Statistics & Analysis Q&A

PHYS428/PHYS576

Advanced Techniques in Experimental Particle Physics

Fred James's lectures

http://preprints.cern.ch/cgi-bin/setlink?base=AT&categ=Academic_Training&id=AT00000799

http://www.desy.de/~acatrain/

Glen Cowan's lectures

http://www.pp.rhul.ac.uk/~cowan/stat_cern.html

Louis Lyons

http://indico.cern.ch/conferenceDisplay.py?confId=a063350

Bob Cousins gave a CMS lecture, may give it more publicly Gary Feldman "Journeys of an Accidental Statistician"

http://www.hepl.harvard.edu/~feldman/Journeys.pdf

http://histfitter.web.cern.ch/histfitter/

Example: Maximum Likelihood Estimator

example: PDF(x) = Gauss(x,μ,σ) → $\mathrm{L}(\mu|x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

→ estimator for μ_{true} from the data measured in an experiment $x_1, \ldots . . x_N$

→fullLikelihood $\mathrm{L}(\mu|x) = \prod_i^N \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$

→typically: $-\mathbf{2ln}\left(\mathbf{L}(\mu|x)\right) = \sum_i^N \left(\frac{(x_i-\mu)^2}{2\sigma^2}\right) + N\frac{1}{\sqrt{2\pi\sigma^2}}$ Note:It'safunctionofμ!

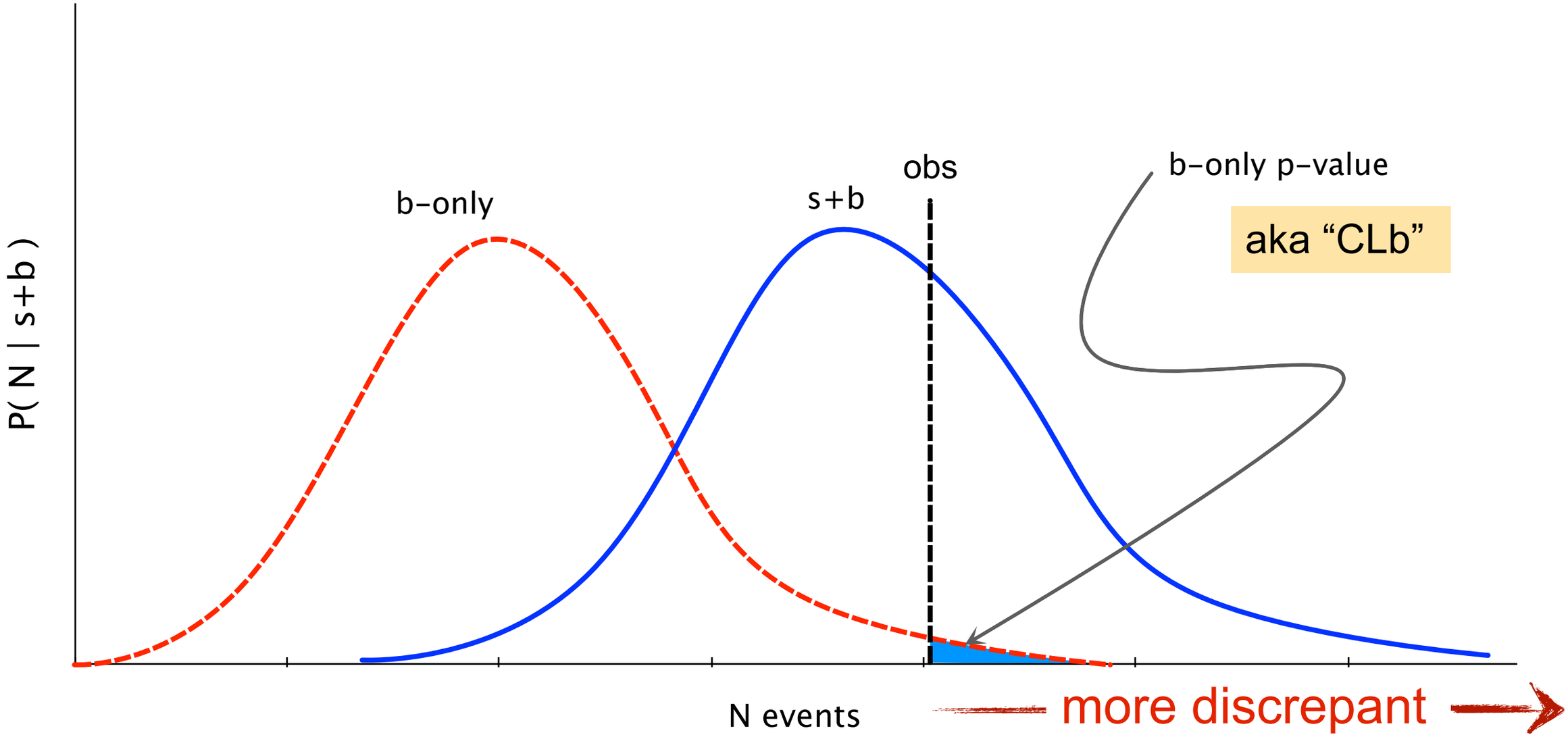
$$-\mathbf{2\Delta ln}\left(\mathbf{L}(\boldsymbol{\mu})\right) = \sum_i^N \left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

→ χ^2 , least squares

Discovery sensitivity: p-value

Discovery: test b-only (null: s=0 vs. alt: s>0)

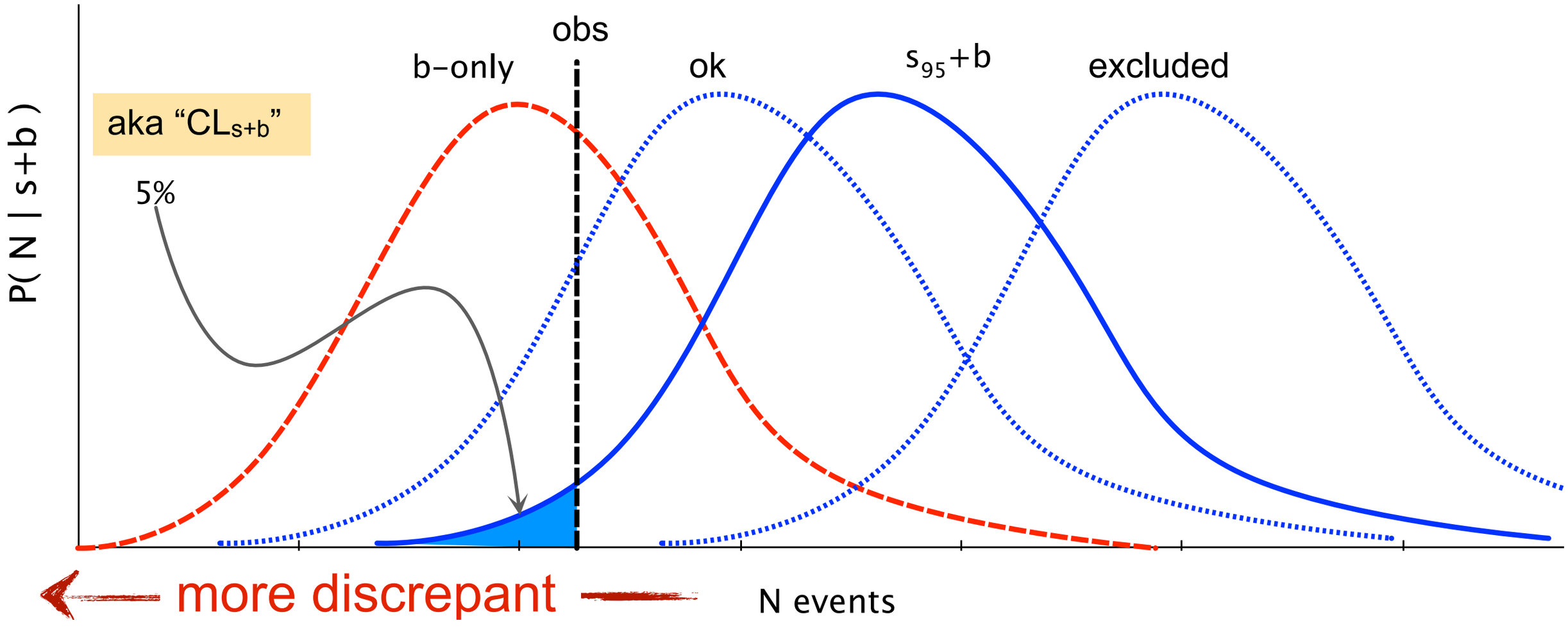
- note, **one-sided** alternative. larger N is “more discrepant”



Upper Limit: Confidence Level

What is meant by “95% upper limit” ? See the picture below?

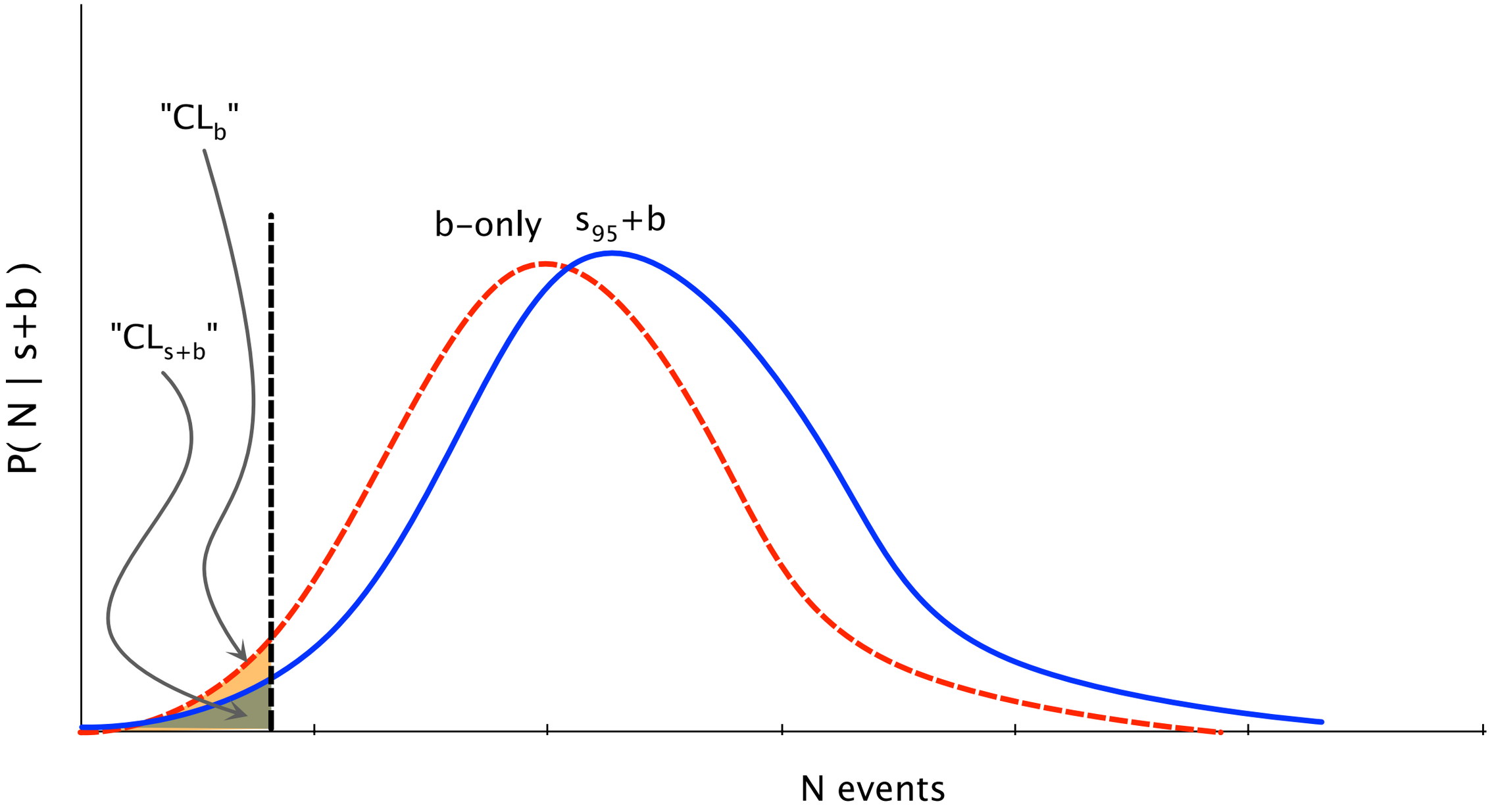
- ie. increase s, until the probability to have data “more discrepant” is < 5%



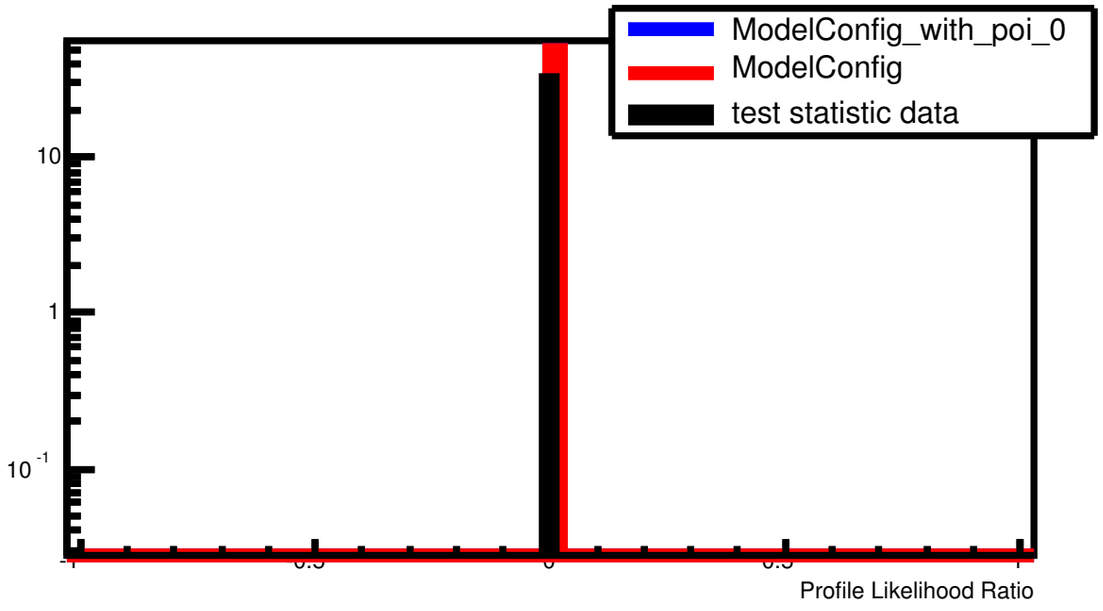
CLs Upper Limit: Confidence Level

CL_s is known to be “conservative” (over-cover): expected limit covers with 97.5%

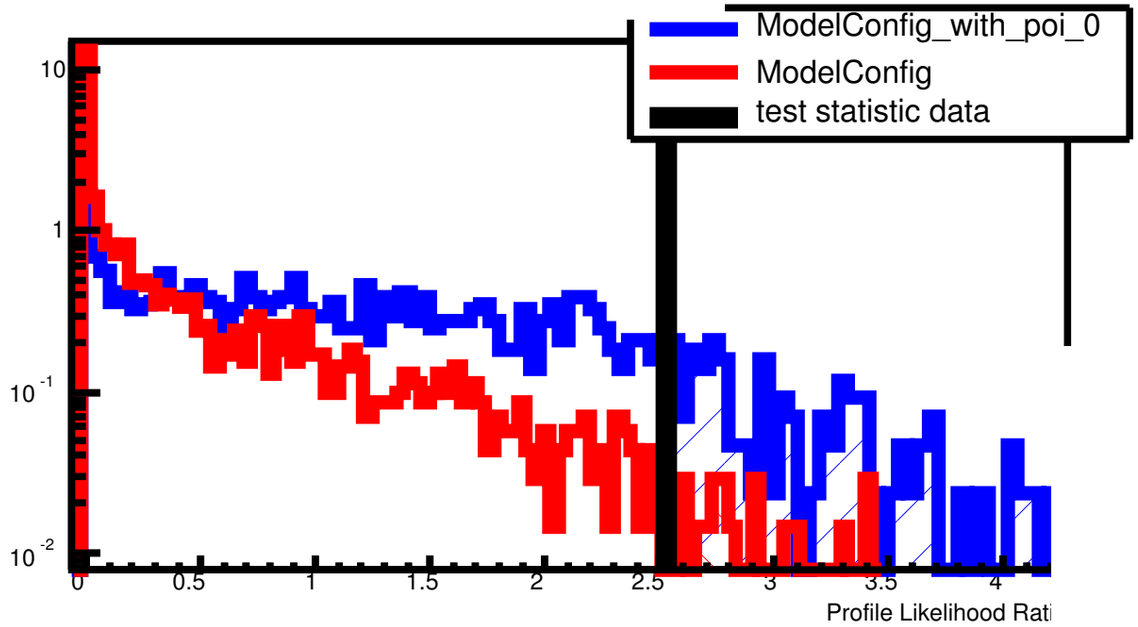
- Note: CL_s is NOT a probability

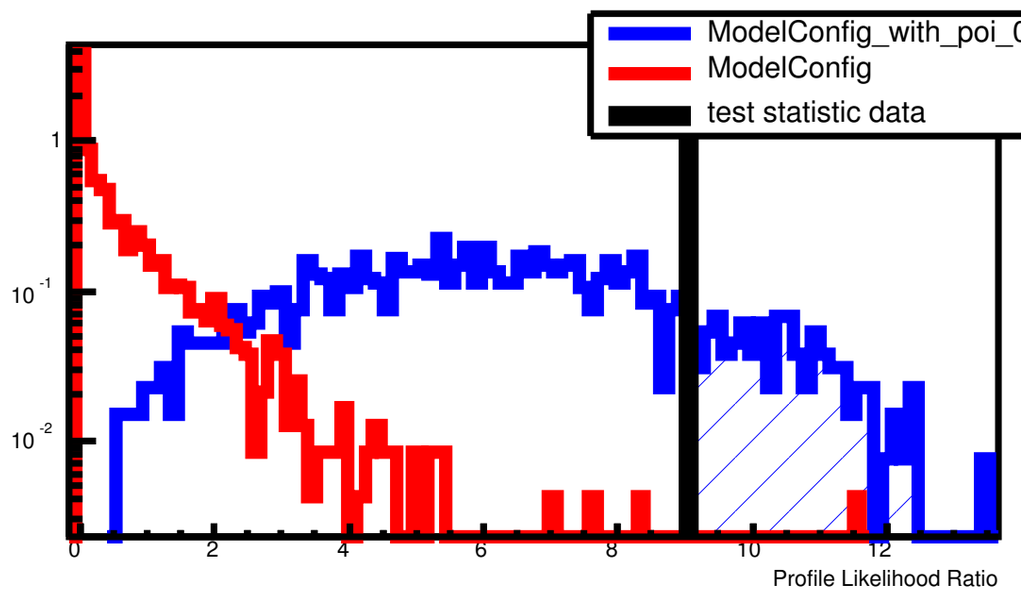


Test Statistics



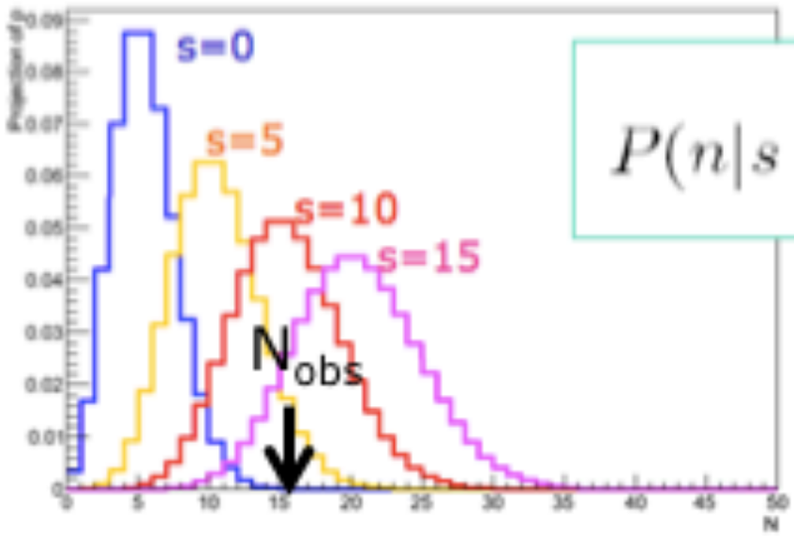
$$q_{\mu_{\mathrm{sig}}} = -2 \log \frac{L(\mu_{\mathrm{sig}}, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}_{\mathrm{sig}}, \hat{\boldsymbol{\theta}})}$$





Likelihood

- **All** fundamental statistical procedures are based on the likelihood function as 'description of the measurement'



$$P(n|s+b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

NB: b is a constant in this example

Definition: the Likelihood is P(observed data|theory)

e.g. L(15|s=0)

e.g. L(15|s=10)

