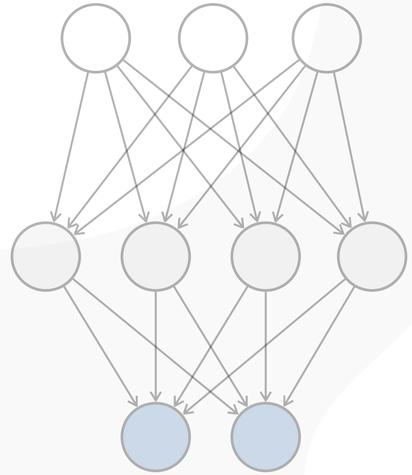
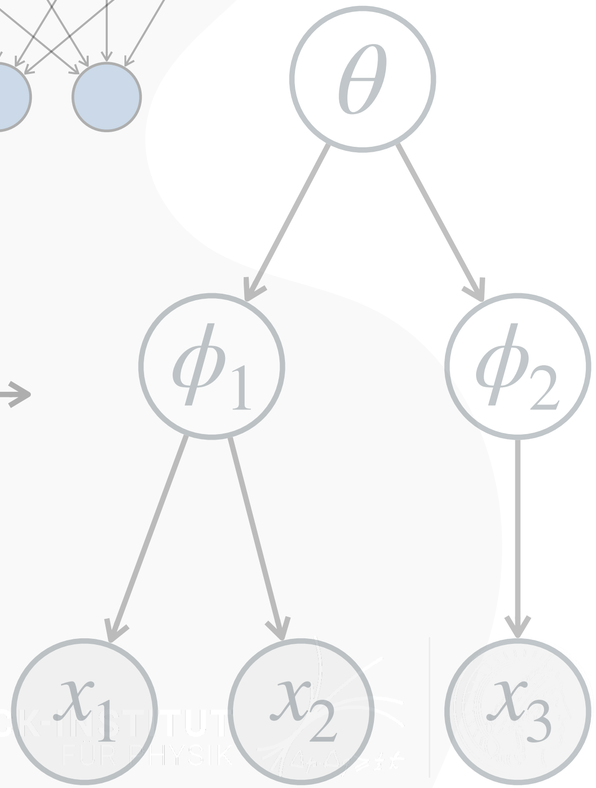
# Statistical Modelling







