Advanced DL in Science

Tobias Golling, University of Geneva

Setting the stage

Personal view

Selection of what I consider interesting & promising

Lots of open-ended questions – like in real research!

At times speculative, provocative, exploratory,…

Let’s make it interactive! [I added some questions for you]

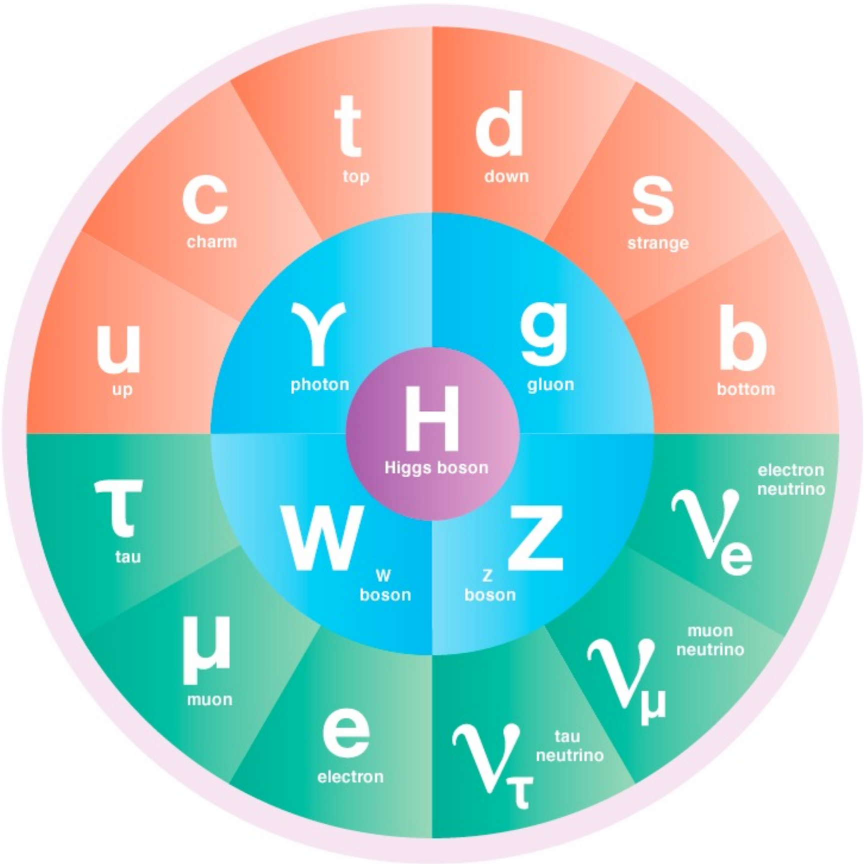
Trying to tell you a story of **AI for Science**

…and teach you a bit of ML on the side

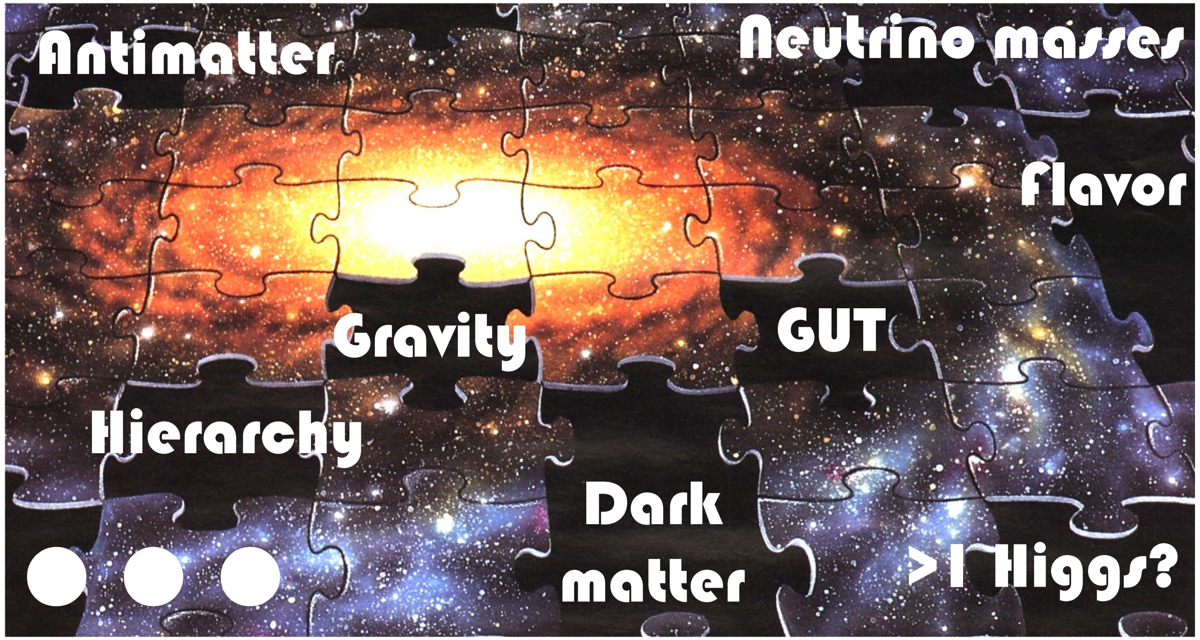
THE HEP BACKSTORY

The SM: blessing & curse…

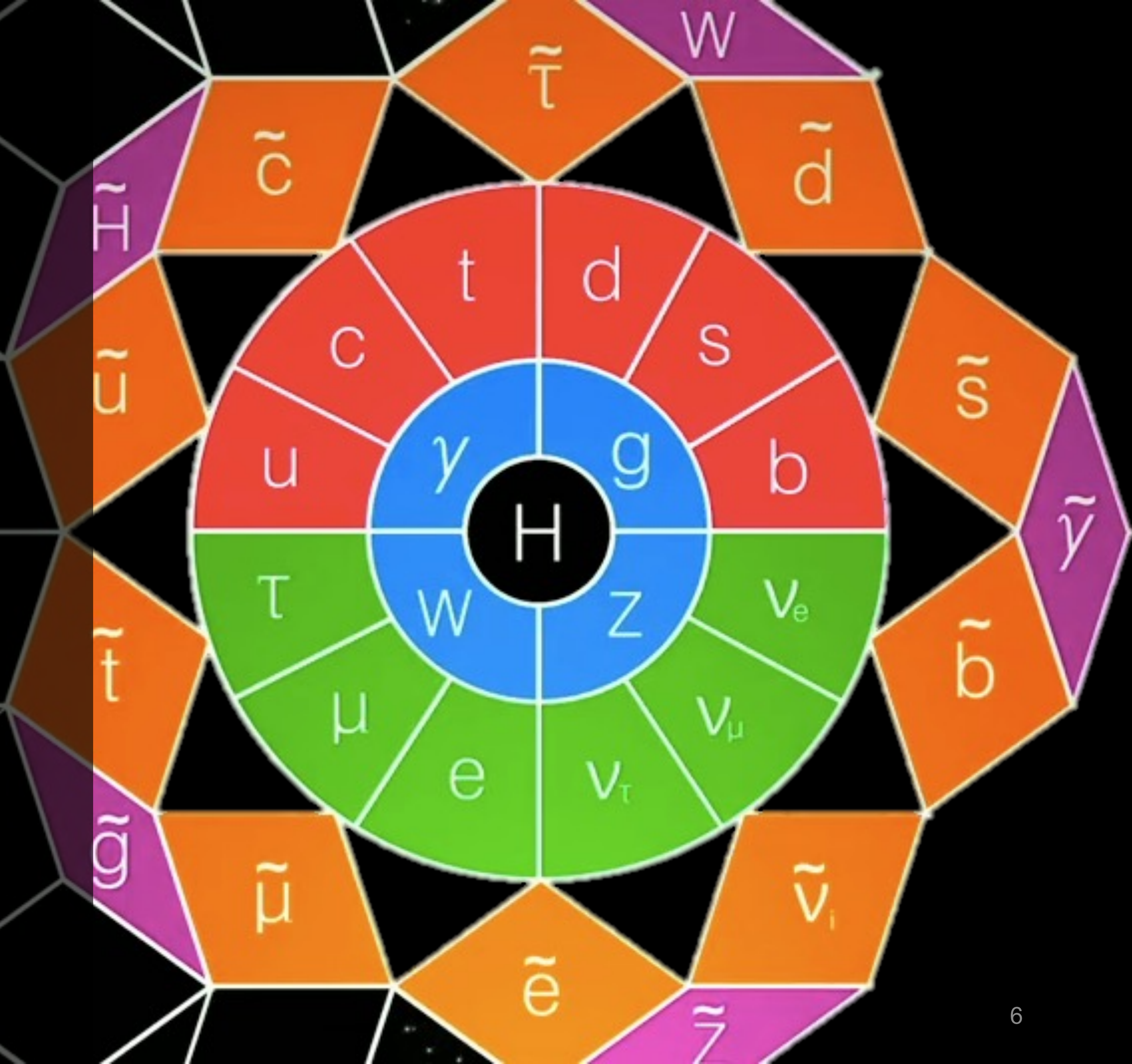
Confirmed by **every** PP experiment **ever** conducted !



Yet, open mysteries remain:



What lies beyond the SM?

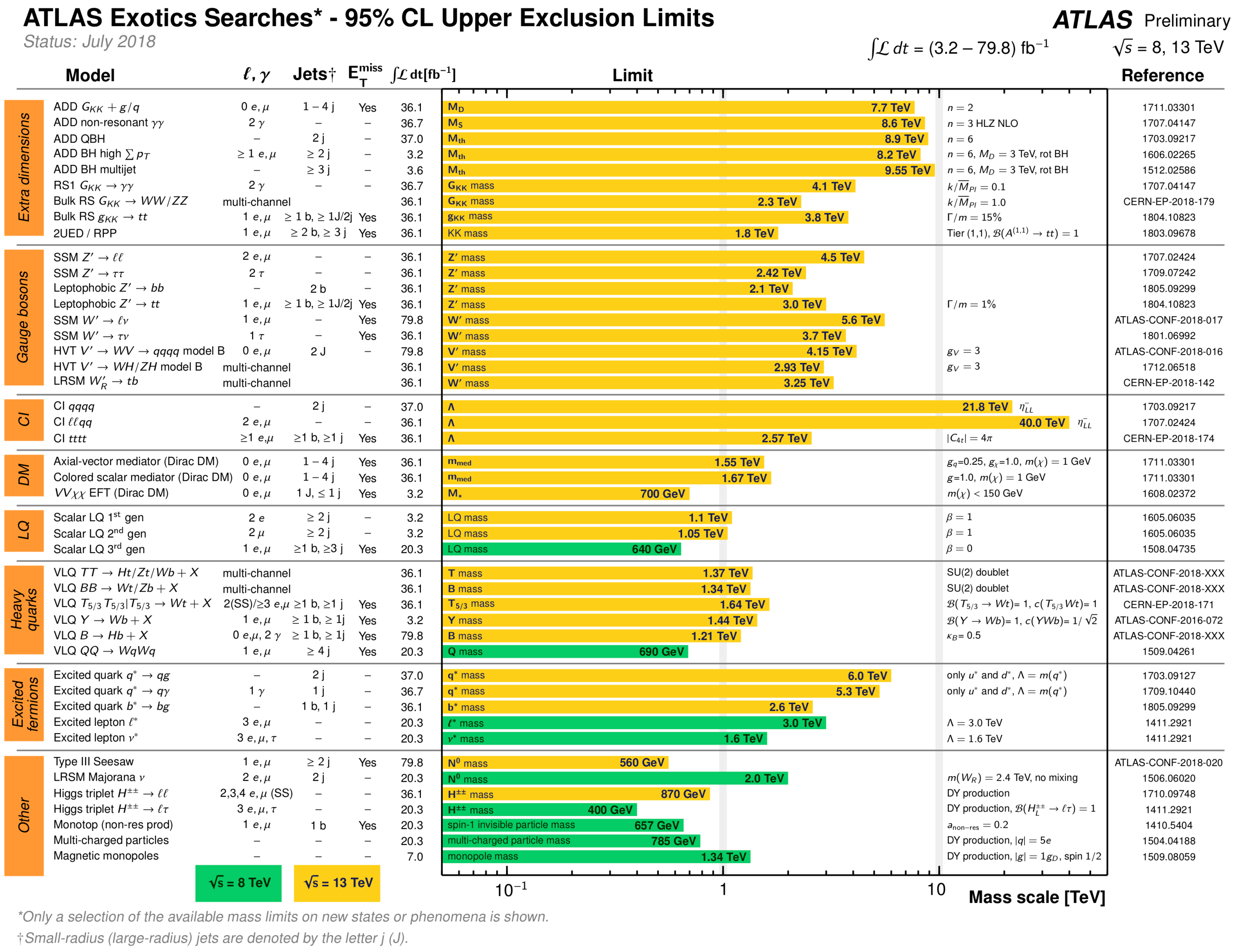


LHC interim evaluation

No sign of

physics beyond the SM

(BSM)



Today’s pulse of particle physics

Exciting new experimental probes !

new LHC upgrades
new gravitational wave signals
new direct detection experiments
new experiments looking for XYZ new physics
new astrophysical probes
new opportunities

The utter lack of discoveries !

A deep theory-fatigue

What next ?

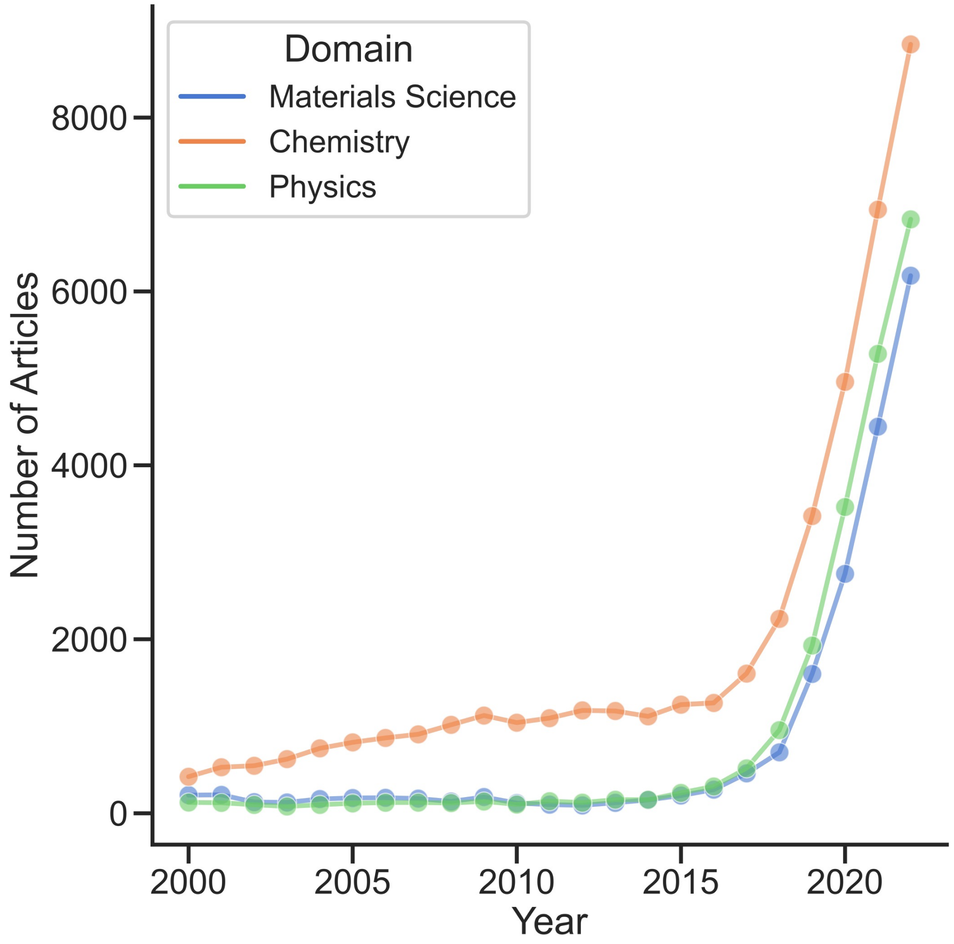
[adopted from Veronica Sanz]

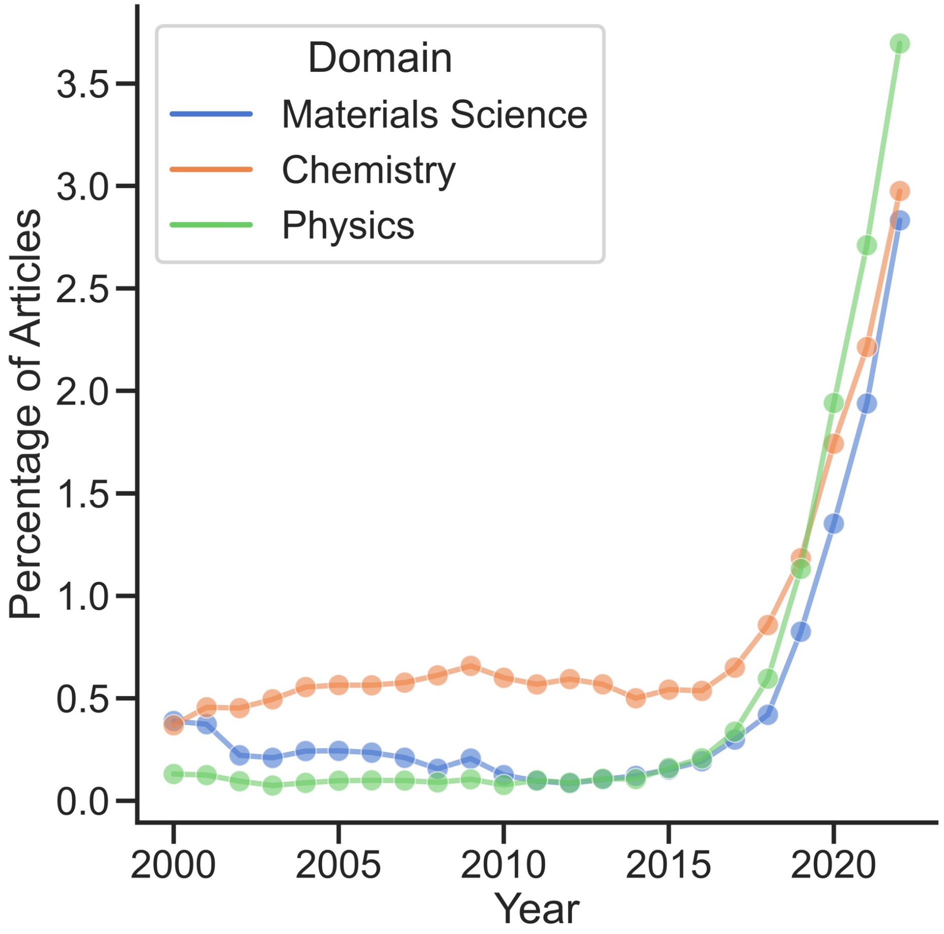
THE RISE OF AI / ML

Machine learning

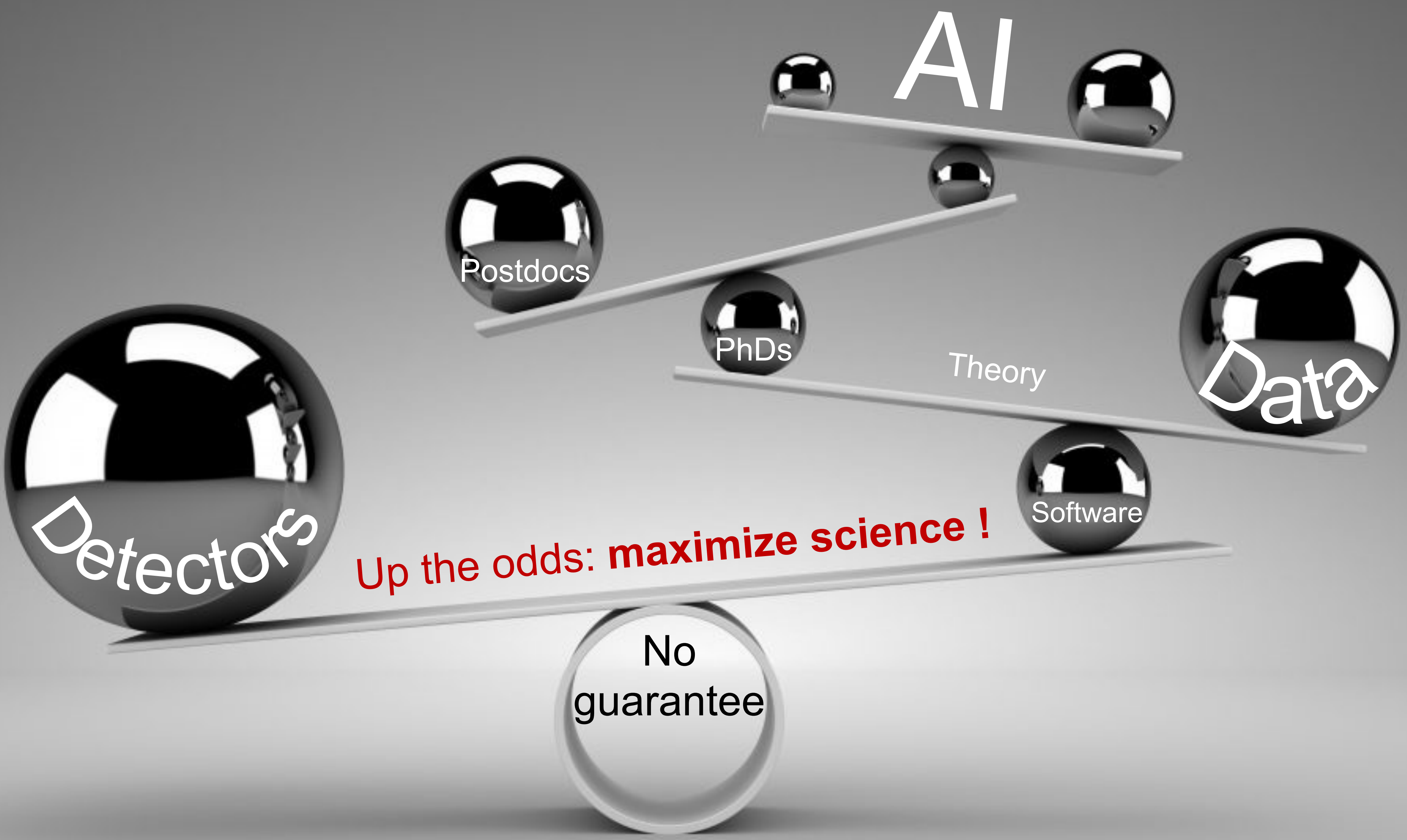
- Statistical algorithms to **model** data & perform **tasks** without explicit instructions
- Thrives on *big data*
- Generalizes to unseen examples

AI/ML in Science has taken off !

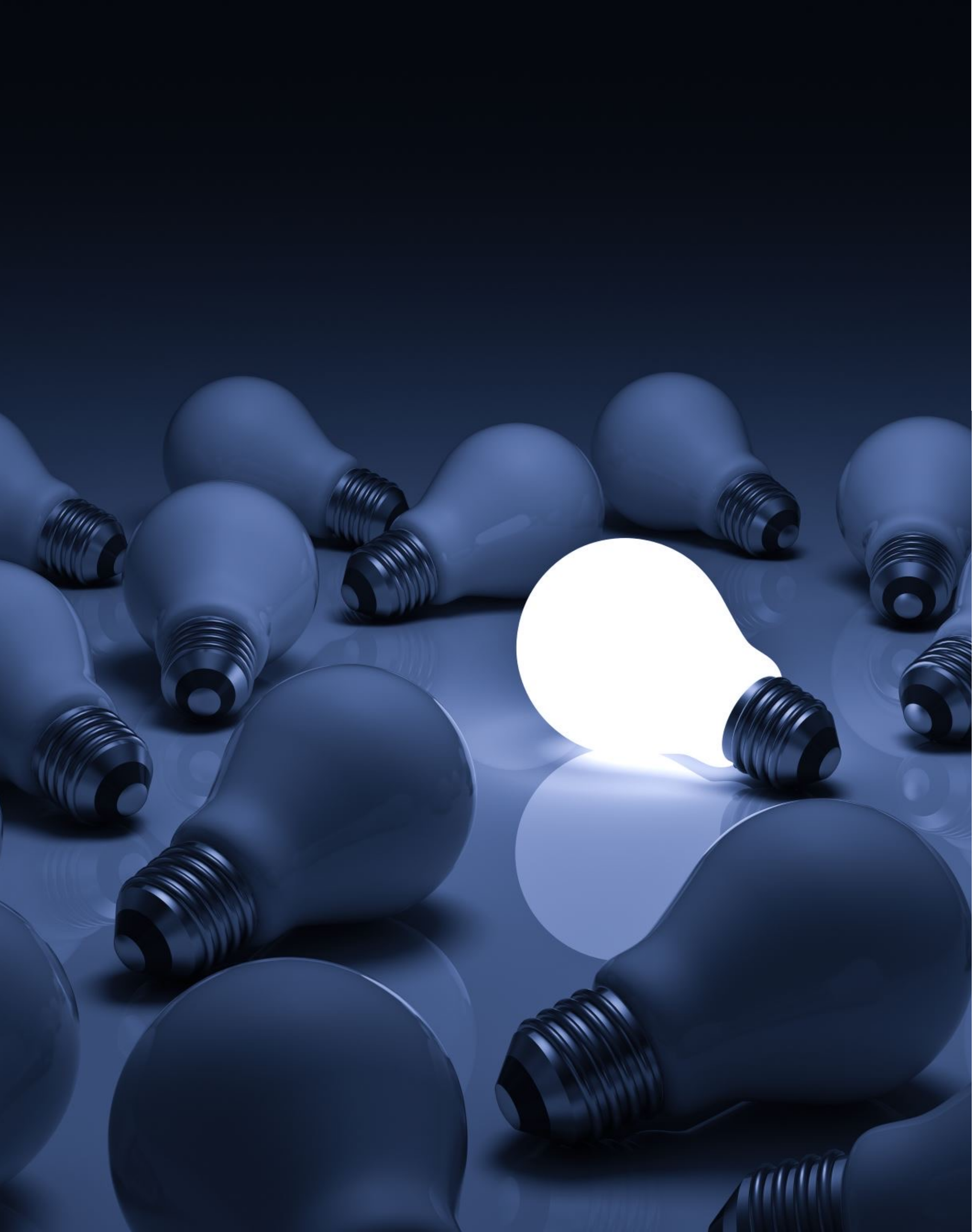




[Ben Blaiszik, "2021 AI/ML Publication Statistics and Charts". Zenodo, Sep. 07, 2022. doi: 10.5281/zenodo.7057437.]



Research is exploration



With great tools comes great responsibility

AI = opportunity

Think big !

WHAT DO YOU THINK?

Why are you a scientist?

Join at

slido.com

#8223 694

https://app.sli.do/event/bg2PknbqUbHxvRJEo34CYE

How do you choose what question to work on?

[don’t answer “because my supervisor told me so”]

Why ML?

[What are your expectations & hopes?]

Someone gives you 100 billion dollars
What science do you invest in?

Why become a scientist?

- Curiosity – learn about & comprehend nature
- You are good at it
- Solve puzzles
- Design tools [to maximize efficiency of science]
- Publish, teaching, outreach,…

What makes fundamental research interesting?

