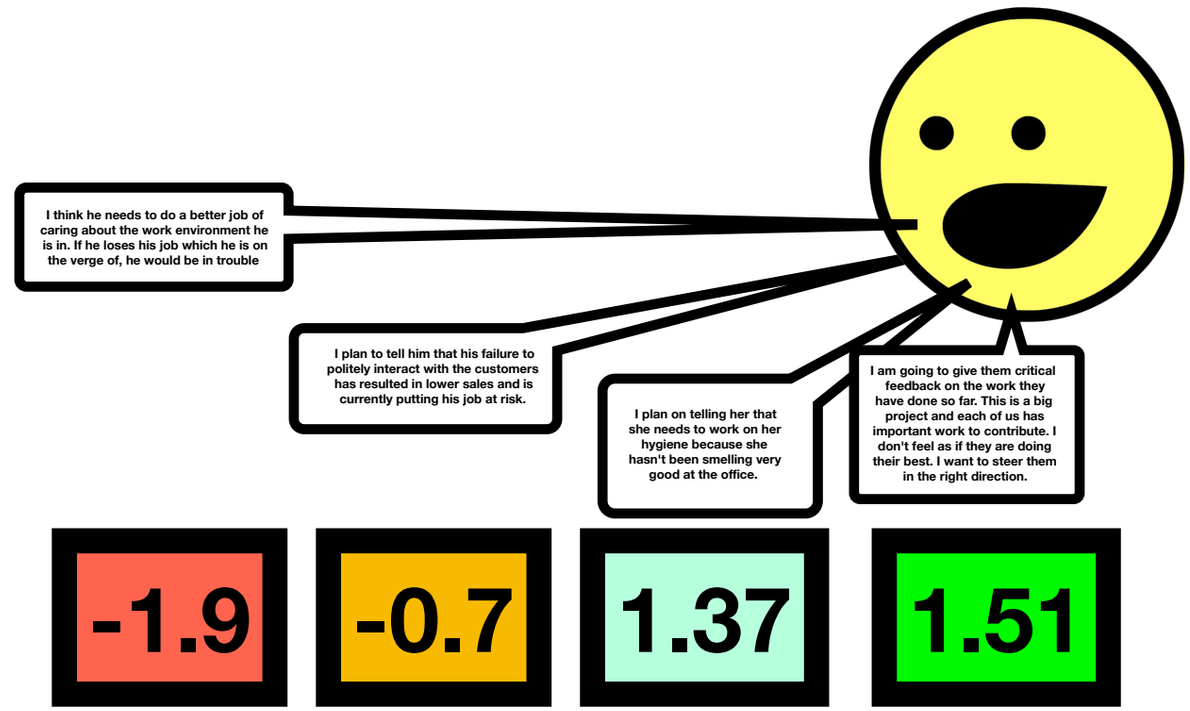
Measurement in Language



How specific was this advice?

How positive was this advice?

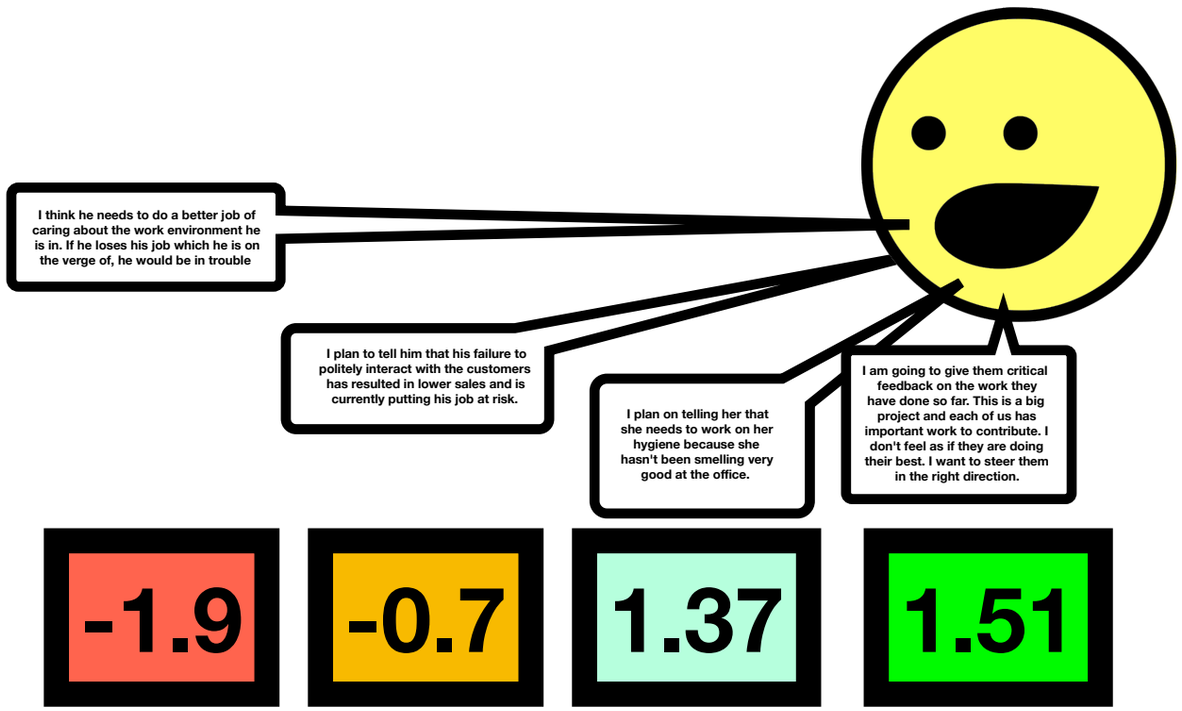
Measurement in Language



Measurement in Language

Humans: Plusses

What we've always been doing



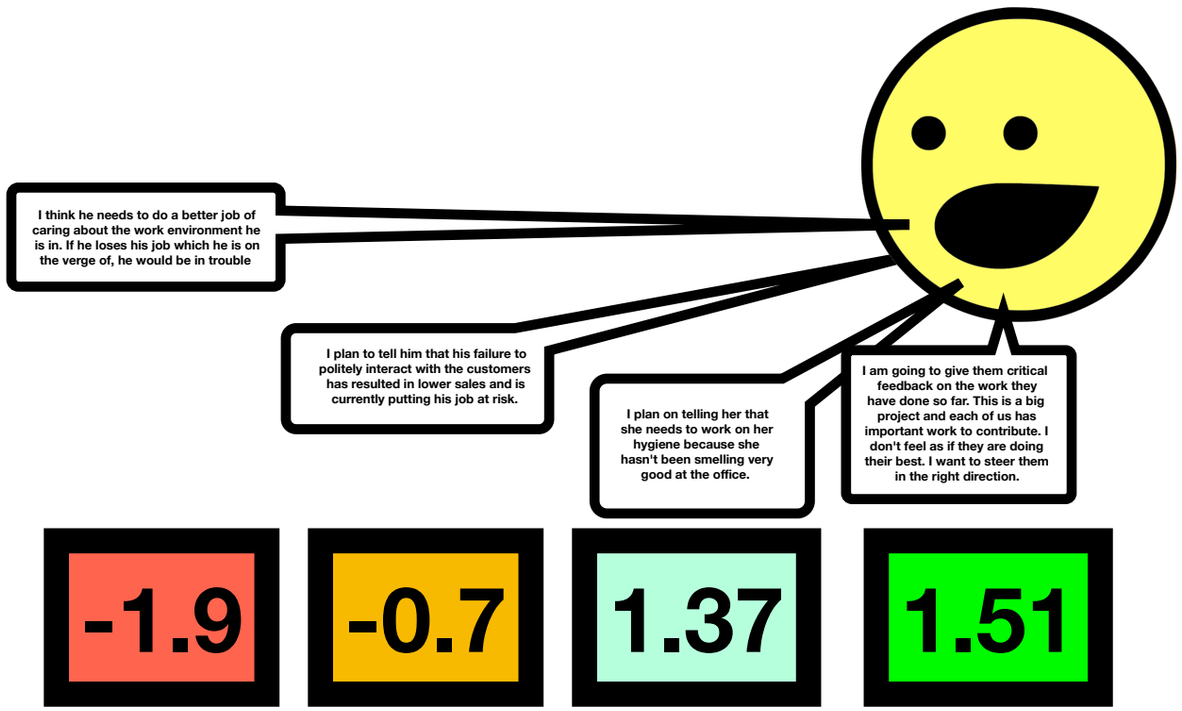
Measurement in Language

Humans: Plusses

What we've always been doing

More accurate than algorithms

for complex tasks



Measurement in Language

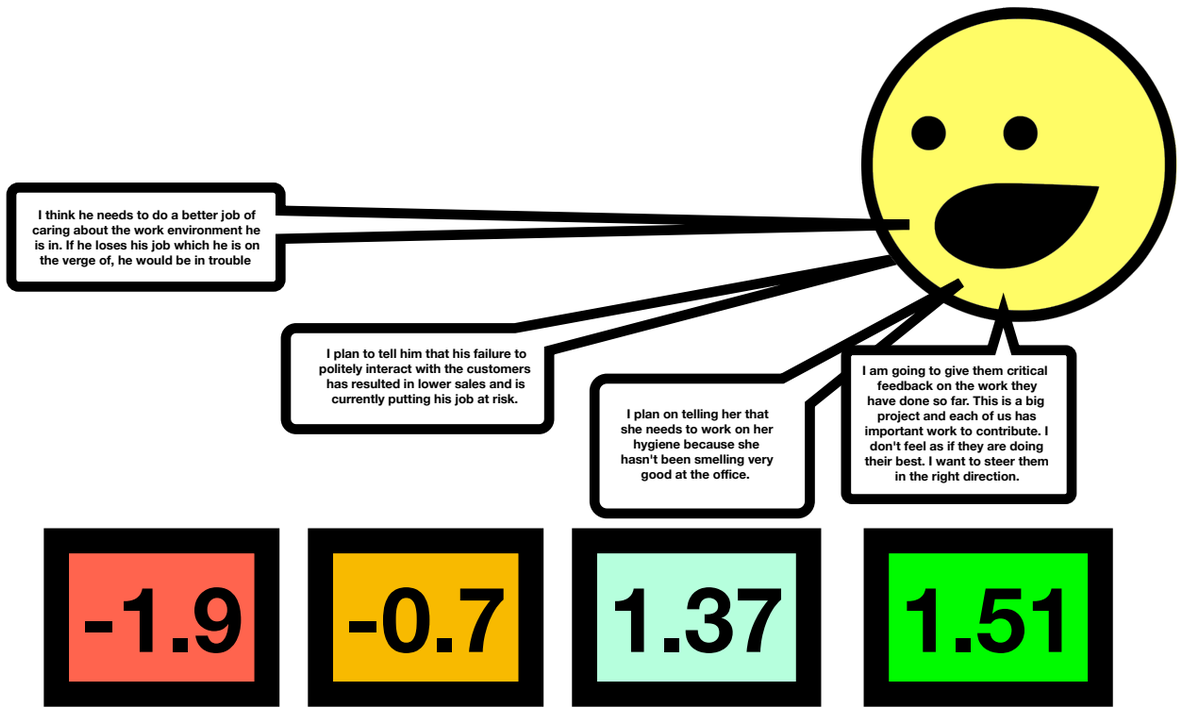
Humans: Plusses

What we've always been doing

More accurate than algorithms

for complex tasks

Can understand context, nuance



Measurement in Language

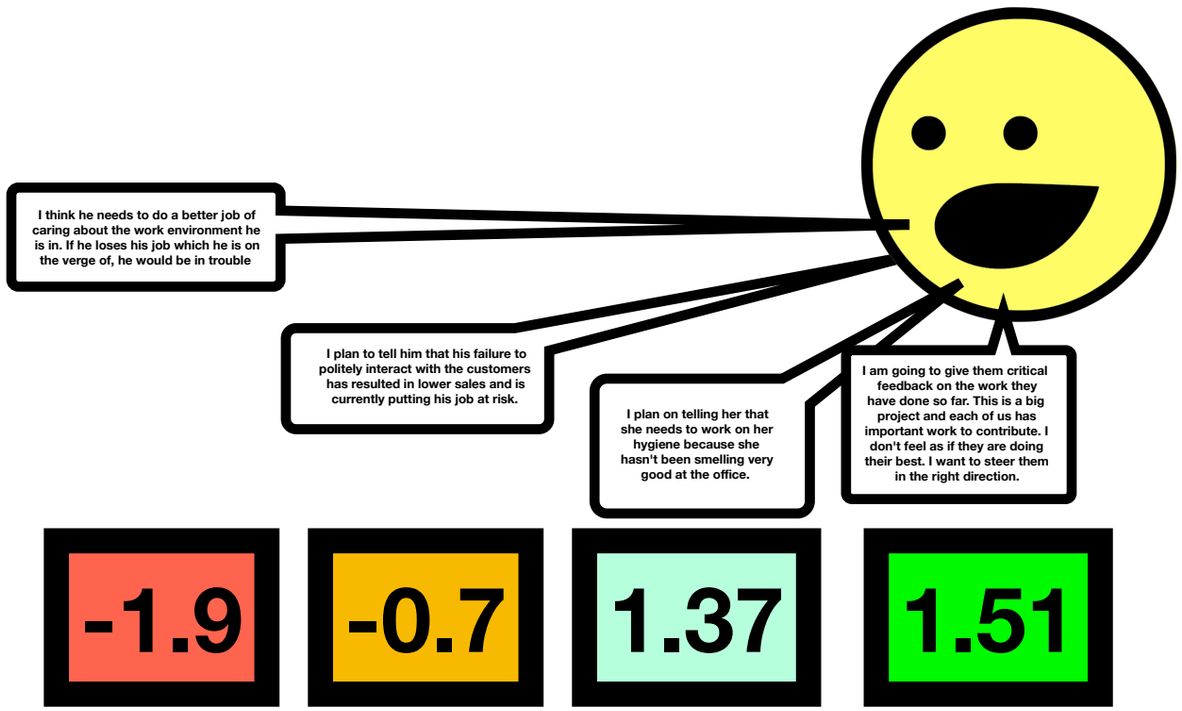
Humans: Plusses

What we've always been doing

More accurate than algorithms

for complex tasks

Can understand context, nuance



Humans: Minuses

High marginal cost of labor

Measurement in Language

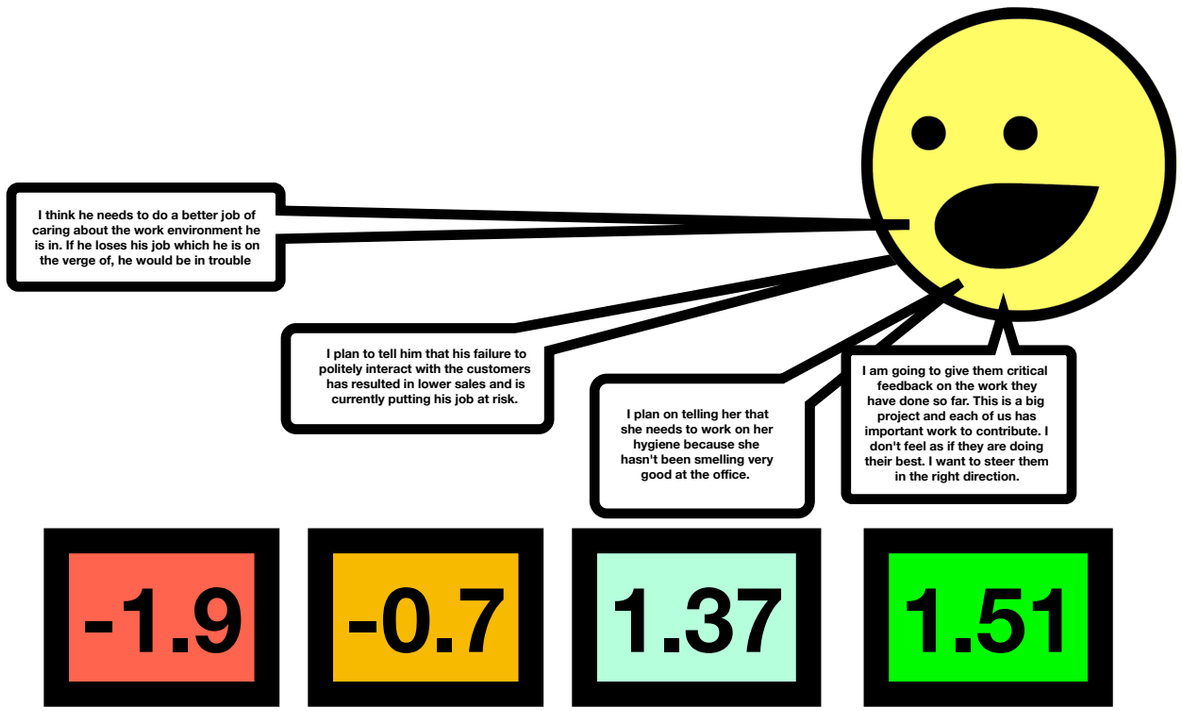
Humans: Plusses

What we've always been doing

More accurate than algorithms

for complex tasks

Can understand context, nuance



Humans: Minuses

High marginal cost of labor

Not reliable

Measurement in Language

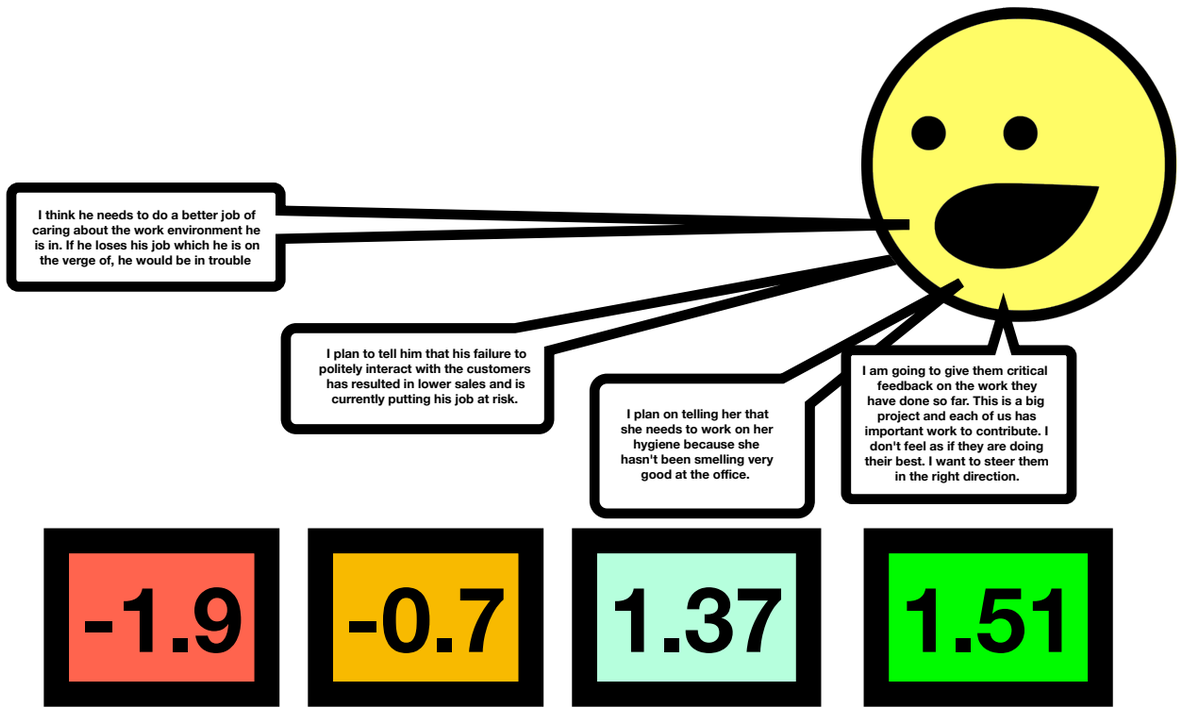
Humans: Plusses

What we've always been doing

More accurate than algorithms

for complex tasks

Can understand context, nuance



Humans: Minuses

High marginal cost of labor

Not reliable

Not transparent

Goals for NLP in Social Science

Goals for NLP in Social Science

Approximate good things about humans

Validate measures in existing theoretical/empirical framework

Measure with roughly the same accuracy as trained RAs

Account for context-specificity

Goals for NLP in Social Science

Approximate good things about humans

- Validate measures in existing theoretical/empirical framework
- Measure with roughly the same accuracy as trained RAs
- Account for context-specificity

Improve on bad things about humans

- Decrease marginal cost of measurement in new texts
- Increase replicability/reliability of measures
- Increase interpretability of resulting models

Goals for NLP in Social Science

There is no "correct model"

Depends on...

Goals for NLP in Social Science

There is no "correct model"

Depends on context, domain, time frame, research question, speakers, goals, sample size, measurement error, total compute, time pressure, audience

Goals for NLP in Social Science

There is no "correct model"

Depends on... a lot of things!

Think economically, start simple

Goals for NLP in Social Science

There is no "correct model"

Depends on... a lot of things!

Think economically, start simple

Is the juice worth
the squeeze?

Using NLP as a Human

NLP is not a substitute for reading
it is a complement for reading

Using NLP as a Human

NLP is not a substitute for reading
it is a complement for reading

If an expert cannot learn the model, an algorithm is unlikely to do better

Using NLP as a Human

NLP is not a substitute for reading
it is a complement for reading

If an expert cannot learn the model, an algorithm is unlikely to do better

Better training data > better algorithm
valid sample
cleaned/structured
human checked/annotated

Some Definitions

Some Definitions

An observation of text is called a *document*

Some Definitions

An observation of text is called a *document*

A group of documents is called a *corpus*

Some Definitions

An observation of text is called a *document*

A group of documents is called a *corpus*

A corpus may contain a complete population of documents,

or it may be a sample

Reasons for sampling: time, resources, privacy

Some Definitions

An observation of text is called a *document*

A group of documents is called a *corpus*

A corpus may contain a complete population of documents,

or it may be a sample

Reasons for sampling: time, resources, privacy

Population of interest - what group of writers/documents/targets

do you want to know about?

Some Definitions

Metadata - what else do you know about the documents?

Some Definitions

Metadata - what else do you know about the documents?

Writer characteristics

Audience characteristics

Stimulus characteristics

Document characteristics

Some Definitions

Metadata - what else do you know about the documents?

Writer characteristics

Audience characteristics

Stimulus characteristics

Document characteristics

Quantity of interest - what are you trying to estimate?

Supervised - map text features onto known variable

Unsupervised - discover new variables from text

Some Definitions

Metadata - what else do you know about the documents?

Writer characteristics

Audience characteristics

Stimulus characteristics

Document characteristics

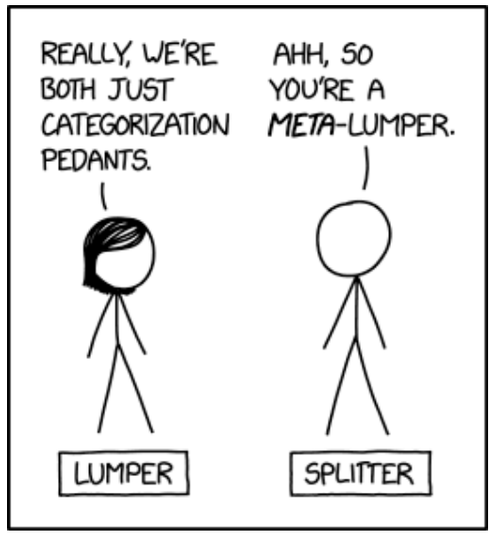
Quantity of interest - what are you trying to estimate?

Supervised - map text features onto known variable

Unsupervised - discover new variables from text

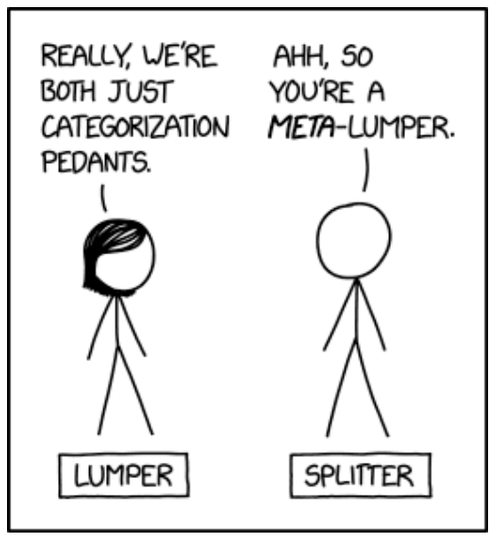
Lumping vs Splitting

A constant tension…..



Lumping vs Splitting

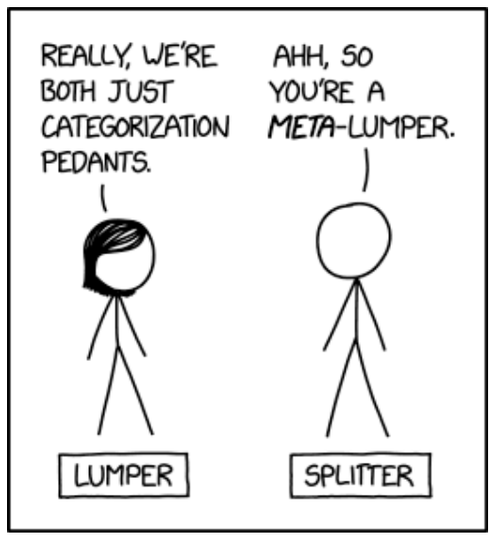
A constant tension…..



Lumping - combining features into a single count

Lumping vs Splitting

A constant tension…..

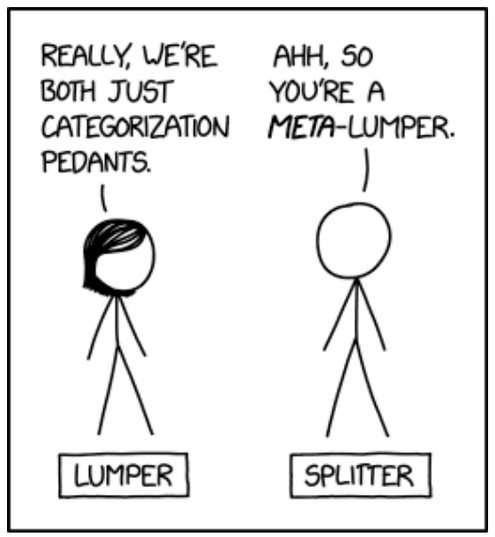


Lumping - combining features into a single count

Splitting - separating features into separate counts

Lumping vs Splitting

A constant tension…..



Lumping - combining features into a single count

Splitting - separating features into separate counts

We will mostly be lumping (but will split later on)

Lumping vs Splitting

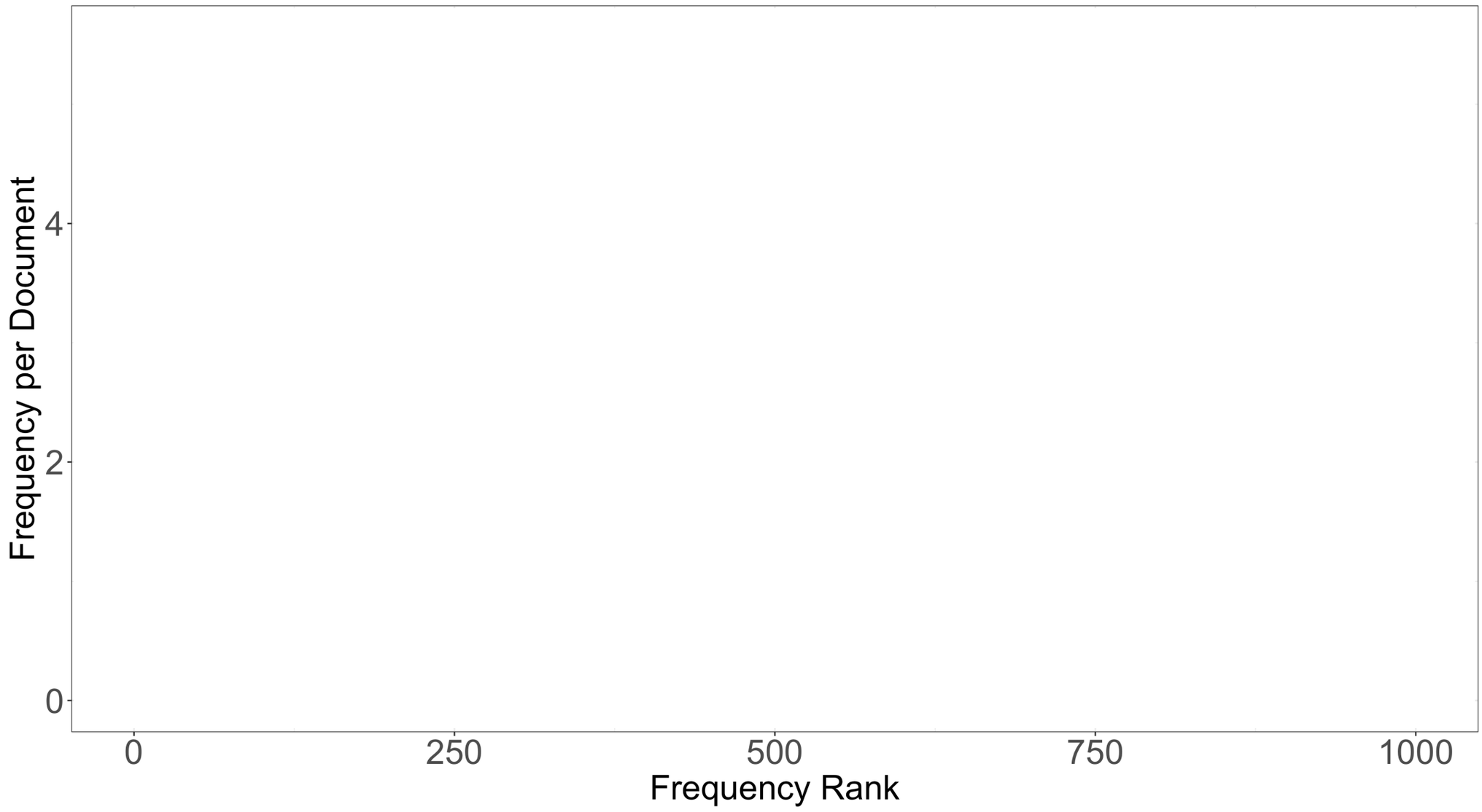
Why do we lump more than split?

"Zipf's law" - most words are very rare!

Lumping vs Splitting

Why do we lump more than split?

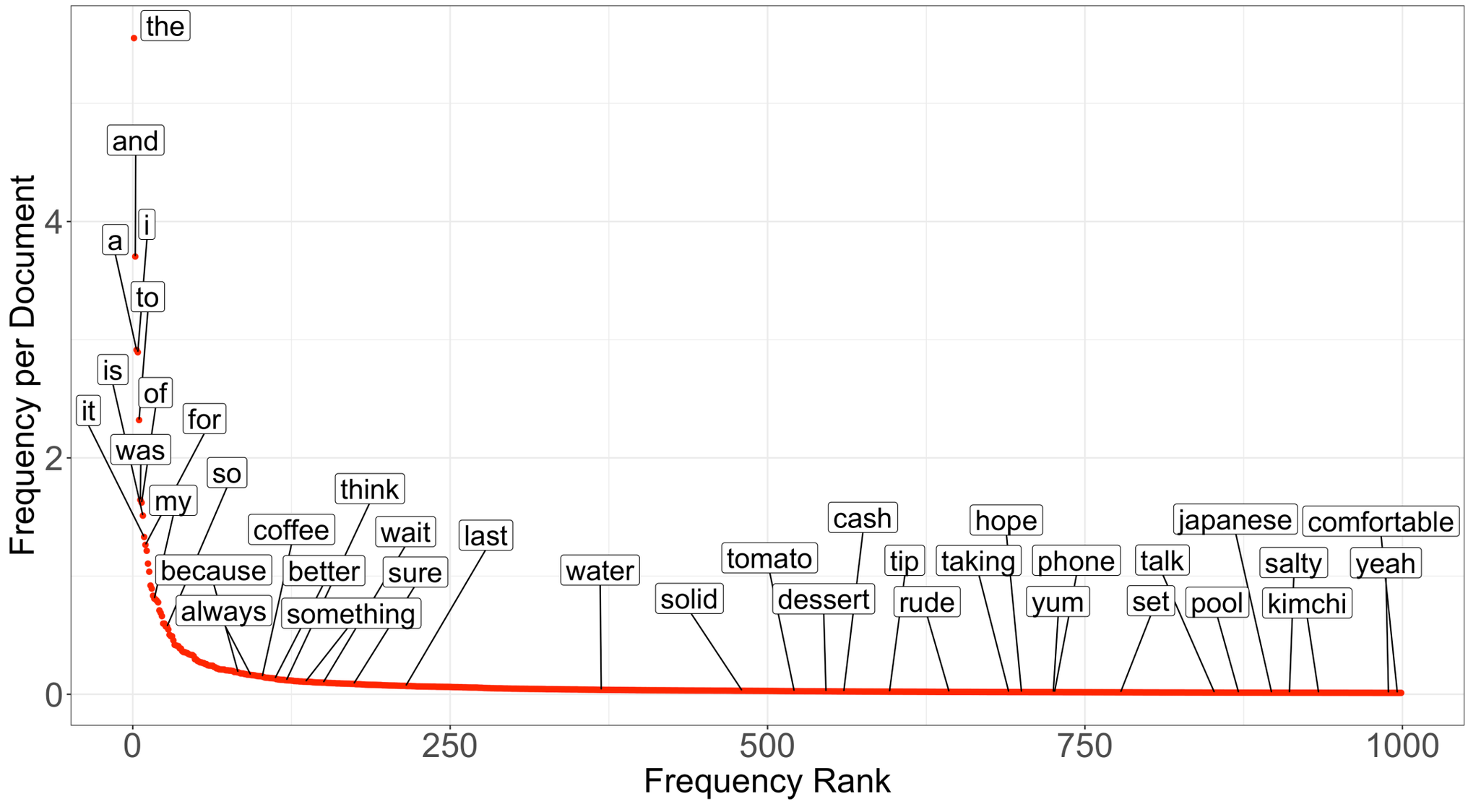
"Zipf's law" - most words are very rare!