Francesca Capel

Max Planck Institute for Physics

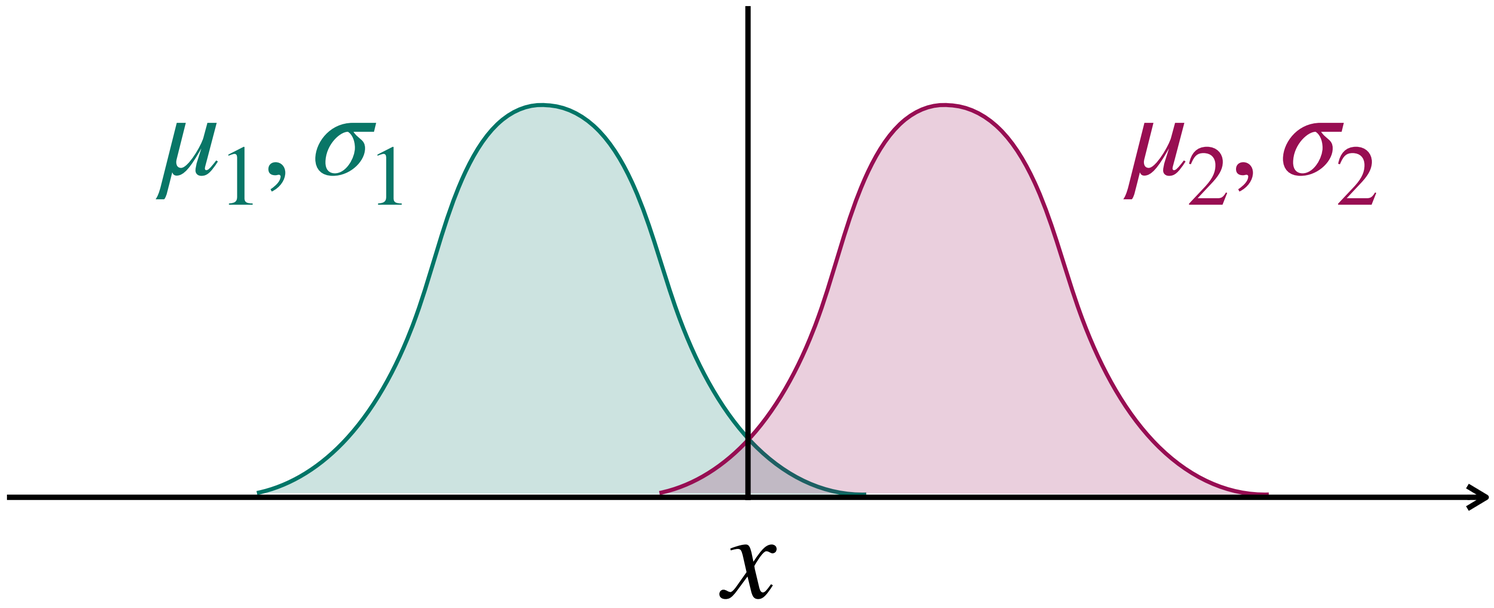

29th March 2025 - TAML Workshop @ ICRR

ICRR
Institute for Cosmic Ray Research
University of Tokyo

MAX-PLANCK-INSTITUT
FÜR PHYSIK

# Statistics and machine learning

A textbook statistics example:




$$\mathscr{L}(x \,|\, \mu, \sigma) = w\, \mathscr{N}(x \,|\, \mu_1, \sigma_1) + (1 - w)\, \mathscr{N}(x \,|\, \mu_2, \sigma_2)$$

# Statistics and machine learning

A realistic ML example:

$$\mathscr{L}(x \mid \text{muffin}, \text{dog}) = w\, \text{p}(x \mid \text{muffin}) + (1 - w)\, \text{p}(x \mid \text{dog})$$

# Statistics and machine learning

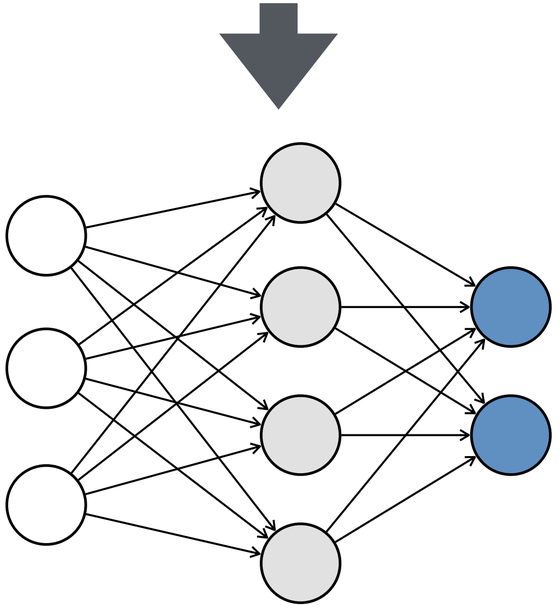
We can think about neural networks as a bunch of weights that figure our how to describe the likelihood

**Training:** Figure out the likelihood (connection between data and model)

**Evaluating:** Maximum likelihood solution

ML in physics: We care about interpretation and uncertainty → No escape from statistics

