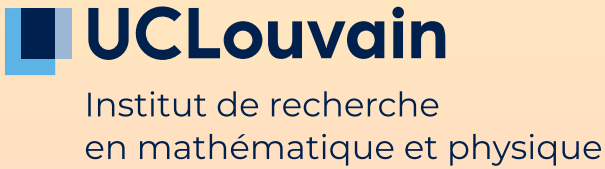
Statistics

or “How to find answers to your questions”

Pietro Vischia[1]

[1] CP3 — IRMP, Université catholique de Louvain



LIP-Lisboa, Statistics Lectures (March 16th and 18th, 2020), Course on Physics at the LHC 2020

Contents

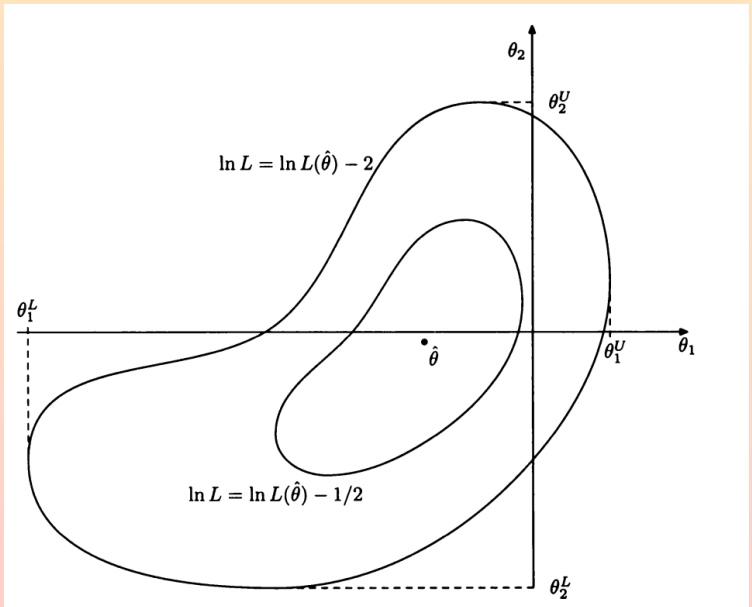
UCLouvain
Institut de recherche
en mathématique et physique

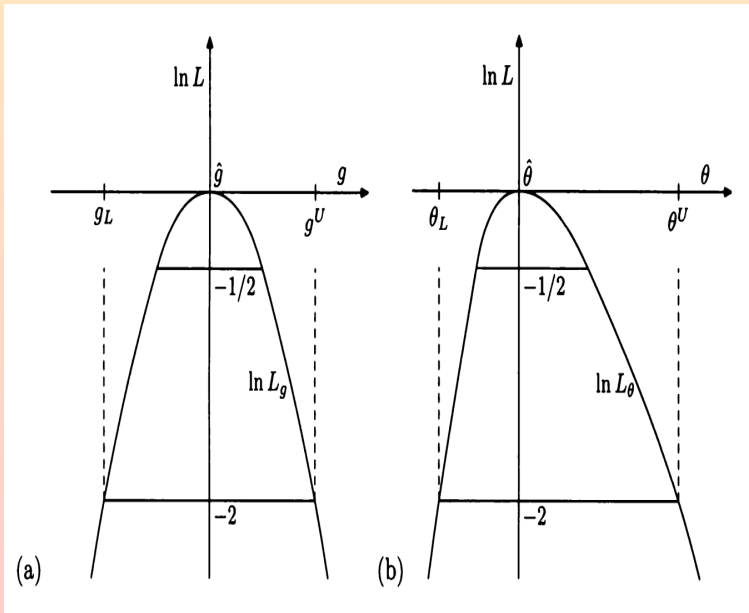
Practicalities

- Schedule: two lessons
  - Monday 16.03, 17h
  - Tuesday 17.03, 17h (this lesson)
- The slides contain links to a few exercises and examples
  - In a longer course there is time to go through them, not in two lessons
  - You are encouraged to play with the exercises offline
- Many interesting references
  - Papers mostly in each slide
  - Some cool books after the summary slide of the second lesson
- Unless stated otherwise, figures belong to P. Vischia, ***** (textbook to be published by Springer in 2021)
- Your feedback is crucial for improving these lectures!

Summary of yesterday

- Theoretical definitions of probability (Kolmogorov, Cox) mostly equivalent
- Practical realizations highlight philosophical differences
  - Frequentist definition: probability is a property of sets of data
  - Bayesian definition: hypotheses and parameters are associated a probability
- Point estimate and interval estimates using the likelihood
  - Statisticians: estimate. Physicists: measurement
  - Parameterize your observable, e.g. w.r.t. reference value ($\mu = \sigma/\sigma_{SM}$), and nuisance parameters $\vec{\theta}$
  - Find a function of μ-only by building likelihood ratio $\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}$
  - Profiling $\hat{\hat{\theta}}(\mu)$ conditional MLEs of the nuisances for each scanned value of μ
  - Can even "freeze" them one by one to extract their contribution to the total uncertainty
- Interval estimate from crossing of the log-likelihood with predetermined values corresponding to Gaussian "sigmas"
  - Log-likehood approximated to gaussian up to $O(1/N)$, therefore probability content slightly larger than the gaussian σ (overcoverage)





Plots from James, 2nd ed.

- Measure N times the same quantity: values x_i and uncertainties σ_i. MLE and variance are:

$$\hat{x}_{ML} = \frac{\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

$$\frac{1}{\hat{\sigma}_x^2} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$$

- The MLE is obtained when each measurement is weighted by its own variance
  - This is because the variance is essentially an estimate of how much information lies in each measurement
- This works if the p.d.f. is known
  - Compare this method with an alternative one that does not assume knowledge of the p.d.f.
  - The second method will be the only one applicable to cases in which the p.d.f. is unknown

Combination of measurements: alternative method 1/

- Take a set of measures sampled from an unknown p.d.f. $f(\vec{x}, \vec{\theta})$
- Compute the expected value and variance of a combination of such measurements described by a function $g(\vec{x})$.
- The expected value and variance of x_i are elementary:

$$\mu = E[x] V_{ij} = E[x_i x_j] - \mu_i \mu_j$$

- If we want to extract the p.d.f. of $g(\vec{x})$, we would normally use the jacobian of the transformation of f to g, but in this case we assumed $f(\vec{x})$ is unknown.

Combination of measurements: alternative method 2/

- We don't know f, but we can still write an expansion in series for it:

$$g(\vec{x}) \simeq g(\vec{\mu}) + \sum_{i=1}^{N} \left(\frac{\partial g}{\partial x_i}\right)\Big|_{x=\mu} (x_i - \mu_i)$$

- We can compute the expected value and variance of g by using the expansion:

$$E\left[g(\vec{x})\right] \simeq g(\mu), \qquad (E[x_i - \mu_i] = 0)$$

$$\sigma_g^2 = \sum_{ij=1}^{N} \left[\frac{\partial g}{\partial x_i}\frac{\partial g}{\partial x_j}\right]\Big|_{\vec{x}=\vec{\mu}} V_{ij}$$

- The variances are propagated to g by means of their jacobian!
- For a sum of measurements, $y = g(\vec{x}) = x_1 + x_2$, the variance of y is $\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}$, which is reduced to the sum of squares for independent measurements

Combination of measurements: example 1/

- Let's compare the two ways of combining measurements, and check the role of the Fisher Information
- Let's estimate the time taken for a laser light pulse to go from the Earth to the Moon and back (in units of Earth-to-Moon-Time EMT)
  - On the Moon we have a receiver built by NASA. It's very good but placed in unfavourable conditions, yielding only a 2% precision on Earth-to-Moon
  - On Earth we have a receiver made out of scrap material. It is however placed in favourable conditions, yielding a 5% precisionon Moon-to-Earth

$$N_{EM} = 0.99 \pm 0.02\ EMT$$

$$N_{ME} = 1.05 \pm 0.05\ EMT$$

- Evidently, the time to moon and back is $N_{EME} = N_{EM} + N_{ME}$, and we can apply Eq. 7: Do it!

Combination of measurements: example 1/

- Let's compare the two ways of combining measurements, and check the role of the Fisher Information
- Let's estimate the time taken for a laser light pulse to go from the Earth to the Moon and back (in units of Earth-to-Moon-Time EMT)
  - On the Moon we have a receiver built by NASA. It's very good but placed in unfavourable conditions, yielding only a 2% precision on Earth-to-Moon
  - On Earth we have a receiver made out of scrap material. It is however placed in favourable conditions, yielding a 5% precisionon Moon-to-Earth

$$N_{EM} = 0.99 \pm 0.02\ EMT$$

$$N_{ME} = 1.05 \pm 0.05\ EMT$$

- Evidently, the time to moon and back is $N_{EME} = N_{EM} + N_{ME}$, and we can apply Eq. 7: Do it!
- Resulting estimate:
  - $N_{EME} = 0.99 + 1.05 \pm \sqrt{0.02^2 + 0.05^2}\ EMT = 2.05 \pm 0.05\ EMT$, corresponding to a precision of $\frac{\sigma_{N_{EME}}}{N_{EME}} \sim 2.4\%$.

Combination of measurements: example 2/

- We now however can argue that over the time it takes for light to go to the Moon and back any environment condition would be roughly constant
- How can we exploit this additional information?

Combination of measurements: example 2/

- We now however can argue that over the time it takes for light to go to the Moon and back any environment condition would be roughly constant
- How can we exploit this additional information?
- We can use this additional information to note that the two estimates N_{EM} and N_{ME} are independent estimates of the same physical quantity $\frac{N_{EME}}{2}$
- Compute N_{EME} and $\sigma(N_{EME})$based on this reasonment

Combination of measurements: example 2/

- We now however can argue that over the time it takes for light to go to the Moon and back any environment condition would be roughly constant
- How can we exploit this additional information?
- We can use this additional information to note that the two estimates N_{EM} and N_{ME} are independent estimates of the same physical quantity $\frac{N_{EME}}{2}$
- Compute N_{EME} and $\sigma(N_{EME})$based on this reasonment
- We can therefore use Eq. 5 to compute $\frac{N_{EME}}{2}$ and multiply the result by 2, obtaining

$$N_{EME} = 2.00 \pm 0.03\ EMT$$

- This estimate corresponds to a precision of only 1.5%!!!
- The dramatic improvement in the precision of the measurement, from 2.4% to 1.5%, is a direct consequence of having used additional information under the form of a relationship (constraint) between the two available measurements.
- A good physicist exploits as many constraints as possible in order to improve the precision of a measurement
  - Sometimes the contraints are arbitrary or correspond to special cases
  - Is is very important to explicitly mention any constraint used to derive a measurement, when quoting the result.

What about asymmetric uncertainties?

- Now suppose my receivers operate by taking data and performing a maximum likelihood fit to estimate N_{EM} and N_{ME}
- Can I combine these two measurements with the two methods seen above?
  - $N_{EM} = 0.99 \pm 0.03$
  - $N_{ME} = 1.10^{+0.05}_{-0.01}$
- For example, $N_{EMT} = 2.09^{+0.06}_{-0.03}$

What about asymmetric uncertainties?

UCLouvain
Institut de recherche
en mathématique et physique

- Now suppose my receivers operate by taking data and performing a maximum likelihood fit to estimate N_{EM} and N_{ME}
- Can I combine these two measurements with the two methods seen above?
  - $N_{EM} = 0.99 \pm 0.03$
  - $N_{ME} = 1.10^{+0.05}_{-0.01}$
- For example, $N_{EMT} = 2.09^{+0.06}_{-0.03}$
- No!
- Why?

What about asymmetric uncertainties?

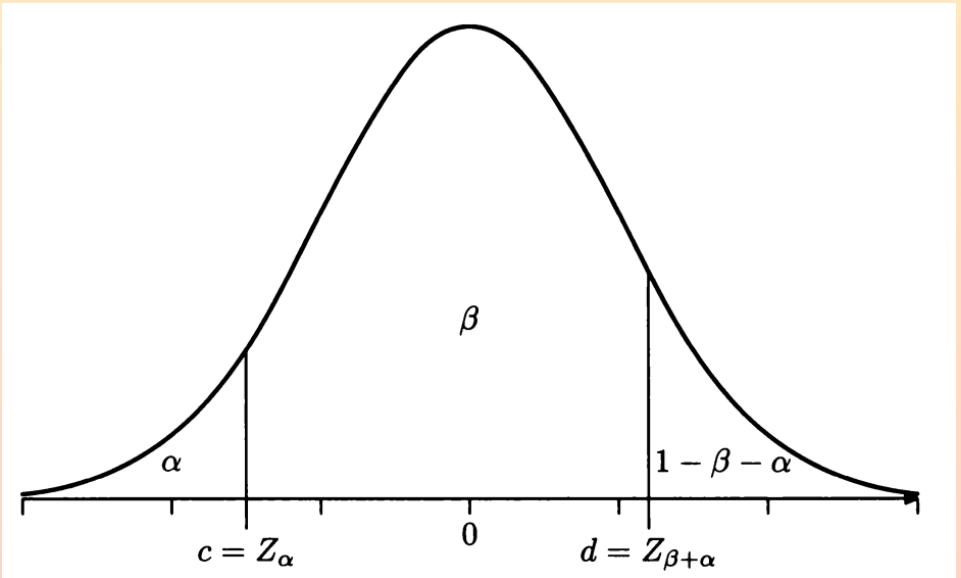
- Now suppose my receivers operate by taking data and performing a maximum likelihood fit to estimate N_{EM} and N_{ME}
- Can I combine these two measurements with the two methods seen above?
  - $N_{EM} = 0.99 \pm 0.03$
  - $N_{ME} = 1.10^{+0.05}_{-0.01}$
- For example, $N_{EMT} = 2.09^{+0.06}_{-0.03}$
- No!
- Why?
- The naïve quadrature of the two uncertainties is wrong!
  - The naïve combination is an expression of the Central Limit Theorem
  - The resulting combination is expected to be more symmetric than the measurements it originates from
  - Symmetric uncertainties usually assume a Gaussian approximation of the likelihood
  - Asymmetric uncertainties? One would need a study of the non-linearity (large biases might be introduced if ignoring this)
- Intrinsic difference between averaging and most probable value
  - Averaging results in average value and variance that propagate linearly
  - Taking the mode (essentially what MLE does) does not add up linearly!
- With asymmetric uncertainties from MLE fits, always combine the likelihoods (better in an individual simultaneous fit)

Confidence Intervals in nontrivial cases

Confidence intervals!

- Confidence interval for θ with probability content β
  - The range $\theta_a < \theta < \theta_b$ containing the true value θ_0 with probability β
  - The physicists sometimes improperly say the uncertainty on the parameter θ
- Given a p.d.f., the probability content is $\beta = P(a \leq X \leq b) = \int_a^b f(X|\theta)dX$
- If θ is unknown (as is usually the case), use auxiliary variable $Z = Z(X, \theta)$ with p.d.f. $g(Z)$ independent of θ
- If Z can be found, then the problem is to estimate interval $P(\theta_a \leq \theta_0 \leq \theta_b) = \beta$
  - Confidence interval
  - A method yielding an interval satisfying this property has coverage

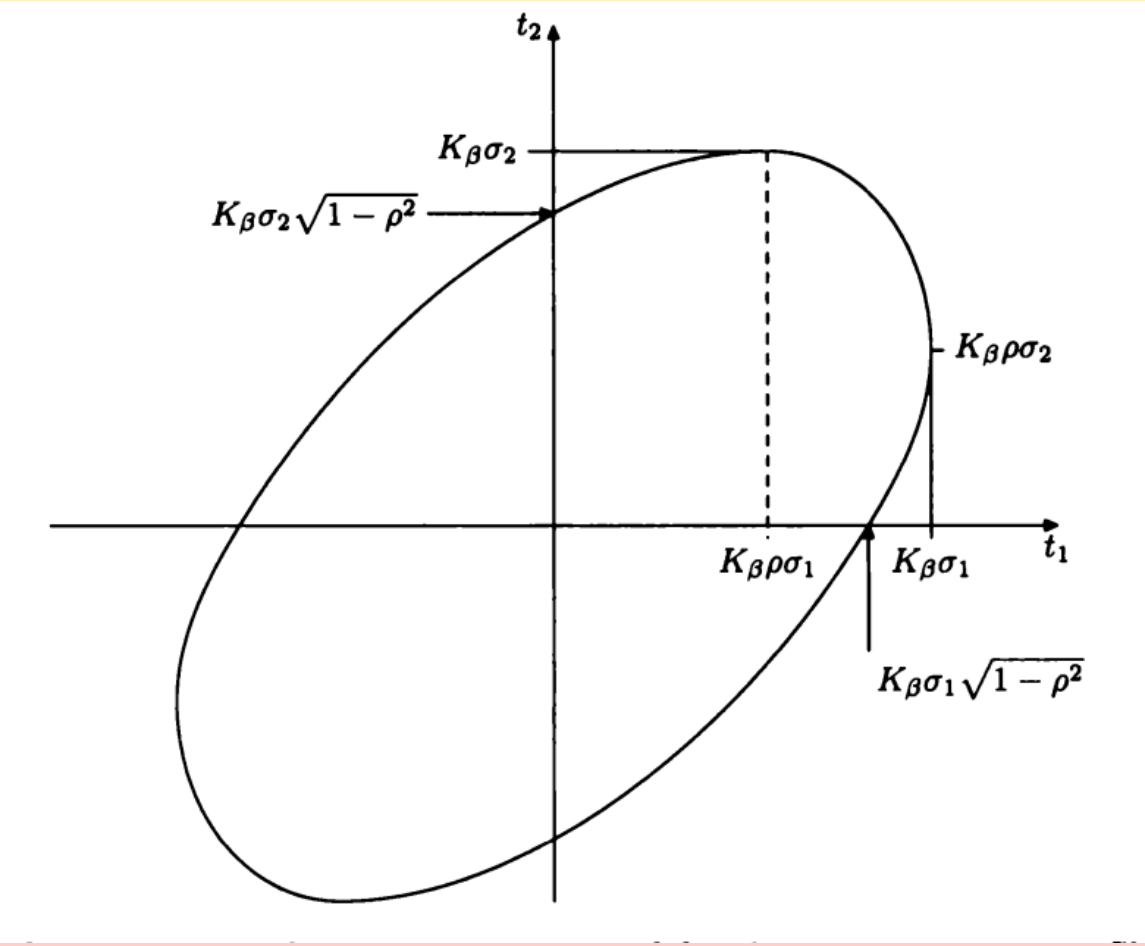
- Example: if $f(X|\theta) = N(\mu, \sigma^2)$ with unknown μ, σ, choose $Z = \frac{X-\mu}{\sigma}$
- Find $[c, d]$ in $\beta = P(c \leq Z \leq d) = \Phi(d) - \Phi(c)$ by finding $[Z_\alpha, Z_{\alpha+\beta}]$
- Infinite interval choices: here central interval $\alpha = \frac{1-\beta}{2}$



Plot from James, 2nd ed.

Confidence intervals in many dimensions

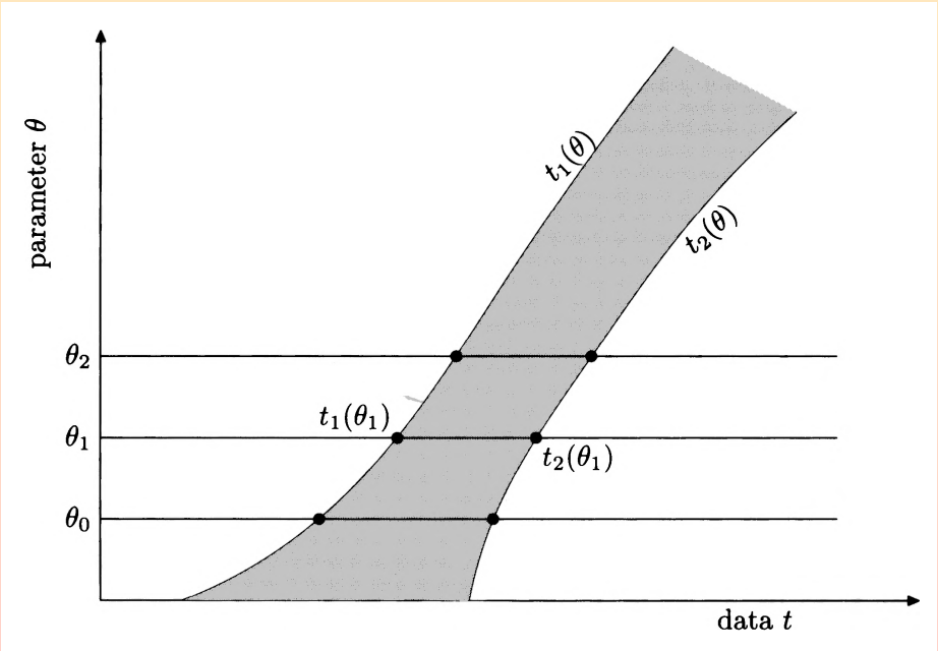
- Generalization to multidimensional $\boldsymbol{\theta}$ is immediate
- Probability statement concerns the whole $\boldsymbol{\theta}$, not the individual θ_i
- Shape of the ellipsoid governed by the correlation coefficient (or the mutual information) between the parameters
- Arbitrariety in the choice of the interval is still present

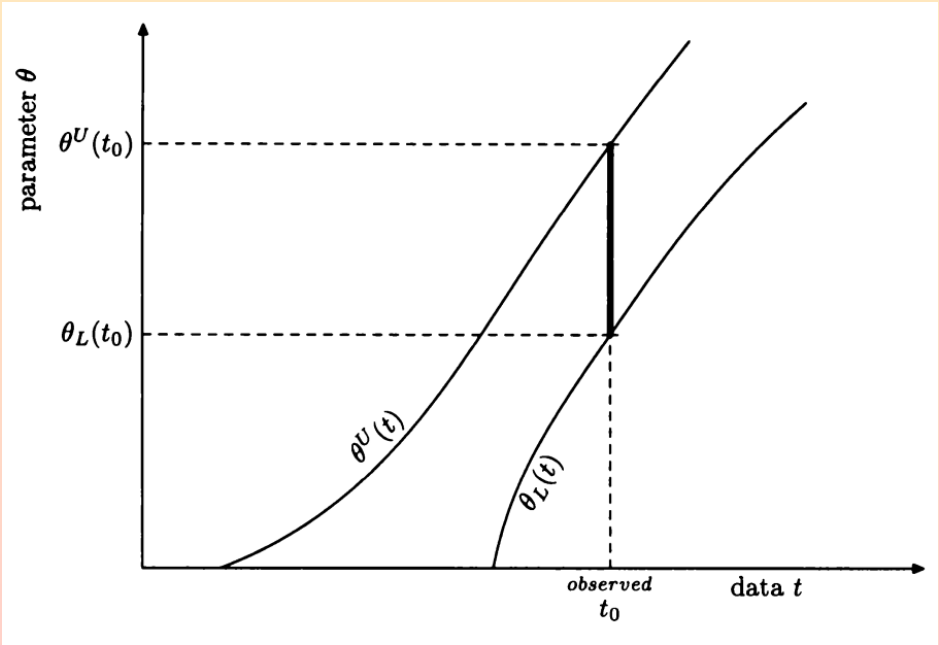


Plot from James, 2nd ed.

Confidence belts: the Neyman construction

- Unique solutions to finding confidence intervals are infinite
  - Central intervals, lower limits, upper limits, etc
- Let's suppose we have chosen a way
- Build horizontally: for each (hypothetical) value of θ, determine $t_1(\theta)$, $t_2(\theta)$ such that $\int_t 1^t 2P(t|\theta)dt = \beta$
- Read vertically: from the observed value t_0, determine $[\theta_L, \theta^U]$ by intersection
  - The resulting interval might be disconnected in severely non-linear cases
- Probability content statements to be seen in a frequentist way
  - Repeating many times the experiment, the fraction of $[\theta_L, \theta^U]$ containing θ_0 is β





Plot from James, 2nd ed.

Coverage

- Coverage probability of a method for calculating a confidence interval $[\theta_1, \theta_2]$: $P(\theta_1 \leq \theta_{true} \leq \theta_2)$
  - Fraction of times, over a set of (usually hypothetical) measurements, that the resulting interval covers the true value of the parameter
  - Can sample with toys to study coverage
- Coverage is not a property of a specific confidence interval!
- The nominal coverage is the value of confidence level you have built your method around (often 0.95)
- When actually derive a set of intervals, the fraction of them that contain θ_{true} ideally would be equal to the nominal coverage
  - You can build toy experiments in each of whose you sample N times for a known value of θ_{true}
  - You calculate the interval for each toy experiment
  - You count how many times the interval contains the true value
- Nominal coverage (CL) and the actual coverage (Co) observed with toys should agree
  - If all the assumptions you used in computing the intervals are valid
  - If they don't agree, it might be that $Co < CL$ (undercoverage) or $Co > CL$ (overcoverage)
  - It's OK to strive to be conservative, but one might be unnecessarily lowering the precision of the measurement
  - When $Co! = CL$ you usually want at least a convergence to equality in some limit

Coverage: the binomial case

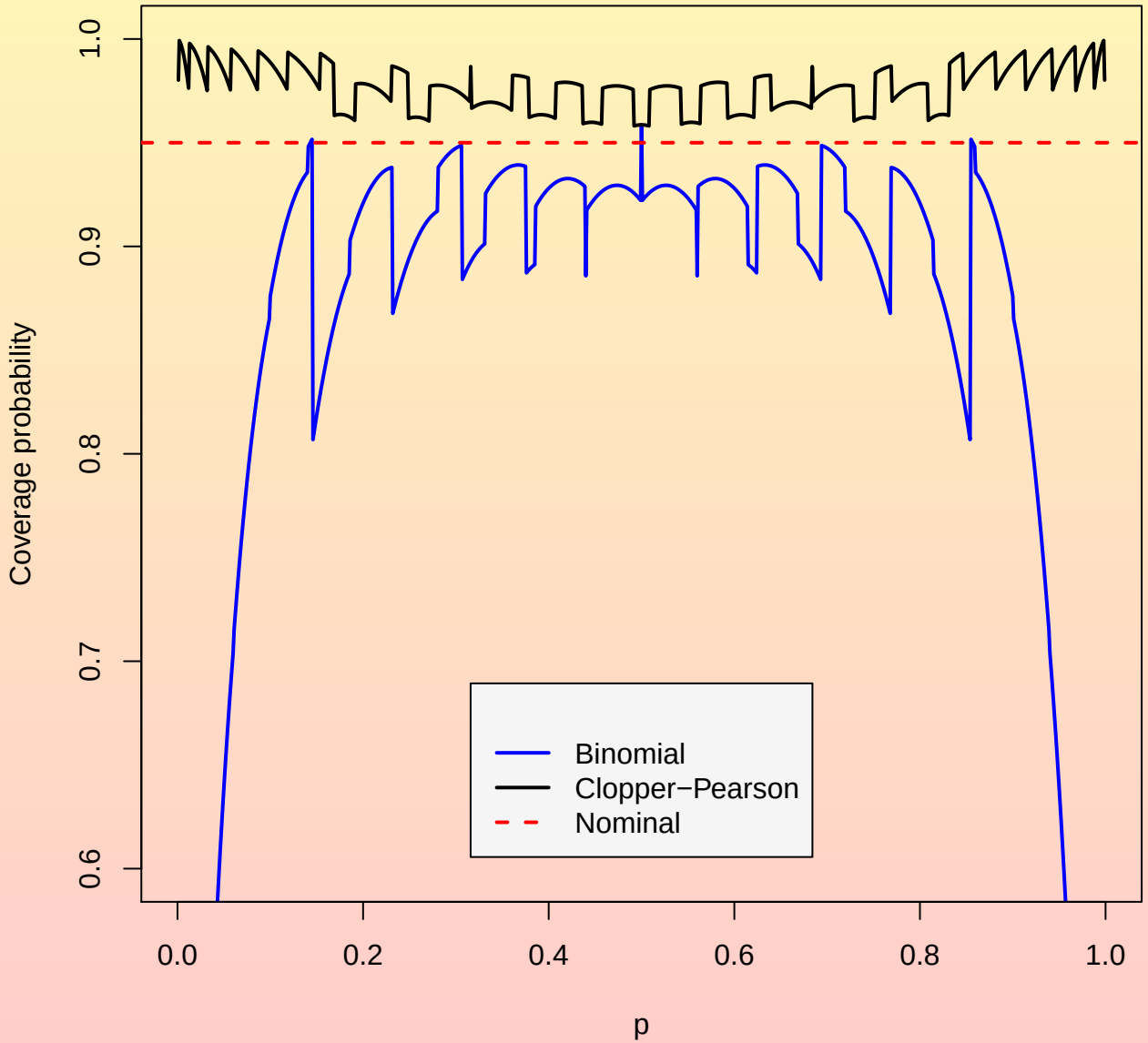
- For discrete distributions, the discreteness induces steps in the probability content of the interval
  - Continuous case: $P(a \leq X \leq b) = \int_a^b f(X|\theta)dX = \beta$
  - Discrete case: $P(a \leq X \leq b) = \sum_a^b f(X|\theta)dX \leq \beta$
- Binomial: find interval (r_{low}, r_{high}) such that $\sum_{r=r_{low}}^{r=r_{high}} \binom{r}{N} p^r (1-p)^{N-r} \leq 1-\alpha$
  - Also, $\binom{r}{N}$ computationally taxing for large r and N
  - Approximations are found in order to deal with the problem
- Gaussian approximation: $p \pm Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{N}}$
- Clopper Pearson: invert two single-tailed binomial tests, designed to overcover
  $\sum_{r=0}^{N} \binom{r}{N} p^n (1-p_{low})^{N-n} \leq \alpha/2$
  $\sum_{r=0}^{N} \binom{r}{N} p^r (1-p_{high})^{N-r} \leq \alpha/2$
  - Single-tailed $\rightarrow$ use $\alpha/2$ instead of α

Coverage: the binomial case

- Gaussian approximation: $p \pm Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{N}}$
- Clopper Pearson: invert two single-tailed binomial tests, designed to overcover
  $\sum_{r=0}^{N} \binom{r}{N} p^n (1-p_{low})^{N-n} \leq \alpha/2$
  $\sum_{r=0}^{N} \binom{r}{N} p^r (1-p_{high})^{N-r} \leq \alpha/2$
  - Single-tailed $\rightarrow$ use $\alpha/2$ instead of α
- Study coverage of intervals from a gaussian approximation and from the Clopper-Pearson method
  wget https://raw.githubusercontent.com/vischia/statex/master/coverageTest.R
  wget https://raw.githubusercontent.com/vischia/statex/master/coverageTest.py
  wget https://raw.githubusercontent.com/vischia/statex/master/coverageTest.ipynb
  - For a given N, calculate intervals for various numbers of successes r, and plot the intervals of p as a function of r
  - Do a coverage test by using the procedure outlined in the previous slide
  - Draw the coverage probability as a function of p
  - Find the issue with the Clopper Pearson implementation in python
  - What happens for different sample sizes N?

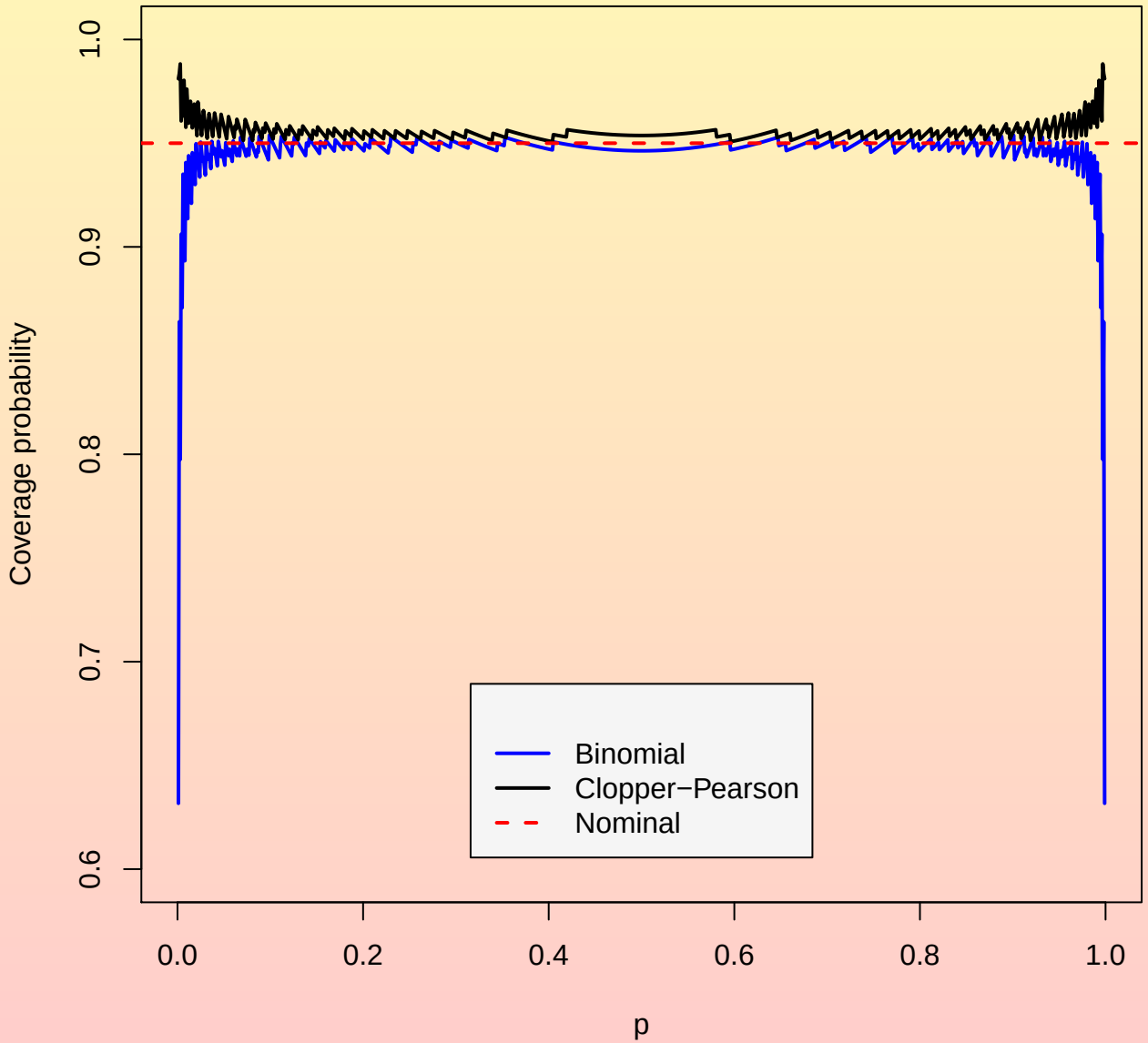
Coverage, $N = 20$

- Gaussian approximation bad for small sample sizes



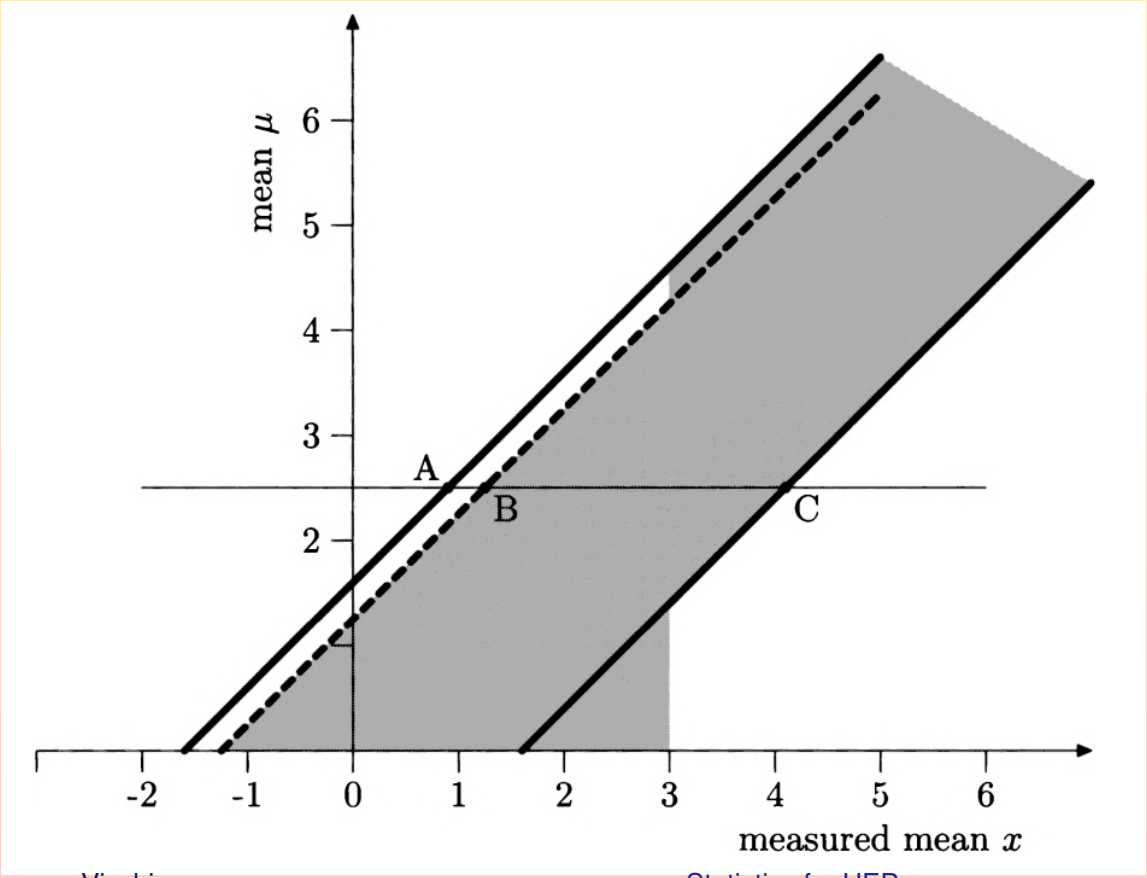
Coverage, $N = 1000$

- Gaussian approximation bad near $p = 0$ and $p = 1$ even for large sample sizes



Upper limits for non-negative parameters

- Gaussian measurement (variance 1) of a non-negative parameter $\mu \sim 0$ (physical bound)
- Individual prescriptions are self-consistent
  - 90% central limit (solid lines)
  - 90% upper limit (single dashed line)
- Other choices are problematic (flip-flopping): never choose after seeing the data!
  - "quote upper limit if x_{obs} is less than 3σ from zero, and central limit above" (shaded)
  - Coverage not guaranteed anymore (see e.g. $\mu = 2.5$)
- Unphysical values and empty intervals: choose 90% central interval, measure $x_{obs} = -2.0$
  - Don't extrapolate to an unphysical interval for the true value of μ!
  - The interval is simply empty, i.e. does not contain any allowed value of μ
  - The method still has coverage (90% of other hypothetical intervals would cover the true value)

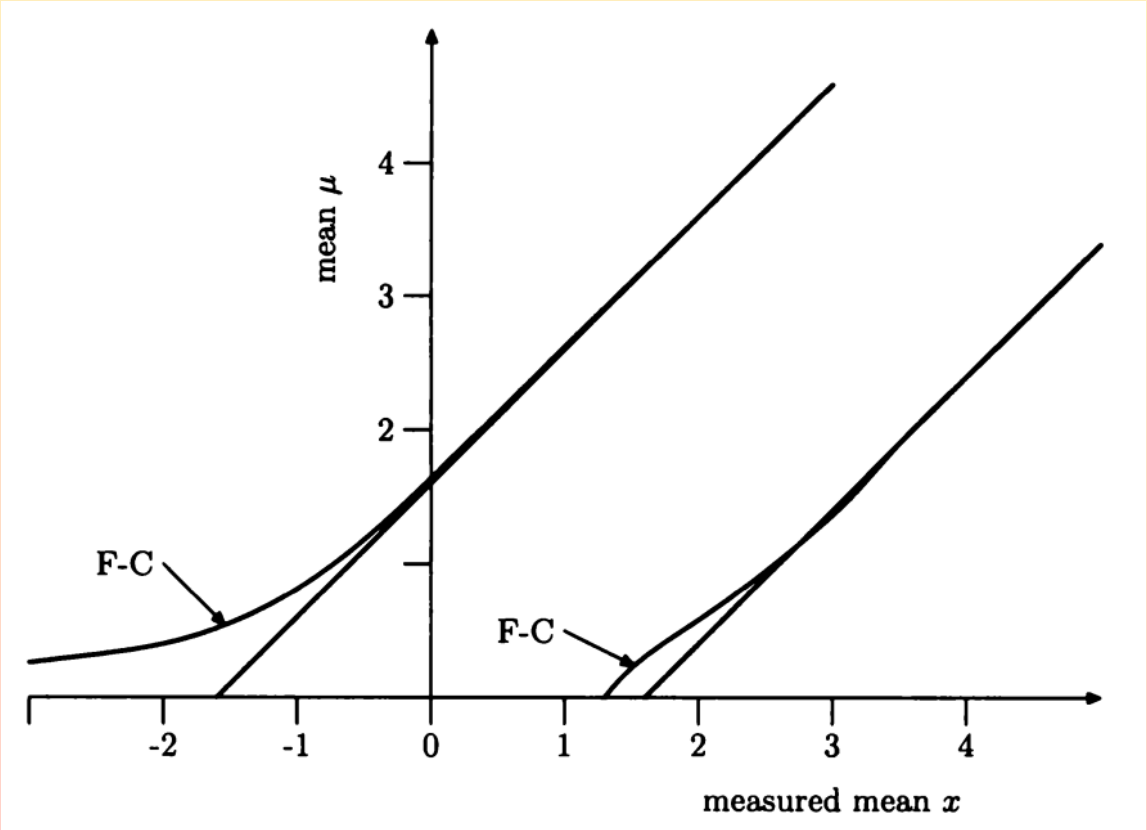


Plot from James, 2nd ed.