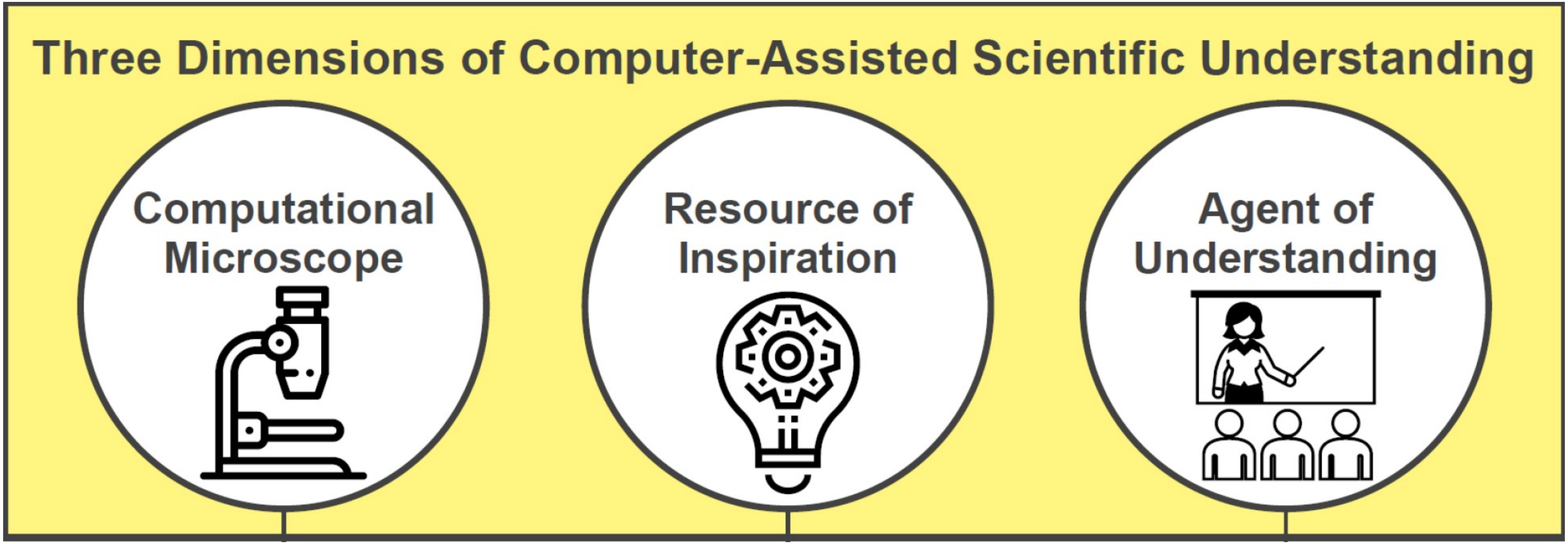
- It connects to **nature**
- You **can** make progress on it [knowledge gain gradient]
- It's **la mode** – someone else thinks it's interesting

[Matt Schwartz, EuCAIFCon 2024]

- Research metric?

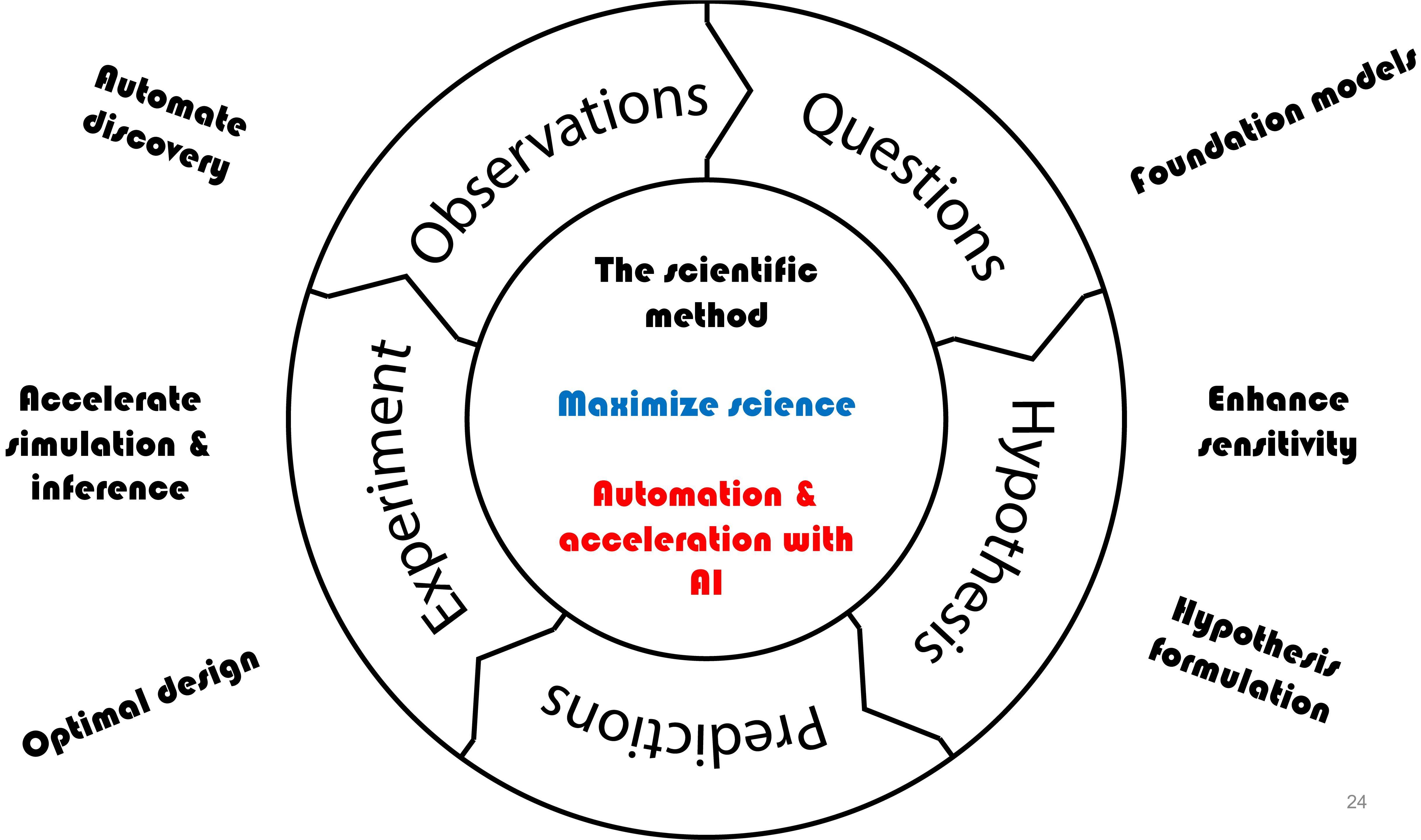
$$\frac{knowledge\ \ gain}{resources\ \ used}$$

What is scientific understanding?



[2204.01467]

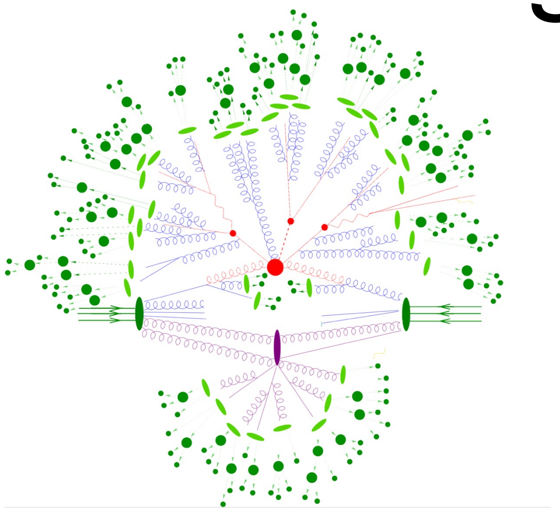
How to optimize the scientific method?


Automate discovery
Foundation models
Observations
Questions
The scientific method
Maximize science
Automation & acceleration with AI
Accelerate simulation & inference
Enhance sensitivity
Experiment
Hypothesis
Predictions
Optimal design
Hypothesis formulation

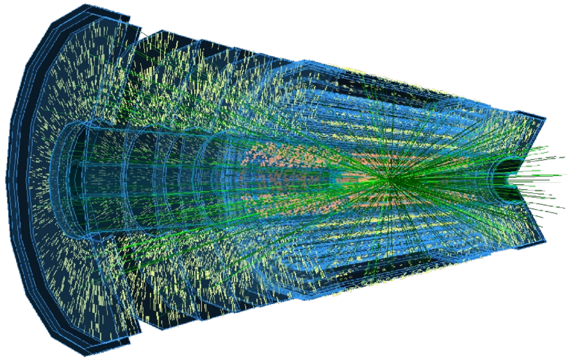
Back to HEP

WHAT IS HOLDING US BACK?

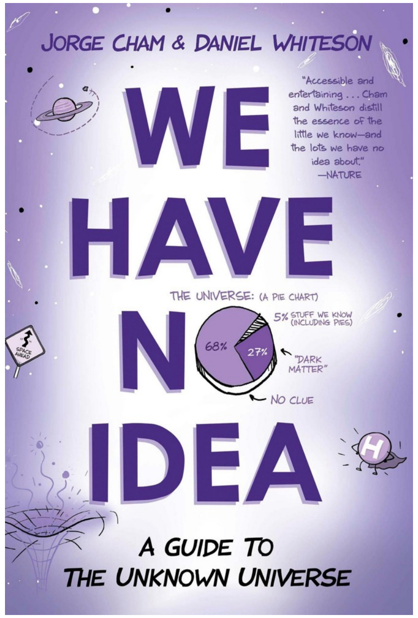
High energy physics challenges



Complex,
high-dimensional
& sparse data



Cannot calculate P(**data**|**theory**)
Slow simulation / limited accuracy



Search for the unknown

Hypothesis testing

Example Higgs boson discovery

H_0: no Higgs

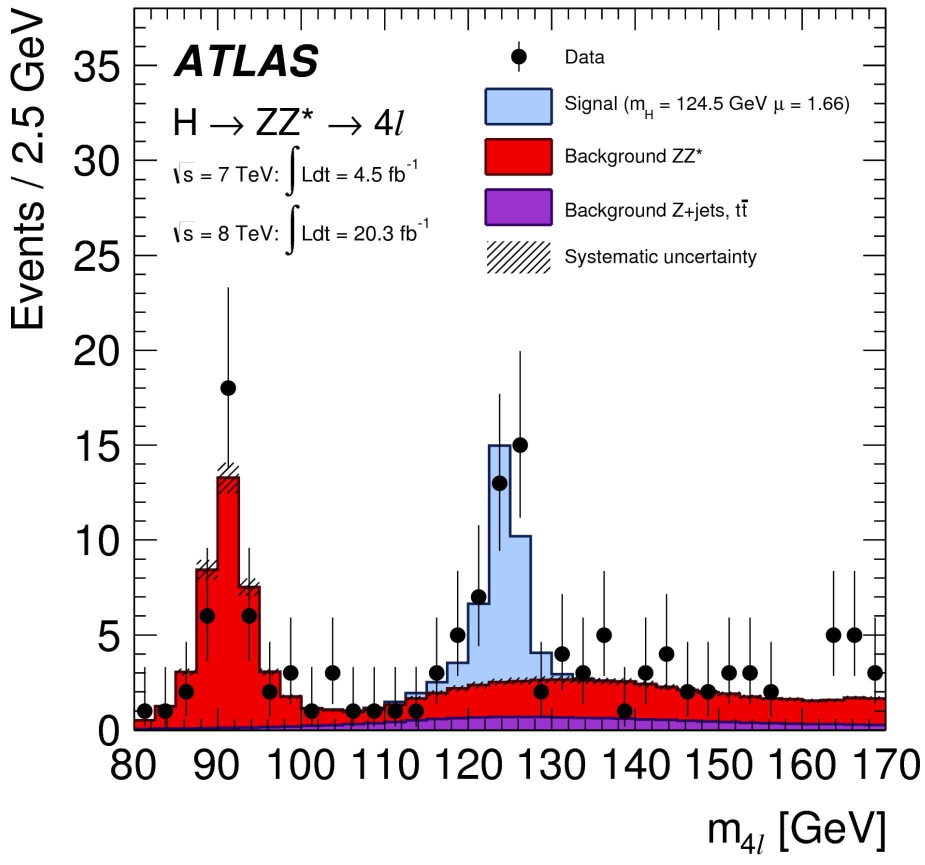
H_1: H_0+Higgs

Neyman-Pearson lemma:

best* statistics is the **likelihood ratio**

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

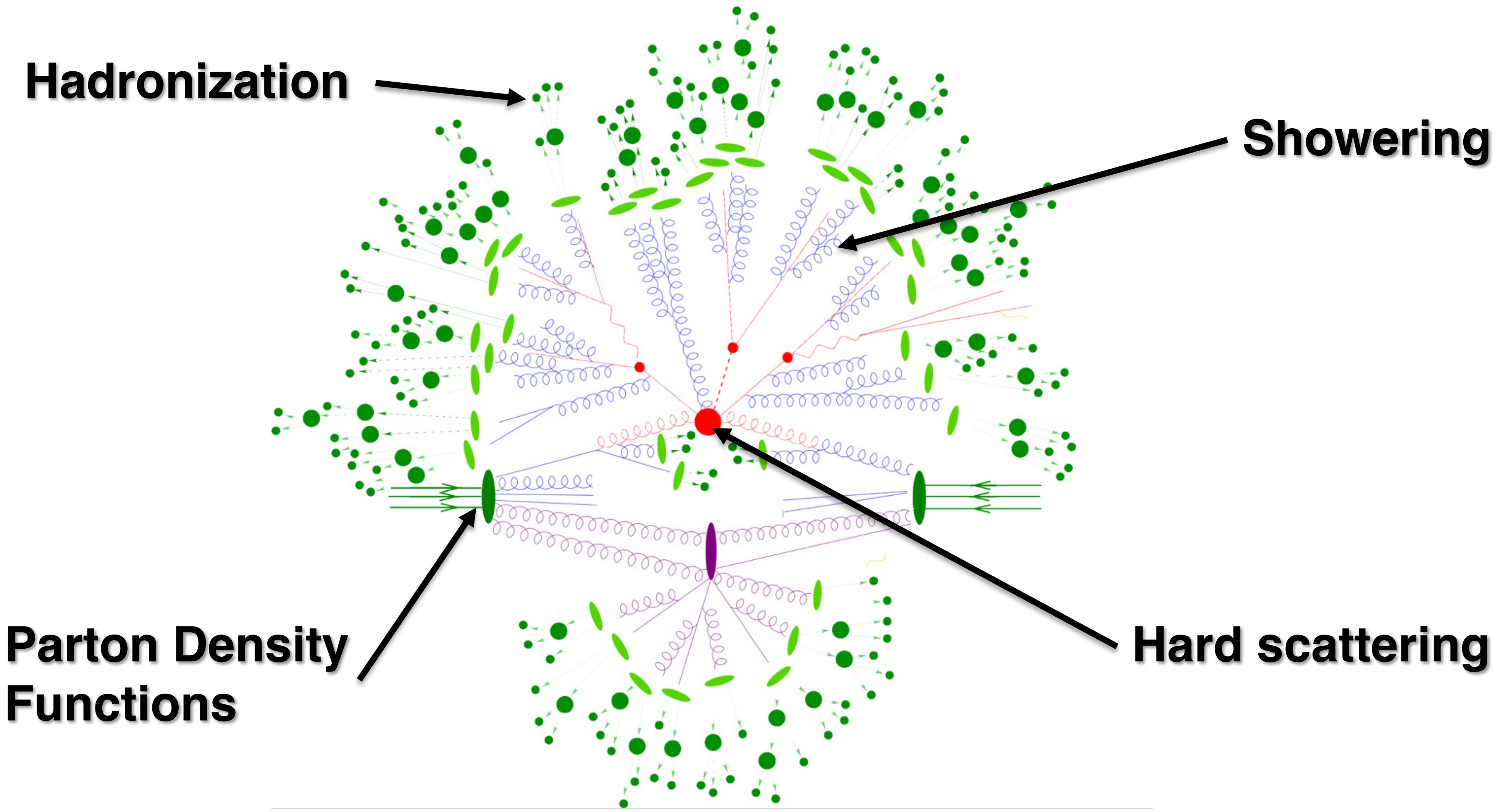
data **theory**



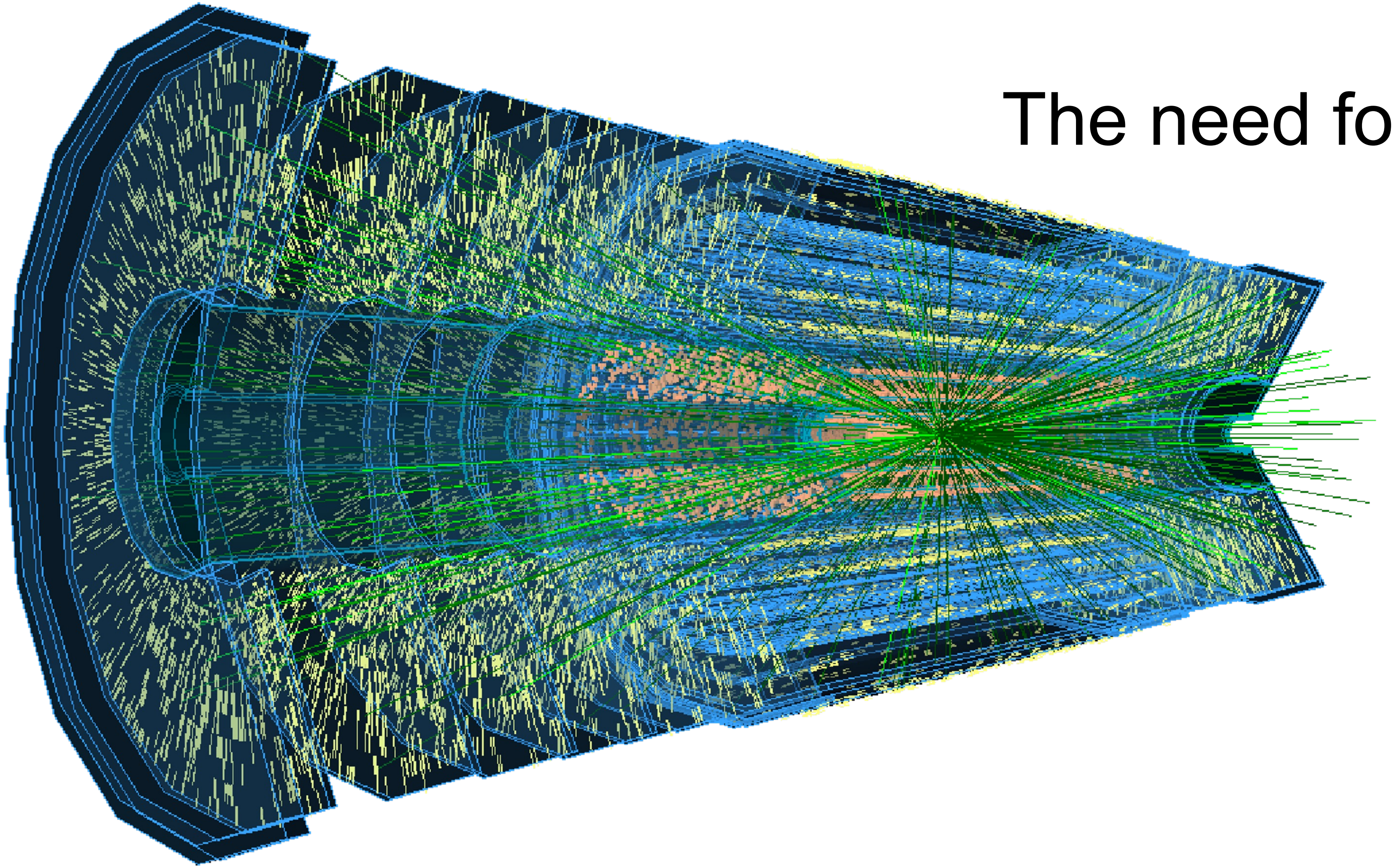
[Phys. Rev. D. 90 (2014) 052004]

*Gives smallest missed discovery rate for fixed false discovery rate

Cannot calculate P(**data**|**theory**)

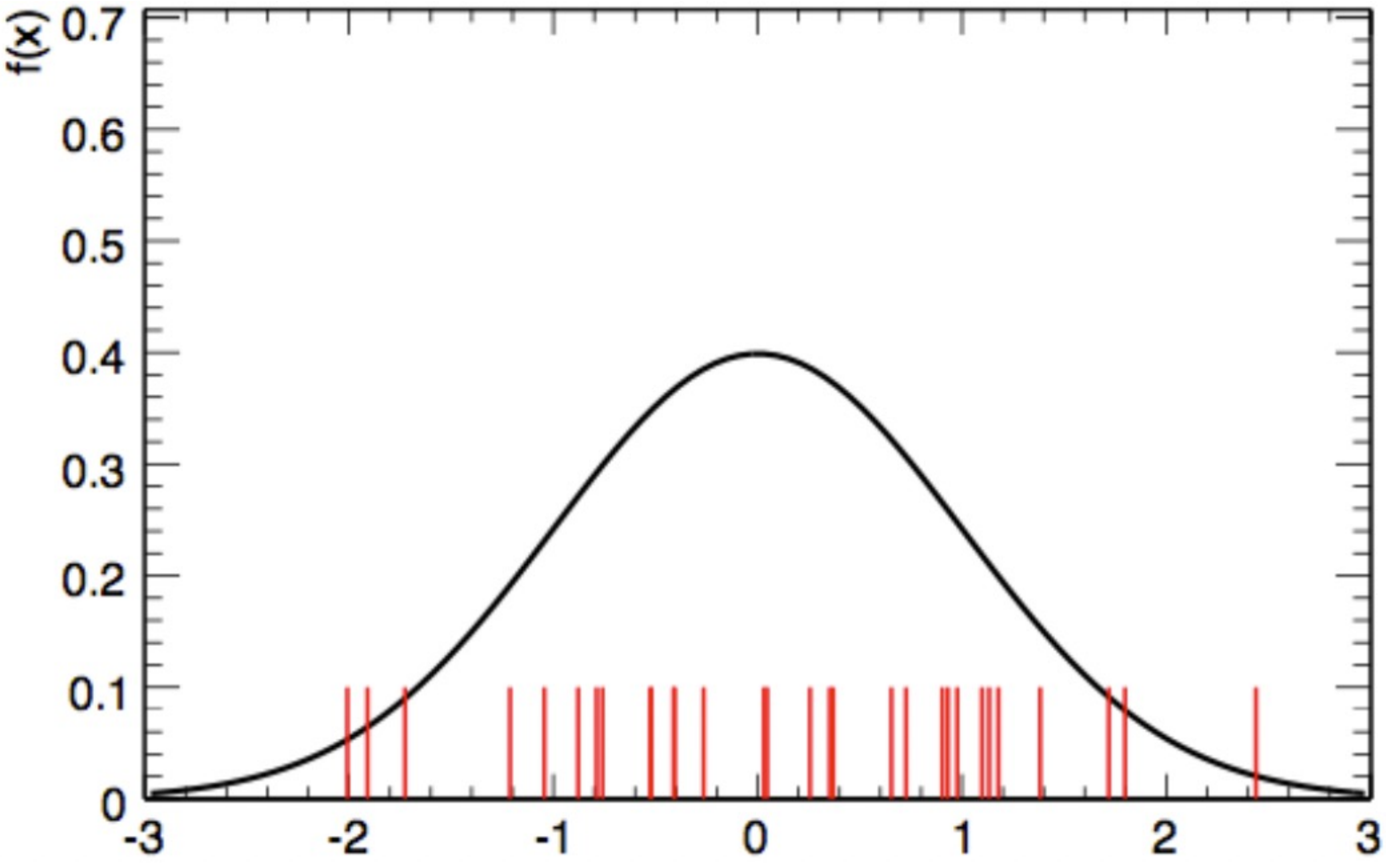


Can **simulate** P(**data**|**theory**)



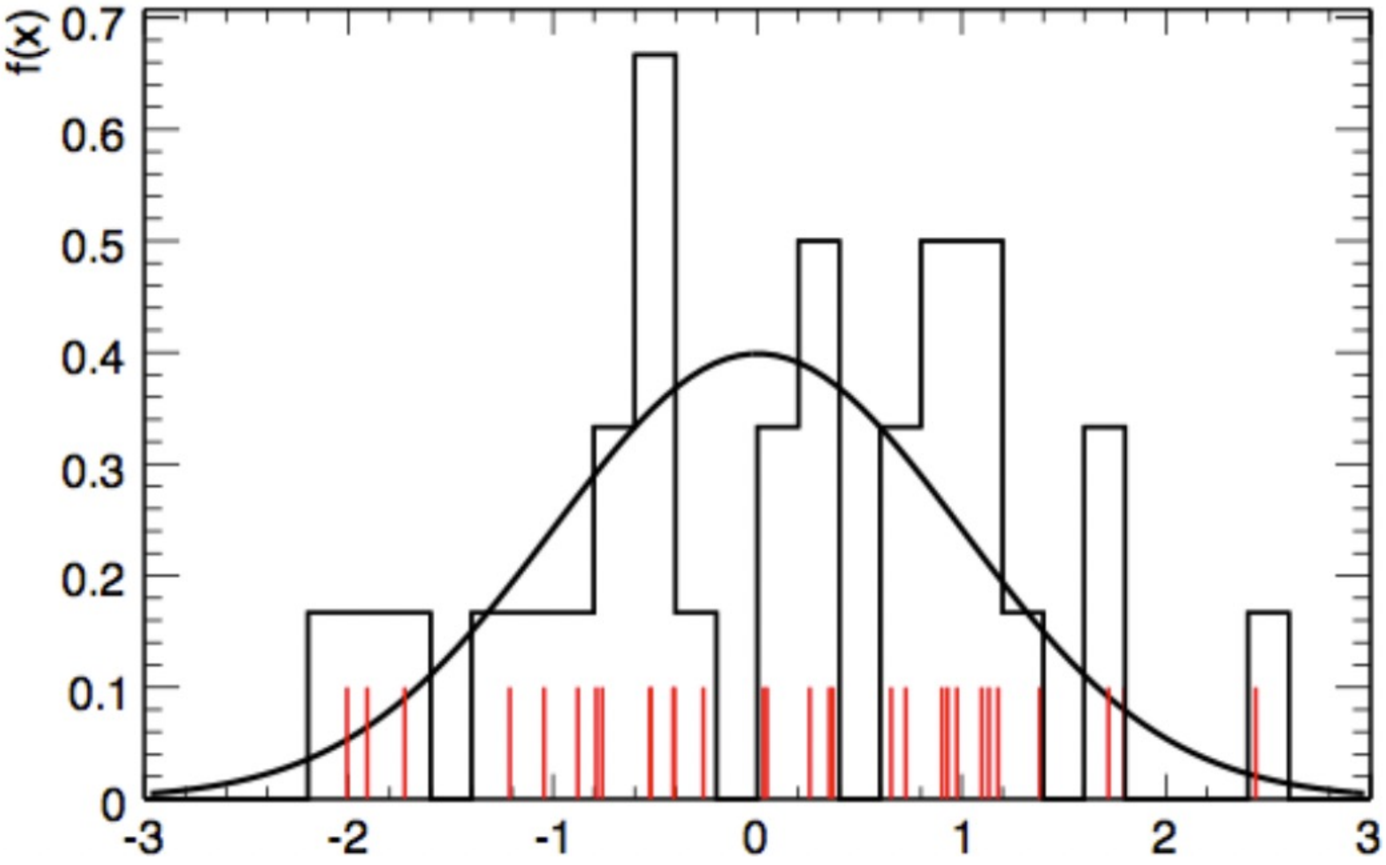
The need for synthetic data:
MC simulation

Tou example: draw events from 1D Gauss



[K. Cranmer]

Histogram ~ P(**data**|**theory**)



[K. Cranmer]

