Parameter estimation — Bayesian analysis primer

Yi (Luna) Chen, Vanderbilt University

Jul 22, 2026, Lehigh ML/AI/Jet School

The Vanderbilt HENP Group's work is supported by US DOE-NP

Bayesian analysis

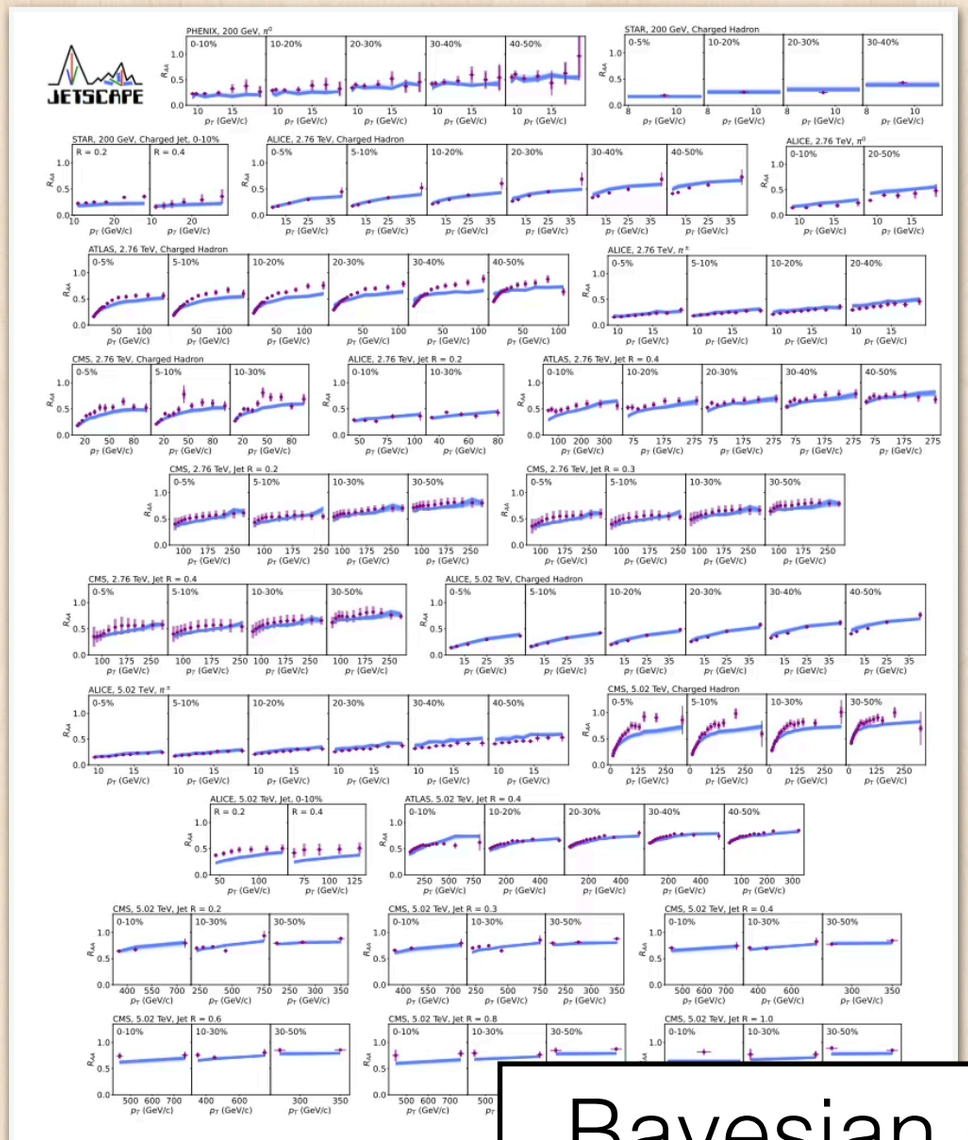
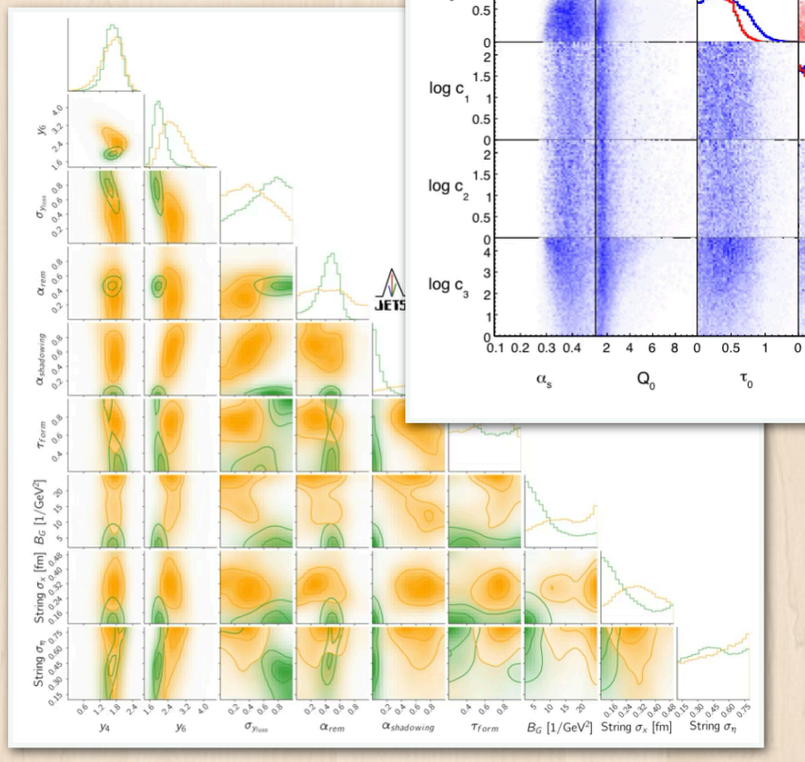
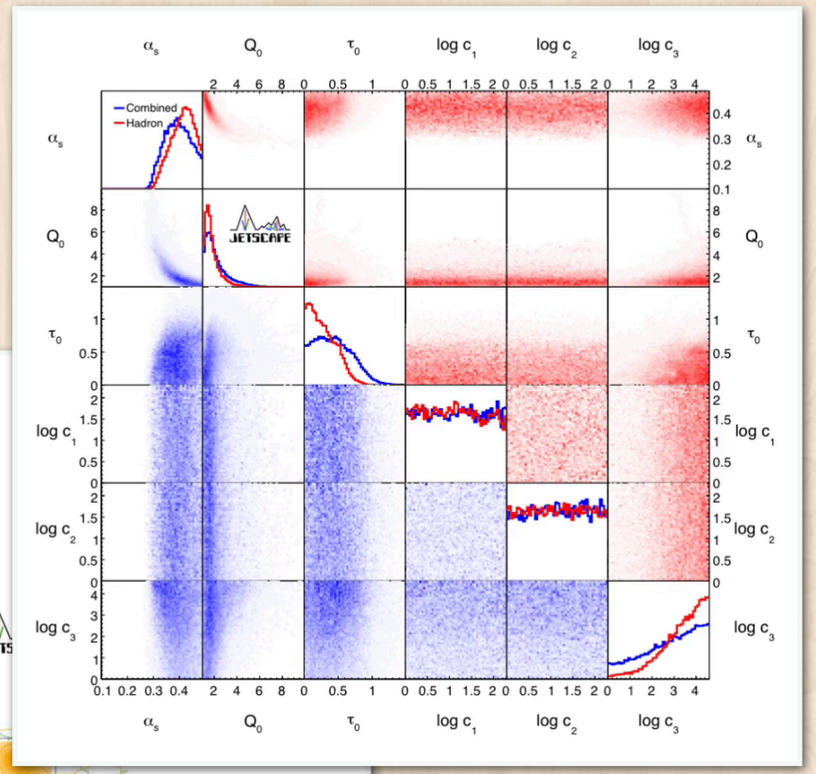


FIG. 11. All measurements of hadron and jet R_{AA} at $\sqrt{s_{NN}} = 0.$
in this analysis, compared to posterior predictive distributions





Bayesian analysis: cohesive parameter estimation from many many data at once

Overview

- We'll develop and make sense of the increasingly prevalent **Bayesian analysis**

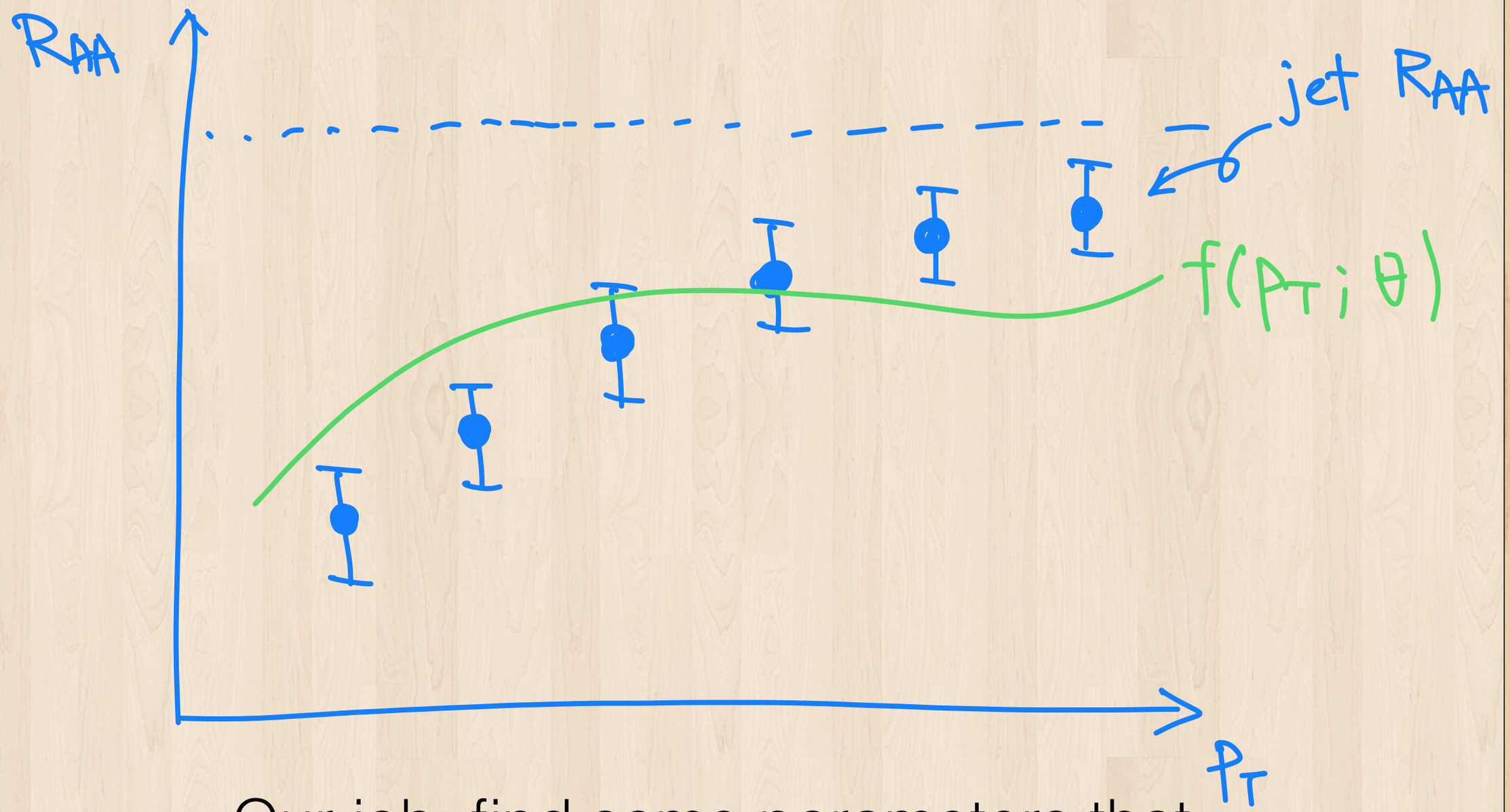
- It looks complicated but at its core it's very simple: we **compare model to data** to find what fits the best

  - Then reality strikes and we replace certain pieces with more sophisticated counterparts

- Today we will build a basic version together without all the fluff — so you know how things are connected

Part 1: Fit

Fits to functions

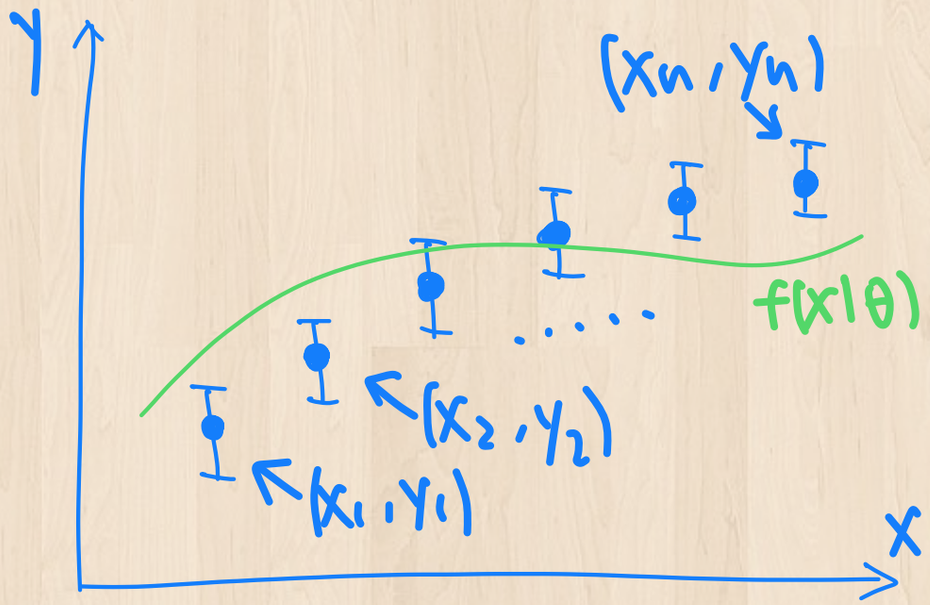


Our job: find some parameters that make $f(p_T;\theta)$ match **data** the best

Least-square fitting

- We have a set of data points $\{x_i, y_i\}$
- We have a function we want to fit to $f(x \mid \theta)$ where the function tells you what the y should be at position x
- We want to minimize distance between the two things
- Least square fitting: find θ that minimizes this function

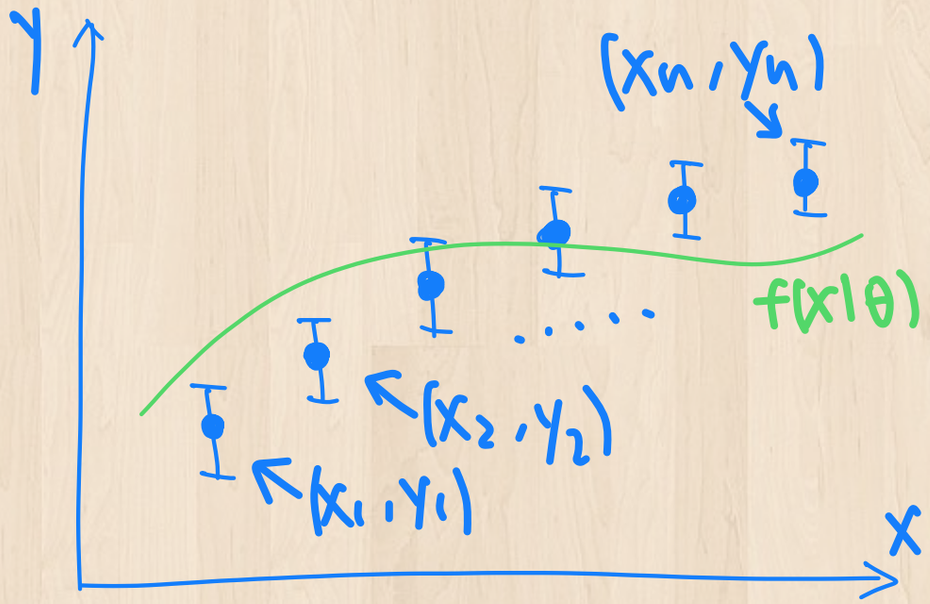
$$\sum_i \left(y_i - f(x_i \mid \theta)\right)^2$$



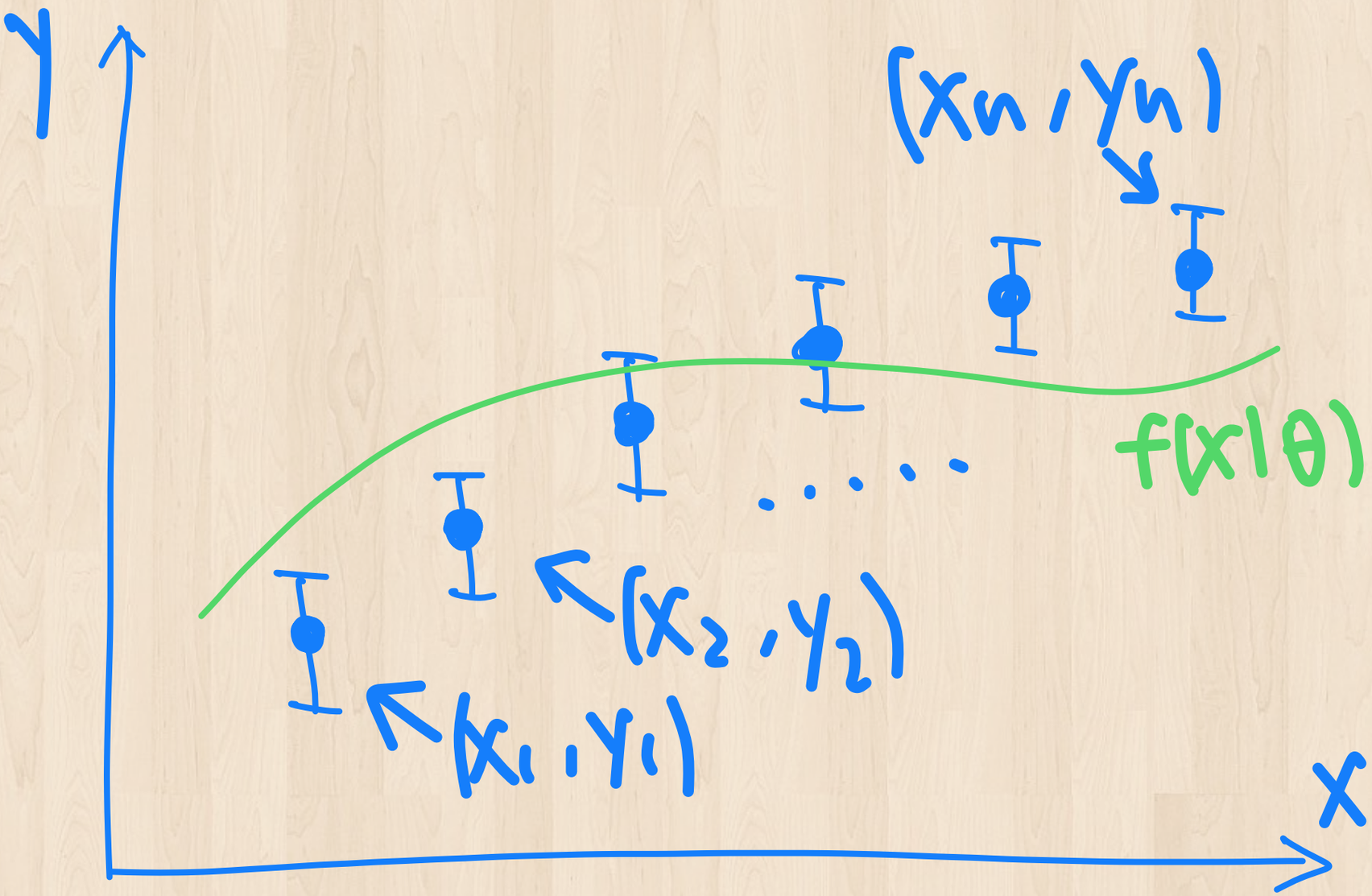
Least-square fitting

- We have a set of data points $\{x_i, y_i \pm \sigma_i\}$
- We have a function we want to fit to $f(x \mid \theta)$ where the function tells you what the y should be at position x
- We want to minimize distance between the two things
- Least square fitting: find θ that minimizes this function

$$\chi^2 = \sum_i \left(\frac{y_i - f(x_i \mid \theta)}{\sigma_i} \right)^2$$



What does it mean?



What's the game plan?

1. Identify the data $\{x_i, y_i\}$ we want to use
2. Find a function $f(x_i|\theta)$ we want to use
3. Define what we mean by "how well things match"
4. Vary θ to minimize the function in item 3.
5. Physics!

What's the game plan?

ask experimentalist or measure yourself

1. Identify the data $\{x_i, y_i\}$ we want to use

check with intuition & calculation

2. Find a function $f(x_i|\theta)$ we want to use

3. Define what we mean by "how well things match"

e.g. χ^2

4. Vary θ to minimize the function in item 3.

some math & algorithm

e.g. RooFit package, $\frac{\partial}{\partial x} = 0$

5. Physics!

What's the game plan?

1. Identify the data $\{x_i, y_i\}$ we want to use
2. Find a function $f(x_i|\theta)$ we want to use
3. Define what we mean by "how well things match"
4. Vary θ to minimize the function in item 3.
5. Physics!

