Modern Bayesian inference in practice

Kevin Kröninger – TU Dortmund University

Terascale Statistics School – February 28th 2025

Outline

Two questions for today

What does the concept of probability mean for you as a person?

→ Scientific reasoning

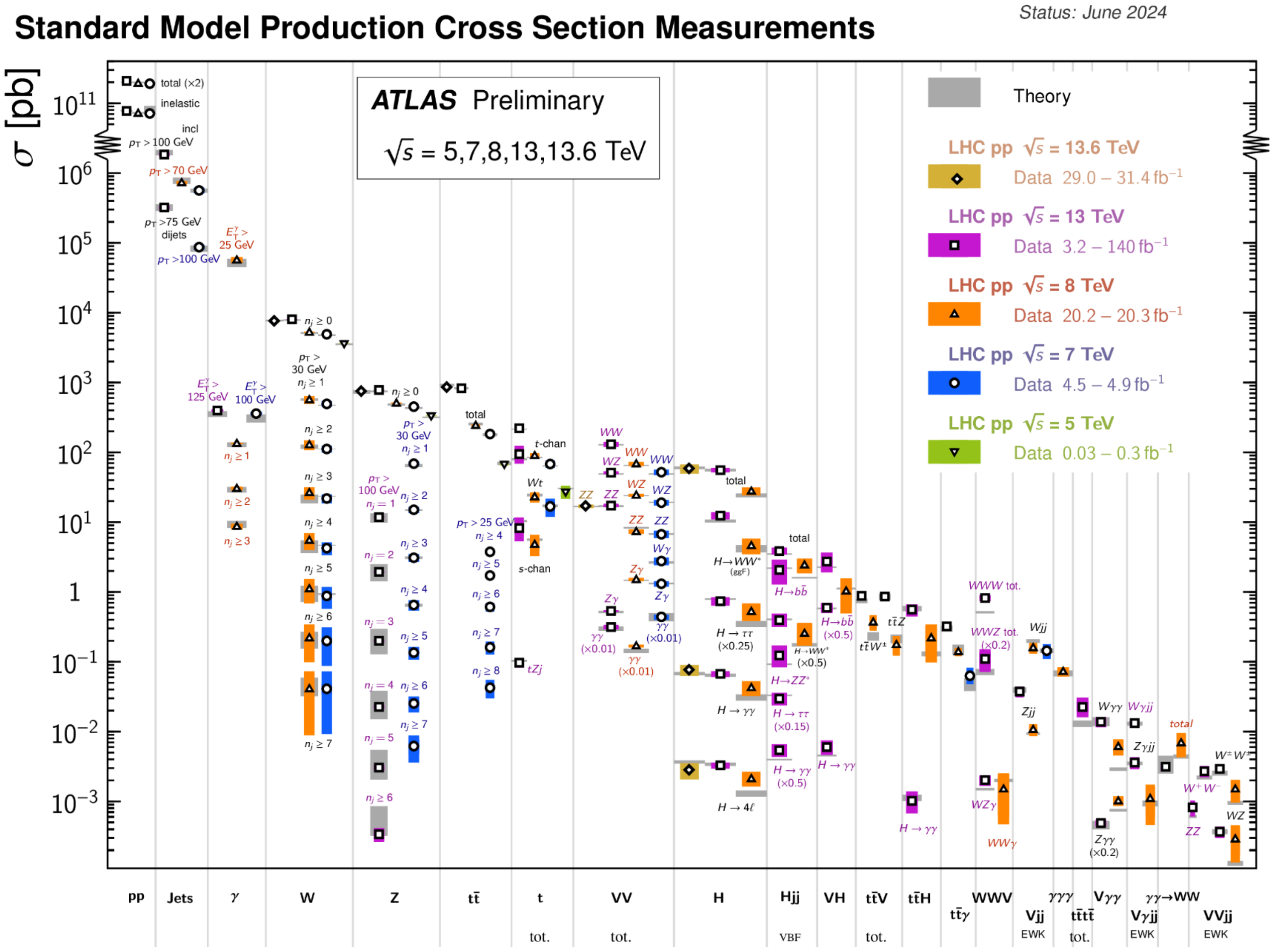
→ Probability and Bayes Theorem

How can Bayesian quantities be computed?

→ Markov Chain Monte Carlo

→ A concrete example

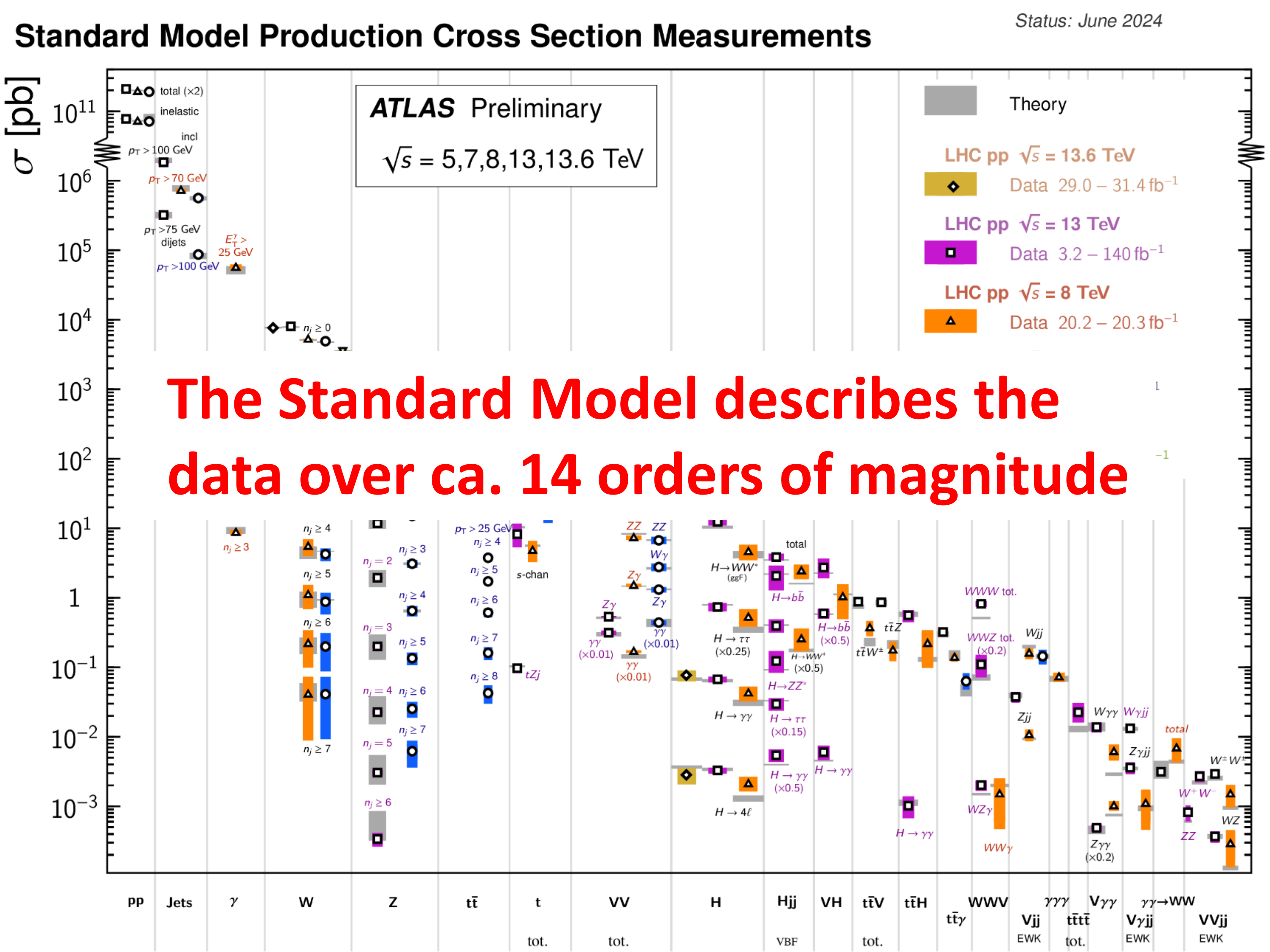
What do you conclude?



Source: ATLAS collaboration

What do you conclude?

technische universität dortmund



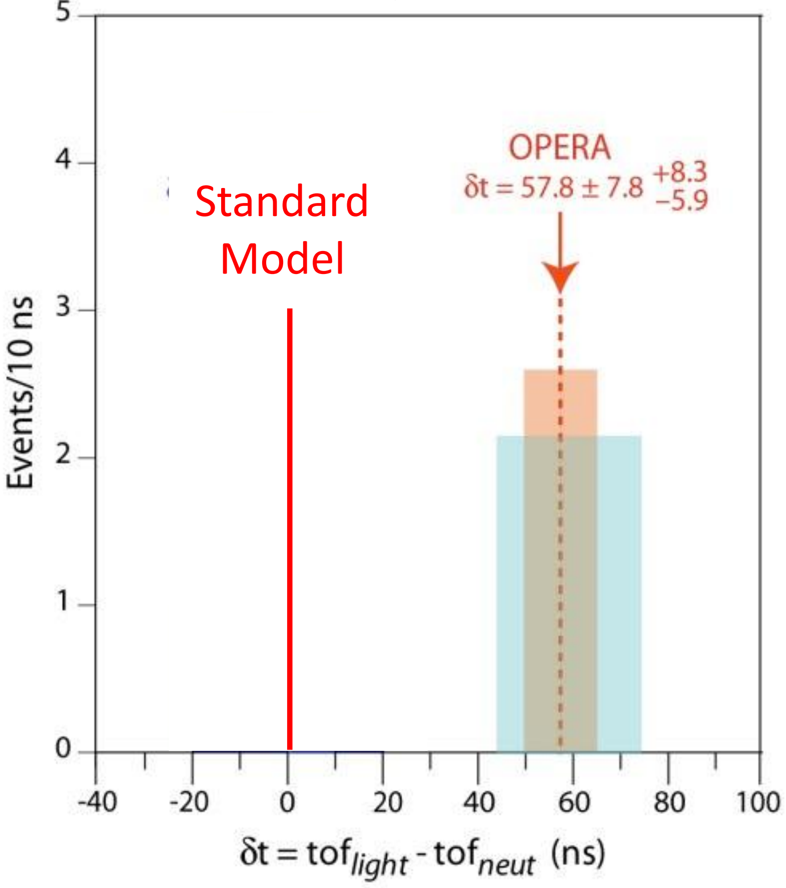
The Standard Model describes the data over ca. 14 orders of magnitude