Unphysical values: Feldman-Cousins

- The Neyman construction results in guaranteed coverage, but choice still free on how to fill probability content
  - Different ordering principles are possible (e.g. central/upper/lower limits)
- Unified approach for determining interval for $\mu = \mu_0$: the likelihood ratio ordering principle
  - Include in order by largest $\ell(x) = \frac{P(x|\mu_0)}{P(x|\hat{\mu})}$
  - $\hat{\mu}$ value of μ which maximizes $P(x|\mu)$ within the physical region
  - $\hat{\mu}$ remains equal to zero for $\mu < 1.65$, yielding deviation w.r.t. central intervals

- Minimizes Type II error (likelihood ratio for simple test is the most powerful test)
- Solves the problem of empty intervals
- Avoids flip-flopping in choosing an ordering prescription



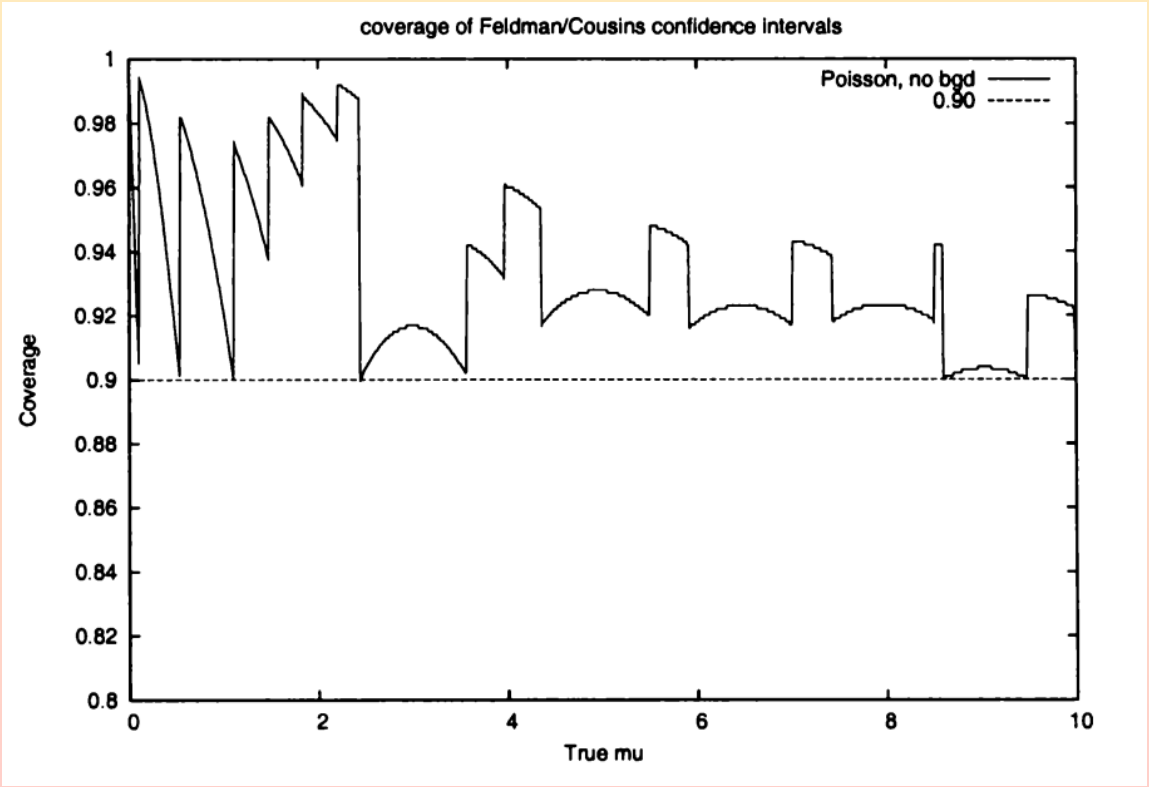
Plot from James, 2nd ed.

Feldman-Cousins in HEP

- The most typical HEP application of F-C is confidence belts for the mean of a Poisson distribution
- Discreteness of the problem affects coverage
- When performing the Neyman construction, will add discrete elements of probability
- The exact probability content won't be achieved, must accept overcoverage

$$\int_{x_1}^{x_2} f(x|\theta)dx = \beta \qquad \rightarrow \qquad \sum_{i=L}^{U} P(x_i|\theta) \geq \beta$$

- Overcoverage larger for small values of μ (but less than other methods)



Plot from James, 2nd ed.

Bayesian intervals

- Often numerically identical to frequentist confidence intervals
  - Particularly in the large sample limit
- Interpretation is different: credible intervals
- Posterior density summarizes the complete knowledge about θ

$$\pi(\theta|\boldsymbol{X}) = \frac{\prod_{i=1}^{N} f(X_i,\theta)\pi(\theta)}{\int \prod_{i=1}^{N} f(X_i,\theta)\pi(\theta)d\theta}$$

- An interval $[\theta_L, \theta^U]$ with content β defined by $\int_{\theta_L}^{\theta^U} \pi(\theta|\boldsymbol{X})d\theta = \beta$
- Bayesian statement! $P(\theta_L < \theta < \theta^U) = \beta$
  - Again, non unique
- Issues with empty intervals don't arise, though, because the prior takes care of defining the physical region in a natural way!
  - But this implies that central intervals cannot be seamlessly converted into upper limits
  - Need the notion of shortest interval
  - Issue of the metric (present in frequentist statistic) solved because here the preferred metric is defined by the prior

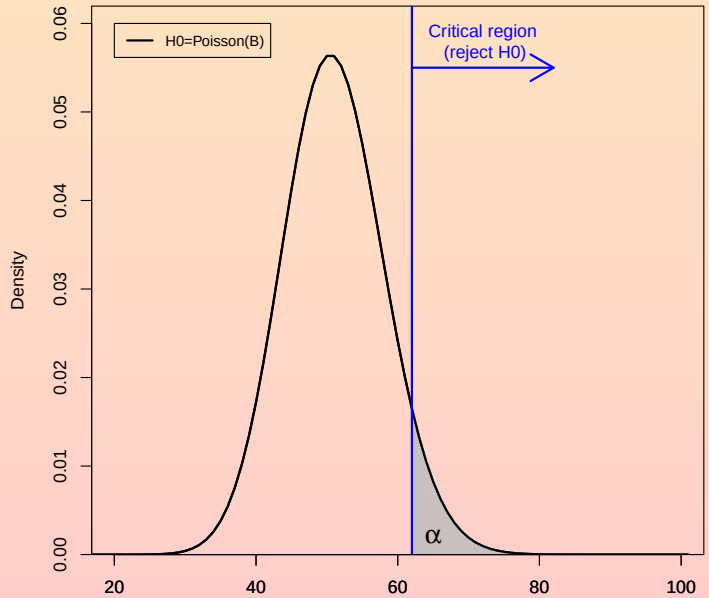
What is an hypothesis...

- Is our hypothesis compatible with the experimental data? By how much?
- Hypothesis: a complete rule that defines probabilities for data.
  - An hypothesis is simple if it is completely specified (or if each of its parameters is fixed to a single value)
  - An hypothesis is complex if it consists in fact in a family of hypotheses parameterized by one or more parameters
- "Classical" hypothesis testing is based on frequentist statistics
  - An hypothesis—as we do for a parameter $\vec{\theta}_{true}$—is either true or false. We might improperly say that $P(H)$ can only be either 0 or 1
  - The concept of probability is defined only for a set of data $\vec{x}$
- We take into account probabilities for data, $P(\vec{x}|H)$
  - For a fixed hypotesis, often we write $P(\vec{x}; H)$, skipping over the fact that it is a conditional probability
  - The size of the vector $\vec{x}$ can be large or just 1, and the data can be either continuos or discrete.

- The hypothesis can depend on a parameter
  - Technically, it consists in a family of hypotheses scanned by the parameter
  - We use the parameter as a proxy for the hypothesis, $P(\vec{x};\theta) := P(\vec{x};H(\theta))$.
- We are working in frequentist statistics, so there is no $P(H)$ enabling conversion from $P(\vec{x}|\theta)$ to $P(\theta|\vec{x})$.
- Statistical test
  - A statistical test is a proposition concerning the compatibility of H with the available data.
  - A binary test has only two possible outcomes: either accept or reject the hypothesis

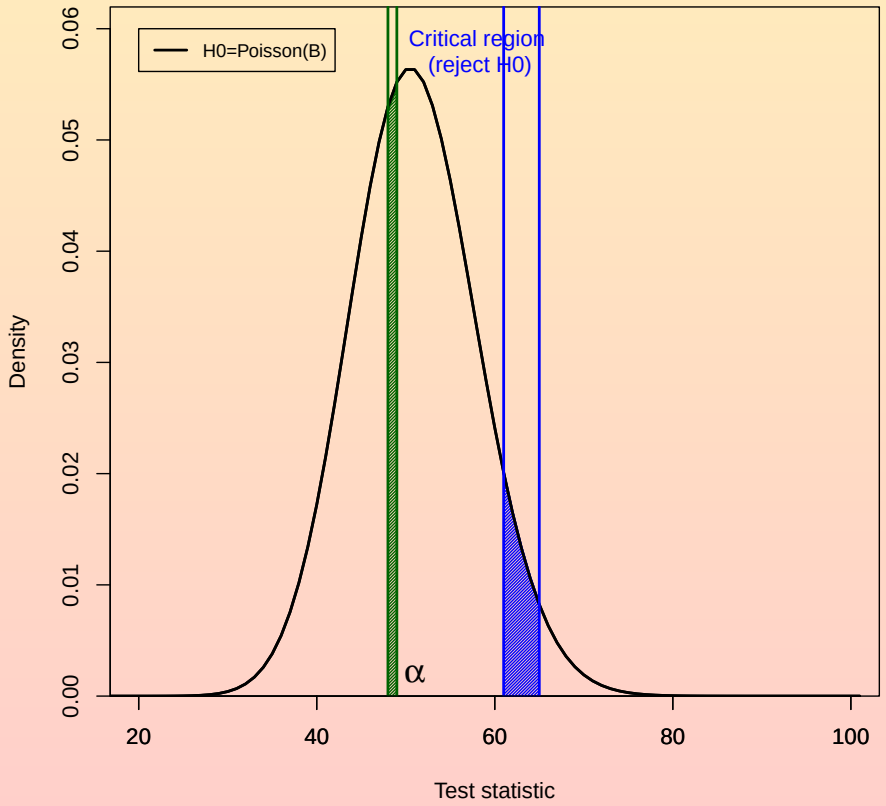
Testing an hypothesis H_0...

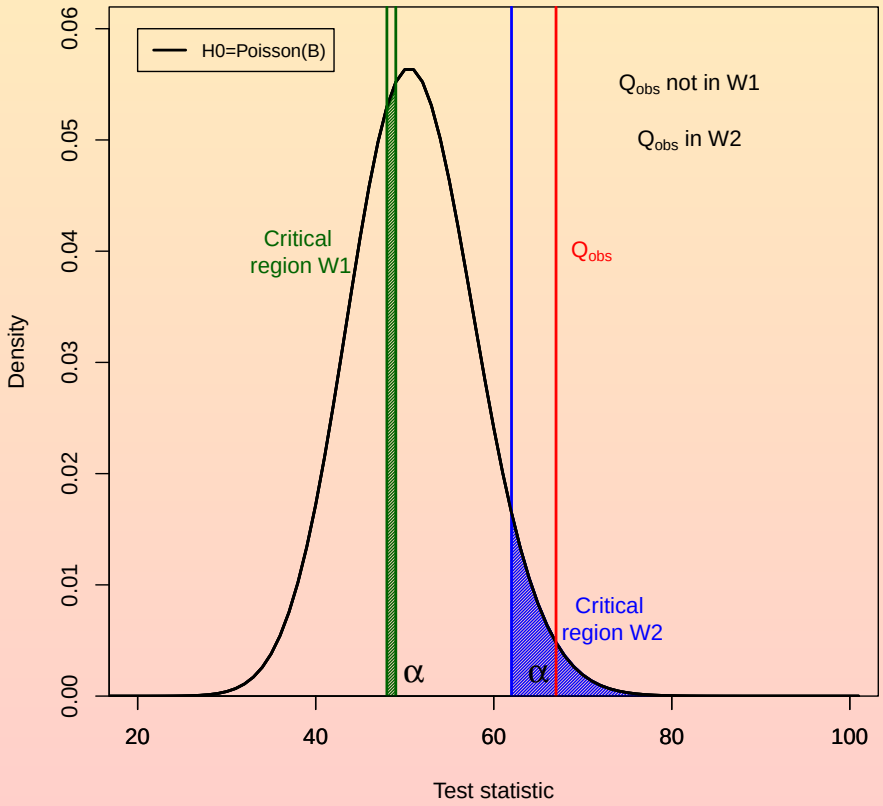
- H_0 is normally the hypothesis that we assume true in absence of further evidence
- Let $\mathbf{X}$ be a function of the observations (called "*test statistic*")
- Let W be the space of all possible values of $\mathbf{X}$, and divide it into
  - A critical region w: observations X falling into w are regarded as suggesting that H_0 is NOT true
  - A region of acceptance $W - w$
- The size of the critical region is adjusted to obtain a desired *level of significance* α
  - Also called *size of the test*
  - $P(X \in w | H_0) = \alpha$
  - α is the (hopefully small) probability of rejecting H_0 when H_0 is actually true
- Once $\mathcal{W}$ is defined, given an observed value $\vec{x}_{obs}$ in the space of data, we define the test by saying that we reject the hypothesis H_0 if $\vec{x}_{obs} \in W$.
- If $\vec{x}_{obs}$ is inside the critical region, then H_0 is rejected; in the other case, H_0 is accepted
  - In this context, accepting H_0 does not mean demonstrating its truth, but simply not rejecting it
- Choosing a small α is equivalent to giving a priori preference to H_0!!!



...while introducing some spice in it

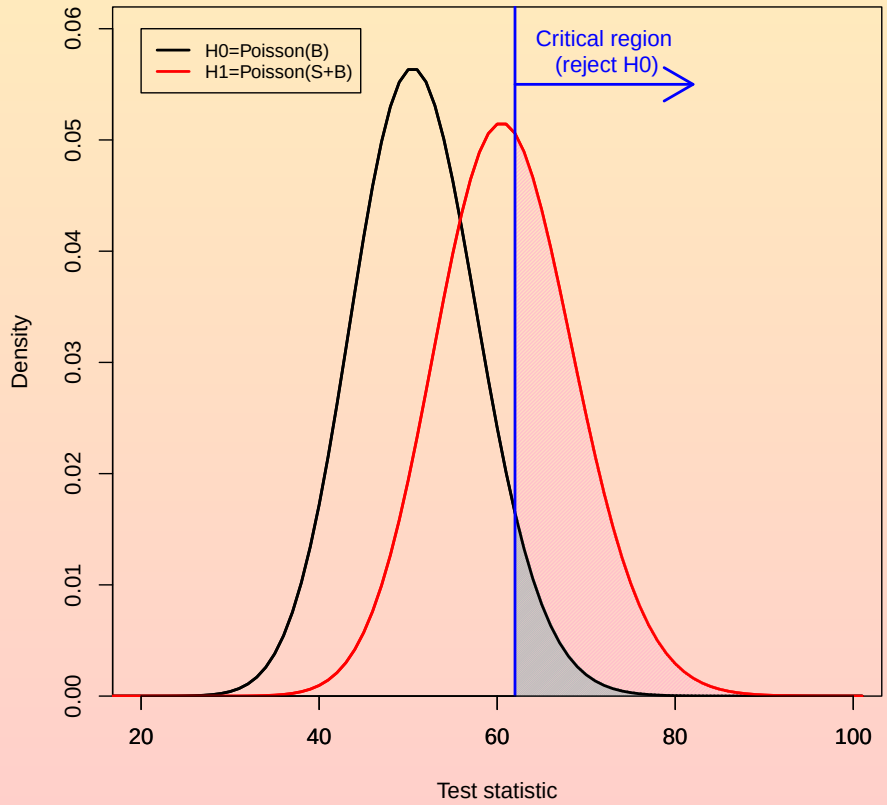
- The definition of $\mathcal{W}$ depends only on its area α, without any other condition
  - Any other area of area α can be defined as critical region, independently on how it is placed with respect to $\vec{x}_{obs}$
  - In particular, for an infinite number of choices of $\mathcal{W}$, the point $\vec{x}_{obs}$—which beforehand was situated outside of $\mathcal{W}$—is now included inside the critical region
  - In this condition, the result of the test switches from accept H_0 to reject H_0
- To remove or at least reduce this arbitrariness in the choice of $\mathcal{W}$, we introduce the alternative hypothesis, H_1





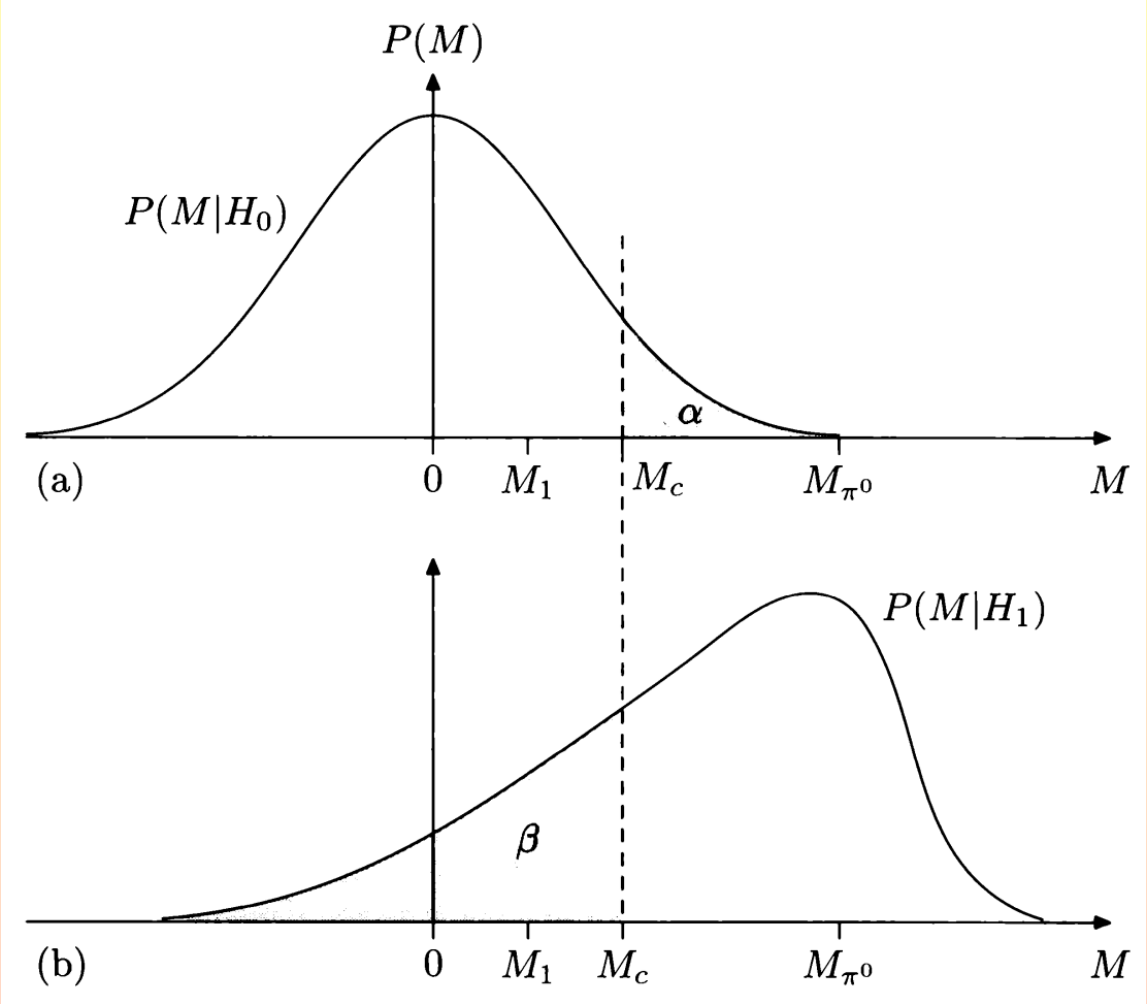
Choose reasonable regions

- Choose a critical region so that $P(\vec{x} \in \mathcal{W}|H_0)$ is α under H_0, and as large as possible under H_1
- Choice of regions is somehow arbitrary, and many choices are not more justified than others
- In Physics, after ruling out an hypothesis we aim at substituting it with one which explains better the data
  - Often H_1 becomes the new H_0, e.g. from (H_0:noHiggs, H_1 =Higgs) to (H_1:Higgs , H_1:otherNewPhysics)
  - We can use our expectations about reasonable alternative hypotheses to design our test to exlude H_0



A small example

- H_0: $pp \to pp$ elastic scattering
- H_1: $pp \to pp\pi^0$
- Compute the missing mass M (as total rest energy of unseen particles)
- Under H_0, $M = 0$
- Under H_1, $M = 135\ MeV$



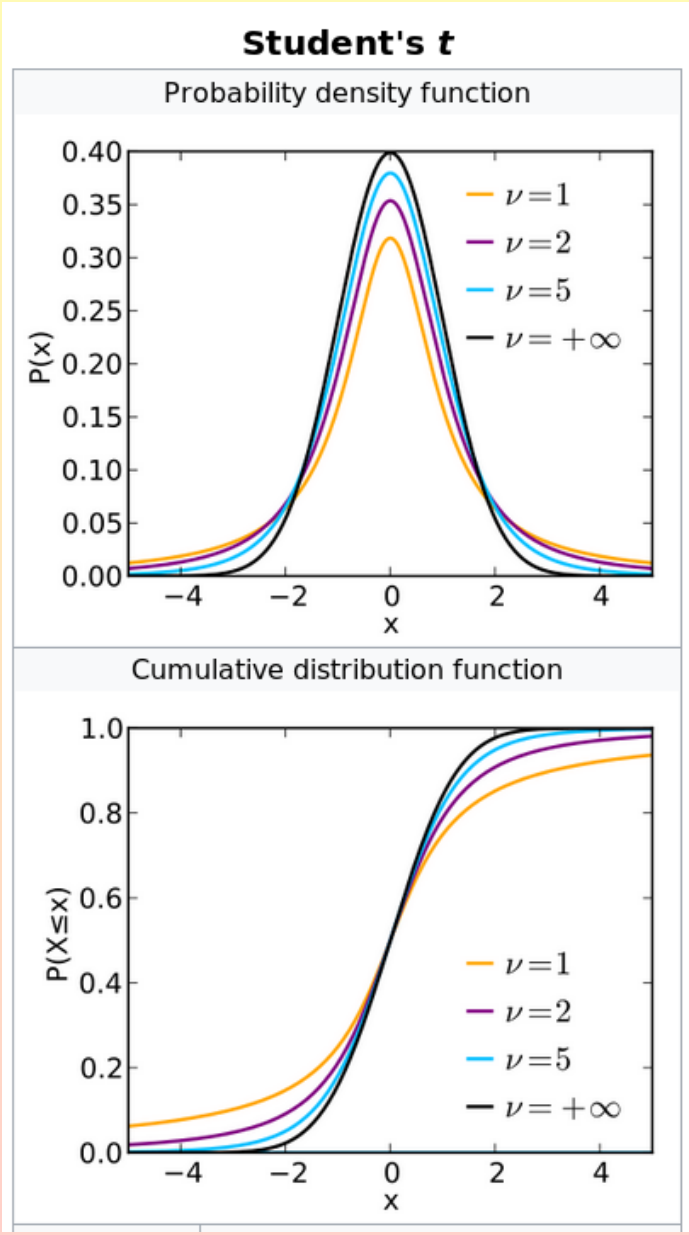
	Choose H_0	Choose H_1
H_0 is true	$1-\alpha$	α (Type I error)
H_1 is true	β (Type II error)	$1-\beta$ (power)

Plot from James, 2nd ed.

A longer example

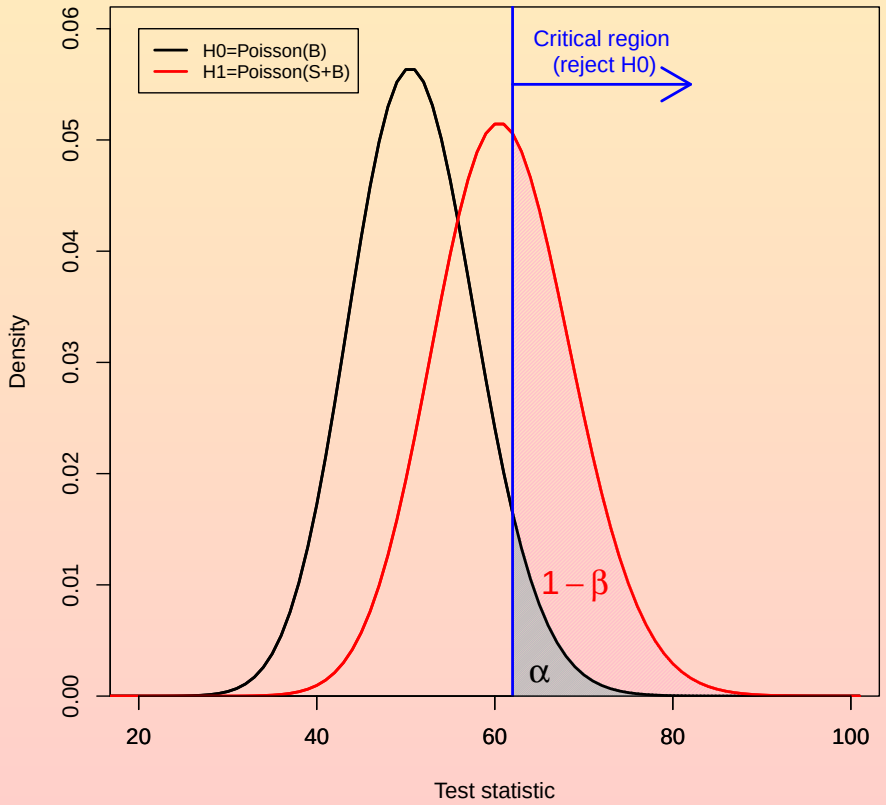
- Student's t distribution
- Test the mean!
- wget hyptest.ipynb

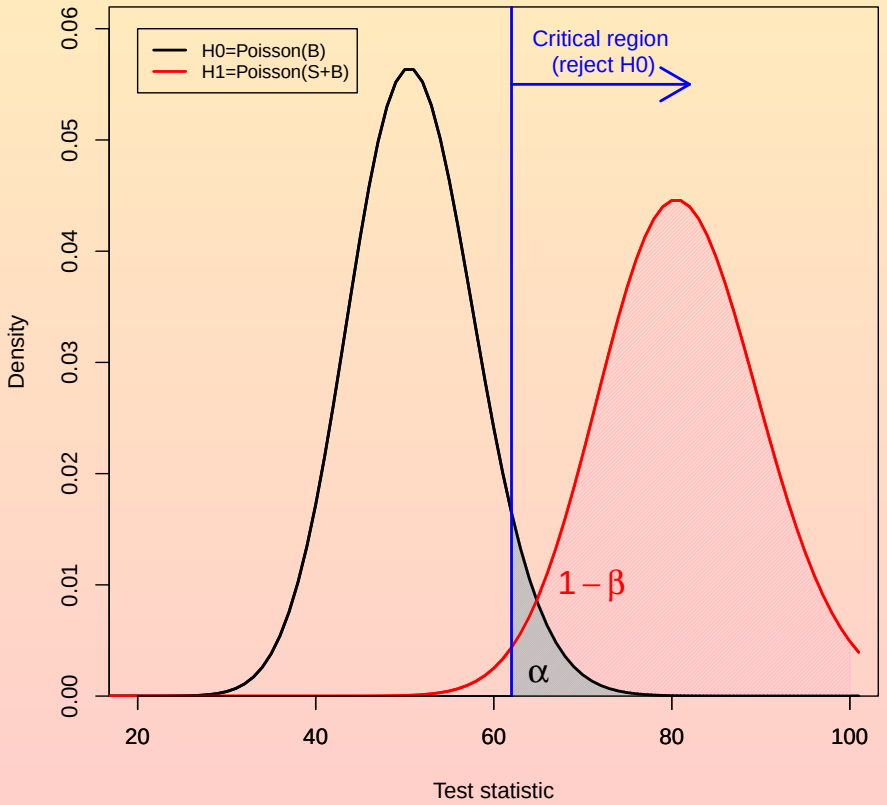
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$



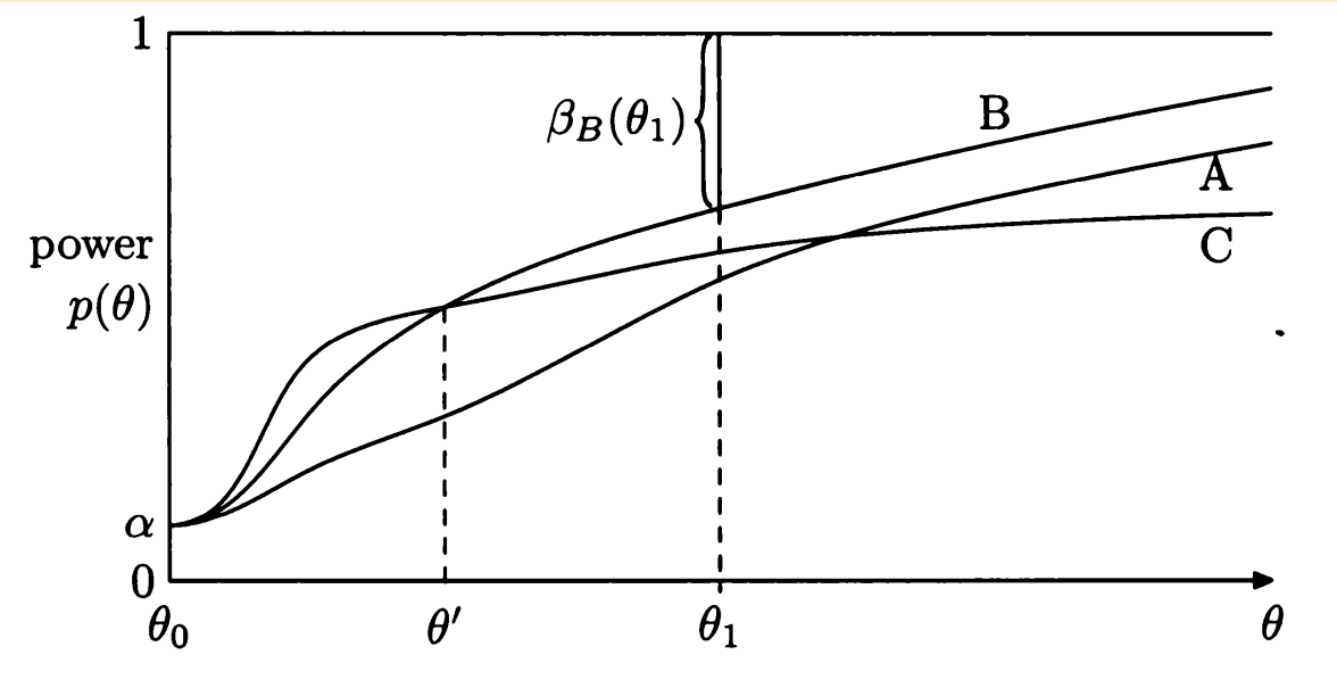
Basic hypothesis testing – 4

- The usefulness of the test depends on how well it discriminates against the alternative hypothesis
- The measure of usefulness is the *power of the test*
  - $P(X \in w|H_1) = 1 - \beta$
  - Power $(1 - \beta)$ is the probabiliity of X falling into the critical region if H_1 is true
  - $P(X \in W - w|H_1) = \beta$
  - β is the probability that X will fall into the acceptance region if H_1 is true
- NOTE: some authors use β where we use $1 - \beta$. Pay attention, and live with it.





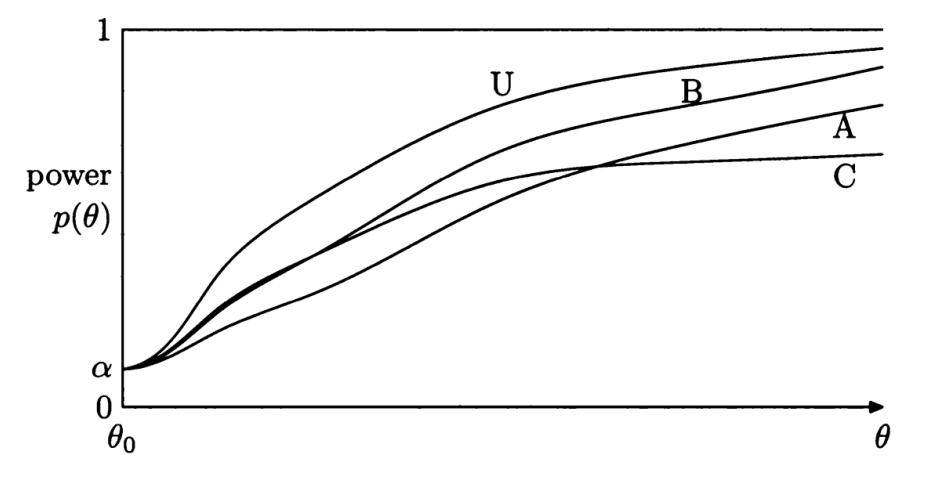
- For parametric (families of) hypotheses, the power depends on the parameter
  - $H_0 : \theta = \theta_0$
  - $H_1 : \theta = \theta_1$
  - Power: $p(\theta_1) = 1 - \beta$
- Generalize for all possible alternative hypotheses: $p(\theta) = 1 - \beta(\theta)$
  - For the null, $p(\theta_0) = 1 - \beta(\theta_0) = \alpha$

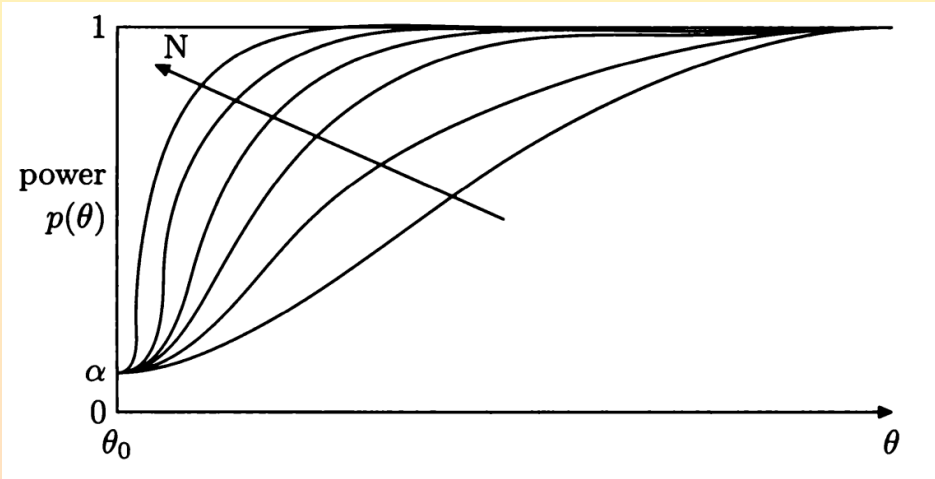


Plot from James, 2nd ed.

