Multiple Overimputation: A Unified Approach to Measurement Error and Missing Data

Matthew Blackwell James Honaker Gary King

July 24th, 2010

We want to convince you of three things

We want to convince you of three things

1. Measurement error is deeply problematic for political science research and current approaches are incorrect or unused.

We want to convince you of three things

1. Measurement error is deeply problematic for political science research and current approaches are incorrect or unused.
2. Missing data is the limiting, most extreme form of measurement error.

We want to convince you of three things

1. Measurement error is deeply problematic for political science research and current approaches are incorrect or unused.
2. Missing data is the limiting, most extreme form of measurement error.
3. We can rework the multiple imputation framework to simultaneously correct for both missing data and measurement error.

PART ONE:

How we deal with measurement error.

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	5	6	6.23	5.92
2	LIBERIA	NA	3	NA	NA
3	SIERRA LEONE	3	3	6.60	NA
4	GHANA	9	6	6.86	12.68
5	TOGO	NA	5	6.27	17.34
6	CAMEROON	6	5	6.93	15.47
7	NIGERIA	5	7	6.88	17.46
8	GABON	6	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	5	6	6.23	5.92
2	LIBERIA	NA	3	NA	NA
3	SIERRA LEONE	3	3	6.60	NA
4	GHANA	9	6	6.86	12.68
5	TOGO	NA	5	6.27	17.34
6	CAMEROON	6	5	6.93	15.47
7	NIGERIA	≈5	7	6.88	17.46
8	GABON	6	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	≈5	6	6.23	5.92
2	LIBERIA	NA	3	NA	NA
3	SIERRA LEONE	≈3	3	6.60	NA
4	GHANA	≈9	6	6.86	12.68
5	TOGO	NA	5	6.27	17.34
6	CAMEROON	≈6	5	6.93	15.47
7	NIGERIA	≈5	7	6.88	17.46
8	GABON	≈6	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	≈5	≈6	6.23	5.92
2	LIBERIA	NA	≈3	NA	NA
3	SIERRA LEONE	≈3	≈3	6.60	NA
4	GHANA	≈9	≈6	6.86	12.68
5	TOGO	NA	≈6	6.27	17.34
6	CAMEROON	≈6	≈5	6.93	15.47
7	NIGERIA	≈5	≈7	6.88	17.46
8	GABON	≈6	≈8	8.19	16.97

Google scholar
measurement error

- Instrumental variables

- Instrumental variables
- Regression calibration

- Instrumental variables
- Regression calibration
- SIMEX

- Instrumental variables
- Regression calibration
- SIMEX
- Semiparametric models

- Instrumental variables
- Regression calibration
- SIMEX
- Semiparametric models
- Mixture models

- Instrumental variables
- Regression calibration
- SIMEX
- Semiparametric models
- Mixture models
- Quasi-likelihood models

- Instrumental variables
- Regression calibration
- SIMEX
- Semiparametric models
- Mixture models
- Quasi-likelihood models
- Denial

- Denial
- Denial
- Denial
- Denial
- Denial
- Denial
- Denial

Why is this the state of the art?

Most existing approaches are

Most existing approaches are application-specific.

Most existing approaches are application-specific.
model dependent.

Most existing approaches are

application-specific.
model dependent.
difficult to implement.

Most existing approaches are

application-specific.
model dependent.
difficult to implement.
inapplicable with multiple variables.

Most existing approaches are

application-specific.
model dependent.
difficult to implement.
inapplicable with multiple variables.
invalid with heteroskadastic errors.

Most existing approaches are

application-specific.
model dependent.
difficult to implement.
inapplicable with multiple variables.
invalid with heteroskadastic errors.
unusable with missing data.

Why is this the state of the art?

Why is this the state of the art?
It's easy and tolerated.

Why is this the state of the art?
It's easy and tolerated.
But it's make believe.

EMBRACE YOUR BAD DATA

EMBRACE YOUR BAD DATA

FIX IT.

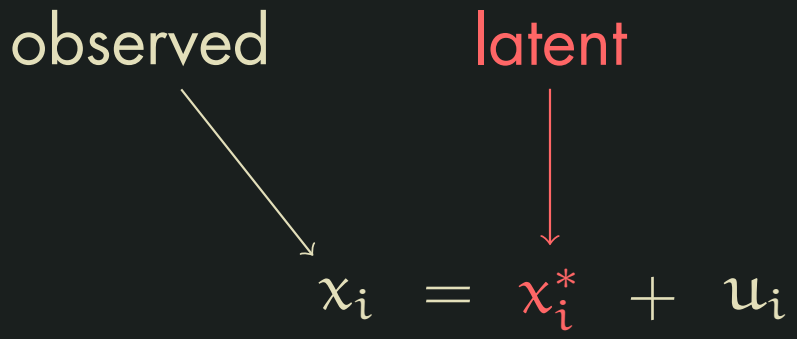
PART TWO:

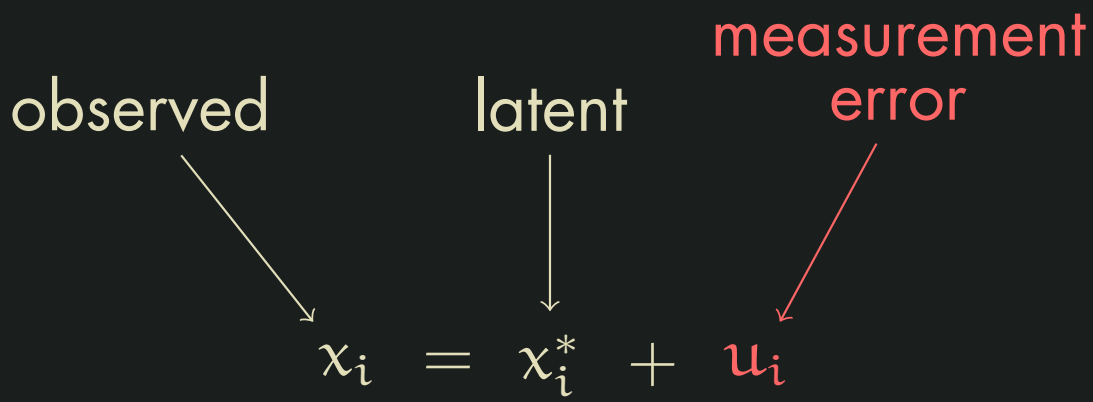
A Brief Review of Measurement Error.

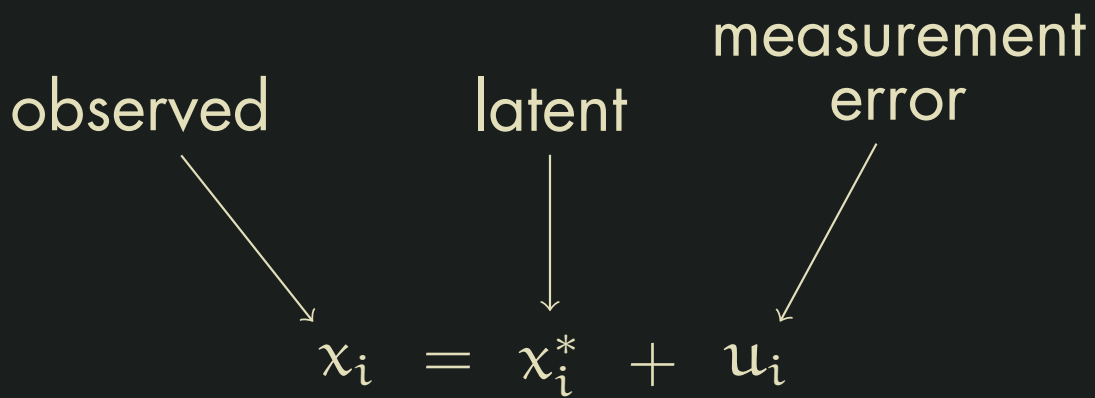
$$x_i = x_i^* + u_i$$

observed

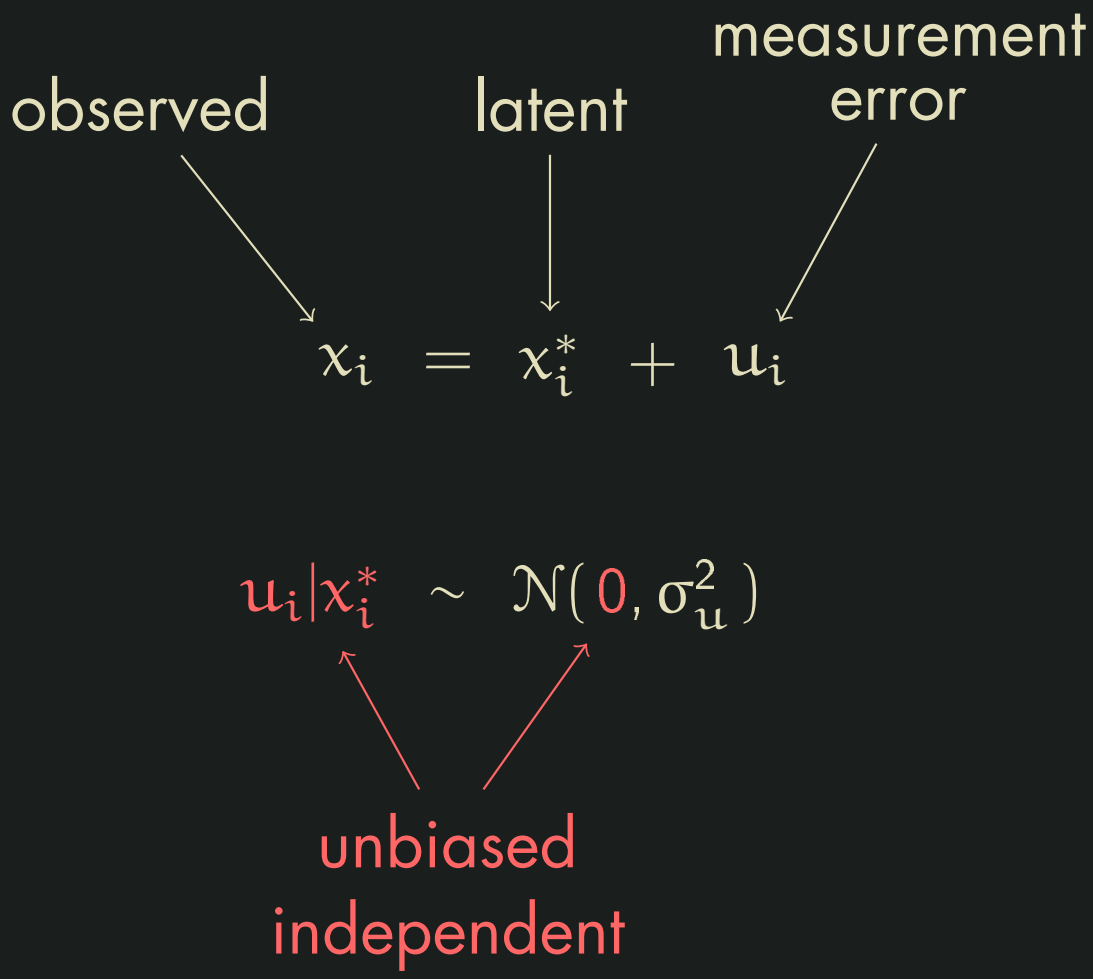
$$x_i = x_i^* + u_i$$

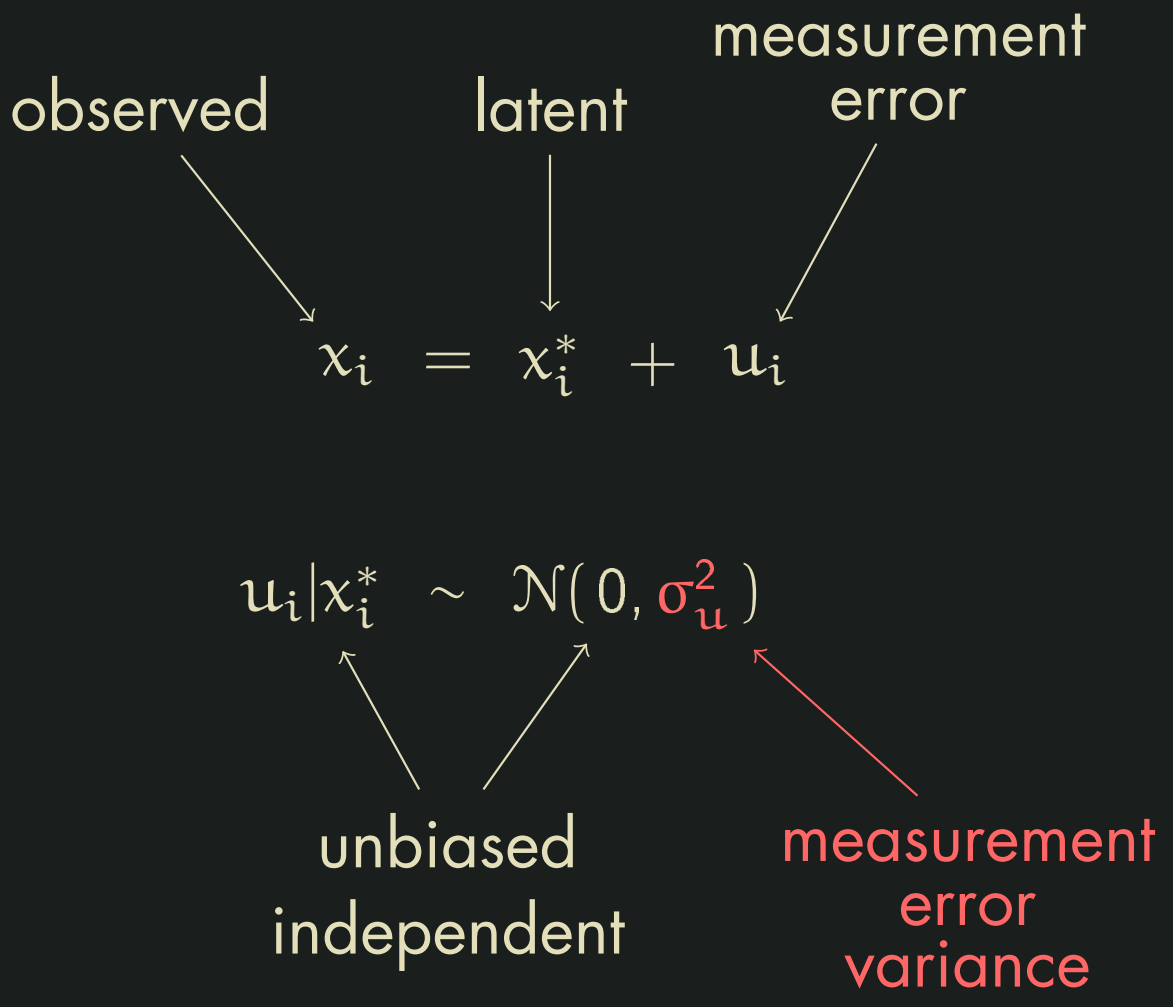

observed
latent
$x_i = x_i^* + u_i$

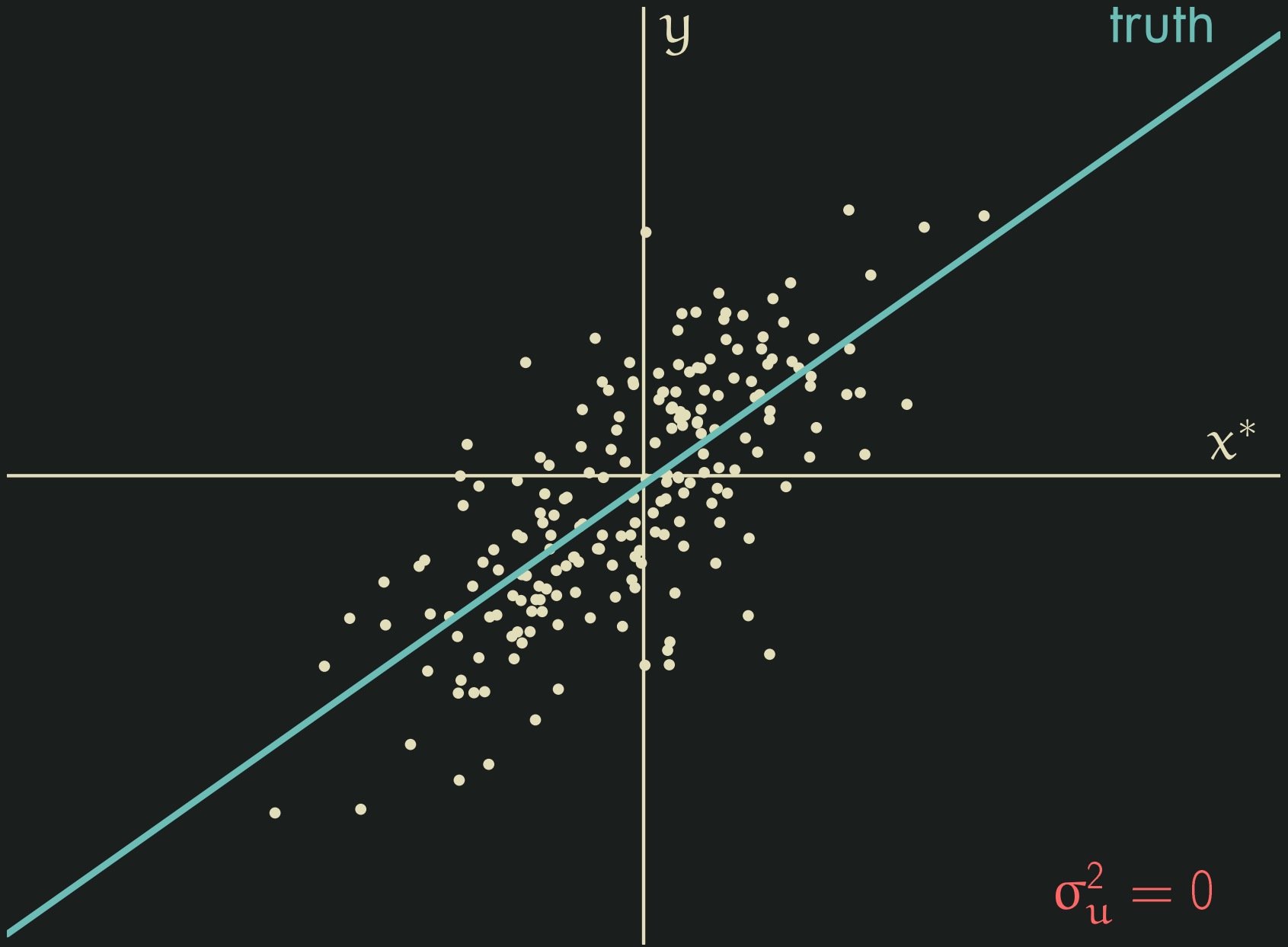

observed
latent
measurement error
$x_i = x_i^* + u_i$

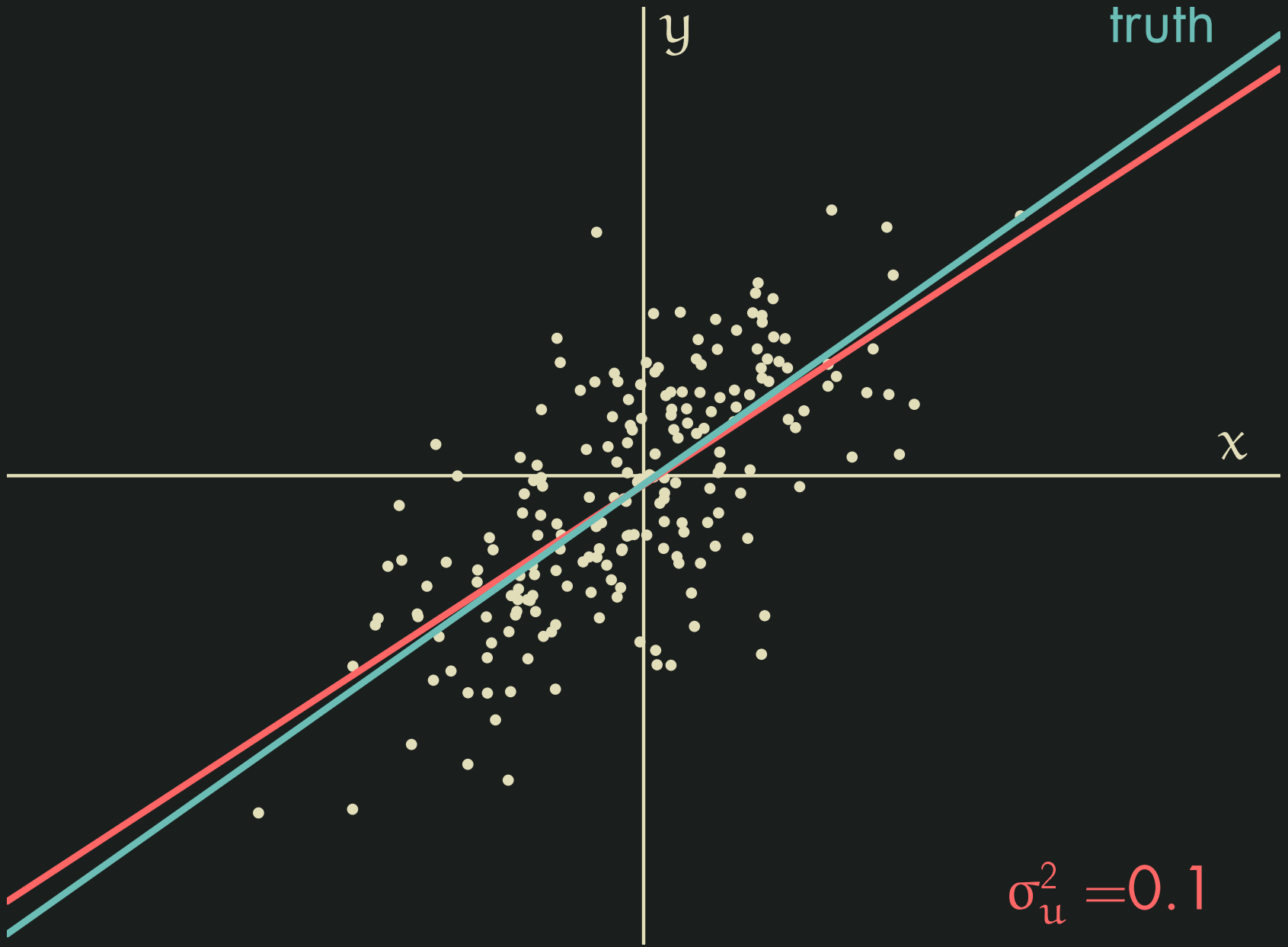


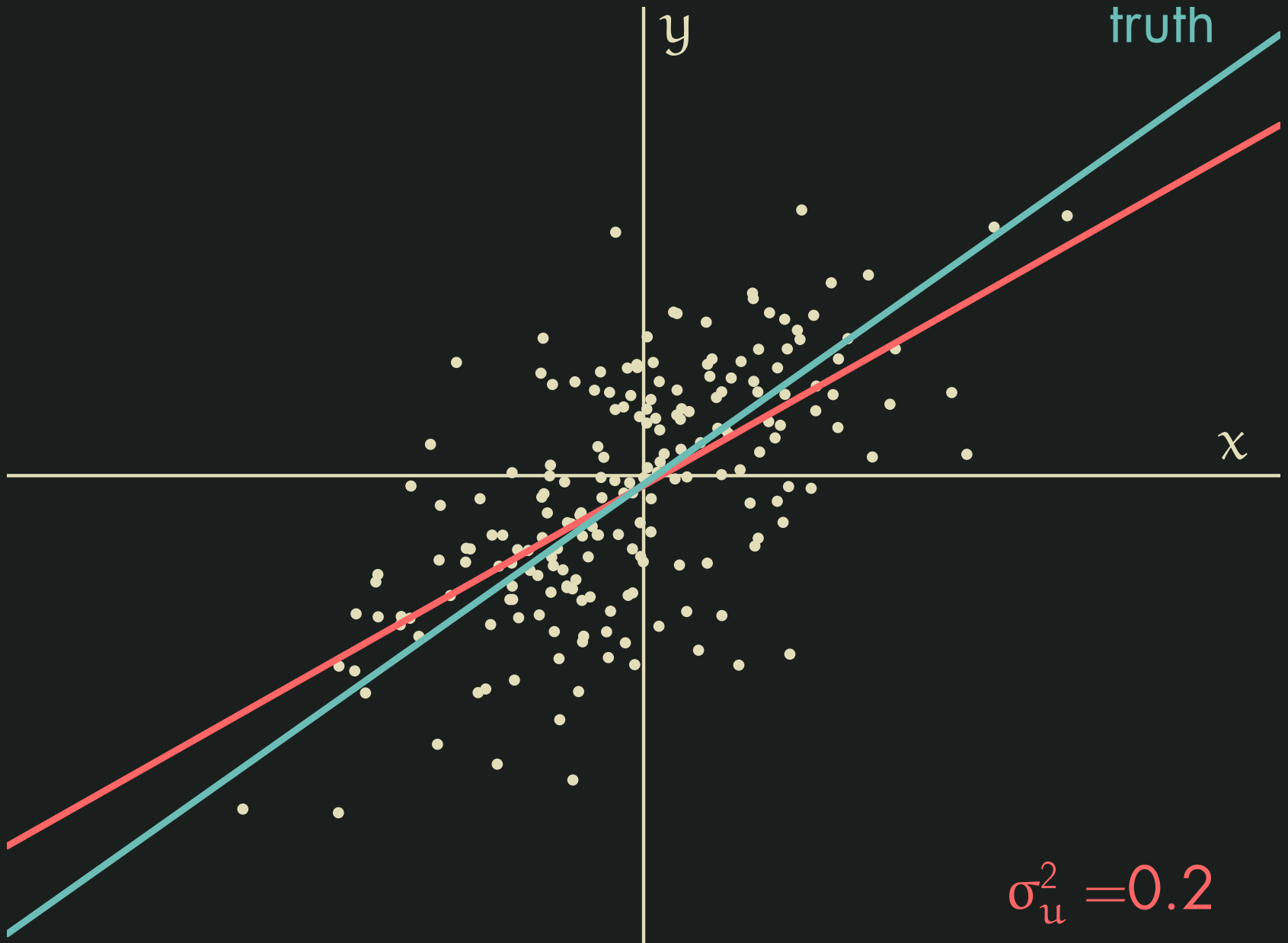
$$u_i | x_i^* \sim \mathcal{N}(0, \sigma_u^2)$$

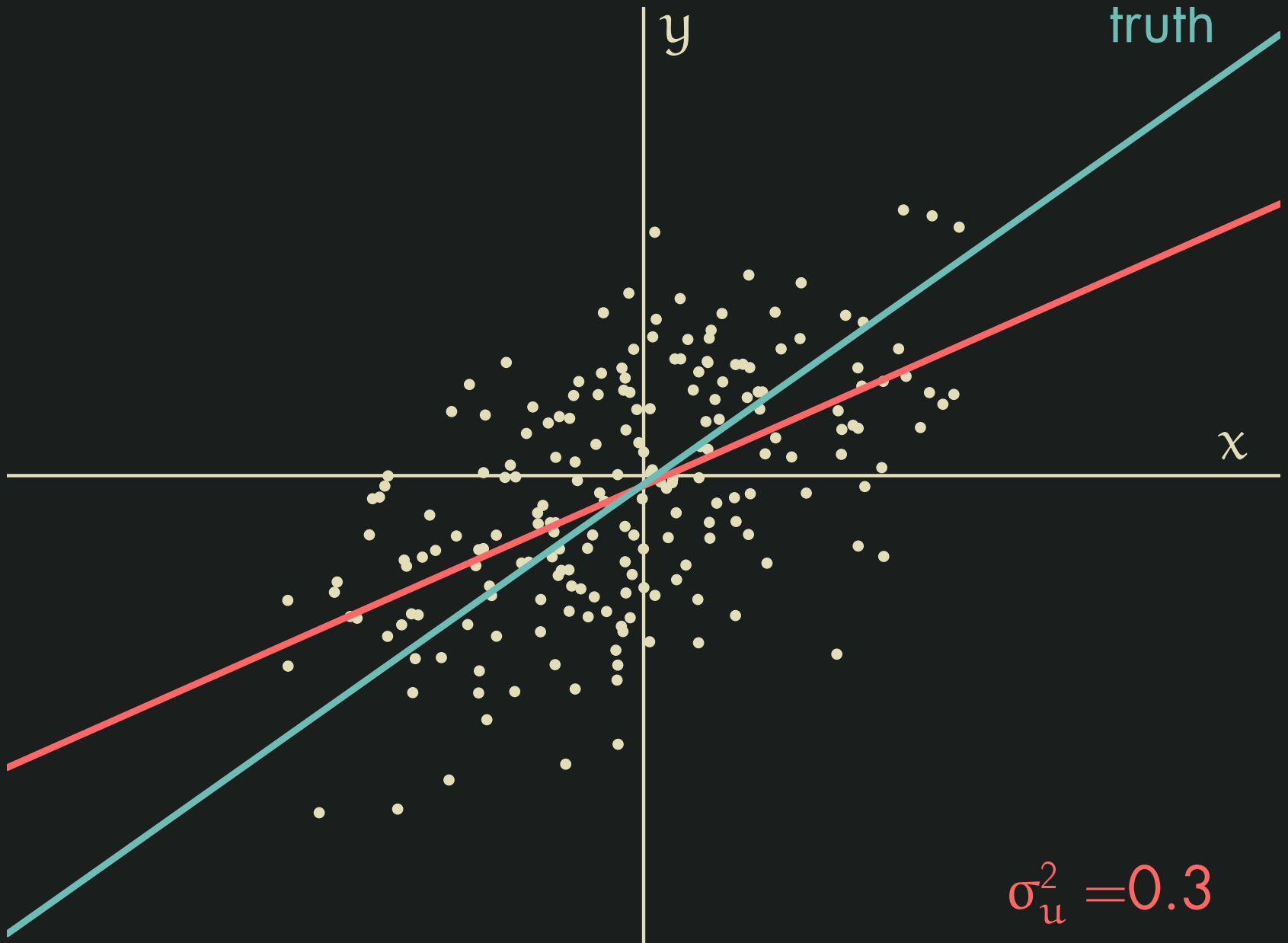

observed
latent
measurement error
$x_i = x_i^* + u_i$
$u_i | x_i^* \sim \mathcal{N}(0, \sigma_u^2)$
unbiased
independent

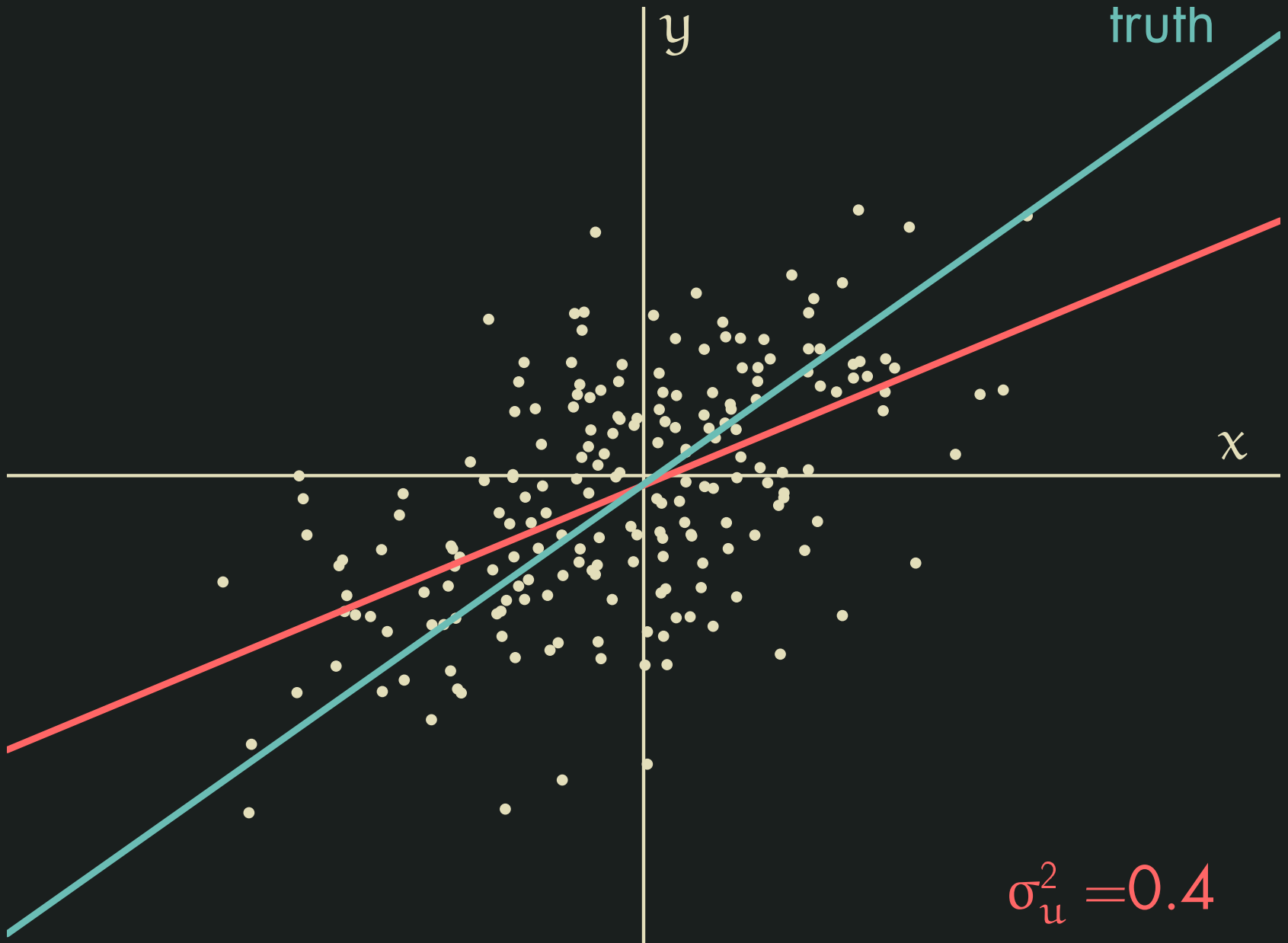

observed
latent
measurement error
$x_i = x_i^* + u_i$
$u_i | x_i^* \sim \mathcal{N}(0, \sigma_u^2)$
unbiased independent
measurement error variance

