Statistics

or “How to find answers to your questions”

Pietro Vischia[1]

[1] CP3 — IRMP, Université catholique de Louvain

CP3—IRMP, Intensive Course on Statistics for HEP, 21/01–18/02 2022

REMEMBER TO START RECORDING

Machine Learning

Lesson 5

Summary

- Schedule: five days of lectures (Every Friday for the next five weeks)
  - 2h morning lecture, virtual coffee break midway (09:30–11:45)
  - 2h (probably less) afternoon exercise session, virtual coffee break midway (13:30–15:45)
- 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
- 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!

- This course provides 3 credits for the UCLouvain doctoral school (CDD Sciences)
  - If you need it recognized by another doctoral school, you have to ask to your school
  - Besides the certificate, I am available at supplying additional information (e.g. detailed schedule) or activity (exam?)
- People connecting online: certificates will be provided by checking connection logs
  - The only way I have to check if you connected to most lectures is to check the Zoom logs
  - Make sure you connect with a recognizable email address (or let me know which unrecognizable address belongs to you)
- This course contributes to the activities of the Excellence of Science (EOS) Be.h network, https://be-h.be/

UCLouvain
Institut de recherche
en mathématique et physique

- I will pop up every now and then some questions
- I will open a link, and you'll be able to answer by going to www.menti.com and inserting a code
- Totally anonymous (no access even for me to any ID information, not even the country): don't be afraid to give a wrong answer!
  - The purpose is making you think, not having 100% correct answers!
- First question of the day is purely a logistics matter
  Question time: ROOT
  - The direct links are accessible to me only: you'll see in your screens the code in a second :)
- The slides of each lecture will be available one minute after the end of the lecture
  - To encourage you to really try answering without looking at the answers

UCLouvain
Institut de recherche
en mathématique et physique

- **Lesson 1 - Fundaments**
  - Bayesian and frequentist probability, theory of measure, correlation and causality, distributions
- **Lesson 2 - Point and Interval estimation**
  - Maximum likelihood methods, confidence intervals, most probable values, credible intervals
- **Lesson 3 - Advanced interval estimation, test of hypotheses**
  - Interval estimation near the physical boundary of a parameter
  - Frequentist and Bayesian tests, CLs, significance, look-elsewhere effect, reproducibility crysis
- **Lesson 4 - Commonly-used methods in particle physics**
  - Unfolding, ABCD, ABC, MCMC, estimating efficiencies
- **Lesson 5 - Machine Learning**
  - Overview and mathematical foundations, generalities most used algorithms, automatic Differentiation and Deep Learning

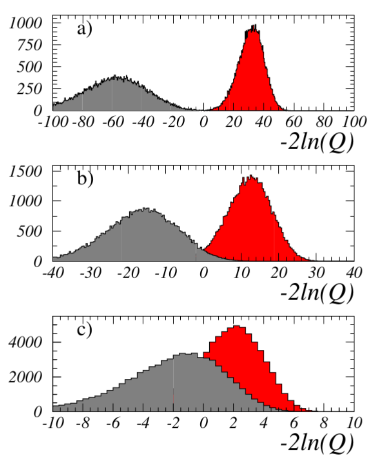
- Identify observables, and a suitable test statistic Q
- Define rules for exclusion/discovery, i.e. ranges of values of Q leading to various conclusions
  - Specify the significance of the statement, in form of confidence level (CL)
- Confidence limit: value of a parameter (mass, xsec) excluded at a 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

- Example: 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
- 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
  - $P(n \leq n_{obs}|n_b \leq n_{obs}, s + b) = \ldots = \frac{P(n \leq n_{obs}|s+b)}{P(n \leq n_{obs}|b)}$

Get your confidence levels right

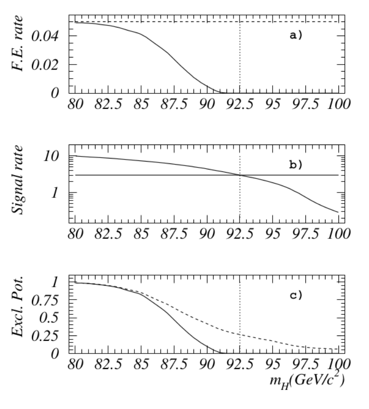
- Find a monotonic Q for increasing signal-like experiments (e.g. likelihood ratio)
- $CL_{s+b} = P_{s+b}(Q \leq Q_{obs})$
  - Small values imply poor compatibility with $S+B$ hypothesis, favouring B-only
- $CL_b = P_b(Q \leq Q_{obs})$
  - Large (close to 1) values imply poor compatibility with B-only,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

UCLouvain
Institut de recherche
en mathématique et physique

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



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

Plot from Read, CERN-open-2000-205

From a scan of CL_s to a limit on a cross section

- Scan the CLs test statistic as a function of the POI (typically the cross section modifier $\mu = \sigma_{obs}/\sigma pred$)
- Find its intersection with the desired confidence level
- (eventually) convert the limit on μ back to a cross section

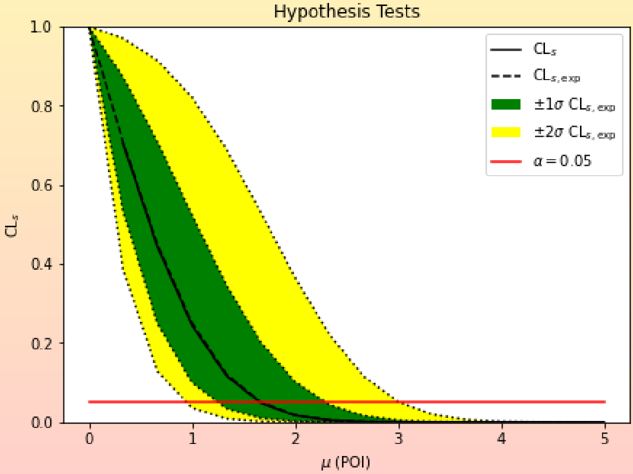
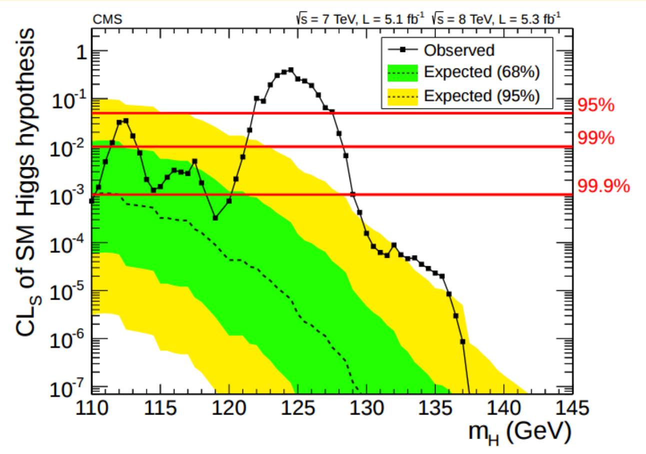


Image from the afternoon exercise on CL_s

From a limit on the cross section to hypothesis testing

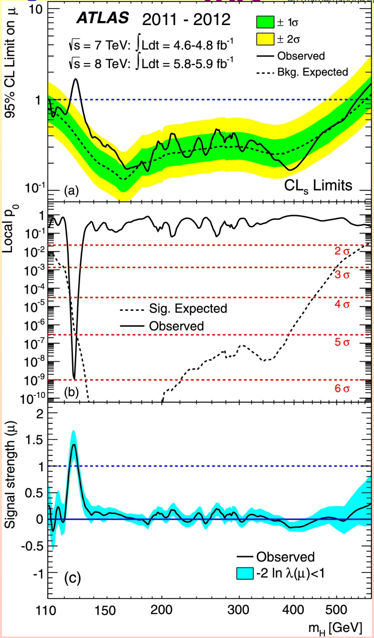
- Apply the CL_s method to each Higgs mass hypothesis
- Show the CL_s test statistic for each value of the fixed hypothesis
- Green/yellow bands indicate the $\pm1\sigma$ and $\pm2\sigma$ intervals for the expected values under B-only hypothesis
  - Obtained by taking the quantiles of the B-only hypothesis



Plot from CMS Higgs discovery paper doi:10.1016/j.physletb.2012.08.021

From a limit on the cross section to hypothesis testing

- CLs limit on μ as a function of mass hypothesis
- p-value of excess
- Fitted signal strength peaks at excess



Plot from ATLAS Higgs discovery paper doi:10.1016/j.physletb.2012.08.020

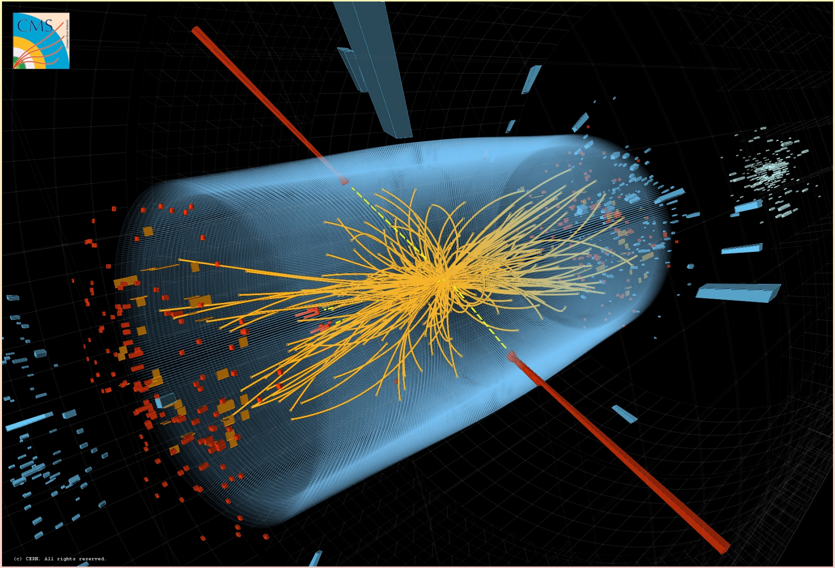
Machine learning

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

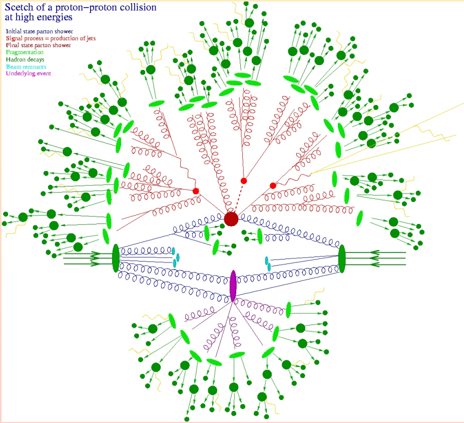


We must efficiently sample from our models

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

- 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



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

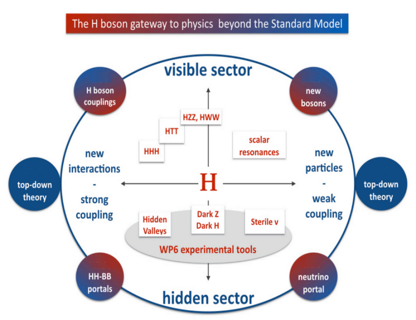
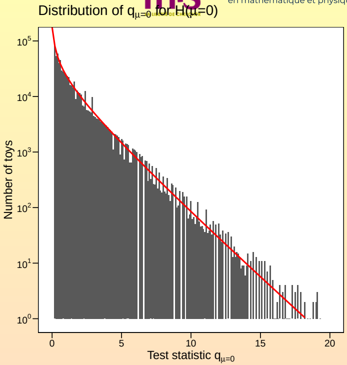
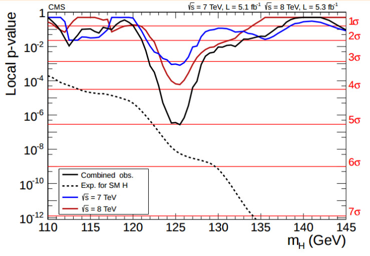


Image by the EOS be.h network

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**



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

- I was told *“this is a black box, we cannot trust it for physics”*
  Comment by one of the researchers assisting to my final summer-student internship seminar
- There was still some diffidence towards machine learning algorithms

Trainee: Pietro Vischia **Year:** 2006

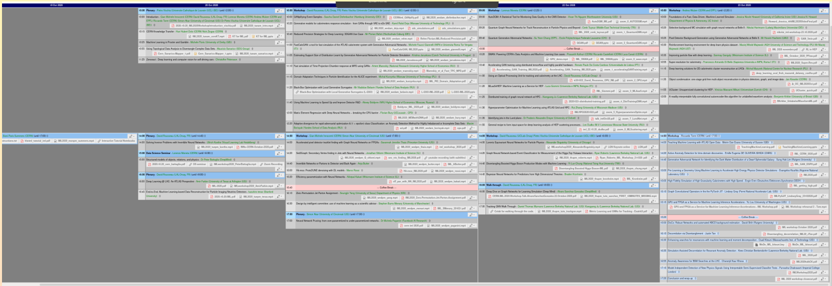
Mentor: Stephan Lammel

A) What did you learn while at Fermilab?

In my period as a Trainee I worked on b-jet energy corrections, thus deepening my knowledge on b-physics and data analysis methods.
In particular, I gained knowledge about neural networks and their possible use in physics.
I indeed used a neural network in the attempt to improve the b-jet energy resolution: this type of improvement is important for obtaining a better measurement of top quark mass and for having more chances of eventually discovering the Higgs boson at CDF.
The work started with gaining familiarity with basics concepts about neural networks, in particular multilayer perceptrons. This was accomplished by studying the software documentation and the work of some physicists who already used neural networks for data analysis. A preliminary work has been done by using some material sent by Brandon Parks (University of Ohio).
I then developed a network and applied it to our Z->b+anti-b and QCD b+anti-b signal. I obtained a substantial improvement of the resolution on b-jet energy by obtaining a scale factor which modifies the measured energy of the quark.