It's the same game plan in Bayesian analysis. Just different choices

Part 2: Bayesian statistics

What is likelihood

Likelihood function = how "likely" a set of parameter is, relative to other parameter values, given observed data

$$\mathscr{L}(\theta \mid x)$$

Parameters of interest

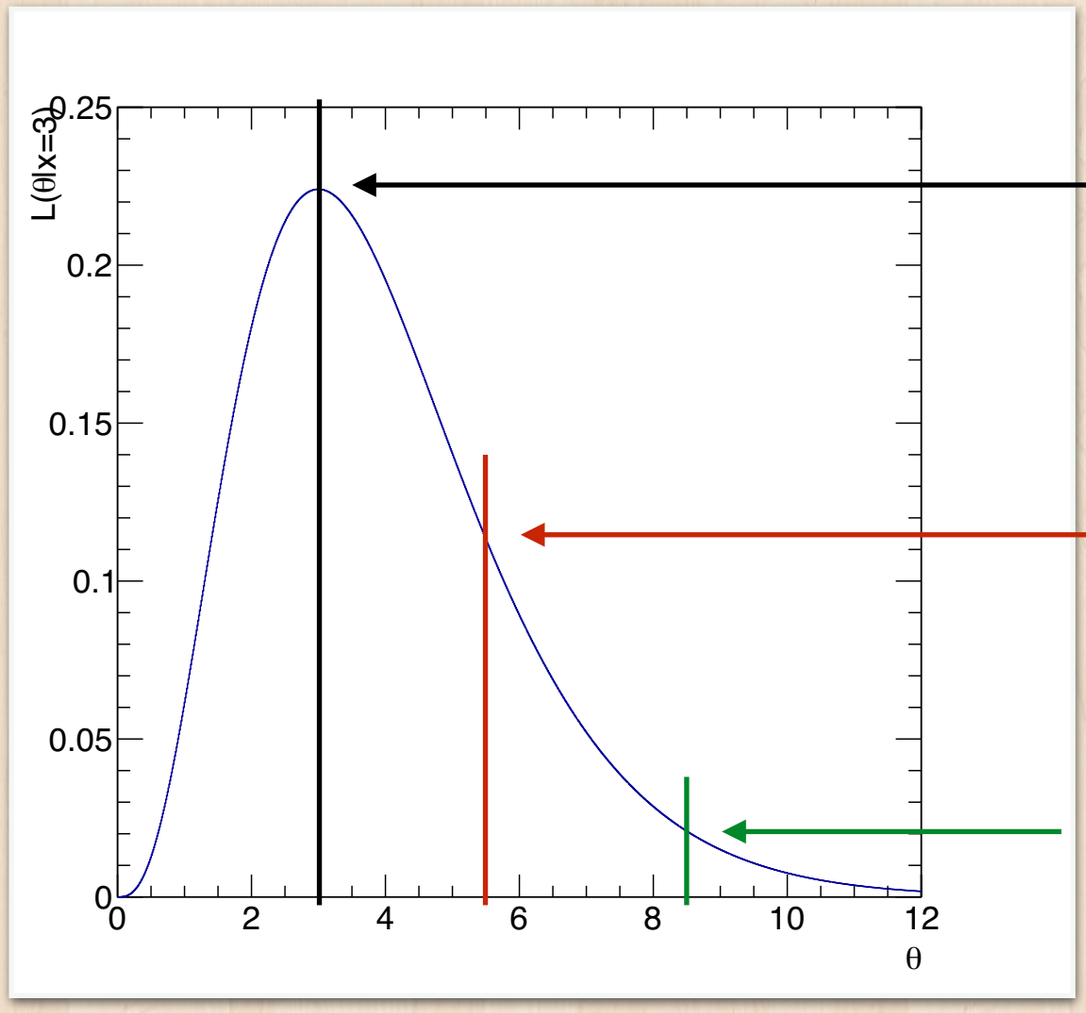
Observed data

It's a function of the parameters of interest, conditioning on what we see in data

Example: counting experiment

Suppose we count number of events and data says 3

$$\mathscr{L}(\theta \,|\, x = 3)$$



Most likely true count is 3

5.5 is about “half as likely”

8.5 is about “10% as likely”

Example: counting experiment

If the expected count is θ, we can write the probability of observing a count of x as

$$P(x|\theta) = \text{Poisson}(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}$$

Function of x, given θ

We write the likelihood function as

$$\mathcal{L}(\theta|x) \equiv \frac{\theta^x e^{-\theta}}{x!}$$

Function of θ, given x

More examples

$$\mathscr{L}(\theta \mid x) \text{ vs } P(x \mid \theta)$$

θ = theory parameter	x = observed data
Probability of head per flip	Heads/tail sequence of coin throw
Higgs boson cross section	Number of events around 125 GeV
Neutrino interaction cross section	Counts in water tank
Energy loss parameter for Partons	Hadron R_{AA}
……	……

Bayesian school of thinking

How to connect the two things?

$$\mathscr{L}(\theta \mid x) \text{ vs } P(x \mid \theta)$$

One is function of θ, one is function of x

In the Bayesian school of thinking we have the posterior probability function *

$$P(\theta \mid x)$$

* Likelihood and Bayesian posterior are not exactly the same thing; the distinction is beyond the scope of this lecture

Bayesian school of thinking

Starting from the probability equality

$$P(a,b) = P(a\,|\,b)P(b)$$

We can “write”

$$P(x,\theta) = P(x\,|\,\theta)P(\theta) = P(\theta\,|\,x)P(x)$$

or

$$P(\vec{\theta}\,|\,\vec{x}) = \frac{P(\vec{x}\,|\,\vec{\theta})P(\vec{\theta})}{P(\vec{x})}$$

Bayes’ theorem

Bayesian school of thinking

Prior knowledge of how likely given parameter is true

Posterior probability $\rightarrow P(\theta \mid x) = \dfrac{P(x \mid \theta)P(\theta)}{P(x)}$

Probability of observing "x", generally speaking

Since experiment is done, $P(x)$ is a constant

$$P(\theta \mid x) \propto P(x \mid \theta)P(\theta) \Big|_{\text{given fixed x}}$$

Prior knowledge

- The Bayesian formalism always involves a “prior” $P(\theta)$, which encodes our prior knowledge on how θ distributes

- It’s both a blessing and a curse — our outcome will always be “biased” by what we know before

- Setting $P(\theta) = 1$ gives us back to the simplest case

  - However $P(\theta) = 1$ does not mean unbiased (why not $P(\theta^2) = 1$? $P(\ln \theta) = 1$?)

Small Bayesian exercise

If it rains, the ground is wet 100% of the time

If it does not rain, the ground is wet 10% of the time

Forecast says 65% chance of rain right now

Given that we see the ground is wet, what is the probability that it is actually raining?

$$P(\theta|x) \propto P(x|\theta)P(\theta)$$

Small Bayesian exercise

$$P(\text{wet} \mid \text{rain}) = 100\%$$
$$P(\text{wet} \mid \text{no rain}) = 10\%$$

$$P(\text{rain}) = 65\%$$

$$P(\text{rain} \mid \text{wet}) \propto P(\text{wet} \mid \text{rain})\, P(\text{rain}) = 0.65$$
$$P(\text{no rain} \mid \text{wet}) \propto P(\text{wet} \mid \text{no rain})\, [1 - P(\text{rain})] = 0.035$$

$$P(\text{rain} \mid \text{wet}) = 0.65 / (0.65 + 0.035) = 94.9\%$$

Small Bayesian exercise

P(wet | rain) = 100%
P(wet | no rain) = 10%

What if forecast says 5%? → P(rain) = 5%

Small Bayesian exercise

$$P(\text{wet} \mid \text{rain}) = 100\%$$
$$P(\text{wet} \mid \text{no rain}) = 10\%$$

What if forecast says 5%? → $P(\text{rain}) = 5\%$

$$P(\text{rain} \mid \text{wet}) \propto P(\text{wet} \mid \text{rain})\, P(\text{rain}) = 0.05$$
$$P(\text{no rain} \mid \text{wet}) \propto P(\text{wet} \mid \text{no rain})\, [1 - P(\text{rain})] = 0.095$$

$$P(\text{rain} \mid \text{wet}) = 0.05 / (0.05 + 0.095) = 34.5\%$$

Small Bayesian exercise

$$P(\text{wet} \mid \text{rain}) = 100\%$$
$$P(\text{wet} \mid \text{no rain}) = 10\%$$

What if forecast says 5%? → $P(\text{rain}) = 5\%$

$$P(\text{rain} \mid \text{wet}) \propto P(\text{wet} \mid \text{rain})\, P(\text{rain}) = 0.05$$
$$P(\text{no rain} \mid \text{wet}) \propto P(\text{wet} \mid \text{no rain})\, [1 - P(\text{rain})] = 0.095$$

Our view of what is happening is affected by prior knowledge. There is no "unbiased" P(rain)

Updating knowledge

Another way to think about the Bayes' formalism is to refine our knowledge about the problem

$$P(\vec{\theta}|\vec{x}) \propto P(\vec{x}|\vec{\theta})P(\vec{\theta})$$

What we know afterwards

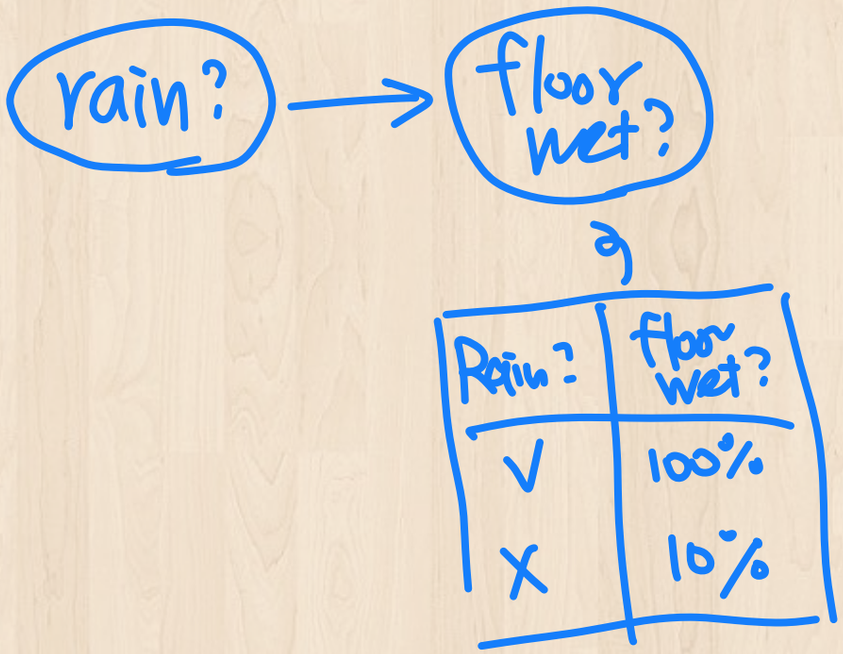
What we know before

$$P(\vec{\theta}|\vec{x}, \vec{x}_{\text{past}}) \propto P(\vec{x}|\vec{\theta}, \vec{x}_{\text{past}})P(\vec{\theta}|\vec{x}_{\text{past}})$$

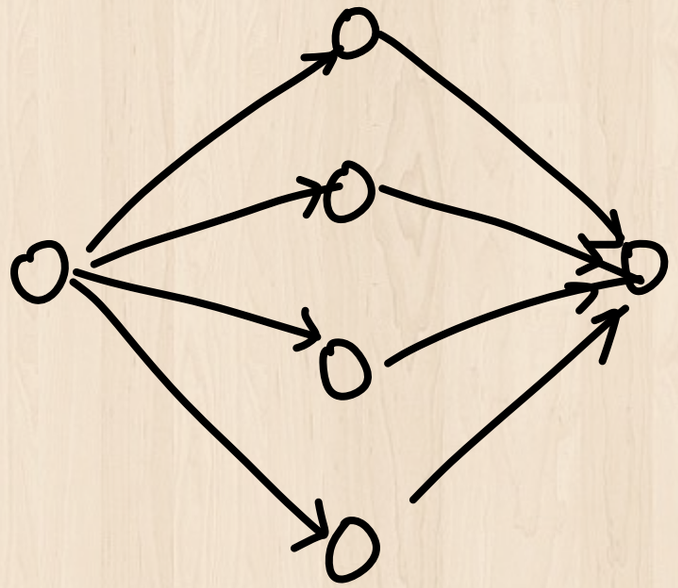
Everything is conditioning on our past knowledge

Beyond “single-layer”

Inferences can be combined and “learned” together
→ Bayesian network



Internal parameters



Prediction →

Inference ←

Training → ←

Bayesian statistics: recap

- The likelihood function $L(\theta|x)$ gives us the relative degree of likelihood as a function of θ, given observed data x

- We can write the posterior $P(\theta|x)$ in terms of the probability distribution $P(x|\theta)$, which is the probability of observing data x, given θ

  - With a prior term $P(\theta)$ — there is no “universally unbiased” choice

Part 3: Building up to (contemporary) Bayesian framework

Back to the game plan

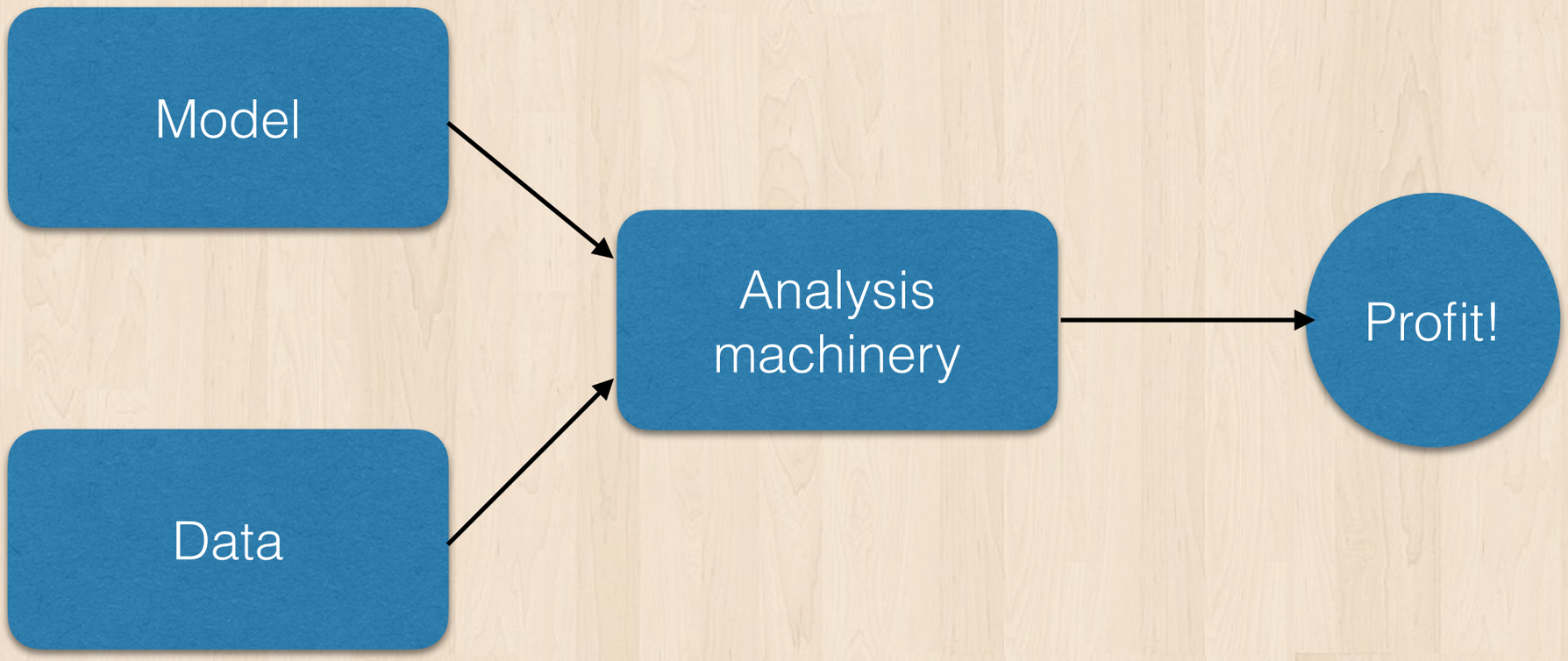
1. Identify the data $\{x_i, y_i\}$ we want to use
2. Find a function $f(x_i|\theta)$ we want to use
3. Define what we mean by “how well things match”
4. Vary θ to minimize the function in item 3.
5. Physics!

Game plan: Bayesian edition

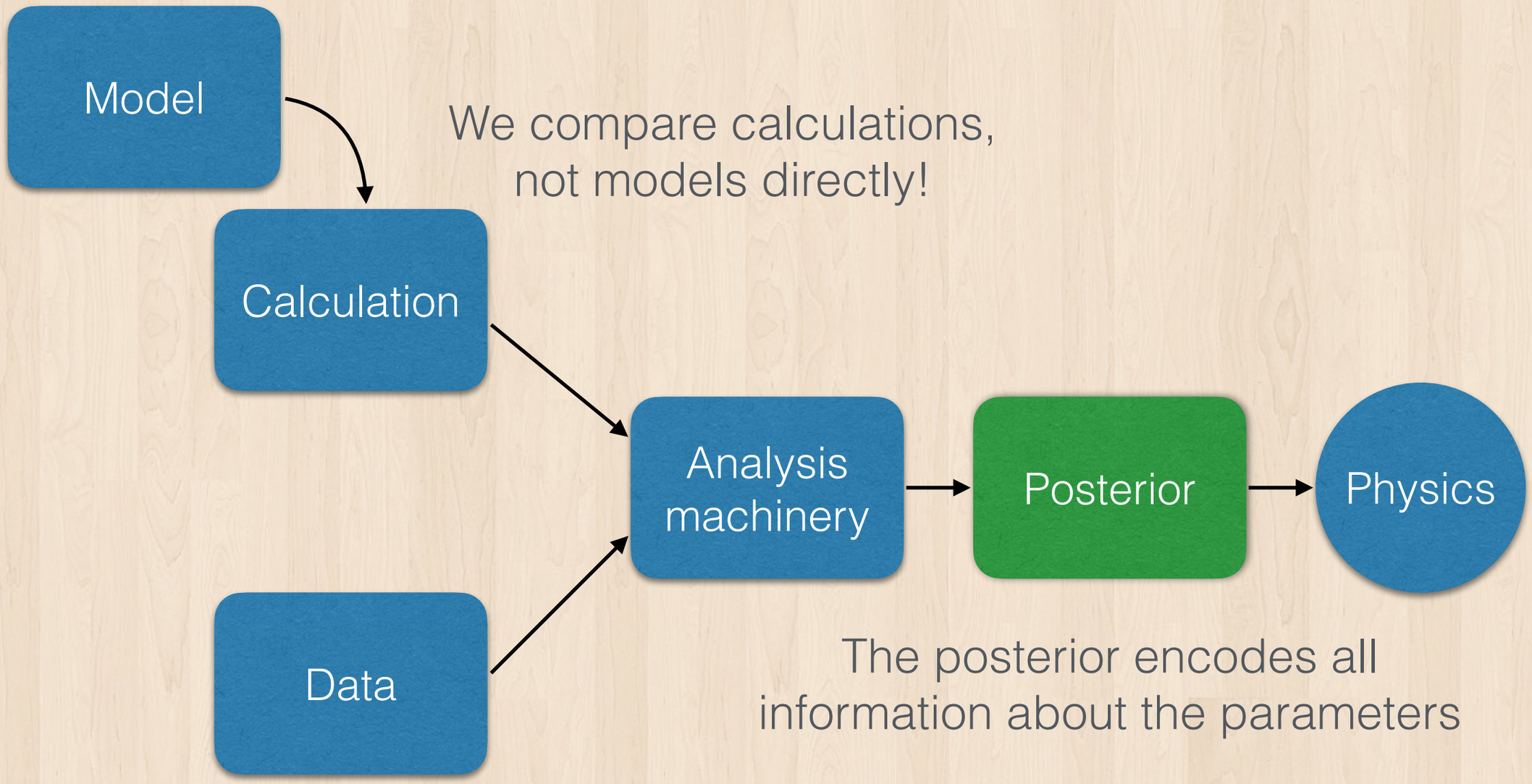
1. Identify the data $\{x_i, y_i\}$ we want to use
2. Find a function $f(x_i|\theta)$ we want to use
3. Define what we mean by "how well things match"
4. Vary θ to minimize the function in item 3.
5. Physics!