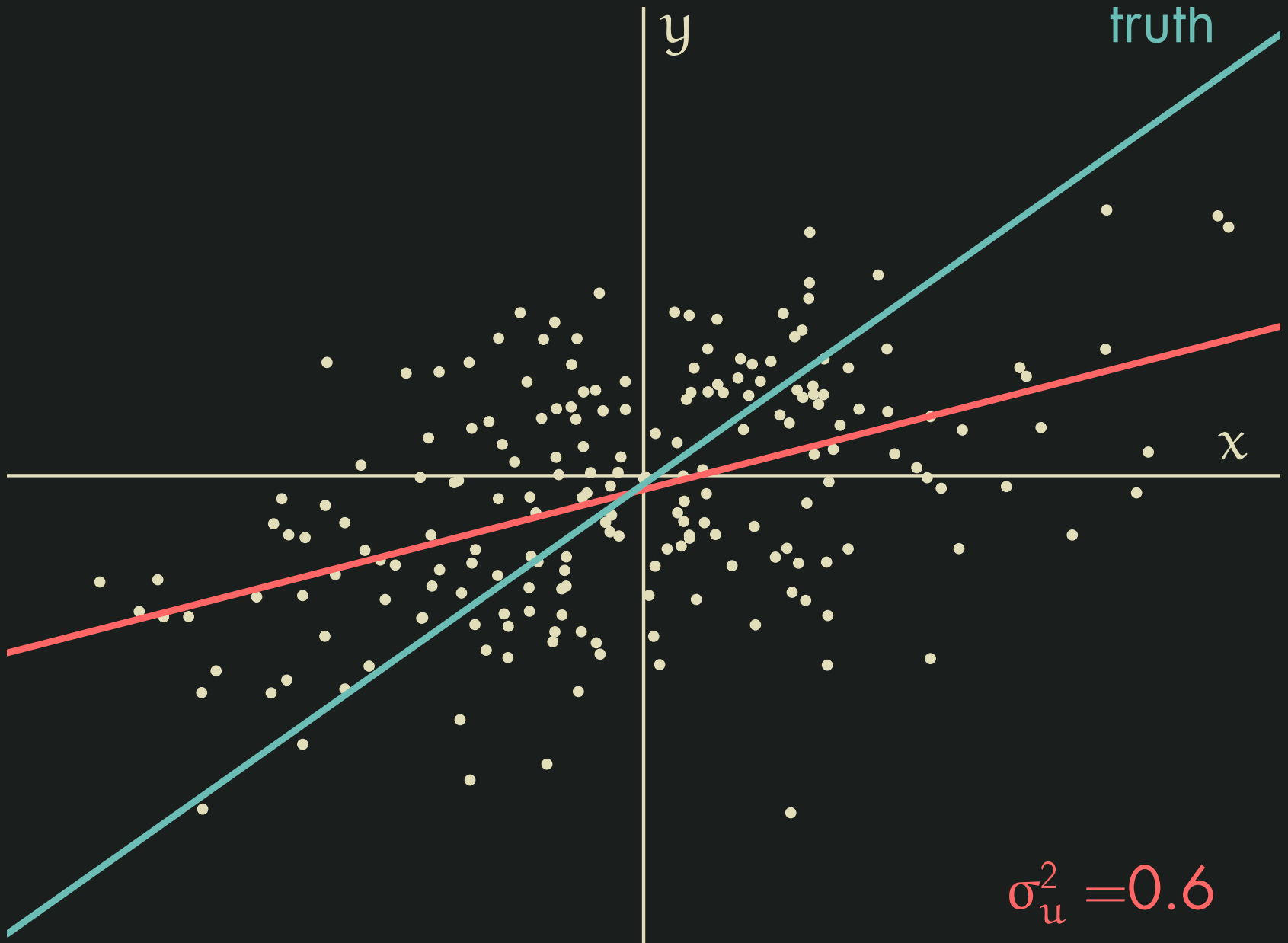

y
truth
x^*
$\sigma_u^2 = 0$

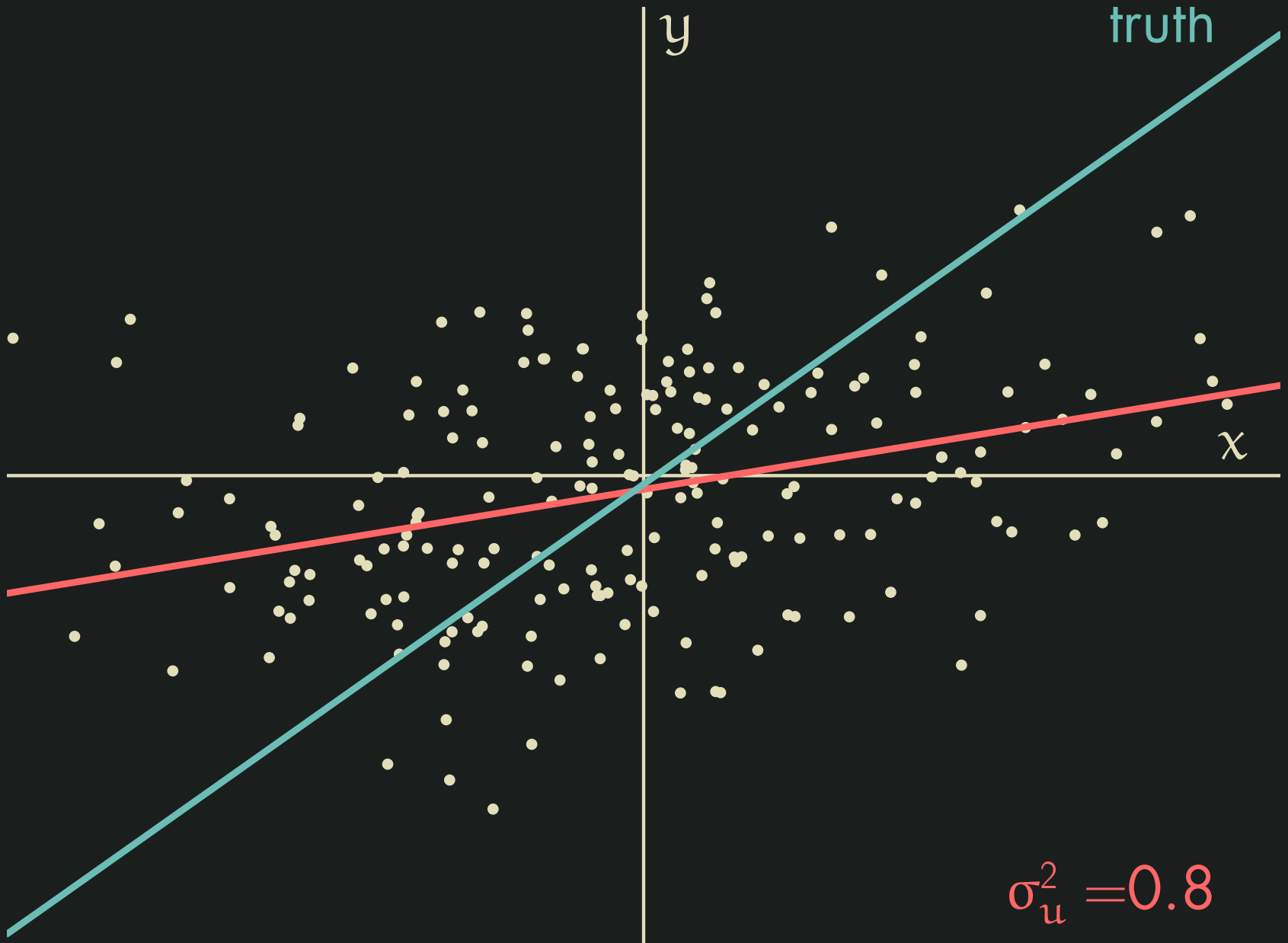

y
truth
x
$\sigma_u^2 = 0.1$

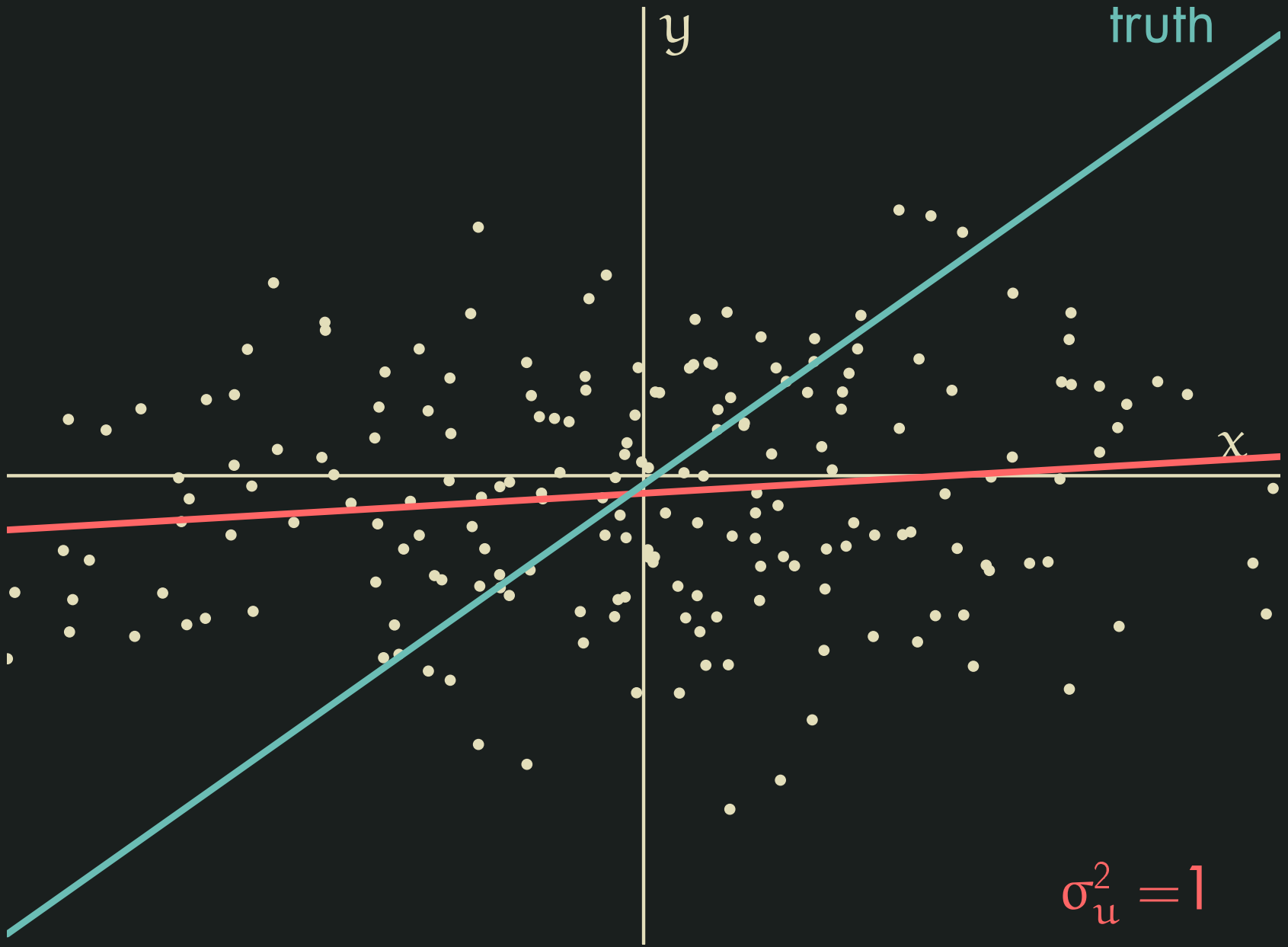

y
truth
x
$\sigma_u^2 = 0.2$

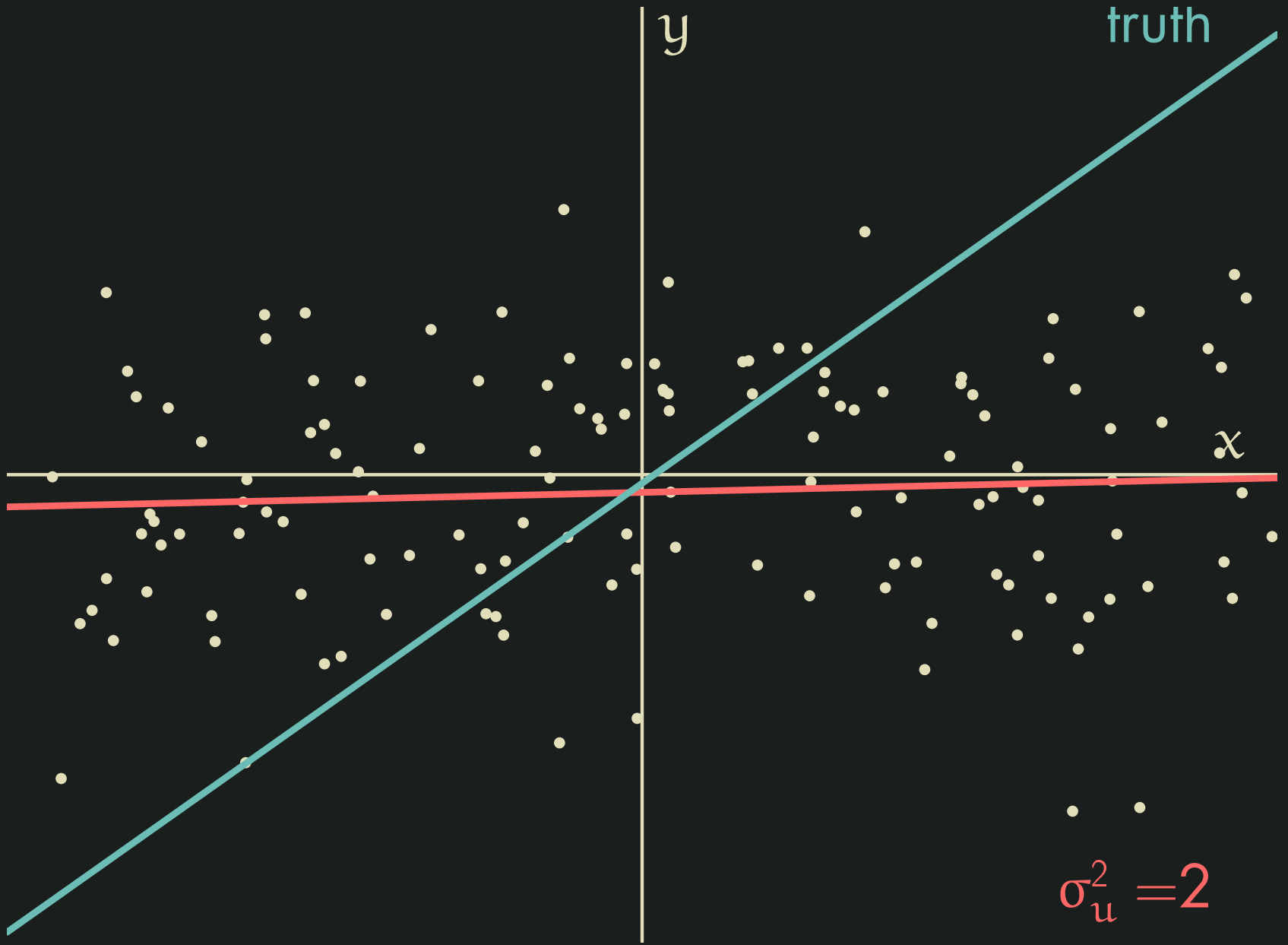

y
truth
x
$\sigma_u^2 = 0.3$


y
truth
x
$\sigma_u^2 = 0.4$


y
truth
x
$\sigma_u^2 = 0.6$


y
truth
x
$\sigma_u^2 = 0.8$


y
truth
x
$\sigma_u^2 = 1$


y
truth
x
$\sigma_u^2 = 2$

ATTENUATION

…only guaranteed in the simplest of cases:

ATTENUATION

…only guaranteed in the simplest of cases:

linear model

ATTENUATION

…only guaranteed in the simplest of cases:

linear model
one mismeasured variable

ATTENUATION

…only guaranteed in the simplest of cases:

linear model
one mismeasured variable
measurement error unrelated to other variables and x^*.

BIAS FROM MEASUREMENT ERROR

In unpredictable directions with most realistic models.

PART THREE:

The connection between missing data and measurement error.

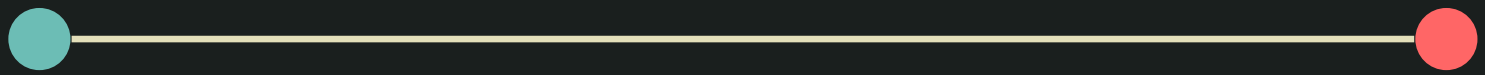
The strict dichotomy of data.

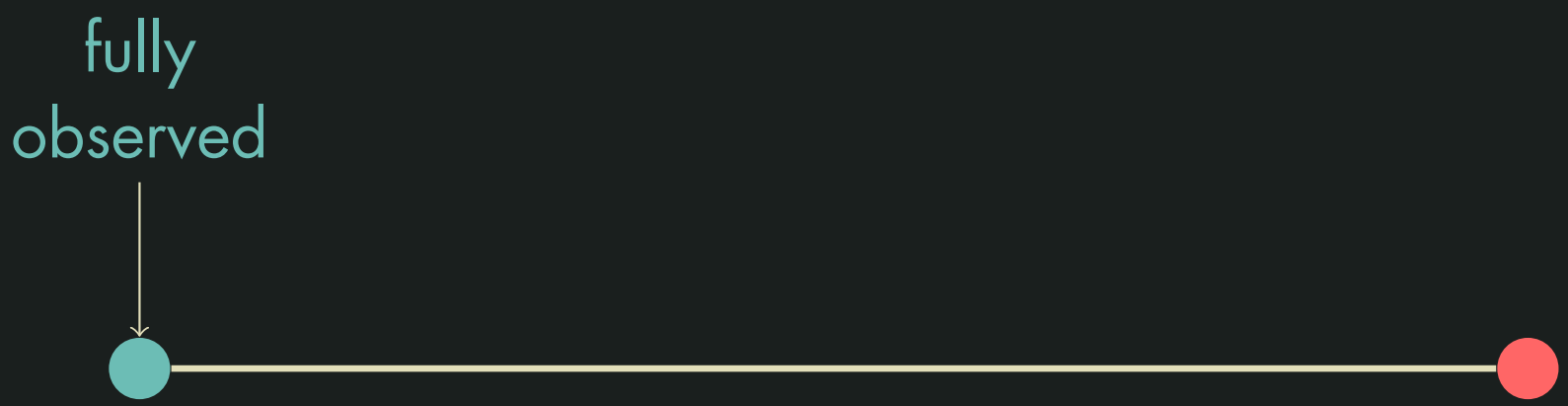
observed missing

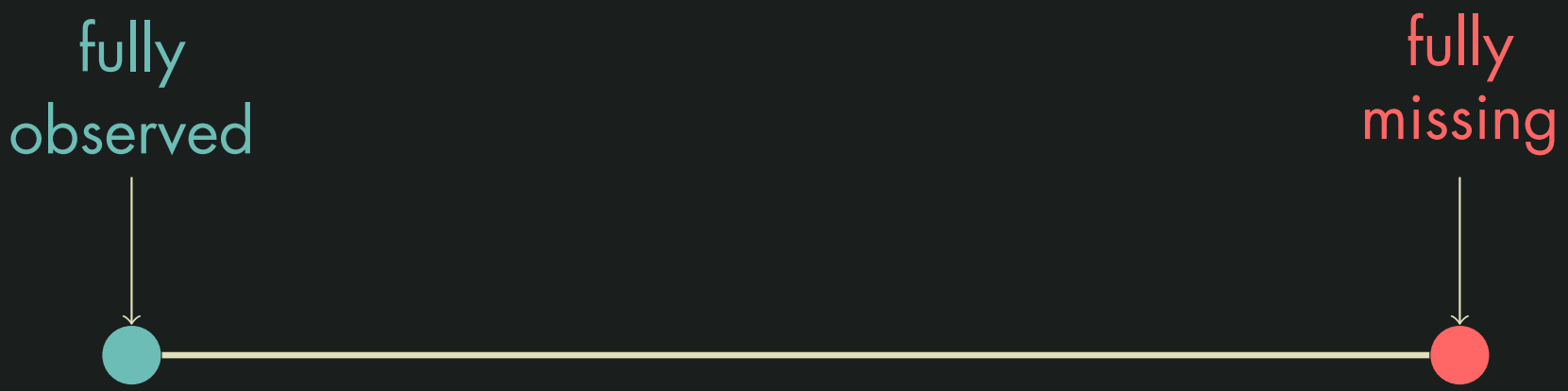
(fully) observed (fully) missing

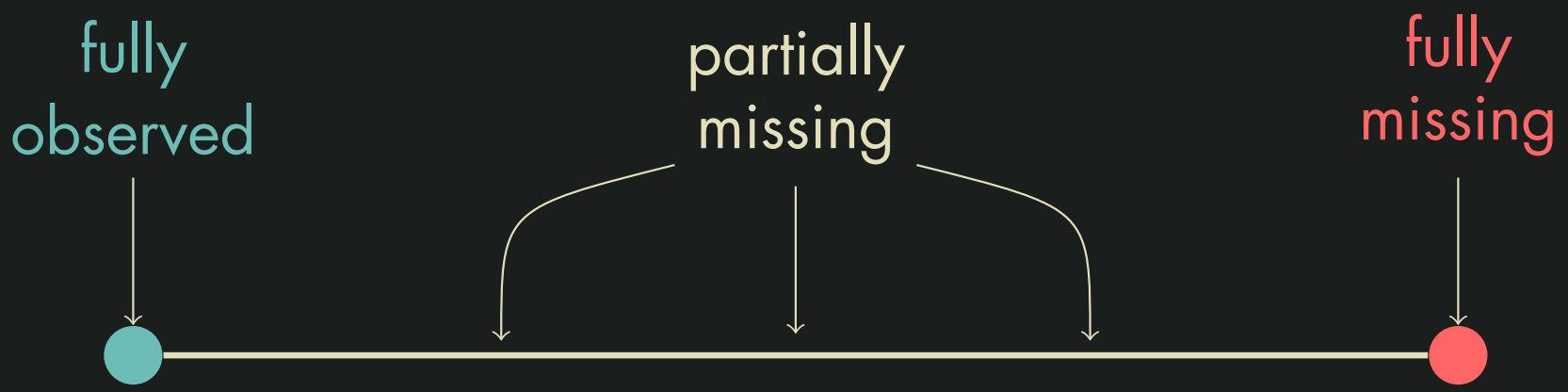
(fully) observed (fully) missing

The false dichotomy of data.

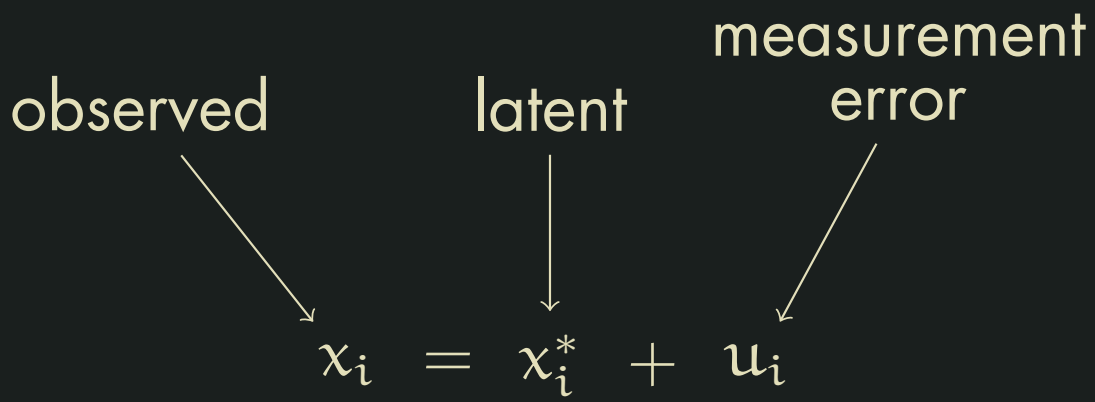



fully
observed


fully
observed
fully
missing

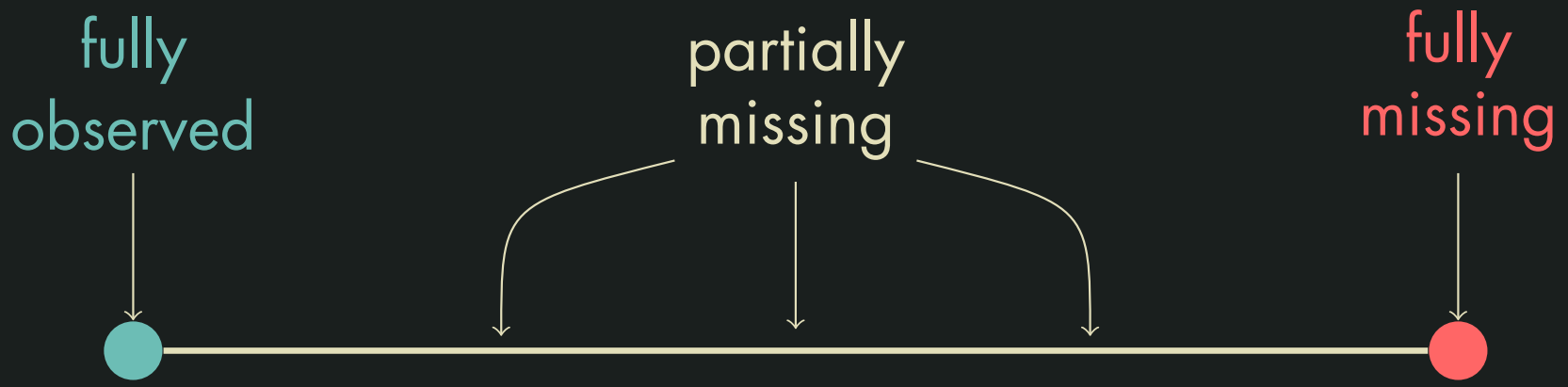

fully
observed
partially
missing
fully
missing

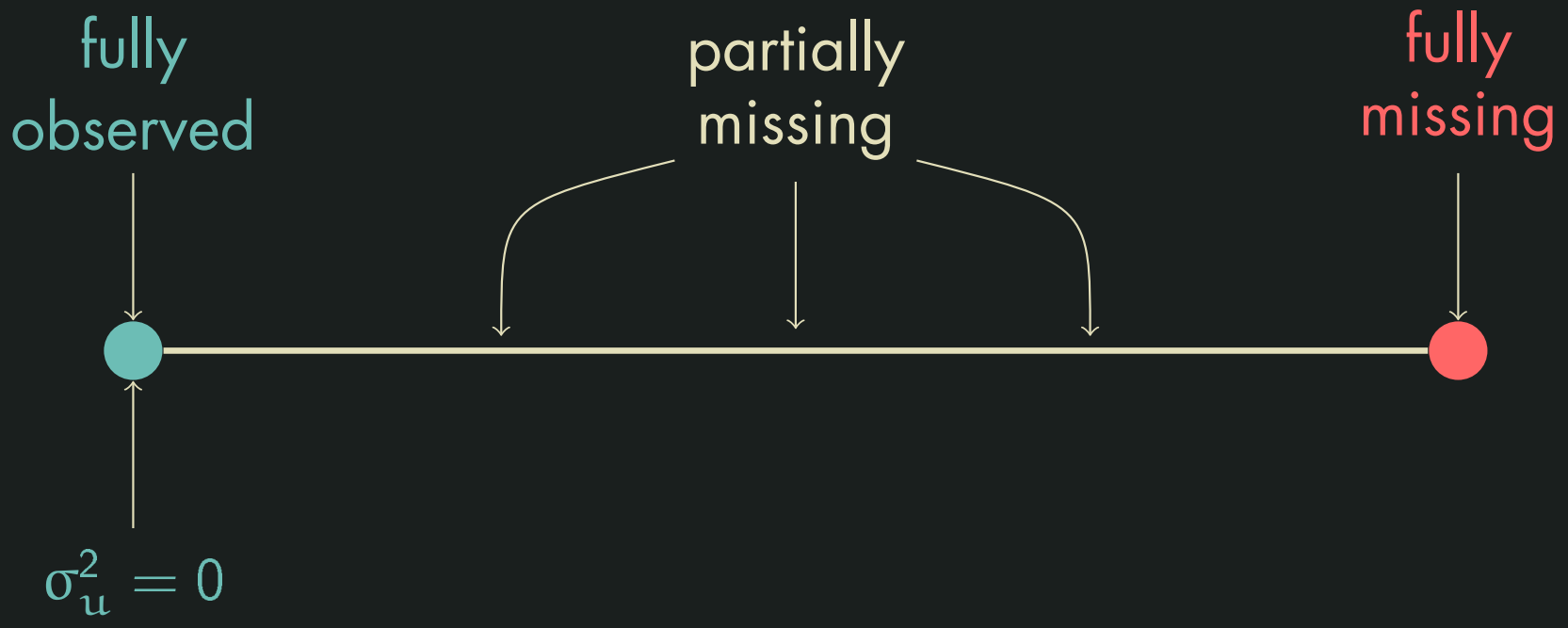
But what is this continuum?

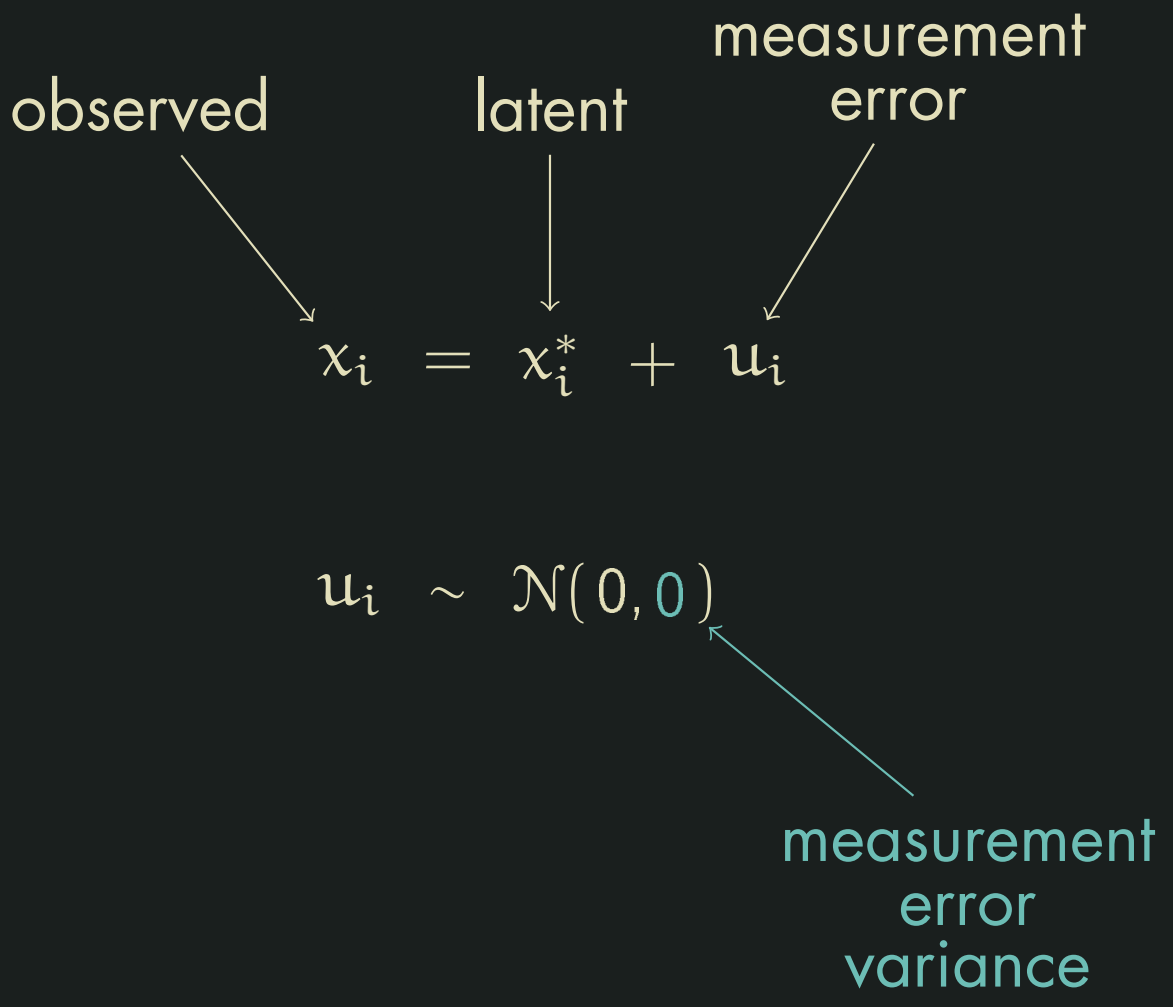


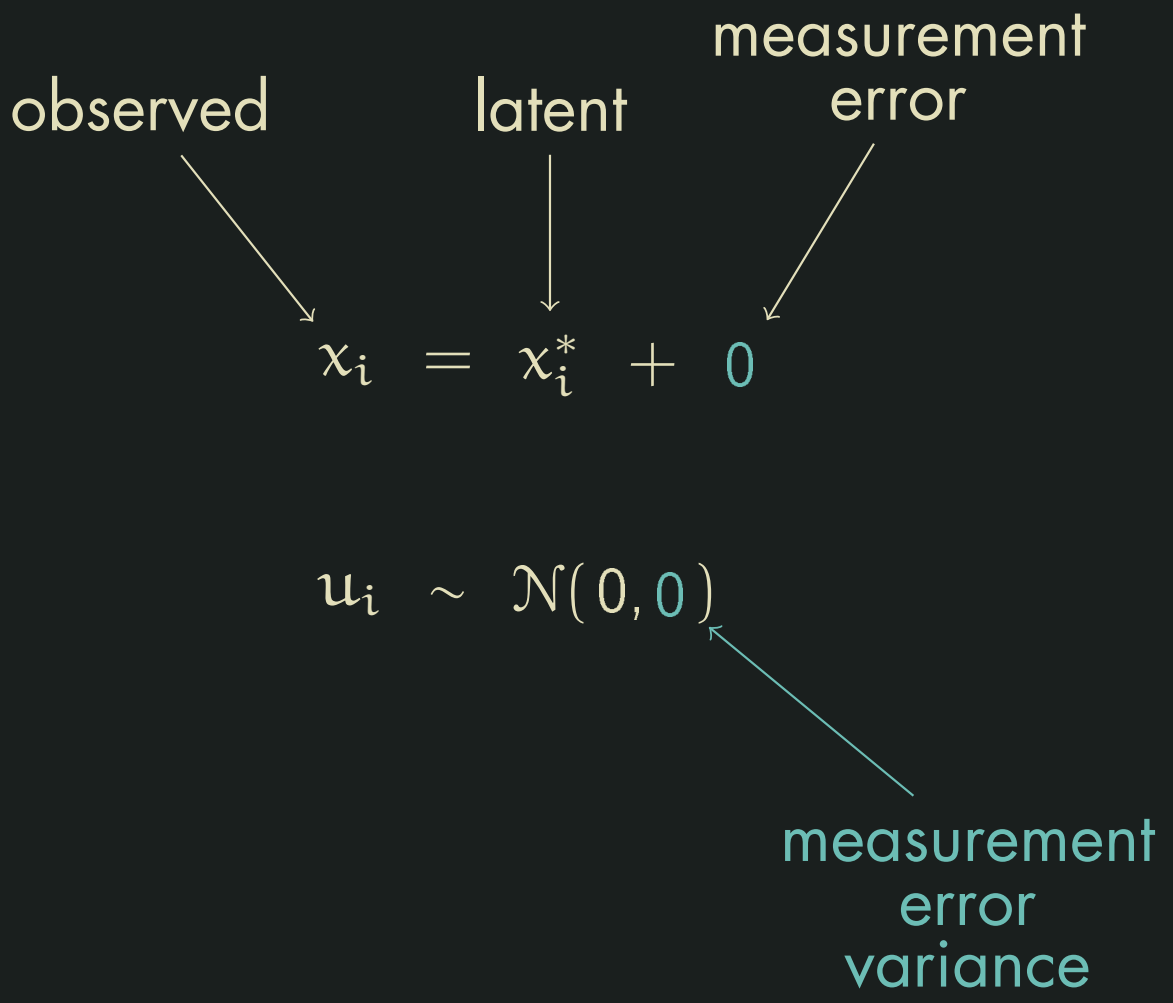
$$u_i | x_i^* \sim \mathcal{N}(0, \sigma_u^2)$$

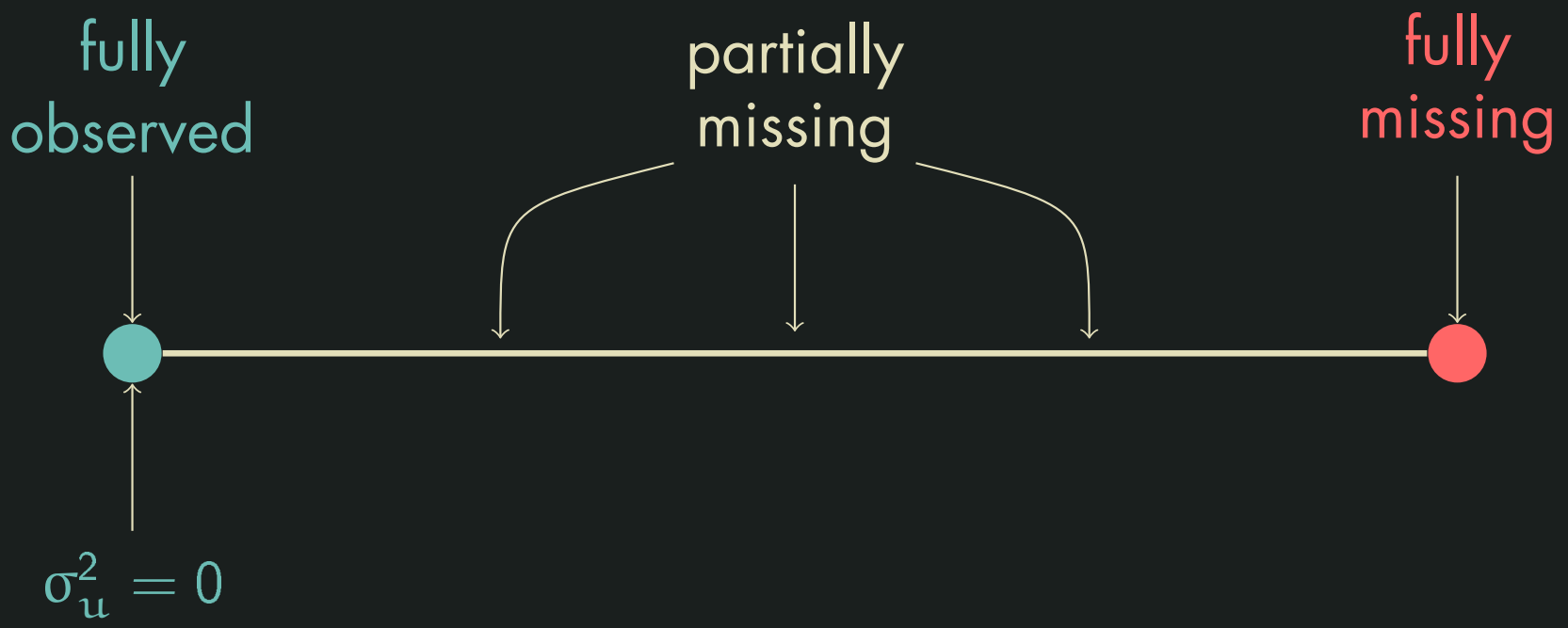
measurement error variance

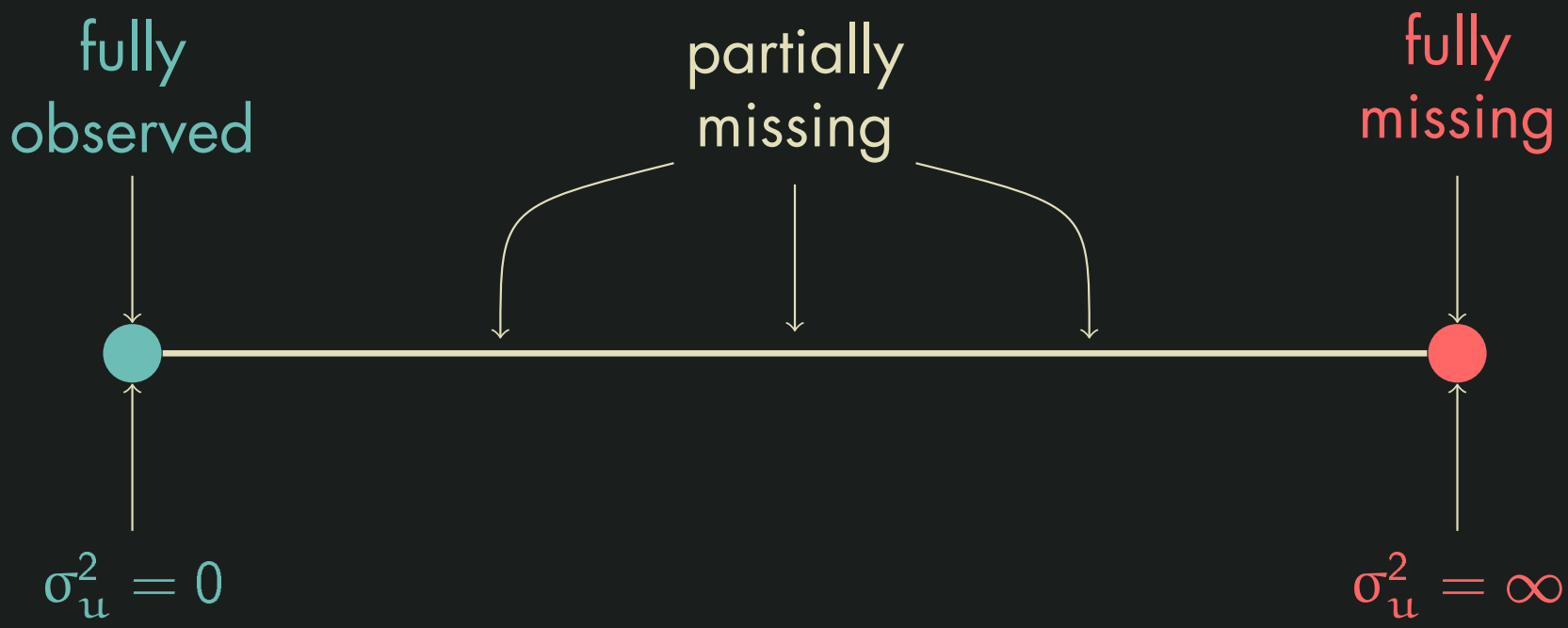

fully
observed
partially
missing
fully
missing

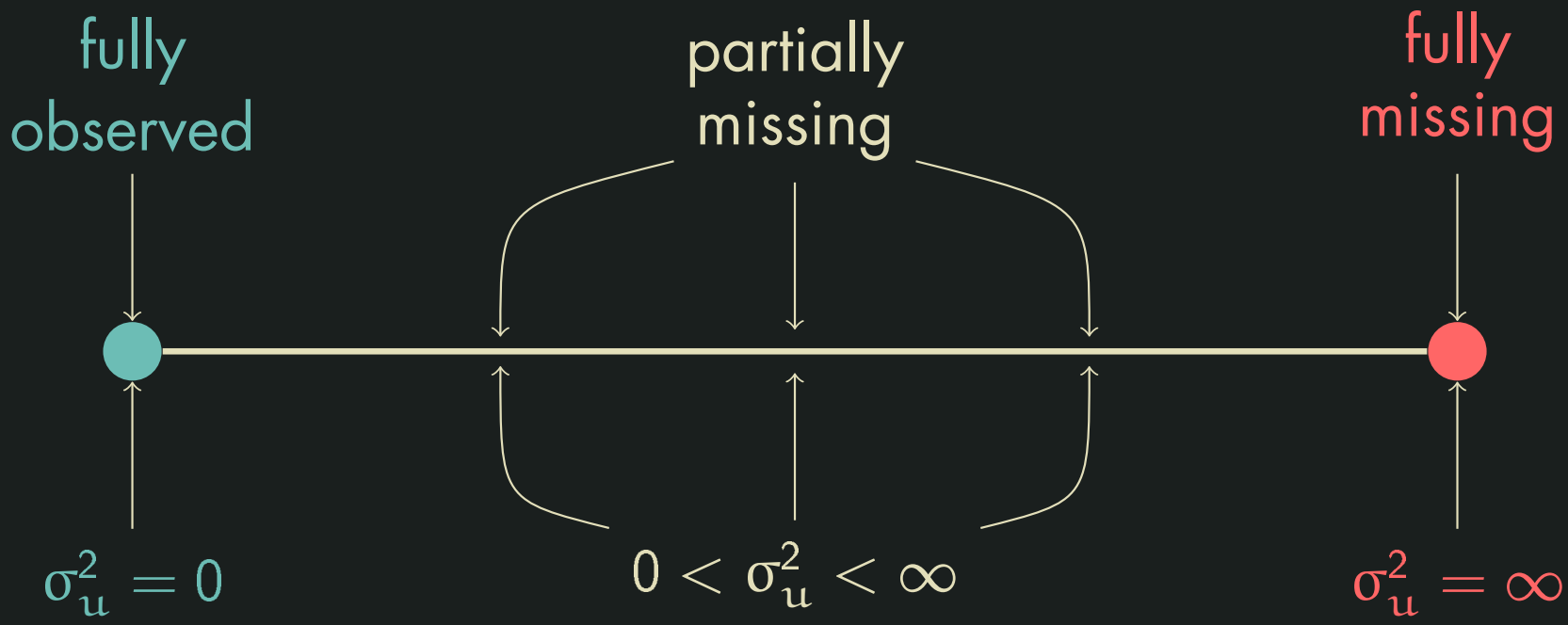

fully
observed
partially
missing
fully
missing
$\sigma_u^2 = 0$


observed
latent
measurement error
$x_i = x_i^* + u_i$
$u_i \sim \mathcal{N}(0, 0)$
measurement error variance

