Statistical Methods in Particle Physics

3. Uncertainty

Heidelberg University, WS 2023/24

Klaus Reygers, Martin Völkl (lectures)
Ulrich Schmidt, (tutorials)

Deductive Reasoning

All ravens are black.

This is a raven.

Therefore it is black.

Deductive Reasoning II

All ravens are black.

This animal is not black.

Therefore it is not a raven.

Inductive Reasoning

This animal is black.

This animal is a raven.

Therefore all ravens are black.

Inductive Reasoning II

Animal 1 is a black raven.

Animal 2 is a black raven.

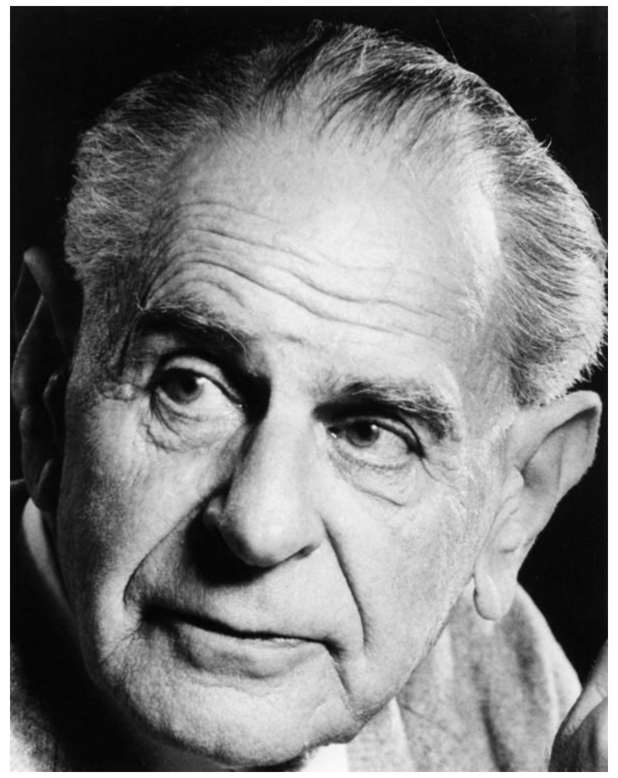
Animal 3 is a black raven.

$\vdots$

Therefore all ravens are black.

The problem of induction

- How can we draw general conclusions based on individual observations?
- If a law always held in the past does this mean it must hold in the future? (Hume)
  - Logic says, such conclusions are invalid.
  - What does it mean to do science then?



source: Wikipedia

Karl Popper (1902–1994)

Karl Popper:

- We cannot prove hypotheses, but we can disprove them (deductive)
- Make hypotheses *falsifiable*, then attempt to find observations that contradict them.
- Hypotheses, which are easy to falsify, but where repeated attempts have not done so have a higher status.

- Occams razor: Everything else being equal, prefer simpler hypotheses.
- It is somewhat more difficult to say what falsification means in the presence of measurement uncertainties

The scientific method

Karl Popper:

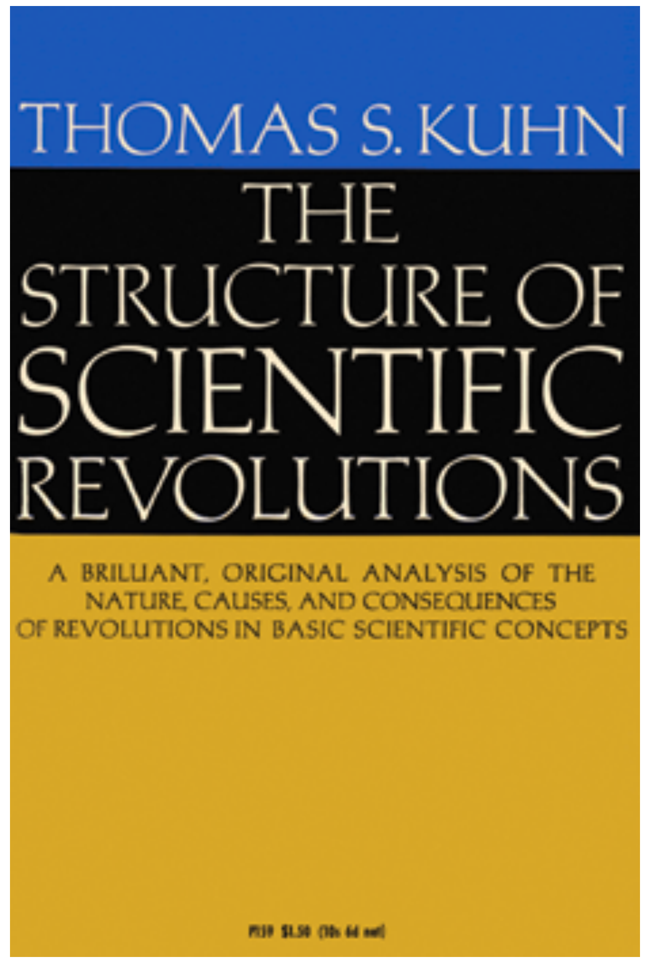
- Find falsifiable hypotheses and try to disprove them

Thomas Kuhn:

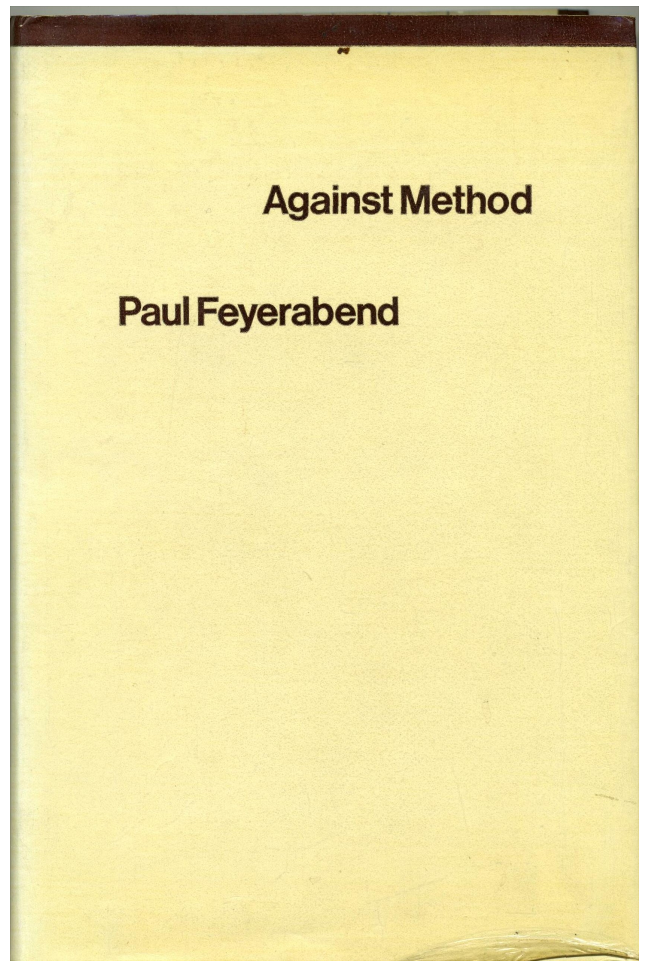
- Actual science works differently.
- Periods of continuity, then paradigm shift, not necessarily from better description of data

Paul Feyerabend:

- Having one single scientific method would inhibit scientific progress
- “Anything goes” - Scientists should pursue any course that seems interesting to them



THOMAS S. KUHN
THE STRUCTURE OF SCIENTIFIC REVOLUTIONS
A BRILLIANT, ORIGINAL ANALYSIS OF THE NATURE, CAUSES, AND CONSEQUENCES OF REVOLUTIONS IN BASIC SCIENTIFIC CONCEPTS



Against Method
Paul Feyerabend

- If so: How to differentiate science and pseudoscience?
- Science has made definite progress in the past

The scientific method II

- No generally agreed upon definition of “scientific method”
- General properties of scientific inquiry:
  - Governed by rational arguments
  - Includes hypotheses that should be self consistent and are usually formulated in mathematical language
  - Models/Theories should make testable predictions
  - Observations can test or motivate hypothesis
  - Cannot arbitrarily discard evidence
  - Honesty in reasoning
  - Discussion, peer review, criticism, adaptation of ideas
  - Progress happens by some kind of consensus

- Not very satisfying. But no solution in this lecture.

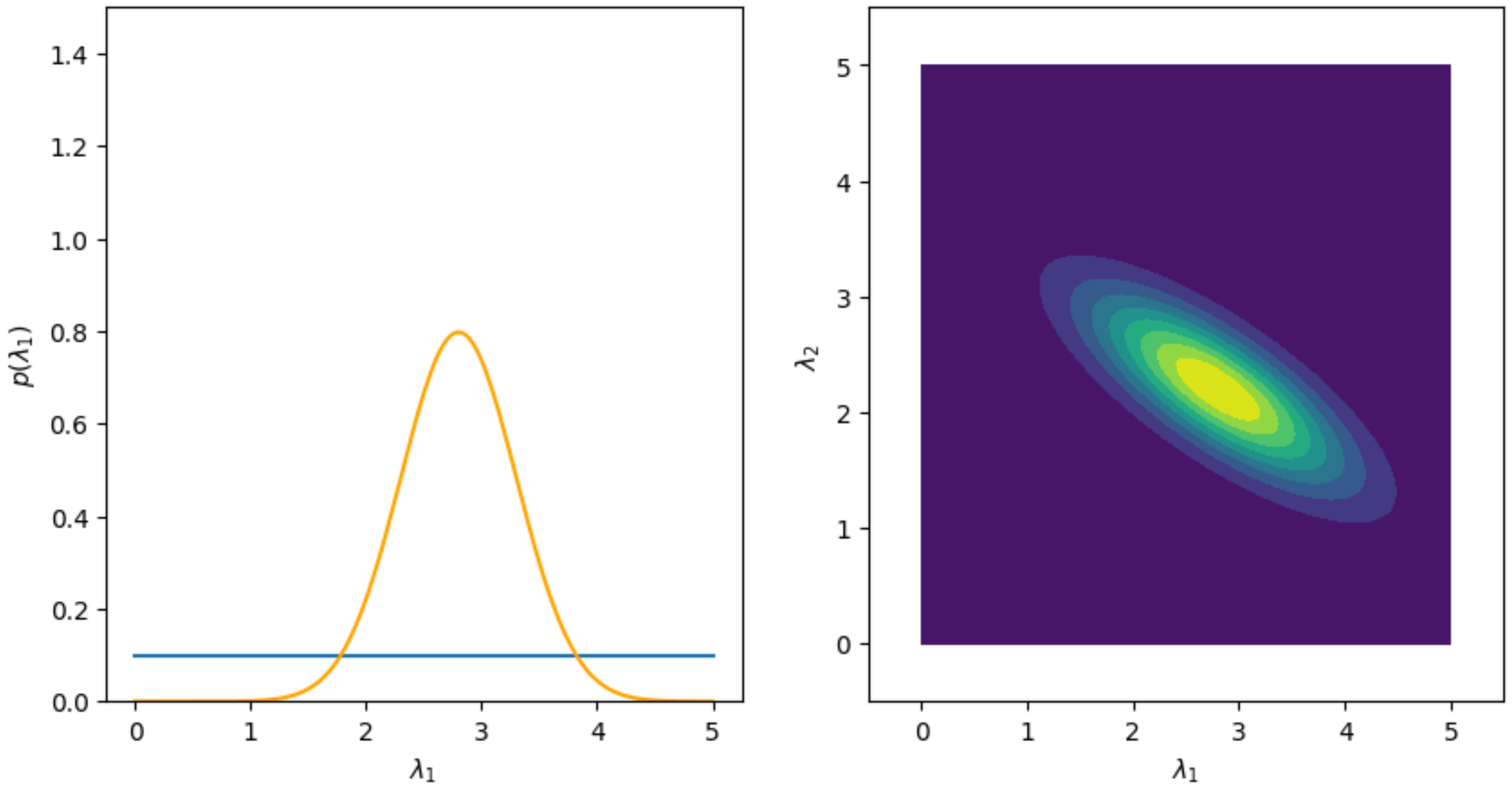
e.g. There is No Scientific Method, Lee Smolin

Occam’s razor

- Prefer simpler hypotheses to more complex ones
- "Entities are not to be multiplied without necessity”
- "Whenever possible, substitute constructions out of known entities for inferences to unknown entities.”- Bertrand Russell
- Often useful for guidance, but vague

Bayesian inference contains a form of Occam’s razor:

- Consider two hypotheses with 1 and 2 unknown parameters, each with total probability 0.5
- Prior probability spread out across more dimensions in second case -> likelihood removes more of it
- Marginal for left hypothesis will be much higher afterwards



Inductive Reasoning III

Animal 1 is a black raven.

Animal 2 is a black raven.

Animal 3 is a black raven.

⋮

Therefore it becomes more plausible that all ravens are black.

Probability as an extension of logic

Basic idea:

- Induction is not valid logical reasoning
- It it what we do in science
- Therefore scientific reasoning is not based on (Aristotelian) logic

- Understanding by Laplace, Bayes, Jeffreys
- Later also Cox, Jaynes and others

"Probability theory is nothing but common sense reduced to calculation." - Pierre-Simon Laplace

Cox Axioms

Axiom 1:

Degrees of plausibility are represented by real numbers

Meaning:

- P(A) is the plausibility of statement A
- P(A|B) is the plausibility of A assuming B
- P(A,B|C) are the plausibility of A AND B assuming C
- P(A)>P(B) means that A is more plausible than B

Cox Axioms

Axiom 2:

Qualitative correspondence with common sense

- If P(A|C')>P(A|C), then also P($\bar{A}$|C')<P($\bar{A}$|C)

Cox Axioms

Axiom 3:

All reasoning must happen consistently

- Logically equivalent statements are equally plausible
- If there are several ways to reason out a conclusion, they must all lead to the same result
- Information may not be ignored; all reasoning is nonideoligical

Cox Axioms

1. Degrees of plausibility are represented by real numbers
2. Qualitative correspondence with common sense
3. All reasoning must happen consistently

Jaynes, 2003

This results in a class of equivalent solutions. One of them has the properties:

- A *true* statement is represented by $P(A) = 1$, a *false* statement is represented by $P(A) = 0$
- $P(A) + P(\bar{A}) = 1$ (sum rule)
- $P(A, B) = P(A \,|\, B) \cdot P(B)$ (product rule)

In this approach, probability theory extends logic. Inductive reasoning becomes Bayesian inference.

We do not prove or disprove hypotheses, but change our state of knowledge about them.

Does a case of the hypothesis support the hypothesis?

(Hempels paradox)

3.1 Error Propagation

Evaluating Estimator Performance

Consistency:

- Does the estimate converge to the true value?

$$\lim_{n\to\infty} \hat{\theta} = \theta$$

Bias:

- Does the average of many measurements converge towards the true value? Otherwise: bias b

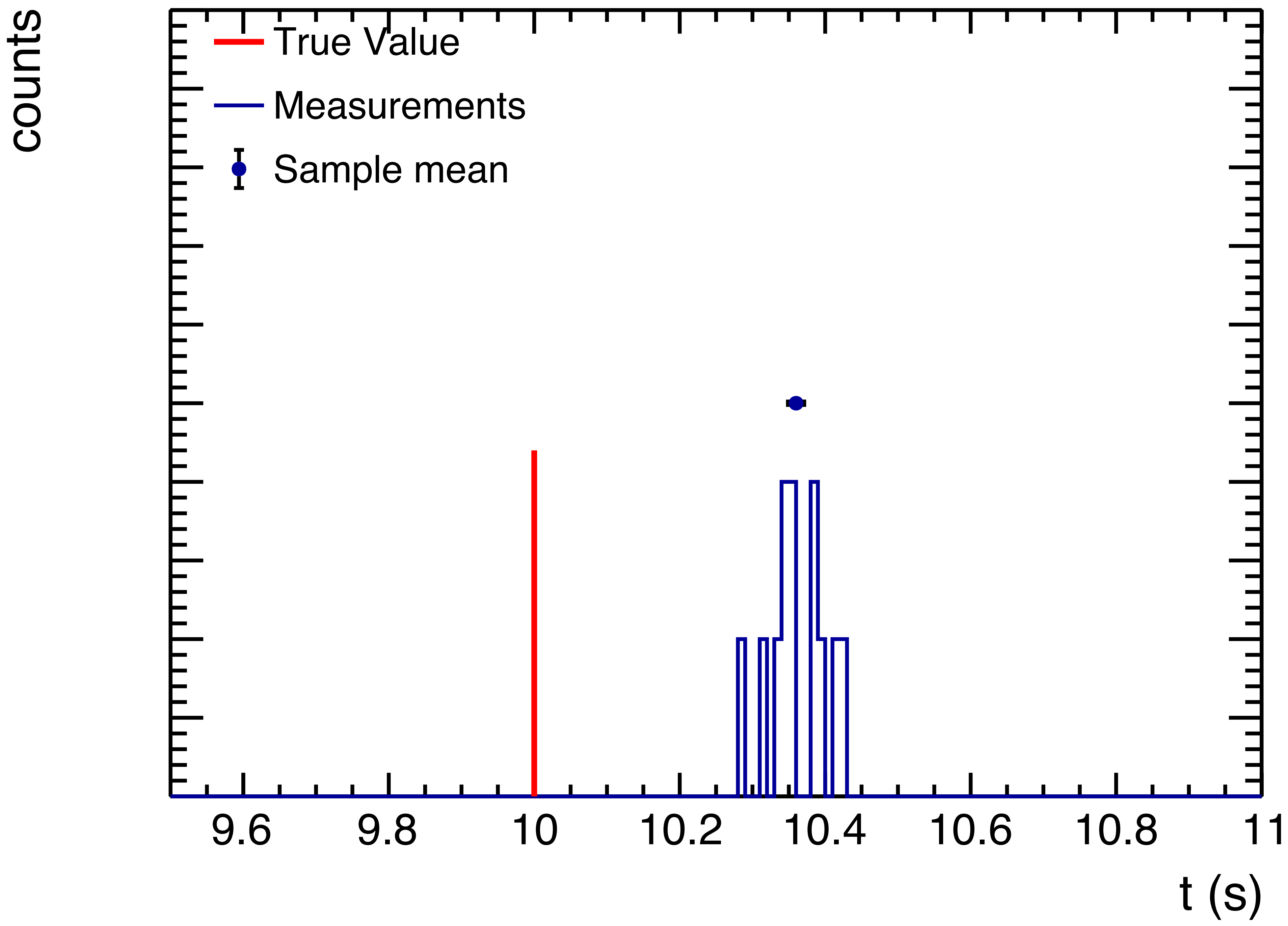
$$E[\hat{\theta}] = \theta$$

Efficiency:

- How small is the uncertainty for a given amount of data and how fast does it decrease with n?

Robustness:

- Does the estimator still work if we are slightly wrong about the assumptions of the data (e.g. in the presence of rare outliers)?