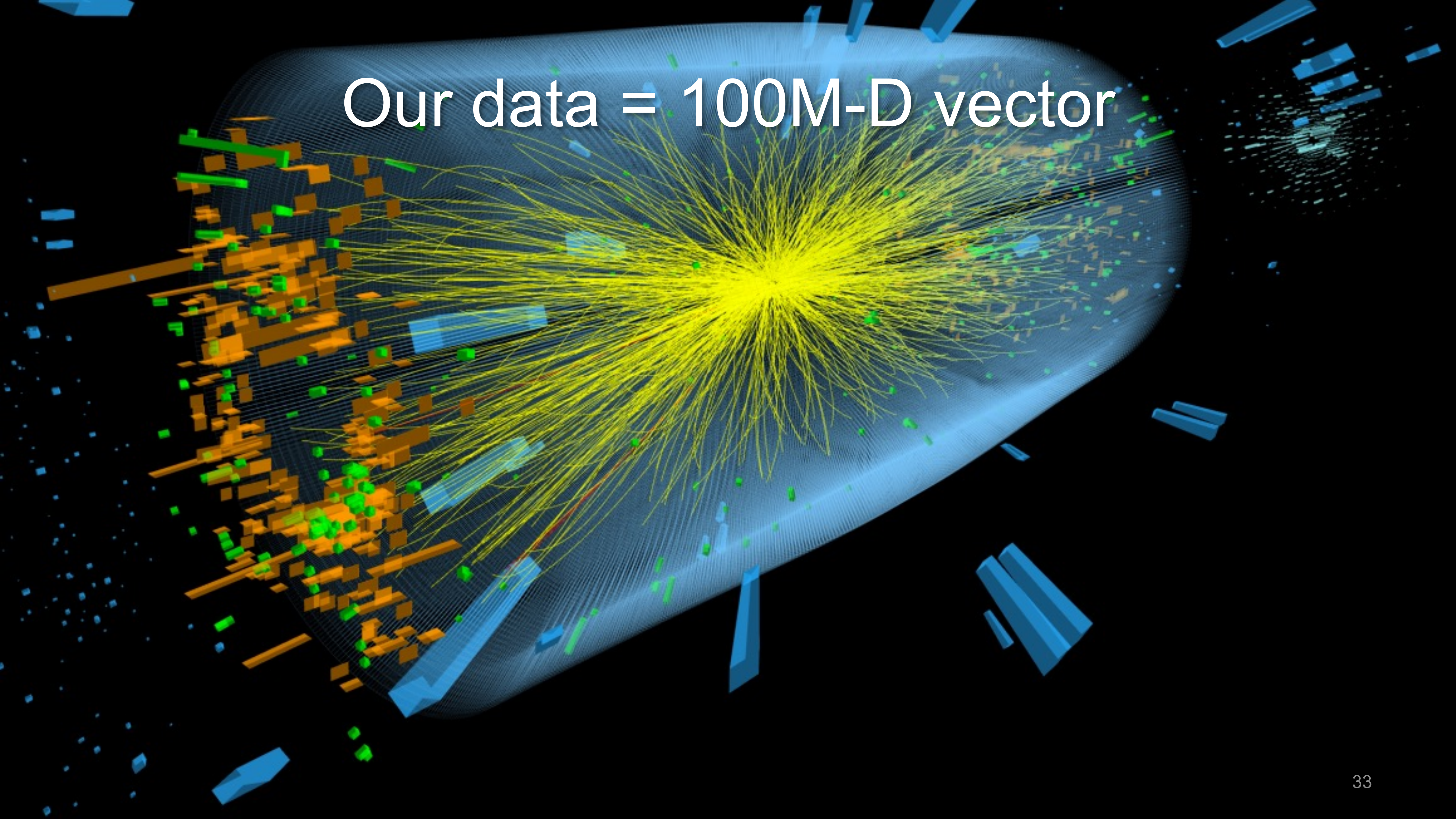
O(100) events needed to describe 1D distribution

And for an n-D distribution?

$$\mathbf{O(100^n)}$$

Curse of dimensionality

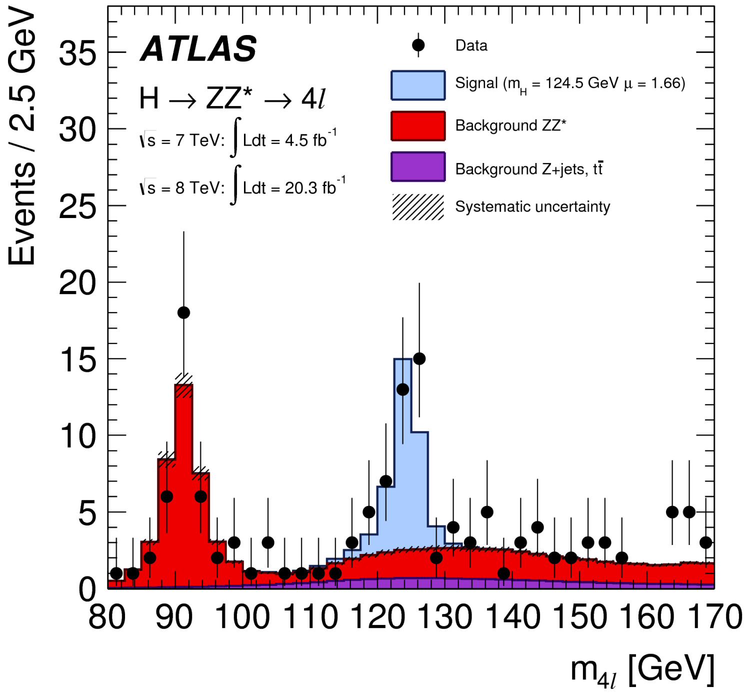
Our data = 100M-D vector



Sufficient test statistics?

Project to O(1) dimension

Meaningful representation



$p(x)$

No guarantee of optimality !

$1D < x < 100M\text{-}D$

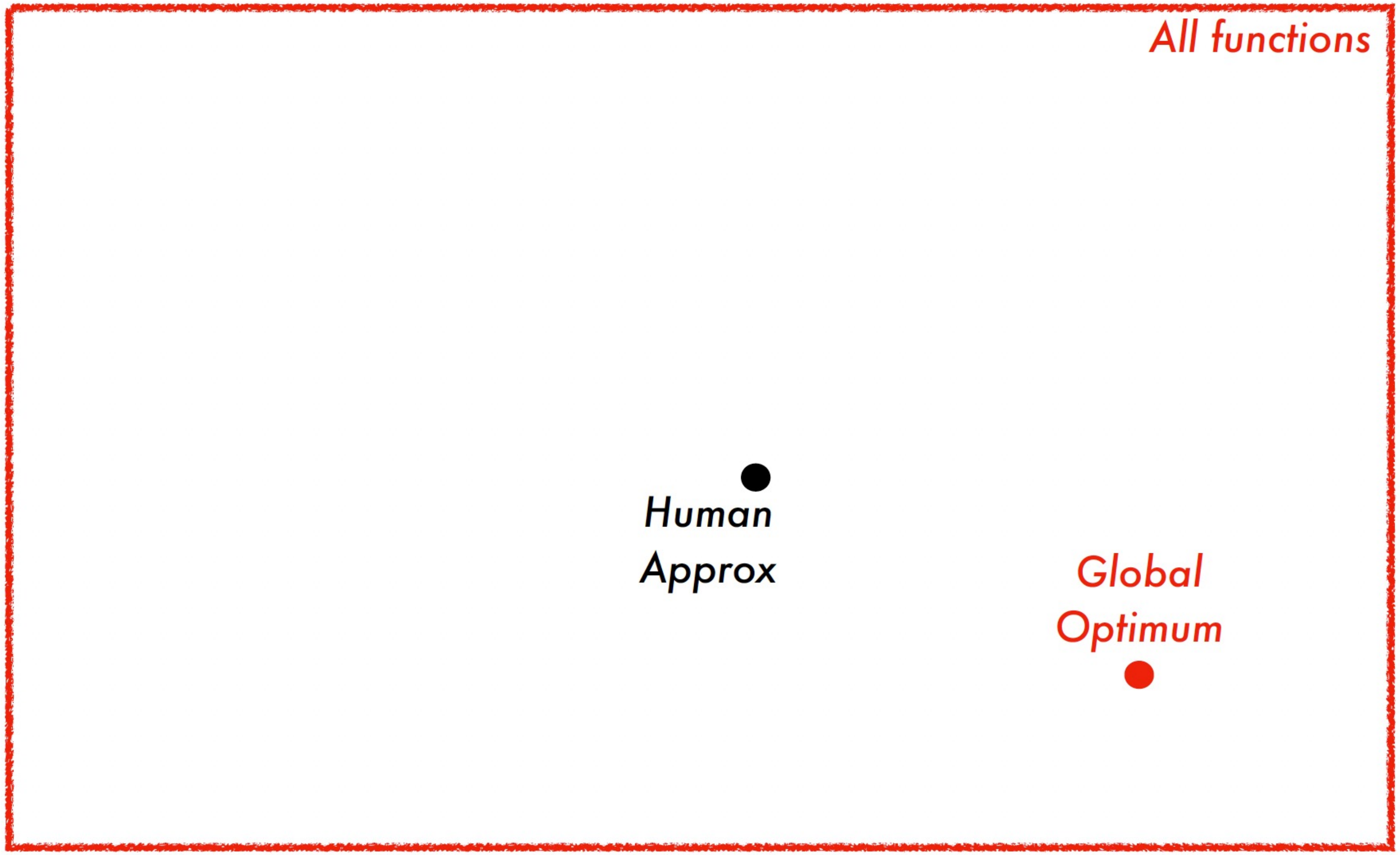
Control dimensionality with ML

ML = learning generic functions

Functional space

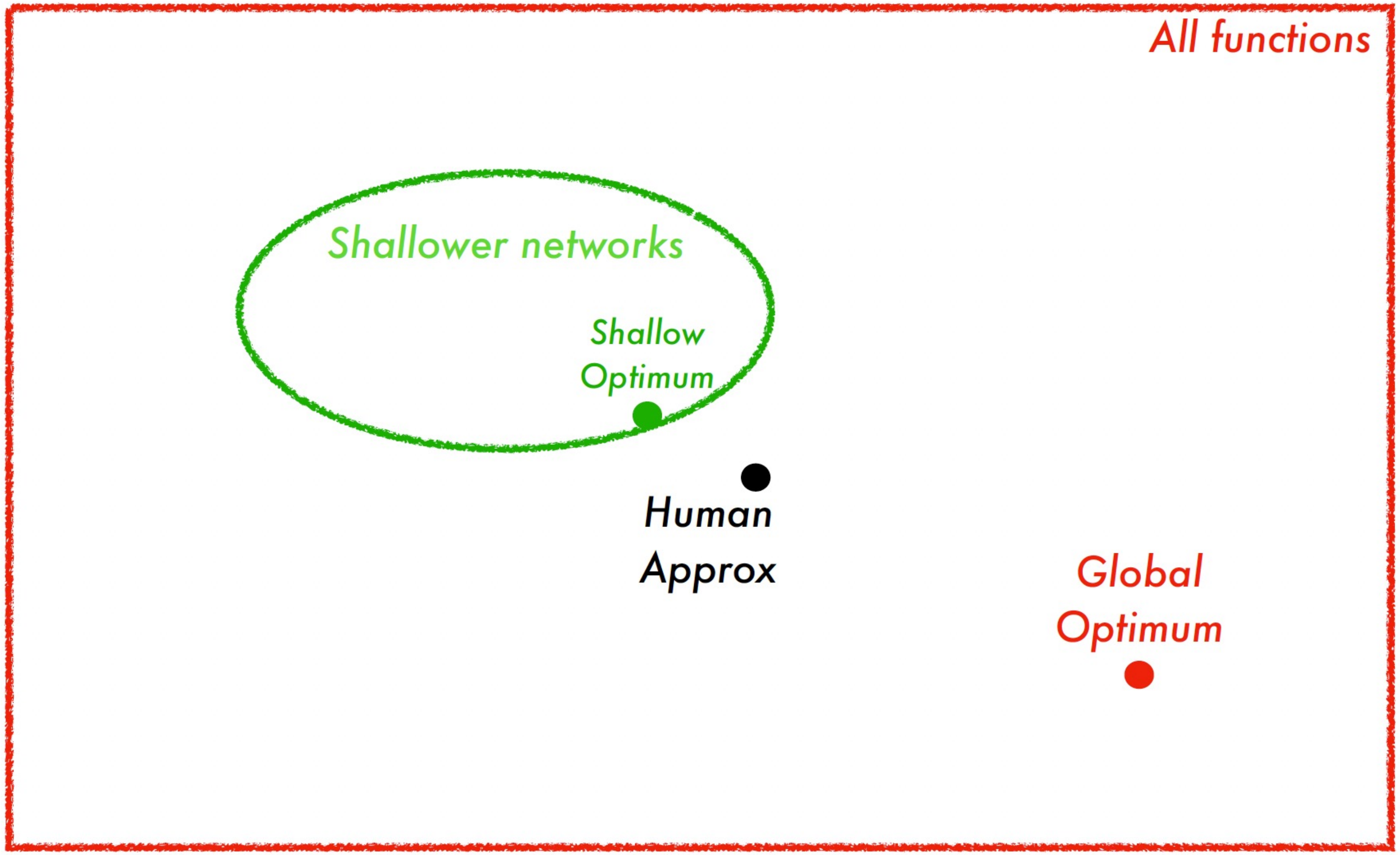
[D. Whiteson]

Human approximation



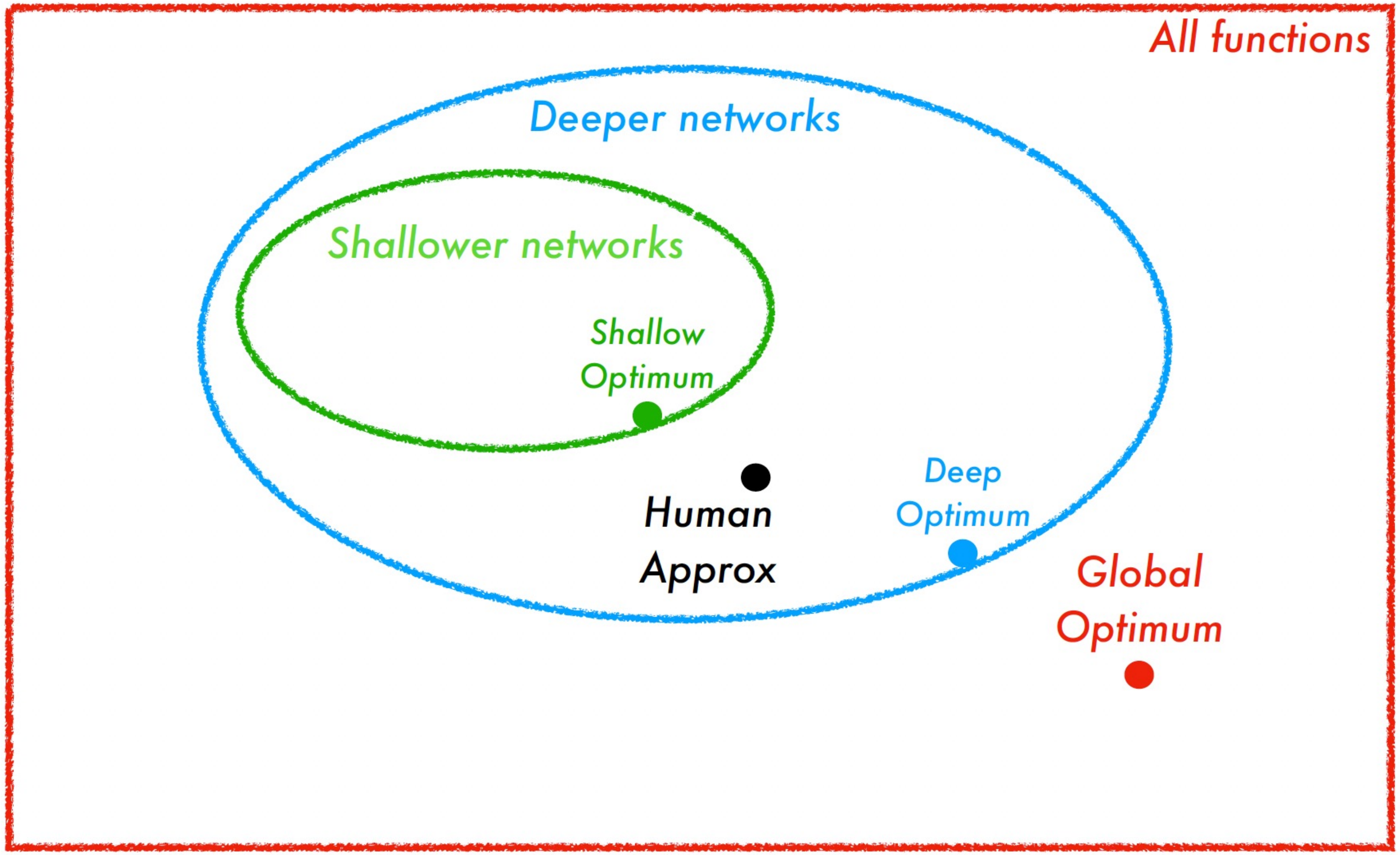
[D. Whiteson]

Shallow space



[D. Whiteson]

Deep space



[D. Whiteson]

Supervised vs unsupervised learning

Supervised

Data: (**x**, **y**)
x is data, **y** is label

Goal: learn mapping **x** → **y**

Examples:
classification, regression

Unsupervised

Data: **x**
x is data, no labels !

Goal: learn underlying (hidden) structure of data

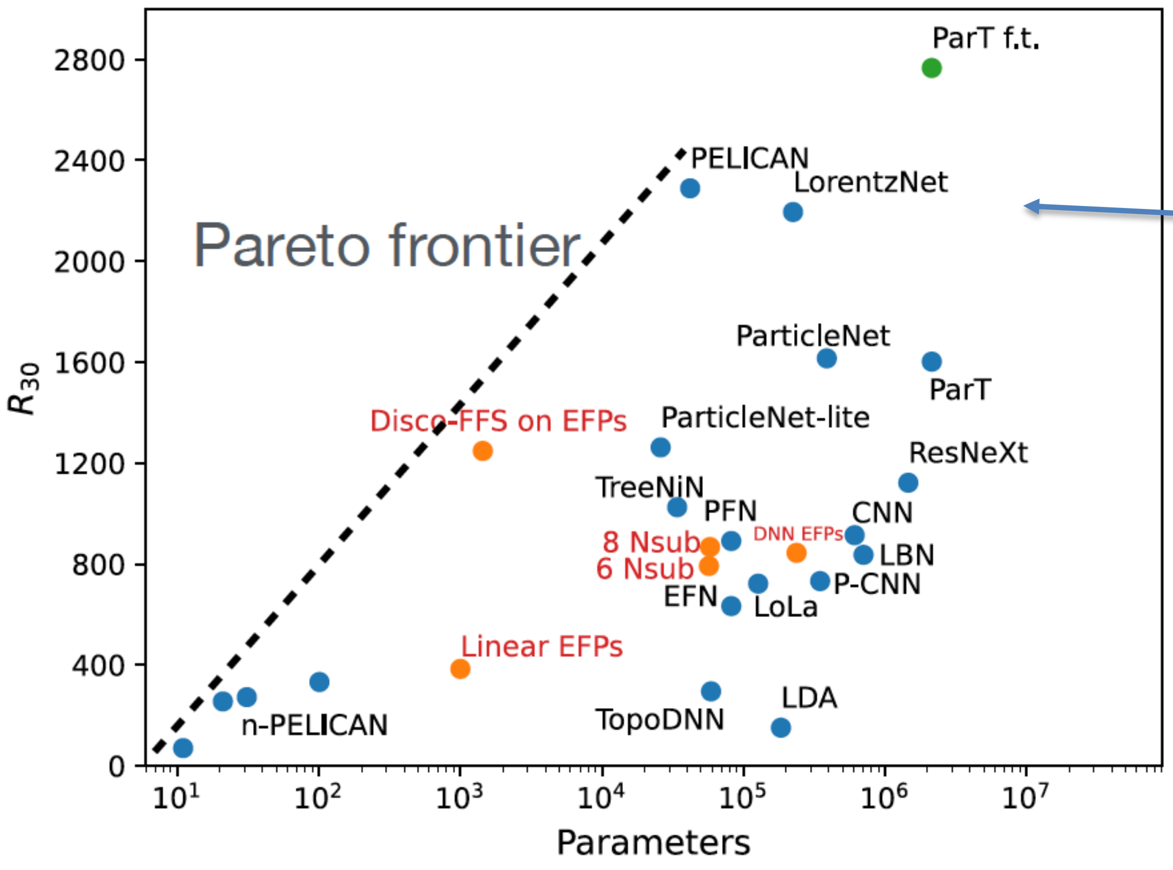
Examples:
Clustering, compression, generation

State of the art of supervised

Example: **classification**

The *frontier* of classification

Top tagger comparison: R_{30} = BG rejection for 30% efficiency vs. #parameters



[G. Kasieczka, EuCAIF 2024]

Transformers rule the world

Inductive bias (Lorentz invariance, symmetries,…):

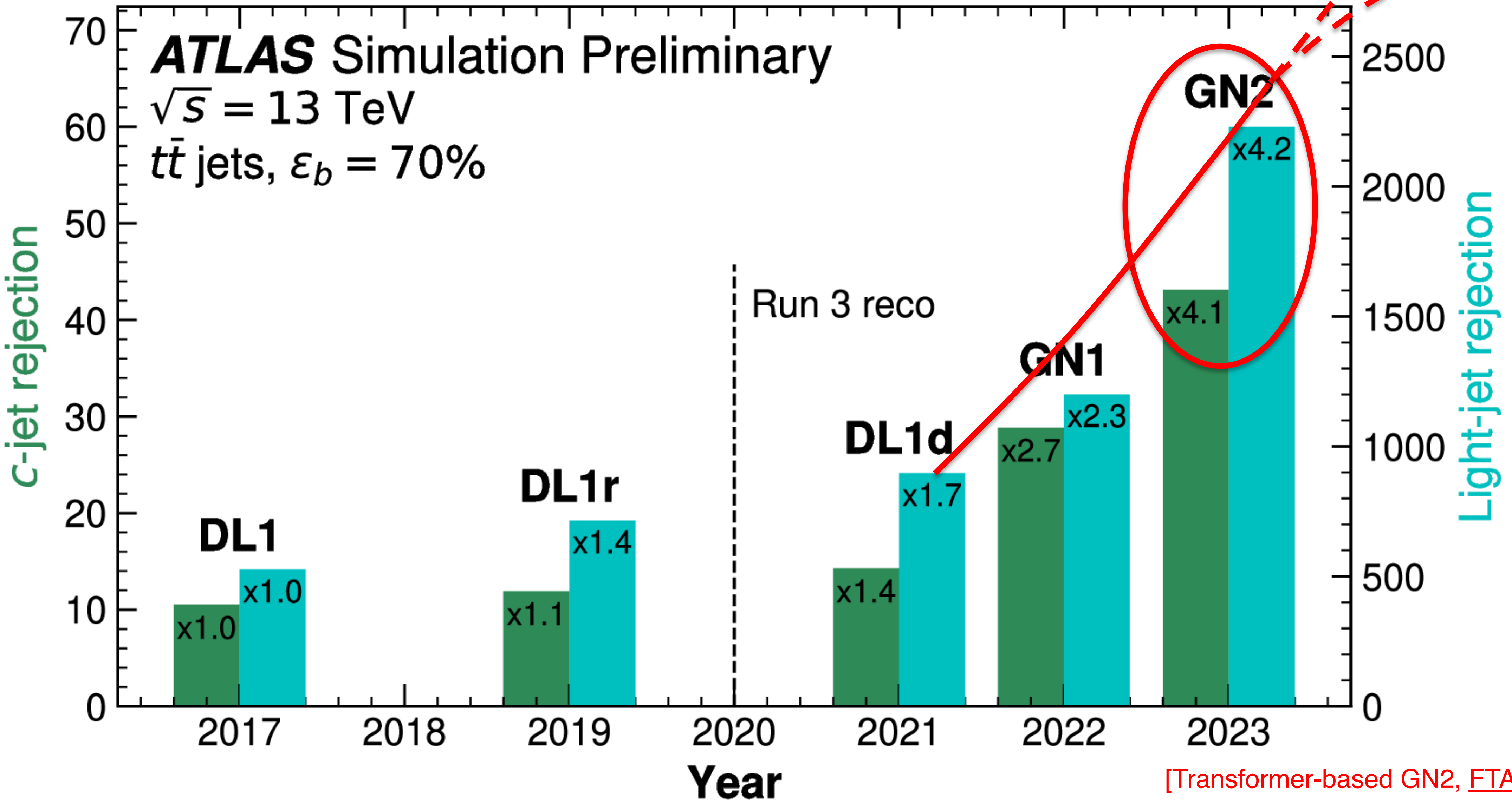
- More parameter efficient
- **BUT** less performant

The bitter lesson vs. heroic domain-specific modeling efforts

Exact symmetries in latent space – hard to learn
Only approximate symmetries in data space

Enhance sensitivity

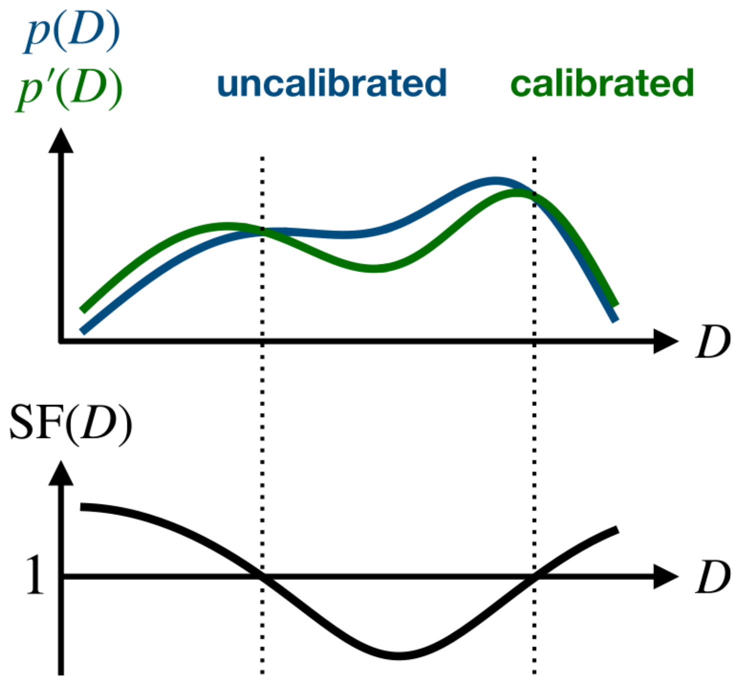
Plateau ?