Lumping vs Splitting

Why do we lump more than split?

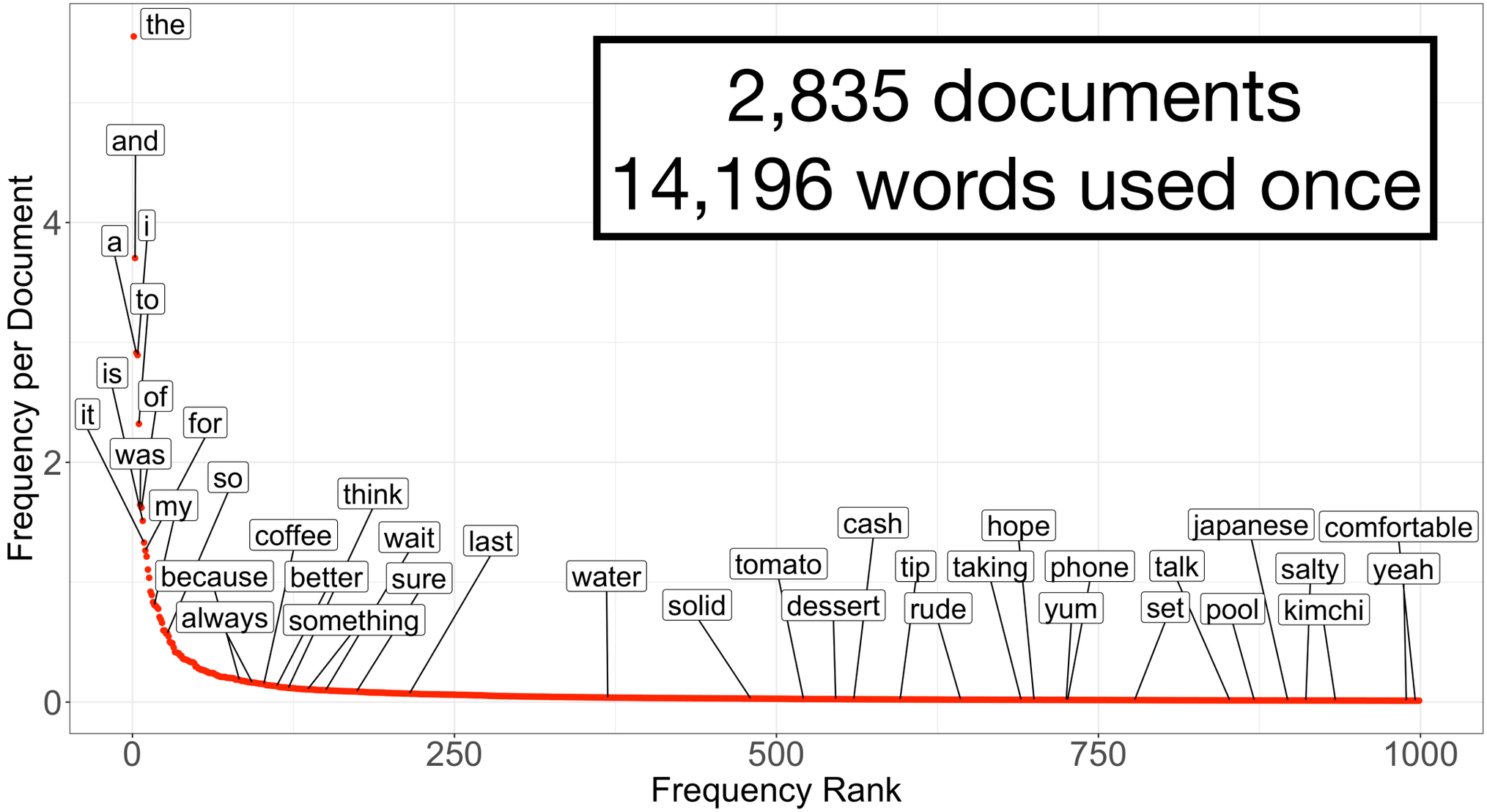
"Zipf's law" - most words are very rare!



Lumping vs Splitting

Why do we lump more than split?

"Zipf's law" - most words are very rare!



Feature Extraction Methods

Bag-of-words approaches

Structural approaches

Feature Extraction Methods

Bag-of-words approaches

- Ngrams
- Dictionaries
- Topic Models

Structural approaches

- Embedding Models
- Parsing Grammar
- Dialogue Acts

Feature Extraction Methods

Bag-of-words approaches

- Ngrams
- Dictionaries
- Topic Models

Structural approaches

- Embedding Models
- Parsing Grammar
- Dialogue Acts

The Bag of Words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1

Document-Feature Matrix

Document-Feature Matrix

The fundamental workhorse of text mining

Document-Feature Matrix

The fundamental workhorse of text mining

Coercing unstructured text into structured data

Document-Feature Matrix

The fundamental workhorse of text mining

Coercing unstructured text into structured data

Imagine a matrix...

Document-Feature Matrix

The fundamental workhorse of text mining

Coercing unstructured text into structured data

Imagine a matrix...

One row for every document

Document-Feature Matrix

The fundamental workhorse of text mining

Coercing unstructured text into structured data

Imagine a matrix...

One row for every document
One column for every feature of text

Document-Feature Matrix

The fundamental workhorse of text mining

Coercing unstructured text into structured data

Imagine a matrix...

One row for every document
One column for every feature of text
Each cell counts the feature i in document j

Document-Feature Matrix

Consider this corpus:

"This is a sentence"
"This is also one sentence"
"This is a document"
"This is not a document"

Document-Feature Matrix

Consider this corpus:

"This is a sentence"
"This is also one sentence"
"This is a document"
"This is not a document"

This	is	a	sentence	also	one	document	not

Document-Feature Matrix

Consider this corpus:

"This is a sentence"
"This is also one sentence"
"This is a document"
"This is not a document"

This	is	a	sentence	also	one	document	not

Document-Feature Matrix

Consider this corpus:

"This is a sentence"
"This is also one sentence"
"This is a document"
"This is not a document"

This	is	a	sentence	also	one	document	not
1	1	1	1	0	0	0	0

Document-Feature Matrix

Consider this corpus:

"This is a sentence"
"This is also one sentence"
"This is a document"
"This is not a document"

This	is	a	sentence	also	one	document	not
1	1	1	1	0	0	0	0
1	1	0	1	1	1	0	0

Document-Feature Matrix

Consider this corpus:

"This is a sentence"
"This is also one sentence"
"This is a document"
"This is not a document"

This	is	a	sentence	also	one	document	not
1	1	1	1	0	0	0	0
1	1	0	1	1	1	0	0
1	1	1	0	0	0	1	0
1	1	1	0	0	0	1	1

Document-Feature Matrix

Consider this corpus:

"I want coffee"
"I want milk"
"I want coffee, not tea"
"I want tea, not coffee"
"I want tea with milk"

I	want	coffee	milk	not	tea	with

Document-Feature Matrix

Consider this corpus:

"I want coffee"
"I want milk"
"I want coffee, not tea"
"I want tea, not coffee"
"I want tea with milk"

I	want	coffee	milk	not	tea	with
1	1	1	0	0	0	0

Document-Feature Matrix

Consider this corpus:

"I want coffee"
"I want milk"
"I want coffee, not tea"
"I want tea, not coffee"
"I want tea with milk"

I	want	coffee	milk	not	tea	with
1	1	1	0	0	0	0
1	1	0	1	0	0	0
1	1	1	0	1	1	0
1	1	1	0	1	1	0
1	1	0	1	0	1	1

Document-Feature Matrix

Consider this corpus:

"I want coffee"
"I want milk"
"I want coffee, not tea"
"I want tea, not coffee"
"I want tea with milk"

I	want	coffee	milk	not	tea	with
1	1	1	0	0	0	0
1	1	0	1	0	0	0
1	1	1	0	1	1	0
1	1	1	0	1	1	0
1	1	0	1	0	1	1

Pre-Processing the Bag

Pre-Processing the Bag

Punctuation

"This is a sentence!"
"This is a sentence?"
"This is a sentence."
"This is?"

Pre-Processing the Bag

Punctuation

"This is a sentence!"
"This is a sentence?"
"This is a sentence."
"This is?"

This	is	a	sentence	xmark	qmark
1	1	1	1	1	0
1	1	0	1	0	1
1	1	1	1	0	0
1	1	0	0	0	1

Pre-Processing the Bag

Punctuation

"This is a sentence!"
"This is a sentence?"
"This is a sentence."
"This is?"

This	is	a	sentence	xmark	qmark
1	1	1	1	1	0
1	1	0	1	0	1
1	1	1	1	0	0
1	1	0	0	0	1

[Same strategy for emojis]

Pre-Processing the Bag

Capitalisation

"This is the sentence."
"The sentence is this."
"This is the mark.”
"This is Mark."

Pre-Processing the Bag

Capitalisation

"This is the sentence."
"The sentence is this."
"This is the mark.”
"This is Mark."

This	is	the	sentence	The	this	mark	Mark
1	1	1	1	0	0	0	0
0	1	0	1	1	1	0	0
1	1	1	0	0	0	1	0
1	1	0	0	0	0	0	1

Pre-Processing the Bag

Word Stemming

Pre-Processing the Bag

Word Stemming

Pre-Processing the Bag

Word Stemming

"This is a sentence."
“This is sentenced.”
“These are sentences."

this	is	a	sentence	sentenced	these	are	sentences
1	1	1	1	0	0	0	0
1	1	0	0	1	0	0	0
0	0	0	0	0	1	1	1

Pre-Processing the Bag

Word Stemming

"This is a sentence."
“This is sentenced.”
“These are sentences."

this	is	a	sentenc	these	are
1	1	1	1	0	0
1	1	0	1	0	0
0	0	0	1	1	1

Pre-Processing the Bag

Contractions

"This is a sentence."
"This isn’t a sentence."
"This is not a sentence."

Pre-Processing the Bag

Contractions

"This is a sentence."
"This isn’t a sentence."
"This is not a sentence."

this	is	a	sentence	isn’t	not
1	1	1	1	0	0
1	0	1	1	1	0
1	1	1	1	0	1

Pre-Processing the Bag

Contractions

"This is a sentence."
"This isn’t a sentence.” -> “This is not a sentence”
"This is not a sentence."

Pre-Processing the Bag

Contractions

"This is a sentence."
"This isn’t a sentence.” -> “This is not a sentence”
"This is not a sentence."

this	is	a	sentence	not
1	1	1	1	0
1	1	1	1	1
1	1	1	1	1

Pre-Processing the Bag

Dropping common “stop” words

"This is a sentence."
"this is also a sentence."
"here is a rare word.”
"here is another word."
"and another sentence.”

Pre-Processing the Bag

Dropping common “stop” words

"This is a sentence."
"this is also a sentence."
"here is a rare word.”
"here is another word."
"and another sentence.”

this	is	a	sentence	also	here	rare	word	another	and
1	1	1	1	0	0	0	0	0	0
1	1	1	1	1	0	0	0	0	0
0	1	1	0	0	1	1	1	0	0
0	1	0	0	0	1	0	1	1	0
0	0	0	1	0	0	0	0	1	1

Pre-Processing the Bag

Dropping common “stop” words

"This is a sentence."
"this is also a sentence."
"here is a rare word.”
"here is another word."
"and another sentence.”

sentence	also	rare	word	another
1	0	0	0	0
1	1	0	0	0
0	0	1	1	0
0	0	0	1	1
1	0	0	0	1

Pre-Processing the Bag

Dropping common “stop” words

i	down	is	should've	she's	more	because	haven	yourself	why	a	didn't
me	in	are	now	her	most	as	haven't	yourselves	how	an	doesn
my	out	was	d	hers	other	until	isn	he	all	the	doesn't
myself	on	were	ll	herself	some	while	isn't	him	any	and	hadn
we	off	be	m	it	such	of	ma	his	both	but	hadn't
our	over	been	o	it's	no	at	mightn	himself	each	if	hasn
ours	under	being	re	its	nor	by	mightn't	she	few	or	hasn't
ourselves	again	have	ve	itself	not	for	mustn	who	s	before	wasn
you	further	has	y	they	only	with	mustn't	whom	t	after	wasn't
you're	then	had	ain	them	own	about	needn	this	can	above	weren
you've	once	having	aren	their	same	against	needn't	that	will	below	weren't
you'll	here	do	aren't	theirs	so	between	shan	that'll	just	to	won
you'd	there	does	couldn	themselves	than	into	shan't	these	don	from	won't
your	when	did	couldn't	what	too	through	shouldn	those	don't	up	wouldn
yours	where	doing	didn	which	very	during	shouldn't	am	should		wouldn't

Pre-Processing the Bag

Dropping common “stop” words

i	down	is	should've	she's	more	because	aven	yourself	why	a	dn't
me	in	are	now	her	most	as	haven't	yourselves	how	an	doesn
my	out	was	d	hers	other	until	isn	he	all	the	doesn't
myself	on	were	ll	herself	some	while	isn't	him	any	and	hadn
we	off	be	m	it	h	of	ma	is	both	but	hadn't
our	over	been	o	it's	no	t	mightn	imself	each	if	hasn
ours	under	being	re	its	nor	y	mightn't	he	few	or	hasn't
ourselves	again	have	ve	itself	not	or	mustn	who	s	befo	wasn
you	further	has	y	they	only	with	mustn't	whom	t	after	wasn't
you're	then	had	ain	them	own	about	needn	is	can	abov	weren
you've	once	having	aren	their	same	against	needn't	at	will	belo	weren't
you'll	here	do	aren't	theirs	so	betwee	shan	hat'll	just	to	won
you'd	there	does	couldn	themselves	than	into	shan't	hese	don	om	won't
your	when	did	couldn't	what	too	through	shouldn	those	don't	p	wouldn
yours	where	doing	didn	which	very	during	shouldn'	am	should		wouldn't

Pre-Processing the Bag

Down-weighting common words

tf-idf

"This is a sentence."
"this is also a sentence."
"here is a rare word."
"here is another word."
"and another sentence."

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

this	is	a	sentence	also	here	rare	word	another	and
1	1	1	1	0	0	0	0	0	0
1	1	1	1	1	0	0	0	0	0
0	1	1	0	0	1	1	1	0	0
0	1	0	0	0	1	0	1	1	0
0	0	0	1	0	0	0	0	1	1

Pre-Processing the Bag

Down-weighting common words

tf-idf

"This is a sentence."
"this is also a sentence."
"here is a rare word."
"here is another word."
"and another sentence."

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

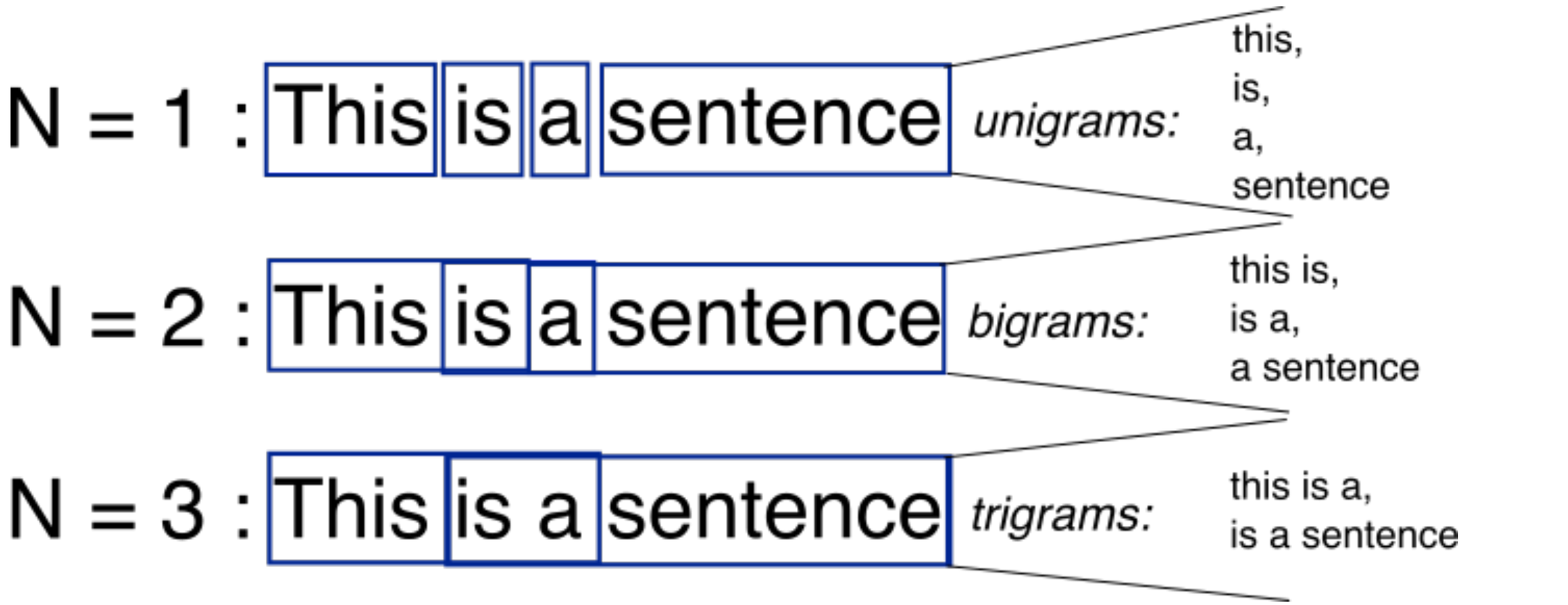
this	is	a	sentence	also	here	rare	word	another	and
0.4	0.1	0.22	0.22	0	0	0	0	0	0
0.4	0.1	0.22	0.22	0.7	0	0	0	0	0
0	0.1	0.22	0	0	0.4	0.7	0.4	0	0
0	0.1	0	0	0	0.4	0	0.4	0.4	0
0	0	0	0.22	0	0	0	0	0.4	0.7

Pre-Processing the Bag

Constructing phrases - “n-grams”

Pre-Processing the Bag

Constructing phrases - “n-grams”



Pre-Processing the Bag

Constructing phrases - “n-grams”

"This is a test."
"This is a guess."

Pre-Processing the Bag

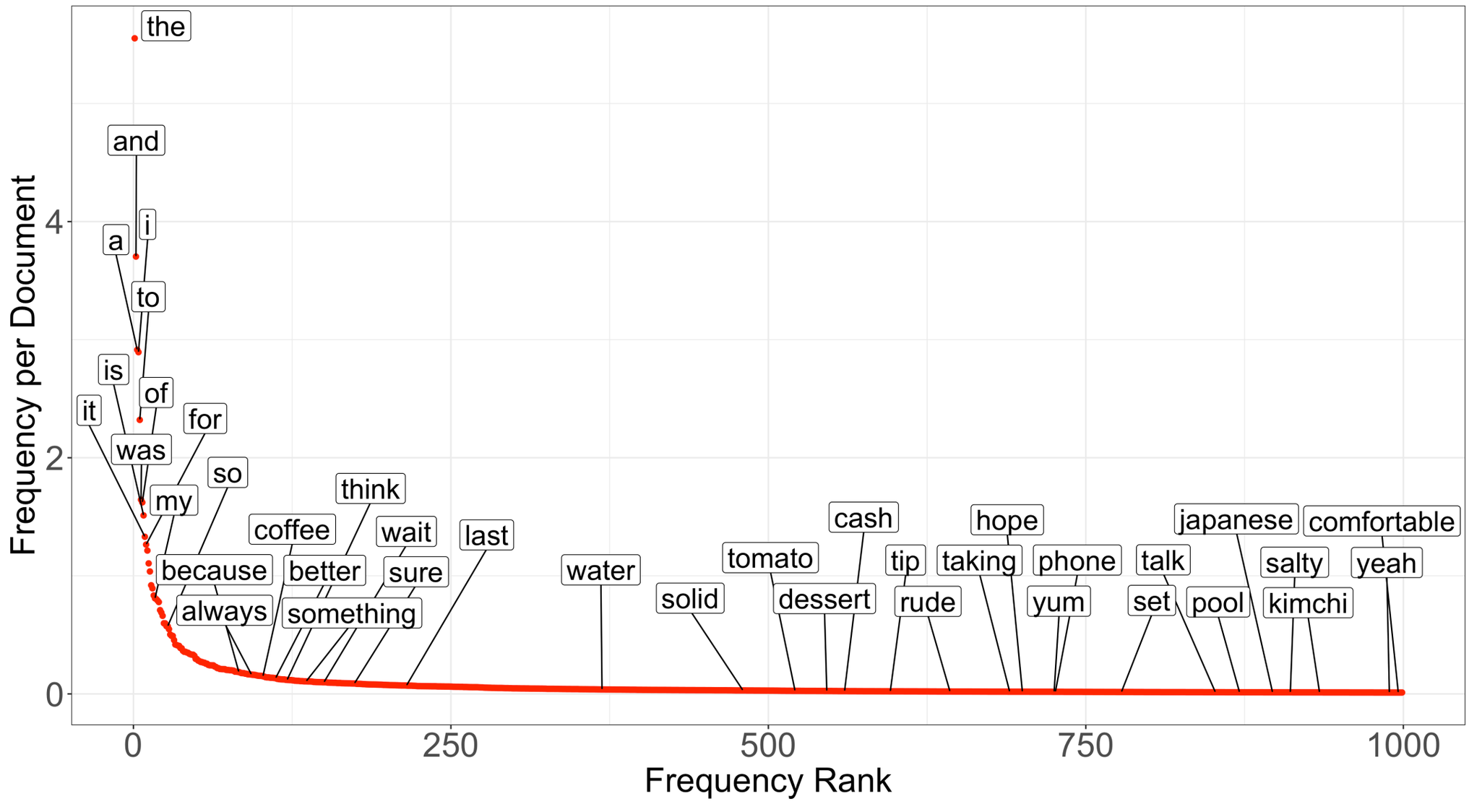
Constructing phrases - “n-grams”

"This is a test."
"This is a guess."

this	is	a	test	this_is	is_a	a_test	this_is_a	is_a_test	guess	a_guess	is_a_guess
1	1	1	1	1	1	1	1	1	0	0	0
1	1	1	0	1	1	0	1	0	1	1	1

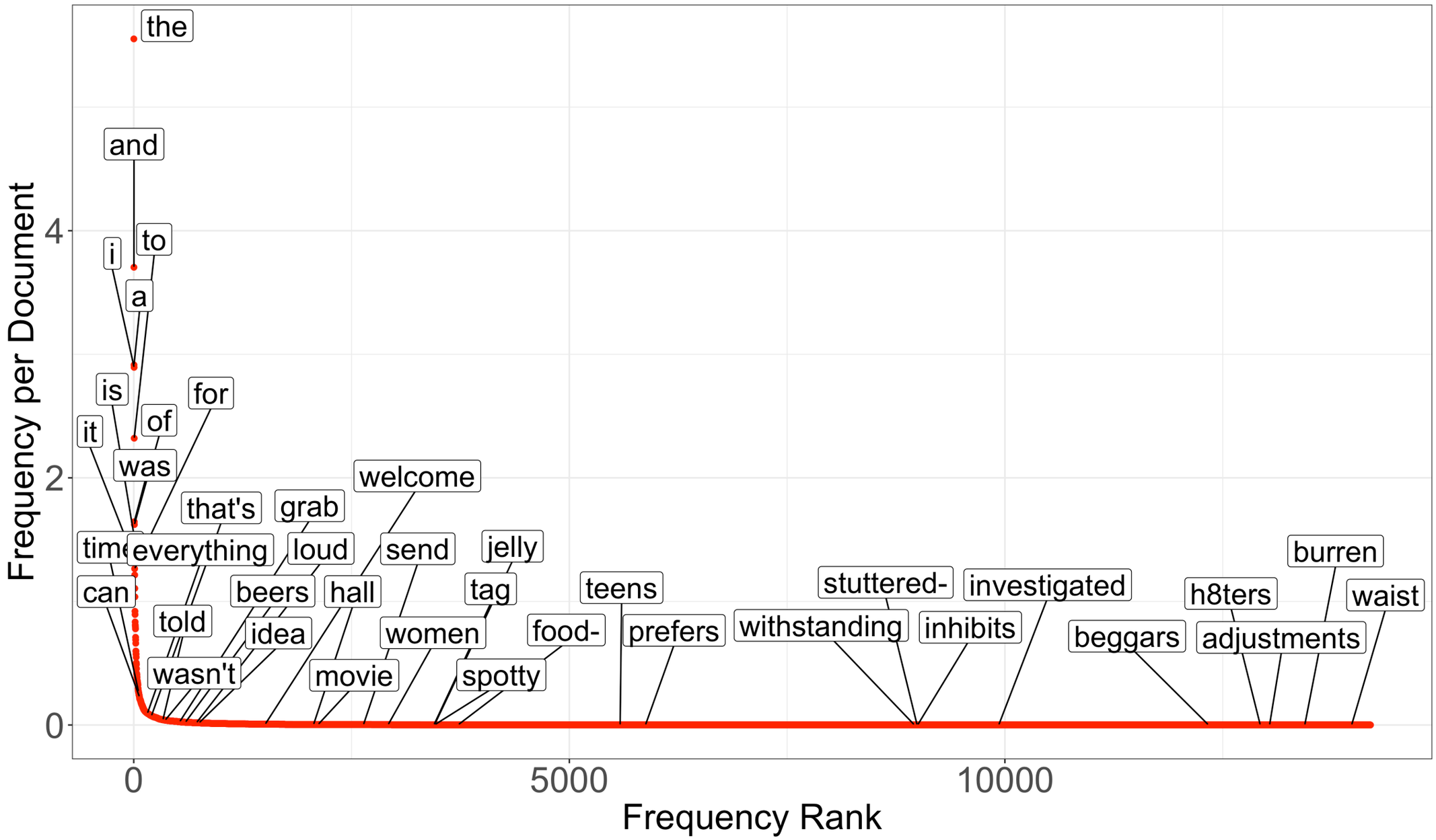
Pre-Processing the Bag

Filtering rare words



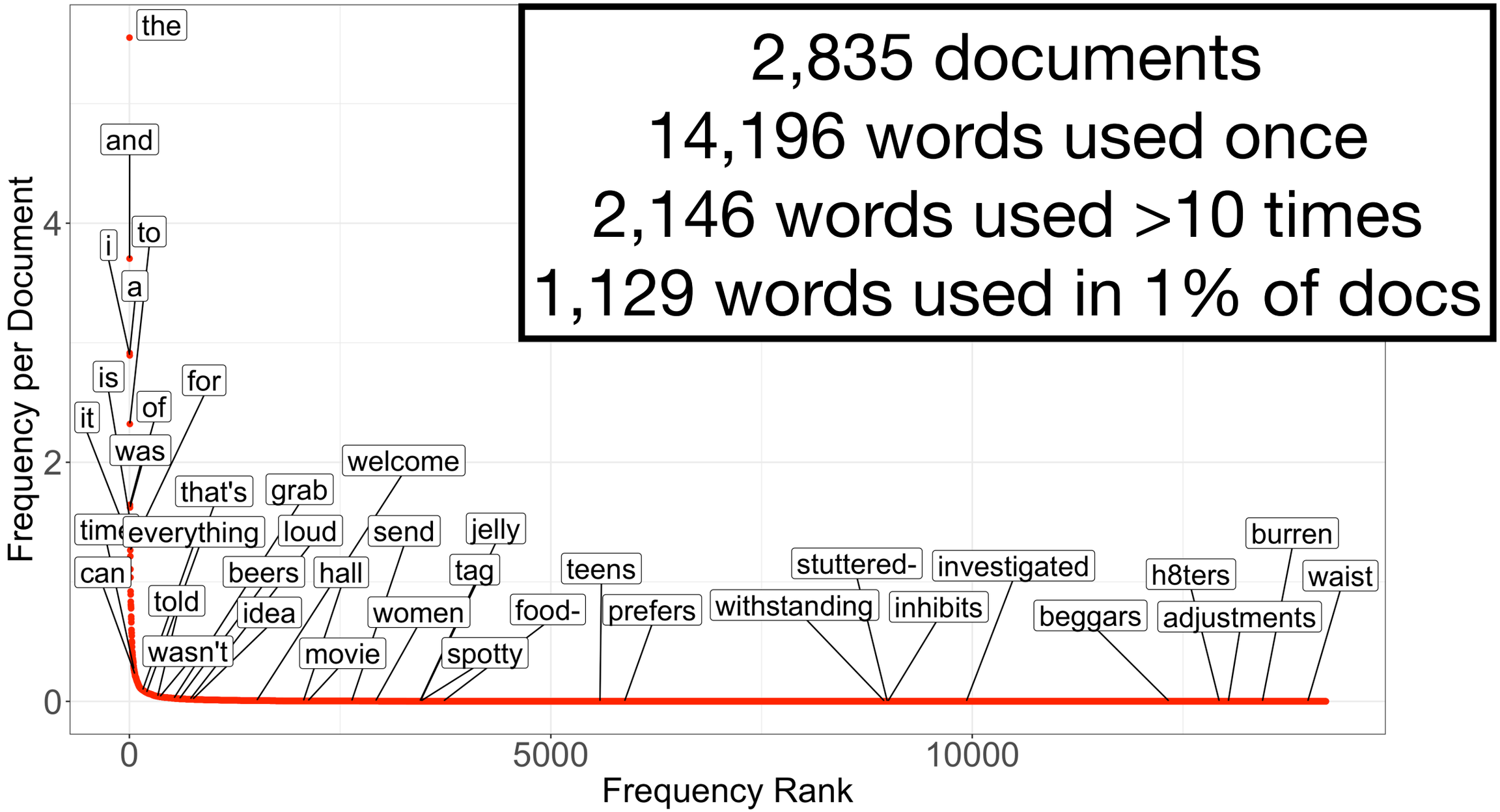
Pre-Processing the Bag

Filtering rare words



Pre-Processing the Bag

Filtering rare words



2,835 documents
14,196 words used once
2,146 words used >10 times
1,129 words used in 1% of docs

Pre-Processing the Bag

Punctuations

Capitalisations

Contractions

Stemming

Common word weighting

Constructing phrases (ngrams)

Rare word filtering

A Linguistic Model

$$\hat{y} = a_0 + e$$

A Linguistic Model

$$\hat{y} = a_0 + x_1 + x_2 + x_3 + \ldots + e$$

Feature Extraction

Select set of observables x

A Linguistic Model

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Feature Extraction

Select set of observables x

Feature Estimation

Determine **β** weights

Usually estimated empirically

Sometimes guesstimated (e.g. equal weights)

Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Determine **β** weights

Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Determine **β** weights

Easy case: supervised learning

Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Determine **β** weights

Easy case: supervised learning

Values of y are known for some documents

Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Determine **β** weights

Easy case: supervised learning

Values of y are known for some documents

List x variables is determined by preprocessing

Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Determine **β** weights

Easy case: supervised learning

Values of y are known for some documents

List x variables is determined by preprocessing

β can be estimated using algorithm

Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Standard approach: linear regression (OLS)

