# Review

D. Jason Koskinen

koskinen@nbi.ku.dk

*Advanced Methods in Applied Statistics*

*Feb - Apr 2021*

Photo by Howard Jackman

# Exam Notes

- Submission is **both**:
  - A nicely written and composed PDF file devoid of code
    - You can create a latex/Word/OpenOffice/etc. template right now and save yourself time
  - The code you used to generate your results
- If you have problems email me. Worst scenario is you get a reply “I am sorry, but I cannot help you with XXXXXX”.
- Especially for Ph.D. students, if you don’t get an exam link via email, I will post the exam on the course webpage within a few minutes of start time and you can email your exam submission(s): code and PDF write-up.
- Problem 1 has been the basically the same question for every AMAS exam, and it will be again this year too.

# Announcements

- I will not be reviewing everything in the course today
    - Some text-heavy slides are included online, but won’t be covered in class.
- An omitted topic in today’s review may appear on the exam

# Likelihoods

$f()$ is commonly the probability distribution function

- The likelihood is the product of the individual probability (or probabilities for multiple parameters) of parameters (θ) which produce the observed outcomes ($x_i$)

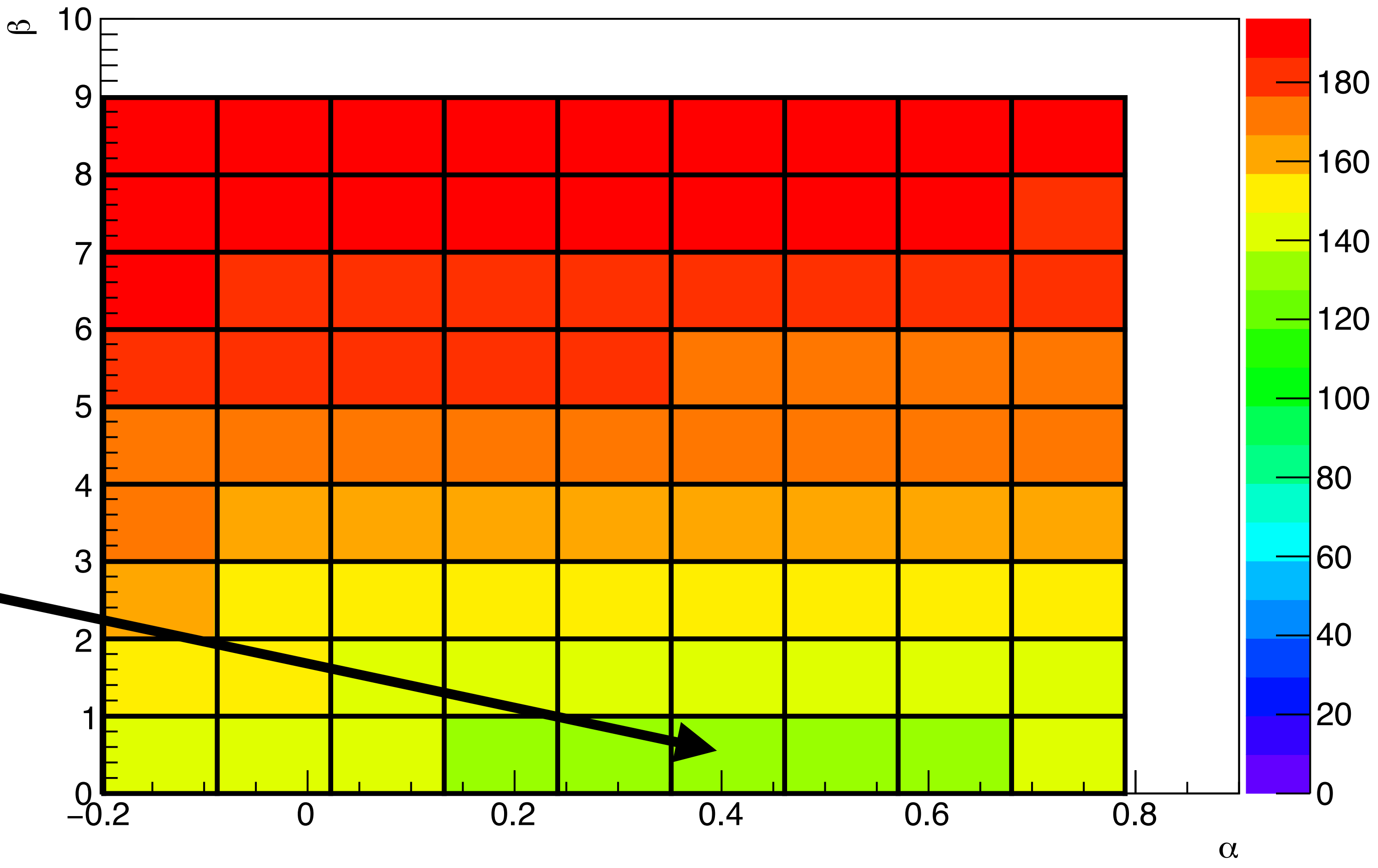
$$\mathcal{L}(\theta) = \prod_{i=0}^{N} f(x_i;\theta)$$

- The likelihood ($\mathscr{L}$ or L) given the observed data ($x_i$) for the parameters (θ) is equal to the probability ($\mathscr{P}$) given the parameters (θ) of getting the observed data ($x_i$)

$$\mathcal{L}(\theta|x) = P(x|\theta)$$

# Raster Scan

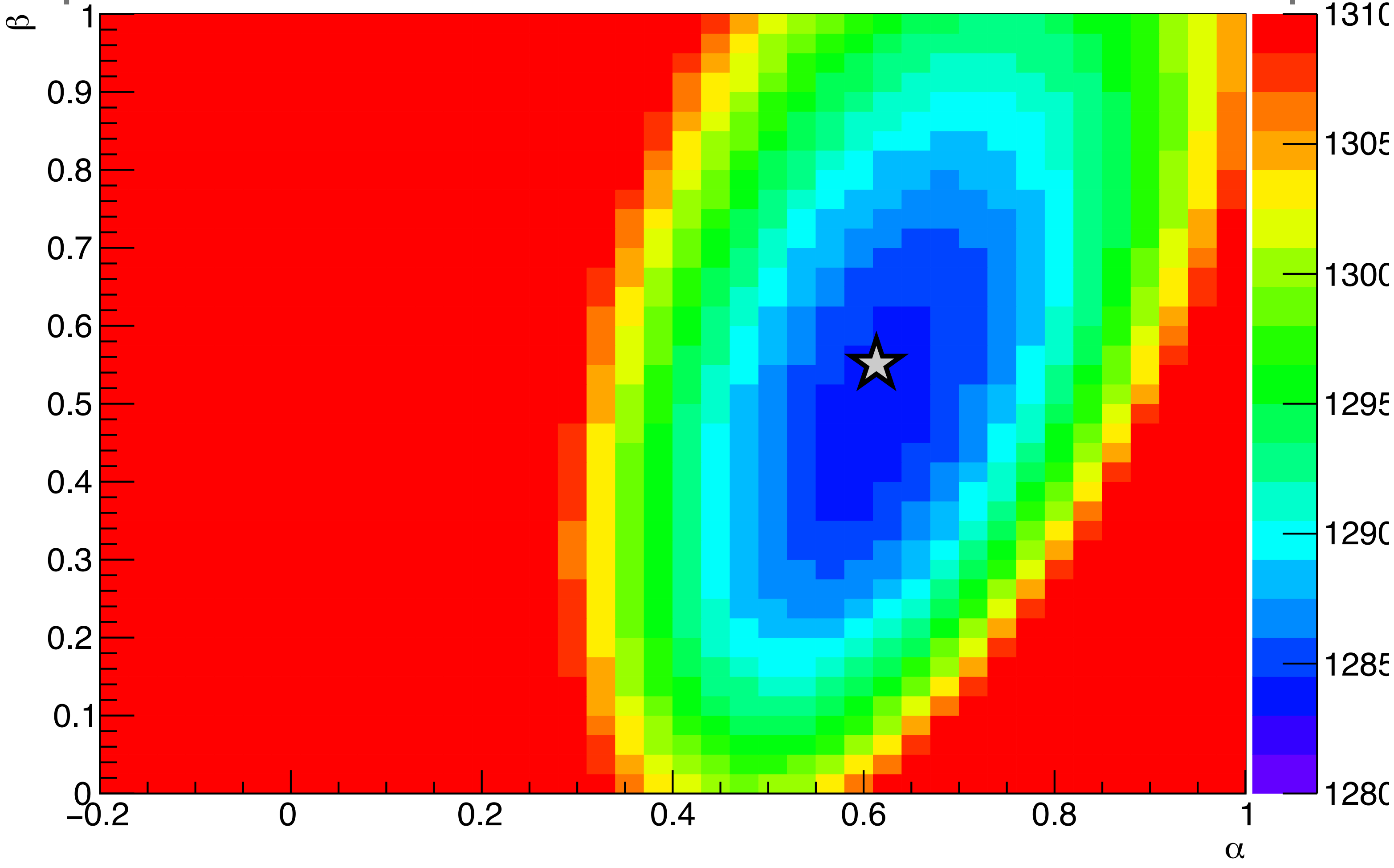
- This is a semi-coarse sampling of the LLH space. Establish which region(s) of the scanned parameter values have the best LLH and start your fit there, or at multiple points near the best LLH.



Start somewhere around here

# Exercise 3 cont.

- Likelihood landscapes are important to visualize and understand… super important. Plot them whenever possible to understand the topology that your minimizer encounters
- For values of **α**=0.6 and **β**=0.5 for the previous formula/PDF make a 2D plot of the likelihood or LLH landscape



Zoomed in

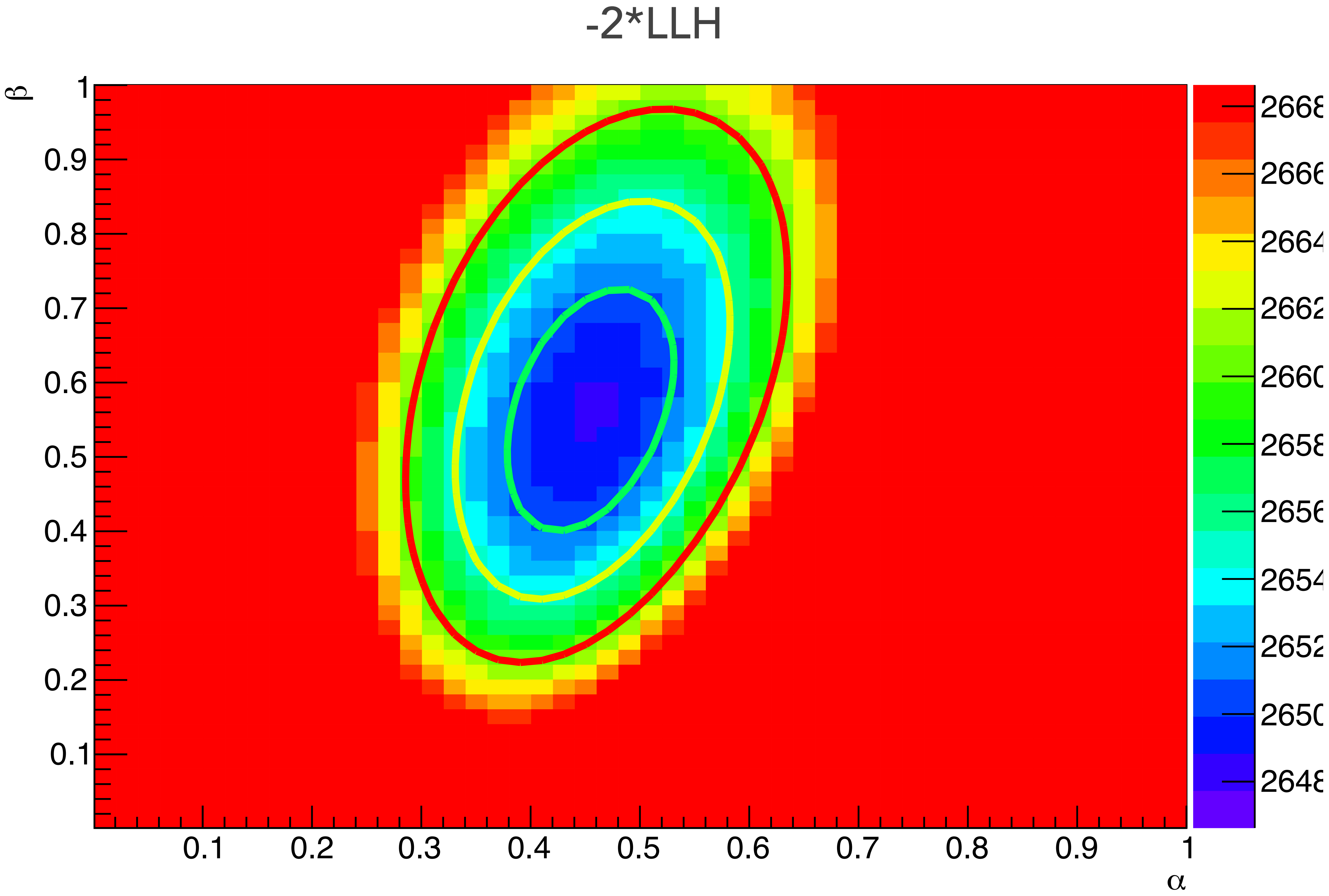
# ln(Likelihood) and 2*LLH

- A change of 1 standard deviation (σ) in the maximum likelihood estimator (MLE) of the parameter θ leads to a decrease in the in the ln(likelihood) of 1/2 for a gaussian distributed estimator
    - Even for a non-gaussian MLE, the 1σ region defined as LLH-1/2 is a good approximation
    - Because the regions defined with ΔLLH=1/2 are consistent with common $\chi^2$ distributions multiplied by 1/2, we often calculate the likelihoods as 2*LLH
- Translates to >1 parameters too, with the appropriate change in 2*LLH confidence values
    - 1 parameter, Δ(2LLH)=1 for 68.3% C.L.
    - 2 parameter, Δ(2LLH)=2.3 for 68.3% C.L.