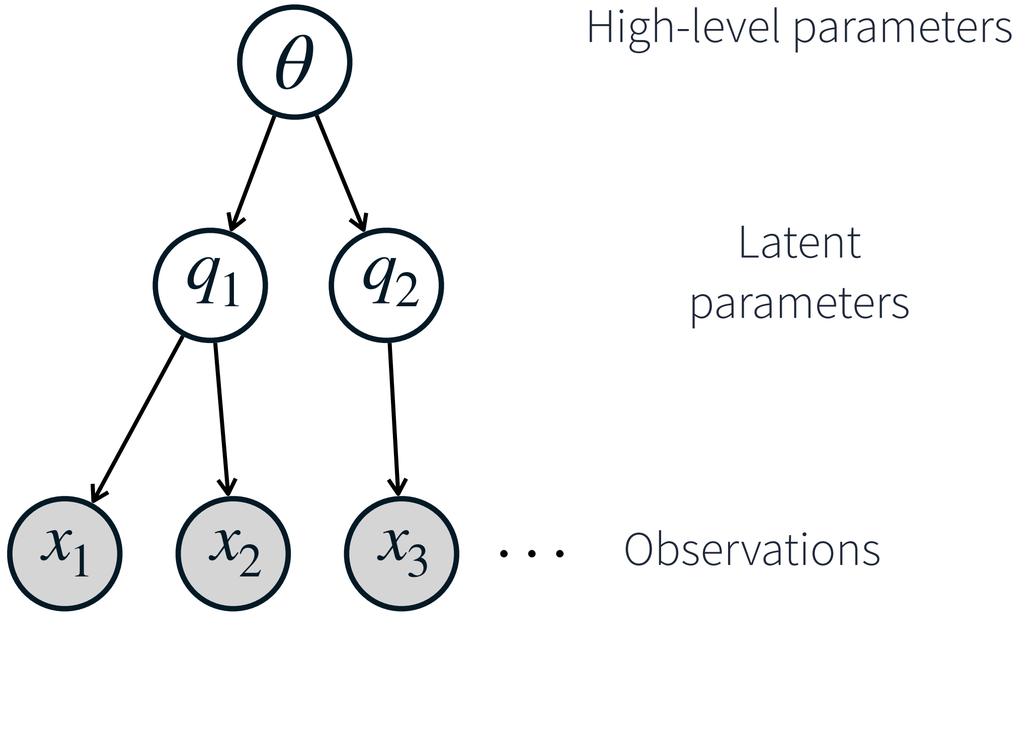

$$\mathscr{L}(x \,|\, \text{muffin}, \text{dog}) = w\, \text{p}(x \,|\, \text{muffin}) + (1 - w)\, \text{p}(x \,|\, \text{dog})$$



# Forward modelling

In astroparticle physics we have “models” for our data, typically in the form of Monte Carlo simulations

- Quantify expected uncertainties
- Map out complexity
- Contains all information needed to derive the likelihood function
- Testbench to evaluate possible simplifications or assumptions

