An Overflight of Statistics in Particle Physics

Data Science & Complexity in Fundamental Physics and the bridge to industry & society

IGFAE, Santiago de Compostela

11-12 June 2026

https://indico.global/event/17473/

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Some thoughts on the use of statistics in High Energy Physics, touching on three questions:

- Foundations of statistics (what is probability?)
- Searching for new phenomena
- Improving accuracy of measurements

Particle Physics – the Big Questions

What are the rules that govern the behaviour of matter at a fundamental level? Current best theory, the Standard Model (SM):

$$\mathcal{L} = \sum_i \overline{\psi}_i \left(i \, \partial\!\!\!/ - m_i - \frac{g m_i H}{2 M_W} \right) \psi_i - \frac{g}{2\sqrt{2}} \sum_i \overline{\Psi}_i \, \gamma^\mu \, (1 - \gamma^5)(T^+ W_\mu^+ + T^- W_\mu^-) \, \Psi_i$$

$$- e \sum_i q_i \, \overline{\psi}_i \, \gamma^\mu \, \psi_i \, A_\mu \; - \frac{g}{2 \cos\theta_W} \sum_i \overline{\psi}_i \, \gamma^\mu (g_V^i - g_A^i \gamma^5) \, \psi_i \, Z_\mu$$

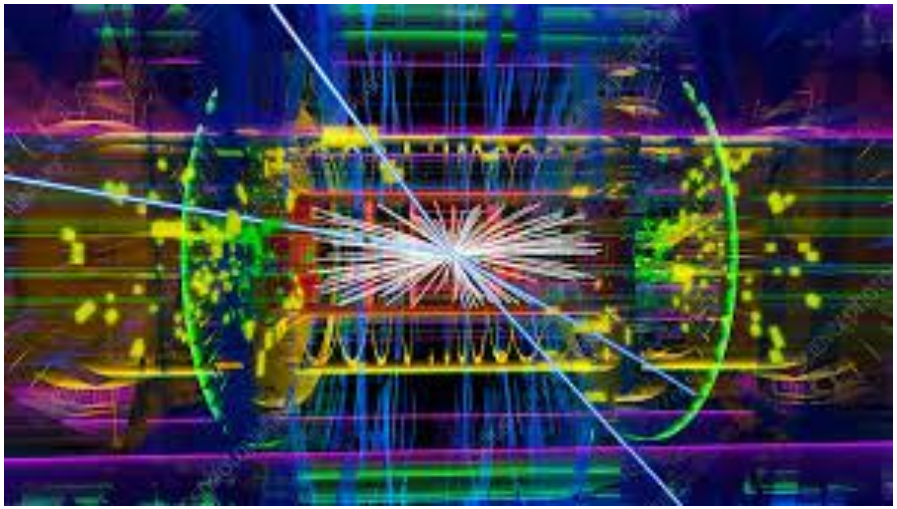
$$+ \sum_q \bar{\psi}_{q,a} (i \gamma^\mu \partial_\mu \delta_{ab} - g_s \gamma^\mu t^C_{ab} \mathcal{A}^C_\mu - m_q \delta_{ab}) \psi_{q,b} - \frac{1}{4} F^A_{\mu\nu} F^{A\,\mu\nu} \quad + \; \ldots$$

leptons, quarks, γ, $W^{\pm}$, Z, g, H

Test SM and other theories e.g. by colliding protons at high energy:

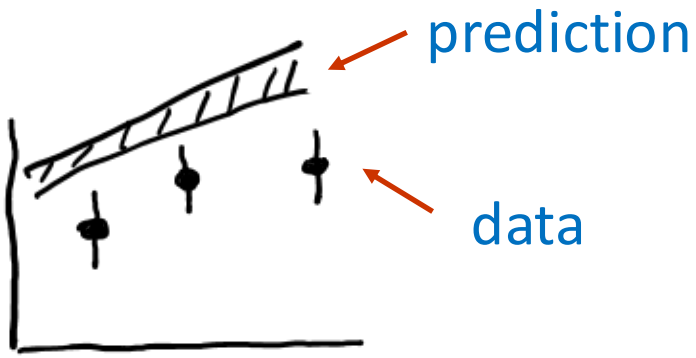
CERN

ATLAS Collaboration



Theory ↔ Statistics ↔ Experiment

Need to compare measurements and theory, both of which have some degree of uncertainty:



To take the uncertainty into account, define probability with a set of mathematical rules (Kolmogorov axioms), and interpret as:

A limiting frequency → frequentist statistics

A degree of belief → Bayesian statistics

Frequentist Statistics

Probability only assigned to outcome of repeatable observation ("data")

A rule that defines the probability for all data outcomes is a hypothesis ("theory")

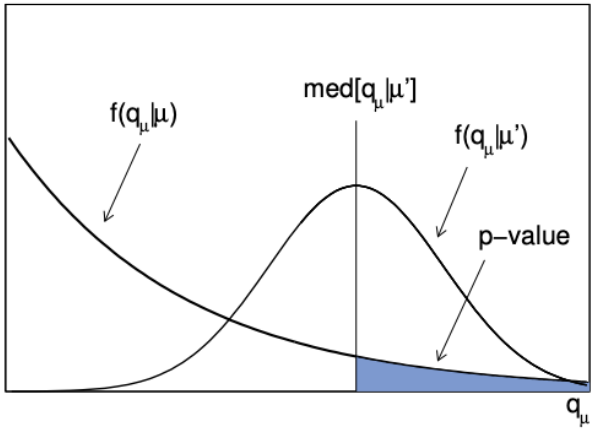
→ cannot talk about probability that a theory is true

So how can we do science?

