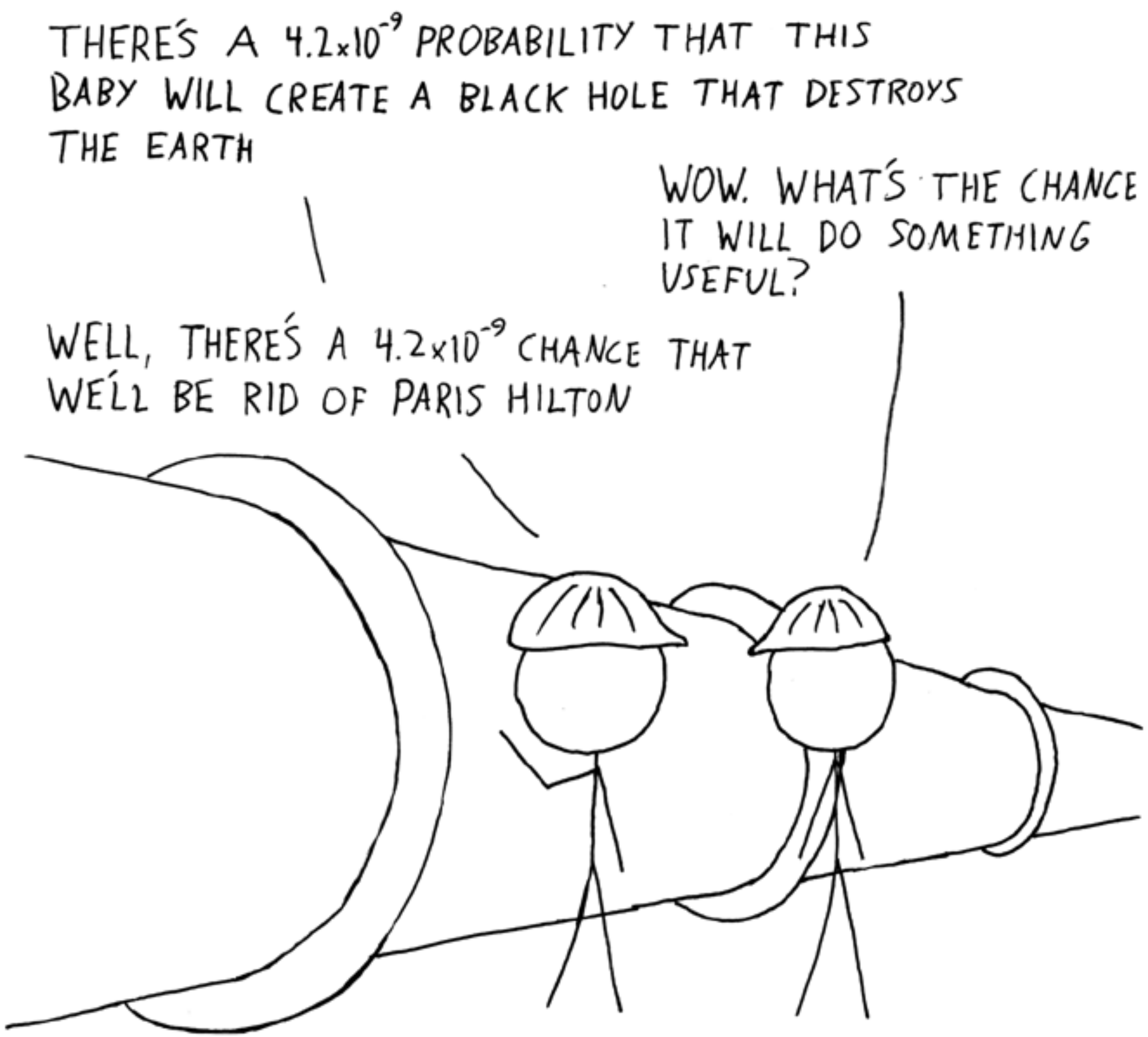
Data Analysis and Bayesian Methods
Lecture 1

Maurizio Pierini

Lecture program

	Day1	Day2	Day3	Day4	Day5
Lecture	Introduction to probability and statistics	Data analysis in a nutshell	Bayesian inference	Bayesian inference	Data analysis beyond hypothesis test

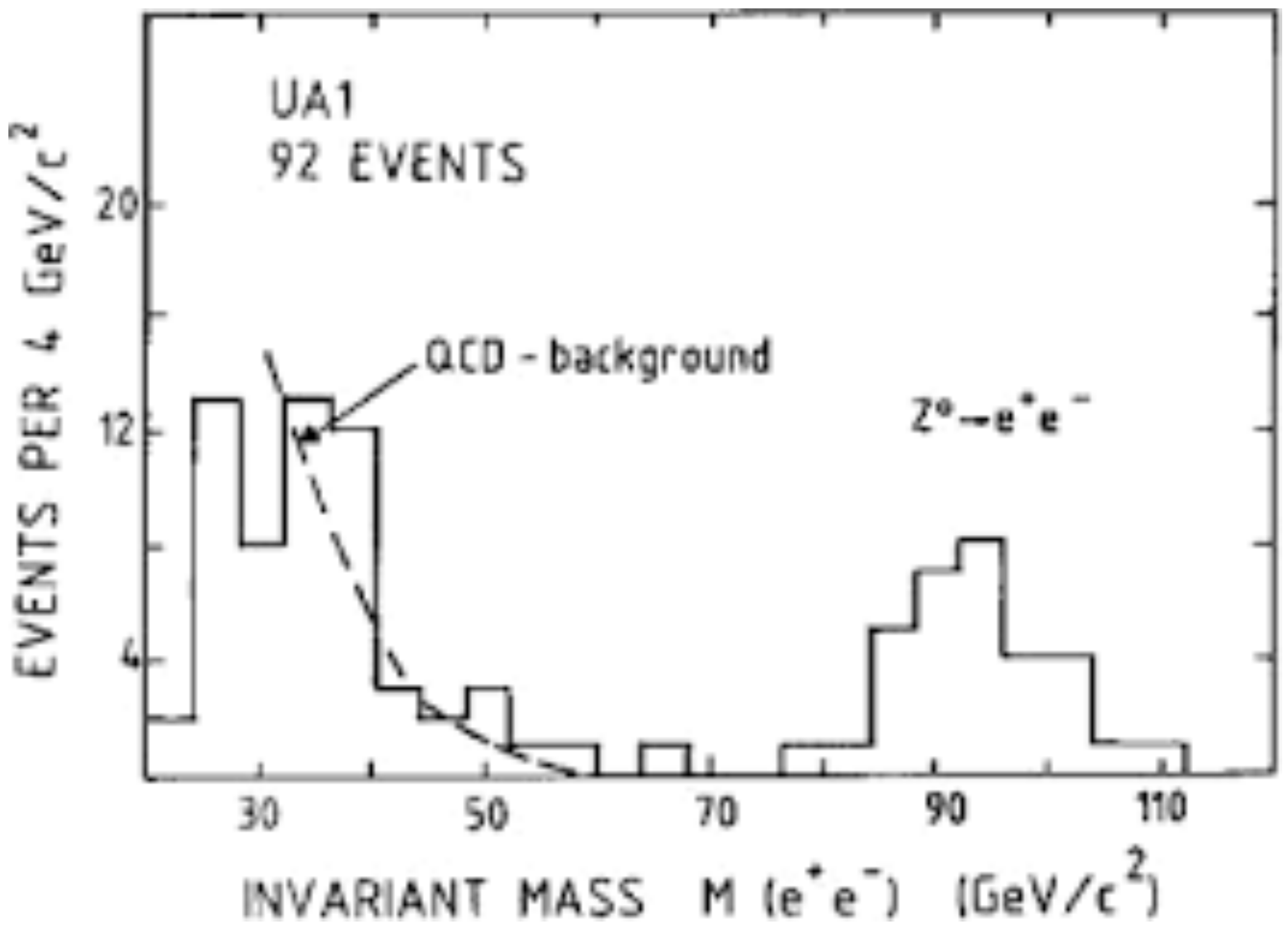
- *The idea of these lectures is to introduce data analysis practices typical of High Energy Physics*
  - *I will use some example from HEP, mostly from LHC, but the scope of these techniques is broader*
  - *I will describe the physics applications when use the examples*
  - *If you are not familiar with something, raise your hand and ask for a clarification*

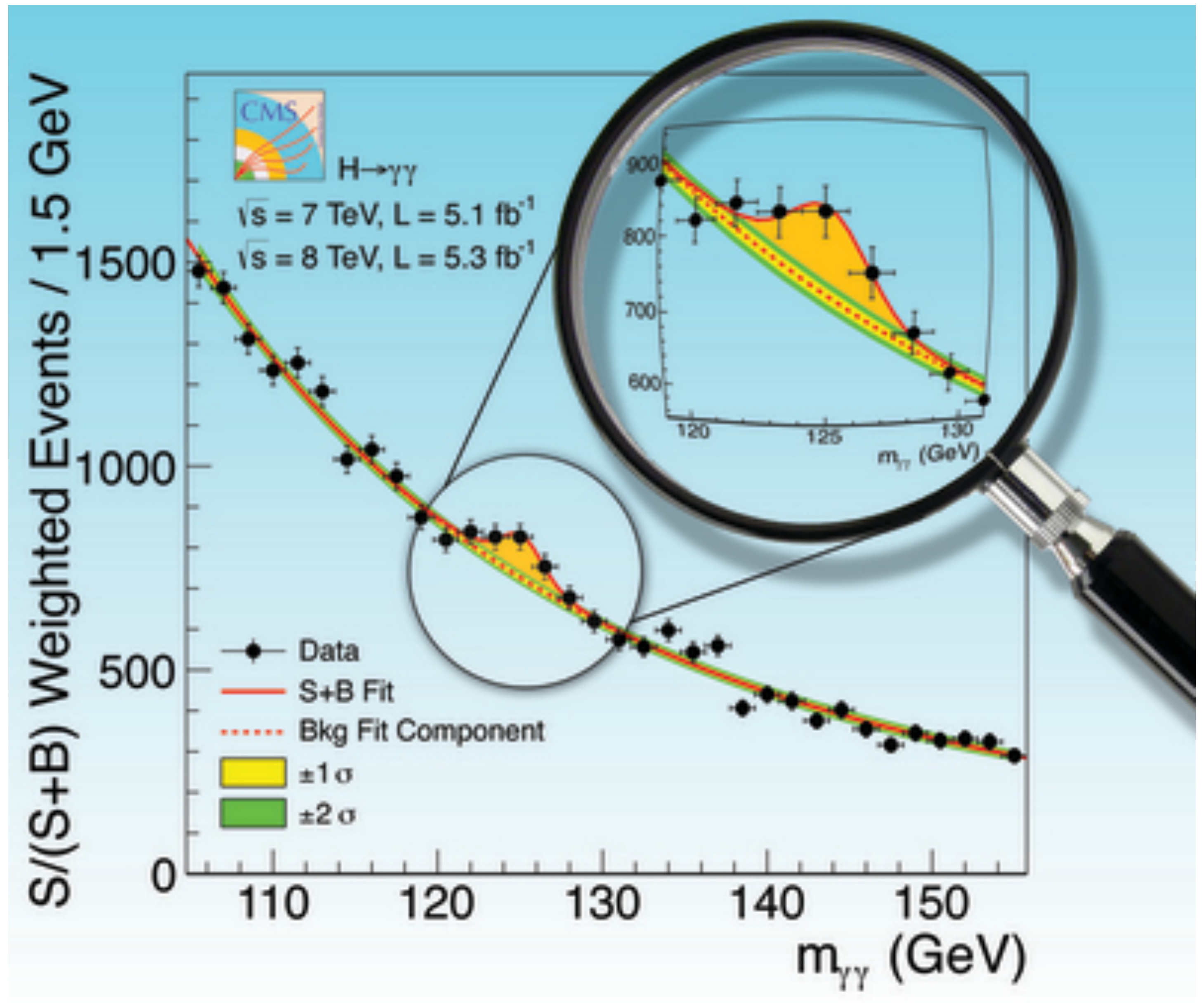


Probability in a nutshell

The starting point

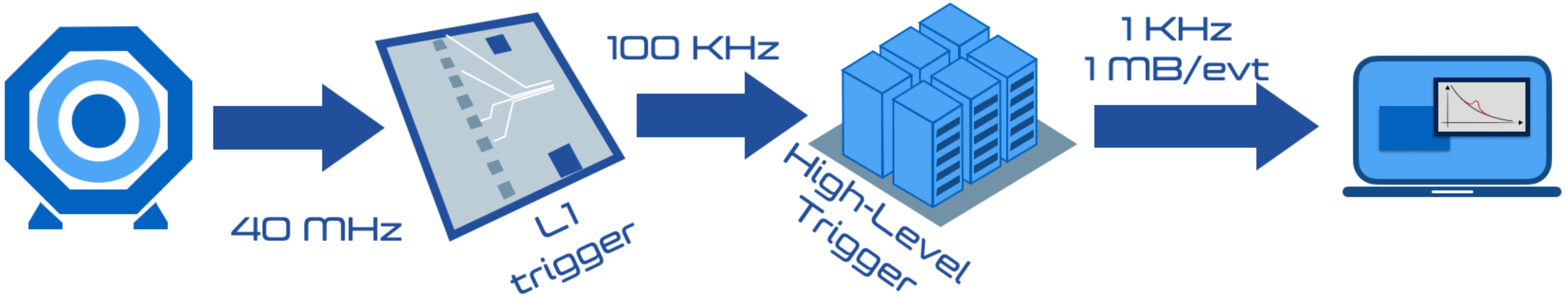
- *Since the early ‘80s, particle physics is dealing with challenging data analysis problems*
  - *Large amount of data created*
  - *Interest in rare processes*
  - *Need to separate interesting events from overwhelming background*
  - *No obvious solution: whatever you do, a perfect separation is just not possible*
- *The days of “I discovered something because I just saw an event” is gone (for us, not for others, see LIGO/VIRGO)*



