Practical Statistics for Particle Physics

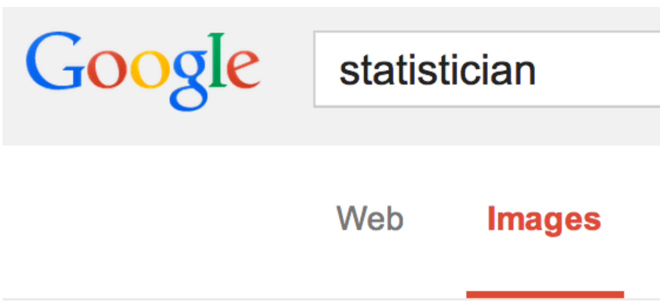
Daniel Whiteson, UC Irvine
HCPSS, 2014

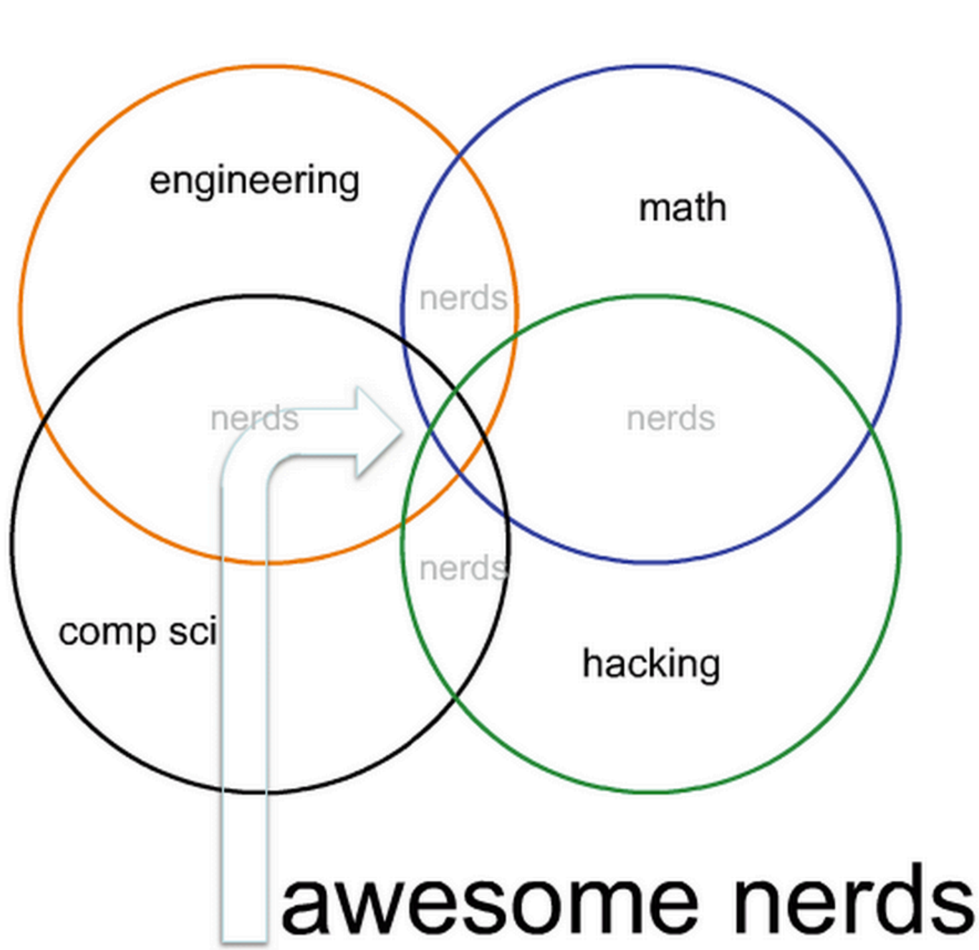
Caveat

I am not a professional statistician!

Caveat

I am not a professional statistician!





Motivation

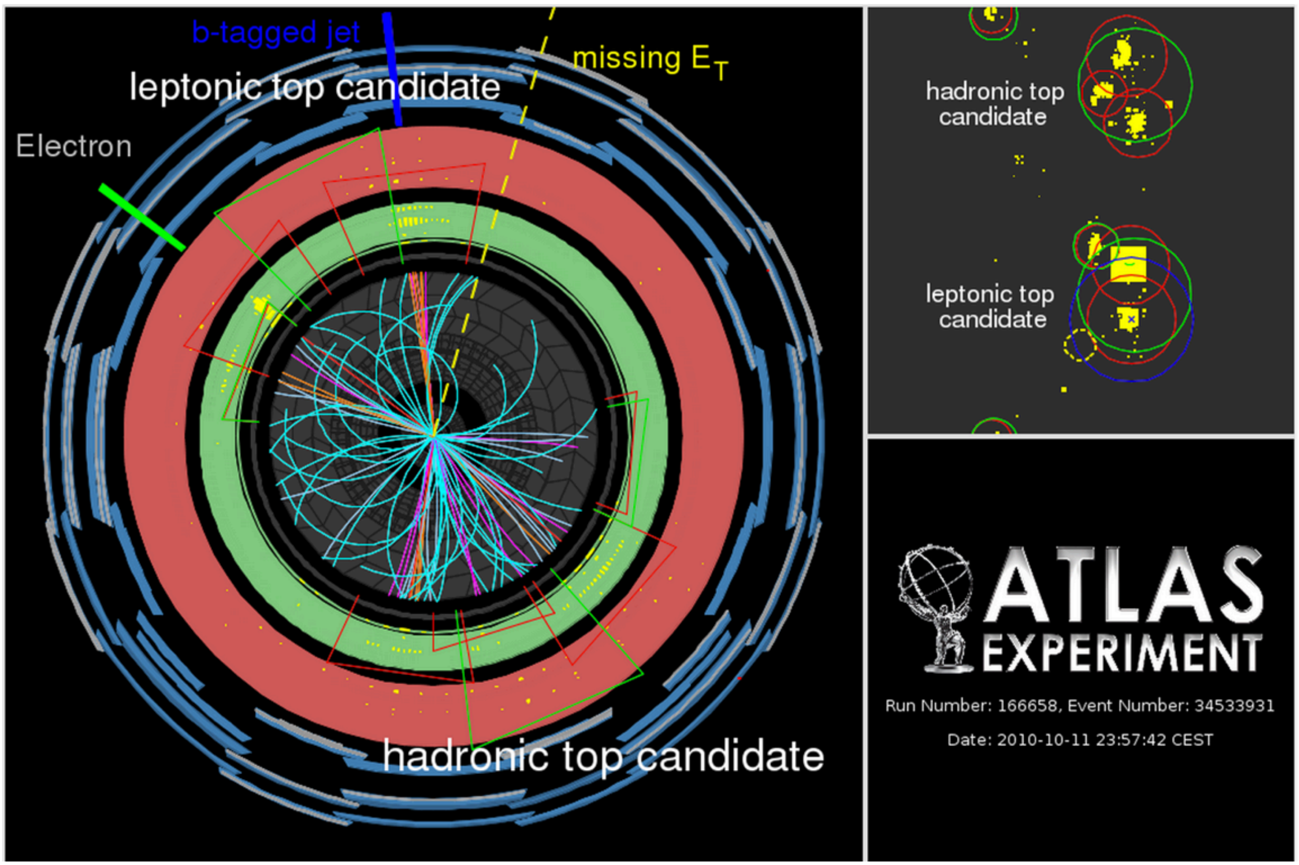
Why do we need statistics?

Why?

"The data were inconclusive, so we applied statistics"

L.Lyons (?)

What's in an event?



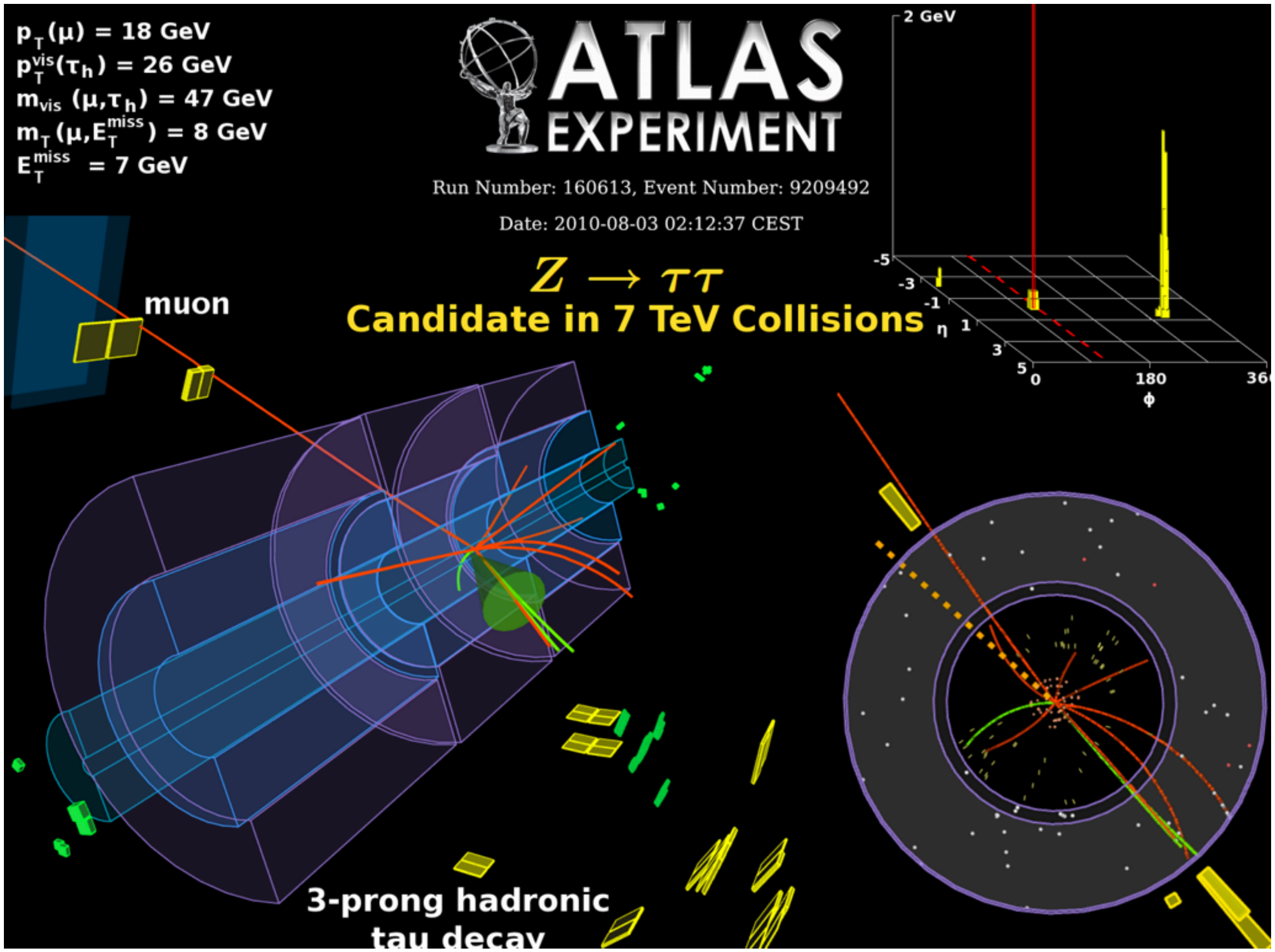
No event can be unambiguously interpreted.

Unambiguous data

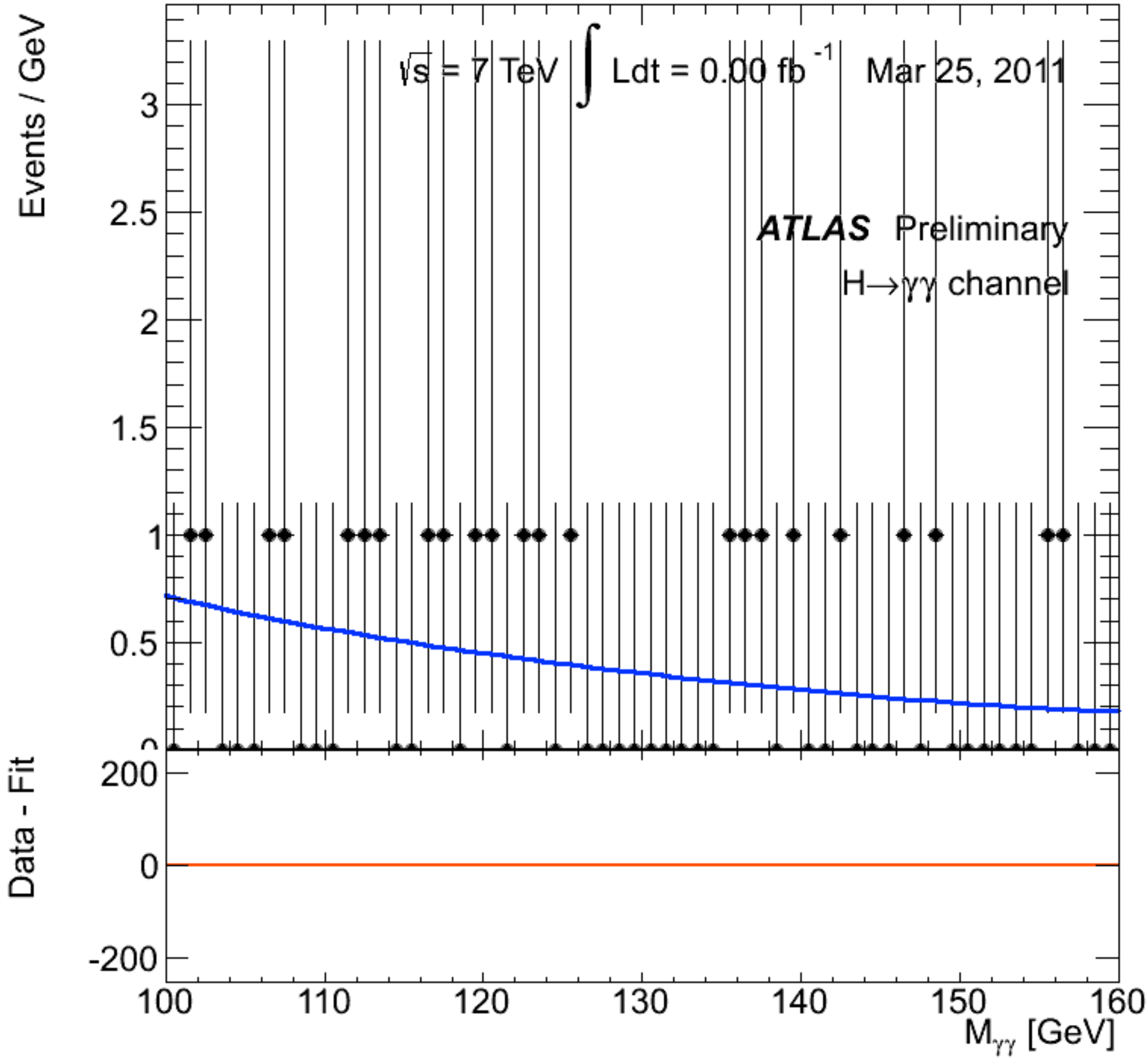
Ok, but see:

http://cerncourier.com/cws/article/cern/54388

Why statistics?



The nature of our data demands it.



