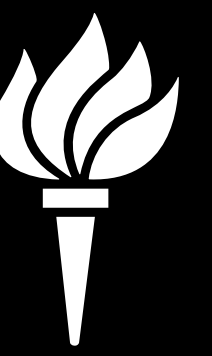




NYU CENTER FOR
DATA SCIENCE

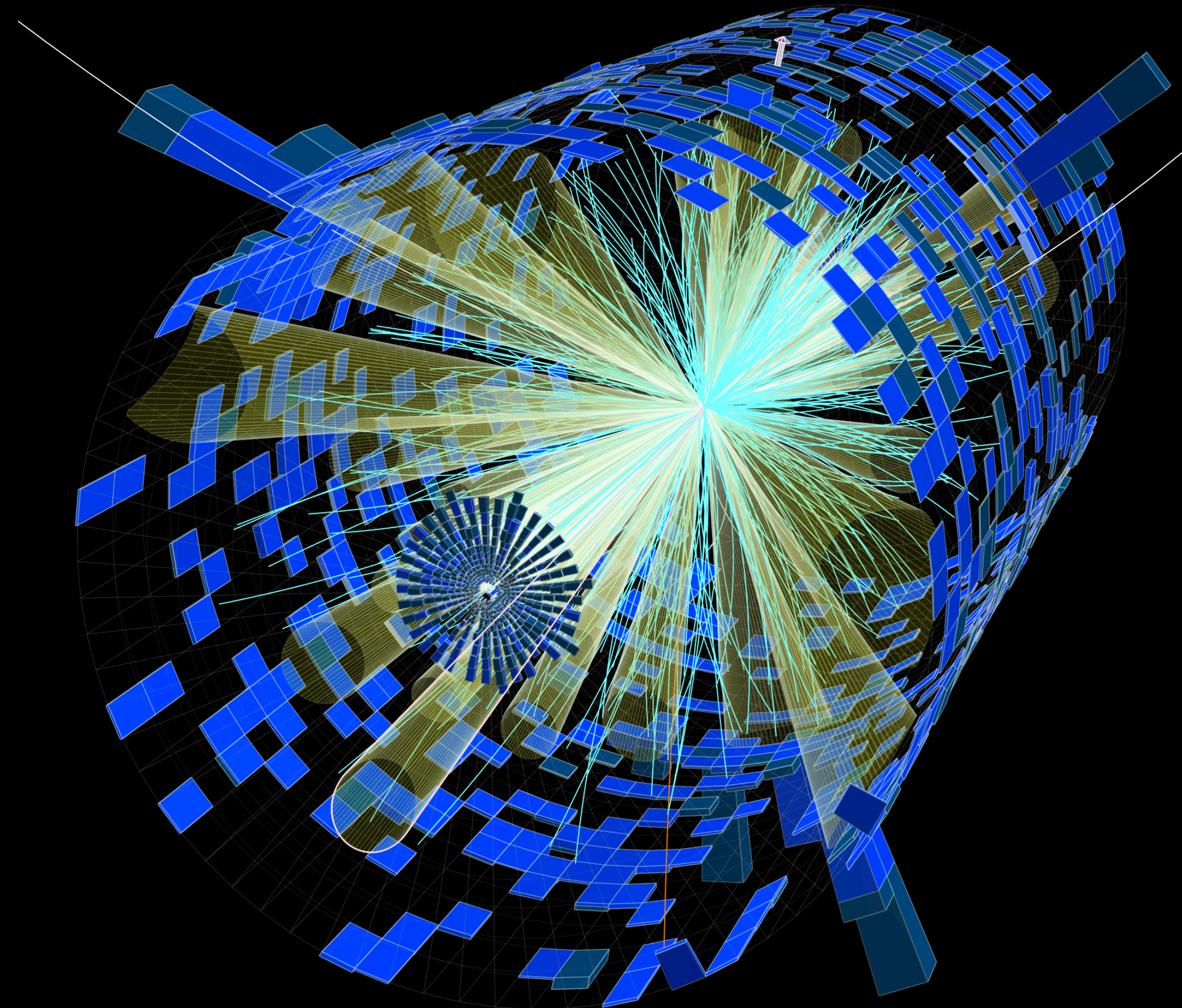
Meta AI

CENTER FOR
COSMOLOGY AND
PARTICLE PHYSICS



NO FREE LUNCH

SETTING OUR EXPECTATIONS FOR MODEL-INDEPENDENT SEARCHES



@KyleCranmer
New York University
Department of Physics
Center for Data Science
CILVR Lab

Pre-Amble

Anomaly detection

Lots of interest recently in anomaly detection — fueled by machine learning

The LHC Olympics 2020

A Community Challenge for Anomaly
Detection in High Energy Physics



Gregor Kasieczka (ed),¹ Benjamin Nachman (ed),^{2,3} David Shih (ed),⁴ Oz Amram,⁵ Anders Andreassen,⁶ Kees Benkendorfer,^{2,7} Blaz Bortolato,⁸ Gustaaf Brooijmans,⁹ Florencia Canelli,¹⁰ Jack H. Collins,¹¹ Biwei Dai,¹² Felipe F. De Freitas,¹³ Barry M. Dillon,^{8,14} Ioan-Mihail Dinu,⁵ Zhongtian Dong,¹⁵ Julien Donini,¹⁶ Javier Duarte,¹⁷ D. A. Faroughy,¹⁰ Julia Gonski,⁹ Philip Harris,¹⁸ Alan Kahn,⁹ Jernej F. Kamenik,^{8,19} Charanjit K. Khosa,^{20,30} Patrick Komiske,²¹ Luc Le Pottier,^{2,22} Pablo Martín-Ramiro,^{2,23} Andrej Matevc,^{8,19} Eric Metodiev,²¹ Vinicius Mikuni,¹⁰ Inês Ochoa,²⁴ Sang Eon Park,¹⁸ Maurizio Pierini,²⁵ Dylan Rankin,¹⁸ Veronica Sanz,^{20,26} Nilai Sarda,²⁷ Uroš Seljak,^{2,3,12} Aleks Smolkovic,⁸ George Stein,^{2,12} Cristina Mantilla Suarez,⁵ Manuel Szewc,²⁸ Jesse Thaler,²¹ Steven Tsan,¹⁷ Silviu-Marian Udrescu,¹⁸ Louis Vaslin,¹⁶ Jean-Roch Vlimant,²⁹ Daniel Williams,⁹ Mikaeel Yunus¹⁸

3	Unsupervised	11
3.1	Anomalous Jet Identification via Variational Recurrent Neural Network	11
3.2	Anomaly Detection with Density Estimation	16
3.3	BuHuLaSpa: Bump Hunting in Latent Space	19
3.4	GAN-AE and BumpHunter	24
3.5	Gaussianizing Iterative Slicing (GIS): Unsupervised In-distribution Anomaly Detection through Conditional Density Estimation	29
3.6	Latent Dirichlet Allocation	33
3.7	Particle Graph Autoencoders	38
3.8	Regularized Likelihoods	42
3.9	UCluster: Unsupervised Clustering	46
4	Weakly Supervised	51
4.1	CWoLa Hunting	51
4.2	CWoLa and Autoencoders: Comparing Weak- and Unsupervised methods for Resonant Anomaly Detection	55
4.3	Tag N' Train	60
4.4	Simulation Assisted Likelihood-free Anomaly Detection	63
4.5	Simulation-Assisted Decorrelation for Resonant Anomaly Detection	68
5	(Semi)-Supervised	71
5.1	Deep Ensemble Anomaly Detection	71
5.2	Factorized Topic Modeling	77
5.3	QUAK: Quasi-Anomalous Knowledge for Anomaly Detection	81
5.4	Simple Supervised learning with LSTM layers	85

A spectrum



A spectrum



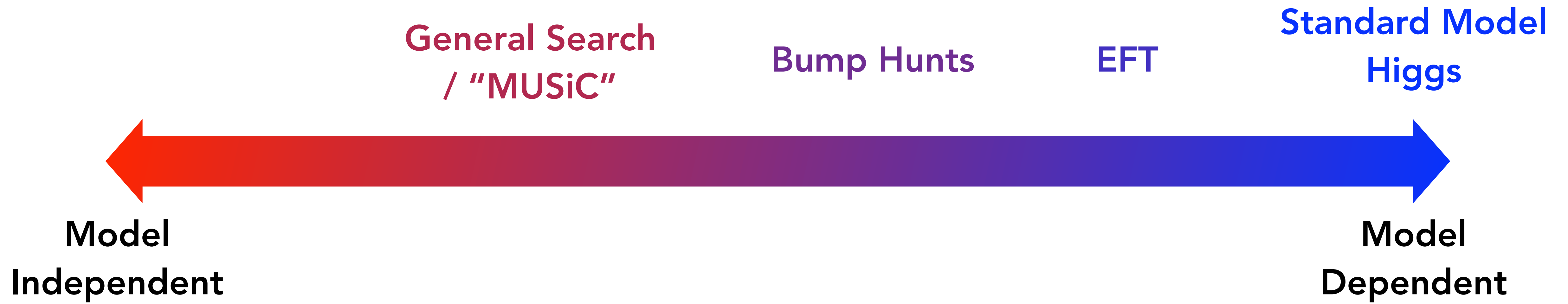
A spectrum



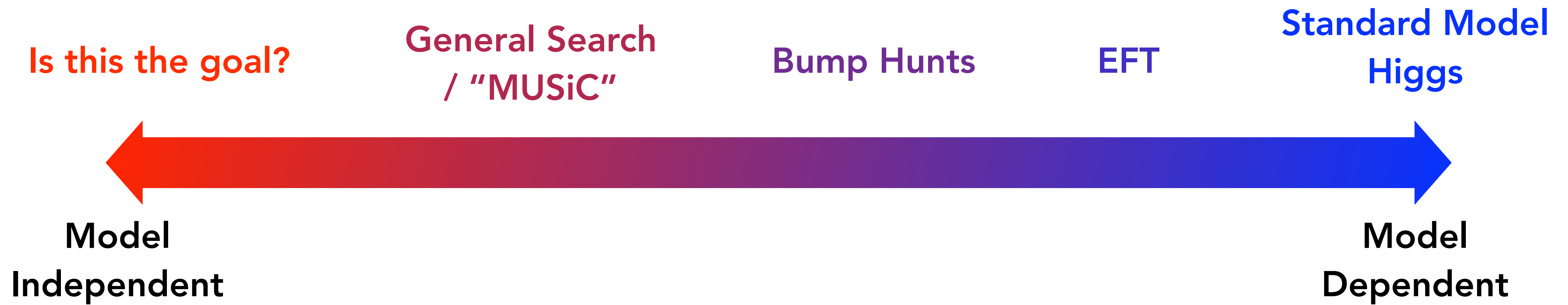
A spectrum



A spectrum



A spectrum



A spectrum



A spectrum



A spectrum



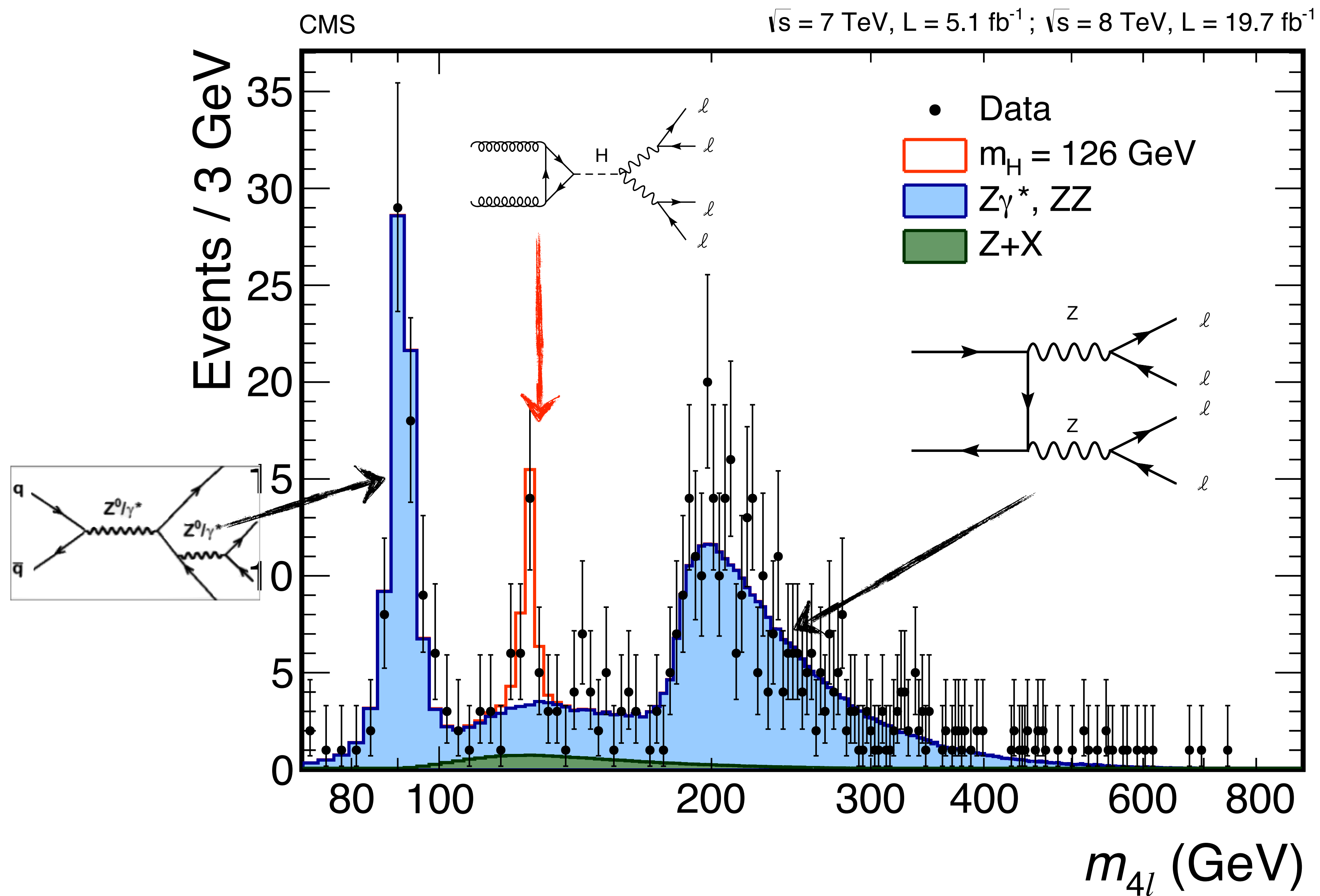
A spectrum



Is a fully model-independent approach our goal?

- What does that mean?
- Is it the right goal? Is it a reasonable goal?

The poster-child for model-dependent searches



Modeling particle physics processes

Theory
parameters
 θ



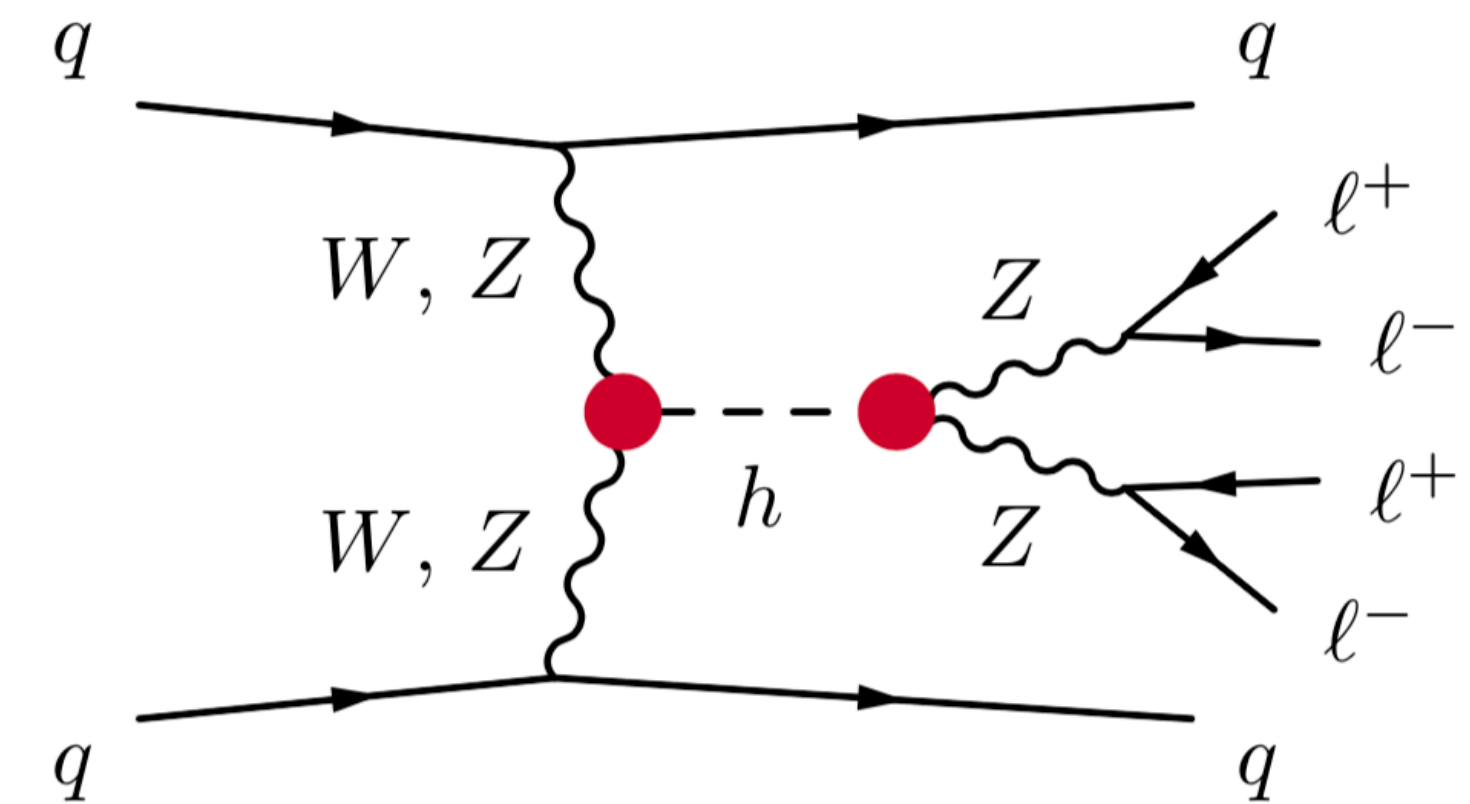
Modeling particle physics processes

Latent variables

Parton-level
momenta

Theory
parameters

z_p ← θ



← Evolution

Modeling particle physics processes

Latent variables

Shower
splittings

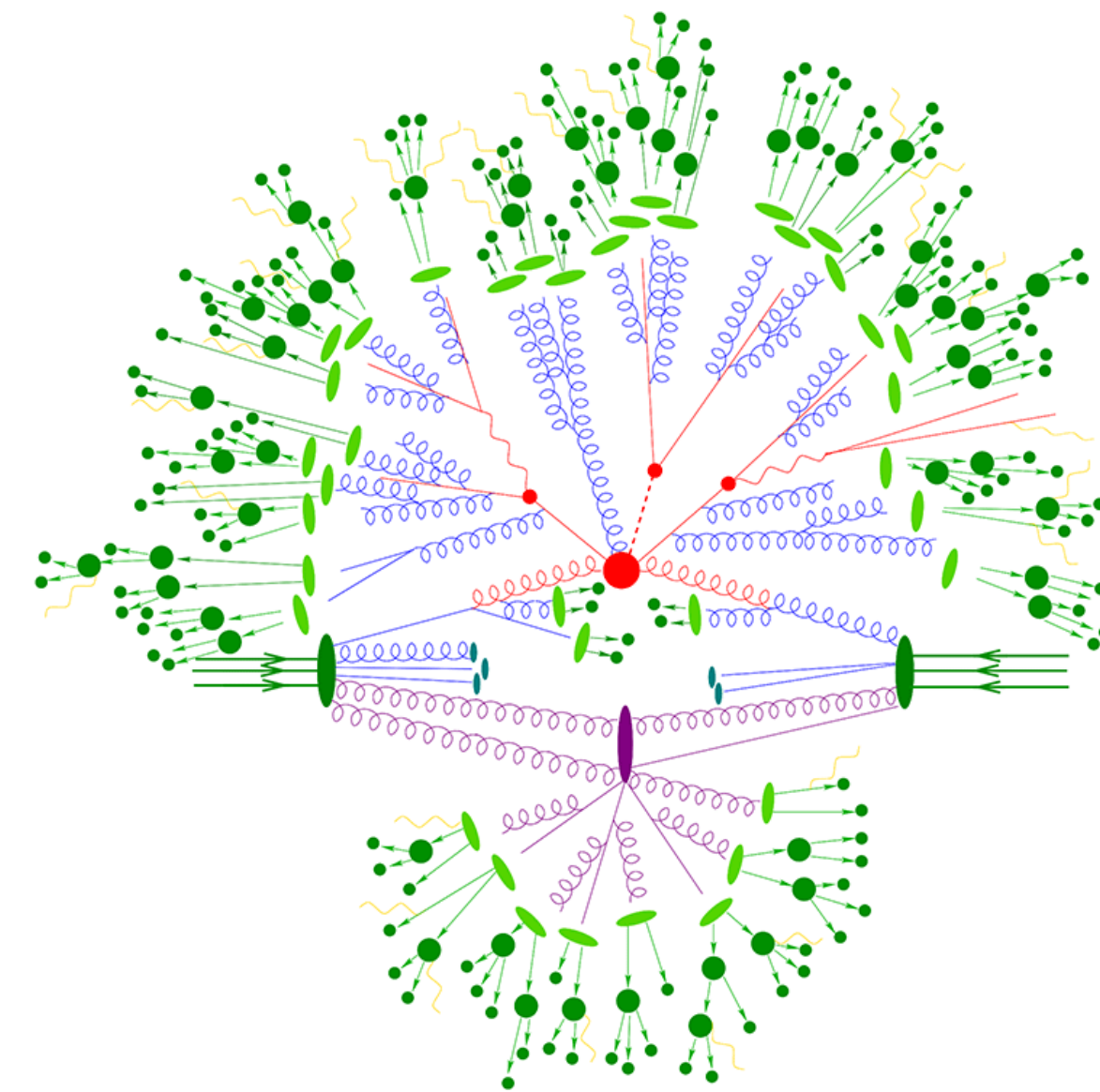
Parton-level
momenta

Theory
parameters

z_s

z_p

θ

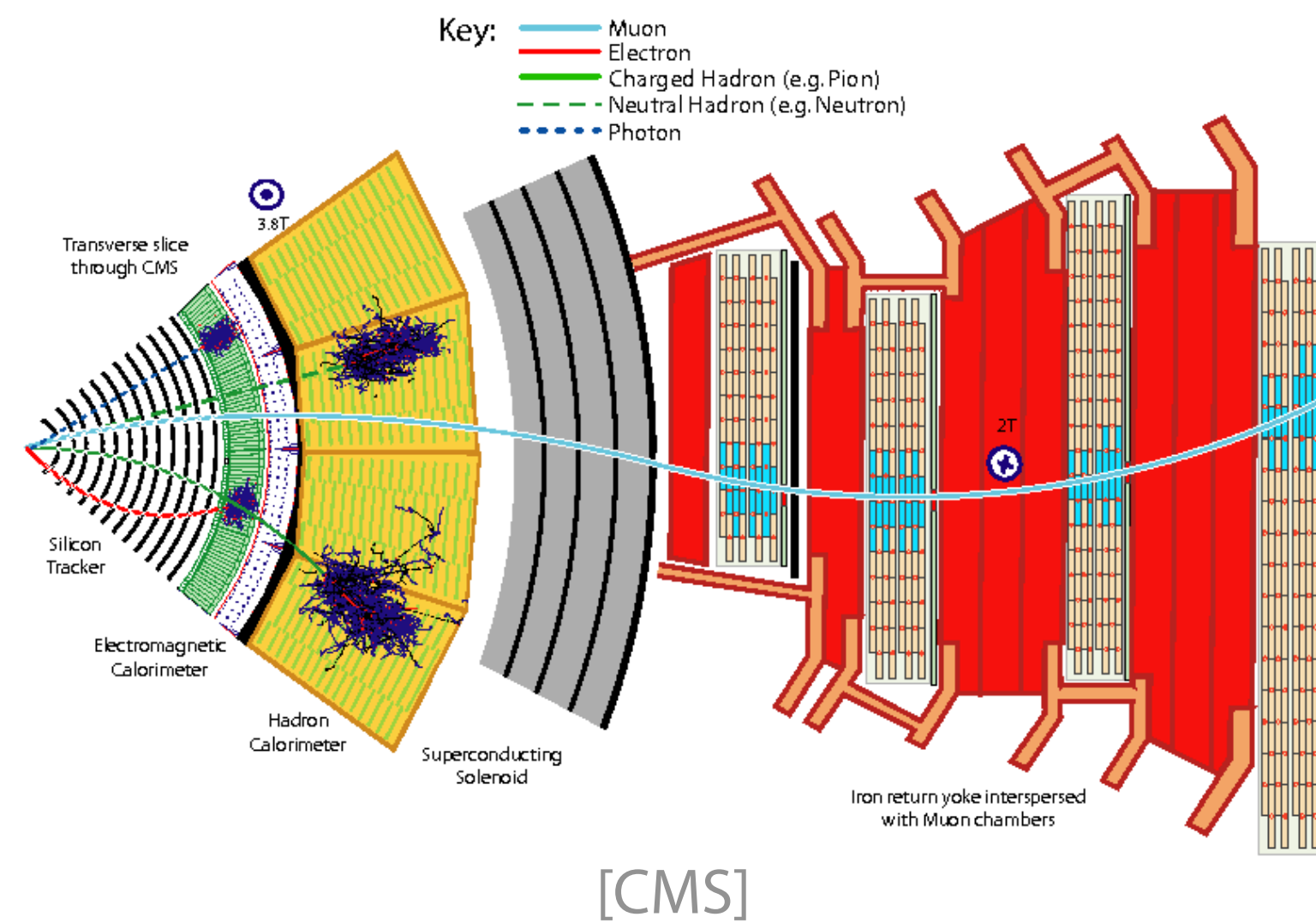
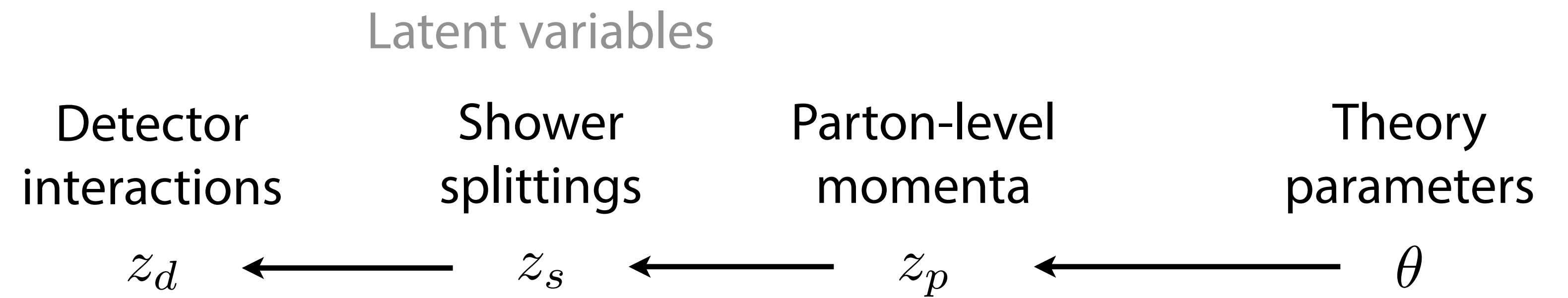


[F. Krauss]

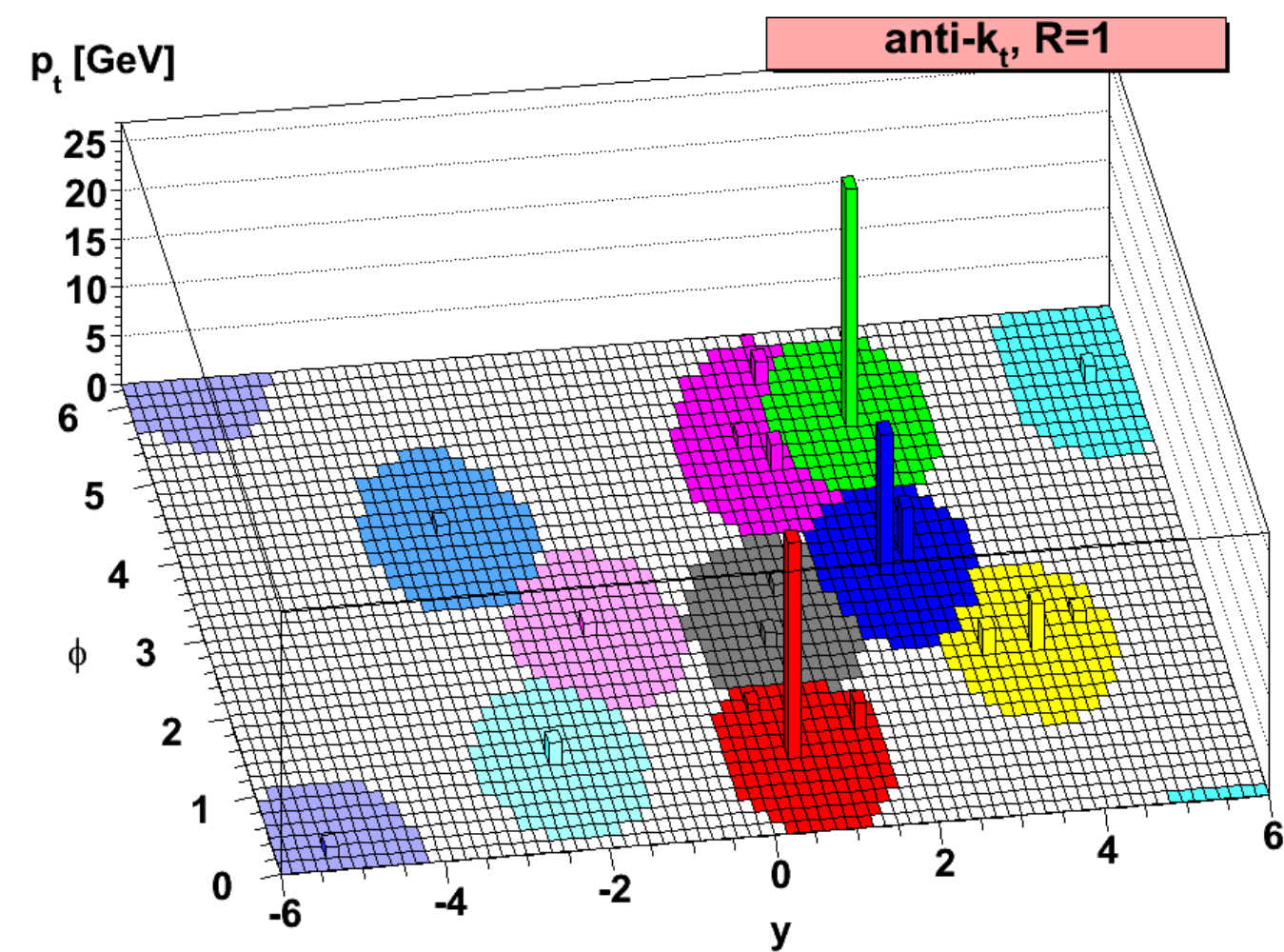
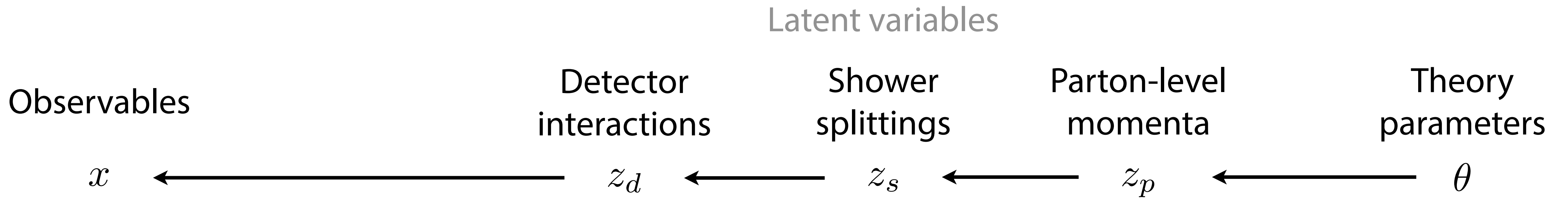


Evolution

Modeling particle physics processes



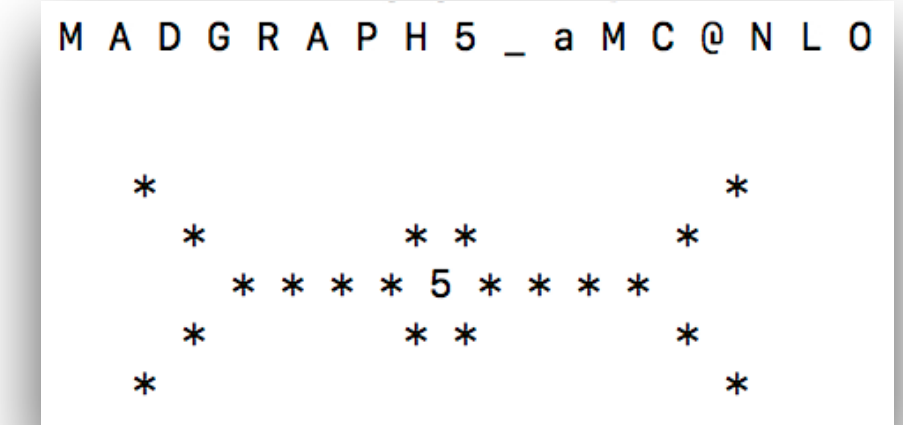
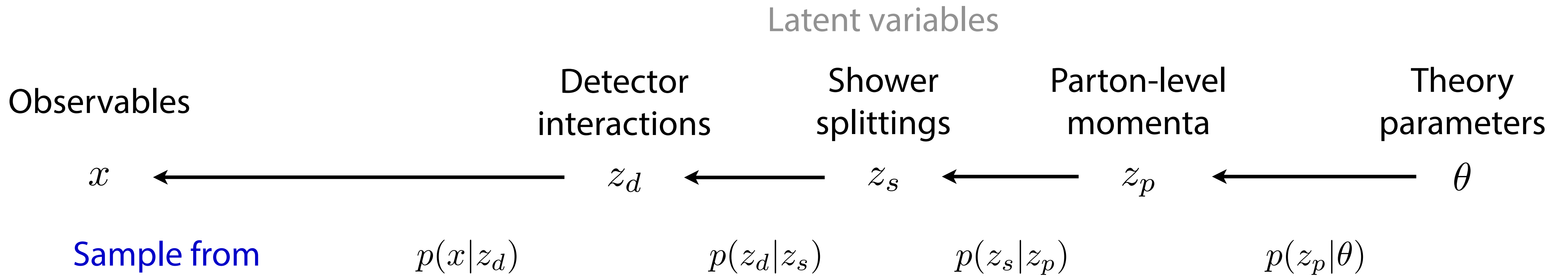
Modeling particle physics processes



[M. Cacciari, G. Salam, G. Soyez 0802.1189]

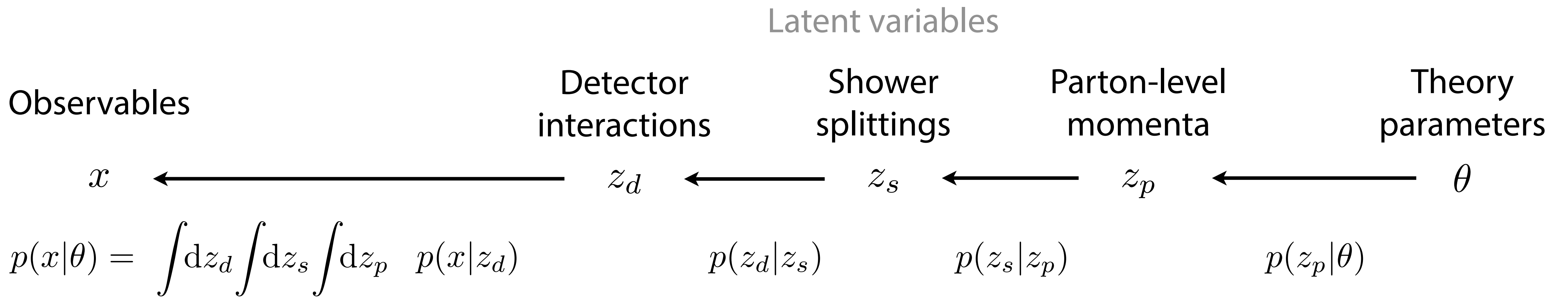


Modeling particle physics processes

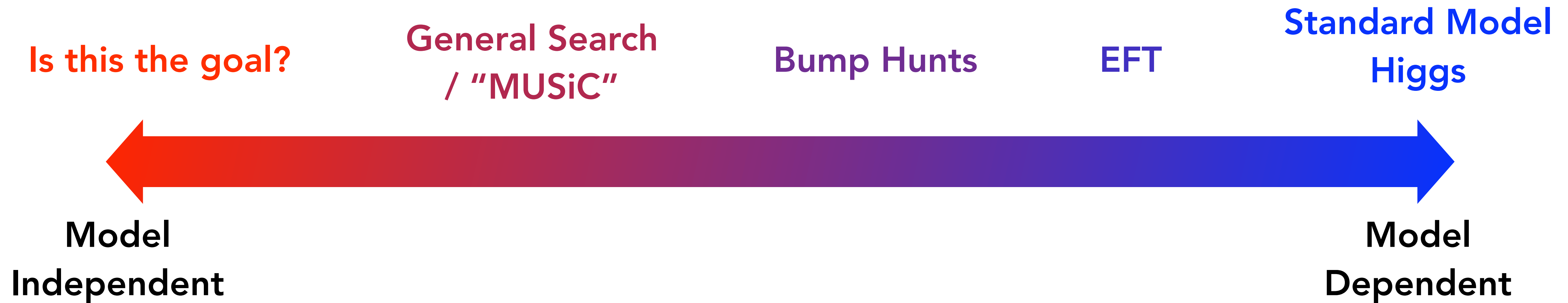


←————— Prediction (simulation)

Modeling particle physics processes



A spectrum



Even our most model-dependent searches have different degrees

- It is easy to take for granted, but let us be pedantic

Beyond the standard model

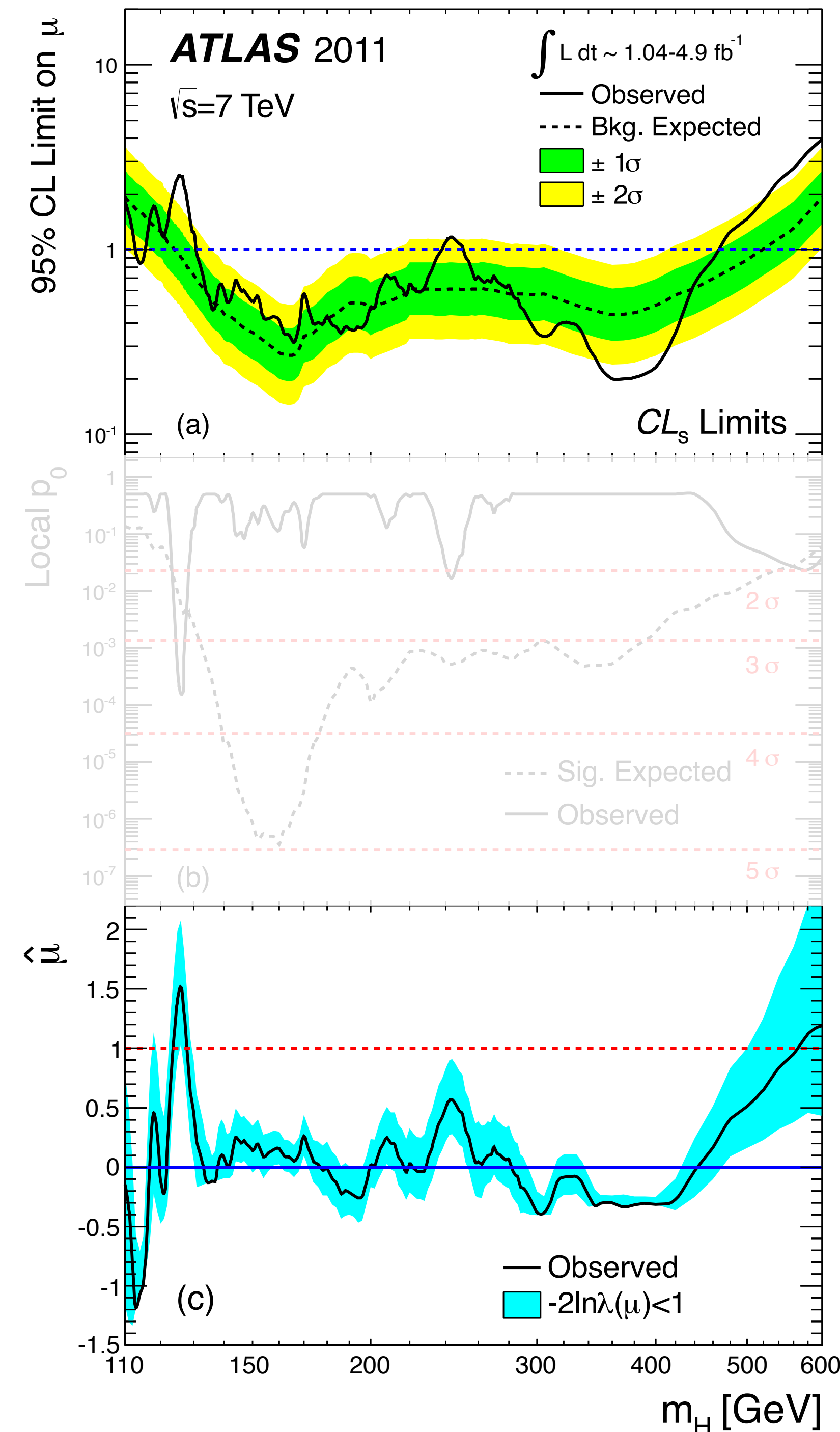
The Standard Model really only had one free parameter (m_H)

- Once m_H is specified, so are the cross-section, branching ratio, and efficiencies

- Signal strength $\mu = \frac{\sigma \cdot BR}{\sigma_{sm} \cdot BR_{sm}}$ and $\mu = 1$ is the SM

So what *is* the model that corresponds to $\mu \neq 1$?