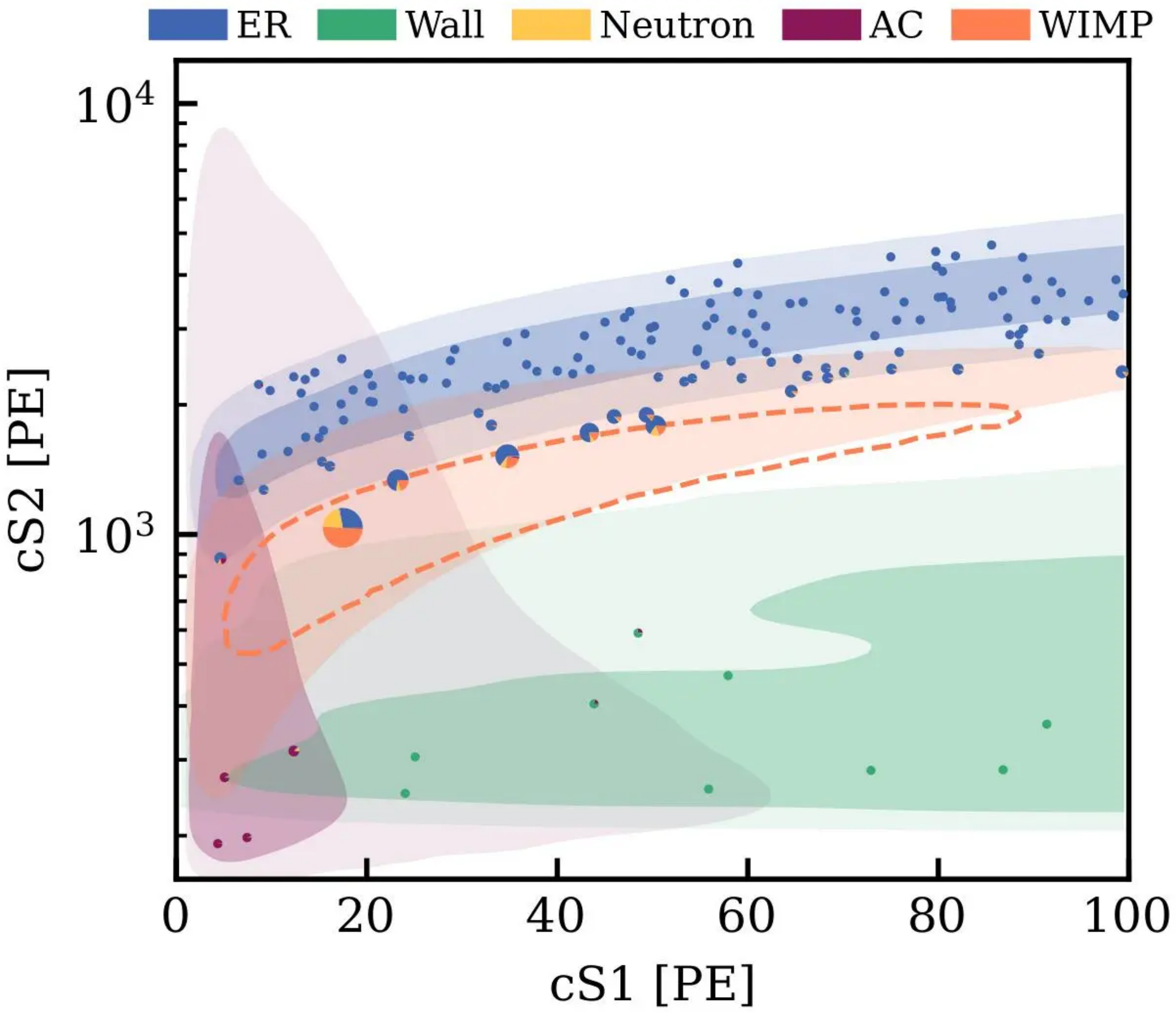


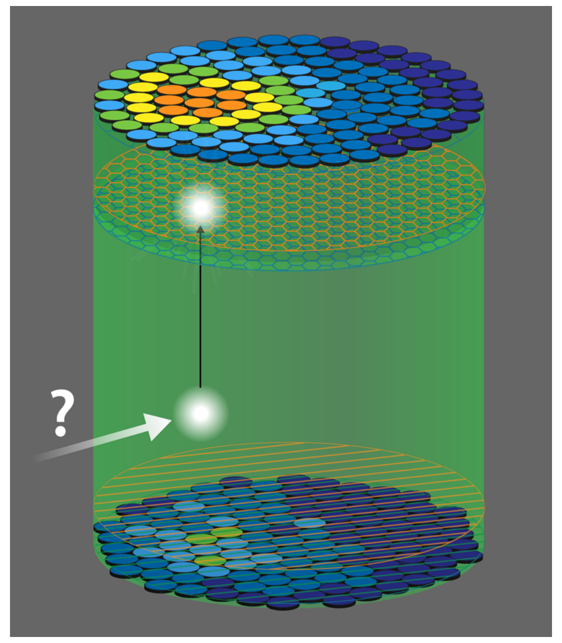
The starting point

- Clearly this is a problem at the LHC



- But this is a problem even with dark matter searches
  - You go below a mountain to look for a quite place
  - But you are looking for a very rare signal
  - Even the most unlikely noise is as probable as your signal
- You need to be able to tell the background from the signal





A counting experiment

- You are searching for Dark Matter. You build a detector underground, screened by any source of natural radiation

- You are waiting for a DM particle to hit your detector and produce an energy deposit

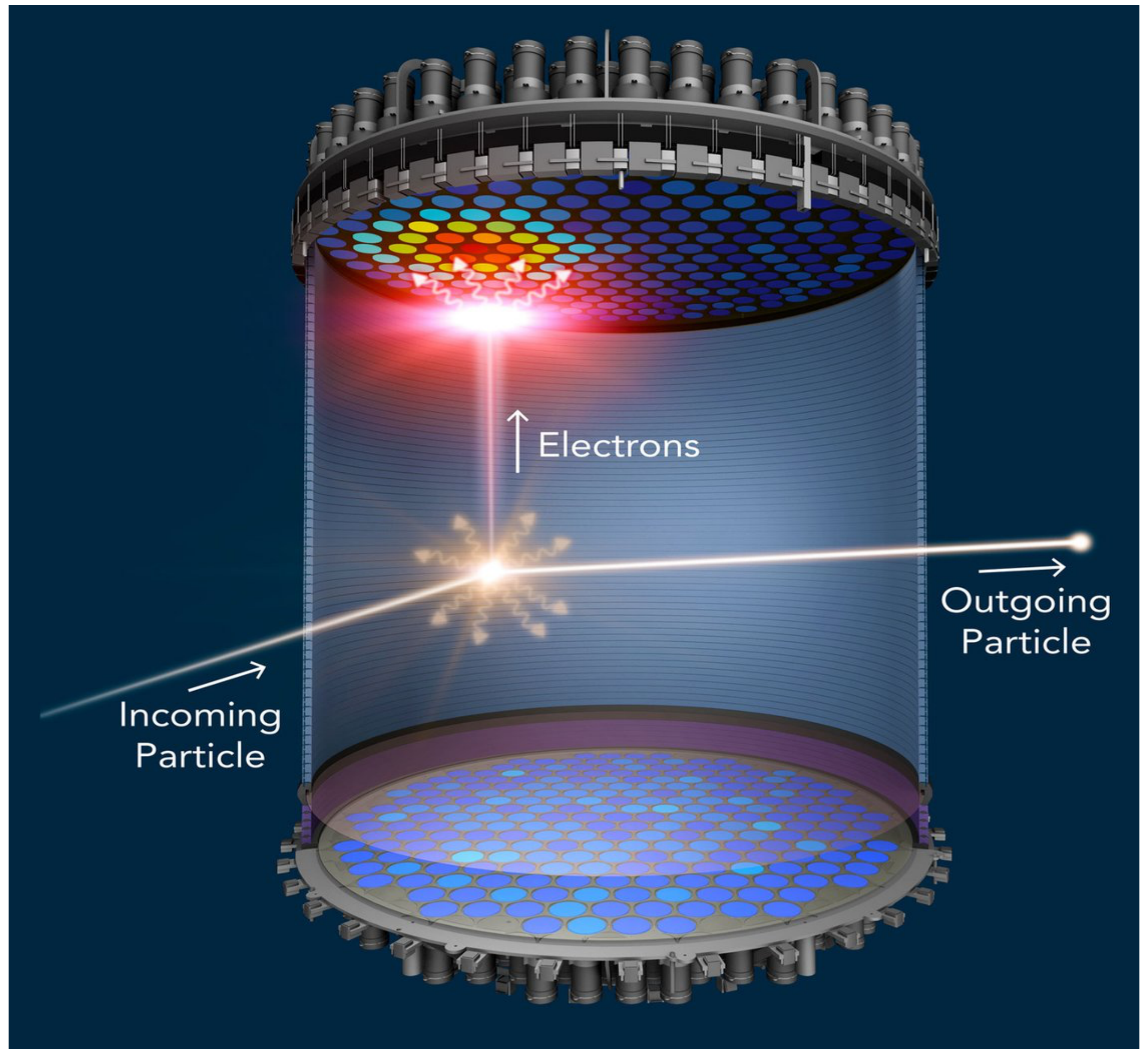
  - Signal = large energy deposit leading to a large electronic signal

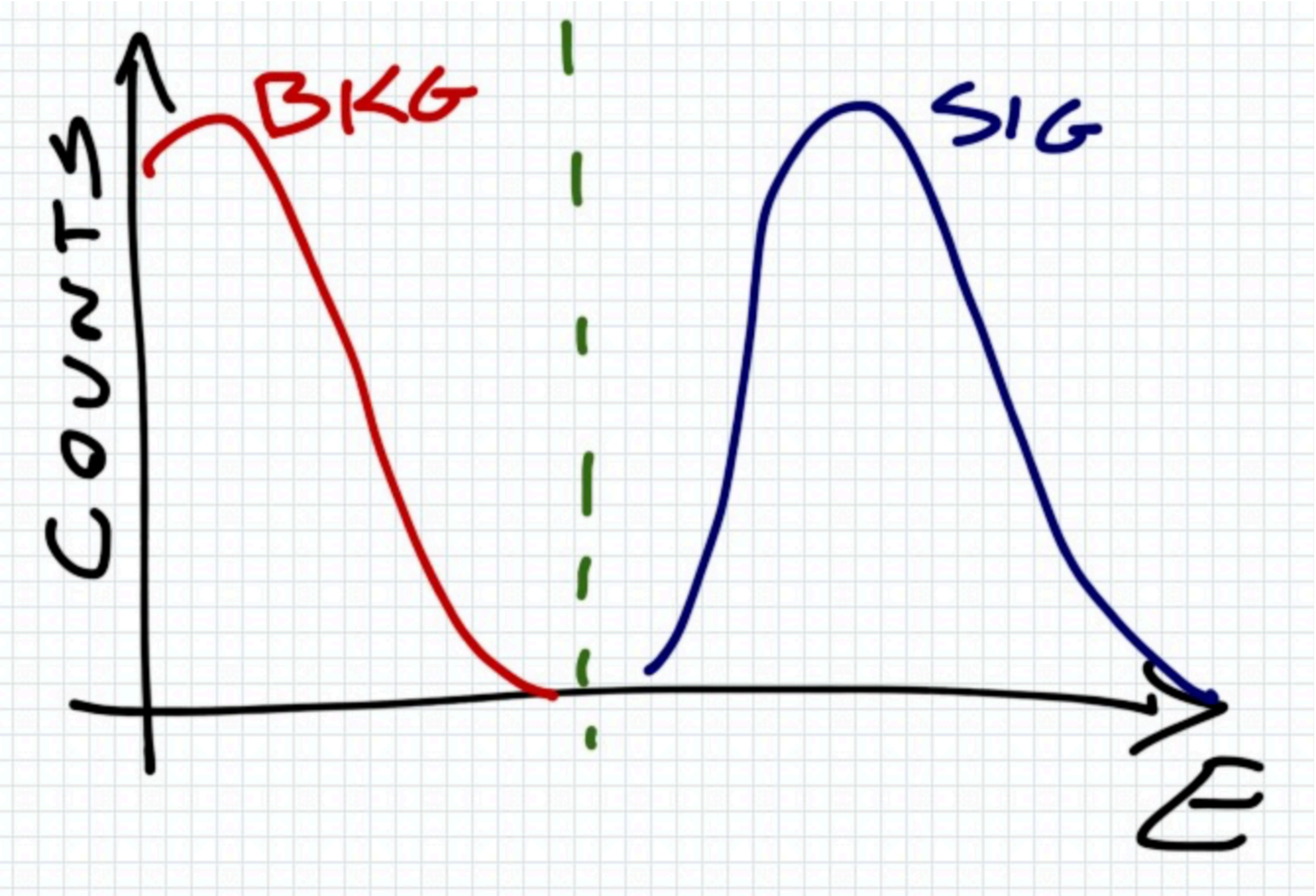
  - Background = some noise

- You count events with large enough electronic signal

  - You see any, you get a Nobel Prize

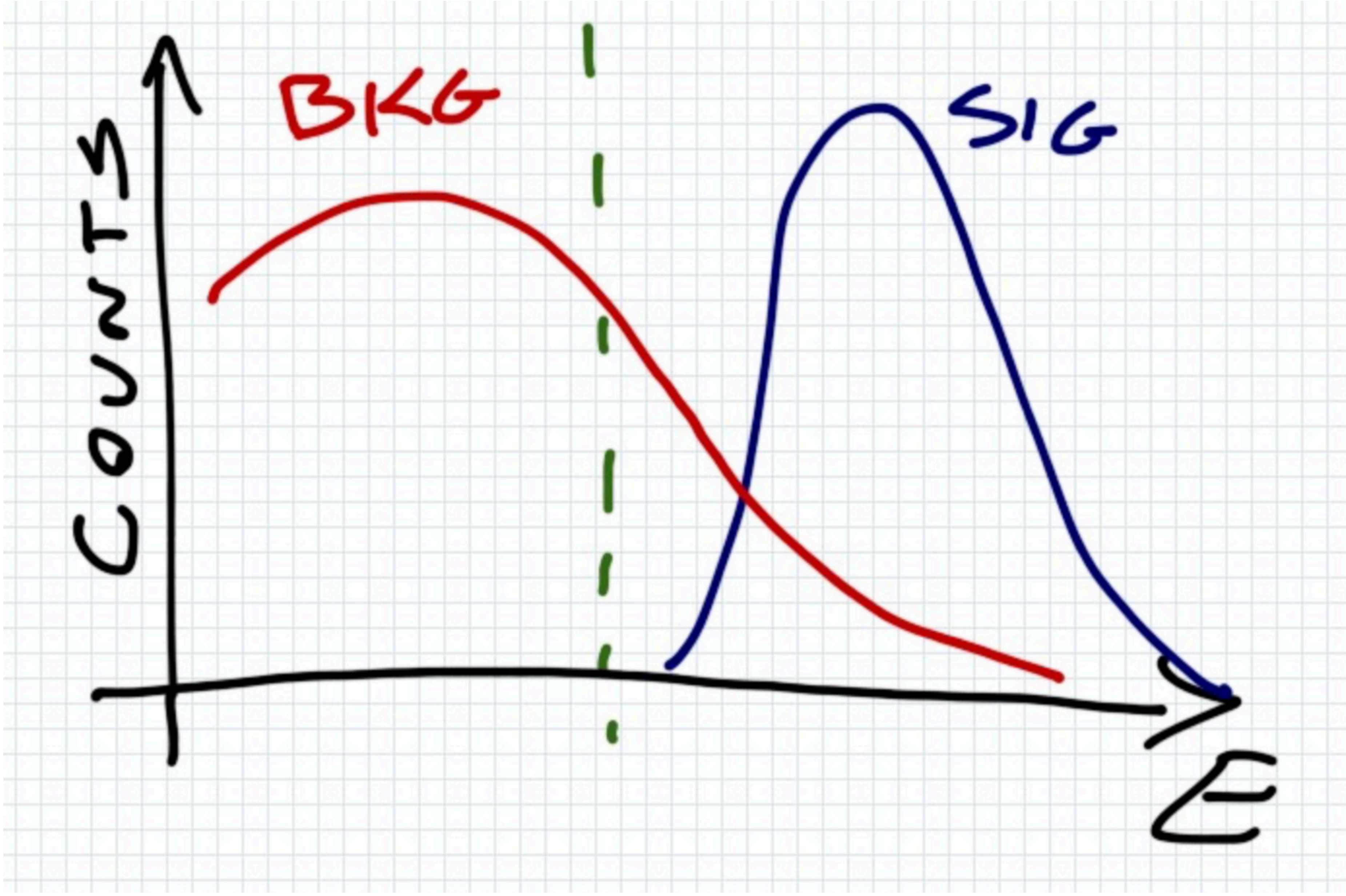
- Why do you need statistics at all?