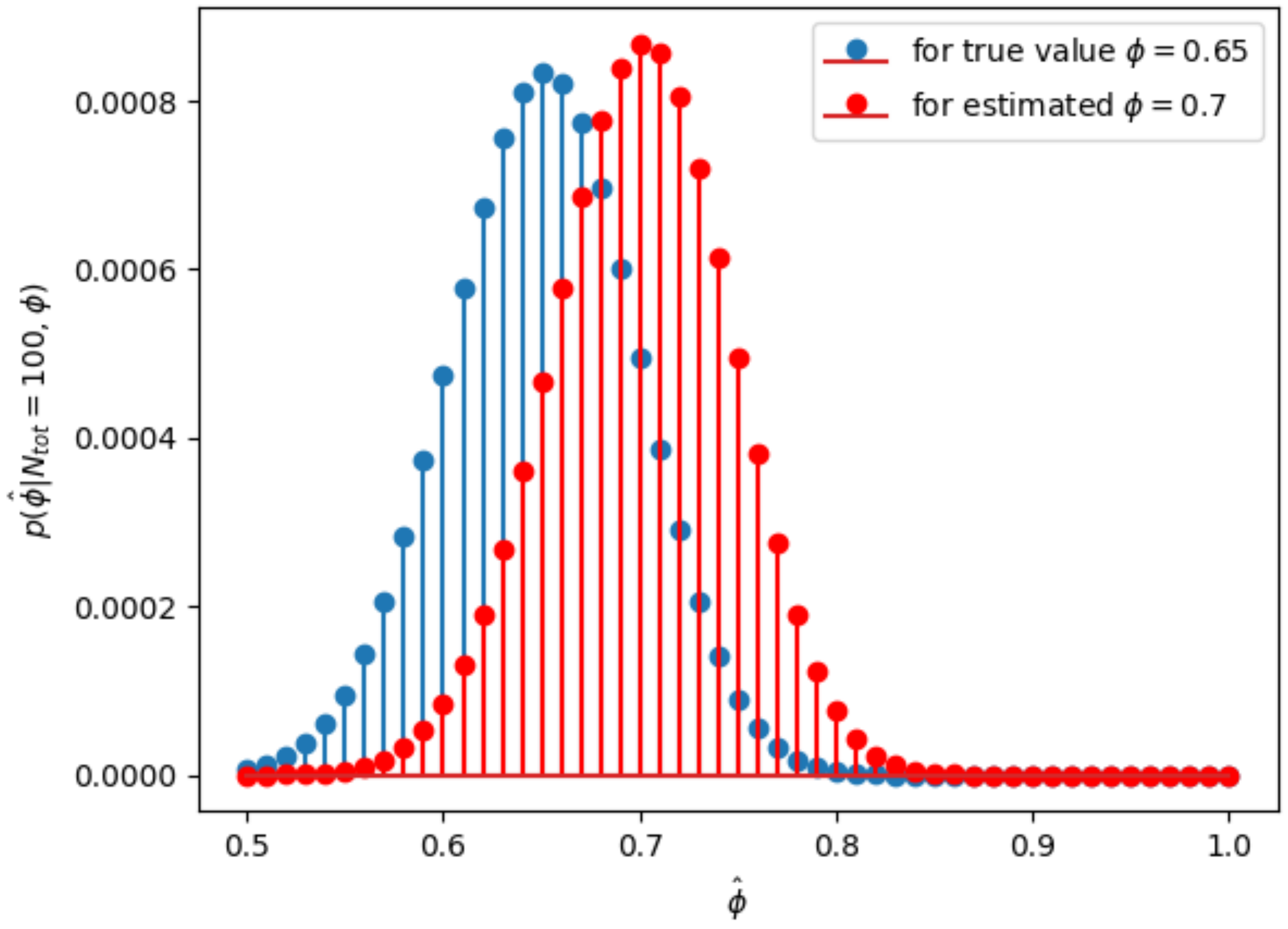
The measured time on a stopwatch may be affected by the reaction time of the experimenter, giving a bias

Minimum Variance Estimators

- Variance and bias of an estimator are often a tradeoff
  - E.g. just guessing a single fixed value has 0 variance, but definitely a bias.
- One optimisation: Allow only unbiased estimators, then try to minimise variance.
- The result is called the *minimum variance unbiased estimator* (MVUE)
- Can be hard to find
- Often quite useful; sometime a lower variance can be a good tradeoff for some bias
- Subtracting the (estimated) bias of an estimator can transform it into an (approximately) unbiased one

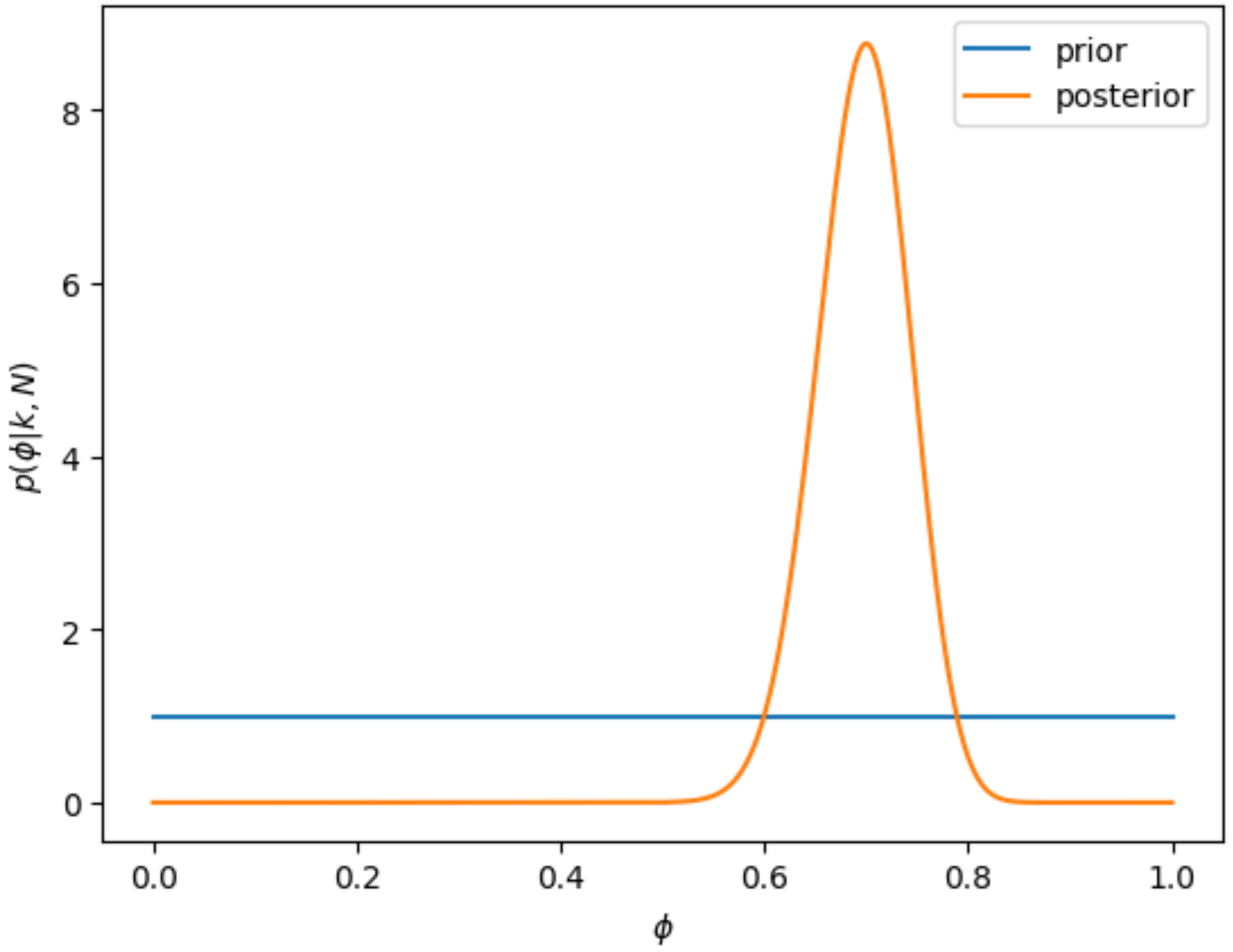
Reminder: Frequentist and Bayesian Uncertainties

Frequentist



- Uncertainty: Standard deviation of estimator fluctuations when repeating experiment a number of times

Bayesian



- Uncertainty: Standard deviation of posterior probability distribution of parameter

Ways to quote uncertainties

$$t = (34.5 \pm 0.7)\ 10^{-3}\ \mathrm{s}$$
$$t = 34.5\ 10^{-3}\ \mathrm{s} \pm 2\,\%$$
$$x = 10.3^{+0.7}_{-0.3}$$
$$m_e = (0.510\,999\,06 \pm 0.000\,000\,15)\ \mathrm{MeV}/c^2$$
$$m_e = 0.510\,999\,06\ (15)\ \mathrm{MeV}/c^2$$
$$m_e = 9.109\,389\,7\,10^{-31}\,\mathrm{kg}\ \pm 0.3\,ppm$$

An uncertainty σ represents some kind of probability distribution (often a Gaussian, if not stated otherwise)

If no further information is given the interval x ± σ corresponds to a one standard deviation. In the Gaussian approximation this contains a probability of 68% ("1σ errors")

Error propagation is simple and difficult

Assume we want to go from variable a to variable $b = f(a)$

Frequentist

- We start with an unbiased estimator for a. Now find an unbiased estimator for b. Estimate the variance of the new estimator.

But:

- Very difficult in practice.
- Would rather just use the estimator $\hat{b} = f(\hat{a})$

Bayesian

- We start with a probability distribution $p_a(a)$ for a. Now transform this into a probability distribution for b with $p_b(b) = p_a(f^{-1}(b))\,|J_{b\to a}|$

But:

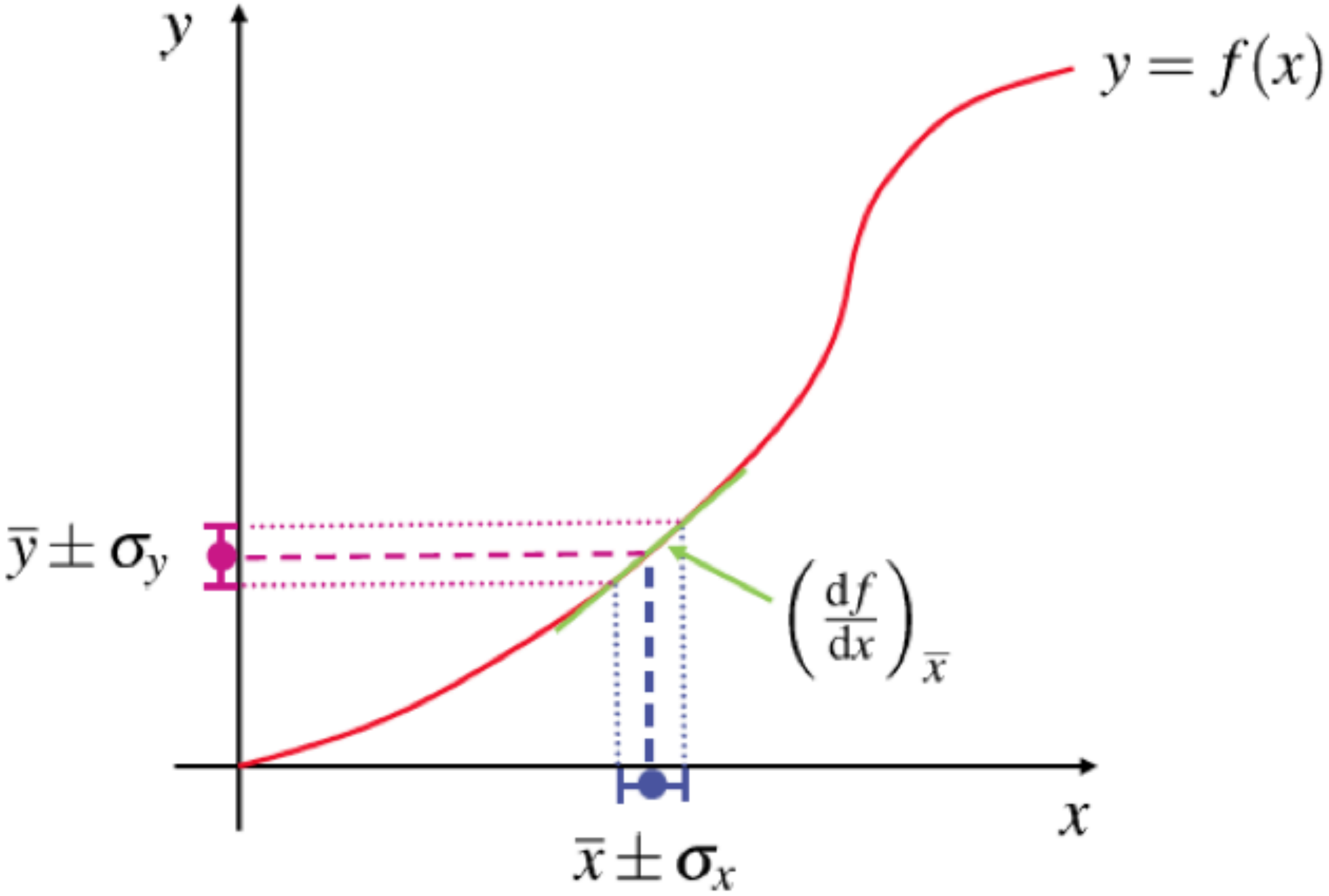
- Can be very difficult to calculate
- The inverse may not have a simple analytical form

Easy conceptually, difficult to calculate → Usually much easier to approximate the result

Linear error propagation: Sometimes applicable …

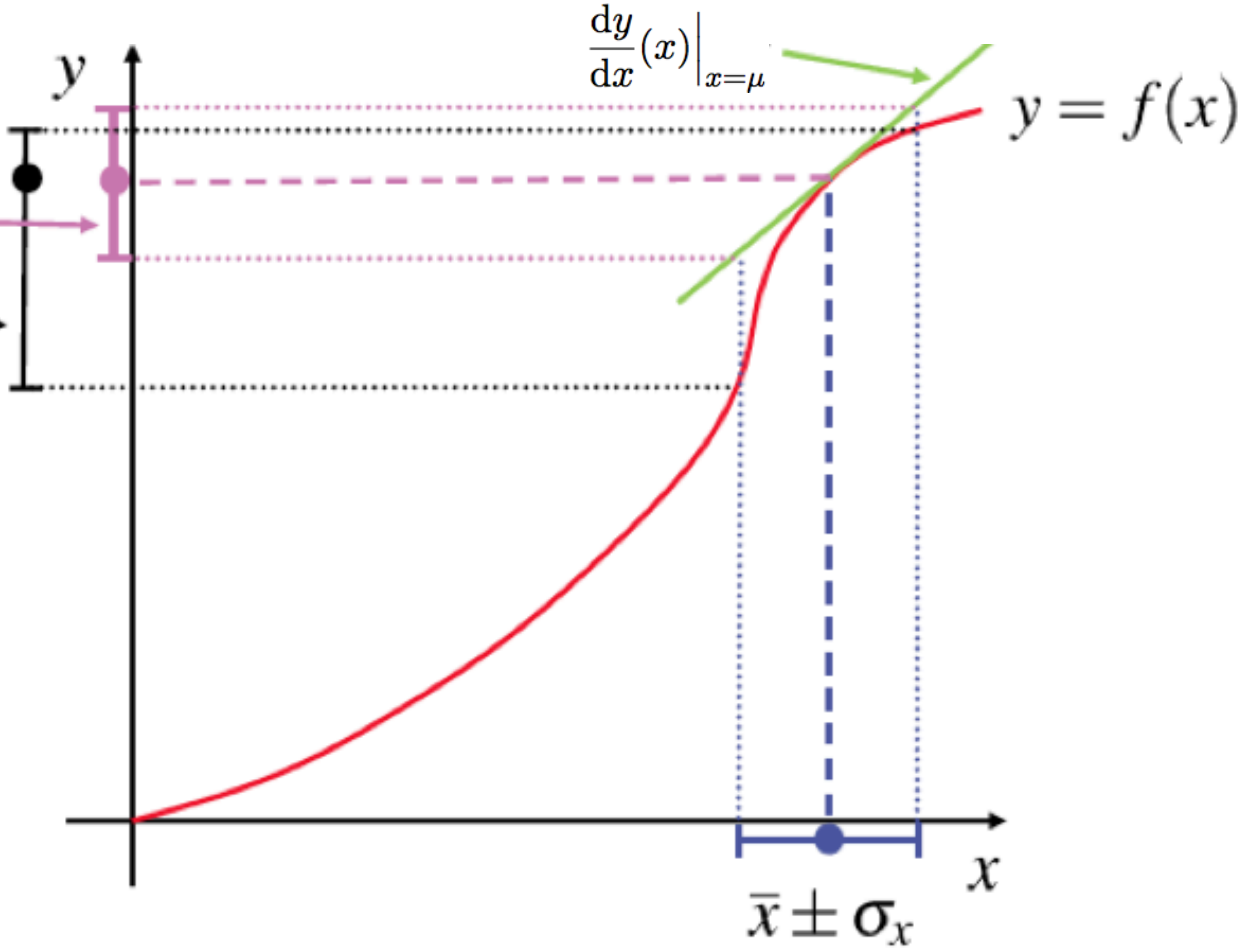
- We know that the transformation $x \to x + \Delta x$ does not change the speed of the variable
- We know that scaling $x \to \alpha x$ changes the variance by a factor α^2
- This is true for both the distribution of the estimator and the posterior
- If f is a linear function, then $\sigma_b = |\alpha|\sigma_a$
- If the function is close to linear in the places where it is of interest to us, the transformation can be approximated

  setting $\sigma_b = \left|\frac{\mathrm{d}f}{\mathrm{d}a}\right| \sigma_a$



Function sufficiently linear within ±σ: linear error propagation applicable

Linear error propagation: Sometimes not applicable …



Calculated

Left and right side transformed differently

In this situation linear error propagation is not applicable

Linear error propagation (general case)

Consider a measurement of values x_i and their covariances:

$$\vec{x} = (x_1, x_2, \ldots, x_n) \qquad V_{ij} = \mathrm{cov}[x_i, x_j]$$

Let y be a function of the x_i: $\quad y = f(\vec{x})$

What is the variance of y?

Approach: Taylor expansion of y around $\vec{\mu}$ where $\quad \mu_i = E[x_i]$

In practice we estimate μ_i by measured value x_i

$$V[y] \equiv \sigma_y^2 = E[y^2] - E[y]^2$$

Linear error propagation formula

Taylor expansion:
$$y(\vec{x}) \approx y(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)$$

E[*y*] is easy:
$$E[y] \approx y(\vec{\mu}) \qquad \text{as} \quad E[x_i - \mu_i] = 0$$

E[*y*²]:
$$E[y^2(\vec{x})] \approx y^2(\vec{\mu}) + 2y(\vec{\mu}) \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} E[x_i - \mu_i]$$
$$+ E\left[\left(\sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)\right)\left(\sum_{j=1}^{n} \left[\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j)\right)\right]$$
$$= y^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Thus:
$$\boxed{\sigma_y^2 = \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}}$$

Matrix notation

Let vector A be given by $\quad \vec{A} = \vec{\nabla} y, \quad \text{i.e., } A_j = \left(\frac{\partial y}{\partial x_j} \right)_{\vec{x}=\vec{\mu}}$

Then: $\quad \sigma_y^2 = \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij} \quad = \quad A^\mathsf{T} V A$

Example: $\quad y = \frac{x_1}{x_2}, \quad A = \begin{pmatrix} 1/x_2 \\ -x_1/x_2^2 \end{pmatrix}$

$$\sigma_y^2 = \left(\frac{1}{x_2}, -\frac{x_1}{x_2^2} \right) \begin{pmatrix} \sigma_1^2 & \mathrm{cov}[x_1, x_2] \\ \mathrm{cov}[x_1, x_2] & \sigma_2^2 \end{pmatrix} \begin{pmatrix} \frac{1}{x_2} \\ -\frac{x_1}{x_2^2} \end{pmatrix}$$

$$= \left(\frac{1}{x_2}, -\frac{x_1}{x_2^2} \right) \begin{pmatrix} \frac{\sigma_1^2}{x_2} - \frac{x_1}{x_2^2} \mathrm{cov}[x_1, x_2] \\ \frac{1}{x_2} \mathrm{cov}[x_1, x_2] - \frac{x_1}{x_2^2} \sigma_2^2 \end{pmatrix} = \frac{1}{x_2^2} \sigma_1^2 + \frac{x_1^2}{x_2^4} \sigma_2^2 - 2 \frac{x_1}{x_2^3} \mathrm{cov}[x_1, x_2]$$

$$\rightarrow \quad \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} - 2 \frac{\mathrm{cov}[x_1, x_2]}{x_1 x_2} = \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} - 2 \frac{\rho \sigma_1 \sigma_2}{x_1 x_2}$$

Linear error proportion: Examples

$$y = ax \quad \rightarrow \quad \sigma_y^2 = a^2\sigma_x^2 \qquad \text{i.e. } \sigma_y = |a|\sigma_x$$

$$y = x^n \quad \rightarrow \quad \frac{\sigma_y^2}{y^2} = n^2\frac{\sigma_x^2}{x^2} \qquad \text{i.e. } \frac{\sigma_y}{y} = |n|\frac{\sigma_x}{x}$$

$$y = x_1 + x_2 \quad \rightarrow \quad \sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\mathrm{cov}[x_1, x_2]$$

Adding variables means the *absolute* errors add in quadrature.

$$y = x_1 - x_2 \quad \rightarrow \quad \sigma_y^2 = \sigma_1^2 + \sigma_2^2 - 2\mathrm{cov}[x_1, x_2]$$

$$y = x_1 x_2 \quad \rightarrow \quad \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{\mathrm{cov}[x_1, x_2]}{x_1 x_2}$$

Multiplying variables means the *relative* errors add in quadrature.