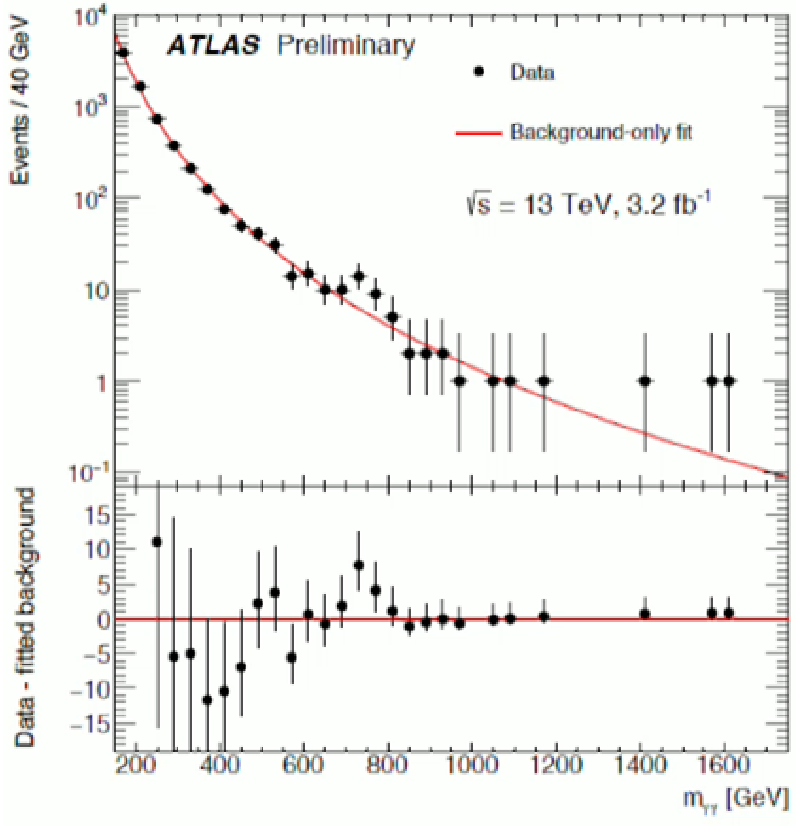
Source: ATLAS collaboration

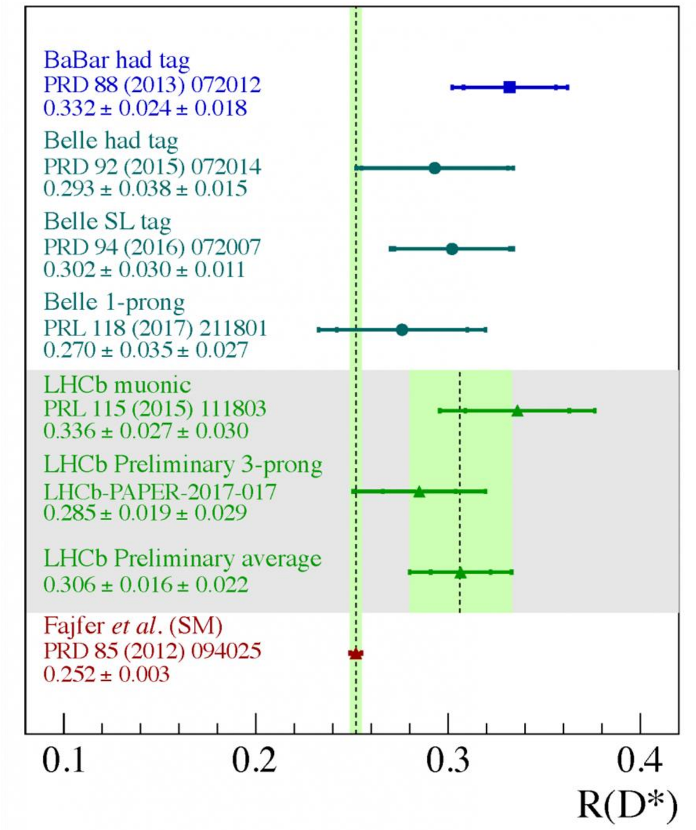
What do you conclude?

technische universität dortmund

2011	2015	2016/17
Phys. Lett. B **713** (2012) 17	*ATLAS-CONF-2015-081*	*Source: LHCb*





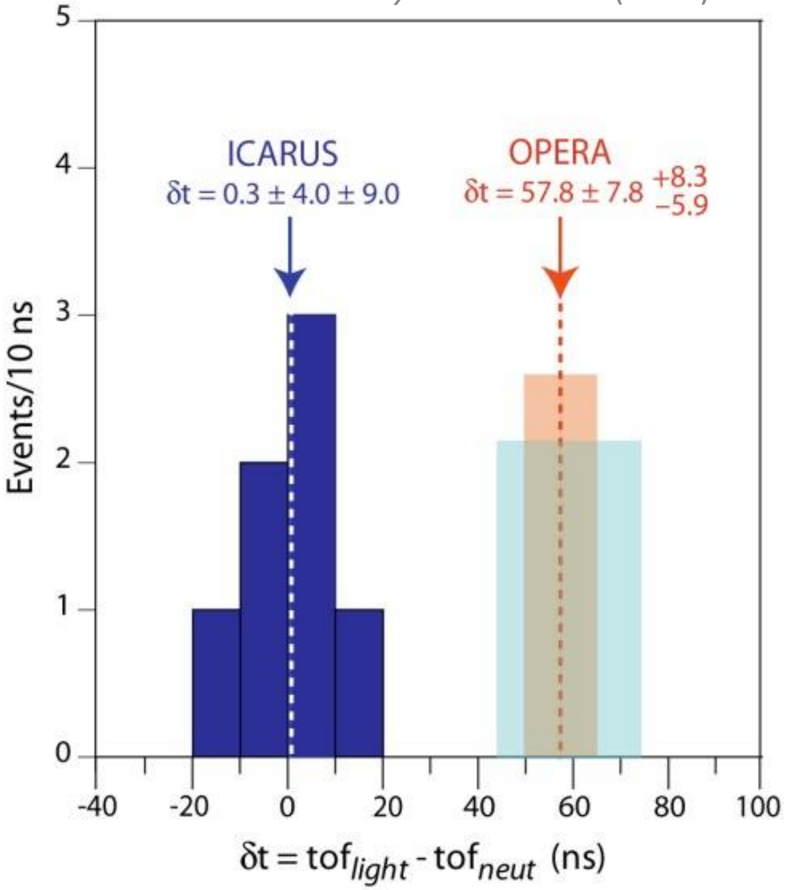


What do you conclude?

technische universität dortmund

2011

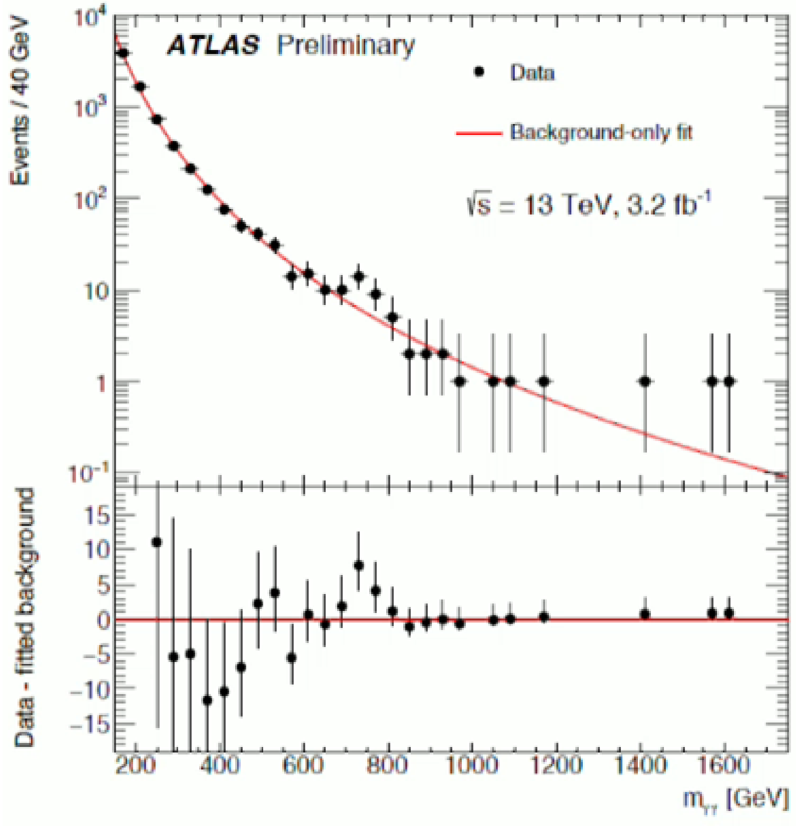
Phys. Lett. B **713** (2012) 17



“No, can't be...”

