CERTIFIED SYSTEM IDENTIFICATION

Marco C. Campi
University of Brescia
Italy

CERTIFIED SYSTEM IDENTIFICATION

TOWARDS DISTRIBUTION-FREE RESULTS

Marco C. Campi

University of Brescia

Italy

with:

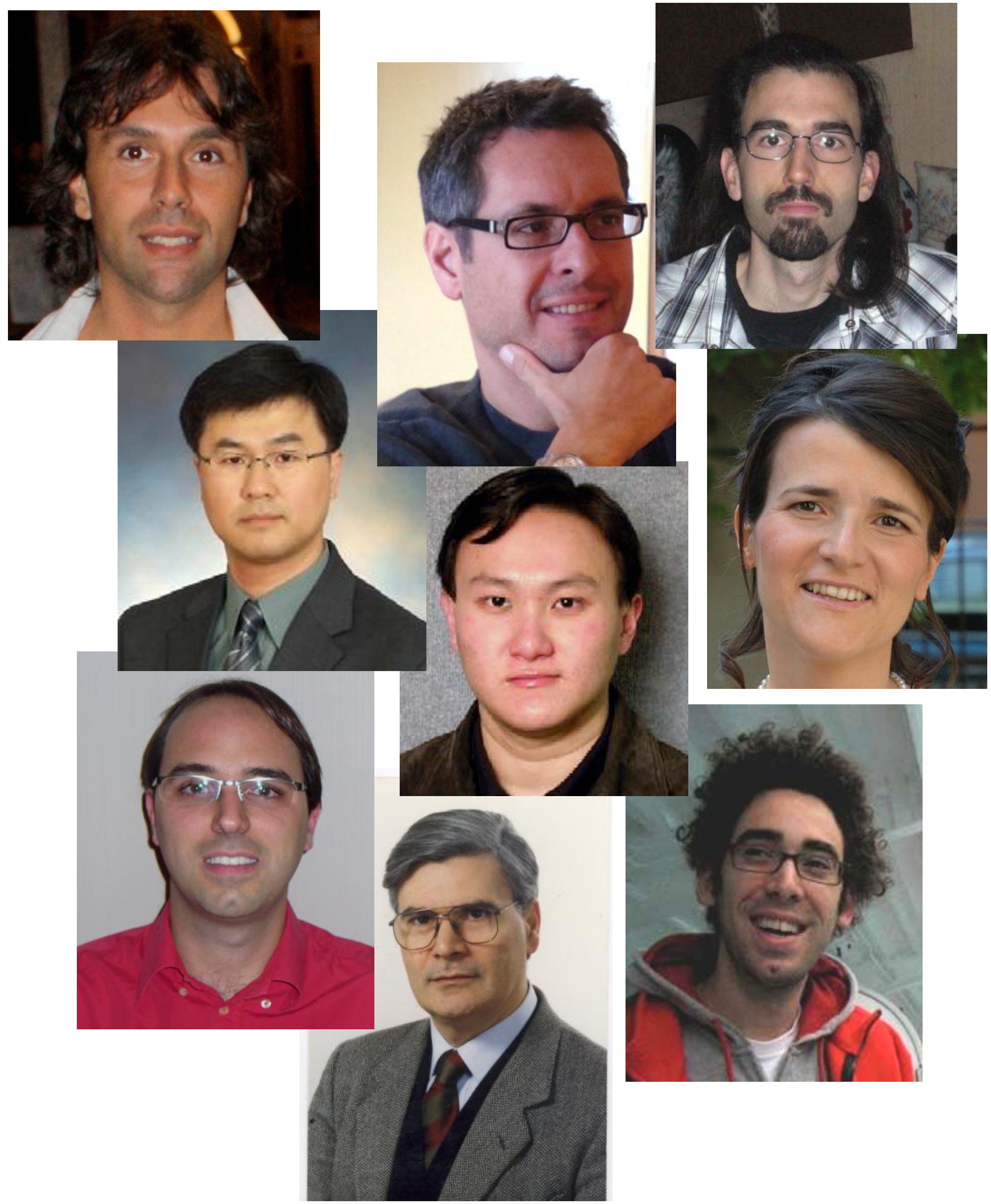
Simone Garatti

Erik Weyer

Balazs Csaji

thanks to :

- *Fabio Baronio*
- *Sergio Bittanti*
- *Giuseppe Calafiore*
- *Algo Care'*
- *Sangho Ko*
- *Su Ki Ooi*
- *Bernardo Pagnoncelli*
- *Maria Prandini*
- *Daniel Reich*



Preamble

Preamble

System identification is the science of constructing models from data

Preamble

System identification is the science of constructing models from data

and

Preamble

System identification is the science of constructing models from data

and

A good system identification method should return models accompanied by certificates guaranteeing the quality of the model

Preamble

System identification is the science of constructing models from data

and

A good system identification method should return models accompanied by certificates guaranteeing the quality of the model

- $\|\theta^o - \widehat{\theta}\| \leq 0.1$
- $y_{t+1} \in$ region $\hat{Y}$ with probability 99%

$\vdots$

Preamble

System identification is the science of constructing models from data

and

A good system identification method should return models accompanied by certificates guaranteeing the quality of the model

Certificates are relevant to the practice of system identification, and are necessary for its scientific use

Preamble

System identification relies on data, and data is the real wealth in a system identification procedure

Preamble

System identification relies on data, and data is the real wealth in a system identification procedure

therefore

Preamble

System identification relies on data, and data is the real wealth in a system identification procedure

therefore

A system identification method should squeeze out all the relevant information contained in the data, for the purpose of constructing models and of certifying their quality

Preamble

This talk is about data, and the quality of models obtained using data

Preamble

This talk is about data, and the quality of models obtained using data

N data points

Preamble

This talk is about data, and the quality of models obtained using data

MODEL

N data points

CERTIFICATE of QUALITY

… a common thread

… a common thread

deterministic, probabilistic $\longrightarrow$ **distribution-free**

... a common thread

deterministic, probabilistic → **distribution-free**

probability exists,
but probability
values do not count

... a common thread

deterministic, probabilistic $\longrightarrow$ **distribution-free**

$$\boxed{\textit{we should look for distribution-free results}}$$

... a common thread

deterministic, probabilistic $\longrightarrow$ **distribution-free**

paradigm shift

A journey through:

- parameter estimation
- prediction
- filtering

A journey through:

- parameter estimation
- prediction
- filtering

an example

an example

$$y_t = \theta^o + n_t$$

an example

$$y_t = \theta^o + n_t$$

10 independent observations

an example

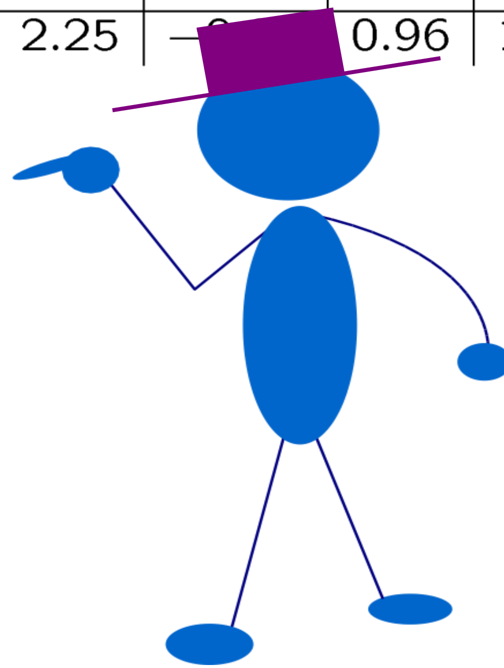
$$y_t = \theta^o + n_t$$

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	−0.66	1.12	1.32	−0.14	2.25	−0.21	0.96	1.28	1.17

an example - deterministic

$$y_t = \theta^o + n_t$$

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	−0.66	1.12	1.32	−0.14	2.25	−[illegible]	0.96	1.28	1.17



<u>Assumption:</u> $|n_t| \leq 2$

an example - deterministic

$$y_t = \theta^o + n_t$$

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	−0.66	1.12	1.32	−0.14	2.25	−0.21	0.96	1.28	1.17

Assumption: $|n_t| \leq 2$

an example - deterministic

$$y_t = \theta^o + n_t$$

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	−0.66	1.12	1.32	−0.14	2.25	−0.21	0.96	1.28	1.17

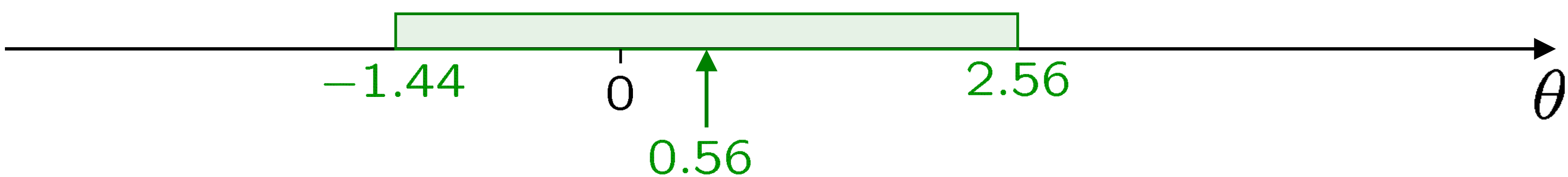
<u>Assumption:</u> $|n_t| \leq 2$

an example - deterministic

$$y_t = \theta^o + n_t$$

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	−0.66	1.12	1.32	−0.14	2.25	−0.21	0.96	1.28	1.17

<u>Assumption:</u> $|n_t| \le 2$

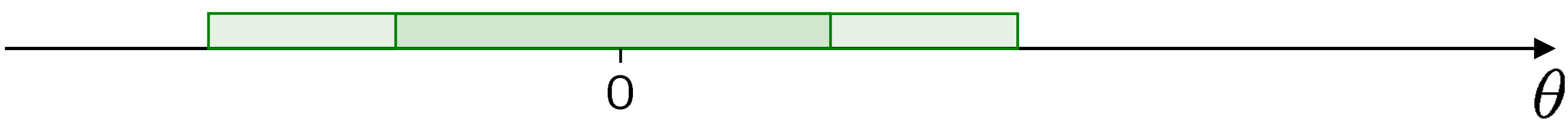


an example - deterministic

$$y_t = \theta^o + n_t$$

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	−0.66	1.12	1.32	−0.14	2.25	−0.21	0.96	1.28	1.17

<u>Assumption:</u> $|n_t| \leq 2$

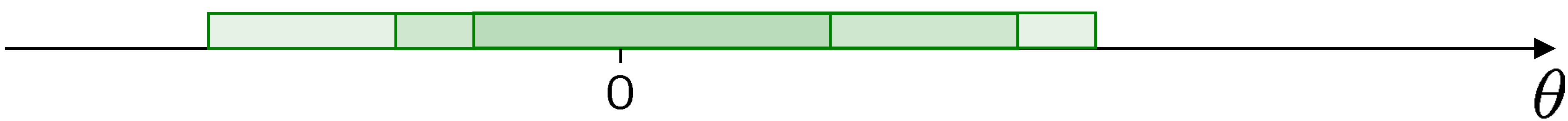


an example - deterministic

$$y_t = \theta^o + n_t$$

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	−0.66	1.12	1.32	−0.14	2.25	−0.21	0.96	1.28	1.17

Assumption: $|n_t| \leq 2$



an example - deterministic

$$y_t = \theta^o + n_t$$

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	−0.66	1.12	1.32	−0.14	2.25	−0.21	0.96	1.28	1.17

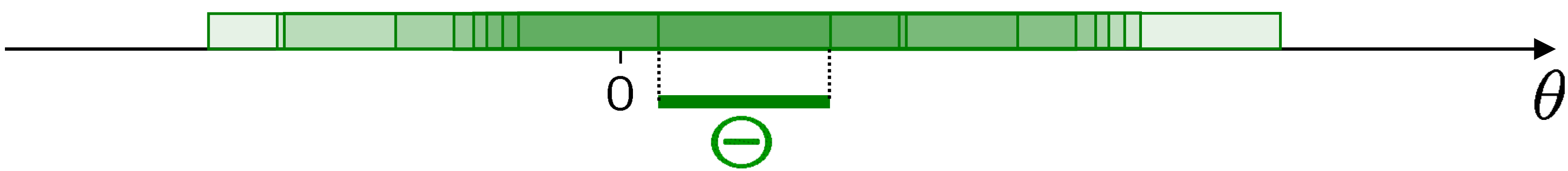
<u>Assumption:</u> $|n_t| \leq 2$

an example - deterministic

$$y_t = \theta^o + n_t$$

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	−0.66	1.12	1.32	−0.14	2.25	−0.21	0.96	1.28	1.17

<u>Assumption:</u> $|n_t| \leq 2$

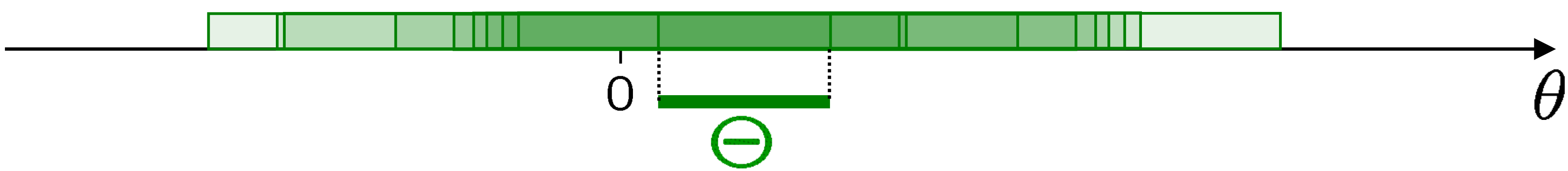


an example - deterministic

$$y_t = \theta^o + n_t$$

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	-0.66	1.12	1.32	-0.14	2.25	-0.21	0.96	1.28	1.17

Assumption: $|n_t| \leq 2$



Claim: $\theta^o \in \Theta$ (deterministic)

an example - probabilistic

an example - probabilistic

are we sure that $|n_t| \leq 2$?

an example - probabilistic

$$y_t = \theta^o + n_t$$

<u>Assumption:</u> $n_t \sim \mathsf{Gaussian}(0, 1)$

an example - probabilistic

$$y_t = \theta^o + n_t$$

Assumption: $n_t \sim \text{Gaussian}(0, 1)$

LS estimate: $\widehat{\theta}_{LS} = \frac{1}{10} \sum_{t=1}^{10} y_t = 0.76$

an example - probabilistic

$$y_t = \theta^o + n_t$$

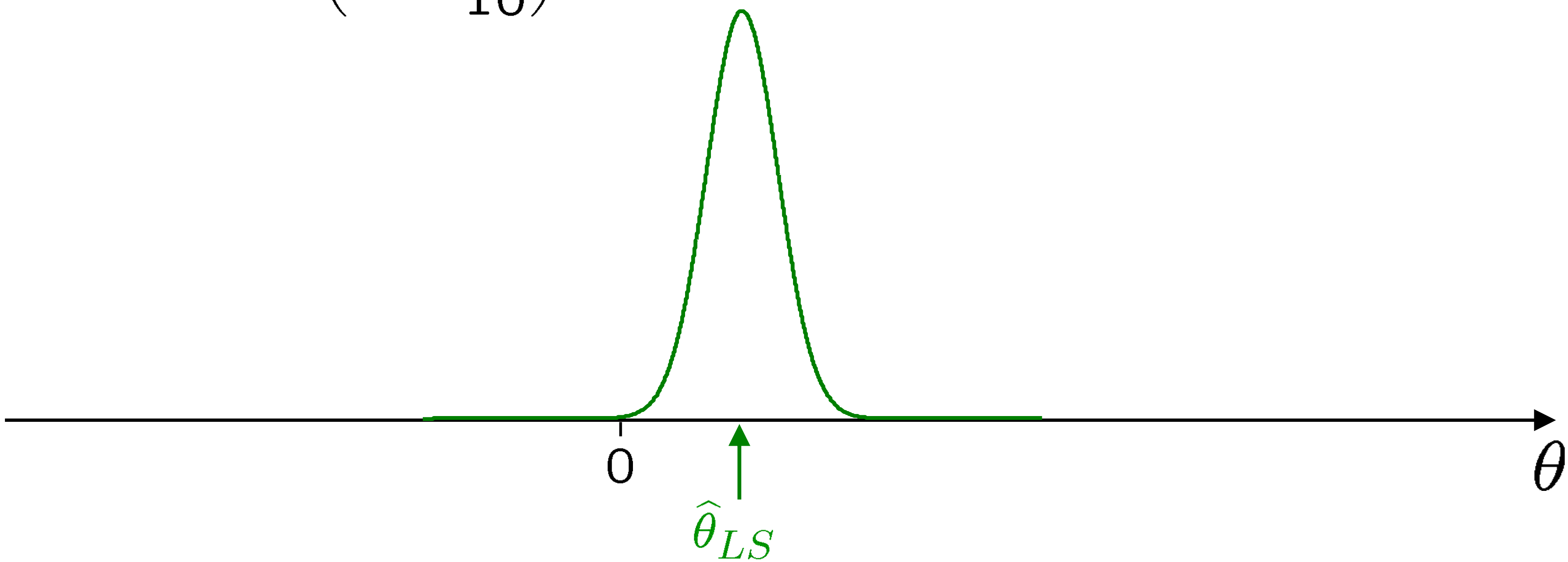
Assumption: $n_t \sim \text{Gaussian}(0, 1)$

LS estimate: $\widehat{\theta}_{LS} = \frac{1}{10} \sum_{t=1}^{10} y_t = 0.76$

$$\theta^o - \widehat{\theta}_{LS} \sim \text{Gaussian}\left(0, \frac{1}{10}\right)$$

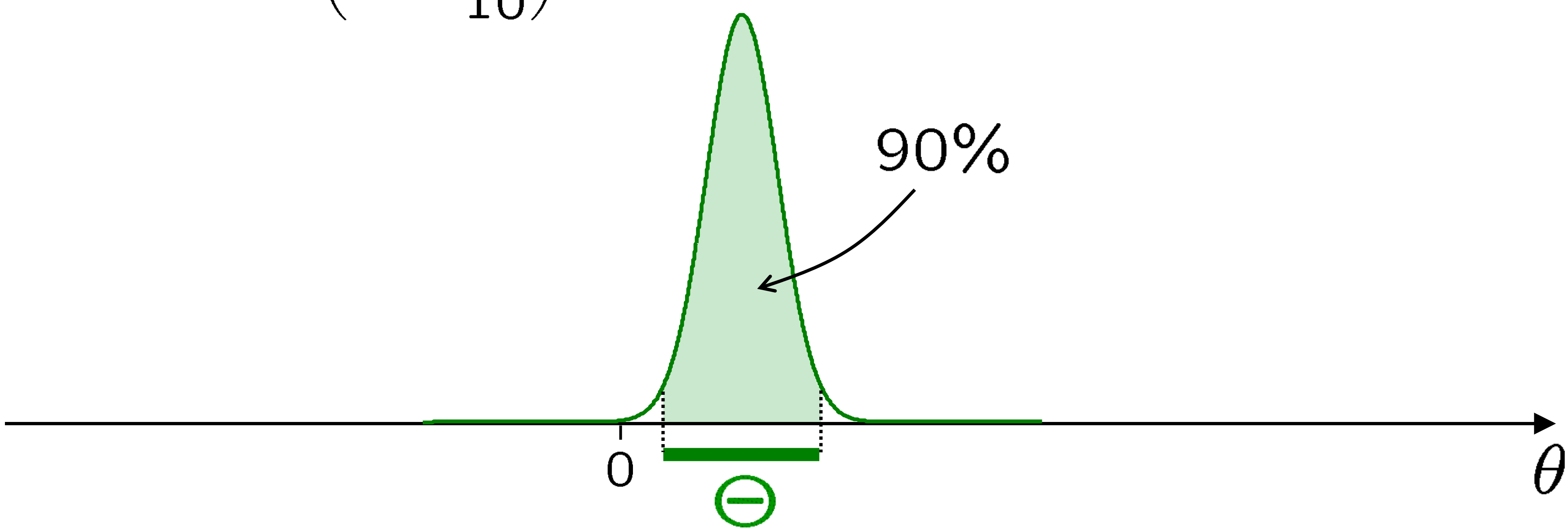
an example - probabilistic

Gaussian $\left(\widehat{\theta}_{LS}, \frac{1}{10}\right)$

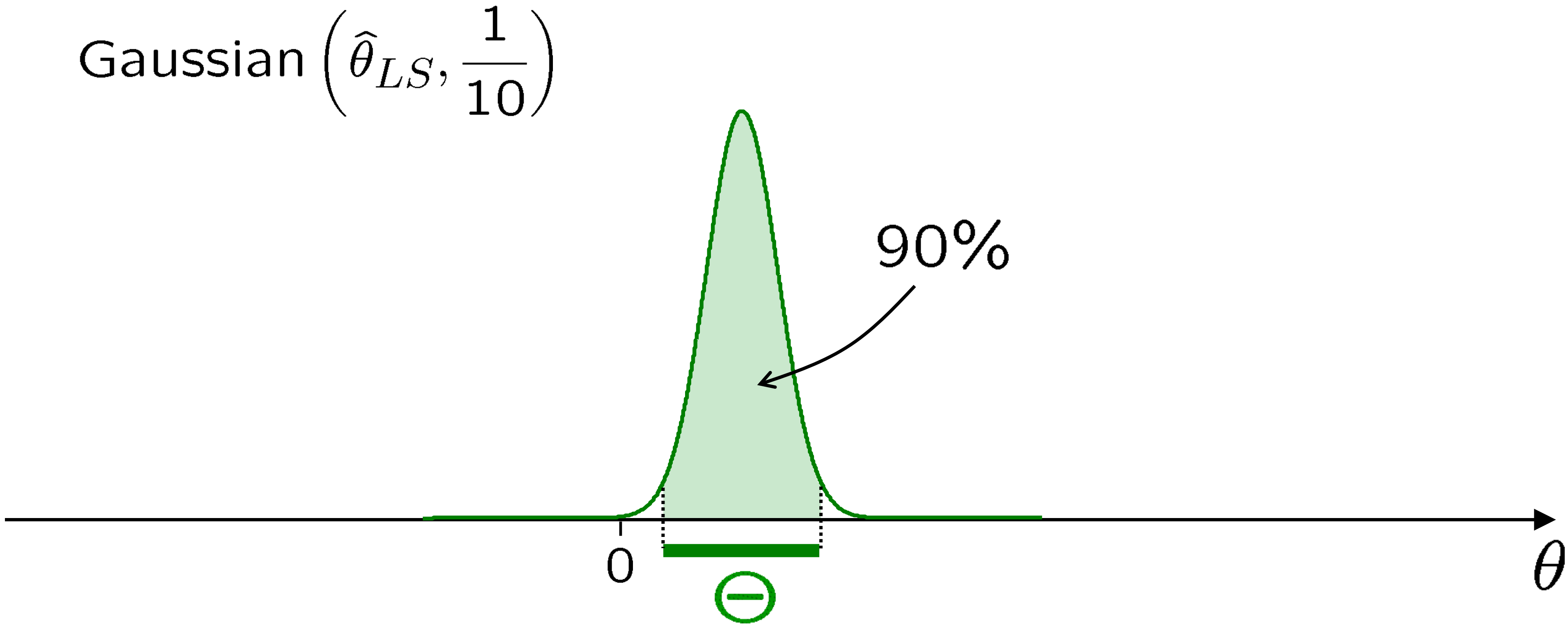


an example - probabilistic

Gaussian $\left(\widehat{\theta}_{LS}, \frac{1}{10}\right)$



an example - probabilistic



Claim: $\theta^o \in \Theta$ with probability 90%

are we sure that

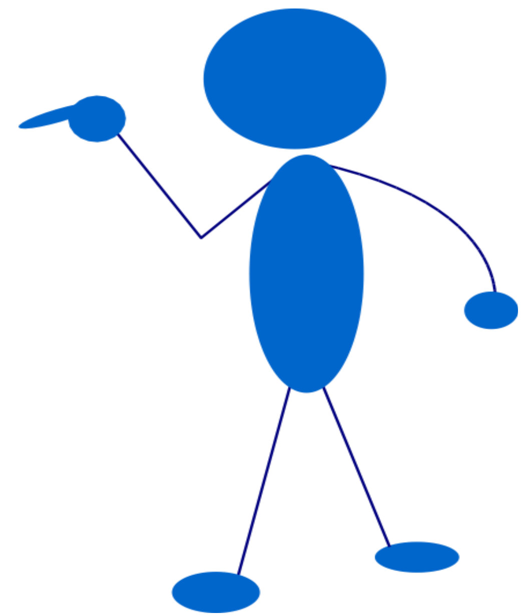
$n_t \sim \text{Gaussian}(0, 1)$?