Sanity checks:

Average of fully correlated measurements:

$$y = \frac{1}{2}(x_1 + x_2)\,, \; \sigma_1 = \sigma_2 \equiv \sigma, \; \rho = 1 \quad \rightsquigarrow \quad \sigma_y = \sigma$$

Difference of fully correlated measurements:

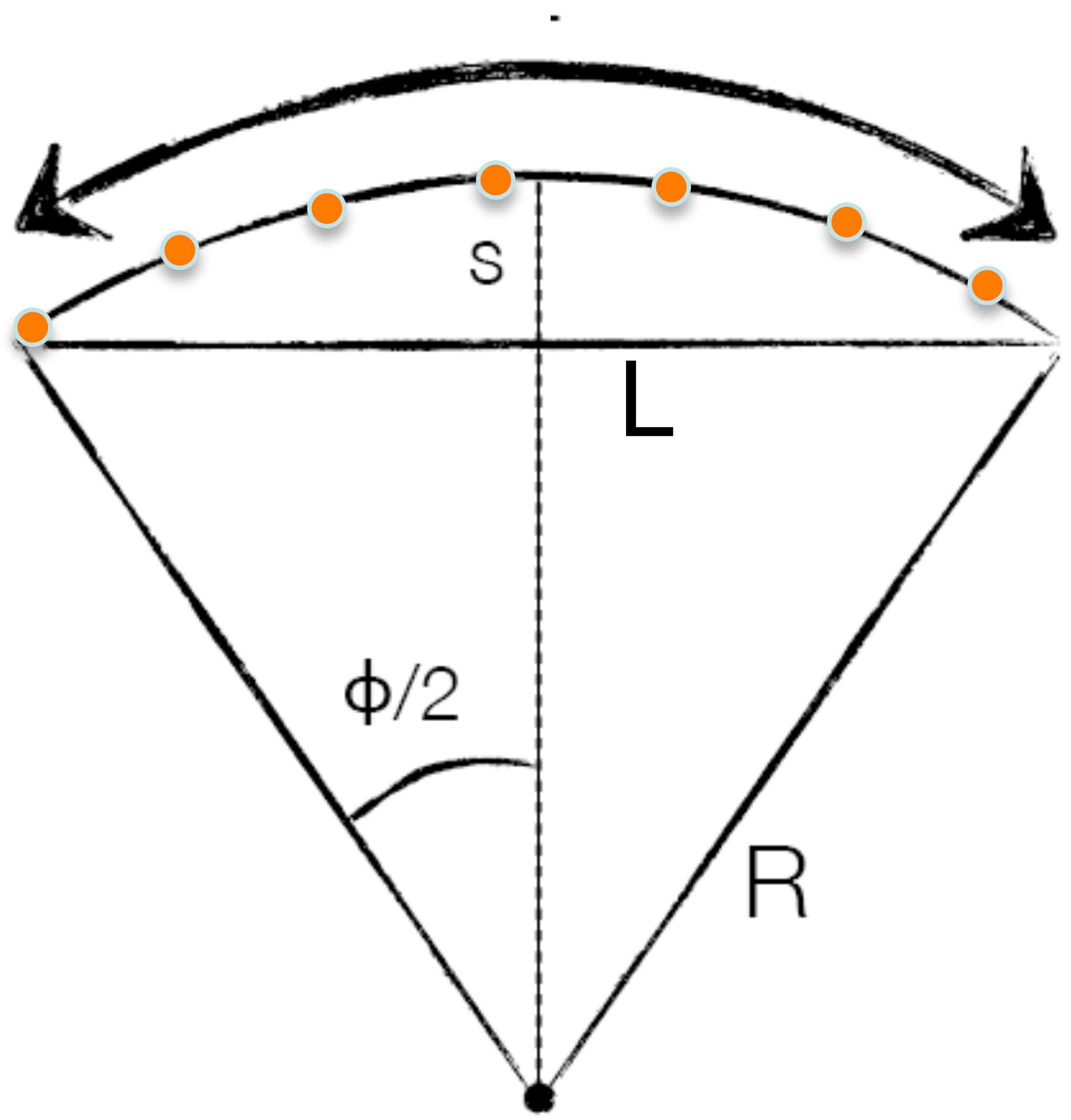
$$y = x_1 - x_2, \; \sigma_1 = \sigma_2 \equiv \sigma, \; \rho = 1$$

$$\rightsquigarrow \quad \sigma_y^2 = 2\sigma^2 - 2\sigma^2 = 0$$

Concrete example: Momentum resolution in tracking

Charged particle moving in constant magnetic field:

$$p_T/\mathrm{GeV} = 0.3 \times B/\mathrm{Tesla} \times R/\mathrm{m}$$



Measurements of space points yields Gaussian uncertainty for sagitta *s* which is related to p_T as

$$R = \frac{L^2}{8s}, \quad p_T = 0.3B\frac{L^2}{8s}$$

Momentum resolution: $$\frac{\sigma_{p_T}}{p_T} = \frac{\sigma_s}{s} = \frac{8p_T}{0.3BL^2}\sigma_s$$

Important features:

- Relative momentum uncertainty proportional to momentum
- Relative uncertainty prop. to uncertainty of coordinate measurement

Example:
ATLAS nominal resolution

$$\left(\frac{\sigma_{p_T}}{p_T}\right)^2 = \underbrace{0.001^2}_{\text{multiple scattering}} + \underbrace{(0.0005p_T)^2}_{\text{track uncertainty}}$$

Linear error propagation for uncorrelated measurements

Special case: the x_i are uncorrelated, i.e., $V_{ij} = \delta_{ij}\sigma_i^2$

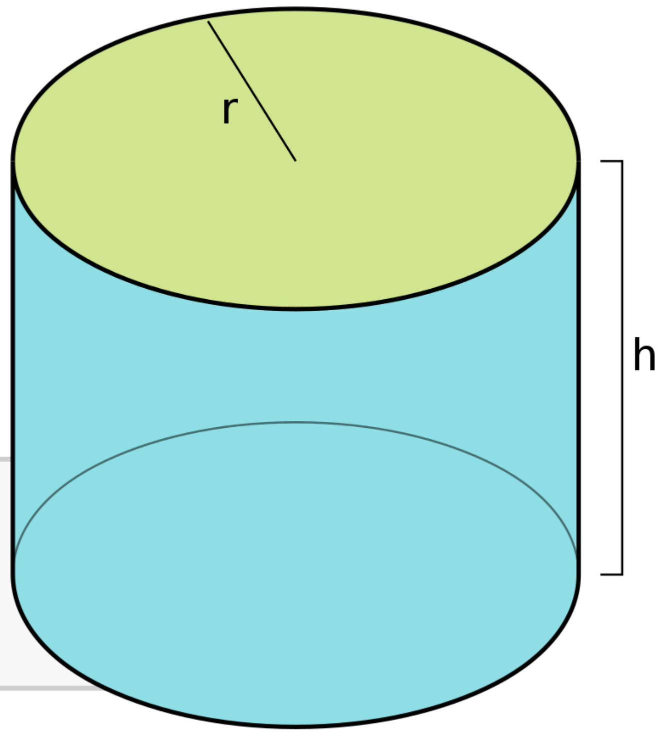
$$\sigma_y^2 = \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]^2_{\vec{x}=\vec{\mu}} \sigma_i^2$$

These formulas are exact only for linear functions.

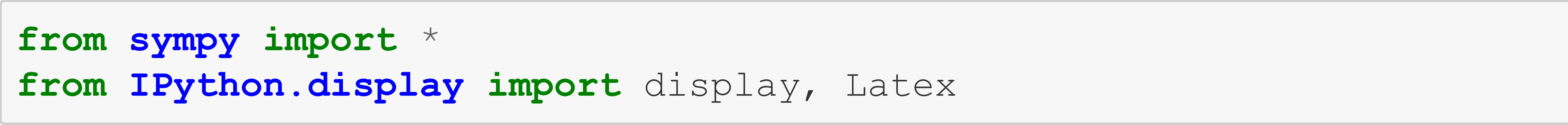
Approximation breaks down if function is nonlinear over a region comparable in size to the σ_i.

Example of Gaussian error propagation: Volume of a cylinder

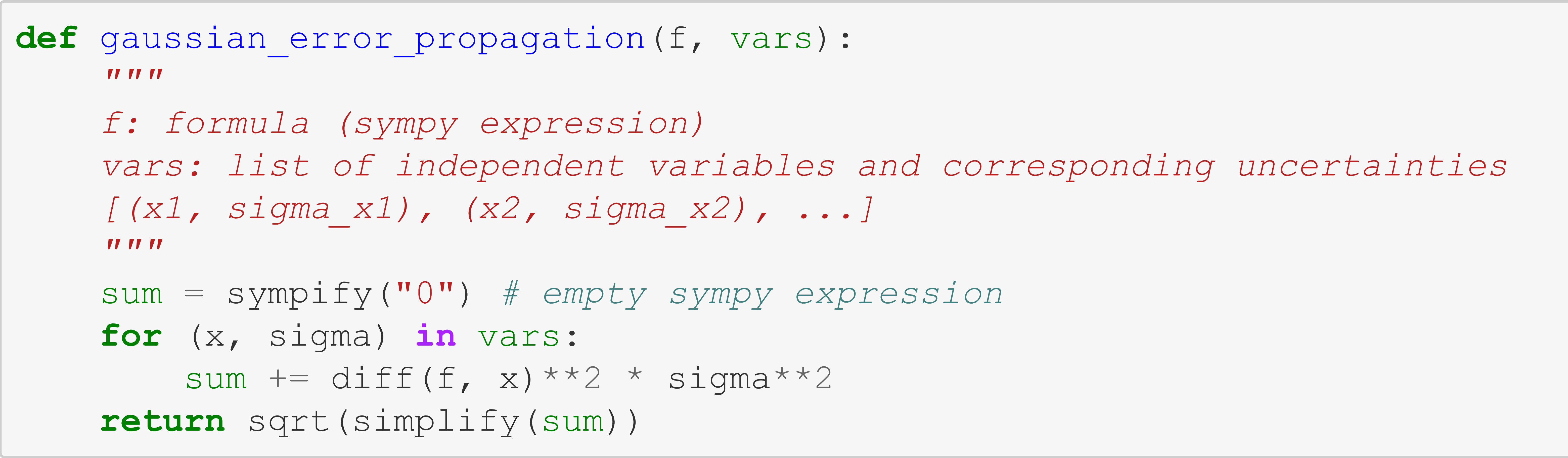
[gaussian_error_propagation.ipynb]



[wikipedia]

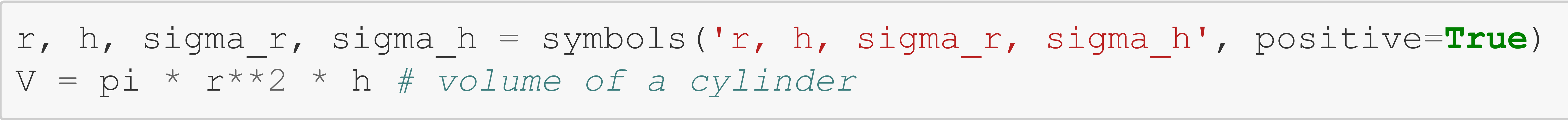


```
from sympy import *
from IPython.display import display, Latex
```

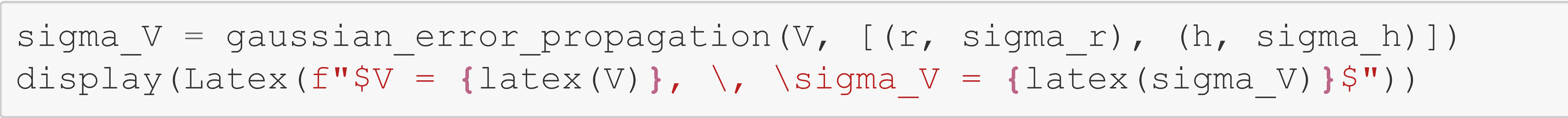


```
def gaussian_error_propagation(f, vars):
    """
    f: formula (sympy expression)
    vars: list of independent variables and corresponding uncertainties
    [(x1, sigma_x1), (x2, sigma_x2), ...]
    """
    sum = sympify("0") # empty sympy expression
    for (x, sigma) in vars:
        sum += diff(f, x)**2 * sigma**2
    return sqrt(simplify(sum))
```

Show usage for a simple example: Volume of a cylinder with radius r and height h:



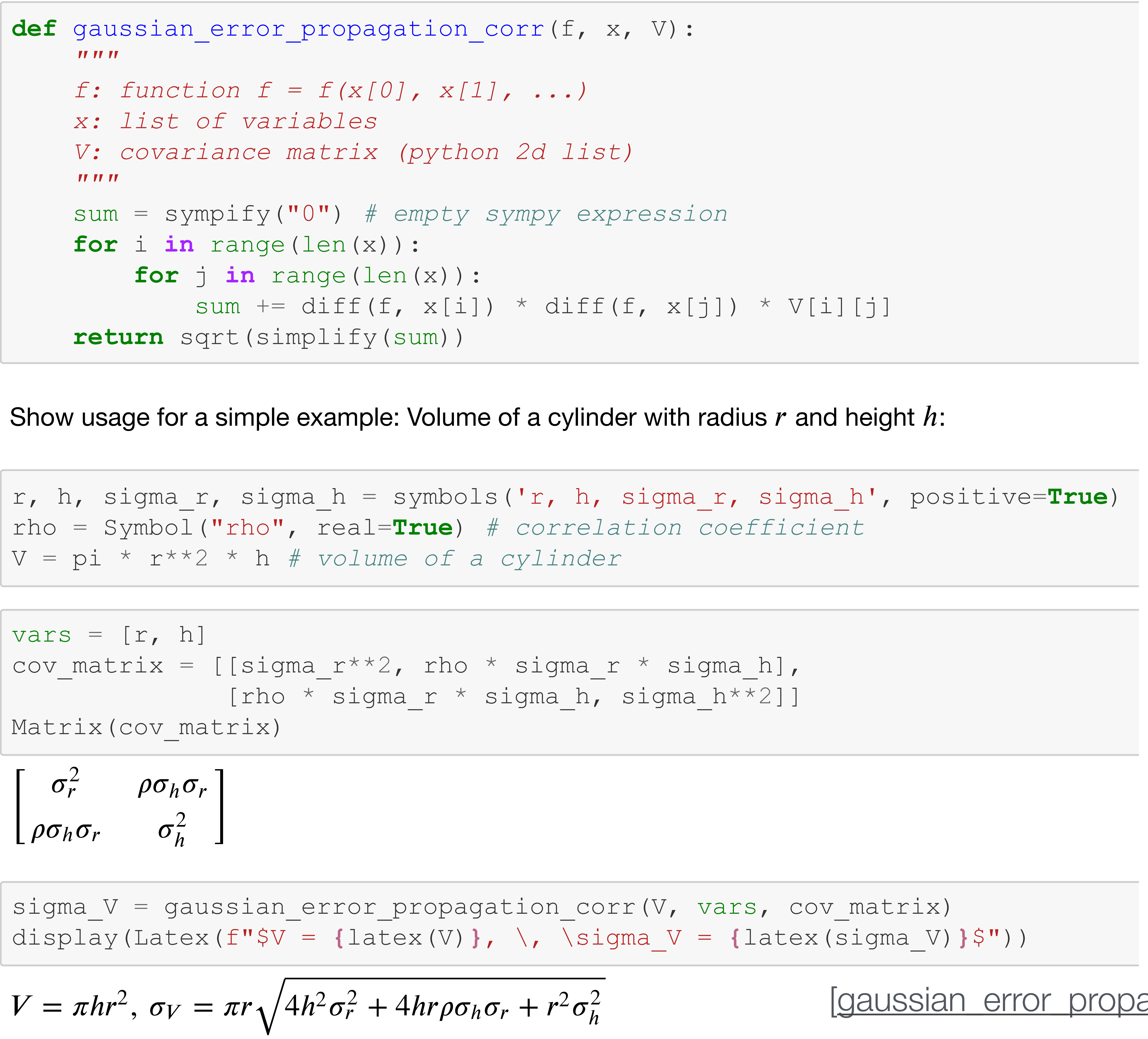
```
r, h, sigma_r, sigma_h = symbols('r, h, sigma_r, sigma_h', positive=True)
V = pi * r**2 * h # volume of a cylinder
```



```
sigma_V = gaussian_error_propagation(V, [(r, sigma_r), (h, sigma_h)])
display(Latex(f"$V = {latex(V)}, \, \sigma_V = {latex(sigma_V)}$"))
```

$$V = \pi h r^2, \, \sigma_V = \pi r \sqrt{4h^2\sigma_r^2 + r^2\sigma_h^2}$$

Example of Gaussian error propagation: Volume of a cylinder (now for correlated *r* and *h*)



```python
def gaussian_error_propagation_corr(f, x, V):
    """
    f: function f = f(x[0], x[1], ...)
    x: list of variables
    V: covariance matrix (python 2d list)
    """
    sum = sympify("0") # empty sympy expression
    for i in range(len(x)):
        for j in range(len(x)):
            sum += diff(f, x[i]) * diff(f, x[j]) * V[i][j]
    return sqrt(simplify(sum))
```

Show usage for a simple example: Volume of a cylinder with radius r and height h:

```python
r, h, sigma_r, sigma_h = symbols('r, h, sigma_r, sigma_h', positive=True)
rho = Symbol("rho", real=True) # correlation coefficient
V = pi * r**2 * h # volume of a cylinder
```

```python
vars = [r, h]
cov_matrix = [[sigma_r**2, rho * sigma_r * sigma_h],
              [rho * sigma_r * sigma_h, sigma_h**2]]
Matrix(cov_matrix)
```

$$\begin{bmatrix} \sigma_r^2 & \rho\sigma_h\sigma_r \\ \rho\sigma_h\sigma_r & \sigma_h^2 \end{bmatrix}$$

```python
sigma_V = gaussian_error_propagation_corr(V, vars, cov_matrix)
display(Latex(f"$V = {latex(V)}, \, \sigma_V = {latex(sigma_V)}$"))
```

$$V = \pi h r^2, \ \sigma_V = \pi r\sqrt{4h^2\sigma_r^2 + 4hr\rho\sigma_h\sigma_r + r^2\sigma_h^2}$$

$$r = 3\,\mathrm{cm}, \sigma_r = 0.1\,\mathrm{cm}$$
$$h = 5\,\mathrm{cm}, \sigma_h = 0.1\,\mathrm{cm}$$
$$V = \pi r^2 h = 141.4\,\mathrm{cm}^3$$

Uncertainty of the cylinder volume *V* depends on the correlation coefficient ρ:

ρ	σ$_V$
–1	6.6 cm³
0	9.8 cm³
1	12.3 cm³

[gaussian_error_propagation_correlated_variables.ipynb]

Linear error propagation: Generalization from $\mathbb{R}^n \rightarrow \mathbb{R}$ to $\mathbb{R}^n \rightarrow \mathbb{R}^m$

Generalization: Consider set of *m* functions:

$$\vec{y}(\vec{x}) = (y_1(\vec{x}), y_2(\vec{x}), \ldots, y_m(\vec{x}))$$

Then:

$$\mathrm{cov}[y_k, y_l] \equiv U_{kl} \approx \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i}\frac{\partial y_l}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}$$

In matrix notation:

$$U = A\,V\,A^\mathsf{T} \qquad A_{ij} = \left[\frac{\partial y_i}{\partial x_j}\right]_{\vec{x}=\vec{\mu}}$$

Reduction of the standard deviation for repeated independent measurements

Consider the average of *n* independent observation x_i:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Expectation values and variance of the measurements:

$$E[x_i] = \mu_i \qquad V[x_i] = \sigma^2$$

Standard deviation of the mean:

$$V[\bar{x}] = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 = \frac{1}{n} \sigma^2 \qquad \to \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Standard deviation of the mean decreases as $1/\sqrt{n}$

Example: Photon energy measurements

The energy resolution of a γ-ray detector used to investigate a decaying nuclear isotope is 50 keV.

- If only one photon is detected the energy of the decay is known to 50 keV
- 100 collected decays: energy of the decay known to 5 keV
- To reach 1 keV one needs to observe 2500 decays

Averaging uncorrelated measurements

Consider two uncorrelated measurements: $x_1 \pm \sigma_1, \; x_2 \pm \sigma_2$

Linear combination:

$$y = w_1 x_1 + w_2 x_2 \qquad \sigma_y^2 = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2$$

Now choose the weights such that σ_y^2 is minimal (under the condition $w_1 + w_2 = 1$):

$$\frac{\partial}{\partial w_i} \sigma_y^2 = 0 \quad \rightarrow \quad w_i = \frac{1/\sigma_i^2}{1/\sigma_1^2 + 1/\sigma_2^2}$$

And for the uncertainty of y we obtain (linear error propagation):

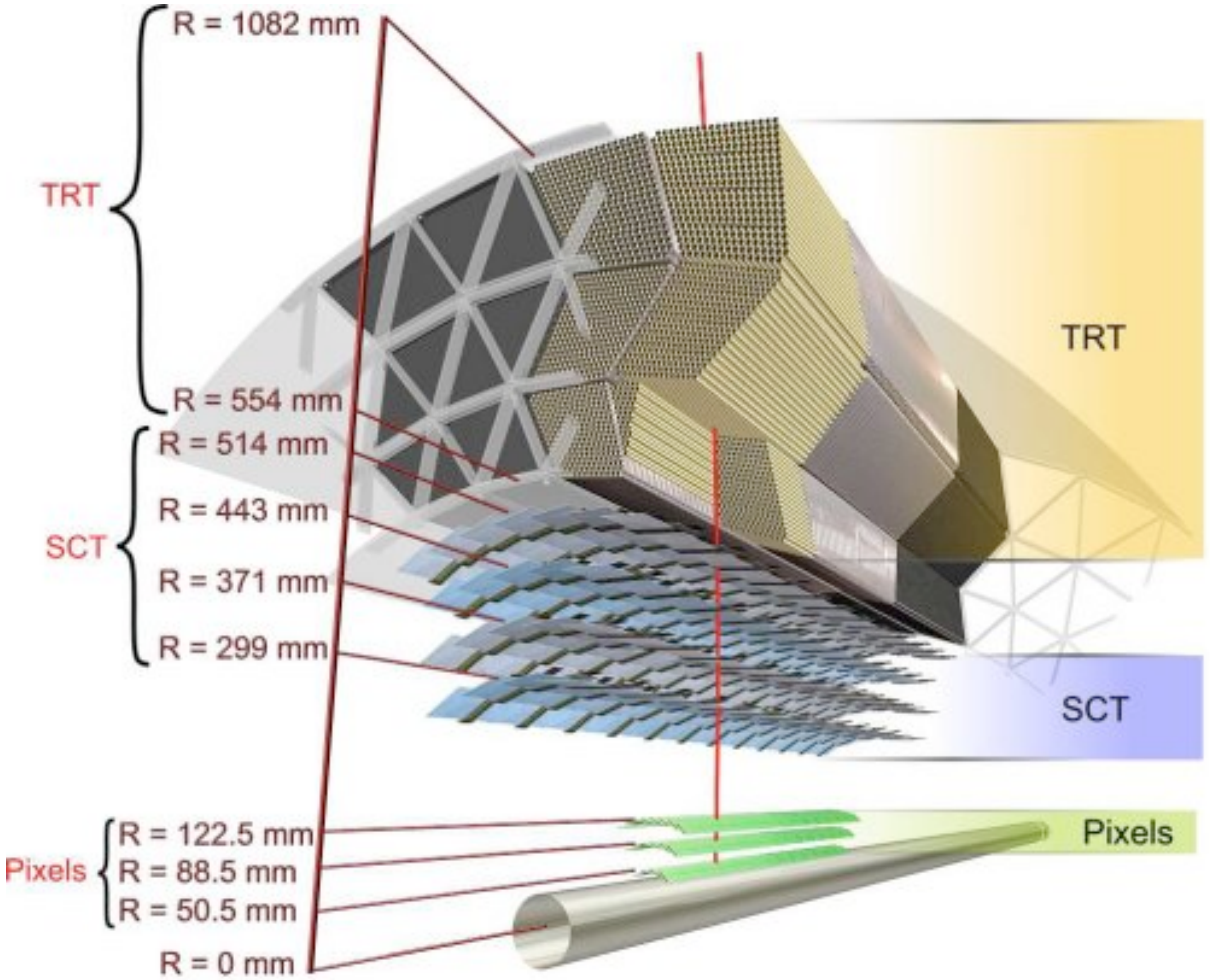
$$\frac{1}{\sigma_y^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$$

In general, for n uncorrelated measurements:

$$y = \sum_{i=1}^{n} w_i x_i, \qquad w_i = \frac{1/\sigma_i^2}{\sum_{j=1}^{n} 1/\sigma_j^2}, \qquad \frac{1}{\sigma_y^2} = \sum_{j=1}^{n} \frac{1}{\sigma_j^2}$$