# Variance/Uncertainty - Using LLH Values

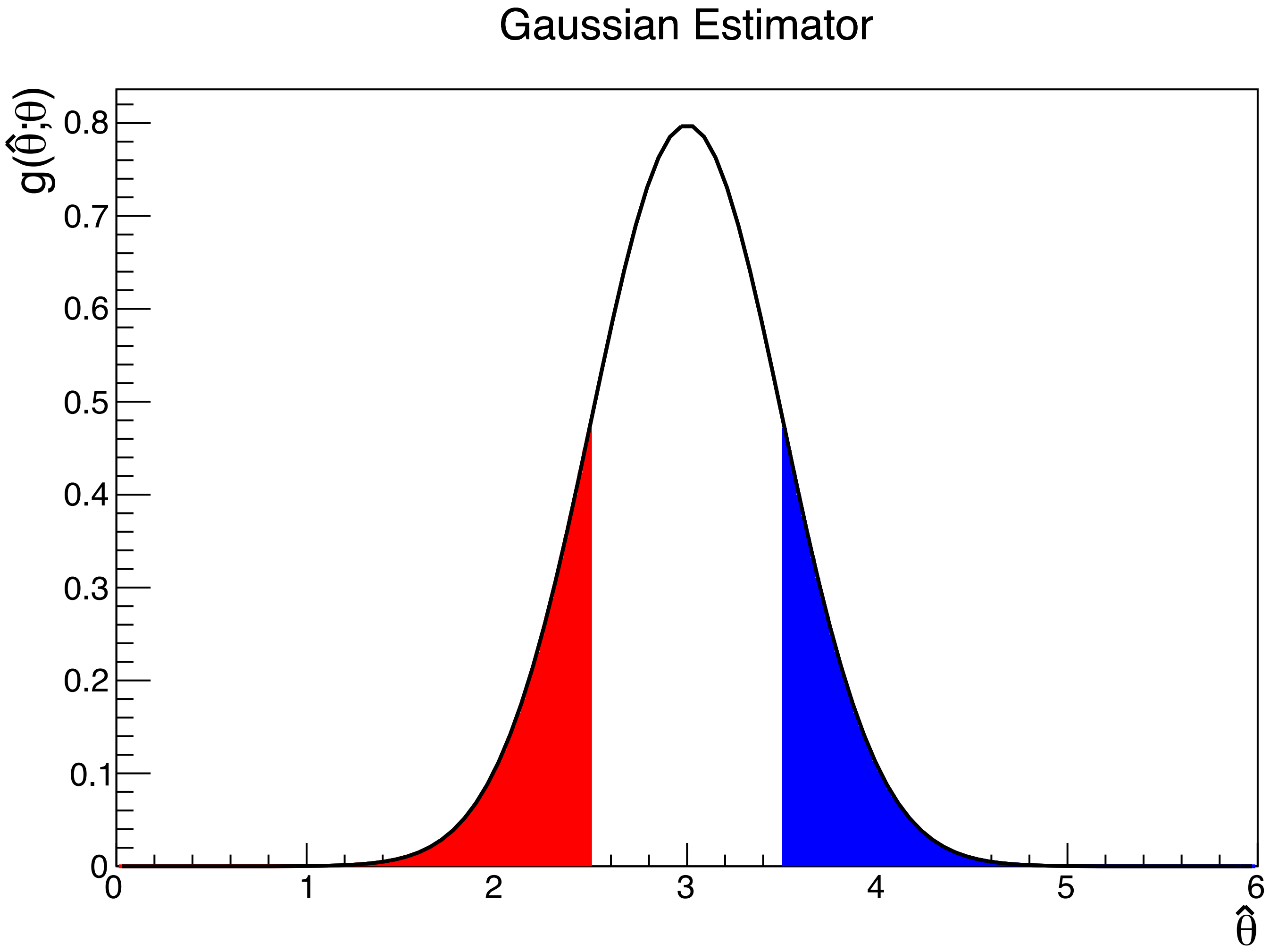
- The LLH (or -2*LLH) landscape provides the necessary information to construct 2+ dimensional confidence intervals, if the respective MLEs are gaussian or well-approximated as gaussian
- Some minimization programs will return the uncertainty on the parameter(s) after finding the best-fit values
    - The .migrad() call in iminuit
    - It is possible to write your own code to do this as well

# Contours on Top of the LLH Space



# Confidence Intervals

- Confidence intervals are often denoted as C.L. or "Confidence Limits/Levels"
- Central limits are different than upper/lower limits
- We can establish uncertainties on our extracted best-fit parameters using likelihoods (hooray!)



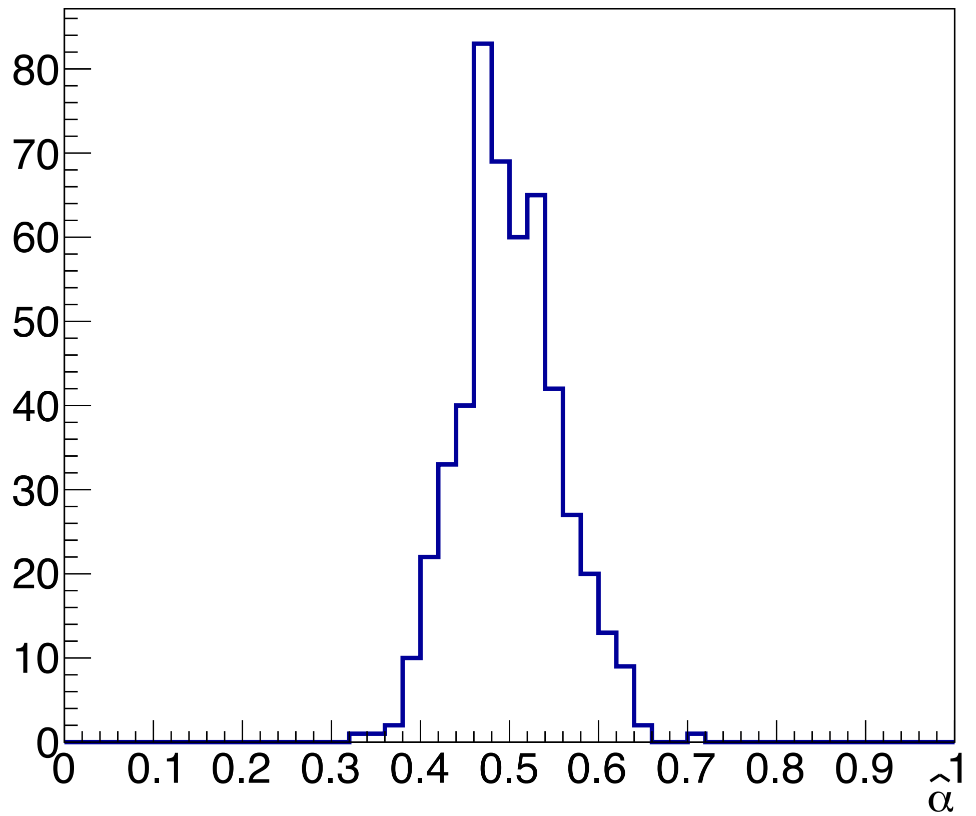
# Brute Force

- If we either did not know, or did not trust, that our estimator(s) are nicely analytic PDFs (gaussian, binomial, poisson, etc.) we can use our pseudo-experiments to establish the uncertainty on our best-fit values
  - Using original PDF, sample from original PDF with injected values of $\hat{\alpha}_{obs}$ and $\hat{\beta}_{obs}$ that were found from our original ‘fit’
  - Fit each pseudo-experiment
  - Repeat
  - Integrate ensuing estimator PDF

To get ±1σ central interval

$$\frac{100\% - 68.27\%}{2} = \int_{C_+}^{\infty} g(\hat{\alpha}; \hat{\alpha}_{obs}) d\hat{\alpha}$$

$$\frac{100\% - 68.27\%}{2} = \int_{-\infty}^{C_-} g(\hat{\alpha}; \hat{\alpha}_{obs}) d\hat{\alpha}$$



# Beyond Parameter Estimation

- Often we want to know if our model fits the data, or vice versa, where we find ourselves in the realm of wanting to test one hypothesis against another
  - Is my event signal or background?
  - In comparison to model $H_1$ can an alternate model $H_0$ be excluded as incompatible with the data?

# Maximum Likelihood Ratio

- An very common test-statistic for the likelihood ratio is:

$$\Lambda(\theta, x_{obs}) = -2\ln\frac{\mathcal{L}(\theta_0|x_{obs})}{\mathcal{L}(\hat{\theta}|x_{obs})}$$

  - Difference between the null hypothesis in the numerator and the alternative hypothesis in the denominator is that the null hypothesis has a **fixed value** of one (or more) of the θ parameters whereas the alternative hypothesis **fits/maximizes** the parameter.

- For a normal distributed, i.e. gaussian, variable the ratio follows the $\chi^2$ distribution,
  - $N_{DOF}$ = difference in dimensionality between the models
  - Also requires that Wilk's Theorem is satisfied

# Wilk's Theorem… Kinda

- As the number of data points approaches infinity, the LLH ratio converges to a $\chi^2$ distribution if $H_0$ is true

$$\Lambda(\theta, x_{obs}) = -2\ln\frac{\mathcal{L}(\theta_0|x_{obs})}{\mathcal{L}(\hat{\theta}|x_{obs})}$$

- But there are regions where the gaussian, and therefore Wilk's and our use of $\chi^2$, breaks down:
    - **Low** number of events where the probability switches from gaussian to poisson
    - **Bounds** on the model parameters, e.g. as n→infinity the parameter does not smoothly vary, but has some truncation or discrete behavior
    - Parameters that have a **near-infinite** variance
    - The null and alternate models are **not** nested

# Bayesian

# Transition to Bayes

- The maximum likelihood approach is both effective and powerful, but does not necessarily take into account any preferences or prior information that may produce a more informed or accurate result
- Thankfully, we have Bayes theorem and Bayesian statistics which make explicit use of prior information
- Bayesian probabilities and statistics can encode an amount of belief in (data, model, systematics, hypothesis, parameters, etc.)

# Bayes' Theorem

- We have Bayes' theorem

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

- or sometimes

(Discrete)

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

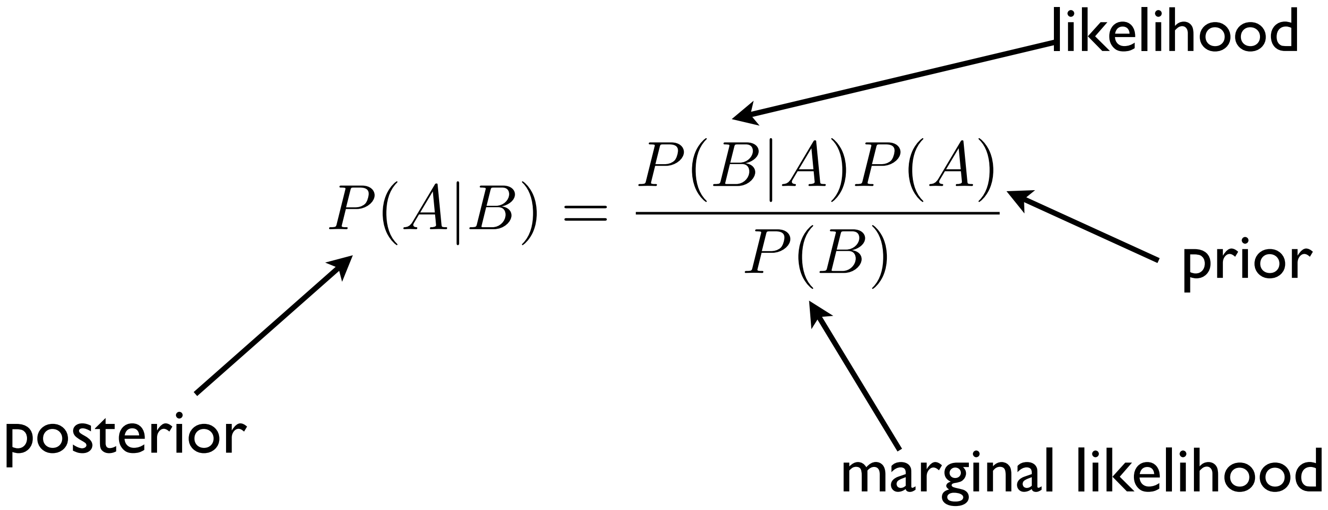
(Continuous)

$$\frac{P(B|A)P(A)}{\int P(B|A)P(A)dA}$$

- Let B be the observed data and A be the model/theory parameters, then we often want the P(A|B); the posterior probability distribution conditional on having observed B.