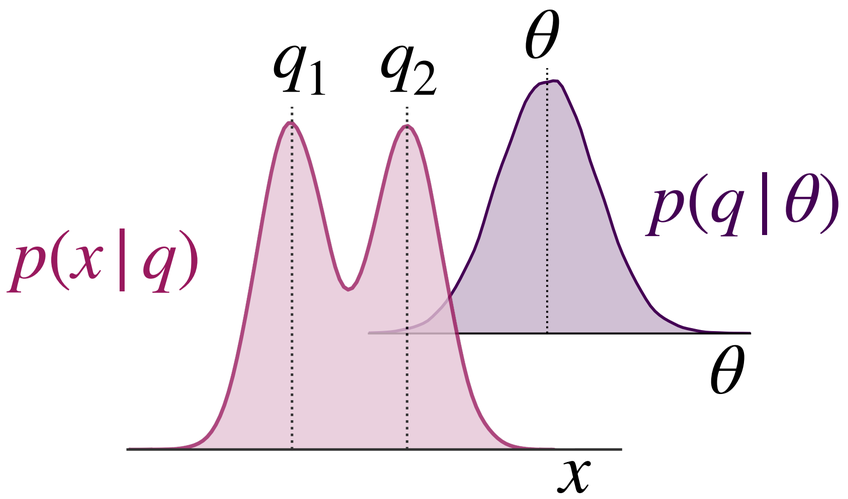


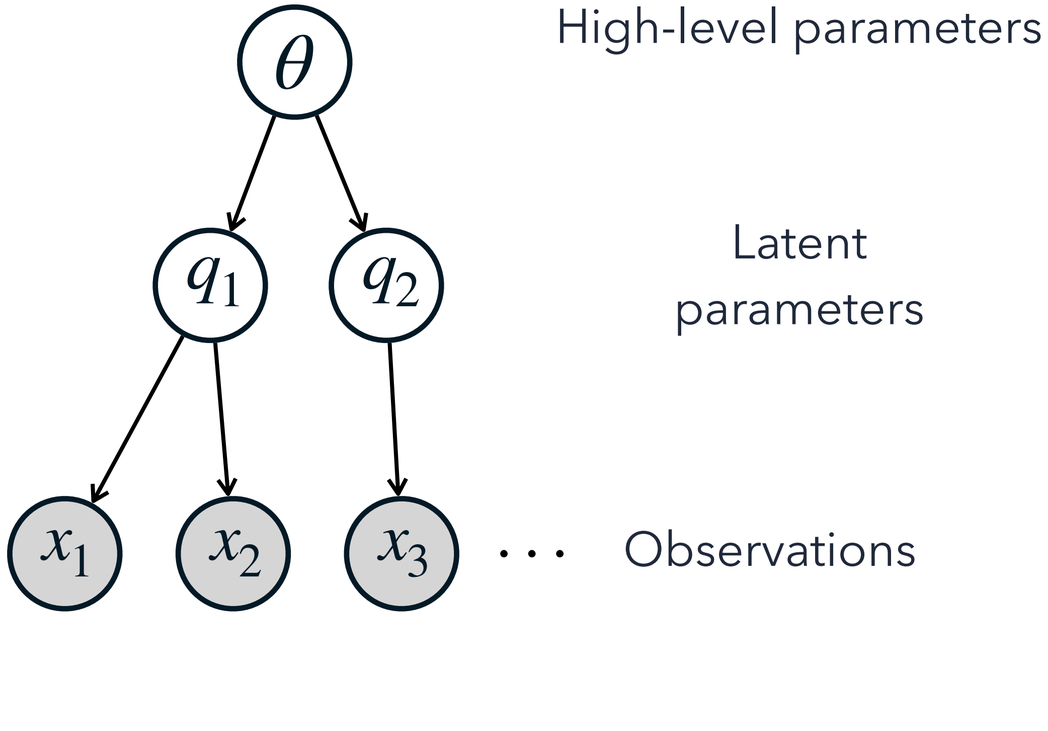
# Forward model = likelihood

Organise the free parameters into a statistical model that describes the **data generating process**

**Lots of free parameters → machine learning is useful (but not only way)**

$$\mathscr{L}(x, \theta) = p(x \mid q)\ p(q \mid \theta) \ldots$$



# Inference in many dimensions

### Markov chain Monte Carlo

Numerically approximate **high-dimensional integrals** (e.g. expectation values, variances of parameters)

Demo: https://chi-feng.github.io/mcmc-demo/app.html

$$\int_\Theta \mathrm{d}\theta\, f(\theta)p(\theta) = \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N} f(\theta_n)$$

Large numbers of free parameters possible

### Hamiltonian Monte Carlo

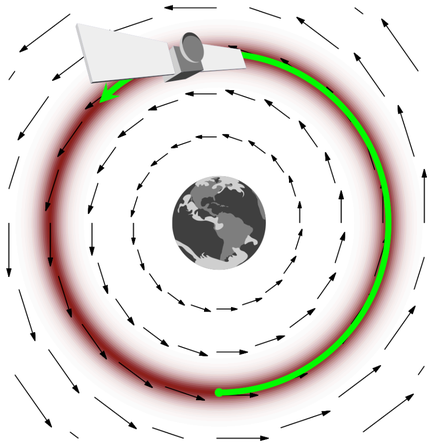
Markov chain Monte Carlo that uses Hamiltonian dynamics to move efficiently through high-dimensional parameter spaces

$$\theta \longrightarrow (\theta, p)$$

$$H(\theta, p) \equiv \log P(\theta, p)$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{\partial H}{\partial p}$$