...and today

- I co-organized the CERN IML workshop (October 19th–23rd, 2020)
  - 951 registered participants
  - 71 contributions



1. ML for data reduction : Application of Machine Learning to data reduction, reconstruction, building/tagging of intermediate object
2. ML for analysis : Application of Machine Learning to analysis, event classification and fundamental parameters inference
3. ML for simulation and surrogate model : Application of Machine Learning to simulation or other cases where it is deemed to replace an existing complex model
4. Fast ML : Application of Machine Learning to DAQ/Trigger/Real Time Analysis
5. ML algorithms : Machine Learning development across applications
6. ML infrastructure : Hardware and software for Machine Learning
7. ML training, courses and tutorials
8. ML open datasets and challenges
9. ML for astroparticle
10. ML for experimental particle physics
11. ML for phenomenology and theory
12. ML for particle accelerators
13. Other

- 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 $\sqrt{d}$, $\mathcal{X} = \mathbb{R}^d$, $d \sim \mathcal{O}(10^6)$

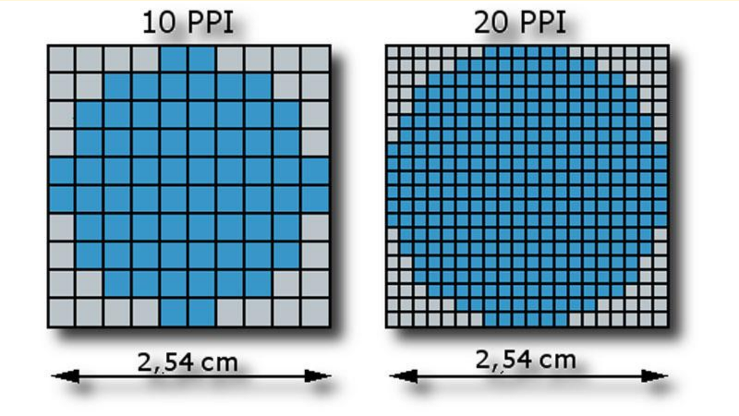


Figure frm scientiamobile.com

UCLouvain
Institut de recherche
en mathématique et physique

- 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

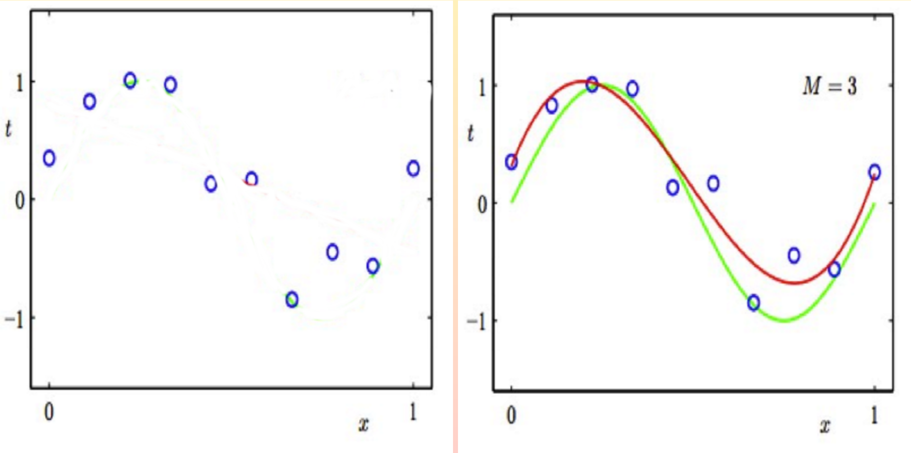
- $f^*: \mathcal{X} \rightarrow \mathbb{R}$, unknown target function
  - In case of multidimensional output to a vector of dimension k, $f^*: \mathcal{X} \rightarrow \mathbb{R}^k$
  - Some loose assumptions (e.g. square-integrable with respect to the ν measure, i.e. finite moments, bounded...)

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

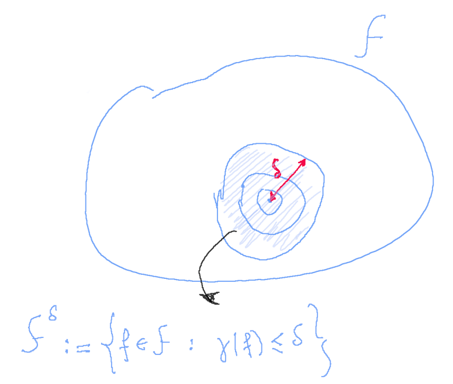
- Goal: predict f^* from a finite i.i.d. sample of points sampled from ν
- Sample: $\left\{x_i, f^*(x_i)\right\}_{i=1,\ldots,n}, x_i \sim \nu$
  - For each of the points x_i, we know the value of the unknown function (our true labels)
  - We want to *interpolate* for any arbitrary x inbetween the labelled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



- 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!

UCLouvain
Institut de recherche
en mathématique et physique

- **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)

UCLouvain
Institut de recherche
en mathématique et physique

- 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 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

UCLouvain
Institut de recherche
en mathématique et physique

- **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

UCLouvain
Institut de recherche
en mathématique et physique

- **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!

UCLouvain
Institut de recherche
en mathématique et physique

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

UCLouvain
Institut de recherche
en mathématique et physique

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

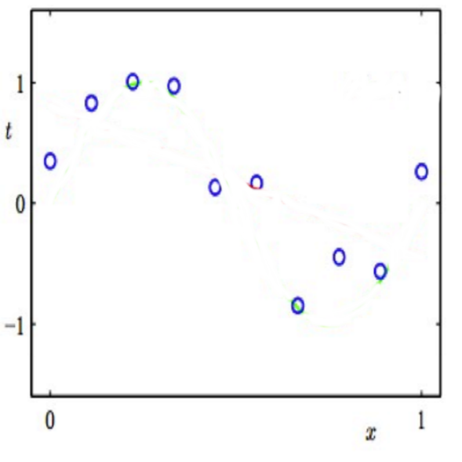
- f^* locally linear, i.e. f^* is Lipschitz
  - $|f^*(x) - f^*(y)| \leq \beta ||x - y||$
  - $Lip(f^*) = inf\left\{\beta; |f^*(x) - f^*(y)| \leq \beta ||x - y|| \text{ is true}\right\}$
  - Lip(f^*) is a measure of smoothness
- Space of functionals that are Lipschitz: $\mathcal{F} = \left\{f : \mathbb{R}^d \rightarrow \mathbb{R}; f \text{ is Lipschitz}\right\}$
- 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

fnrs LA LIBERTÉ DE CHERCHER
UCLouvain
Institut de recherche
en mathématique et physique

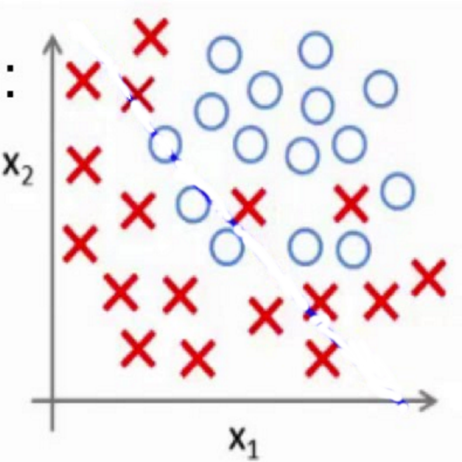
- $\forall \epsilon > 0$, find $f \in \mathcal{F}$ such that $\|f - f^{*}\| \leq \epsilon$ from n i.i.d. samples
  - n: sample complexity, “how many more samples to I need to make the error a given amount of times smaller”
- If f^{*} is Lipschitz, it can be demonstrated that $n \sim \epsilon^{-d}$
  - 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?

To describe the data points?
(regression)

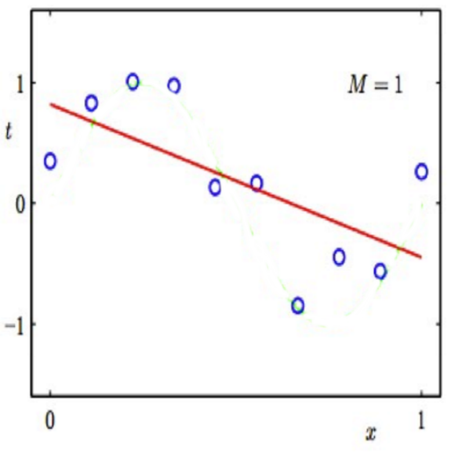


To separate into two classes?
(classification)

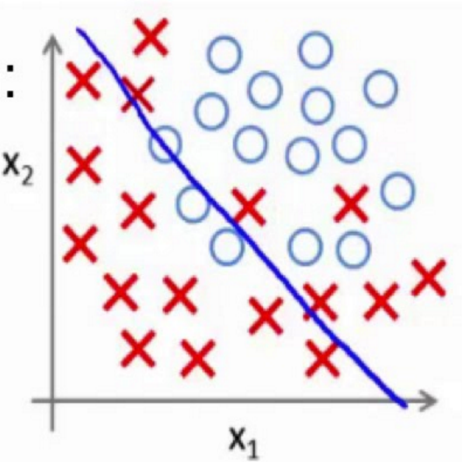


Images by Victor Lavrenko

To describe the data points?
(regression)

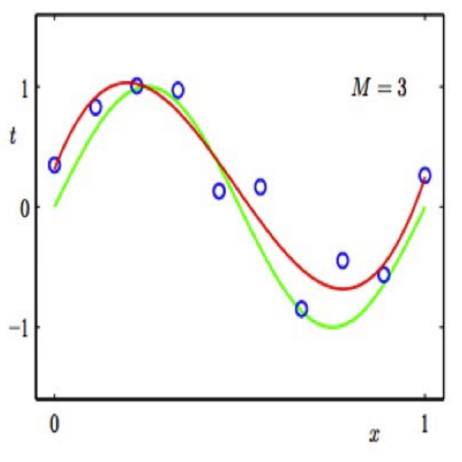


To separate into two classes?
(classification)

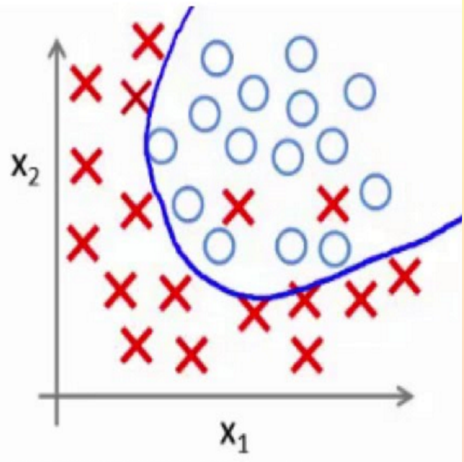


Images by Victor Lavrenko

To describe the data points?
(regression)

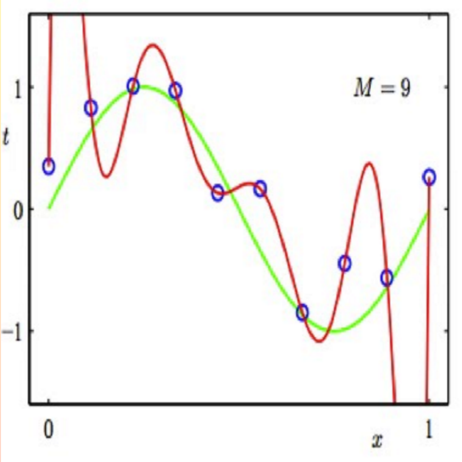


To separate into two classes?
(classification)

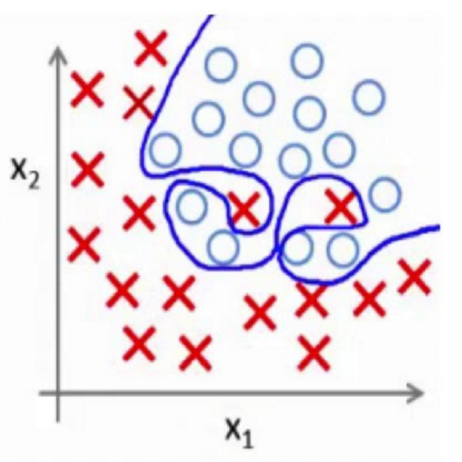


Images by Victor Lavrenko

To describe the data points?
(regression)

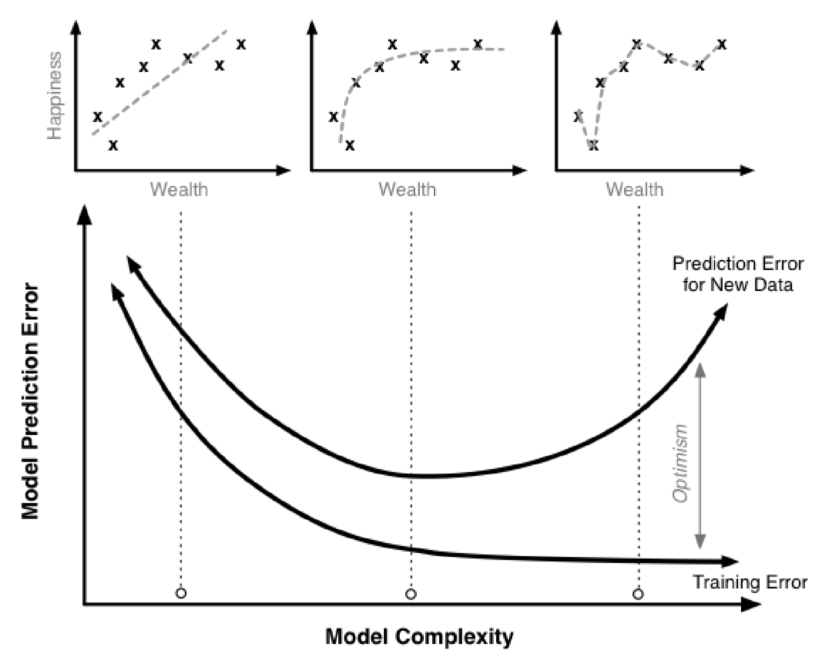


To separate into two classes?
(classification)



