Statistics

or “How to find answers to your questions”

Pietro Vischia[1]

[1] CP3 — IRMP, Université catholique de Louvain

LIP LHC Physics Course 2021, March 15th–17th

Practicalities

UCLouvain
Institut de recherche
en mathématique et physique

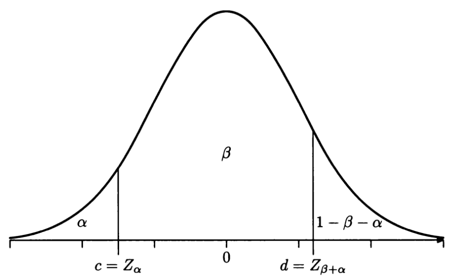
- Schedule: two days × two hours
- This has evolved to being an excerpt from a longer course (about 10h)
  - You can find additional material, as well as a set of exercises, at https://agenda.irmp.ucl.ac.be/event/4097/
  - The exercises, in particular, are designed to show the inner workings of several techniques: try to run them! You can find them at https://github.com/vischia/intensiveCourse_public
- Many interesting references, nice reading list for your career
  - Papers mostly cited in the topical slides
  - Some cool books cited here and there and in the appendix
- These slides include some material that we won't able to cover today
  - Mostly to provide some additional details without having to refer to the full course
  - Slides with this advanced material are those with the title in red
- Unless stated otherwise, figures belong to P. Vischia for inclusion in my upcoming textbook on Statistics for HEP
  (textbook to be published by Springer in 2021)
  - Or I forgot to put the reference, let me know if you spot any figure obviously lacking reference, so that I can fix it
  - I cannot put the recordings publicly online as “massive online course”, so I will distribute them only to registered participants, and have to ask you to not record yourself. I hope you understand.
- Your feedback is crucial for improving these lectures (a feedback form will be provided at the end of the lectures)!
  - You can also send me an email during the lectures: if it is something I can fix for the next day, I'll gladly do so!

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

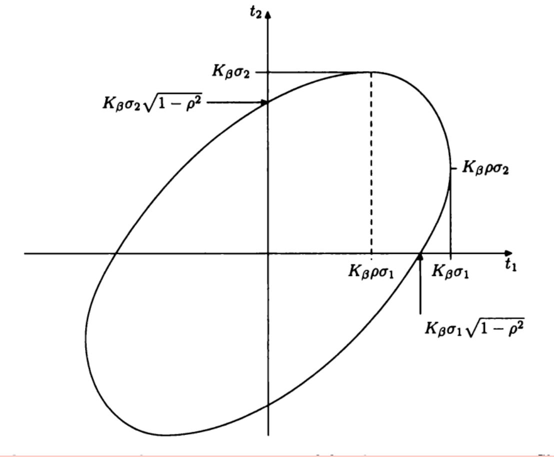
- Example: if $f(X|\theta) = N(\mu, \sigma^2)$ with unknown μ, σ, choose $Z = \frac{X-\mu}{\sigma}$
- Find $[c, d]$ in $\beta = P(c \leq Z \leq d) = \Phi(d) - \Phi(c)$ by finding $[Z_\alpha, Z_{\alpha+\beta}]$
- Infinite interval choices: here central interval $\alpha = \frac{1-\beta}{2}$



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.

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!
- **Coverage is a property of the method you use to compute your confidence interval**
  - It is calculated from the sampling distribution of your confidence intervals
- 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

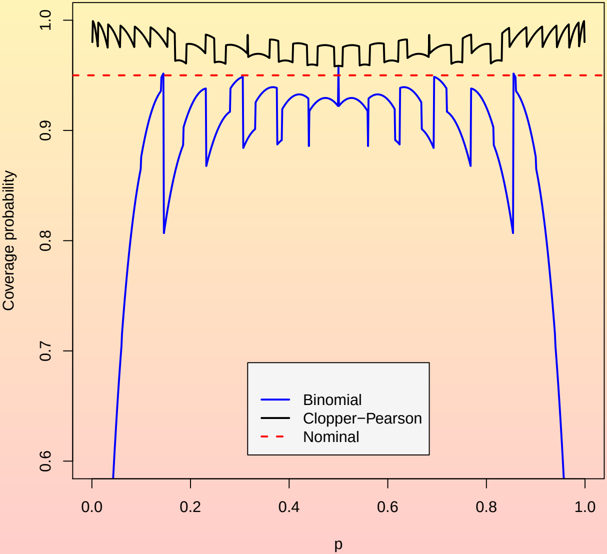
- For discrete distributions, the discreteness induces steps in the probability content of the interval
    - Continuous case: $P(a \leq X \leq b) = \int_a^b f(X|\theta)dX = \beta$
    - Discrete case: $P(a \leq X \leq b) = \sum_a^b f(X|\theta)dX \leq \beta$
- Binomial: find interval (r_{low}, r_{high}) 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
  $\sum_{r=0}^{N} \binom{r}{N} p^n (1-p_{low})^{N-n} \leq \alpha/2$
  $\sum_{r=0}^{N} \binom{r}{N} p^r (1-p_{high})^{N-r} \leq \alpha/2$
    - Single-tailed → 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 → use $\alpha/2$ instead of α
- This afternoon we will study the coverage of intervals from a gaussian approximation and from the Clopper-Pearson method
- We will also study the coverage of intervals obtained from crossings with ΔlnL
- Question time: Coverage

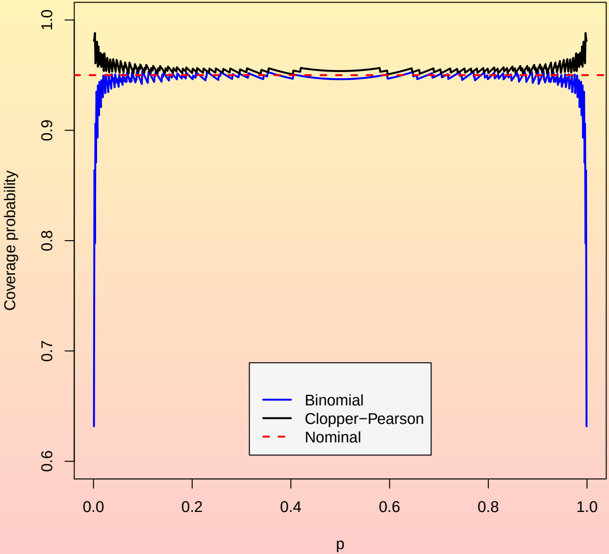
Coverage, $N = 20$

- Gaussian approximation bad for small sample sizes



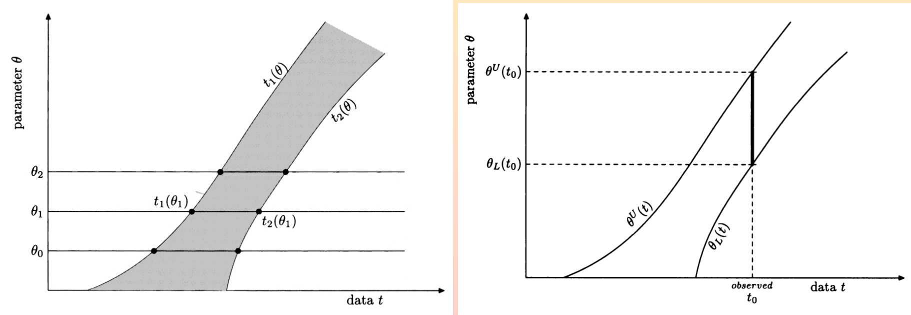
Coverage, $N = 1000$

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



Confidence belts: the Neyman construction

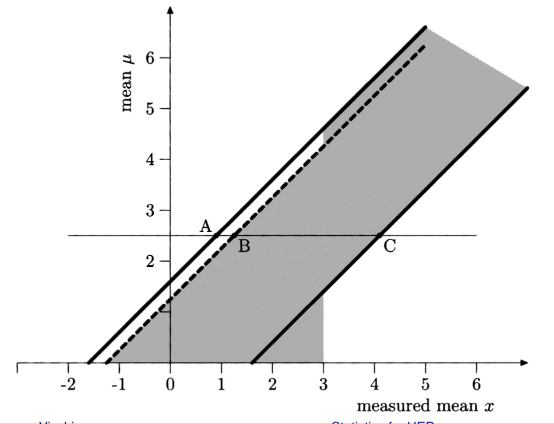
- Unique solutions to finding confidence intervals are infinite
  - Central intervals, lower limits, upper limits, etc
- Let's suppose we have chosen a way
- Build horizontally: for each (hypothetical) value of θ, determine $t_1(\theta)$, $t_2(\theta)$ such that $\int_{t_1}^{t_2} P(t|\theta)dt = \beta$
- 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.

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)

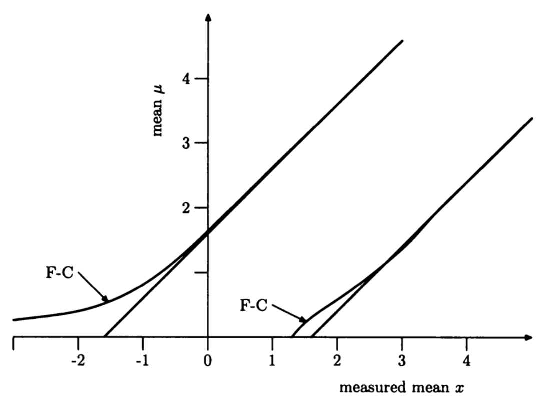


Plot from James, 2nd ed.

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



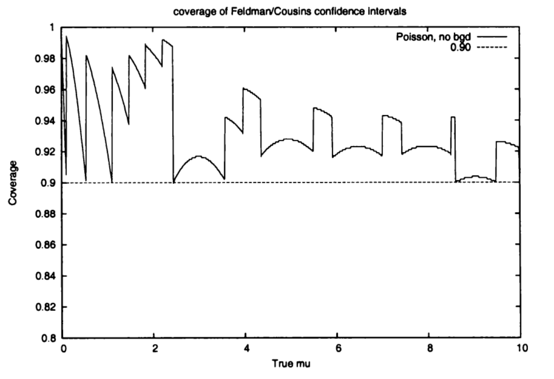
Plot from James, 2nd ed.

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

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

- Sometimes you may want to summarize the prior with estimates of its location and of its dispersion
  - For the location, you can use mode or median (see tomorrow's lecture)
- 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

Bayesian intervals and coverage

- What about computing the frequentist coverage for Bayesian intervals?
- Question time: Coverage Bayes

Bayesian intervals and coverage

- What about computing the frequentist coverage for Bayesian intervals?
- Question time: Coverage Bayes
- Even if you are not interested in frequentist methods, it can be useful! Certainly it doesn't hurt
- Knowing the sampling properties of a method can always give insights or work as a cross-check of the method
- Particularly given that typically Bayesian and frequentist answers tend to converge in the high-N limit
  - Except for hypothesis tests, we'll find out later today

Image from the Statistical Statistics Memes Facebook Page

Test of Hypotheses

What is an hypothesis...

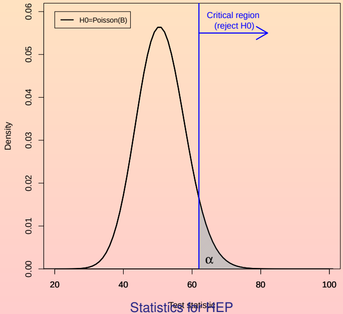
- Is our hypothesis compatible with the experimental data? By how much?
- Hypothesis: a complete rule that defines probabilities for data.
  - An hypothesis is simple if it is completely specified (or if each of its parameters is fixed to a single value)
  - An hypothesis is complex if it consists in 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.

...and how do we test it?

- 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

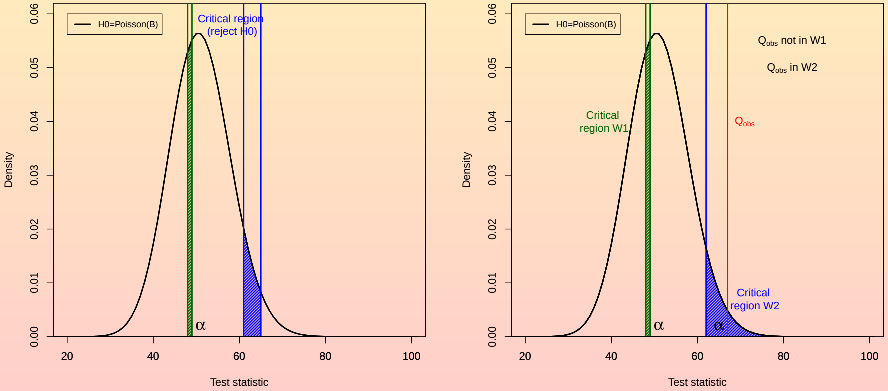
Testing an hypothesis H_0...

- 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!!!



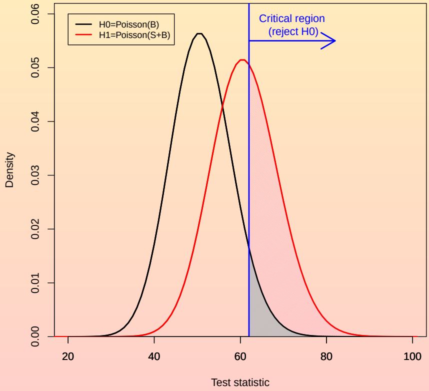
...while introducing some spice in it

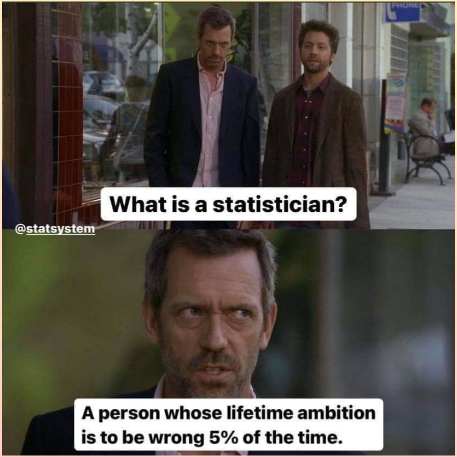
- The definition of $\mathcal{W}$ depends only on its area α, without any other condition
  - Any other area of area α can be defined as critical region, independently on how it is placed with respect to $\vec{x}_{obs}$
  - In particular, for an infinite number of choices of $\mathcal{W}$, the point $\vec{x}_{obs}$—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 <u>accept</u> H_0 to <u>reject</u> 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

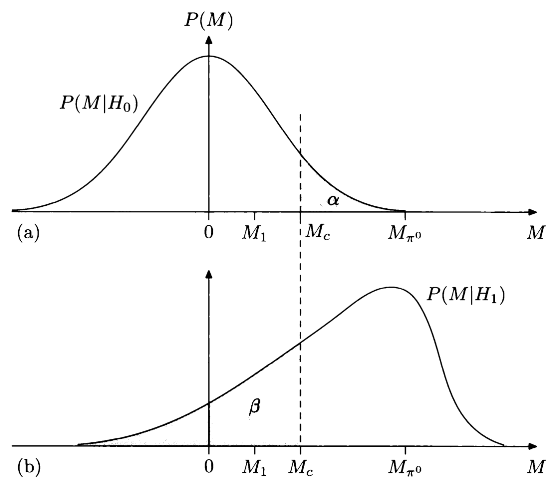



What is a statistician?
@statsystem
A person whose lifetime ambition is to be wrong 5% of the time.

Could not find source for the meme

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, $M = 135\ MeV$

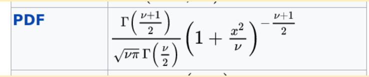


	Choose H_0	Choose H_1
H_0 is true	$1-\alpha$	α (Type I error)
H_1 is true	β (Type II error)	$1-\beta$ (power)

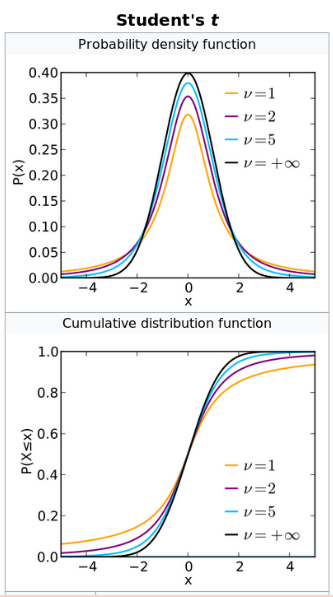
Plot from James, 2nd ed.

A longer example

- Student’s t distribution
- Test the mean!
- Will not run it this afternoon, you can check it at home hyptest.ipynb

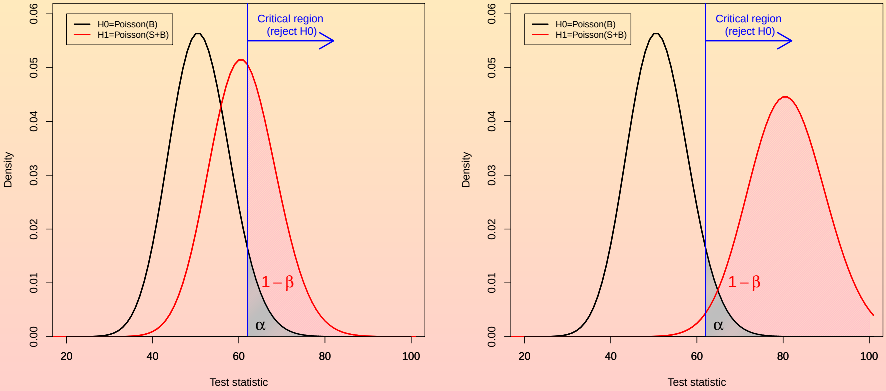


PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$



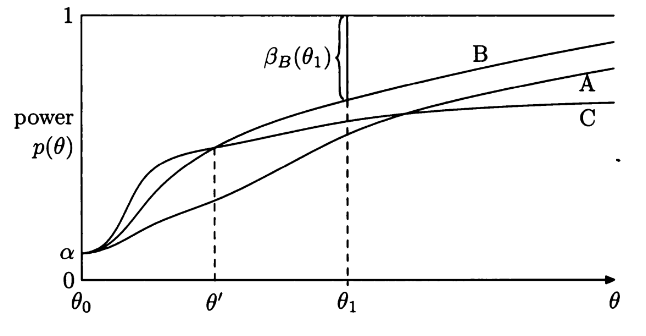
Basic hypothesis testing – 4

- 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.



Comparing tests

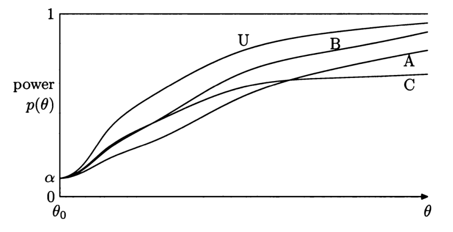
- For parametric (families of) hypotheses, the power depends on the parameter
  - $H_0 : \theta = \theta_0$
  - $H_1 : \theta = \theta_1$
  - Power: $p(\theta_1) = 1 - \beta$
- Generalize for all possible alternative hypotheses: $p(\theta) = 1 - \beta(\theta)$
  - For the null, $p(\theta_0) = 1 - \beta(\theta_0)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

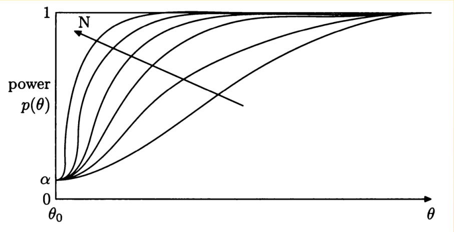


Plot from James, 2nd ed.

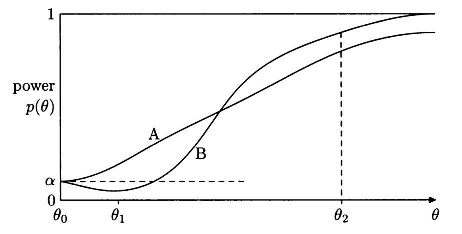
Properties of tests

- 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 $\theta_1]$
- 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

Image from the Statistical Statistics Memes Facebook Page

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$\psi = \psi^*$.

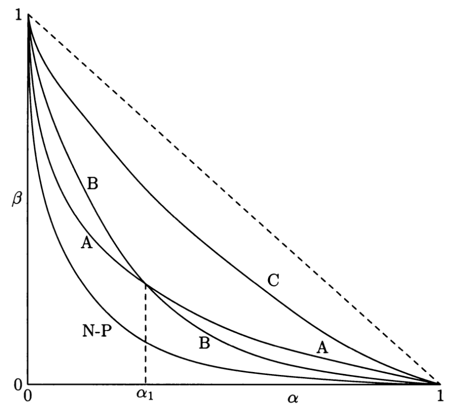


Table 10.4. A cost function.

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

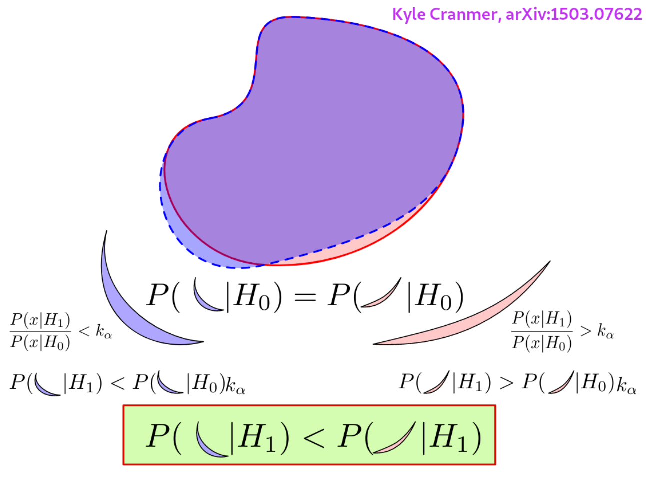
Demonstrating the Neyman-Pearson lemma

- We want to prove that $\ell(\mathbf{X}, \theta_0, \theta_1) := \frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} \geq c_\alpha$ gives the best acceptance region

Image from Evan Vucci, Shutterstock, meme is mine

Demonstrating the Neyman-Pearson lemma

- We want to prove that $\ell(\mathbf{X}, \theta_0, \theta_1) := \frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} \geq c_\alpha$ gives the best region
  - Critical region from NP (red contour), demonstrate that any other region (blue contour) has less power
  - Take out a wedge region and add it e.g. to the other side
  - Regions must have equal area under H_0 (tests with same size)
  - Being on different sides of the red contour, under H_1 data is less likely in the added region than in the removed one
  - Less probability to reject the null → test based on the new contour is less powerful!

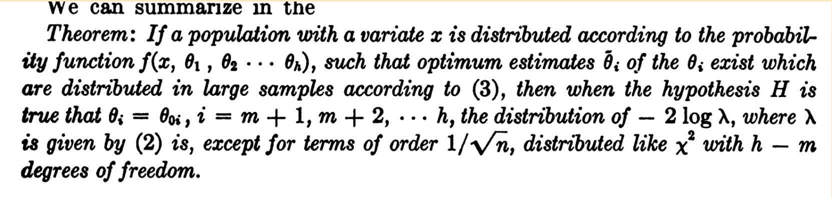


Kyle Cranmer, arXiv:1503.07622

$P(\,|H_0) = P(\,|H_0)$
$\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha$
$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$
$P(\,|H_1) < P(\,|H_0)k_\alpha$
$P(\,|H_1) > P(\,|H_0)k_\alpha$
$P(\,|H_1) < P(\,|H_1)$

Intermezzo: the Wilks theorem

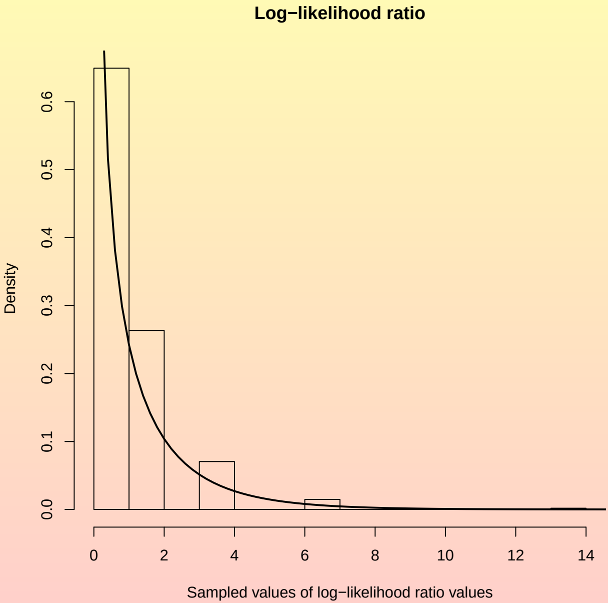
- The likelihood ratio is commonly used
- As any test statistic in the market, in order to select critical regions based on confidence levels it is necessary to know its distribution
  - Run toys to find its distribution (very expensive if you want to model extreme tails)
  - Find some asymptotic condition under which the likelihood ratio assumes a simple known form
- Wilks theorem: when the data sample size tends to ∞, the likelihood ratio tends to $\chi^2(N-N_0)$
  - Exercise yesterday afternoon



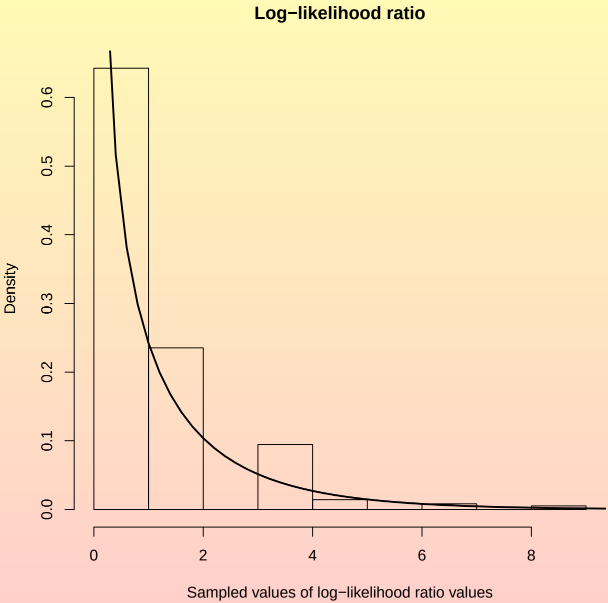
We can summarize in the

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 $\bar{\theta}_i$ 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.*

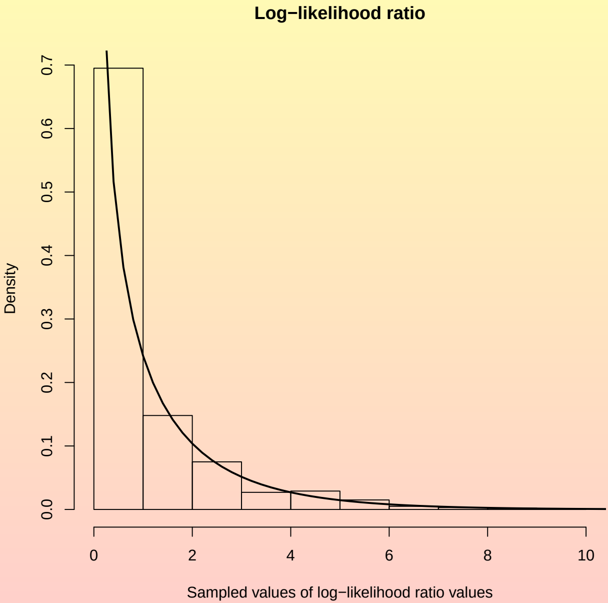
Verifying the Wilks theorem: N=2



Verifying the Wilks theorem: N=10



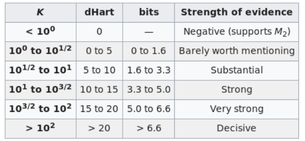
Verifying the Wilks theorem: N=100



Bayesian model selection — two models...