2015

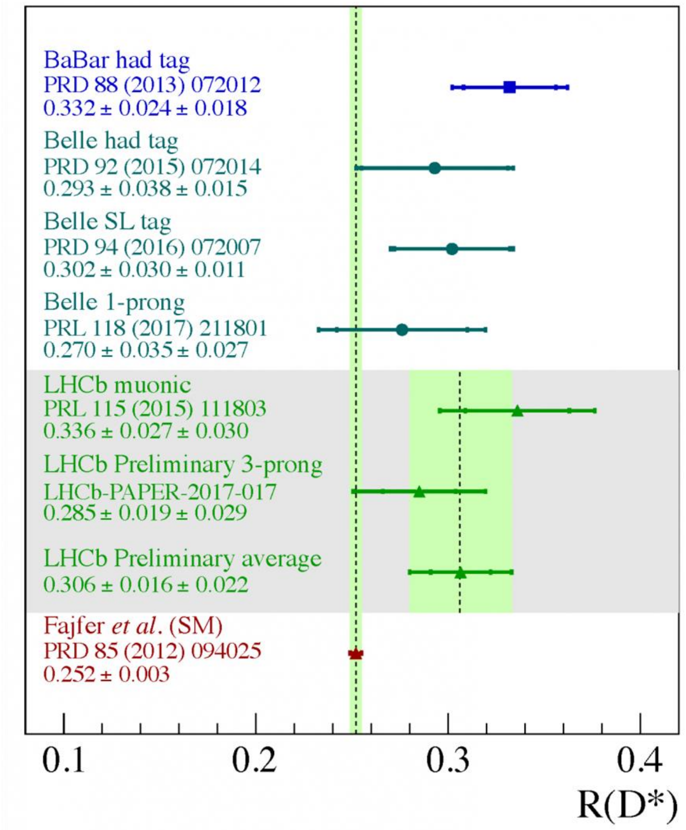
ATLAS-CONF-2015-081



“Could have been, but wasn't confirmed”

2016/17

Source: LHCb



“Mh, really?”

Scientific reasoning

Scientific reasoning

technische universität dortmund

Typical questions (and tasks) in (HEP) data analysis:

Given a model, what are the values of its free parameters?

→ **Parameter estimation**

Is the given model consistent with the data?

→ **Goodness-of-fit**

Which of the many models available describes the data best?

→ **Model comparison**

➔ **All three tasks are associated with meaningful probabilities**

Scientific reasoning

technische universität dortmund

Deductive reasoning

- Used when making predictions from a model
- Application in HEP:
  - Premise: model with parameters → Conclusion: predictions of observables
- Example:
  - Premise: The SM is correct (Lagrangian, perturbation theory, Fermis Golden rule, …)
  - Conclusion: a clear prediction for the cross section of a process, e.g. pp → Higgs
- Comments:
  - Given a model, the outcome is uniquely specified
  - **No need to argue, it's math!**

Scientific reasoning

Inductive reasoning

- Used when choosing a model
- Application in HEP:
  - Premise: model A with parameters → Conclusion: predictions of observables O
  - Revise the logic: Starting from a measurement O, what does it say about the model? Not much since it could have been A → O, A’ → O, A’’ → O, ...
- Validity of model A?
  - If we know all models, and only A results in O, then we know that A is true.
  - Otherwise, **we can not verify the model**.
  - Can try to **falsify the model**: if we observe something that contradicts the model, it can not be true
- Can we know which model is true? No!

Scientific reasoning

technische universität
dortmund

Inductive reasoning

HOME > SCIENCE > VOL. 338, NO. 6114 > A PARTICLE CONSISTENT WITH THE HIGGS BOSON OBSERVED WITH THE ATLAS DETECTOR AT THE LARGE HADRON…

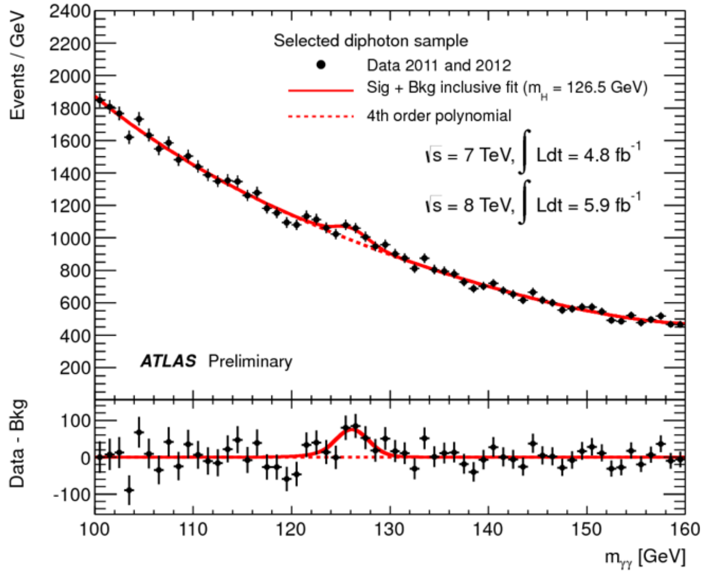
ARTICLE

A Particle Consistent with the Higgs Boson Observed with the ATLAS Detector at the Large Hadron Collider

THE ATLAS COLLABORATION , G. AAD, T. ABAJYAN, B. ABBOTT, J. ABDALLAH, S. ABDEL KHALEK, A.A. ABDELALIM, O. ABDINOV, R. ABEN, […], AND L. ZWALINSKI

+2923 authors Authors Info & Affiliations

SCIENCE • 21 Dec 2012 • Vol 338, Issue 6114 • pp. 1576-1582 • DOI: 10.1126/science.1232005



Are you sure it is *the* Higgs boson?

Scientific reasoning

technische universität dortmund

Knowledge

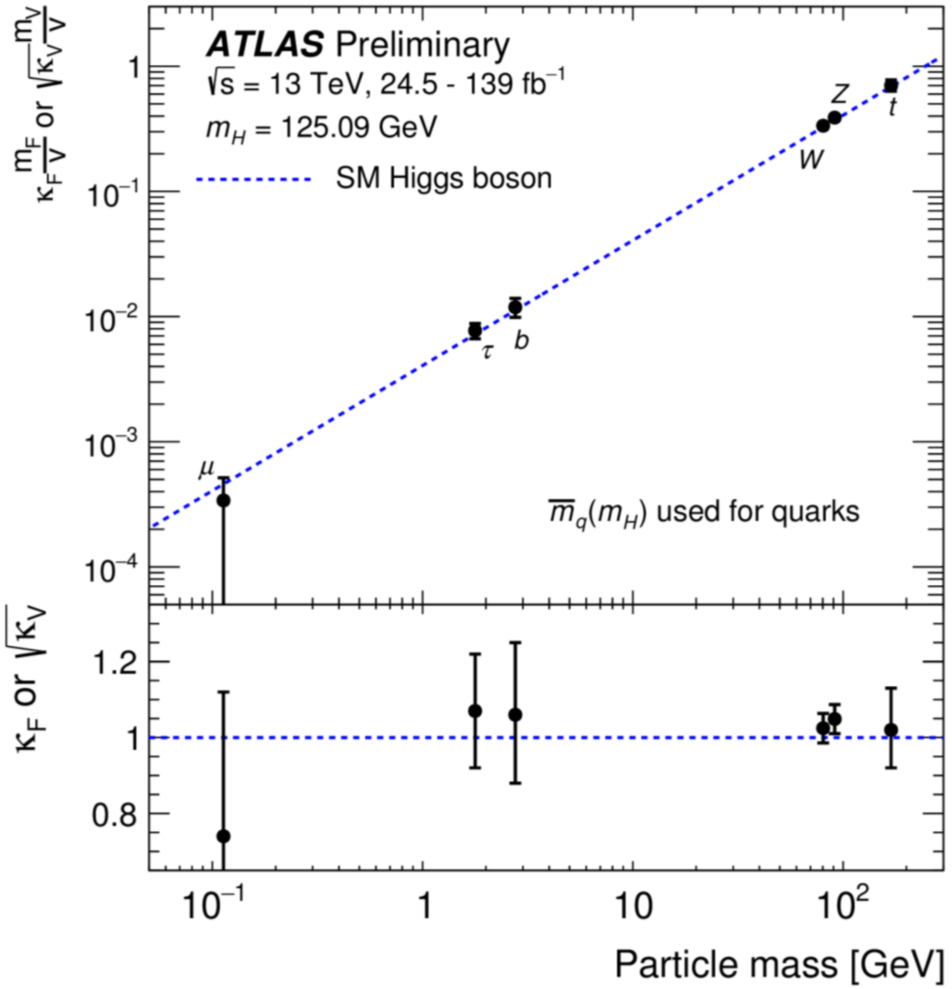
- Analytical philosophy (Plato?): knowledge is **justified true belief**.
- *S* knows that *P* if and only if
  - *P* is true,
  - *S* believes *P* to be true, and
  - this belief is justified.
- Discussion known as the Gettier problem
- Justification comes from observations:
  - Test model predictions
  - The more tests are passed, the greater the belief in the model...

Scientific reasoning

technische universität dortmund

Knowledge

- Analytical philosophy (Plato?): knowledge is **justified true belief**.
- *S* knows that *P* if and only if
  - *P* is true,
  - *S* believes *P* to be true, and
  - this belief is justified.
- Discussion known as the Gettier problem
- Justification comes from observations:
  - Test model predictions
  - The more tests are passed, the greater the belief in the model...