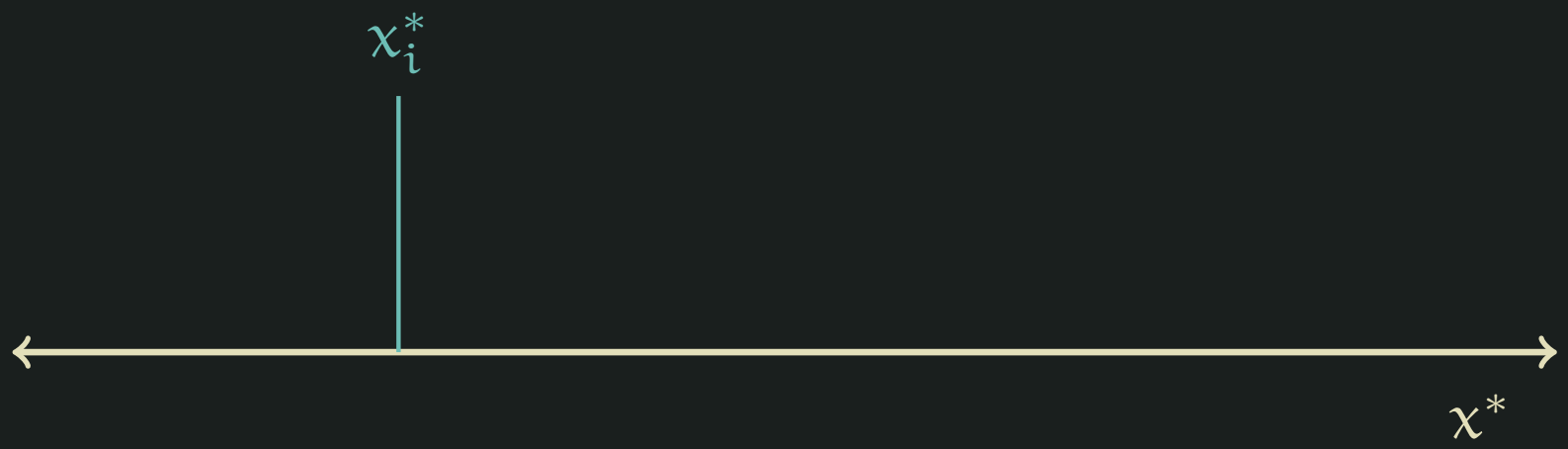

observed
latent
measurement error
$x_i = x_i^* + 0$
$u_i \sim \mathcal{N}(0, 0)$
measurement error variance

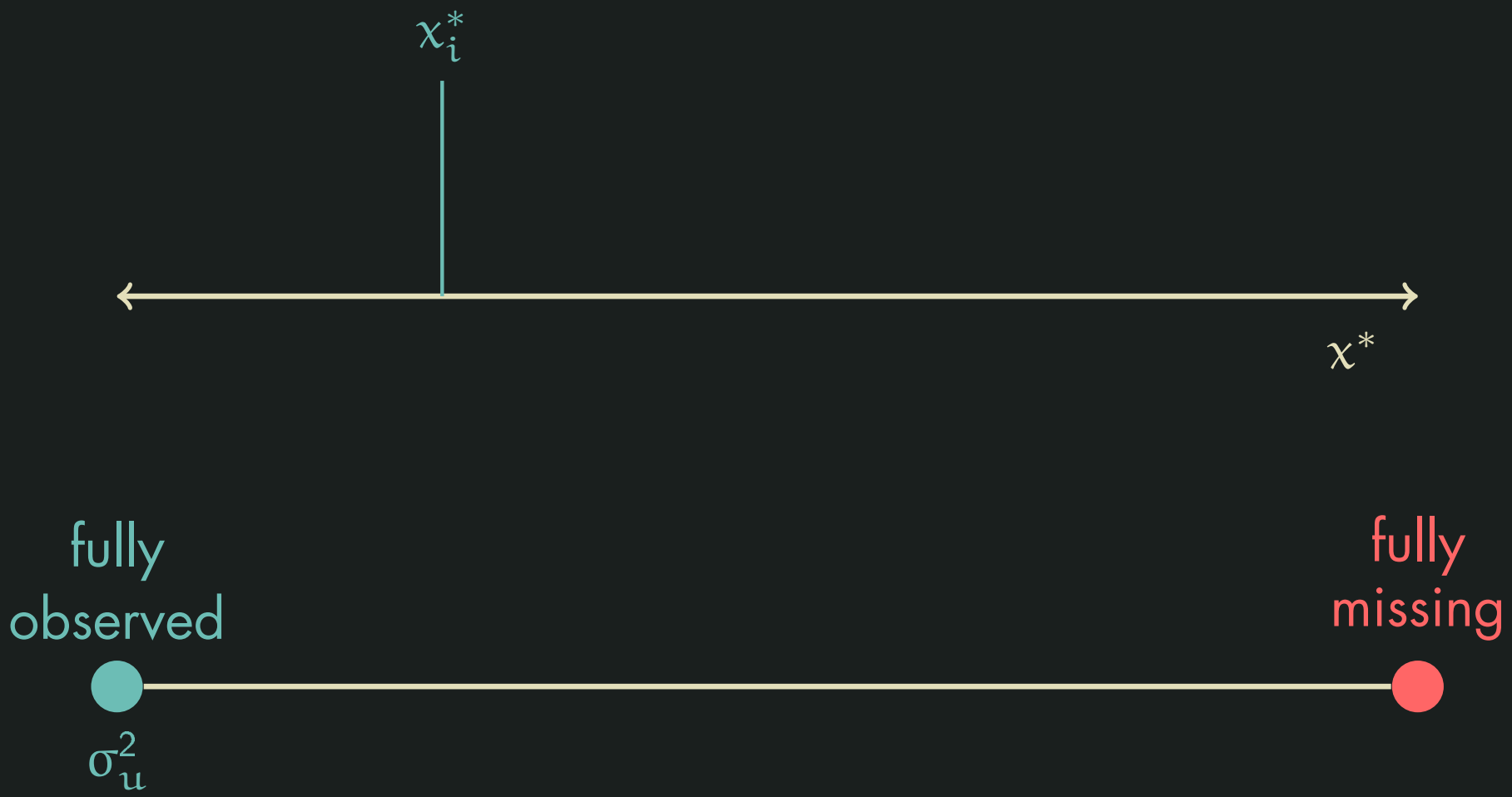

fully
observed
partially
missing
fully
missing
$\sigma_u^2 = 0$

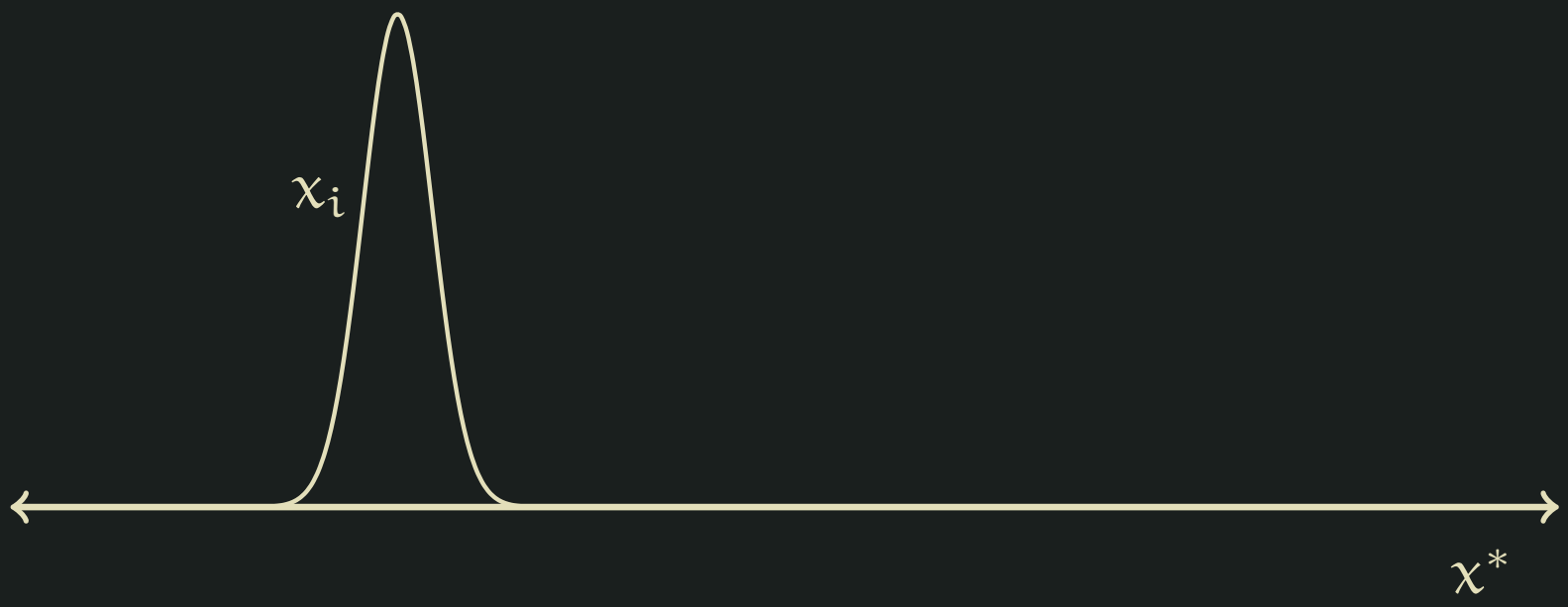

fully
observed
partially
missing
fully
missing
$\sigma_u^2 = 0$
$\sigma_u^2 = \infty$


fully
observed
partially
missing
fully
missing
$\sigma_u^2 = 0$
$0 < \sigma_u^2 < \infty$
$\sigma_u^2 = \infty$

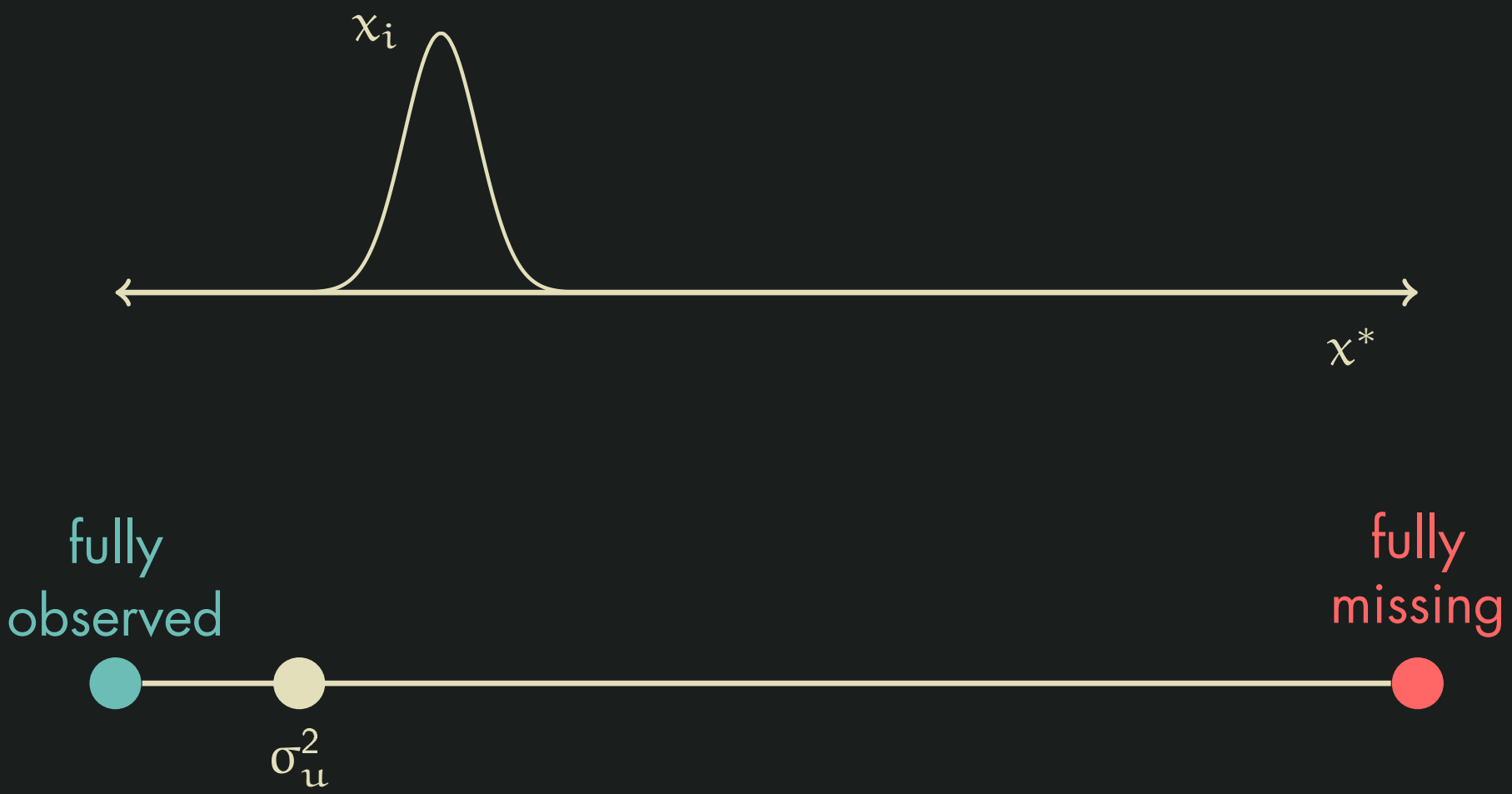
Missing data is the most
extreme case of measurement error.

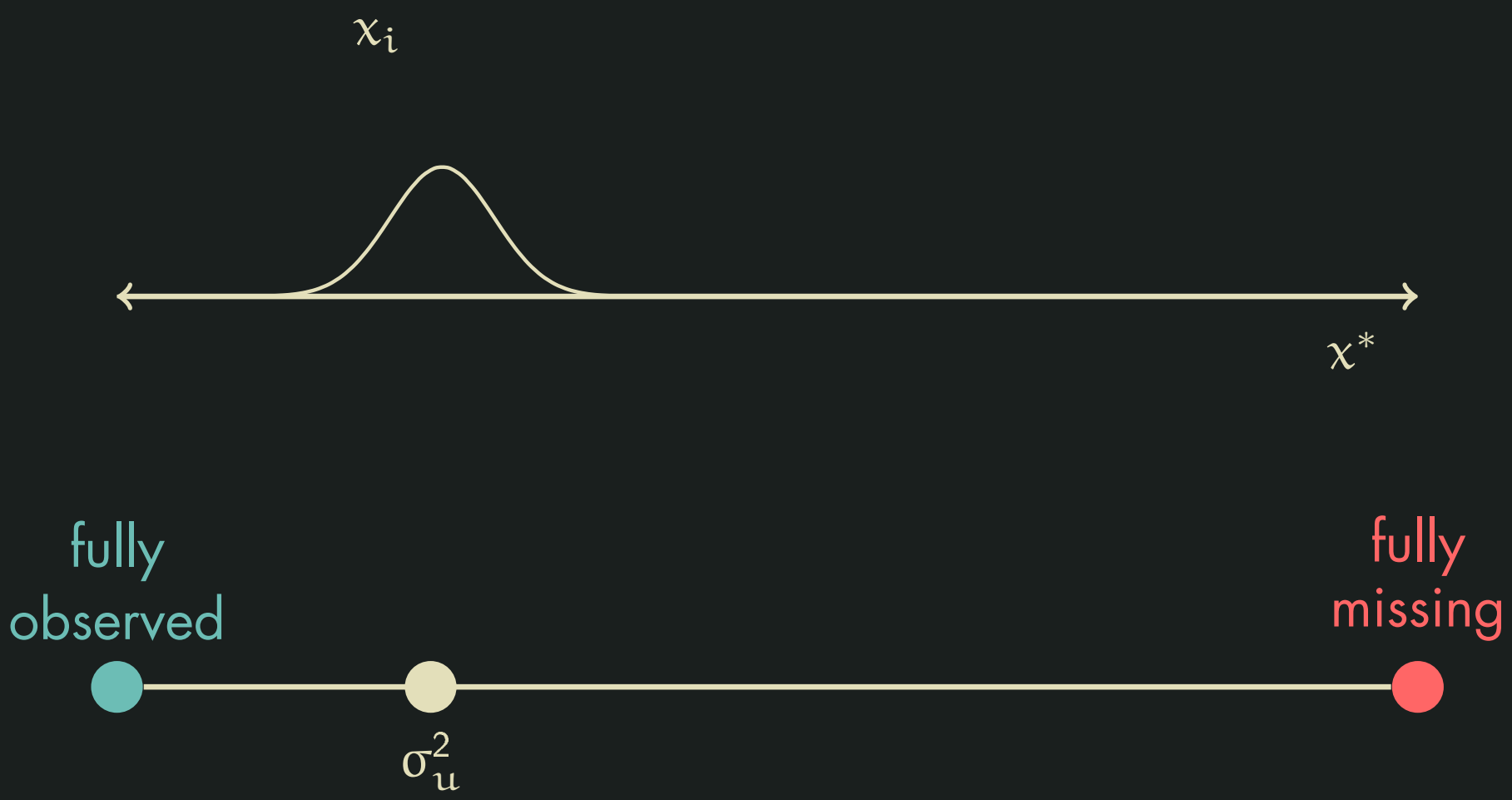

x_i^*
x^*

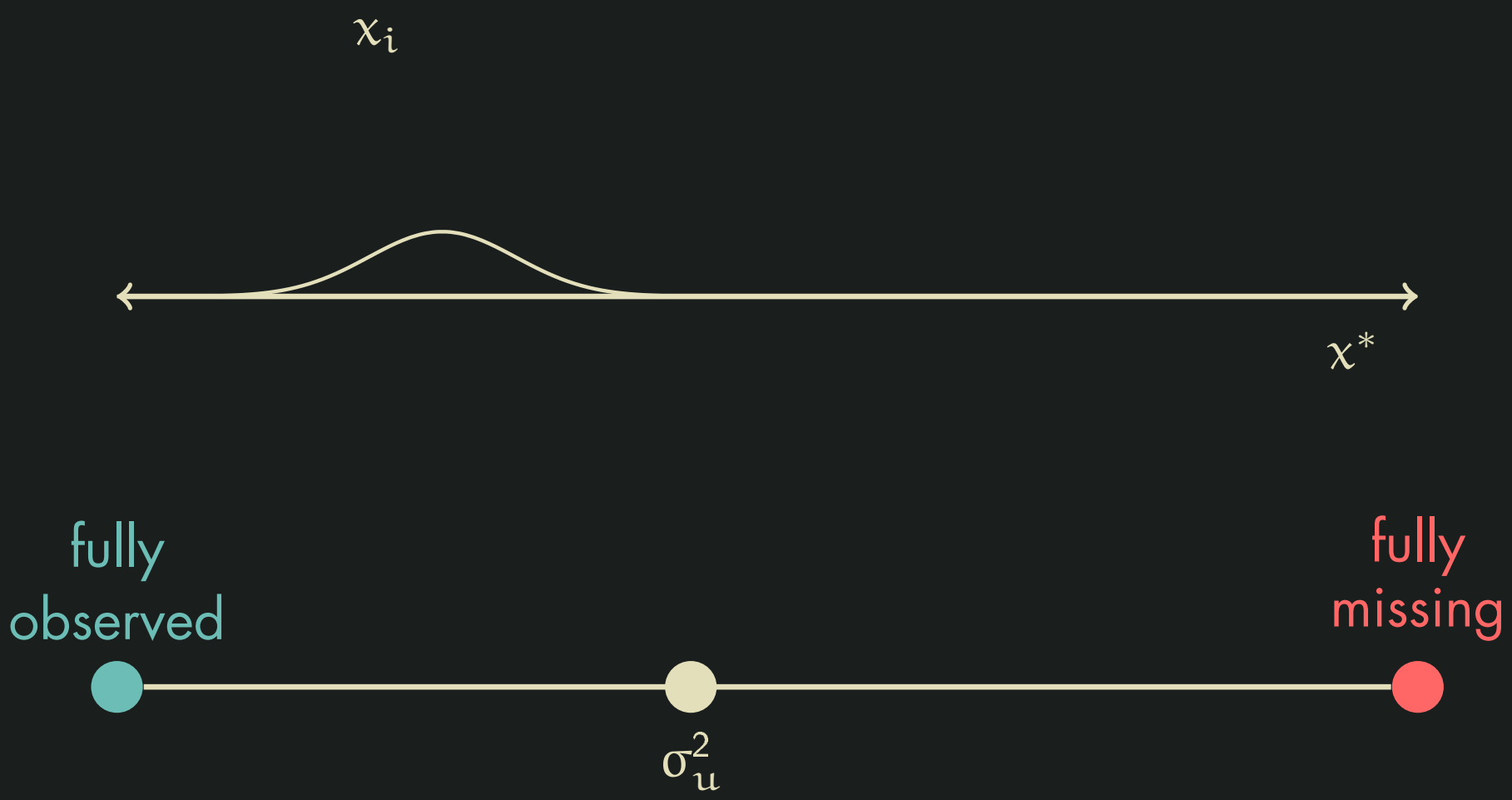

x_i^*
x^*
fully observed
fully missing
σ_u^2

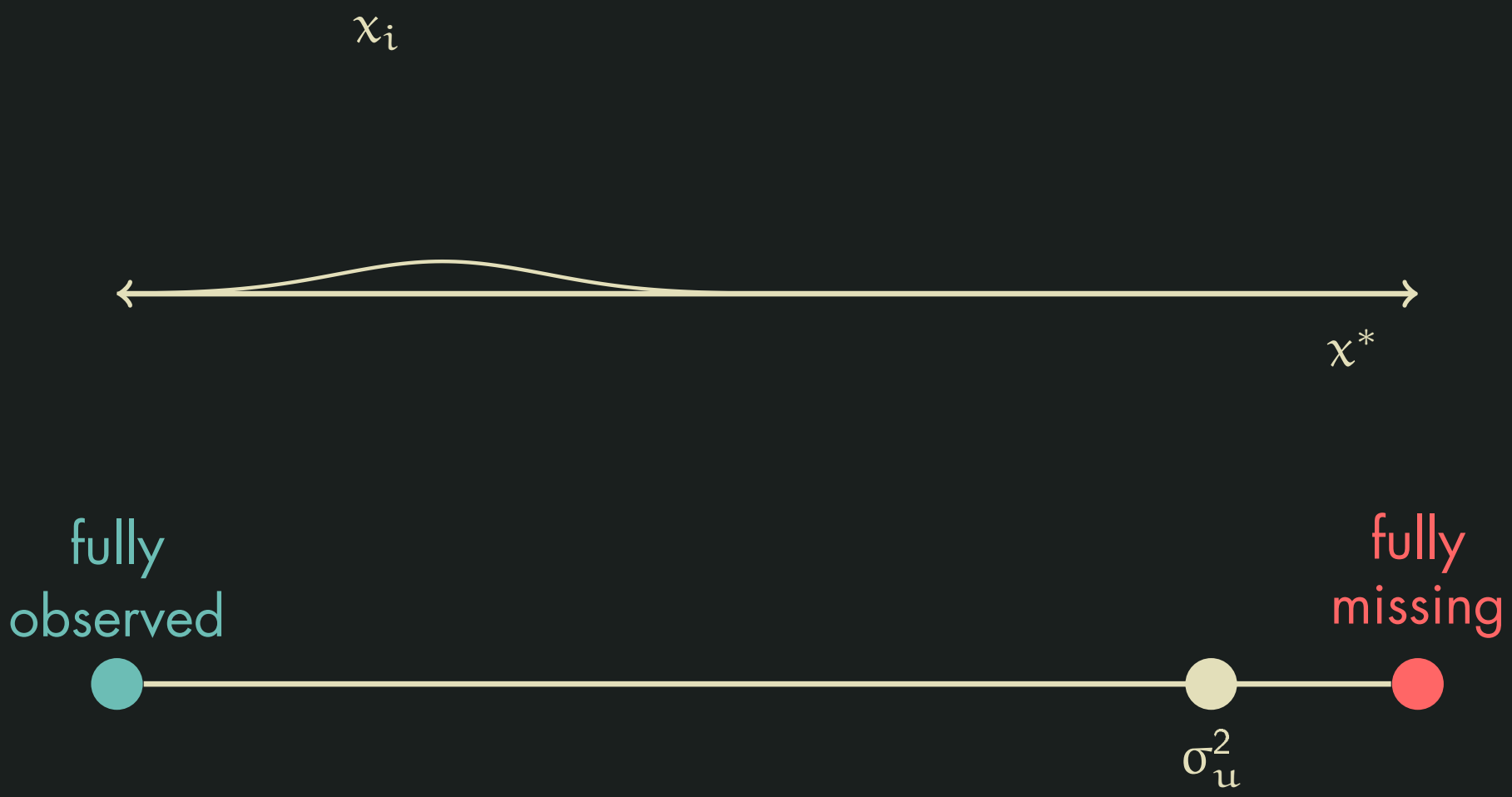

x_i
x^*

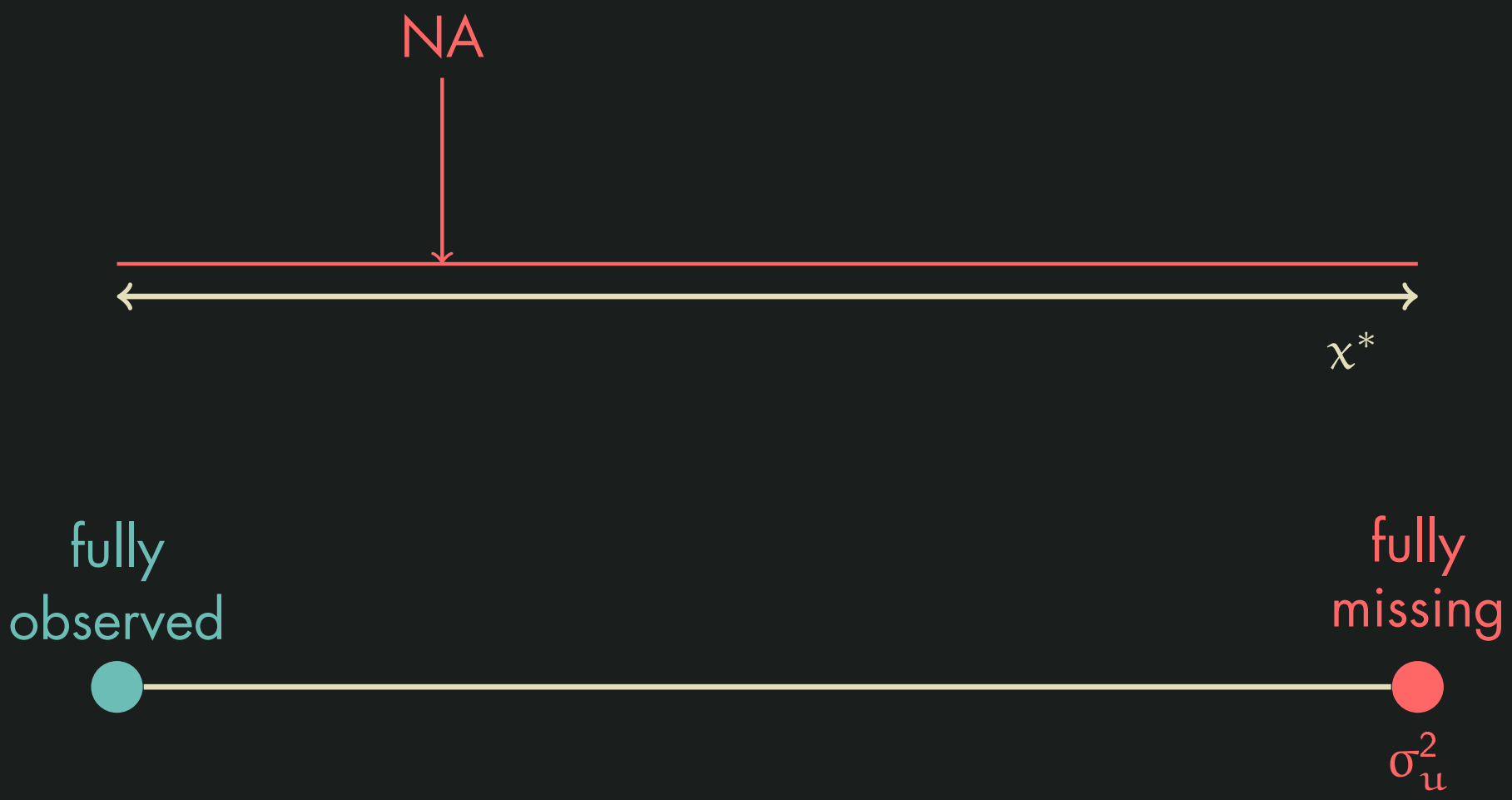
fully observed
fully missing
σ_u^2


x_i
x^*
fully observed
fully missing
σ_u^2


x_i
x^*
fully observed
fully missing
σ_u^2


x_i
x^*
fully observed
fully missing
σ_u^2


x_i
x^*
fully observed
fully missing
σ_u^2


NA
x^*
fully observed
fully missing
σ_u^2

PART FOUR:

Multiple Overimputation.

Multiple Overimputation
extends the multiple imputation framework
to correct for measurement error.

Missing Data

APPLICATION-SPECIFIC METHODS:

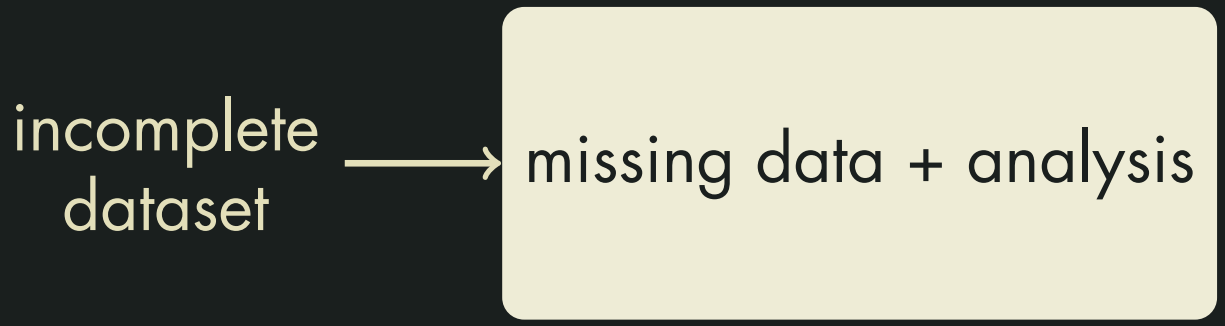
Missing Data

APPLICATION-SPECIFIC METHODS:

incomplete
dataset

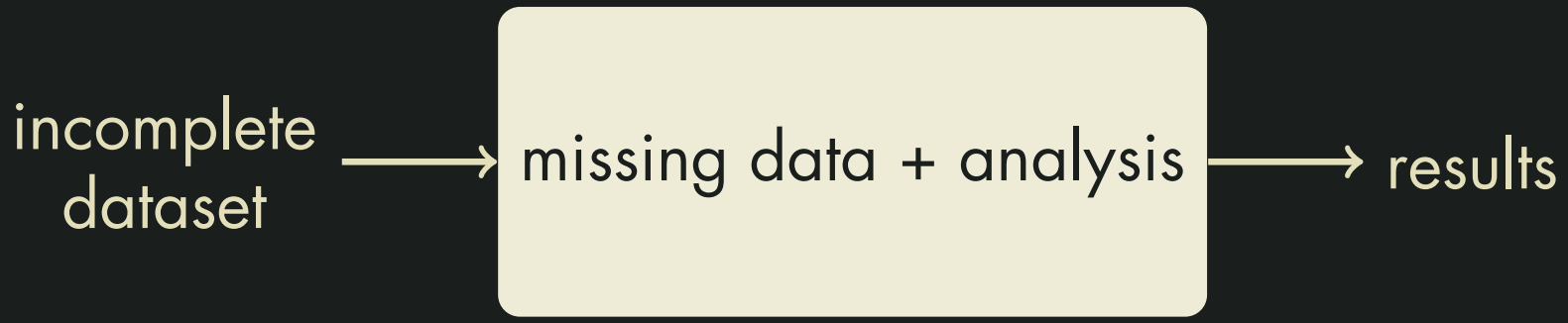
Missing Data

APPLICATION-SPECIFIC METHODS:



Missing Data

APPLICATION-SPECIFIC METHODS:



Missing Data

APPLICATION-SPECIFIC METHODS:

Missing Data

APPLICATION-SPECIFIC METHODS:

MULTIPLE IMPUTATION:

Missing Data

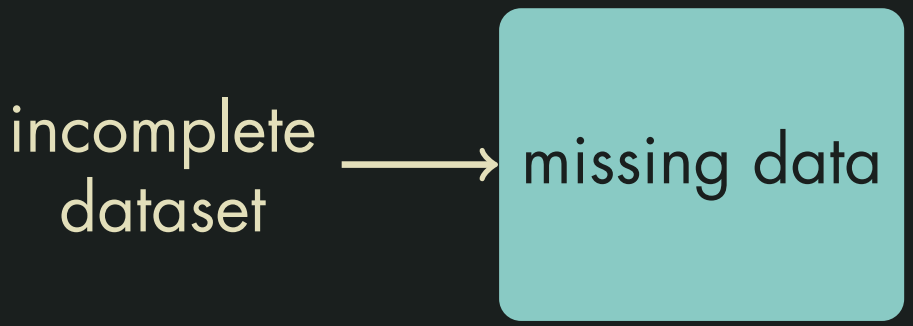
APPLICATION-SPECIFIC METHODS:

MULTIPLE IMPUTATION:

Missing Data

APPLICATION-SPECIFIC METHODS:

MULTIPLE IMPUTATION:



Missing Data

APPLICATION-SPECIFIC METHODS:

MULTIPLE IMPUTATION:

Missing Data

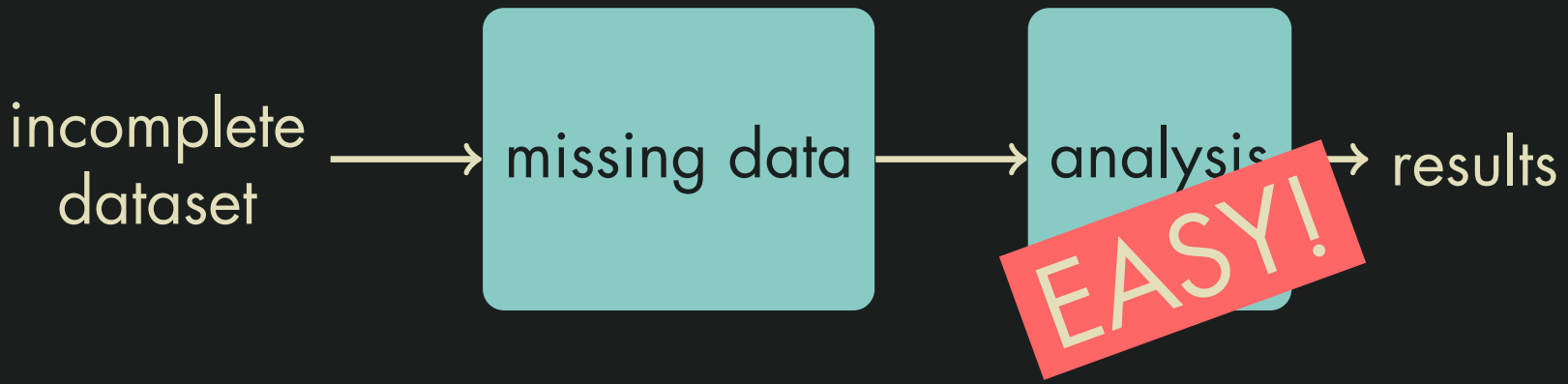
APPLICATION-SPECIFIC METHODS:

MULTIPLE IMPUTATION:

Missing Data

APPLICATION-SPECIFIC METHODS:

MULTIPLE IMPUTATION:



Missing Data

APPLICATION-SPECIFIC METHODS:

MULTIPLE IMPUTATION:

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

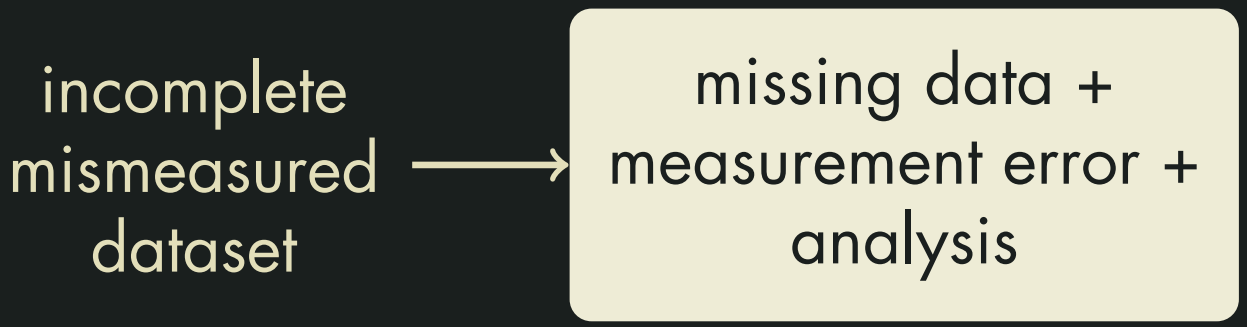
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

incomplete
mismeasured
dataset

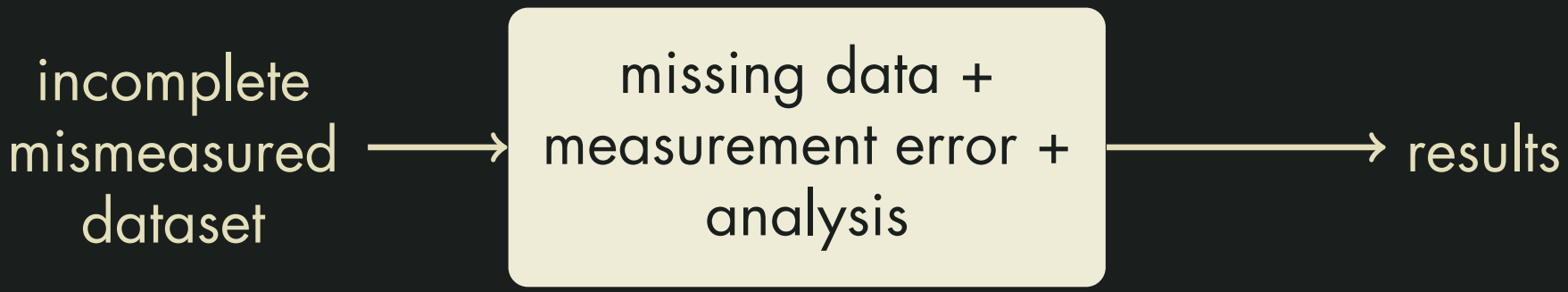
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:



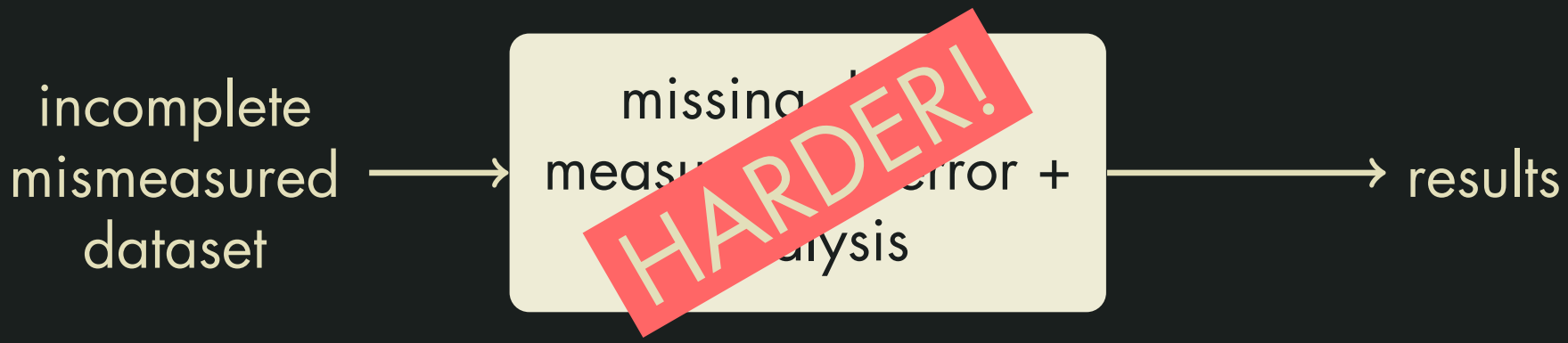
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:



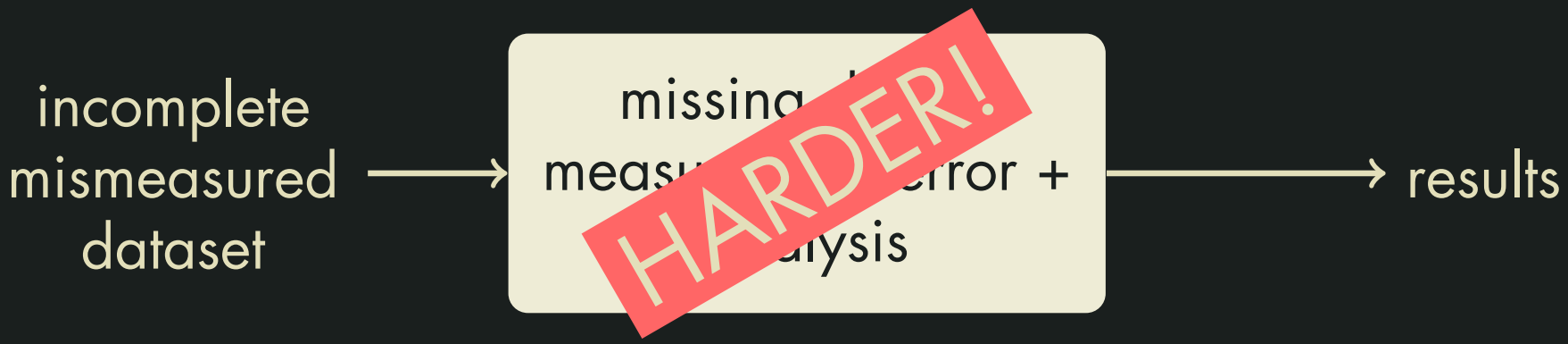
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:



Missing Data and Measurement Error

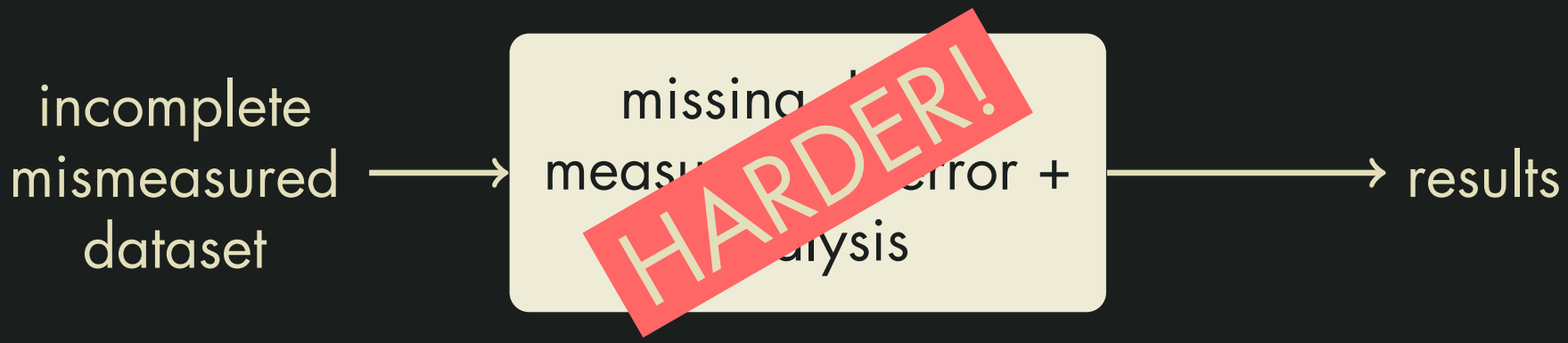
APPLICATION-SPECIFIC METHODS:



MULTIPLE OVERIMPUTATION:

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

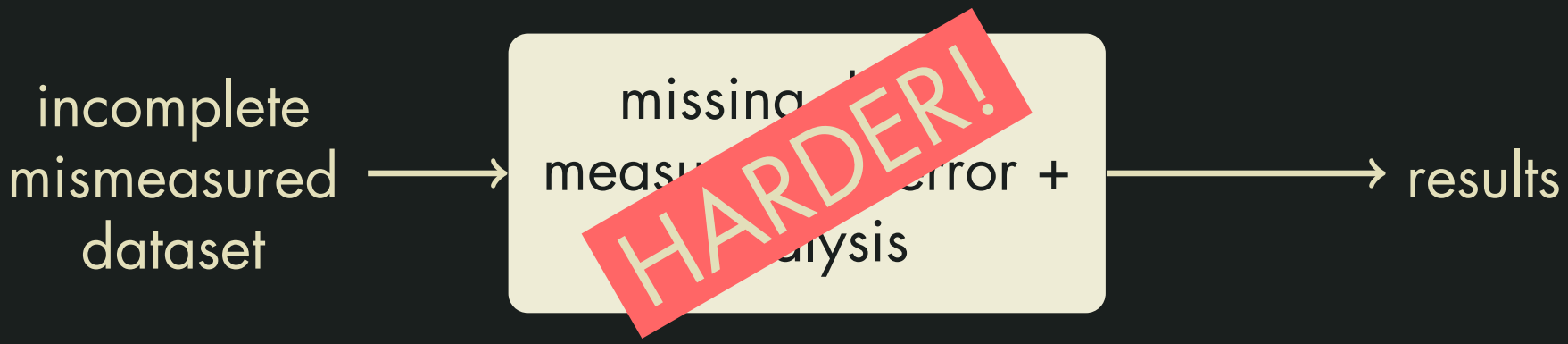


MULTIPLE OVERIMPUTATION:

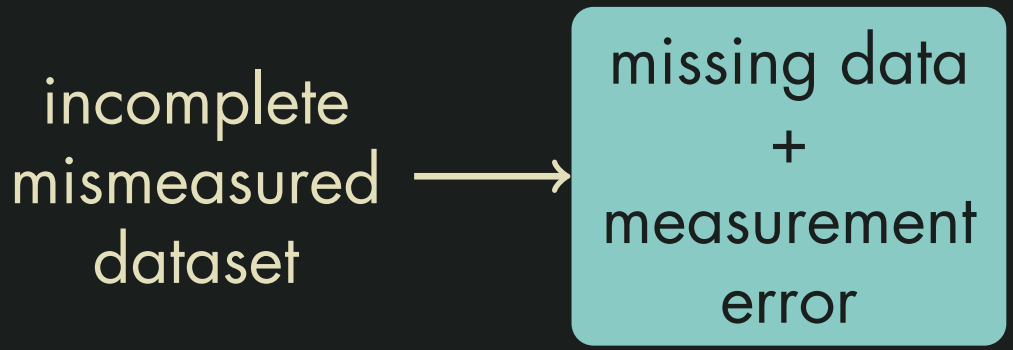
incomplete
mismeasured
dataset

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

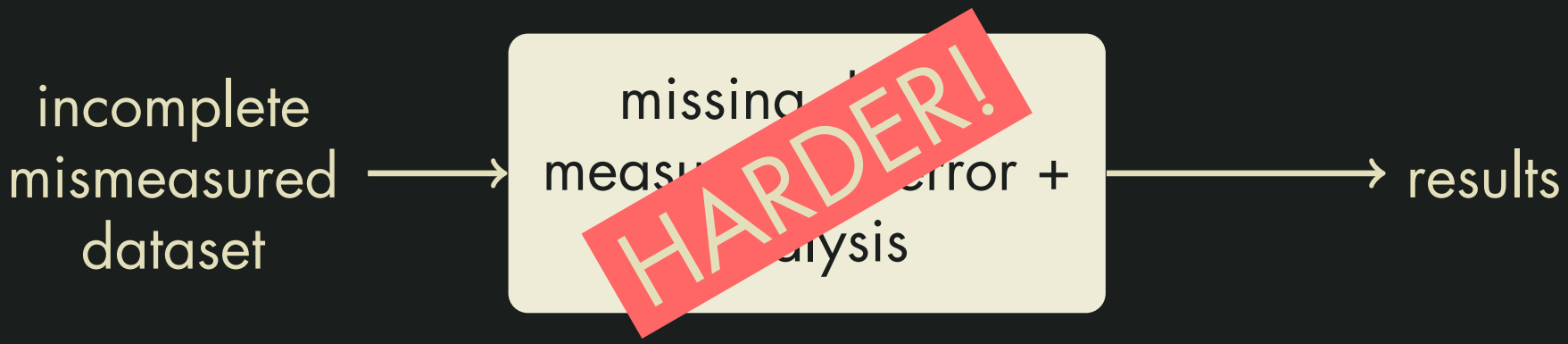


MULTIPLE OVERIMPUTATION:

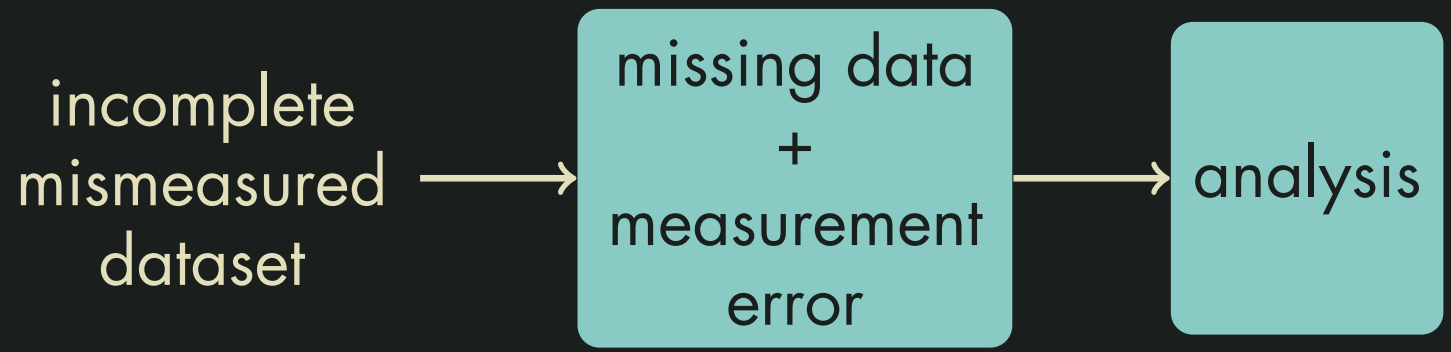


Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

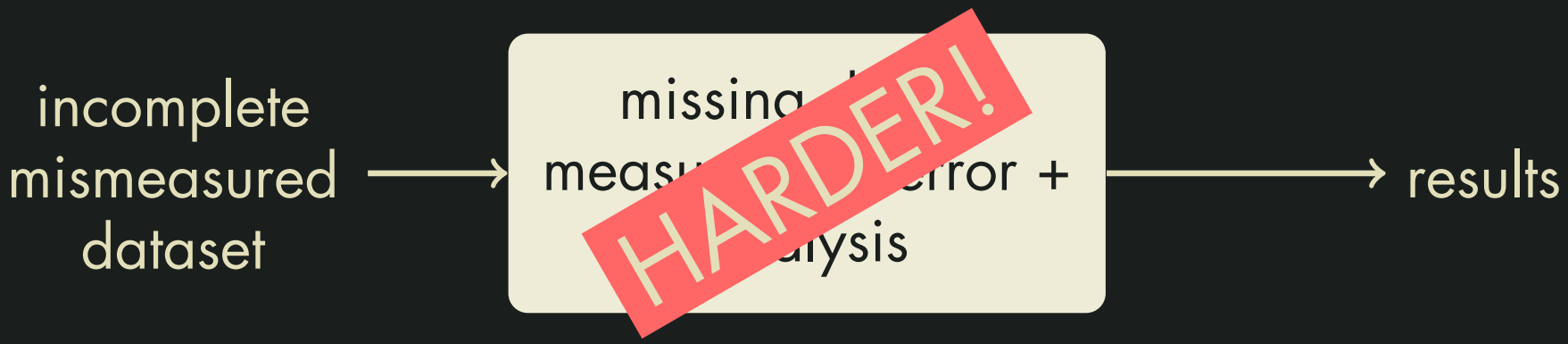


MULTIPLE OVERIMPUTATION:

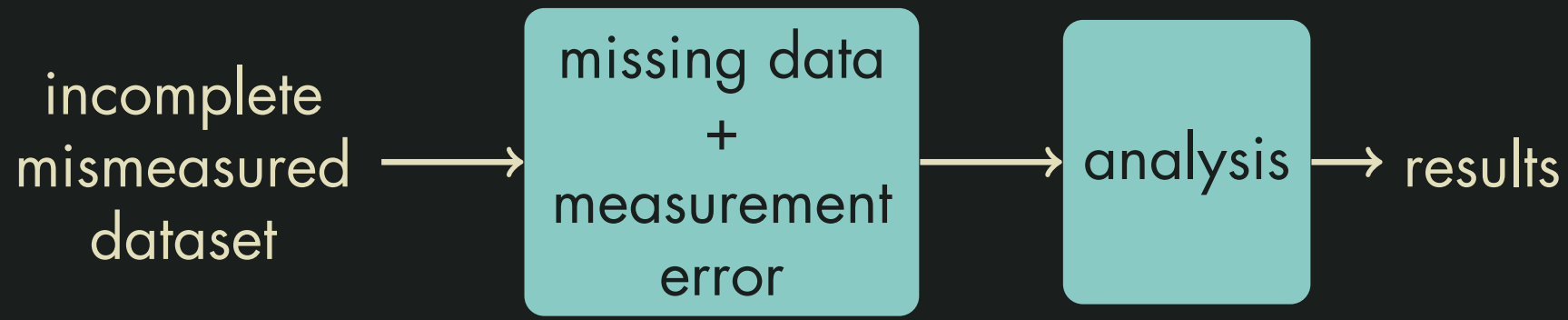


Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

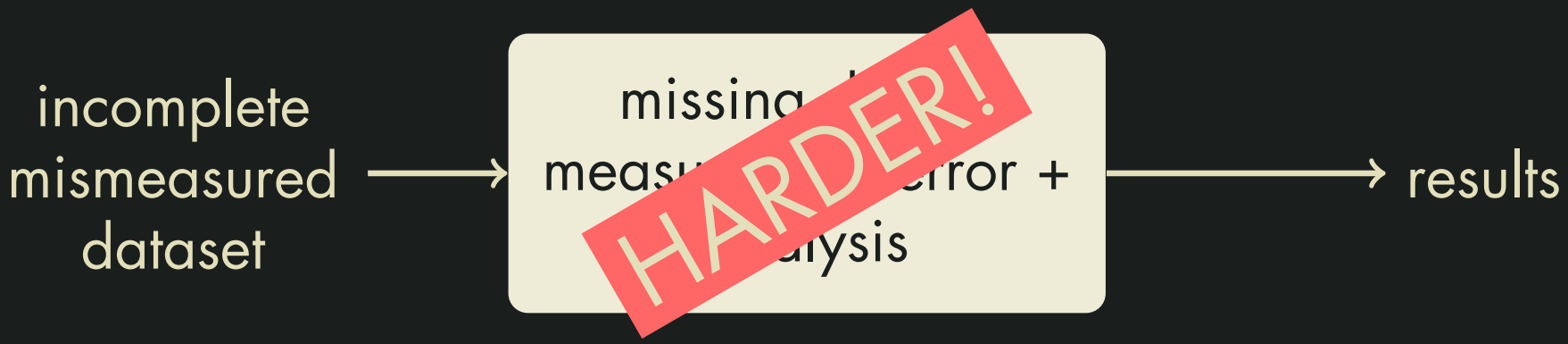


MULTIPLE OVERIMPUTATION:

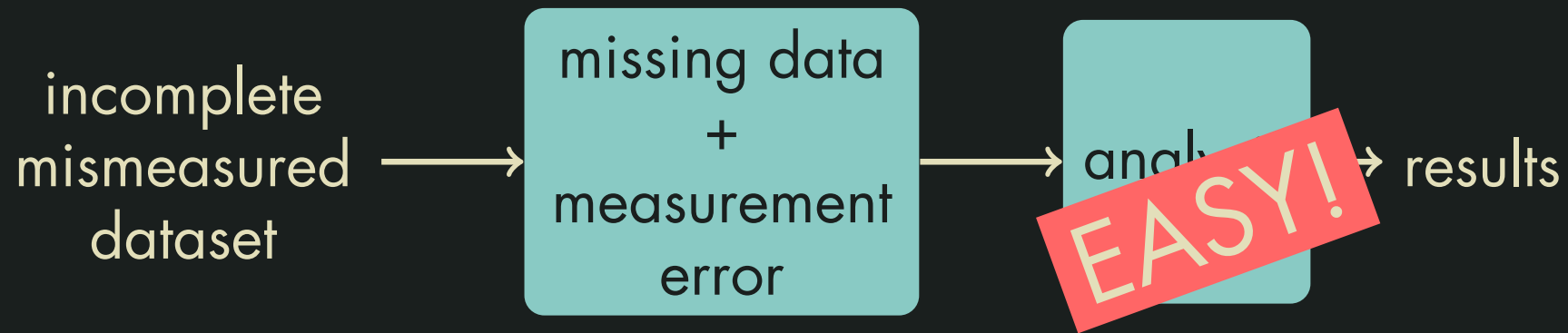


Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

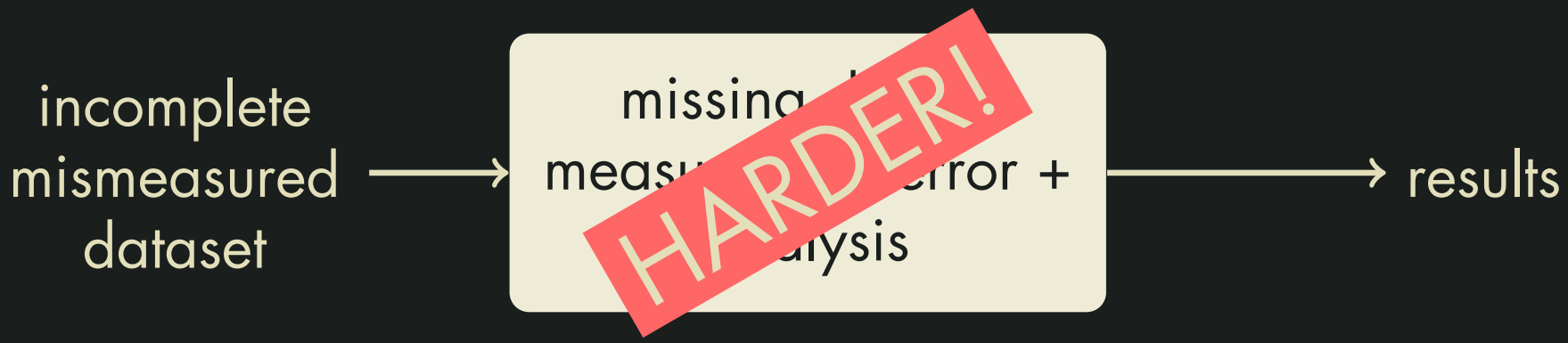


MULTIPLE OVERIMPUTATION:



Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:



MULTIPLE OVERIMPUTATION:

What MO allows you to do:

What MO allows you to do:

political science.

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	≈9	6	6.23	5.92
2	LIBERIA	NA	3	NA	NA
3	SIERRA LEONE	≈3	3	6.60	NA
4	GHANA	≈9	6	6.86	12.68
5	TOGO	NA	5	6.27	17.34
6	CAMEROON	≈6	5	6.93	15.47
7	NIGERIA	≈5	7	6.88	17.46
8	GABON	≈6	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	≈9	6	6.23	5.92
2	LIBERIA	∧	3	∧	∧
3	SIERRA LEONE	≈3	3	6.60	∧
4	GHANA	≈9	6	6.86	12.68
5	TOGO	∧	5	6.27	17.34
6	CAMEROON	≈6	5	6.93	15.47
7	NIGERIA	≈5	7	6.88	17.46
8	GABON	≈6	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

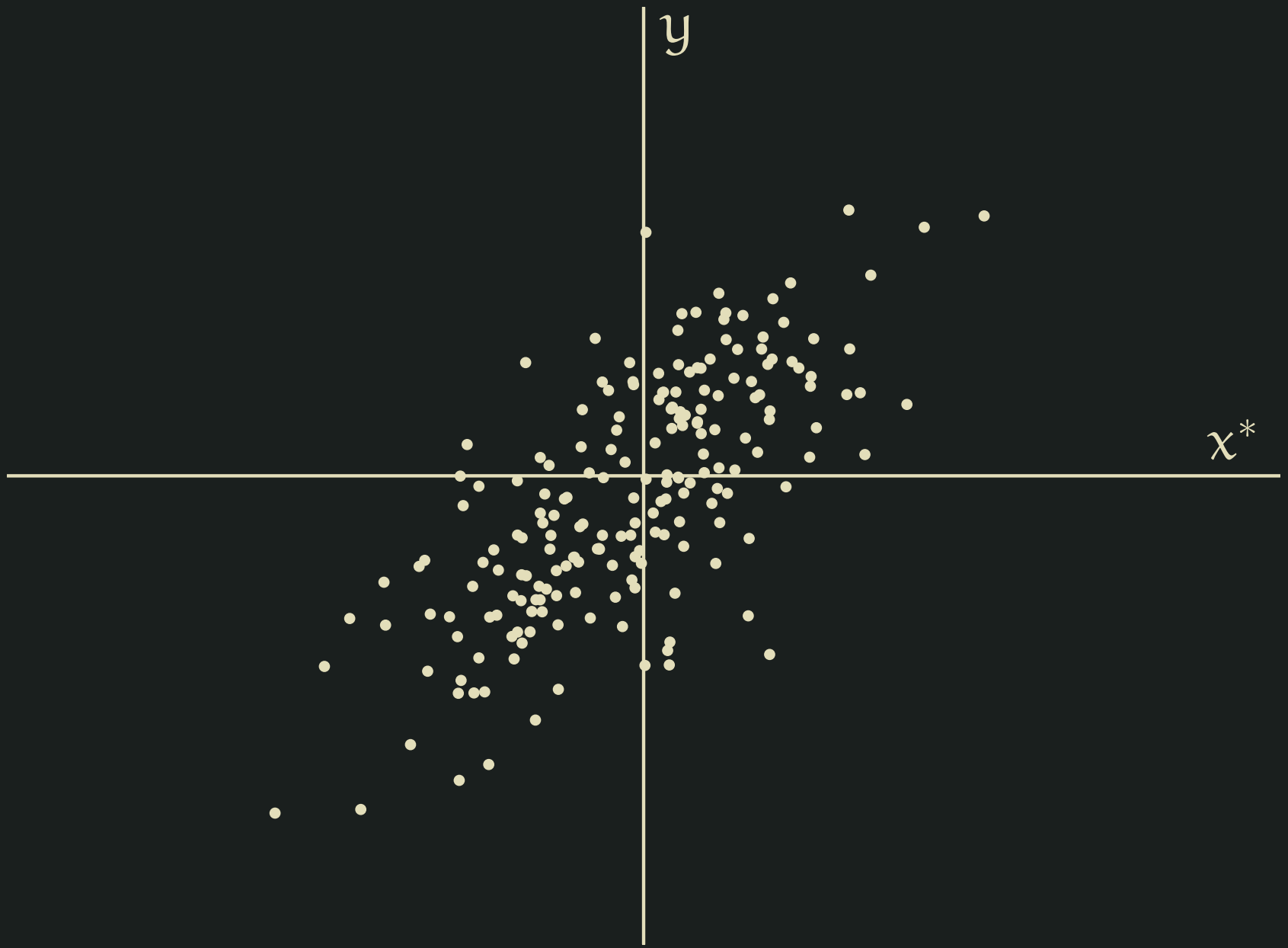
	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

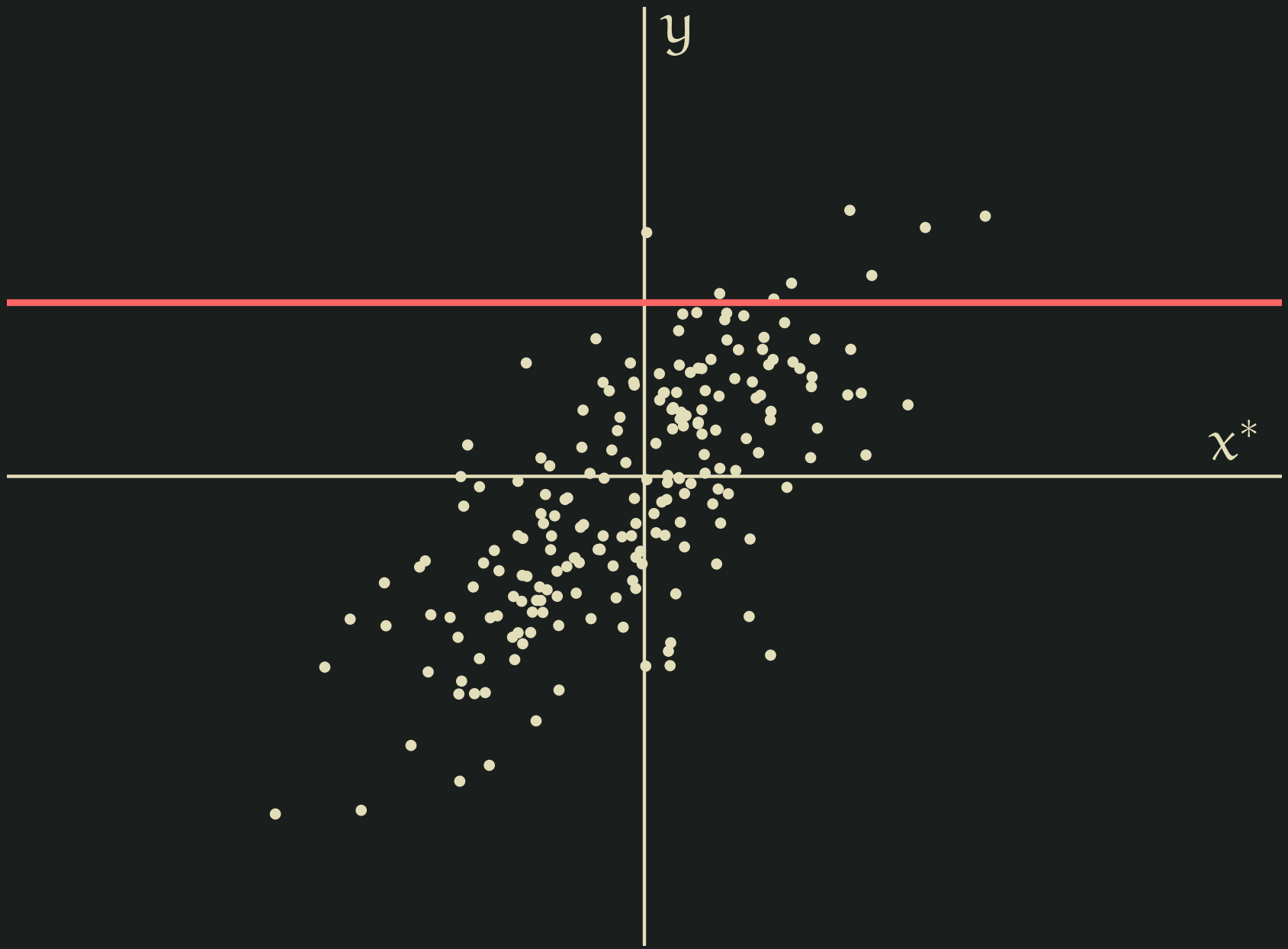
	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO		6	6.23	5.92
2	LIBERIA		3		
3	SIERRA LEONE		3	6.60	
4	GHANA		6	6.86	12.68
5	TOGO		5	6.27	17.34
6	CAMEROON		5	6.93	15.47
7	NIGERIA		7	6.88	17.46
8	GABON		8	8.19	16.97

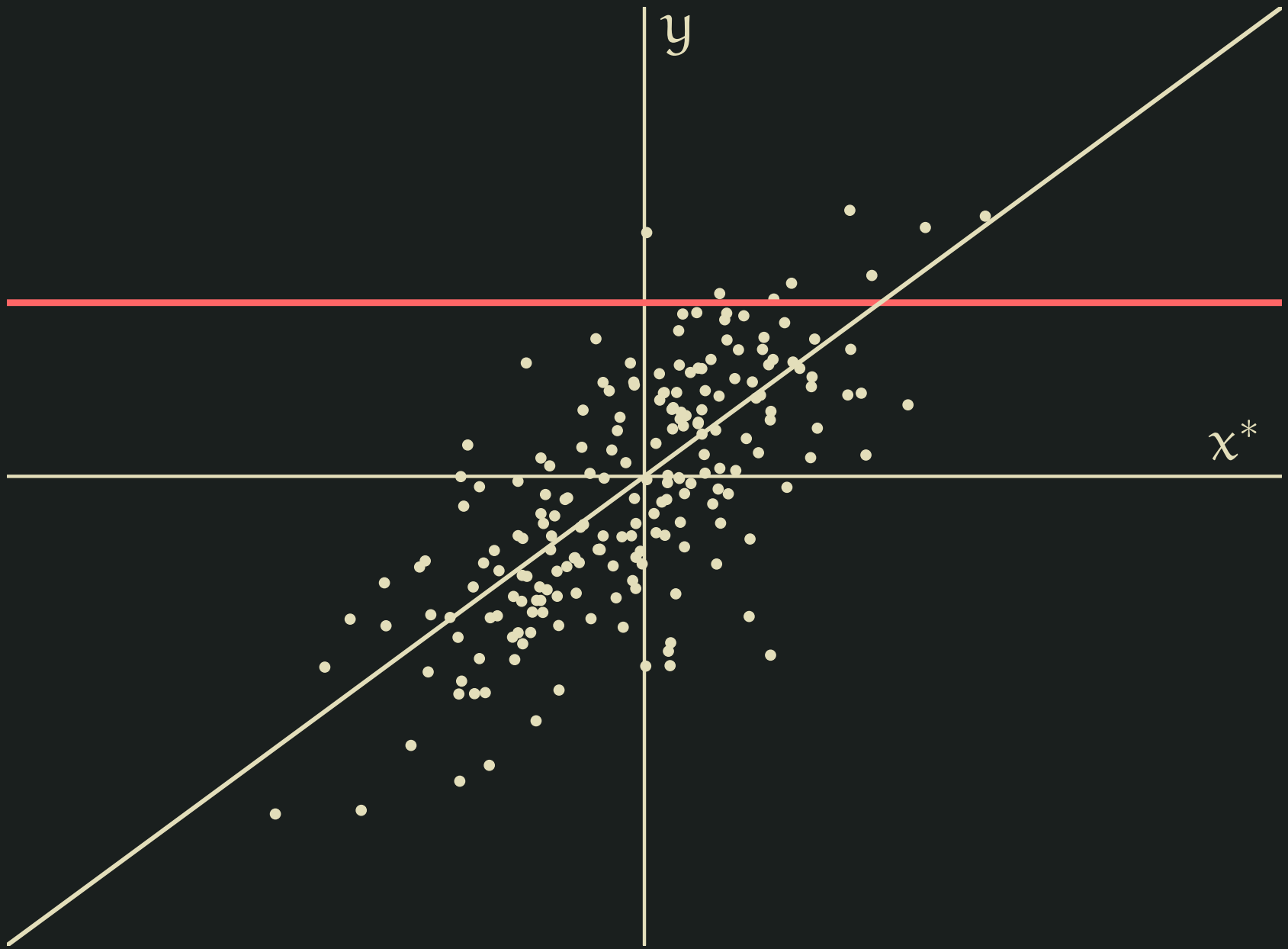
Run whatever analysis model you wanted to run.

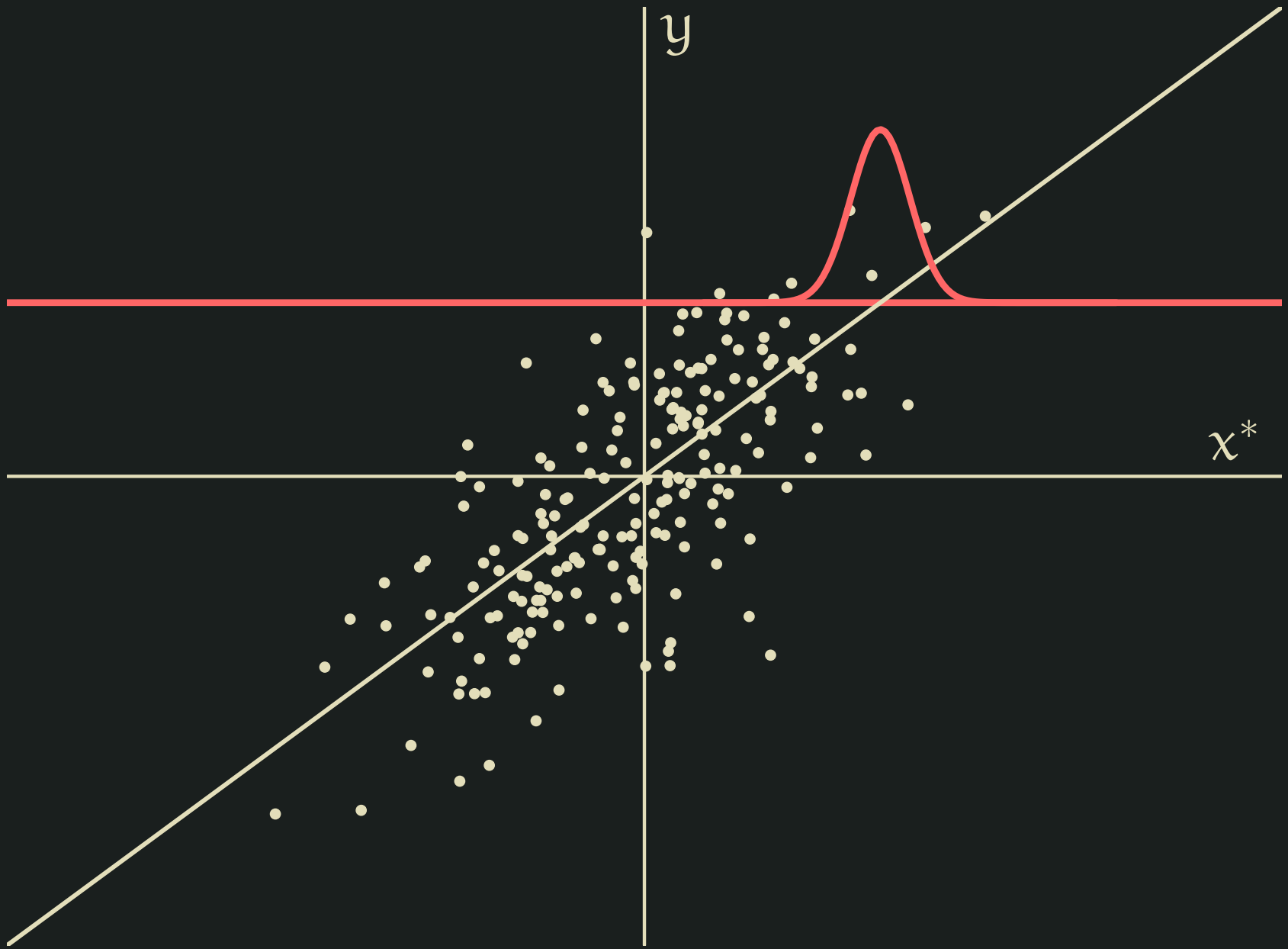
Run whatever analysis model you wanted to run.
$(\times 5)$

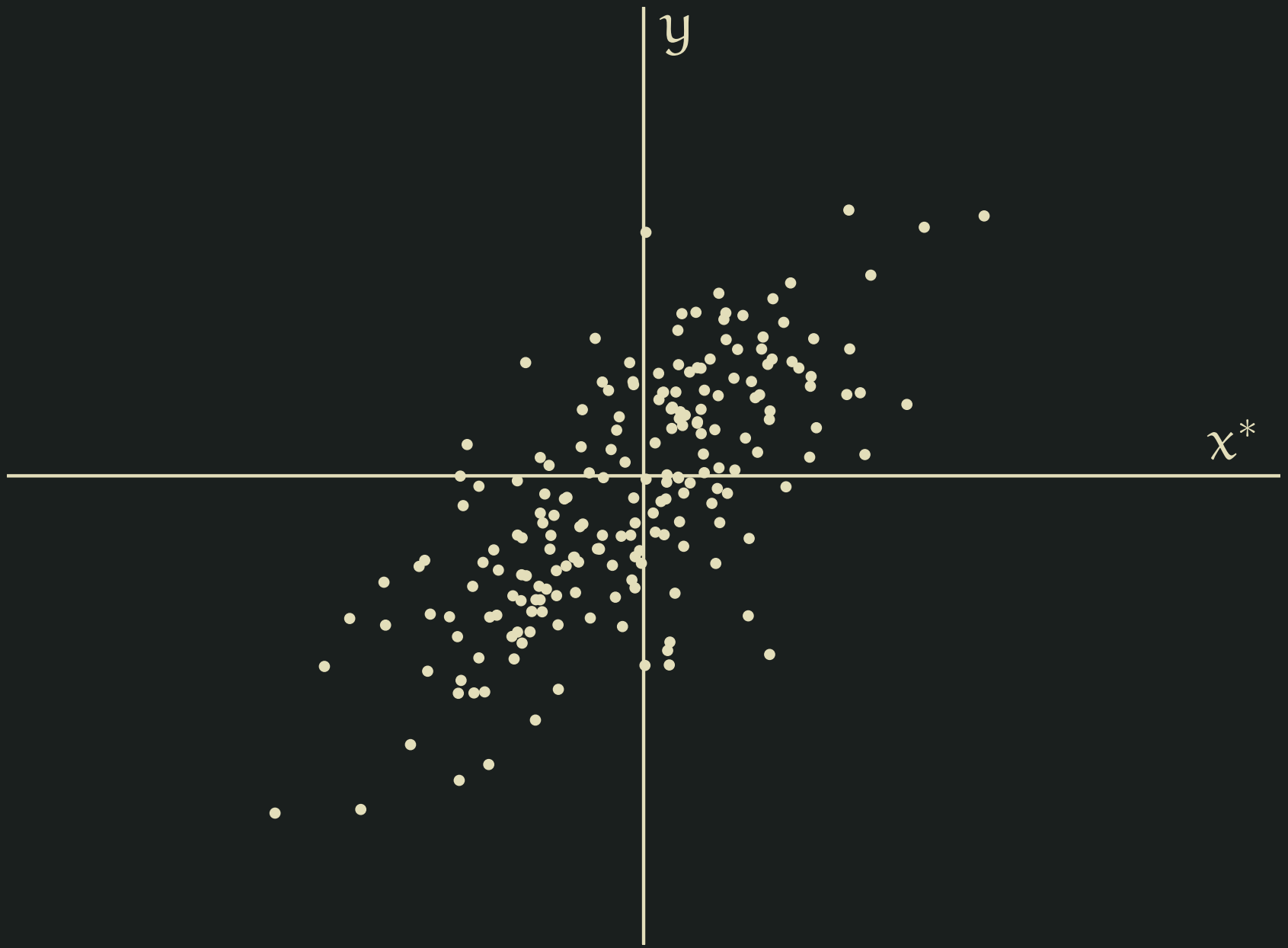
But how does it work?