A frequentist favours a theory if it predicts a high probability for data similar to what we got (other theory predicts lower).

Examples: maximum-likelihood,
least-squares estimation,
confidence intervals,
p-values,...

$p_H = P(\text{data as extreme as observed or more} | H)$



Bayesian Statistics

Extends interpretation of probability to include degree of belief, use this for hypotheses (theories).

Need to start by stating degree of belief in the hypothesis before seeing the data (the prior).

Probability of data given the theory is the likelihood.

Bayes theorem updates the prior to give the posterior probability:

posterior likelihood prior

$$P(\text{theory}|\text{data}) = \frac{P(\text{data}|\text{theory})P(\text{theory})}{P(\text{data})}$$

Requires prior probability (subjective).

Thomas Bayes (1701-1761)

Particle Physics vs Astro

Particle Physics: mainly frequentist

Number of particle collisions very large

Astrophysics/cosmology: mainly Bayesian

One Universe

In practice, the conclusions point towards a similar direction.

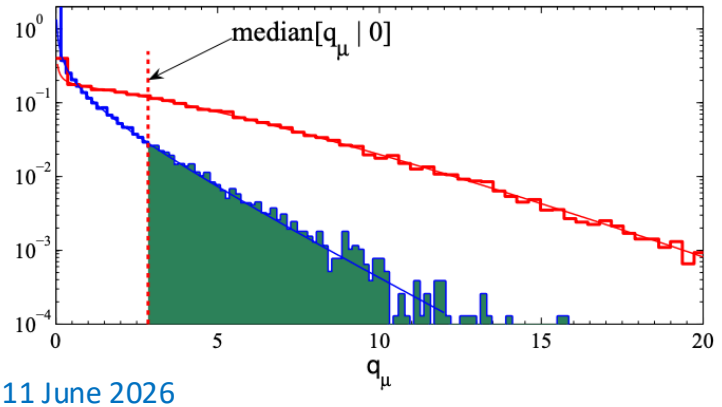
But the computational steps involved are very different:

Bayesians marginalise over unwanted parameters (MCMC sampling of parameter space); compute marginal likelihoods

$$p(\mu|\boldsymbol{x}) = \int p(\mu, \boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta}$$

$$P(\boldsymbol{x}|H) = \int P(\boldsymbol{x}|H, \boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

Frequentists compute sampling distributions of estimators and test statistics (asymptotic methods, Monte Carlo sampling of data space)



Conclusion on Frequentist vs Bayesian

Frequentist vs Bayesian debate going on for more than 2 centuries.

Frequentist:

Cannot talk about Prob(theory),

surprisingly subtle issues with confidence limits, *p*-values

Bayesian:

Can talk about Prob(theory), requires prior probabilities.

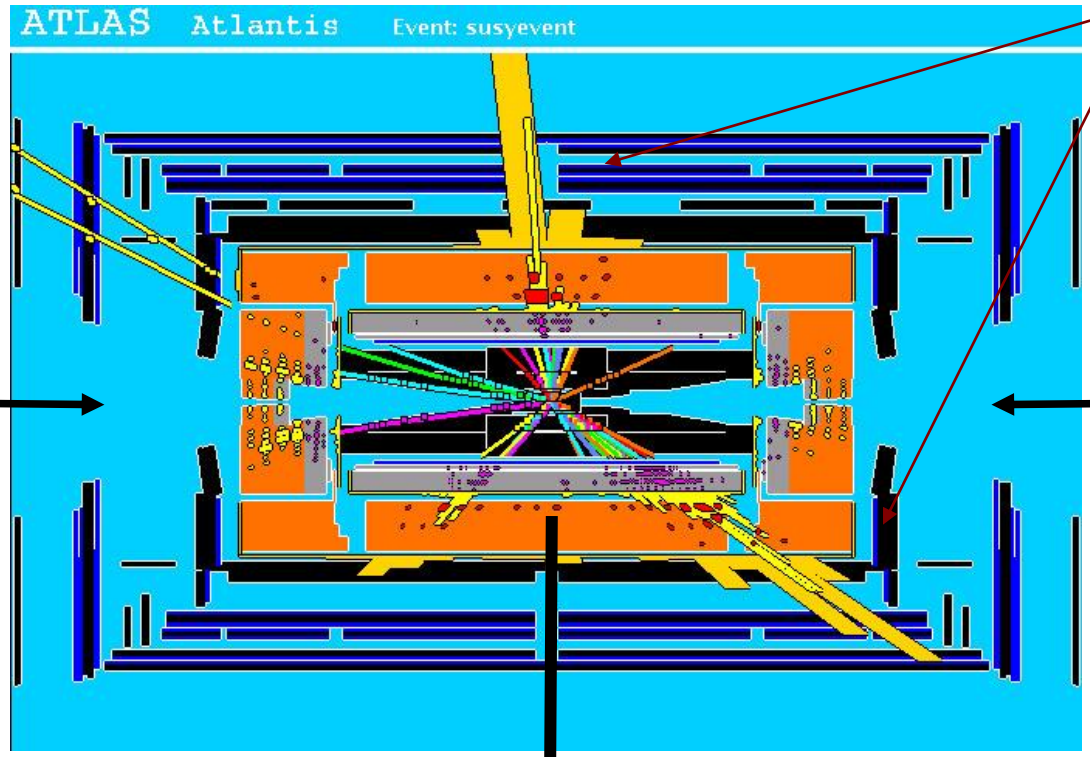
Both approaches have their own computational challenges.