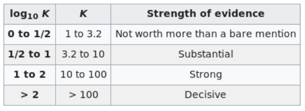
- The parameter θ might be predicted by two models M_0 and M_1: $P(\theta|\vec{x},M) = \frac{P(\vec{x}|\theta,M)P(\theta|M)}{P(\vec{x}|M)}$
  - A step further than yesterday in writing down the Bayes theorem: now multiple conditioning
  - $P(\vec{x}|M) = \int P(\vec{x}|\theta,M)P(\theta|M)d\theta$: *Bayesian 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 × 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



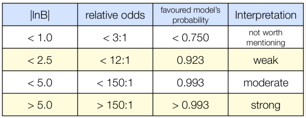
K	dHart	bits	Strength of evidence
$< 10^0$	0	—	Negative (supports M_2)
10^0 to $10^{1/2}$	0 to 5	0 to 1.6	Barely worth mentioning
$10^{1/2}$ to 10^1	5 to 10	1.6 to 3.3	Substantial
10^1 to $10^{3/2}$	10 to 15	3.3 to 5.0	Strong
$10^{3/2}$ to 10^2	15 to 20	5.0 to 6.6	Very strong
$> 10^2$	> 20	> 6.6	Decisive

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



$\lvert \ln B \rvert$	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

Bayesian model selection — ...with many models

Bayesian model comparison of 193 models
Higgs inflation as reference model

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

Martin,RT+14

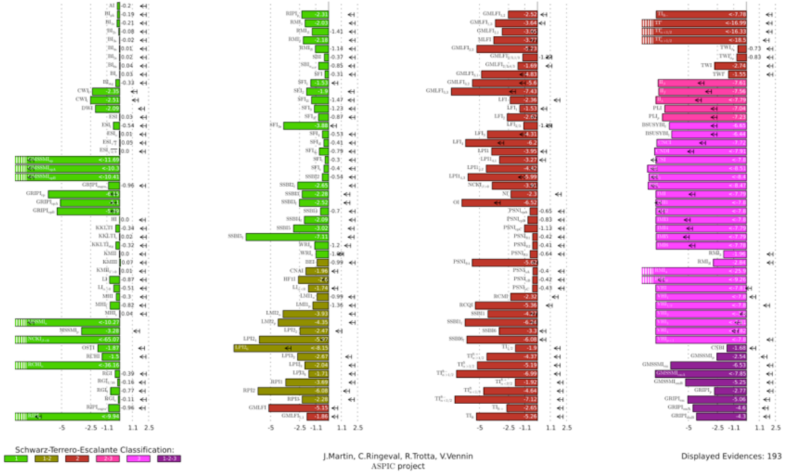
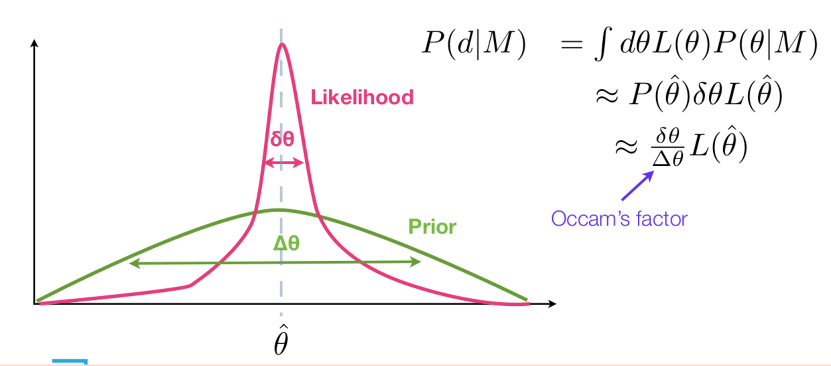


Image from Roberto Trotta, Chair Lemaitre Lectures 2018

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



$$
\begin{aligned}
P(d|M) &= \int d\theta L(\theta)P(\theta|M) \\
&\approx P(\hat{\theta})\delta\theta L(\hat{\theta}) \\
&\approx \frac{\delta\theta}{\Delta\theta} L(\hat{\theta})
\end{aligned}
$$

Occam's factor

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

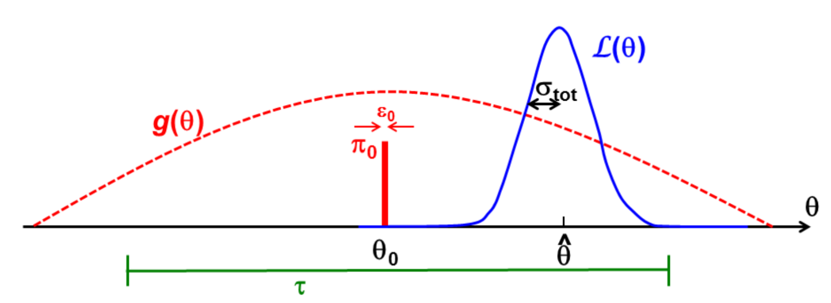


Image from Cousins, doi:10.1007/s11229-014-0525-z

How does HEP looks for new physics?

- 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!

Testing hypotheses near the boundary: Zech

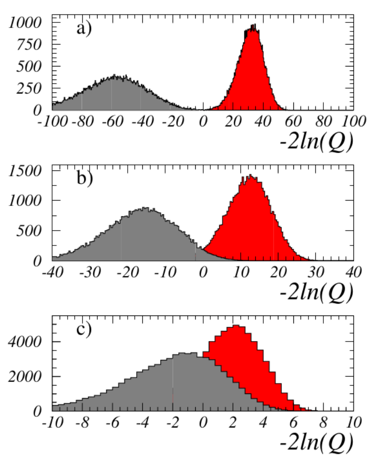
- Counting experiment: observe n events
- Assume they come from Poisson processes: $n \sim Pois(s+b)$, with known b
- Set limit on s given n_{obs}
- Exclude values of s for which $P(n \leq n_{obs}|s+b) \leq \alpha$ (guaranteed coverage $1-\alpha$)
- $b = 3$, $n_{obs} = 0$
  - Exclude $s + b \leq 3$ at 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

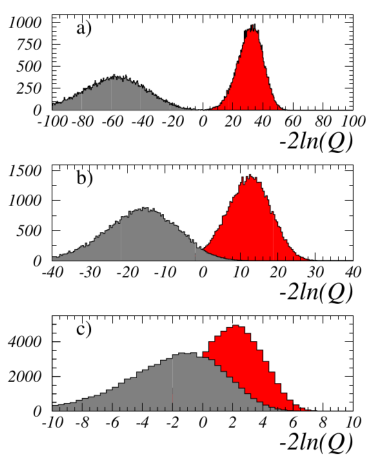
- Find a monotonic Q for increasing signal-like experiments (e.g. likelihood ratio)
- $CL_{s+b} = P_{s+b}(Q \leq Q_{obs})$
  - Small values imply poor compatibility with $S+B$ hypothesis, favouring B-only
- $CL_b = P_b(Q \leq Q_{obs})$
  - Large (close to 1) values imply poor compatibility with B-only, 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

Get your confidence levels right

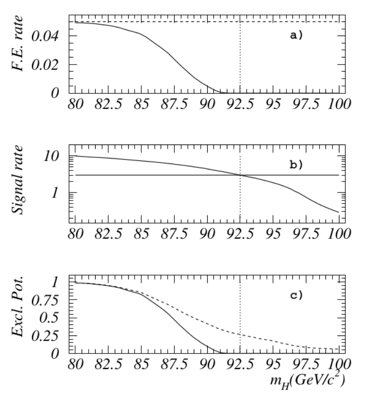
- Find a monotonic Q for increasing signal-like experiments (e.g. likelihood ratio)
- $CL_{s+b} = P_{s+b}(Q \leq Q_{obs})$
  - Small values imply poor compatibility with $S+B$ hypothesis, favouring B-only
- $CL_b = P_b(Q \leq Q_{obs})$
  - Large (close to 1) values imply poor compatibility with B-only, 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

- $CL_s := \frac{CL_{s+b}}{CL_b}$ x
- 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!

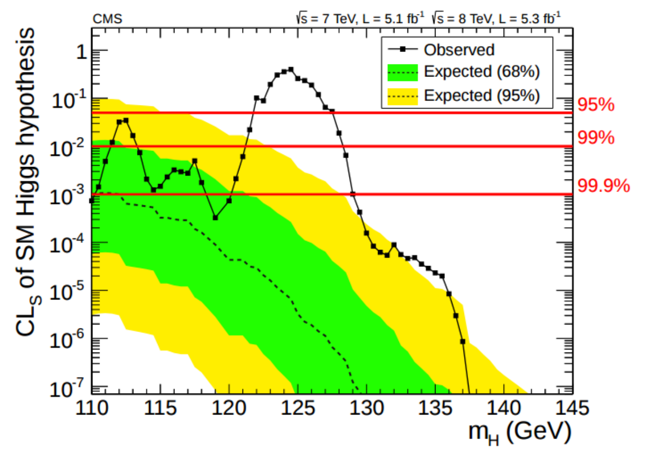


Dashed: CL_{s+b}
Solid: CL_s
$S < 3$: exclusion for a B-free search $\equiv 0$

Plot from Read, CERN-open-2000-205

That's what we used for the Higgs discovery!

- 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

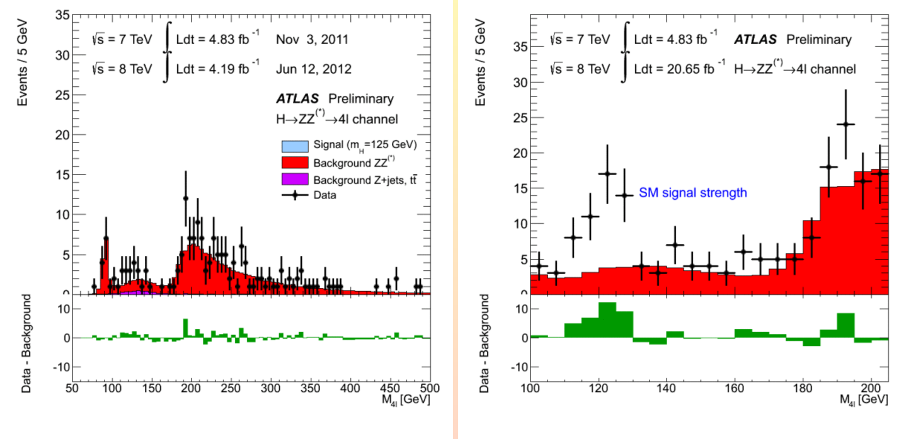


Plot from Higgs discovery paper

Deal with CLs!

- This afternoon we'll play with CLs!

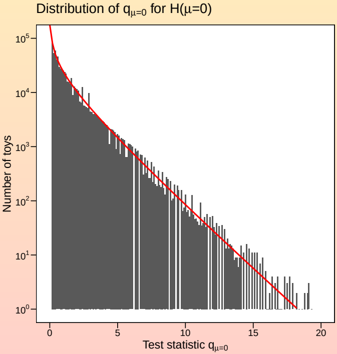
I have an excess, do I?

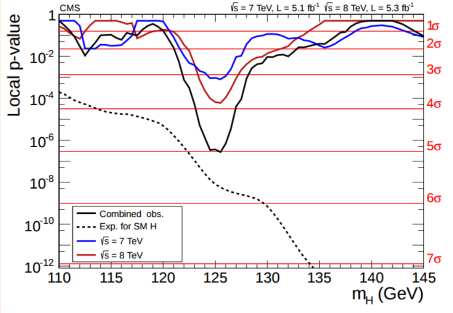


Plot from https://cds.cern.ch/record/2230893

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 $p = \frac{1}{2}P_{\chi_1^2}(Z^2)$

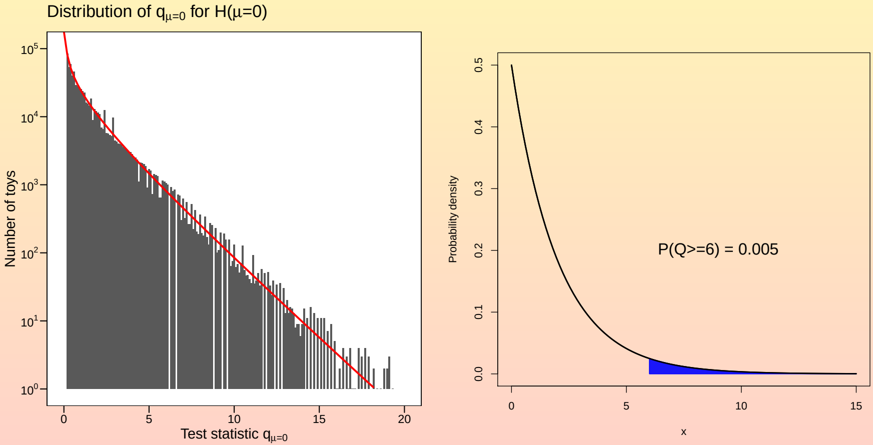




Left plot by Pietro Vischia, right plot from ATL-PHYS-PUB-2011-011 and Higgs discovery paper

Look only at the null hypothesis!

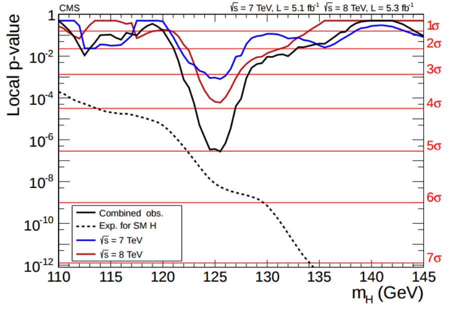
- Probability of obtaining a fluctuation with test statistic q_{obs} or larger, under the null hypothesis H_0
  - Distribution of test statistic under H_0 either with toys or asymptotic approximation (if N_{obs} is large, then $q \sim \chi^2(1)$)

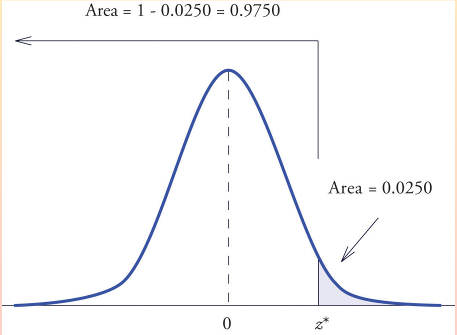


Plots from Vischia—in preparation with Springer

And the sigmas?

- Just an artifact to convert p-values to easy-to-remember $\mathcal{O}(1)$ numbers
  - 1σ: $p = 0.159$
  - 3σ: $p = 0.00135$
  - 5σ: $p = 0.000000285$
- No approximation involved, just a change of units to gaussian variances: one-sided tail area $\frac{1}{2\pi}\int_x^\infty e^{-\frac{t^2}{2}}dt = p$
  - p-value must be **flat** under the null, or interpretation is invalidated
- HEP: usually interested in one-sided deviations (upper fluctuations)
  - Most other disciplines interested in two-sided effects (e.g. 2σ: $p_{2sided} = 0.05$)



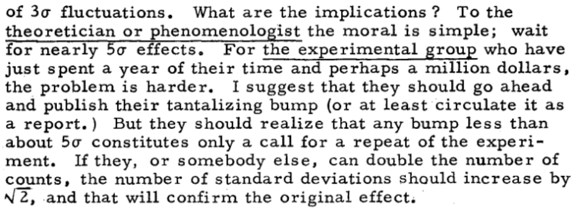


Left: ATLAS Collaboration, Right: https://saylordotorg.github.io/

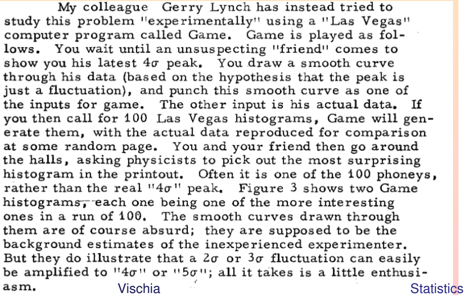
- Question time: Significance

Fluctuations in HEP? The proposal of a 5σ criterion

- Rosenfeld, 1968 (https://escholarship.org/uc/item/6zm2636q) *Are there any Far-out Mesons or Baryons?*
  - “In summary of all the discussion abouve, I conclude that each of our 150,000 annual histograms is capable of generating somewhere between 10 and 100 deceptive upward fluctuations [...] (we) should expect several 4σ and hundreds of 3σ fluctuations”



of 3σ fluctuations. What are the implications? To the theoretician or phenomenologist the moral is simple; wait for nearly 5σ effects. For the experimental group who have just spent a year of their time and perhaps a million dollars, the problem is harder. I suggest that they should go ahead and publish their tantalizing bump (or at least circulate it as a report.) But they should realize that any bump less than about 5σ constitutes only a call for a repeat of the experiment. If they, or somebody else, can double the number of counts, the number of standard deviations should increase by $\sqrt{2}$, and that will confirm the original effect.



My colleague Gerry Lynch has instead tried to study this problem "experimentally" using a "Las Vegas" computer program called Game. Game is played as follows. You wait until an unsuspecting "friend" comes to show you his latest 4σ peak. You draw a smooth curve through his data (based on the hypothesis that the peak is just a fluctuation), and punch this smooth curve as one of the inputs for game. The other input is his actual data. If you then call for 100 Las Vegas histograms, Game will generate them, with the actual data reproduced for comparison at some random page. You and your friend then go around the halls, asking physicists to pick out the most surprising histogram in the printout. Often it is one of the 100 phoneys, rather than the real "4σ" peak. Figure 3 shows two Game histograms, each one being one of the more interesting ones in a run of 100. The smooth curves drawn through them are of course absurd; they are supposed to be the background estimates of the inexperienced experimenter. But they do illustrate that a 2σ or 3σ fluctuation can easily be amplified to "4σ" or "5σ"; all it takes is a little enthusiasm.

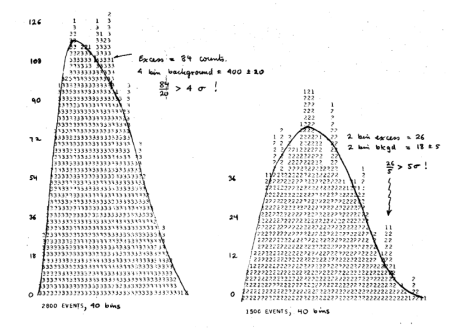
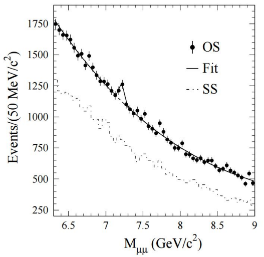
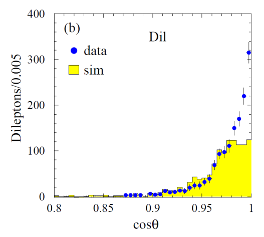


Fig. 3. Two "Las Vegas" histograms generated by G. Lynch's program GAME.

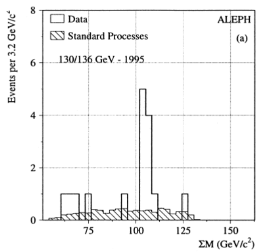
HEP has a history of unconfirmed effects

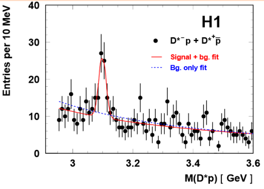
- 3.5σ (2005, CDF) in dimuon (candidate bottom squark, doi:/10.1103/PhysRevD.72.092003)





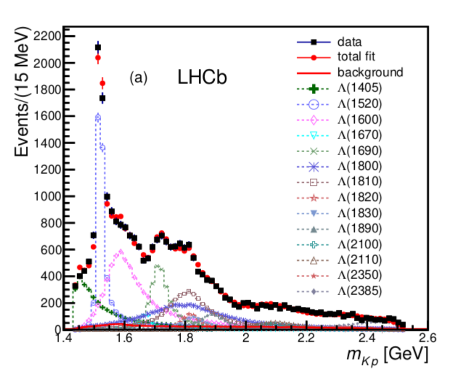
- $\sim 4\sigma$ (1996, Aleph) in four-jet (Higgs boson candidate, doi:/10.1007/BF02906976)
- 6σ (2004, H1) (narrow $\bar{c}$ baryon state, doi:/10.1016/j.physletb.2004.03.012)
    - H1 speaks of “Evidence”, not confirmed.