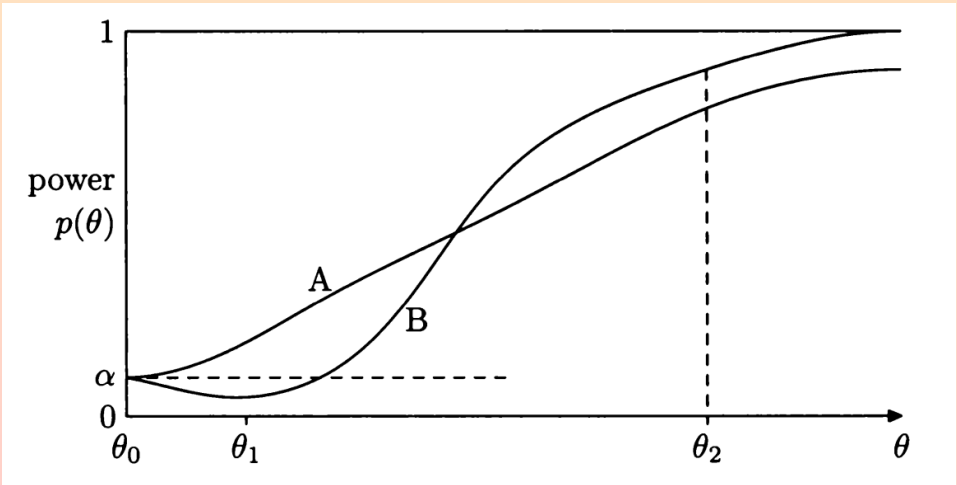
Properties of tests

- More powerful test: a test which at least as powerful as any other test for a given θ
- Uniformly more powerful test: a test which is the more powerful test for any value of θ
  - A less powerful test might be preferrable if more robust than the UMP[1]
- If we increase the number of observations, it makes sense to require consistency
  - The more observations we add, the more the test distinguishes between the two hypotheses
  - Power function tends to a step function for $N \to \infty$





- Biased test: $argmin(p(\theta)) \neq \theta_0$
- More likely to accept H_0 when it is false than when it is true
- Big no-no for θ_0 vs θ_1]
- Still useful (larger power) for θ_0 vs θ_2



Plot from James, 2nd ed.

[1] Robust: a test with low sensitivity to unimportant changes of the null hypothesis

Play with Type I (α) and Type II (β) errors freely

- Comparing only based on the power curve is asymmetric w.r.t. α
- For each value of $\alpha = p(\theta_0)$, compute $\beta = p(\theta_1)$, and draw the curve
  - Unbiased tests fall under the line $1 - \beta = \alpha$
  - Curves closer to the axes are better tests
- Ultimately, though, choose based on the cost function of a wrong decision
  - Bayesian decision theory

$$h(\mathbf{X}|\theta, \phi, \psi) = \theta f(\mathbf{X}|\phi) + (1 - \theta) g(\mathbf{X}, \psi)$$

d_0 : No choice is possible; results are ambiguous
d_1, ϕ^* : Family was $f(\mathbf{X}|\phi)$, with $\phi = \phi^*$
d_2, ψ^* : Family was $g(\mathbf{X}|\psi)$, with $\psi = \psi^*$.

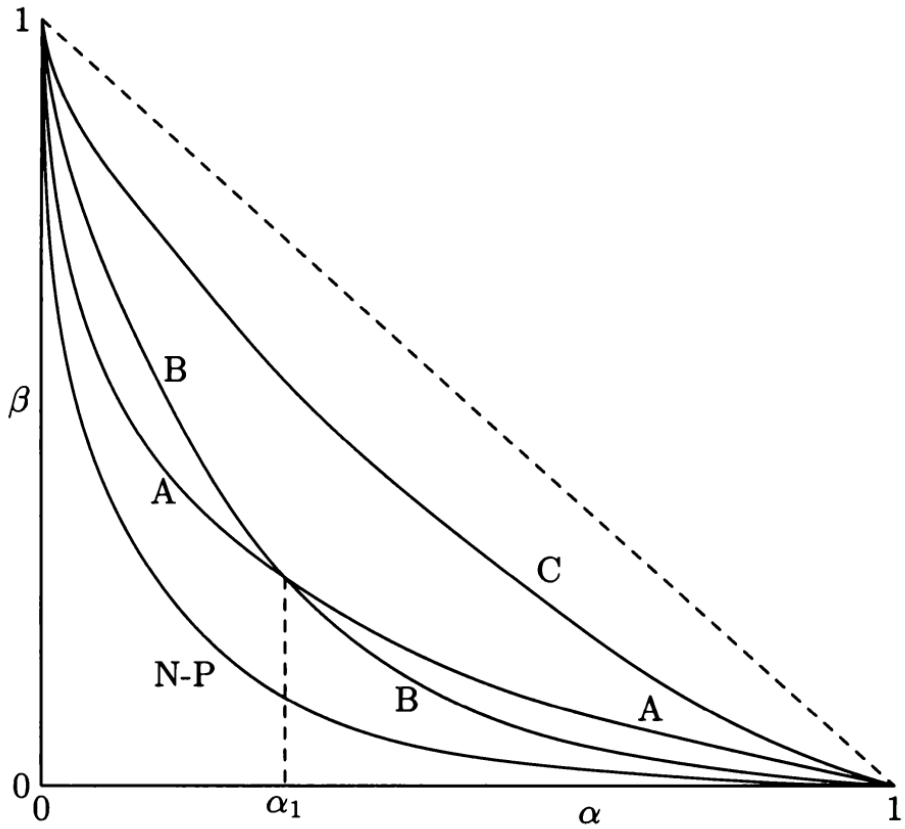


Table 10.4. A cost function.

Decisions	True state of nature	
	$\theta = \theta_1 = 1, \phi$	$\theta = \theta_2 = 0, \psi$
d_0	β_1	β_2
d_1, ϕ^*	$\alpha_1(\phi^* - \phi)^2$	γ_1
d_2, ψ^*	γ_2	$\alpha_2(\psi^* - \psi)^2$

Plot from James, 2nd ed.

Find the most powerful test

- Testing simple hypotheses H_0 vs H_1, find the best critical region
- Maximize power curve $1-\beta = \int_{w_\alpha} f(\mathbf{X}|\theta_1)d\mathbf{X}$, given $\alpha = \int_{w_\alpha} f(\mathbf{X}|\theta_0)d\mathbf{X}$
- The best critical region w_α consists in the region satisfying the likelihood ratio equation

$$\ell(\mathbf{X}, \theta_0, \theta_1) := \frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} \geq c_\alpha$$

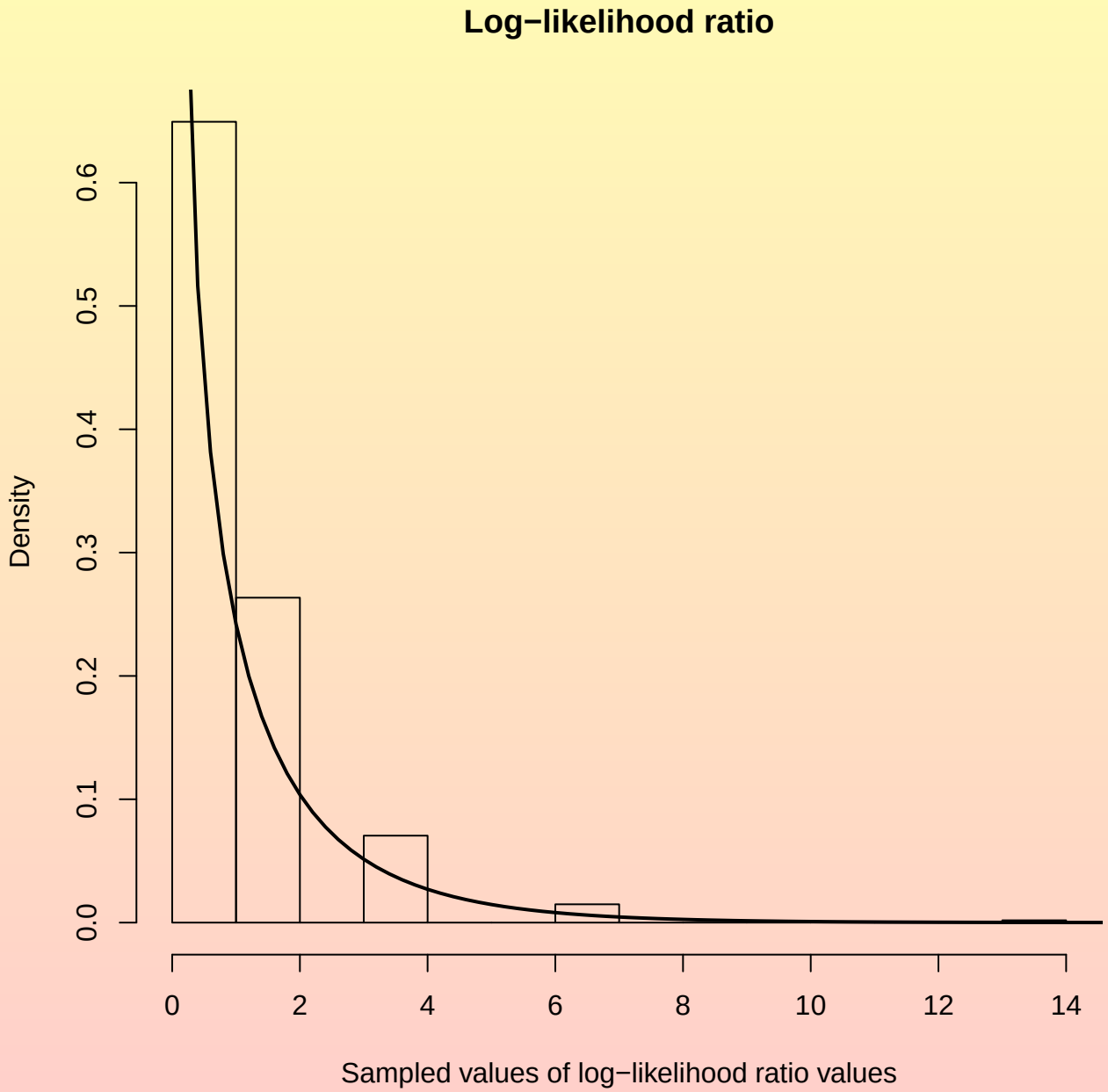
- The criterion, called Neyman-Pearson test is therefore
  - If $\ell(\mathbf{X}, \theta_0, \theta_1) > c_\alpha$ then choose H_1
  - If $\ell(\mathbf{X}, \theta_0, \theta_1) \leq c_\alpha$ then choose H_0
- The likelihood ratio must be calculable for any $\mathbf{X}$
  - The hypotheses must therefore be completely specified simple hypotheses
  - For complex hypotheses, ℓ is not necessarily optimal

- The likelihood ratio is commonly used
- As any test statistic in the market, in order to select critical regions based on confidence levels it is necessary to know its distribution
  - Run toys to find its distribution (very expensive if you want to model extreme tails)
  - Find some asymptotic condition under which the likelihood ratio assumes a simple known form
- Wilks theorem: when the data sample size tends to ∞, the likelihood ratio tends to $\chi^2(N - N_0)$
  - Check if it's actually true!
    wget https://raw.githubusercontent.com/vischia/statex/master/wilks.R
    wget https://raw.githubusercontent.com/vischia/statex/master/wilks.ipynb

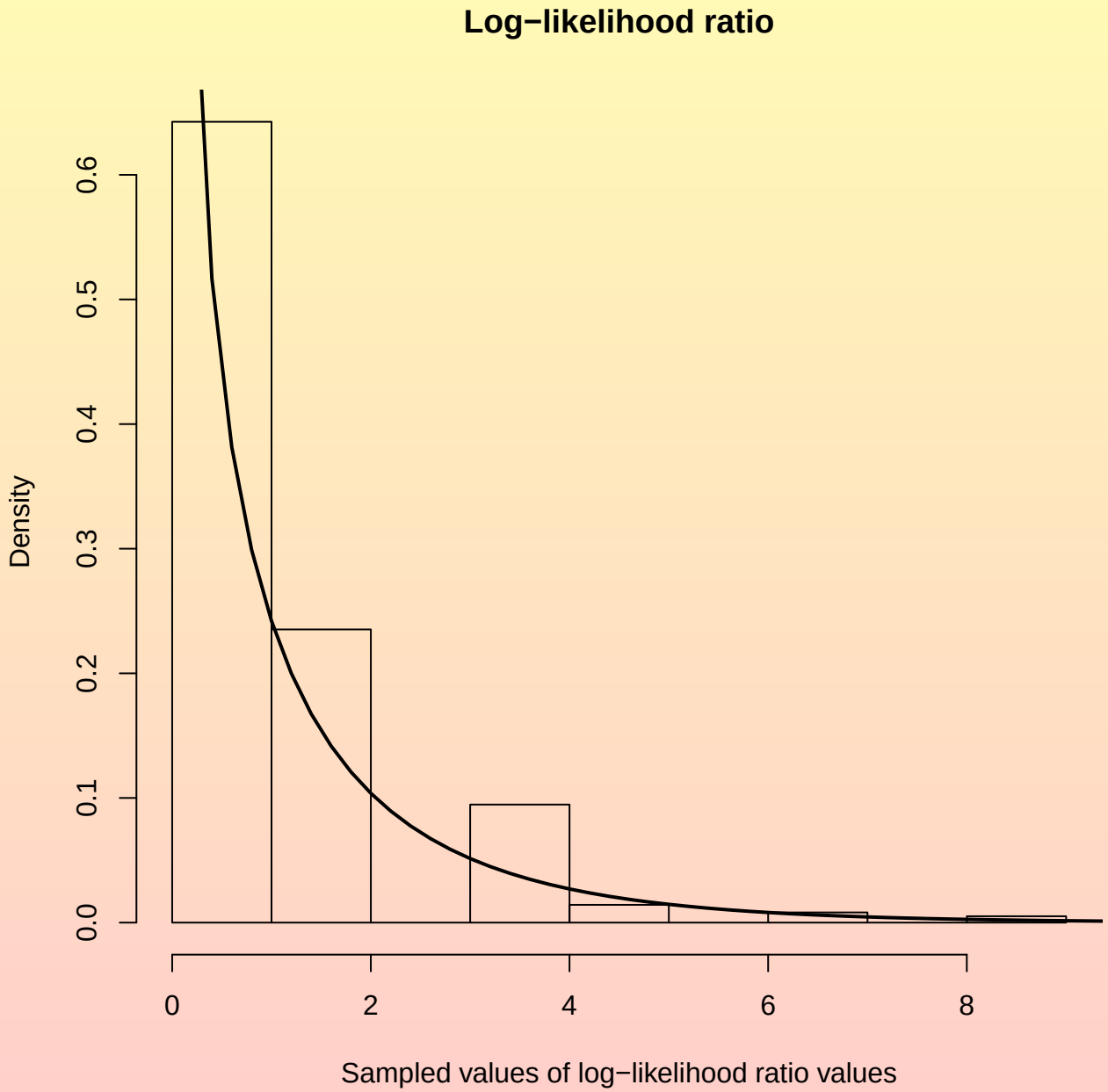
We can summarize in the

Theorem: If a population with a variate x is distributed according to the probability function $f(x, \theta_1, \theta_2 \cdots \theta_h)$, such that optimum estimates $\tilde{\theta}_i$ of the θ_i exist which are distributed in large samples according to (3), then when the hypothesis H is true that $\theta_i = \theta_{0i}$, $i = m + 1, m + 2, \cdots h$, the distribution of $-2 \log \lambda$, where λ is given by (2) is, except for terms of order $1/\sqrt{n}$, distributed like χ^2 with $h - m$ degrees of freedom.

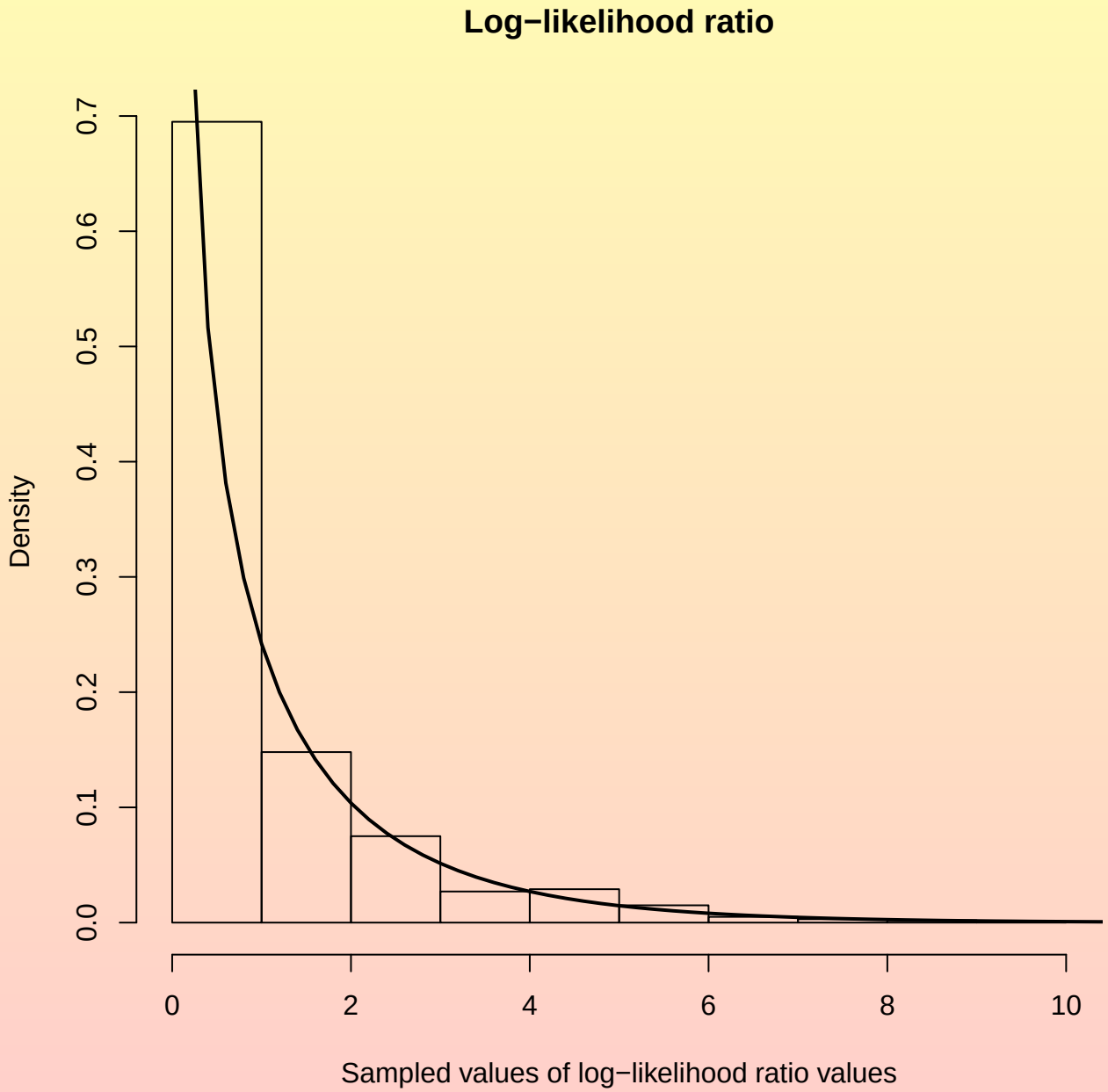
Verifying the Wilks theorem: N=2



Verifying the Wilks theorem: N=10



Verifying the Wilks theorem: N=100



Bayesian model selection — two models...

- The parameter θ might be predicted by two models M_0 and M_1: $P(\theta|\vec{x},M) = \frac{P(\vec{x}|\theta,M)P(\theta|M)}{P(\vec{x}|M)}$
  - A step further than yesterday in writing down the Bayes theorem: now multiple conditioning
  - $P(\vec{x}|M) = \int P(\vec{x}|\theta,M)P(\theta|M)d\theta$: *Bayesian evidence* or *model likelihood*
- Posterior for M_0: $P(M_0|\vec{x}) = \frac{P(\vec{x}|M_0)\pi(M_0)}{P(\vec{x})}$
- Posterior for M_1: $P(M_1|\vec{x}) = \frac{P(\vec{x}|M_1)\pi(M_1)}{P(\vec{x})}$
- The *odds* indicate relative preference of one model over the other
- Posterior odds: $\frac{P(M_0|\vec{x})}{P(M_1|\vec{x})} = \frac{P(\vec{x}|M_0)\pi(M_0)}{P(\vec{x}|M_1)\pi(M_1)}$
  - Posterior odds = Bayes Factor $\times$ prior odds
- $B_{01} := \frac{P(\vec{x}|M_0)}{P(\vec{x}|M_1)}$
- Various slightly different scales for the Bayes Factor
  - Interesting: deciban, unit supposedly theorized by Turing (according to IJ Good) as *the smallest change of evidence human mind can discern*

Jeffreys

K	dHart	bits	Strength of evidence
$< 10^0$	0	—	Negative (supports M_2)
10^0 to $10^{1/2}$	0 to 5	0 to 1.6	Barely worth mentioning
$10^{1/2}$ to 10^1	5 to 10	1.6 to 3.3	Substantial
10^1 to $10^{3/2}$	10 to 15	3.3 to 5.0	Strong
$10^{3/2}$ to 10^2	15 to 20	5.0 to 6.6	Very strong
$> 10^2$	> 20	> 6.6	Decisive

Kass and Raftery

$\log_{10} K$	K	Strength of evidence
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Trotta

|lnB|	relative odds	favoured model's probability	Interpretation
< 1.0	< 3:1	< 0.750	not worth mentioning
< 2.5	< 12:1	0.923	weak
< 5.0	< 150:1	0.993	moderate
> 5.0	> 150:1	> 0.993	strong

Images from Wikipedia and from Roberto Trotta, Chair Lemaitre Lectures 2018

UCLouvain
Institut de recherche
en mathématique et physique

Bayesian model comparison of 193 models
Higgs inflation as reference model

$\ln(\mathcal{E}/\mathcal{E}_{\rm HI})$

Martin,RT+14

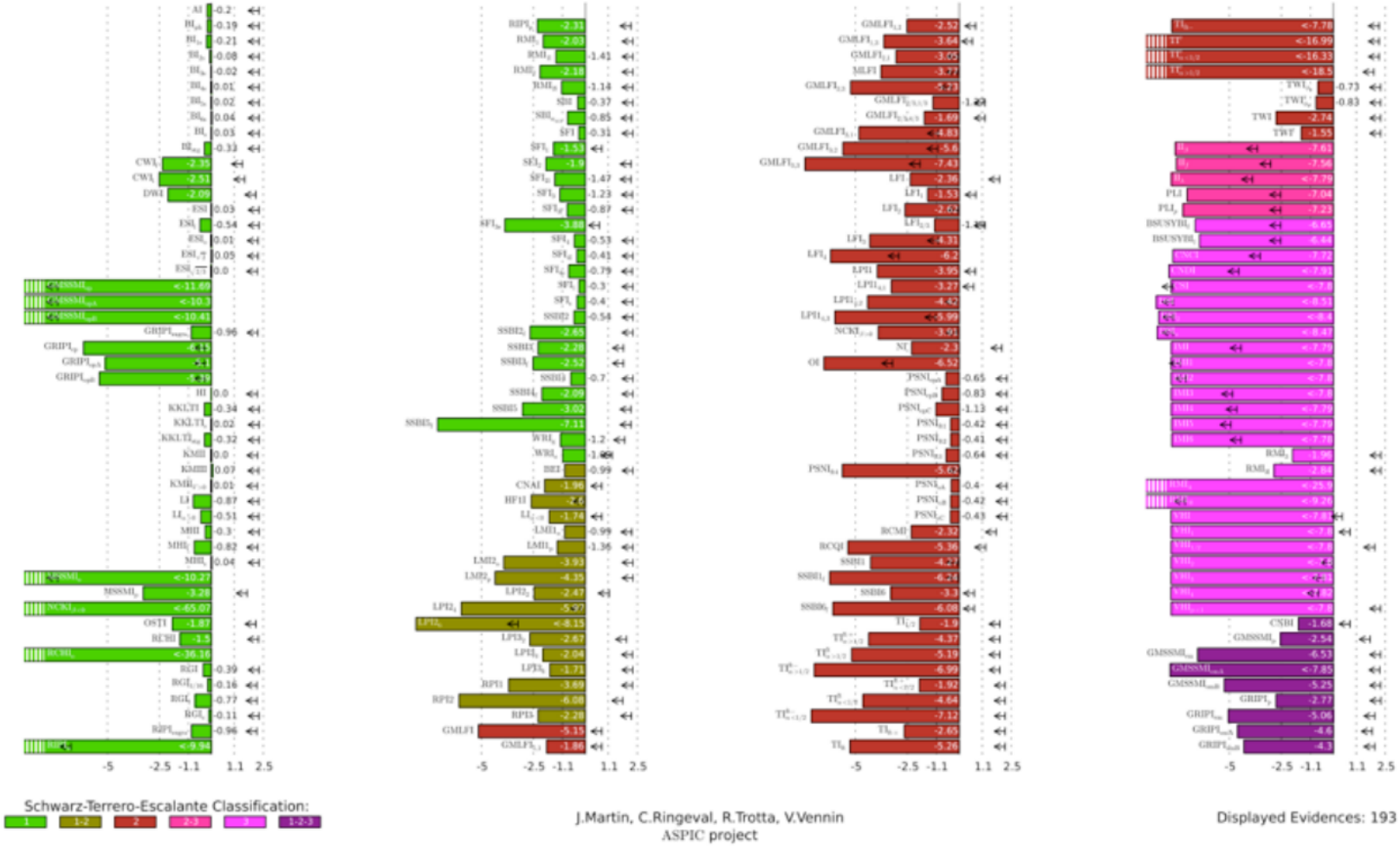
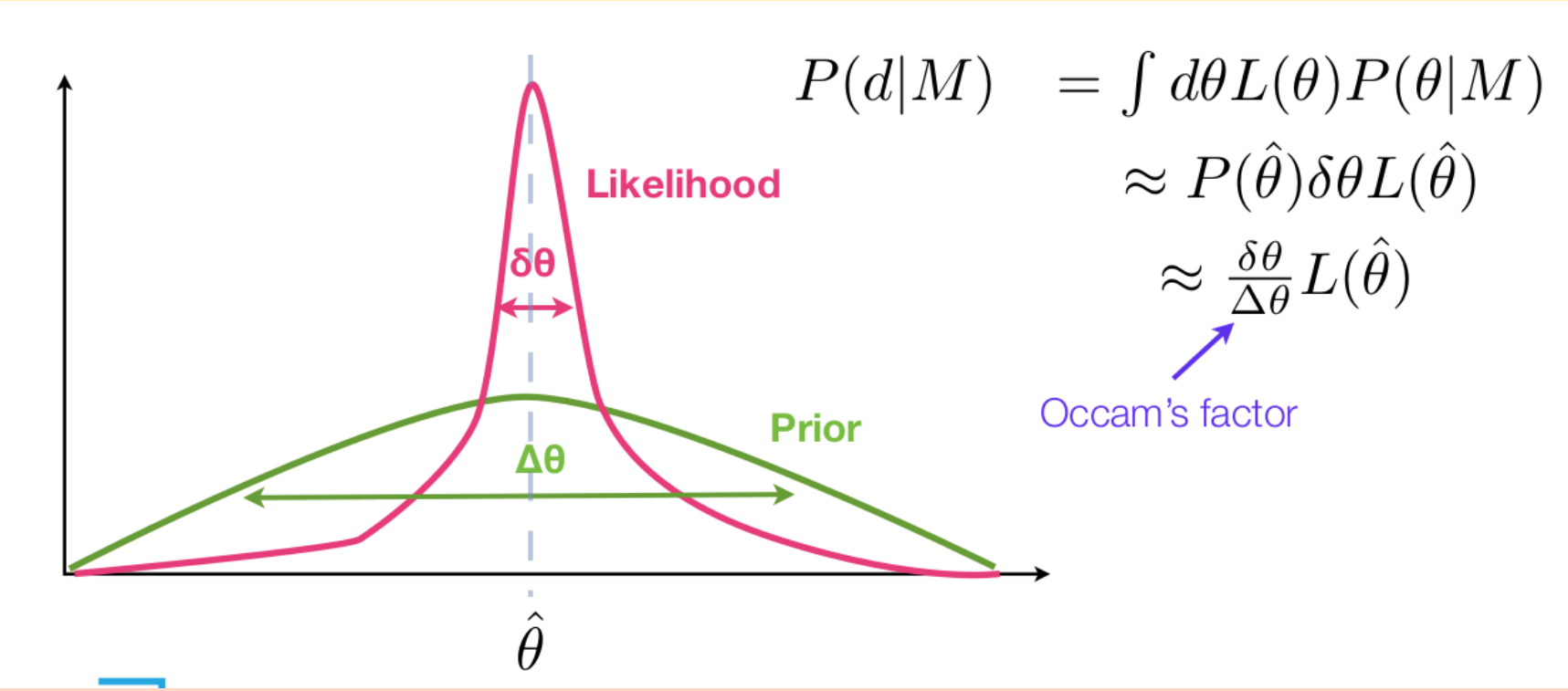


Image from Roberto Trotta, Chair Lemaitre Lectures 2018

Bayesian model selection — Discourage nonpredictive models

- The Bayes Factor also takes care of penalizing excessive model complexity
- Highly predictive models are rewarded, broadly-non-null priors are penalized



From Roberto Trotta, Chair Lemaitre Lectures 2018

Bayes vs p-values: the Jeffreys-Lindley paradox

- Data X (N data sampled from $f(x|\theta)$)<
  - H_0:$\theta = \theta_0$. Prior: π_0 (non-zero for point mass, Dirac's δ, counting measure)
  - H_1: $\theta ! = \theta_0$. Prior: $\pi_1 = 1 - \pi_0$ (usual Lebesgue measure)
- Conditional on H_1 being true:
  - Prior probability density $g(\theta)$
  - If $f(x|\theta) \sim Gaus(\theta, \sigma^2)$, then the sample mean $\bar{X} \sim Gaus(\theta, \sigma_{tot} = \sigma/N)$
- Likelihood ratio of H_0 to best fit for H_1: $\lambda = \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\hat{\theta})} = exp(-Z^2/2) \propto \frac{\sigma_{tot}}{\tau} B_{01}$; $Z := \frac{\hat{\theta}-\theta_0}{\sigma_{tot}}$
  - λ disfavours the null hypothesis for large significances (small p-values), independent of sample size
  - B_{01} includes σ_{tot}/τ (Ockham Factor, penalizing H_1 for imprecise determination of θ), sample dependent!
- For arbitrarily large Z (small p-values), λ disfavours H_0, while there is always a N for which B_{01} favours H_0 over H_1

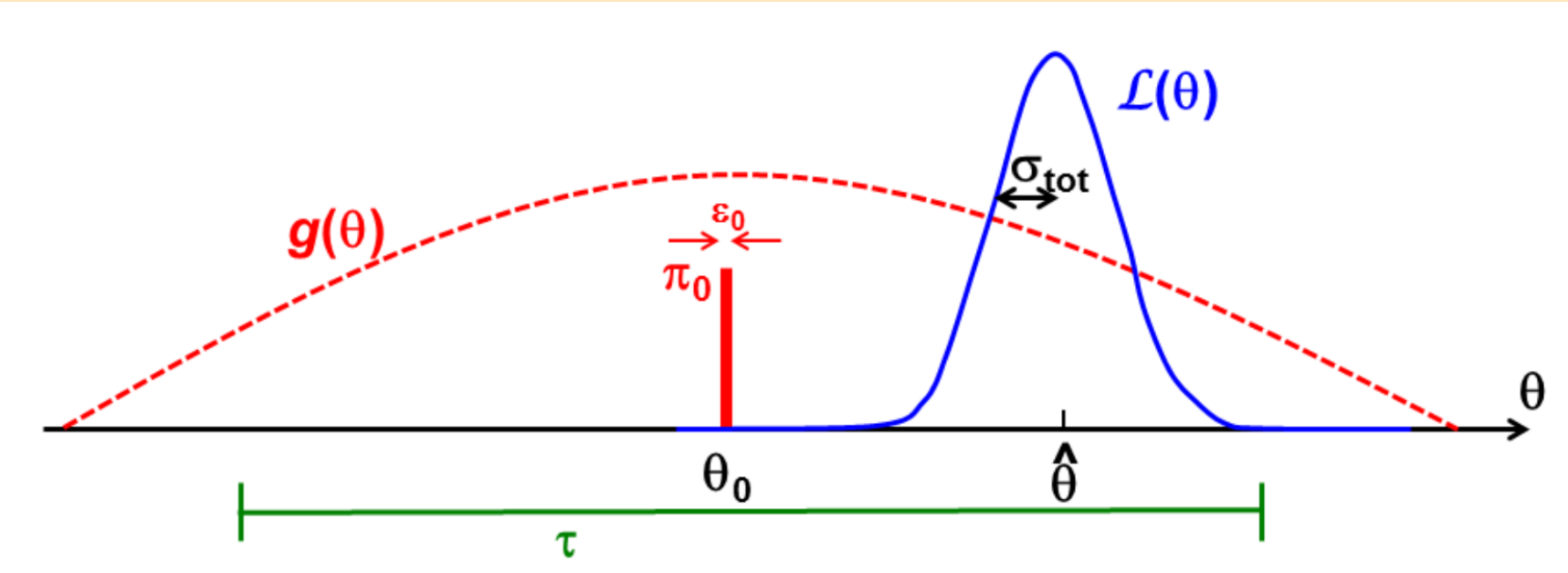


Image from Cousins, doi:10.1007/s11229-014-0525-z

Testing hypotheses near the boundary: Zech

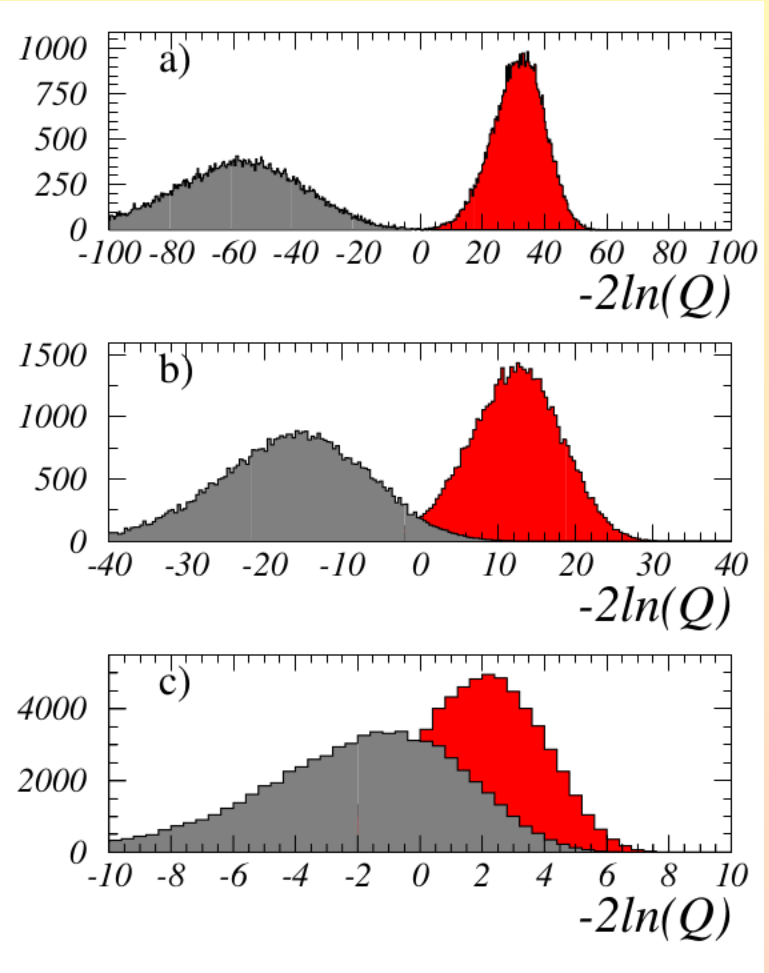
- Counting experiment: observe n events
- Assume they come from Poisson processes: $n \sim Pois(s+b)$, with known b
- Set limit on s given n_{obs}
- Exclude values of s for which $P(n \leq n_{obs}|s+b) \leq \alpha$ (guaranteed coverage $1-\alpha$)
- $b = 3$, $n_{obs} = 0$
  - Exclude $s+b \leq 3$ at 95%CL
  - Therefore excluding $s \leq 0$, i.e. **all** possible values of s (can't distinguish b-only from very-small-s)
- Zech: let's condition on $n_b \leq n_{obs}$ (n_b unknown number of background events)
  - For small n_b the procedure is more likely to undercover than when n_b is large, and the distribution of n_b is independent of s
  - $P(n \leq n_{obs}|n_b \leq n_{obs}, s+b) = \ldots = \frac{P(n \leq n_{obs}|s+b)}{P(n \leq n_{obs}|b)}$

Beyond coverage: CLs

- Goal: seamless transition between exclusion, observation, discovery (historically for the Higgs)
  - Exclude Higgs as strongly as possible in its absence (in a region where we would be sensitive to its presence)
  - Confirm its existence as strongly as possible in its presence (in a region where we are sensitive to its presence)
  - Maintain Type I and Type II errors below specified (small) levels
- Identify observables, and a suitable test statistic Q
- Define rules for exclusion/discovery, i.e. ranges of values of Q leading to various conclusions
  - Specify the significance of the statement, in form of confidence level (CL)
- Confidence limit: value of a parameter (mass, xsec) excluded at a given confidence level CL
  - A confidence limit is an upper(lower) limit if the exclusion confidence is greater(less) than the specified CL for all values of the parameter below(above) the confidence limit
- The resulting intervals are neither frequentist nor bayesian!