Example: Averaging uncorrelated measurements

p_T of a particle in three subsystems of the ATLAS detector:



detector	p_T (GeV)
pixel detector	20 ± 2
semiconductor tracker	21 ± 1
transition radiation tracker	22 ± 4

Weighted average:

(20.86 ± 0.87) GeV

$$p_T = \frac{\frac{20\,\text{GeV}}{4\,\text{GeV}^2} + \frac{21\,\text{GeV}}{1\,\text{GeV}^2} + \frac{22\,\text{GeV}}{16\,\text{GeV}^2}}{\frac{1}{4\,\text{GeV}^2} + \frac{1}{1\,\text{GeV}^2} + \frac{1}{16\,\text{GeV}^2}} = 20.86\,\text{GeV}$$

$$\sigma_{p_T} = \left[\frac{1}{4\,\text{GeV}^2} + \frac{1}{1\,\text{GeV}^2} + \frac{1}{16\,\text{GeV}^2}\right]^{-1/2} = 0.87\,\text{GeV}$$

Weighted average from Bayesian approach

Consider two measurements x_1 and x_2 with Gaussian uncertainties σ_1 and σ_2. In a Bayesian approach the probability distribution for the true value μ is given by

$$p(\mu) \propto L(x_1, x_2|\mu)\pi(\mu)$$

Assuming a flat prior $\pi(\mu) \equiv 1$ and independence of the two measurements one obtains

$$\begin{aligned} p(\mu) &\propto L(x_1|\mu)L(x_2|\mu) \\ &= G(x_1; \mu, \sigma_1)G(x_2; \mu, \sigma_2) \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{(\mu - x_1)^2}{\sigma_1^2} + \frac{(\mu - x_2)^2}{\sigma_2^2}\right)\right] \end{aligned}$$

In one case the resulting Gaussian comes from a product, in the other from a convolution - both give Gaussians again

The product of the two Gaussians gives a Gaussian with mean

$$\mu = w_1 x_1 + w_2 x_2 \quad \text{where } w_i = \frac{1/\sigma_i^2}{1/\sigma_1^2 + 1/\sigma_2^2}$$

and standard deviation

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$$

→ same result as before

Monte Carlo error propagation

Example:
Ratio of two Gaussian distributed quantities

$$x = 5 \pm 1$$

$$y = 5 \pm 1$$

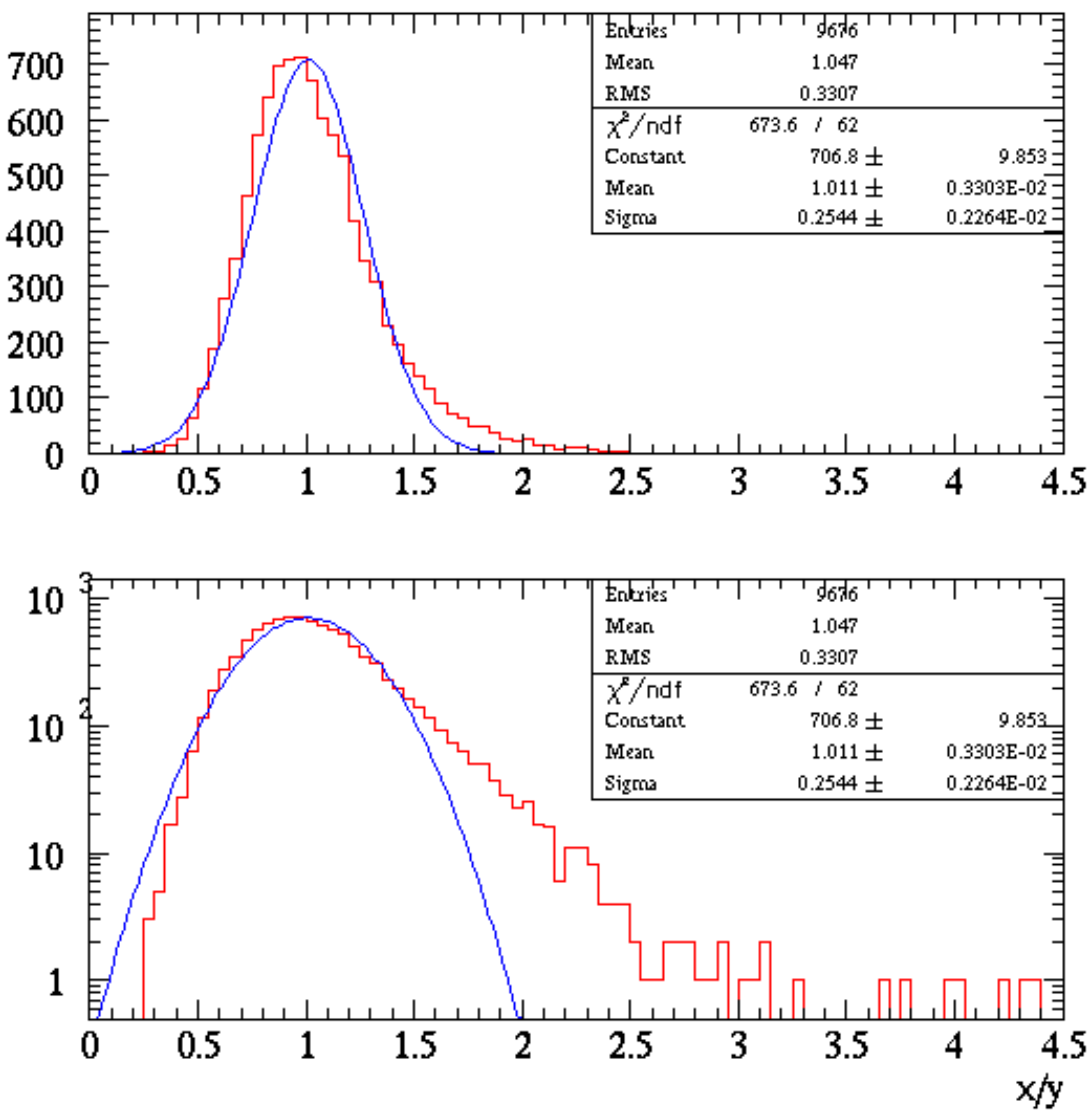
Approach: draw values for *x* and *y* many times and fill histogram with ratios

Standard linear error prop.:

$$R = 1 \pm 0.28$$

Mean and rms of histogram:

$$R = 1.05 \pm 0.33$$



Rule of thumb: ratio of two Gaussians will be approximately Gaussian if fractional uncertainty is dominated by numerator, and denominator cannot be small compared to numerator

Classification of Uncertainties

Classification of Errors/Uncertainties

1. Mistakes

- The experimenter did something wrong

2. Statistical Uncertainties

- Uncertainties in the result due to fluctuations which can be addressed by statistical methods.

3. Systematic Uncertainties

- Uncertainties from any other sources

Nothing deep here. Distinction useful order thoughts

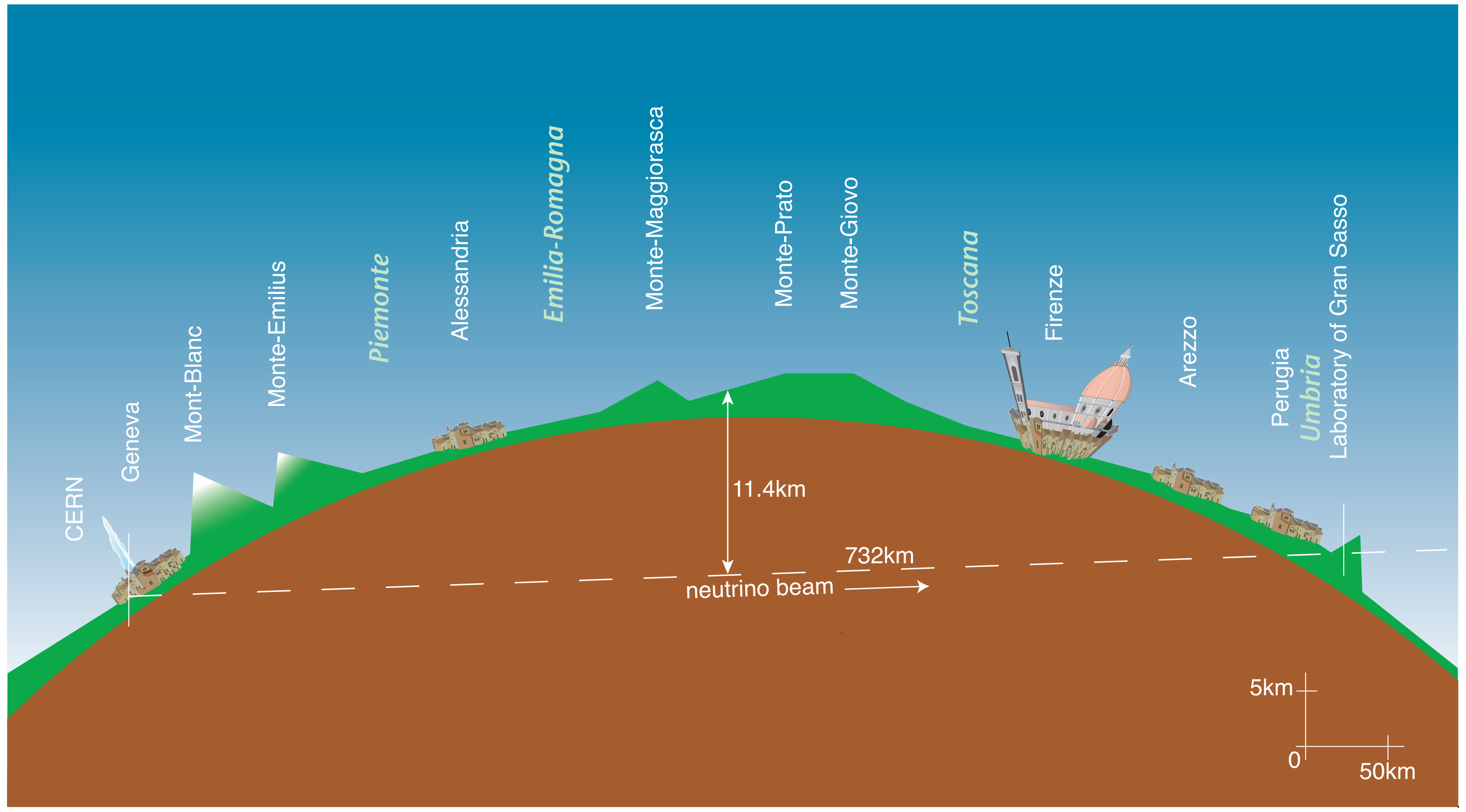
Example for experimental mistakes

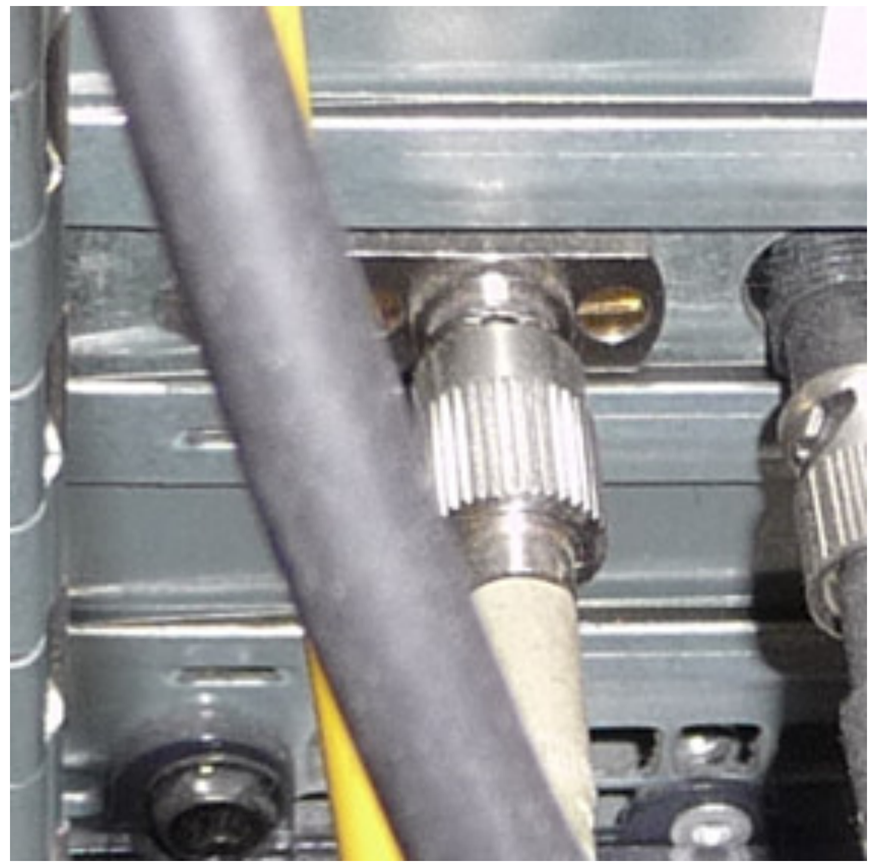
1. Mistakes

- The experimenter did something wrong

Superluminal neutrinos - OPERA experiment

- OPERA measures neutrinos produced at CERN
- September 2011: OPERA announced measurement of faster than light neutrinos
- 6 standard deviations of difference
- Turned out to be due to loose fiberoptic cable

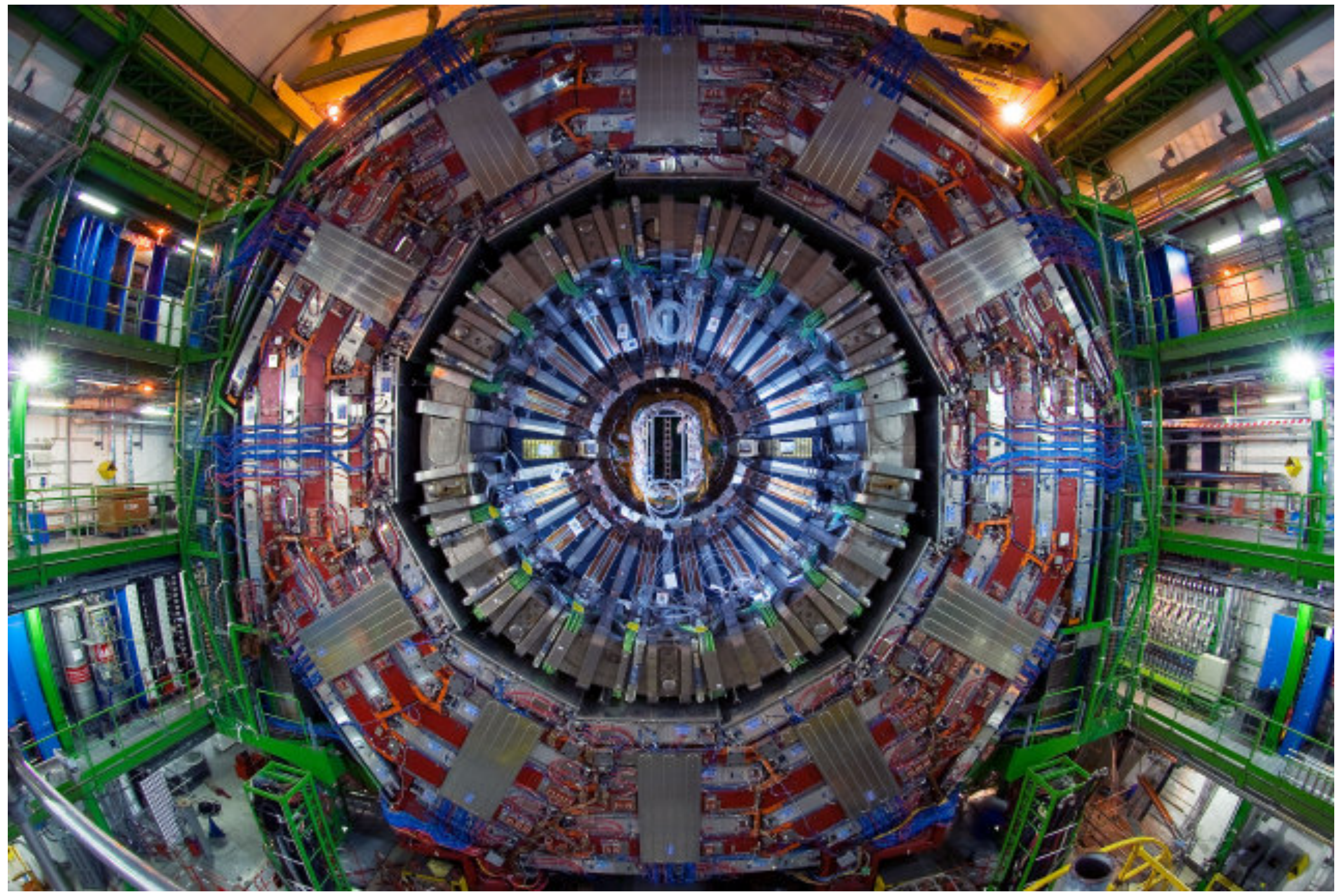


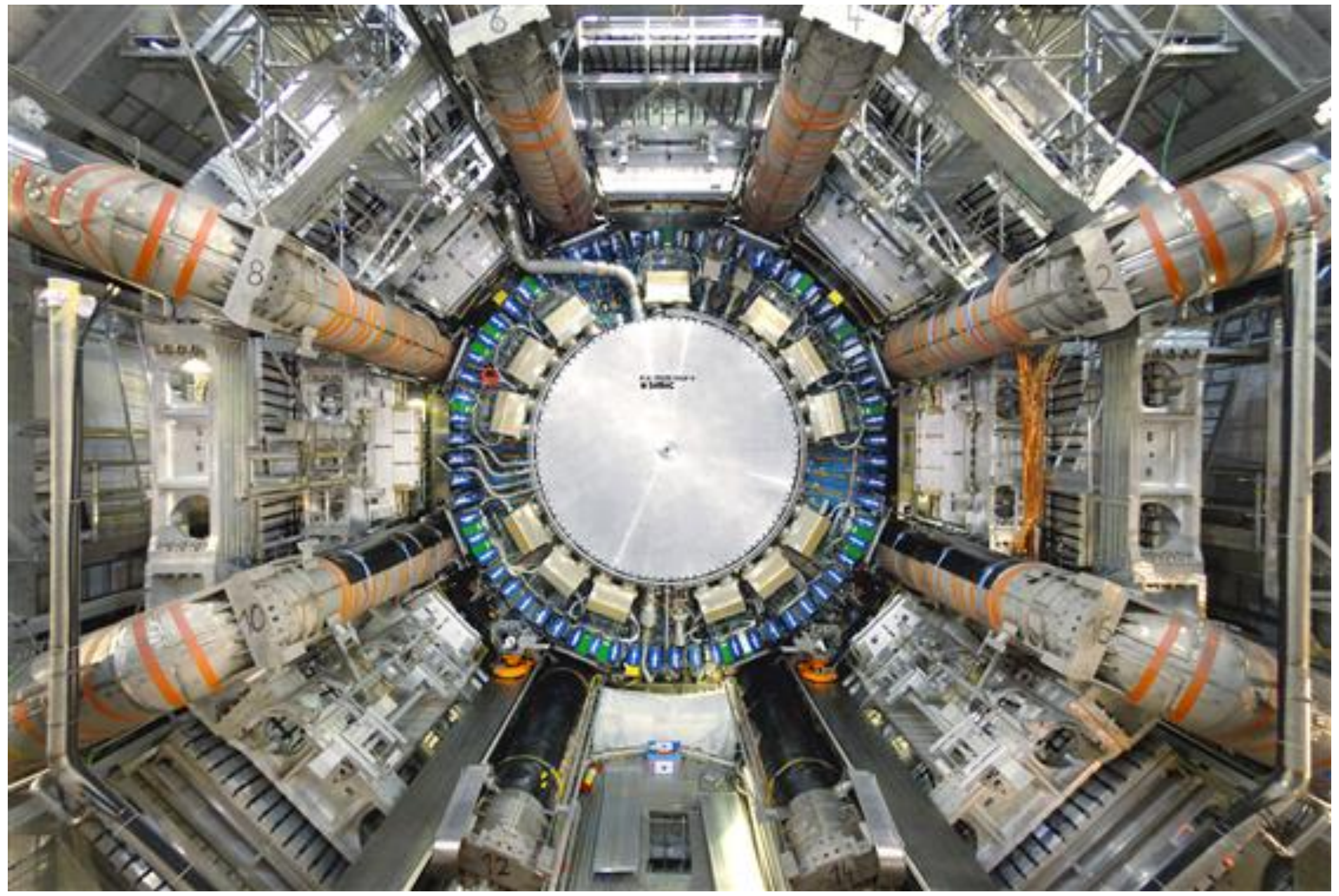


How to deal with mistakes?

Avoid!