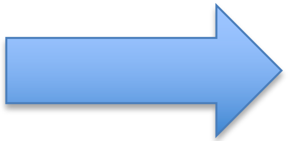
[Transformer-based GN2, FTAG-2023-01]

Domain adaptation: calibrate synthetic to real data

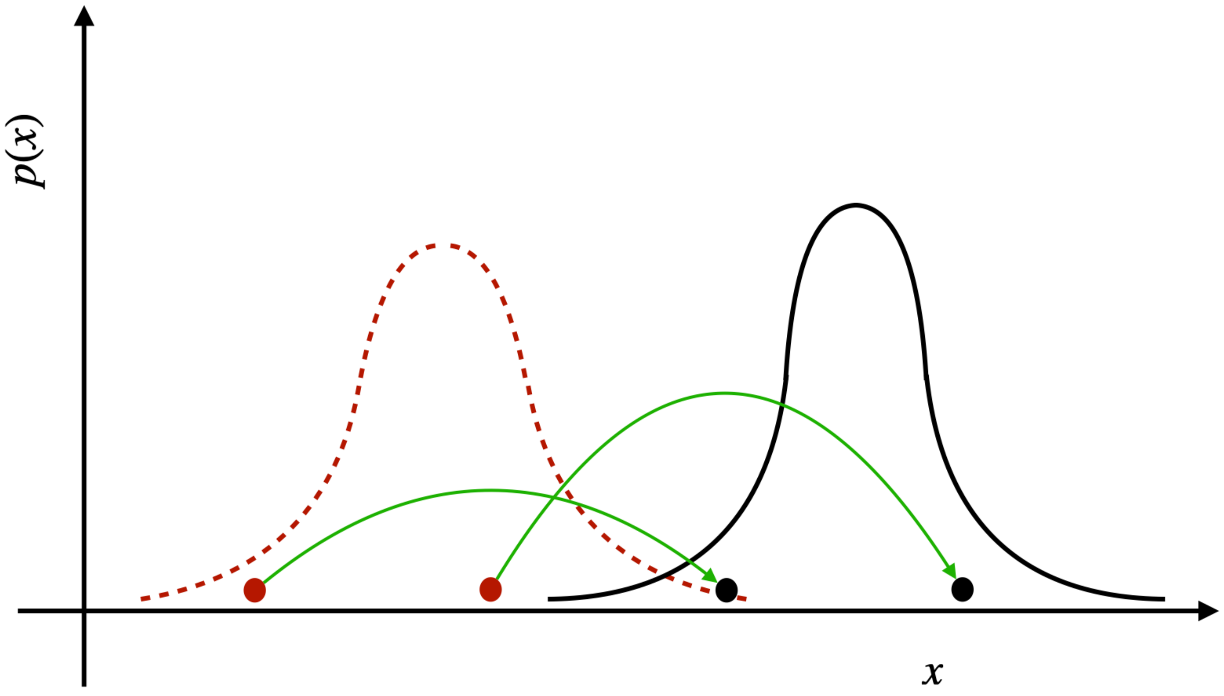
1. Reweighting with ratios ["scale factors"]
- Non-overlapping support
- Battle curse of dimensionality*



(b)



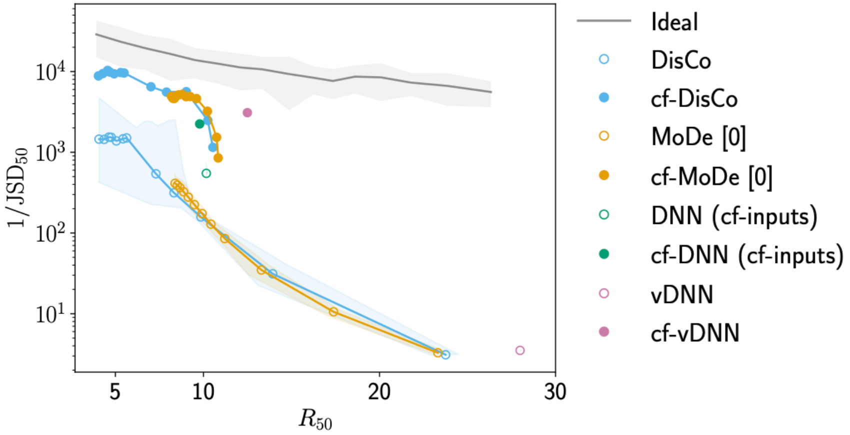
2. "Transport your problems away"



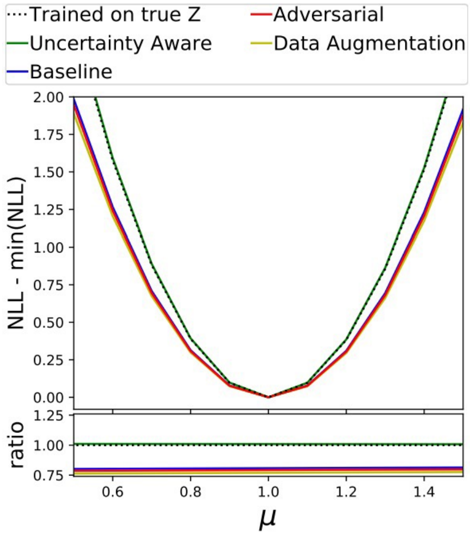
[*e.g. using classifiers [1506.02169](1506.02169)]

[[2107.08648](2107.08648)]

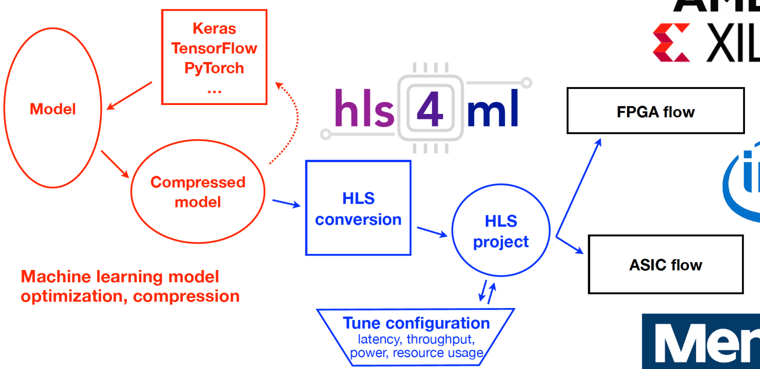
Many more challenges



Decorrelation
[*Ethical AI* in Science]
[e.g. 2211.02486]
More in a few slides!



Making scientific decisions in the presence of uncertainties
[e.g. 2105.08742]



Offline → online
[On-the-edge, 1804.06913, hls4ml]

INTERPRETABILITY

Which face is real?



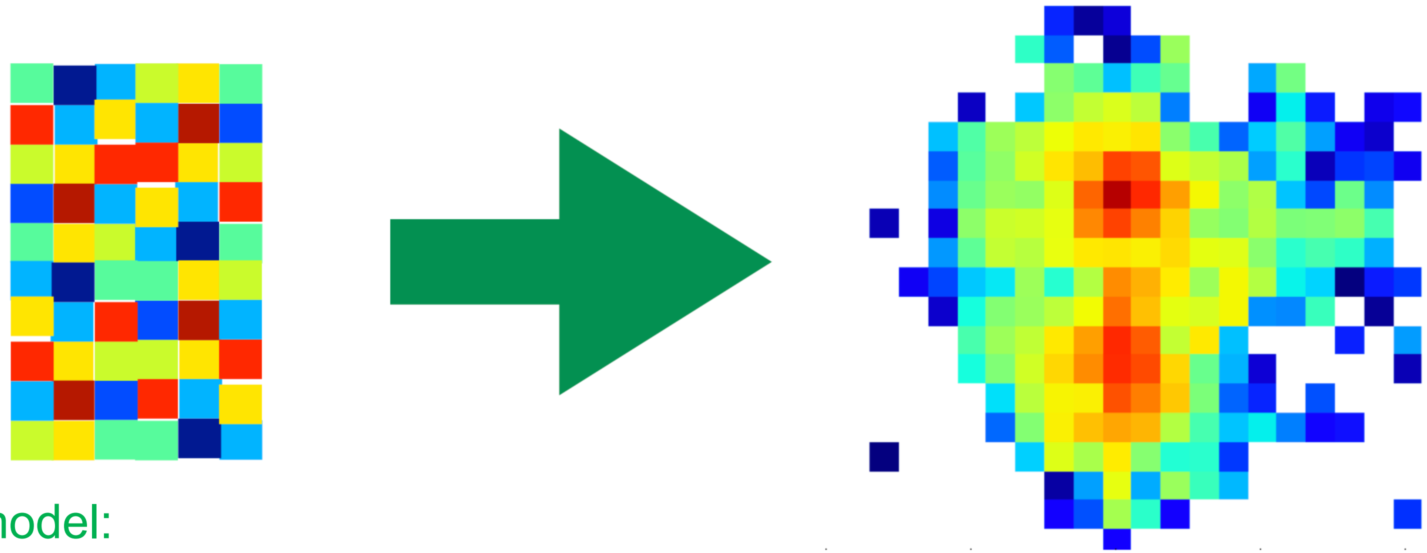
https://thispersondoesnotexist.com

Why generative models?

- Density estimation & outlier detection
- Data compression
- Mapping from one domain to another
  - Language translation, text-to-speech,…
- Representation learning
- Understanding the data
- …

The ML toolbox: generative models

Fast surrogate model* which maps random numbers to structure



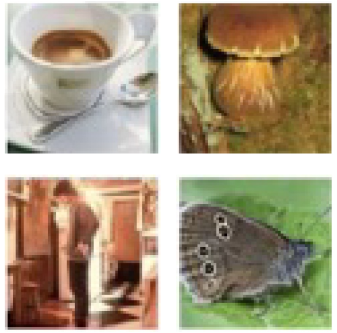
$$p_{\text{model}} \approx p_{\text{data}}$$

*Deep generative NN model:

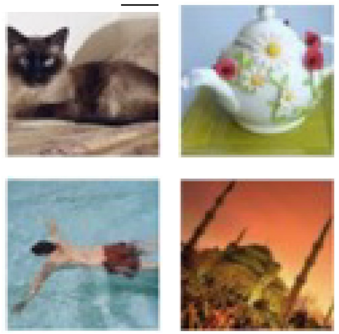
- Variational Autoencoders (VAEs)
- Generative Adversarial Network (GANs)
- Normalizing Flows (NFs)
- Diffusion models

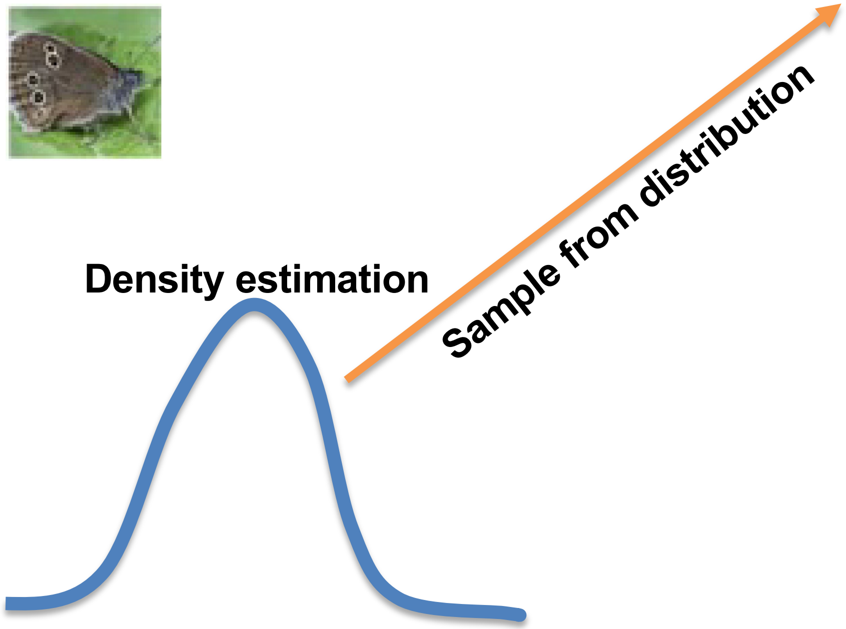
Example: image generation

Train on $p_{data}(x)$



Generate new samples $p_{model}(x)$





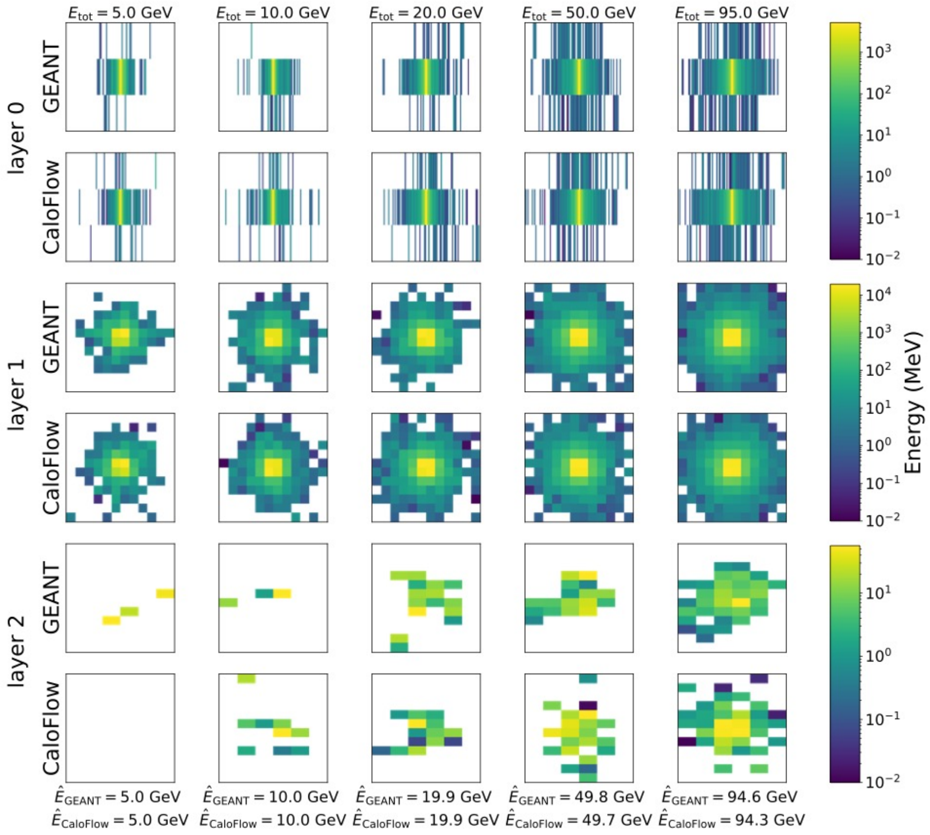
$$p_{\mathrm{model}} \approx p_{\mathrm{data}}$$

Faces

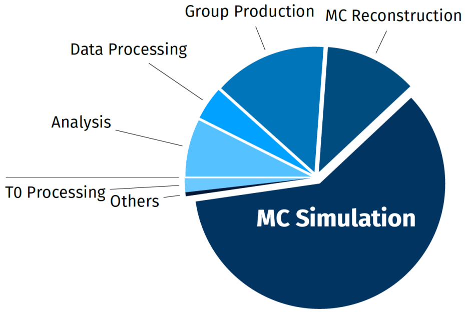
Detector images

[Karras et al., 2018]

$\longleftrightarrow$



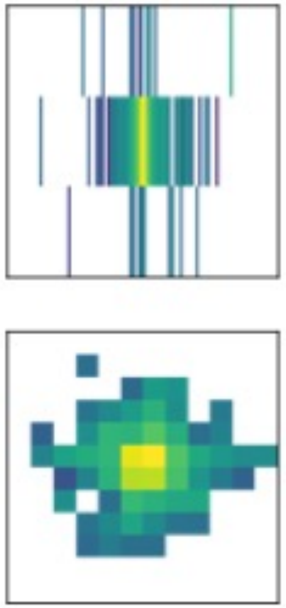
[2210.06204 ,CaloFlow,…]



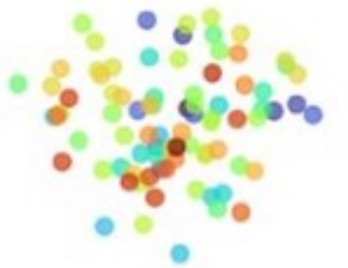
<u>Considerations:</u>

- Fast
- High fidelity
- Sample rare events [tails]
- Conditional sampling [e.g. p(model | prompt)]

Images → Point cloud



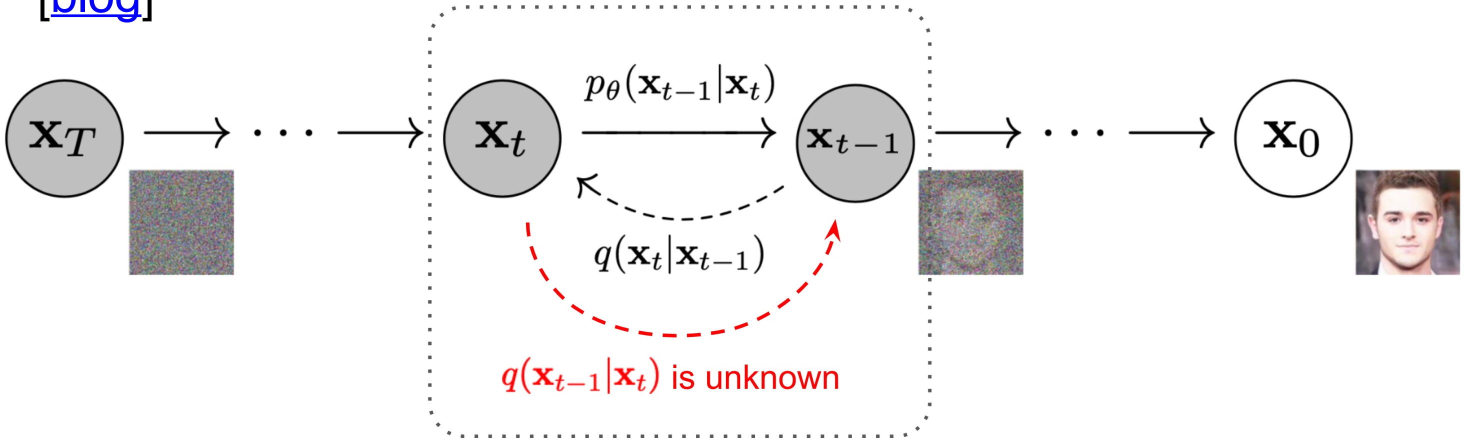
Decouple modeling
from detector geometry



- Addresses sparsity issue
- Promotes portable solutions

Point cloud diffusion

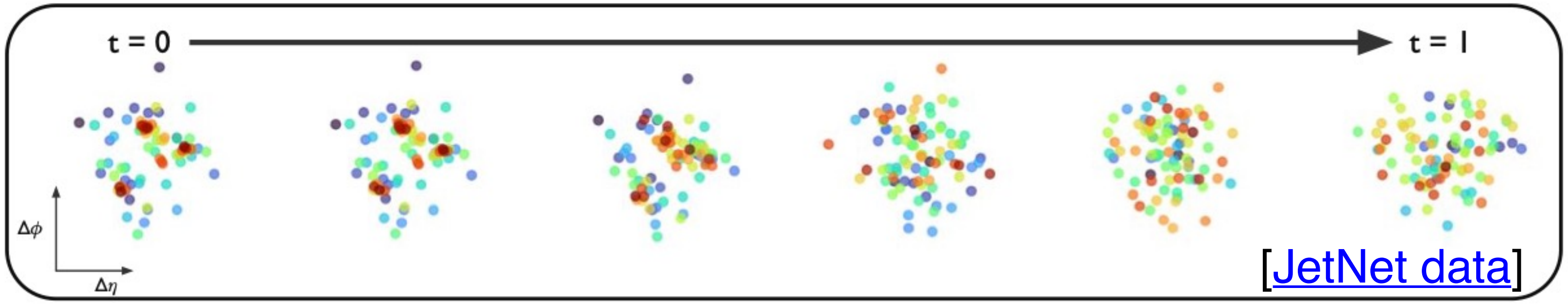
[blog]



Gradually add Gaussian noise (right-to-left=forward)

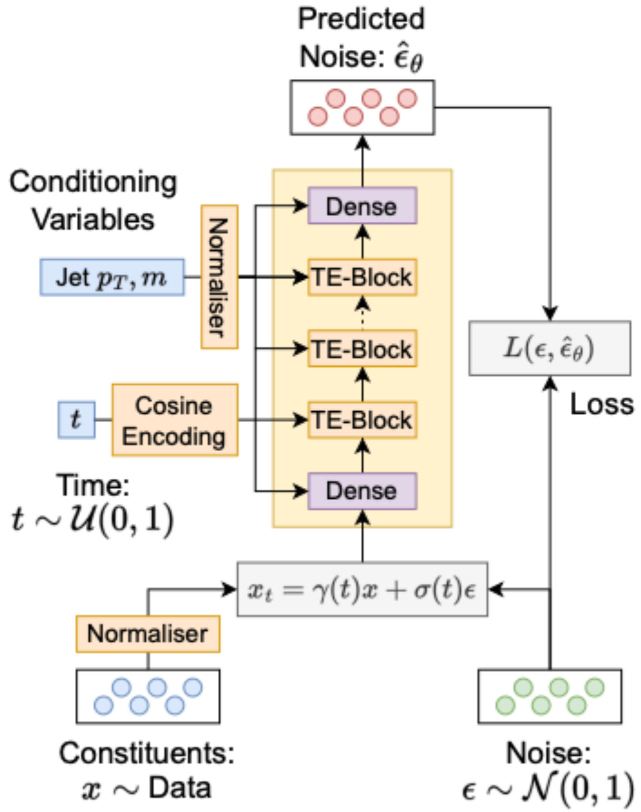
Reverse "learn the noise"

1000 → 100 → ~few steps (over last ~year)



[JetNet data]

[PC-JeDi]



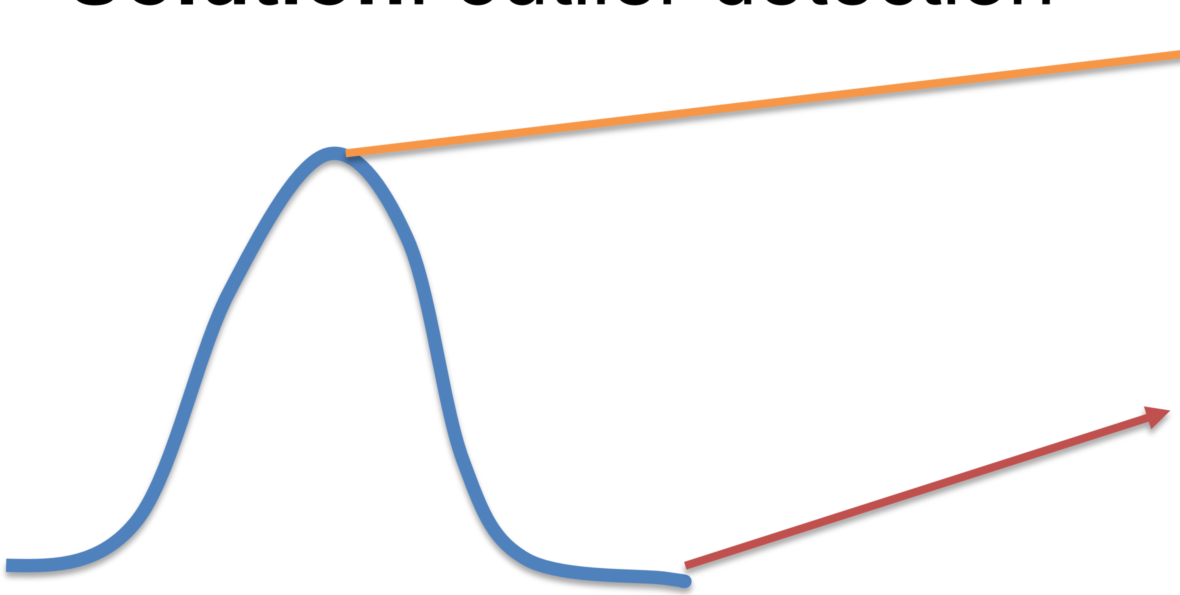
Transformer Encoder (TE) Block

[See also 2206.11898,…]

Outlier detection

Task: detect new or rare events

Solution: outlier detection



Example: **autonomous driving**

Normal data: sunny, highway, straight road

Outliers:

Edge cases

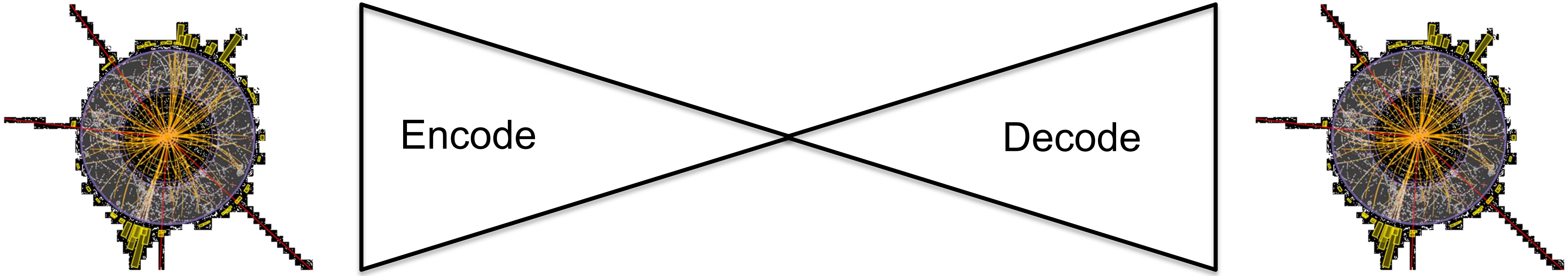
Harsh weather

Pedestrians

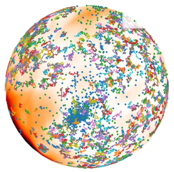
Outlier detection with autoencoders – does it work?

Not ready for prime time!

Train on *normal* (=SM)



Poor reconstruction = *anomaly*



[NAE]

<u>Challenges</u>:
- Outlier in high-dimensional space
- Performance ~ leading feature separation
- How to add physics priors without becoming supervised

Jet level [1808.08979, 1808.08992, 2007.01850, 2301.04660…]
Event level [1806.02350, 2105.14027…]