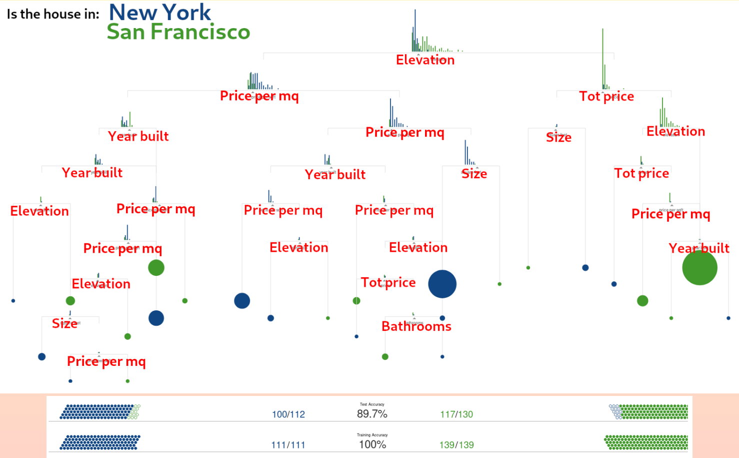
Images by Victor Lavrenko

Avoid overtraining



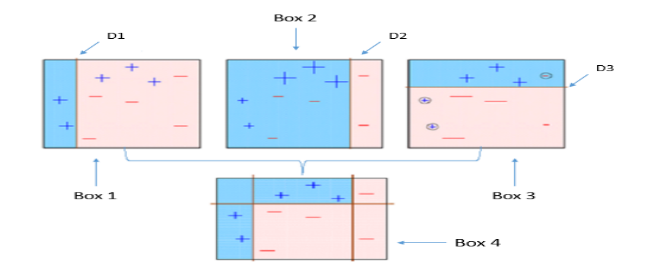
From can't remember where

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

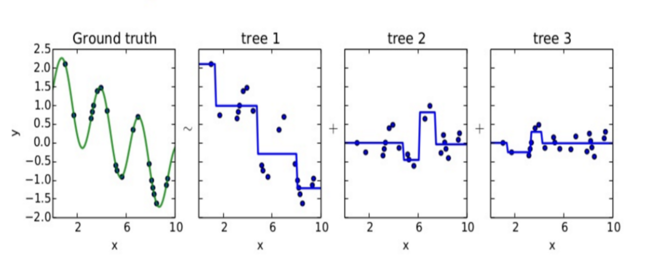


From http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

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

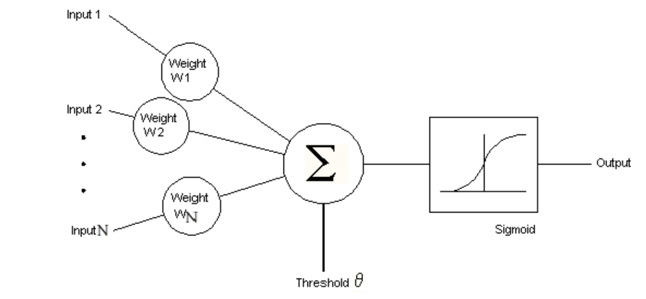


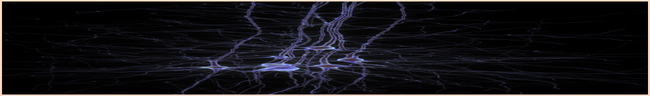
- GradientBoost: fit the new predictor to the residual errors of the previous one



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

UCLouvain
Institut de recherche
en mathématique et physique

- 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”

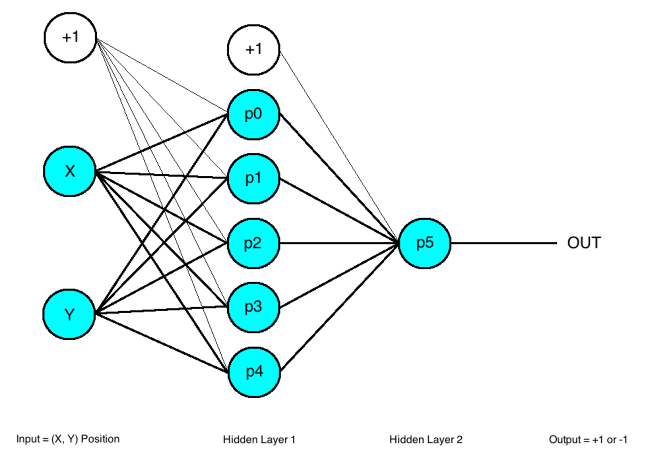


Image https://www.cs.utexas.edu/ teammco/misc/mlp/mlp.png

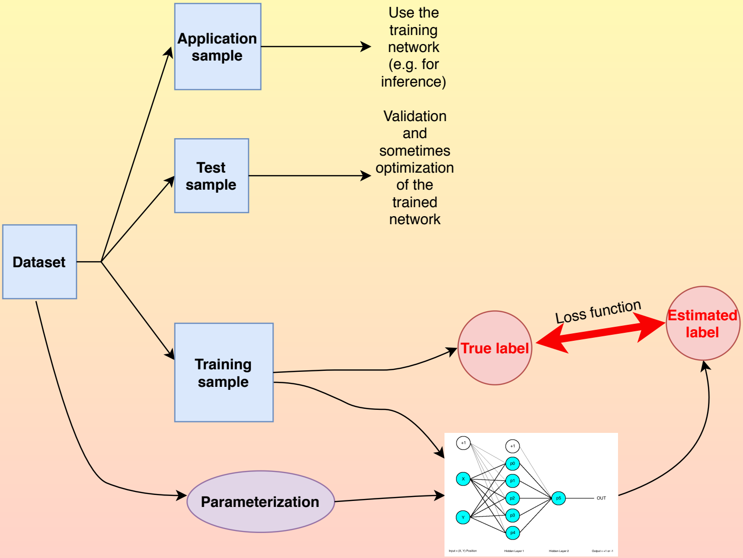
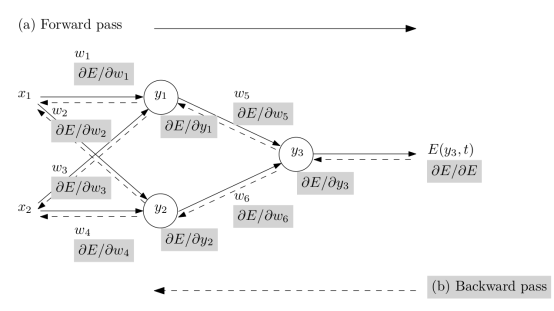
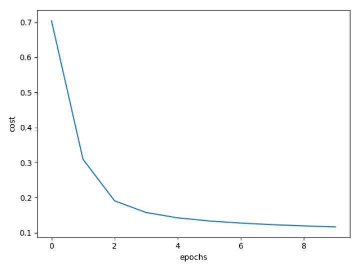


Image copyright Vischia, 2019

- 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





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

UCLouvain
Institut de recherche
en mathématique et physique

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

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)

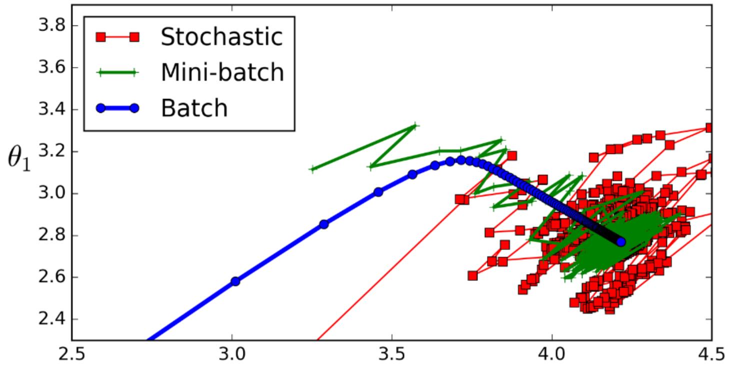
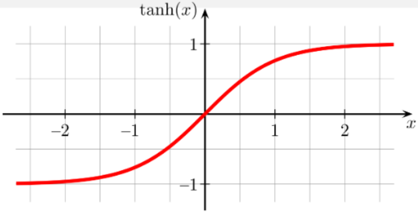
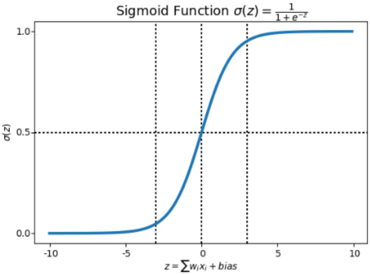


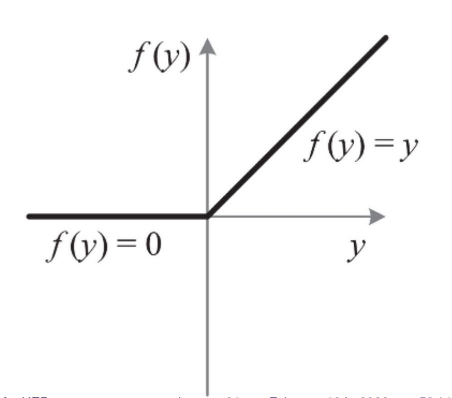
Image from a talk by W. Verbeke

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

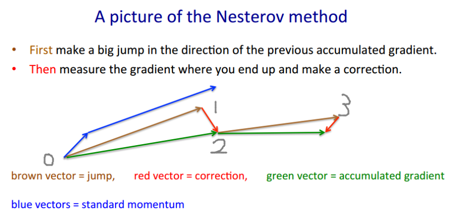






Improve algorithm to follow the gradient

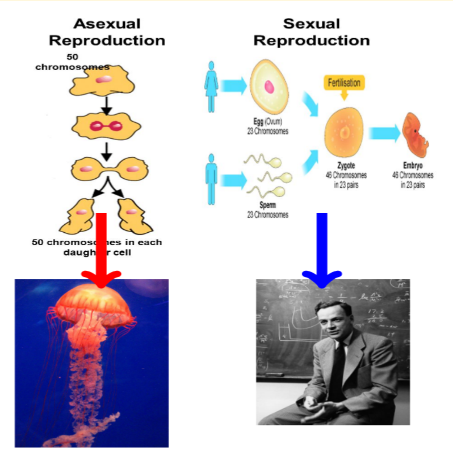
- Mostly nonconvex optimization: very complicated problem, convergence in general not guaranteed
- Nesterov momentum: big jumps followed by correction seem to help!
- Adaptive moments: gradient steps decrease when getting closer to the minimum (avoids overshooting)

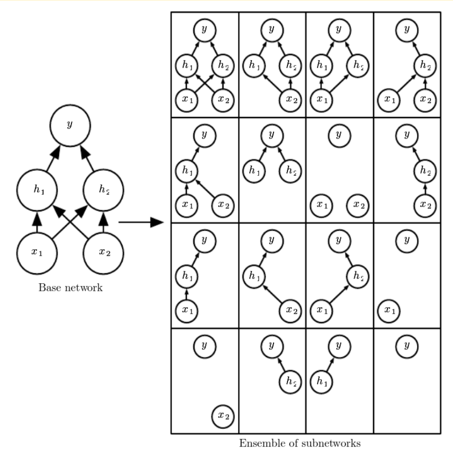


Nesterov Video

Diagram by Geoffrey Hinton, animation by Alec Radford

- 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





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

UCLouvain
Institut de recherche
en mathématique et physique

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

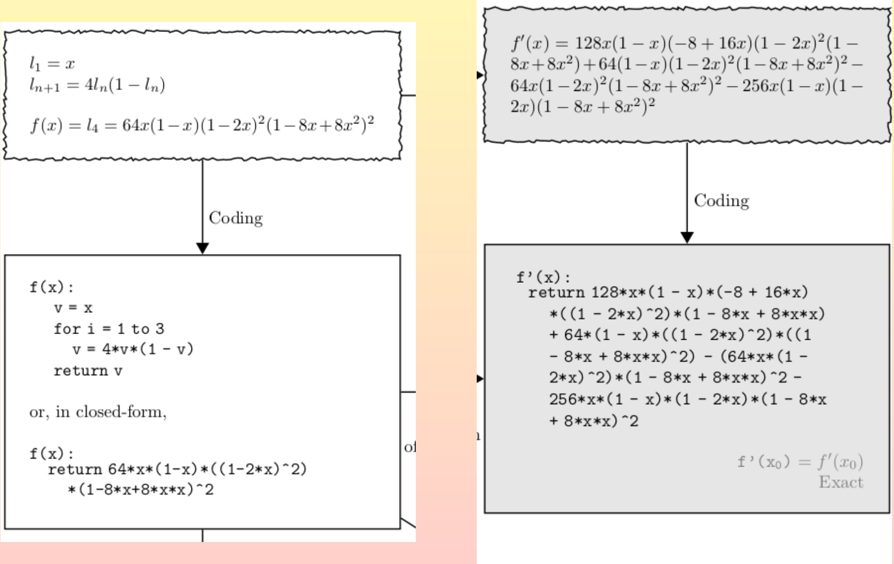
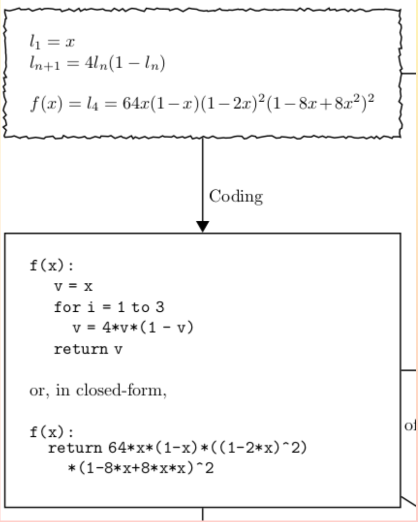


Image from Güneş Baydin et al, JMLR 18 (2018) 1–43

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)



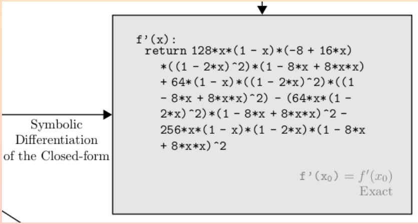
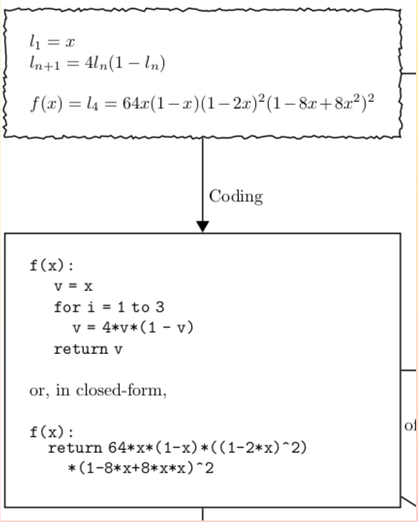


