Statistical Methods in Particle Physics

1. Basic Concepts

Heidelberg University, WS 2023/24

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

Nobel laureates in Chemistry, Physics, Medicine, and Economics

Rank	Entity	Laureates/ 10 million
—	*Faroe Islands*	202.065
1	Saint Lucia	55.659
2	Luxembourg	33.880
3	Switzerland	24.967
4	Austria	20.567
5	Denmark	17.378
6	Sweden	17.029
7	United Kingdom	16.373
8	Norway	14.944
9	Hungary	12.385 (17,7)
10	Germany	11.180
11	Netherlands	11.121
12	United States	10.711
13	Israel	10.647
14	Cyprus	8.410
15	New Zealand	6.316
16	France	5.979
17	Canada	5.953
—	*European Union*[9]	5.574
18	Finland	5.413

- What makes the Faroe Islands and St. Lucia so efficient at producing great scientists?

Nobel laureates in Chemistry, Physics, Medicine, and Economics

Rank	Entity	Nobel laureates[1]	Population [2]	Laureates/ 10 million
—	*Faroe Islands*	1	49,489	202.065
1	Saint Lucia	1	179,667	55.659
2	Luxembourg	2	590,321	33.880
3	Switzerland	22	8,811,524	24.967
4	Austria	19	8,751,820	20.567
5	Denmark	10	5,754,356	17.378
6	Sweden	17	9,982,709	17.029
7	United Kingdom	109	66,573,504	16.373
8	Norway	8	5,353,363	14.944
9	Hungary	12 (17)[8]	9,688,847	12.385 (17,7)
10	Germany	92	82,293,457	11.180
11	Netherlands	19	17,084,459	11.121
12	United States	350	326,766,748	10.711
13	Israel	9	8,452,841	10.647
14	Cyprus	1	1,189,085	8.410
15	New Zealand	3	4,749,598	6.316
16	France	39	65,233,271	5.979
17	Canada	22	36,953,765	5.953
—	*European Union*[9]	247	443,123,600	5.574
18	Finland	3	5,542,517	5.413

- What makes the Faroe Islands and St. Lucia so efficient at producing great scientists?
- Consider not just estimate, but also uncertainty

1.0 Organization

Contents

Learning goals and required knowledge

This course is a natural follow-up to PEP4 for **Bachelor students** interested in Particle Physics. **Master students** are invited to attend this lecture in parallel or after the Particle Physics course.

Learning goals

- Get to know and apply the toolbox of statistical methods used in particle physics
- Understand error bars and confidence limits as reported in publications
- Solid understanding of maximum likelihood and least squares fits
- From measurement to message: which conclusion can you draw from your data (and which not)?
- Learn to apply machine learning methods

Required knowledge

- Basic understanding of experimental particle physics (as taught in the bachelor's course)
- Basic knowledge of python is helpful

Practical information

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

- Website
  - https://uebungen.physik.uni-heidelberg.de/vorlesung/20232/1733
- Lecture
  - Thursday 16:15-17:45
  - Break?
- Exam
  - There will be a written exam at the end of the semester
  - Refers to contents of lectures and exercises
  - 60% of the points of the homework sheets required to be eligible to write the exam

Tutorials

- Mondays, 16:15 – 17:45
- Weekly problem sheets, to be handed in (uploaded) on Thursday before 10:00
- Includes programming exercises - Python using Jupyter Notebooks
- First Tutorial, 23.10., introduction to Python
- Solutions ideally in groups of 2

- See next slide for detailed schedule

Statistical methods in particle physics — WS 2023/24

Week	Date Monday	Date Friday	Tutorial Mon., 16:15-17:45	Lecture Thu., 16:15-17:45	Hand out sheet Thu., 18:00	Hand in sheet Thu., 12:00	Comment
1	16.10.23	20.10.23	—	Lecture 1	Sheet 1	—	
2	23.10.23	27.10.23	Tutorial 1 / Python	Lecture 2	Sheet 2	Sheet 1	
3	30.10.23	03.11.23	Tutorial 2 / Sheet 1	Lecture 3	Sheet 3	Sheet 2	
4	06.11.23	10.11.23	Tutorial 3 / Sheet 2	Lecture 4	Sheet 4	Sheet 3	
5	13.11.23	17.11.23	Tutorial 4 / Sheet 3	Lecture 5	Sheet 5	Sheet 4	
6	20.11.23	24.11.23	Tutorial 5 / Sheet 4	Lecture 6	Sheet 6	Sheet 5	
7	27.11.23	01.12.23	Tutorial 6 / Sheet 5	Lecture 7	Sheet 7	Sheet 6	
8	04.12.23	08.12.23	Tutorial 7 / Sheet 6	Lecture 8	Sheet 8	Sheet 7	
9	11.12.23	15.12.23	Tutorial 8 / Sheet 7	Lecture 9	Sheet 9	Sheet 8	
10	18.12.23	22.12.23	Tutorial 9 / Sheet 8	—	—	—	
11	08.01.24	12.01.24	Tutorial 10 / test exam	Lecture 10	Sheet 10	Sheet 9	
12	15.01.24	19.01.24	Tutorial 11 / Sheet 9	Lecture 11	—	Sheet 10	
13	22.01.24	26.01.24	Tutorial 12 / Sheet 10	Lecture 12			
14	29.01.24	02.02.24	Tutorial 13 /				Lernwoche
15	05.02.24	09.02.24					Klausurwoche

Useful books

- G. Cowan, Statistical Data Analysis
- L. Lista, Statistical Methods for Data Analysis in Particle Physics
- Behnke, Kroeninger, Schott, Schoerner-Sadenius: Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods
- C. Pruneau, Data Analysis Techniques for Physical Scientists
- R. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences
- Bohm, Zech, Introduction to Statistics and Data Analysis for Physicist [available online]
- Blobel, Lohrmann: Statistische Methoden der Datenanalyse (in German), [free ebook]
- L. Lyons:
  Statistics for Nuclear and Particle Physicists (Cambridge University Press)
- F. James, Statistical Methods in Experimental physics
- W. Metzger, Statistical Methods in Data Analysis [available online]

Further Material

- Glen Cowan: http://www.pp.rhul.ac.uk/~cowan/stat_course.html
- Scott Oser: http://www.phas.ubc.ca/~oser/p509/
- Terascale Statistics School: https://indico.desy.de/indico/event/25594/other-view?view=standard
- Particle Data Group reviews on Probability and Statistics
  - https://pdg.lbl.gov/2020/reviews/rpp2020-rev-probability.pdf
  - https://pdg.lbl.gov/2020/reviews/rpp2020-rev-statistics.pdf

- E.T. Jaynes, Probability Theory, the Logic of Science - Bayesian view
- Nate Silver, The Signal and the Noise - Statistics and Probability in the “real world”

1.1 Introduction

Why bother with statistical methods?

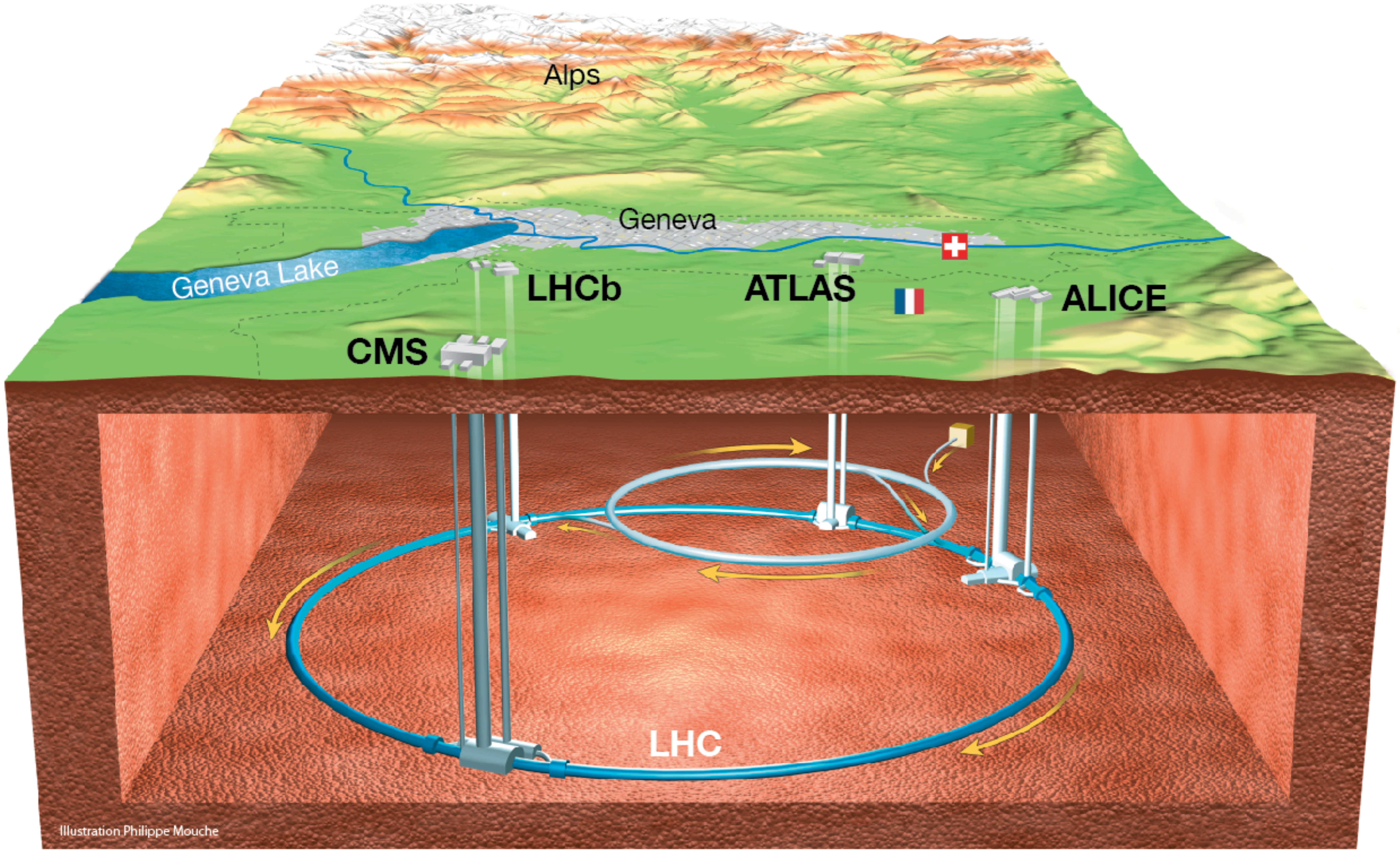
Grand Forks, North Dakota (1997)

Levee height: 50 ft Flood prediction: 49 ft ±9 ft

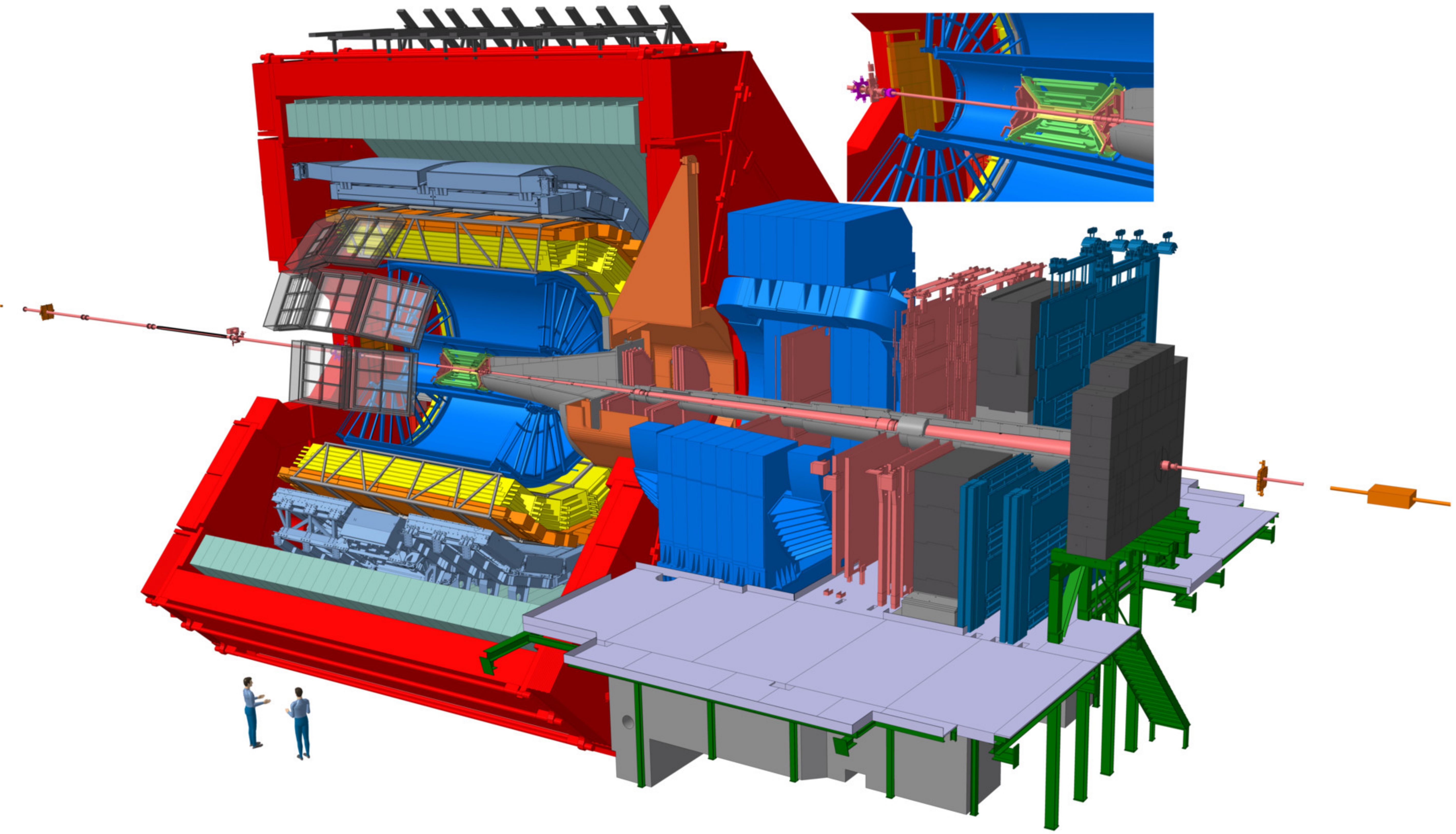
Grand Forks, North Dakota (1997)

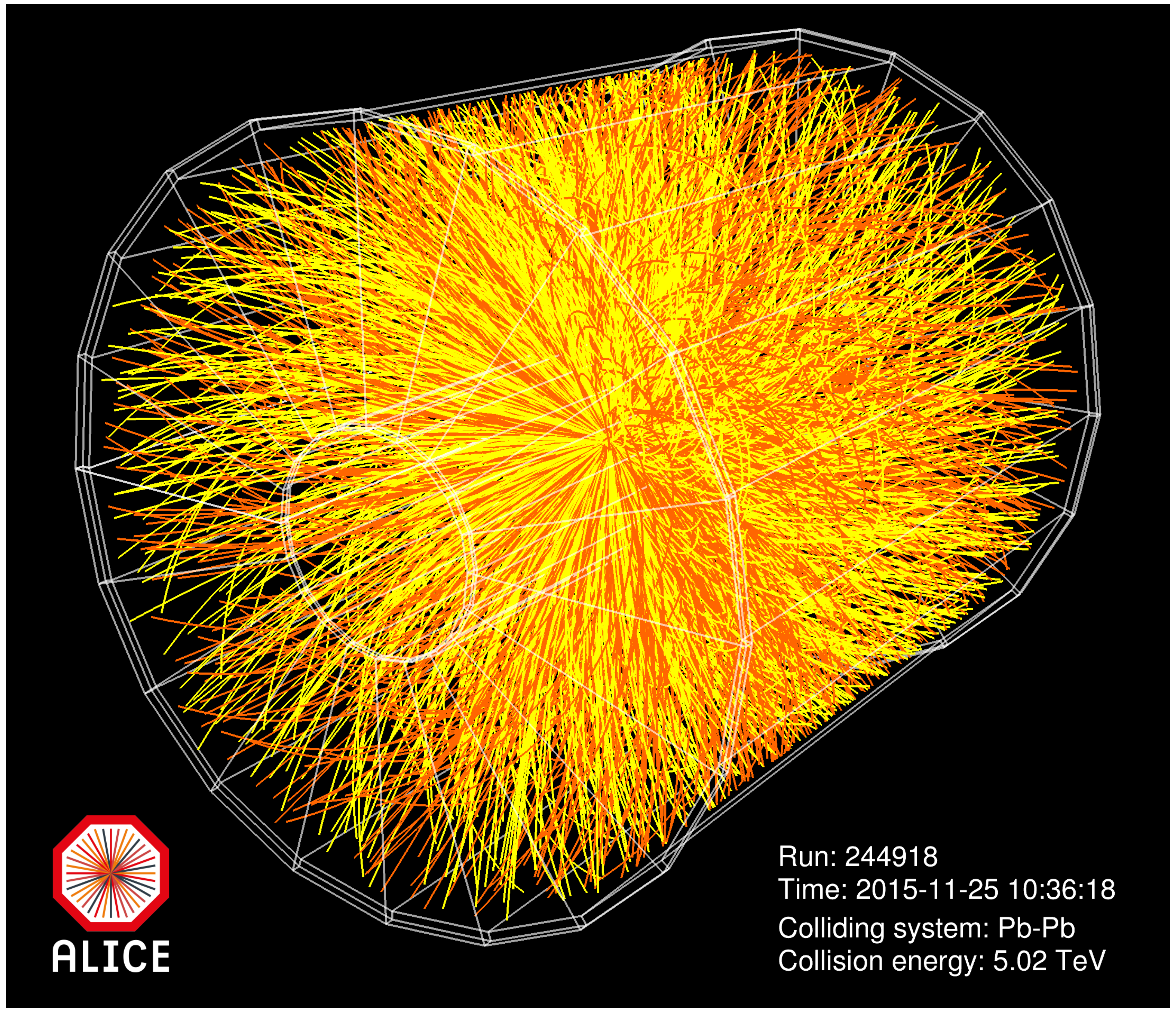
Levee height: 50 ft   Flood prediction: 49 ft ± 9 ft

The CERN LHC


Alps
Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
LHC
Illustration Philippe Mouche

ALICE

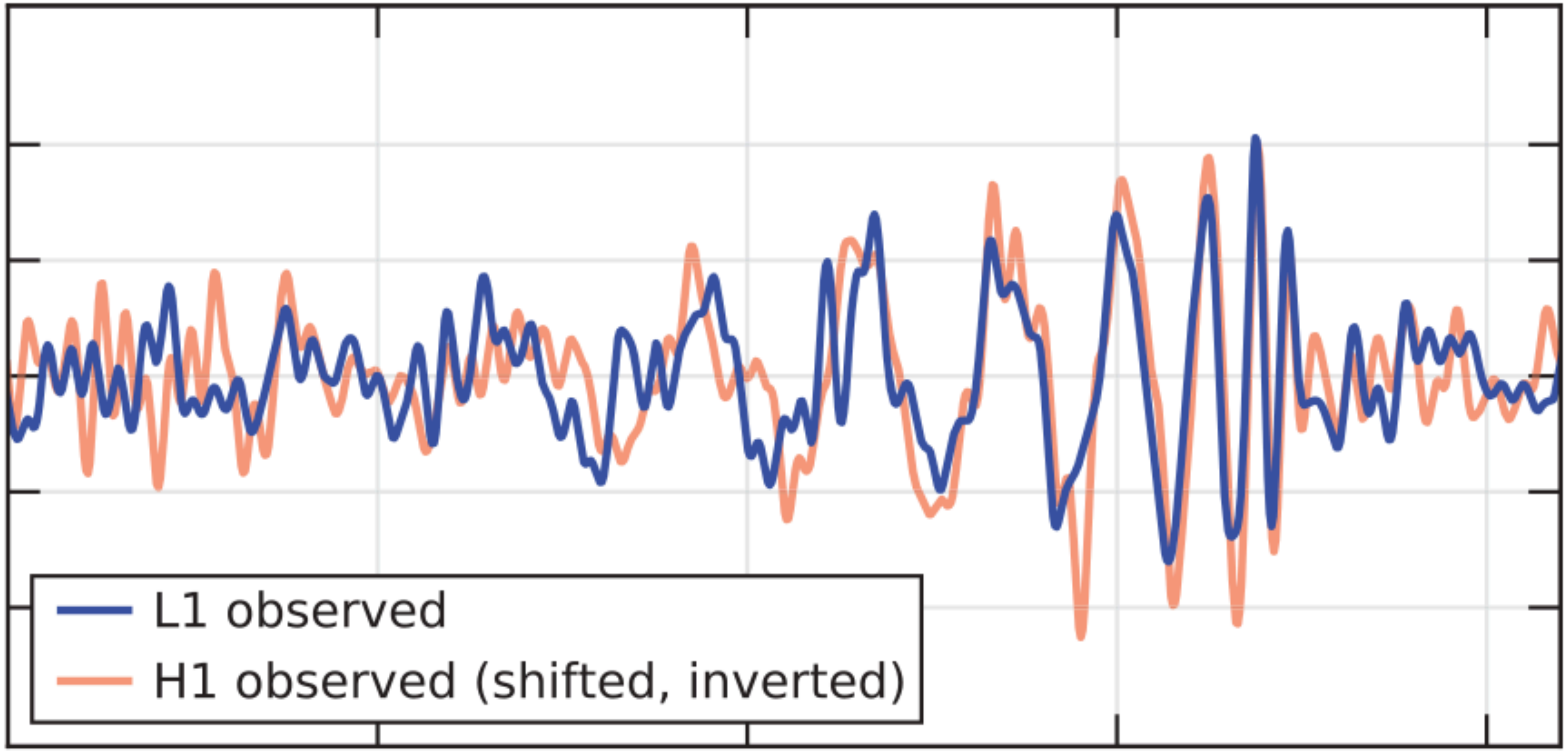




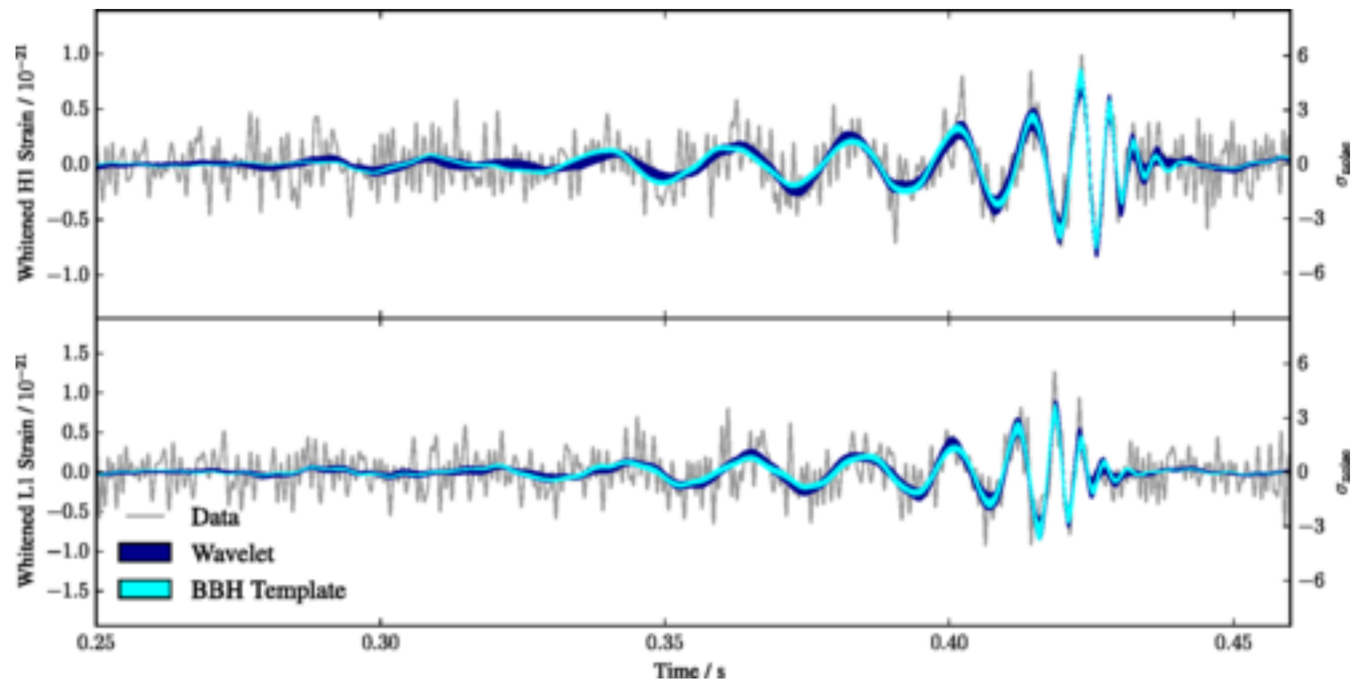
- Lead nuclei with $\gamma \approx 1000$
- Many 1000s of particles per collision
- How to extract physics from that?