Separating Sig from Bkg

- Real life is not like that: whatever is your fiducial region (your cut on E) you never expect exactly 0 background
  - And even if you know exactly the number of bkg events you expect (e.g., $\lambda = 1$ event) you cannot jump to conclusions if you see $> \lambda$ events
  - This is because statistical fluctuations happen
    - If you toss the same coin ten times, you expect 5 heads, but you might see 4, or 6, etc
  - What is the probability of seeing k events when you expect λ?



Bernoulli's process

- *You pick k items out of a bag with N items and you ask a yes/no question*
  - *does my event come with energy E above a threshold?*
  - *is my ball red?*
- *Let's call*
  - *p: probability that the answer is Yes*
  - *q = 1-p : probability that the answer is No*

Axiomatic Probability

Probability is a set function *P(E)* that assigns to every event *E* a number called the "probability of *E*" such that:

1. The probability of an event is greater than or equal to zero

$$P(E) \geq 0$$

2. The probability of the sample space is one

$$P(\Omega) = 1$$

Binomial distribution

- *Probability of one out of one item being Y*

$$P(k=1\,|\,N=1) = p$$

- *Probability of one out of two items being Y [order not important!]*

$$P(k=1\,|\,N=2) = \frac{pq+qp}{p^2+pq+qpq^2} = 2pq$$

- *Probability of k out of N items being Y [order not important!]*

$$P(k\,|\,N) = \frac{n!}{k!(N-k)!}p^k q^{N-k}$$

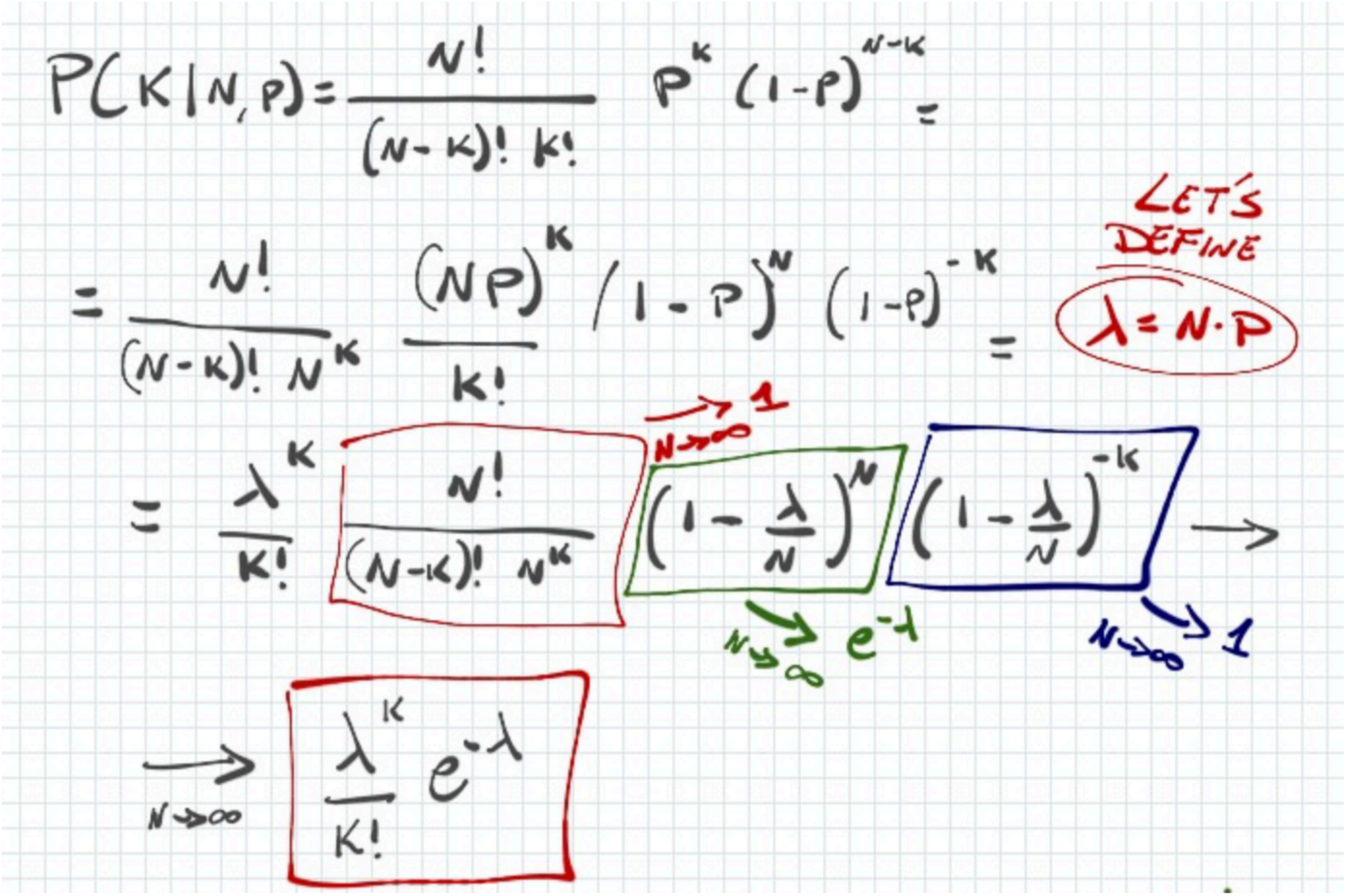
Probability that the selected event is obtained k times out of the total of N trials.

Probabability that something other than the chosen event will occur in all the other trials.

The "combination" expression, which is the permutation relationship (the number of ways to get k occurrences of the selected event) divided by k ! (the number of different orders in which the k events could be chosen, assuming they are distinguishable).

Limit of rare events

- For N→∞ with p→0 so that Np stays finite, the Binomial distribution takes the form of a Poisson distribution
- This is the distribution followed by your counting experiment for a very hard cut on the recorded energy (i.e., for a very small number of expected background events)

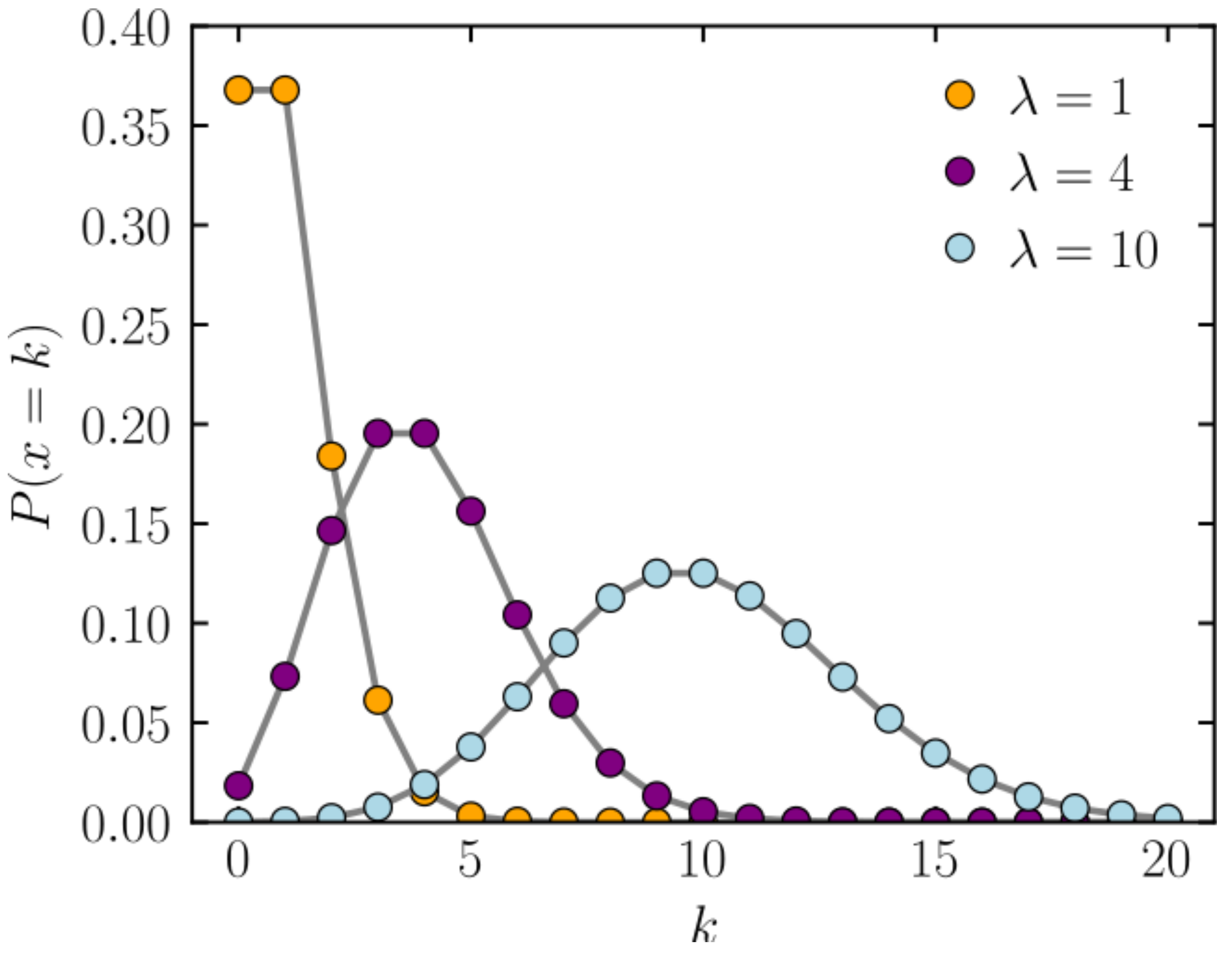


Poisson distribution

$$f(k;\lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- k is the unknown (the outcome of our experiment counting). It takes integer values by construction

- λ is the parameter determining the distribution shape and it is related to the most probable outcome of our counting experiment. It might be an integer, but in general it is a real number



Expectation value

- *What is the most probable outcome of our experiment?*
  - *For a given value of λ, the probability of seeing k=0, 1, 2, etc. depends on the value of the Poisson distribution*
  - *So, we can compute the expectation value of k as a weighted average of all the possible outcomes of the experiment, waited by the value of the Poisson distribution*

Expectation value

$$E[k|\lambda] = \frac{0\cdot P(0|\lambda) + 1\cdot P(1|\lambda) + 2\cdot P(2|\lambda) + \ldots}{P(0|\lambda) + P(1|\lambda) + P(2|\lambda) + \ldots} =$$

$$= \frac{\sum_{k=0}^{\infty} \lambda P(k|\lambda)}{\sum_{k=0}^{\infty} P(k|\lambda)} =$$

$$= \frac{e^{-\lambda} \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!}}{e^{-\lambda} \sum \frac{\lambda^k}{k!}} = \frac{\lambda \sum_{k=1}^{\infty} \frac{k\cdot\lambda^{k-1}}{k\cdot(k-1)!}}{e^{\lambda}} = \frac{\lambda e^{\lambda}}{e^{\lambda}} = \boxed{\lambda}$$

Expectation value

DEFINITION:

FOR A GENERIC $P(k|\alpha)$

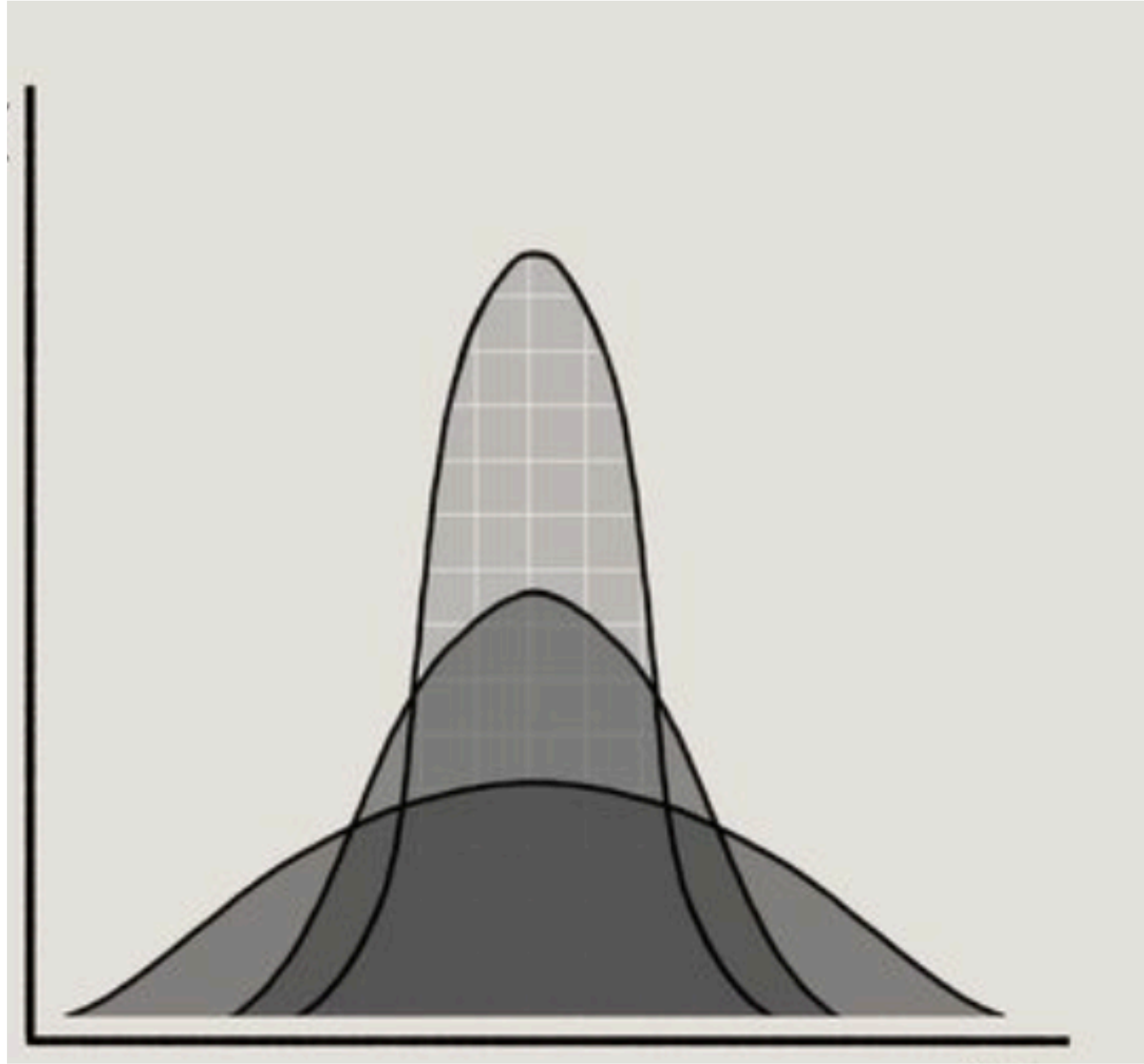
$$E[k|\alpha] = \frac{\sum_k k\,P(k|\alpha)}{\sum P(k|\alpha)}$$

FOR A GENERIC $P(x|\alpha)$

$$E[x|\alpha] = \frac{\int dx\; x\,P(x|\alpha)}{\int dx\; P(x|\alpha)}$$

Variance

- *E[x] is not enough to characterize a distribution*
  - *distributions with same E[x] can be very different*
- *It is convenient to have a measure of the dispersion of points around E[x]*
  - *One typically introduced the variance (aka mean square error)*



$$Var[x] = E[(x - E[x])^2] = E[x^2] - E[x]^2$$

Variance of a Poisson dist.