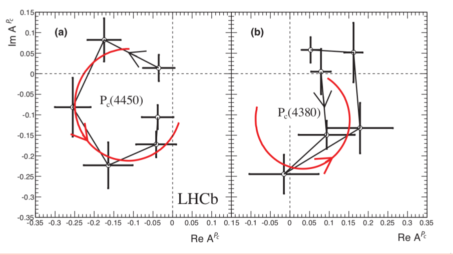




The revenge of the pentaquarks

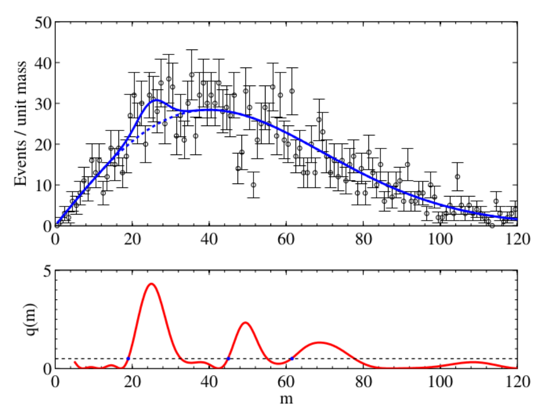
- 9σ and 12σ (2015, LHCb): pentaquarks! (doi:/10.1103/PhysRevLett.115.072001)
  - Several cross-checks (fit to mass spectrum, fit with non-resonant components, evolution of complex amplitude in Argand diagrams)
  - Mass measurement, soft statement: “Interpreted as resonant states they must have minimal quark content of $ccuud$, and would therefore be called charmonium-pentaquark states.
- One remark: quoting significances above about 5–6σ is meaningless
  - Asymptotic approximation not trustable (tail effects). Can run lots of toys but...
  - ...cannot possibly trust knowing your systematic uncertainties to that level





The Look-elsewhere effect — 1

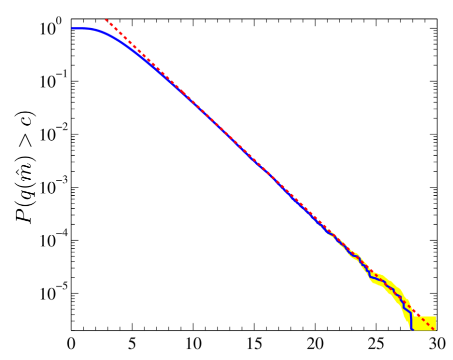
- Searching for a resonance X of arbitrary mass
  - H_0 = no resonance, the mass of the resonance is not defined (Standard Model)
  - $H_1 = H(M \neq 0)$, but there are infinite possible values of M
- Wilks theorem not valid anymore, no unique test statistic encompassing every possible H_1
- Quantify 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)

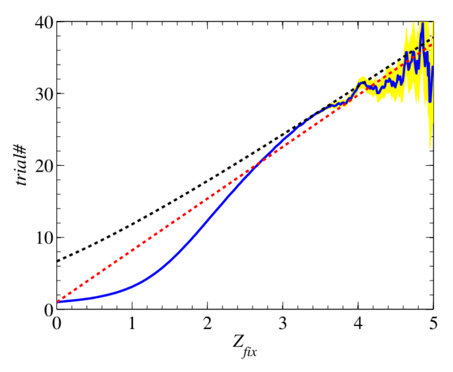


Plot from Gross-Vitells, 10.1140/epjc/s10052-010-1470-8

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*

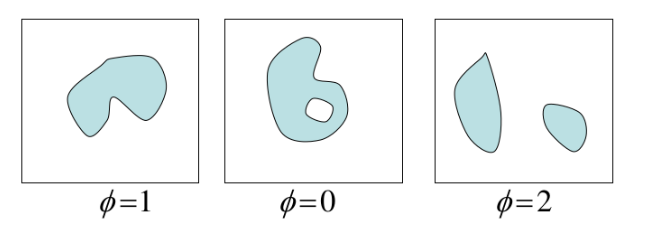


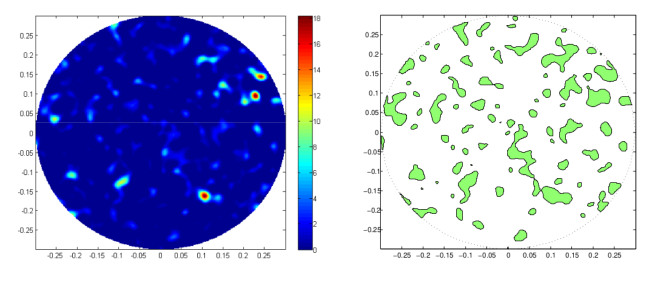


Plot from Gross-Vitells, 10.1140/epjc/s10052-010-1470-8

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

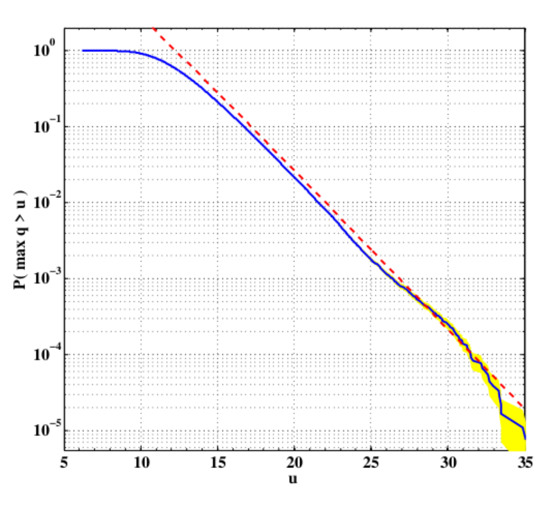




Plot from Gross-Vitells, 10.1016/j.astropartphys.2011.08.005

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

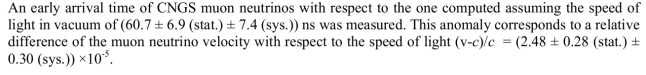
- Asymptoticity holds also for the 2D effect, as desired
  - Dashed red lines: prediction based on the formula with upcrossings
  - Blue: $200k$ toys (pseudoexperiments)



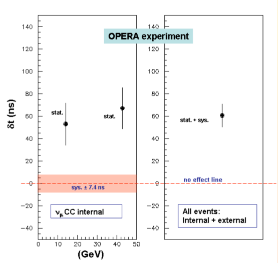
Plot from Gross-Vitells, 10.1016/j.astropartphys.2011.08.005

When there is no LEE, you still need to make sure your systematics are right

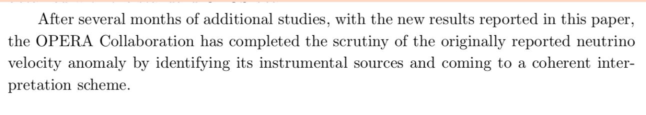
- In 2011 OPERA (arXiv:1109.4897v1) reported superluminal neutrino speed, with 6.0σ significance...



An early arrival time of CNGS muon neutrinos with respect to the one computed assuming the speed of light in vacuum of (60.7 ± 6.9 (stat.) ± 7.4 (sys.)) ns was measured. This anomaly corresponds to a relative difference of the muon neutrino velocity with respect to the speed of light (v-c)/c = (2.48 ± 0.28 (stat.) ± 0.30 (sys.)) $\times 10^{-5}$.



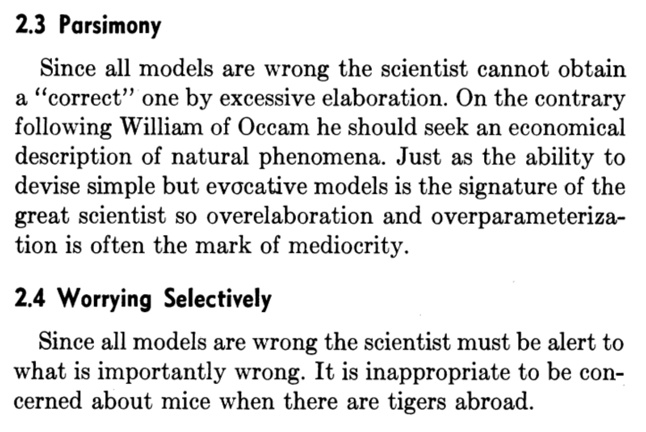
- ...but they had a loose cable connector (doi:/10.1007/JHEP10(2012)093)



After several months of additional studies, with the new results reported in this paper, the OPERA Collaboration has completed the scrutiny of the originally reported neutrino velocity anomaly by identifying its instrumental sources and coming to a coherent interpretation scheme.

Truth and models: all models are wrong

- Box (https://www.jstor.org/stable/2286841) warns that any model is an approximation



2.3 Parsimony

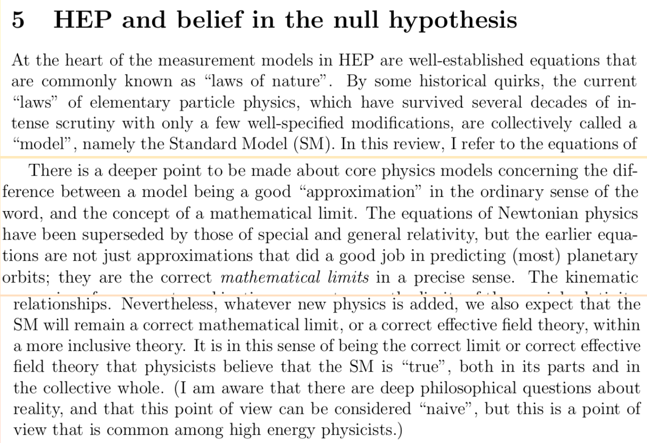
Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

2.4 Worrying Selectively

Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

Truth and models: HEP is special

- Cousins (doi:/10.1007/s11229-014-0525-z) notes HEP is in a privileged position when compared with social or medical sciences



5 HEP and belief in the null hypothesis

At the heart of the measurement models in HEP are well-established equations that are commonly known as "laws of nature". By some historical quirks, the current "laws" of elementary particle physics, which have survived several decades of intense scrutiny with only a few well-specified modifications, are collectively called a "model", namely the Standard Model (SM). In this review, I refer to the equations of

There is a deeper point to be made about core physics models concerning the difference between a model being a good "approximation" in the ordinary sense of the word, and the concept of a mathematical limit. The equations of Newtonian physics have been superseded by those of special and general relativity, but the earlier equations are not just approximations that did a good job in predicting (most) planetary orbits; they are the correct *mathematical limits* in a precise sense. The kinematic

relationships. Nevertheless, whatever new physics is added, we also expect that the SM will remain a correct mathematical limit, or a correct effective field theory, within a more inclusive theory. It is in this sense of being the correct limit or correct effective field theory that physicists believe that the SM is "true", both in its parts and in the collective whole. (I am aware that there are deep philosophical questions about reality, and that this point of view can be considered "naive", but this is a point of view that is common among high energy physicists.)

- Others (Gelman, Raftery, Berger, Bernardo) argue that a point null is impossible (at most "small")

Ground truth, models, and point nulls

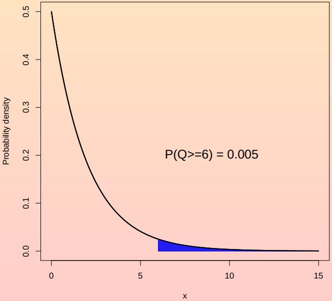
- I think a point or almost-point null is related to our simplifications rather than with a claim on reality
- Some disciplines deal with phenomena which cannot (yet) be explained from first principles
  - Maybe one day we will have a full quasi-deterministic model of a whole body or brain
  - Certainly so far most models are attempts at finding a functional form for the relationship between two variables
- Some disciplines (HEP) have to do with phenomena which can be explained from first principles
  - These principles are *reasonable* but not necessarily the best or the only possible ones
  - No guarantee that they reflect a universal truth
  - Arguing that the vast experimental agreement of the SM implies ground truth behaves based on our principles sounds a bit wishful thinking
  - What can be claimed is that the vast experimental agreement warrants the use of point or quasi-point nulls
- Box’s view on models, and the Occam’s Razor, should still lead considerations on model choices
  - A version of the Occam’s Razor is even implemented in Bayesian model selection
- Still, to avoid interpreting fluctuations as real effects all disciplines should strive—when possible—to describe causal relationships rather than correlations

Goodness of fit tests

Also the three most common types of hypothesis testing also mention somewhere about duality test-confregion

The χ^2 distribution: why degrees of freedom?

- Sample randomly from a Gaussian p.d.f., obtaining X_1 y X_2
- $Q = X_1^2 + X_2^2$ (or in general $Q = \sum_{i=1}^{N} X_i^2$) is itself a random variable
  - What is $P(Q \geq 6)$? Just integrate the $\chi^2(N = 2)$ distribution from 6 to ∞
- Depends only on N!
  - If we sample 12 times from a Gaussian and compute $Q = \sum_{i=1}^{12} X_i^2$, then $Q \sim \chi^2(N = 12)$
- Theorem: if $Z_1, ..., Z_N$ is a sequence of normal random variables, the sum $V = \sum_{i=1}^{N} Z_i^2$ is distributed as a $\chi^2(N)$
  - The sum of squares is closely linked to the variance $E[(X - \mu)^2] = E[X^2] - \mu^2$ from Eq. **??**
- The χ^2 distribution is useful for goodness-of-fit tests that check how much two distributions diverge point-by-point
- It is also the large-sample limit of many distributions (useful to simplify them to a single parameter)



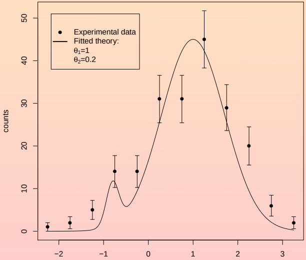
The χ^2 distribution: goodness-of-fit tests 1/

- Consider a set of M measurements $\{(X_i, Y_i)\}$
  - Suppose Y_i are affected by a random error representable by a gaussian with variance σ_i
- Consider a function $g(X)$ with predictive capacity, i.e. such that for each i we have $g(X_i) \sim Y_i$
- Pearson's χ^2 function related to the difference between the prediction and the experimental measurement in each point

$$\chi^2_P := \sum_{i=1}^{M} \left[\frac{Y_i - g(X_i)}{\sigma_i} \right]^2 \qquad (1)$$

- Neyman's χ^2 is a similar expression under some assumptions
  - If the gaussian error on the measurements is constant, it can be factorized
  - If Y_i represent event counts $Y_i = n_i$, then the errors can be approximated with $\sigma_i \propto \sqrt{n_i}$

$$\chi^2_N := \sum_{i=1}^{M} \frac{\left(n_i - g(X_i)\right)^2}{n_i} \qquad (2)$$



The χ^2 distribution: goodness-of-fit tests 2/

- If $g(X_i) \sim Y_i$ (i.e. $g(X)$ reasonably predicts the data), then each term of the sum is approximately 1
- Consider a function of $\chi^2_{N,P}$ and of the number of measurements M
  - $E[f(\chi^2_{N,P}, M)] = M$
  - The function is analytically a χ^2:

$$f(\chi^2, M) = \frac{2^{-\frac{M}{2}}}{\Gamma\left(\frac{N}{2}\right)} \chi^{N-2} e^{-\frac{\chi^2}{2}} \tag{3}$$

  - The cumulative of f is

$$1 - cum(f) = P(\chi^2 > \chi^2_{obs} | g(x) \text{ is the correct model}) \tag{4}$$

- Comparing χ^2 with the number of degrees of freedom M, we therefore have a criterion to test for goodness-of-fit
  - For a given M, the p.d.f. is known ($\chi^2(M)$) and the observed value can be computed and compared with it
  - Null hypothesis: there is no difference between prediction and observation (i.e. g fits well the data)
  - Alternative hypothesis: there is a significant difference between prediction and observation
  - Under the null, the sum of squares is distributed as a $\chi^2(M)$
  - p-values can be calculated by integration of the χ^2 distribution

$$\begin{aligned} \frac{\chi^2}{M} &\sim 1 \Rightarrow \text{g(X) approximates well the data} \\ \frac{\chi^2}{M} &>> 1 \Rightarrow \text{poor model (increases } \chi^2\text{), or statistically improbable fluctuation} \\ \frac{\chi^2}{M} &<< 1 \Rightarrow \text{overestimated } \sigma_i\text{, or fraudulent data, or statistically improbable fluctuation} \end{aligned} \tag{5}$$

The χ^2 distribution: goodness-of-fit tests 3/

- $\chi^2(M)$ tends to a Normal distribution for $M \to \infty$
  - Slow convergence
  - It is generally not a good idea to substitute a χ^2 distribution with a Gaussian
- The goodness of fit seen so far is valid only if the model (the function $g(X)$) is fixed
- Sometimes the model has k free parameters that were not given and that have been fit to the data
- Then the observed value of χ^2 must be compared with $\chi^2(N')$, with $N' = N - k$ degrees of freedom
  - $N' = N - k$ are called reduced degrees of freedom
  - This however works only if the model is linear in the parameters
  - If the model is not linear in the parameters, when comparing χ^2_{obs} with $\chi^2(N-k)$ then the p-values will be deceptively small!
- Variant of the χ^2 for small datasets: the G-test
  - $g = 2\sum O_{ij} ln(O_{ij}/E_{ij})$
  - It responds better when the number of events is low (Petersen 2012)

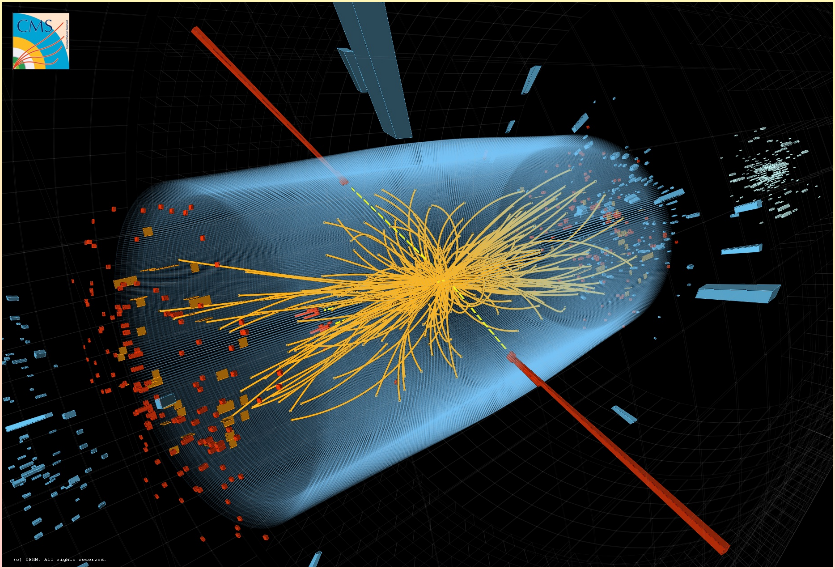
Machine learning

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 (Springer 2017)

We must efficiently collect and well reconstruct data

- ~ 40 MHz (millions per second) collision photos
  - Can store and reconstruct only a few of them



We must efficiently sample from our models

$$P(\mathbf{x}|\alpha) = \boxed{\frac{1}{A_\alpha \sigma_\alpha}} \int \boxed{\mathrm{d}\Phi(y)\frac{\mathrm{d}x_1 \mathrm{d}x_2}{x_1 x_2 s}\; f(x_1) f(x_2)} \boxed{|\mathcal{M}_\alpha(y, x_1, x_2)|^2} \boxed{W(\mathbf{x}|y)\, \epsilon_\alpha(y)}$$

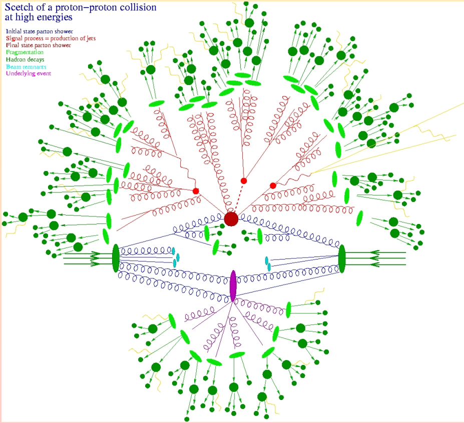
Normalization factor

We collide protons and that is a mess

Linear operator in Hamiltonian formalism, describes the physics process

Detector response, experimental efficiencies

- Costly MonteCarlo simulations, sampling from these high-dimensional probability density functions


Scetch of a proton−proton collision
at high energies
Initial state parton shower
Signal process = production of jets
Final state parton shower
Fragmentation
Hadron decays
Beam remnants
Underlying event

We must improve the tools for detailed studies (e.g. EFT, differential)

- The Standard Model leaves some questions open
  - What is the origin of the Higgs mechanism? The Higgs field vacuum expectation (246 GeV) very far from Planck scale (quantum gravity): hierarchy problem
  - Origin of the observed neutrino masses? Most explanations of neutrino non-zero masses and mixing are beyond the SM
  - Dark Matter: a new, hidden sector of particles and forces?
  - Is the Higgs boson discovered in 2012 the Standard Model one?
- The study of Higgs boson physics is crucial for many of these topics
  - New scalar bosons (e.g. charged Higgs bosons) by simple extensions of the Higgs sector of the SM
  - **Slight deviations** from the expected properties of the observed Higgs boson could reveal signs for new physics

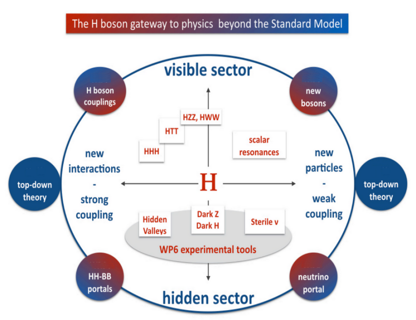
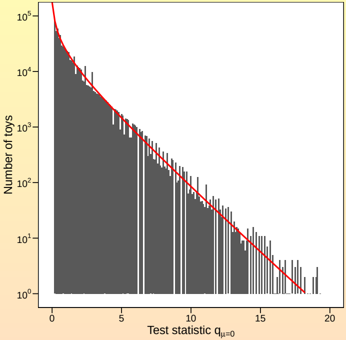


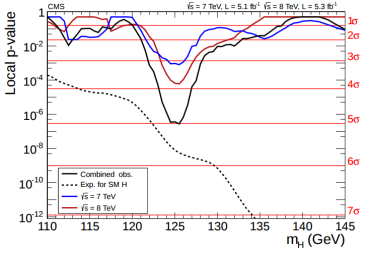
Image by the EOS be.h network

Ultimately, we must improve our inference: the end goal!

- Statistical inference to make statements about parameters of our models
- New physics?
    - Probability of extreme fluctuation under the null measures significance of excess
    - Function of other parameters under investigation (e.g. Higgs boson mass in 2012)
- Systematic uncertainties induce variations in the number of events in the search region
    - We account for them in our statistical procedures at the hypothesis testing stage
- Often machine learning techniques are employed to optimize the analysis at early stages: **systematic uncertainties not accounted for in the optimization**

Distribution of $q_{\mu=0}$ for $H(\mu=0)$





Images from Phys. Lett. B 716 (2012) 30 and P. Vischia, ***** (textbook to be published by Springer in 2021)

Mathematical formalism

- Let’s formalize the concept of learning from data
- We’ll look into the formalism mostly for supervised learning
- Fore more mathematical details, see arXiv:1712.04741 and Joan Bruna’s lectures online

The elements of supervised learning: input space

- $\mathcal{X}$: a high-dimensional input space
  - The challenges come from the high dimensionality!
- If all dimensions are real-valued, $\mathbb{R}^d$
  - For square images of side $\sqrt{d}$, $\mathcal{X} = \mathbb{R}^d$, $d \sim \mathcal{O}(10^6)$

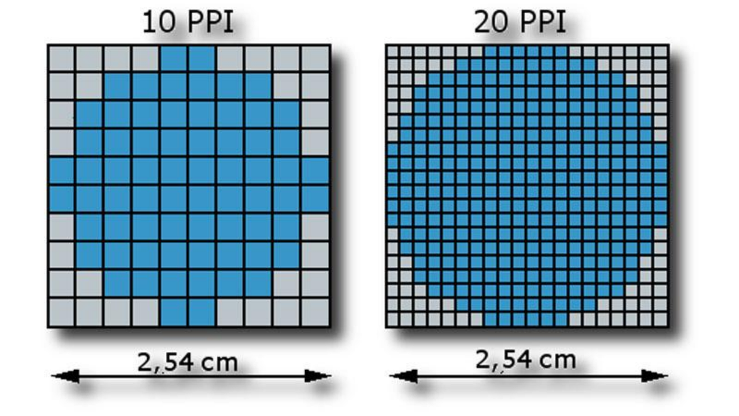


Figure frm scientiamobile.com

The elements of supervised learning: data probability distribution

- ν: unknown data probability distribution
  - We can sample from it to obtain an arbitrary amount of data points
  - We are not allowed to use any analytic information about it in our computations

The elements of supervised learning: the target function

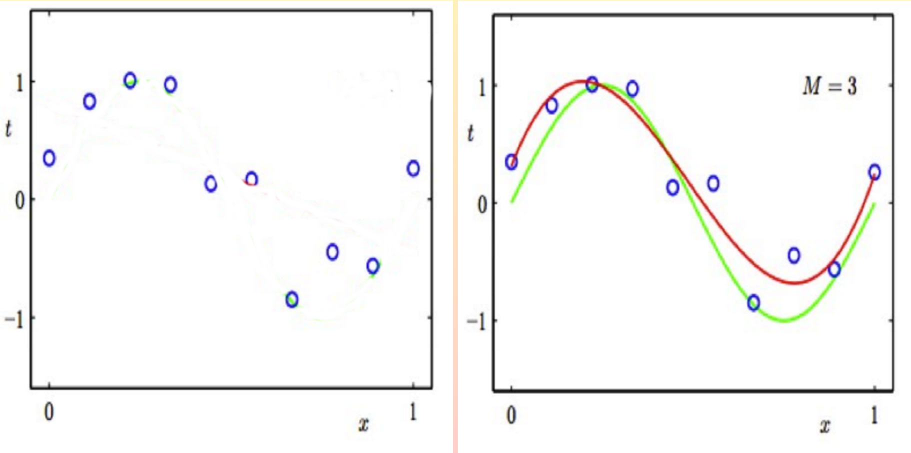
- $f^*:\mathcal{X} \to \mathbb{R}$, unknown target function
  - In case of multidimensional output to a vector of dimension k, $f^*:\mathcal{X} \to \mathbb{R}^k$
  - Some loose assumptions (e.g. square-integrable with respect to the ν measure, i.e. finite moments, bounded...)

How do we learn? The loss functional

- $L[f] = \mathbb{E}_{\nu}\left[l\Big(f(x), f^{*}(x)\Big)\right]$
  - The metric that tells us how good our predictions are
- The function $l(\cdot,\cdot)$ is a given expression, e.g. regression loss, logistic loss, etc
  - In this lecture, typically it is the L^2 norm: $\|f - f^{*}\|_{L^2(\mathcal{X},\nu)}$

Learning goal

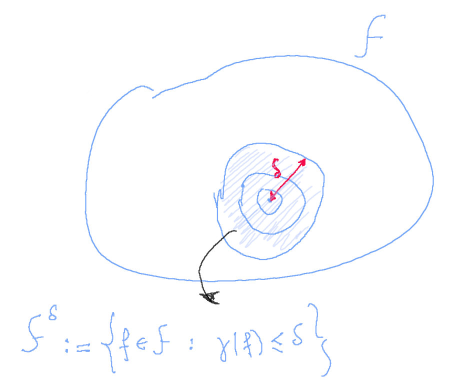
- Goal: predict f^* from a finite i.i.d. sample of points sampled from ν
- Sample: $\left\{x_i, f^*(x_i)\right\}_{i=1,\dots,n}$, $x_i \sim \nu$
  - For each of the points x_i, we know the value of the unknown function (our true labels)
  - We want to *interpolate* for any arbitrary x inbetween the labelled x_i...
  - ...in million of dimensions!



Images by Victor Lavrenko

The space of possible solutions

- The space of functionals that can potentially solve the problem is vast: $\mathcal{F} \subseteq \left\{ f : \mathcal{X} \to \mathbb{R} \right\}$ (hypothesis class)
- We need a notion of complexity to “organize” the space
- $\gamma(f), f \in \mathcal{F}$: *complexity* of f
    - It can for example be the norm, i.e. we can augment the space $\mathcal{F}$ with the norm
- When the complexity is defined via the norm, $\mathcal{F}$ is highly organized: Banach space!
    - The simplest function according to the norm criterion is the 0 function
    - If we increase the complexity by increasing the norm, we obtain **convex balls** $\left\{ f \in \mathcal{F}; \gamma(f) \leq \delta \right\} =: \mathcal{F}^{\delta}$
    - Convex minimization is considerably easier than non-convex minimization


$\mathcal{F}$
δ
$\mathcal{F}^{\delta} := \{ f \in \mathcal{F} : \gamma(f) \leq \delta \}$

Empirical Risk Minimization

- For each element of $\mathcal{F}$, a measure of how well it's interpolating the data
- Empirical risk: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} |f(x_i) - f^*(x_i)|^2$
  - $|\cdot|$ is the empirical loss. If it's the norm, then $\hat{L}(f)$ is the empirical Mean Square Error
  - If you find an analogy with least squares method, it's because for one variable it's exactly that!