Why bother with statistical methods?

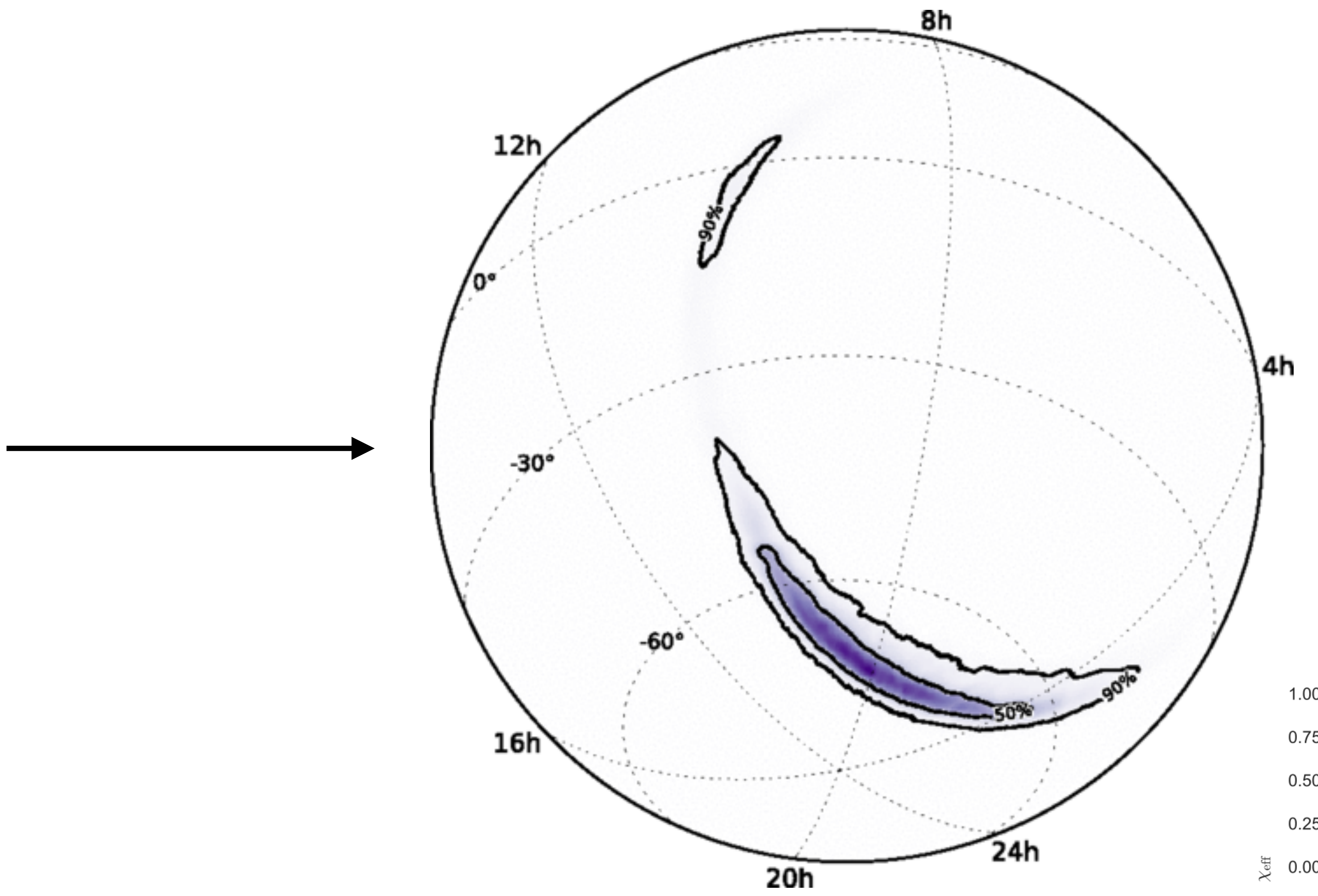
14.09.2015, Livingston and Hanford

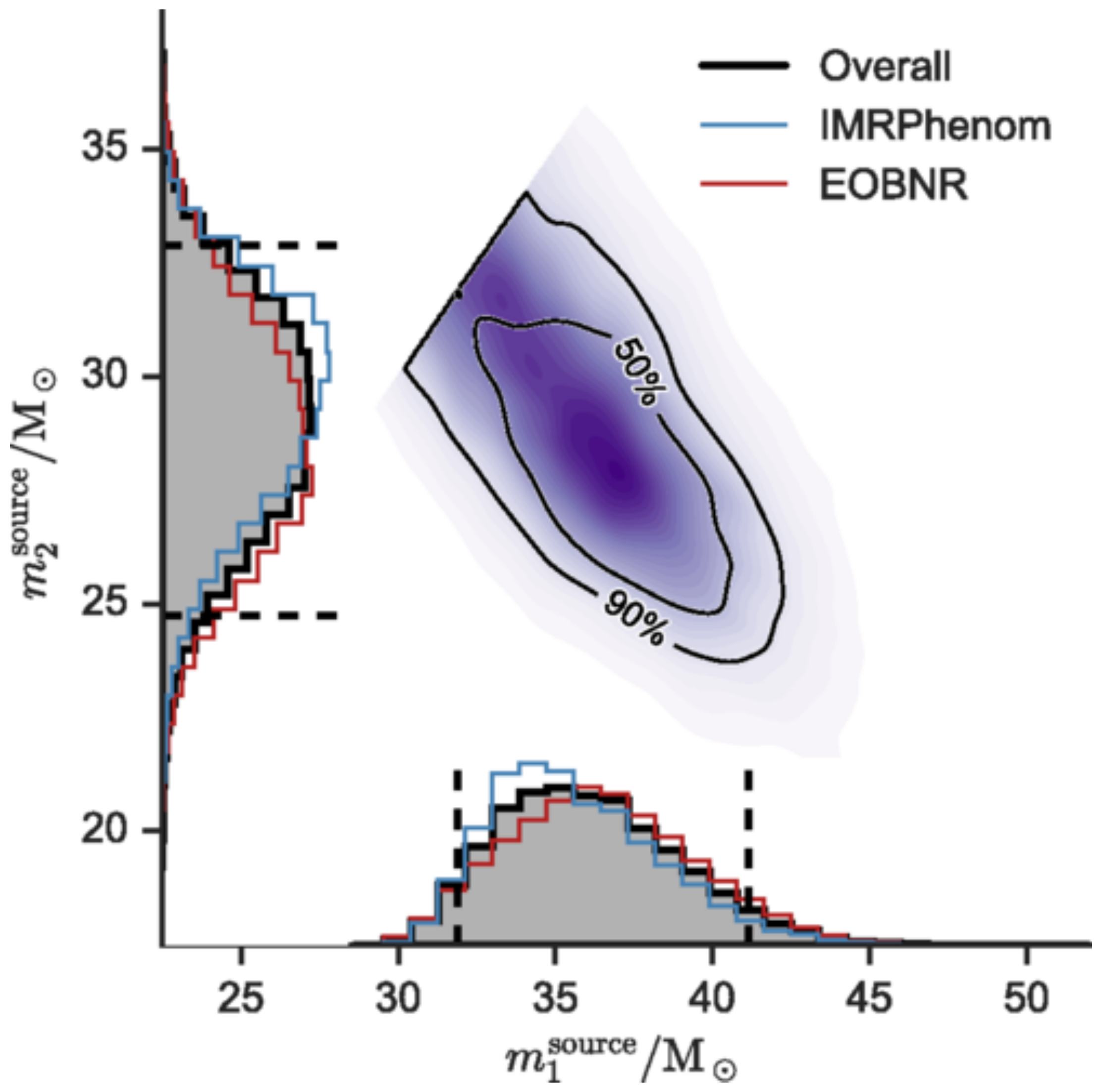


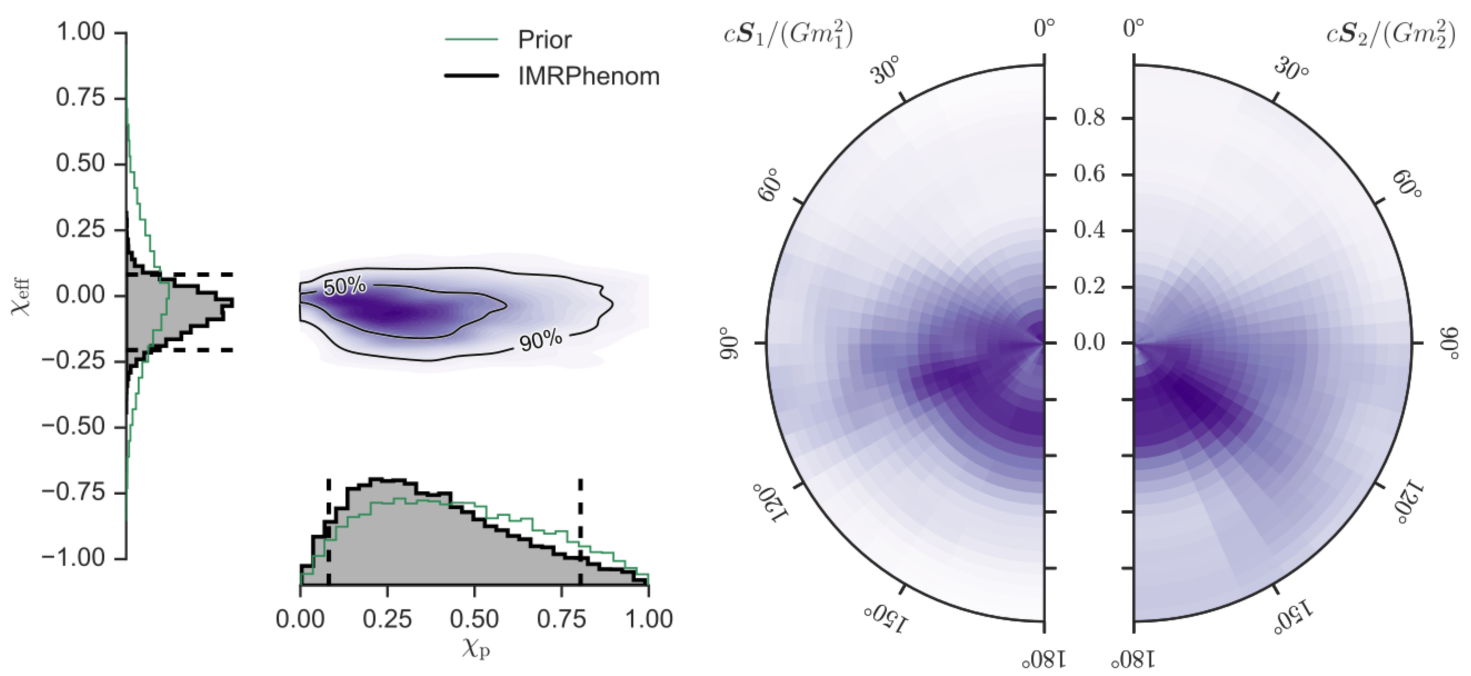
Why bother with statistical methods?



LIGO GW measurement



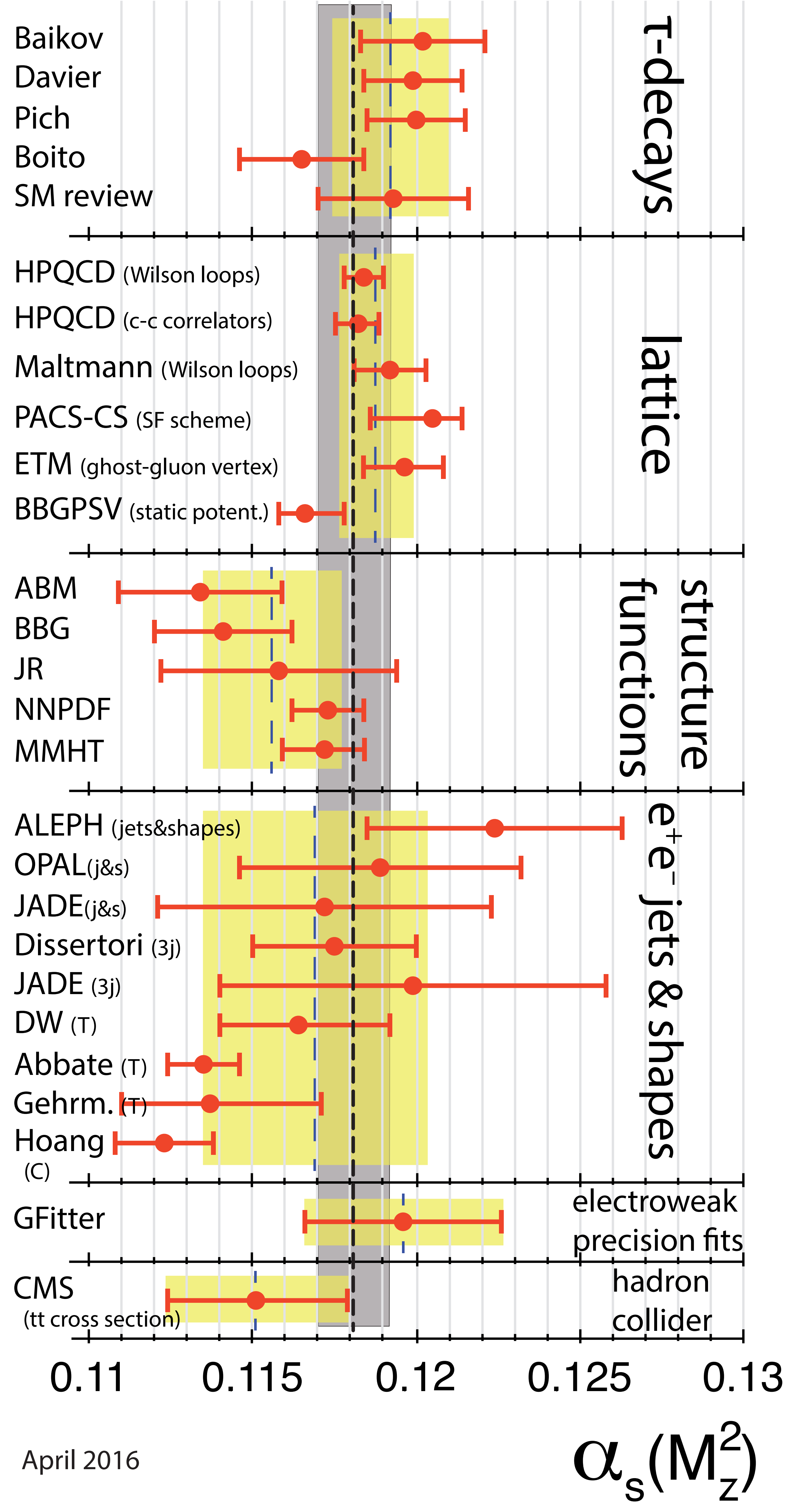




How to connect a measurement with the knowledge it provides?

Error and Uncertainty

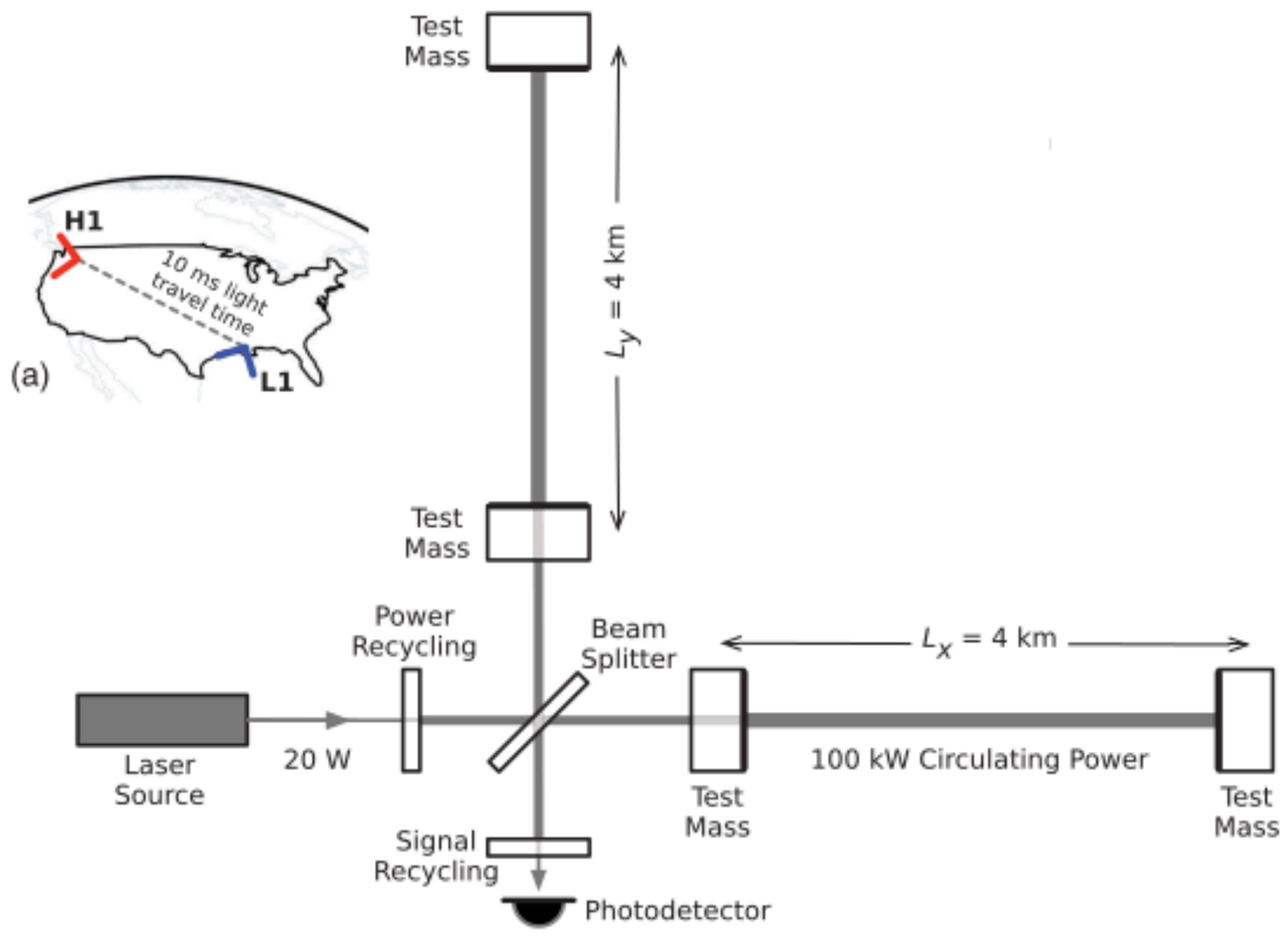
- “Error calculation”, “Error bars”, “Measurement error” - frequently used
- Also frequently used: Uncertainty
- Definition for this lecture:
- **Error** of a measurement**:** The difference between the true and measured value
- **Uncertainty:** Available information about the difference between true and measured value - usually described as an interval (e.g. $a \pm \Delta_a$)