- *The Variance of Poisson distribution is equal to its expectation value*
- *It is convenient to introduce the Root Mean Square (RMS) = $\sqrt{Var}$, since it has the same "units" as the mean and it quantifies the "statistical dispersion" around it*

$$E[k^2] = \frac{\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k k^2}{k!}}{\underbrace{\left(\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\right)}_{e^{\lambda}} e^{-\lambda}} = \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k k}{(k-1)!} =$$

$$= \lambda e^{-\lambda}\left[\sum_{k=1}^{\infty} \frac{\lambda^{k-1}(k-1)}{(k-1)!} + \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}}_{e^{\lambda}}\right] =$$

$$= \lambda e^{-\lambda}\left[\lambda \underbrace{\sum_{k=2} \frac{\lambda^{k-2}}{(k-2)!}}_{e^{\lambda}} + e^{\lambda}\right] = \lambda(\lambda+1) = \boxed{\lambda^2+\lambda}$$

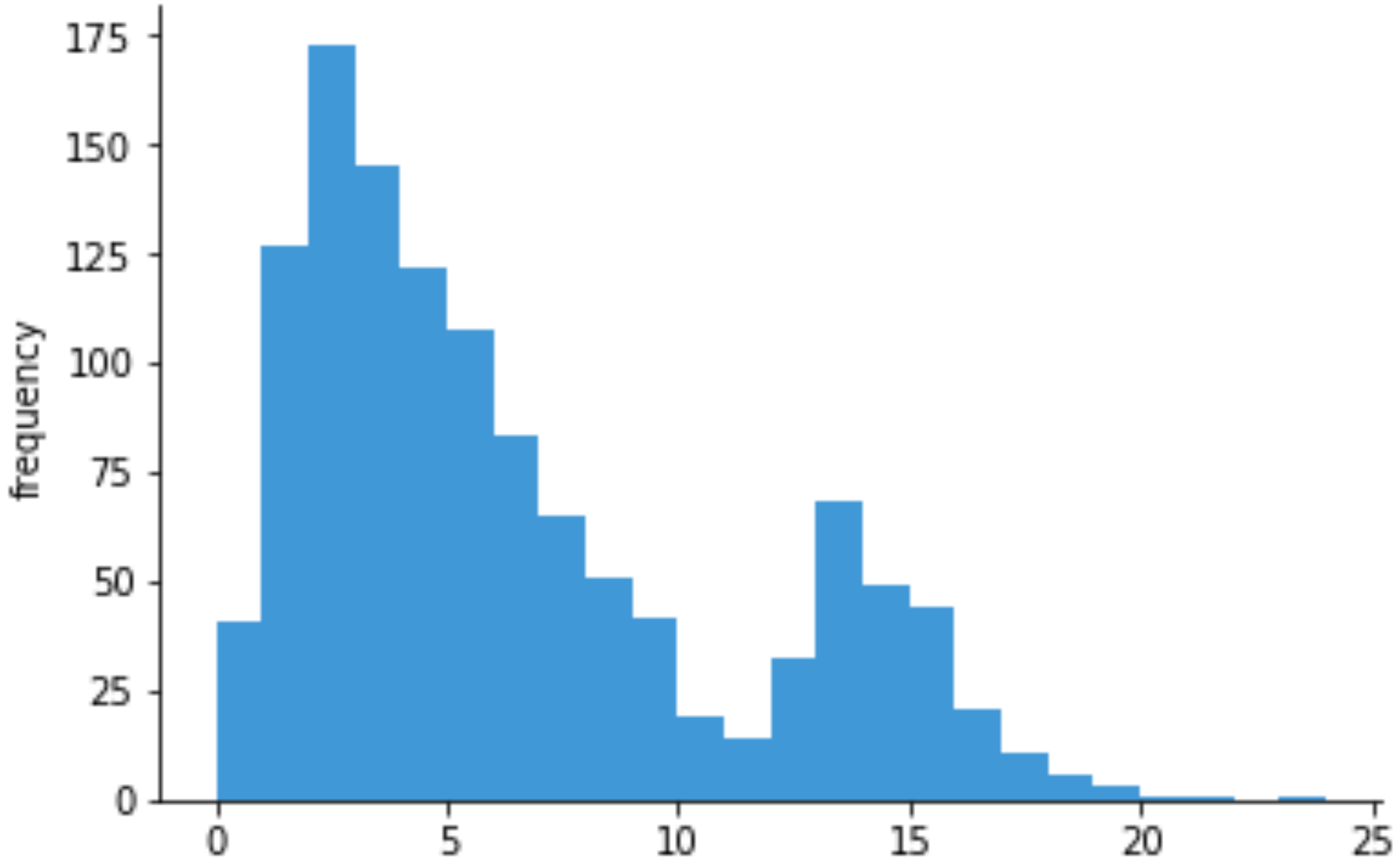
$$E[(k-E[k])^2] = E[k^2] - E[k]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Expectation value and variance

Function	Distribution	E[x]	Var[x]
Poisson	$P(k\|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$	λ	λ
Binomial	$P(k\|p,N) = \frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}$	pN	p(1-p)N
Gaussian	$G(x\|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

Histogram uncertainty

- *The number of entries in a histogram bin can be computed as a Y/N question (Bernoulli process)*
  - *The for large p_i, the bin counting follows a binomial distribution*
    - *expected count =* $Np_i \pm \sqrt{Np_i(1-p_i)}$
  - *For small p_i, the bin counting follows a Poisson distribution*
    - *expected counts =* $Np_i \pm \sqrt{Np_i}$
  - *In both cases, the relative uncertainty on the expected counting decreases* $\propto 1/\sqrt{N}$ *(which is why experiments take more data to increase precision)*



Asymptotic limit: Gaussian

$$\ln\left(P(k|\lambda)\right) = \ln\left(\frac{\lambda^k e^{-\lambda}}{k!}\right) \simeq \ln\left(\frac{\lambda^k e^{-\lambda}}{k^k e^{-k}\sqrt{2\pi k}}\right) =$$

STIRLING'S APPROXIMATION $x! \simeq x^x e^{-x}\sqrt{2\pi x}$

$$= k\ln\lambda - \lambda - k\ln k + k - \ln\sqrt{2\pi k} =$$

$$= \dots = -\frac{y^2}{2\lambda} + \frac{y^3}{6\lambda^2} - \ln\sqrt{2\pi(y+\lambda)} \simeq$$

$$\simeq -\frac{y^2}{2\lambda}$$

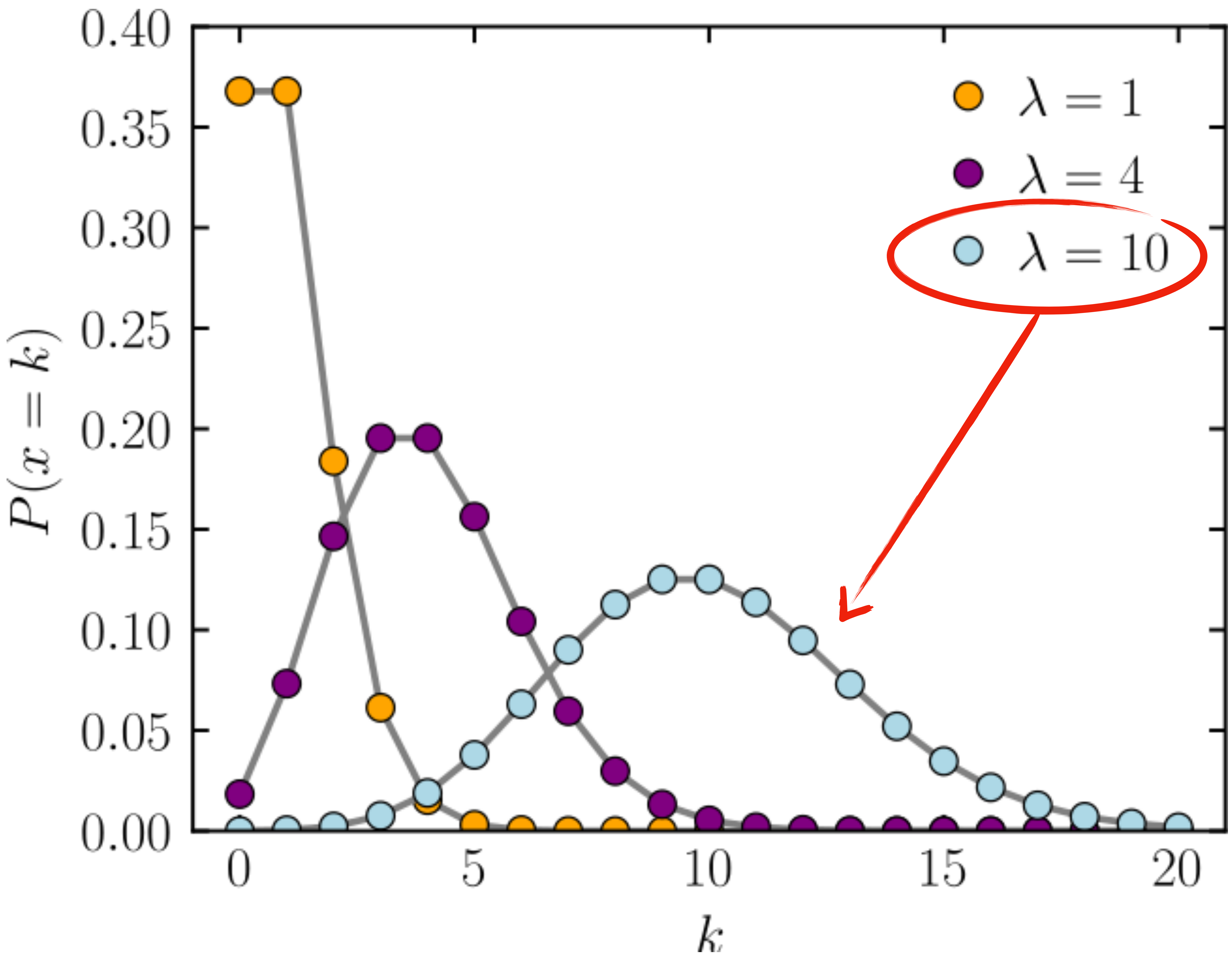
$y = k - \lambda$ WITH $\lambda \to \infty$, $k \sim O(\lambda)$ $\Rightarrow y \ll \lambda$

$$\ln(1+\varepsilon) \simeq \varepsilon - \varepsilon^2 + \frac{\varepsilon^3}{3} \dots$$

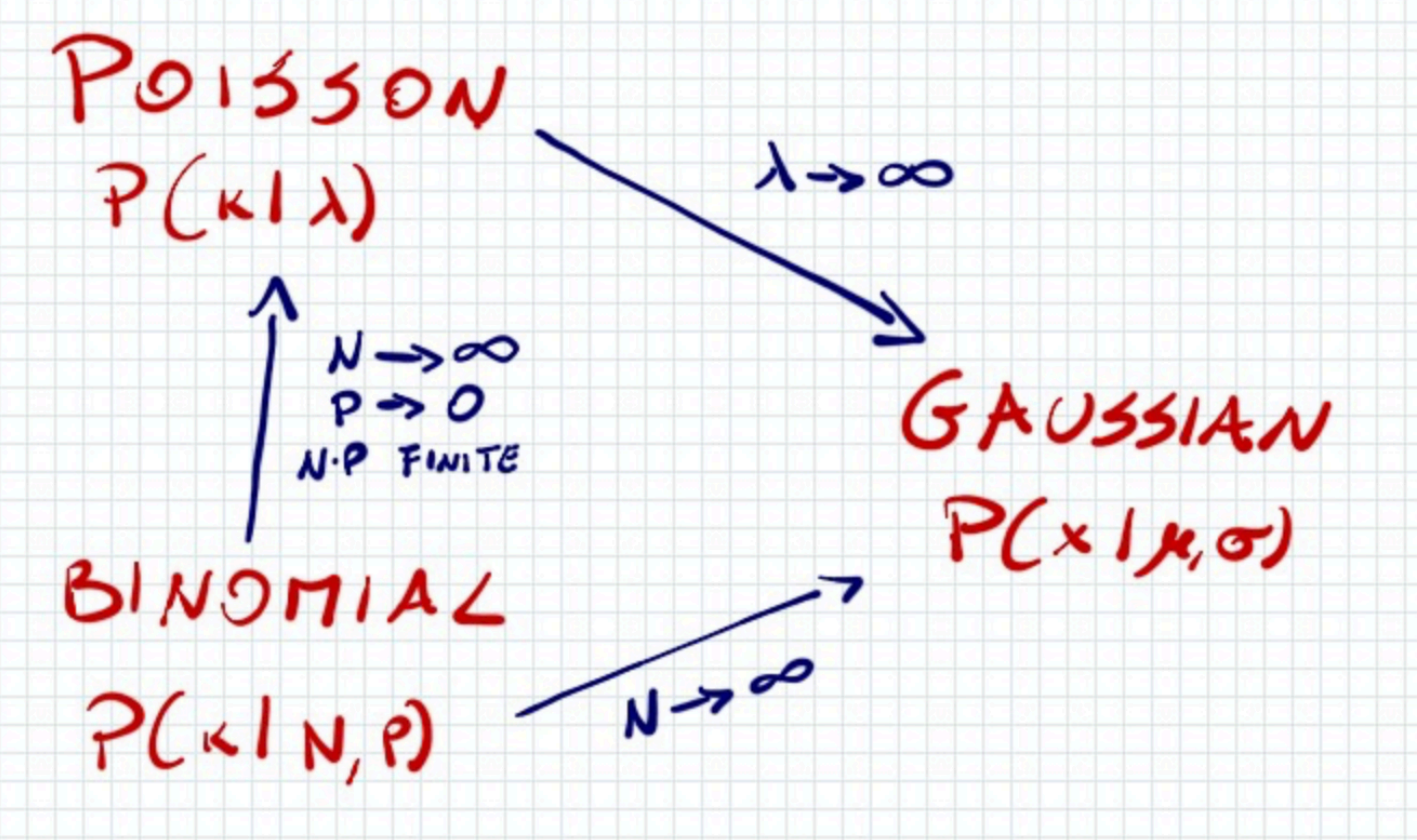
$$P(y|\lambda) = e^{-\frac{y^2}{2\lambda}}$$

GAUSSIAN

How big is big λ?