a paradigm shift

an example - distribution-free (almost)

$$y_t = \theta^o + n_t$$

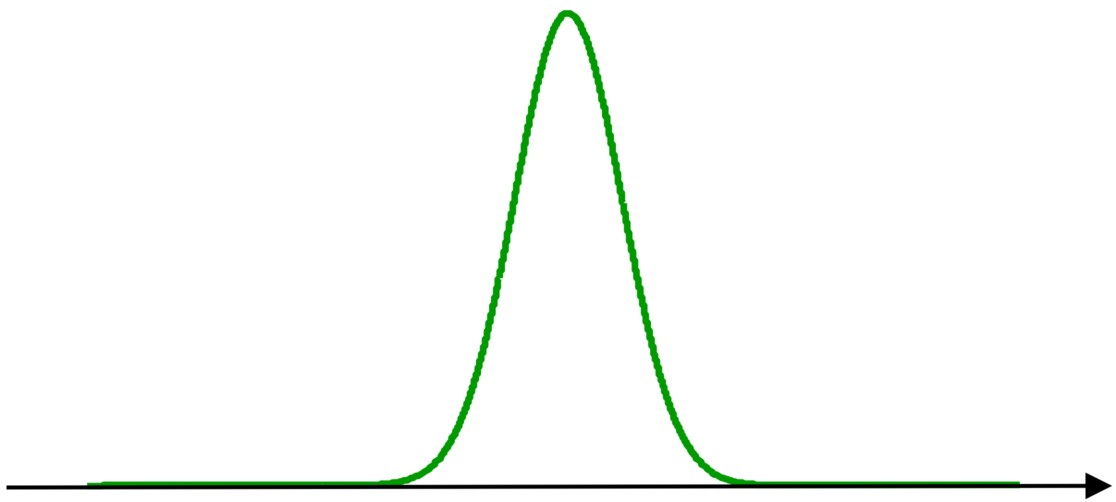
<u>Assumption:</u> n_t has an unknown distribution symmetric around zero



an example - distribution-free (almost)

$$y_t = \theta^o + n_t$$

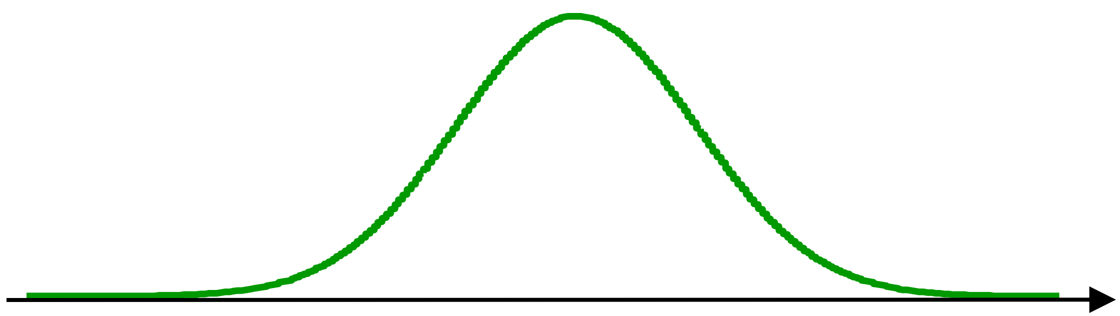
<u>Assumption:</u> n_t has an unknown distribution symmetric around zero



an example - distribution-free (almost)

$$y_t = \theta^o + n_t$$

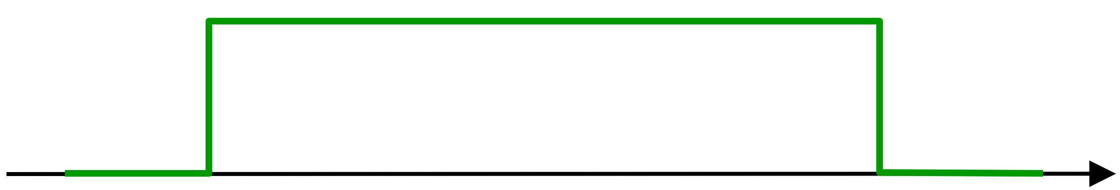
Assumption: n_t has an unknown distribution symmetric around zero



an example - distribution-free (almost)

$$y_t = \theta^o + n_t$$

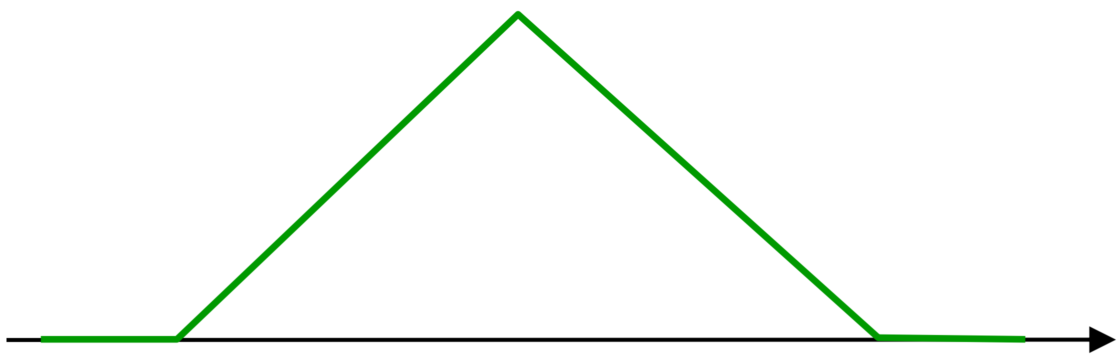
Assumption: n_t has an unknown distribution symmetric around zero



an example - distribution-free (almost)

$$y_t = \theta^o + n_t$$

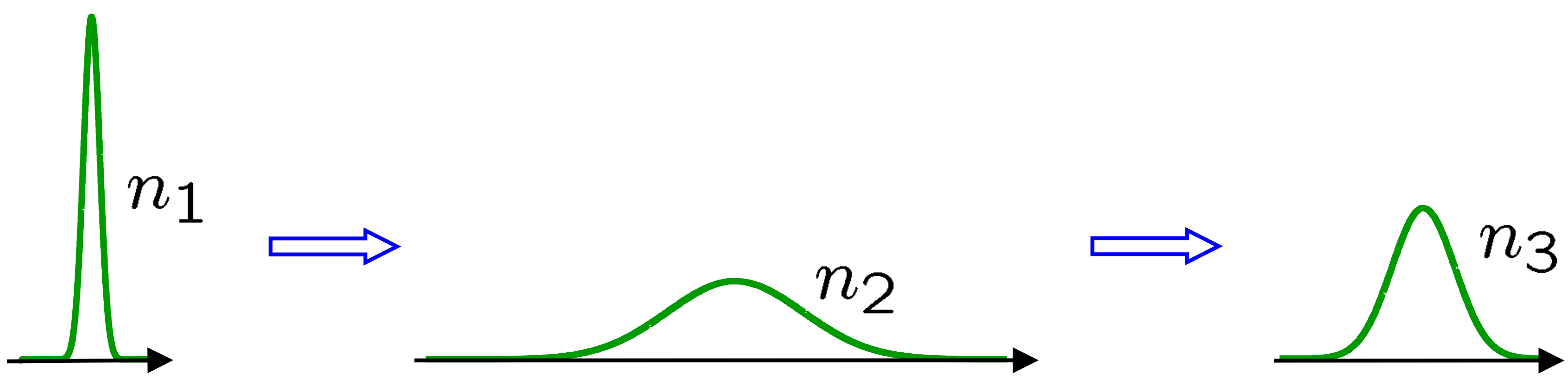
<u>Assumption:</u> n_t has an unknown distribution symmetric around zero



an example - distribution-free (almost)

$$y_t = \theta^o + n_t$$

<u>Assumption:</u> n_t has an unknown distribution symmetric around zero



an example - distribution-free (almost)

$$y_t = \theta^o + n_t$$

Assumption: n_t has an unknown distribution symmetric around zero

an example - distribution-free (almost)

$$y_t = \theta^o + n_t$$

Assumption: n_t has an unknown distribution symmetric around zero

$$\left[\sum_{t=1}^{10} (y_t - \theta)\right]^2 = \text{TEST PARABOLA}$$

an example - distribution-free (almost)

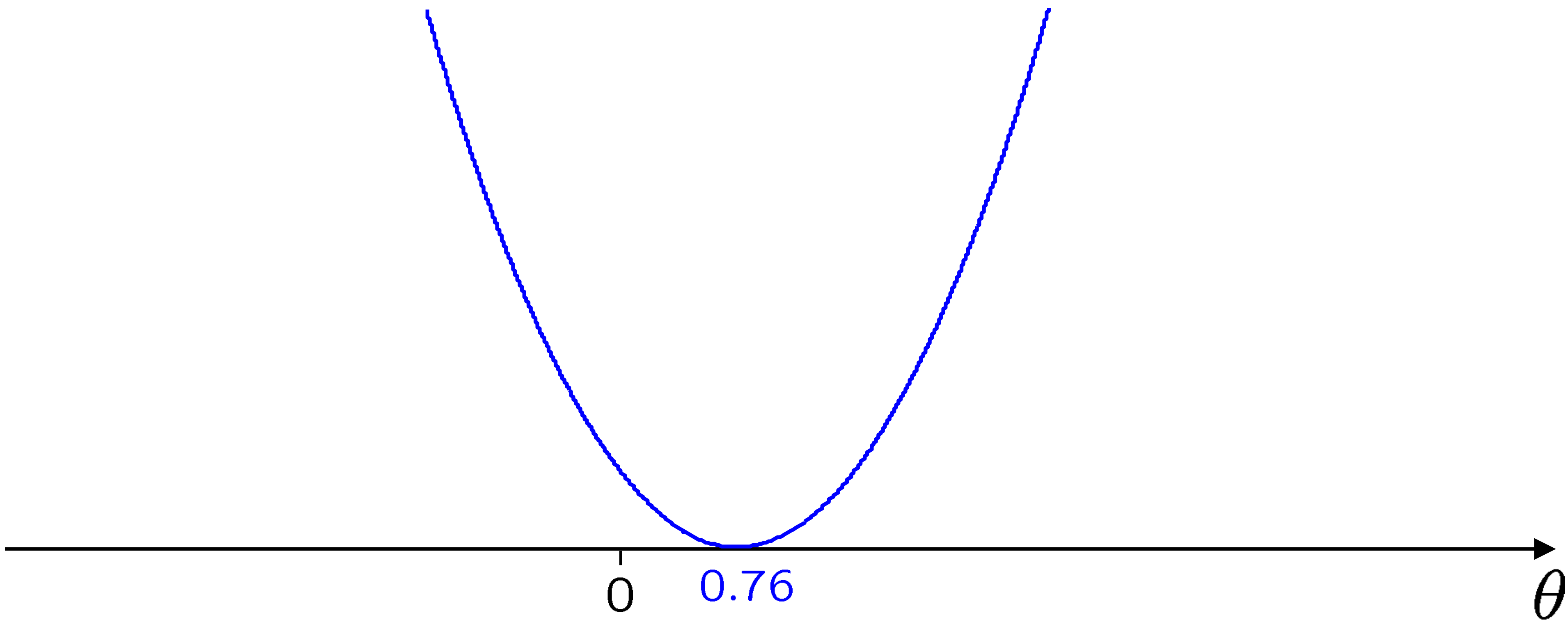
$$y_t = \theta^o + n_t$$

Assumption: n_t has an unknown distribution symmetric around zero

$$\left[\sum_{t=1}^{10} (y_t - \theta)\right]^2 = \text{TEST PARABOLA}$$

$$\text{vertex} = \frac{1}{10}\sum_{t=1}^{10} y_t = \hat{\theta}_{LS}$$

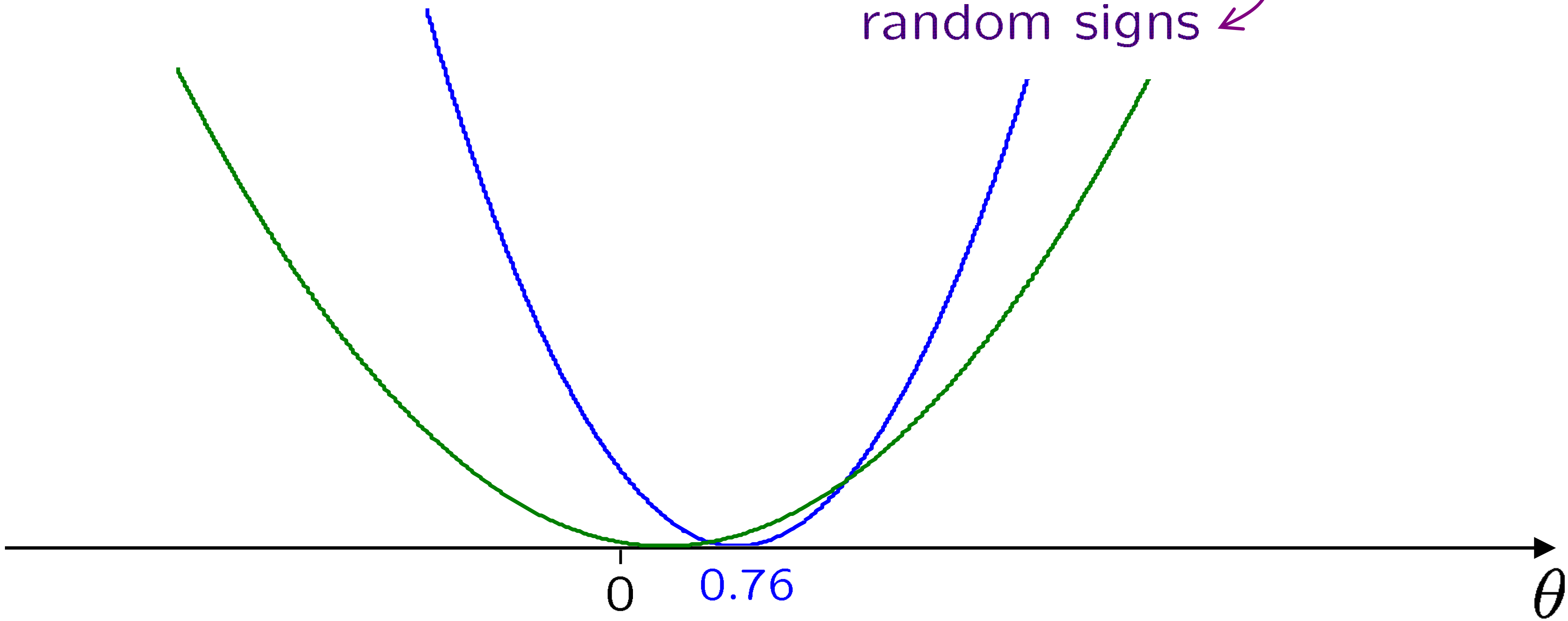
an example - distribution-free (almost)



an example - distribution-free (almost)

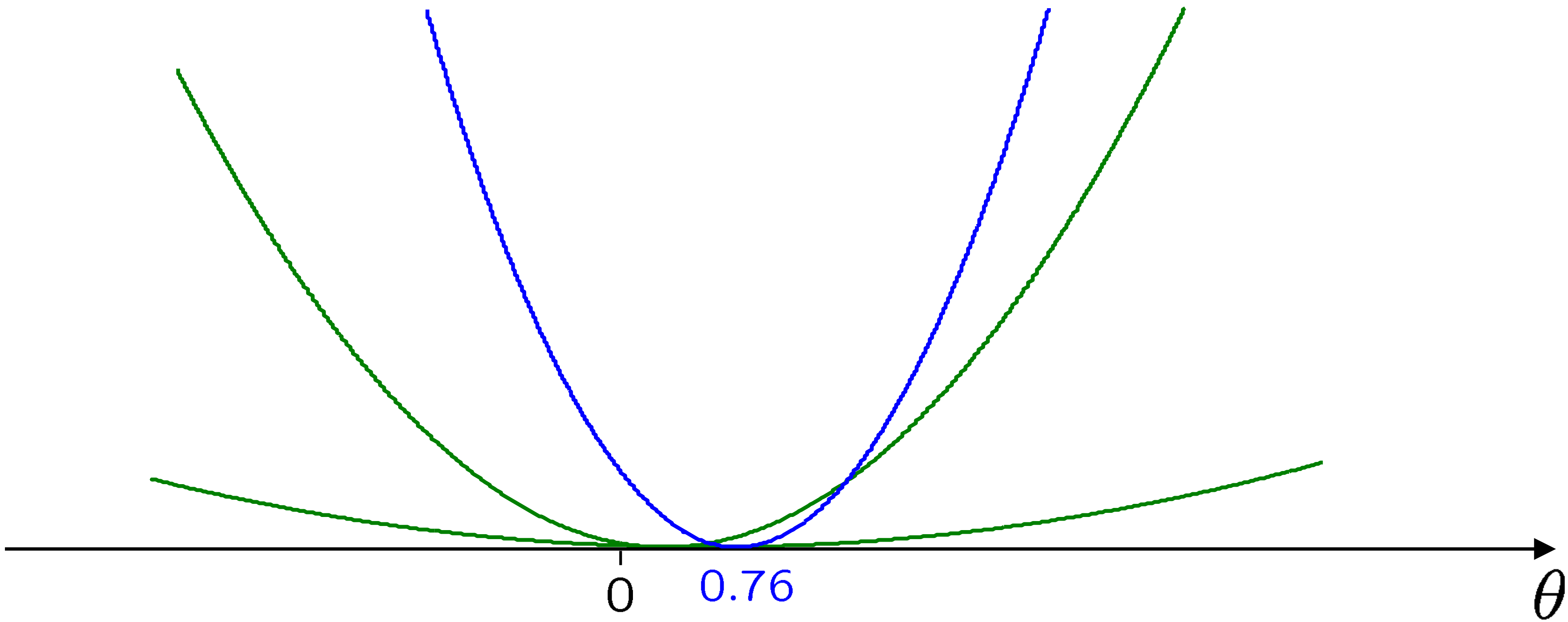
$$\text{PARABOLA} = \left[\sum_{t=1}^{10} \pm (y_t - \theta)\right]^2$$

random signs



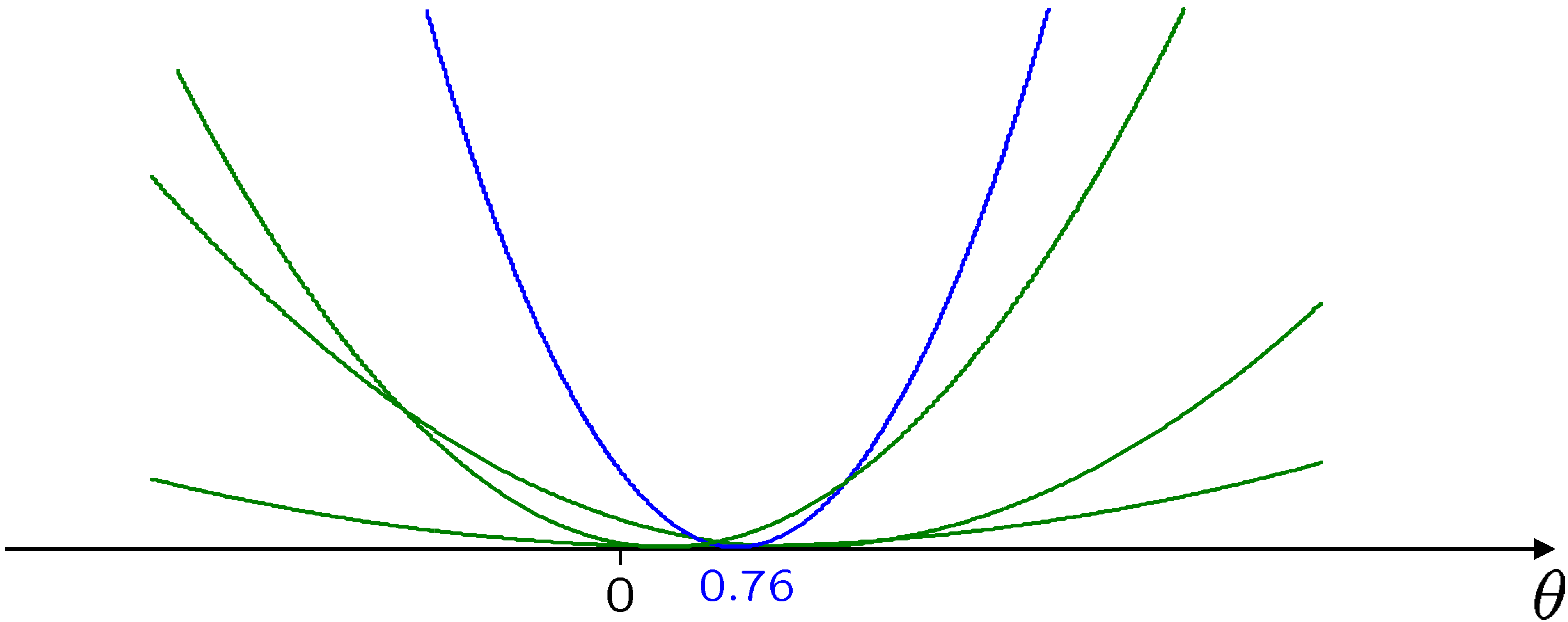
an example - distribution-free (almost)

$$2^0 \text{ PARABOLA} = \left[\sum_{t=1}^{10} \pm(y_t - \theta)\right]^2$$



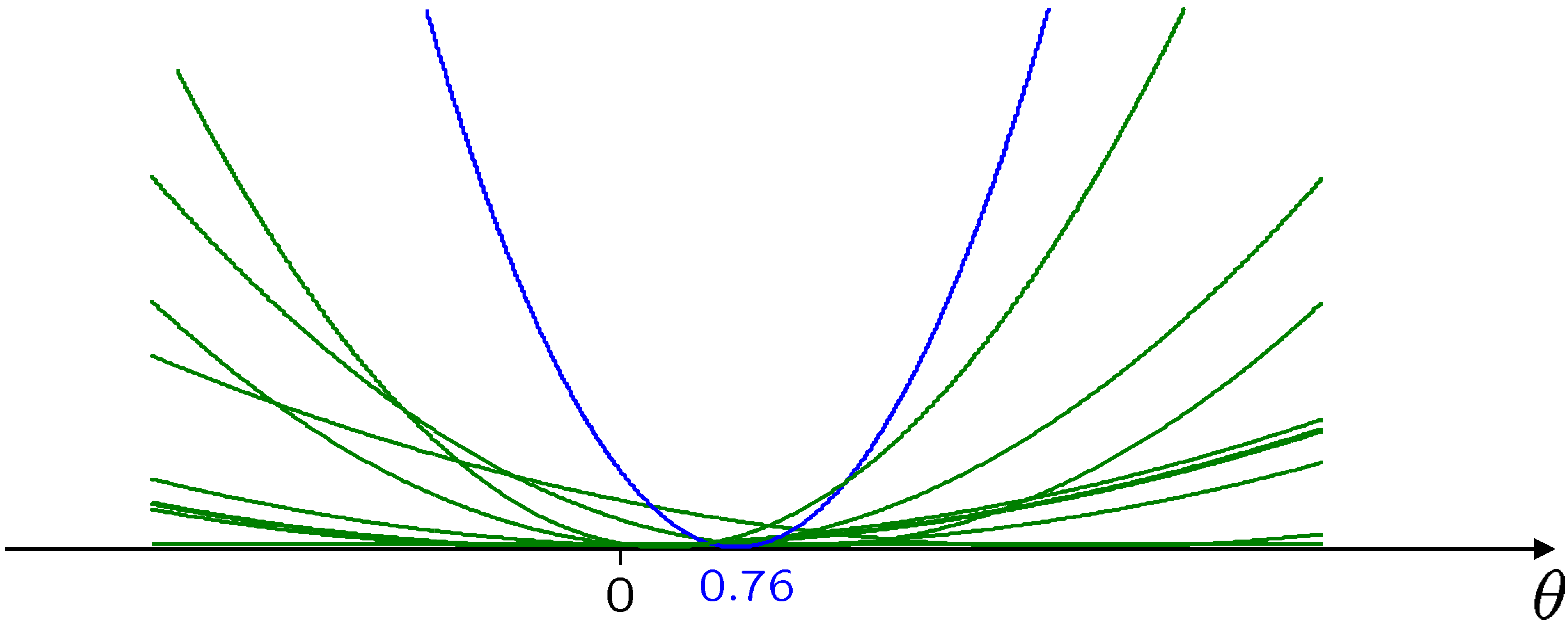
an example - distribution-free (almost)

$$3^0 \text{ PARABOLA} = \left[\sum_{t=1}^{10} \pm(y_t - \theta)\right]^2$$



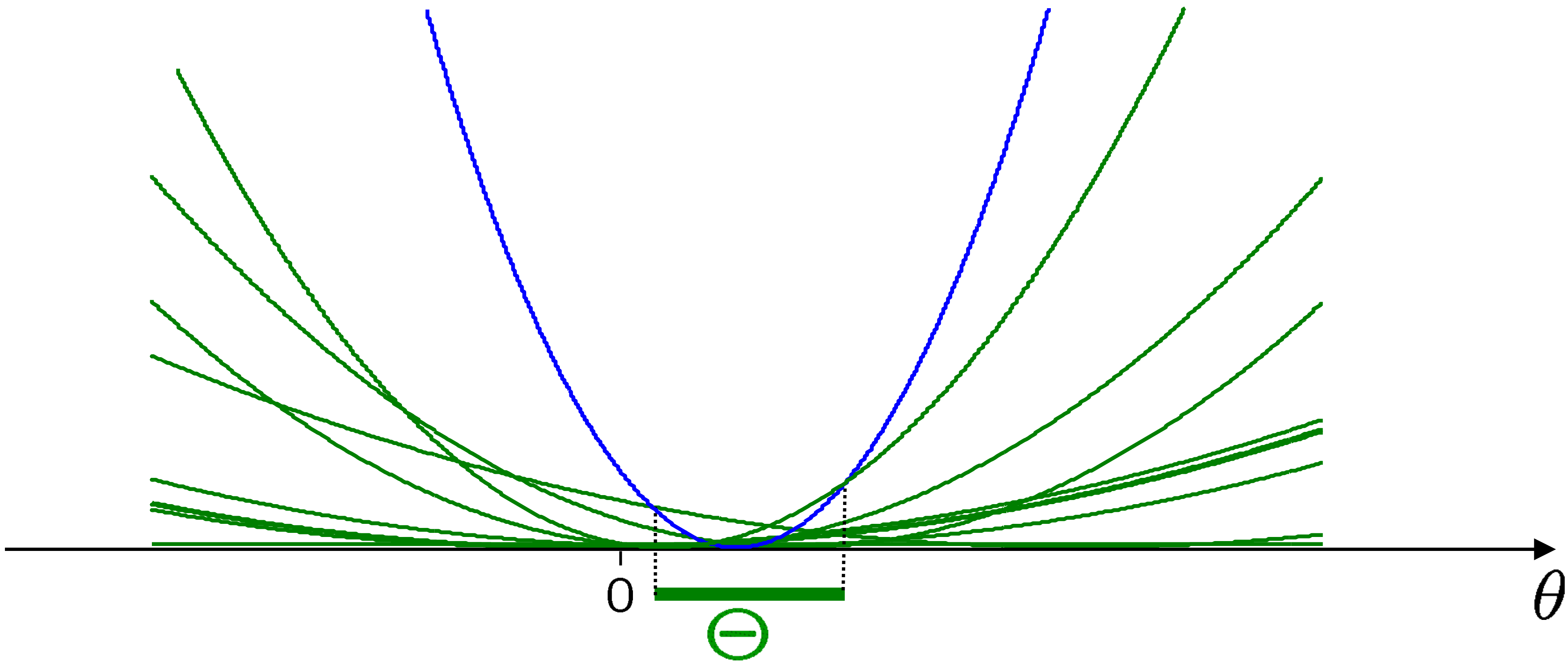
an example - distribution-free (almost)