Image from Güneş Baydin et al, JMLR 18 (2018) 1–43

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



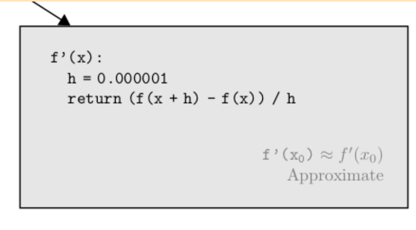
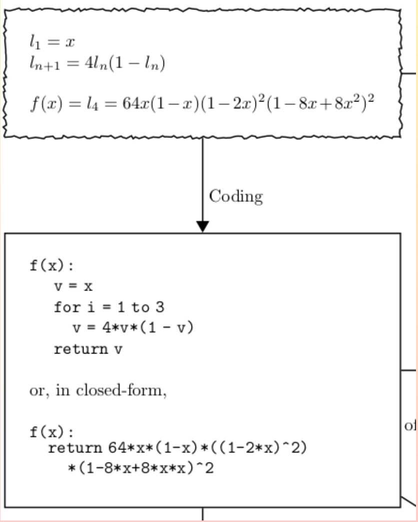


Image from Güneş Baydin et al, JMLR 18 (2018) 1–43

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



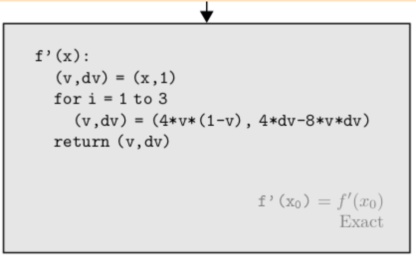


Image from Günes Baydin et al, JMLR 18 (2018) 1–43

The power of autodifferentiation

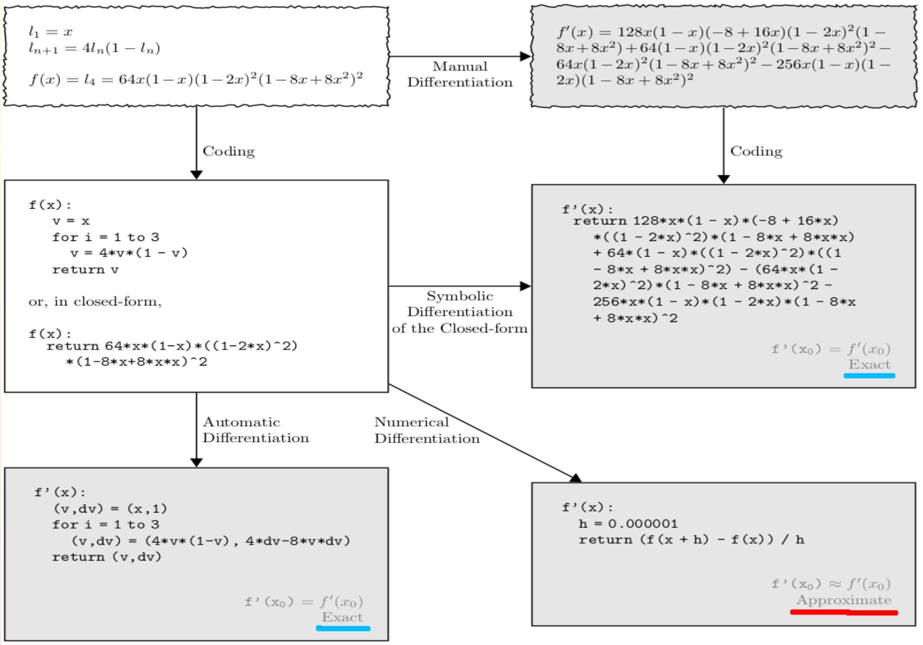


Image from Güneş Baydin et al, JMLR 18 (2018) 1–43

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

$\dot{v}_{-1}$	$= \dot{x}_1$	$= 1$
$\dot{v}_0$	$= \dot{x}_2$	$= 0$
$\dot{v}_1$	$= \dot{v}_{-1}/v_{-1}$	$= 1/2$
$\dot{v}_2$	$= \dot{v}_{-1} \times v_0 + \dot{v}_0 \times v_{-1}$	$= 1 \times 5 + 0 \times 2$
$\dot{v}_3$	$= \dot{v}_0 \times \cos v_0$	$= 0 \times \cos 5$
$\dot{v}_4$	$= \dot{v}_1 + \dot{v}_2$	$= 0.5 + 5$
$\dot{v}_5$	$= \dot{v}_4 - \dot{v}_3$	$= 5.5 - 0$
$\boldsymbol{\dot{y}}$	$\boldsymbol{= \dot{v}_5}$	**= 5.5**

Reverse mode

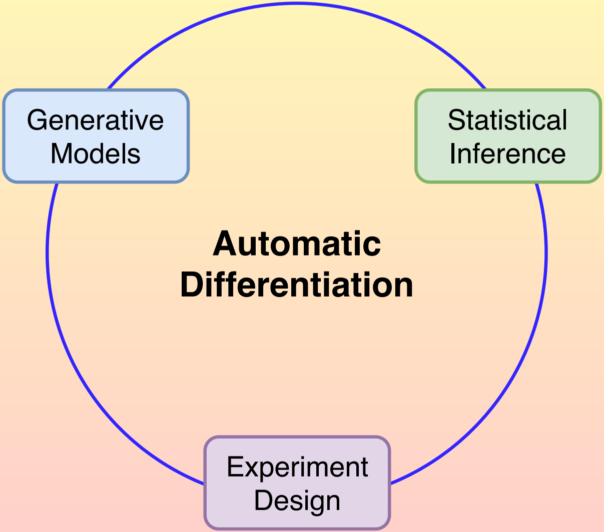
- Associate with each intermediate v_i an adjoint
  $\bar{v} = \frac{\partial y_j}{\partial v_i}$
- Run forwards and backwards as in backpropagation
- Single pass for $f : \mathbb{R}^n \to \mathbb{R}$ (functions with many inputs)
- Must store several values

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

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


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

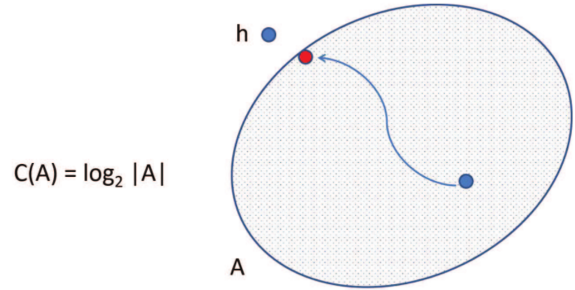


Figure 1. Learning framework where h is the function to be learnt and A is the available class of hypothesis or approximating functions. The cardinal capacity is the logarithm base two of the number, or volume, of the functions contained in A.

Plot from Baldi and Vershynin, arXiv:1901.00434

Separate prompt from fake leptons...

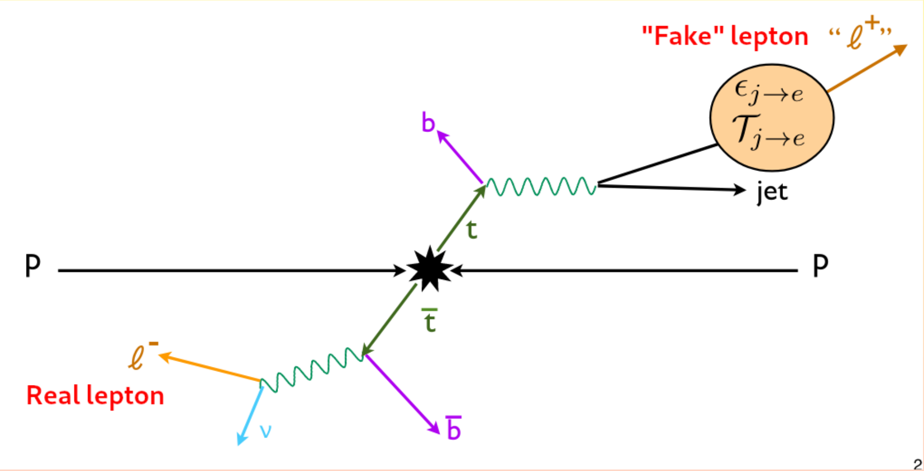
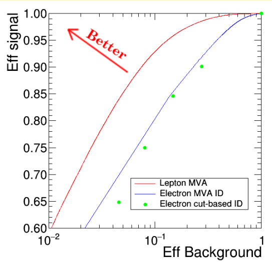
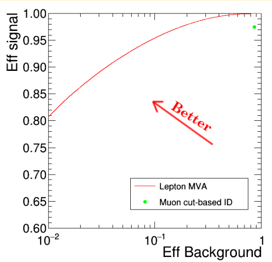


Image edited from David Curtin's talk at MC4BSM-2014

...is not very difficult

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





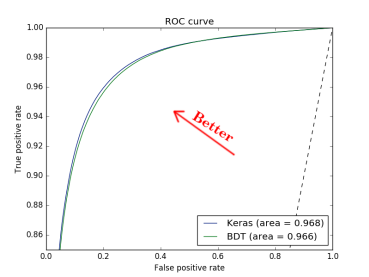
Plots by S.S. Cruz

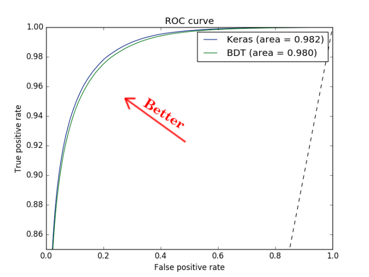
Corte	Background Fondo: *Fakes* *Valor*	Δ	$\frac{Fakes}{Fakes(t\bar{t}H)}$ (%)	Total *Valor*	Δ	Signal Señal ($t\bar{t}$H) *Valor*	Δ	Datos *Valor*	Δ	Predicción/Datos	$\frac{S}{\sqrt{F+(\Delta F)^2}}$
> 0.97	14	7	5	246	22	46	5	363	19	0.804	1.694
> 0.95	177	71	60	471	75	62	7	524	23	1.017	0.788
$t\bar{t}H$	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

UCLouvain
Institut de recherche
en mathématique et physique

Neural networks can approximate any continuous real-valued function

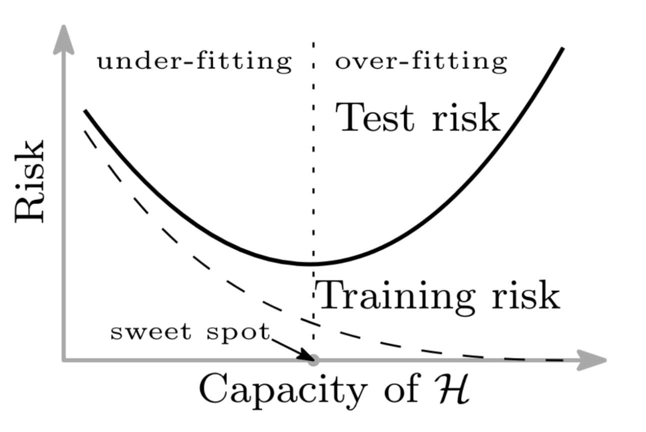
- A feed-forward network with sigmoid activation functions can approximate any continuos real-valued function. Cybenko, G. (1989)
- Any failure in mapping a function comes from inadequate choice of weights or insufficient number of neurons. Hornik et al (1989), Funahashi (1989)
- Derivatives can be approximated 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

Learning as a tradeoff...

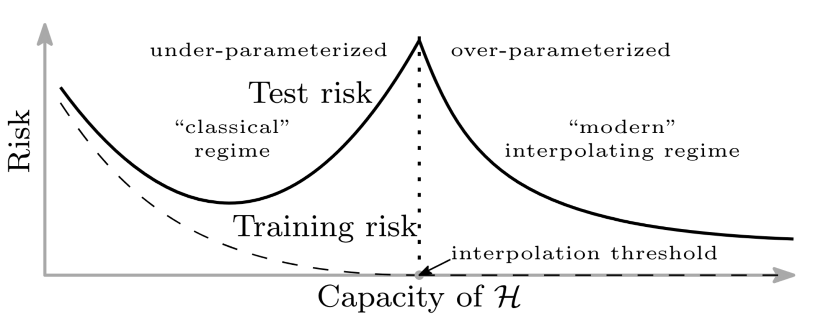
“a model with zero training erroris overfit to the training data and will typically generalize poorly” (Hastie, Tibshirani, Friedman)



From Belkin et al. arXiv:1812.11118

"the learned predictors achieve (near) perfect fits to the training data—i.e., interpolation. Although the learned predictors obtained at the interpolation threshold typically have high risk, we show that increasing the function class capacity beyond this point leads to decreasing risk, typically going below the risk achieved at the sweet spot in the 'classical' regime."

- The general idea is that increasing the class of allowed functions, it's more likely to find a smooth function with lower norm (complexity): Occam's razor



From Belkin et al. arXiv:1812.11118

The concept of loss function actually comes from statistics!