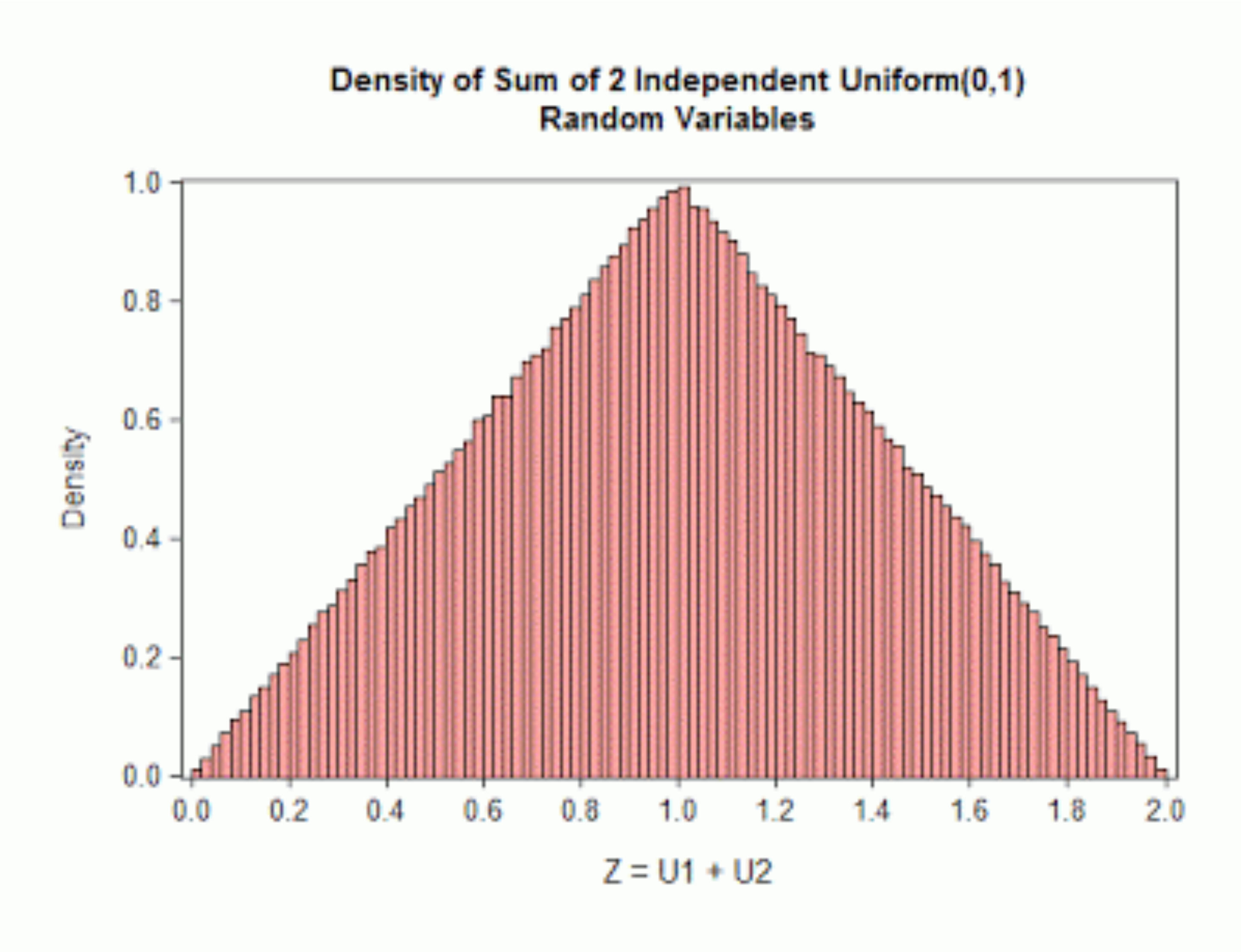
The special role of Gaussian

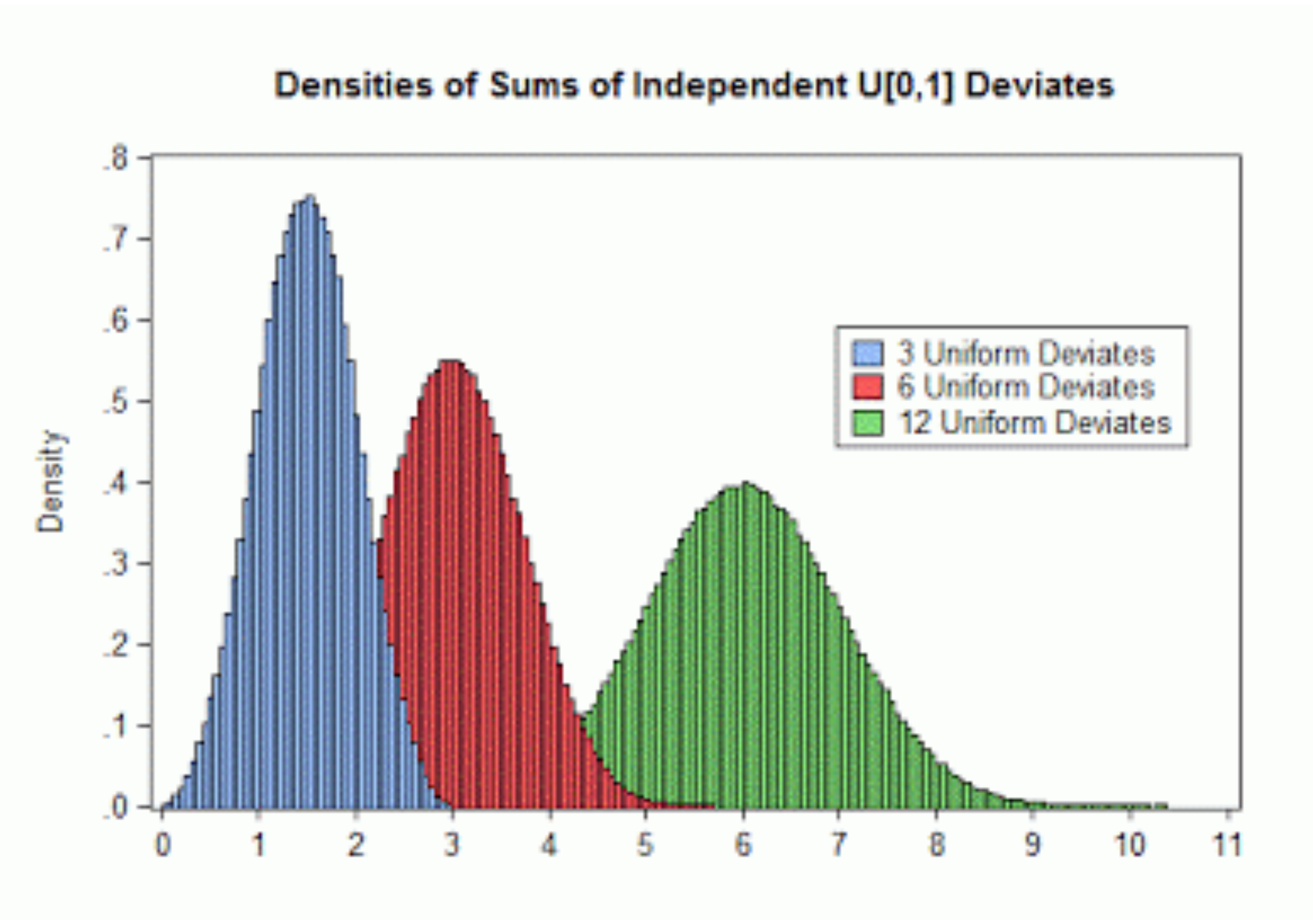


The special role of Gaussian

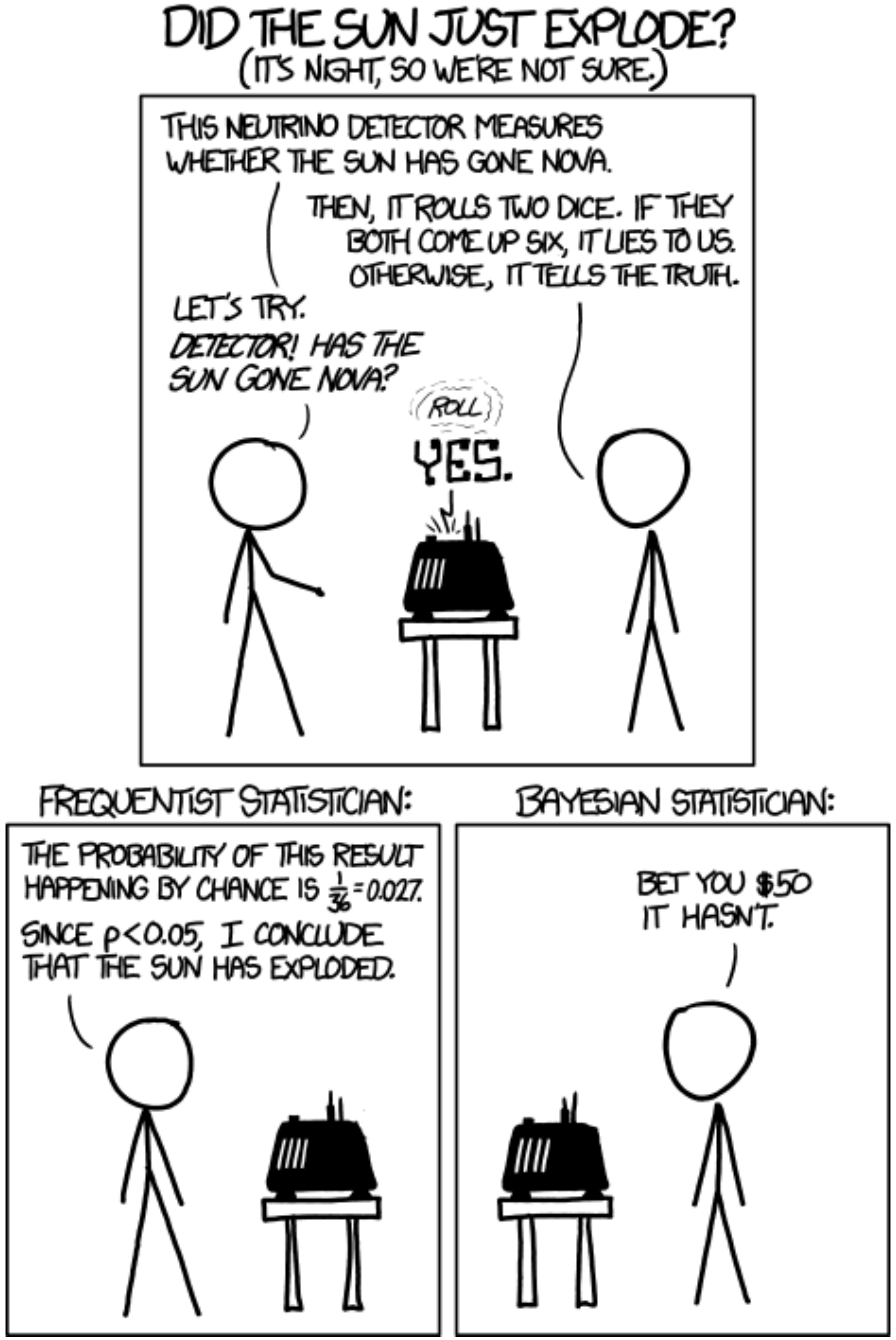
- *The central limit theorem establishes the role of the Gaussian distribution as the asymptotic limit of a much broader class of problems*

In probability theory, the central limit theorem establishes that, in many situations, when independent random variables are summe up, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. (from Wikipedia)





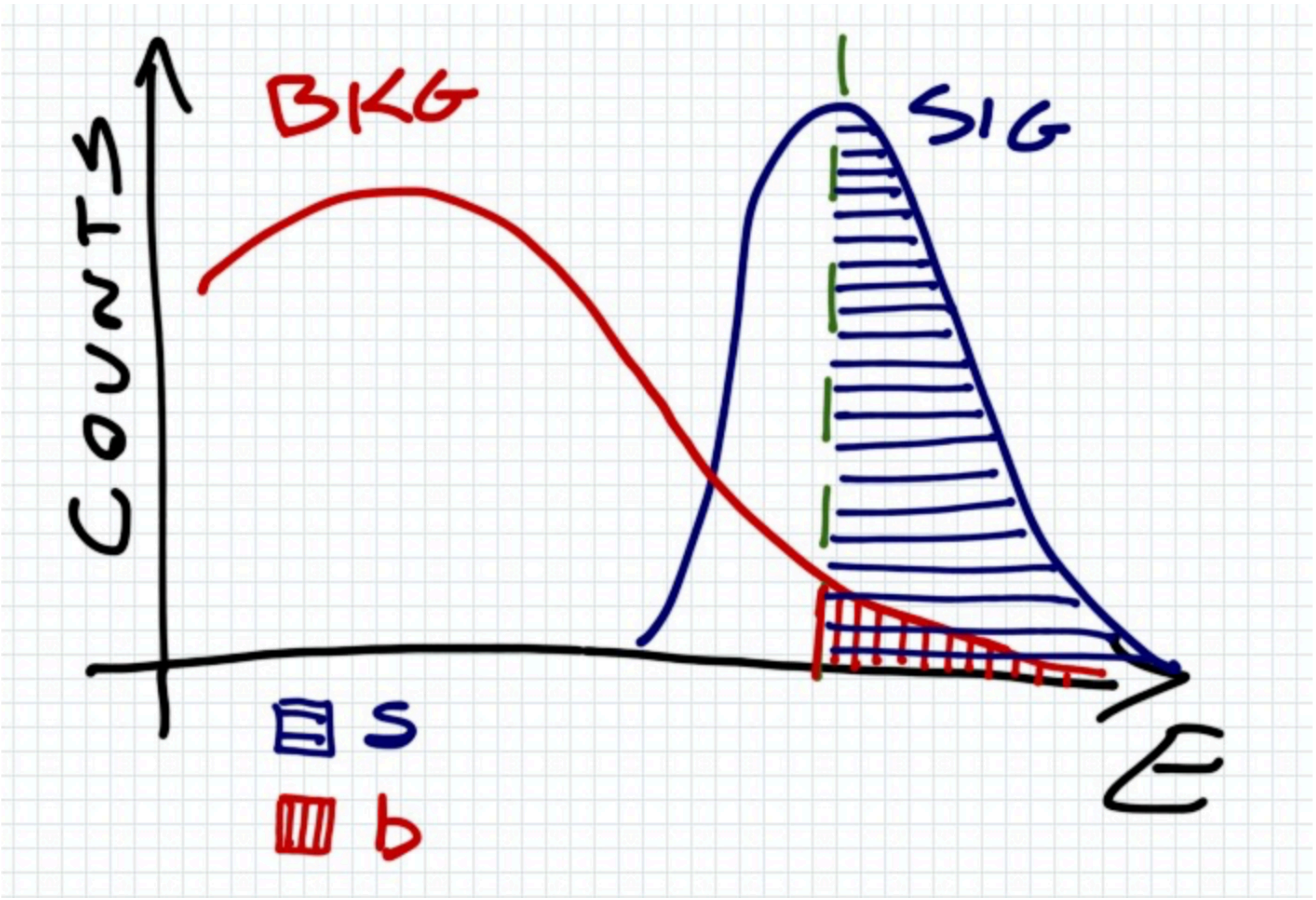
- *In practice, in a counting experiment one has to deal with*
  - *The intrinsic variation (statistical uncertainty) associated with the spread of the distribution (Poisson, Binomial, etc.)*
  - *The systematic uncertainty, associated to the uncertainty on the knowledge of the expectation. This is typically the result of many contributions -> it tends to have a Gaussian behavior*



Statistics in a nutshell

From Probability Model to Likelihood

- We have a discriminating quantity, in our case the energy E
- We apply a threshold and count values above threshold
- The integral of the background distribution above threshold sets the expected background count
- In absence of a signal, we expect to observe a number of counts distributed around b and following a Poisson distribution (we typically cut tight enough for the expected yield to be small) P(n|b)
- In presence of a signal, we expect that the observed counting distributed according to a Poisson P(n|s+b) (signal, if exists, is rare, so s is also small)
- How do we know if what we observe favours the BKG-only hypothesis P(n|b) or the SIG+BKG hypothesis P(n|s+b)?