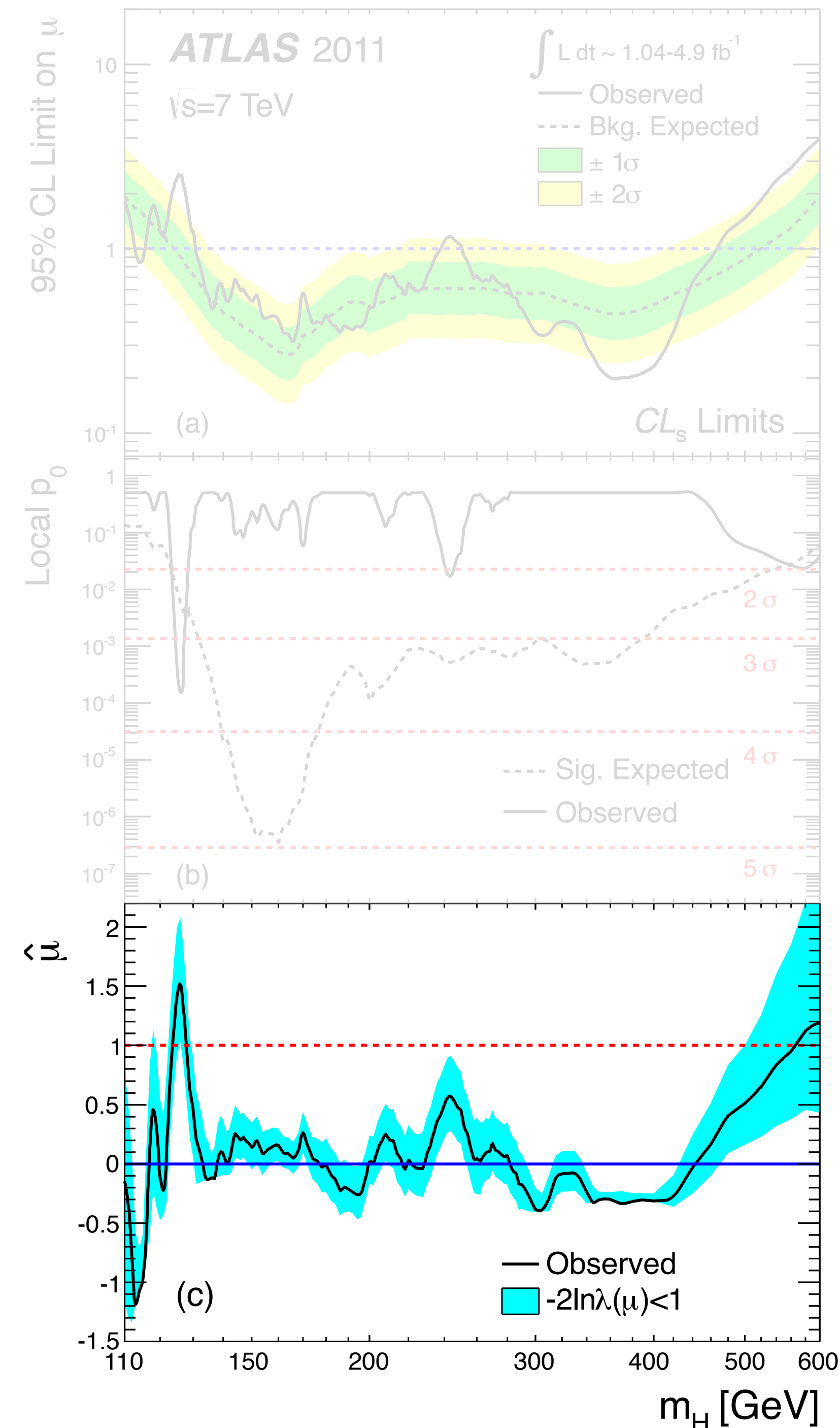
- It is a well-defined statistical model
- Just scale signal template proportionally
- But it isn't a model defined by quantum field theory
- (Yes, there are some EFTs that map to it)



Beyond the standard model

Here we even consider $\mu < 0$, which would correspond to a negative number of signal events.

- That doesn't make sense physically
- The statistical model is well defined as long as the total number of events is positive
- It indicates a deficit of events
- In other cases, destructive quantum mechanical interference might lead to such a deficit of events



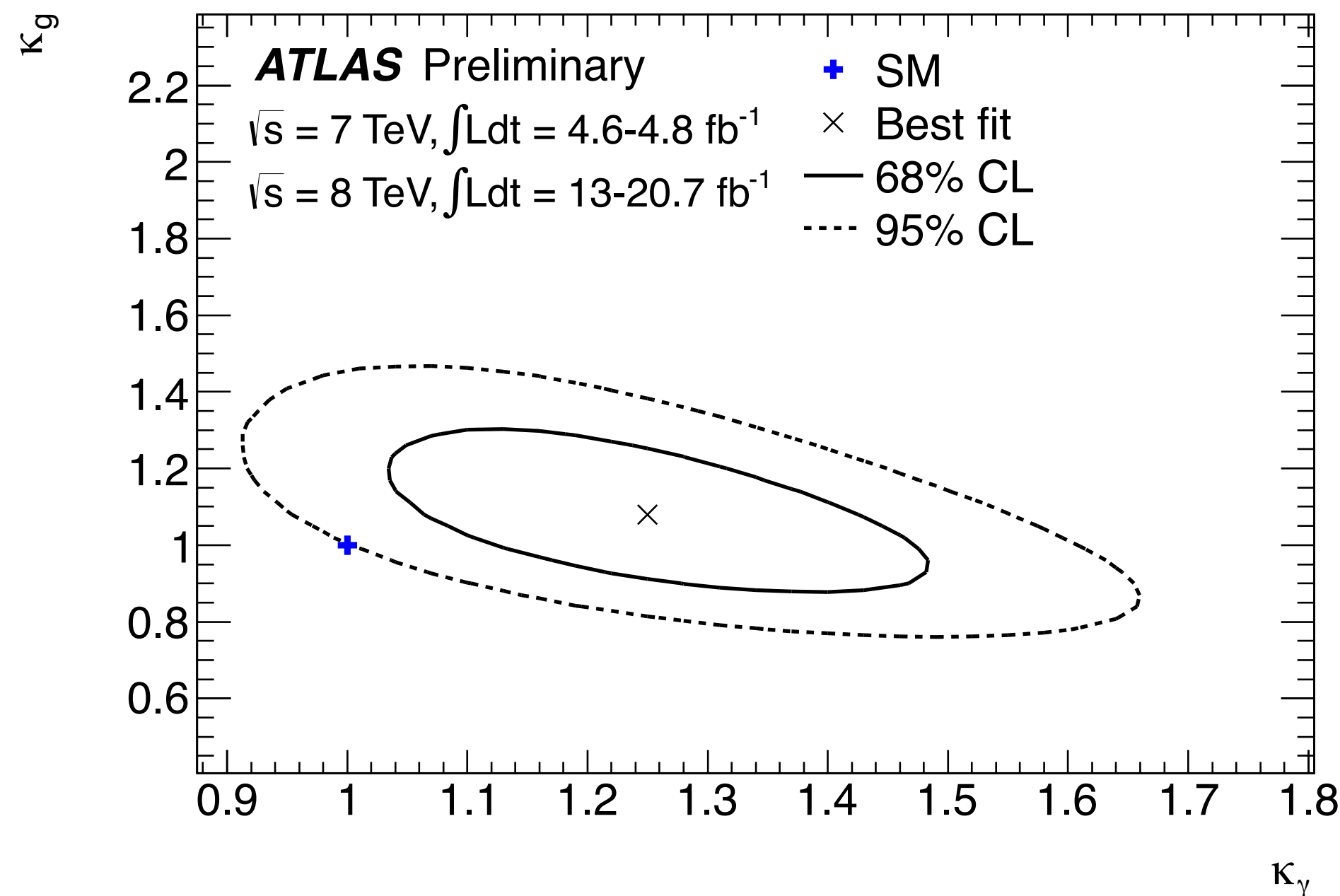
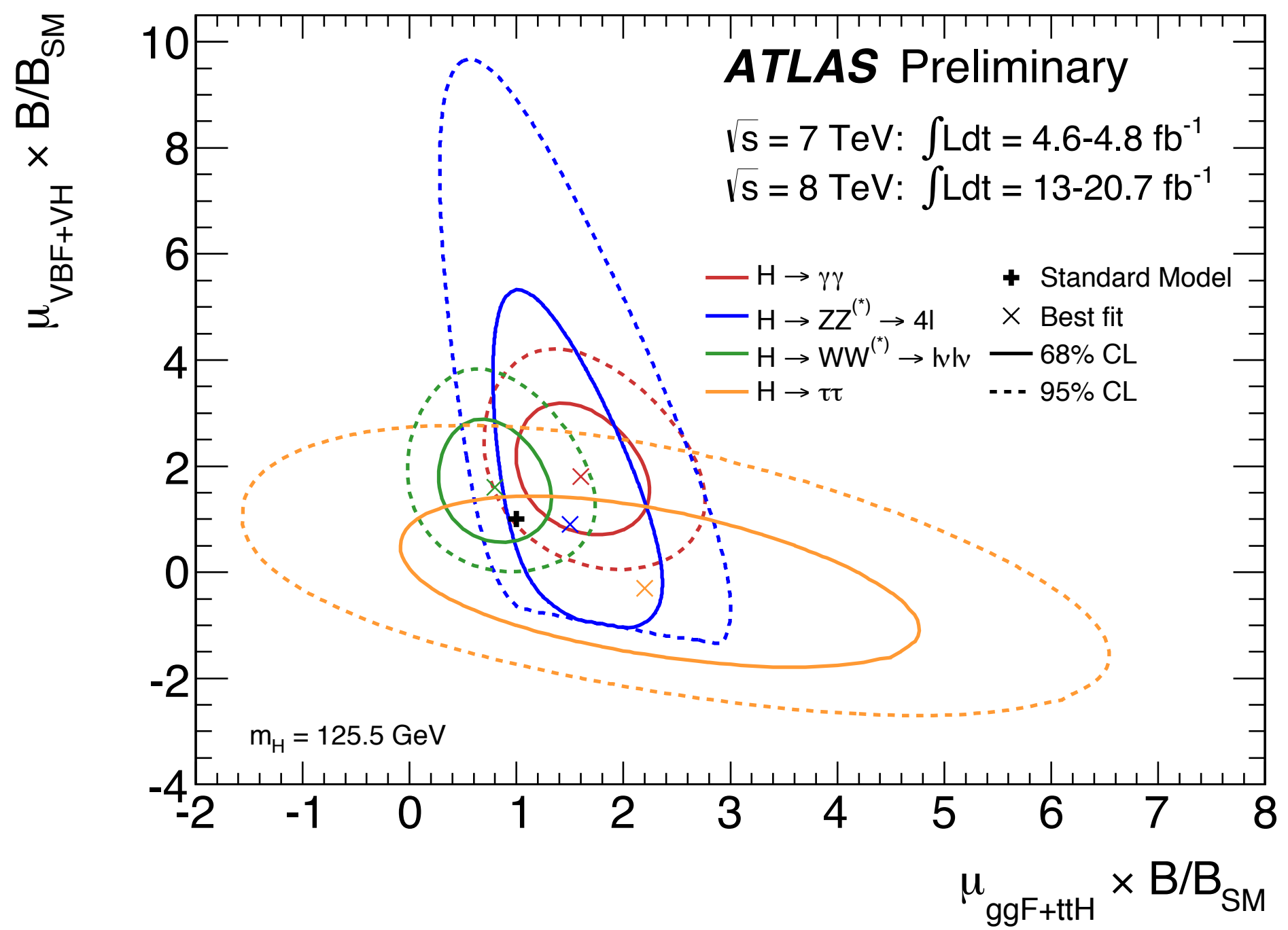
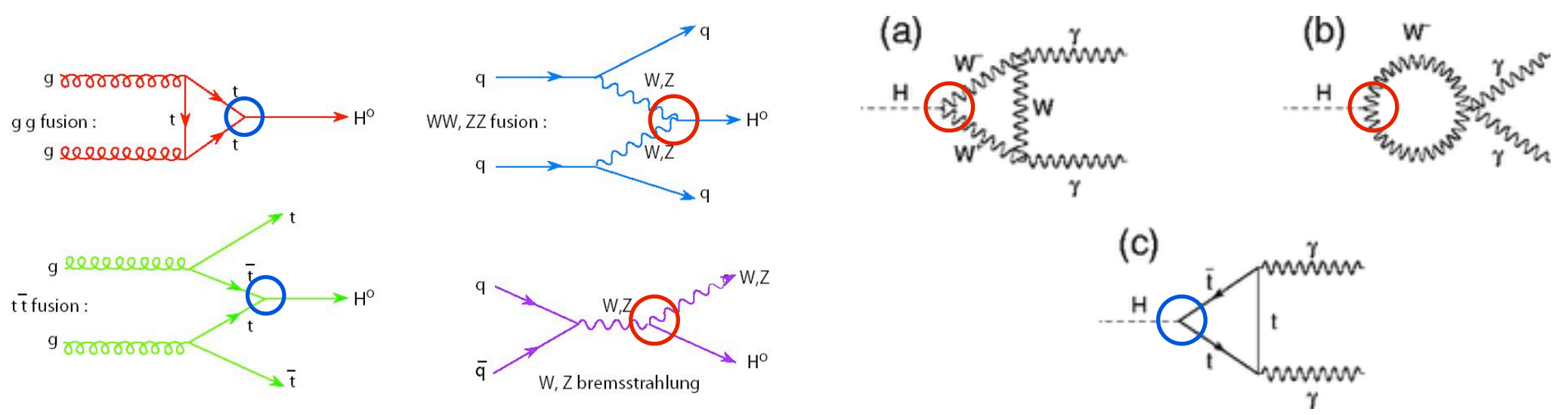
Beyond the standard model

Many production and decay modes

- Can consider deviations from the SM
- Not a valid QFT, but it was practical

Production →	$\gamma\gamma$	ZZ	WW	$Z\gamma$	gg	bb	$\tau\tau$	

Decay →

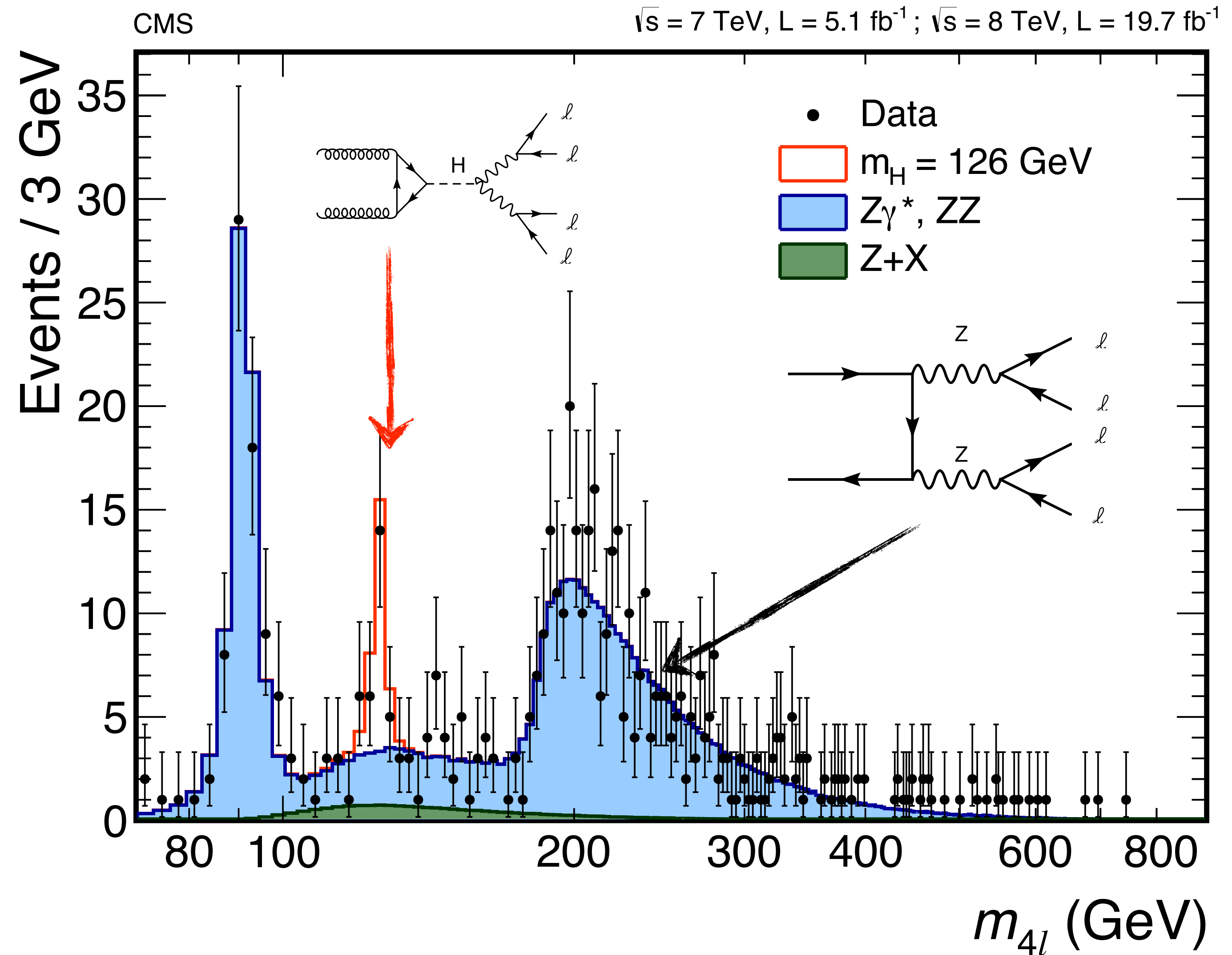


Background only?

Also, what is the “background-only” hypothesis here?

- The Higgs is needed for the SM to work
- “SM background-only” without the Higgs isn’t meaningful
- We don’t have an unique, operationally defined, consistent QFT to serve as the “null hypothesis”

In practice, “ignore” Higgs component of the SM prediction



What if there was no Higgs boson?

In the run up to the SSC and LHC, arguments based on "No-Lose Theorem"

- Either we will see a light Higgs in high energy collisions, or
- We will see strong WW , WZ , ZZ scattering

Generic prediction, but details depend on specific theory

May 1985

LBL-19470
UCB-PTH-85/19

THE TeV PHYSICS OF STRONGLY INTERACTING W's AND Z's *

Michael S. Chanowitz

Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

Mary K. Gaillard

Lawrence Berkeley Laboratory
and
Department of Physics
University of California
Berkeley, California 94720

ABSTRACT

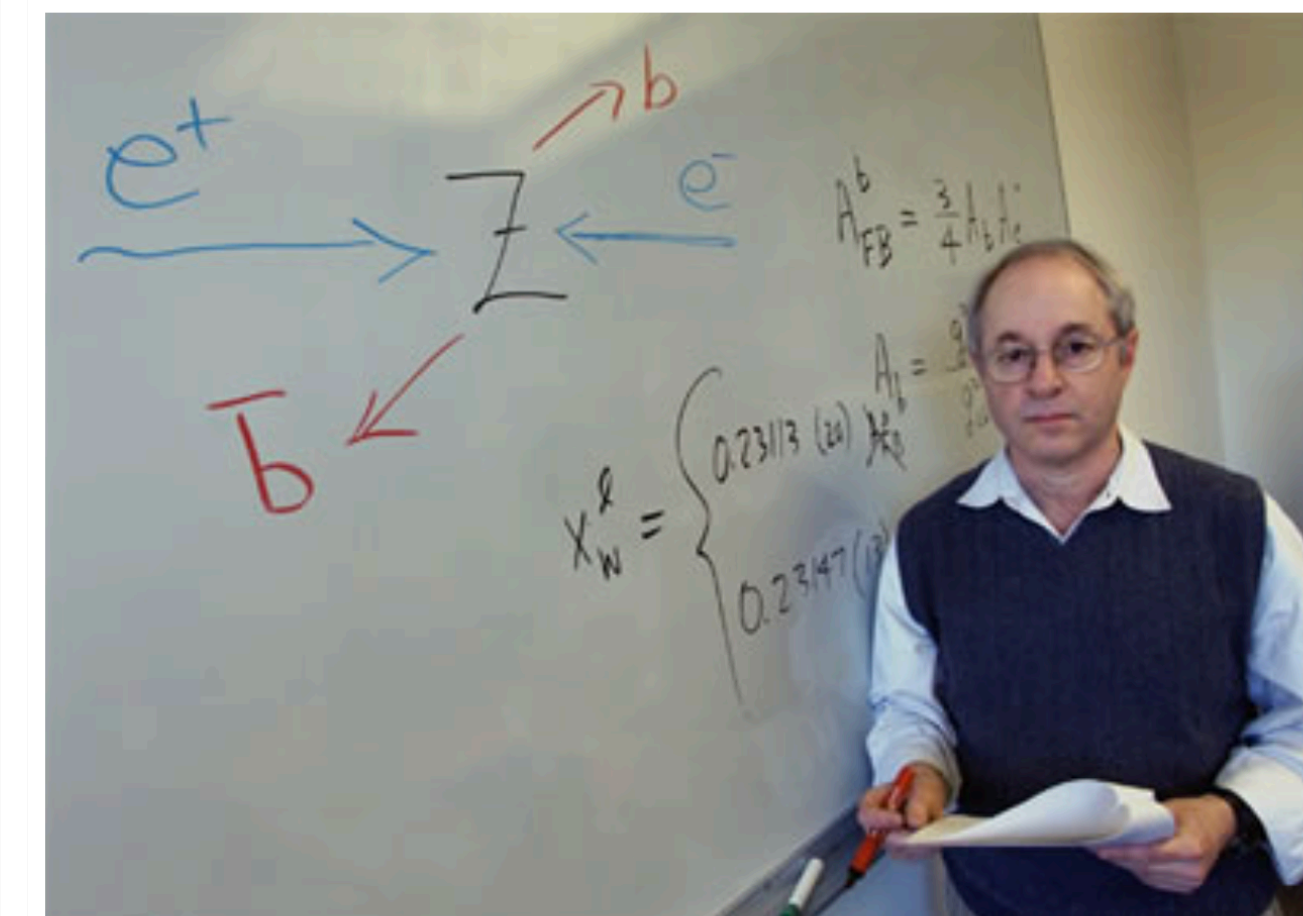
There are two possibilities for electroweak symmetry breaking: either there is a scalar particle much lighter than 1 TeV or the longitudinal components of W and Z bosons interact strongly at center of mass energies of order 1 TeV or more. We study the general signatures of a strongly interacting W, Z system and conclude that these two possibilities can be unambiguously distinguished by a hadron collider facility capable of observing the enhanced production of WW, WZ and ZZ pairs that will occur if W 's and Z 's have strong interactions. Detection of the enhanced signal over background requires hadron collisions at a center of mass energy of order $\sqrt{s} = 40 \text{ TeV}$ and an integrated luminosity of order 10^{40} cm^{-2} . With these parameters we predict 3800 to 6000 gauge boson pairs satisfying cuts for which only 2600 pairs would be produced in the absence of strong interactions.

As our results draw on the global chiral $SU(2)$ symmetry of the scalar sector of the standard $SU(2) \times U(1)$ model, we give an extended proof, to all orders in the generalized renormalizable gauge, that high energy amplitudes of longitudinal W 's and Z 's are well approximated by amplitudes of the corresponding unphysical scalars. The results are applicable to the broad class of strong interaction models that admit a global chiral $SU(2)$ symmetry.

What if there is no Higgs boson?

Higgs or Not - ATLAS will solve the mystery of mass

30 November 2011 | By [Michael Chanowitz](#)



Michael Chanowitz is a theoretical physicist at Lawrence Berkeley National Laboratory. He is the author with Mary K. Gaillard of the heavily cited 1985 paper entitled: "The TeV Physics of Strongly Interacting W's and Z's." While ATLAS and CMS are narrowing the allowed mass regions where a Higgs boson may be found, Chanowitz addresses what would be the impact of not finding the Higgs. (Image: ATLAS Experiment)

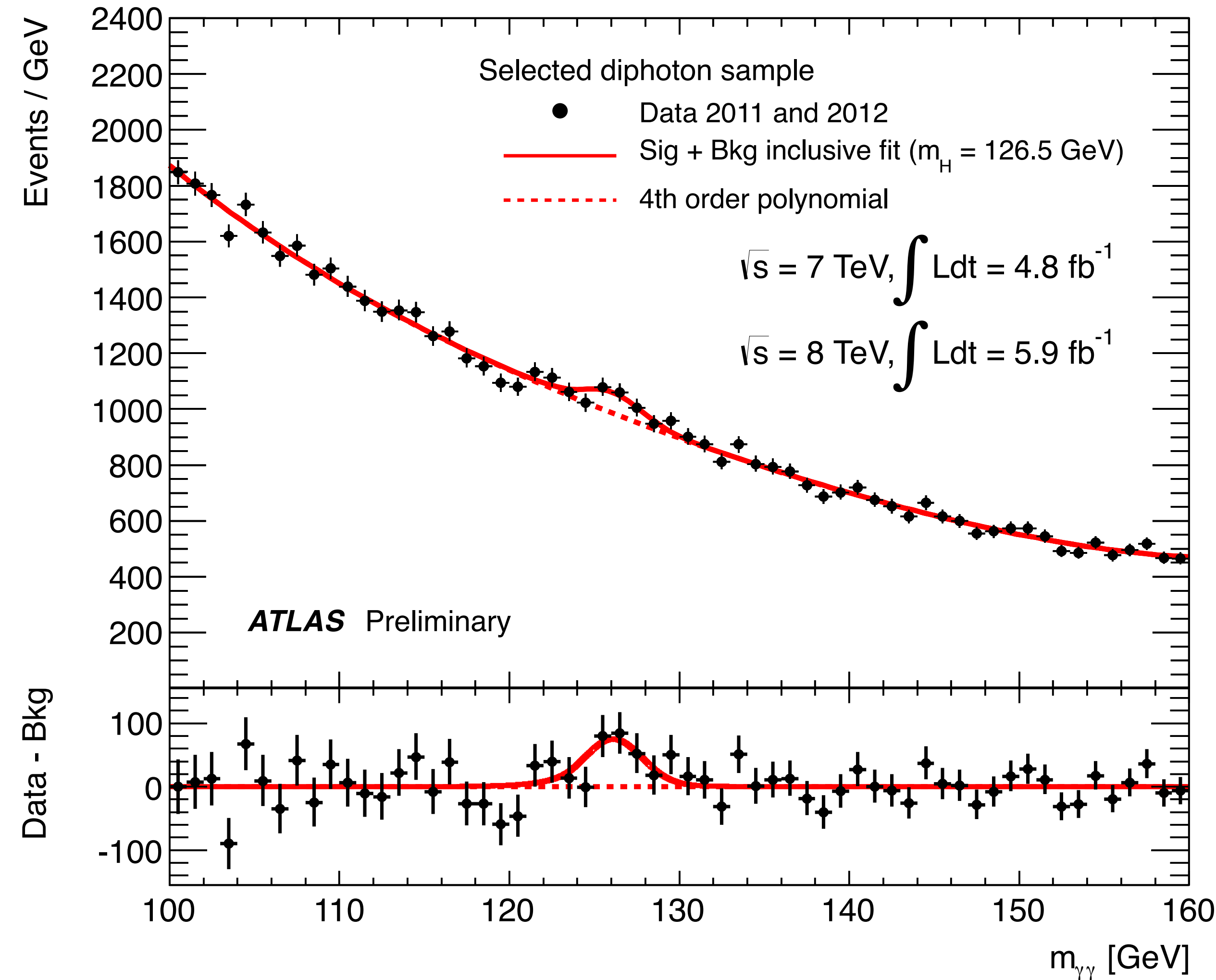
What is the background-only model here?

In the case of Higgs to two photons, the background-only hypothesis isn't really based on QFT at all

- There is a sizable background from jets faking photons, which depends on details of jets and detector performance, hard to predict from first principles
- Instead, we fit the background with a smooth function

The null hypothesis is a 4th order polynomial!

- This choice was informed / validated with simulated data, but we should recognize it for what it is



Takeaway

The main point of the slides above is that:

- Statistical model used for hypothesis test was always well defined, but
- the connection of that statistical model to quantum field theory varies
 - Many reasonable assumptions that we have become used to as a field
 - Easy to take for granted and be blind to them
 - Or we can use these as baby-steps for a more “model independent” strategy by loosening the connection to QFT while maintaining some intuitive notions for what we mean by background and signal (or null / alternate hypothesis)

Searching without an alternate

(aka Goodness of fit / Out of Distribution Detection / Anomaly Detection)

isn't a well defined goal

(It is underspecified)

Anomaly detection

Lots of interest recently in anomaly detection — fueled by machine learning

- Formally the same as Goodness-of-Fit or Out-of-Distribution detection

The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics



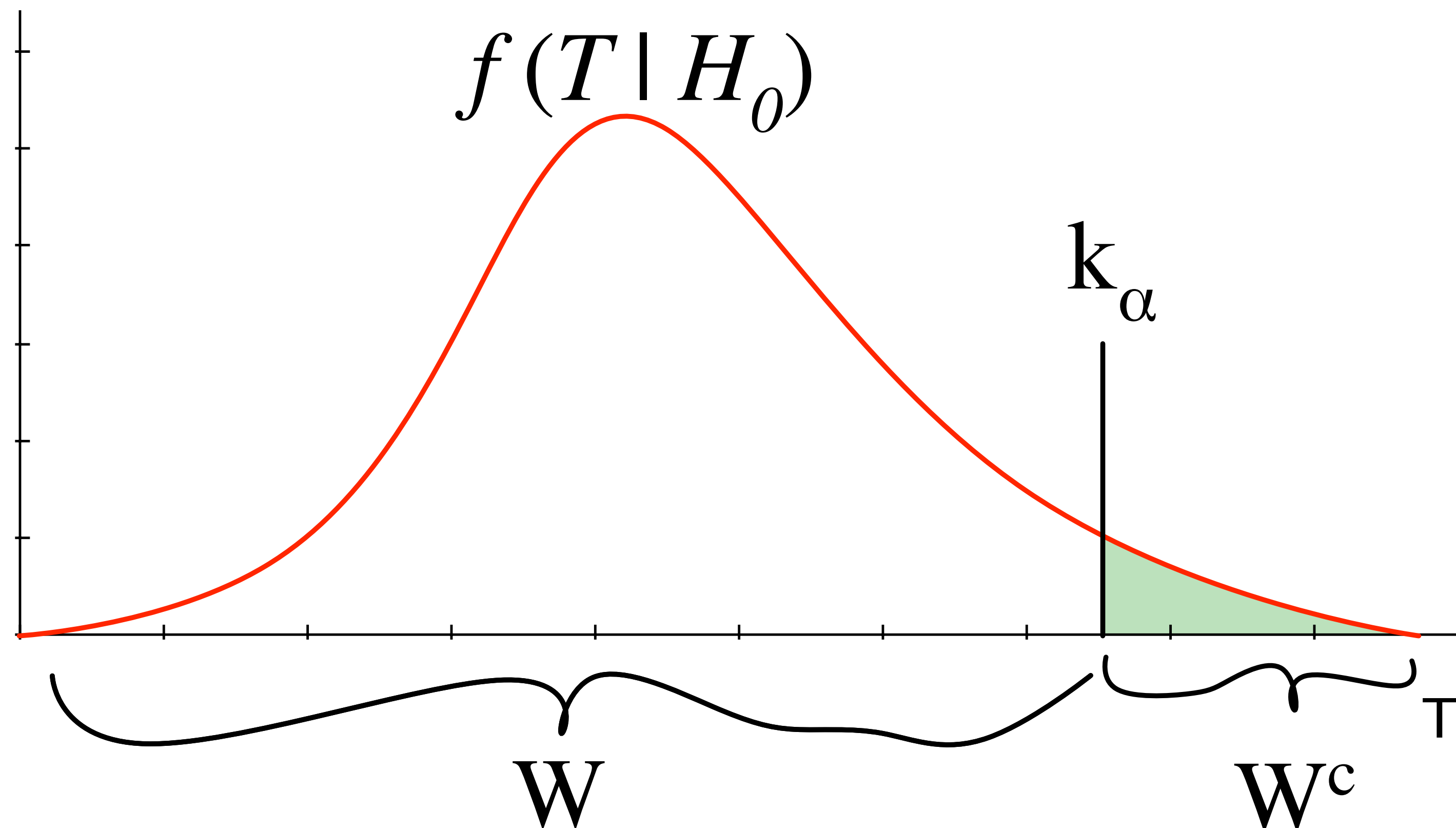
Gregor Kasieczka (ed),¹ Benjamin Nachman (ed),^{2,3} David Shih (ed),⁴ Oz Amram,⁵ Anders Andreassen,⁶ Kees Benkendorfer,^{2,7} Blaz Bortolato,⁸ Gustaaf Brooijmans,⁹ Florencia Canelli,¹⁰ Jack H. Collins,¹¹ Biwei Dai,¹² Felipe F. De Freitas,¹³ Barry M. Dillon,^{8,14} Ioan-Mihail Dinu,⁵ Zhongtian Dong,¹⁵ Julien Donini,¹⁶ Javier Duarte,¹⁷ D. A. Faroughy,¹⁰ Julia Gonski,⁹ Philip Harris,¹⁸ Alan Kahn,⁹ Jernej F. Kamenik,^{8,19} Charanjit K. Khosa,^{20,30} Patrick Komiske,²¹ Luc Le Pottier,^{2,22} Pablo Martín-Ramiro,^{2,23} Andrej Matevc,^{8,19} Eric Metodiev,²¹ Vinicius Mikuni,¹⁰ Inês Ochoa,²⁴ Sang Eon Park,¹⁸ Maurizio Pierini,²⁵ Dylan Rankin,¹⁸ Veronica Sanz,^{20,26} Nilai Sarda,²⁷ Uroš Seljak,^{2,3,12} Aleks Smolkovic,⁸ George Stein,^{2,12} Cristina Mantilla Suarez,⁵ Manuel Szewc,²⁸ Jesse Thaler,²¹ Steven Tsan,¹⁷ Silviu-Marian Udrescu,¹⁸ Louis Vaslin,¹⁶ Jean-Roch Vlimant,²⁹ Daniel Williams,⁹ Mikaeel Yunus¹⁸

3	Unsupervised	11
3.1	Anomalous Jet Identification via Variational Recurrent Neural Network	11
3.2	Anomaly Detection with Density Estimation	16
3.3	BuHuLaSpa: Bump Hunting in Latent Space	19
3.4	GAN-AE and BumpHunter	24
3.5	Gaussianizing Iterative Slicing (GIS): Unsupervised In-distribution Anomaly Detection through Conditional Density Estimation	29
3.6	Latent Dirichlet Allocation	33
3.7	Particle Graph Autoencoders	38
3.8	Regularized Likelihoods	42
3.9	UCluster: Unsupervised Clustering	46
4	Weakly Supervised	51
4.1	CWoLa Hunting	51
4.2	CWoLa and Autoencoders: Comparing Weak- and Unsupervised methods for Resonant Anomaly Detection	55
4.3	Tag N' Train	60
4.4	Simulation Assisted Likelihood-free Anomaly Detection	63
4.5	Simulation-Assisted Decorrelation for Resonant Anomaly Detection	68
5	(Semi)-Supervised	71
5.1	Deep Ensemble Anomaly Detection	71
5.2	Factorized Topic Modeling	77
5.3	QUAK: Quasi-Anomalous Knowledge for Anomaly Detection	81
5.4	Simple Supervised learning with LSTM layers	85

Goodness-of-fit

Intuitively: does the null hypothesis H_0 fit the data?

- Pick some "test statistic" T (e.g. chi-square) and can compute the p-value
- If the p-value is small, reject the null

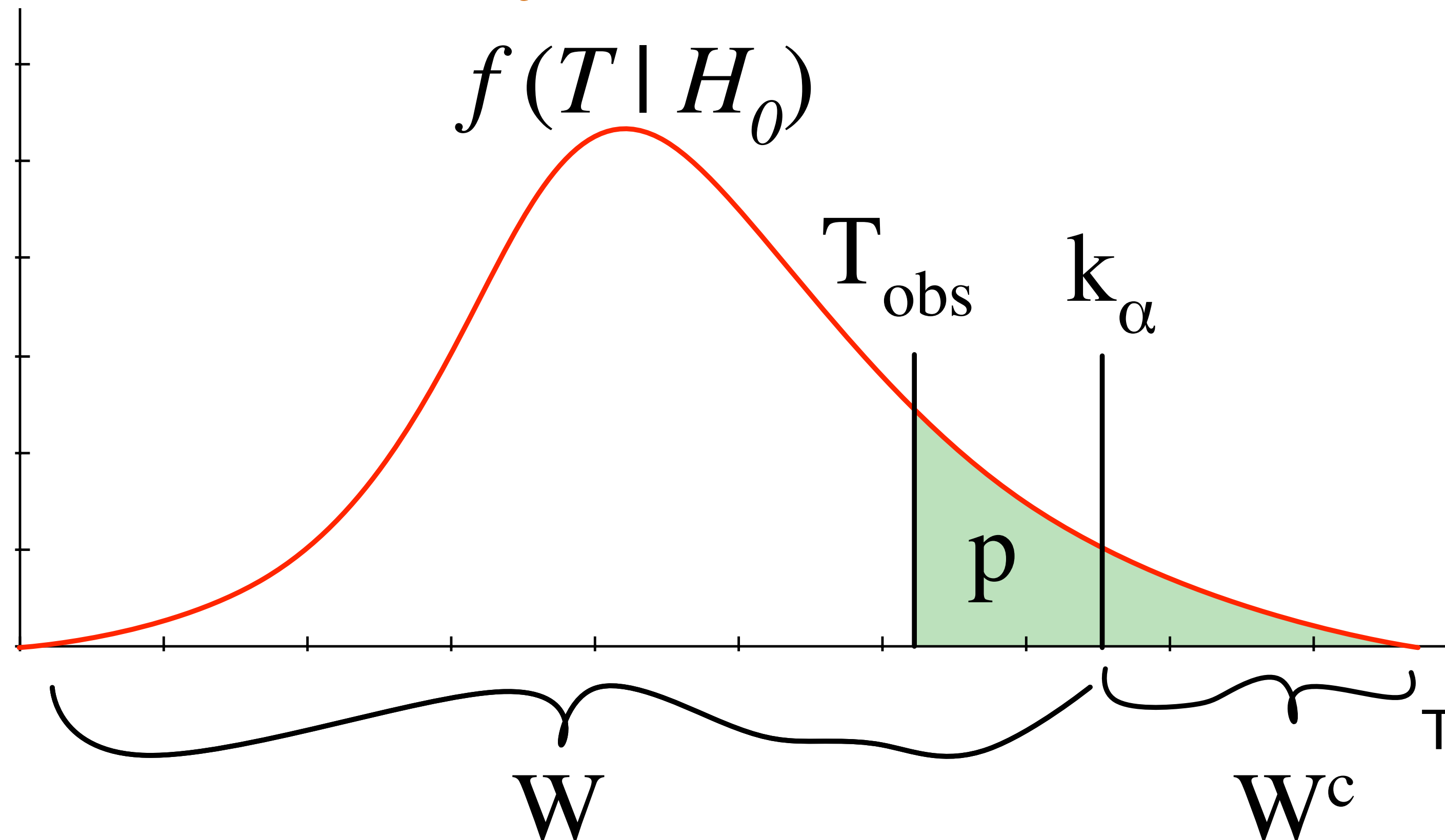


$$p = \int_{T_0}^{\infty} f(T | H_0)$$

Goodness-of-fit

Intuitively: does the null hypothesis H_0 fit the data?