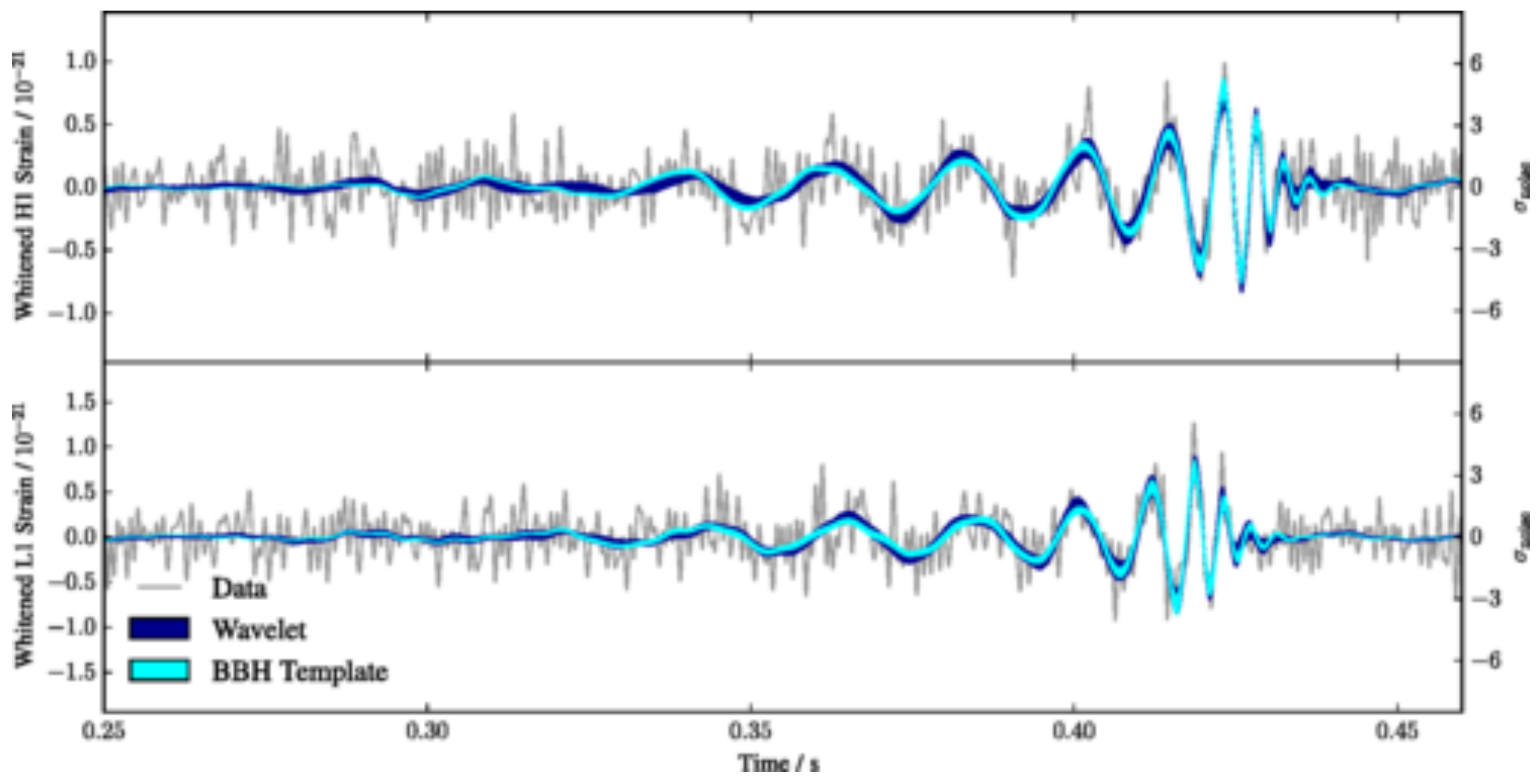


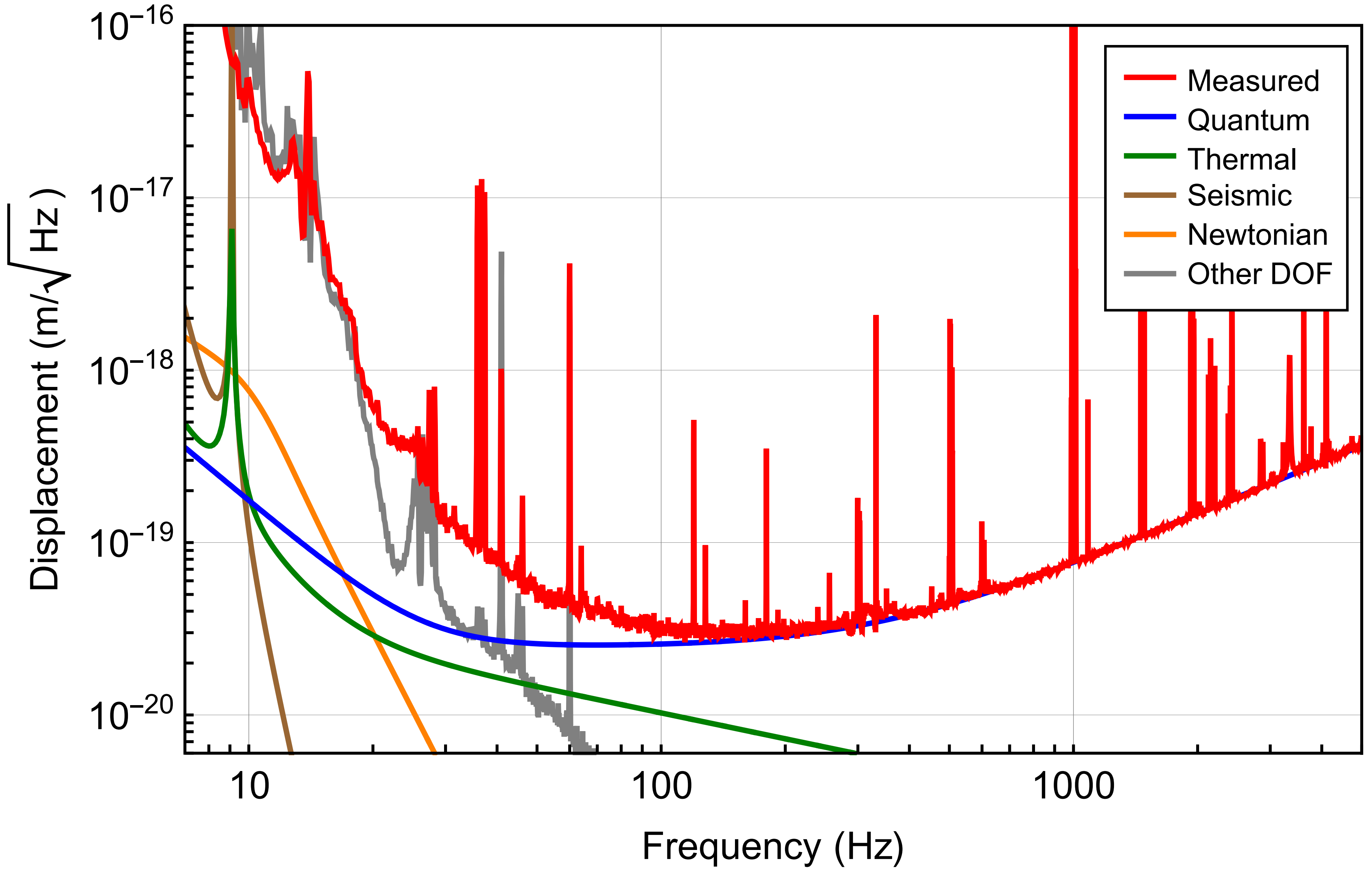
The Review of Particle Physics (2017), Particle Data Group

Examples for sources of uncertainty: LIGO

- Quantum fluctuations in the photon number
- Fluctuations in pressure of photons on the mirrors
- Vibration in wires
- Anthropogenous noise
- Brownian motion of detector setup
- Alternating current etc.





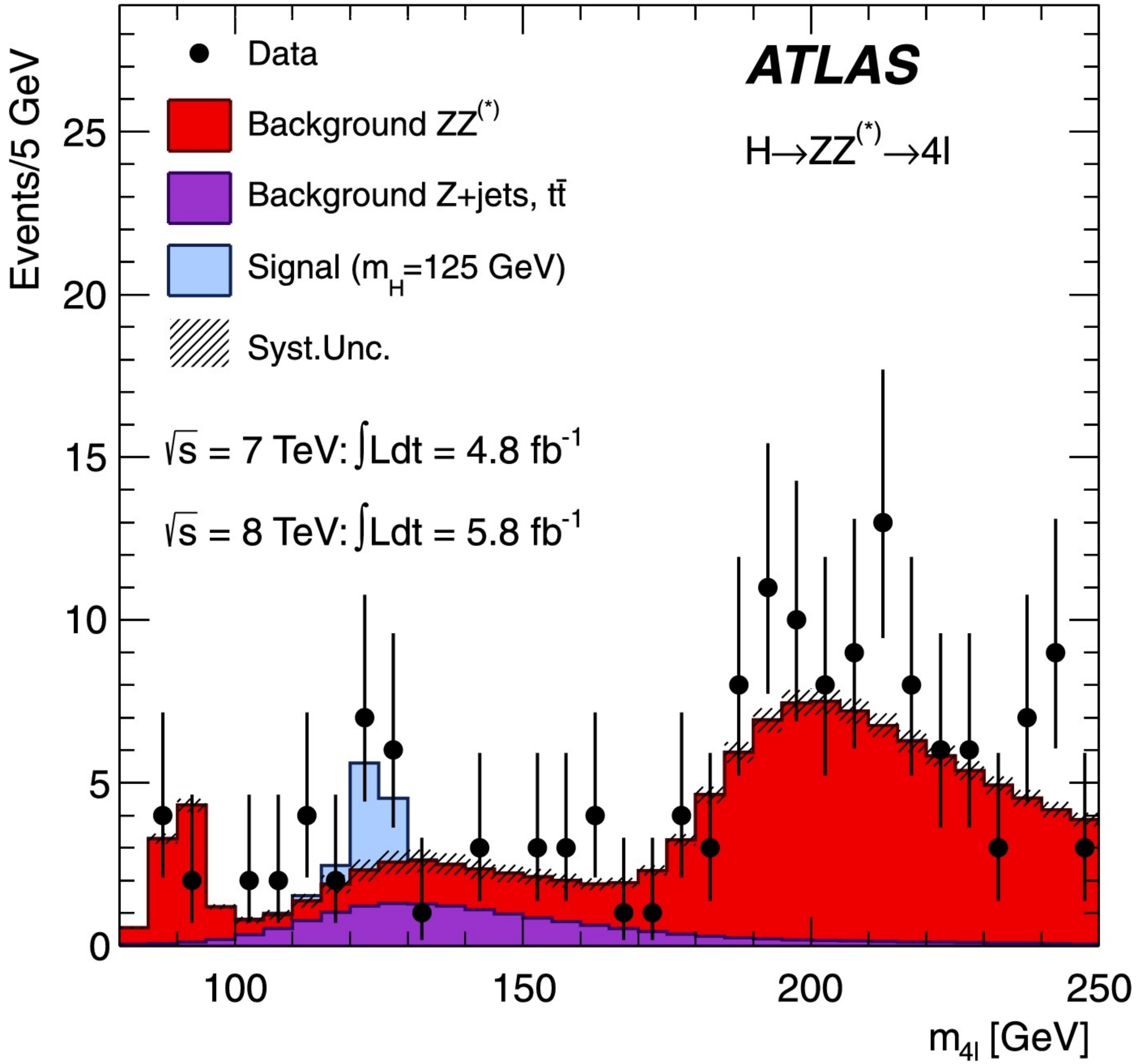


GW150914: The Advanced LIGO Detectors in the Era of First Discoveries, LIGO Collaboration PRL 116, 131103 (2016)

Examples for sources of uncertainty: ATLAS

- Fluctuations in the measured values
- Uncertainty of the expected background from other sources

→Uncertainties can come from quantum processes, classical processes, imperfect knowledge, etc.



Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, ATLAS Collaboration

Probability

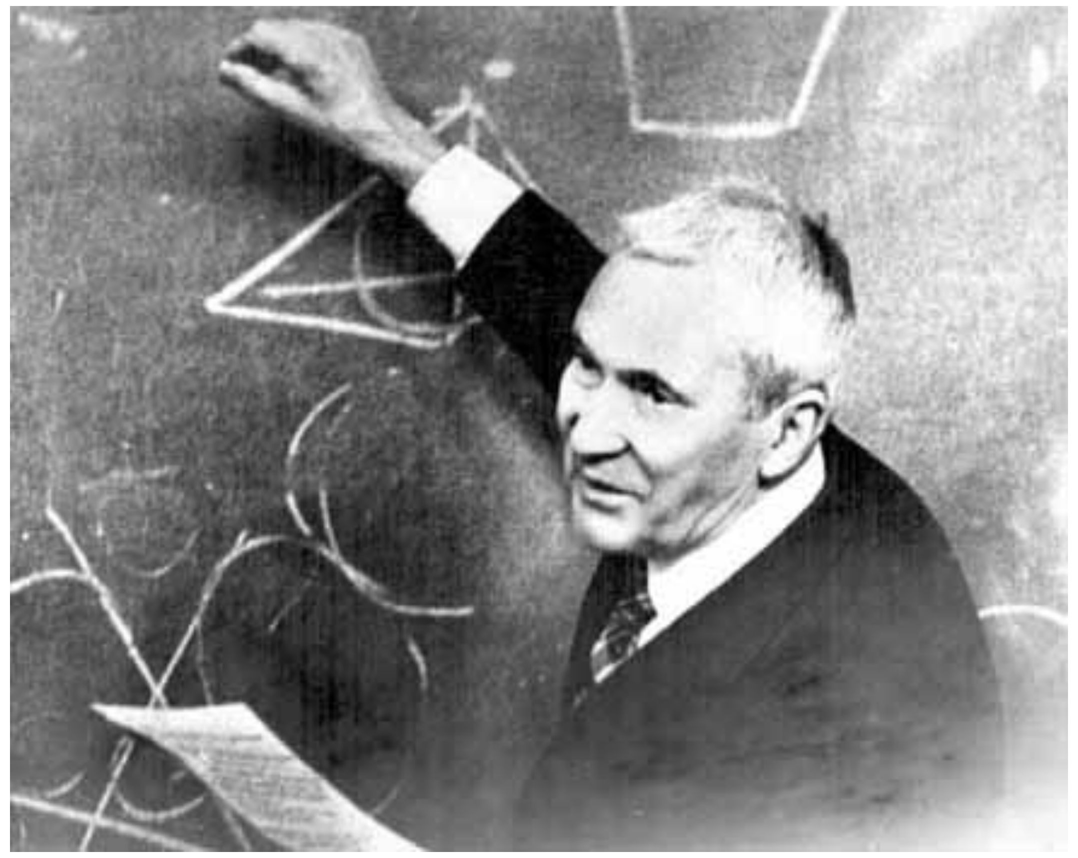
Uncertainty and Probability

Consider the following statements:

- I will probably manage to finish the project by the end of the week.
- The probability of two dice throws each yielding "6” is 1/36.
- The extinction of the dinosaurs was probably caused by an asteroid impact.
- The probability of rain for tomorrow is 75%.
- The true value is probably in the interval $a \pm \sigma_a$.
- The patient probably has the flu.
- You will probably not win the lottery this week.
- The probability for Germany to become European Champion in 2024 is 10%.

The term *probability* is used in a variety of ways.

Mathematical definition of probability



Kolmogorov, 1933

Let A be an event. Then probability is a number obeying three conditions, the *Kolmogorov axioms*:

1. $P(A) \geq 0$ (non-negative real number)
2. $P(S) = 1$, where S is the set of all A, the sample space
3. $P(A \cup B) = P(A) + P(B)$ for any A, B which are exclusive, i.e., $A \cap B = 0$

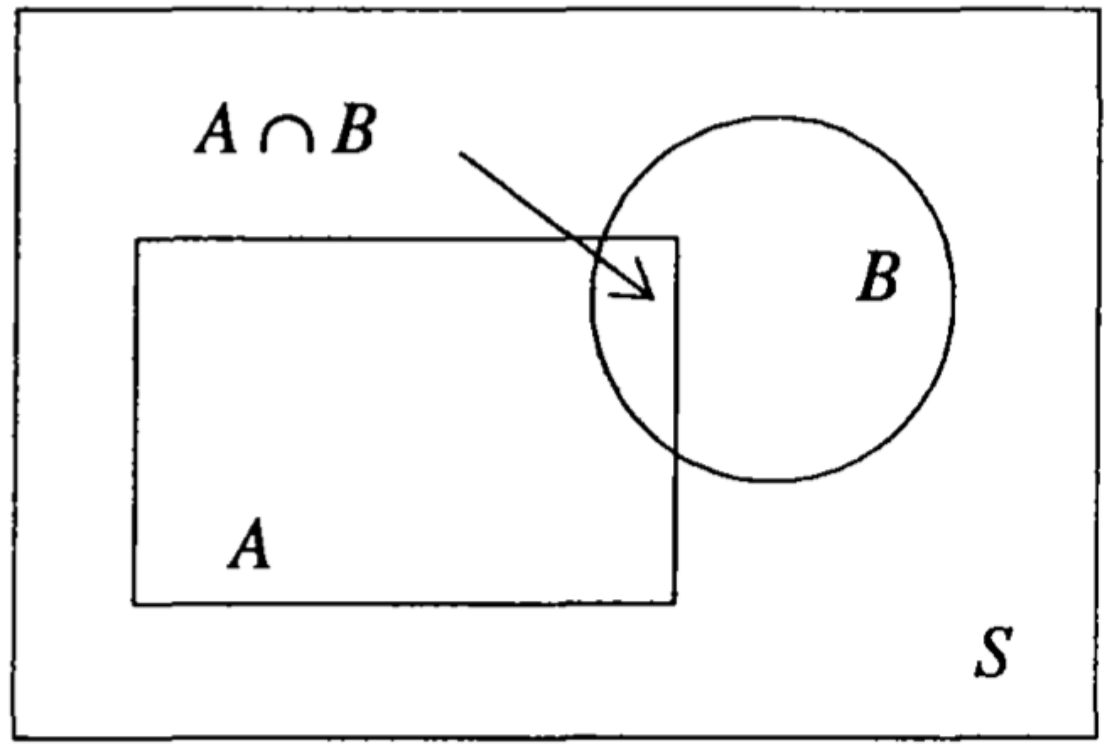
From these axioms further properties can be derived, e.g.:

$P(\bar{A}) = 1 - P(A)$

$P(\varnothing) = 0$

if $A \subset B$ then $P(A) \leq P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$



- Does not assign any meaning to "P" - Not useful for physics.
- But useful for proving theorems of probability

Classical Probability

Assume symmetry between a number N of possible outcomes

The probability of one outcome is then $\frac{1}{N}$

Probability of a favourable outcome is then just counting. E.g. 18 red numbers out of 37 means:

$$P(\text{red}) = \frac{18}{37}$$

Not helpful for physics - typically no symmetry between all cases

Frequentist Interpretation of Probability

(Occasionally also referred to as “classical”.)

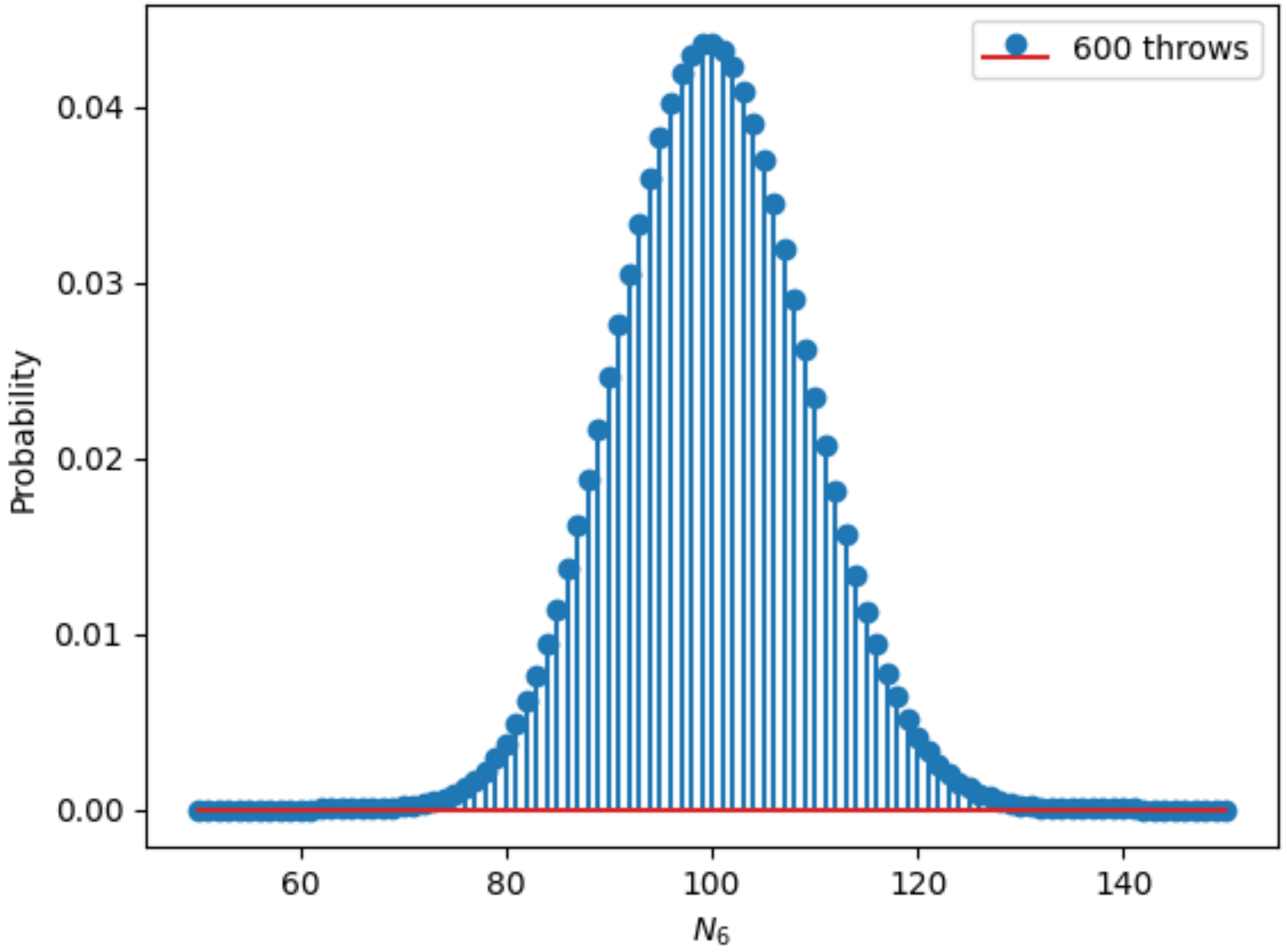
- Take some process with a classical probability p for a success
- Make sure it is possible to make an independent repetition of this
- When repeating a large number of times, there is a high probability that

$$\frac{N_{\text{success}}}{N_{\text{tries}}} \approx p$$

- This can be turned around to *define* probability as

$$p \equiv \lim_{N\to\infty} \frac{N_{\text{success}}}{N}$$

→ Now probabilities are defined without needing any symmetry; this can be used in physics!