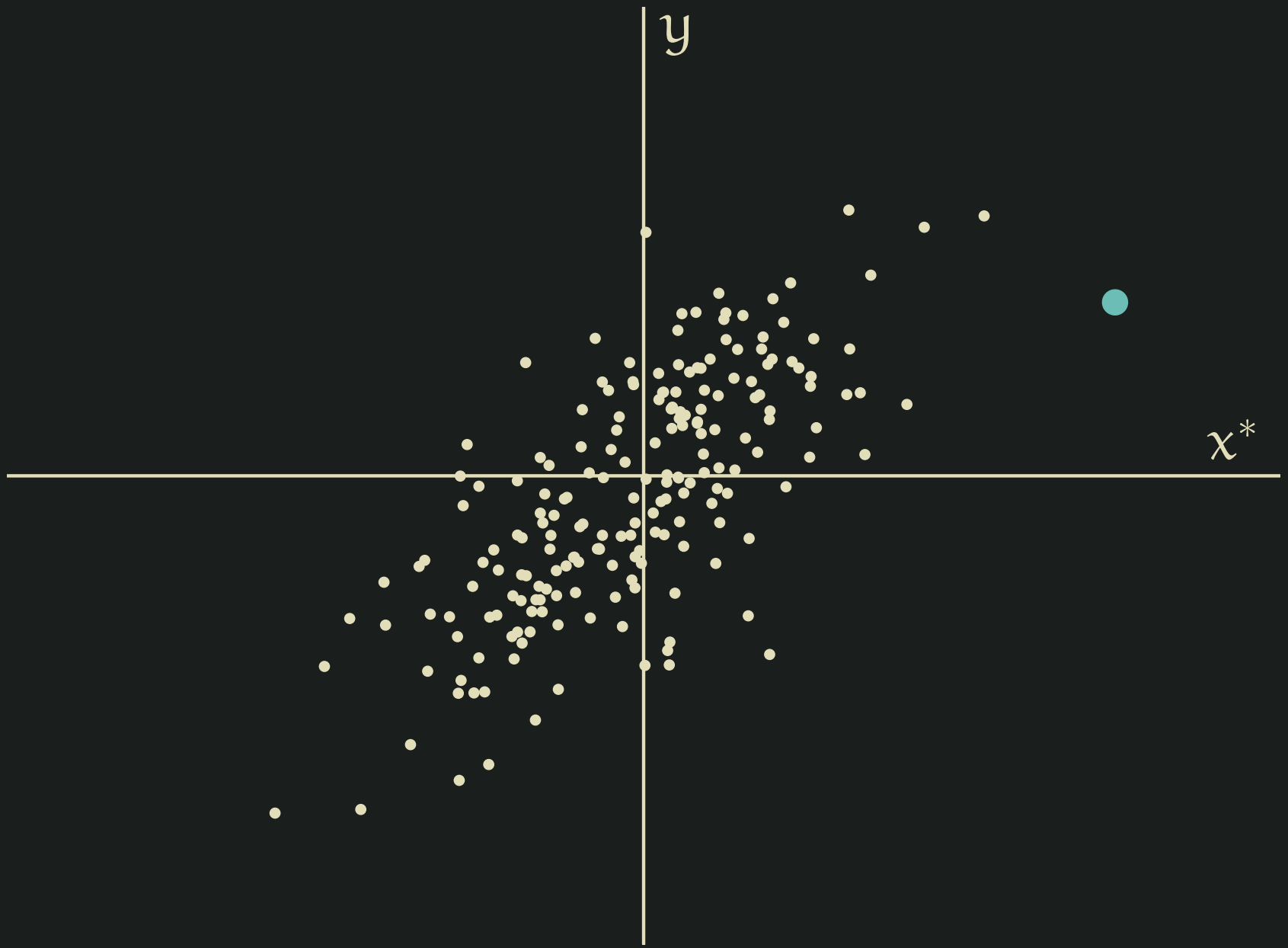

y
x*

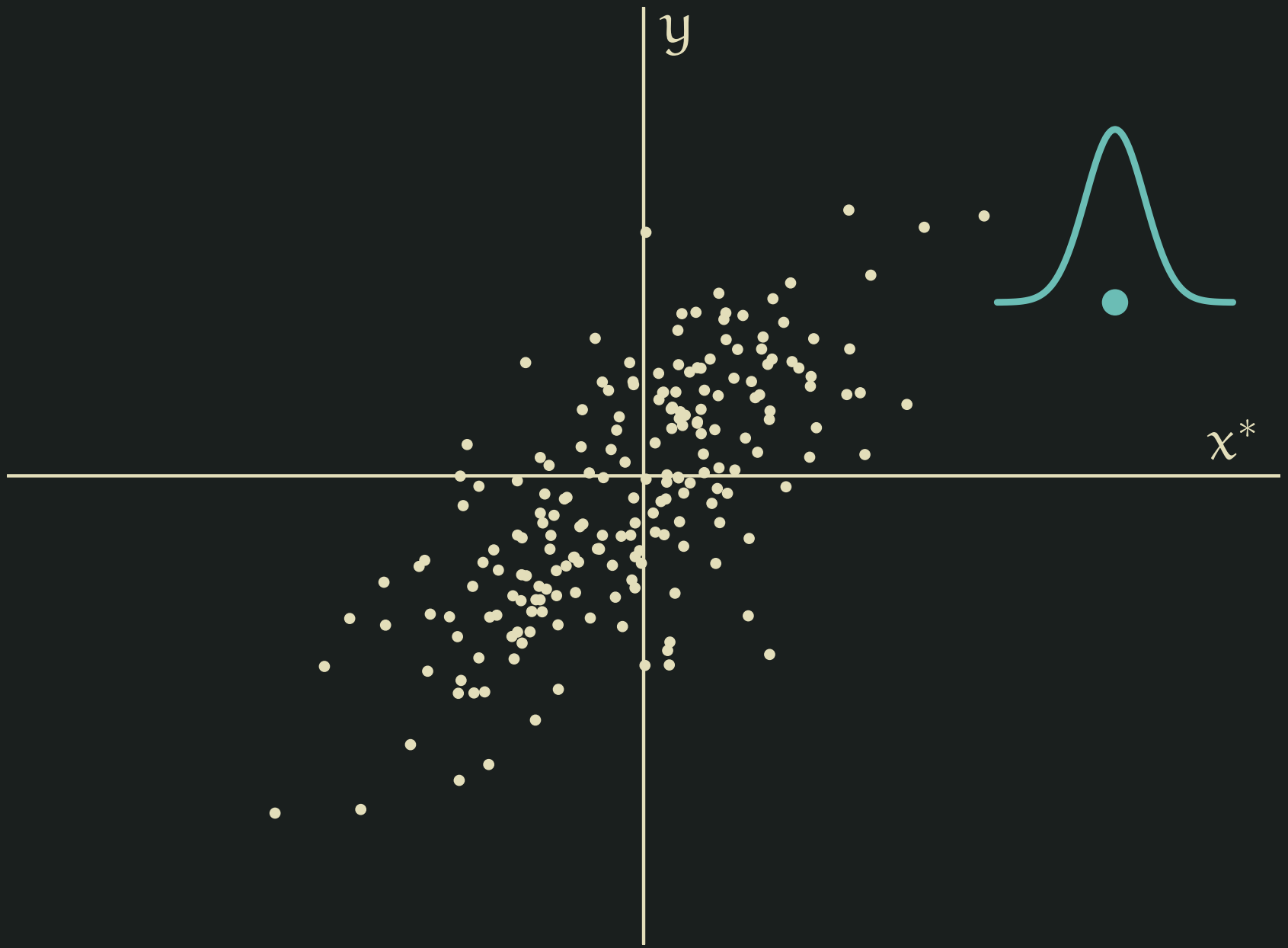

y
x^*

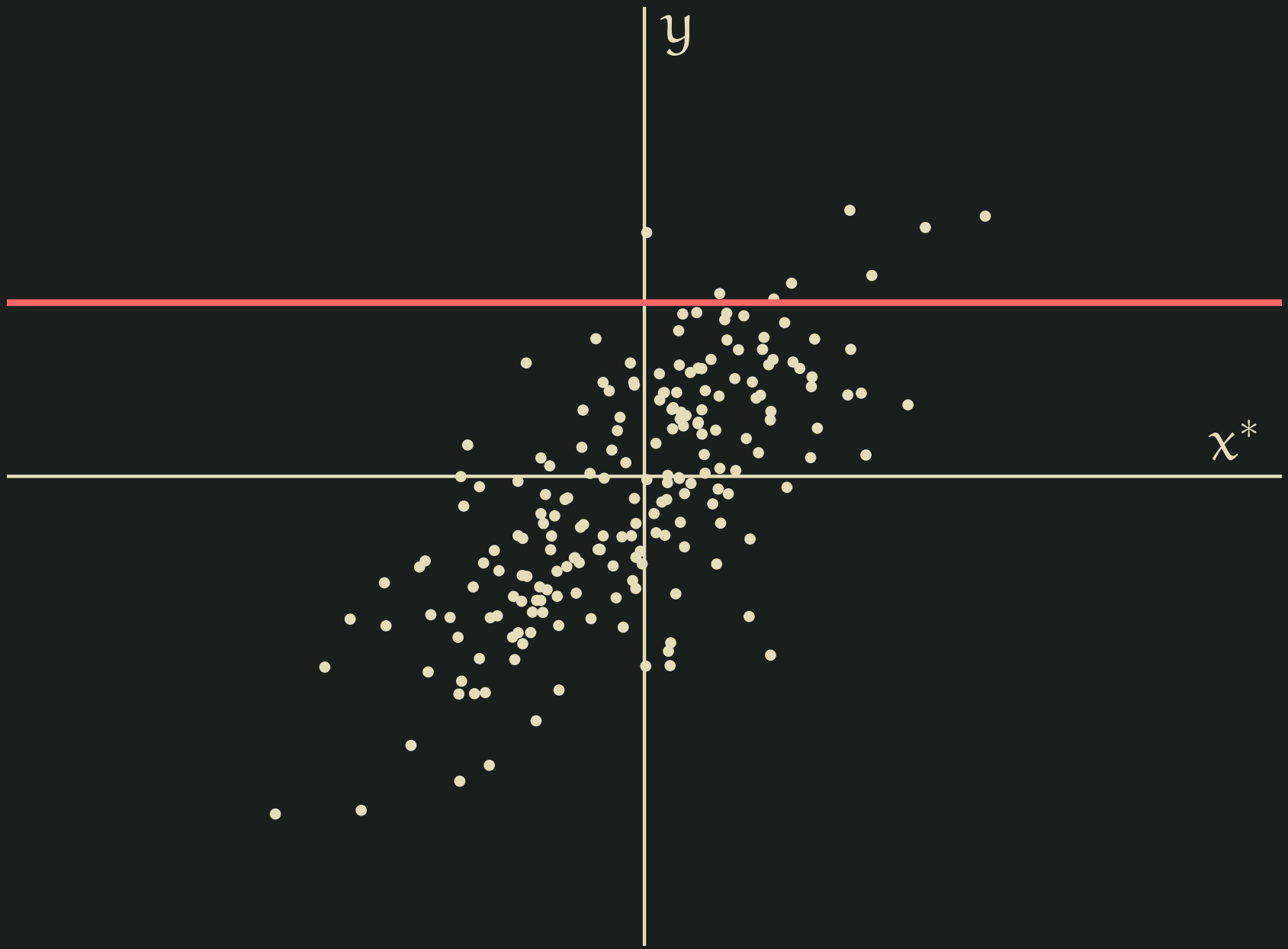

y
x^*

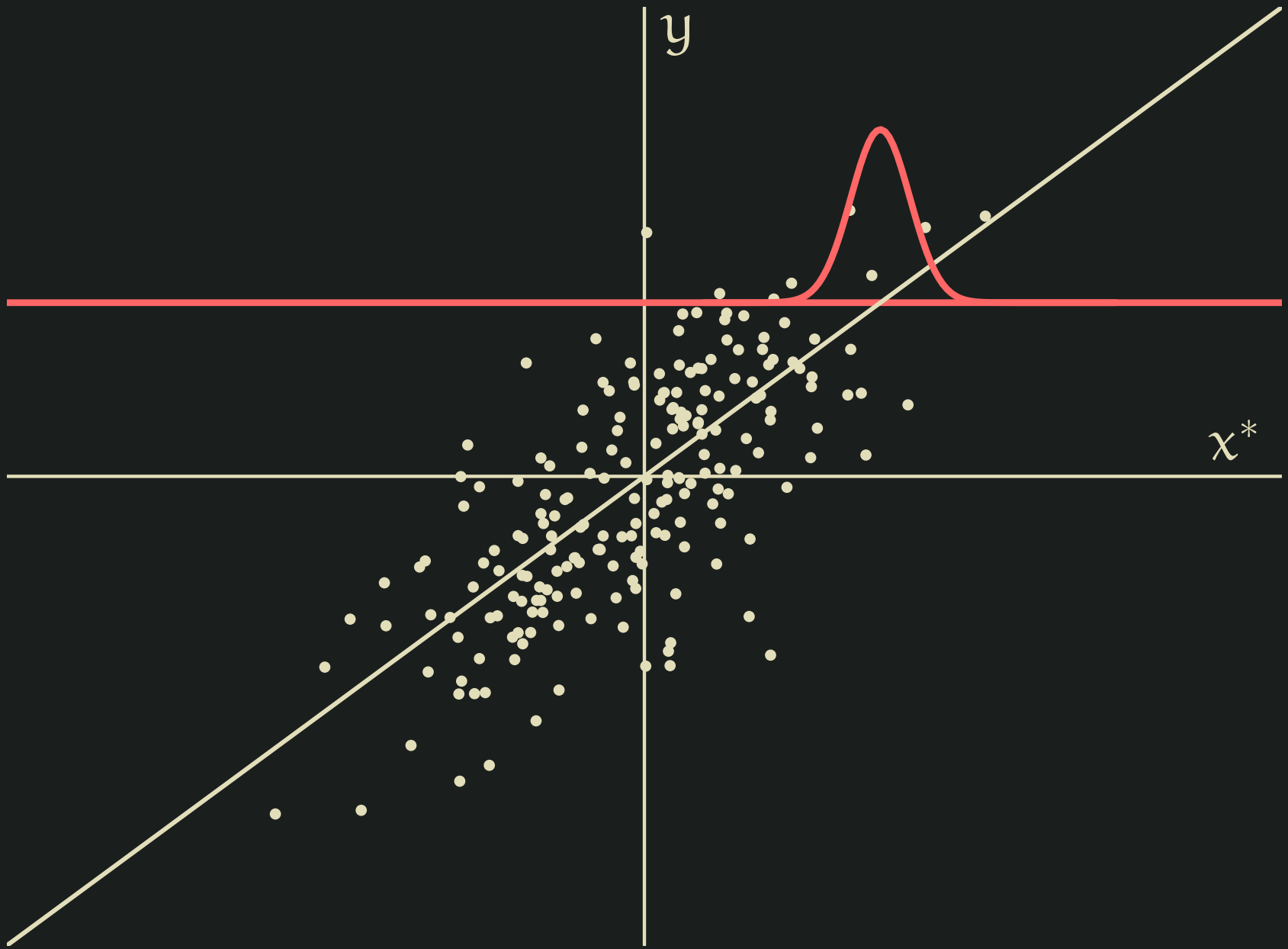

y
x^*

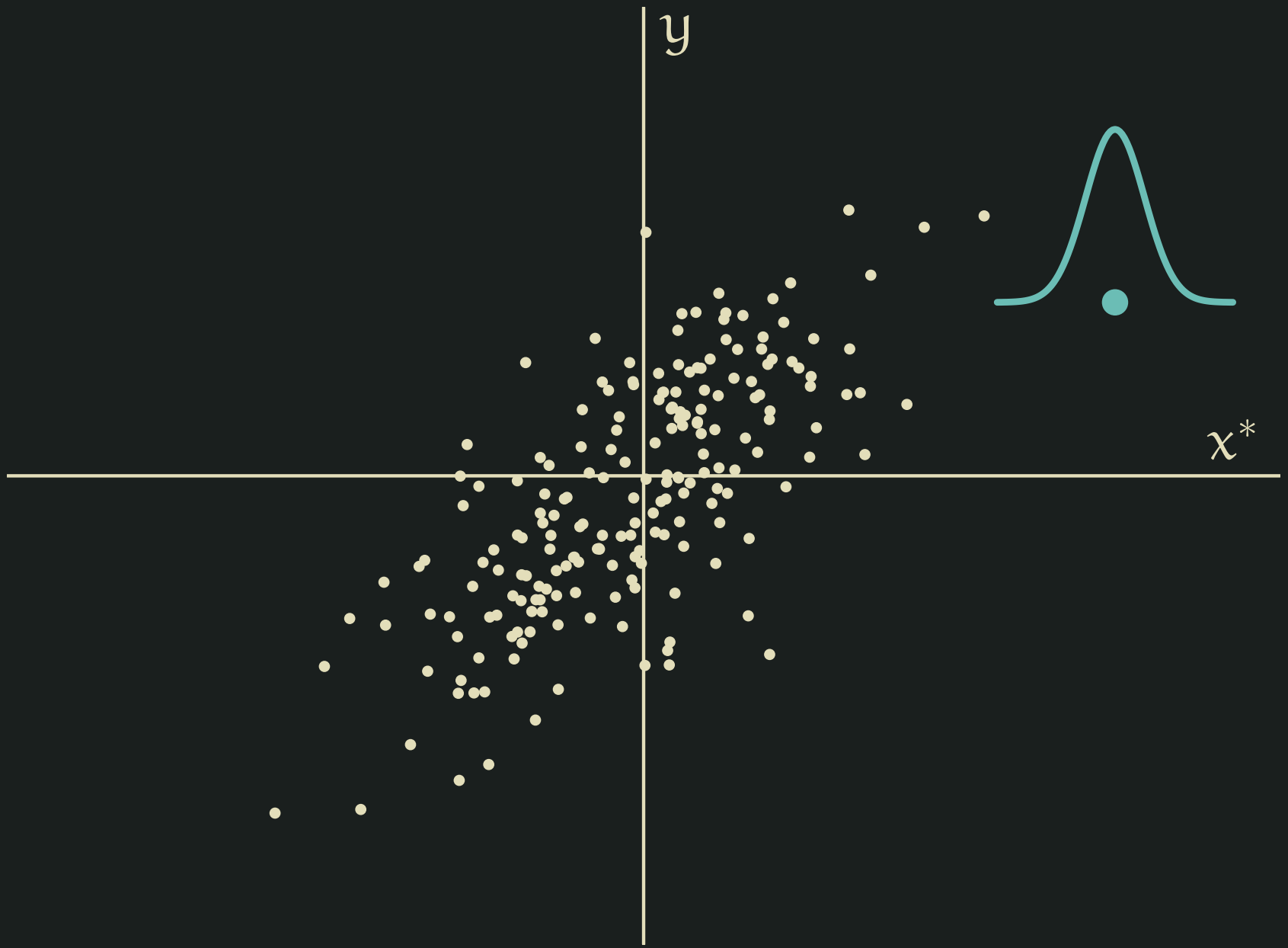

y
x*

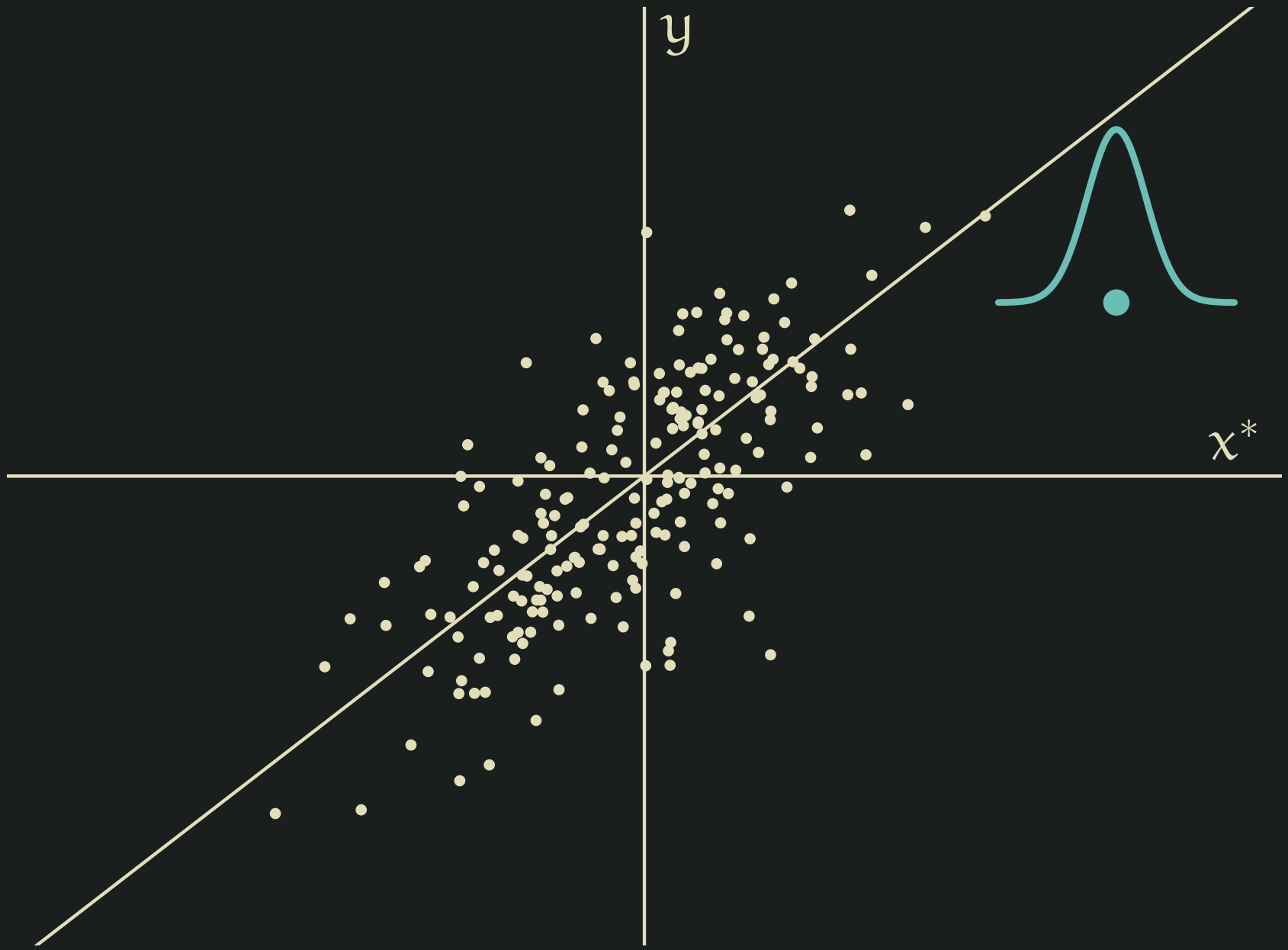

y
x^*

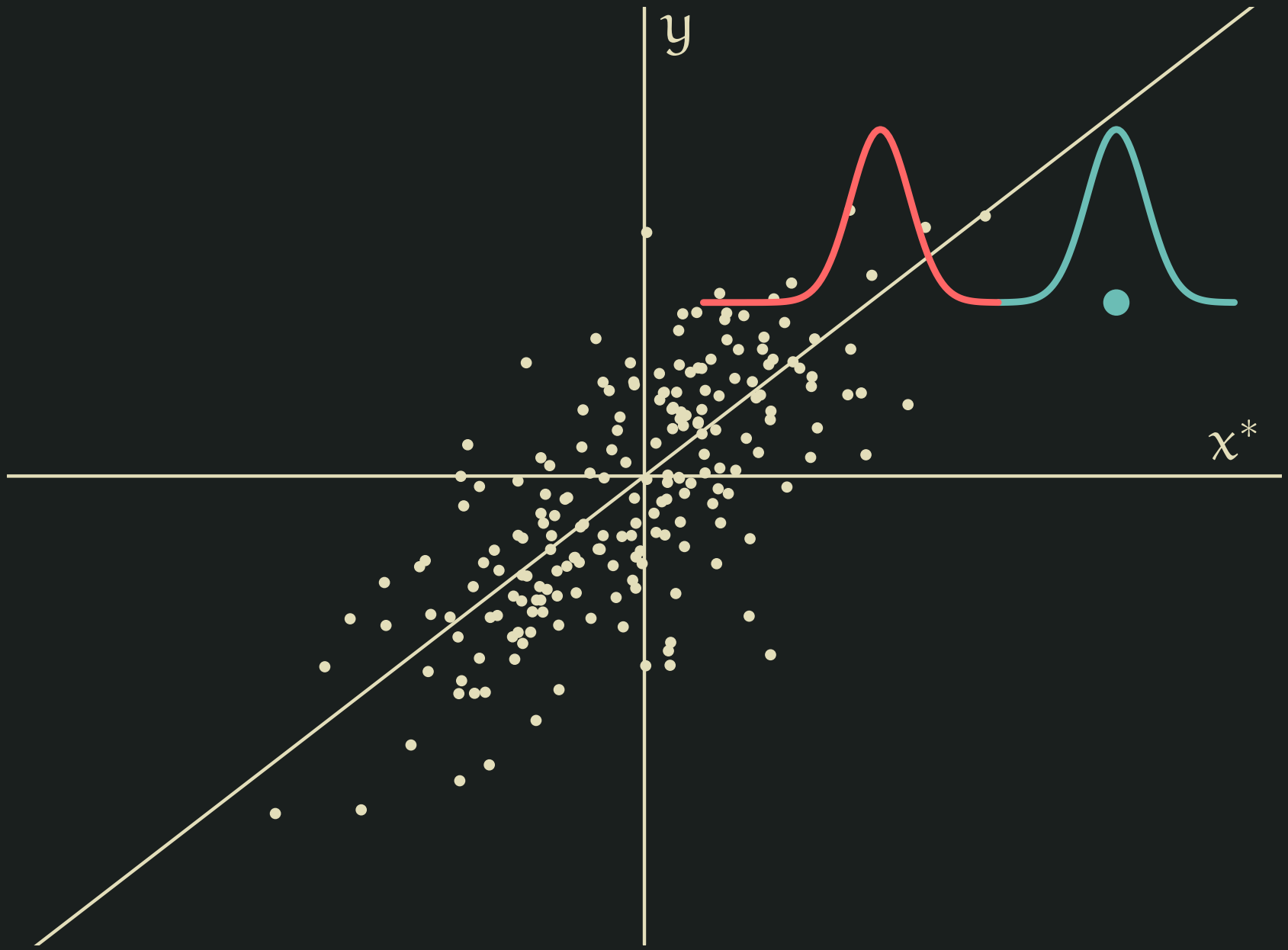

y
x^*

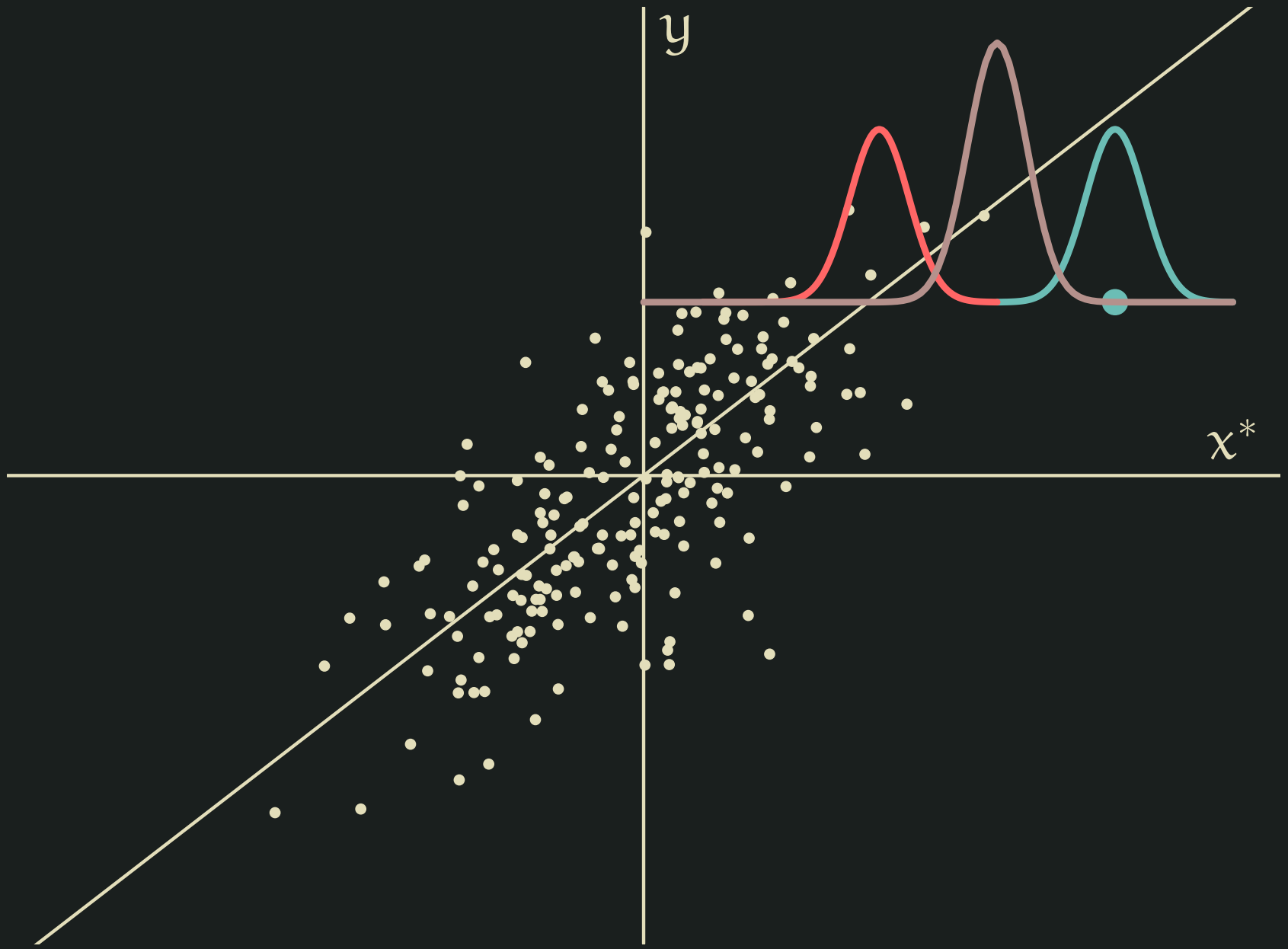

y
x^*


y
x^*


y
x^*


y
x^*


y
x^*


y
x^*

ARBITRARY PATTERNS OF MISMEASUREMENT & MISSINGNESS:

ARBITRARY PATTERNS OF MISMEASUREMENT & MISSINGNESS:

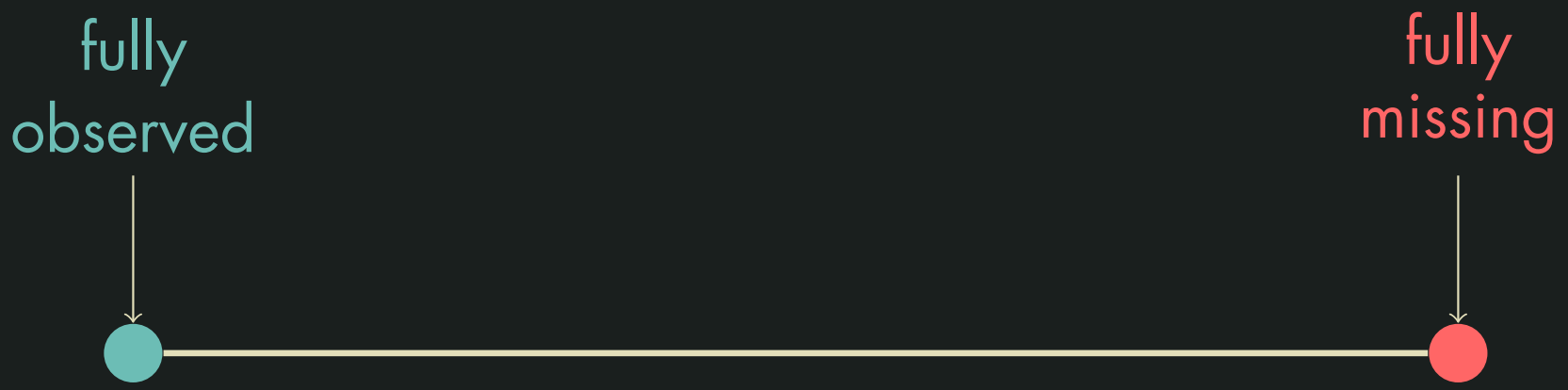
	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	4	≈6	6.23	5.92
2	LIBERIA	NA	3	NA	NA
3	SIERRA LEONE	3	3	6.60	NA
4	GHANA	≈9	6	6.86	12.68
5	TOGO	NA	5	6.27	17.34
6	CAMEROON	≈6	5	6.93	15.47
7	NIGERIA	≈5	7	≈6.88	17.46
8	GABON	≈6	8	≈8.19	≈16.97

PART FIVE:

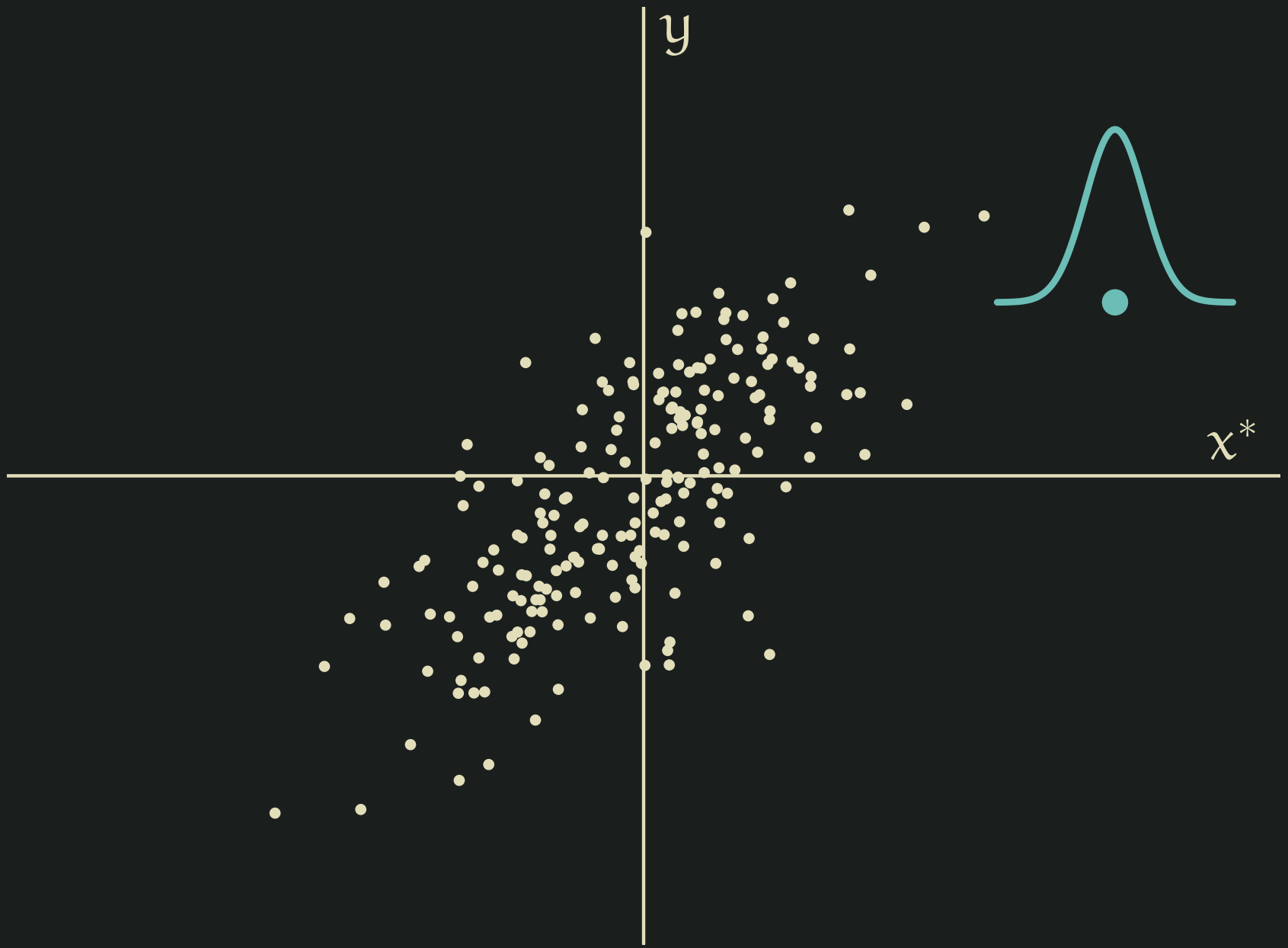
Establishing the measurement error variance.

The problem of identification.