Formalizing the minimization of a functional in a given space

- **Constraint form:** $\min_{f \in \mathcal{F}^{\delta}} \hat{L}(f)$.
  - Not trivial
- **Penalized form:** $\min_{f \in \mathcal{F}} \hat{L}(f) + \lambda \gamma(f)$.
  - More typical
  - λ is the price to pay for more complex solutions. Depends on the complexity measure
- **Interpolant form:** $\min_{f \in \mathcal{F}} \gamma(f)$ s.t. $\hat{L}(f) = 0 \iff f(x_i) = f^*(x_i) \qquad \forall i$
  - In ML, most of the times there is no noise, so $f(x_i)$ is exactly the value we expect there (i.e. we really know that x_i is of a given class, without any uncertainty)
  - The interpolant form exploits this (*"give me the least complex elements in $\mathcal{F}$ that interpolates"*)

- These forms are not completely equivalent. The penalized form to be solved requires averaging a full set of penalized forms, so it's not completely equivalent
- There is certainly an implicit correspondence between δ and λ (the larger λ, the smaller δ and viceversa)

The Fundamental Theorem of Machine Learning

- We want to relate the result of the empirical risk minimization (ERM) with the prediction
    - Let's use the constraint form
- Let's assume we have solved the ERM at a precision ϵ (we are ϵ-away from...) we then have $\hat{f} \in \mathcal{F}^{\delta}$ such that $\hat{L}(\hat{f}) \leq \epsilon + \min_{f \in \mathcal{F}^{\delta}} \hat{L}(f)$
- How good is $\hat{f}$ at predicting f^*? In other words, what's the true loss?
    - Can use the triangular inequality

$$L(\hat{f}) - \inf_{f \in \mathcal{F}} L(f) \leq \inf_{f \in \mathcal{F}^{\delta}} L(f) - \inf_{f \in \mathcal{F}} L(f)$$

Approximation error

(how appropriate is my measure of complexity)

$$+2 \sup_{f \in \mathcal{F}^{\delta}} \left| L(f) - \hat{L}(f) \right|$$

Statistical error

(impact of having the empirical loss instead of the true loss)

$$+\epsilon$$

Optimization error

The Error Market

- The minimization is regulated by the parameter δ (the size of the ball in the space of functions)
- Changing δ results in a **tradeoff** between the different errors
  - Very small δ makes the statistical error blow up
- We are better at doing convex optimization (easier to find minimum), but even then the optimization error ϵ will not be negligible
  - ϵ: how much are ou willing to spend in resources to minimize $\hat{L}(f)$
  - We kind of control it!
  - If the other errors are smaller than ϵ, then it makes sense to spend resources to decrease it
  - Otherwise, don’t bother

Bottou and Bousquet, 2008, Shalev-Shwartz, Ben-David

The big questions

- **Approximation:** we want to design “good” spaces $\mathcal{F}$ to approximate f^* in high-dimension
    - Rather profound problem, on which we still struggle
- **Optimization:** how to design algorithms to solve the ERM in general
    - We essentally have ONE answer: Question Time: The Optimization Problem

The big questions

- **Approximation:** we want to design “good” spaces $\mathcal{F}$ to approximate f^* in high-dimension
  - Rather profound problem, on which we still struggle
- **Optimization:** how to design algorithms to solve the ERM in general
  - We essentally have ONE answer: Question Time: The Optimization Problem
  - Gradient Descent!

The Curse of Dimensionality

- How many samples do we need to estimate f^* depending on assumptions on its regularity?
- Question time: Curse of Dimensionality

The Curse of Dimensionality

- How many samples do we need to estimate f^* depending on assumptions on its regularity?
- Question time: Curse of Dimensionality
- f^* constant → 1 sample
- f^* linear $\rightarrow d$ samples
  - Space of functionals is $\mathcal{F} = \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R}; f(x) =< x, \theta > \right\} \simeq \mathbb{R}^d$ (isomorphic)
  - It's essentially like solving a system of linear equations for the linear form $< x_i, \theta^* >$
  - d equations, d degrees of freedom
- The reason why it's so easy is that linear functions are regular at a global level
  - Knowing the function locally tells us automatically the properties everywhere

Locally linear functions

- f^* locally linear, i.e. f^* is Lipschitz
  - $|f^*(x) - f^*(y)| \leq \beta \|x - y\|$
  - $Lip(f^*) = inf \left\{ \beta ; |f^*(x) - f^*(y)| \leq \beta \|x - y\| \text{ is true} \right\}$
  - Lip(f^*) is a measure of smoothness
- Space of functionals that are Lipschitz: $\mathcal{F} = \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R} ; f \text{ is Lipschitz} \right\}$
- We want a normed space to parameterize complexity, so we convert to a Banach space
  - $\gamma(f) := max(Lip(f), |f|_\infty)$
  - The parameterization of complexity **is** the Lipschitz constant

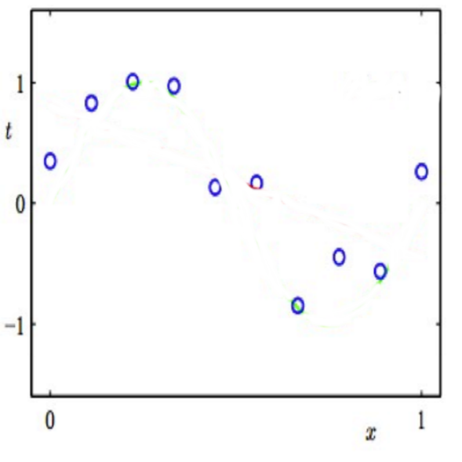
Formalization of the prediction problem

- $\forall \epsilon > 0$, find $f \in \mathcal{F}$ such that $\|f - f^*\| \leq \epsilon$ from n i.i.d. samples
  - n: sample complexity, “how many more samples to I need to make the error a given amount of times smaller”
- If f^* is Lipschitz, it can be demonstrated that $n \sim \epsilon^{-d}$
  - Upper bound: approximate f with its value in the closest of the sampled data points, find out expected error $\sim \epsilon^2$, upper bound is exponential
  - Lower bound: maximum discrepancy (the worst case scenario): unless you sample exponential number of data points, knowing $f(x_i)$ for all of them doesn’t let youwell approximate outside

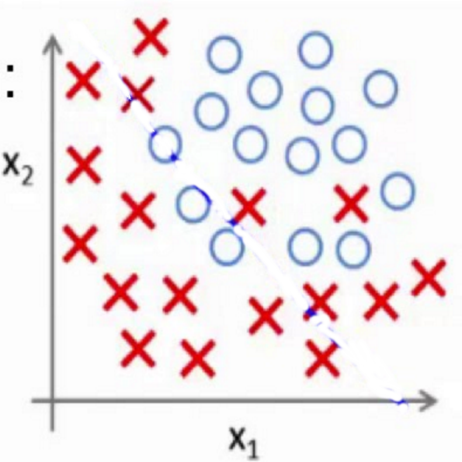
Enough of the math?

What’s the best function

To describe the data points?
(regression)



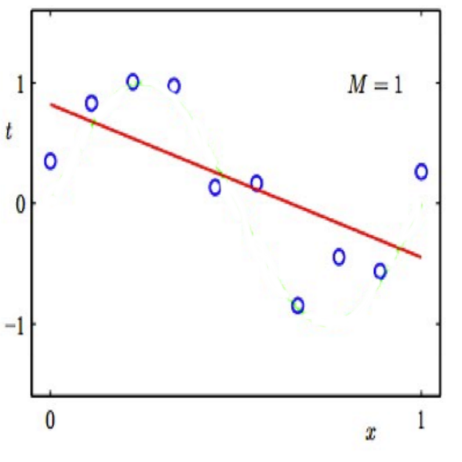
To separate into two classes?
(classification)



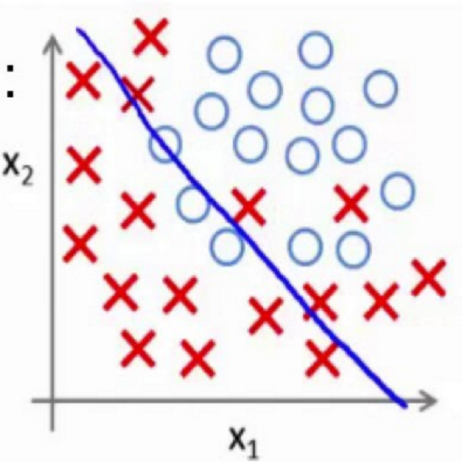
Images by Victor Lavrenko

What’s the best function

To describe the data points?
(regression)



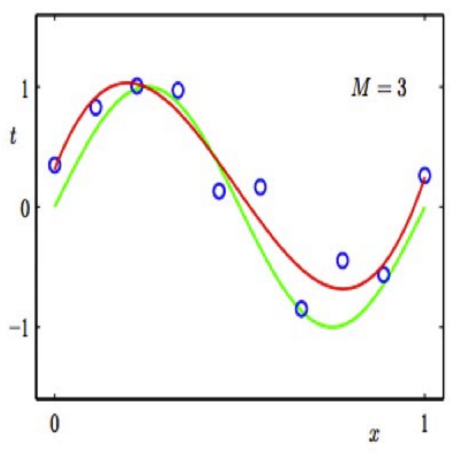
To separate into two classes?
(classification)



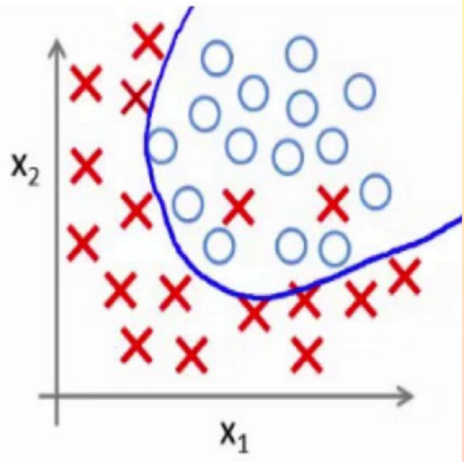
Images by Victor Lavrenko

What’s the best function

To describe the data points?
(regression)



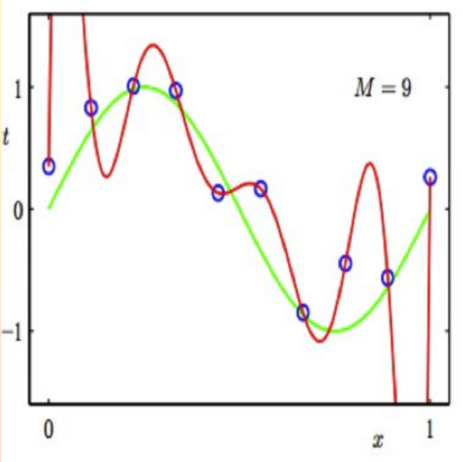
To separate into two classes?
(classification)



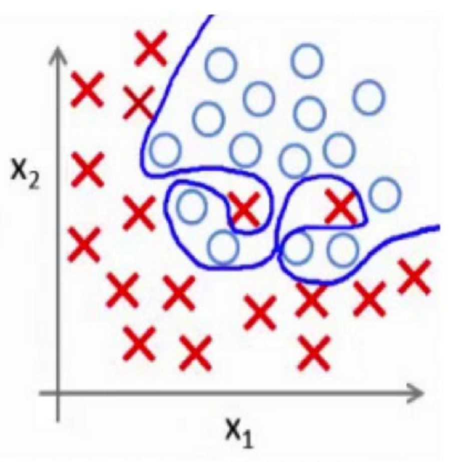
Images by Victor Lavrenko

What’s the best function

To describe the data points?
(regression)

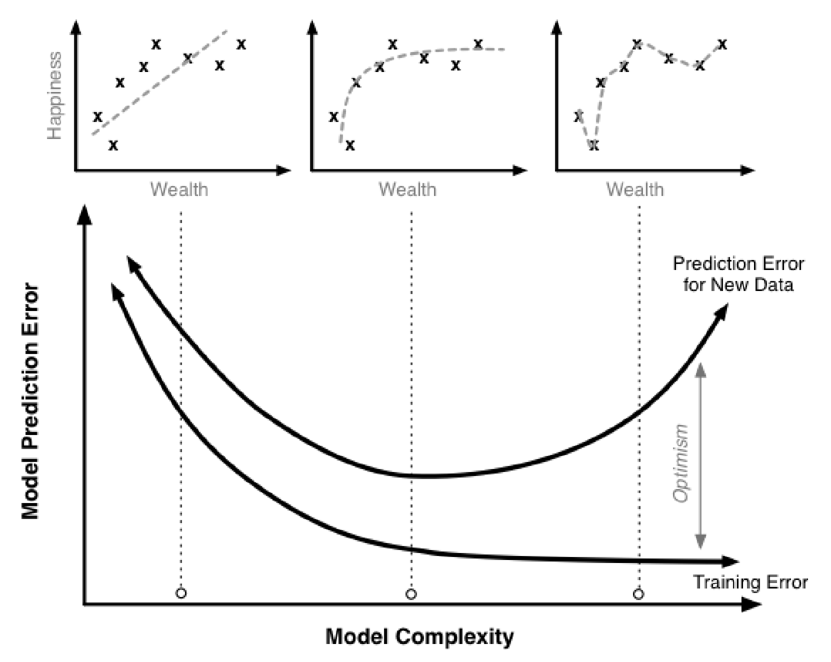


To separate into two classes?
(classification)



Images by Victor Lavrenko

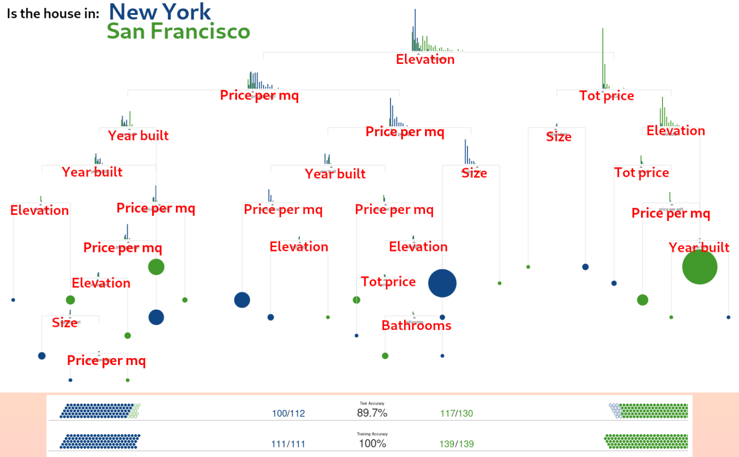
Avoid overtraining



From can't remember where

Simplest methods: Decision trees

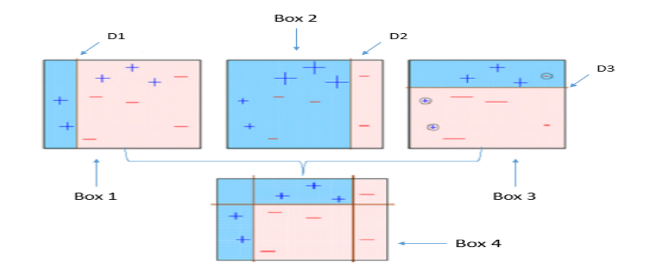
- Rather simple technique inspired by the standard approach of classifying events by selecting thresholds on several variables



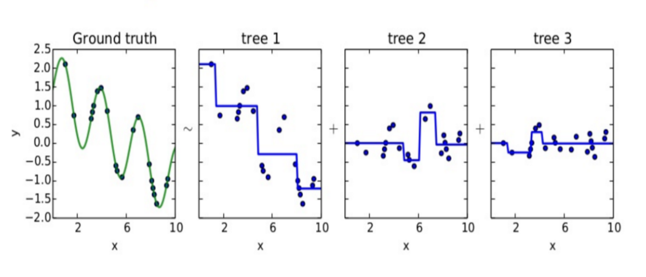
From http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Boosted decision trees)

- Ada(ptive)Boost: increase at each iteration the importance of events incorrectly classified in the previous iteration

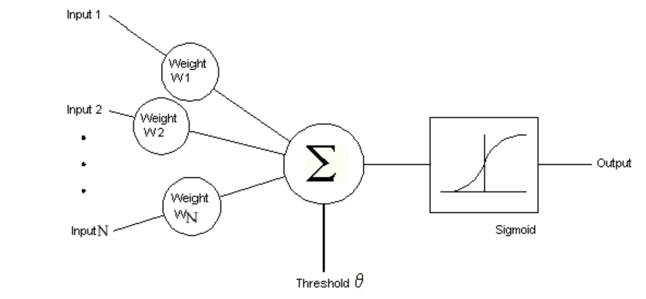


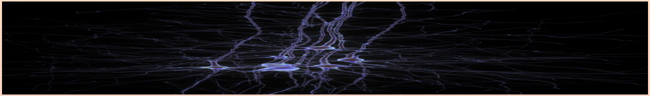
- GradientBoost: fit the new predictor to the residual errors of the previous one



Neurons

- Perceptron: simplest mathematical model of a neuron
  - Activation function provides nonlinearity in the response
  - A network of these can demonstrably approximate any (insert loose conditions here) function





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

Neural networks...

- Connecting neurons into a network
- Fully-connected networks: the most common a few decades ago
- Each weight is a free parameter that must be determined during the “training”

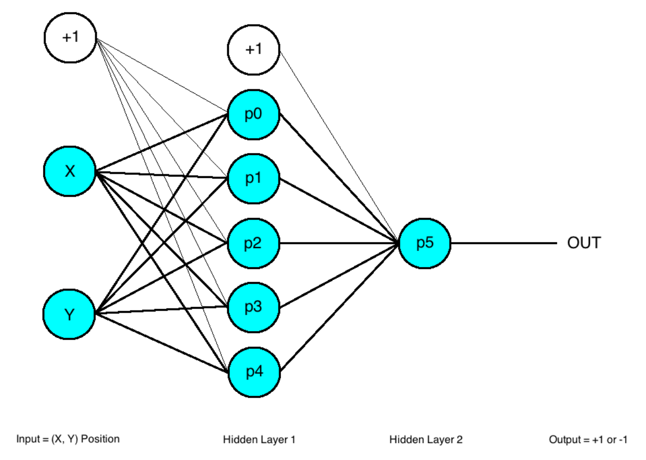
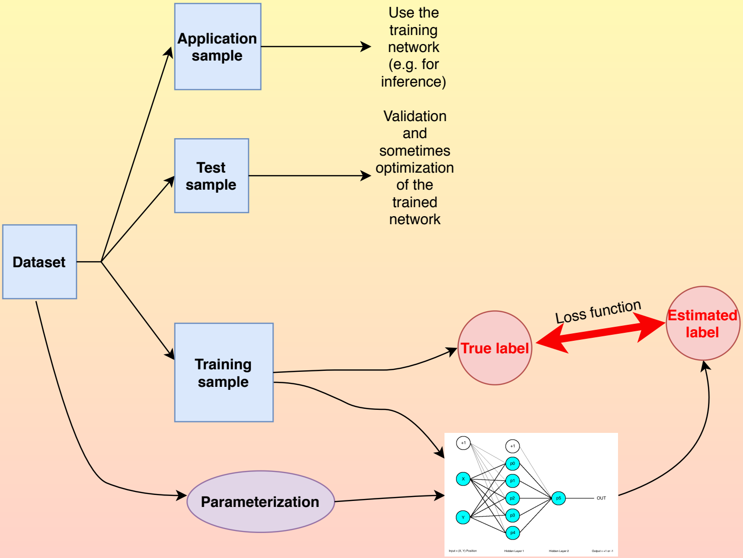


Image https://www.cs.utexas.edu/ teammco/misc/mlp/mlp.png

... and how to train them



Dataset

Application sample

Use the training network (e.g. for inference)

Test sample

Validation and sometimes optimization of the trained network

Training sample

True label

Loss function

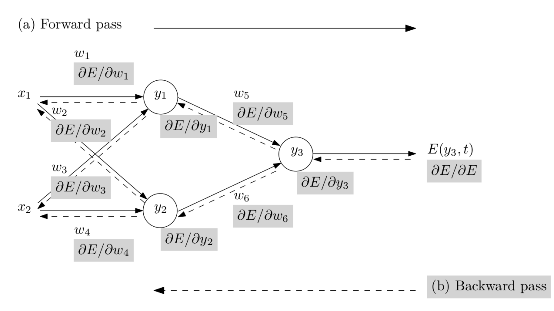
Estimated label

Parameterization

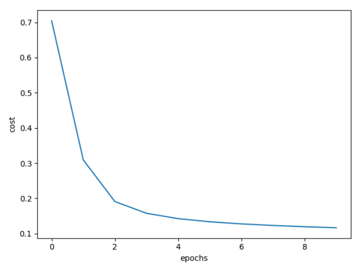
Image copyright Vischia, 2019

Loss function and backpropagation

- Adjust the parameters of each neuron and connection by backpropagating the difference between the estimated and the true output
- Differentiation and matrix (tensor) operations; dedicated software, automatic differentiation frameworks (e.g. tensorflow)
- Minimization of a cost (loss) function; the loss function can be tweaked to optimize w.r.t. several different objectives



(a) Forward pass
w_1
$\partial E/\partial w_1$
x_1
w_2
$\partial E/\partial w_2$
y_1
$\partial E/\partial y_1$
w_5
$\partial E/\partial w_5$
y_3
$\partial E/\partial y_3$
$E(y_3, t)$
$\partial E/\partial E$
w_3
$\partial E/\partial w_3$
x_2
y_2
$\partial E/\partial y_2$
w_6
$\partial E/\partial w_6$
w_4
$\partial E/\partial w_4$
(b) Backward pass



Images from Güneş Baydin et al, JMLR 18 (2018) 1–43 and http://www.adeveloperdiary.com

Make the training (mostly) possible

- In real problems, it's not guaranteed that a simple gradient descent can find $argmin(Loss)$
- Several techniques to help the process to happen

First look at your data! And sometimes preprocess them.