"A particle consistent with the Higgs boson" → "The Higgs boson"

Scientific reasoning

technische universität dortmund

Knowledge

- How do we gain knowledge?
  - Set up models and specify their parameters (check arXiv.org!)
  - Derive (deductively) predictions from the models
  - Good model: falsifiable, make predictions which can be proven wrong (Z' vs. SUSY vs. string theory)
  - Use data to gain knowledge about the models and their parameters
- Examples:
  - Special relativity predicts time dilation. Atmospheric muons can thus be observed on the earth's surface.
  - Neutrino postulation: Pauli was hesitant to publish his neutrino idea because he thought it would be difficult to discover.
- **Can we quantify the knowledge about a model? Yes, use probabilities**

Probability and Bayes Theorem

Probability and Bayes Theorem

technische universität dortmund

Axioms and interpretations

Kolmogorov axioms: start from a set S

1. For each subset *A*, assign probability *P(A)* between 0 and 1
2. Probability *P(S)*=1
3. For disjunct subsets *A* and *B*:

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

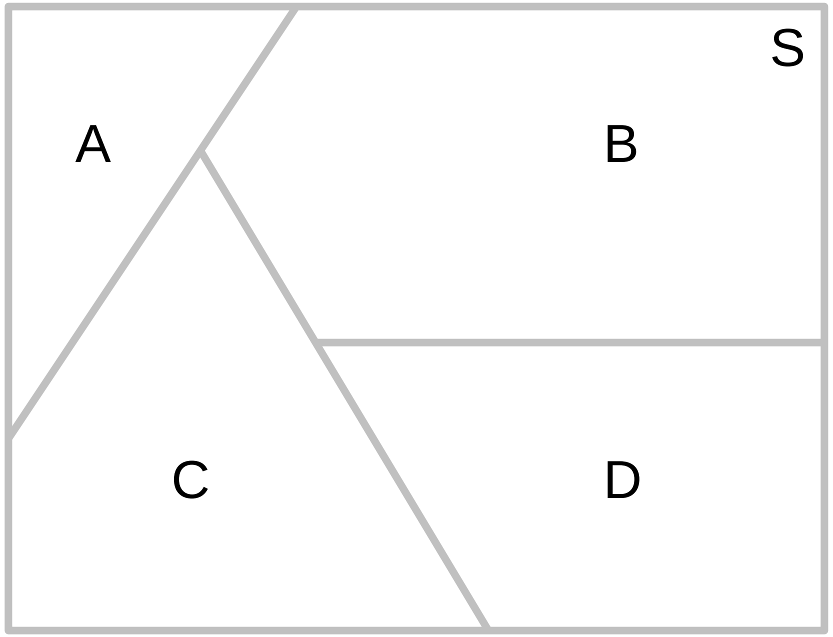
Nice mathematic formulation, but meaningless!

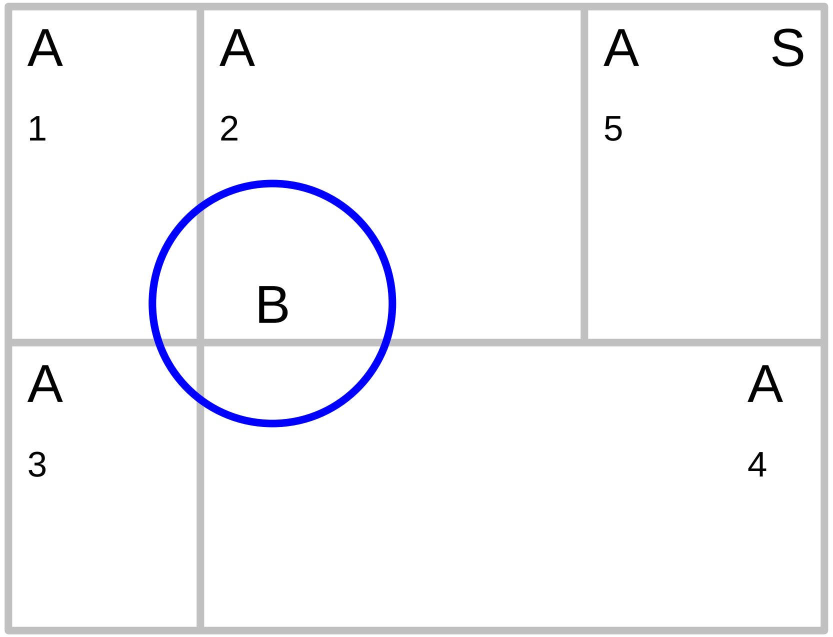
Law of total probability:

$$P(B) = \sum P(B|A_i) \cdot P(A_i)$$

Probability and Bayes Theorem

Kolmogorov axioms





$$P(A) \geq 0, \dots \quad A \cap B = \emptyset, \dots$$

$$S = A \cup B \cup C \cup D$$

$$P(S) = P(A) + P(B) + P(C) + P(D) = 1$$

$$P(B) = \sum P(B|A_i) \cdot P(A_i)$$

$$= P(B|A_1) \cdot P(A_2) + \dots + P(B|A_5) \cdot P(A_5)$$

= 0

Probability and Bayes Theorem

technische universität dortmund

Bayesian interpretation

- Subsets correspond to hypotheses, i.e. a model with a particular value of the parameter.
- Probability is understood as **degree-of-belief** (or **state-of-knowledge**) for this hypothesis to be true
- Interpretation fully consistent with Kolmogorov axioms.
- Gives meaning to the term probability.
- Examples:
  - Probability that it will rain tomorrow
  - SM and the masses of particles

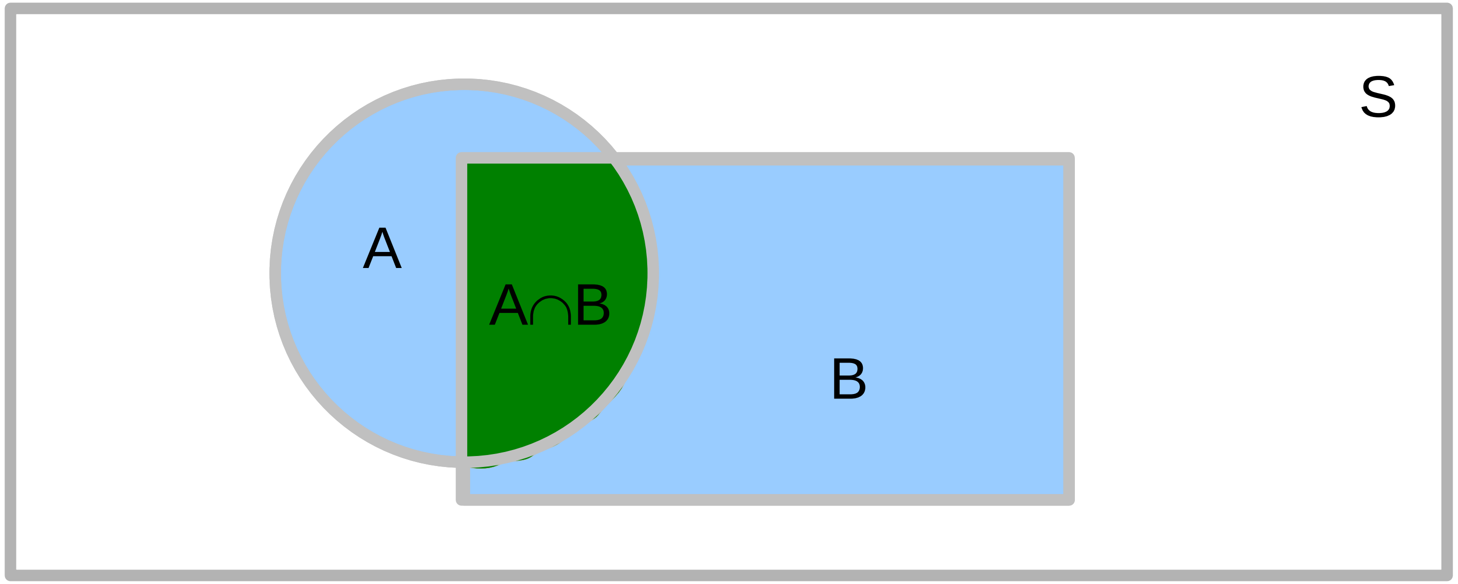
Probability and Bayes Theorem

Bayes Theorem

$$P(A|B) \cdot P(B) = P(A \cap B) = P(B|A) \cdot P(A)$$

$$\Leftrightarrow \qquad P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$\Leftrightarrow \qquad P(A|B) = \frac{P(B|A) \cdot P(A)}{\sum P(B|A_i) \cdot P(A_i)}$$



Probability and Bayes Theorem

technische universität dortmund

Bayes Theorem

$$P(A|B) \cdot P(B) = P(A \cap B) = P(B|A) \cdot P(A)$$

$$\Leftrightarrow \quad P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$\Leftrightarrow \quad P(A|B) = \frac{P(B|A) \cdot P(A)}{\sum P(B|A_i) \cdot P(A_i)}$$

Here:

$$P(\text{model} \mid \text{data}) \sim P(\text{data}|\text{model}) \cdot P(\text{model})$$

P(model | data): posterior probability (induction)

P(data | model): probability of the data, likelihood (deduction)

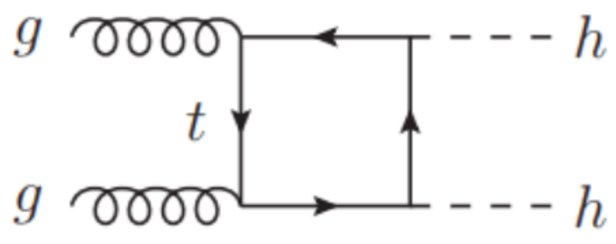
P(model): prior probability

In words: "My degree-of-belief of the model (given the data) is x%"

Probability and Bayes Theorem

technische universität dortmund

Bayes Theorem – An example: counting experiment



- Consider the search for a rare process, e.g. two-Higgs-production
- Assume that there is no background
- The number of observed events *N* is Poisson-distributed around the expected number λ:

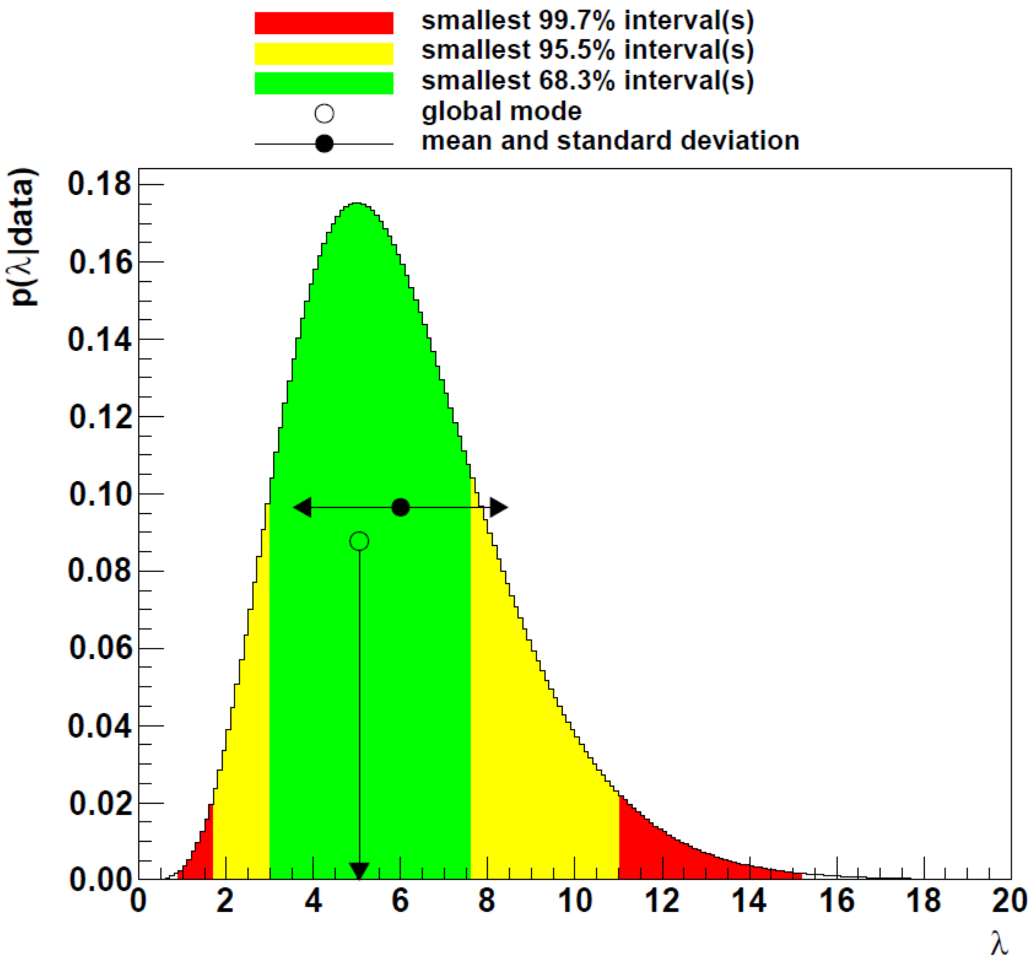
$$P(N|\lambda) = \frac{\lambda^N}{N!} e^{-\lambda}$$

- The expected number of events can be estimated using Bayes' Theorem.
- Since λ is continuous, so is the probability (density):

$$P(\lambda|N) = \frac{\frac{\lambda^N}{N!} e^{-\lambda} \cdot p(\lambda)}{\int_0^{\lambda_{\max}} \frac{\lambda^N}{N!} e^{-\lambda} \cdot p(\lambda) \mathrm{d}\lambda}$$

Bayes Theorem – An example: counting experiment

- Assume further N=5 observed events
- What can we put as prior for λ?
- For now: assume all values to be equally likely (uniform prior)
- Result: “The parameter value lies in an interval [3, 7.5] with 68.3% probability.”

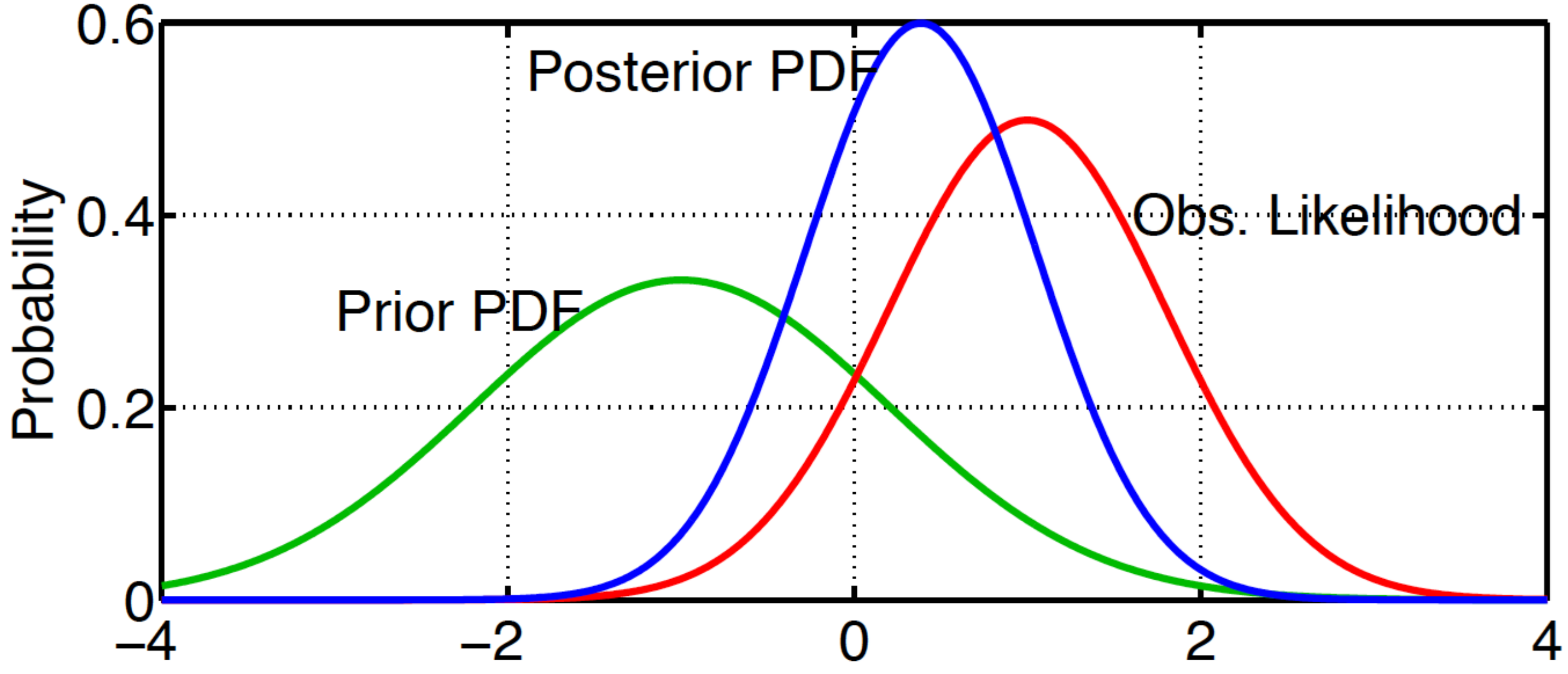


Pro: there is no funny construction!

Con: What if we have chosen a different prior?

Probability and Bayes Theorem

Prior probabilities



"A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule" – Stephen Senn, Statistician & Bayesian Skeptic (mostly)

Probability and Bayes Theorem

Prior probabilities

- Where does prior knowledge come from?
- Prior knowledge can come from
  - personal degree-of-belief (“gut feeling”, “physics intuition”, prejudice (“do you *want* SUSY to be true?”, …),
  - theoretical considerations (masses cannot be negative, charges are quantized, …),
  - auxiliary measurements, …
  - … good arguments … (in the best case)
- **Examples:**
  - Physical constraints, e.g. positive cross-sections, unitarity, …
  - Measurement of background contributions in counting experiments
  - Considering other types of measurements, e.g. searches for Dark Matter at colliders vs. direct-detection experiments

Probability and Bayes Theorem

Prior probabilities

- Elegant update of knowledge:
  - posterior of one experiment can be prior of another experiment.
  - Natural way to combine measurements.
  - Reflects human way of learning

$$P(\mathrm{model}|\mathrm{DS1}) \sim P(\mathrm{DS1}|\mathrm{model}) \cdot P(\mathrm{model})$$

$$P(\mathrm{model}|\mathrm{DS2}) \sim P(\mathrm{DS2}|\mathrm{model}) \cdot P'(\mathrm{model})$$

$$P'(\mathrm{model}) = P(\mathrm{DS1}|\mathrm{model}) \cdot P(\mathrm{model})$$

$$\begin{aligned} P(\mathrm{model}|\mathrm{DS2}) &= P(\mathrm{DS1}|\mathrm{model}) \cdot P(\mathrm{DS2}|\mathrm{model}) \cdot P(\mathrm{model}) \\ &= P(\mathrm{model}|\mathrm{DS1},\mathrm{DS2}) \end{aligned}$$

Probability and Bayes Theorem

technische universität dortmund

Prior probabilities - criticism

- **Priors are subjective**
  - Yes, but it is made explicit
  - Objective Bayesian movement, try to find objective priors
  - **reference priors** minimize the "information"
- **Prior depends on parametrization**
  - Example: lifetime τ vs. decay constant $\lambda=1/\tau$
  - **Jeffreys prior** invariant under re-parameterization