Get your confidence levels right

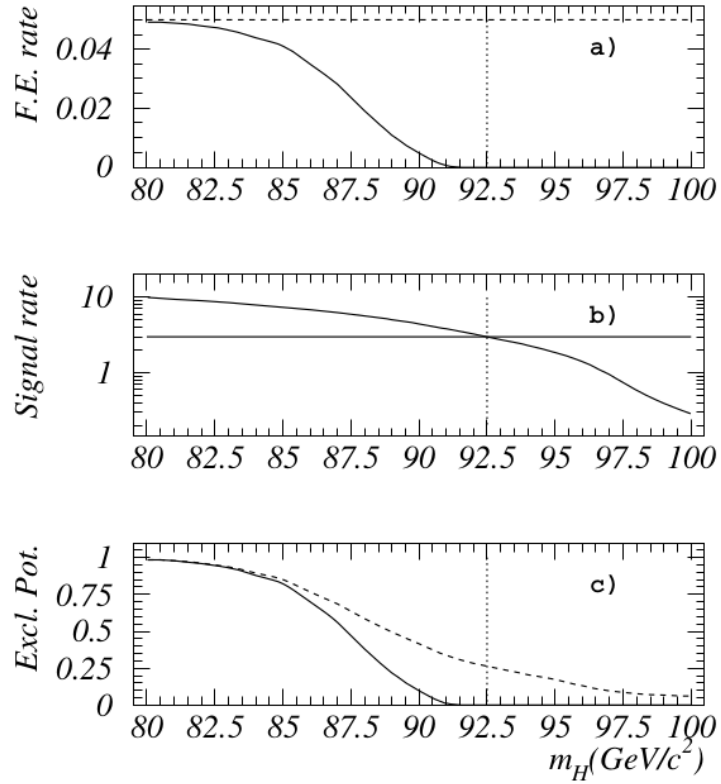
- Find a monotonic Q for increasing signal-like experiments (e.g. likelihood ratio)
- $CL_{s+b} = P_{s+b}(Q \leq Q_{obs})$
  - Small values imply poor compatibility with $S + B$ hypothesis, favouring B-only
- $CL_b = P_b(Q \leq Q_{obs})$
  - Large (close to 1) values imply poor compatibility with B-only, favouring $S + B$
- What to do when the estimated parameter is unphysical?
  - The same issue solved by Feldman-Cousins
  - If there is also underfluctuation of backgrounds, it's possible to exclude even zero events at 95%CL!
  - It would be a statement about future experiments
  - Not enough information to make statements about the signal
- Normalize the $S + B$ confidence level to the B-only confidence level!



Plot from Read, CERN-open-2000-205

Avoid issues at low signal rates

- $CL_s := \frac{CL_{s+b}}{CL_b}$
- Exclude the signal hypothesis at confidence level CL if $1 - CL_s \leq CL$
- Ratio of confidences is not a confidence
  - The hypotetical false exclusion rate is generally less than the nominal $1 - CL$ rate
  - CL_s and the actual false exclusion rate grow more different the more $S + B$ and B p.d.f. become similar
- CL_s increases coverage, i.e. the range of parameters that can be exclude is reduced
  - It is more conservative
  - Approximation of the confidence in the signal hypothesis that might be obtained if there was no background
- Avoids the issue of CL_{s+b} with experiments with the same small expected signal
  - With different backgrounds, the experiment with the larger background might have a better expected performance
- Formally corresponds to have $H_0 = H(\theta! = 0)$ and test it against $H_1 = H(\theta = 0)$
  - Test inversion!

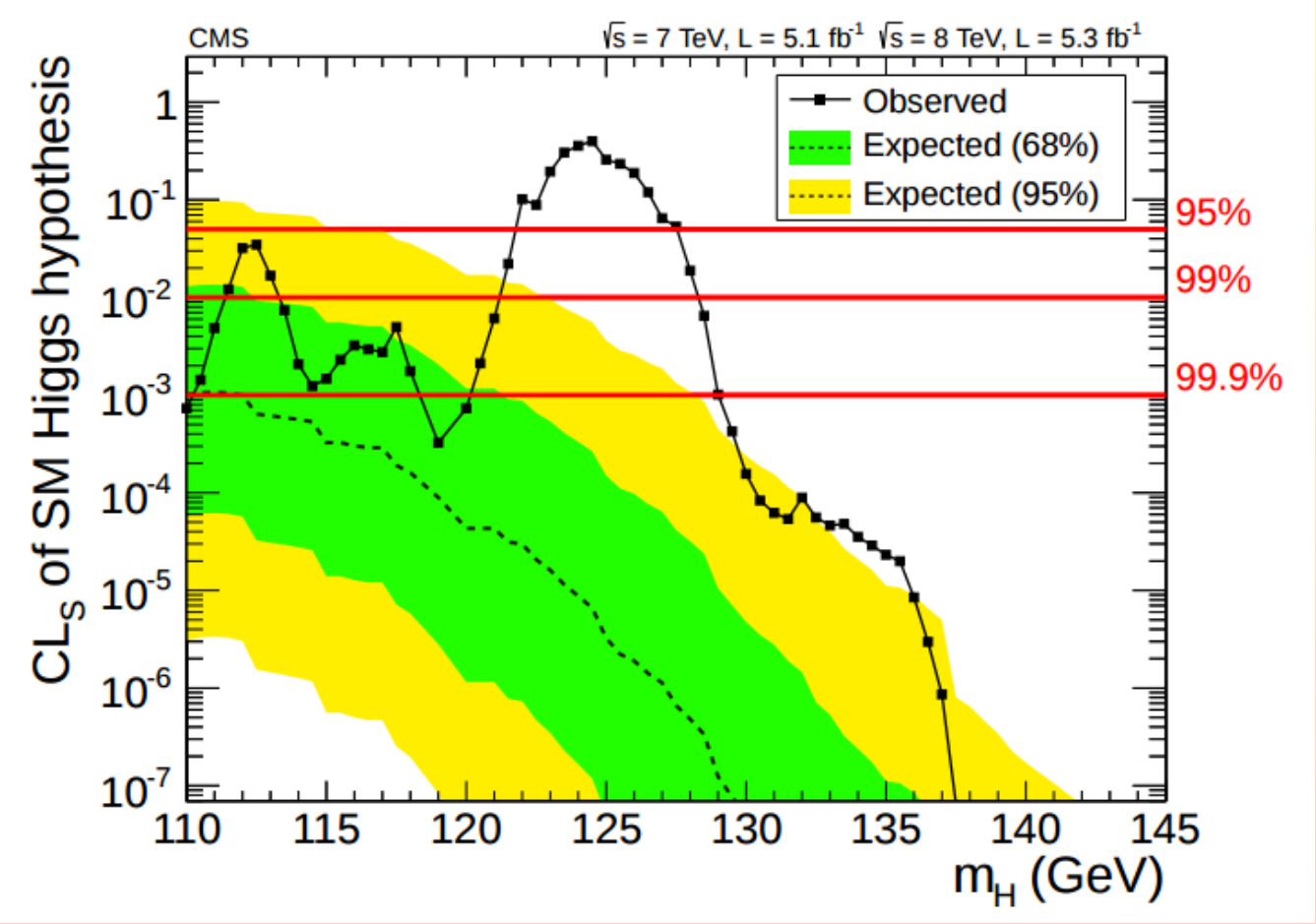


Dashed: CL_{s+b}
Solid: CL_s
$S < 3$: exclusion for a B-free search $\equiv 0$

Plot from Read, CERN-open-2000-205

A practical example: Higgs discovery - 1

- Apply the CL_s method to each Higgs mass point
- Green/yellow bands indicate the $\pm 1\sigma$ and $\pm 2\sigma$ intervals for the expected values under B-only hypothesis
  - Obtained by taking the quantiles of the B-only hypothesis



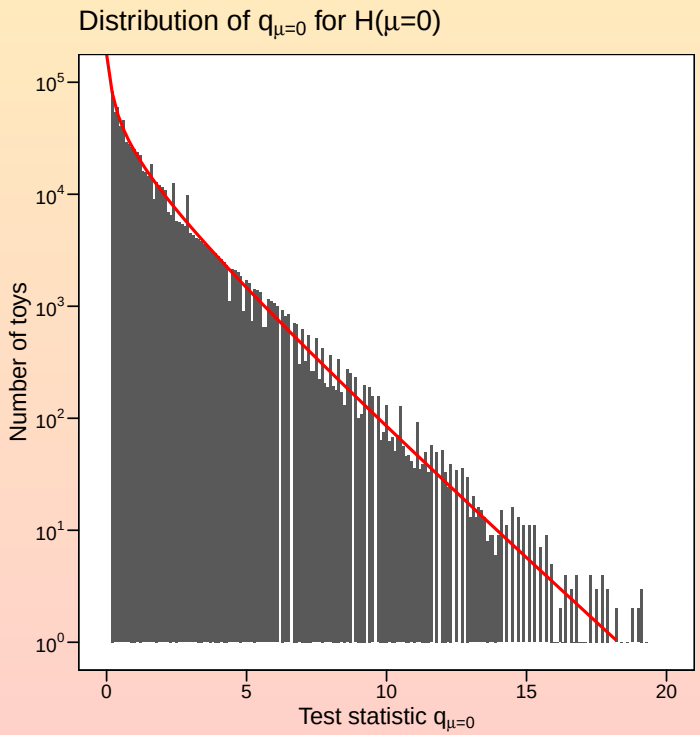
Plot from Higgs discovery paper

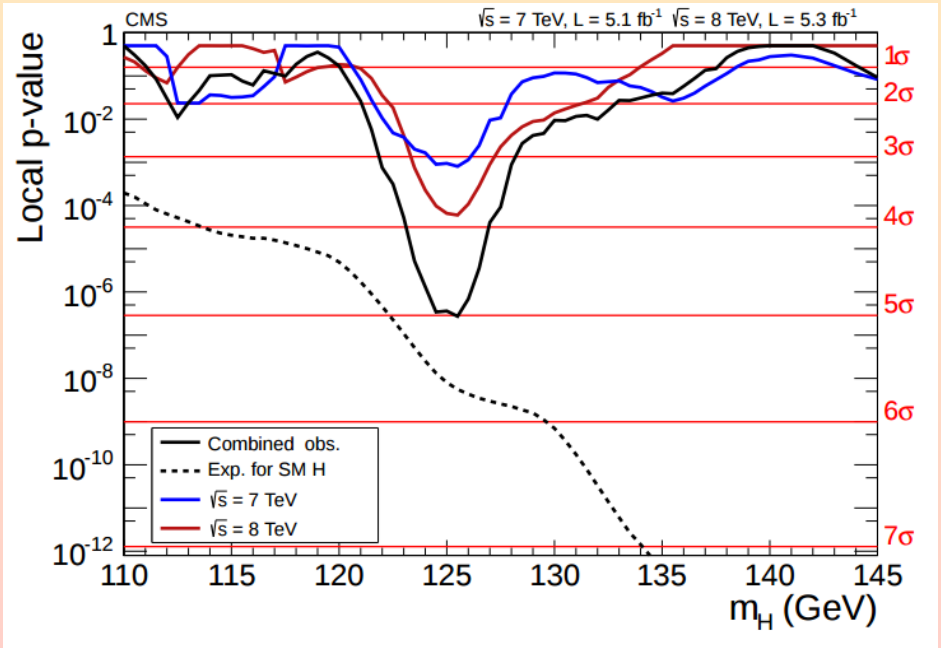
Deal with CLs!

- Now let's play with CLs!
- wget https://raw.githubusercontent.com/vischia/statex/master/cls_counting.ipynb
- You will need to install the first two (the other two are for the next exercises)
  - `pip3 install pyhf -user`
  - `pip3 install uproot -user`
  - `pip3 install -user pyunfold`
  - `pip3 install -user seaborn`

Quantifying excesses

- Quantify the presence of the signal by using the background-only p-value
  - Probability that the background fluctuates yielding and excess as large or larger of the observed one
- For the mass of a resonance, $q_0 = -2log\frac{\mathcal{L}(data|0,\hat{\theta}_0)}{\mathcal{L}(data|\hat{\mu},\hat{\theta})}$, with $\hat{\mu} \geq 0$
  - Interested only in upwards fluctuation, accumulate downwards one to zero
- Use pseudo-data to generate background-only Poisson counts and nuisance parameters θ_0^{obs}
  - Use distribution to evaluate tail probability $p_0 = P(q_0 \leq q_0^{obs})$
  - Convert to one-sided Gaussian tail areas by inverting $p = \frac{1}{2}P_{\chi_1^2}(Z^2)$

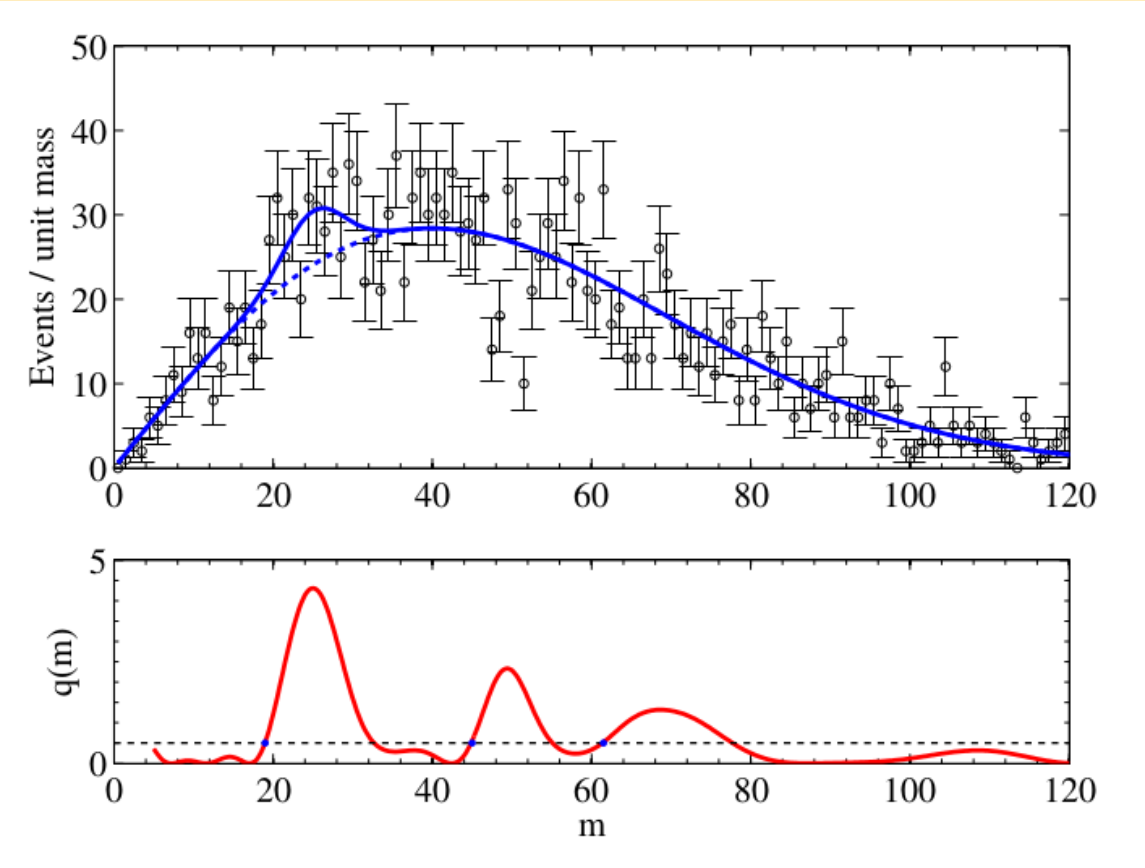




Left plot by Pietro Vischia, right plot from ATL-PHYS-PUB-2011-011 and Higgs discovery paper

The Look-elsewhere effect — 1

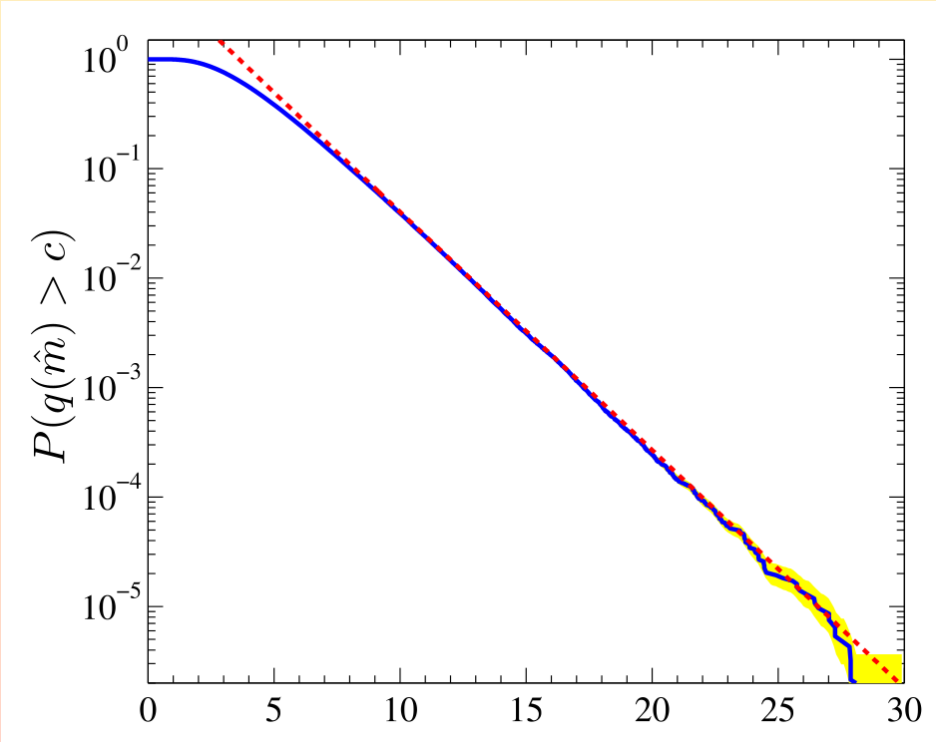
- Searching for a resonance X of arbitrary mass
  - $H_0 =$ no resonance, the mass of the resonance is not defined (Standard Model)
  - $H_1 = H(M \neq 0)$, but there are infinite possible values of M
- Wilks theorem not valid anymore, no unique test statistic encompassing every possible H_1
- Quantify the compatibility of an observation with the B-only hypothesis
  - $q_0(\hat{m_X}) = \max_{m_X} q_0(m_X)$
  - Write a global p-value as $p_b^{global} := P(q_0(\hat{m_X}) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi^2_1}(u)$
  - u fixed confidence level
  - Crossings (Davis, Biometrika 74, 33–43 (1987)) , computable using pseudo-data (toys)

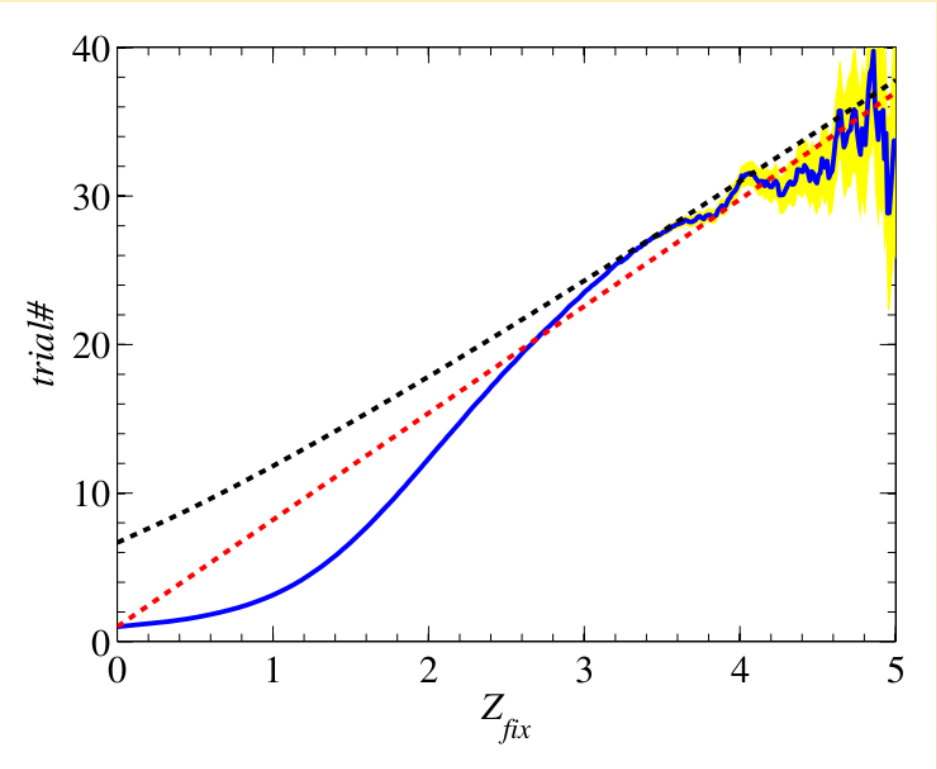


Plot from Gross-Vitells, 10.1140/epjc/s10052-010-1470-8

The Look-elsewhere effect — 2

- Ratio of local (excess right here) and global (excess anywhere) p-values: trial factor
- Asymptoticly linear in the number of search regions and in the fixed significance level
  - Dashed red lines: prediction based on the formula with upcrossings
  - Blue: 10^6 toys (pseudoexperiments)
- Here *asymptotic* means *for increasingly smaller tail probabilities*

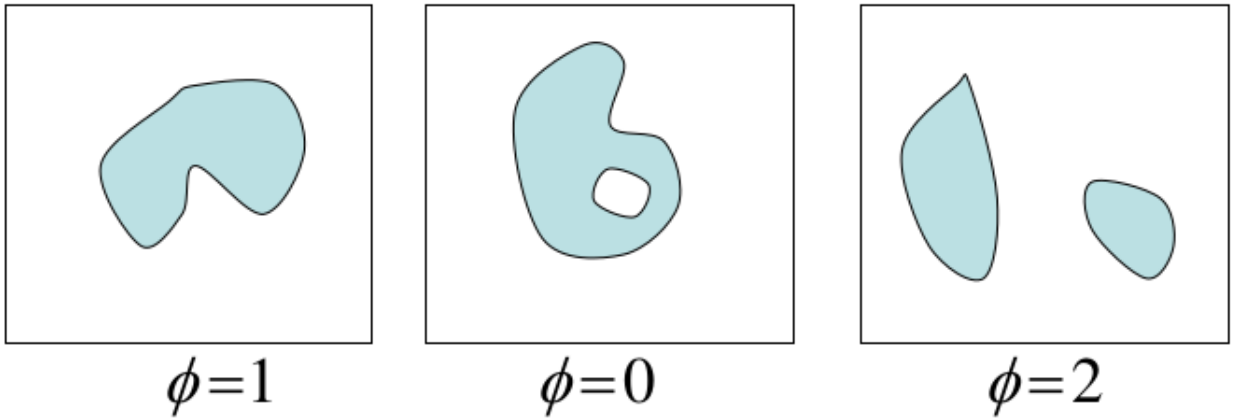


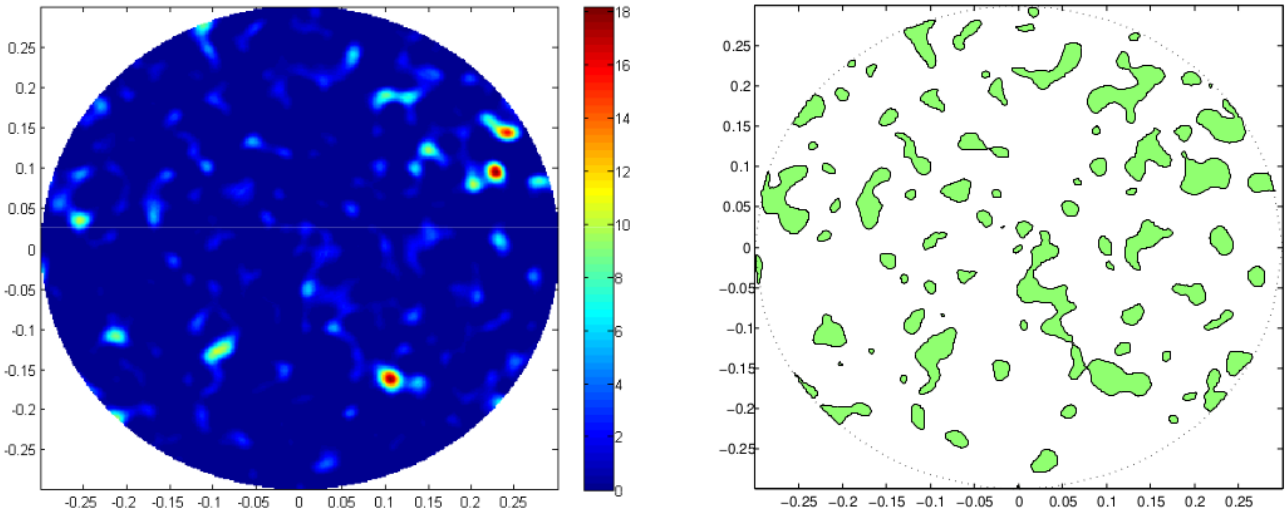


Plot from Gross-Vitells, 10.1140/epjc/s10052-010-1470-8

The Look-elsewhere effect, now also in 2D — 1

- Extension to two dimensions requires using the theory of random fields
  - Excursion set: set of points for which the value of a field is larger than a threshold u
  - Euler characteristics interpretable as number of disconnected regions minus number of holes

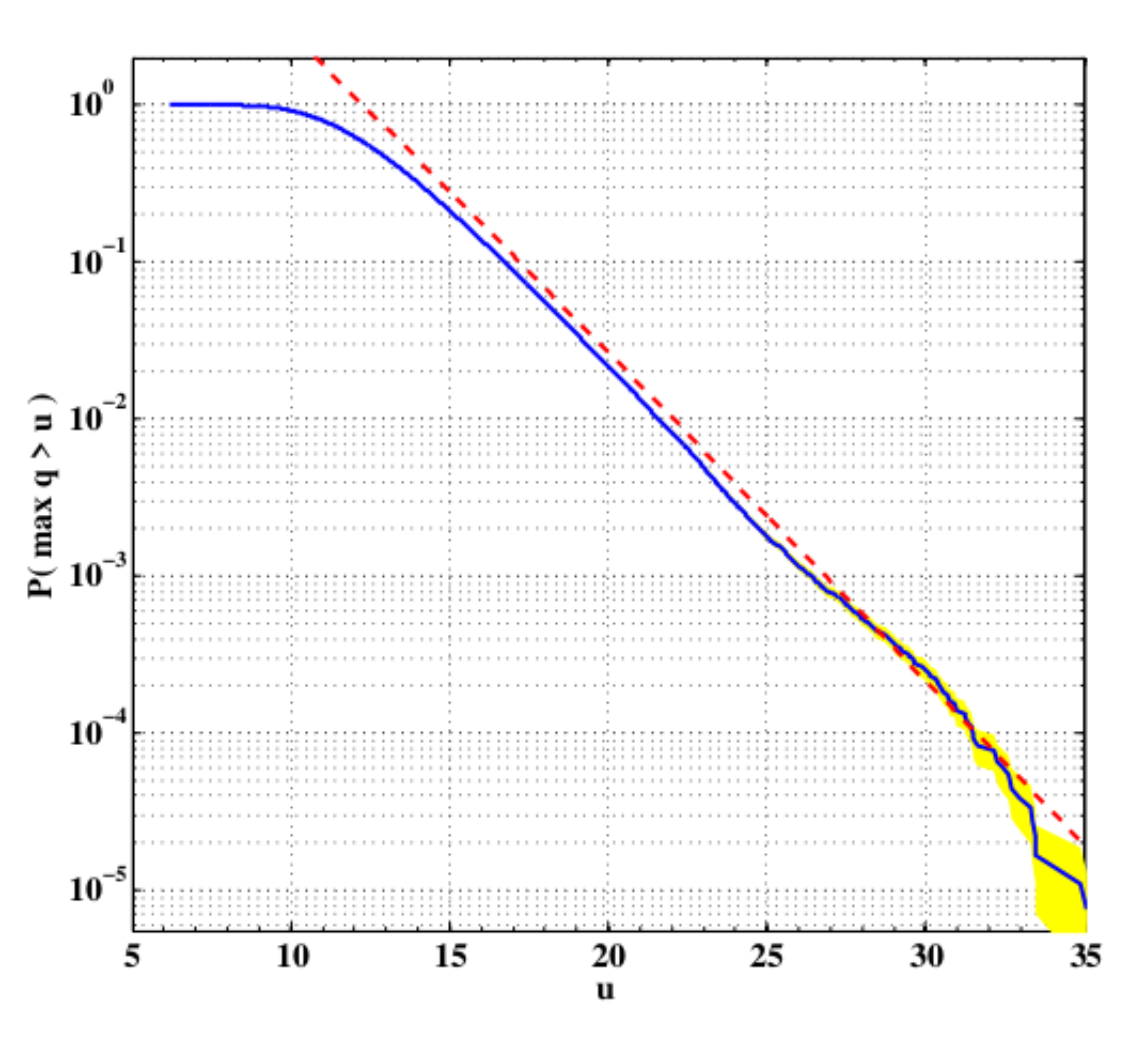




Plot from Gross-Vitells, 10.1016/j.astropartphys.2011.08.005

The Look-elsewhere effect, now also in 2D — 2

- Asymptoticity holds also for the 2D effect, as desired
  - Dashed red lines: prediction based on the formula with upcrossings
  - Blue: $200k$ toys (pseudoexperiments)



Plot from Gross-Vitells, 10.1016/j.astropartphys.2011.08.005

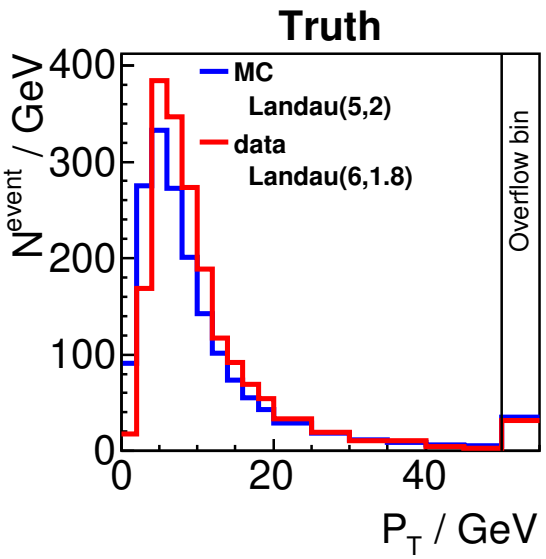
Measuring differential distributions

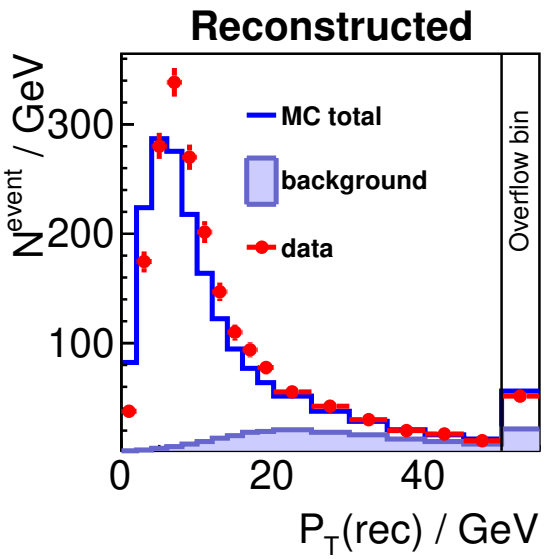
Unfolding: the problem

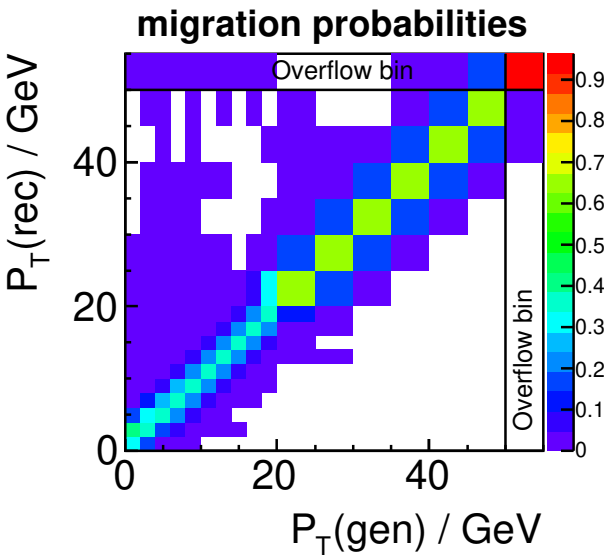
- Unfolding it's about how to invert a matrix that should not be inverted

$$\mathcal{L} = (\mathbf{y} - \mathbf{A}\mathbf{x})^T \mathbf{V}_{\mathbf{yy}} (\mathbf{y} - \mathbf{A}\mathbf{x}),$$

- Observations y, to be transformed in the theory space into x
  - Model the detector as a response matrix
  - Invert the response to convert experimental data to theory space distributions
  - Usually to compare with models in the theory space
- The best solution is to fold any new theory and make comparisons in the experimental data space







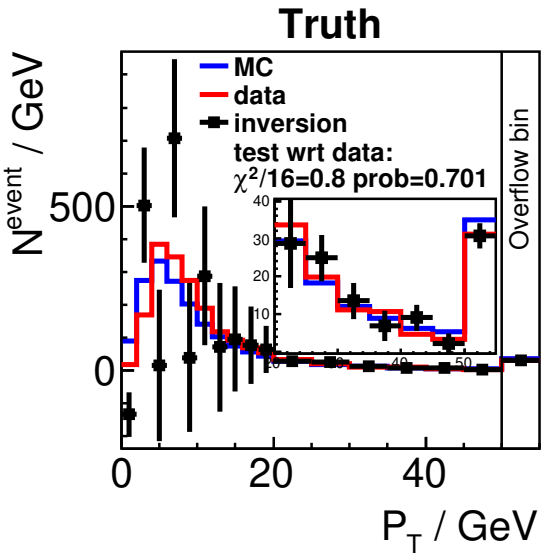
Plot from ArXiv:1611.01927

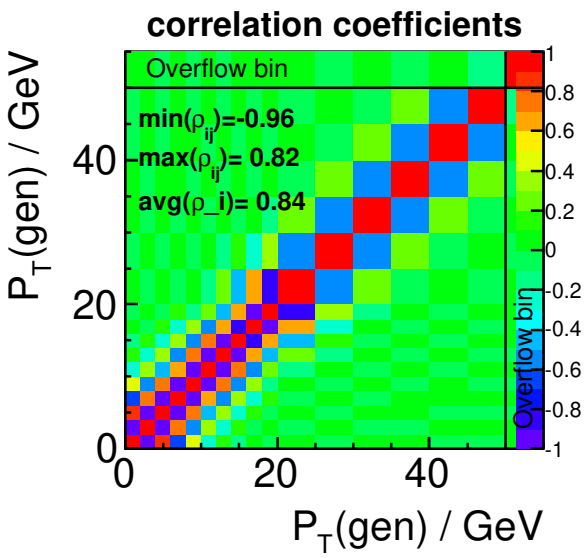
Unfolding: naïve solutions

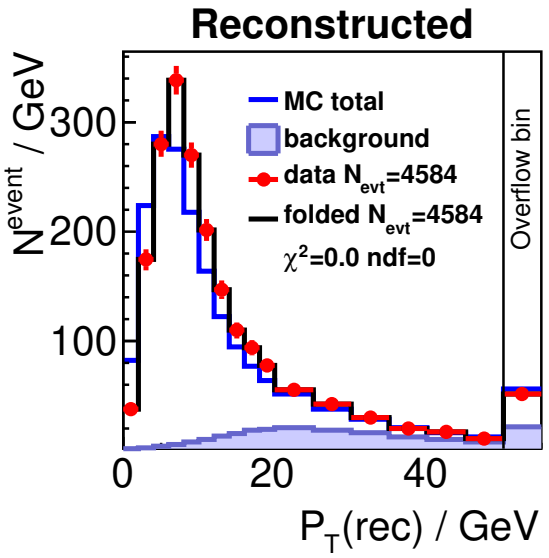
- Bin-by-bin correction factors $\hat{x}_i = (y_i - b_i)\frac{N_i^{\text{gen}}}{N_i^{\text{rec}}}$; disfavoured
  - Heavy biases due to the underlying MC truth
  - Yields the wrong normalization for the unfolded distribution
- Invert the response matrix $\hat{\boldsymbol{x}} = \boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b})$
  - Only for square matrices, but always unbiased
  - Oscillation patterns (small determinants in matrix inversion)
  - Patterns also seen as large negative $\rho_{ij} \sim -1$ near diagonal
  - Result is correct within uncertainty envelope given by $\boldsymbol{V}_{xx}$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

determinant



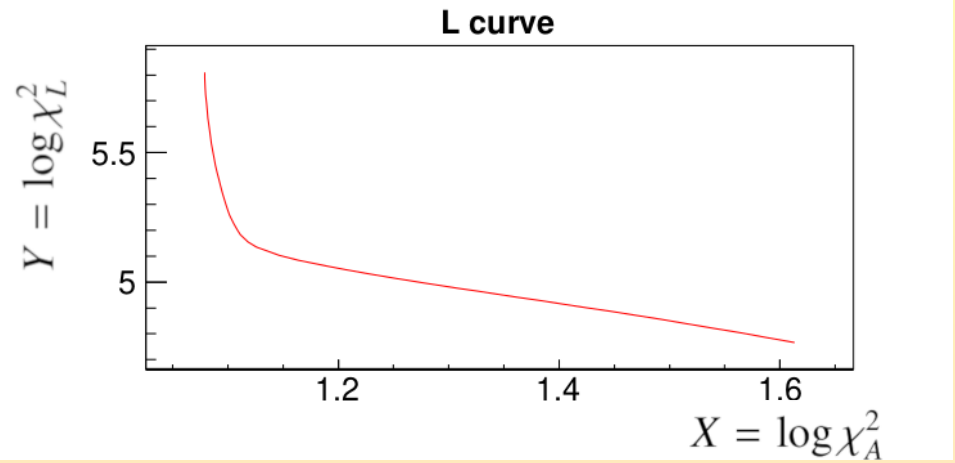




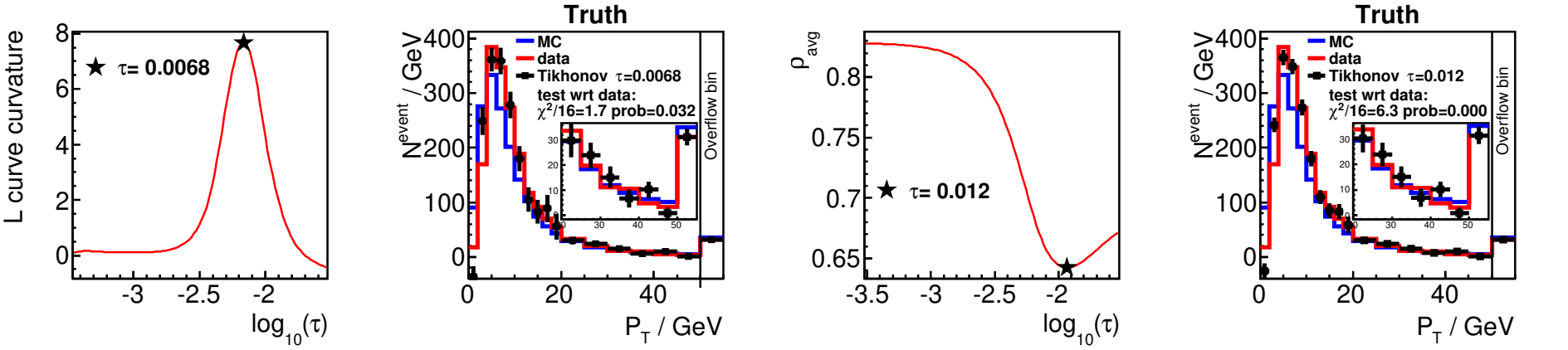
Cartoon from https://www.mathsisfun.com/algebra/matrix-inverse.html, plots from ArXiv:1611.01927