- Concentration while performing experiment
- Good understanding of setup and measurement details
- Check of setup and measurement principle by independent observers
- Replication studies





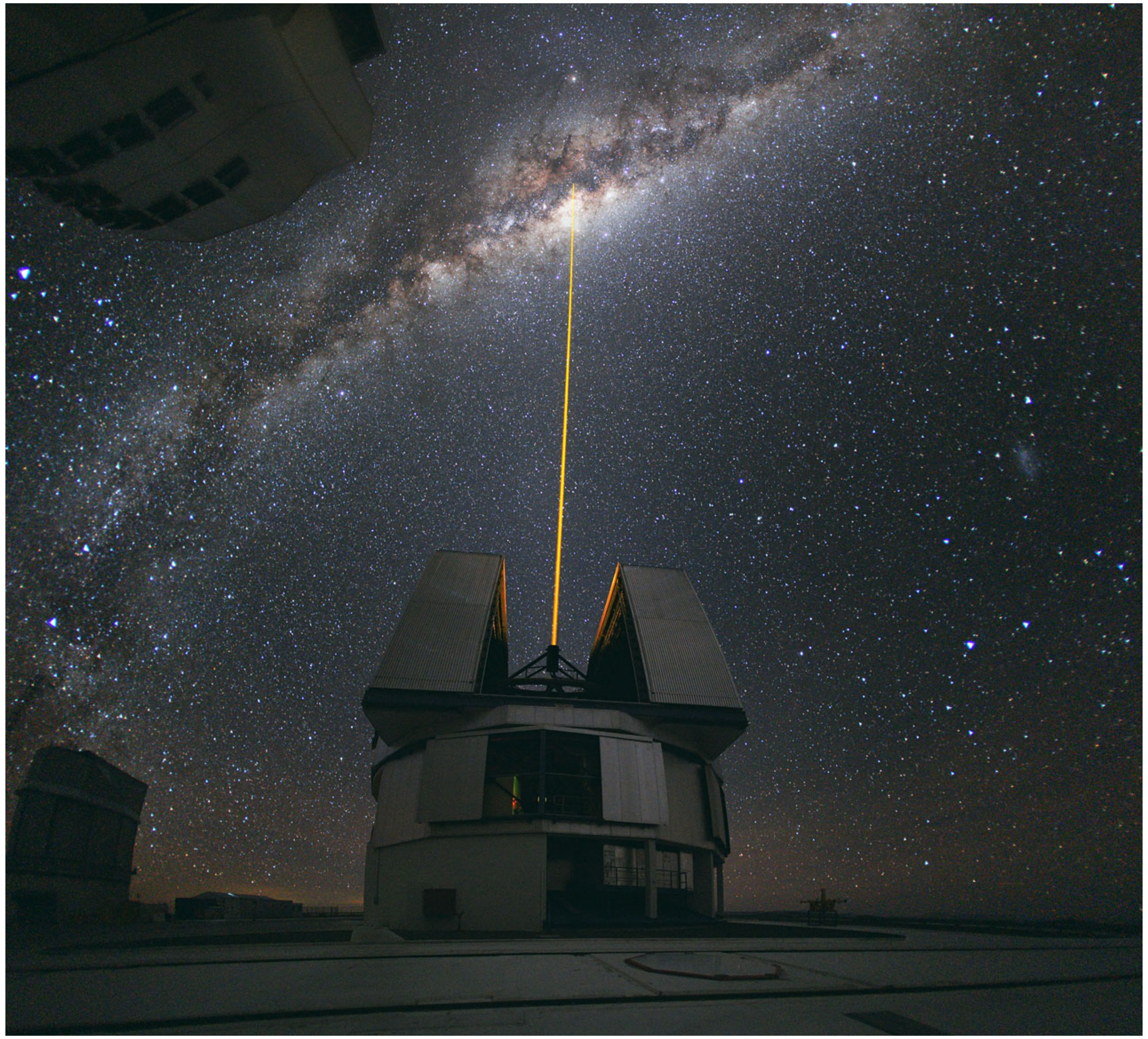
Examples for statistical uncertainties

2. Statistical Uncertainties

- Uncertainties in the result due to fluctuations which can be addressed by statistical methods.

- Number of decays in radioactive material
- Detector signal of particle
- Variation of reaction time using stopwatch
- Atmospheric distortion

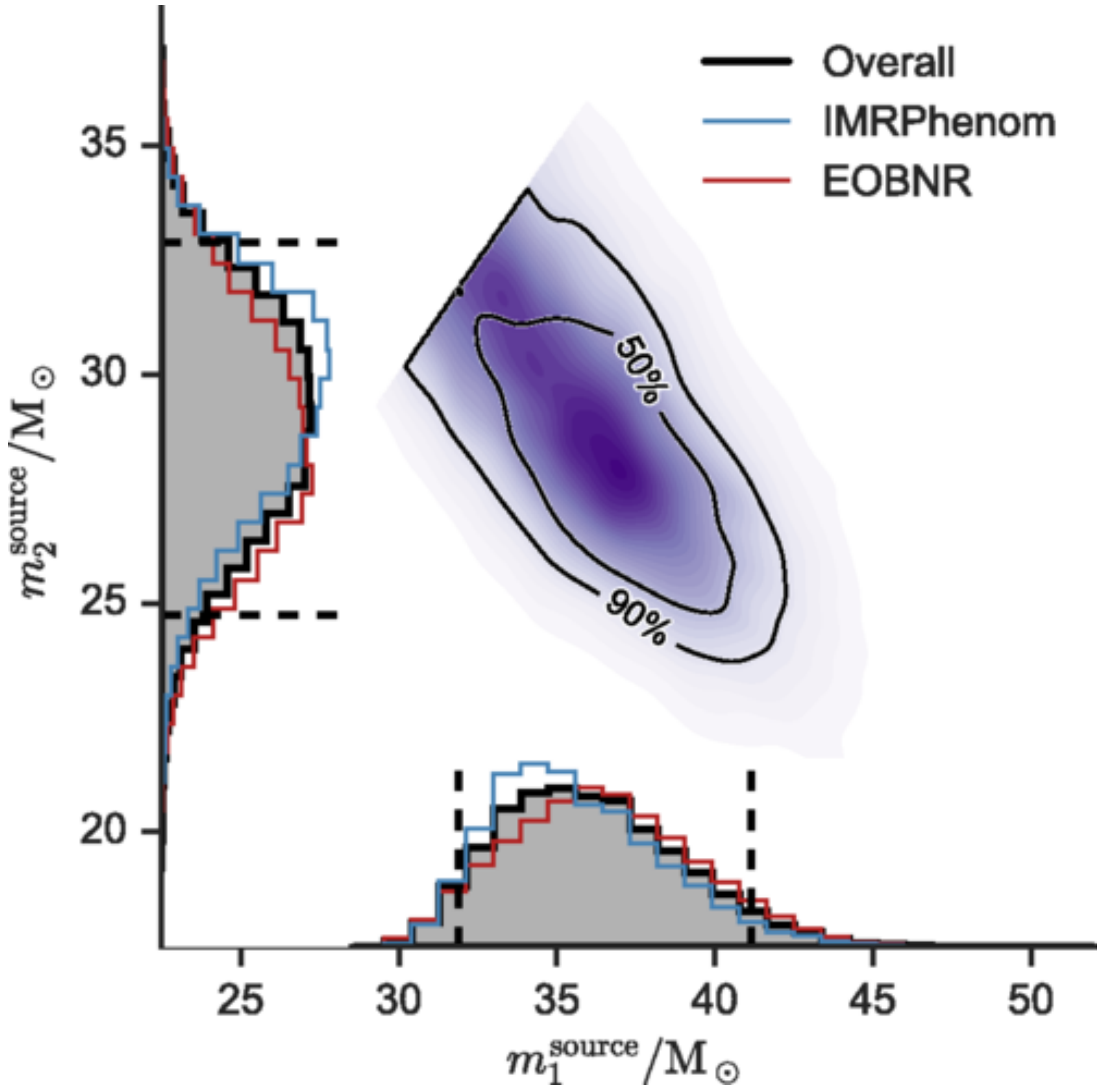
- Usually not possible to correct
- Usually possible to decrease by collecting more data



VLT

How to deal with statistical uncertainties?

- Main part of this lecture
- Useful: When repeating measurement, statistical fluctuation usually independent
- This allows for simple mathematical description
- Usually: The more knowledge you have about the nature of the fluctuations, the better you can deal with them!

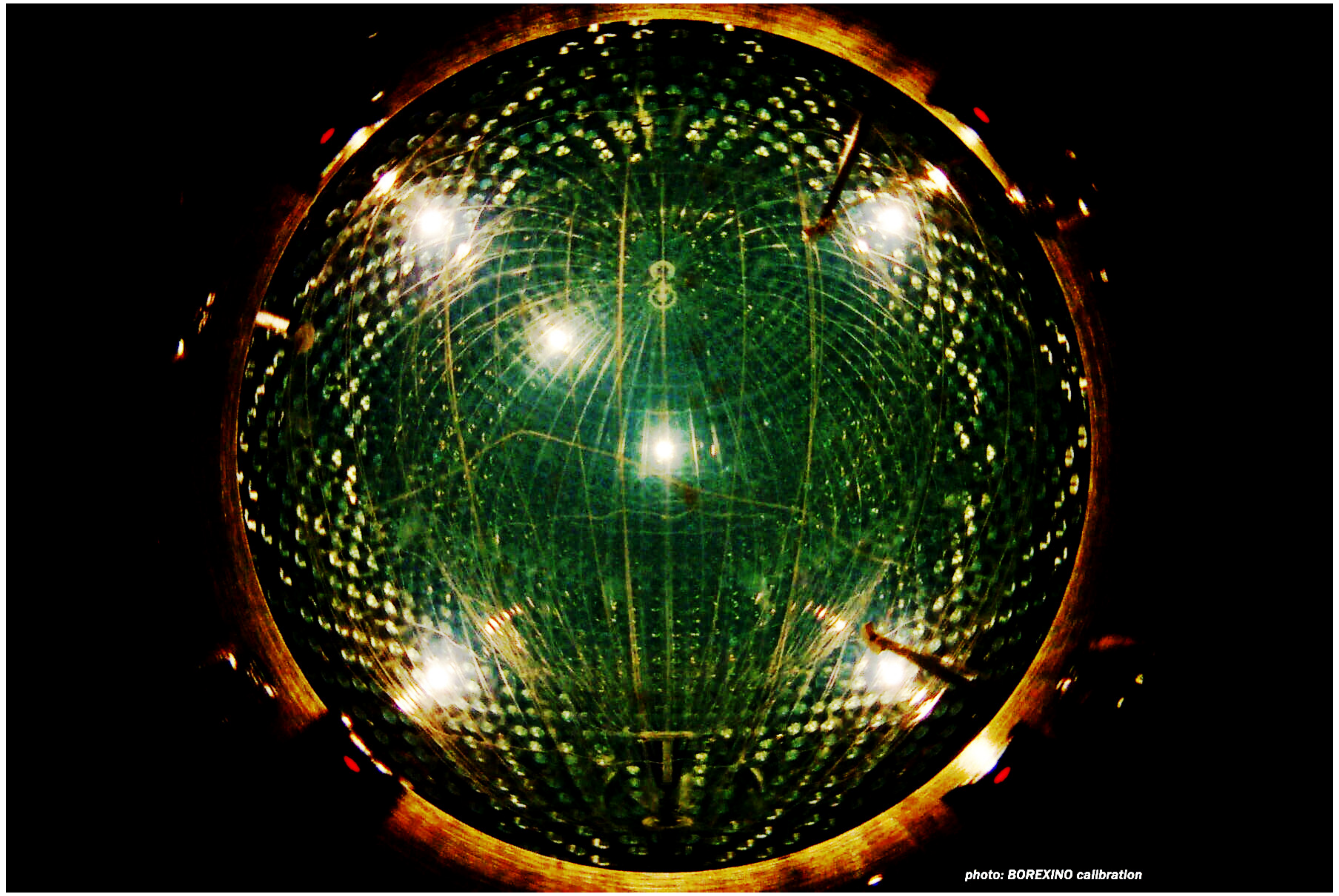


3.2 Systematic Uncertainties

Examples of systematic uncertainties

- Borexino experiment searching for neutrino oscillations in Gran Sasso underground laboratory
- Balloon of scintillator liquid, light measured by photomultipliers
- Turned out to have a small leak, causing deformation - effect on measurement

- Other:
- Thermal expansion of equipment
- Pressure change in gas detectors
- Model assumptions (e.g. neglecting friction)
- Theory uncertainties (e.g. for subtracting backgrounds)
- Imperfect calibration, e.g. of a calorimeter



Borexino experiment

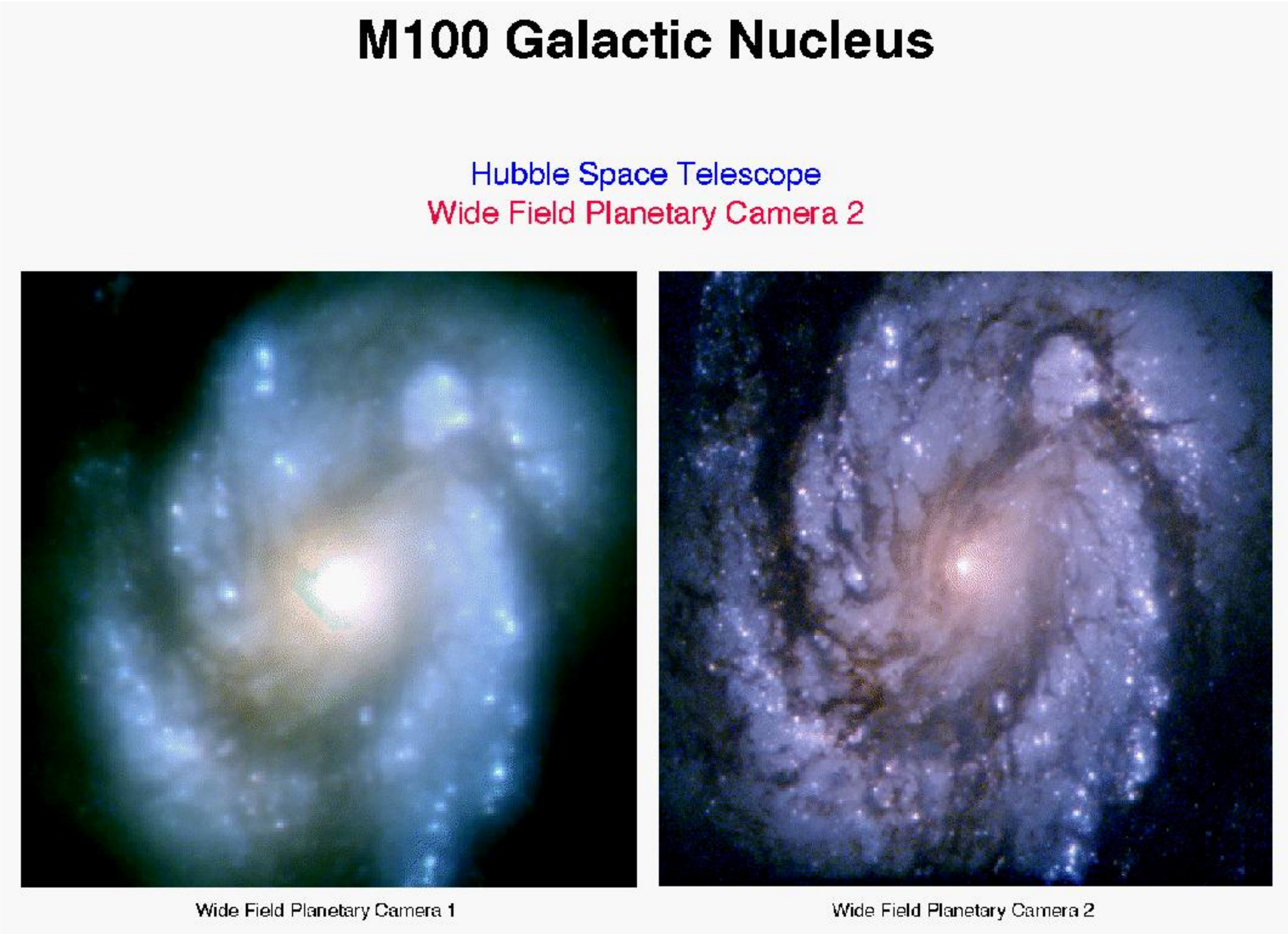
How to deal with systematic uncertainties?

Systematic effect: Cause of some deviation from the correct result

Systematic error: Size of the deviation

Systematic uncertainty: Estimate of the deviation

R. Barlow
“Systematic Errors, Fact and Fiction,” hep-ex/0207026


M100 Galactic Nucleus
Hubble Space Telescope
Wide Field Planetary Camera 2
Wide Field Planetary Camera 1
Wide Field Planetary Camera 2

hubblesite.org

Generally:

1. Find and understand all relevant systematic effects
2. Correct them as far as possible
3. Uncertainty in the size of the correction is then the systematic uncertainty

How to deal with systematic uncertainties?

- Different effects for different experiments - hard to give general approach
- Often systematic uncertainty main difficulty
- Much neglected in statistics literature
- In data analyses in high energy physics: statistical uncertainties often straightforward, systematics need most of the time
- Strongly helped by experience
- Studying examples helps
- Often hard to estimate effects precisely

Bayesian approach to systematic uncertainties

- Systematic uncertainties usually do not independently fluctuate from measurement to measurement - hard do describe in frequentist terms
- Some systematics are Bayesian only - description of knowledge about system

Quantity of interest: θ, prior knowledge: $\pi(\theta)$

Likelihood depends parameter ν ("nuisance parameter")

We simply treat θ and ν as an unknown parameters:

$$P(\theta, \nu | \mathsf{data}) \propto L(\mathsf{data} | \theta, \nu) \pi(\theta, \nu)$$

"Bayesians lose no sleep over systematics" (lecture S. Oser)

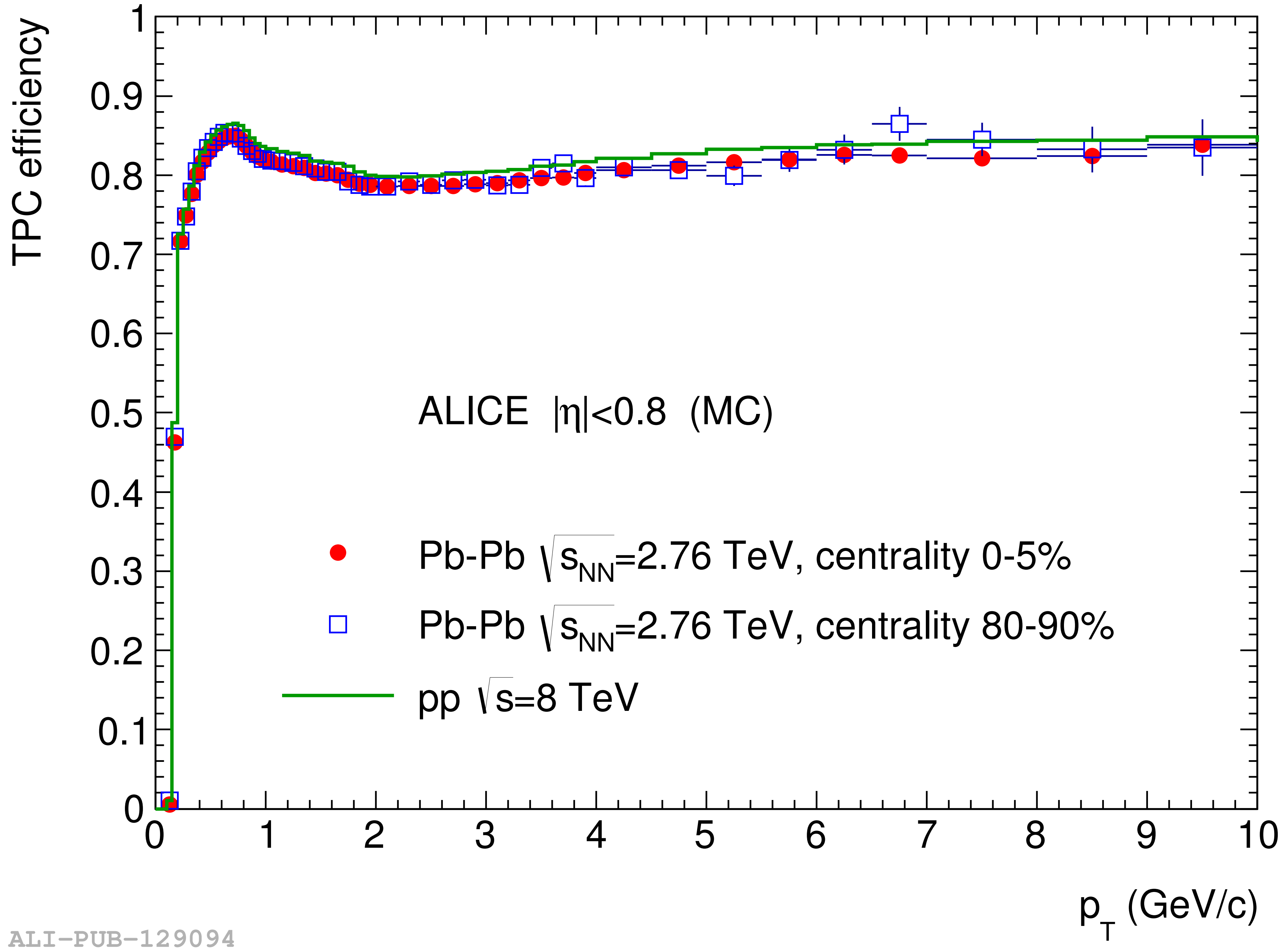
As we are only interested in θ, we marginalize by integrating over ν:

$$P(\theta) = \int P(\theta, \nu) \, d\nu$$

Prior knowledge on ν often is the result of a calibration measurement.