- It may be useful to transform the input features before feeding them to the network
  - Heavily problem-dependent, sometimes it helps sometimes not or hinders
- Imputation: if you have missing features, can replace them with some possibly meaningful values
  - Vastly delicate, a field of its own. Very risky, in HEP you don't usually need to do it
- Categorical variables: use numbers only if they have a meaning!
  - Number of jets: OK to use 0, 1, 2. Other categorical region (e.g. "is in signal region" vs "is in control region") must use one-hot encoding: [[1,0], [0, 1]]
- Range reduction: it is difficult for a ML algorithm to learn across vastly different ranges
  - Take the logarithm (useful in particular for the target labels)
  - Minmax scaling (transform each feature so that it is in a given range (e.g. from [0, 5000] to [0,1])
  - Standardization (transform each event x_i by $(x_i - \mu)/\sigma$), where μ and σ are mean and variance of x_i. Or other similar transformations)
- Rotate to privileged directions can help ML algorithms to avoid learning these privileged directions first
  - PCA: find a new basis of the space while minimizing average square distance of points to each vector of the basis (the individual dimensions of the data will be linearly uncorrelated)
  - This merely saves you learning cycles (the ML algoritm will eventually learn the transformation before learning actual characteristics of the features, it will just take longer)

Choose sampling scheme for loss and backpropagation

- Batch: compute on the whole training set (for large sets becomes too costly)
- Stochastic: compute on one sample (large noise, difficult to converge)
- Mini-batch: use a relatively small sample of data (tradeoff)

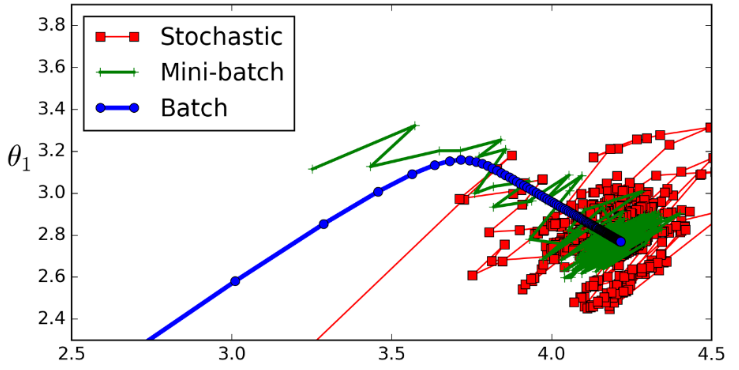
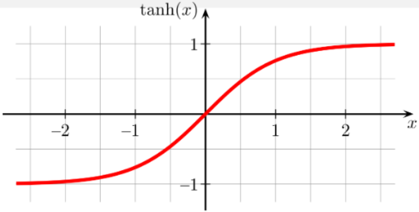
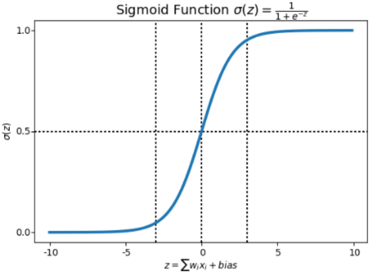


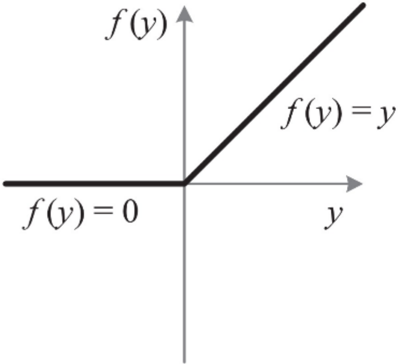
Image from a talk by W. Verbeke

Choose your activation function wisely

- *tanh* and *sigmoid* used a lot in the past
  - Seemed desirable to constrain neuron output to $[0, 1]$
  - For deep networks, vanishing gradients
  - *sigmoid* still used for output of the networks (outputs interpretable as probability)
- ReLU: a generally good choice for modern problems
  - Tricky cases may require variants

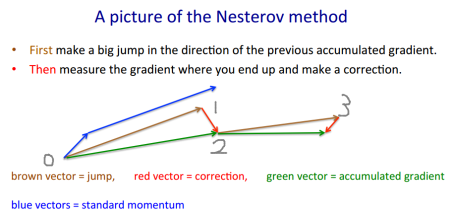






Improve algorithm to follow the gradient

- Mostly nonconvex optimization: very complicated problem, convergence in general not guaranteed
- Nesterov momentum: big jumps followed by correction seem to help!
- Adaptive moments: gradient steps decrease when getting closer to the minimum (avoids overshooting)



A picture of the Nesterov method

- First make a big jump in the direction of the previous accumulated gradient.
- Then measure the gradient where you end up and make a correction.

brown vector = jump, red vector = correction, green vector = accumulated gradient

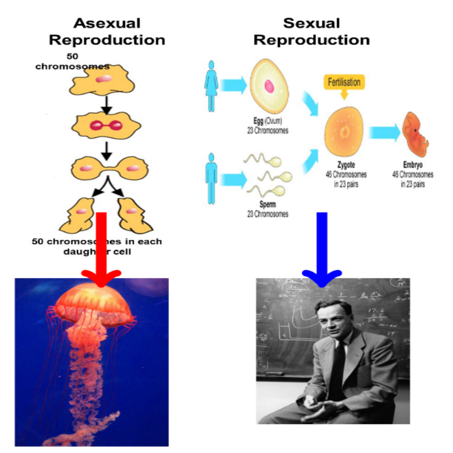
blue vectors = standard momentum

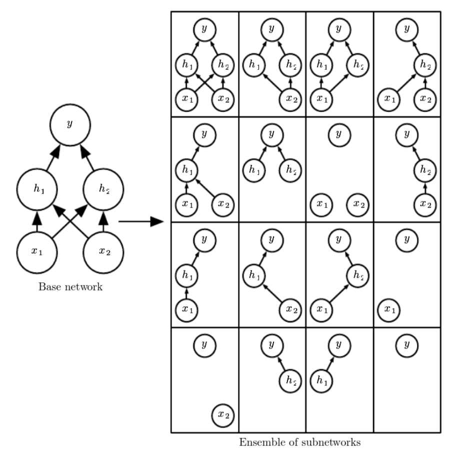
Nesterov Video

Diagram by Geoffrey Hinton, animation by Alec Radford

Regularization

- Batch normalization
  - Normalize (transform by $(x - \bar{x})/var(x)$) each input coming from previous layer over the (mini-)batch
  - Stabilizes response and reduces dependence among layers
- Dropoup: randomly shut down nodes in training
  - Avoids a weight to acquire too much importance
  - Inspired in genetics


Asexual Reproduction
50 chromosomes
50 chromosomes in each daughter cell
Sexual Reproduction
Egg (Ovum) 23 Chromosomes
Sperm 23 Chromosomes
Fertilisation
Zygote 46 Chromosomes in 23 pairs
Embryo 46 Chromosomes in 23 pairs



Images from a talk by W. Verbeke (likely originally #theInternet) and from Goodfellow-Bengio-Courville book

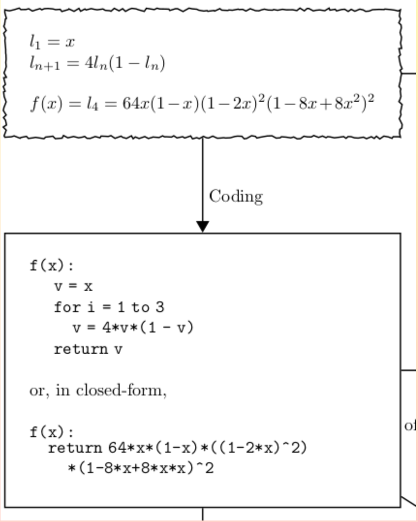
Computing derivatives in computer programs

1. Manual calculation, followed by explicit coding
2. Symbolic differentiation with expression manipulation (e.g. `Mathematica`)
3. Numerical differentiation with finite-difference approximations
4. Automatic (algorithmic) differentiation (AD): *autodiff*

- Question Time: Best Differentiation

Manual differentiation

- Manual calculation, followed by explicit coding
    - Time consuming and prone to error, require a closed-form model



$l_1 = x$
$l_{n+1} = 4l_n(1-l_n)$

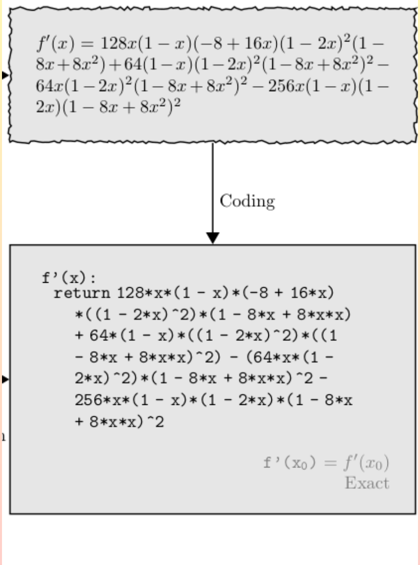
$f(x) = l_4 = 64x(1-x)(1-2x)^2(1-8x+8x^2)^2$

Coding

```
f(x):
  v = x
  for i = 1 to 3
    v = 4*v*(1 - v)
  return v
```

or, in closed-form,

```
f(x):
 return 64*x*(1-x)*((1-2*x)^2)
   *(1-8*x+8*x*x)^2
```



$f'(x) = 128x(1-x)(-8+16x)(1-2x)^2(1-8x+8x^2)+64(1-x)(1-2x)^2(1-8x+8x^2)^2-64x(1-2x)^2(1-8x+8x^2)^2-256x(1-x)(1-2x)(1-8x+8x^2)^2$

Coding

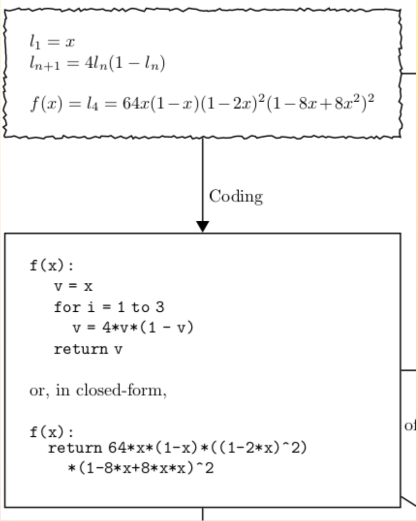
```
f'(x):
 return 128*x*(1 - x)*(-8 + 16*x)
   *((1 - 2*x)^2)*(1 - 8*x + 8*x*x)
   + 64*(1 - x)*((1 - 2*x)^2)*((1
   - 8*x + 8*x*x)^2) - (64*x*(1 -
   2*x)^2)*(1 - 8*x + 8*x*x)^2 -
   256*x*(1 - x)*(1 - 2*x)*(1 - 8*x
   + 8*x*x)^2
```

f'(x_0) $= f'(x_0)$
Exact

Image from Güneş Baydin et al, JMLR 18 (2018) 1–43

Symbolic differentiation

- Symbolic differentiation with expression manipulation (e.g. `Mathematica`, `Theano`)
    - Complex expressions, require a closed-form model
    - Sometimes can just minimize the problem without requiring derivative calculation
    - Nested duplications produce exponentially large symbolic expressions (*expression swell*, slow to evaluate)

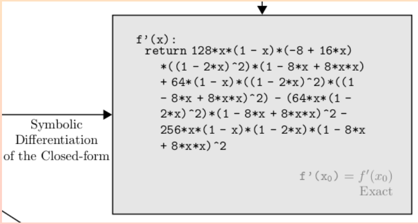


$$l_1 = x$$
$$l_{n+1} = 4l_n(1-l_n)$$
$$f(x) = l_4 = 64x(1-x)(1-2x)^2(1-8x+8x^2)^2$$

```
f(x):
  v = x
  for i = 1 to 3
    v = 4*v*(1 - v)
  return v
```

or, in closed-form,

```
f(x):
 return 64*x*(1-x)*((1-2x)^2)
   *(1-8*x+8*x*x)^2
```



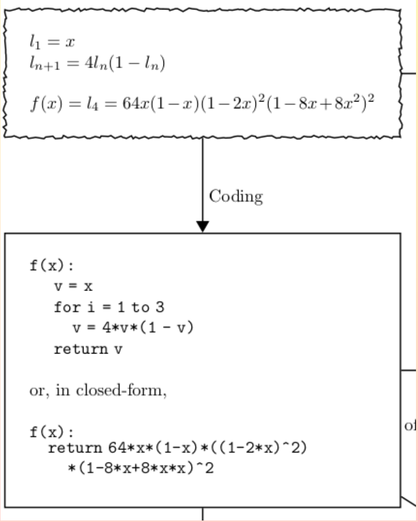
```
f'(x):
 return 128*x*(1 - x)*(-8 + 16*x)
   *((1 - 2*x)^2)*(1 - 8*x + 8*x*x)
   + 64*(1 - x)*((1 - 2*x)^2)*((1
   - 8*x + 8*x*x)^2) - (64*x*(1 -
   2*x)^2)*(1 - 8*x + 8*x*x)^2 -
   256*x*(1 - x)*(1 - 2*x)*(1 - 8*x
   + 8*x*x)^2
```

$f'(x_0) = f'(x_0)$
Exact

Image from Güneş Baydin et al, JMLR 18 (2018) 1–43

Standard numerical differentiation

- Numerical differentiation with finite-difference approximations
  - Rounding errors and truncation errors can make it very inaccurate
  - Mitigation techniques that cancel first-order errors are computationally costly
  - Accuracy must be traded off for performance for high dimensionalities



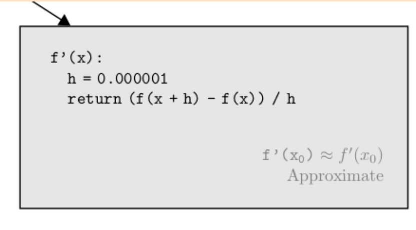
$l_1 = x$
$l_{n+1} = 4l_n(1 - l_n)$

$f(x) = l_4 = 64x(1-x)(1-2x)^2(1-8x+8x^2)^2$

```
f(x):
   v = x
   for i = 1 to 3
      v = 4*v*(1 - v)
   return v
```

or, in closed-form,

```
f(x):
  return 64*x*(1-x)*((1-2*x)^2)
     *(1-8*x+8*x*x)^2
```

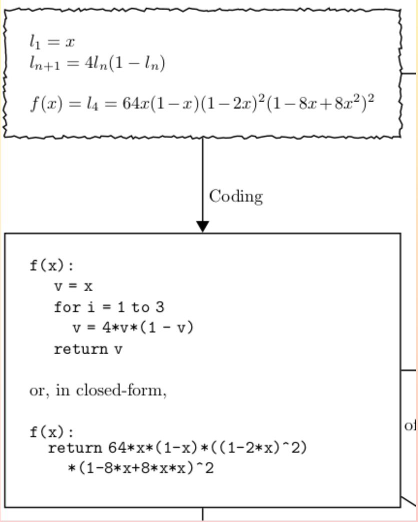


```
f'(x):
  h = 0.000001
  return (f(x + h) - f(x)) / h
```

Image from Güneş Baydin et al, JMLR 18 (2018) 1–43

Automatic differentiation

- Automatic (algorithmic) differentiation (AD): *autodiff*
  - Class of techniques to generate numerical derivative evaluations during code execution rather than derivative expressions
  - Accurate at machine precision with small constant overhead and asymptotic efficiency
  - No need to rearrange the code in a closed-form expression
  - Reverse AD generalizes the common chain-rule-based neural network backpropagation

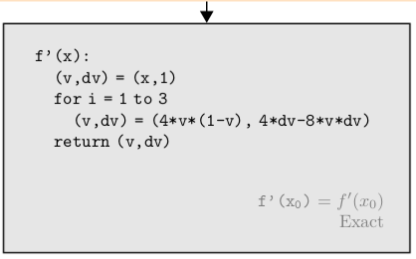


$$l_1 = x$$
$$l_{n+1} = 4l_n(1 - l_n)$$
$$f(x) = l_4 = 64x(1-x)(1-2x)^2(1-8x+8x^2)^2$$

```
f(x):
  v = x
  for i = 1 to 3
    v = 4*v*(1 - v)
  return v
```

or, in closed-form,

```
f(x):
 return 64*x*(1-x)*((1-2*x)^2)
   *(1-8*x+8*x*x)^2
```

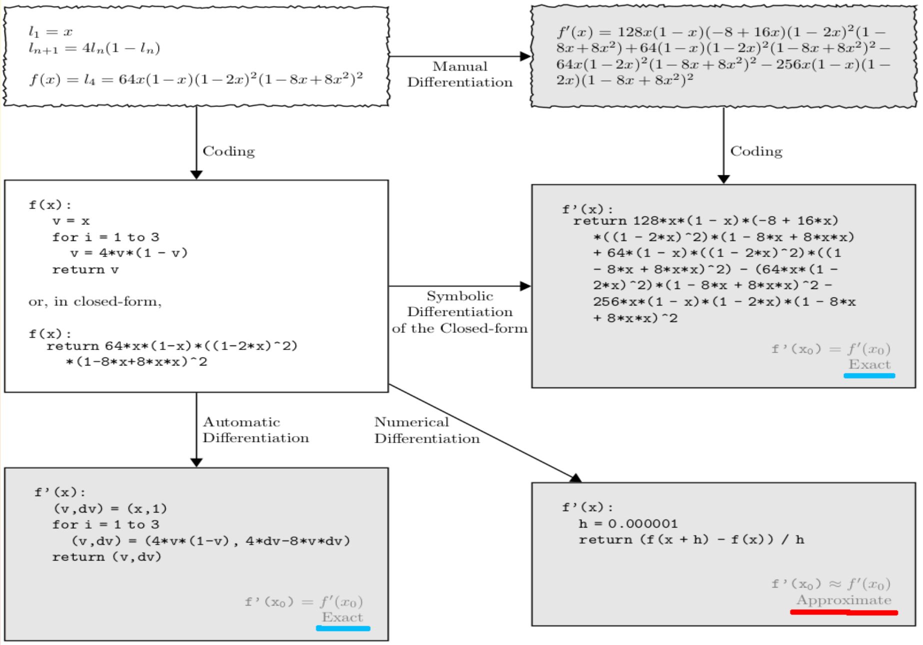


```
f'(x):
  (v,dv) = (x,1)
  for i = 1 to 3
    (v,dv) = (4*v*(1-v), 4*dv-8*v*dv)
  return (v,dv)
```

$\texttt{f'}(x_0) = f'(x_0)$
Exact

Image from Günes Baydin et al, JMLR 18 (2018) 1–43

The power of autodifferentiation



$l_1 = x$
$l_{n+1} = 4l_n(1-l_n)$

$f(x) = l_4 = 64x(1-x)(1-2x)^2(1-8x+8x^2)^2$

$f'(x) = 128x(1-x)(-8+16x)(1-2x)^2(1-8x+8x^2)+64(1-x)(1-2x)^2(1-8x+8x^2)^2-64x(1-2x)^2(1-8x+8x^2)^2-256x(1-x)(1-2x)(1-8x+8x^2)^2$

```
f(x):
   v = x
   for i = 1 to 3
     v = 4*v*(1 - v)
   return v
```

or, in closed-form,

```
f(x):
  return 64*x*(1-x)*((1-2*x)^2)
    *(1-8*x+8*x*x)^2
```

```
f'(x):
 return 128*x*(1 - x)*(-8 + 16*x)
   *((1 - 2*x)^2)*(1 - 8*x + 8*x*x)
   + 64*(1 - x)*((1 - 2*x)^2)*((1
   - 8*x + 8*x*x)^2) - (64*x*(1 -
   2*x)^2)*(1 - 8*x + 8*x*x)^2 -
   256*x*(1 - x)*(1 - 2*x)*(1 - 8*x
   + 8*x*x)^2
```

f'(x_0) $= f'(x_0)$
Exact

```
f'(x):
  (v,dv) = (x,1)
  for i = 1 to 3
    (v,dv) = (4*v*(1-v), 4*dv-8*v*dv)
  return (v,dv)
```

f'(x_0) $= f'(x_0)$
Exact

```
f'(x):
  h = 0.000001
  return (f(x + h) - f(x)) / h
```

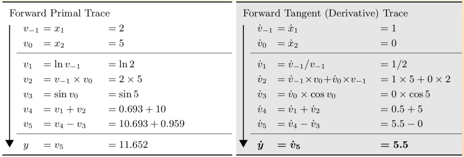
f'(x_0) $\approx f'(x_0)$
Approximate

Image from Güneş Baydin et al, JMLR 18 (2018) 1–43

The two modes of autodiff

Forward mode

- Associate with each intermediate v_i a derivative $\dot{v} = \frac{\partial v_i}{\partial x_1}$
- Apply the chain rule
- Single pass for $f : \mathbb{R} \to \mathbb{R}^n$
- n passes for $f : \mathbb{R}^n \to \mathbb{R}$

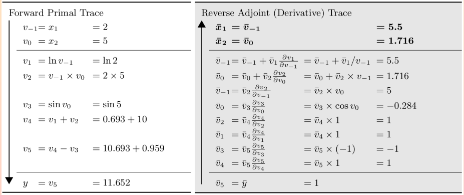


Forward Primal Trace		
v_{-1}	$= x_1$	$= 2$
v_0	$= x_2$	$= 5$
v_1	$= \ln v_{-1}$	$= \ln 2$
v_2	$= v_{-1} \times v_0$	$= 2 \times 5$
v_3	$= \sin v_0$	$= \sin 5$
v_4	$= v_1 + v_2$	$= 0.693 + 10$
v_5	$= v_4 - v_3$	$= 10.693 + 0.959$
y	$= v_5$	$= 11.652$

Forward Tangent (Derivative) Trace		
$\dot{v}_{-1}$	$= \dot{x}_1$	$= 1$
$\dot{v}_0$	$= \dot{x}_2$	$= 0$
$\dot{v}_1$	$= \dot{v}_{-1}/v_{-1}$	$= 1/2$
$\dot{v}_2$	$= \dot{v}_{-1} \times v_0 + \dot{v}_0 \times v_{-1}$	$= 1 \times 5 + 0 \times 2$
$\dot{v}_3$	$= \dot{v}_0 \times \cos v_0$	$= 0 \times \cos 5$
$\dot{v}_4$	$= \dot{v}_1 + \dot{v}_2$	$= 0.5 + 5$
$\dot{v}_5$	$= \dot{v}_4 - \dot{v}_3$	$= 5.5 - 0$
$\boldsymbol{\dot{y}}$	$\boldsymbol{= \dot{v}_5}$	$\mathbf{= 5.5}$

Reverse mode

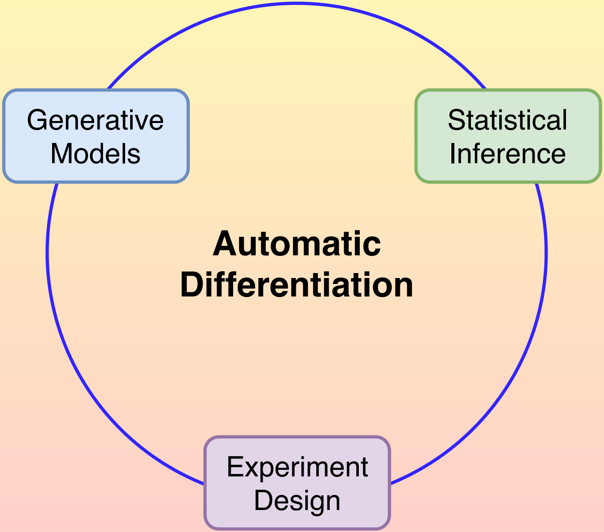
- Associate with each intermediate v_i an adjoint $\bar{v} = \frac{\partial y_j}{\partial v_i}$
- Run forwards and backwards as in backpropagation
- Single pass for $f : \mathbb{R}^n \to \mathbb{R}$ (functions with many inputs)
- Must store several values



Forward Primal Trace		
v_{-1}	$= x_1$	$= 2$
v_0	$= x_2$	$= 5$
v_1	$= \ln v_{-1}$	$= \ln 2$
v_2	$= v_{-1} \times v_0$	$= 2 \times 5$
v_3	$= \sin v_0$	$= \sin 5$
v_4	$= v_1 + v_2$	$= 0.693 + 10$
v_5	$= v_4 - v_3$	$= 10.693 + 0.959$
y	$= v_5$	$= 11.652$

Reverse Adjoint (Derivative) Trace			
$\boldsymbol{\bar{x}_1}$	$\boldsymbol{= \bar{v}_{-1}}$		$\mathbf{= 5.5}$
$\boldsymbol{\bar{x}_2}$	$\boldsymbol{= \bar{v}_0}$		$\mathbf{= 1.716}$
$\bar{v}_{-1}$	$= \bar{v}_{-1} + \bar{v}_1 \frac{\partial v_1}{\partial v_{-1}}$	$= \bar{v}_{-1} + \bar{v}_1/v_{-1}$	$= 5.5$
$\bar{v}_0$	$= \bar{v}_0 + \bar{v}_2 \frac{\partial v_2}{\partial v_0}$	$= \bar{v}_0 + \bar{v}_2 \times v_{-1}$	$= 1.716$
$\bar{v}_{-1}$	$= \bar{v}_2 \frac{\partial v_2}{\partial v_{-1}}$	$= \bar{v}_2 \times v_0$	$= 5$
$\bar{v}_0$	$= \bar{v}_3 \frac{\partial v_3}{\partial v_0}$	$= \bar{v}_3 \times \cos v_0$	$= -0.284$
$\bar{v}_2$	$= \bar{v}_4 \frac{\partial v_4}{\partial v_2}$	$= \bar{v}_4 \times 1$	$= 1$
$\bar{v}_1$	$= \bar{v}_4 \frac{\partial v_4}{\partial v_1}$	$= \bar{v}_4 \times 1$	$= 1$
$\bar{v}_3$	$= \bar{v}_5 \frac{\partial v_5}{\partial v_3}$	$= \bar{v}_5 \times (-1)$	$= -1$
$\bar{v}_4$	$= \bar{v}_5 \frac{\partial v_5}{\partial v_4}$	$= \bar{v}_5 \times 1$	$= 1$
$\bar{v}_5$	$= \bar{y}$	$= 1$	

Applications of autodiff computation



Generative Models

Statistical Inference

Automatic Differentiation

Experiment Design

Using deep neural networks is not always necessary

- Each realization of a machine learning algorithm has a certain complexity
- *Capacity* can be defined as the upper bound to the number of bits that can be stored in the network during learning
  - Transfer of (Fisher or Shannon) information from the training data to the weights of the synapses
- Sometimes the problem does not need the capacity of a neural network, and simpler algorithms are enough
  - Identifying true leptons from leptons produced in b hadron decays is an example

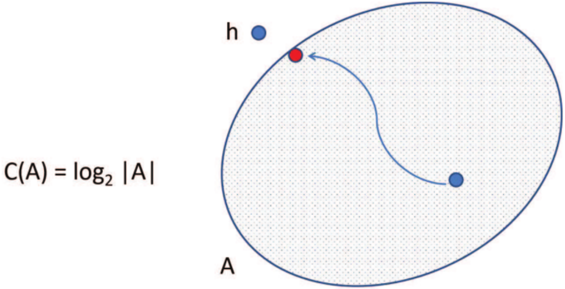


Figure 1. Learning framework where h is the function to be learnt and A is the available class of hypothesis or approximating functions. The cardinal capacity is the logarithm base two of the number, or volume, of the functions contained in A.

Plot from Baldi and Vershynin, arXiv:1901.00434

Separate prompt from fake leptons...