Debiasing

Uncover **underlying feature space** in data

Possibility to correct for existing biases in the data

vs

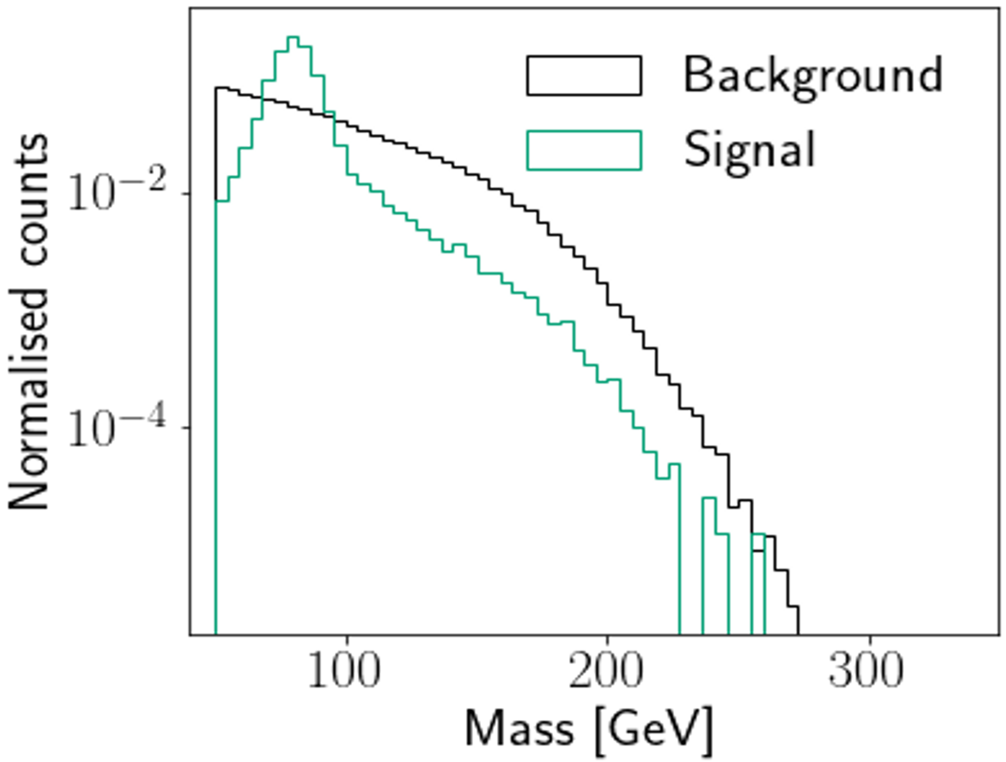
Homogeneous skin color & pose

Diverse skin color, pose, illumination,…

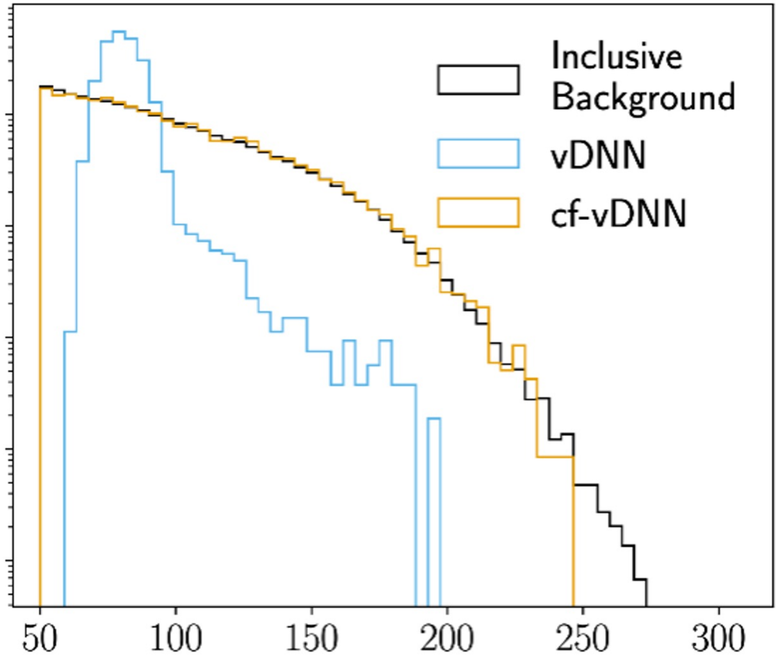
⇒ **Fairness, decorrelation**

Issue: background sculpting for bump hunting

Signal



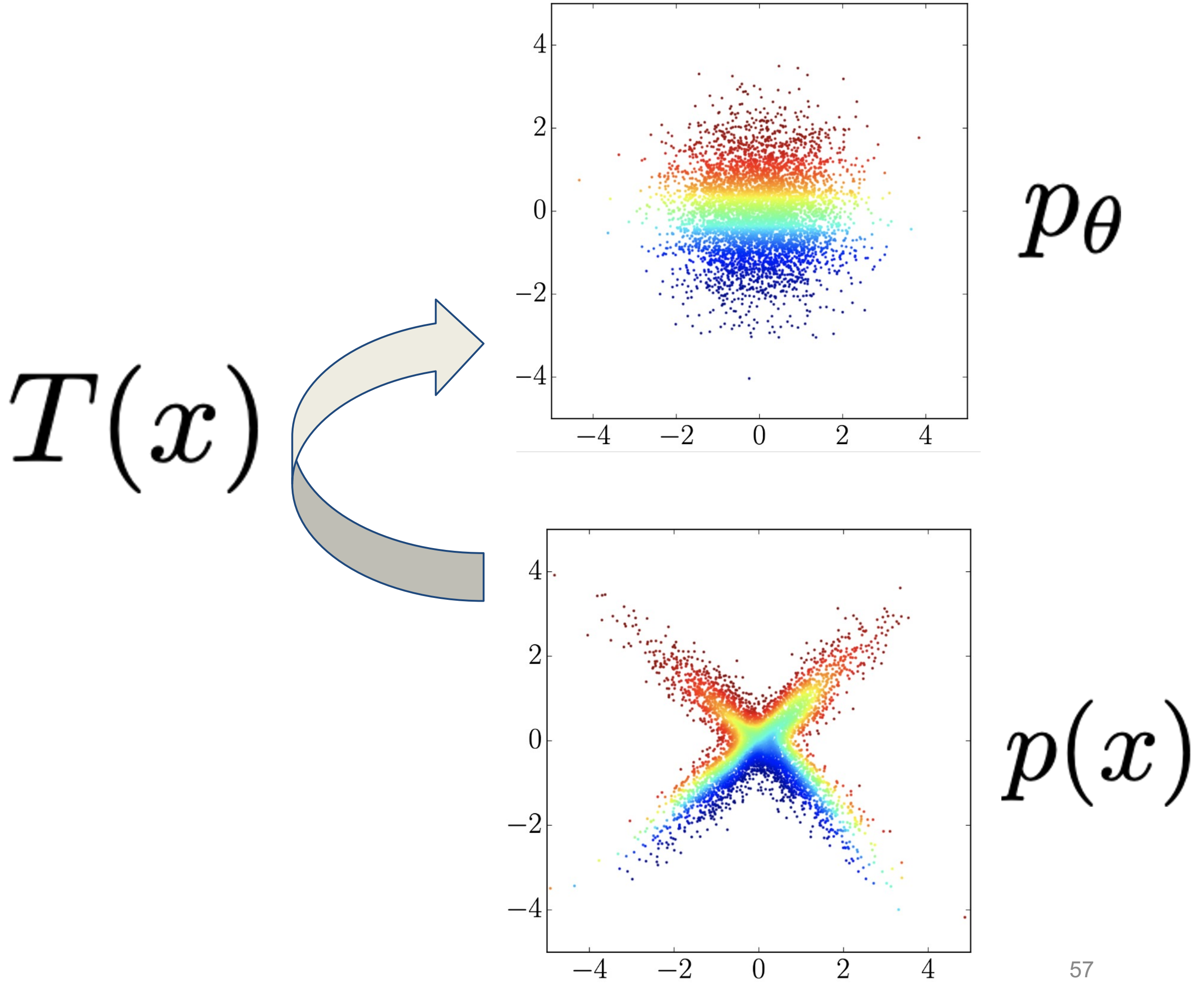
Background after cut on classifier



Goal: decorrelate background from mass

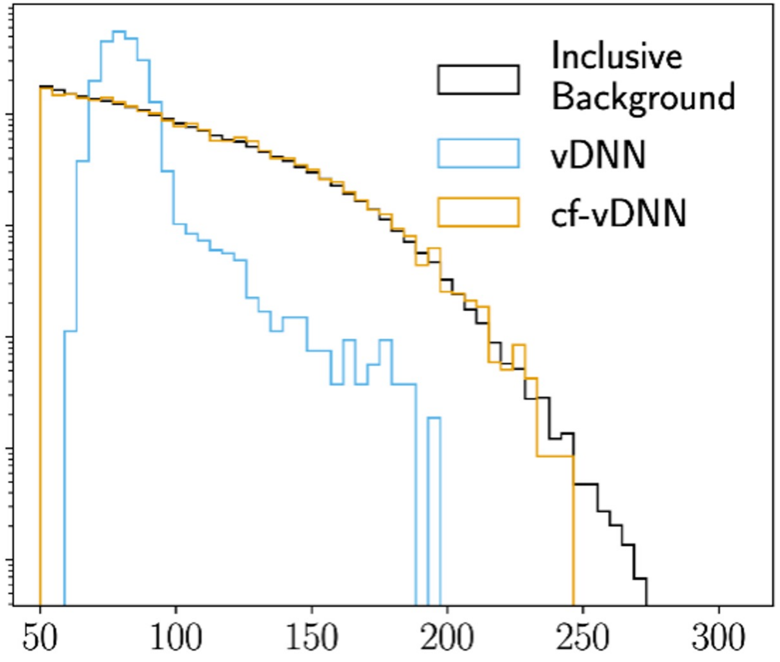
Decorrelation with normalizing flows

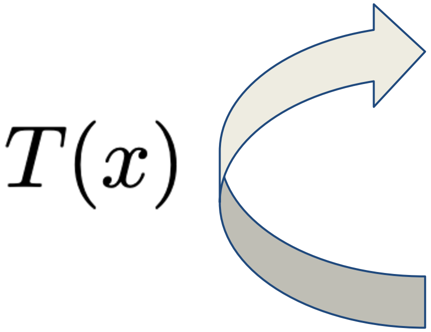
- A flow is a map between distributions
- It is invertible: no change in separation power !
- Can be made conditional

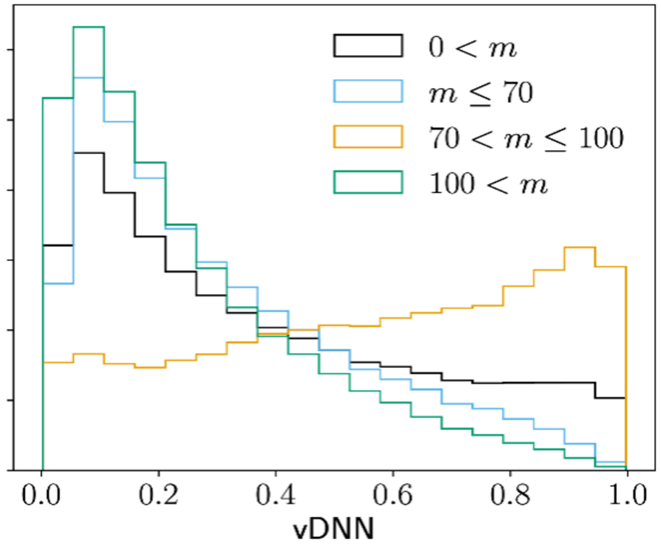


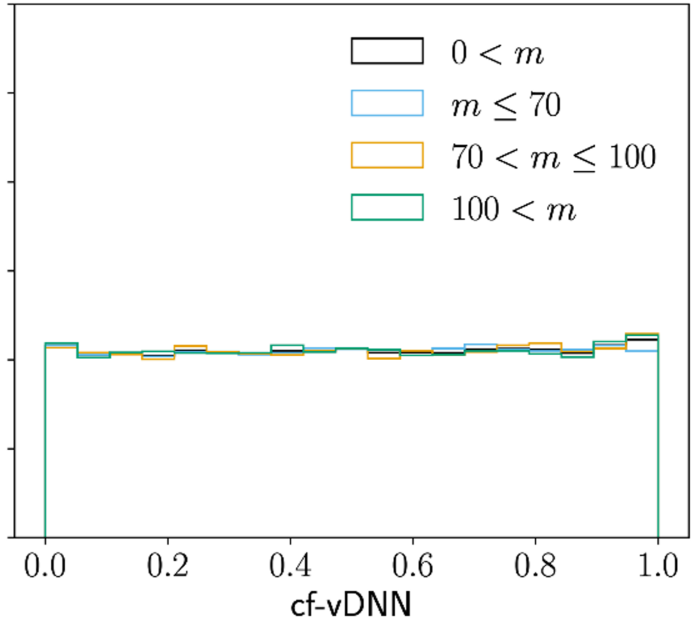
Train a flow to learn p(vDNN | m)

Decorrelated background



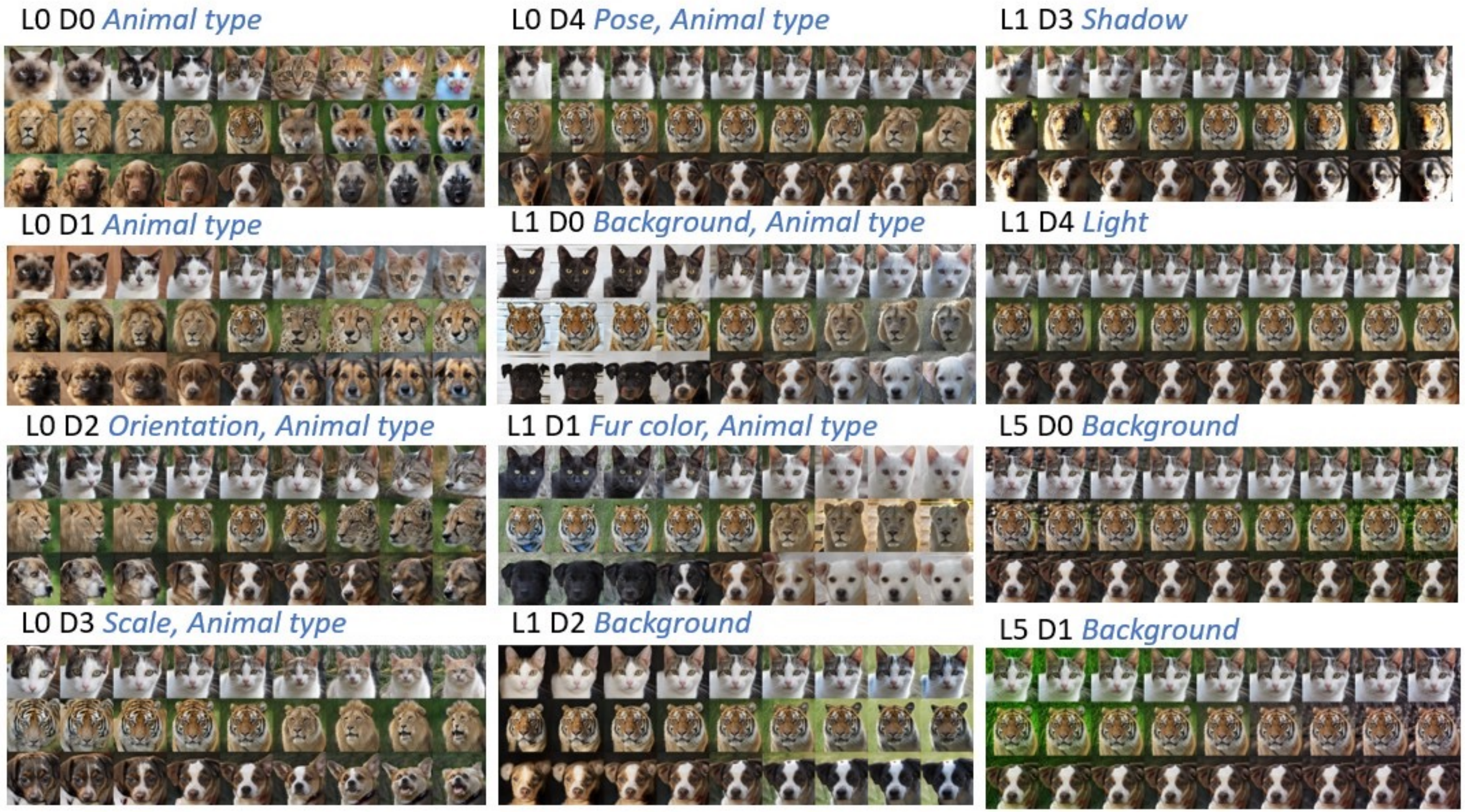




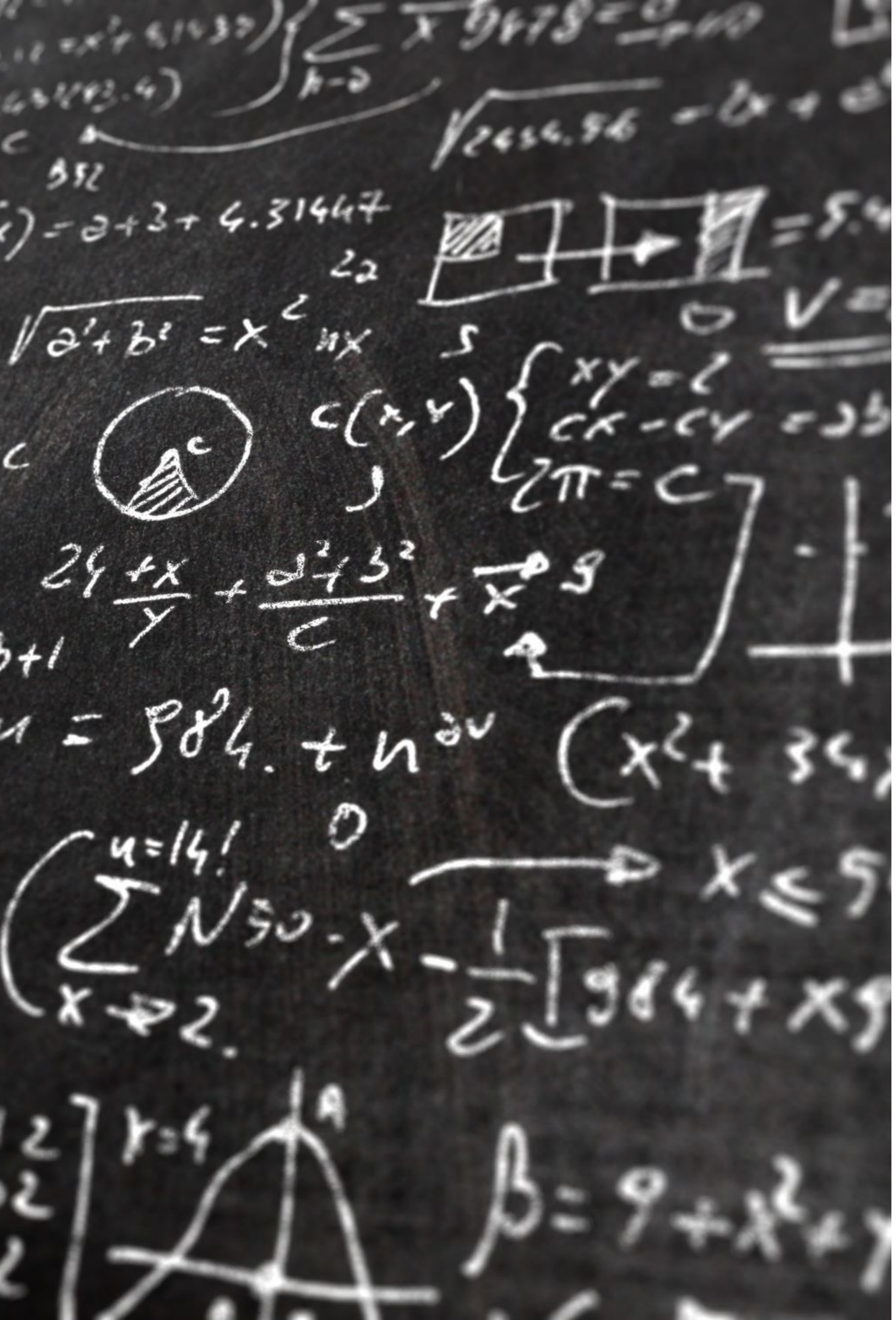


[2211.02486, 2307.05187]

Explore & **interpret** learned feature space

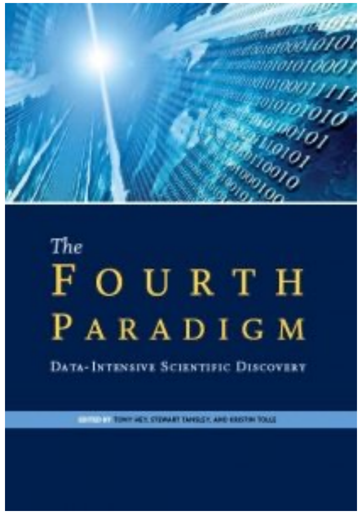


[InfoSCC-GAN]



The 4 paradigms of scientific discovery

1. Observation of natural phenomena
2. Theoretical models of nature
3. Numerical computation
4. Data-intensive scientific discovery



Jim Gray

"All physics is known, but the equations are too difficult to be solved."

- Paul Dirac, 1929 (paraphrased)

The 5^{th} paradigm:

tackle this challenge with ML

Long list of hard-to-model systems

- Schrödinger’s equation for “large" systems
- Quantum chemistry
- Material design (~10^{180} stable materials)
- Molecule design (~10^{60} small-molecule drug candidates)
- Fluid dynamics
- Weather
- Climate
- Fusion
- The Universe
- **LHC collisions: particles interactions with the ATLAS detector**
- **Experiment design**
- …

Challenge: how to explore these vast spaces?

The 5th paradigm: new ML simulators

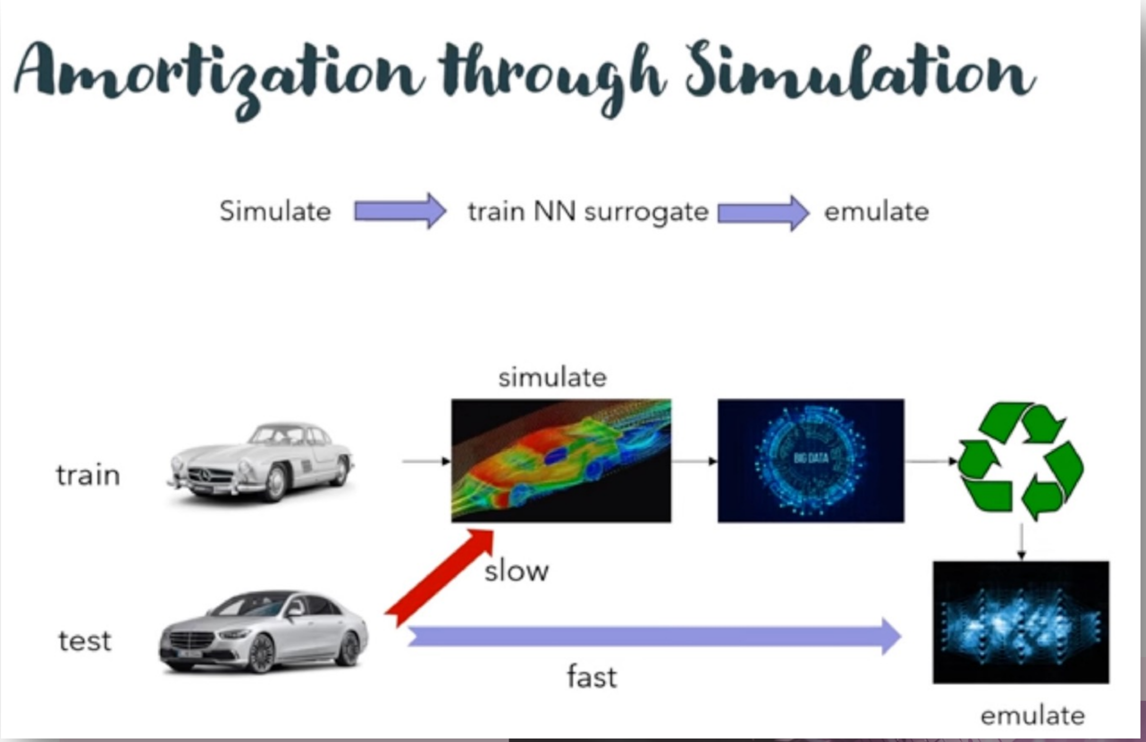
- *Classical* simulators are great !
  - They encapsulate ALL our understanding to model complex systems

- But they are typically
  - Prohibitively costly
  - Non-differentiable

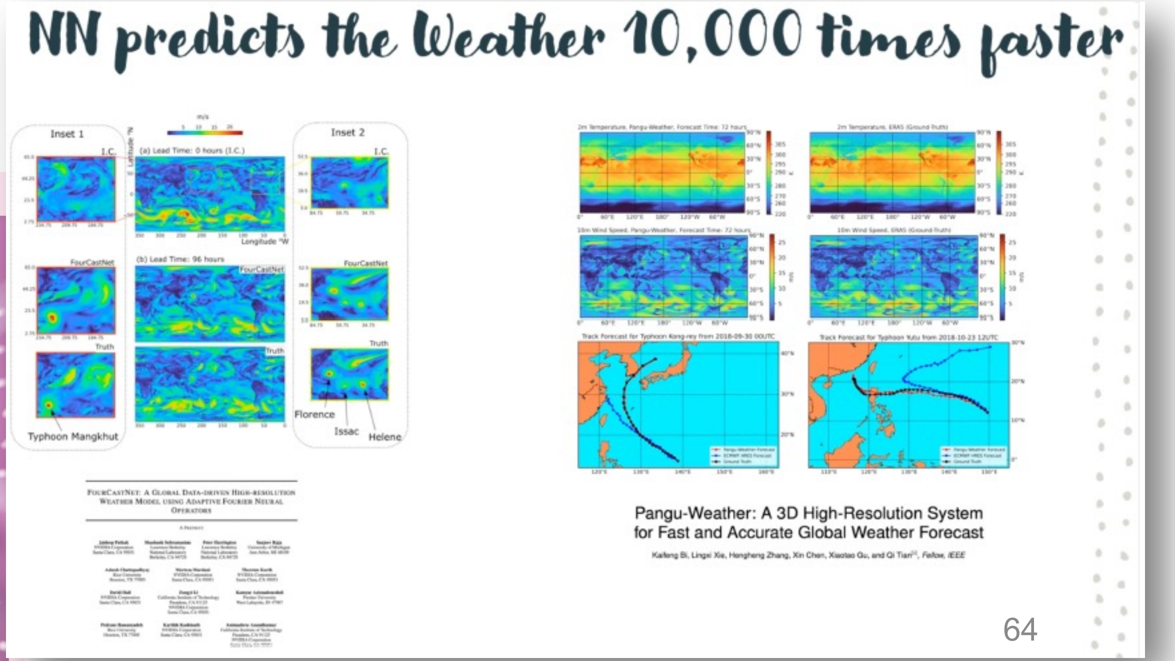
- **ML to rescue: generative models** as effective emulators
  - Data representation (image, time series, point cloud)
  - Inductive bias (symmetries)
  - Differentiable for automation (design, optimisation, inference)
  - Low computing cost

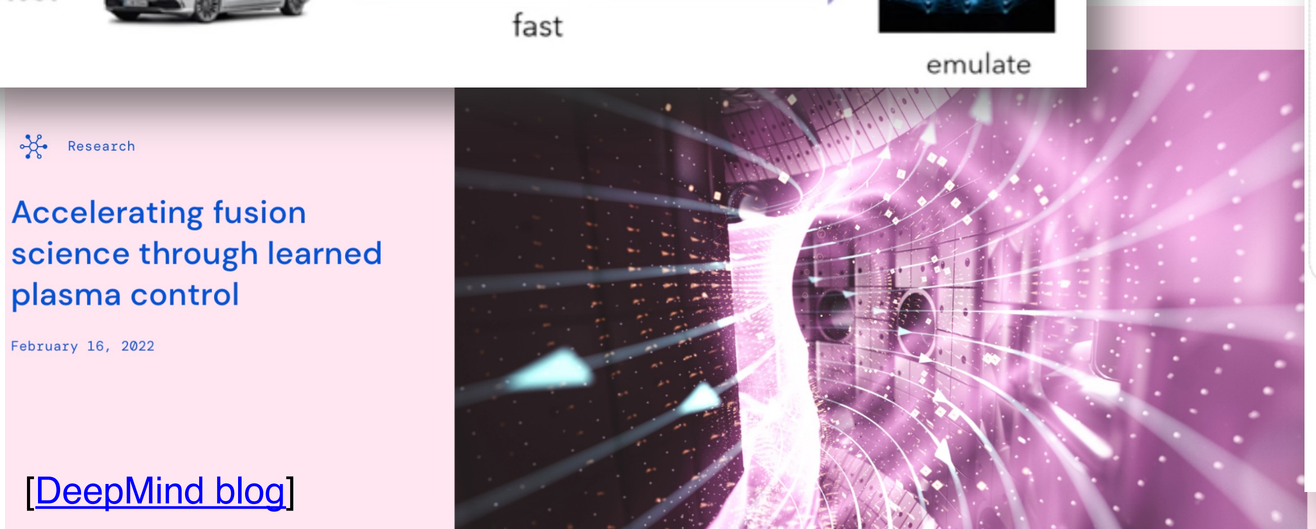
[Bishop, Microsoft]

Simulation & emulators



Max Welling
@ EuCAIFCon 2024

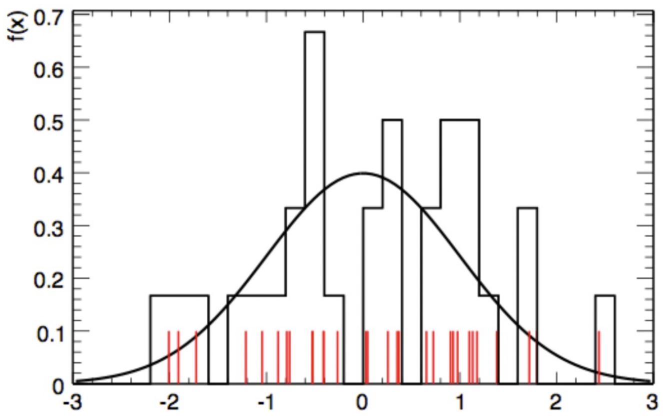




[DeepMind blog]

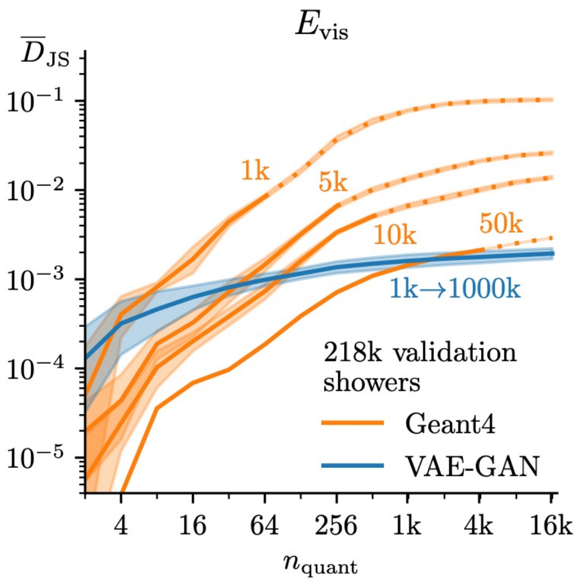
What does a GenModel actually learn?

- Data memorization?
  - Overfitting does NOT seem to be a problem



- The mechanism underlying this amazing performance is **poorly understood**

- Related to underlying **strong inductive bias**
  - [GANplification]



D_{JS} Jensen-Shannon divergence: How well reproduce the truth density

Smaller is better

Train GenModel on 1k and produce 1000k examples

~better than 50k real data

Implicit inductive bias

Generative models let us estimate the probabilities of data occurrences

– even in *non-populated* regions of data space (sparsity)

Evaluation of generative models

- **Comparing multivariate (high-dim joint) distributions is *hard***
- No *best* GOF test with power against *all* alternative hypotheses
  - Need to know *relevant* alternatives p_1
- Set of *practical* tests to establish *trust*
  - Reproducible
  - Standardizable (image, point cloud)
  - Computationally efficient
  - Interpretable
  - Pragmatic: *good enough* if OK for task at hand
- Typical problems
  - GAN: mode collapse [diversity]
  - VAE: blurriness [quality]

$$p_{\text{model}} \overset{?}{\approx} p_{\text{data}}$$

[2211.10295]

Search for the Unknown

BSM stubbornly resists discovery

ATLAS + CMS = O(1000) search papers

O(8’000) person years

~2 years per analysis

Average of ~4 people

Best use of resources ?