Unfolding: regularization 1/

$$\chi^2_{\text{TUnfold}} = \chi^2_A + \tau^2 \chi^2_L$$
$$\chi^2_A = (\boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{b} - \boldsymbol{y})^{\mathsf{T}} (\boldsymbol{V_{yy}})^{-1} (\boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{b} - \boldsymbol{y})$$
$$\chi^2_L = (\hat{\boldsymbol{x}} - \boldsymbol{x_B})^{\mathsf{T}} \boldsymbol{L}^{\mathsf{T}} \boldsymbol{L} (\hat{\boldsymbol{x}} - \boldsymbol{x_B})$$



- Choose τ corresponding to maximum curvature of L-curve
- Or minimize the global $\rho_{\text{avg}} = \frac{1}{M_x} \sum_{j=1}^{M_x} \rho_j$
  - Often results in stronger regularization than L-curve



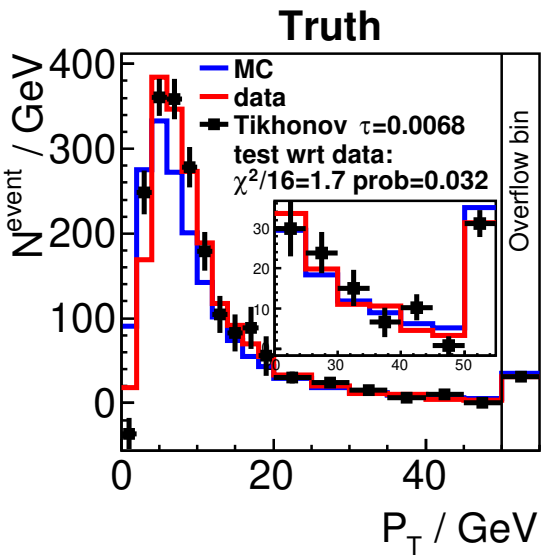
Plots from ArXiv:1611.01927

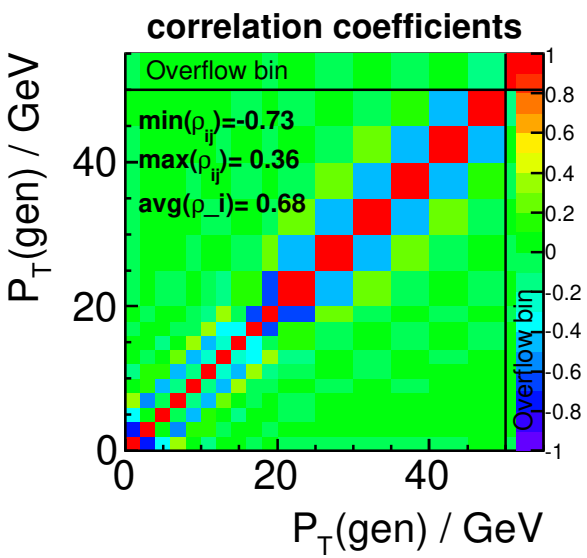
Unfolding: regularization 2/

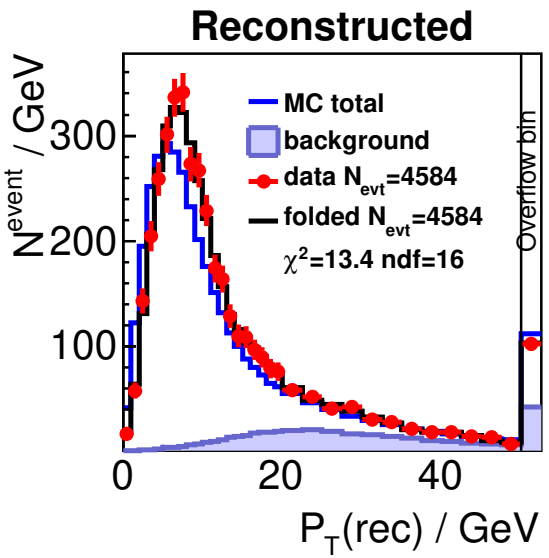
UCLouvain
Institut de recherche en mathématique et physique

$$\mathcal{L}(\mathbf{x}, \lambda) = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3,$$
$$\mathcal{L}_1 = (\mathbf{y} - \mathbf{A}\mathbf{x})^T \mathbf{V}_{\mathbf{yy}} (\mathbf{y} - \mathbf{A}\mathbf{x}),$$
$$\mathcal{L}_2 = \tau^2 (\mathbf{x} - f_b \mathbf{x_0})^T (\mathbf{L}^T \mathbf{L}) (\mathbf{x} - f_b \mathbf{x_0}),$$
$$\mathcal{L}_3 = \lambda (Y - \mathbf{e}^T \mathbf{x}),$$
$$Y = \sum_i y_i,$$
$$e_j = \sum_i A_{ij}.$$

- $\mathbf{y}$: observed yields
- $\mathbf{A}$: response matrix
- $\mathbf{x}$: the unfolded result
- $\mathcal{L}_1$: least-squares minimization ($V_{ij} = e_{ij}/e_{ii}e_{jj}$ correlation coefficients)
- $\mathcal{L}_2$: regularization with strength τ
- Bias vector $f_b\mathbf{x_0}$: reference with respect to which large deviations are suppressed
- $\mathcal{L}_3$; area constraint (bind unfolded normalization to the total yields in folded space)







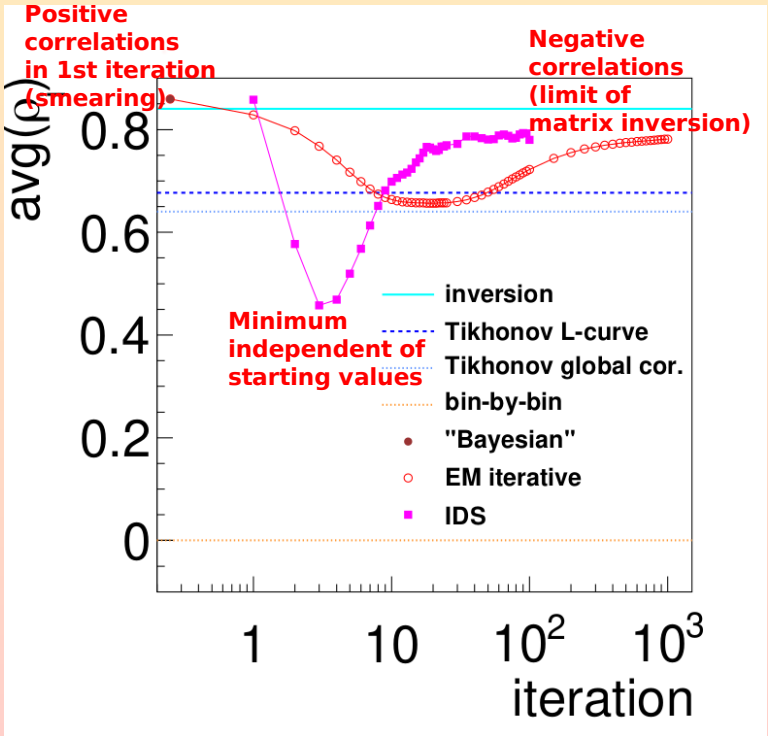
Plots from ArXiv:1611.01927

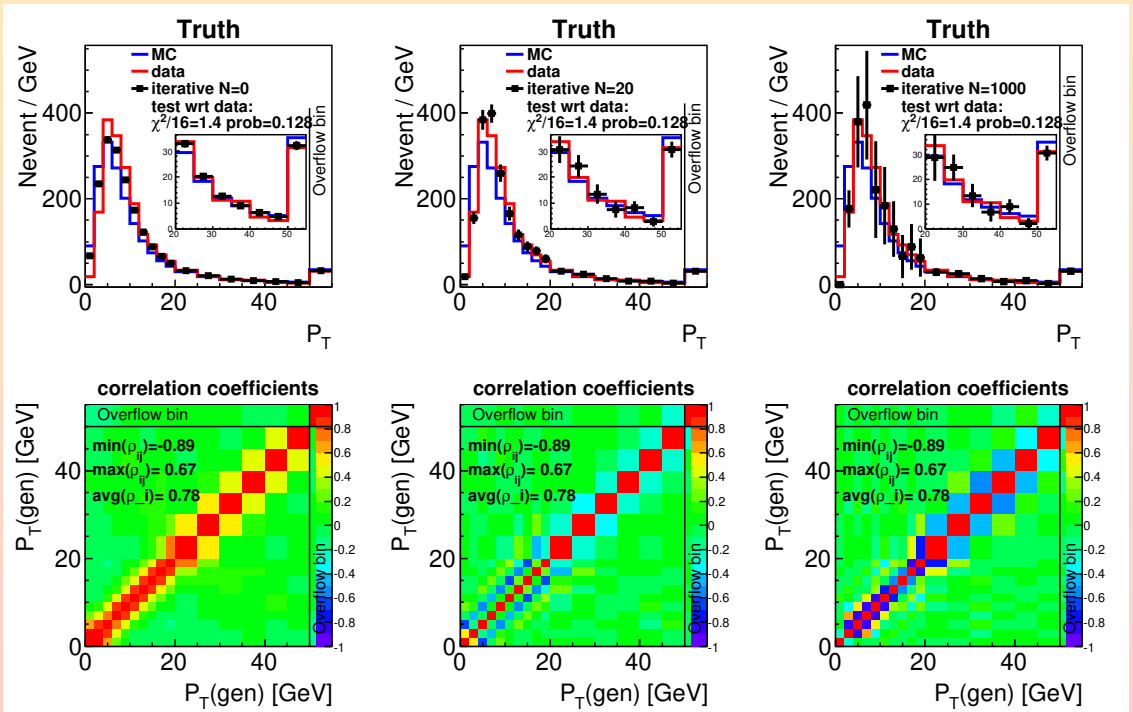
Unfolding: Iterative Unfolding

- Iterative improvement over the result of a previous iteration;

$$x_j^{(n+1)} = x_j^{(n)} \sum_{i=1}^{M} \frac{A_{ij}}{\epsilon_j} \frac{y_i}{\sum_{k=1}^{N} A_{ik} x_k^{(n)} + b_i}$$

  - It converges (slowly, $N_{iter} \sim N_{bins}^2$) to the MLE of the likelihood for independent Poisson-distributed y_i
  - Not necessarily unbiased for correlated data (does not make use of covariance of input data V_{yy})
- In HEP most people don't iterate until convergence
  - Fixed N_{iter} is often used; the dependence on starting values provides regularization
- Intrinsically frequentist method
  - for $N_{iter} \to \infty$ converges to matrix inversion, if all $\hat{x}_j$ from matrix inversion are positive
  - $N_{iter} = 0$ sometimes called improperly "Bayesian" unfolding (the author, D'Agostini, is Bayesian)
- Don't use software defaults!!! (e.g. some software has $N_{iter} = 4$)
  - Minimizing the global ρ is a good objective criterion, but there are others (Akaike information, etc)





Plots from ArXiv:1611.01927

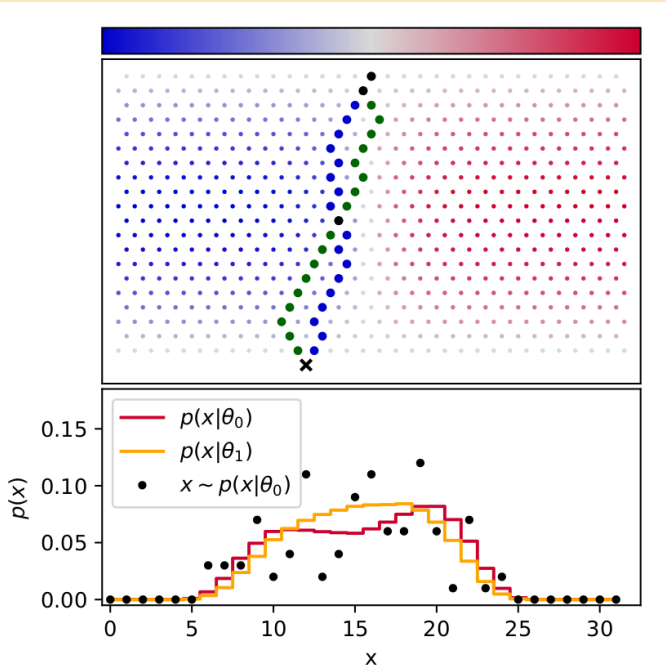
Exercise time!

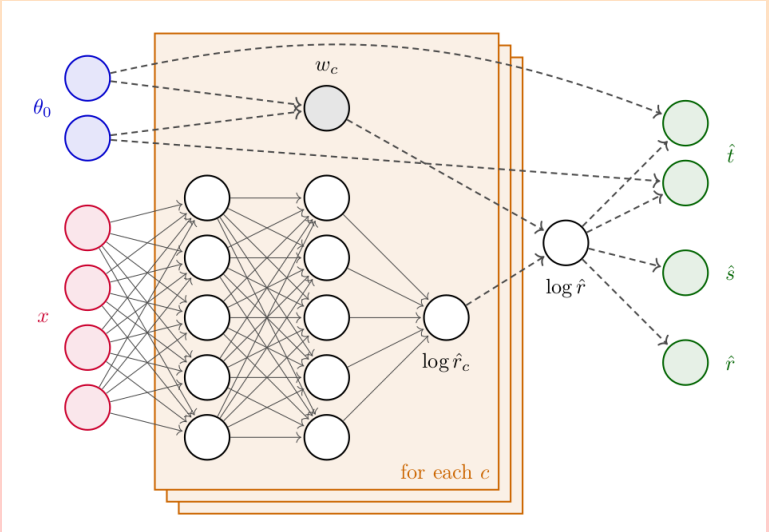
UCLouvain
Institut de recherche
en mathématique et physique

- I don't really have to add anything to the wonderful `pyunfold` tutorials: https://github.com/jrbourbeau/pyunfold/tree/master/docs/source/notebooks
- Basic unfolding
  wget tutorial.ipynb
- Change your prior!
  wget user_prior.ipynb
- Regularization
  wget regularization.ipynb
- Multivariate unfolding
  wget multivariate.ipynb
- You can get them all by running
  the pyunfold/https://raw.githubusercontent.com/vischia/statex/master/pyunfold/get.sh script from the exercises repository

What if we don't have a likelihood?

- Likelihood $p(x|\theta) = \int dz p(x,z|\theta) = \int dz p_x(x|\theta,z) \prod_i p_i(z_i|\theta, z_{<i})$
  - Latent states sampled from $z_i \sim p_i(z_i|\theta, z_{<i})$
  - Final output sampled from $x \sim p_x(x|\theta, z)$
  - Observables x from particle generator; dependency on latent zs (matrix element, parton shower, detector...)
- Want to do inference in θ given a $p(x|\theta)$ which is intractable; likelihood trick;
  - Train a classifier (NN) to separate samples from $p(x|\theta_0)$ and $p(x|\theta_1)$
  - Likelihood ratio between θ_0 and θ_1 by inverting the minimization of the binary cross-entropy loss
- Joint score $t(x,z|\theta_0)$ and likelihood ratio $r(x,z|\theta_0,\theta_1)$ computable from simulated samples
  - Train parameterized estimators, then likelihood ratio is the minimum of loss function
  - Or local approximation, then the score is a sufficient statistic for inference
- Rewrite the EFT likelihood in a basis in which it is a mixture model
- Calculate the full true parton-level likelihood starting from N simulated events
  - Obtain a sufficient statistic for inference; exploit all available information!
  - Inference not limited anymore by the size of the generated samples

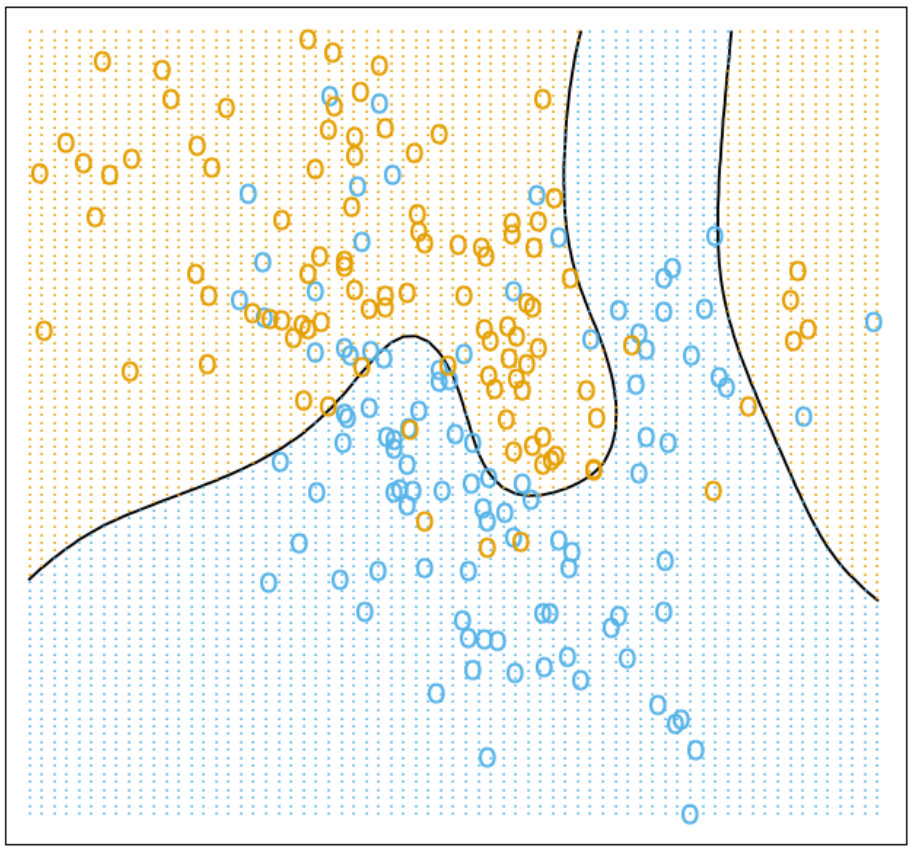


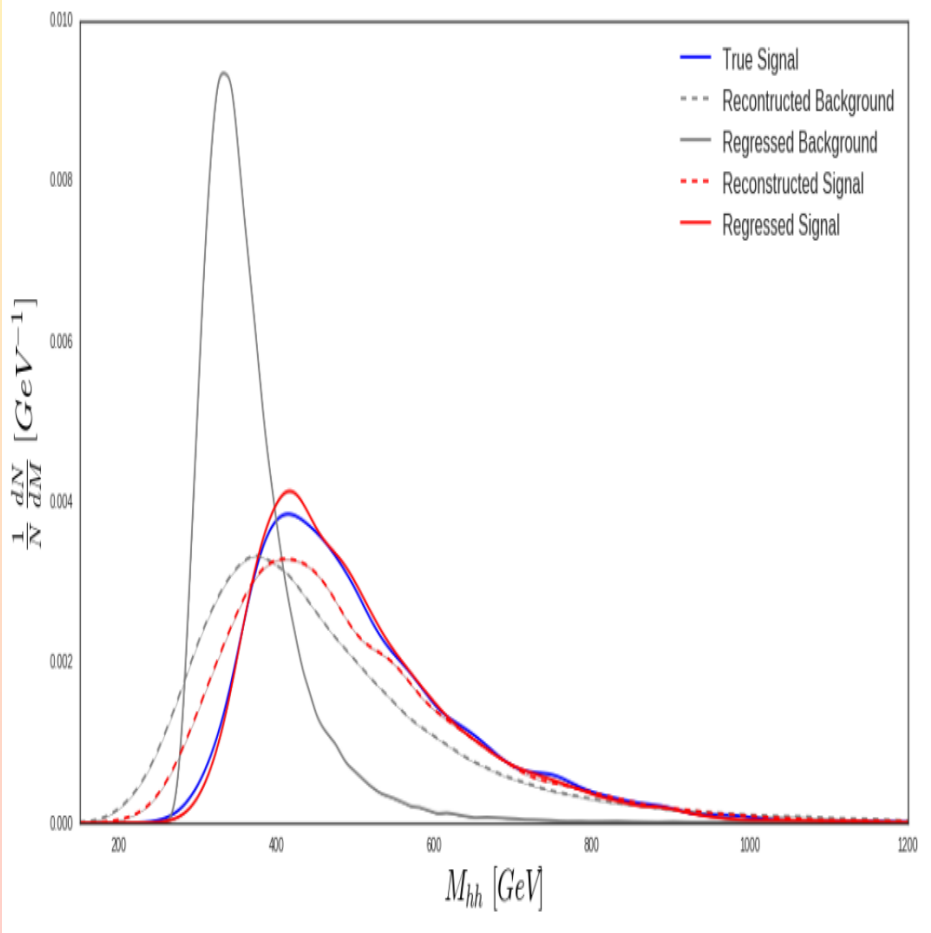


Images from arXiv:1805.12244 and arXiv:1805.00020

Machine Learning: a general definition

- *Vast amounts of data are being generated in many fields, and the statistician's job is to make sense of it all: to extract important patterns and trends, and understand "what the data says." We call this learning from data.* (Hastie, Tibshirani, Friedman, Springer2017)
  - Classification into categories
  - Regression of physical observables
- Well-defined mathematical problems
- Well-defined validation procedures

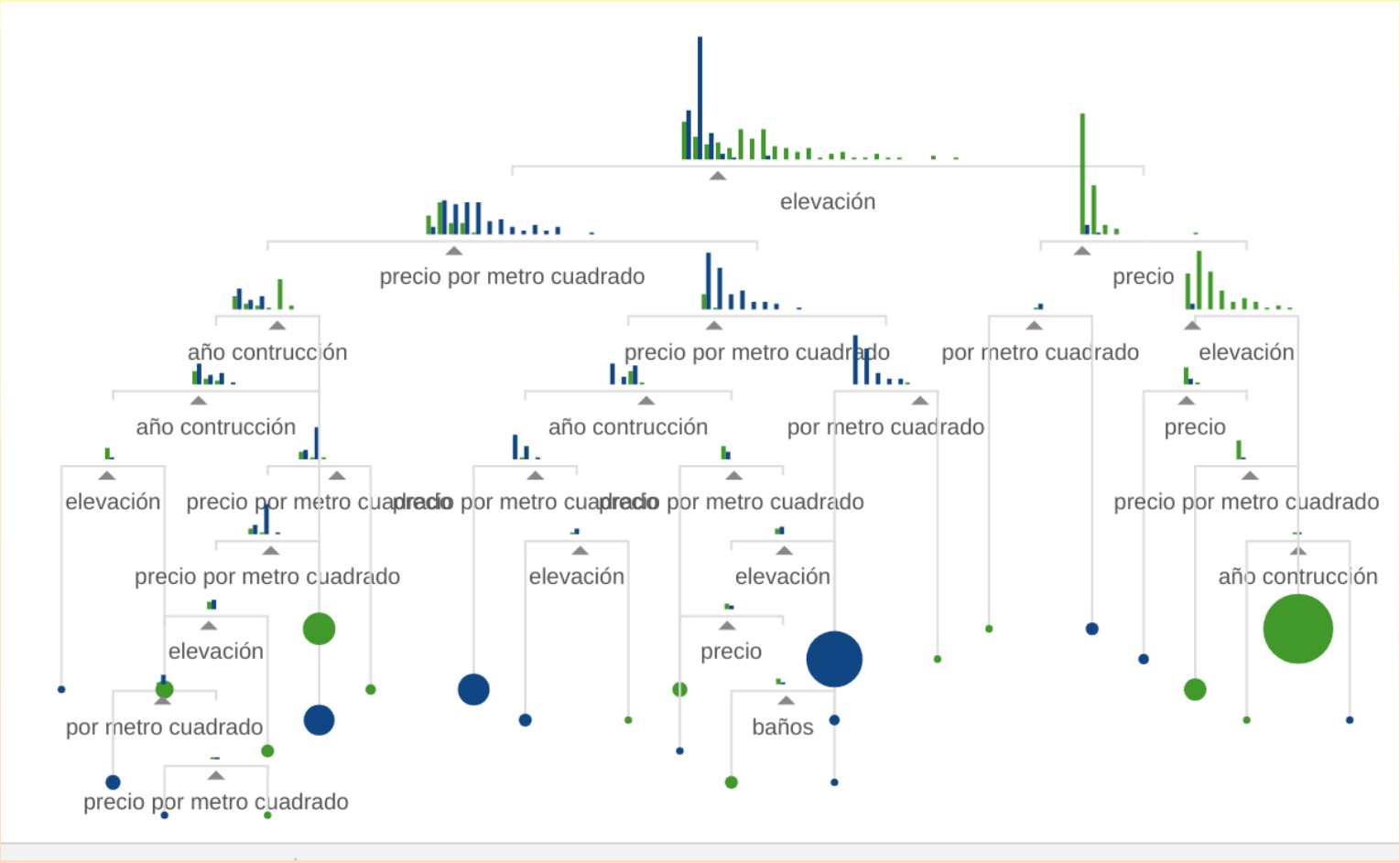




Figures from Hastie, Tibshirani, Friedman, Springer 2017, and from AMVA4NewPhysics deliverable 1.1 public report

Classify events with a decision tree (Decision Tree)

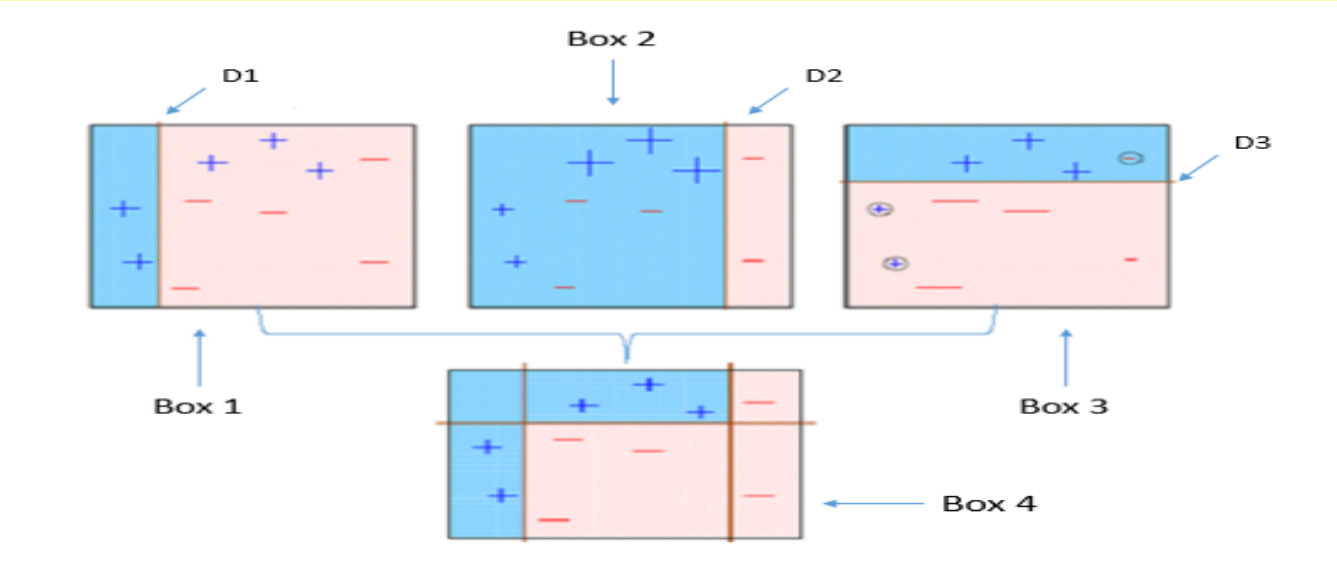
UCLouvain
Institut de recherche
en mathématique et physique



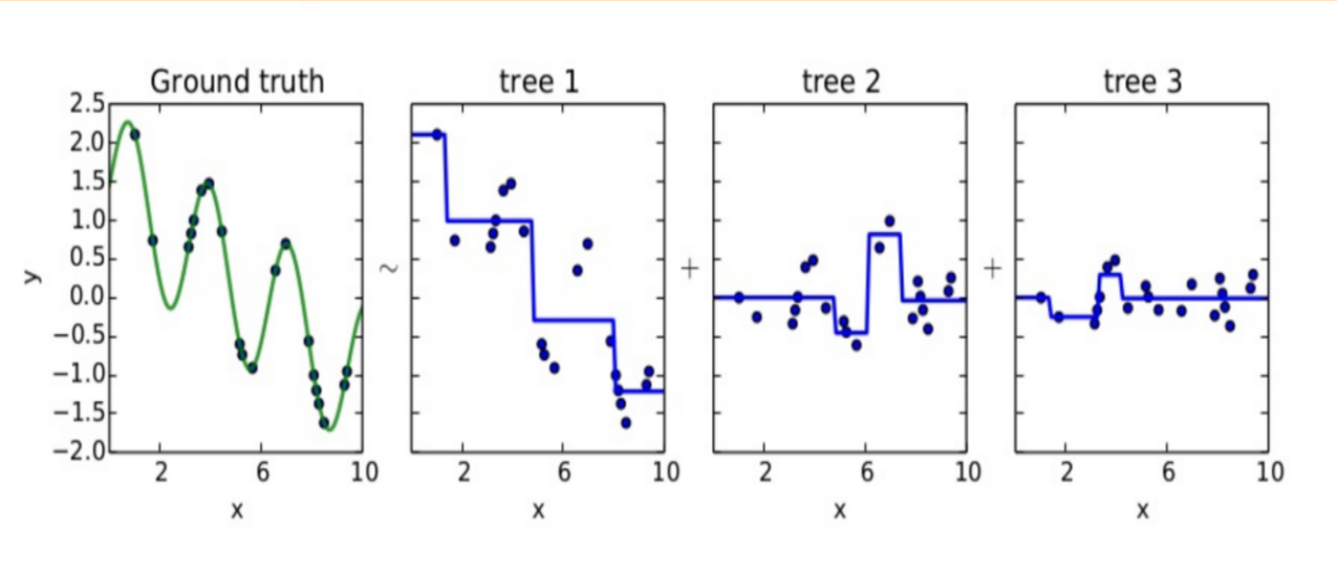
From http://www.r2d3.us/una-introduccion-visual-al-machine-learning-1/

Boosted decision trees

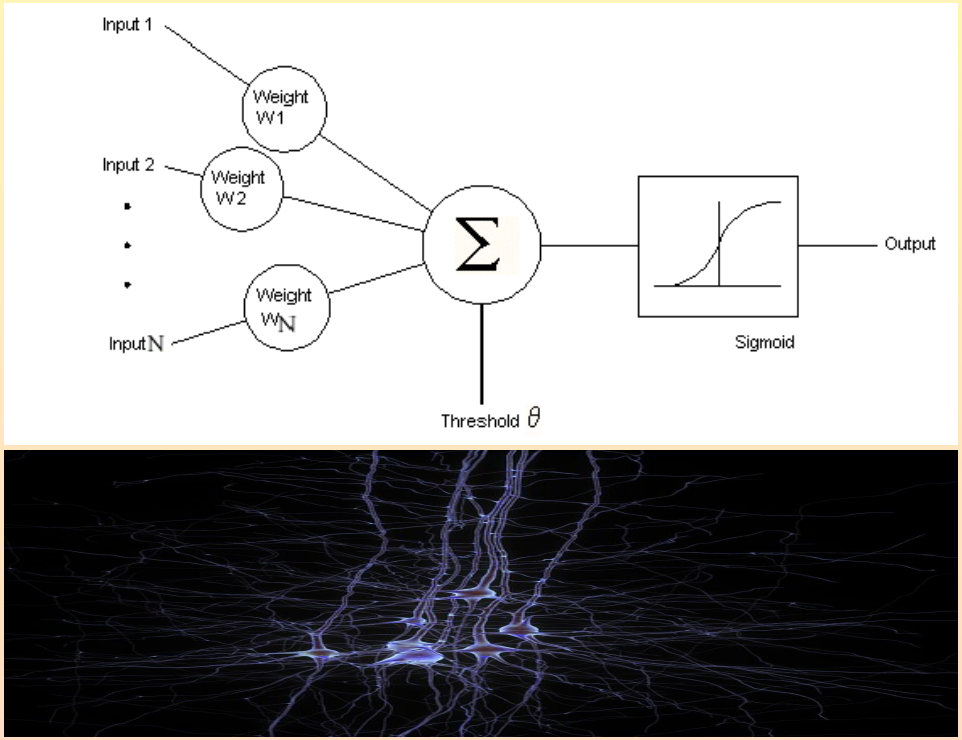
- Ada(ptive)Boost: increase at each iteration the importance of events which were badly-classified at previos iteration



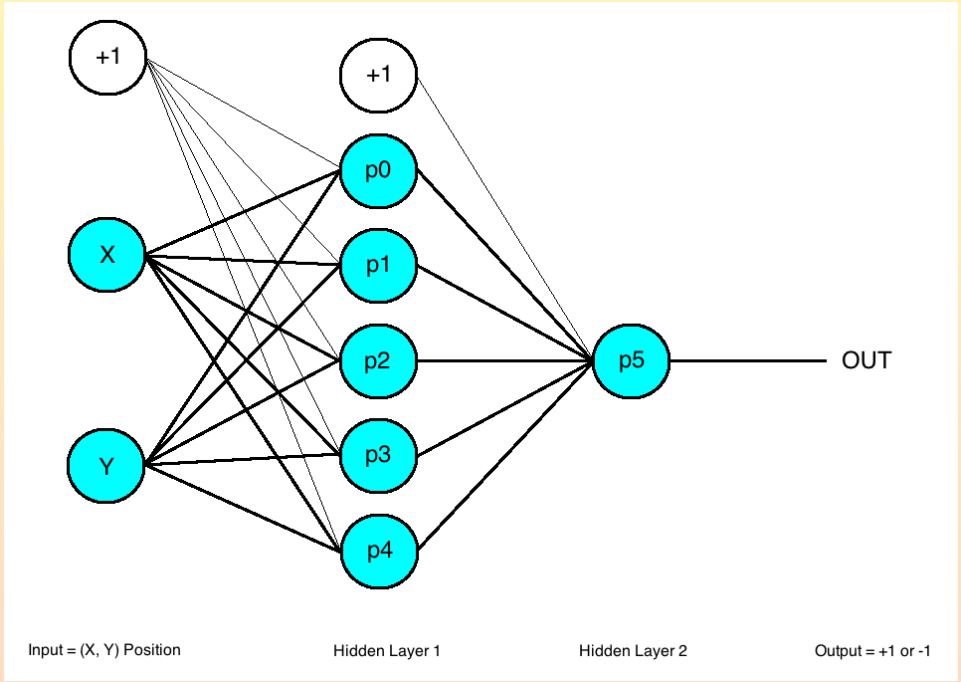
- GradientBoost: fit the new predictor to the residual errors of the previous one



Neural networks...