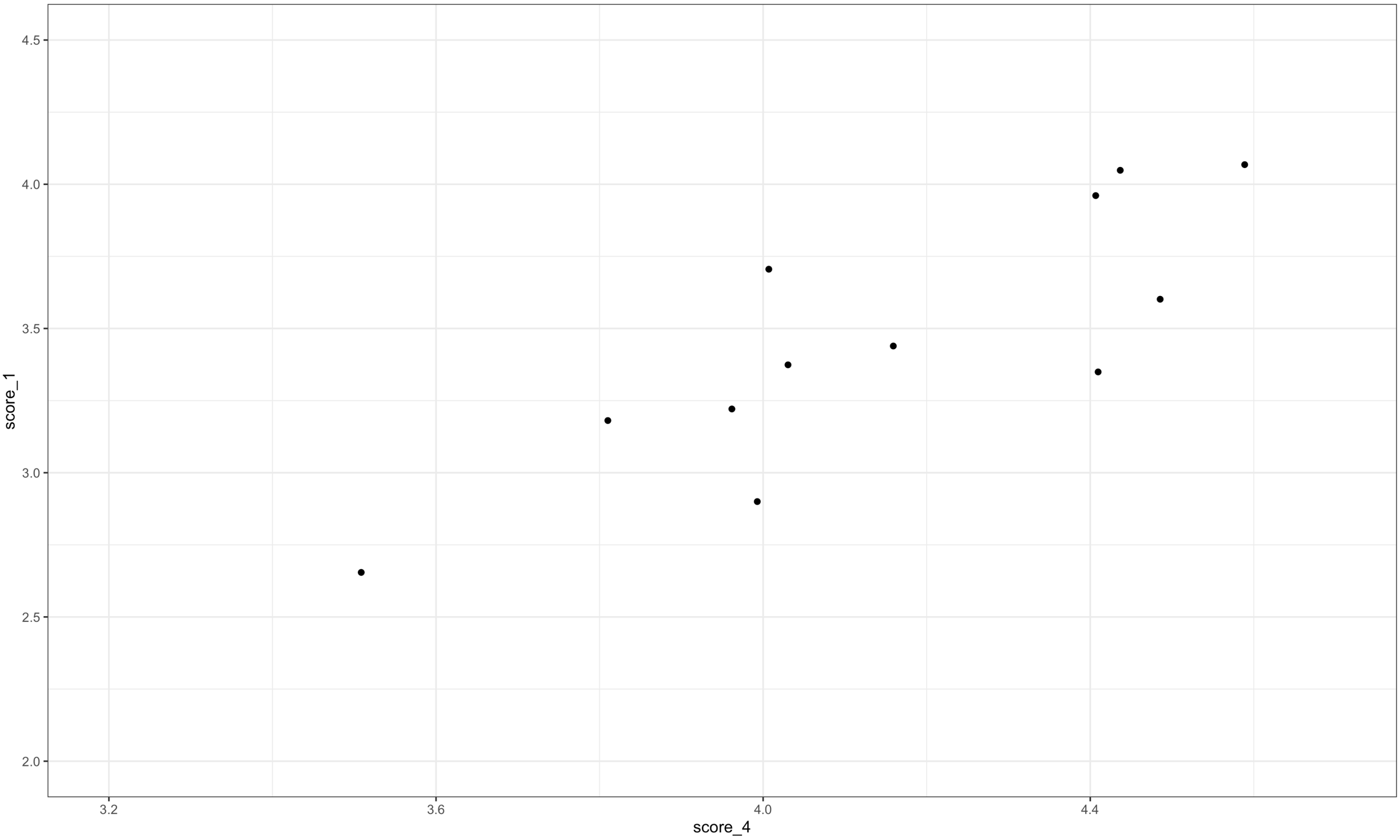
Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

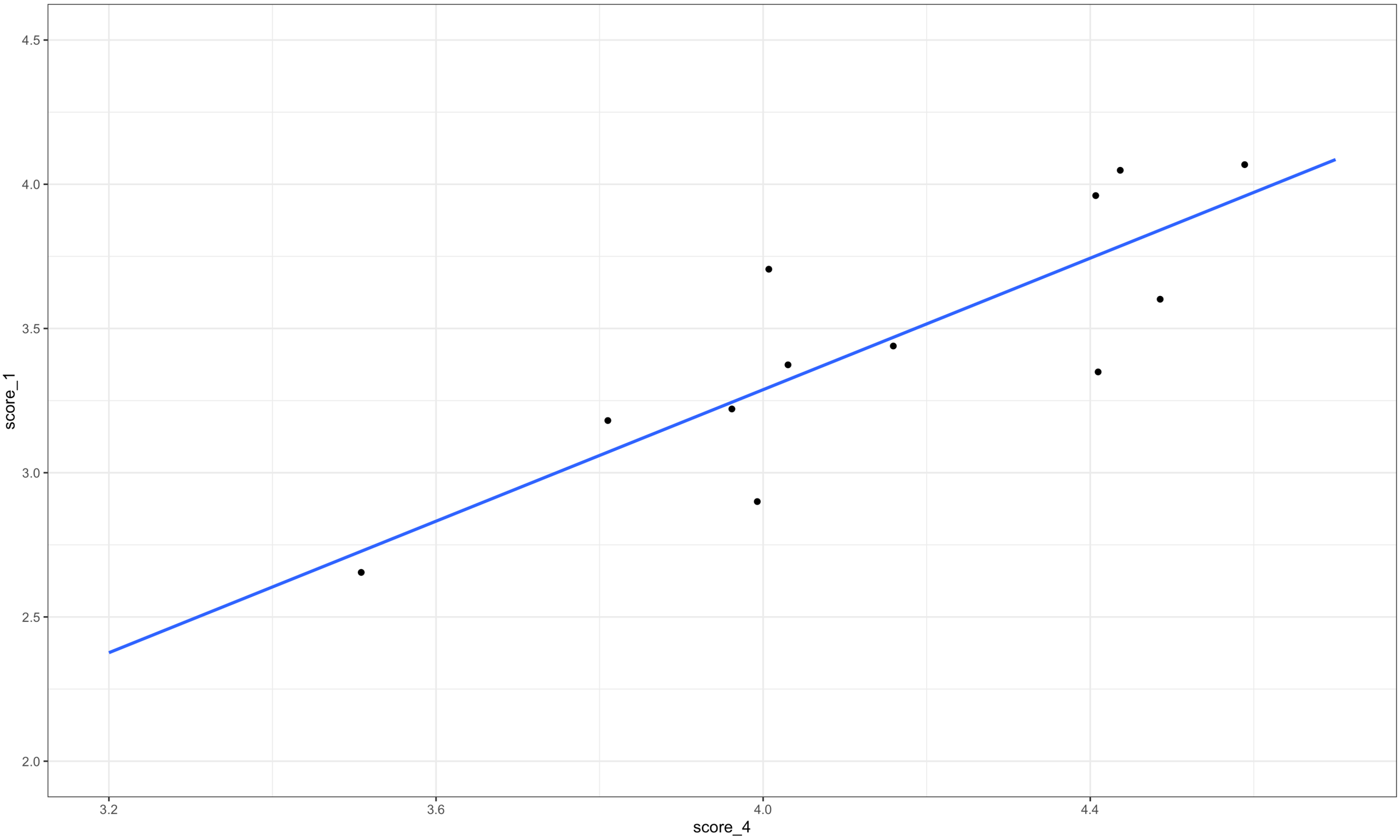
Standard approach: linear regression (OLS)

Why? BLUE - Best Linear Unbiased Estimator

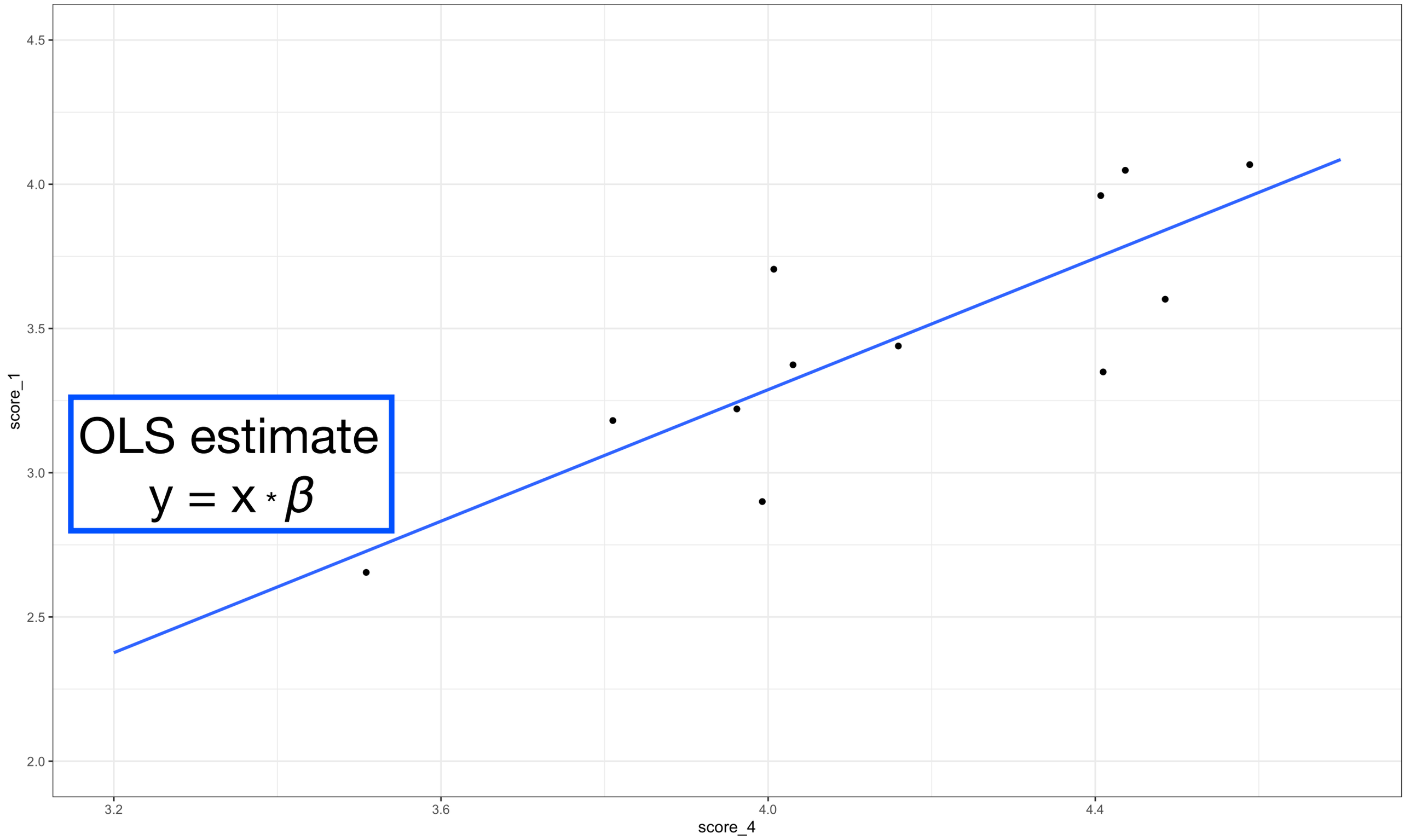
OLS Algorithm



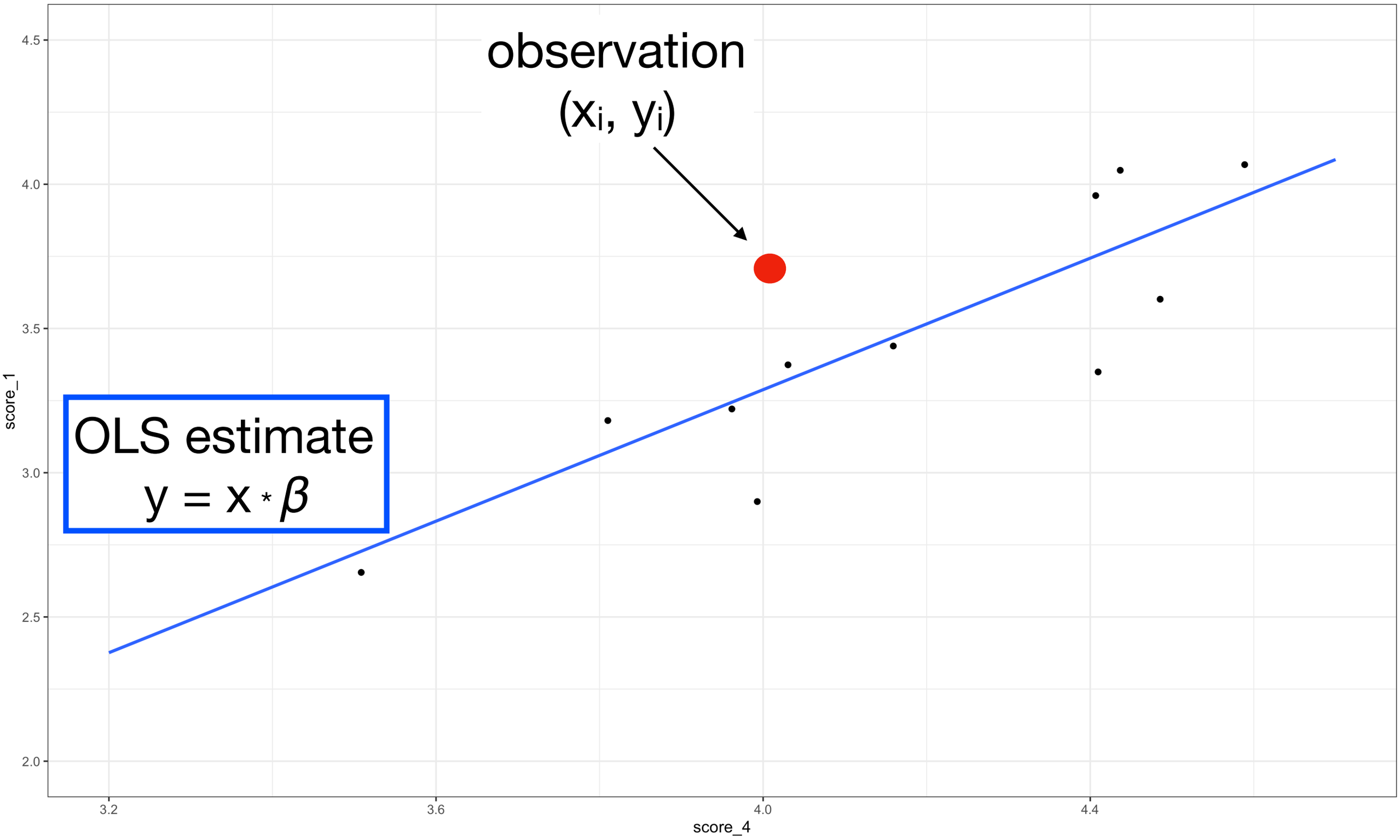
OLS Algorithm



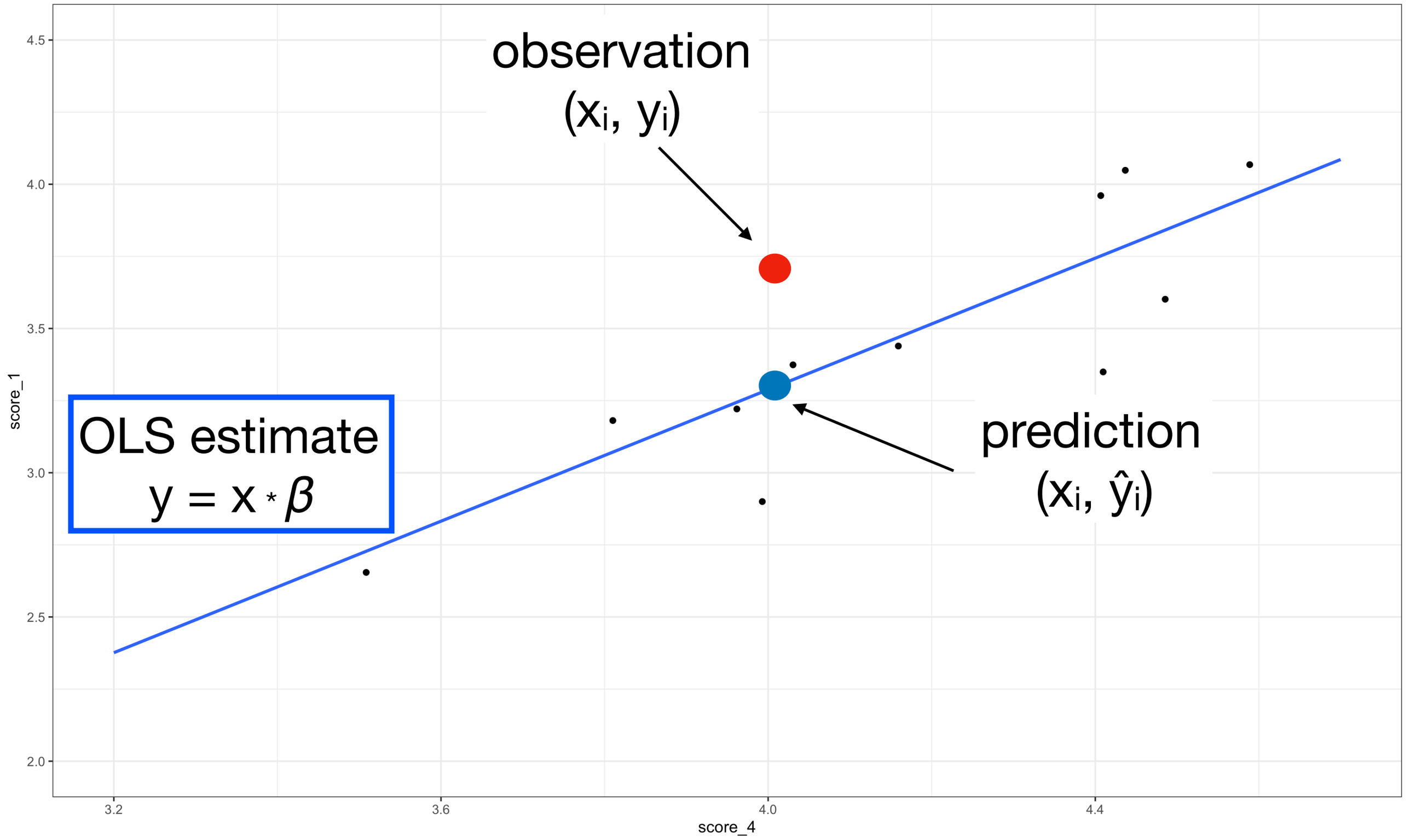
OLS Algorithm



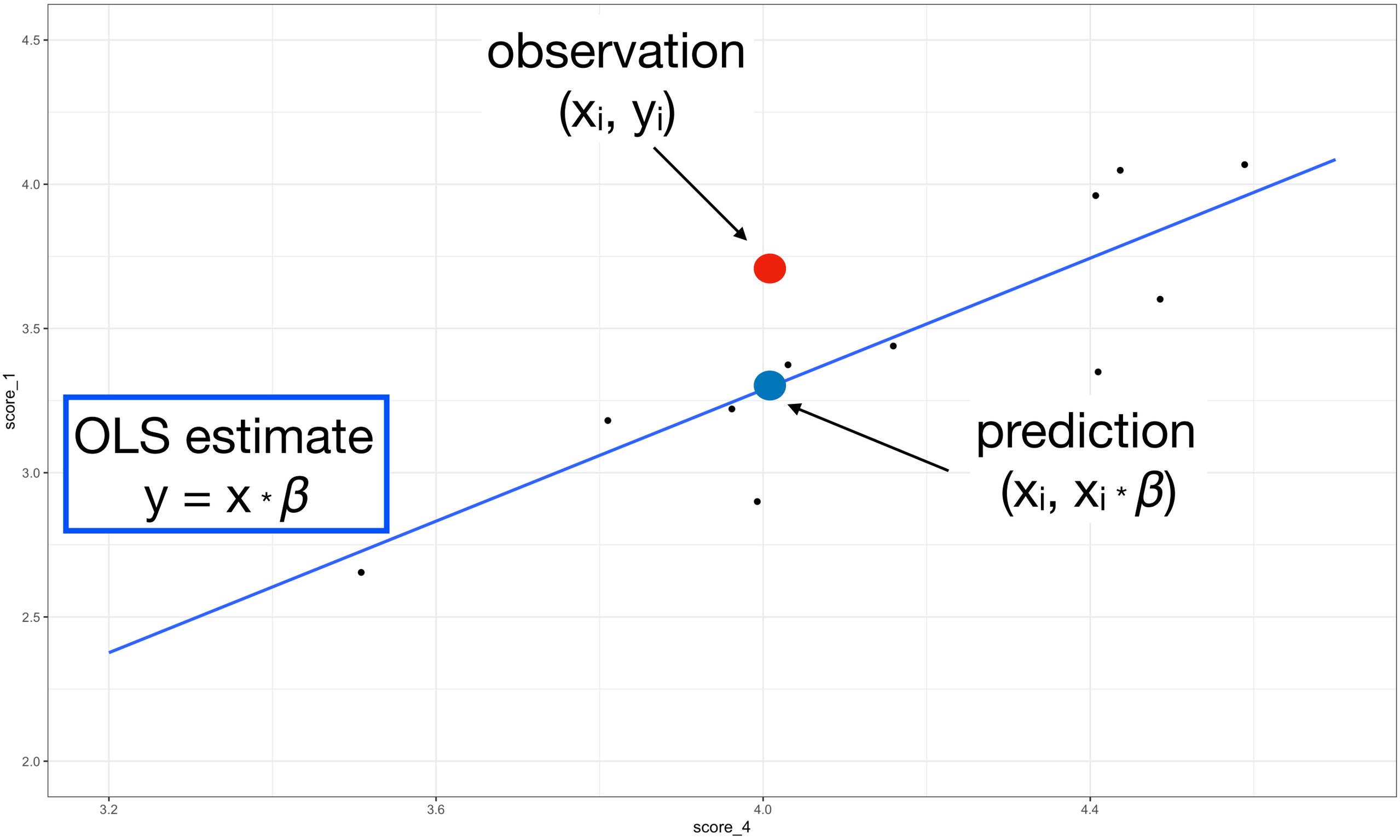
OLS Algorithm



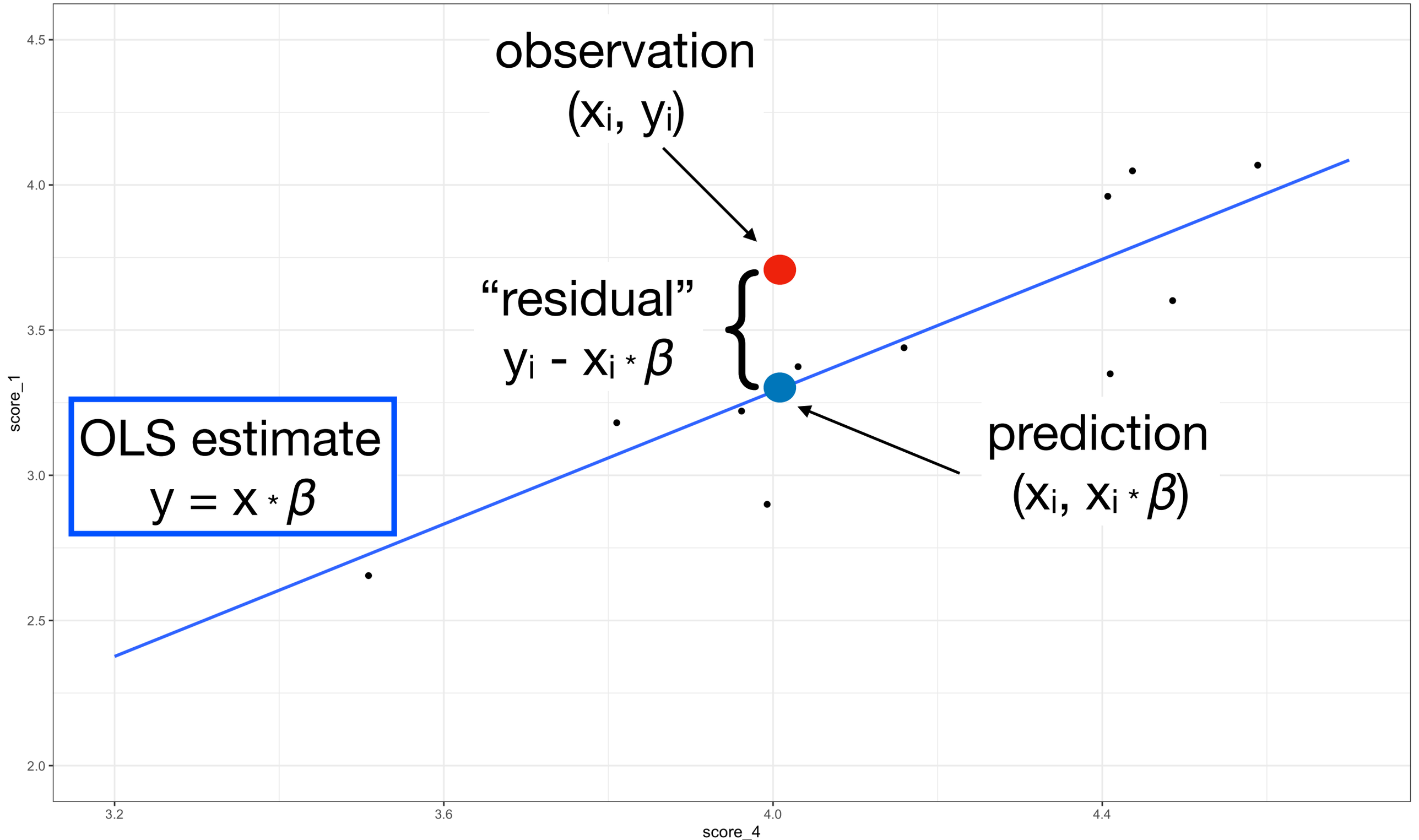
OLS Algorithm



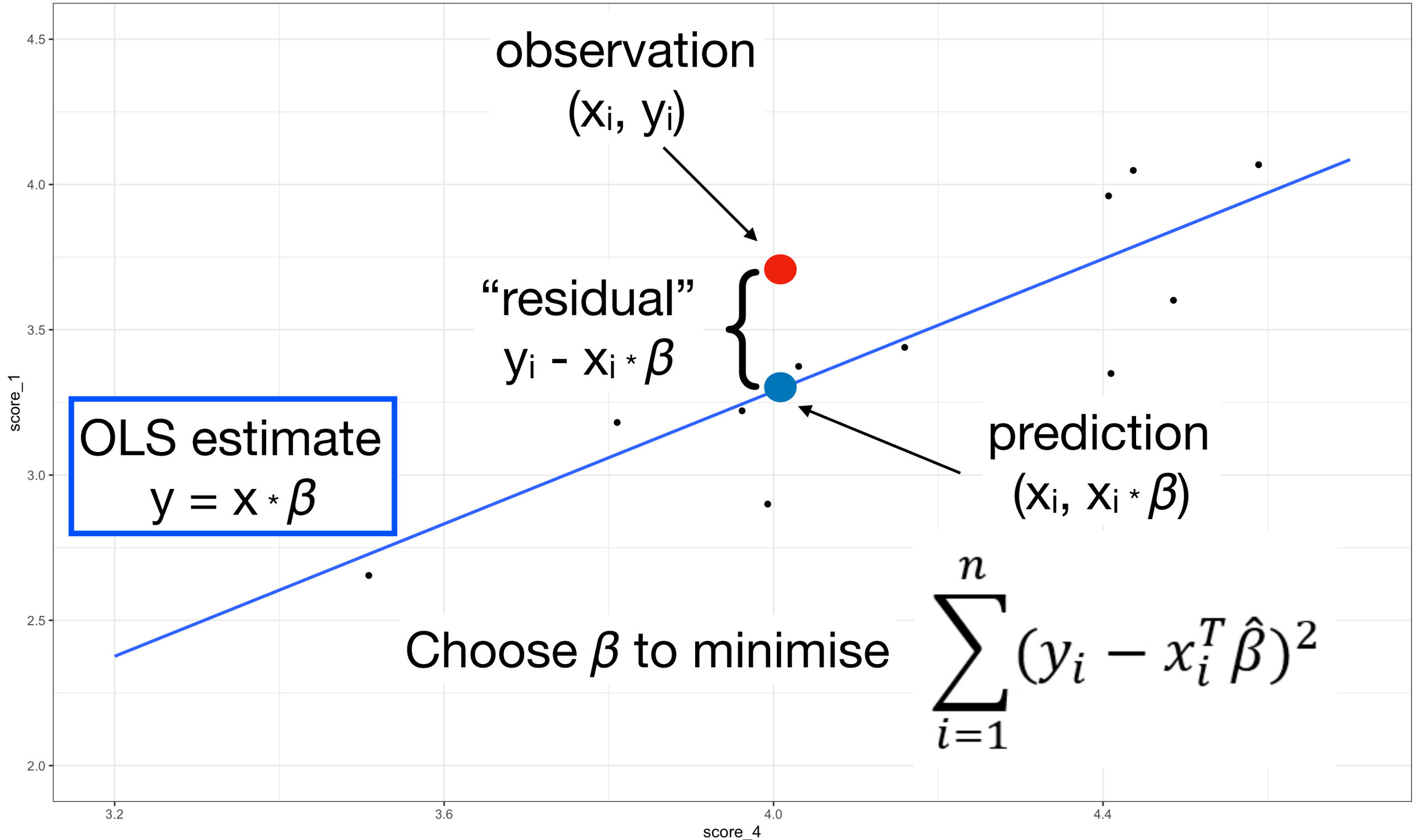
OLS Algorithm



OLS Algorithm



OLS Algorithm



Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Standard approach: linear regression (OLS)

Why? BLUE - Best Linear Unbiased Estimator

Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Standard approach: linear regression (OLS)

Why? BLUE - Best Linear Unbiased Estimator

Problems - truth is neither linear, nor independent

Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Standard approach: linear regression (OLS)

Why? BLUE - Best Linear Unbiased Estimator

Problems - truth is neither linear, nor independent

- world is high-dimensional (more x than n!)

Feature Estimation

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

Standard approach: linear regression (OLS)

Why? BLUE - Best Linear Unbiased Estimator

Problems - truth is neither linear, nor independent

- world is high-dimensional (more x than n!)

Implies -> easy to overfit model

What is the objective?

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

$\hat{y}$	β
Prediction Utility	**Intervention Utility**
What y should I expect?	*If I change x, what will y do?*

What is the objective?

$$\hat{y} = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

$\hat{y}$

Prediction Utility

What y should I expect?

$\hat{\beta}$

Intervention Utility

If I change x, what will y do?