Frequentist statistics

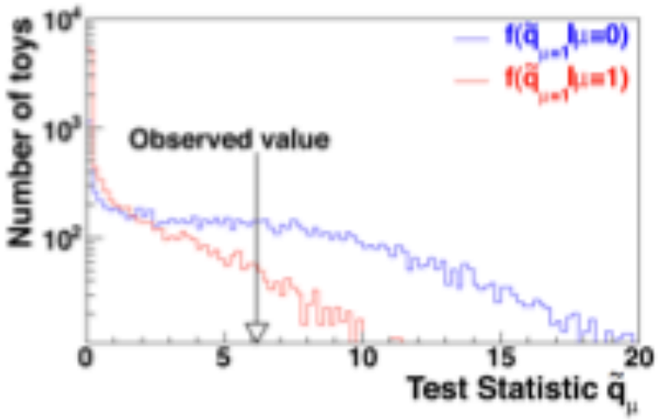
Bayesian statistics

Maximum Likelihood

Likelihood

Frequentist statistics

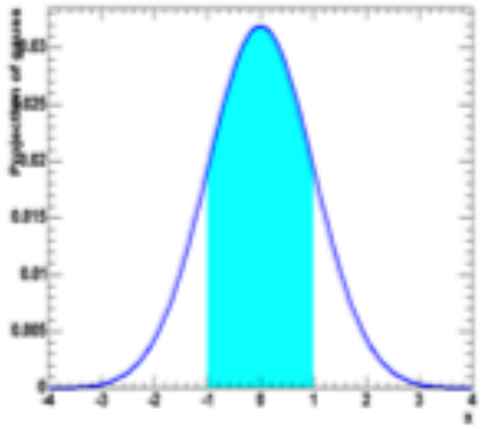
$$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N} \mid \mu)}{L(\vec{N} \mid \hat{\mu})}$$



Confidence interval or p-value

Bayesian statistics

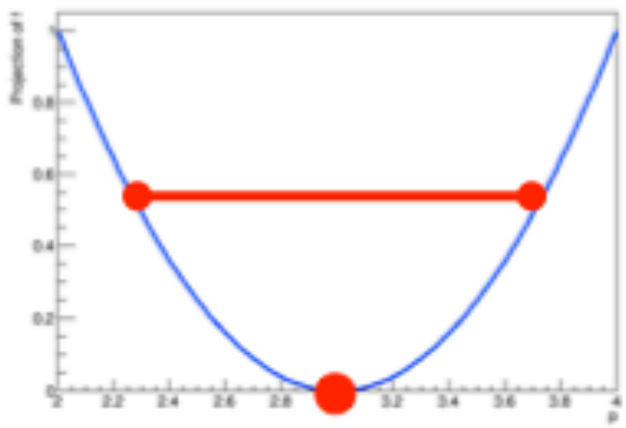
$$P(\mu) \propto L(x \mid \mu) \cdot \pi(\mu)$$



Posterior on s or Bayes factor

Maximum Likelihood

$$\left.\frac{d \ln L(\vec{p})}{d\vec{p}}\right|_{p_i = \hat{p}_i} = 0$$

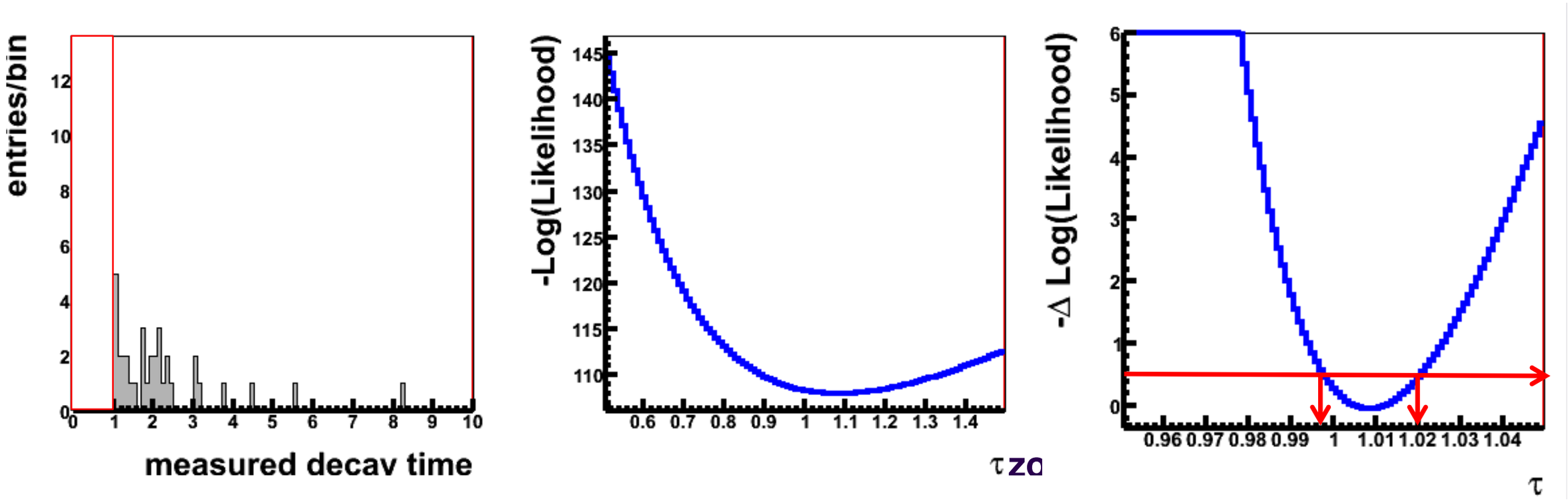


s = x ± y

Maximum Likelihood Estimator

$$-ln\left(L(\theta)\right) \approx -ln\left(L(\hat{\theta})\right) - \underbrace{\left[\frac{dln(L\theta)}{d\theta}\right]_{\hat{\theta}} (\theta - \hat{\theta})}_{\text{minimum}} - \frac{1}{2}\left[\frac{d^2 ln(L\theta)}{d\theta^2}\right]_{\hat{\theta}} (\theta - \hat{\theta})^2 + \ ..$$

$$-ln\left(L(\theta)\right) \approx -ln\left(L(\hat{\theta})\right) + \frac{1}{2\sigma^2}(\theta - \hat{\theta})^2$$



read off parabolic $\boldsymbol{ln}\ (\boldsymbol{L})$ curve:

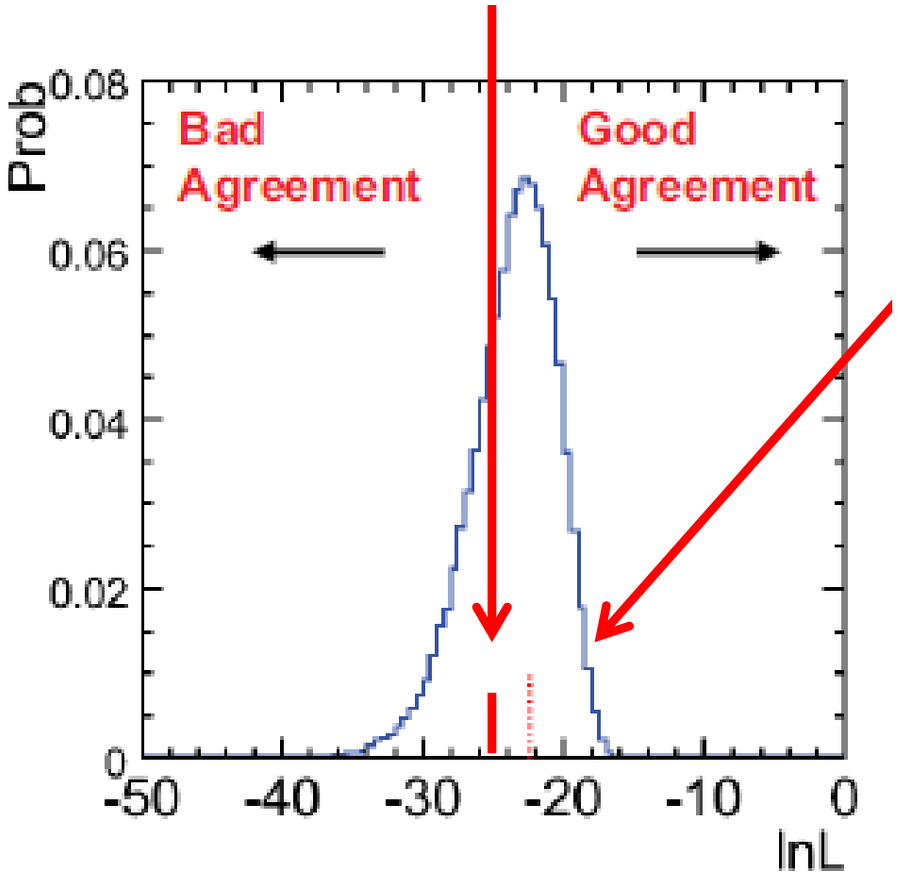
$$-ln\left(L(\hat{\theta} \pm \sigma_\theta)\right) = -ln\left(L(\hat{\theta})\right) + \frac{1}{2}$$

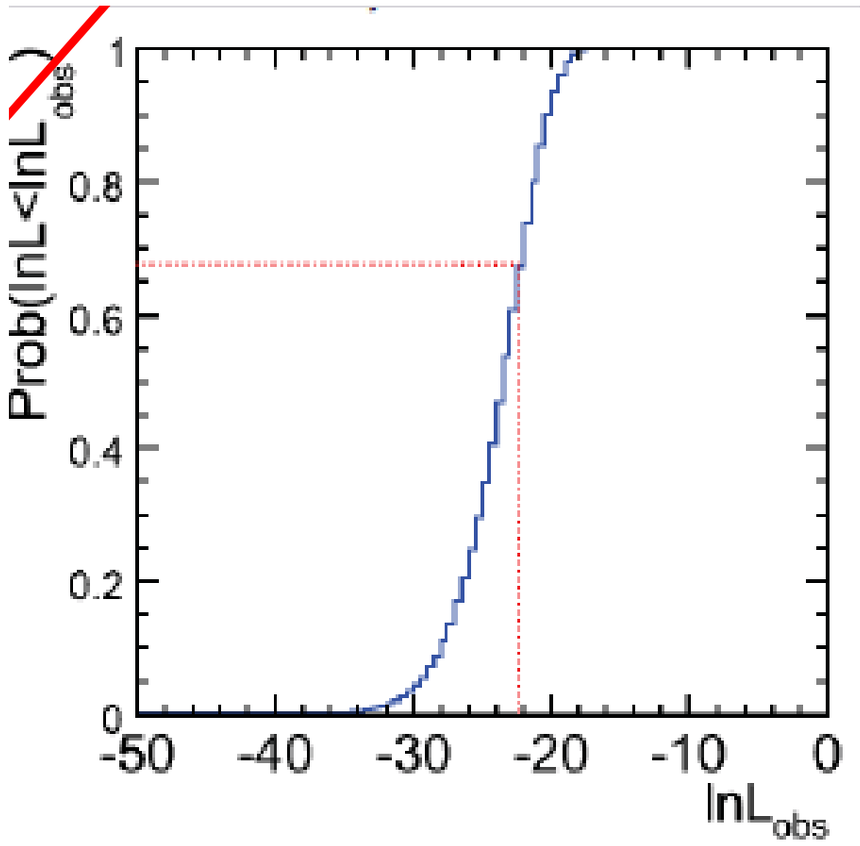
Goodness-of-fit

- So far we know the "uncertainty" on the fitted value of $\boldsymbol{\theta}$, but...
- did the fitted model "really" describe the data?
- The value of the lnL (log Likelihood) at the minimum does not "mean anything" → calibrate!

→ determine the distribution of lnL fit results with Monte Carlo toys!

→ check your "data"-fit





Goodness-of-fit

- Easier with Gaussian distributed variables

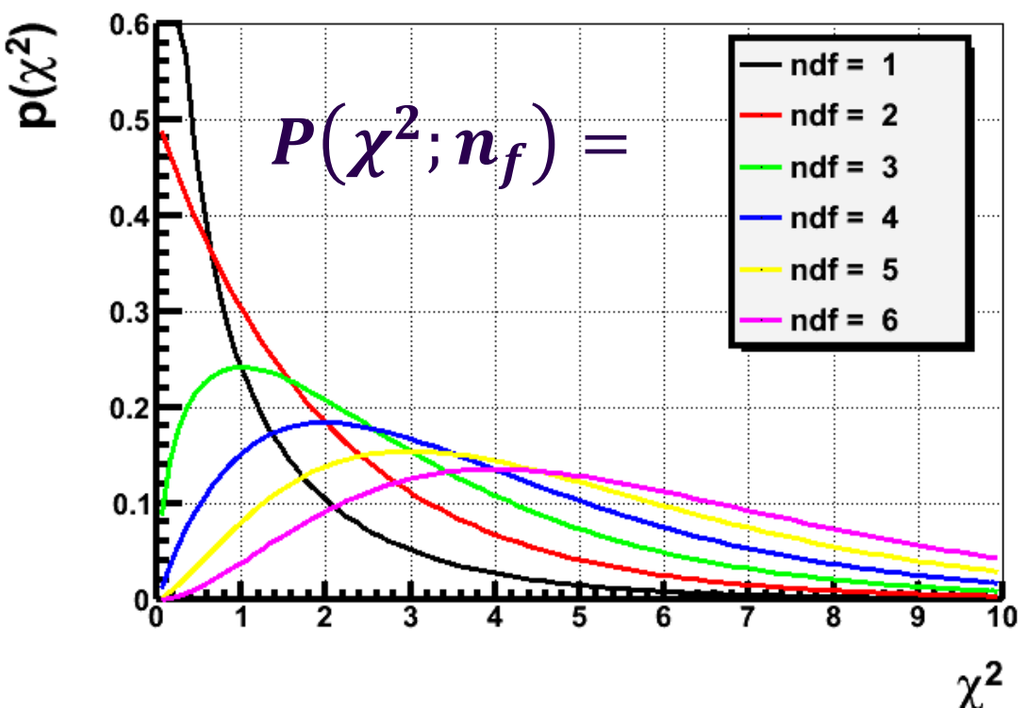
→ least square fit → $\boldsymbol{\chi^2}$

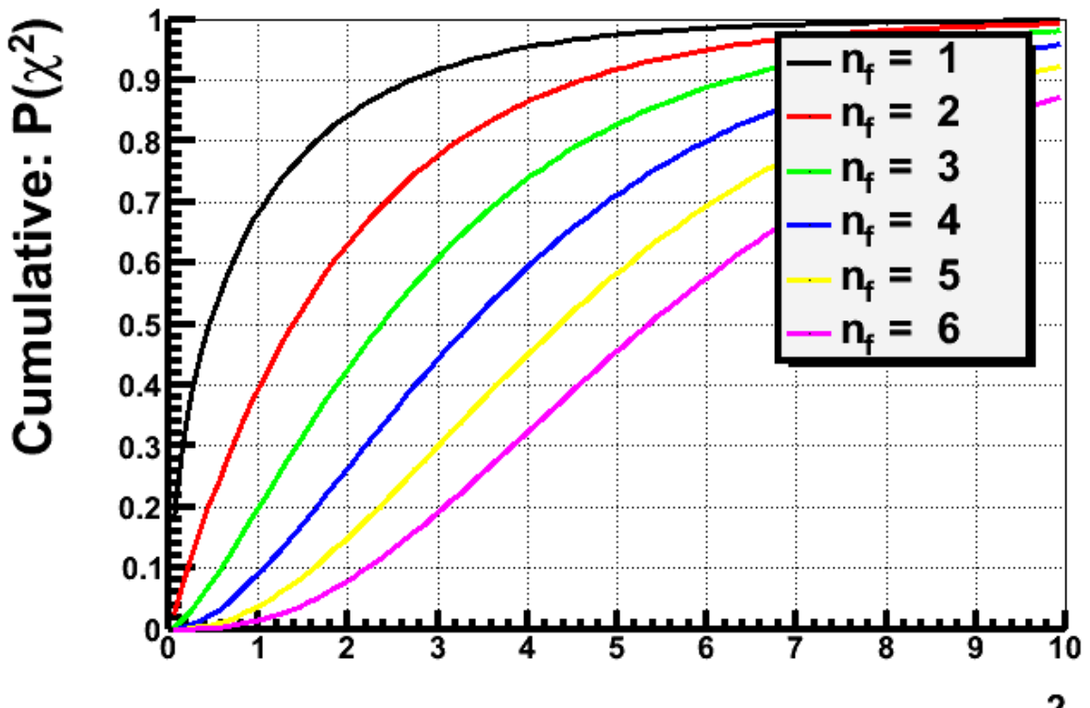
$$\chi^2 = \sum_i^n \left(\frac{(\widehat{\mu_i} - \mu_i(\theta))^2}{2\sigma^2} \right)$$

has known distribution: E[$\boldsymbol{\chi^2}$]= $\boldsymbol{n_f}$: #number of "degree of freedom"

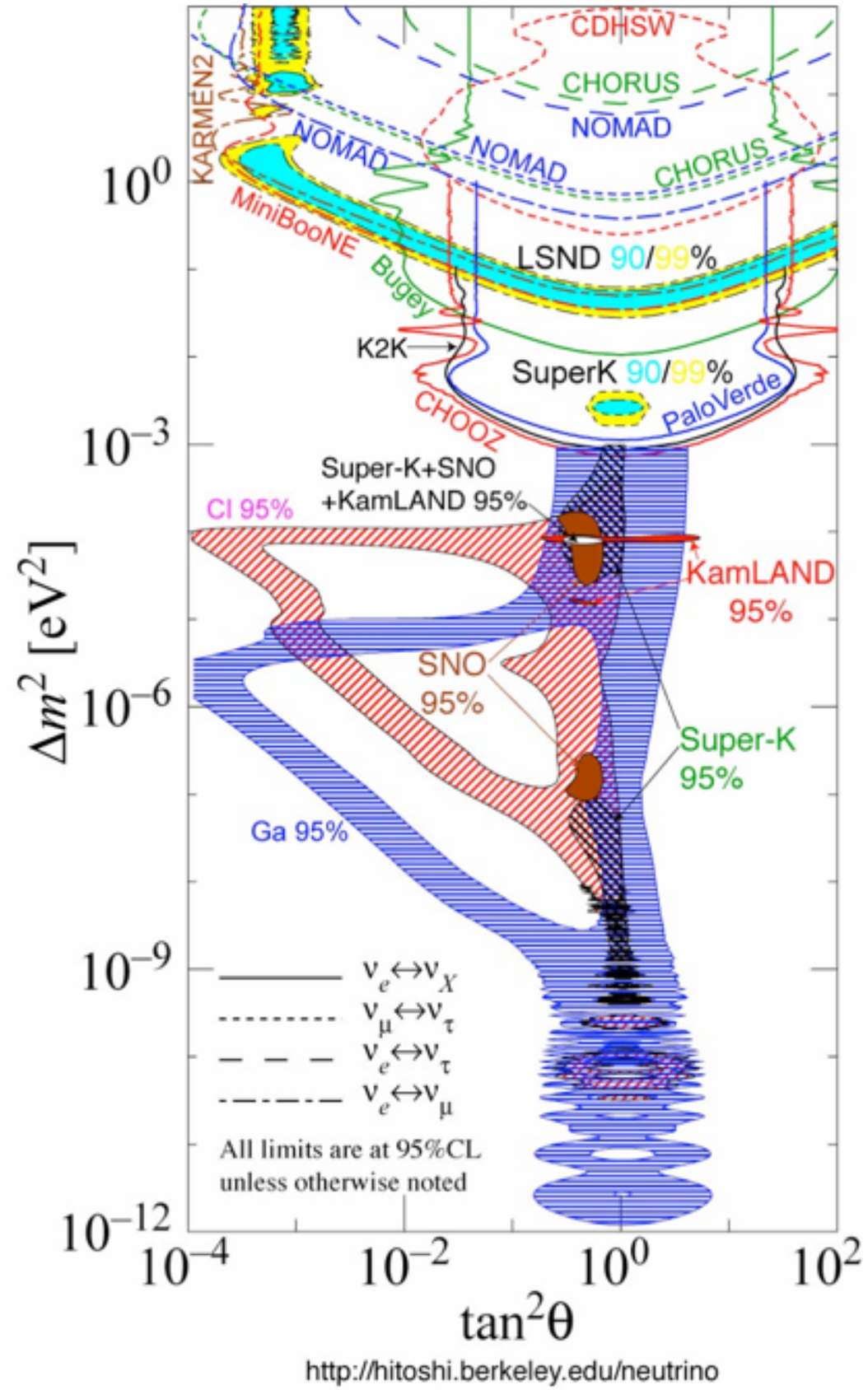
i.e. $\boldsymbol{n - \#fitted\ parameters}$

Chi2 Probability: The 1-cumulative distr. of P($\boldsymbol{\chi^2}$,$\boldsymbol{n_f}$) **distribution**
how often to expect "worse" fit result (i.e. with larger $\boldsymbol{\chi^2\ value}$ at min.)