From Probability Model to Likelihood

- ***Probability***: *When we introduced distributions, we started from known distributions (e.g., a Poisson on known λ) and we tried to characterize a typical experiment outcome*

- ***Hypothesis Testing***: *Now we inverted the problem: we know the experiment outcome (e.g., we counted events above threshold during a one-year run) and we ask ourselves which of two λ values (bkg-only or sig+bkg) they come from*

- ***Inference***: *we could also just ask what is the value of λ more compatible with the observation (trivial question in this case - right? - but not in general). This is a typical application of maximum likelihood fits and a regression problem in Machine Learning*

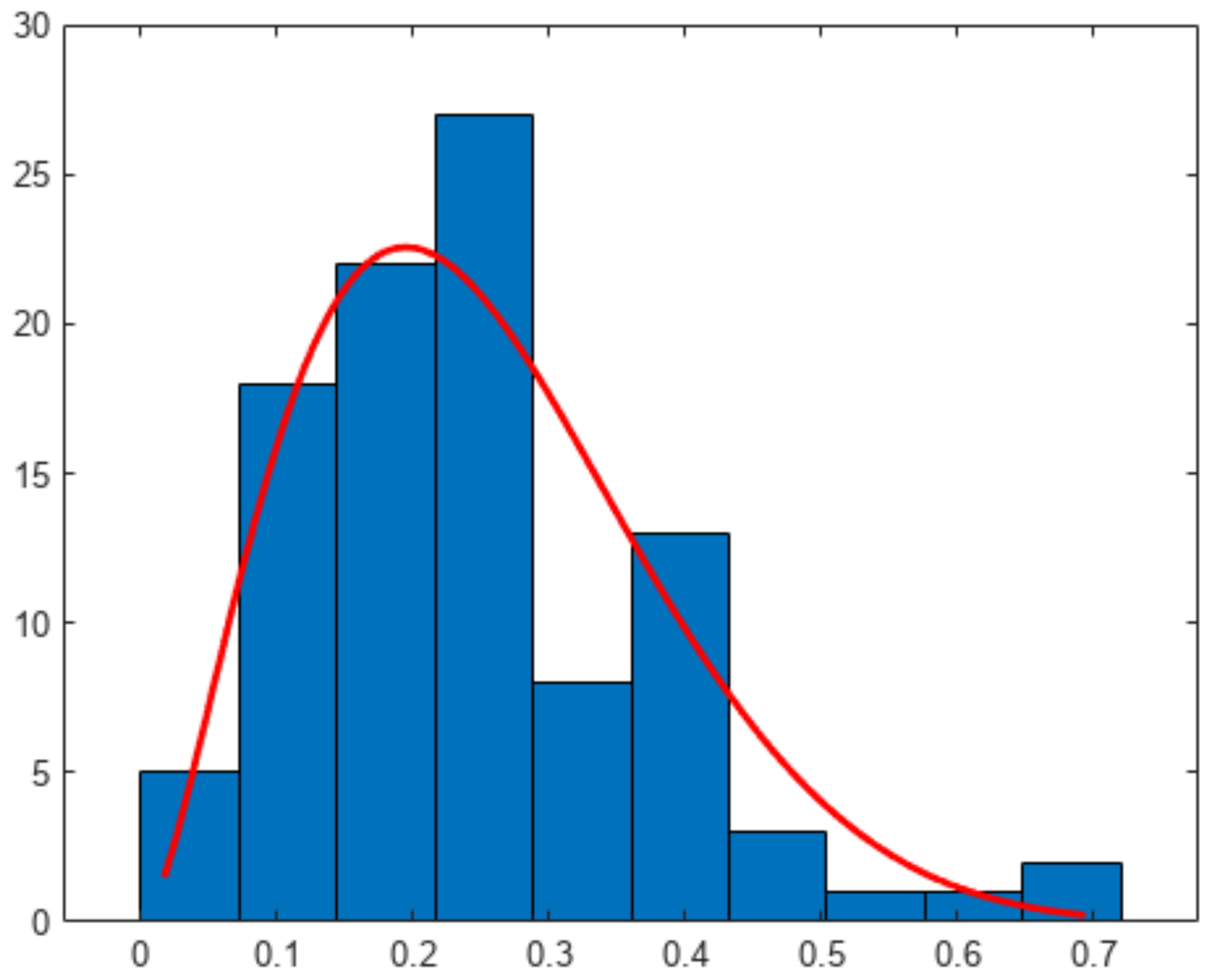
Likelihood

$$\Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- *Given a statistical model (e.g., our Poisson of known λ and unknown k), we can assess probabilities. Pr is a function of k*
- *Given a class of statistical models for k, function of unknown λ, we have a likelihood model*
  - *Formally the same function but a much different object*
  - *The counting is given (observed) and the mean is unknown → A likelihood is a function of λ, given the observed k*

Likelihood

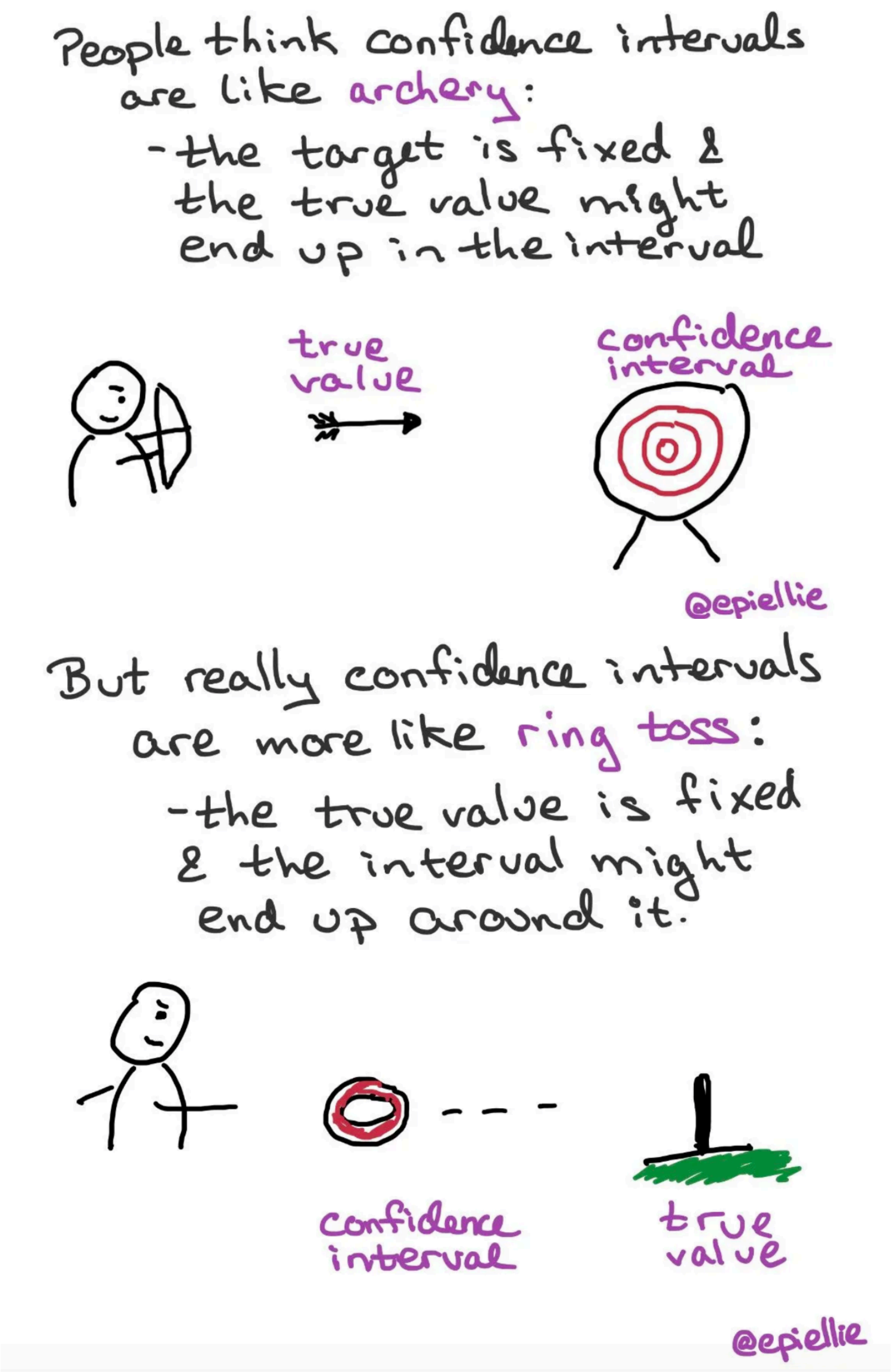
- Let's imagine a histogram of a quantity x and a curve b(x) predicting the amount of expected background
  - for each bin centre x_i we can compute $b_i = b(x_i)$
  - the b_i values will depend on a set of parameters that describe the curve $y = b(x)$
- In each bin, we observe some counting n_i
- The likelihood of the model is given by



$$\mathscr{L}(\vec{n}\,|\,\vec{\alpha}) = \prod_i P(n_i\,|\,b_i(\vec{\alpha})) = \prod_i P(n_i\,|\,b(x_i\,|\,\vec{\alpha})) = \prod_i \frac{e^{-b(x_i|\vec{\alpha})}\,b(x_i\,|\,\vec{\alpha})^{n_i}}{n_i!}$$

Two Approaches

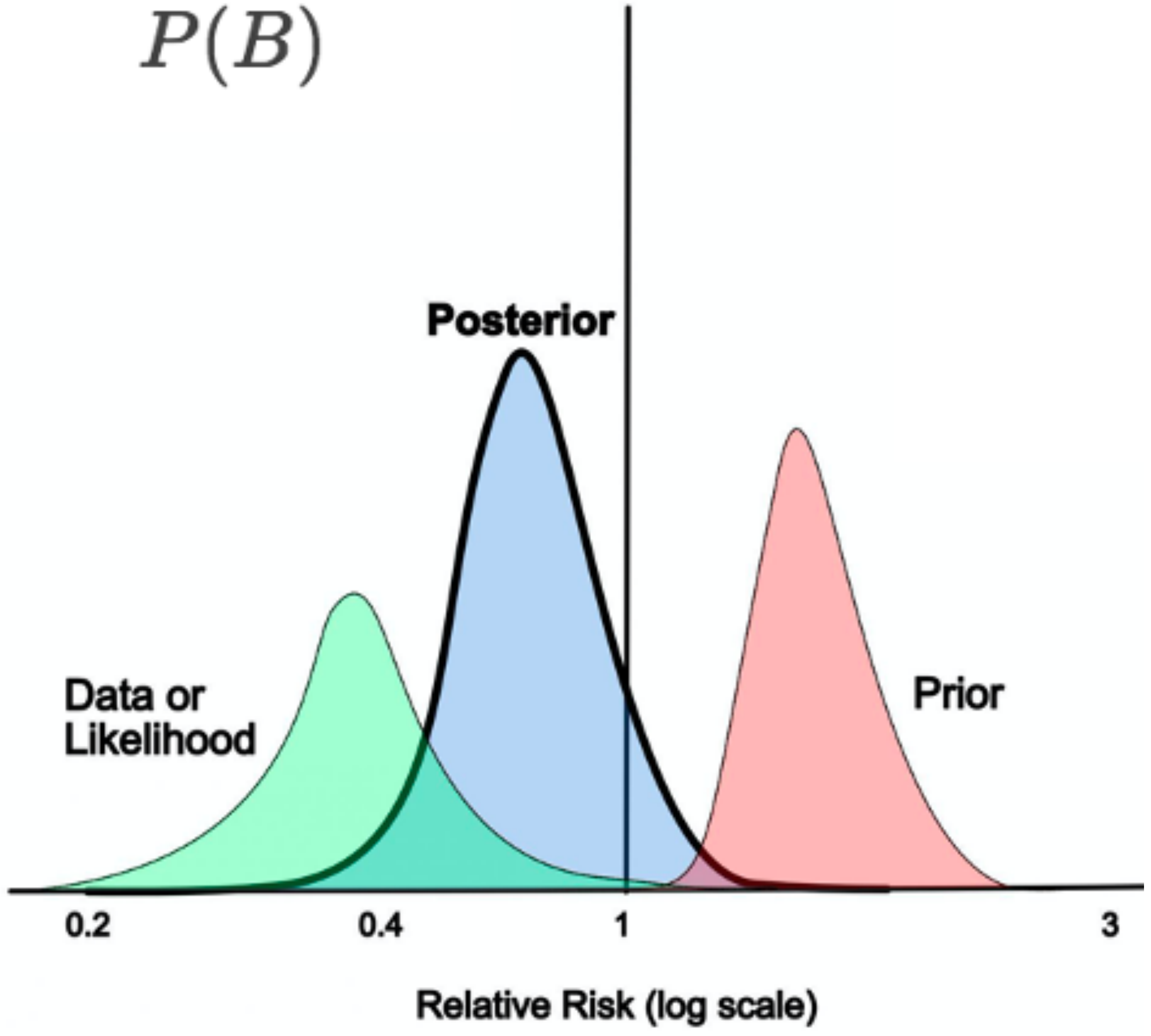
- ***Frequentist:***
  - *Frequentist statistics is a type of statistical inference that draws conclusions from sample data by emphasising the frequency or proportion of the data*
  - *Given an unaccessible true value the outcome of a measurement, frequentist statistics assess how typical the outcome is*
    - *The result is a confidence interval, defined based on a given probability (confidence level) that the true value is contained in an interval built as specified*



Two Approaches

- ***Bayesian:***
  - *Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes' theorem. Unique for Bayesian statistics is that all observed and unobserved parameters in a statistical model are given a joint probability distribution, termed the prior and data distributions*
  - *Given an accessible true value and the outcome of a measurement, Bayesian statistics assesses a probability range (credibility interval) for the true value, based on the measurement outcome and prior knowledge of the true value*