The modern view is that each approach answers different, relevant questions, and one can use both.

In practice the methods are chosen by sociology/politics, but I still hold out hope of a more mixed approach.

Statistics for Discovery of a Signal

An important goal of Particle Physics is to see if some extensions to the Standard Model are perhaps correct, e.g., Supersymmetry. If it is, we expect proton-proton collision events like this (simulated) one:



high p_T muons

high p_T jets of hadrons

p

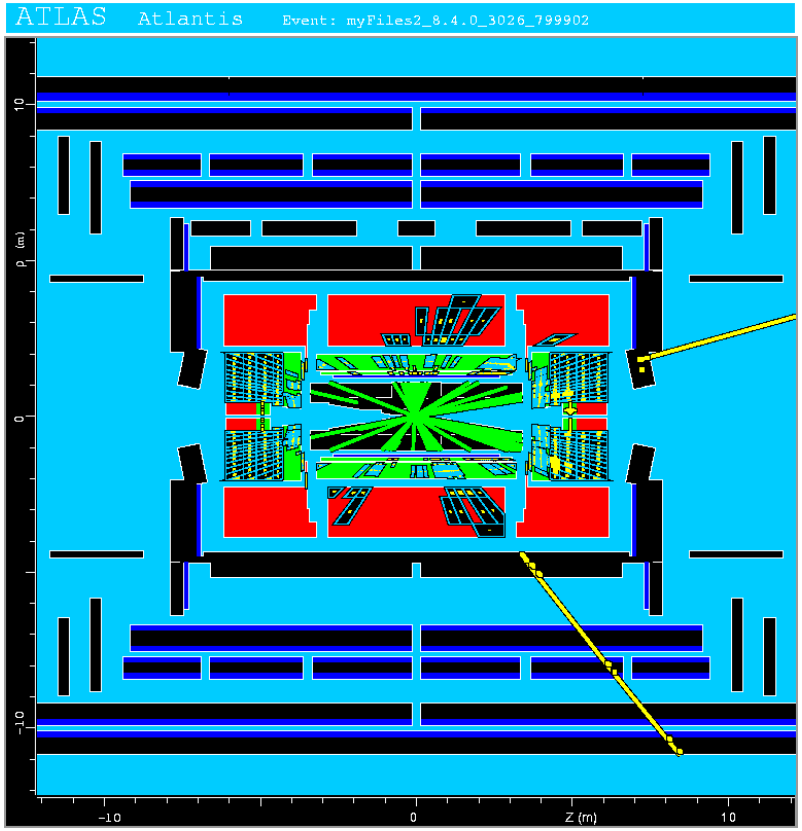
p

missing transverse energy

Background Events

But there are collision events produced by known Standard Model processes that can easily mimic the signal, e.g.,

Simulated top-anti-top event, with features similar to those expected from SUSY signal.



For each event, measure a "feature vector" $\boldsymbol{x} = (x_1, \ldots, x_n)$:

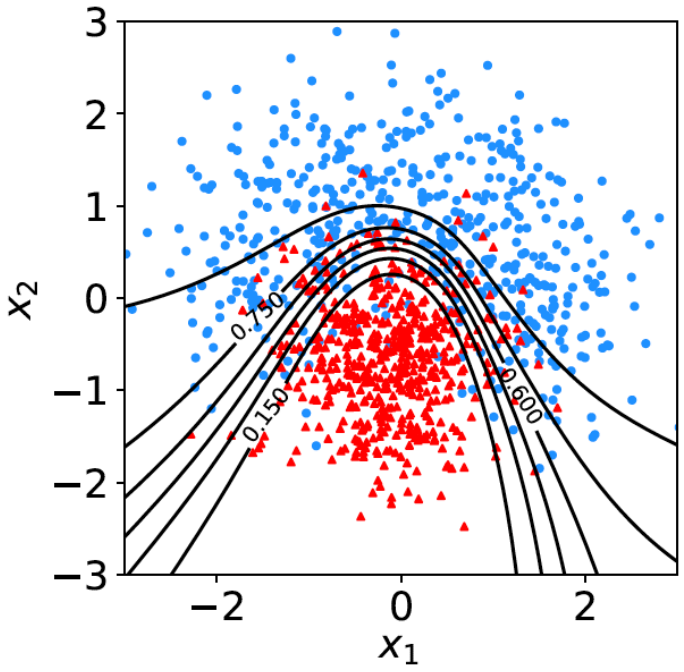
x_1 = energy of jet

x_2 = p_T of muon

x_3 = missing energy

⋮

Decision function / test statistic

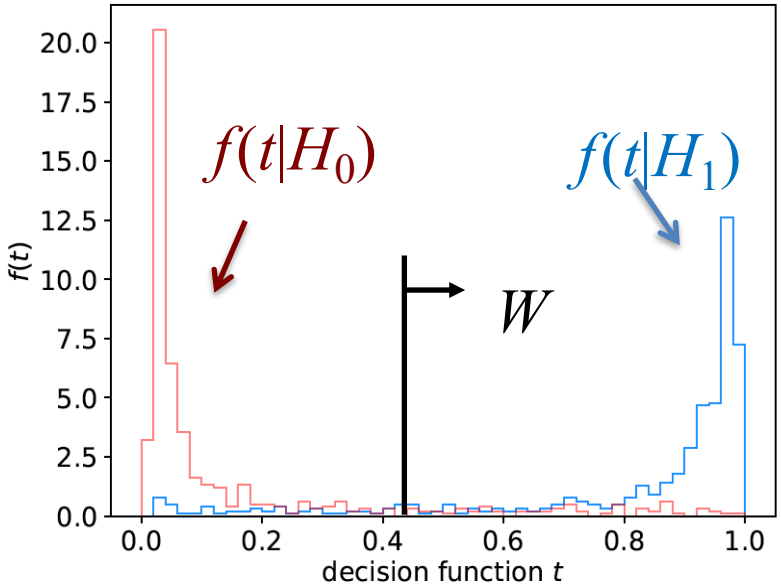


Distribution of feature vectors expected for signal (blue) and background (red) events.

To obtain a decision boundary, define a scalar test statistic:

$$t(x_1, \ldots, x_n) = t_c$$

defines decision boundary



Optimal Test Statistic

Neyman-Pearson lemma: optimal test statistic is the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}$$

But we usually don't have explicit formulae for the pdfs $f(\boldsymbol{x}|\mathrm{s}), f(\boldsymbol{x}|\mathrm{b})$, so for a given $\boldsymbol{x}$ we can't evaluate the statistic.

Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data:

generate $\boldsymbol{x} \sim f(\boldsymbol{x}|\mathrm{s})$ $\quad \rightarrow \quad$ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

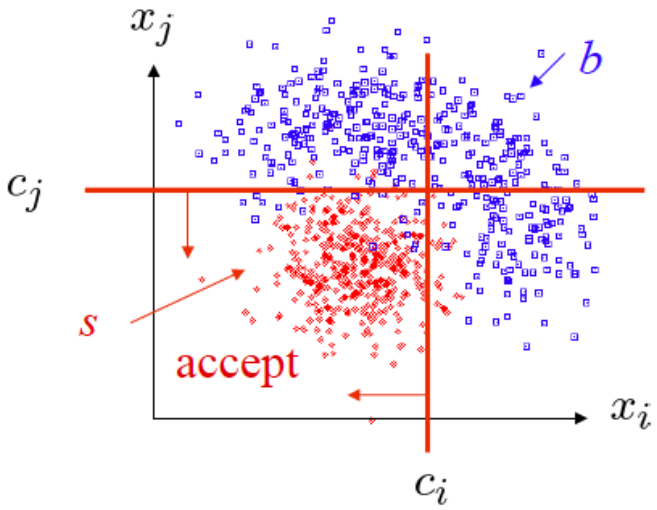
generate $\boldsymbol{x} \sim f(\boldsymbol{x}|\mathrm{b})$ $\quad \rightarrow \quad$ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

This gives samples of "training data" with events of known type.

- Can be expensive (1 fully simulated LHC event ~ 1 CPU minute).