Bayesian Posterior function
$P(\theta|x)$

The 2000 ft. view



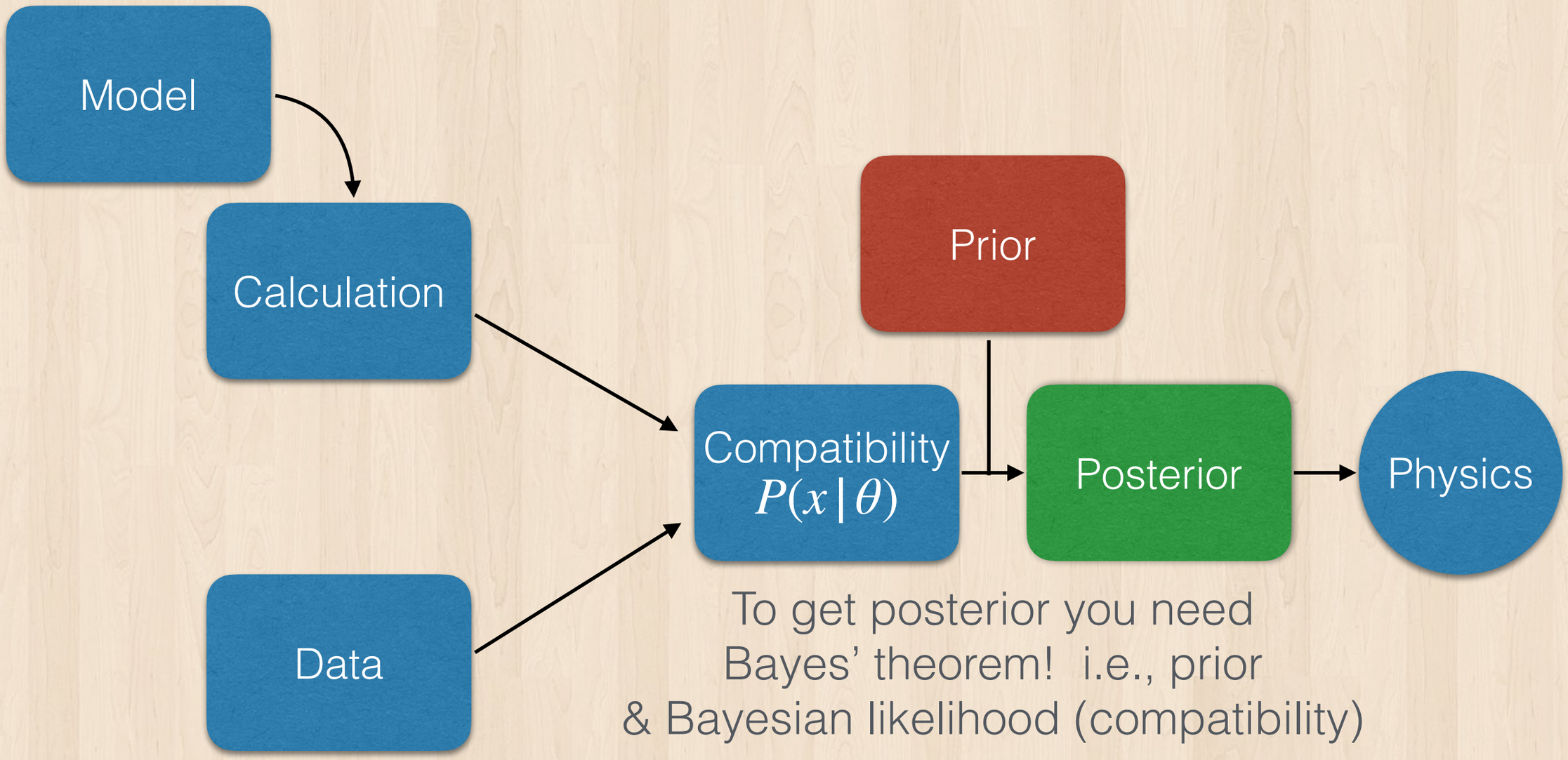
The 200 ft. view



We compare calculations, not models directly!

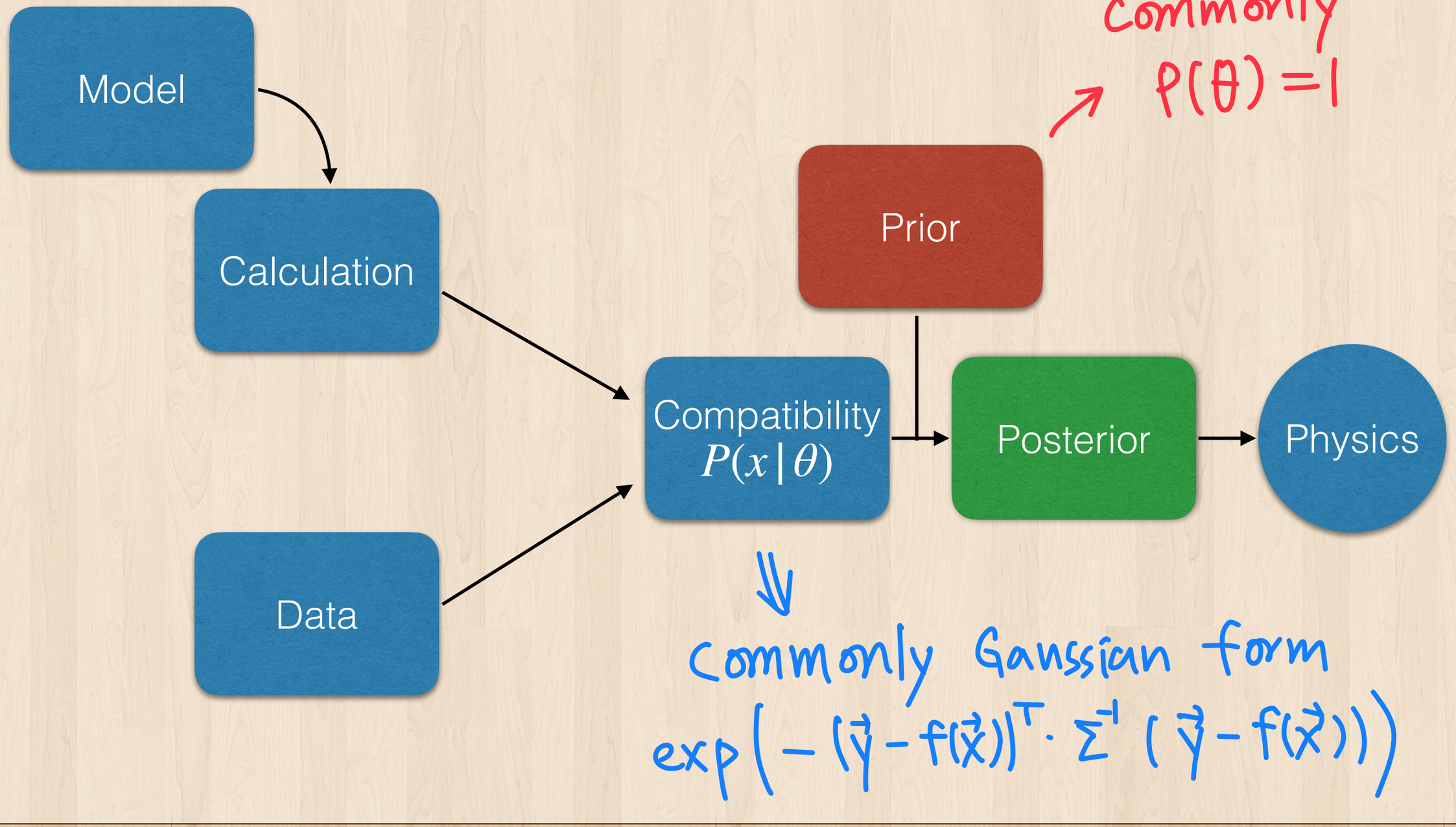
The posterior encodes all information about the parameters

The 50 ft. view



To get posterior you need
Bayes' theorem! i.e., prior
& Bayesian likelihood (compatibility)

The 50 ft. view



The core of the analysis

- Get the **posterior function**
  - Choose a **prior**
  - Get the **Bayesian likelihood** as a (continuous) function of theory parameters
  - Interpolate theory calculations (e.g., GP)
  - Prepping data as input
  - Make a choice of the compatibility function

Or the likelihood if non-Bayesian

Bayes' theorem:

$$P(\theta|x) \propto P(x|\theta)P(\theta)$$

Some discussion

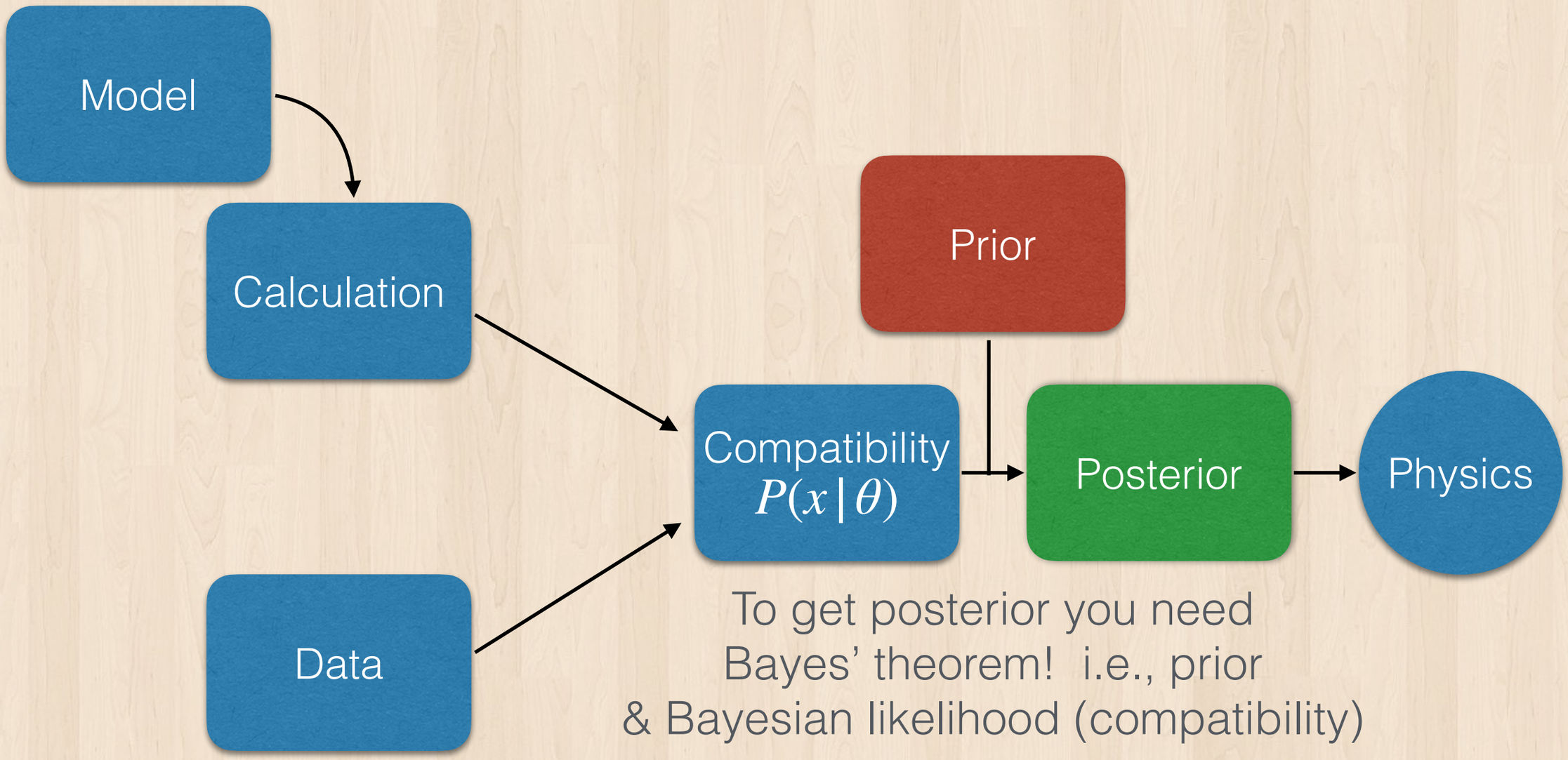
- Up to this point, we described all the **core** of the contemporary **Bayesian analysis**
  - The first part of the workfest we will implement and understand this base version of the Bayesian analysis
- Everything from this point on is **improvements**: we use fancier methodology to replace some items to make them run faster/better/etc
  - Most are out of necessity, but nevertheless the game plan is the same, we just do things in fancier ways

Now let's do module 1

Part 3': Building up to (contemporary) Bayesian framework

Dealing with constraints

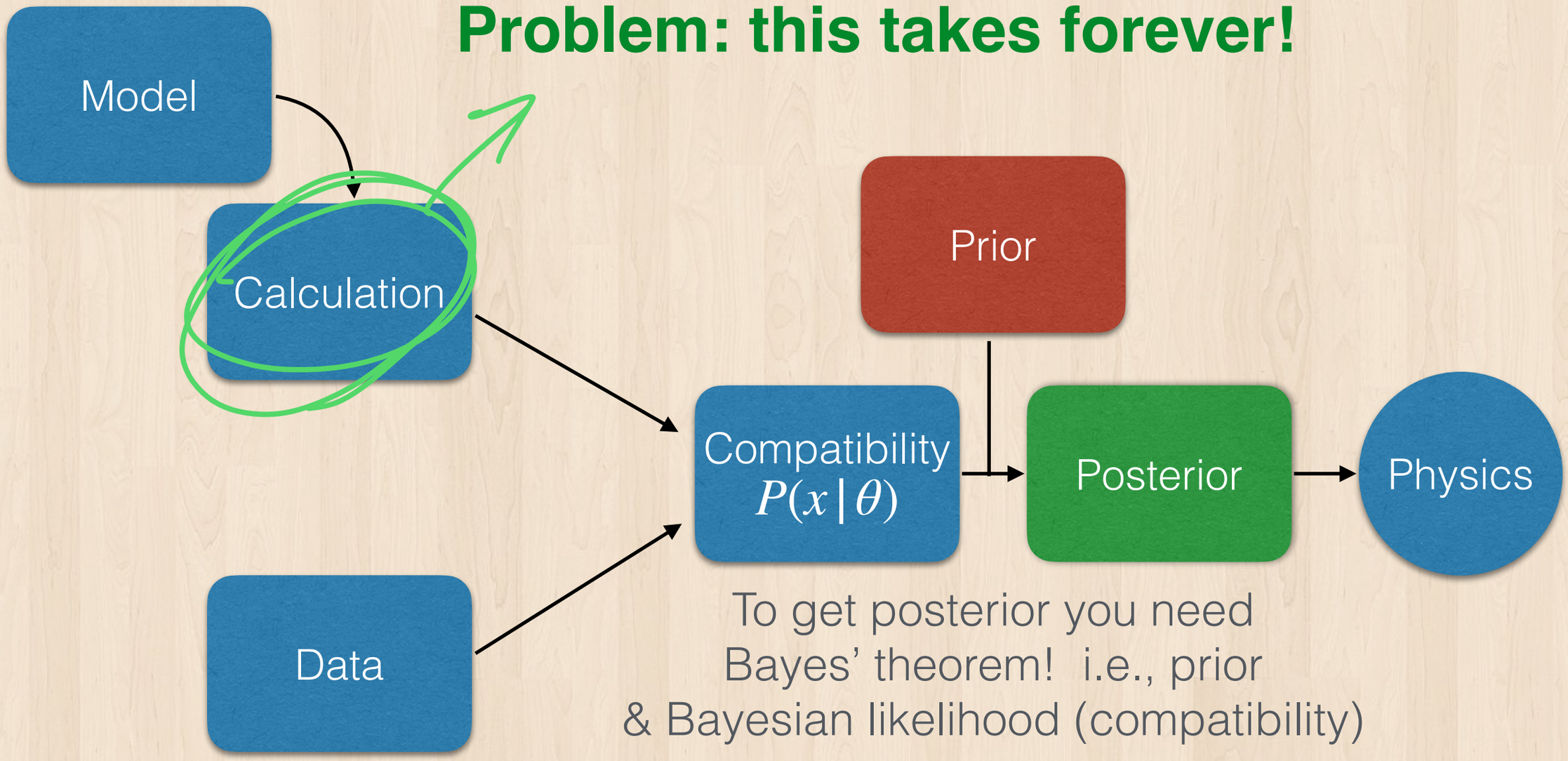
The 50 ft. view



To get posterior you need
Bayes' theorem! i.e., prior
& Bayesian likelihood (compatibility)

The 50 ft. view

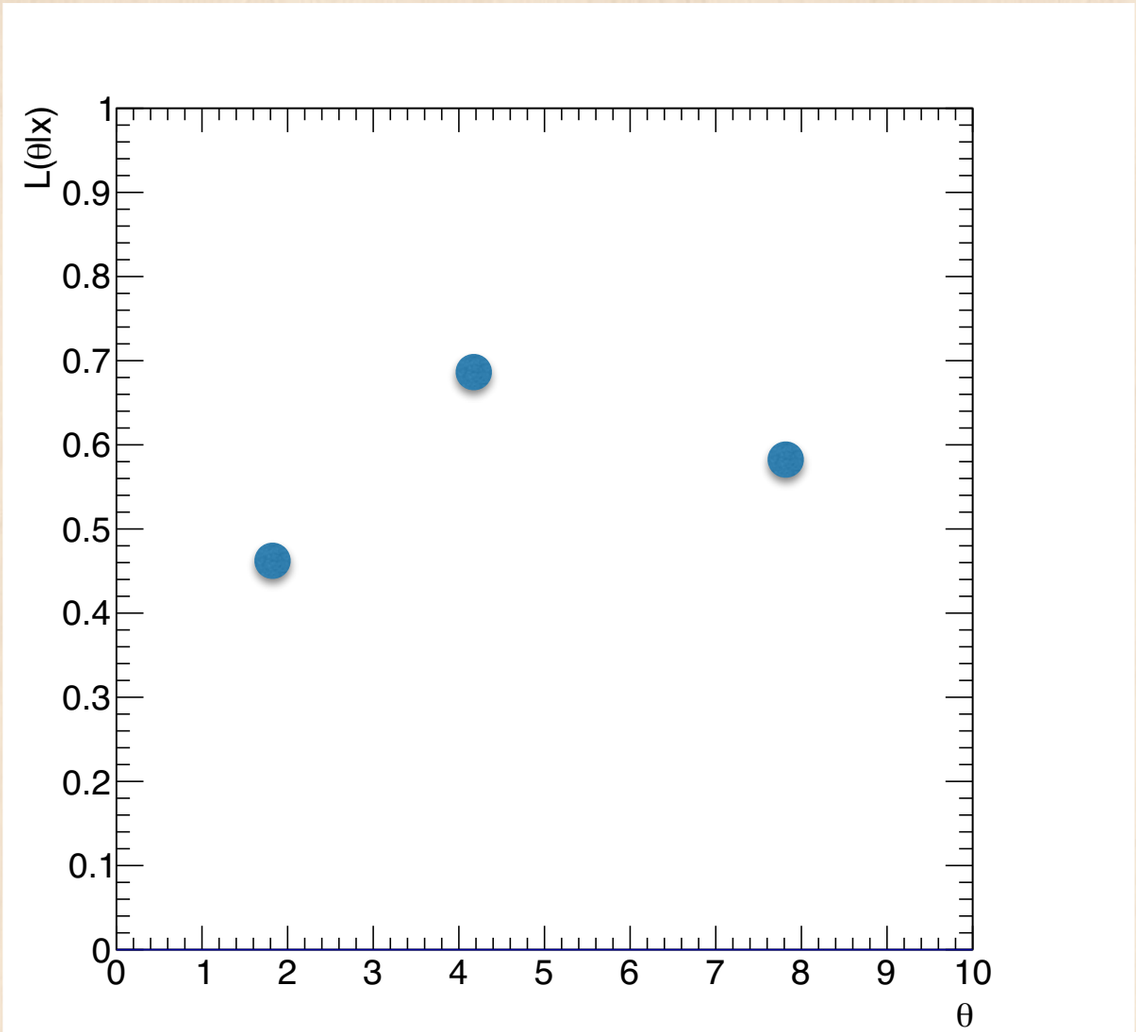
Problem: this takes forever!



To get posterior you need
Bayes' theorem! i.e., prior
& Bayesian likelihood (compatibility)

Computing-intensive case

What can we do, if likelihood on one point takes weeks—months to calculate on a computing cluster?