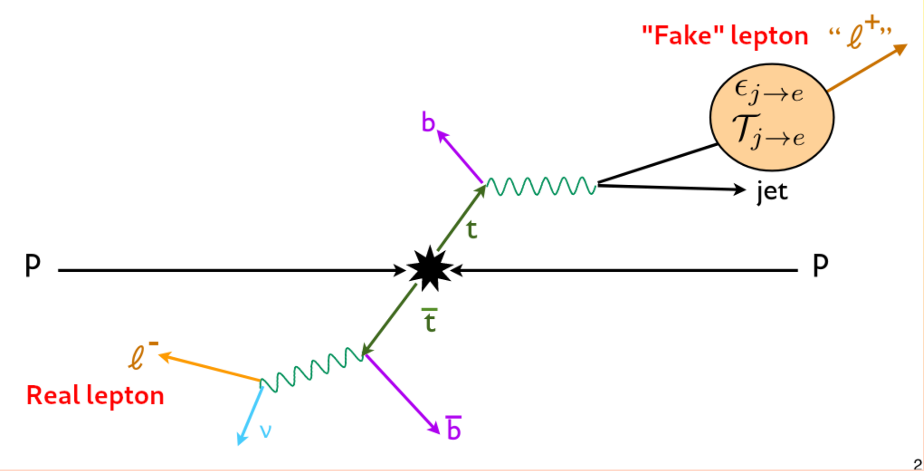
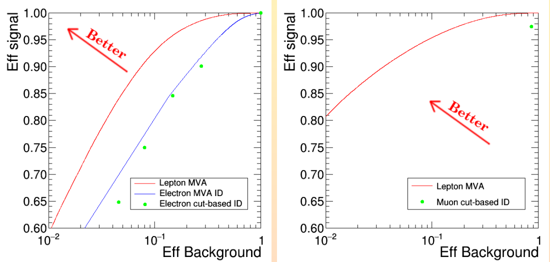


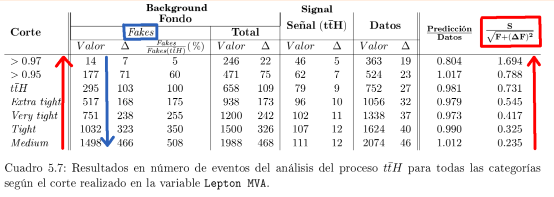
Image edited from David Curtin's talk at MC4BSM-2014

...is not very difficult

- Baseline algorithms: select particular ranges of discriminant observables
- BDT-based MVA ID improves substantially w.r.t baseline algorithms



Plots by S.S. Cruz

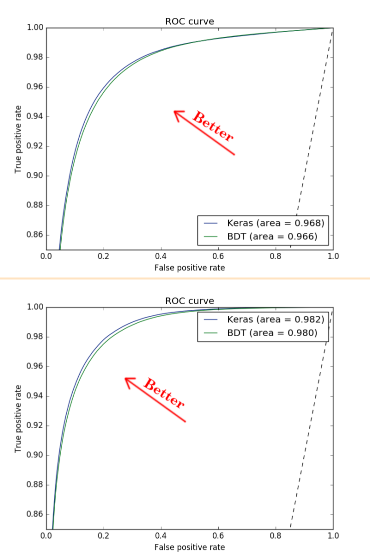


Corte	Background Fondo					Signal Señal ($t\bar{t}H$)		Datos		$\frac{\text{Predicción}}{\text{Datos}}$	$\frac{S}{\sqrt{F+(\Delta F)^2}}$
	Fakes			Total							
	Valor	Δ	$\frac{Fakes}{Fakes(t\bar{t}H)}$ (%)	*Valor*	Δ	*Valor*	Δ	*Valor*	Δ		
> 0.97	14	7	5	246	22	46	5	363	19	0.804	1.694
> 0.95	177	71	60	471	75	62	7	524	23	1.017	0.788
$t\bar{t}H$	295	103	100	658	109	79	9	752	27	0.981	0.731
Extra tight	517	168	175	938	173	96	10	1056	32	0.979	0.545
Very tight	751	238	255	1200	242	102	11	1338	37	0.973	0.417
Tight	1032	323	350	1500	326	107	12	1624	40	0.990	0.325
Medium	1498	466	508	1988	468	111	12	2074	46	1.012	0.235

Cuadro 5.7: Resultados en número de eventos del análisis del proceso $t\bar{t}H$ para todas las categorías según el corte realizado en la variable `Lepton MVA`.

Table by Víctor Rodríguez Bouza

- Deep neural network (DNN) does not help much w.r.t. BDT



Plots by Antonio Márquez García

Sometimes complexity is not the main point

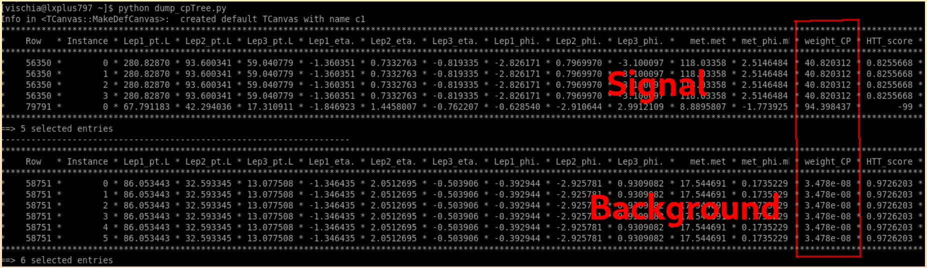
Neural networks can approximate any continuous real-valued function

- A feed-forward network with sigmoid activation functions can approximate any continuos real-valued function. Cybenko, G. (1989)
- Any failure in mapping a function comes from inadequate choice of weights or insufficient number of neurons. Hornik et al (1989), Funahashi (1989)
- Derivatives can be approximated as well as the functions, even in case of non-differentiability (e.g. piecewise differentiable functions). Hornik et al (1990)
- These results are valid even with other classes of activation functions. Light (1992), Stinchcombe and White (1989), Baldi (1991), Ito (1991), etc

Neural networks can be used to build fully invertible models

- The backpropagation algorithm is a special case of *automatic differentiation*
- A fully invertible model is a powerful tool that can be used for many frontier applications in particle physics

Difficult tasks for humans may be easy for artificial networks



```
[vischia@lxplus797 ~]$ python dump_cpTree.py
Info in <TCanvas::MakeDefCanvas>:  created default TCanvas with name c1
*************************************************************************************************************************************************************************************************
*    Row   * Instance * Lep1_pt.L * Lep2_pt.L * Lep3_pt.L * Lep1_eta. * Lep2_eta. * Lep3_eta. * Lep1_phi. * Lep2_phi. * Lep3_phi. *   met.met * met_phi.m * weight_CP * HTT_score *
*************************************************************************************************************************************************************************************************
*    56350 *        0 * 280.82870 * 93.600341 * 59.040779 * -1.360351 * 0.7332763 * -0.819335 * -2.826171 * 0.7969970 * -3.100097 * 118.03358 * 2.5146484 * 40.820312 * 0.8255668 *
*    56350 *        1 * 280.82870 * 93.600341 * 59.040779 * -1.360351 * 0.7332763 * -0.819335 * -2.826171 * 0.7969970 * -3.100097 * 118.03358 * 2.5146484 * 40.820312 * 0.8255668 *
*    56350 *        2 * 280.82870 * 93.600341 * 59.040779 * -1.360351 * 0.7332763 * -0.819335 * -2.826171 * 0.7969970 * -3.100097 * 118.03358 * 2.5146484 * 40.820312 * 0.8255668 *
*    56350 *        3 * 280.82870 * 93.600341 * 59.040779 * -1.360351 * 0.7332763 * -0.819335 * -2.826171 * 0.7969970 * -3.100097 * 118.03358 * 2.5146484 * 40.820312 * 0.8255668 *
*    79791 *        0 * 67.791183 * 42.294036 * 17.310911 * -1.846923 * 1.4458007 * -0.762207 * -0.628540 * -2.910644 * 2.9912109 * 8.8895807 * -1.773925 * 94.398437 *       -99 *
*************************************************************************************************************************************************************************************************
==> 5 selected entries
--------------------------------------------------------------------
*************************************************************************************************************************************************************************************************
*    Row   * Instance * Lep1_pt.L * Lep2_pt.L * Lep3_pt.L * Lep1_eta. * Lep2_eta. * Lep3_eta. * Lep1_phi. * Lep2_phi. * Lep3_phi. *   met.met * met_phi.m * weight_CP * HTT_score *
*************************************************************************************************************************************************************************************************
*    58751 *        0 * 86.053443 * 32.593345 * 13.077508 * -1.346435 * 2.0512695 * -0.503906 * -0.392944 * -2.925781 * 0.9309082 * 17.544691 * 0.1735229 * 3.478e-08 * 0.9726203 *
*    58751 *        1 * 86.053443 * 32.593345 * 13.077508 * -1.346435 * 2.0512695 * -0.503906 * -0.392944 * -2.925781 * 0.9309082 * 17.544691 * 0.1735229 * 3.478e-08 * 0.9726203 *
*    58751 *        2 * 86.053443 * 32.593345 * 13.077508 * -1.346435 * 2.0512695 * -0.503906 * -0.392944 * -2.925781 * 0.9309082 * 17.544691 * 0.1735229 * 3.478e-08 * 0.9726203 *
*    58751 *        3 * 86.053443 * 32.593345 * 13.077508 * -1.346435 * 2.0512695 * -0.503906 * -0.392944 * -2.925781 * 0.9309082 * 17.544691 * 0.1735229 * 3.478e-08 * 0.9726203 *
*    58751 *        4 * 86.053443 * 32.593345 * 13.077508 * -1.346435 * 2.0512695 * -0.503906 * -0.392944 * -2.925781 * 0.9309082 * 17.544691 * 0.1735229 * 3.478e-08 * 0.9726203 *
*    58751 *        5 * 86.053443 * 32.593345 * 13.077508 * -1.346435 * 2.0512695 * -0.503906 * -0.392944 * -2.925781 * 0.9309082 * 17.544691 * 0.1735229 * 3.478e-08 * 0.9726203 *
*************************************************************************************************************************************************************************************************
==> 6 selected entries
```

Signal

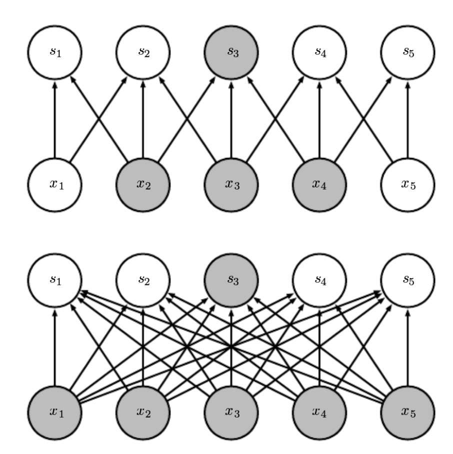
Background

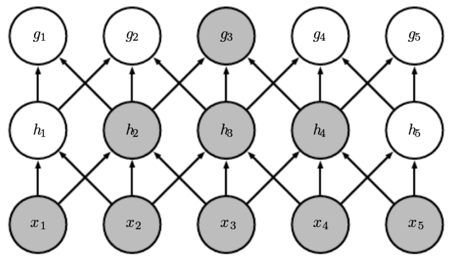
Image by Pietro Vischia

Easy tasks for humans may be very difficult for artificial networks

Image from indiatimes.com

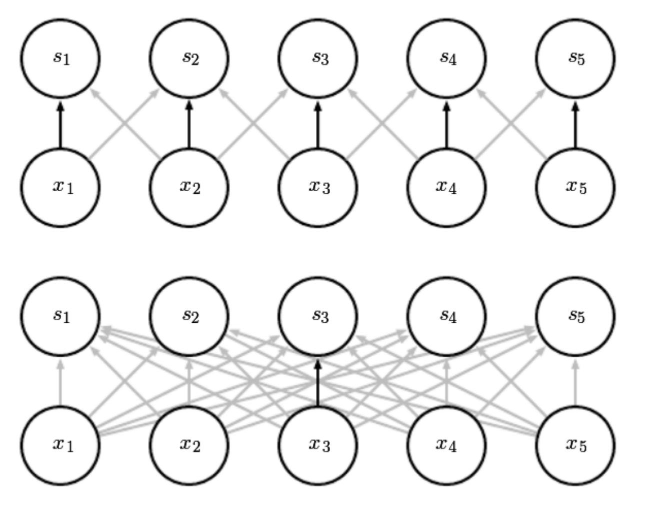
Sparse connectivity and receptive fields


s_1 s_2 s_3 s_4 s_5
x_1 x_2 x_3 x_4 x_5
s_1 s_2 s_3 s_4 s_5
x_1 x_2 x_3 x_4 x_5


g_1 g_2 g_3 g_4 g_5
h_1 h_2 h_3 h_4 h_5
x_1 x_2 x_3 x_4 x_5

Images from https://www.deeplearningbook.org/

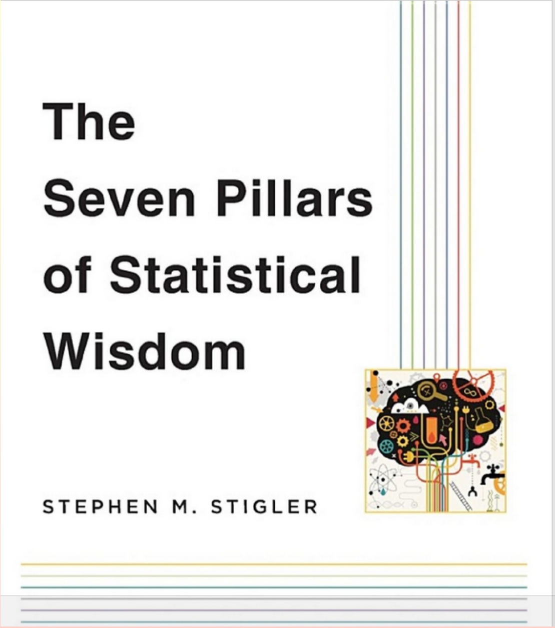
Parameter sharing and equivariance to translations



Images from https://www.deeplearningbook.org/

The first pillar

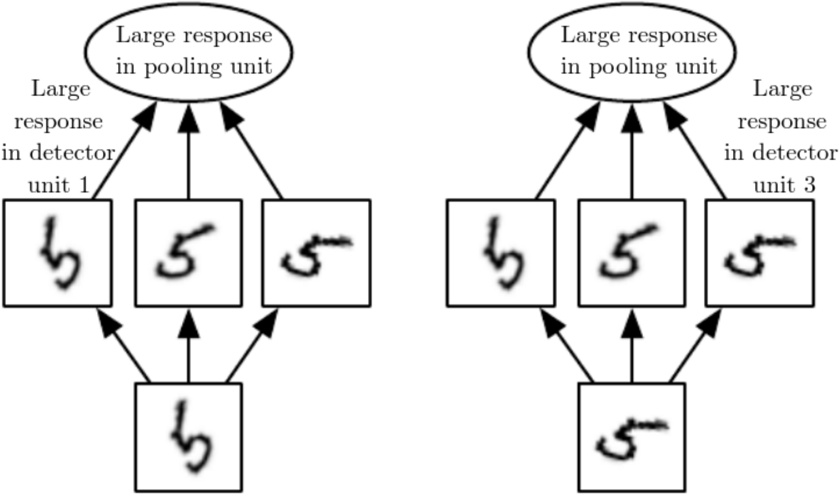
- **Aggregation**
- Information
- Likelihood
- Intercomparison
- Regression
- Design
- Residual



The Seven Pillars of Statistical Wisdom

STEPHEN M. STIGLER

Pooling and invariance for generic transformations



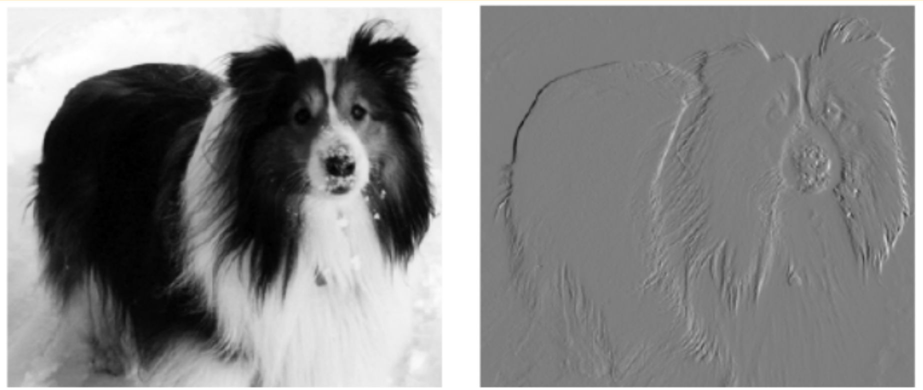
Large response in pooling unit
Large response in detector unit 1
Large response in pooling unit
Large response in detector unit 3

Images from https://www.deeplearningbook.org/

Convolution makes it easier to learn transformations

- Standard fully connected network: 8 billion matrix entries, 16 billion floating-point operations
- Convolutional network: 2 matrix entries, 267960 floating-point operations
  - 4 billion times more efficient in representing the transformation
  - 60000 times more efficient computationally

Edge detection



Images from https://www.deeplearningbook.org/

LeNet

- LeNet (Yann LeCun 1998, http://yann.lecun.com/exdb/lenet/)

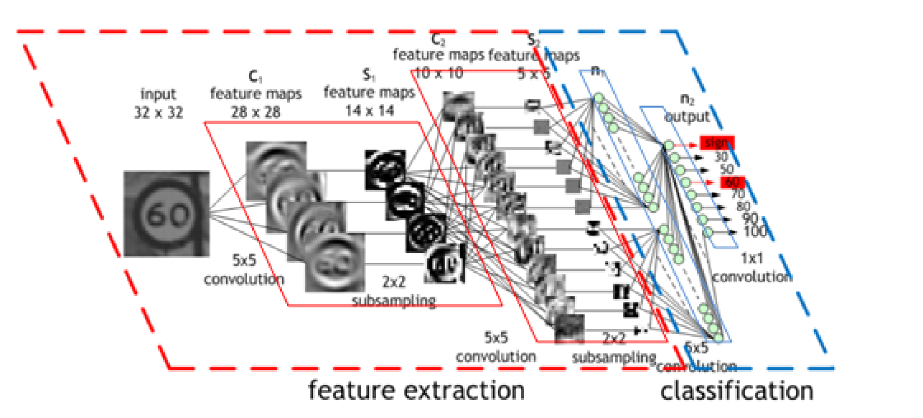
LeNet

LeNet

- LeNet (Yann LeCun 1998, http://yann.lecun.com/exdb/lenet/)

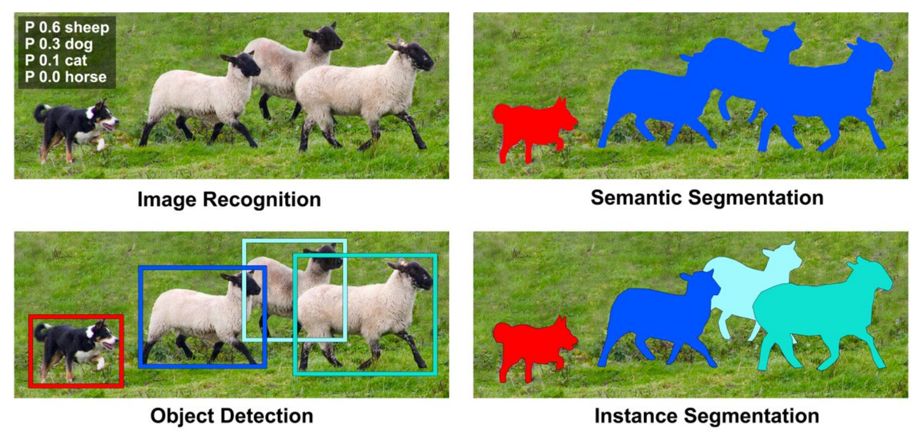
LeNet

Learning in stages



From http://parse.ele.tue.nl/education/cluster0

Segmentating objects



Reconstructing jets

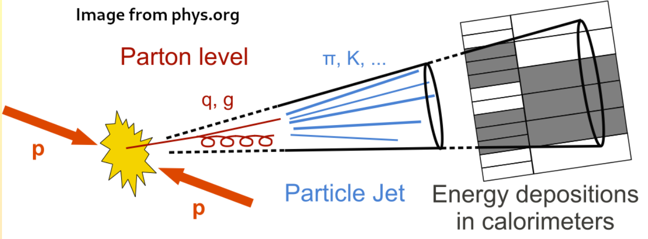
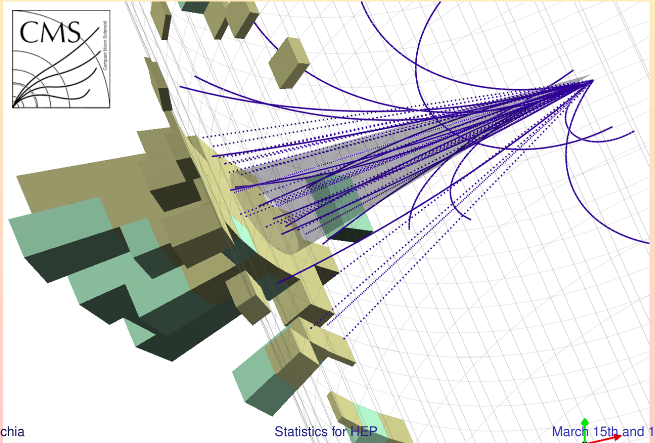
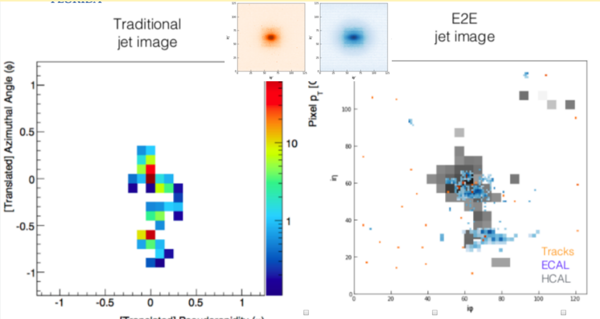


Image from phys.org

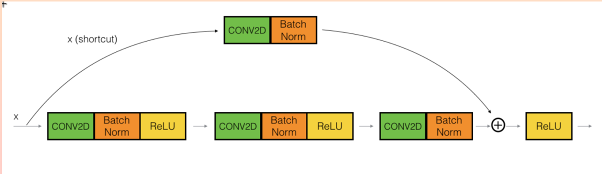


End-to-end jet reconstruction

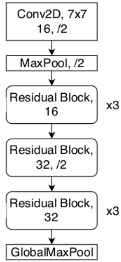
- Build images by projecting different layers into a single one
- Treat the result as an image with Res(idual)Net(works)
- Role of tracks in jet reco from network matches physics we know



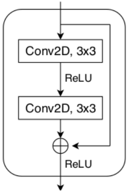
Andrews et al., 2019



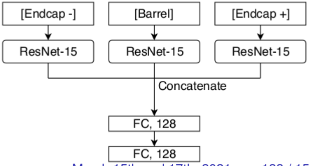
X_shortcut goes through convolution block



(a) ResNet-15



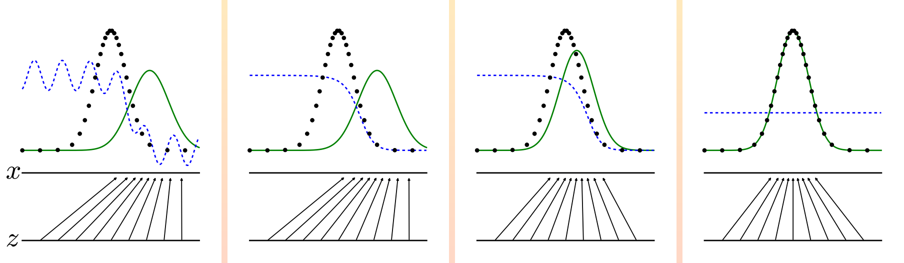
(b) The Residual block with skip connection.



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

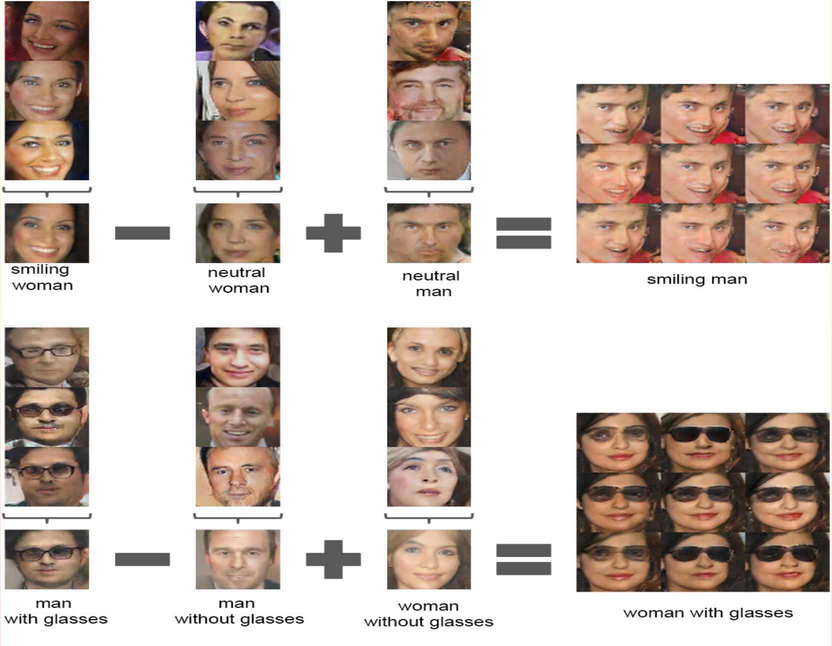
Unsupervised learning (no data label is given)

- Train two networks
- Green network: tries to capture the shape of the data
- Blue network: estimates the probability that an event comes from data rather than the green network
- Strategy: Green tries to fool Blue
  (Javier C. says: Green is Barcelona FC, Blue is Real Madrid)



From https://arxiv.org/abs/1406.2661

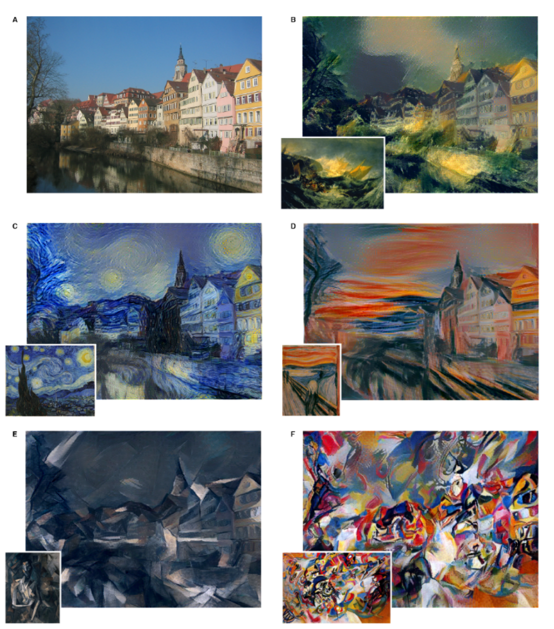
Can pick elements and combine them into new images



From https://github.com/Newmu/dcgan_code/blob/master/images/faces_arithmetic_collage.png

If you can write a loss function for it, you can learn it

- C = mathematical representation of **content**
- S = mathematical representation of **style**
- Loss = distance[S(reference) - S(generated image)
  + distance[C(original image) - C(generated image)



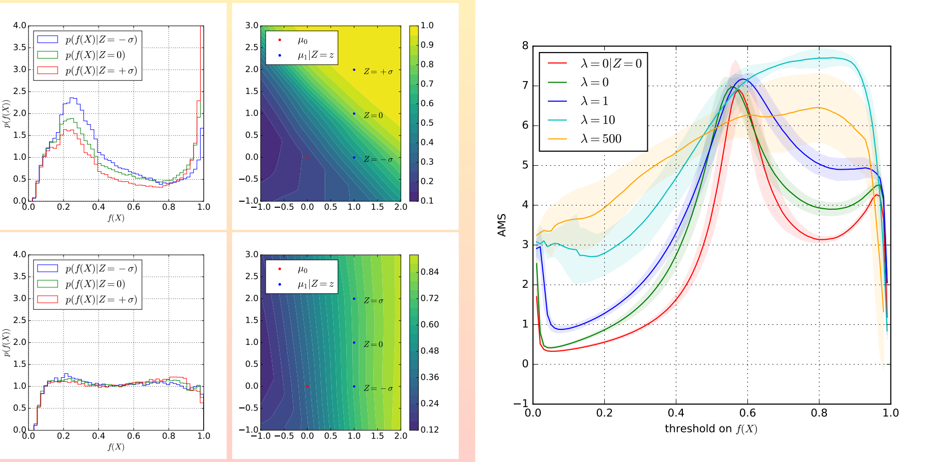
From https://arxiv.org/abs/1508.06576

This person does not exist!