√s = 7 TeV ∫ Ldt = 0.00 fb -1 Mar 25, 2011
ATLAS Preliminary
H→γγ channel
Events / GeV
3
2.5
2
1.5
1
0.5
0
Data - Fit
200
0
-200
100
110
120
130
140
150
160
Mγγ [GeV]

Other lectures

Kyle Cranmer:

http://indico.cern.ch/event/117033/material/slides/1?contribId=19

https://indico.cern.ch/event/243641/

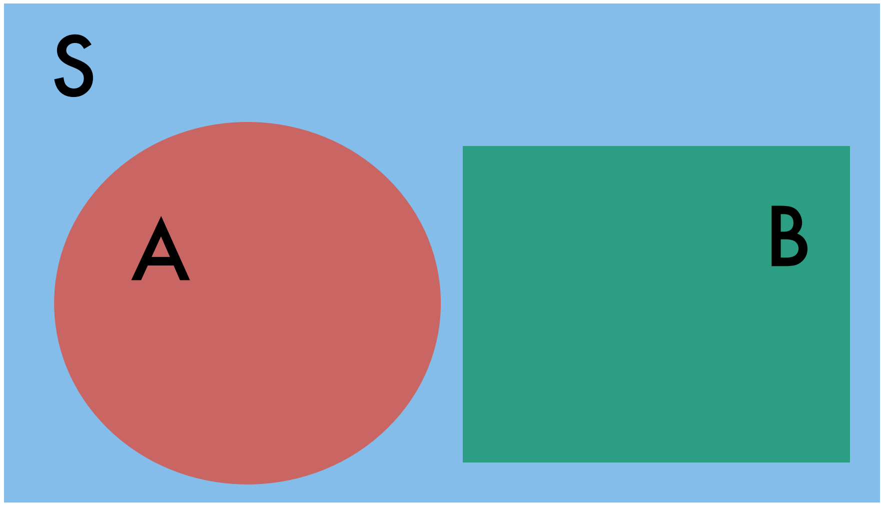
(I have borrowed many of his drawings)

Outline

I. Mathematical preliminaries
II. Fitting
III. Data models
IV. Hypothesis testing
V. Tools and examples

Mathematics

Probability

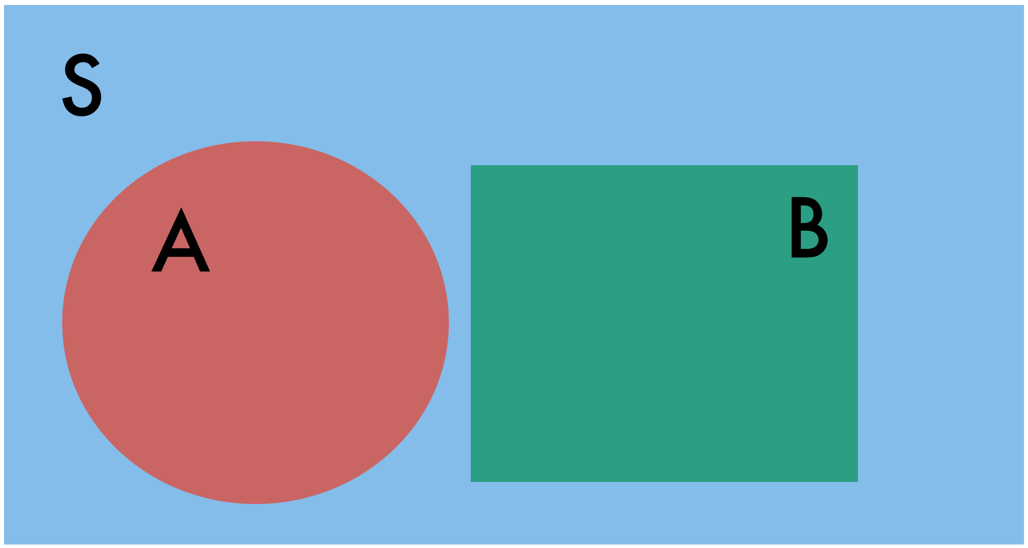


P is a probability over the space S iff

- for every subspace A, $P(A) \geq 0$
- for disjoint subspaces A and B, $P(A or B) = P(A) + P(B)$.
- $P(S) = 1$

Probabilities are between zero and one.

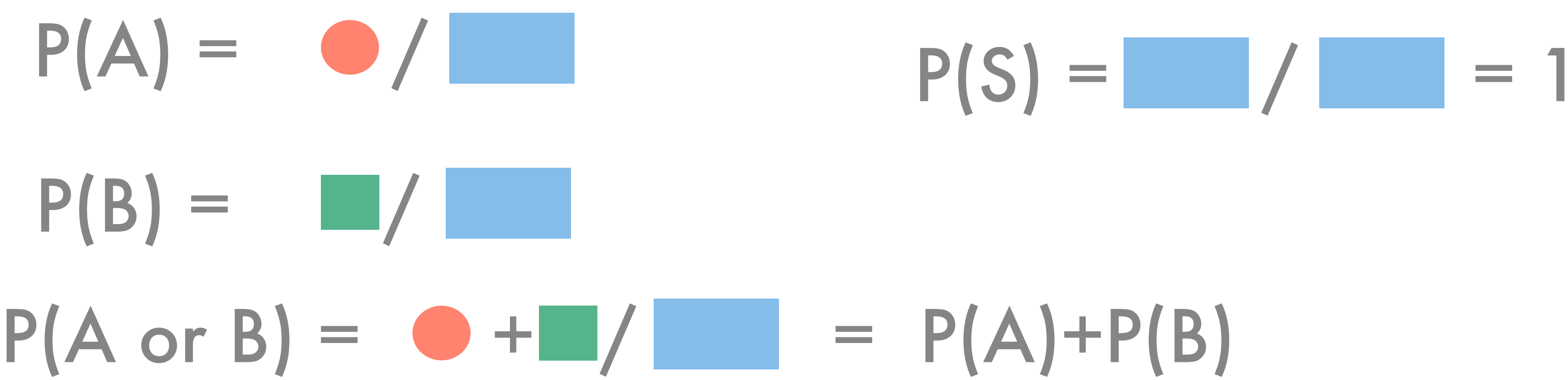
examples



P is a probability over the space S iff

- for every subspace A, $P(A) \geq 0$
- for disjoint subspaces A and B, $P(A or B) = P(A) + P(B)$.
- $P(S) = 1$

Probabilities are between zero and one.



Conditional Prob

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

Need to consider the various cases A_i
then the probability of B in each
of these cases.

Practical application

Conditional Prob

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

Russian Roulette

A_1-A_5 = no bullet

A_6 = bullet

P(Ai) = 1/6

P(death|no bullet) = 0.0001

P(death|bullet) = 1-0.0001

P(death) = 0.0001*5/6 + 1/6*(0.9999) = ~1/6

Probability Density

$$P(x \in [x, x + dx]) = f(x)dx$$

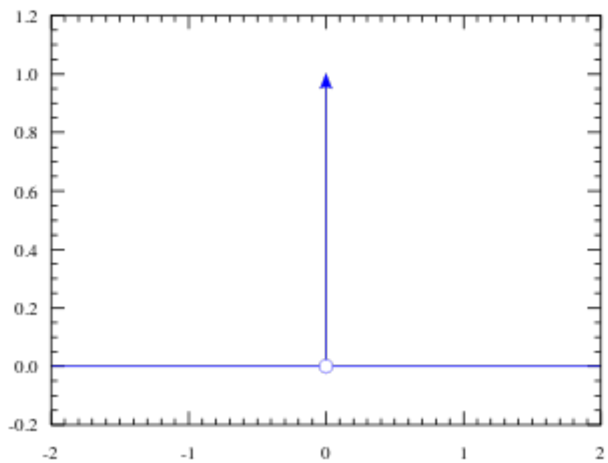
Note f(x) is **not** a probability,
can have **any** positive value.

But must be normalized:

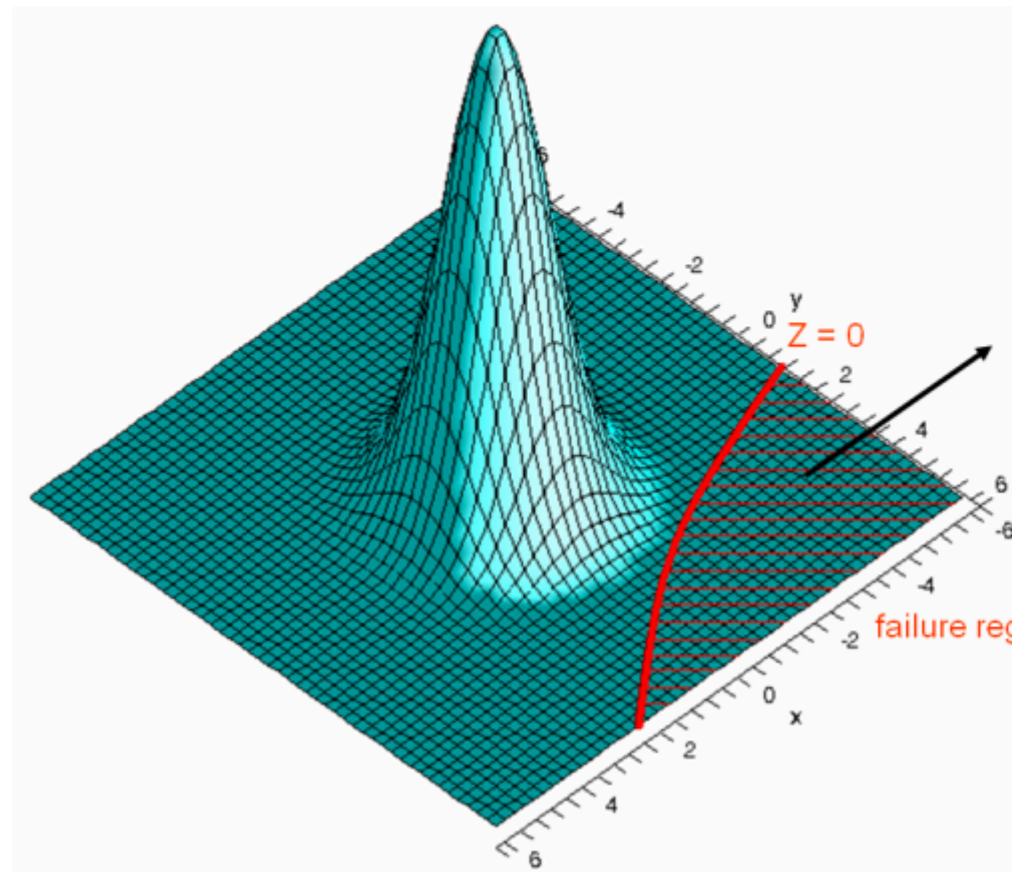
$$\int f(x)dx = 1$$

examples

Delta function



Other



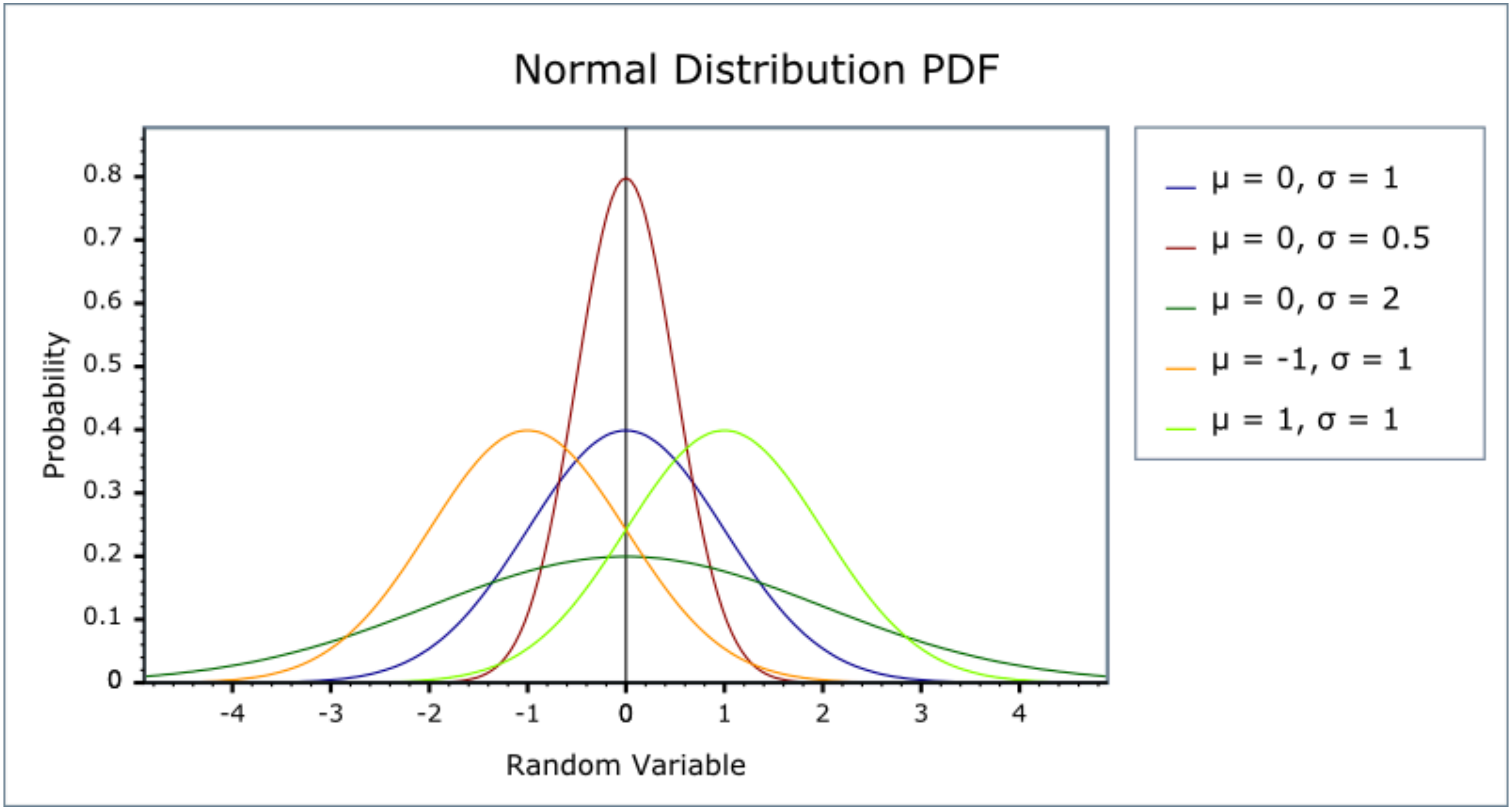
Parametric pdfs

Family of PDFs

$$G(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Described by parameters: σ,μ

examples



Described by parameters: σ, μ

PDFs and Likelihoods

PDF:

For fixed parameters, gives probability density of various possible data.

Likelihood:

For fixed data, gives relative likelihood of various parameters

Likelihood

Variation of pdf w.r.t to parameters, for fixed data

$$L(\sigma, \mu) = P(data|\sigma, \mu)$$

Note that it is not normalized

$$\int L(\sigma, \mu) \neq 1$$

Fitting

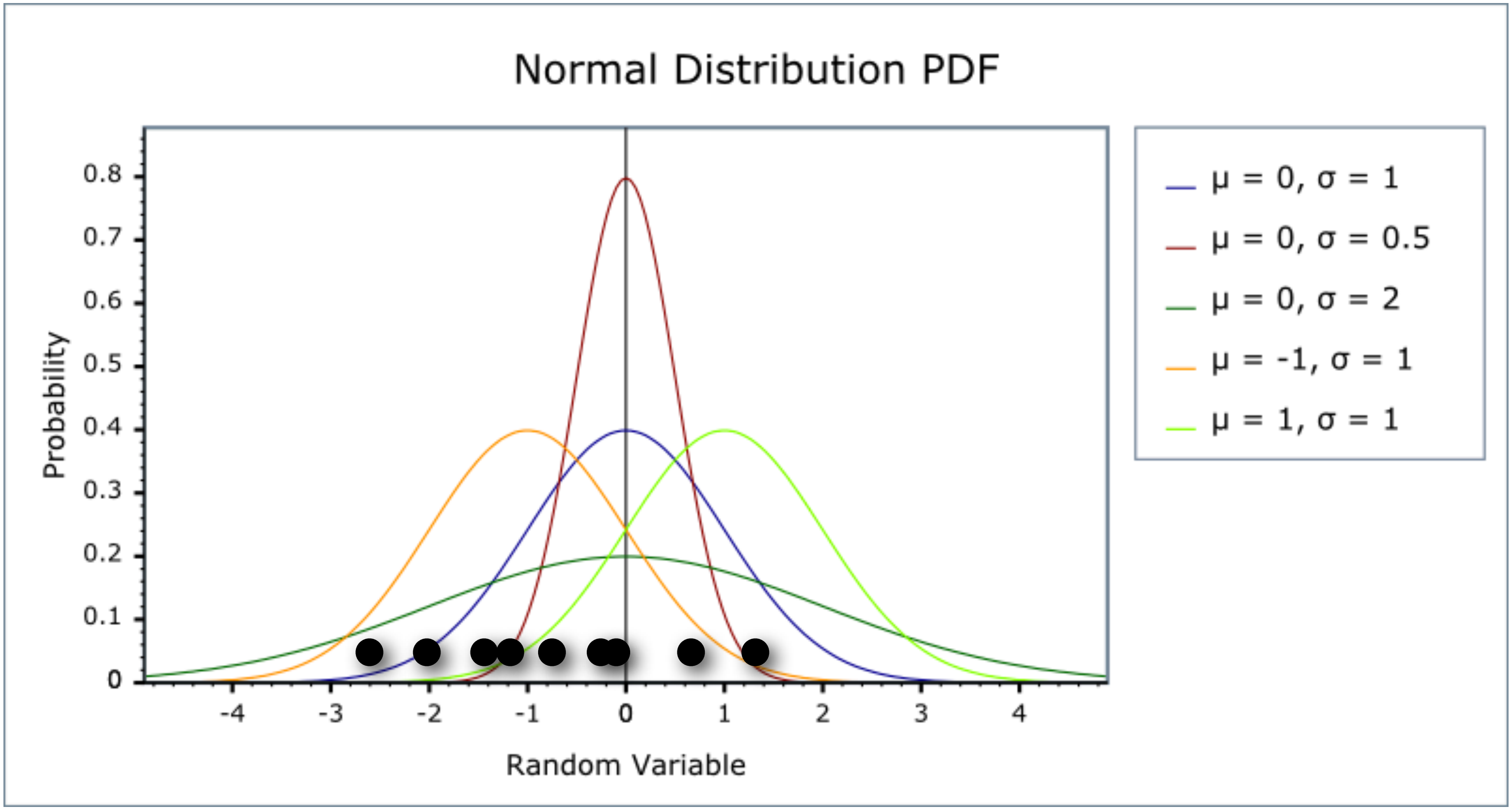
The problem

$$x = (x_1, ..., x_N)$$ Your data

$$\theta = (\theta_1, ..., \theta_n)$$ Your parameters

<u>Problem</u>: Find parameters which are most likely to have generated your data