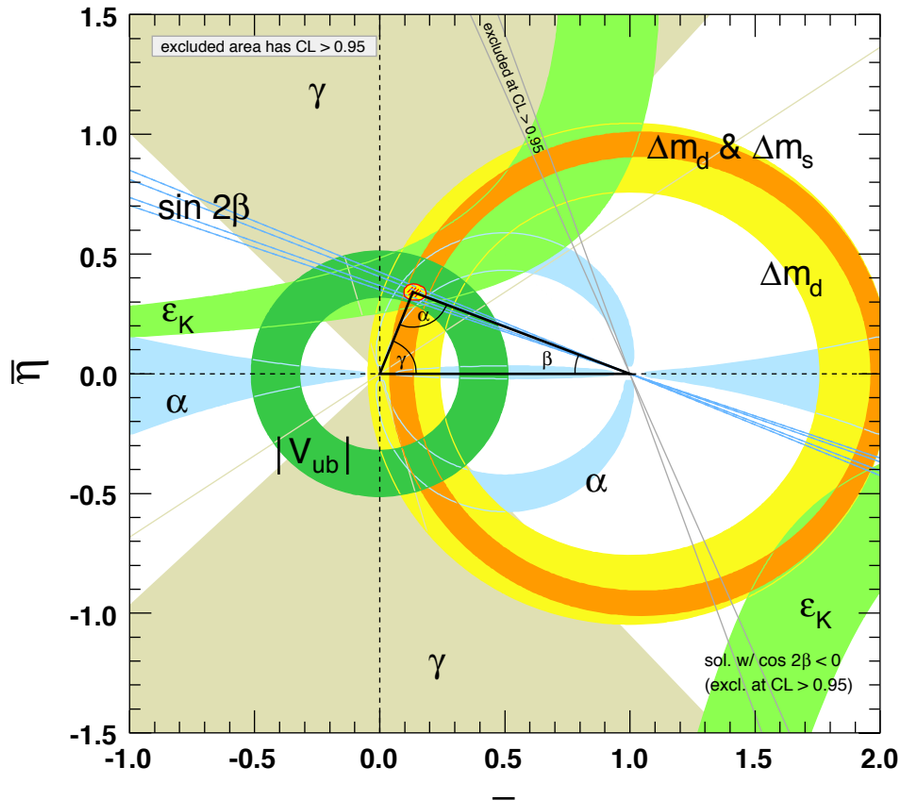


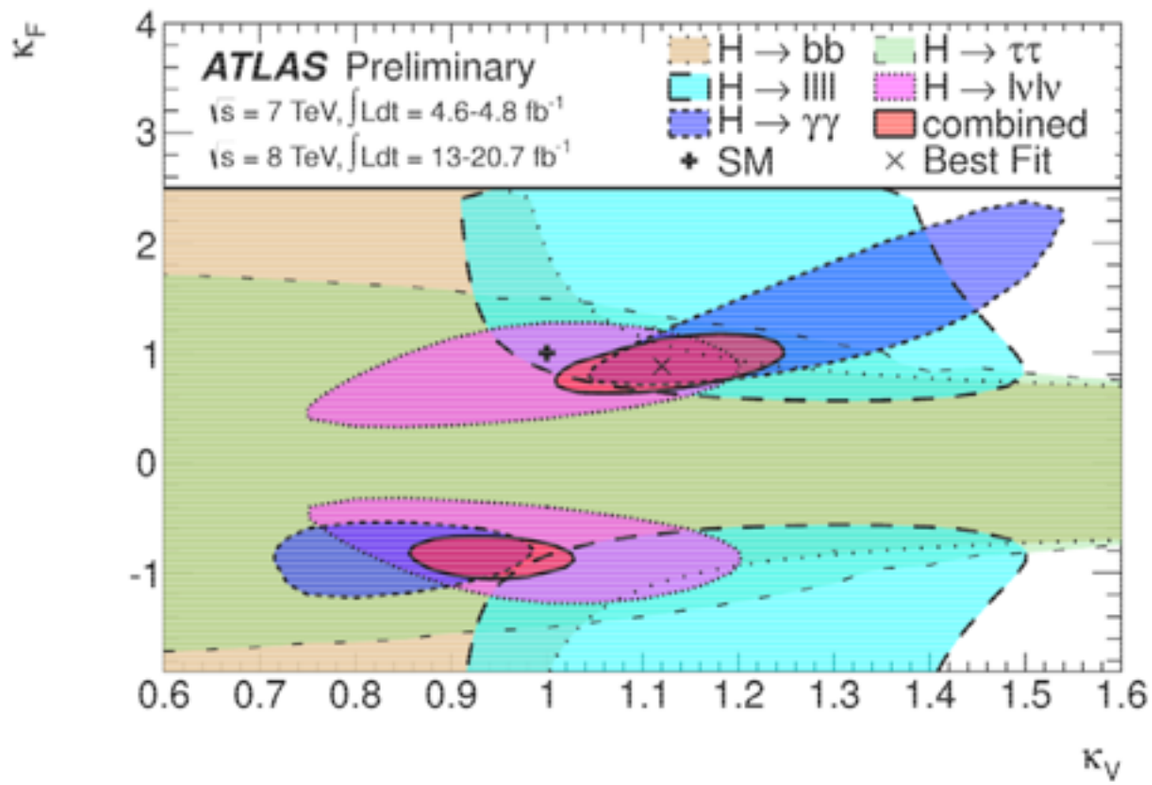


Examples of Confidence Intervals



http://hitoshi.berkeley.edu/neutrino

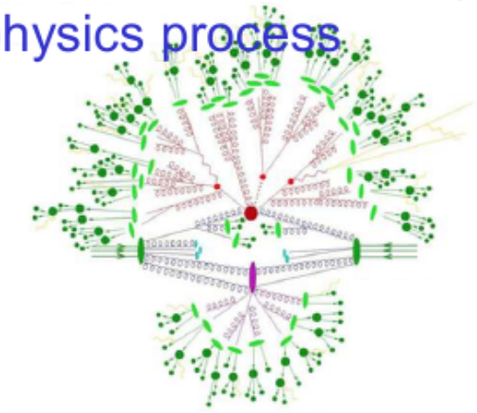




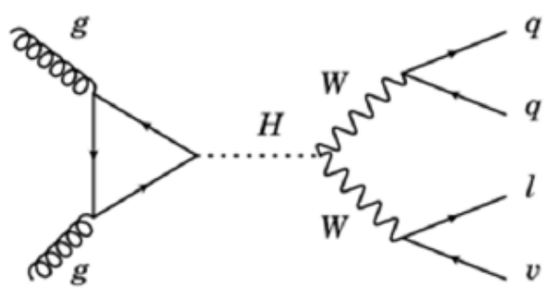
Physics Analysis

Particle Physics Workflow

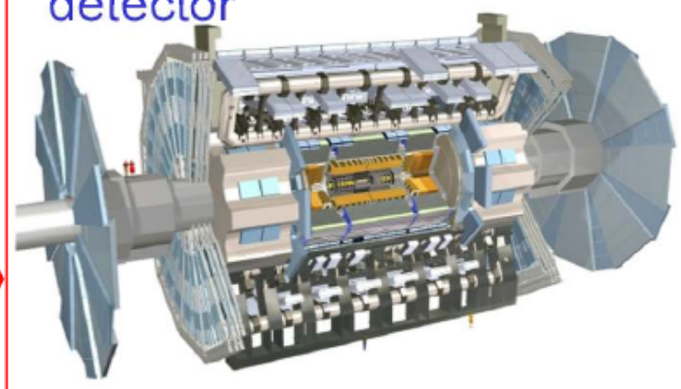
Simulation of 'soft physics' physics process



Simulation of high-energy physics process

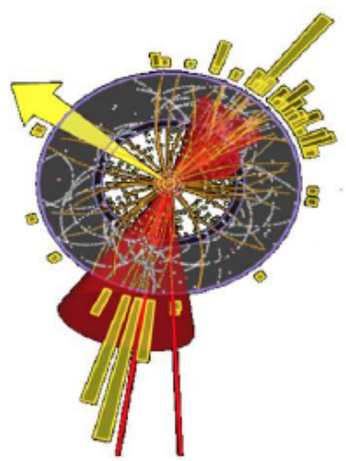


Simulation of ATLAS detector



LHC data

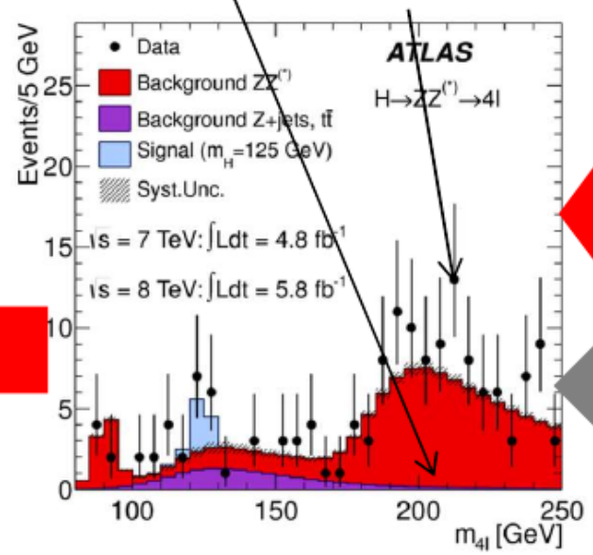
Reconstruction of ATLAS detector



Analysis Event selection

$P(m_{4l}|SM[m_H])$

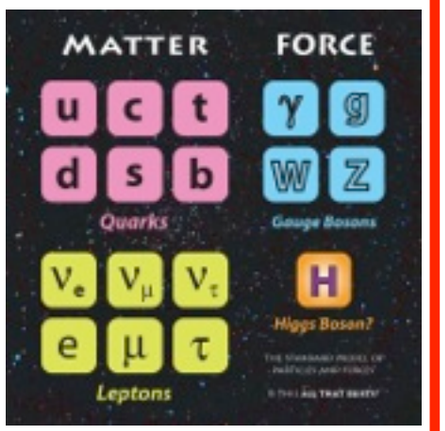
Observed m_{4l}

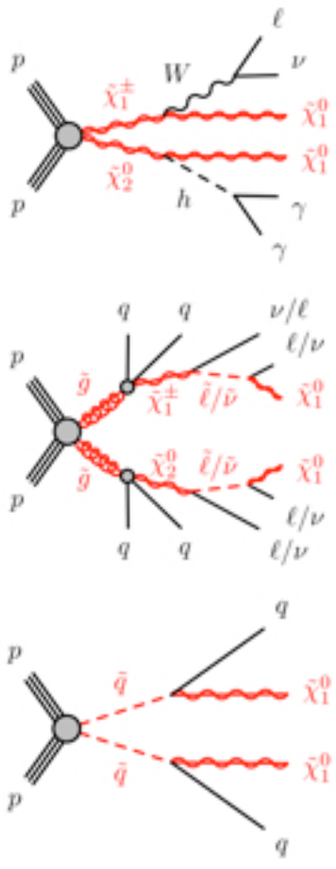


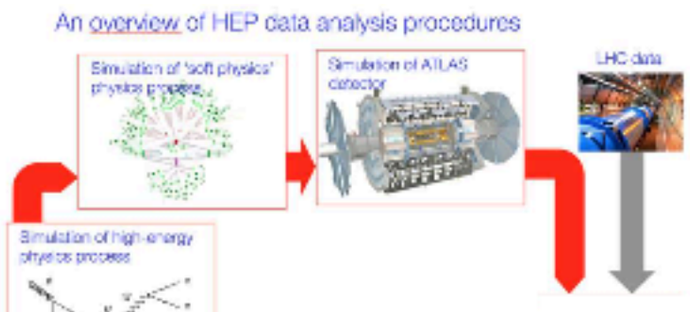
prob(data|SM)

W. Verkerke

Analysis Workflow







HEP workflow: ***analysis view***

MC Simulated Events (sig kg)

All available "real data"

N-tuples
Cut-flows,
Multi-variate analysis (NN,BDT)
ROOT, TMVA, NeuroBayes

Helps to define selection

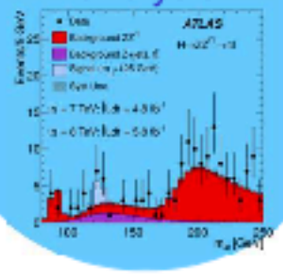
Event selection (cuts, NN, BDT)

Final Event Selection (MC)

Final Event Selection (data)

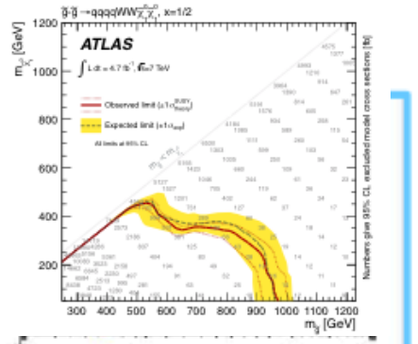
Signal, background models
Likelihood models,
MINUIT, RooFit
RooStats, MCLimit