# Bayes for Parameter Estimation

- We apply prior information not just for discrete probabilities, but for probability distributions as well
- Remember that for Bayesian analyses we include all possible values of the parameter, i.e. θ
    - This means for the PDF, it will **not** be calculated at a single value of θ, but over a suitable range



likelihood

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

prior

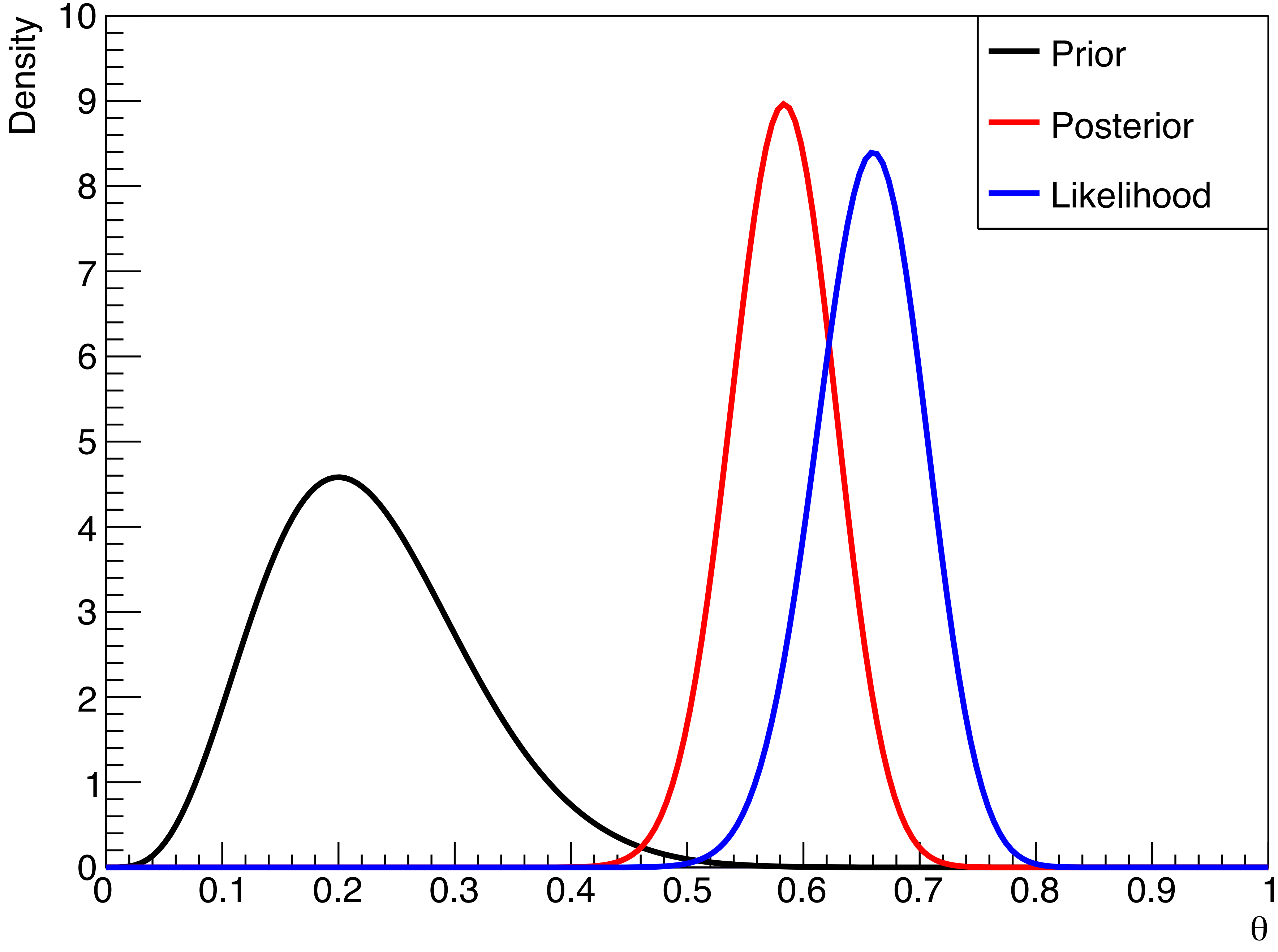
posterior

marginal likelihood

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

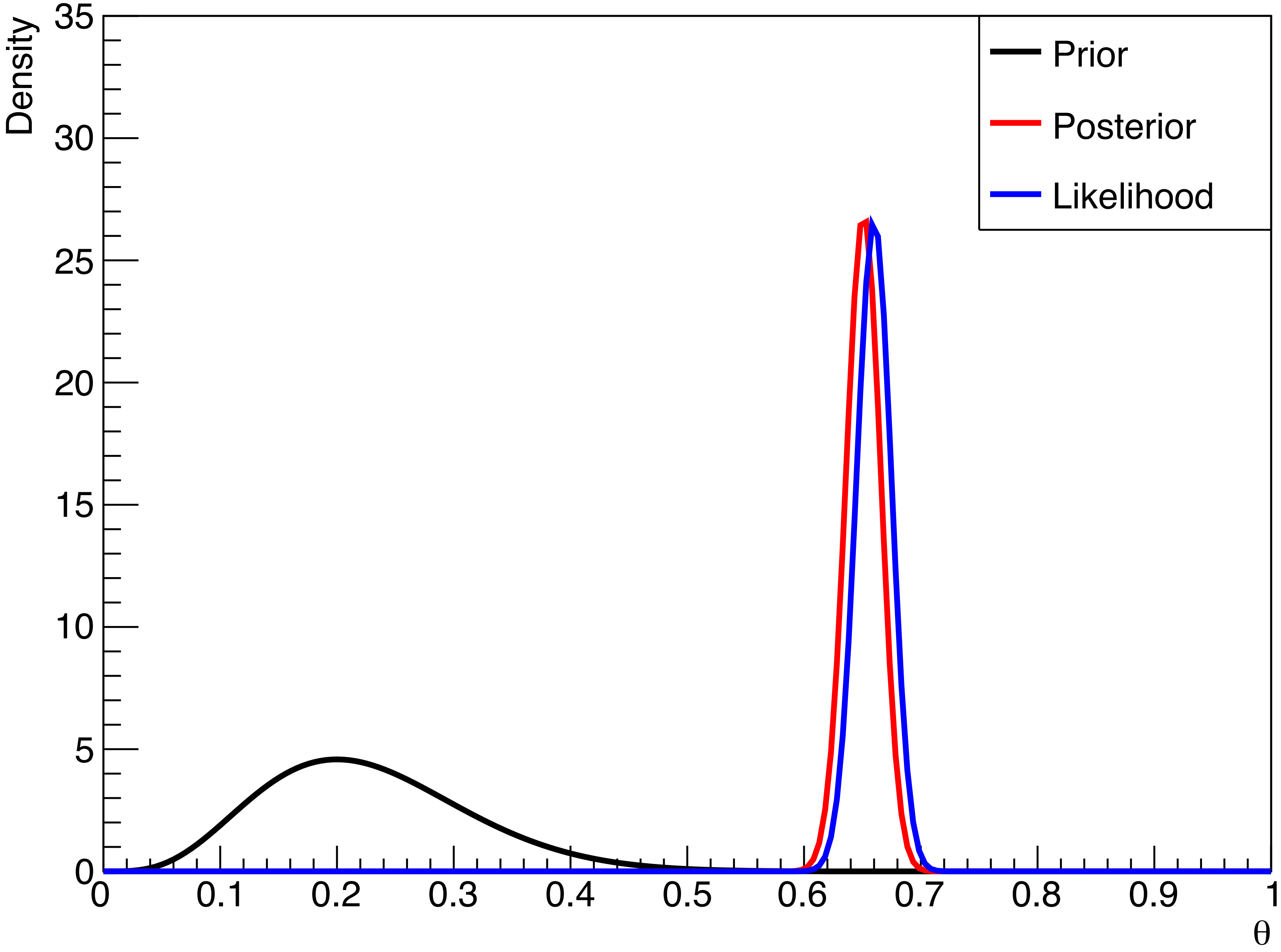
# Exercise #1 plot (from MCMC Lecture)

- Coin flipping bias with n throws/flips, but now in Bayesian style where we want the prob. of coming up heads (θ)



# Exercise #1 (cont.)

- With 10x more statistics, an obvious feature pops up, i.e. that as n→infinity the *maximum a posterior* (MAP) approaches the *maximum likelihood estimator* (MLE)



# Numerical Limitations

- The previous example had only 1 parameter (θ) and 1 prior. When dealing with more parameters, the computational load approx. increases exponentially with the number of parameters.
  - For summation, or integration via Monte Carlo sampling, the number of points (n) grows as $\mathcal{O}(n^d)$ if n points are used to cover each parameter (d)
  - It's possible to tune the number of scan or Monte Carlo points, but then the number of points necessary for calculation is the product of the number of points:

$$\prod_{i=1}^{d} n_i$$

# Markov Chains for Bayes' Stuff

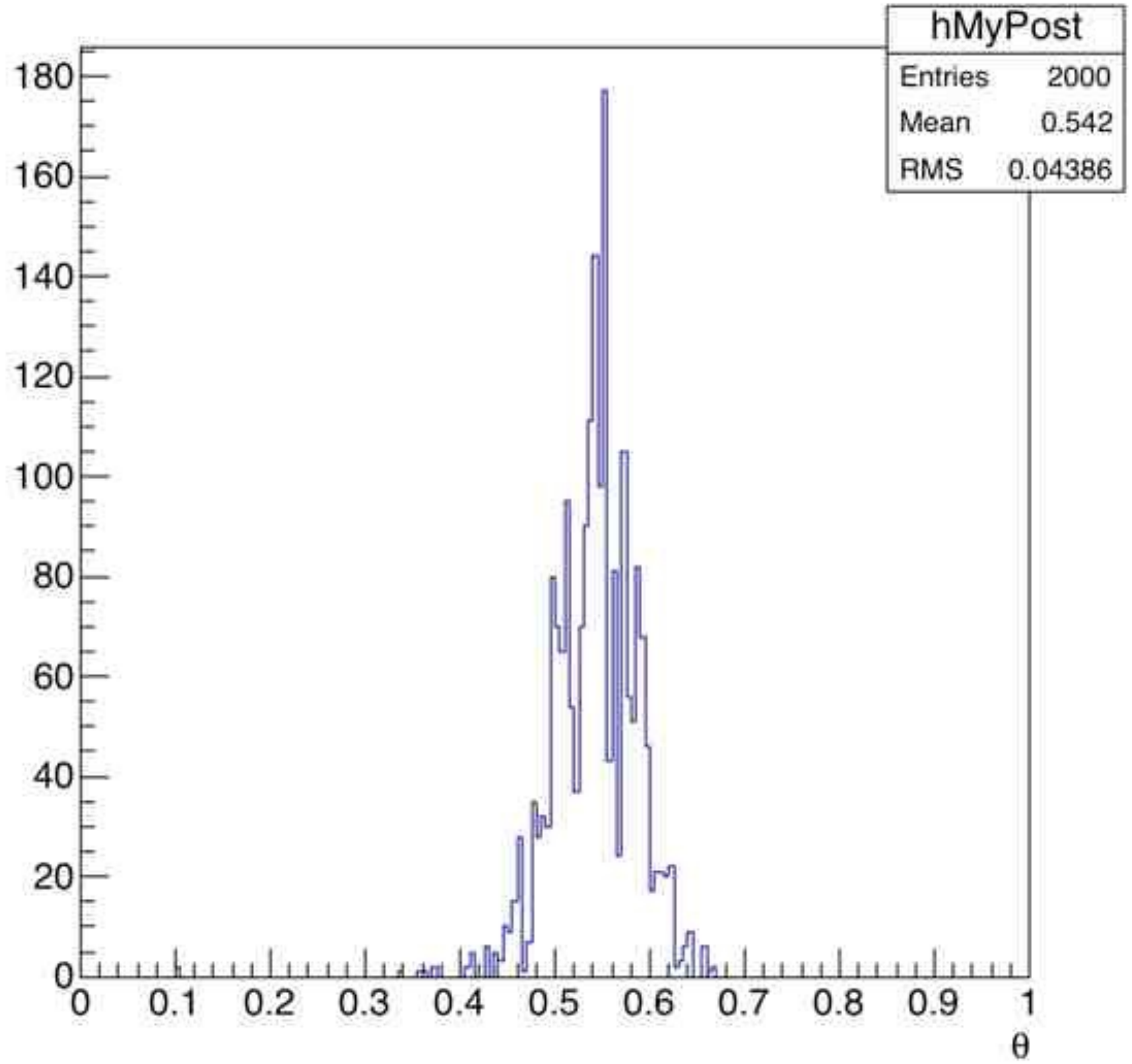
- So how does a Markov chain help with establishing Bayesian posterior distributions?
- Markov chains will asymptotically approach a stable distribution, and we can give the Markov chain a distribution that is representative of the posterior. Remember that,