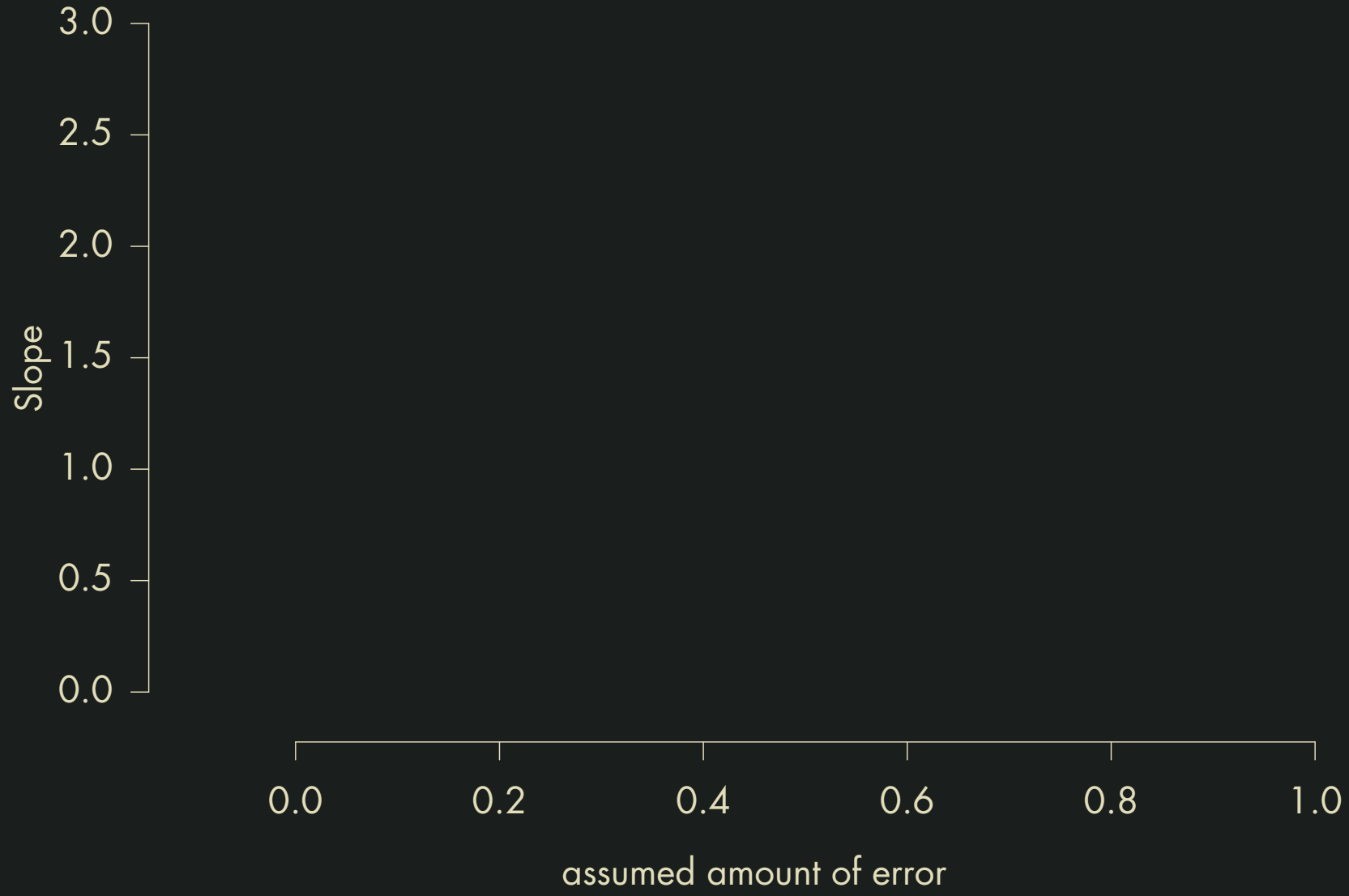
CHOOSE A VALUE OF σ_u^2

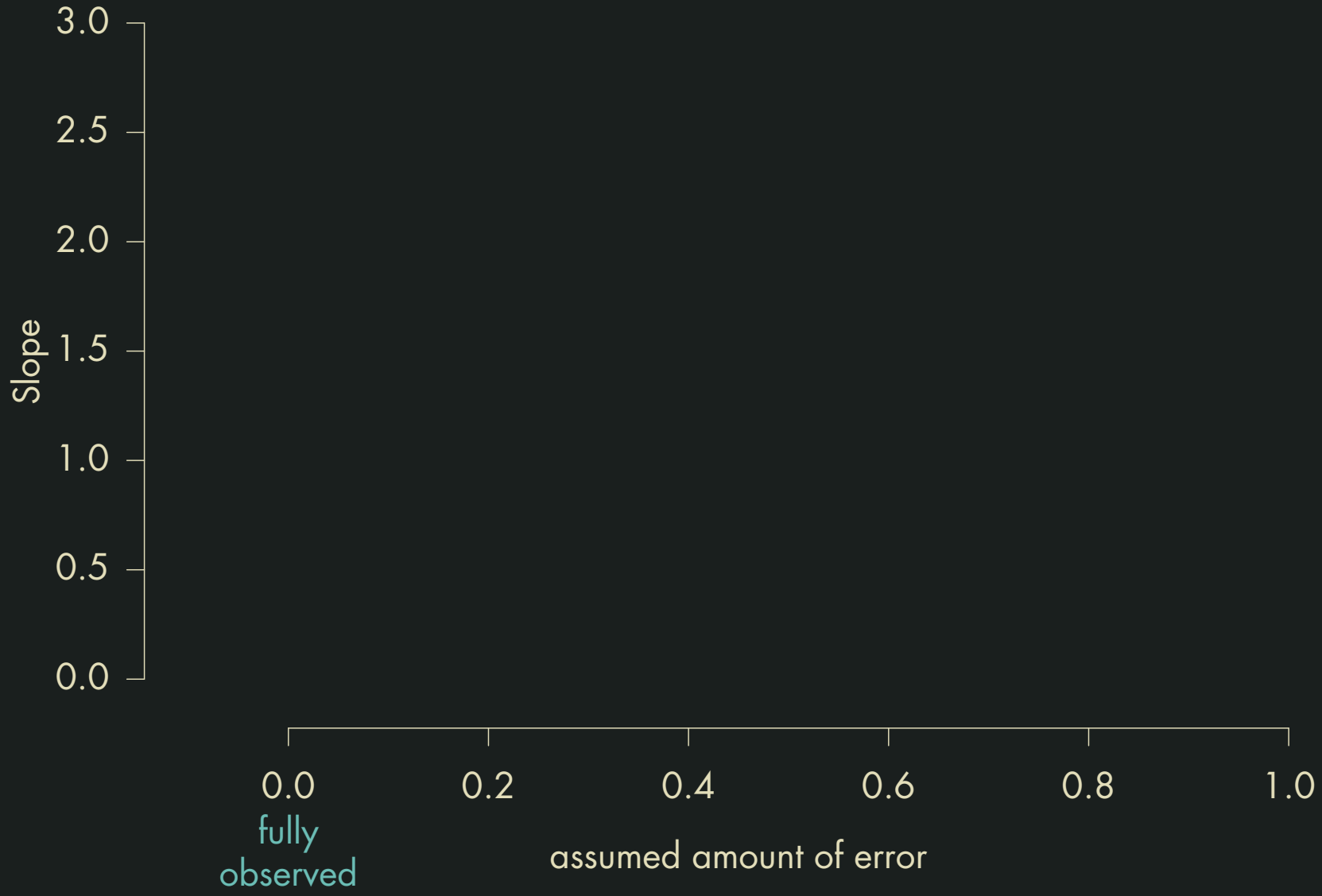


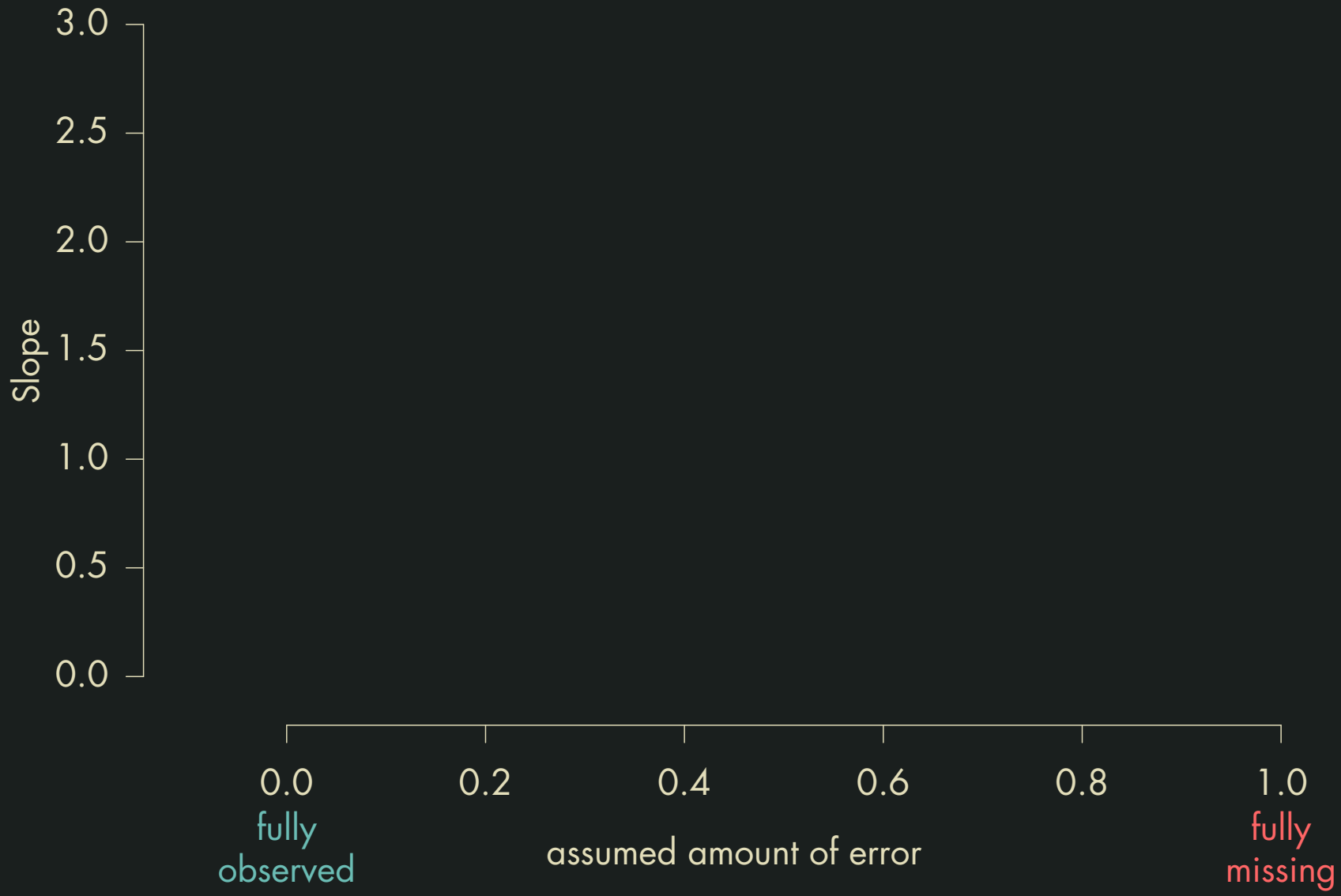
CHOOSE A VALUE OF σ_u^2

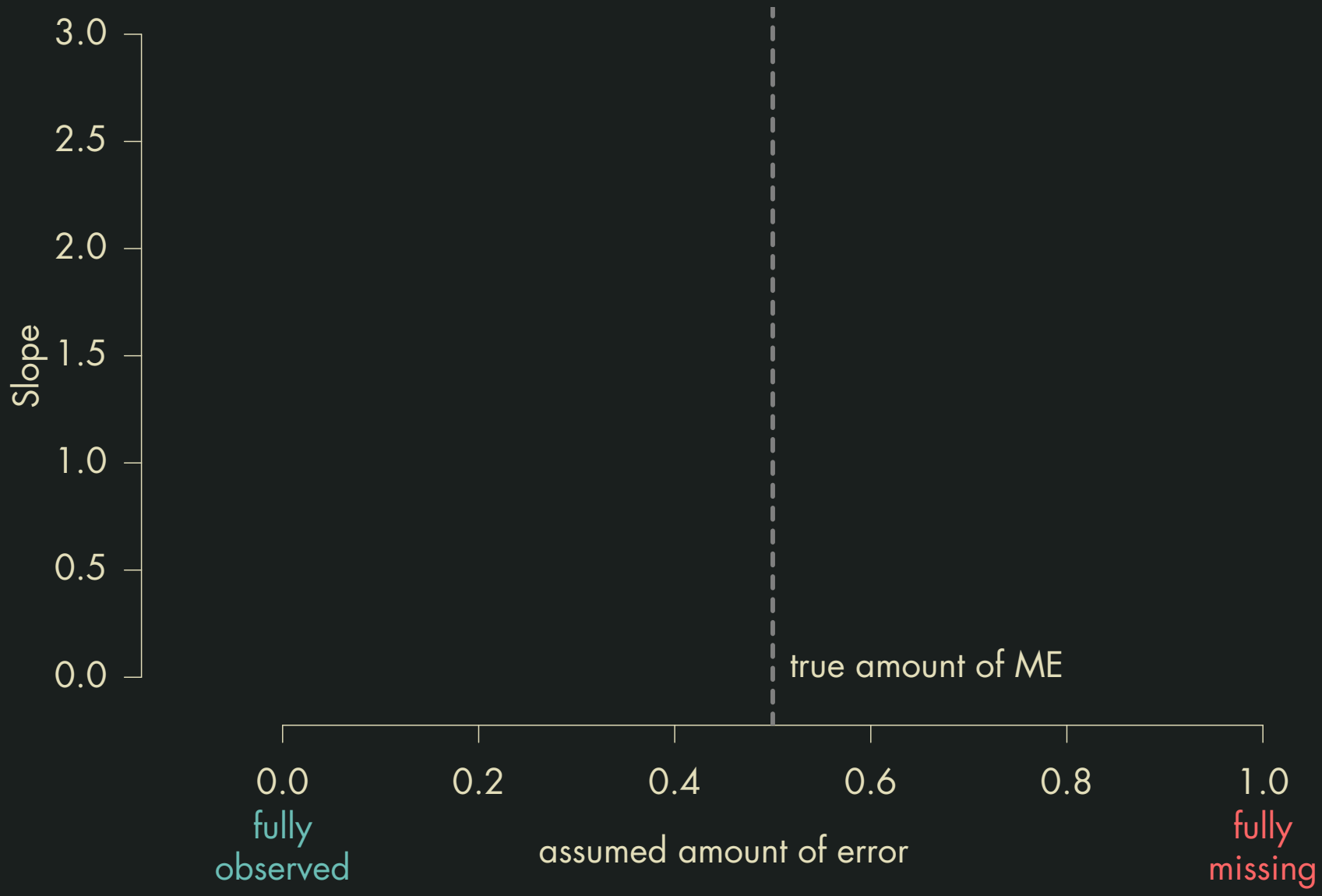

y
x^*

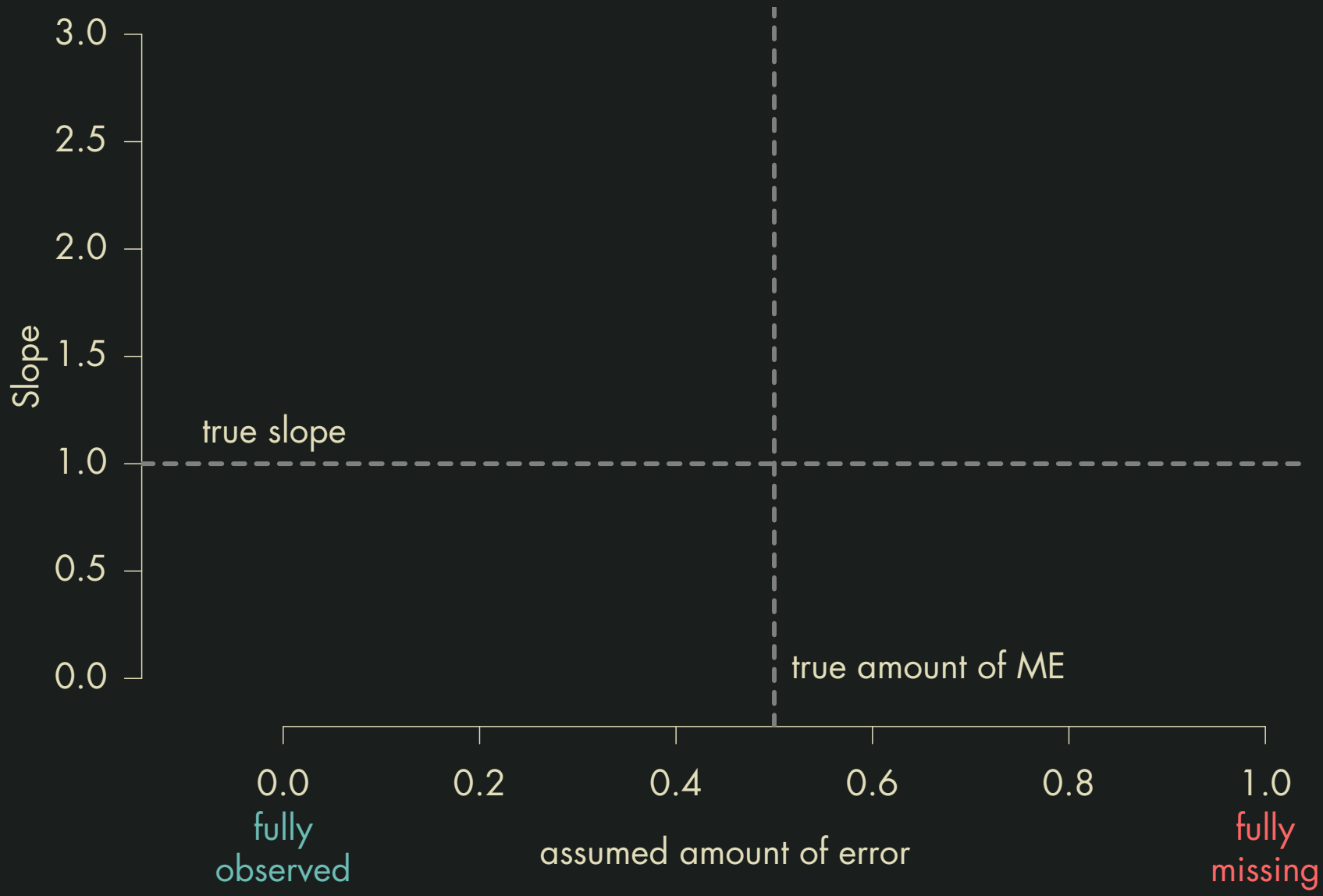
Some simulations.

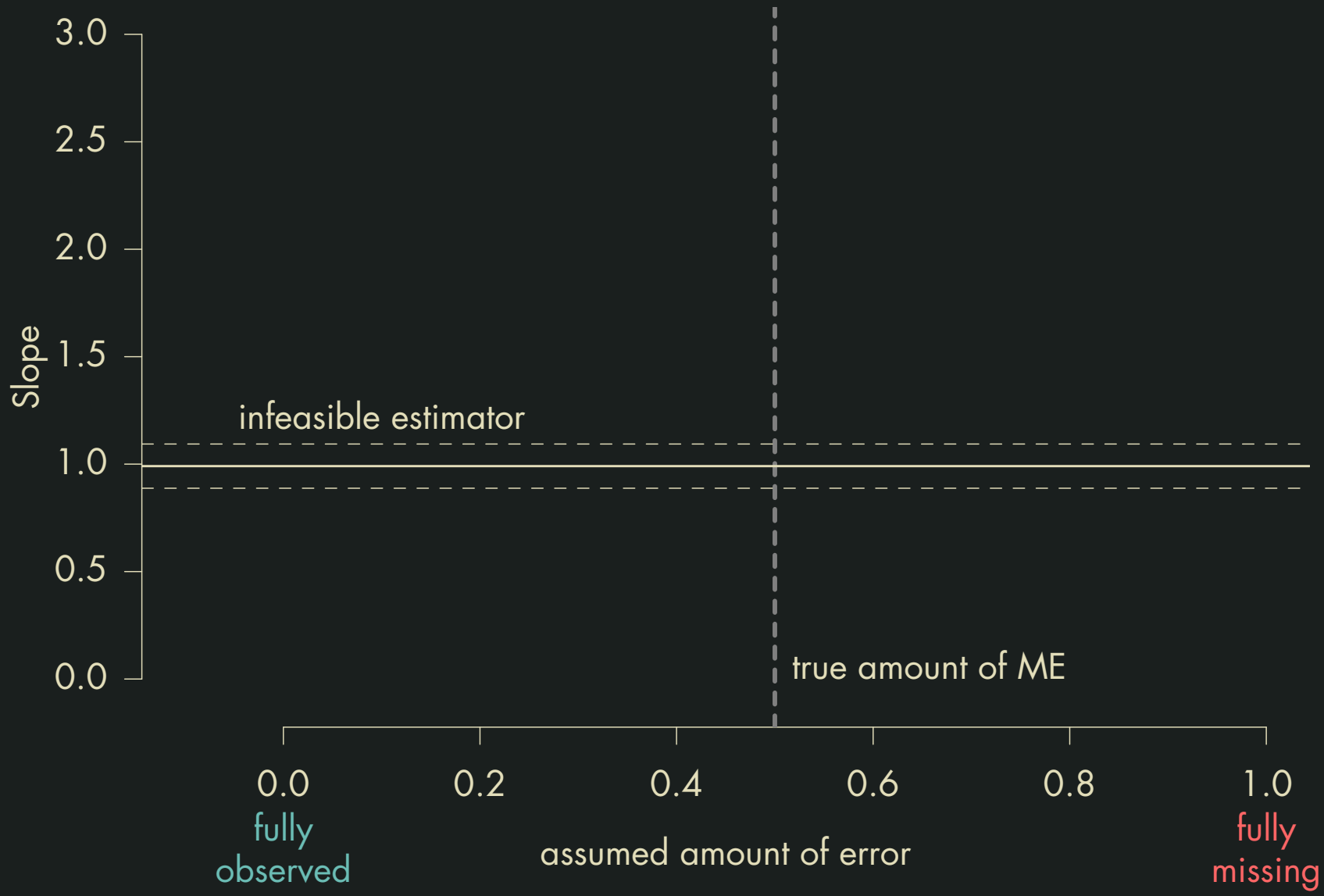

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
0.0
0.2
0.4
0.6
0.8
1.0
assumed amount of error

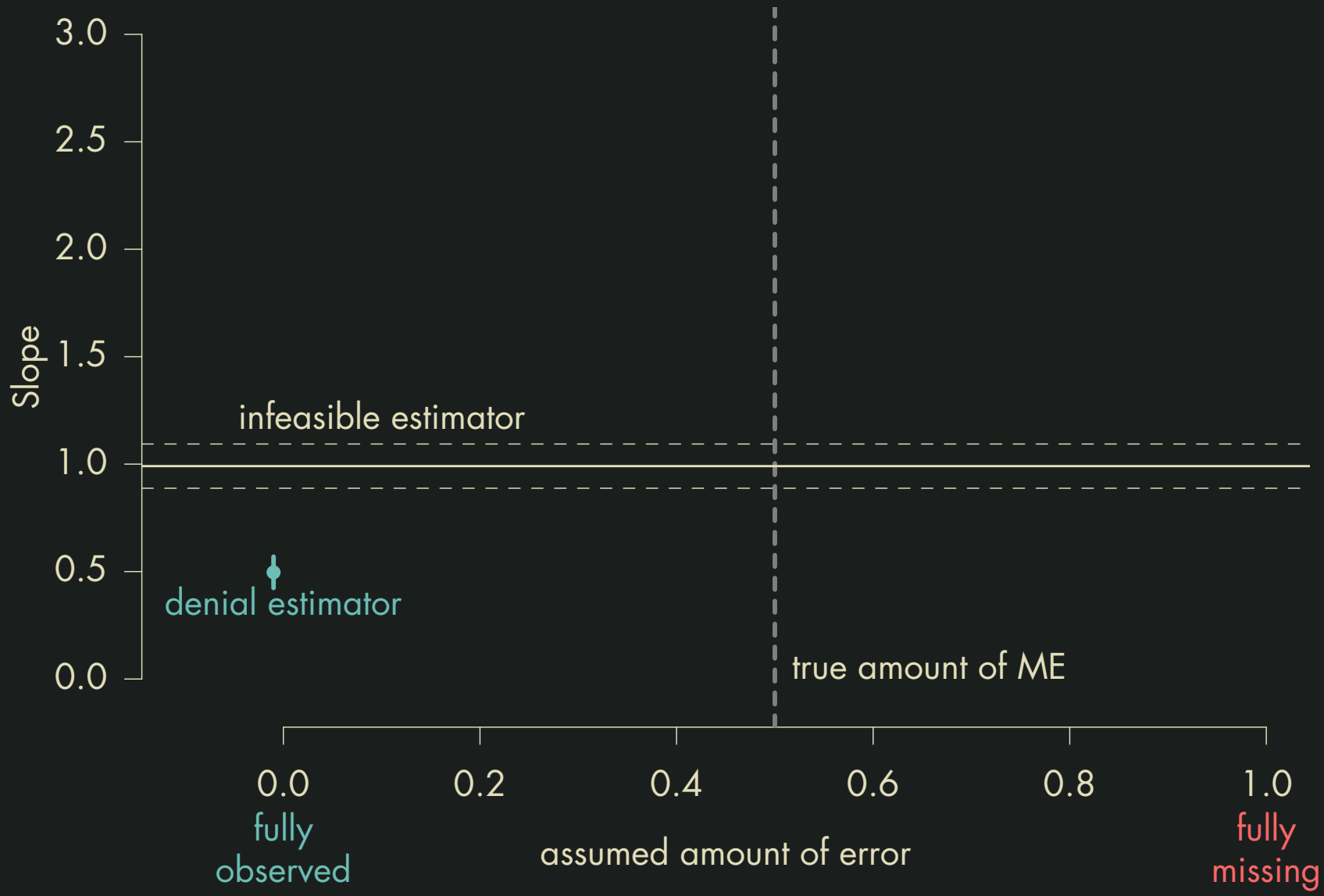

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error

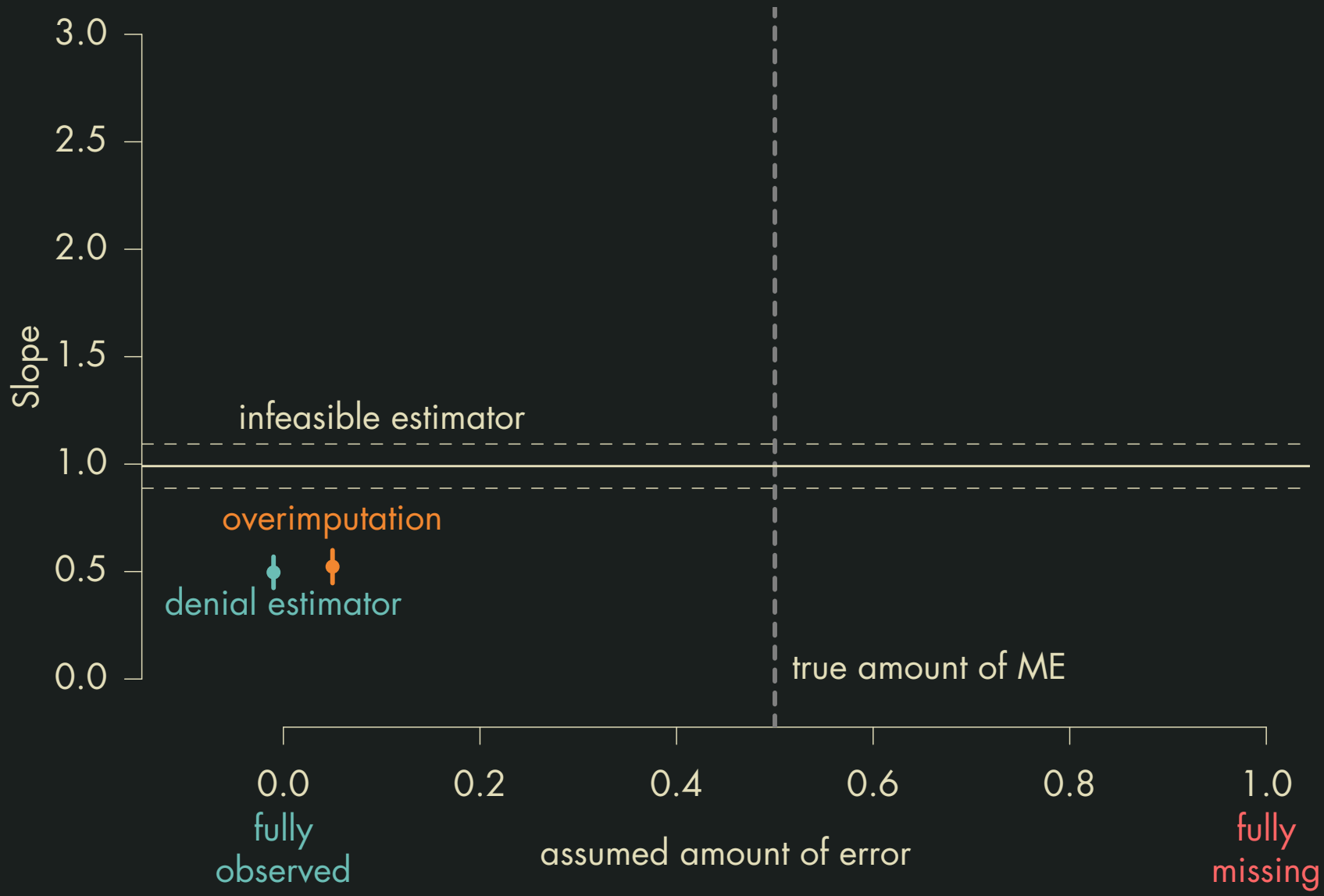

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
0.0
fully
observed
0.2
0.4
0.6
0.8
1.0
fully
missing
assumed amount of error

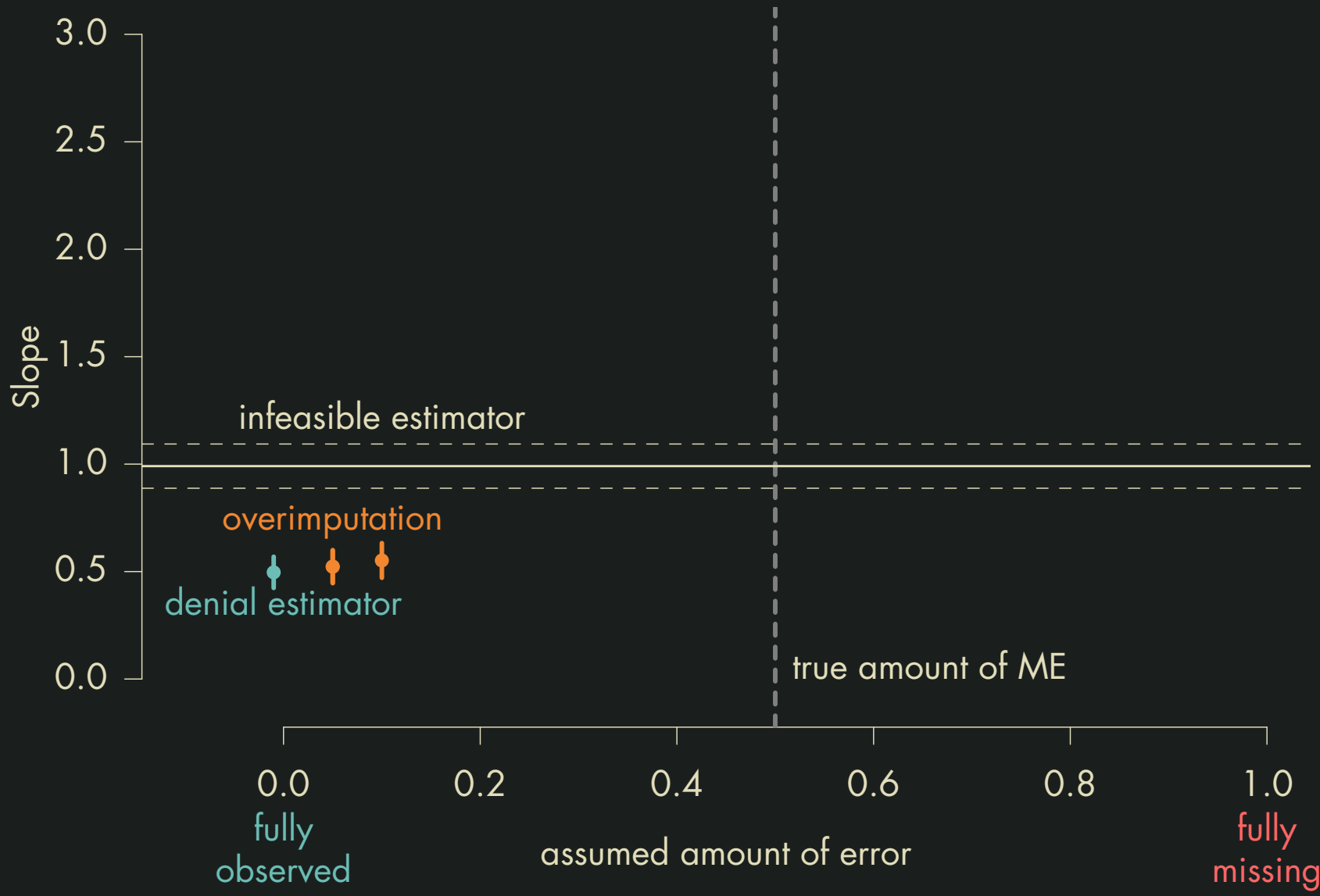

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
true amount of ME
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error
fully missing

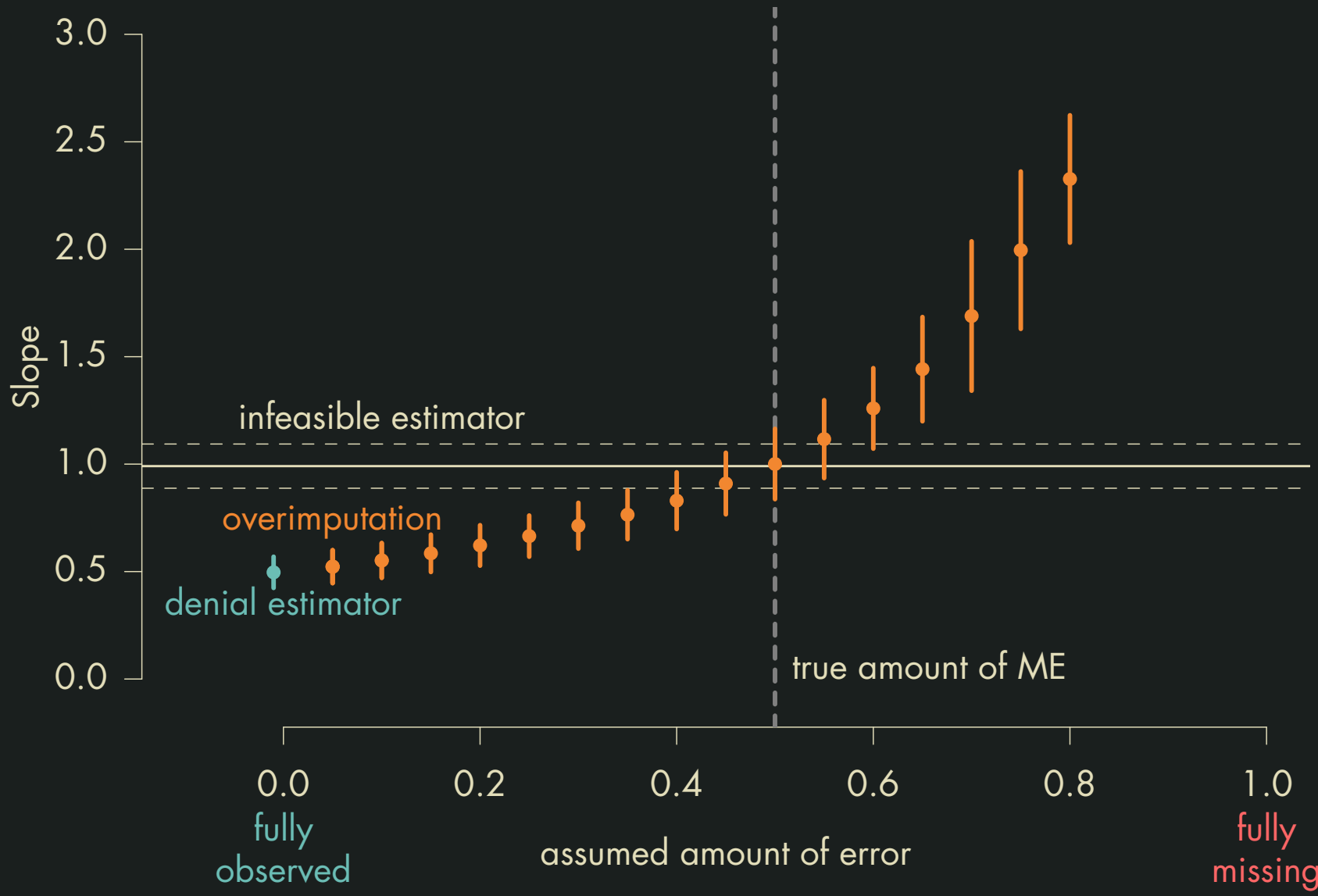

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
true slope
true amount of ME
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error
fully missing

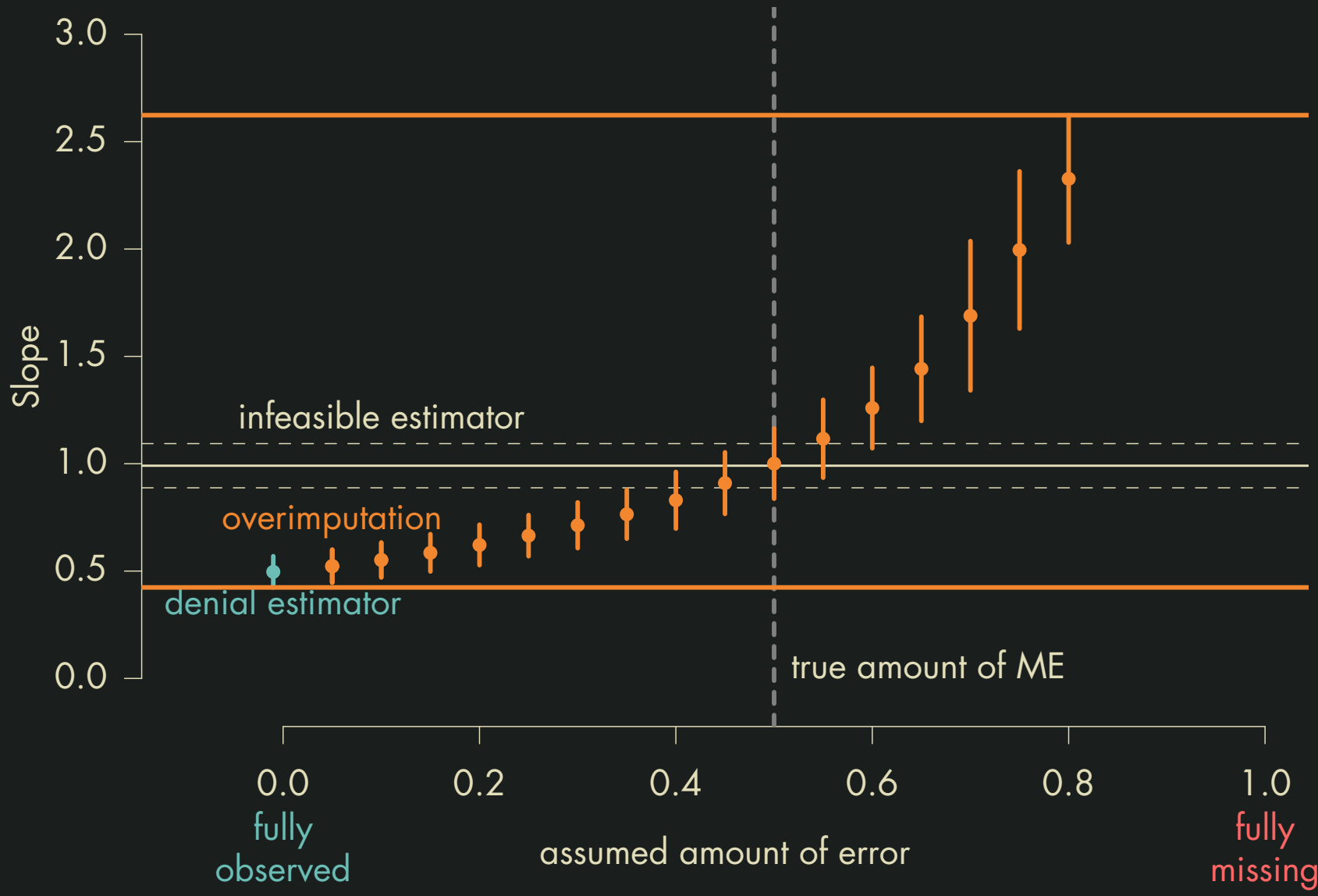

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
true amount of ME
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error
fully missing


3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
denial estimator
true amount of ME
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error
fully missing

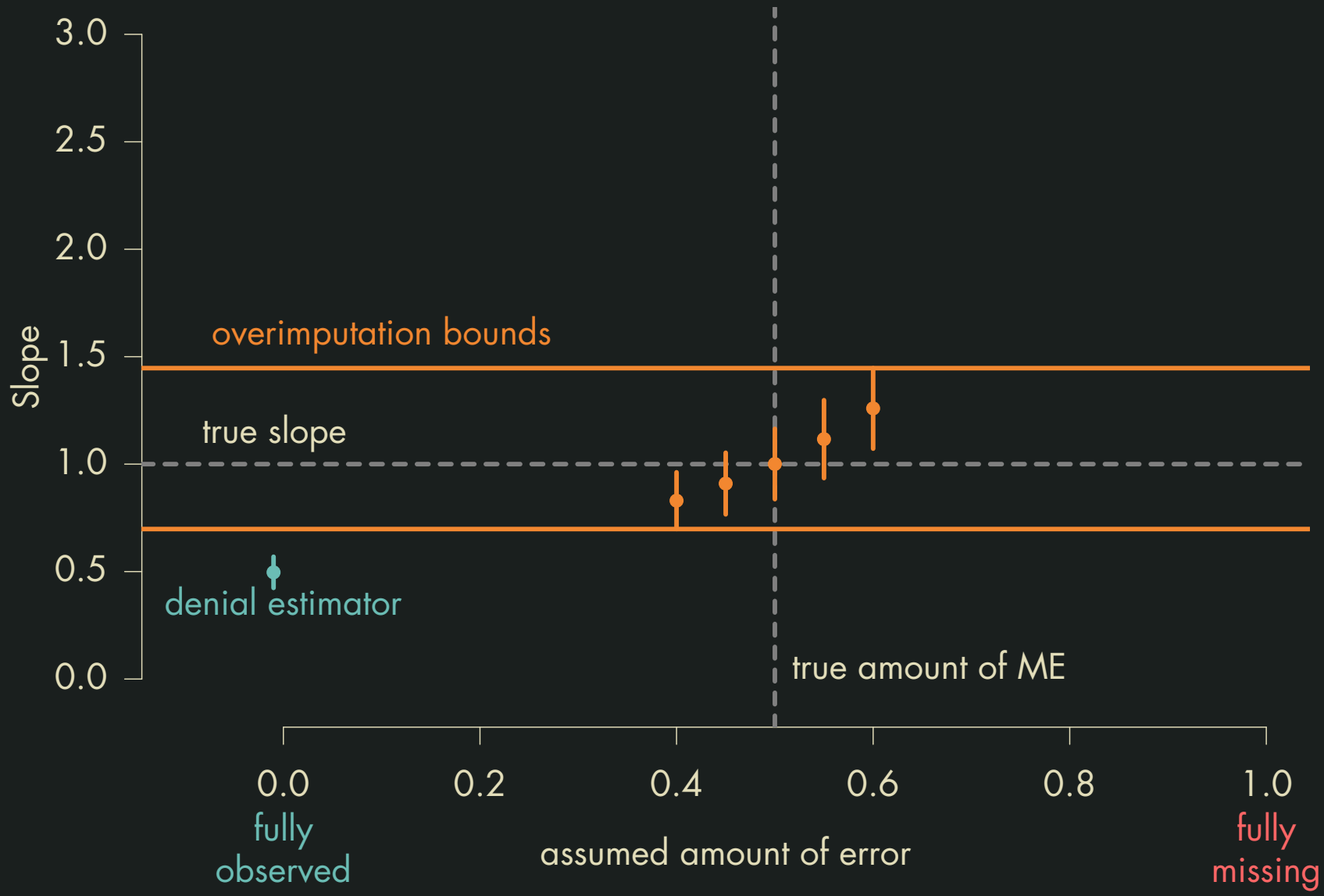

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
overimputation
denial estimator
true amount of ME
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error
fully missing