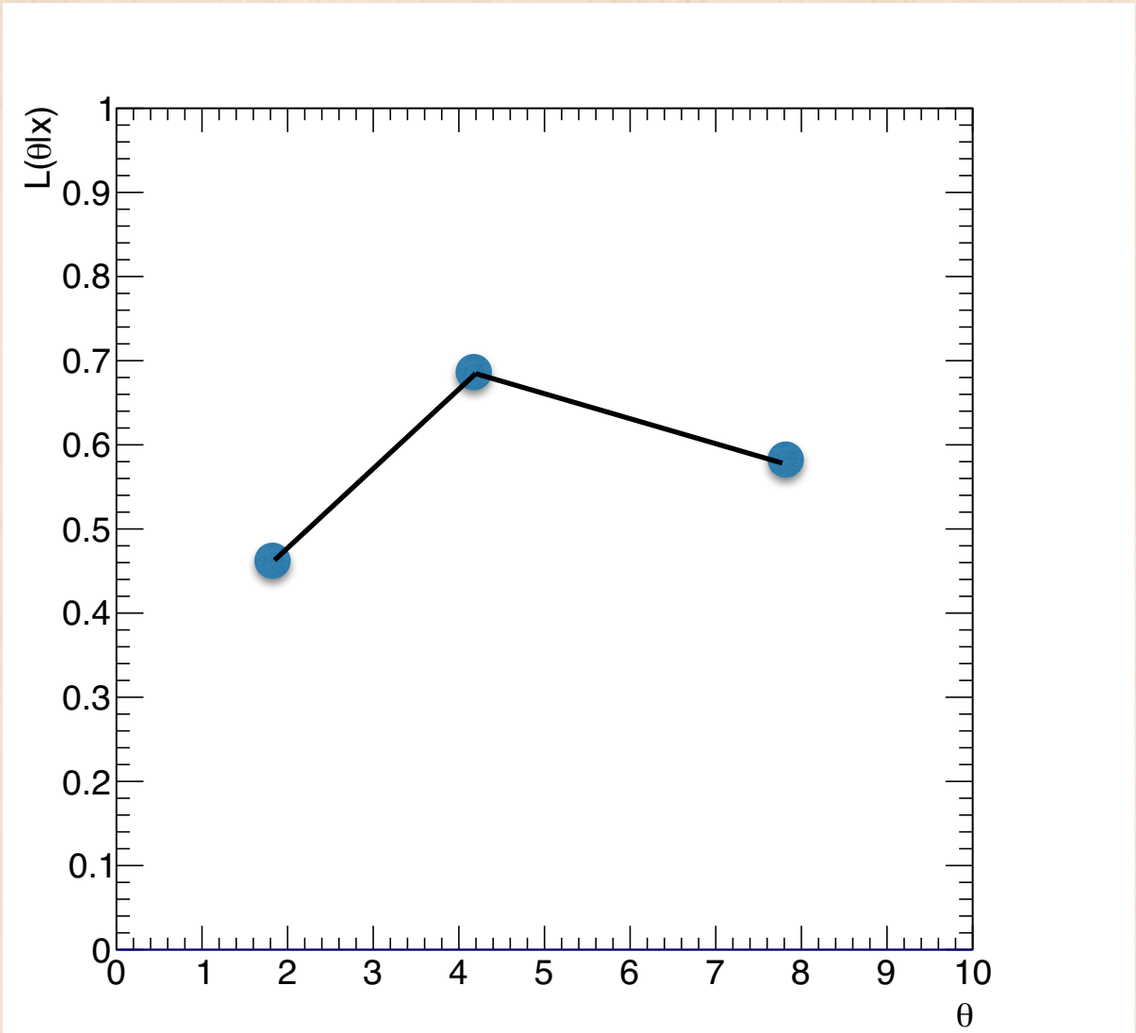


Each of these points take a month to get

Things won't work well with this latency

Computing-intensive case

We pick nicely spaced points (“design points”), evaluate the likelihood on those, and interpolate



If the points are picked well, the interpolated function should approach L(θ|x)

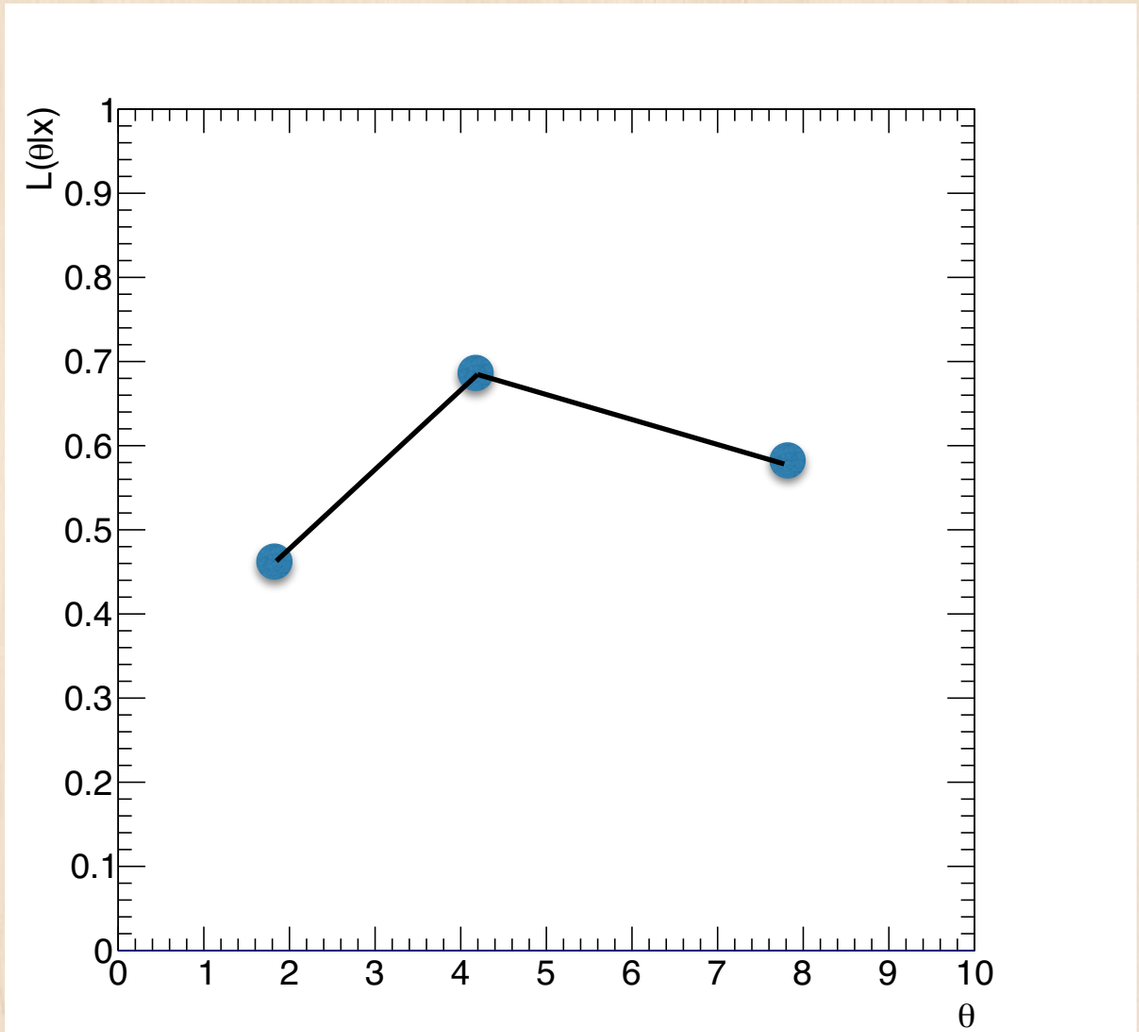
“Latin hypercube”: an algorithm to sample N dimensional space ~uniformly

How to interpolate?

<u>Straight line / spline:</u>
works ~well for 1D

Generalization to more dimensions not straightforward

But not ideal because of kinks etc



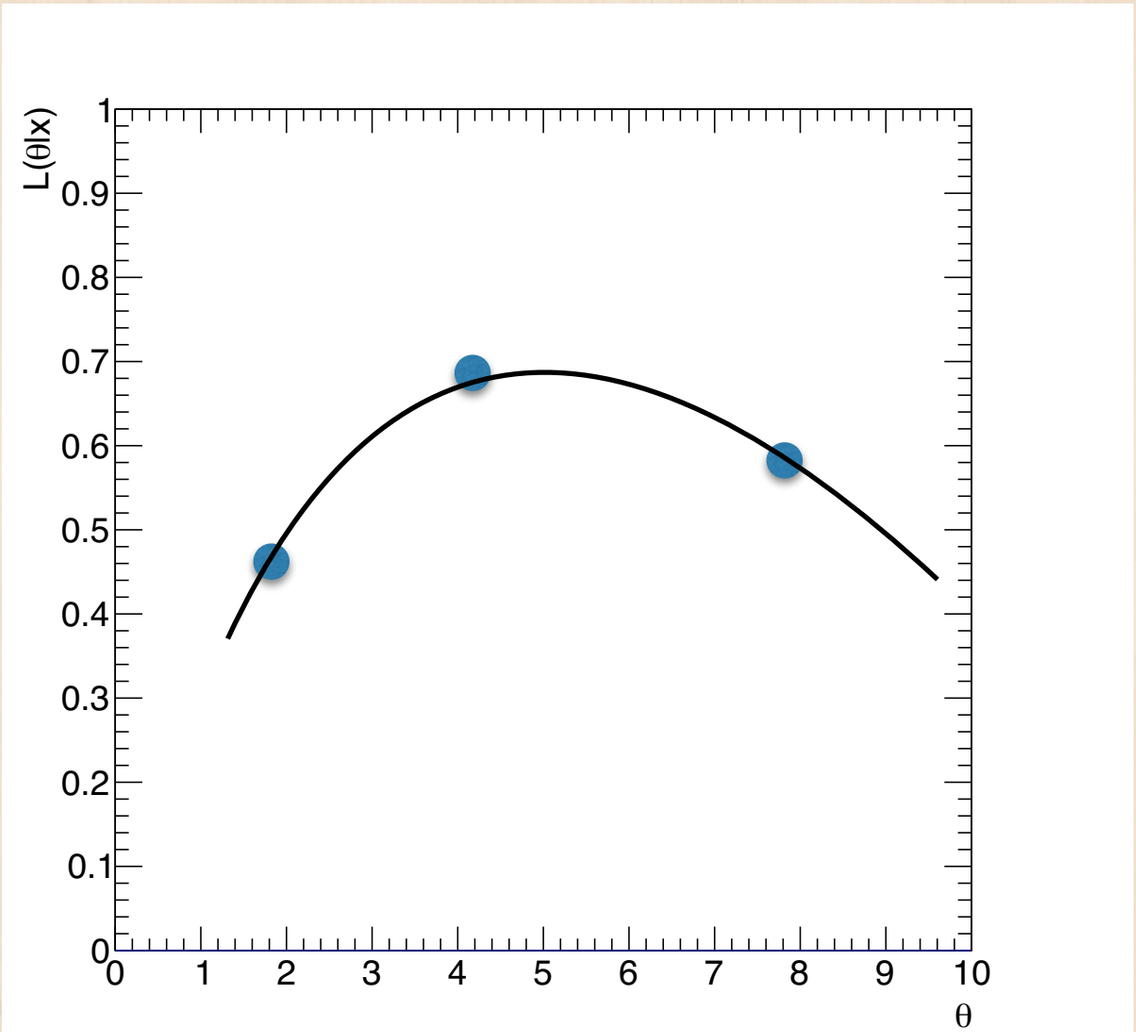
Now let's do module 2

How to interpolate?

Fit a function

Good choice if there is a well-motivated functional form

The choice of function is too important, and can bias the result if chosen poorly



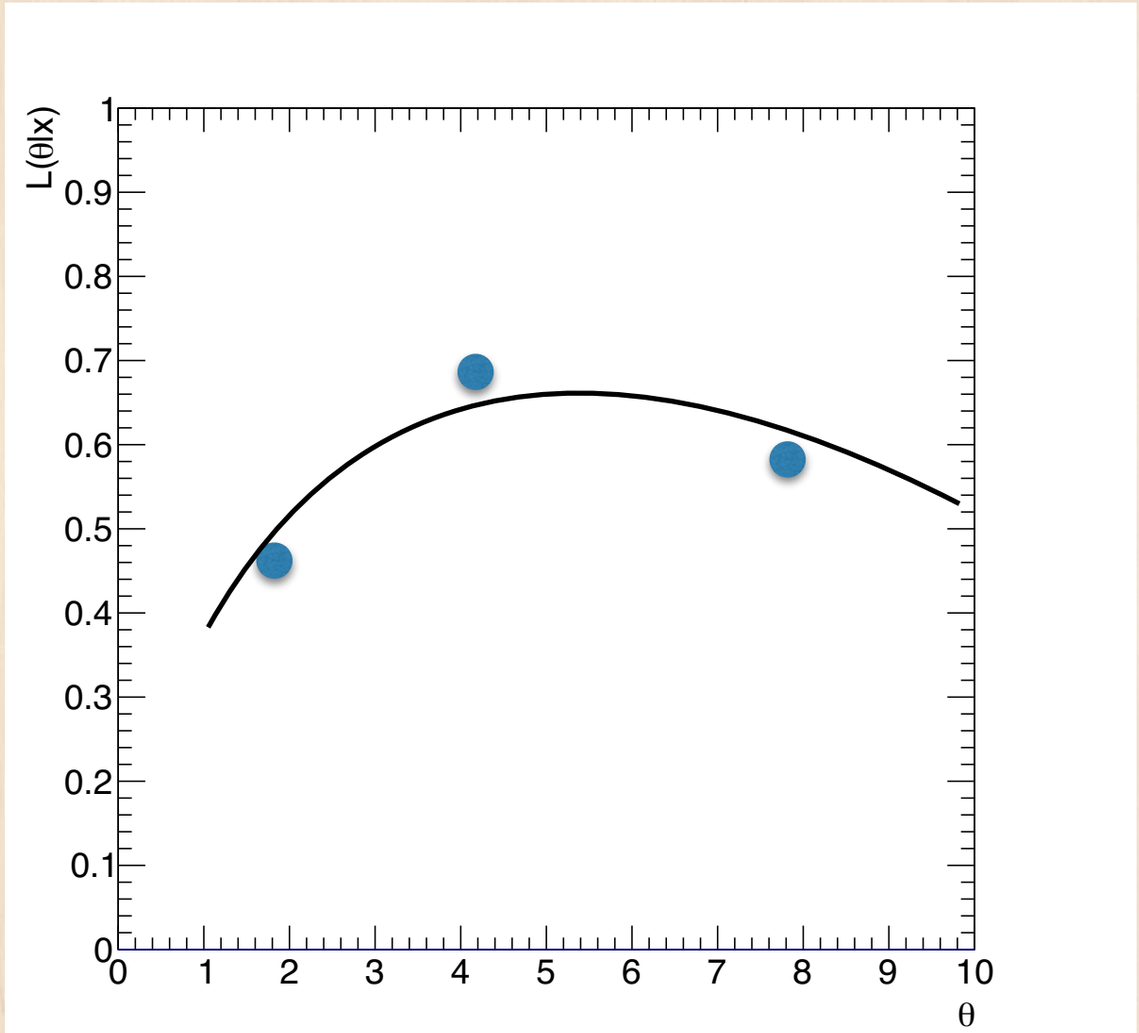
How to interpolate?

<u>Closest neighbor average</u>

Average of neighbors with weights depending on distance

Easily generalized to higher dimensions

Smooths the likelihood function — may not be ideal



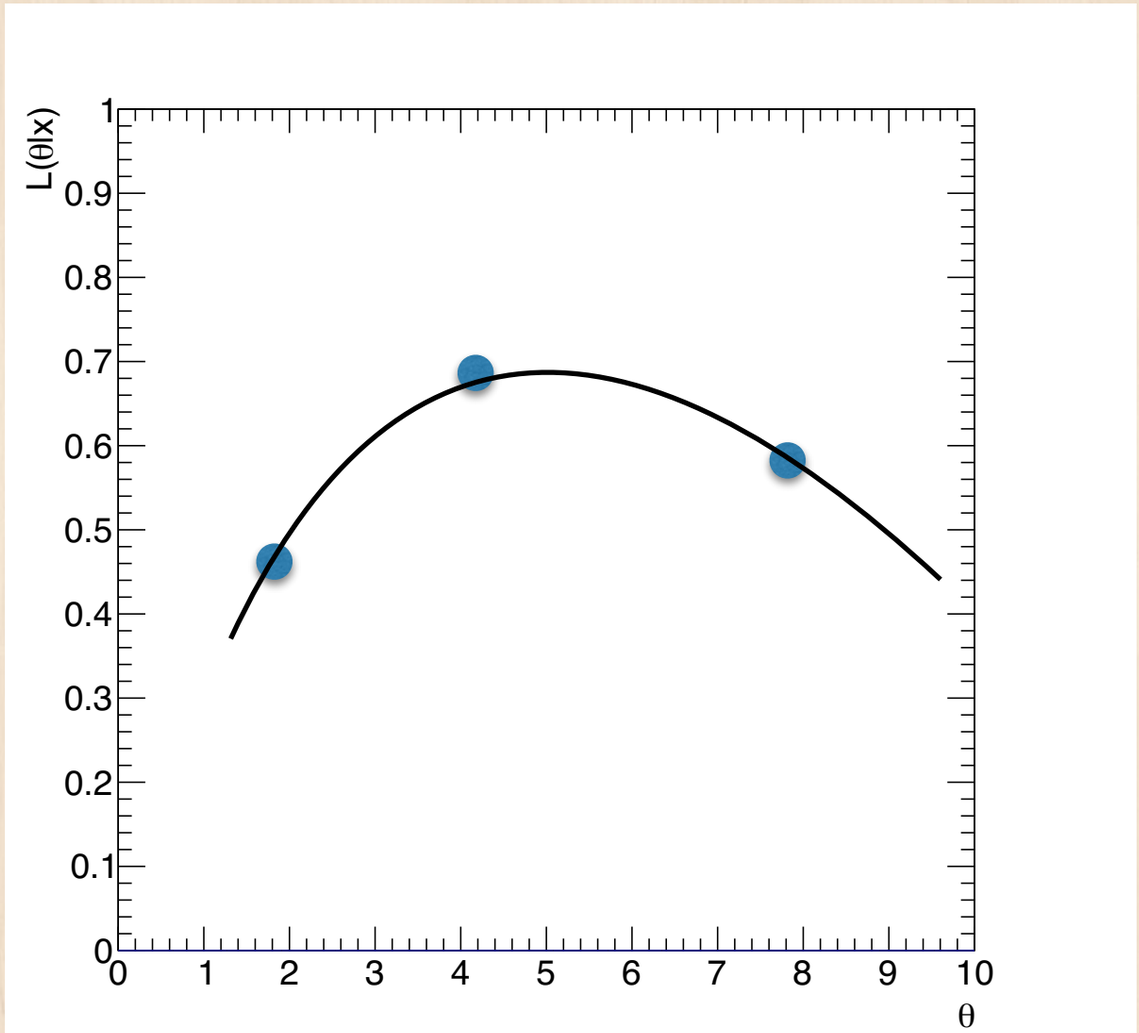
How to interpolate?

"Gaussian process emulator" (GPE)

Interpolates points without needing to assume a global functional form

Can easily be adapted to higher dimensions

Gives "interpolation error"

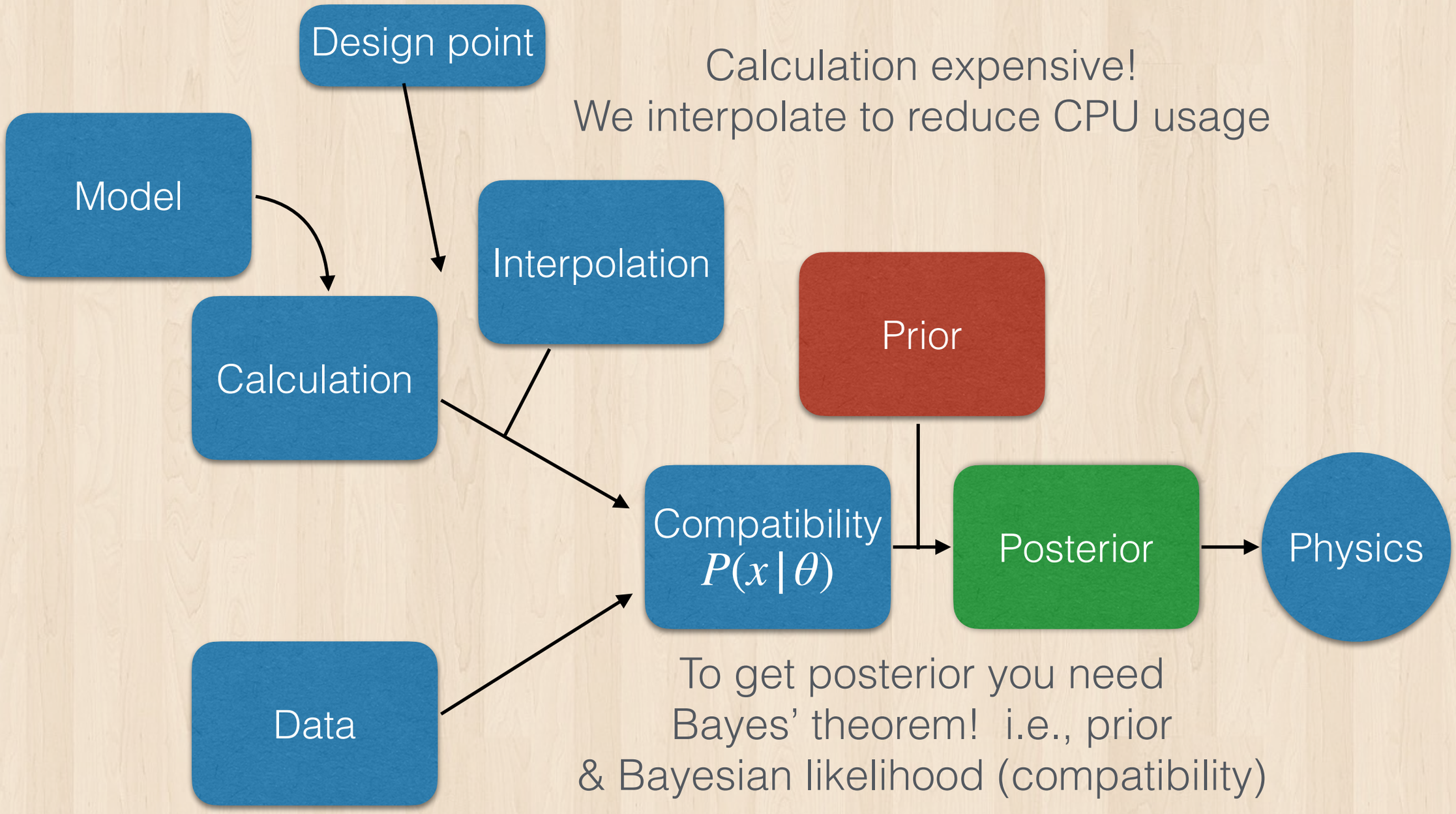


Now let's do module 3

Interpolation: recap

- There are many ways we can build the function
  - Analytical functions if we are extremely lucky
- When it becomes complicated, approximations have to be made
- For example in the case of computing-intensive calculations, we can pick points and interpolate
  - Gaussian process emulator (GPE) is one of the good ways to do this