The problem


Normal Distribution PDF
Probability
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Random Variable
μ = 0, σ = 1
μ = 0, σ = 0.5
μ = 0, σ = 2
μ = -1, σ = 1
μ = 1, σ = 1

Max Likelihood Method

$x = (x_1, ..., x_N)$ Your data

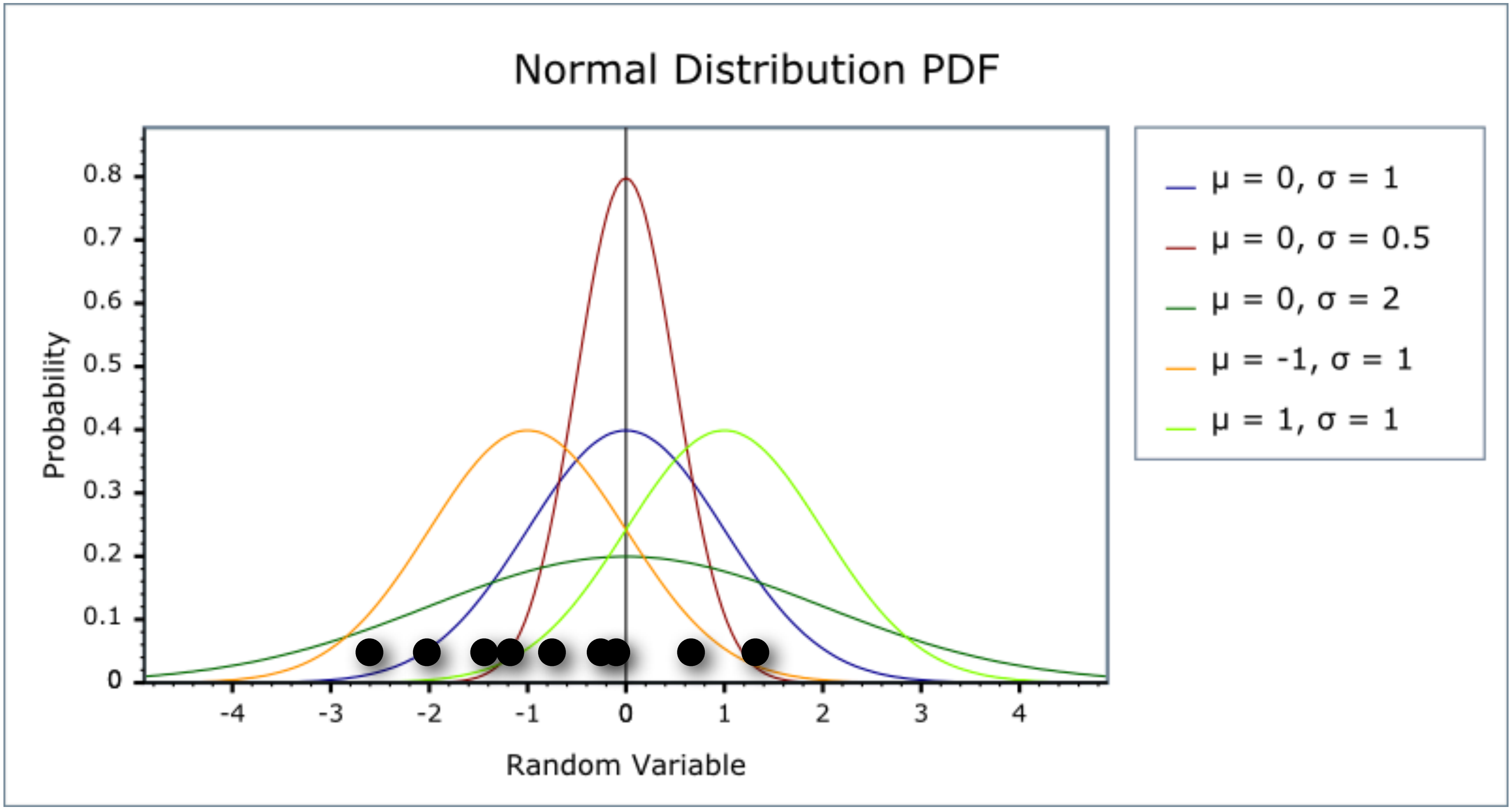
$\theta = (\theta_1, ..., \theta_n)$ Your parameters

Your likelihood

$$L(\theta) = \prod_{i=1}^{N} f(x_i; \theta)$$

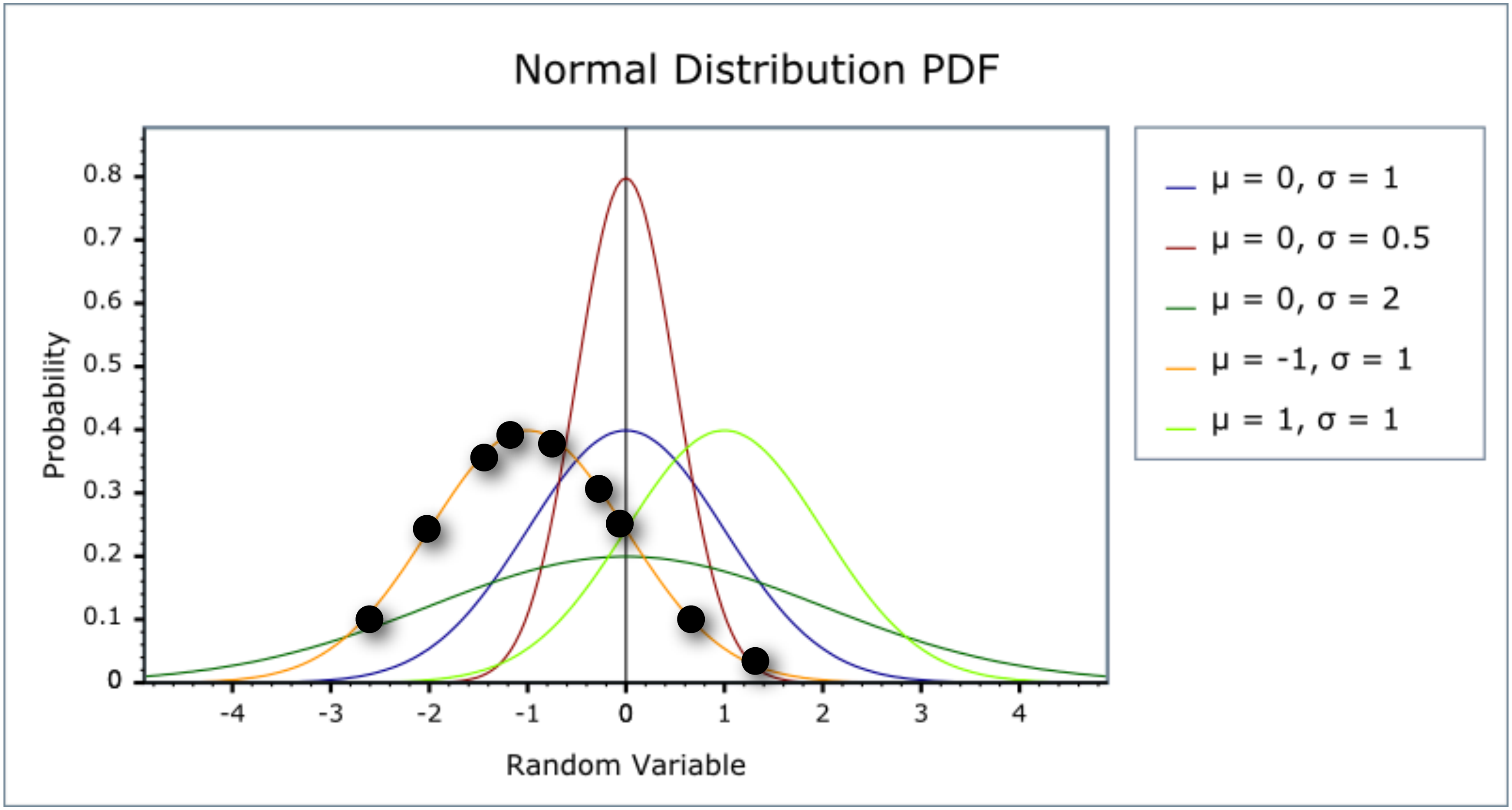
<u>Method</u>: maximize L w.r.t. parameters!

The problem



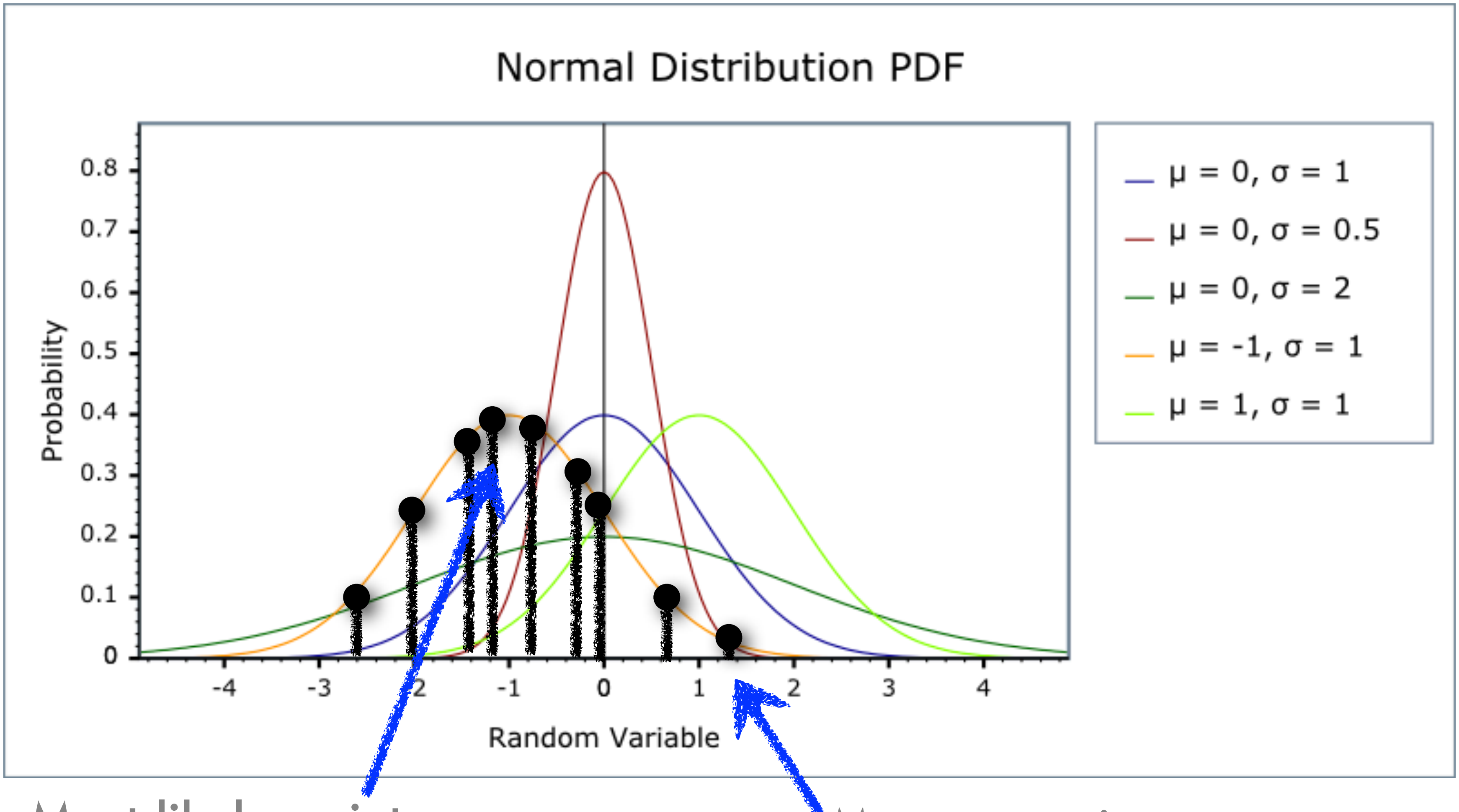
Normal Distribution PDF
Probability
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Random Variable
μ = 0, σ = 1
μ = 0, σ = 0.5
μ = 0, σ = 2
μ = -1, σ = 1
μ = 1, σ = 1

The problem



Normal Distribution PDF
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Probability
-4
-3
-2
-1
0
1
2
3
4
Random Variable
μ = 0, σ = 1
μ = 0, σ = 0.5
μ = 0, σ = 2
μ = -1, σ = 1
μ = 1, σ = 1