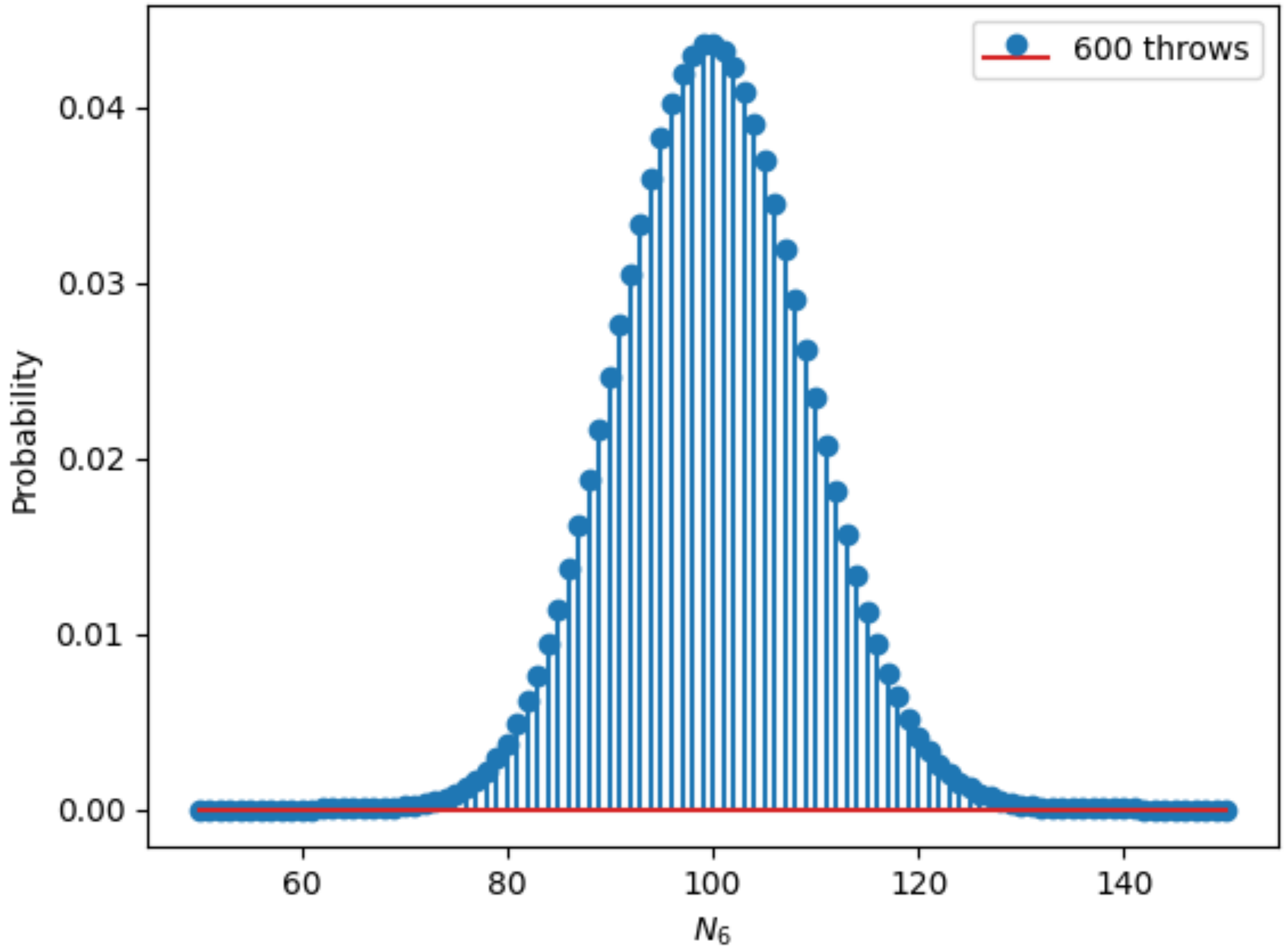


Frequentist Interpretation of Probability

$$p \equiv \lim_{N\to\infty} \frac{N_{\text{success}}}{N}$$

Some caveats:

- This definition requires a random process as prerequisite
  - Probabilities *only* exist for random processes
  - Tricky to define “random” without using probability
- Only works if process can be independently repeated many times



Things that are not probabilities in the frequentist sense:

- Probability of rain tomorrow
- Probability for sports team to win the next game
- Probability for a hypothesis to be true
- Probability for a physical parameter to be in some range

Bayesian Interpretation of Probability

A probability expresses a state of knowledge

For example:

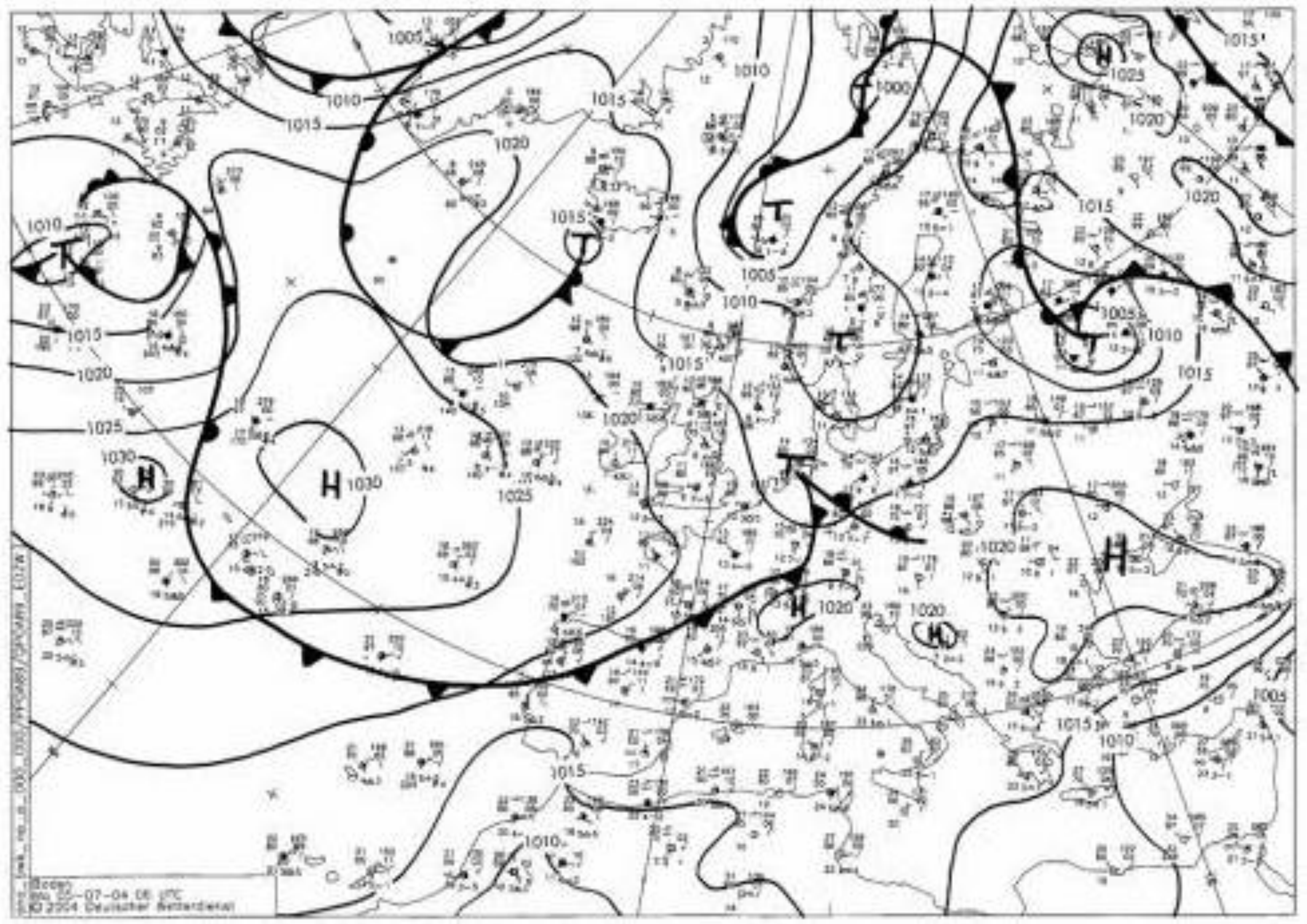
$p($"There will be rain tomorrow"$)$
$p($"Dark matter is made of WIMPS"$)$

Expresses degree of belief

To quantify: Consider the odds at which a bet would be rational

New information changes (updates) the state of knowledge

Since different people might have different states of knowledge about something, probabilities are *subjective*



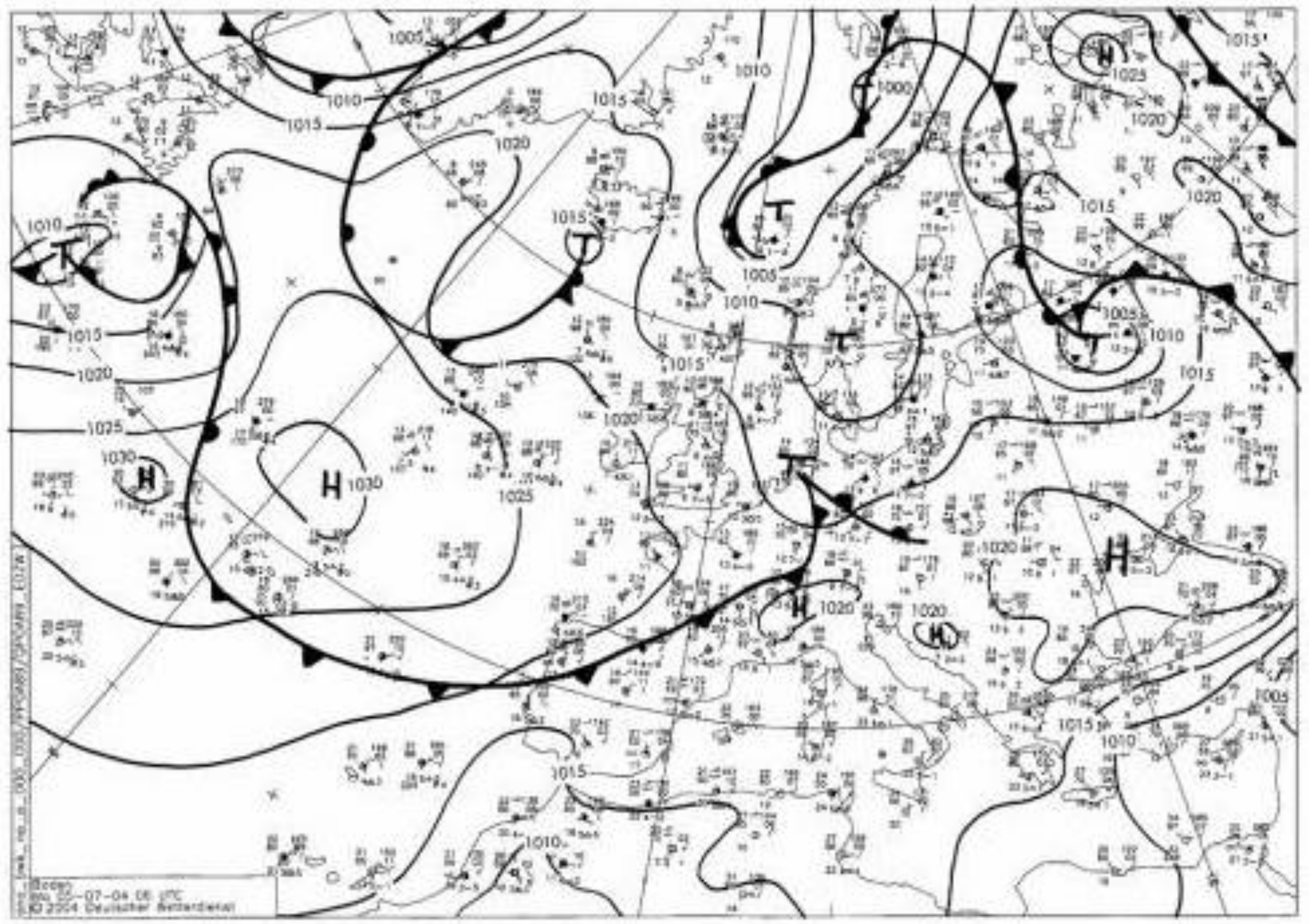
Bayesian Interpretation of Probability

Under symmetric conditions: $p = 1/N$ is only rational assessment of probability → contains classical definition

For a random process with known average frequency, $p = \lim_{N\to\infty} \frac{N_{\text{success}}}{N}$ is the only rational assessment of probability → contains frequentist definition

For other cases, we have a *prior* probability (before all data) and this is modified as we learn more information

→ it is not so obvious what this prior should be



Problem of old evidence: What happens if we only formulate the hypothesis after the evidence has been found?

Should we always draw the same conclusions from the same data?

I tell you that this die is weighted and 6 will appear much more often. You throw the die 10 times and “6” appears five times.

I tell you that with my psychic powers I can often predict dice rolls. You throw a die 10 times and I correctly predict the result five times.

“Extraordinary claims require extraordinary evidence.”
- Carl Sagan

Some definitions

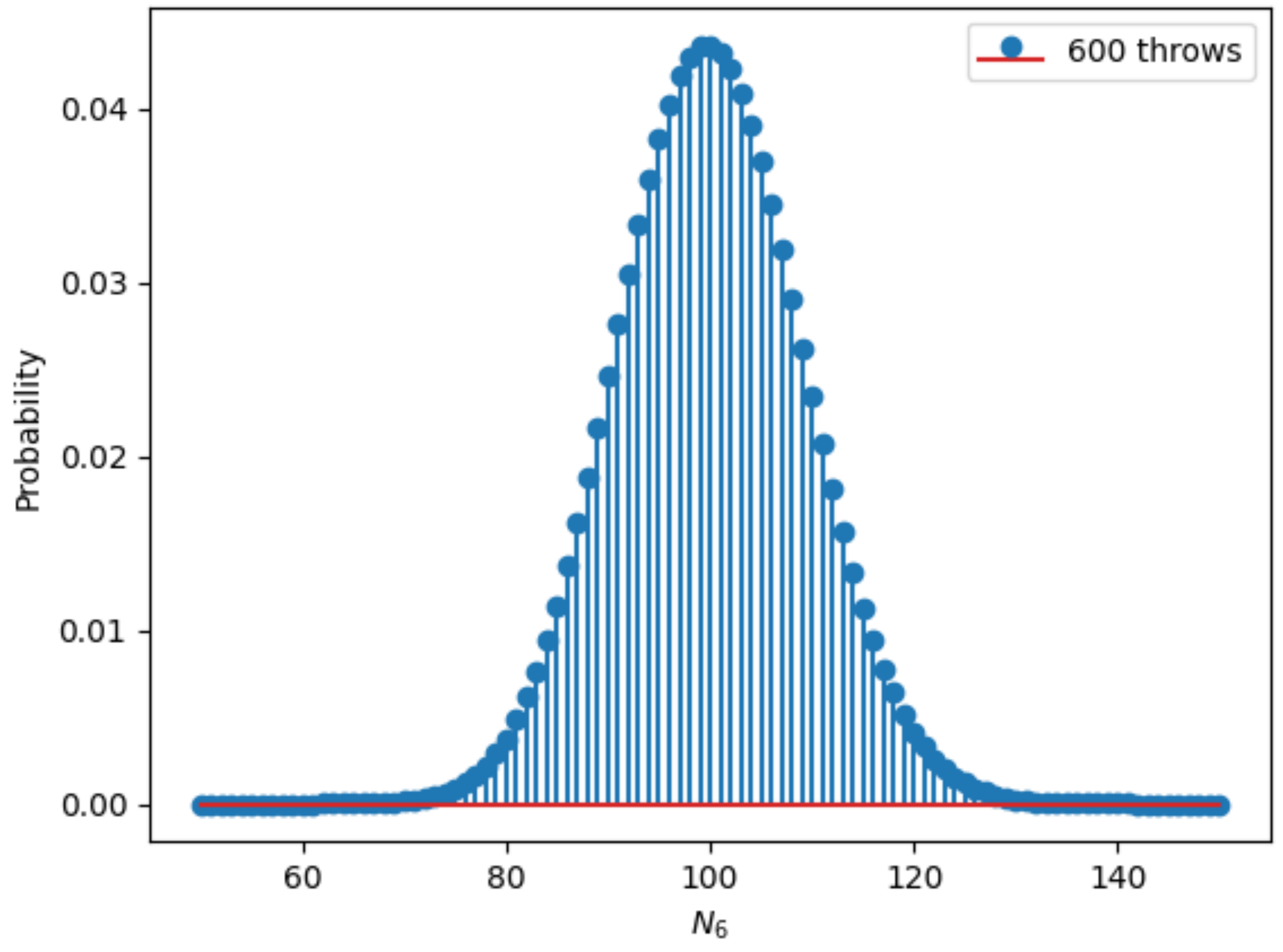
Probability density functions

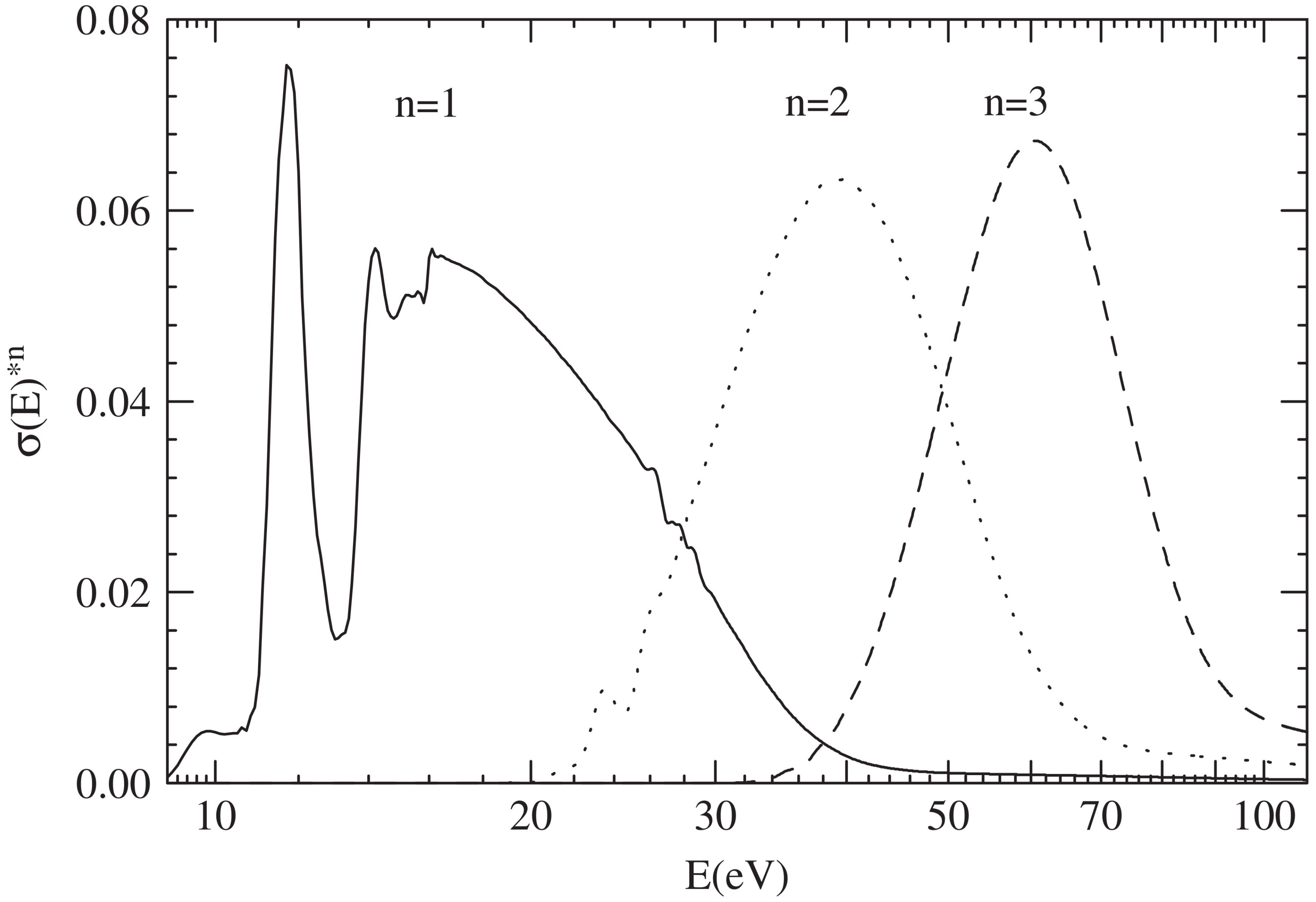
- Discrete probability distributions - probabilities for for all possible E_i events out of a set of possible events S
- Probabilities sum to 1: $\sum_i p(E_i) = 1$
- Can generalise to continuous case as usual. E.g. for random variable x:

$\int_{x_1}^{x_2} p(x)\ \mathrm{d}x$ is the probability for x to be in the range $[x_1, x_2]$

$$\int p(x)\ \mathrm{d}x = 1$$

- $p(x)$ is called the *probability density function* (pdf)
- From now on we will always use p to signify probabilities: $p(A)$ for discrete and $p(\lambda)$ for continuous cases