$$9^0 \text{ PARABOLA} = \left[\sum_{t=1}^{10} \pm(y_t - \theta)\right]^2$$



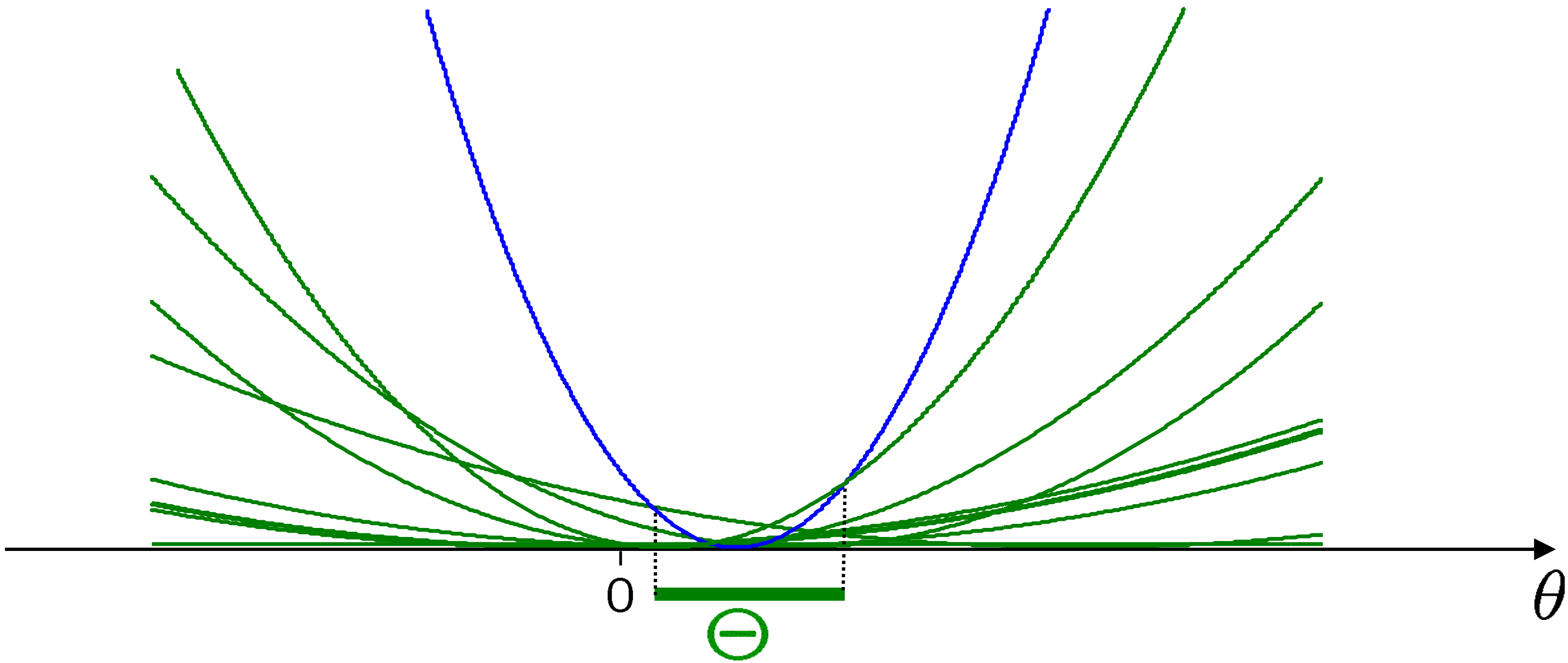
an example - distribution-free (almost)

$$9^0 \text{ PARABOLA} = \left[\sum_{t=1}^{10} \pm(y_t - \theta)\right]^2$$



an example - distribution-free (almost)

$$9^0 \text{ PARABOLA} = \left[\sum_{t=1}^{10} \pm(y_t - \theta)\right]^2$$



Claim: $\theta^o \in \Theta$ with probability 90%

Proof:

TEST PARABOLA($\theta = \theta^O$) $= \left[\sum^{10} \ldots)\right]^2 = \left[\sum_{t=1}^{10} (\theta^O + n_t - \theta^O)\right]^2$

$\ldots n_t \Big]^2$

1^0 PARABOLA($\theta = \ldots \left[\sum_{t=1} \pm n_t\right]^2$

all parabolas ... same probability to be at top

very simple!

there is a lot of information in the data that traditional methods do not exploit *there is a lot of information in the data that traditional methods do not exploit*

let the data speak!

SPS = Sign Perturbed Sums (with Csaji, Weyer)

SPS = Sign Perturbed Sums (with Csaji, Weyer)

$$y_t = \theta_1^o u_{t-1} + \theta_2^o u_{t-2} + n_t$$

$$\text{TEST PARABOLA} = \left\| \sum_{t=1}^{N} \begin{bmatrix} u_{t-1} \\ u_{t-2} \end{bmatrix} (y_t - \theta_1 u_{t-1} - \theta_2 u_{t-2}) \right\|^2$$

$$\text{vertex} = \widehat{\theta}_{LS}$$

$$\text{n-th PARABOLA} = \left\| \sum_{t=1}^{N} \pm \begin{bmatrix} u_{t-1} \\ u_{t-2} \end{bmatrix} (y_t - \theta_1 u_{t-1} - \theta_2 u_{t-2}) \right\|^2$$

SPS = Sign Perturbed Sums (with Csaji, Weyer)

- $A^o(z^{-1})y_t = B^o(z^{-1})u_t + n_t$

SPS = Sign Perturbed Sums (with Csaji, Weyer)

- $A^o(z^{-1})y_t = B^o(z^{-1})u_t + n_t$

- $y_t = G^o(z^{-1})u_t + H^o(z^{-1})n_t$

Example

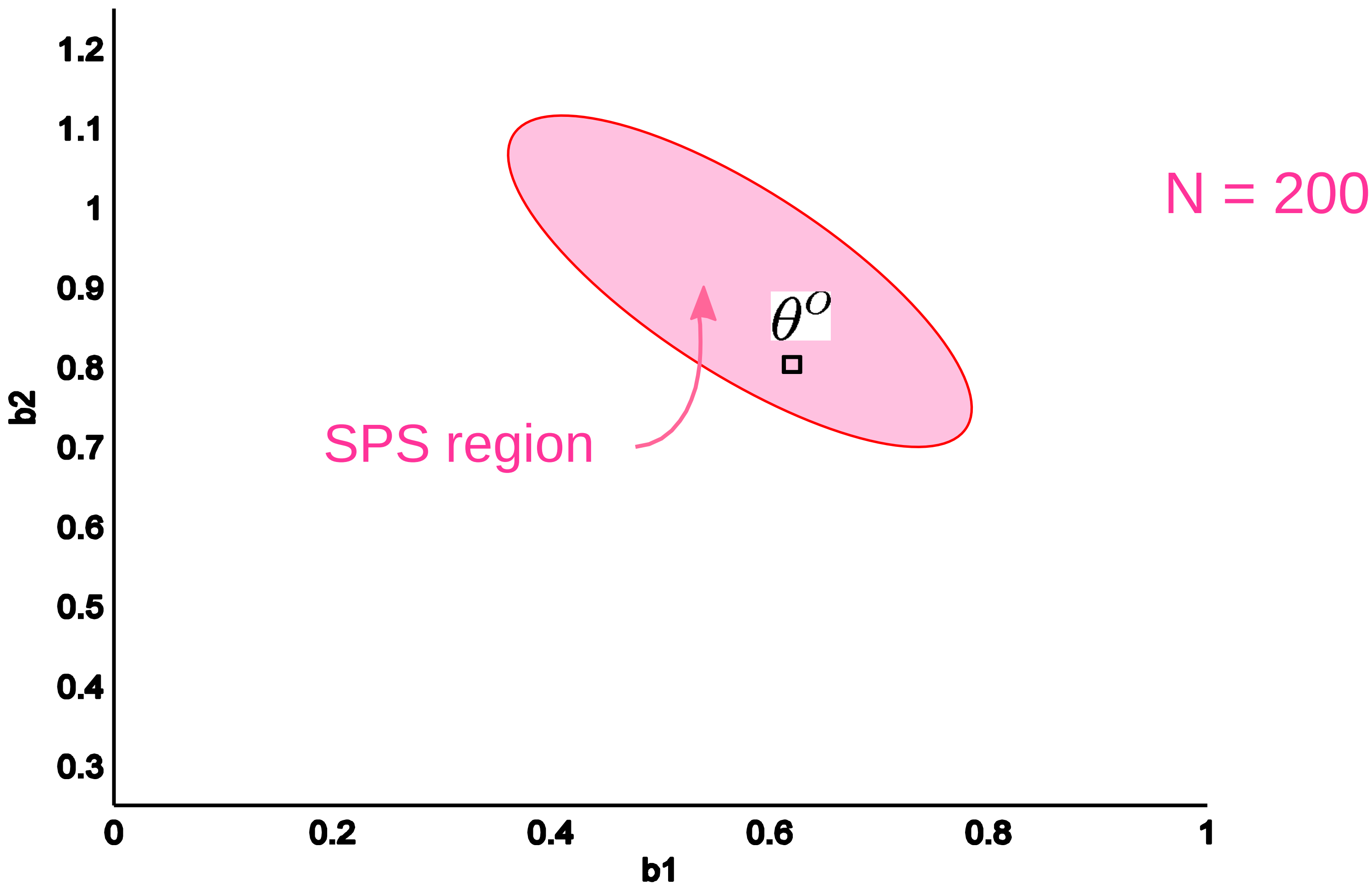
$$y_t = 0.6u_{t-1} + 0.8u_{t-2} + n_t$$

$$u_t = 0.8u_{t-1} + v_t$$

Example

$$y_t = 0.6u_{t-1} + 0.8u_{t-2} + n_t$$

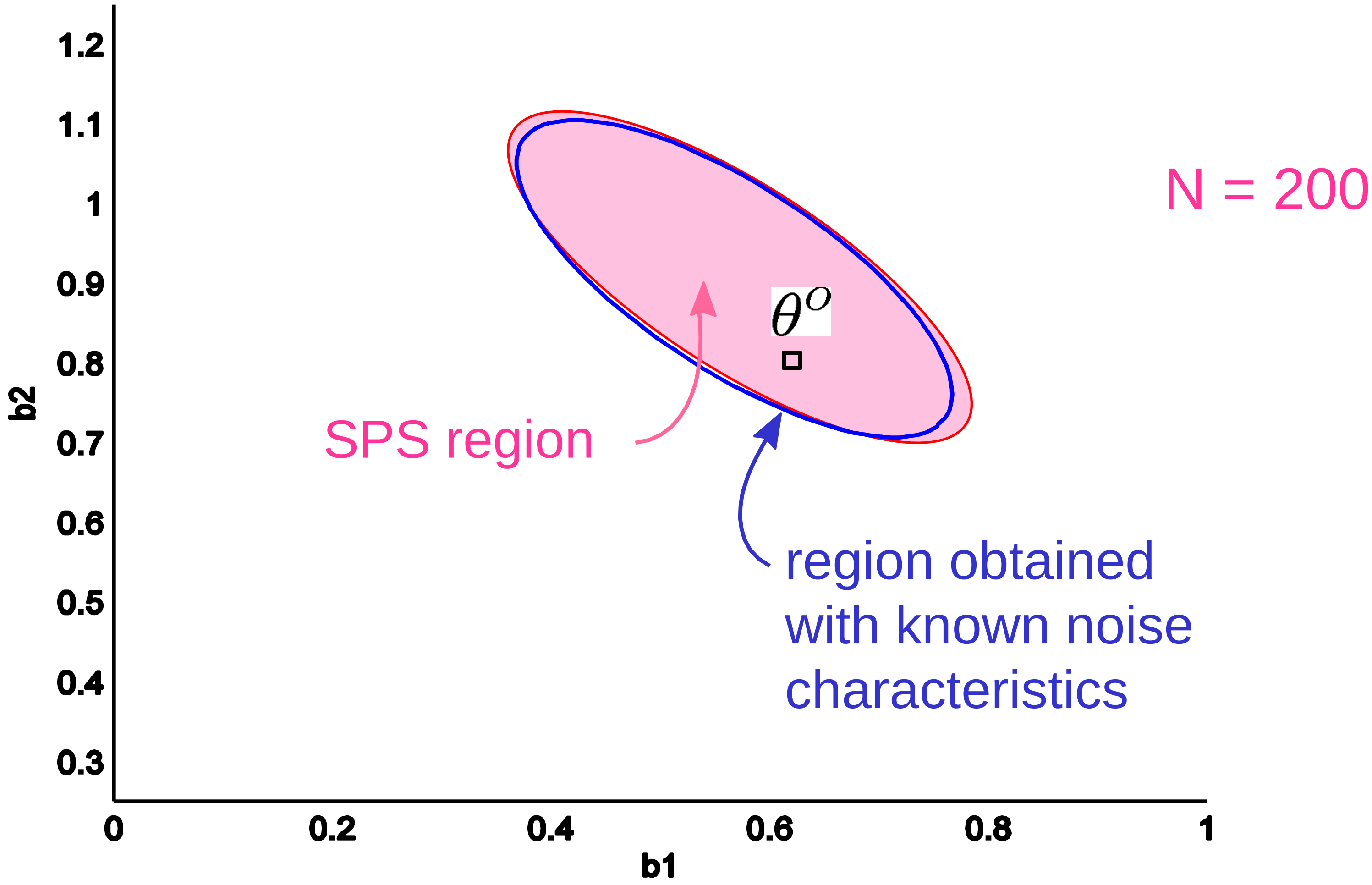
$$u_t = 0.8u_{t-1} + v_t$$



Example

$$y_t = 0.6u_{t-1} + 0.8u_{t-2} + n_t$$

$$u_t = 0.8u_{t-1} + v_t$$



. . . mid-talk summary

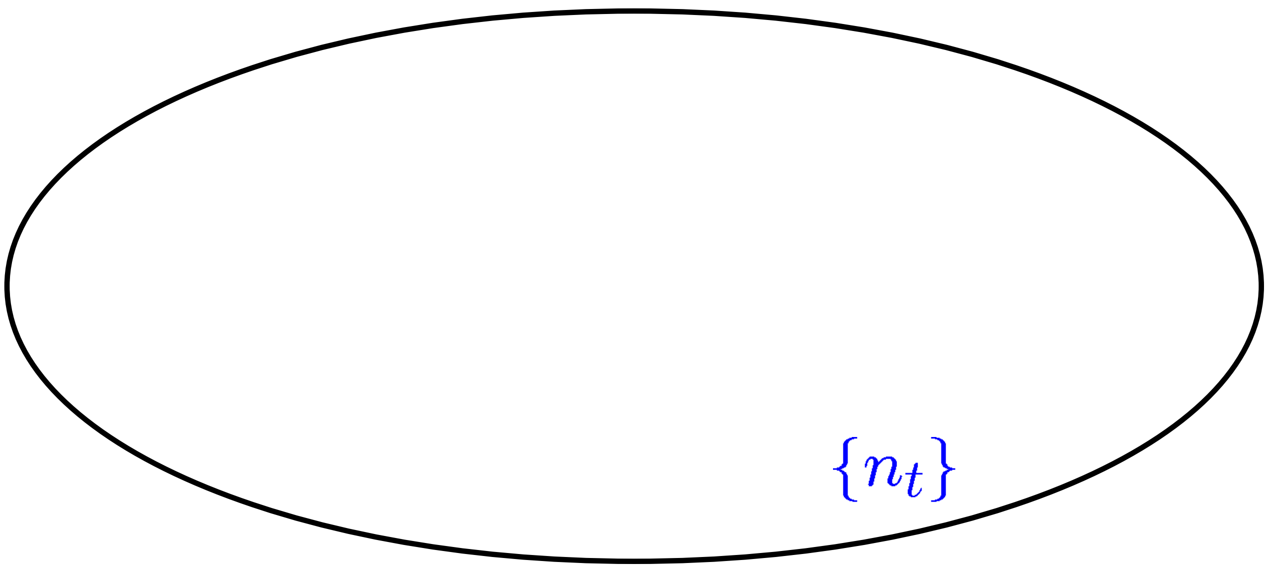
deterministic

Claim: $\theta^o \in \Theta$

deterministic

Claim: $\theta^o \in \Theta$

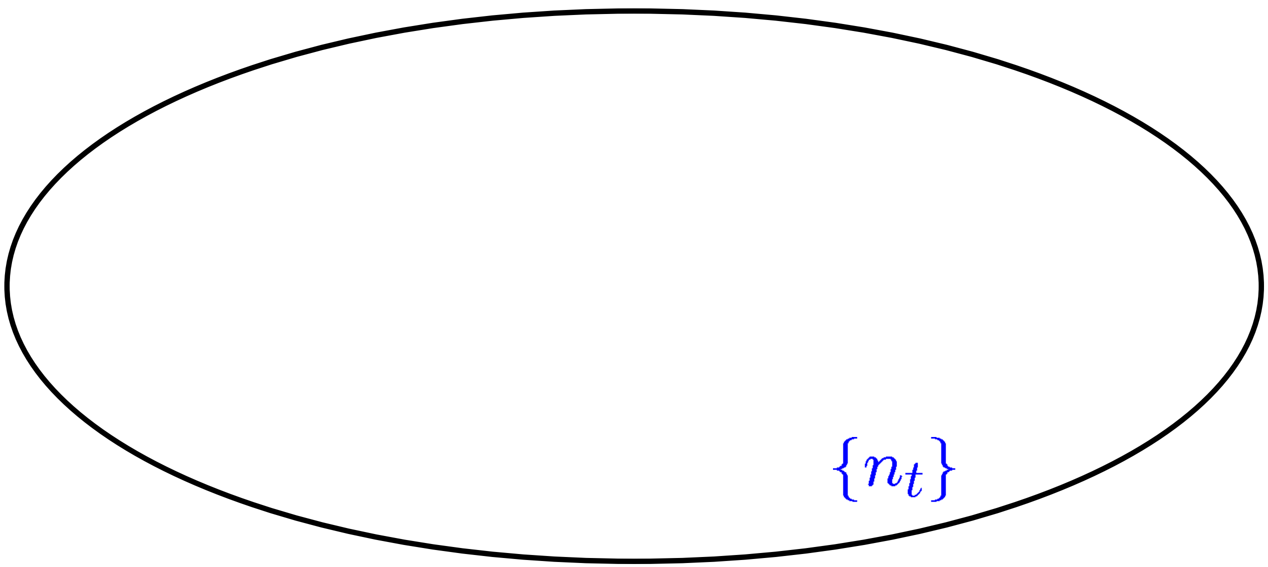
set-theoretic statement



deterministic

Claim: $\theta^o \in \Theta$

set-theoretic statement

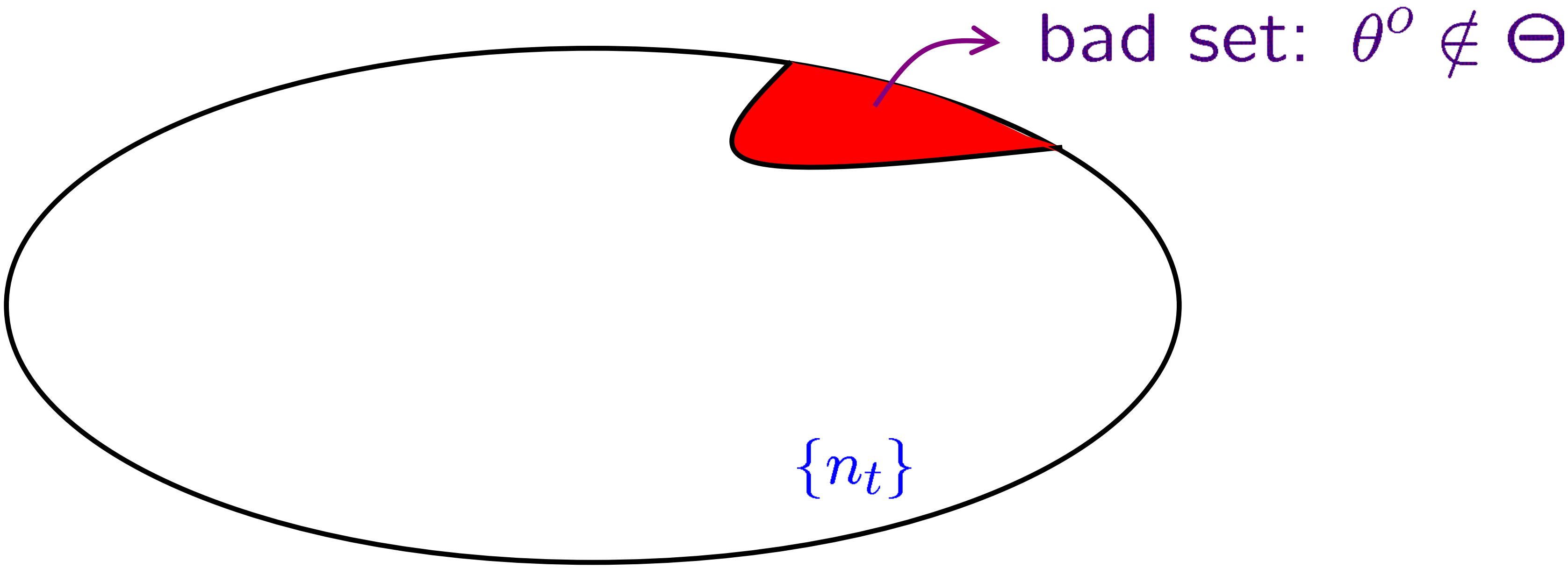


rigid deterministic assumptions

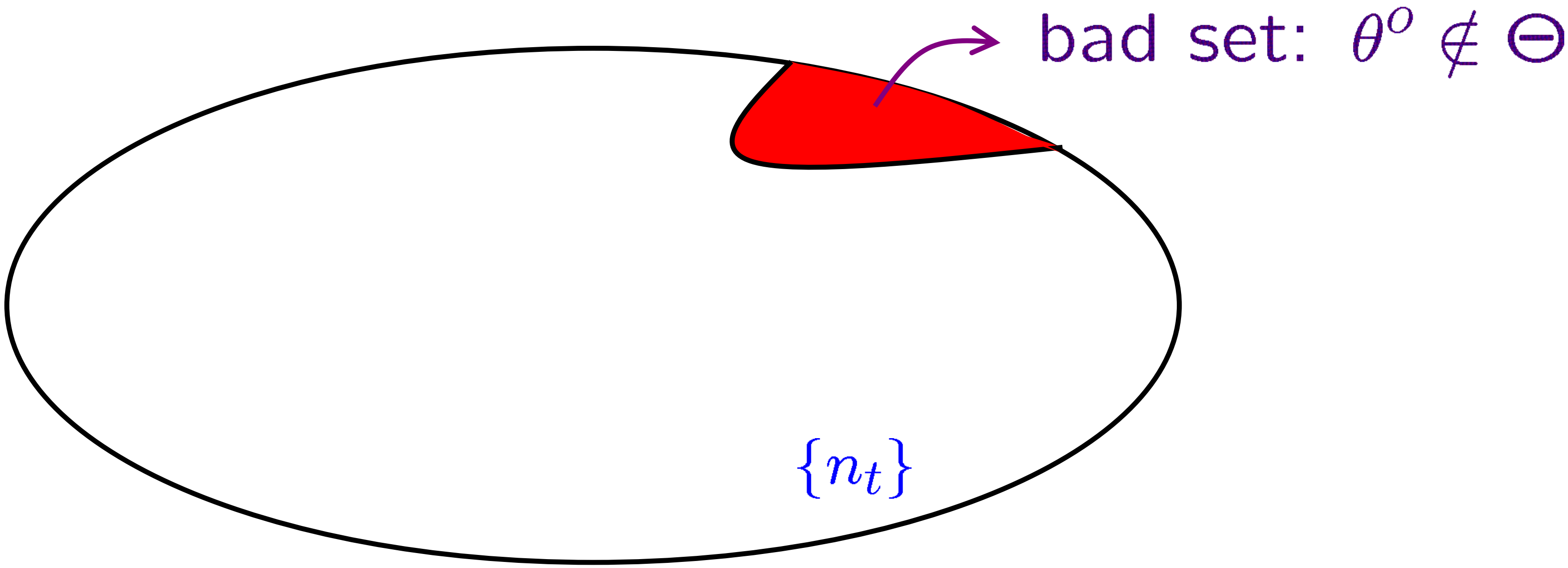
Mark Twain: “what gets us into trouble is not what we don't know. It's what we know for sure that just ain't so.”