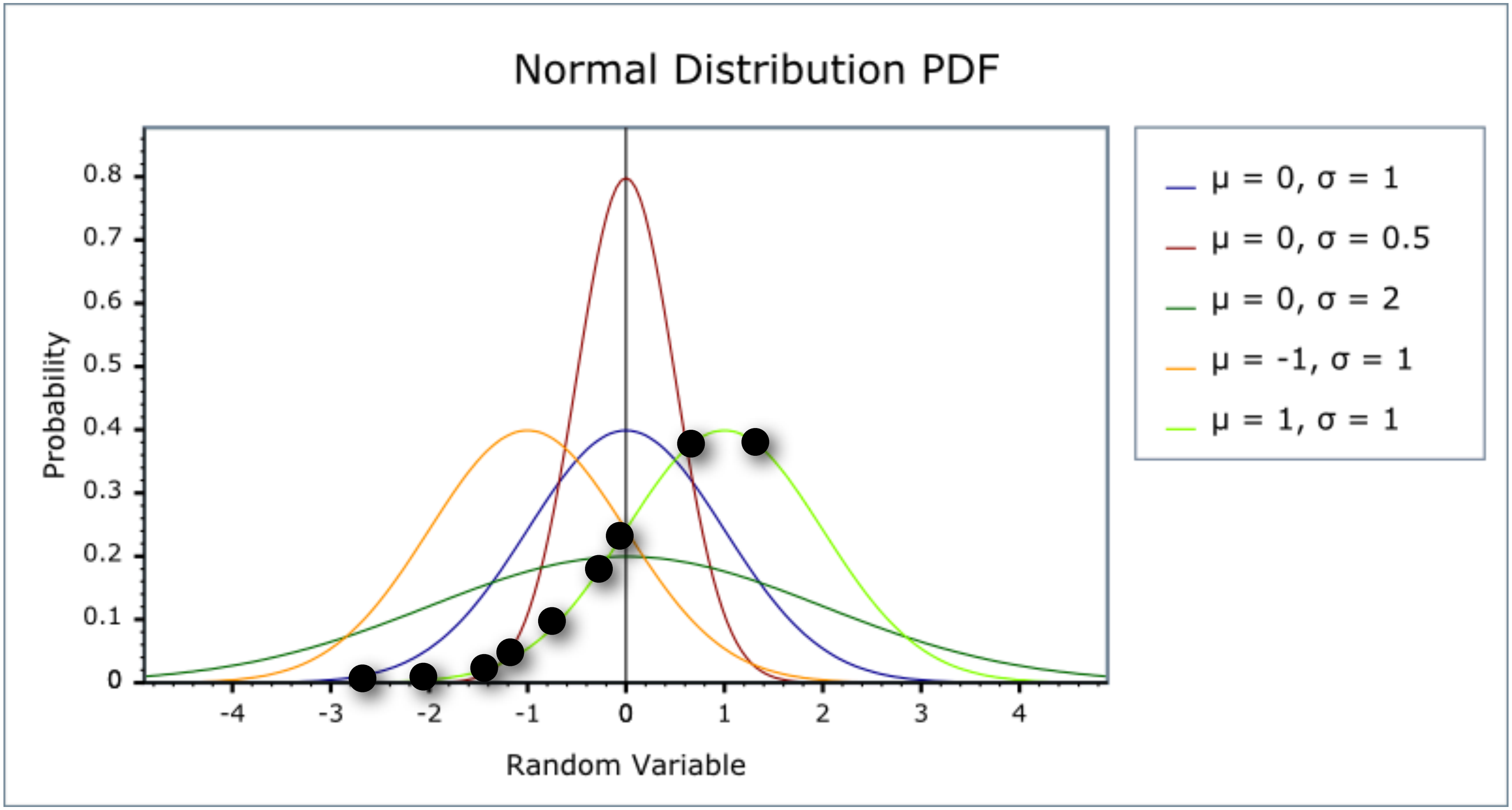
The problem



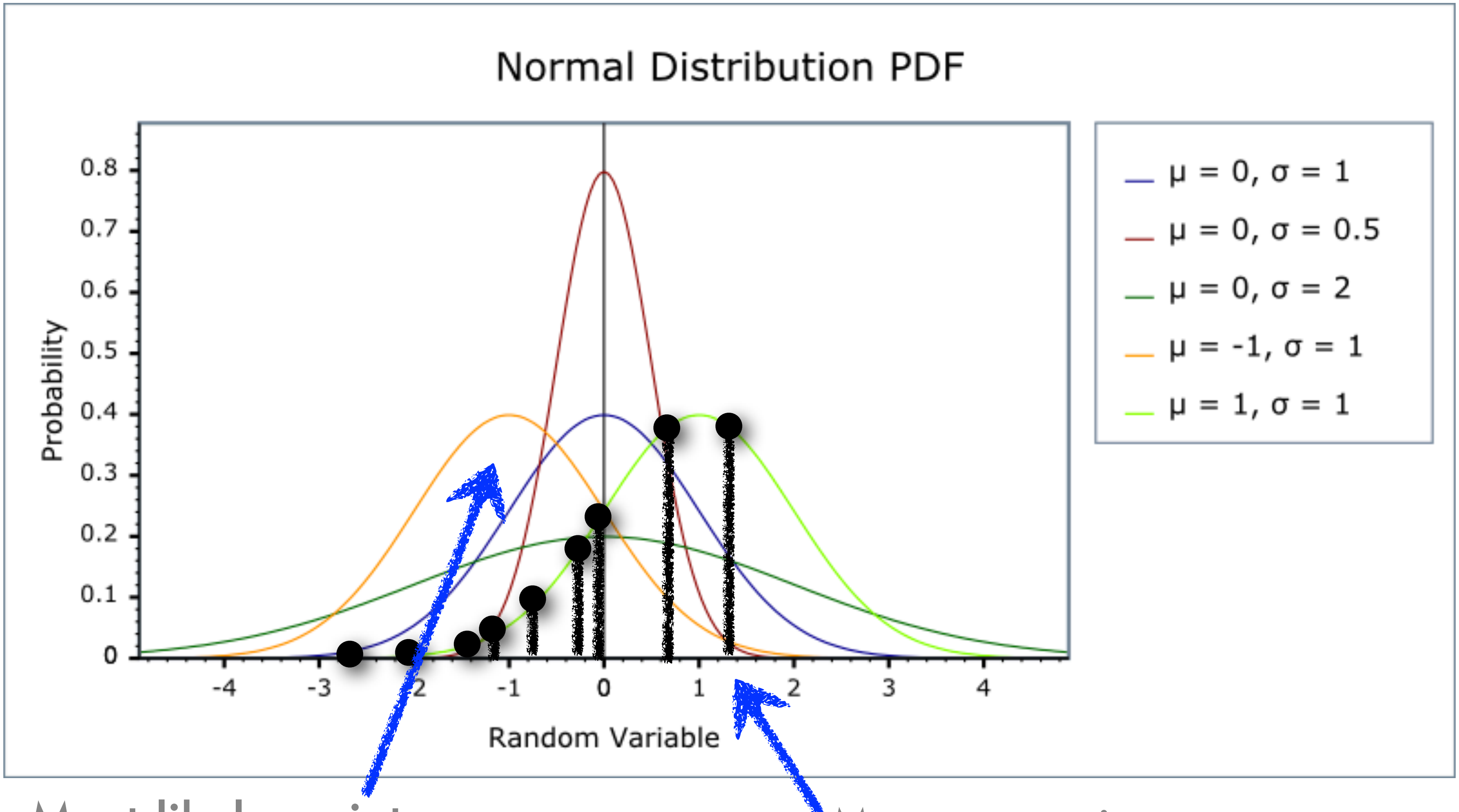
Most likely points
give large L

Most rare points
give small L

The problem


Normal Distribution PDF
Probability
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Random Variable
μ = 0, σ = 1
μ = 0, σ = 0.5
μ = 0, σ = 2
μ = -1, σ = 1
μ = 1, σ = 1

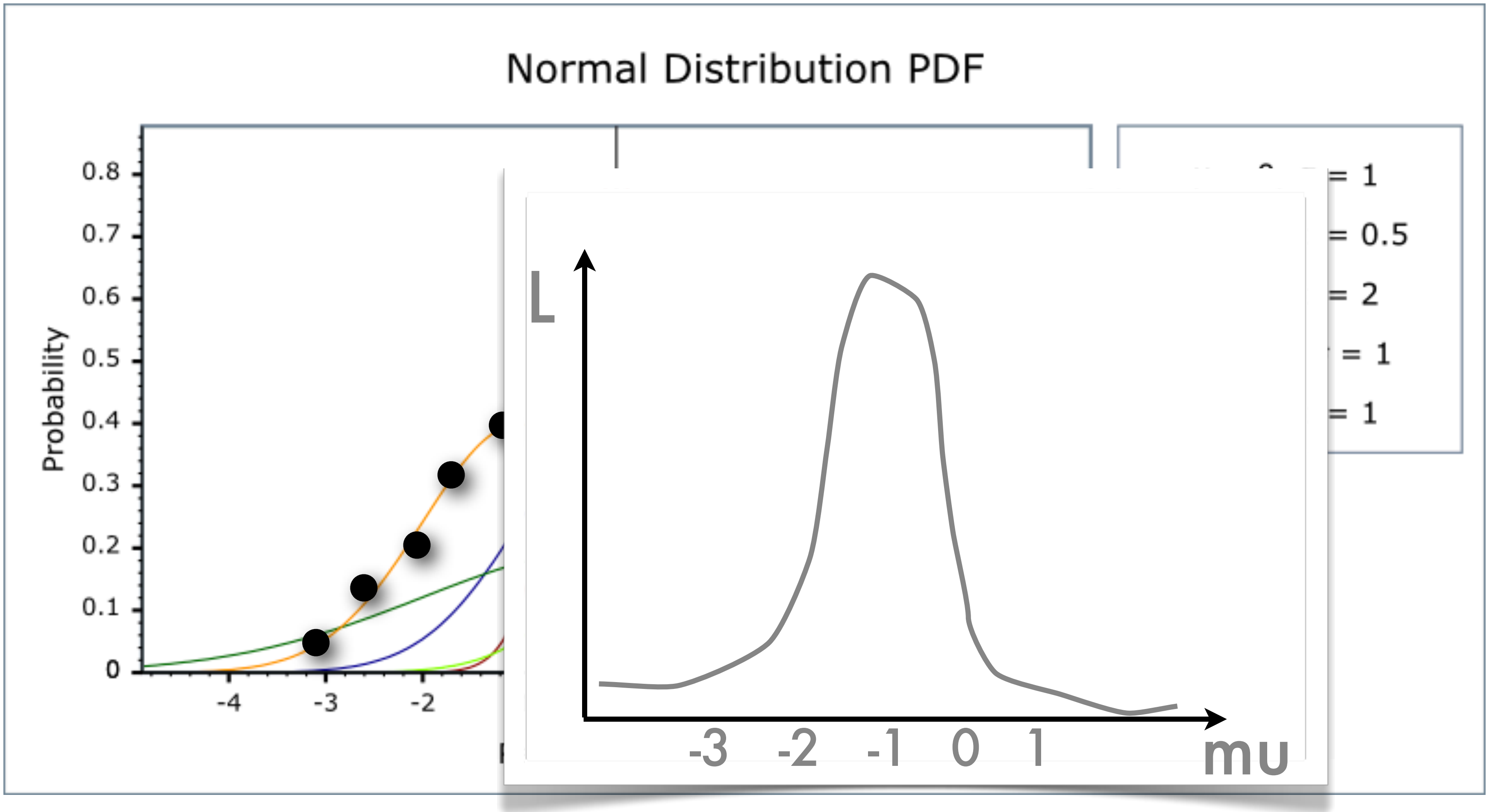
The problem



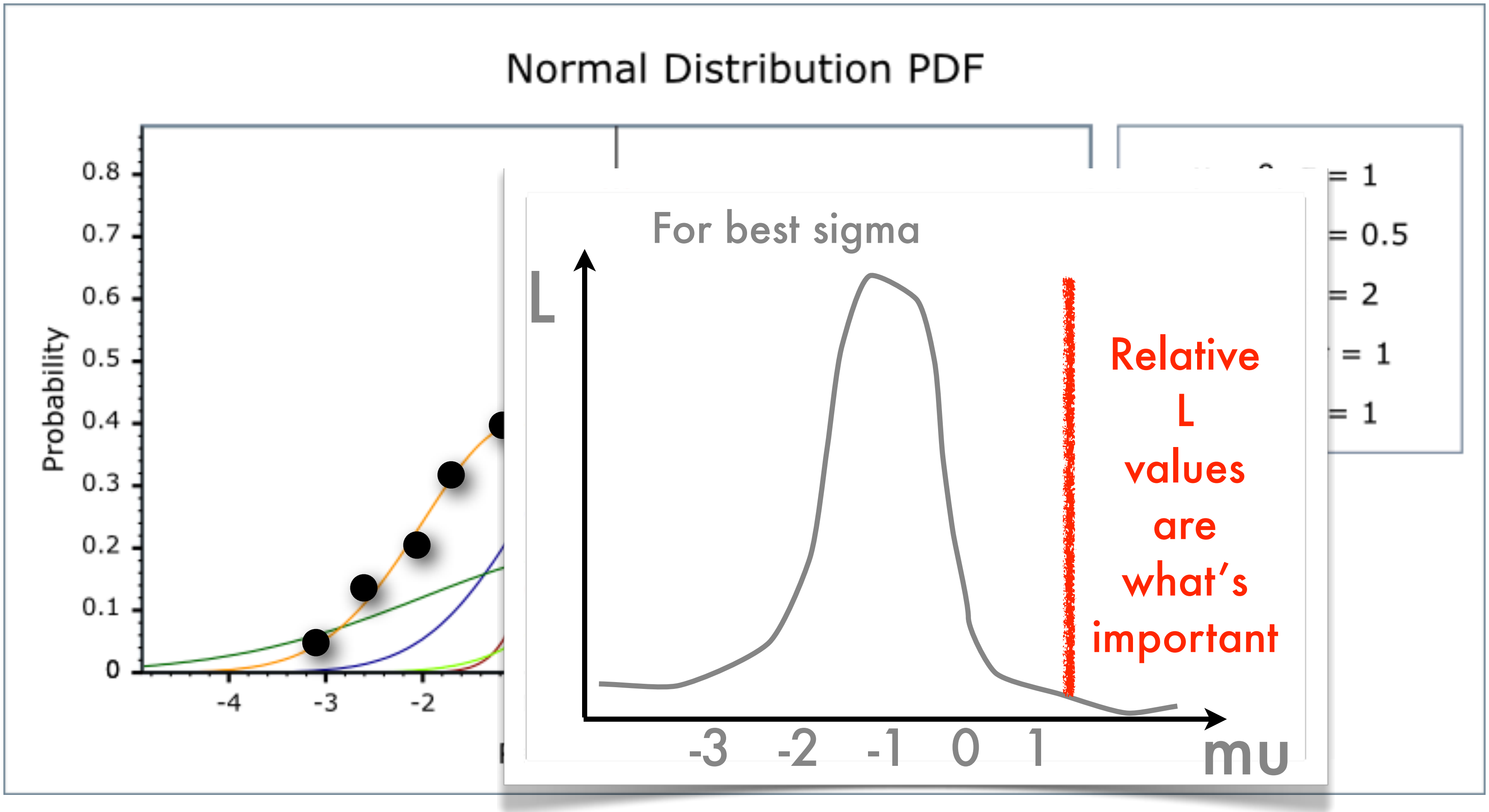
Most likely points give small L

Most rare points give large L

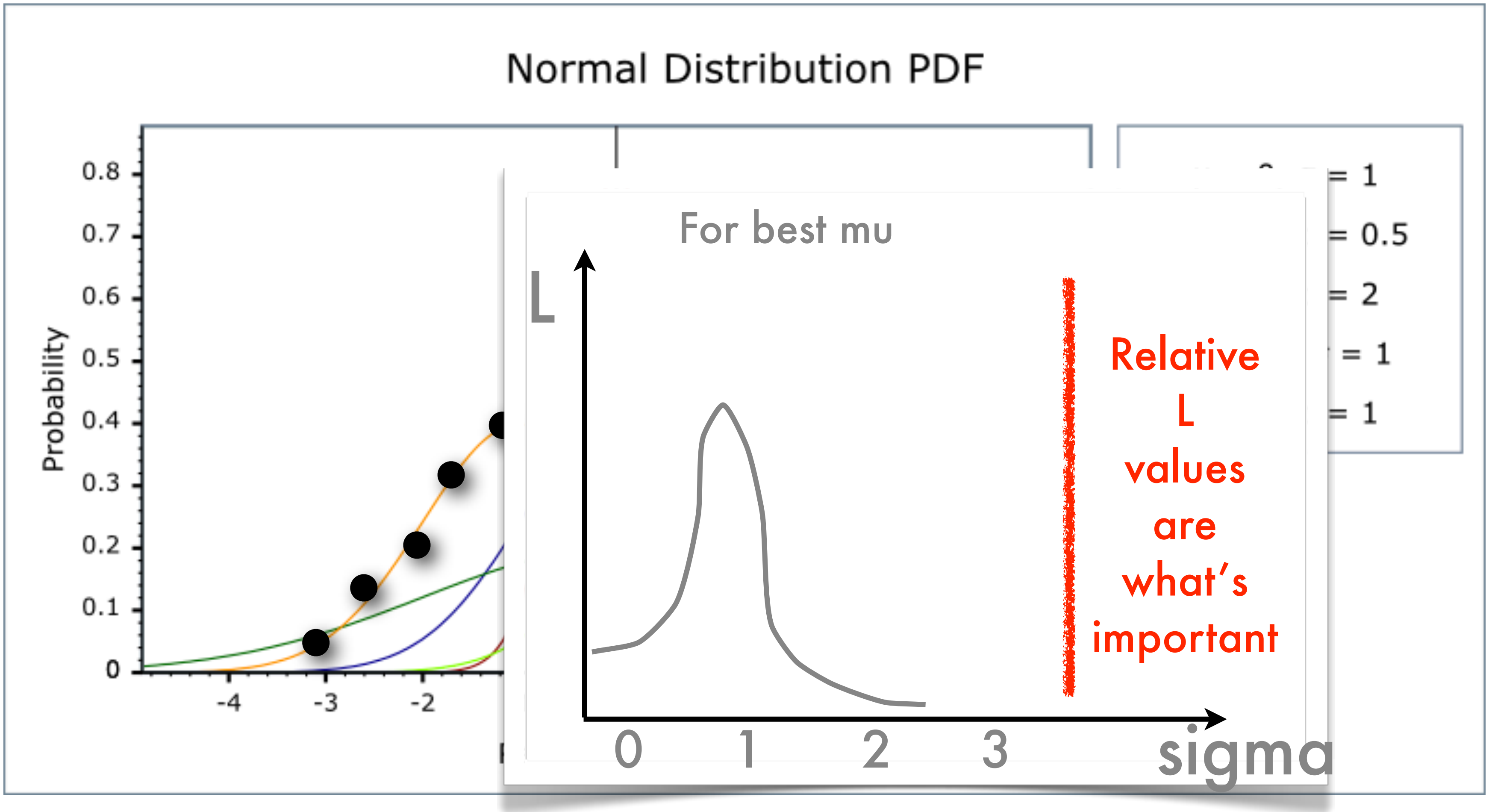
Max likelihood fitting



Max likelihood fitting

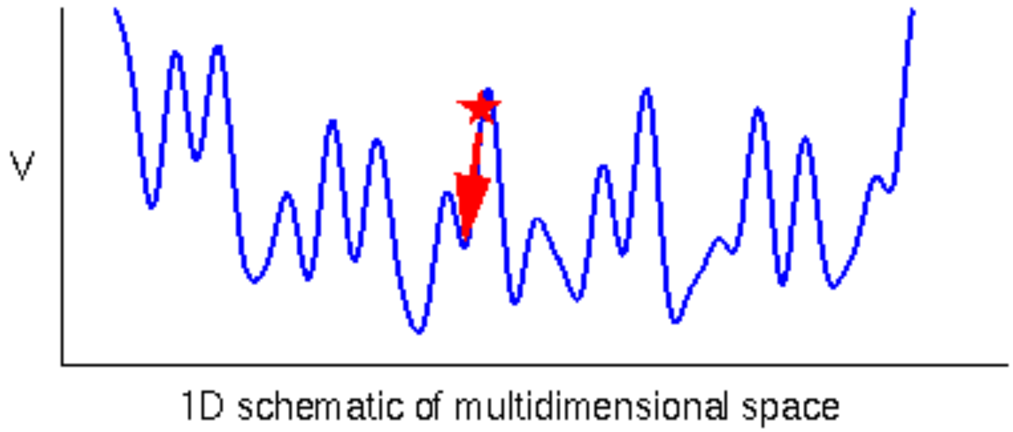


Max likelihood fitting



Minimization

Finding likelihood maximum is non-trivial



No ability to predict functional form of function, enormous space

In general, not a solved problem

Heuristic strategies need to start close to solution.