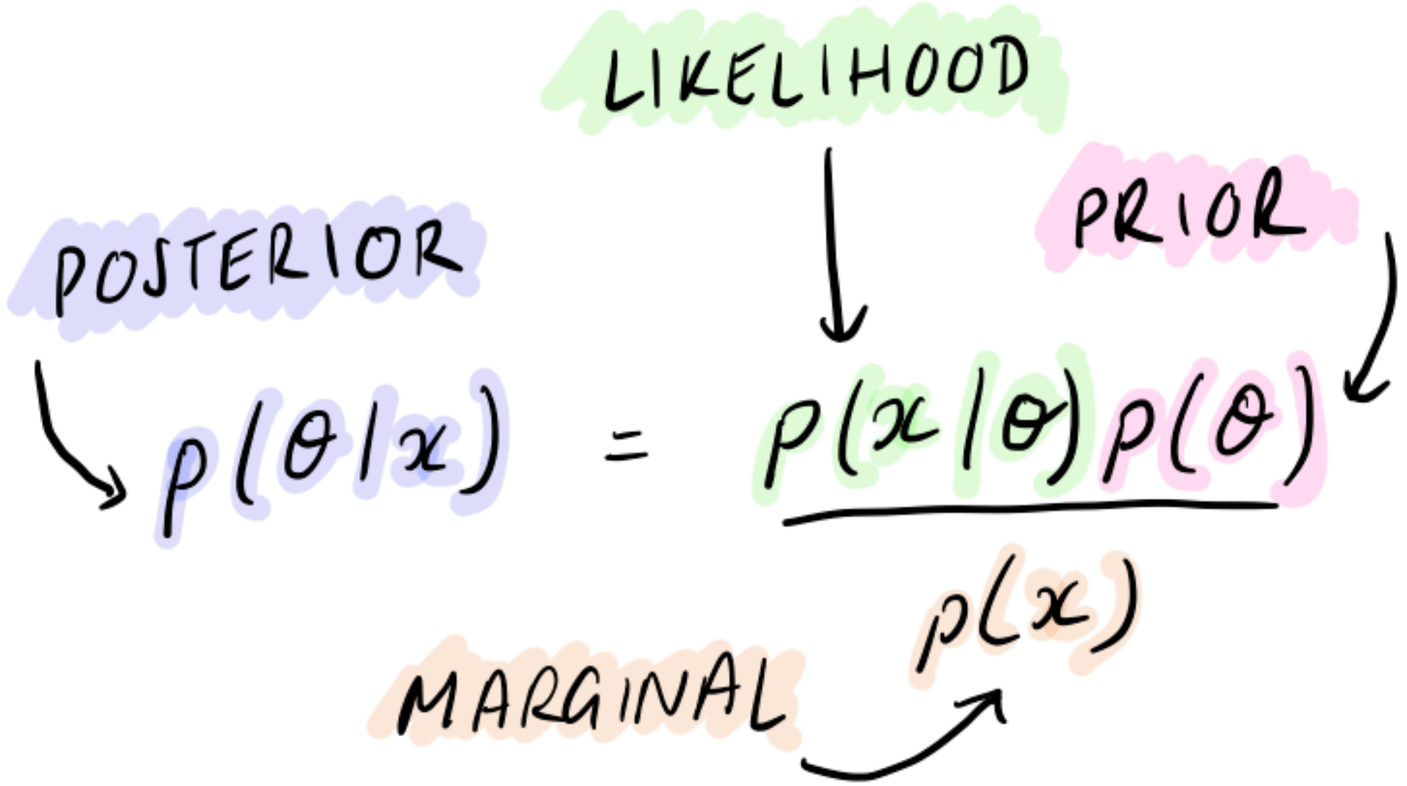
$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial H}{\partial \theta}$$



[Betancourt (2014)]

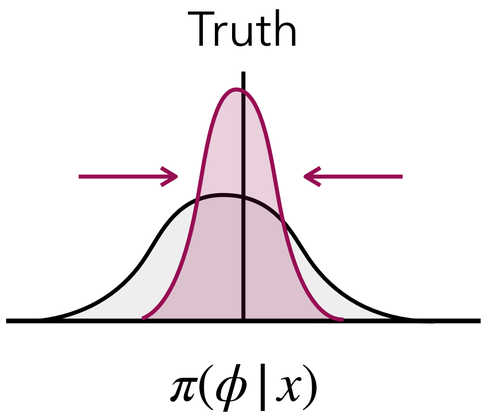
Uncertainty quantification for free

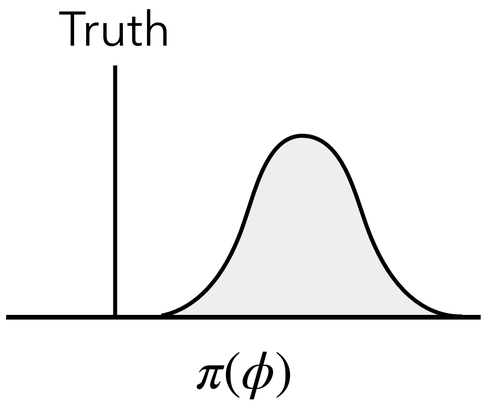
# Bayesian Inference



# Systematic uncertainties

[Sinervo PHYSTAT 2003; Heinrich+Lyons Annu. Rev. Nucl. Part. Sci. 2007]




