Statistics for HEP

Invited lectures, 12th Course on Physics of the LHC
(LIP, Lisboa, Portugal)

Dr. Pietro Vischia
pietro.vischia@cern.ch
@pietrovischia

UNIVERSIDAD DE OVIEDO
ICTEA

Lecture 1

Probability and statistics

Practicalities

- Significantly restructured with respect to the past years
  - Lecture 1: Probability and Statistics (minus hypothesis testing)
  - Lecture 2: Machine Learning (and hypothesis testing)
- More detailed material in my twenty-hours intensive course
  - It may be useful if you tried out the exercises, at your pace!
- Many references here and there, and in the last slide
  - Try to read some of the referenced papers!
  - Unreferenced stuff copyrighted P. Vischia for inclusion in my (finally) upcoming textbook

Statistics answers questions

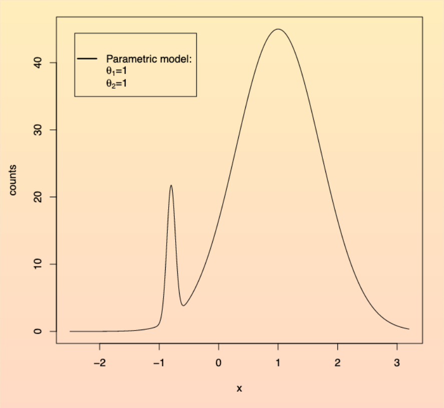
The quality of the answer depends on the quality of the question

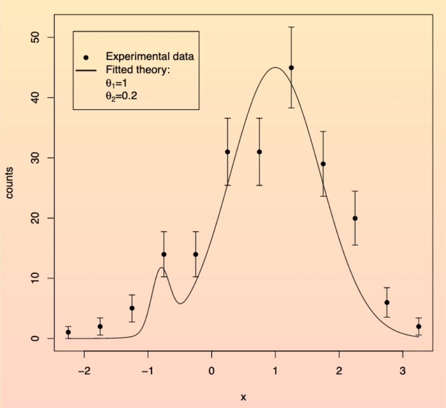
—

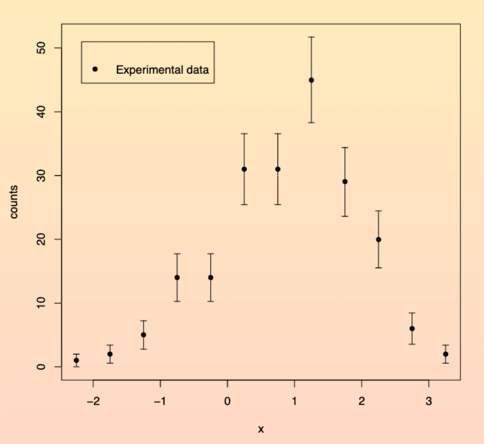
Image by ``The Tiger Lillies''

...in a mathematical way

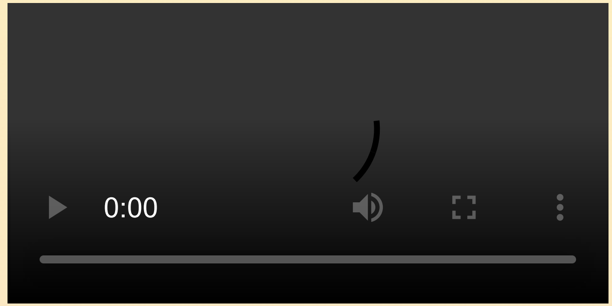
- Theory
  - Approximations
  - Free parameters
- Statistics
  - Estimate parameters
  - Quantify uncertainty
  - Test theories
- Experiment
  - Random fluctuations
  - Mismeasurements (detector effects, etc)





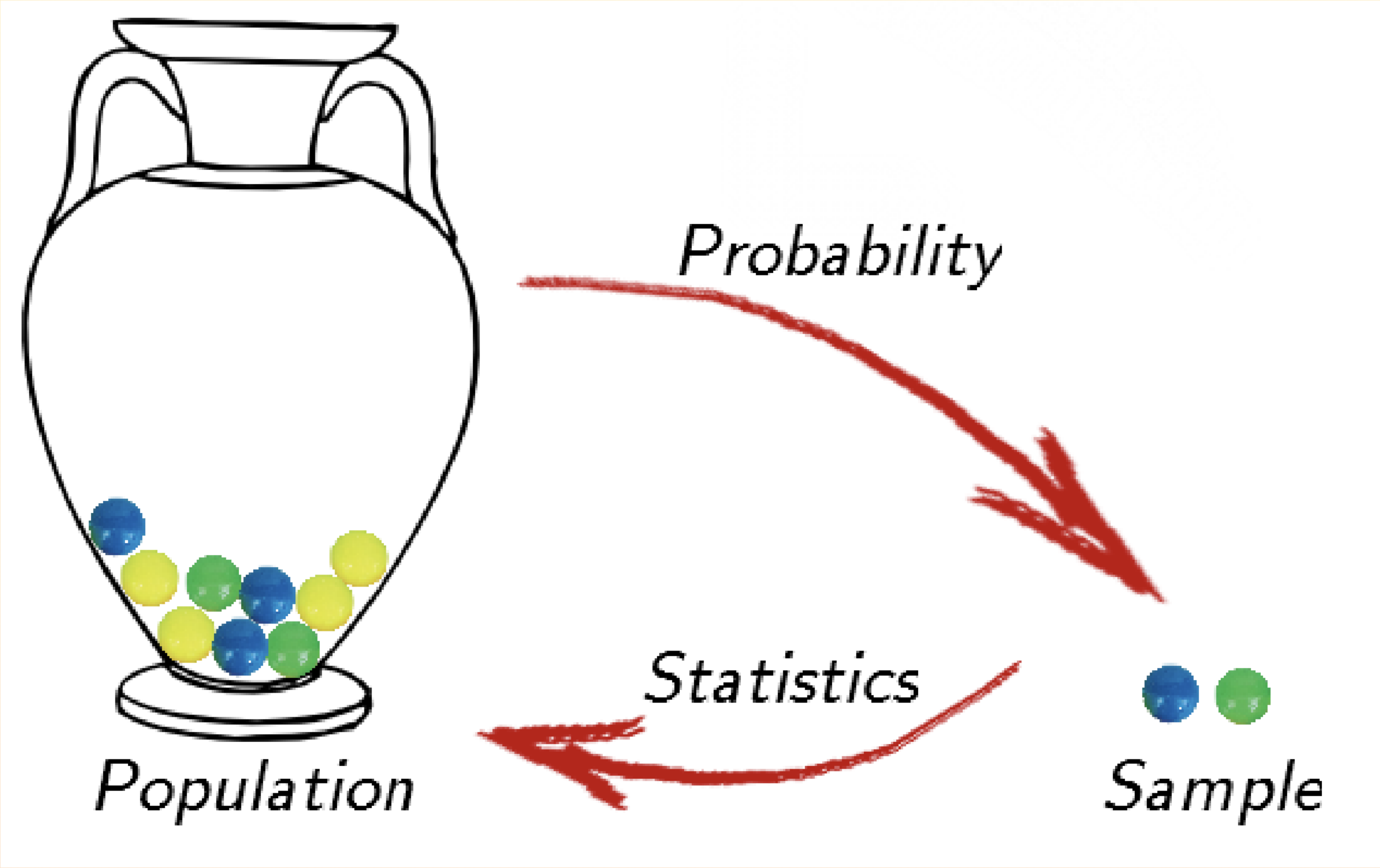


Why does Statistics work?



—
Thomas Garrity, video from YouTube

Probability and Statistics

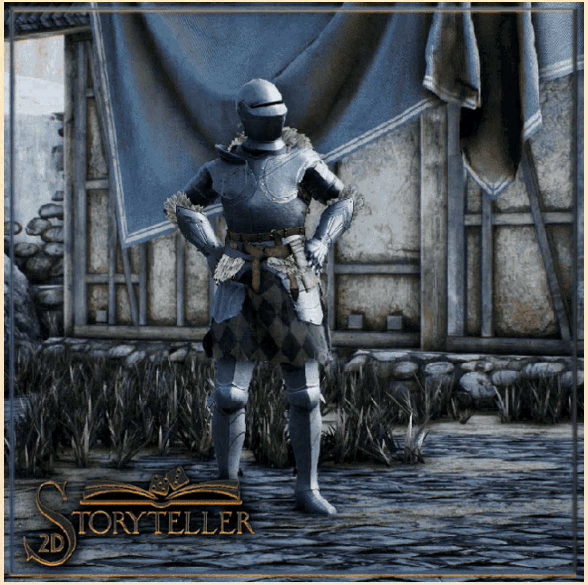


—
Picture by Rob Hall

Random Experiments

- A well-defined procedure that produces an observable outcome x that is not perfectly known

- S is the set of all possible outcomes
- S must be simple enough that we can tell whether $x \in S$ or not
- If we obtain the outcome x, then we say the event defined by $x \in S$ has occurred



- Repetitions of the experiment must happen under uniform conditions

—
GIF from Tenor.com

Axiomatic definition of probability (Kolmogorov)

- $(\Omega, \mathcal{F}, P)$: measure space
  - a set Ω with associated field (σ-algebra) $\mathcal{F}$ and measure P
  - Define a random event $A \in \mathcal{F}$ (A is a subset of Ω)

then:

1. The probability of A is a real number $P(A) \geq 0$
2. If $A \cap B = \emptyset$, then $P(A + B) = P(A) + P(B)$
3. $P(\Omega) = 1$ (probability measures are finite)

—
Image from Wikipedia, I think

Axiomatic definition for propositions (Cox and Jaynes)

- Cox, 1946: start from reasonable premises about propositions
  - $A|B$ is the plausibility of the proposition A given a related proposition B
  - $\sim A$ the proposition $not - A$, i.e. answering "no" to "is A wholly true?"
  - $F(x, y)$ is a function of two variables
  - $S(x)$ a function of one variable
- Two postulates concerning propositions
  - $C \cdot B|A = F(C|B \cdot A, B|A)$
  - $\sim V|A = S(B|A)$, i.e. $(B|A)^m + (\sim B|A)^m = 1$
- Jaynes demonstrated that these axioms are formally equivalent to the Kolmogorov ones
  - Continuity as infinite states of knowledge rather than infinite subsets

Game Theory

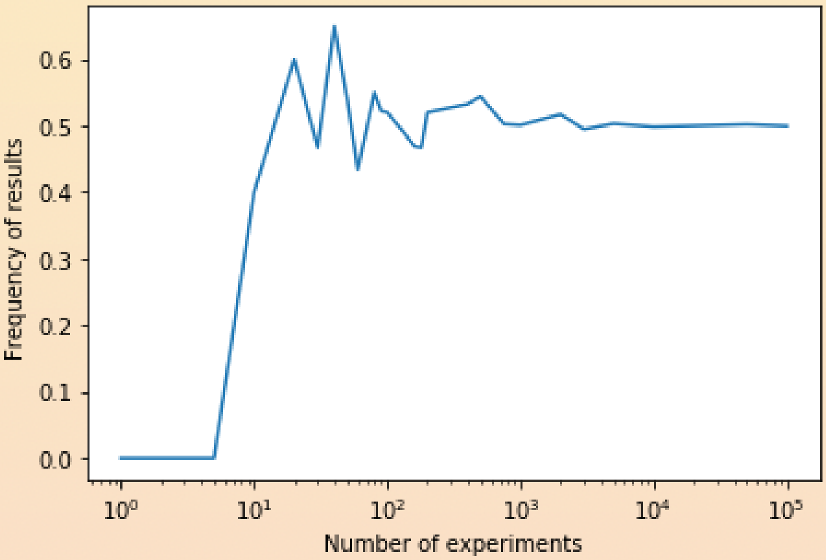
Independence

Frequentist realization

- Repeat an experiment N times, obtain n times the outcome X
- Probability as empirical limit

$$P(X) = \lim_{N\to\infty} \frac{n}{N}$$

Hand	Distinct Hands	Frequency	Probability	Cumulative probability	Odds	Mathematical expression of absolute frequency
Royal flush	1	4	0.000154%	0.000154%	649,739 : 1	$\binom{4}{1}$
Straight flush (excluding royal flush)	9	36	0.00139%	0.0014%	72,192 : 1	$\binom{10}{1}\binom{4}{1} - \binom{4}{1}$
Four of a kind	156	624	0.0240%	0.0256%	4,164 : 1	$\binom{13}{1}\binom{12}{1}\binom{4}{1}$
Full house	156	3,744	0.1441%	0.17%	693 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}$
Flush (excluding royal flush and straight flush)	1,277	5,108	0.1965%	0.367%	508 : 1	$\binom{13}{5}\binom{4}{1} - \binom{10}{1}\binom{4}{1}$
Straight (excluding royal flush and straight flush)	10	10,200	0.3925%	0.76%	254 : 1	$\binom{10}{1}\binom{4}{1}^5 - \binom{10}{1}\binom{4}{1}$
Three of a kind	858	54,912	2.1128%	2.87%	46.3 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{2}\binom{4}{1}^2$
Two pair	858	123,552	4.7539%	7.62%	20.0 : 1	$\binom{13}{2}\binom{4}{2}^2\binom{11}{1}\binom{4}{1}$
One pair	2,860	1,098,240	42.2569%	49.9%	1.37 : 1	$\binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3$
No pair / High card	1,277	1,302,540	50.1177%	100%	0.995 : 1	$\left[\binom{13}{5} - 10\right]\left[\binom{4}{1}^5 - 4\right]$
Total	7,462	2,598,960	100%	—	0 : 1	$\binom{52}{5}$



—
Left: lost the reference, if you have info please let me know!. Right: P. Vischia

Subjective ("Bayesian") realization

- $P(X)$ is the subjective degree of belief in the outcome of a random experiment (in X being true)
  - Update your degree of belief after an experiment
- De Finetti: operative definition, based on the concept of coherent bet
  - Assume that if you bet on X, you win a fixed amount of money if X happens, and nothing (0) if X does not happen

$$P(X) := \frac{\text{The largest amount you are willing to bet}}{\text{The amount you stand to win}}$$

- Coherence is when the bet is fair, i.e. it doesn't guarantee an average profit/loss

Dutch book

Book	Odds	Probability	Bet	Payout
Trump elected	Even (1 to 1)	$1/(1+1) = 0.5$	20	$20+20=40$
Clinton elected	3 to 1	$1/(1+3) = 0.25$	10	$10+30=40$
All outcomes	---	$0.5+0.25=0.75$	30	40

Random variables...

- Numeric label for each element in the space of possible outcomes
  - In Physics, we usually assume Nature is continuous, and discreteness comes from our experimental limitations
- Work with probability density functions (p.d.f.s) normalized with respect to the interval

$$f(X) := \lim_{\Delta X \to 0} \frac{P(X)}{\Delta X}$$

$$P(a < X < b) := \int_a^b f(X) dX$$

... in many dimensions

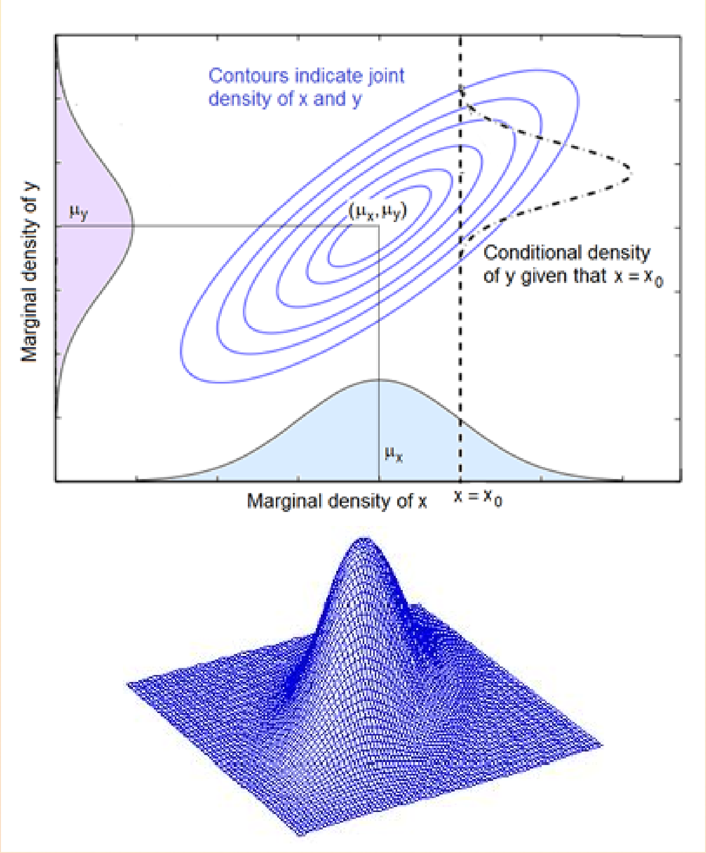
- Joint pdf for many variables: $f(X, Y, \dots)$

- Marginal pdf
  integrate over the uninteresting variables

$$f_X(X) := \int f(X,Y)dY$$

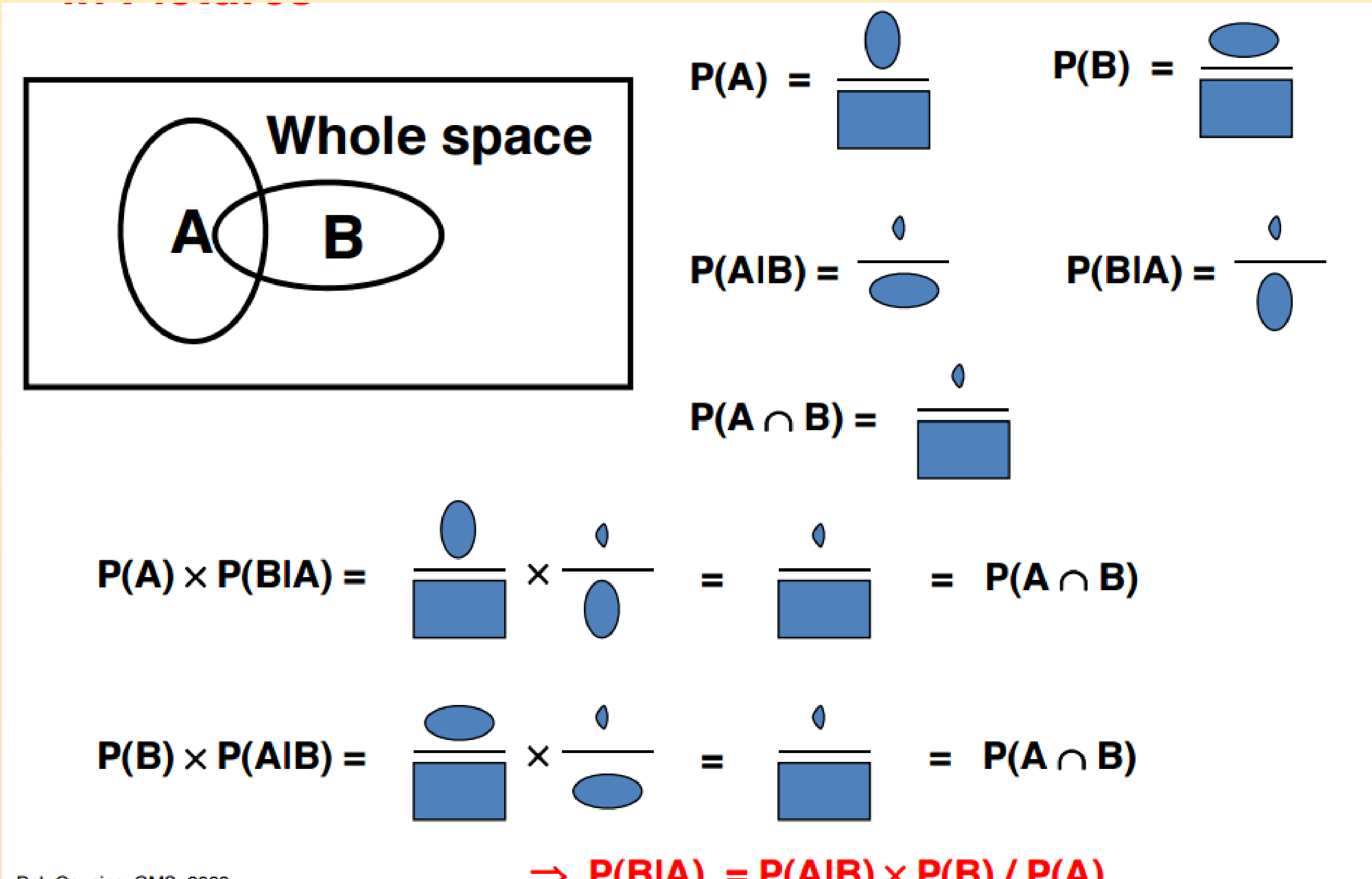
- Conditional pdf
  fix the value of the uninteresting variables

$$f(X|Y) := \frac{f(X,Y)}{f_Y(Y)}$$



—
Figure from prioritysystems.com

Bayes Theorem



- Venn diagrams were also the basis of Kolmogorov approach (Jaynes, 2003)

Diagram by Cousins, 2008 and 2018

Independence

- Two events A and B are independent if $P(AB) = P(A)P(B)$
  - Can be assumed (e.g. assume that coin tosses are independent)
  - Can be derived (verifying that equality holds)
  - E.g. if $A = \{2, 4, 6\}$, $B = \{1, 2, 3, 4\}$, we have $P(AB) = 1/3 = P(A)P(B)$
- Two disjoint outcomes with positive probability cannot be independent

$$P(AB) = P(\emptyset) = 0 \neq P(A)P(B) > 0$$

Law of Total Probability

- Bayes theorem is valid for any probability measure

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

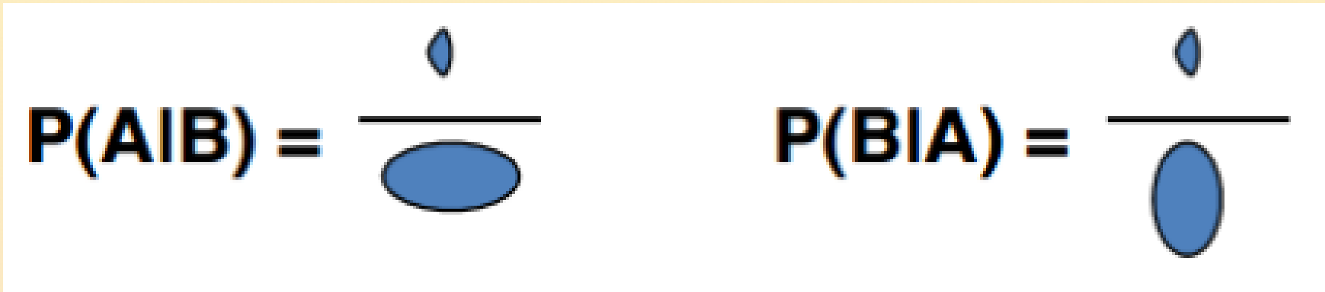
- Useful decomposition by partitioning S in disjoint sets A_i
  - $\cap A_i A_j = 0 \qquad \forall i, j$
  - $\cup_i A_i = S$

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$

- The Bayes theorem becomes

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

A Word of Advice



$$P(A|B) \neq P(B|A)$$

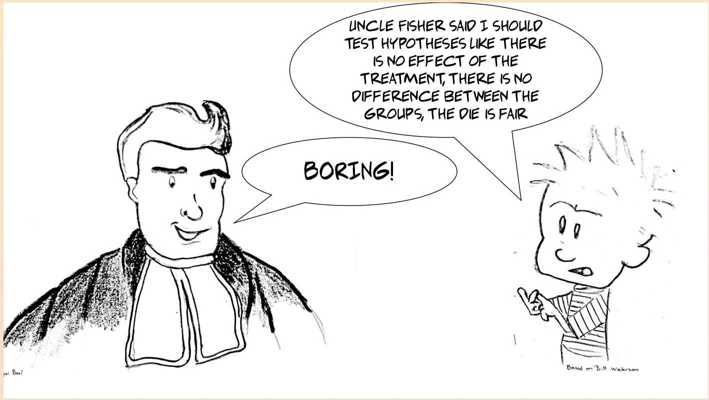
- $P(have\, TOEFL|speak\ English)$ is very small, say $<< 1\%$
- $P(speak\ English|have\, TOEFL)$, is (hopefully) $\sim 100\%$

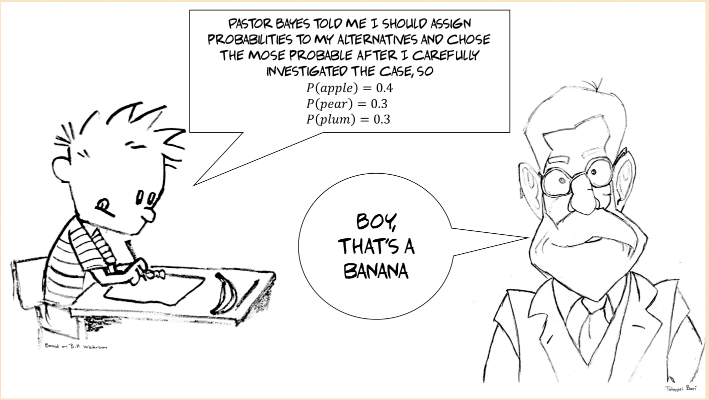
Another Word of Advice

—
Image from r/dataisugly

P(outcome), P(hypothesis)

- Frequentist probability (Fisher) always refers to outcomes in repeated experiments
    - $P(hypothesis)$ is undefined
    - Criticism: statistical procedures rely on complicated constructions (pseudodata from hypotetical experiments)
- Bayesian probability assigns probabilities also to hypotheses
    - Statistical procedures intrinsically simpler
    - Criticism: subjectivity





Intrinsically different statements

- The probability for the hypothesis to be true, given the observed data I collected, is 80%
- The probability that, when sampling many times from the hypothesis, I would obtain pseudodata similar to the data I have observed is 80%

Some history

- Bayes' 1763 (posthumous) article explains the theorem in a game of pool
- A full system for subjective probabilities was (likely independently) developed and used by Laplace
- Laplace in a sense is the actual father of Bayesian statistics


—
Stigler (1996) and McGrayne (2011)

The Obligatory COVID-19 slide

- Mortal disease
  - D: the patient is diseased (sick)
  - H: the patient is healthy
- Diagnostic test
  - $+$: the patient flags positive to the disease
  - $-$: the patient flags negative to the disease
- A very good test
  - $P(+|D) = 0.99$
  - $P(+|H) = 0.01$

The Obligatory COVID-19 slide

- Mortal disease
  - D: the patient is diseased (sick)
  - H: the patient is healthy
- Diagnostic test
  - $+$: the patient flags positive to the disease
  - $-$: the patient flags negative to the disease
- A very good test
  - $P(+|D) = 0.99$
  - $P(+|H) = 0.01$
- You take the test and you flag positive: do you have the disease?

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D)+P(+|H)P(H)}$$

- We need the incidence of the disease in the population, $P(D)$!
  - $P(D) = 0.001$ (very rare disease): then $P(D|+) = 0.0902$, which is fairly small
  - $P(D) = 0.01$ (only a factor 10 more likely): then $P(D|+) = 0.50$, which is pretty high
  - $P(D) = 0.1$: then $P(D|+) = 0.92$, almost certainty!

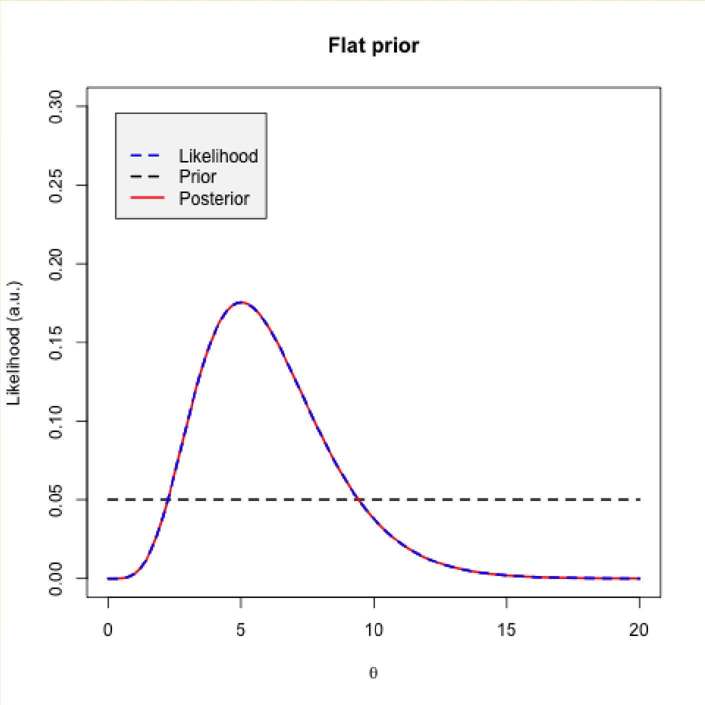
Naming Bayes

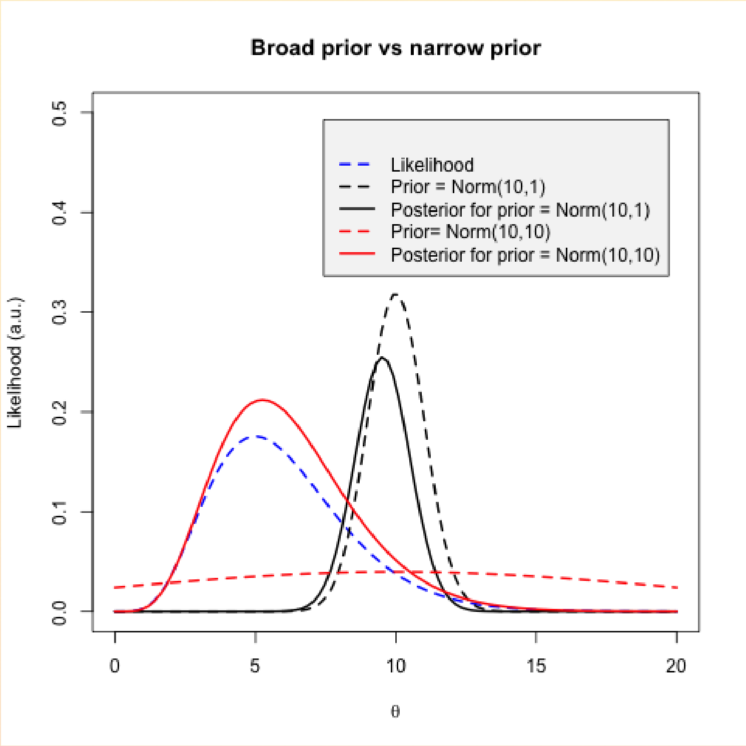
$$P(H|\vec{X}) := \frac{P(\vec{X}|H)\pi(H)}{P(\vec{X})}$$

- $\vec{X}$, the vector of observed data
- $P(\vec{X}|H)$, the likelihood function, encoding the result of the experiment
- $\pi(H)$, the probability we assign to H before the experiment
- $P(\vec{X})$, the probability of the data
  - usually expressed using the law of total probability
    $$\sum_i P(\vec{X}|H_i) = 1$$
  - often omitted when normalization is not important, i.e. searching for mode rather than integral
    $$P(H|\vec{X}) \propto P(\vec{X}|H)\pi(H)$$
- $P(H|\vec{X})$, the posterior probability, after the experiment
  - For a parametric $H(\theta)$, often written $P(\theta)$

Prior, Likelihood, and Posterior

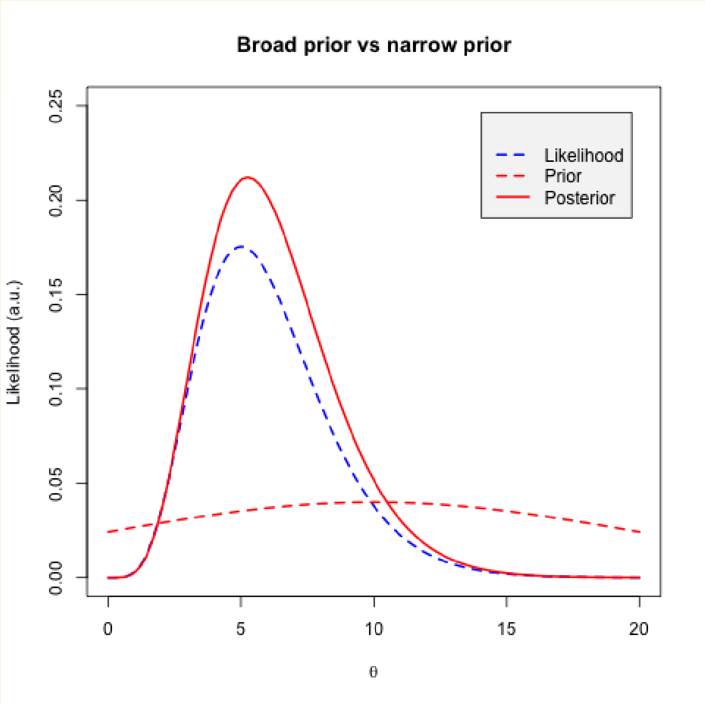
- Likelihood is always the same: usually it is the frequentist answer

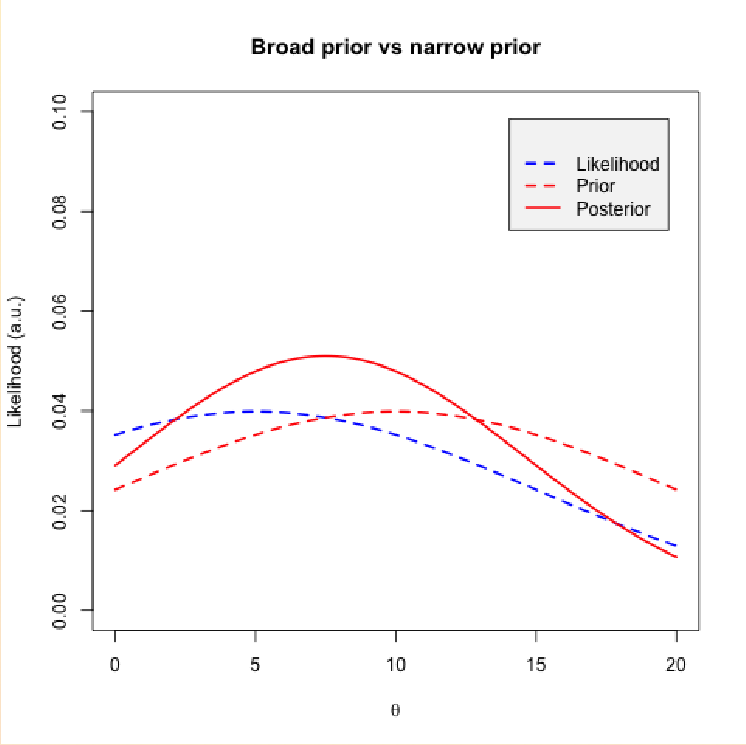




Prior, Likelihood, and Posterior

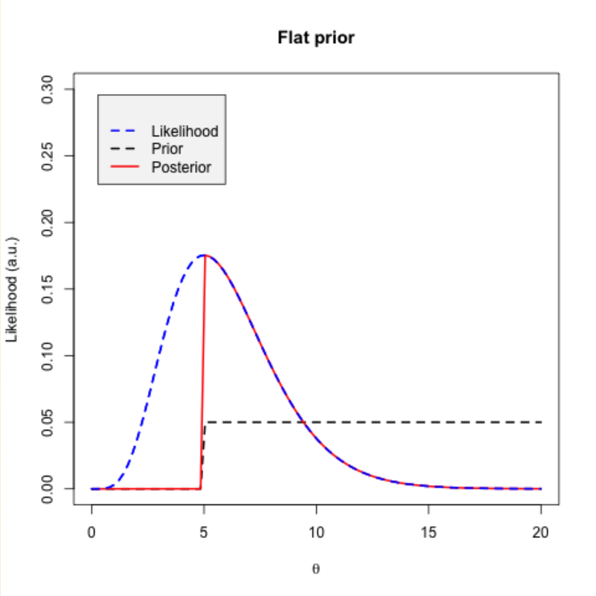
- Likelihood is always the same: usually it is the frequentist answer