[Shikma Bressler, Hammers & Nails ’22]

Who thinks we can do better?

№ 1 Priority

Maximize Discovery Potential

Given available (human & compute) resources

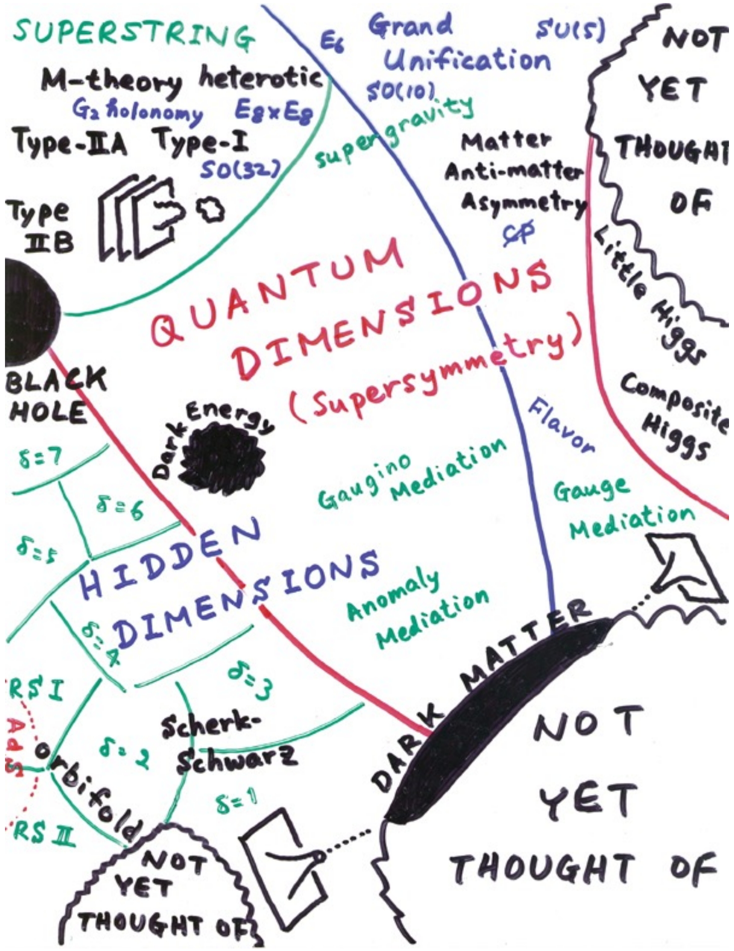
How much **signature** space have we explored?

	e	μ	τ	q/g	b	t	γ	Z/W	H	BSM $\to$ SM$_1$ $\times$ SM$_1$				BSM $\to$ SM$_1$ $\times$ SM$_2$				BSM $\to$ complex			
										q/g	γ/π^0's	b	$\cdots$	tZ/H	bH		$\cdots$	$\tau qq'$	eqq'	$\mu qq'$	$\cdots$
e	[37,38]	[39,40]	[39]	$\varnothing$	$\varnothing$	$\varnothing$	[41]	[42]	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	[43,44]	$\varnothing$	
μ		[37,38]	[39]	$\varnothing$	$\varnothing$	$\varnothing$	[41]	[42]	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	[43,44]	
τ			[45,46]	$\varnothing$	[47]	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$		[48,49]	$\varnothing$	$\varnothing$	
q/g				[29,30,50,51]	[52]	$\varnothing$	[53,54]	[55]	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$	
b					[29,52,56]	[57]	[54]	[58]	[59]	$\varnothing$	$\varnothing$	$\varnothing$		[60]	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$	
t						[61]	$\varnothing$	[62]	[63]	$\varnothing$	$\varnothing$	$\varnothing$		[64]	[60]	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$	
γ							[65,66]	[67–69]	[68,70]	$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$	
Z/W								[71]	[71]	$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$	
H									[72,73]	[74]	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$	
BSM $\to$ SM$_1$ $\times$ SM$_1$: q/g										$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$	
BSM $\to$ SM$_1$ $\times$ SM$_1$: γ/π^0's											[75]	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$	
BSM $\to$ SM$_1$ $\times$ SM$_1$: b												[76,77]		$\varnothing$	$\varnothing$	$\varnothing$		$\varnothing$	$\varnothing$	$\varnothing$	
$\vdots$																					
$\vdots$																					

Vast signature space **un**explored

[1907.06659]

How to quantify *coverage*?



- What theory prior? [Bayesian vs. Frequentist]
- How to interpret “model-agnostic” null results?
- Pragmatic metric?

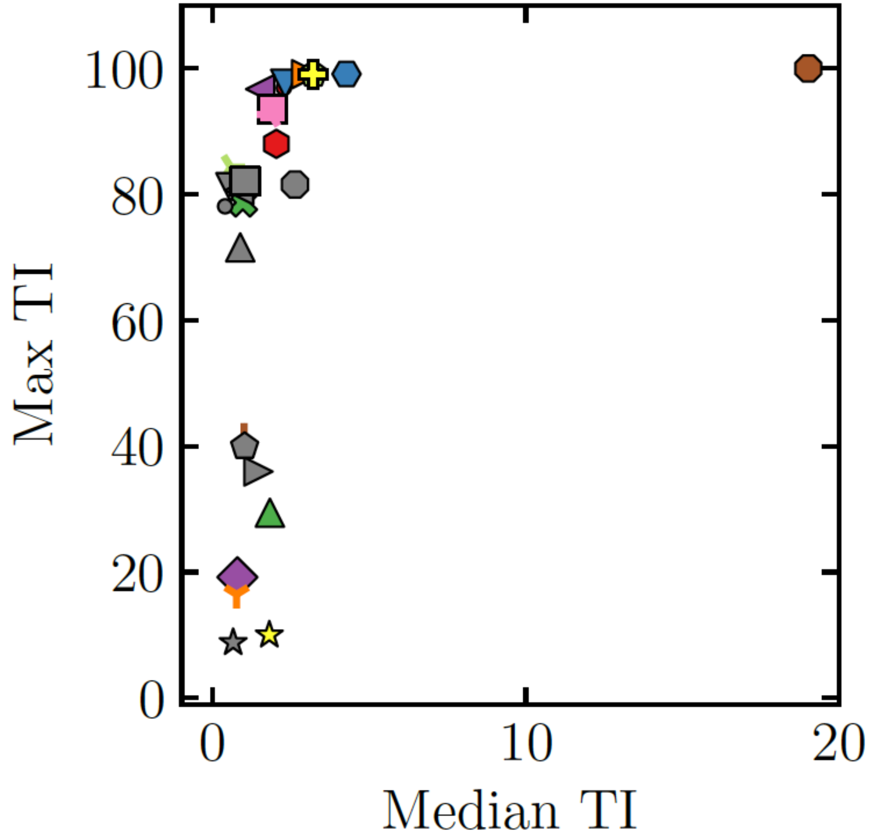
$$\frac{Maximum\ model\ space\ coverage}{Minimal\ set\ of\ searches} \sim \frac{\#\ excluded\ signal\ points}{\#\ searches}$$

  - Benchmark dependence
  - Correlation of final states [stop, 3rd gen LQ, $t\bar{t}$+MET]
  - Volume in *embedded space* [2208.05484]

What is the next best search given all existing search results?

Poor man's assessment: *benchmarking*

Compromise:
average over "many" benchmarks

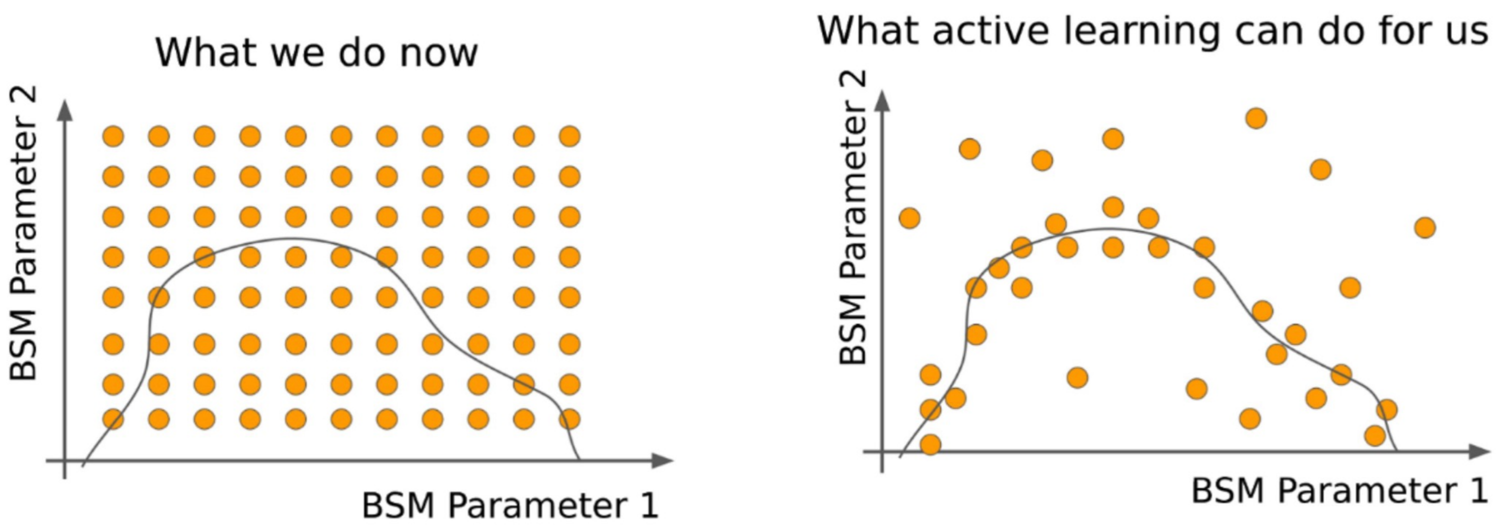


Make analysis **reinterpretable** for any future benchmark

[Need to provide models & likelihoods]

[2105.14027]

Smart sampling with active learning



Thrives on high-dimensional theory space

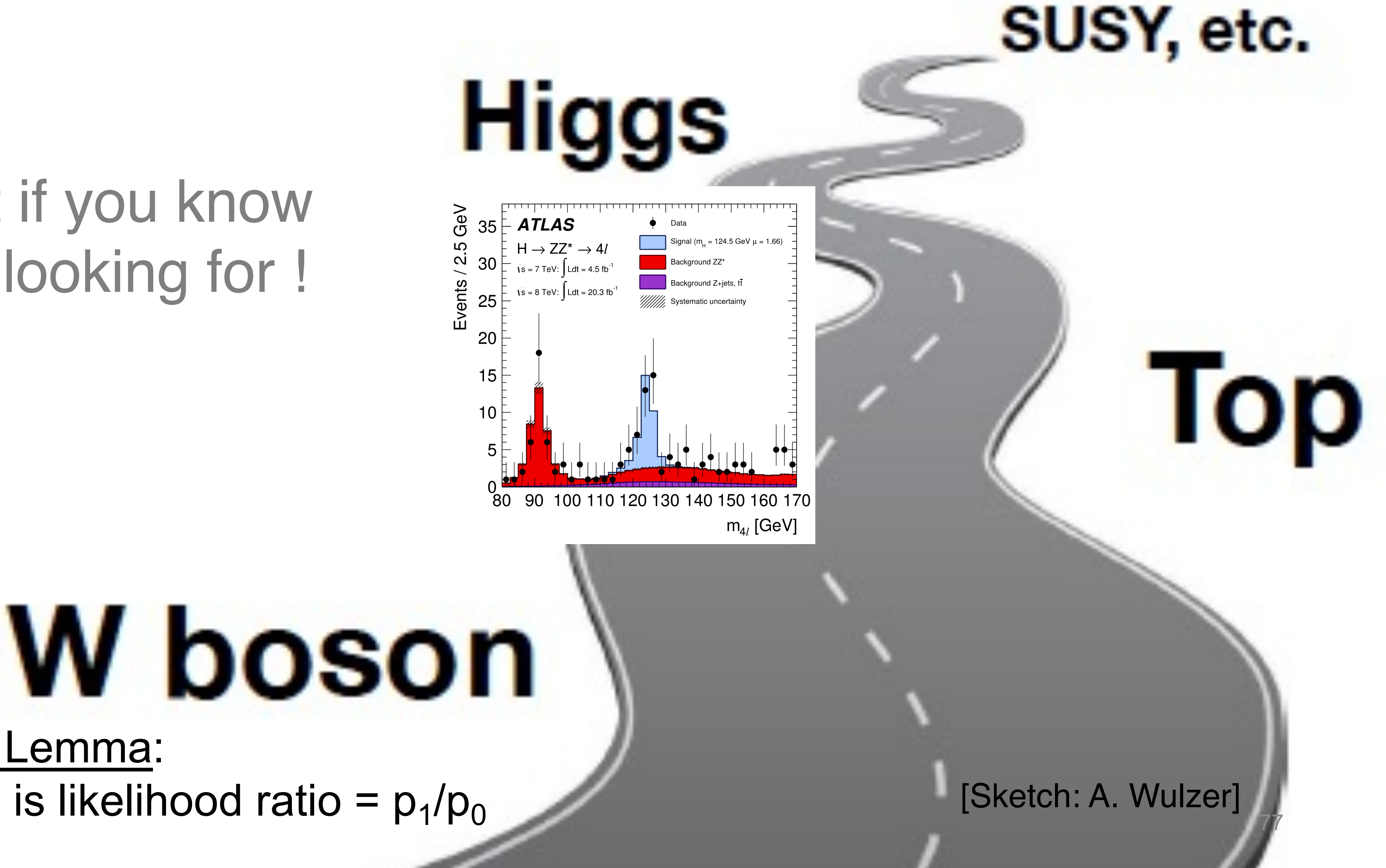
→ Simulate on demand

[Regress on upper limit with GP for ATLAS mono-H(bb) search, ACAT22]   [ATL-PHYS-PUB-2022-045]

Search for the Unknown
Part 2

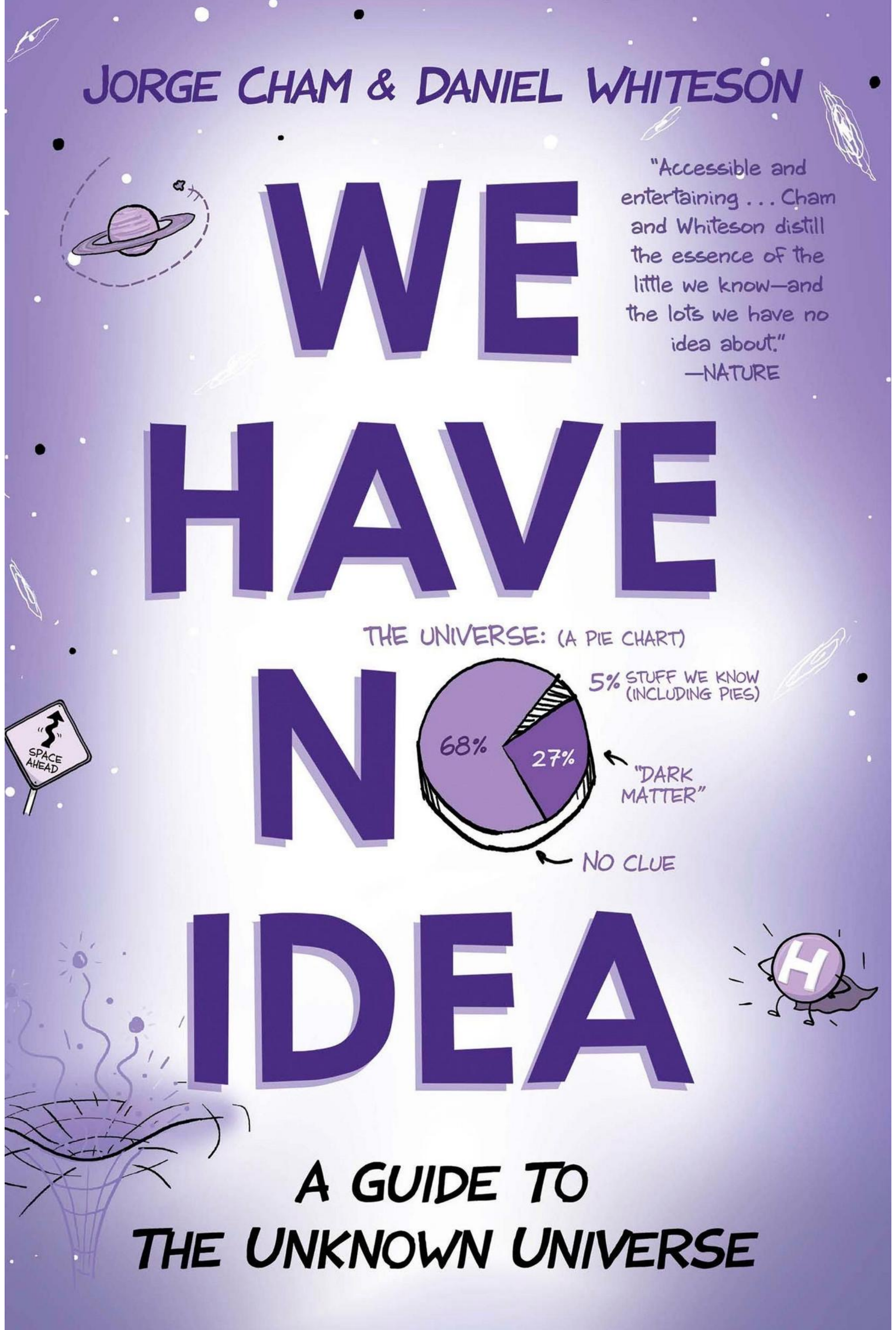
Our powerhorse: **2**-hypothesis test

Works great if you know what you're looking for !



<u>Neyman-Pearson Lemma</u>:
Best test statistics is likelihood ratio = p_1/p_0

[Sketch: A. Wulzer]



We don't know what we're looking for

No *trust* in p_1 = playing the lottery!

p_0 = SM

p_1 = *everything else*

Ideal test: p_0 & p_1 known (with high stats)
→ 1 single LH test (sufficient test statistics)

Optimal if:
- p_0 ~ SM
- p_1 ~ *true* BSM

Add realism:
- Finite stats
- Mismodeling
- p_1 NOT known

→ *Factorize* problem

$$\frac{p_1}{p_0}$$

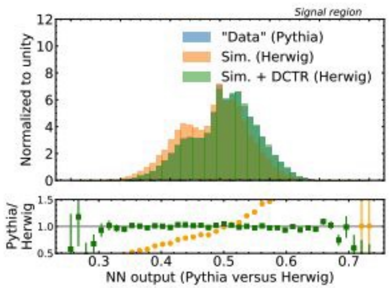
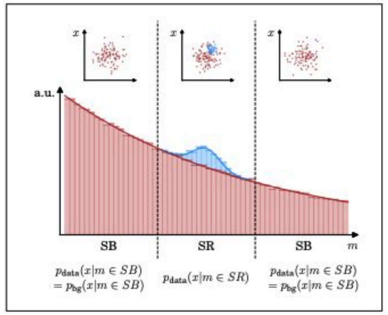
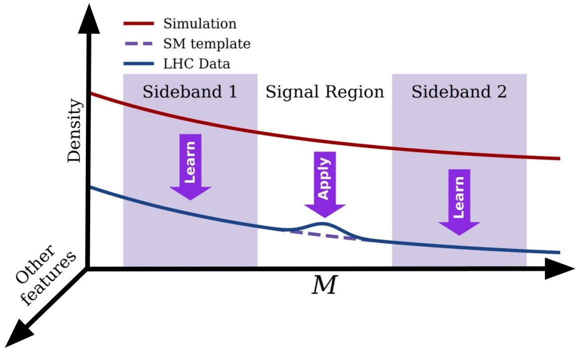
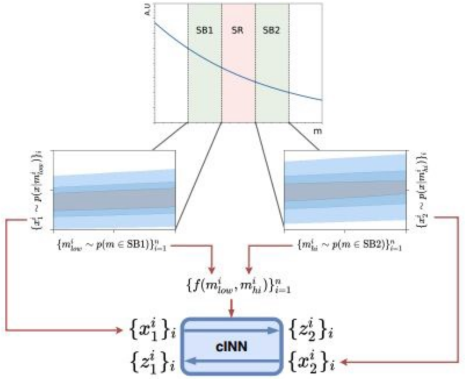
p_0 ~ MC:
- Limited accuracy
- Limited statistics

p_1 assumptions inform
- Event selection
- Feature choice

p_0 ~ in-situ BG estimate:
- CATHODE
- CURTAINS
- SALAD
- FETA

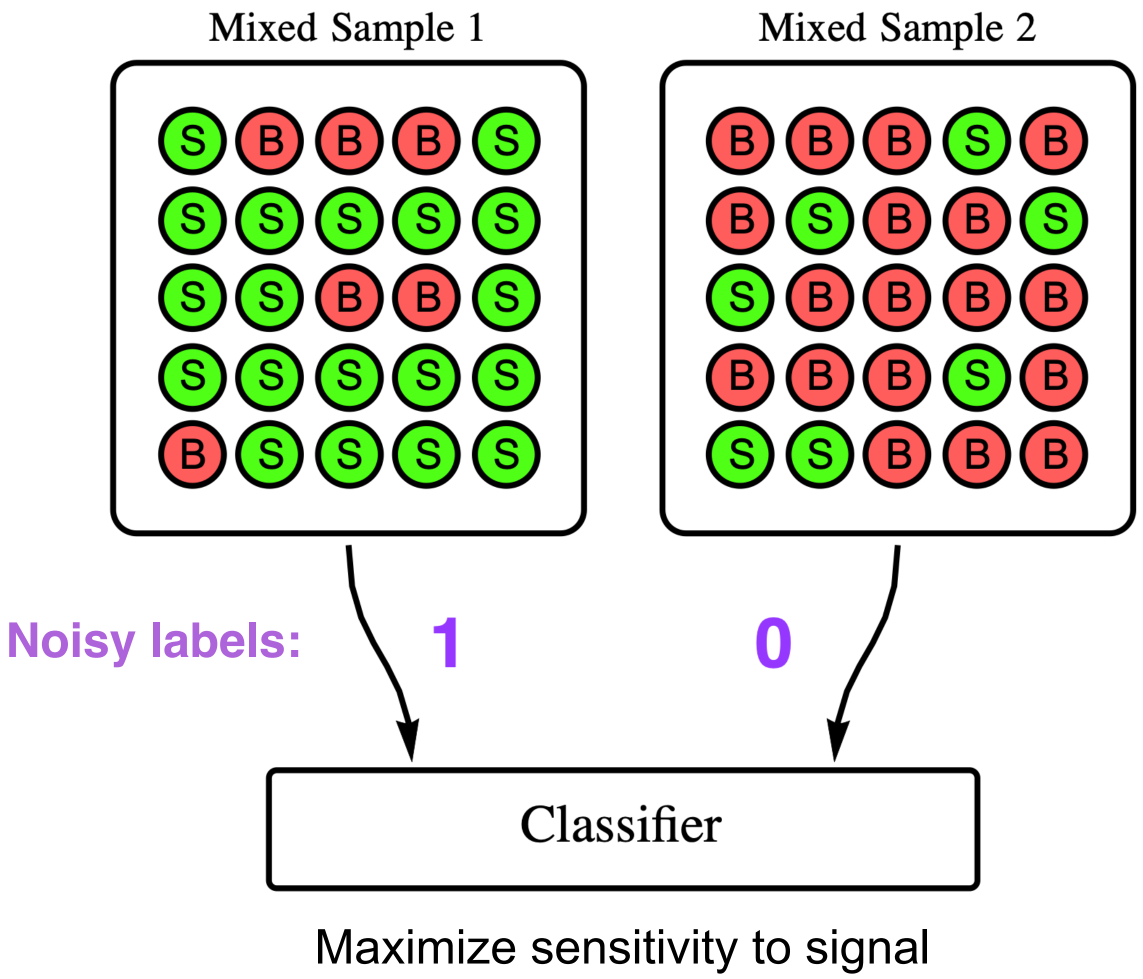
p_1 choices:
- Simplified MC model
- Parametric model (fit, NN,…)
- Learn p_1 from data → NPLM
- Approximate LH ratio with CWoLa classifier

Learning high-D background templates*

	Learn from simulation	Learn from data (SB)
Modeling the likelihood ratio	[SALAD] 	[CATHODE]*  [*see also [LaCATHODE] & [ANODE]]
Morphing the features	[FETA] 	[CURTAINs] & [Flow4Flows] 

[*Fidelity of simulation alone insufficient]

Classification without labeling (CWoLa)



Abandon notion of *event label*

Noisy labels to be S or B

Bump hunt [1902.02634]

ATLAS analysis [2005.02983]

Beyond resonances

e.g. symmetries [2203.07529]

Gedankenexperiment

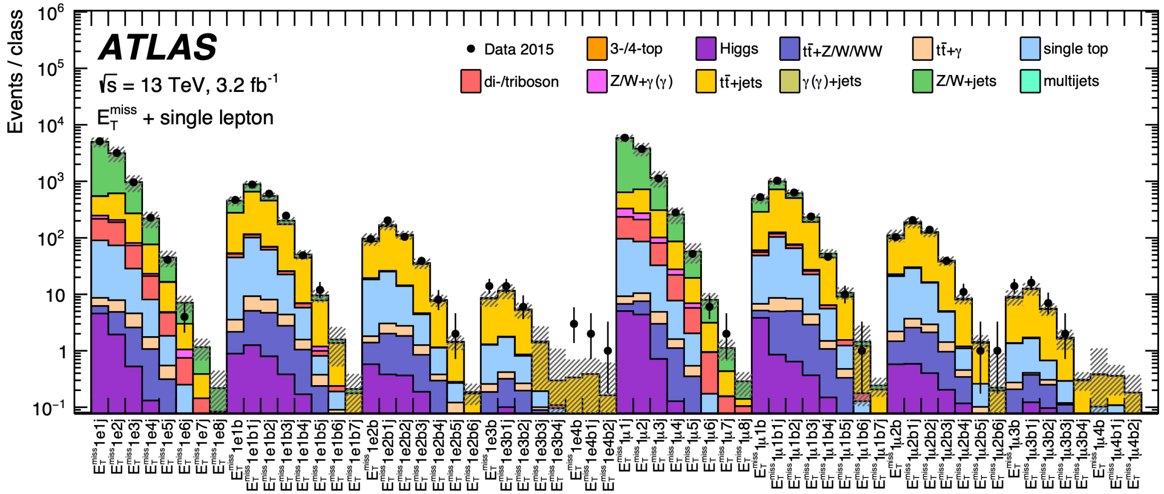
Fast parametrized model exists for all possible p_1 **including the true p_1**

$$\frac{p_1}{p_0}$$

Known perfectly

How to design the optimal search strategy?

A question of **automation**



10^5 signal region [1807.07447]

- One classifier?
- Event selection?
- Feature space?
- Data slicing [# tests]?
  - Look elsewhere effect

- Dial up/down the physics prior?