1 Systematics that can be constrained

2 Uncertain assumptions in measurement/analysis

3 Uncertain theoretical framework

Calibration

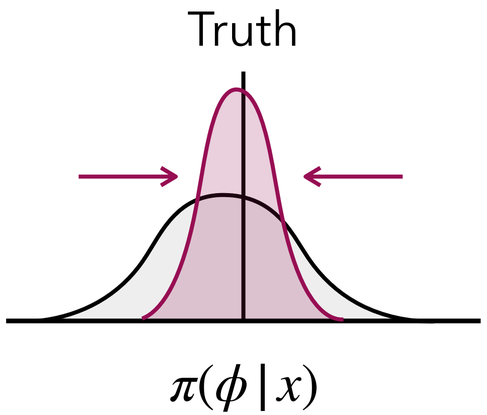
Data

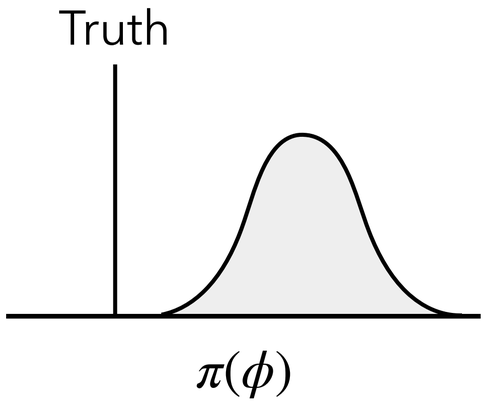
Model misspecification

Theory/ phenomenology

# Systematic uncertainties

[Sinervo PHYSTAT 2003; Heinrich+Lyons Annu. Rev. Nucl. Part. Sci. 2007]



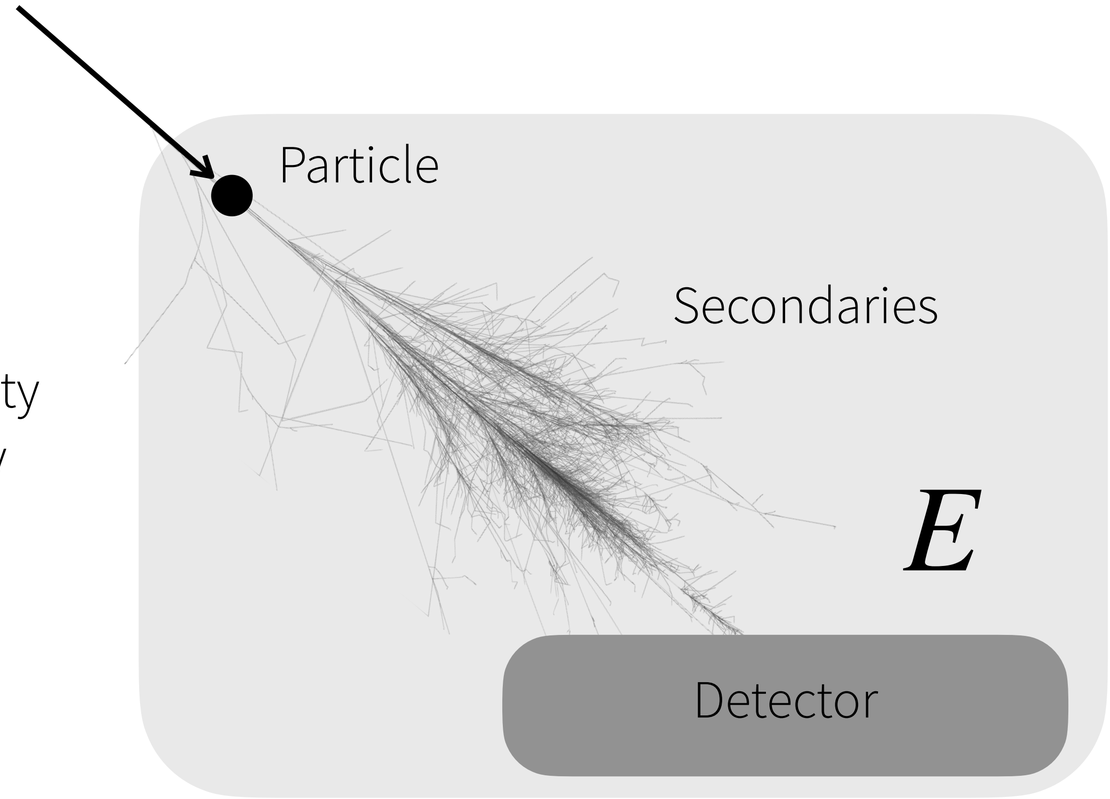

1 Systematics that can be constrained — GOOD 😀

2 Uncertain assumptions in measurement/analysis — BAD 😕

3 Uncertain theoretical framework — UGLY 😱

# Today’s session

Example: A toy particle detector that measures energy indirectly




Interaction probability depends on energy

Goal:

Measure particle spectrum

# Today’s session

- Forward folding
- Goodness-of-fit
- Systematic uncertainties