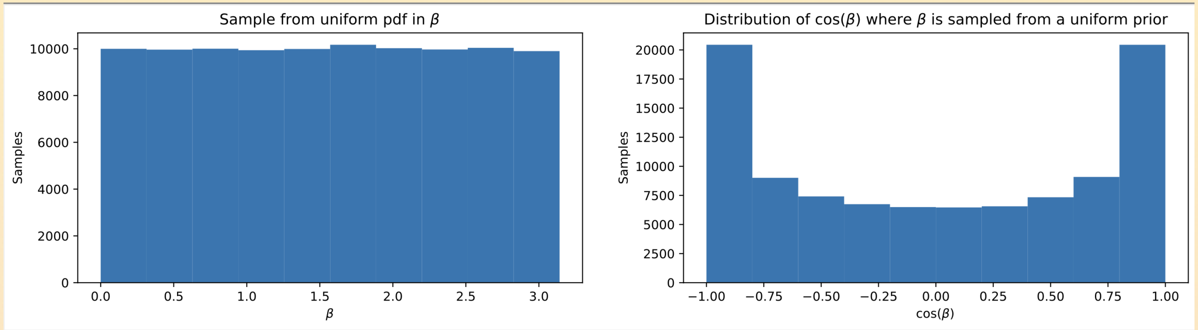
Priors to represent boundaries

- Can encode physical boundaries in the model
  - positivity of the mass of a particle
  - cross section is positive definite
- Strong assumptions on the model can hide weaknesses or anomalies
  - a transition probability such as V_{tb} is defined in $[0, 1]$ only if you assume the standard model



Representing ignorance

- Ignorance depends on the parameterization



- Elicitation of expert opinion

- *Jeffreys priors*
  - *Compute* information *on the parameter*
  - *Find a parameterization that keeps it constant*

—

Figure on expert elicitation from Mikkola (2021)

Information (Fisher)

- Information should increase with the number of observations
  - 2x data, 2x information (if data are independent)
- Information should be conditional on the hypothesis we are studying
  - $I = I(\theta)$, irrelevant data should carry zero information on θ
- Information should be related to precision
  - Larger information should lead to better precision

- Formal equivalence with other definitions (e.g. Shannon)

The Likelihood Principle

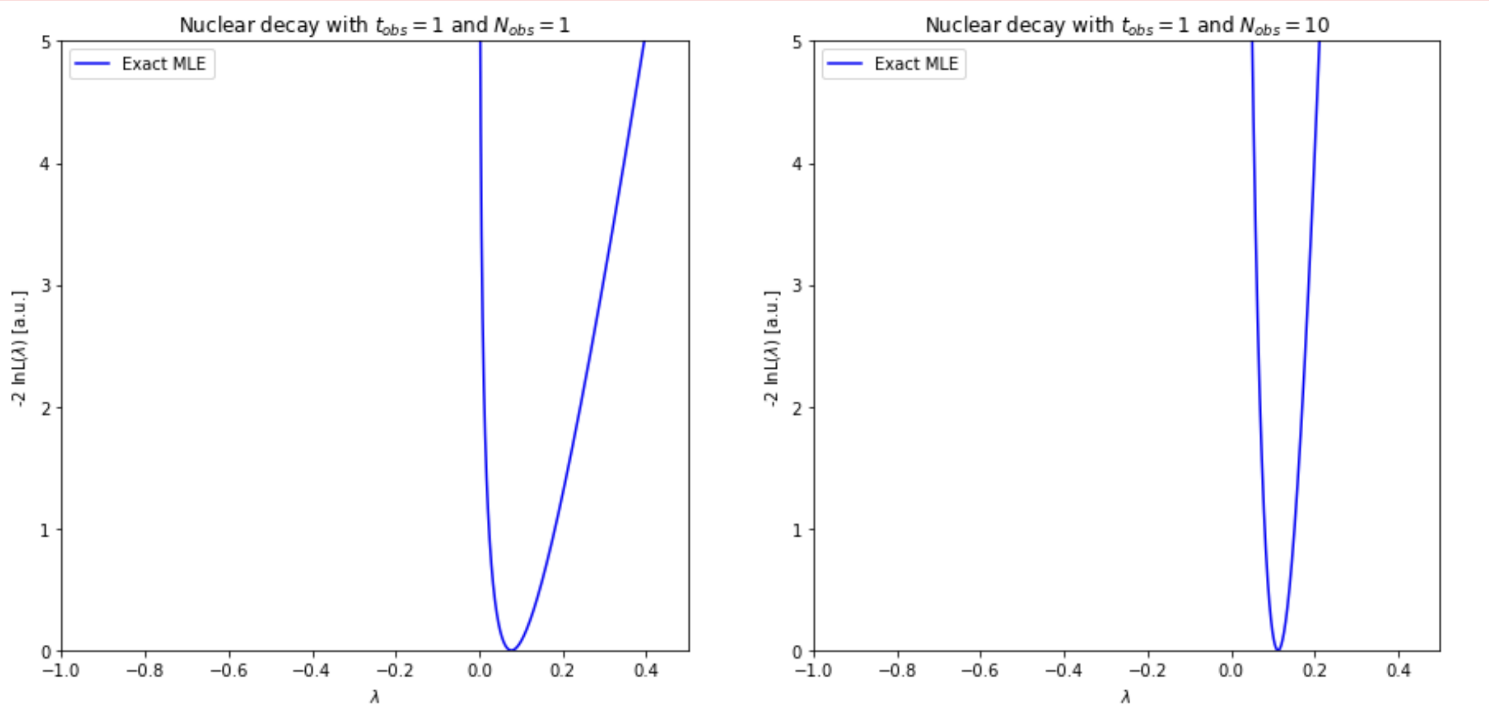
- Data sample $\vec{x}_{obs}$

$$\mathcal{L}(\vec{x};\theta) = P(\vec{x}|\theta)|_{\vec{x}obs}$$

- The likelihood function $L(\vec{x};\theta)$ contains all the information available in the data sample relevant for the estimation of θ
  - Automatically satisfied by Bayesian statistics: $P(\theta|\vec{x}) \propto L(\vec{x};\theta) \times \pi(\theta)$
  - Frequentist typically make inference in terms of hypothetical data (likelihood not the only source of information)
- Does randomness arise from our imperfect knowledge or is it an intrinsic property of Nature?

Likelihood and Fisher Information

- Define Fisher information via the curvature of the likelihood function, $\frac{\partial^2 \mathcal{L}(X;\theta)}{\partial\theta^2}$
  - Larger when there are more data
  - Conditional on the parameter studied
  - Larger when the spread is smaller (larger precision)



More formally...

- Score: $S(X;\theta) = \frac{\partial}{\partial\theta} lnL(X;\theta)$
- Fisher information as variance of the score

$$I(\theta) = E\Big[\Big(\frac{\partial}{\partial\theta} lnL(X;\theta)\Big)^2 |\theta_{true}\Big] = \int \Big(\frac{\partial}{\partial\theta} lnf(x|\theta)\Big)^2 f(x|\theta)dx \geq 0$$

- Under some regularity conditions (twice differentiability, differentiability of integral, support indep. on θ)

$$I(\theta) = -E\Big[\Big(\frac{\partial^2}{\partial\theta^2} lnL(X;\theta)\Big)^2 |\theta_{true}\Big]$$

Jeffreys Priors and Information

- Reparameterization: $\theta \to \theta'(\theta)$, when $\pi(\theta') := E\left[\left(\frac{\partial lnN}{\partial\theta'}\right)^2\right]$

$$\pi(\theta) = \pi(\theta')\left|\frac{d\theta'}{d\theta}\right| \propto \sqrt{E\left[\left(\frac{\partial lnN}{\partial\theta'}\right)^2\right]\left|\frac{\partial\theta'}{\partial\theta}\right|} = \sqrt{E\left[\left(\frac{\partial lnL}{\partial\theta'}\frac{\partial\theta'}{\partial\theta}\right)^2\right]}$$

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

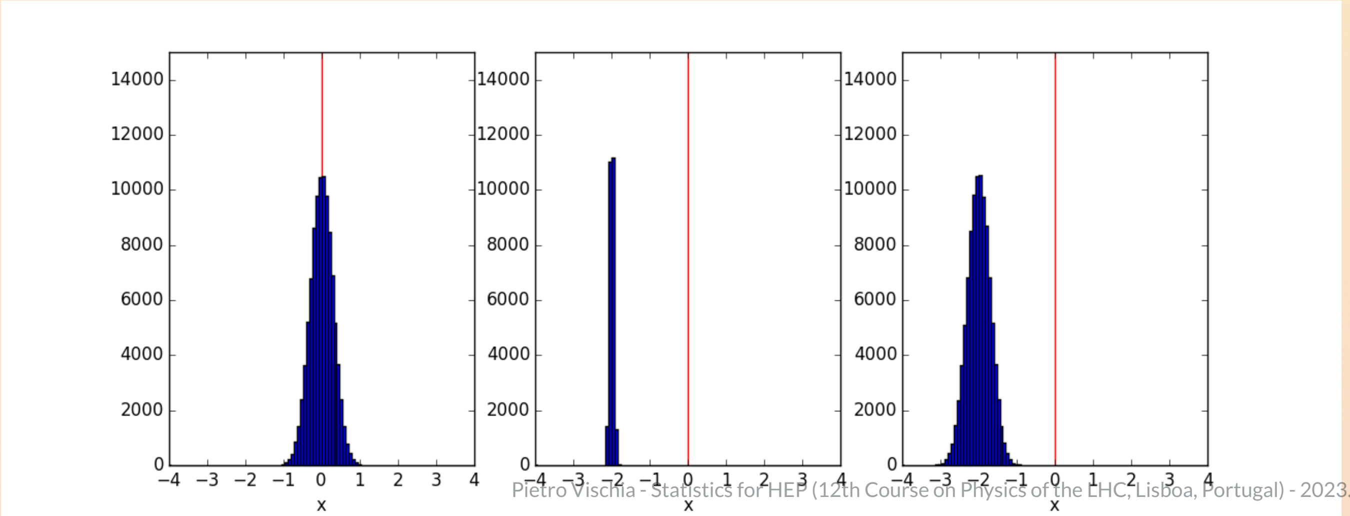
- To keep information constant, define prior via the information
  - Location parameters: uniform prior
  - Scale parameters: prior $\propto \frac{1}{\theta}$
  - Poisson processes: prior $\propto \frac{1}{\sqrt{\theta}}$
- The authors of STAN maintain a nice set of recommendations on priors

Location and Dispersion

- Draw inference on a population using a sample of experiment outcomes
  - Location ("where are most values concentrated at?")
  - Dispersion ("how spread are the values around the center?")

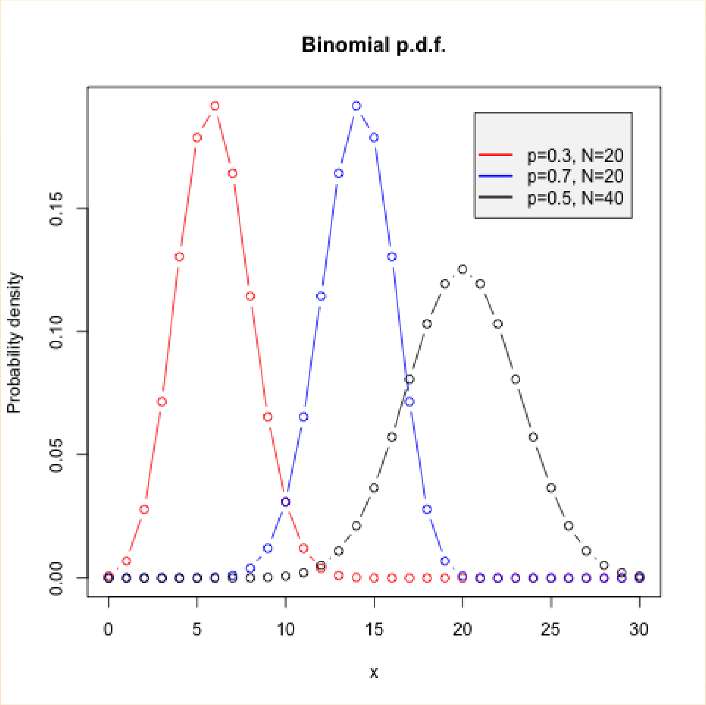
- Types of uncertainty
  - Error: deviation from the true value (bias)
  - Uncertainty: spread of the sampling distribution

- Sources of uncertainty
  - Random ("statistical"): randomness manifests as distribution spread
  - Systematic: wrong measurement manifests as bias



Binomial Distribution

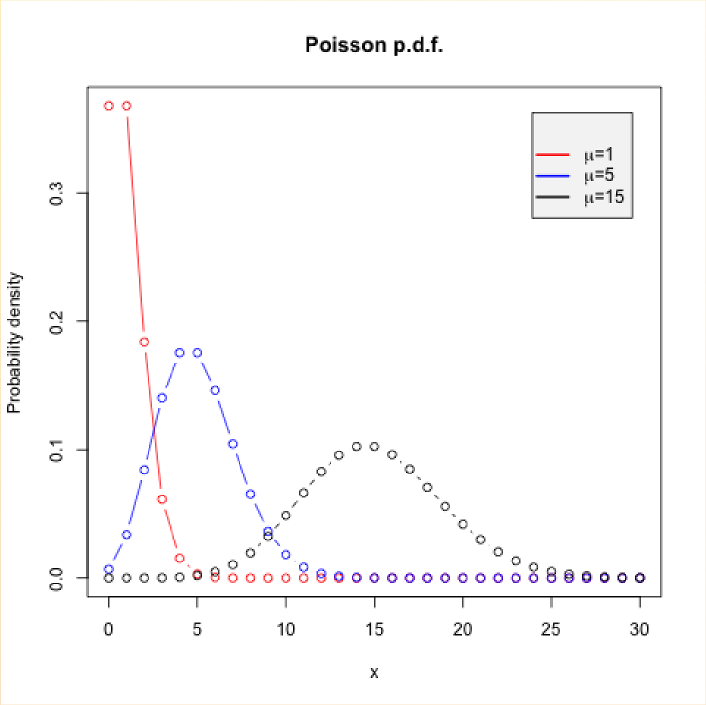
- Discrete variable: r, positive integer $\leq N$
- Parameters:
  - N, positive integer
  - $p, 0 \leq p \leq 1$
- Probability function: $P(r) = \binom{N}{r} p^r (1-p)^{N-r}, r = 0, 1, \ldots, N$
- $E(r) = Np, V(r) = Np(1-p)$
- Usage: probability of finding exactly r successes in N trials



- The distribution of the number of events in a single bin of a histogram is binomial (if the bin contents are independent)

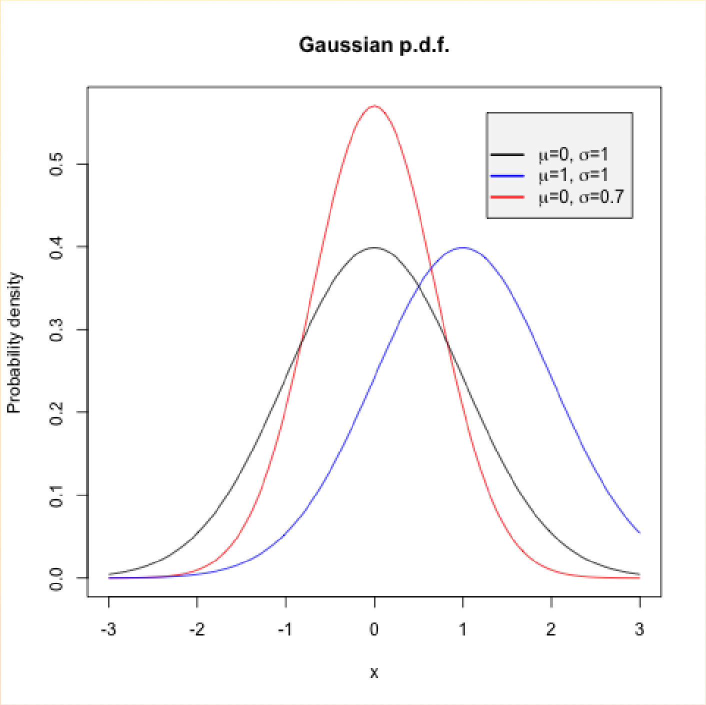
Poisson Distribution

- Discrete variable: r, positive integer
- Parameter: μ, positive real number
- Probability function: $P(r) = \frac{\mu^r e^{-\mu}}{r!}$
- $E(r) = \mu, V(r) = \mu$
- Usage: probability of finding exactly r events in a given amount of time, if events occur at a constant rate.



Gaussian ("Normal") Distribution

- Variable: X, real number
- Parameters:
  - μ, real number
  - σ, positive real number
- Probability function: $f(X) = N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2} \right]$
- $E(X) = \mu$, $V(X) = \sigma^2$
- Usage: describes the distribution of independent random variables. It is also the high-something limit for many other distributions

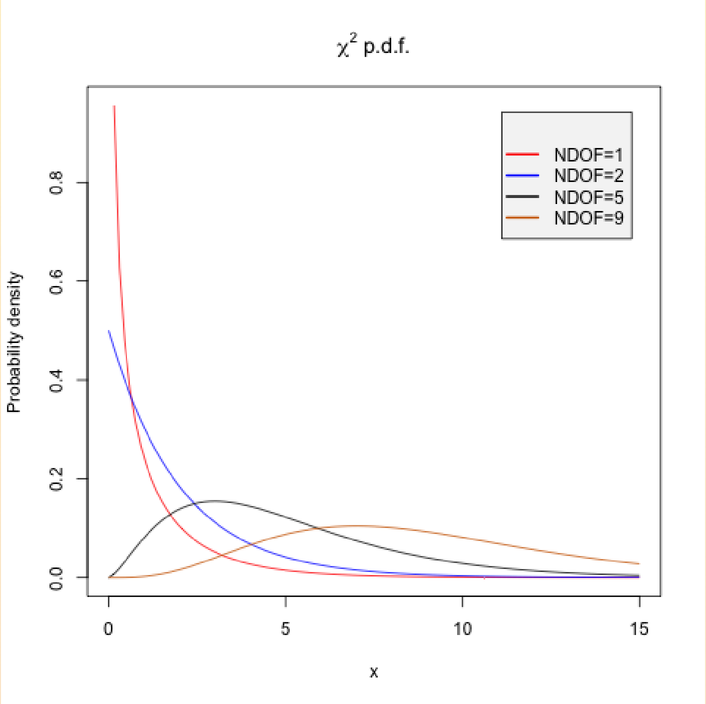


χ^2 distribution

- Parameter: integer $N > 0$ {\em degrees of freedom}
- Continuous variable $X \in \mathcal{R}$
- p.d.f., expected value, variance

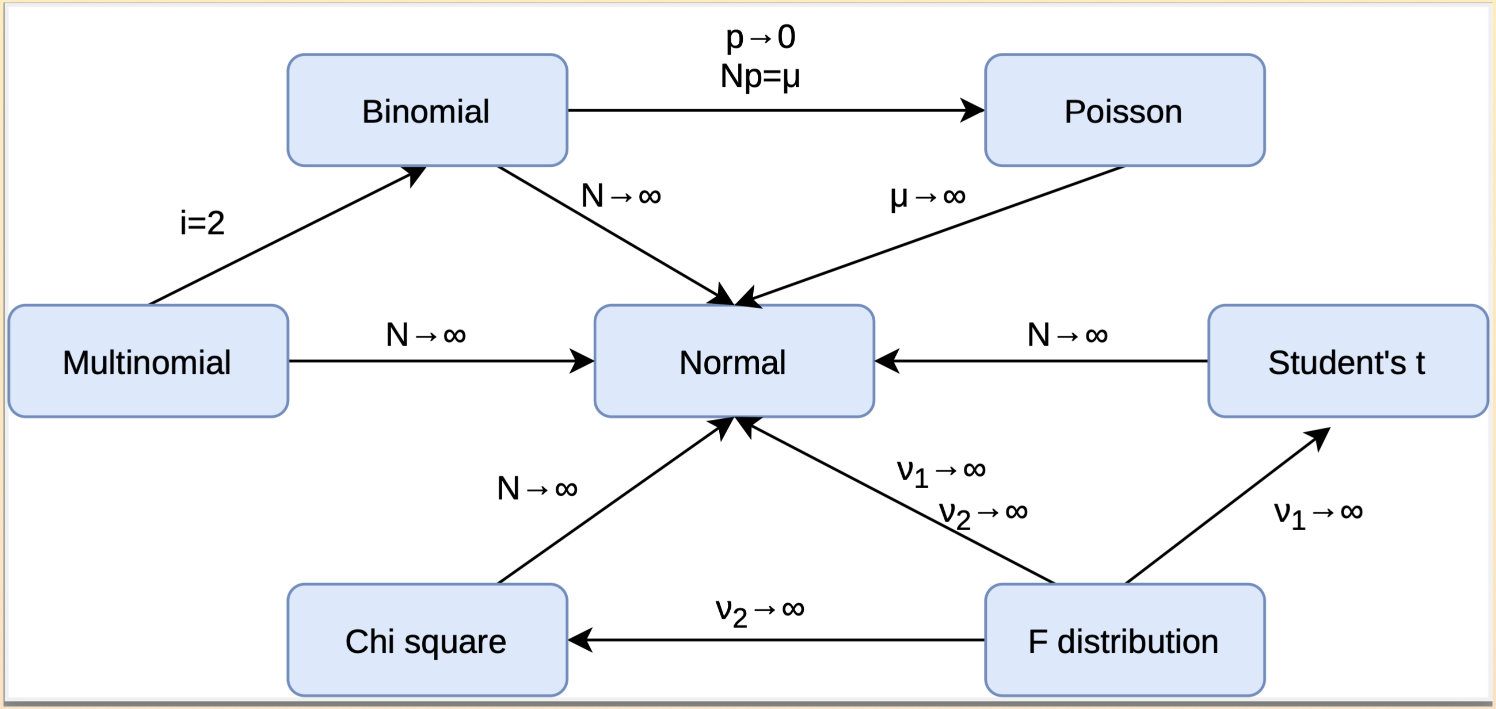
$$f(X) = \frac{\frac{1}{2}\left(\frac{X}{2}\right)^{\frac{N}{2}-1} e^{-\frac{X}{2}}}{\Gamma\left(\frac{N}{2}\right)}$$
$$E[r] = N$$
$$V(r) = 2N$$

- It describes the distribution of the sum of the squares of a random variable, $\sum_{i=1}^{N} X_i^2$



- Reminder: $\Gamma() := \frac{N!}{r!(N-r)!}$

Asymptotically



Estimate location and dispersion

- Expected value: $E[X] := \int_\Omega X f(X) dX$ (or $E[X] := \sum_i X_i P(X_i)$ in the discrete case)
  - Extended to generic functions of a random variable: $E[g] := \int_\Omega g(X) f(X) dX$
- Mean of X is $\mu := E[X]$
- Variance of X is $\sigma_X^2 := V(X) := E[(X-\mu)^2] = E[X^2] - (E[X])^2 = E[X^2] - \mu^2$
- Extension to more variables is trivial, and gives rise to the concept of
- Covariance (or error matrix) of two variables: $V_{XY} = E\big[(X-\mu_X)(Y-\mu_Y)\big] = E[XY] - \mu_X\mu_Y = \int XY f(X,Y) dX dY - \mu_X\mu_Y$
  - Symmetric, and $V_{XX} = \sigma_X^2$
  - Correlation coefficient $\rho_{XY} = \frac{V_{XY}}{\sigma_X \sigma_Y}$

Yes...

- ρ_{XY} is related to the angle in a linear regression of X on Y (or viceversa)

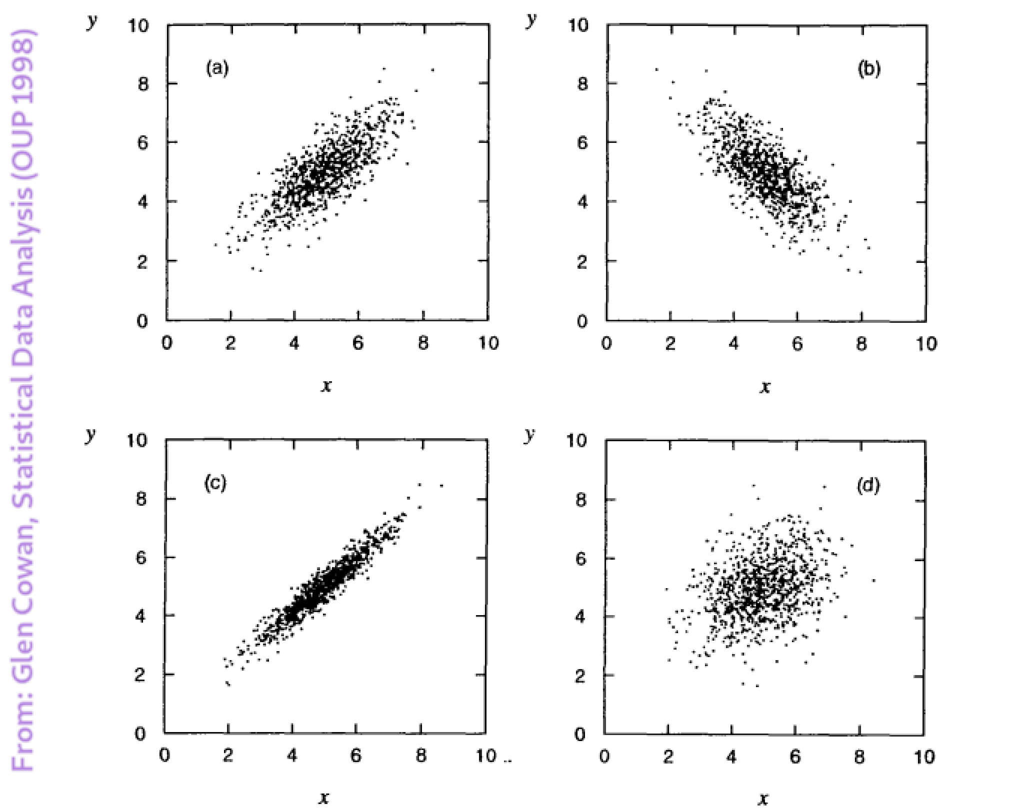
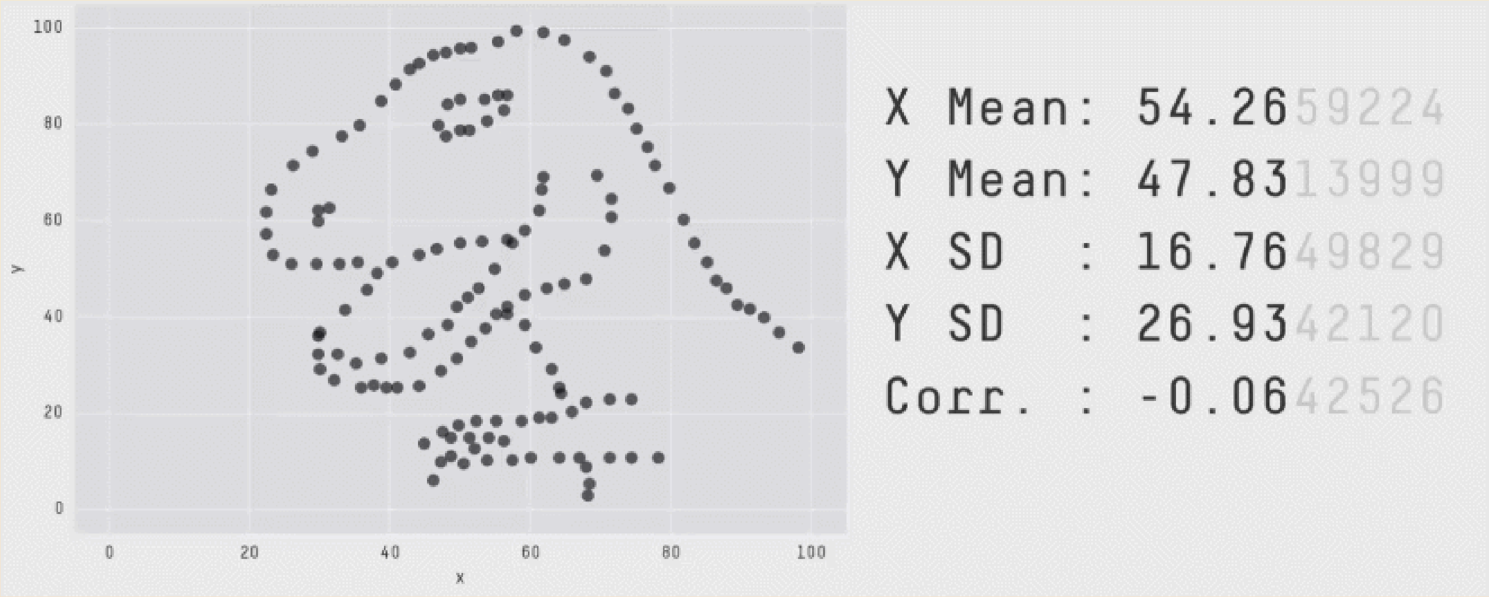


Fig. 1.9 Scatter plots of random variables x and y with (a) a positive correlation, $\rho = 0.75$, (b) a negative correlation, $\rho = -0.75$, (c) $\rho = 0.95$, and (d) $\rho = 0.25$. For all four cases the standard deviations of x and y are $\sigma_x = \sigma_y = 1$.

... but:

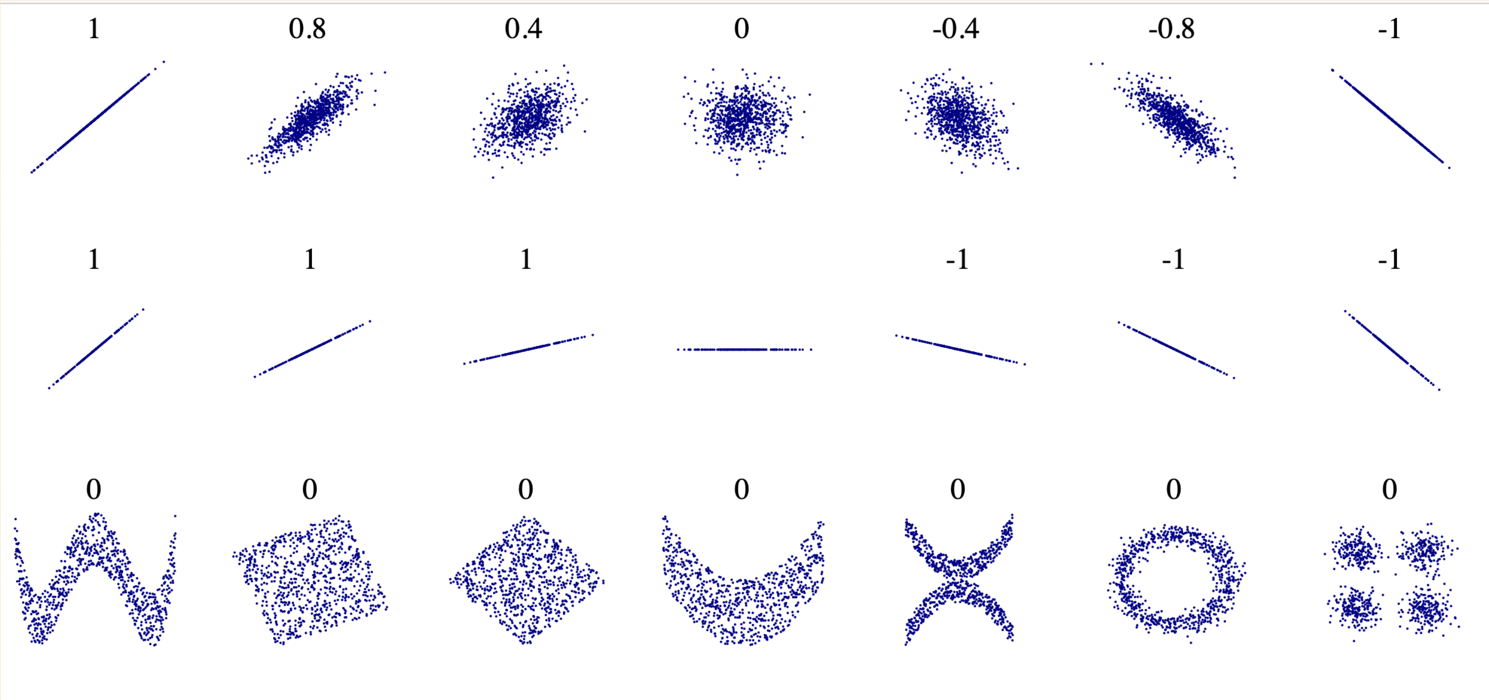
- Several nonlinear correlations may yield the same ρ_{XY} (and other summary statistics)



—
Datasaurus by Alberto Cairo, gif from David Smith

Linear correlation is weak

- X and Y are independent if the occurrence of one does not affect the probability of occurrence of the other
  - X, Y independent $\Longrightarrow \rho_{XY} = 0$
  - $\rho_{XY} = 0 \not\Rightarrow X, Y$ independent



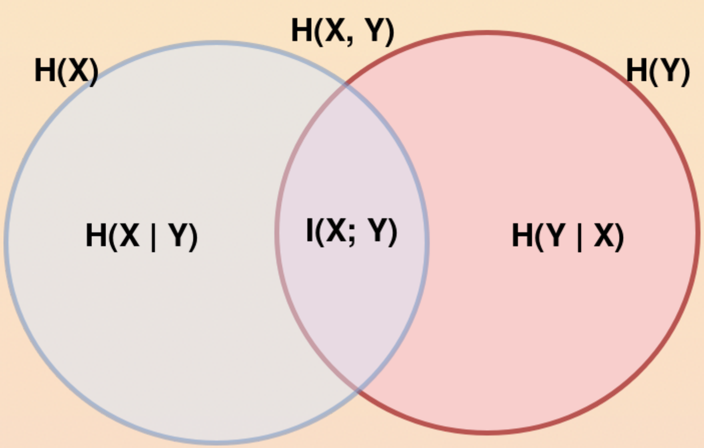
—

Figure by Denis Boigelot

Mutual information

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right)$$

- General notion of correlation linked to the information that X and Y share
  - Symmetric: $I(X;Y) = I(Y;X)$
  - $I(X;Y) = 0$ if and only if X and Y are totally independent

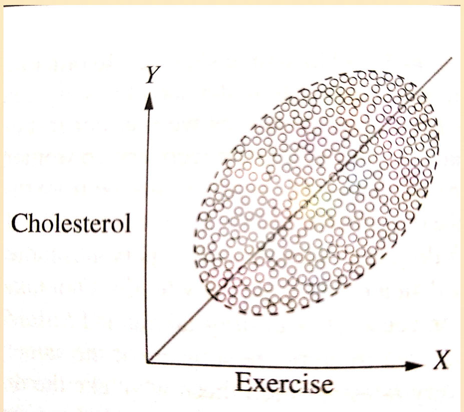


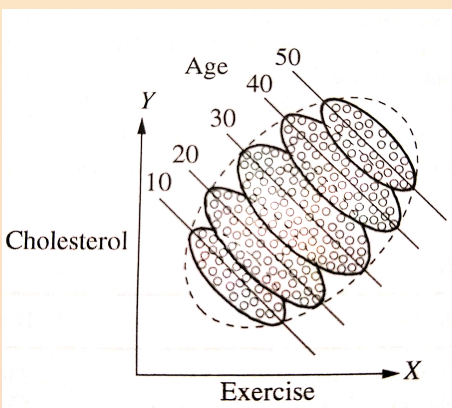
- Related to entropy

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

Causal inference

- Disentangle with interventions on Directed Acyclic Graphs





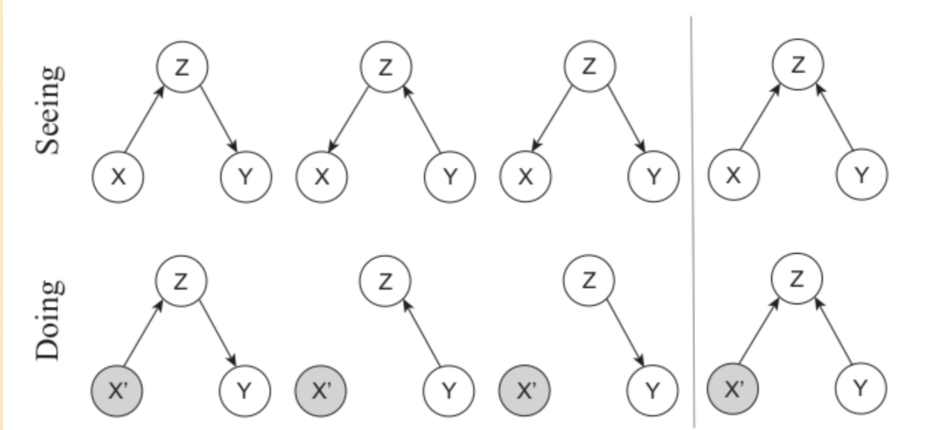
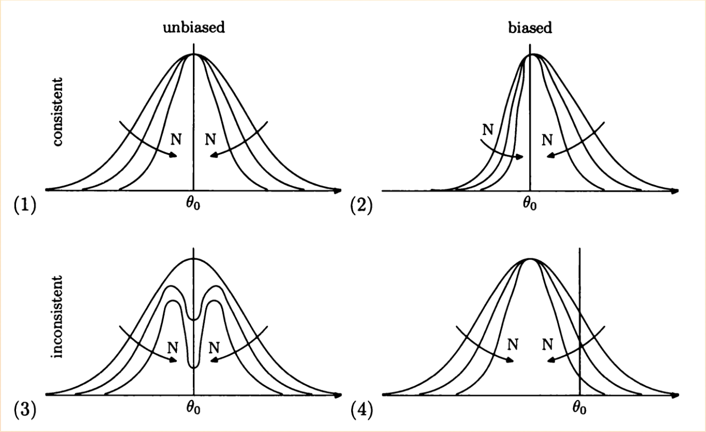


Figure 6. Seeing: DAGs are used to encode conditional independencies. The first three DAGs encode the same associations. *Doing*: DAGs are causal. All of them encode distinct causal assumptions.