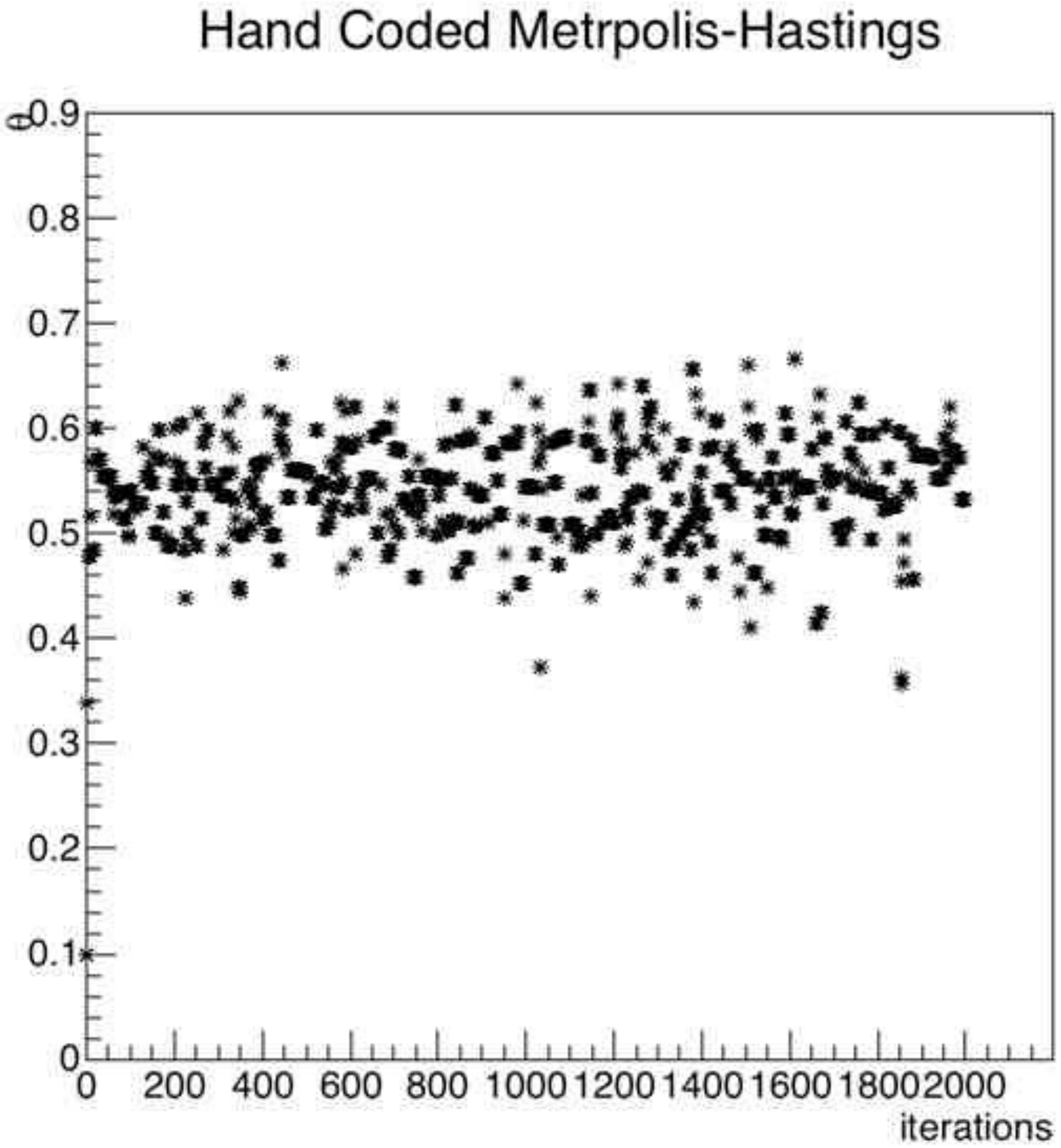
$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

- So using Markov Chain Monte Carlo, the chain can start at points that are not typical of the actual posterior (which we may not know well), but after enough Monte Carlo iterations it should converge to the posterior
- Markov Chain Monte Carlo is the solution

# Exercise #3 (cont.) (from MCMC lecture)

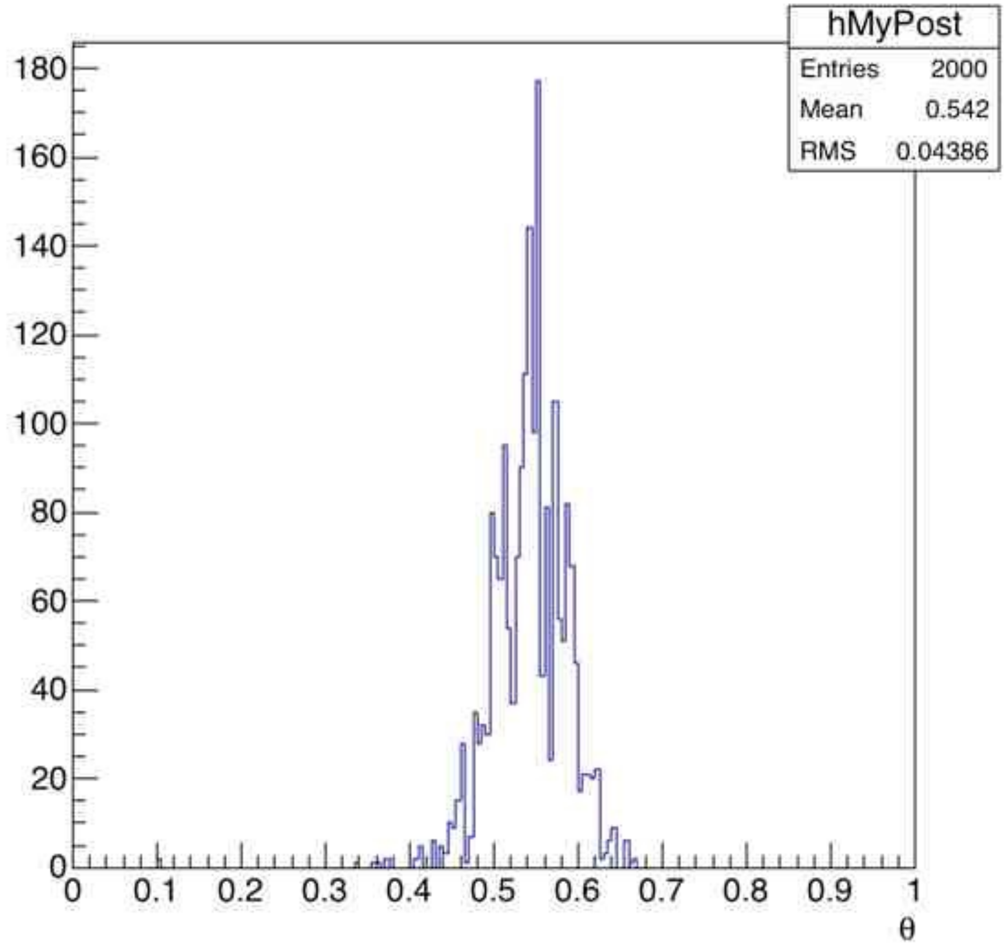
- For 2000 iterations plot Markov Chain Monte Carlo samples as a function of iteration, as well as a histogram of the samples, i.e. the posterior distribution.

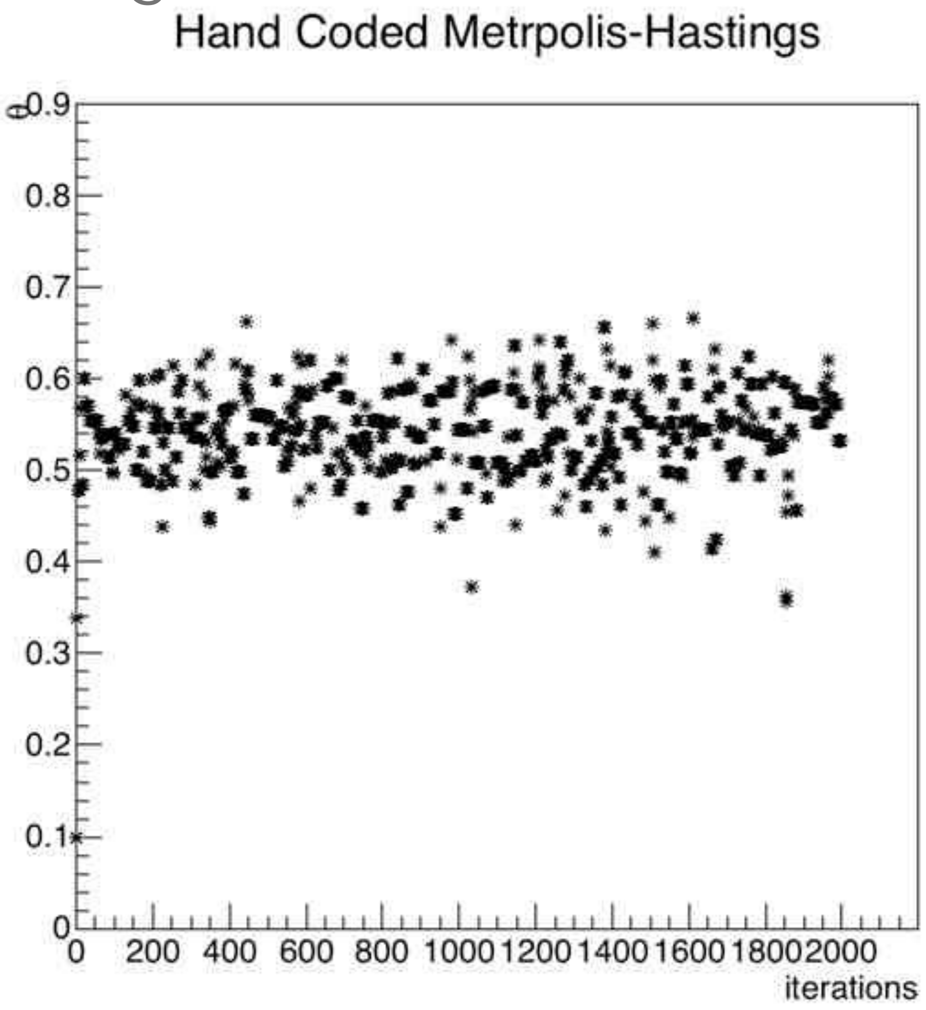




# Why the Posterior?

- The posterior distribution in the Bayesian framework provides not only the most likely value of our parameter of interest, i.e. the **maximum a posterior value**, but also the **uncertainty**. The width of the posterior gives the parameter uncertainty.
- For the example below, if 68.3% of the posterior MCMC iterations occur from 0.5 to 0.59, then that is the 1σ uncertainty range.