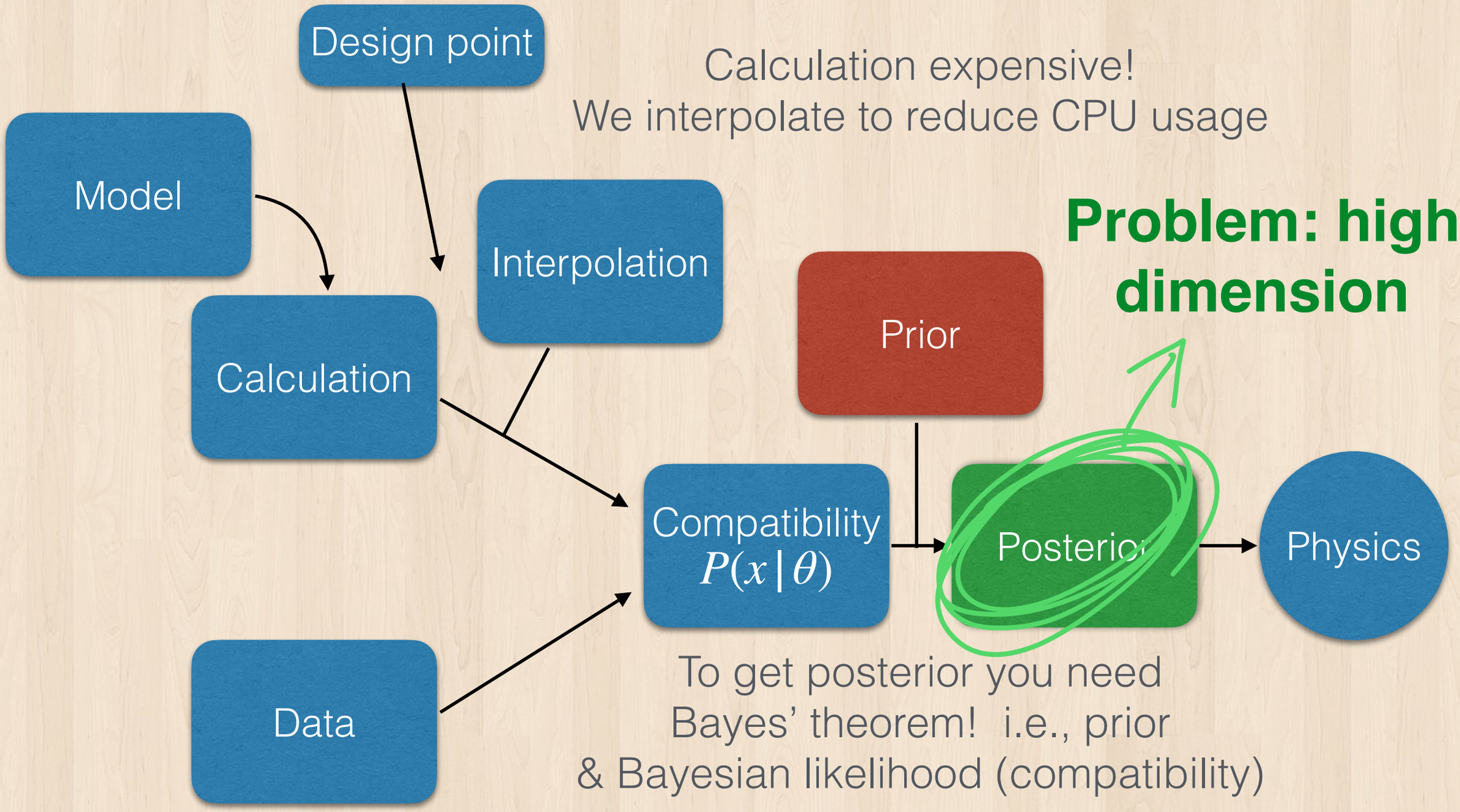
The 20 ft. view



Calculation expensive!
We interpolate to reduce CPU usage

To get posterior you need
Bayes' theorem! i.e., prior
& Bayesian likelihood (compatibility)

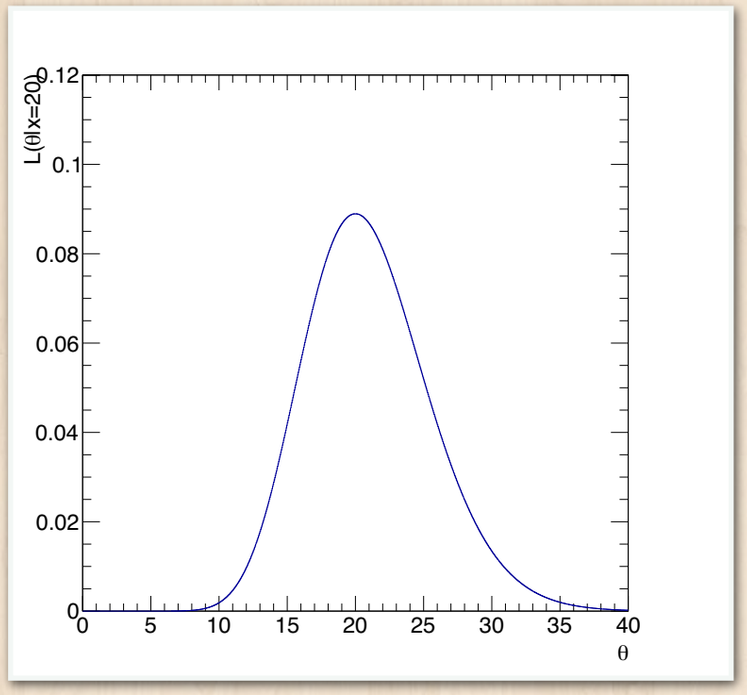
The 20 ft. view



Large dimension

- If we are lucky enough to write down the analytical form, things are easy — we can derive many things
- What if we have a large number of parameters?
- We build the function **numerically**
  - For example to get the Higgs CP property, we have the Higgs mass shape, anomalous couplings, etc
  - CMS “MD” method in $H \to 4\ell$: 12 observables, 10+ parameters
- One potential way is to **create a set of samples that distributes according to the posterior function**

Sampling



Sampling →

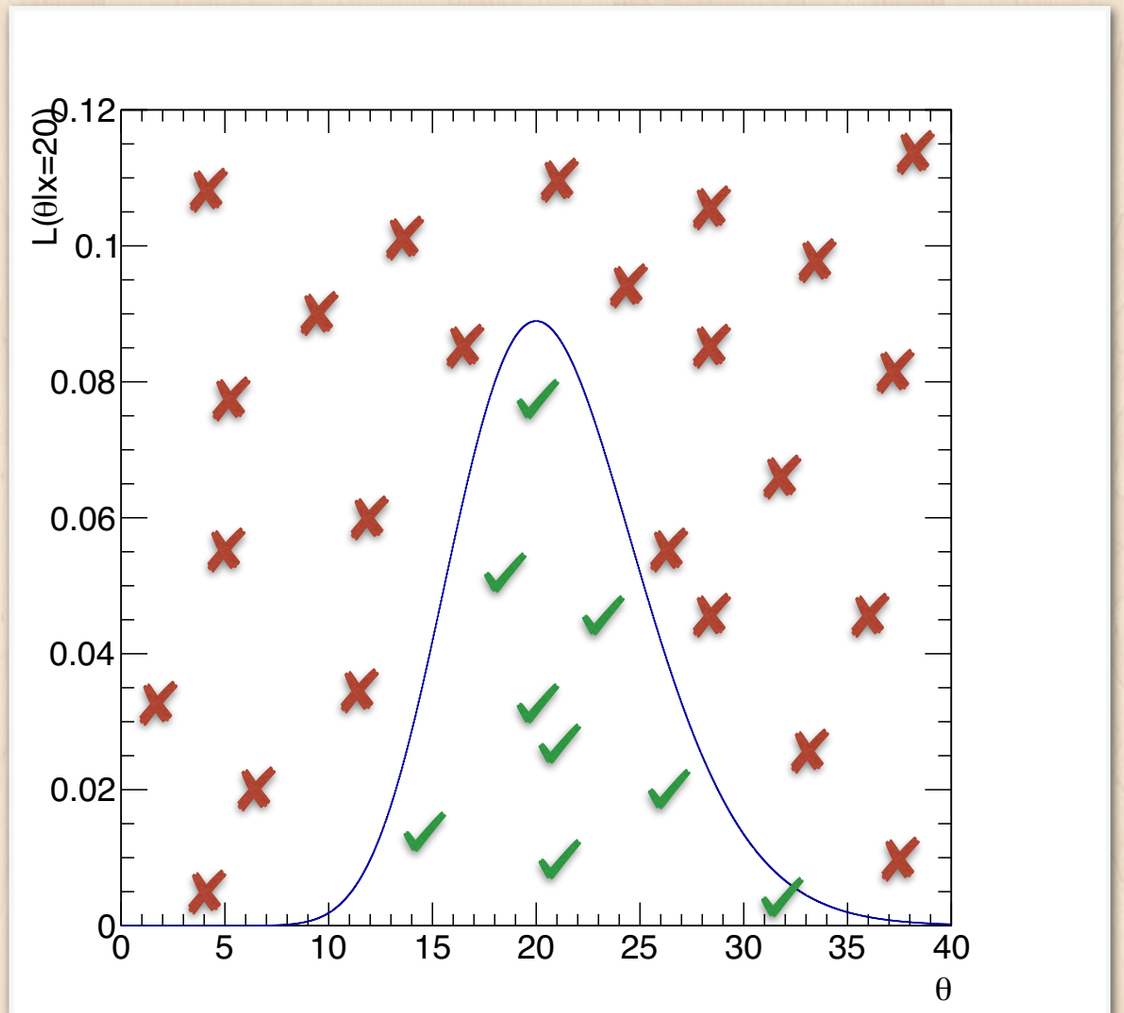
← Histogram

{20.1, 20.9, 17.5, 15.2, 30.8, 19.7, 20.3, …}

In other words, we create a large set of numbers, when plotted as a histogram, gives us back the function

Then we can analyze the samples without worrying too much about the costly calculations

Sampling: how?



Conceptually simplest way: shoot darts

Randomly pick points on the plot, and collect the θ values of the ones falling under the curve

This would work — though not necessary efficient

MCMC

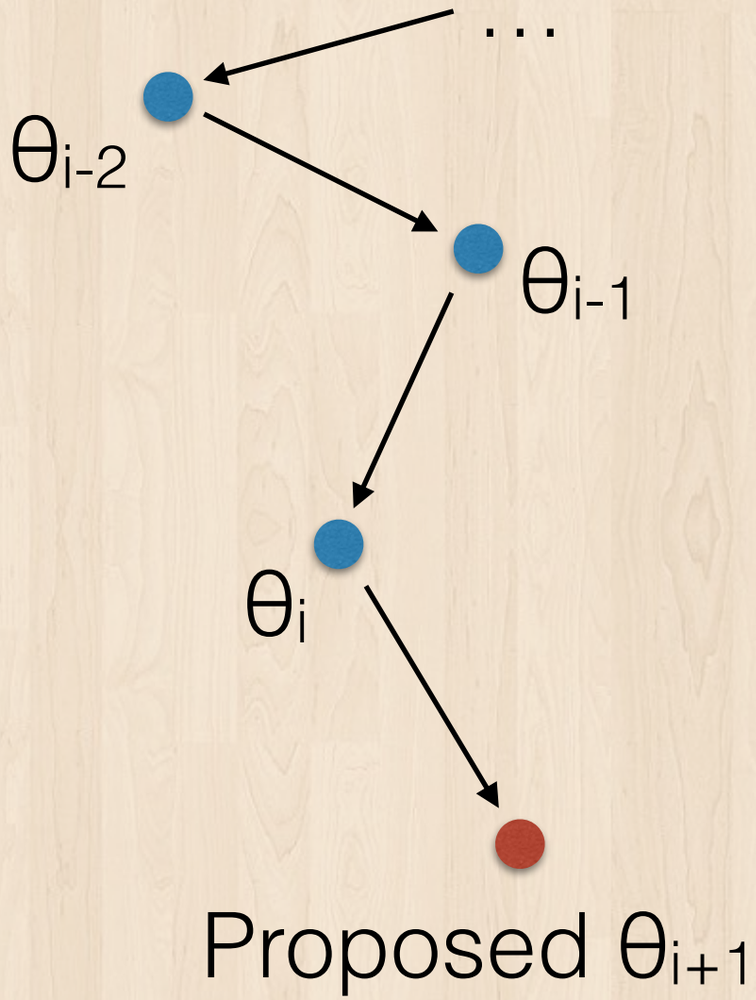
- Markov-Chain Monte-Carlo (MCMC) is another way to achieve the same thing

- It walks through the phase space with some special algorithm: $\theta_i \rightarrow \theta_{i+1} \rightarrow \theta_{i+2} \rightarrow \ldots$

  - Such that <u>asymptotically</u>, $\{\theta_i\}$ approaches $P(\theta|x)$

  - We call $\{\theta_i\}$ a “chain”. Ie., a chain of samples

ex. Metropolis algorithm



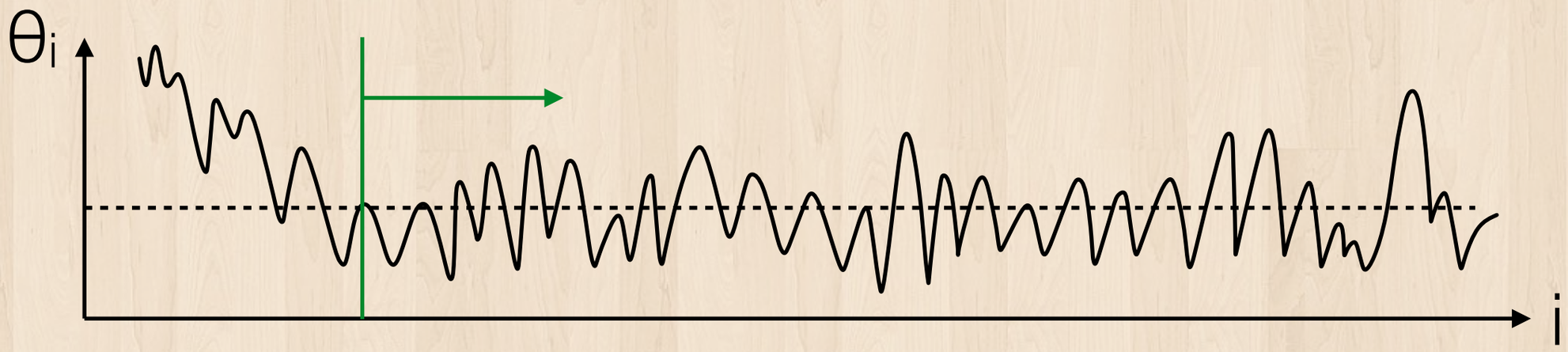
1. Pick a proposed location to move to
2. Evaluate likelihood $P(\theta_i)$ and P(proposal)
   A. If $P(\theta_i) < P(\text{proposal})$, go
   B. Otherwise, throw a dice to see if we move — with probability $P(\text{proposal})/P(\theta_i)$

Direct consequence: samples are very correlated!

{…, 1, 1, 1, 2, 2, 5, 3, …}

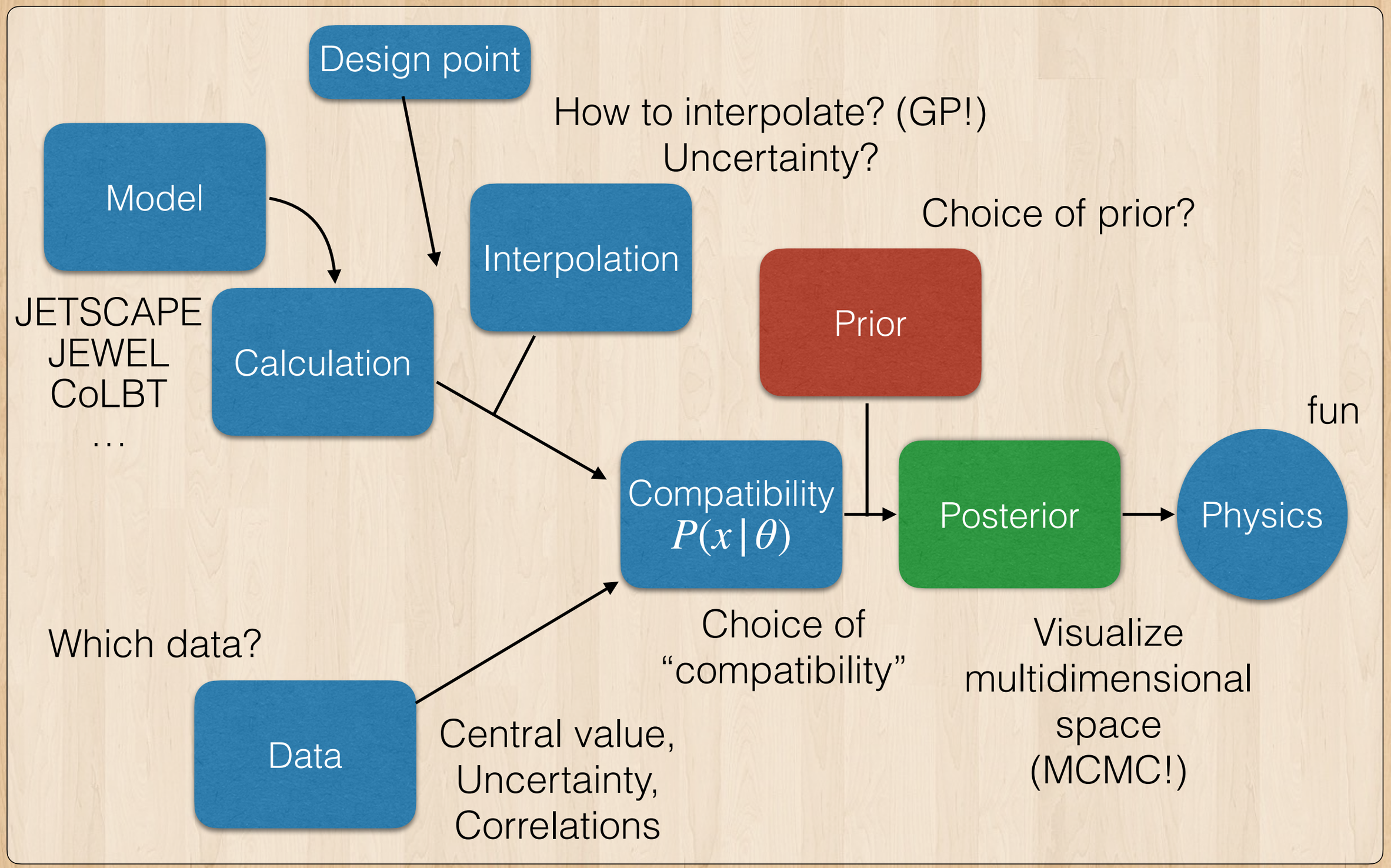
“Burn-in”

- The MCMC will only approach the desired distribution <u>asymptotically</u>
- In other words, when we let the chain go on for a long time, eventually, it will approach the distribution
- This also means that the initial steps do not necessarily follow the posterior