Example 1: Efficiency correction

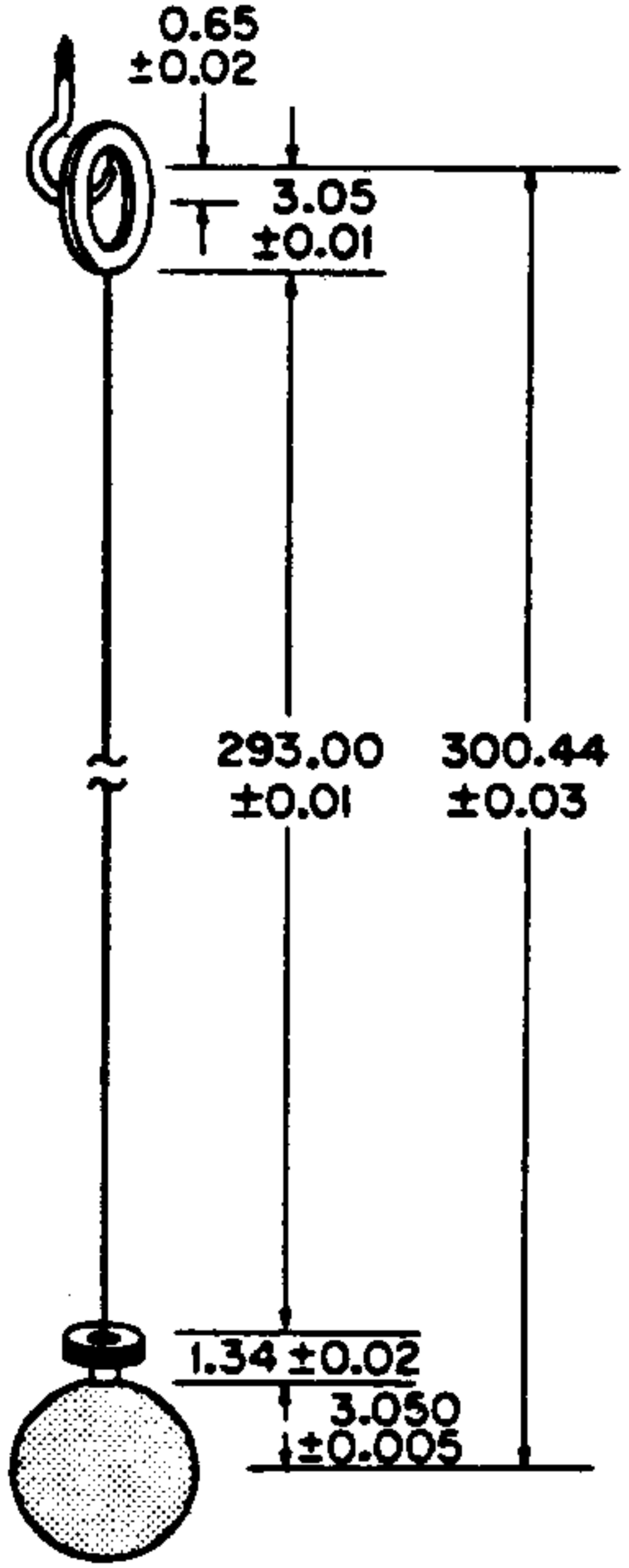
- Want to calculate e.g. cross section, measure particle number
- $\langle n \rangle_{meas} = \langle n \rangle_{true} \cdot \epsilon$, with ϵ the efficiency $\rightarrow \langle n \rangle_{true} = \langle n \rangle_{meas} / \epsilon$
- Estimate ϵ from (Monte Carlo) simulations
  - Uncertainty from binomial fluctuations
  - Uncertainty from how well the simulation describes the real world
- Bayesian: This results in some probability distribution $p(\epsilon)$; from this we get $p(\langle n \rangle_{true}, \epsilon)$, and we marginalize out ϵ
- Frequentist: Define repetition of experiment as repeating both the measurement of n and the simulation - now both effects can be considered as frequencies
  - This does not include the uncertainty in the simulation describing the real world
- Ask: Which detector effects are relevant? And how are they implemented in the MC code?



Example 2: Pendulum

- $g \approx (2\pi/T_0)^2 \cdot L$

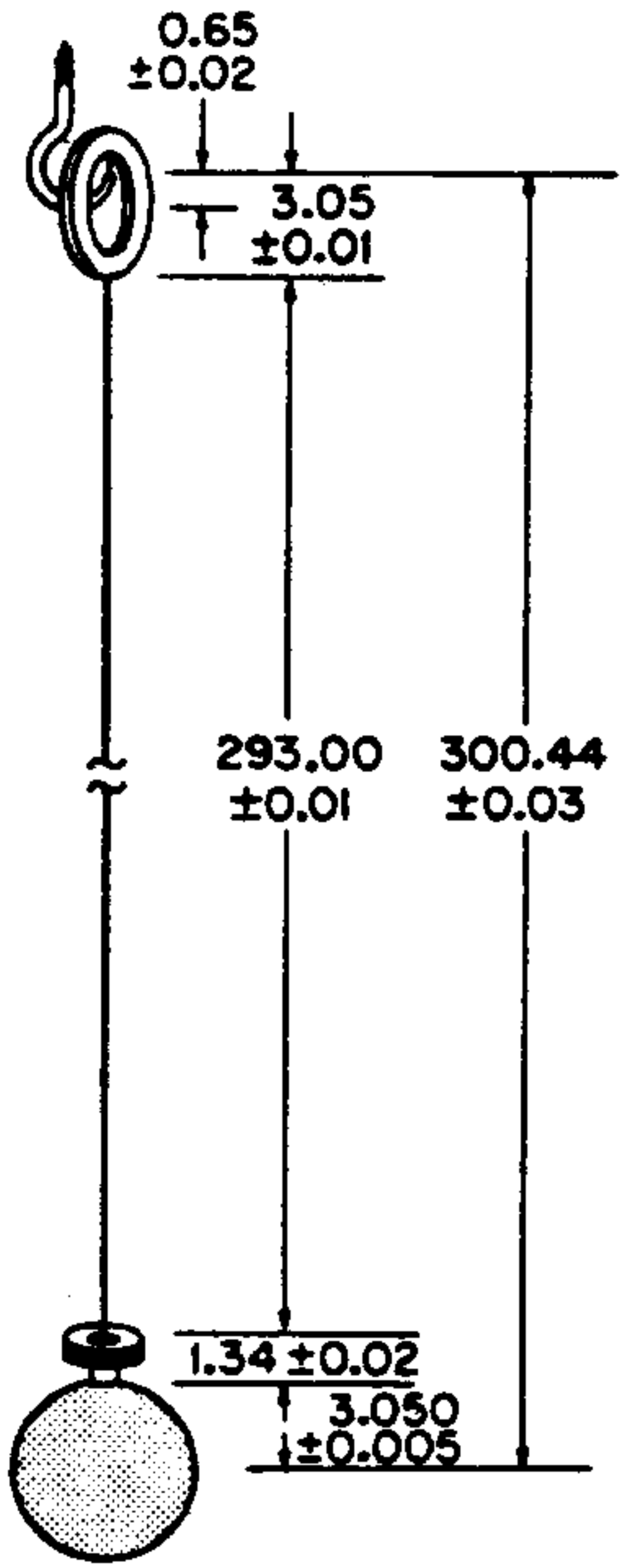
1. Not a harmonic oscillator at finite amplitude
2. Weight is not a point - angular momentum
3. Buoyancy of weight in atmosphere
4. Damping of oscillations by air
5. Mass of the wire
6. Elasticity of wire
7. Increased effective weight by air being dragged along (non-dissipative)



The pendulum – Rich physics from a simple system, R. Nelson, M. Olsson

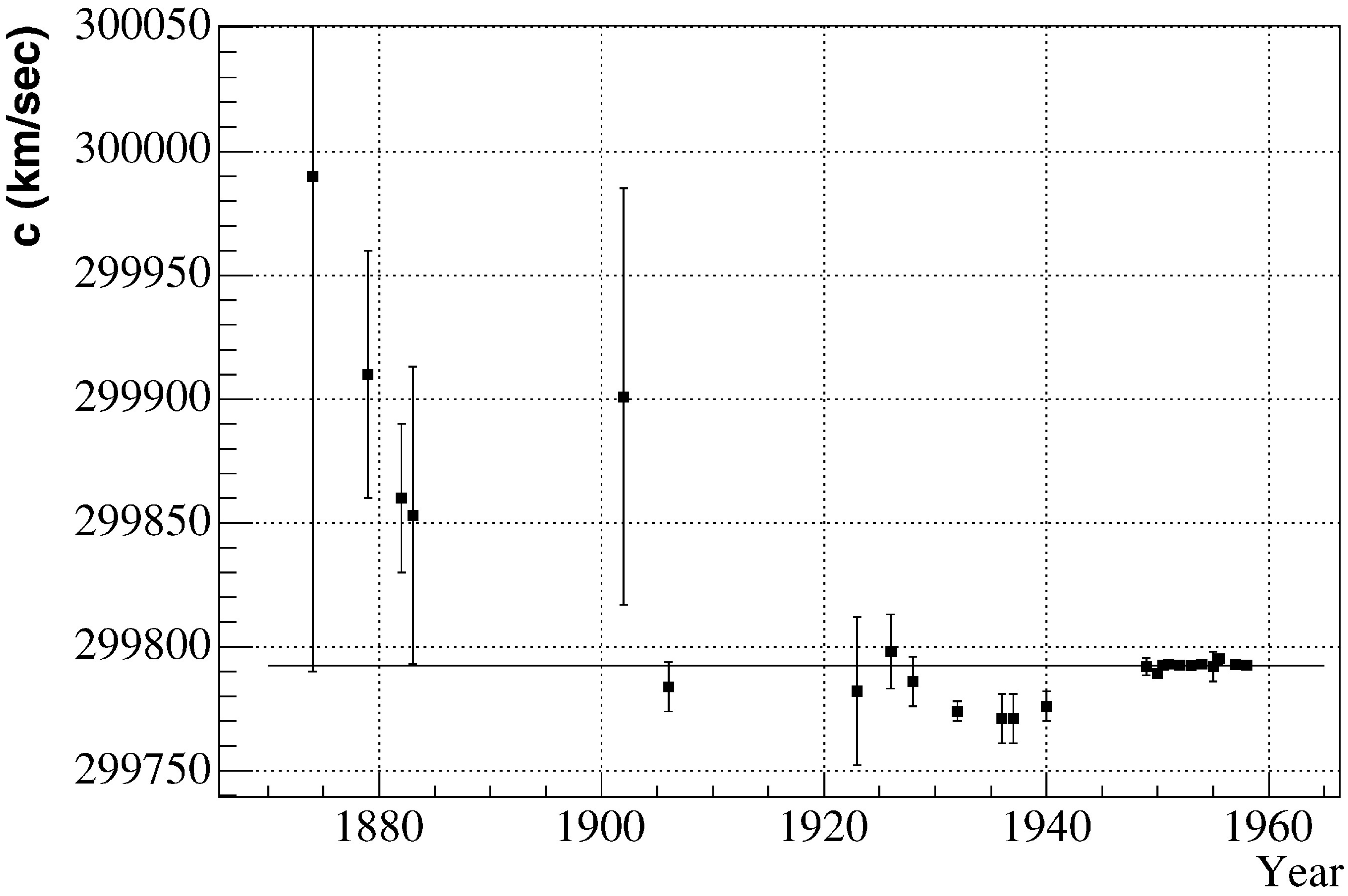
Example 2: Pendulum

- Which effects are dominant?
- Calculate effects
  - Fully
  - Estimate via approximation
  - Just some maximum possible effect?
- Try to decrease effects
  - With a longer wire
  - With smaller amplitude
  - With denser material
  - In a vacuum



The pendulum – Rich physics from a simple system, R. Nelson, M. Olsson

Speed of light measurements vs. year of publication

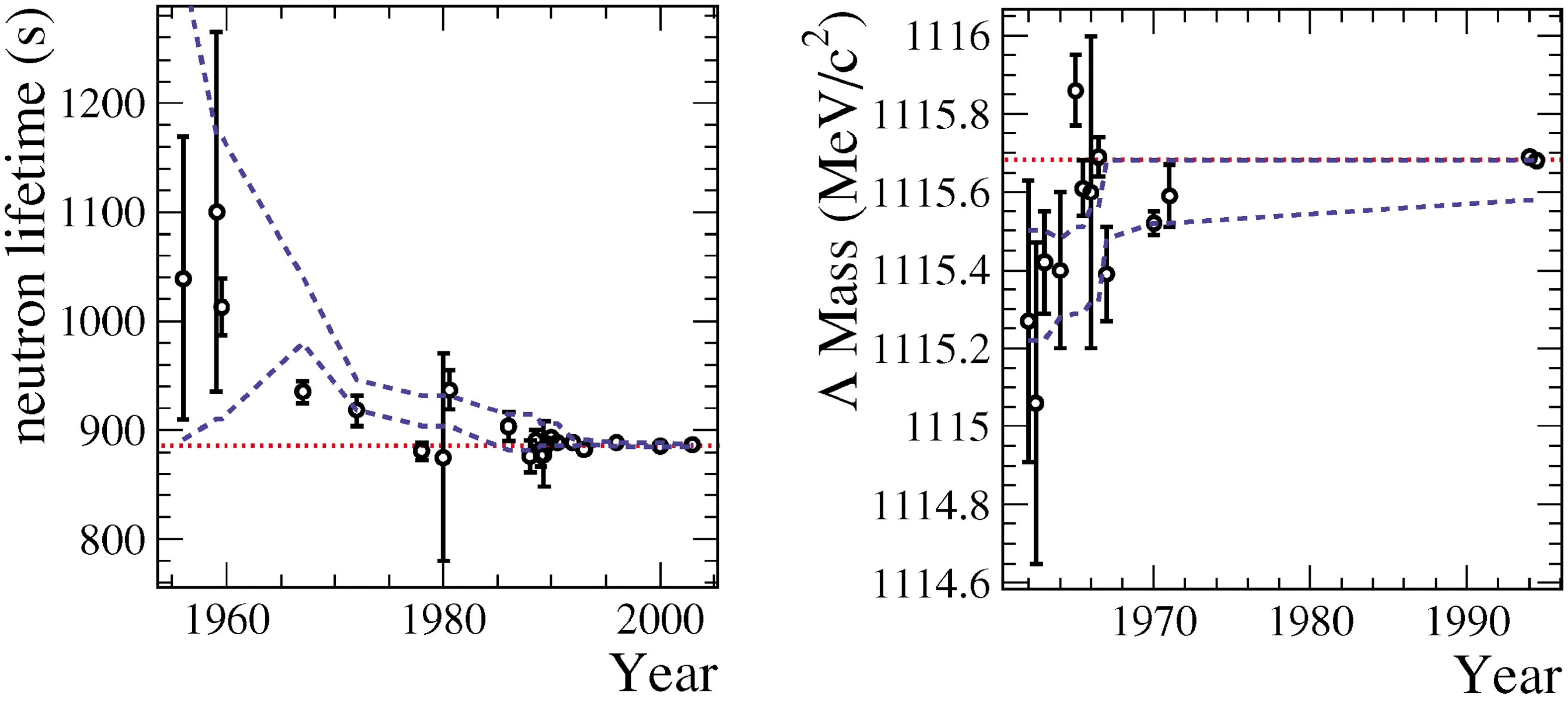


Klein JR, Roodman A. 2005.
Annu. Rev. Nucl. Part. Sci. 55:141–63

Experimenter's bias?

Klein JR, Roodman, A. 2005,
Annu. Rev. Nucl. Part. Sci. 55:141–63

Do researchers unconsciously work towards a certain value?



Possible bias:

the investigator searches for the source or sources of such errors, and continues to search until she/he gets a result close to the accepted value.

Then he/she stops!

Some amount of paranoia can be useful!

Example 3: Thermal expansion

- Measurement may be affected by thermal expansion
- Calibration at one temperature, measurement at another
- Similarly: Particle detector signals depend on temperature, atmospheric pressure etc.

Possibility 1: We also have a temperature measurement

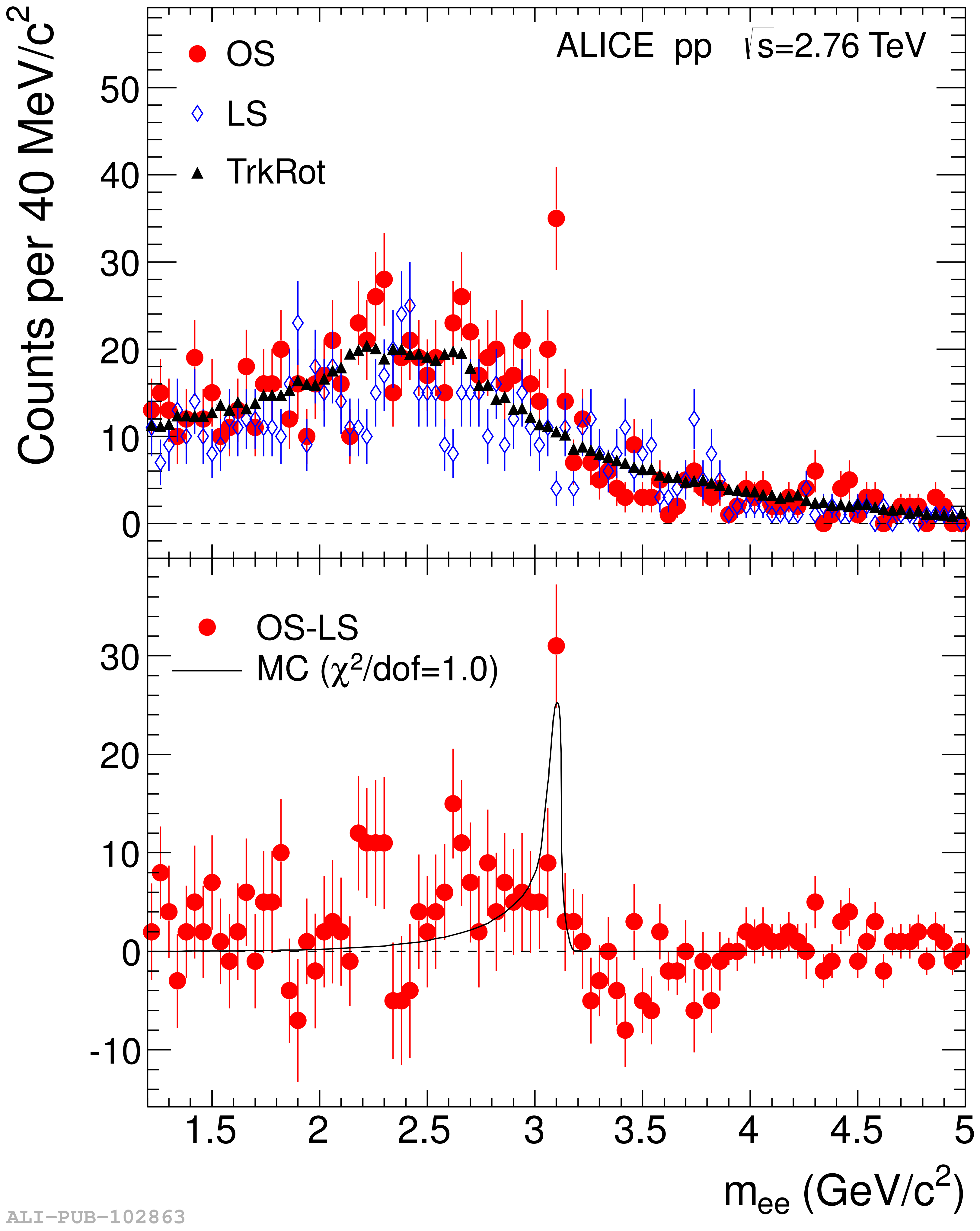
- Calculate effect from thermal expansion and correct for it
- Propagate uncertainty of temperature measurement and expansion coefficient

Possibility 2: We do not have a temperature measurement

- Consider possible range of temperatures in experiment hall (summer, winter, open windows etc.)
- Propagate uncertainty to final result

Example 4: Invariant mass background

- Measurement of $\mathrm{J}/\psi \rightarrow e^+e^-$
- Background estimated from measuring e^+e^+ and e^-e^- pairs
- How good is this estimate?
- Hard to list every contributing effect
- Instead try another estimate based on track rotations (TrkRot)
- If systematic effects are very different, the difference can tell us about systematic error



Inclusive J/ψ production in pp collisions at $\sqrt{s}$ = 2.76 TeV, ALICE Collaboration

Handling discrete systematic uncertainties

Example: choice of model used to determine a correction *R*

With 1 preferred model and one other, quote $R_1 \pm |R_1 - R_2|$

With 2 models of equal status, quote $\dfrac{R_1 + R_2}{2} \pm \dfrac{|R_1 - R_2|}{\sqrt{2}}$

n equal models, quote $\bar{R} \pm \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(R_i - \overline{R})^2} = \sqrt{\dfrac{n}{n-1}(\overline{R^2} - \overline{R}^2)}$

Two extreme models, quote $\dfrac{R_1 + R_2}{2} \pm \dfrac{|R_1 - R_2|}{\sqrt{12}}$

Example 5: Resistor

- These have tolerances printed on them
- But better to measure them in addition, vastly decreases uncertainty

Sanity / Consistency checks

R. Barlow
"Systematic Errors, Fact and Fiction," hep-ex/0207026

Look for systematic *effects* by repeating the analysis with changes which *should* make no difference:

Data subsets

Magnet up/down

Different selection cuts

Different histogram bin sizes and fit ranges

Different Event Generator for efficiency calculation

Look for impossibilities

If a check passes the test:
move on and do not add the discrepancy to the systematic uncertainty

If a check fails: try to identify the reason. Only as very last resort, add contribution to total systematic uncertainty. This might underestimate the real uncertainty.