Prior probabilities - remarks

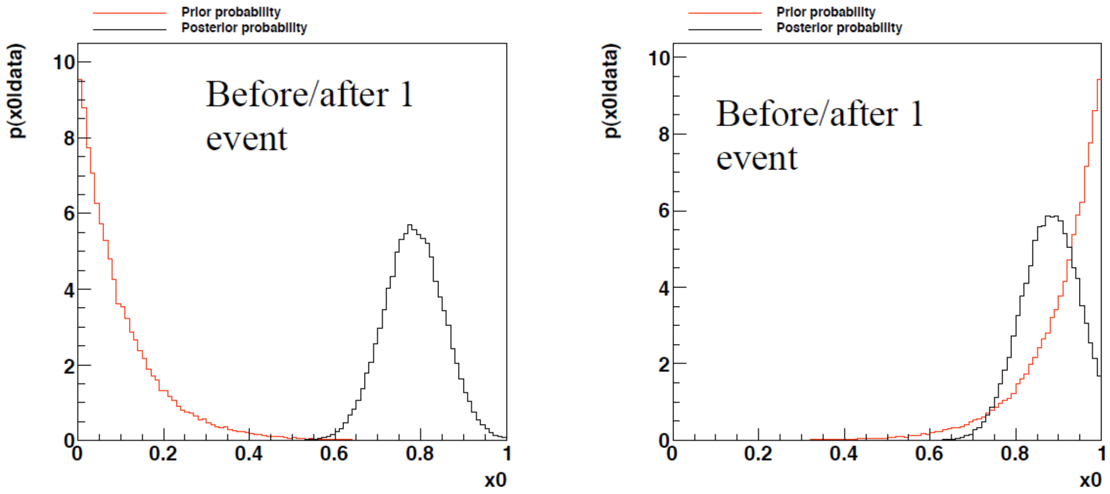
- **Choice of (initial) prior should not play a strong role.**
- Difficult to formulate one prior for a collaboration of about 3.000 people
- Practical solution: Requote your result with different priors ("the optimist", "the pessimist", "the ignorant", ...)
- **Write down your prior!**

Probability and Bayes Theorem

technische universität dortmund

Prior probabilities – an example

- Assume a model with a free parameter $x_0 = 0.75$
- Likelihood: Gaussian with mean value between 0 and 1 and std. dev. of 0.1
- Priors: optimistic vs. pessimistic

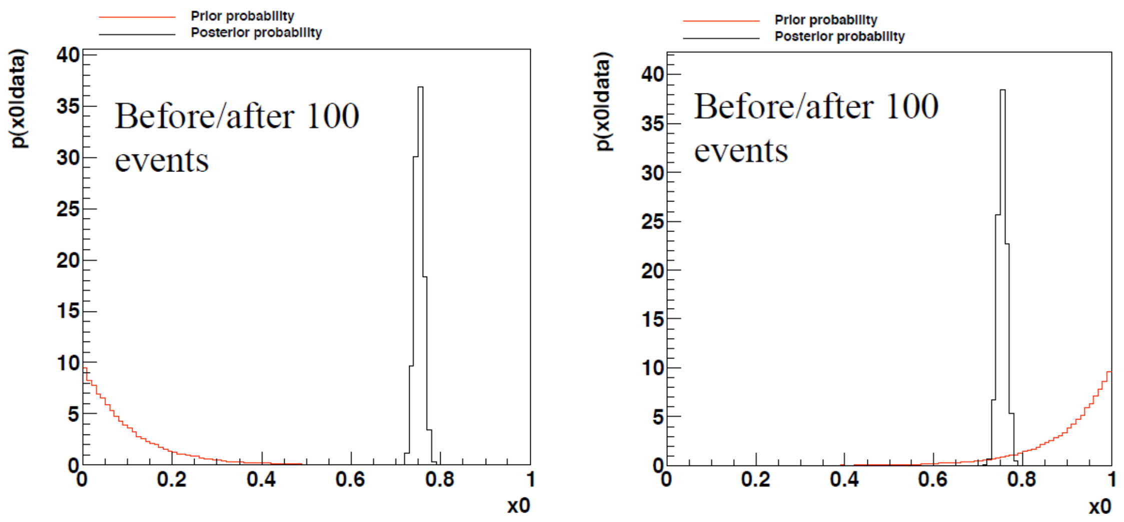


→ Slightly different posteriors after one event.

Probability and Bayes Theorem

Prior probabilities – an example

- Assume a model with a free parameter x0=0.75
- Likelihood: Gaussian with mean value between 0 and 1 and std of 0.1
- Priors: optimistic vs. pessimistic



→ About the same posterior after 100 events

Markov Chain Monte Carlo

Markov Chain Monte Carlo

technische universität dortmund

Numerical considerations

- Point estimate for parameters:
  - Maximization of posterior
  - Typical tool: Minuit
  - Also: Simulated annealing
- Calculation of marginal distributions:
  - Analytical solutions usually difficult
  - Numerical integration methods
- Sampling methods:
  - Hit&miss, simple Monte Carlo, ...
  - Importance sampling
  - Markov Chain Monte Carlo (MCMC)
- Model comparison:
  - Analytical solutions usually difficult
  - Numerical integration methods, e.g. VEGAS

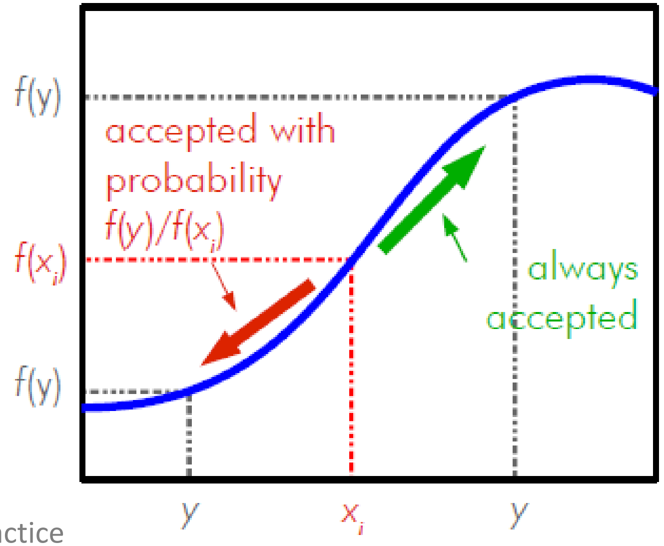
MCMC Methods and computing resources have made Bayesian computation possible

Markov Chain Monte Carlo

How does Markov Chain Monte Carlo work?

- Output of Bayesian analyses are posterior probability densities, i.e., functions of an arbitrary number of parameters (dimensions)
- Sampling large dimensional functions is difficult
- Idea: use random walk biased towards region of larger values (probabilities)
- Metropolis algorithm: N. Metropolis *et al.*, J. Chem. Phys. **21** (1953) 1087

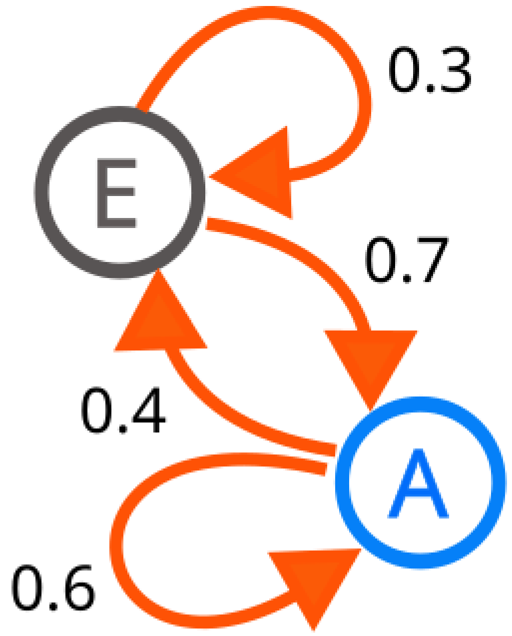
- Start at some randomly chosen x_i
- Randomly generate y around x_i
- If $f(y) > f(x_i)$ set $x_{i+1} = y$
- If $f(y) < f(x_i)$ set $x_{i+1} = y$ with prob. $p=f(y)/f(x_i)$
- If y is not accepted set $x_{i+1} = x_i$
- Start over



Markov Chain Monte Carlo

Remarks on MCMC

- In general: MCMC is a class of algorithms that is used for drawing samples from a distribution in the form of a **Markov Chain**
- The elements of the Markov Chain approximate the underlying distribution
- There are a lot for concrete MCMC algorithms on the market, e.g. Metropolis, Gibbs sampling, Hamiltonian Monte Carlo, etc.
- Efficient algorithms for high-dimensional sampling, but comes with some limitations, e.g. auto-correlation and multimodal distributions



Source: Wikipedia (Joxemai4)

Markov Chain Monte Carlo

technische universität dortmund

MCMC for Bayesian inference

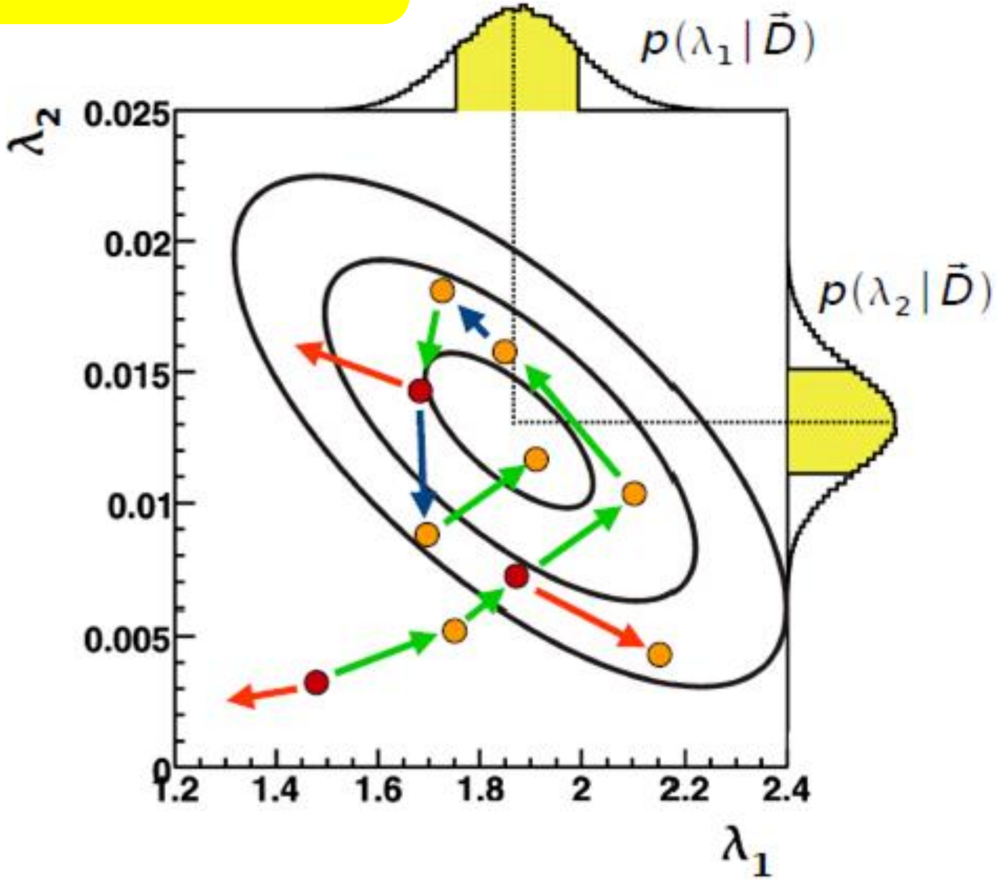
- Use MCMC to sample the posterior probability, i.e.