- Decision-theoretic approach (C.P. Robert, "The Bayesian Choice"): a statistical model involves three spaces
  - $\mathcal{X}$: observation space
  - Θ: parameter space
  - $\mathcal{D}$: decision (action) space
- **Statistical inference:** *"taking a decision $d \in \mathcal{D}$ related to the parameter $\theta \in \Theta$ based on the observation $x \in \mathcal{X}$, x and θ being related by the distribution $f(x|\theta)$"*
  - Typically, d consists in estimating a function $h(\theta)$ as accurately as possible
- Decision theory: the accuracy of each action can be quantified, leading to a reward r with utility function $U(r)$, typically assuming a rational decision-maker
- Utility function ultimately depends on θ and d, and where random factors are involved $U(\theta, d) = \mathbb{E}_{\theta,d}\left[U(r)\right]$
  - A measure of proximity between the proposed estimate d and the true value $h(\theta)$

Frequentist loss, Bayesian loss

- Loss function: $L(\theta, d) = -U(\theta, d)$
  - Represents intuitively the loss or error in which you incur when you make a bad decision (a bad estimation of the target function)
  - Lower bound at 0: avoids “infinite utility” paradoxes (St. Petersburg paradox, martingale-based stragegies)
  - Generally impossible to uniformly minimize in d the loss for θ unknown
- Frequentist loss (risk) is integrated on $\mathcal{X}$: $R(\theta, \delta) = \mathbb{E}_\theta \left[L(\theta, \delta(x)) \right]$
  - $\delta(\cdot)$ is an **estimator** of θ (e.g. MLE)
  - Compare estimators, find the best estimator based on long-run performance for all values of unknown θ
  - Issues: based on long run performance (not optimal for x_{obs}); repeatability of the experiment; no total ordering on the set of estimators
- Bayesian loss: is integrated on Θ: $\rho(\pi, d|x) = \mathbb{E}^\pi \left[L(\theta, d) | x \right]$
  - Posterior loss averages the error over the posterior distribution of θ conditional on x_{obs}
  - Can use the conditionality because x_{obs} is known!
  - Can also integrate the frequentist risk; integrated risk $r(\pi, \delta) = \mathbb{E}^\pi \left[R(\theta, \delta) \right]$ averaged over θ according to π (total ordering)

Losses in ANNs and Bayesian networks

- Standard ANN training essentially is a frequentist MLE
  - NN weights: true, unknown values
  - Data: random variable
- Bayesian networks treat weights ω as random (latent) variables, and condition on the observed data
  - Obtain $p(\omega|data)$ starting from prior belief $\pi(\omega)$ and likelihood $p(data|\omega)$
  - Predictions obtained as expectation values, $E_p[f] = \int f(\omega)p(\omega|data)d\omega$, averaging f weighting by the posterior
  - Marginalization leads to essentially learning the generative model (the pdfs), leading to interpretability

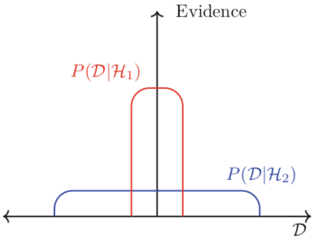
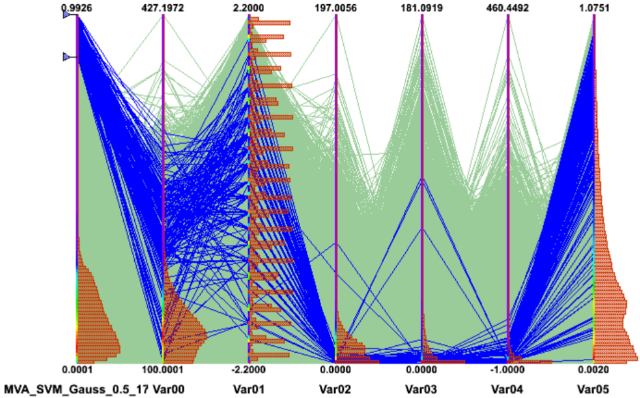


Fig. 3.4 Graphical illustration of how the evidence plays a role in investigating different model hypotheses. The simple model $\mathcal{H}_1$ is able to predict a small range of data with greater strength, while the more complex model $\mathcal{H}_2$ is able to represent a larger range of data, though with lower probability. Adapted from [45, 46]

Image from doi:10.1007/978-3-030-42553-1

- **Permutation importance:** the decrease in a model score when a single feature value is randomly shuffled (scikit-learn docs) (akin to impacts for profile likelihood fits)
- **Shapley values:** based on game theory (see other contribution)
- **Correlation-based:** e.g. parallel coordinates in TMVA: look where each variable is mapped to/correlated with

Journal of Physics: Conference Series **219** (2010) 032010 doi:10.1088/1742-6596/219/3/032010



UCLouvain
Institut de recherche en mathématique et physique

- **Bayesian Information Criterion:** $BIC = n_{free\ params} ln(n_{data}) - 2ln(\hat{L})$
  - Parameter θ predicted by two models M_0 and M_1: $P(\theta|\vec{x}, M) = \frac{P(\vec{x}|\theta,M)P(\theta|M)}{P(\vec{x}|M)}$
  - Apply Bayes theorem to Bayesian evidence (Model likelihood): $P(\vec{x}|M) = \int P(\vec{x}|\theta, M)P(\theta|M)d\theta$
  - Posterior odds: $\frac{P(M_0|\vec{x})}{P(M_1|\vec{x})} = \frac{P(\vec{x}|M_0)\pi(M_0)}{P(\vec{x}|M_1)\pi(M_1)}$
  - Can rewrite posteriors in terms of BIC, equivalent
- **Minimum Description Length (MDL):** Kolmogorov complexity (length of minimum program needed to describe the data)
  - $for\ i = 1\ to\ 2500; do\ \{print'0001'\}; halt$
  - $print'1010010101000101110010000100001011100111000010101001 01\ldots'; halt$
- **Structural risk minimization:** complexity as Vapnik-Chervonkensis class (largest number of shattered points)
  - Build a nested sequence of models with increasing VC complexity h
  - Write a probabilistic upper bound for the regression error: $err \leq f(h/N)$
  - Choose model with smallest value of the upper bound

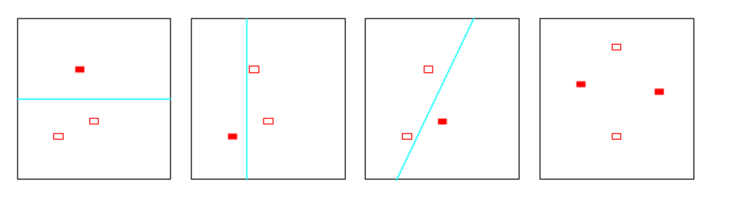
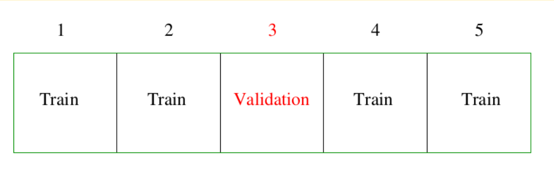


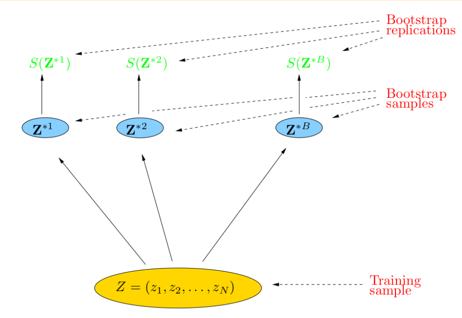
Image from Hastie, Tibshirani, Friedman

Model assessment by focussing on prediction error

- **Cross-validation:** useful when data are scarce
  - Split the data into K parts ("folds")
  - For the kth part, fit the model to the other K-1 folds, and calculate test error as error on predicting the kth part data
  - Do this for all k, then combine the K estimates of the prediction error
  - Choose K
    - K=N (leave-one-out), unbiased but high variance (training sets are basically the same)
    - Low K (5–10): Lower variance, but maybe bias (folds not representative of the data set)



- **Boostrap:** a general tool to assessing statistical accuracy
  - Estimate the variance on the statistic $S(Z)$ (Z are the data)
  - Can be used as model assessment tool, or to improve an estimator
  - **Bagging** to combine weak learners (ensemble learning)



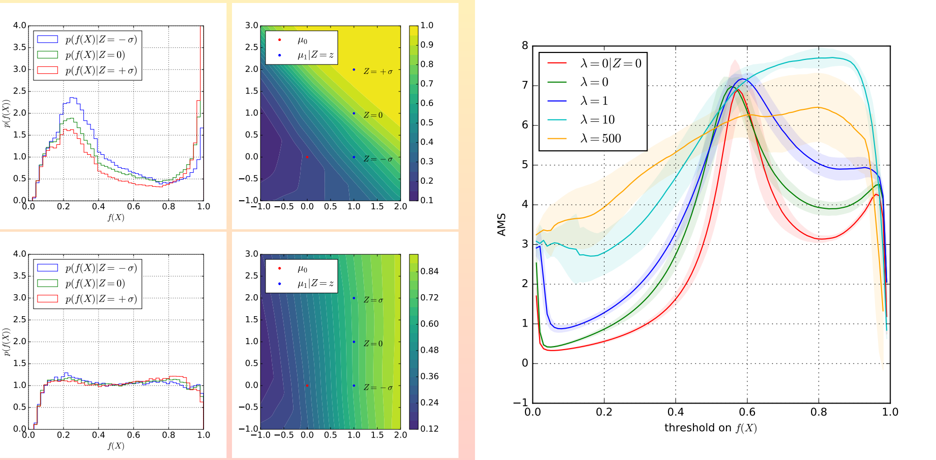
Images from Hastie, Tibshirani, Friedman

- Same method that we use for any other feature (e.g. yields, invariant mass)
  - Consider the ML algorithm as an additional feature of the data
  - For any given variation caused by a systematic uncertainty, compute alternative values for all the input features
  - Compute the ML algorithm output based on the varied features
  - Use the varied shapes as uncertainties e.g. in combine
- Issue: training the ML algorithm finds the MLE for the nominal sample
  - What we are truly interested in is the MLE **given the presence of systematic uncertainties**
  - This is a different optimization problem, with a different optimal solution
  - General property of joint optimization: $\arg\min_{a,b} f(a,b) \neq \left(\arg\min_a f(a,b), \arg\min_b f(a,b) \right)$

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

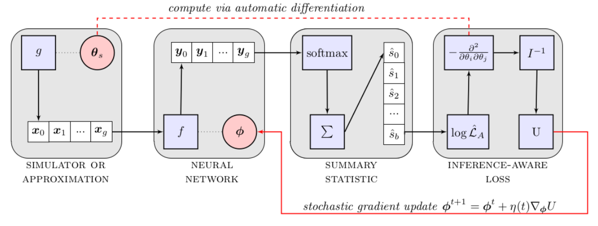
$$E_{\lambda}(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) - \lambda \mathcal{L}_r(\theta_f, \theta_r)$$

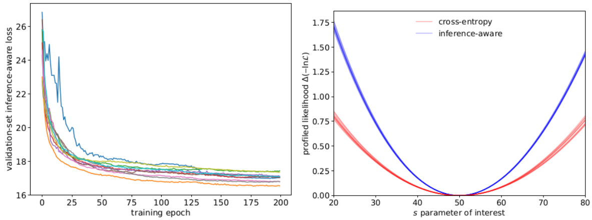


From Louppe-Kagan-Cranmer, arXiv:1611.01046

INFERNO: inference-aware neural optimization in one slide

- Build a nonparametric likelihood function based on the simulation, and use it as summary statistic
- Minimize the expected variance of the parameter of interest
  - Obtain the Fisher information matrix via automatic differentiation, and use it as loss function!
  - For (asymptotically) unbiased estimators, Rao-Cramér-Frechet (RCF) bound $V[\hat{\theta}] \sim \frac{1}{\hat{\theta}}$
  - Constraints from auxiliary measurements (i.e. systematic uncertainties) included out of the box in the covariance matrix!





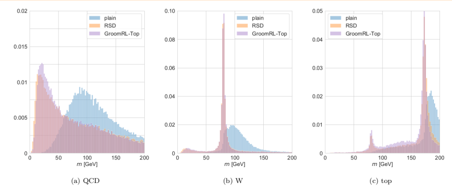
(a) inference-aware training loss

(b) profile-likelihood comparison

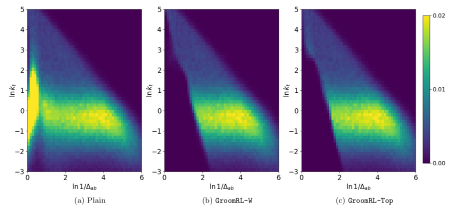
From De Castro-Dorigo, arXiv:1806.04743, and AMVA4NewPhysics deliverable 1.4 public report

Publish the details of your models

- Machine learning ultimately is based on statistical theory
- As with statistics, we must strive to use well-grounded methods, and thoroughly document them
- This is particularly true when we are interested in interpretability
  - In some cases, can check interpretation directly (e.g. fat jet grooming with reinforcement learning cross-checked in the Lund plane)
- Often we don't provide enough information
  - We may be concerned with reproducibility (e.g. "maintaining the competitive edge")
  - However, the details of the ML algorithm setup should not be considered dangerous or questionable
- We should strive to describe our algorithms, particularly when we are developing or applying a novel or yet-unexplored-in-practice method!
  - Some thoughts at doi:10.1016/j.revip.2020.100046



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



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

Images from arXiv:1903.09644

Difficult tasks for humans may be easy for artificial networks

```
[vischia@lxplus797 ~]$ python dump_cpTree.py
Info in &lt;TCanvas::MakeDefCanvas&gt;:  created default TCanvas with name c1
```