Blind analyses

Klein JR, Roodman, A. 2005,
Annu. Rev. Nucl. Part. Sci. 55:141–63

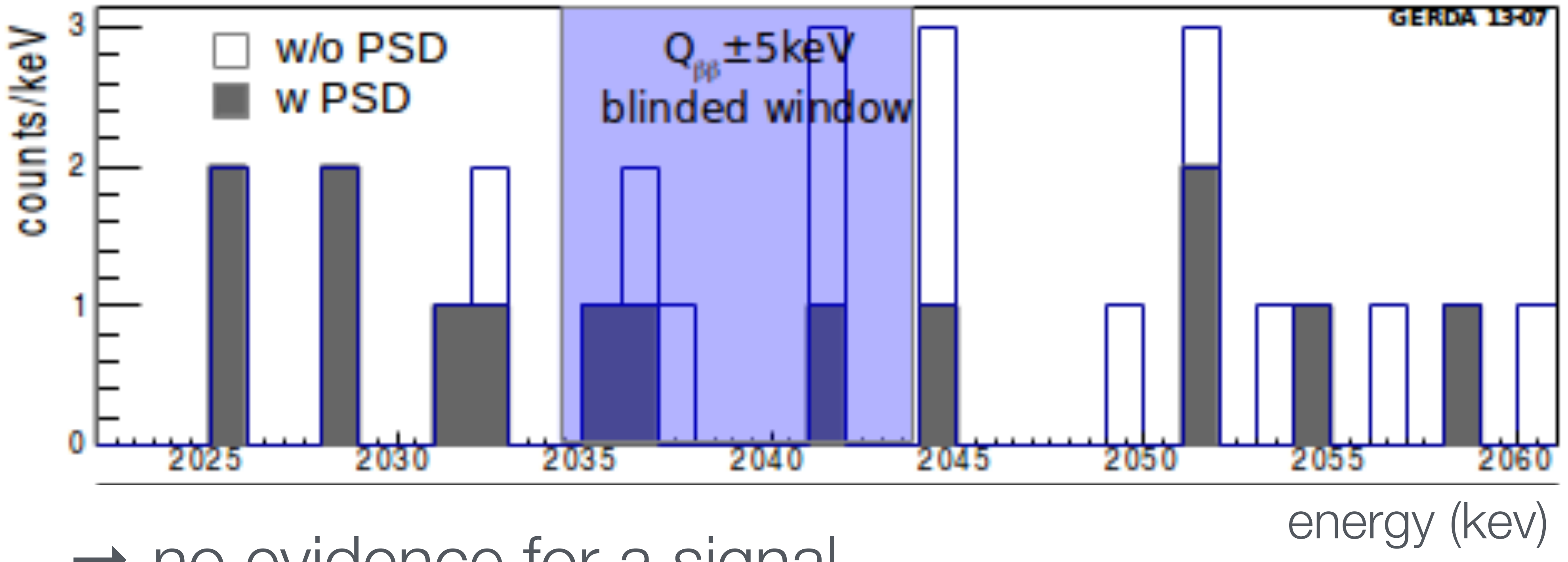
Avoid experimenter's bias by hiding certain aspects of the data.

Things that can be hidden in the analysis:

- The signal events, when the signal occurs in a well-defined region of the experiment's phase space.
- The result, when the numerical answer can be separated from all other aspects of the analysis.
- The number of events in the data set, when the answer relies directly upon their count.
- A fraction of the entire data set.

Example: GERDA experiment

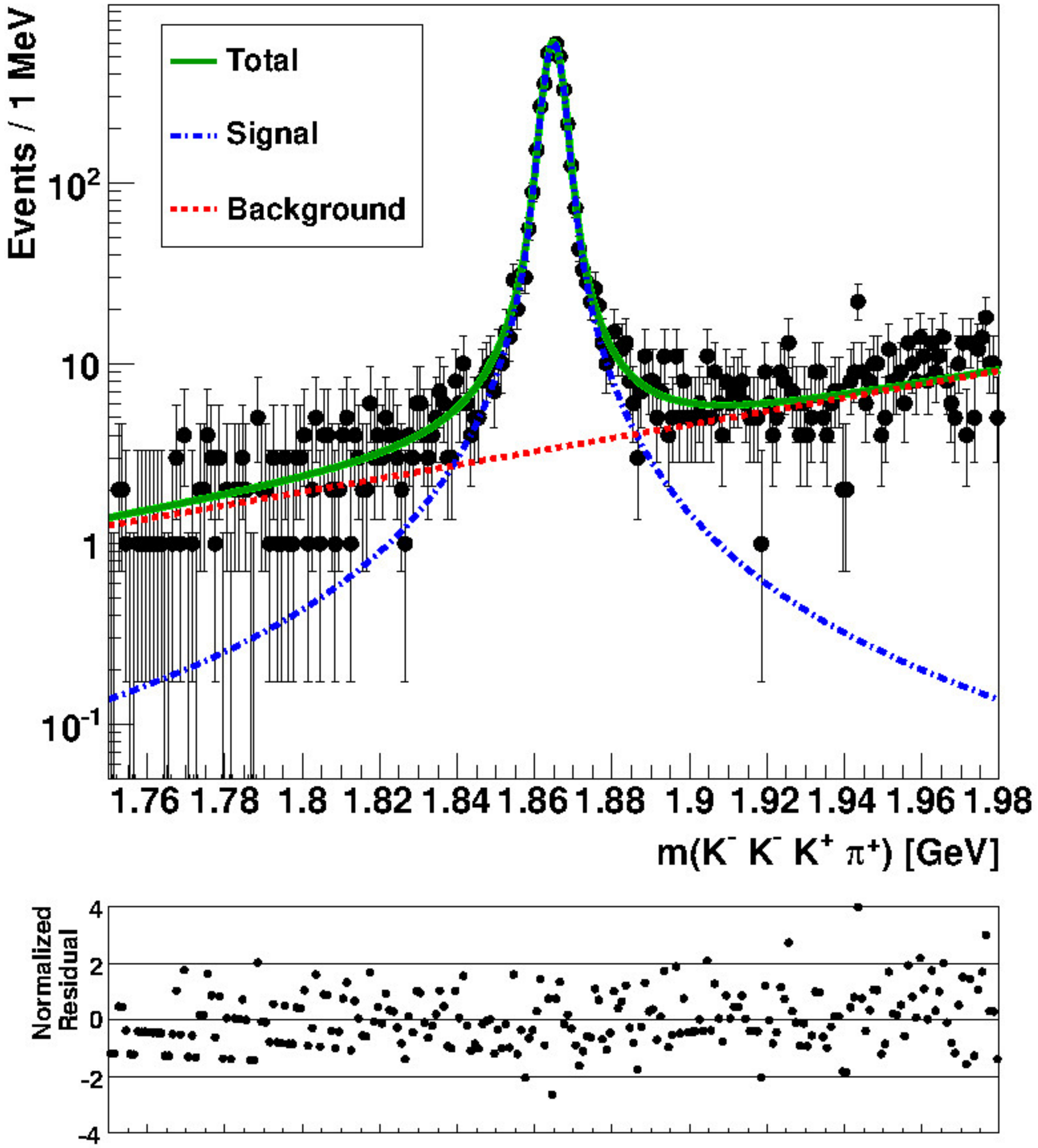
- search for neutrinoless double beta decay
- Signal: sharp peak
- Background model fixed prior to unblinding of signal region



→ no evidence for a signal

Example 6: D^0 mass

- Largest contribution found to be mass of kaons
- Take mass and uncertainty from literature (e.g. PDG)
- Propagate to D meson mass
- Here: Repeated analysis with different mass assumptions



Measurement of the mass of the D^0 meson, BABAR Collaboration

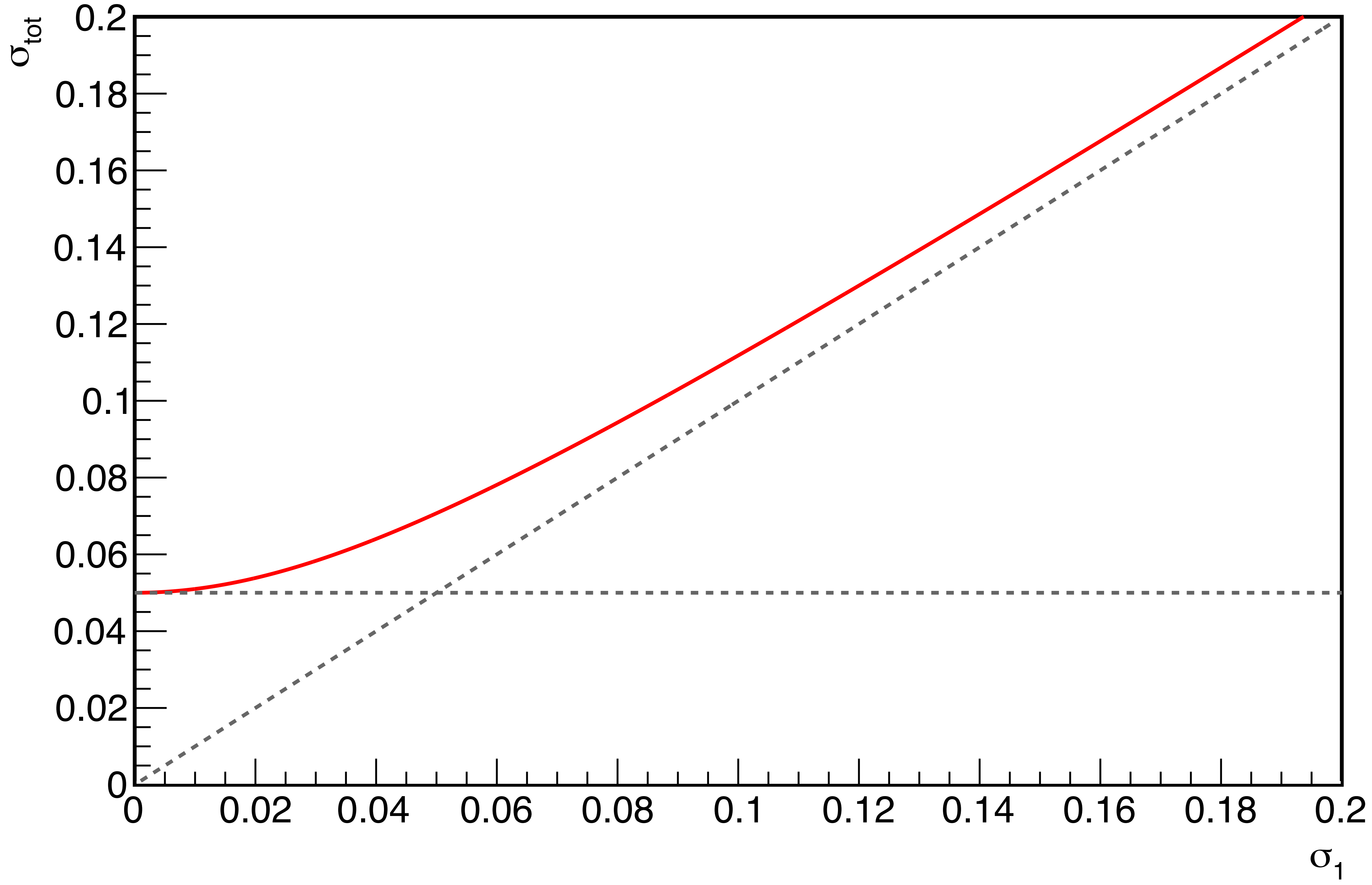
Combination of systematic uncertainties

Systematic uncertainties are usually given as standard deviations ($x \pm \sigma_x$), corresponding to a 68% probability.

Other meaning (e.g. maximum extent uncertainty) this should be explicitly stated.

In most cases one tries to find independent sources of systematic uncertainties. These independent uncertainties are therefore added in quadrature:

$$\sigma_{\mathrm{tot}}^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$$



Often a few source dominate the systematic uncertainty

→ No need to work to hard on precisely estimating the small uncertainties

Systematic uncertainties: Covariance matrix approach (I)

Consider two measurement x_1 and x_2 with with individual random uncertainties $\sigma_{1,r}$ and $\sigma_{2,r}$ and a common systematic uncertainty σ_s:

$$x_i = x_{\mathrm{true}} + \Delta x_{i,\mathrm{r}} + \Delta x_{\mathrm{s}}$$

$$\langle \Delta x_{i,\mathrm{r}} \rangle = 0, \quad \langle \Delta x_{\mathrm{s}} \rangle = 0,$$
$$\langle (\Delta x_{i,\mathrm{r}})^2 \rangle = \sigma_{i,\mathrm{r}}^2, \quad \langle (\Delta x_{\mathrm{s}})^2 \rangle = \sigma_s^2$$

Variance:

$$\begin{aligned} V[x_i] &= \langle x_i^2 \rangle - \langle x_i \rangle^2 \\ &= \langle (x_{\mathrm{true}} + \Delta x_{i,\mathrm{r}} + \Delta x_{\mathrm{s}})^2 \rangle - \langle x_{\mathrm{true}} + \Delta x_{i,\mathrm{r}} + \Delta x_{\mathrm{s}} \rangle^2 \\ &= \langle (\Delta x_{i,\mathrm{r}} + \Delta x_{\mathrm{s}})^2 \rangle \\ &= \sigma_{i,r}^2 + \sigma_s^2 \end{aligned}$$

Covariance:

$$\begin{aligned} \mathrm{cov}[x_1, x_2] &= \langle x_1 x_2 \rangle - \langle x_1 \rangle \langle x_2 \rangle \\ &= \ldots \\ &= \sigma_s^2 \end{aligned}$$

Systematic uncertainties: Covariance matrix approach (II)

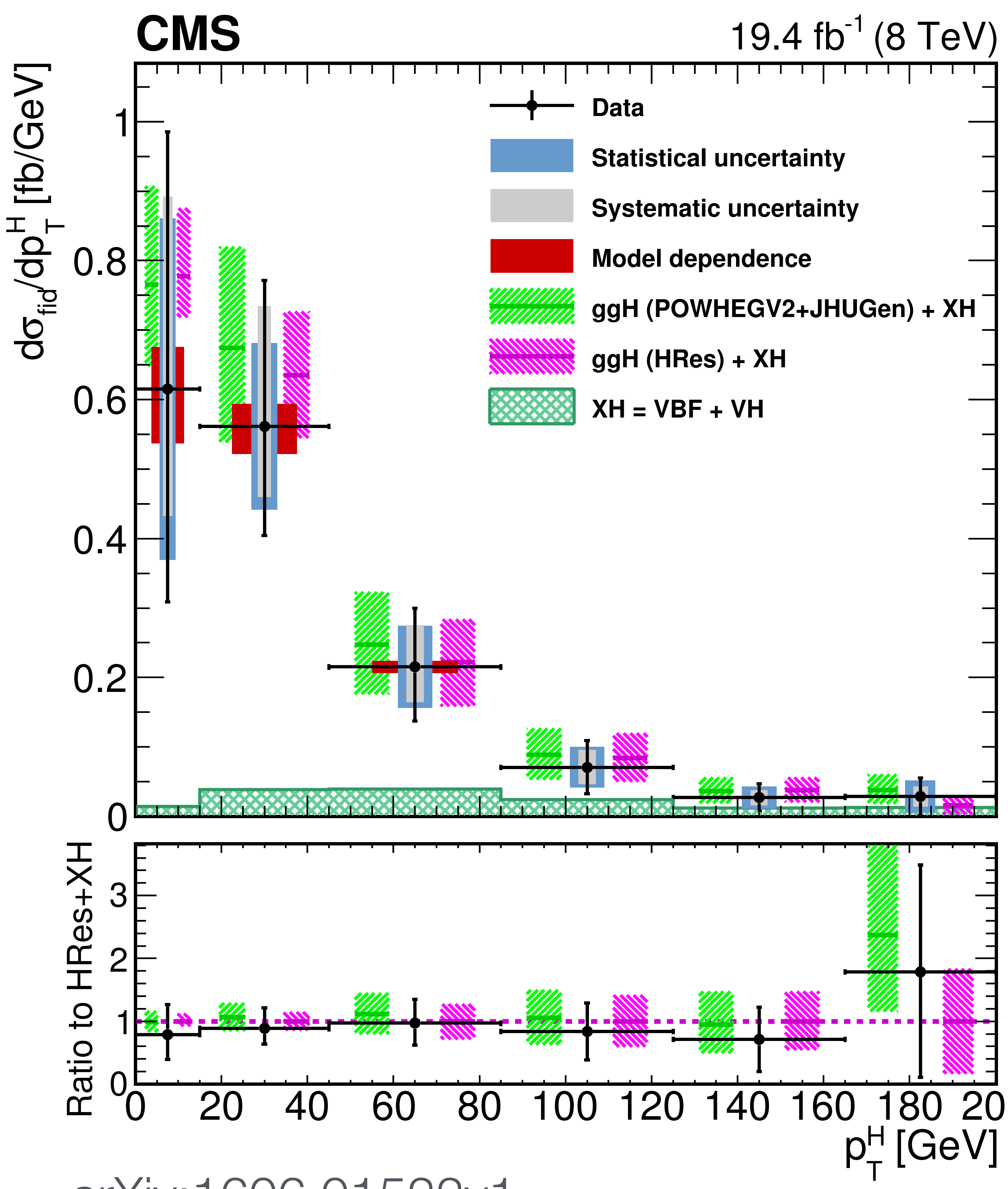
Covariance matrix for x_1 and x_2:

$$V = \begin{pmatrix} \sigma_{1,r}^2 + \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_{2,r}^2 + \sigma_s^2 \end{pmatrix}$$

This also works when the uncertainties are quoted as relative uncertainties:

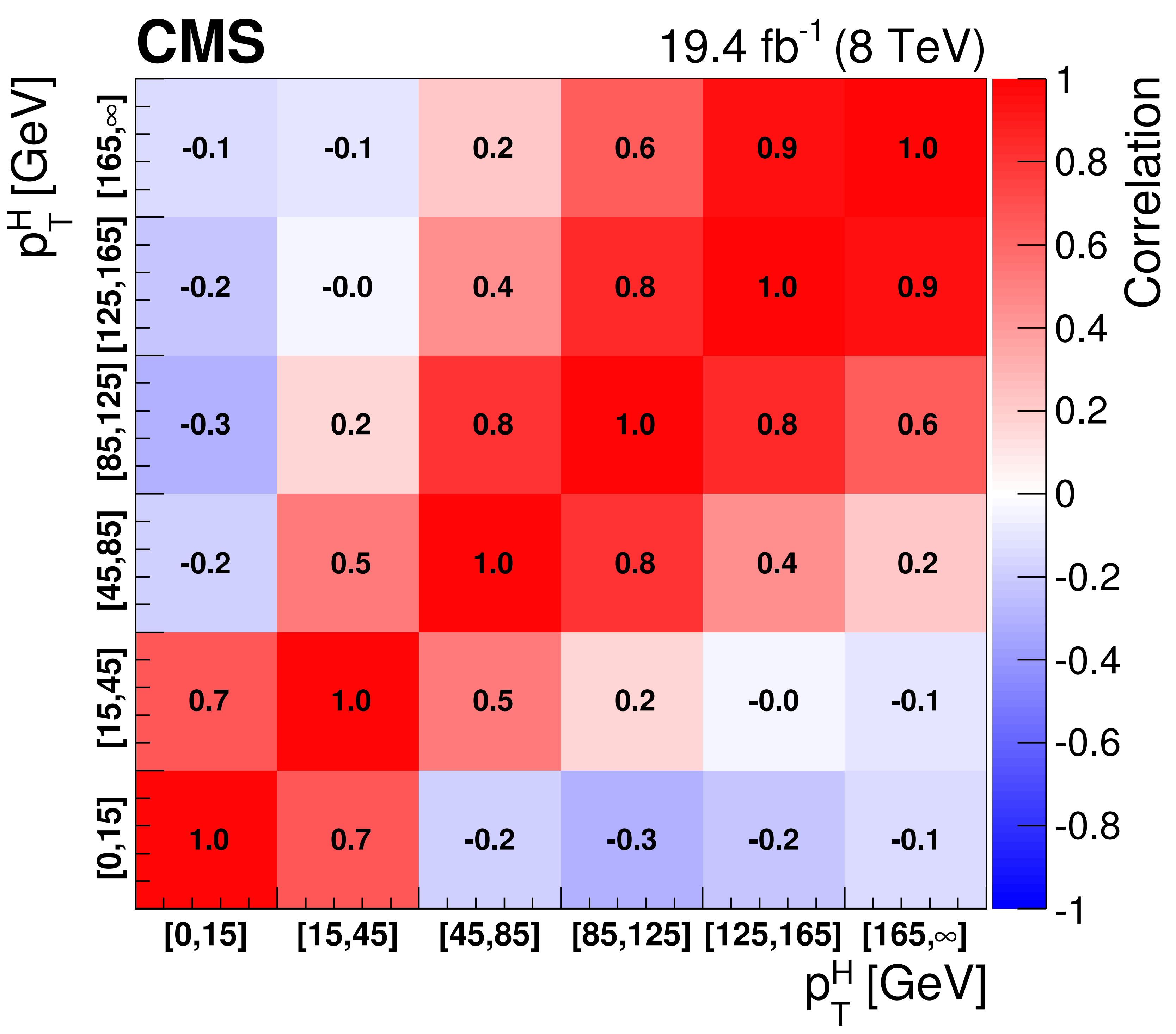
$$\sigma_s = \varepsilon x \qquad \rightsquigarrow \qquad V = \begin{pmatrix} \sigma_{1,r}^2 + \varepsilon^2 x_1^2 & \varepsilon^2 x_1 x_2 \\ \varepsilon^2 x_1 x_2 & \sigma_{2,r}^2 + \varepsilon^2 x_1^2 \end{pmatrix}$$

Example:
Transverse momentum spectrum of the Higgs boson



arXiv:1606.01522v1

Correlation matrix of the p_T bins:



	[0,15]	[15,45]	[45,85]	[85,125]	[125,165]	[165,∞]
[165,∞]	-0.1	-0.1	0.2	0.6	0.9	1.0
[125,165]	-0.2	-0.0	0.4	0.8	1.0	0.9
[85,125]	-0.3	0.2	0.8	1.0	0.8	0.6
[45,85]	-0.2	0.5	1.0	0.8	0.4	0.2
[15,45]	0.7	1.0	0.5	0.2	-0.0	-0.1
[0,15]	1.0	0.7	-0.2	-0.3	-0.2	-0.1

$$\rho_{i,j} = \frac{V_{i,j}}{\sigma_i \sigma_j}, \qquad V = \text{covariance matrix}$$

Weighted average of correlated data points

Consider *n* data points y_i with covariance matrix *V*: $\vec{y} = (y_1, y_2, \dots, y_n)$

One can calculate a weighted average λ by minimizing

$$\chi^2(\lambda) = (\vec{y} - \vec{\lambda})^\mathsf{T} V^{-1} (\vec{y} - \vec{\lambda})$$

$\vec{\lambda} := (\lambda, \lambda, \dots, \lambda)$

One obtains (here without calculation):

$$\hat{\lambda} = \sum_{i=1}^{N} w_i y_i \qquad w_i = \frac{\sum_{j=1}^{n} (V^{-1})_{i,j}}{\sum_{k,l=1}^{n} (V^{-1})_{k,l}}$$

Variance results from error propagation:

$$\sigma_{\hat{\lambda}}^2 = \vec{w}^\mathsf{T} V \vec{w} = \sum_{i,j=1}^{n} w_i V_{ij} w_j$$

Minimizing the χ^2 gives the *best linear unbiased estimate* (BLUE) → linear unbiased estimator with the lowest variance

- BLUE combination may be biased if uncertainties not known or are estimated from measured values
- Improvement: iterative approach (rescaling uncertainties based on previous iteration)

Special case:
Weighted average of two correlated measurements

Consider two measurements with covariance matrix *V* (ρ = correlation coeff.):

$$y_1,\ y_2 \qquad V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Applying the formulas from the previous slide:

$$V^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix} \qquad \hat{\lambda} = wy_1 + (1-w)y_2$$

$$w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \qquad V[\hat{\lambda}] = \sigma^2 = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

equivalently:

$$\frac{1}{\sigma^2} = \frac{1}{1-\rho^2}\left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2}\right]$$

Summary of systematic uncertainties

- Large variety of effects - hard to give general recipes
- Systematics do not mean mistakes
- Typical approaches:
  - Decrease uncertainties (e.g. with ancillary measurements, improved setup …)
  - Estimate and correct effects
  - Compare different methods of the same analysis
  - Estimate magnitude of effect
- Quadratic sum means, dominant contributions should get the most attention
  - Often more important not to overlook a large effect rather than having a precise estimate of a smaller one
- A systematic effect that is not found becomes a mistake.