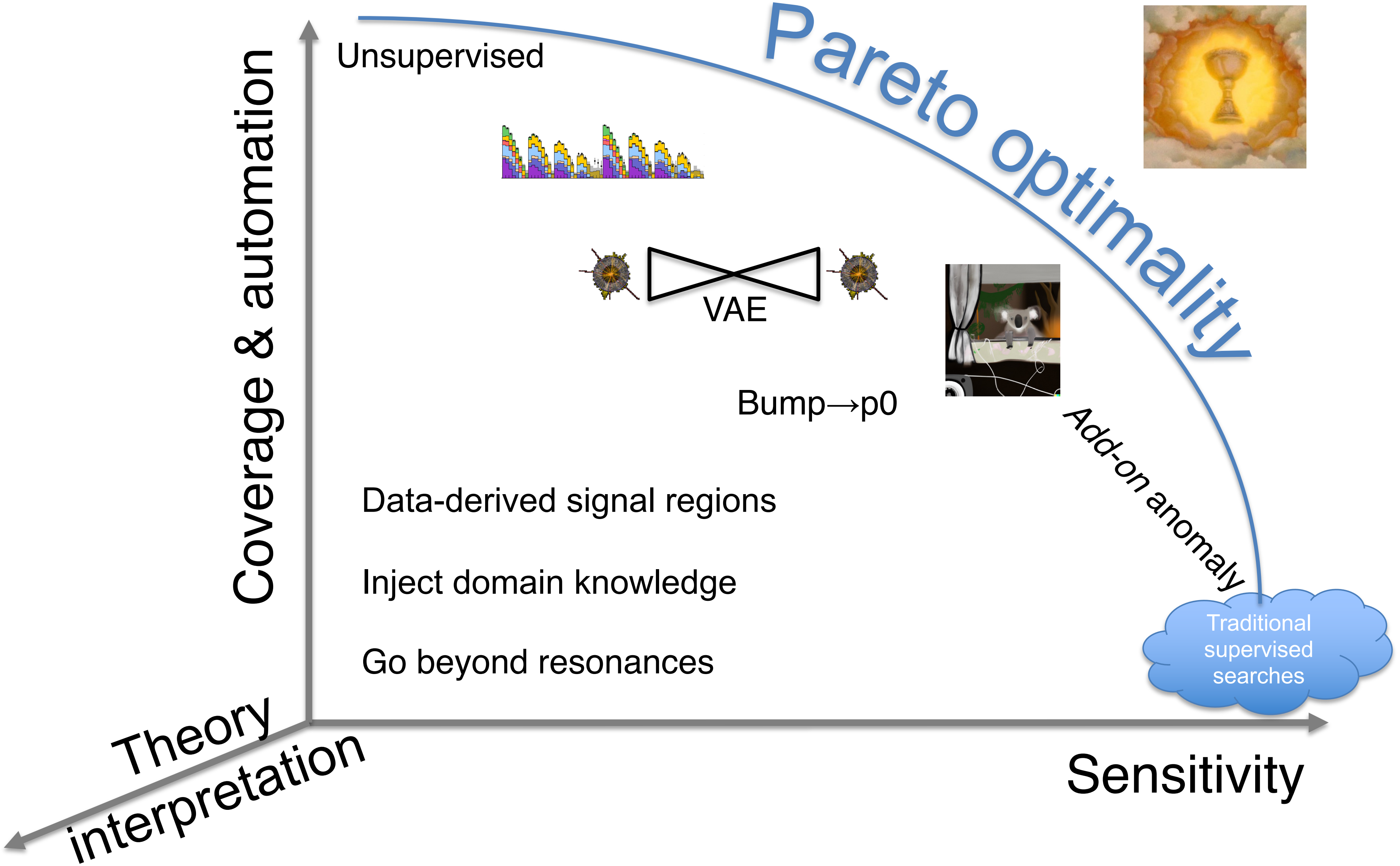
- Interpretation w/o benchmarks?

Lots of open questions & room for YOU to make a BIG impact

Optimal search for the unknown

- Trade-off between **generality** and **specificity**
- Knob to tune pareto optimality between the endpoints: **supervised** & **unsupervised**
- What **metric** to assess performance – should **not** be known models
- What’s the **follow-up strategy** after an “anomalous” signal ?
  - Balance cost of follow-up against frequency alerts ?

Diverse Search Strategy



Concept → Production

[Innovate → Exploit]

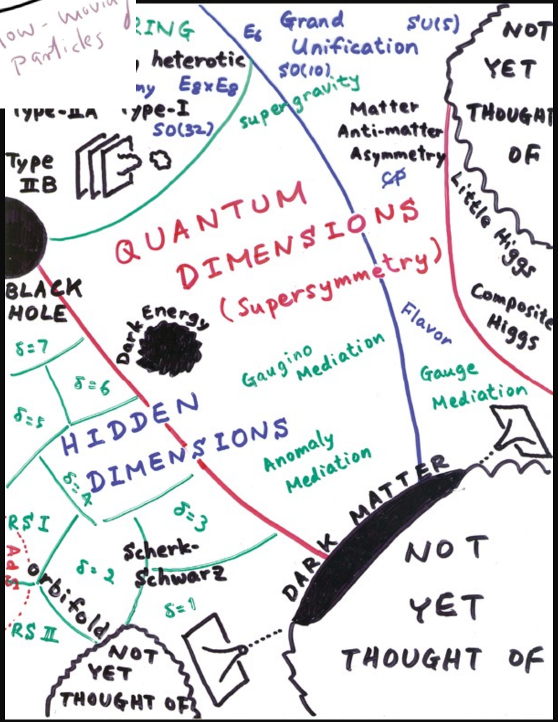
Use in Experiment!

Scientists model the world

The usefulness of a common latent space

Signatures

Theory



All models are wrong, but some are are useful.

– GEORGE BOX

Useful in what sense?

Statistical Modeling

X → Nature → Y

Data → Model → Evaluation

Interpretable [e.g. SM]

↕

Black box [AI]

- Prediction
- Decision-making
- Discovery
- Explanation

[Leo Breiman 2001 on statistical modeling: the two cultures]

Geoff Hinton

There exists some "universal" learning algorithm that can learn ***anything***: language, vision, speech, etc. The brain is based on it, and we're working on uncovering it. (Hint: the brain uses neural networks)

Expert knowledge:
full specific knowledge

Machine Learning

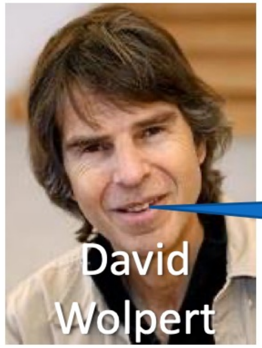
none

more data →

Use data to fit specific model

Expert Systems : **Physics laws**
(no data at all)

no free lunch



There is no "free lunch": no learning is possible without *some* prior assumption about the structure of the problem (prior knowledge)

5. THE USE OF DATA MODELS

Statisticians in applied research consider data modeling as the template for statistical analysis: Faced with an applied problem, think of a data model. This enterprise has at its heart the belief that a ~~statistician~~, by imagination and by looking at the data, can invent a reasonably good parametric class of models for a complex mechanism devised by nature. Then parameters are estimated and conclusions are drawn. But when a model is fit to data to draw quantitative conclusions:

physicist

- The conclusions are about the model's mechanism, and not about nature's mechanism.

It follows that:

- If the model is a poor emulation of nature, the conclusions may be wrong.

[Breiman 2001]

7.1 A New Research Community

In the mid-1980s two powerful new algorithms for fitting data became available: neural nets and decision trees. A new research community using these tools sprang up. Their goal was predictive accuracy. The community consisted of young computer scientists, physicists and engineers plus a few aging statisticians. They began using the new tools in working on complex prediction problems where it was obvious that data models were not applicable: speech recognition, image recognition, nonlinear time series prediction, handwriting recognition, prediction in financial markets.

[Breiman 2001]

Their interests range over many fields that were once considered happy hunting grounds for statisticians and have turned out thousands of interesting research papers related to applications and methodology. A large majority of the papers analyze real data. The criterion for any model is what is the predictive accuracy. An idea of the range of research of this group can be got by looking at the *Proceedings of the Neural Information Processing Systems Conference* (their main yearly meeting) or at the *Machine Learning Journal*.

[Breiman 2001]

Model vs. task culture

Model culture

Data first [e.g. LHC]

Goal: build a data model (x, y)

Evaluate: generalization

Task culture

Task first [e.g. cats vs. dogs]

Goal: best possible task fulfillment

Evaluate: on task

Used to drive most of the AI breakthroughs [e.g. AlphaFold]

Physicists do both !

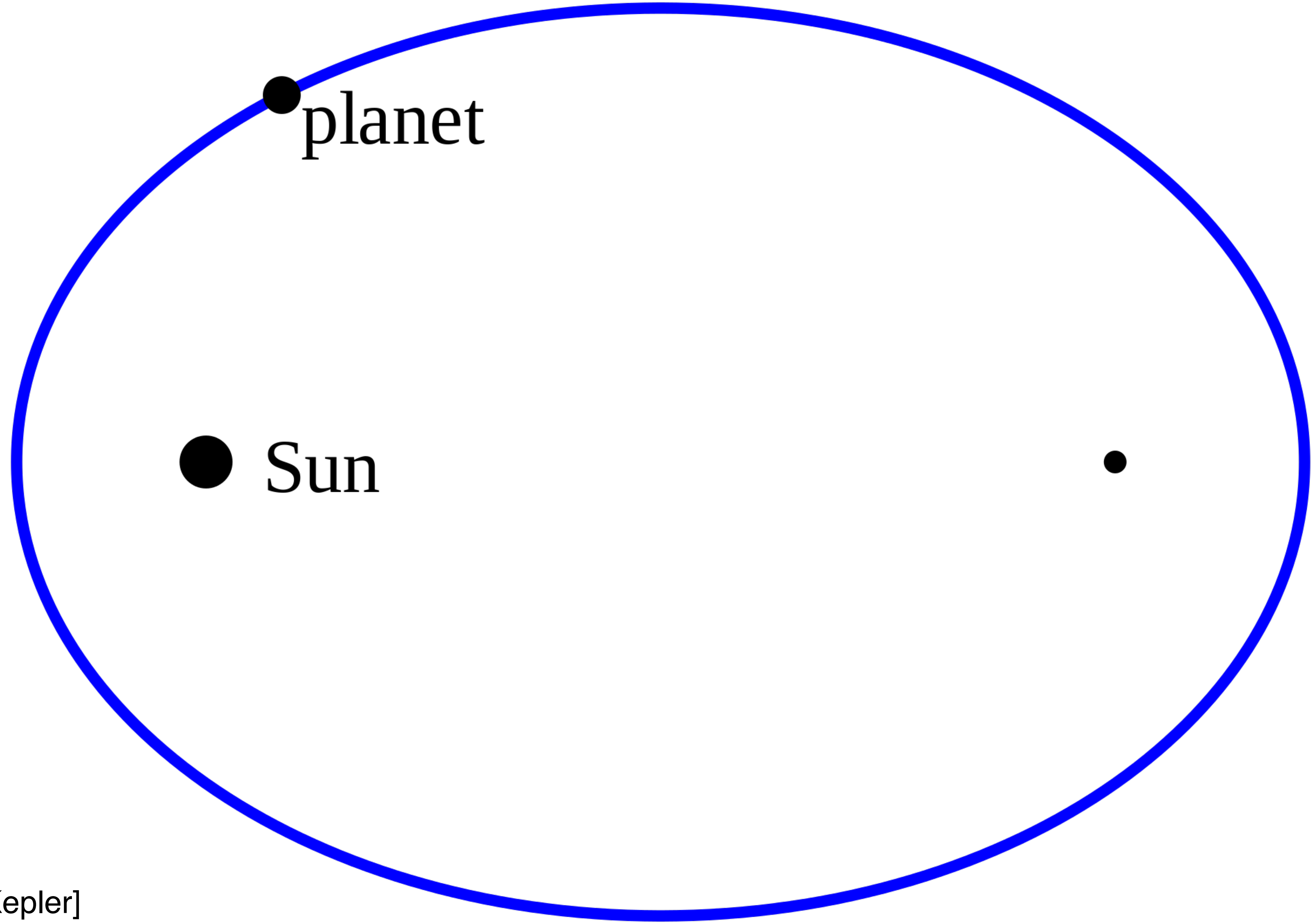
The *essence* of science?

Prediction machine

Finding new regularities

Learning saves computational resources

Reduce dimensionality of problem



[Example: Kepler]

Learning *meaningful* **latent** representations

Plato: myth of the cave

<u>The quest of science</u>:

Learn true underlying objects (latent variables)

from observed data (shadows)

The promise of foundation models

THE WORLD BEYOND OUR **CURRENT** REACH OF SCIENCE

Access to this *hidden* world with AI?


Will we ever be able to talk to whales?

Imagine an **AI oracle**…

- …which would give the true answer to any question !

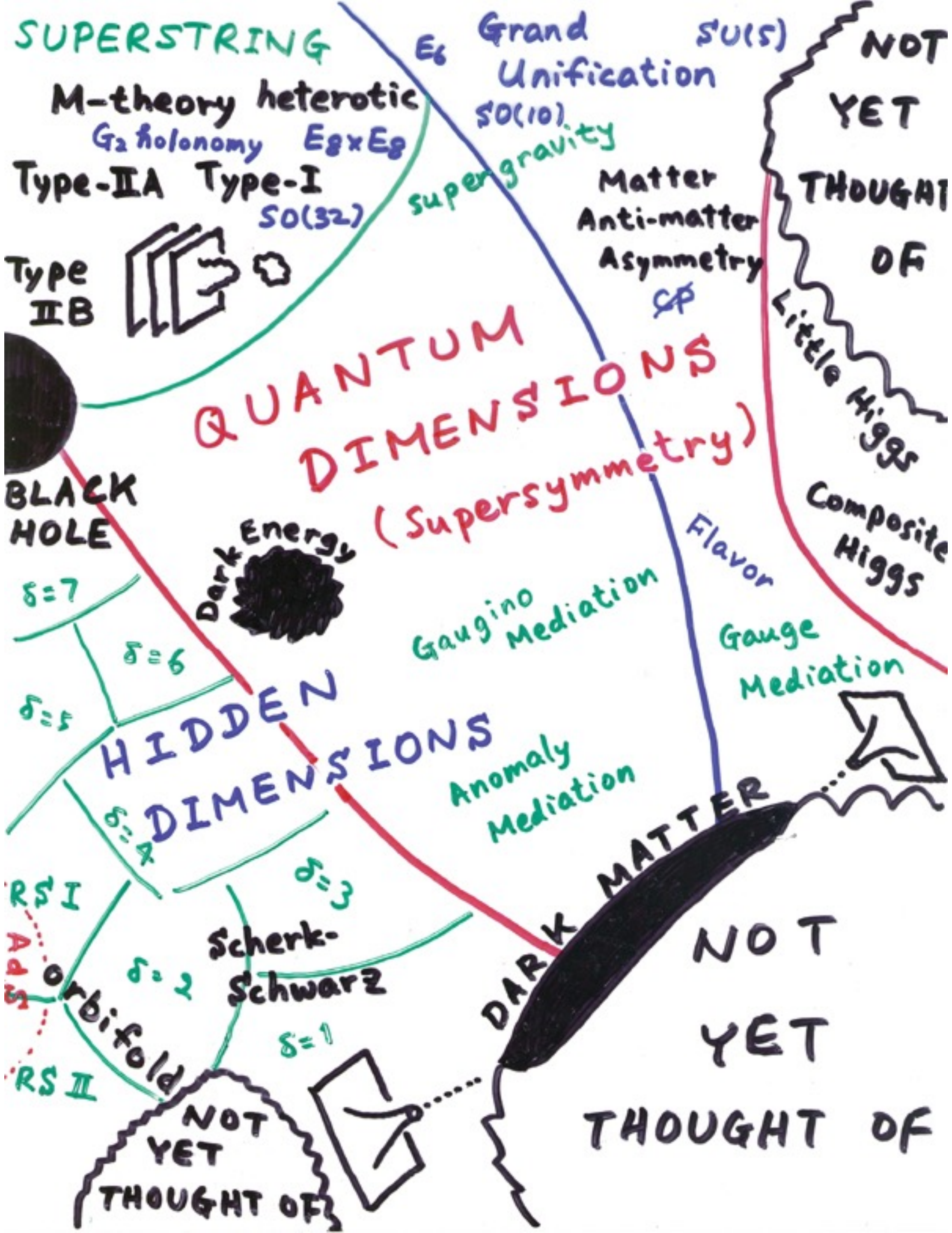
- Such an oracle would revolutionize science and technology as we know them

- Would scientists be satisfied?
  - No. But what if it's the best we will get?
  - What if the *theory of everything* is "beyond human comprehension"
  - Does it matter if a **human** or an **AI** writes a popular science book
    - And explains it to you like a 5-year-old?

Recap: what is a generative model?

An **implicit model** that describes how data was generated [probability density]

[there is no model-less model]
[ChatGPT = implicit model of human-language text]
[DALL.E2 = implicit model of natural images]

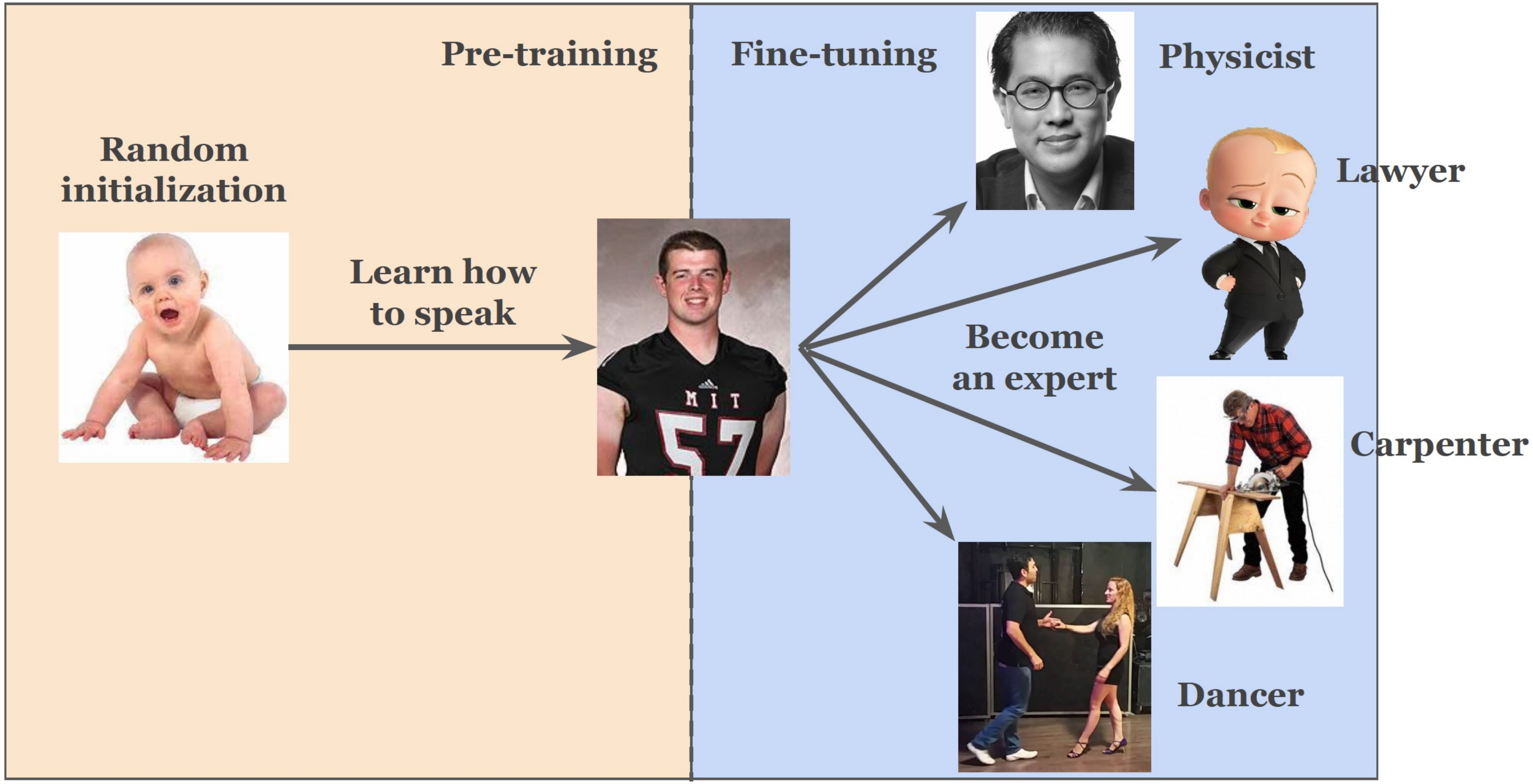


…and we want a *model* of our natural world !

Theory-driven: human ingenuity

Data-learned: foundation models

The idea of a foundation model (FM)



1. **Pre-train** on big unlabeled data

2. **Fine-tune** on labeled data + transfer learning

[Image credit: Kazuhiro Terao]

Characteristic features of a FM

Pre-train using SSL – no labels needed !

Encode in meaningful data representation

Transferrable & finetunable: adopt to multiple downstream tasks

Multimodality: common embedding / no pairing needed

FMs = stochastic generative models with high expressiveness and outstanding interpolation and generalization power in ultra-sparse training data spaces of high dimensionality.

Pre-training

[2002.05709]

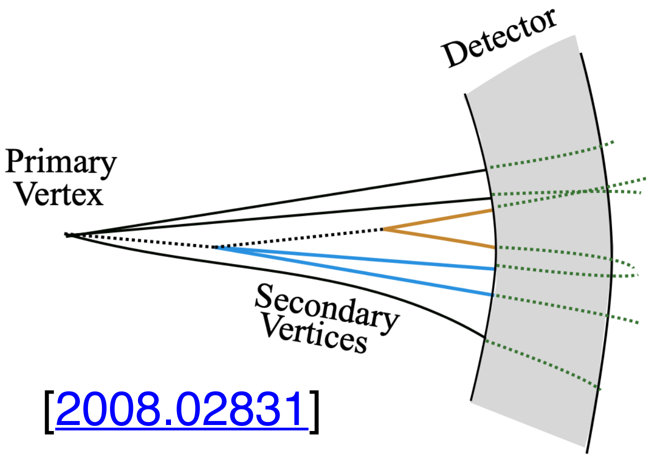
Augmentation [Re-sim]

Masking [next word prediction]

Novel physics-inspired training schemes?

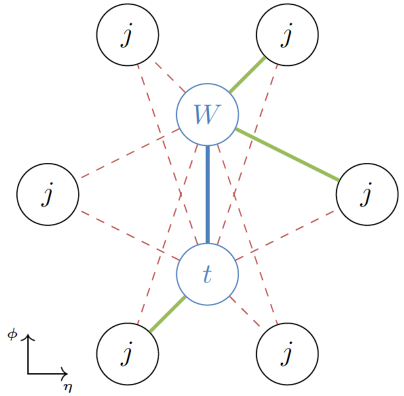
Train using auxiliary tasks?

Encode physics [*flexible* prior]



[2008.02831]

Evaluation: go beyond downstream task?



[2303.13937]

(b) Topograph

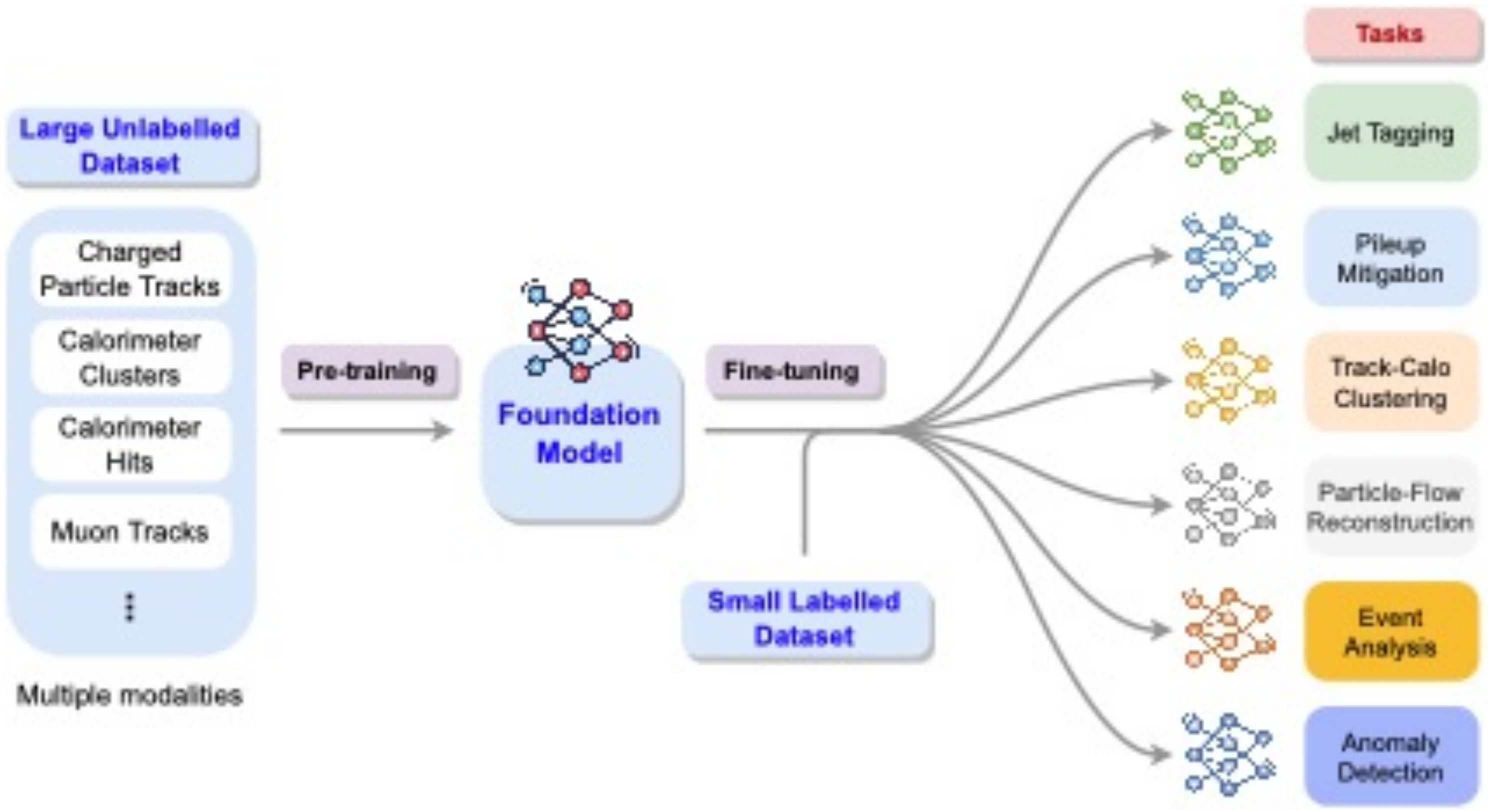
Example: masked particle modeling

Pre-training task:

Mask & predict constituents of a jet

Fine-tune for downstream tasks:

- Classification
- Weak supervision
- …



[2401.13537]

We have **our own** embedding spaces

Reconstruction = common embedding space of our data

Theory space = multi-modal common embedding space

What do FMs add to this?

- **End-to-end**
- **Differentiable**
- **Democratize** AI – commonly trained
- **Common** model across subdetectors, experiments, domains,…

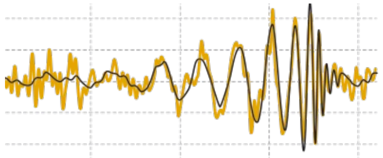
FM trend: **task** culture → **model** culture

- Implicit model of the data
- Task-specific → generalizes across tasks
- **Model first** – then downstream tasks

- **AI oracle ⇔ interpretability**
  - Machine *understands* & explains it to *5-year-old [us]*
  - Symbolic regression → map to our *simple* description of math symbols
  - Limitation of human brain:
    - Humans can only hold 5-9 concepts in working memory at once [length of equation]
    - 2D visualization for human eyes ⇔ model of the universe

ML interpretability for science

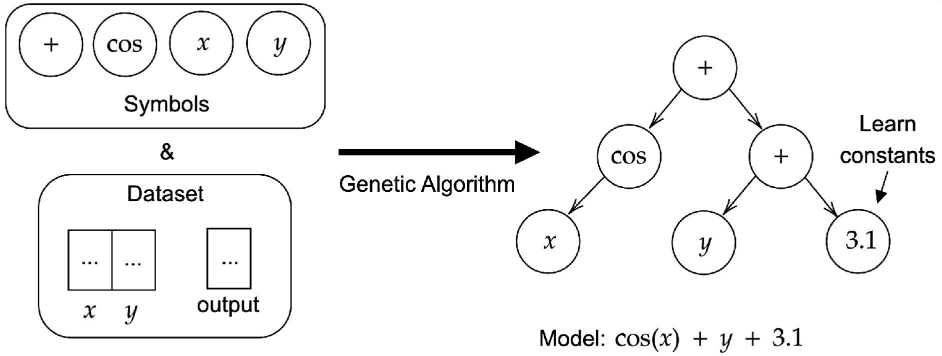
Science



$$h = \frac{2G}{c^4}\frac{1}{r}\frac{\partial^2 Q}{\partial t^2}$$

Computer vision

???