remove the stringent deterministic assumptions

remove the stringent deterministic assumptions

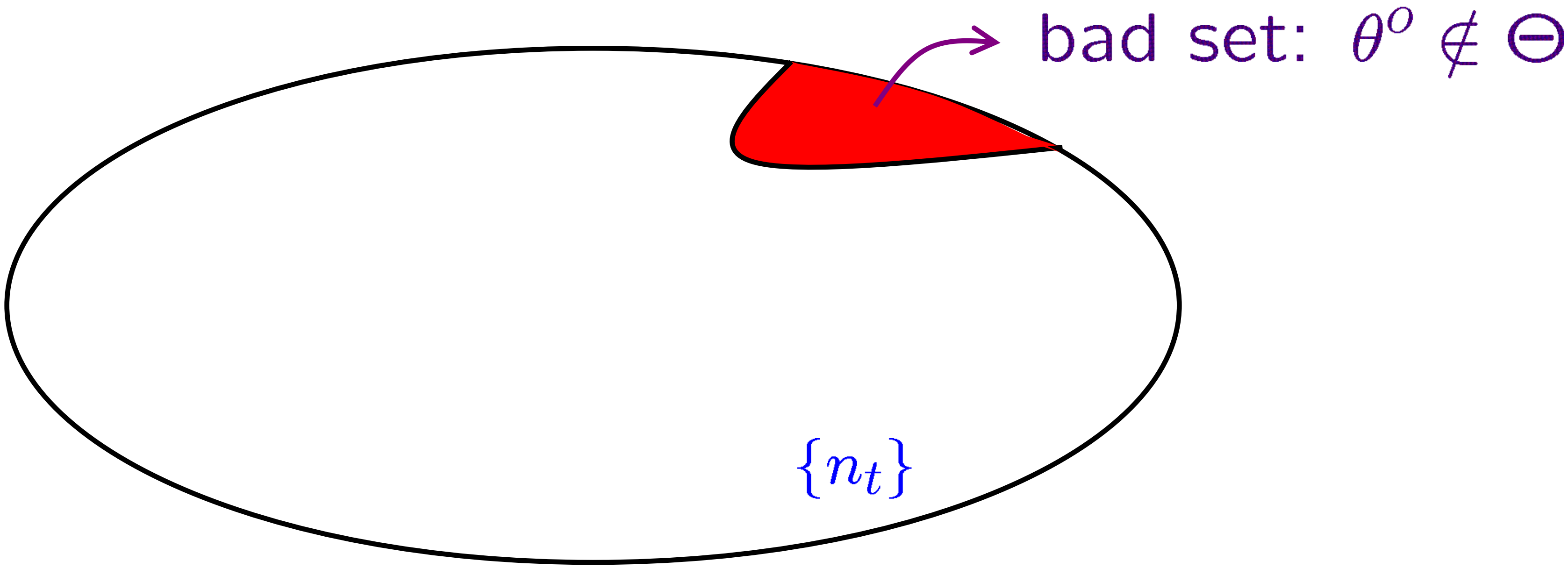


remove the stringent deterministic assumptions



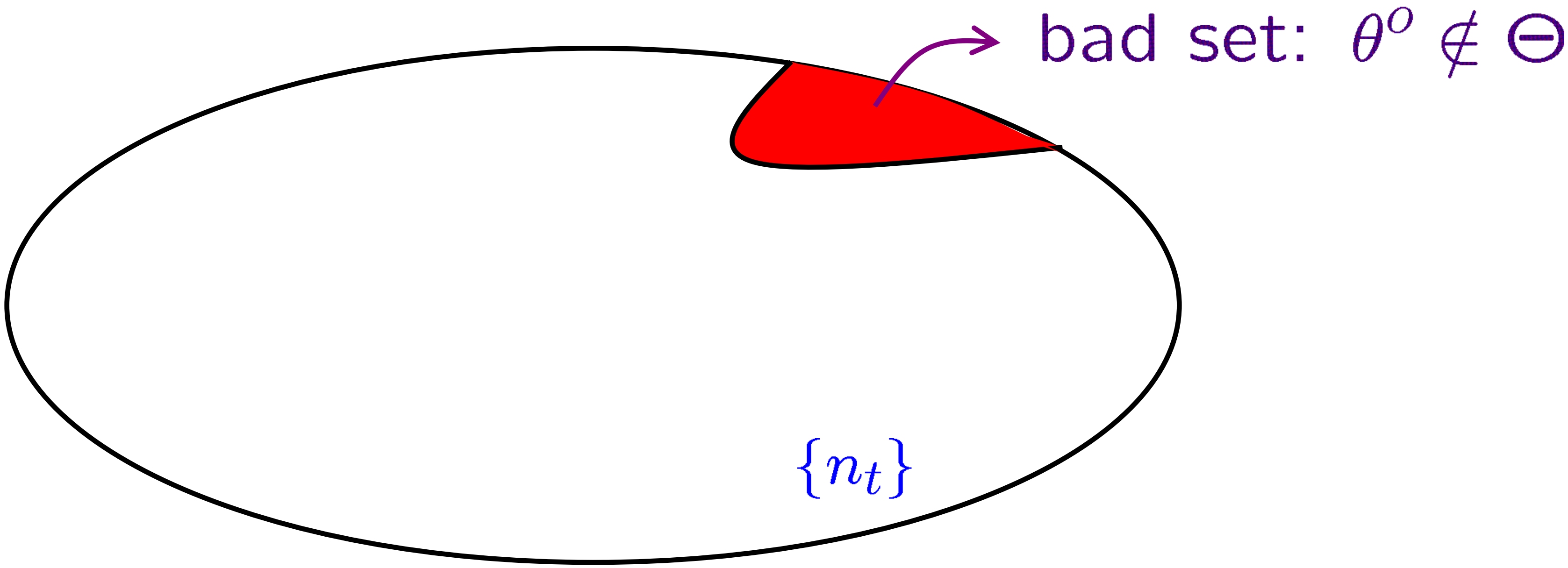
a probability is needed to tackle the challenge of being quantitative

remove the stringent deterministic assumptions



Claim: $\theta^o \in \Theta$ with probability 90%

remove the stringent deterministic assumptions



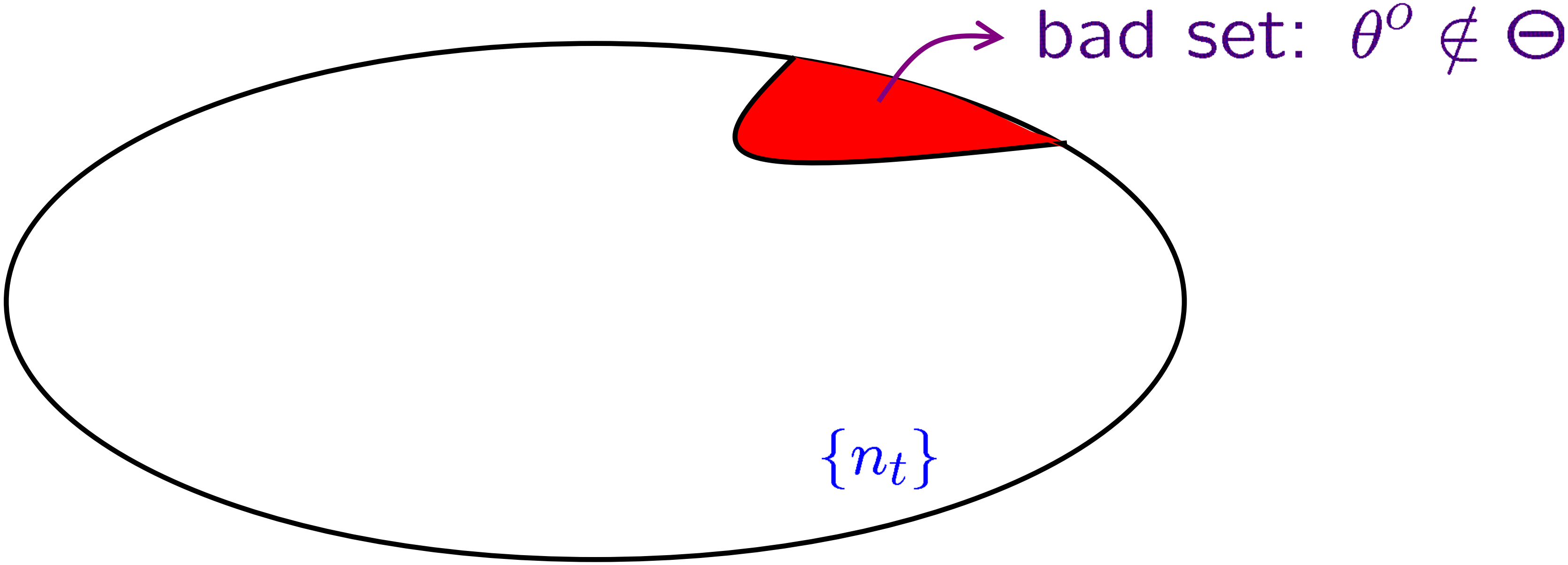
Claim: $\theta^o \in \Theta$ with probability 90%

strong probabilistic priors are used

Jan Willems: “where would the numerical values of this probability come from?”

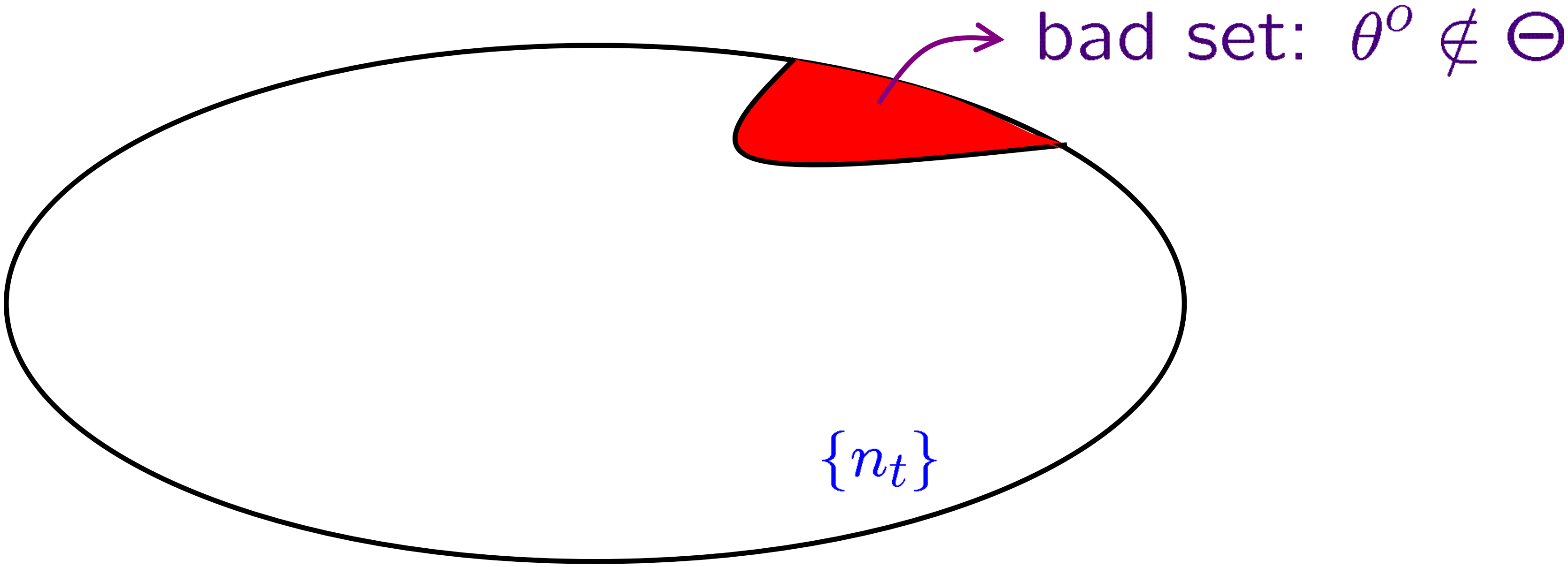
move towards a distribution-free approach

move towards a distribution-free approach



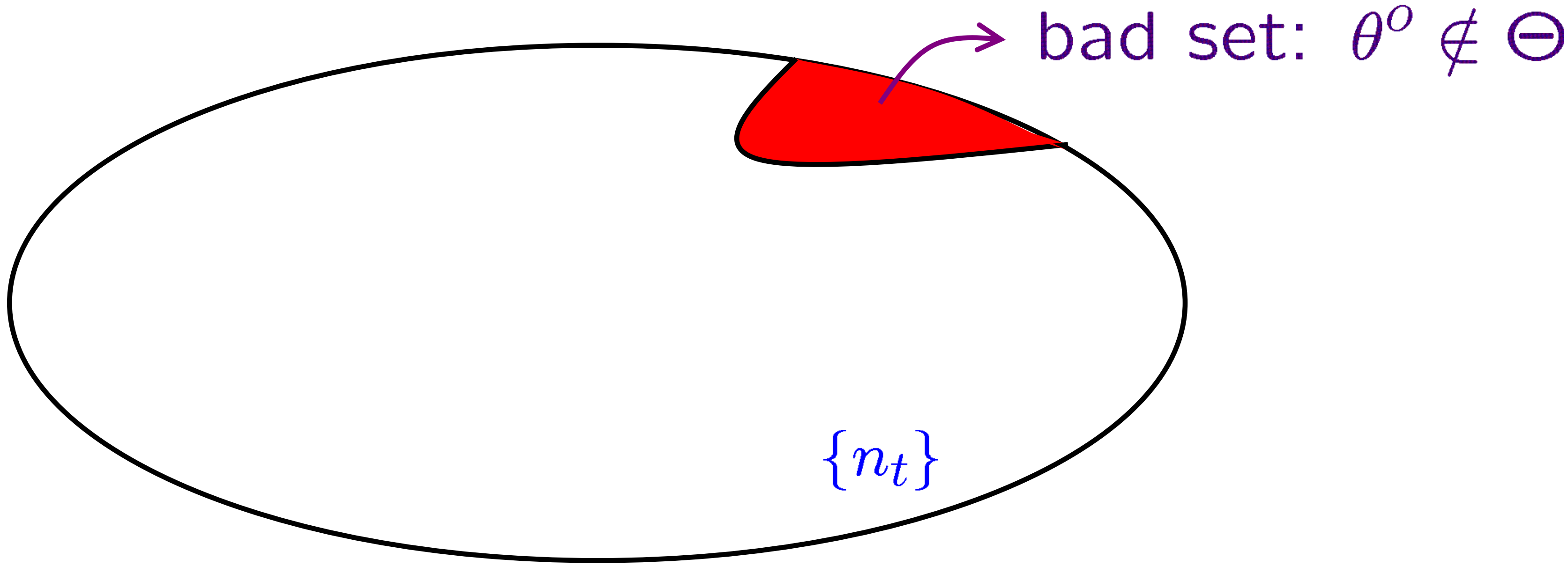
Claim: $\theta^o \in \Theta$ with probability 90%

move towards a distribution-free approach

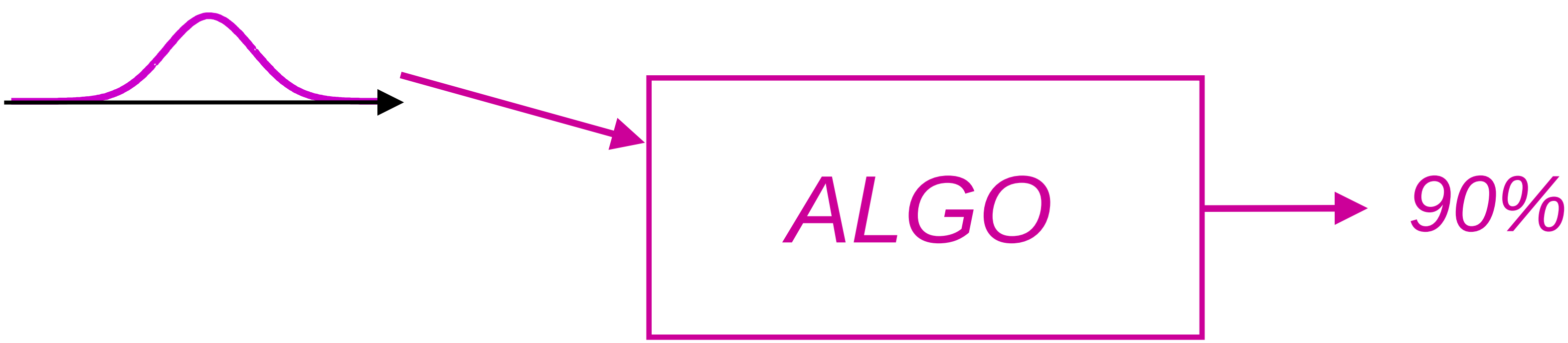


<u>Claim:</u> $\theta^o \in \Theta$ with probability 90%

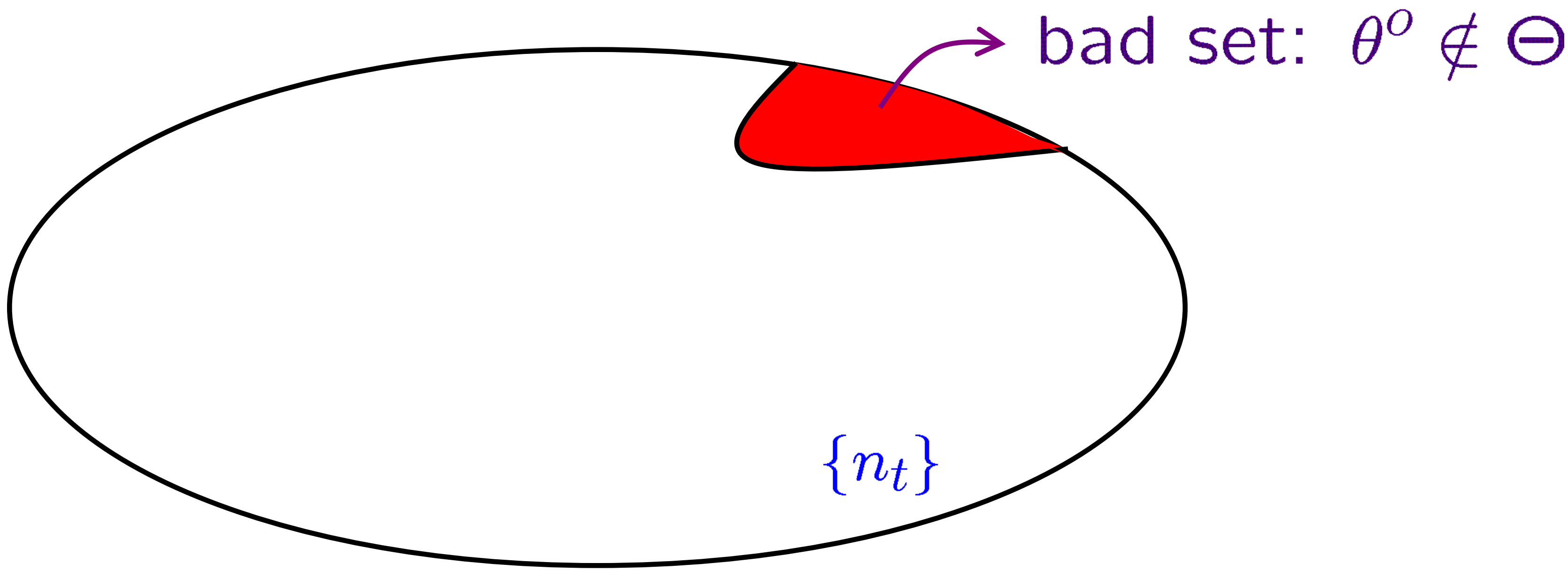
move towards a distribution-free approach



<u>Claim:</u> $\theta^o \in \Theta$ with probability 90%



move towards a distribution-free approach



Claim: $\theta^o \in \ominus$ with probability 90%

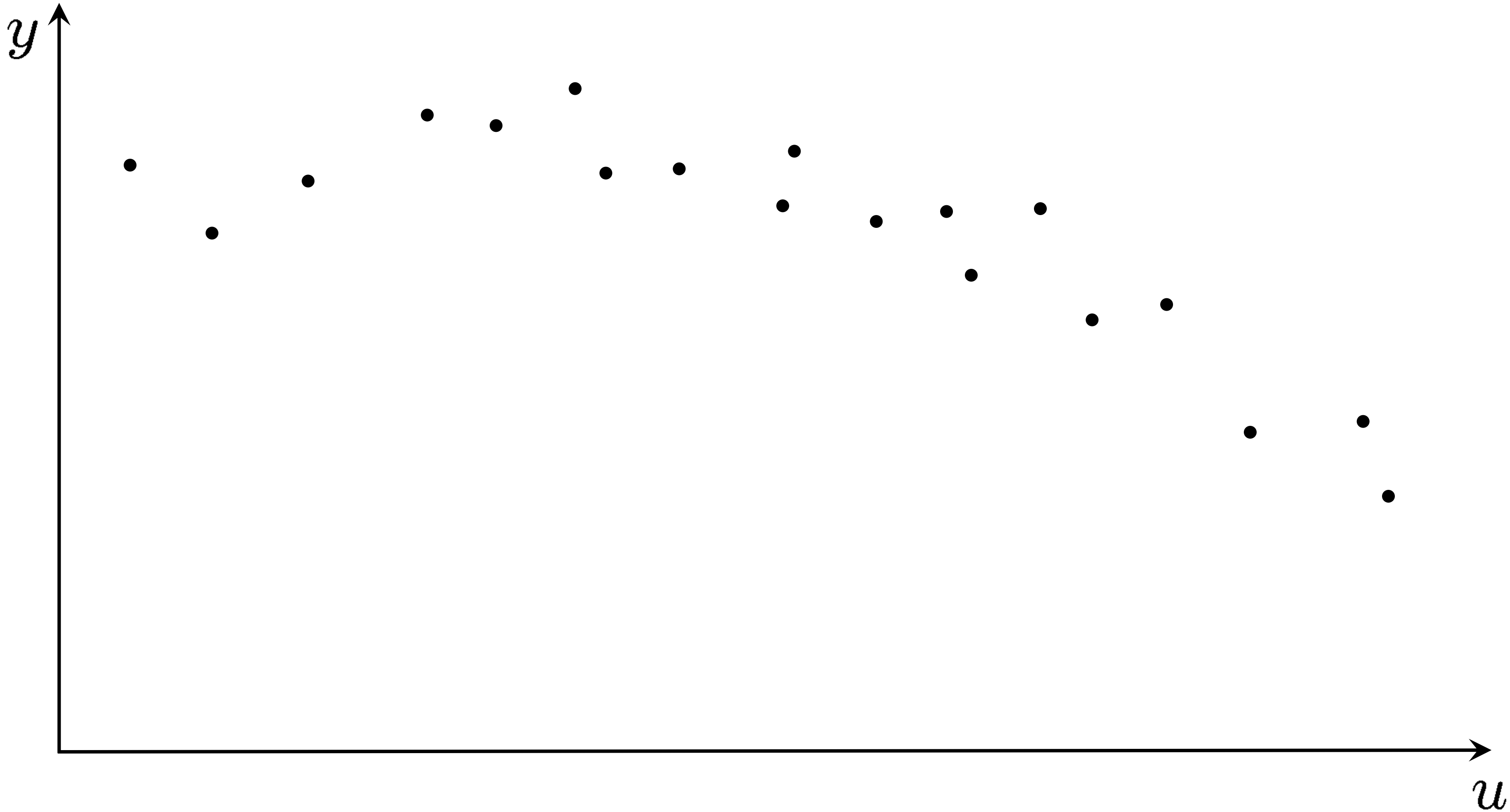
A journey through:

- parameter estimation
- prediction
- filtering

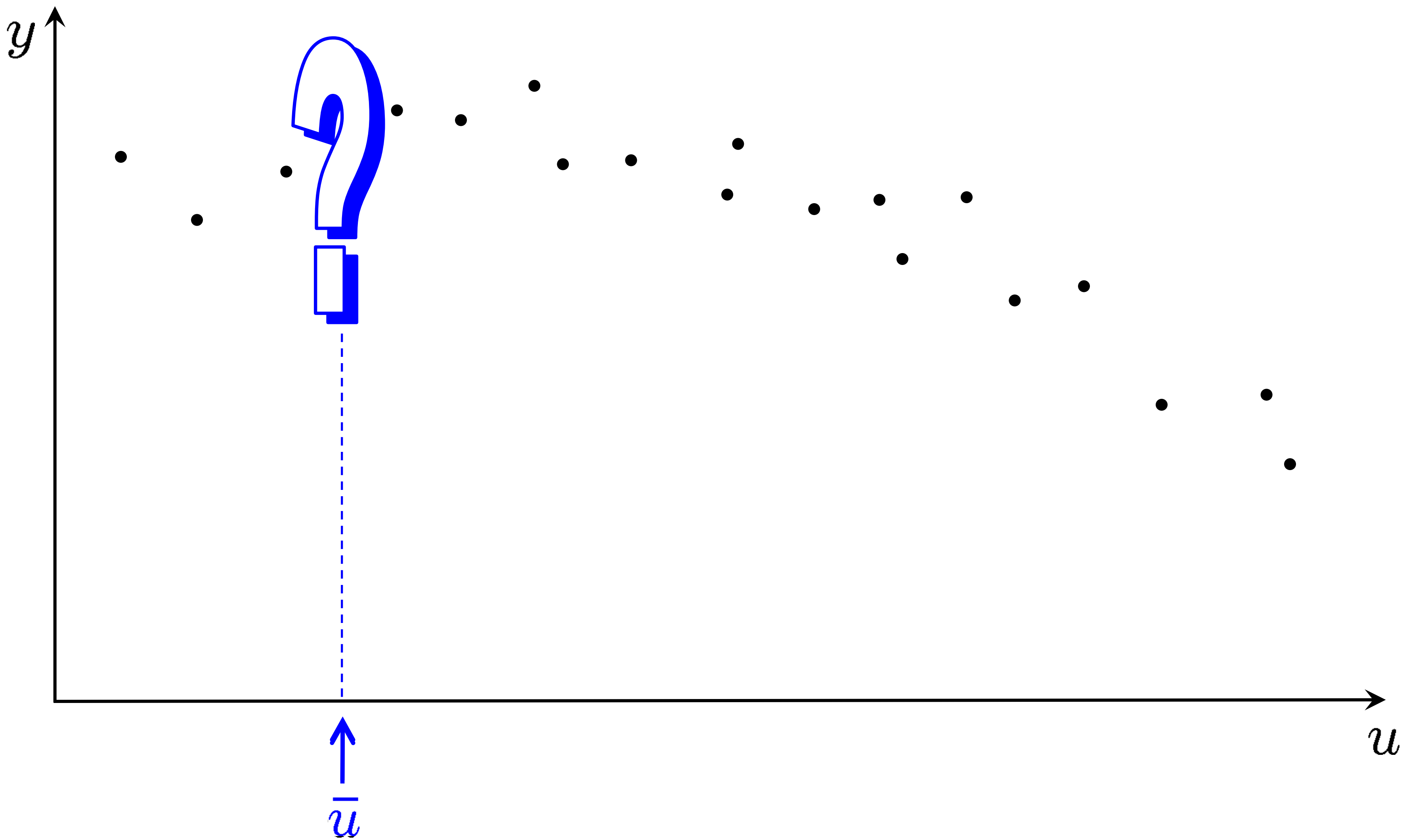
A journey through:

- parameter estimation
- prediction
- filtering

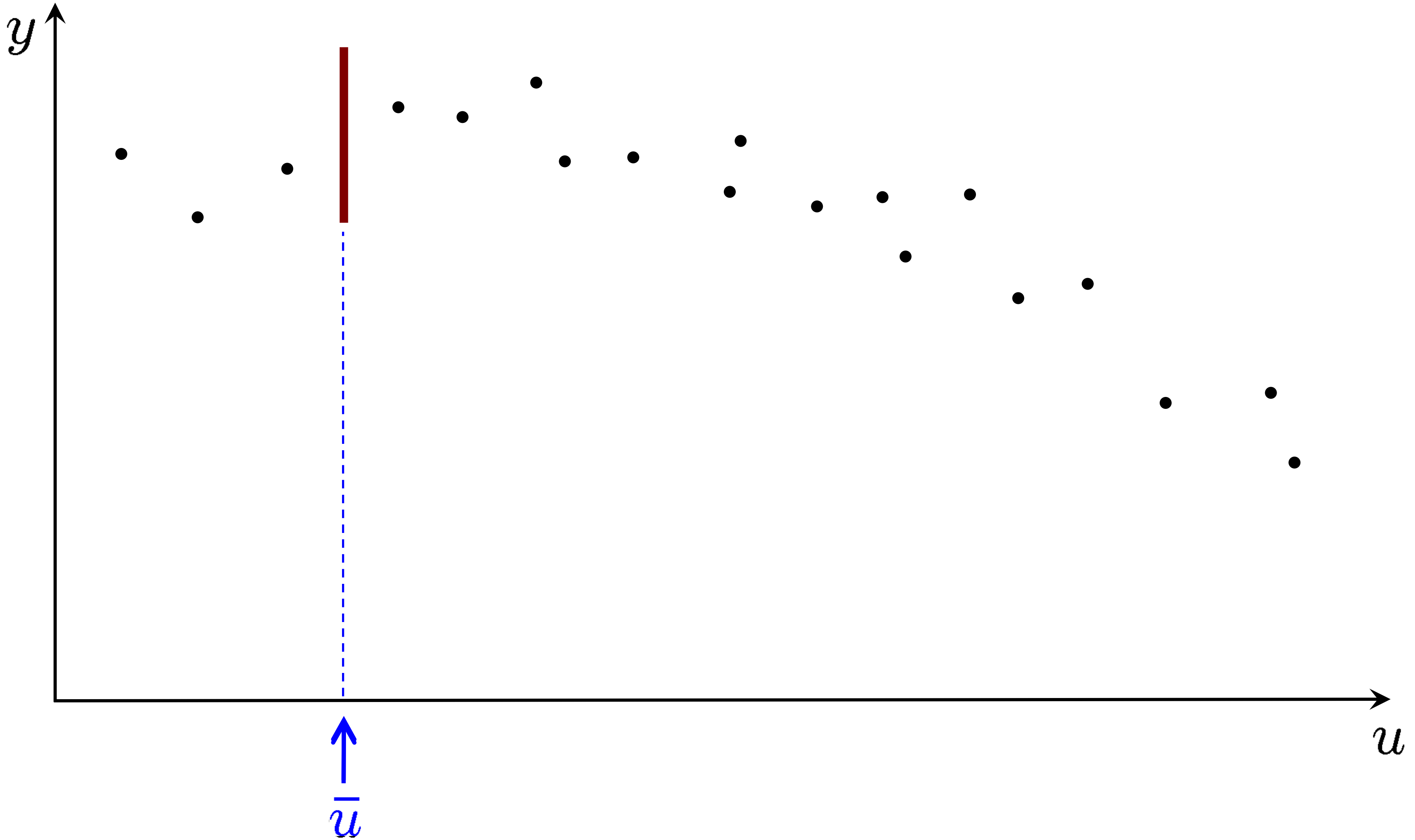
an example



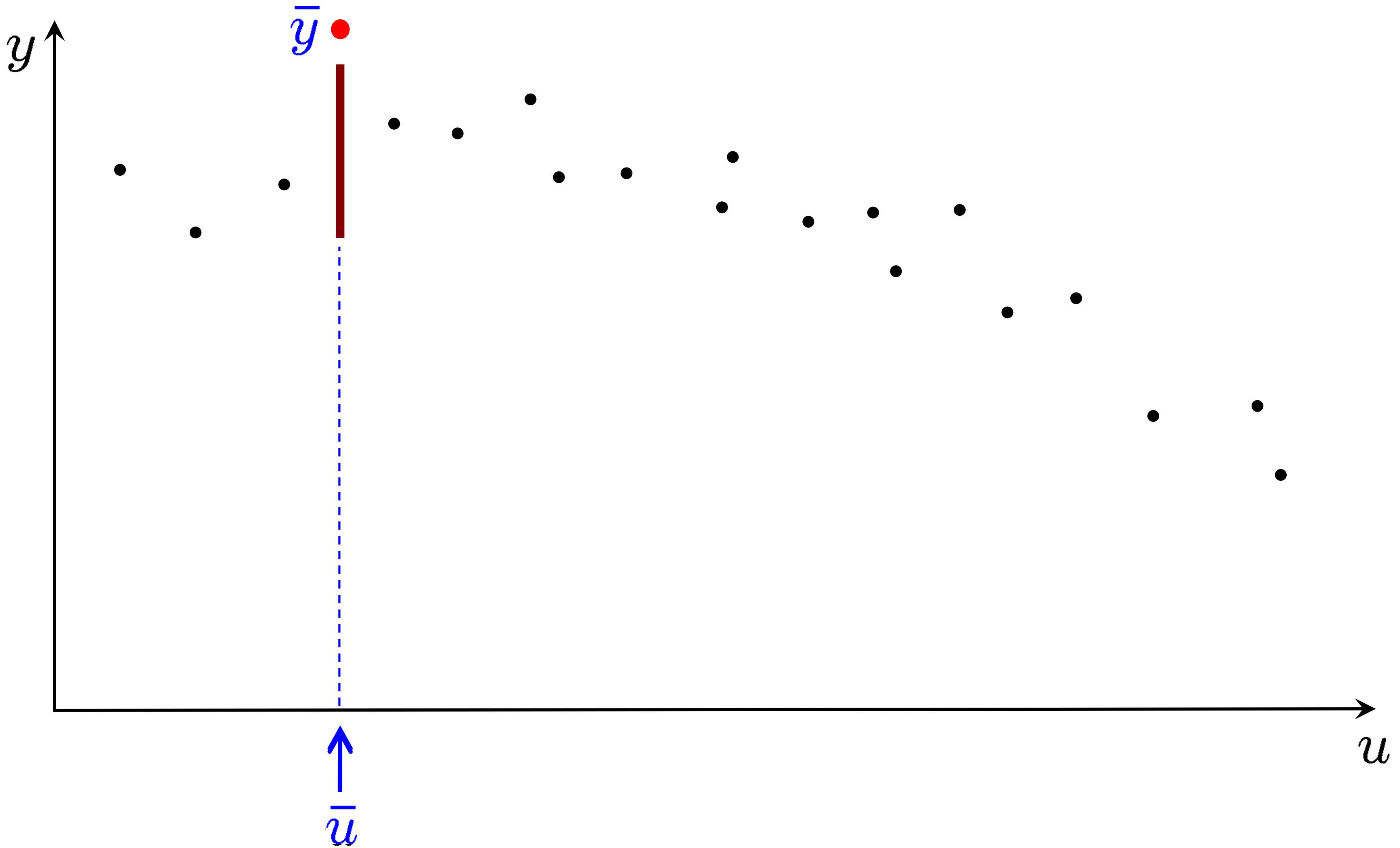
an example



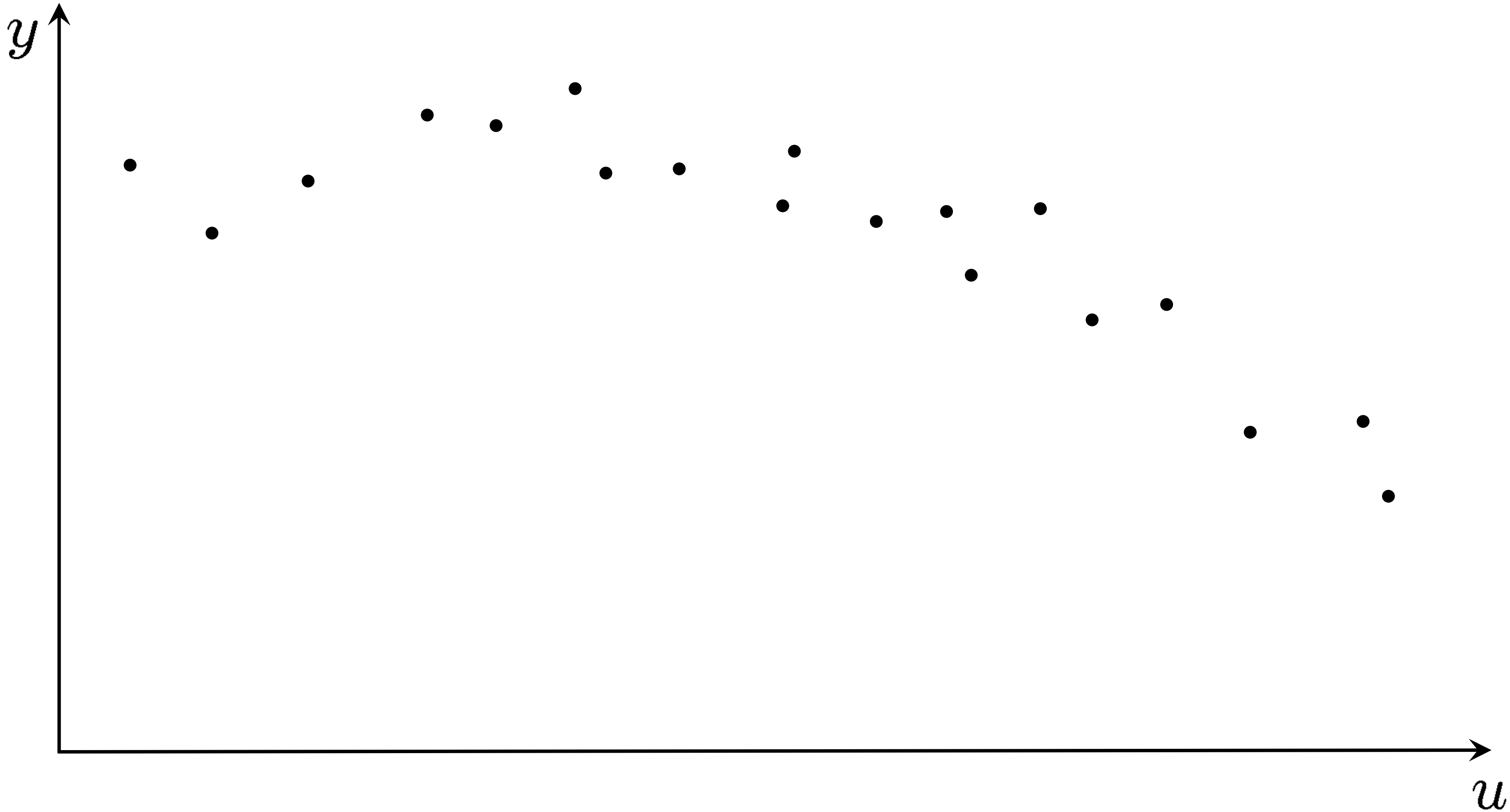
an example



an example



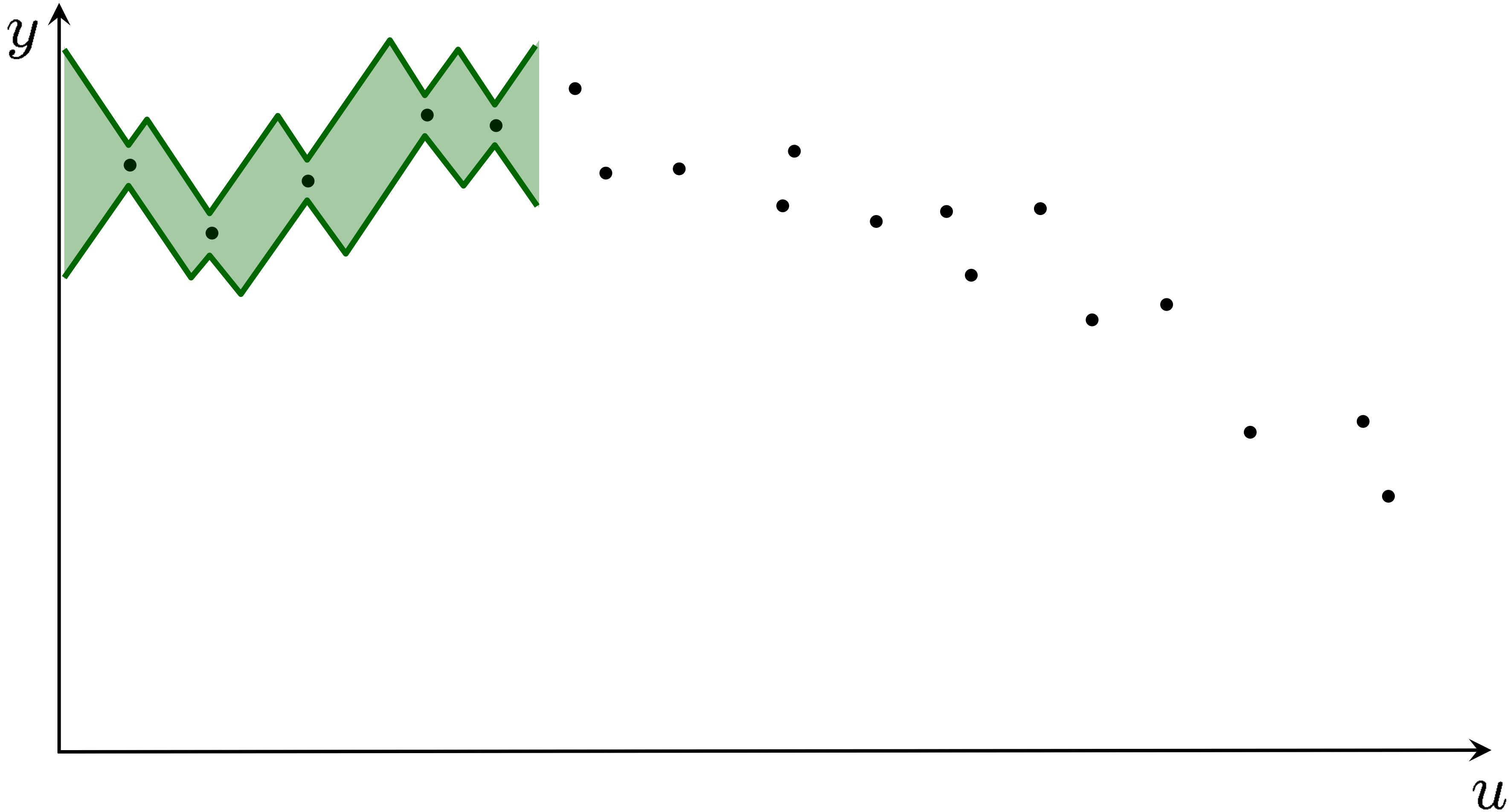
an example



deterministic assumption: $y = f(u) + n$ $\quad \left|\dfrac{\partial f}{\partial u}\right| \leq 1$

$|n| \leq 0.5$

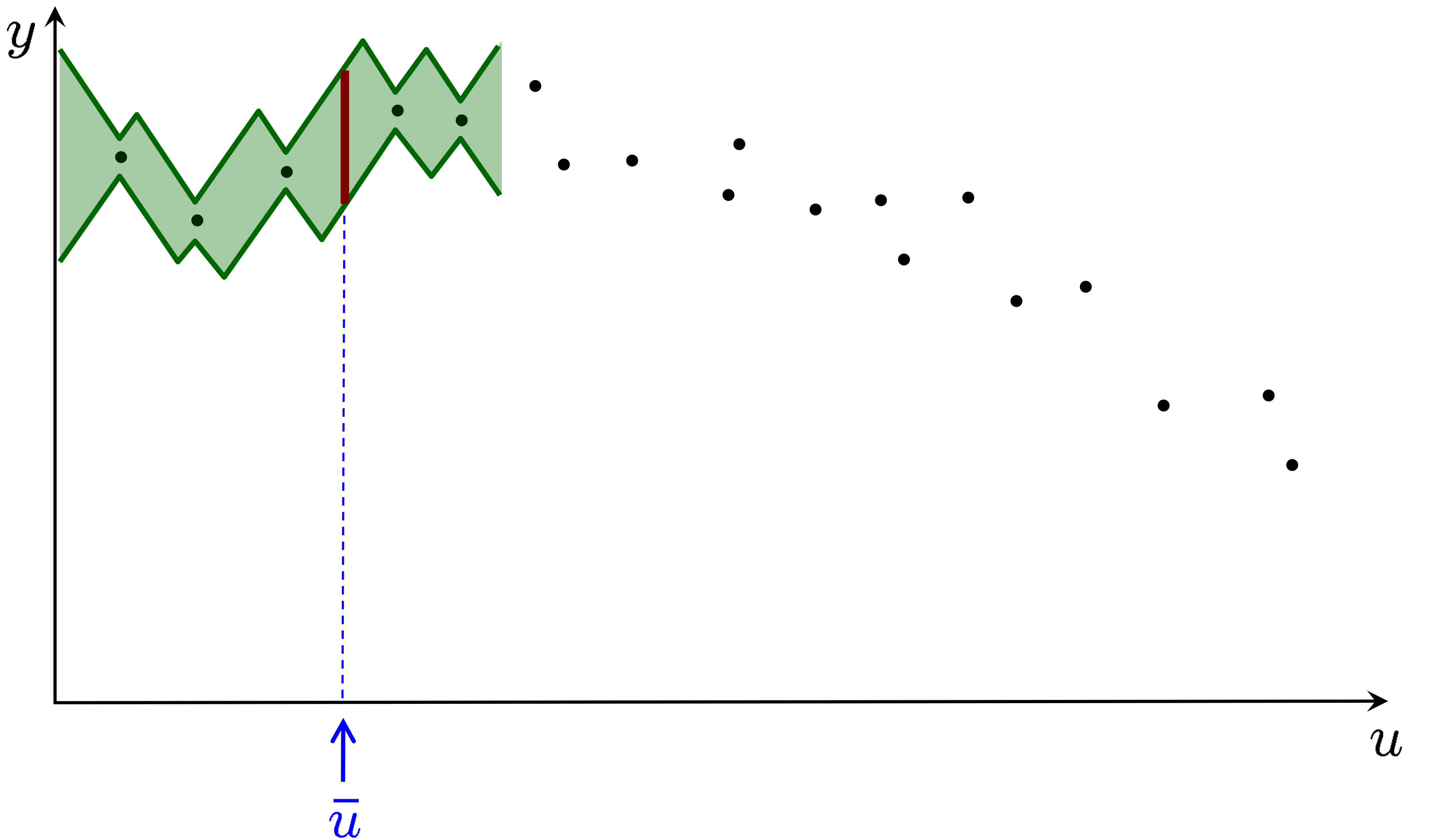
an example



deterministic assumption: $y = f(u) + n$ $\left|\frac{\partial f}{\partial u}\right| \leq 1$

$|n| \leq 0.5$

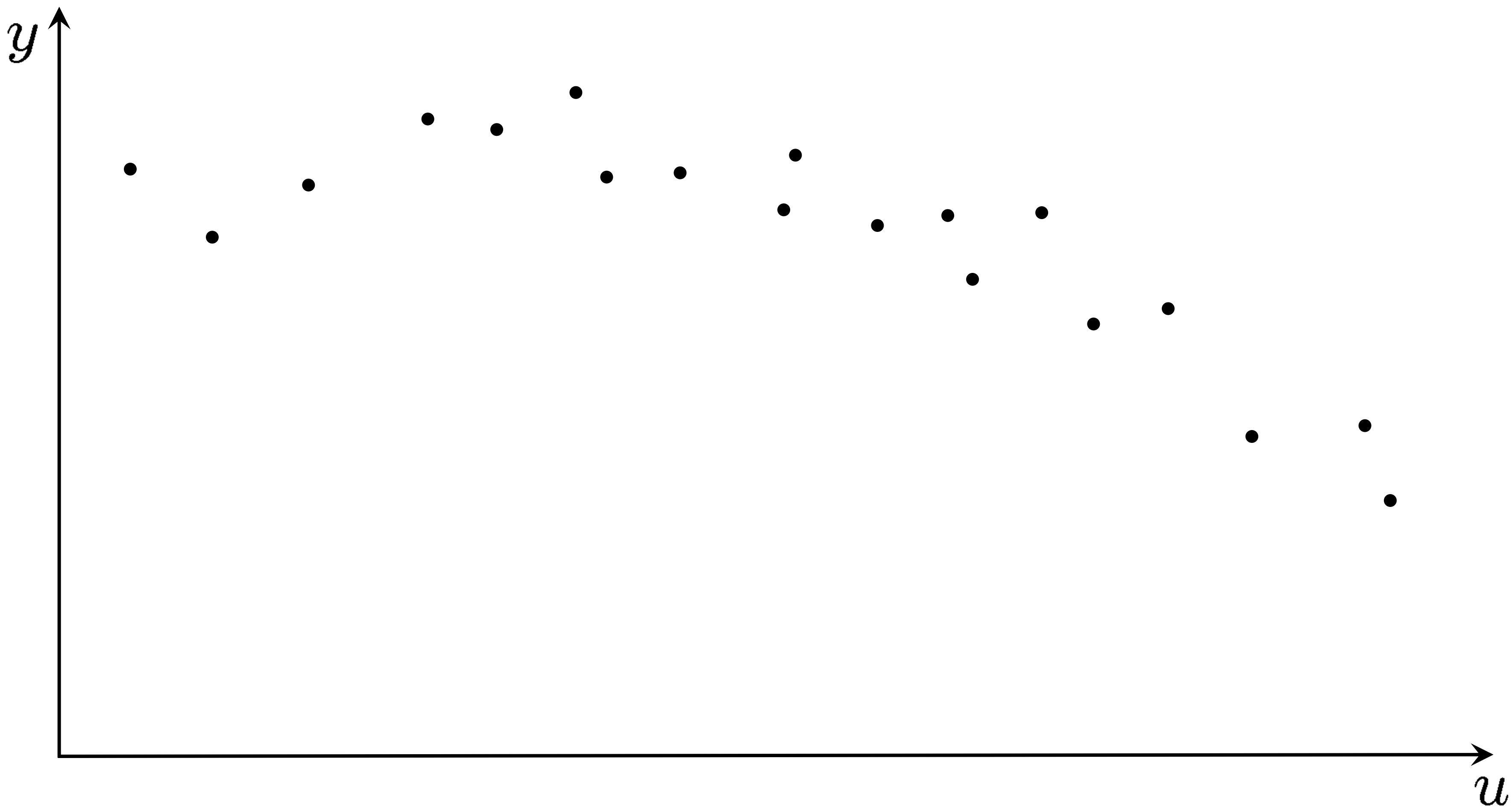
an example



do we really believe in these priors?