# Bayesian Complication

- Unlike the maximum likelihood approach, where we normally just have to know the -2*LLH value which can be converted to a probability, the Bayesian approach can be more resource intensive
- In order to get a 5σ confidence limit, we need approx. 1.7M stable posterior points/iterations

# Smoothing, Interpolating, and Estimation

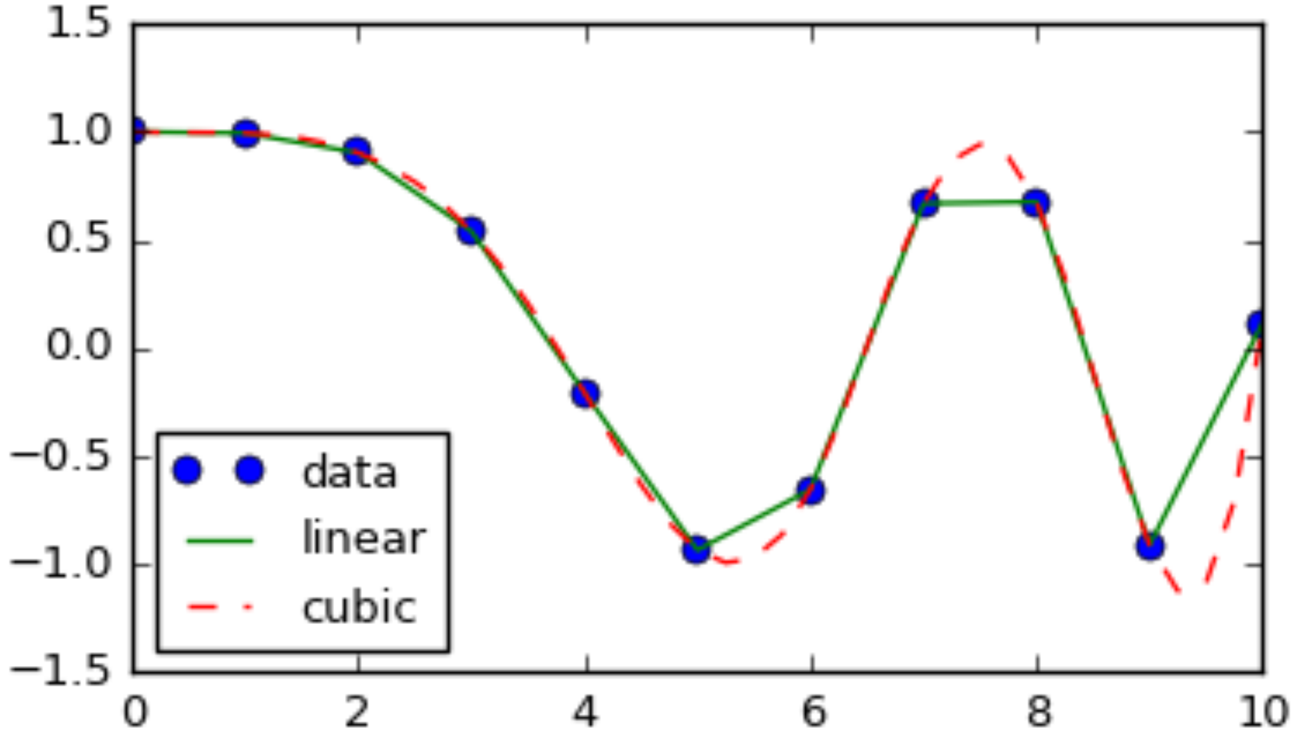
# -

# Splines and Kernel Density Estimation

# Common Spline Types

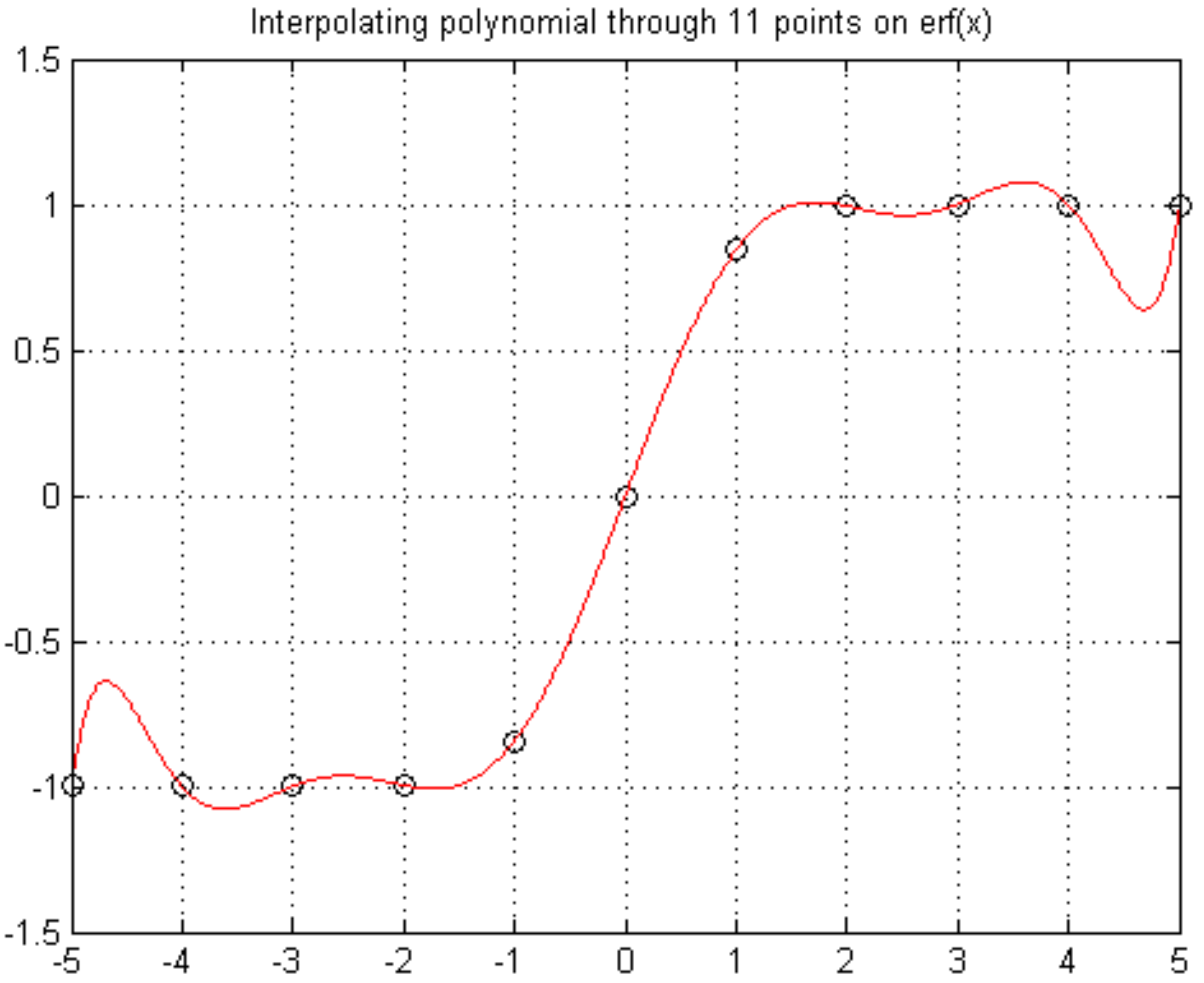
*Scipy interpolate

- Linear splines are continuous across the data points, but do not match the 1st or 2nd derivative at the knots
- Quadratic splines (not shown) match the 1st derivative but not necessarily the 2nd
- Cubic splines are continuous and match the 1st and 2nd derivative at the knots
- Hermite splines - Continuous cubic splines matching the 1st derivative but not necessarily the 2nd

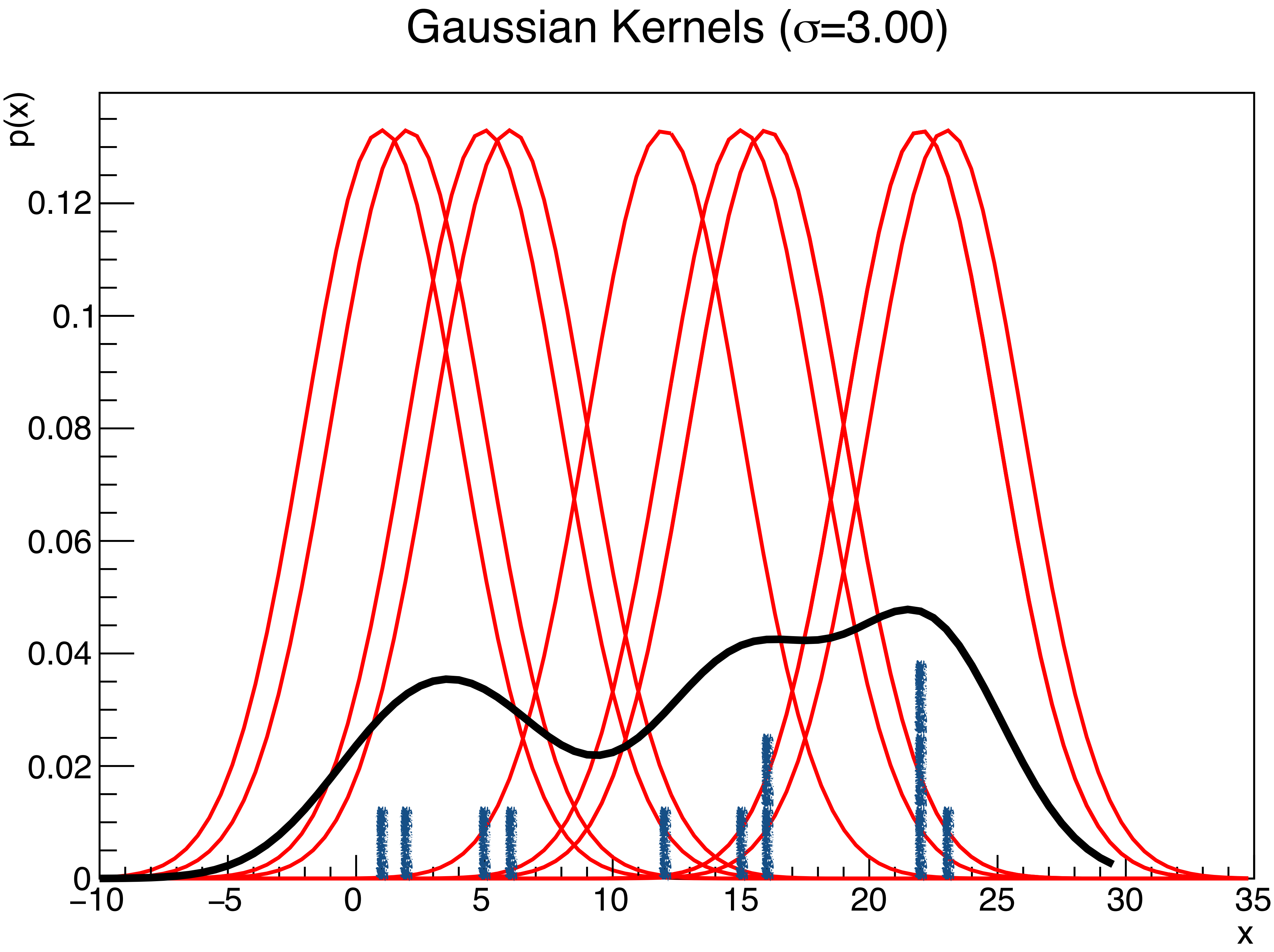


# Polynomial Interpolation

- A problem referred to as 'ringing' is pronounced in polynomial interpolations.



# Data Driven Density Estimation



Data = [1,2,5,6,12,15,16,16,22,22,22,23]

# Kernel Density Estimator

- The generic KDE expression can be expressed as:

$$P_{KDE}(\vec{x}) = \frac{1}{N}\sum_{n=1}^{N} K(\vec{x})$$

- A gaussian kernel is:

$$K(\vec{x},\sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^{D}} e^{-\frac{||\vec{x}-\vec{x}_n||^2}{2\sigma^2}}$$

- The kernel at each data point contributes a non-zero probability from [-∞,+∞] smoothly with decreasing weight as a function of distance
    - Each data point and corresponding kernel integrate to 1 over the whole parameter space