Estimators

- $x = (x_1, ..., x_N)$ of N statistically independent observations $x_i \sim f(x)$
  - Determine some parameter θ of $f(x)$
  - x, θ in general are vectors
- Estimator is a function of the observed data that returns numerical values $\hat{\theta}$ for the vector θ.

- (Asymptotic) Consistency: $\lim_{N\to\infty} \hat{\theta} = \theta_{true}$
- Unbiasedness: the bias is zero
  - Bias: $b := E[\hat{\theta}] - \theta_{true}$
  - If bias known: $\hat{\theta}' = \hat{\theta} - b$, so $b' = 0$
- Efficiency: smallest possible $V[\hat{\theta}]$

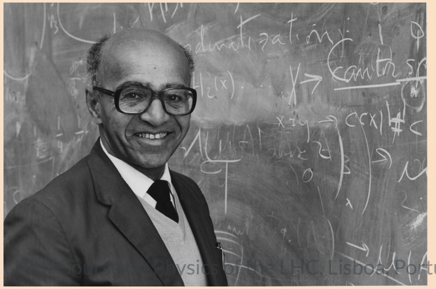


- Robustness: insensitivity from small deviations from the underlying p.d.f.

—
Image from James, 2nd ed.

Sufficient statistic

- Test statistic: a function of the data (a quantity derived from the data sample)
- $X \sim f(X|\theta)$, then $T(X)$ is sufficient for θ if $f(X|T)$ is independent of θ
- T carries as much information about θ as the original data X
  - Data X with model M and statistic $T(X)$ with model M' provide the same inference
- Rao-Blackwell theorem: if $g(X)$ is an estimator for θ and T is sufficient, then $E[g(X)|T(X)]$ is never a worse estimator of θ
  - Build a ballpark estimator $g(X)$, then condition on some $T(X)$ to obtain a better estimator
- Sufficiency Principle: if $T(X) = T(Y)$, then X and Y provide same inference about θ
  - Implications for data storage, computation requirements, etc.



—

Images from AmStat magazine and illinois.edu

The Maximum Likelihood Method

- $x = (x_1, ..., x_N)$ of N statistically independent observations $x_i \sim f(x)$

$$L(x;\theta) = \prod_{i=1}^{N} f(x_i, \theta)$$

- Maximum-likelihood estimator is θ_{ML} such that

$$\theta_{ML} := argmax\theta\Big(L(x,\theta)\Big)$$

- Numerically, best to minimize: $-lnL(x;\theta) = -\sum_{i=1}^{N} lnf(xi, \theta)$
  - Fred James' Minuit's MINOS routine powers e.g. RooFit
- The MLE is:
  - Consistent: $\lim_{N\to\infty} \theta_{ML} = \theta_{true}$;
  - Unbiased: only asymptotically. $\vec{b} \propto \frac{1}{N}$, so $\vec{b} = 0$ only for $N \to \infty$;
  - Efficient: $V[\theta_{ML}] = \frac{1}{I(\theta)}$
  - Invariant under $\psi = g(\theta)$: $\hat{\psi}_{ML} = g(\theta_{ML})$

MLE for Nuclear Decay

- Nuclear decay with half-life τ
  $f(t;\tau) = \frac{1}{\tau}e^{-\frac{t}{\tau}}$
  $E[f] = \tau$
  $V[f] = \tau^2$

- Sample $t_i \sim f(t;\tau)$, obtaining $f(t_1, ...t_N; \tau) = \prod_i f(t_i;\tau) = L(\tau)$

$$\frac{\partial lnL(\tau)}{\partial\tau} = \sum_i \Big(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \Big) \equiv 0 \qquad \Longrightarrow \qquad \hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- Unbiased: $b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$

- Variance depends on samples: $V[\hat{\tau}] = V\Big[\frac{1}{N}\sum_i t_i\Big] = \frac{1}{N^2}\sum_i V[t_i] = \frac{\tau^2}{N}$

Estimator	Consistent	Unbiased	Efficient
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1+...+t_N}{N}$	Yes	Yes	Yes
$\hat{\tau} = \frac{t_1+...+t_N}{N-1}$	Yes	No	No
$\hat{\tau} = t_i$	No	Yes	No

Bias-variance tradeoff

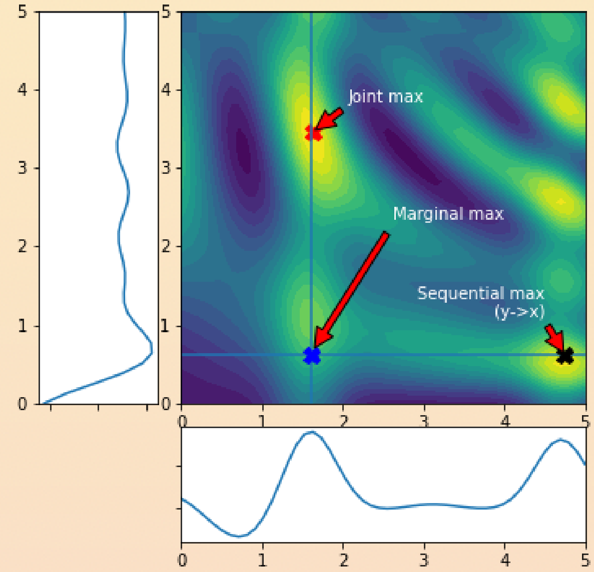
- Cannot have both zero bias and the smallest variance
- Information acts on the curvature of the likelihood, which represents the precision
  - Information is a limiting factor for the variance
- Rao-Cramer-Frechet (RCF) bound

$$V[\hat{\theta}] \geq \frac{(1+\partial b/\partial\theta)^2}{-E\left[\partial^2 lnL/\partial\theta^2\right]}$$

- Fisher Information Matrix

$$I_{ij} = E\left[\partial^2 lnL/\partial\theta_i\partial\theta_j\right]$$

$$argmin_{x,y}\Big(f(x,y)\Big)_y \neq argmin_y\Big(f(x,y)\Big)$$



Approximate variance

$$V[\hat{\theta}] \geq \frac{\left(1+\frac{\partial b}{\partial \theta}\right)^2}{-E\left[\frac{\partial^2 lnL}{\partial \theta^2}\right]}$$

- MLE is efficient and asymptotically unbiased

$$V[\theta_{ML}] \simeq \frac{1}{-E\left[\frac{\partial^2 lnL}{\partial \theta^2}\right]}\Bigg|\theta = \theta ML$$

- For a Gaussian pdf $f(x;\theta) = N(\mu, \sigma)$

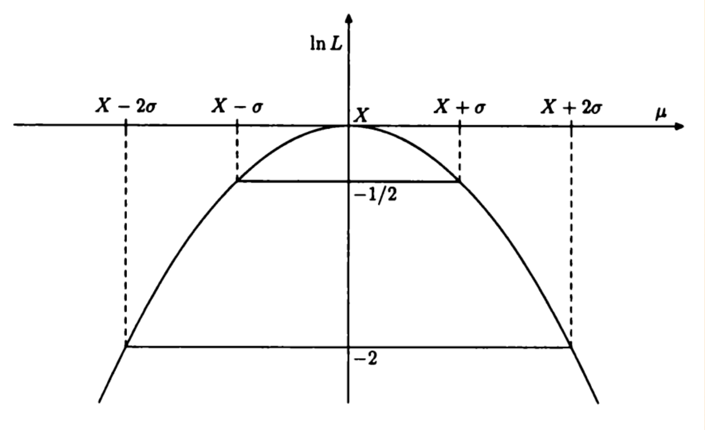
$$L(\theta) = ln\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]$$

- $L(\theta_{1\sigma}) - \hat{\theta}_{ML} = 1/2$, and the area enclosed in $[\theta_{ML} - \sigma, \theta_{ML} + \sigma]$ will be 68.3%.

Confidence interval

- An interval with a fixed probability content
  $$P\Big((\theta_{ML} - \theta_{true})^2 \leq \sigma)\Big) = 68.3\%$$
  $$P(-\sigma \leq \theta_{ML} - \theta_{true} \leq \sigma) = 68.3\%$$
  $$P(\theta_{ML} - \sigma \leq \theta_{true} \leq \theta_{ML} + \sigma) = 68.3\%$$

- Practical prescription
  - Point estimate by computing the MLE
  - Confidence interval by taking the range delimited by the crossings of the likelihood function with $\frac{1}{2}$ (for 68.3% probability content, or 2 for 95% probability content), etc)



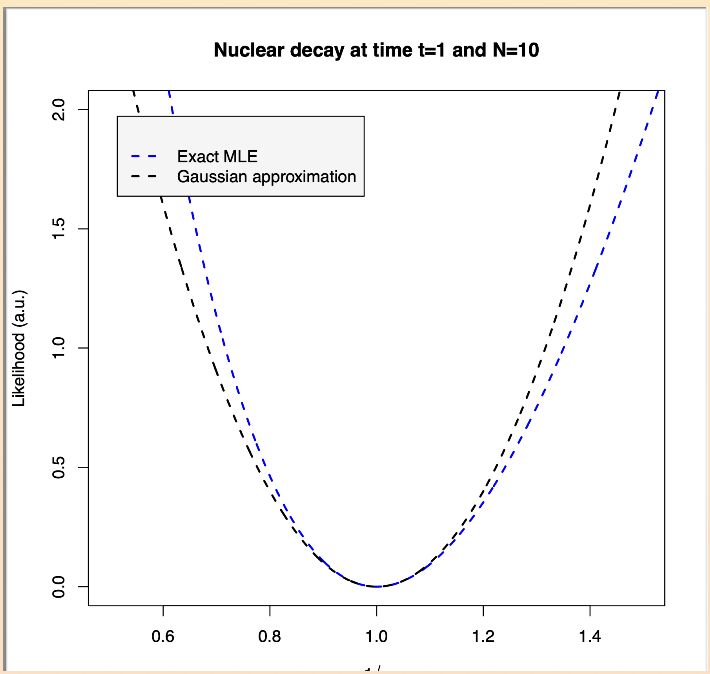
- MLE is invariant for monotonic transformations of θ
  - Likelihood crossings can be used also for asymmetric likelihood functions
  - Intervals exact only to $\mathcal{O}(\frac{1}{N})$

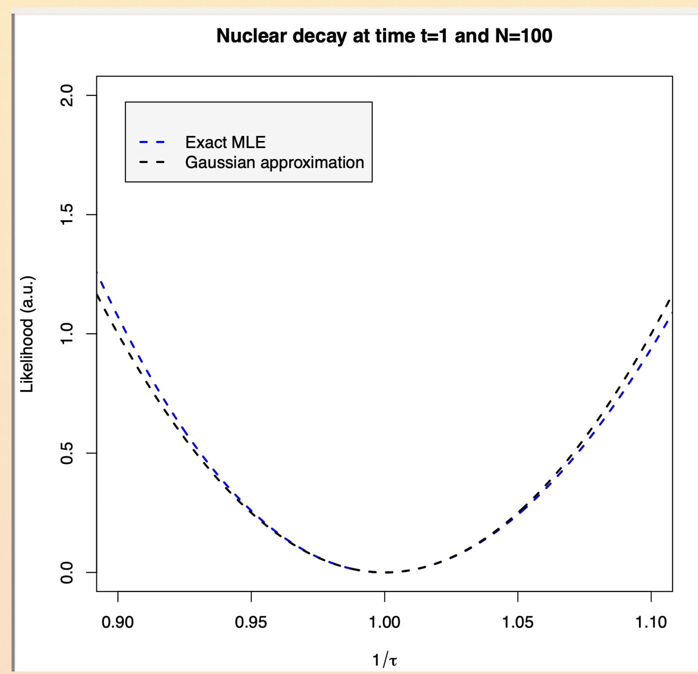
–
Image from James, 2nd ed.

Normal approximation

- Good only to $\mathcal{O}(\frac{1}{N})$:

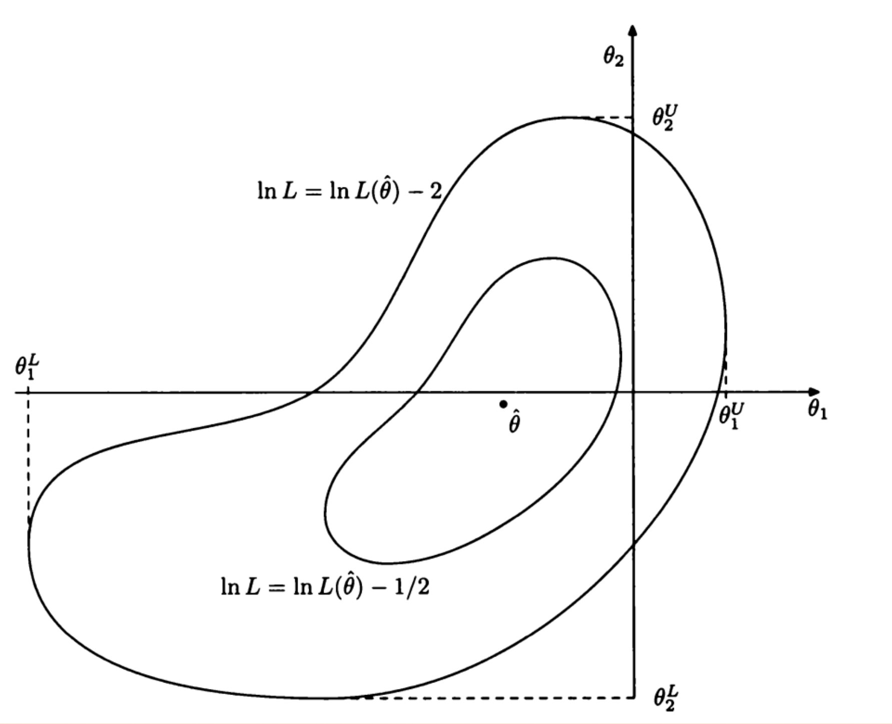
$$L(x;\theta) \propto exp\Big[-\frac{1}{2}(\theta-\theta_{ML})^T H(\theta-\theta_{ML})\Big]$$





Likelihood in many dimensions

- Elliptical contours correspond to gaussian Likelihoods
  - The closer to MLE, the more elliptical the contours, even in nonlinear problems
  - Minimizers just follow the contour regardless of nonlinearity
- Crossings (contours) adapted to areas under N-dimensional gaussians



—
Plots from James, 2ed.

Profiling for systematic uncertainties

- Once upon a time, cross sections were: $\sigma = \frac{N_{data} - N_{bkg}}{\epsilon L}$
  - N_{sig} estimated from $N_{data} - N_{bkg}$ for the measured integrated luminosity L
  - Uncertainties in the acceptance ϵ propagated to the result for σ
- Nowadays, $p(x|\mu, \theta)$ pdf for the observable x to assume a certain value in a single event
  - $\mu := \frac{\sigma}{\sigma_{pred}}$ parameter of interest
  - θ nuisance parameters representing all the uncertainties affecting the measurement
  - Many events: $\prod_{e=1}^{n} p(x_e|\mu, \theta)$
- The number of events in the data set is however a Poisson random variable itself!
  - Marked Poisson Model $f(X|\nu(\mu, \theta), \mu, \theta) = Pois(n|\nu(\mu, \theta)) \prod_{e=1}^{n} p(x_e|\mu, \theta)$

—
Pleasant read Vischia, 2019

Uncertainties as nuisance parameters

- Incorporate systematic uncertainties as nuisance parameter θ (Conway, 2011)
  - constraint interpreted as (typically Gaussian) prior coming from the auxiliary measurement
- MLE still depends on nuisance parameters: $\hat{\mu} := argmax_{\mu}\mathcal{L}(\mu, \theta; X)$

$$\mathcal{L}(\boldsymbol{n}, \boldsymbol{\alpha}^{\mathbf{0}}|\mu, \boldsymbol{\alpha}) = \prod_{i \in bins} \mathcal{P}(n_i|\mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j \in syst} \mathcal{G}(\alpha_j^0|\alpha_j, \delta\alpha_j)$$

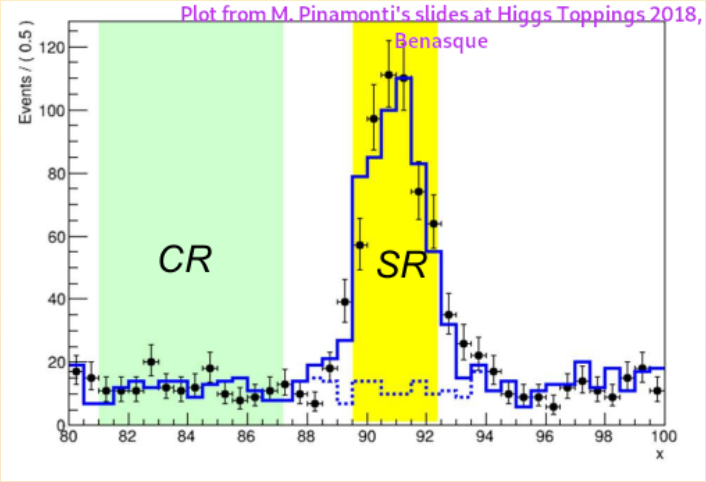
$$\downarrow$$

$$\mathcal{L}(\boldsymbol{n}, 0|\mu, \boldsymbol{\alpha}) = \prod_{i \in bins} \mathcal{P}(n_i|\mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j \in syst} \mathcal{G}(0|\alpha_j, 1)$$

Sidebands

- Sideband measurement

$$L_{SR}(s,b) = Poisson(N_{SR} \mid s+b)$$
$$L_{CR}(b) = Poisson(N_{CR} \mid \tilde{\tau} \cdot b)$$



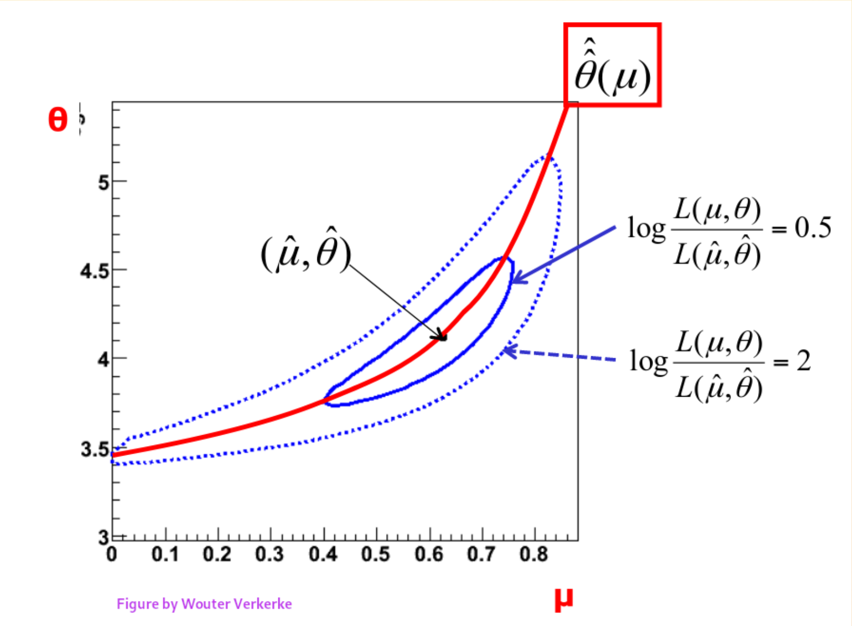
$$\mathcal{L}_{full}(s,b) = \mathcal{P}(N_{SR}|s+b) \times \mathcal{P}(N_{CR}|\tilde{\tau} \cdot b)$$

- Example subsidiary measurement of the background rate:
  - 8% systematic uncertainty in the MC rates
  - $\tilde{b}$: measured background rate
  - $\mathcal{G}(\tilde{b}|b, 0.08)\, \mathcal{L}_{full}(s,b) = \mathcal{P}(N_{SR}|s+b) \times \mathcal{G}(\tilde{b}|b, 0.08)$

The Likelihood Ratio:

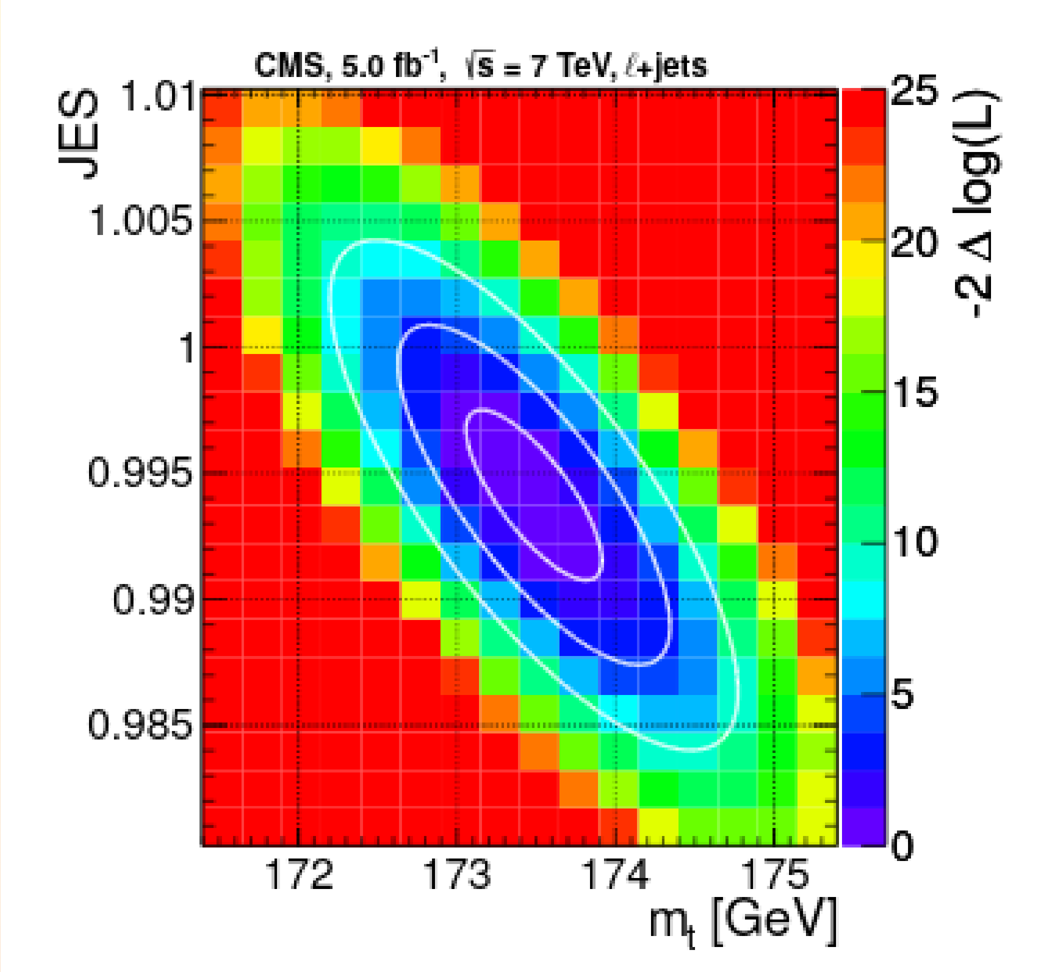
$$\lambda(\mu) := \frac{\mathcal{L}(\mu,\hat{\hat{\theta}})}{\mathcal{L}(\hat{\mu},\hat{\theta})}$$

- Profiling: eliminate dependence on θ by taking conditional MLEs
  - Bayesian marginalize Demortier, 2002



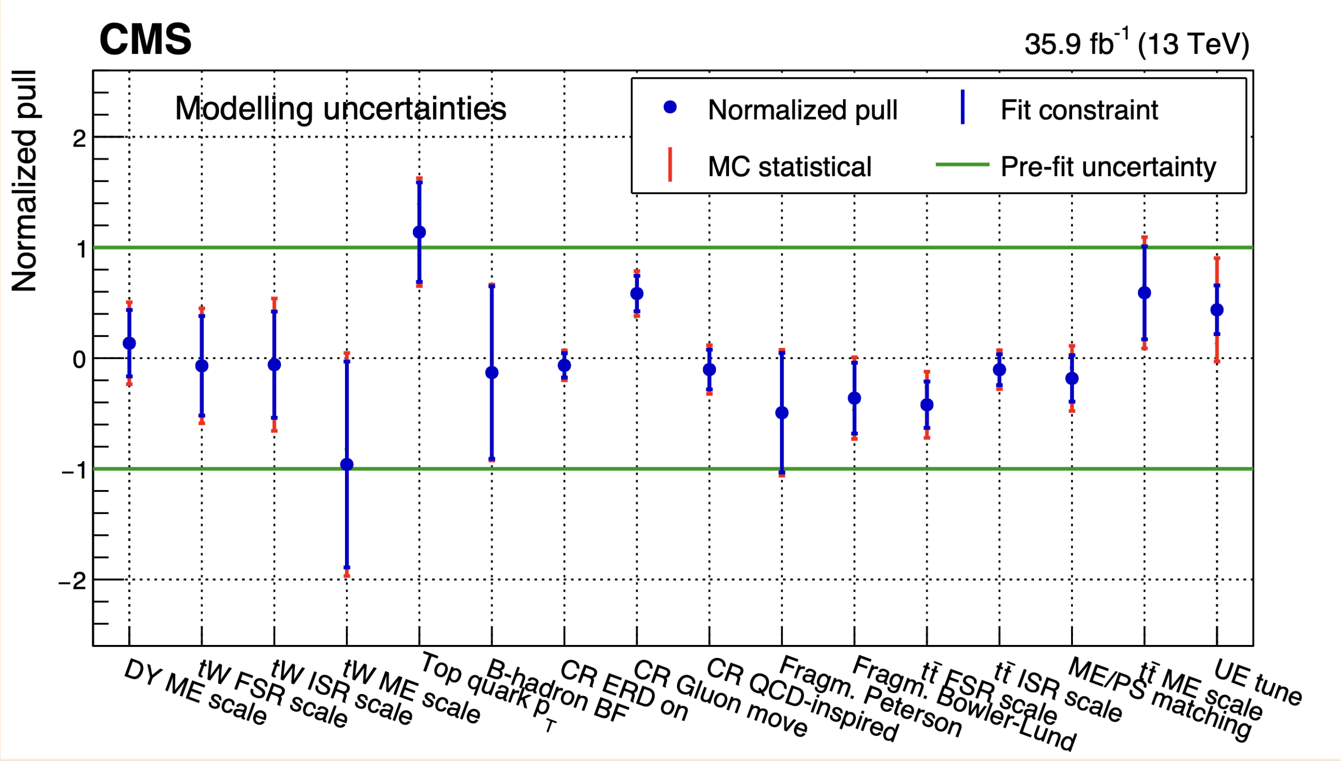
- $\lambda(\mu)$ distribution by toy data, or use Wilks theorem: $\lambda(\mu) \sim exp\Big[-\frac{1}{2}\chi^2\Big]\Big(1+\mathcal{O}(\frac{1}{\sqrt{N}})\Big)$ under some regularity conditions

What is a nuisance parameter?



Pulls and Constraints

- Pull: difference of the post-fit and pre-fit values of the parameter, normalized to the pre-fit uncertainty: $pull := \frac{\hat{\theta}-\theta}{\delta\theta}$
- Constraint: the ratio between the post-fit and the pre-fit uncertainty in the nuisance parameter.



Correlation and Significance

- What worries you the most?
  - A pull with very small constraint: $\theta_{prefit} = 0 \pm 1, \theta_{postfit} = 1 \pm 0.9$
  - The same pull with a strong constraint: $\theta_{prefit} = 0 \pm 1, \theta_{postfit} = 1 \pm 0.2$

Correlation and Significance

- What worries you the most?
  - A pull with very small constraint: $\theta_{prefit} = 0 \pm 1, \theta_{postfit} = 1 \pm 0.9$
  - The same pull with a strong constraint: $\theta_{prefit} = 0 \pm 1, \theta_{postfit} = 1 \pm 0.2$
- Compare the shift to its uncertainty
- Indipendent measurements: the compatibility C is

$$C = \Delta\theta / \sigma_{\Delta\theta} = \frac{\theta_2 - \theta_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

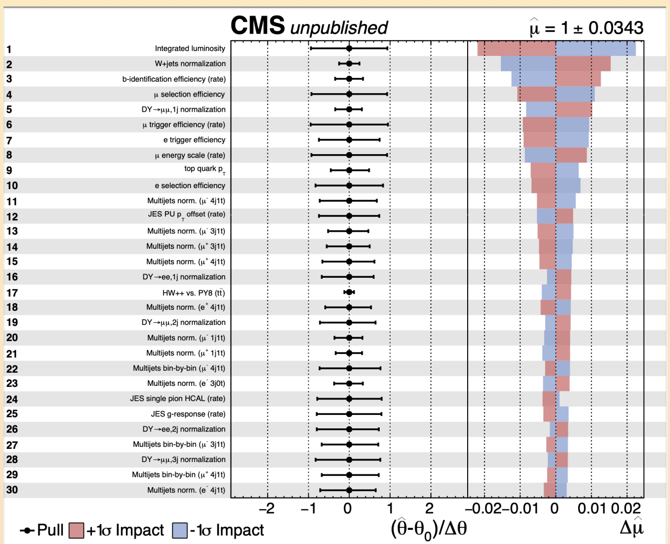
- First case $C = 0.74$, second case $C = 0.98$ (larger, still within uncertainty)
- These are not independent measurements! Worst-case scenario formula:

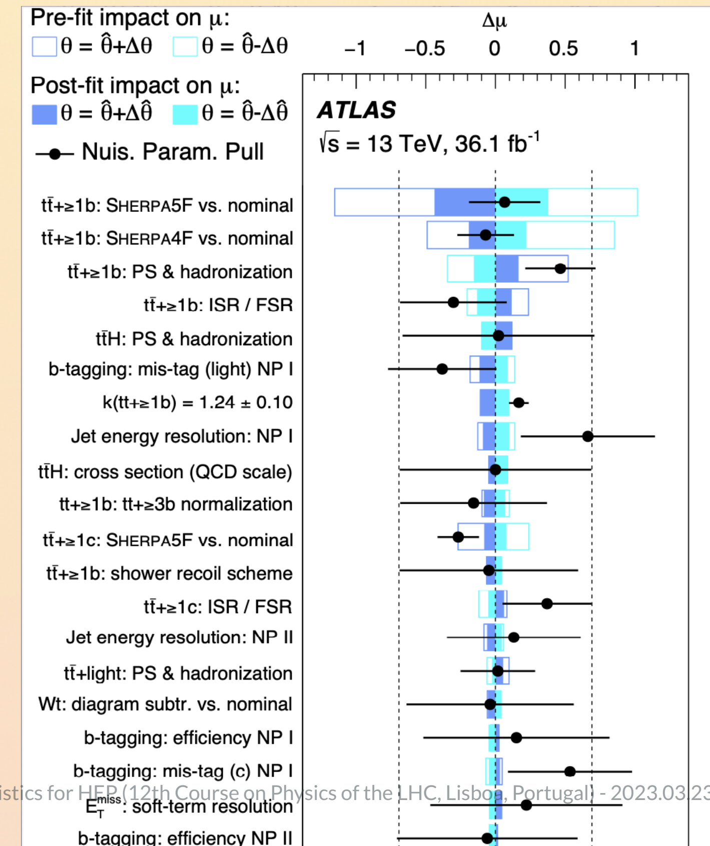
$$C = \Delta\theta / \sigma_{\Delta\theta} = \frac{\theta_2 - \theta_1}{\sqrt{\sigma_1^2 - \sigma_2^2}}$$

- First case, $C = 2.29$, second case $C = 1.02$
- The same pull is more significant if there is (almost no) constraint!!!

Impacts on the post-fit μ

- Fix each θ to its post-fit value $\hat{\theta}$ plus/minus its pre(post)fit uncertainty $\delta\theta$ ($\delta\hat{\theta}$)
- Reperform the fit for μ
- Impact is $\hat{\mu} - \hat{\mu}(\hat{\theta})$ (should give perfect result on Asimov dataset)

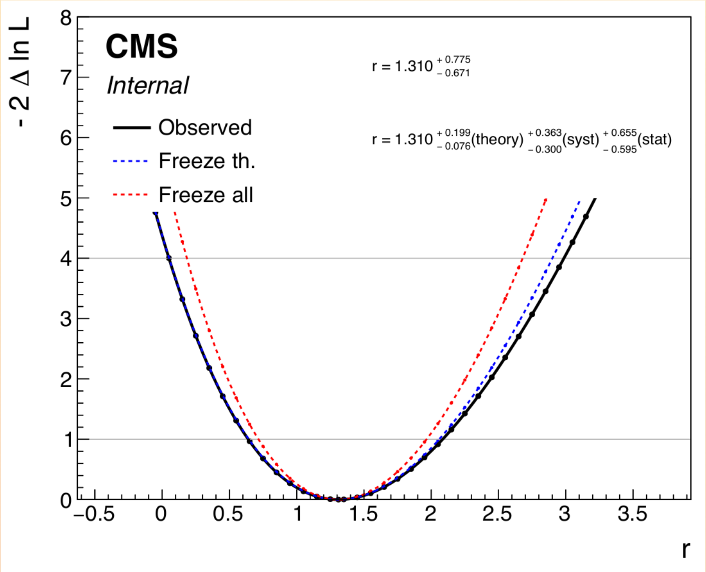




Breakdown of uncertainties

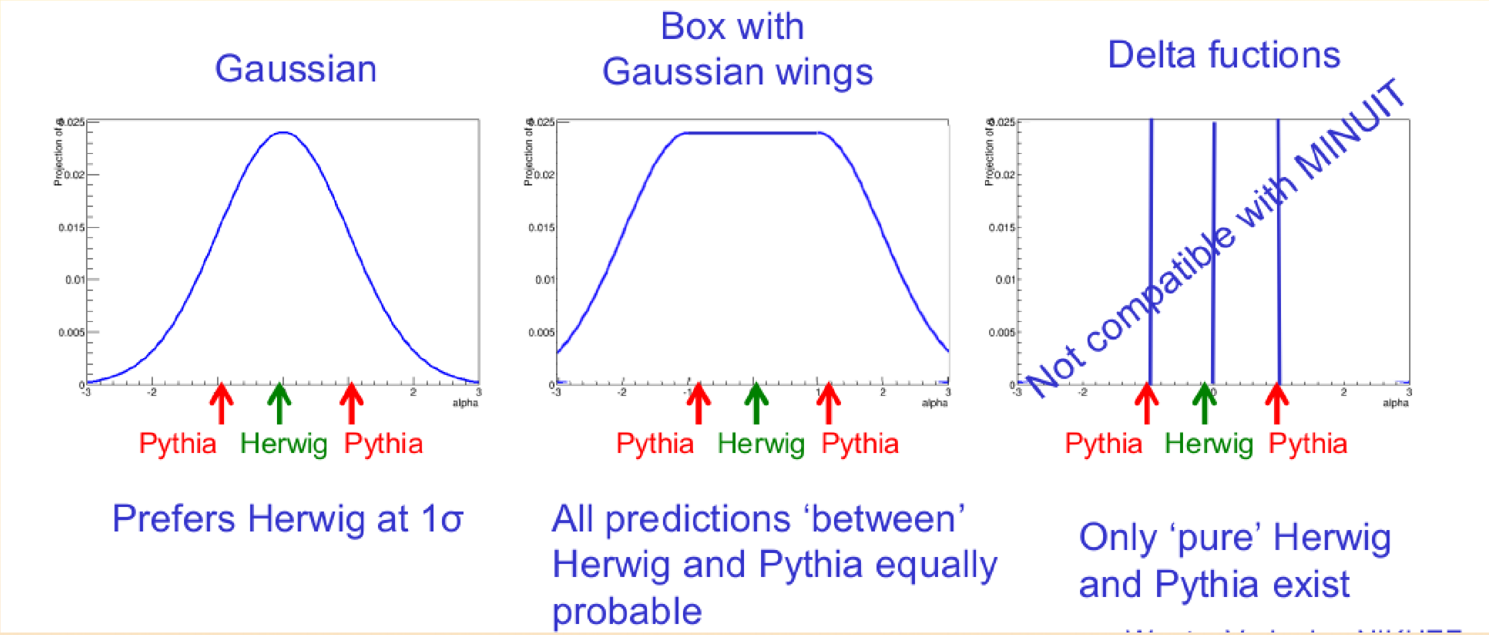
- Amount of uncertainty on μ imputable to a given source of uncertainty
  - Modern version of Fisher's formalization of the ANOVA concept
  - the constituent causes fractions or percentages of the total variance which they together produce (Fisher, 1919)
  - the variance contributed by each term, and by which the residual variance is reduced when that term is removed (Fisher, 1921)

- Freeze a set of θ_i to $\hat{\theta}_i$
- Repeat the fit, uncertainty on μ is smaller
- Contribution of θ_i to the overall uncertainty as squared difference
- Statistical uncertainty by freezing all nuisance parameters



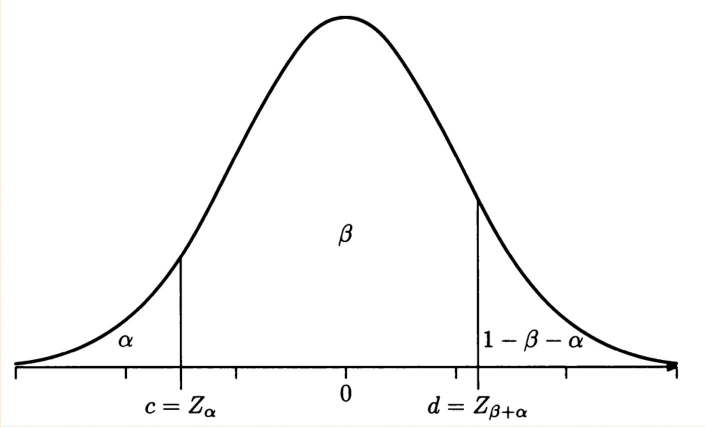
— Toy data

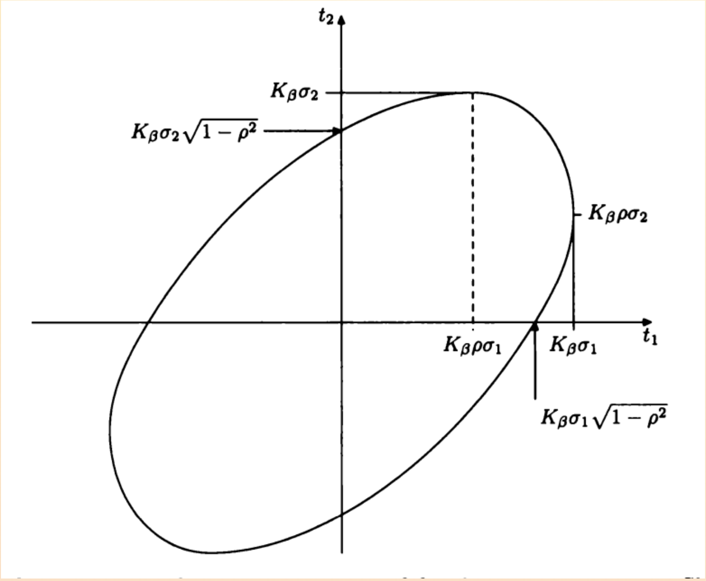
Which is the "correct" constraint?