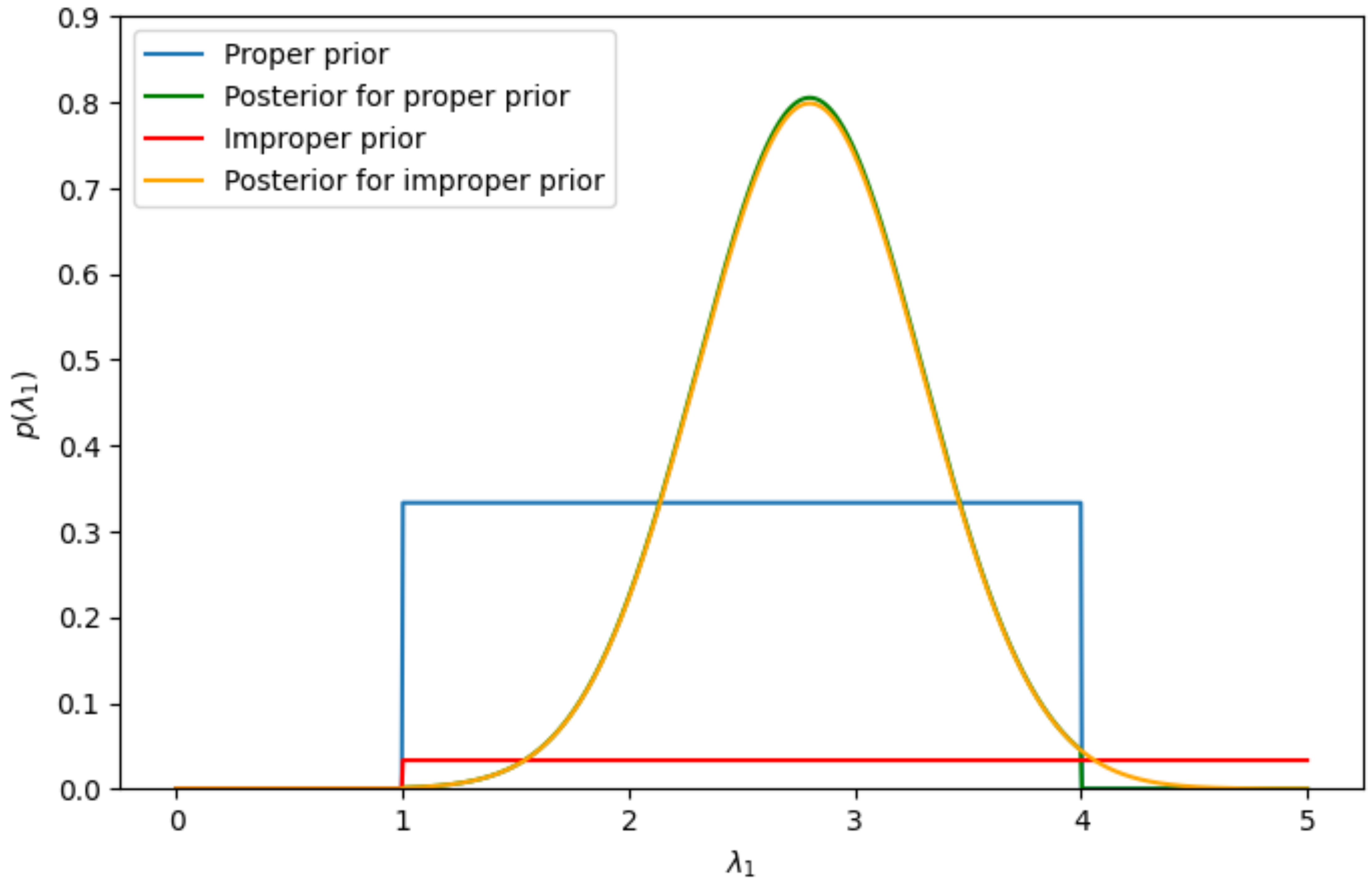
3.3 Priors

Proper and improper priors

- Probability distributions must be normalised → unnormalizable distributions are not allowed

$$p(x\,|\,m) \sim \frac{p(m\,|\,x)\;p(x)}{\int p(m\,|\,x)\;p(x)\,\mathrm{d}x}$$

- Often: Even if prior is not normalisable prior*likelihood is
- Mathematically convenient
- Non-normalizable priors are called *improper priors*
- Often simplifies calculation, occasionally leads to problems

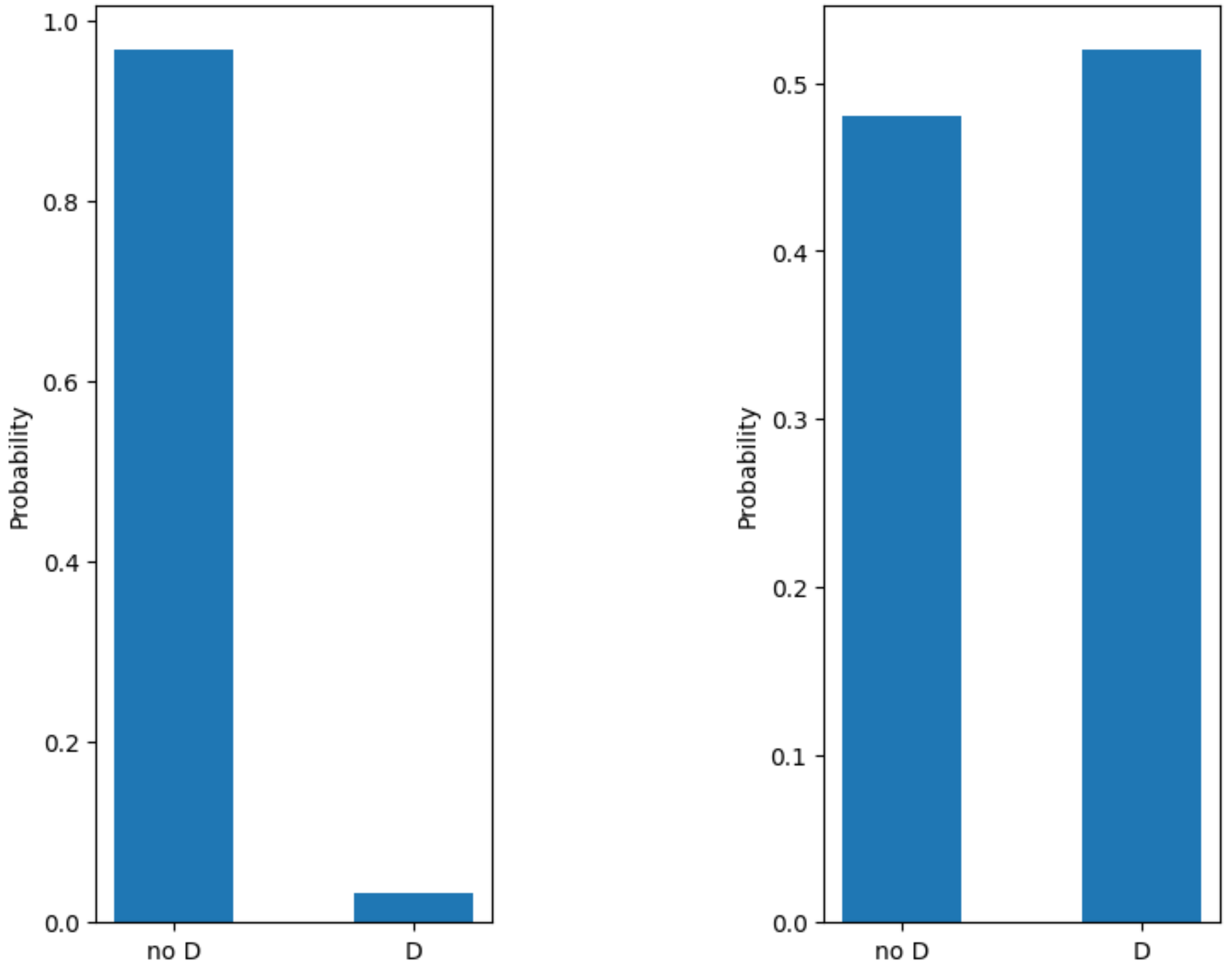


The precise value of the upper edge does not change the result much

Informative and uninformative priors

- Priors contain our information before analysing the data
- If they do, they are called informative
- If we try to find distributions that represent “no information”, these are called uninformative priors
- Intermediate case: We have some information, but not enough to fully constrain the prior (e.g. approximate scale of some parameter)

- Intuitively we would say that more even distribution of probabilities represents less information
- This should certainly be true if our knowledge is symmetric in reordering the states
- Can we come up with a more general principle?



Information content

- Idea: Define information in some result
  - Unsurprising results contain less information, thus information must decrease with the probability of the result
  - Information should add linearly $I_{1,2} = I_1 + I_2$
  - The common probability of two independent events is the product $p(x, y) = p(x)p(y)$
- This motivates the definition of the self information

$$I(x) = \log \frac{1}{p(x)}$$

- In information theory, often base 2 logarithms

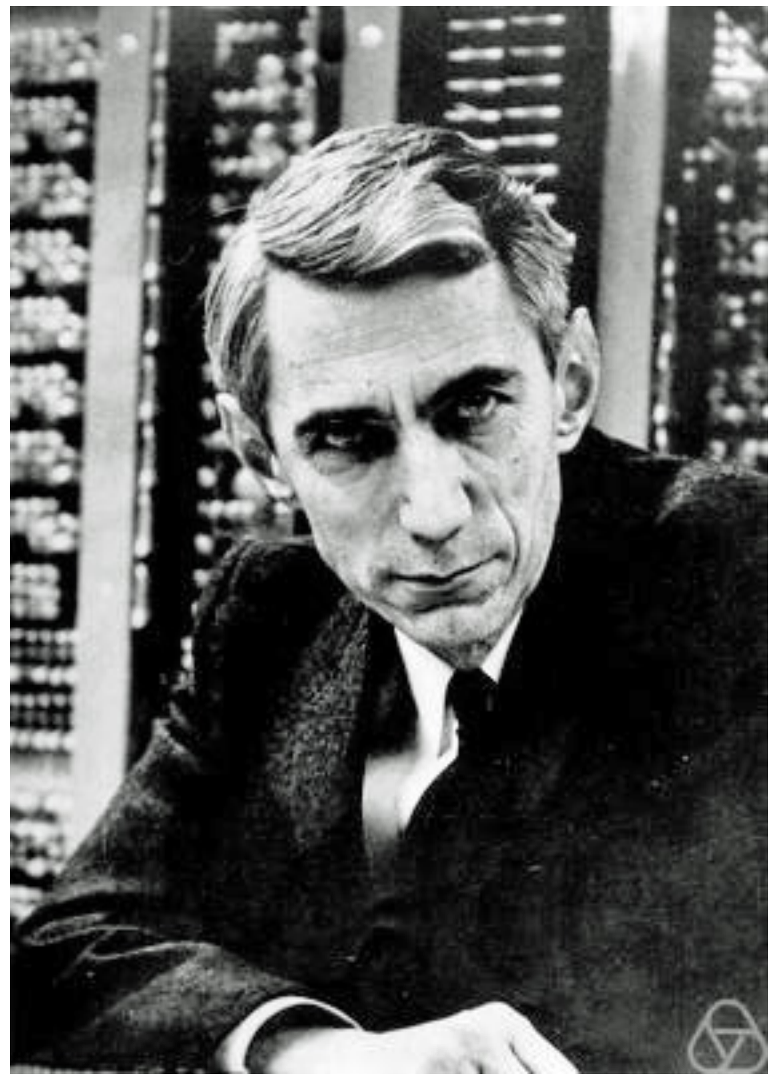
The maximum entropy principle

- How much information is contained in a distribution can be assessed by the expectation value of the information $\langle I(x) \rangle$

$$S = -\sum_i p_i \log p_i$$

- This is called the *Shannon entropy* or information entropy

- Now look for priors which maximise this quantity. These are called maximum entropy priors, MAXENT, PME, …
- Jaynes warns not to associate to much philosophical meaning here, in particular to the concept of “information”.
- For now, we just see what the result is



Claude Shannon (1916-2001)

Maximum entropy example

- We have some discrete values n and associated p_n
- Chose p_i such that $S = -\sum_n p_n \log p_n$ is maximised while $\sum_i p_i = 1$
- Can introduce Lagrange multipliers and set derivative of $-\sum_n p_n \log p_n + \lambda \left(\sum p_n - 1 \right)$ to 0

$$0 = -\log p_n - 1 + \lambda$$

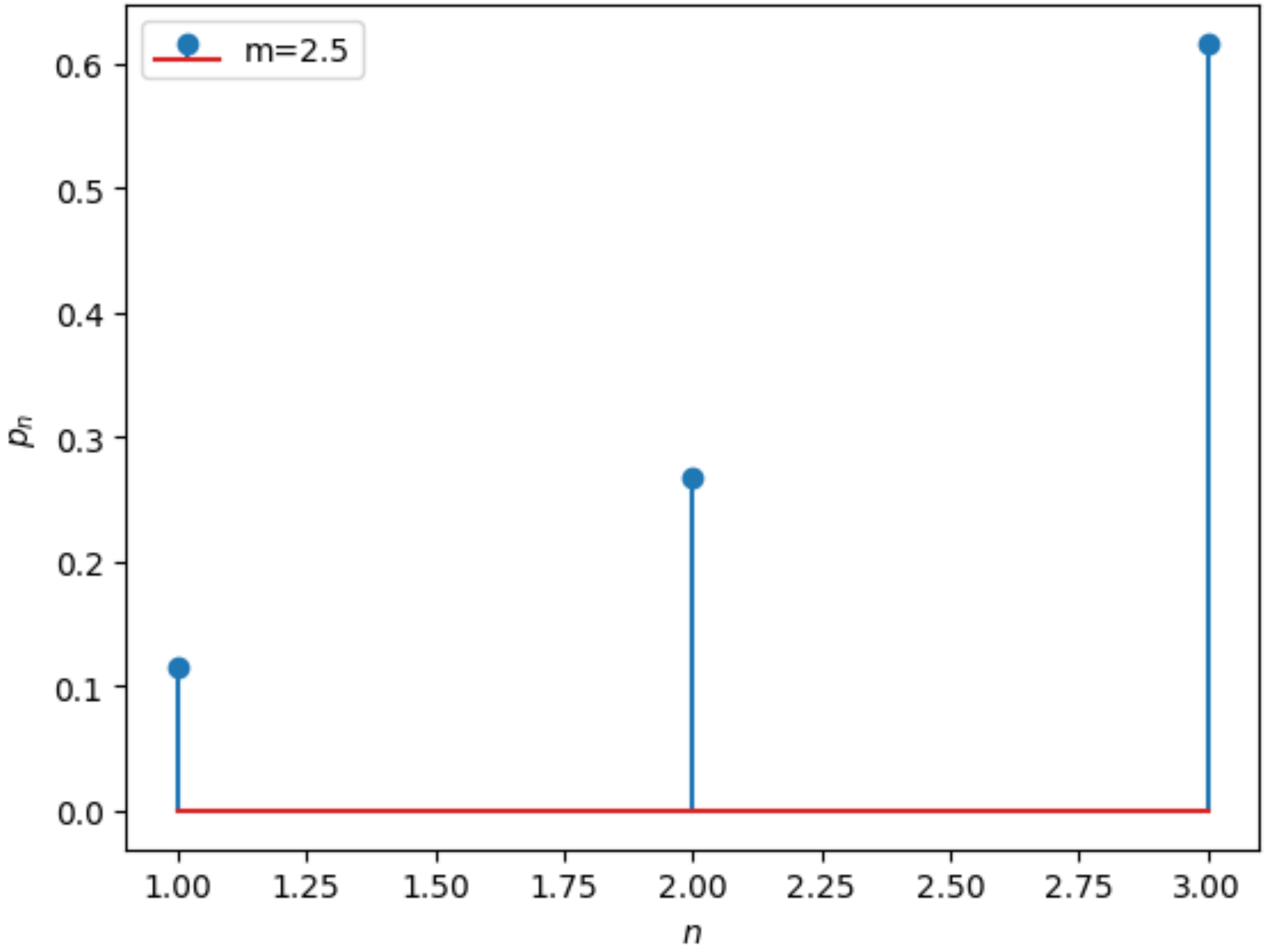
- Thus all p_n must be equal!

Maximum entropy with known average

- We assume that the n are integer values
- Our prior information is an average m, but nothing more
- Additional constraint: $\sum_n np_n = m$
- Now look at derivatives of

$$-\sum_n p_n \log p_n + \lambda \left(\sum_n p_n - 1 \right) + \vartheta \left(\sum_n \frac{np_n}{N} - m \right)$$

- $0 = -\log p_n - 1 + \lambda + n\frac{\vartheta}{N}$
- So the p_n follow an exponential law $p_n \sim \exp(\alpha n)$



Maximum Entropy for continuous distributions

- To go to continuous case, increase density of states
- Result depends on how exactly this is done - no direct translation
- However, consider another distribution with q_n all equal
- Then $S_{SJ} = -\sum_n p_n \log(p_n/q_n)$ would still be maximised
- This can be generalised as $S_c = -\int p(x)\, \log \frac{p(x)}{m(x)} \mathrm{d}x$
- Kullback-Leibler divergence / relative entropy
- $m(x)$ tells us the distribution of the continuous limit of the q_i

- If we know $m(x)$, the “uninformative” distribution without constraints, then we maximise S_c to get the corresponding distribution with constraints
- For constant $m(x)$:
  - Constraint on the mean → $p(x)$ is exponential; Constraint on mean and variance → $p(x)$ is Gaussian

Scaling priors

- If translation of a variable changes nothing in our knowledge, then $p(x)$ is constant

$$p(x) = p(x + \Delta x) \implies p(x) = \text{const}\,.$$

- If rescaling of the problem changes nothing in our knowledge, then $p(x)$ is the inverse of x

$$p(x) = \frac{1}{\alpha} p\left(\frac{x}{\alpha}\right) \implies p(x) \sim \frac{1}{x}$$

- This gives a density that is constant in the logarithm

$$p_{\log x}(\log x) = \text{const}\,.$$

Jeffreys' prior

- Jeffreys: Result should be the same independent of parametrisation
- Uncertainty depends on what can be extracted from data
- Thus prior should depend on likelihood, which has transformation properties

Jeffreys' prior (non-informative prior) for a model $L(\vec{x}|\vec{\theta})$ of the measurement:

$$\pi(\vec{\theta}) \propto \sqrt{I(\vec{\theta})} \qquad\qquad I(\vec{\theta}) = \det\left[\left\langle \frac{\partial \ln L(\vec{x}|\vec{\theta})}{\partial \theta_i} \frac{\partial \ln L(\vec{x}|\vec{\theta})}{\partial \theta_j} \right\rangle\right]$$

determinant of the Fisher information matrix

invariant under re-parameterization

expectation value evaluated by inte$\vec{x}$grating over all possible results

Examples:

PDF parameter	Jeffreys' prior
Poissonian mean μ	p(μ) ∝ 1/√μ
Gaussian mean μ	p(μ) ∝ 1

Should our prior information depend on what measurement we do?

Jeffreys' prior: Example

Exponential distribution: $$L(t \mid \tau) = \frac{1}{\tau} e^{-t/\tau}$$

Jeffreys' prior: $$\pi(\tau) \propto \sqrt{I(\tau)} = \sqrt{\mathrm{E}\left[\left(\frac{d}{d\tau} \ln L(t \mid \tau)\right)^2\right]}$$

$$\frac{d}{d\tau} \ln L(t|\tau) = -\frac{1}{\tau} + \frac{t}{\tau^2}$$

$$E\left[\left(\frac{t}{\tau^2} - \frac{1}{\tau}\right)^2\right] = E\left[\left(\frac{t-\tau}{\tau^2}\right)^2\right] = \frac{1}{\tau^4} V[t] = \frac{\tau^2}{\tau^4} = \frac{1}{\tau^2}$$

$\rightsquigarrow \quad \pi(\tau) \propto \frac{1}{\tau}$ (prior distribution)

Conjugate priors - examples

- Conjugate prior: posterior=prior*likelihood is from the same class of functions
- For ease of calculation
- A flat distribution is a special case of all
  - e.g. Gaussian with $\sigma \to \infty$

Likelihood	Conjugate Prior
Binomial	Beta
Poisson	Gamma
Gaussian	Gaussian
Exponential	Gamma

Summary priors

- The problem of assigning priors is quite difficult
- Some approaches try to remove the discussion, giving rules for which distribution to take → objective Bayes
- In many practical applications the prior is obvious
- In some of the most interesting ones it is not