- Pick some "test statistic" T (e.g. chi-square) and can compute the p-value
- If the p-value is small, reject the null



$$p = \int_{T_o}^{\infty} f(T | H_0)$$

Goodness-of-fit

Problem: There is no unique choice for the test statistic, giving rise to a large number of goodness-of-fit tests

- Can ask about the "power" of a GoF test to detect a given alternate

		Actual condition	
		Guilty	Not guilty
Decision	Verdict of 'guilty'	True Positive power	False Positive (i.e. guilt reported unfairly) Type I error
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) Type II error	True Negative

actually guilty ↔ new physics
 verdict guilty ↔ claim discovery

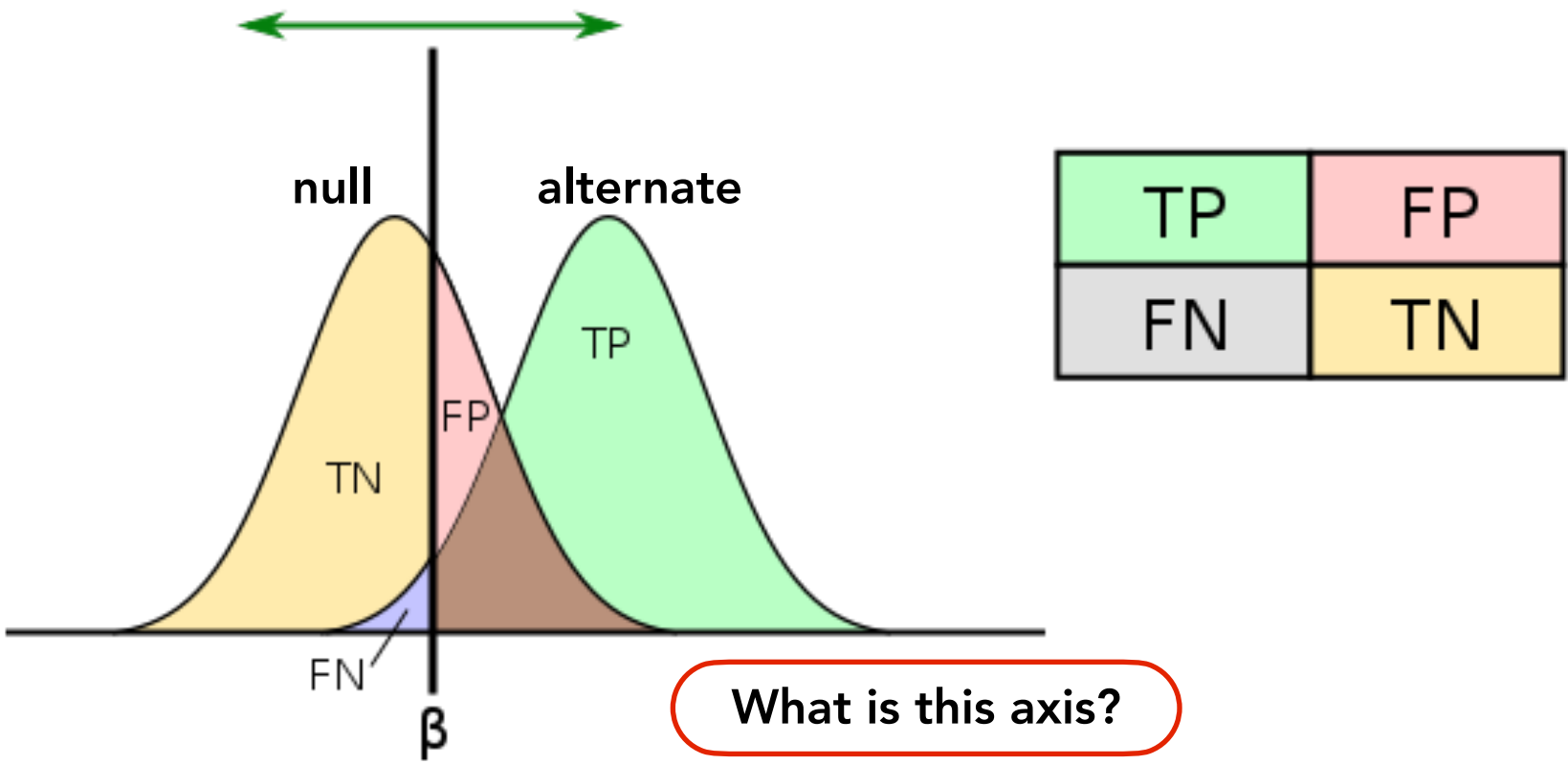
Supplemental Studies for Simultaneous Goodness-of-Fit Testing

Dr. Wolfgang Rolke, Dept. of Mathematical Sciences, University of Puerto Rico
 December 7, 2020

Abstract

Testing to see whether a given data set comes from some specified distribution is among the oldest types of problems in Statistics. Many such tests have been developed and their performance studied. The general result has been that while a certain test might perform well, aka have good power, in one situation it will fail badly in others. This is not a surprise given the great many ways in which a distribution can differ from the one specified in the null hypothesis. It is therefore very difficult to decide a priori which test to use. The obvious solution is not to rely on any one test but to run several of them. This however leads to the problem of simultaneous inference, that is, if several tests are done even if the null hypothesis were true, one of them is likely to reject it anyway just by random chance. In this paper we present a method that yields a p value that is uniform under the null hypothesis no matter how many tests are run. This is achieved by adjusting the p value via simulation. We present a number of simulation studies that show the uniformity of the p value and others that show that this test is superior to any one test if the power is averaged over a large number of cases.

Keywords: Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilk, Neyman Smooth test, Power, Monte Carlo Simulation

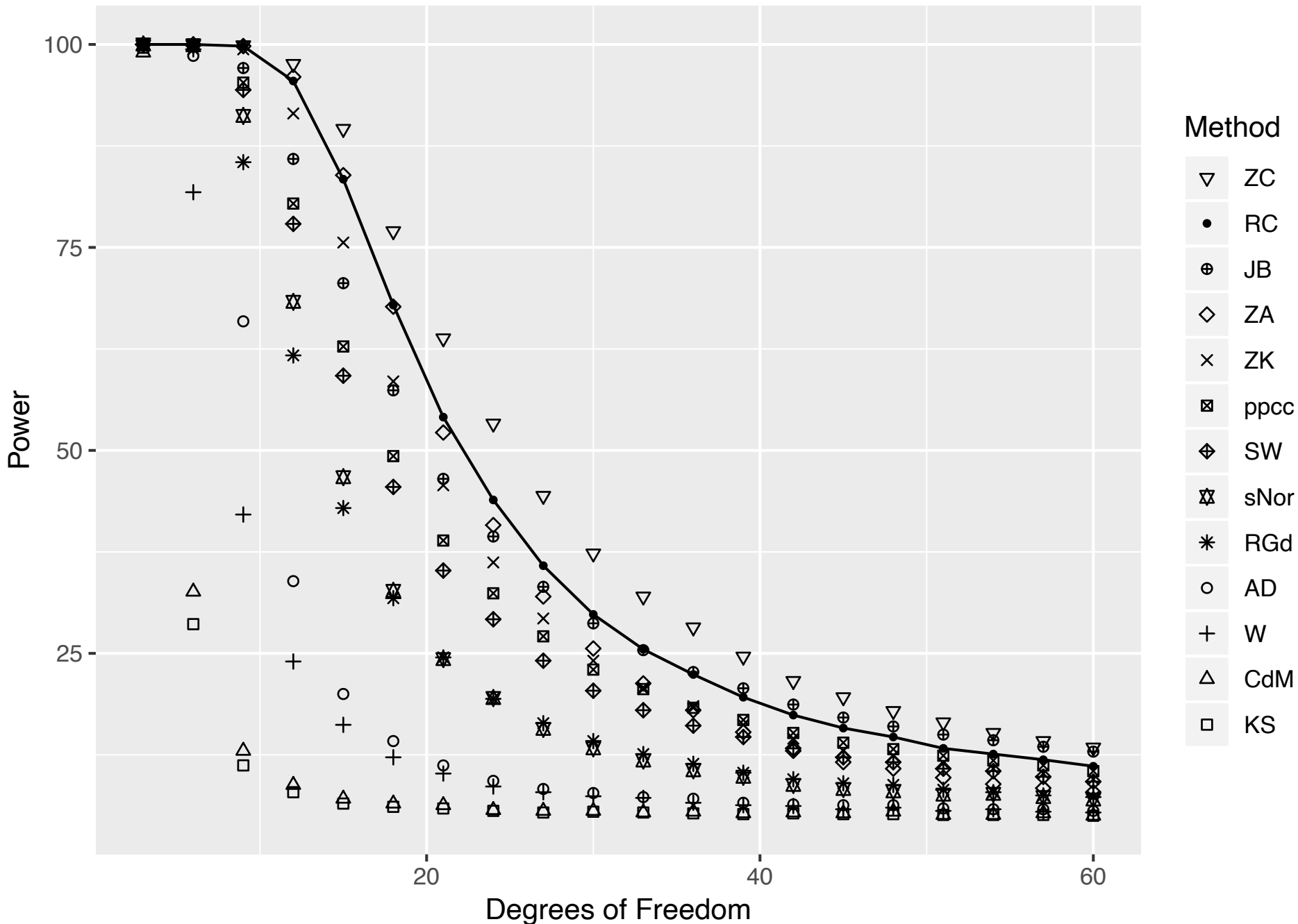


What is this axis?

Goodness-of-fit

Problem: There is no unique choice for the test statistic, giving rise to a large number of goodness-of-fit tests

- Can ask about the "power" of a GoF test to detect a given alternate



Supplemental Studies for Simultaneous Goodness-of-Fit Testing

Dr. Wolfgang Rolke, Dept. of Mathematical Sciences, University of Puerto Rico

December 7, 2020

Abstract

Testing to see whether a given data set comes from some specified distribution is among the oldest types of problems in Statistics. Many such tests have been developed and their performance studied. The general result has been that while a certain test might perform well, aka have good power, in one situation it will fail badly in others. This is not a surprise given the great many ways in which a distribution can differ from the one specified in the null hypothesis. It is therefore very difficult to decide a priori which test to use. The obvious solution is not to rely on any one test but to run several of them. This however leads to the problem of simultaneous inference, that is, if several tests are done even if the null hypothesis were true, one of them is likely to reject it anyway just by random chance. In this paper we present a method that yields a p value that is uniform under the null hypothesis no matter how many tests are run. This is achieved by adjusting the p value via simulation. We present a number of simulation studies that show the uniformity of the p value and others that show that this test is superior to any one test if the power is averaged over a large number of cases.

Keywords: Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilk, Neyman Smooth test, Power, Monte Carlo Simulation

The Neyman-Pearson Lemma

In 1928-1938 Neyman & Pearson developed a theory in which one must consider competing Hypotheses:

- the Null Hypothesis H_0 (background only)
- the Alternate Hypothesis H_1 (signal-plus-background)

Given some probability that we wrongly reject the Null Hypothesis

$$\alpha = P(x \notin W | H_0)$$

(Convention: if data falls in W then we accept H_0)

Find the region W such that we minimize the probability of wrongly accepting the H_0 (when H_1 is true)

$$\beta = P(x \in W | H_1)$$

The Neyman-Pearson Lemma

The region W that minimizes the probability of wrongly accepting H_0 is just a contour of the Likelihood Ratio

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

Any other region of the same size will have less power

The likelihood ratio is an example of a **Test Statistic**, eg. a real-valued function that summarizes the data in a way relevant to the hypotheses that are being tested

The Neyman-Pearson Lemma

You can also read the Neyman-Pearson lemma backwards to reverse engineer an alternate "H₁" for which a given GoF test statistic $T(x)$ is powerful

$$\frac{p(x | H_1)}{p(x | H_0)} = LR(x) \quad \longrightarrow \quad p(x | \text{"}H_1\text{"}) \propto T(x) \cdot p(x | H_0)$$

An impossibility result

However, for any GoF test statistic $T(x)$ there is also an entire family of alternates where the distribution of $T(x)$ is the same as for the null

- e.g. the GoF test is just doing random guessing
- See: <https://arxiv.org/abs/2107.06908>

Applies to “out-of-distribution detection” and “anomaly detection” as well

Understanding Failures in Out-of-Distribution Detection with Deep Generative Models

Lily H. Zhang¹ Mark Goldstein¹ Rajesh Ranganath¹

2.1. OOD Detection as Goodness-of-fit Testing

In its unconstrained form, OOD detection can be formalized as a single-sample hypothesis test (Nalisnick et al., 2019b; Serrà et al., 2020; Wang et al., 2020); given a sample \mathbf{x} , the test decides whether to reject the null hypothesis that a sample was drawn from the data distribution P , in favor of an alternative hypothesis that the sample came from a distribution other than P :

$$\begin{aligned} H_0 &: \mathbf{x} \sim P \\ H_A &: \mathbf{x} \sim Q \in \mathcal{Q}, P \notin \mathcal{Q}. \end{aligned}$$

2.2. OOD Detection as a Single-Sample Distributional Test is Impossible

OOD detection defined as a single-sample goodness-of-fit test is a challenging classification task given that the out-distributions are unknown. To remove the effect of misestimation, we consider test statistics which can use knowledge of the true in-distribution P via its density or probability function, denoted $\phi_p : \mathcal{X} \rightarrow \mathbb{R}$. We now present an impossibility result: no test can do well against all alternatives.

Proposition 1. *Let P be the distribution under the null hypothesis H_0 . Let μ be the measure associated with the distribution of test statistic $\phi_p(\mathbf{x})$ under the null. Then, assuming the conditional $\mathbf{x} | \phi_p(\mathbf{x})$ is not degenerate on a μ -non-measure zero set, there exists a set of alternative distributions $Q \in \mathcal{Q}$ where $Q \neq_d P$ and the test has power equal to the false positive rate. In other words, the test does no better than random guessing.*

Proof. See Appendix A. The proof sketch is as follows: First we construct distributions $Q \in \mathcal{Q}$ for which the distribution of $\phi_p(\mathbf{x})$ is the same but the distribution of $\mathbf{x} | \phi_p(\mathbf{x})$ differs when $\mathbf{x} \sim P$ and $\mathbf{x} \sim Q$ for all $\phi_p(\mathbf{x})$ in a non-measure-zero set Φ . This implies $q(\mathbf{x}) \neq_d p(\mathbf{x})$. We show that the power of the test for any rejection rule for such a pair P, Q is equal to the false positive rate for all false positive rates, which is equivalent to random guessing. \square

No Free Lunch



No Free Lunch

No Free Lunch Theorem for Search & Optimization

A similar impossibility theorem exists in machine learning (formulated here as an optimization problem or search)

- “any two optimization algorithms are equivalent when their performance is averaged across all possible problems”.
- “We have dubbed the associated results NFL theorems because they demonstrate that if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.”

The Lack of A Priori Distinctions Between Learning Algorithms

David H. Wolpert

*The Santa Fe Institute, 1399 Hyde Park Rd.,
Santa Fe, NM, 87501, USA*

This is the first of two papers that use off-training set (OTS) error to investigate the assumption-free relationship between learning algorithms. This first paper discusses the senses in which there are no a priori distinctions between learning algorithms. (The second paper discusses the senses in which there are such distinctions.) In this first paper it is shown, loosely speaking, that for any two algorithms A and B, there are “as many” targets (or priors over targets) for which A has lower expected OTS error than B as vice versa, for loss functions like zero-one loss. In particular, this is true if A is cross-validation and B is “anti-cross-validation” (choose the learning algorithm with largest cross-validation error). This paper ends with a discussion of the implications of these results for computational learning theory. It is shown that one *cannot* say: if empirical misclassification rate is low, the Vapnik–Chervonenkis dimension of your generalizer is small, and the training set is large, then with high probability your OTS error is small. Other implications for “membership queries” algorithms and “punting” algorithms are also discussed.

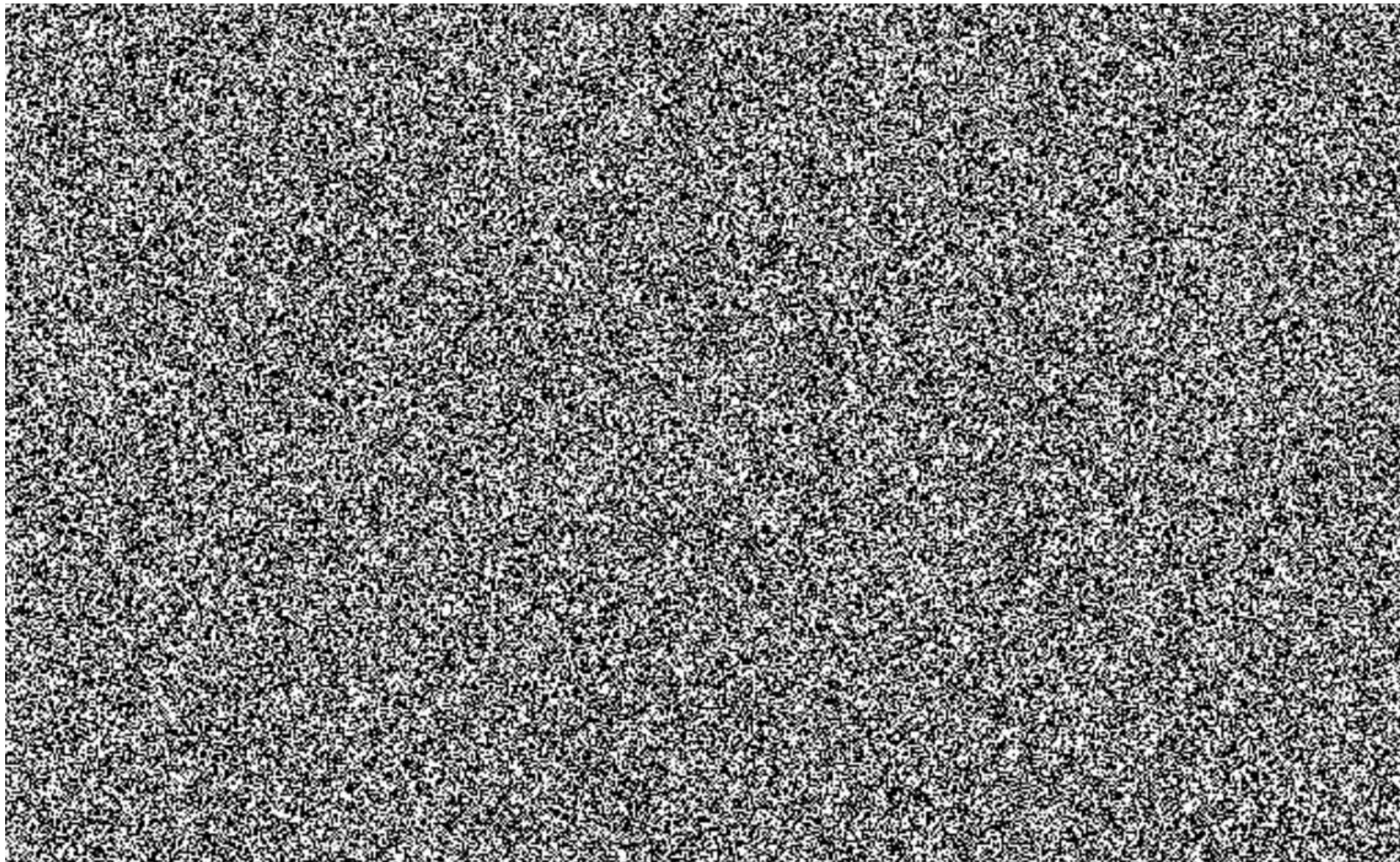
“Even after the observation of the frequent conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience.”
David Hume, in *A Treatise of Human Nature*, Book I, part 3, Section 12.

Real world data has structure

But real world data has structure inherited from the causal mechanism that generated it

- If we bias our models away from irrelevant, unphysical possibilities we can do better

Random image
(no structure)



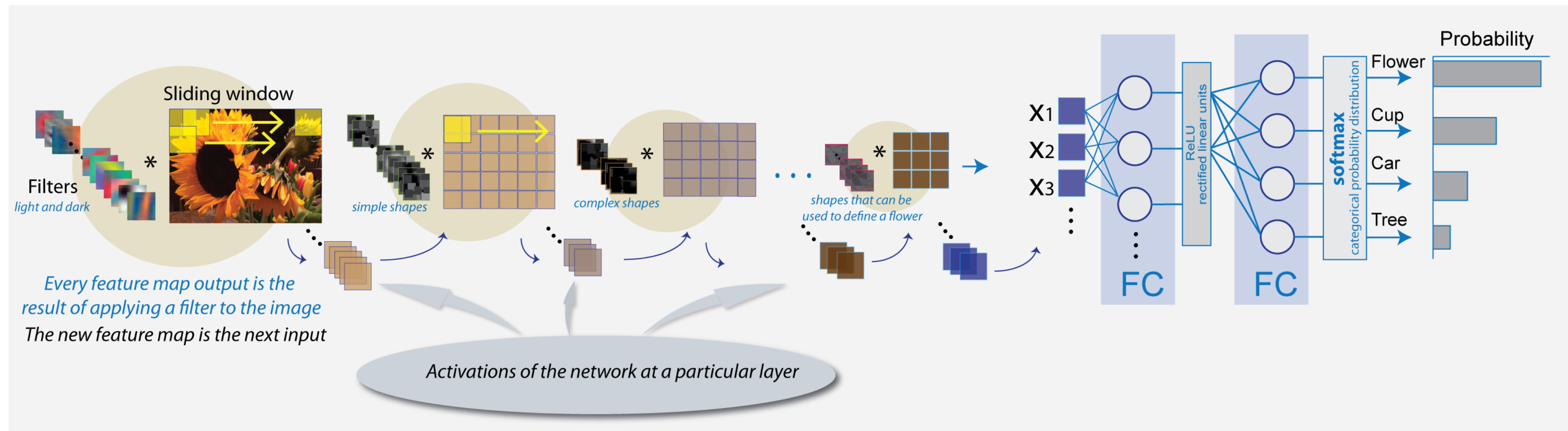
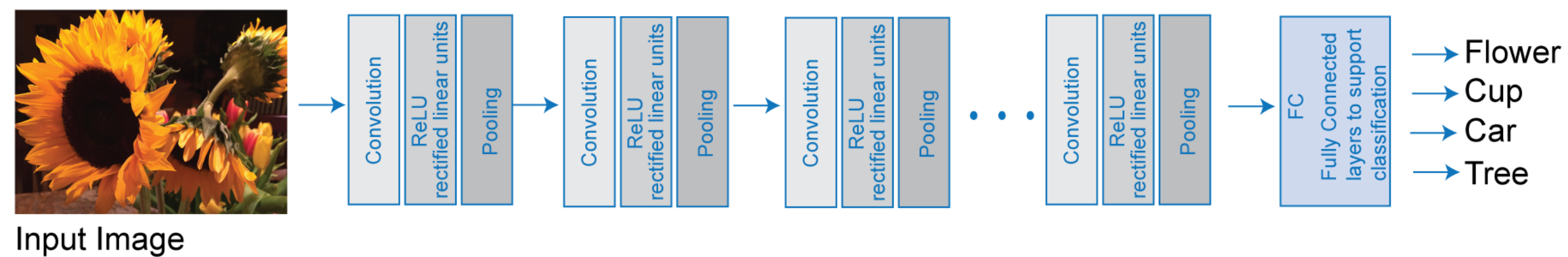
Natural image
(rich structure)



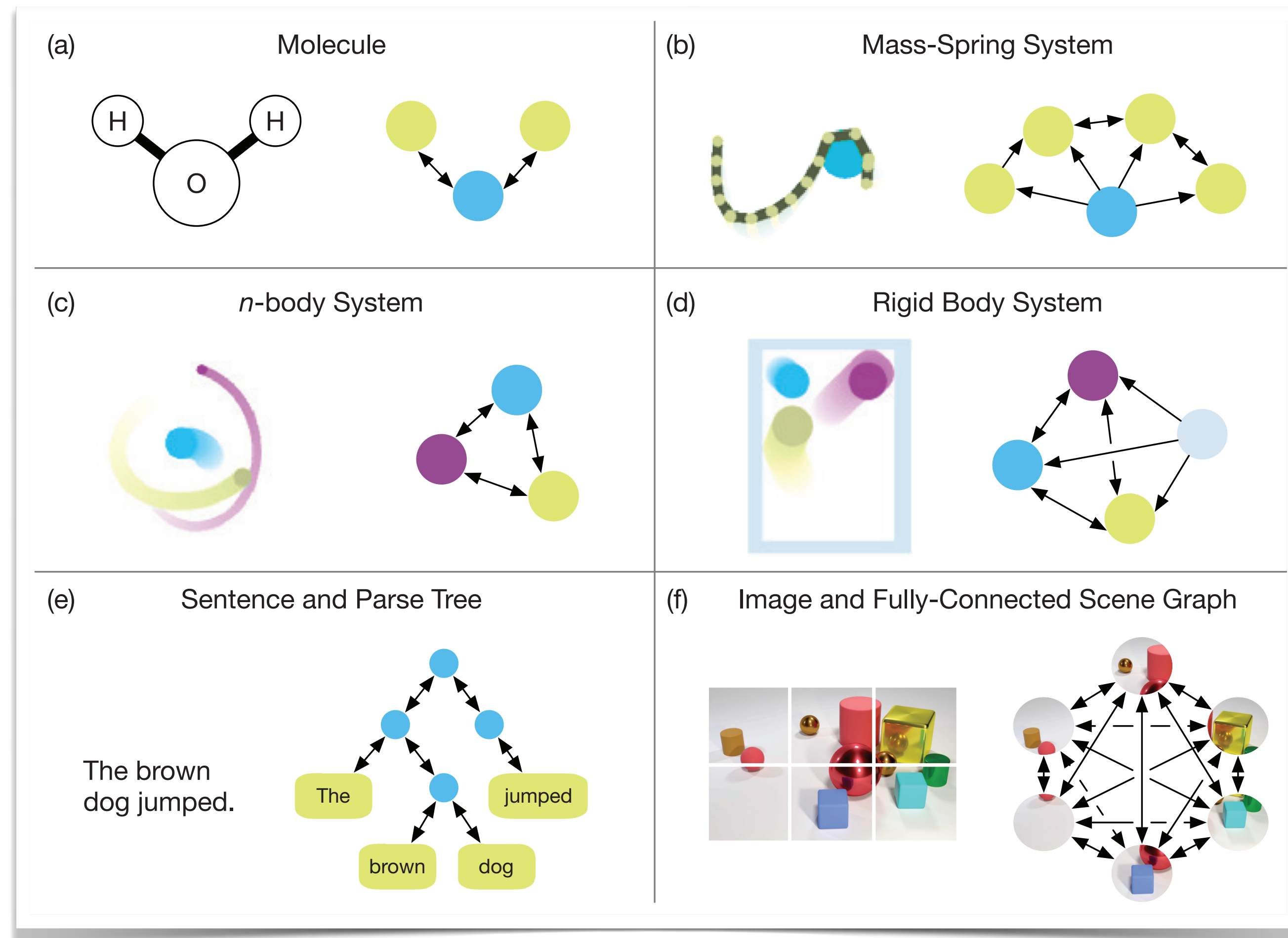
Inductive Bias

A major tool of deep learning: convolutional neural networks

- the world is compositional \Rightarrow hierarchical architecture
- images are translationally invariant \Rightarrow convolutions



Insight of data generating process informs inductive bias on architecture



Inductive Bias

Compositionality

Relationships

Symmetry

Causality



Is bias bad?

The term “bias”

The term “**bias**” is highly overloaded and used in many ways

- Generally, “**bias**” carries a negative connotation or is pejorative
- “To be biased” is considered bad

In the context of model-independent searches, we often hear “bias” being used informally, e.g.

- “We chose a model-independent approach to avoid theory bias”
- “To remove theory bias and model-dependence in ...”

But what does this mean more formally? ... and is it bad?

Estimators & Bias

Given some statistical model $p(x|\alpha)$ and a set of observations $\{x_i\}$ often one wants to estimate the true value of α (assuming the model is true).

An **estimator** is function of the data written $\hat{\alpha}(x_1, \dots, x_n)$

- Since the data are random, so is the resulting estimate
- one can compute **expectation** of the estimator $E[\hat{\alpha}(x)|\alpha] = \int \hat{\alpha}(x)f(x|\alpha)dx$

Properties of estimators:

- **bias** $E[\hat{\alpha}(x)|\alpha] - \alpha$ ("unbiased" means bias of estimator 0 for all true α)
- **variance** $E[(\hat{\alpha}(x) - \alpha)^2|\alpha] = \int (\hat{\alpha}(x) - \alpha)^2 f(x|\alpha)dx$

Bias as a term

Relaxing the language a bit, one might think of “bias” as:

- When average result from procedure doesn't recover the ground-truth target
- Preferring **a priori** one option to another without explicit evidence

In a Bayesian language, one would usually use Bayes theorem

- $\text{Posterior}(\text{theory} \mid \text{data}) \propto \text{Likelihood}(\text{data} \mid \text{theory}) \cdot \text{Prior}(\text{theory})$
- In general, Bayesian approaches are “biased” towards the prior

Cramér-Rao Bound

The minimum variance bound on an unbiased estimator is given by the Cramér-Rao bound:

$$\text{cov}[\hat{\theta}|\theta_0]_{ij} \geq I_{ij}^{-1}(\theta_0)$$

Expected error of best-fit parameter Inverse of Fisher information

Where I is the Fisher information matrix

$$I_{ij}[\theta] = -\mathbb{E} \left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta_i \partial \theta_j} \middle| \theta \right]$$

Maximum Likelihood Estimators *asymptotically* reach this bound

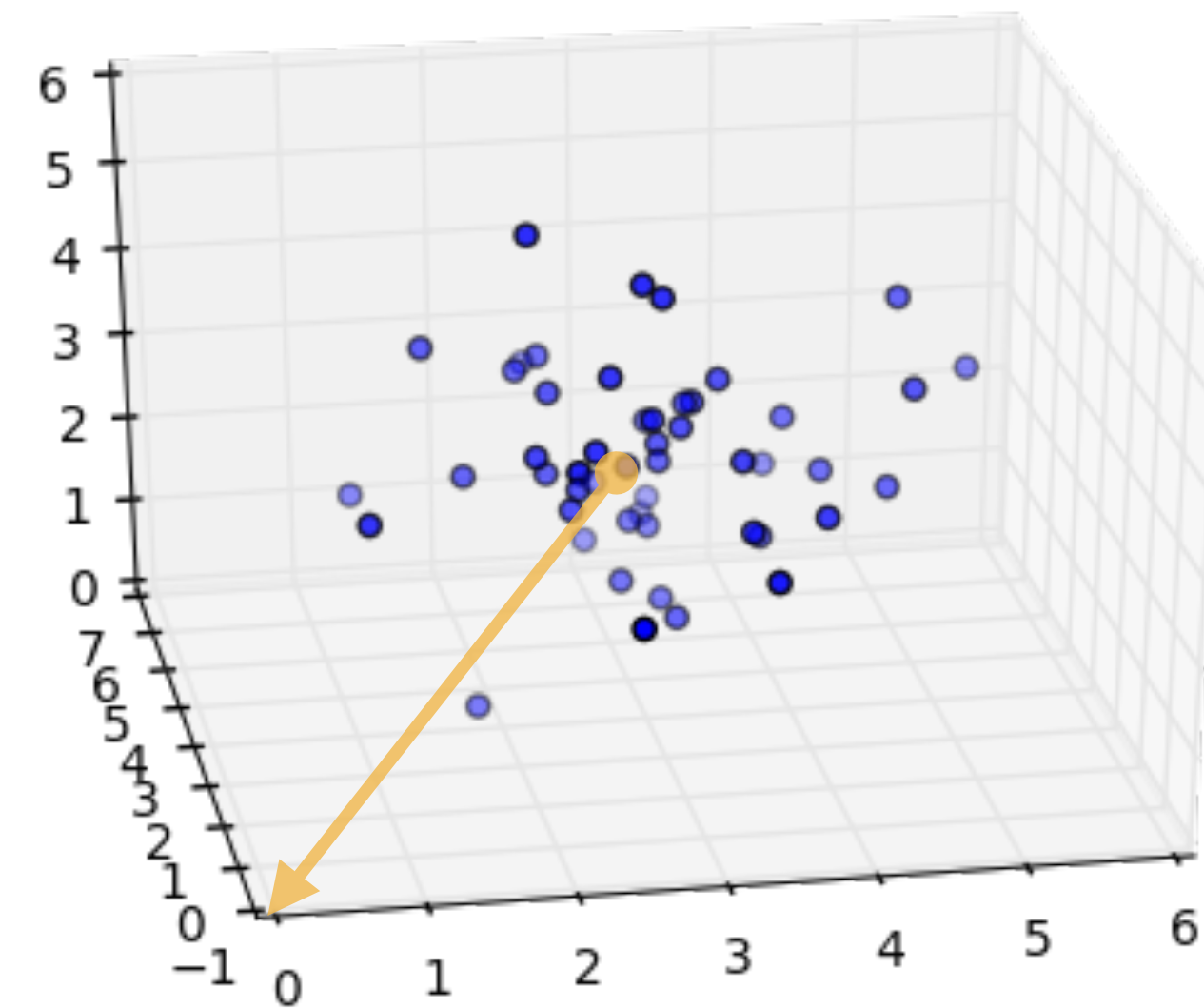
James-Stein Estimator

Consider a standard multivariate Gaussian distribution for \vec{x} in n dimensions centered around $\vec{\mu}$

$$f(\vec{x}|\vec{\mu}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_i)^2}{2}\right).$$

Goal: minimize mean-squared error

$$MSE[\hat{\vec{\mu}}] = E[||\hat{\vec{\mu}} - \vec{\mu}||^2]$$



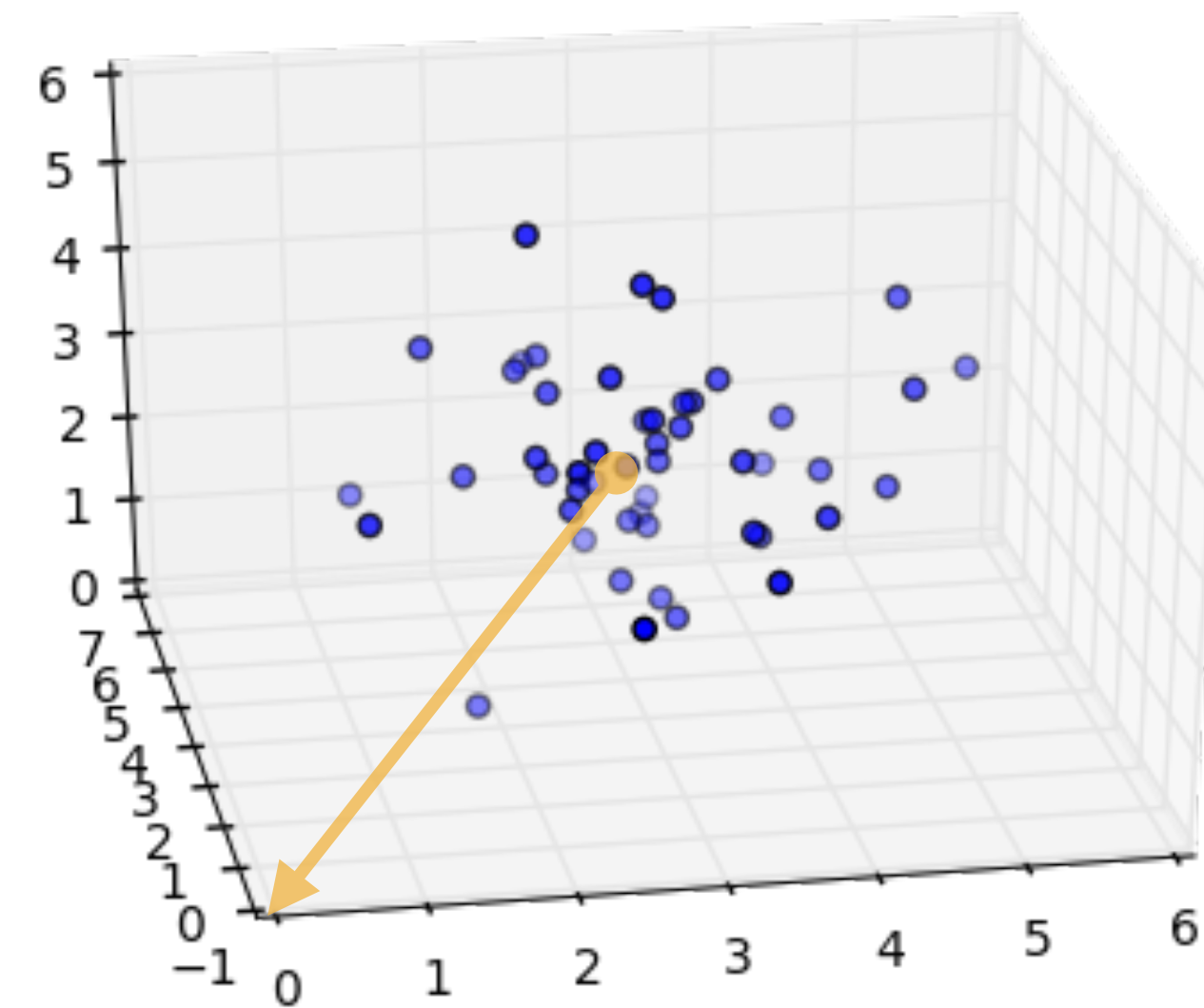
James-Stein Estimator

Consider a standard multivariate Gaussian distribution for \vec{x} in n dimensions centered around $\vec{\mu}$

$$f(\vec{x}|\vec{\mu}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_i)^2}{2}\right).$$

Goal: minimize mean-squared error

$$MSE[\hat{\vec{\mu}}] = E[||\hat{\vec{\mu}} - \vec{\mu}||^2]$$



MLE (unbiased)

$$\hat{\vec{\mu}}_{MLE} = \bar{\vec{x}} = \frac{1}{m} \sum_{j=1}^m \vec{x}_j$$

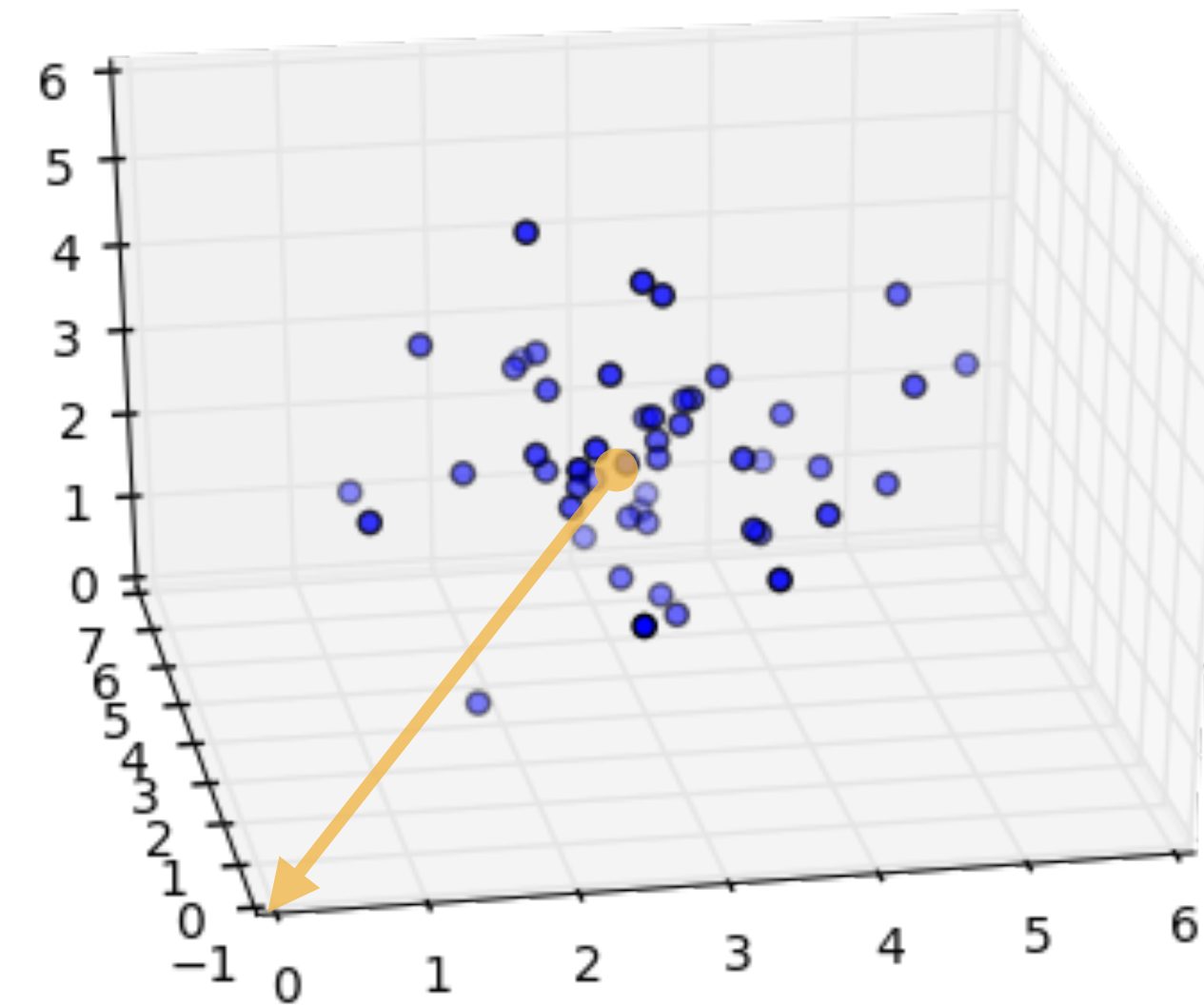
James-Stein Estimator

Consider a standard multivariate Gaussian distribution for \vec{x} in n dimensions centered around $\vec{\mu}$

$$f(\vec{x}|\vec{\mu}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_i)^2}{2}\right).$$

Goal: minimize mean-squared error

$$MSE[\hat{\vec{\mu}}] = E[||\hat{\vec{\mu}} - \vec{\mu}||^2]$$



MLE (unbiased)

$$\hat{\vec{\mu}}_{MLE} = \bar{\vec{x}} = \frac{1}{m} \sum_{j=1}^m \vec{x}_j$$

James-Stein (weird)