3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
overimputation
denial estimator
true amount of ME
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error
fully missing

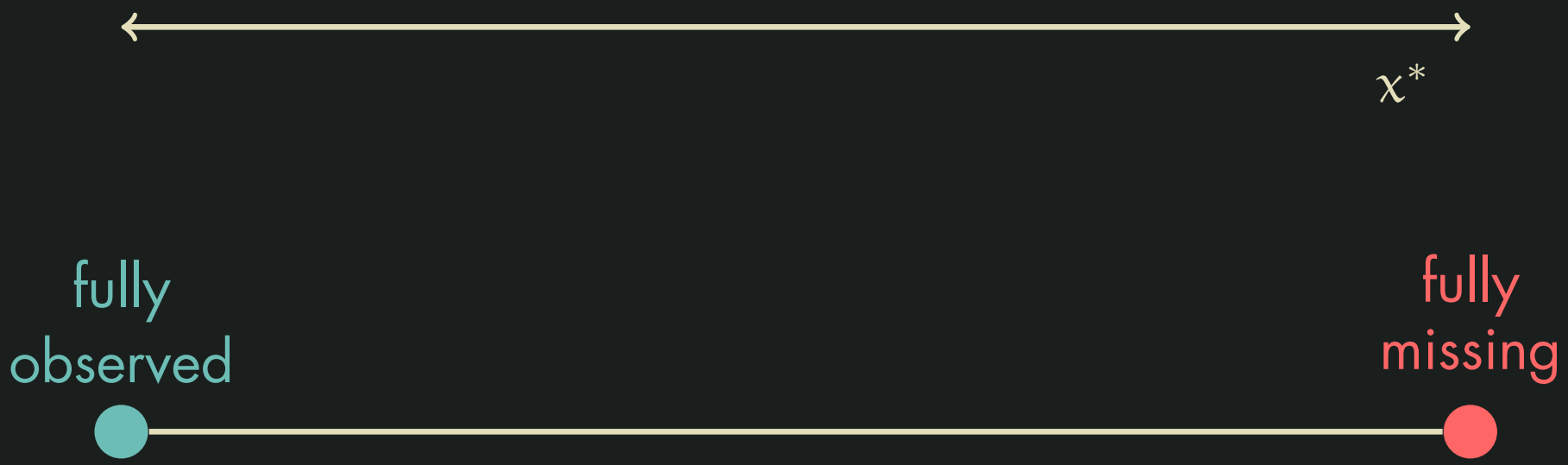

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
overimputation
denial estimator
true amount of ME
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error
fully missing


3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
overimputation
denial estimator
true amount of ME
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error
fully missing

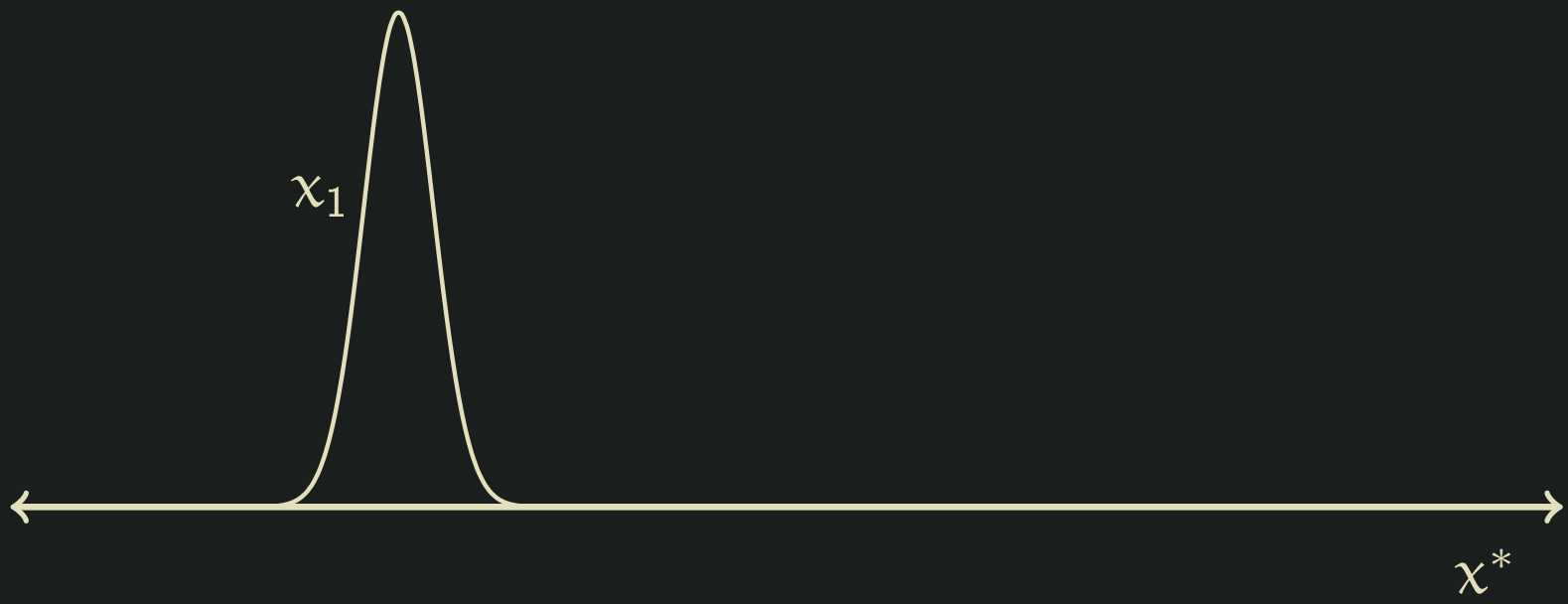
CHOOSE A RANGE OF σ_u^2


3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
overimputation bounds
true slope
denial estimator
true amount of ME
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error
fully missing

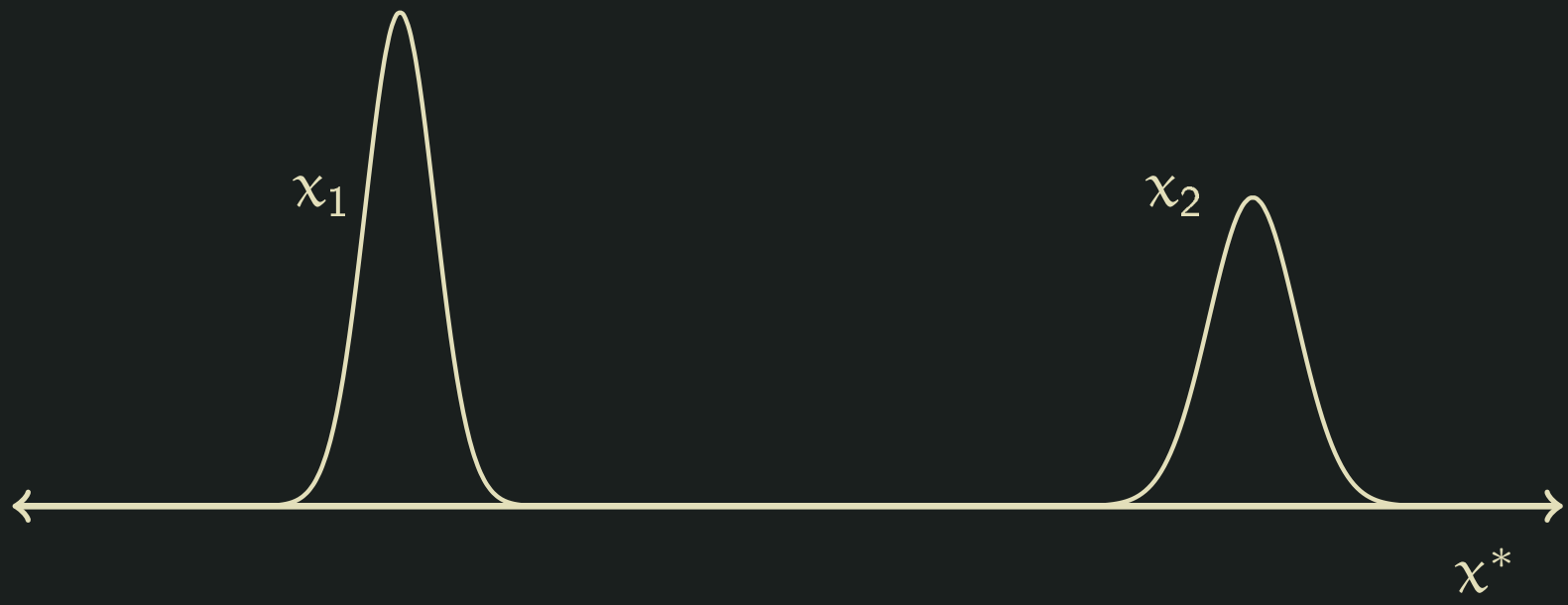
HETEROSKEDASTIC ERRORS

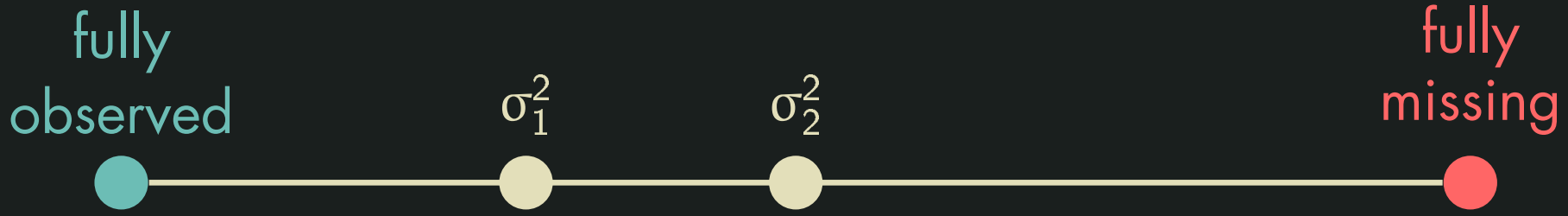


HETEROSKEDASTIC ERRORS

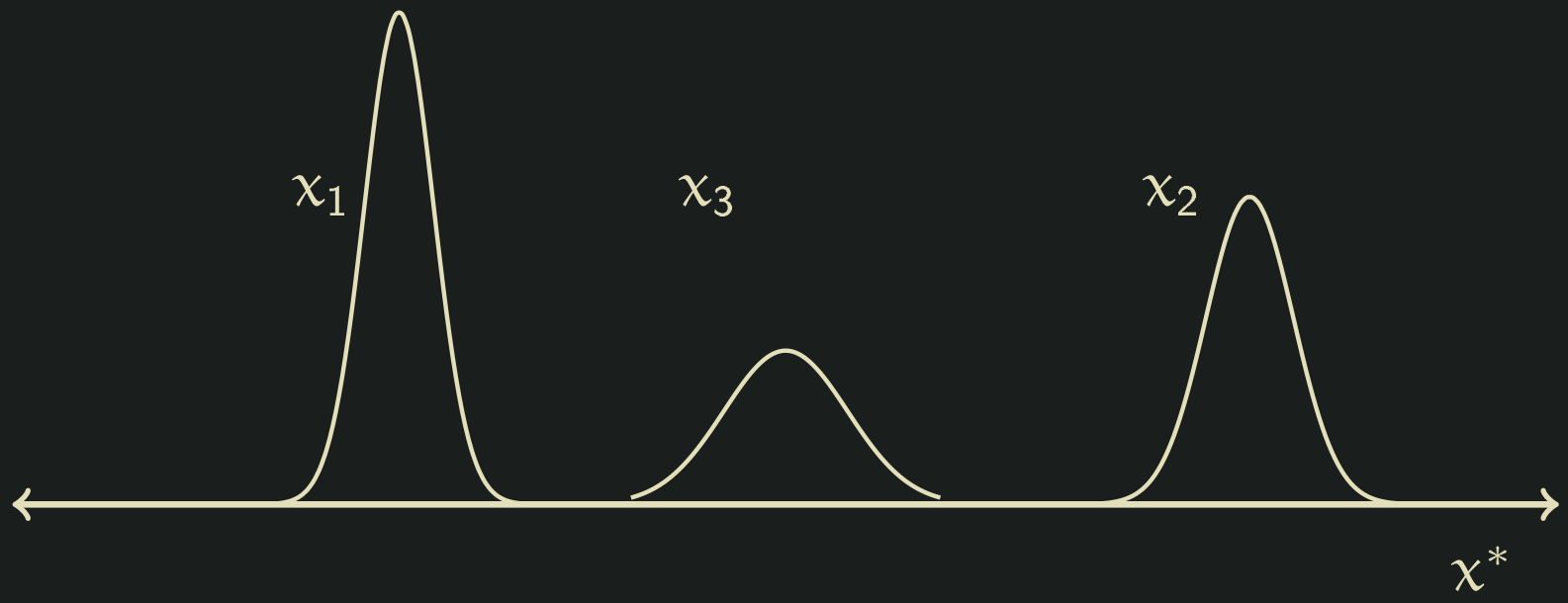


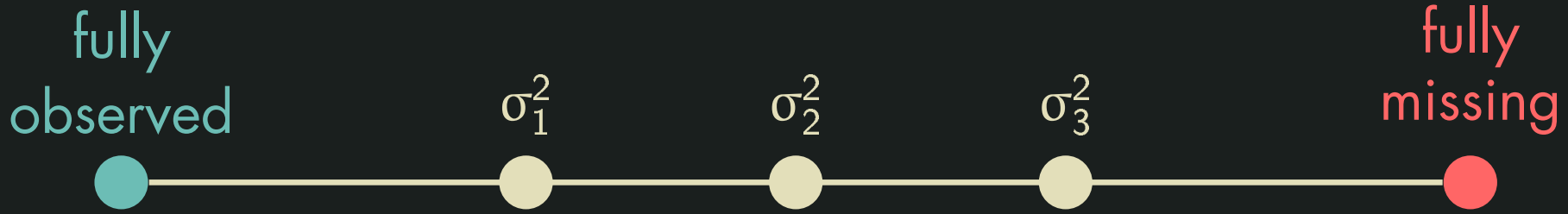
HETEROSKEDASTIC ERRORS



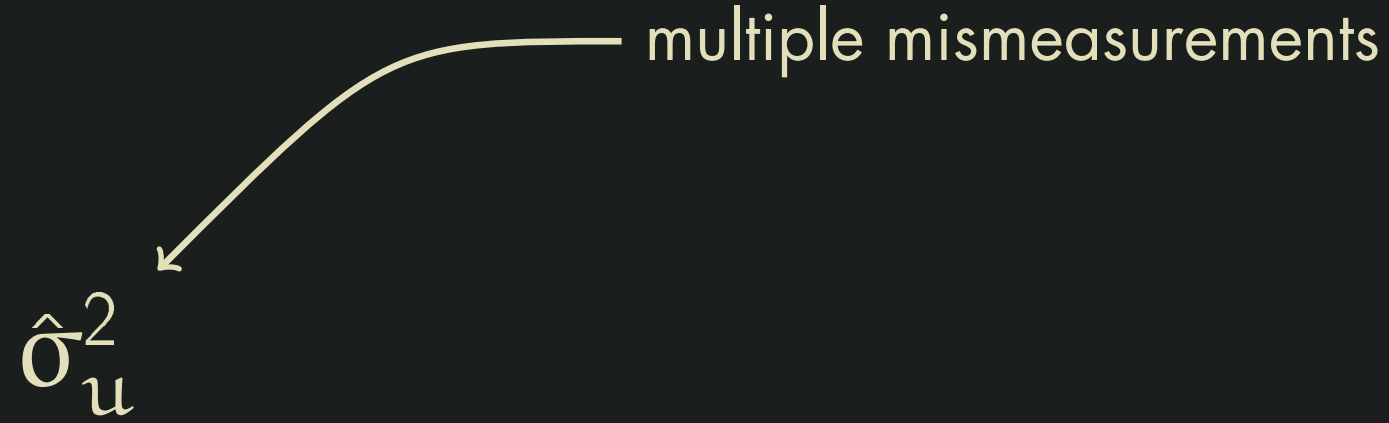


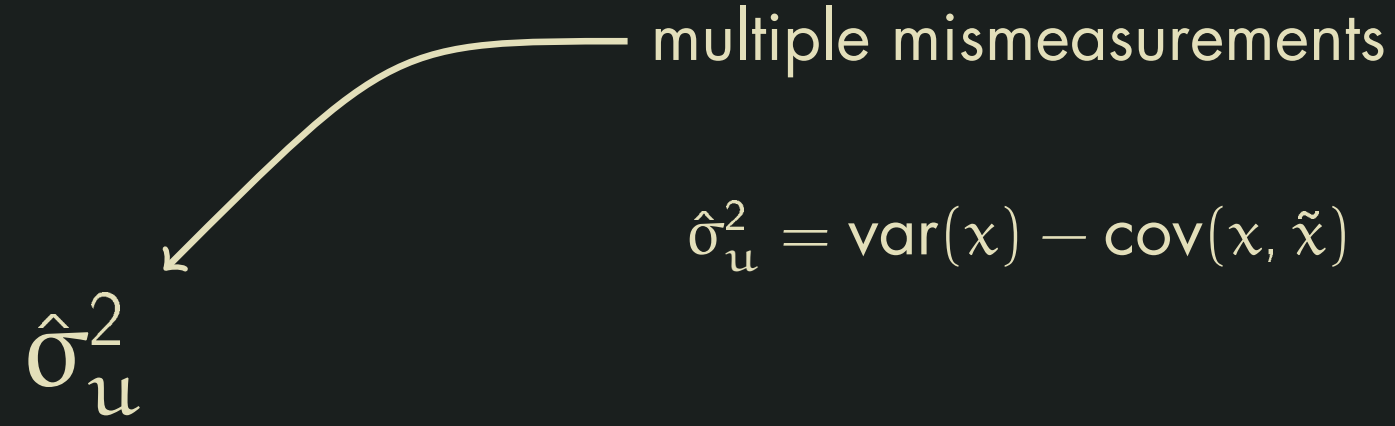
HETEROSKEDASTIC ERRORS

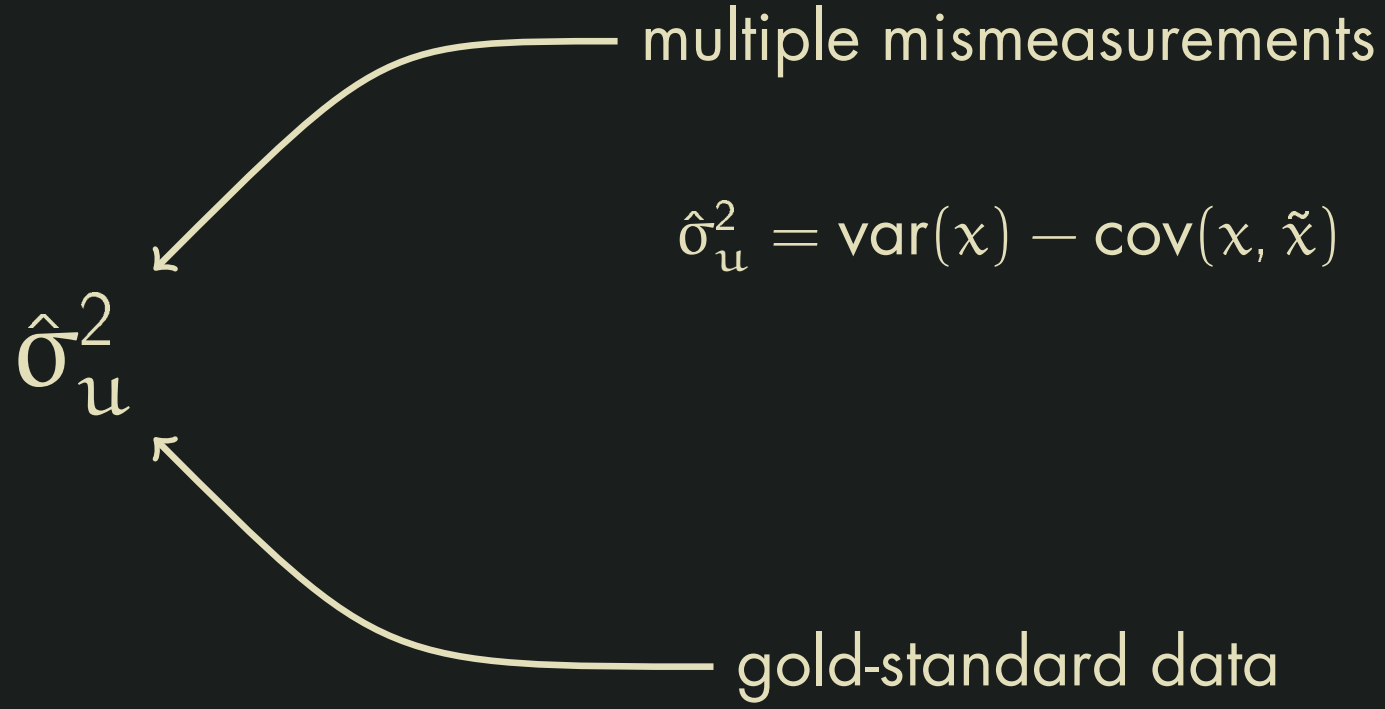


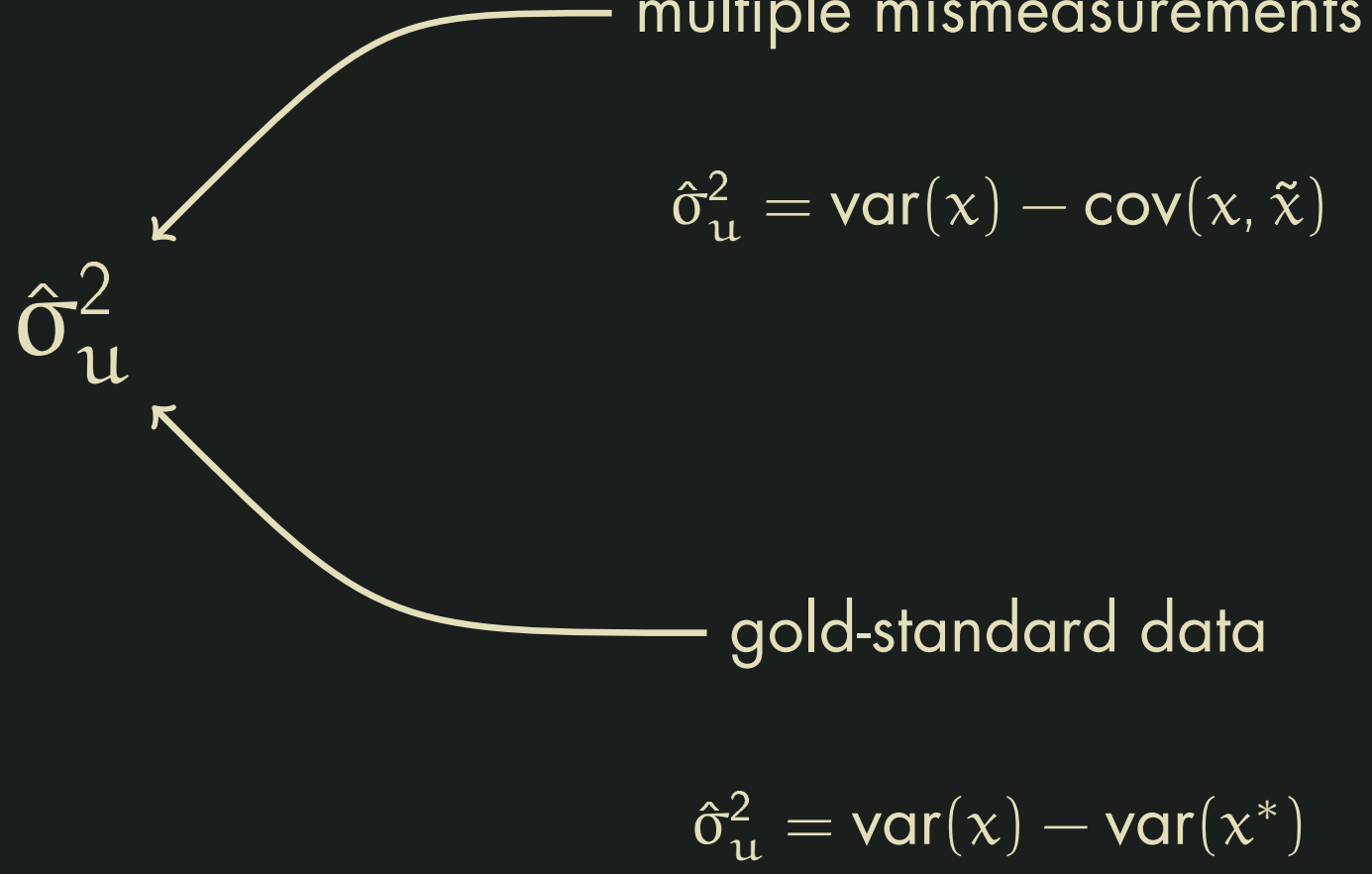


$\hat{\sigma}_u^2$


multiple mismeasurements
$\hat{\sigma}_u^2$


multiple mismeasurements
$\hat{\sigma}^2_u = \text{var}(x) - \text{cov}(x, \tilde{x})$
$\hat{\sigma}^2_u$

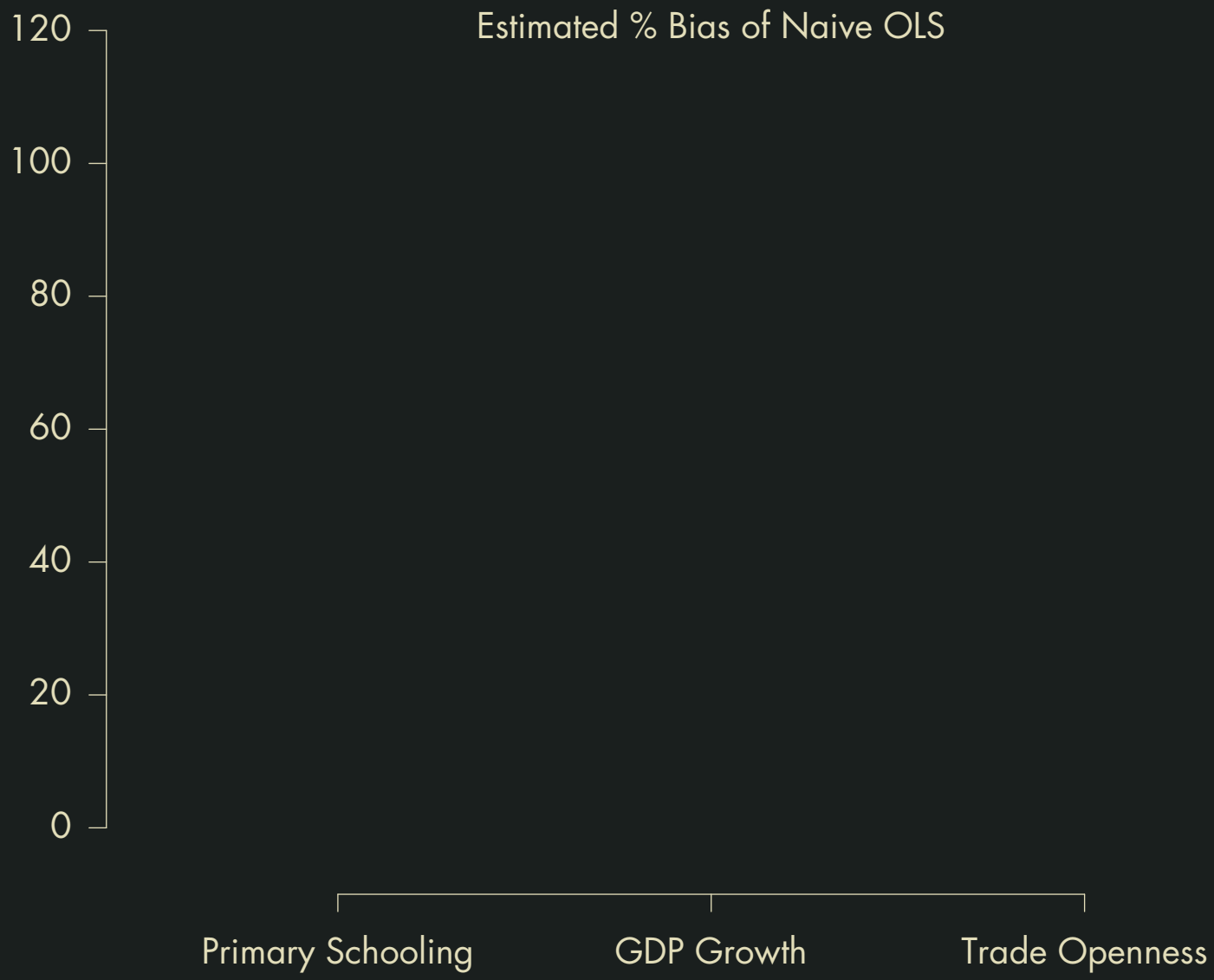

multiple mismeasurements
$\hat{\sigma}^2_u = \text{var}(x) - \text{cov}(x, \tilde{x})$
$\hat{\sigma}^2_u$
gold-standard data


multiple mismeasurements
$\hat{\sigma}_u^2 = \text{var}(x) - \text{cov}(x, \tilde{x})$
$\hat{\sigma}_u^2$
gold-standard data
$\hat{\sigma}_u^2 = \text{var}(x) - \text{var}(x^*)$

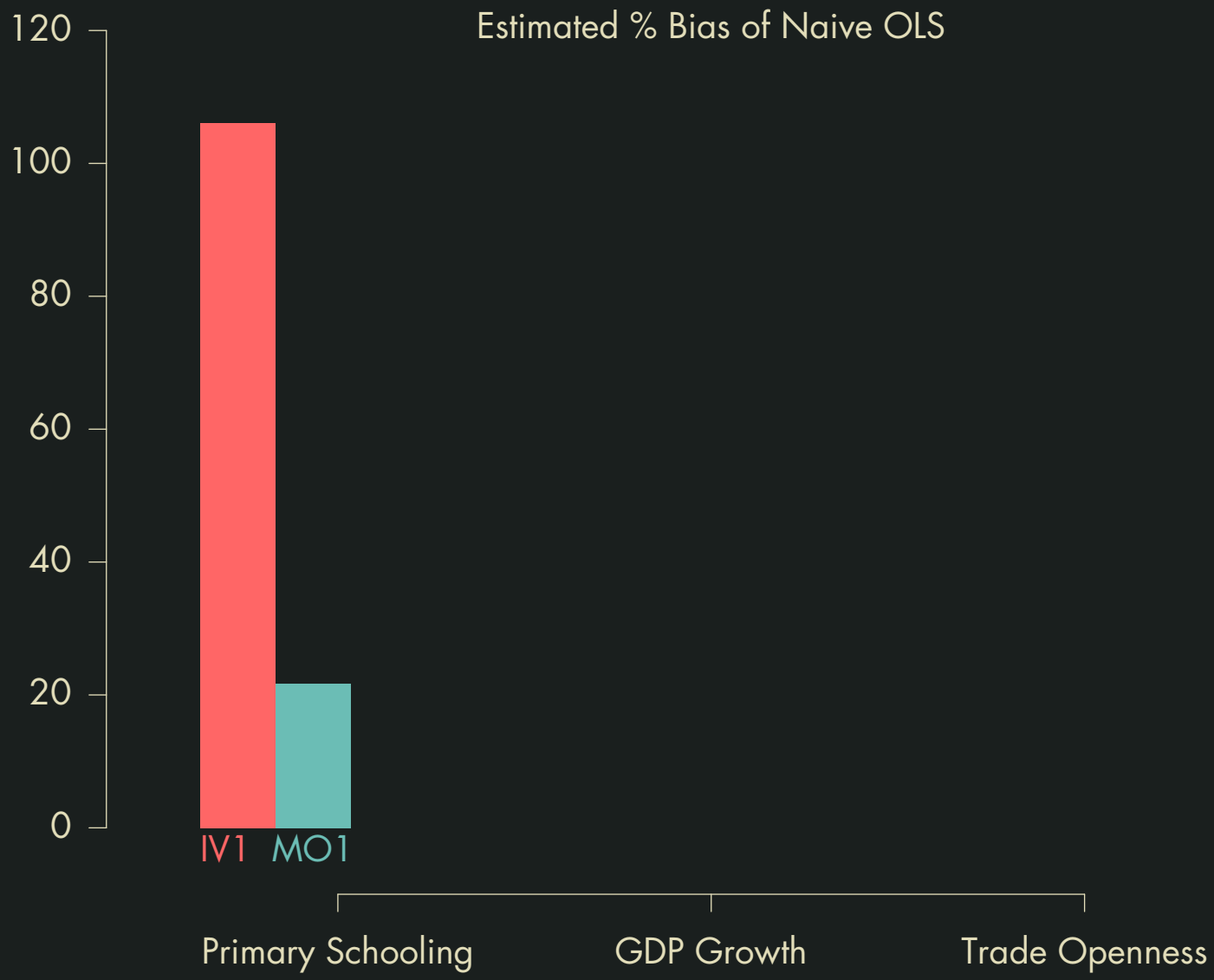
PART SIX:

Application and contrast with other methods.