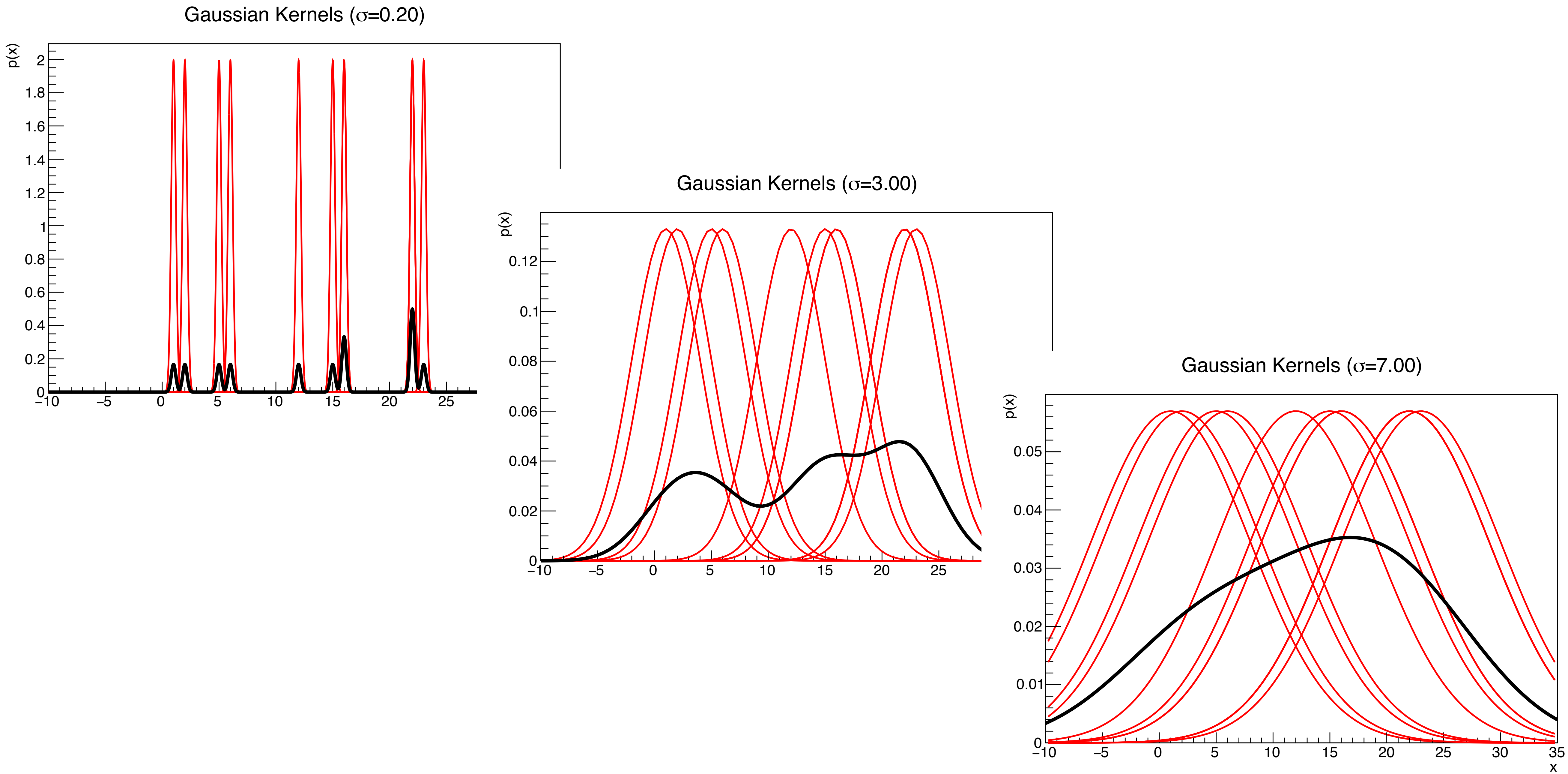
# Comment on KDE Normalizations

$$P_{KDE}(\vec{x}) = \frac{1}{N}\sum_{n=1}^{N} K(\vec{x}) \qquad K(\vec{x},\sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^D} e^{-\frac{||\vec{x}-\vec{x}_n||^2}{2\sigma^2}}$$

- The 1/N normalizes the KDE for the number of events
- No normalization terms in this kernel choice depend on values of $\vec{x}$
- The kernel is **always** normalized to 1

# Kernel Bandwidth

- Every KDE  is, unfortunately, strongly influenced by the kernel bandwidth, which is a user defined free parameter


Gaussian Kernels (σ=0.20)

Gaussian Kernels (σ=3.00)

Gaussian Kernels (σ=7.00)

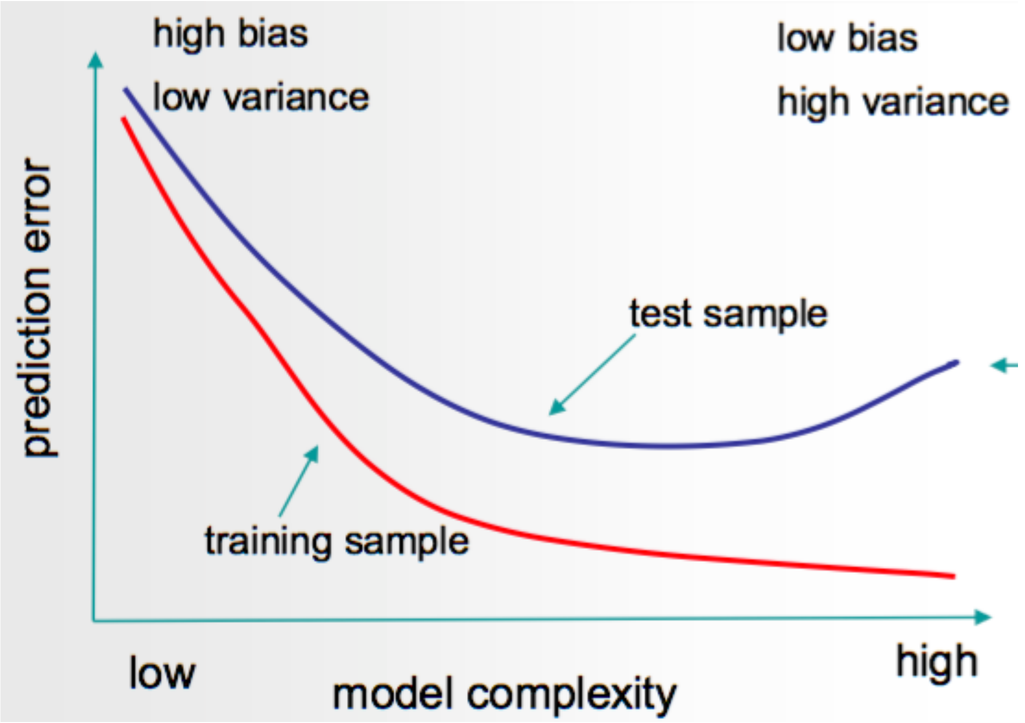
# Multivariate Method and Boosted Decision Tree

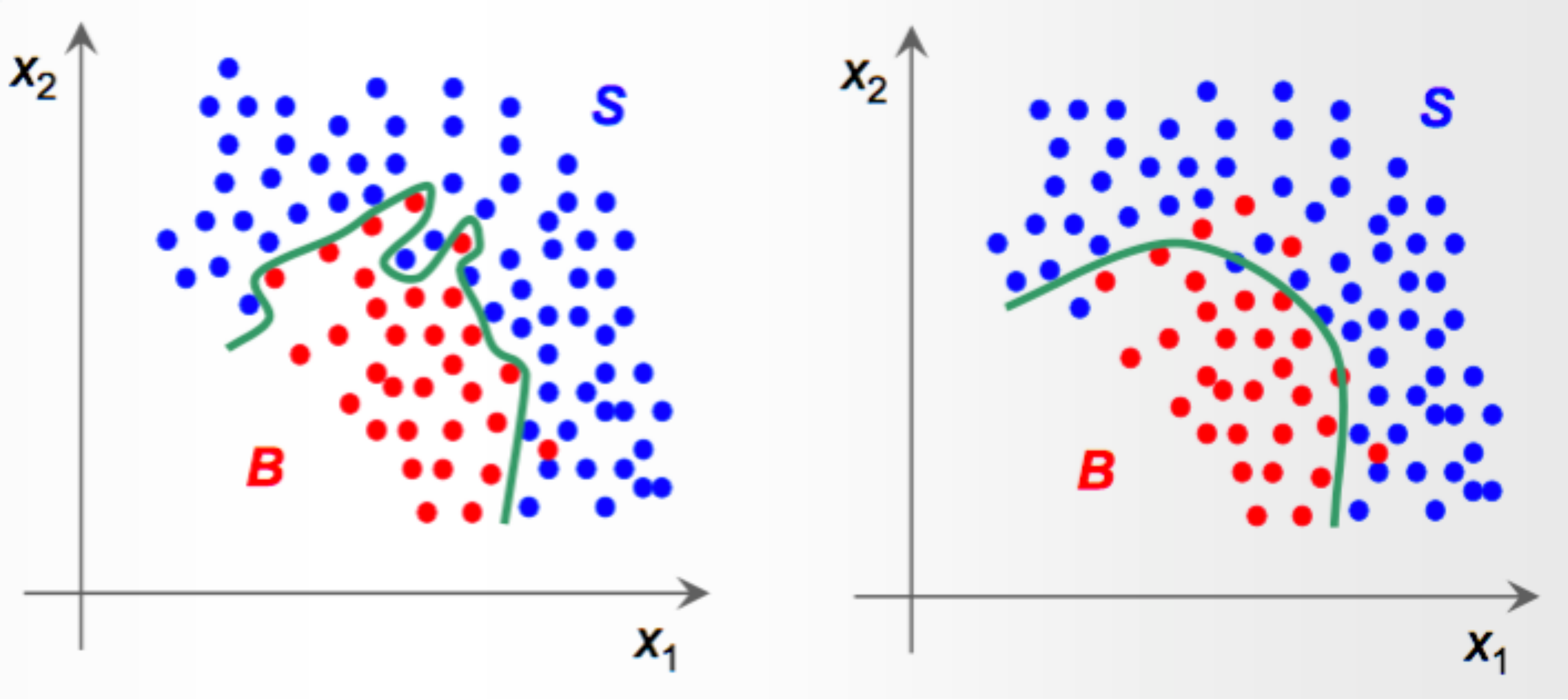
# “Simple” Problems

- Using likelihoods to separate background from signal is not always feasible
  - Likelihood may be too complicated for analytic or Monte Carlo evaluation
  - High dimensionality makes Monte Carlo computationally expensive
- Data sets which are linearly separable in variables, e.g. between signal and background, have useful tools for doing such a separation (Fisher Discriminant)
- For linear and non-linear classification scenarios and/or where the available separators are weak, there is a class of multivariate tools
  - k-Nearest Neighbor
  - Random Forest
  - Artificial Neural Networks
  - Support Vector Machine (can be a linear regression classifier too)
  - **(Boosted) Decision Trees**
  - etc.

# Overtraining

- Machine Learning algorithms can be overly optimized wherein statistical fluctuations from the training data are wrongly characterized as true features of the distributions
  - Deficit of training data statistics versus number of variables or complexity
  - Model flexibility, e.g. many free parameters

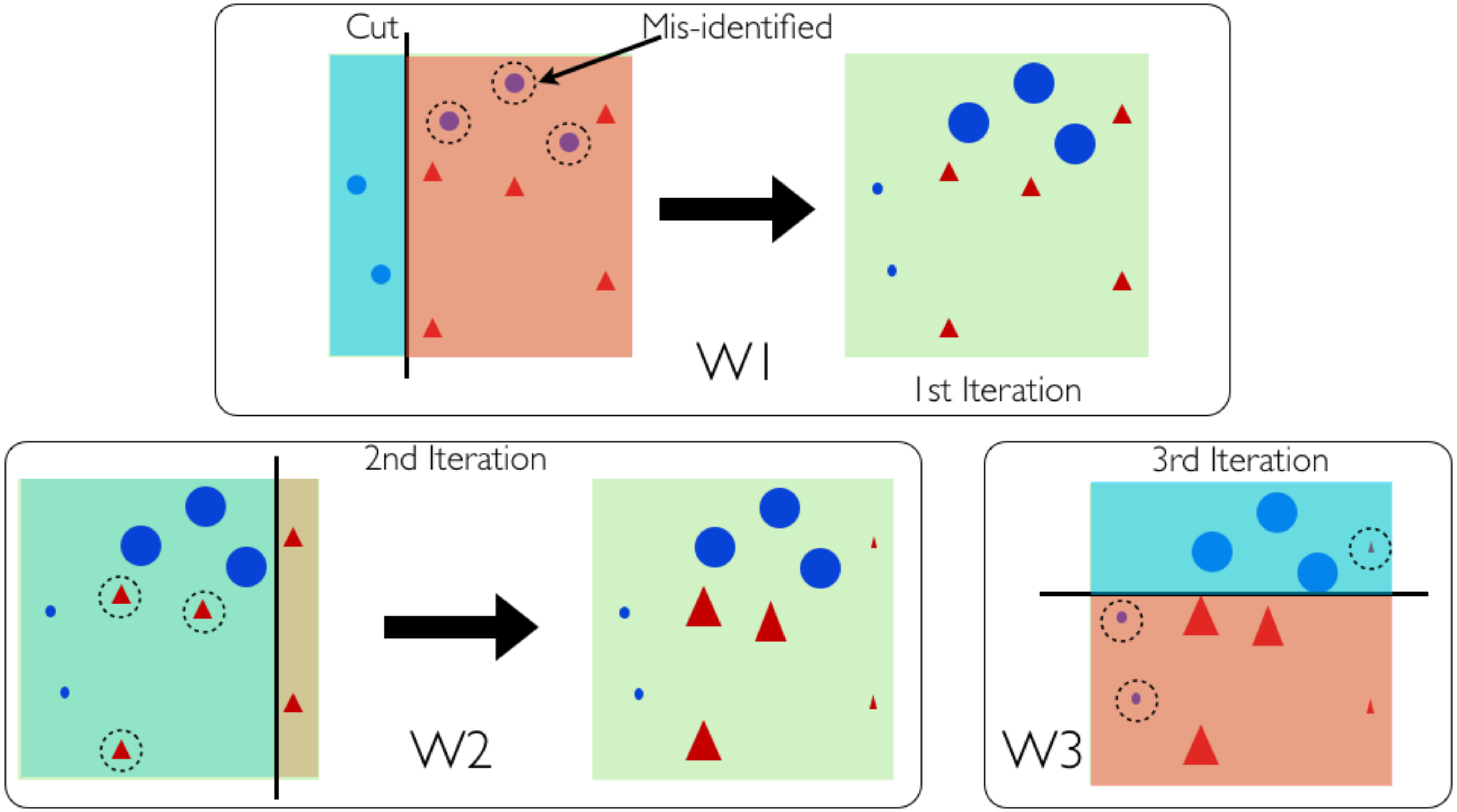




*H. Voss (MPIK)