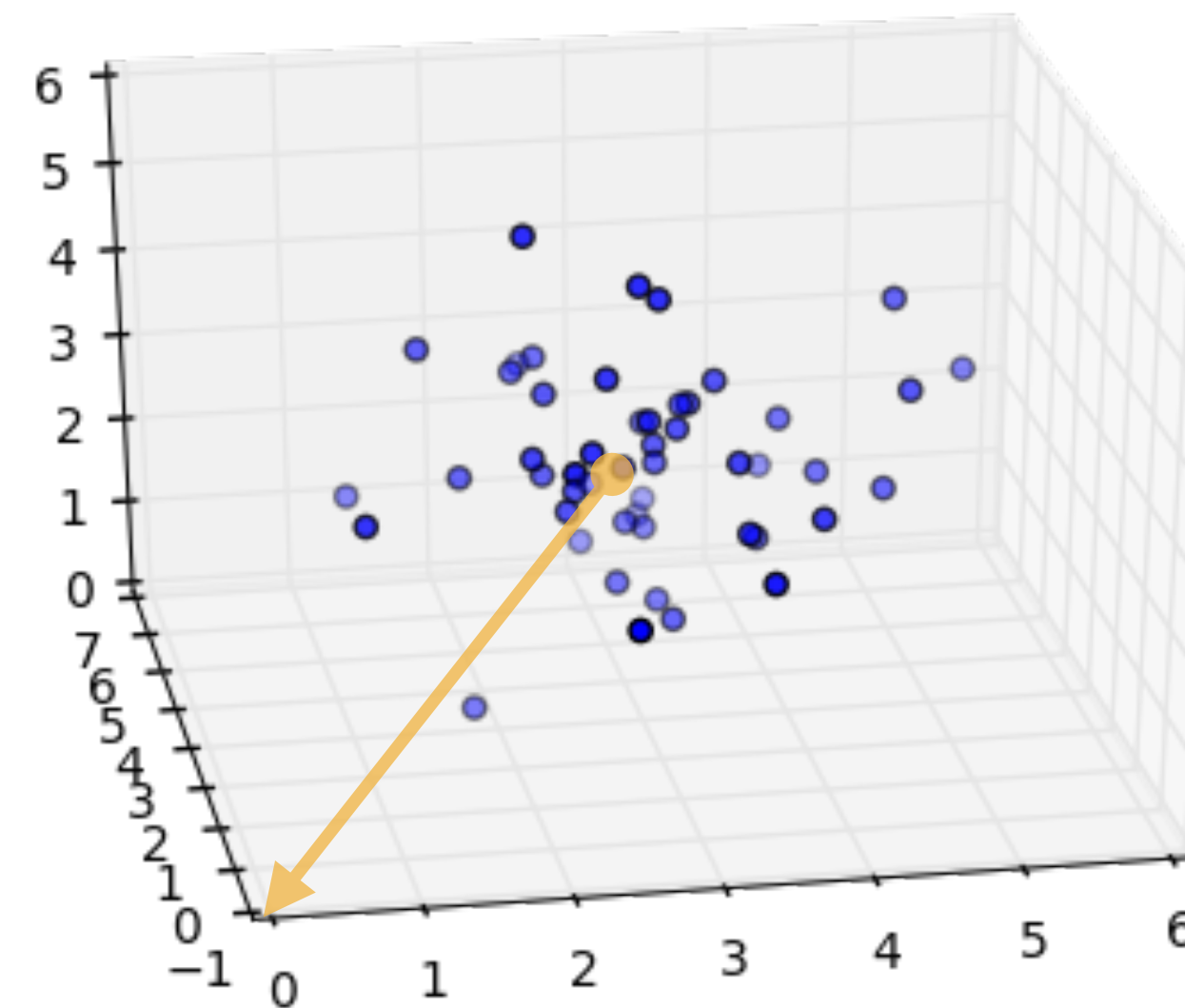
$$\hat{\vec{\mu}}_{JS} = \left(1 - \frac{n-2}{||\bar{\vec{x}}||^2}\right) \bar{\vec{x}}$$

James-Stein Estimator

The James-Stein estimator seems like a horrible suggestion

$$\hat{\mu}_{JS} = \left(1 - \frac{n-2}{\|\bar{x}\|^2}\right) \bar{x}$$

- clearly biased (MLE is not)
 - shifts towards origin is not translationally invariant
- $$x \rightarrow x' = x + \Delta$$



James-Stein Estimator

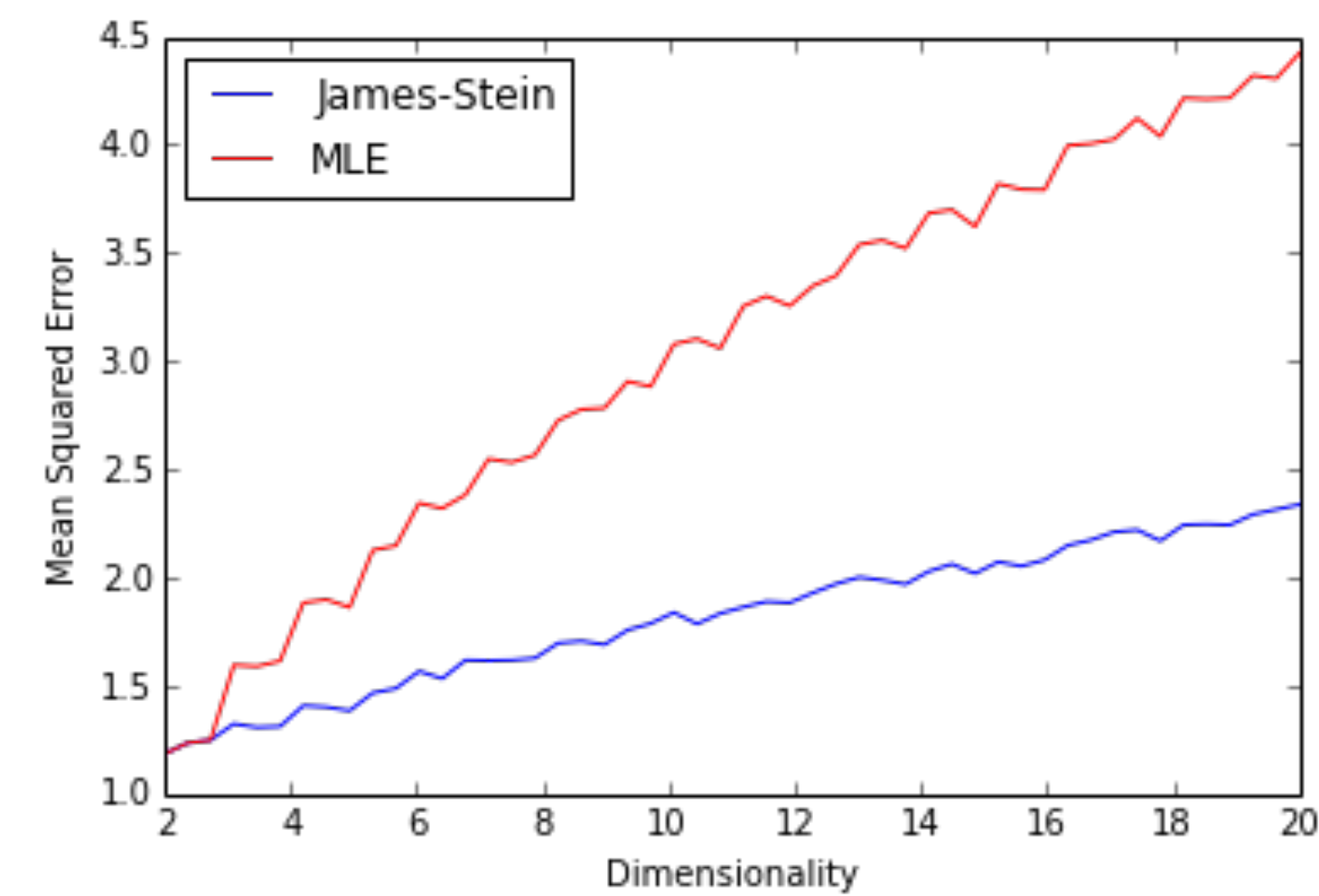
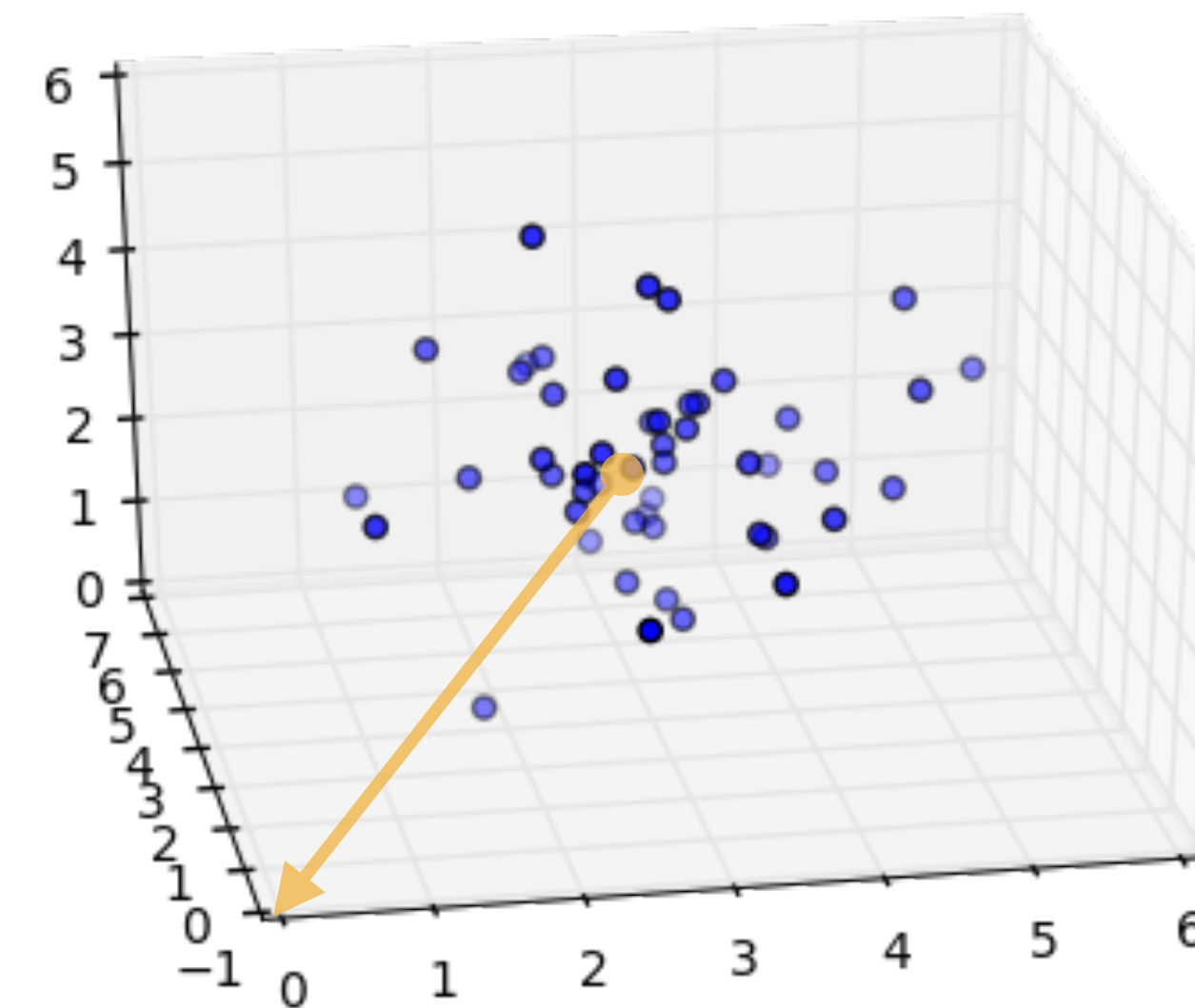
The James-Stein estimator seems like a horrible suggestion

$$\hat{\mu}_{JS} = \left(1 - \frac{n-2}{\|\bar{x}\|^2}\right) \bar{x}$$

- clearly biased (MLE is not)
 - shifts towards origin is not translationally invariant
- $$x \rightarrow x' = x + \Delta$$

Yet, it has smaller mean squared error than MLE for $n > 2$!

- it "dominates" the MLE



Bias - Variance Tradeoff

Best understood in terms of Bias - Variance tradeoff

Most physicist are allergic to the idea of a biased estimator

- try to find unbiased estimator with smallest variance
- hence importance of Cramér-Rao bound

But what if we just want to minimize the mean-squared error?

$$MSE[\hat{\mu}|\mu] = E[(\hat{\mu} - \mu)^2] | \mu]$$

it decomposes like this

$$MSE[\hat{\mu}|\mu] = \text{Var}[\hat{\mu}|\mu] + (\text{Bias}[\hat{\mu}|\mu])^2$$

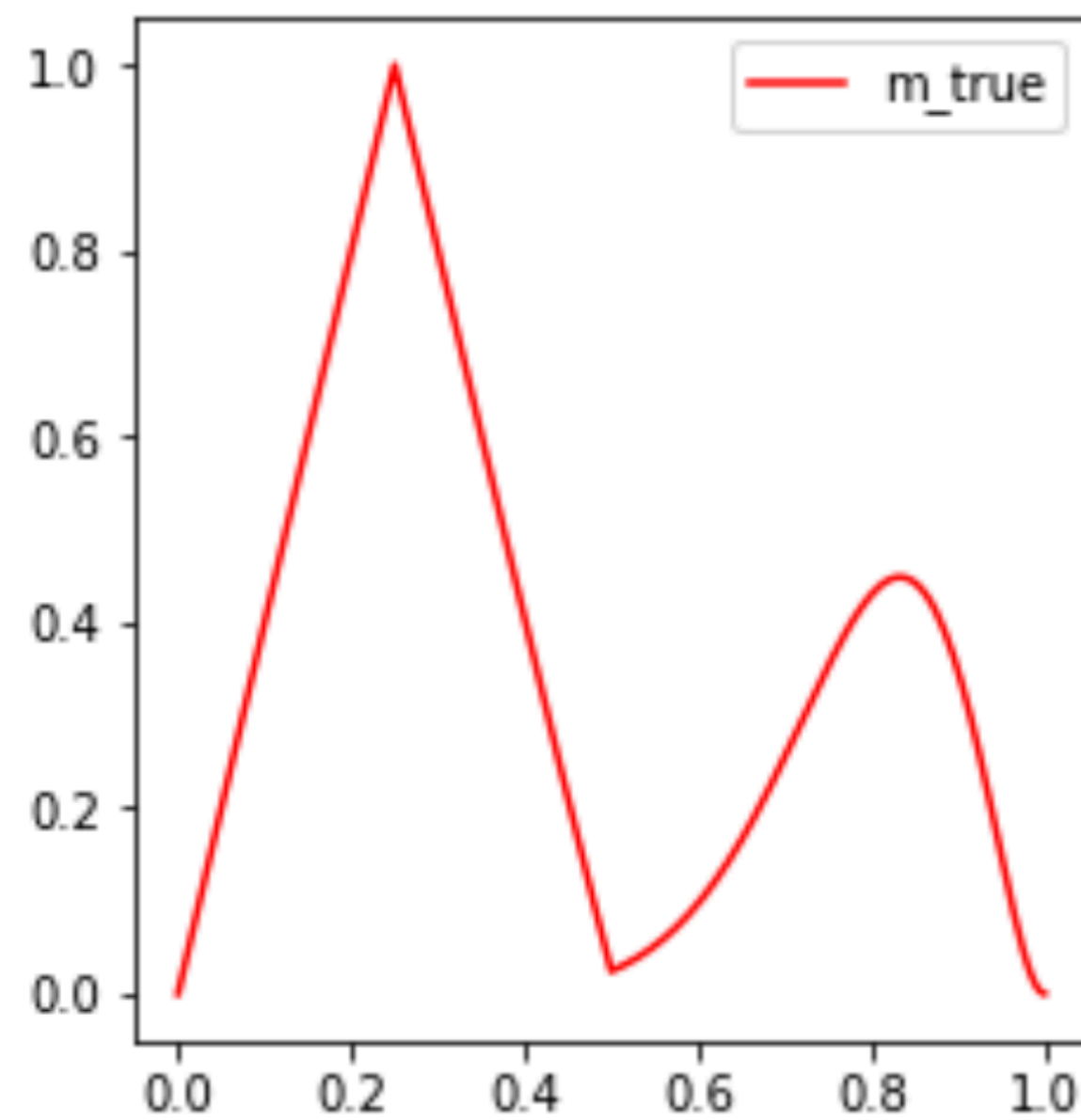
So it encodes some relative weight to bias and variance. **Need to think harder!**

Unfolding & Regularization

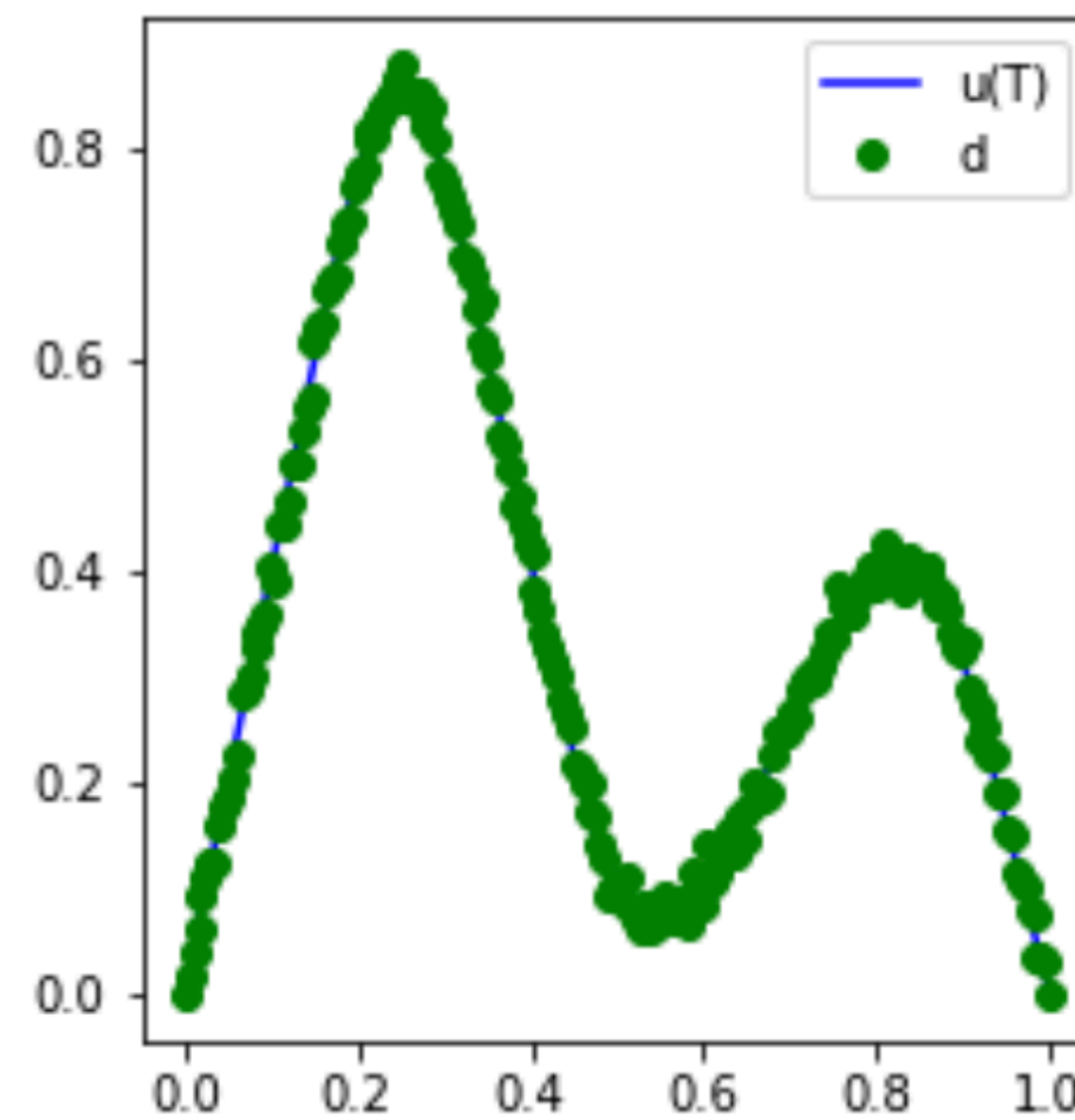
The maximum likelihood solution for an unfolding problem yields highly oscillatory solutions

- The inverse of transfer matrix has high condition number, the problem is “ill-posed”
- **Solution:** Tikhonov regularization
 - Yields more physical solutions, smaller MSE, but they are biased (bias-variance trade off)

True distribution

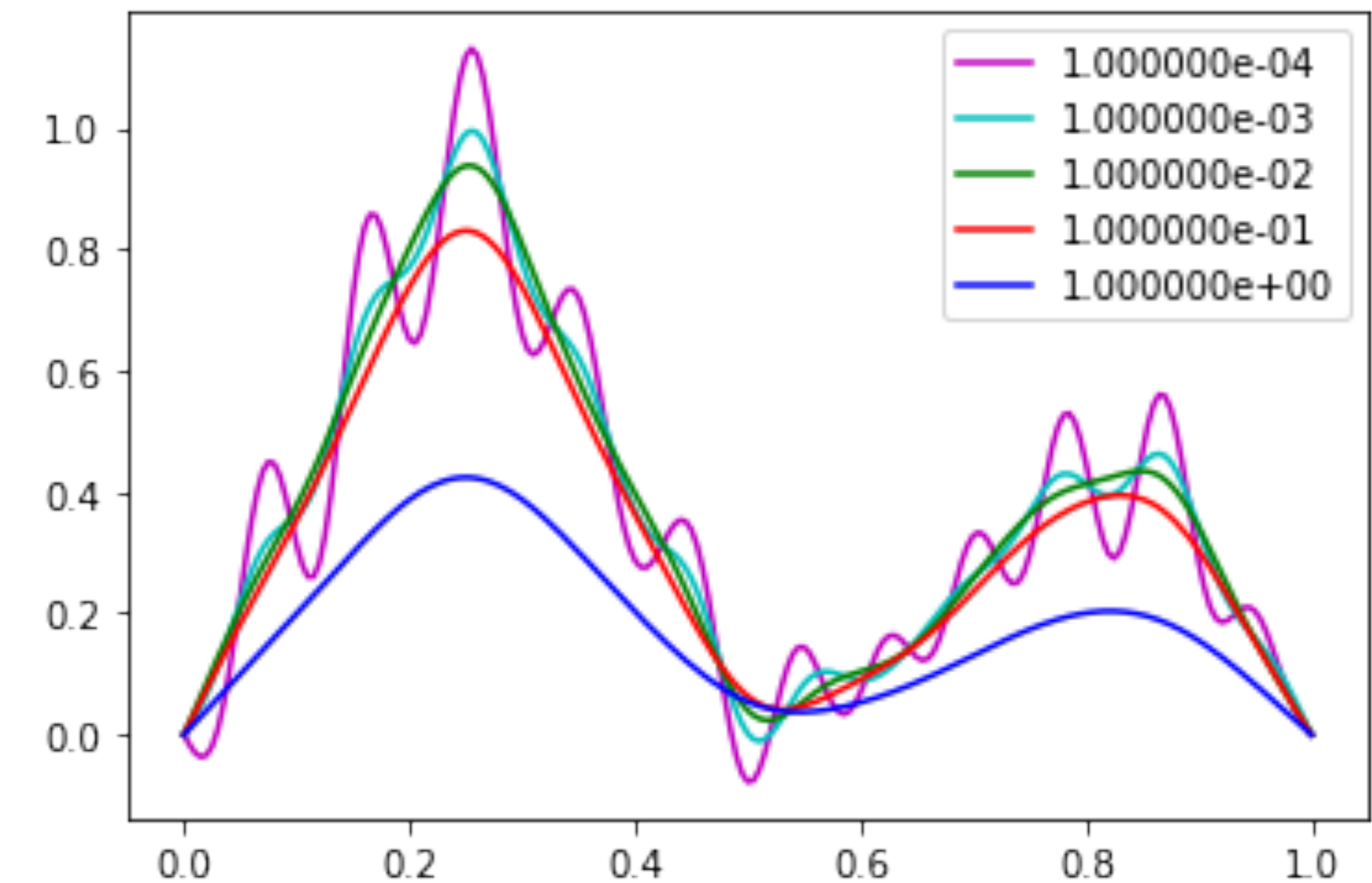


Noisy data



UNFOLDING

Unfolded distribution



Regularization

Fitting 10 data points to polynomials of degree M

- Intuitive example of "overfitting" if M is large
- Lower order polynomial a "hard" form of regularization / inductive bias
- Leads to bias if true solution isn't a low-order polynomial

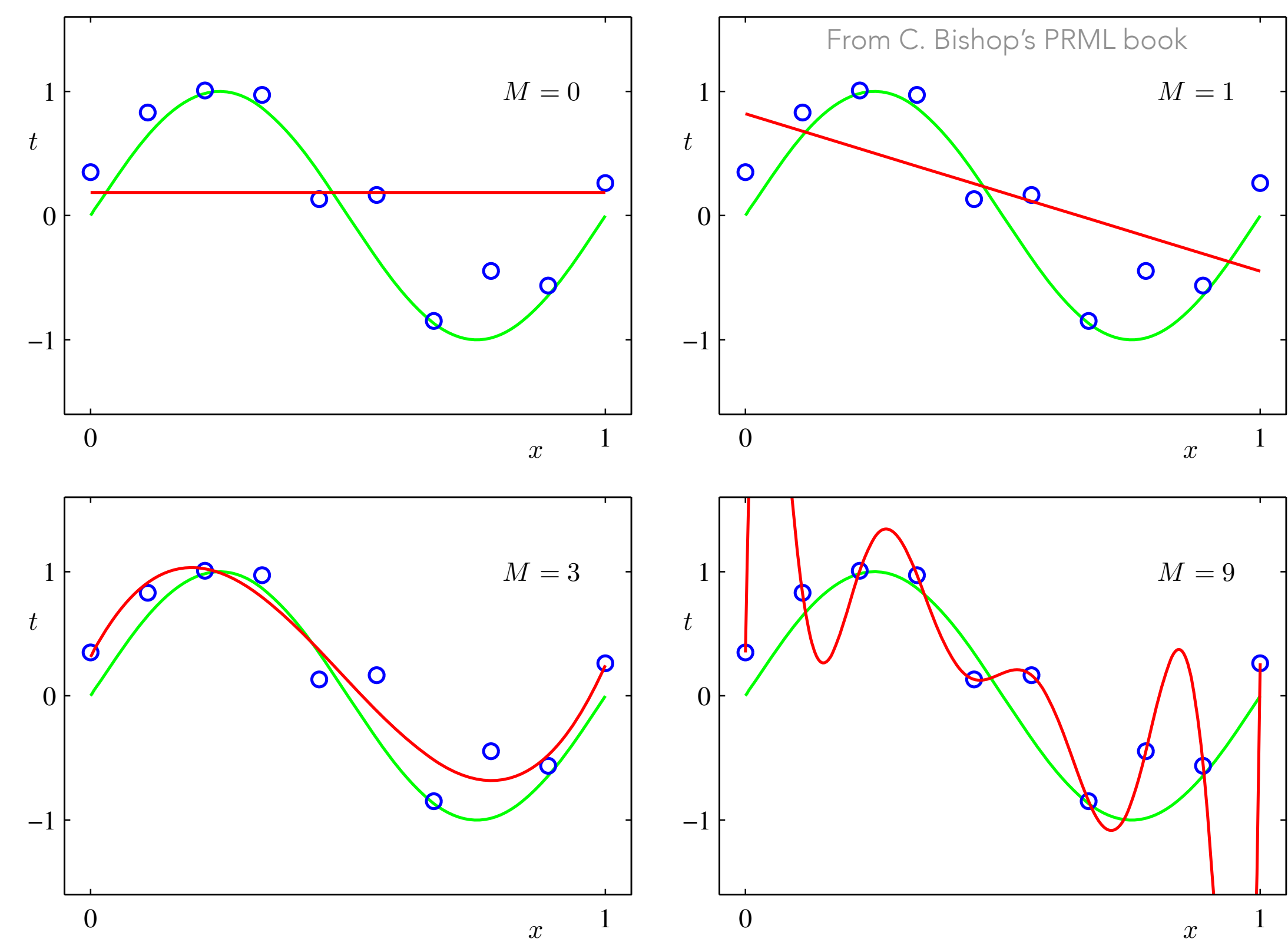


Figure 1.4 Plots of polynomials having various orders M , shown as red curves, fitted to the data set shown in Figure 1.2.

Regularization

Fitting 10 data points to polynomials of degree M

- Intuitive example of "overfitting" if M is large
- Lower order polynomial a "hard" form of regularization / inductive bias
- Leads to bias if true solution isn't a low-order polynomial

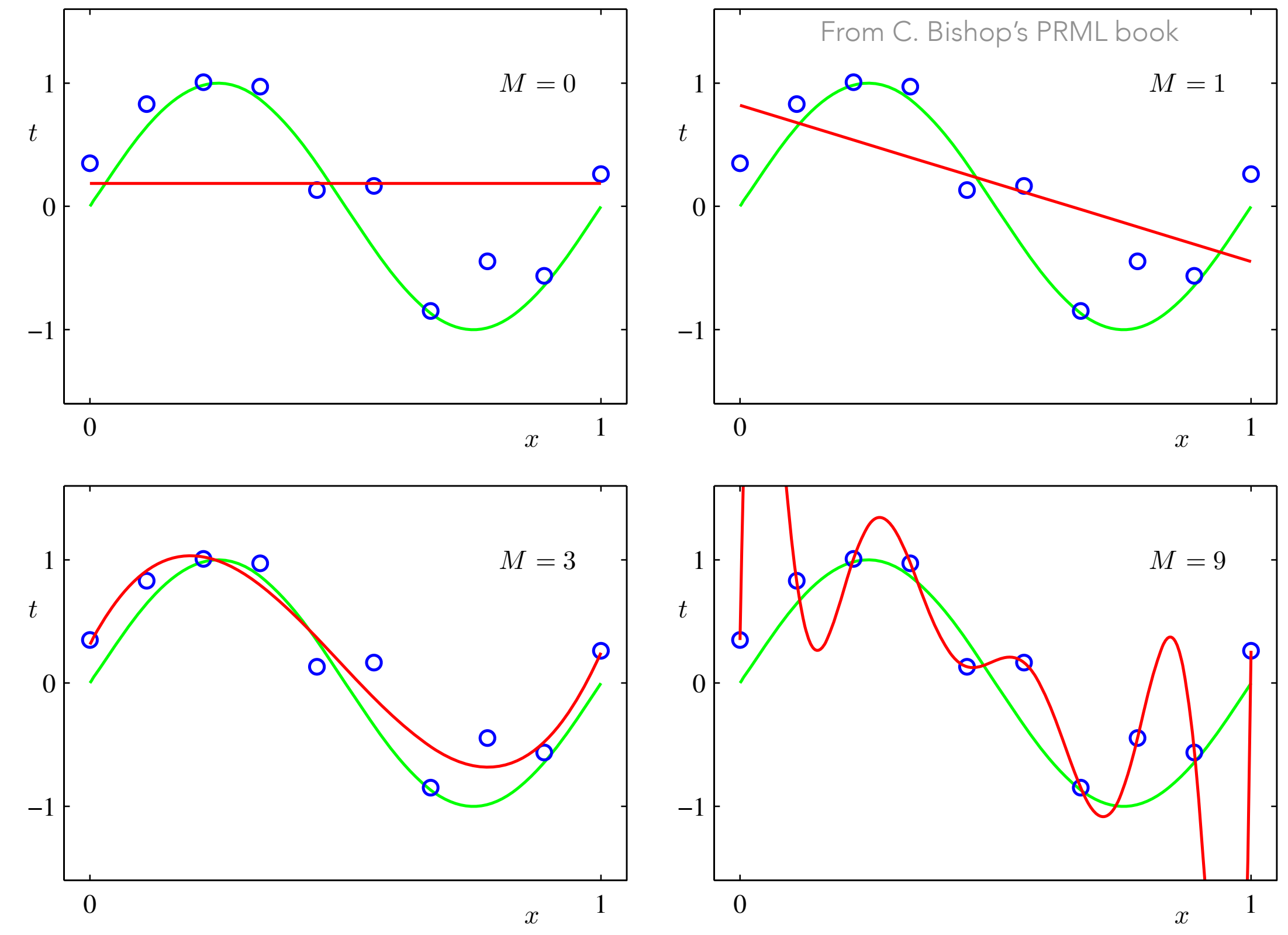


Figure 1.4 Plots of polynomials having various orders M , shown as red curves, fitted to the data set shown in Figure 1.2.

Alternatively, allow higher order polynomials and regularize coefficients \mathbf{w} to keep them small.

- Penalized least squares / ridge regression

- Shrinkage / bias $\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$

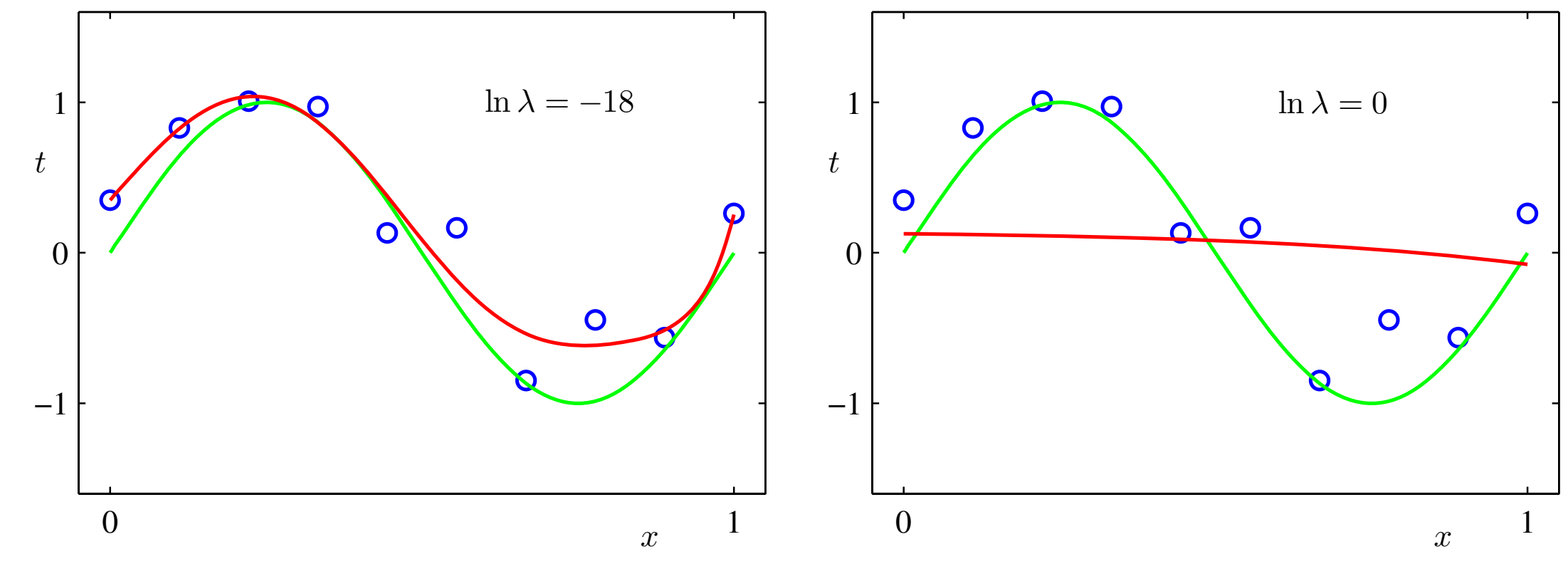


Figure 1.7 Plots of $M = 9$ polynomials fitted to the data set shown in Figure 1.2 using the regularized error function (1.4) for two values of the regularization parameter λ corresponding to $\ln \lambda = -18$ and $\ln \lambda = 0$. The case of no regularizer, i.e., $\lambda = 0$, corresponding to $\ln \lambda = -\infty$, is shown at the bottom right of Figure 1.4.

Gaussian Processes

A more extreme version of this strategy is to use Gaussian Processes

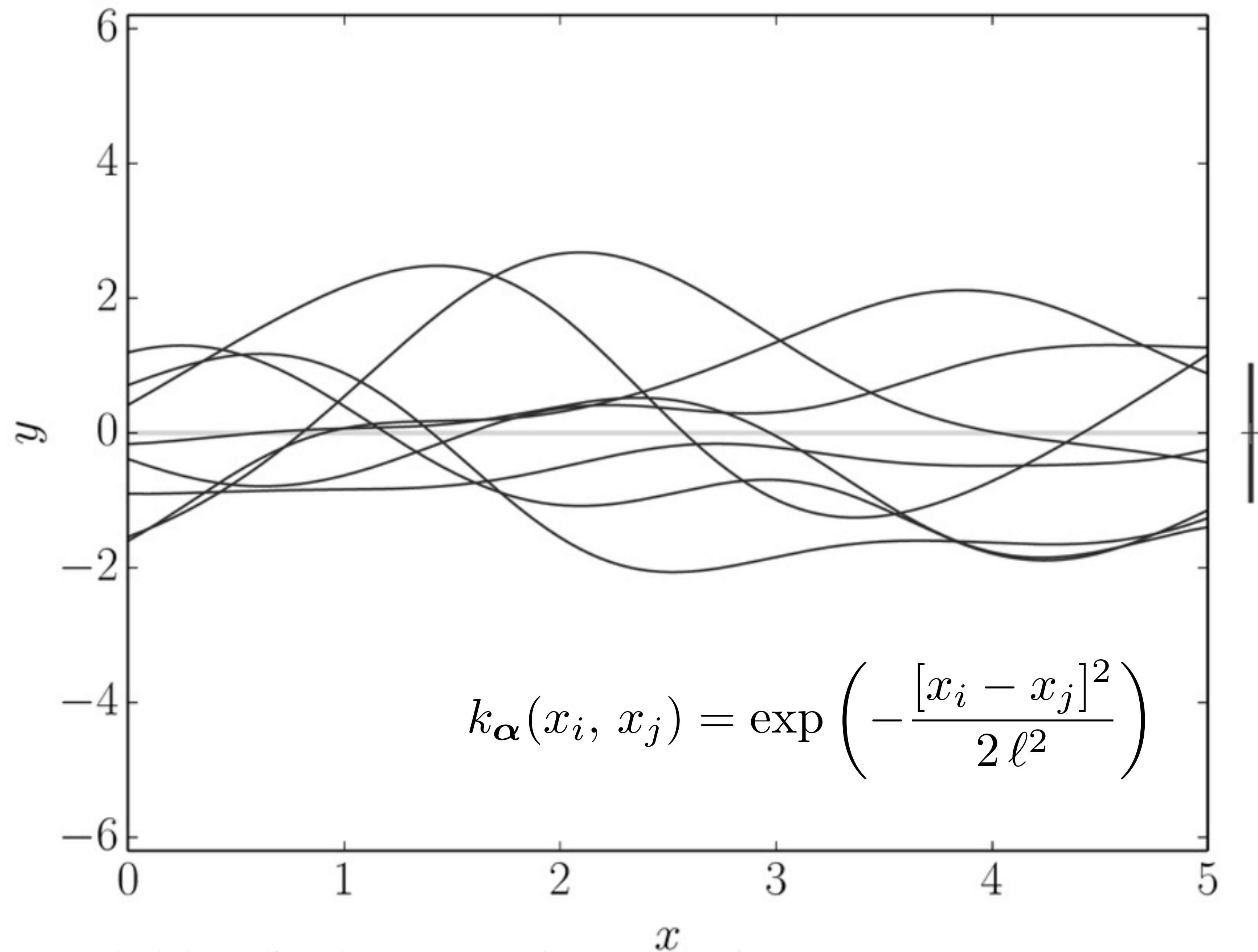
- Consider **all possible functions** (no analytic, parametric form assumed)
- Then put a prior over space of all possible functions defined by:
 - A Mean function $\mu(x)$
 - A covariance kernel $\Sigma(x, x')$ which quantifies $\text{cov}[f(x), f(x')]$
 - $f(x) \sim GP(\mu, \Sigma)$