Confidence intervals

- Probability content: solve $\beta = P(a \leq X \leq b) = \int_a^b f(X|\theta)dX$ for a and b
  - A method yielding interval with the desired β, has coverage





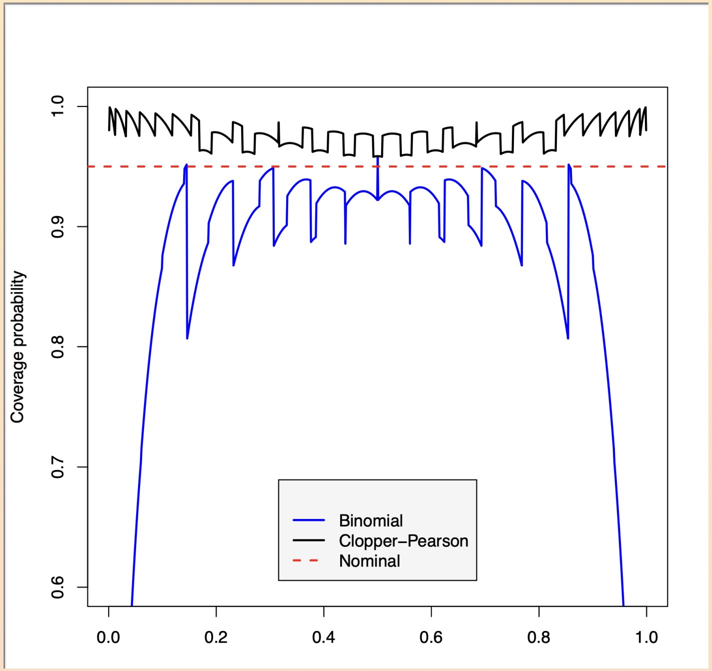
—
Plots from James, 2nd ed.

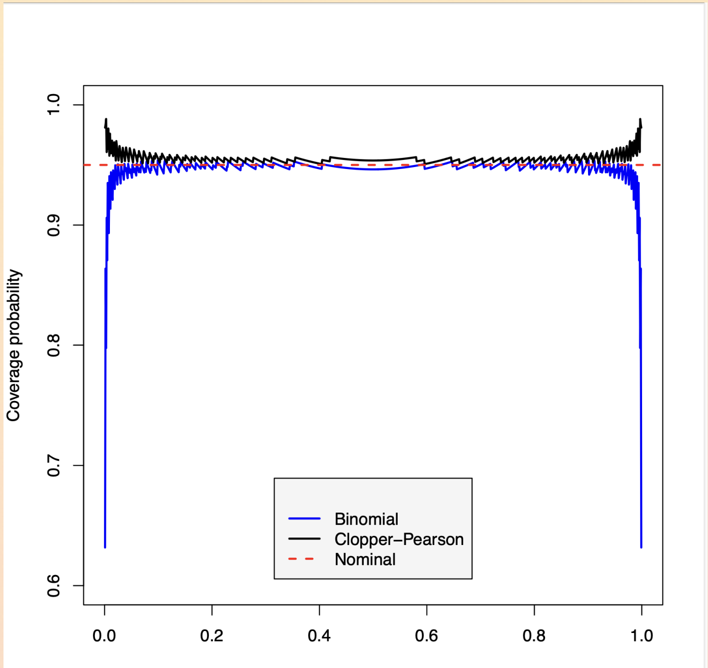
Checking for coverage

- Operative definition of coverage probability
  - Fraction of times, over a set of (usually hypothetical) measurements, that the resulting interval covers the true value of the parameter
  - Obtain the sampling distribution of the confidence intervals using toy data
- Nominal coverage: the one you have built your method around
- Actual coverage: the one you calculate from the sampling distribution
  - Toy experiment: sample N times for a known value of θ_{true}
  - Compute interval for each experiment
  - Count fractions of intervals containing θ_{true}
- Nominal and actual coverage should agre if all assumptions of method are valid
  - Undercoverage: intervals smaller than proper ones
  - Overcoverage: intervals larger than proper ones

Discrete Case

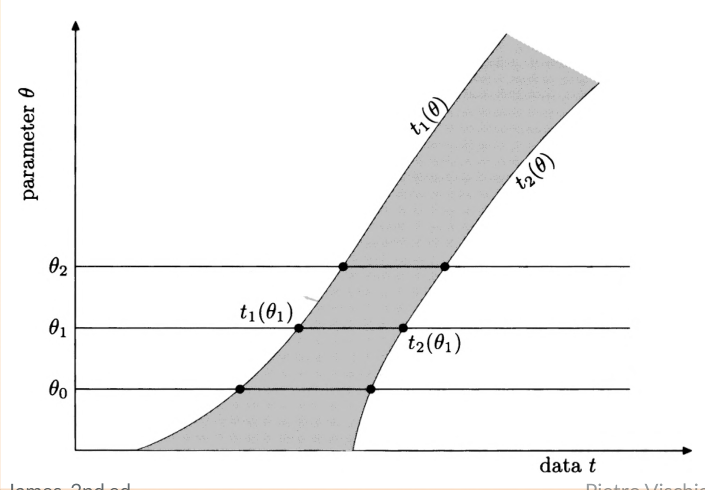
- Probability content $P(a \leq X \leq b) = \sum_a^b f(X|\theta)dX \leq \beta$
- Binomial: find (r_{low}, r_{high}) such that $\sum_{r=r_{low}}^{r=r_{high}} \binom{r}{N} p^r (1-p)^{N-r} \leq 1-\alpha$
  - Gaussian approximation: $p \pm Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{N}}$
  - Clopper Pearson: invert two single-tailed binomial tests

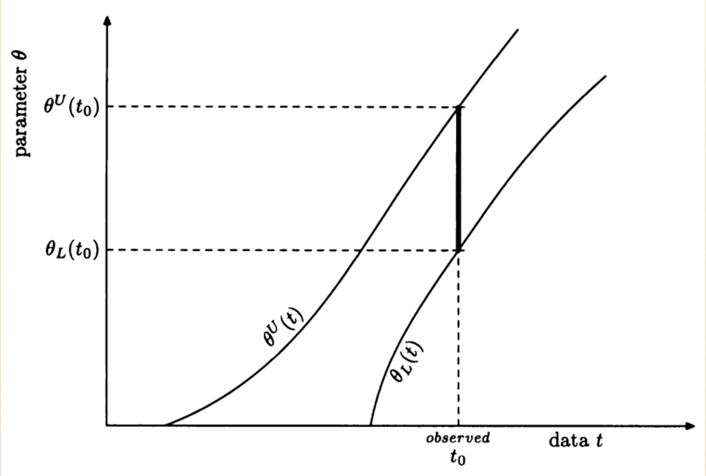




The Neyman construction

- Unique solutions to finding confidence intervals are infinite
  - 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
- Intrinsically frequentist procedure



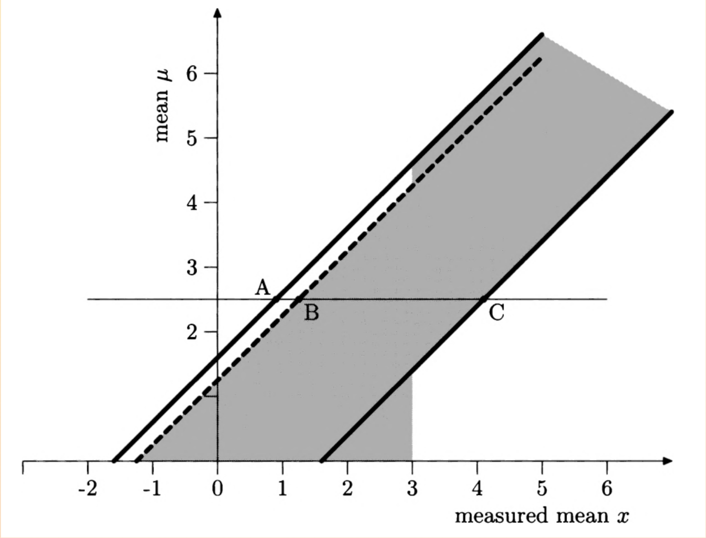


—
Figures from James, 2nd ed.

Flip-flopping

- Gaussian measurement (variance 1) of $\mu > 0$ (physical bound)
- Individual prescriptions are self-consistent
  - 90% central limit (solid lines)
  - 90% upper limit (single dashed line)

- Mixed choices (after looking at data) are problematic
- Unphysical values and empty intervals: choose 90% central interval, measure $x_{obs} = -2.0$
  - Interval empty, yet with the desired coverage

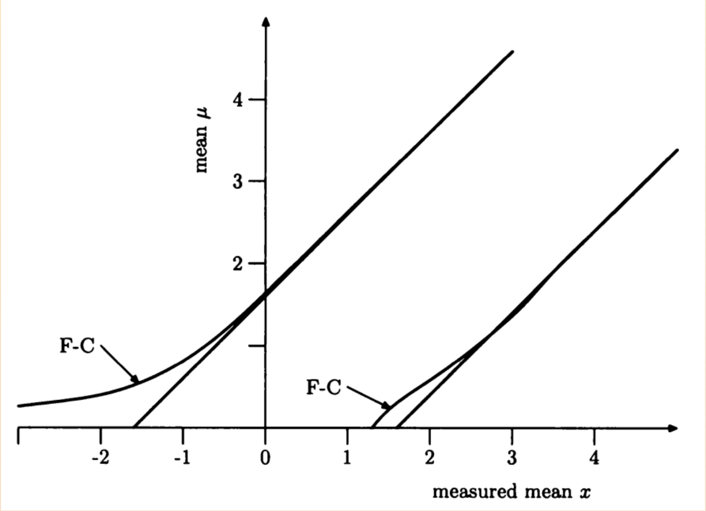


—
Figures from James, 2nd ed.

The Feldman-Cousins Ordering Principle

- Unified approach for determining interval for $\mu = \mu_0$
  - 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

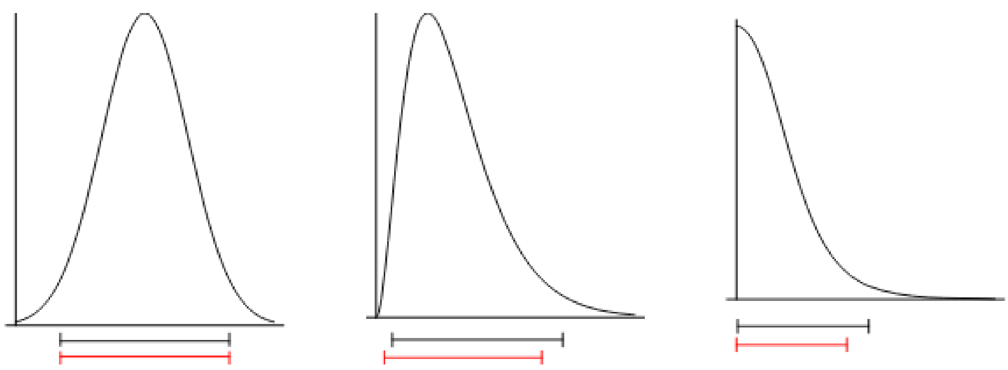
- Minimizes Type II error (likelihood ratio for simple test is the most powerful test)
- Solves the problem of empty intervals
- Avoids flip-flopping in choosing an ordering prescription



Bayesian intervals

- Often numerically identical to frequentist confidence intervals
  - Much simple derivation
  - Interpretation is different: {\em credible intervals}
  - Posterior density summarizes the complete knowledge about θ
- Highest Probability Density intervals
  - Work out of the box for multimodal distributions and for physical constraints

Fig. 1 Simple examples of central (*black*) and highest probability density (*red*) intervals. The intervals coincide for a symmetric distribution, otherwise the HPD interval is shorter. The three examples are a normal distribution, a gamma with shape parameter 3, and the marginal posterior density for a variance parameter in a hierarchical model. (Color figure online)

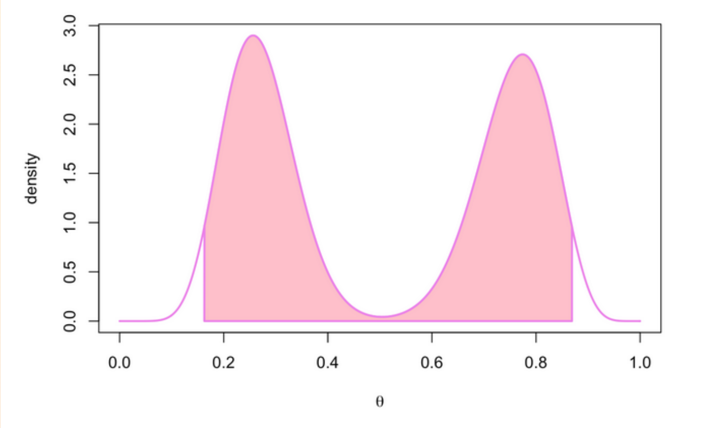


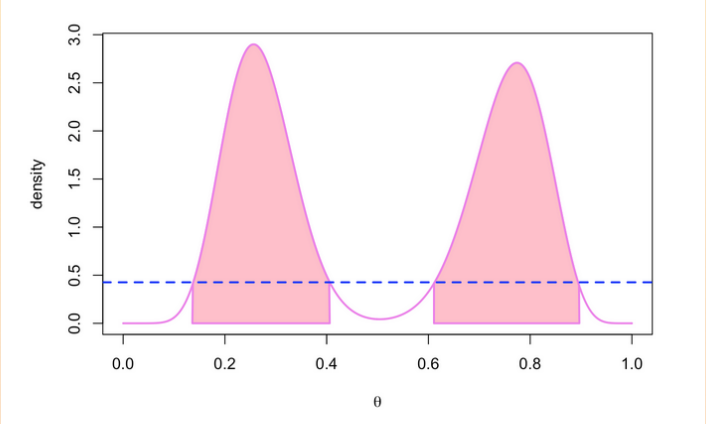
—

Figure from Stat Comput (2015) 25:809--819

Bayesian intervals

- Often numerically identical to frequentist confidence intervals
  - Much simple derivation
  - Interpretation is different: {\em credible intervals}
  - Posterior density summarizes the complete knowledge about θ
- Highest Probability Density intervals
  - Work out of the box for multimodal distributions and for physical constraints

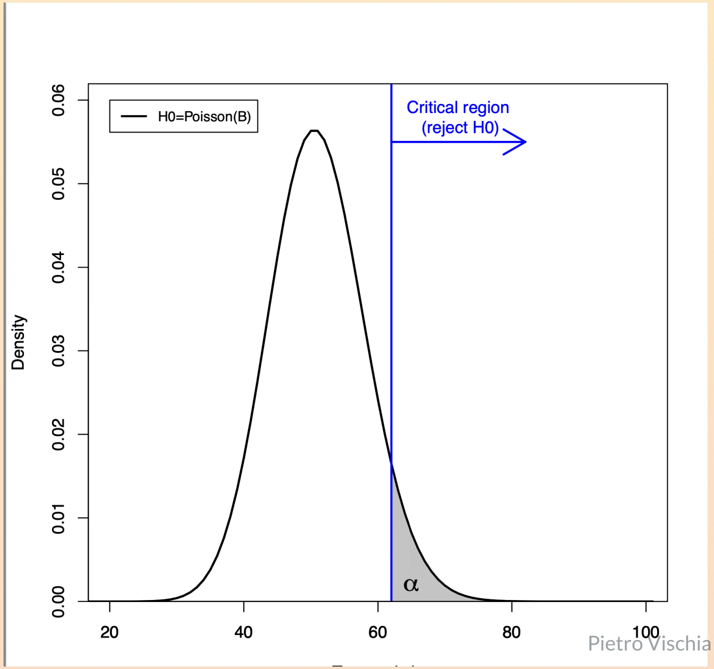


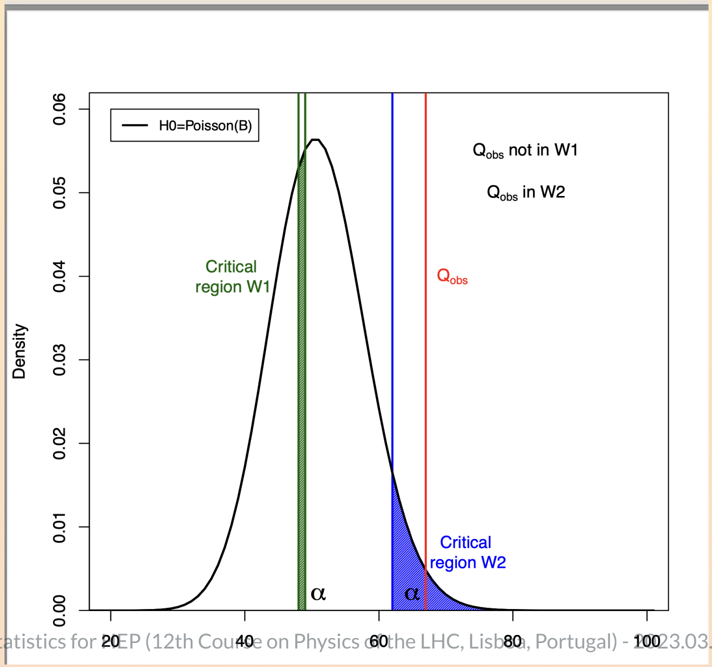


—
Figure from vioshyvo.github.io

Test of hypotheses

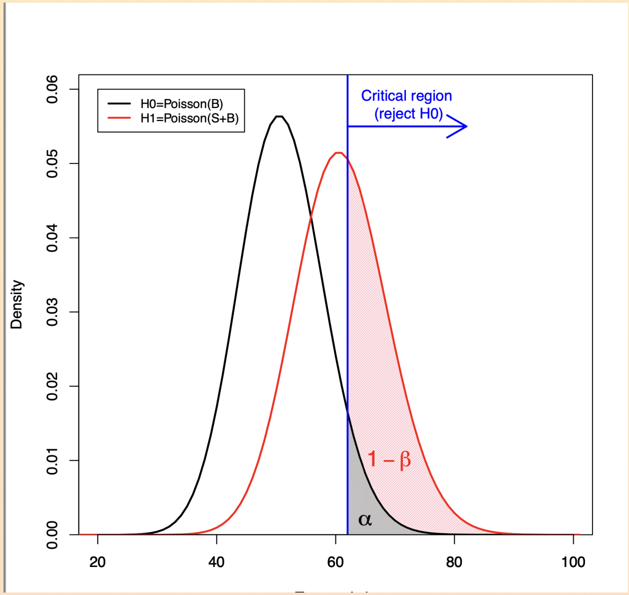
- Hypothesis: a complete rule that defines probabilities for data.
- Statistical test: a proposition on compatibility of H_0 with the available data.
  - $X \in \Omega$ a test statistic
  - Critical region W: if $X \in W$, reject H_0, Acceptance region>: if $X \in \Omega - W$, accept H_0
  - Level of significance (size of the test): $P(X \in W | H_0) = \alpha$

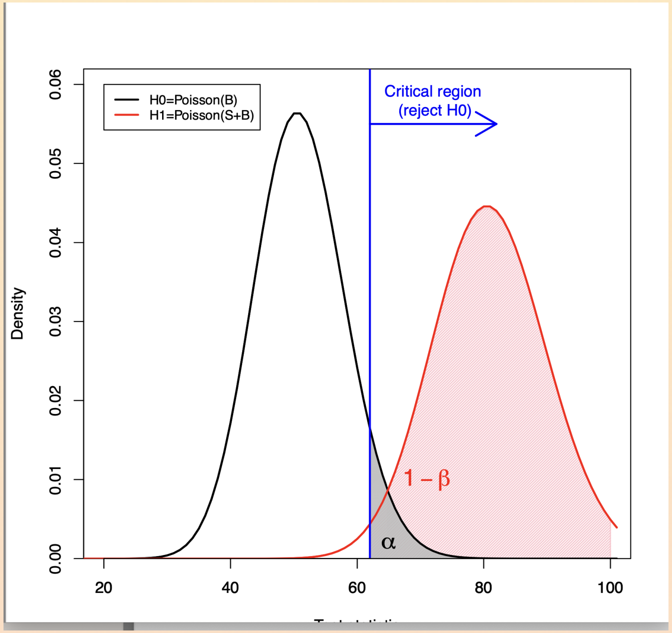




Alternative hypothesis and power

- Need an alternative to solve ambiguities
- Power of the test
  - $P(X \in W | H_1) = 1 - \beta$
  - Power β is such that $P(X \in \Omega - W | H_1) = \beta$

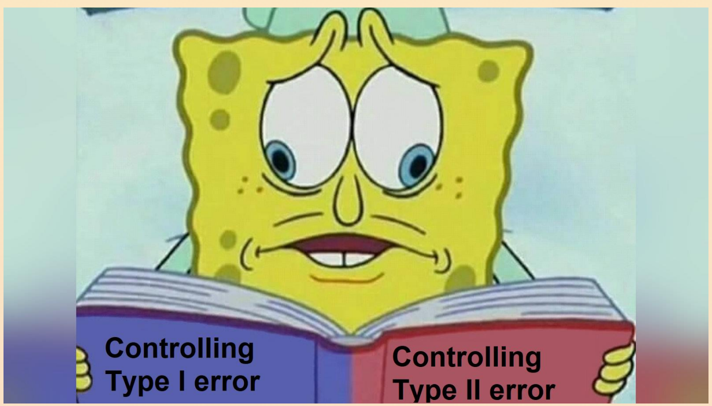


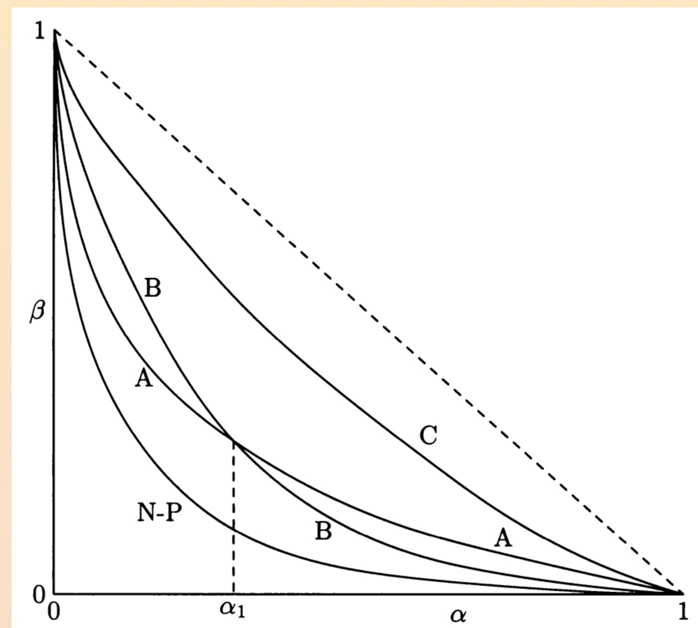


Families of Tests

- Varying α and β results in families of tests
- In one dimension, likelihood ratio (Neyman-Pearson) test is the most powerful test, given by

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





—
Images from Statistical Statistics Memes and James, 2nd ed.

Bayesian Model Selection

- M_0 and M_1 predict θ> : $P(\theta|x, M) = \frac{P(x|\theta,M)P(\theta|M)}{P(x|M)}$
    - Bayesian evidence (Model likelihood) $P(x|M) = \int P(x|\theta, M)P(\theta|M)d\theta$
    - Posterior for M_0: $P(M_0|x) = \frac{P(x|M_0)\pi(M_0)}{P(x)}$, posterior for M_1: $P(M_1|x) = \frac{P(x|M_1)\pi(M_1)}{P(x)}$
    - Posterior odds: $\frac{P(M_0|x)}{P(M_1|x)} = \frac{P(x|M_0)\pi(M_0)}{P(x|M_1)\pi(M_1)}$
    - Bayes factor: $B_{01} := \frac{P(x|M_0)}{P(x|M_1)}$
    - Posterior odds = Bayes Factor $\times$ prior odds
- Turing (IJ Good, 1975): deciban as the smallest change of evidence human mind can discern

Jeffreys

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

Kass and Raftery

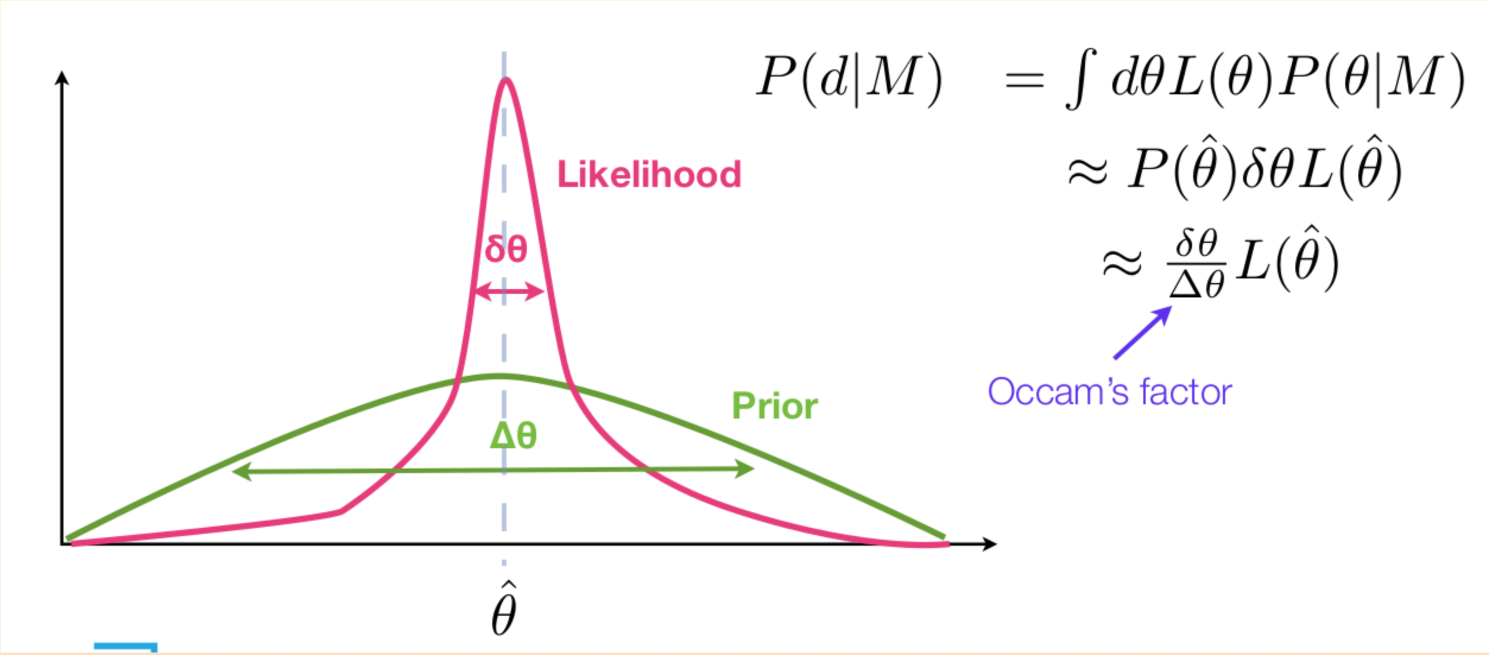
$\log_{10} K$	K	Strength of evidence
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Trotta

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

Discourage nonpredictive models

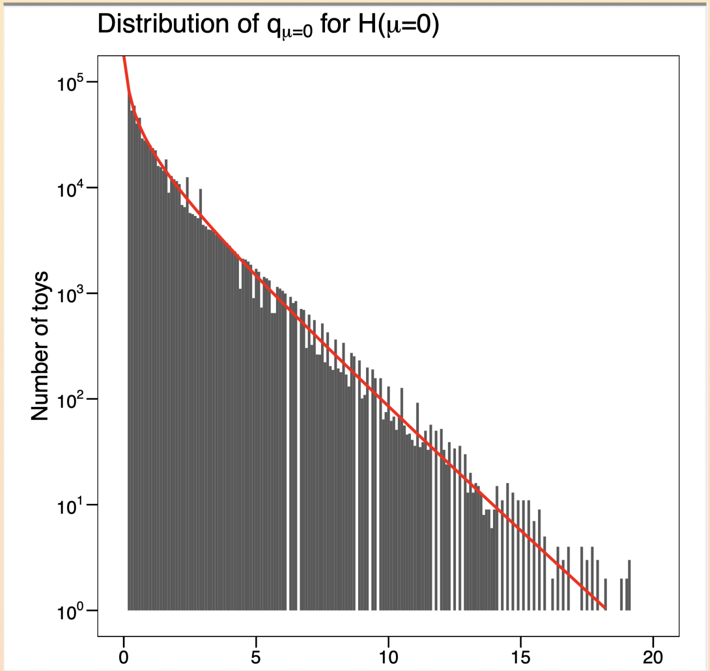
- The Bayes Factor penalizes excessive model complexity
- Highly predictive models are rewarded, broadly-non-null priors are penalized

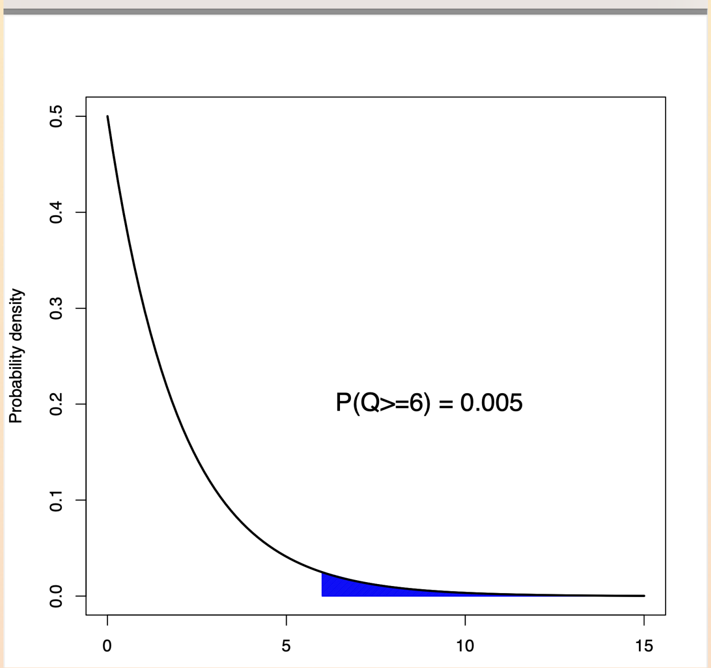


—
Image from Trotta, Chair Lemaitre 2018

P-values

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

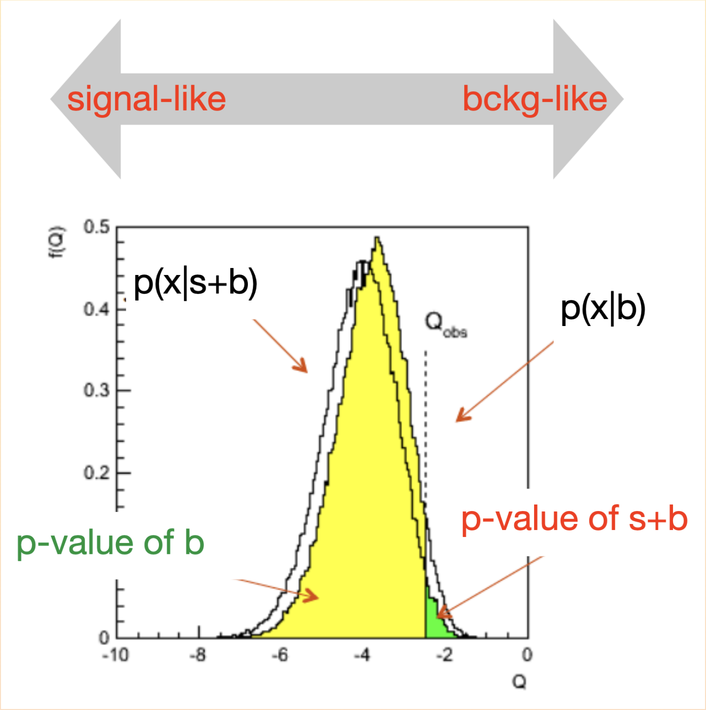




Beyond frequentism: CLs

- $CL_s := \frac{CL_{s+b}}{CL_b}$
- Exclude the signal hypothesis at confidence level CL if $1 - CL_s \leq CL$