Statistical analysis

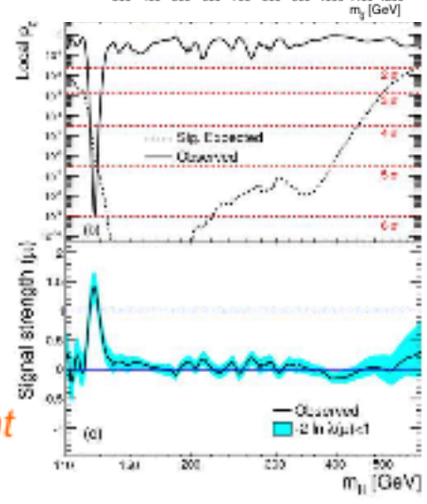


Final Result

Limit



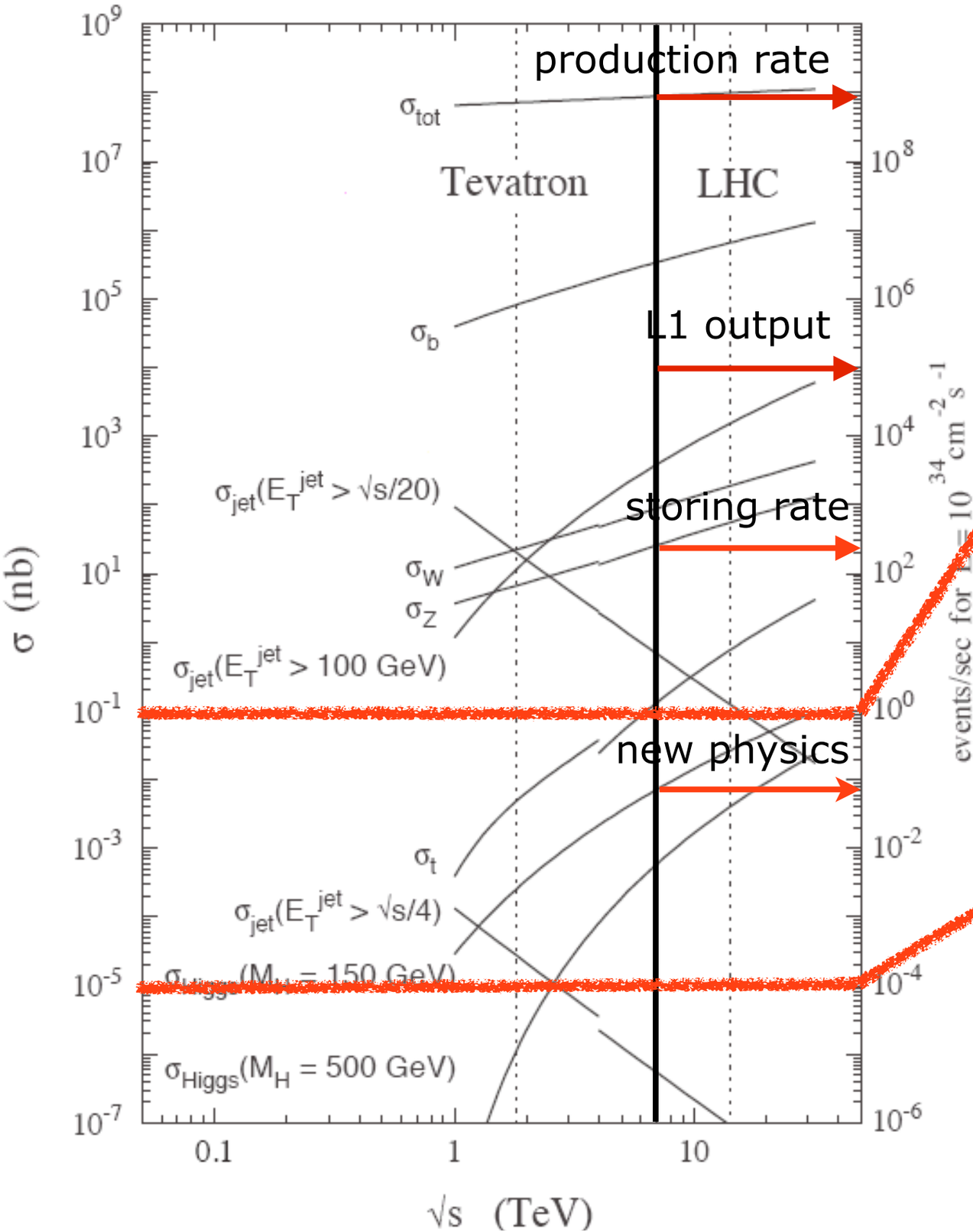
Discovery

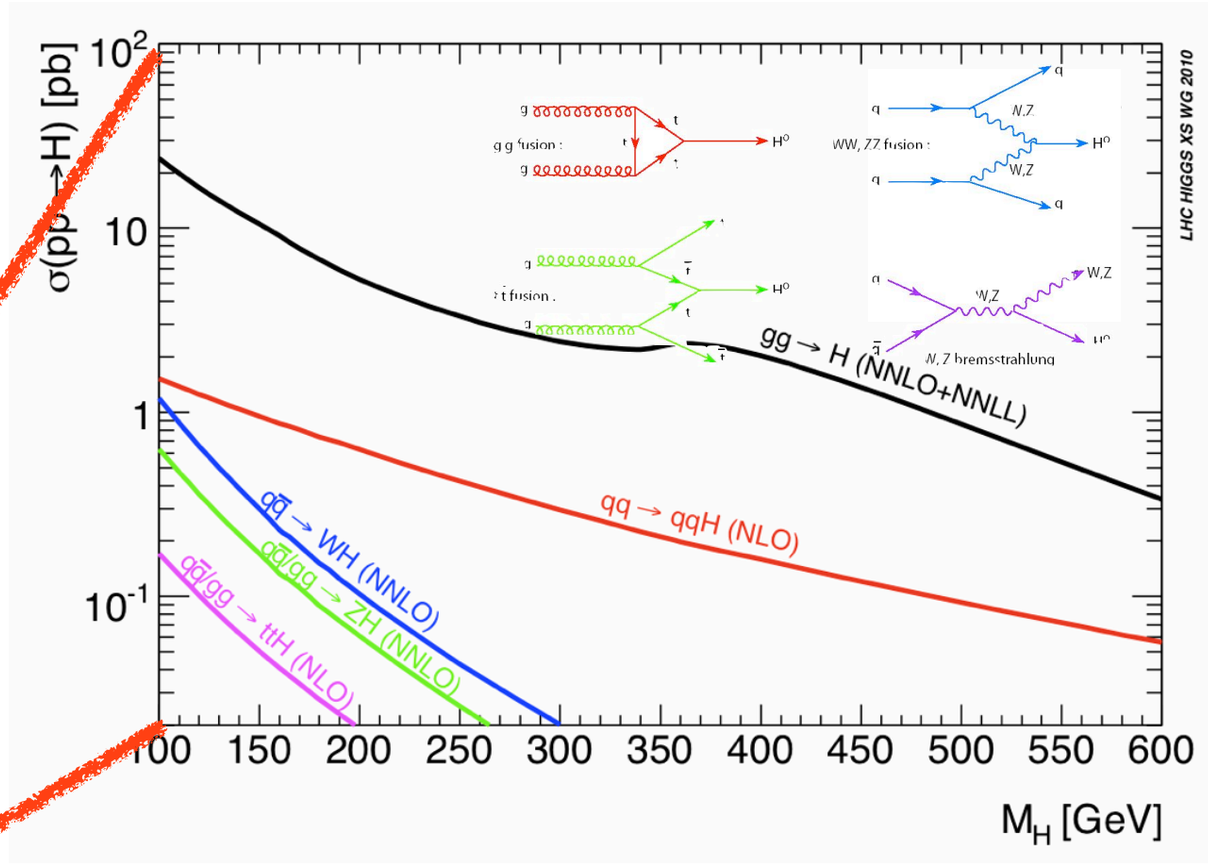


Measurement

W. Verkerke

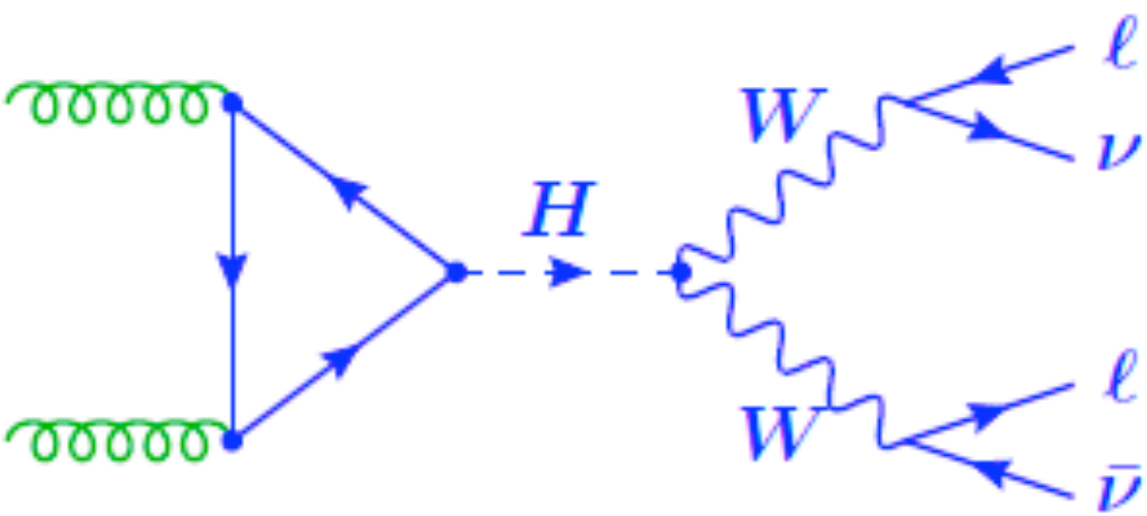
Extremely Rare Production processes

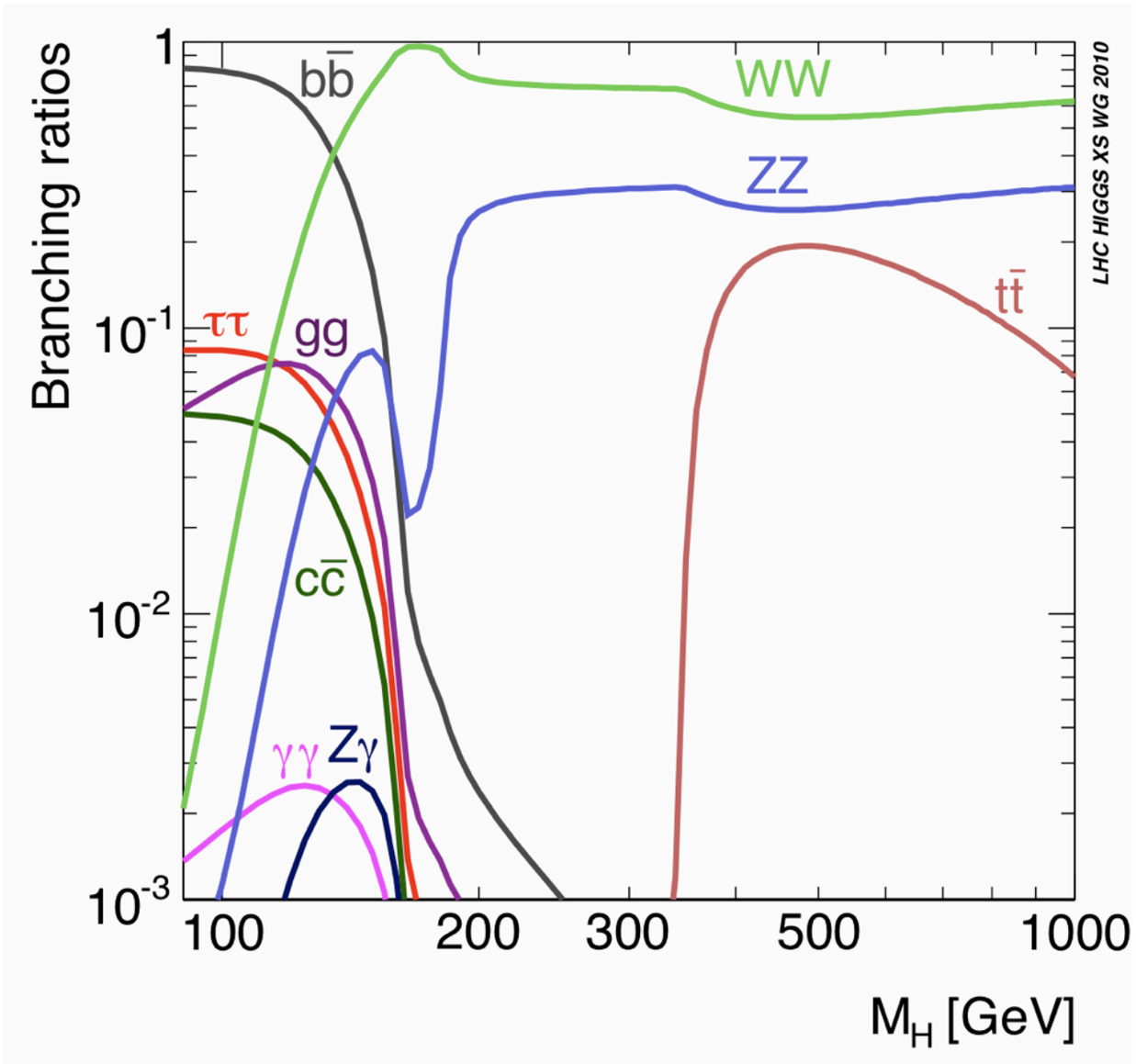




H > WW > lvlv: Signal

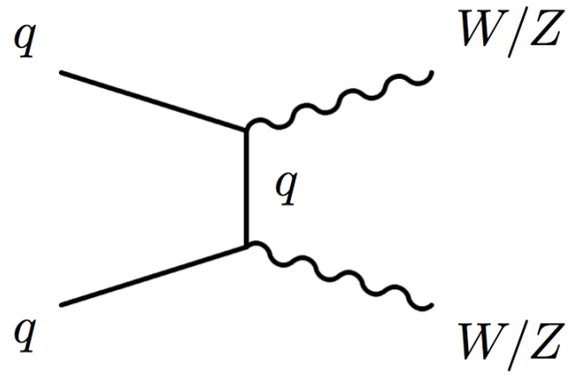
$$gg \to H \to WW \to \ell\bar{\nu}\bar{\ell}\nu$$

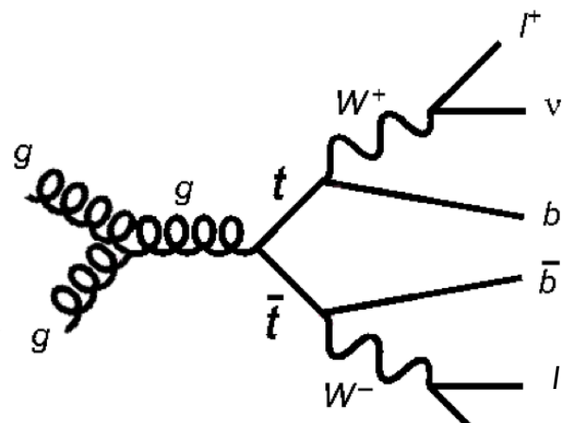


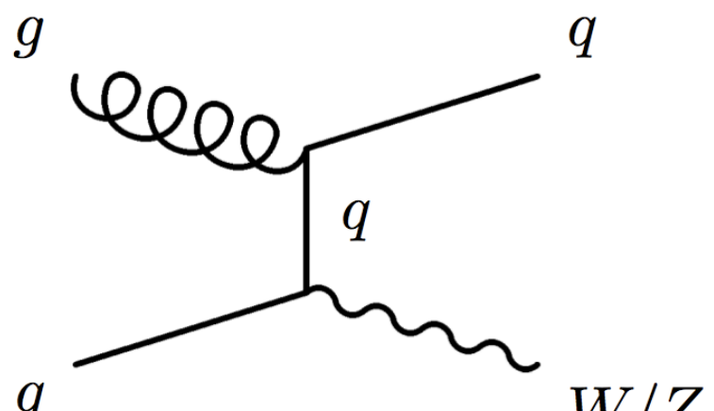


Backgrounds

- two identified leptons + missing energy in the final state



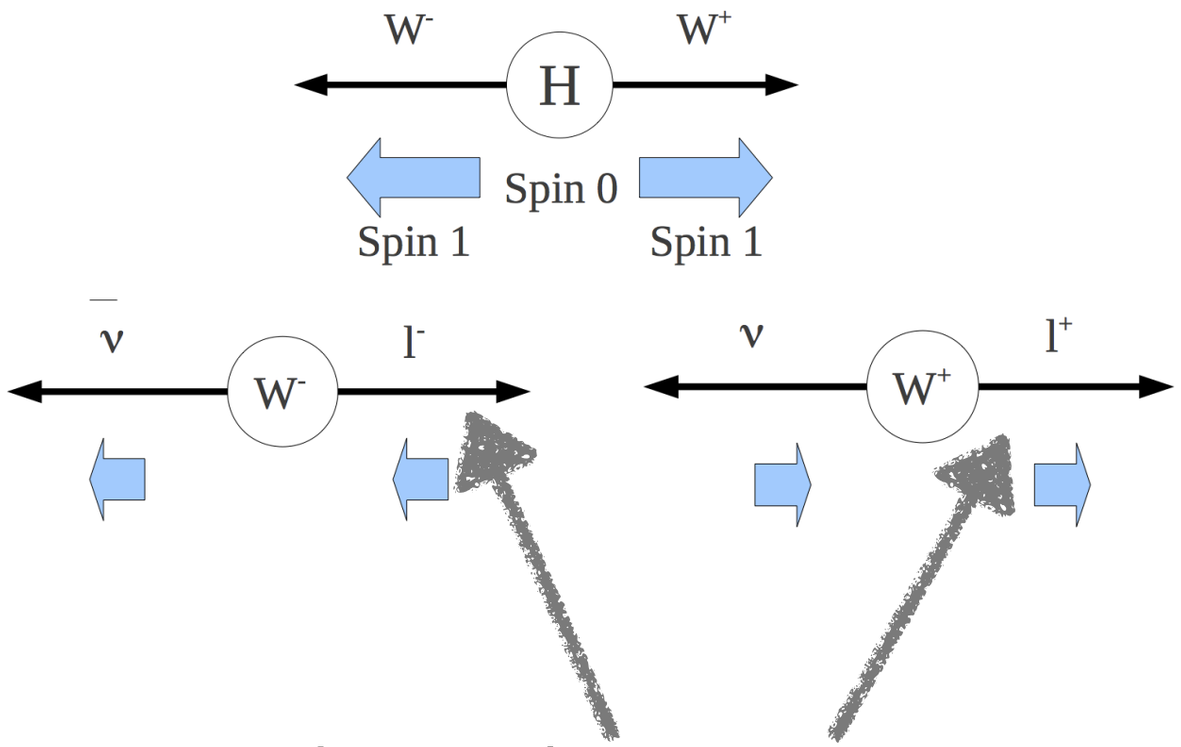




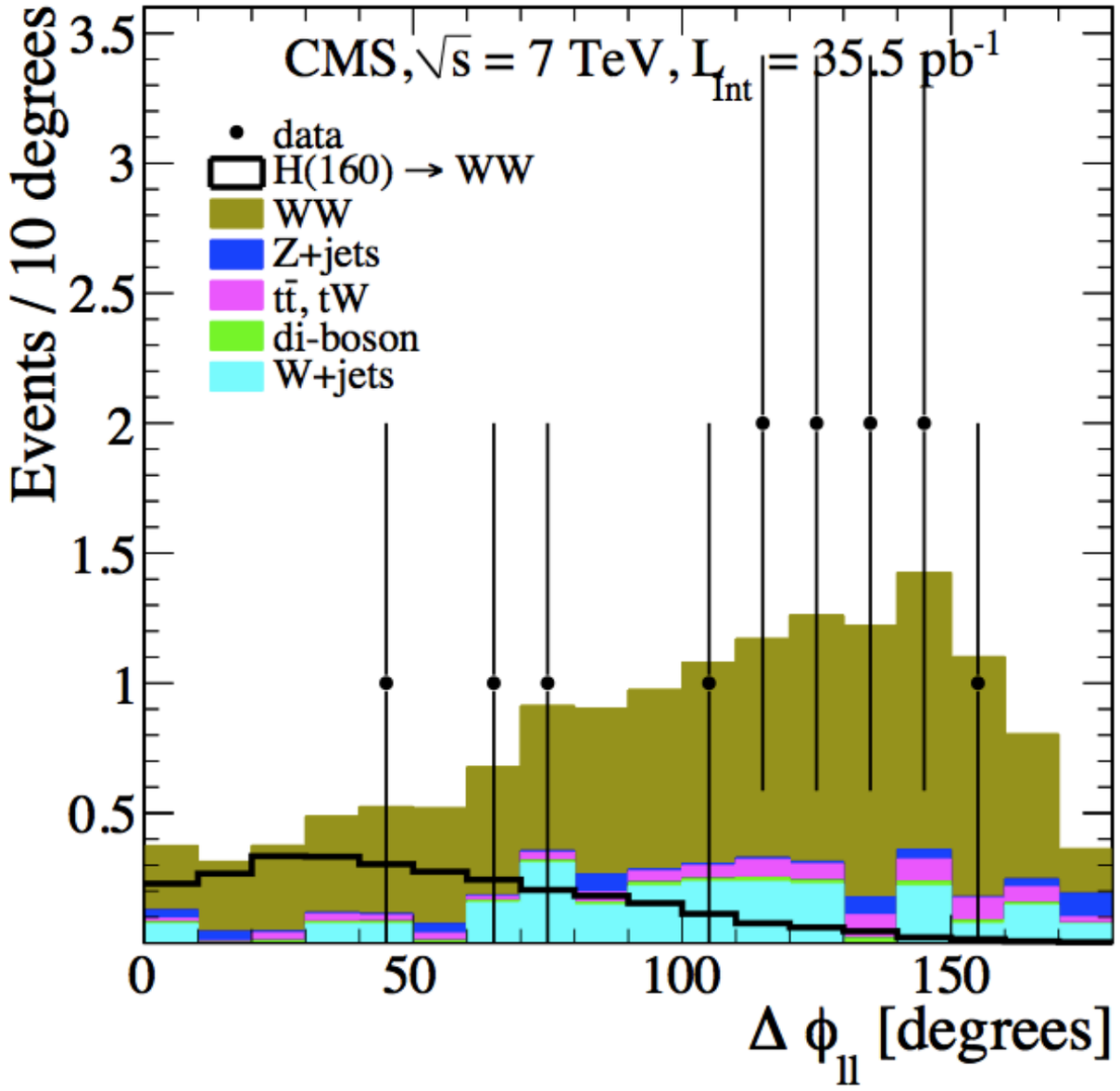
$\sigma = 4.5$ pb	$\sigma = 15$ pb	W: $\sigma = 31\ 10^3$ pb Z: $\sigma = 3.5\ 10^3$ pb
irreducible: same final state of the signal, exploit different kinematics of the production	there are two additional b-jets in the detector, due to the top decay, veto on jets (or on b-jets)	jets in the detector can give a lepton-like signature (non prompt leptons, or fake leptons form track+calo deposit): very high cross section

Signal Features

example of a discriminating variable;

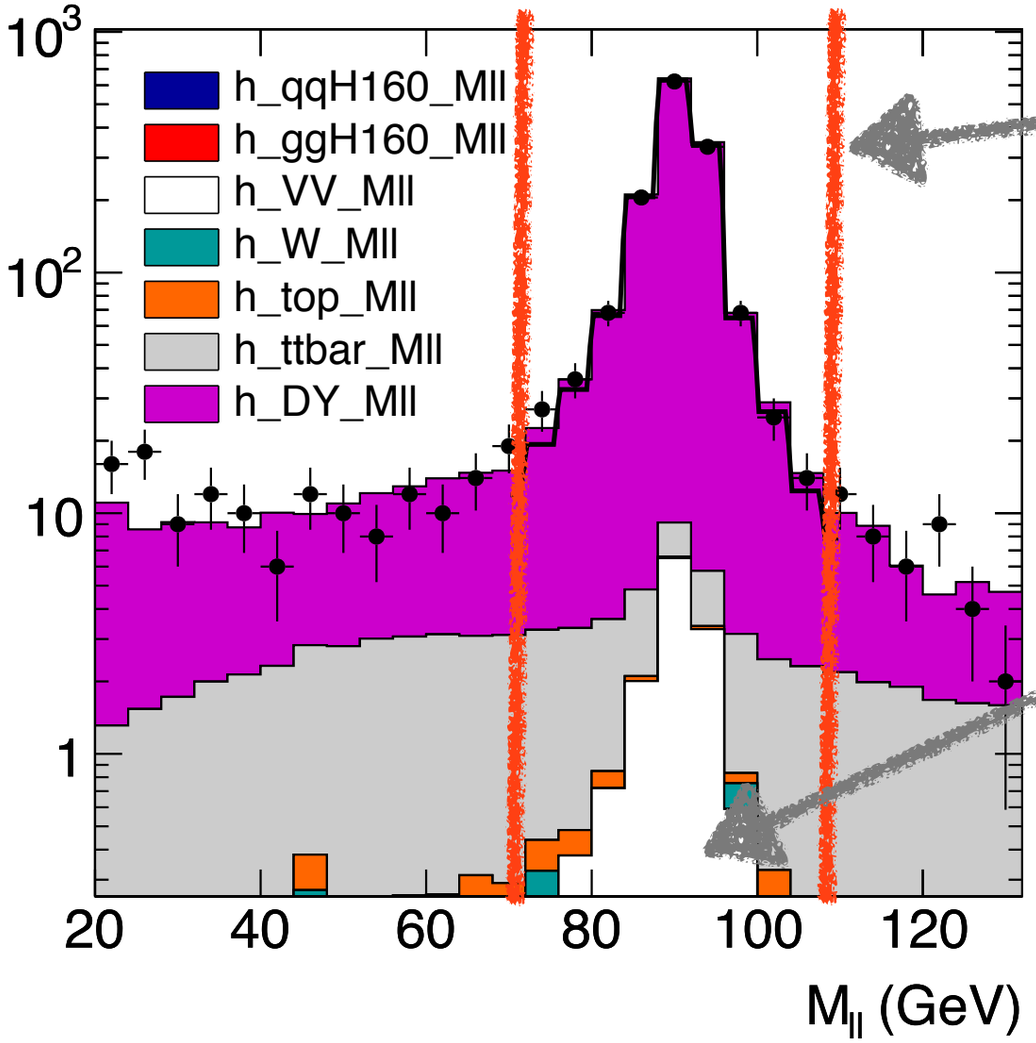


the two leptons are close to each other



Cross-section measurement

$$\sigma = \frac{N_{obs} - N_{bkg}}{\varepsilon \cdot \int \mathcal{L} dt}$$



the Z contribution is dominant under the Z resonance peak

the uncertainty is due to the statistics available and the contamination of other samples