Signal

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	[illegible]	[illegible]	2.5146484	40.820312	0.8255668
56350	3	280.82870	93.600341	59.040779	-1.360351	0.7332763	-0.819335	-2.826171	0.7969970	[illegible]	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
```

Background

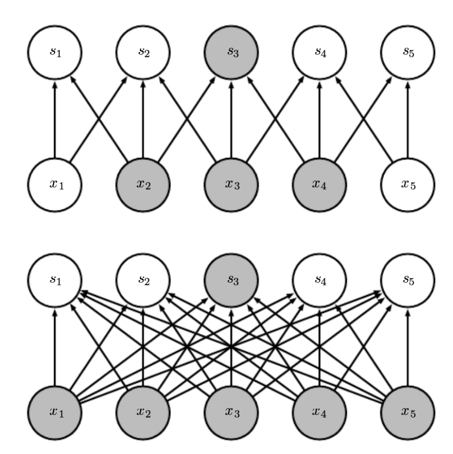
Row	Instance	Lep1_pt.L	Lep2_pt.L	Lep3_pt.L	Lep1_eta.	Lep2_eta.	Lep3_eta.	Lep1_phi.	Lep2_phi.	Lep3_phi.	met.met	met_phi.m	weight_CP	HTT_score
58751	0	86.053443	32.593345	13.077508	-1.346435	2.0512695	-0.503906	-0.392944	-2.925781	0.9309082	17.544691	0.1735229	3.478e-08	0.9726203
58751	1	86.053443	32.593345	13.077508	-1.346435	2.0512695	-0.503906	-0.392944	-2.925781	0.9309082	17.544691	[illegible]	3.478e-08	0.9726203
58751	2	86.053443	32.593345	13.077508	-1.346435	2.0512695	-0.503906	-0.392944	[illegible]	[illegible]	[illegible]	[illegible]	3.478e-08	0.9726203
58751	3	86.053443	32.593345	13.077508	-1.346435	2.0512695	-0.503906	-0.392944	[illegible]	[illegible]	[illegible]	[illegible]	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

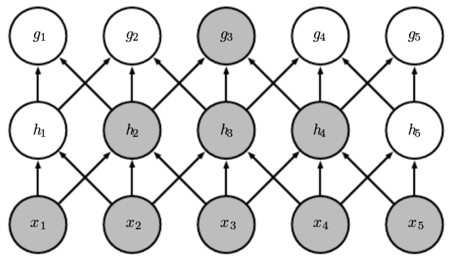
```
==> 6 selected entries
```

Image by Pietro Vischia

Image from indiatimes.com

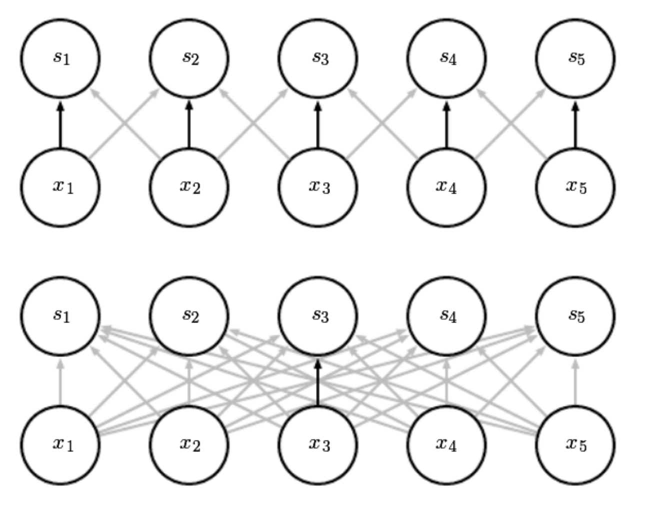
UCLouvain
Institut de recherche
en mathématique et physique





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

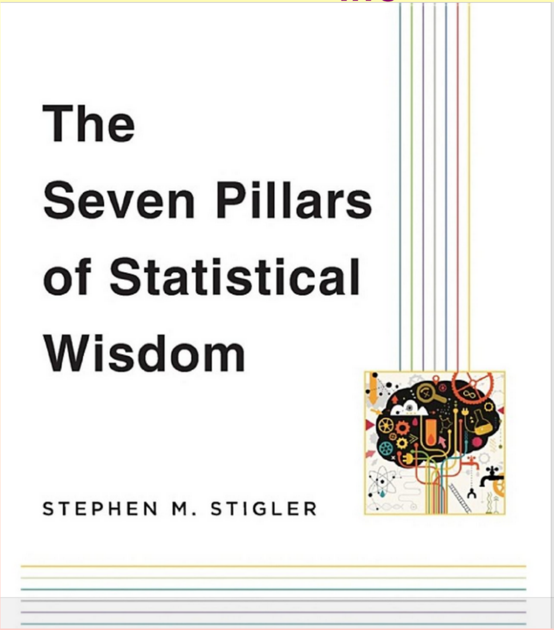
Parameter sharing and equivariance to translations

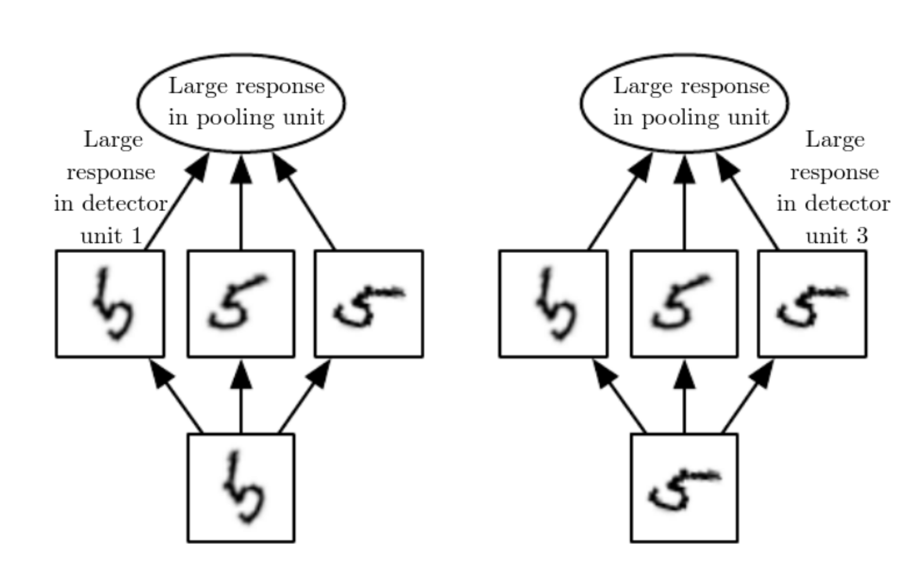


Images from https://www.deeplearningbook.org/

The first pillar

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





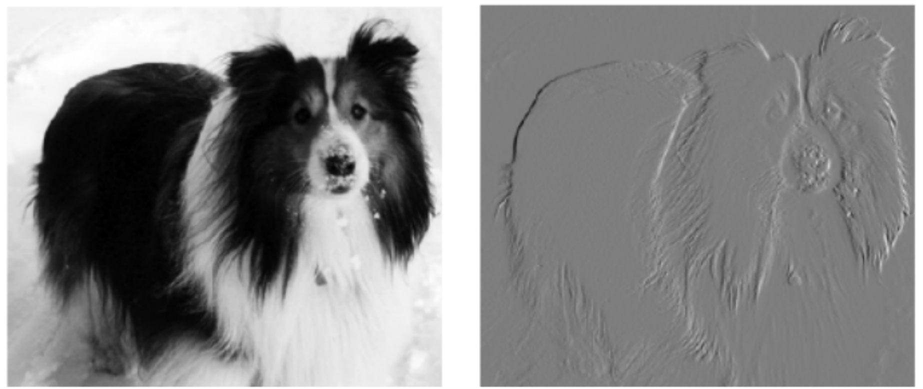
Images from https://www.deeplearningbook.org/

Convolution makes it easier to learn transformations

fnrs LA LIBERTÉ DE CHERCHER
UCLouvain
Institut de recherche
en mathématique et physique

- 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/

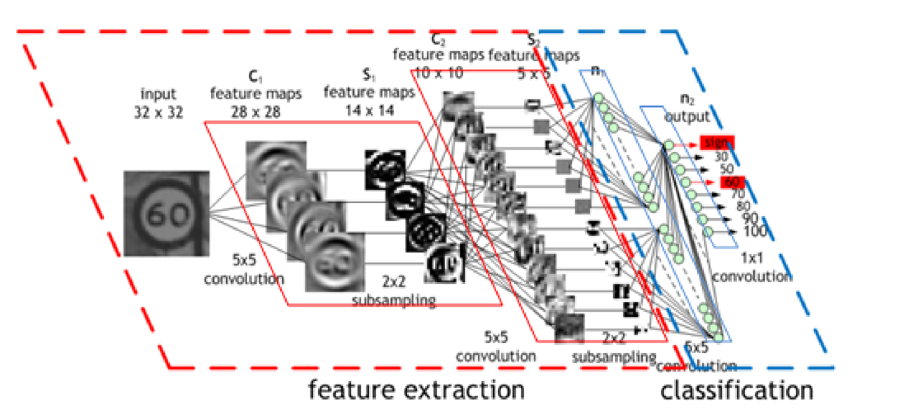
LeNet

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

LeNet

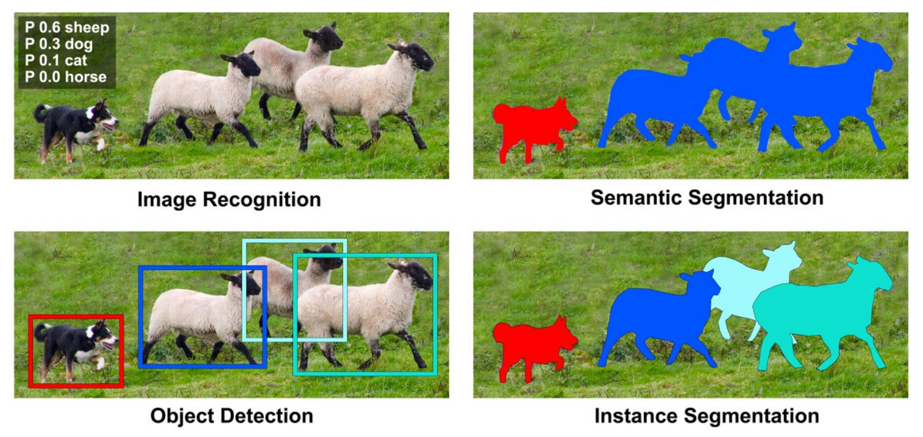
- LeNet (Yann LeCun 1998, http://yann.lecun.com/exdb/lenet/)

LeNet

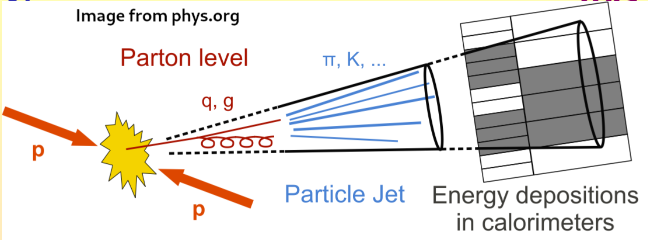


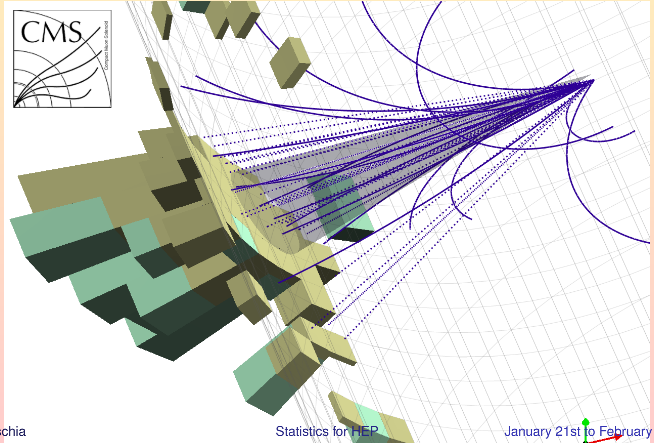
From http://parse.ele.tue.nl/education/cluster0

Segmentating objects



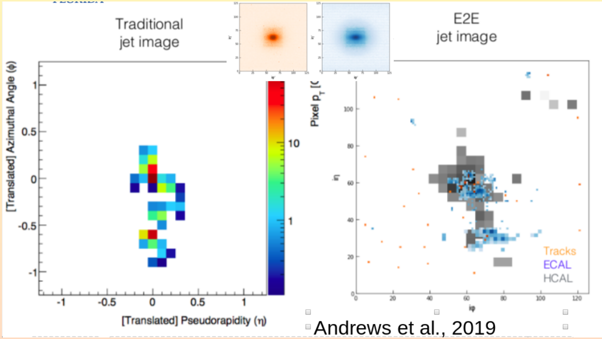
Reconstructing jets

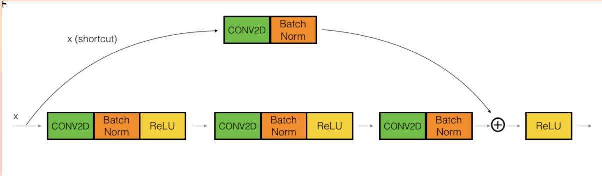


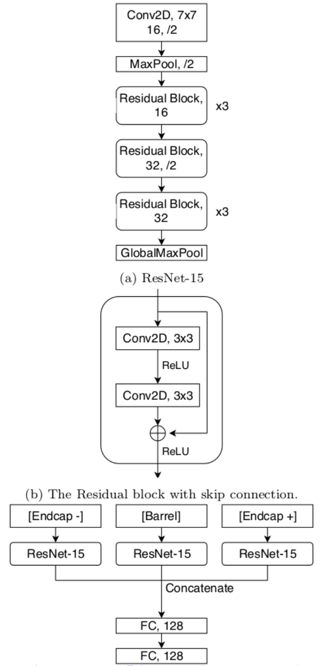


End-to-end jet reconstruction

- 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

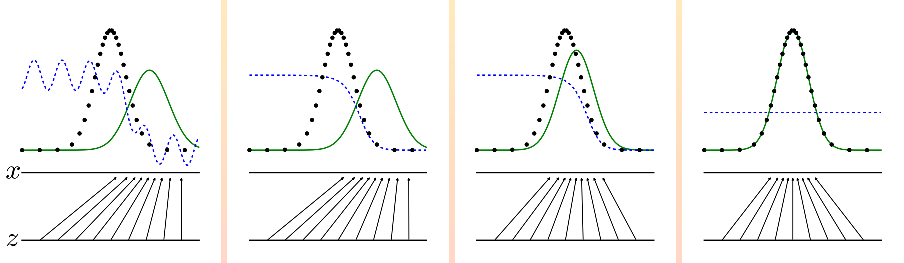






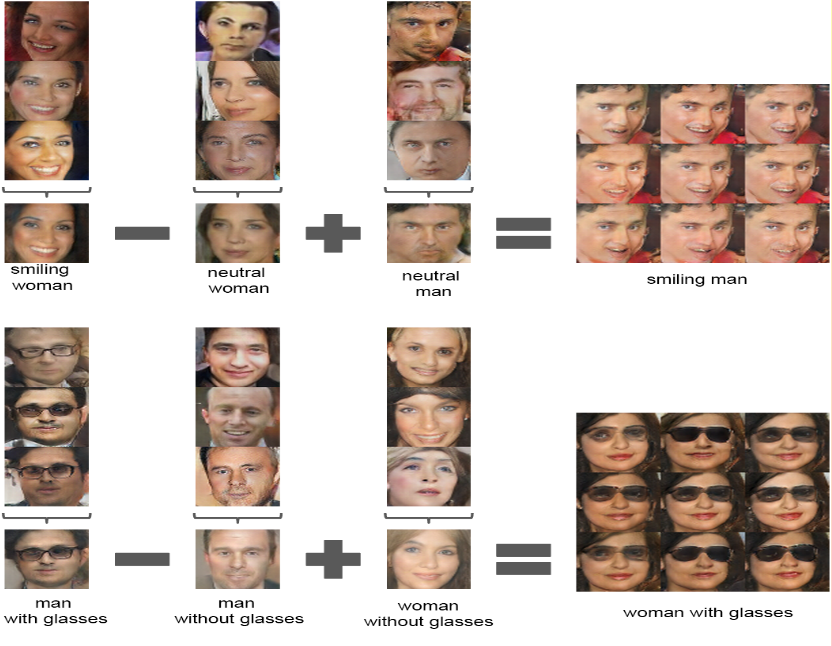
arXiv:1902.08276, S. Gleyzer's talk at 3rd IML workshop, Priya Dwivedi

- 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