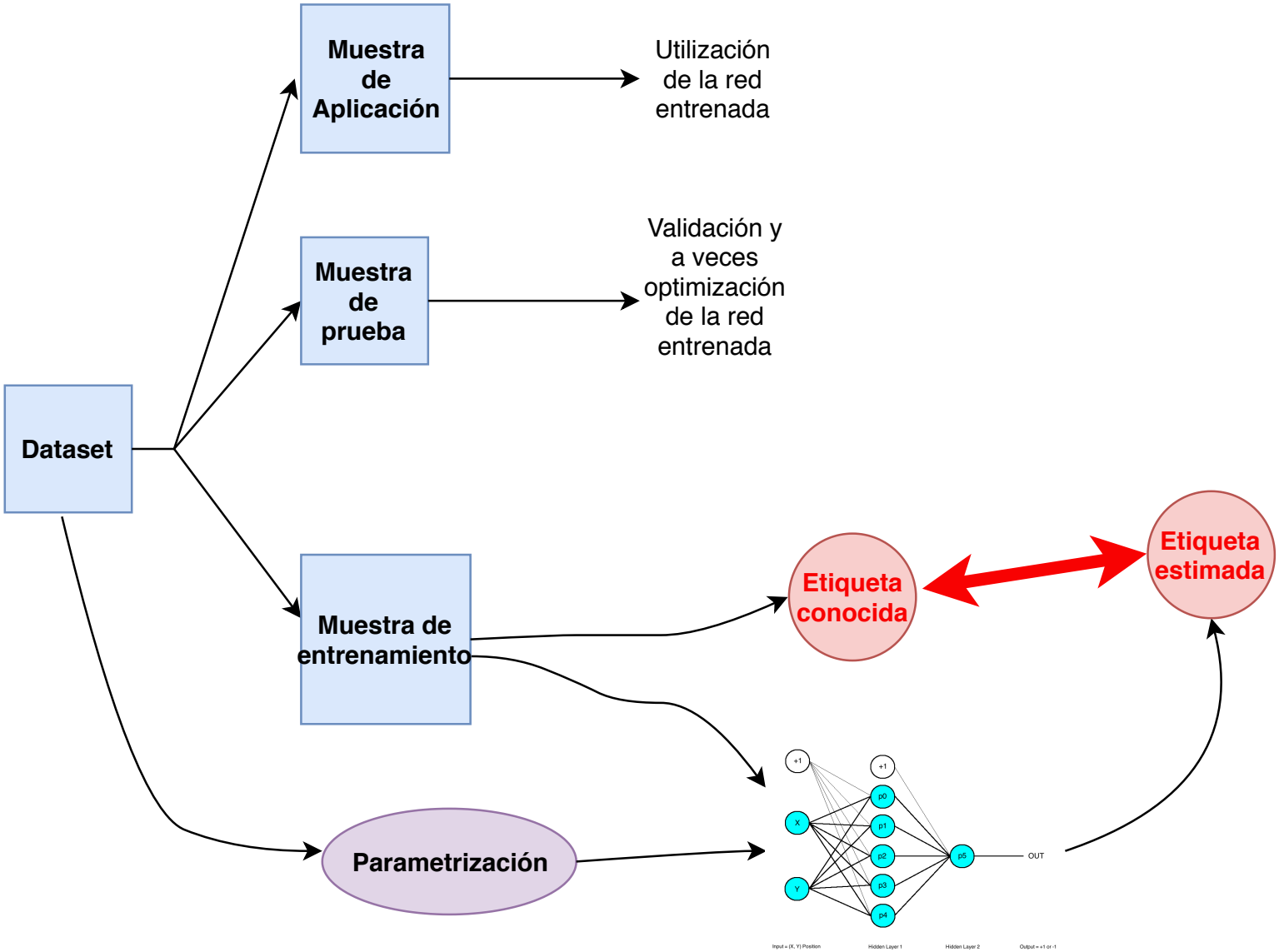


From http://homepages.gold.ac.uk/nikolaev/perceptr.gif and https://i.pinimg.com/originals/e3/fa/f5/e3faf5e2a977f98db1aa0b191fc1030f.jpg



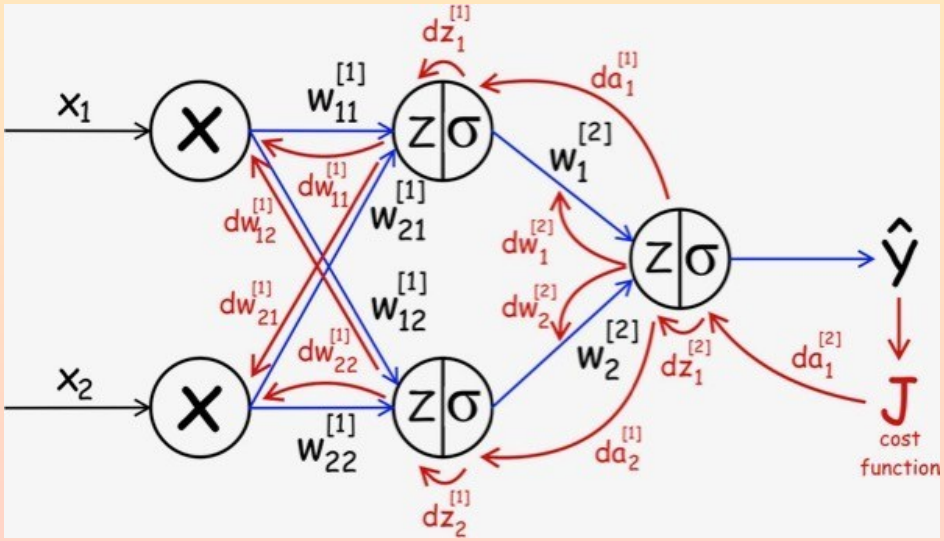
From https://www.cs.utexas.edu/ teammco/misc/mlp/mlp.png

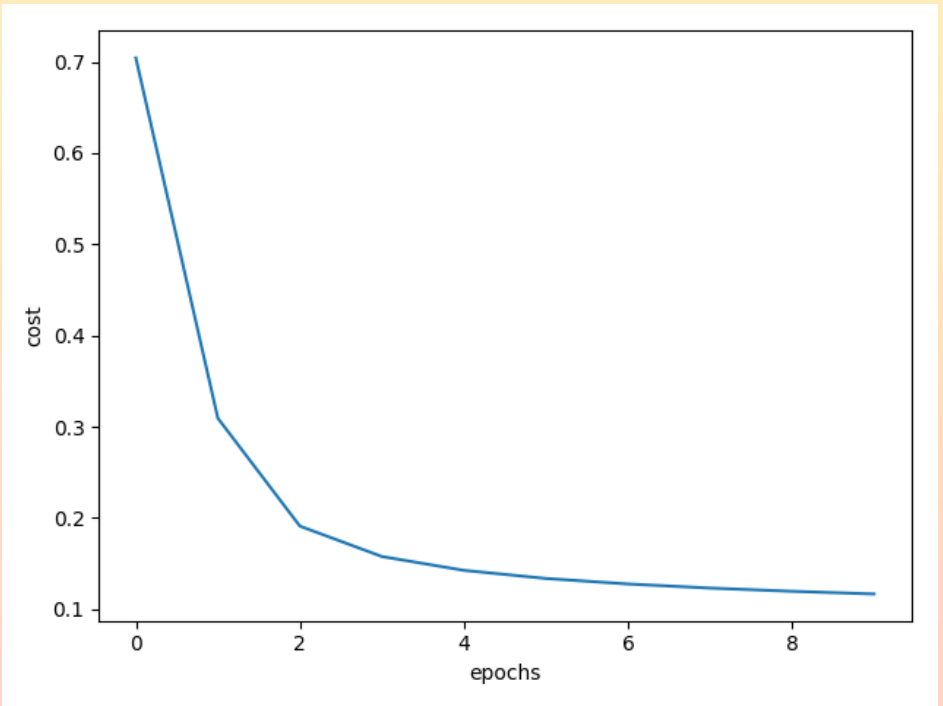
... and how to train them



Loss function and backpropagation

- Adjust the parameters of each neuron and each connetion, back-propagating to the inputs the error in the final classification
- Differentiation and matrix (tensor) calculus; dedicated software, autodifferentiation frameworks (e.g. tensorflow)
- Minimization of a loss function, which can be designed to optimize with different objectives in mind

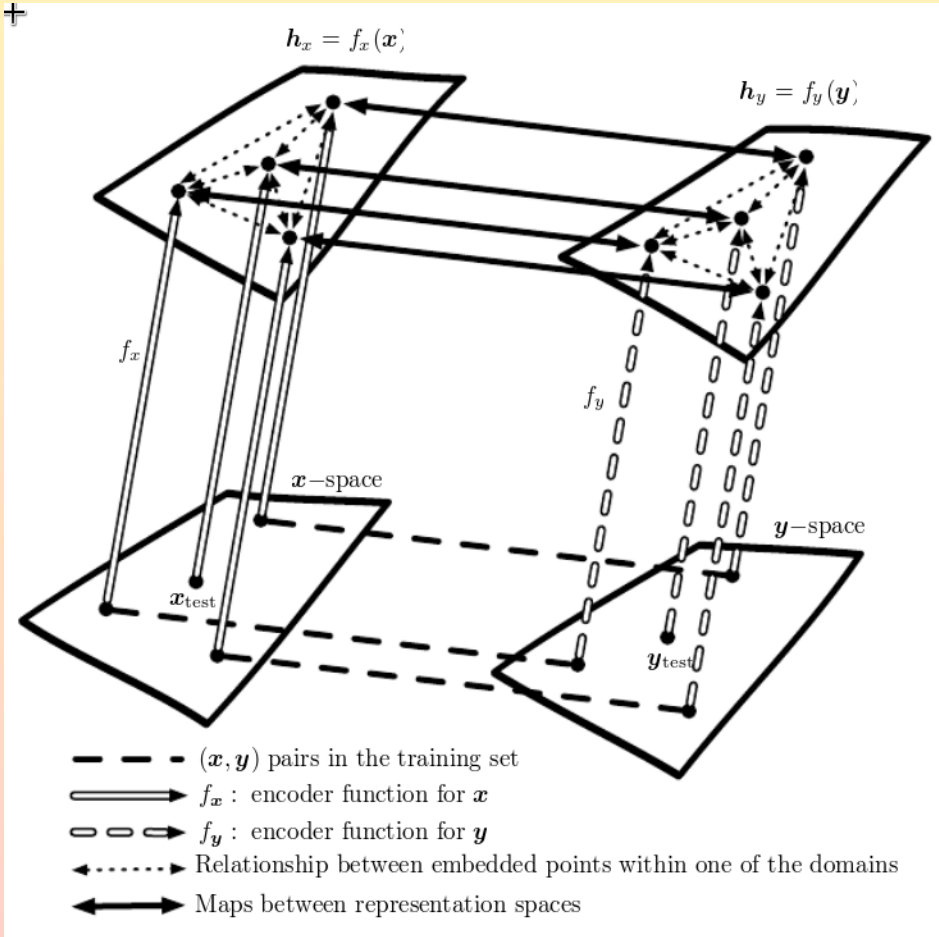


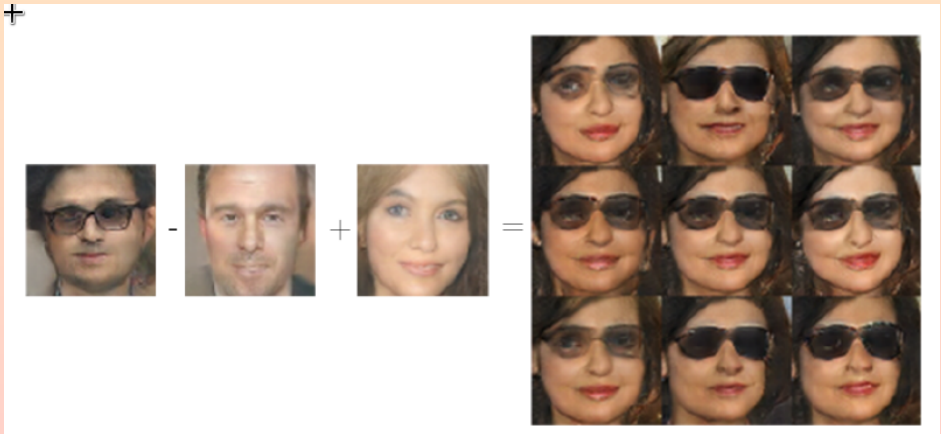


Images from http://www.adeveloperdiary.com

The era of mathematical representations

- Change representation of a problem (metric of the space)
- Sometimes gives access to discriminating power which would be inaccessible (or very difficult to pick up)
- Disentanglement of concepts



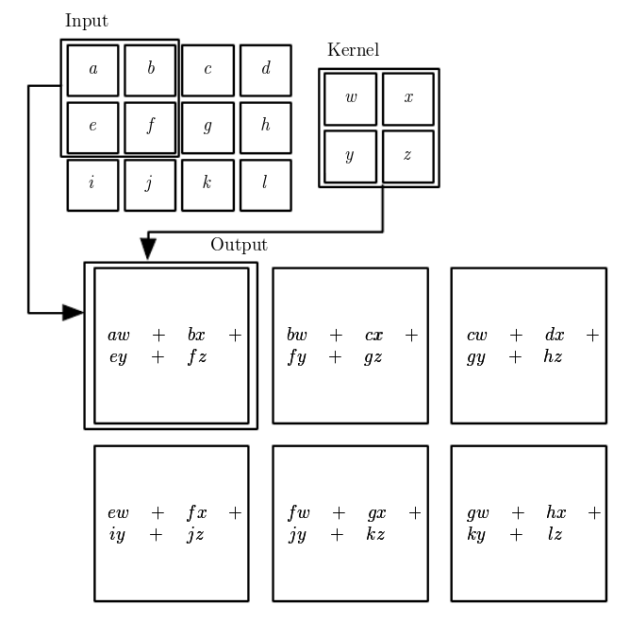


Images from http://www.deeplearningbook.org

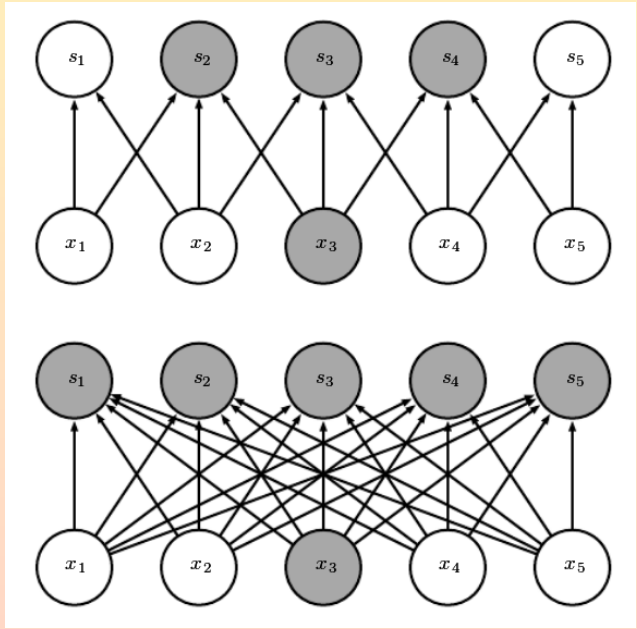
Deep learning and sparse connections

- Sparse connections, parameter-sharing between different portions of the network
- Efficient: less parameters, easier differentiation
- Abstraction of properties (e.g. recognize the same object in different places of the image)

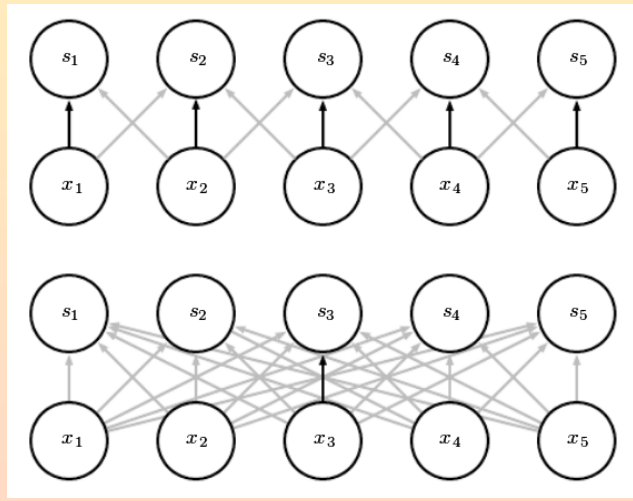
Convolution



Sparse connections



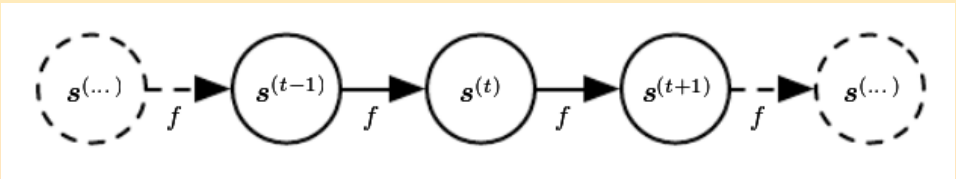
Shared parameters

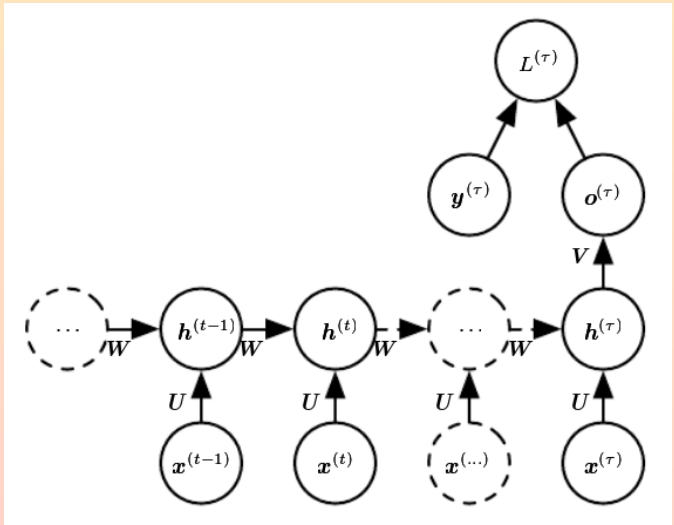


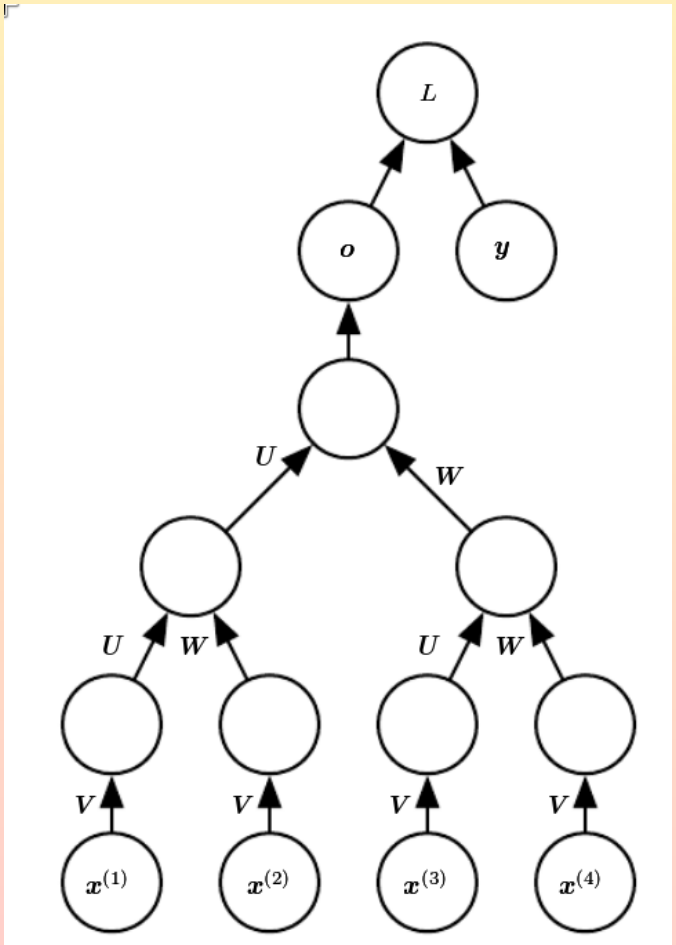
Images from http://www.deeplearningbook.org

Model sequences: recurrent networks

- From recurrent networks...
  - Sequencies of data with a common parameter (e.g. time, for time series)
  - Recognize elements in different places of sequencies with different length (e.g. words in sentences)
  - Varios ways of building networks (one output at each step, or a single final ouput, etc)
- ...to recursive networks
  - Generalization to deep tree
  - Reduced depth, helps identifying long-range dependencies (b/ween distant elements)
  - Applications to data structures processing, language structures, computer vision







Images from http://www.deeplearningbook.org

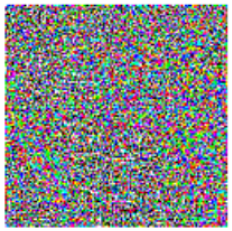
Adversarial networks

- Many networks exhibit human-level performance (e.g. image classification)
- Focus on the badly classified images (to understand if the network has human-level understanding)
- Examples with extremely small differences (indistinguishable for humans) result in misclassification by network 100% of the times!
- Train two networks at the same time, one trying to fool the other
  - Green network: tries to capture the shape of data
  - Blue network: estimates the probability that a point comes from data instead of from the green network
  - Strategy: Green network tries to make Blue network misfunction
    (some people say: Green network is Sporting Lisboa, Blue network is Benfica)

$\boldsymbol{x}$

$y =$"panda"
w/ 57.7%
confidence

$+ .007 \times$



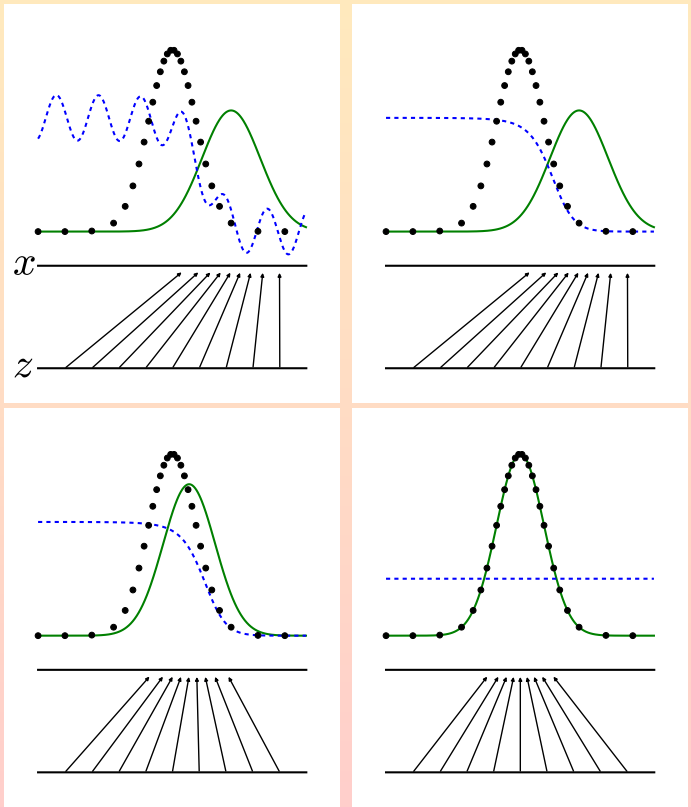
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
w/ 8.2%
confidence

$=$

$\boldsymbol{x} + \epsilon\,\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
w/ 99.3 %
confidence

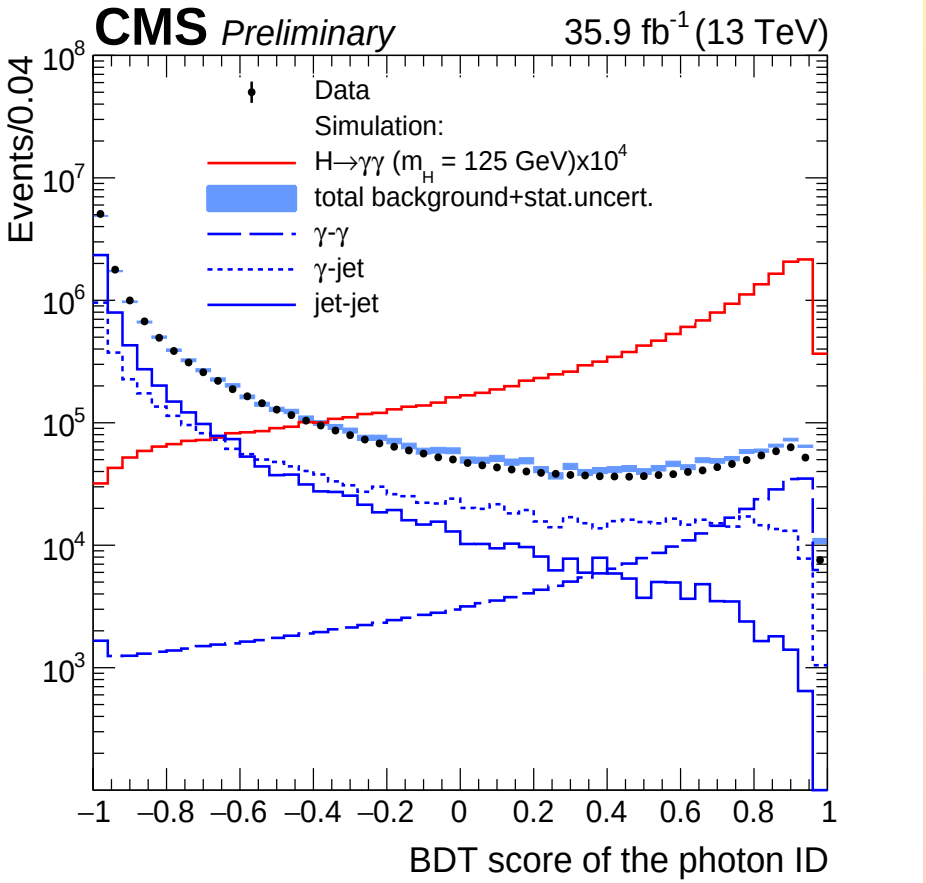


Object ID

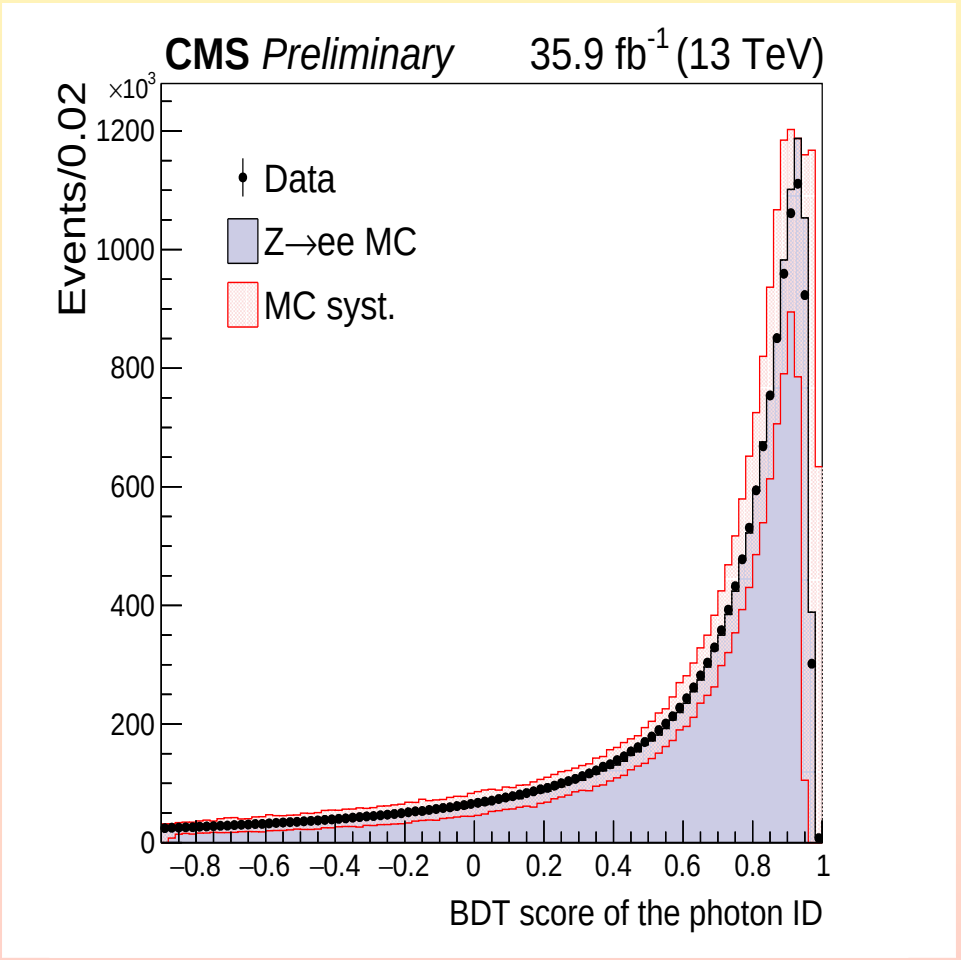
BDTs for object identification: the case of $H \to \gamma\gamma$

- Object identification done with ML techniques since the Higgs discovery
- Classification problem (e.g. real photons vs objects misidentified as photons)

γ identification score for the lowest-score photons



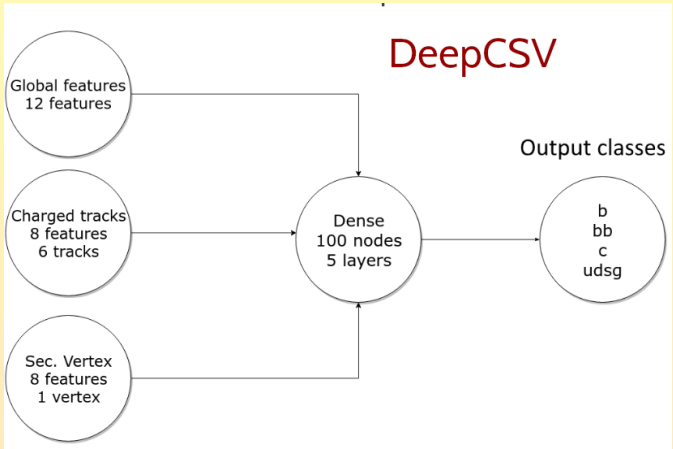
Validation in $Z \to ee$ events

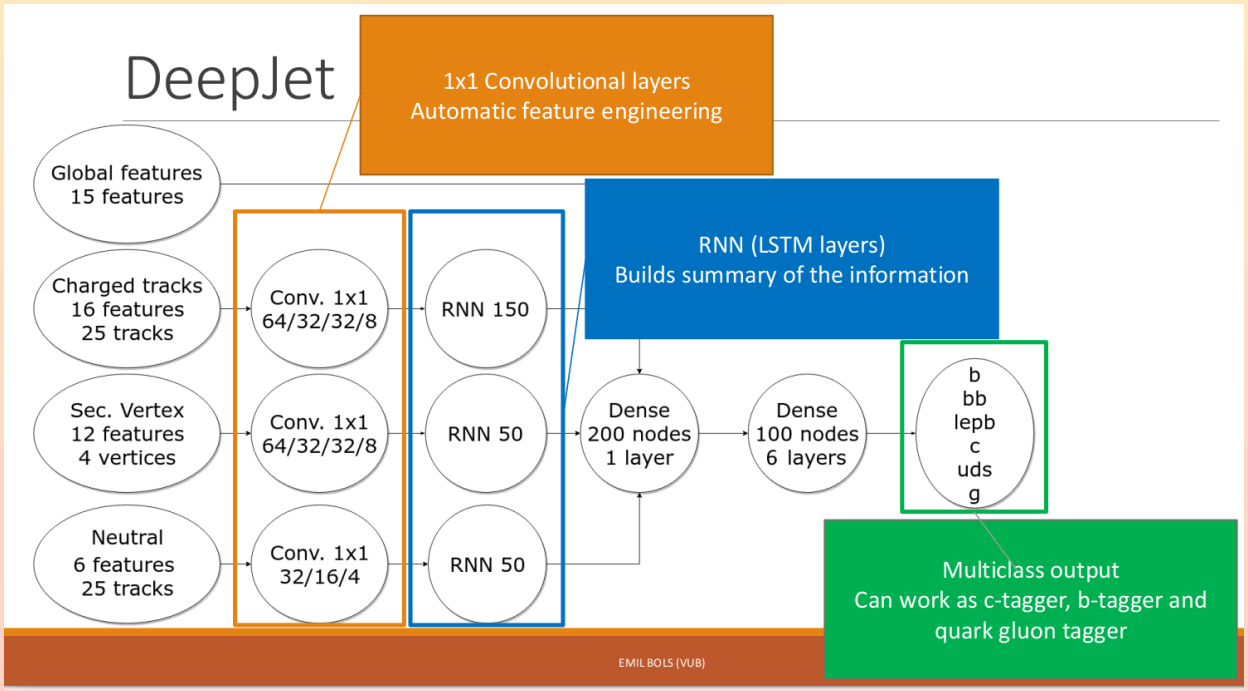


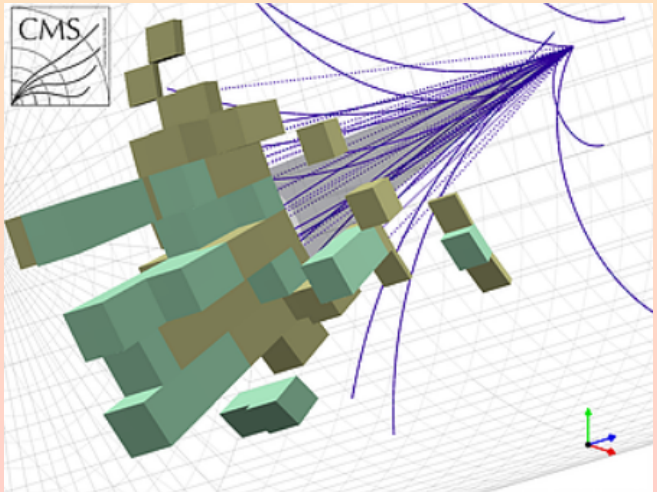
Plots from CMS-PAS-HIG-16-040

Object ID enters the era of mathematical representations — 1

- Identification of jets from bquarks (b tagging) at CMS
  - CSV (Run I and first part of Run II): BDT sensitive to the presence of secondary vertices
- DeepCSV: similar inputs, generic DNN
- Domain knowledge informs the choice of the better mathematical representation
  - Main criterion to choose the classification technique
- What's the best representation for jets?
  - Convolutional networks for images
  - Structure based on individual particles





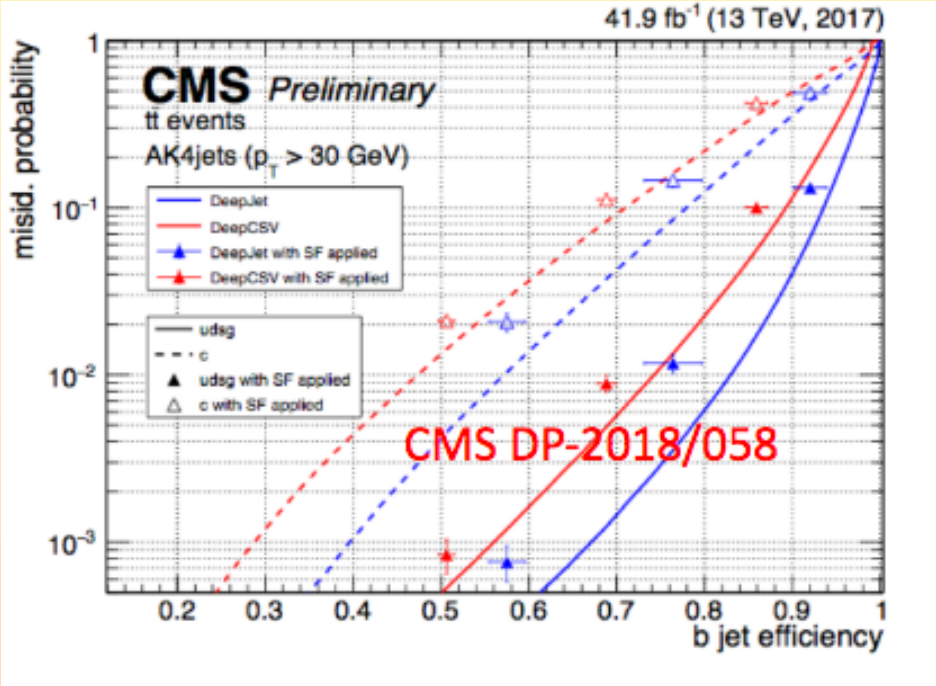


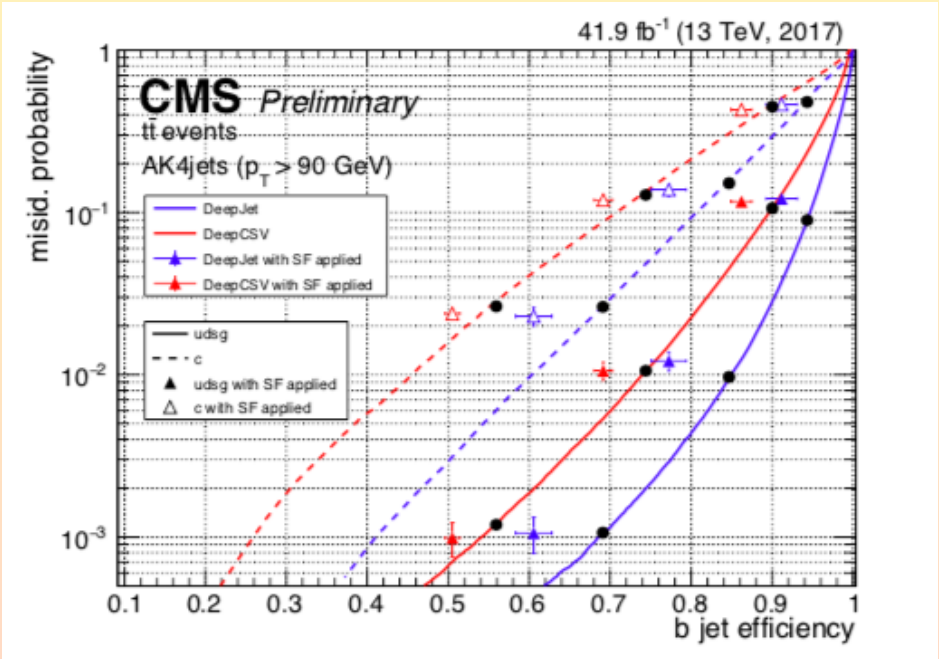
CMS DeepJet, plot from Emil Bols' talk at IML workshop

Object ID enters the era of mathematical representations — 2

UCLouvain
Institut de recherche
en mathématique et physique

- Clear gain even with respect to using a generic DNN (DeepCSV)



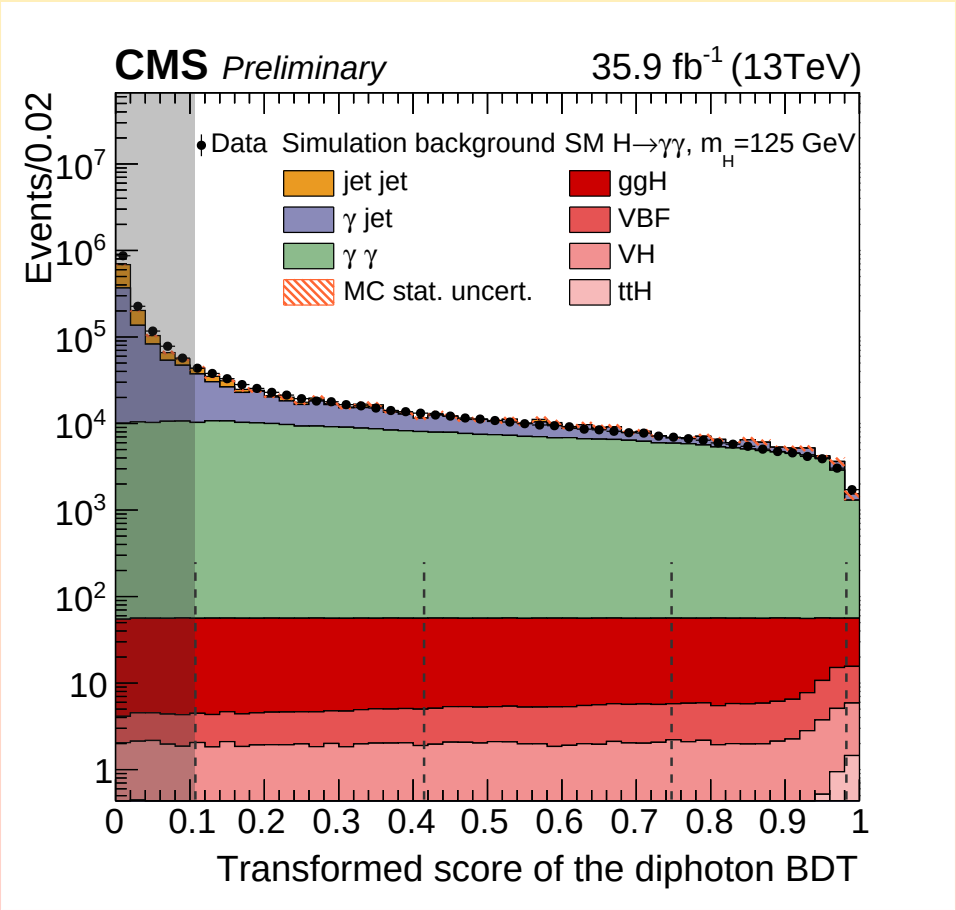


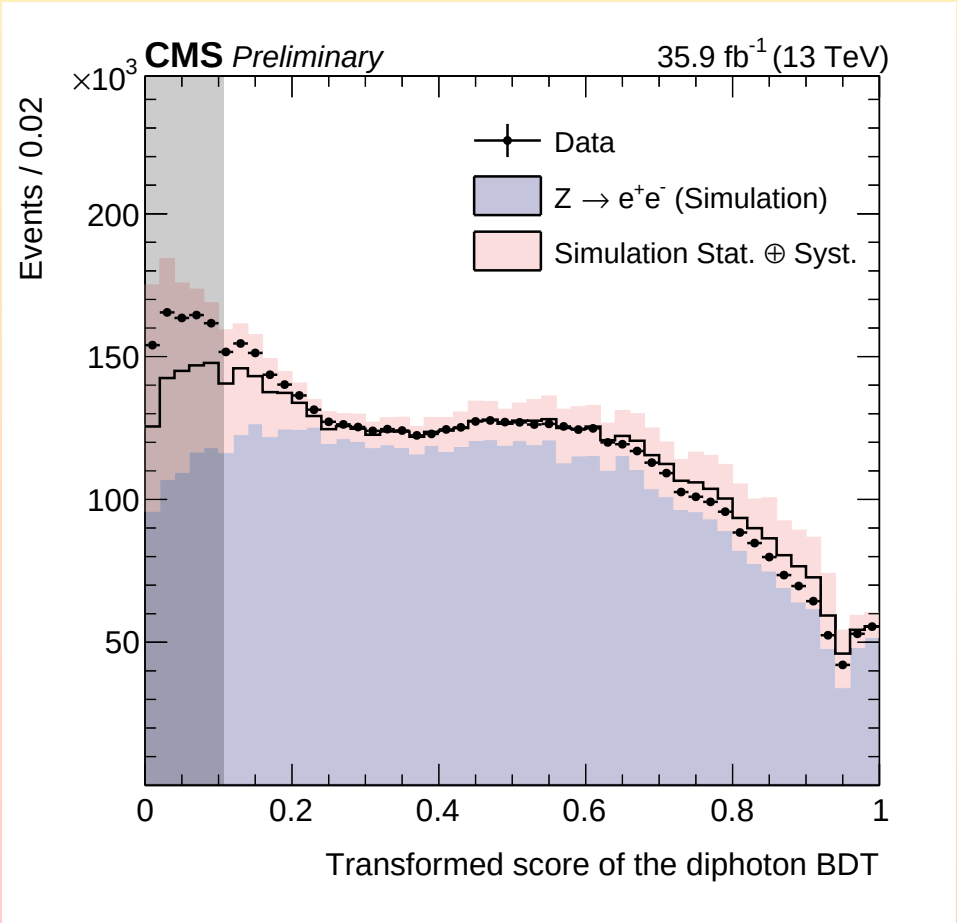
CMS DeepJet

Combining MVA ID for object identification

UCLouvain
Institut de recherche en mathématique et physique

- Dedicated BDT, one score for each event, representing the mass resolution of the diphoton system
  - The photon ID BDT output is used as an input
  - High score for diphoton pairs with kinematic properties similar to signal, good mass resolution, and high individual γ ID score
- Validated in $Z \to ee$ events where electrons are reconstructed as photons