Physicist then models the mean $\mu(x)$ and covariance kernel $\Sigma(x, x')$

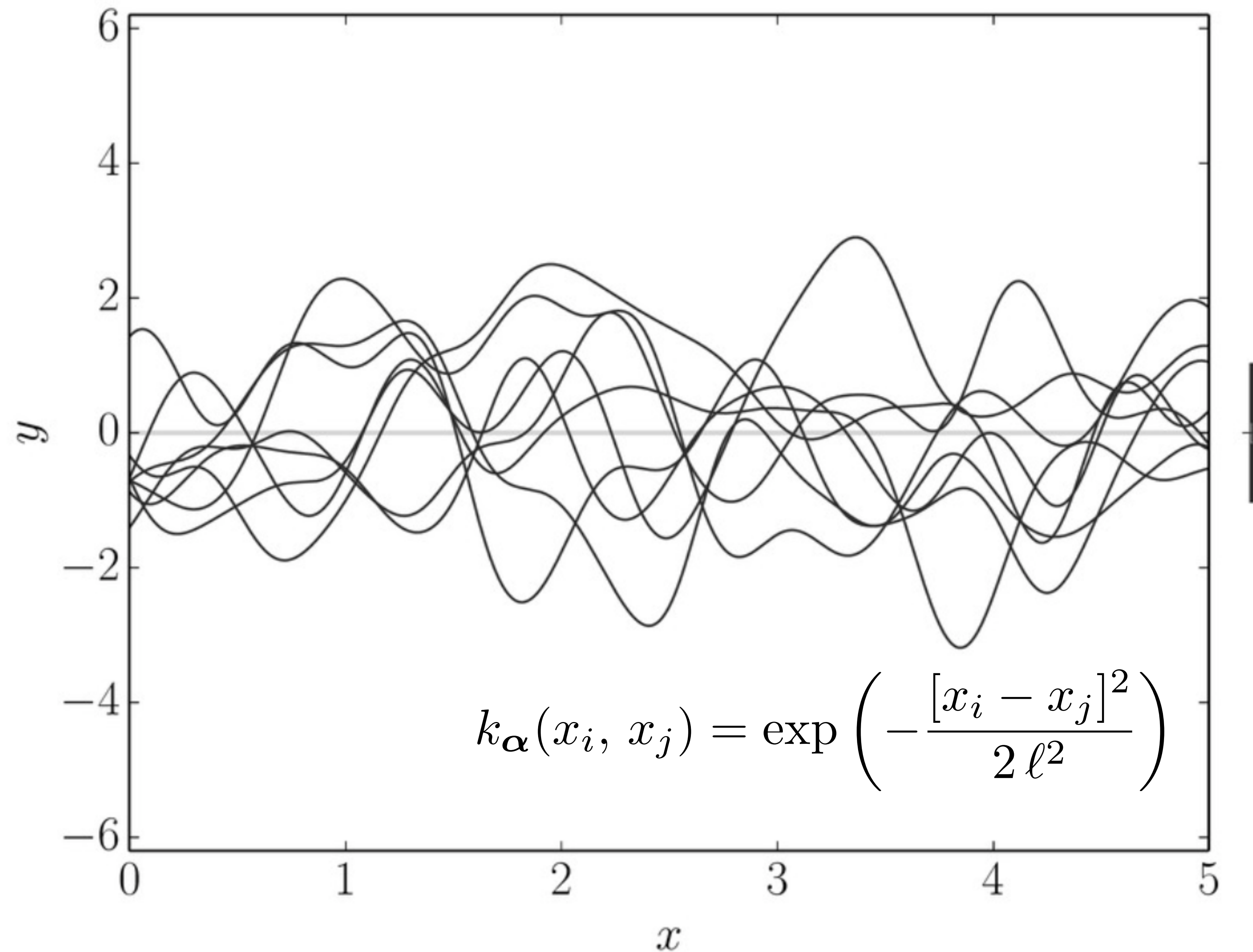
- Fit of GP model to data has explicit, unique answer (just linear algebra)

$$k_{\alpha}(x_i, x_j) = \exp\left(-\frac{[x_i - x_j]^2}{2\ell^2}\right)$$

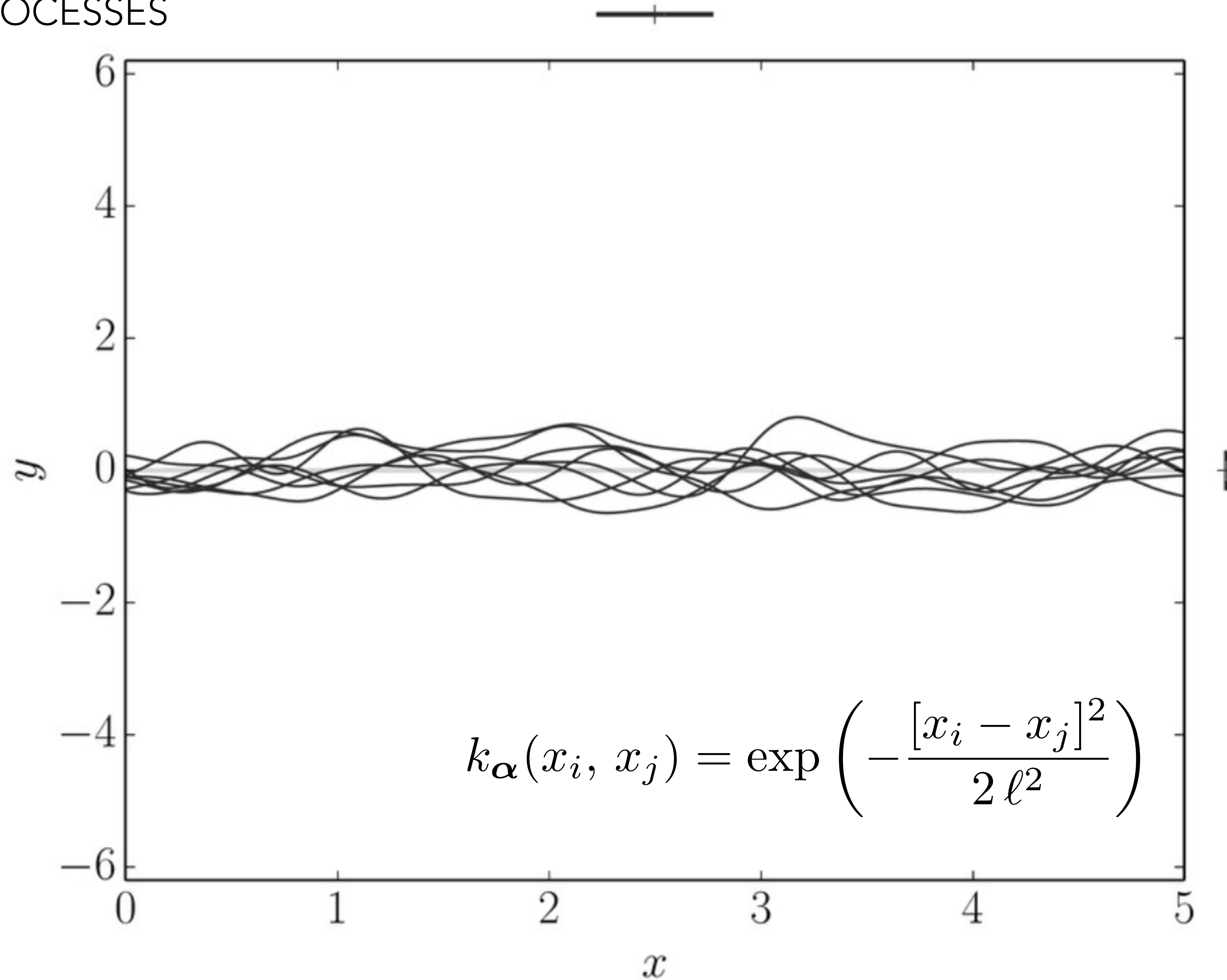
GAUSSIAN PROCESSES

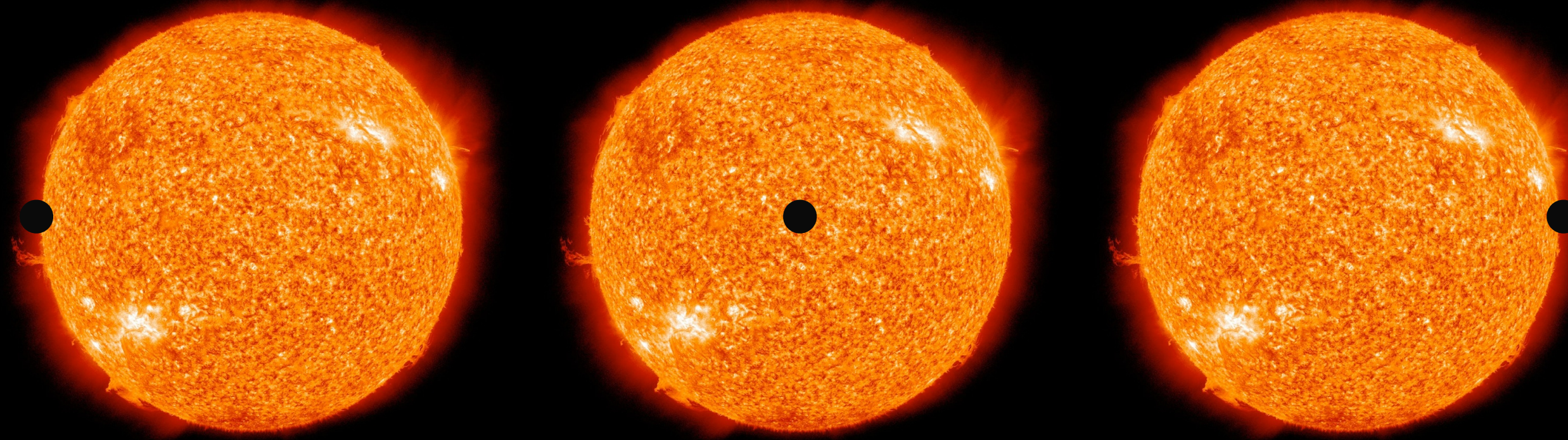


GAUSSIAN PROCESSES

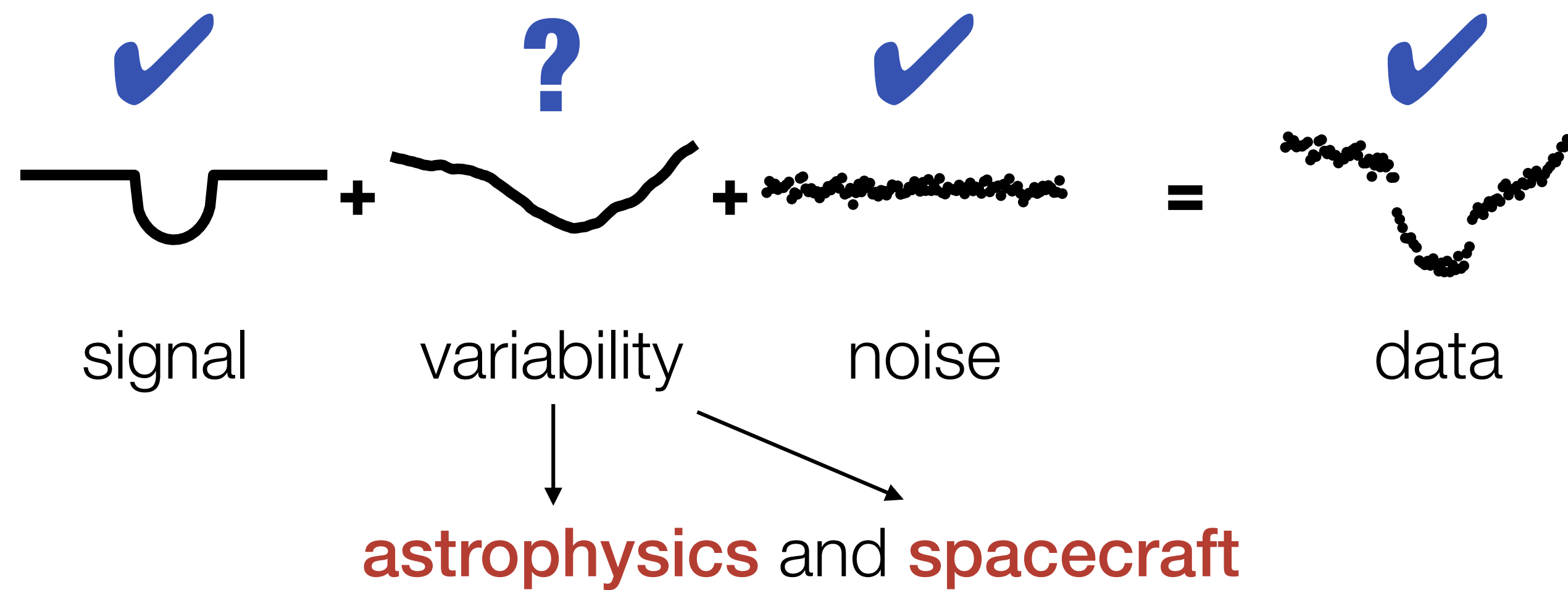


GAUSSIAN PROCESSES

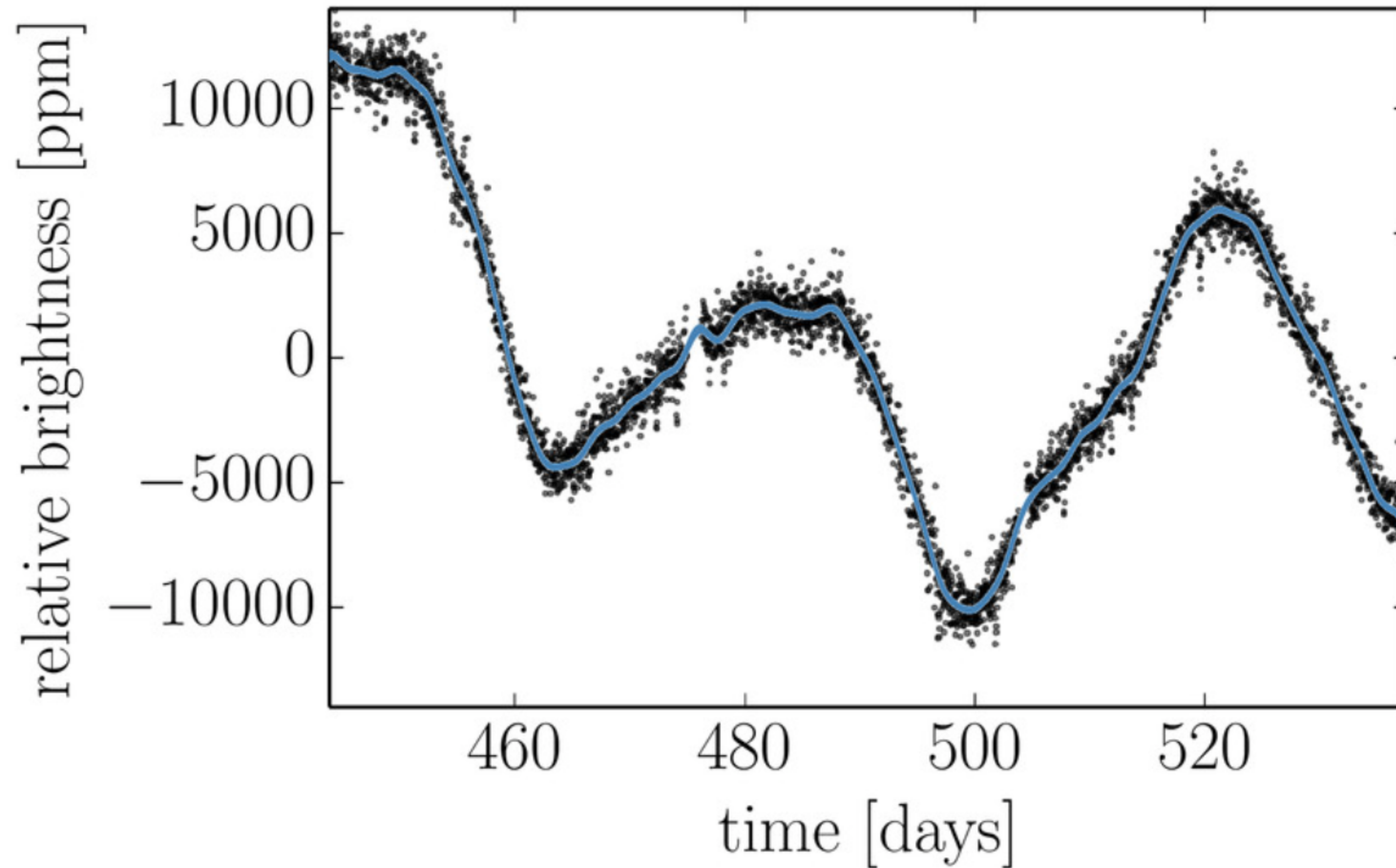




The **anatomy** of a **transit** observation

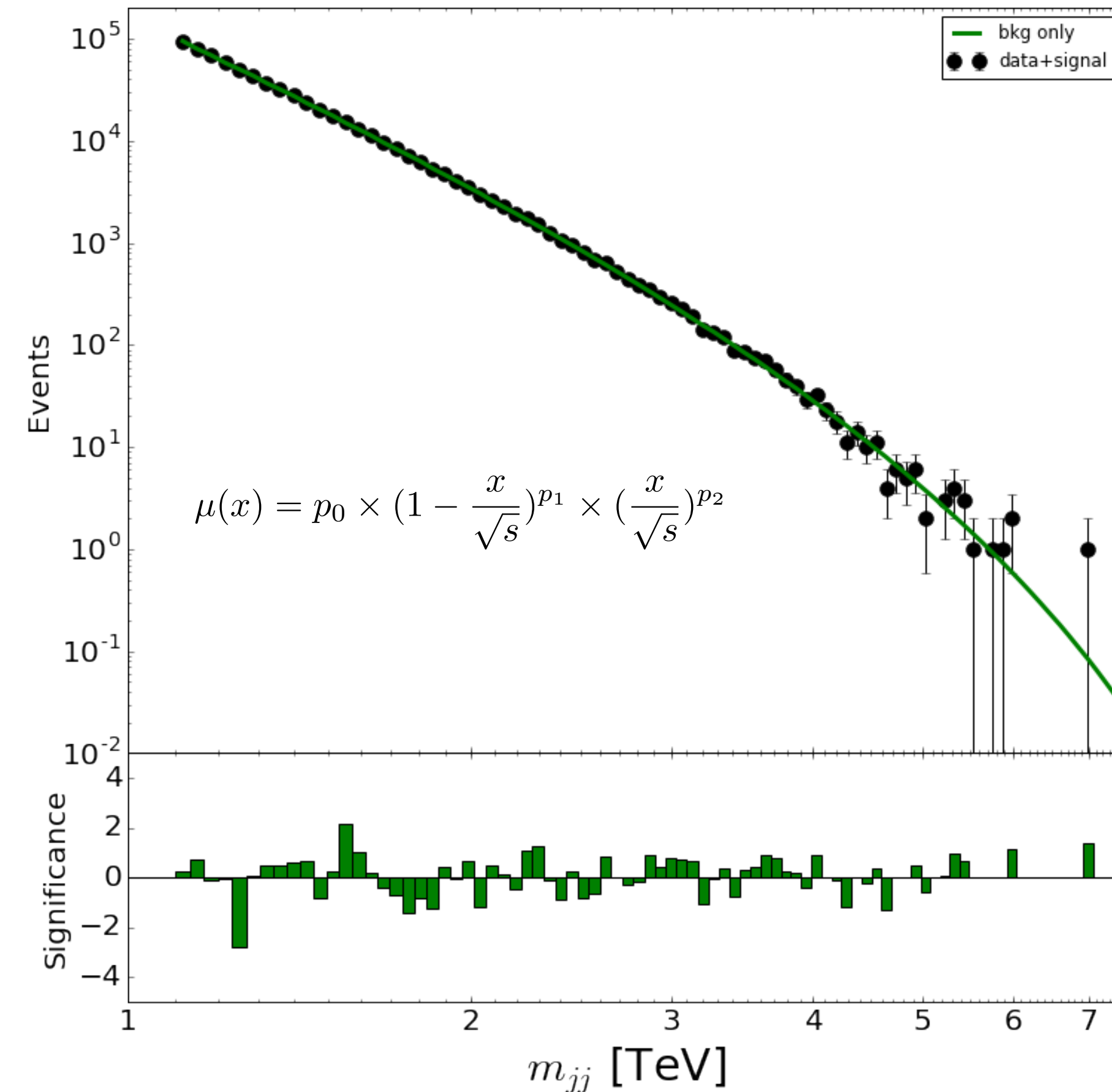
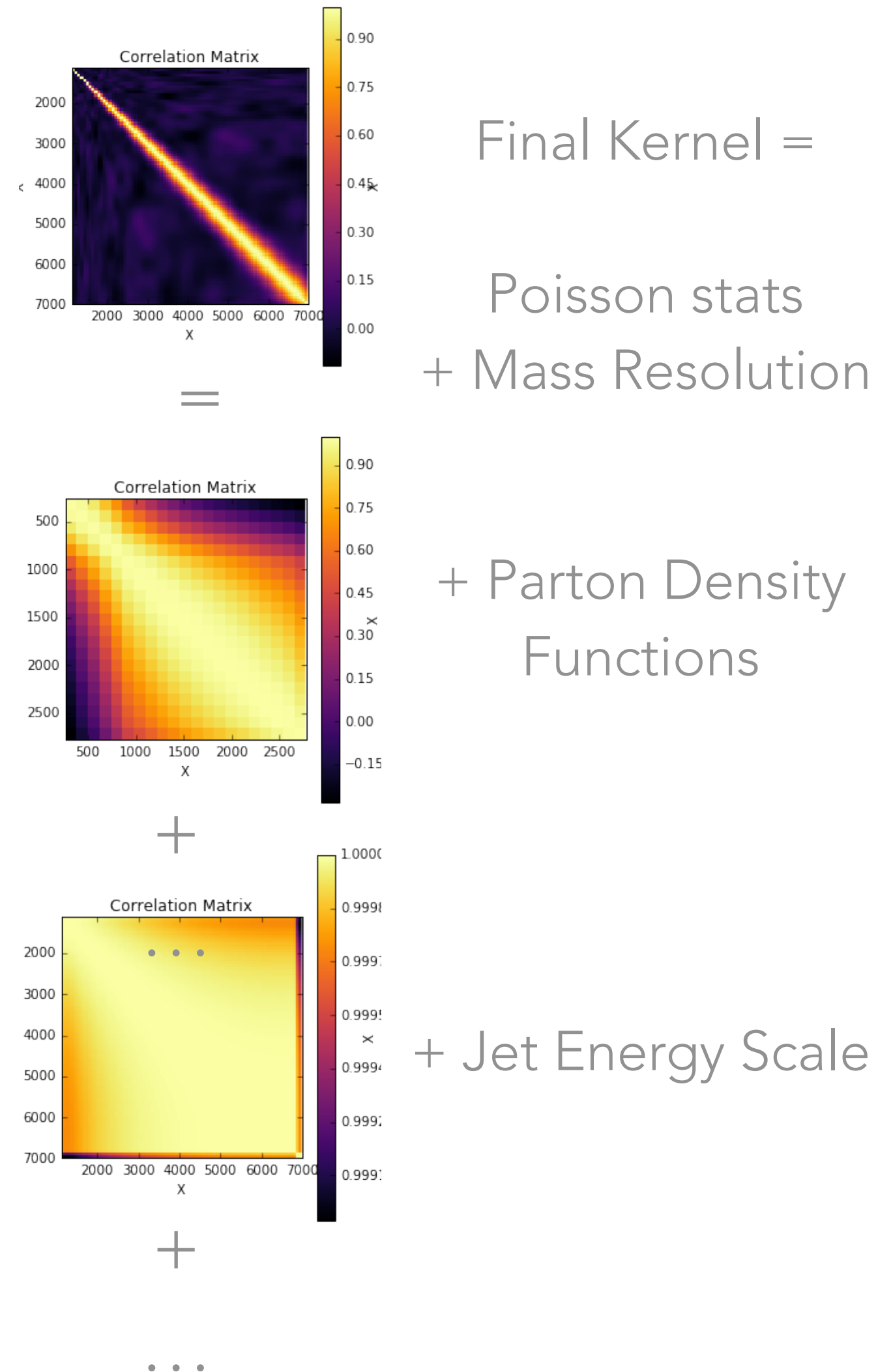


An exoplanet Example



Gaussian Processes for HEP

Instead of fitting the dijet spectrum with an ad hoc 3-5 parameter function, use GP with kernel motivated from physics



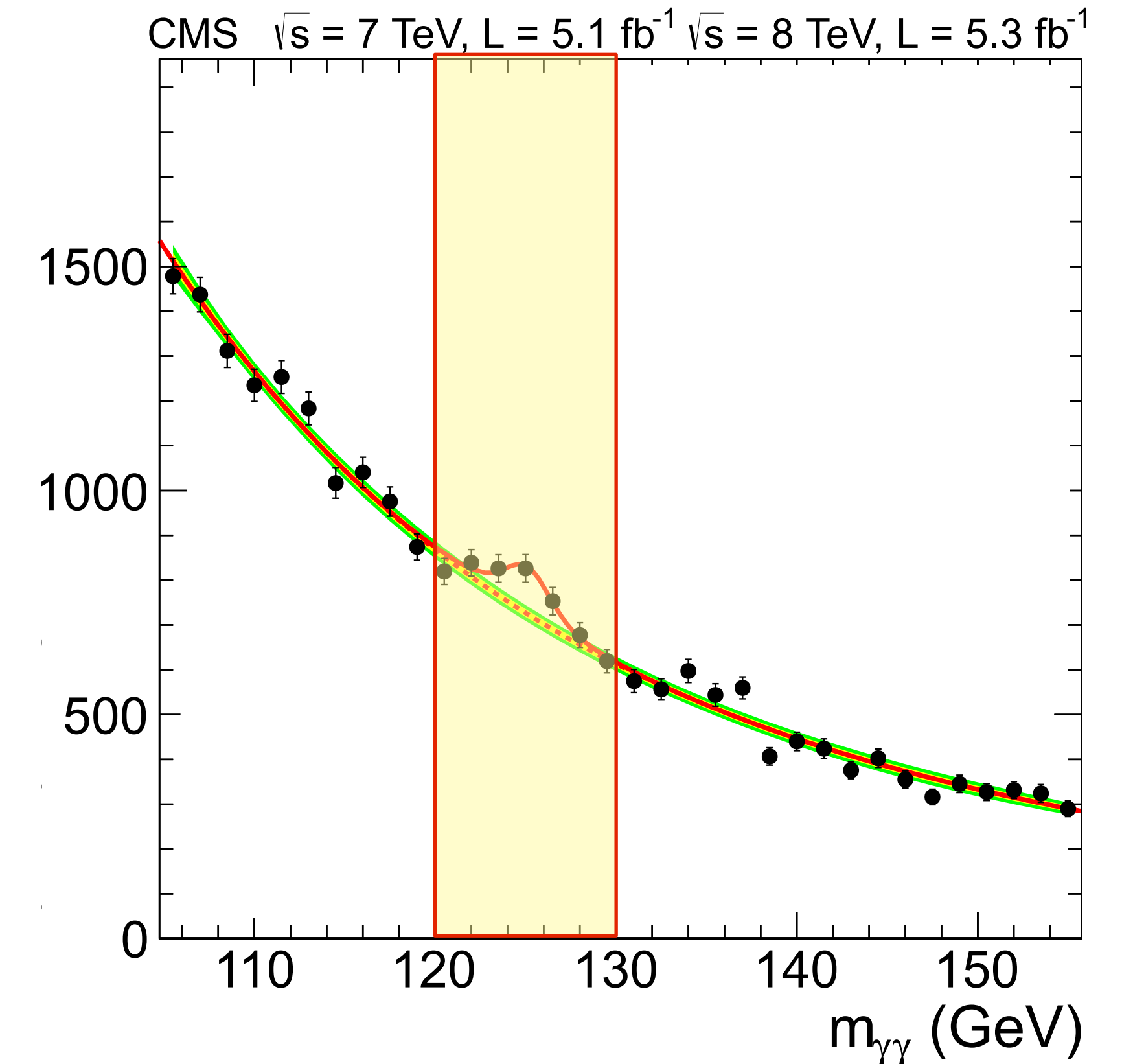
Bump Hunts

Classic bump hunt scans across a mass window looks for an excess in a localized region

- (usually 2-3x the mass resolution)
- Very mild "bias" on type of signal models
- Number counting in the window, no signal shape

A narrow resonance search can add sensitivity by using shape information

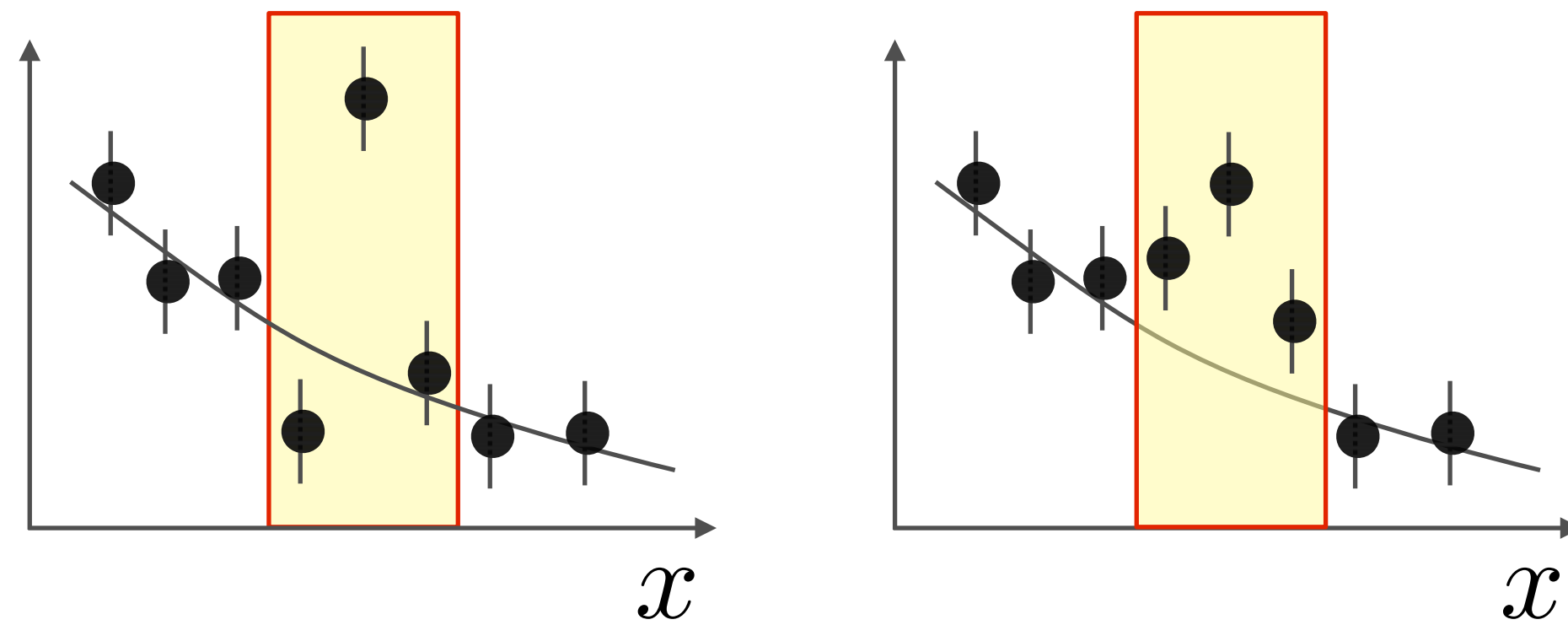
- Excess should be consistent with resolution
- Model dependence (width \ll resolution)



Gaussian Process for localized signals

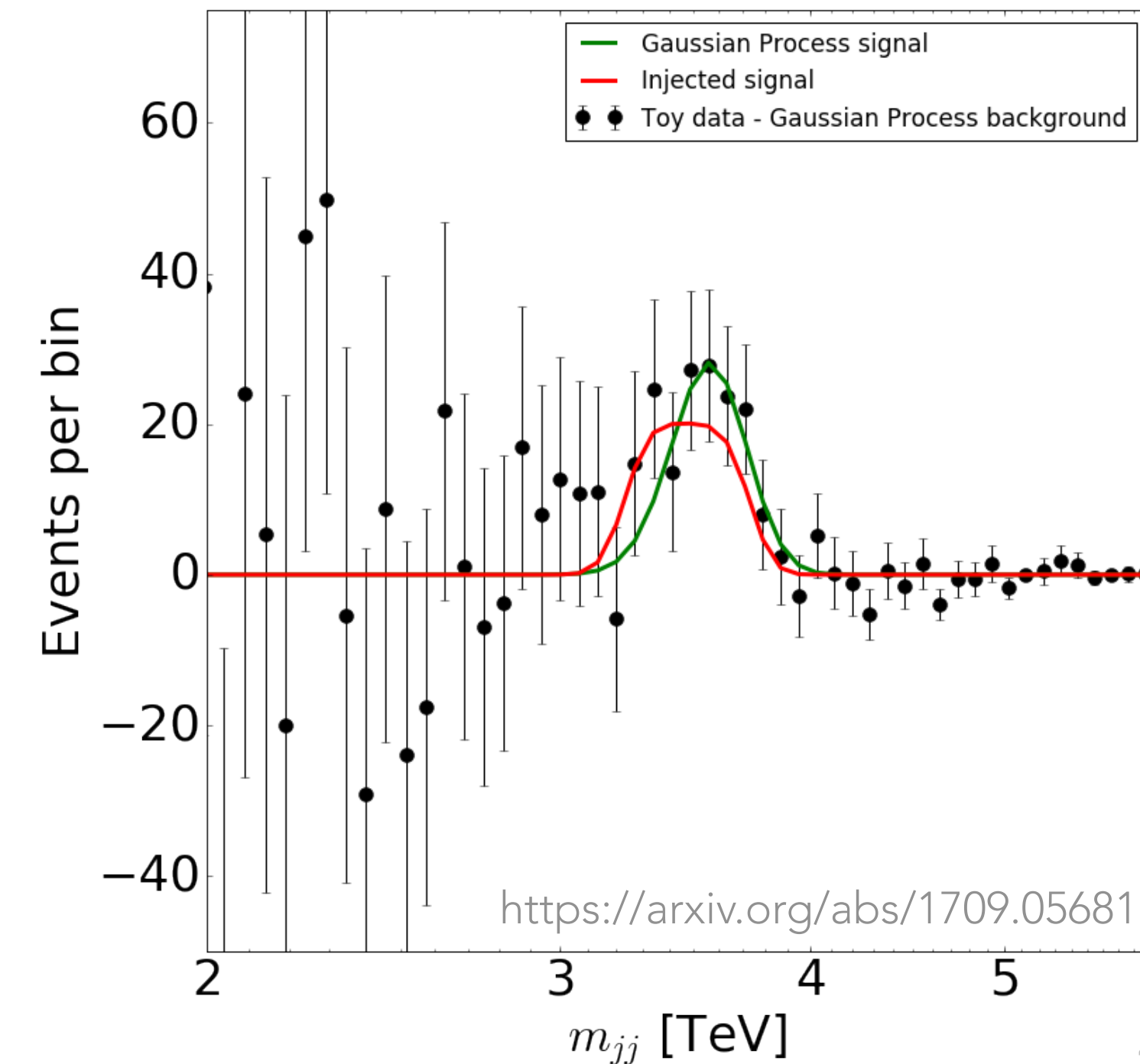
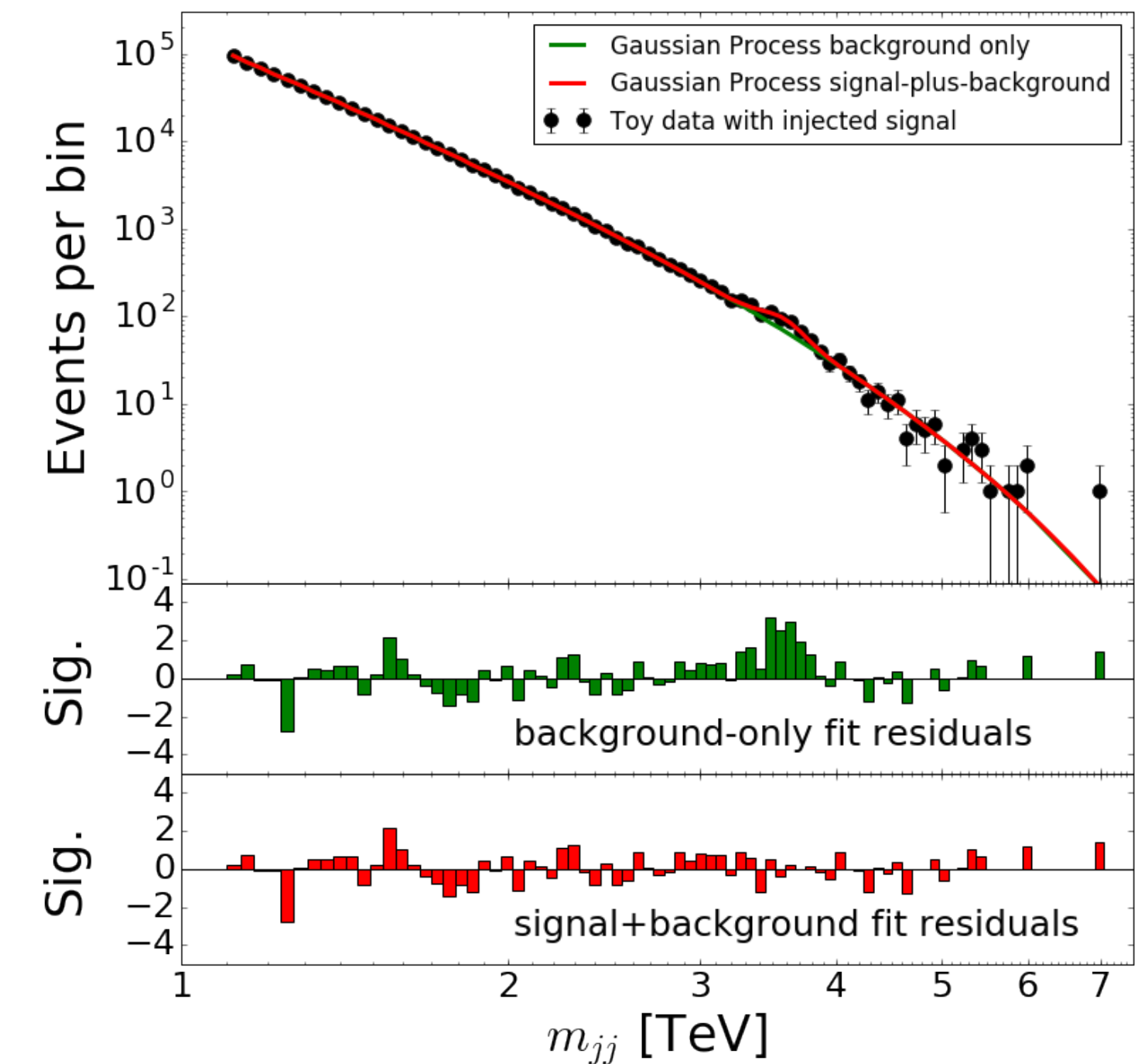
The classic bump hunt will not distinguish between these two situations with same number of events in mass window

- **Left** is not physical, width of excess \ll resolution
- **Right** is physical



With Gaussian processes we can specify signal to be a localized excess of width t centered around m and mass resolution l without having to specify the exact shape of the signal

$$\Sigma(x, x') = A e^{-\frac{1}{2}(x-x')^2/l^2} e^{-\frac{1}{2}((x-m)^2+(x'-m)^2)/t^2}$$



<https://arxiv.org/abs/1709.05681>

Information

References (44)

Citations (0)

Files

Plots

Modeling Smooth Backgrounds and Generic Localized Signals with Gaussian Processes

Meghan Frate, Kyle Cranmer, Saarik Kalia, Alexander Vandenberg-Rodes, Daniel Whiteson

Sep 17, 2017 - 14 pages

e-Print: [arXiv:1709.05681](https://arxiv.org/abs/1709.05681) [physics.data-an] | [PDF](#)

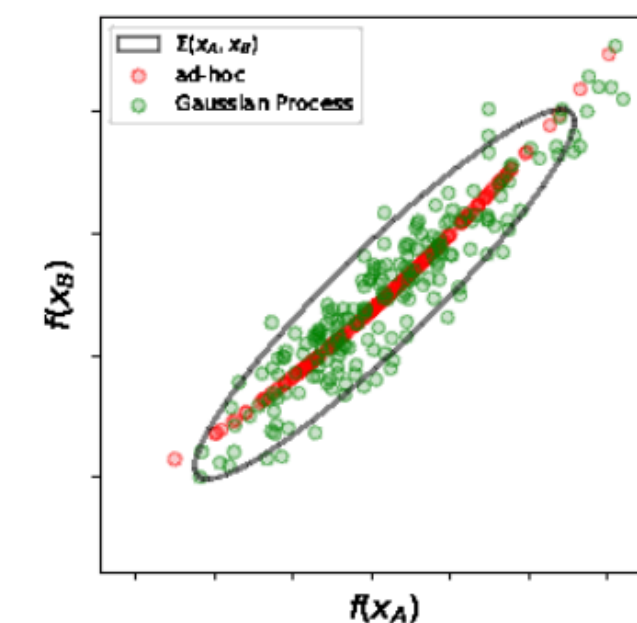
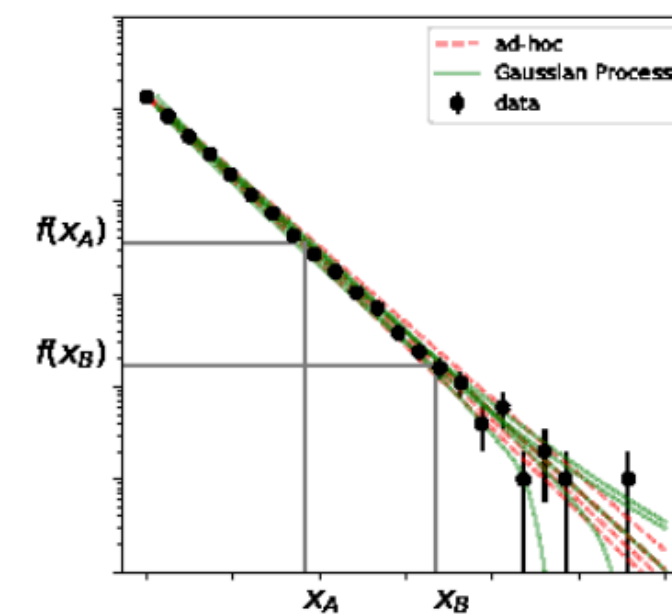
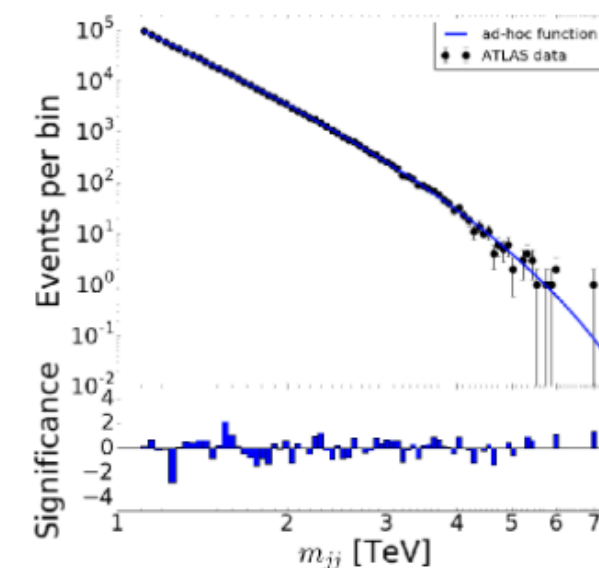
Abstract (arXiv)

We describe a procedure for constructing a model of a smooth data spectrum using Gaussian processes rather than the historical parametric description. This approach considers a fuller space of possible functions, is robust at increasing luminosity, and allows us to incorporate our understanding of the underlying physics. We demonstrate the application of this approach to modeling the background to searches for dijet resonances at the Large Hadron Collider and describe how the approach can be used in the search for generic localized signals.