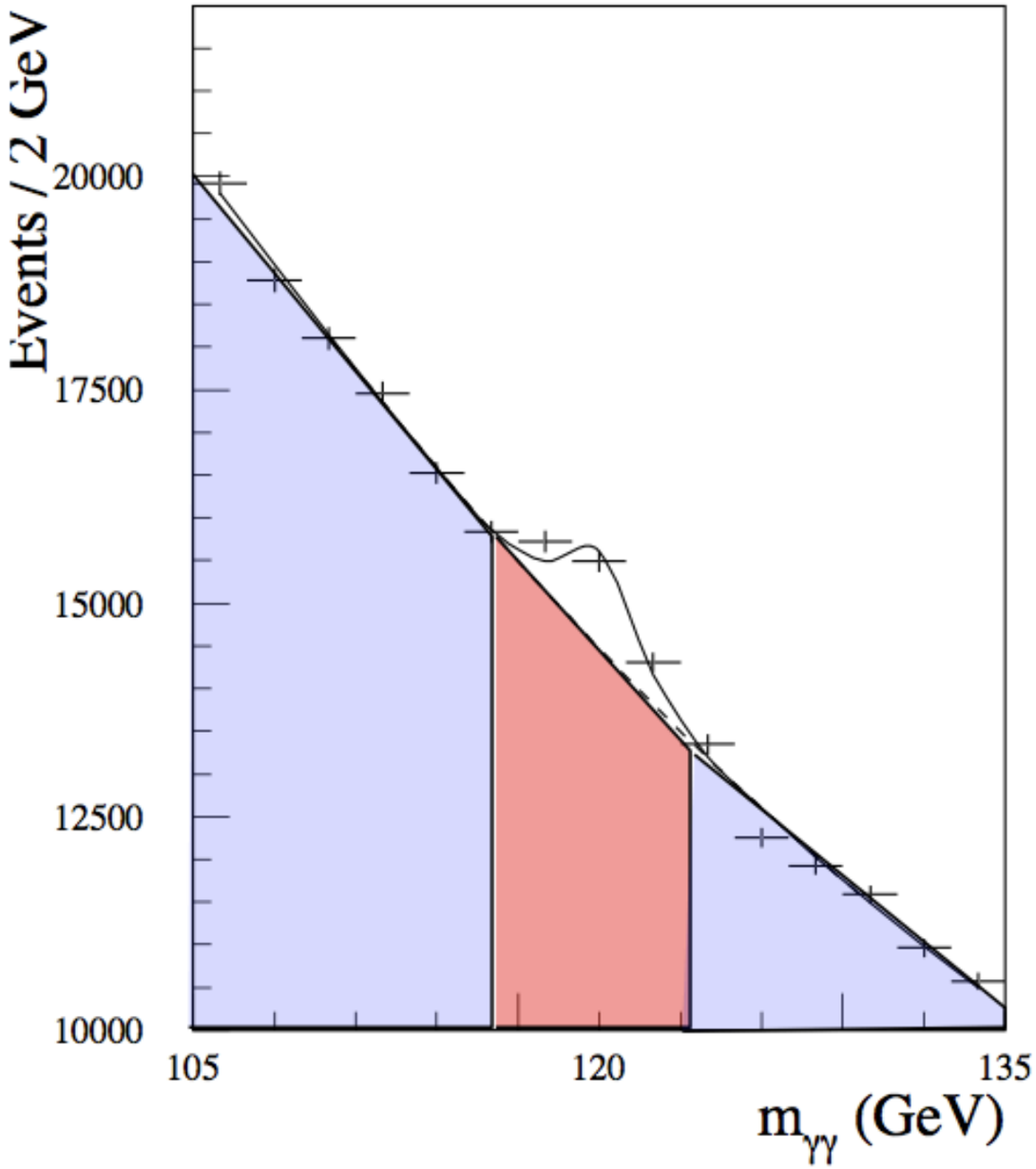
evaluate the **impact on the analysis**: probably does not need the same precision as a cross-section measurement

Side-bands method

- when the background is **expected to behave smoothly**, for example in case of random combinations

- assume a simple shape, and **extrapolate the background under the signal peak** from the sides

- **fit the distribution** with signal shape (a resonance) and background (exponential, linear)



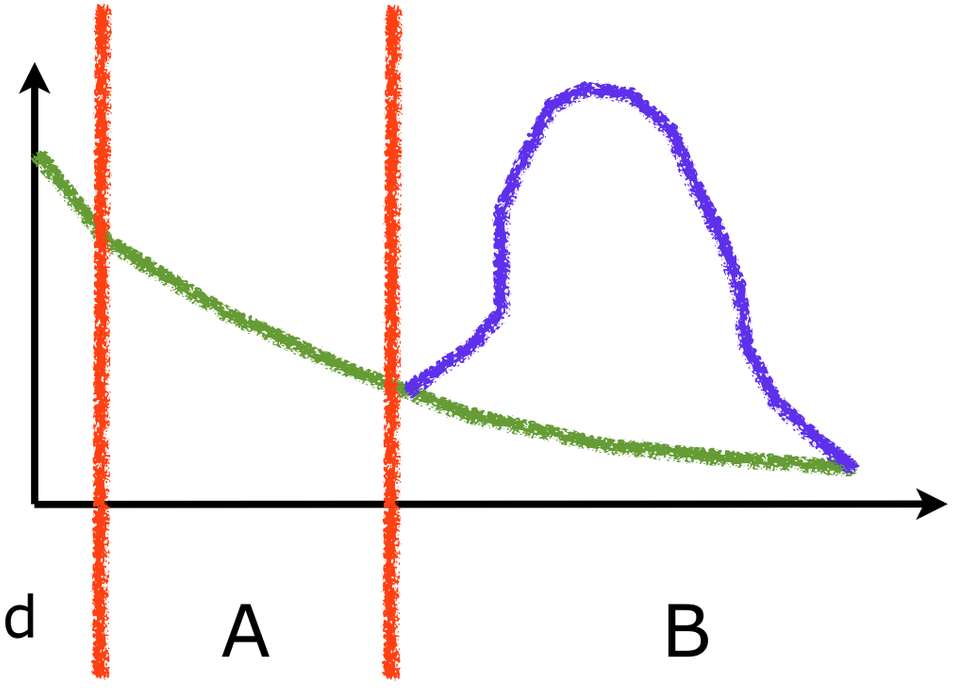
Control region

- assume the **knowledge of the background shape**, at a certain level of the analysis
- fit the shape to data in a **signal-free region**, where that background is dominant and extrapolate it to the signal region
- in case of low statistics, **count the number of events** and extrapolate

$$N^{bkg-A}_{\text{inferred}} = N^{bkg-B}_{DATA} \left(\frac{N^{bkg-A}_{MC}}{N^{bkg-B}_{MC}} \right)$$

uncertainty due to the data statistics and other systematic sources

uncertainty due to: background model, control region contamination, propagation of other systematic sources, MC stats

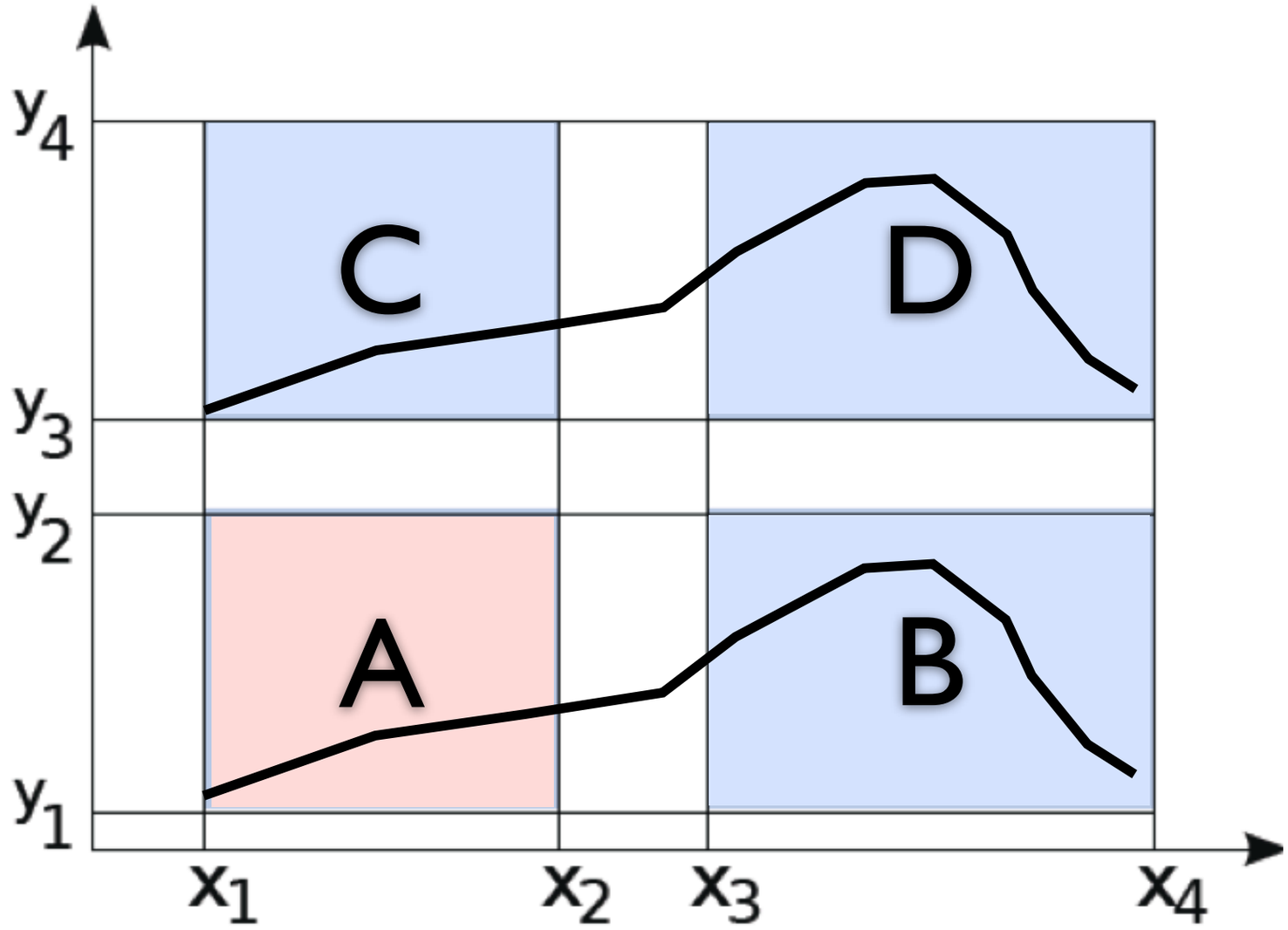


ABCD method

$$N_A^{bkg} = N_B^{bkg} \frac{N_C^{bkg}}{N_D^{bkg}}$$

uncertainty due to the data statistics and other systematic sources

uncertainty due to: control region contamination, propagation of other systematic sources

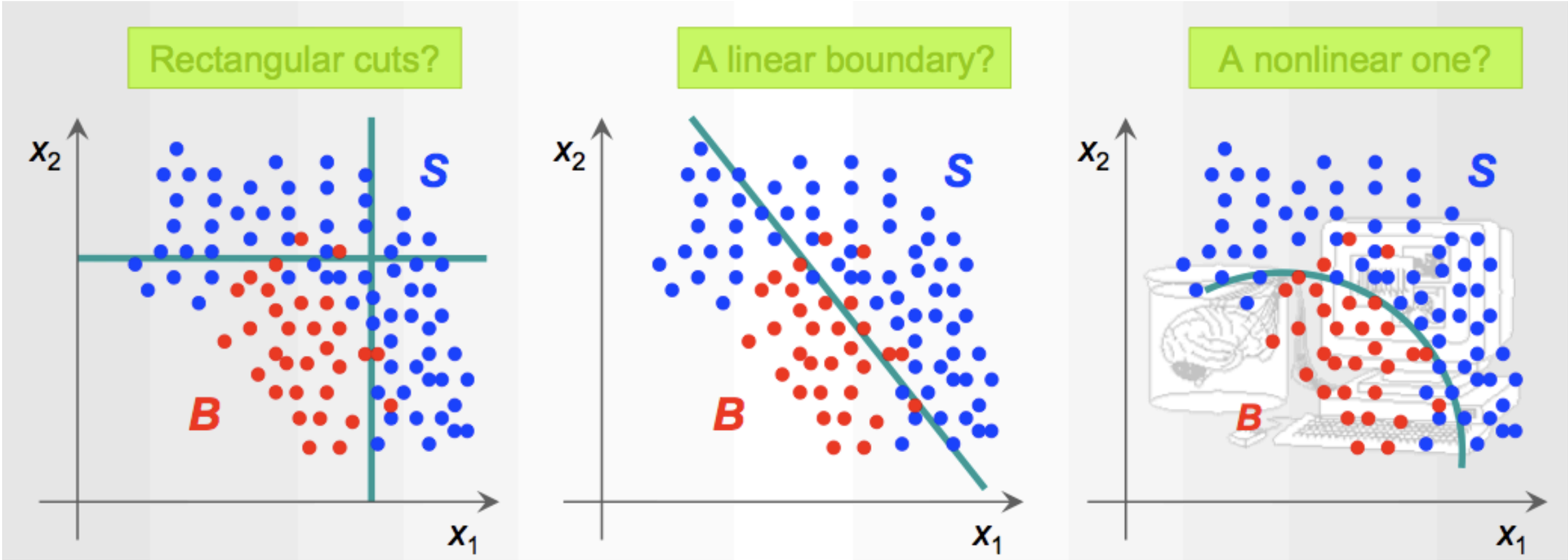


- assume the bkg pdf to be factorized: $f^{bkg}(x,y) = f_x^{bkg}(x) \cdot f_y^{bkg}(y)$
- the **correlation check** done with simulation is a less stringent requirement than the good description of the shape
- in case of low statistics, **count the number** of events and extrapolate

Multivariate Techniques

- **Toolkit for Multivariate Data Analysis with ROOT,** http://tmva.sourceforge.net/
  H. Voss, **Multivariate Data Analysis and Machine Learning in High Energy Physics**

Signal vs Background

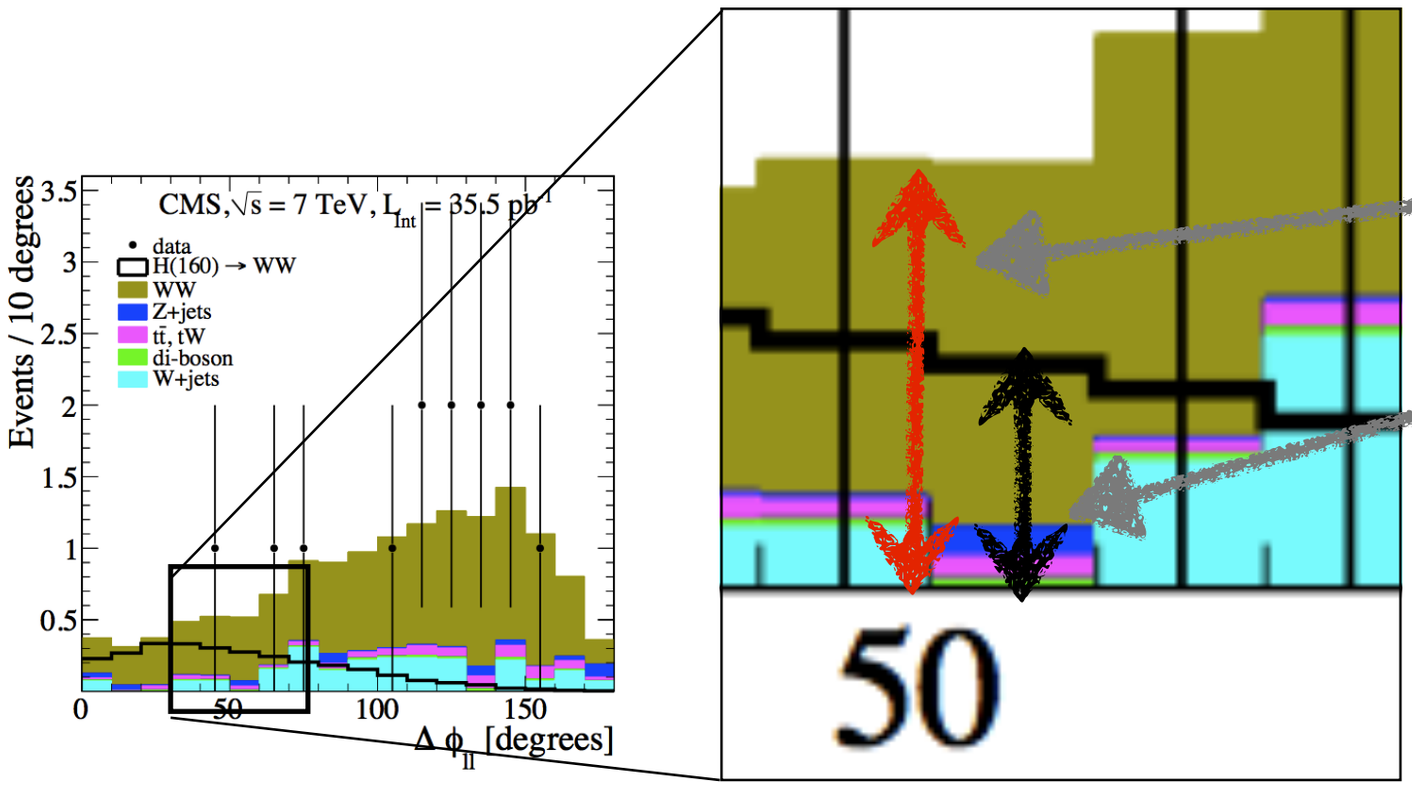


- rectangular selections do not fully exploit the **topology of the events**
- build **more sophisticated discriminants** to separate signal from backgrounds
- need a **good knowledge** of both signal and background
- need high Monte Carlo **statistics**

Likelihood Discriminant

search for a classification of the events, that maps the set of the analysis variables into a single one

$$y_i = f(\vec{x}_i) : \mathbb{R}^n \rightarrow \mathbb{R}$$



$$R_L(i) = \frac{L_S(i)}{L_B(i) + L_S(i)}$$

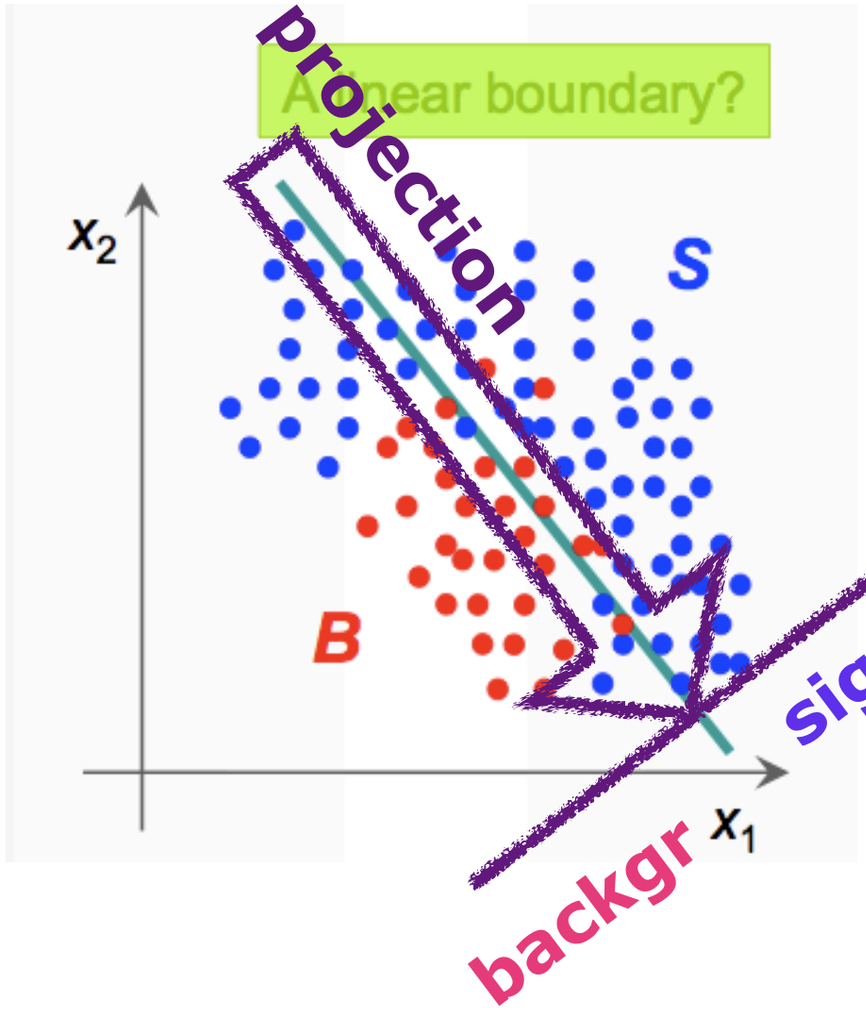
estimates for each event the confidence of being signal-like

For more, uncorrelated, variables: "easily" built
For linearly correlated variables, first decorrelate them

$$L_S(i) = f_S(\vec{x}_i) = \prod_{j \epsilon (\mathrm{vars})} f_{S,j}(x_{ij})$$

Fisher Discriminant

- **project high-dimensional dataset onto a line** and perform classification in this one-dimensional space
- **optimization**: maximize the distance between means, while minimizing the variance within each class
- very effective with **linear correlations**



build the linear combination:

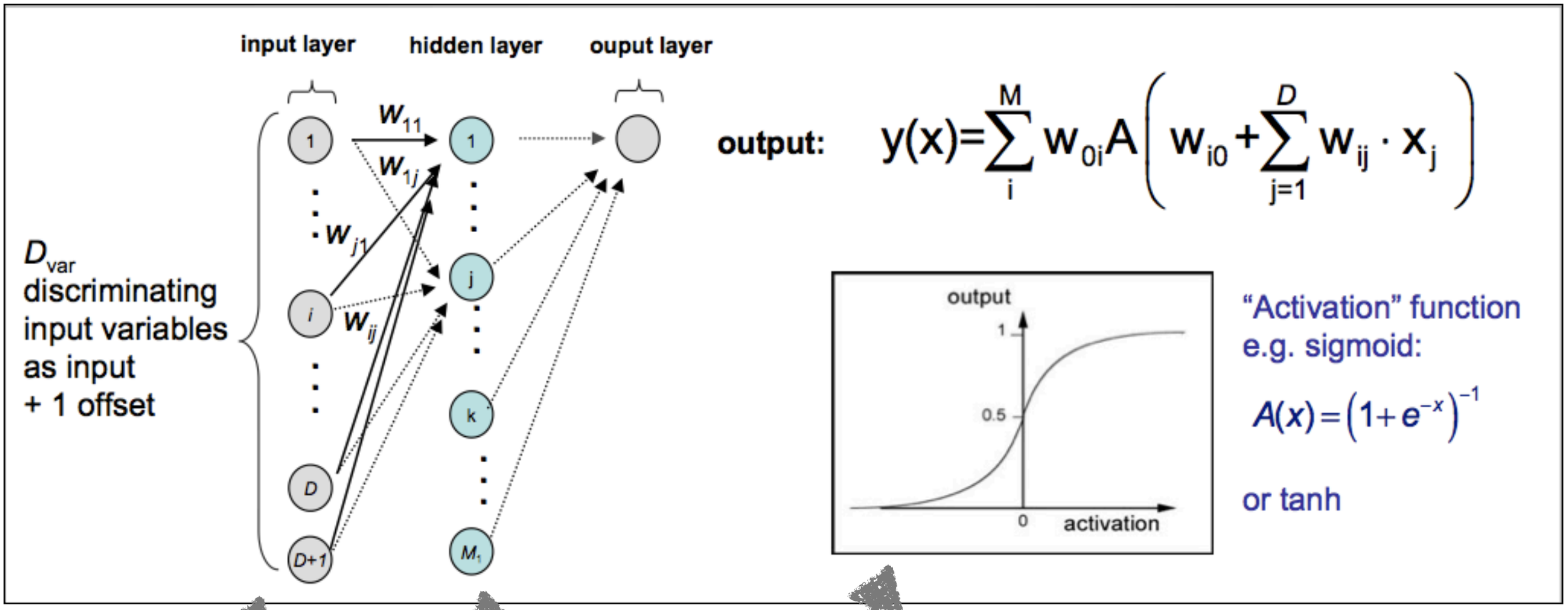
$$y(\vec{x}, \vec{w}) = w_0 + \sum x_i \cdot w_i$$

find the best weight by minimizing the criterion:

$$J(\vec{w}) = \frac{\left(\langle y_S \rangle - \langle y_B \rangle\right)^2}{\sigma^2_{y_S} + \sigma^2_{y_B}}$$

Neural Network

- to cope with non-linear correlations, try a more sophisticated combination of the inputs



input variables, none of them is a smoking gun

factors of the "base" in which the non linear y is decomposed

the non-linear base element

need to find the weights, i.e. to train the neural network

On the training

loss function: how many times I make a mistake in the classification

$$L(w) = \sum_{i}^{events} (y(x_i) - y(C))^2$$

$y(x_i)$: predicted event type; $y(C)$: true event type

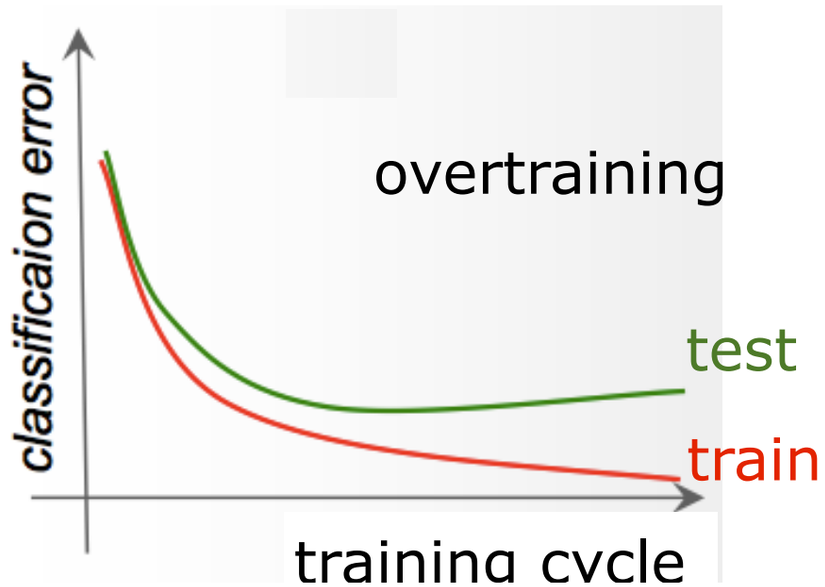
values of y:
1 = signal
0 = background

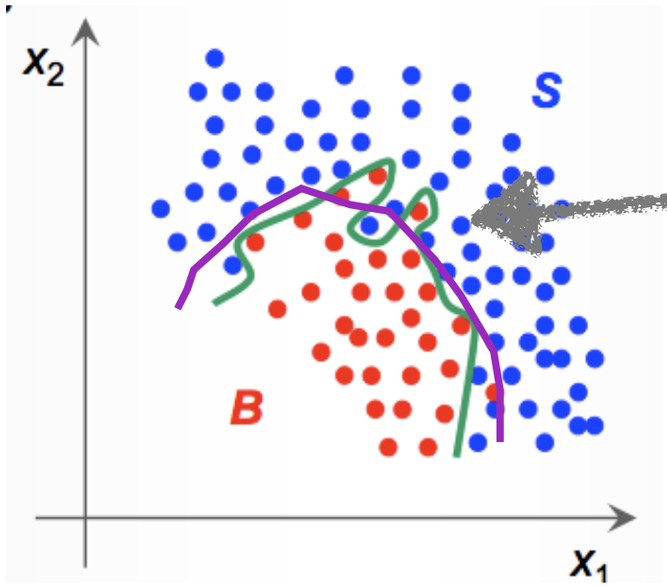
minimize the loss function:

- start from random weights
- change them according to the L gradient
- loop several times on the training samples

$$w^{n+1} = w^n + \eta \cdot \vec{\nabla}_w L \quad (w)$$

divide the simulated sample into **training** and **testing**,
continue the training until the performances on the training stabilize,
stop before the ones on the testing worsen

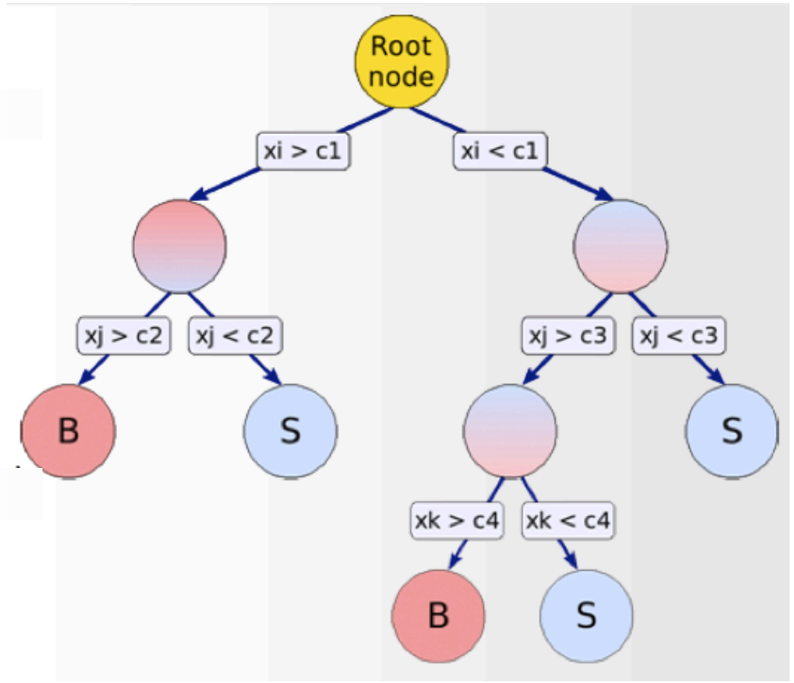


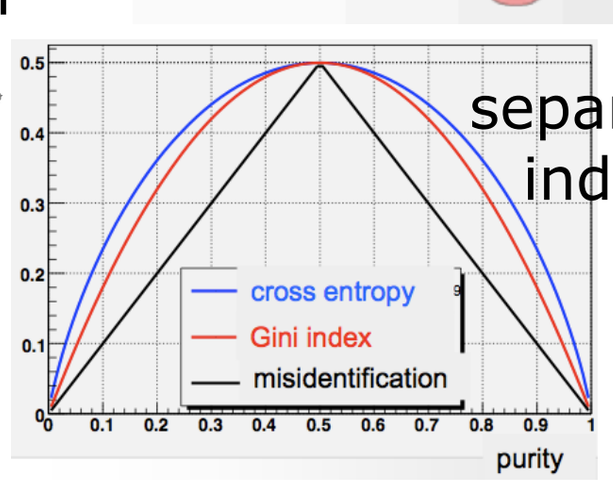


in overtraining, the NN is adapting to statistical fluctuations of the training sample

Boosted Decision Trees

- rank the variables in terms of **discriminating power**
- apply **subsequent selections** in each of the variables
- **stop** when:
  - minimal #events per node
  - maximum number of nodes
  - maximum depth specified
  - a split doesn't give a minimum separation gain





separation indices

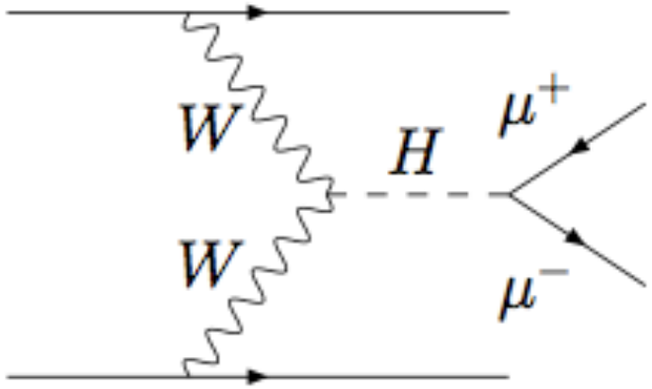
- in each final node (leaf) return S/B discrimination (discrete or continuous)
- independent of monotonous variable transformations
- immune against outliers
- weak variables are ignored
- very sensitive to statistical fluctuations in training data

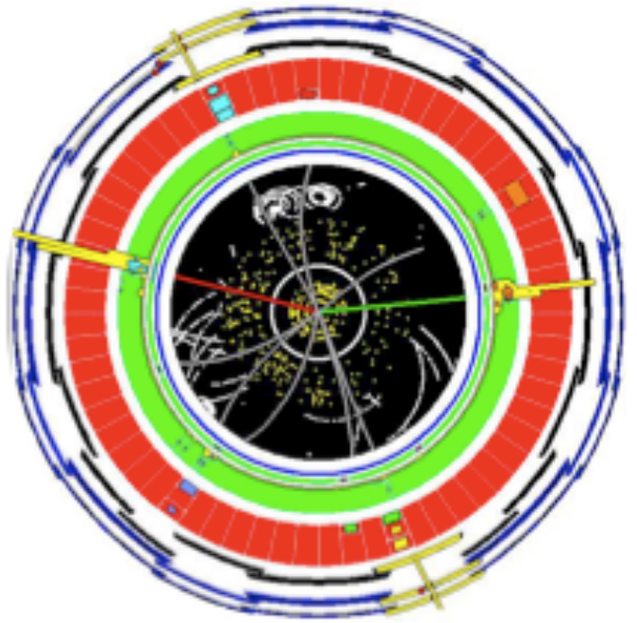
boosting: combine a whole forest of Decision Trees, derived from the same sample, e.g. using different event weights.

Matrix Element

- the MVA techniques is **describe the final state topology** with a parametrization, built on the simulation
- matrix elements are this description, at the level of the physical process

need to include the effects due to the detector for each physics object considered





$$P(\mathbf{x}|M_t) = \frac{1}{N}\int d\Phi\ |\mathcal{M}_{t\bar{t}}(p;M_t)|^2 \prod_{jets} f(p_i,j_i) f_{PDF}(q_1) f_{PDF}(q_2)$$

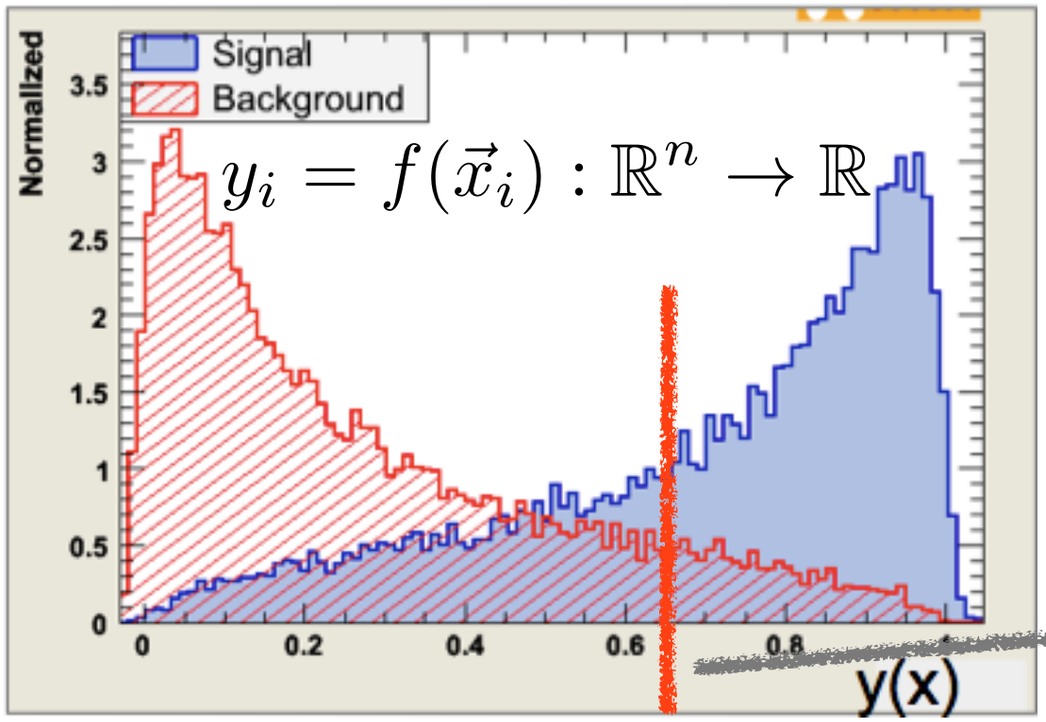
Phase-space Integral

Matrix Element

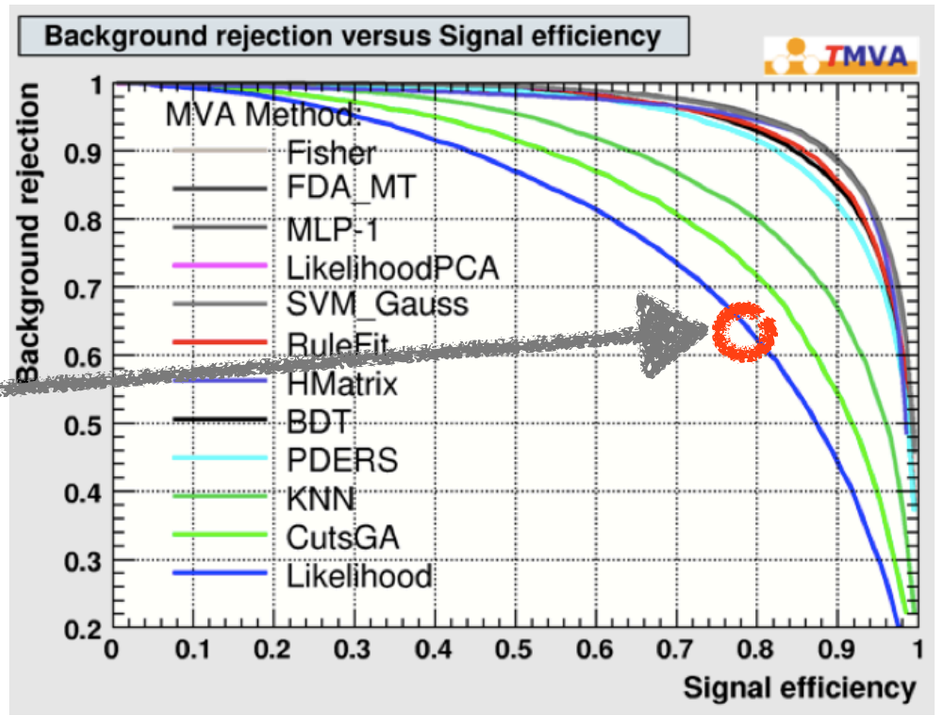
Transfer Functions

calculate the probability for each background and **build a likelihood ratio**

Event Selection Optimization



for each discriminant, now make the choice: what is signal, what is background?



Receiver Operating Characteristics (ROC)
Curve: how efficiency versus purity

- choose the working point by maximizing a figure of merit:

$$\frac{S}{\sqrt{B}}$$

search: sensitivity over background fluctuations

$$\frac{S}{\sqrt{S+B}}$$

known signal: sensitivity over fluctuations of the total sample

$$\frac{S}{\sqrt{B+(\Delta B)^2}}$$

search: sensitivity over background fluctuations plus systematics

Method Comparison

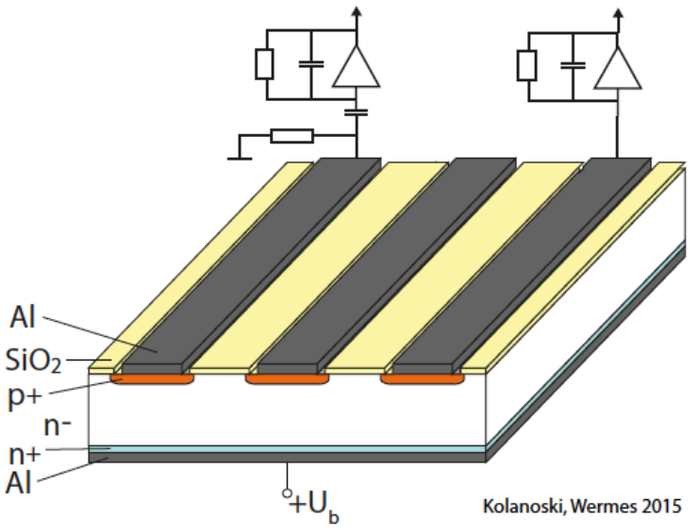
- useful table for the choice of the method to be used, among the ones provided by TMVA

Criteria		Classifiers								
		Cuts	Likeli-hood	PDERS / k-NN	H-Matrix	Fisher	MLP	BDT	RuleFit	SVM
Perfor-mance	no / linear correlations	😐	☺	☺	😐	☺	☺	😐	☺	☺
	nonlinear correlations	😐	☹	☺	☹	☹	☺	☺	😐	☺
Speed	Training	☹	☺	☺	☺	☺	😐	☹	😐	☹
	Response	☺	☺	☹/😐	☺	☺	☺	😐	😐	😐
Robust-ness	Overtraining	☺	😐	😐	☺	☺	☹	☹	😐	😐
	Weak input variables	☺	☺	☹	☺	☺	😐	😐	😐	😐
Curse of dimensionality		☹	☺	☹	☺	☺	😐	☺	😐	😐
Transparency		☺	☺	😐	☺	☺	☹	☹	☹	☹