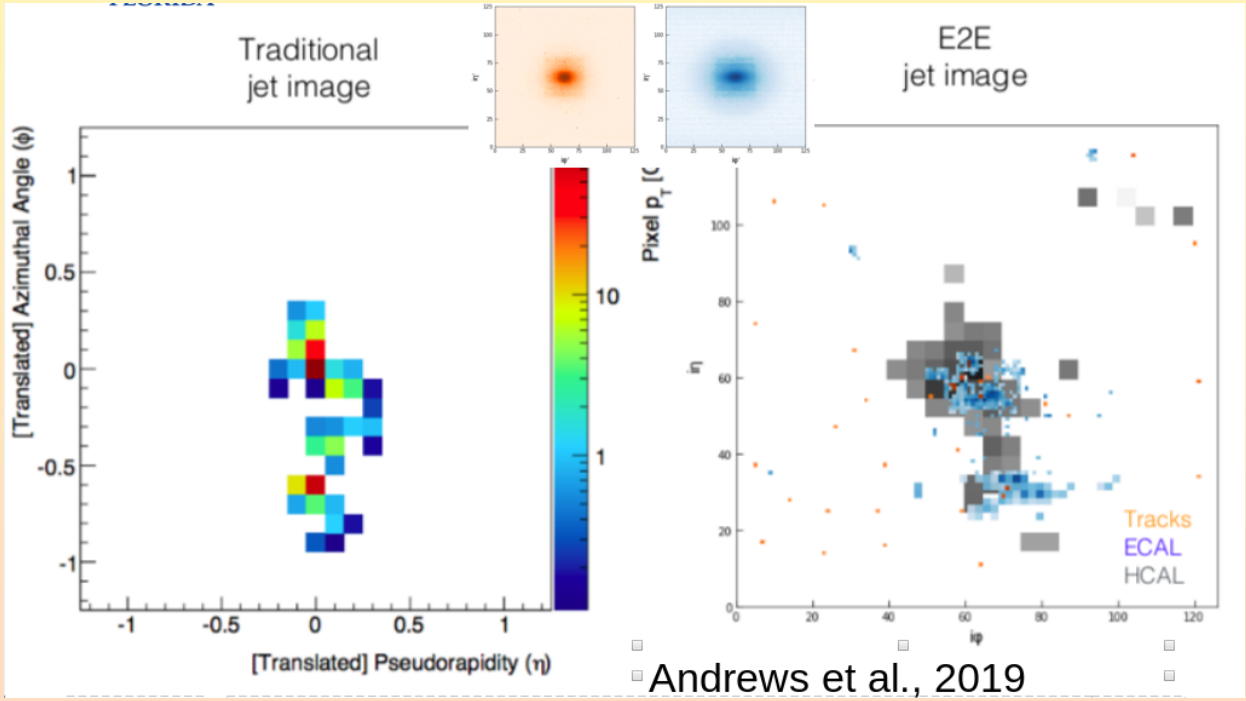




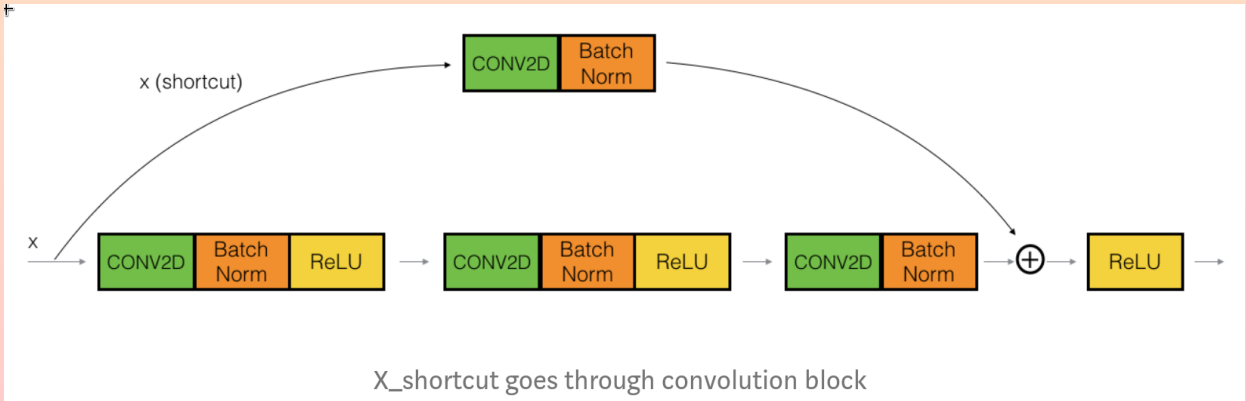
Plots from CMS-PAS-HIG-16-040

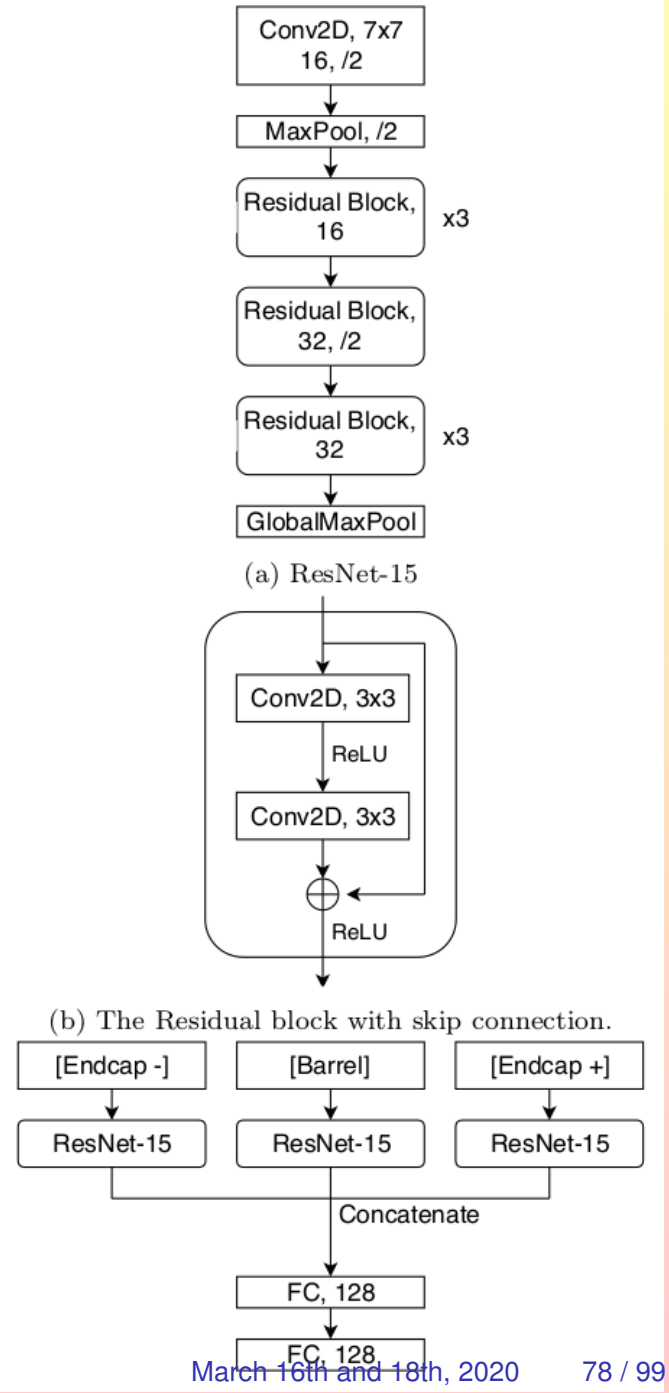
End-to-end reconstruction of jets

- Project detector layers in a single map
- Treat as an image: Res(idual)Net(works)
- Role of tracks in the reconstruction by the network is the same as we expect from the physics we know



Andrews et al., 2019





arXiv:1902.08276, S. Gleyzer's talk at 3rd IML workshop, Priya Dwivedi

Signal extraction

Separate signal from background using selection cuts

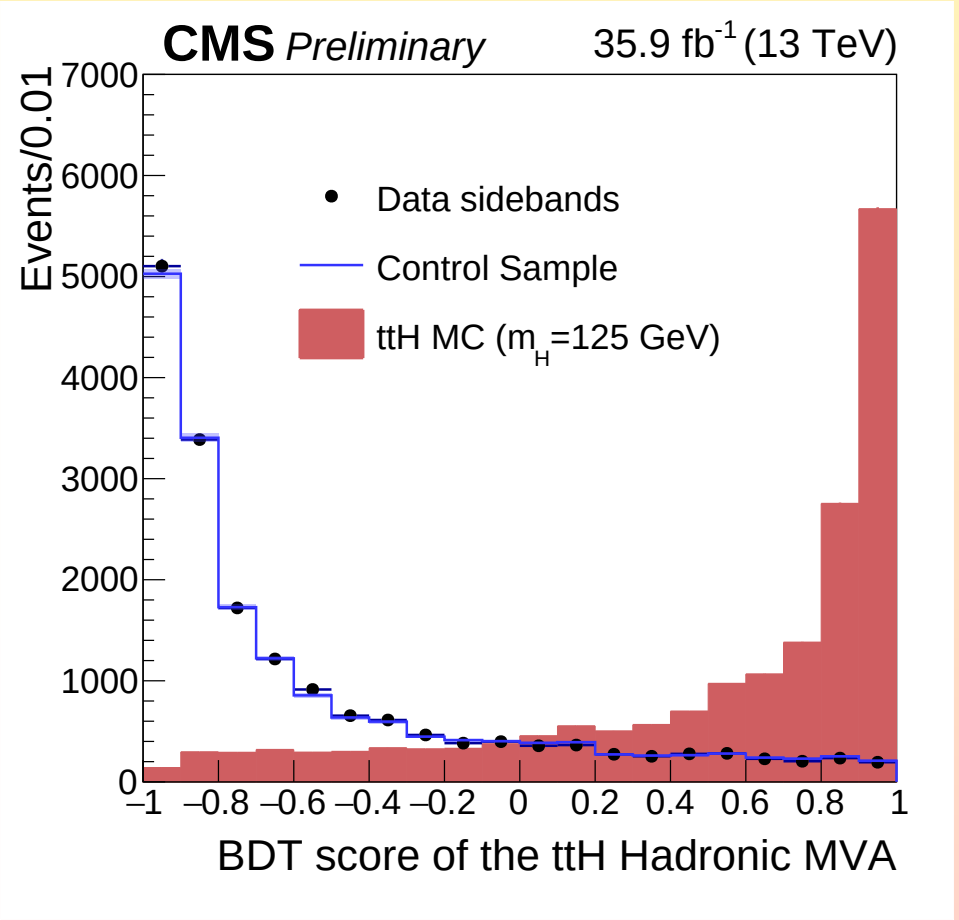
- High fraction of correct events in $t\bar{t}H$ categories by removing events from the dataset
- Delicate: removing events based on MVA output introduces tricky dependency on simulation
  - Dangerous, e.g. prevents from using unfolding results in comparisons with non-SM processes
- In both channels, remove events with low diphoton-BDT score
  - Threshold optimized simultaneously with $\gamma\gamma$-ID score, maximizing expected precision on signal strength

$t\bar{t}H$ leptonic

- ≥ 1 e/μ
- ≥ 2 jets
- ≥ 1 btagged jet

$t\bar{t}H$ hadronic

- ≥ 3 jets
- ≥ 1 btagged jet
- 0 e/μ
- BDT classifier (inputs: N_{jets}, $p_{\mathrm{T}}^{leadjet}$, lead and sublead btag scores)

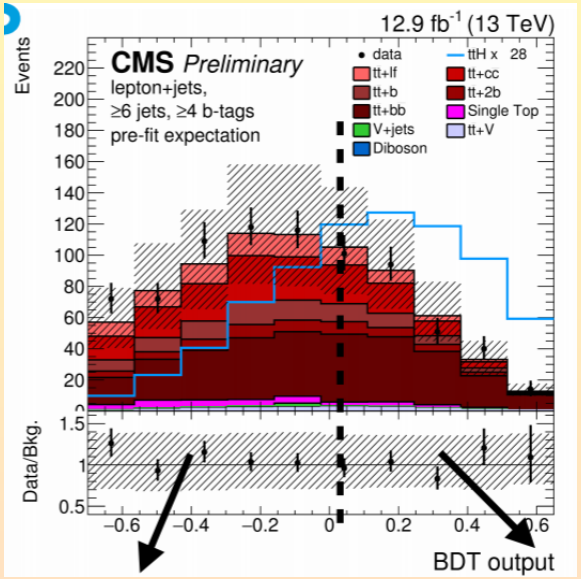


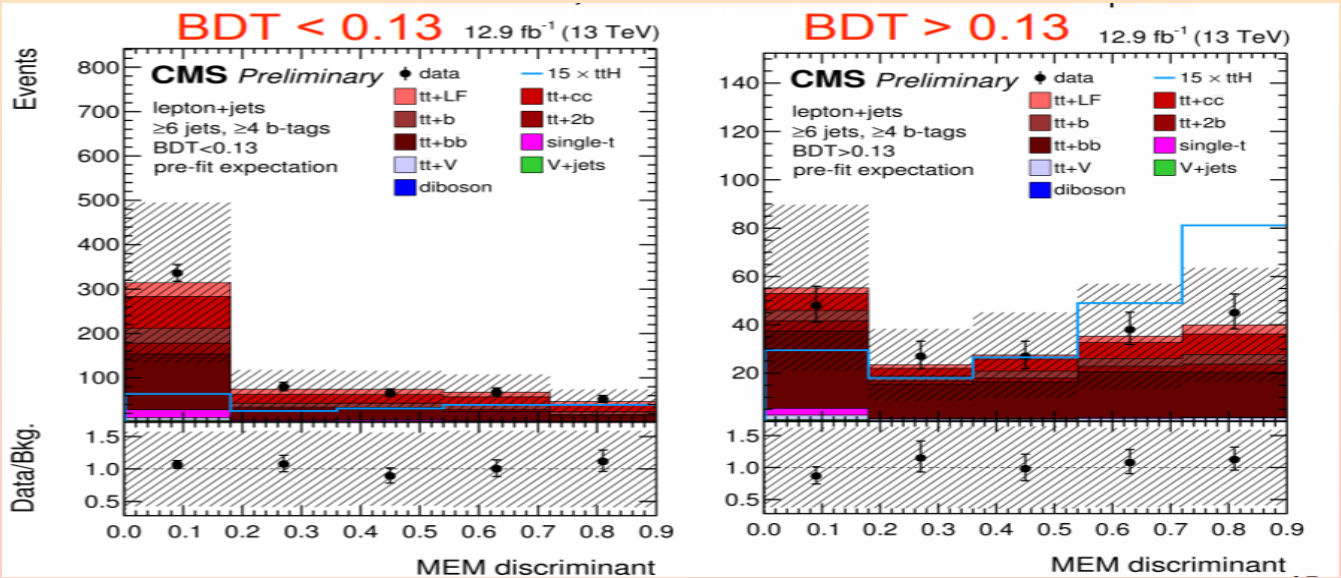
Evt Cat.	SM 125 GeV Higgs boson expected signal										
	Total	ggH	VBF	ttH	bbH	tHq	tHW	WH lep	ZH lep	WH had	ZH had
ttH Had.	5.85	10.99 %	0.70 %	77.54 %	2.02 %	4.13 %	2.02 %	0.09 %	0.05 %	0.63 %	1.82 %
ttH Lep.	3.81	1.90 %	0.05 %	87.48 %	0.08 %	4.73 %	3.04 %	1.53 %	1.15 %	0.02 %	0.02 %

Plots from CMS-PAS-HIG-16-040

Separate signal from background using all events

- Increase sensitivity by keeping the full MVA score distribution, possibly separating it into regions
  - Different fraction of signal/background
  - Constrain normalization or uncertainties in background-dominated regions

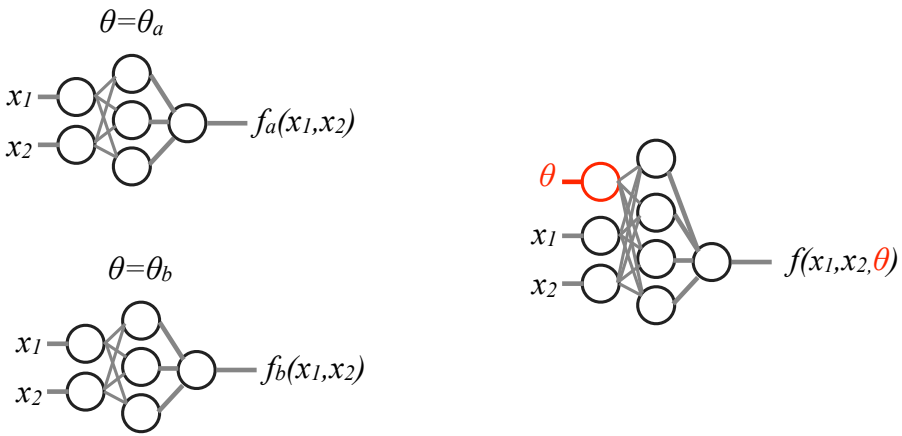


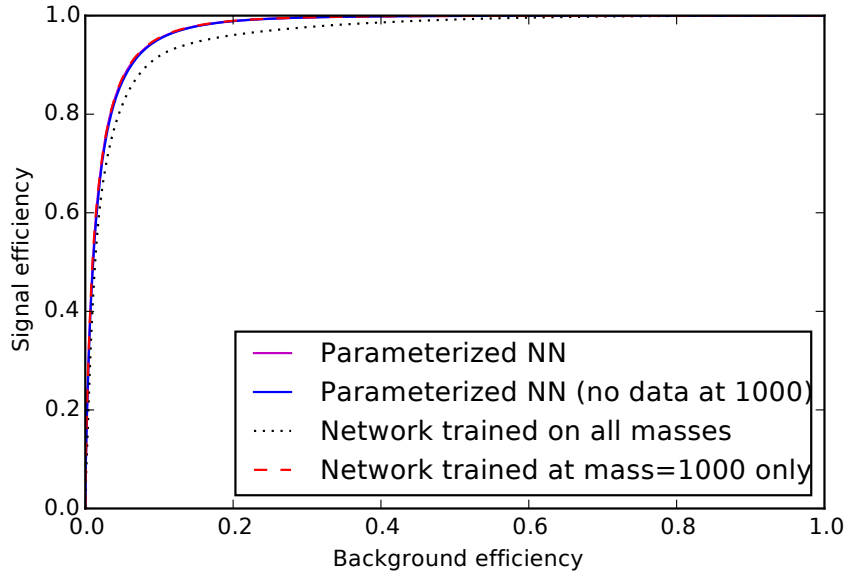


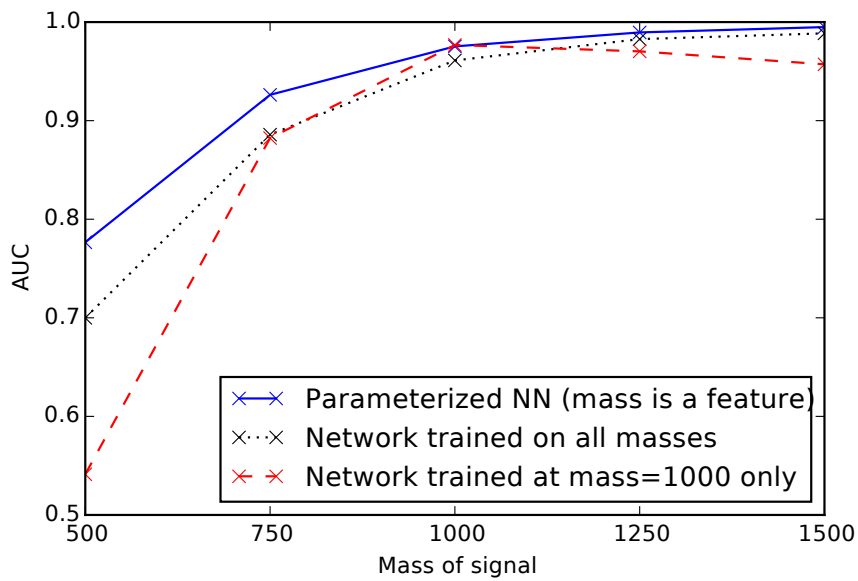
From $t\bar{t}H$ ($b\bar{b}$), CMS-PAS-HIG-16-004

Unknown parameters? Parameterized Machine Learning can help you!

- Classifier sensitive to the value of the parameter
  - Train using as an input the true value of the parameter (signal) or a random value (background)
  - Evaluate in slices at fixed values of the parameter
- Equal or better than training for individual values, and permits interpolation!
- We already use it!!
  - First application in: `CMS-HIG-17-006`
  - Recent application: CMS-HIG-18-004, arXiv:1908.09206 ☺





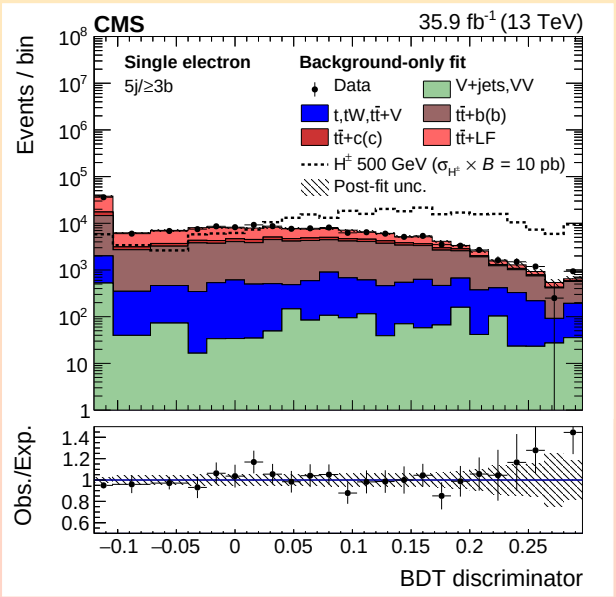


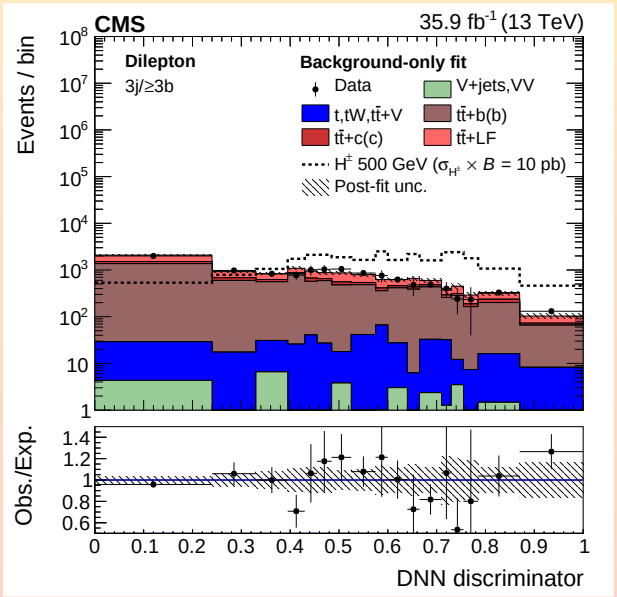
From Baldi *et al.* arXiv:1601.07913

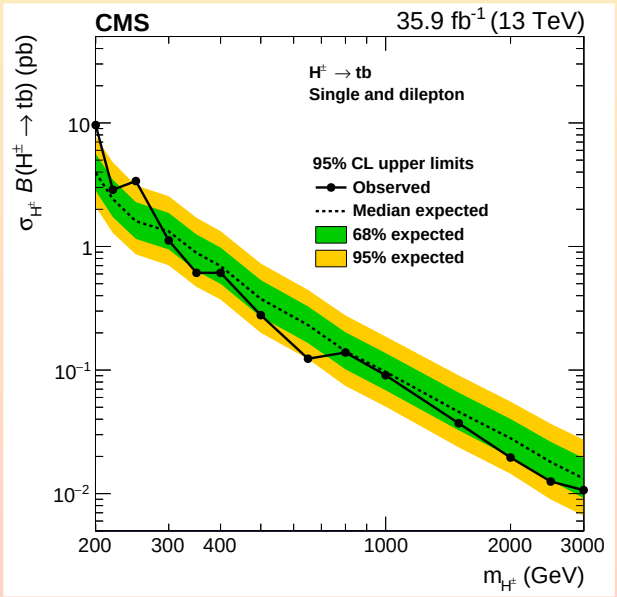
Different techniques are “better” for different situations

UCLouvain
Institut de recherche
en mathématique et physique

- Each classification or regression problem is a distinct problem
  - Choice of the algorithm dictated e.g. by the structure of data and the complexity of the problem (network capacity)
- Sometimes not trivial: CMS-HIG-18-004, arXiv:1908.09206 ☺
  - 20–40% improvement w.r.t. single-variable result (H_T) usando BDT (single lepton) and parameterized DNN (dilepton)
  - DNN: more sensitive at low mass, where the BDT has not enough capacity to discriminate similar topologies ($t\bar{t}$ vs $H^{\pm}$)





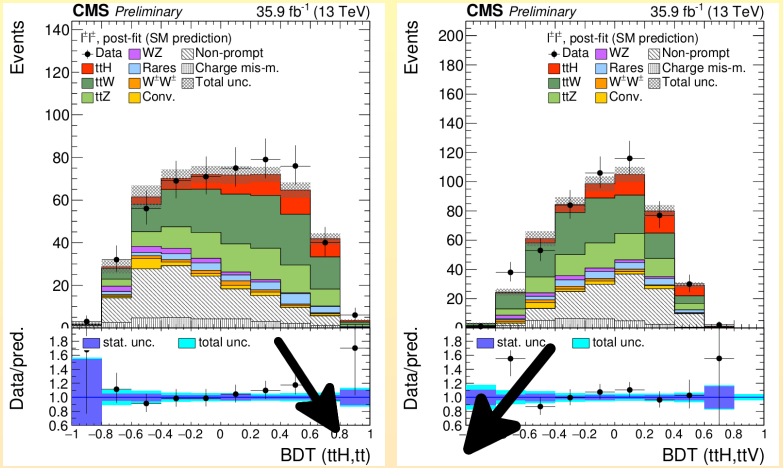


Reduce complexity: how many BDTs do you have?

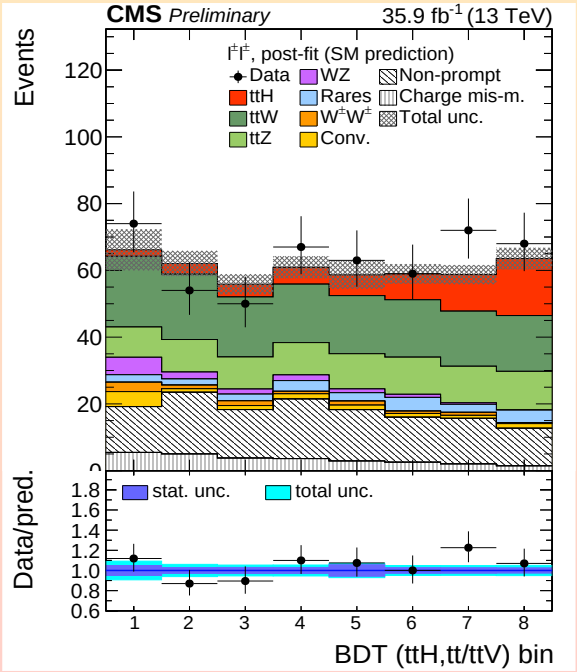
UCLouvain
Institut de recherche en mathématique et physique

- $t\bar{t}H$ multilepton: two different classifiers
  - BDT1: $t\bar{t}H$ vs $t\bar{t}$
  - BDT2: $t\bar{t}H$ vs $t\bar{t}V$
- Finely partition the 2D plane (BDT1, BDT2)
  - Use a training sample to calculate binning
  - Apply to the application sample used for inference
- Define the target N_{bins} with clustering techniques (k-means)
- Finally separate regions based on empirical likelihood
  - Likelihood ratio approximated by $\frac{S}{B}$
  - Ordering from the Neyman-Pearson lemma
  - Quantile-based binning

BDT classifier output (2LSS)



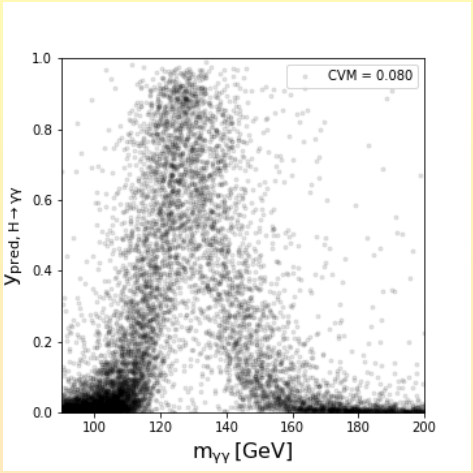
Final 1D discriminator (2LSS)

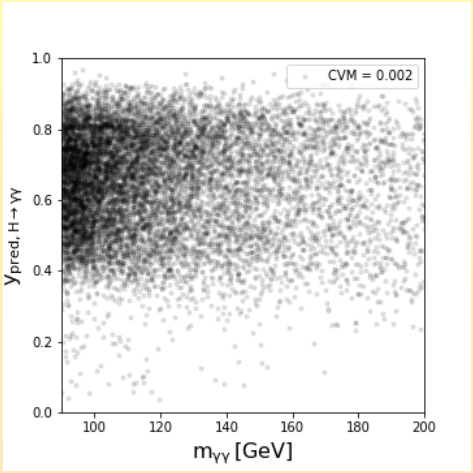


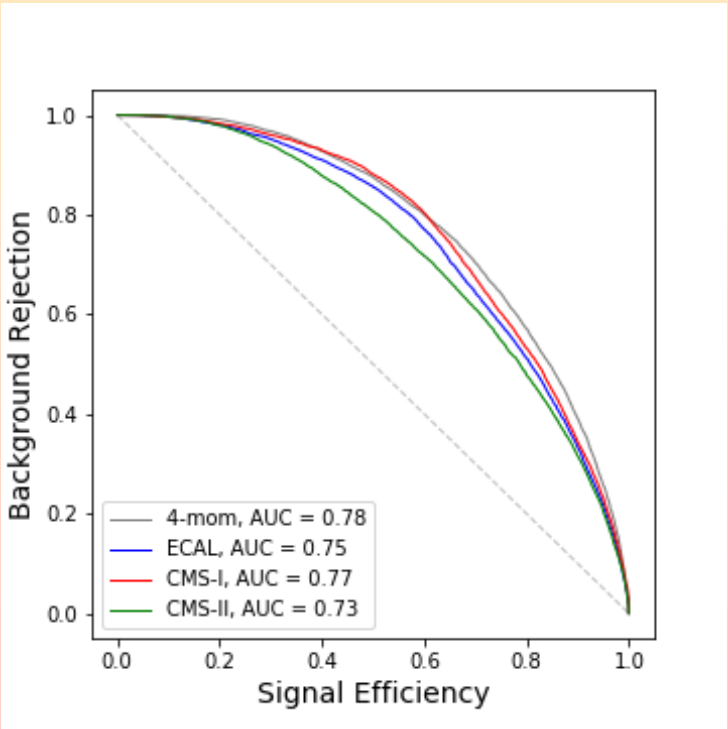
CMS-PAS-HIG-17-004, part of CMS-HIG-17-018: evidence for ttH production in multilepton final states

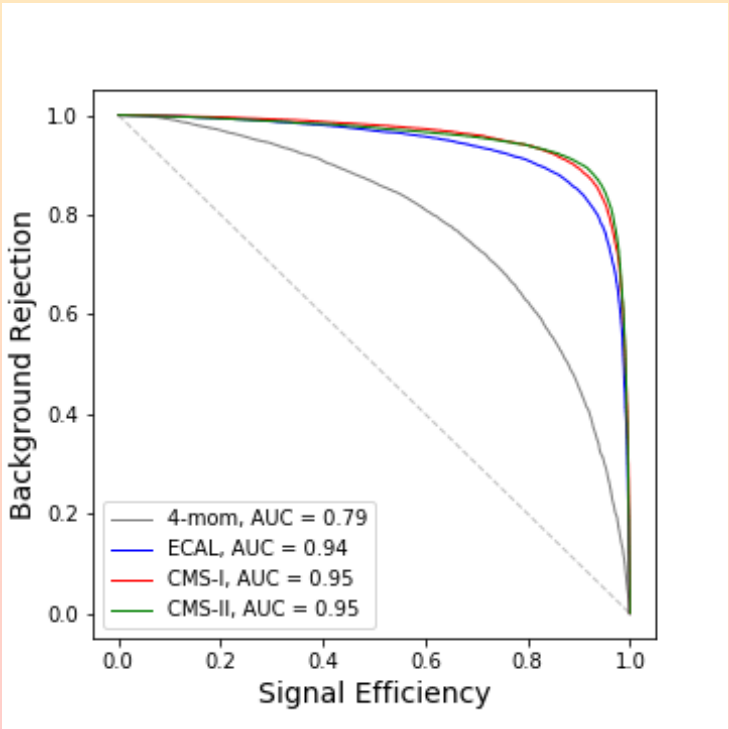
End-to-end event classification

- Low-level data representation
  - Tracker, electromagnetic calorimeter, hadronic calorimeter
  - Various possible geometries
- Mass decorrelation to avoid structure sculpting
  - Transform $E_{\gamma\gamma}$ in units of $M_{\gamma\gamma}$
  - Extension of pivoting technique
- Training with a 3-classes ResNet ($H \to \gamma\gamma$, $\gamma\gamma$, γ+jet)
- Statistically-limited technique









From arXiv:1807.11916

What if you don’t know your signal?