From https://thispersondoesnotexist.com/: try it out!

Reduce the impact of systematic uncertainties on our results

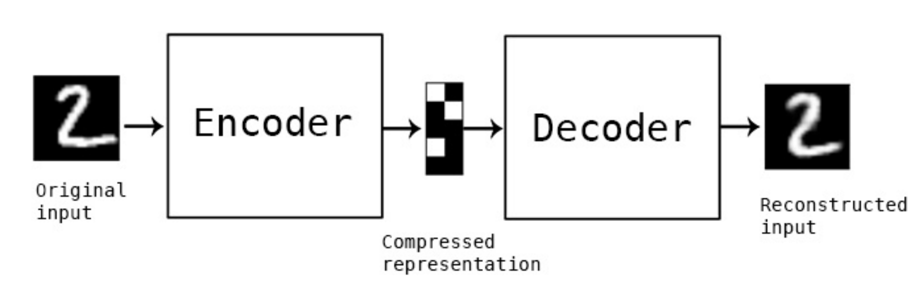
- Adversarial networks used to build pivot quantities
  - Quantities that are invariant in some parameter (typically a nuisance parameter representing a source of uncertainty)
- Best Approximate Mean Significance as tradeoff optimal/pivotal
  $E_\lambda(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) -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



From Louppe-Kagan-Cranmer, arXiv:1611.01046

Autoencoders...

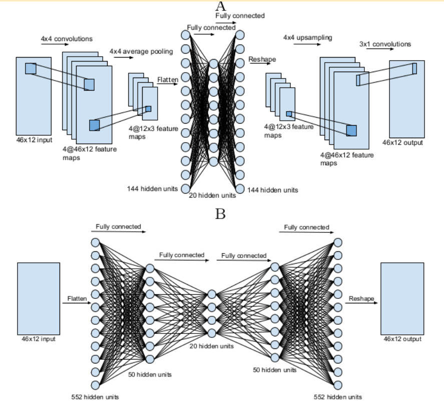
- Learn how to transform an object into almost itself

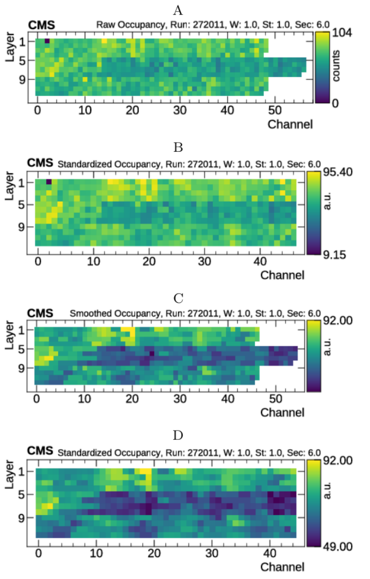


From Chollet and Allaire, Deep Learning With R

...in physics

- Use it to spot objects that are different from those you have trained on
- CMS Muon Chamber detectors modelled as geographic layered maps
  - Map is an image: use **convolutional** autoencoders
  - Local approach (independent layers): spot anomalies in a layer
  - Regional approach (simultaneusly across the layers): spot intra-chamber issues

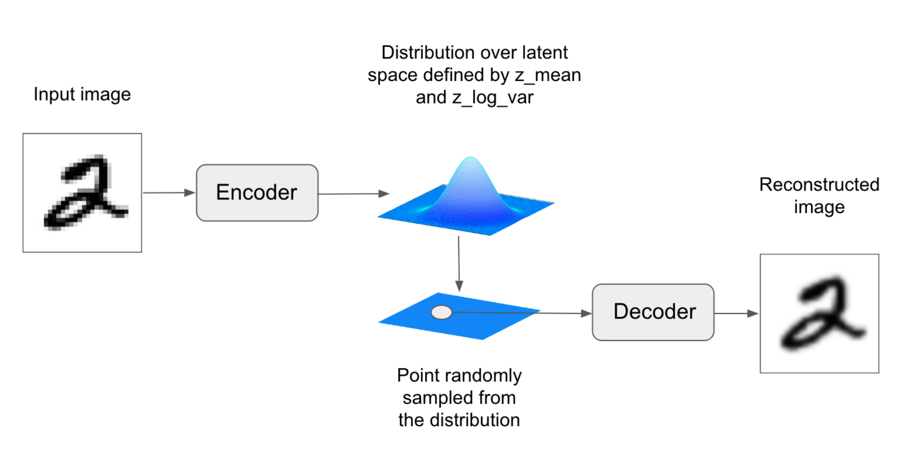




From arXiv:1808.00911

...and Variational autoencoders

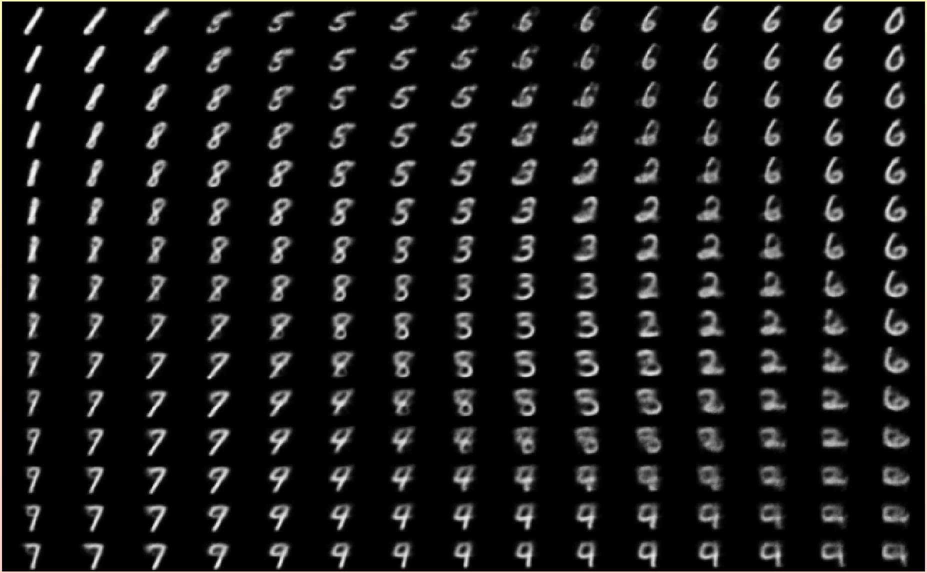
- Learn a space of continous representations of the inputs


Input image
Encoder
Distribution over latent space defined by z_mean and z_log_var
Point randomly sampled from the distribution
Decoder
Reconstructed image

From Chollet and Allaire, Deep Learning With R

...and Variational autoencoders

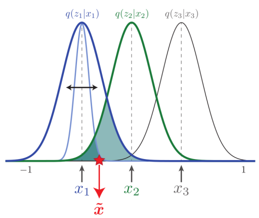
- "How do I transform a 1 into a 0?"
- Space directions have a meaning! "four-ness", "one-ness"

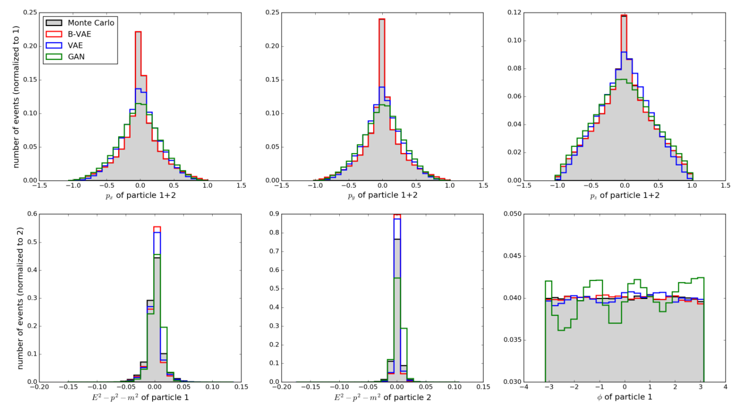


From Chollet and Allaire, Deep Learning With R

...also in physics

- Fast generation of collision events in a multidimensional phase space
- Balancing goodness-of-reconstruction and overlap in latent space
  - B-VAE $Loss = \frac{1}{M}\sum_{i=1}^{M}(1-B)\cdot MSE + B\cdot D_{KL}$.
- Works better than a GAN!



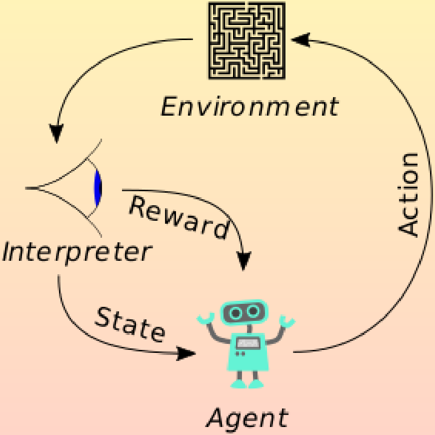


Plots from arXiv:1804.03599 and arXiv:1901.00875

We figure out images

- What about adding a time component?
- A single network is not complex enough for driving a car
- What if we permit a network to modify itself?

- Reinforcement Learning
- “Q” is the letter denoting the reward function for an action



By Megajuice - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=57895741

...is what you do to train your pets

From The Auckland Dog Coach

From videogames...

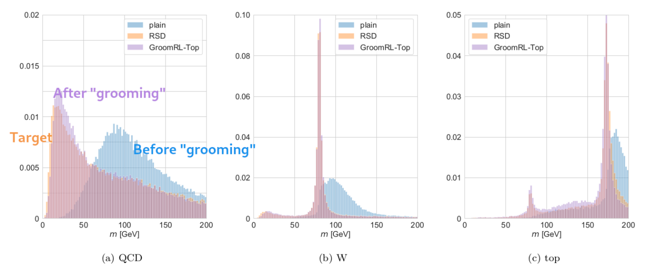
- ATARI Blackout (Google Deep Mind)
- *https://deepmind.com/research/dqn/*

...to self driving cars...

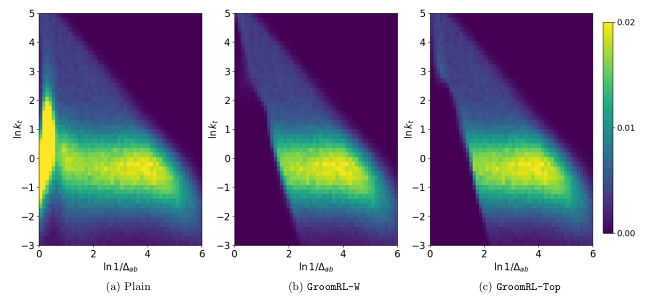
- *https://www.youtube.com/watch?v=MqUbdd7ae54*
- Build your own simulated driver: http://selfdrivingcars.mit.edu/deeptraffic/

...to physics

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



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

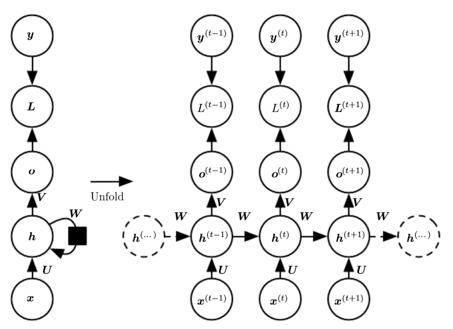


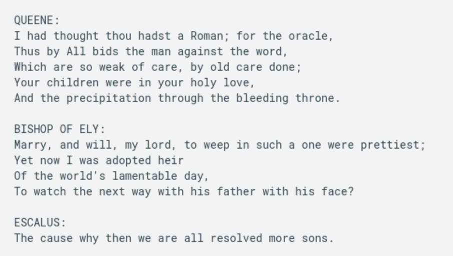
(a) Plain (b) GroomRL-W (c) GroomRL-Top

Images from arXiv:1903.09644

Learn sequences

- Recurrent architectures insert a “time” component: learn sequences!
  - In general a dimension that is supposed to be ordered (time, position of words in a sentence, etc)
- Can even learn how to generate Shakespearian text
  - With Markov Chains, the results are rather worse: https://amva4newphysics.wordpress.com/2016/09/20/hermione-had-become-a-bit-pink/



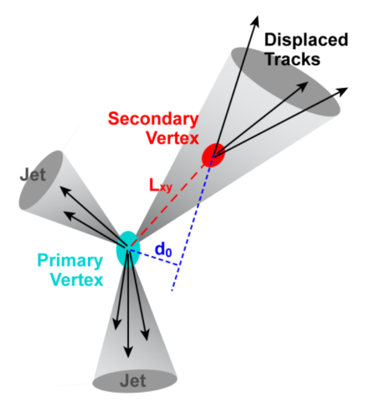


```
QUEENE:
I had thought thou hadst a Roman; for the oracle,
Thus by All bids the man against the word,
Which are so weak of care, by old care done;
Your children were in your holy love,
And the precipitation through the bleeding throne.

BISHOP OF ELY:
Marry, and will, my lord, to weep in such a one were prettiest;
Yet now I was adopted heir
Of the world's lamentable day,
To watch the next way with his father with his face?

ESCALUS:
The cause why then we are all resolved more sons.
```

From https://www.deeplearningbook.org/ and https://www.tensorflow.org/tutorials/text/text_generation

Identifying jets from b quarks...

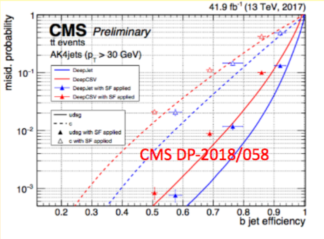
- Quarks produced in proton-proton collisions give rise to collimated “jets” of particles
- Bottom quarks travel for a while before fragmenting into jets

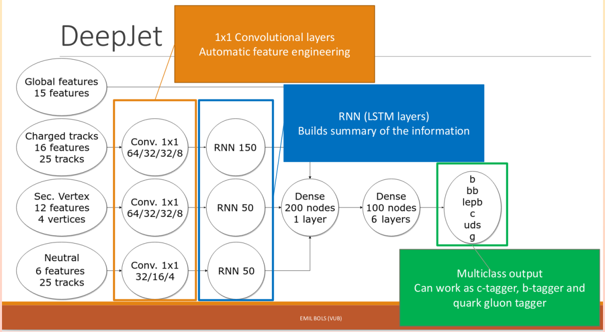


Plot from D0

...requires combining image and sequential processing!

- b tagging at CMS
  - CSV (Run I and early Run II): BDT sensitive to secondary vertexes
- DeepCSV: similar inputs, generic DNN
- Domain knowledge can inform the representation used!
  - Leading criterion for choice of technique for the classifier
- What is the best representation for jets?
  - Convolutional networks for images
  - Particle-based structure





DeepJet

1x1 Convolutional layers
Automatic feature engineering

Global features
15 features

Charged tracks
16 features
25 tracks

Sec. Vertex
12 features
4 vertices

Neutral
6 features
25 tracks

Conv. 1x1
64/32/32/8

Conv. 1x1
64/32/32/8

Conv. 1x1
32/16/4

RNN 150

RNN 50

RNN 50

RNN (LSTM layers)
Builds summary of the information

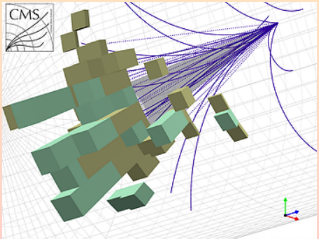
Dense
200 nodes
1 layer

Dense
100 nodes
6 layers

b
bb
lepb
c
uds
g

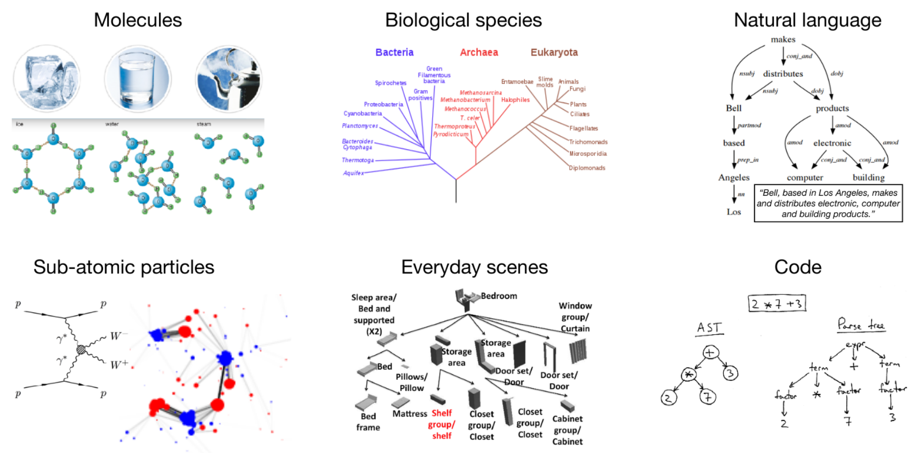
Multiclass output
Can work as c-tagger, b-tagger and quark gluon tagger

EMIL BOLS (VUB)



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

Learn graph of the underlying structure...



"Bell, based in Los Angeles, makes and distributes electronic, computer and building products."

From Peter Battaglia's talk at the IML2020 Workshop

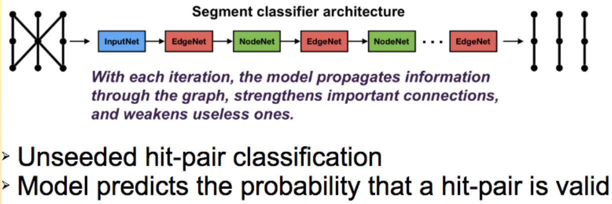
...until the structure is learned

Water

Video from https://sites.google.com/view/learning-to-simulate

Tracking

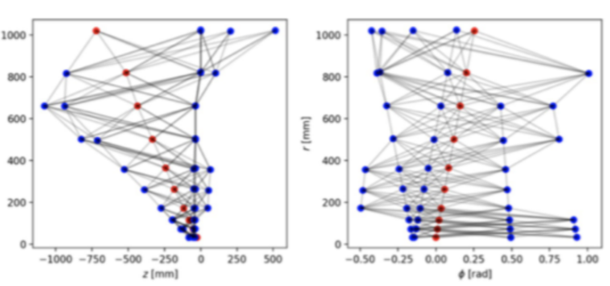
- Graph networks to literally connect the dots

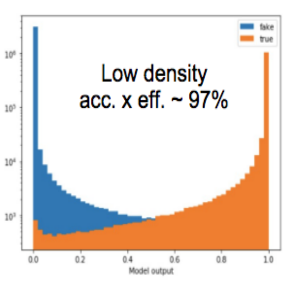


Segment classifier architecture

With each iteration, the model propagates information through the graph, strengthens important connections, and weakens useless ones.

- Unseeded hit-pair classification
- Model predicts the probability that a hit-pair is valid



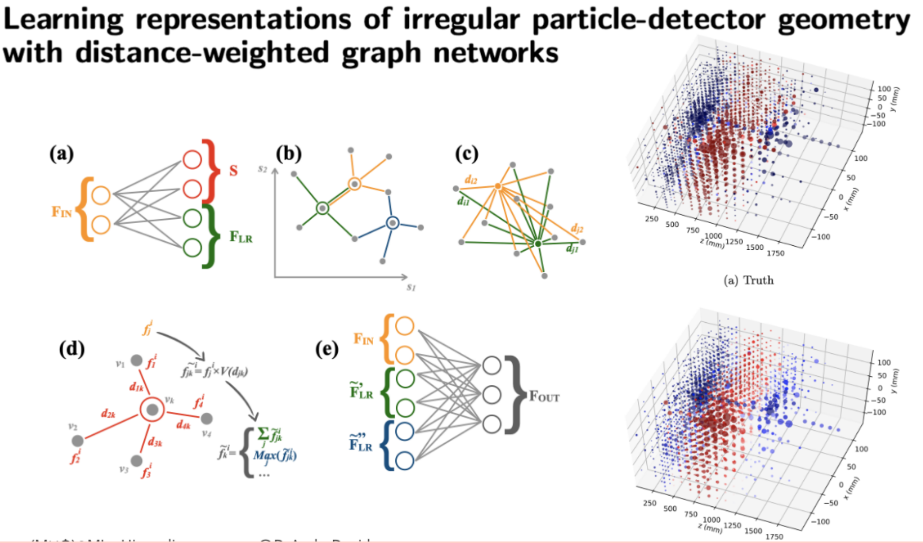


Low density
acc. x eff. ~ 97%

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

High-granularity calorimeter

- 600m^2 of sensors, 50 layers: 6 million cells with ∼3mm spatial resolution
  - Some square cells, some exagonal cells
  - Non-projective geometry



Learning representations of irregular particle-detector geometry with distance-weighted graph networks

Image from a talk by André David and the HGCAL team

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

Backup

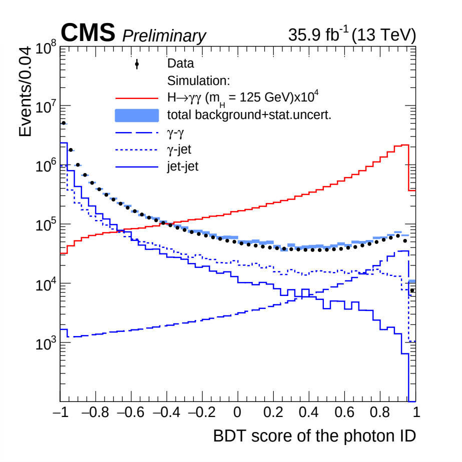
Object ID

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

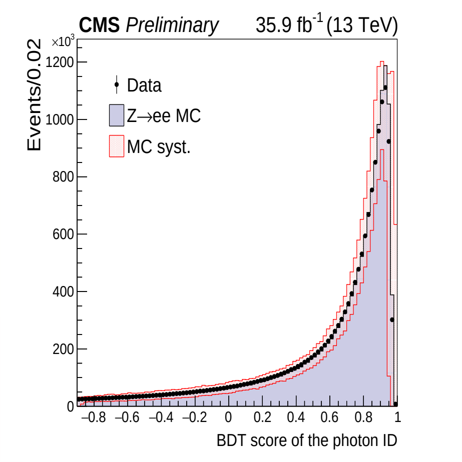
UCLouvain
Institut de recherche
en mathématique et physique

- 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



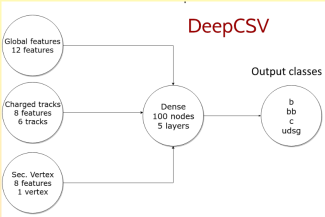
Validation in $Z \to ee$ events



Plots from CMS-PAS-HIG-16-040

Object ID enters the era of mathematical representations — 1

- Identification of jets from $b$quarks (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