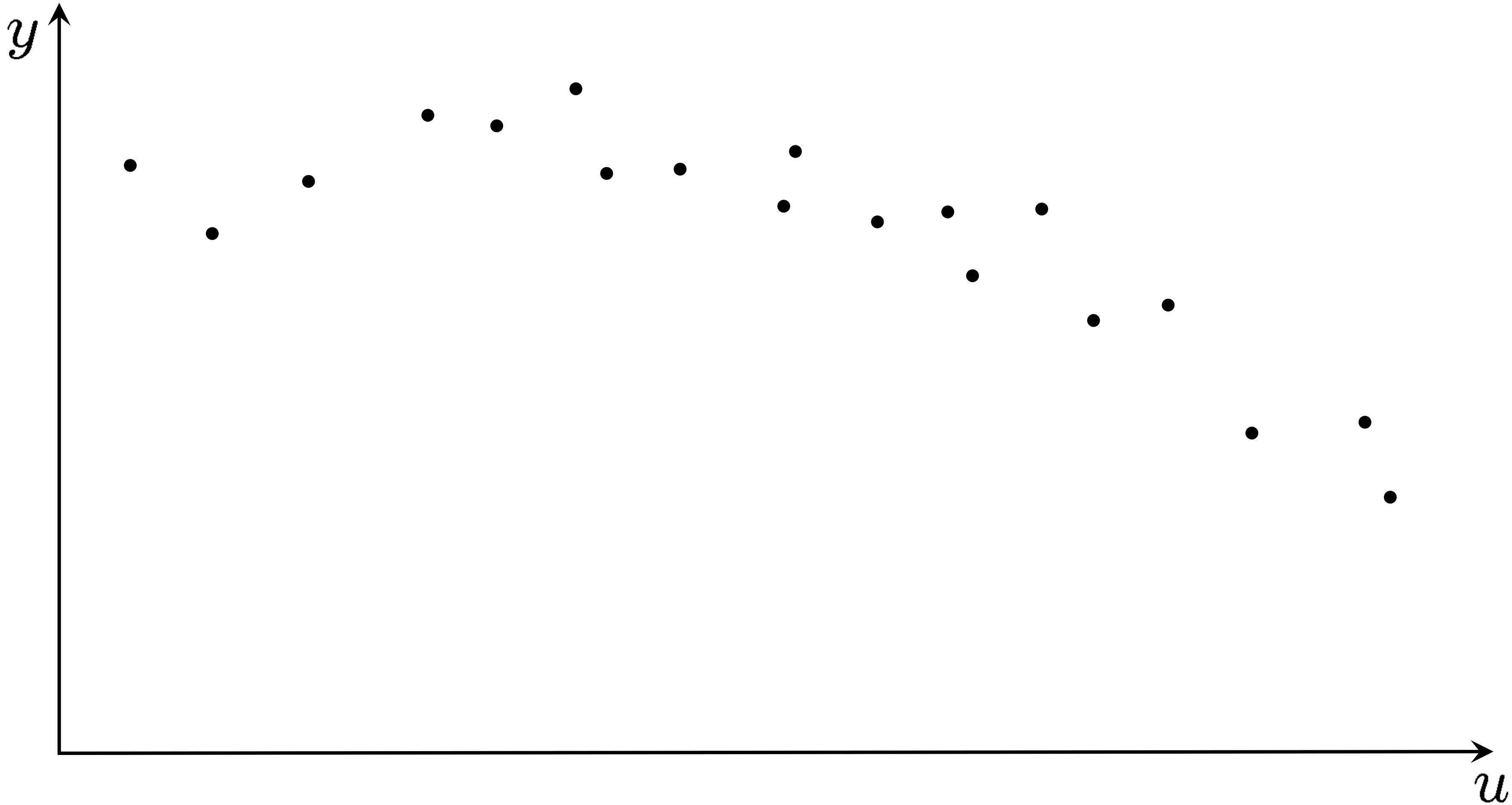
a paradigm shift

an example

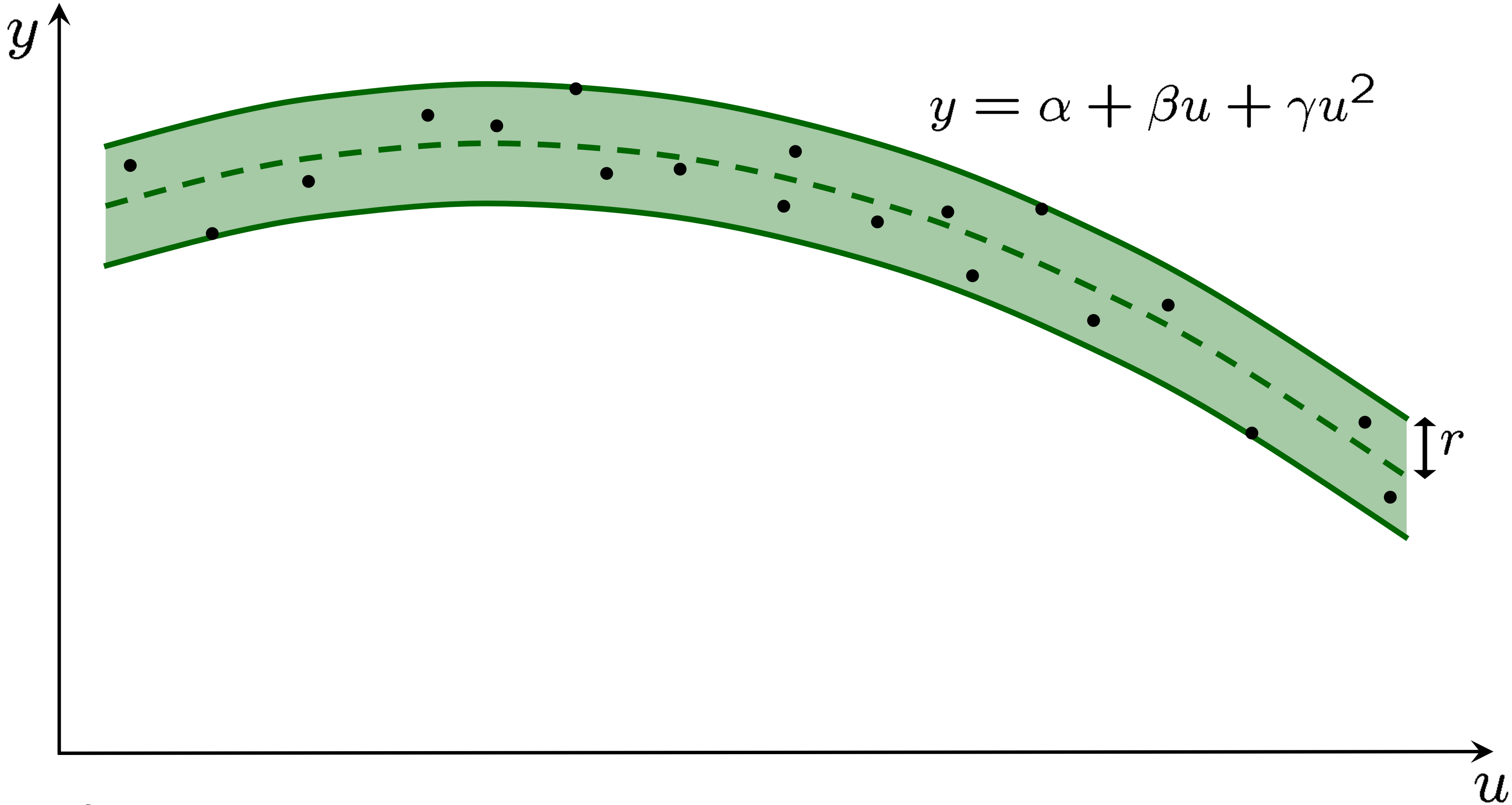


no assumptions

an example



an example

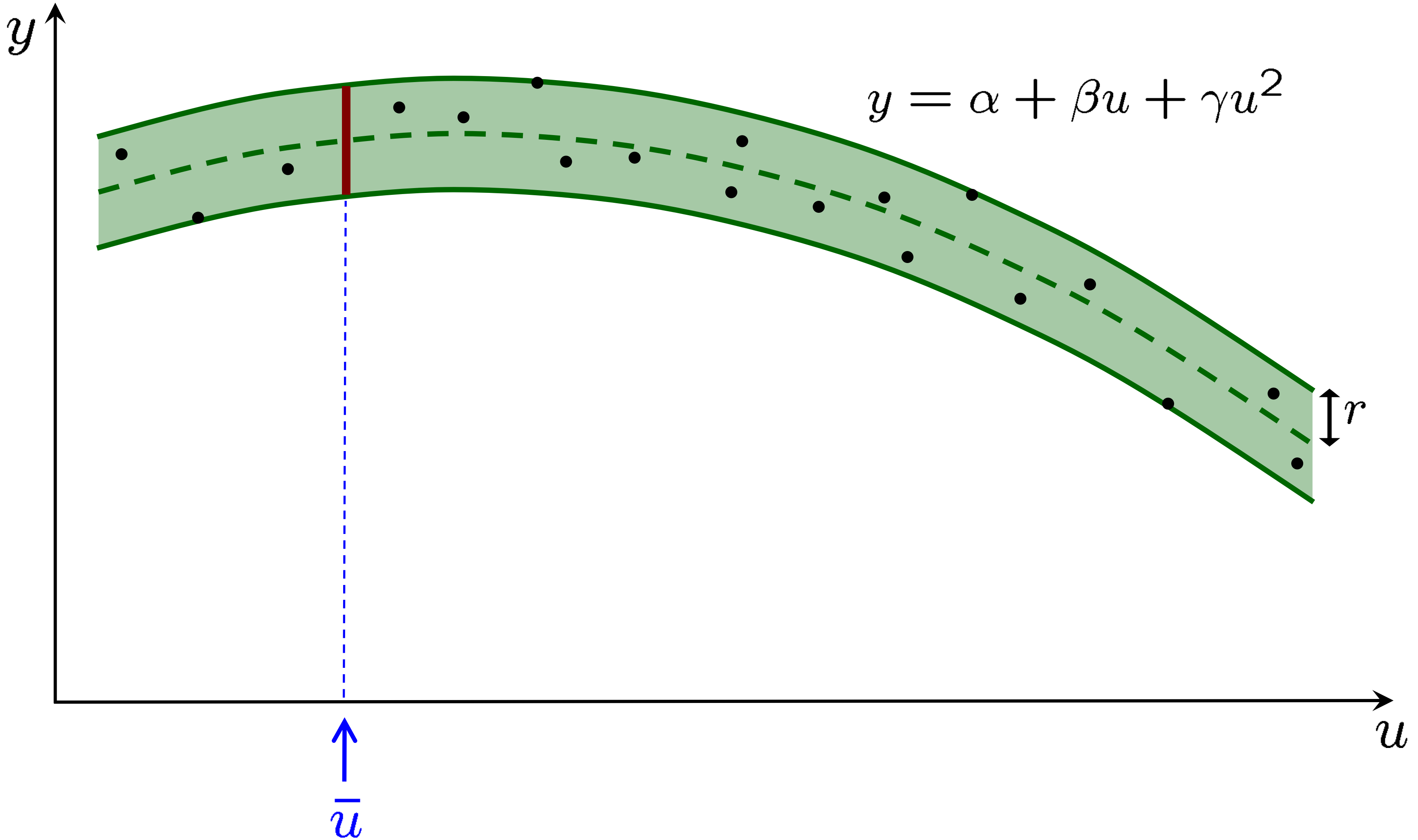


$$\min \quad r$$

$$\text{subject to:} \quad |y_i - [\alpha + \beta u_i + \gamma u_i^2]| \leq r, \quad i = 1, \ldots, 19$$

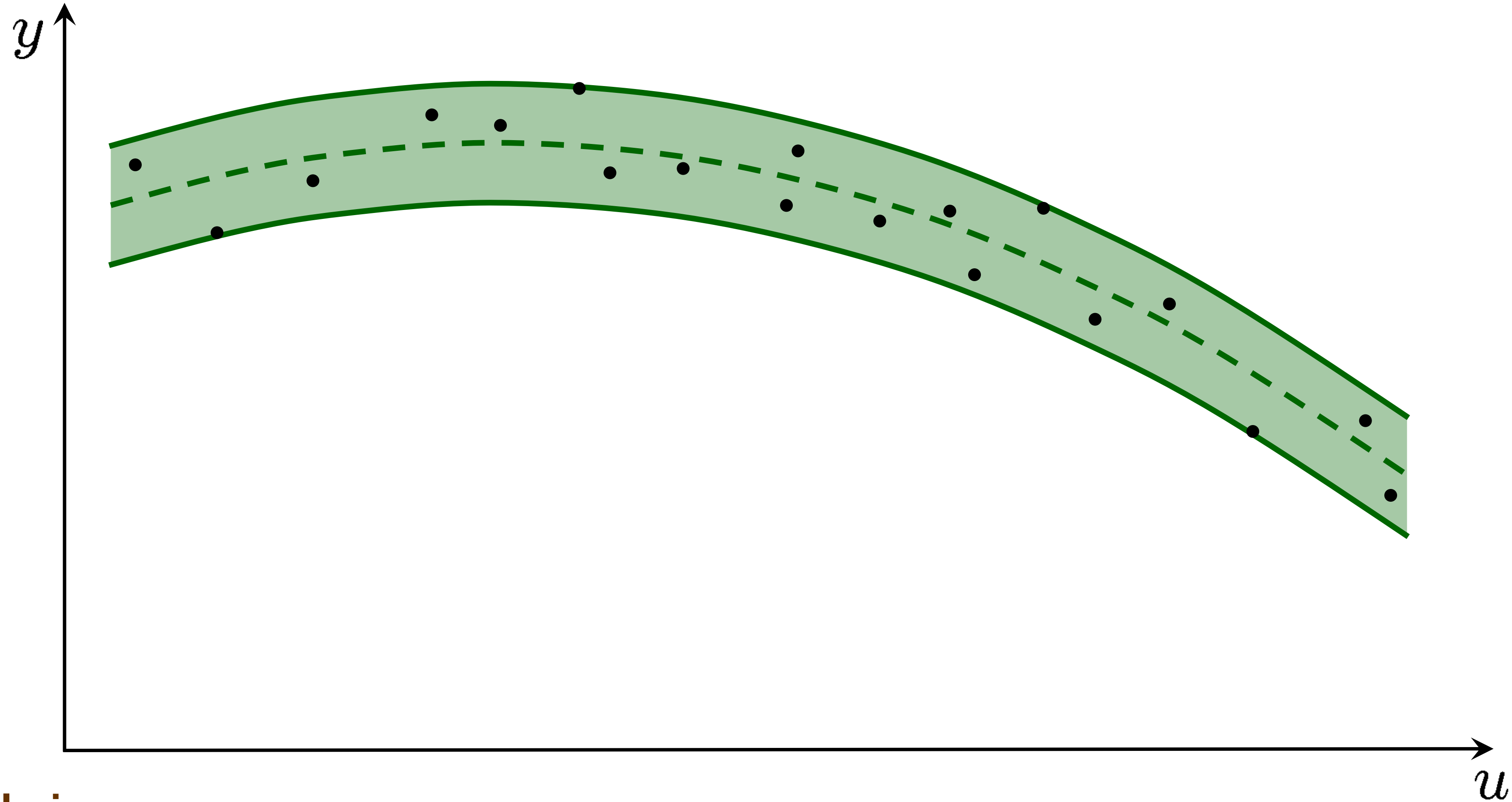
[Campi, Calafiore, Garatti, Automatica, 2009]

an example



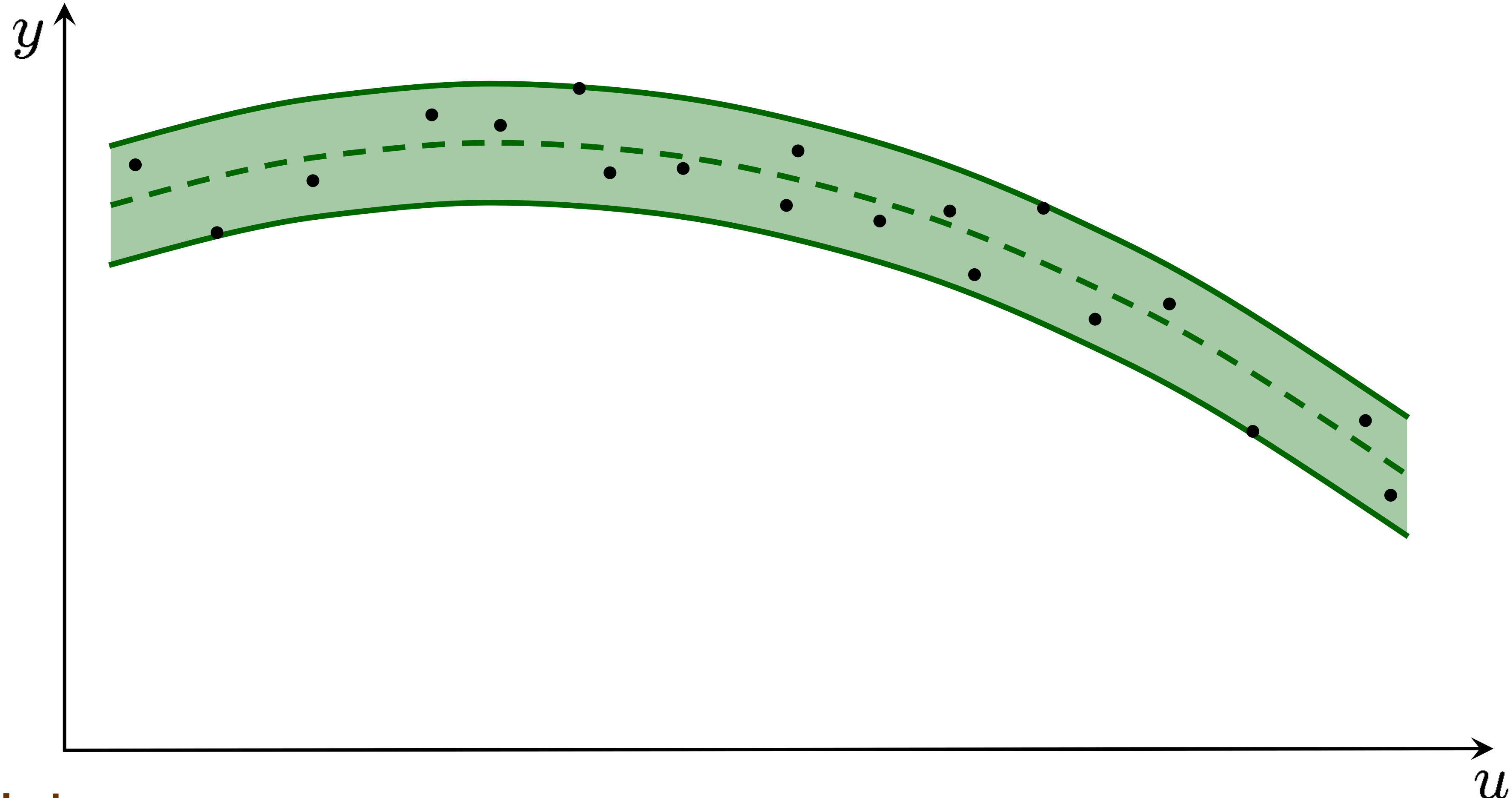
[Campi, Calafiore, Garatti, Automatica, 2009]

an example



Claim: The prediction is correct with exact probability 80%, irrespective of the probability with which data are generated

an example



Claim: The prediction is correct with exact probability 80%, irrespective of the probability with which data are generated

totally distribution-free!

Theorem [Campi, Calafiore, Garatti, Automatica, 2009]

k = degrees of freedom of centerline
N = # of data points

The layer predicts correctly with exact probability p if

$$N = \frac{k+p}{1-p}$$

Theorem [Campi, Calafiore, Garatti, Automatica, 2009]

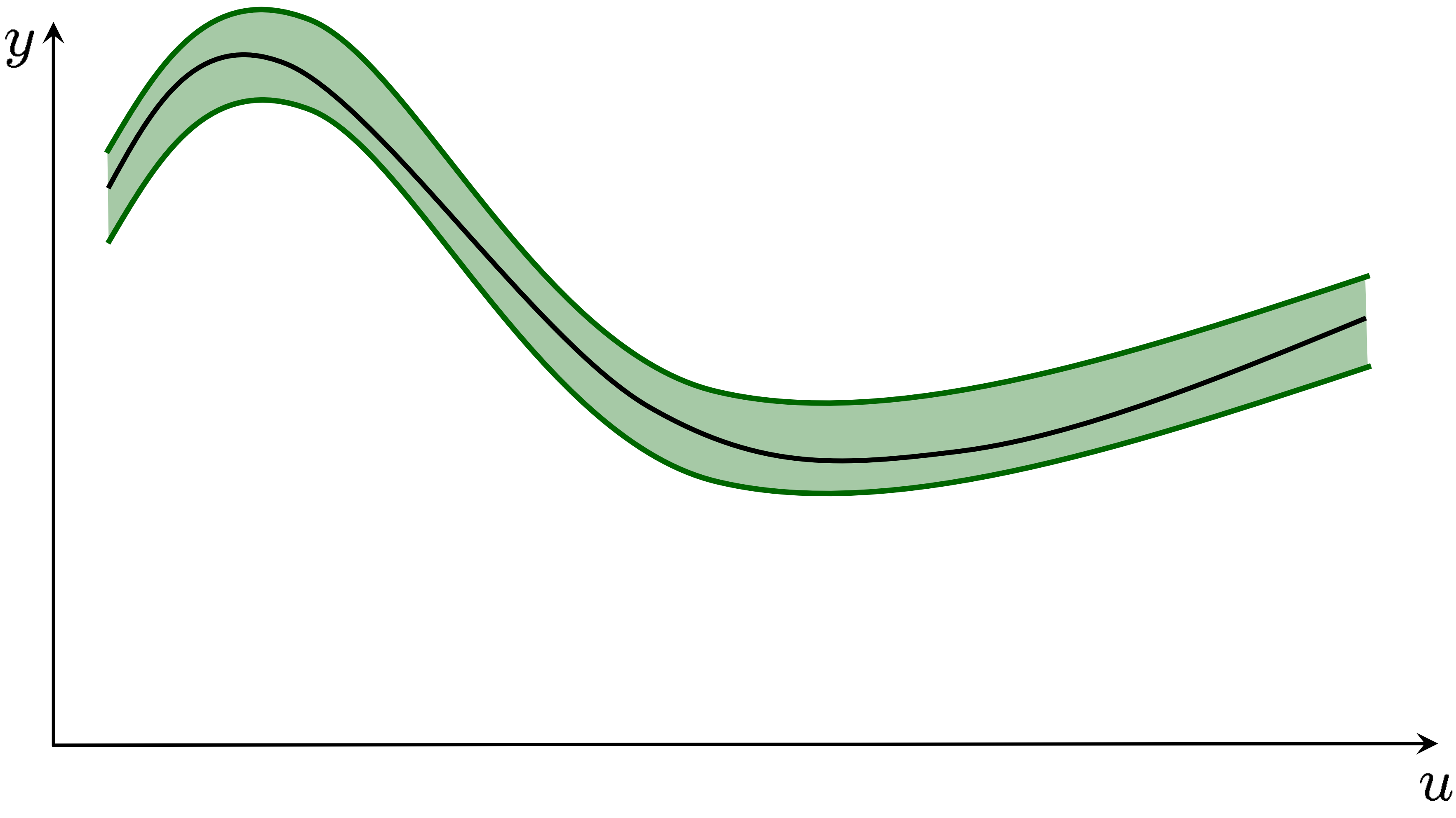
k = degrees of freedom of centerline
N = # of data points

The layer predicts correctly with exact probability p if

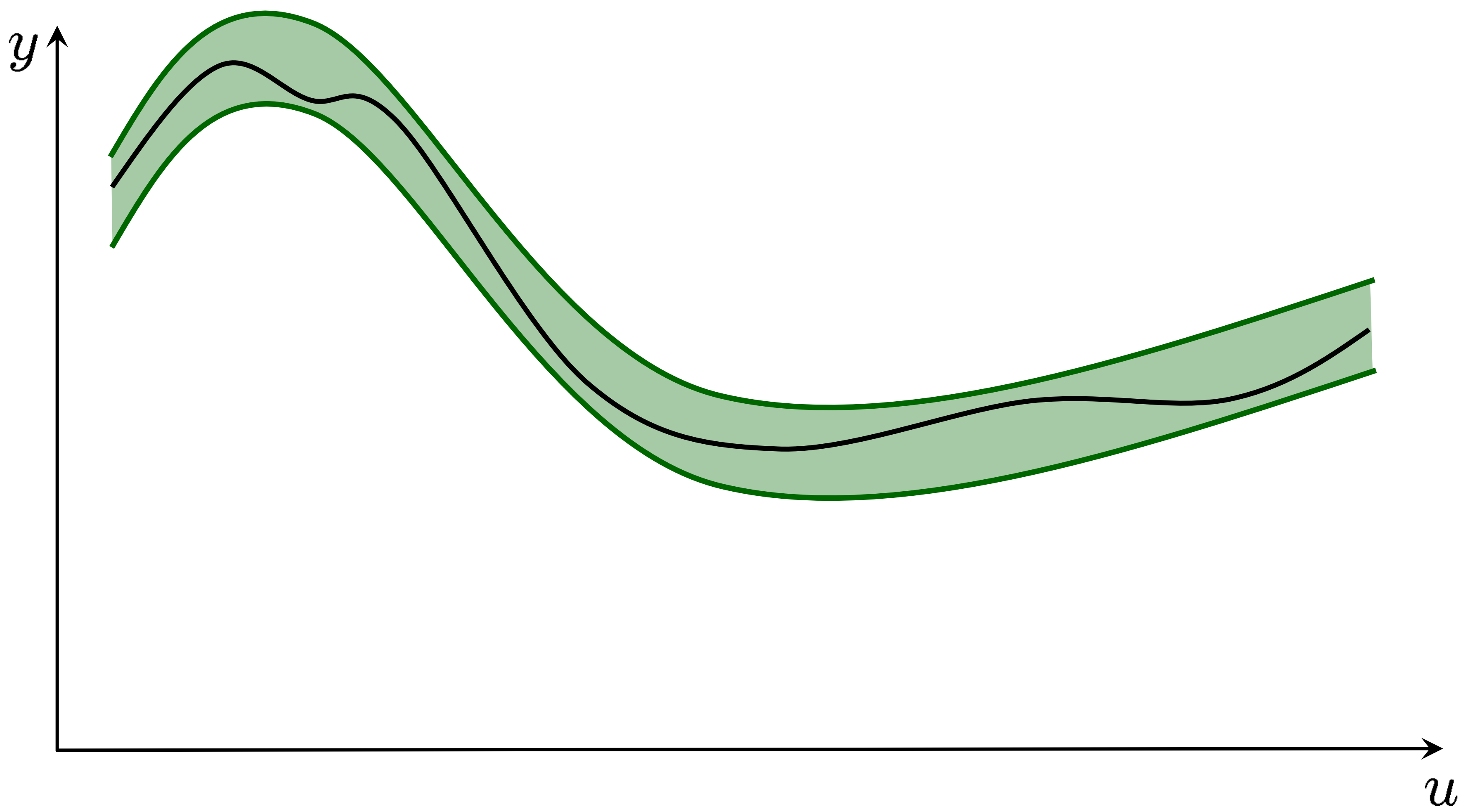
$$N = \frac{k+p}{1-p}$$

$p = 80\%$ $\Rightarrow$ $\frac{k+p}{1-p} = \frac{3+0.8}{1-0.8} = 19$

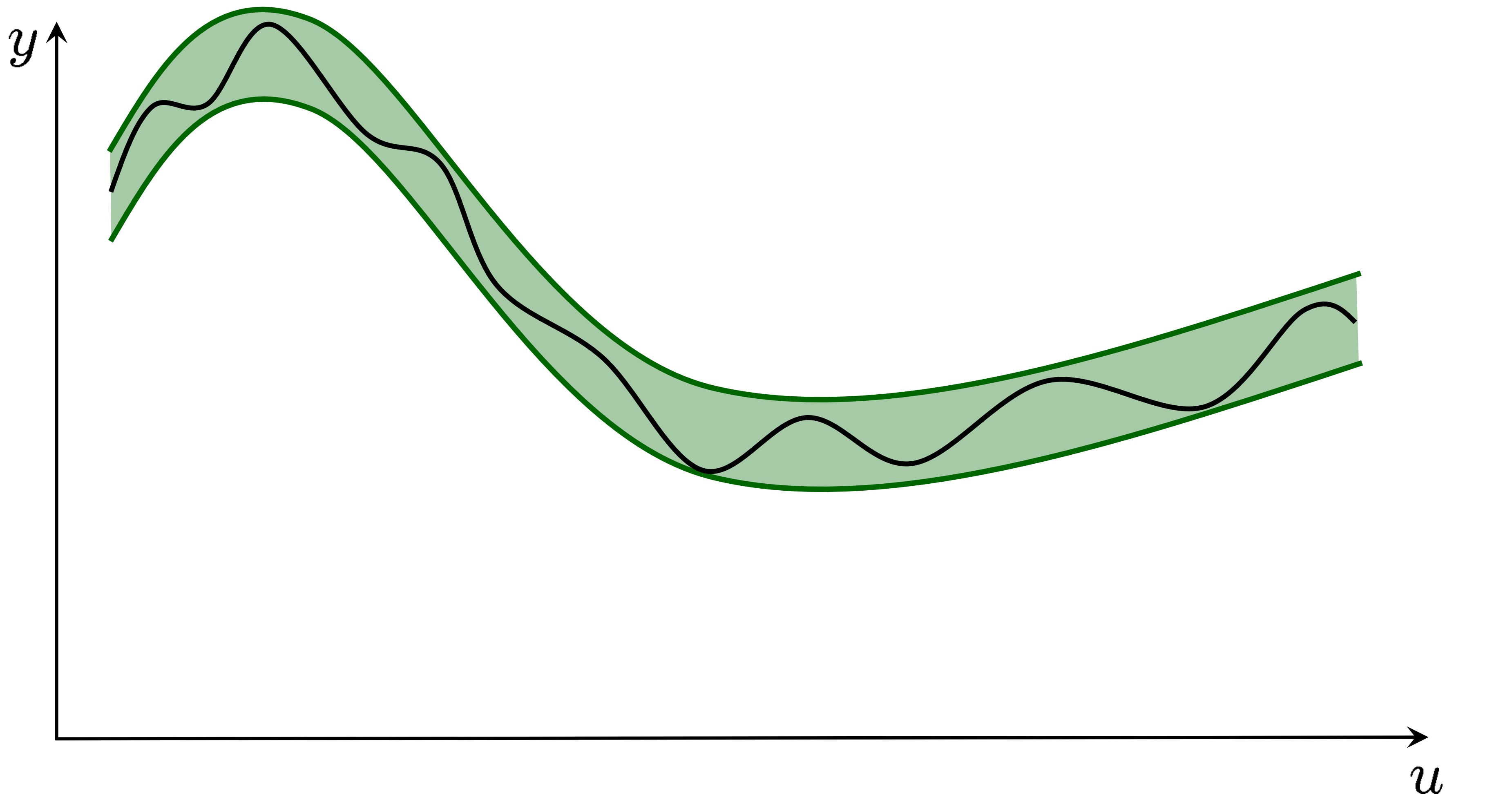
why is this possible?