Gaussian processes)

- Multivariate gaussian associated to a set of random variables ($N_{dim} = N_{random\ variables}$)
  - *Kernel* as a similarity measure between bin centers (counts) and a *averaging function*

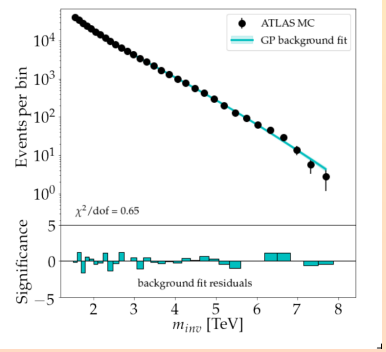
$$\mu(x) = 0\,, \tag{9}$$

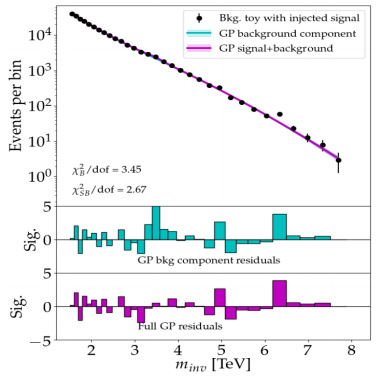
$$\Sigma_B(x,x') = A\ \exp\left(\frac{d-(x+x')}{2a}\right)\sqrt{\frac{2l(x)l(x')}{l(x)^2+l(x')^2}}\exp\left(\frac{-(x-x')^2}{l(x)^2+l(x')^2}\right), \tag{10}$$

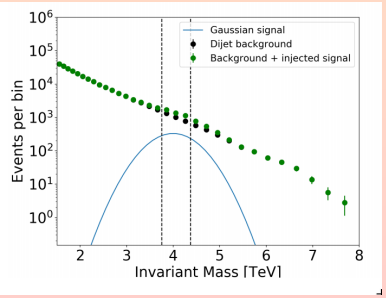
$$\Sigma_S(x,x') = C\ \exp\left(-\frac{1}{2}(x-x')^2/k^2\right)\exp\left(-\frac{1}{2}\left((x-m)^2+(x'-m)^2\right)/t^2\right), \tag{11}$$

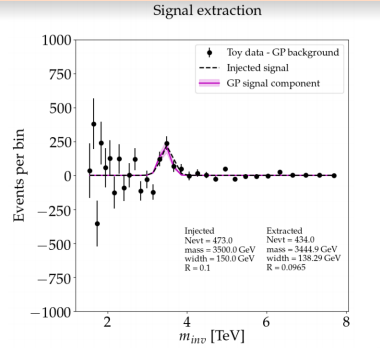
- Signal is not parameterized
- Hyperparameters fixed by the B-only fit

- S: residual of B-subtraction







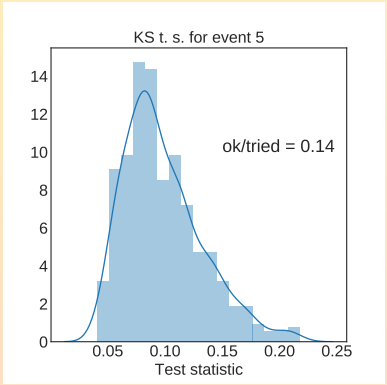


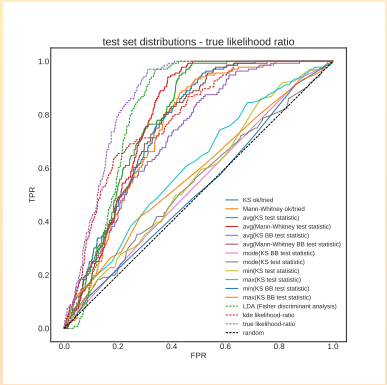
AMVA4NewPhysics deliverable 2.5 public report

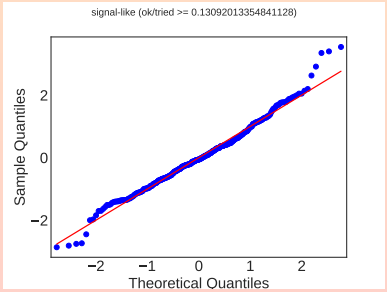
Inverse Bagging

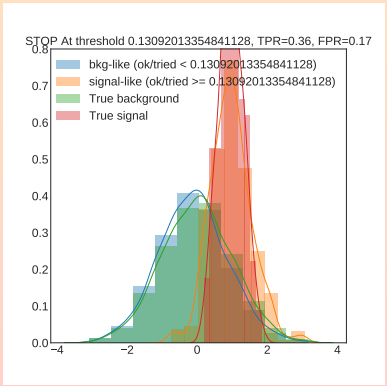
UCLouvain
Institut de recherche en mathématique et physique

- Data: mixture model with small S
- Classification based on sample properties
  - Compare bootstrapped samples with reference (pure B)
  - Use Metodiev theorem to translate inference into signal fraction
- Validate with LR y LDA
  - Promising results









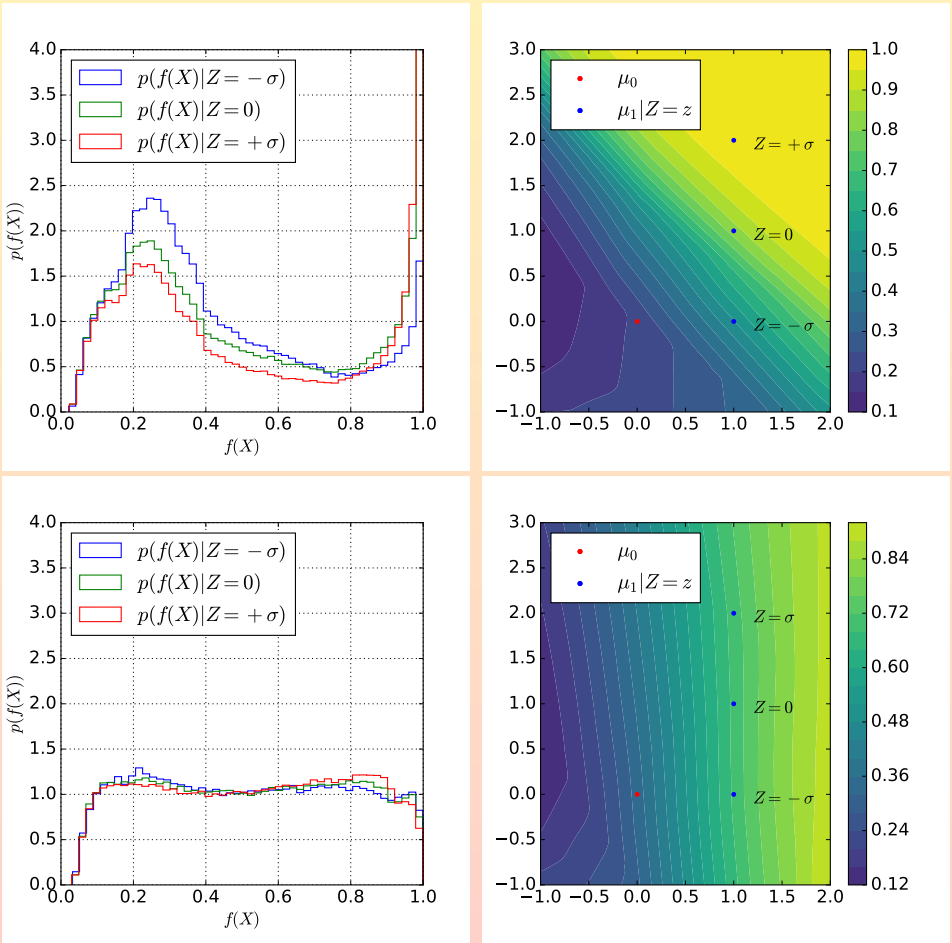
Vischia-Dorigo arXiv:1611.08256, doi:10.1051/epjconf/201713711009, and P. Vischia's talk at EMS2019

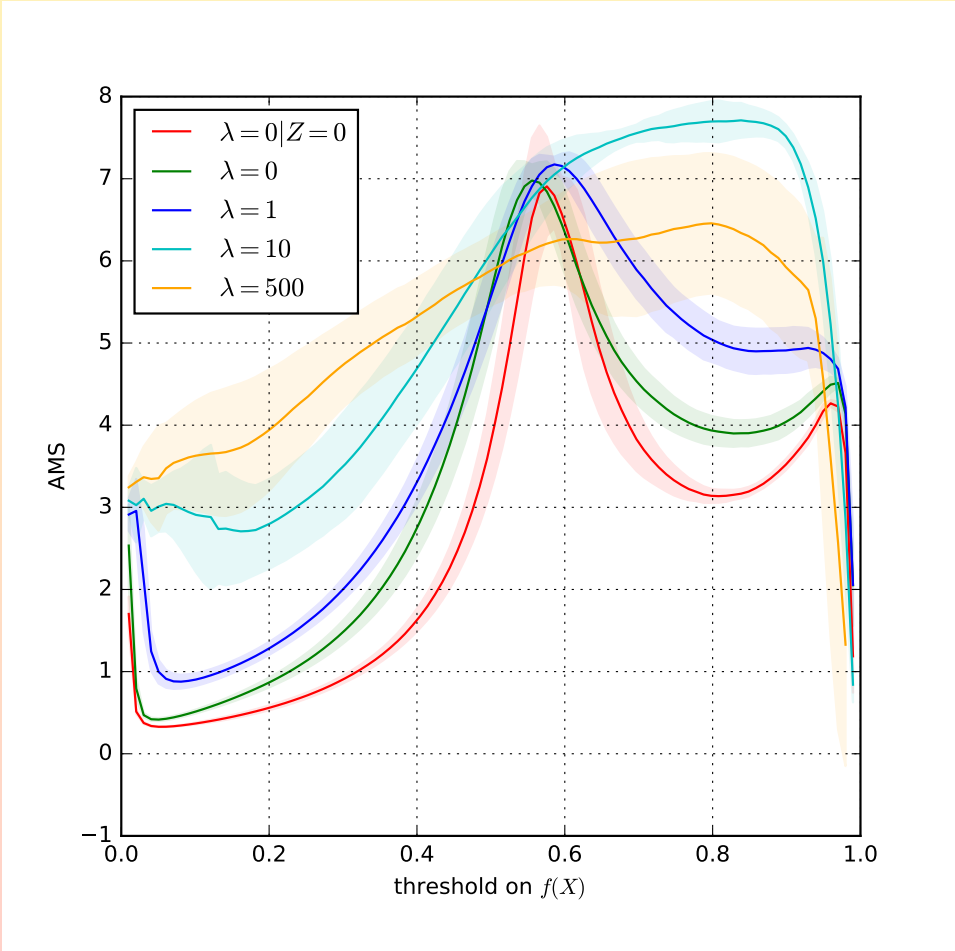
What about the uncertainties?

Can we reduce the impact of uncertainties on our results?

UCLouvain
Institut de recherche
en mathématique et physique

- Adversarial networks used to build pivot quantities
  - Quantities invariant in some parameter (typically nuisance parameter representing an uncertainty)
- Best Approximate Mean Significance as tradeoff optimal/pivotal
  $E_\lambda(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) - \lambda \mathcal{L}_r(\theta_f, \theta_r)$





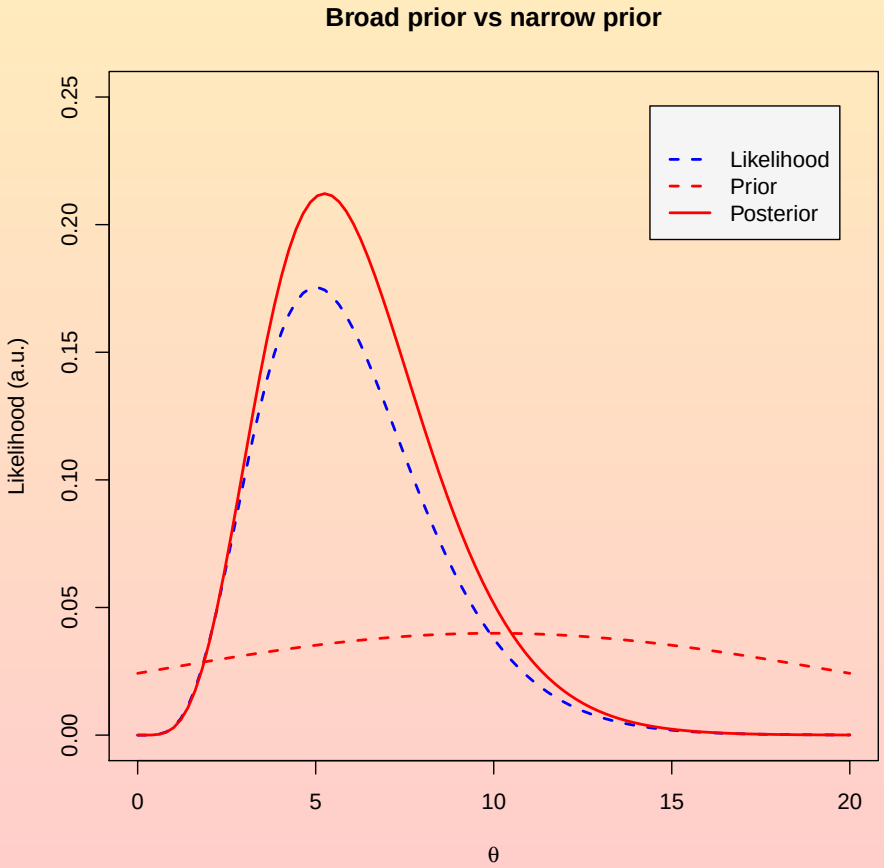
From Louppe-Kagan-Cranmer, arXiv:1611.01046

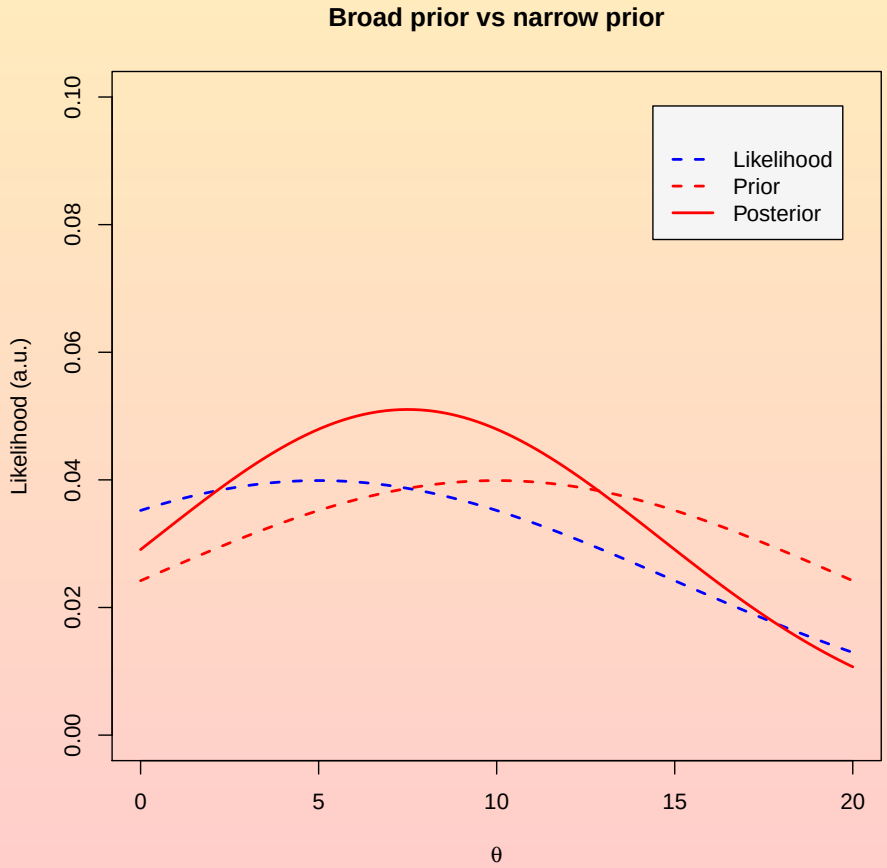
Reminder: likelihood function and Fisher information

- The (second) derivative of the likelihood function is connected to the quantity of information you can extract from data

$$I(\theta) = -E\left[\frac{\partial^2 lnL}{\partial\theta^2}\right] = E\left[\left(\frac{\partial lnL}{\partial\theta}\right)^2\right]$$

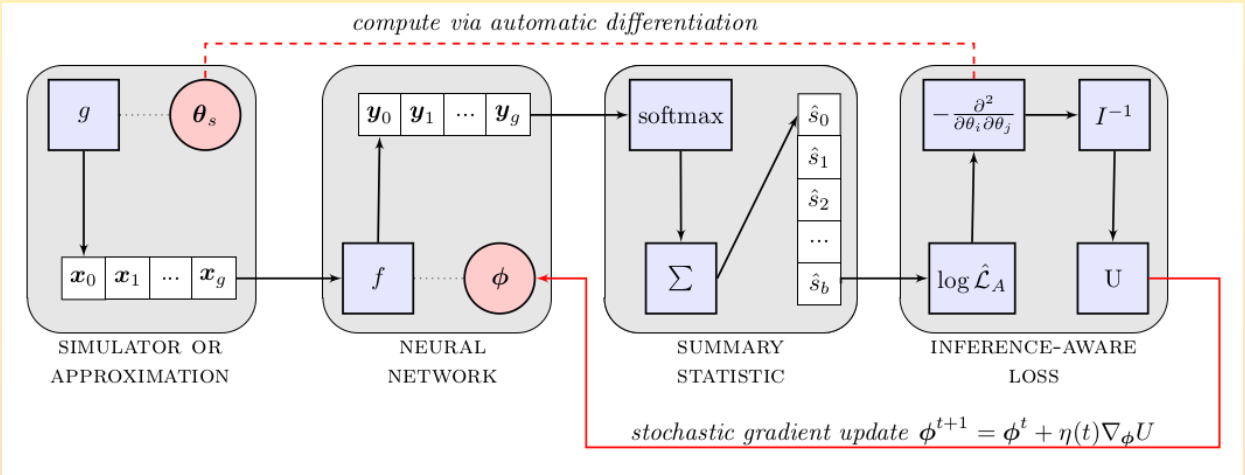
- The likelihood function contains all the information that you can extract from data on the parameter θ
- A narrow likelihood function carries more information than a broader one

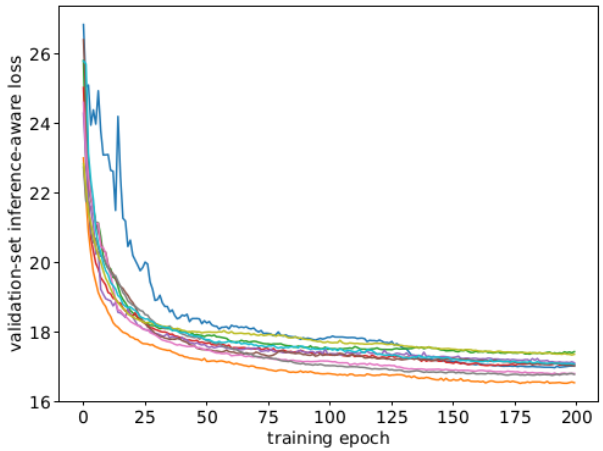




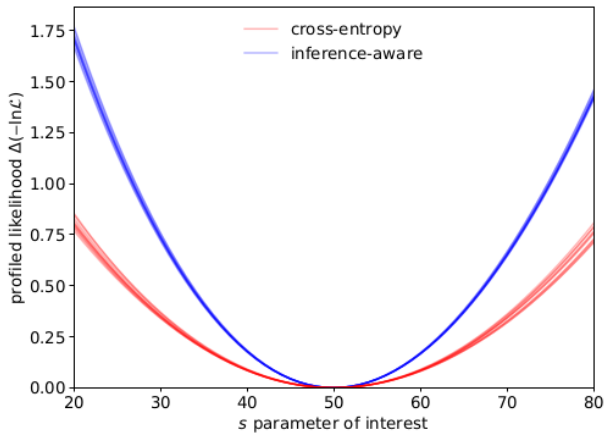
INFERNO: inference-aware neural optimization

- Build non-parametric likelihood function based on simulation, use it as summary statistic
- Minimize the expected variance of the parameter of interest
  - Obtain the Fisher information matrix with automatic differentiation, and use it as loss function
  - For (asymptotically) unbiased estimators, Rao-Cramér-Frechet (RCF) bound $V[\hat{\theta}] \sim \frac{1}{\hat{\theta}}$ (see my Monday lesson)
  - Constraints via auxiliary measurements (typically on nuisance parameters) included in covariance matrix out of the box





(a) inference-aware training loss



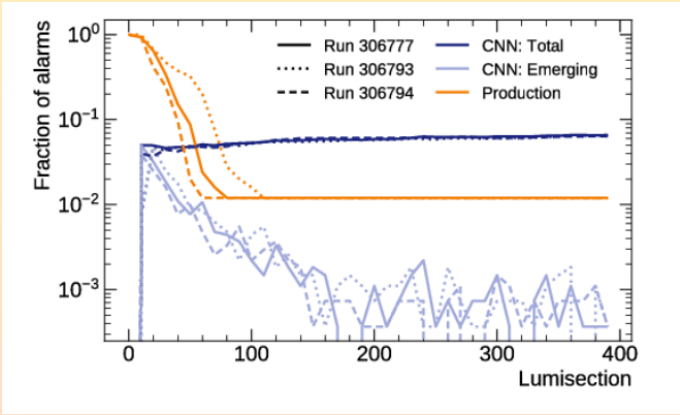
(b) profile-likelihood comparison

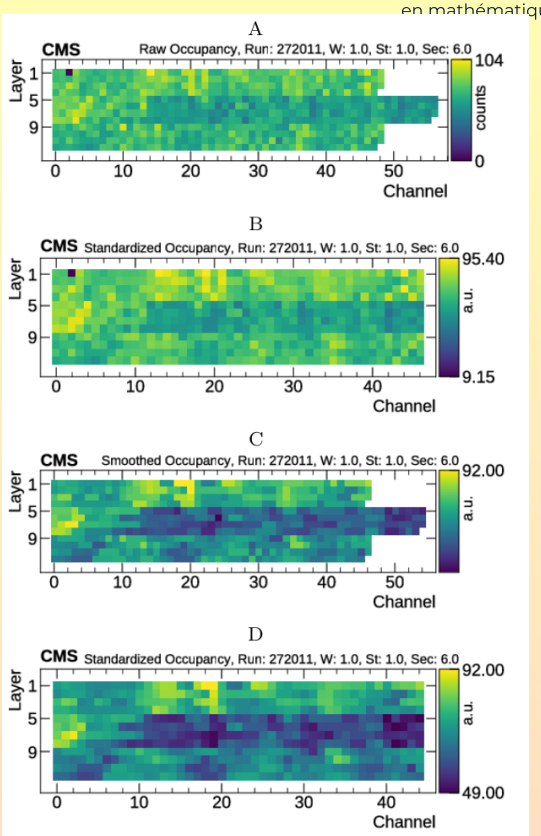
From De Castro-Dorigo, arXiv:1806.04743, and AMVA4NewPhysics deliverable 1.4 public report

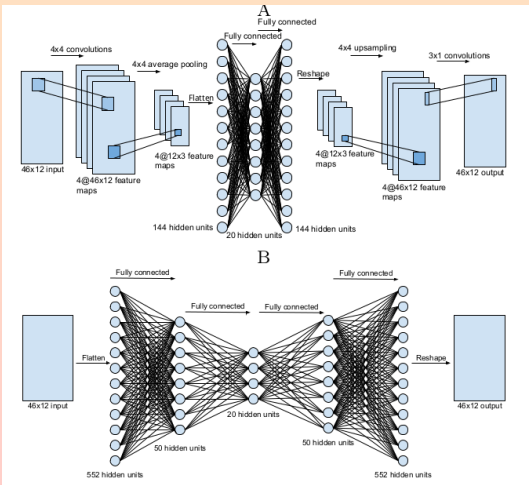
Which data should we take?

What if we don't know which data to take?

- Represent data as geographically-organized images
  - Local focus: detector layers treated independently
  - Regional focus: detector layers treated independently but simultaneously (spot problems between layers)
- Autoencoders (noise detection, dimensionality reduction)
  - Encode the inputs to the hidden layer
  - Decode the hidden layer to an approximate representation of the inputs

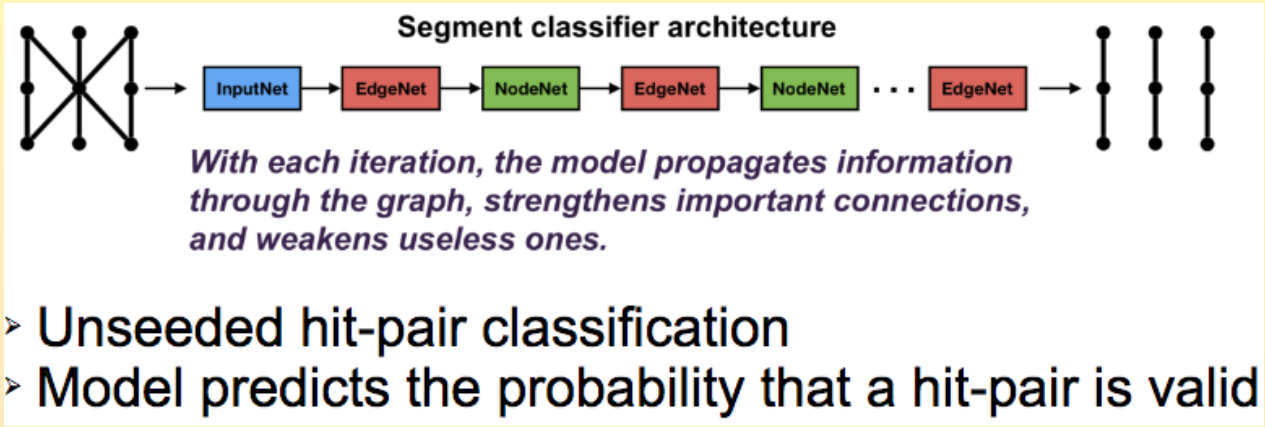


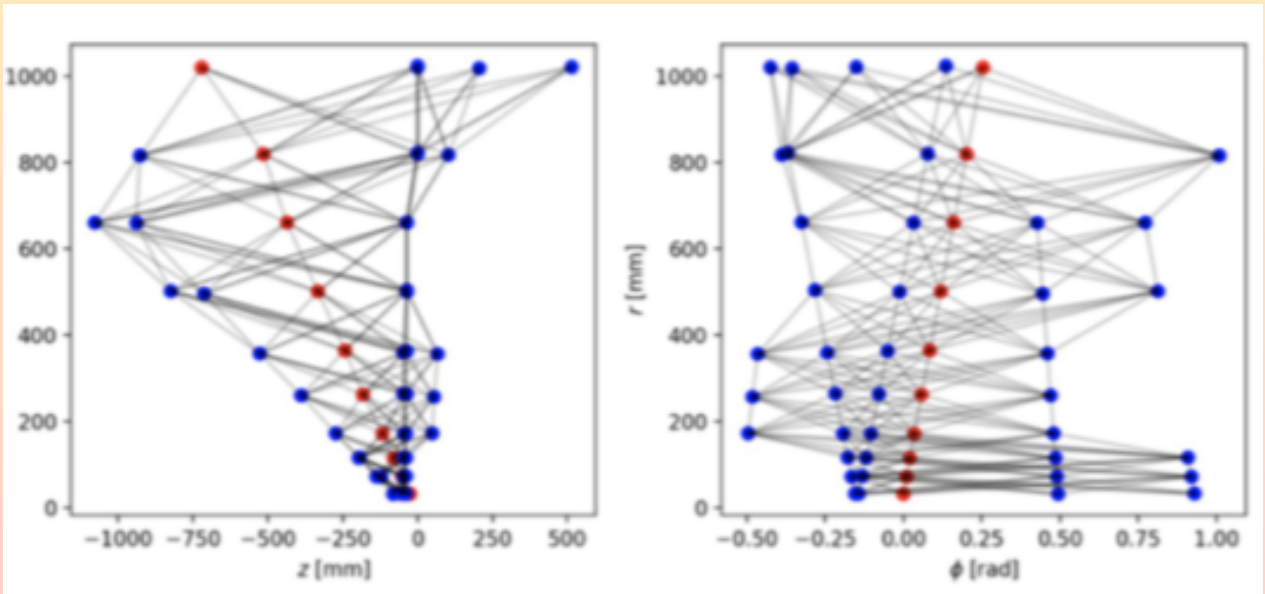


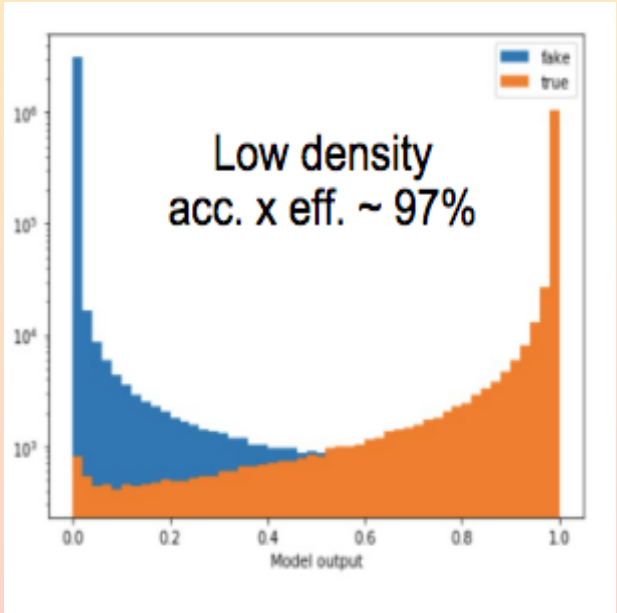


From arXiv:1808.00911

- Graph networks to literally connet the dots



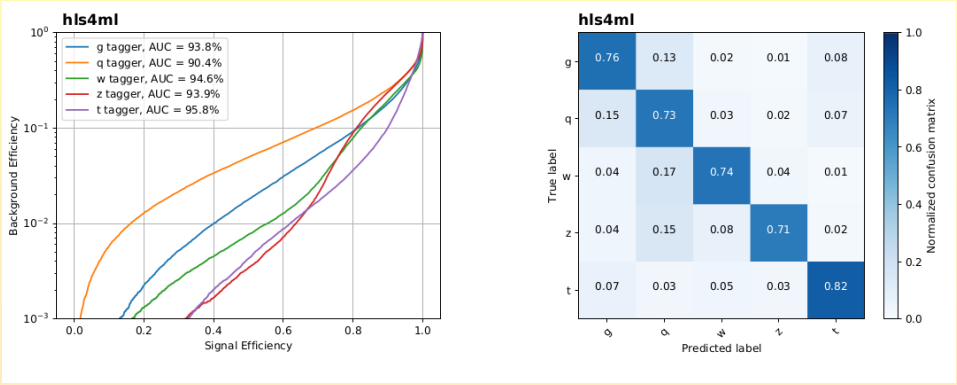


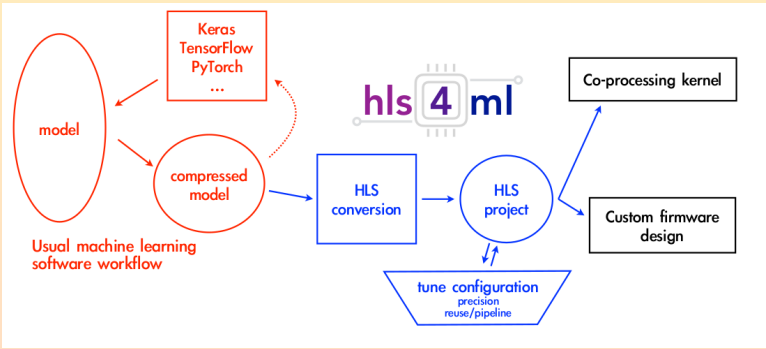


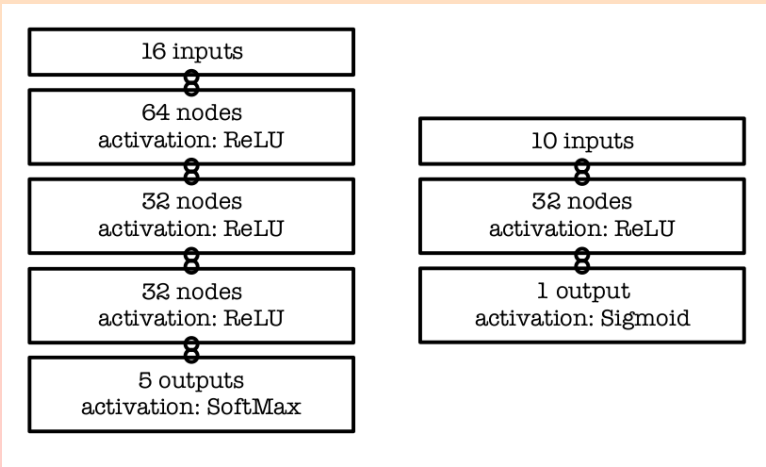
The HEP.TrkX project, S. Gleyzer's talk at 3rd IML workshop

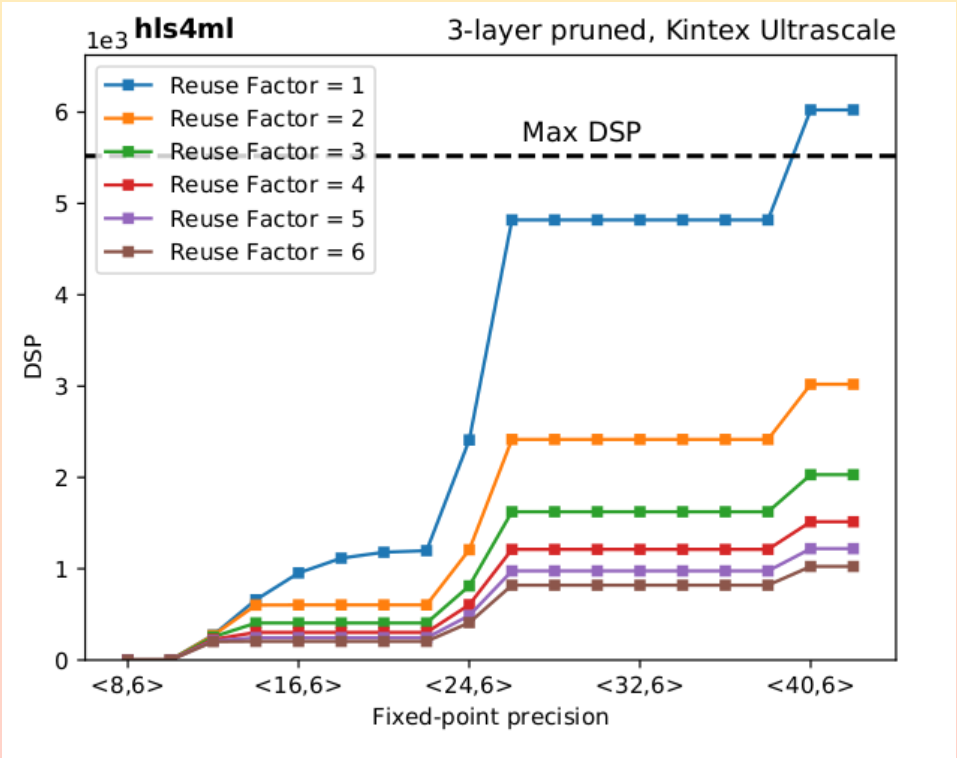
What if you need to do it quickly?

- Real-time event processing requires low-latency and low-power-consumption hardware: FPGAs
- Case study: classify structures inside jets (jet substructure)
- Compression, quantization, parallelization digital signal processing (arithmetic) blocks (DSPs),





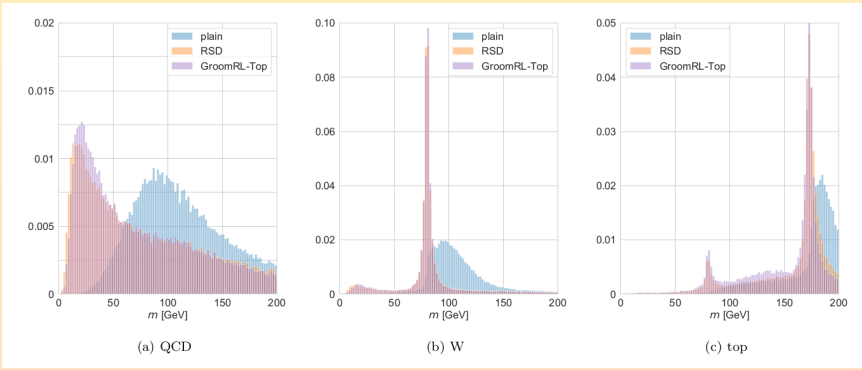




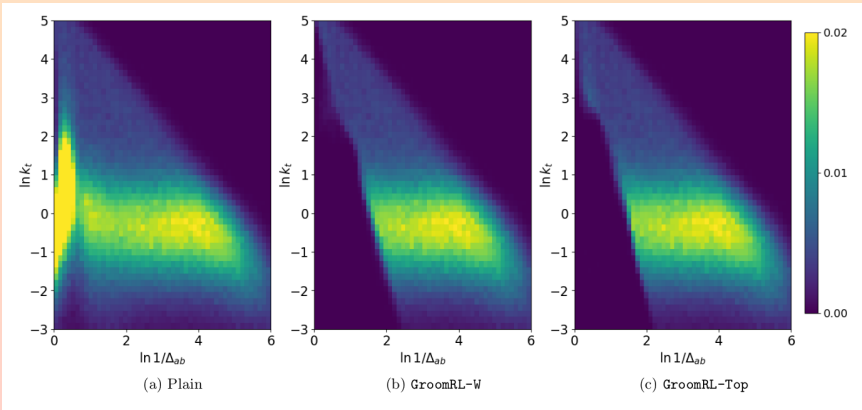
From arXiv:1804.06913

Next? Probably Deep Q learning (reinforcement learnin)

- Boosted objects decay to collimated jets reconstructed as a single jet
- Fat jet grooming: remove soft wide-angle radiation not associated with the underlying hard substructure



(a) QCD (b) W (c) top



(a) Plain (b) GroomRL-W (c) GroomRL-Top

Images from arXiv:1903.09644 and The Auckland Dog Coach

- Statistics is about answering questions
  - ...and posing the questions in an appropriate way
- Foundations
  - Mathematical definition of probability
  - Bayesian and Frequentist realizations
- How wide is the table?: Point estimates and the method of maximum likelihood
- Is it really that wide, or am I somehow uncertain about it?: Interval estimates
  - Maximum likelihood
  - Neyman construction
  - Feldman-Cousins ordering
  - Coverage
- Is the table a standard-size ping-pong table or not? Testing hypotheses
  - Frequentist hypothesis testing, and some mention to the Bayesian one
  - I need no toy: the Wilks theorem
  - Upper limits and the CL_s prescription
- Can I decouple my result from my instrumentation? Unfolding
- How can I exploit learning algorithms? Machine Learning
  - Machine learning is a well defined mathematical technique
  - Used in many flavours across all the spectrum of tasks in HEP
- Are you satisfied? Tell me more by clicking here https://forms.gle/XntoBLdDoUmqZYcL7

THANK YOU VERY MUCH FOR ATTENDING!!

This course has already improved on the fly thanks to you!
I'll take any further feedback and trasforming into improvements for the next edition!

Non-exhaustive list of references

- Frederick James: Statistical Methods in Experimental Physics - 2nd Edition, World Scientific
- Glen Cowan: Statistical Data Analysis - Oxford Science Publications
- Louis Lyons: Statistics for Nuclear And Particle Physicists - Cambridge University Press
- Louis Lyons: A Practical Guide to Data Analysis for Physical Science Students - Cambridge University Press
- E.T. Jaynes: Probability Theory - Cambridge University Press 2004
- Annis?, Stuard, Ord, Arnold: Kendall's Advanced Theory Of Statistics I and II
- Pearl, Judea: Causal inference in Statistics, a Primer - Wiley
- R.J.Barlow: A Guide to the Use of Statistical Methods in the Physical Sciences - Wiley
- Kyle Cranmer: Lessons at HCP Summer School 2015
- Kyle Cranmer: Practical Statistics for the LHC - http://arxiv.org/abs/1503.07622
- Roberto Trotta: Bayesian Methods in Cosmology - https://arxiv.org/abs/1701.01467
- Harrison Prosper: Practical Statistics for LHC Physicists - CERN Academic Training Lectures, 2015 https://indico.cern.ch/category/72/
- Christian P. Robert: The Bayesian Choice - Springer
- Sir Harold Jeffreys: Theory of Probability (3rd edition) - Clarendon Press
- Harald Crámer: Mathematical Methods of Statistics - Princeton University Press 1957 edition

THANKS FOR THE ATTENTION!

Backup