Susceptible to local minima

Nuisance parameters

$$L(\theta) = \prod_{i=1}^{N} f(x_i; \theta)$$

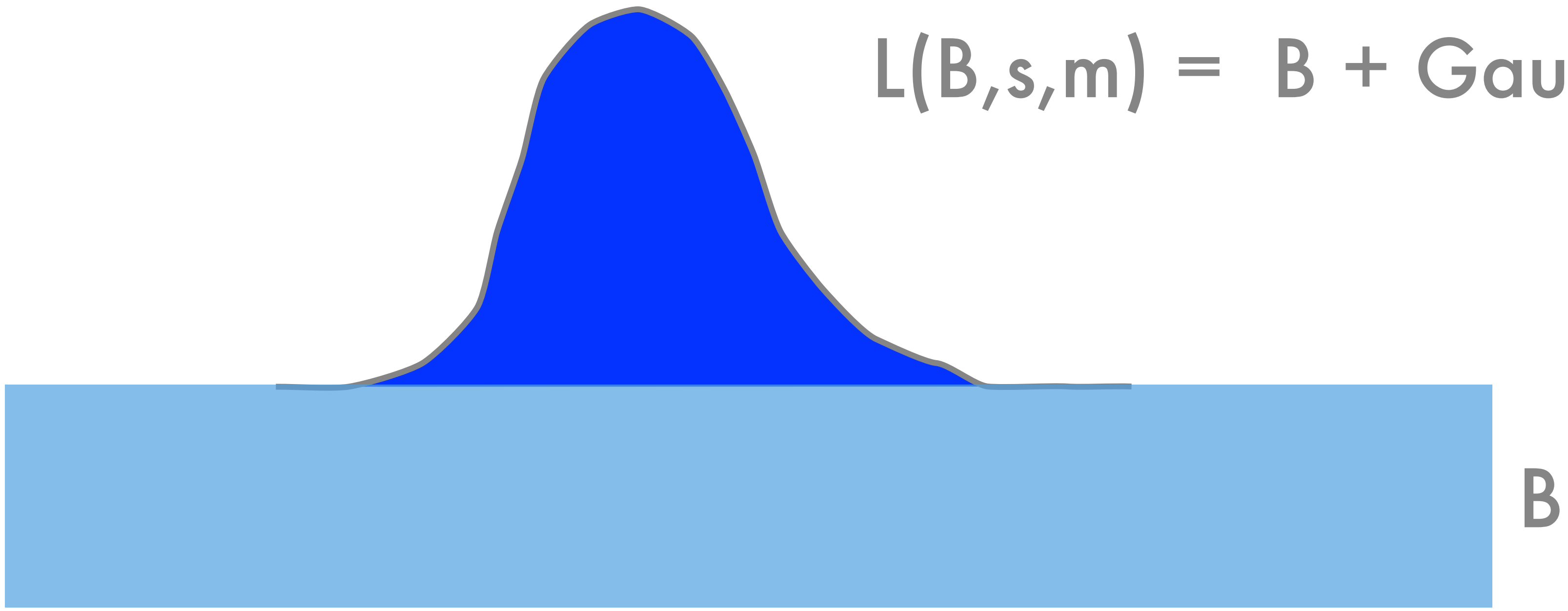
$$\theta = (\theta_1, ..., \theta_n)$$

Likelihood can have several parameters

The ones we care about: *Parameter of Interest*

The ones we don't: *nuisance parameters*

example



L(B,s,m) = B + Gauss(s,mu)

Background level under peak

One (wo)man's POI is another's NP

Binned lhood

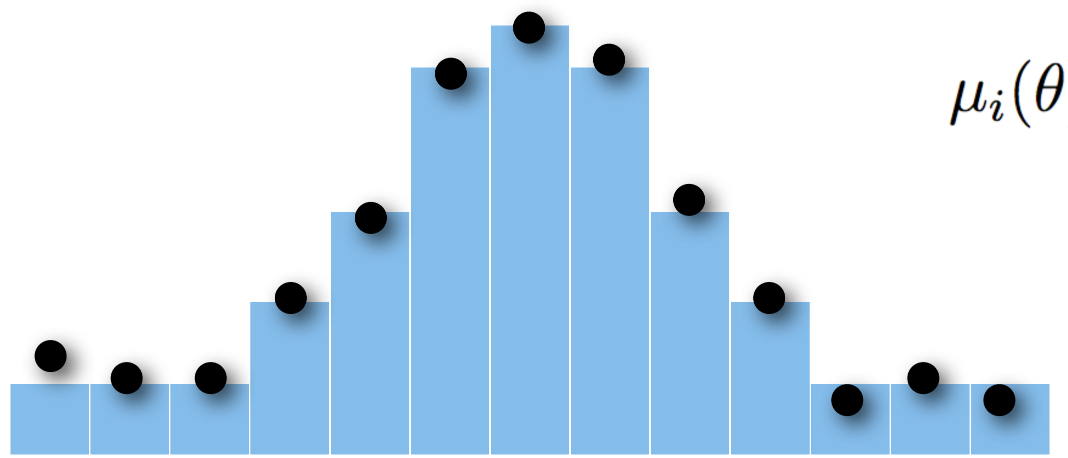
$$L(\theta) = \prod_{bini=1}^{N} Pois(n_i|\mu_i(\theta))$$

$\mu_i(\theta)$ is the predicted value in the bin

Binned likelihood

$$L(\theta) = \prod_{bin i=1}^{N} Pois(n_i|\mu_i(\theta))$$

$\mu_i(\theta)$ is the predicted value in the bin



pros:

(1) fast, no need to loop over all data.

(2) sometimes don't have unbinned PDF

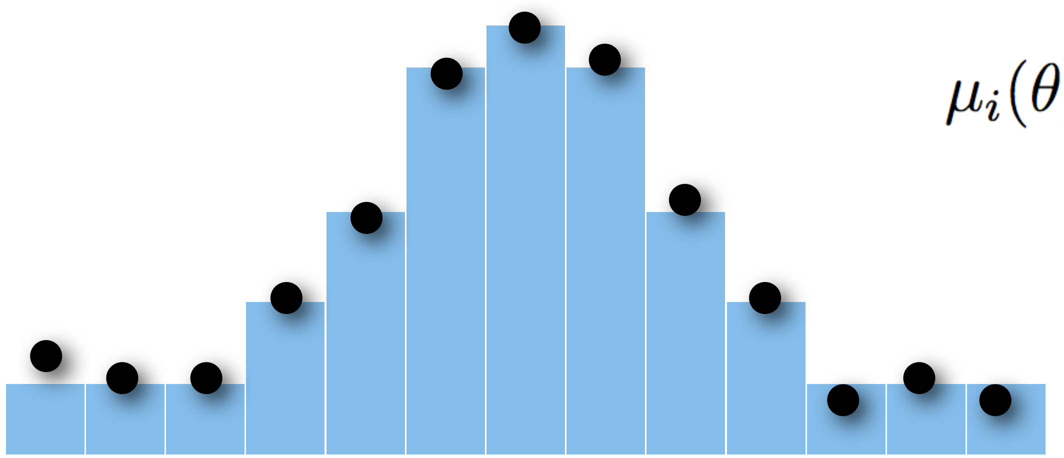
cons:

beware overly large or small bins (approaches unbinned as bin size →0)