Particle Physics meets Machine Learning

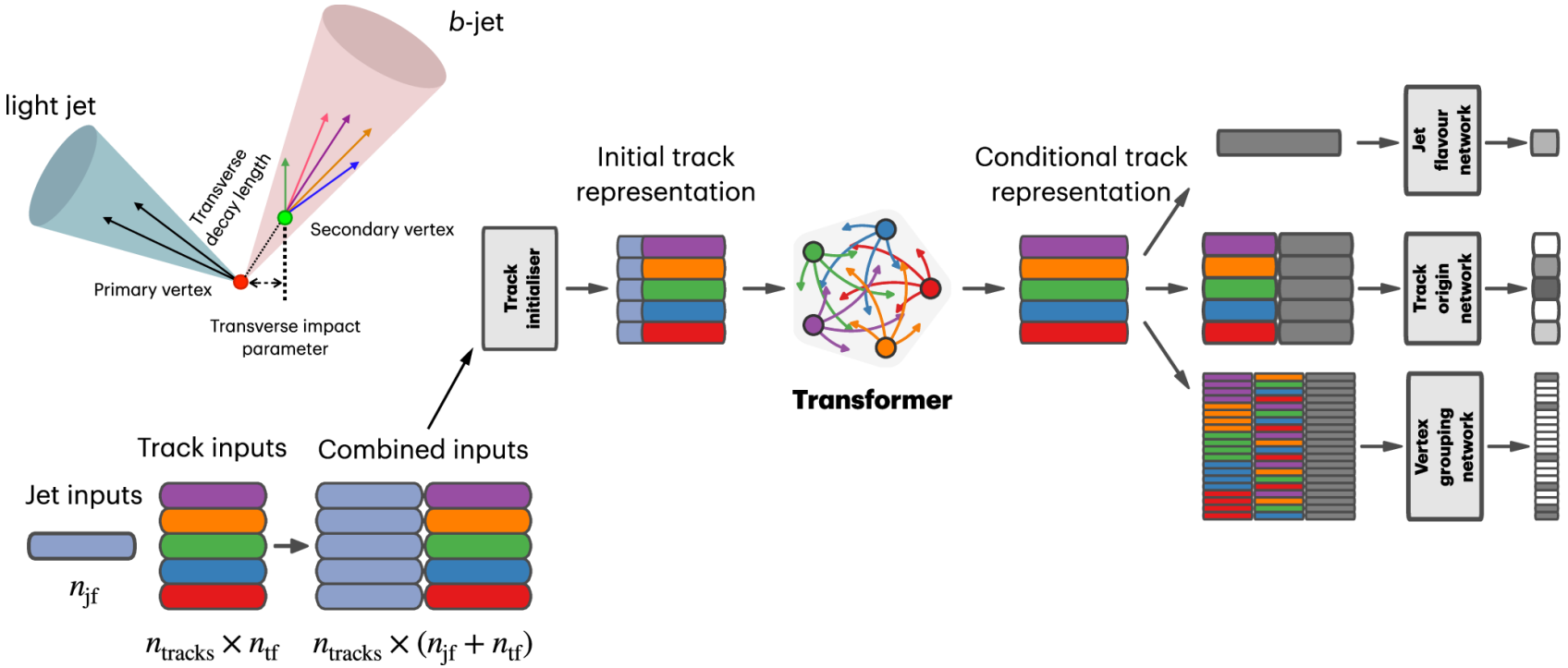
Particle Physics has had a long history of "cut-based" analyses:



Eventually adopted methods from Machine Learning:

1990s	Fisher Discriminants, Neural Networks
Early 2000s	Boosted Decision Trees, Support Vector Machines
From mid 2010s	Deep Neural Nets, modern architectures,...

Transformer-based NN for b-jet tagging



Track+jet becomes "token" in transformer-based DNN architecture.

"Attention" quantifies how these tracks relate to others in same jet.

Resulting ability to identify a jet as originating from a b-quark with background rejection improved by factor of 1.8 to 3.5.

Systematic uncertainties / modeling

Setup: measure data $\boldsymbol{y}$, want to estimate parameter μ.

Problem: $P(\boldsymbol{y}|\mu)$ in general approximate $\rightarrow$ systematic error in estimate of μ.

Solution: fix the model, usual by including further adjustable (nuisance) parameter(s) θ: $P(\boldsymbol{y}|\mu) \rightarrow P(\boldsymbol{y}|\mu,\theta)$

This inflates the statistical uncertainty on estimate of μ unless then nuisance parameters are themselves well constrained.

Usual procedure:

$$P(\boldsymbol{y}|\mu) \rightarrow P(\boldsymbol{y}, u|\mu, \theta) = P(\boldsymbol{y}|\mu, \theta)P(u|\theta)$$

Original likelihood

New likelihood

Auxiliary measurement u, often $u \sim \text{Gauss}(\theta, \sigma_u)$

”Errors on errors”

The uncertainties on estimated systematic errors (“errors on errors”) can in general play an important role in many analyses, see:

G. Cowan, Eur. Phys. J. C (2019) 79:133; arXiv:1809.05778

G. Cowan, , EPJ Web of Conferences 258, 09002 (2022); arXiv:2107.02652

E. Canonero, A. Brazzale and G. Cowan, Eur. Phys. J. C (2023) 83:1100; arXiv:2304.10574

E. Canonero, G. Cowan, Eur. Phys. J. C (2025) 85: 156; arXiv:2407.05322

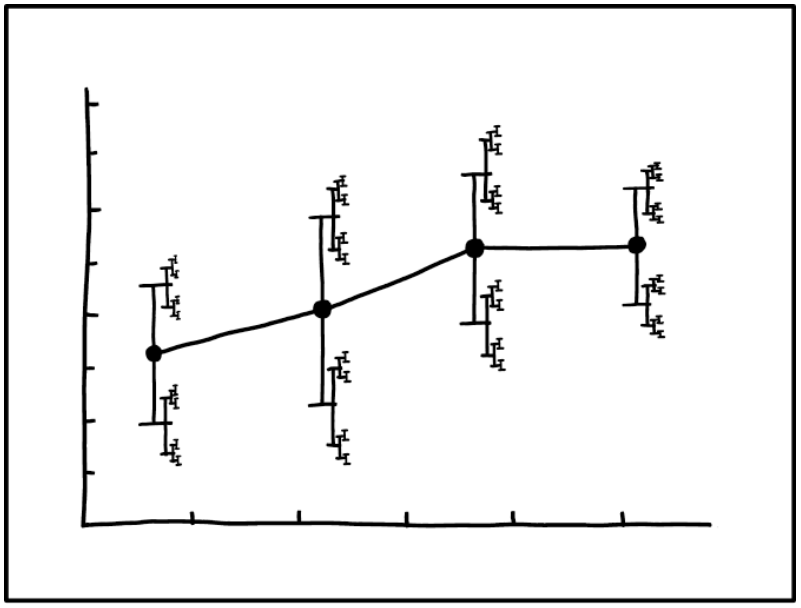
It turns out that models that use errors on errors have qualitatively new, interesting, desirable features:

Sensitivity to outliers reduced.