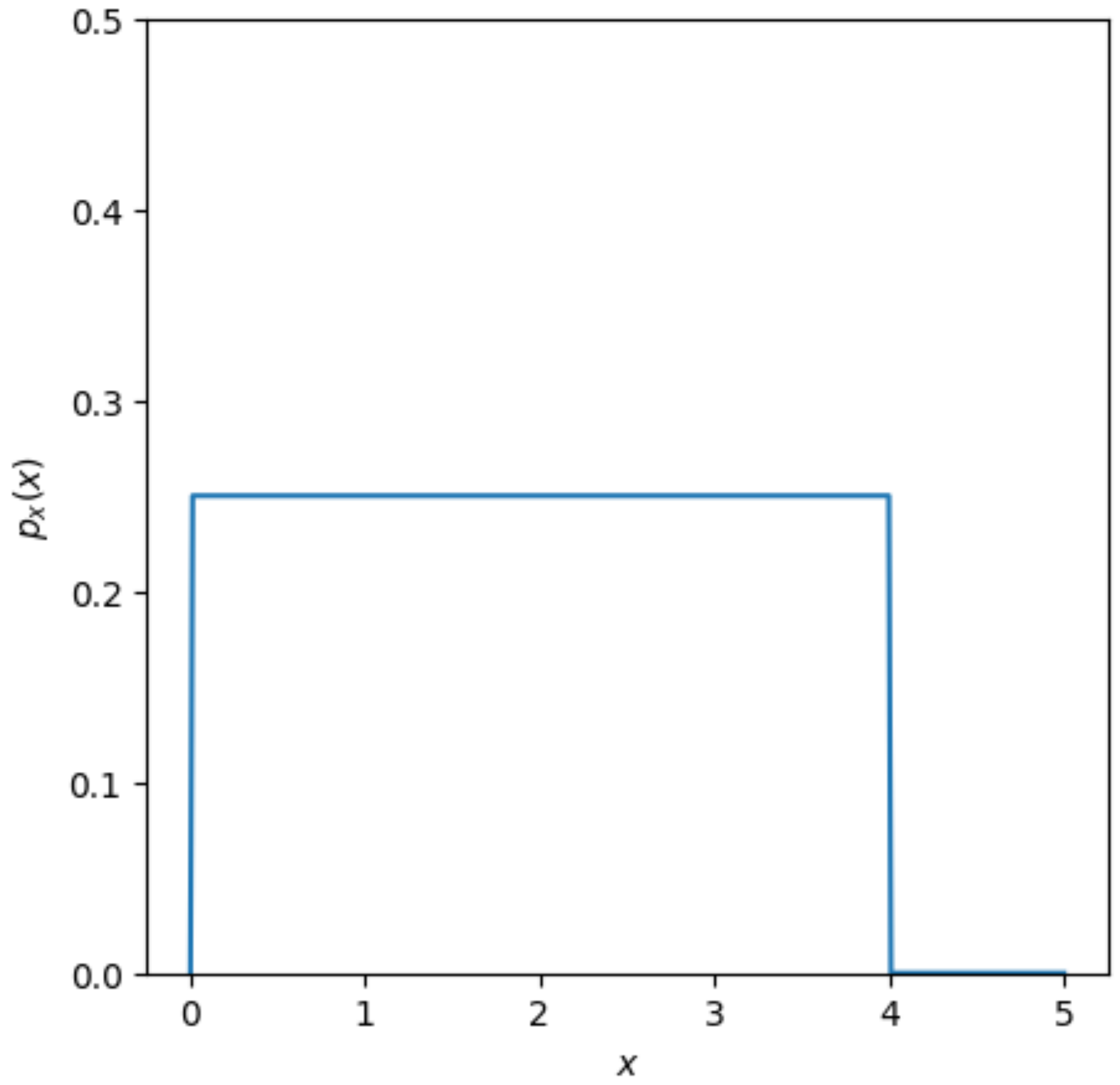


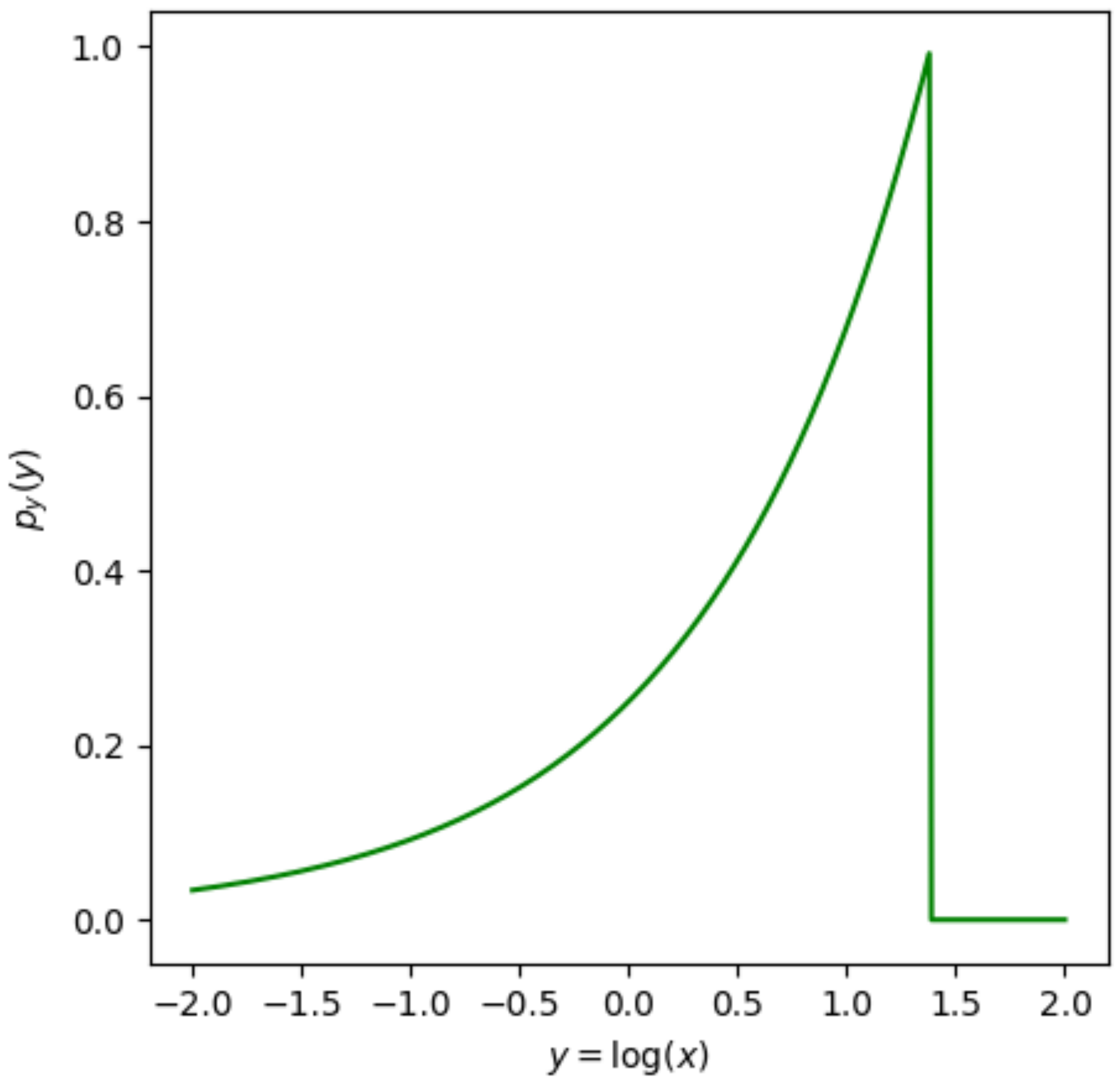
Reparametrization

- Start with probability distribution $p_\lambda(\lambda)$ for parameter λ
- Want to find distribution $p_\phi(\phi)$ for parameter $\phi = \phi(\lambda)$ with inverse $\lambda = \lambda(\phi)$
- Transformation:

$$p_\phi(\phi) = \left| \frac{d\lambda}{d\phi} \right| \; p_\lambda(\lambda(\phi))$$

- To memorize: $p_\phi(\phi) = \frac{dp}{d\phi}$ "=" $\frac{d\lambda}{d\phi} \frac{dp}{d\lambda}$
- Also useful $\left| \frac{d\lambda}{d\phi} \right| = \frac{1}{\left| \frac{d\phi}{d\lambda} \right|}$





Multivariate probability distributions

- Signify as $p(A, B)$, meaning $P(A \cap B)$, the probability that A and B are both true
  - Example: $p(N_1, N_2)$, the probability that out of two dice the first gives N_1 and the second gives N_2
  - $\sum_{N_1, N_2} p(N_1, N_2) = 1$
- Similar for densities, written as: $p(x_1, x_2)$
  - Example: $p(x, y)$, the probability that a particle is measured at the coordinates x and y
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 = 1$

Marginalization

- To remove one variable, sum over all its possible values
  $$p(N_1) = \sum_{N_2} p(N_1, N_2)$$
- Example of two dice, then $p(N_1, N_2) = 1/36$, each of the 36 outcomes is equally likely. Now $p(N_1) = \sum_{N_2=1}^{6} p(N_1, N_2) = 1/6$
- Similarly, for the continuous case: $p(x_1) = \int p(x_1, x_2)\, \mathrm{d}x_2$
- This is called *marginalization*

Conditional probabilities

- Write $p(A \mid B)$ to mean *the probability of A assuming B*
  - E.g.: $p(\text{"There will be rain tomorrow"} \mid \text{"There was rain today"})$
- *Product rule*: $p(A, B) = p(A \mid B)\ p(B)$
  - E.g.: The probability that tomorrow the chess player wins their game and has a good breakfast is the probability that they have a good breakfast multiplied by the probability that they win the game assuming they had a good breakfast

Bayes' Theorem

Bayes' Theorem

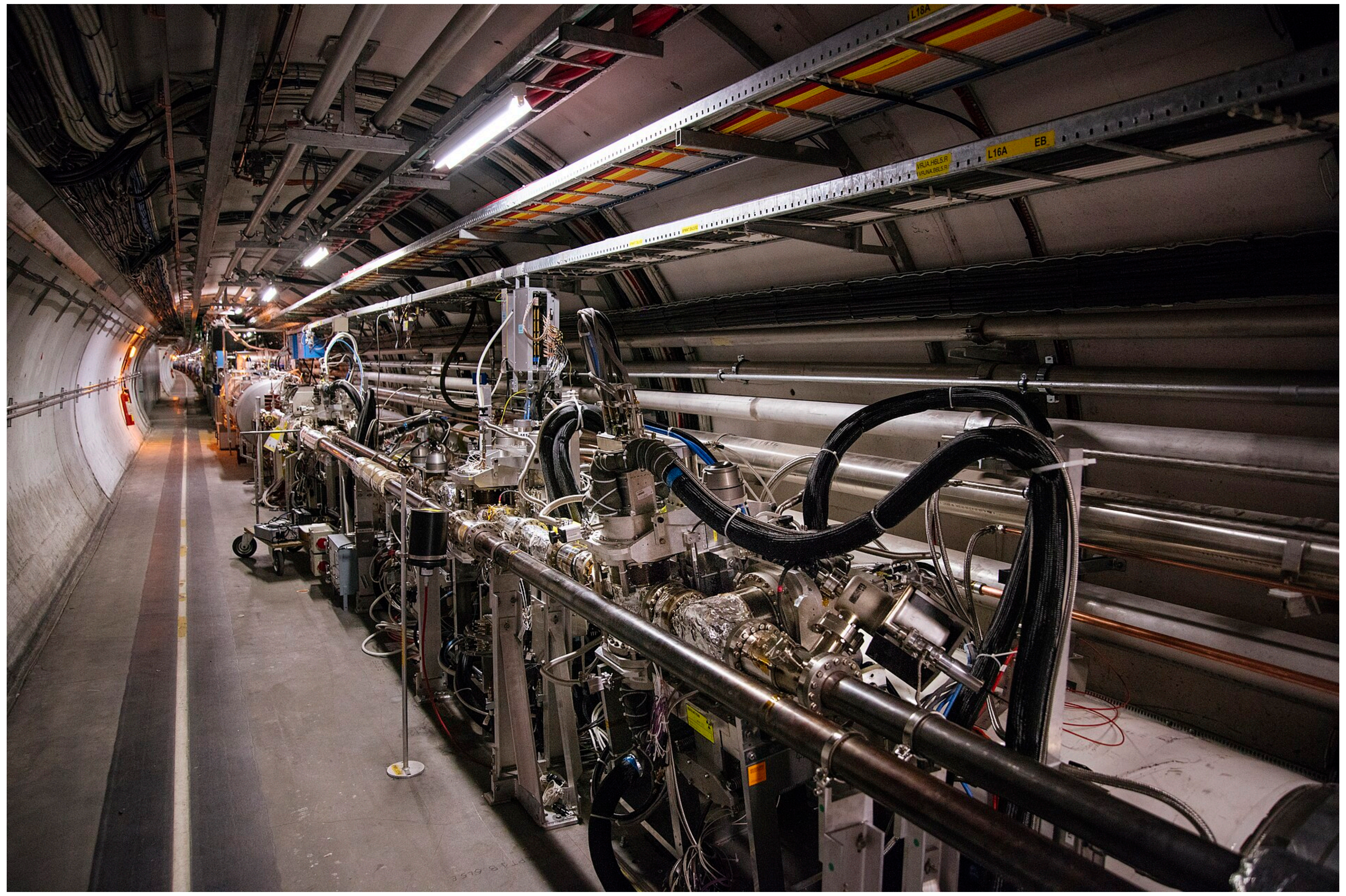
- Take a probability distribution with two variables: $p(A, B)$
- Product rule: $p(A, B) = p(A\,|\,B)\; p(B)$
- But also: $p(A, B) = p(B\,|\,A)\; p(A)$
- Combining these two, we get:

$$p(A\,|\,B) = \frac{p(B\,|\,A)\; p(A)}{p(B)}$$

- This is called *Bayes' Theorem*
- We can connect the two conditional probabilities

Bayesian inference I

- Consider $p(N_{\text{meas}} \,|\, \sigma)$, where N_{meas} is the number of particles measured in the detector and σ is the production cross section
- The number measured depends on σ, but also the detector efficiency, the measurement accuracy, quantum fluctuations etc.
- $p(N_{\text{meas}} \,|\, \sigma)$ tells us how for a particular σ the measurement would fluctuate when we repeat it
- Want to know the opposite: What does a measurement of N tell us about σ?
- Bayesian probability: $p(\sigma)$ is our state of knowledge about σ
- So we need to calculate $p(\sigma \,|\, N_{\text{meas}})$: the probability of σ assuming we measured N_{meas}



TOTEM Experiment at the LHC

Bayesian inference II

- Want to calculate probability of statements $A \in \{A_1, A_2, \dots\}$ based on a measurement $B \in \{B_1, B_2, \dots\}$
- Bayes' theorem:

$$p(A \mid B) = \frac{p(B \mid A)\ p(A)}{p(B)}$$

- $p(A \mid B)$ is the probability of the statement given the measurement. This is called the **posterior**
- $p(A)$ is the probability of the statement without knowing about the measurement, this is called the **prior**
- $p(B \mid A)$ is the probability of the measurement given the statement, this is called the **likelihood**
- $p(B)$ is a normalisation and can be rewritten as

$$p(B) = \sum_i p(A_i, B) = \sum_i p(B \mid A_i)\ p(A_i)$$

Example of using Bayes' theorem: Test for a rare disease

Base probability (for anyone) to have a disease D:

$$p(\mathrm{D}) = 0.001$$
$$p(\text{no D}) = 0.999$$

Consider a test for the disease: result is positive or negative (+ or −):

"sensitivity" ⟶ $p(+\,|\,\mathrm{D}) = 0.98$

$p(-\,|\,\mathrm{D}) = 0.02$

$p(+\,|\,\text{no D}) = 0.03$

"specificity" ⟶ $p(-\,|\,\text{no D}) = 0.97$

Suppose your result is +. How worried should you be?

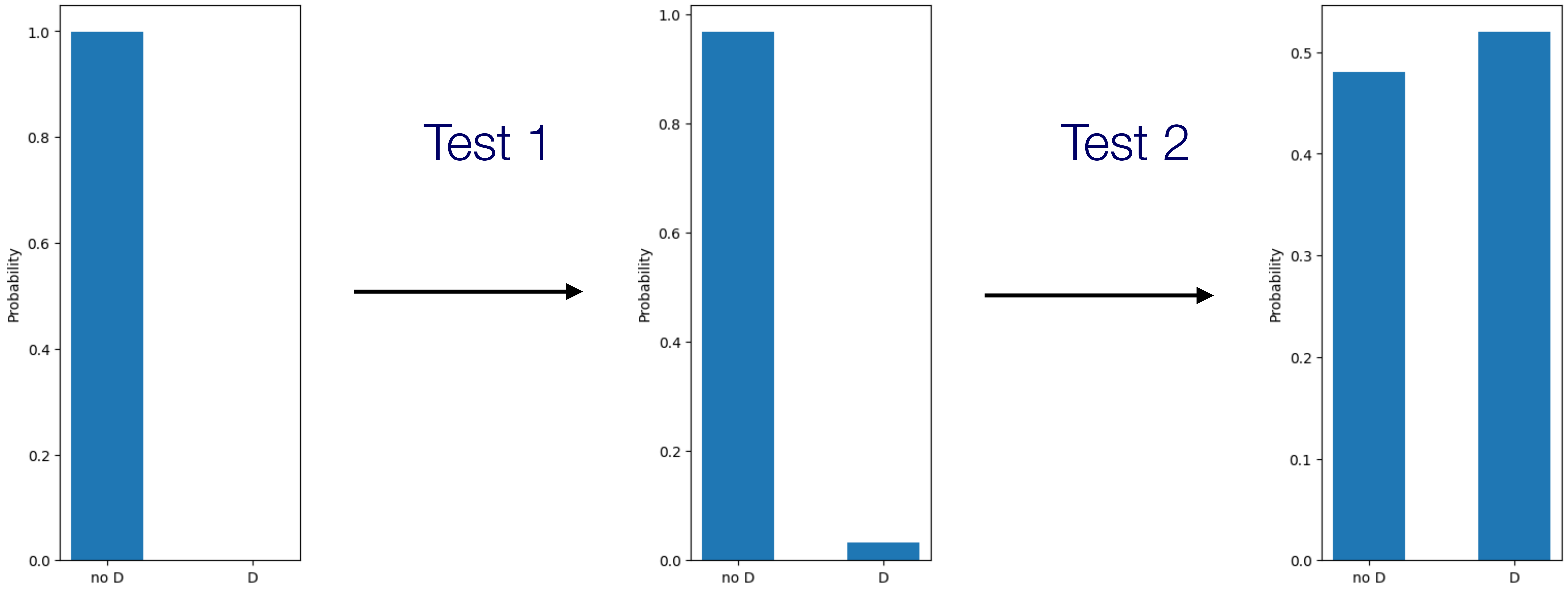
$$p(\mathrm{D}\,|\,+) = \frac{p(+\,|\,\mathrm{D})\; p(\mathrm{D})}{p(+)} = \frac{p(+\,|\,\mathrm{D})\; p(\mathrm{D})}{p(+\,|\,\mathrm{D})\; p(\mathrm{D}) + p(+\,|\,\text{no D})\; p(\text{no D})}$$
$$= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} = 0.032$$

Probability for you to have the disease is 3.2%, i.e., you're probably ok.

Remark: false positives **not** a relevant issue in statistics of Corona cases
(in case of a positive result usually double checks are made resulting in very high specificity)

What if we do a second test?