Binned likelihood

$$L(\theta) = \prod_{bin i=1}^{N} Pois(n_i|\mu_i(\theta))$$

$\mu_i(\theta)$ is the predicted value in the bin

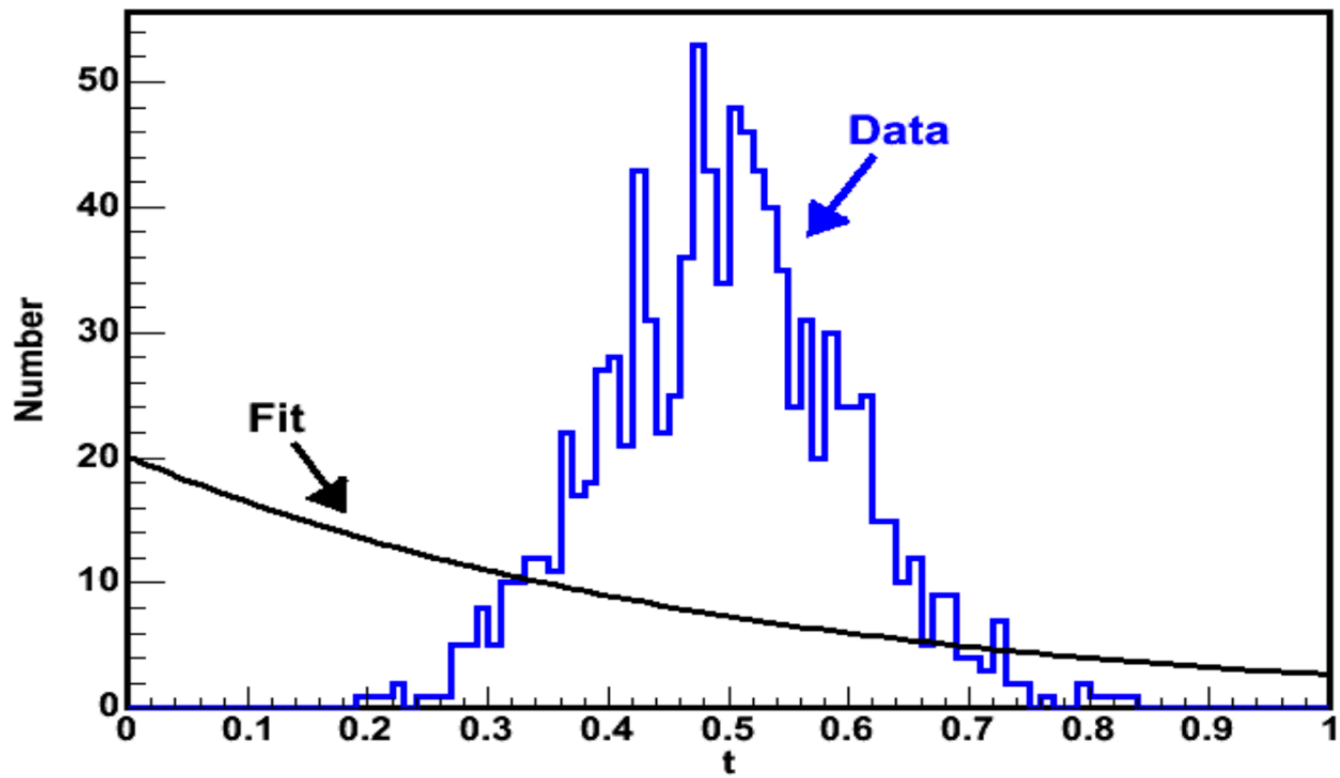


<u>how to choose binning?</u>

approach experimental resolution

ensure all bins have valid predicted value

Goodness of fit



Likelihood will find best fit

but will **not** tell you how **well** it is fit

Chi-squared fitting

data

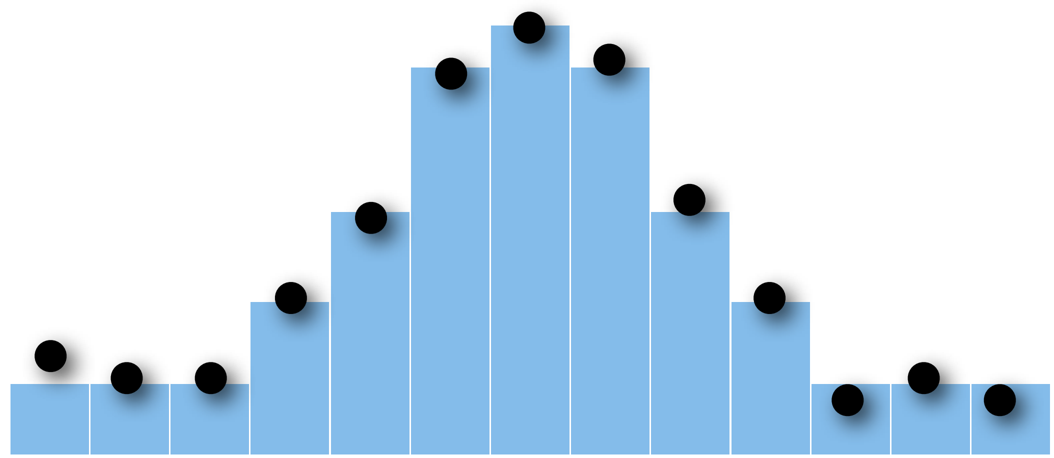
Prediction

$$\chi^2(\theta) = \sum_{i}^{N} \frac{y_i - F(x_i, \theta)}{\sigma_i^2}$$

uncertainty

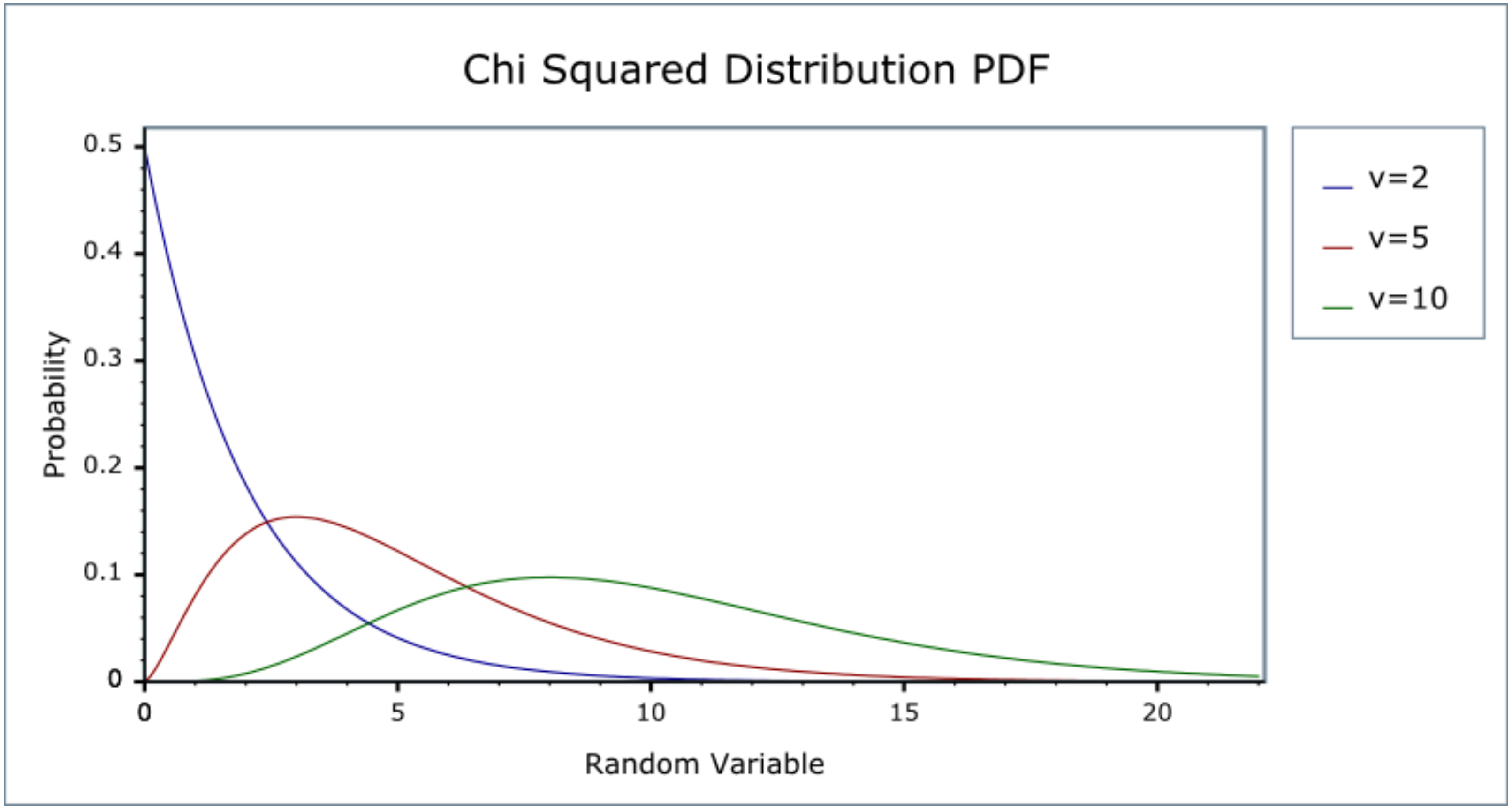
example

$$\chi^2(\theta) = \sum_i^N \frac{y_i - F(x_i, \theta)}{\sigma_i^2}$$



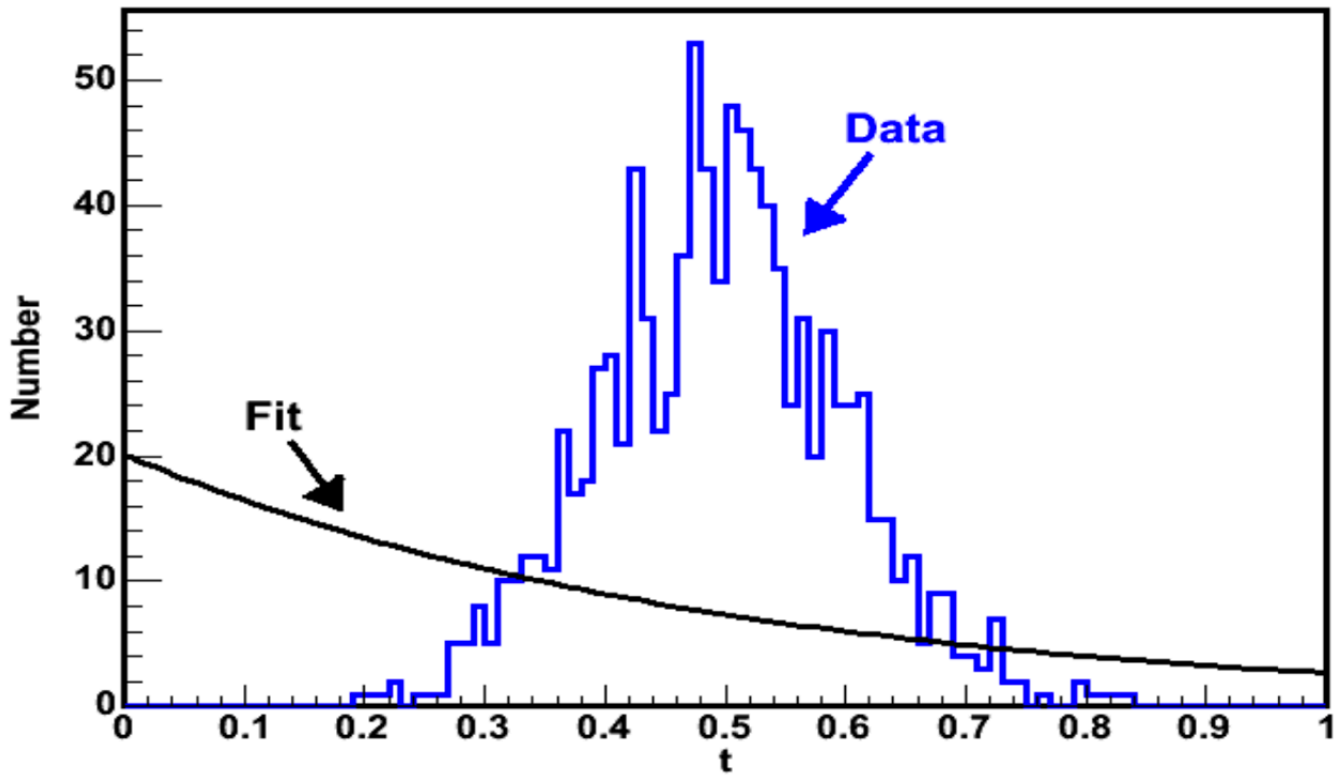
Difference between predicted and observed values

PDF of chi-squared



You can tell the quality of the fit.

goodness of fit



chi-squared value will be very large
chi-squared prob will be very small

Mendel's data

http://nih.gov/about/director/ebiomed/mendel.htm

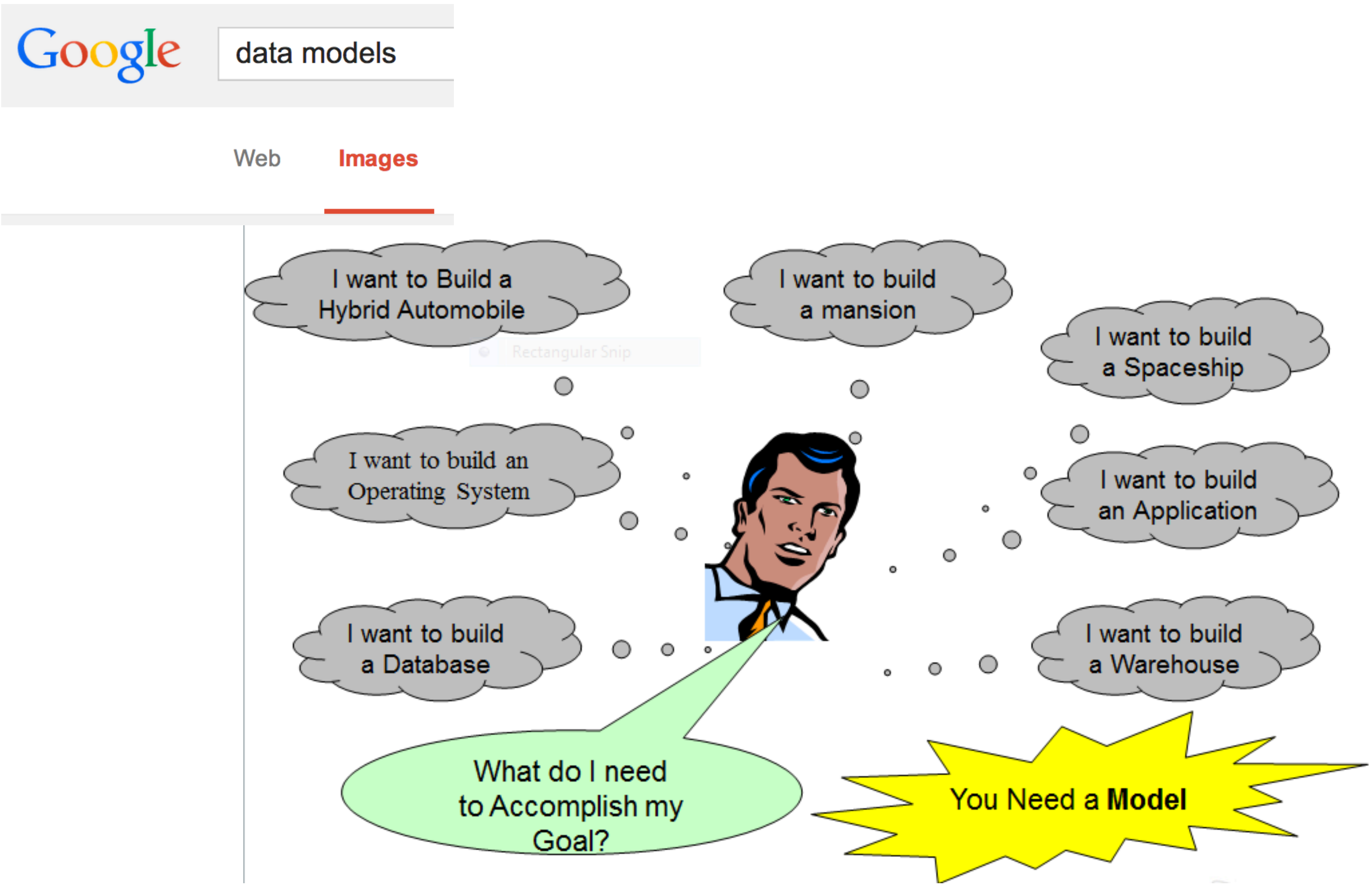
1. A cursory look at Mendel's various observations soon makes a statistically literate person notice that they come, over and over again, uncomfortably close to Mendel's expectations. As Edwards put it, "one can applaud the lucky gambler; but when he is lucky again tomorrow, and the next day, and the following day, one is entitled to become a little suspicious" [5]. The precise calculations are still under dispute, but the best current estimate suggests that results as close as or closer to expectations as the ones reported by Mendel would occur in only 1 out of 33,000 replications [6, p. 921]. In other words, it is virtually inconceivable that Mendel obtained his "good" results by pure chance.

http://www.genetics.org/content/175/3/975.full.pdf+html

Altogether, the experiments yielded 399 parents classified as heterozygous and 201 parents classified as homozygous. Fisher noted that the expected values from Equation 1 are 377.5 and 222.5, respectively. A chi-square test yields the test statistic 3.31, which has an associated *P*-value of 0.069. This does not differ significantly from Fisher's expectation; nevertheless, it fanned Fisher's suspicion because he writes, "a deviation as fortunate as Mendel's is to be expected once in twenty-nine trials" (Fisher 1936, pp. 125–126).

Data models

Data models


data models
Web
Images
I want to Build a Hybrid Automobile
I want to build a mansion
I want to build a Spaceship
I want to build an Operating System
I want to build an Application
I want to build a Database
I want to build a Warehouse
What do I need to Accomplish my Goal?
You Need a **Model**

Models & Statistics

Physics:

our model of the expected results of the experiment

f(data | theory)

Provides:

- PDF for data as a function of POI, NPs
- generate pseudo-data
- fix data to get lhood

Models & Statistics

Physics:

our model of the expected results of the experiment

f(data | theory)

Provides:

- PDF for data as a function of POI, NPs
- generate pseudo-data
- fix data to get lhood

Statistics:

use model and data to make statistical statements about POIs:

- parameter estimates
- hypothesis testing
- confidence interval

Models & Statistics

Physics:

our ... ected

resu ... ent

f(

Prov

- PD ...

function of POI, NPs

- generate pseudo-data
- fix data to get lhood

Approaches:

MC-driven

fast-MC driven

Data-driven

Effective models

Statistics:

use m

to ma

stater ... Is:

- para ... es
- hypo ...
- confidence interval

Approaches:

Frequentist

Bayesian

F-B Hybrid

Models & Statistics

Physics:

our ... ected

resu ... ent

f(

Approaches:

MC-driven

fast-MC driven

Data-driven

Effective models

Prov

- PD

function of POI, NPs

- generate pseudo-data
- fix data to get lhood

Statistics:

use m

to mo

state ... OIs:

- para ... es
- hyp
- confidence interval

Approaches:

Frequentist

Bayesian

F-B Hybrid

Models & Statistics

Physics:

our ... ected

resu ... ent

f(

Prov

- PD...

function of POI, NPs

- generate pseudo-data
- fix data to get lhood

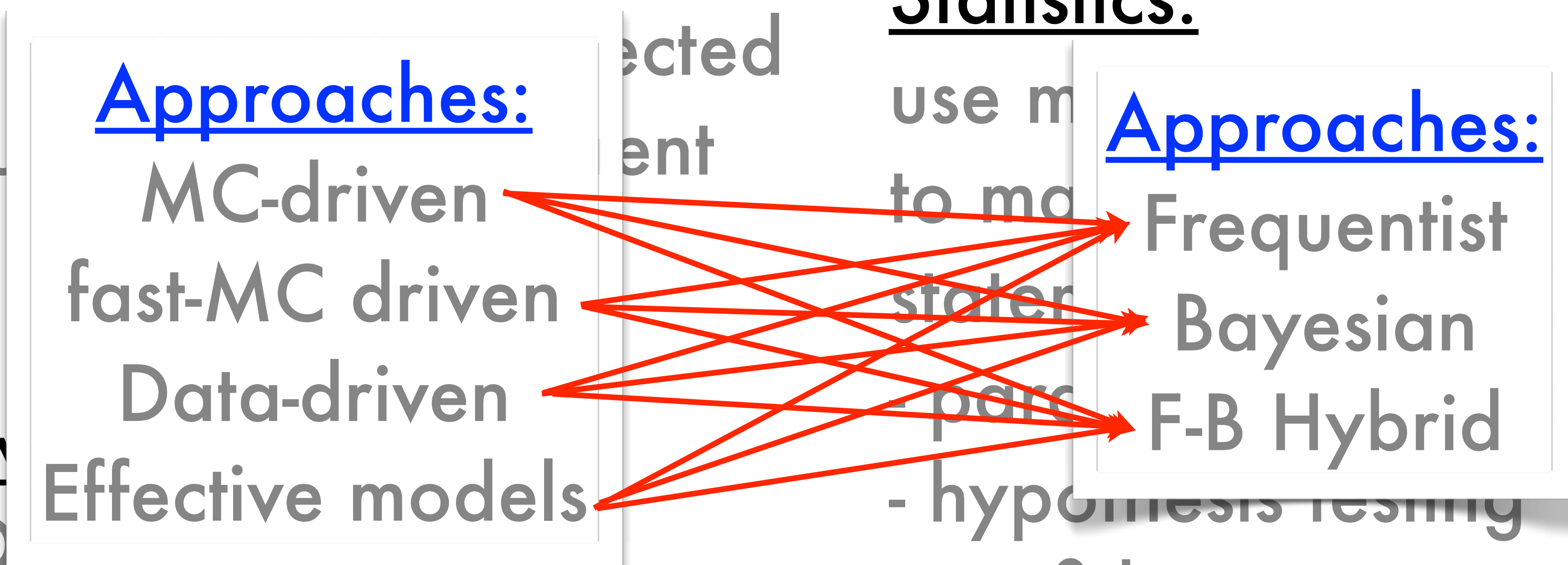
Statistics:

use m

to mo

stater ... OIs:

- para ... es
- hypo...
- confidence interval



Upshot

Model building is distinct from stat interpretation

Note: some stats packages have
implied model choices
(eg MC limit uses histograms, so no unbinned PDFs)

Quality of your **result** comes
from the quality of the **model**

This idea: K. Cranmer (I *think*)