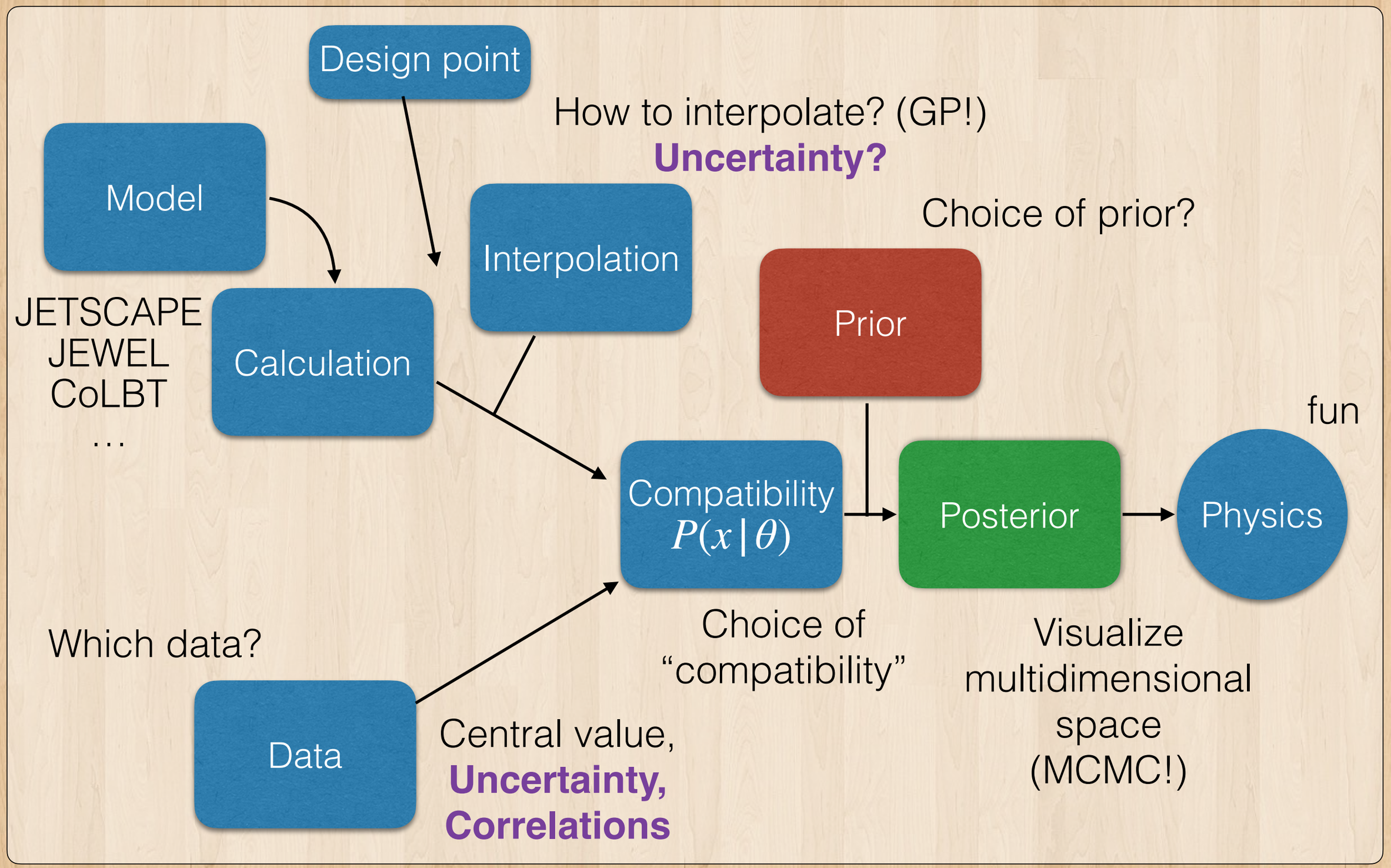
Now let’s do module 4

The 14 ft. view



Part 4: Dealing with uncertainties

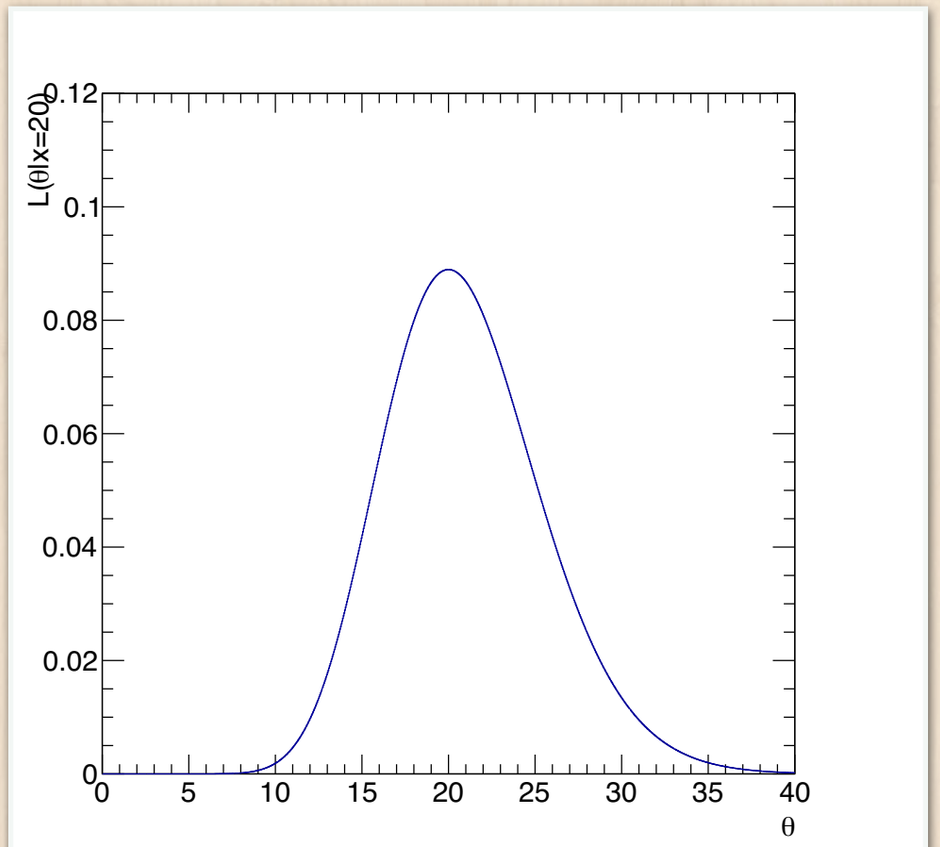
The 14 ft. view



“Description”

To quote numbers we are **describing** the function

$P(\theta \mid x) \rightarrow$ mean, RMS, most probable point, …



Example: counting, x = 20

What	Value
Most probable	20
Mean	20.99
RMS	4.57 (~sqrt(20))
Skewness	0.41

“Description”

- Uncertainty, too, is a description
- The big take home message: in experimental physics, **everything always has a underlying distribution, and the numbers we quote are descriptions that characterizes the function**
- When we say we measure 25 ± 5, we are **describing** the underlying function (following some prescription)
  - For example it could mean that most probable value is 25, and the 68.27% most likely interval is [20, 30]
  - Or it could mean that the range [20, 30] has likelihood value above $1/\sqrt{e}$ (~60.65%) of the peak value
  - Or that the RMS of the distribution is 5, …etc etc

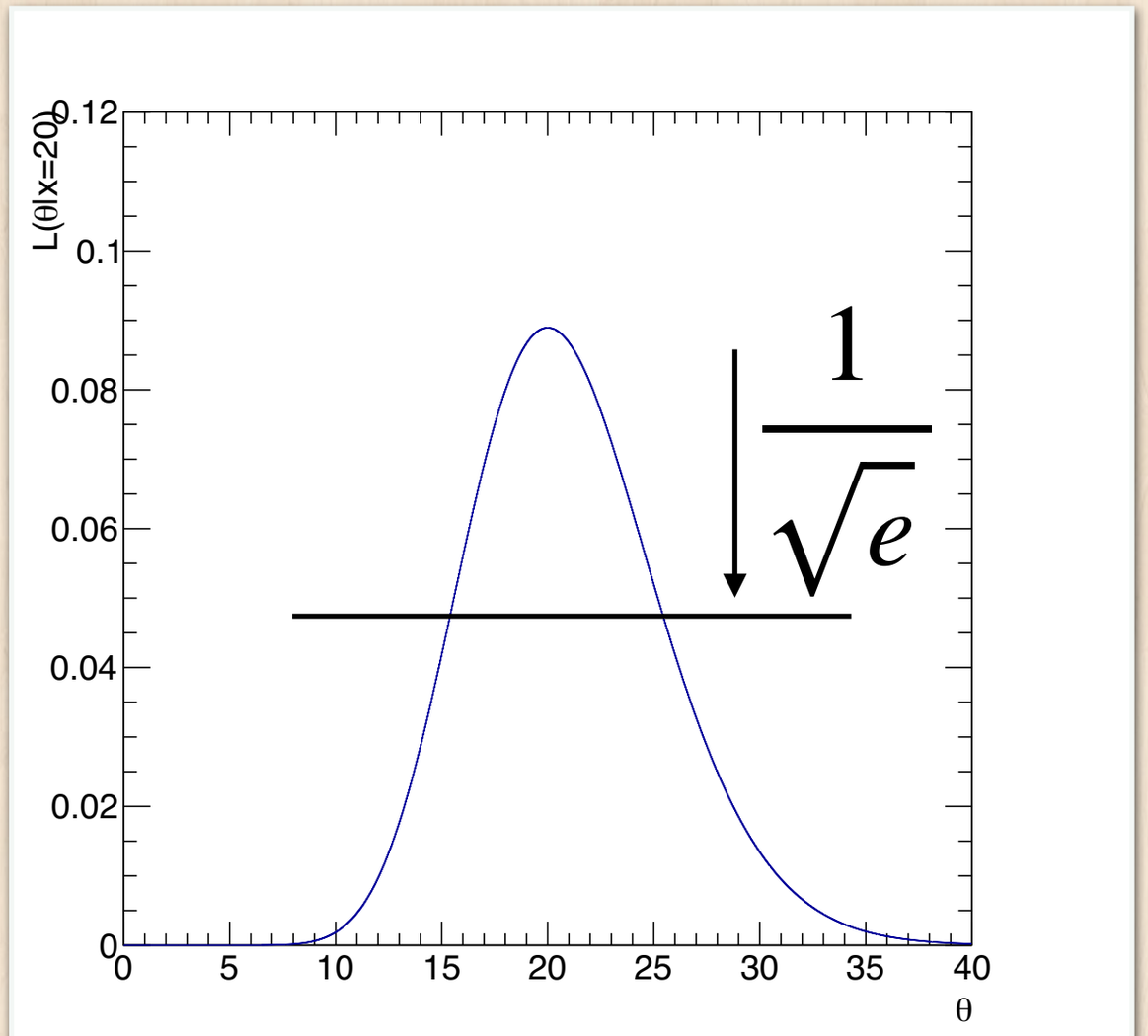
Typically when the distribution is Gaussian these all agree

“Description”

- Try to guess what underlying distribution is assumed and what is quoted!
  - When someone varies jet energy scale by 2% and quote the difference in jet mass distribution
  - When someone smears jet by 5% and quote the difference in jet p_T distribution
  - When someone use an alternative fit function (exponential instead of polynomial) and quote the difference in extracted yield

Example: Poisson statistics

Range of parameters with $L > 1/\sqrt{e} \times L_{max}$ (*)



$$x = 1 \rightarrow \theta \in [0.30, 2.35]$$

$$x = 20 \rightarrow \theta \in [15.9, 24.8]$$

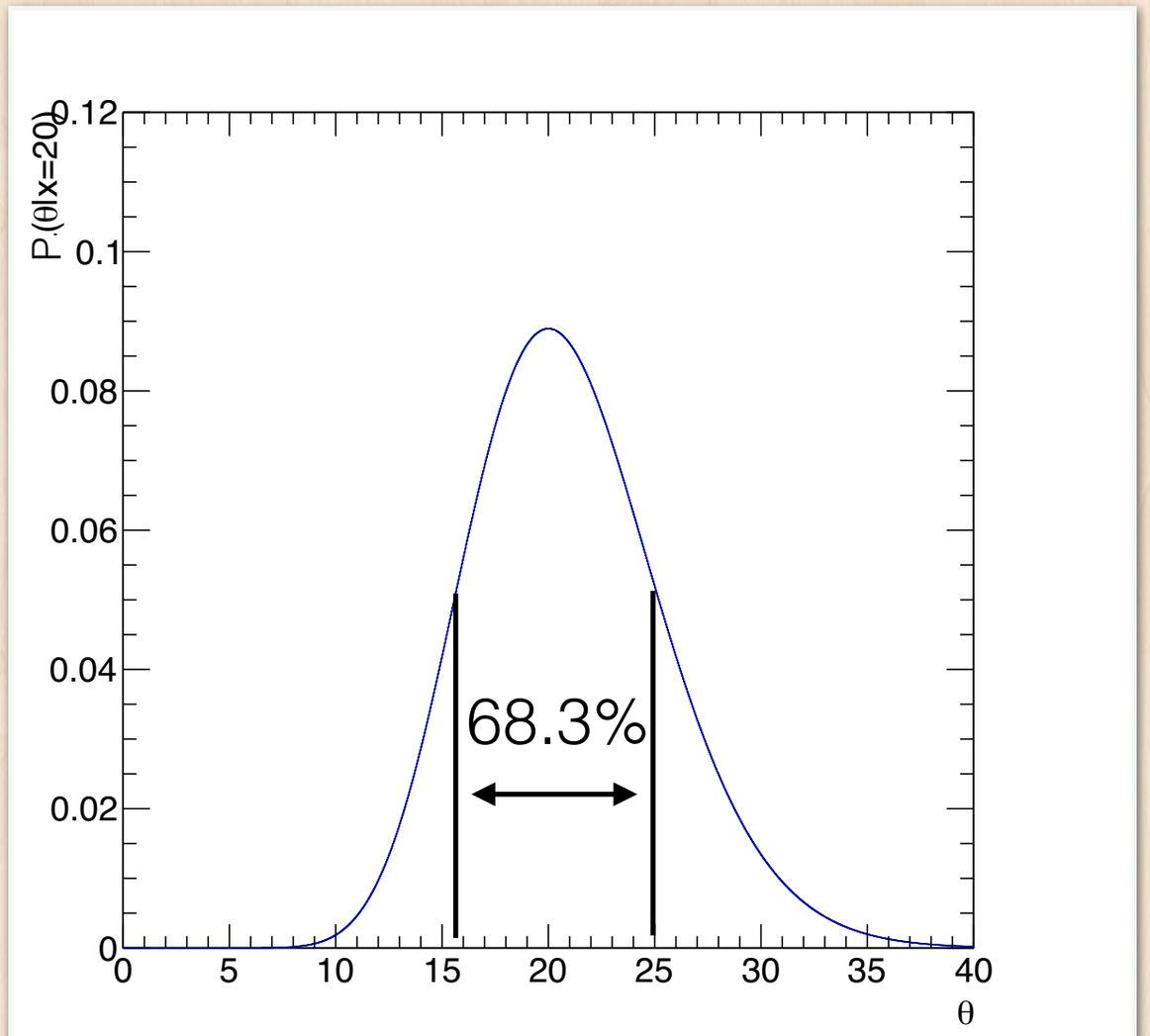
$$x = 100 \rightarrow \theta \in [90.3, 110.3]$$

$$x = 0 \rightarrow \theta \in [?, ?]$$

(*) Or more commonly known as $-2 \ln \Delta L = 1$

Example: Poisson statistics

Most likely parameters enclosing 68.3% of the area



$$x = 1 \rightarrow \theta \in [0.27, 2.50]$$

$$x = 20 \rightarrow \theta \in [15.8, 24.8]$$

$$x = 100 \rightarrow \theta \in [90.3, 110.3]$$

$$x = 0 \rightarrow \theta \in [?, ?]$$

Going beyond: tails matter!

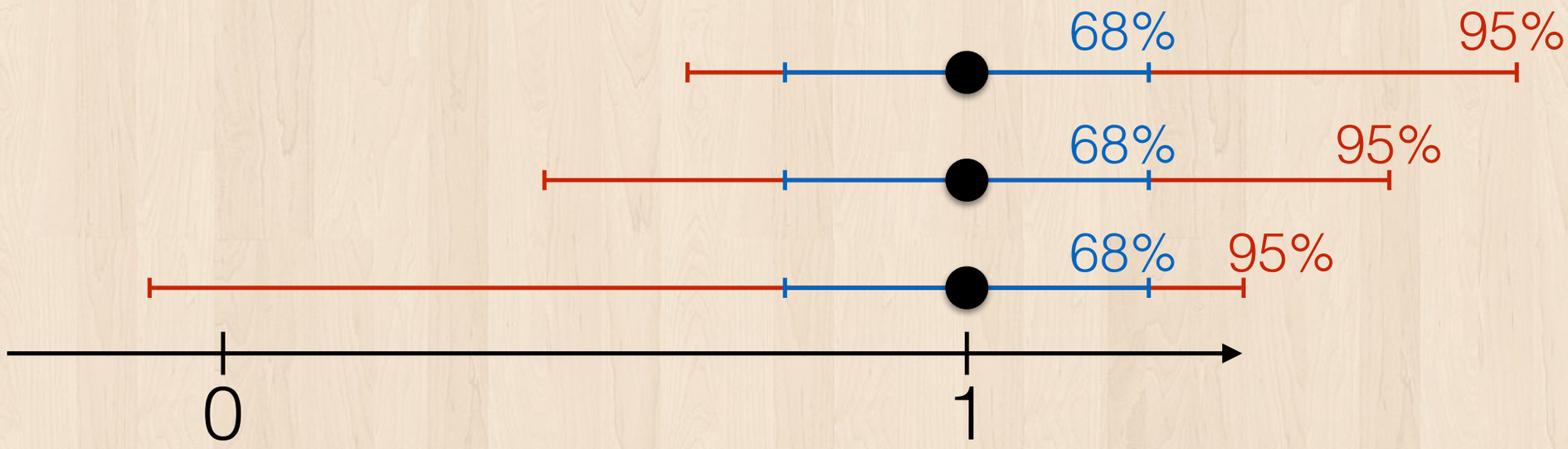
- The uncertainty is just one number that characterizes the width of the (core) distribution
- Often we need to go beyond the single uncertainty number
- Is 1.00 ± 0.25 compatible with 0?

Going beyond: tails matter!

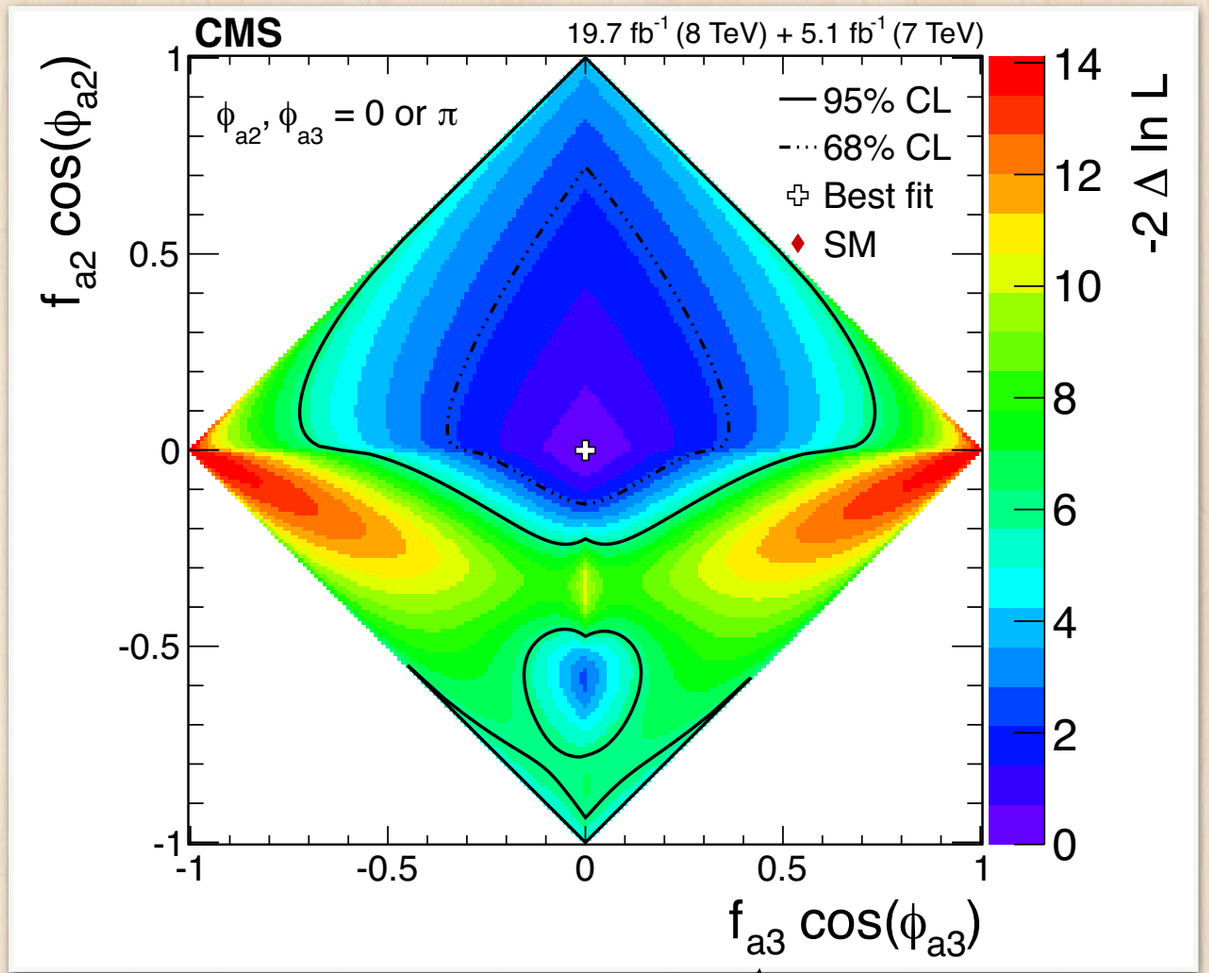
- The uncertainty is just one number that characterizes the width of the (core) distribution

- Often we need to go beyond the single uncertainty number

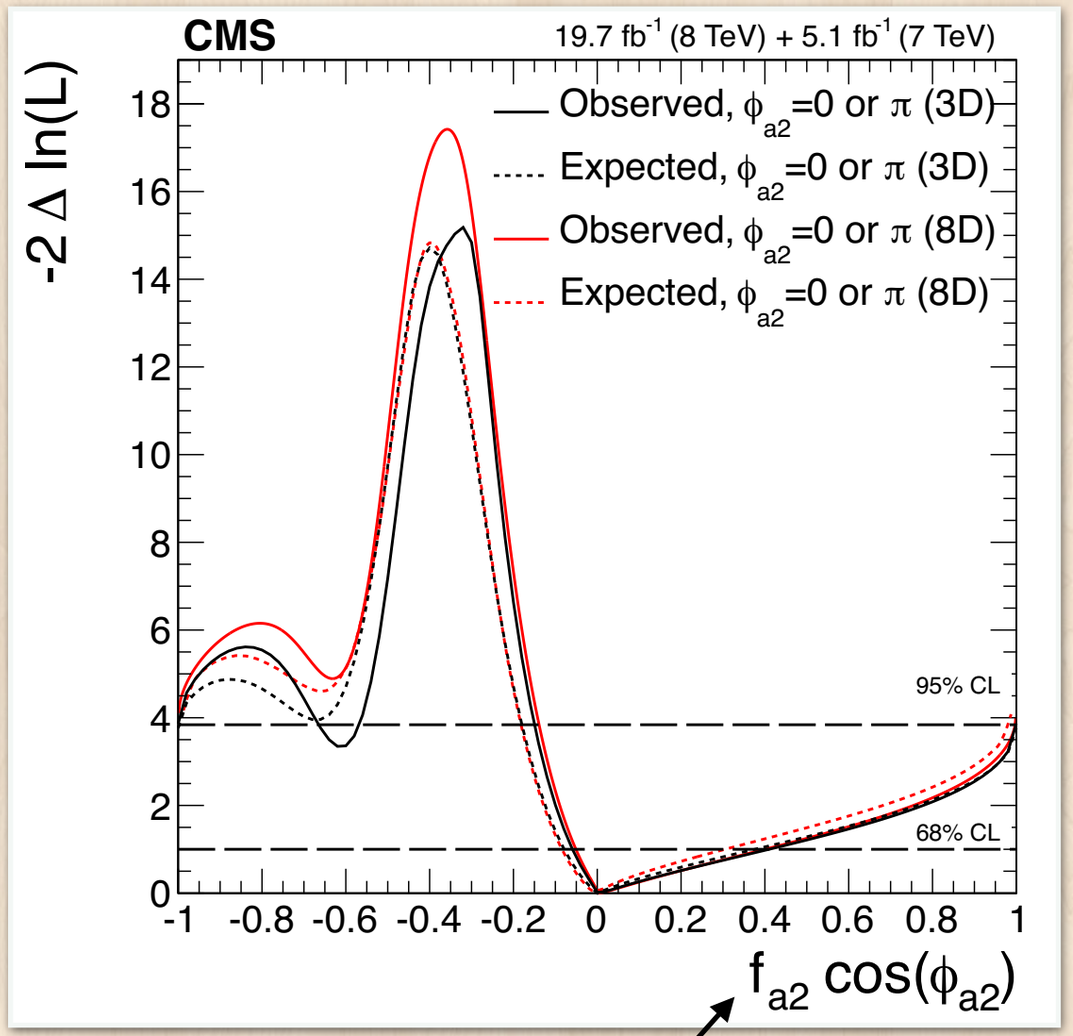
- Is 1.00 ± 0.25 compatible with 0? We don't know!