Note: *Temporary entry*

Note: 14 pages, 16 figures

Keyword(s): INSPIRE: [background](#) | [CERN LHC Coll](#) | [dijet](#) | [resonance](#) | [data analysis method](#) | [Gauss model](#) | [statistics](#) | [statistical analysis](#)



[Show more plots](#)

Record added 2017-09-19, last modified 2017-10-07

Takeaways

Main takeaways from those slides:

- Bias-variance tradeoff: Allowing for some bias may lead to better solutions
- Regularization is a technique for biasing models towards well behaved solutions
- We already do this in several settings where the thing we are estimating is more complicated than a single number (e.g. unfolding)
- We can build models for the signal that aren't based on QFT, but on other descriptive properties
 - e.g. smooth, localized excess comparable with detector resolution

Statistical Decision Theory

Statistical Decision theory

Θ - States of nature;

- Background-only is true
- BSM Theory 1 is true
- BSM Theory 2 is true

Statistical Decision theory

Θ - States of nature;

- Background-only is true
- BSM Theory 1 is true
- BSM Theory 2 is true

X - possible observations;

- Data from LHC and other experiments

Statistical Decision theory

Θ - States of nature;

- Background-only is true
- BSM Theory 1 is true
- BSM Theory 2 is true

X - possible observations;

- Data from LHC and other experiments

A - action to be taken

- claim a discovery
- build a muon collider
- build next hadron collider

Statistical Decision theory

Θ - States of nature;

- Background-only is true
- BSM Theory 1 is true
- BSM Theory 2 is true

X - possible observations;

- Data from LHC and other experiments

A - action to be taken

- claim a discovery
- build a muon collider
- build next hadron collider

$p(x|\theta)$ - **statistical model** (likelihood);

- Predictions of QFT + detector simulation etc.

$\pi(\theta)$ - **prior**

- **You don't need this for frequentist statistical statements, but you will probably need it for making decisions !**

Statistical Decision theory

Θ - **States of nature**;

- Background-only is true
- BSM Theory 1 is true
- BSM Theory 2 is true

X - **possible observations**;

- Data from LHC and other experiments

A - **action to be taken**

- claim a discovery
- build a muon collider
- build next hadron collider

$p(x|\theta)$ - **statistical model** (likelihood);

- Predictions of QFT + detector simulation etc.

$\pi(\theta)$ - **prior**

- **You don't need this for frequentist statistical statements, but you will probably need it for making decisions !**

$\delta: X \rightarrow A$ - **decision rule** (take some action based on observation)

- Some data analysis pipeline (either model-dependent or model-independent) that might claim "discovery"
- The community planning process (e.g. Snowmass, European strategy, etc.); Lab decisions

$L: \Theta \times A \rightarrow \mathbb{R}$ - **loss function**, real-valued function true parameter and action

- Usually not made explicit.
- Claim discovery when new physics is there +++; Claim discovery when no new physics - - -; Build collider that doesn't discover what was anticipated ???

Statistical Decision theory

- Θ - **States of nature**; X - **possible observations**; A - **action to be taken**
- $p(x|\theta)$ - **statistical model** (likelihood); $\pi(\theta)$ - **prior**
- $\delta: X \rightarrow A$ - **decision rule** (take some action based on observation)
- $L: \Theta \times A \rightarrow \mathbb{R}$ - **loss function**, real-valued function true parameter and action

Statistical Decision theory

- Θ - **States of nature**; X - **possible observations**; A - **action to be taken**
- $p(x|\theta)$ - **statistical model** (likelihood); $\pi(\theta)$ - **prior**
- $\delta: X \rightarrow A$ - **decision rule** (take some action based on observation)
- $L: \Theta \times A \rightarrow \mathbb{R}$ - **loss function**, real-valued function true parameter and action

$$R(\theta, \delta) = E_{p(x|\theta)}[L(\theta, \delta)] - \text{risk}$$

- Function of both θ and δ . We don't know true value of θ !
- If $R(\theta, \delta_1) < R(\theta, \delta_2)$ for all θ , then δ_1 "dominates" δ_2 , and δ_2 is "inadmissible"
- But usually one rule is better for some θ , while the other is better for other values of θ
- **Mini-max strategy**: choose δ that minimizes risk over all θ — very conservative.

$$r(\pi, \delta) = E_{\pi(\theta)}[R(\theta, \delta)] - \text{Bayes risk} \quad (\text{expectation over } \theta \text{ w.r.t. prior and possible observations})$$

- **Bayes rule**: choose δ that minimize Bayes risk (w.r.t. prior π).
- Also averages over potential data, so you can choose δ before seeing the data X

$$\rho(\pi, \delta | x) = E_{\pi(\theta|x)}[L(\theta, \delta(x))] - \text{expected loss} \quad (\text{expectation over } \theta \text{ w.r.t. posterior } \pi(\theta|x))$$

- Here decision is conditioned on the data you actually collect. Still depends on prior π .

Statistical Decision theory

- Θ - **States of nature**; X - **possible observations**; A - **action to be taken**
- $p(x|\theta)$ - **statistical model** (likelihood); $\pi(\theta)$ - **prior**
- $\delta: X \rightarrow A$ - **decision rule** (take some action based on observation)
- $L: \Theta \times A \rightarrow \mathbb{R}$ - **loss function**, real-valued function true parameter and action

$$R(\theta, \delta) = E_{p(x|\theta)}[L(\theta, \delta)] - \text{risk}$$

- Function of both θ and δ . We don't know true value of θ !
- If $R(\theta, \delta_1) < R(\theta, \delta_2)$ for all θ , then δ_1 "dominates" δ_2 , and δ_2 is "inadmissible"
- But usually one rule is better for some θ , while the other is better for other values of θ
- **Mini-max strategy**: choose δ that minimizes risk over all θ — very conservative.

$$r(\pi, \delta) = E_{\pi(\theta)}[R(\theta, \delta)] - \text{Bayes risk} \quad (\text{expectation over } \theta \text{ w.r.t. prior and possible observations})$$

- **Bayes rule**: choose δ that minimize Bayes risk (w.r.t. prior π).
- Also averages over potential data, so you can choose δ before seeing the data X

$$\rho(\pi, \delta | x) = E_{\pi(\theta|x)}[L(\theta, \delta(x))] - \text{expected loss} \quad (\text{expectation over } \theta \text{ w.r.t. posterior } \pi(\theta|x))$$

- Here decision is conditioned on the data you actually collect. Still depends on prior π .

- **Priors used for decision making are subtly different than priors for making statistical statements about the data.**
- It's our risk / loss, so natural we get to use our own prior when making decisions.
- Usually there implicitly in human decisions
- Without prior, what is the principle?

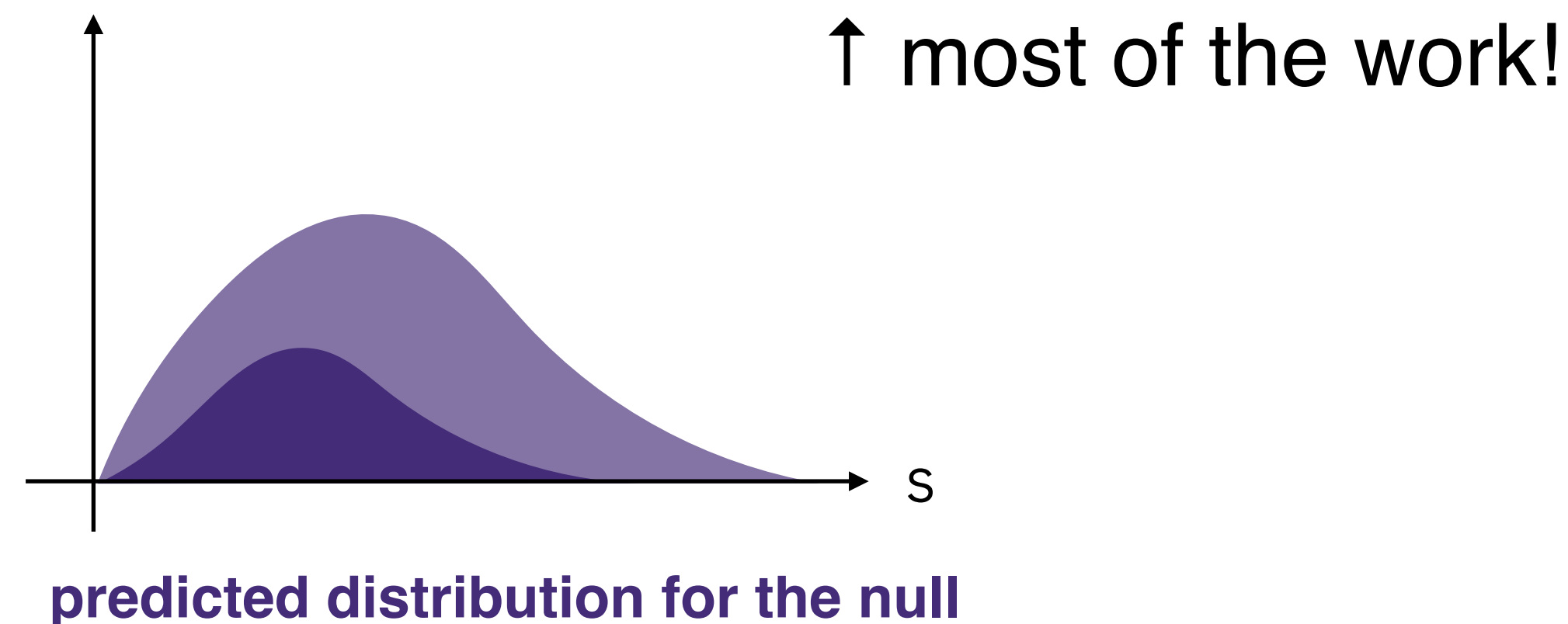
Not optimal, but not wrong

Thumbnail Sketch of Analysis

We select a small subset of the collision events relevant for testing the hypotheses we are considering.

And we design a summary statistic \mathbf{s} that can distinguish between different hypotheses we are considering.

- Then we run simulated collisions through the pipeline to make the prediction for the null or “background-only” hypothesis and quantify systematic uncertainties

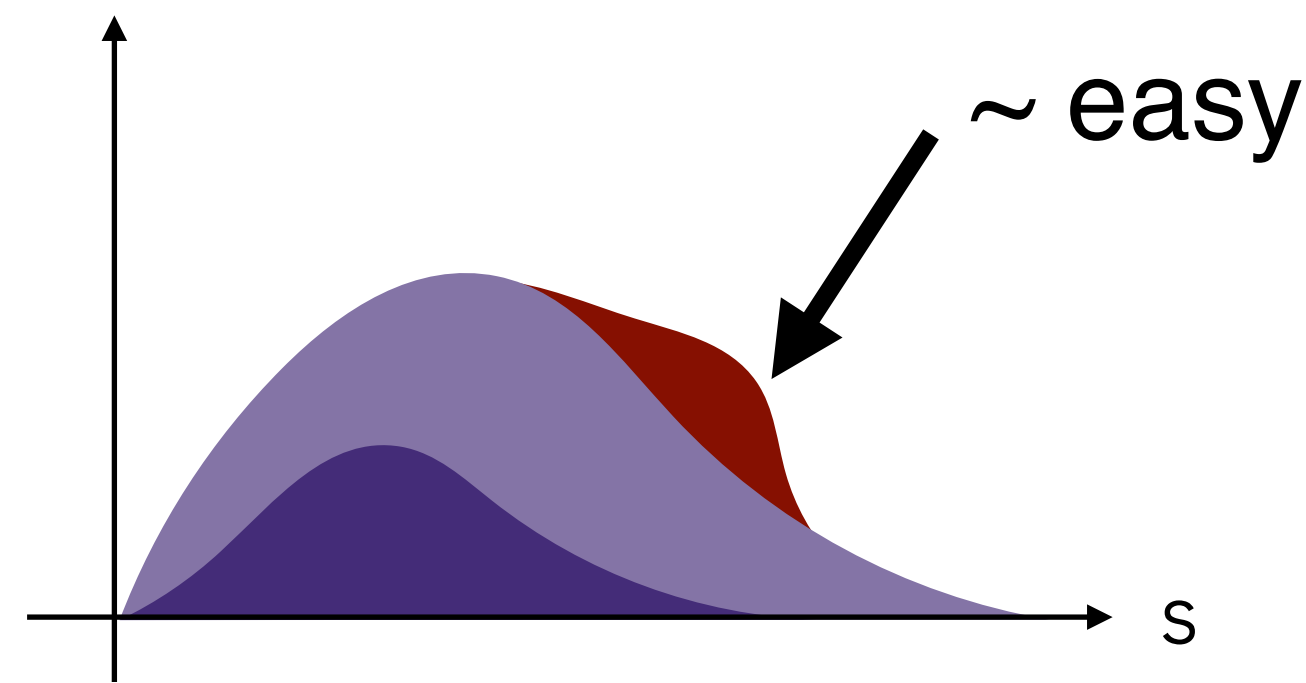


Thumbnail Sketch of Analysis

We select a small subset of the collision events relevant for testing the hypotheses we are considering.

And we design a summary statistic \mathbf{s} that can distinguish between different hypotheses we are considering.

- Then we run simulated collisions for a hypothetical particle or interaction to make the prediction for alternate or “signal-plus-background” model



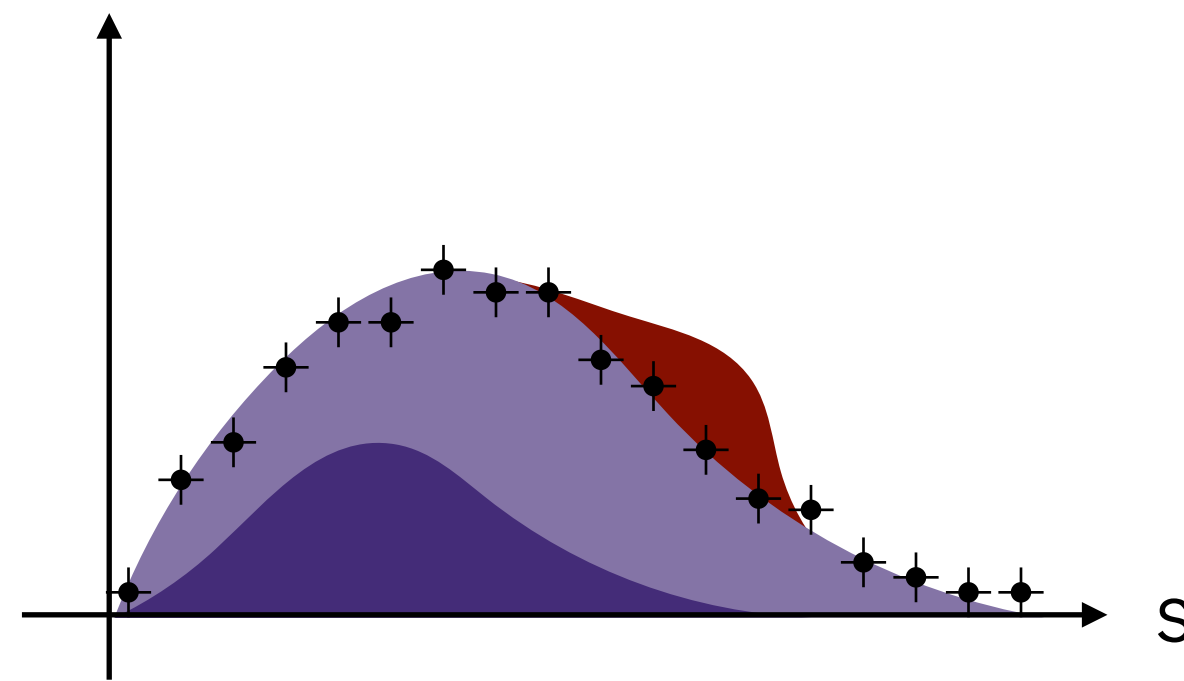
predicted distribution for the alternate in **Model A**

Thumbnail Sketch of Analysis

We select a small subset of the collision events relevant for testing the hypotheses we are considering.

And we design a summary statistic \mathbf{s} that can distinguish between different hypotheses we are considering.

- Then we add the observed data



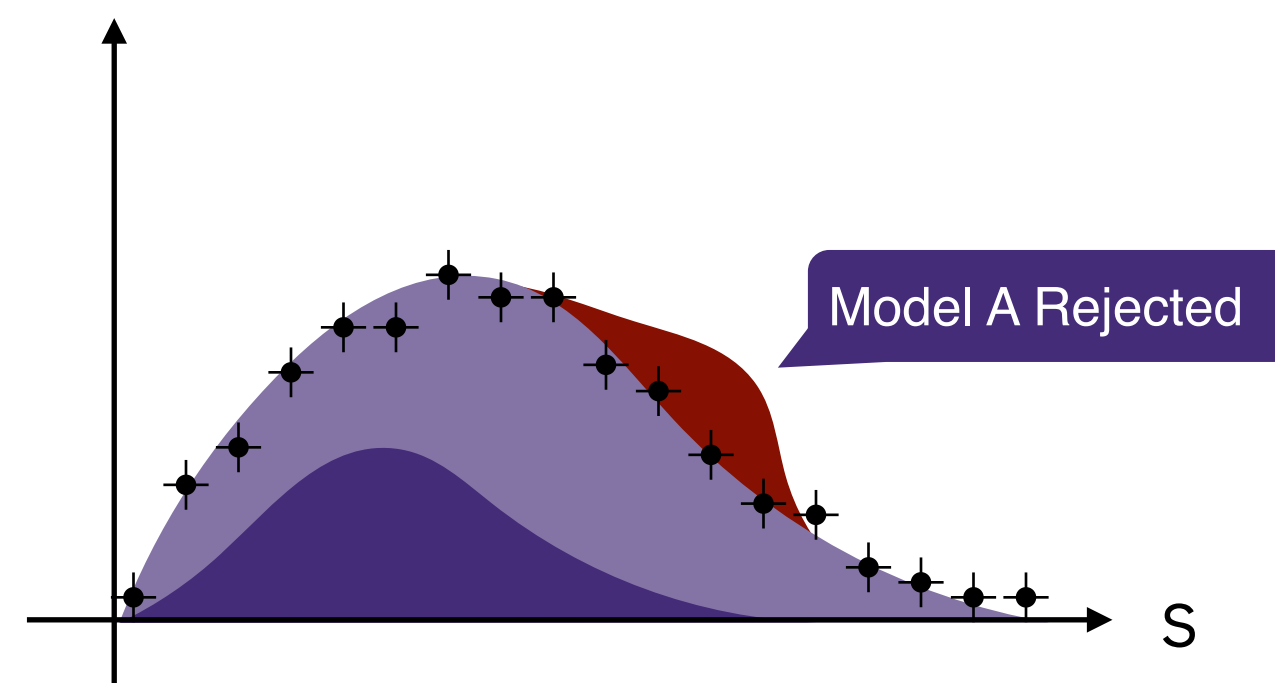
observed **data** + predicted distribution for the alternate in **Model A**

Thumbnail Sketch of Analysis

We select a small subset of the collision events relevant for testing the hypotheses we are considering.

And we design a summary statistic \mathbf{s} that can distinguish between different hypotheses we are considering.

- Then we test the hypothesis and write a paper



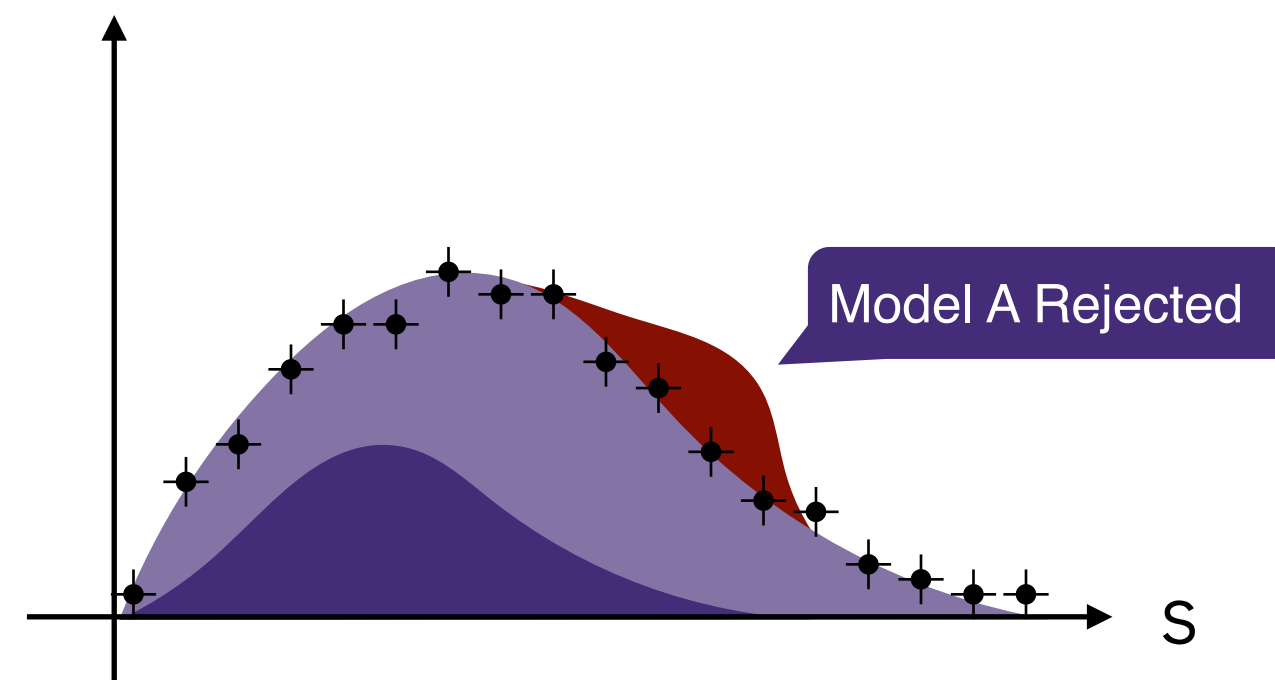
observed **data** + predicted distribution for the alternate in **Model A**

Thumbnail Sketch of Analysis

We select a small subset of the collision events relevant for testing the hypotheses we are considering.

And we design a summary statistic \mathbf{s} that can distinguish between different hypotheses we are considering.

- ... and graduate students graduate, analysis code rots, and it would be difficult to reproduce or reuse this work

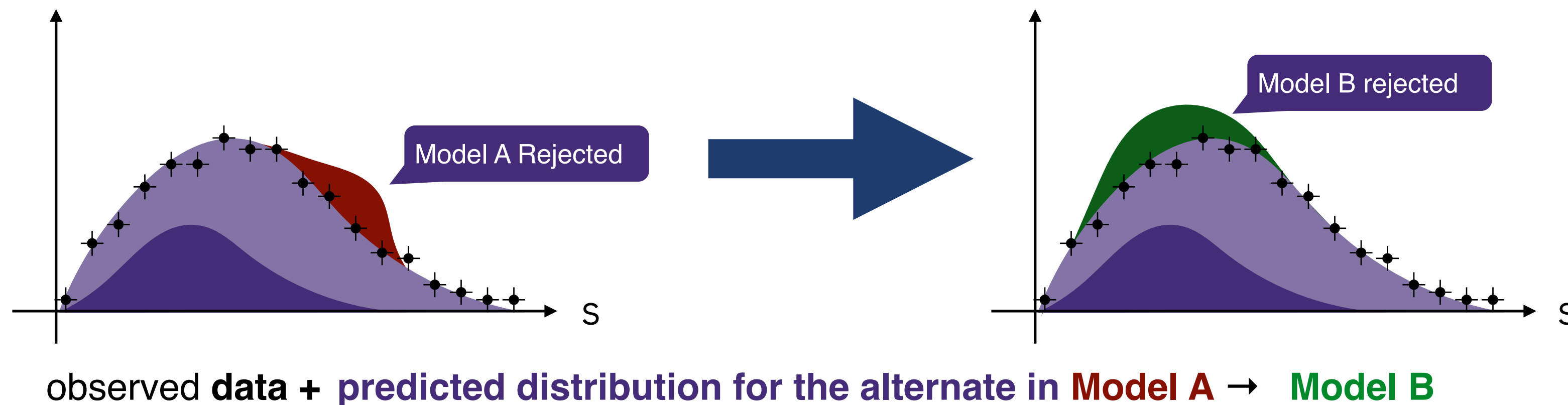


observed **data** + predicted distribution for the alternate in **Model A**

Reinterpretation

If we can capture the definition of the summary $\mathbf{s}(\mathbf{x})$ and the event selection, then we can reuse the existing analysis

- We just need to run simulated events for **Model B** through the pipeline and test the new signal+background alternate hypothesis
- In that sense, the original analysis isn't "model-dependent"
- Not optimal, but not wrong



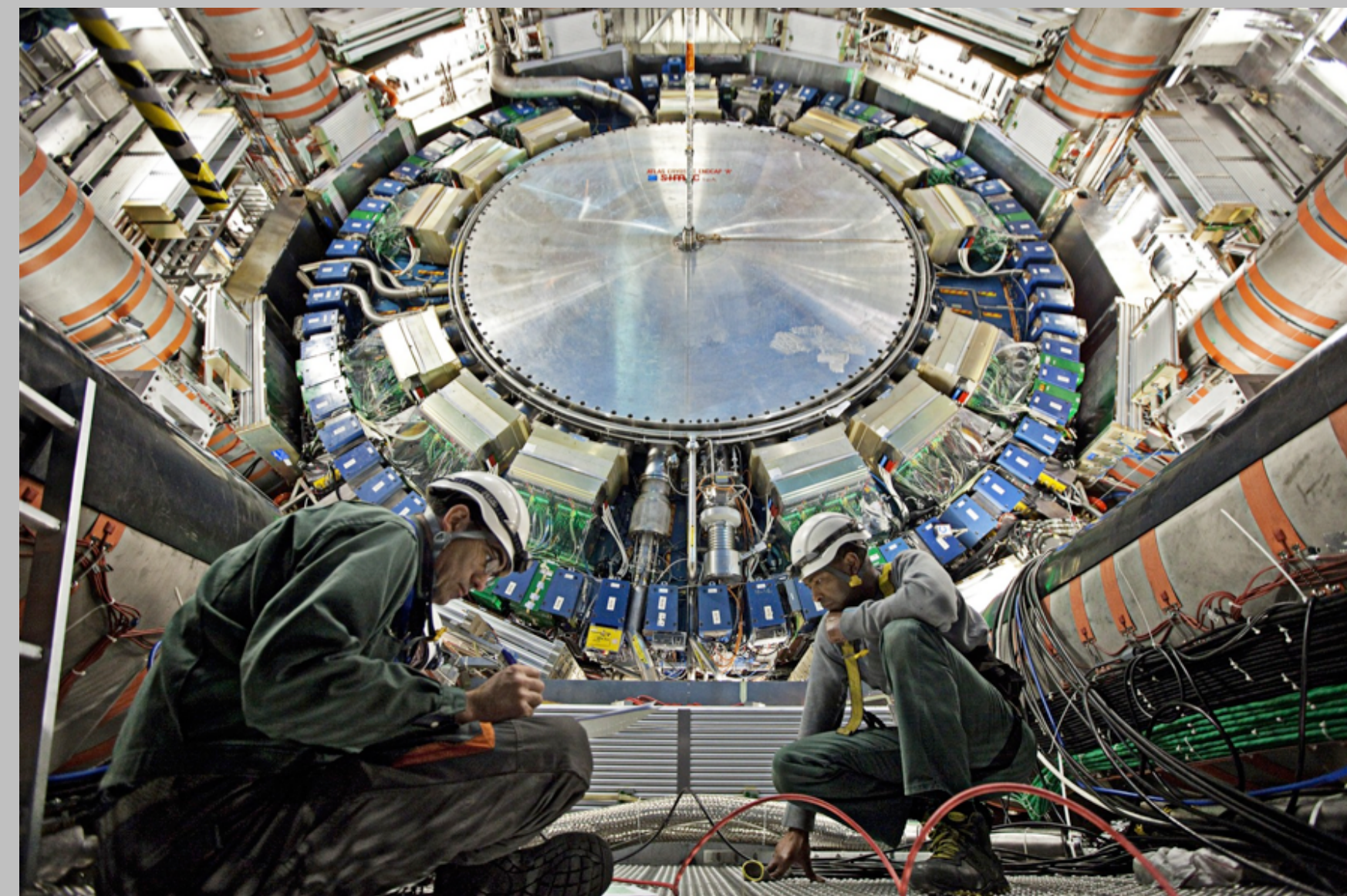
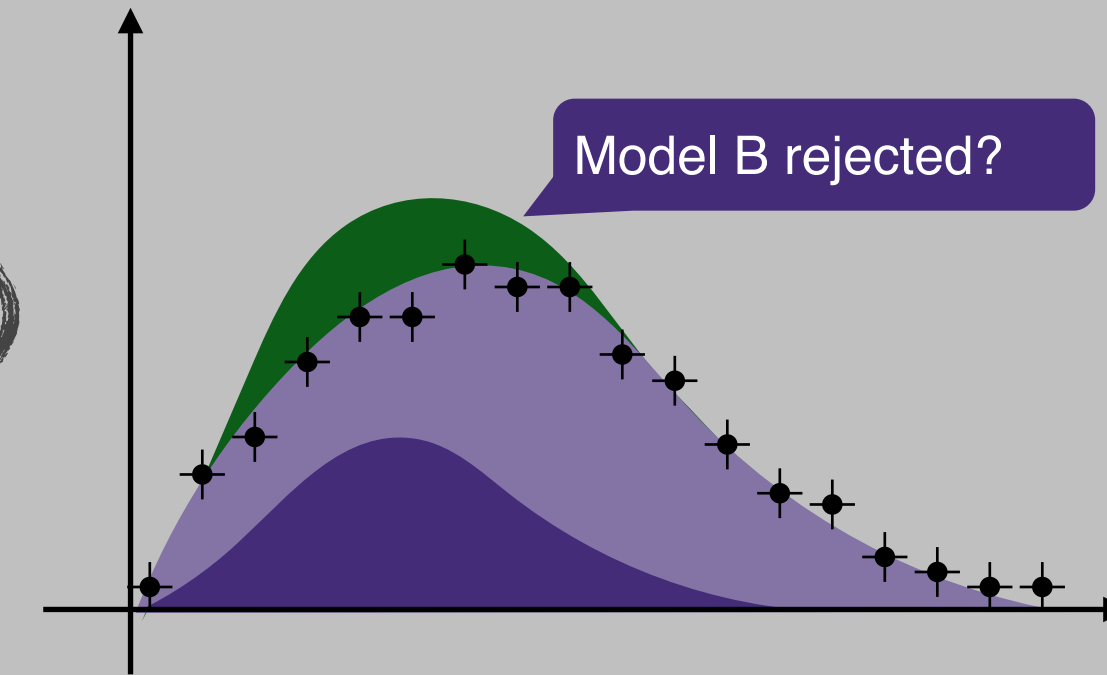
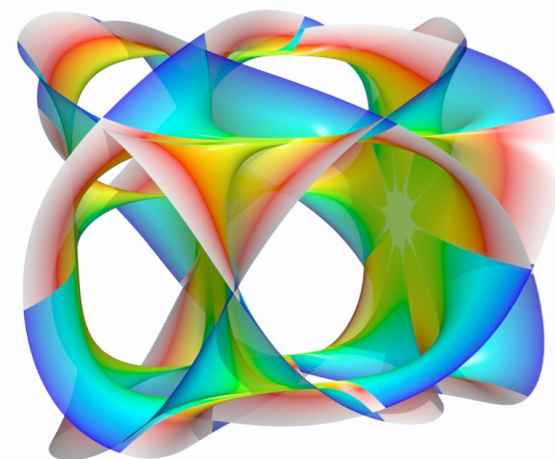
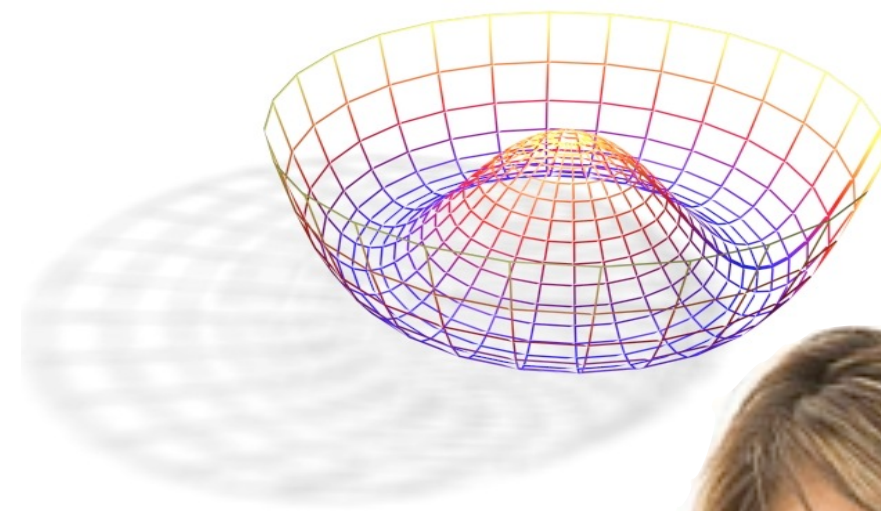
THEORY

SERVICE

$$\begin{aligned} \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\ & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\ & + \underbrace{\frac{1}{2} |(i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi|^2 - V(\phi)}_{\text{W}^\pm, Z, \gamma, \text{ and Higgs masses and couplings}} \\ & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}} \end{aligned}$$

Q

A



We proposed RECAST framework in Oct 2010

- People said it couldn't be done, our workflows are too complicated
- Hard to get effort to work on it.



RECAST

Extending the Impact of Existing Analyses

Kyle Cranmer and Itay Yavin

Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003



ABSTRACT: Searches for new physics by experimental collaborations represent a significant investment in time and resources. Often these searches are sensitive to a broader class of models than they were originally designed to test. We aim to extend the impact of existing searches through a technique we call *recasting*. After considering several examples, which illustrate the issues and subtleties involved, we present RECAST, a framework designed to facilitate the usage of this technique.

RECAST in action

ATLAS has started using RECAST to reinterpret SUSY and exotics searches



- Also relevant for exotic BSM Higgs scenarios



ATLAS PUB Note
ATL-PHYS-PUB-2019-032
11th August 2019


RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two b -quarks



The ATLAS Collaboration

The reinterpretation of a search for dark matter produced in association with a Higgs boson decaying to b -quarks performed with RECAST, a software framework designed to facilitate the reinterpretation of existing searches for new physics, is presented. Reinterpretation using RECAST is enabled through the sustainable preservation of the original data analysis as re-executable declarative workflows using modern cloud technologies and integrated with the wider CERN Analysis Preservation efforts. The reinterpretation targets a model predicting dark matter production in association with a hypothetical dark Higgs boson decaying into b -quarks where the mass of the dark Higgs boson m_χ is a free parameter, necessitating a faithful reinterpretation of the analysis. The dataset has an integrated luminosity of 79.8 fb^{-1} and was recorded with the ATLAS detector at the Large Hadron Collider at a centre-of-mass energy of $\sqrt{s} = 13 \text{ TeV}$. Constraints on the parameter space of the dark Higgs model for a fixed choice of dark matter mass $m_\chi = 200 \text{ GeV}$ exclude model configurations with a mediator mass up to 3.2 TeV .

© 2019 CERN for the benefit of the ATLAS Collaboration.
Reproduction of this article or parts of it is allowed as specified in the CC-BY-4.0 license.

ATL-PHYS-PUB-2019-032
12 August 2019





ATLAS PUB Note
ATL-PHYS-PUB-2020-007
27th March 2020


Reinterpretation of the ATLAS Search for Displaced Hadronic Jets with the RECAST Framework

The ATLAS Collaboration

A recent ATLAS search for displaced jets in the hadronic calorimeter is preserved in RECAST and thereafter used to constrain three new physics models not studied in the original work. A Stealth SUSY model and a Higgs-portal baryogenesis model, both predicting long-lived particles and therefore displaced decays, are probed for proper decay lengths between a few cm and 500 m. A dark sector model predicting Higgs and heavy boson decays to collimated hadrons via long-lived dark photons is also probed. The cross-section times branching ratio for the Higgs channel is constrained between a few millimetres and a few metres, while for a heavier 800 GeV boson the constraints extend from tenths of a millimetre to a few tens of metres. The original data analysis workflow was completely captured using virtualisation techniques, allowing for an accurate and efficient reinterpretation of the published result in terms of new signal models following the RECAST protocol.

© 2020 CERN for the benefit of the ATLAS Collaboration.
Reproduction of this article or parts of it is allowed as specified in the CC-BY-4.0 license.

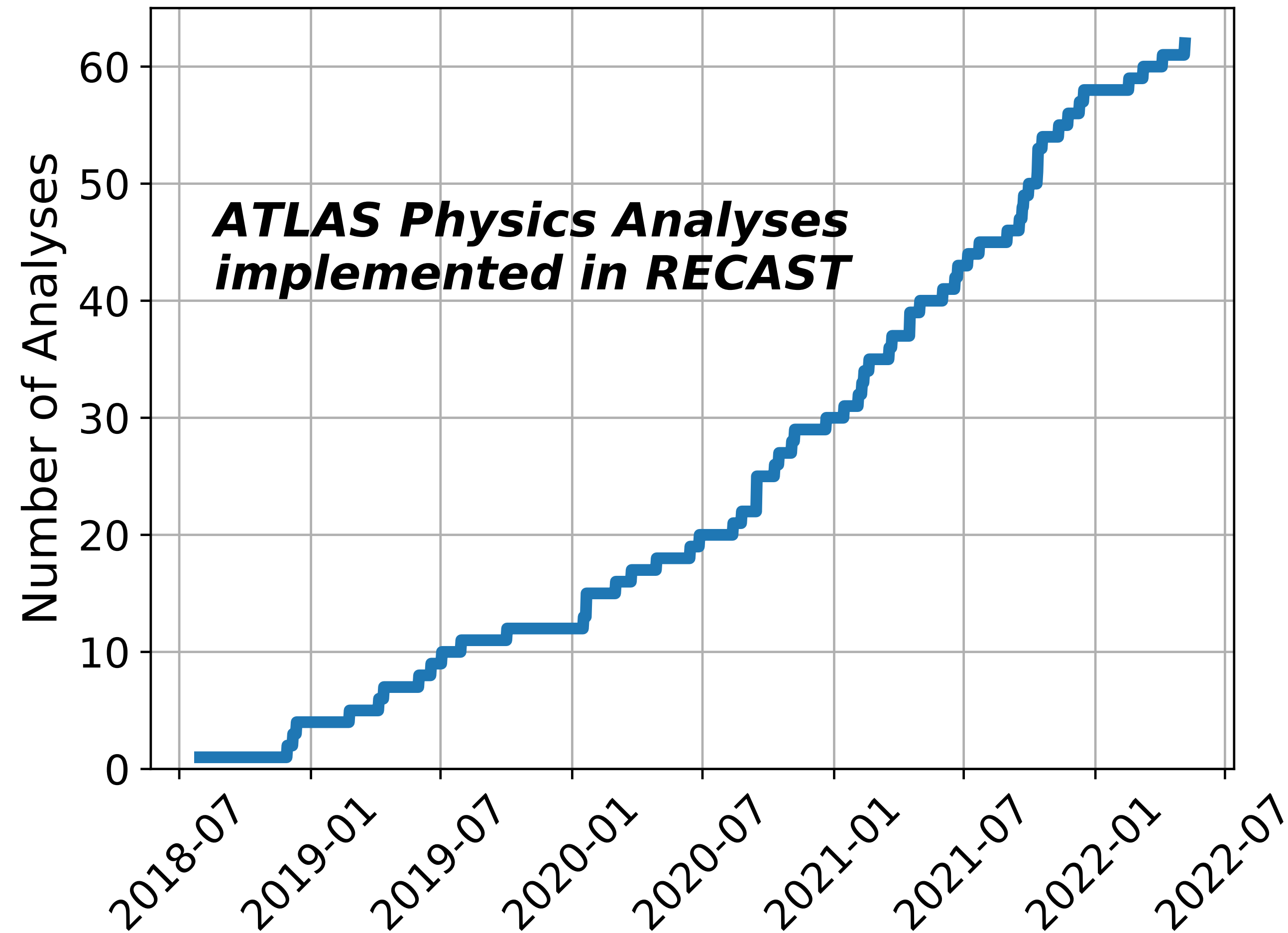
ATL-PHYS-PUB-2020-007
28/03/2020



RECAST in action

ATLAS has started using RECAST to reinterpret SUSY and exotics searches

- Also relevant for exotic BSM Higgs scenarios



CMS Physics Analysis Summary

Contact: cms-pag-conveners-exotica@cern.ch 2020/05/21

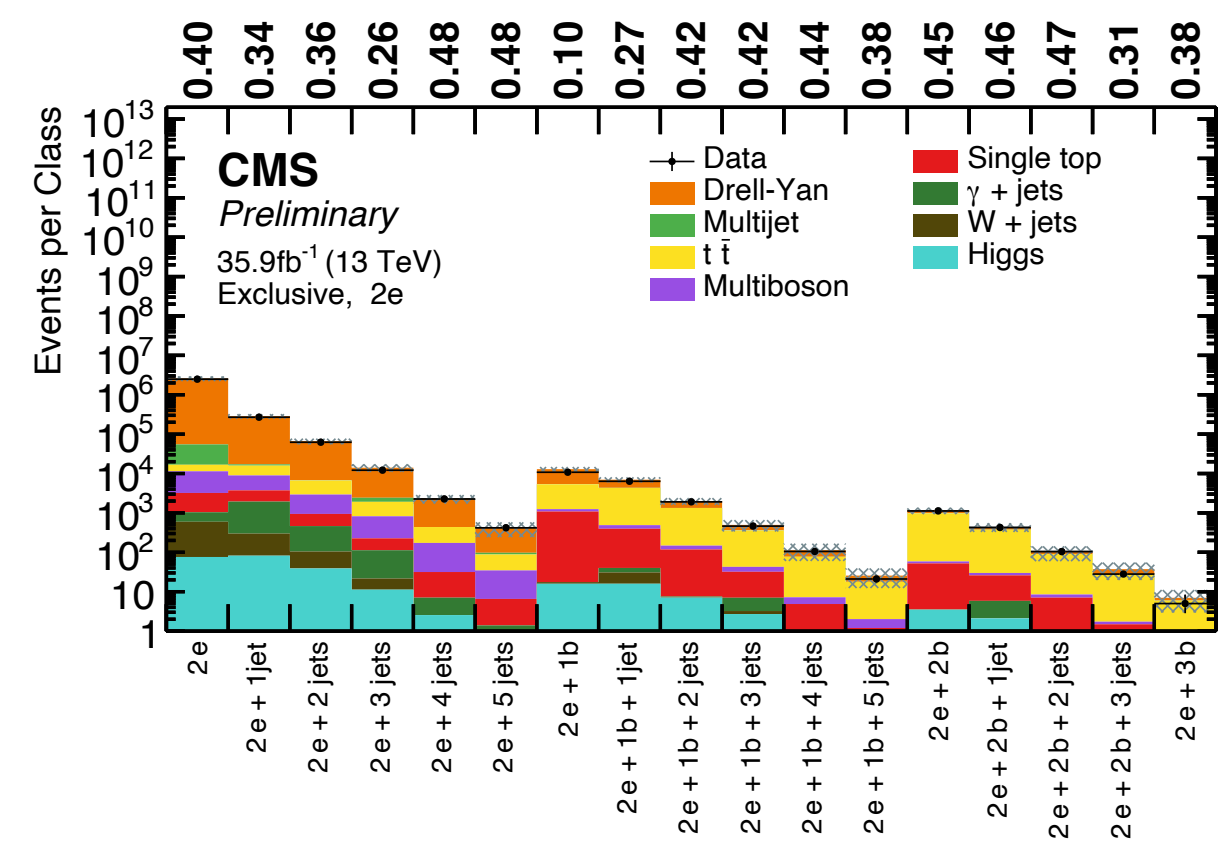
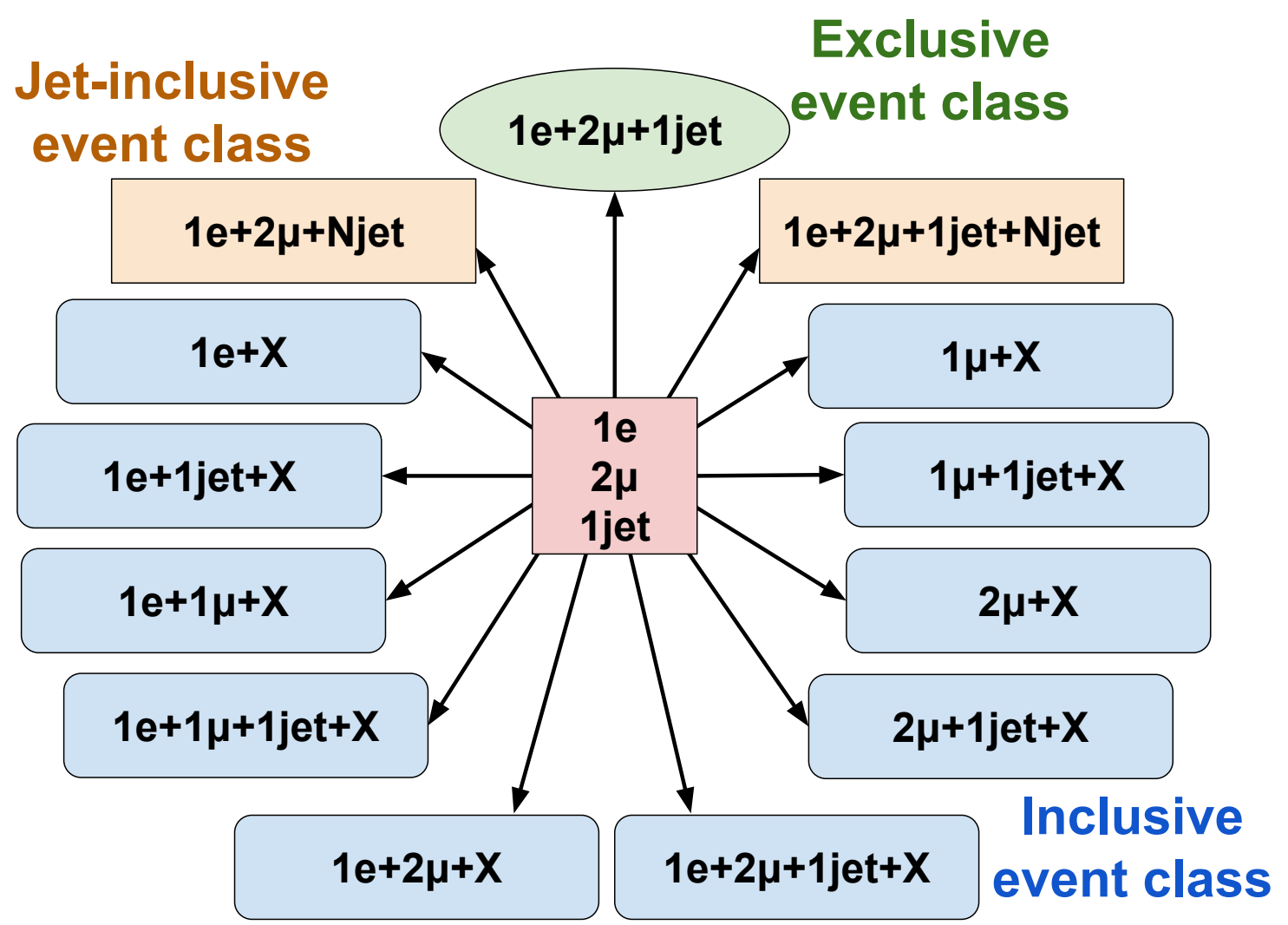
MUSiC, a model unspecific search for new physics, in pp collisions at $\sqrt{s} = 13$ TeV

The CMS Collaboration

These signature based searches have some sensitivity, but it is often unclear how to interpret them

- Do they exclude a particular theory?