Building a Prediction Model

Key insight: "regularization"

Building a Prediction Model

Key insight: "regularization"

- Prioritise the best y at the expense of β
- Overfitting from large β -> estimate smaller β

Building a Prediction Model

Key insight: "regularization"

- Prioritise the best y at the expense of β
- Overfitting from large β -> estimate smaller β

Workhorse Model: the LASSO

Least Angle Shrinkage & Selection Operator

(see Tibshirani, 1996; Friedman, Hastie & Tibshirani, 2010)

Building a Prediction Model

Key insight: "regularization"

- Prioritise the best y at the expense of β
- Overfitting from large β -> estimate smaller β

Workhorse Model: the LASSO

Least Angle Shrinkage & Selection Operator

(see Tibshirani, 1996; Friedman, Hastie & Tibshirani, 2010)

LASSO Regression

Choose β to minimise $\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|$

LASSO Regression

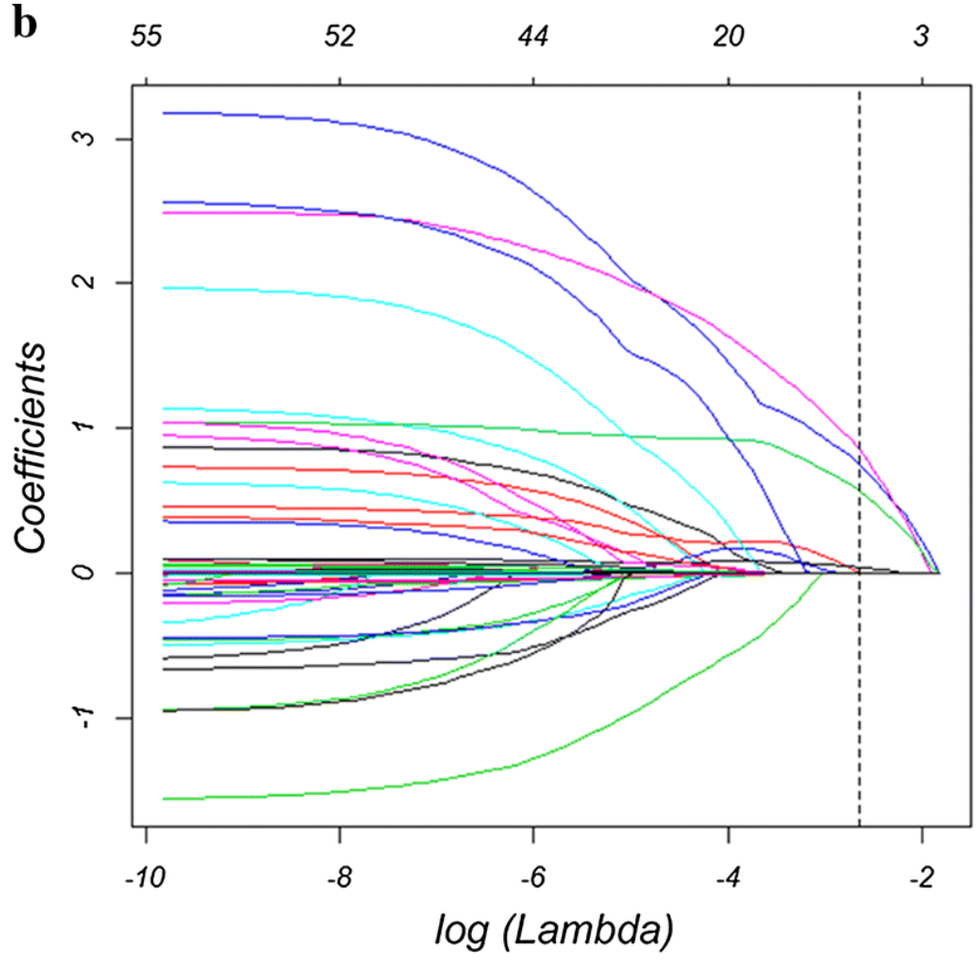
Choose β to minimise $\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|$

LASSO Regression

Choose β to minimise $\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|$

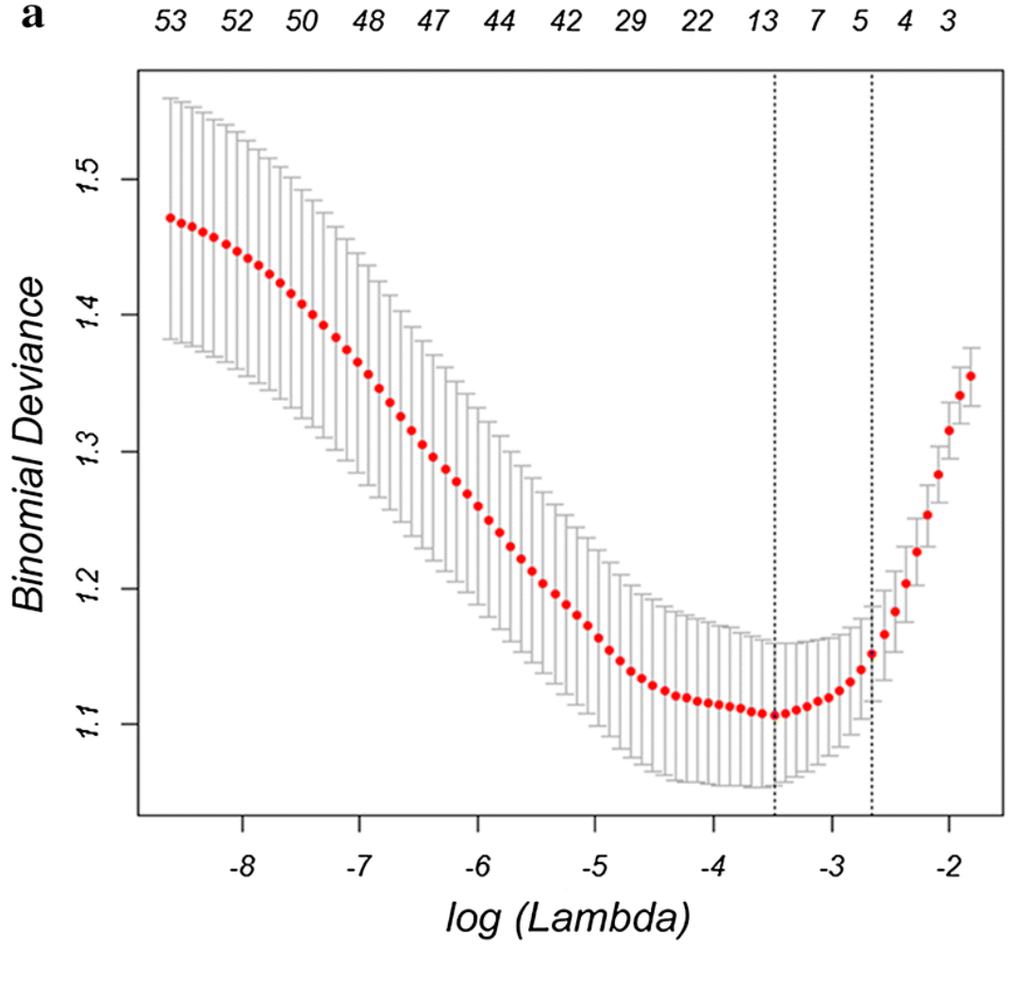
LASSO Regression

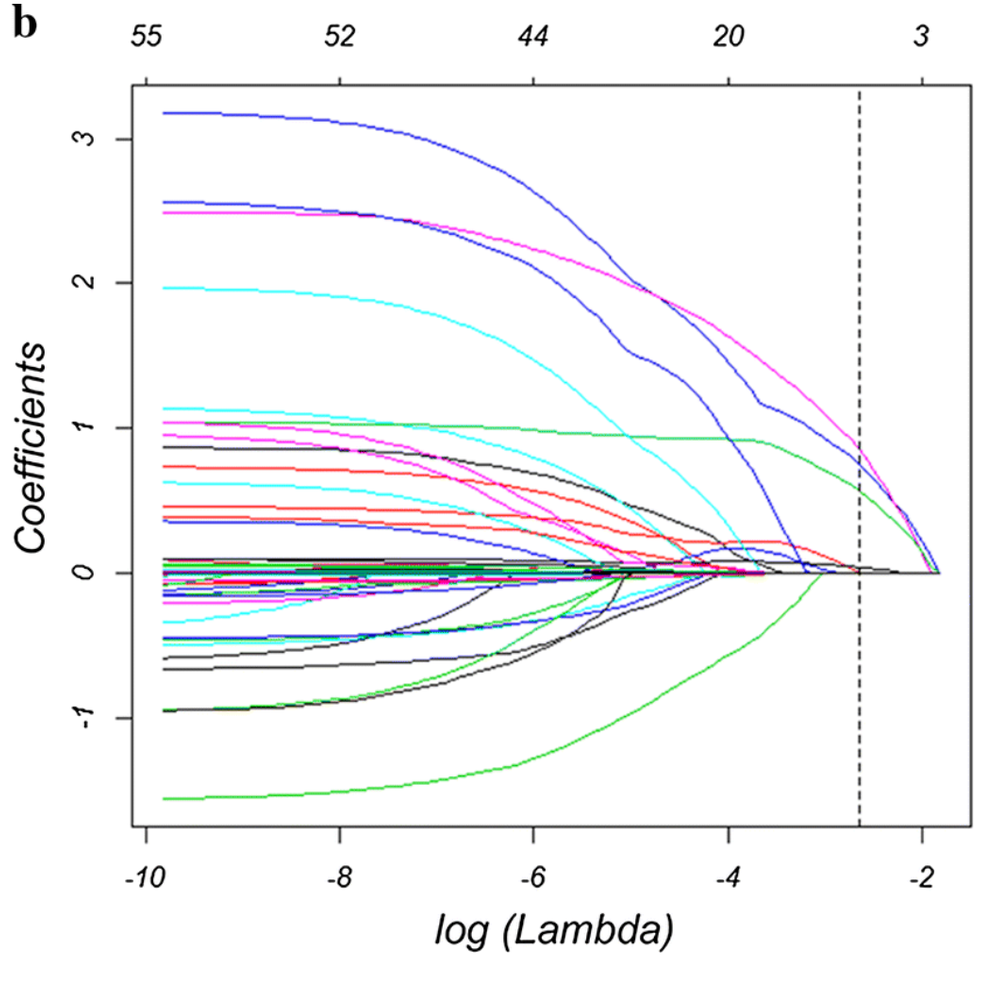
Choose β to minimise $\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda \sum_{j=1}^{m}|\hat{\beta}_j|$



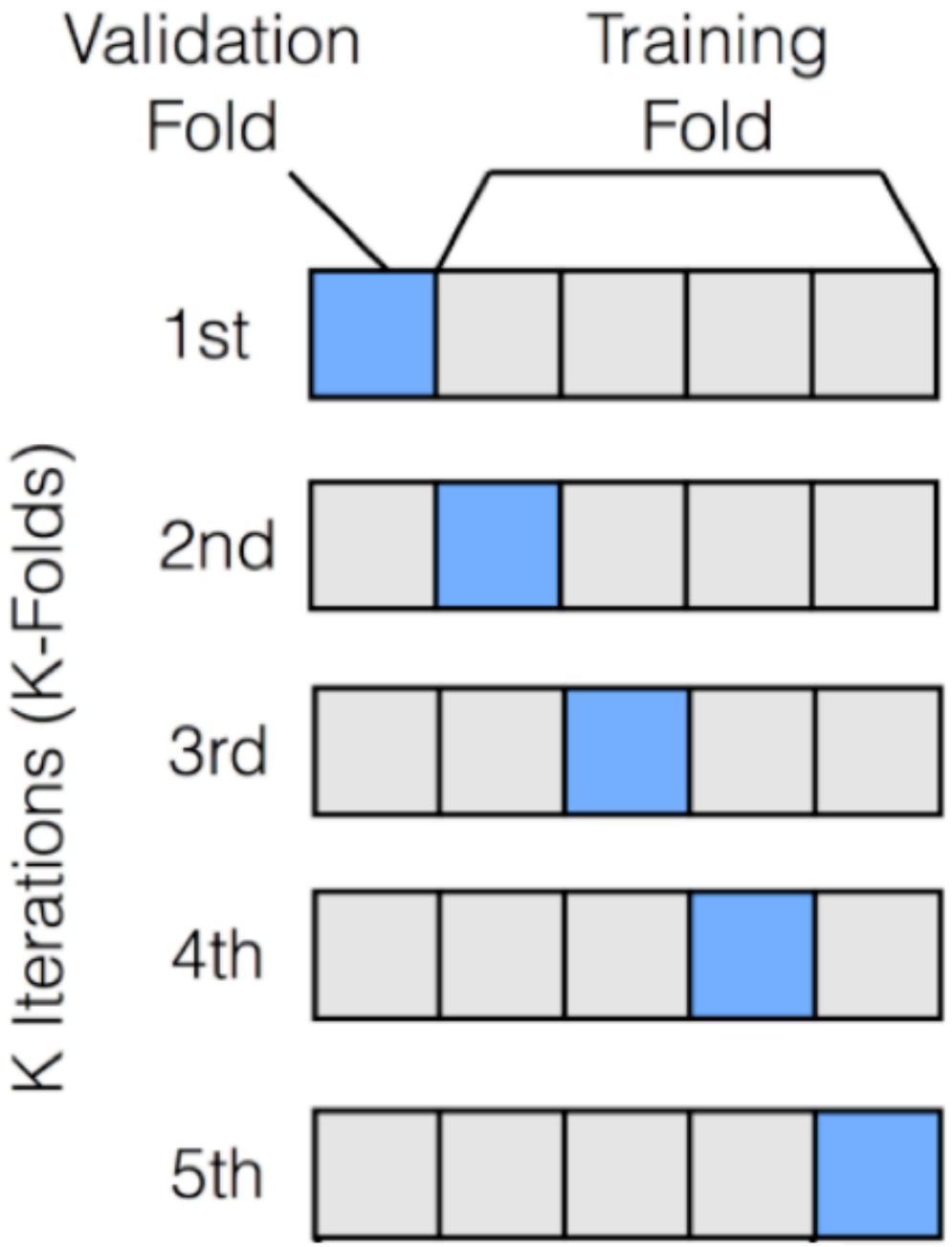
LASSO Regression

Choose β to minimise $\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda \sum_{j=1}^{m}|\hat{\beta}_j|$

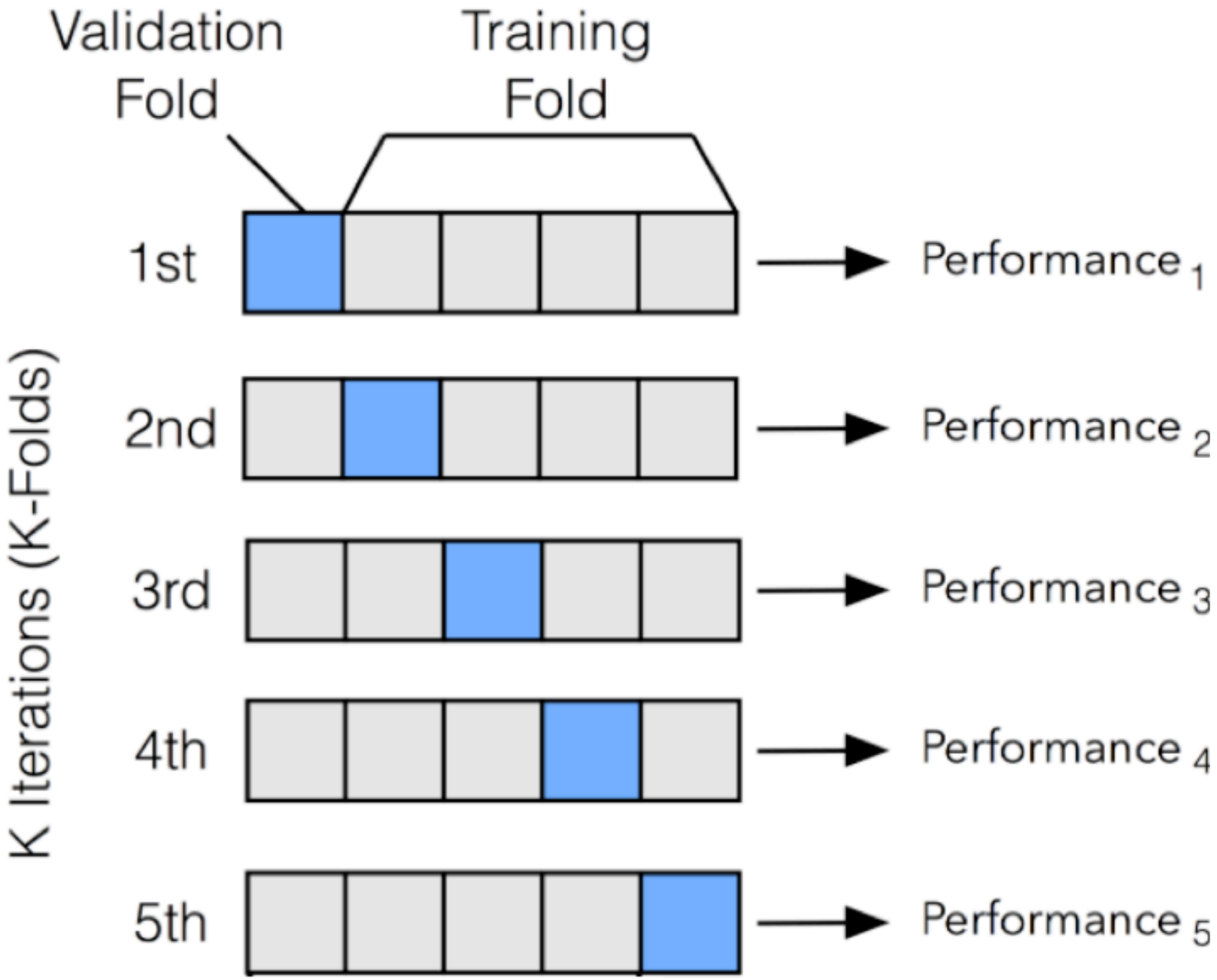




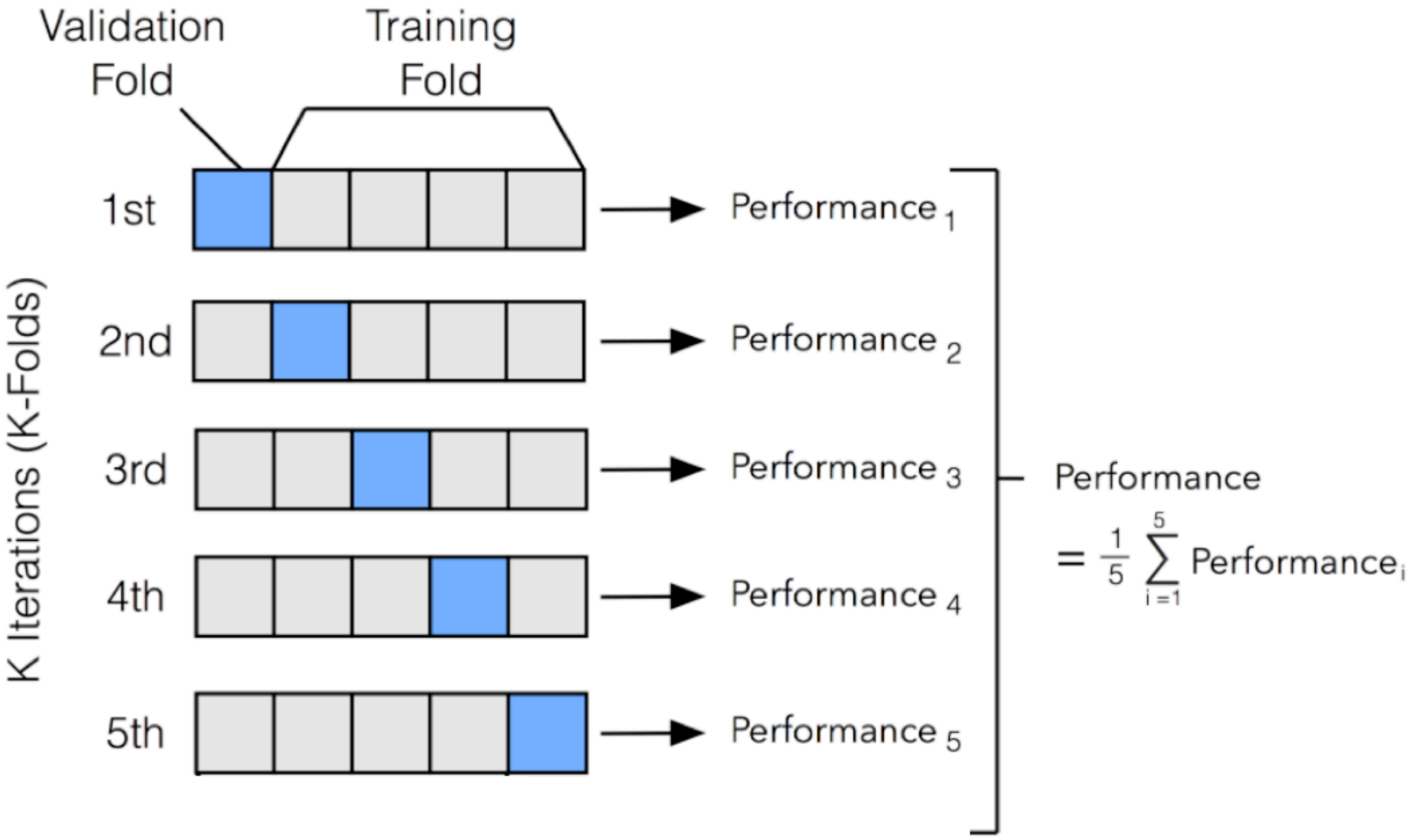
Cross-Validation



Cross-Validation



Cross-Validation

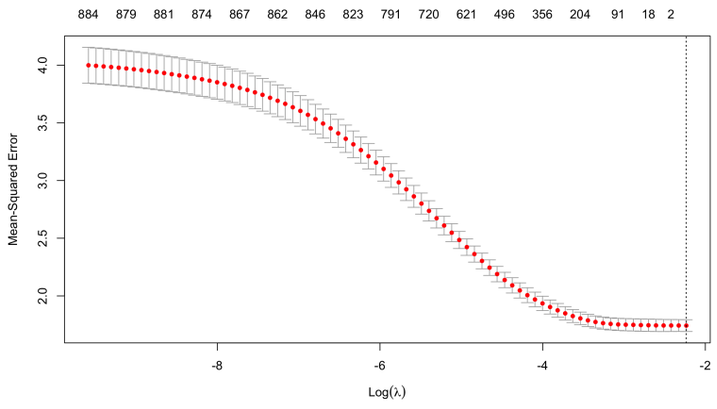


LASSO Regression

Rules of thumb for model evaluation

LASSO Regression

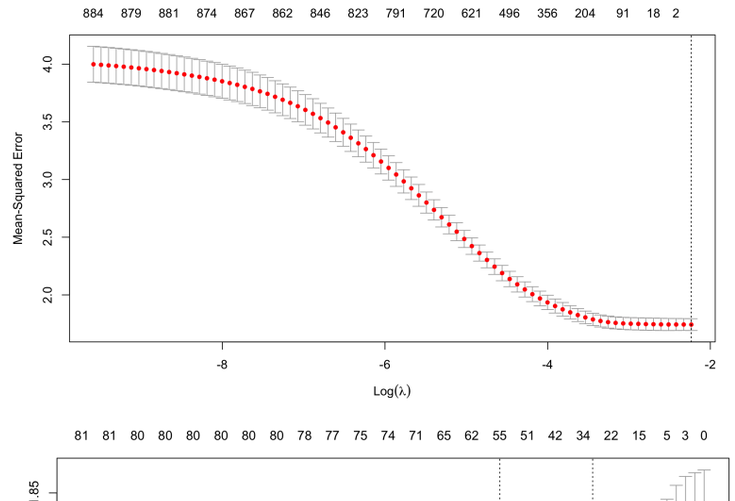
Rules of thumb for model evaluation



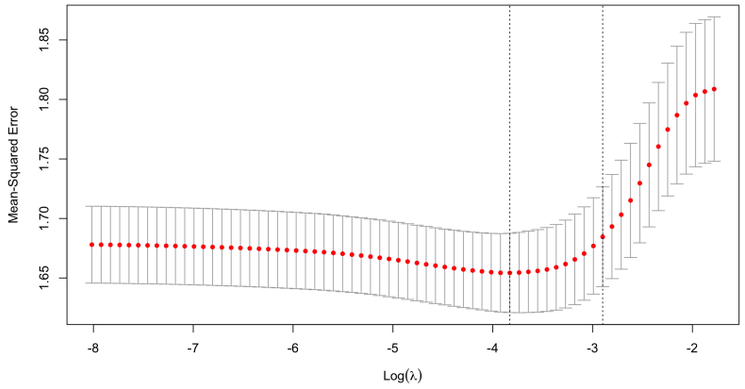
None of your features work, start over

LASSO Regression

Rules of thumb for model evaluation



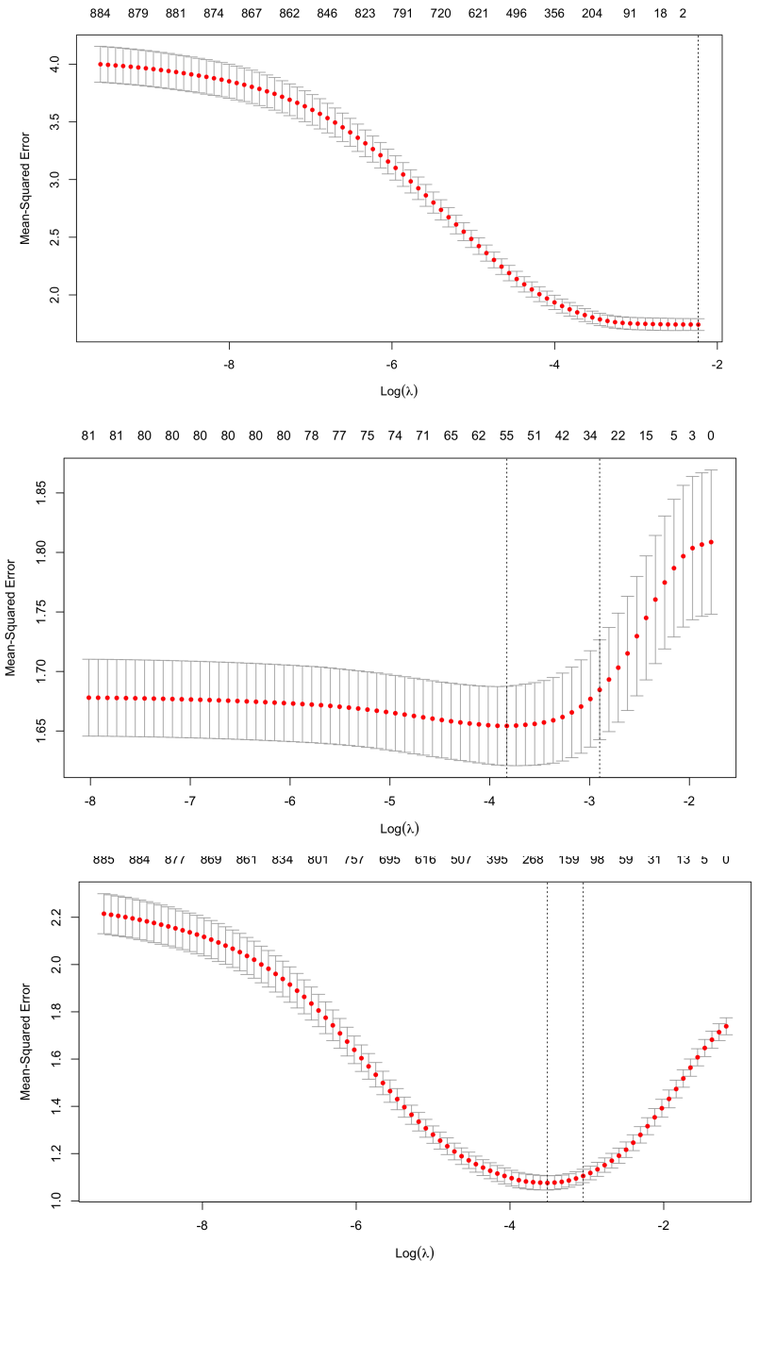
None of your features work, start over



All of your features work, add more

LASSO Regression

Rules of thumb for model evaluation



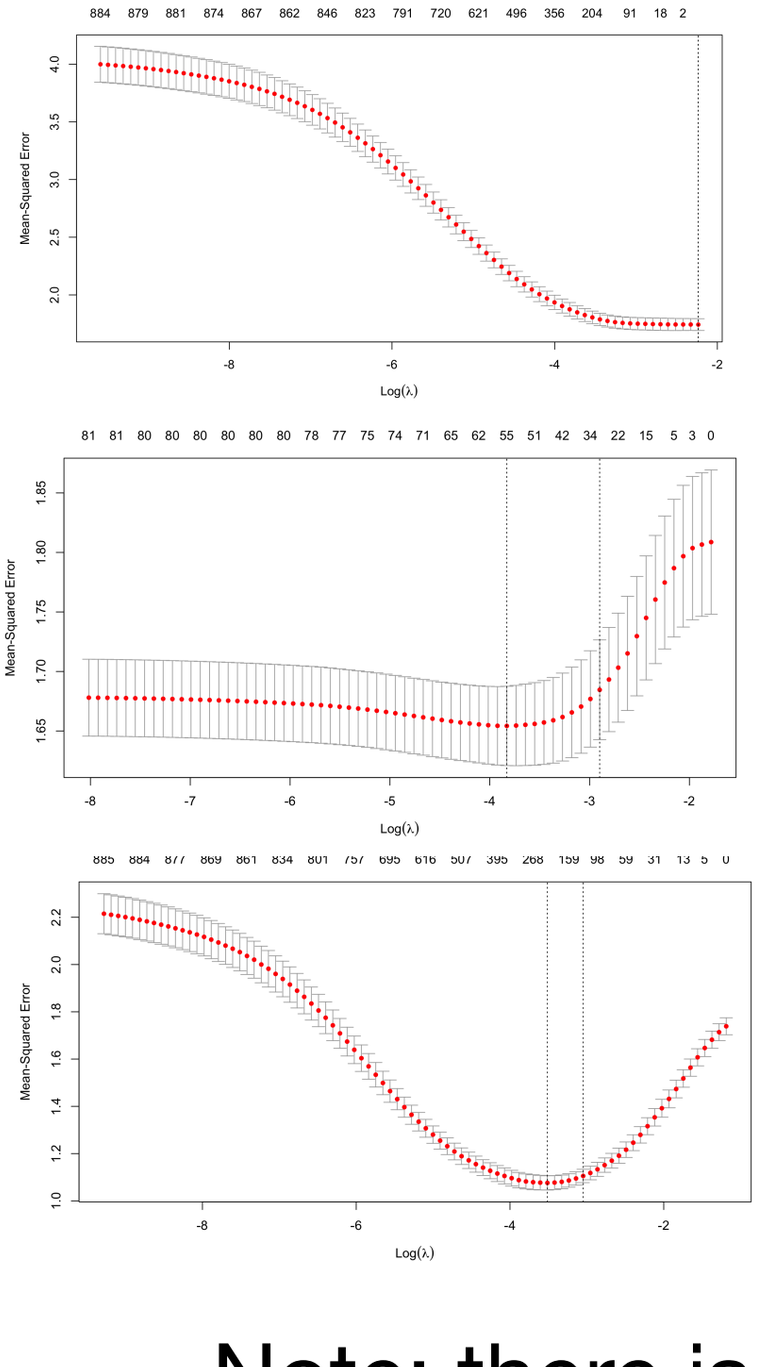
None of your features work, start over

All of your features work, add more

Some of your features work - good

LASSO Regression

Rules of thumb for model evaluation



None of your features work, start over

All of your features work, add more

Some of your features work - good

Note: there is always room for better features!

Why LASSO?

Why LASSO?

Easy to interpret

Why LASSO?

Easy to interpret

“Bet on sparsity” - if only a few independent features matter, LASSO will find them

Why LASSO?

Easy to interpret

“Bet on sparsity” - if only a few independent features matter,
LASSO will find them
if many and/or collinear features matter,
no algorithm will figure it out

Why LASSO?

Easy to interpret

“Bet on sparsity” - if only a few independent features matter,
LASSO will find them
if many and/or collinear features matter,
no algorithm will figure it out

What about SVMs, random forests, NNs, etc. ?

Why LASSO?

Easy to interpret

“Bet on sparsity” - if only a few independent features matter,
LASSO will find them
if many and/or collinear features matter,
no algorithm will figure it out

What about SVMs, random forests, NNs, etc. ?

Good for finding complex features (e.g. interactions)

Why LASSO?

Easy to interpret

“Bet on sparsity” - if only a few independent features matter,
LASSO will find them
if many and/or collinear features matter,
no algorithm will figure it out

What about SVMs, random forests, NNs, etc. ?

Good for finding complex features (e.g. interactions)
Also good at this - humans, theory

Why LASSO?

Easy to interpret

"Bet on sparsity" - if only a few independent features matter,
LASSO will find them
if many and/or collinear features matter,
no algorithm will figure it out

What about SVMs, random forests, NNs, etc. ?

Good for finding complex features (e.g. interactions)
Also good at this - humans, theory
Downsides - VERY data intensive
- results are hard to interpret

Evaluating Accuracy

How do we measure accuracy?

Evaluating Accuracy

How do we measure accuracy?

Depends on..