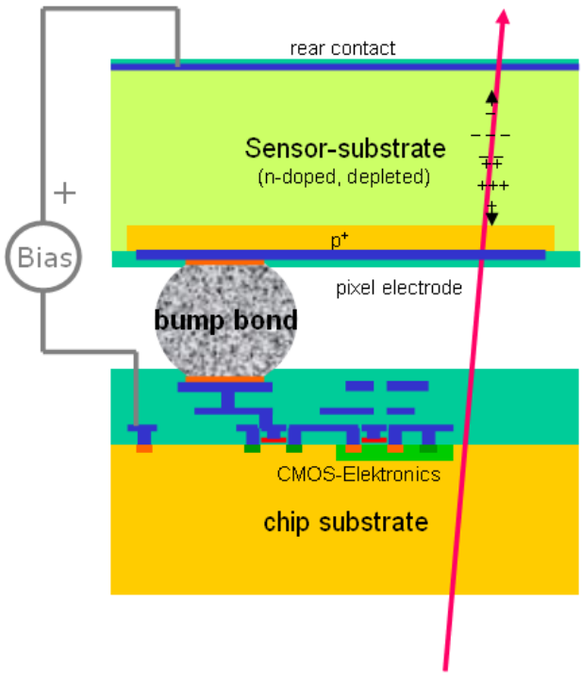
Capacitance of Silicon Pixels

https://arxiv.org/abs/hep-ex/0003032.pdf

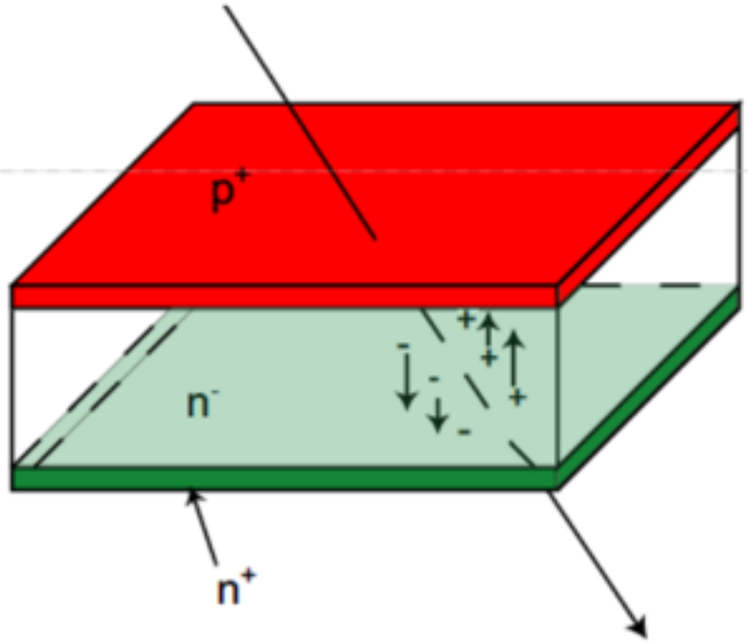
Semiconductor detectors



Microstrip Detector



Pixel Detector



in Si bulk fully depleted

- w_i = 3.65 eV per e/h
- a high energy particle
  → ~ 80 e/h per µm
- all charge collected
- ~ 20 000 e/h per 250 µm
  = 3 fC

TCAD Simulation: capacitance calculation

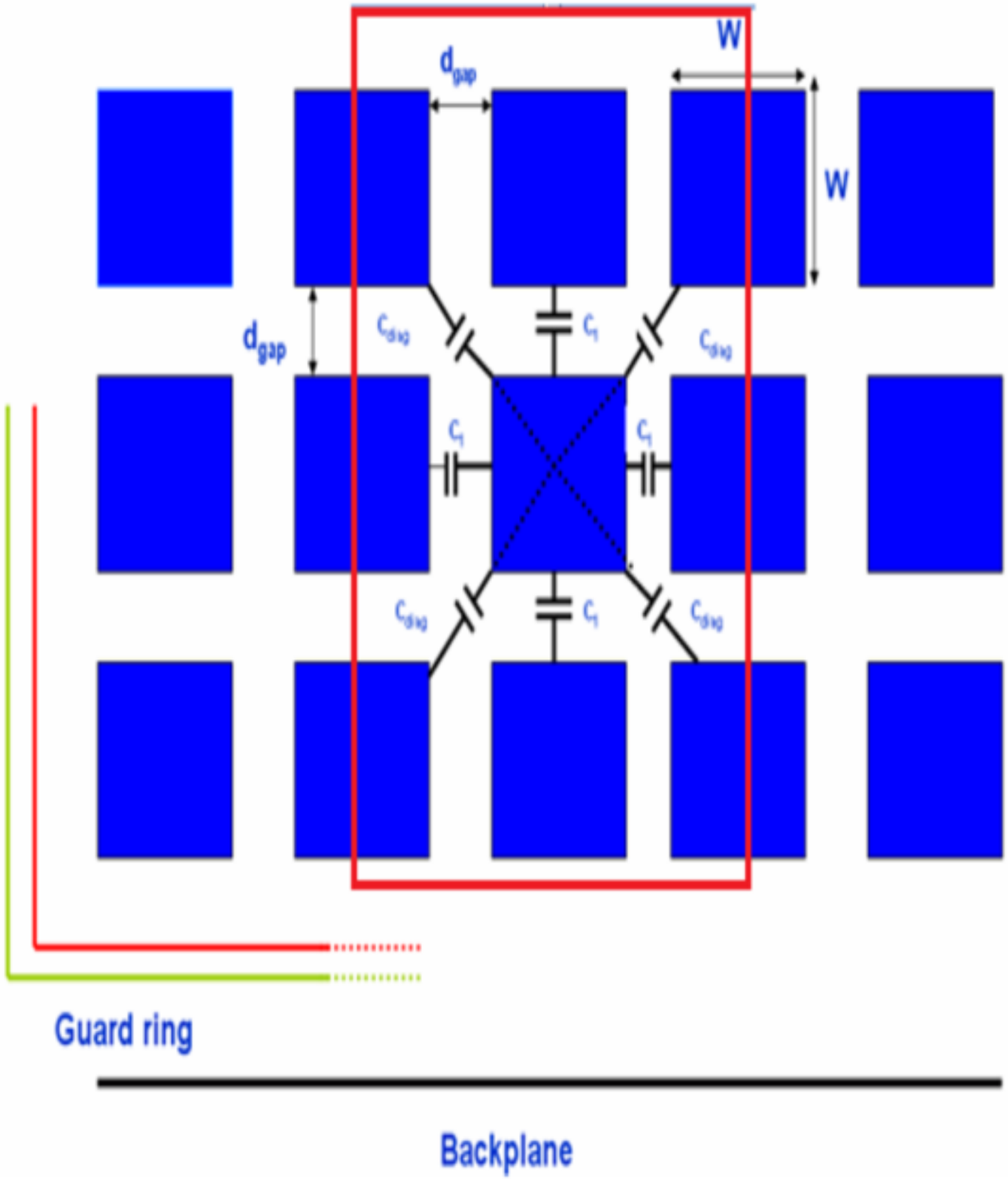


Figure.1: Layout of 5x5 sensor pixel array. The simulated portion is marked by red colour.

Capacitance calculations in p^+n silicon pixel sensors using three dimensional TCAD simulation approach

Ajay K. Srivastava[a,1] E. Fretwurst[a], R.Klanner[a]

[a]*Institute for Experimental Physics, University of Hamburg, Hamburg 22761, Germany*

Internal note (within AGIPD collaboration)

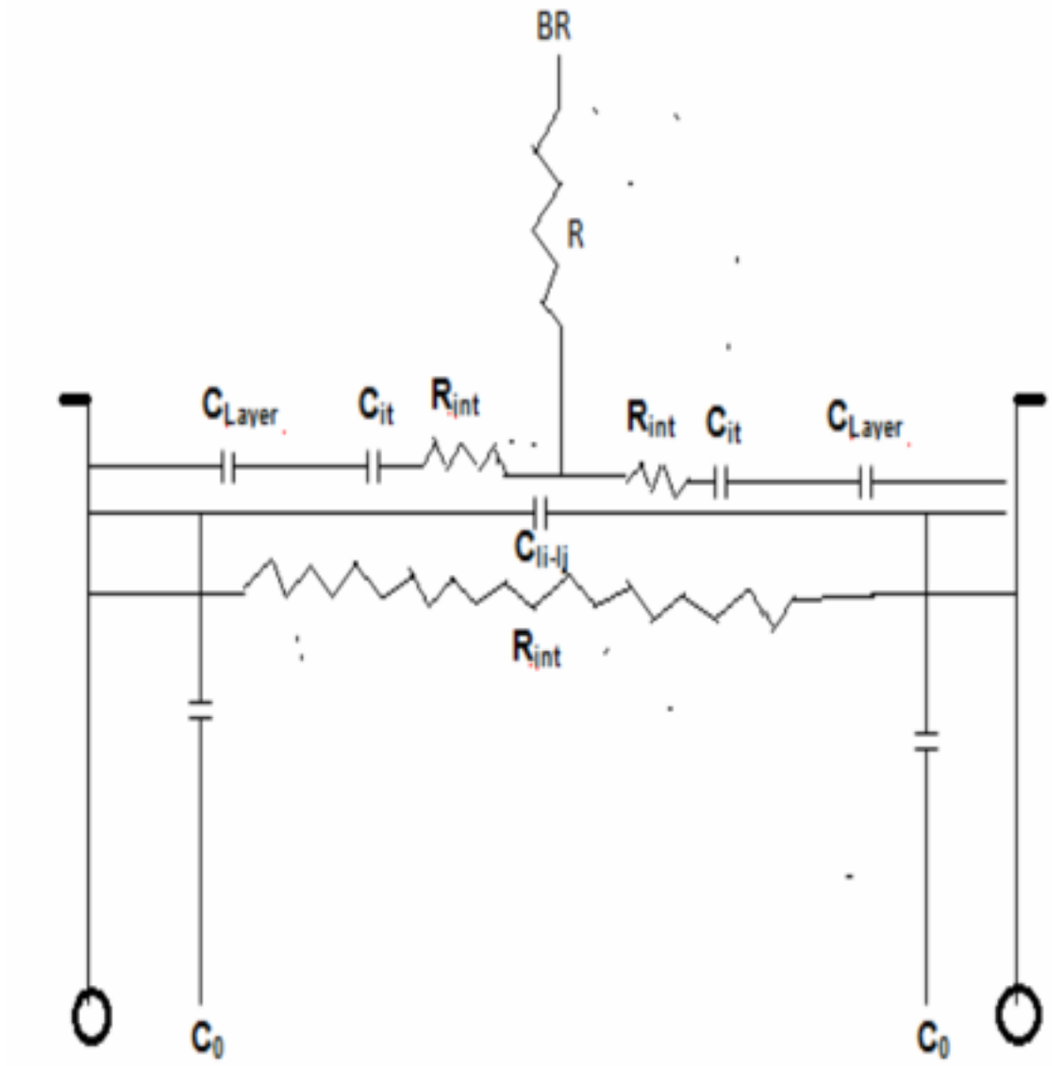


Figure.2: Two lumped model for the C/V curve description of two adjacent strip/pixel sensor.

Simualtion Results

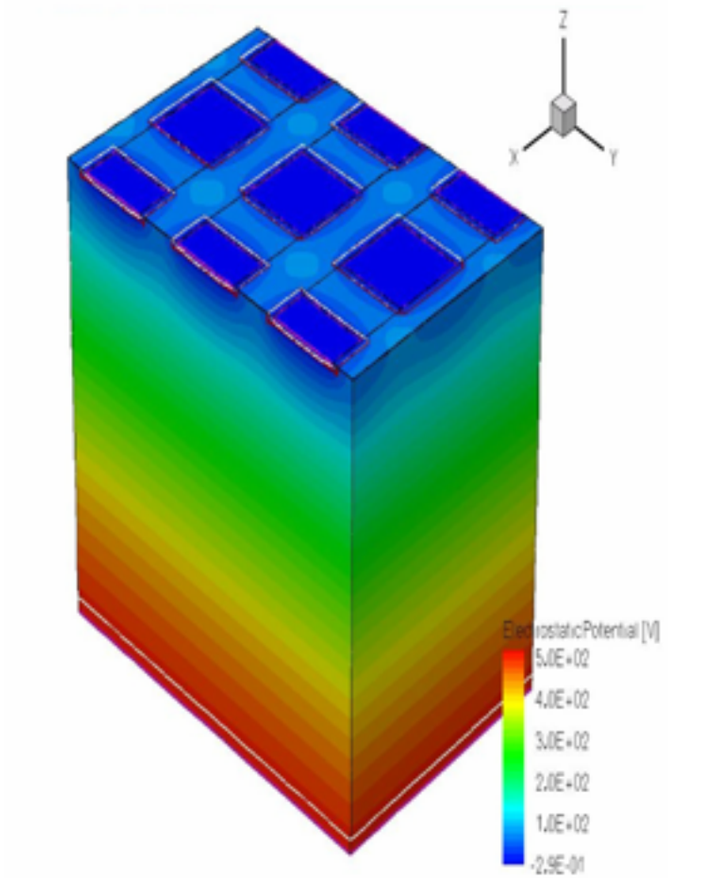


Table.1: Comparison of 3-D simulated capacitances with analytical expressions for 80 µm p^+ pixel gap.

Capacitances	Analytical calculation [4] in fF	Simulation [S] in fF	Error [%]
C_0	7.68	8	5
C_1	5.099	5.48	22
C_{diag}	1.355	1.94	34

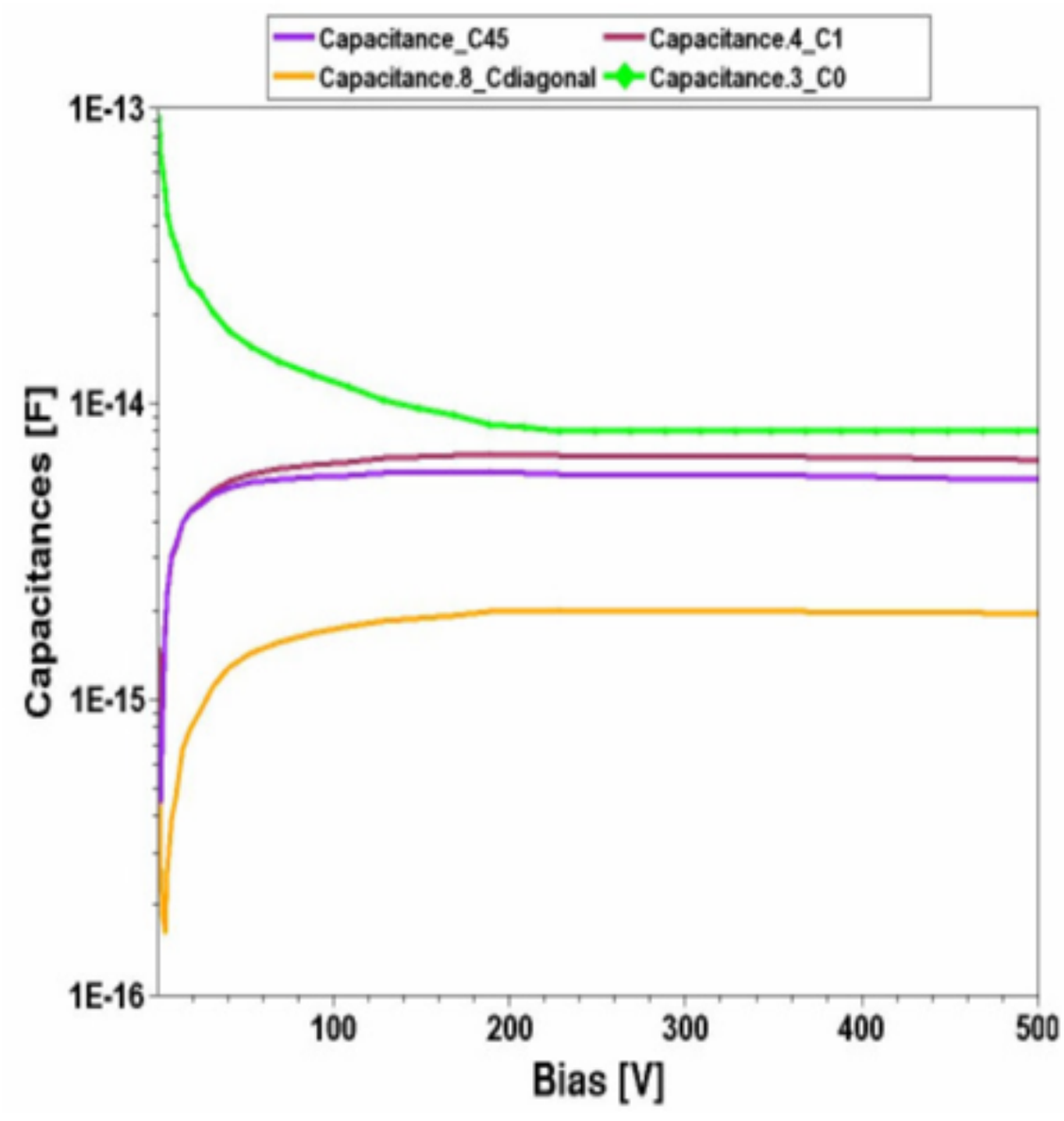


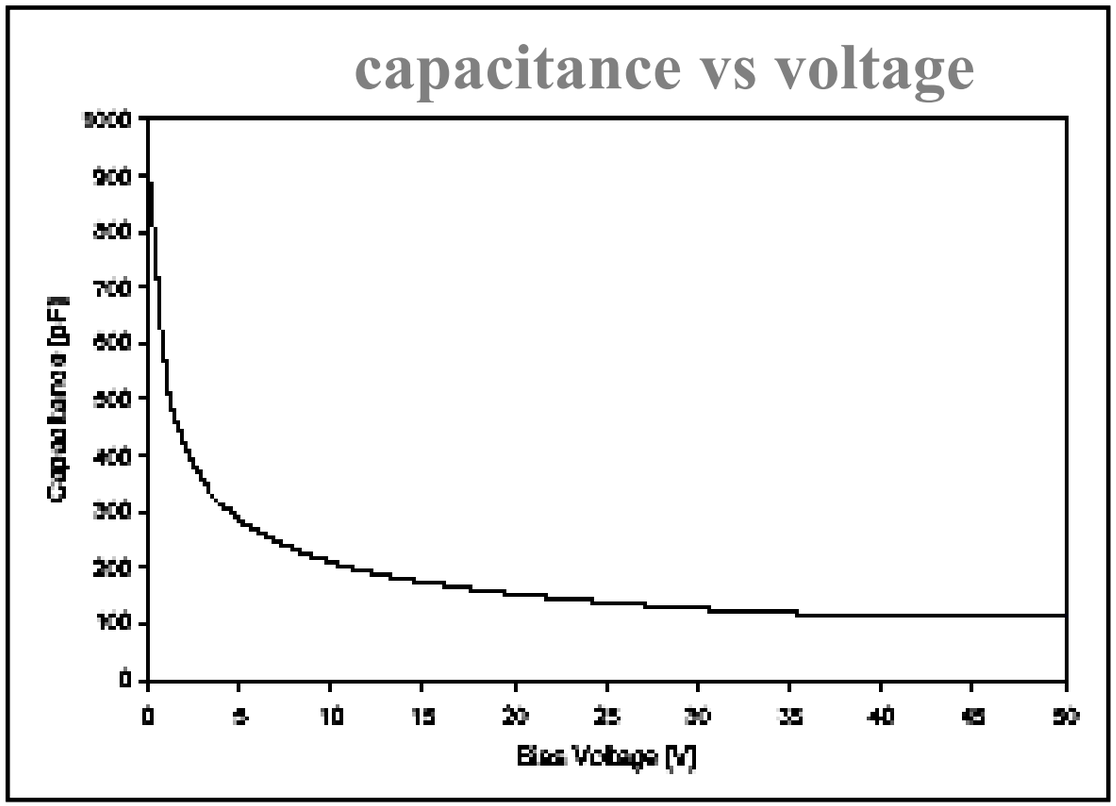
Figure.4: Interpixel capacitance as a function of the appli voltage at 1 MHz.