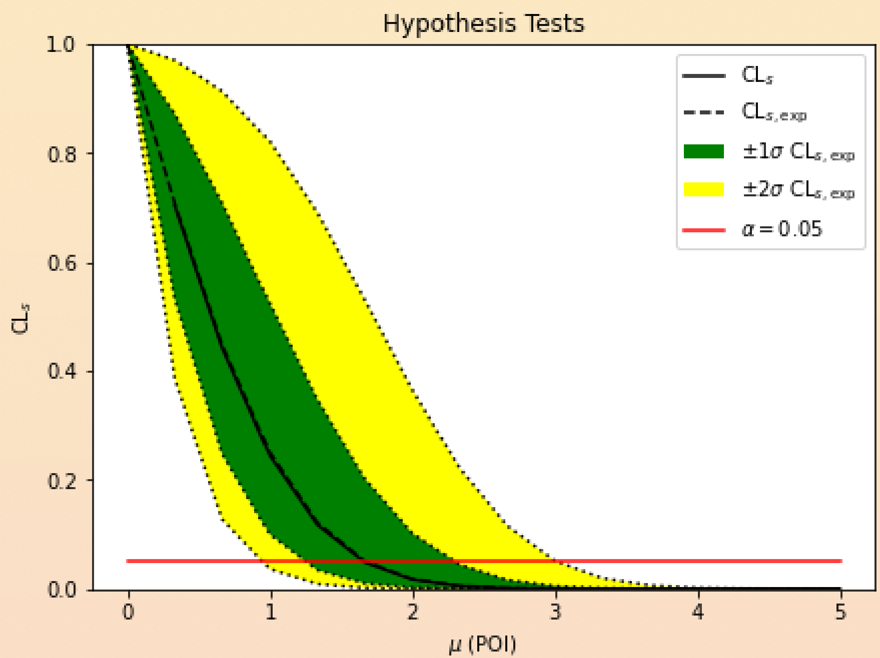
- Ratio of p-values is not a p-value
- Denominator prevents excluding signals for which there is no sensitivity
- Formally corresponds to have $H_0 = H(\theta! = 0)$ and test it against $H_1 = H(\theta = 0)$



—
Image by Diego Tonelli

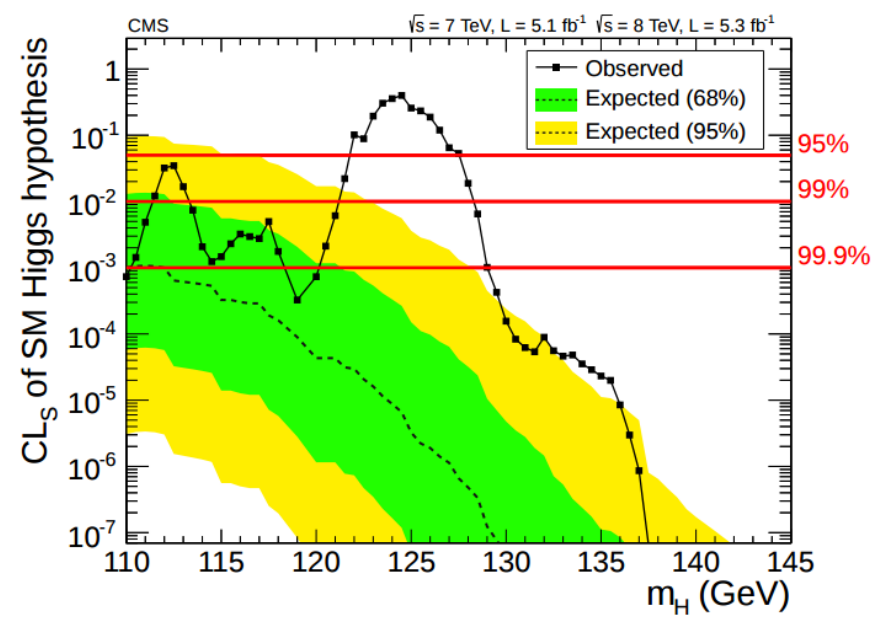
From a scans to limits

- Scan the \$CL$steststatisticasafunctionofthePOI(typically$\mu = \sigma{obs}/\sigma{pred}\$)
- Find intersection with the desired confidence level
- (eventually) convert the limit on μ back to a cross section



From a limit to hypothesis testing

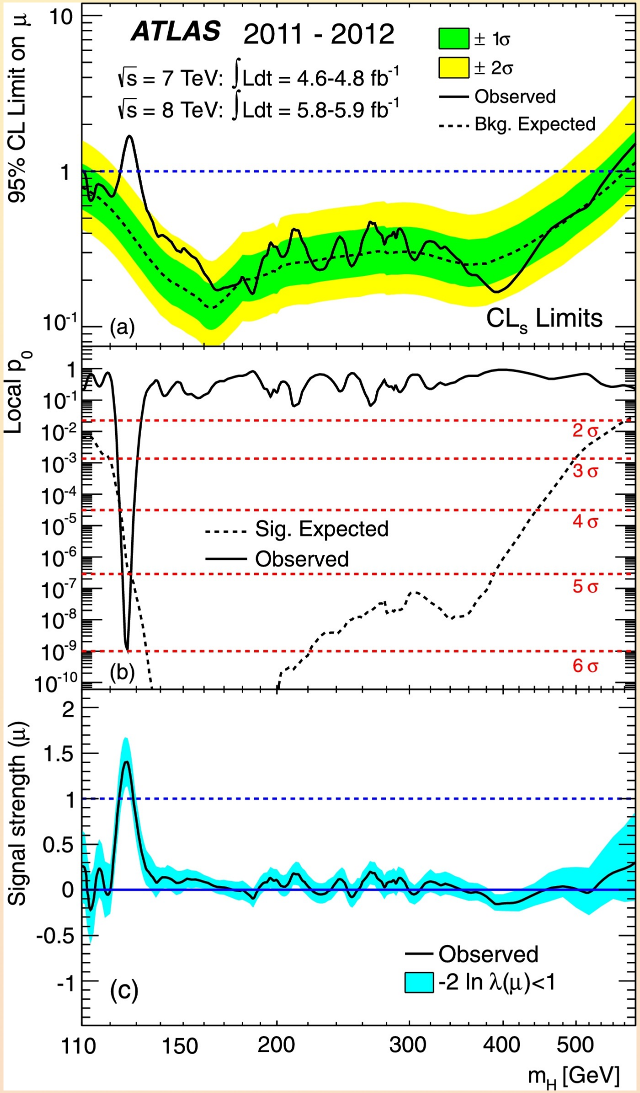
- Apply the CL_s method to each Higgs mass hypothesis
- Show the CL_s test statistic for each value of the fixed hypothesis
- Green/yellow bands indicate the $\pm 1\sigma$ and $\pm 2\sigma$ intervals for the expected values under B-only hypothesis
  - Obtained by taking the quantiles of the B-only hypothesis



—
Plot from the CMS Higgs discovery paper

From a limit to hypothesis testing

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

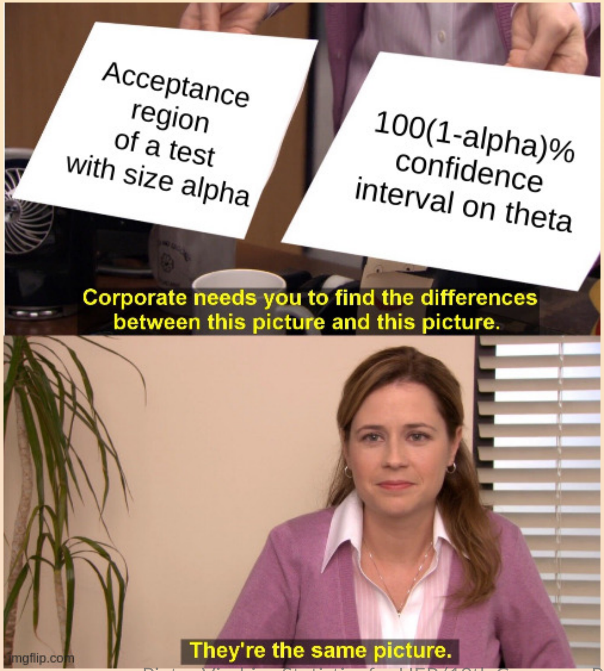


Plot from the ATLAS Higgs discovery paper

Duality

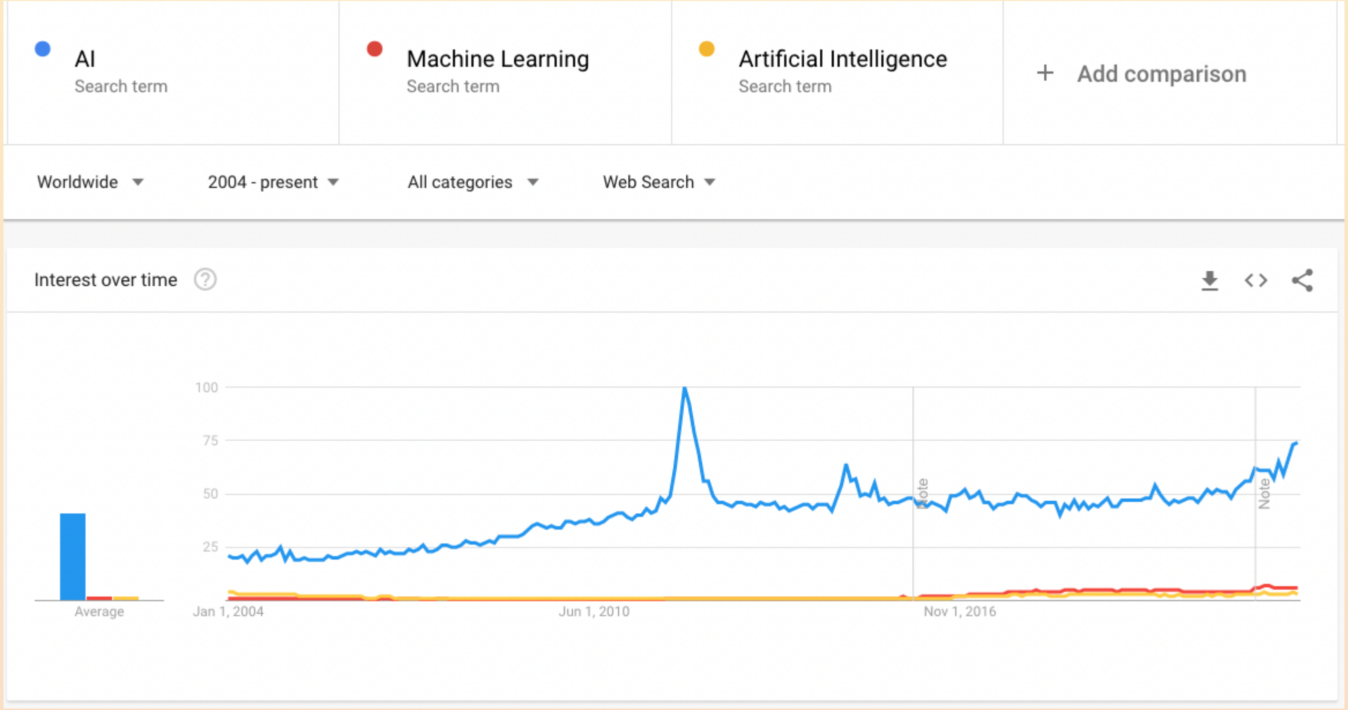
Meme generated with memegenerator

- Acceptance region set of values of the test statistic for which we don't reject H_0 at significance level α
- $100(1-\alpha)\%$ confidence interval: set of *values of the parameter θ for which we don't reject H_0 (if H_0 is assumed true)



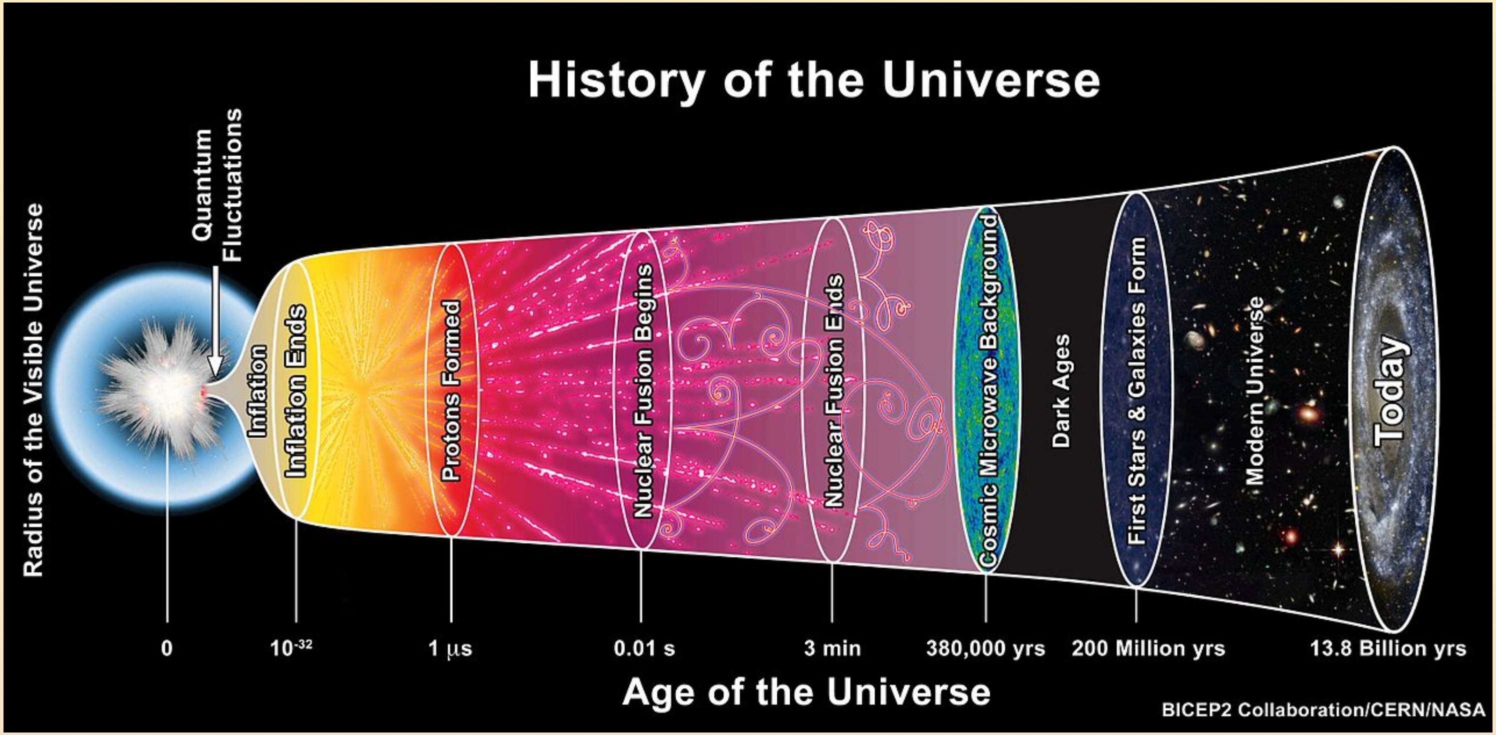
AI, the eternal buzzword?

- Artificial Intelligence (AI)
- Machine Learning (ML)
- Statistical Learning



—
Image produced by using Google Trends

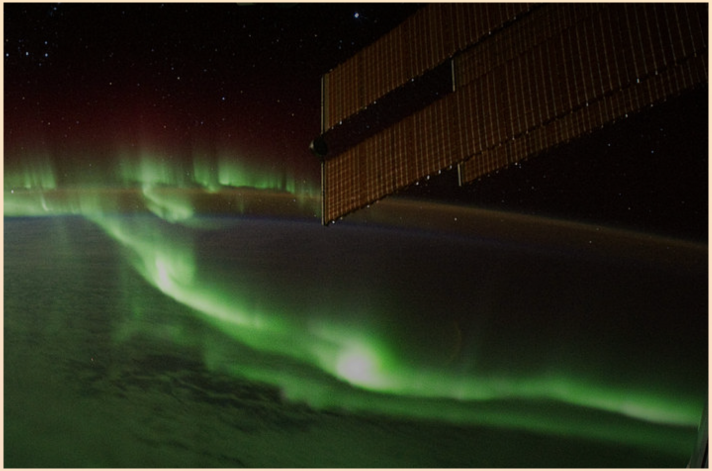
We Try to Understand the Universe



—
Image from CERN

Fundamental $\longrightarrow$ Applied

- X/γ detectors (Xrays, PET)
- Hadron therapy and proton CT
- Vacuum technology
- Cryogenics
- Art
- WWW
- MRI
- GPS
- Satellites
- Solar panels
- Airport security scanners
- Space watch (avoid asteroids)



—
Images from CERN

Applied $\longrightarrow$ Fundamental

—
Image from Pixabay

AI: from Applied to Fundamental

Bulletin of Mathematical Biology Vol. 52, No. 1/2, pp. 99–115, 1990.
Printed in Great Britain.

0092–8240/90$3.00 + 0.00
Pergamon Press plc
Society for Mathematical Biology

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*

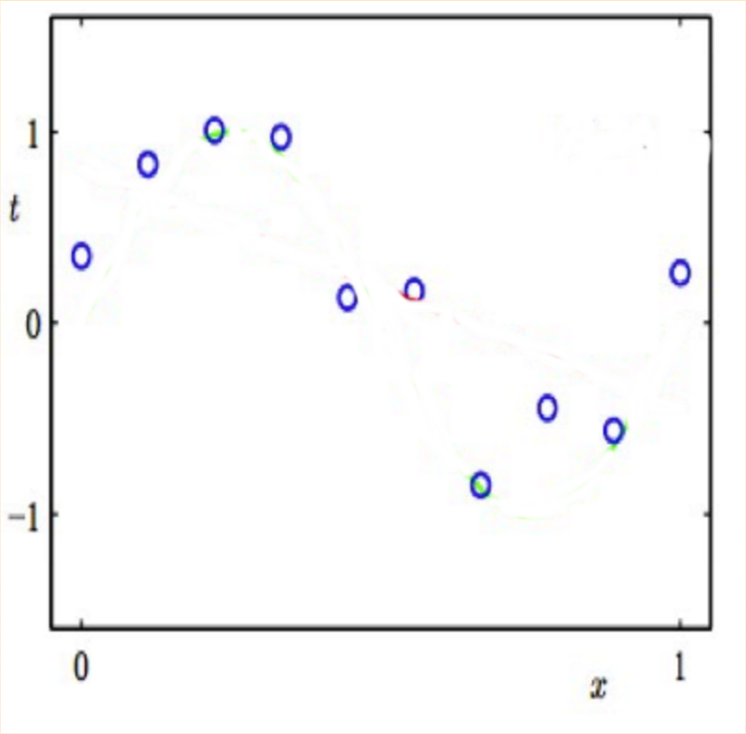
■ WARREN S. MCCULLOCH AND WALTER PITTS
University of Illinois, College of Medicine,
Department of Psychiatry at the Illinois Neuropsychiatric Institute,
University of Chicago, Chicago, U.S.A.

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

Understanding Data

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)

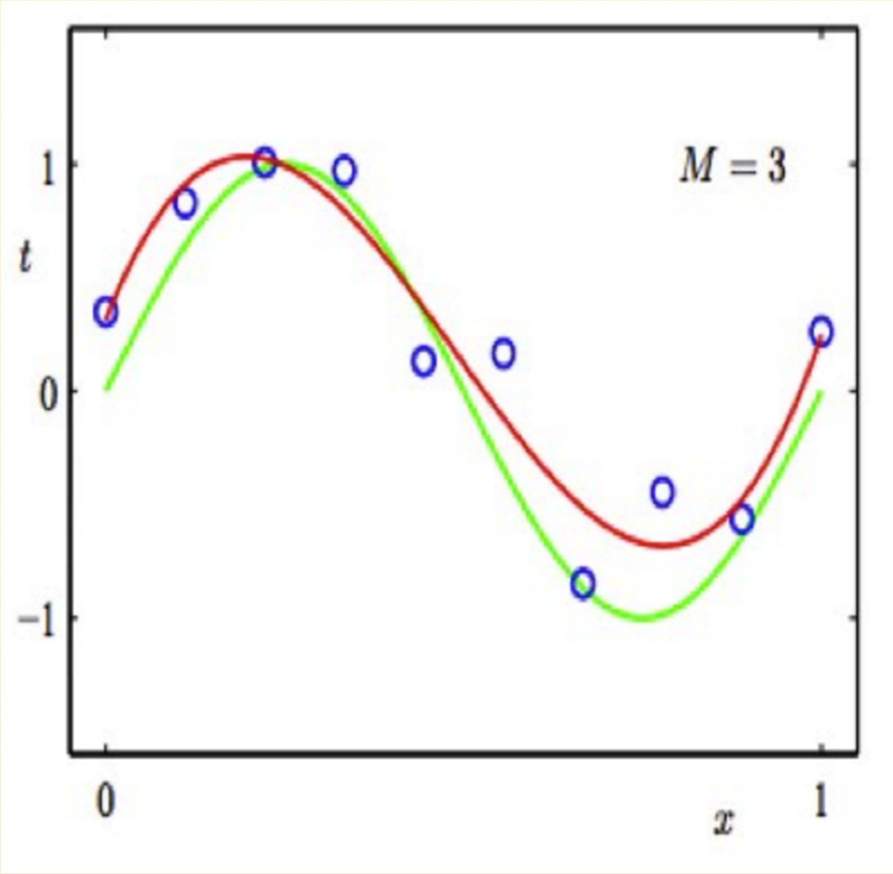


—

Image by Victor Lavrenko

Functions Describe the World

- Interpolation

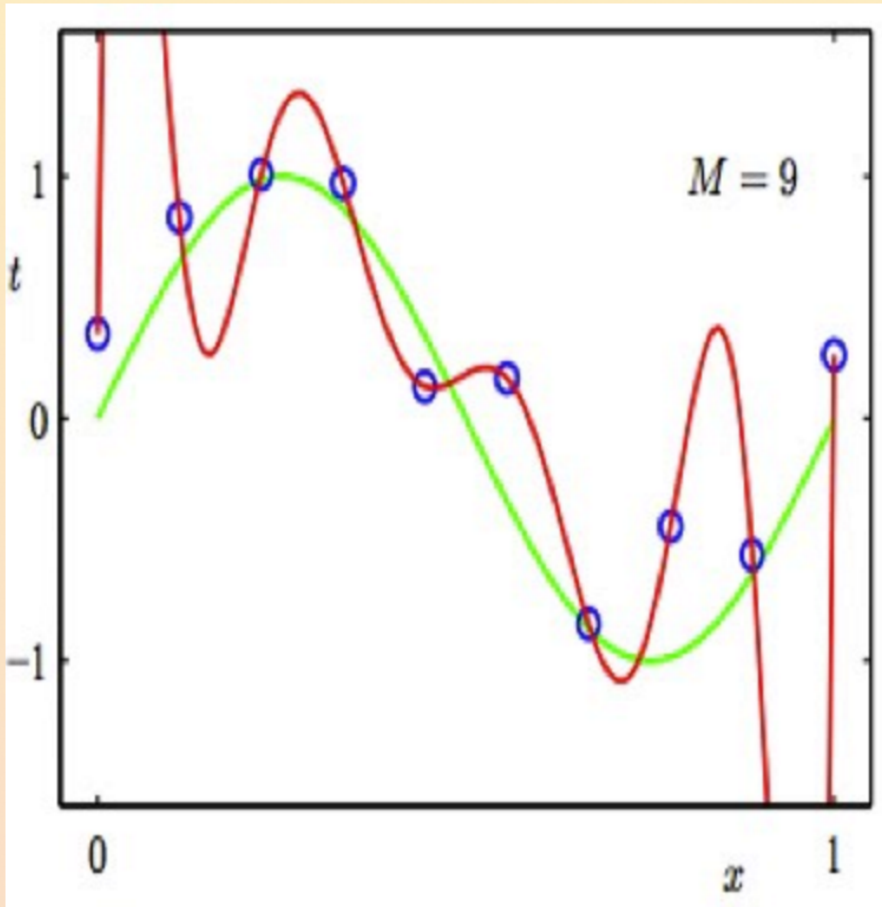


—

Image by Victor Lavrenko

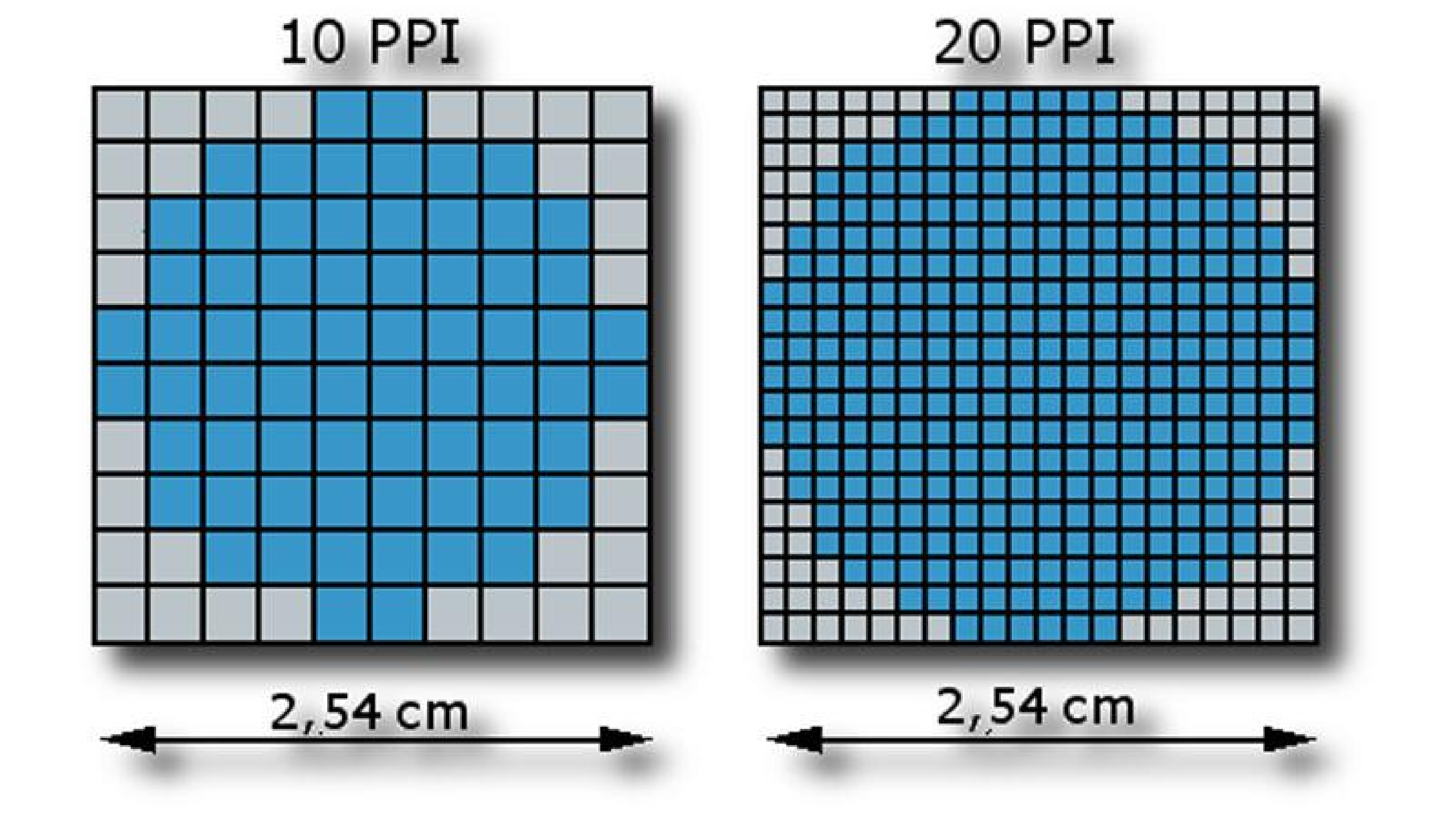
Sometimes too well

- Generalization



—
Image by Victor Lavrenko

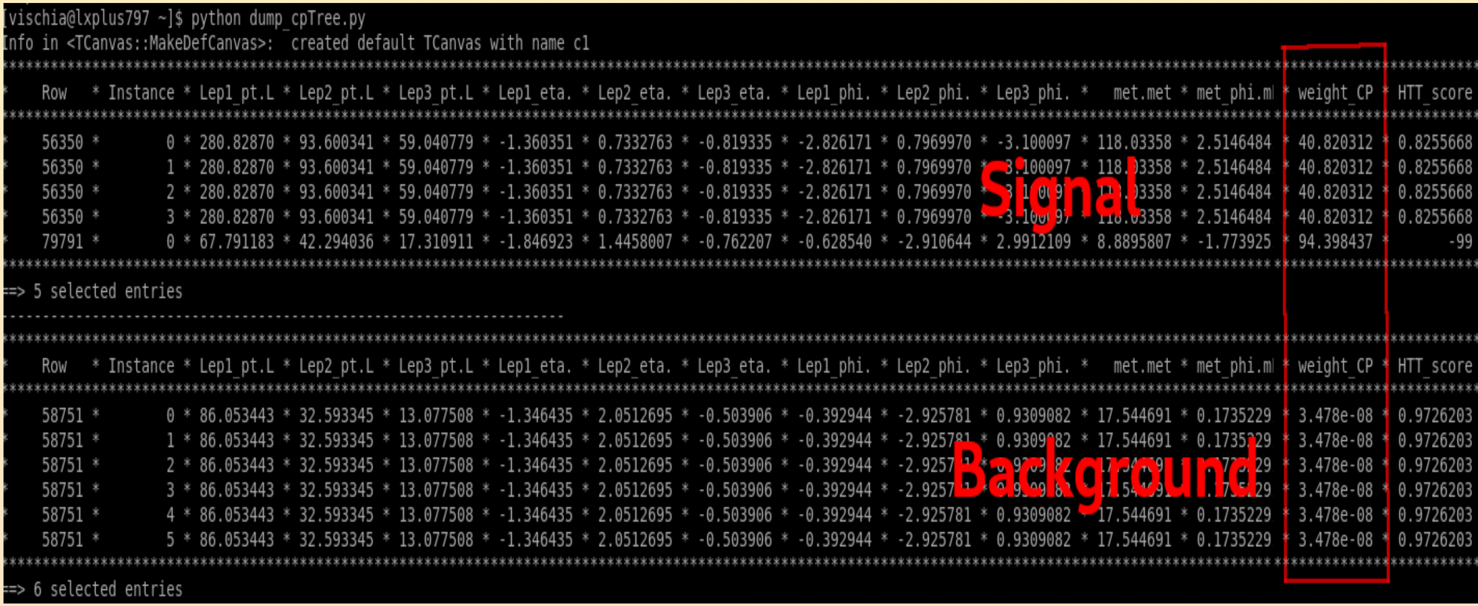
Think in Millions of Dimensions



—

Image from scientiamobile.com

Easy or difficult?



[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	[illegible]	[illegible]	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]	[illegible]	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	[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

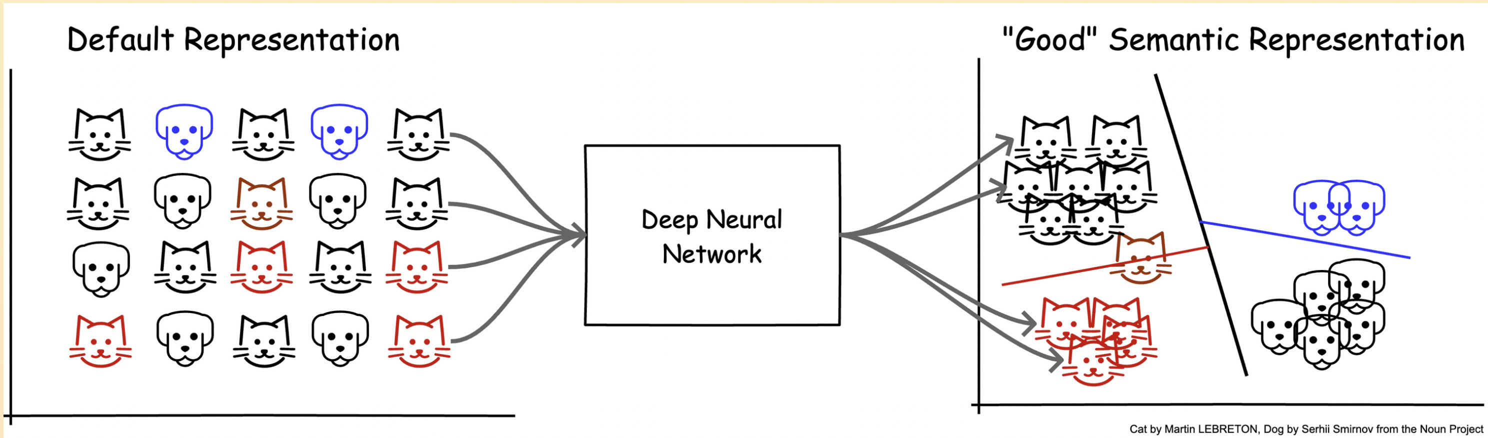
Easy or difficult?