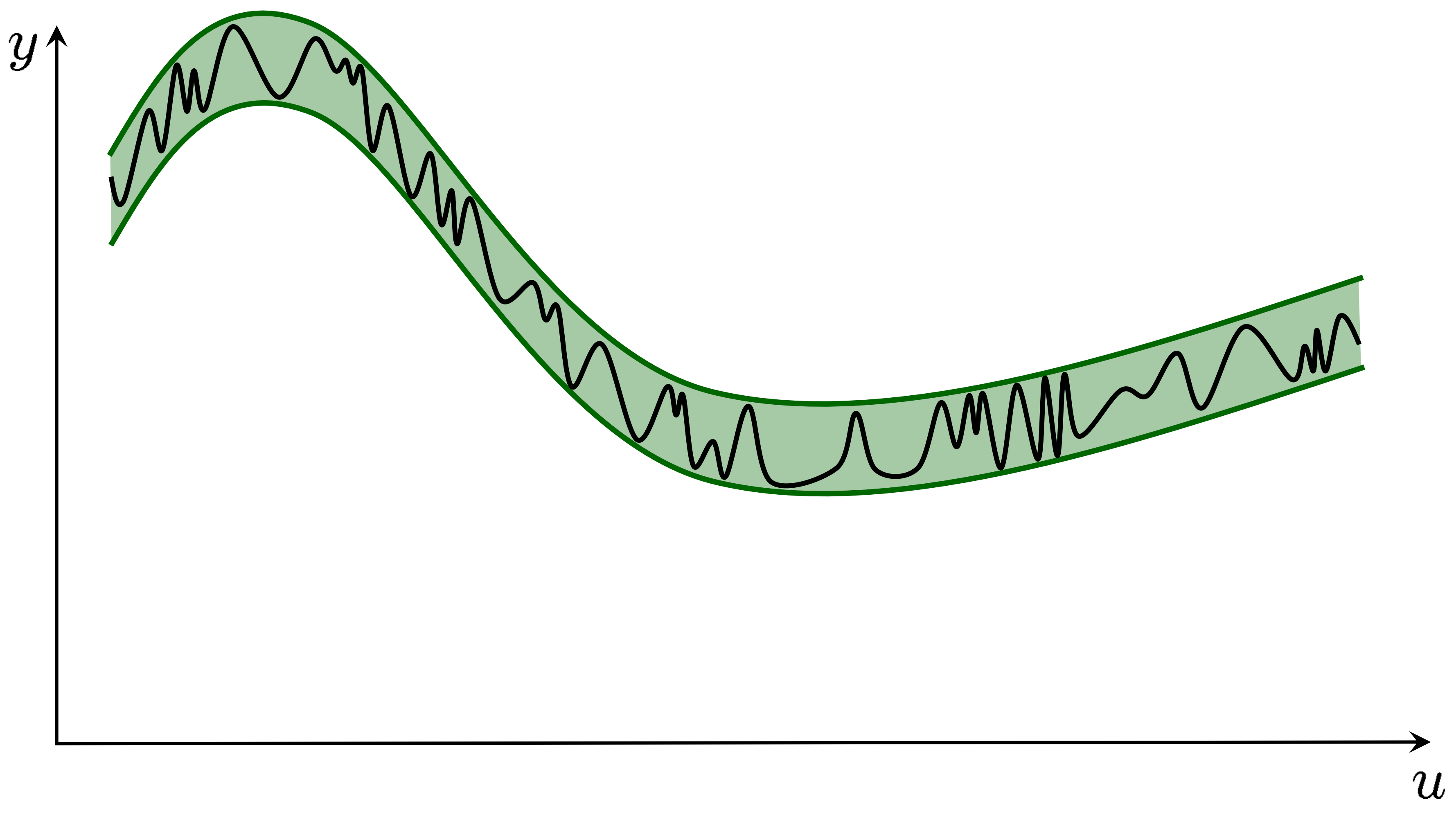
y
u



y
u

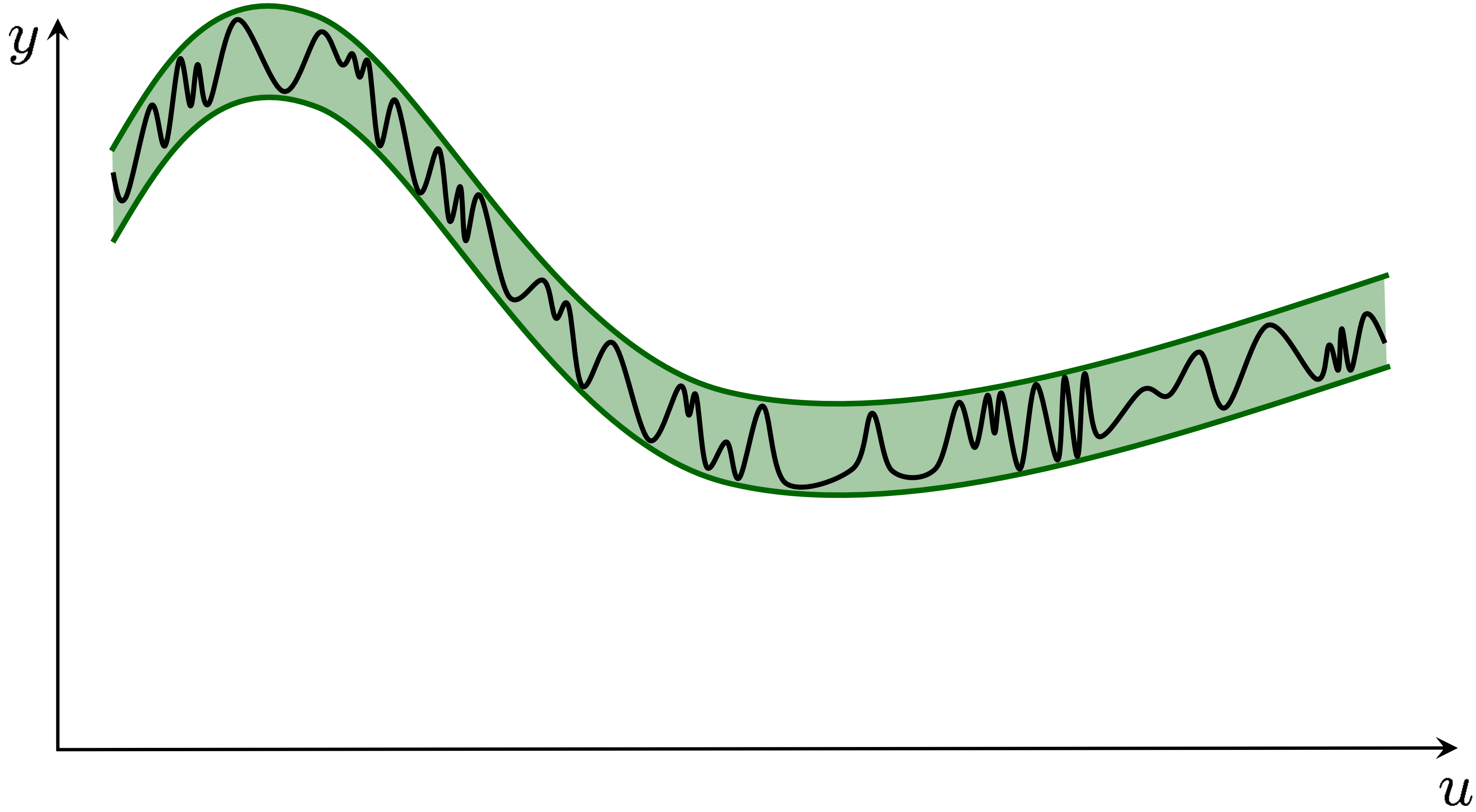


y
u

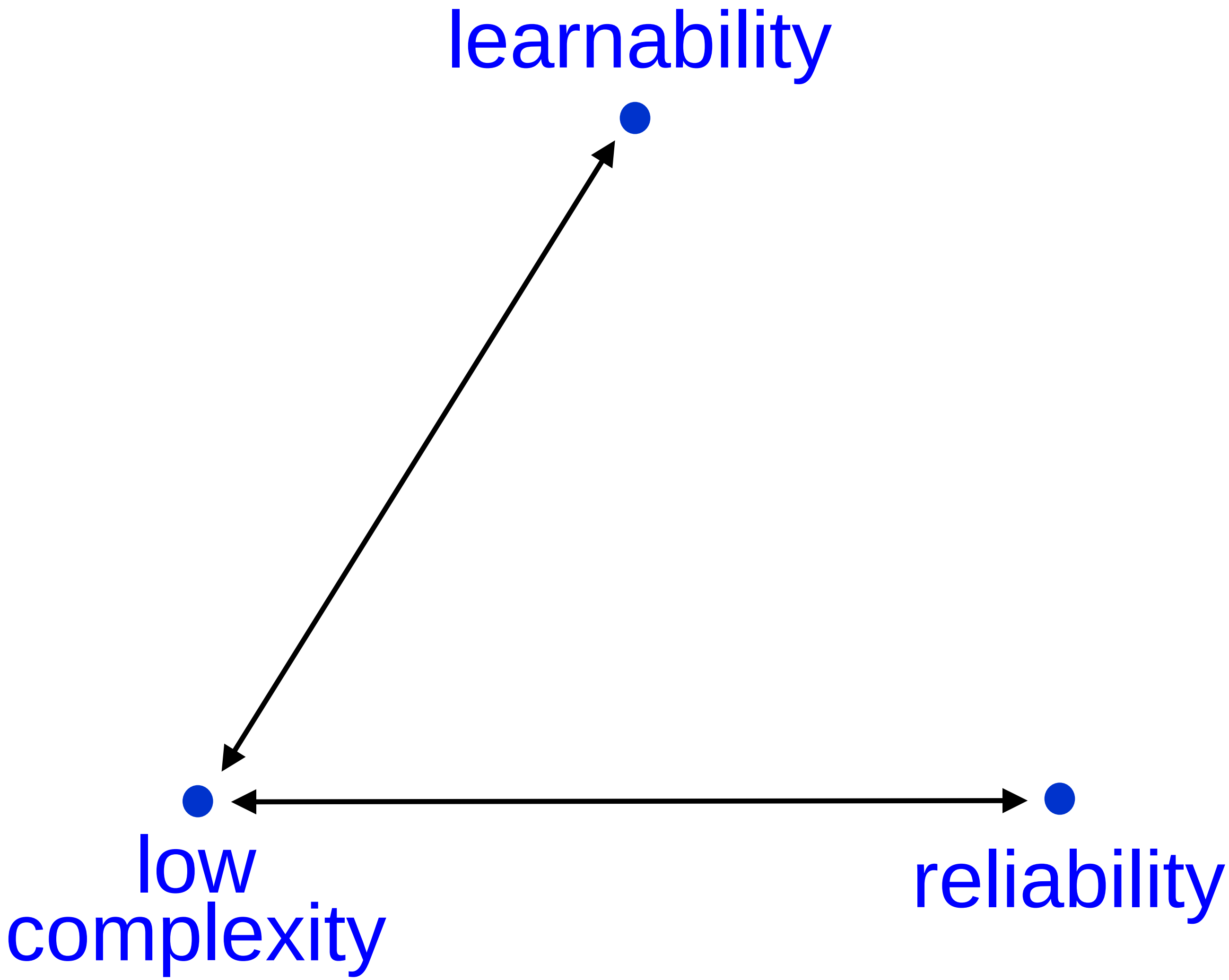


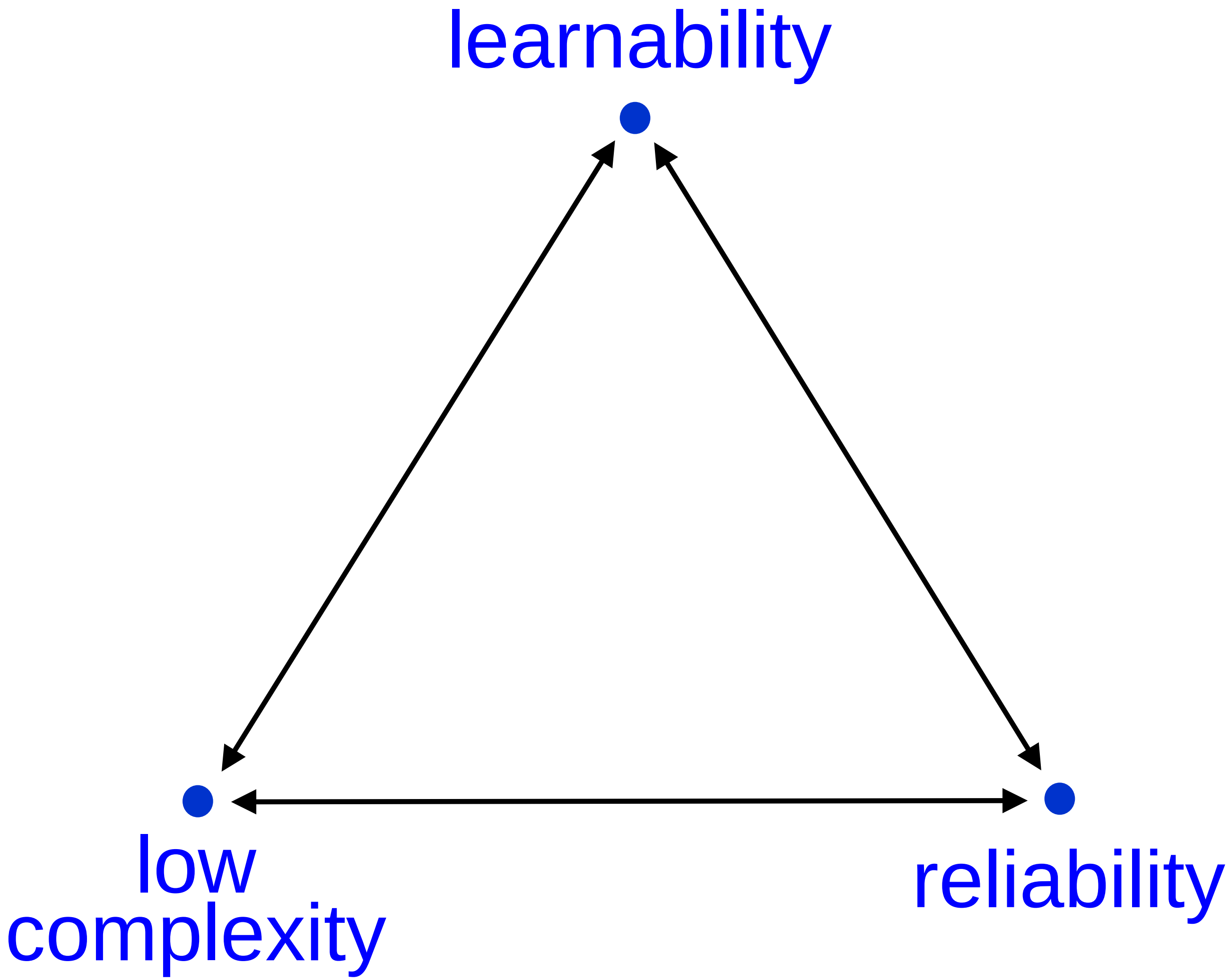
y
u

a simple predictor can be reliable for complex data generation mechanisms



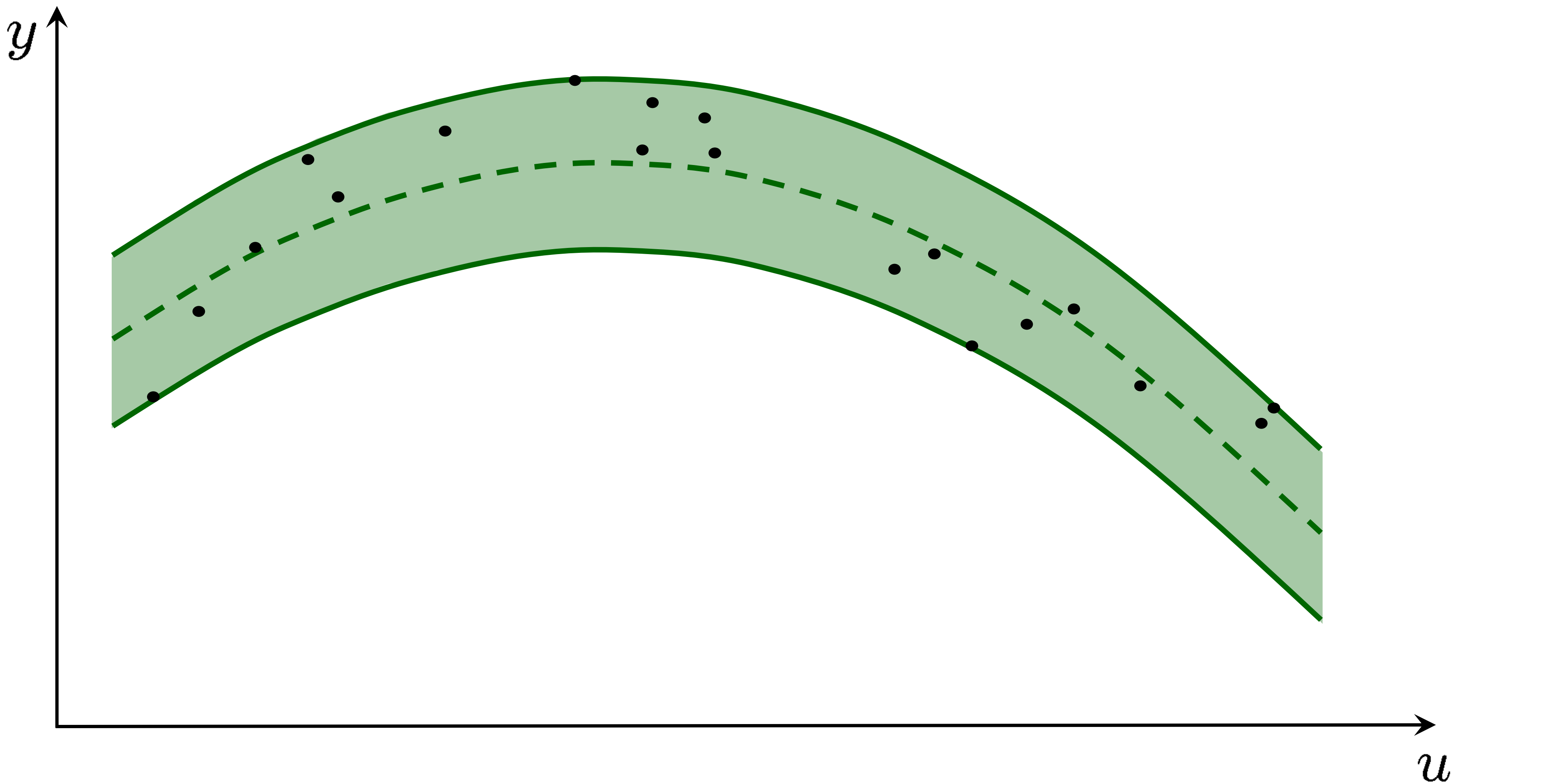

low
complexity
reliability


learnability
low
complexity
reliability

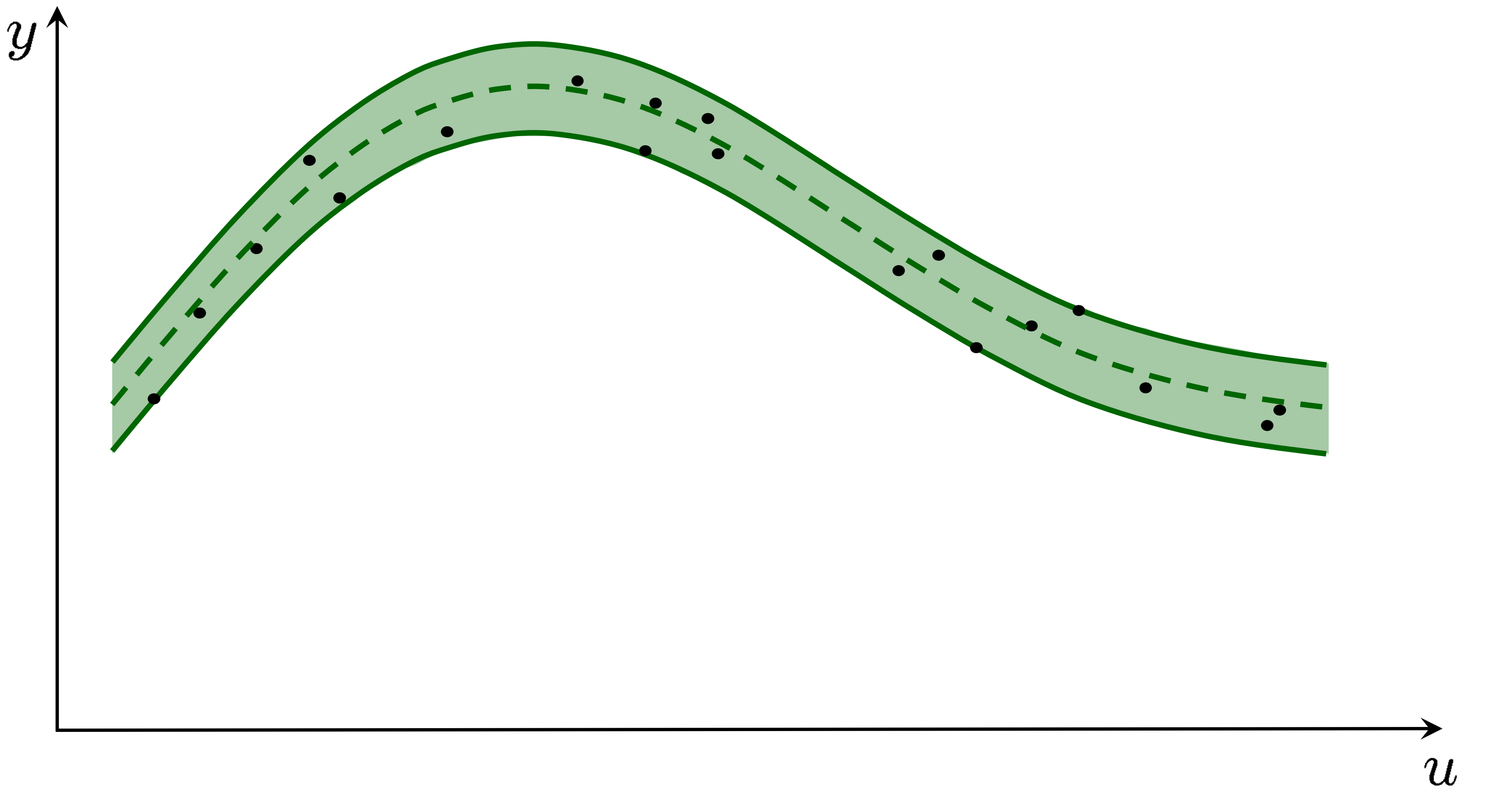

learnability
low
complexity
reliability

what is the role of priors?