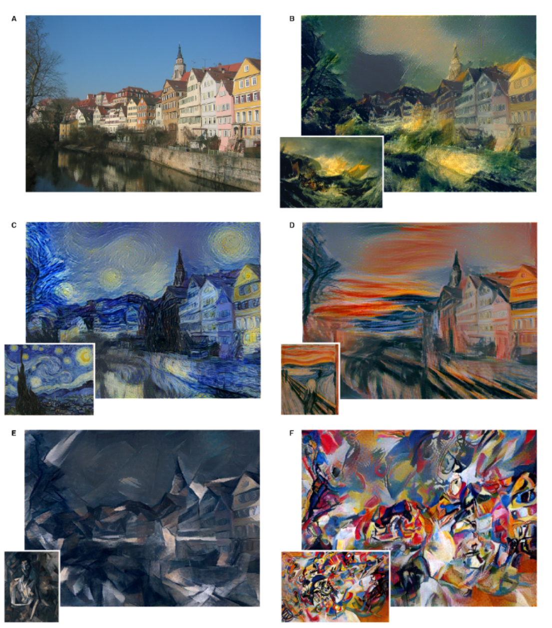
Can pick elements and combine them into new images



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)



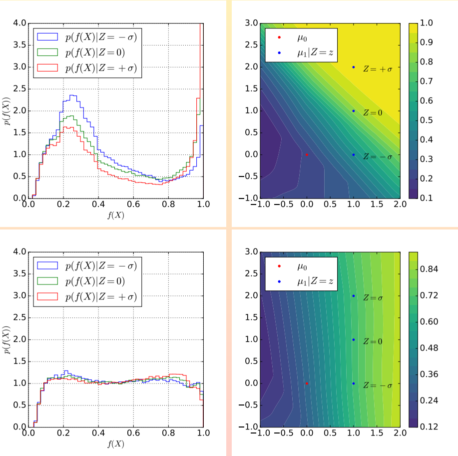
From https://arxiv.org/abs/1508.06576

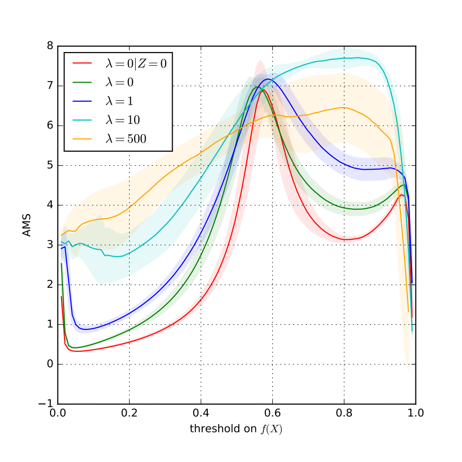
This person does not exist!

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

UCLouvain
Institut de recherche
en mathématique et physique

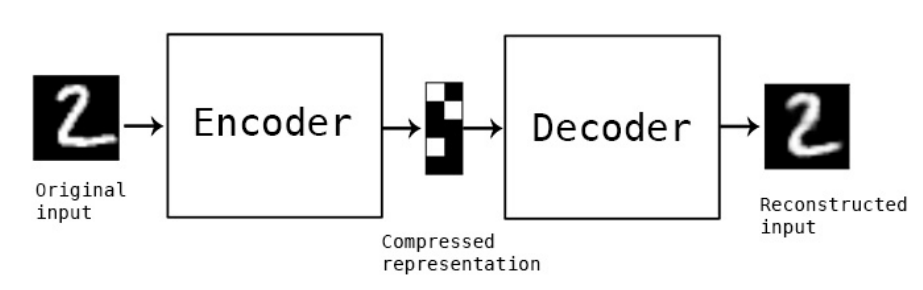
- Adversarial networks used to build pivot quantities
  - Quantities that are invariant in some parameter (typically a nuisance parameter representing a source of uncertainty)
- Best Approximate Mean Significance as tradeoff optimal/pivotal
  $E_\lambda(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) -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





From Louppe-Kagan-Cranmer, arXiv:1611.01046

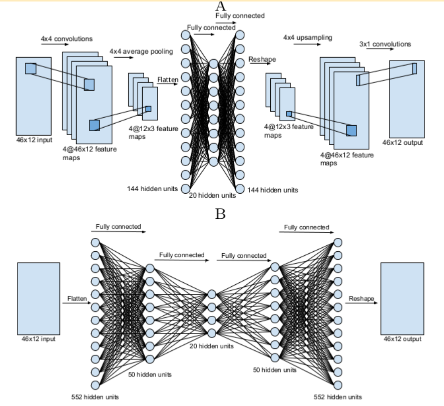
- Learn how to transform an object into almost itself

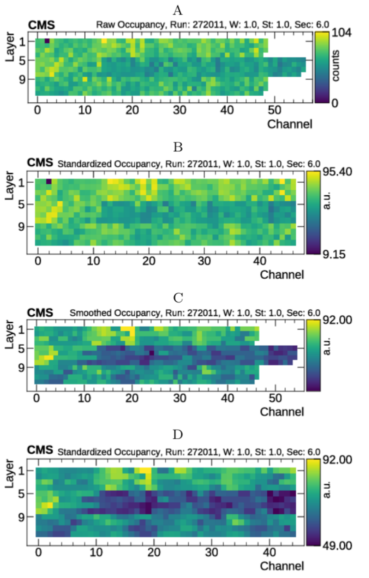


From Chollet and Allaire, Deep Learning With R

...in physics

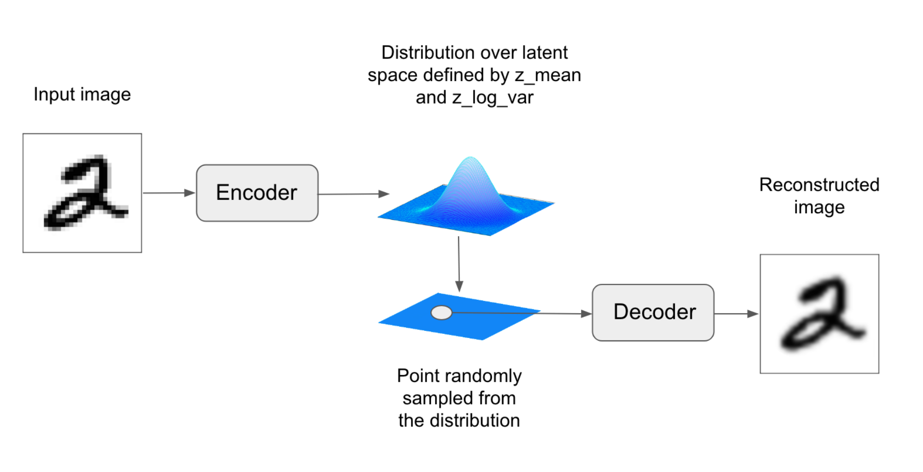
- Use it to spot objects that are different from those you have trained on
- CMS Muon Chamber detectors modelled as geographic layered maps
  - Map is an image: use **convolutional** autoencoders
  - Local approach (independent layers): spot anomalies in a layer
  - Regional approach (simultaneusly across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





From arXiv:1808.00911

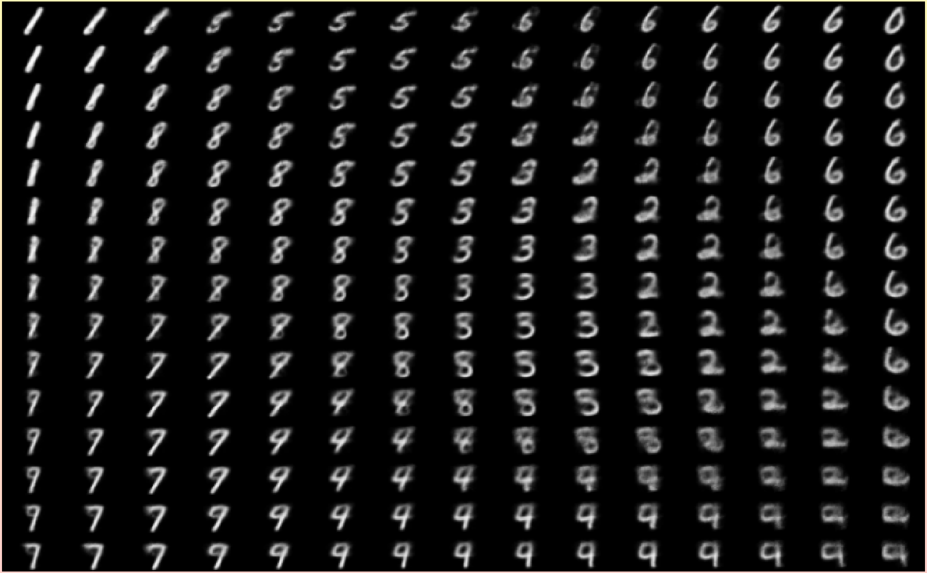
- Learn a space of continous representations of the inputs



From Chollet and Allaire, Deep Learning With R

...and Variational autoencoders

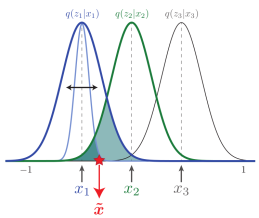
- "How do I transform a 1 into a 0?"
- Space directions have a meaning! "four-ness", "one-ness"

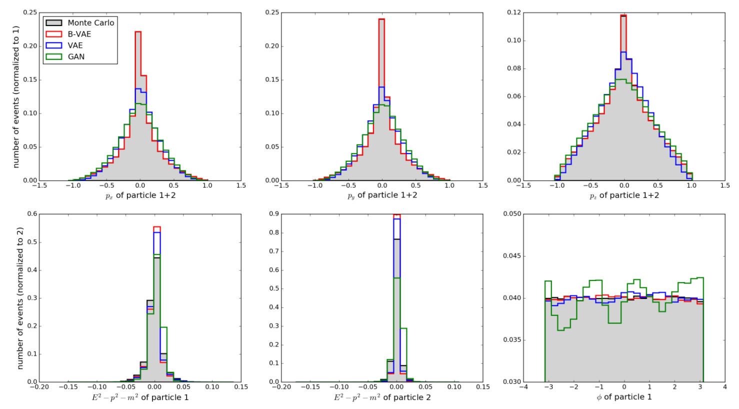


From Chollet and Allaire, Deep Learning With R

...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!



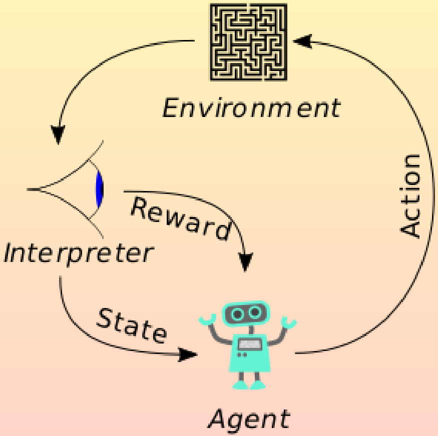


Plots from arXiv:1804.03599 and arXiv:1901.00875

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...

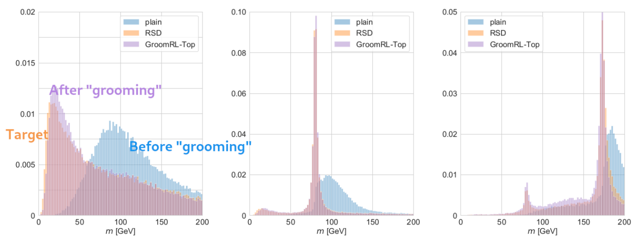
- ATARI Blackout (Google Deep Mind)
- *https://deepmind.com/research/dqn/*

...to self driving cars...

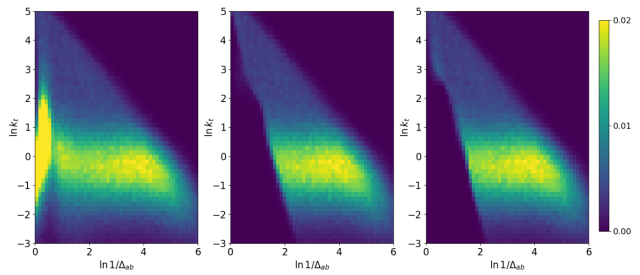
- *https://www.youtube.com/watch?v=MqUbdd7ae54*
- Build your own simulated driver: http://selfdrivingcars.mit.edu/deeptraffic/

...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



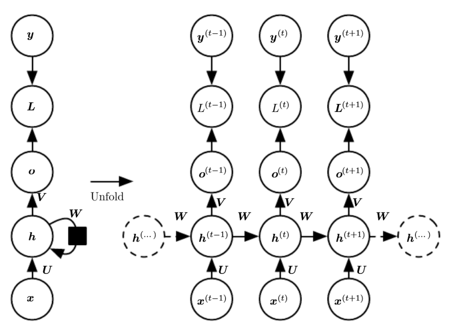
(a) QCD (b) W (c) top



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

Images from arXiv:1903.09644

- 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/



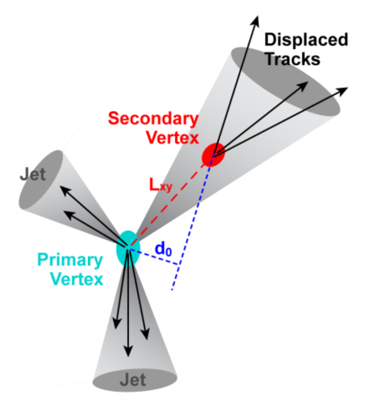