Distribution of predictions

Distribution of target variable

Relative value of mistakes

Evaluating Accuracy

Binary target, binary prediction

Evaluating Accuracy

Binary target, binary prediction

Text	Prediction	Label
This is great!	0	1
This is awesome!	1	1
I was thrilled	1	1
I loved it	0	1
This was terrible.	0	0
I hated it	0	0

Evaluating Accuracy

Binary target, binary prediction

Text	Prediction	Label	Accuracy
This is great!	0	1	0
This is awesome!	1	1	1
I was thrilled	1	1	1
I loved it	0	1	0
This was terrible.	0	0	1
I hated it	0	0	1

Evaluating Accuracy

Binary target, binary prediction

Text	Prediction	Label	Accuracy
This is great!	0	1	0
This is awesome!	1	1	1
I was thrilled	1	1	1
I loved it	0	1	0
This was terrible.	0	0	1
I hated it	0	0	1

Accuracy = 4/6 ~ 67%

Evaluating Accuracy

Binary target, binary prediction

Prediction	Label
0	1
1	1
1	1
0	1
0	0
0	0

Evaluating Accuracy

Binary target, binary prediction

Prediction	Label
0	1
1	1
1	1
0	1
0	0
0	0

“Confusion matrix”

Evaluating Accuracy

Binary target, binary prediction

Prediction	Label
0	1
1	1
1	1
0	1
0	0
0	0

"Confusion matrix"

	Label	
Prediction	0	1
0	2	2
1	0	2

Evaluating Accuracy

Binary target, binary prediction

Prediction	Label
0	1
1	1
1	1
0	1
0	0
0	0

"Confusion matrix"

	Label	
Prediction	0	1
0	2	2
1	0	2

Accuracy

= (2+2)/(2+2+0+2)

~ 67%

Evaluating Accuracy

Categorical target, categorical prediction

Evaluating Accuracy

Categorical target, categorical prediction

True Label

Prediction

	Mexican	Indian	Chinese
Mexican			
Indian			
Chinese			

Evaluating Accuracy

Categorical target, categorical prediction

True Label

Prediction

	Mexican	Indian	Chinese
Mexican	401	19	80
Indian	55	212	101
Chinese	41	27	367

Evaluating Accuracy

Categorical target, categorical prediction

True Label

Prediction

	Mexican	Indian	Chinese
Mexican	401	19	80
Indian	55	212	101
Chinese	41	27	367

(401 + 212 + 367) /
(401 + 19 + 80 + 55 + 212 + 101 + 41 + 27 + 367) = 75%

Evaluating Accuracy

Binary target, continuous prediction

Evaluating Accuracy

Binary target, continuous prediction

Text	Prediction	Label
This is great!	0.23	1
This is awesome!	0.68	1
I was thrilled	0.84	1
I loved it	0.13	1
This was terrible.	0.17	0
I hated it	0.49	0

Evaluating Accuracy

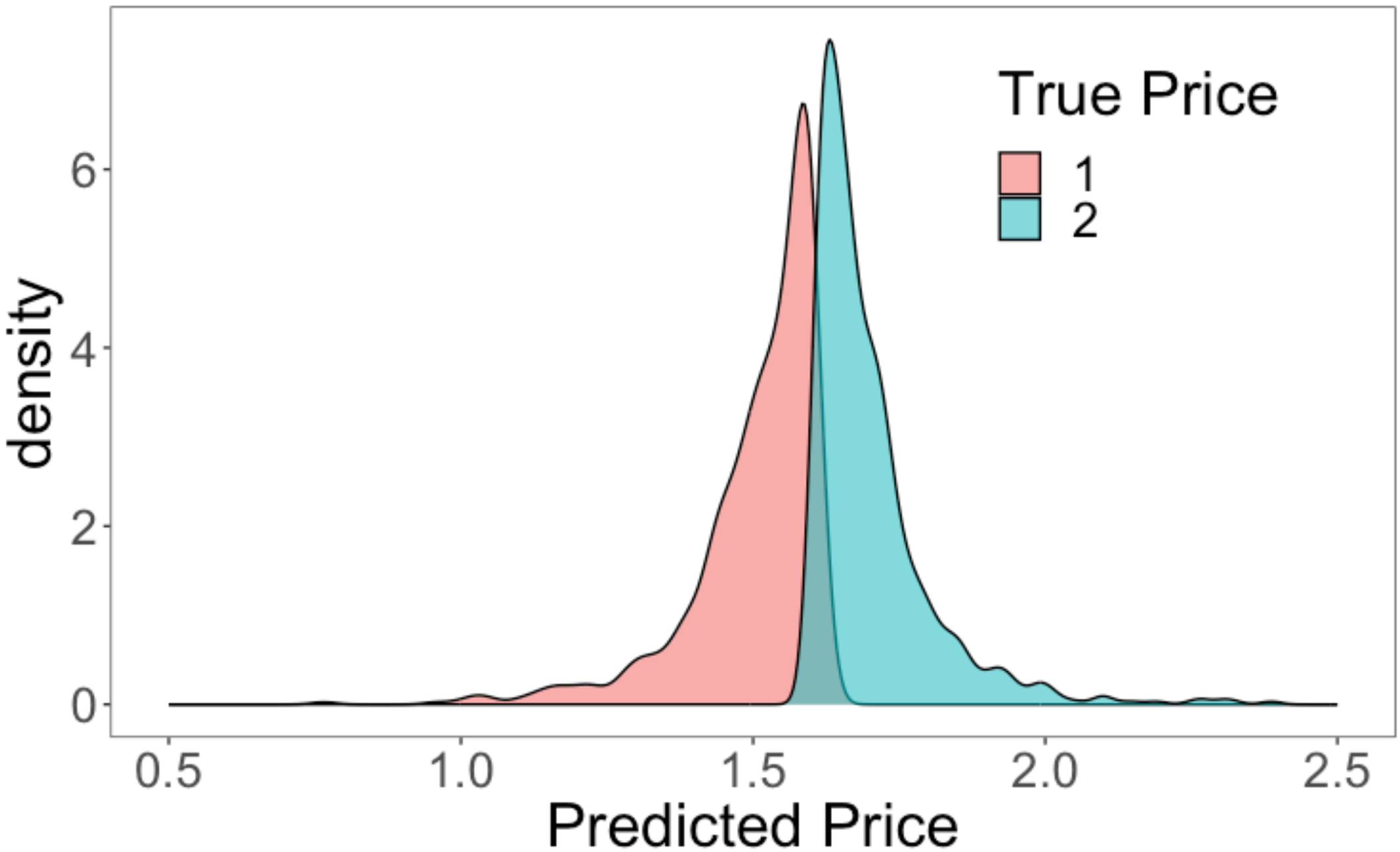
Binary target, continuous prediction

Text	Prediction	Label
This is great!	0.23	1
This is awesome!	0.68	1
I was thrilled	0.84	1
I loved it	0.13	1
This was terrible.	0.17	0
I hated it	0.49	0

One approach - choose a threshold to split predictions

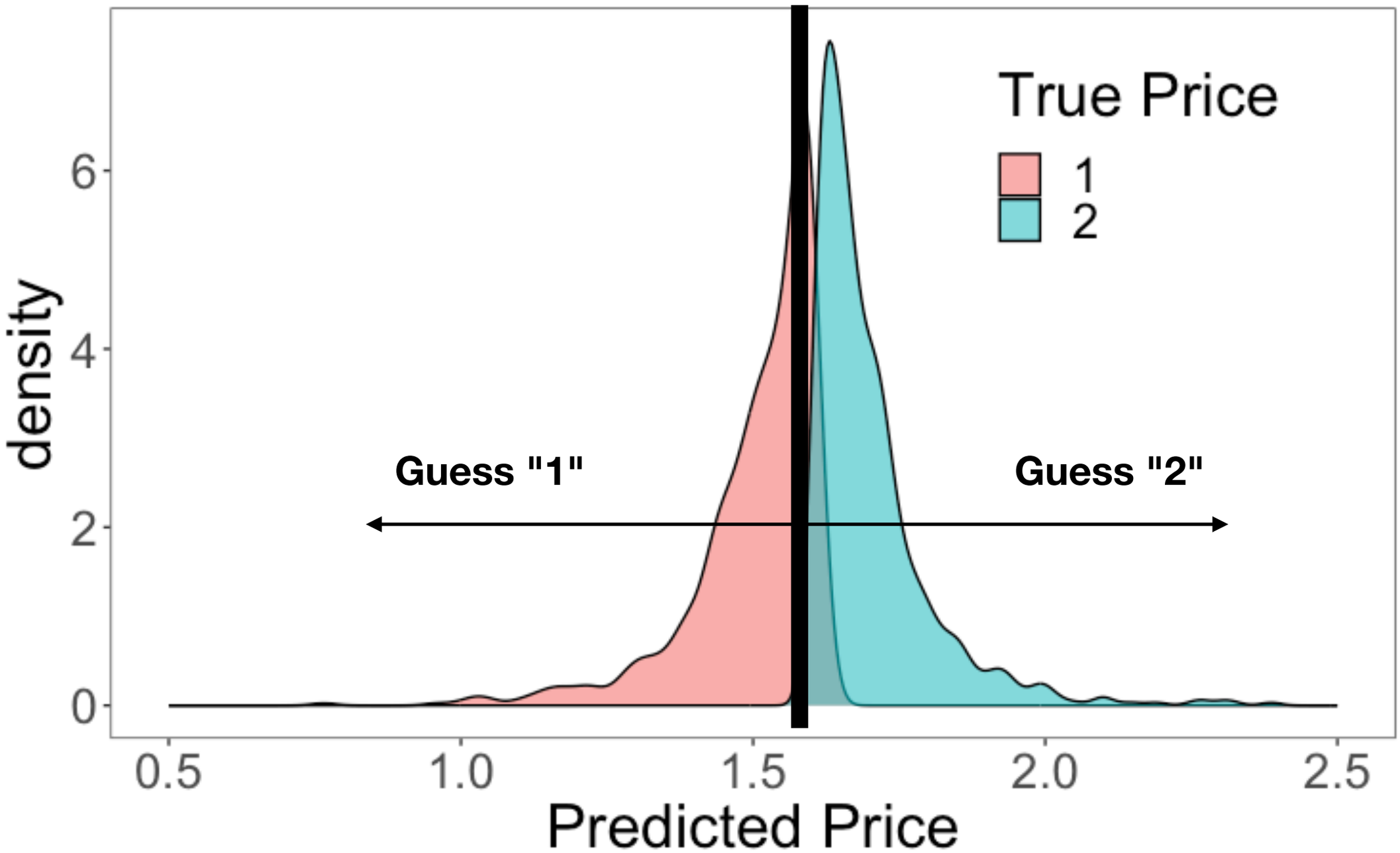
Evaluating Accuracy

Binary target, continuous prediction



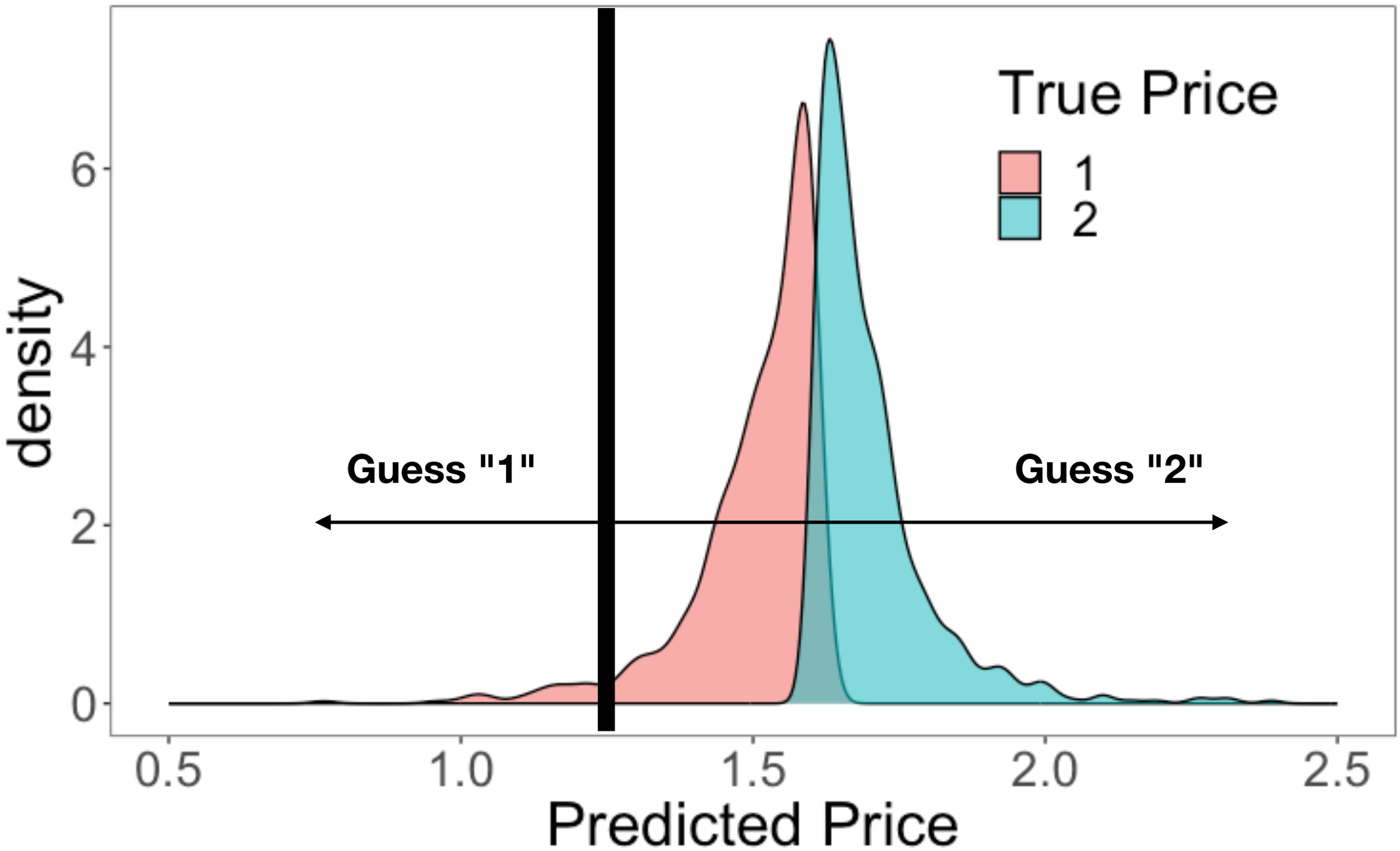
Evaluating Accuracy

Binary target, continuous prediction



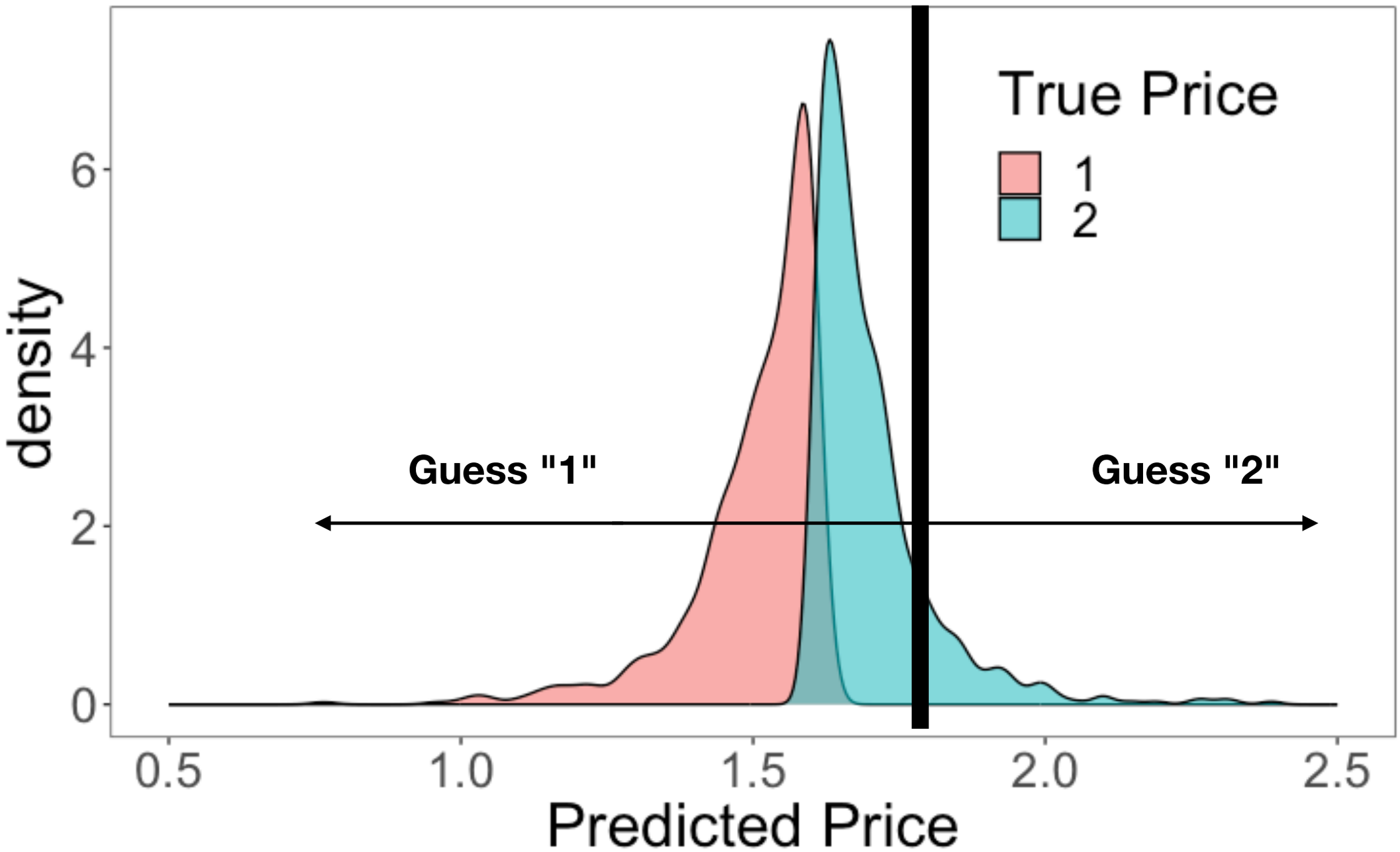
Evaluating Accuracy

Binary target, continuous prediction



Evaluating Accuracy

Binary target, continuous prediction



Evaluating Accuracy

Binary target, continuous prediction

Text	Prediction	Label
This is great!	0.23	1
This is awesome!	0.68	1
I was thrilled	0.84	1
I loved it	0.13	1
This was terrible.	0.17	0
I hated it	0.49	0

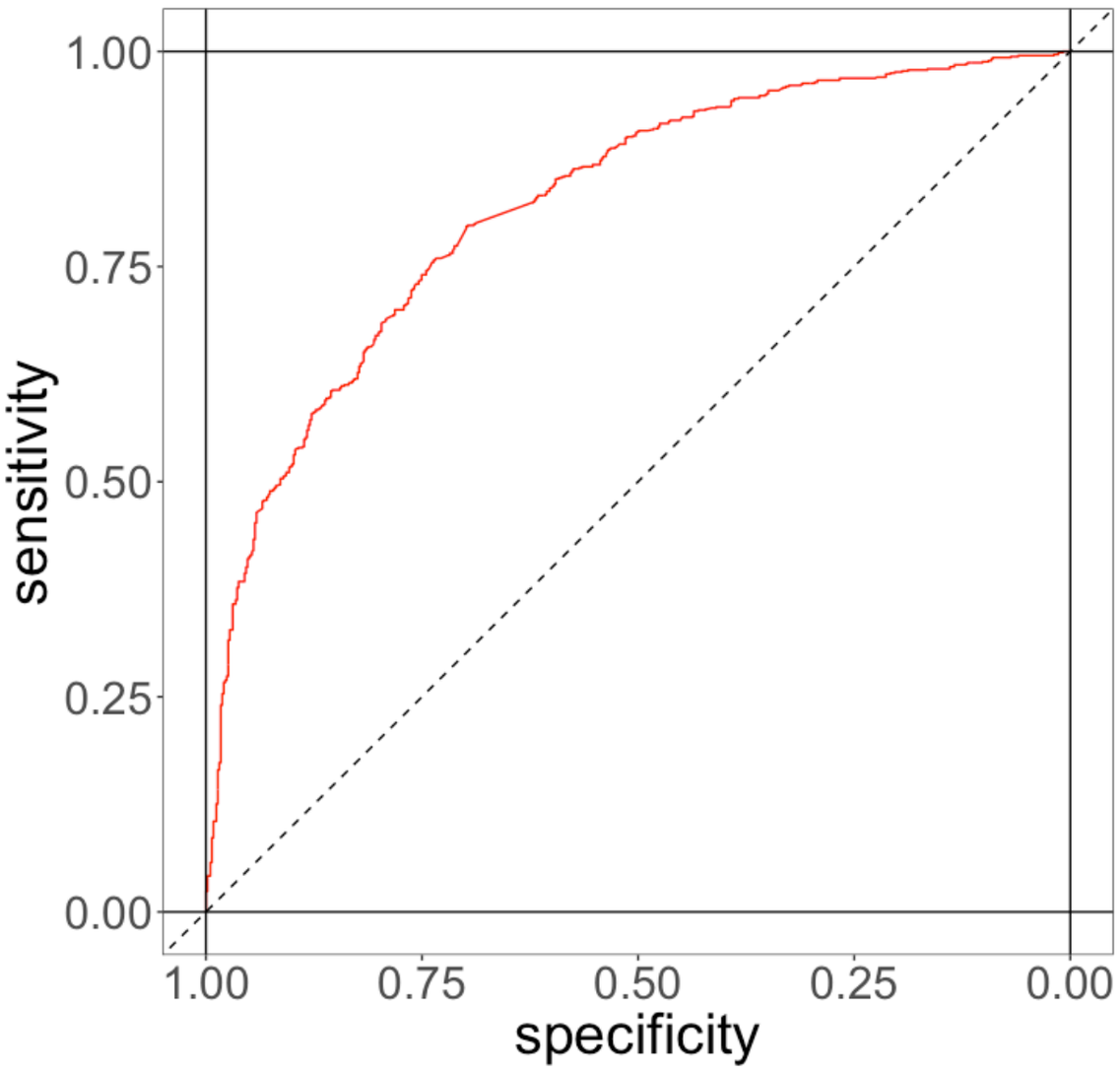
One approach - choose a threshold to split predictions

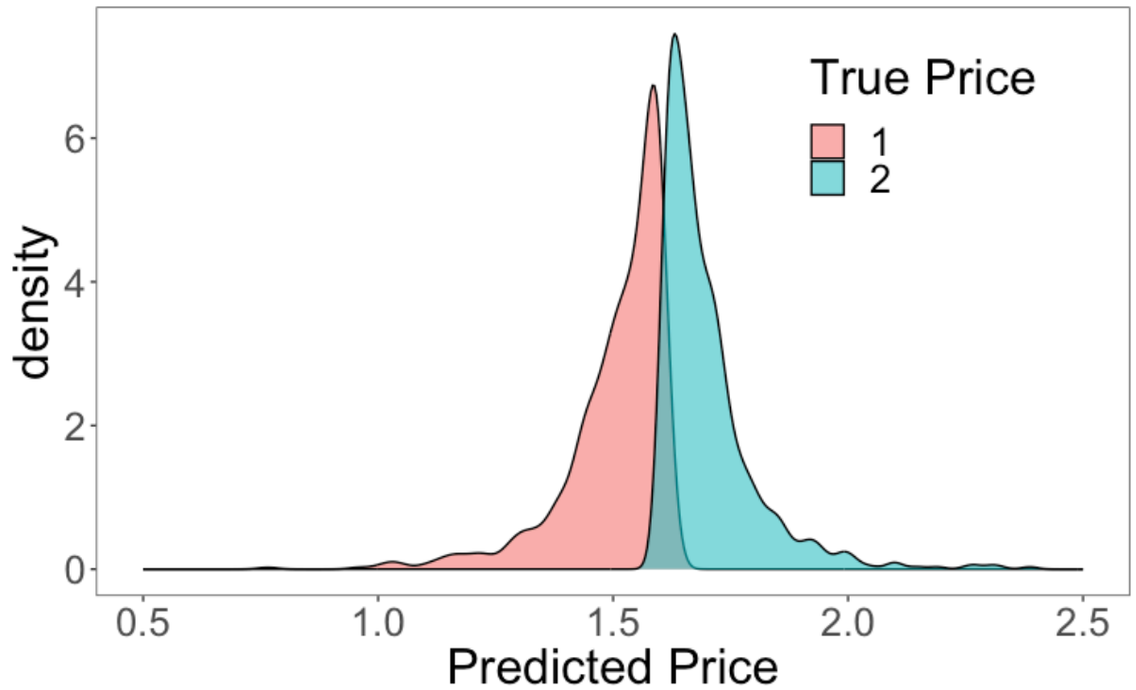
More commonly - test across all possible thresholds