Models

Full MC
Fast MC
Effective models
Data-driven models

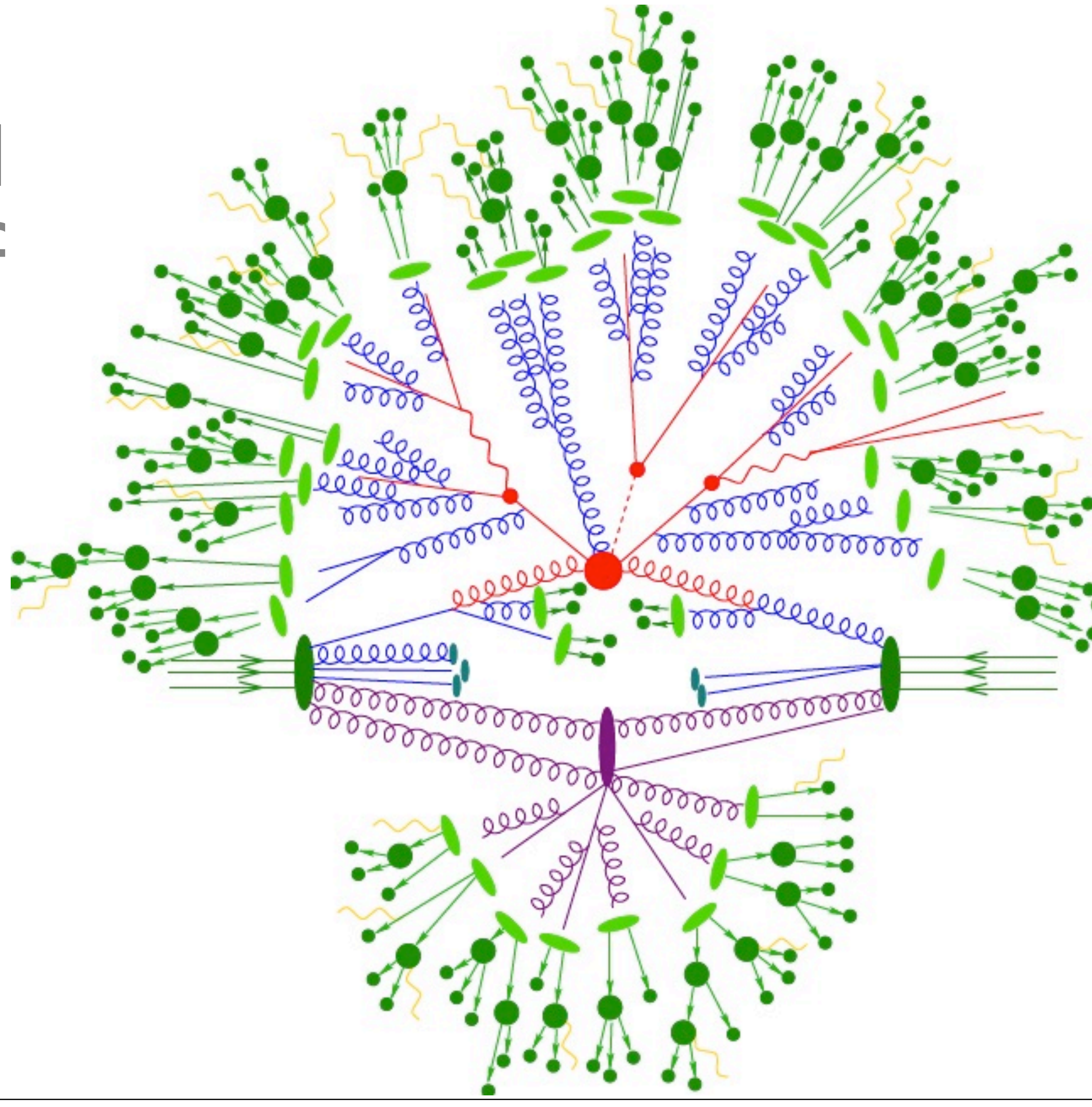
Full MC Models

Full MC model

We have a good understanding of of the pieces

Do we have

f(data|theory)?

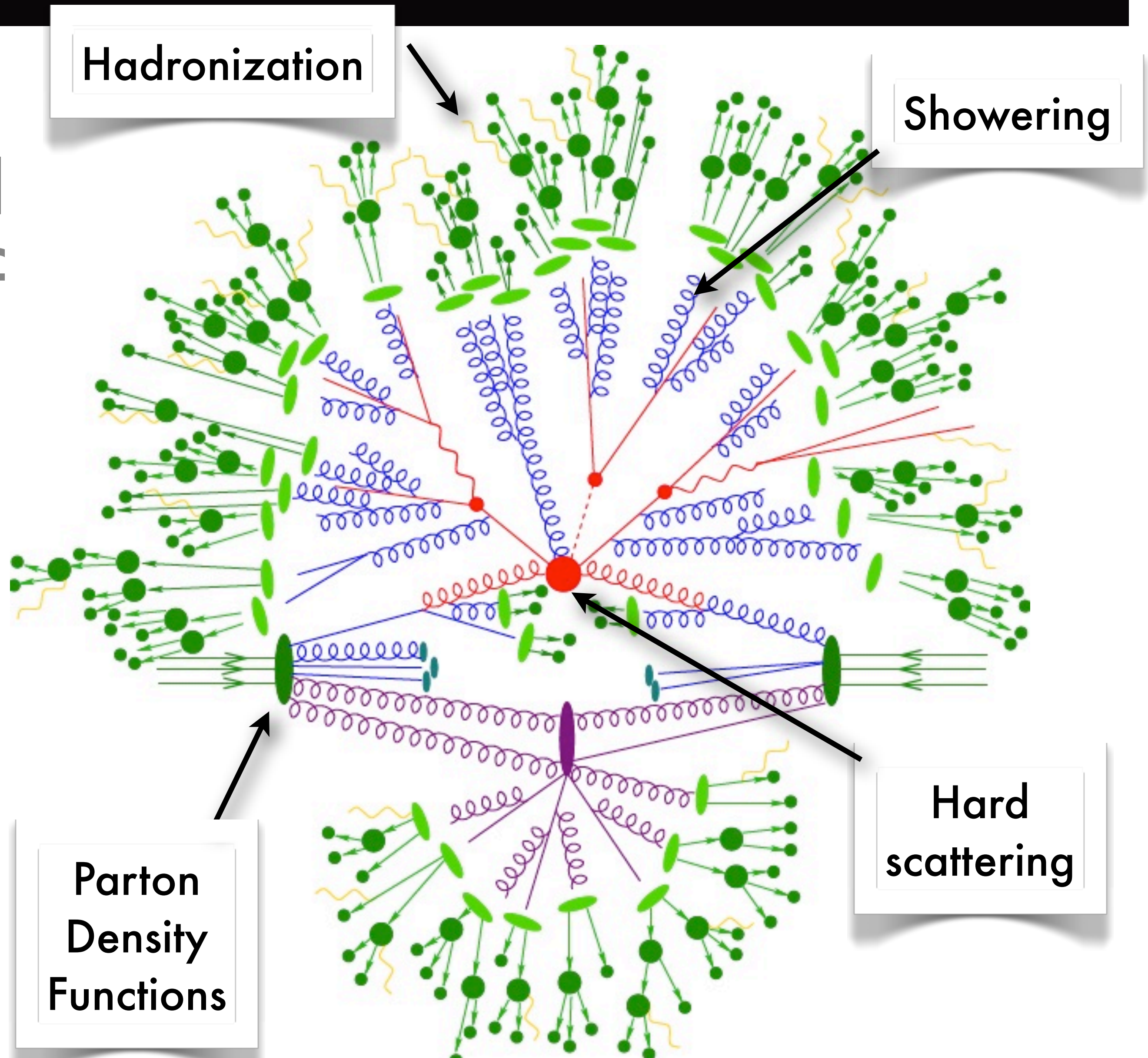


Full MC model

We have a good understanding of of the pieces

Do we have

f(data|theory)?



The dream

What would

f(data|theory)

look like?

The dream

f(data | final-state particles P)

x f(final state particles P | showered particles S)

x f(showered particles S | hard scatter products M)

x f(hard scatter products M | theory)

Sum over all possible intermediate P,S,M

The dream

Detector Response

f(data | final-state particles P)

Hadronization

x f(final state particles P | showered particles S)

Showering

x f(showered particles S | hard scatter products M)

x f(hard scatter products M | theory)

Parton Density Functions

Hard scattering

Sum over all e intermediate P,S,M

The dream

f(hard scatter products M | theory)

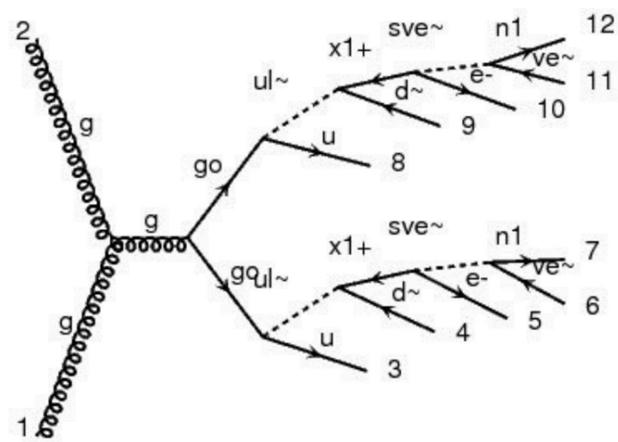
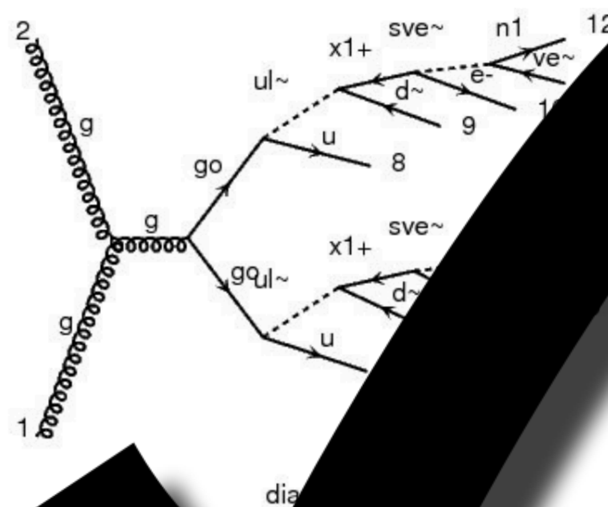


diagram 1

Theory well defined
automatic calculators exist
for almost any (B)SM theory

The dream

f(hard scatter products M | theory)



The ... defined

automati... ulators exist

for almost any (B)SM theory

The nightmare

f(data | final-state particles P)

x f(final state particles P | showered particles S)

x f(showered particles S | hard scatter products M)

We have: solid understanding of microphysics

We need: analytic description of high-level physics

[redacted]

The solution

<u>We have</u>: solid understanding of microphysics
<u>We need</u>: analytic description of high-level physics
<u>But</u>: only heuristic lower-level approaches exist

Iterative simulation strategy, no overall PDF

<u>Iterative approach</u>
(1) Draw events from f(M|theory)
(2) add random showers
(3) do hadronization
(4) simulate detector

The solution

<u>We have</u>: solid understanding of microphysics
<u>We need</u>: analytic description of high-level physics
<u>But</u>: only heuristic lower-level approaches exist

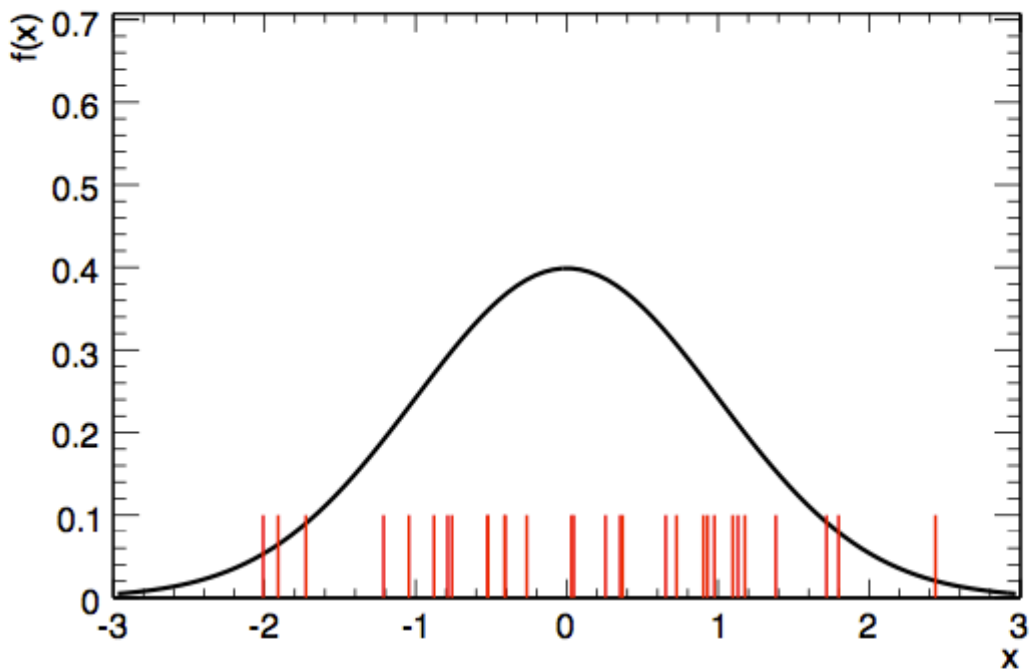
Iterative simulation strategy, no overall PDF

<u>What do we get</u>
Arbitrarily large samples of events drawn from f(data|theory), but not the PDF itself

The problem

Don't know PDF, have events drawn from PDF

$$f_{emp} = \frac{1}{N}\sum_{i}^{N}\delta(x - x_i)$$



(K. Cranmer)

Need to recreate PDF

What do we need?

Want:

our model of the expected results of the experiment

f(data | theory)

Provides:

- PDF for data as a function of POI, NPs
- generate pseudo-data
- fix data to get lhood

We have:

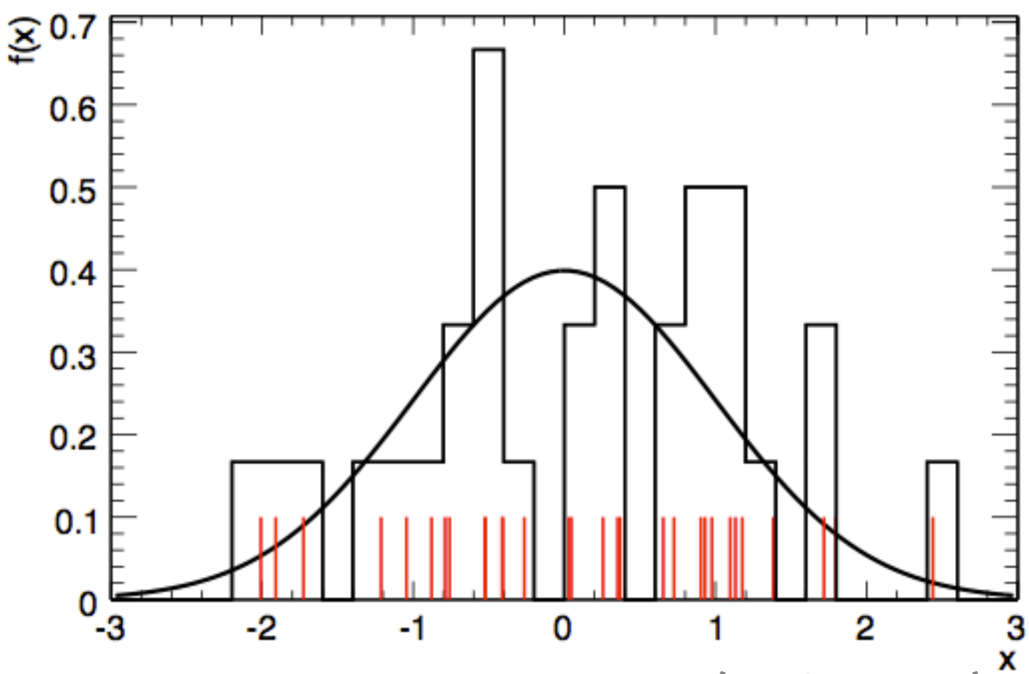
A tool that can generate sample event data

How do we use that to build our PDF?