Example of tails



Size of CP-odd HZZ term



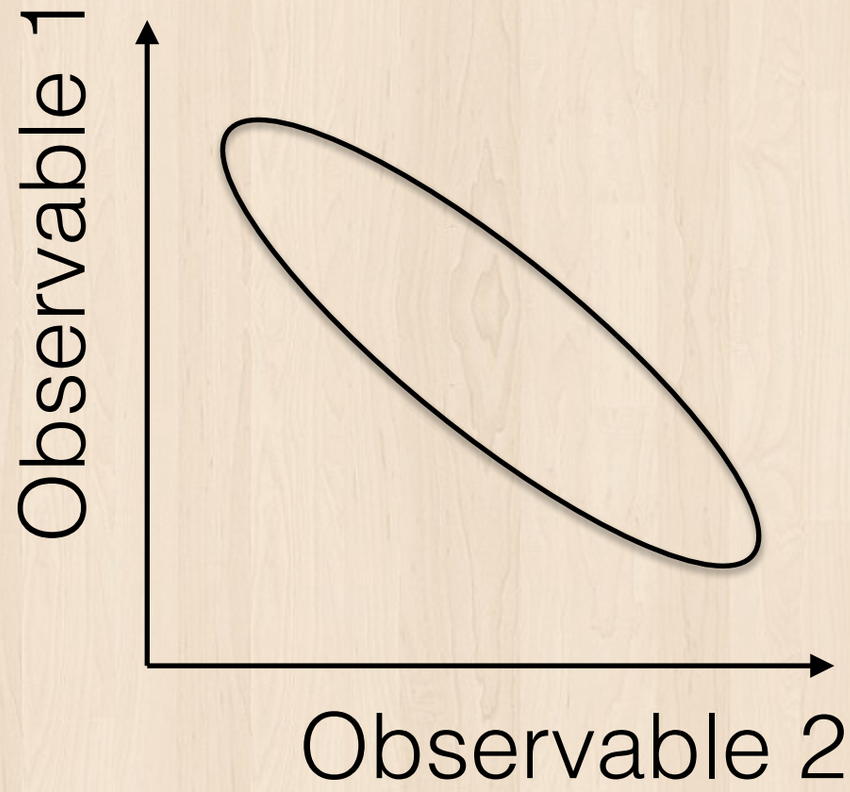
Size of higher order CP-even HZZ term

Uncertainties are correlated

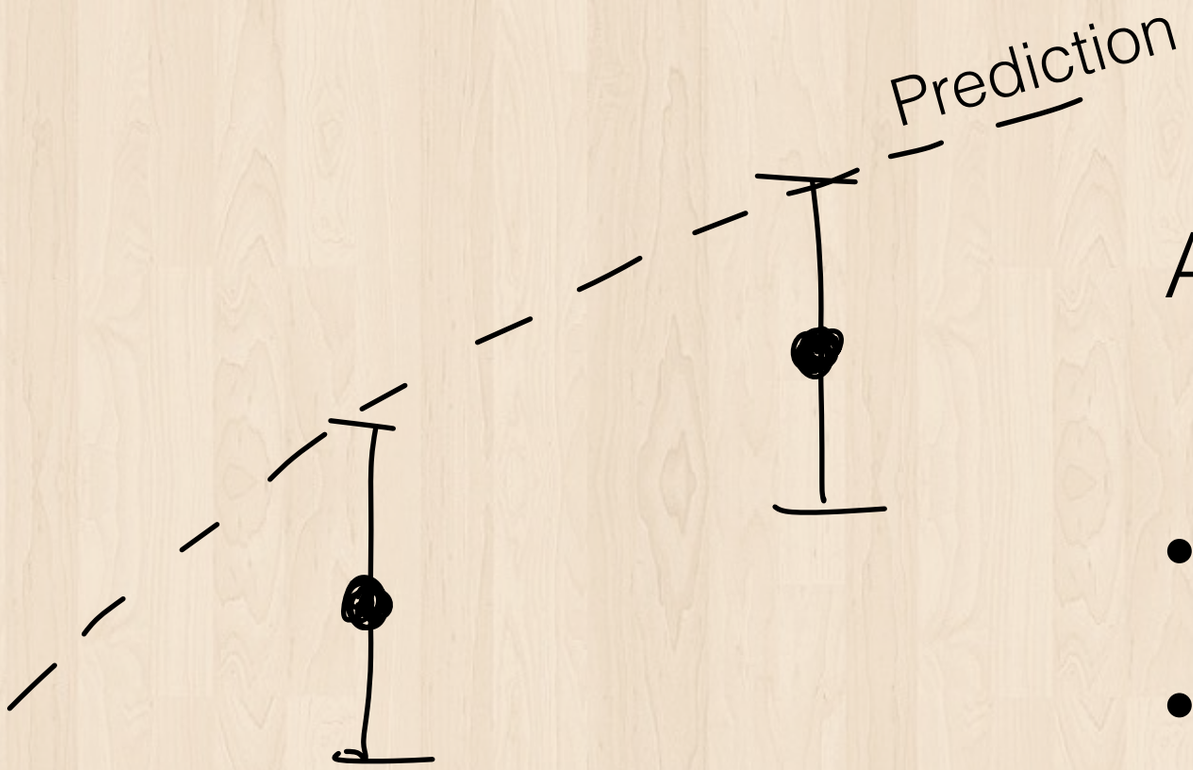
- Uncertainties in general are correlated to some degree

- For example if you measure branching fraction to two different channels

- Even statistical uncertainties are correlated if you unfold



Correlation matters…a lot



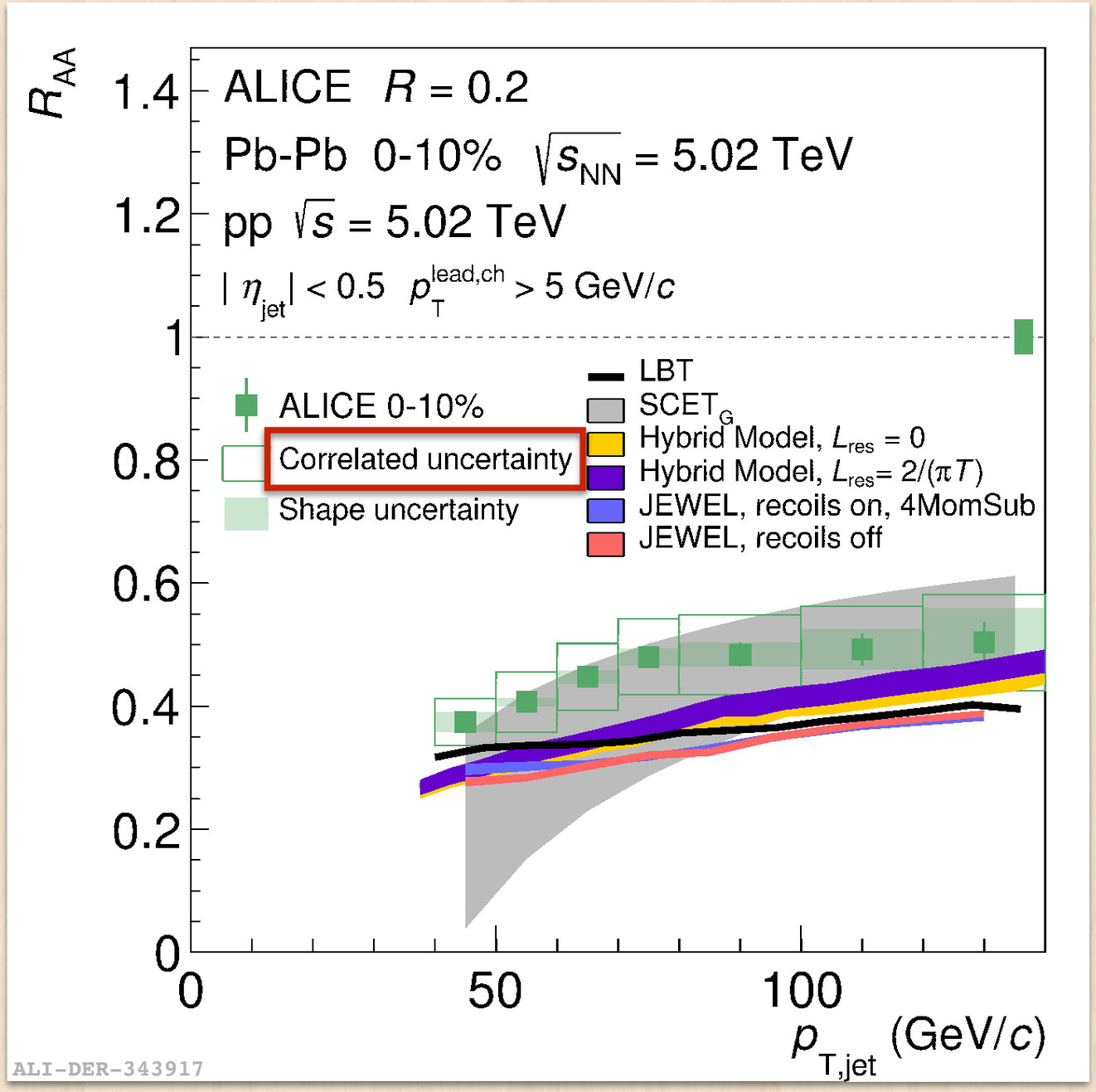
Correlation is key!

Agreement depends on uncertainty correlation

- Fully Correlated: 1σ
- Non-correlated: 2σ
- Anti-correlated: >2σ

Faithfully capturing the correlation is crucial

Correlation is everywhere

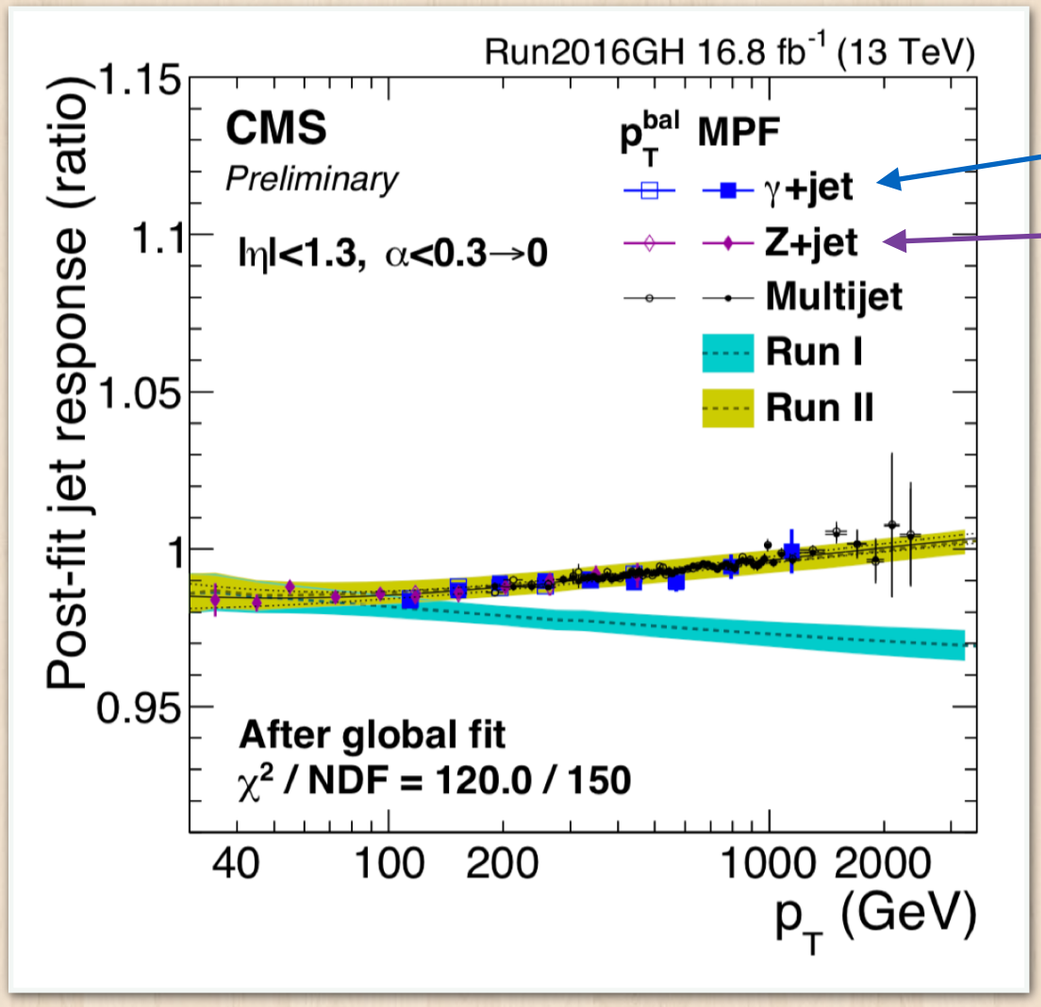


Uncertainties between neighbor PT bins are naturally correlated

For example from jet energy scale, resolution, etc.

Correlation is everywhere

Residual calibration of jet energy scale in data



Jet energy scale from

photon+jet events

Z+jet events

Independent analyses: different people, code, …

Electrons and photons both use electromagnetic calorimeters → uncertainties correlated

Real-life example

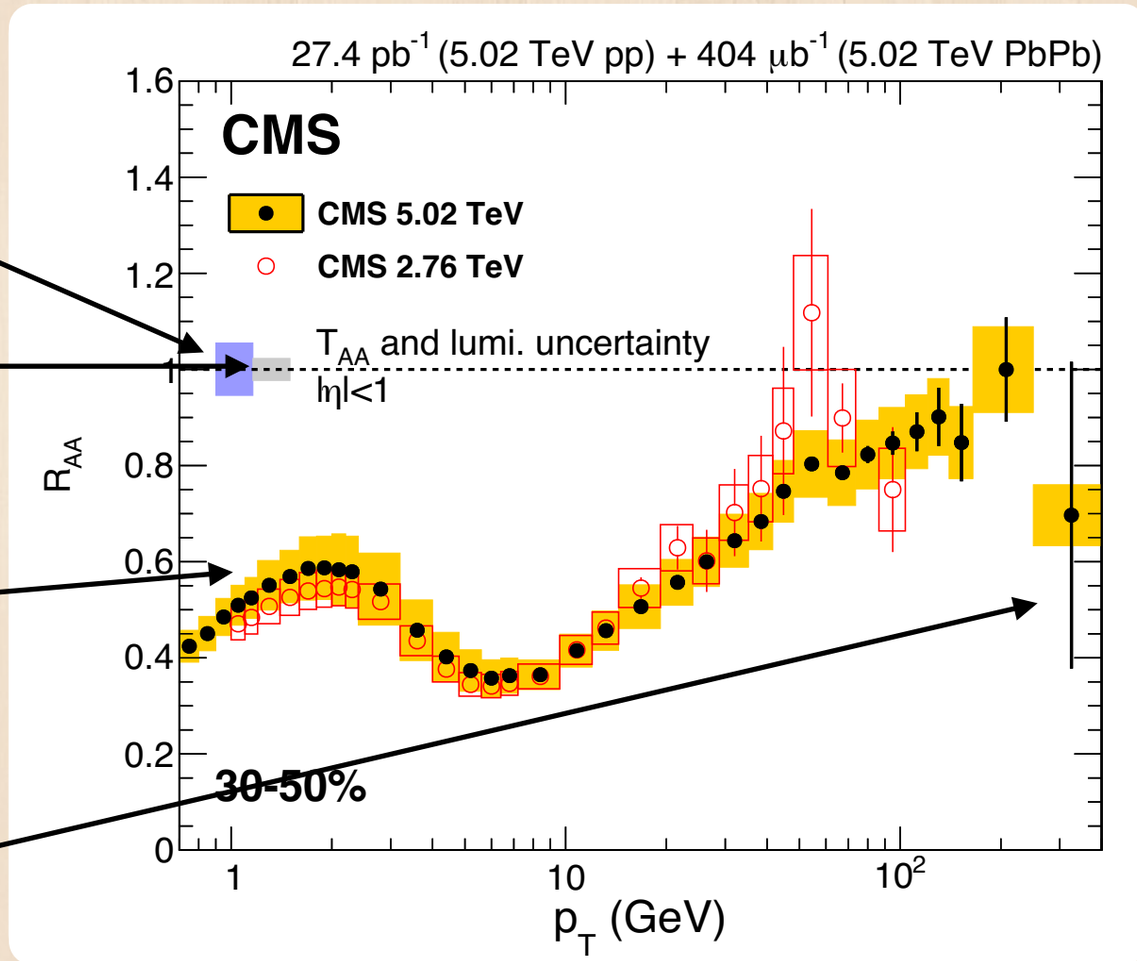
Could be correlated even across experiments!