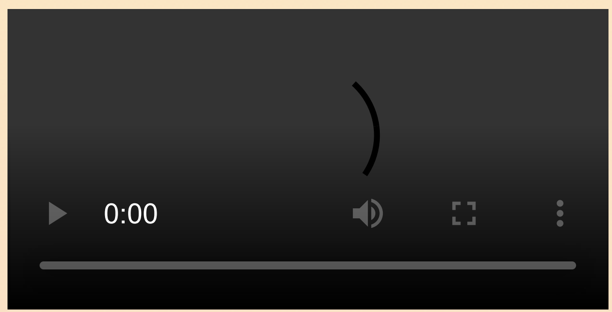
Image from indiatimes.com

Mapping Improves Understanding

—
Tabula Rogeriana by Al Idrisi (1154) 1929 reproduction

Representations Make Tasks Easier

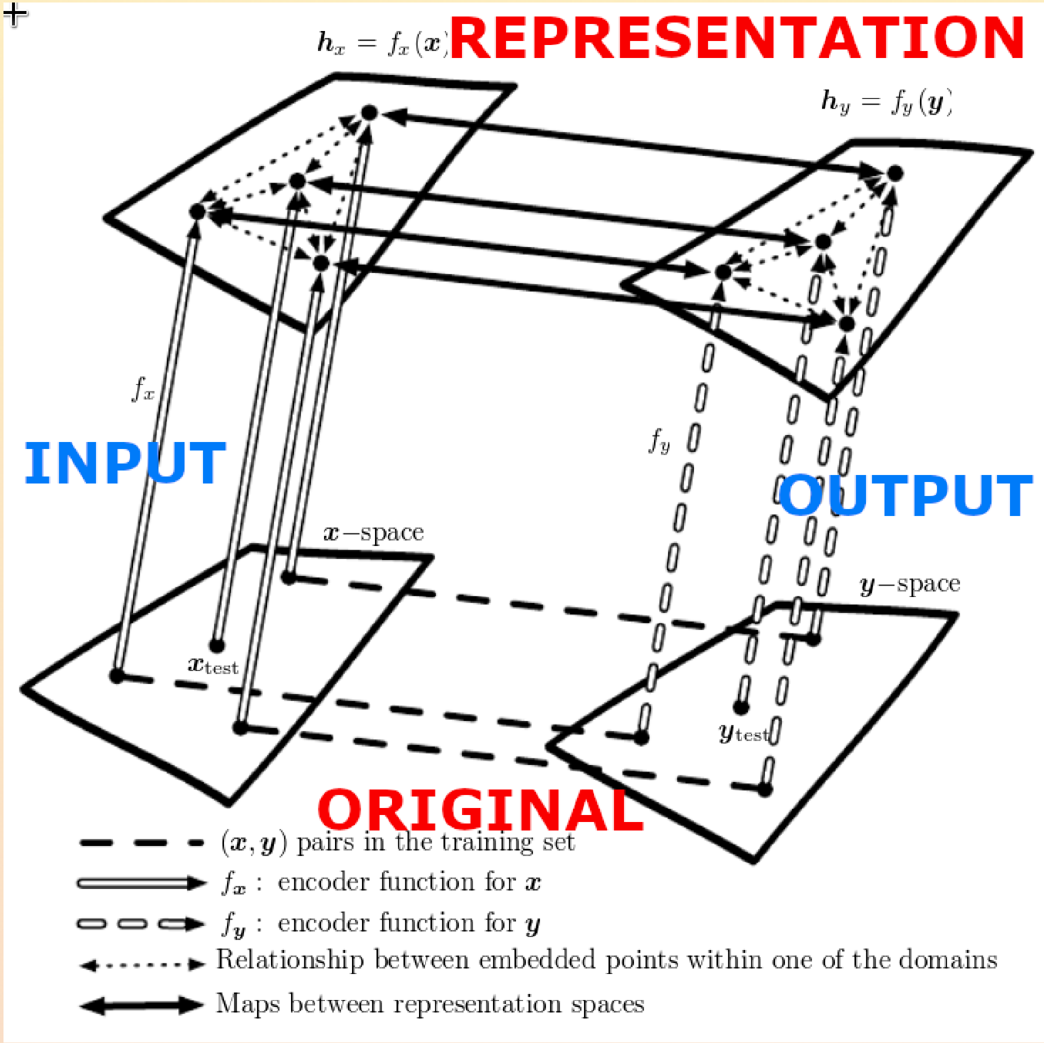




—

Animation and picture from FastForward Labs

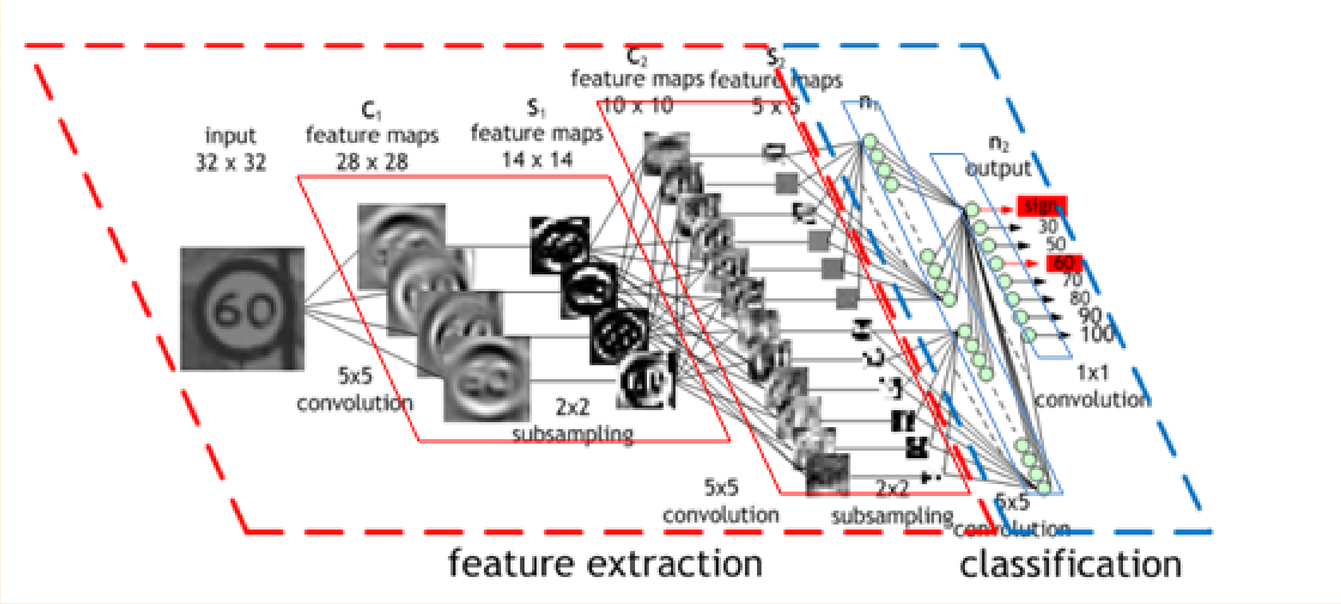
Learn Representations



—

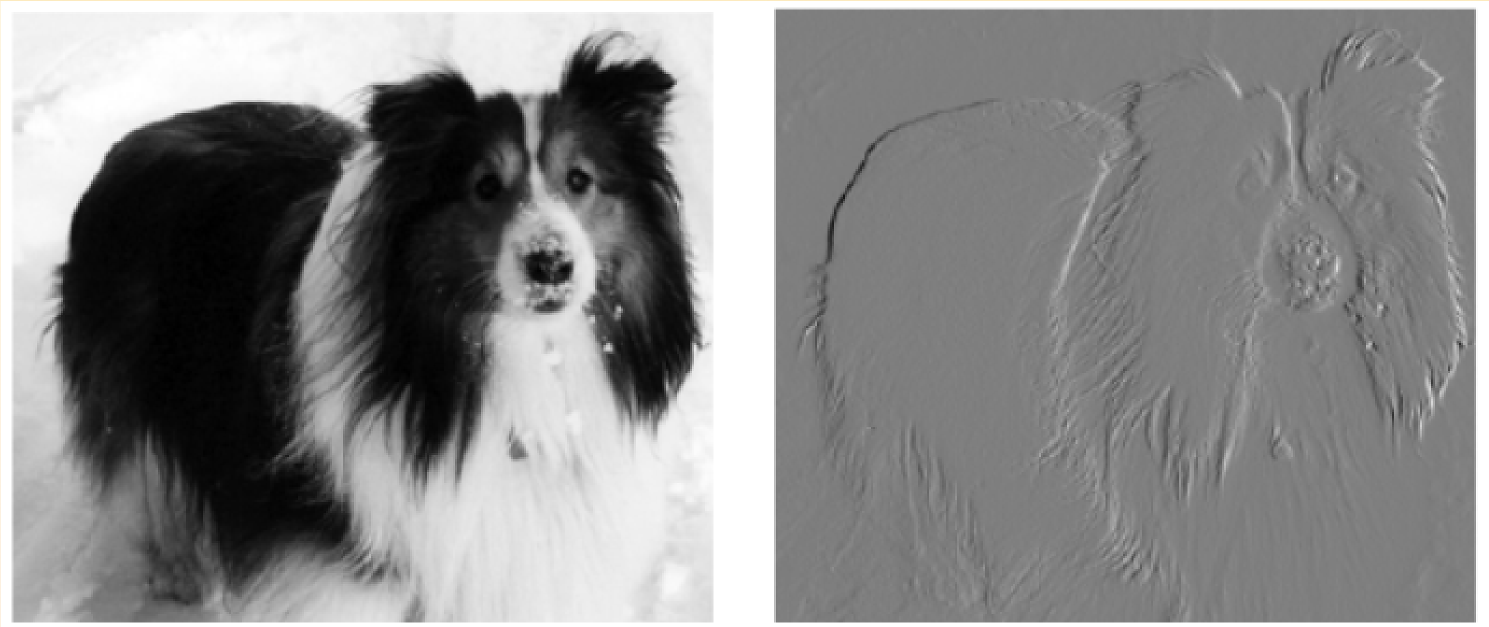
Images from Goodfellow, Bengio, Courville, 2016

Convolutional networks



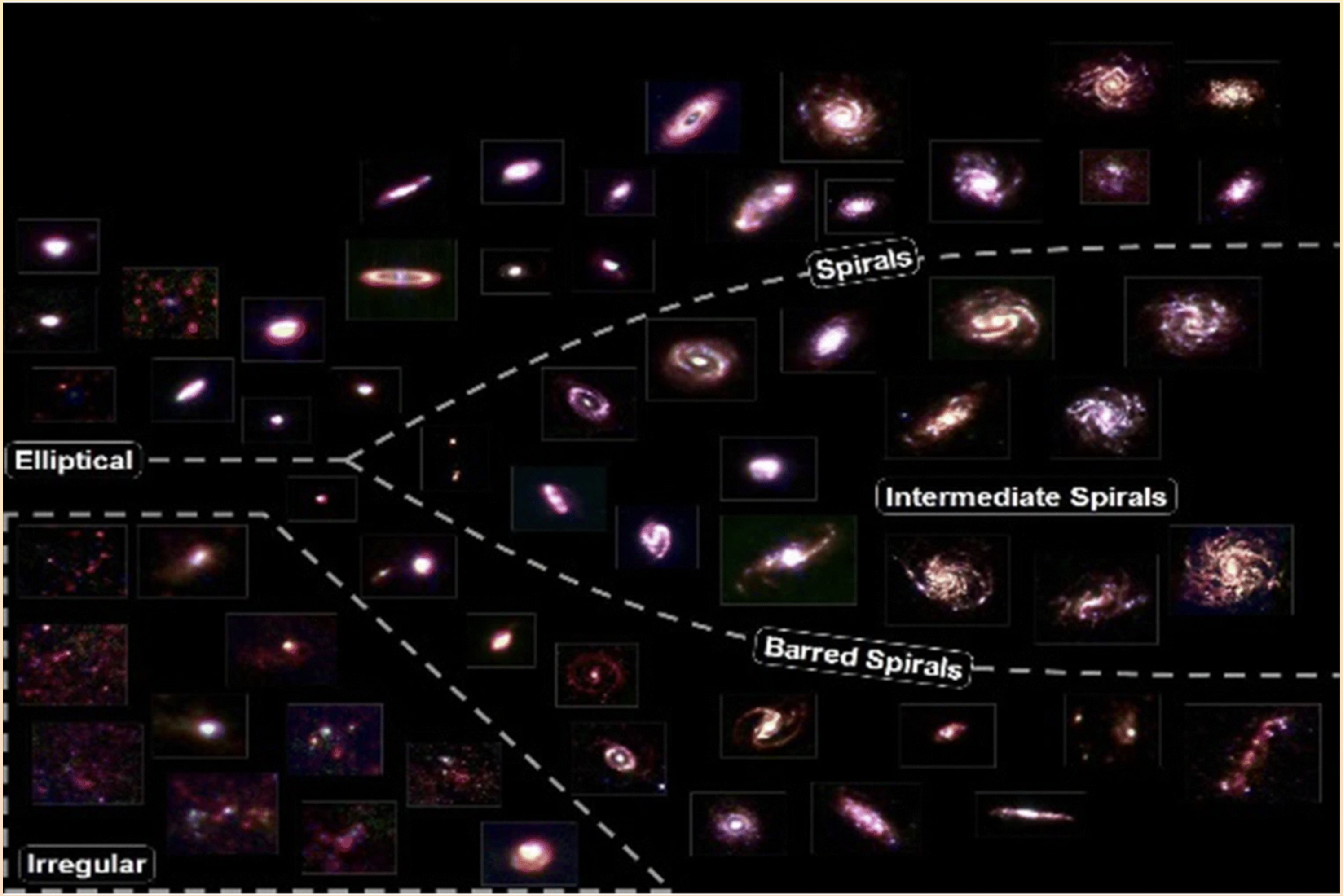
—
Images from tue.nl and Yann LeCun's LeNet

Intermediate representations



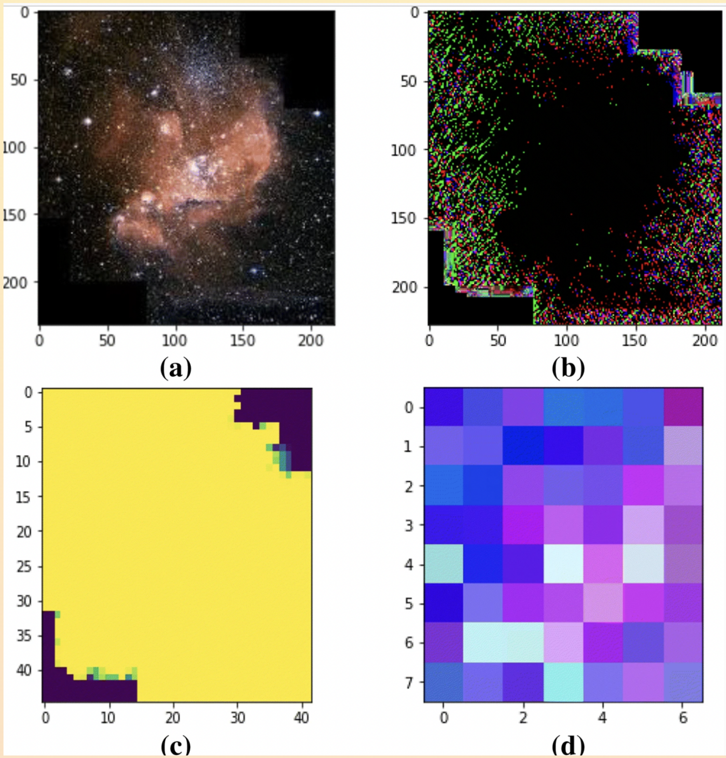
—
Images from Goodfellow, Bengio, Courville, 2016

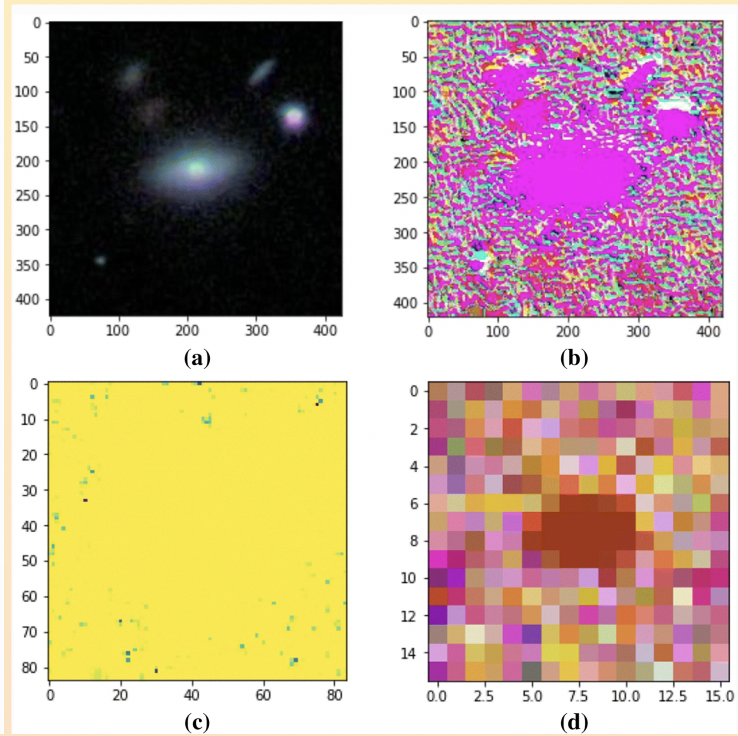
Morphology of galaxies



—
Image from Earth Science Informatics volume 13, pages 601–617 (2020)

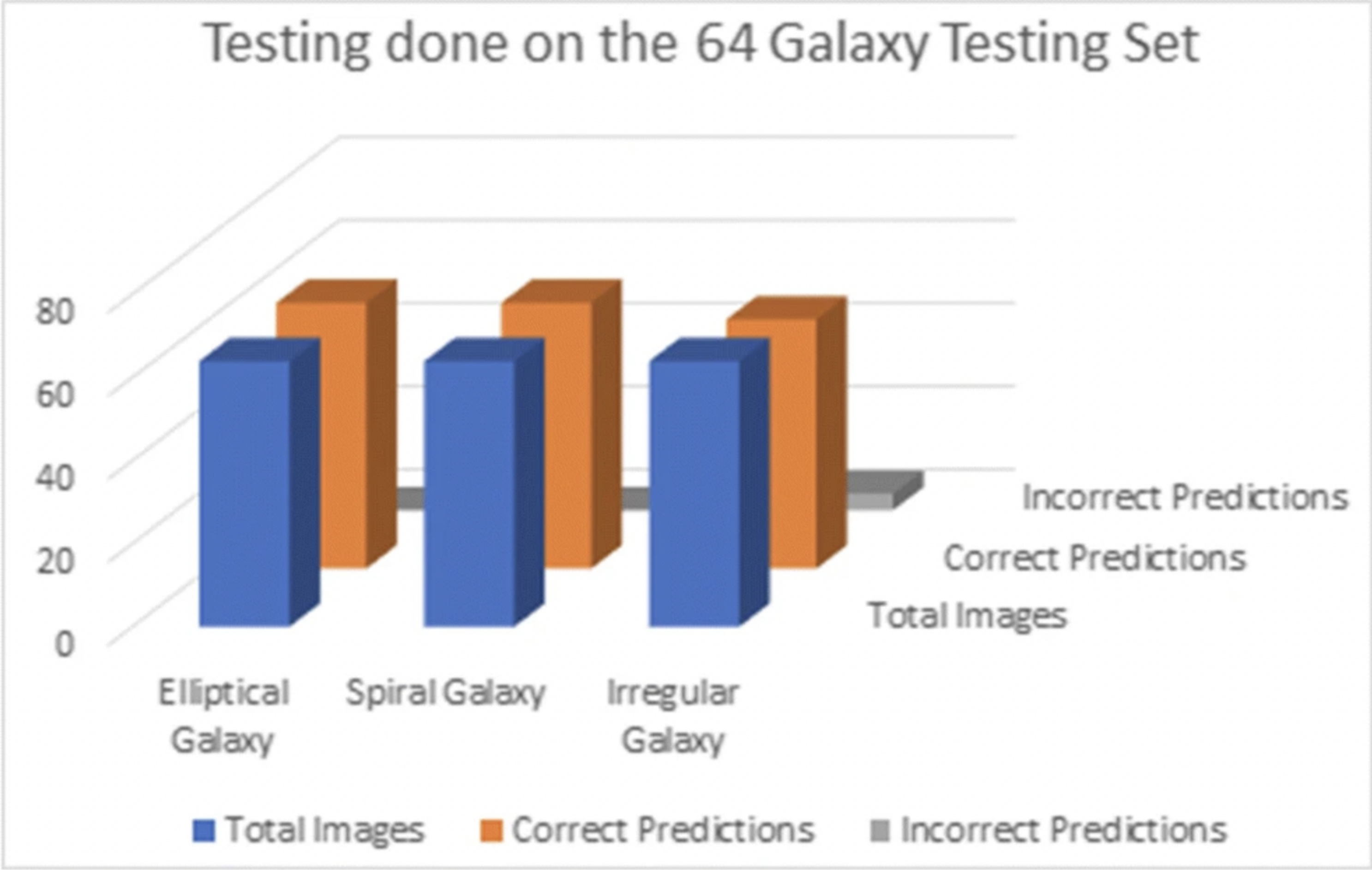
Representations of galaxies...





—
Image from Earth Science Informatics volume 13, pages 601–617 (2020)

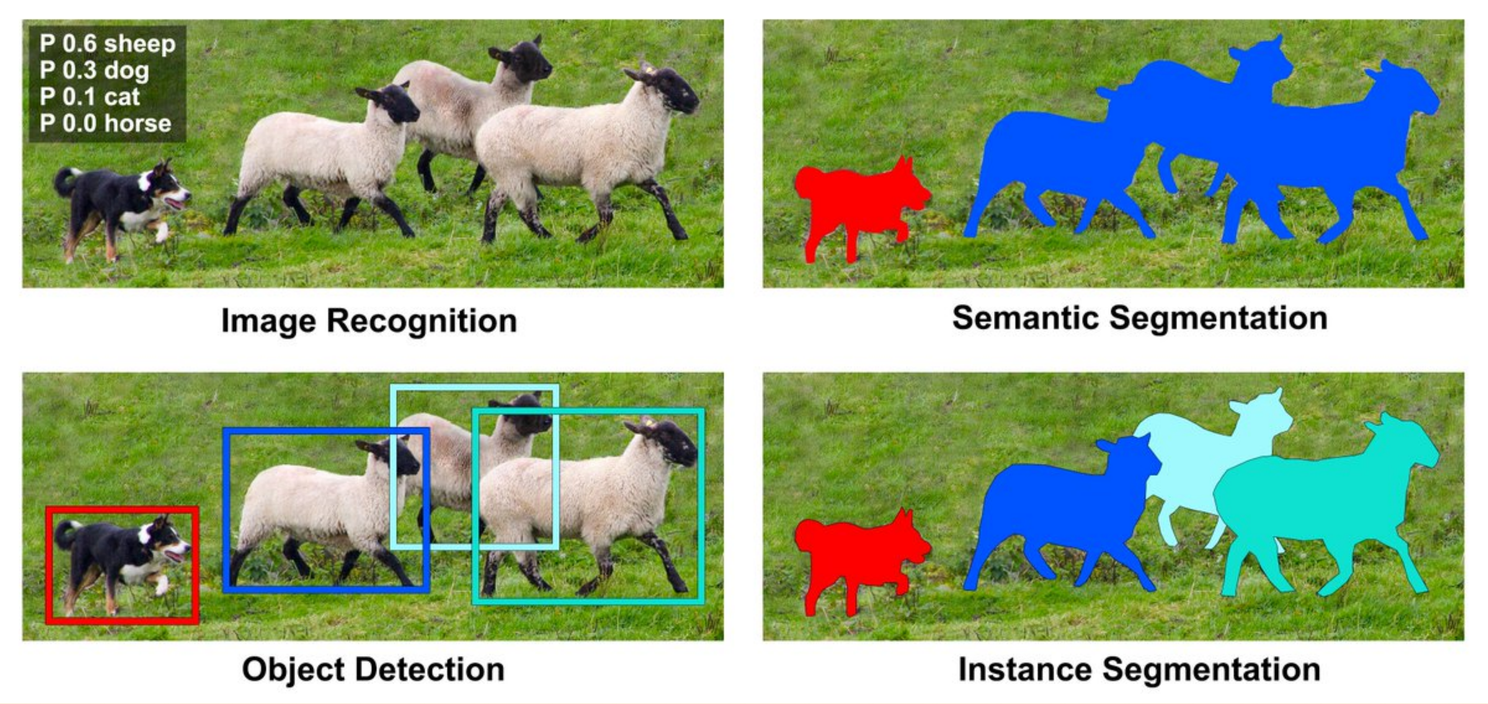
...work pretty well



—

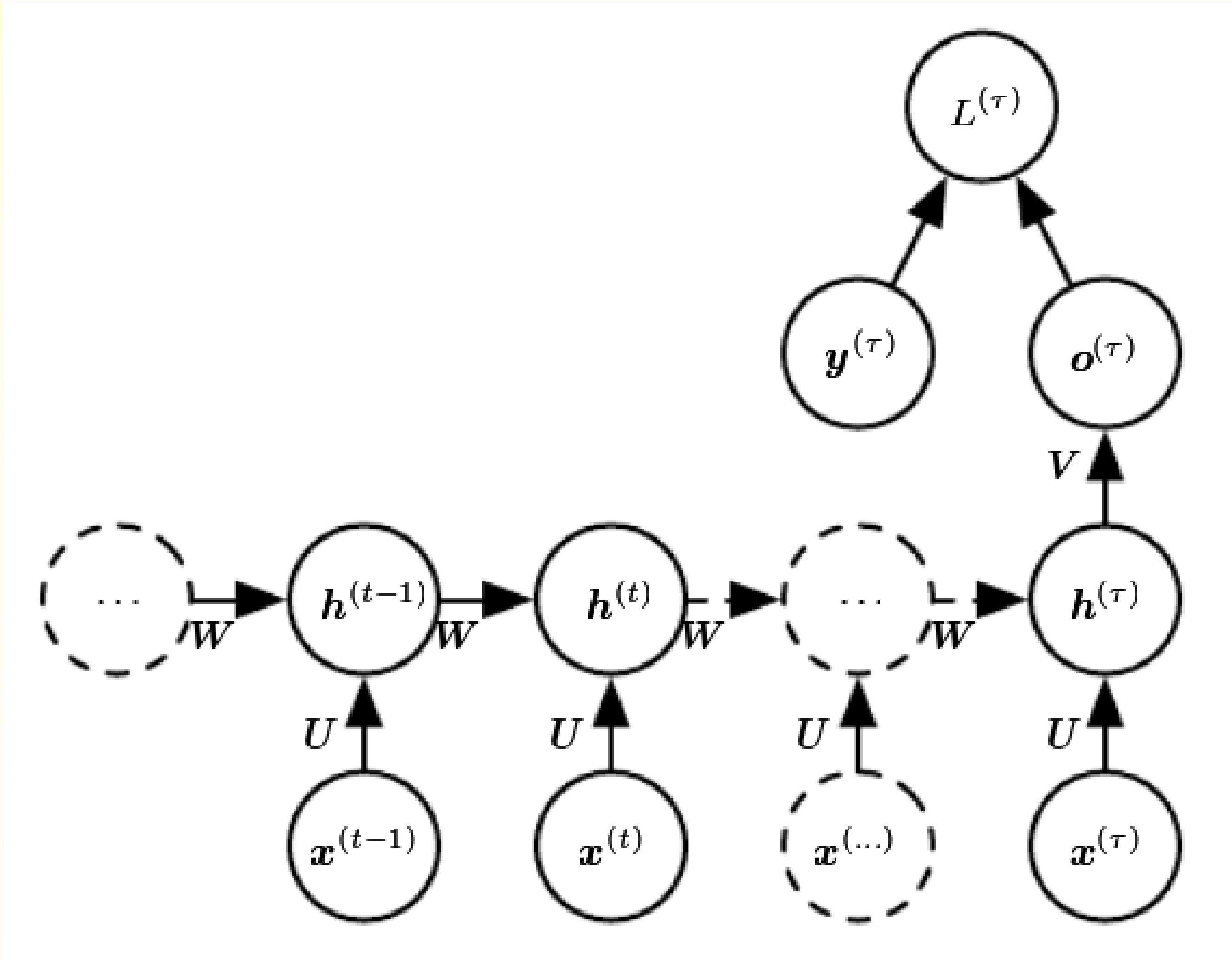
Image from Earth Science Informatics volume 13, pages 601–617 (2020)

Semantic representations



—
Image from ai-pool.com

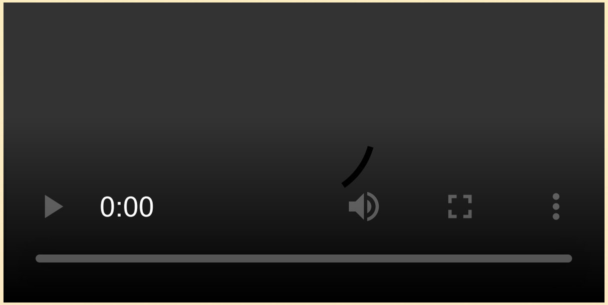
Account for the time component

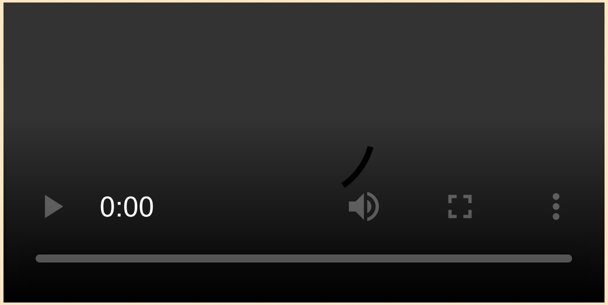


—

Images from Goodfellow, Bengio, Courville, 2016

Real-time segmentation



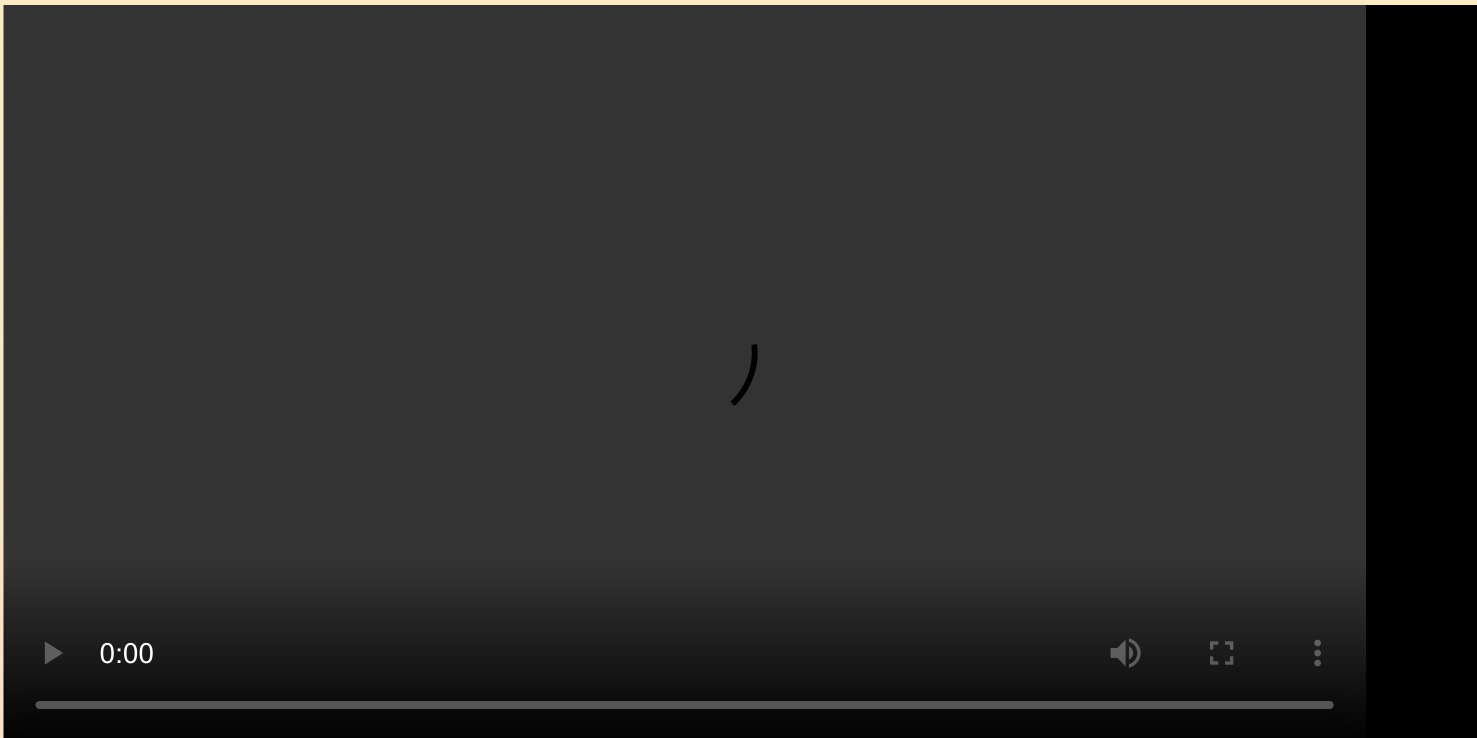


—

Videos from YouTube, autonomous driving and YouTube, cancer research

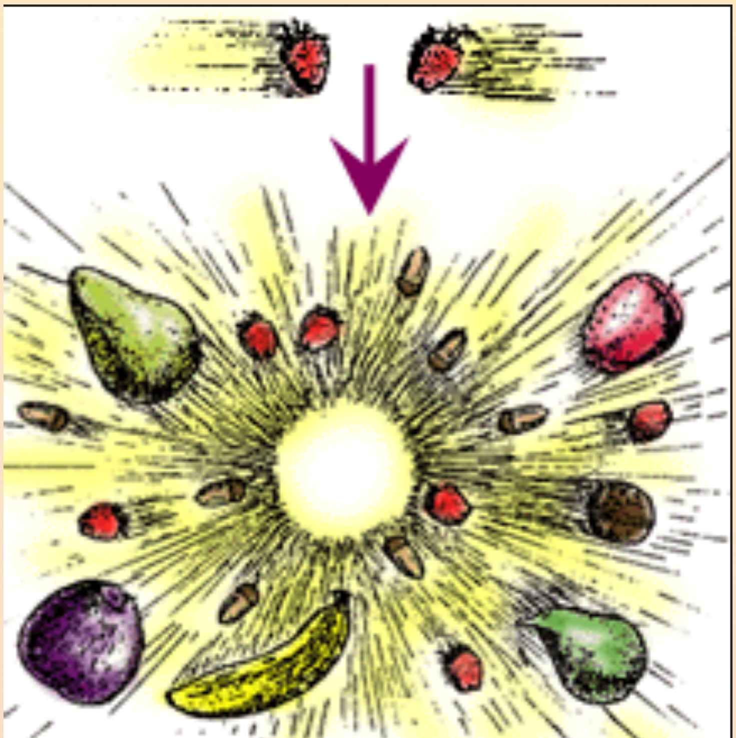
Collide Protons...