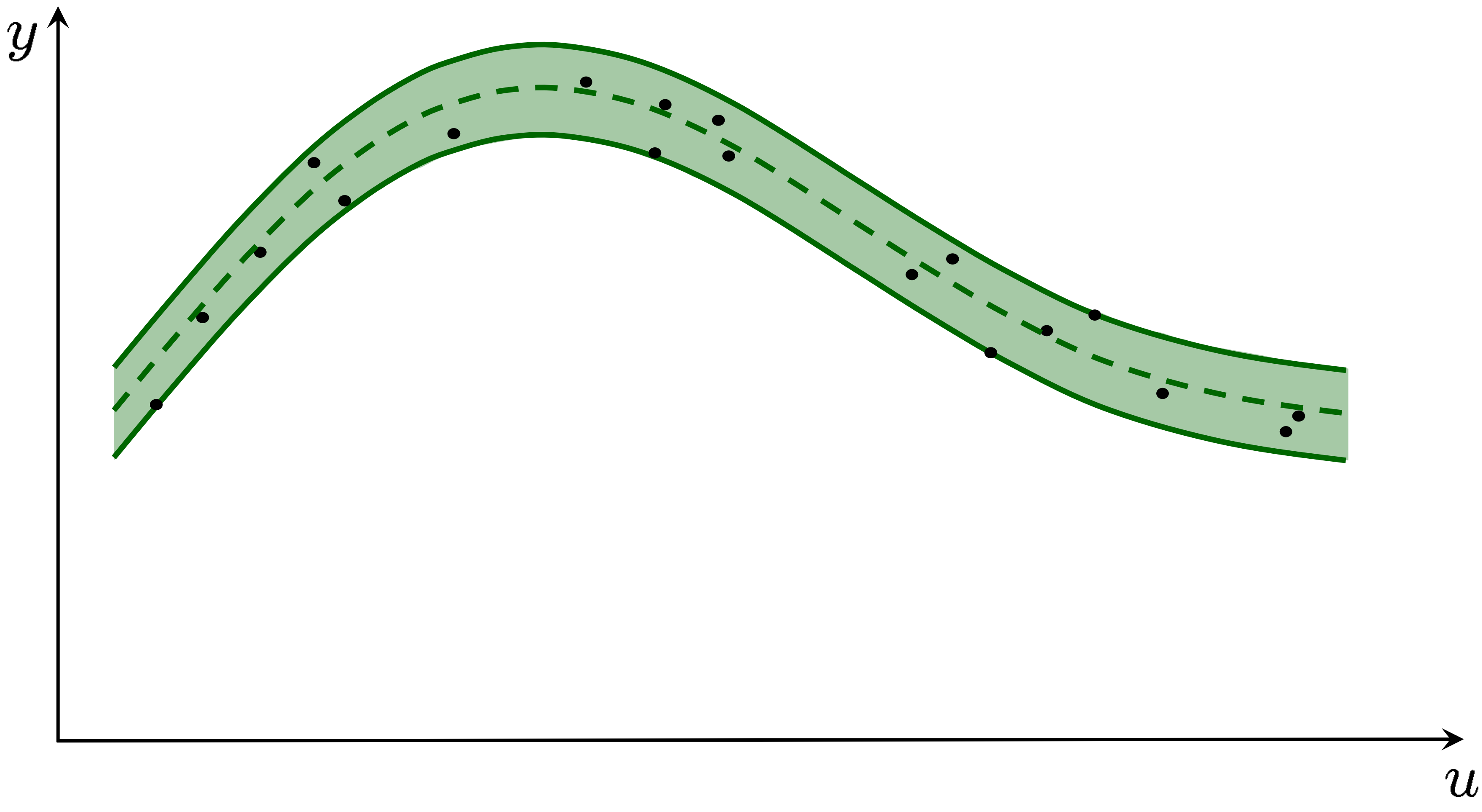
what is the role of priors?



what is the role of priors?



what is the role of priors?



“thickness”

an interesting fact:

an interesting fact:

- “reliable” is a property with 2 arguments:

  reliable(model, data generation mechanism);

an interesting fact:

- “reliable” is a property with 2 arguments:
  reliable(model, data generation mechanism);

- “thickness” is a property with 1 argument:
  thickness(model).

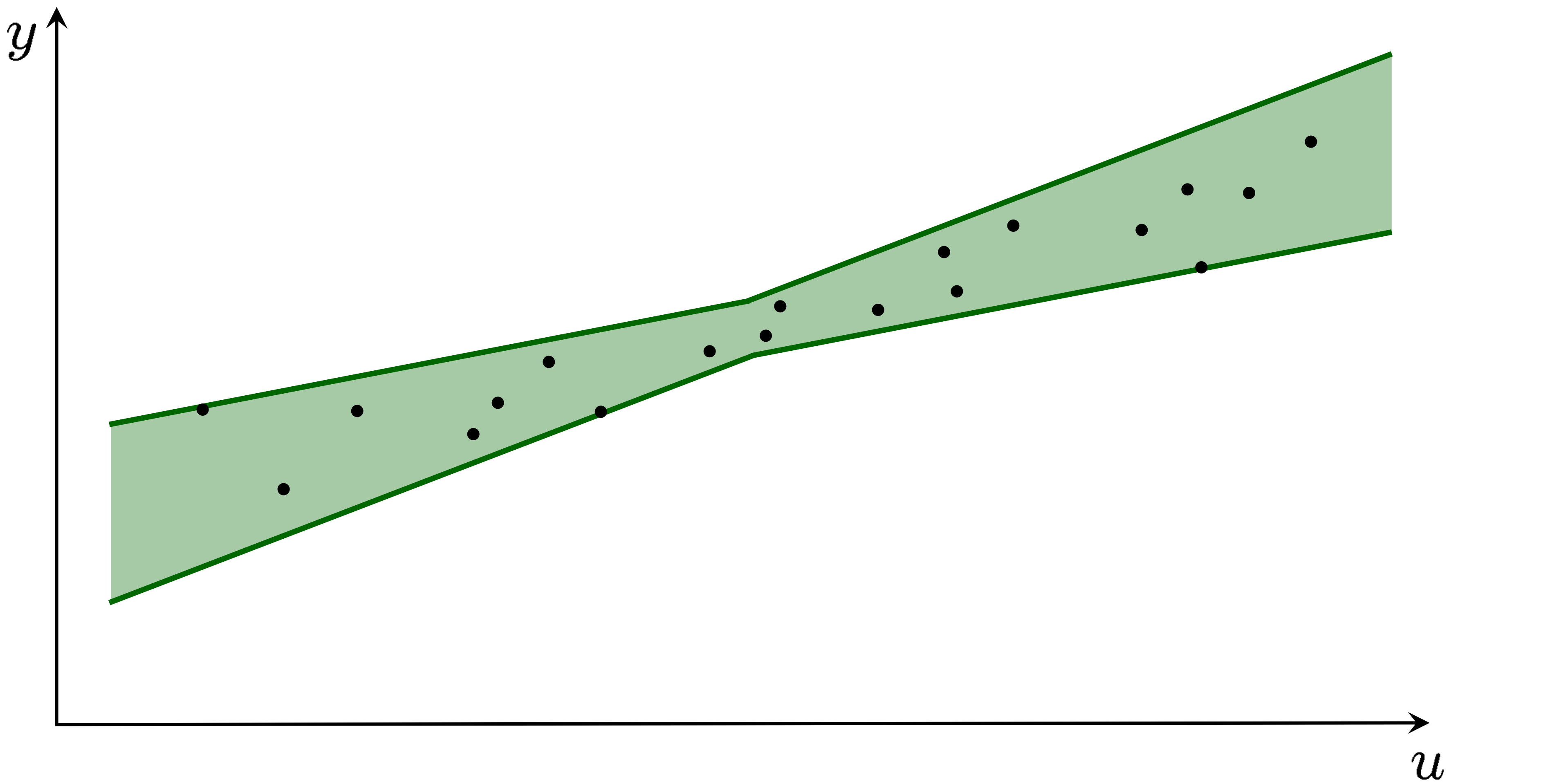
an interesting fact:

- “reliable” is a property with 2 arguments:
  reliable(model, data generation mechanism);

- “thickness” is a property with 1 argument:
  thickness(model).

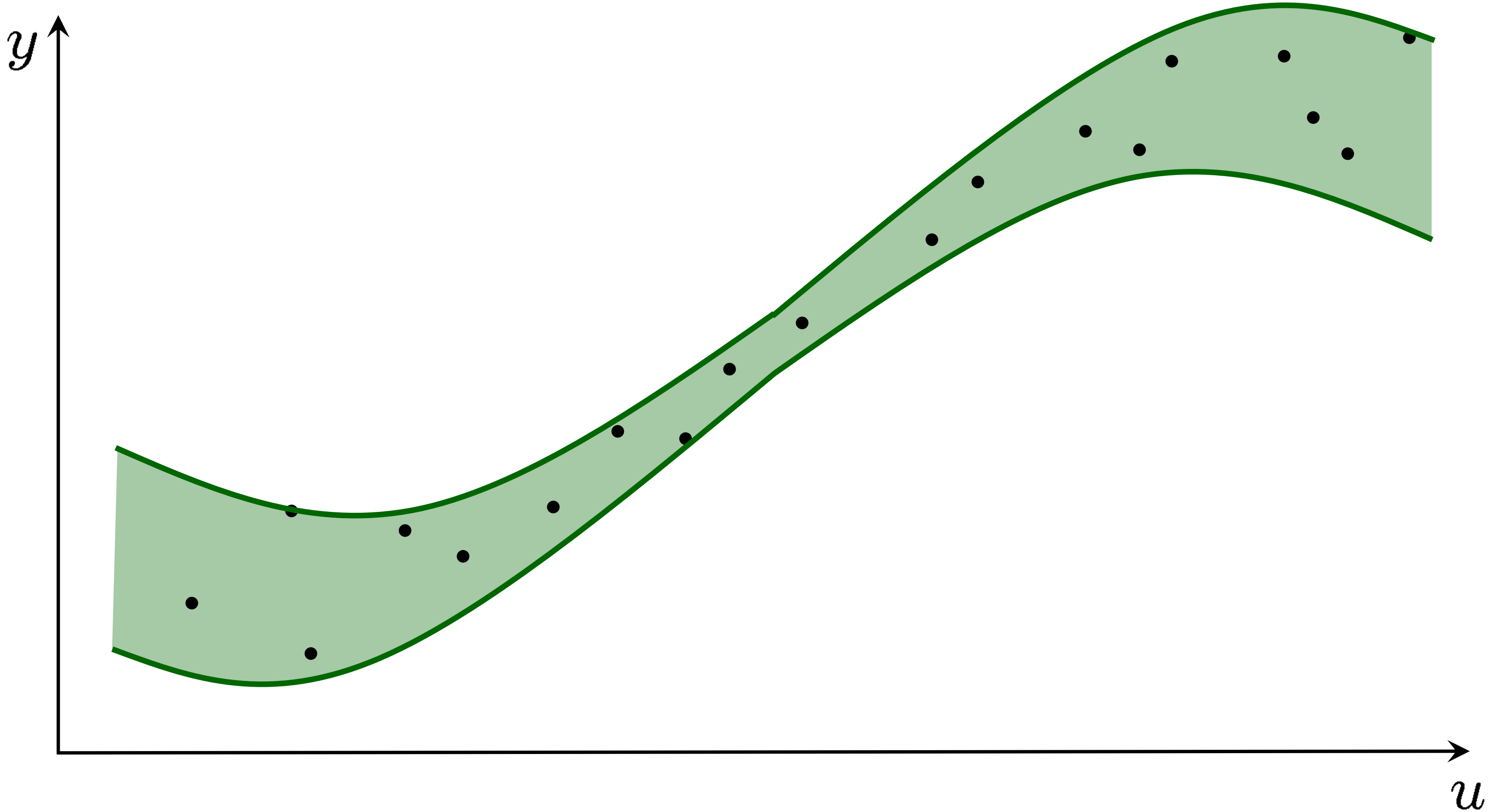
So: thickness can be inspected, while reliability has to be guaranteed by a theory valid under general assumptions.

algorithms for ...



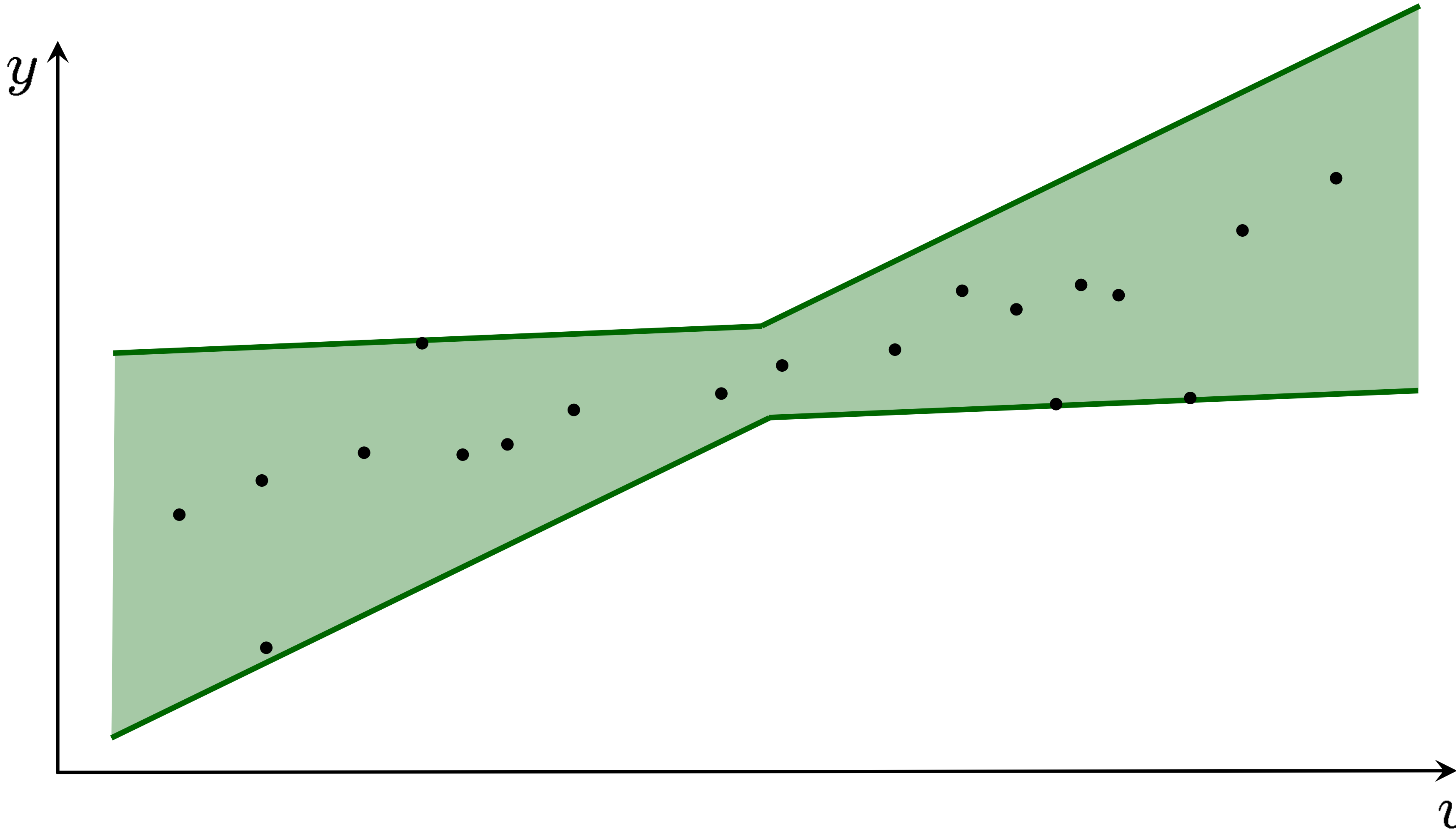
[Campi, Calafiore, Garatti, Automatica, 2009]

algorithms for …



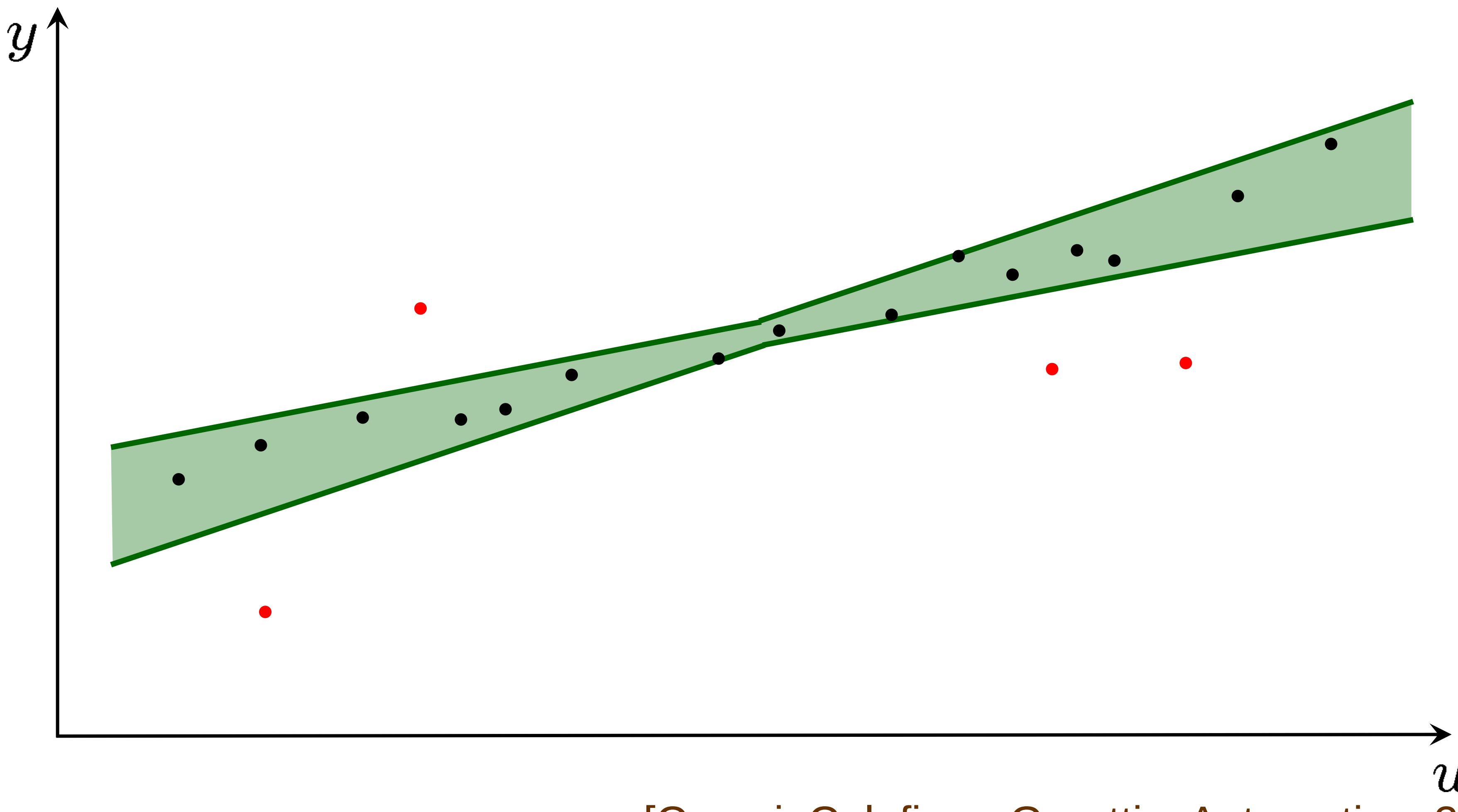
[Campi, Calafiore, Garatti, Automatica, 2009]

algorithms for ...



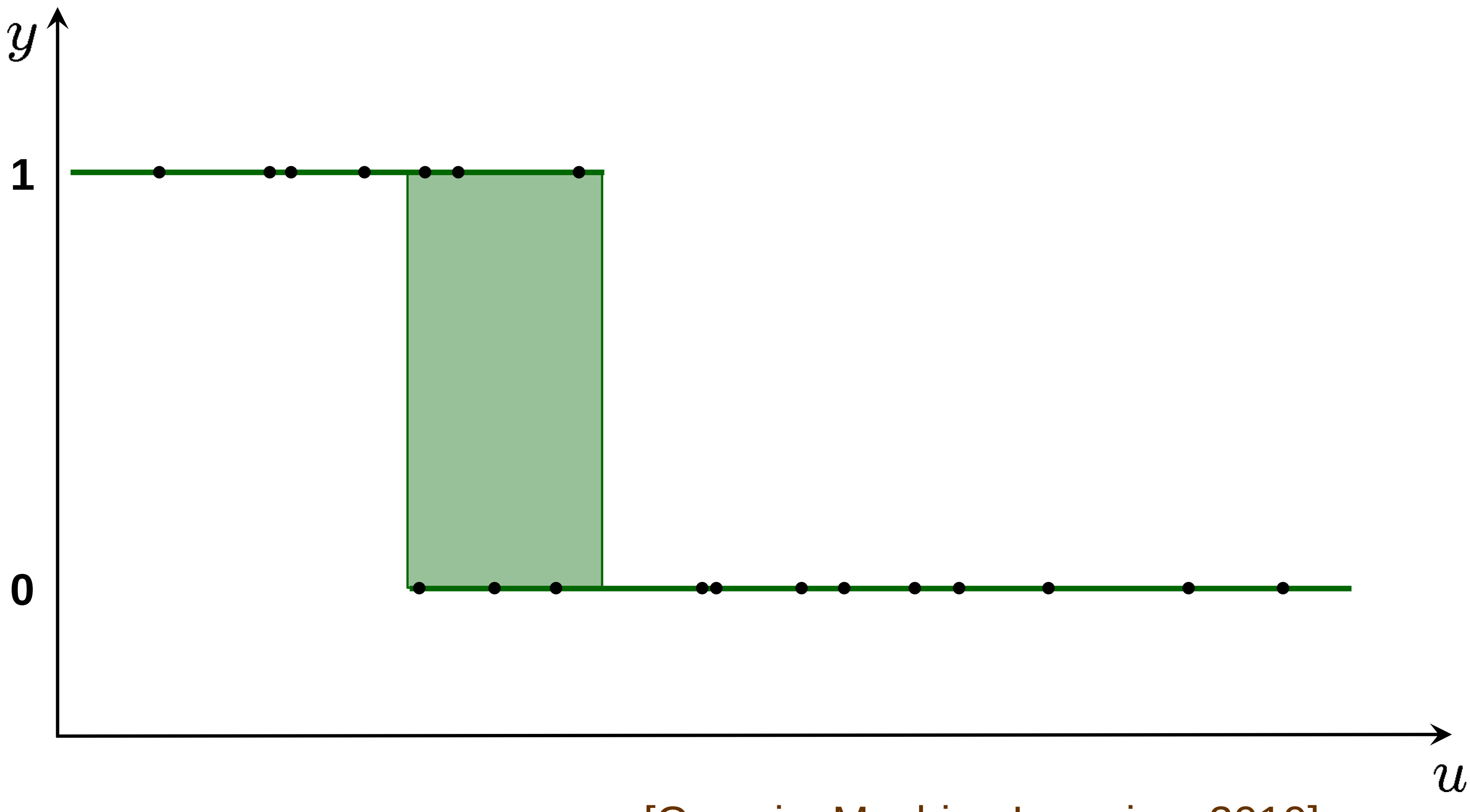
[Campi, Calafiore, Garatti, Automatica, 2009]

algorithms for ...



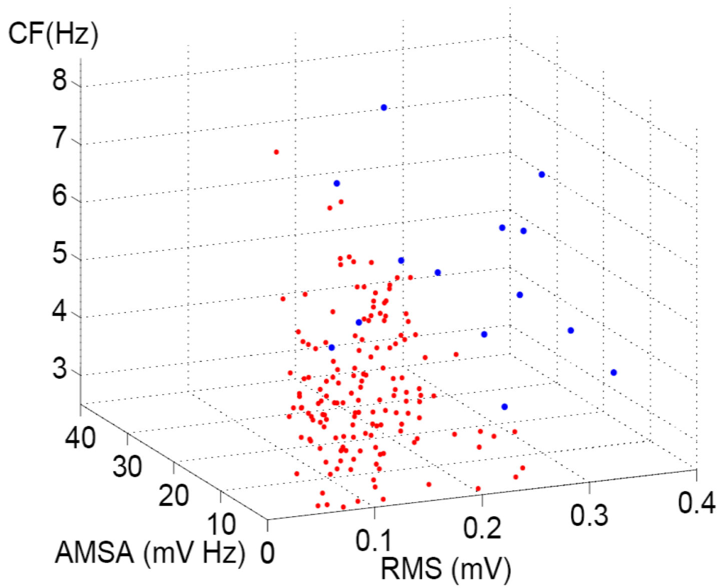
[Campi, Calafiore, Garatti, Automatica, 2009]

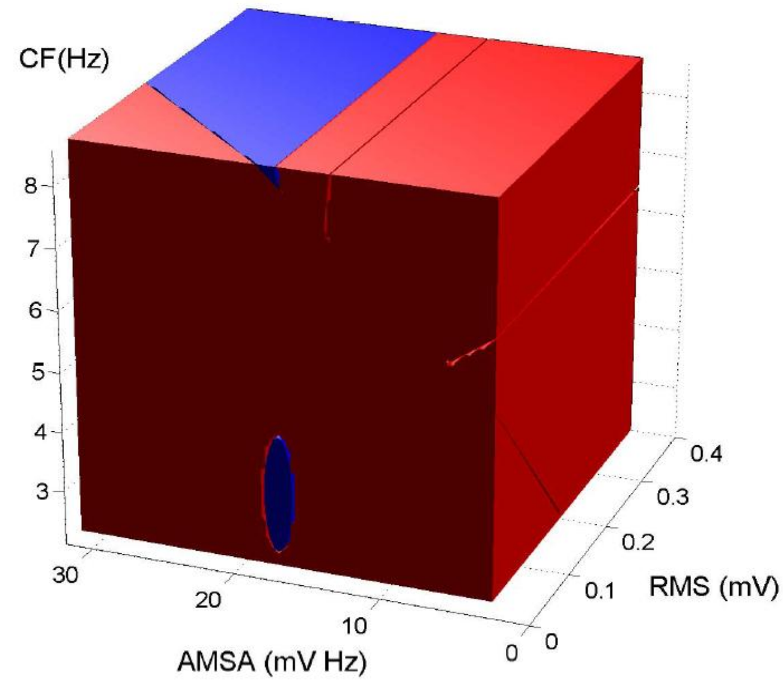
algorithms for …



[Campi, Machine Learning, 2010]

example: defibrillation





[Campi, Machine Learning, 2010]

much remains to be done in prediction:

- more general constructions
- dynamical case

A journey through:

- parameter estimation
- prediction
- filtering

A journey through:

- parameter estimation
- prediction
- filtering

beautiful opportunities for research!

Conclusions

Distribution-free results are theoretically interesting, and relevant to the practice of system identification

Conclusions

Distribution-free results are theoretically interesting, and relevant to the practice of system identification

Data contains more information than what available algorithms can exploit

Conclusions

Distribution-free results are theoretically interesting, and relevant to the practice of system identification

Data contains more information than what available algorithms can exploit

There is a need to explore the real intrinsic limitations posed by information in the data beyond the limitations enforced by the existing algorithms

Conclusions

Distribution-free results are theoretically interesting, and relevant to the practice of system identification

Data contains more information than what available algorithms can exploit

There is a need to explore the real intrinsic limitations posed by information in the data beyond the limitations enforced by the existing algorithms

It's a wonderful world to explore, which offers incredible opportunities for research

Thank you!