DeepCSV

Global features
12 features

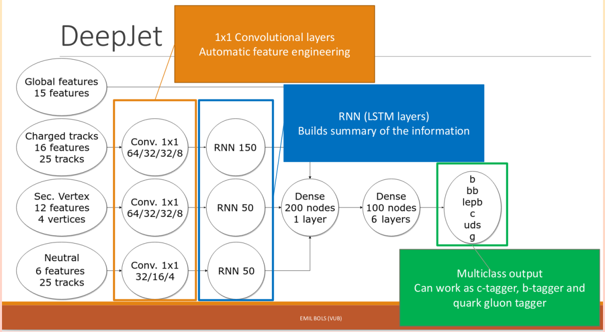
Charged tracks
8 features
6 tracks

Sec. Vertex
8 features
1 vertex

Dense
100 nodes
5 layers

Output classes

b
bb
c
udsg



DeepJet

1x1 Convolutional layers
Automatic feature engineering

Global features
15 features

Charged tracks
16 features
25 tracks

Sec. Vertex
12 features
4 vertices

Neutral
6 features
25 tracks

Conv. 1x1
64/32/32/8

Conv. 1x1
64/32/32/8

Conv. 1x1
32/16/4

RNN 150

RNN 50

RNN 50

RNN (LSTM layers)
Builds summary of the information

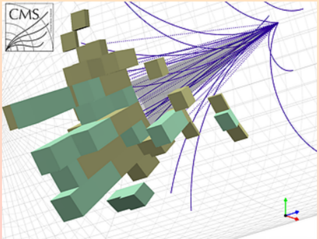
Dense
200 nodes
1 layer

Dense
100 nodes
6 layers

b
bb
lepb
c
uds
g

Multiclass output
Can work as c-tagger, b-tagger and quark gluon tagger

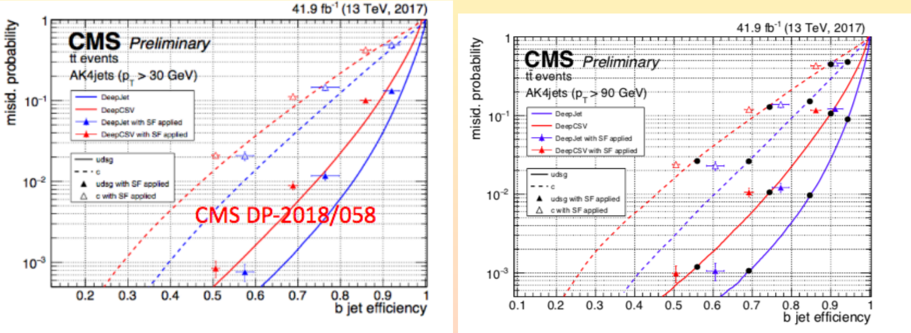
EMIL BOLS (VUB)



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

Object ID enters the era of mathematical representations — 2

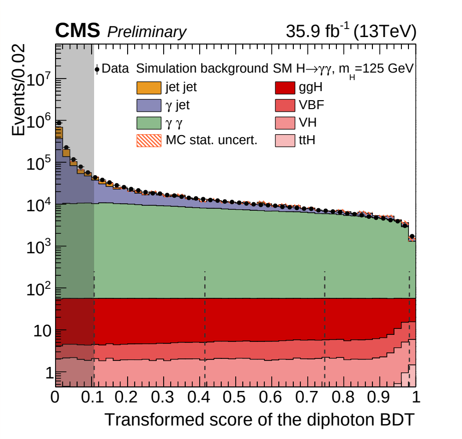
- Clear gain even with respect to using a generic DNN (DeepCSV)

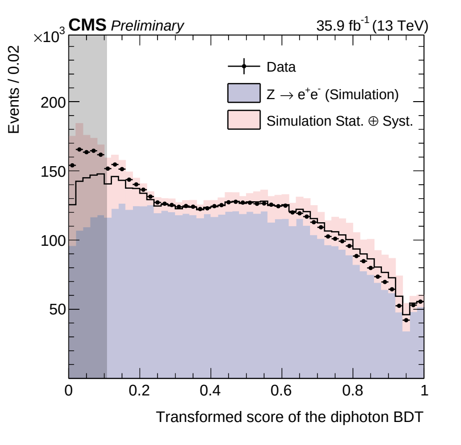


CMS DeepJet

Combining MVA ID for object identification

- 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 $Z \rightarrow ee$ events where electrons are reconstructed as photons

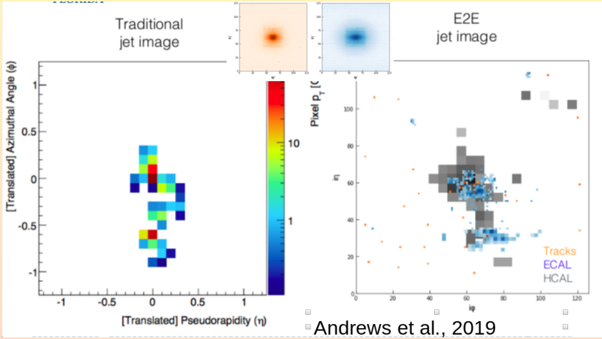




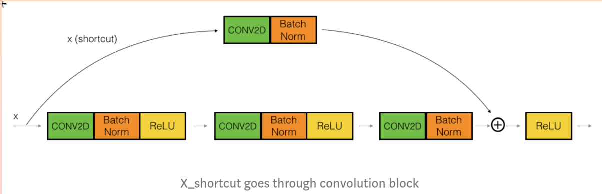
Plots from CMS-PAS-HIG-16-040

End-to-end reconstruction of jets

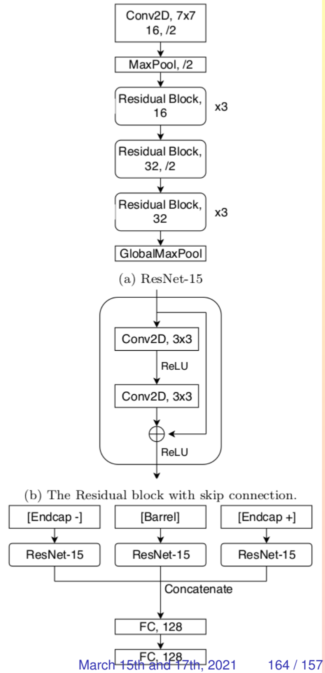
- Project detector layers in a single map
- Treat as an image: Res(idual)Net(works)
- Role of tracks in the reconstruction by the network is the same as we expect from the physics we know



Andrews et al., 2019



X_shortcut goes through convolution block



(a) ResNet-15

(b) The Residual block with skip connection.

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

Signal extraction

Separate signal from background using selection cuts

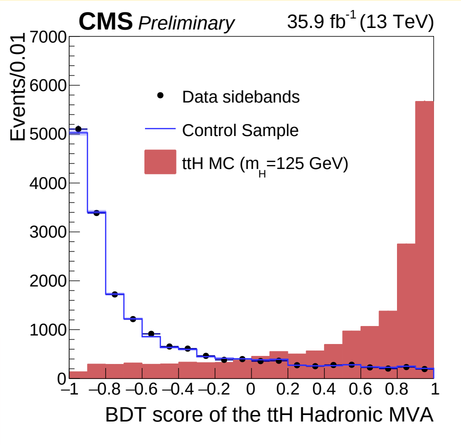
- High fraction of correct events in $t\bar{t}H$ categories by removing events from the dataset
- Delicate: removing events based on MVA output introduces tricky dependency on simulation
  - Dangerous, e.g. prevents from using unfolding results in comparisons with non-SM processes
- In both channels, 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}, $p_{T}^{leadjet}$, lead and sublead btag scores)

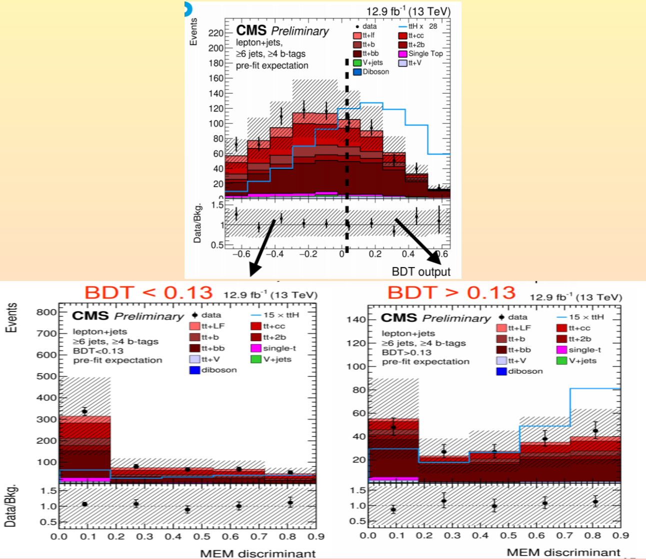


Evt Cat.	SM 125 GeV Higgs boson expected signal										
	Total	ggH	VBF	ttH	bbH	tHq	tHW	WH lep	ZH lep	WH had	ZH had
ttH Had.	5.85	10.99 %	0.70 %	77.54 %	2.02 %	4.13 %	2.02 %	0.09 %	0.05 %	0.63 %	1.82 %
ttH Lep.	3.81	1.90 %	0.05 %	87.48 %	0.08 %	4.73 %	3.04 %	1.53 %	1.15 %	0.02 %	0.02 %

Plots from CMS-PAS-HIG-16-040

Separate signal from background using all events

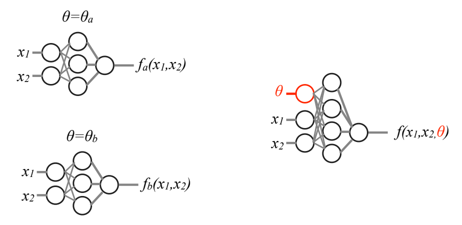
- Increase sensitivity by keeping the full MVA score distribution, possibly separating it into regions
  - Different fraction of signal/background
  - Constrain normalization or uncertainties in background-dominated regions

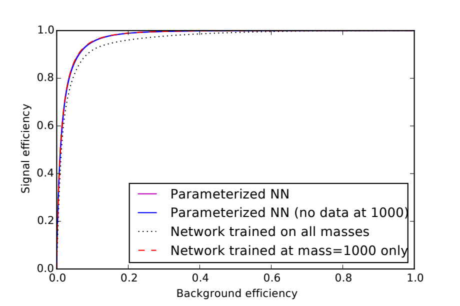


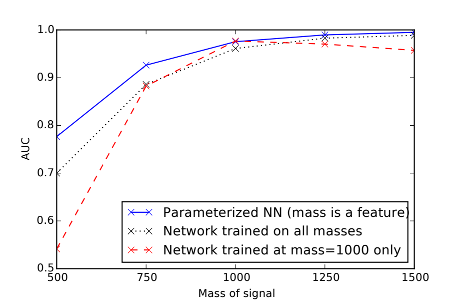
From $t\bar{t}H$ (bb): CMS-PAS-HIG-16-004

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 ☺



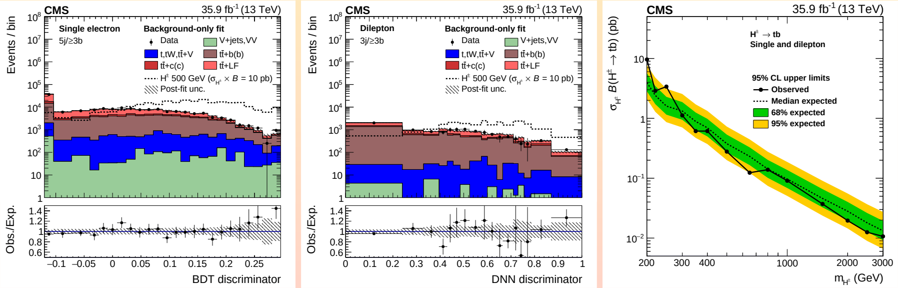




From Baldi *et al.* arXiv:1601.07913

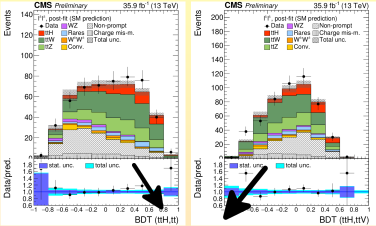
Different techniques are “better” for different situations

- 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 ($\mathrm{t\bar{t}}$ vs $\mathrm{H}^{\pm}$)



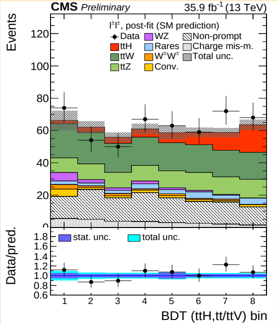
Reduce complexity: how many BDTs do you have?

BDT classifier output (2LSS)



- $t\bar{t}H$ 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

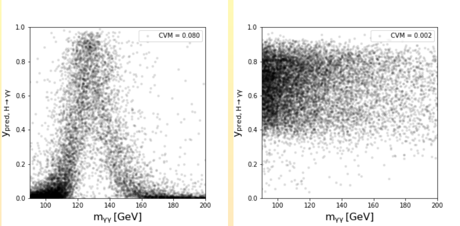
Final 1D discriminator (2LSS)

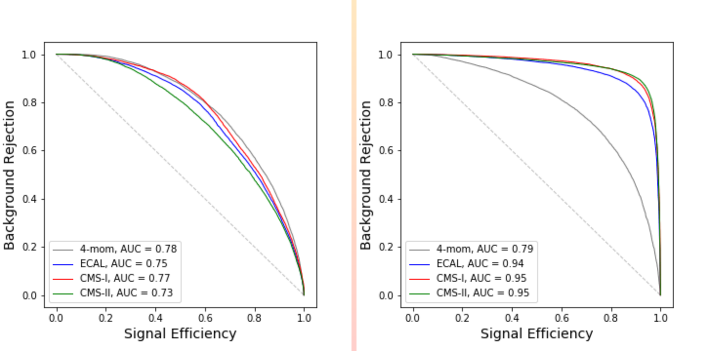


CMS-PAS-HIG-17-004, part of CMS-HIG-17-018: evidence for ttH production in multilepton final states

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 ($\mathrm{H} \to \gamma\gamma$, $\gamma\gamma$, γ+jet)
- Statistically-limited technique





From arXiv:1807.11916

What if you don’t know your signal?

Gaussian processes)

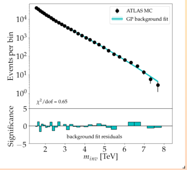
- Multivariate gaussian associated to a set of random variables ($N_{dim} = N_{random\ variables}$)
  - *Kernel* as a similarity measure between bin centers (counts) and a *averaging function*

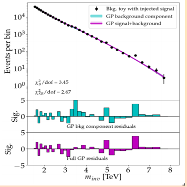
$$\mu(x) = 0\,, \tag{9}$$

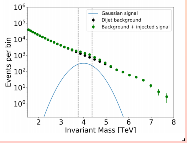
$$\Sigma_B\left(x, x'\right) = A\ \exp\left(\frac{d-\left(x+x'\right)}{2a}\right)\sqrt{\frac{2l(x)l\left(x'\right)}{l(x)^2+l\left(x'\right)^2}}\exp\left(\frac{-\left(x-x'\right)^2}{l(x)^2+l\left(x'\right)^2}\right)\,, \tag{10}$$

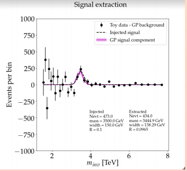
$$\Sigma_S(x, x') = C\ \exp\left(-\frac{1}{2}\left(x-x'\right)^2/k^2\right)\exp\left(-\frac{1}{2}\left((x-m)^2+\left(x'-m\right)^2\right)/t^2\right)\,, \tag{11}$$

-
  - Signal is not parameterized
  - Hyperparameters fixed by the B-only fit
- S: residual of B-subtraction





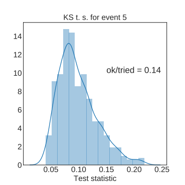


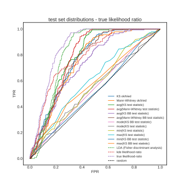


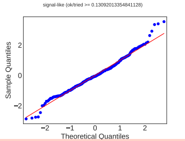
AMVA4NewPhysics deliverable 2.5 public report

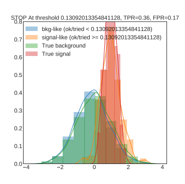
Inverse Bagging

- 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









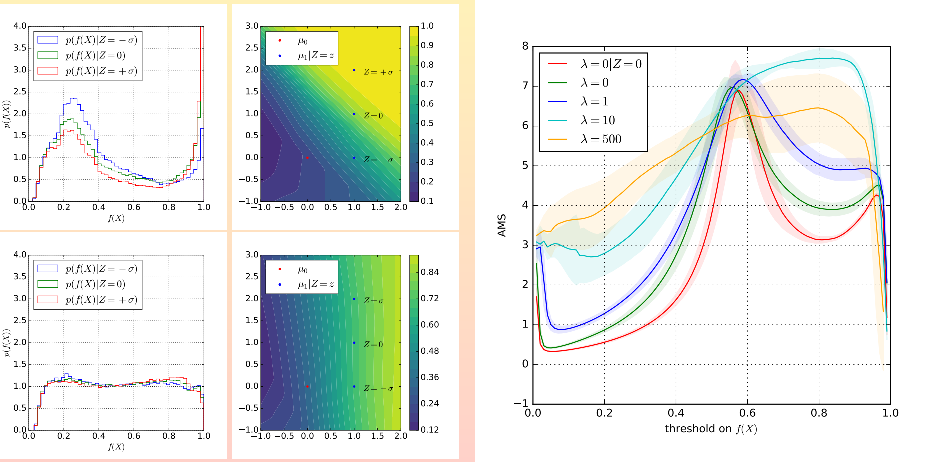
Vischia-Dorigo arXiv:1611.08256, doi:10.1051/epjconf/201713711009, and P. Vischia's talk at EMS2019

What about the uncertainties?

Can we reduce the impact of uncertainties on our results?

- 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

  $E_\lambda(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) - \lambda \mathcal{L}_r(\theta_f, \theta_r)$



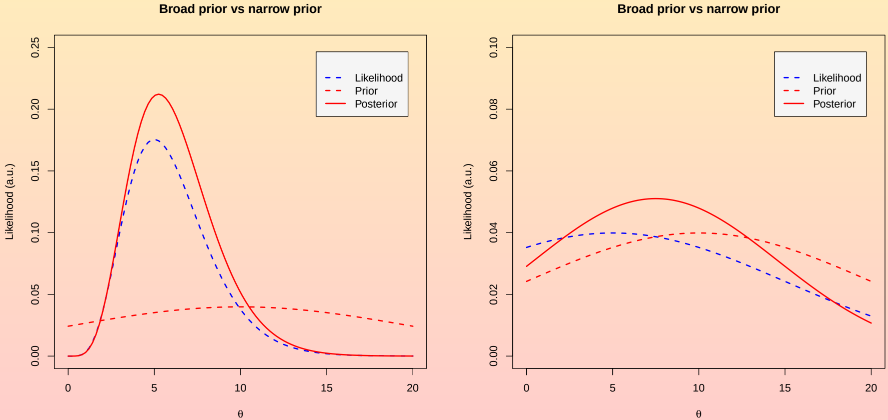
From Louppe-Kagan-Cranmer, arXiv:1611.01046

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

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

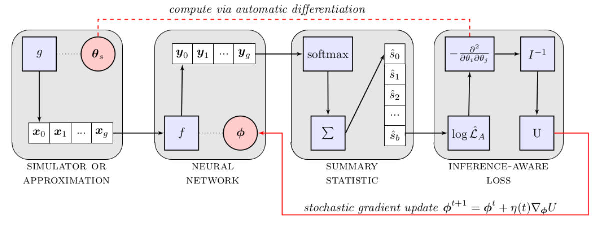
- The likelihood function contains all the information that you can extract from data on the parameter θ
- A narrow likelihood function carries more information than a broader one

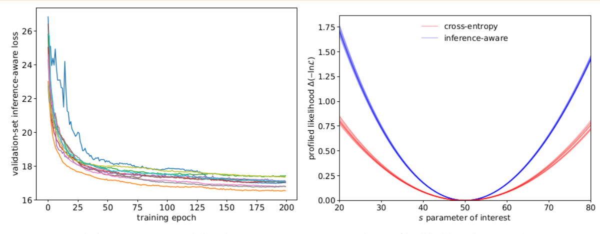


From Vischia, book in preparation

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


compute via automatic differentiation
g
θ_s
x_0 x_1 ... x_g
SIMULATOR OR APPROXIMATION
y_0 y_1 ... y_g
f
ϕ
NEURAL NETWORK
softmax
Σ
$\hat{s}_0$ $\hat{s}_1$ $\hat{s}_2$ ... $\hat{s}_b$
SUMMARY STATISTIC
$-\frac{\partial^2}{\partial\theta_i\partial\theta_j}$
I^{-1}
$\log \hat{\mathcal{L}}_A$
U
INFERENCE-AWARE LOSS
stochastic gradient update $\phi^{t+1} = \phi^t + \eta(t)\nabla_\phi U$



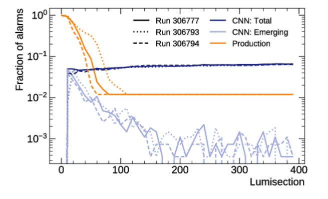
(a) inference-aware training loss (b) profile-likelihood comparison

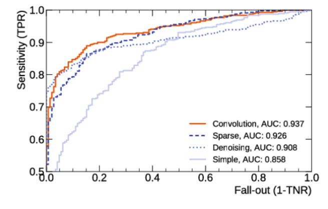
From De Castro-Dorigo, arXiv:1806.04743, and AMVA4NewPhysics deliverable 1.4 public report

Which data should we take?

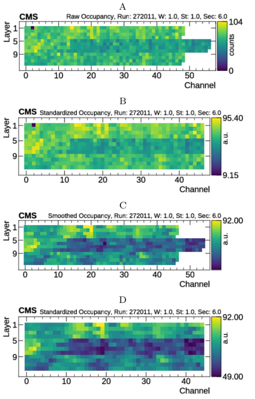
What if we don’t know which data to take?

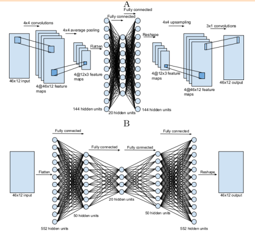
- Represent data as geographically-organized images
  - Local focus: detector layers treated independently
  - Regional focus: detector layers treated independently but simultaneously (spot problems between layers)
- Autoencoders (noise detection, dimensionality reduction)
  - Encode the inputs to the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





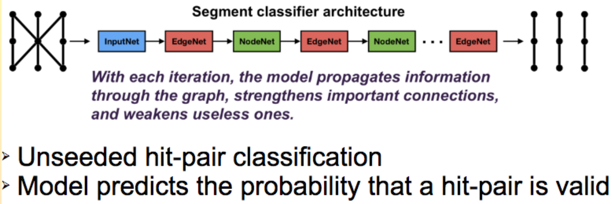
From arXiv:1808.00911





Tracking

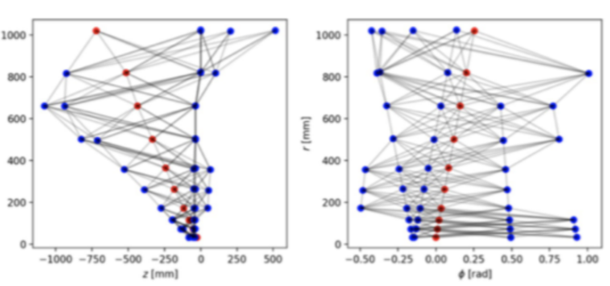
- Graph networks to literally connet the dots

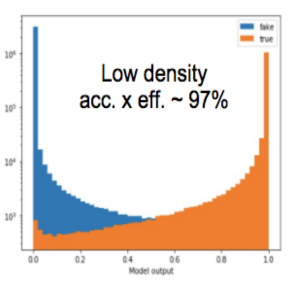


Segment classifier architecture

With each iteration, the model propagates information through the graph, strengthens important connections, and weakens useless ones.

- Unseeded hit-pair classification
- Model predicts the probability that a hit-pair is valid



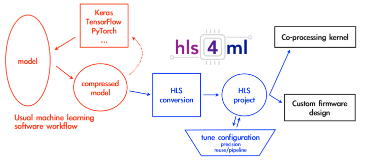


Low density
acc. x eff. ~ 97%

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),



Keras
TensorFlow
PyTorch
...

model

compressed model

Usual machine learning software workflow

hls 4 ml

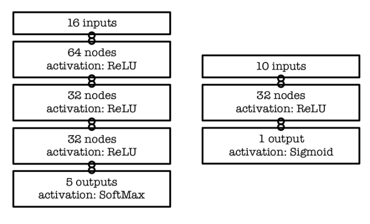
HLS conversion

HLS project

tune configuration
precision
reuse/pipeline

Co-processing kernel

Custom firmware design



16 inputs

64 nodes
activation: ReLU

32 nodes
activation: ReLU

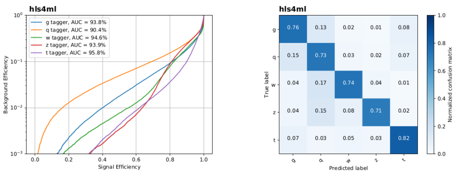
32 nodes
activation: ReLU

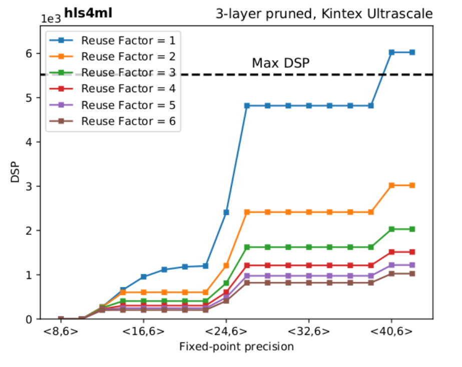
5 outputs
activation: SoftMax

10 inputs

32 nodes
activation: ReLU

1 output
activation: Sigmoid





From arXiv:1804.06913