```
QUEENE:
I had thought thou hadst a Roman; for the oracle,
Thus by All bids the man against the word,
Which are so weak of care, by old care done;
Your children were in your holy love,
And the precipitation through the bleeding throne.

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

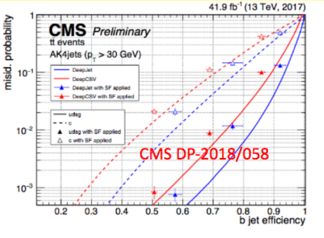
- Quarks produced in proton-proton collisions give rise to collimated “jets” of particles
- Bottom quarks travel for a while before fragmenting into jets

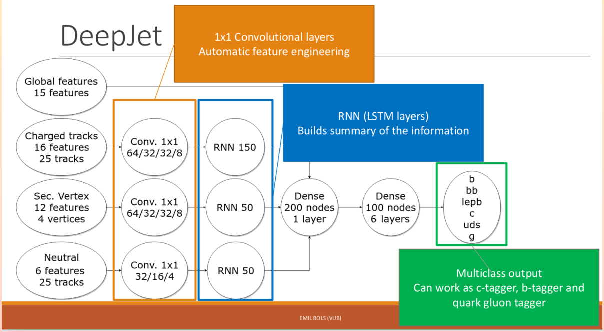


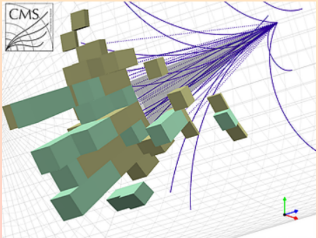
Plot from D0

...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



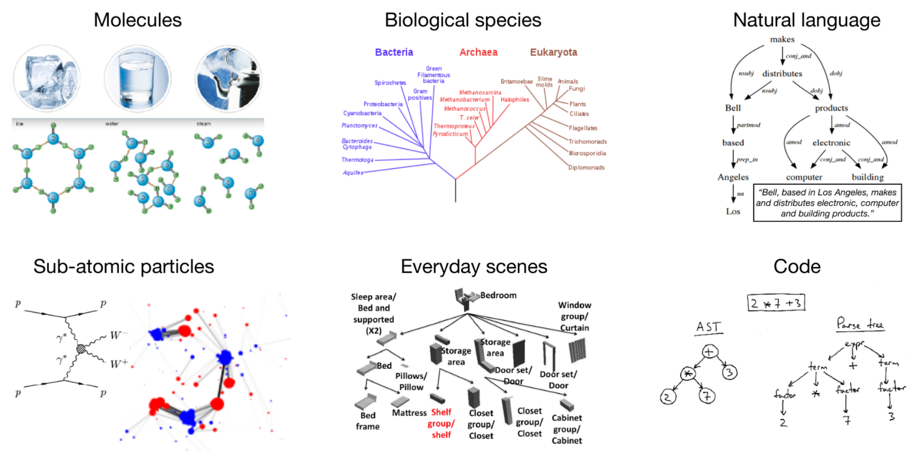




CMS DeepJet, plot from Emil Bols' talk at IML workshop

Learn graph of the underlying structure...

UCLouvain
Institut de recherche
en mathématique et physique

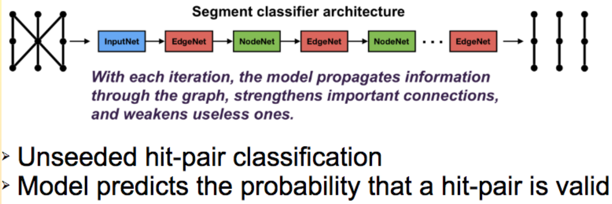


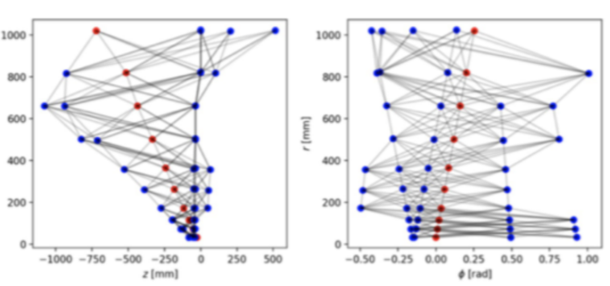
From Peter Battaglia's talk at the IML2020 Workshop

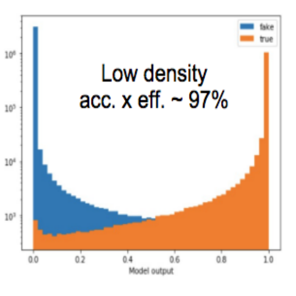
Water

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

- Graph networks to literally connect the dots







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

High-granularity calorimeter

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

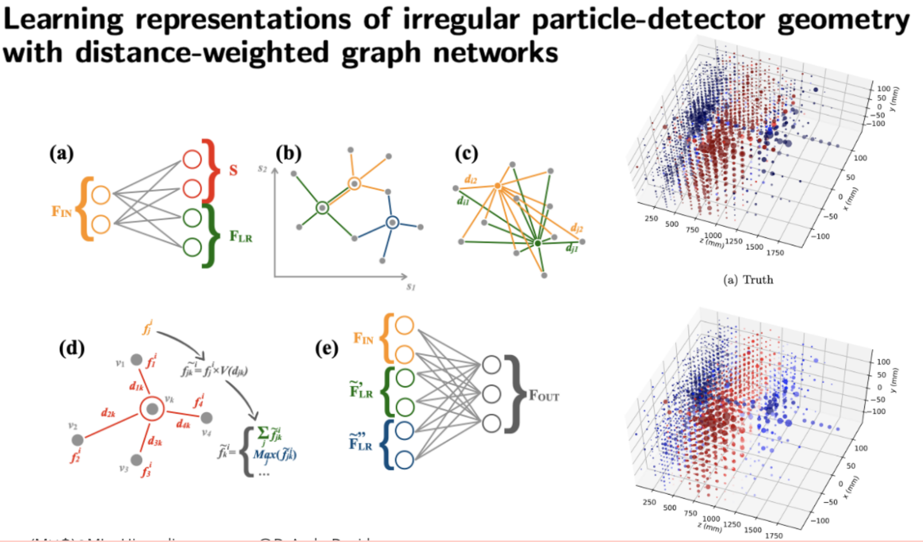


Image from a talk by André David and the HGCAL team

GPT3

- GPT3: autoregressive model with 175 billion parameters
  - Non-recurrent, attention-based (non-fixed-length sequences)
  - Standard RNN-based autoencoders have problems due to fixed-length (different languages have different information density)

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

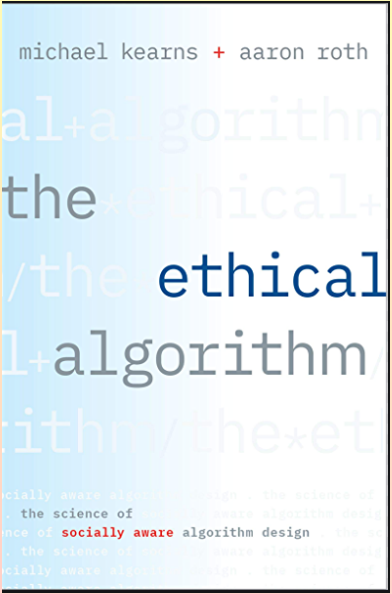
From https://arxiv.org/abs/2005.14165

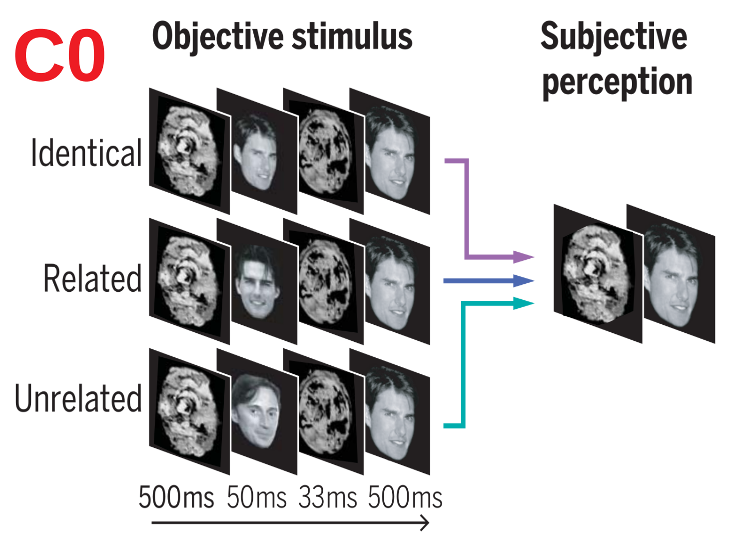
HUGE potential for misuse

UCLouvain
Institut de recherche
en mathématique et physique

- Is an image real or fake?
- Is a video real or fake?
- Is a text real or fake?
- If a self-driving car kills someone, who's fault is that?
- You can be tracked anywhere
- Your behaviour can be modelled and exploited

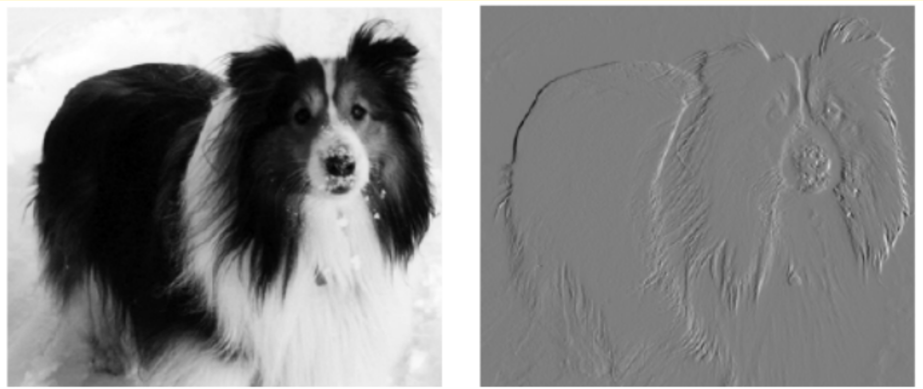
HUGE potential for misuse




C0
Objective stimulus
Subjective perception
Identical
Related
Unrelated
500ms 50ms 33ms 500ms

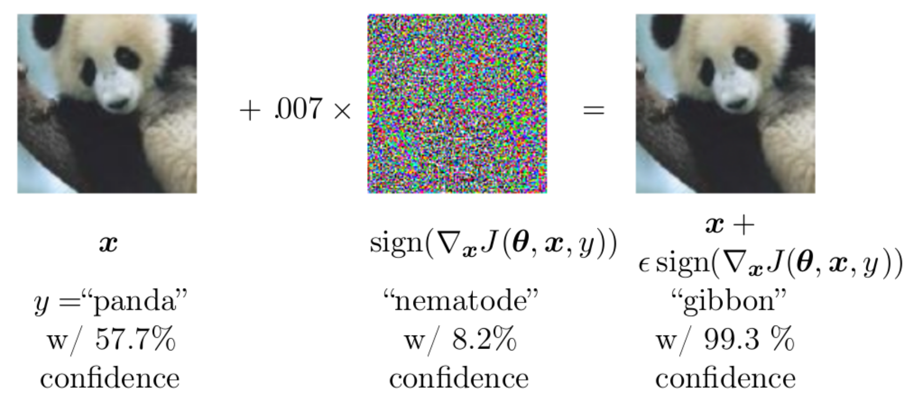
Information selection for a purpose


C2



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

...turns out we don't...



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

...not even in text generation

UCLouvain
Institut de recherche
en mathématique et physique

Q: If m x q changes to q x m , what does p a b m change to?
A: m b a p

Q: If m x q changes to q x m , what does y r q l v change to?

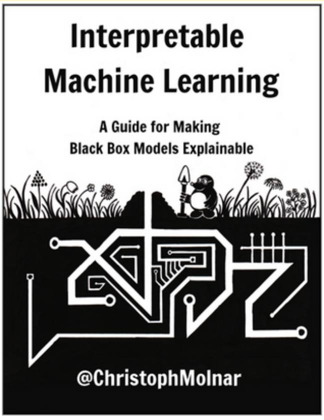
GPT-3 never gave the "reversal" answer on any of the five trials. Here are its answers:

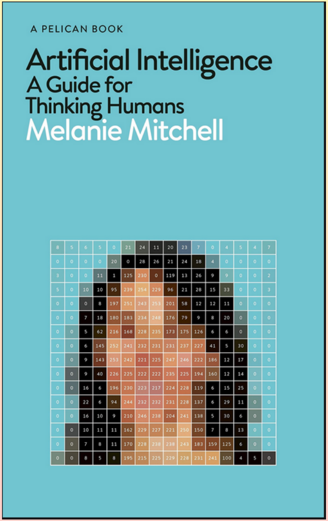
l q r y v
r l y q v
l y r q v
r y l v q
l y r q v

From https://medium.com/@melaniemitchell.me/follow-up-to-can-gpt-3-make-analogies-b202204bd292

The Importance of Interpretability

- A few resources
  - Intepretability: Christoph Molnar, Interpretable Machine Learning
  - Artificial “intelligence”: Melanie Mitchell, Artificial Intelligence: A Guide for Thinking Humans





ARE YOU READY TO INCLUDE MACHINE LEARNING IN YOUR RESEARCH?

Not before having coded a neural network from scratch!!!

This afternoon's session!

Now, if time allows: overview of the use of neural networks in physics

- I am a big fan of feedback: you'll receive in the next couple days a questionnaire
  - You'll receive it **at the email address you used for registering**
  - I'd be grateful if you could answer to the questions
  - There are also free fields for more articulated suggestions
- I will update the list of references of the last slide later today and reupload

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

- I hope this course has helped in broadening the spectrum of techniques you will consider using in the future
- Or at least that it has clarified some of the underlying concepts for techniques you already use!

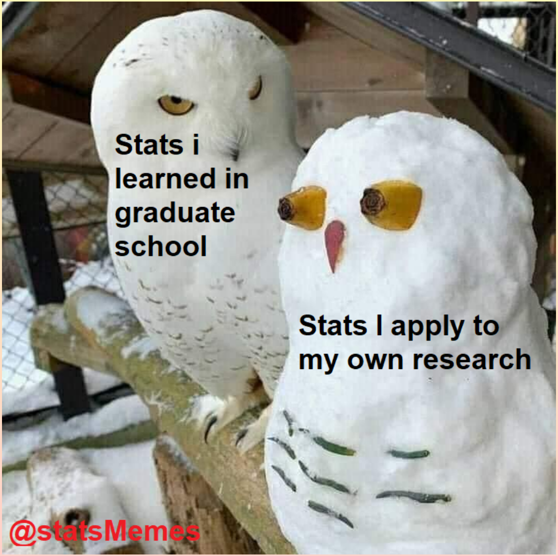


Image from the Statistical Statistics Memes Facebook Page

THANK YOU VERY MUCH FOR ATTENDING!!

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

Non-exhaustive list of references

fnrs LA LIBERTÉ DE CHERCHER — UCLouvain Institut de recherche en mathématique et physique

- 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