- Second, *independent* test is also positive
- Same strength $p(+_{(2)}\,|\,\mathrm{D}) = 0.98$, $p(-_{(2)}\,|\,\mathrm{no\ D}) = 0.97$
- The prior is now the posterior of the previous result (current state of knowledge)
- $p(\mathrm{D}\,|\,+) = \dfrac{0.98 \times 0.032}{0.98 \times 0.032 + 0.03 \times 0.968} = 0.52$



Bayesian and Frequentist interpretation of result

Bayesian:

- Frequency of disease in population is prior
- Probability for *this* patient to have disease is valid concept
  - $p(\mathrm{D} \,|\, +) = 0.032$ is the probability for this patient to have the disease, this encodes the uncertainty

Frequentist:

- Probability for *this* patient to have disease is not a valid concept - there is no random process
- Probability for a patient randomly drawn from the population to have disease is a valid concept
- Two possible statements:
  - "If we randomly select a person from the population, then the people testing positive have a probability of 0.032 of having the disease."
  - "If a patient is healthy, we would get a positive test with a probability of 0.03"
- Neither are probabilities for *this particular person* to have the disease

Criticisms of the probability interpretations

(Subjective) Bayesian interpretation:

- Gives us what we want from measurements: The degree of certainty in a particular statement
- Straightforward: Always apply Bayes' theorem and that is all
- Close to everyday thinking

but:

- "Subjective" estimates have no place in science
- Result may strongly depend on prior - how do we quantify the prior state of our knowledge
- Usually requires stating all possible hypotheses explicitly

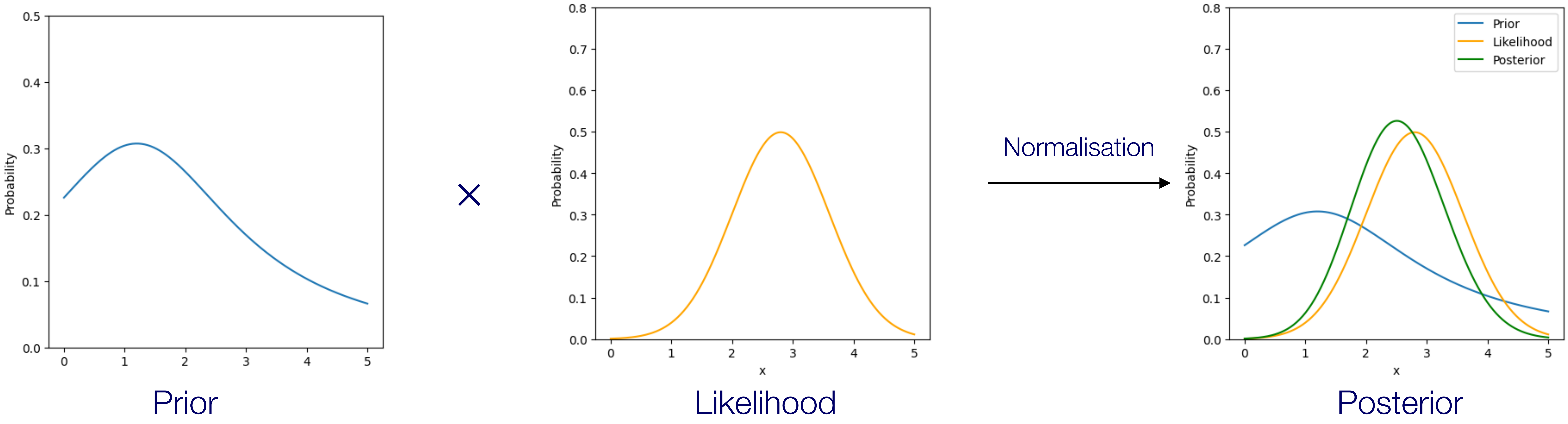
Frequentist interpretation:

- Gives objective results
- Variety of methods; some don't require explicitly stating all hypotheses
- Essentially how we think about QM

but:

- Need to reformulate question to get any answer at all
- $n \to \infty$ actually not possible in practice (e.g. a coin may deform)
- p is not an intrinsic property of A, it depends on the how the ensemble of possible outcomes was constructed (do we consider redoing the test or also re-choosing a person)
- Many different methods follow different (possibly contradictory) paradigms

Bayesian inference for a continuous parameter



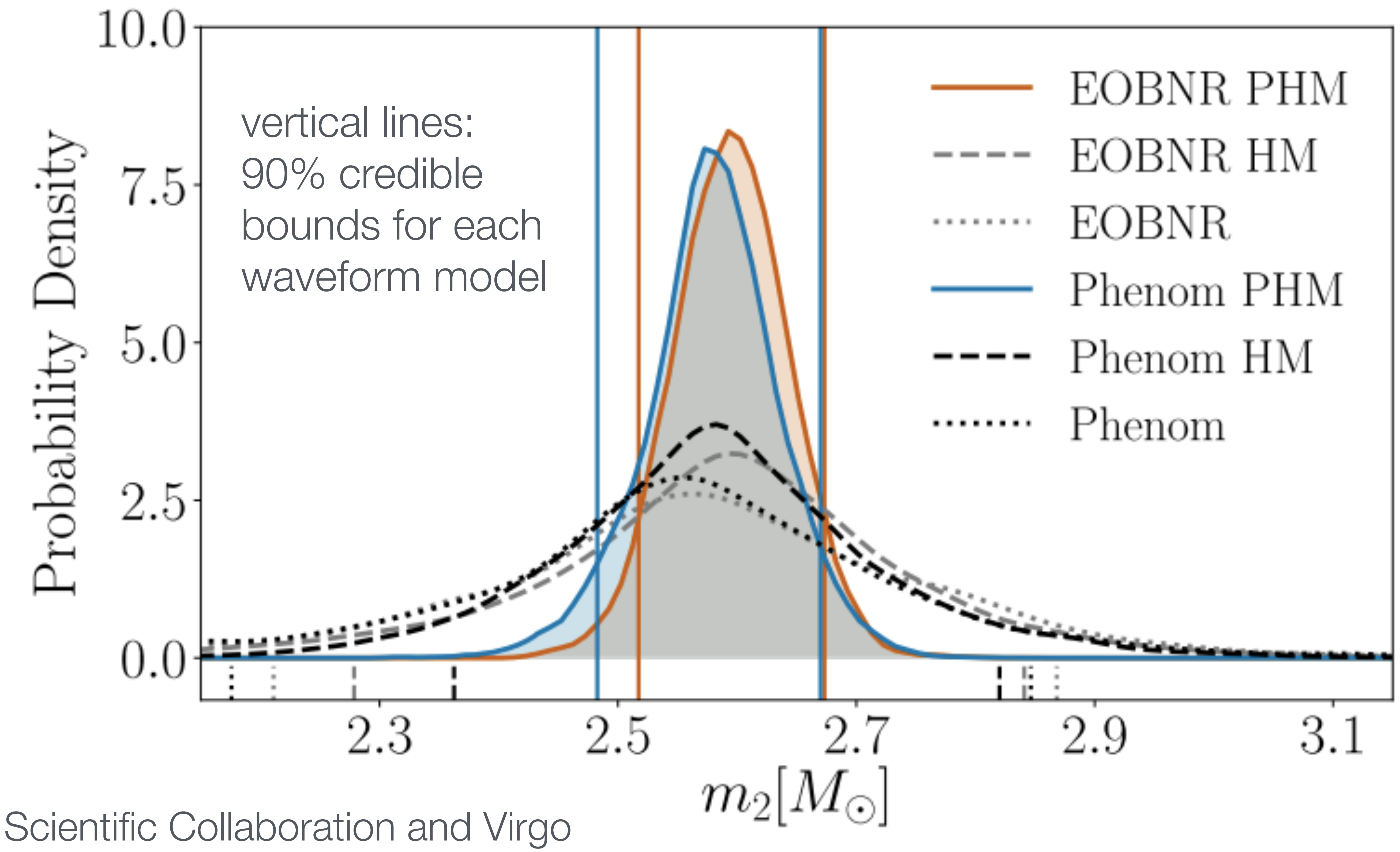
- Bayes' theorem for a parameter x with measurement(s) m:

$$p(x \mid m) = \frac{p(m \mid x)\ p(x)}{\int p(m \mid x)\ p(x)\,\mathrm{d}x}$$

- Posterior usually much narrower than prior. Narrow distribution → more information content

Example of a posterior distribution

GW190814: Gravitational waves from the coalescence of a 23 solar mass Black Hole with a 2.6 solar mass compact object



Posterior distribution for mass of the lighter objects:

LIGO Scientific Collaboration and Virgo Collaboration:
The Astrophysical Journal Letters, 896:L44 (20pp), 2020 June 20

Measured value and uncertainty from position and width of posterior distribution

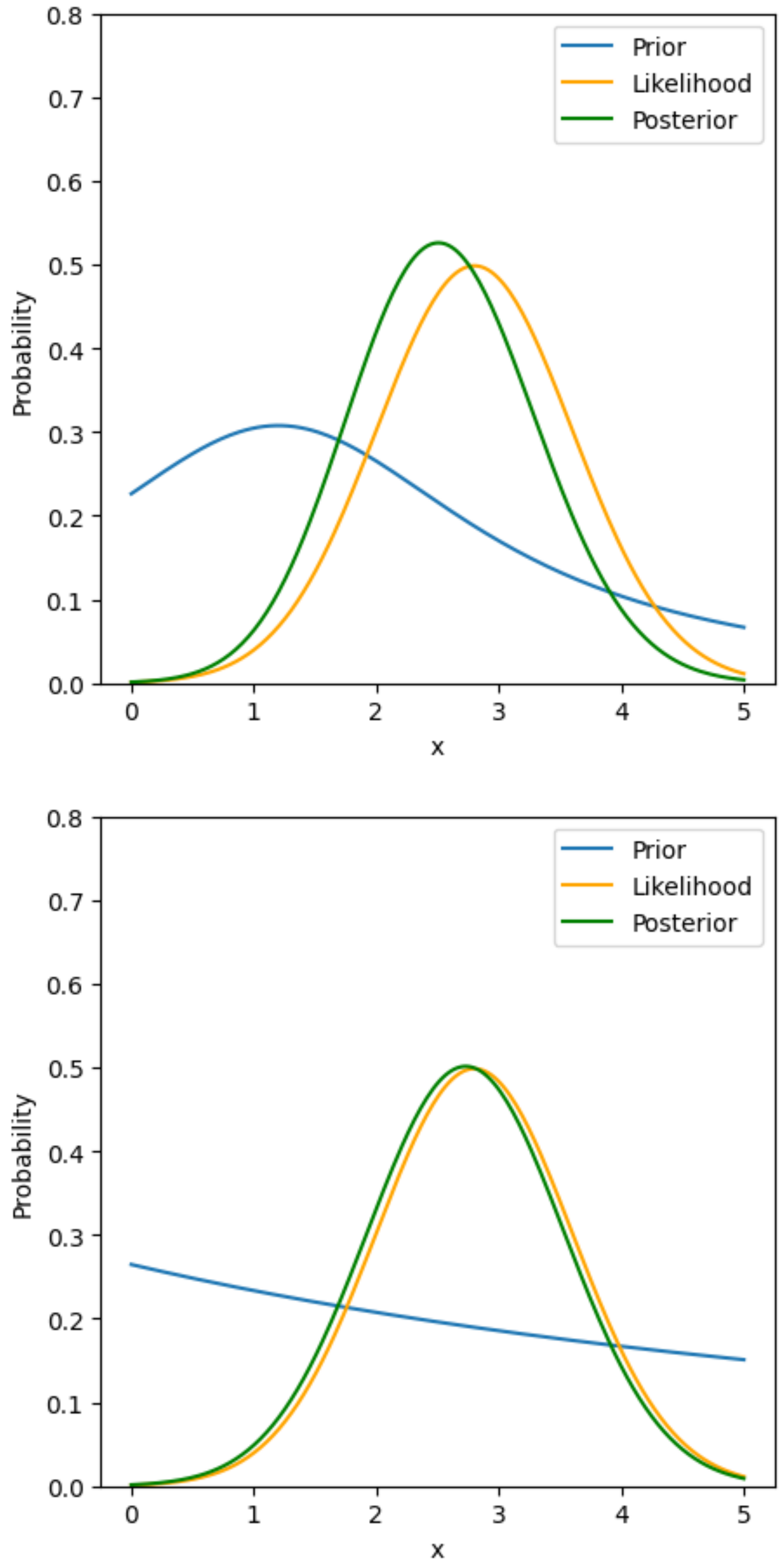
Where does the prior come from?

- Result depends on prior - where does it come from
- What does it mean if we "know nothing"?
- Complex question, will revisit later

Three main approaches:

1. Select prior to represent current (subjective) state of knowledge
2. Select prior based on transformation/scaling properties (e.g. Jeffrey's priors, maximum entropy principle)
3. Select priors for convenience of calculation (e.g. conjugate priors)

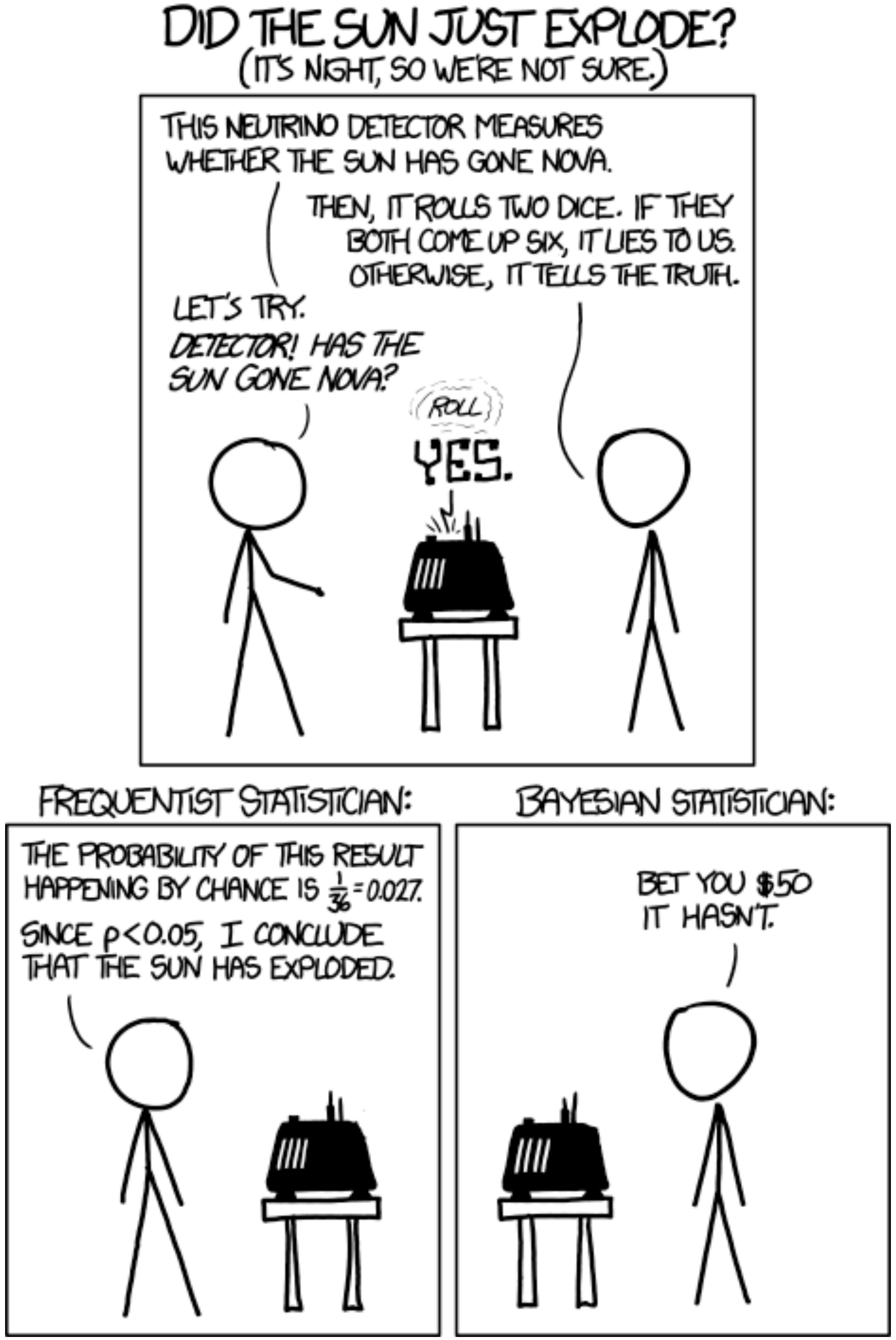
- The good news: If the prior distribution is wide, the posterior only has weak dependence on it



Bayesian versus Frequentist Probability

	Bayesian	Frequentist
Meaning of probability	degree of belief	relative frequency
Probability applies to	random variables, parameters and hypotheses	random variables only
Data analysis	Data changes state of knowledge	Data speaks for itself
Unphysical / empty intervals	excluded by prior	can occur
Final statement	posterior probability distribution	parameter values, hypothesis test (*p*-value), confidence interval …
Nuisance parameters	Marginalization	Various methods, e.g., profile likelihood, hard
Systematics	Included naturally	Separate concept
Difficulties	Describing state of knowledge as prior distribution	Interpretation of result

Bayesian versus Frequentism



https://xkcd.com/1132/

1.2 Describing the Data

Goal of describing data

- Data often series of measurements, e.g. $m_1, m_2, m_3, \ldots$
- How to summarize in few numbers?
- Similarly: Summarize distribution, e.g. posterior

- Usually two main numbers: Where is the distribution and how wide is it?
- Also called summary statistics

Histograms

Histogram:

- representation of the frequencies of the numerical outcome of a random phenomenon
- Number of entries found in different (often equidistant) intervals

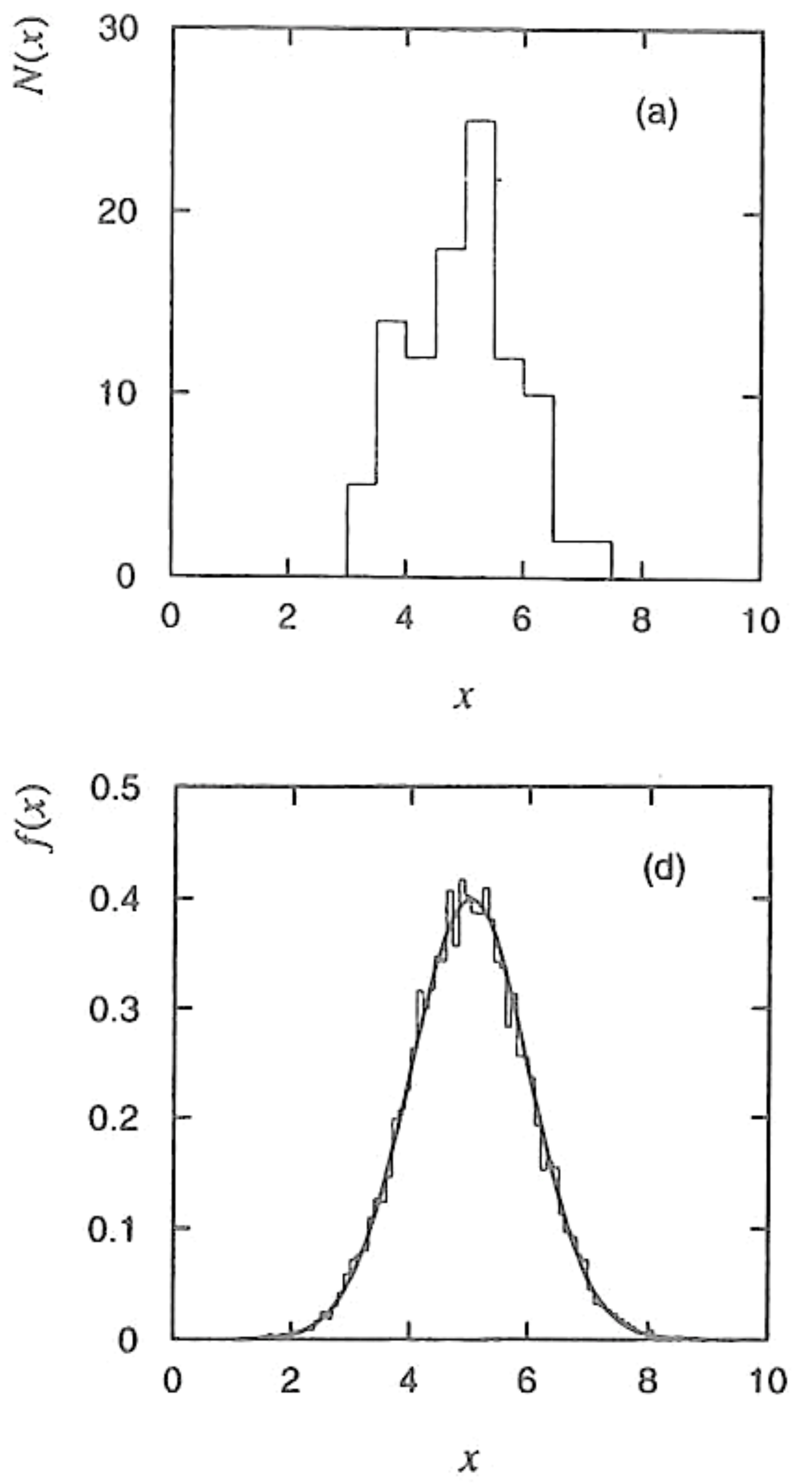
pdf = histogram for

- infinite data sample
- zero bin width
- normalized to unit area

$$f(x) = \frac{N(x)}{n\Delta x}$$

$n = \text{total number of entries}$

$\Delta x = \text{bin width}$



Location summary statistics: Mean, Median, and Mode

Mean:

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

"sample mean"

$$\mu \equiv \langle x \rangle \equiv \int x\,P(x)\,\mathrm{d}x$$

"expectation value"

Median:

Value for which half of measured values are above and half below

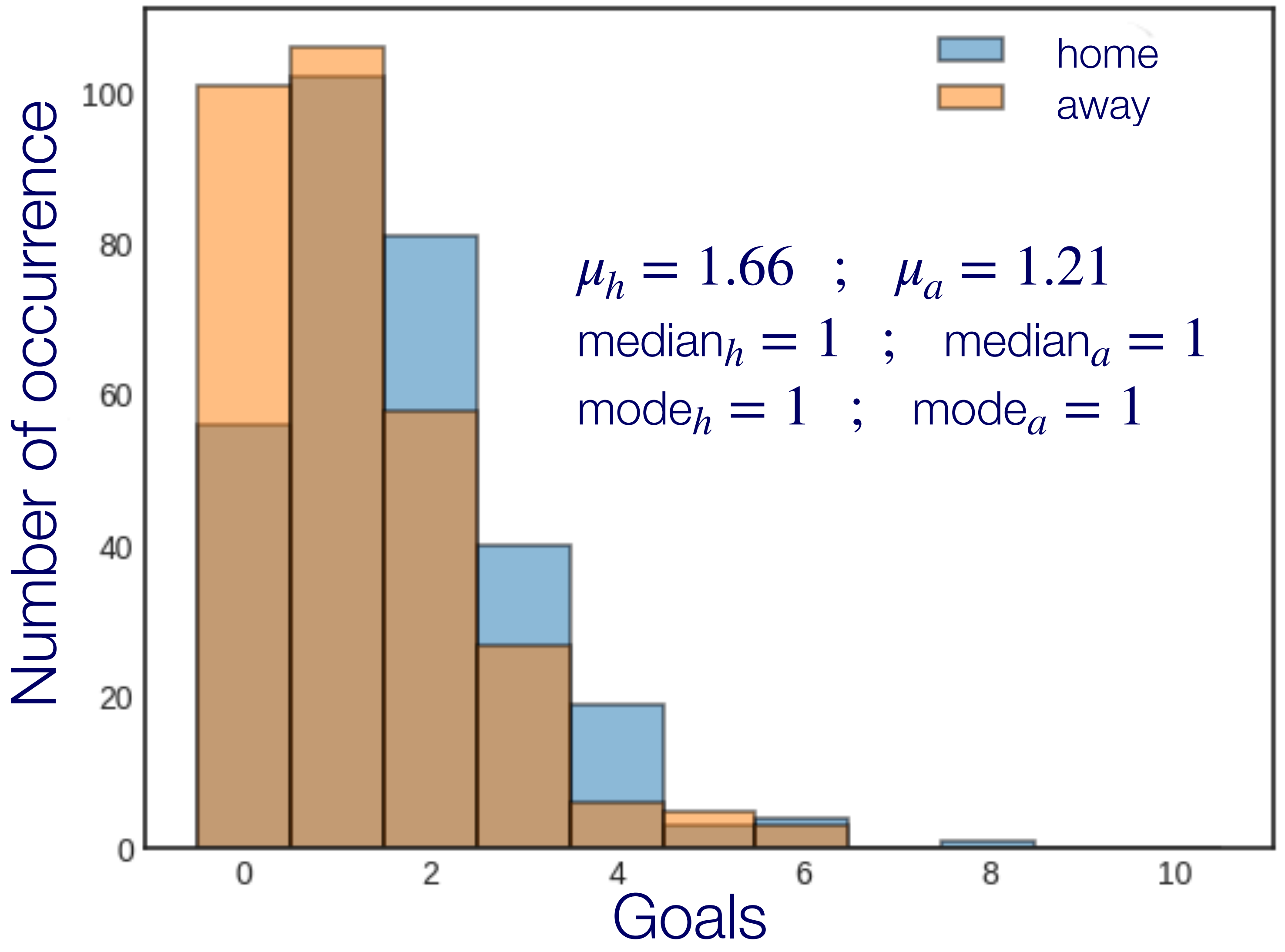
$$\int_{-\infty}^{x_\text{med}} p(x)\,\mathrm{d}x = \int_{x_\text{med}}^{\infty} p(x)\,\mathrm{d}x = \frac{1}{2}$$