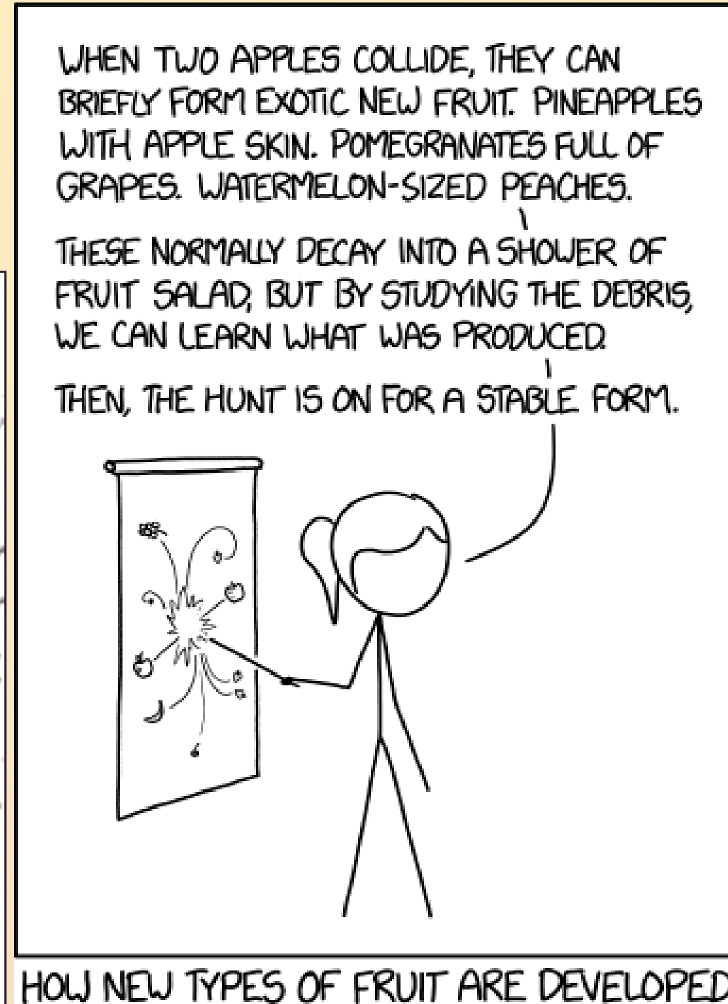
- The kinetic energy of two 88k tons aircraft carriers, each at 10km/h
- Packed into a transverse section of 16 micron



—
Video from YouTube

To See What Happens

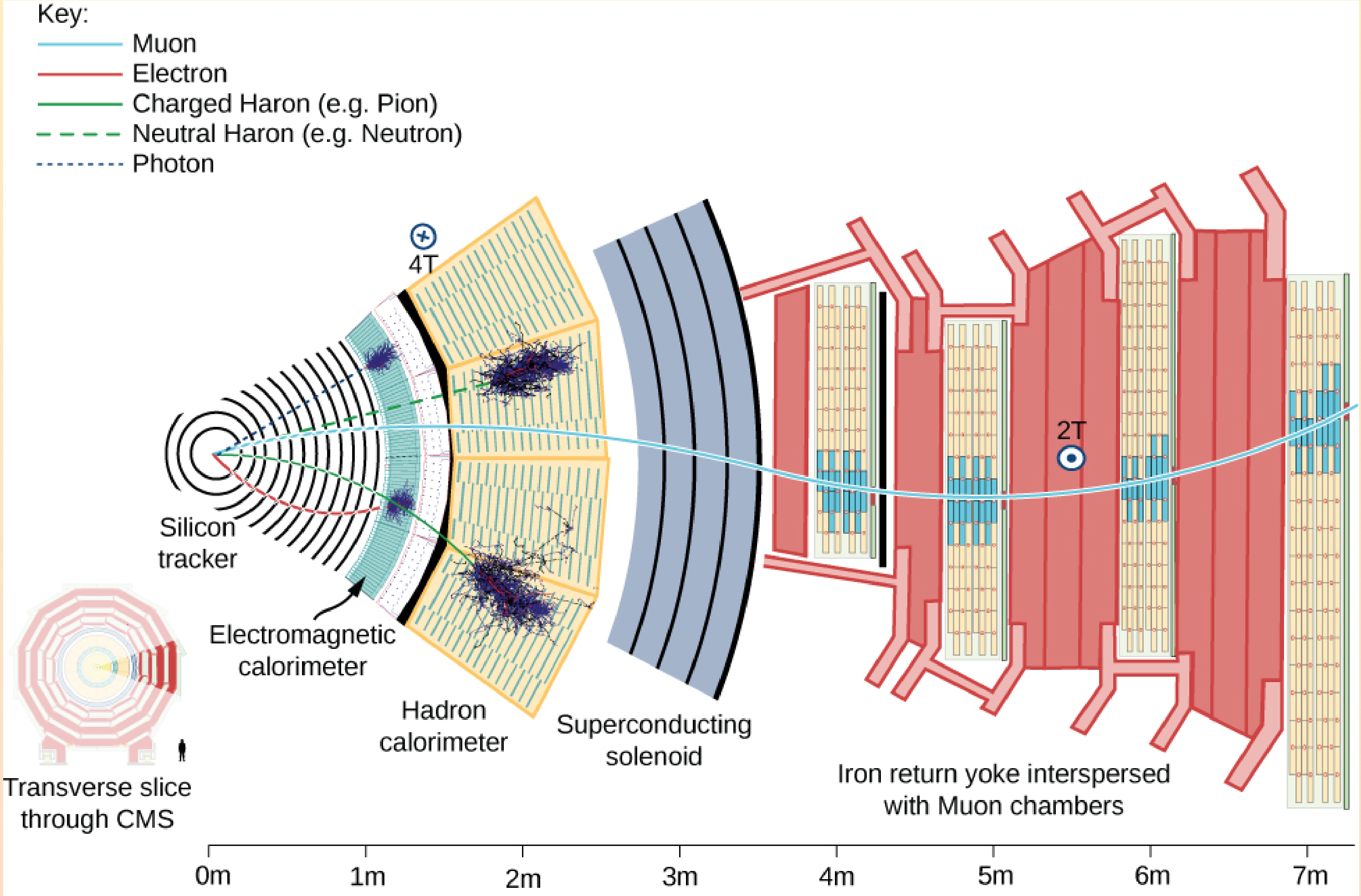




—

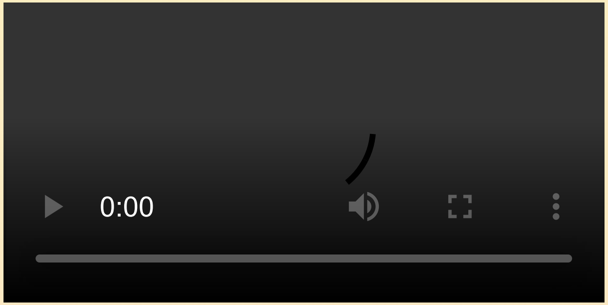
Images from ParticleAdventure.org and XKCD.com

40 Million Times per Second



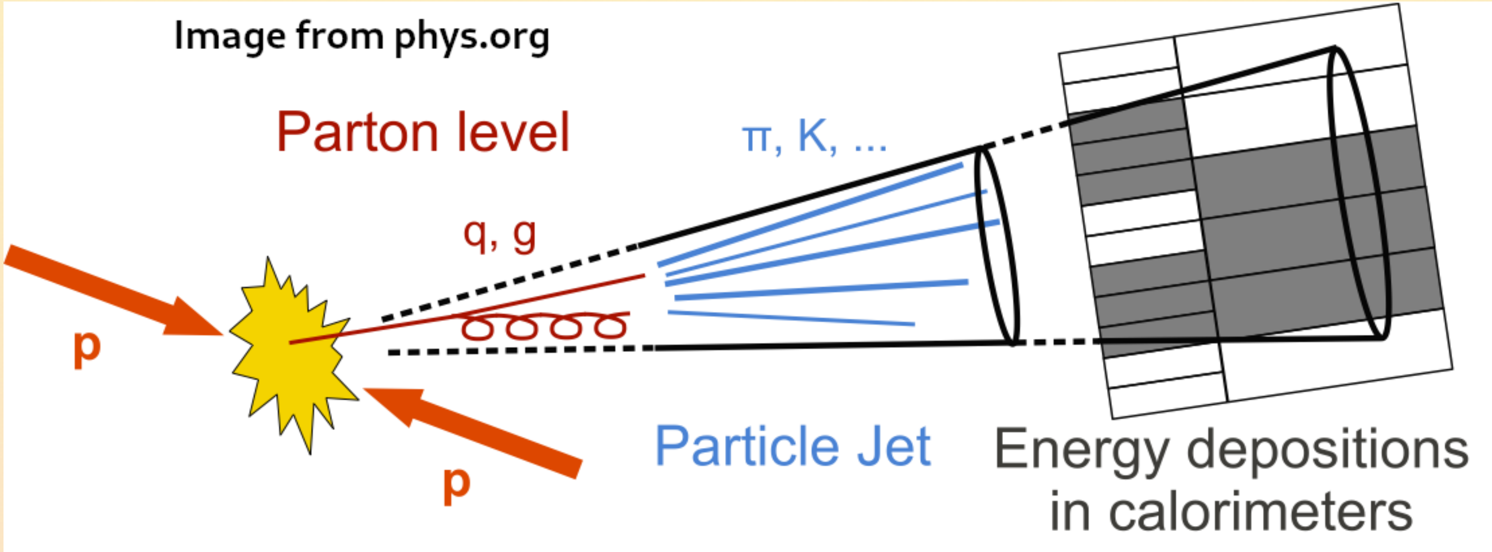
—
Images from the CMS Collaboration

Quark fragmentation



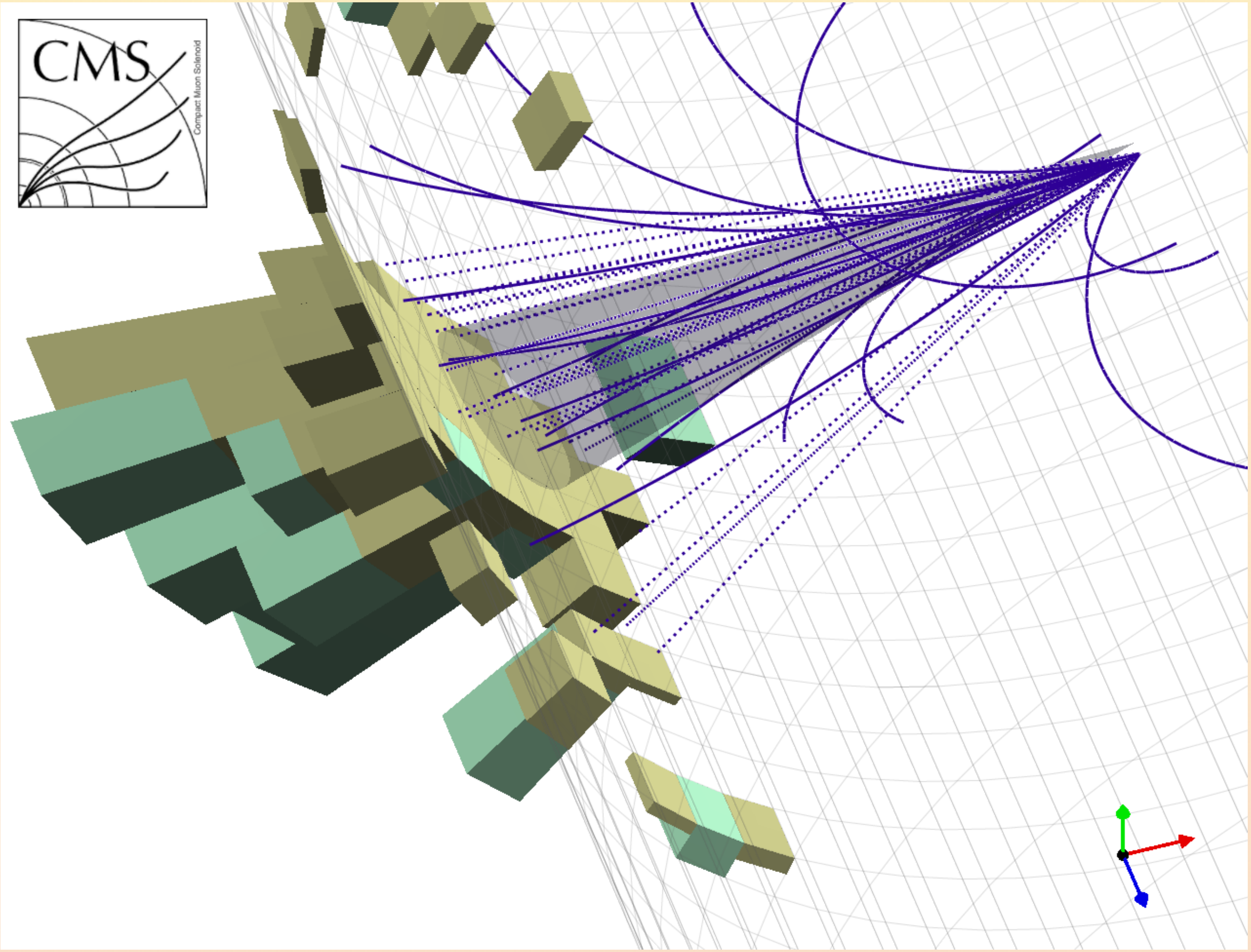
—
Video from the ATLAS experiment YouTube channel

Jets...



—
Image from phys.org

...in the detector...



—
Image from the CMS Collaboration

...exploiting relationships

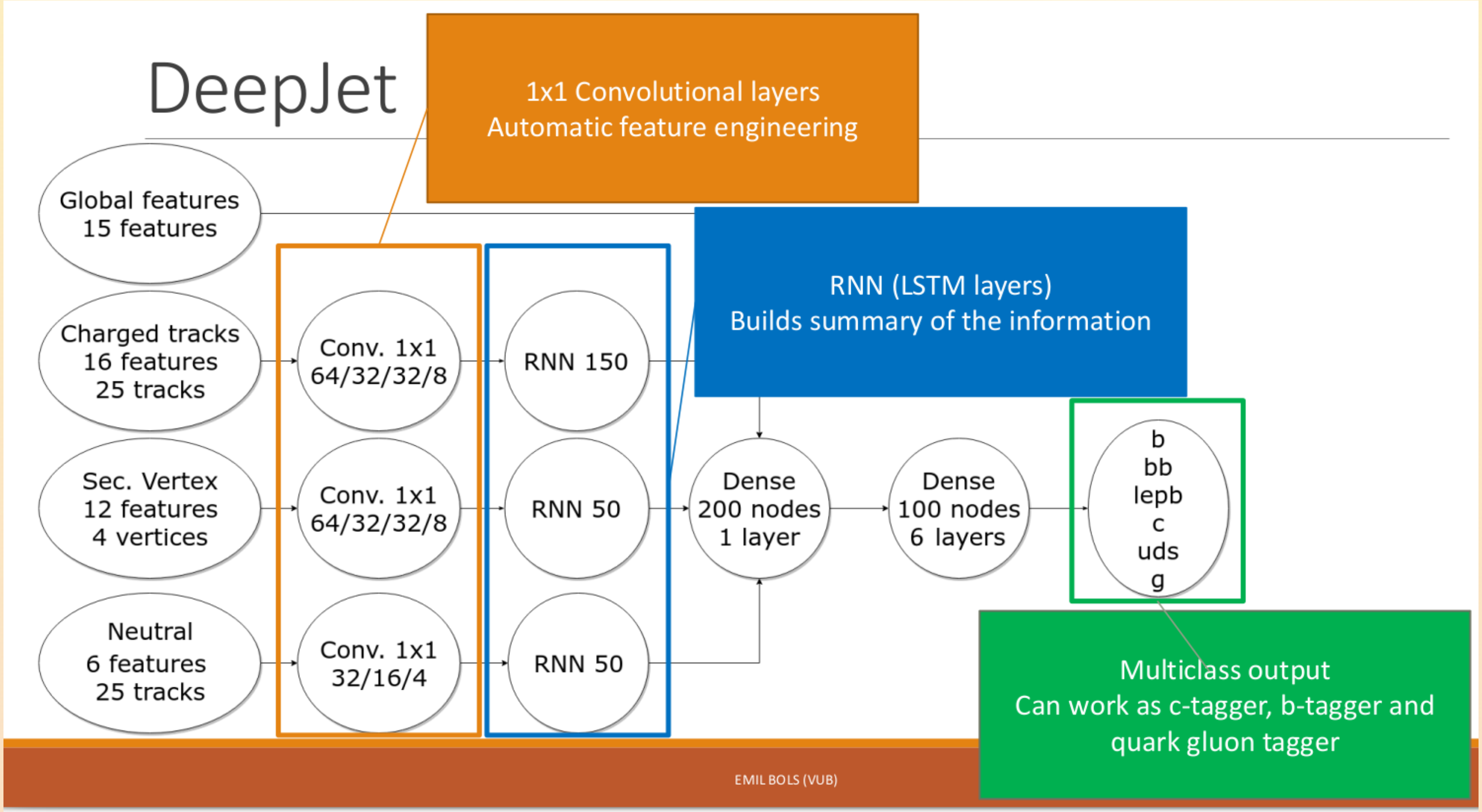
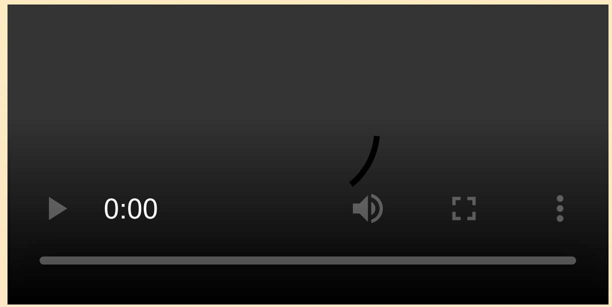


Diagram from a talk by Emil Bols

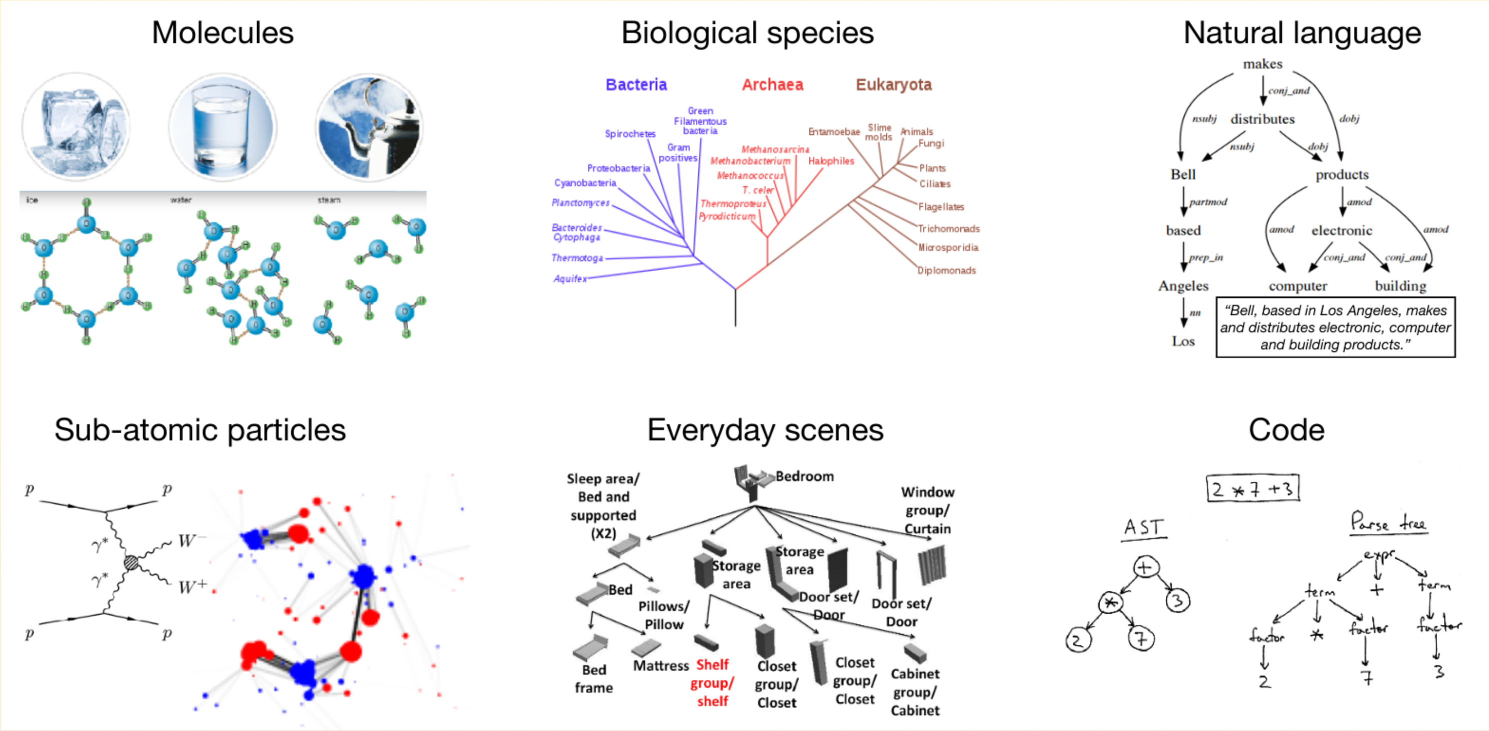
Plug the Physics into the AI

- Physics-aware differential equations solving



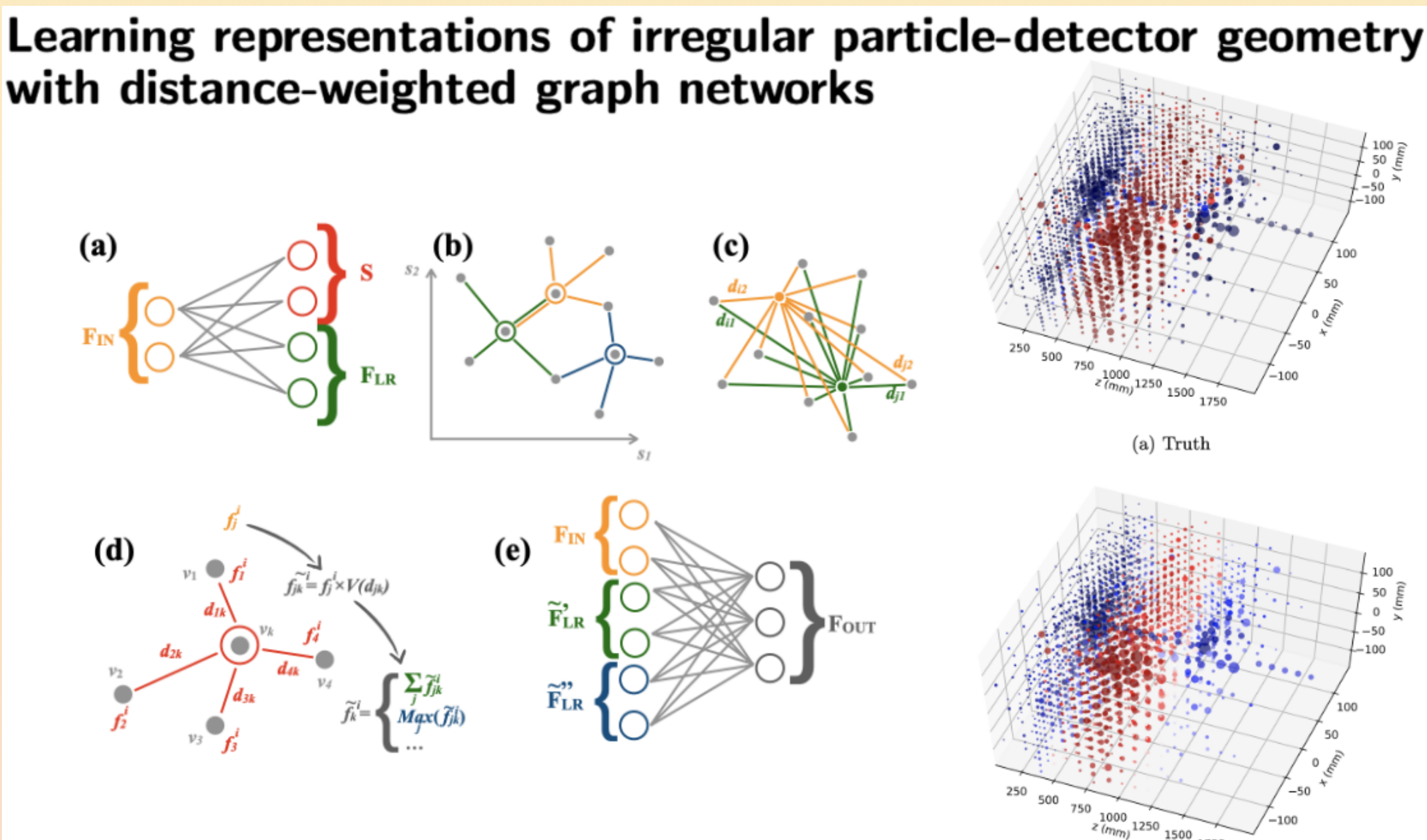
—
Animation from 2202.06988

Graphs Represent Structure

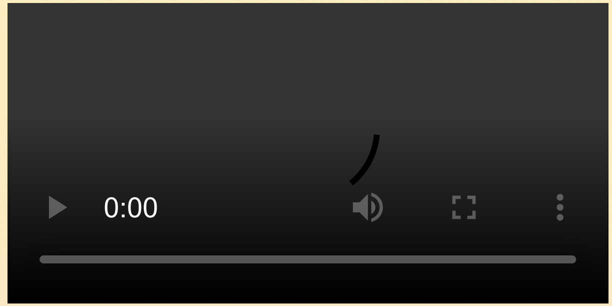


CMS High-granularity calorimeter

- 6 million cells with $\sim 3mm$ spatial resolution, over $600m^2$ of sensors
- Non-projective geometry



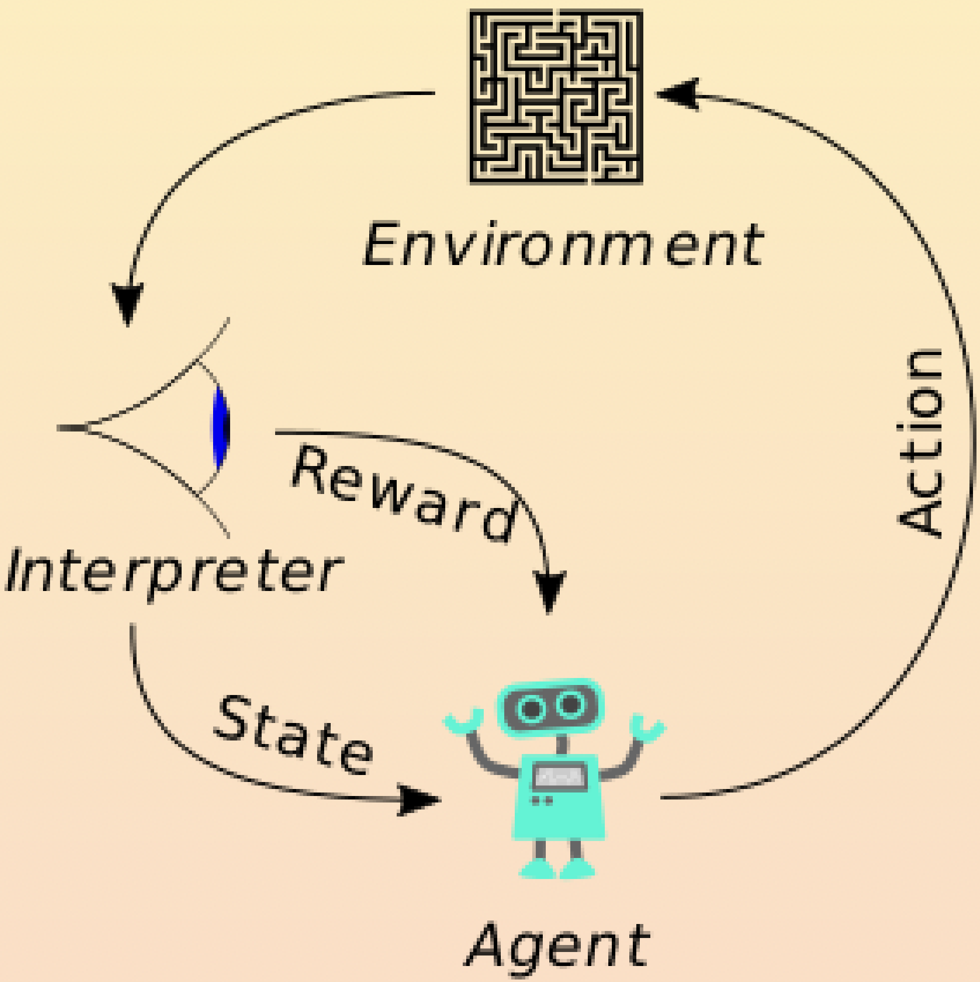
Graphs for water simulation



—

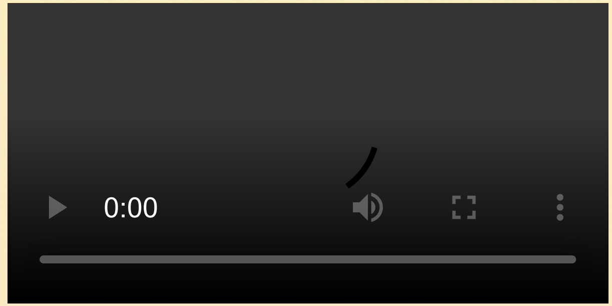
Video by Peter Battaglia's team

Reinforcement learning



—
Image by Megajuice on Wikimedia

From Videogames...



—
Image from YouTube video

...to Physics

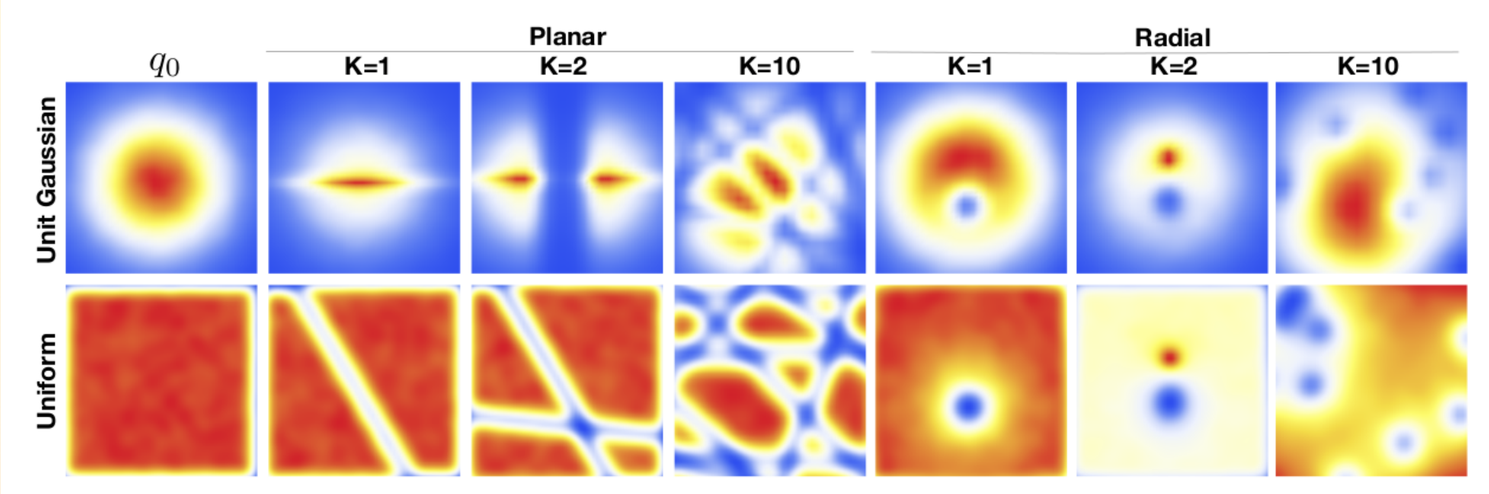
- Reward models consistent with the observed quark properties

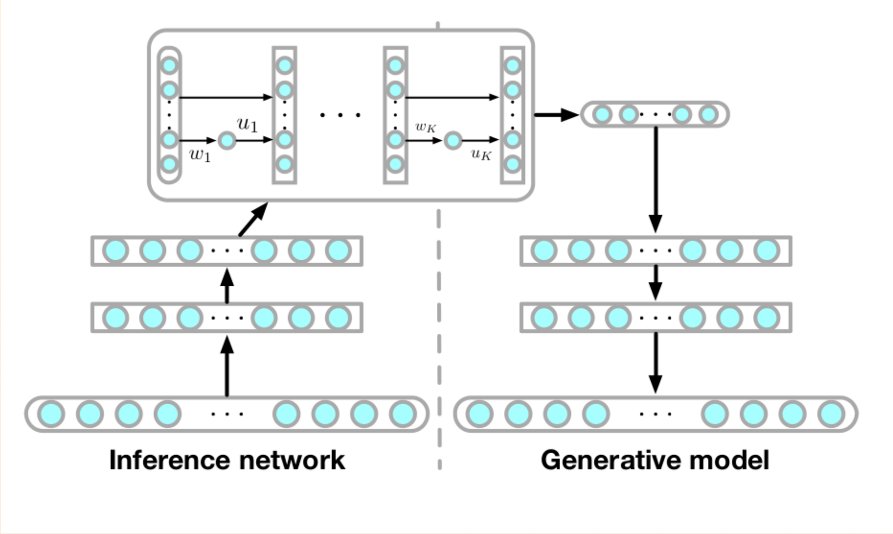
charges	$\mathcal{Q} = \left(\begin{array}{c ccc ccc ccc c c} & Q_1 & Q_2 & Q_3 & u_1 & u_2 & u_3 & d_1 & d_2 & d_3 & H & \phi \\ \hline q & 6 & 4 & 3 & -2 & 2 & 4 & -3 & -1 & -1 & 1 & 1 \end{array} \right)$
$\mathcal{O}(1)$ coeff.	$(a_{ij}) \simeq \begin{pmatrix} -1.975 & 1.284 & -1.219 \\ 1.875 & -1.802 & -0.639 \\ 0.592 & 1.772 & 0.982 \end{pmatrix}$ $(b_{ij}) \simeq \begin{pmatrix} -1.349 & 1.042 & 1.200 \\ 1.632 & 0.830 & -1.758 \\ -1.259 & -1.085 & 1.949 \end{pmatrix}$
VEV, Value	$v_1 \simeq 0.224\,, \qquad \mathcal{V}(\mathcal{Q}) \simeq -0.598$
charges	$\mathcal{Q} = \left(\begin{array}{ccc ccc ccc c c} Q_1 & Q_2 & Q_3 & u_1 & u_2 & u_3 & d_1 & d_2 & d_3 & H & \phi \\ \hline 1 & 2 & 0 & -1 & -3 & 1 & -3 & -5 & -4 & 1 & 1 \end{array} \right)$
$\mathcal{O}(1)$ coeff.	$(a_{ij}) \simeq \begin{pmatrix} -0.601 & 1.996 & 0.537 \\ -0.976 & -1.498 & -1.156 \\ 1.513 & 1.565 & 0.982 \end{pmatrix}$ $(b_{ij}) \simeq \begin{pmatrix} 0.740 & -1.581 & -1.664 \\ -1.199 & -1.383 & 0.542 \\ 0.968 & 0.679 & -1.153 \end{pmatrix}$
VEV, value	$v_1 \simeq 0.158\,, \qquad \mathcal{V}(\mathcal{Q}) \simeq -0.621$

—

Table from 2103.04759

Invertible networks





Correct Detector Noise

- Correct detector observation noise to recover source distribution

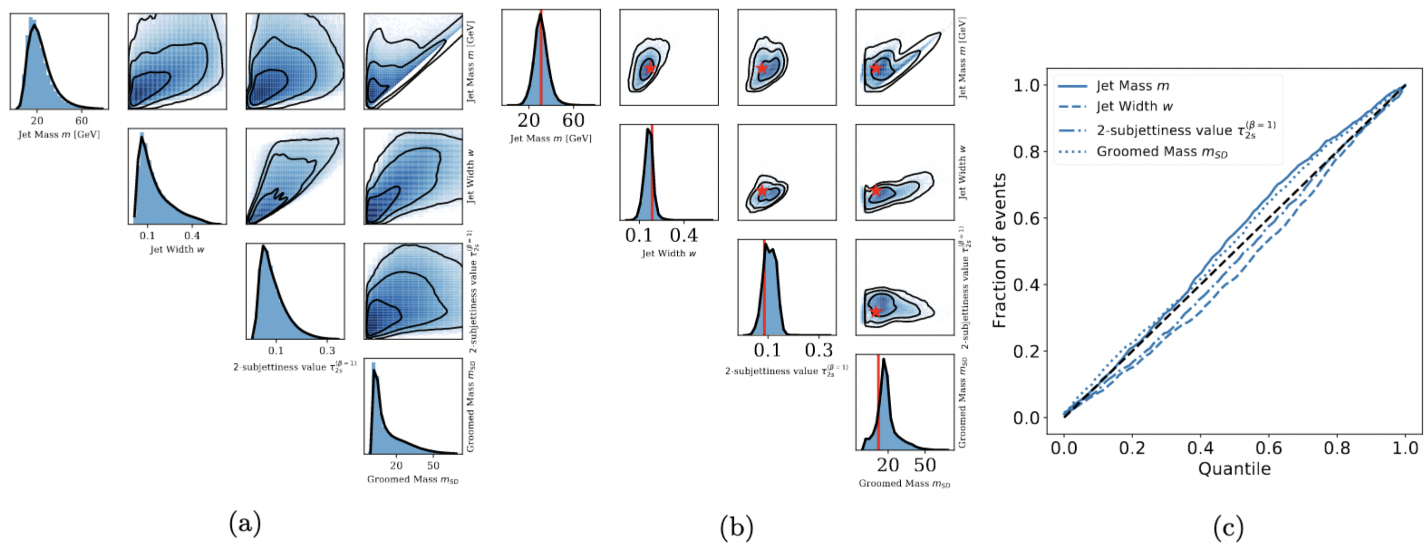
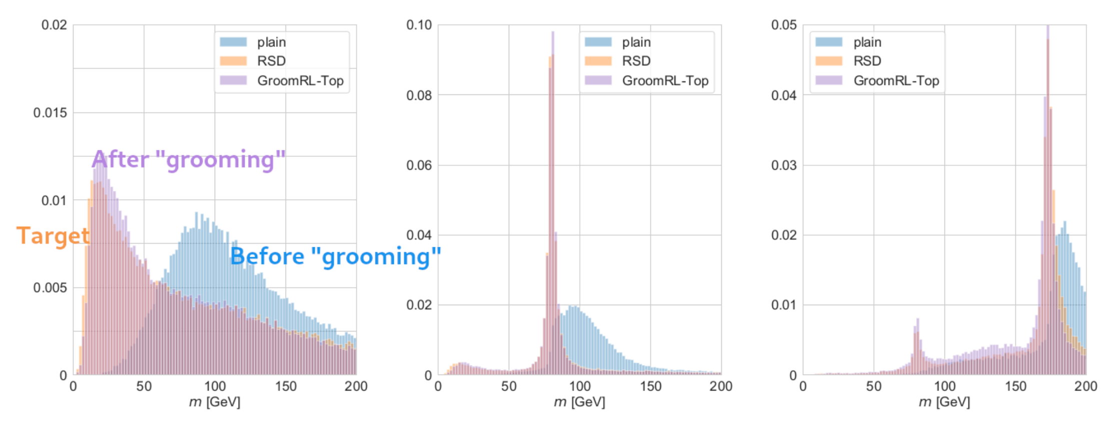


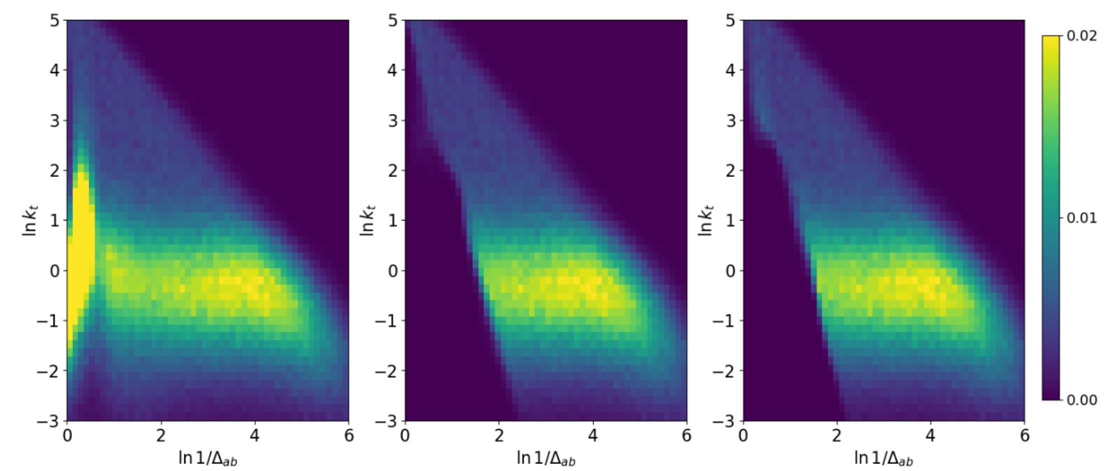
Figure 5: Neural Empirical Bayes for detector correction in collider physics. (a) The source distribution $p(\mathbf{x})$ is shown in blue against the estimated source distribution $q_\theta(\mathbf{x})$ in black. (b) Posterior distribution obtained with rejection sampling, with generating source sample $\mathbf{x}$ indicated in red. (c) Calibration curves for each jet property obtained with rejection sampling on 10000 observations. In (a) and (b), contours represent the 68-95-99.7% levels.