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



Frequentist and Bayesian in practice

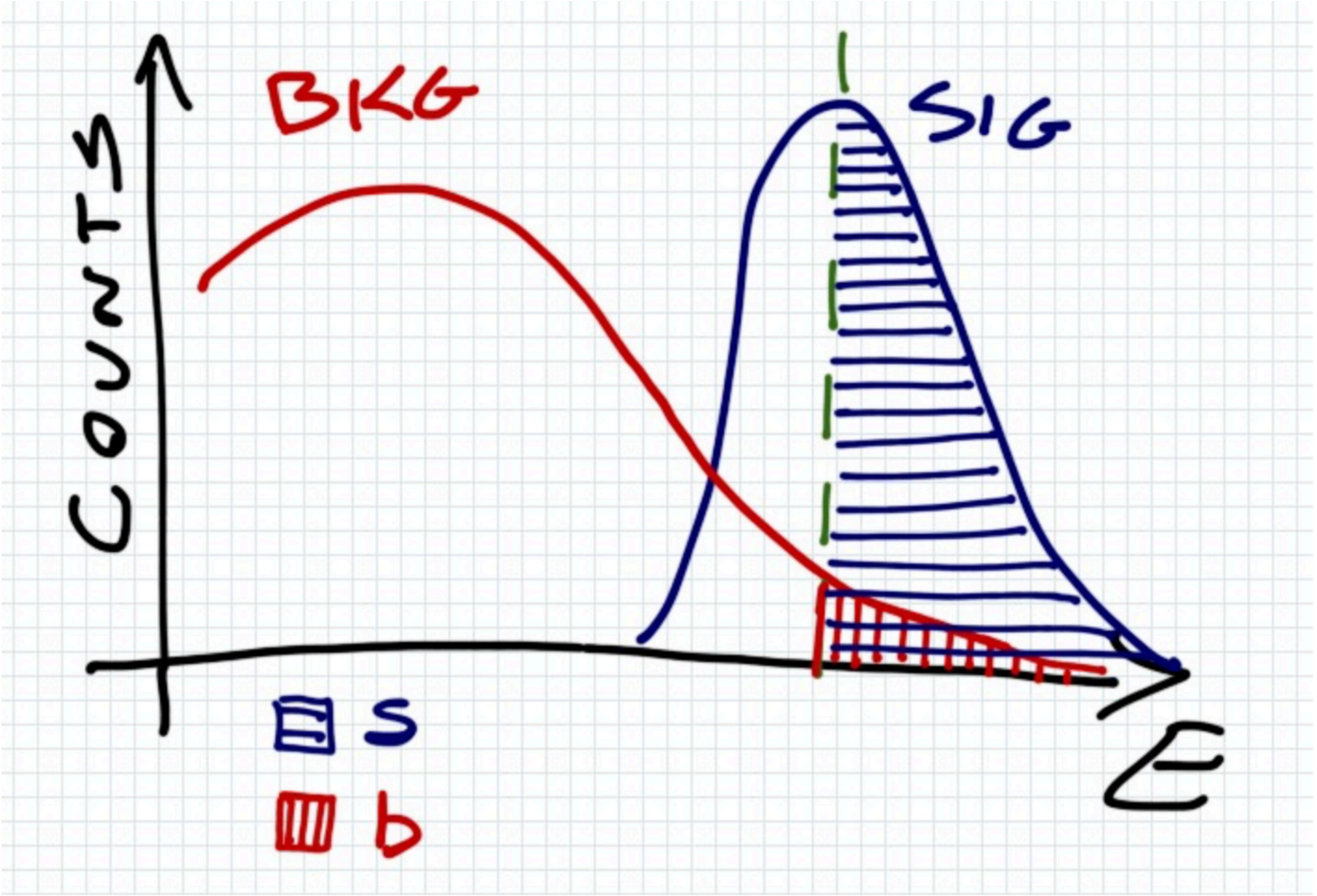
- *A frequentist in HEP would*
  - *build the likelihood*
  - *modify it to include the modeling of nuisance parameters (the systematic uncertainty)*
  - *use profiling to remove the dependence on nuisance parameters*
  - *use a maximum-likelihood estimator of the parameter of interest to report a measurement, etc*
- *A bayesian in HEP would*
  - *build the likelihood*
  - *build the posterior as a product of likelihood and the priors of the nuisance parameters and the parameters of interest*
  - *use marginalization to remove dependence on nuisance parameters*
- *Let' see this in practice with out counting experiment*

F: Building the likelihood

- *The process likelihood is a Poisson function*

$$\mathscr{L} = P(n \mid \lambda_B + \lambda_S)$$

- *The "full likelihood" including systematics has three terms*
  - *The "real" likelihood*
  - *The constraint on the background expected yield.*
  - *Additional systematic uncertainties on the signal (we neglect them here to keep discussion simple)*



$$\mathscr{L}_{HEP} = P(n \mid \lambda_B + \lambda_S) G(\bar{\lambda}_B \mid \lambda_B, \sigma_{\lambda_B})$$

F: Maximum Likelihood estimation

- *We are given a likelihood model* $\mathscr{L}(D|w)$ *and some data D*
  - *D is known,* w *are unknown*
- *We want to find the* $\hat{w}$ *values that would make our data D the most probable outcome of the experiment*
  - *If we knew these* $\hat{w}$ *values, the probability of observing D is maximal (here D is unknown and* $\hat{w}$ *is known)*
  - *You can convince yourselves that*

$$\hat{w} = \arg\max_{w} \mathscr{L}(D|w)$$

F: Maximum Likelihood estimation

- A full likelihood model would depend on the parameter of interest (signal yield, if any) and some nuisance parameter (the amount of background)

$$\mathscr{L}_{HEP}(n \mid \lambda_S) = P(n \mid \lambda_B + \lambda_S) G(\bar{\lambda}_B \mid \lambda_B, \sigma_{\lambda_B})$$

- To learn about the parameter of interest, one has to remove the dependence on nuisance

$$\hat{\mathscr{L}}_{HEP}(n \mid \lambda_S) = \max_{\lambda_B} P(n \mid \lambda_B + \lambda_S) G(\bar{\lambda}_B \mid \lambda_B, \sigma_{\lambda_B})$$

- At that point, one can estimate the parameter of interest

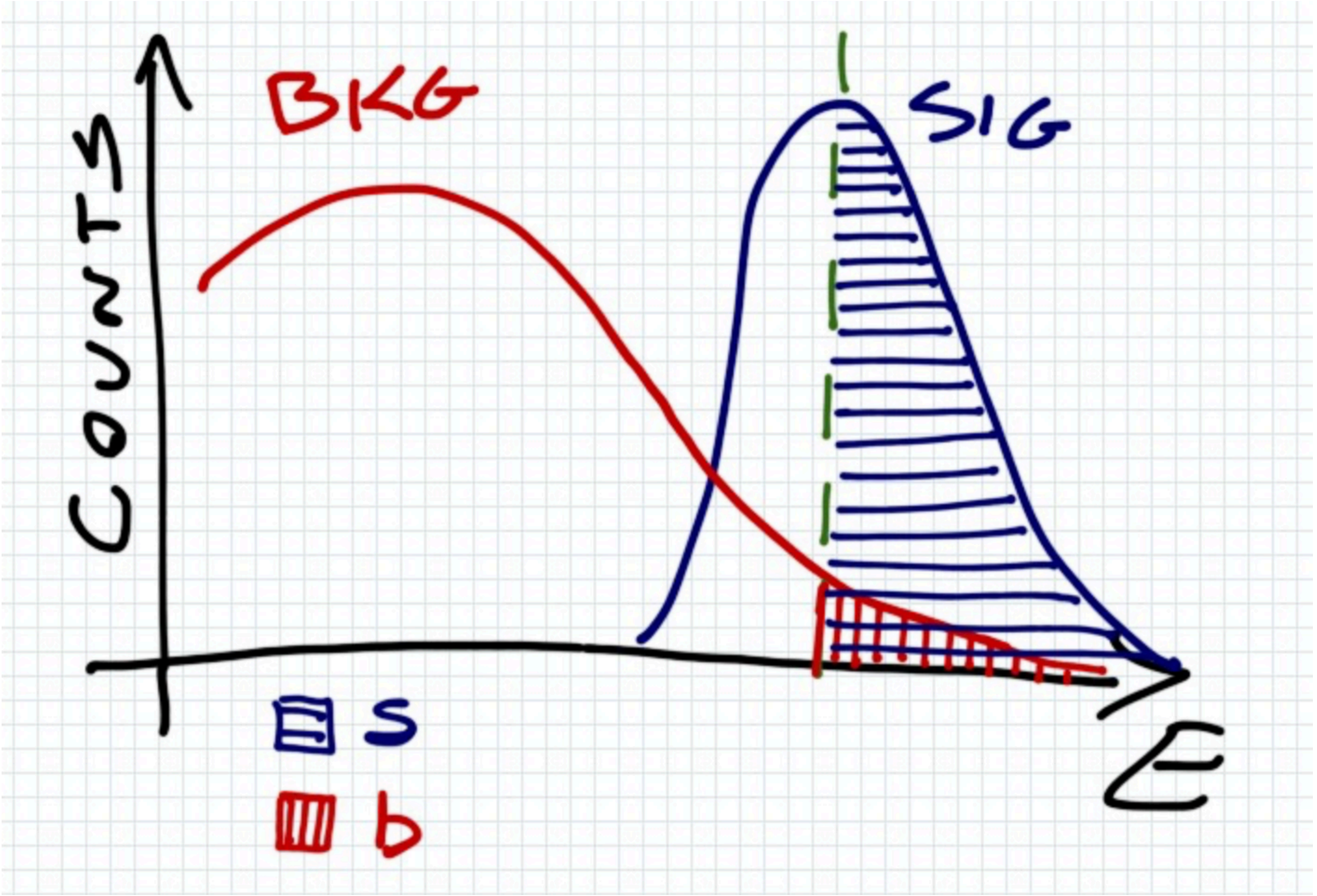
$$\bar{\lambda}_s = \arg\max_{\lambda_S} \hat{\mathscr{L}}_{HEP}(n \mid \lambda_S)$$

B: Building the likelihood

- *The process likelihood is a Poisson function*

$$\mathscr{L} = P(n \mid \lambda_B + \lambda_S)$$

- *The posterior probability density function, including systematics, has three terms*
  - *The "real" likelihood*
  - *The constraint on the signal expected yield*
  - *The constraint on the background expected yield.*



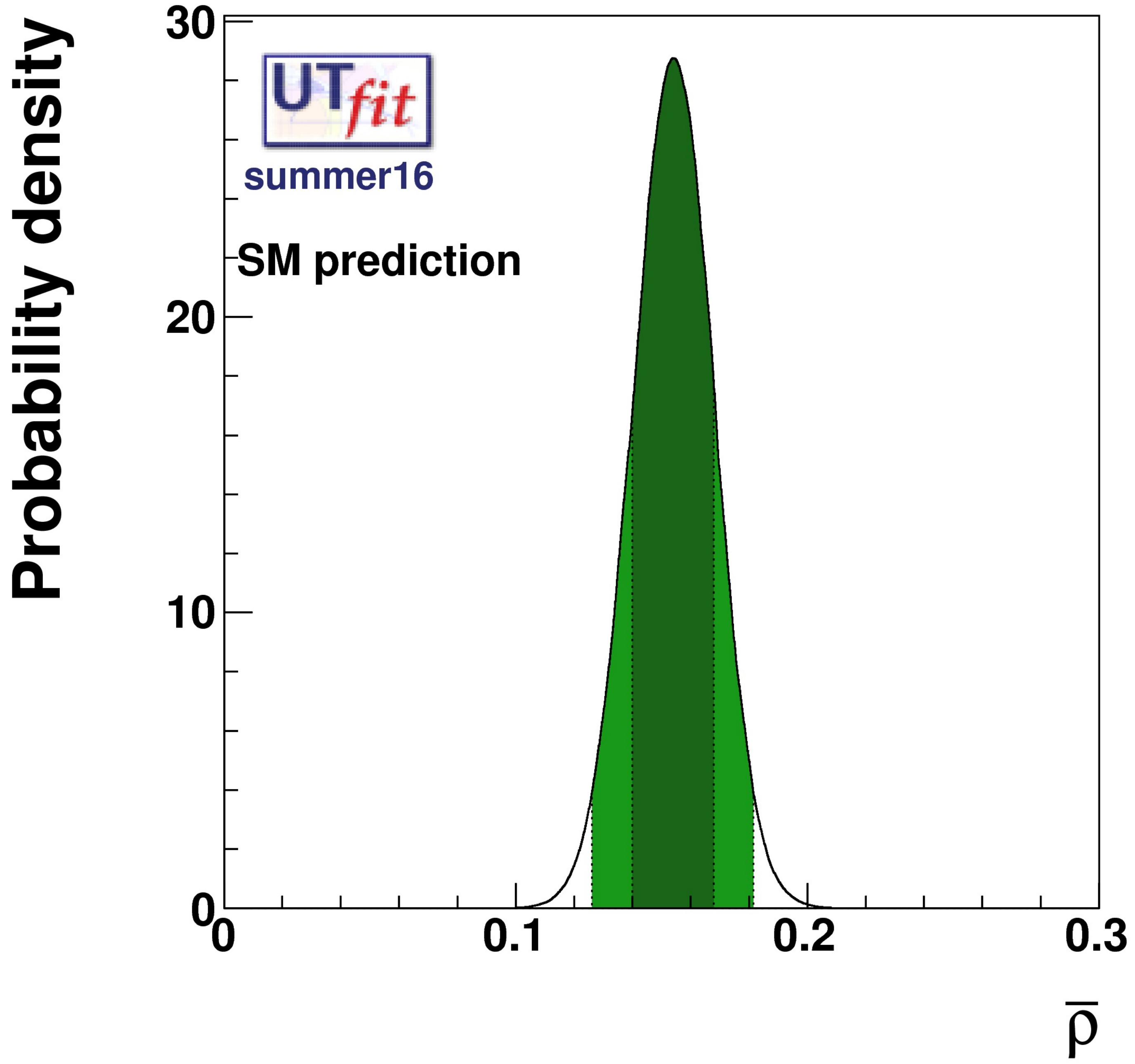
$$P(\lambda_B, \lambda_S \mid n) = \frac{P(n \mid \lambda_B + \lambda_S)\Pi(\lambda_S)G(\lambda_B \mid \bar{\lambda}_B, \sigma_{\lambda_B})}{\int d\lambda_S d\lambda_B P(n \mid \lambda_B + \lambda_S)\Pi(\lambda_S)G(\lambda_B \mid \bar{\lambda}_B, \sigma_{\lambda_B})}$$

Posterior pdf and credibility interval

- *One can remove dependence on nuisance with probability theory, summing over all possible λ_B values, weighted by their probabilities (marginalization)*