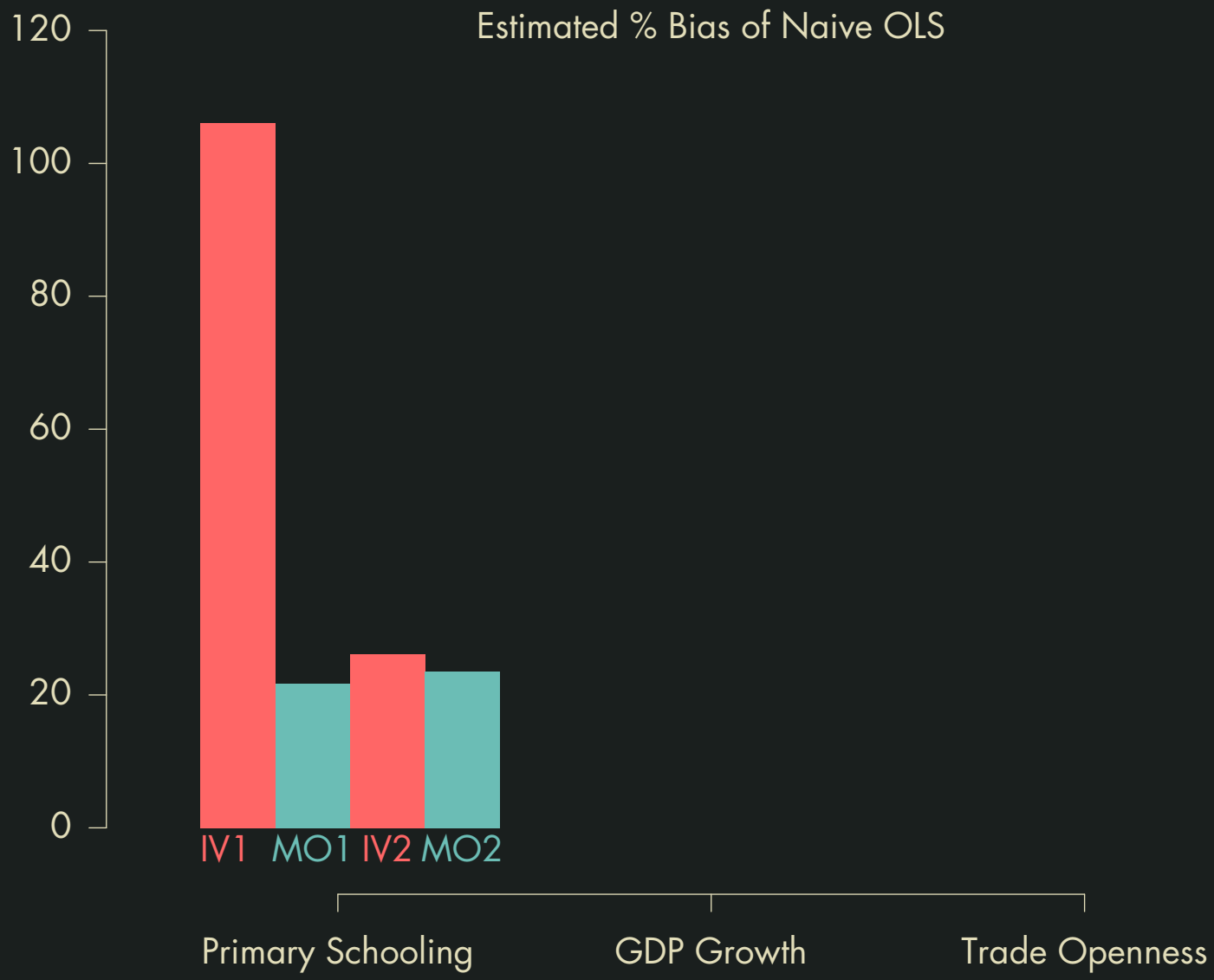
MO is less model dependent



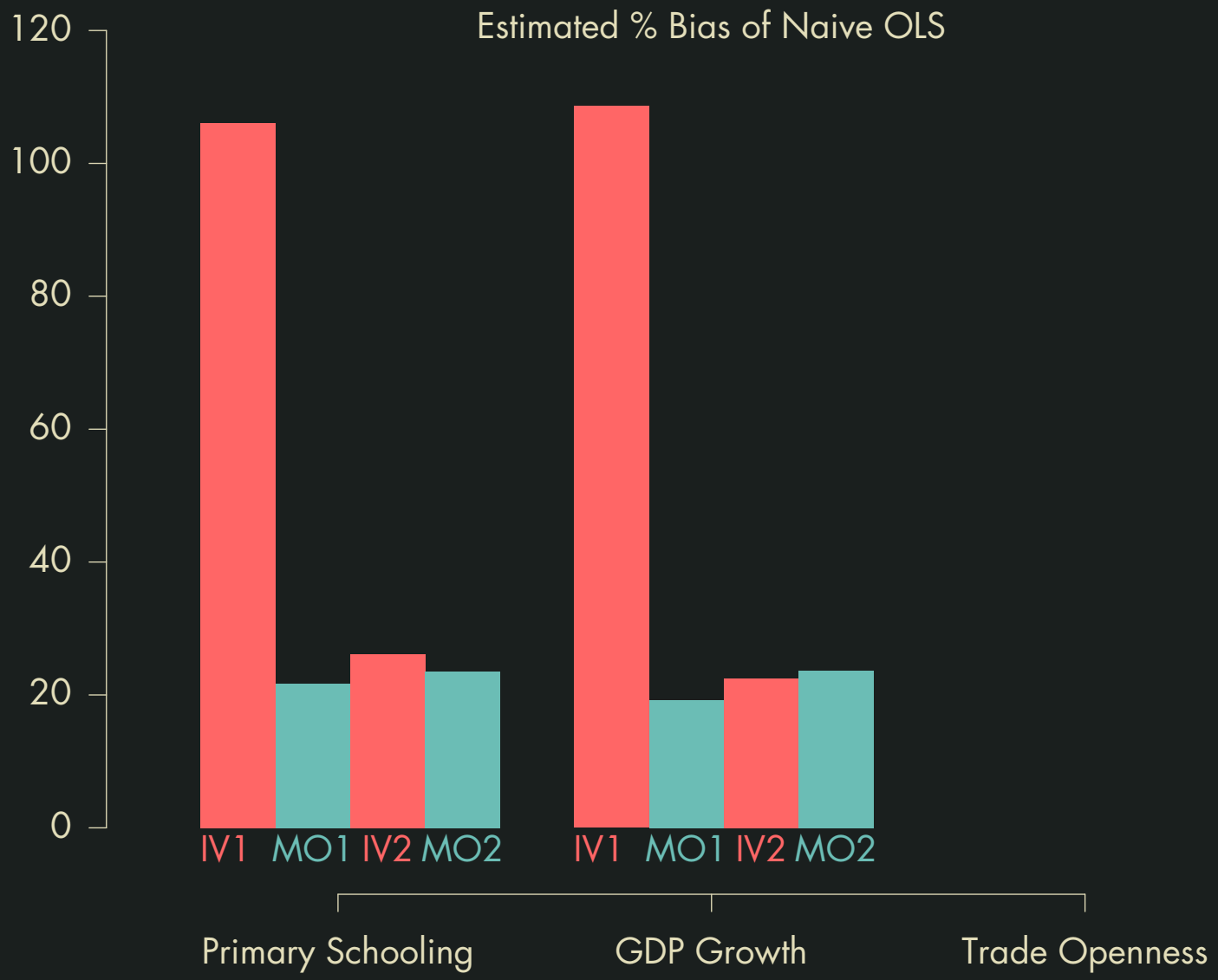
MO is less model dependent



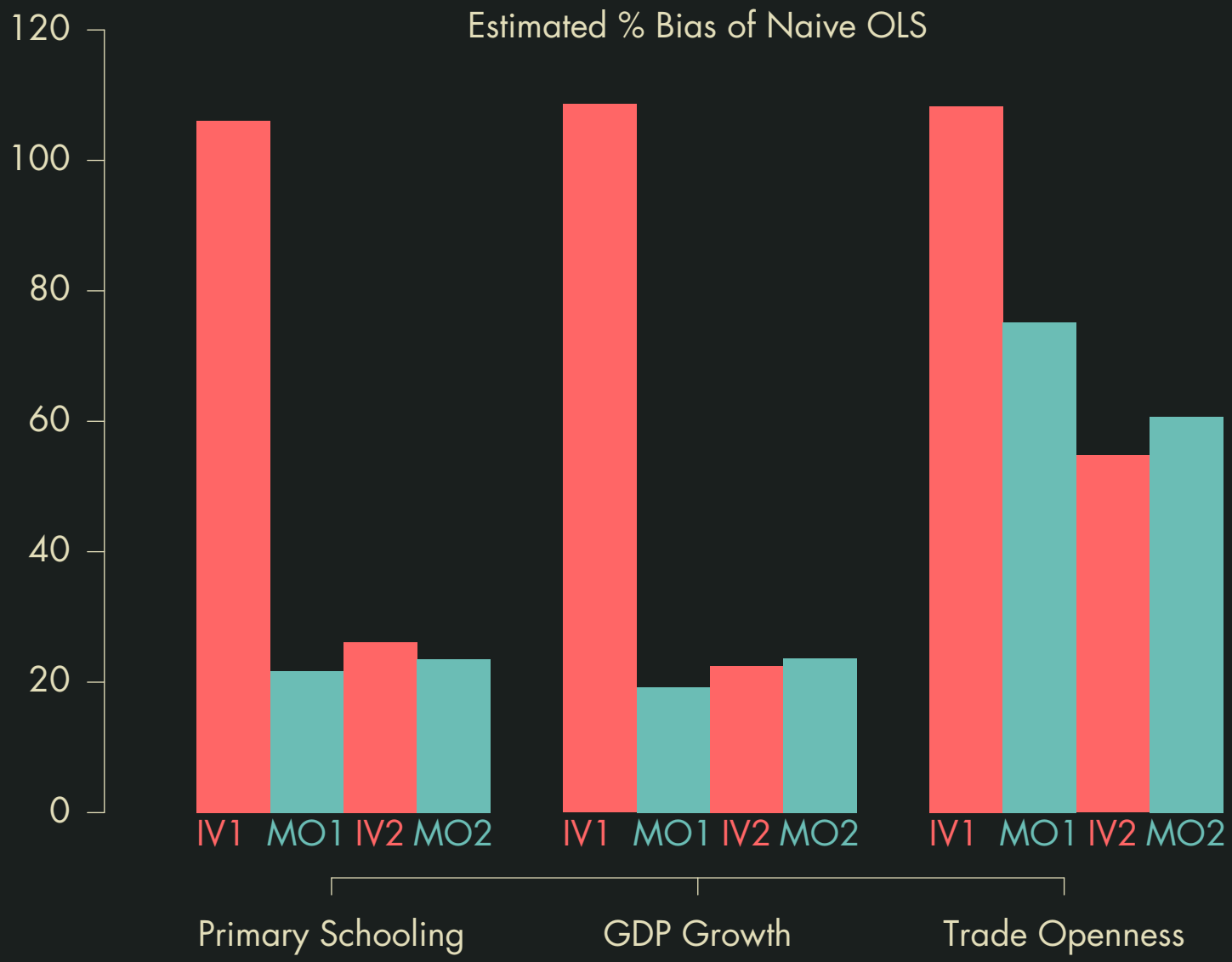
MO is less model dependent



MO is less model dependent

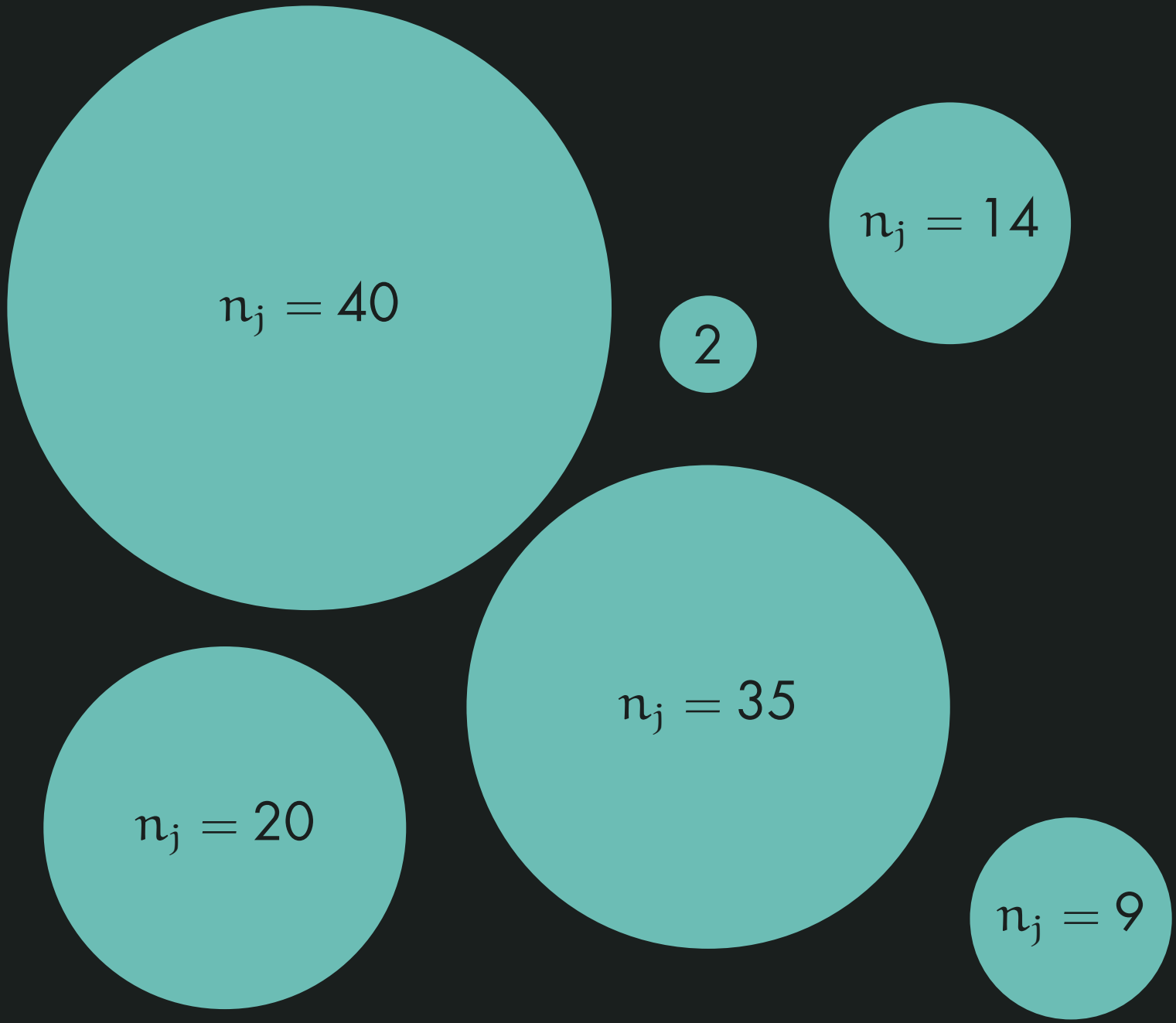


MO is less model dependent

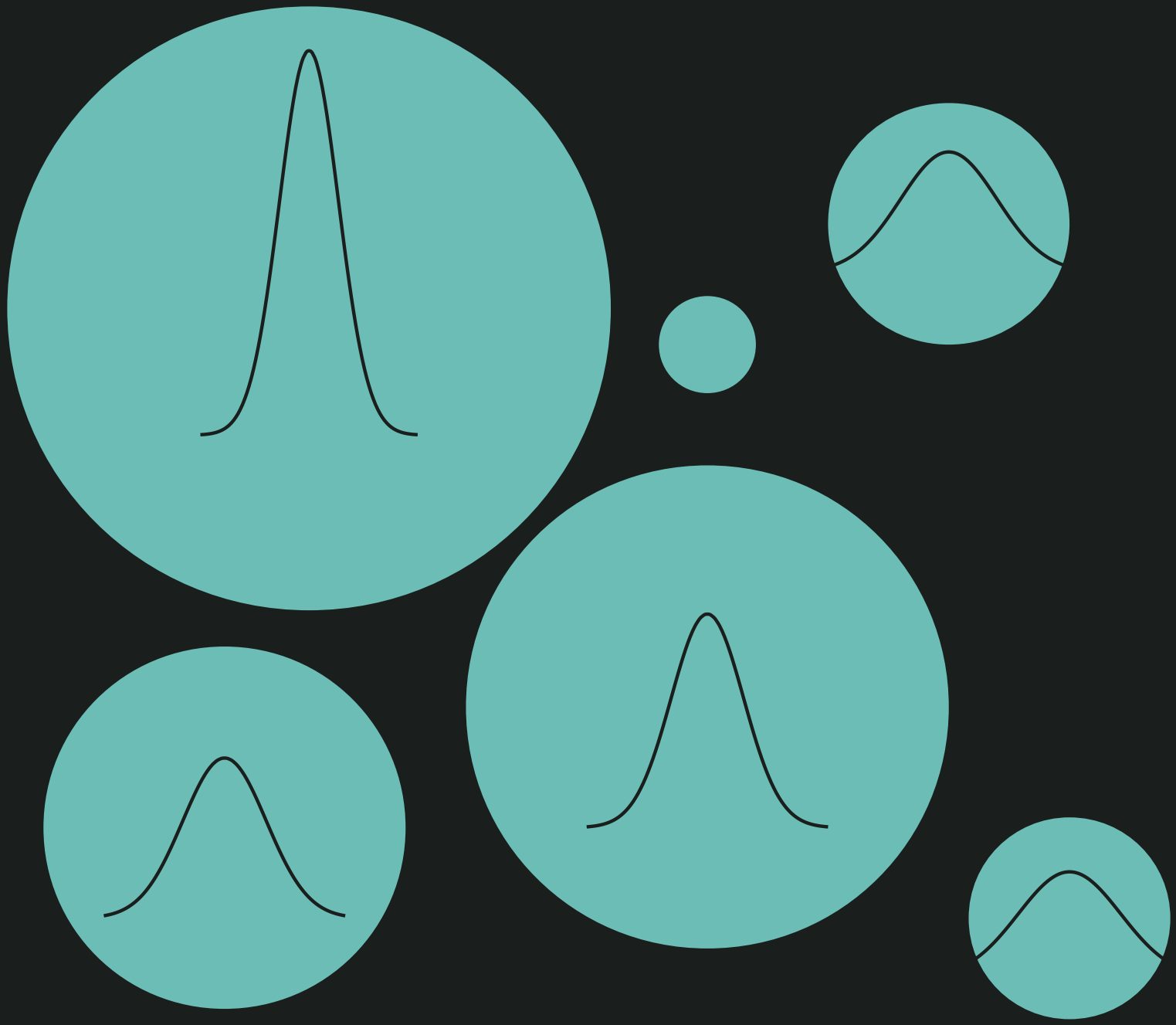


Small Area Estimation

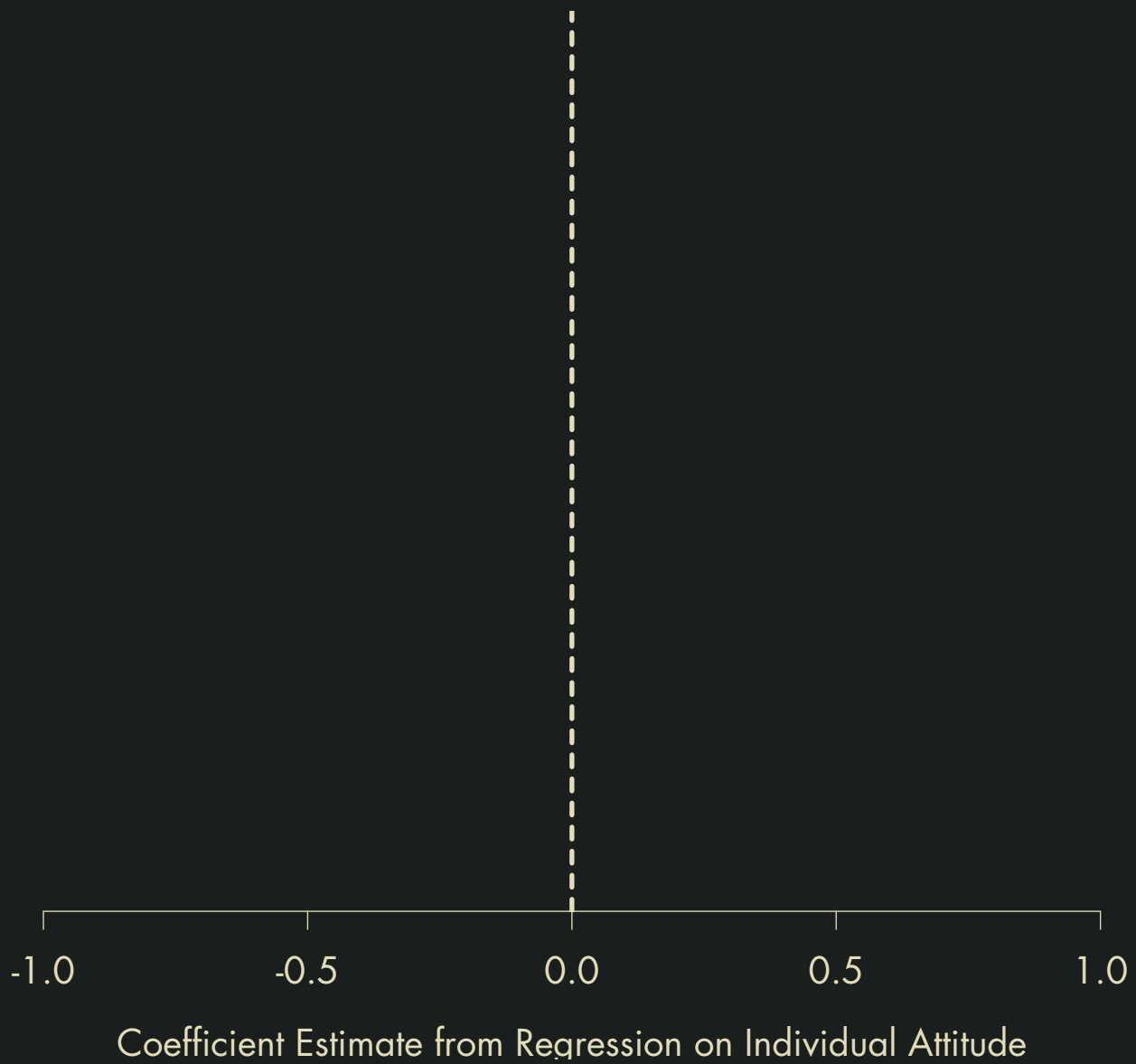
Small Area Estimation



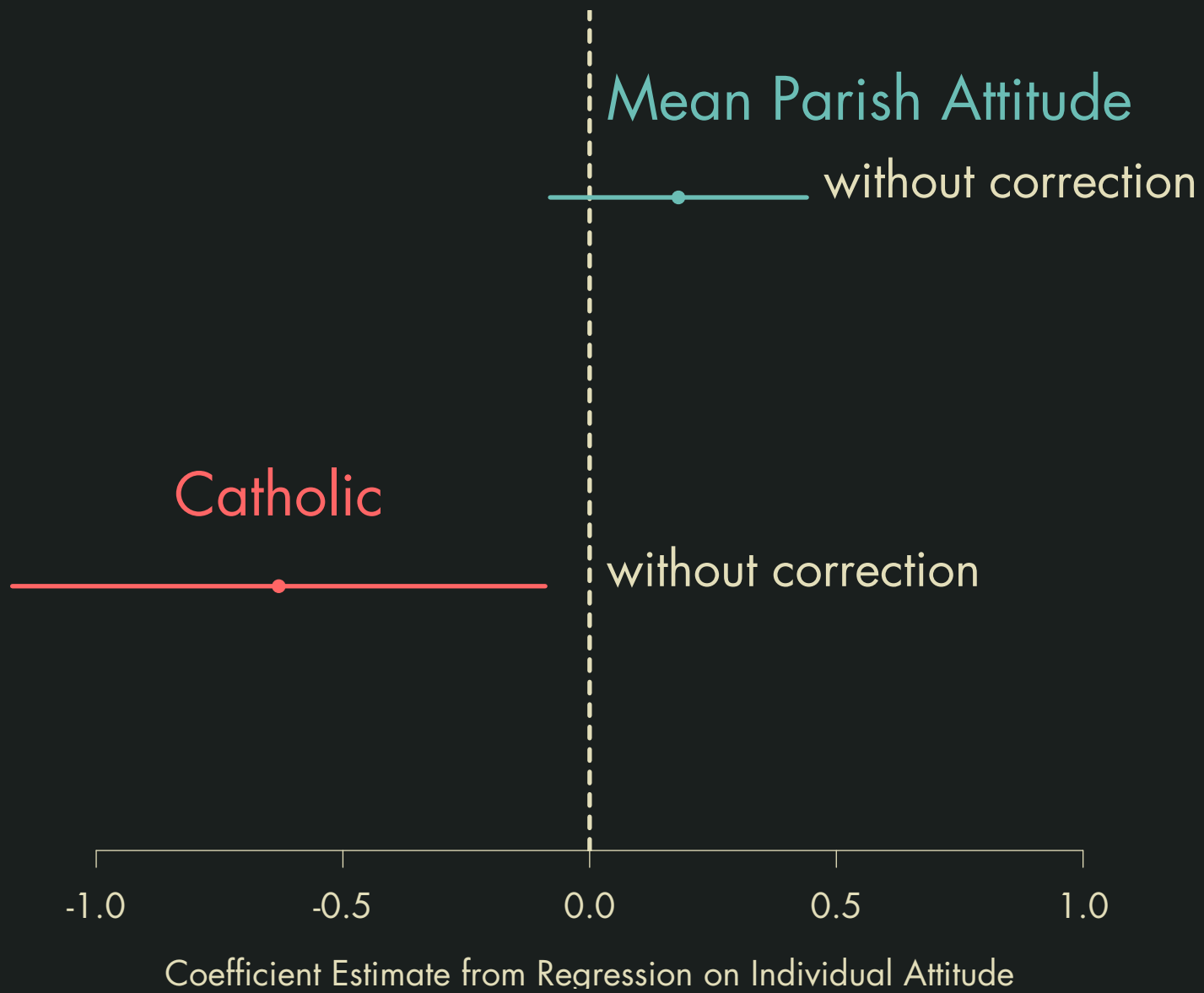
Small Area Estimation



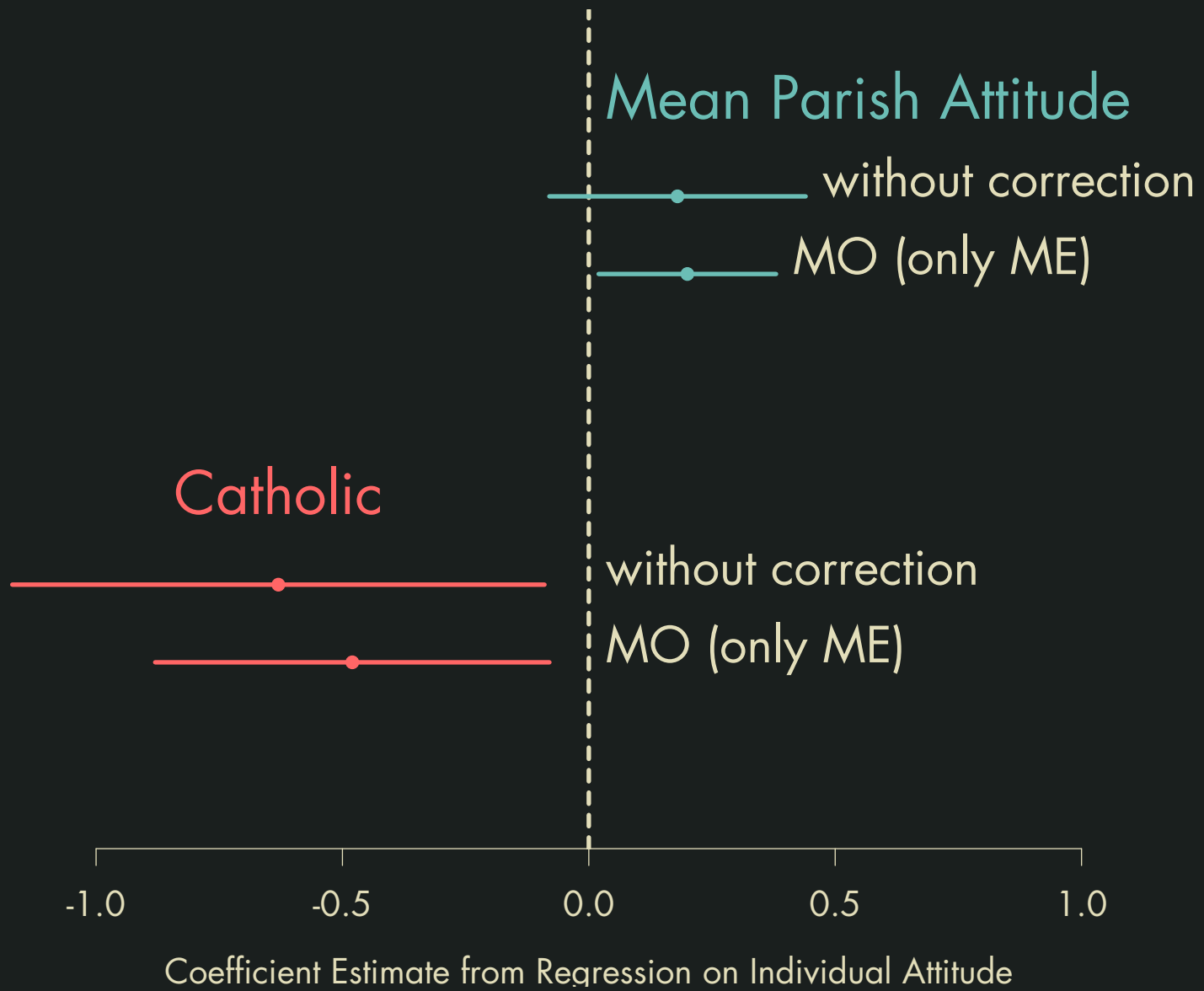
MO makes a difference



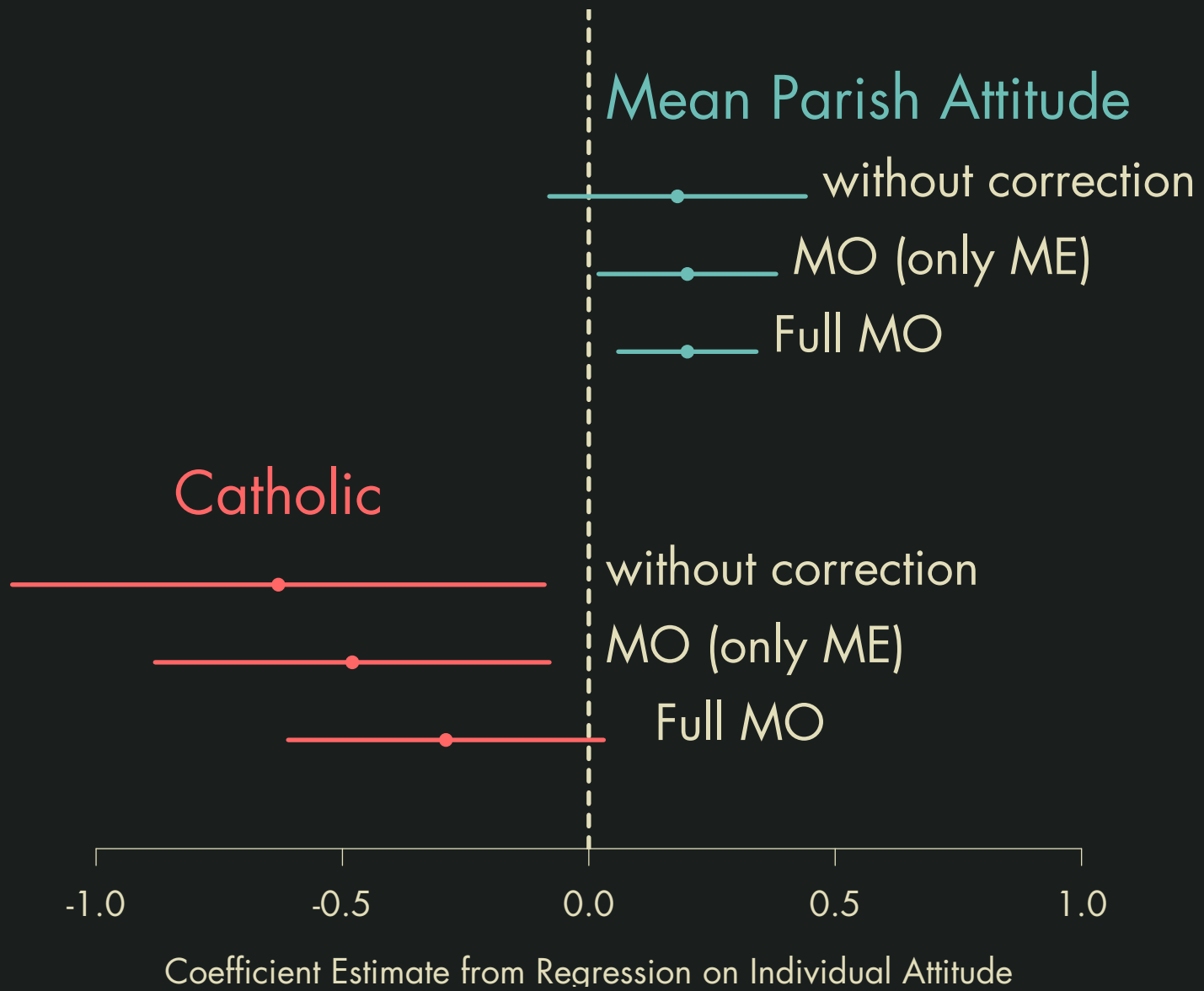
MO makes a difference



MO makes a difference

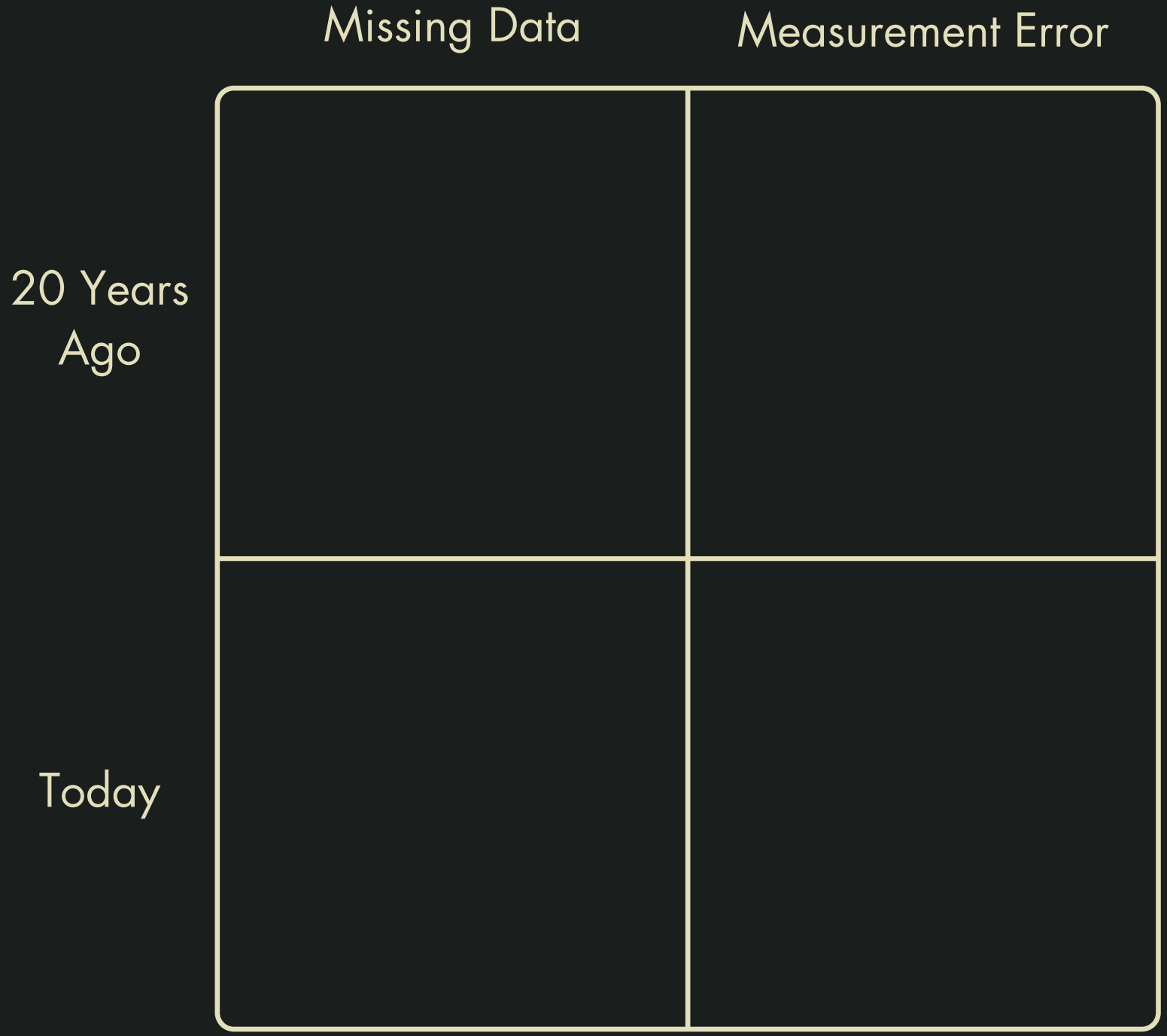


MO makes a difference



PART SEVEN:

A tale of two literatures.


Missing Data
Measurement Error
20 Years Ago
Today

Missing Data
Measurement Error
TAILORED METHODS:
TAILORED METHODS:
20 Years Ago
Today

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent	TAILORED METHODS: Model dependent
Today		

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement	TAILORED METHODS: Model dependent Difficult to implement
Today		

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today		

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today	MULTIPLE IMPUTATION:	

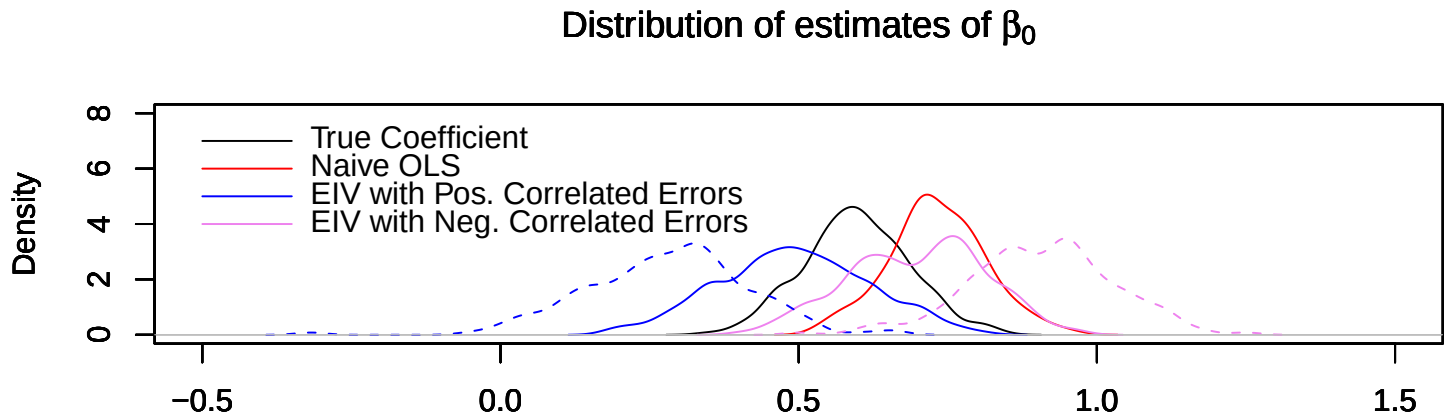
	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today	MULTIPLE IMPUTATION: Broadly applicable	

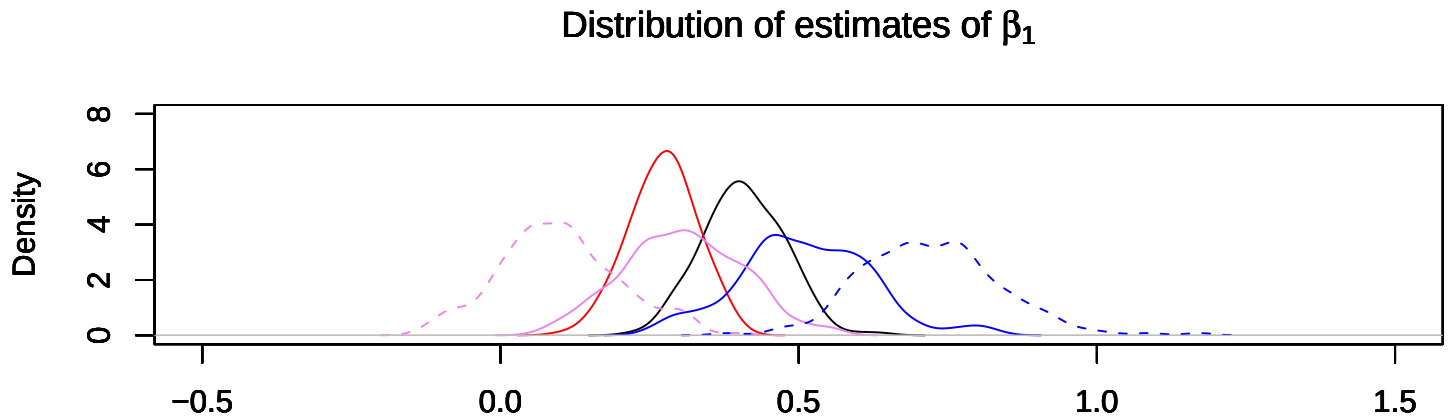
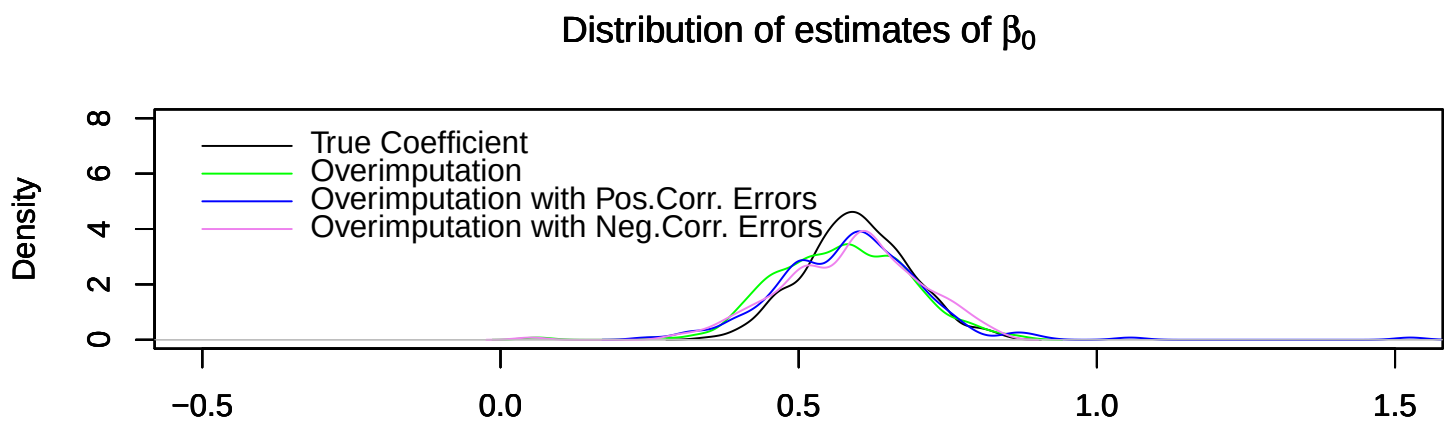
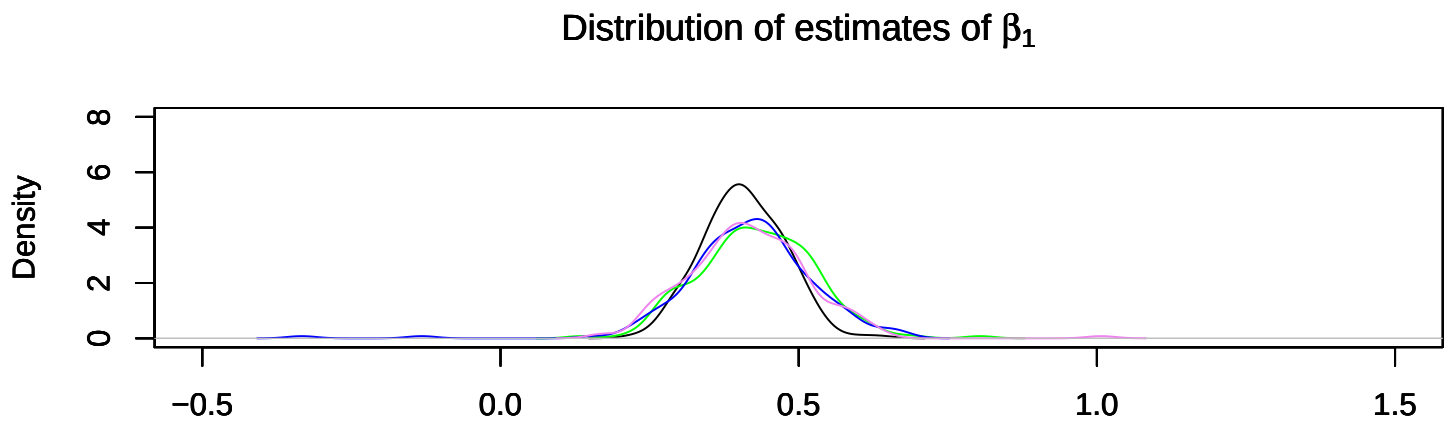
	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today	MULTIPLE IMPUTATION: Broadly applicable Easy to implement	

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today	MULTIPLE IMPUTATION: Broadly applicable Easy to implement Widely used.	

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today	MULTIPLE IMPUTATION: Broadly applicable Easy to implement Widely used.	MULTIPLE OVERIMPUTATION

What if measurement error is correlated with the outcome? (Instrumental Variables)





What if measurement error is correlated with the outcome? (Multiple Overimputation)