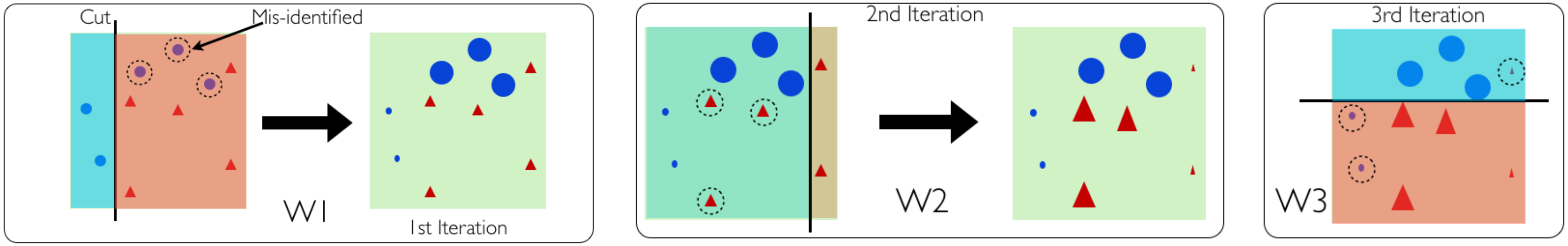
# AdaBoost Boosted Decision Trees

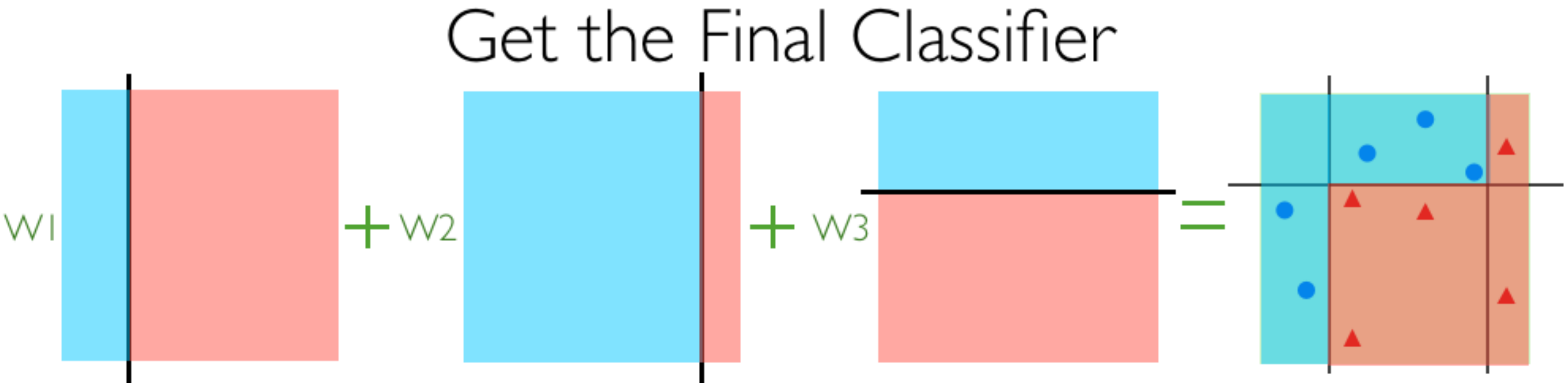
- Past the first one, each iterative boosted decision tree (classifier) is trained on the 'same' events. But now, the events have weights according to whether they were previously wrongly classified. Also, the regions have weights (W1, W2, W3 in the example below) corresponding to the classification error.



# Boosted Decision Trees

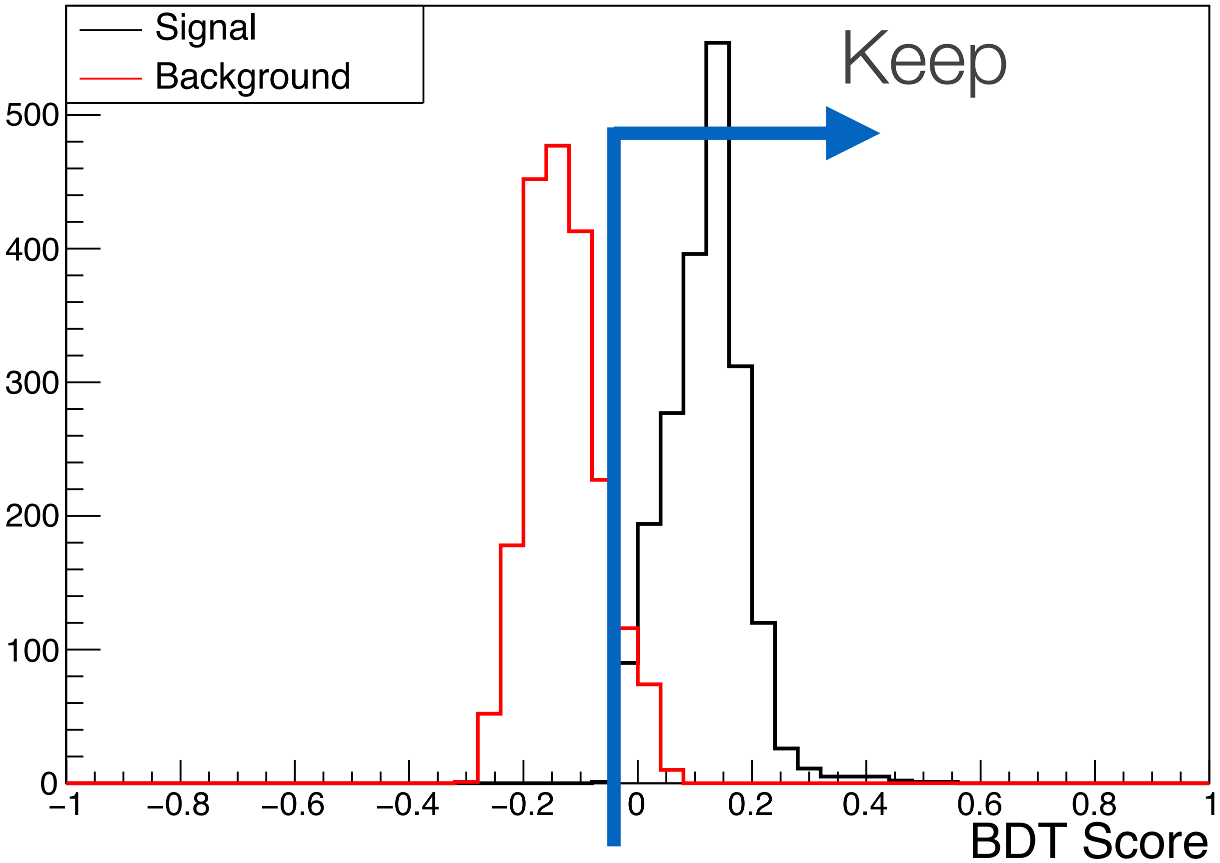
- The combined classifier is the weighted average from all trees for the different regions
- Works very well "out-of-the-box"





# Boosted Decision Tree Classifier

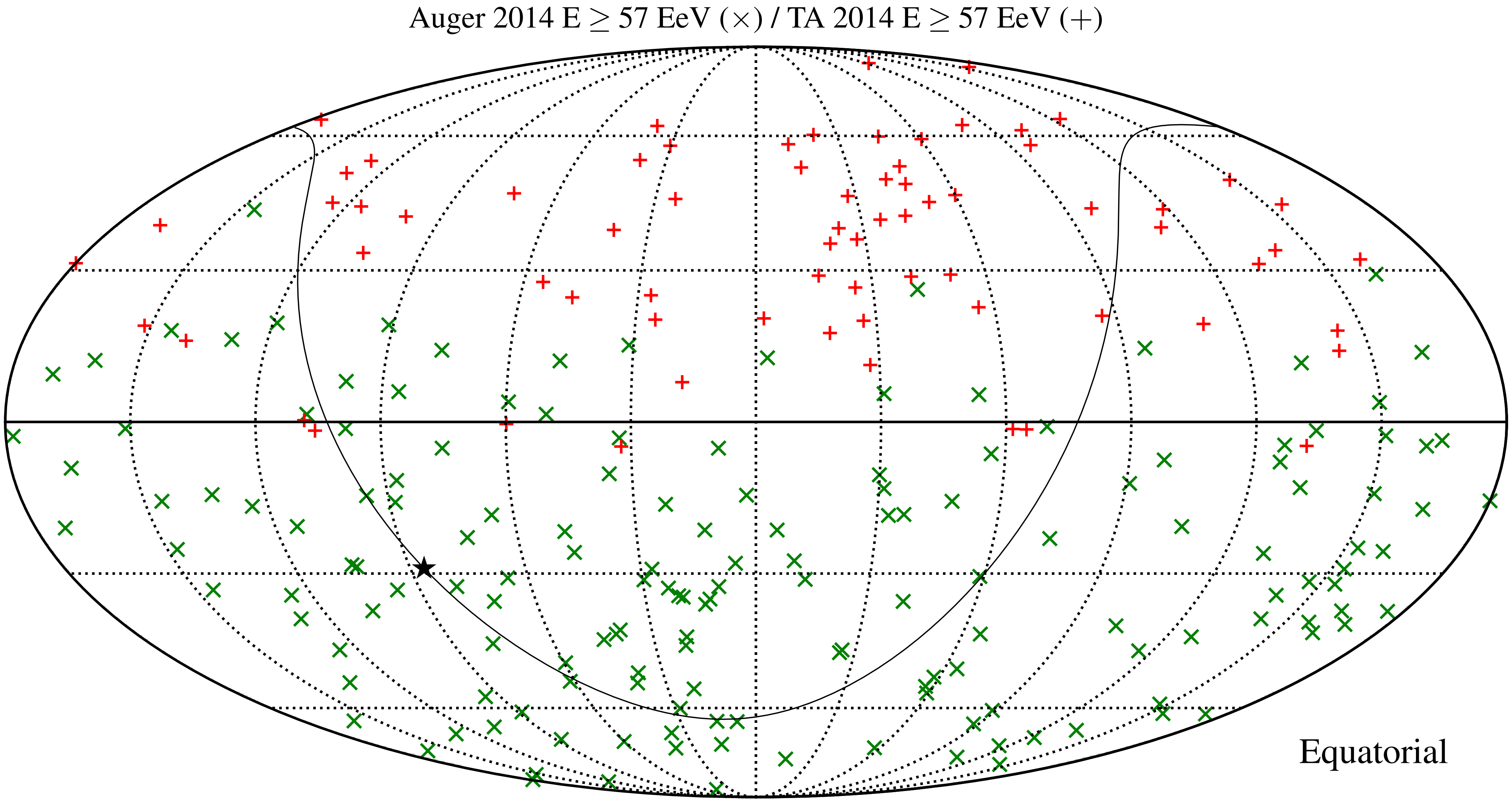
- After training, and hopefully testing, the BDT can generate a score when run over new data that allows signal/ background separation
    - More negative values are background
    - Place a cut at some score to get desired purity and efficiency



We are using the BDT as a classifier and want a decision about whether a **new** data event is more similar to class-A or class-B, e.g. signal or background. We use a "BDT Score" which is here the BDT decision score.

# Auto-Correlation and Statistical Tests

# Example: Arrival Direction of Cosmic Rays



Anisotropies in the arrival directions of ultra-high energy cosmic rays (data from the observatories Telescope Array (TA) and Auger).

# Auto-Correlation

- So far, we have only looked into local excesses in individual bins.
- This method was not sensitive to the correlation between events, *e.g.* in neighbouring bins or in small clusters.
- Consider $N_{\rm tot}$ events distributed on a sphere with position $\mathbf{n}_i$ (unit vector).
- For two events with label $i$ and $j$ ($i \neq j$) we can define an angular distance:

$$\cos\varphi_{ij} = \mathbf{n}_i \cdot \mathbf{n}_j$$

- The **cumulative two-point auto-correlation function** is defined as

$$\mathcal{C}(\{\mathbf{n}_i\},\varphi) = \frac{2}{N_{\rm tot}(N_{\rm tot}-1)}\sum_{i=1}^{N_{\rm tot}}\sum_{j=1}^{i-1}\Theta(\cos\varphi_{ij} - \cos\varphi) \qquad (2)$$

with **step function** $\Theta(x) = 1$ for $x \geq 0$ and $\Theta(x) = 0$ for $x < 0$.

➜ This expression counts the pairs of events within angular distance $\varphi$.

- **Note :** The step function $\Theta()$ is sometimes referenced as the Heaviside function.