Evaluating Accuracy

Binary target, continuous prediction

Receiver Operator Characteristic (ROC) curve

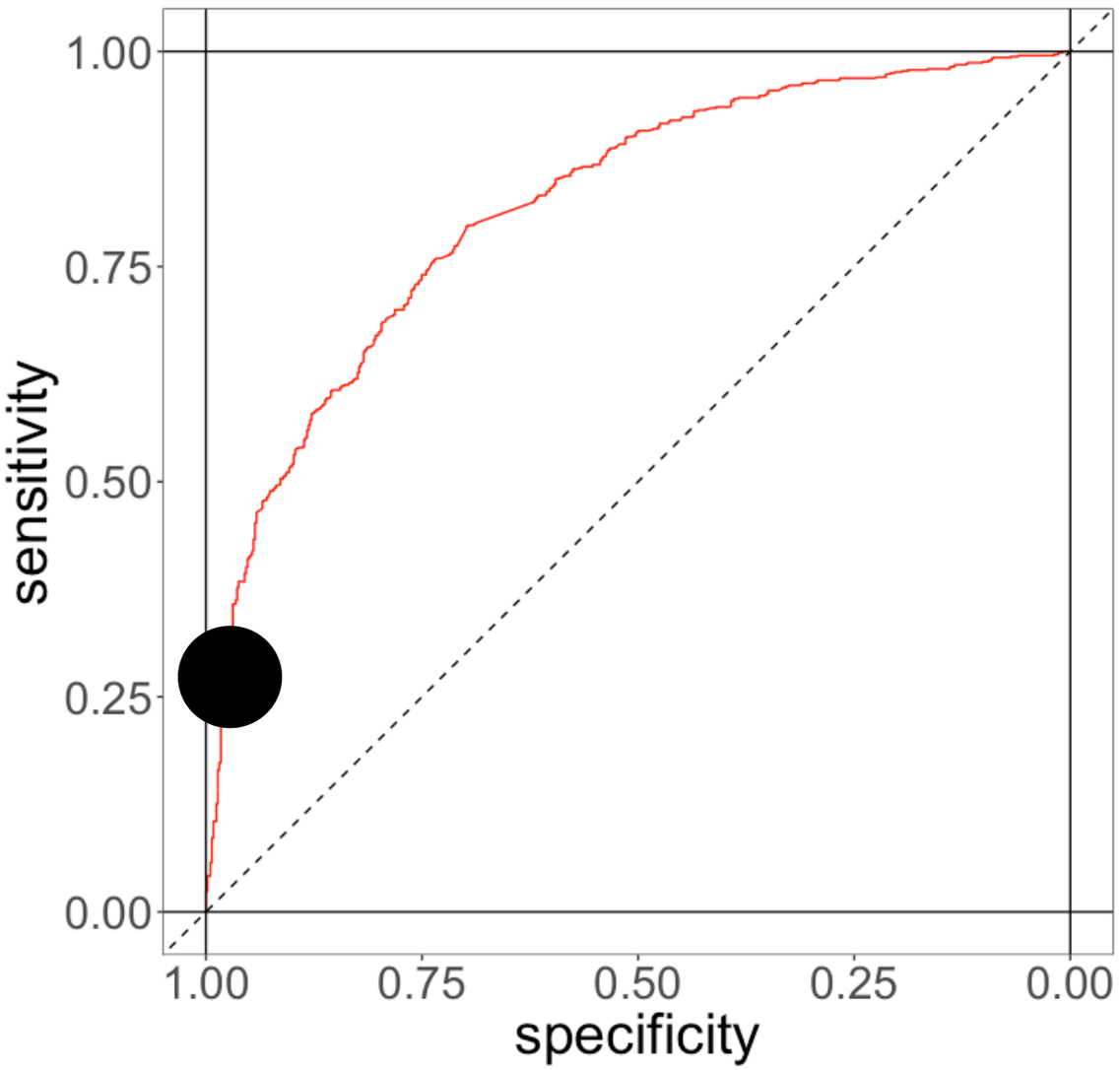


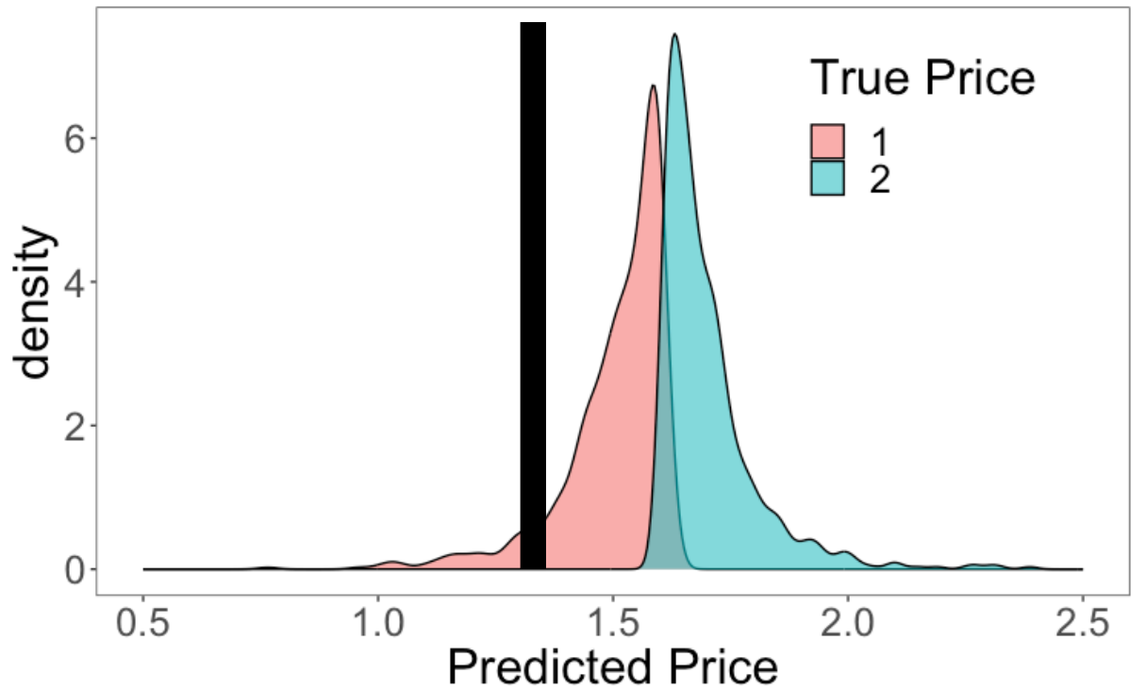


Evaluating Accuracy

Binary target, continuous prediction

Receiver Operator Characteristic (ROC) curve

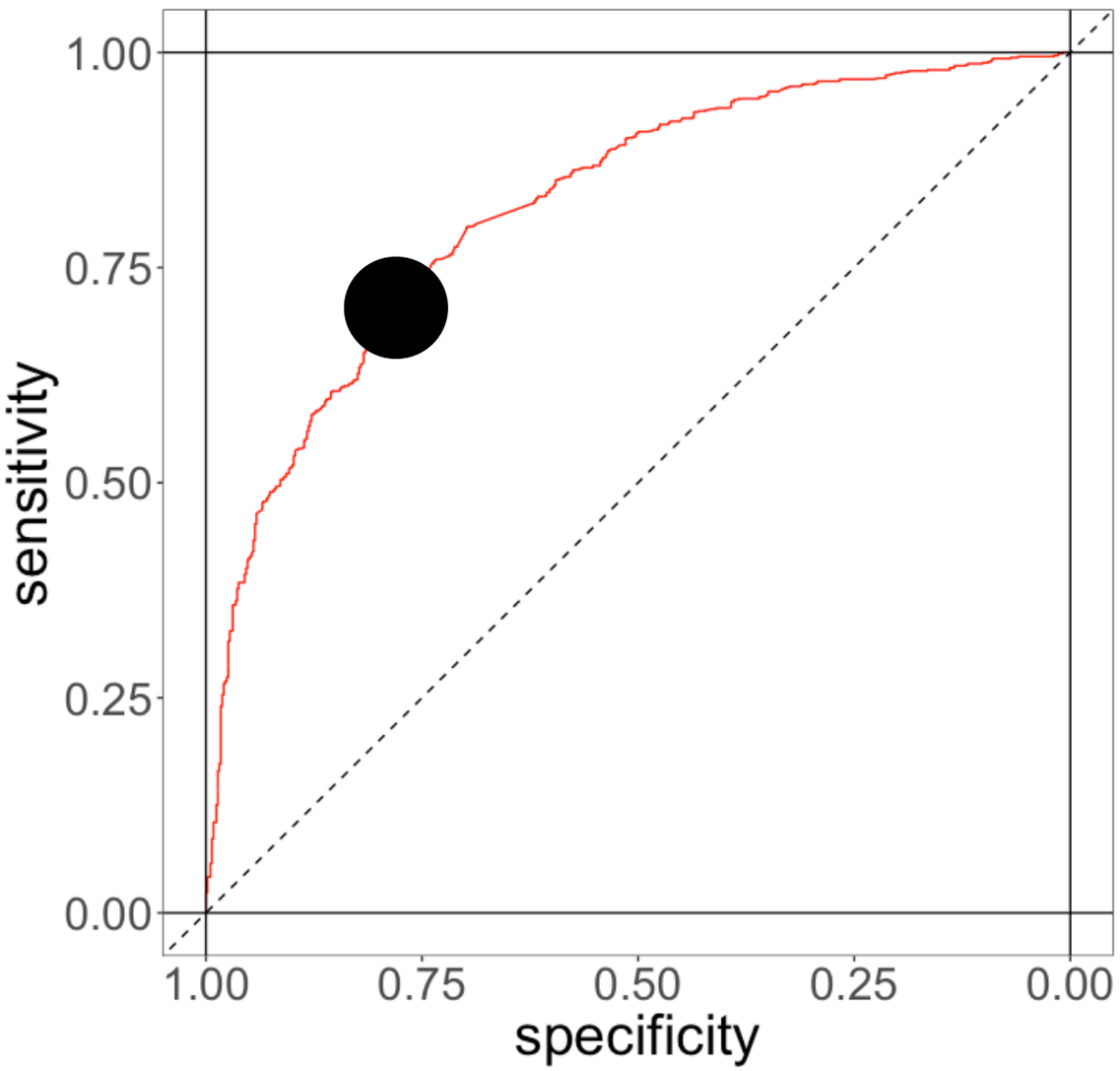


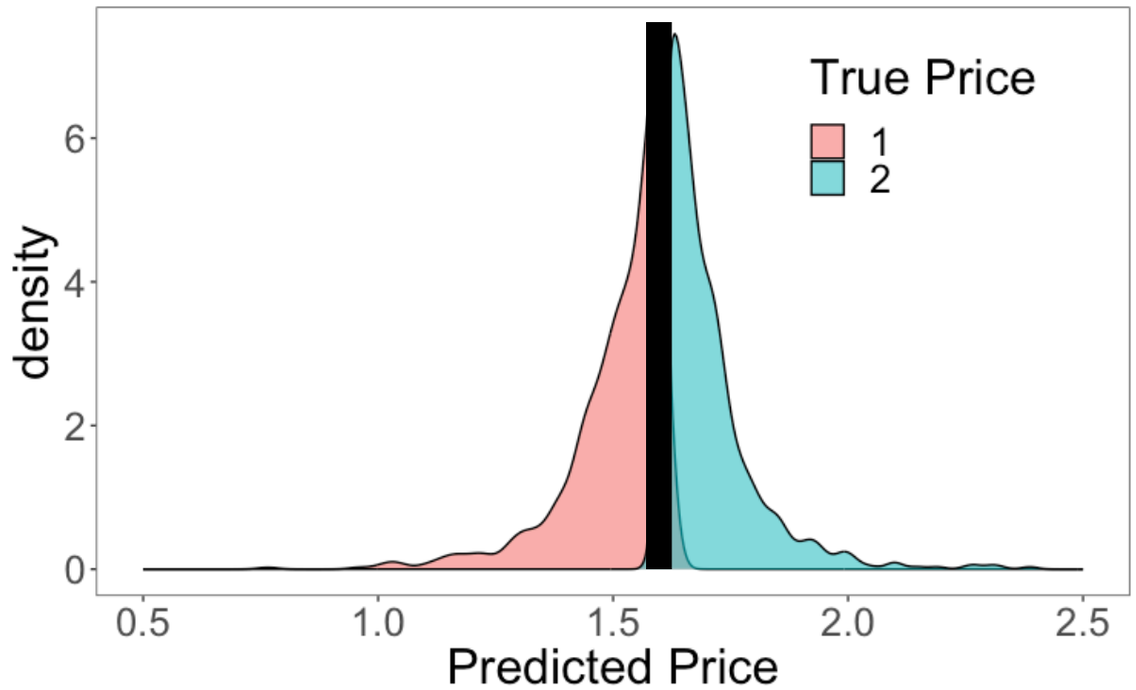


Evaluating Accuracy

Binary target, continuous prediction

Receiver Operator Characteristic (ROC) curve

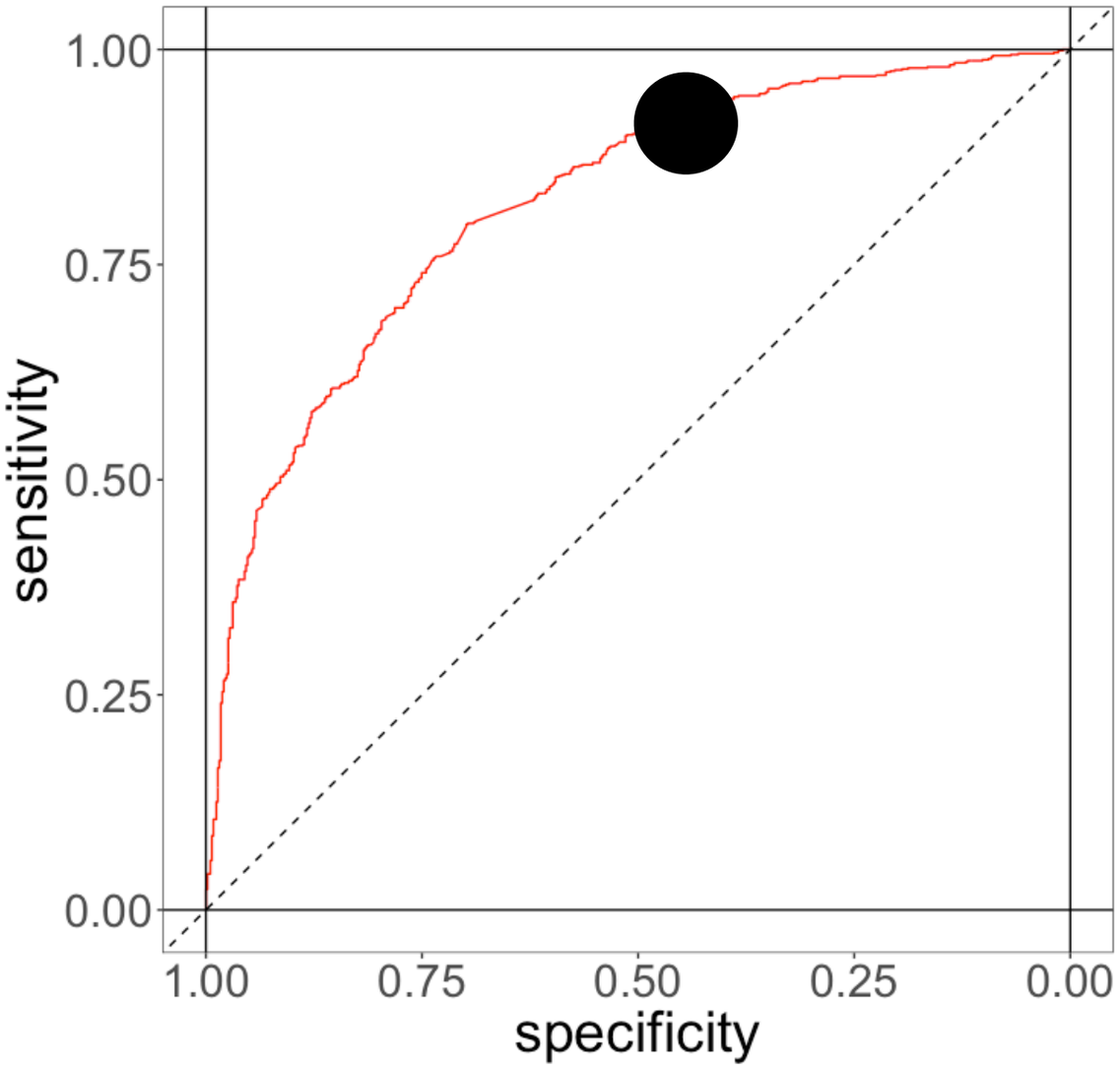


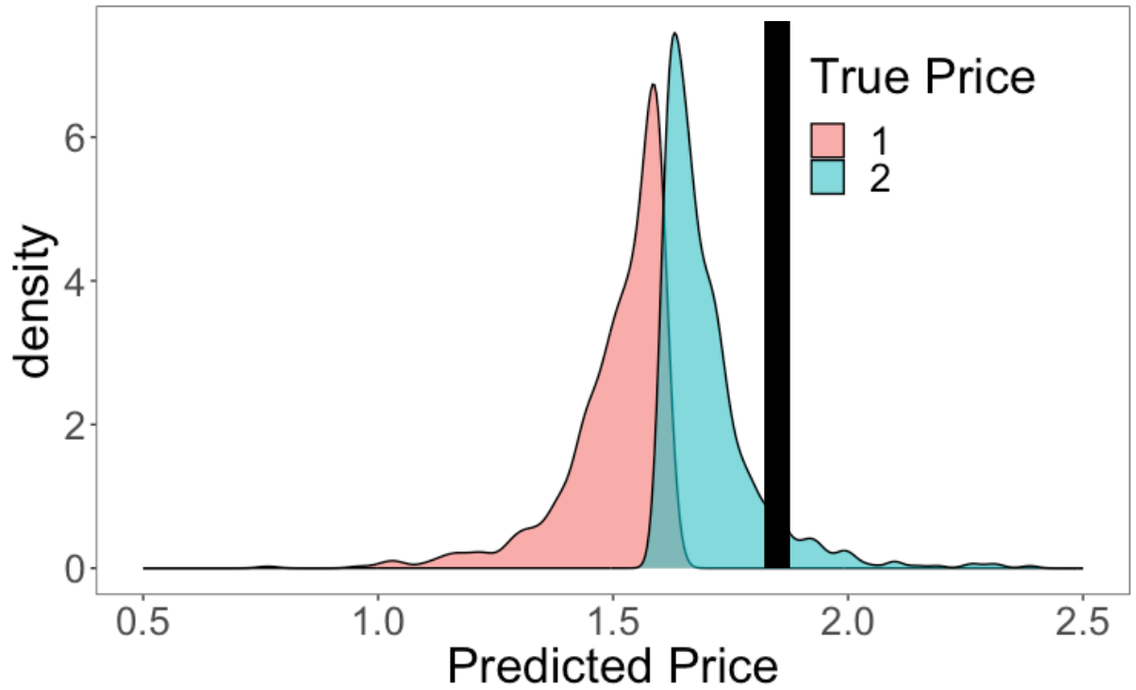


Evaluating Accuracy

Binary target, continuous prediction

Receiver Operator Characteristic (ROC) curve

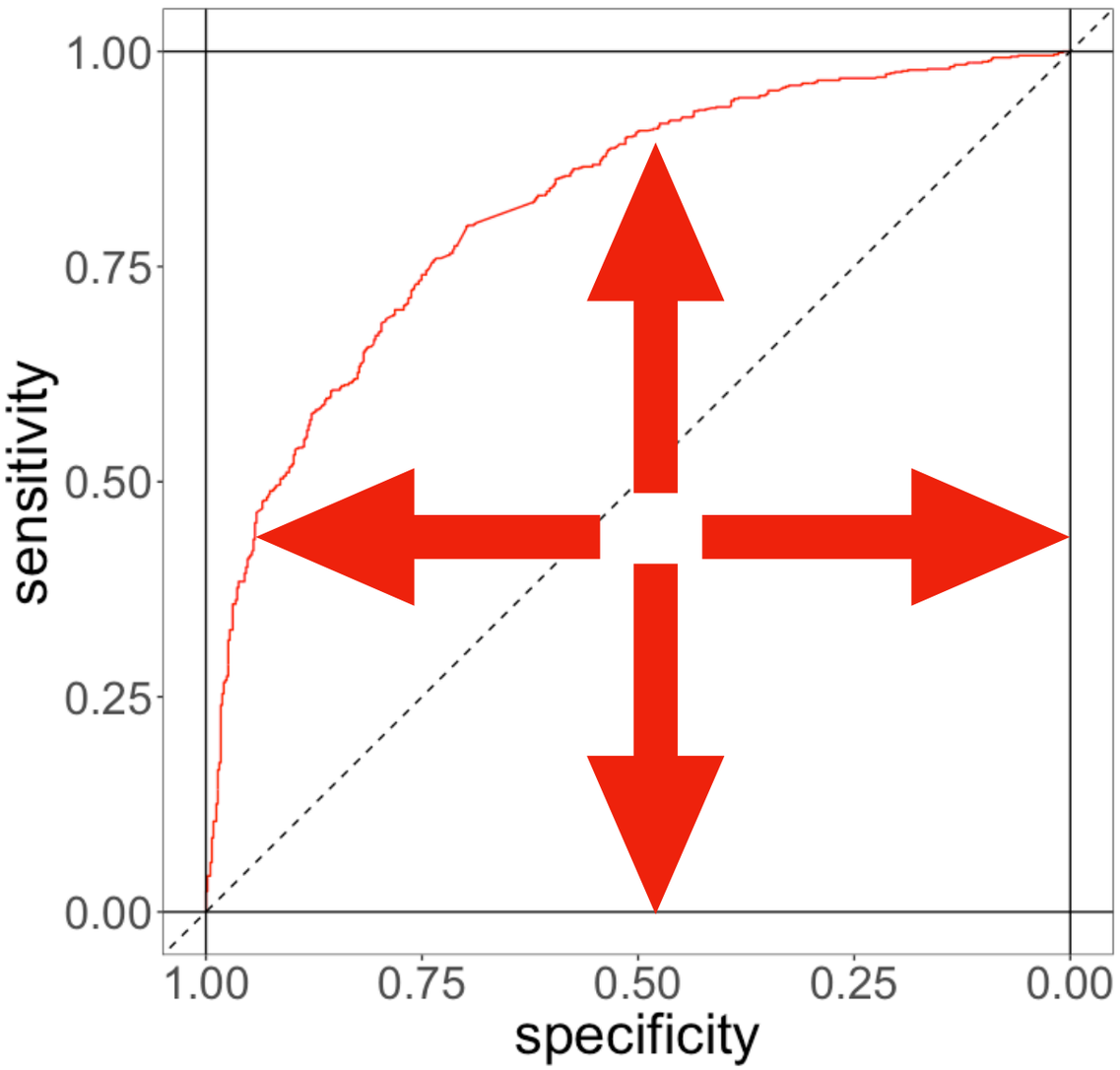


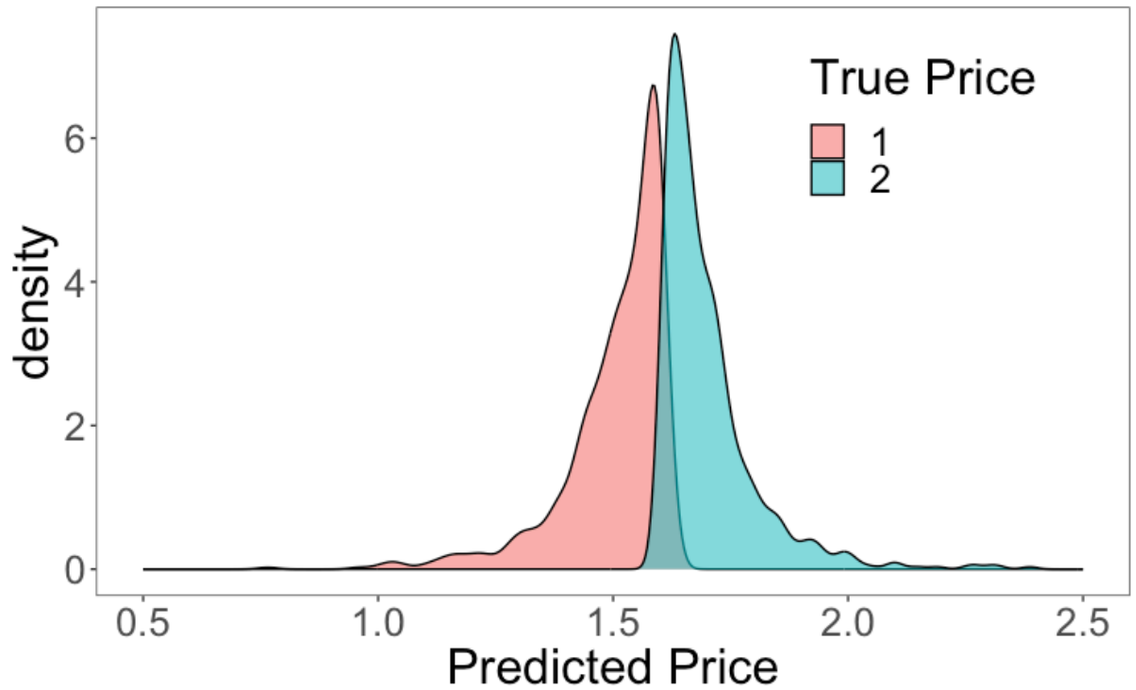


Evaluating Accuracy

Binary target, continuous prediction

Receiver Operator Characteristic (ROC) curve





Accuracy = AUC
"Area Under the Curve"

Evaluating Accuracy

Continuous target, continuous prediction

Kendall's Tau (sometimes "Probability of superiority")

Sentence	Text	Prediction	Target
A	This is	4.2	7
B	This is	3.4	3
C	I hated it	3.1	4
D	I loved it	4.5	6

Pair		Target	Prediction	Accuracy
A	B			
A	C			
A	D			
B	C			
B	D			
C	D			

Evaluating Accuracy

Continuous target, continuous prediction

Kendall's Tau (sometimes "Probability of superiority")

Sentence	Text	Prediction	Target
A	This is	4.2	7
B	This is	3.4	3
C	I hated it	3.1	4
D	I loved it	4.5	6

Pair		Target	Prediction	Accuracy
A	B	>		
A	C	>		
A	D	>		
B	C	<		
B	D	<		
C	D	<		

Evaluating Accuracy

Continuous target, continuous prediction

Kendall's Tau (sometimes "Probability of superiority")

Sentence	Text	Prediction	Target
A	This is	4.2	7
B	This is	3.4	3
C	I hated it	3.1	4
D	I loved it	4.5	6

Pair		Target	Prediction	Accuracy
A	B	>	>	
A	C	>	>	
A	D	>	<	
B	C	<	>	
B	D	<	<	
C	D	<	<	

Evaluating Accuracy

Continuous target, continuous prediction

Kendall's Tau (sometimes "Probability of superiority")

Sentence	Text	Prediction	Target
A	This is	4.2	7
B	This is	3.4	3
C	I hated it	3.1	4
D	I loved it	4.5	6

Pair		Target	Prediction	Accuracy
A	B	>	>	1
A	C	>	>	1
A	D	>	<	0
B	C	<	>	0
B	D	<	<	1
C	D	<	<	1

Evaluating Accuracy

Continuous target, continuous prediction

Kendall's Tau (sometimes "Probability of superiority")

Sentence	Text	Prediction	Target
A	This is	4.2	7
B	This is	3.4	3
C	I hated it	3.1	4
D	I loved it	4.5	6

Pair		Target	Prediction	Accuracy
A	B	>	>	1
A	C	>	>	1
A	D	>	<	0
B	C	<	>	0
B	D	<	<	1
C	D	<	<	1

Accuracy
= 4/6 ~ 67%

Evaluating Accuracy

Continuous target, continuous prediction

Kendall's Tau (sometimes "Probability of superiority")

Sentence	Text	Prediction	Target
A	This is	4.2	7
B	This is	3.4	3
C	I hated it	3.1	4
D	I loved it	4.5	6

Pair		Target	Prediction	Accuracy
A	B	>	>	1
A	C	>	>	1
A	D	>	<	0
B	C	<	>	0
B	D	<	<	1
C	D	<	<	1

Accuracy
= 4/6 ~ 67%

Why?

- non-parametric correlation, no assumptions
- metric is easy to compare and interpret

Evaluating Accuracy

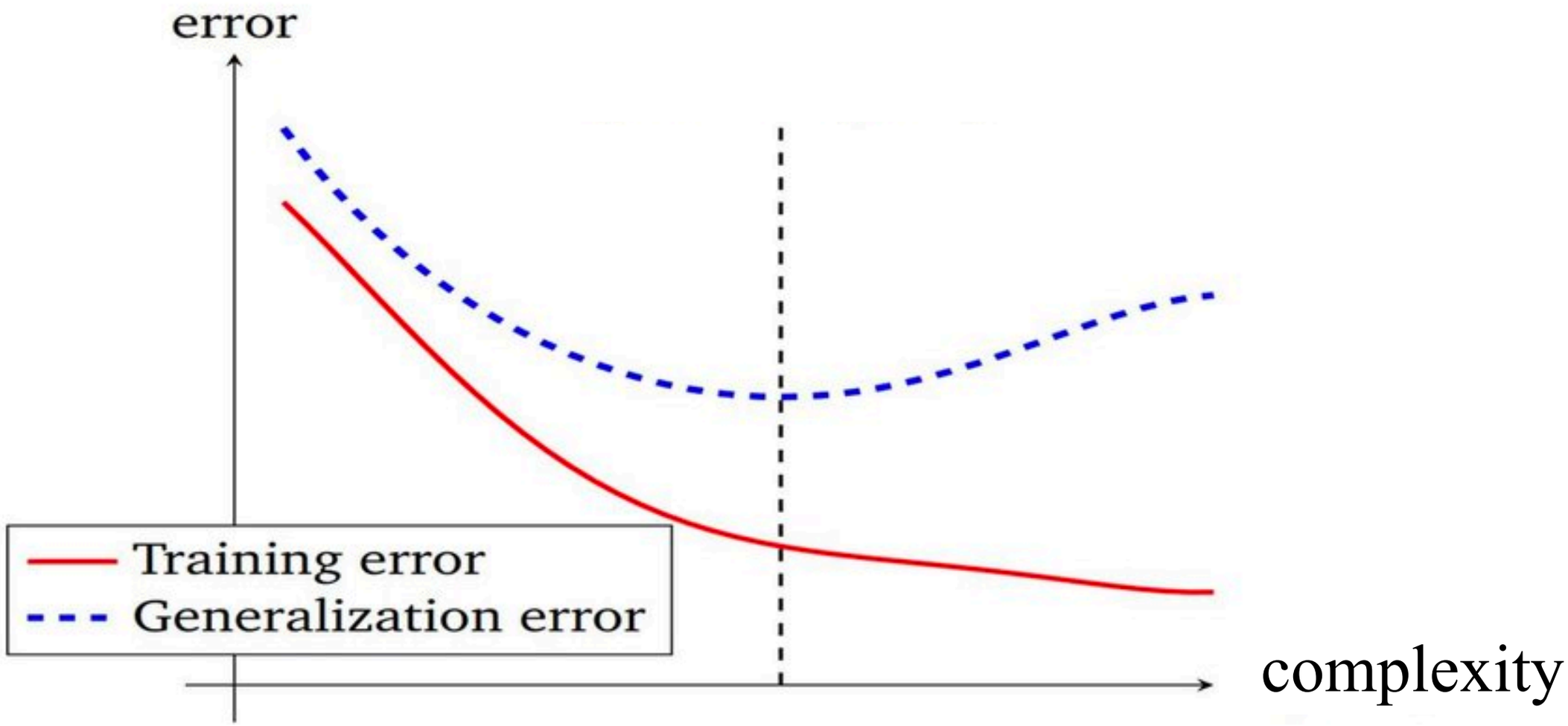
Key insight: "out of sample prediction”

Overfitting to training data will hurt accuracy on test data

Evaluating Accuracy

Key insight: "out of sample prediction”

Overfitting to training data will hurt accuracy on test data



Evaluating Accuracy

How to hold out?

Evaluating Accuracy

How to hold out?

Simplest approach: random split

Evaluating Accuracy

How to hold out?

Simplest approach: random split

Common problems: - data-intensive
- risk of a bad split

Evaluating Accuracy

How to hold out?

Simplest approach: random split

Common problems: - data-intensive
- risk of a bad split

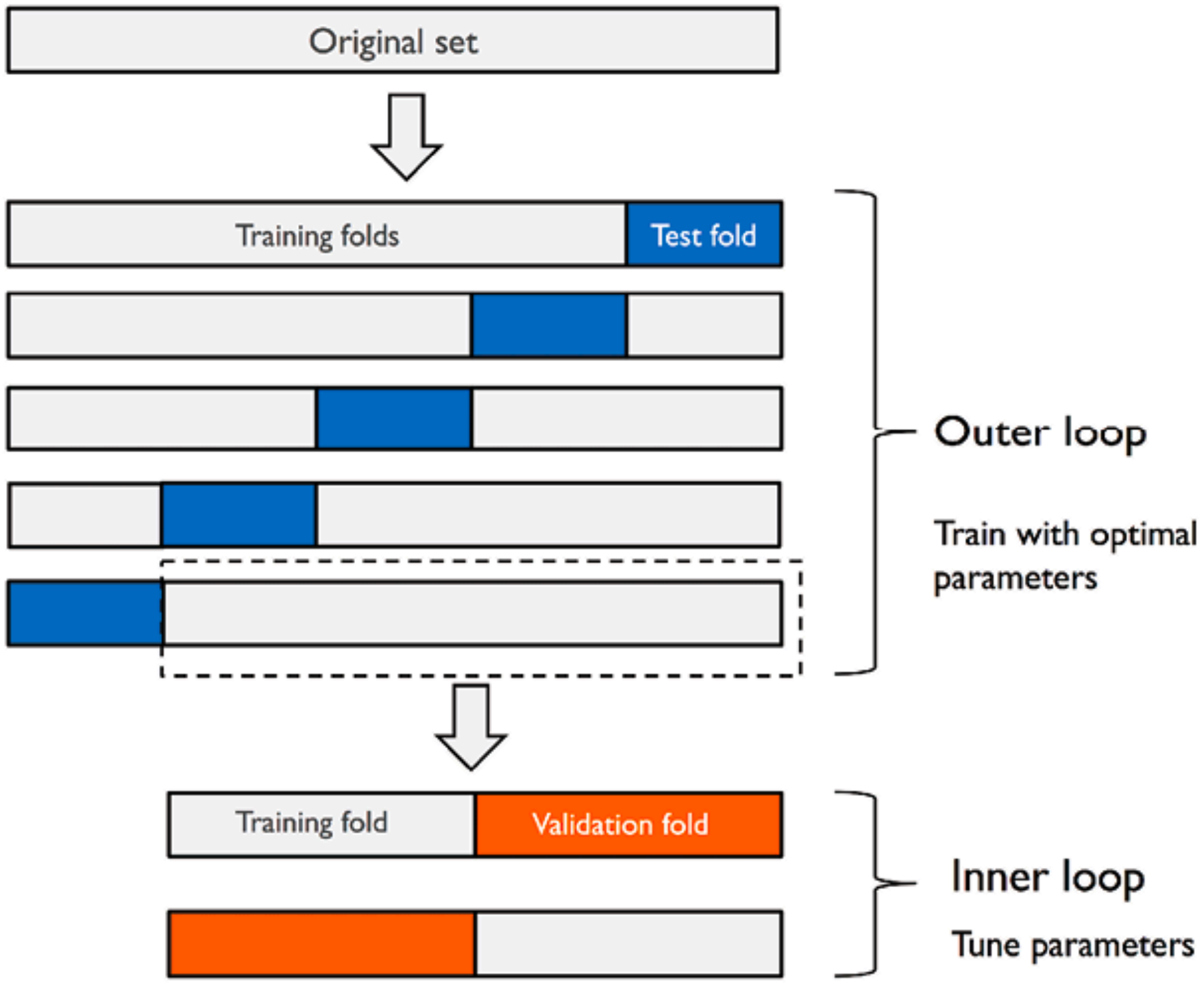
Solutions:

cross-validation (Stone, 1974; Poldrack et al., 2020)

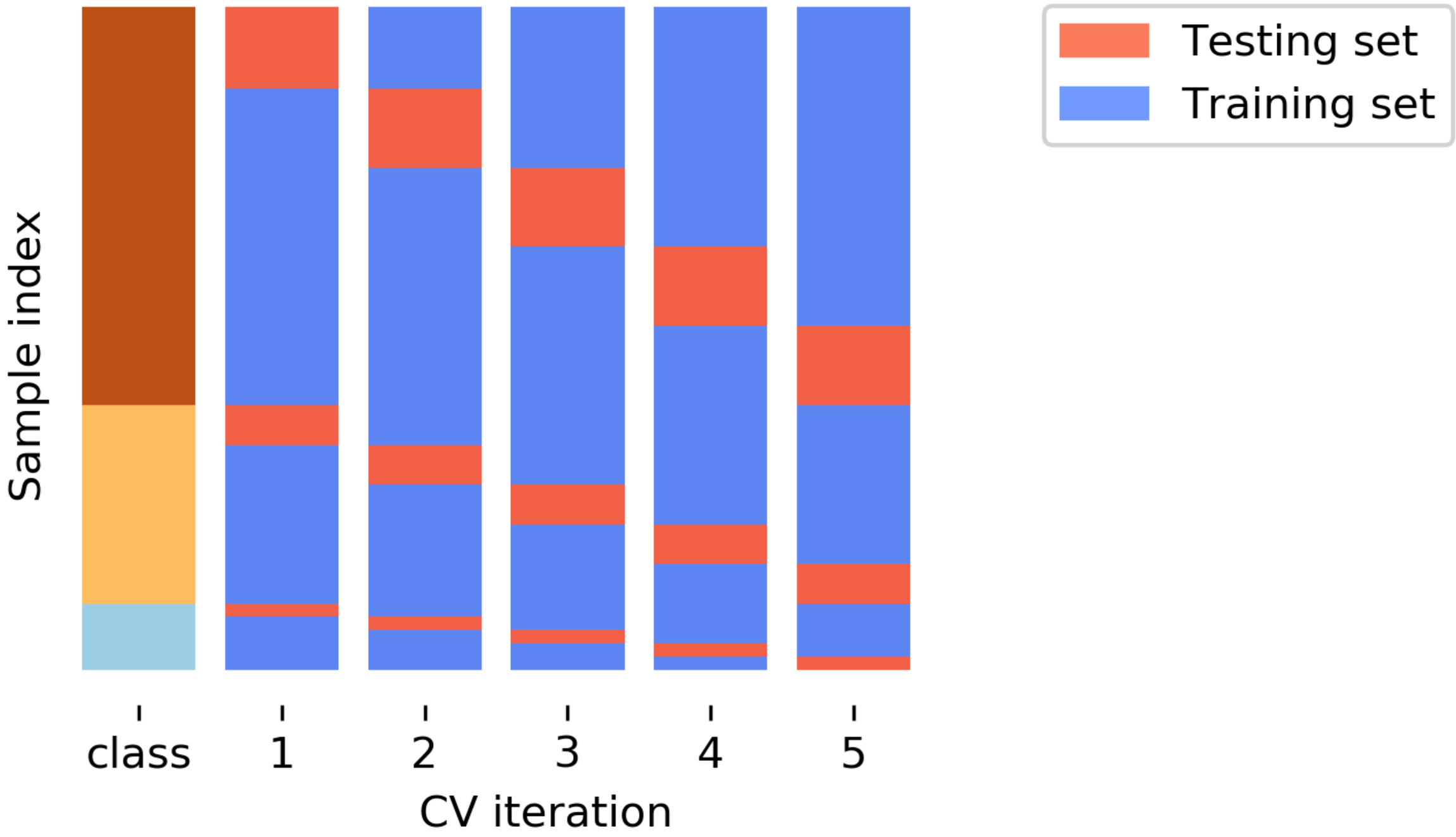
- nested CV

- stratified CV

Nested Cross-Validation



Stratified Cross-Validation



Evaluating Accuracy

How to hold out?