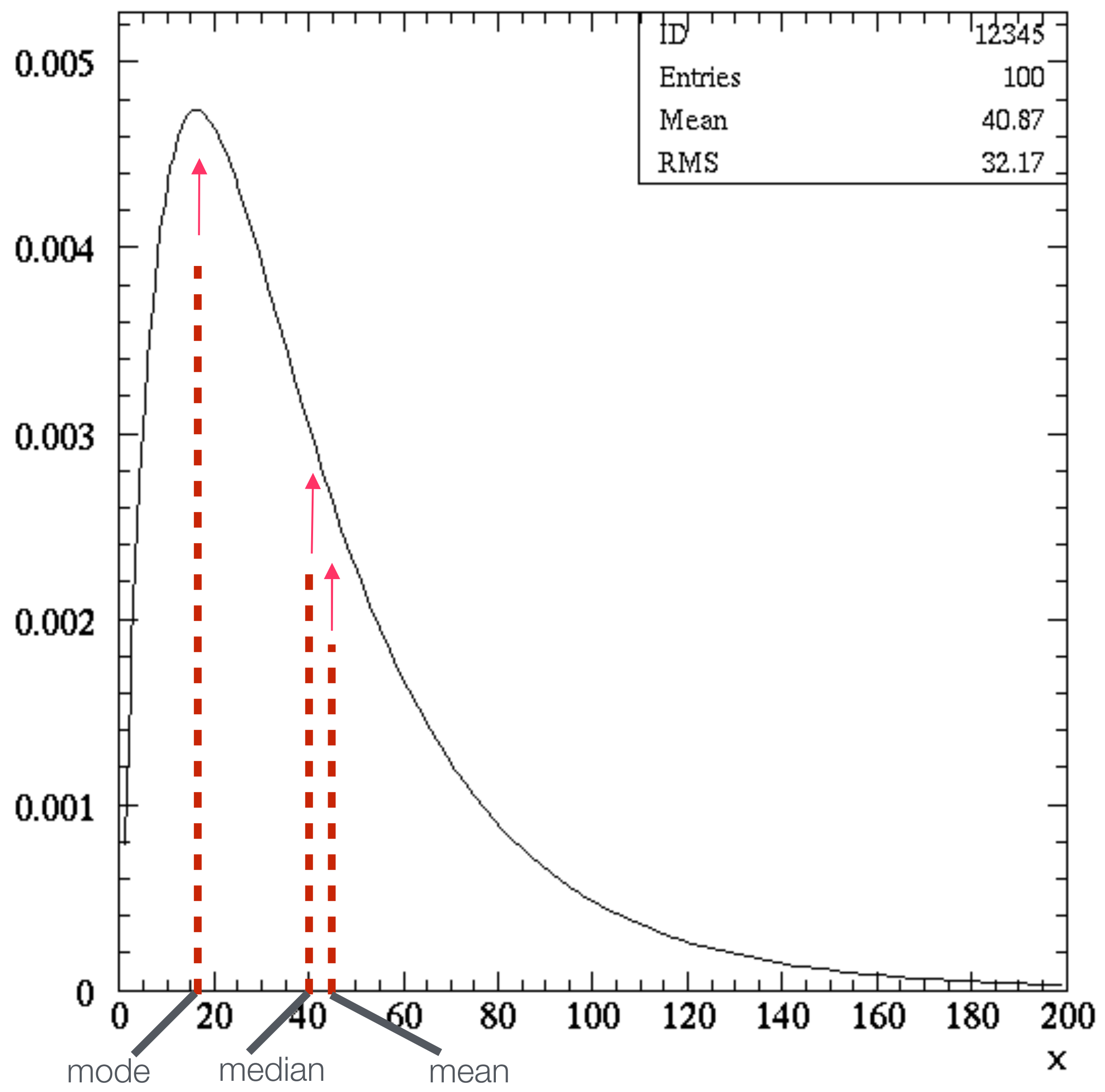
Mode:

Most frequently measured value

Maximum of distribution

There are also others, e.g. centre between lowest and highest measured value





Dispersion summary statistics

With respect to a position parameter x_0

Mean absolute difference:

$$MD = \frac{\sum_i \left| x_i - x_0 \right|}{N}$$

$$MD = \int_{-\infty}^{\infty} p(x) \; \left| x - x_0 \right| \, \mathrm{d}x$$

Root-mean-square deviation:

$$rms = \sqrt{\frac{\sum_i \left(x_i - x_0\right)^2}{N}}$$

$$rms = \sqrt{\int_{-\infty}^{\infty} p(x) \; \left(x - x_0\right)^2 \mathrm{d}x}$$

- The mean absolute difference is minimised when x_0 is the median
- The rms is minimised when x_0 is chosen to be the mean
  - *Standard deviation* σ

Variance and standard deviation

Variance of a distribution:

$$V(x) = \int \mathrm{d}x\, P(x)(x-\mu)^2 = E[(x-\mu)^2]$$

expectation value

$$V(x) = \int \mathrm{d}x\, P(x) x^2 - 2\mu \underbrace{\int \mathrm{d}x\, P(x) x}_{=\mu} + \mu^2 \int \mathrm{d}x\, P(x) = \langle x^2 \rangle - \mu^2 = \langle x^2 \rangle - \langle x \rangle^2$$

Sample variance:

$$V(x) = \frac{1}{N} \sum_i (x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

This formula underestimates the variance of underlying distribution as it uses the mean calculated from data!

Use this if you have to estimate the mean from data (***unbiased sample variance***):

$$\hat{V}(x) = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

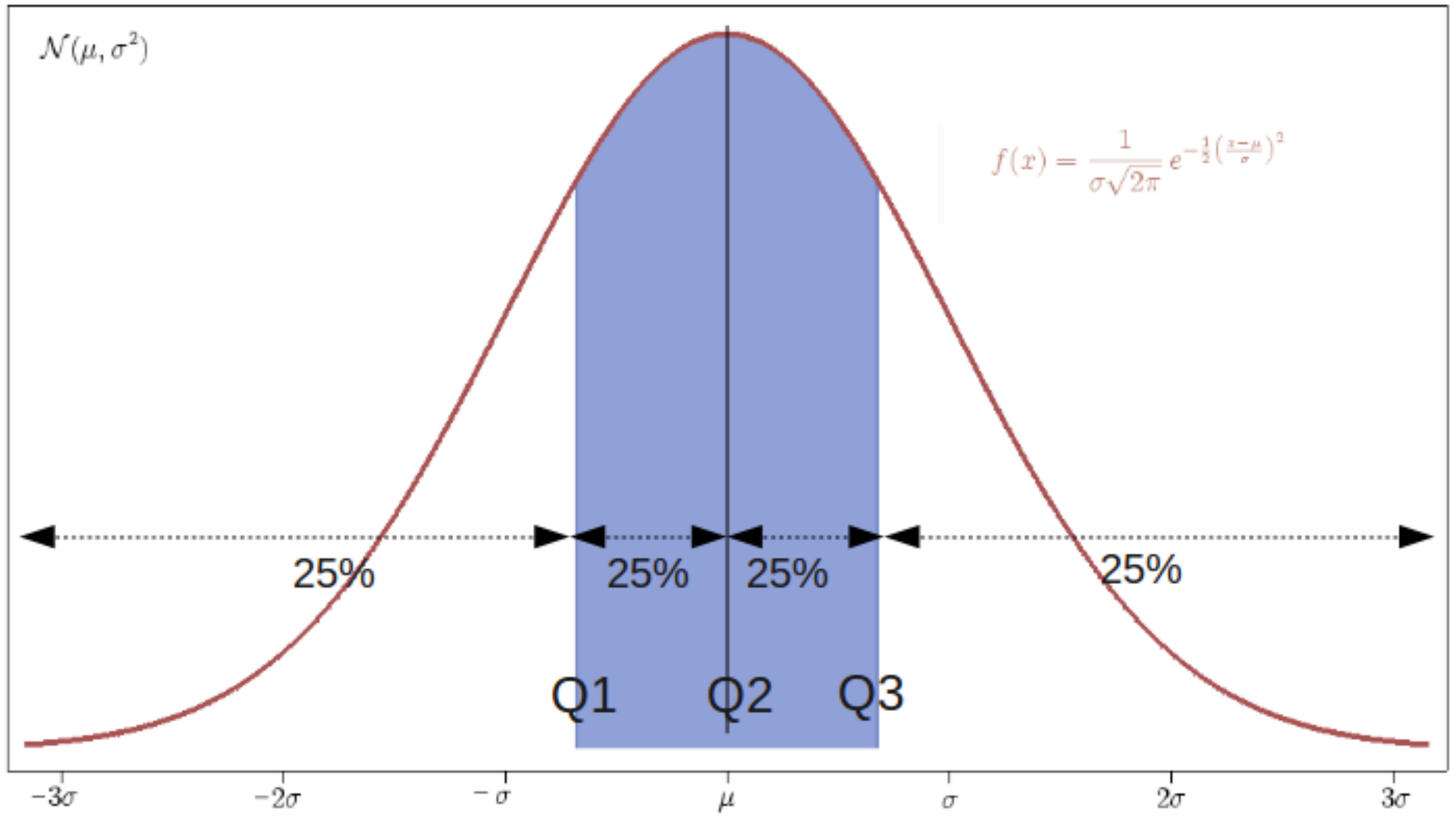
Use this if you know the true mean μ:

$$V(x) = \frac{1}{N} \sum_i (x_i - \mu)^2$$

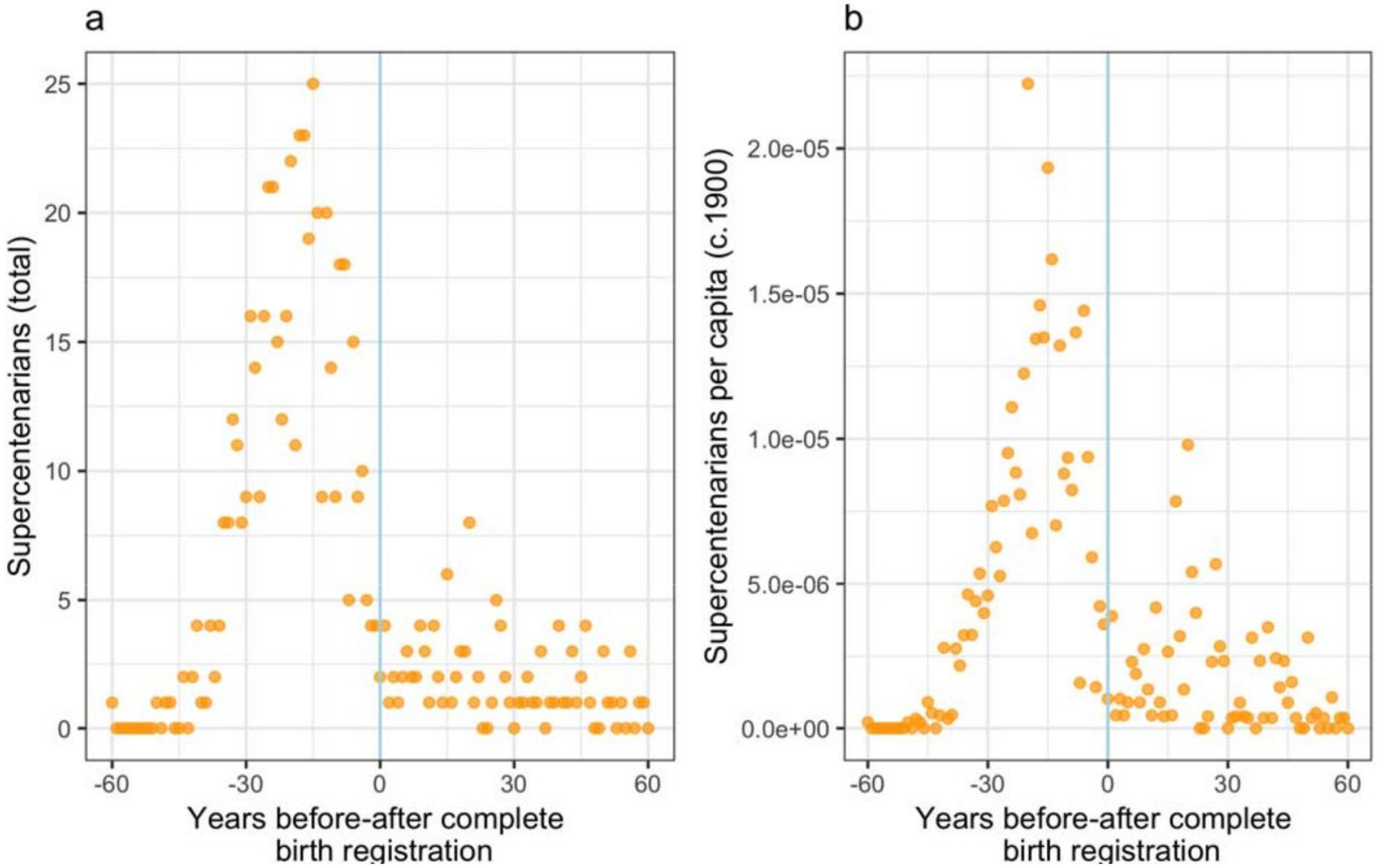
Standard deviation: $\sigma = \sqrt{V(x)}$

Quantiles

- Summarise both position and dispersion simultaneously
- Split range of distribution into intervals corresponding to equal probabilities
- For $n = 2$, the intervals are separated by the median



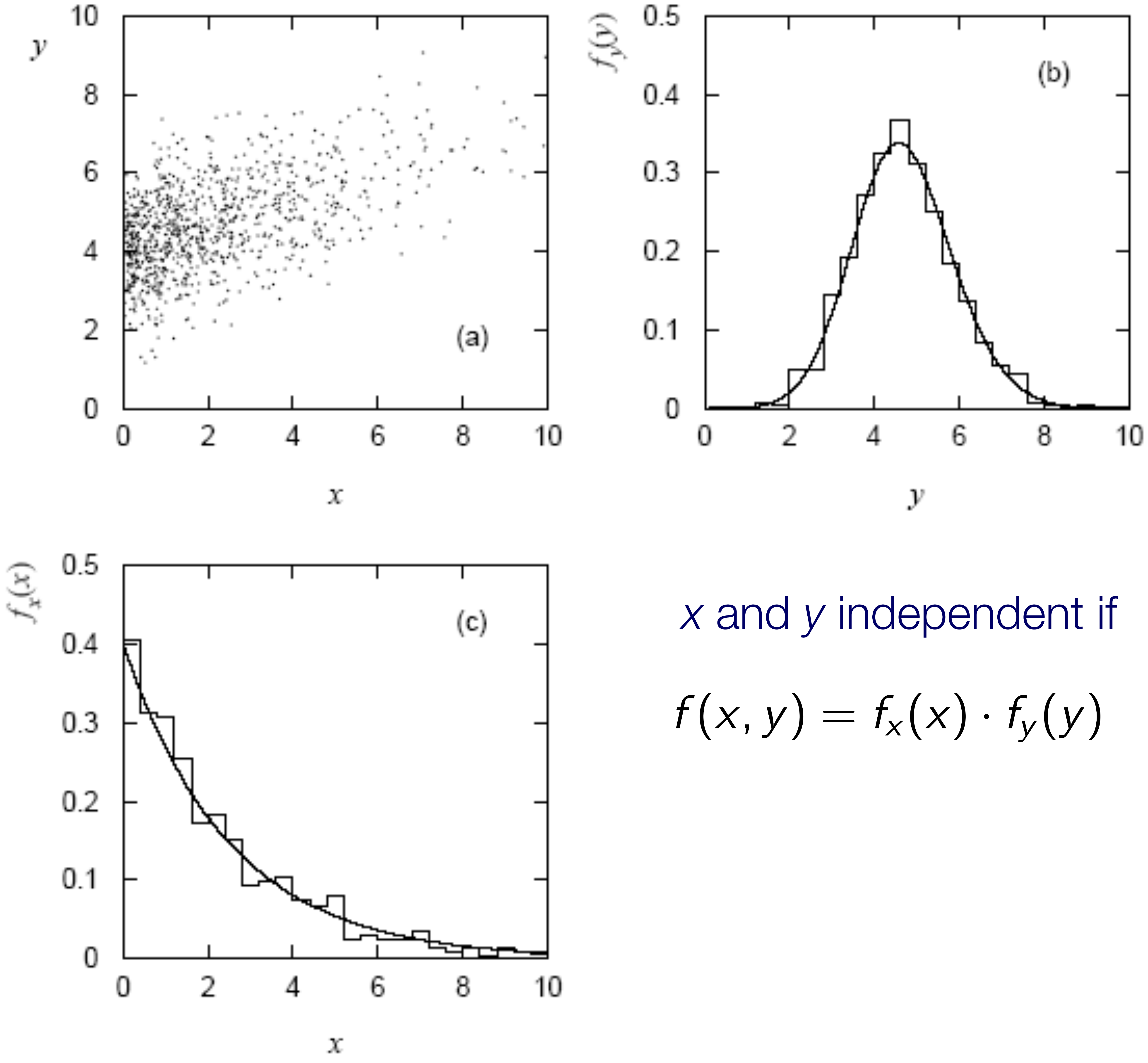
Supercentenarians



Saul Justin Newman, https://doi.org/10.1101/704080 (preprint)

Correlation and dependence

- Can have distributions of more than one variable
- Two variables are *dependent* or *correlated* if knowledge of one changes the state of knowledge about the other
- Otherwise they are independent
- We often consider *linear correlation*
- Stronger correlation means: knowledge about one variable gives more information about the other



x and y independent if

$$f(x,y) = f_x(x) \cdot f_y(y)$$

Covariance and correlation

- Generalisation of variance, *Covariance* ($\mu_x := \langle x \rangle, \mu_y := \langle y \rangle$):

$$\mathrm{cov}[x,y] = E[(x-\mu_x)(y-\mu_y)]$$

$\mathrm{cov}[x,x] = \sigma_x^2$ $\mathrm{cov}[y,y] = \sigma_y^2$ is the variance

Pearson (linear) *correlation coefficient*:

$$\rho_{xy} = \frac{\mathrm{cov}[\mathrm{x},\mathrm{y}]}{\sigma_x \sigma_y}$$

Gives values in range $[-1,1]$

x, y independent, i.e., $f(x,y) = f_x(x) \cdot f_y(y)$:

$$f_x(x) = \int dy\, f(x,y)$$

$$f_y(y) = \int dx\, f(x,y)$$

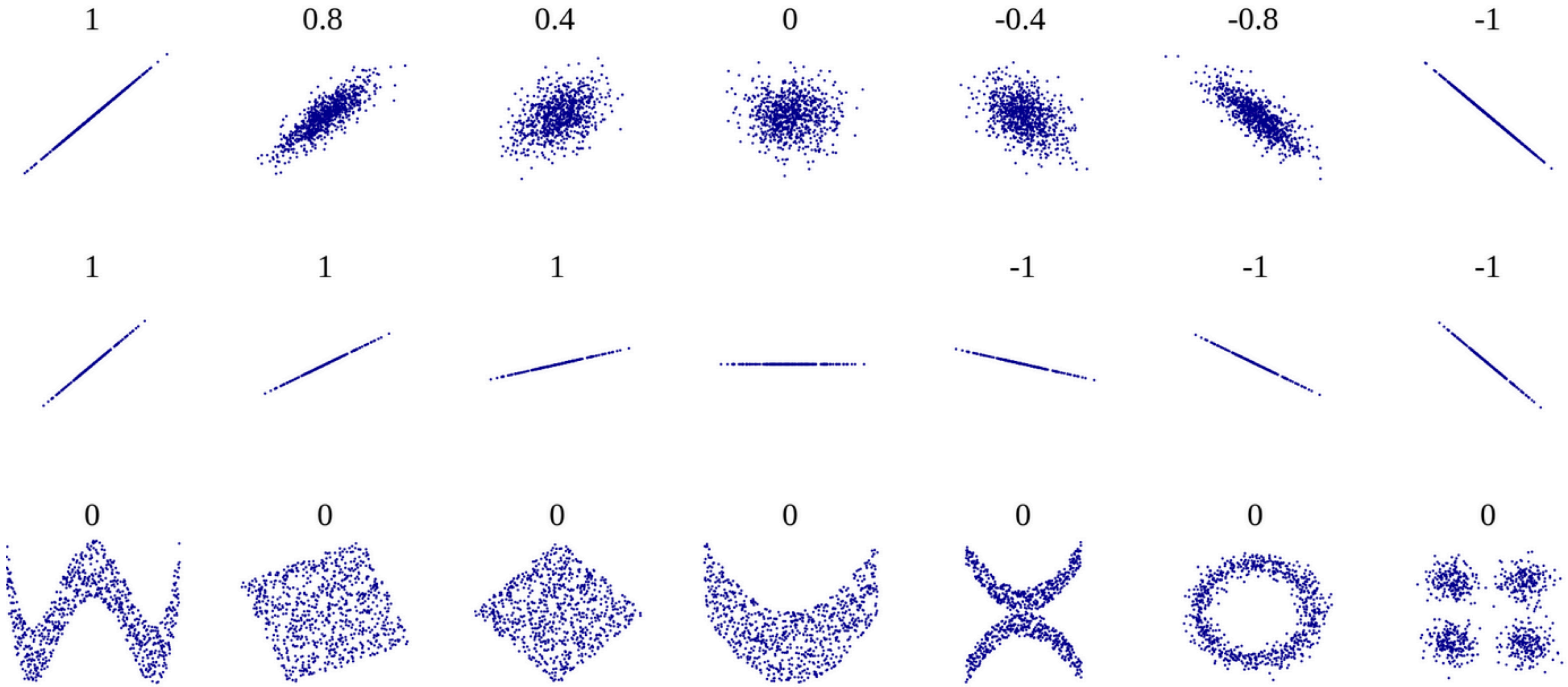
$$E[(x-\mu_x)(y-\mu_y)] = \int \mathrm{d}x \int \mathrm{d}y\, (x-\mu_x)(y-\mu_y) f(x,y)$$

$$= \int (x-\mu_x) f_x(x)\, \mathrm{d}x \int (y-\mu_y) f_y(y)\, \mathrm{d}y = 0$$