# Kolmogorov-Smirnov (KS) Test

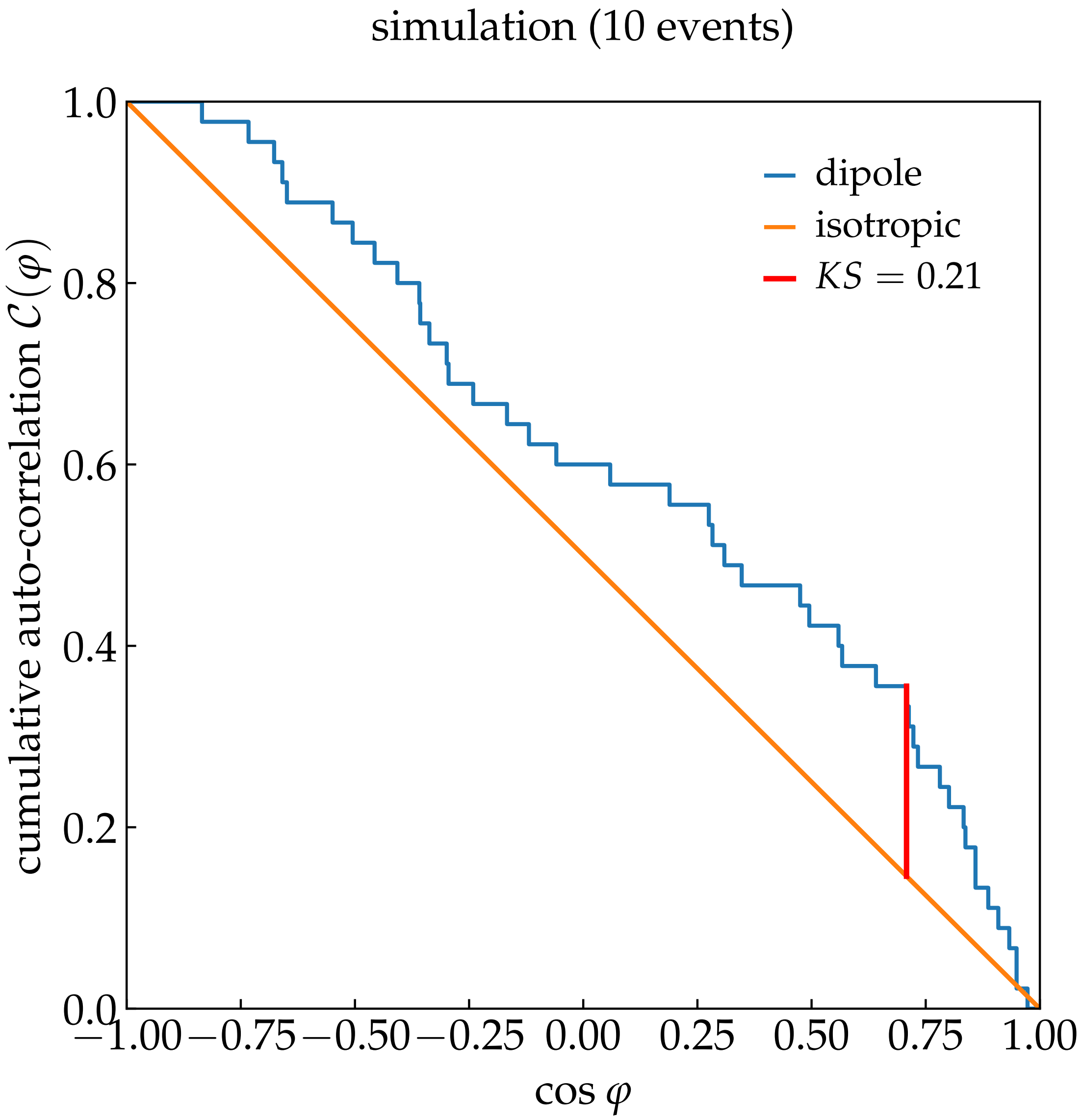
- We want to define a quantity that is a statistical measure for the difference between the empirical distribution and background distribution.
- Area between two curves?

$$\int \mathrm{d}\cos\varphi \left|\mathcal{C}(\{\mathbf{n}_i\},\varphi) - \mathcal{C}_{\mathrm{iso}}(\varphi)\right|$$

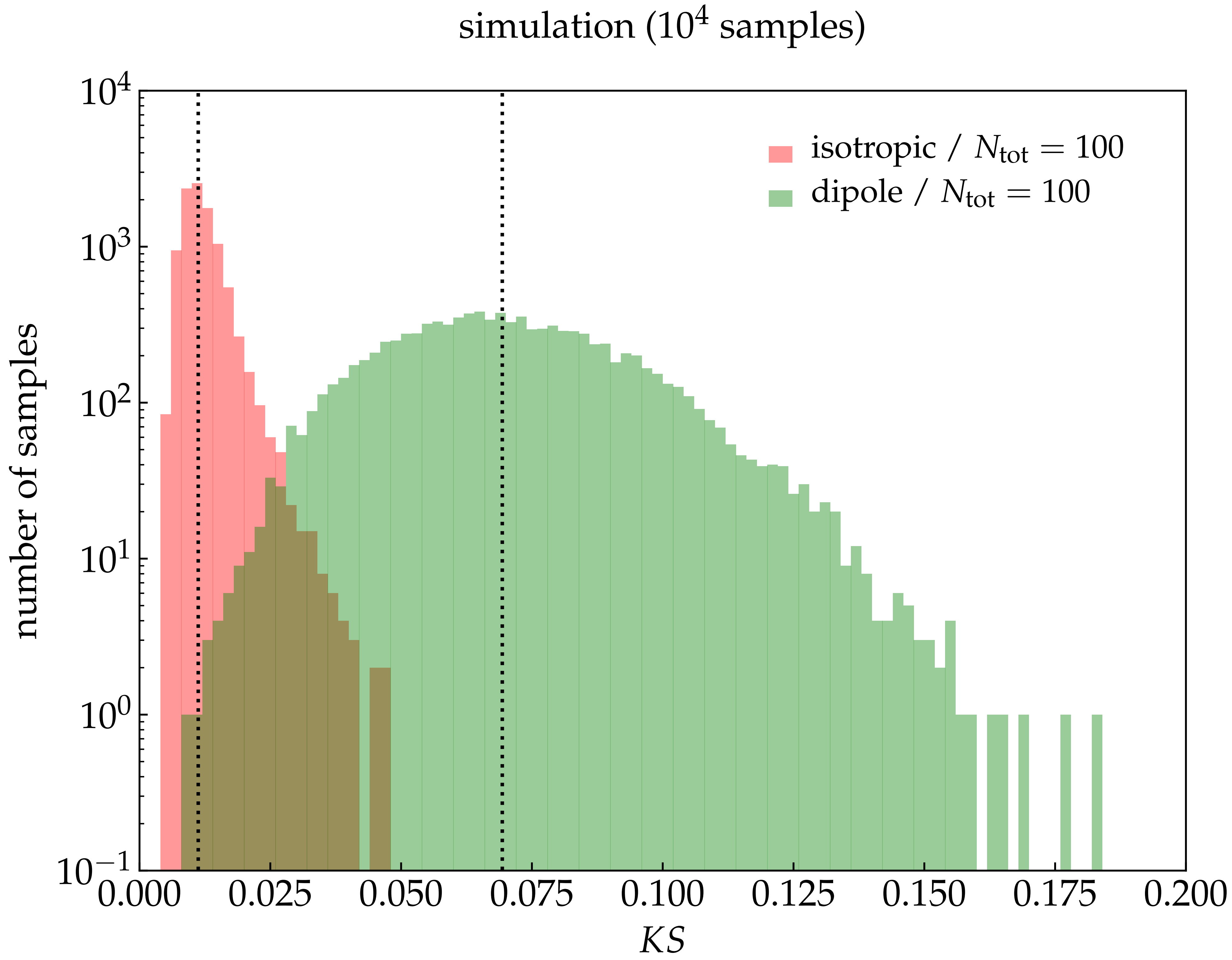
- Or, more general ($L^p$ norm)?

$$\left[\int \mathrm{d}\cos\varphi \left|\mathcal{C}(\{\mathbf{n}_i\},\varphi) - \mathcal{C}_{\mathrm{iso}}(\varphi))\right|^p\right]^{\frac{1}{p}}$$

- **Kolmogrov-Smirnov:** $p \to \infty$.



# Kolmogorov-Smirnov (KS) Test



for python code see : `KS_produce.py` & `KS_show.py`

# Nested Sampling

# Pure Mystic Beauty

- Nested sampling for Bayesian inference is a more recent development and can handle very complicated posterior/likelihood landscapes
- Most recent topic covered, so no review here…

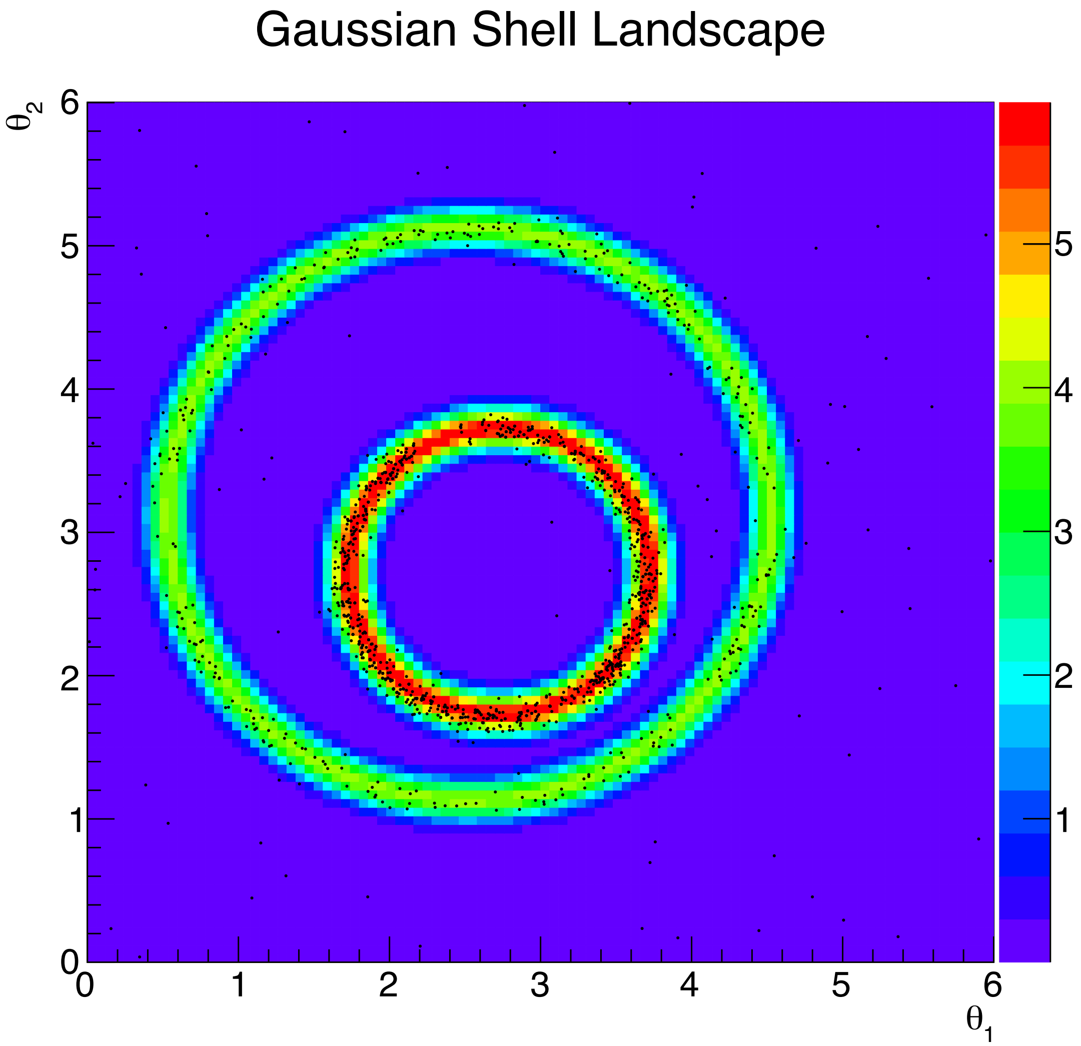
Any sufficiently advanced technology is
indistinguishable from magic.
- Arthur C. Clarke

# Exercise Nested Nested Cylinder

- Using the following likelihood for the two cylinders plot the underlying likelihood and posterior distribution:

$$\mathcal{L}(\vec{\theta}) = \mathrm{circ}(\vec{\theta}; \vec{c}_1, r_1, \sigma_1) + 1.5\ \mathrm{circ}(\vec{\theta}; \vec{c}_2, r_2, \sigma_2)$$

  - $c_1$=( 2.5, 3.1) and $c_2$=( 2.7, 2.7) and $r_1$=2 and $r_2$=1



# Fin