Pairing with RECAST addresses this



Results of the Model Unspecific Search in CMS (MUSiC) using data recorded by the CMS detector at the LHC, during proton-proton collisions at a center of mass energy of $\sqrt{s} = 13$ TeV in 2016 and corresponding to an integrated luminosity of 35.9 fb^{-1} , are presented. The MUSiC analysis aims to search for anomalies that could be probed as signatures for phenomena beyond the standard model, and is based on the comparison of data with the expectation according to the standard model, determined from simulations, in several hundred final states and multiple kinematic distributions. Events containing at least one lepton are classified based on their final state topology, and an automated search algorithm subsequently surveys the data for deviations from the expectation. The sensitivity of the search is validated using multiple methods. No significant deviations beyond the expectations have been found. For a wide range of final state topologies, good agreement is found between the data and simulation of the standard model.

RECAST + STXS overcomes model dependence

Different analysis strategies

- Highly optimised analyses targeting specific properties / operators
 - “best possible” sensitivity
 - very model specific
- Fiducial and differential cross section measurements
 - minimise model dependence
 - relatively restricted sensitivity (hard to combine different channels)
 - re-interpretable outside experiment
- Differential measurements in experimentally sensitive observables per production mode (STXS)
 - model dependence from production mode definition
 - easy combination of different Higgs decay channels → sensitivity to large number of EFT operators
 - re-interpretable outside experiment

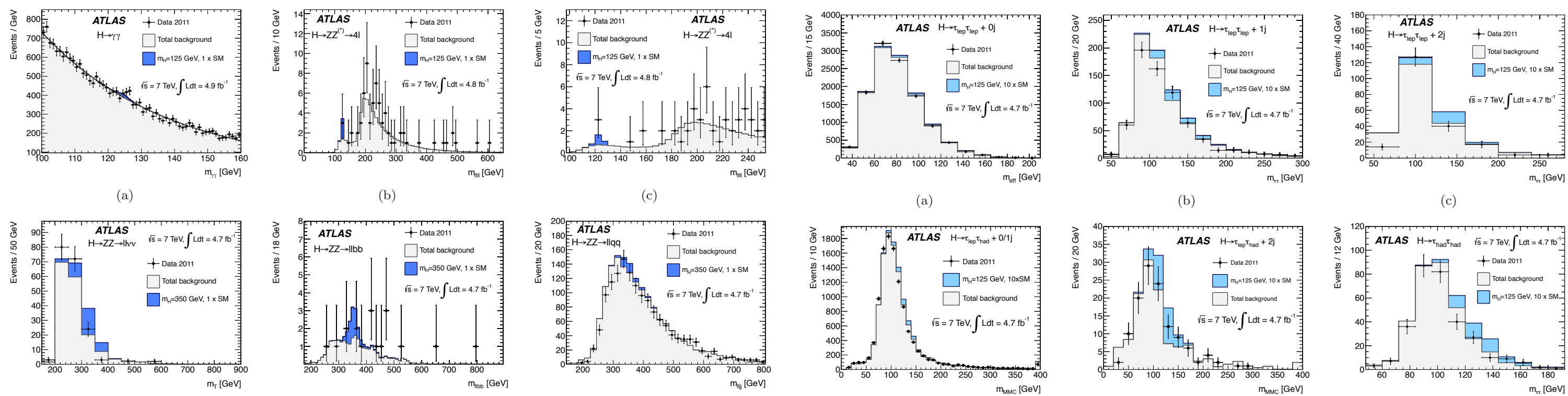
The model dependence in STXS mainly connected to how results are conveyed.

- The phase space regions are just phase space regions, they don't assume any model
- Paired with RECAST one could reinterpret any model using the STXS phase space regions

RECAST & Combinations

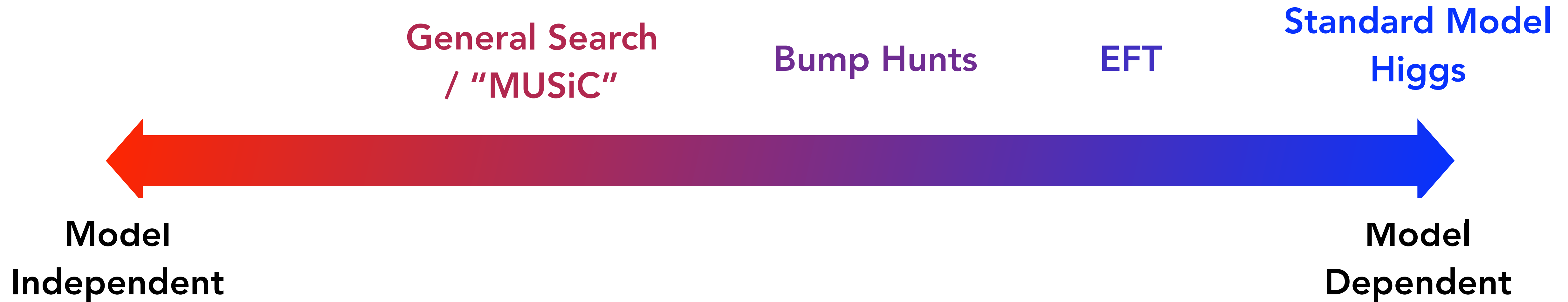
Combining multiple searches is a strategy to enhance sensitivity

- But a protocol is needed to combine different analyses
- A likelihood-based combination is a natural protocol, but it requires knowing how the signal will populate all the different analyses
- A model-independent combination isn't unique and may hurt sensitivity
- With RECAST can run any signal through each analysis and then combine



$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G} | \alpha) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c | \nu_c(\alpha)) \prod_{e=1}^{n_c} f_c(x_{ce} | \alpha) \right] \cdot \prod_{p \in \mathcal{S}} f_p(a_p | \alpha_p)$$

The spectrum revisited



Gaussian Processes allow us to specify model in a language other than QFT that captures intuitive physics. Other approaches along these lines are possible & should be developed.

RECAST allows us to reuse analyses for other purposes, decouple original motivation

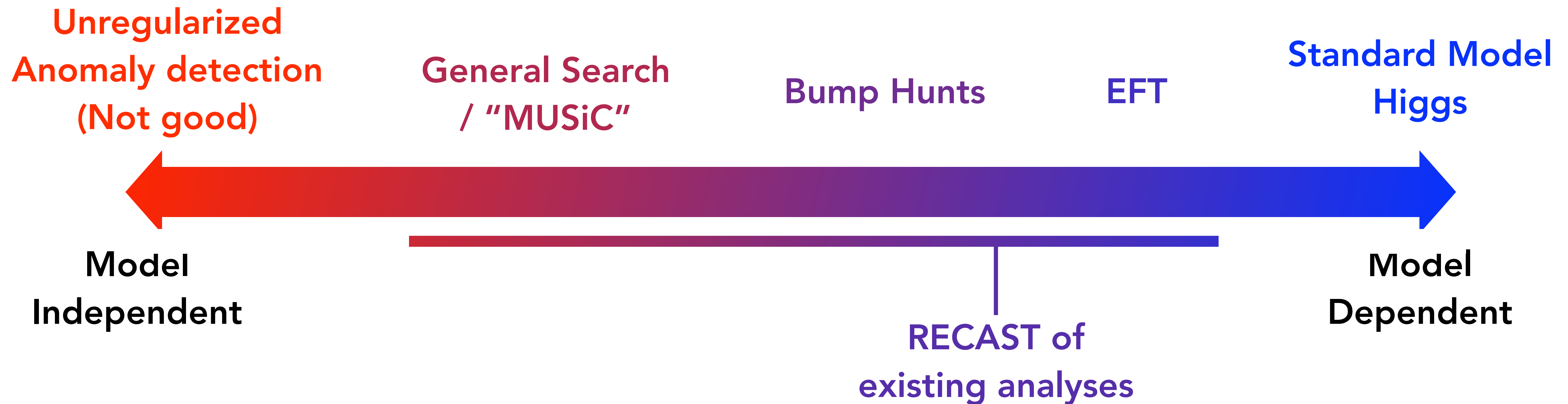
The spectrum revisited



Gaussian Processes allow us to specify model in a language other than QFT that captures intuitive physics. Other approaches along these lines are possible & should be developed.

RECAST allows us to reuse analyses for other purposes, decouple original motivation

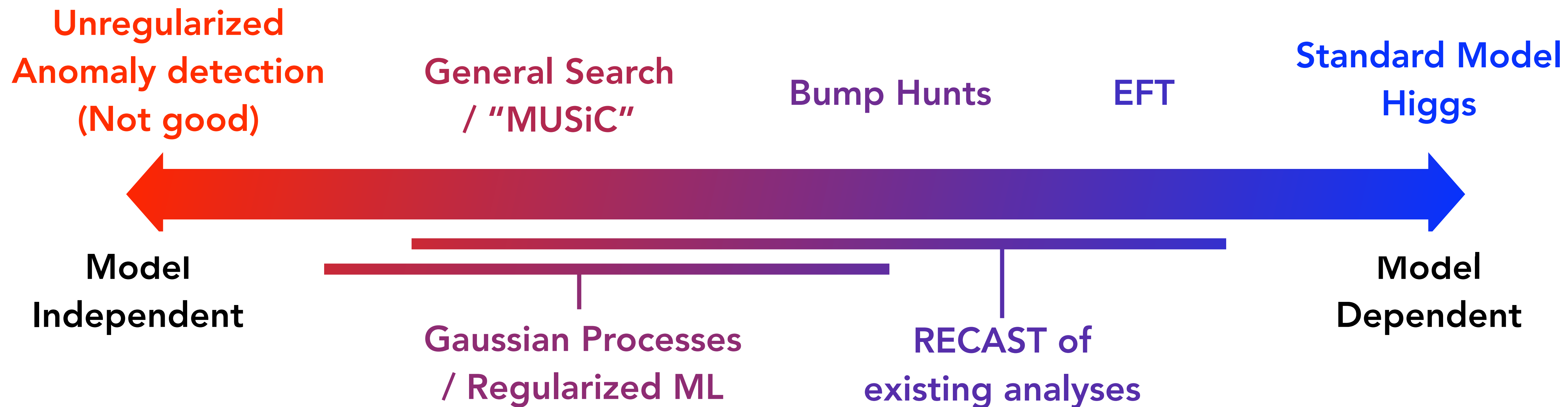
The spectrum revisited



Gaussian Processes allow us to specify model in a language other than QFT that captures intuitive physics. Other approaches along these lines are possible & should be developed.

RECAST allows us to reuse analyses for other purposes, decouple original motivation

The spectrum revisited



Gaussian Processes allow us to specify model in a language other than QFT that captures intuitive physics. Other approaches along these lines are possible & should be developed.

RECAST allows us to reuse analyses for other purposes, decouple original motivation

Backup

Reconstructing the Higgs by exploiting causal structure

Don't believe the media:

$$E \neq mc^2$$

Reconstructing the Higgs by exploiting causal structure

Don't believe the media:

$$E \neq mc^2$$

What Einstein really said:

$$E^2 = (mc^2)^2 + (|\vec{p}|c)^2$$

Reconstructing the Higgs by exploiting causal structure

Don't believe the media:

$$E \neq mc^2$$

What Einstein really said:

$$E^2 = (mc^2)^2 + (|\vec{p}|c)^2$$

Every physics student knows energy and momentum are conserved

$$E_{\text{Higgs}} = E_{\text{before}} = E_{\text{after}} = \sum_i E_i$$
$$\vec{p}_{\text{Higgs}} = \vec{p}_{\text{before}} = \vec{p}_{\text{after}} = \sum_i \vec{p}_i$$

Reconstructing the Higgs by exploiting causal structure

Don't believe the media:

$$E \neq mc^2$$

What Einstein really said:

$$E^2 = (mc^2)^2 + (|\vec{p}|c)^2$$

Every physics student knows energy and momentum are conserved

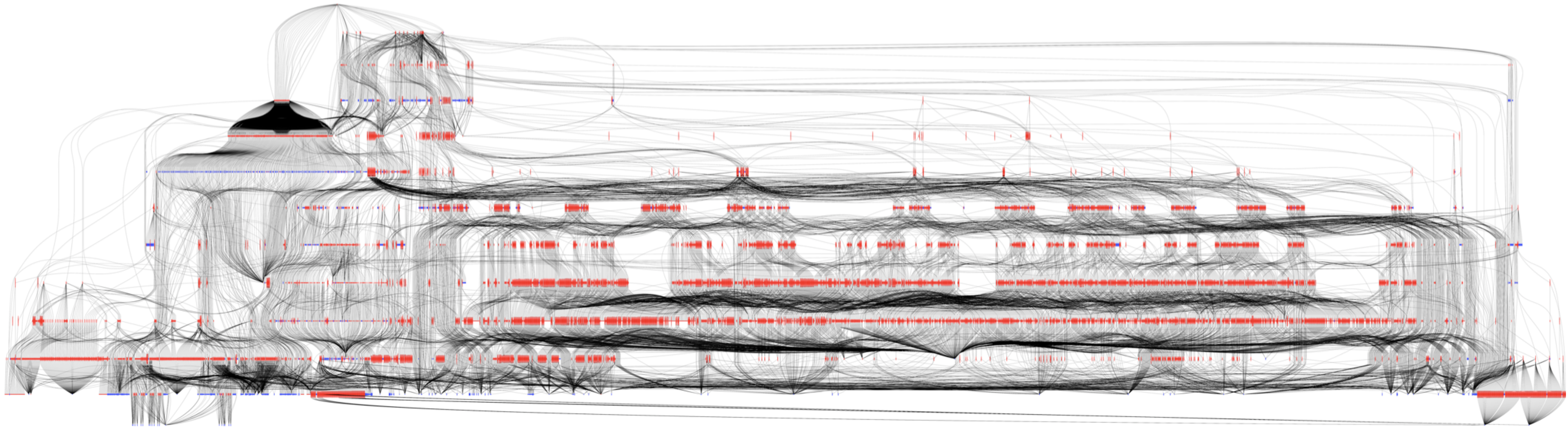
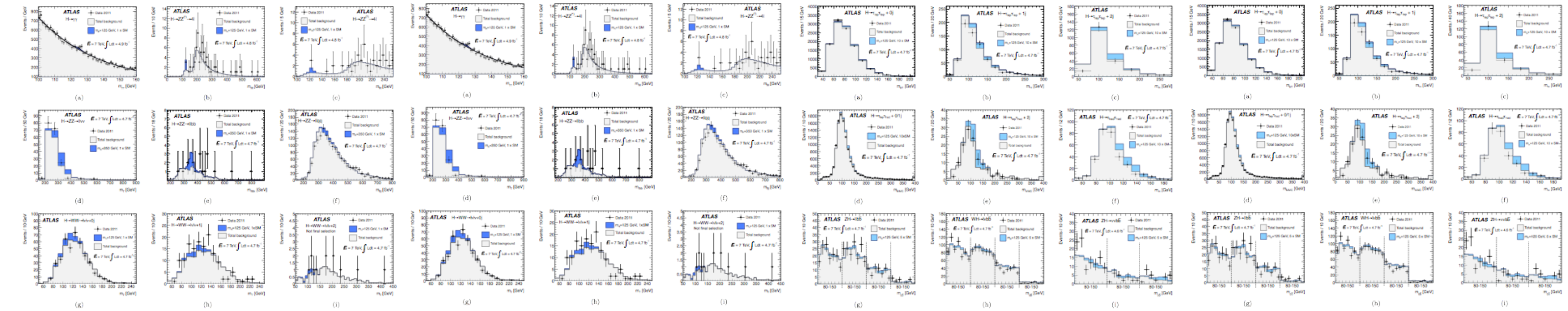
$$E_{\text{Higgs}} = E_{\text{before}} = E_{\text{after}} = \sum_i E_i$$

$$\vec{p}_{\text{Higgs}} = \vec{p}_{\text{before}} = \vec{p}_{\text{after}} = \sum_i \vec{p}_i$$

Thus, we can estimate the mass of the Higgs particle with

$$m_H = \sqrt{E_{\text{after}}^2/c^4 - |\vec{p}_{\text{after}}|^2/c^2}$$

Collaborative Statistical Modeling



$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G} | \boldsymbol{\alpha}) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c | \nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce} | \boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathcal{S}} f_p(a_p | \alpha_p)$$

Pendulum

$$P(\text{theory} \mid \text{data}) \propto P(\text{data} \mid \text{theory}) P(\text{theory})$$

— Thomas Bayes



Traditional approaches in physics

- hand-crafted data analysis
- largely guided by expert knowledge and theoretical insights



Big Data & Deep Learning

- eschew expert knowledge
- end-to-end learning
- data-driven

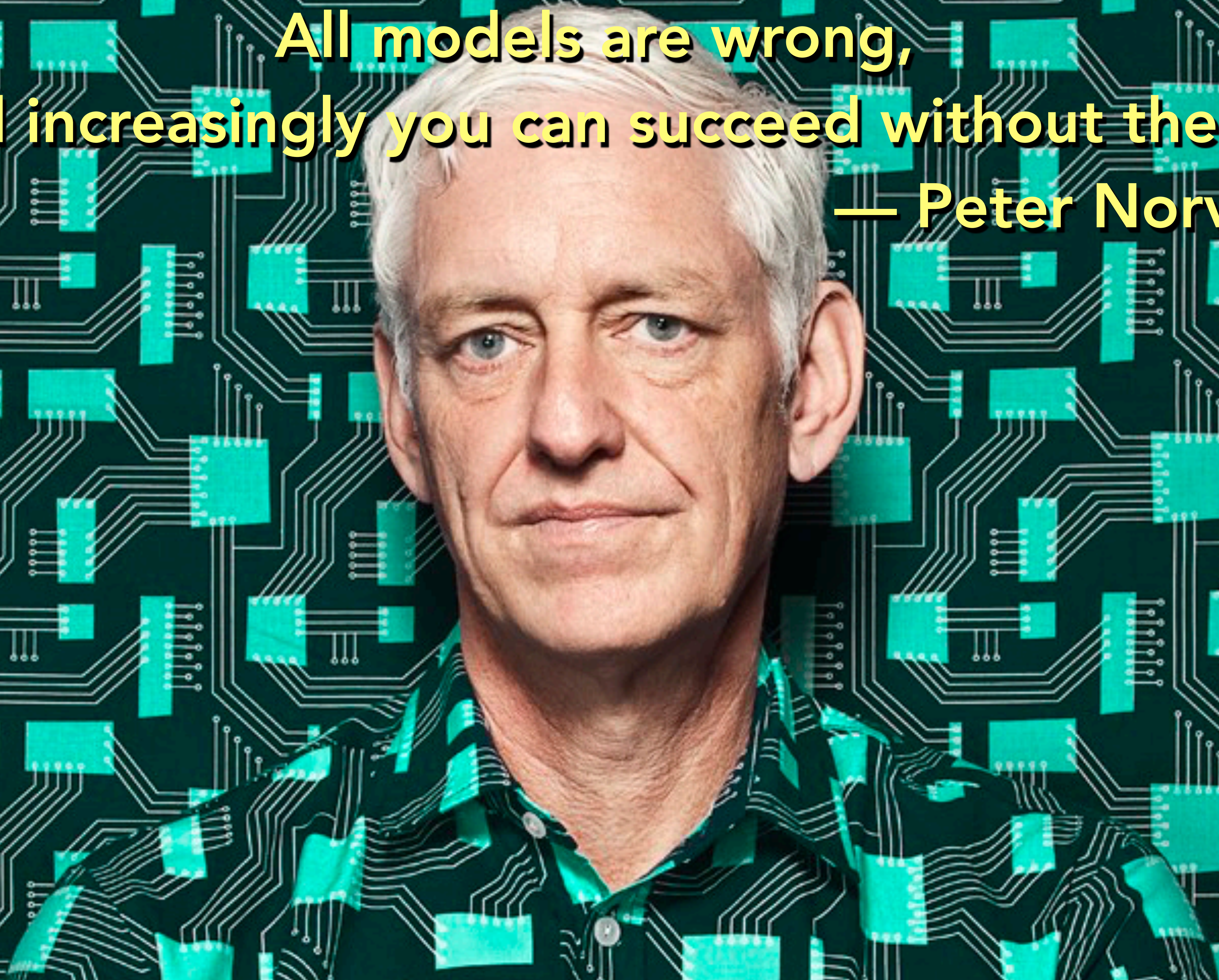


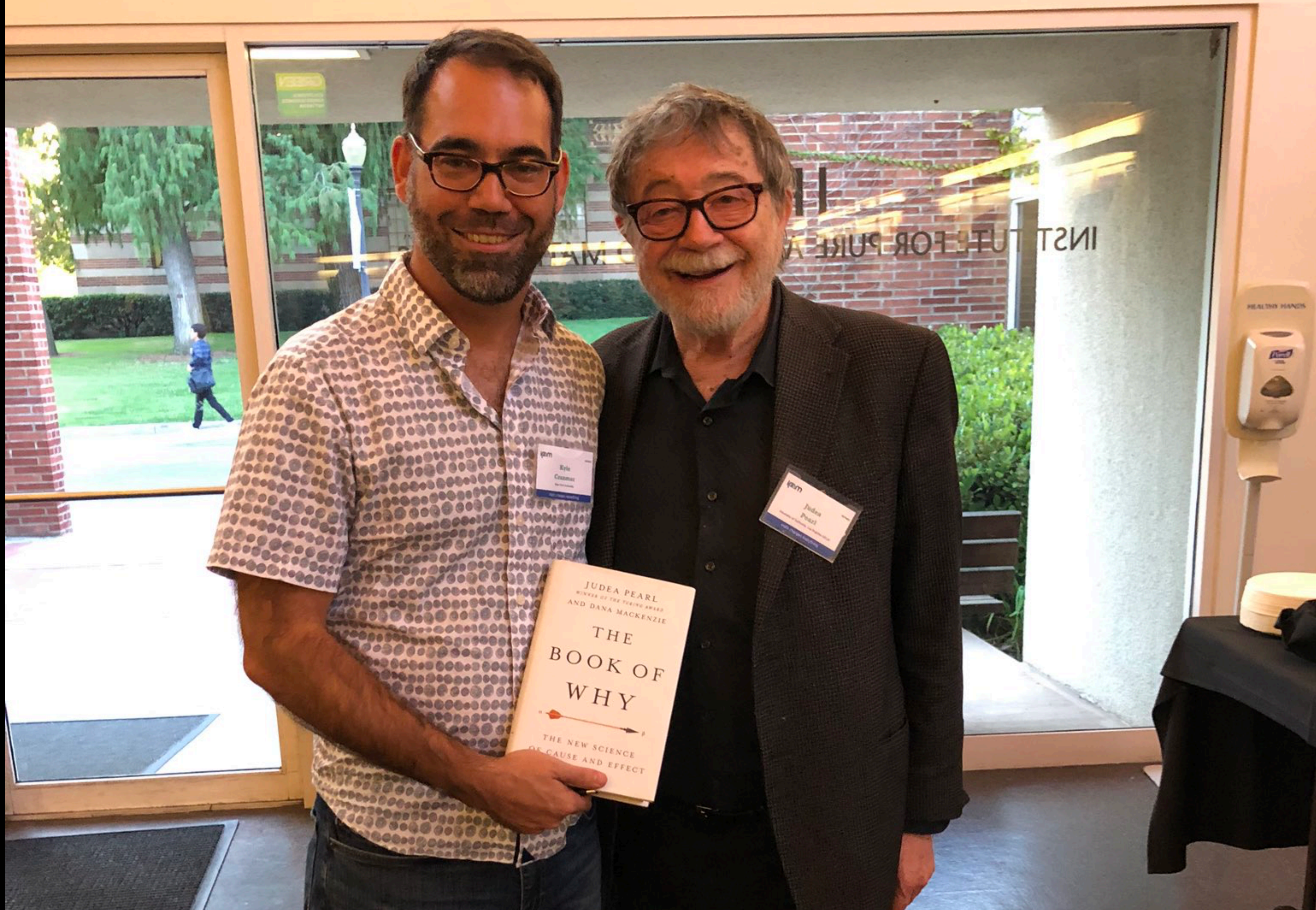
THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

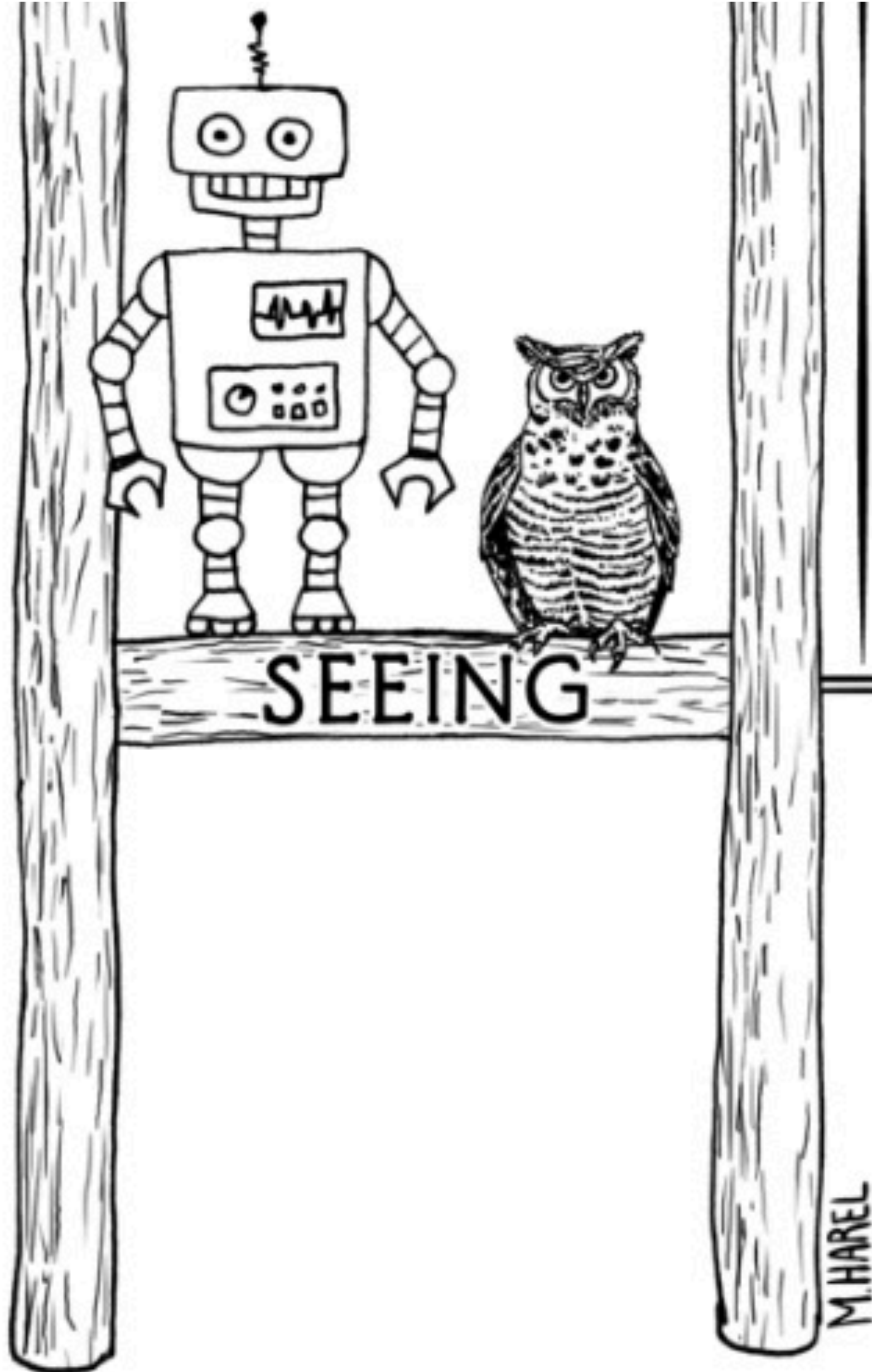


**All models are wrong,
and increasingly you can succeed without them.**

— Peter Norvig





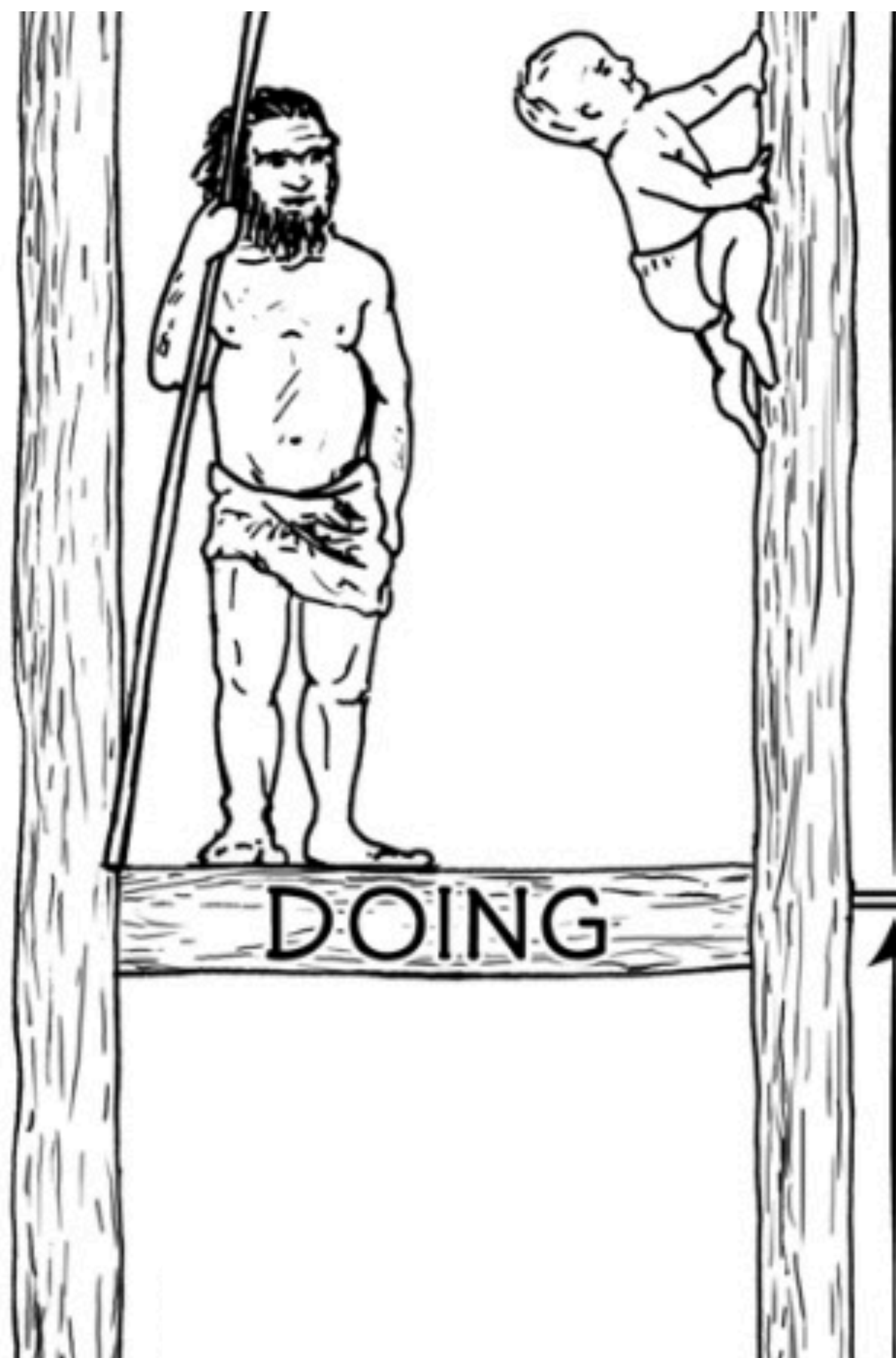


1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the
election results?