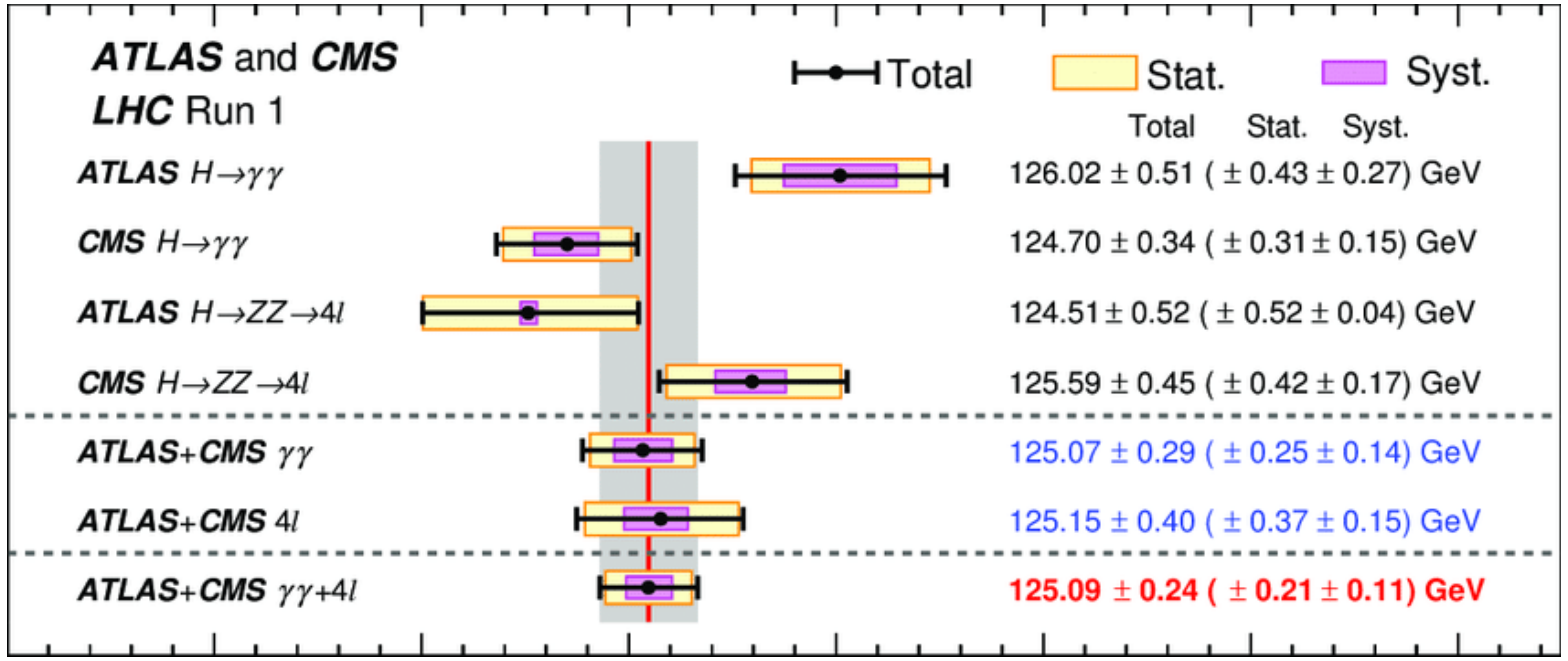
$$P(\lambda_S | n) = \int d\lambda_B P(\lambda_S, \lambda_B | n)$$

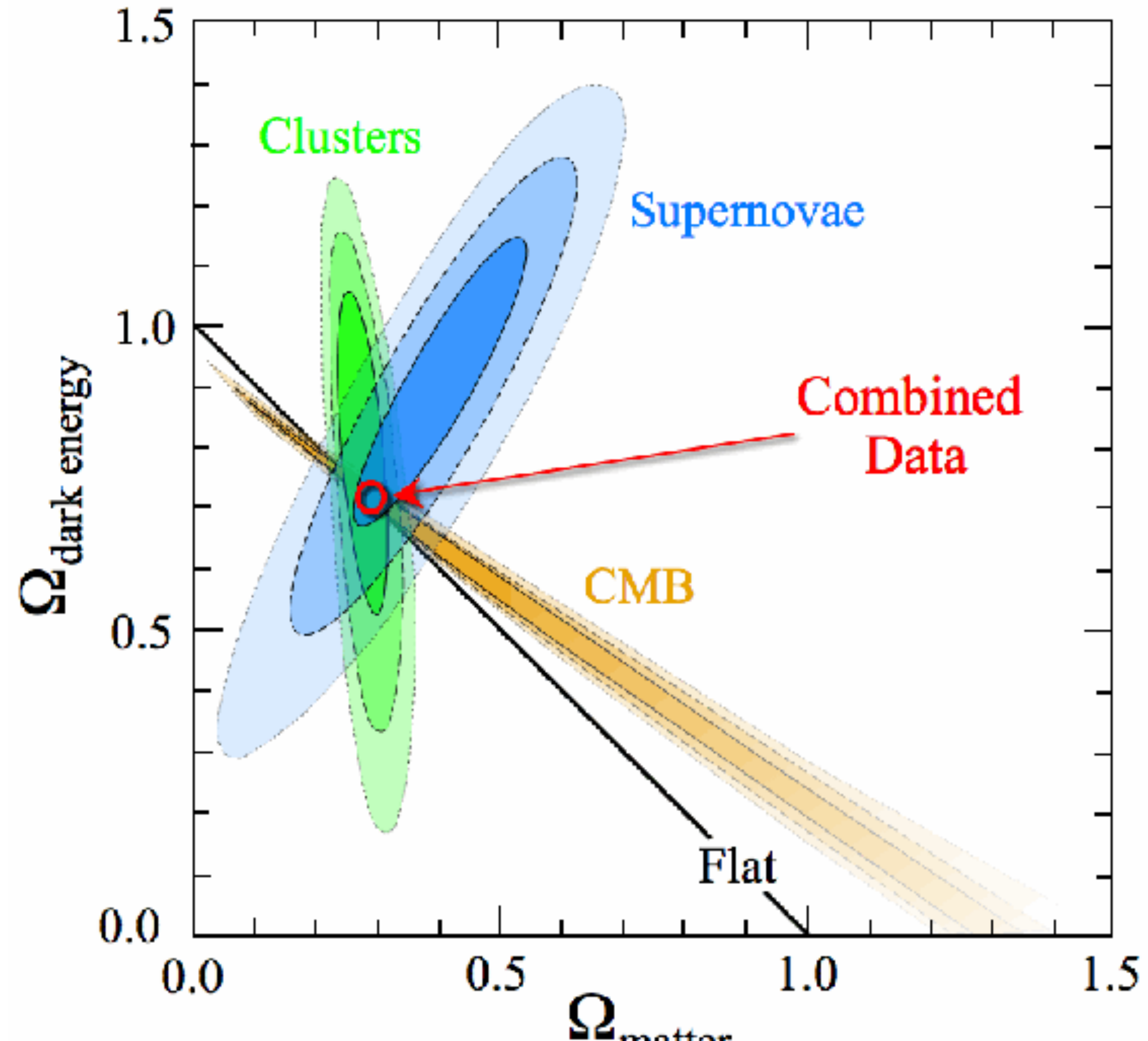
- *At that point, any statement on λ_S can be made using integration of the posterior*
  - *median value*
  - *68% probability region around the median*
  - *etc.*



When to use what

- Frequentist statistics makes a lot of sense when you are facing multiple equiprobable experiments
  - Ideal tool to handle HEP data
  - But keep in mind that any statement is about your experiment, not about the true values of fundamental quantities
- Bayesian statistics is good to invert the information gathered by the experiment into knowledge on true values of fundamental quantities
  - But a prior is needed here. No free lunch





The prior problem

- ***For a Frequentist***
  - *the need to introduce prior knowledge in Bayesian statistics is a problem. A measurement outcome should not depend on who makes the measurement*
- ***For a Bayesian***
  - *the absence of a framework to introduce prior knowledge is a limitation of the frequentist statistics: there is no rigorous way to add information about nuisance parameters and systematic uncertainties*
- ***They are both right***
  - *Frequentists are focused on Bayesians having to model ignorance (what's the prior on something you never measured before? A flat prior? Flat on what? on x, x^2? Mind the Jacobian)*
  - *Bayesians are focused on Frequentists not being able to model knowledge on nuisance parameters (if you know that your signal efficiency for the model you are excluding would have been $(20 \pm 1)\%$, how do you tell your likelihood?)*

The difference in practice

- The real fact is
  - that only a few people attempt Bayesian analyses with HEP data
  - But everyone uses a posterior with flat prior (notice different role of λ_B and $\bar{\lambda}_B$ as arguments in G)

$$\mathscr{L}_{HEP} = P(n \mid \lambda_B + \lambda_S) G(\bar{\lambda}_B \mid \lambda_B, \sigma_{\lambda_B})$$

$$P(\lambda_B, \lambda_S \mid n) = \frac{P(n \mid \lambda_B + \lambda_S)\Pi(\lambda_S) G(\lambda_B \mid \bar{\lambda}_B, \sigma_{\lambda_B})}{\int d\lambda_S d\lambda_B P(n \mid \lambda_B + \lambda_S)\Pi(\lambda_S) G(\lambda_B \mid \bar{\lambda}_B, \sigma_{\lambda_B})}$$

The difference in practice

- The same expression as long as
  - the prior Π is flat
  - The function G is symmetric for exchange of the λ_B and its estimate $\bar{\lambda}_B$ (e.g., a Gaussian, which is what people typically use)

$$\mathscr{L}_{HEP} = P(n \,|\, \lambda_B + \lambda_S) G(\bar{\lambda}_B \,|\, \lambda_B, \sigma_{\lambda_B})$$

$$P(\lambda_B, \lambda_S \,|\, n) \propto P(n \,|\, \lambda_B + \lambda_S) \Pi(\lambda_S) G(\lambda_B \,|\, \bar{\lambda}_B, \sigma_{\lambda_B})$$

- HEP physicists are Bayesians that use profiling on the posterior and think that they are doing frequentist statistics

A Hybrid approach

- *Using profiling on a posterior is considered very bad practice in Bayesian statistics (see paradoxes documented in literature)*
- *As a matter of fact, frequentist's concern is the issue with the prior on the parameter of interest*
- *Professional literature suggests to*
    - *Marginalize the posterior wrt the nuisance parameters*
    - *Apply frequentist methods (e.g., confidence interval derivation) on the marginalised function of the parameter of interest, w/o a prior on it*

$$\mathscr{L}_{hybrid}(n\,|\,\lambda_S) = \int d\lambda_B P(n\,|\,\lambda_B + \lambda_S) G(\bar{\lambda}_B\,|\,\lambda_B, \sigma_{\lambda_B})$$

- *Why don't we do it?*

Integration is hard and we are having a

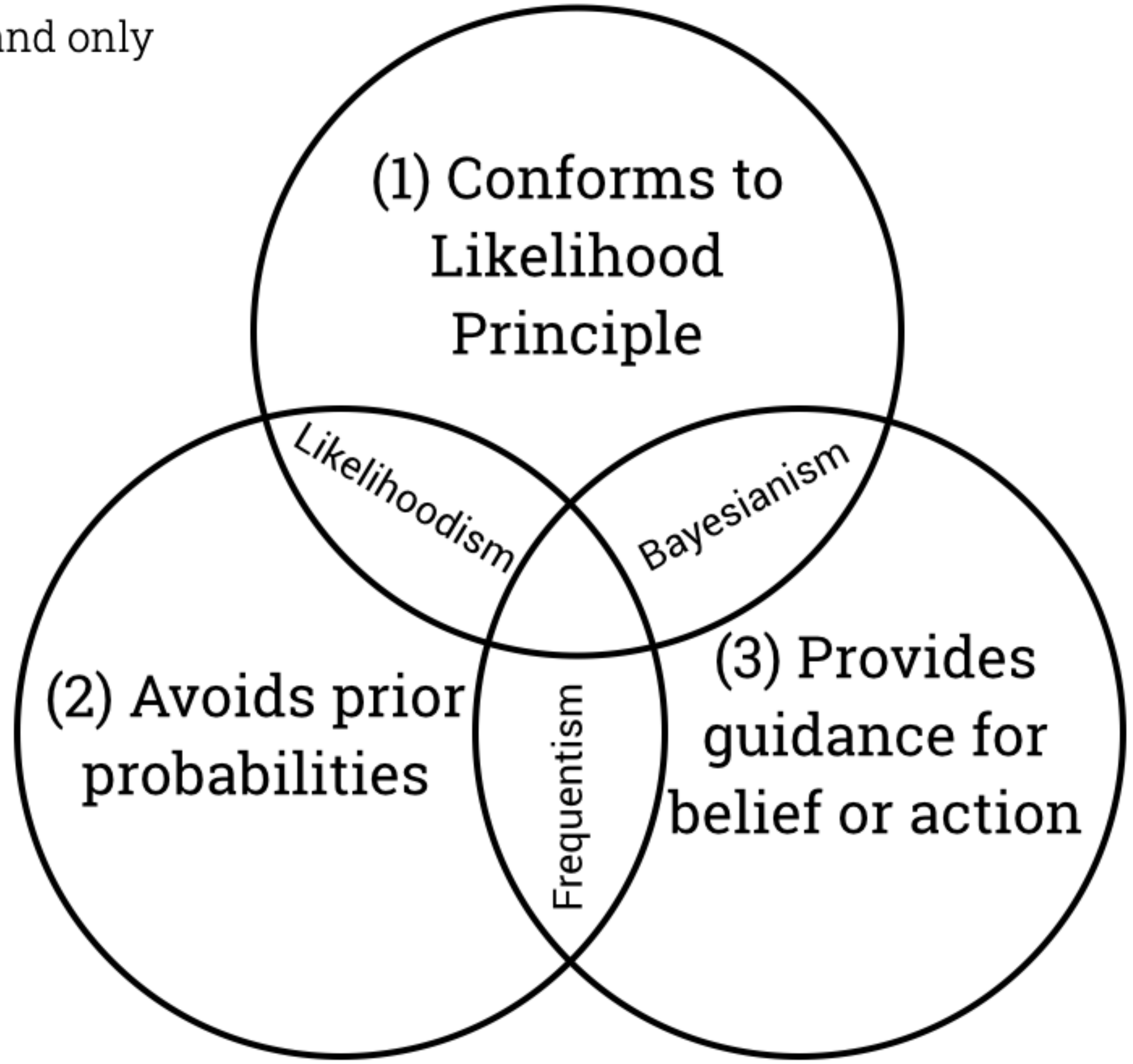
What are we?

https://gandenberger.org/2014/07/28/intro-to-statistical-methods-2/

Likelihoodists use likelihood functions to characterize data as evidence. Their primary interpretive tool is the *Law of Likelihood*, which says that E favors H_1 over H_2 if and only if their likelihood ratio $\mathcal{L} = \Pr(E|H_1)/\Pr(E|H_2)$ on E is greater than 1, with $\mathcal{L}$ measuring the degree of favoring. Two major advantages of this approach are (1) it

Bayesians use likelihood functions to update probability distributions in accordance with Bayes's theorem. Their approach fits nicely with the likelihoodist approach in that the ratio of the "posterior probabilities" (that is, the probabilities after updating on the evidence) $\Pr(H_1|E)/\Pr(H_2|E)$ on E equals the ratio of the prior probabilities $\Pr(H_1)/\Pr(H_2)$ times the likelihood ratio $\mathcal{L} = \Pr(E|H_1)/\Pr(E|H_2)$. The Bayesian approach conforms to

Frequentists use likelihood functions to design experiments that are in some sense guaranteed to perform well in repeated applications in the long run, no matter what the truth may be. Frequentist tests, for instance, control both the probability of rejecting the "null hypothesis" if it is true (often at the 5% level) and the probability of failing to reject it if it is false to a degree that one would hate to miss (often at the 20% level). They violate the



Now that we know everything about statistics…

Summary

- *We introduce basic probability notions to build a probability model (with the specific example of a counting experiment in mind)*
- *We built the likelihood from the probability model*
- *We reviewed how a frequentist and a bayesian would approach the problem*
- *The rest tomorrow…*