Confidence intervals sensitive to goodness of fit.

Effect on goodness of fit, p-values, significance.

https://xkcd.com/2110/



Including the "errors on errors" in a fit

Let $\sigma_u \to s_u$ become an *estimate* of the systematic uncertainty,

σ_u becomes an adjustable parameter,

Analyst must supply ε = relative "error on the error"

The usual log-likelihood based on $u \sim \text{Gauss}(\theta, \sigma_u)$

$$\ln L(\mu, \theta) = \ln P(\boldsymbol{y}|\mu, \theta) - \frac{1}{2}\frac{(u-\theta)^2}{\sigma_u^2}$$

becomes

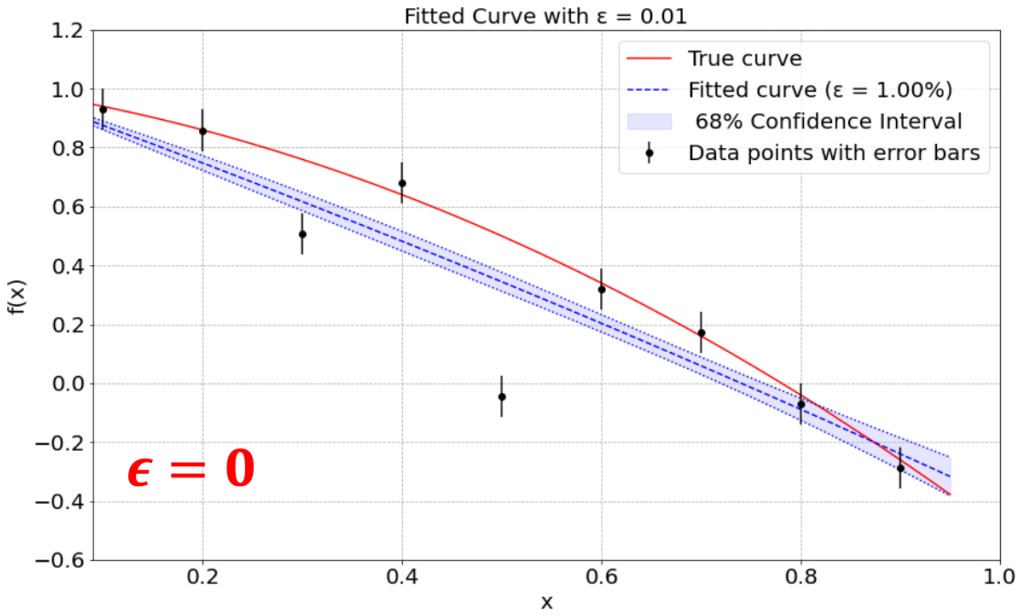
$$\ln L(\mu, \theta) = \ln P(\boldsymbol{y}|\mu, \theta) - \frac{1}{2}\left(1 + \frac{1}{2\varepsilon^2}\right) \ln\left[1 + 2\varepsilon^2 \frac{(u-\theta)^2}{s_u^2}\right]$$

Fitting with outliers (e.g., parton fits)