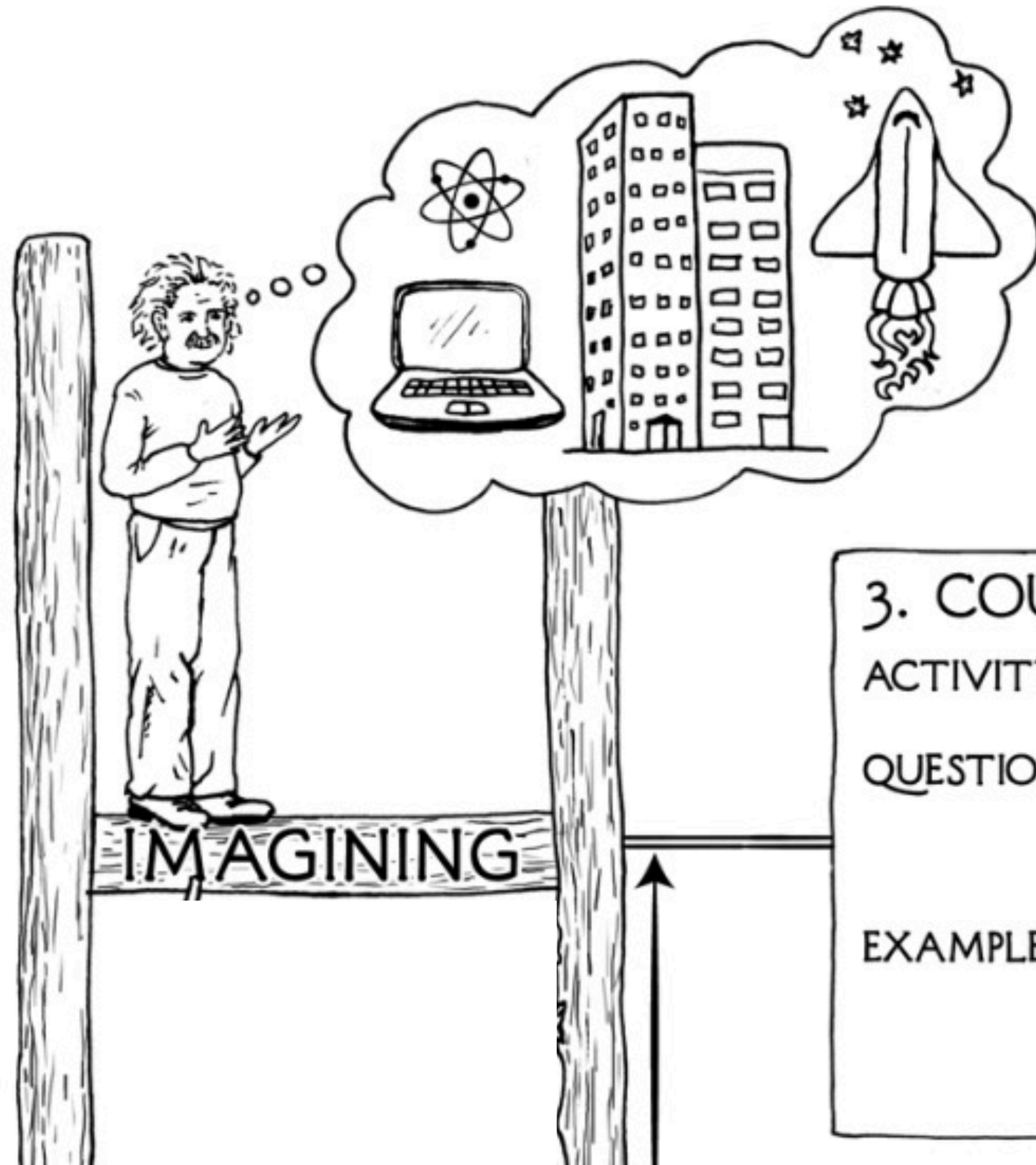


2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do ...? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done ...? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache?
Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

A toy example

Ferenc Huszár

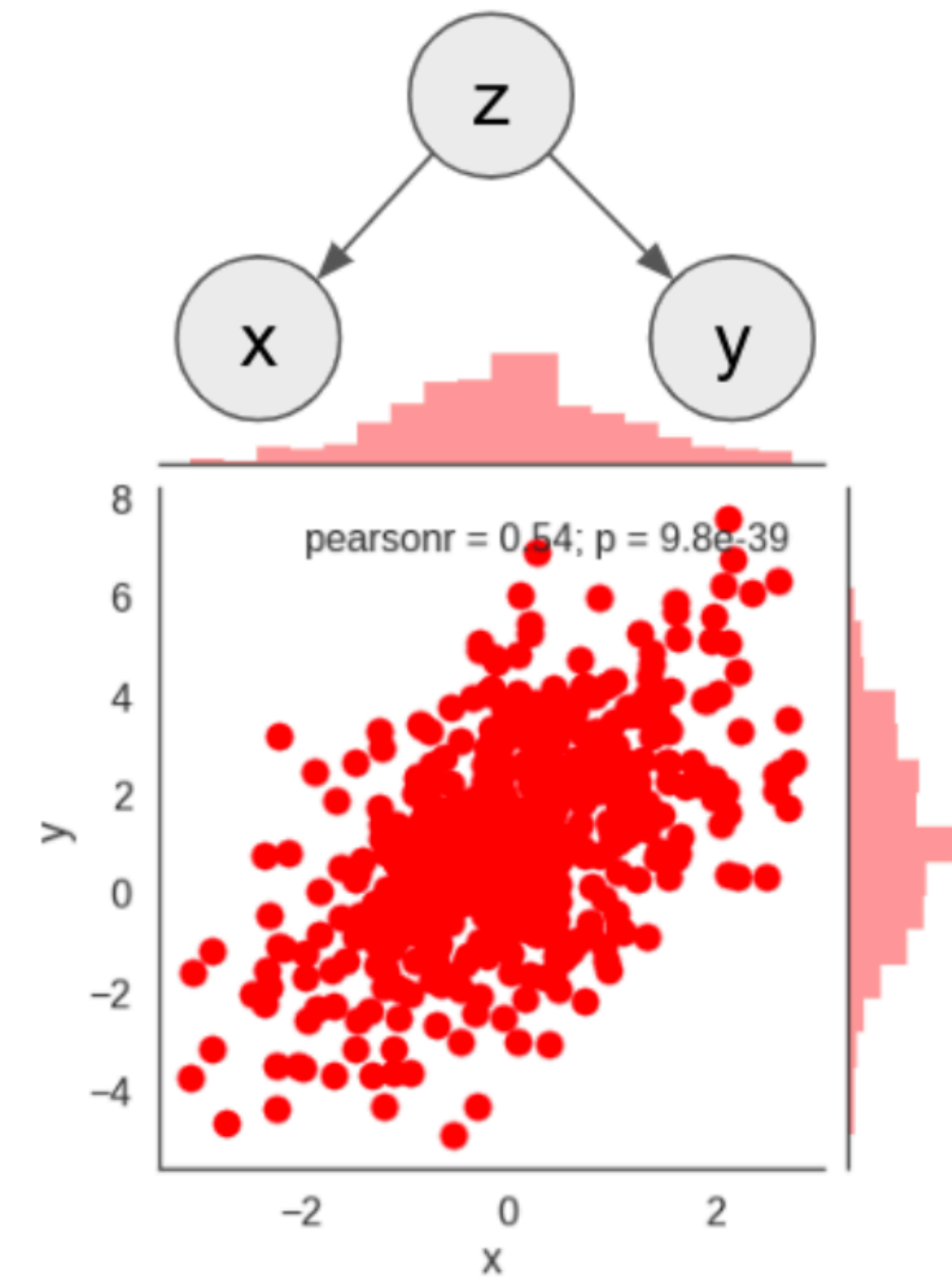
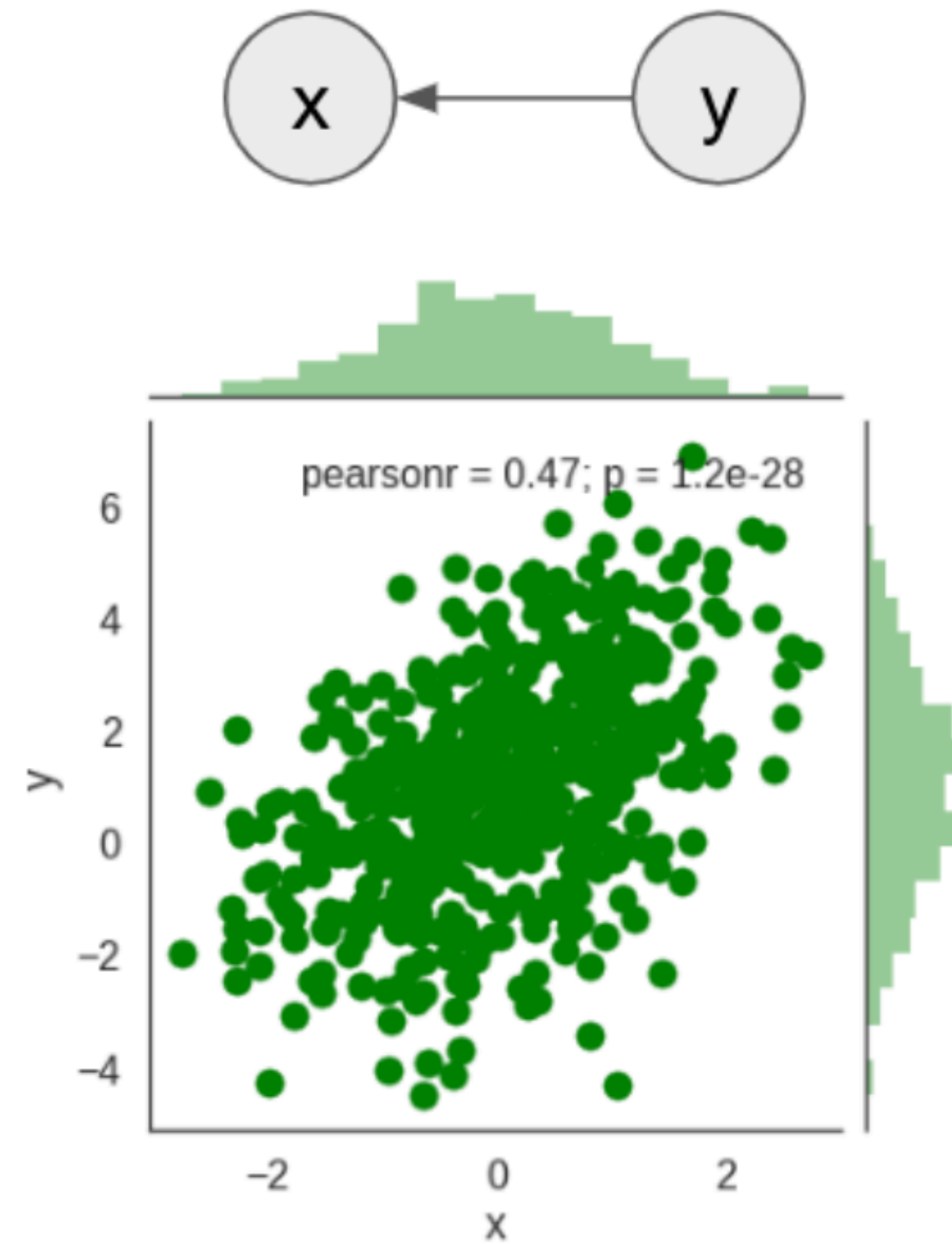
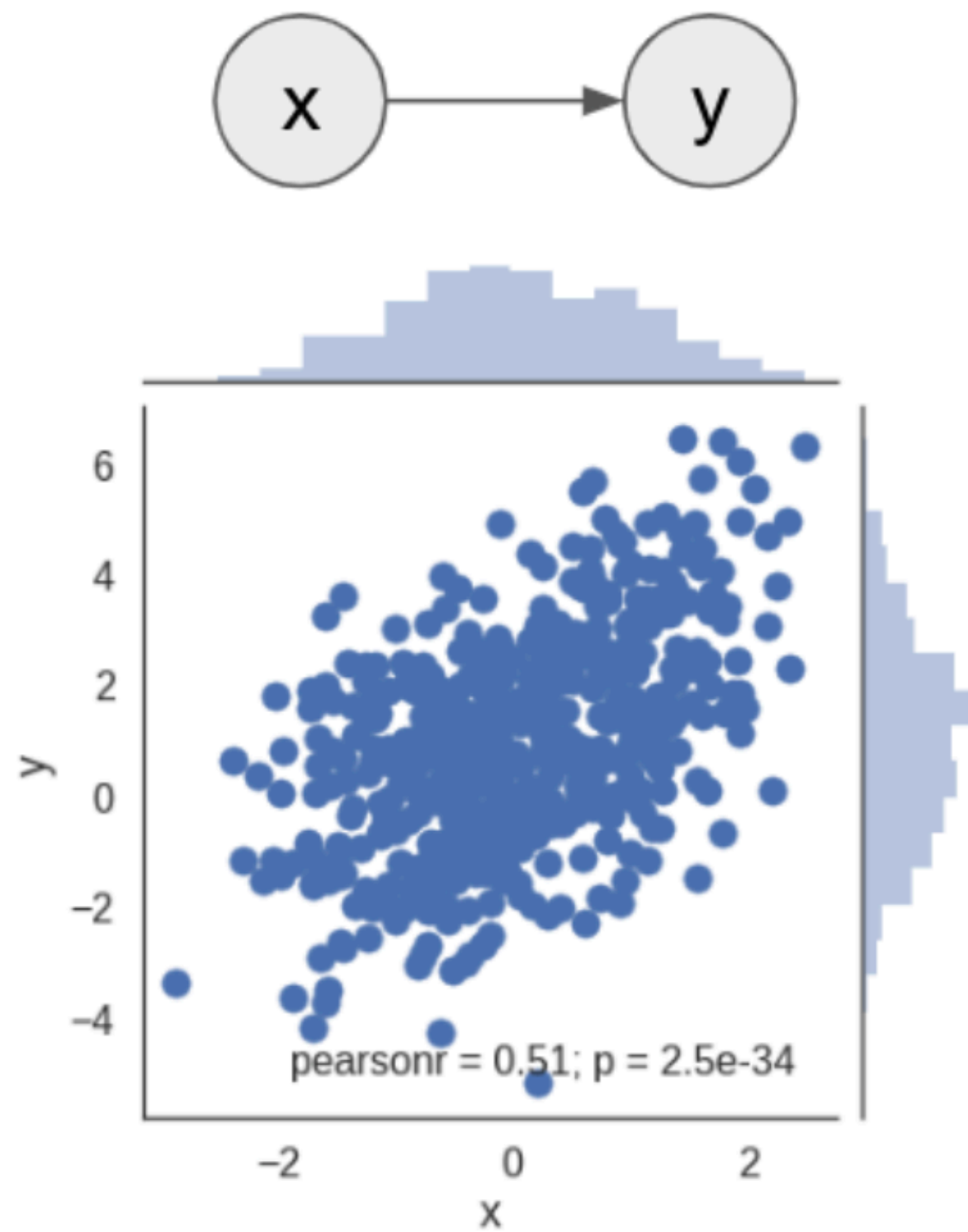


inFERENCE

```
x = randn()  
y = x + 1 + sqrt(3)*randn()
```

```
y = 1 + 2*randn()  
x = (y-1)/4 + sqrt(3)*randn()/2
```

```
z = randn()  
y = z + 1 + sqrt(3)*randn()  
x = z
```

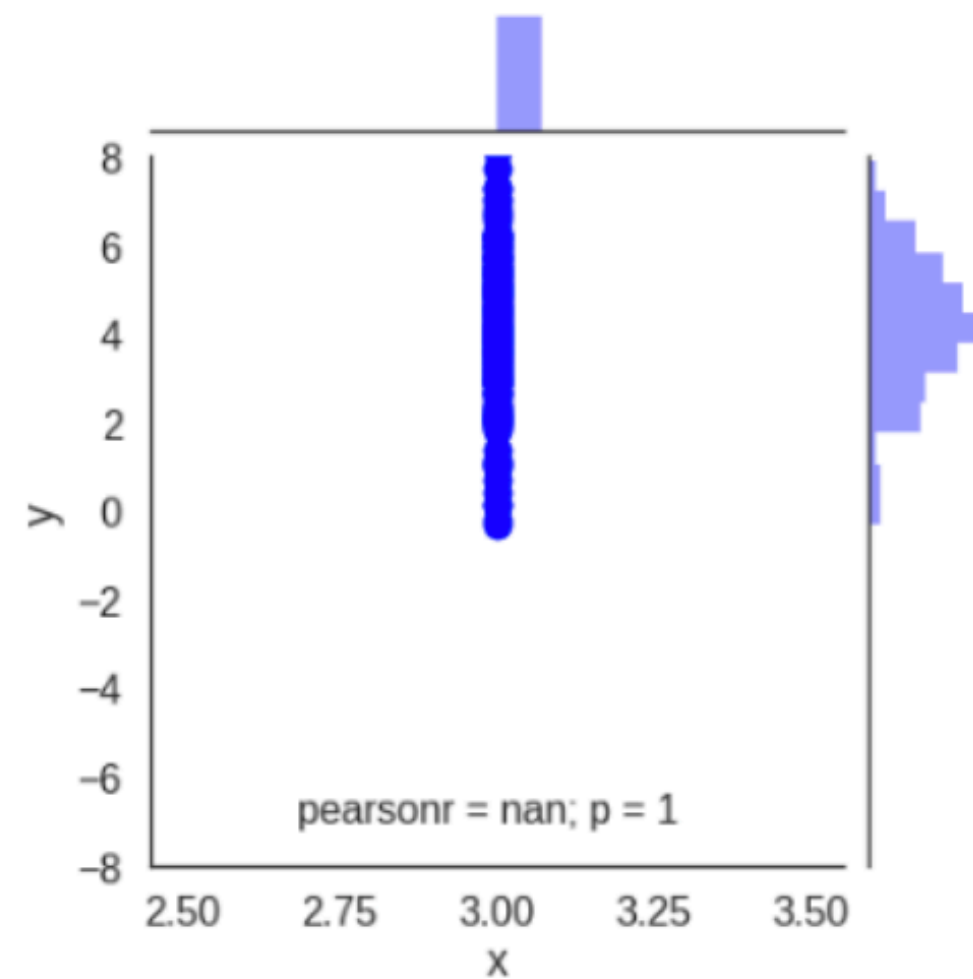


A toy example



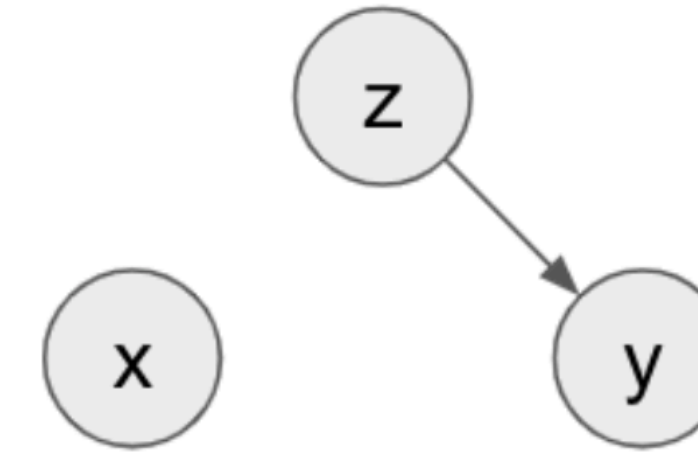
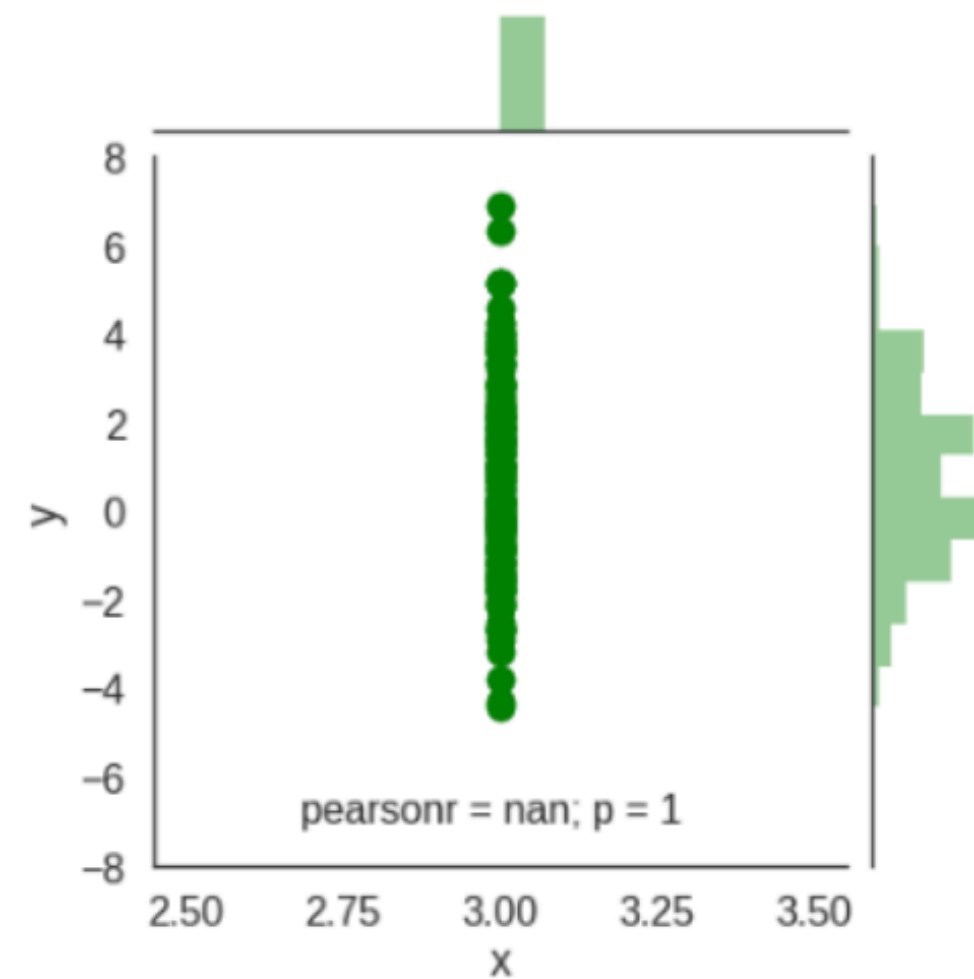
$$P(y|do(X)) = p(y|x)$$

```
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3
```



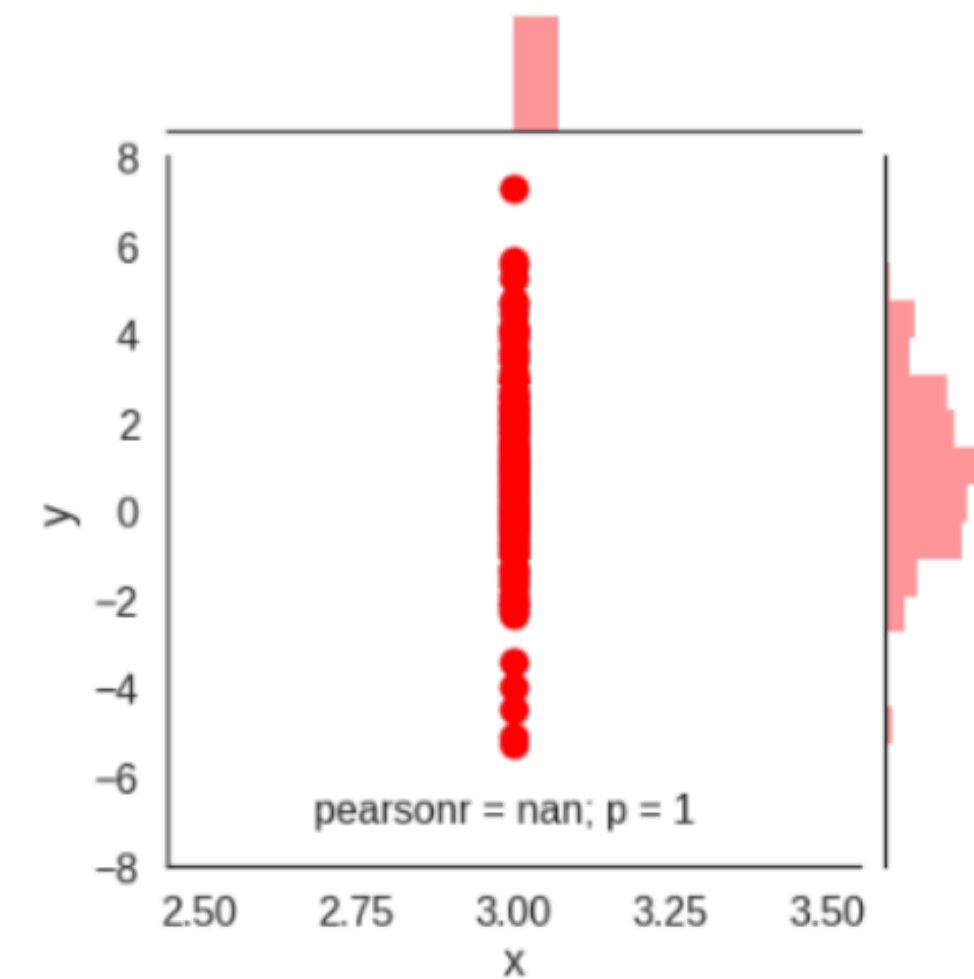
$$P(y|do(X)) = p(y)$$

```
y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3
```



$$P(y|do(X)) = p(y)$$

```
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3
```



STATISTICAL DECISION THEORY
&
JAMES-STEIN ESTIMATOR

Cramér-Rao Bound

The minimum variance bound on an unbiased estimator is given by the Cramér-Rao bound:

$$\text{cov}[\hat{\theta}|\theta_0]_{ij} \geq I_{ij}^{-1}(\theta_0)$$

Expected error of best-fit parameter Inverse of Fisher information

Fisher information matrix (is also a Riemannian metric!)

$$I_{ij}[\theta] = -\mathbb{E} \left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta_i \partial \theta_j} \middle| \theta \right]$$

Maximum Likelihood Estimators *asymptotically* reach this bound

STATISTICAL DECISION THEORY IN 1 SLIDE

Θ - States of nature; X - possible observations;

A - action to be taken

$f(x|\theta)$ - statistical model; $\pi(\theta)$ - prior

$\delta: X \rightarrow A$ - **decision rule** (take some action based on observation)

$L: \Theta \times A \rightarrow \mathbb{R}$ - **loss function**, real-valued function true parameter and action

$R(\theta, \delta) = E_{f(x|\theta)}[L(\theta, \delta)]$ - **risk**

Decision rule that minimizes risk depends on unknown value of θ

$r(\pi, \delta) = E_{\pi(\theta)}[R(\theta, \delta)]$ - **Bayes risk** (expectation over θ w.r.t. prior and possible observations)

Your risk, your prior

CRAMÉR-RAO BOUND

The minimum variance bound on an estimator is given by the Cramér-Rao inequality:

- ▶ simple univariate case:

$$\text{Var}[\hat{\theta}|\theta] = E[(\hat{\theta} - E[\hat{\theta}|\theta])^2 | \theta]$$

- ▶ For an unbiased estimator the Cramér-Rao bound states

$$\text{Var}[\hat{\theta}|\theta] \geq \frac{1}{I(\theta)}$$

- ▶ where $I(\theta)$ is the Fisher information

$$(I(\theta))_{i,j} = E \left[\frac{\partial}{\partial \theta_i} \ln f(X; \theta) \frac{\partial}{\partial \theta_j} \ln f(X; \theta) \middle| \theta \right].$$

- ▶ General form for multiple parameters:

$$\text{cov}[\hat{\theta}|\theta]_{ij} \geq I_{ij}^{-1}(\theta)$$

Maximum Likelihood Estimators *asymptotically* reach this bound

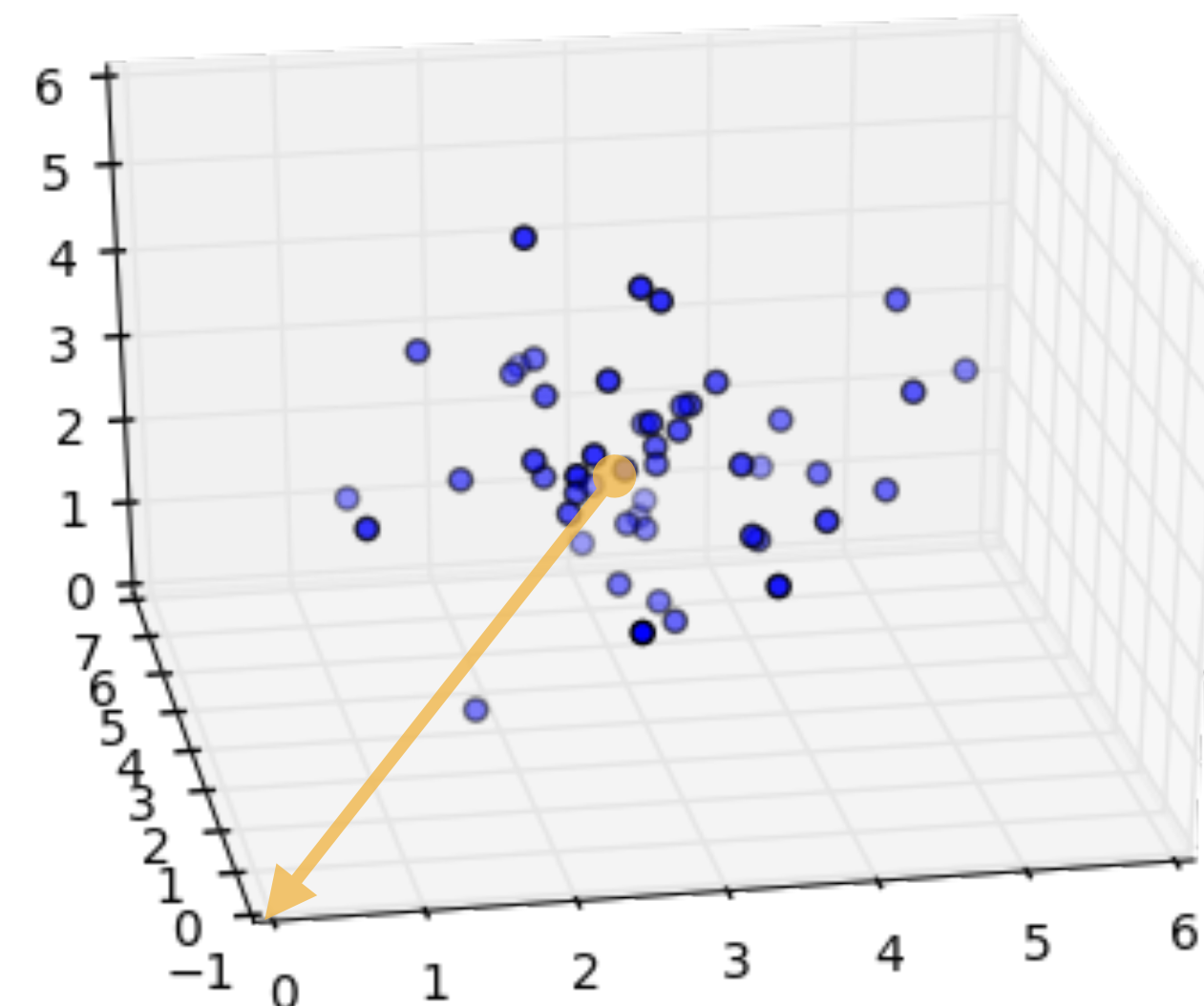
JAMES-STEIN ESTIMATOR

Consider a standard multivariate Gaussian distribution for \vec{x} in n dimensions centered around $\vec{\mu}$

$$f(\vec{x}|\vec{\mu}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_i)^2}{2}\right).$$

Goal: minimize mean-squared error

$$MSE[\hat{\vec{\mu}}] = E[||\hat{\vec{\mu}} - \vec{\mu}||^2]$$



MLE (unbiased)

$$\hat{\vec{\mu}}_{MLE} = \bar{x} = \frac{1}{m} \sum_{j=1}^m \vec{x}_j$$

James-Stein (weird)

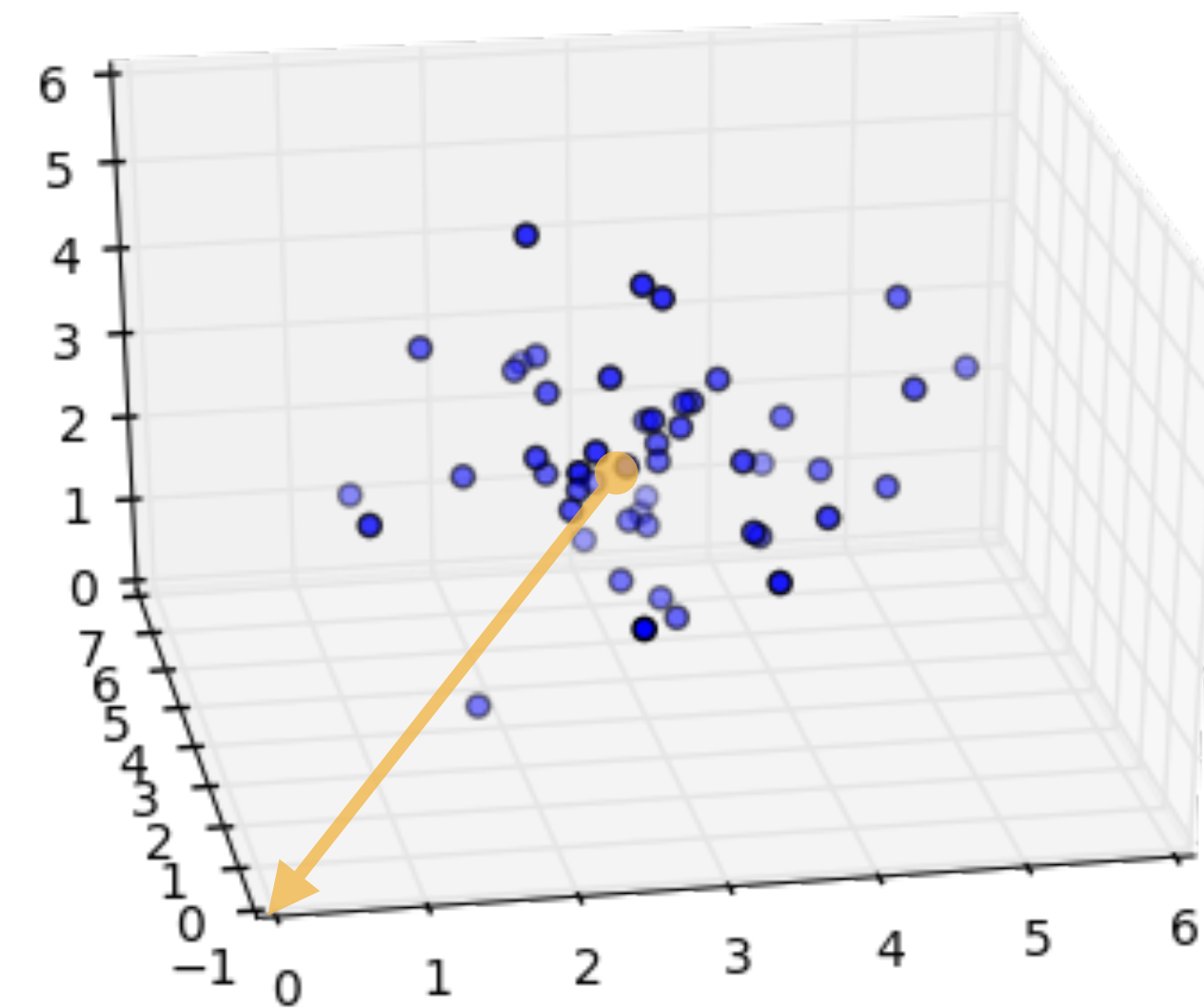
$$\hat{\mu}_{JS} = \left(1 - \frac{n-2}{||\bar{x}||^2}\right) \bar{x}$$

JAMES-STEIN ESTIMATOR

The James-Stein estimator seems like a horrible suggestion

$$\hat{\mu}_{JS} = \left(1 - \frac{n-2}{\|\bar{x}\|^2}\right) \bar{x}$$

- clearly biased (MLE is not)
- shifts towards origin is not translationally invariant
 $x \rightarrow x' = x + \Delta$

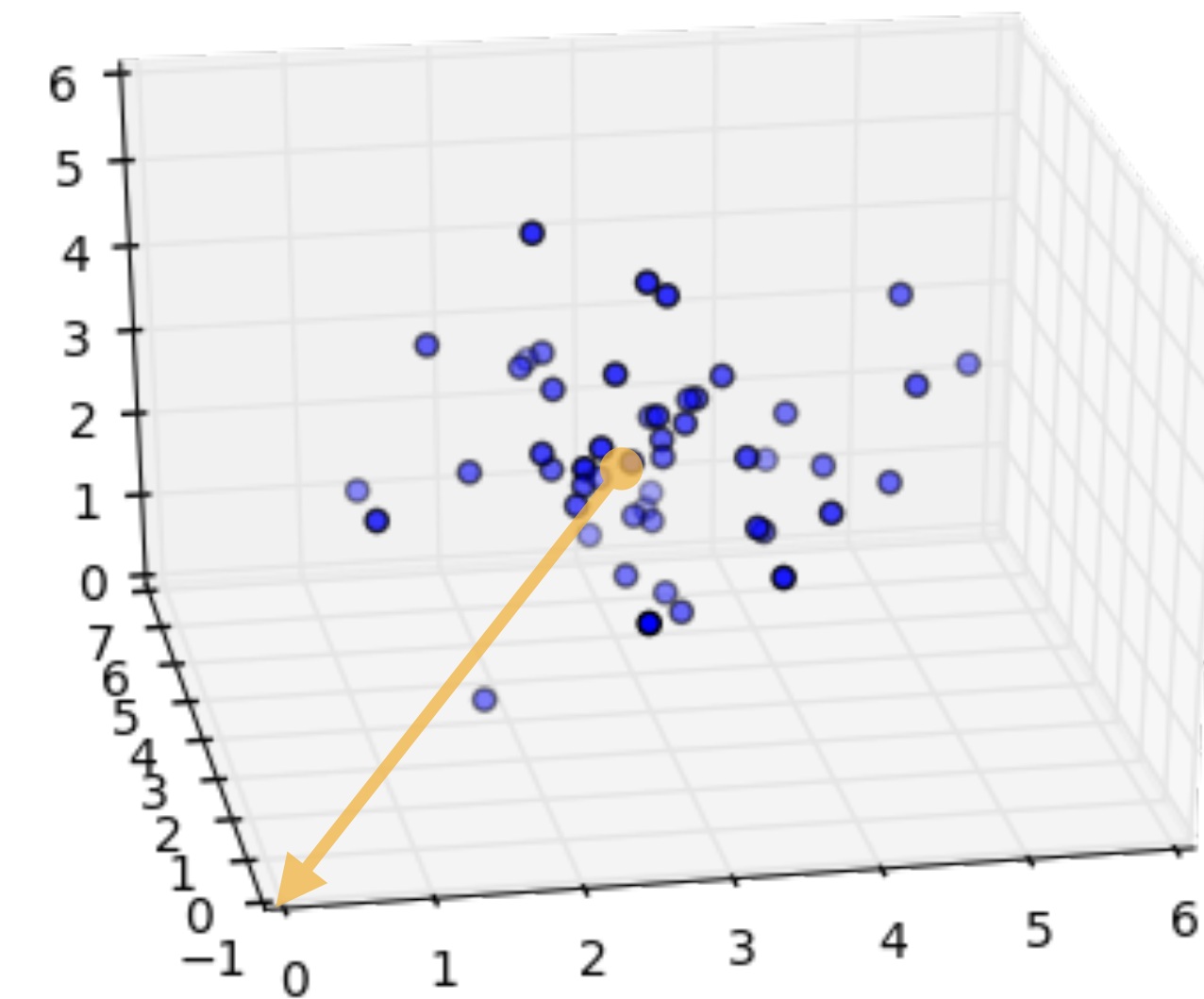


JAMES-STEIN ESTIMATOR

The James-Stein estimator seems like a horrible suggestion

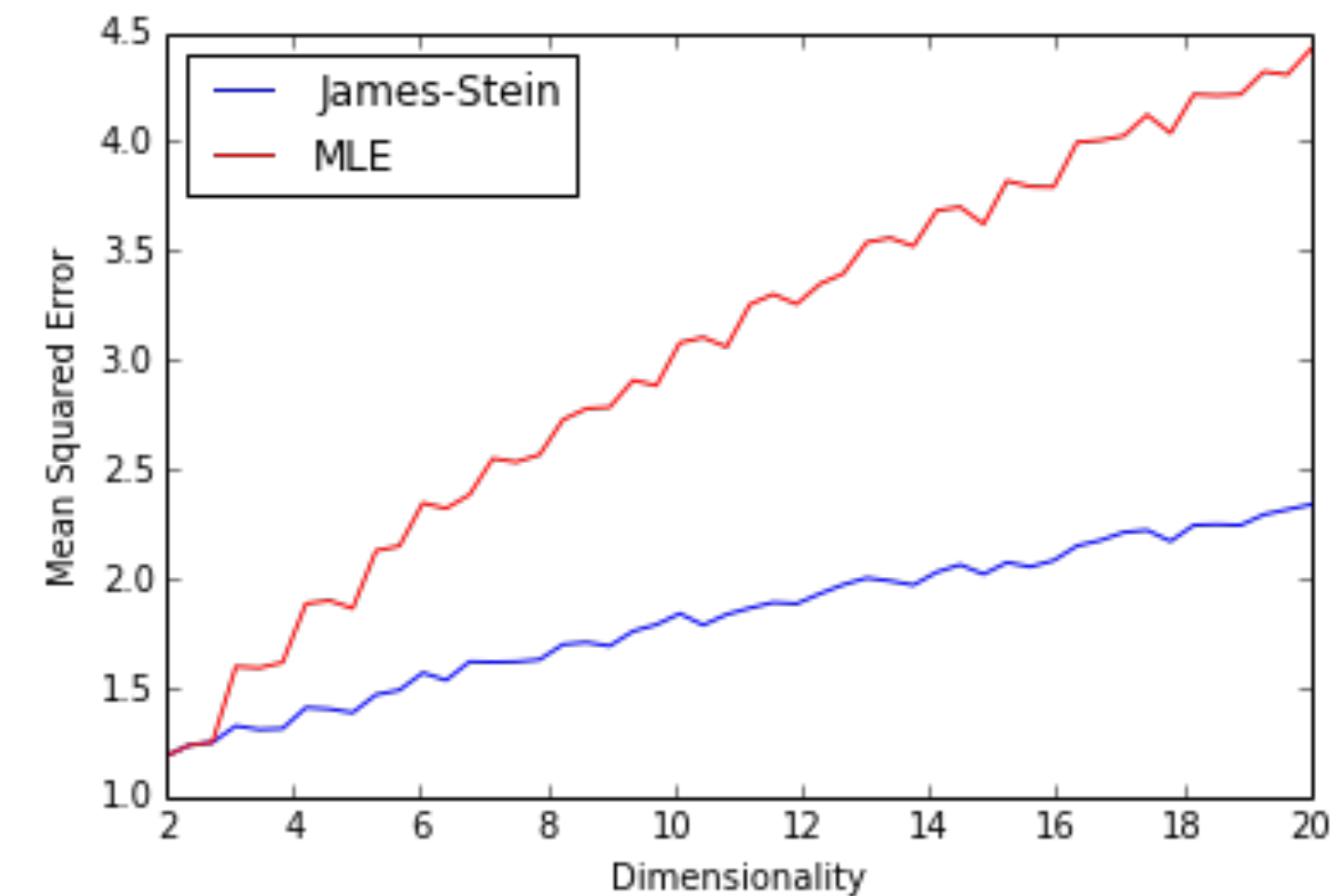
$$\hat{\mu}_{JS} = \left(1 - \frac{n-2}{\|\bar{x}\|^2}\right) \bar{x}$$

- clearly biased (MLE is not)
- shifts towards origin is not translationally invariant
 $x \rightarrow x' = x + \Delta$



Yet, it has smaller mean squared error than MLE for $n > 2$!

- it "dominates" the MLE



BIAS/VARIANCE TRADEOFF

We introduced Bias and Variance of estimators

$$\text{Var}[\hat{\mu}|\mu] = E[(\hat{\mu} - E[\hat{\mu}|\mu])^2] | \mu]$$

Most physicist are allergic to the idea of a biased estimator

- try to find unbiased estimator with smallest variance
- hence importance of Cramér-Rao bound

But what if we just want to minimize the mean-squared error?

$$MSE[\hat{\mu}|\mu] = E[(\hat{\mu} - \mu)^2] | \mu]$$

it decomposes like this

$$MSE[\hat{\mu}|\mu] = \text{Var}[\hat{\mu}|\mu] + (\text{Bias}[\hat{\mu}|\mu])^2$$

So it encodes some relative weight to bias and variance. Think harder!

STATISTICAL DECISION THEORY IN 1 SLIDE

Θ - States of nature; X - possible observations; A - action to be taken

$f(x|\theta)$ - statistical model; $\pi(\theta)$ - prior

$\delta: X \rightarrow A$ - **decision rule** (take some action based on observation)

$L: \Theta \times A \rightarrow \mathbb{R}$ - **loss function**, real-valued function true parameter and action

$R(\theta, \delta) = E_{f(x|\theta)}[L(\theta, \delta)]$ - **risk**

- A decision δ^* rule **dominates** a decision rule δ if and only if $R(\theta, \delta^*) \leq R(\theta, \delta)$ for all θ , and the inequality is strict for some θ .
- A decision rule is **admissible** if and only if no other rule dominates it; otherwise it is inadmissible

$r(\pi, \delta) = E_{\pi(\theta)}[R(\theta, \delta)]$ - **Bayes risk** (expectation over θ w.r.t. prior and possible observations)

$\rho(\pi, \delta | x) = E_{\pi(\theta|x)}[L(\theta, \delta(x))]$ - **expected loss** (expectation over θ w.r.t. posterior $\pi(\theta|x)$)

- δ' is a (generalized) Bayes rule if it minimizes the expected loss
- under mild conditions every admissible rule is a (generalized) Bayes rule (**with respect to some prior** — possibly an improper one and not necessarily your prior — that favors distributions where that rule achieves low risk). Thus, in frequentist decision theory it is sufficient to consider only (generalized) Bayes rules.
- Conversely, while Bayes rules with respect to proper priors are virtually always admissible, generalized Bayes rules corresponding to improper priors need not yield admissible procedures. Stein's example is one such famous situation.