$$f\left(\vec{\lambda}\right) = p\left(D\middle|\vec{\lambda}\right)p\left(\vec{\lambda}\right)$$

- Marginalization of posterior:

$$p(\lambda_i|D) = \int p\left(D\middle|\vec{\lambda}\right)p\left(\vec{\lambda}\right)\prod_j^n \mathrm{d}\lambda_j$$

- Fill a histogram with just one coordinate while sampling
- Error propagation: calculate any function of the parameters while sampling



A concrete example

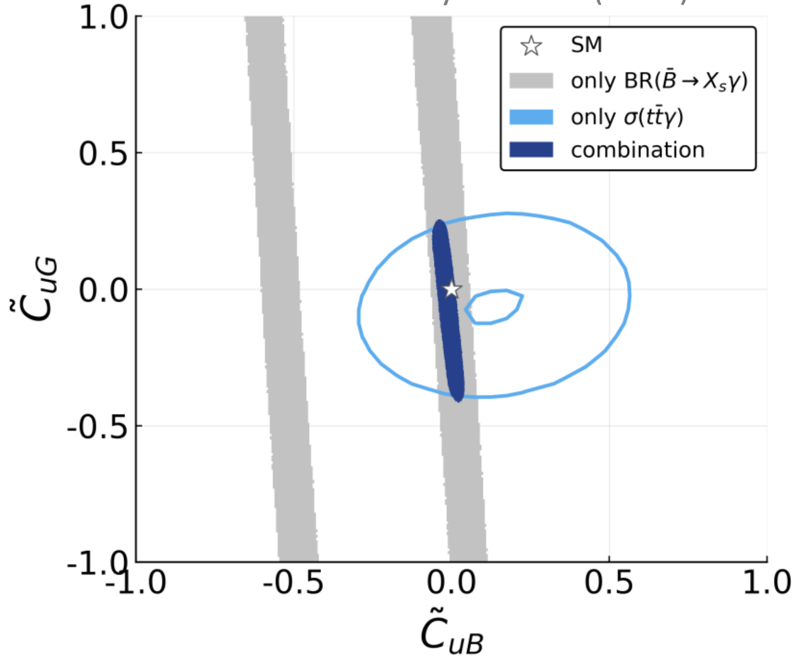
A concrete example

technische universität dortmund

Applications

- Various applications of Bayesian reasoning
- Focus on problems for which the posterior is not well behaved
  - Large (>3) number of dimensions
  - Multimodal distributions
  - Distributions with non-linear correlations
- Examples include
  - Fitting of Wilson coefficients in SMEFT (O(14) parameters and O(30) measurements)
  - **Tuning of Monte Carlo generators** (8 parameters and O(100) measurements)
  - Using MCMC for event generation (>13 parameters)

Eur. Phys. J. C **80** (2020) 136

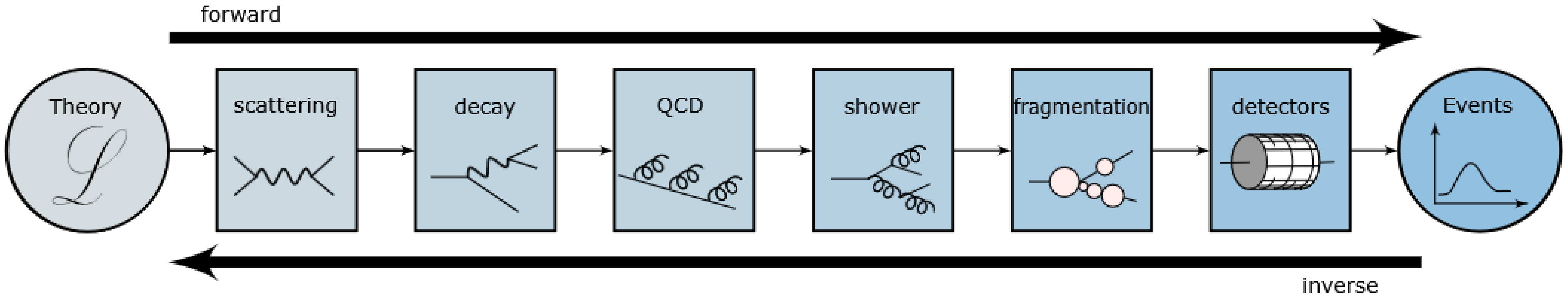


A concrete example

technische universität
dortmund

Event generators

- Programs to calculate an ensemble of collider events on the parton and/or particle level, e.g. Pythia, Herwig, MadGraph, Sherpa, …
- Input: usually prescription for the SM with a certain precision (mostly NLO plus some corrections), rare decays, special processes, BSM physics, …
- Come with free parameters that need adjustment ("tuning") to data



A concrete example

technische universität dortmund

Example: Tuning the Herwig event generator

- Herwig: Multi-purpose event generator including showering
- Consider only a subset of the **free parameters: $\vec{\lambda}$**

Parameter	Range	Default
AlphaQCD	$[0.1000, 0.1417]$	0.1181
IRcutoff (GeV)	$[0.5004, 1.5012]$	1.0080
m_g (GeV)	$[0.7445, 1.1400]$	0.9500
m_s (GeV)	$[0.4734, 0.5000]m_g$	0.4500
ClMax (GeV)	$[1.9334, 5.8000]$	3.8667
ClPow	$[0.8295, 2.4885]$	1.6590
ClSmr	$[0.1719, 0.8593]$	0.3437
PSplit	$[0.3450, 1.0348]$	0.6899

JINST 18 (2023) 10033

Side remark:
this is prior knowledge

- Consider O(100) **data distributions from LEP (e^+e^-): $\vec{D}$**

A concrete example

Procedure

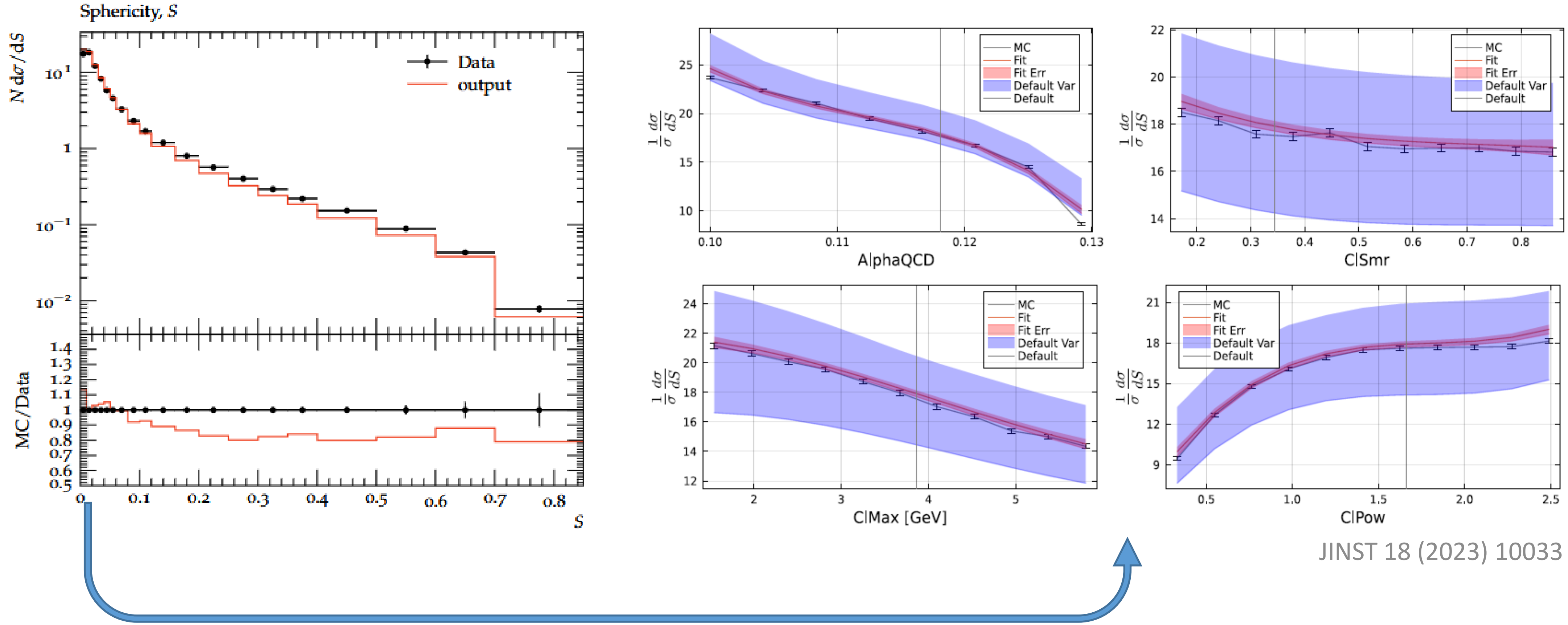
1. Parameterize observables as a function of the free parameters
2. Formulate a likelihood
3. Formulate your prior knowledge
4. Perform the fit
5. Recalculation of the observables assuming the tuned values