MC events to PDF

Approach 1: histogram

$$f_{hist}^{w,s}(x) = \frac{1}{N}\sum_i h_i^{w,s}$$



(K. Cranmer)

Curse of Dimensionality

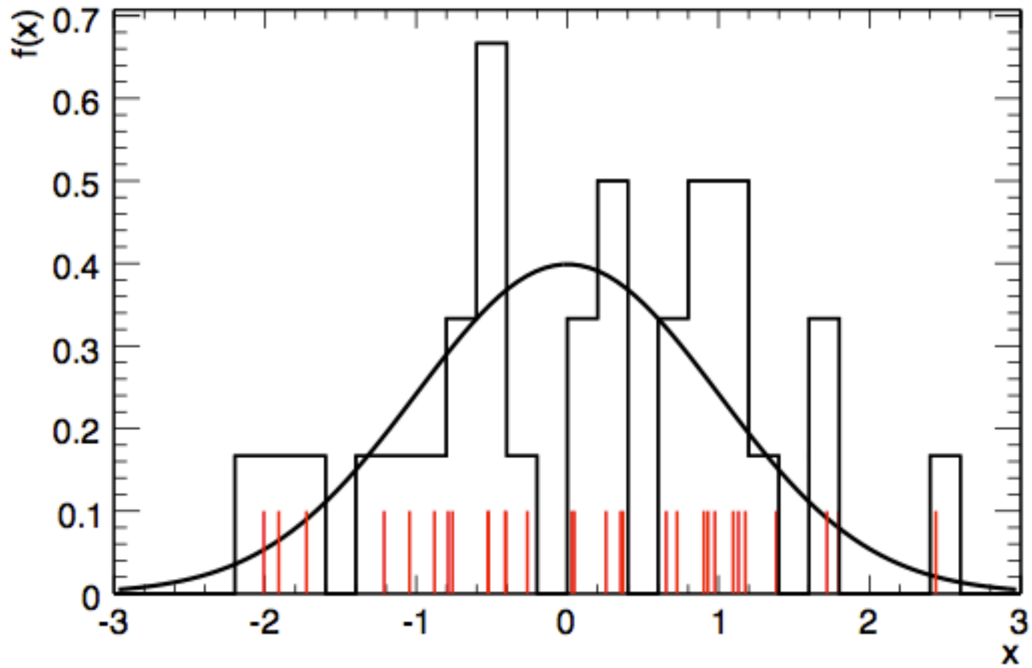
How many events do you need to describe a 1D distribution? O(100)

An n-D distribution?

$O(100^n)$

!!

$$f^{w,s}_{hist}(x) = \frac{1}{N}\sum_i h_i^{w,s}$$



(K. Cranmer)

The nightmare

f(data | final-state particles P)

x f(final state particles P | showered particles S)

x f(showered particles S | hard scatter products M)

"data" is a 100M-d vector!

The nightmare

f(data|final-state particles P)

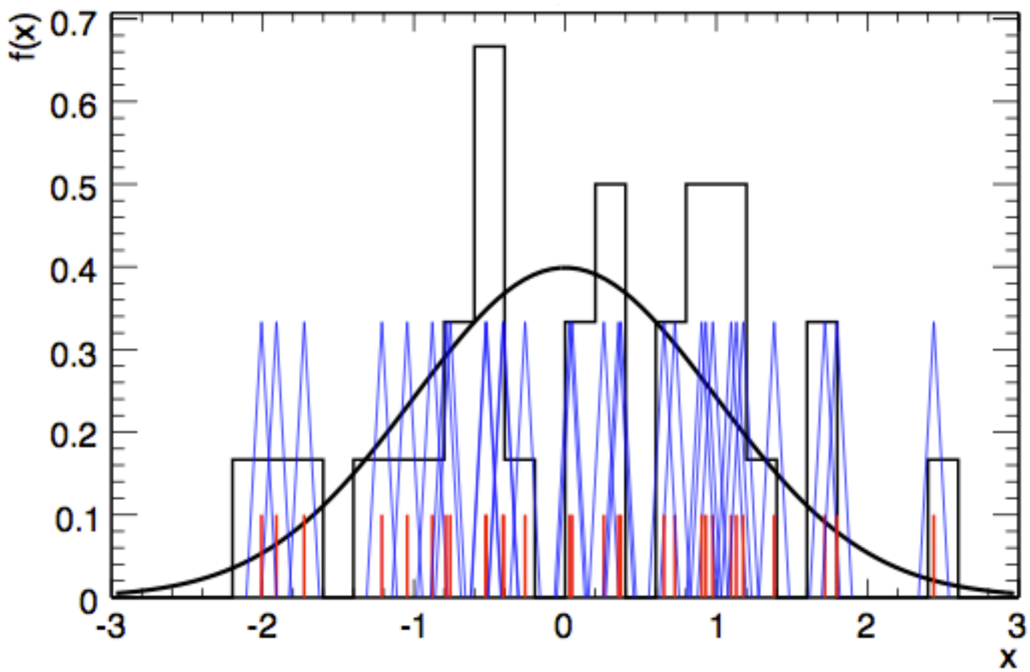
x f(final stat ... red particles S)

x f(showered ... atter products M)

“ ... vector!

MC events to PDF

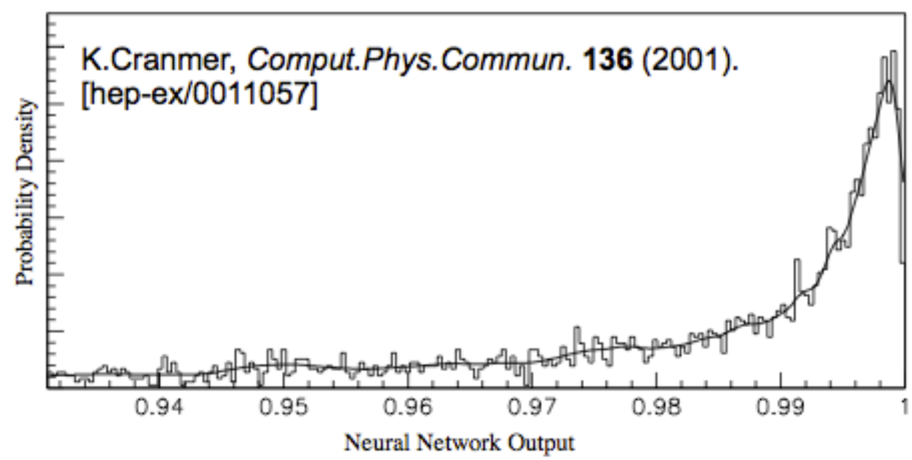
Approach 2: probability density estimates



(K. Cranmer)

Prob Density Estimate

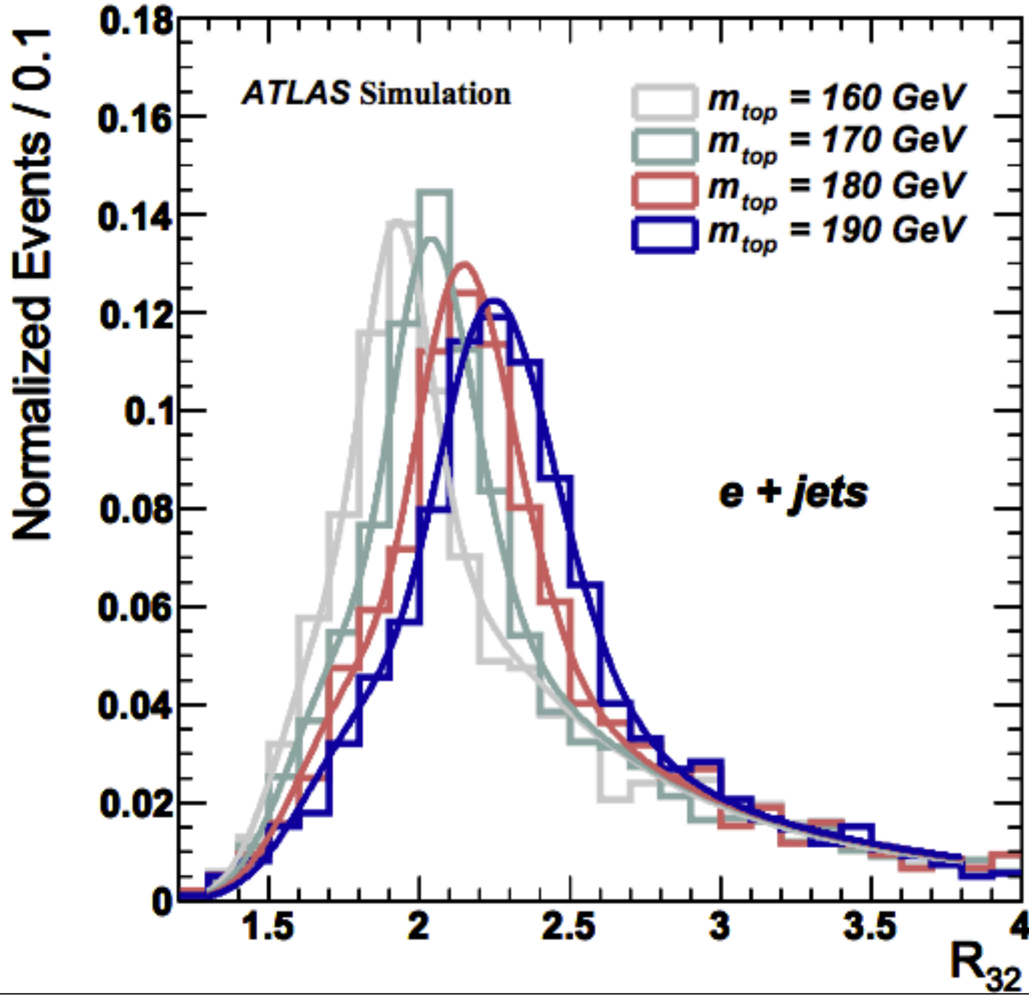
Approach 2: probability density estimates



More effective use of events,
require fewer events to make smooth prediction

MC events to PDF

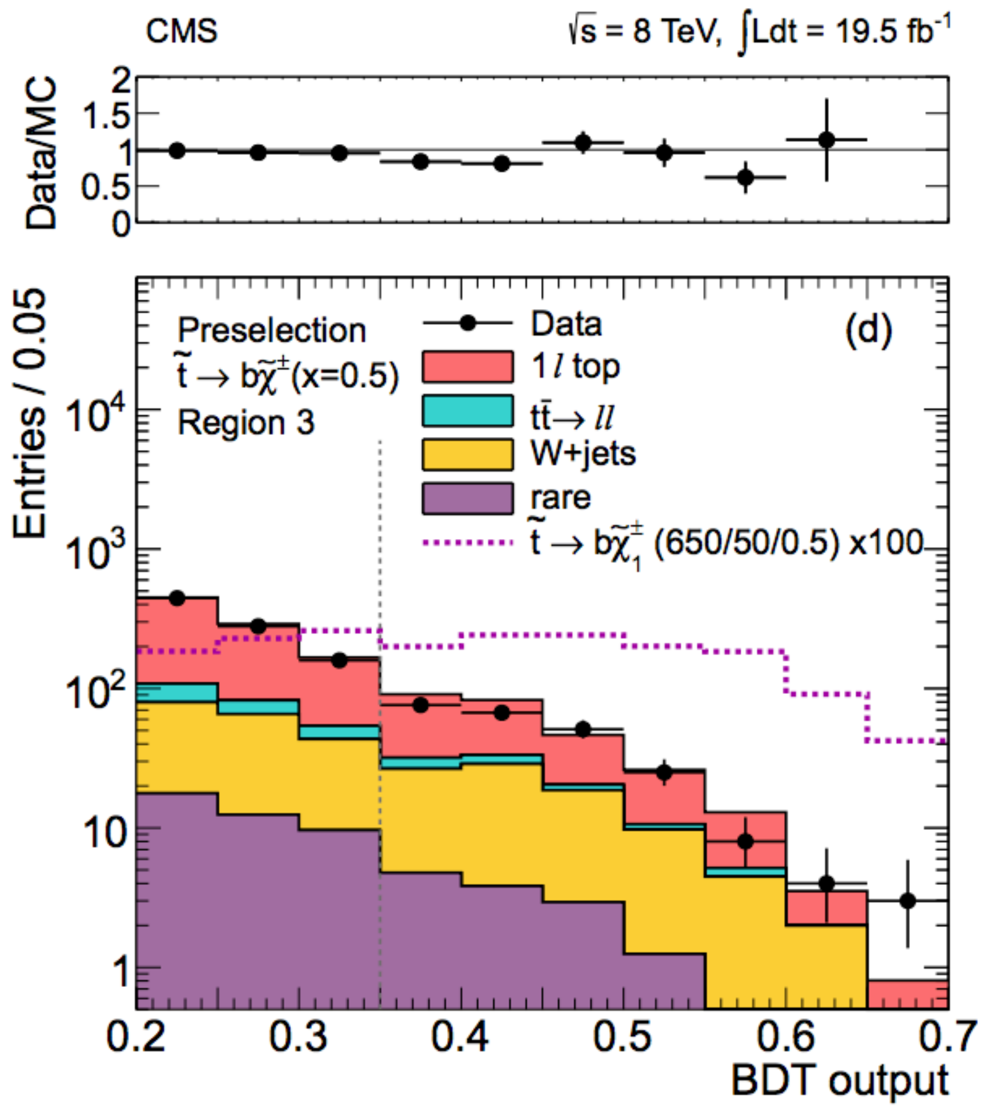
Approach 3: parametric description



Fit function to sample events

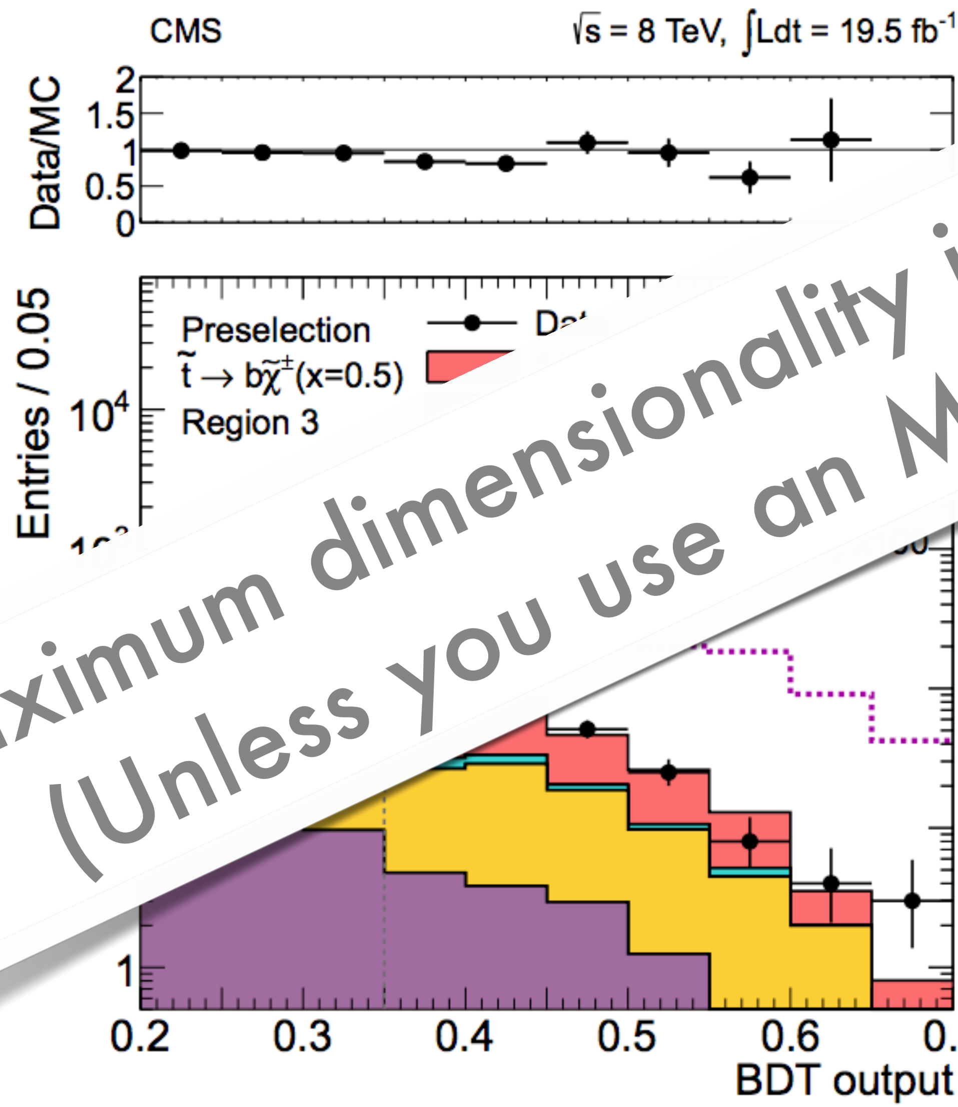
Templates vary as a function of POI

Full MC example



1308.1586

Full MC example



Maximum dimensionality is ~ 4d
(Unless you use an MVA)

1308.1586