T_{AA}

Luminosity

Other Systematic Uncertainties

Statistical Uncertainty



Modeling the correlation

In order to model correlation, first we need to choose a compatibility function between data and calculation

For example we can choose the multivariate Gaussian

$$-\ln \mathscr{L} \sim \frac{1}{2}(\vec{x} - \vec{x}_\theta)^T \Sigma^{-1} (\vec{x} - \vec{x}_\theta) + \dots$$

Data

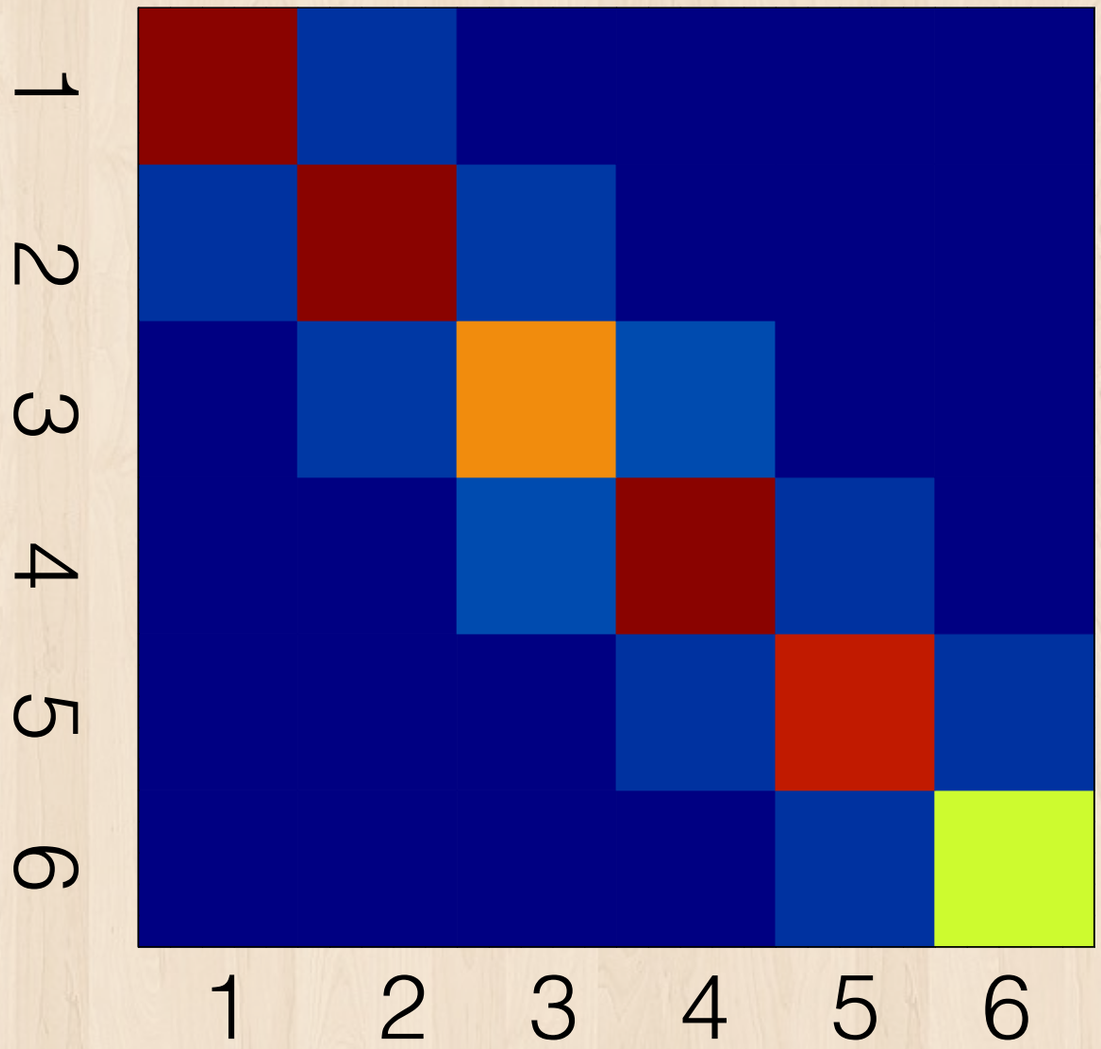
Matrix with uncertainty information

Calculation

ie., Bayesian likelihood

Covariance matrix

Σ: Covariance matrix $\quad \Sigma_{ij} = E[(X_i - \overline{X}_i)(X_j - \overline{X}_j)]$



The analog of variance in multi-dimension

$$-\frac{1}{2}(\vec{x} - \vec{x}_\theta)^T \Sigma^{-1} (\vec{x} - \vec{x}_\theta)$$

$$-\frac{(x - x_\theta)^2}{2\sigma^2}$$

Off-diagonal entries = correlated

Difference sources

$\Sigma^{\text{int.}}$

From interpolation etc.

Uncertainty from choice of Bayesian analysis details

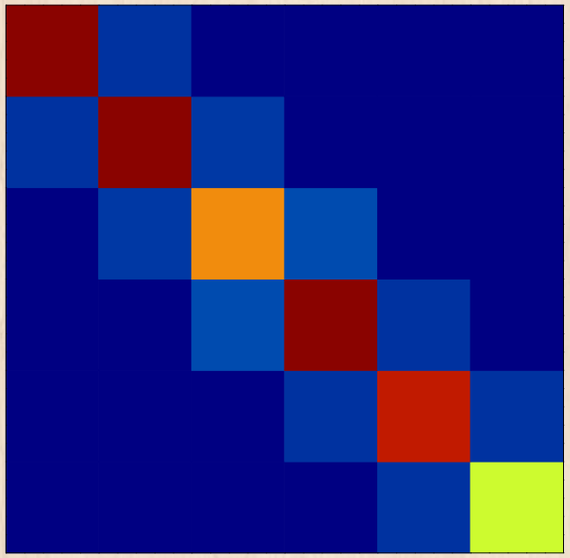
$\Sigma^{\text{exp.}}$

From experiments

Do the best we can to reproduce ~~guess~~ this

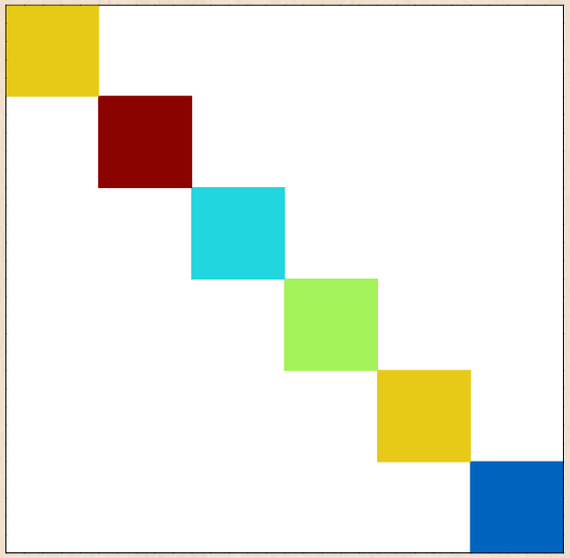
Building the matrix

Full Covariance



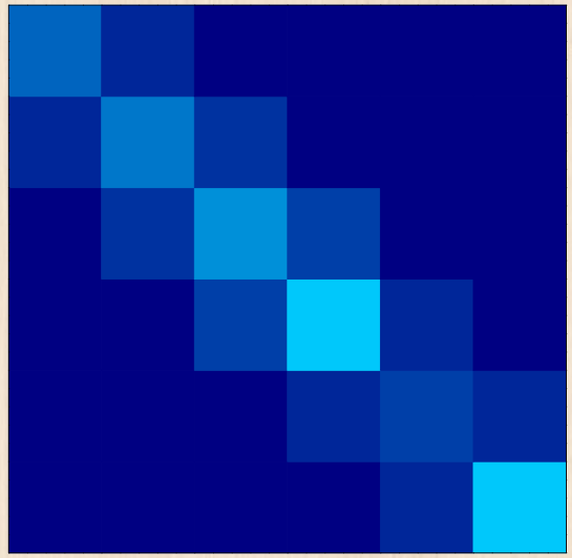
=

Uncorrelated



+

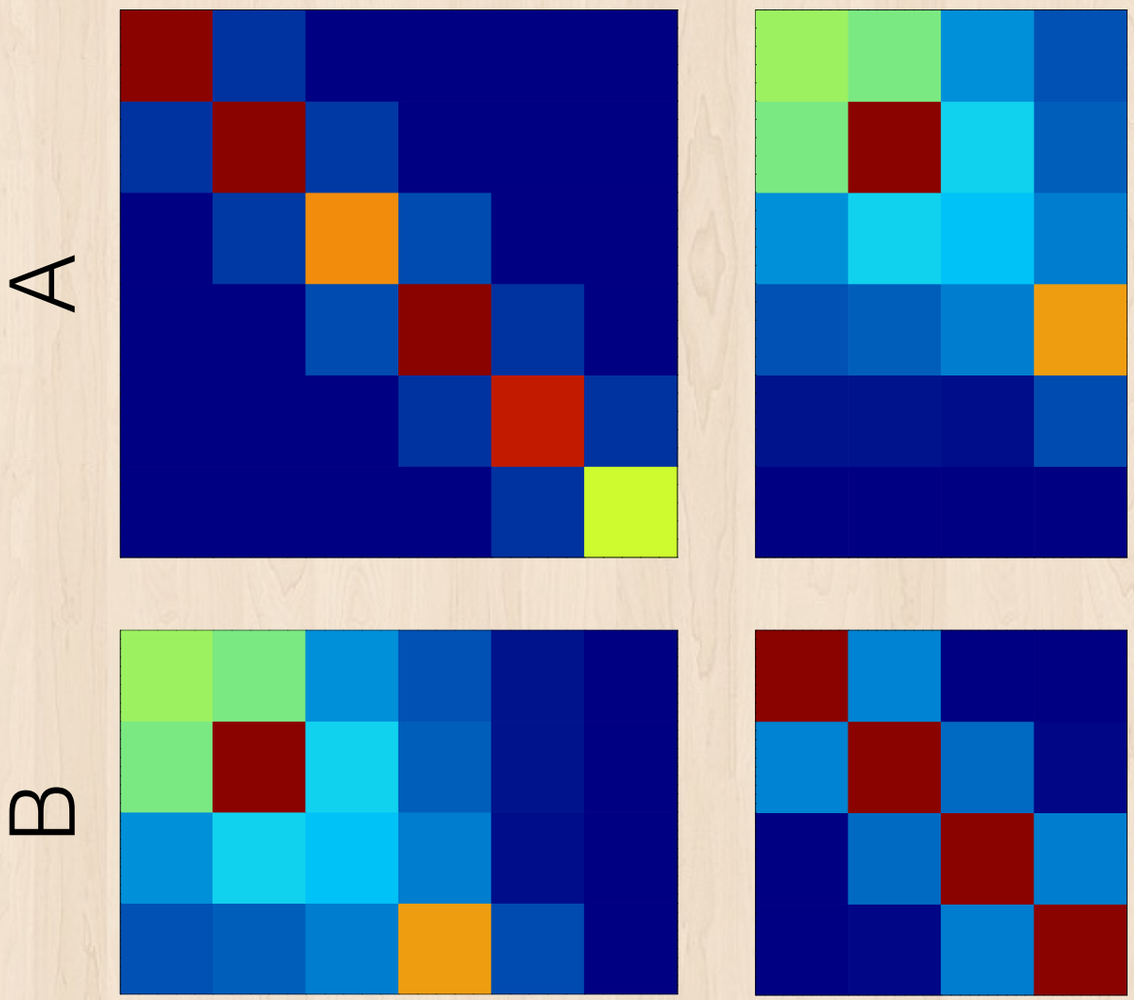
Somewhat correlated



+ …

Add up the covariance matrices source-by-source

Capture Correlations



Some uncertainties are correlated across measurements

For example luminosity uncertainty from the same experiment

...

Guesses

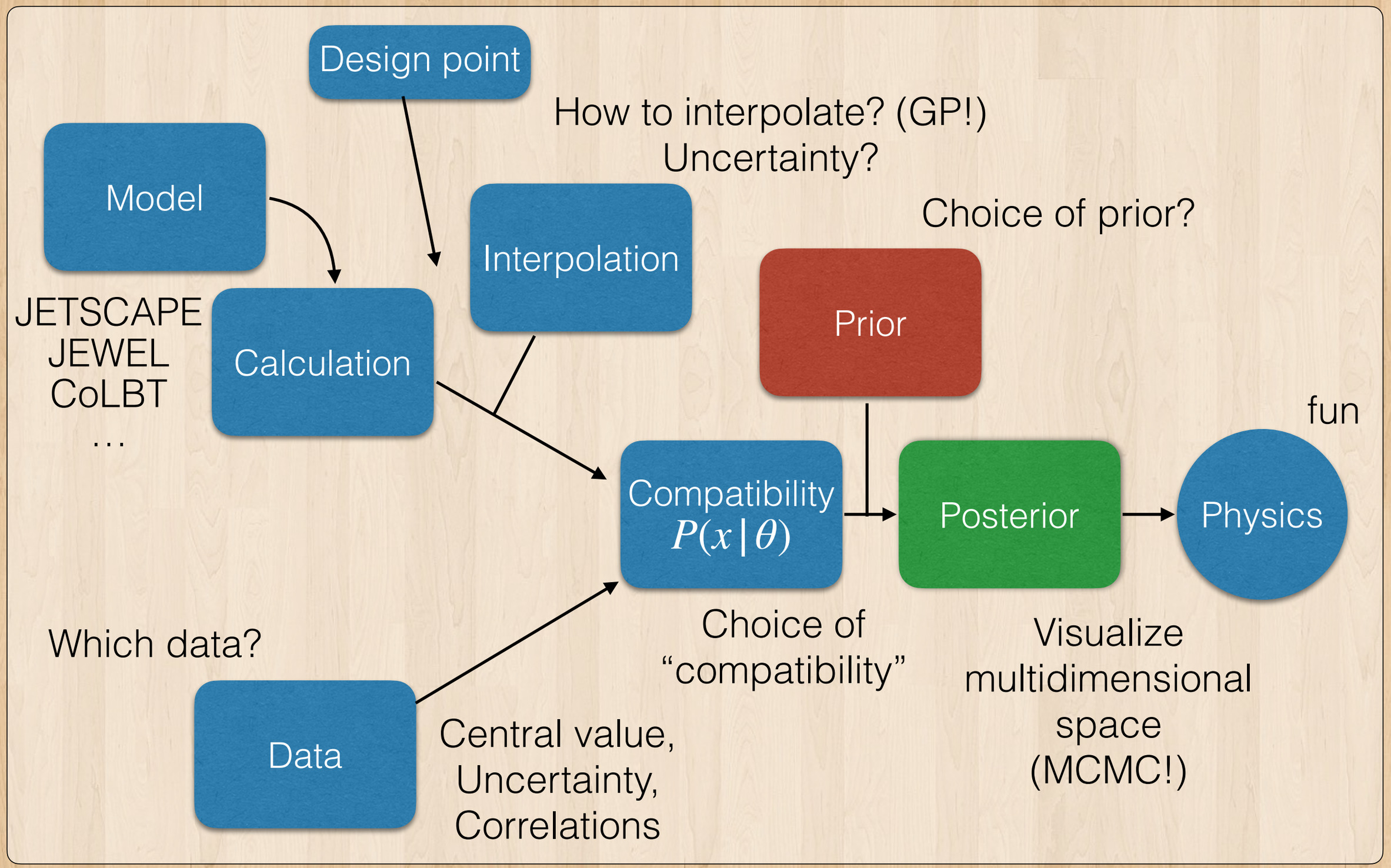
- Information from experiments is limited
- In many cases we are forced to make guesses
- The specific guesses then become either…
  - caveat in the analysis we perform, or
  - additional uncertainties on extracted parameters if we systematically explore valid guesses

Recap: uncertainties

- Many sources of experimental uncertainties (and modeling)
  - Correlation matters!
  - Tails beyond 1σ matter!
- Sometimes we have to make some guesses…

Now let's do module 5

The 14 ft. view



What's the game plan?

1. Identify the data $\{x_i, y_i\}$ we want to use
2. Find a function $f(x_i|\theta)$ we want to use
3. Define what we mean by "how well things match"
4. Vary θ to minimize the function in item 3.
5. Physics!

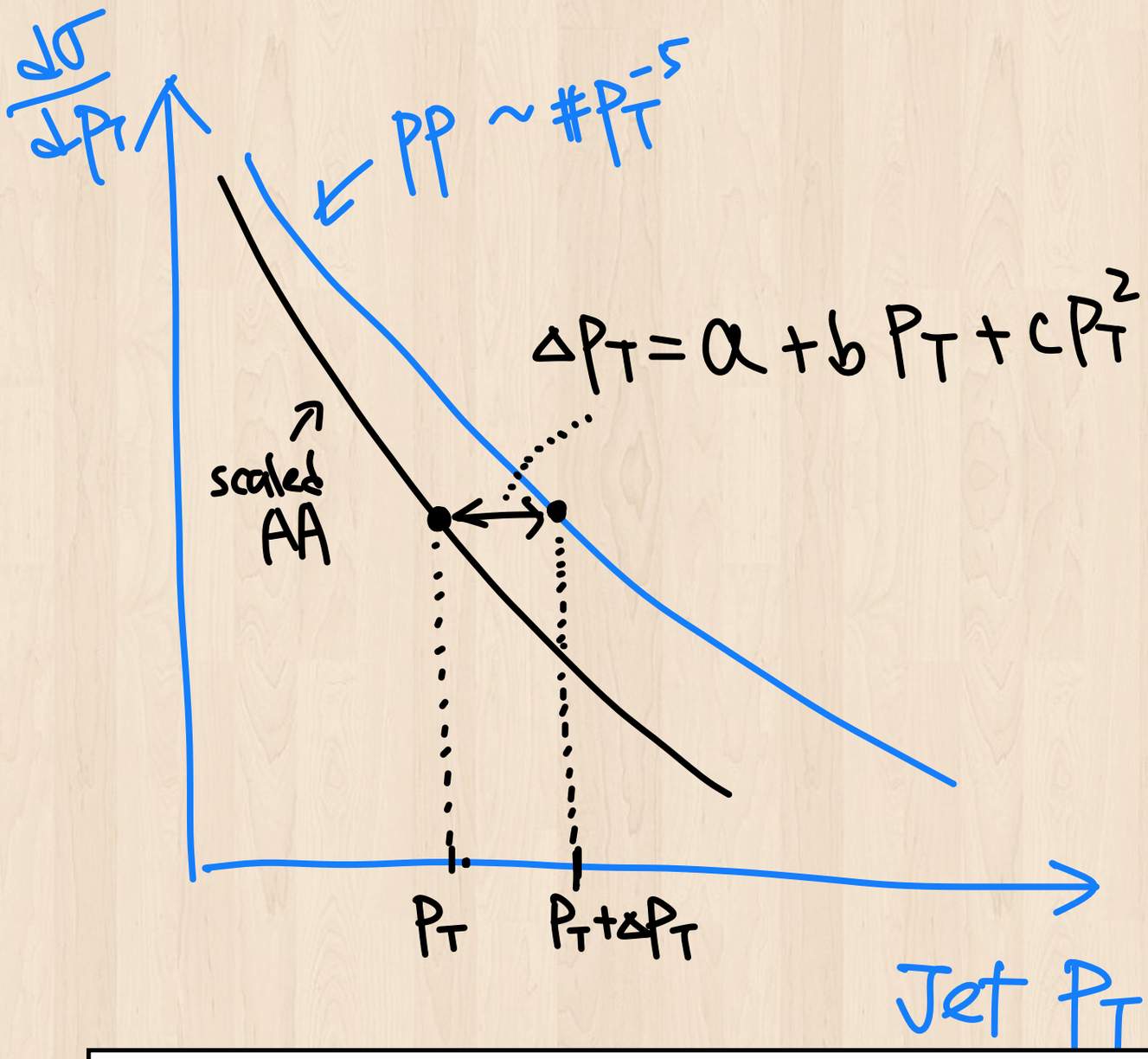
Long exercise

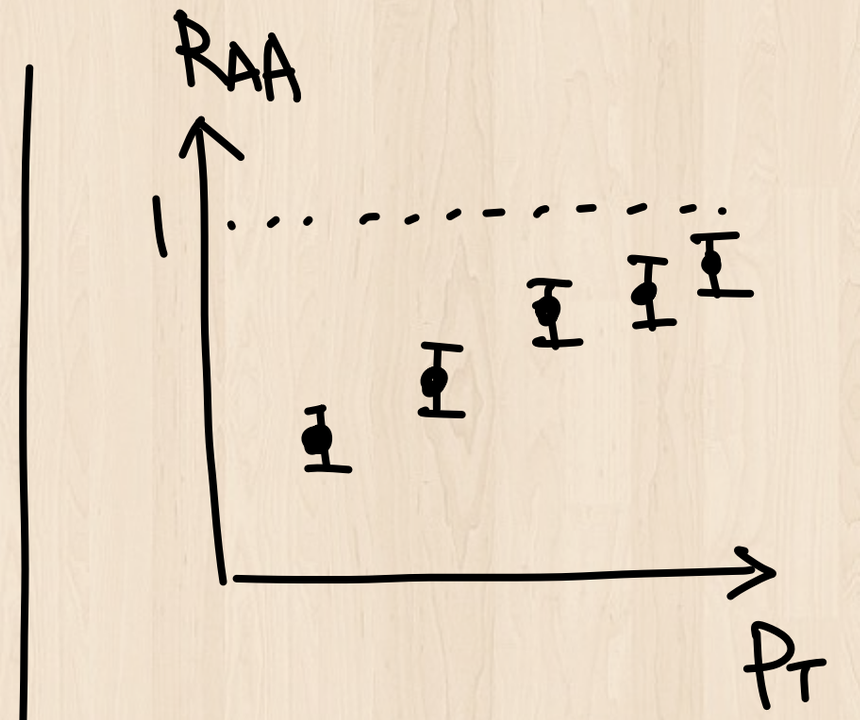
Long exercise

- Module 6 notebook has everything we have talked about so far (except the uncertainty correlations)
  - Please check that you can run things as-is
- Then — we will attempt to do a fun long exercise!
- We will fit for the amount of jet energy loss to create inclusive jet R_{AA} (some mock up data), assuming that things just move horizontally

Setup [effort level: ⭐⭐]

Data is in
`long_exercise.txt`





goal: learn (a, b, c)

a nice intermediate step would be to do $a + b \cdot p_T$

Now let's start from module 6

Note

for 3D parameter space this needs to be three 1.0's

```
22 print("scaled theory-value range:", np.min(t
23
24 gpe_list = []
25 training_warnings = []
26 for ix, x_value in enumerate(x_theory):
27     kernel = ConstantKernel(1.0, constant_va
28         length_scale=[1.0, 1.0, 1.0],
29         length_scale_bounds=(1e-2, 1e2),
30         nu=1.5,
31     )
32     gpe = GaussianProcessRegressor(
33         kernel=kernel,
34         alpha=1e-8,
```

This is Gaussian process block

Stretch goals

- If you are done with the long exercise and are bored, consider these expansions
  - What happens when the uncertainties are correlated? [effort level: ⭐] Anti-correlated? [effort level: ⭐]
  - Let's split out systematic and statistical uncertainties. Build covariance separately and add. What happens with different assumptions? [effort level: ⭐]
- Let's get some photon-jet $x_{J\gamma}$ data from HEPData. Can you fit the same model to the data? Do you get compatible results? [effort level: ⭐⭐⭐]
- Can you fit for BOTH at the same time? What do you see? [effort level: ⭐]

Part 5: Beyond the baseline workflow

Backup Slides Ahead