Sophisticated approach: non-random split

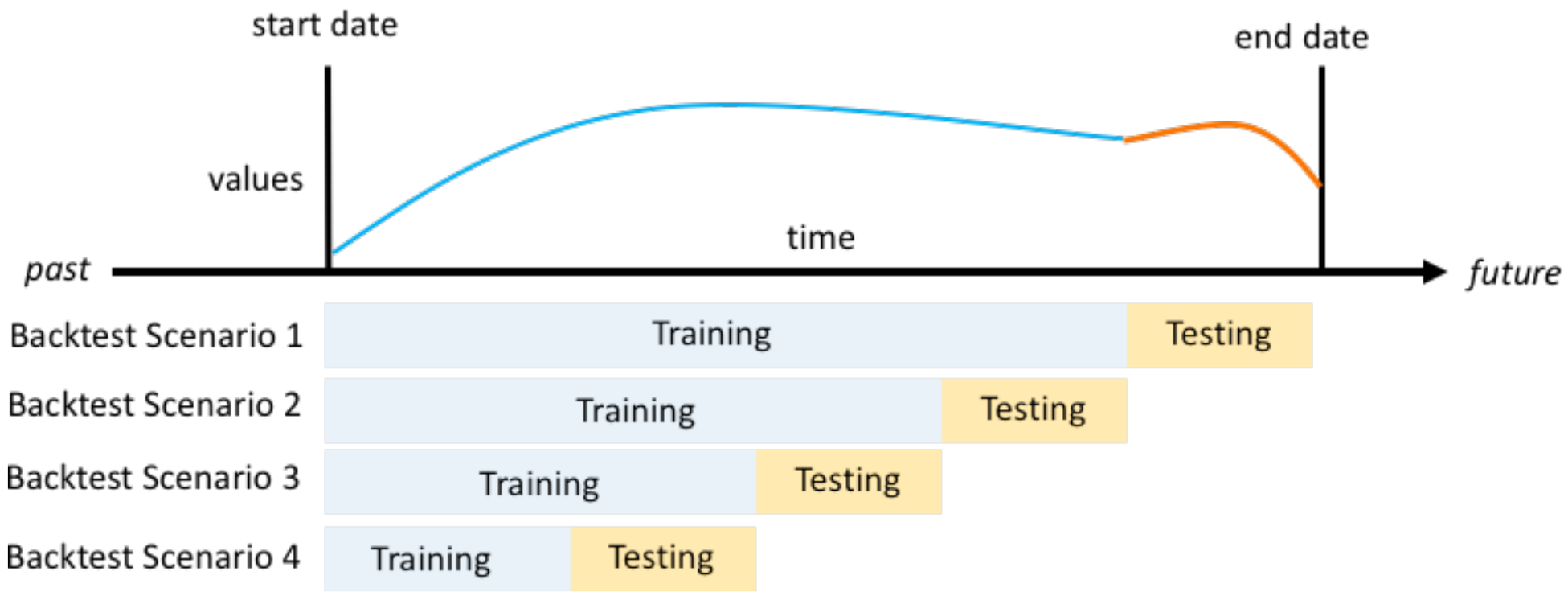
Evaluating Accuracy

How to hold out?

Sophisticated approach: non-random split

Reasons to manually select hold-out:

Time-varying effects (e.g. forecasting)



Evaluating Accuracy

How to hold out?

Sophisticated approach: non-random split

Reasons to manually select hold-out:

Time-varying effects (e.g. forecasting)

Context-varying effects (almost everything)

Evaluating Accuracy

How to hold out?

Sophisticated approach: non-random split

Reasons to manually select hold-out:

Time-varying effects (e.g. forecasting)

Context-varying effects (almost everything)

Key concept: **“transfer learning”**

How well does your model generalise?

Transfer Learning

Do people enjoy different things about high-, medium-, and low-price restaurants?

Transfer Learning

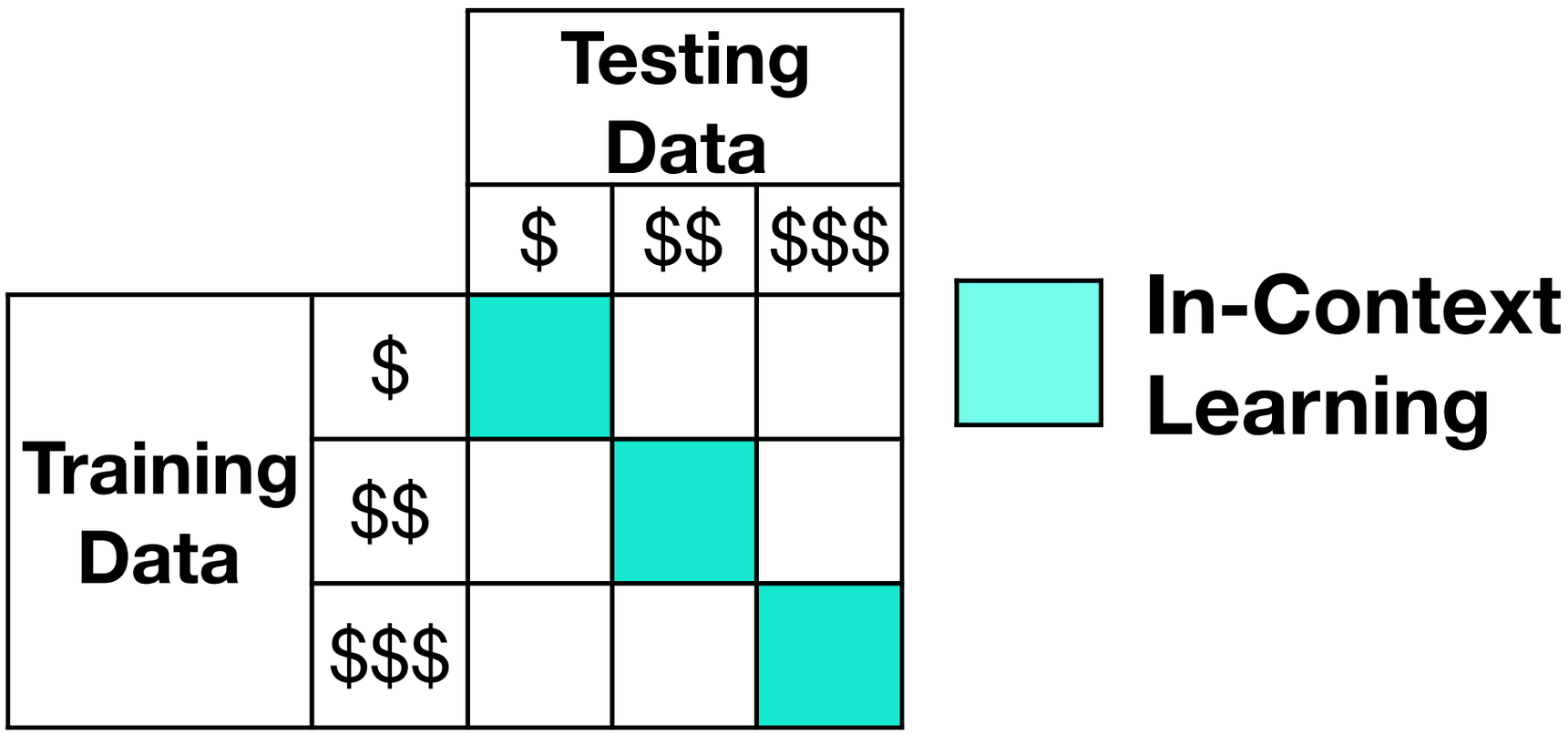
Do people enjoy different things about high-, medium-, and low-price restaurants?

- split sample by prices
- train three models to predict star rating
- test accuracy within and across prices

Transfer Learning

Do people enjoy different things about high-, medium-, and low-price restaurants?

- split sample by prices
- train three models to predict star rating
- test accuracy within and across prices

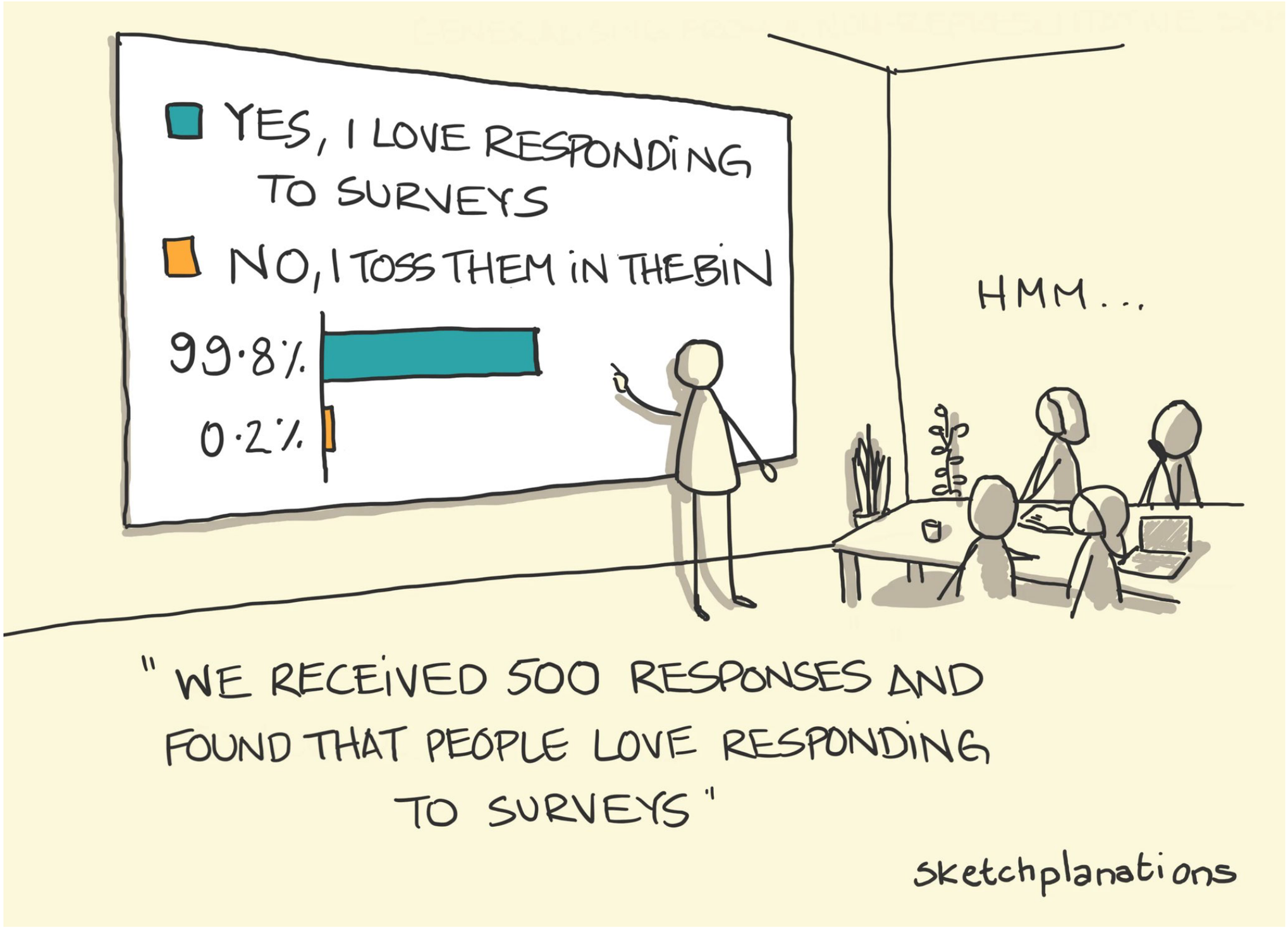


Transfer Learning

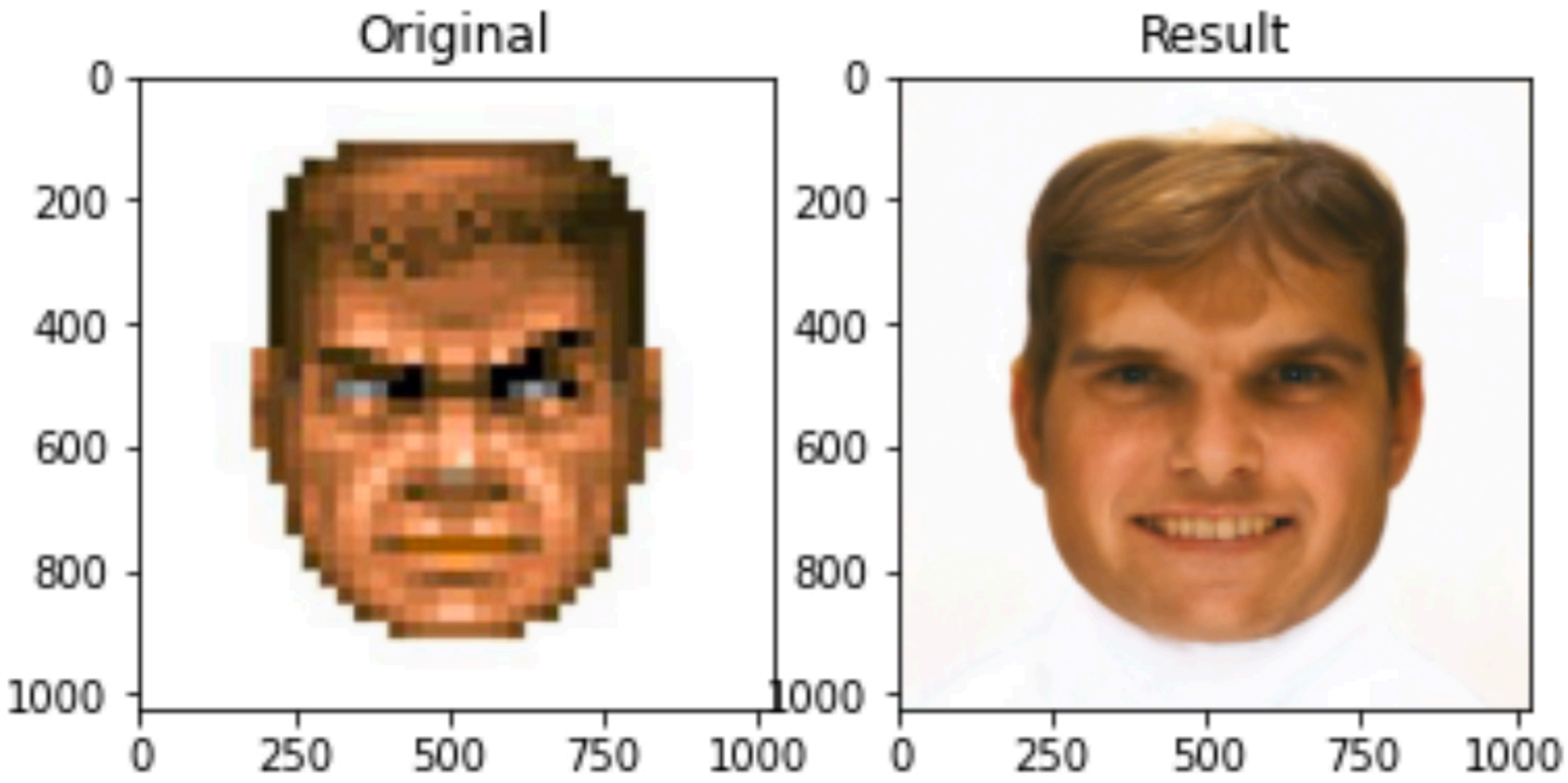
Do people enjoy different things about high-, medium-, and low-price restaurants?

- split sample by prices
- train three models to predict star rating
- test accuracy within and across prices

Sampling Biases

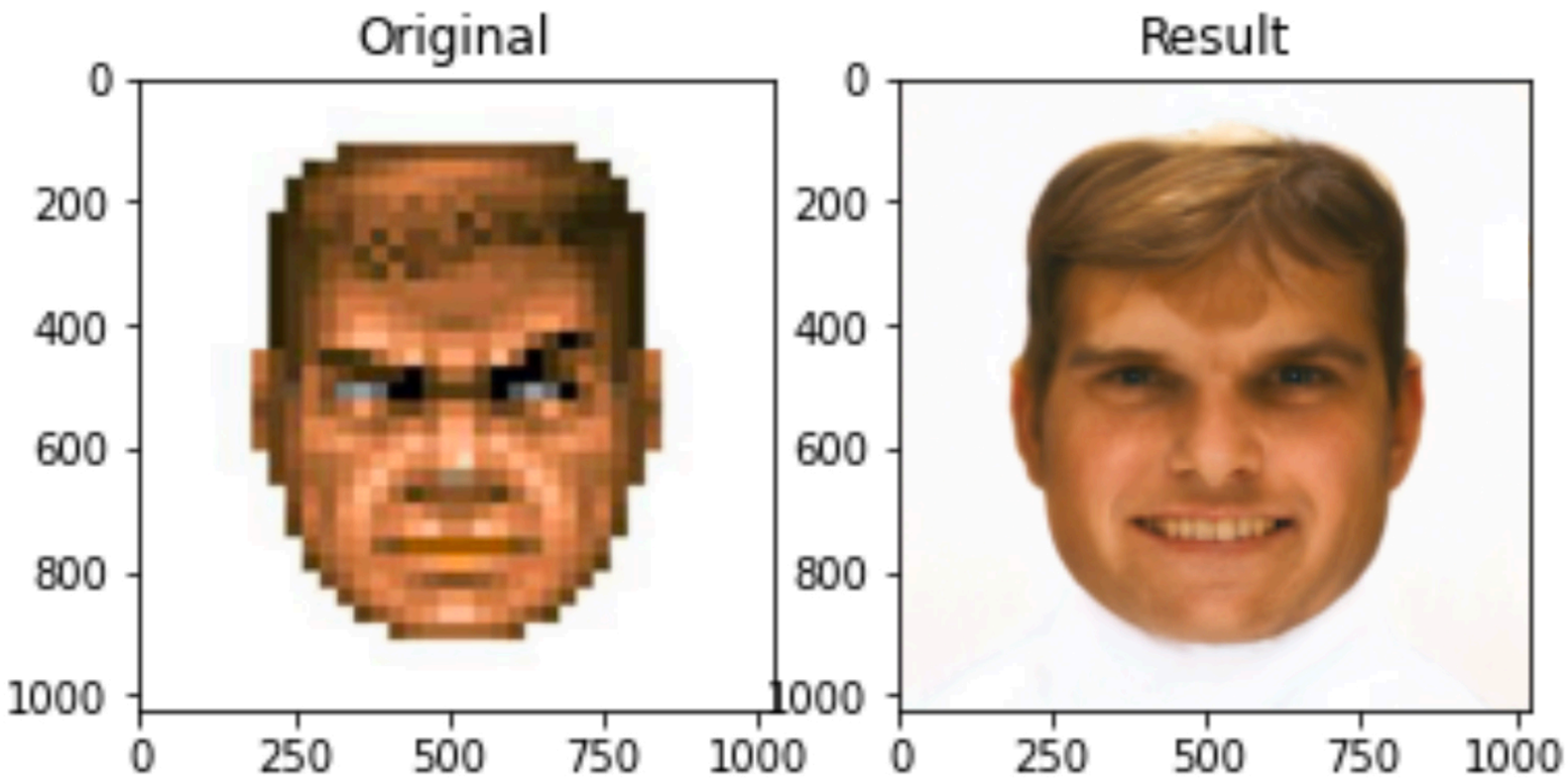


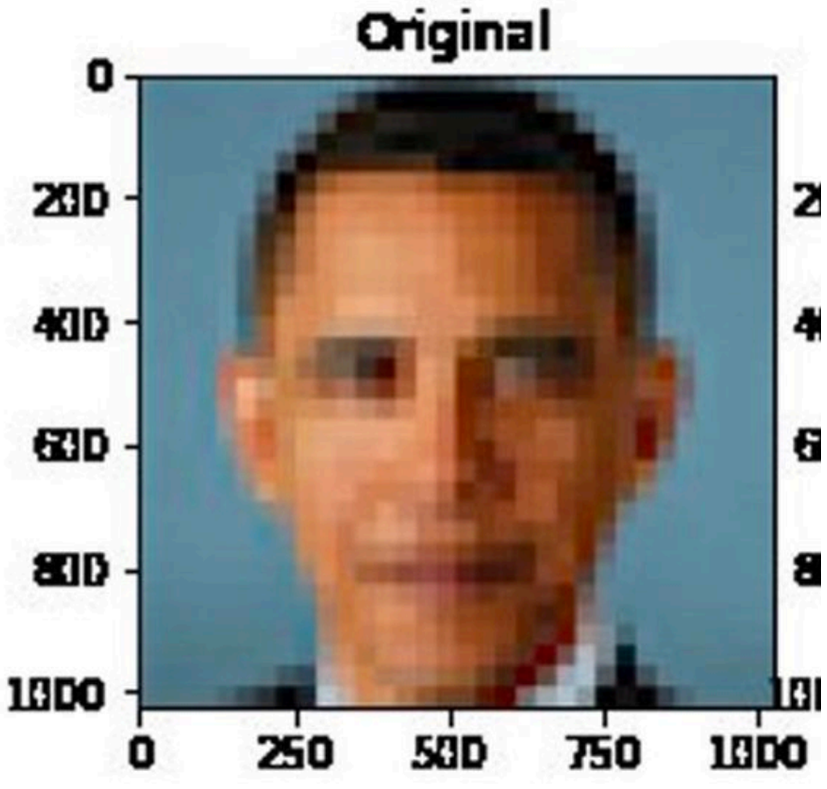
Sampling Biases



Menon et al., 2020

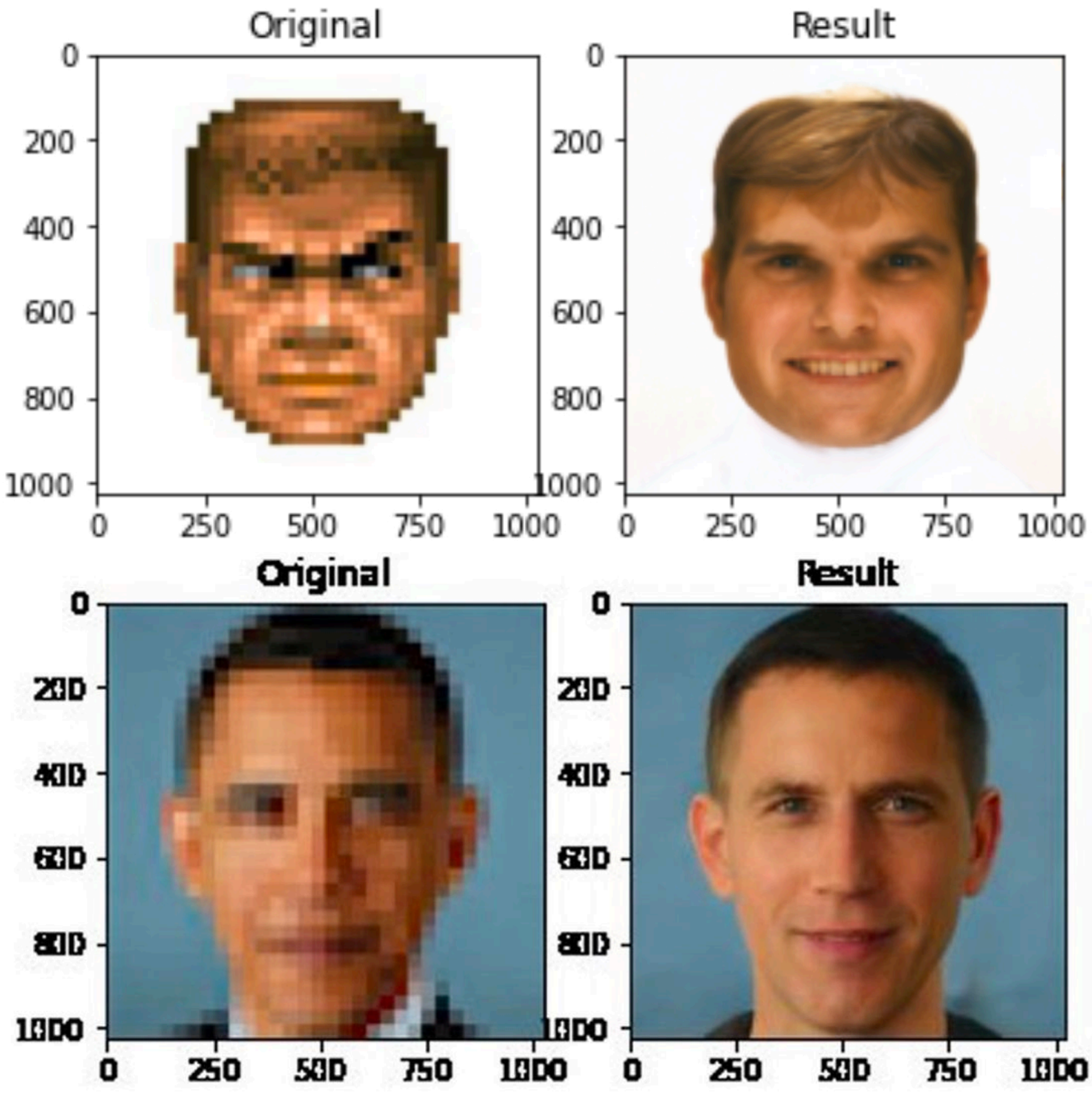
Sampling Biases





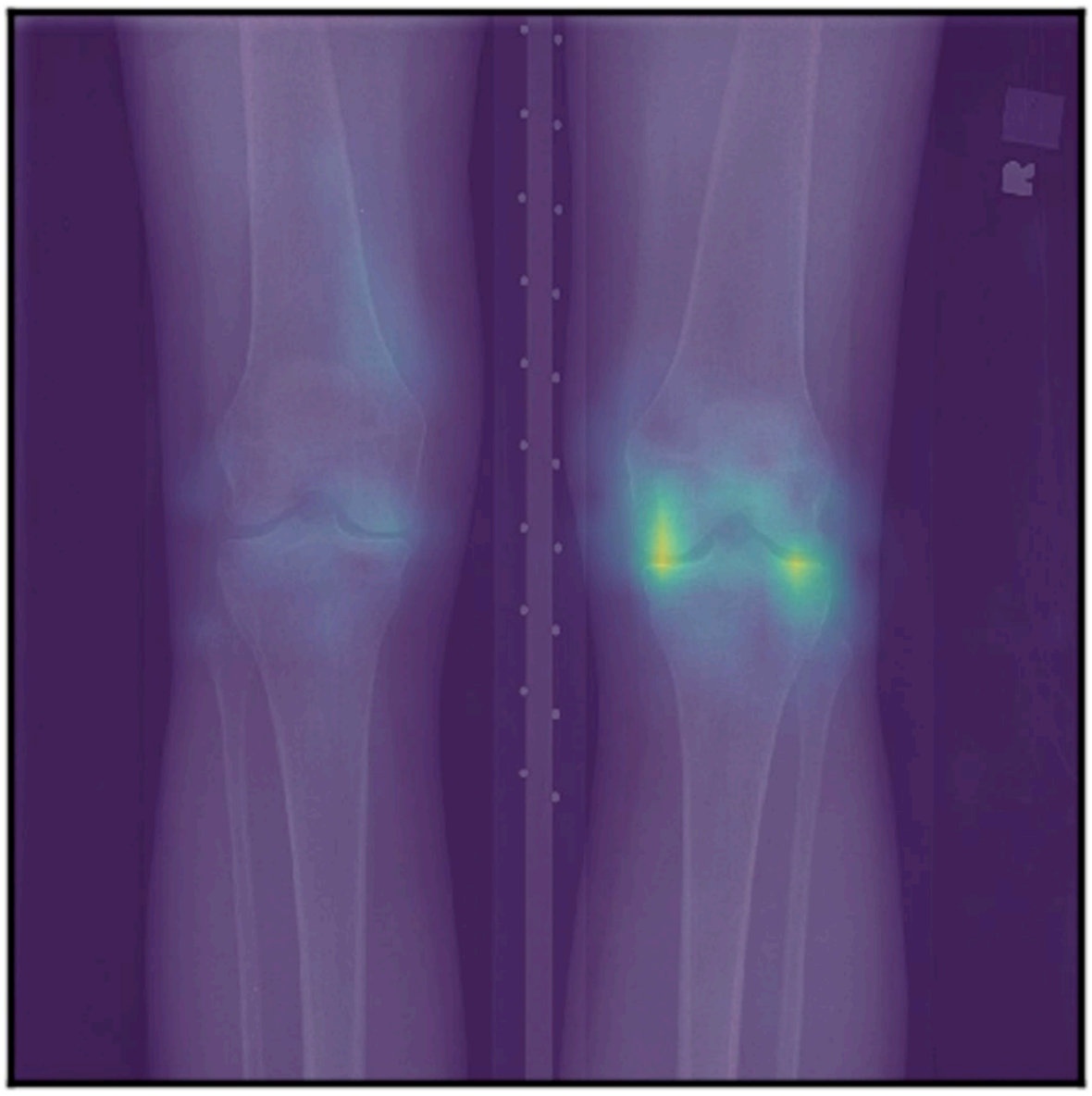
Menon et al., 2020

Sampling Biases



Menon et al., 2020

Sampling Biases



Pierson et al., 2021

Sampling Biases

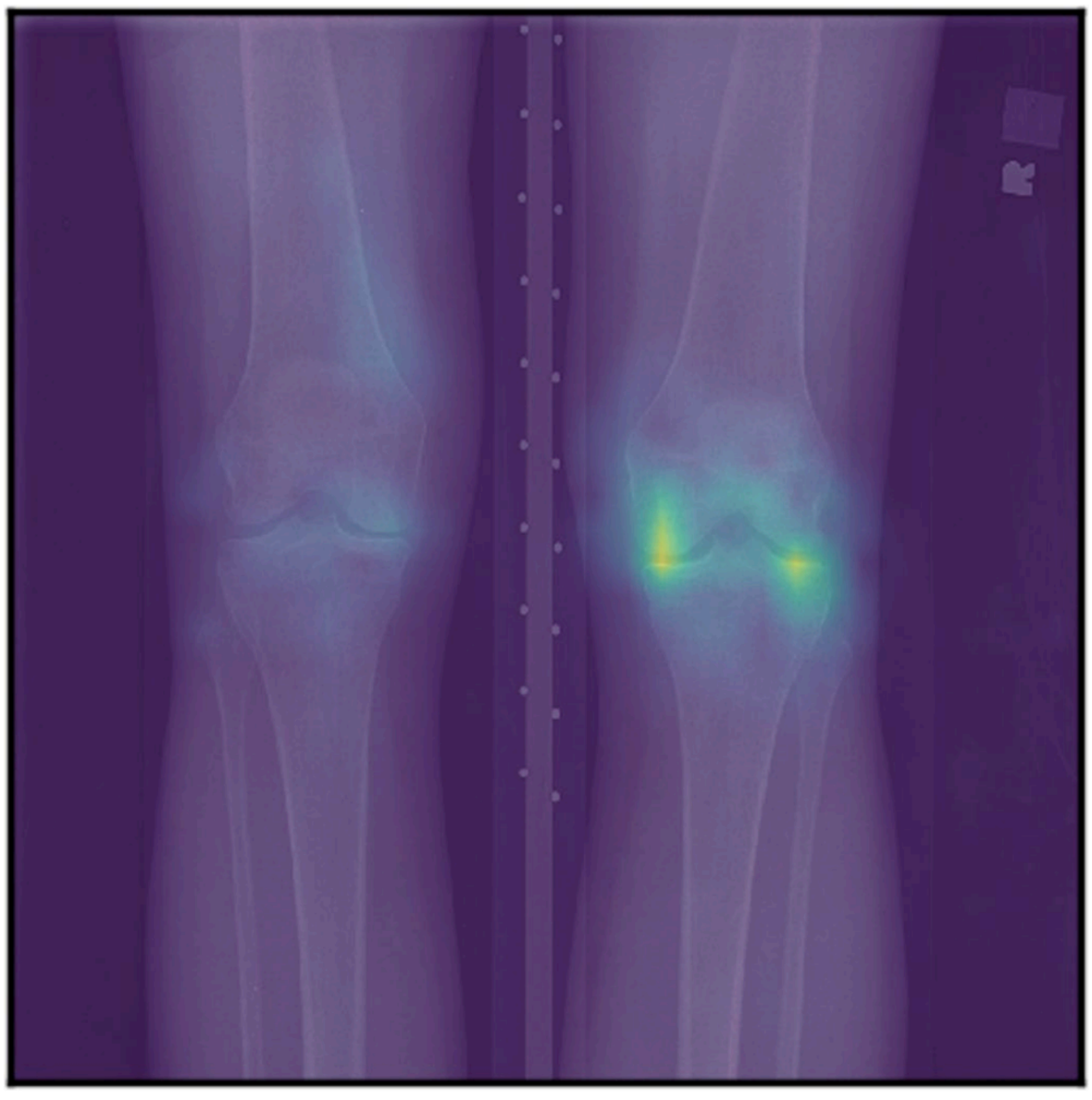
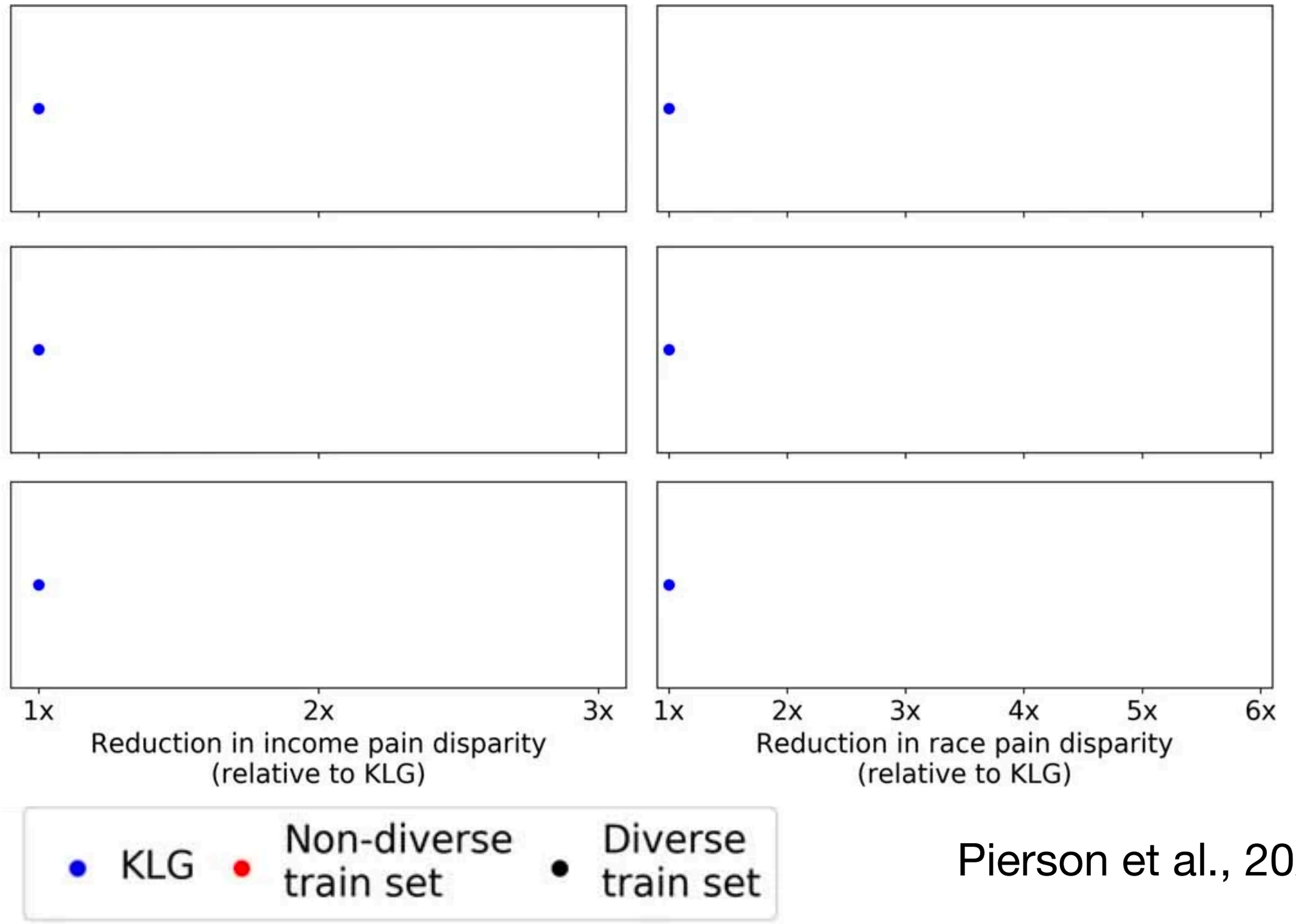