$\rightarrow$ $\mathrm{cov}[x,y] = 0$ (N.B. converse not always true)

Correlation coefficient

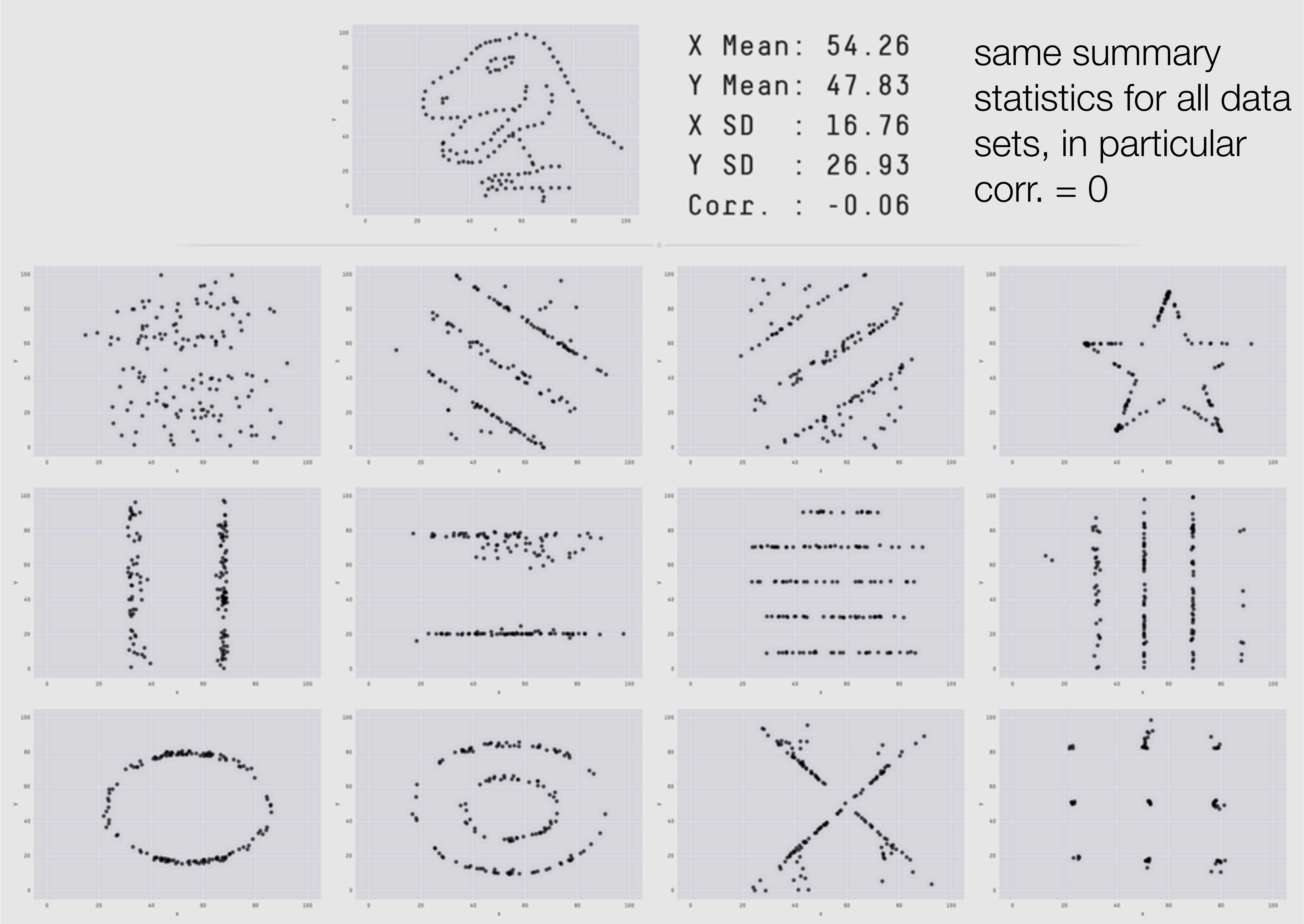
https://en.wikipedia.org/wiki/Correlation_and_dependence



$$\mathrm{cov}[x, y] = E[(x - \mu_x)(y - \mu_y)]$$

Never trust summary statistics alone; always visualize your data

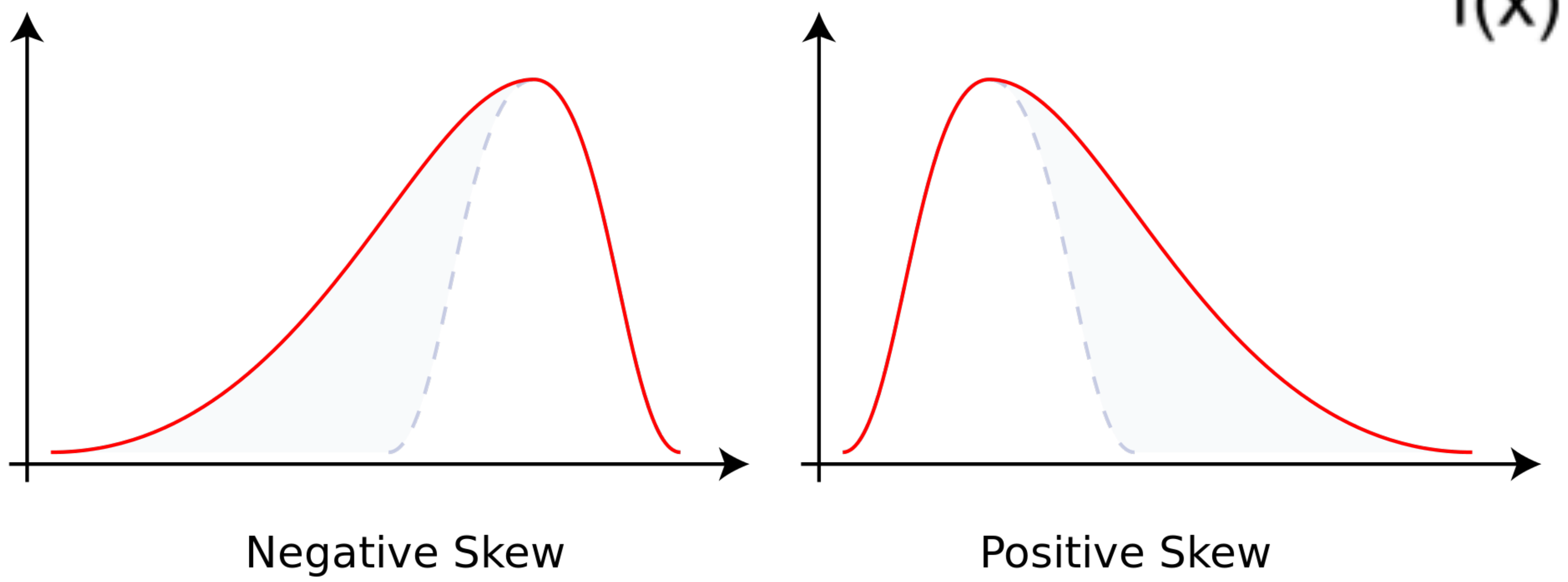
https://www.autodeskresearch.com/publications/samestats



same summary statistics for all data sets, in particular corr. = 0

Higher moments

Skewness: $$\gamma_1 = \left\langle \left(\frac{x - \langle x \rangle}{\sigma} \right)^3 \right\rangle$$



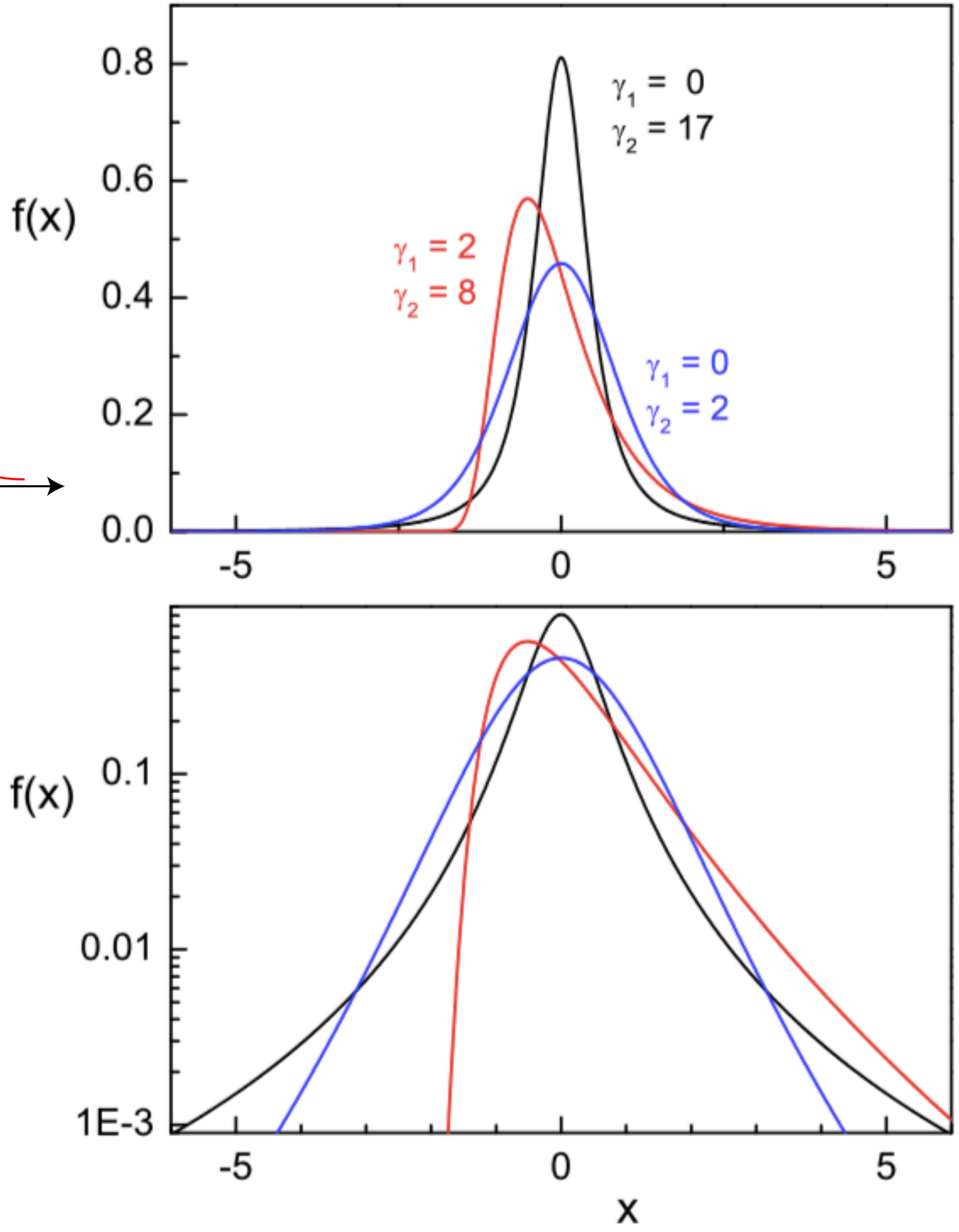
https://en.wikipedia.org/wiki/Skewness

Symmetric distribution have skewness equal to zero

Kurtosis: $$\beta_2 = \left\langle \left(\frac{x - \langle x \rangle}{\sigma} \right)^4 \right\rangle$$

$$\gamma_2 = \beta_2 - 3$$

defined such that $\gamma_2 = 0$ for the normal distribution



Correlation ≠ Causation

Examples of illogically inferring causation from correlation

https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

Example 1 ("reverse causality"):

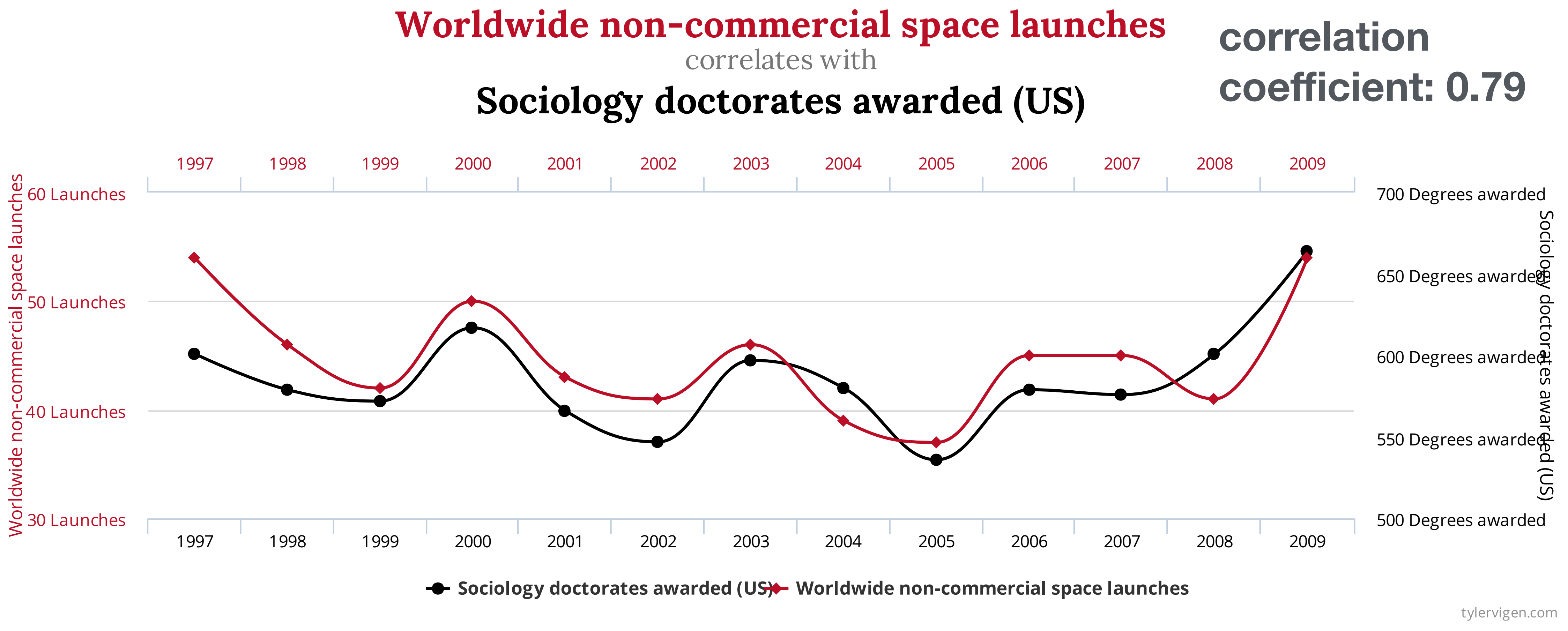
- The faster windmills are observed to rotate, the more wind is observed to be.
- Therefore wind is caused by the rotation of windmills.

Example 2 ("third factor C causes both A and B"):

- Sleeping with one's shoes on is strongly correlated with waking up with a headache.
- Therefore, sleeping with one's shoes on causes headache.

Example 3 ("relationship is coincidental"):

- many examples on tylervigen.com ("spurious correlations")



correlation coefficient: 0.79

What makes nobel prize winners?

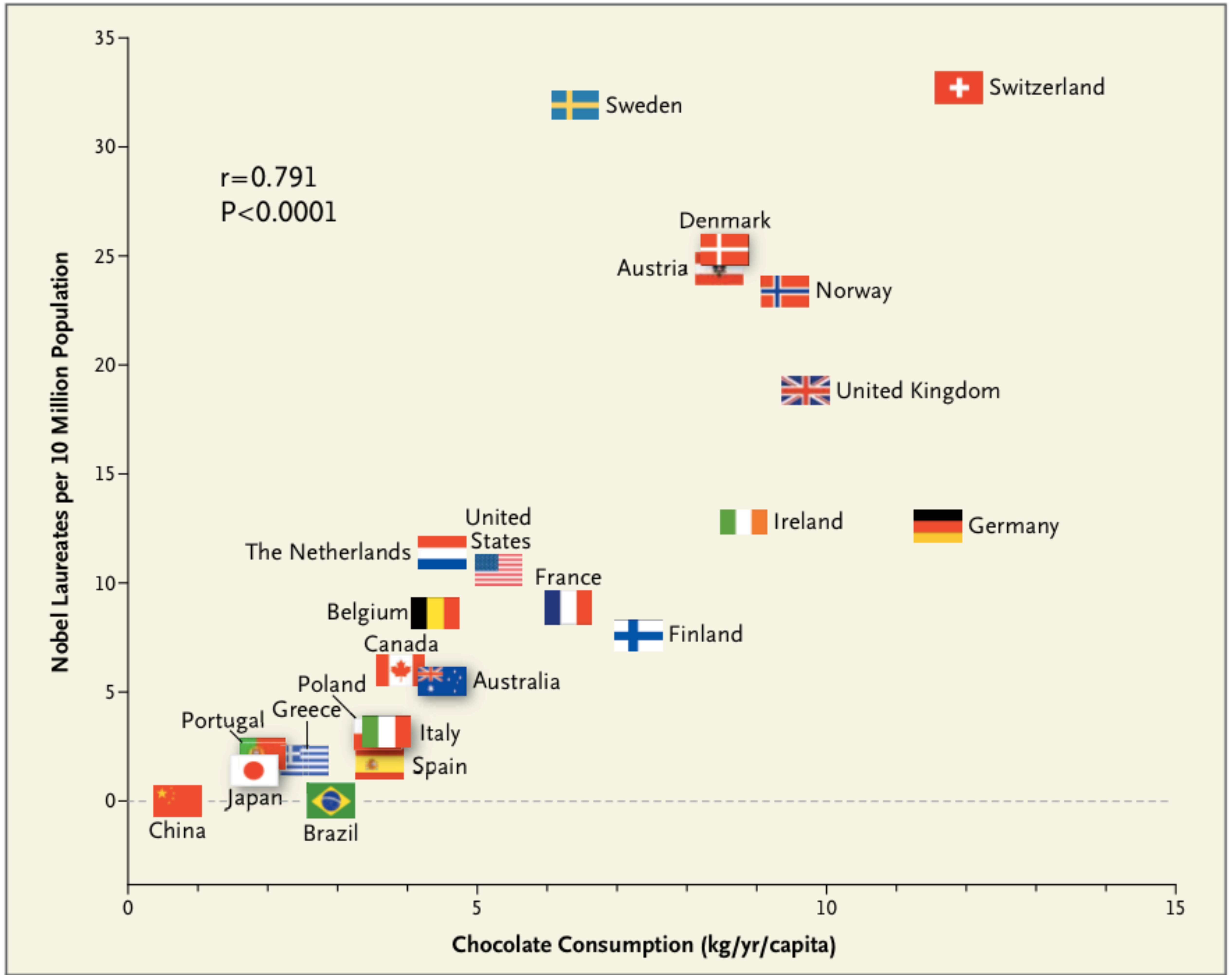


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

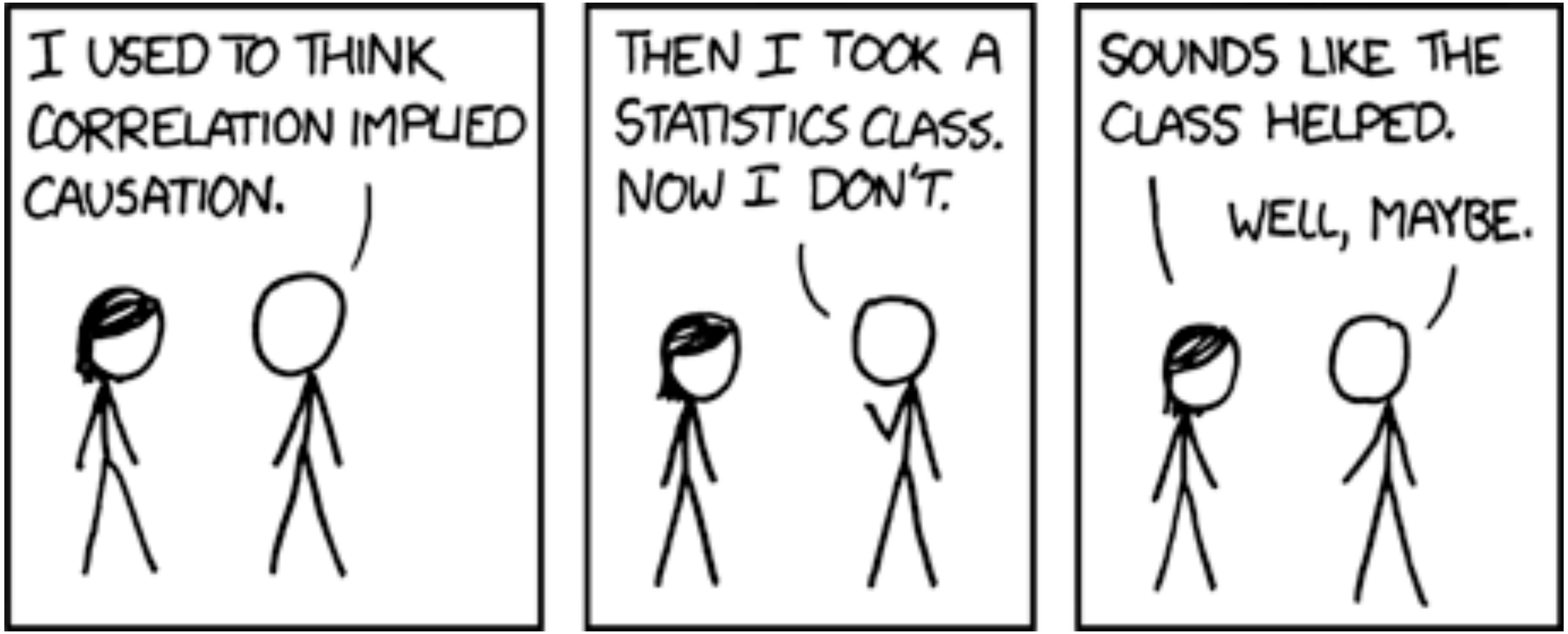
Correlation coefficient: 0.791

Improved cognitive function associated with a regular intake of flavonoids???

Probably not …

F. Messerli, 2012, New England Journal of Medicine, 2012

Correlation ≠ Causation



"Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'."

https://xkcd.com/552/