A concrete example

technische universität dortmund

1. Parameterize observables as a function of the free parameters

- **Predictions $\vec{f}\left(\vec{\lambda}\right)$** are parameterized by multidimensional cubic polynomials



JINST 18 (2023) 10033

A concrete example

technische universität dortmund

2. Formulate a likelihood

- Assume (symmetric) Gaussian uncertainties for each measurement
- Consider possible correlations among measurements, e.g. from luminosity, detector effects, theory uncertainties
- Likelihood:

$$\ln L\left(\vec{D}\middle|\vec{\lambda}\right) = -\frac{1}{2}\left[\vec{D} - \vec{f}\left(\vec{\lambda}\right)\right]^T \cdot M^{-1} \cdot \left[\vec{D} - \vec{f}\left(\vec{\lambda}\right)\right]$$

data (vector of measured values)

parameters

covariance matrix

prediction as a function of the parameters

3. Formulate your prior knowledge

- Assume uniform priors for all parameters

A concrete example

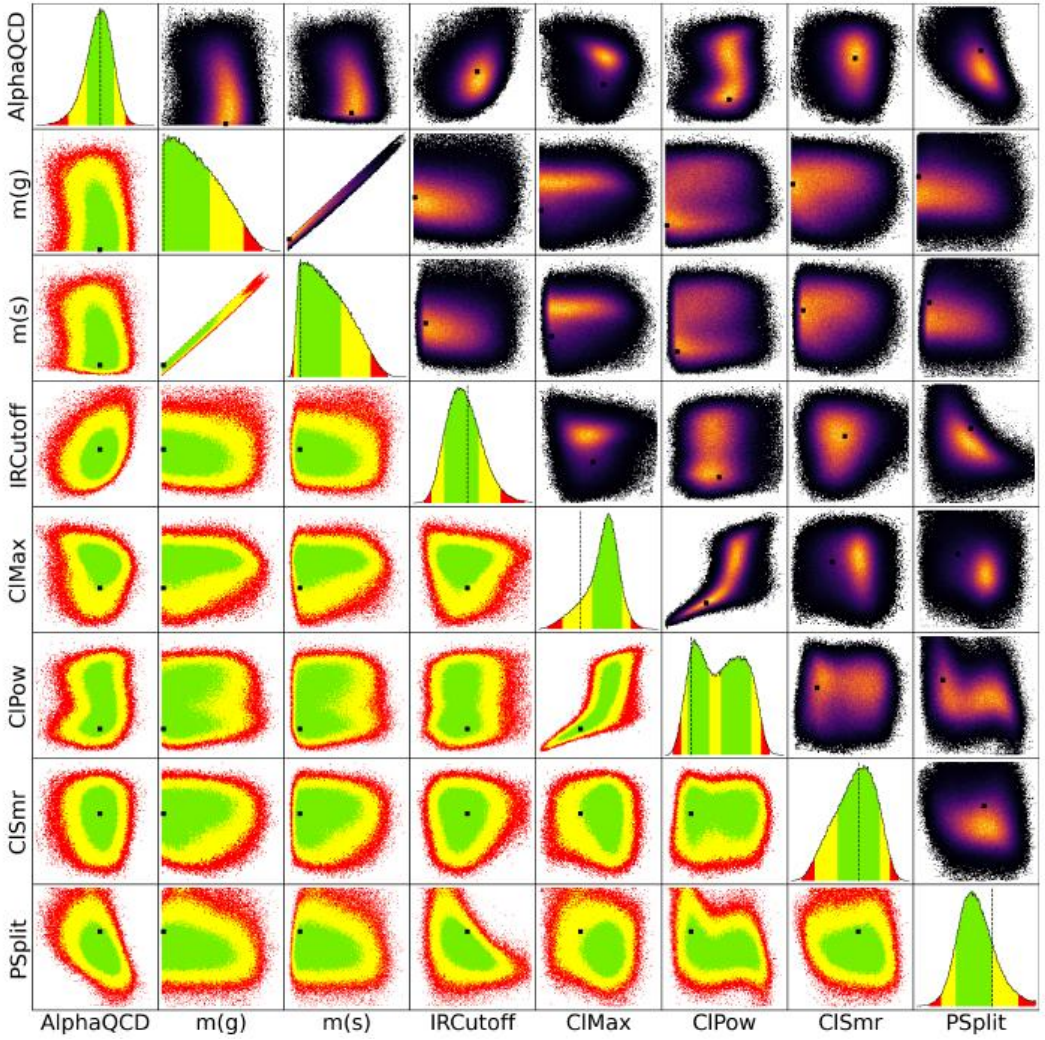
technische universität dortmund

4. Perform the fit

BAT.jl
Bayesian Analysis Toolkit

https://github.com/bat/BAT.jl

Parameter	Global mode	Marginal mode	Smallest 68% interval
AlphaQCD	0.115	0.115	[0.112, 0.118]
IRCutoff (GeV)	0.879	0.755	[0.580, 1.020]
m_g(GeV)	0.709	0.738	[0.700, 0.955]
m_s(GeV)	0.353	0.375	[0.346, 0.470]
ClMax (GeV)	2.591	4.025	[3.200, 4.750]
ClPow	0.823	0.910	[0.740, 1.260], [1.540, 2.260]
ClSmr	0.675	0.725	[0.480, 0.885]
PSplit	0.868	0.728	[0.615, 0.865]



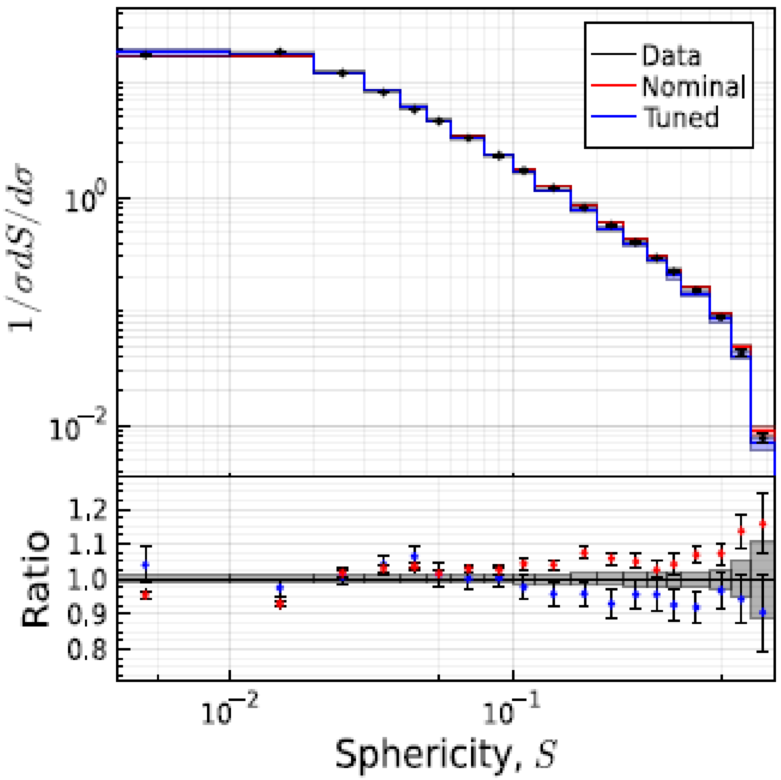
JINST 18 (2023) 10033

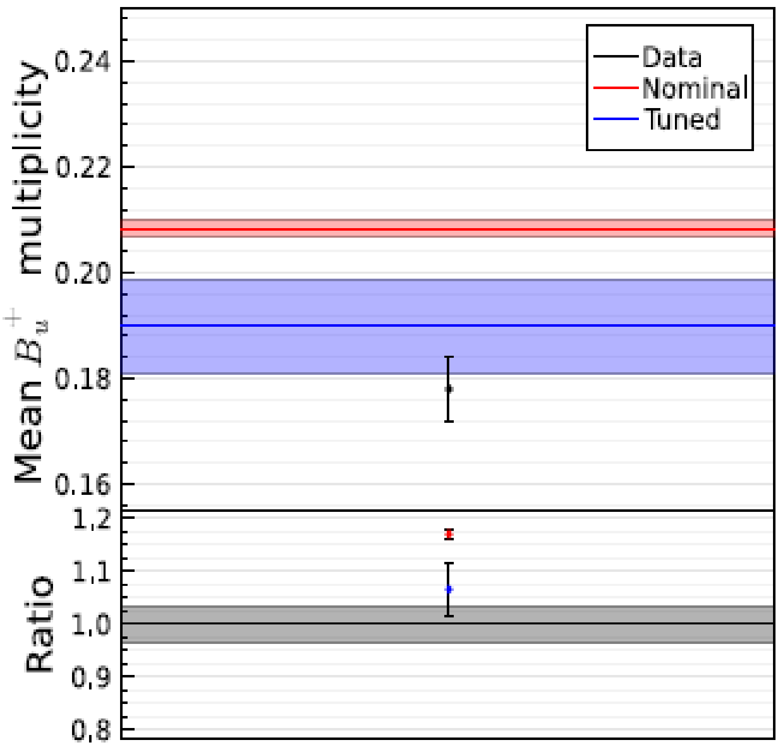
A concrete example

technische universität dortmund

5. Recalculation of the observables assuming the tuned values

- Use the global mode and do uncertainty propagation respecting the correlations (“uncertainty on the tuned values”)





JINST 18 (2023) 10033

A concrete example

technische universität dortmund

Some remarks

- Tuning of event generators is parameter estimation
  - Point estimation
  - Interval estimation
- One could do more, e.g.
  - Hypothesis testing: does the event generator describe the data well?
  - Model comparison: which event generator does a better job in describing the data?
  - Data quality: are all data sets consistent with each other?
  - ...
- **Very good application for Bayesian reasoning!**

Conclusion

technische universität
dortmund

Two answers for today

What does the concept of probability mean for you as a person?

→ Scientific reasoning

→ Probability and Bayes Theorem

How can Bayesian quantities be computed?

→ Markov Chain Monte Carlo

→ A concrete example