—
Images from 2011.05836

Interpretability



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

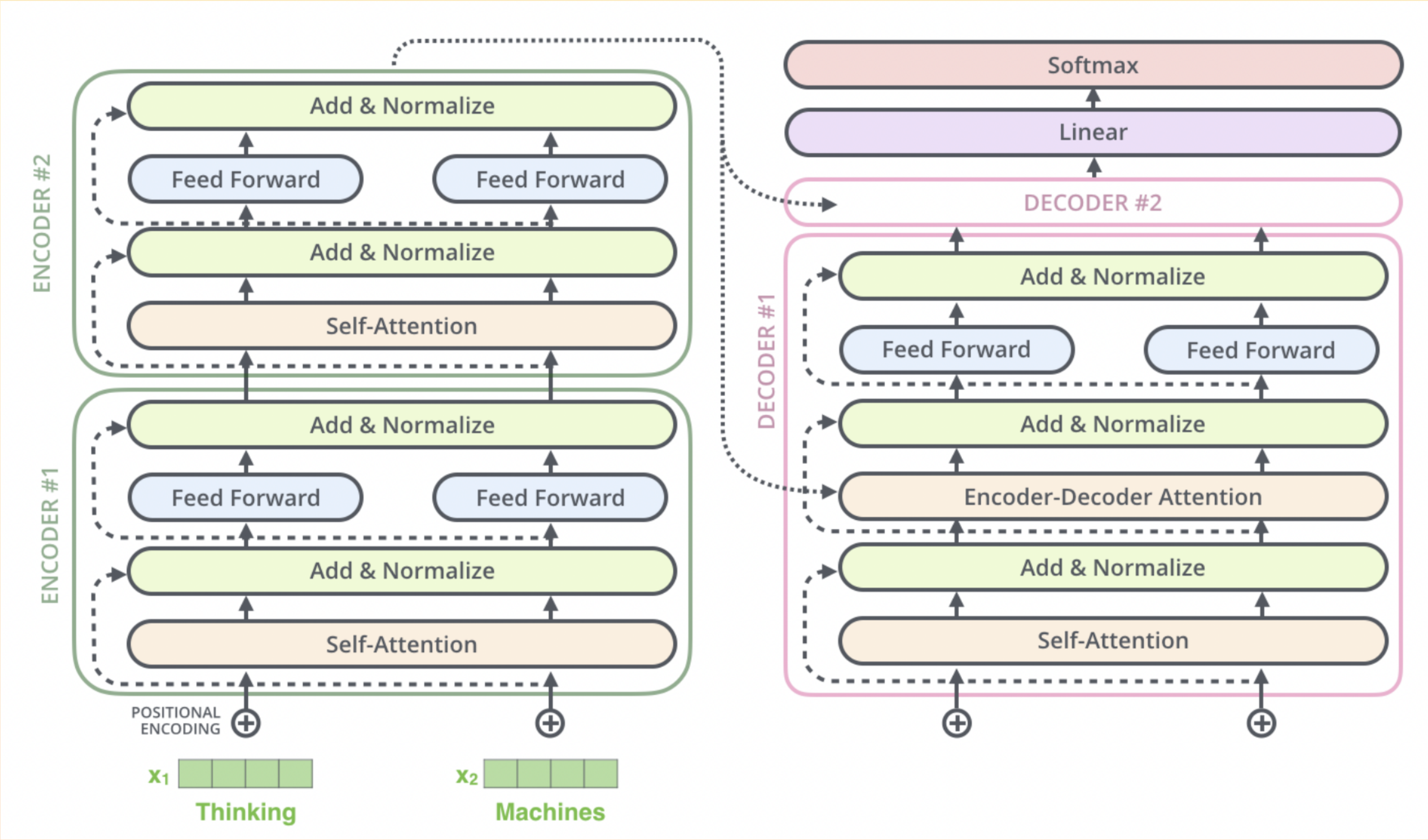


(a) Plain (b) GroomRL-W (c) GroomRL-Top

—
Images from 1903.09644

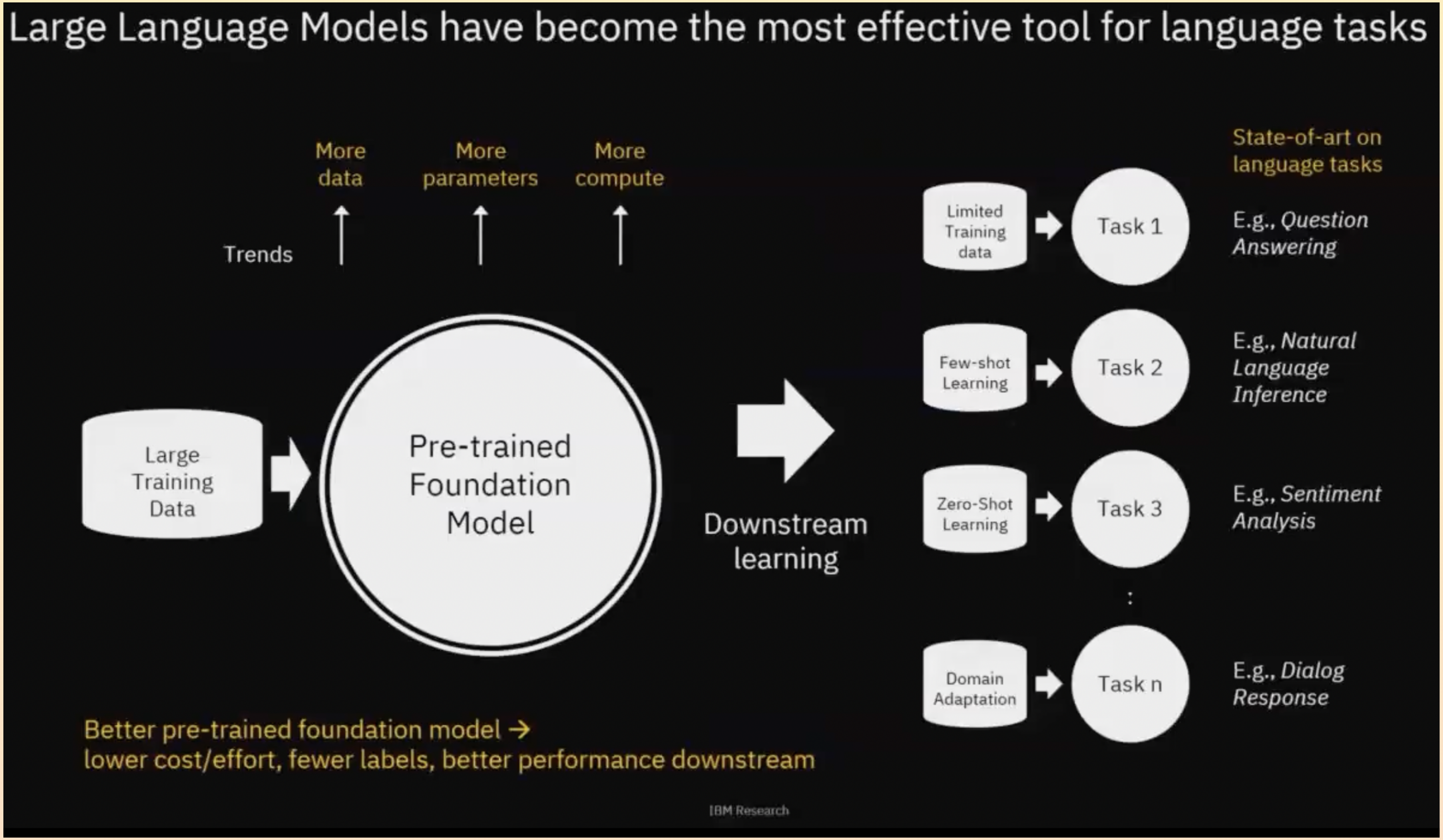
Transformers

- The engine behind GPT3



—
Image by Jay Alammar

Transformers

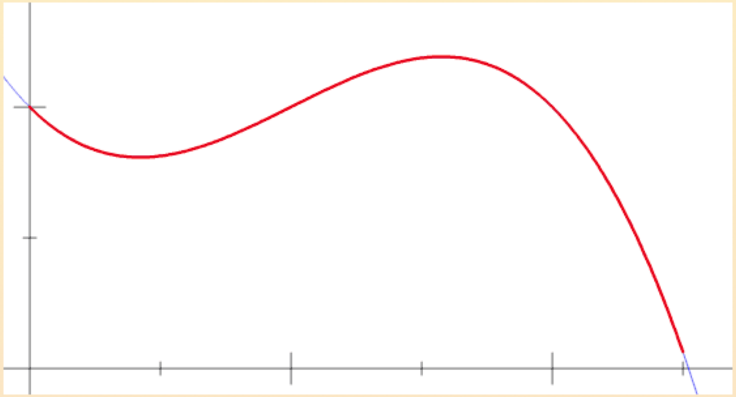


—

Image by John R. Smith (IBM Research)

Translate Problems into Solutions

- Symbolic integration: find the analytic formula for the area of the curve



	Integration (BWD)	ODE (order 1)	ODE (order 2)
Mathematica (30s)	84.0	77.2	61.6
Matlab	65.2	-	-
Maple	67.4	-	-
Beam size 1	98.4	81.2	40.8
Beam size 10	99.6	94.0	73.2
Beam size 50	99.6	97.0	81.0

—
Table by Lample, Charton (2019), image by Julia Inozemtseva

Quantum ML

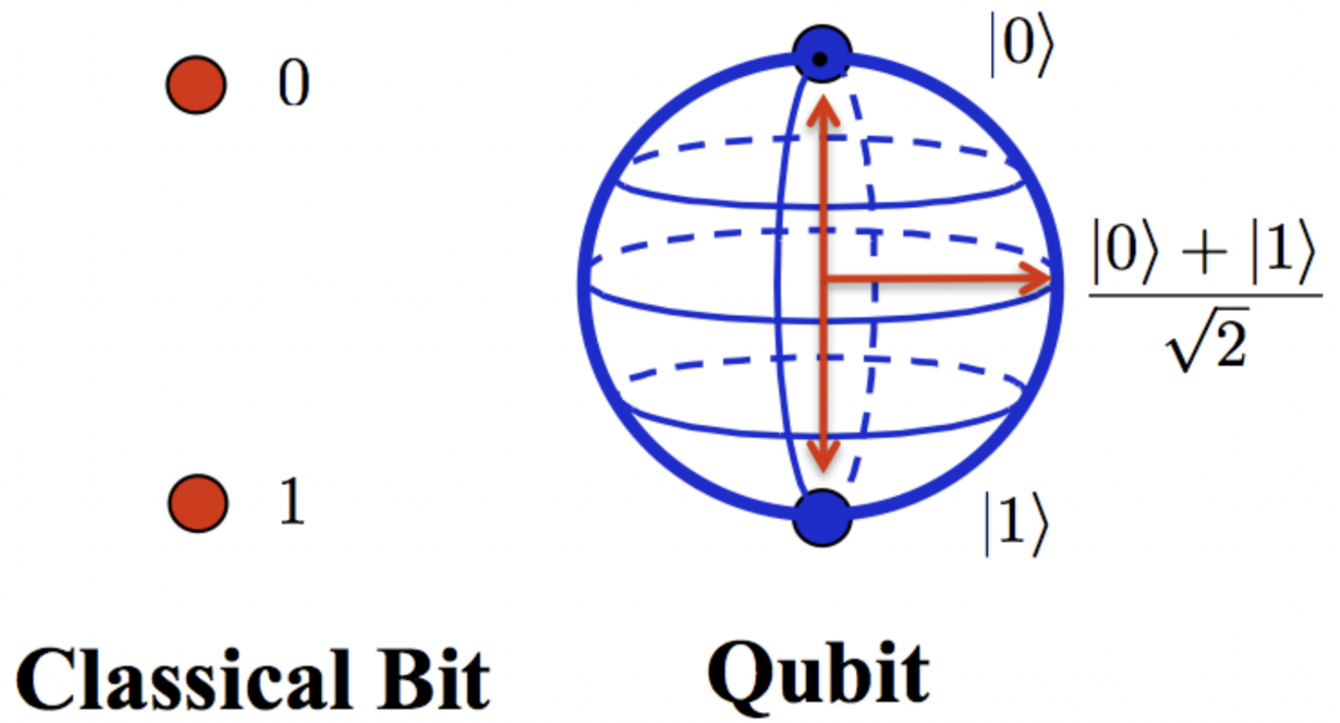
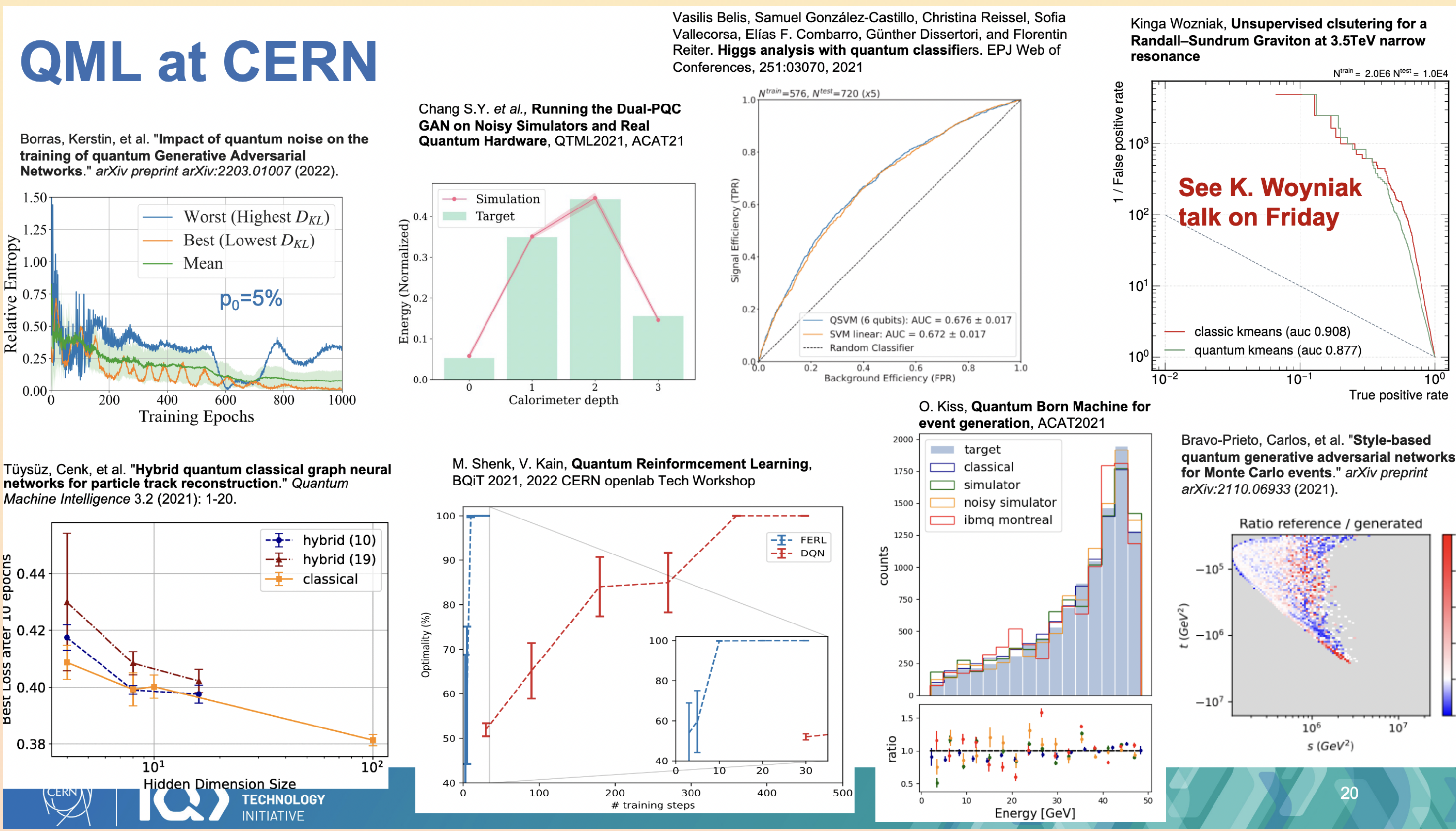


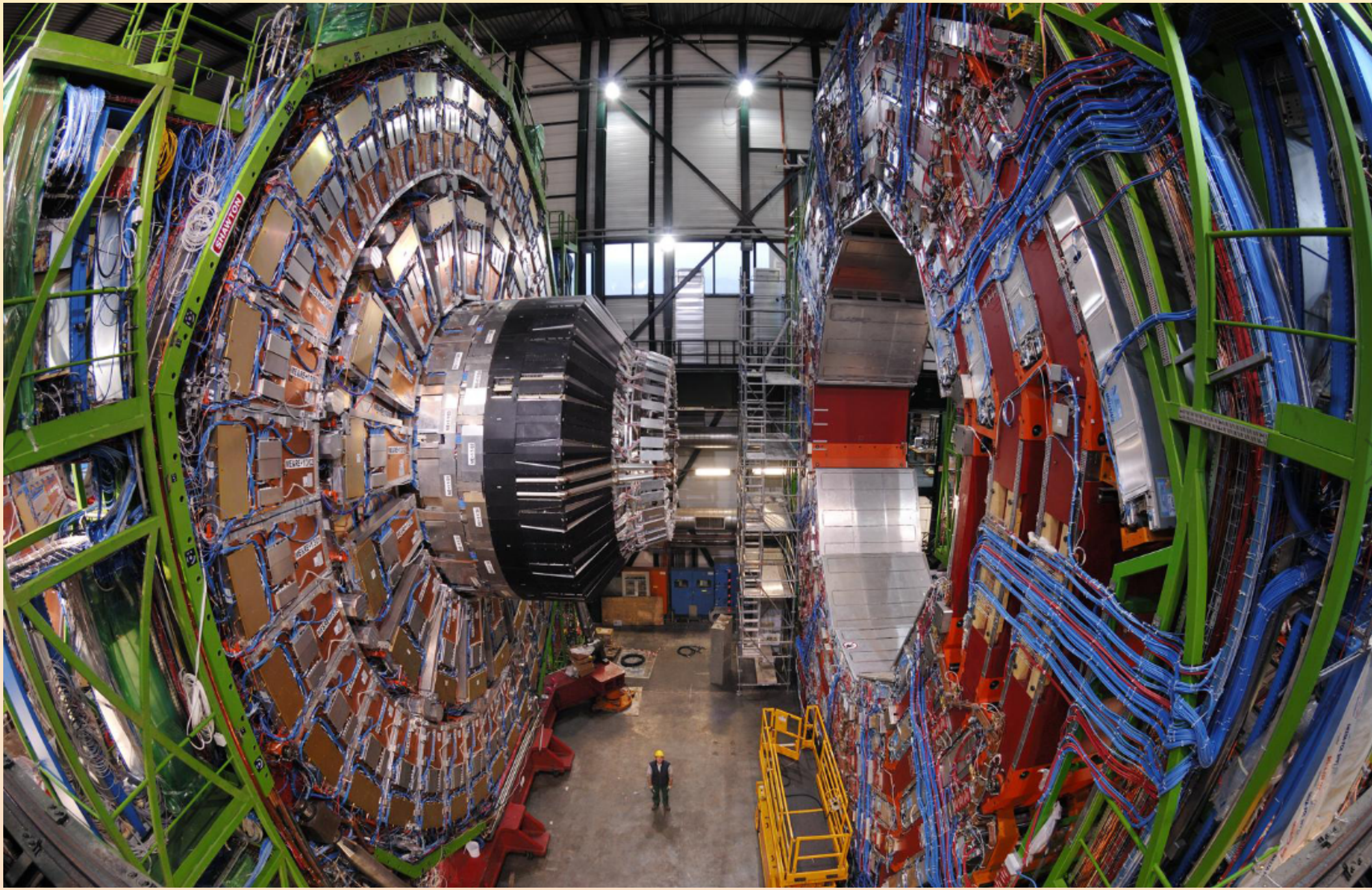
Image credits: https://prateekvjoshi.com/

QML in fundamental research

- So far, mostly as good as classical methods
- Must identify use cases where a quantum approach can be more effective

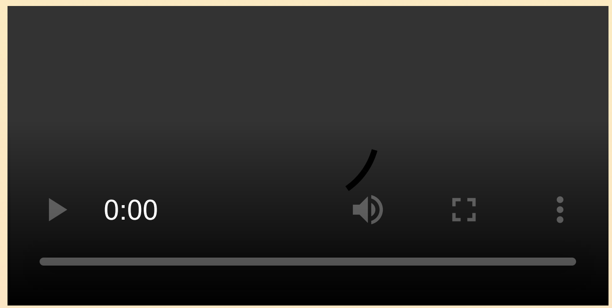


The Detectors of the Present



The Detectors of the Future

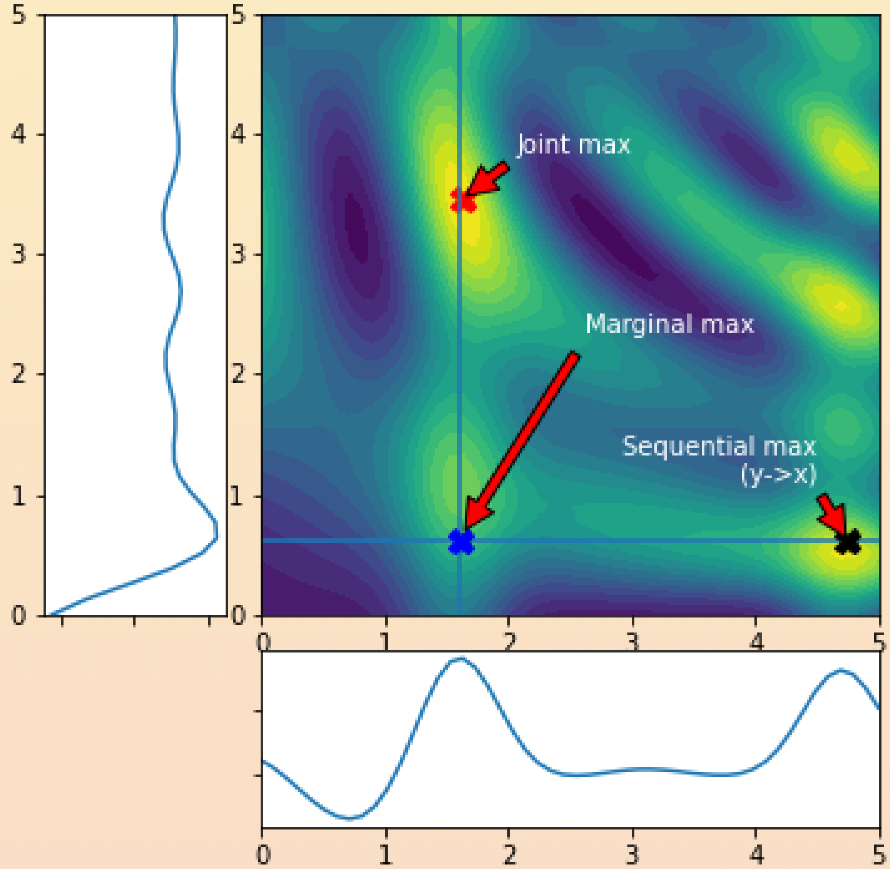
- Optimize/learn by finding the minimum of a function $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$
- Output represents performance on a physics goal or a constraint (e.g. cost)



—
Animation by Alec Radford

The Detectors of the Future

- Joint optimization yields in general different solution than optimization of individual features



—
Illustration (c) P. Vischia, book in preparation (Springer Nature)

The Detectors of the Future

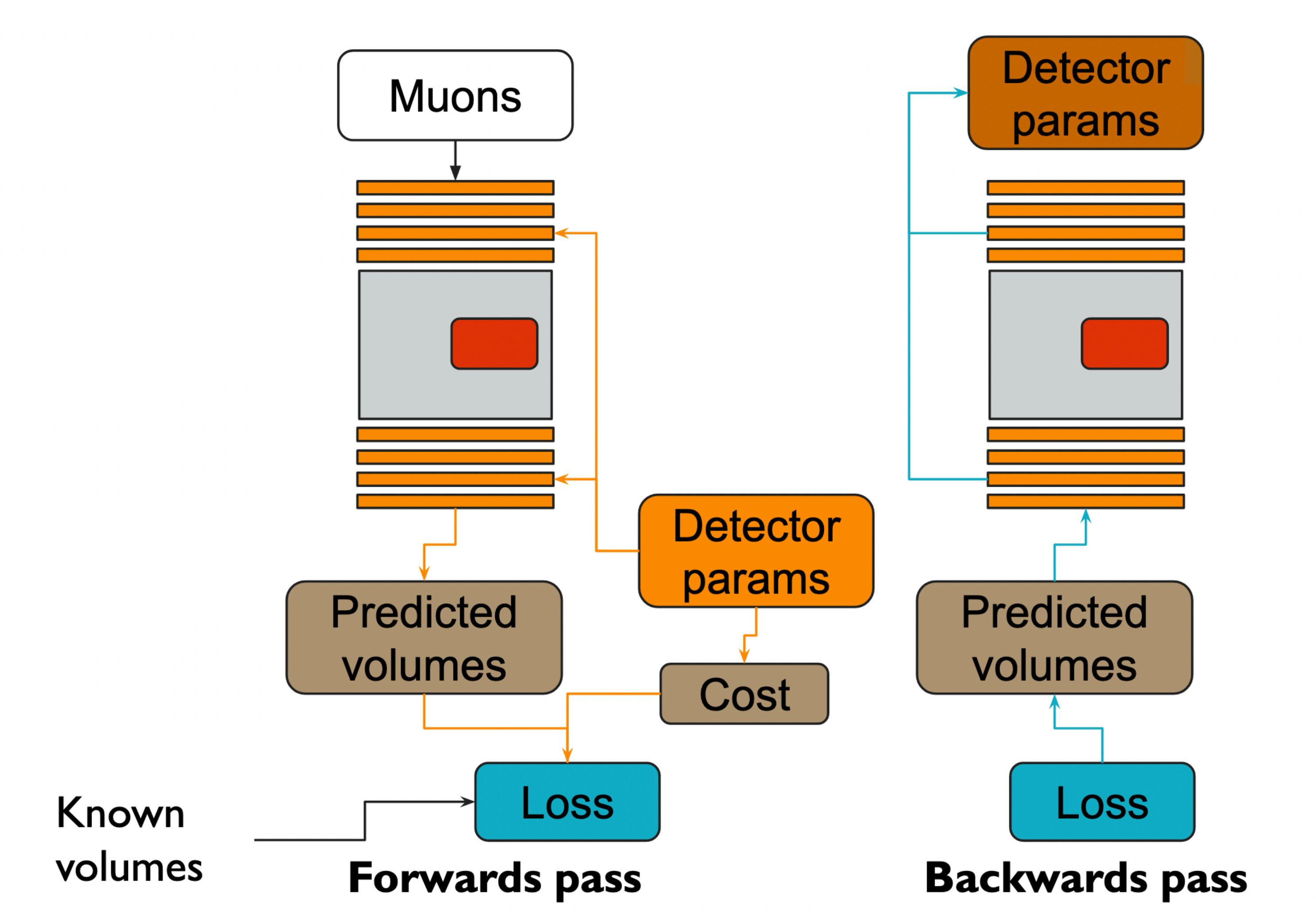


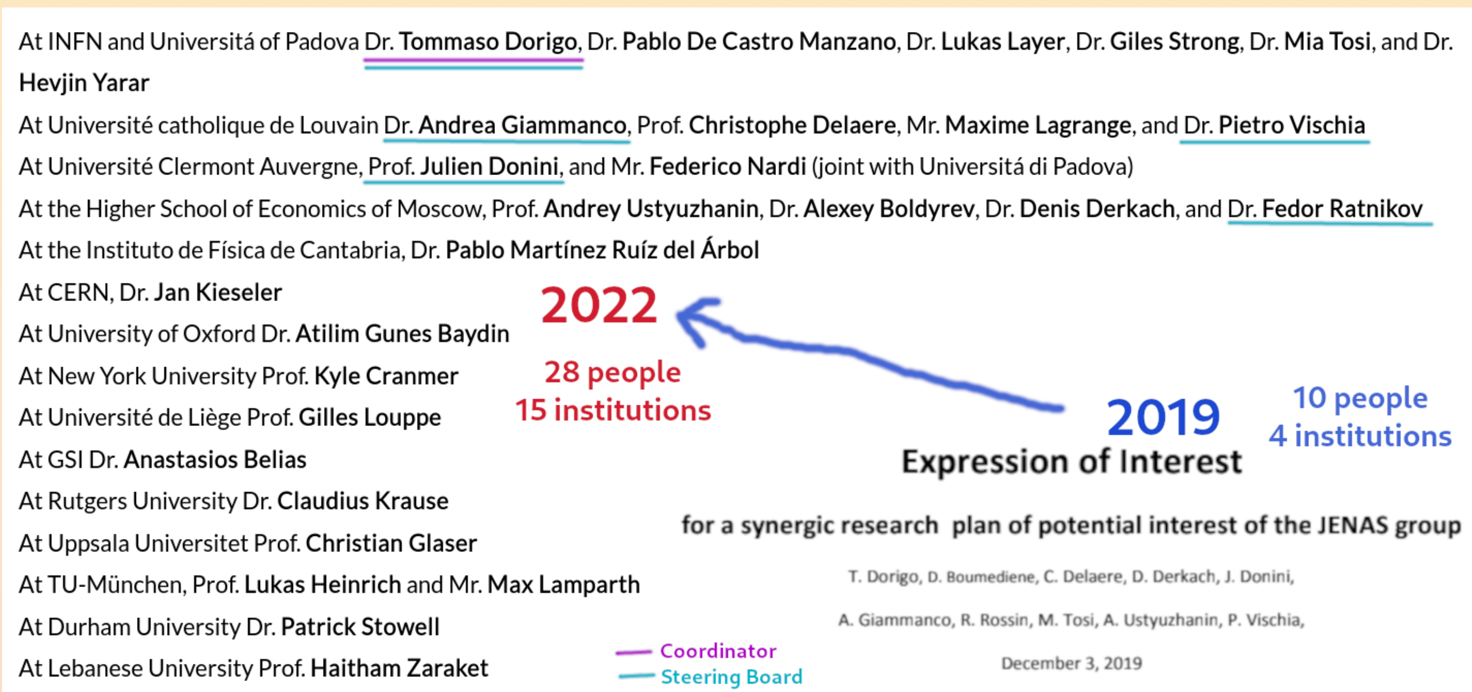
Illustration from the TomOpt paper draft

MODE

- Creation (11.2020) and rapid expansion of the MODE Collaboration

https://mode-collaboration.github.io/

- Joint effort of particle physicists, nuclear physicists, astrophysicists, and computer scientists



At INFN and Universitá of Padova Dr. **Tommaso Dorigo**, Dr. **Pablo De Castro Manzano**, Dr. **Lukas Layer**, Dr. **Giles Strong**, Dr. **Mia Tosi**, and Dr. **Hevjin Yarar**

At Université catholique de Louvain Dr. **Andrea Giammanco**, Prof. **Christophe Delaere**, Mr. **Maxime Lagrange**, and Dr. **Pietro Vischia**

At Université Clermont Auvergne, Prof. **Julien Donini**, and Mr. **Federico Nardi** (joint with Universitá di Padova)

At the Higher School of Economics of Moscow, Prof. **Andrey Ustyuzhanin**, Dr. **Alexey Boldyrev**, Dr. **Denis Derkach**, and Dr. **Fedor Ratnikov**

At the Instituto de Física de Cantabria, Dr. **Pablo Martínez Ruíz del Árbol**

At CERN, Dr. **Jan Kieseler**

At University of Oxford Dr. **Atilim Gunes Baydin**

At New York University Prof. **Kyle Cranmer**

At Université de Liège Prof. **Gilles Louppe**

At GSI Dr. **Anastasios Belias**

At Rutgers University Dr. **Claudius Krause**

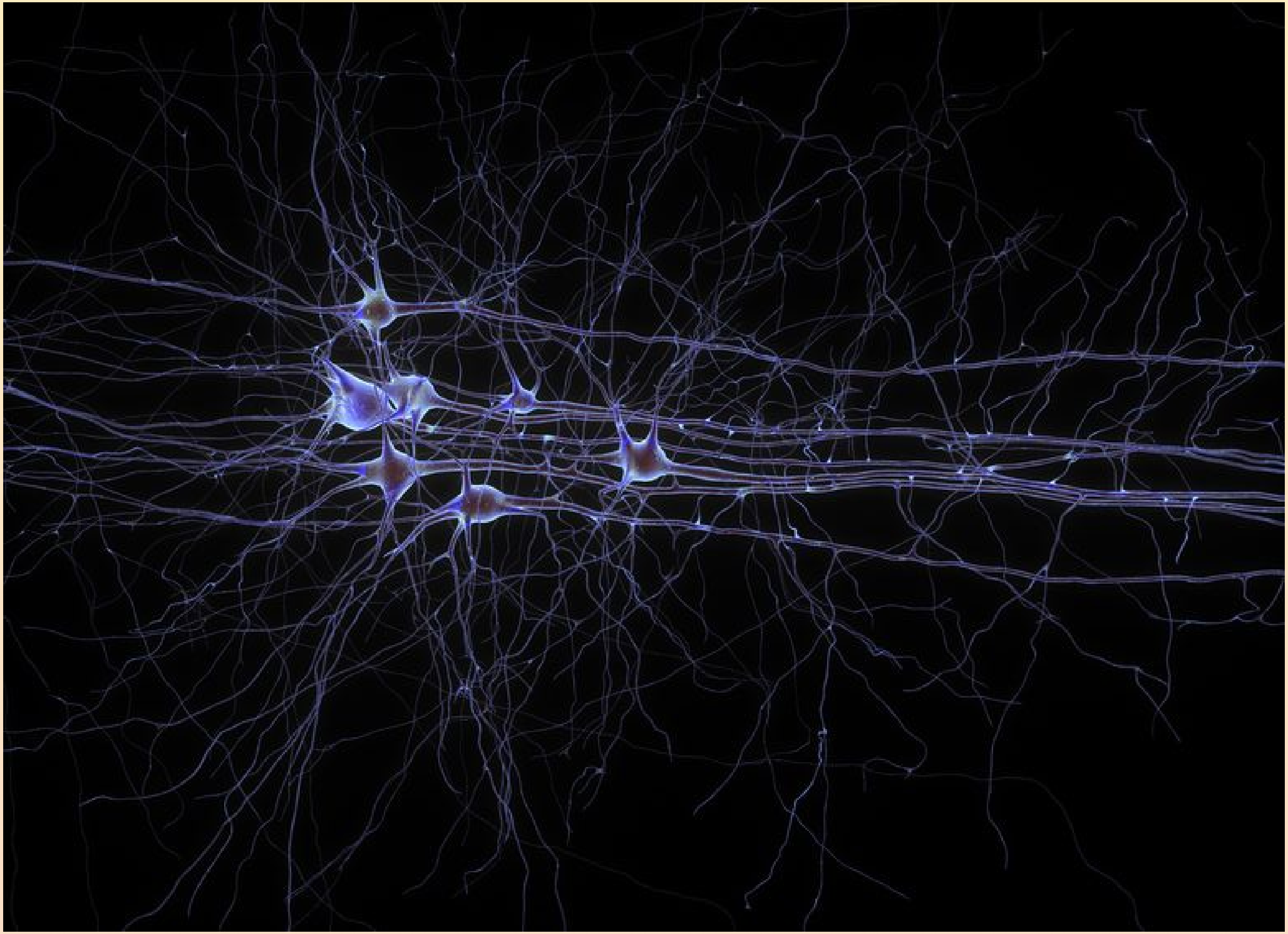
At Uppsala Universitet Prof. **Christian Glaser**

At TU-München, Prof. **Lukas Heinrich** and Mr. **Max Lamparth**

At Durham University Dr. **Patrick Stowell**

At Lebanese University Prof. **Haitham Zaraket**

Artificial Brain



—
Image from pinimg.com

Artificial General Intelligence?

Source of the images lost in the sands of time

Not yet.

My efforts in the next years

- Exploration of high-dimensional spaces via gradient descent, eventually powered by quantum algorithms
- Realistic neurons (spiking networks on neuromorphic circuits)
- Applications to experiment design and to heart diseases
- If you are interested in collaborating, drop me an email!