Fitting of a curve: compatible measurements

ROYAL HOLLOWAY UNIVERSITY OF LONDON

- Fit of a quadratic function with ***two outliers***



$$y_i \sim f(x_i) + \theta_i$$

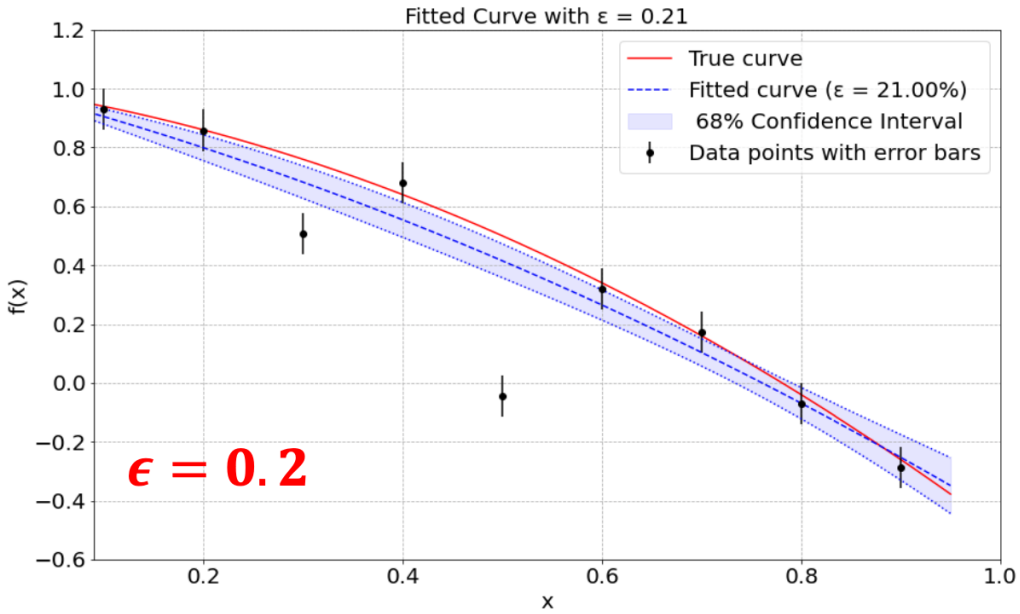
Params of interest

$$f(x_i) = ax_i^2 + bx + c$$

Fitting of a curve: compatible measurements

ROYAL HOLLOWAY UNIVERSITY OF LONDON

- Fit of a quadratic function with ***two outliers***



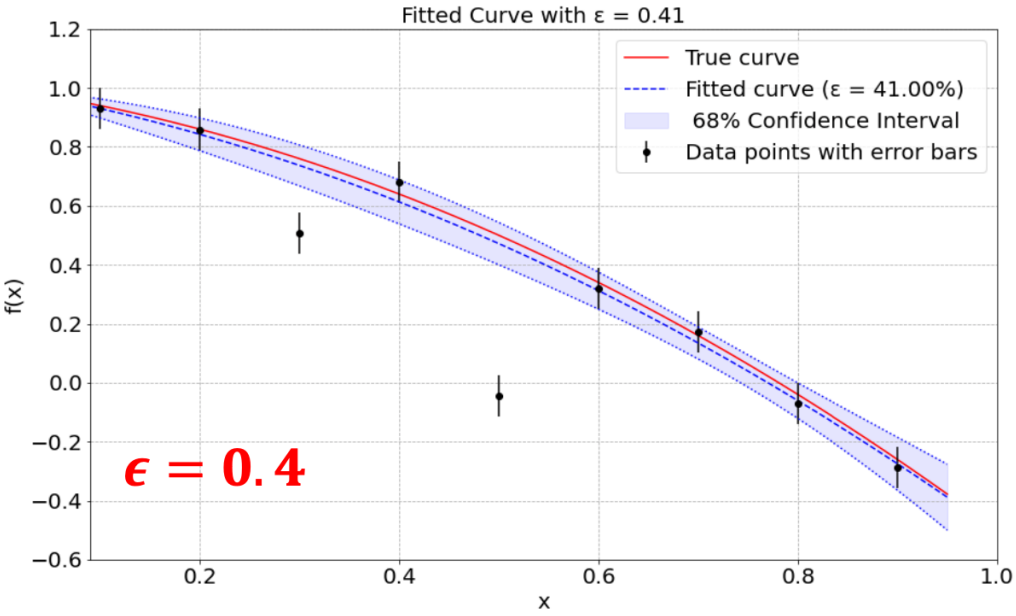
$$y_i \sim f(x_i) + \theta_i$$

Params of interest

$$f(x_i) = ax_i^2 + bx + c$$

Fitting of a curve: compatible measurements

- Fit of a quadratic function with *two outliers*



$$y_i \sim f(x_i) + \theta_i$$

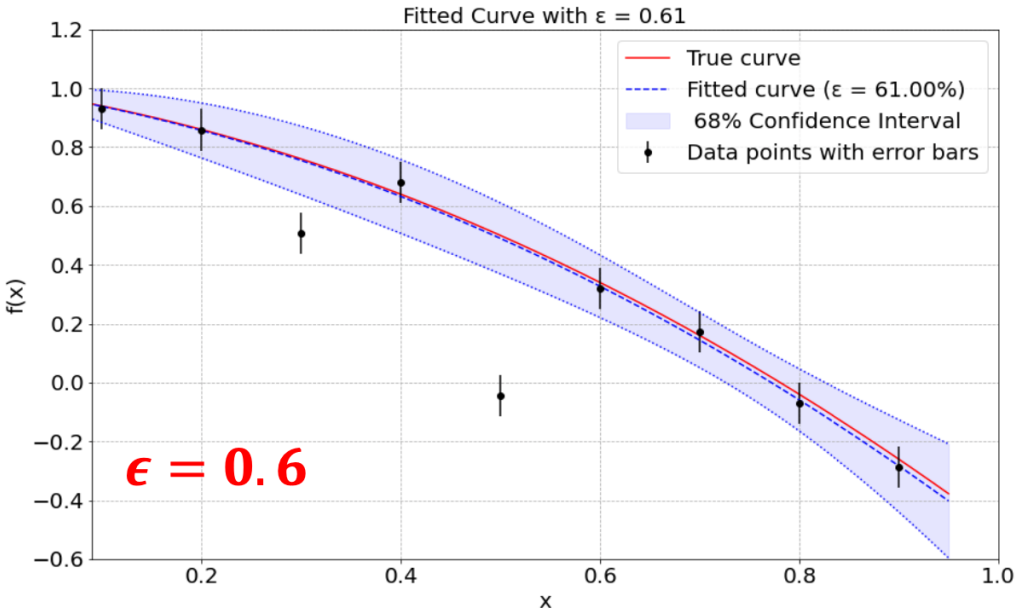
Params of interest

$$f(x_i) = ax_i^2 + bx + c$$

Fitting of a curve: compatible measurements

ROYAL HOLLOWAY UNIVERSITY OF LONDON

- Fit of a quadratic function with ***two outliers***



$$y_i \sim f(x_i) + \boldsymbol{\theta_i}$$

$$f(x_i) = \boldsymbol{a}x_i^2 + \boldsymbol{b}x + \boldsymbol{c}$$

Params of interest

Conclusion: The model is sensitive to internal compatibility of the data

Final thoughts

Statistical data analysis has made significant progress in recent decades; here touched only on three areas:

Attention to subtleties related to foundational issues

Adoption of modern ML methods

Continued development of accurate statistical models

For an overview of recent history, see e.g., PhyStat25 Symposium https://indico.cern.ch/event/1465837/ which concluded that (M. Kuusela, B. Nachman):

> ”Phystatistics” has emerged as a proper subfield, analogous to biostatistics, with problems requiring distinct expertise, approaches and thinking.

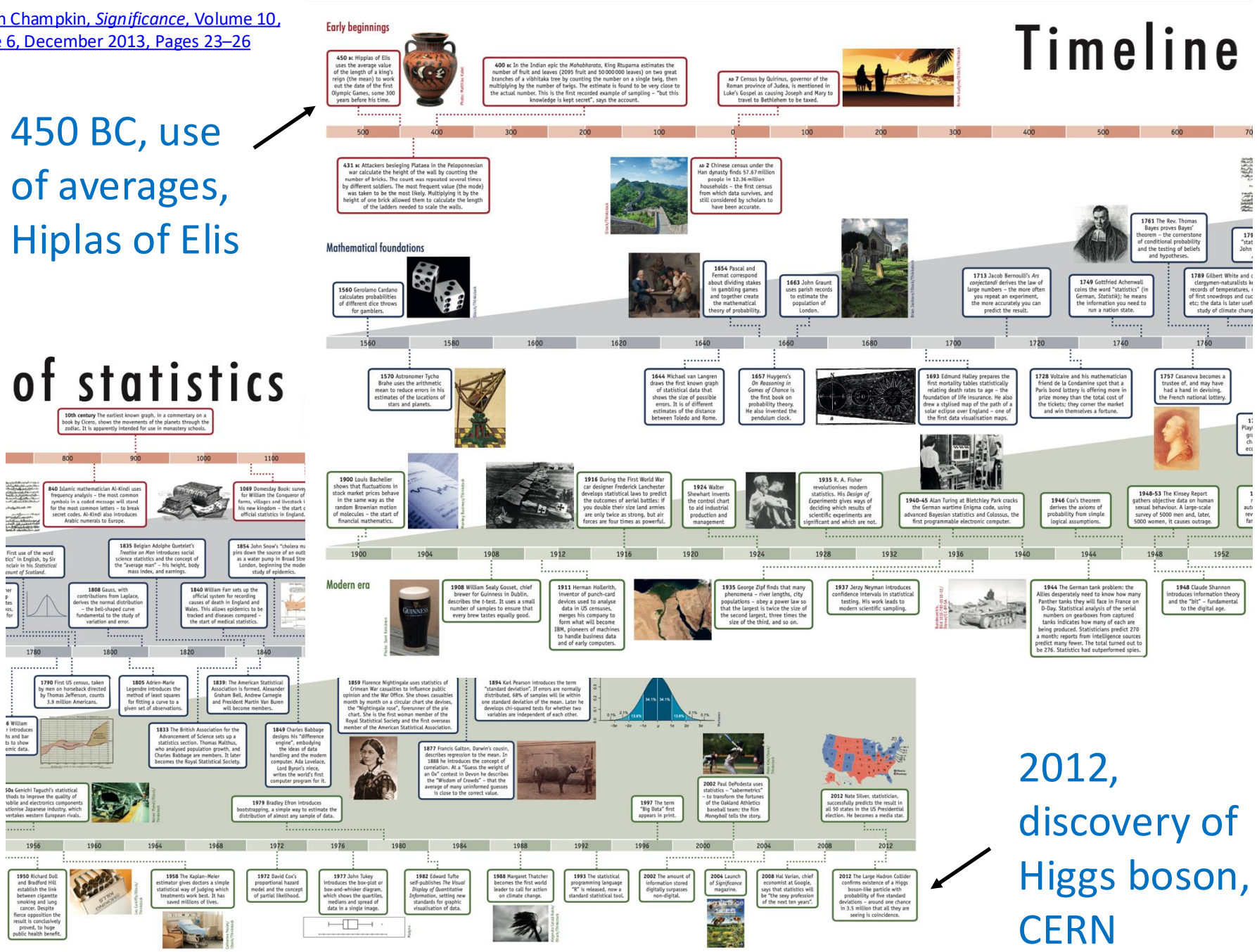
Will Phystatistics become a sandbox of ML/AI algorithms? I think not, because of the fundamental role played by statistics in the scientific method – we will always need to confront our theories with measurements in a way that incorporates all relevant uncertainties.

Extra Slides

Julian Champkin, *Significance*, Volume 10, Issue 6, December 2013, Pages 23–26

Timeline of statistics

450 BC, use of averages, Hiplas of Elis



2012, discovery of Higgs boson, CERN