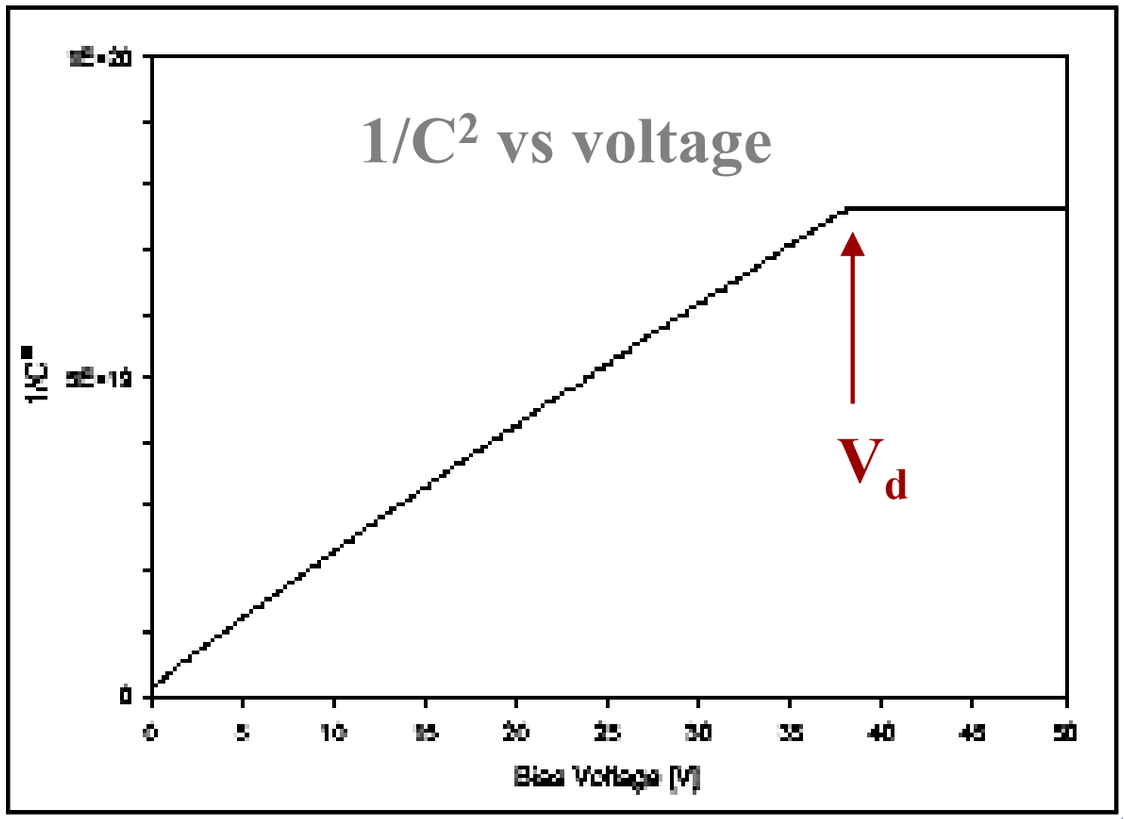
Detector Capacitance

- Capacitance is similar to parallel-plate capacitor
- Fully depleted detector capacitance defined by geometric capacitance

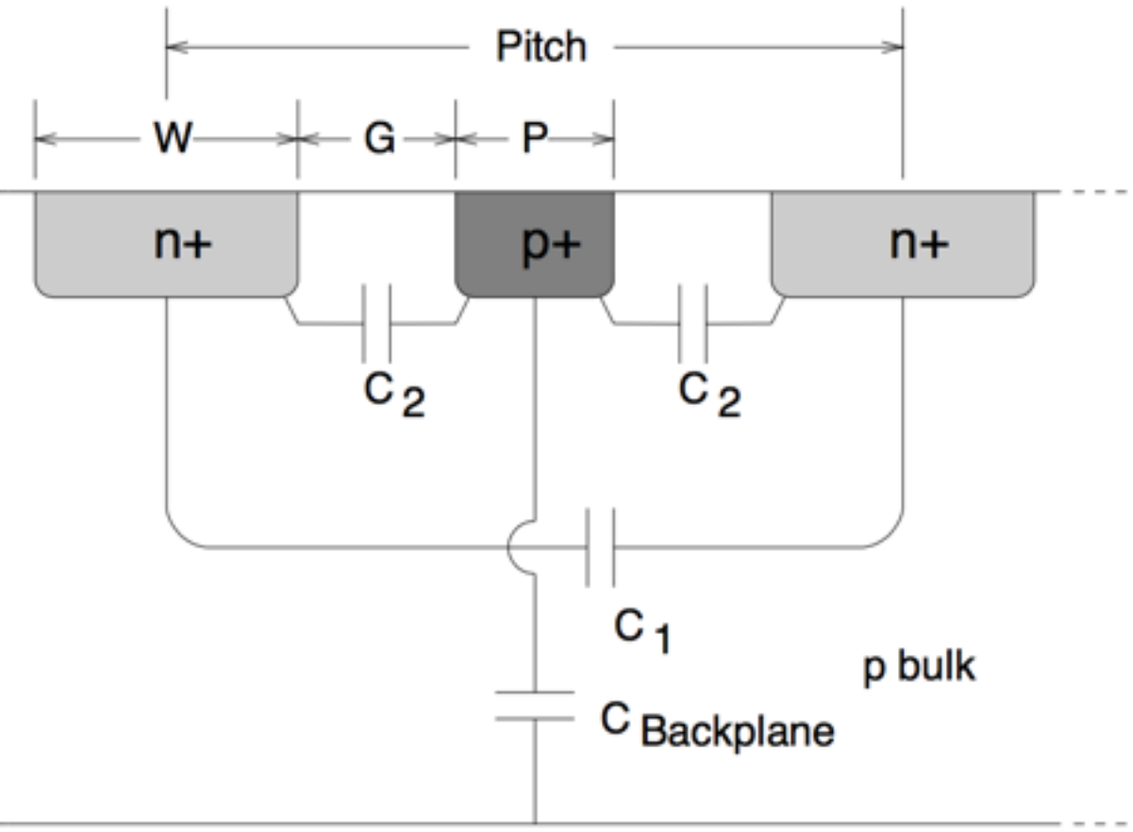
$$C = \sqrt{\frac{\varepsilon_0 \varepsilon_r}{2\mu\rho|V|}} \cdot A$$

One normally measures the depletion behaviour (finds the depletion voltage) by measuring the capacitance versus reverse bias voltage.





Capacitance measurements



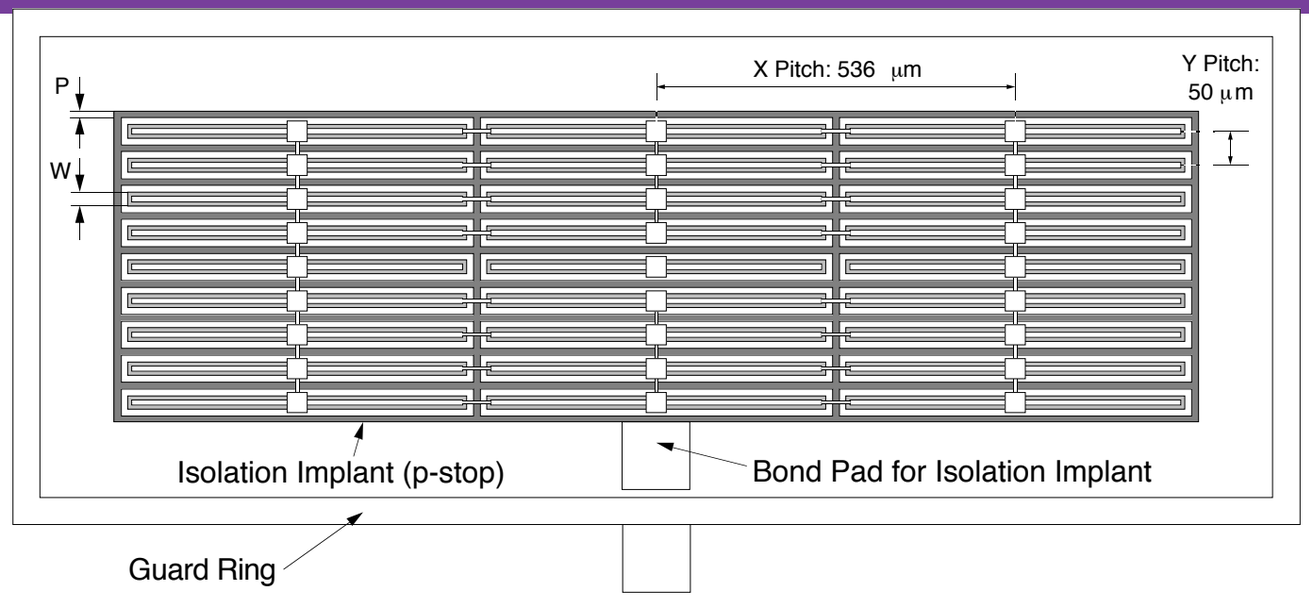
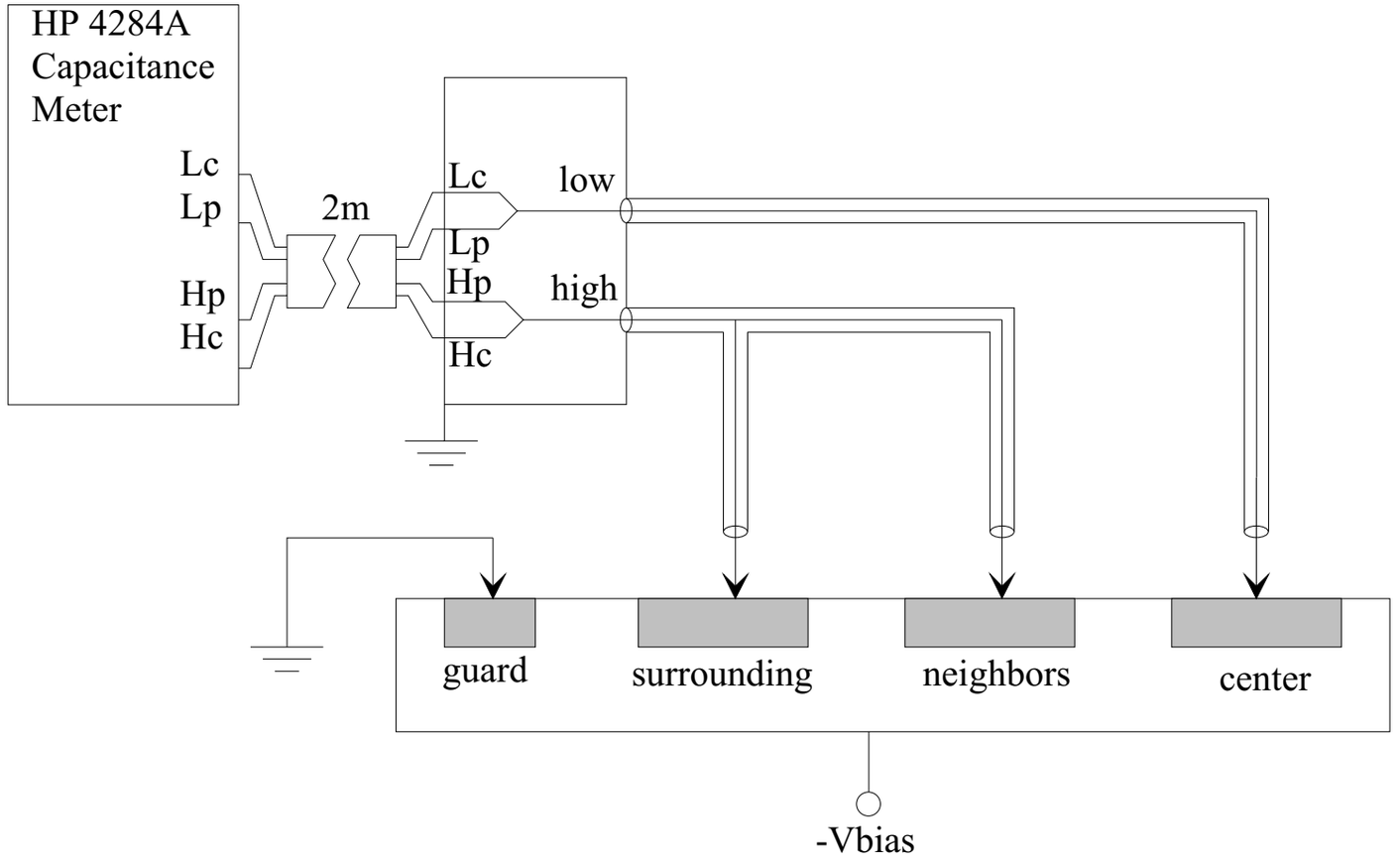


Figure 2: One pixel array in the LBNL p-type test structure.



Results

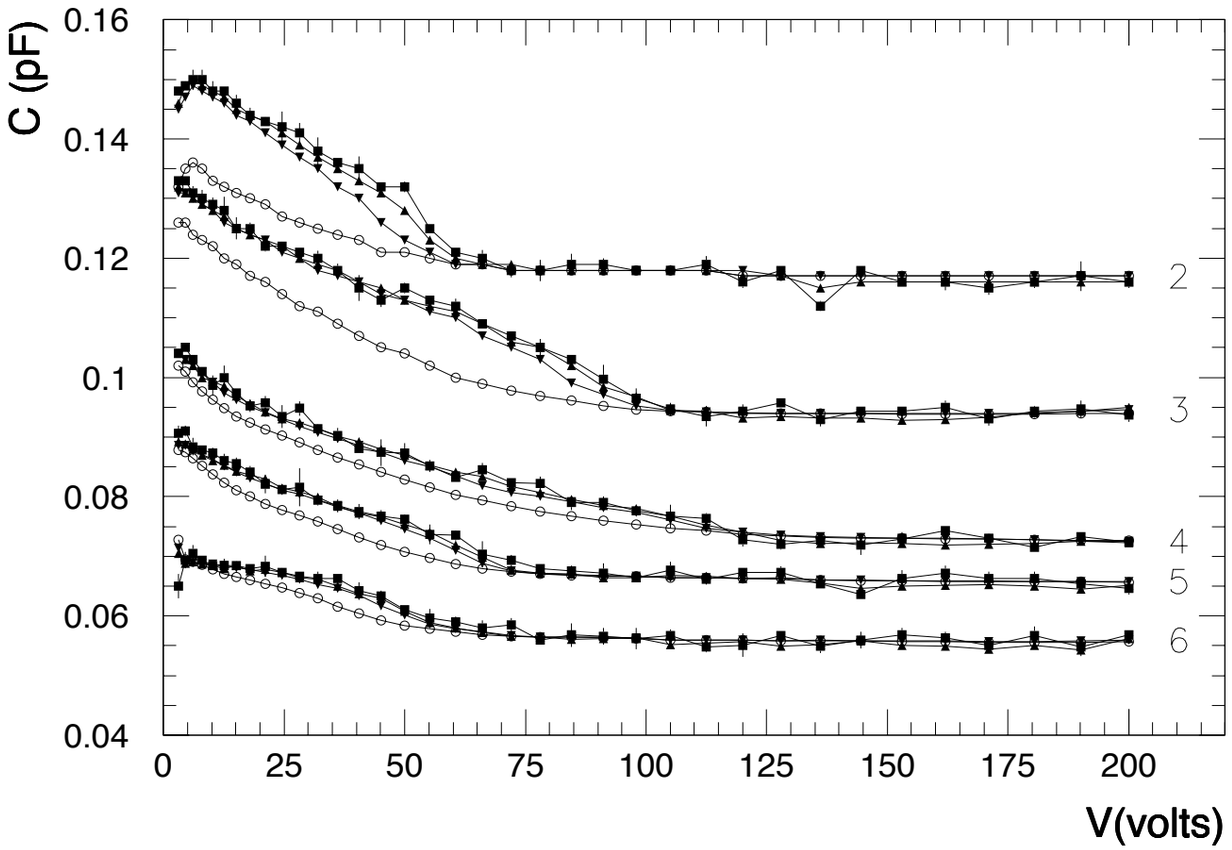


Figure 9: Inter-pixel capacitance of an unirradiated n-type LBNL detector. The multiple curves represent measurements made at frequencies 3, 10, and 100 kHz and 1 MHz. The families of curves labelled 2–6 show measurements of the arrays with the corresponding numbers (see Table 1 for their characteristics). The open circles are the 1 MHz data.

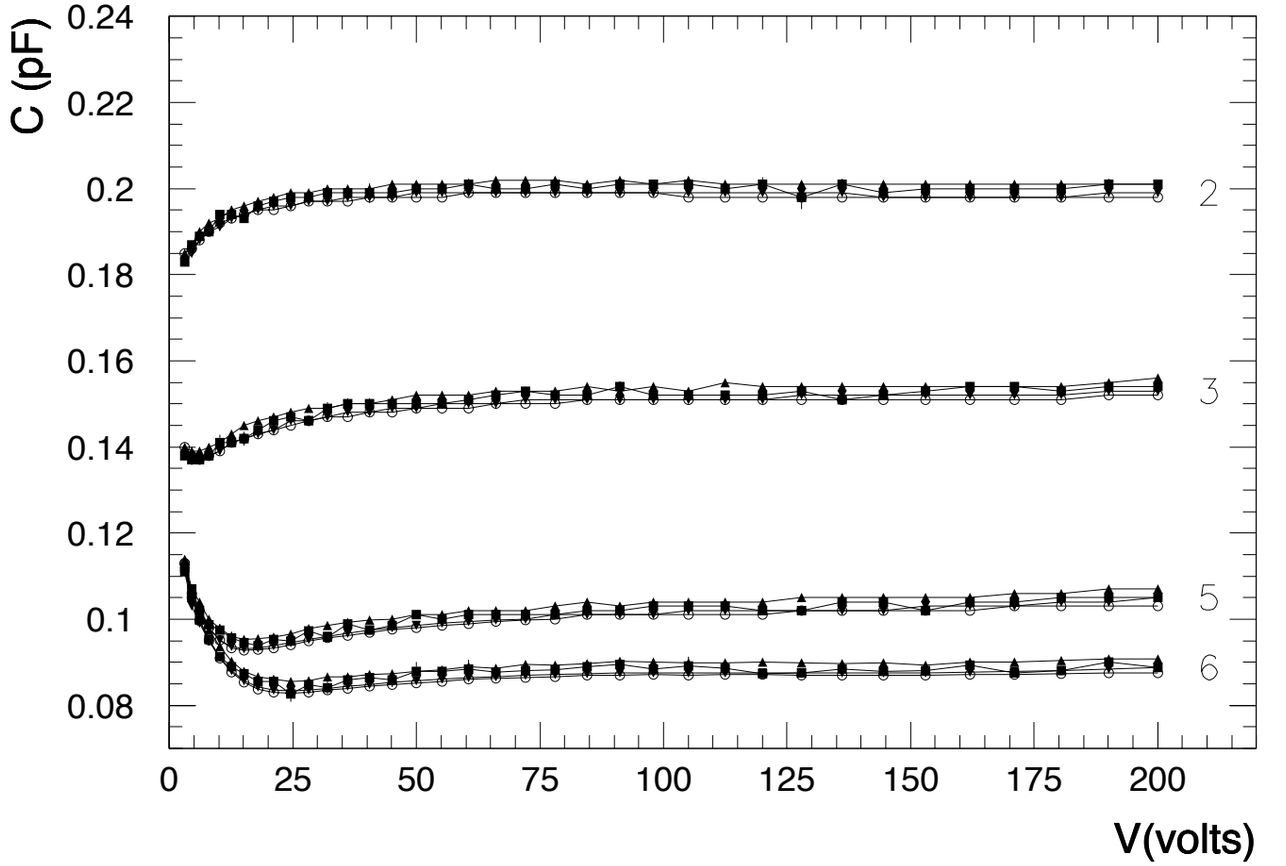
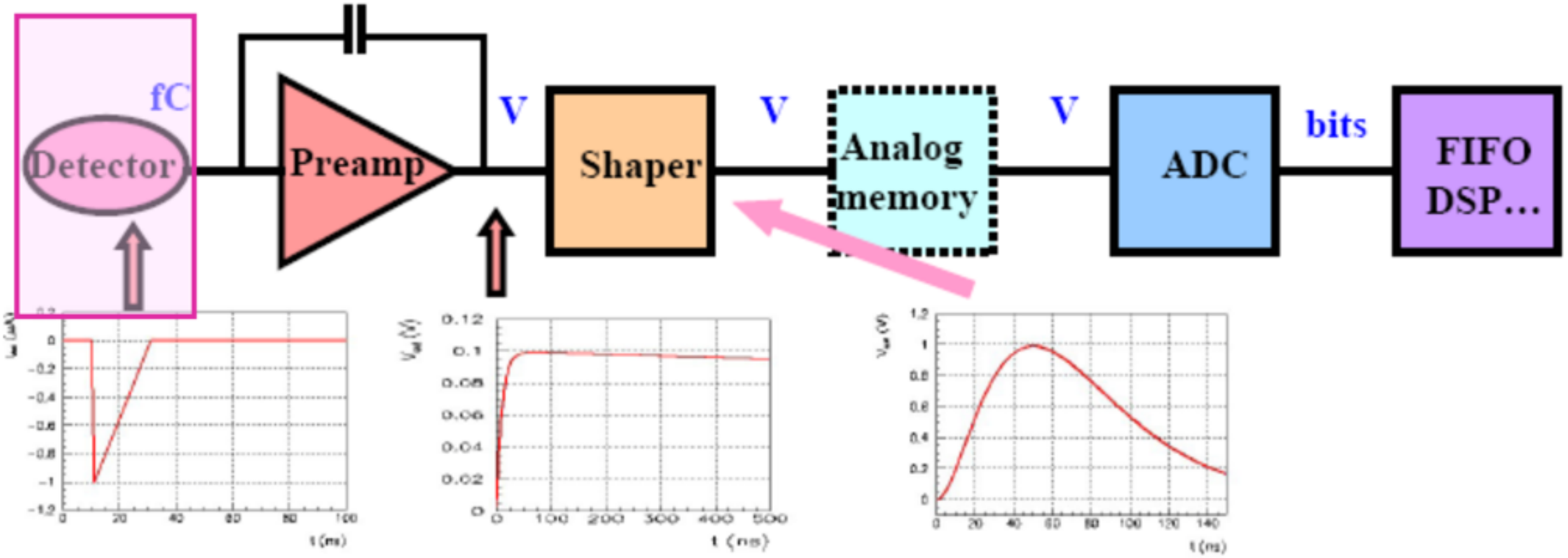


Figure 10: Inter-pixel capacitance of an unirradiated p-type LBNL detector. The multiple curves represent measurements made at frequencies 3, 10, and 100 kHz and 1 MHz. The families of curves labelled 2, 3, 5, and 6 show measurements of the arrays with the corresponding numbers (see Table 1 for their characteristics).

Readout chain

Most front-ends follow a similar architecture



- Very small signals (fC) -> need amplification
- Measurement of amplitude (ADCs)
- Thousands to millions of channels