Table 3 | Potential eligibility for surgery: comparing KLG and ALG-P

	Knees potentially eligible for surgery (%)		Knees in severe pain and not eligible for surgery (%)	
	Using KLG	Using ALG-P	Using KLG	Using ALG-P
Black	11% (7%, 15%)	22% (17%, 27%)	51% (45%, 57%)	40% (34%, 46%)
Lower-income	10% (8%, 12%)	13% (10%, 15%)	36% (33%, 40%)	34% (31%, 38%)
Lower-education	9% (7%, 11%)	14% (11%, 16%)	38% (35%, 42%)	33% (30%, 37%)

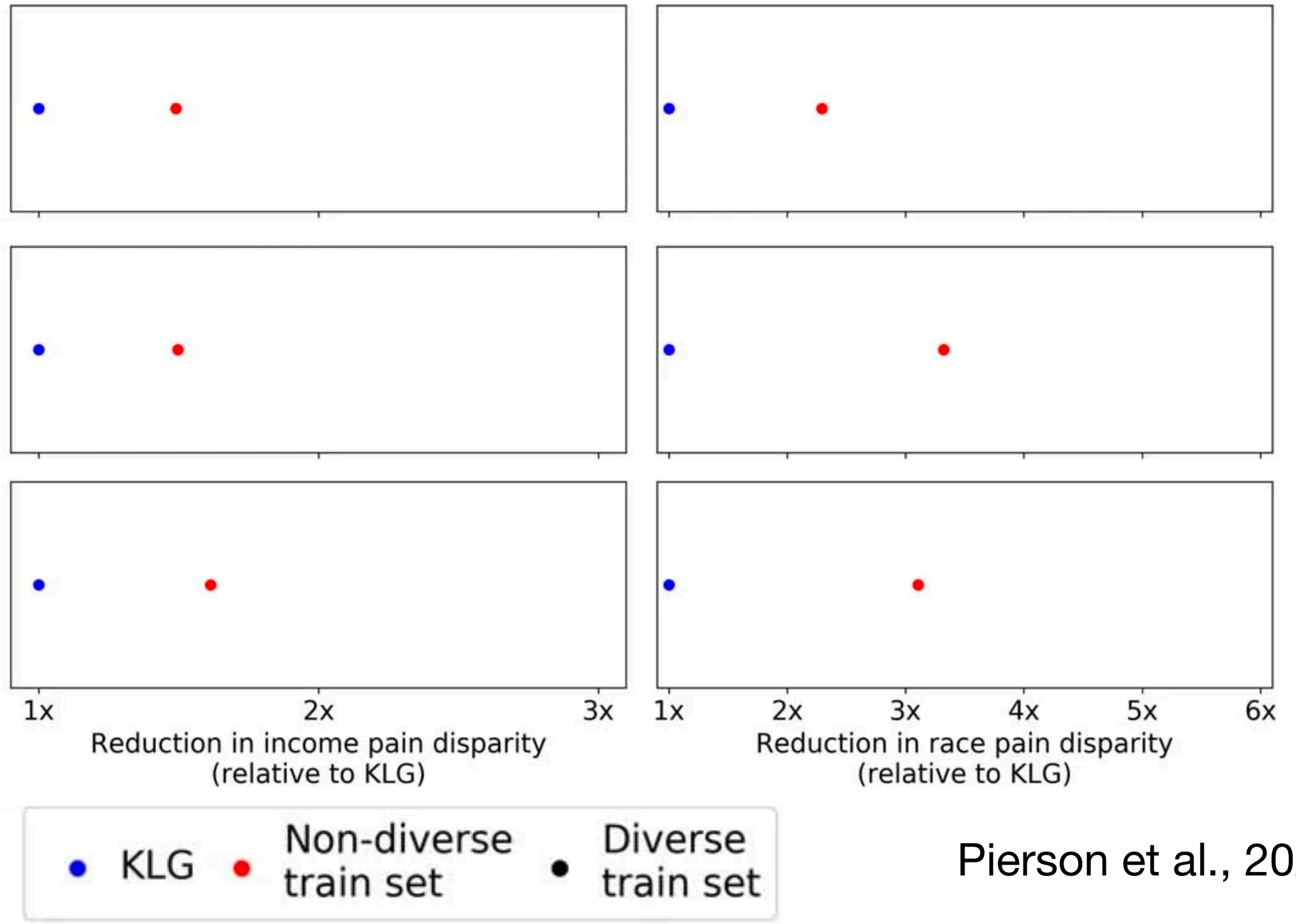
Pierson et al., 2021

Sampling Biases



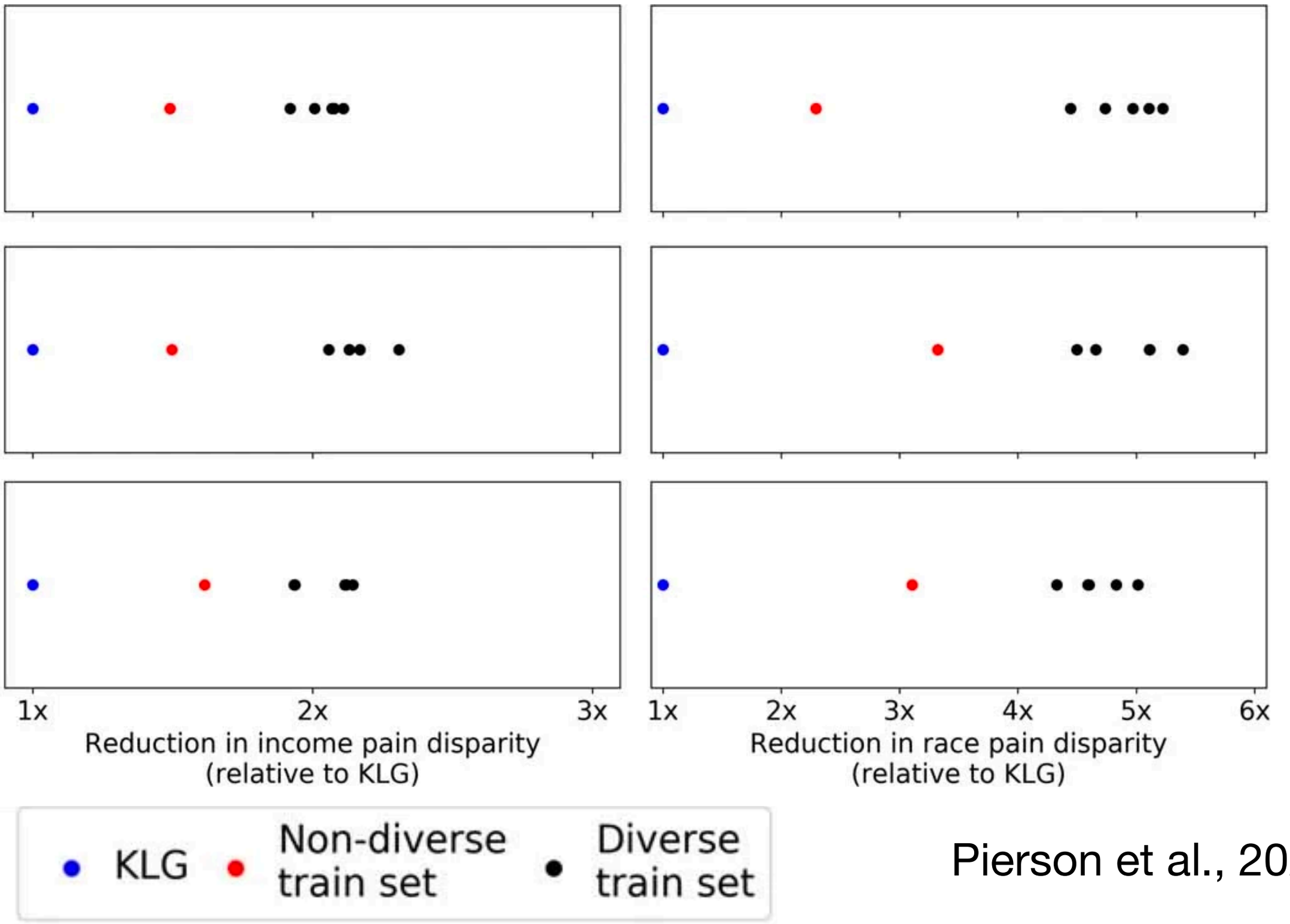
Pierson et al., 2021

Sampling Biases



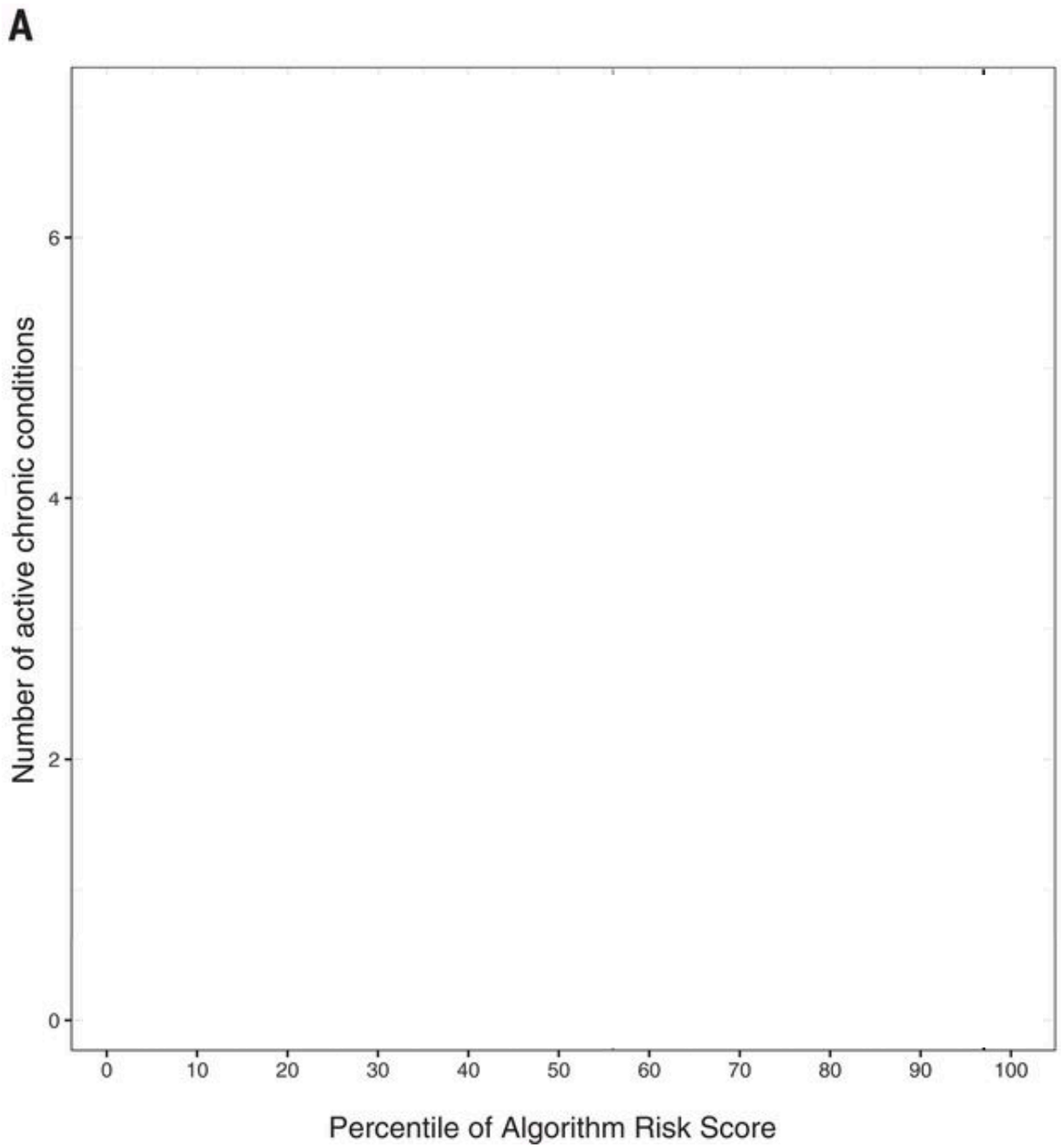
Pierson et al., 2021

Sampling Biases



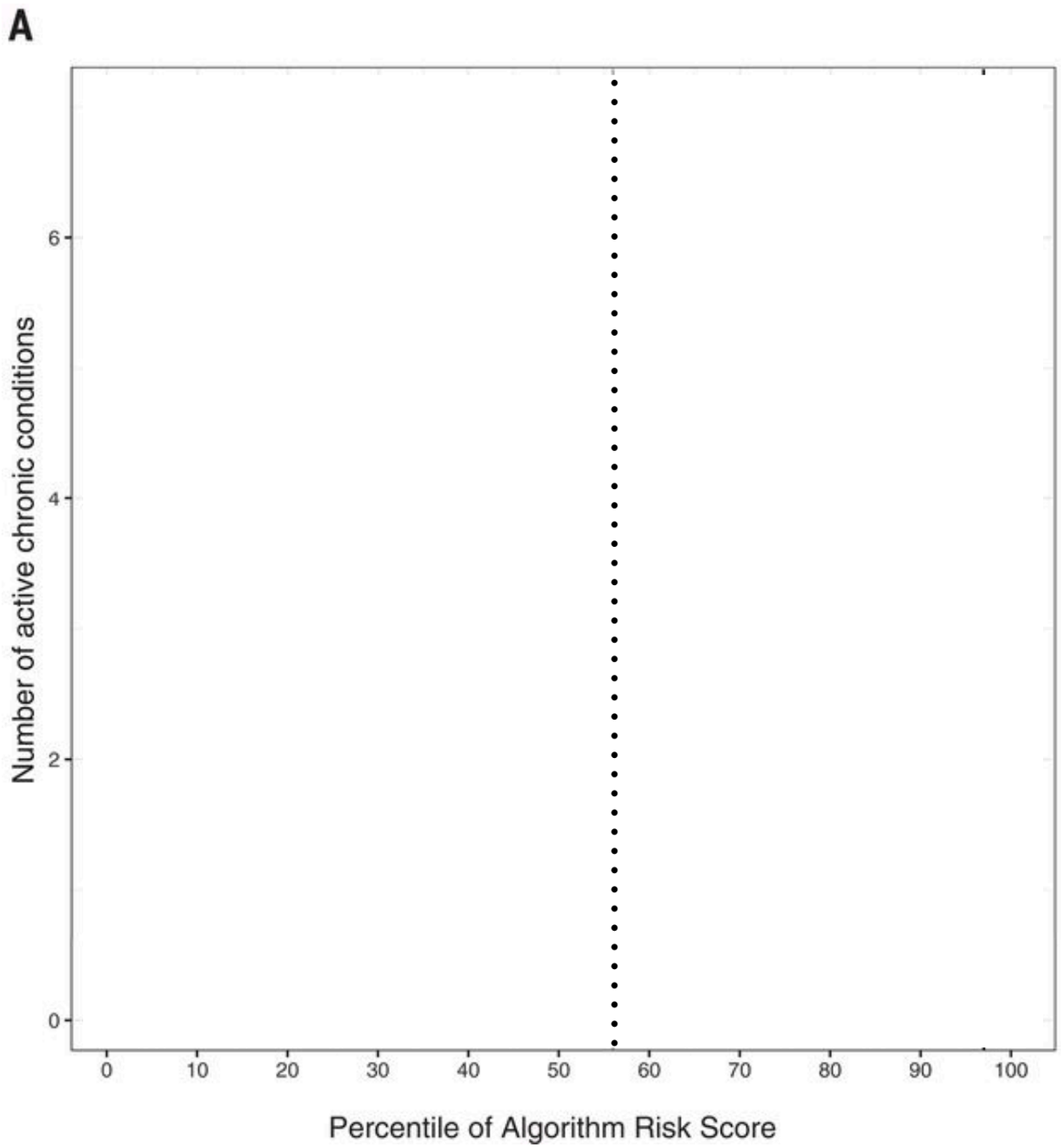
Pierson et al., 2021

Sampling Biases



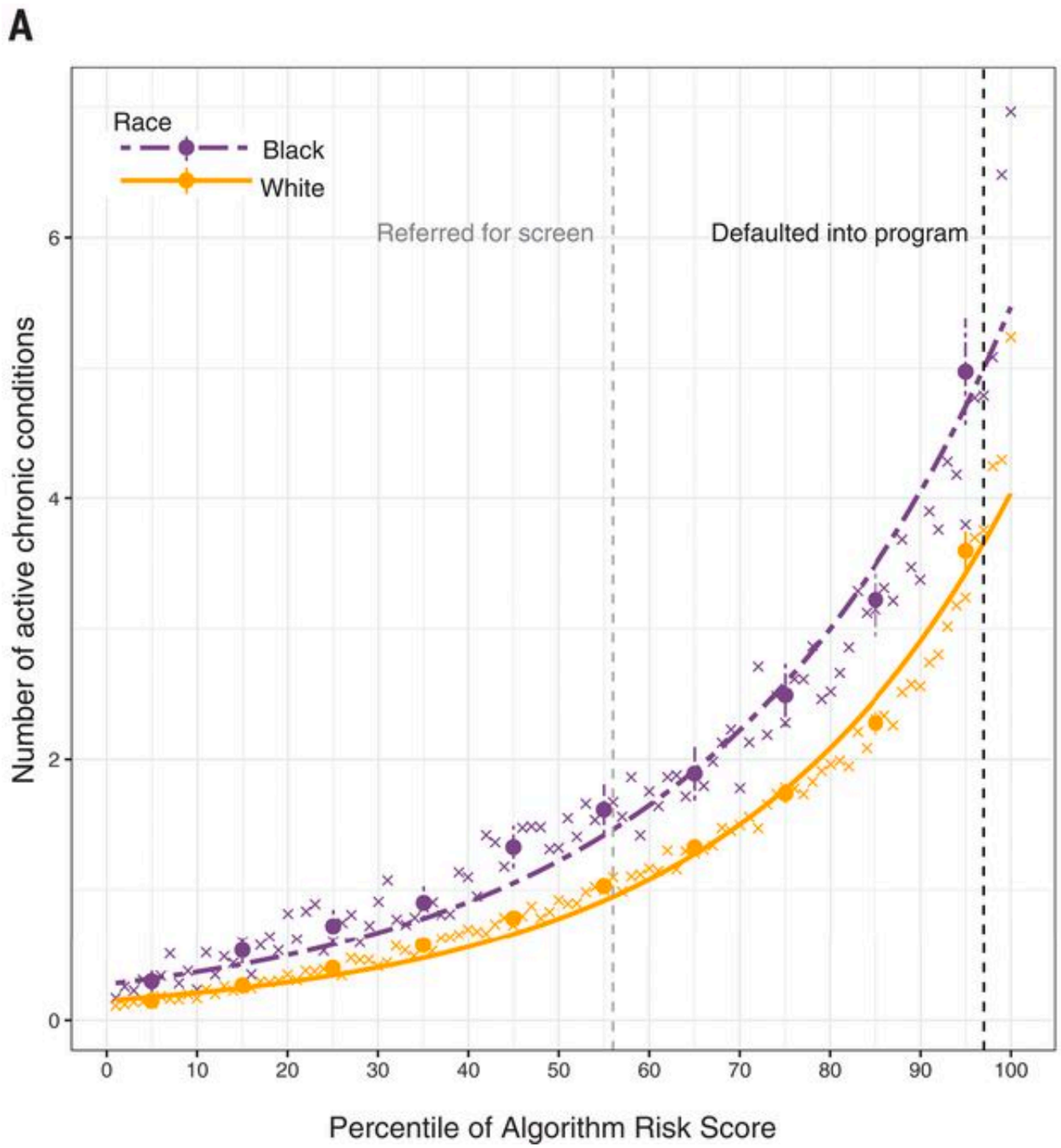
Obermeyer et al., 2019

Sampling Biases



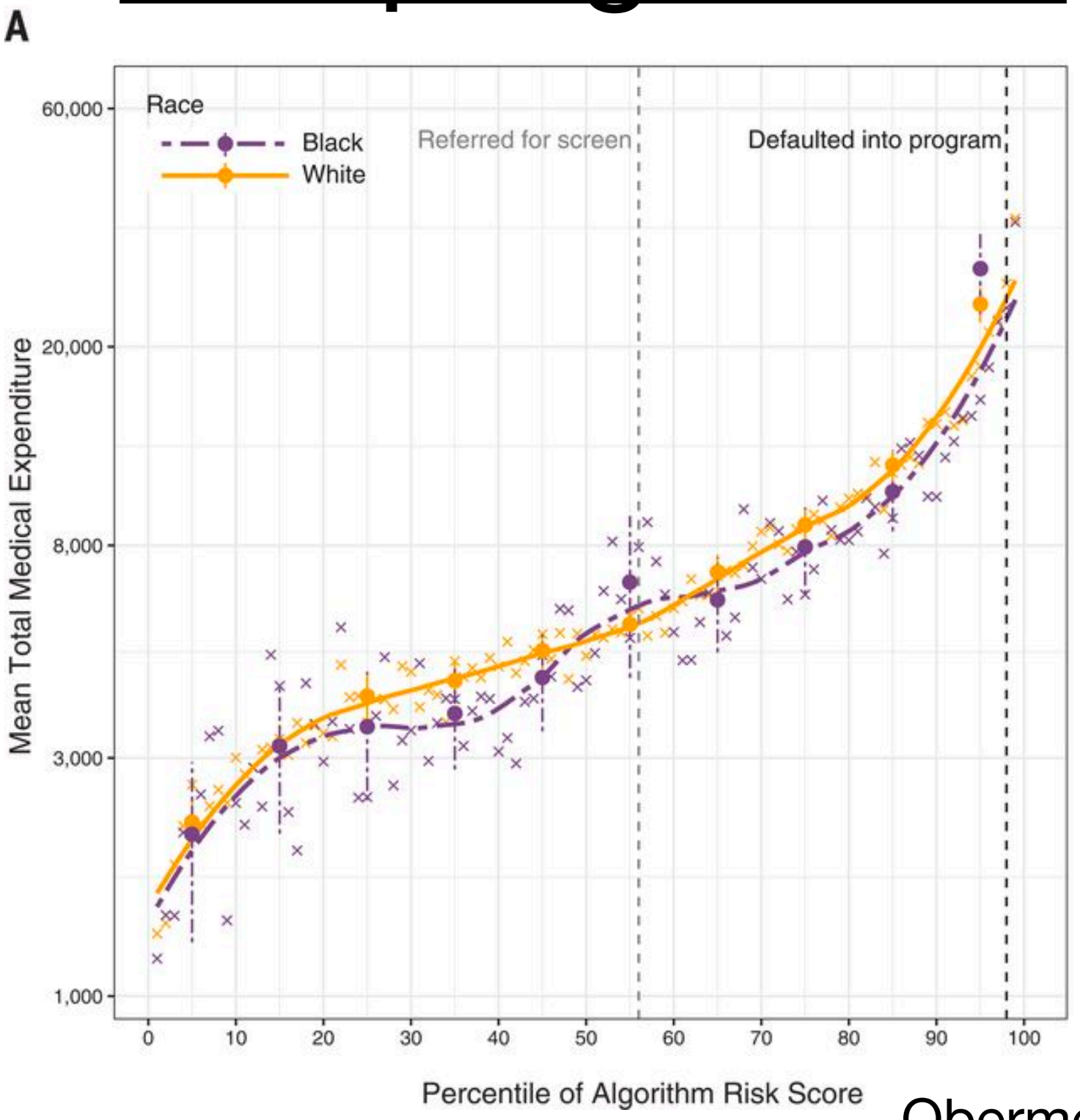
Obermeyer et al., 2019

Sampling Biases



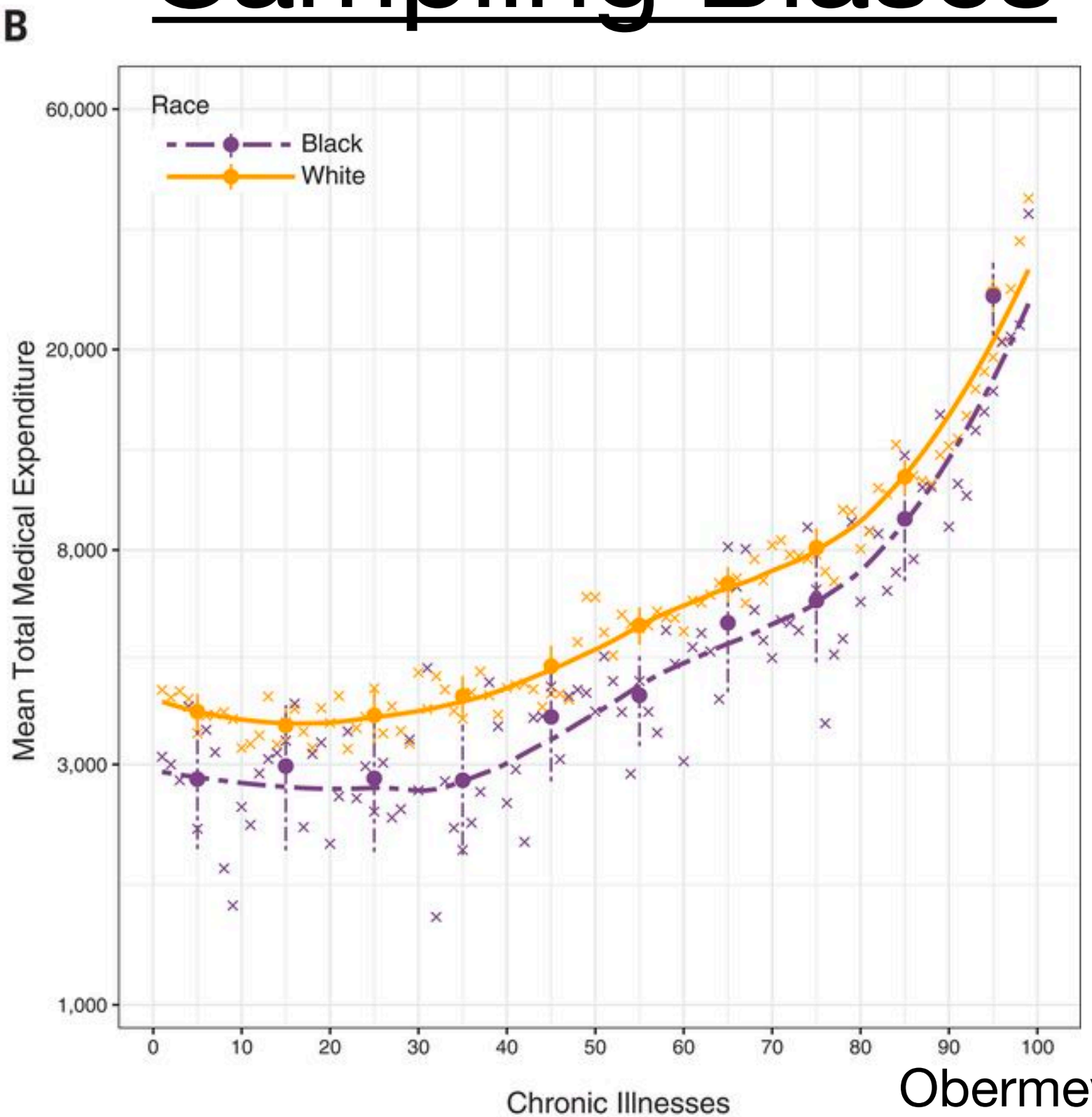
Obermeyer et al., 2019

Sampling Biases



Obermeyer et al., 2019

Sampling Biases



Obermeyer et al., 2019

Model Building Recipe

Define quantities and populations of interest

Determine feature set (much more on this later)

Determine hold-out procedure

Estimate weights on features in training data

Calculate accuracy on held-out data

Goals for NLP in Social Science

There is no "correct model"

Depends on...

Goals for NLP in Social Science

There is no "correct model"

Depends on context, domain, time frame, research question, speakers, goals, sample size, measurement error, total compute, time pressure, audience

Goals for NLP in Social Science

There is no "correct model"

Depends on... a lot of things!

Think economically, start simple

Goals for NLP in Social Science

There is no "correct model"

Depends on... a lot of things!

Think economically, start simple

Is the juice worth
the squeeze?

Using NLP as a Human

NLP is not a substitute for reading
it is a complement for reading

Using NLP as a Human

NLP is not a substitute for reading
it is a complement for reading

If an expert cannot learn the model, an algorithm is unlikely to do better

Using NLP as a Human

NLP is not a substitute for reading
it is a complement for reading

If an expert cannot learn the model, an algorithm is unlikely to do better

Better training data > better algorithm
valid sample
cleaned/structured
human checked/annotated