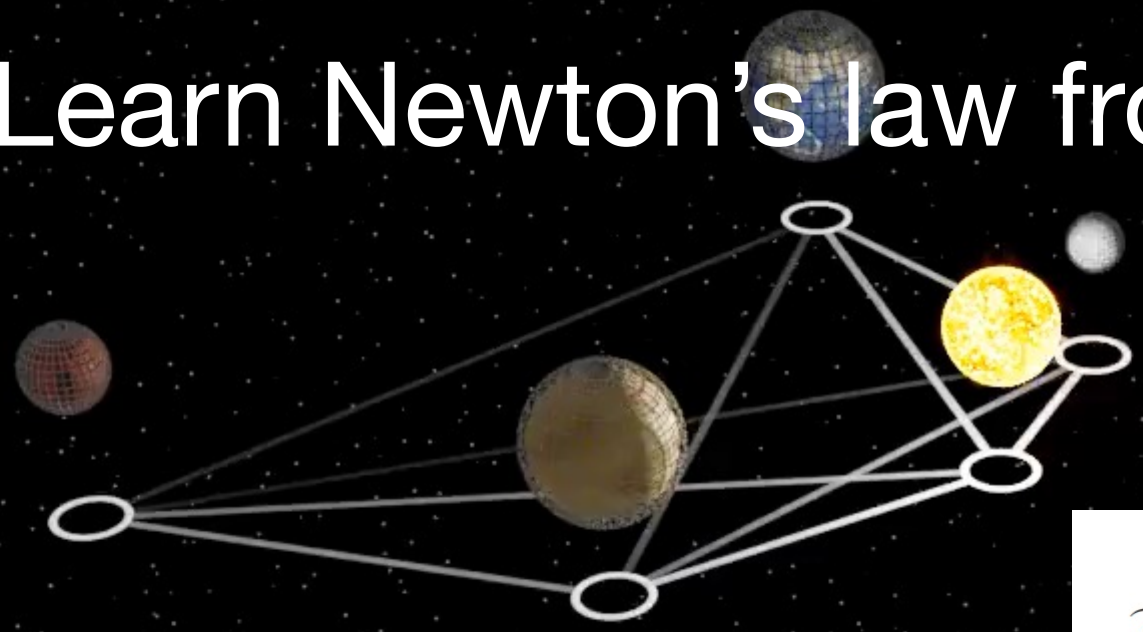


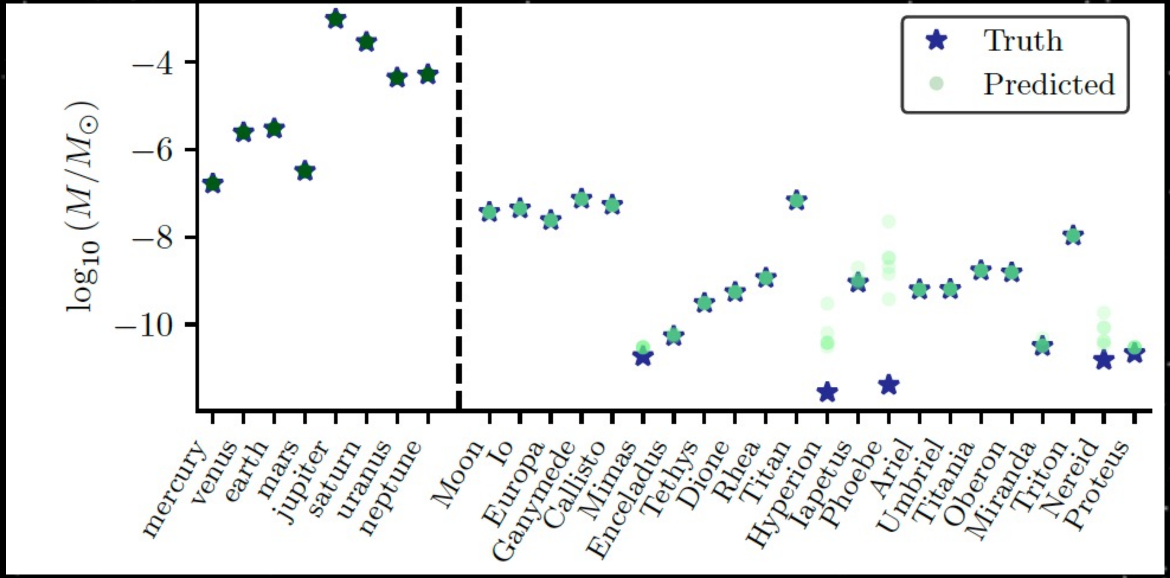
Analytic models often generalize better than NN
Symbolic regression as inductive bias

NN weights [black box] → PySR → **Analytic expression** [insights]

[Miles Cranmer – Hammers & Nails 2022]

Learn Newton's law from solar system





GNN → PySR → Learn masses + dynamics

[2202.02306](2202.02306)

True

Predicted

Search for the Unknown
Part 3


LOOKING UNDER THE LAMPPOST
LOST YOUR KEYS?
YEAH, I LOST THEM OVER THERE BUT THE LIGHT'S BETTER HERE
sketchplanations

Foundation models for discovery

Common / portable model [efficient]

Accelerate with surrogate models

Automation Automation Automation

Technology → automation → human evolution

The Industrial Revolution (18th Century):
Manual work ⇒ operating machinery

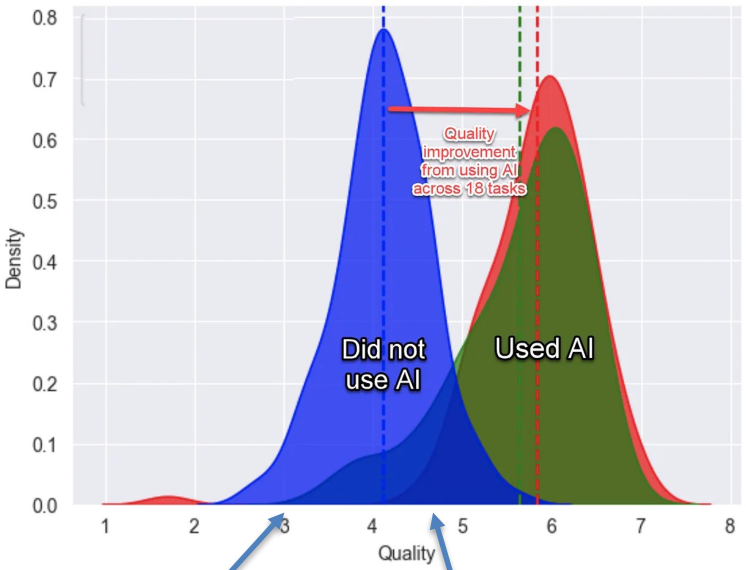
The Information Age (Late 20th Century):
Specialized knowledge in programming, data analysis, & more

AI & ML (21st Century):
Need for human expert knowledge, human-AI collaboration, (personalized) human augmentation, building trust [e.g. diagnosis]

[https://www.linkedin.com/pulse/paradox-ai-how-automation-amplifies-need-expertise-rishi-sharma]

AI → skill evolution & skill leveler

LLMs augment human intelligence:



Distribution of output quality across all the tasks. The blue group did not use AI, the green and red groups used AI, the red group got some additional training on how to use AI.

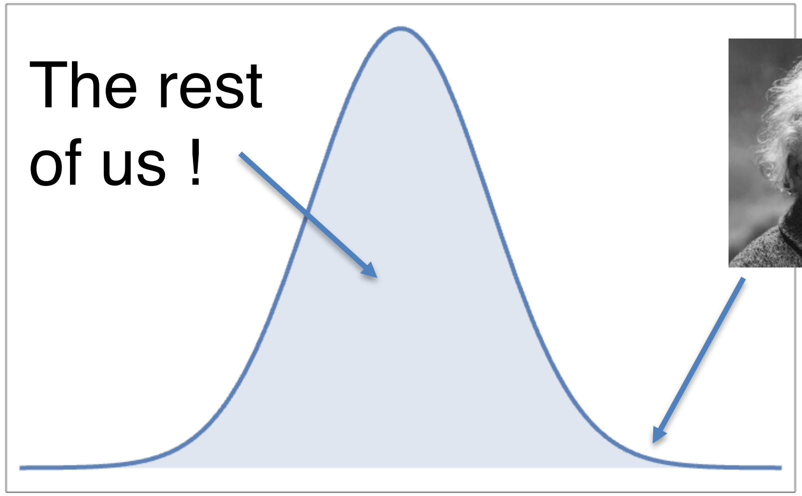
bottom half improved 43%

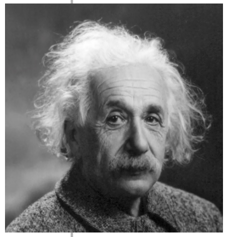
top half improved 17%

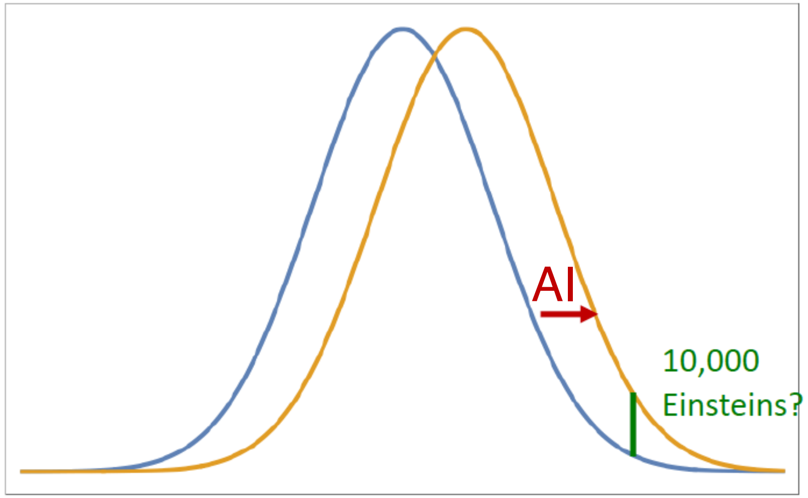
⇒ Skill leveler

Dell'Acqua et al, “Navigating the Jagged Technological Frontier” (Harvard Buisness School, 2023)

Applied to physics







[Matt Schwartz, EuCAIFCon 2024]

Humans are *mixtures of experts*

- Specialize in subset of input data
- Jointly perform complex tasks

- Same trend in AI:
  - Increased model capacity
  - Reduced computational burden
  - Faster training

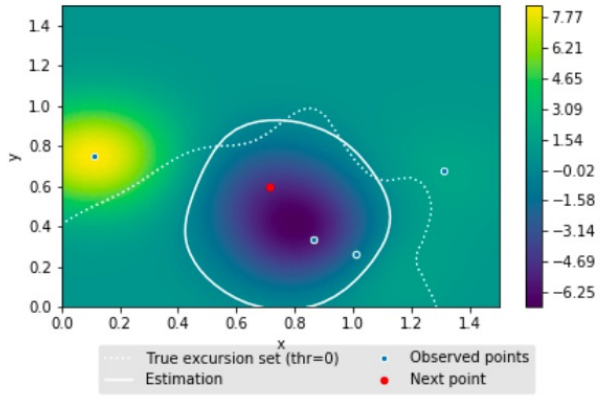
What if secrets of nature are NOT in *our current data?*

$$p(\text{theory} \mid \text{data}) = \frac{p(\text{data} \mid \text{theory})\, p(\text{theory})}{p(\text{data})}$$

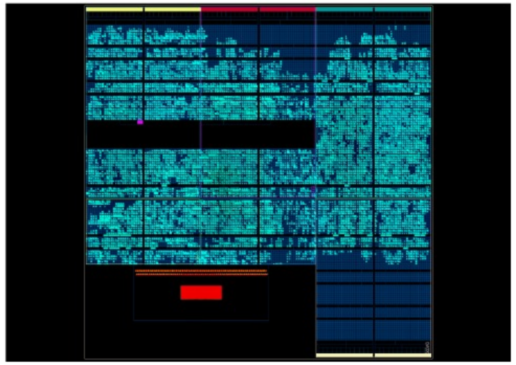
[Lukas Heinrich - Detector design using differential programing]

Ultimate goal:
Learning about Nature

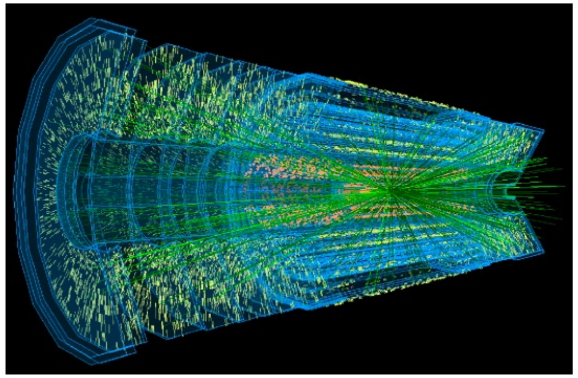
Optimizing the science output



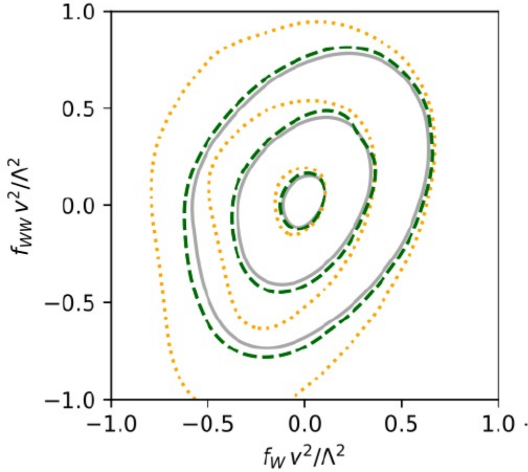
Optimal Theory Exploration



Optimal Data Taking / Experiment Operations



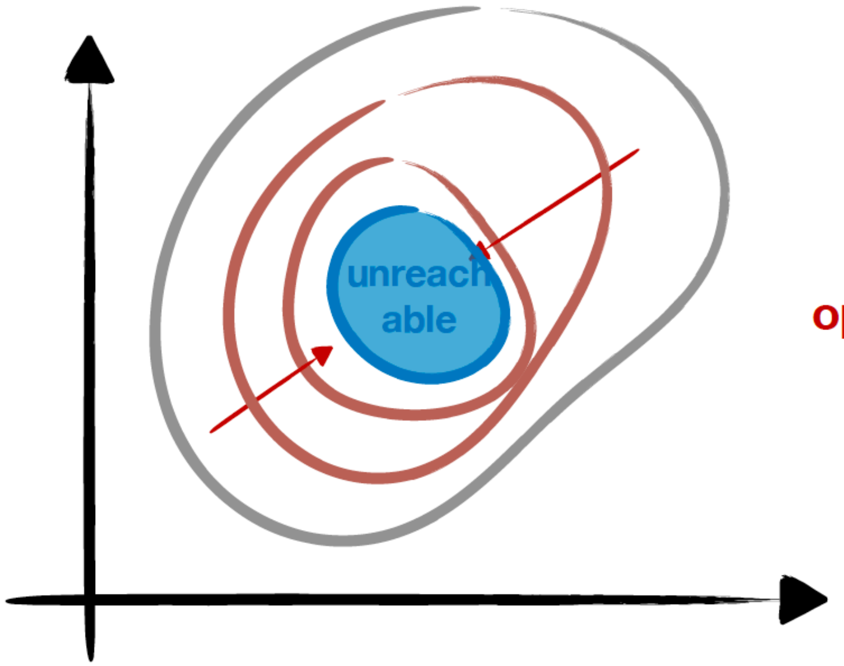
Optimal Reconstruction



Optimal Analysis

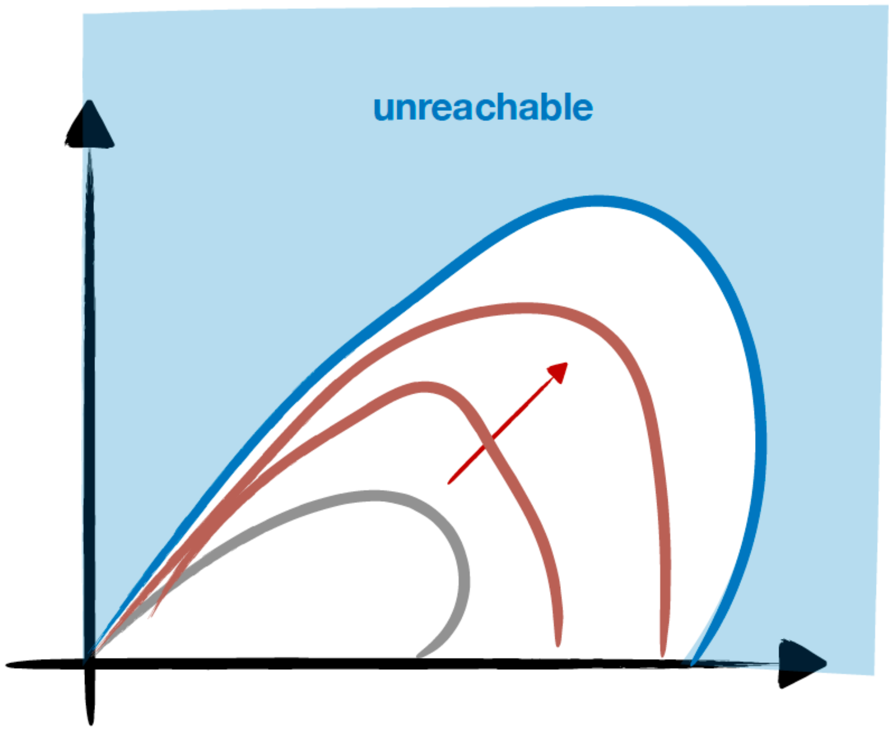
[Lukas Heinrich - Detector design using differential programing]

Natural limit: true posterior p(theory | data)



unoptimized

optimized (e.g. w/ ML)



Measurements
(e.g. Higgs Couplings)

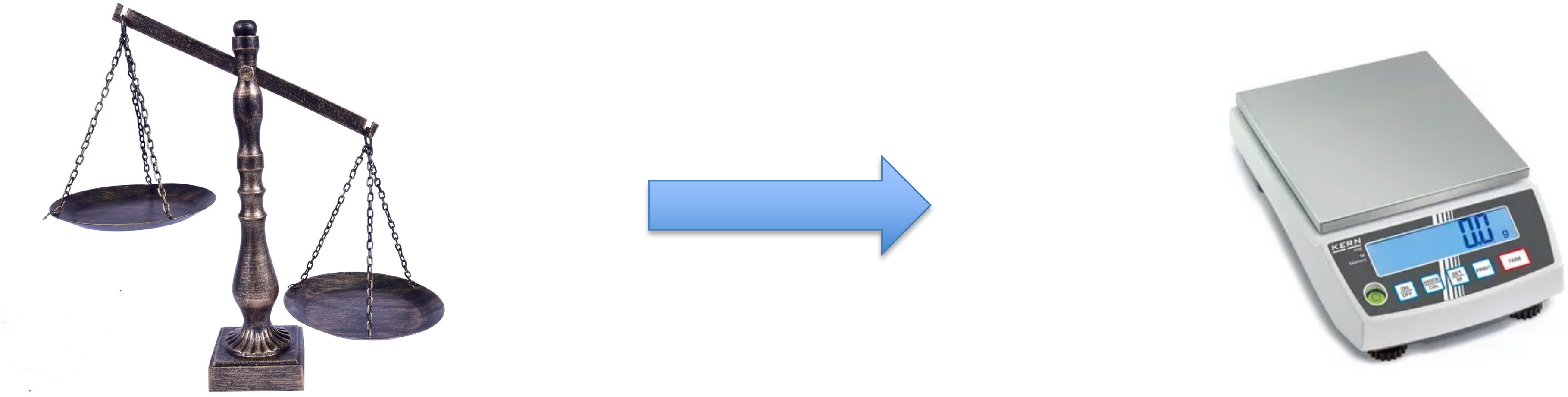
Searches
(e.g. Supersymmetry)

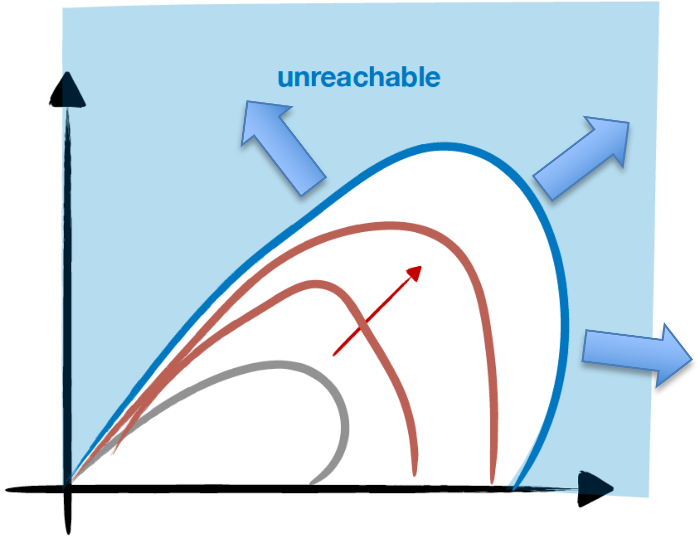
Need better data

[Lukas Heinrich - Detector design using differential programing]

Design new ***optimal*** detector

to optimize *p*(theory | data)





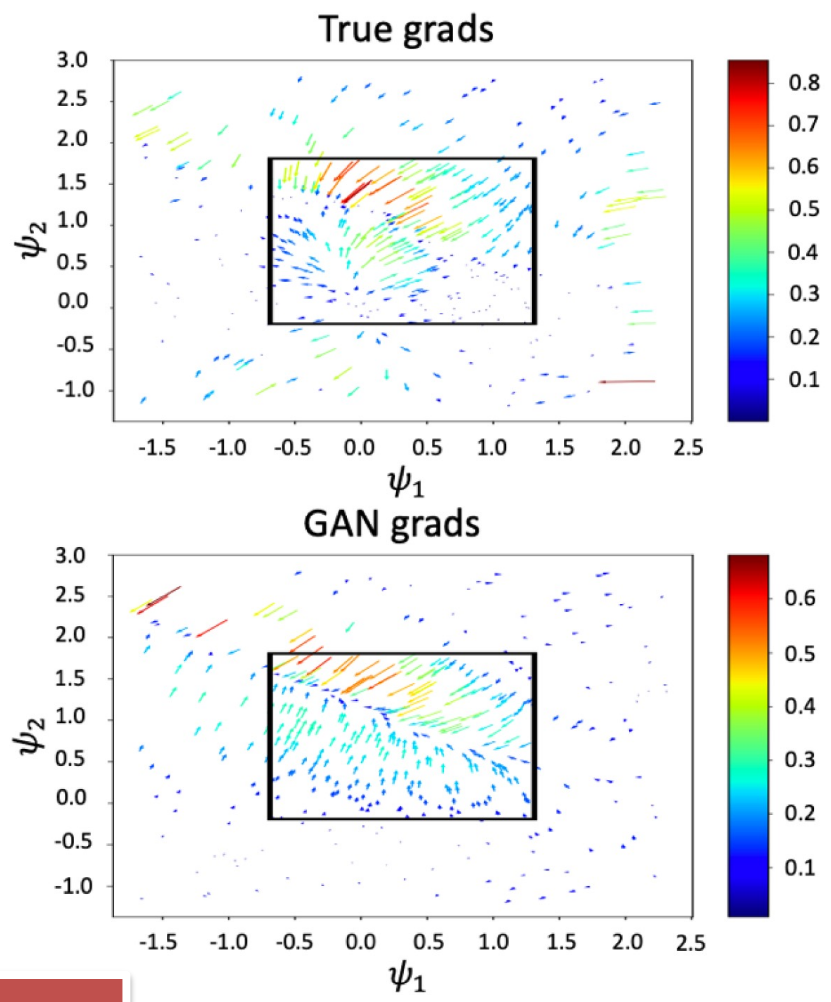
[[Lukas Heinrich - Detector design using differential programing]]

Need **design-conditional** model *p*(x | θ,**D**)

Approximate *p*(x | θ,D) using **generative model**
→ **Fast**
→ **Differentiable**

Challenge:
p(x | **D**) without already exploring all design space **D**

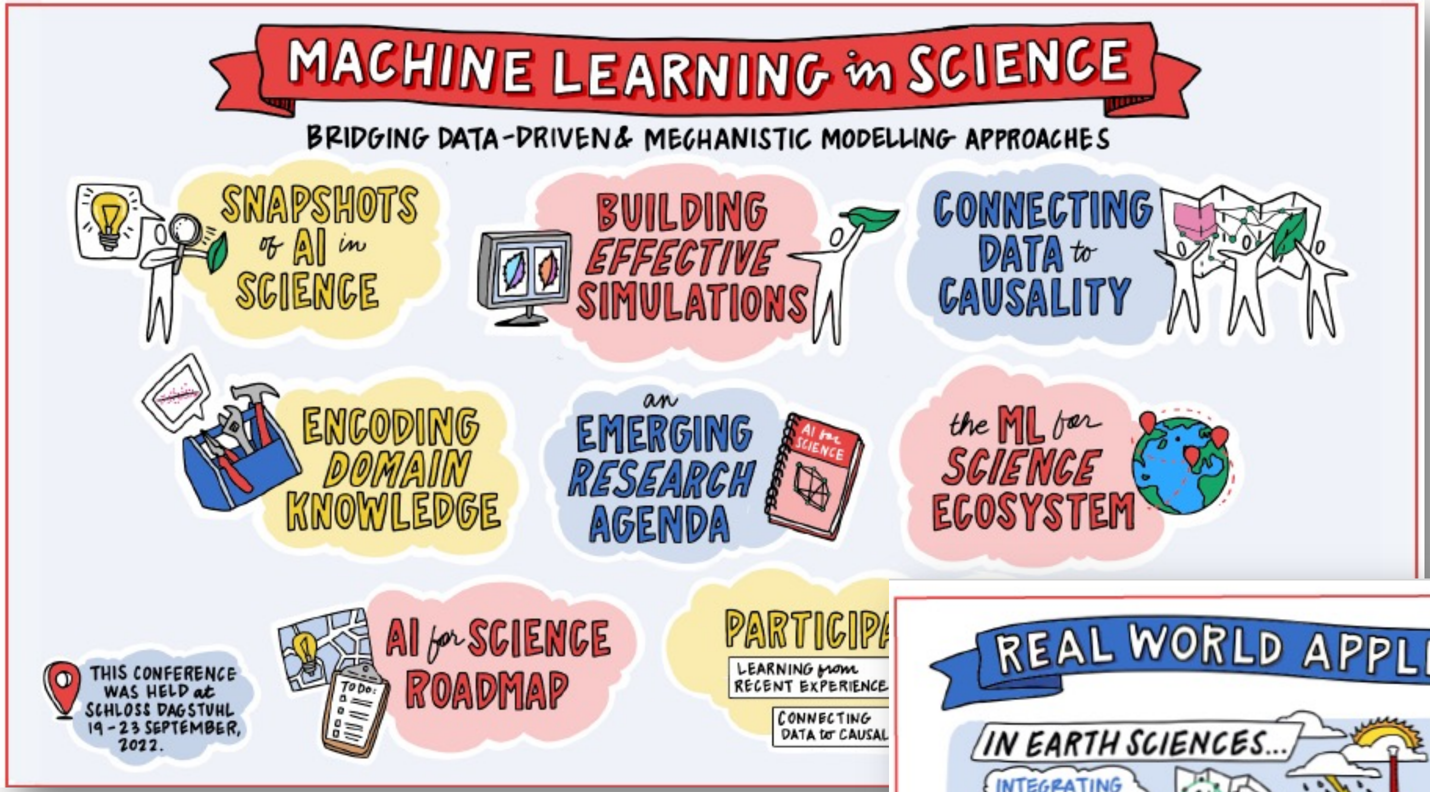
Solution:
train local models as you optimize [2002.04632]

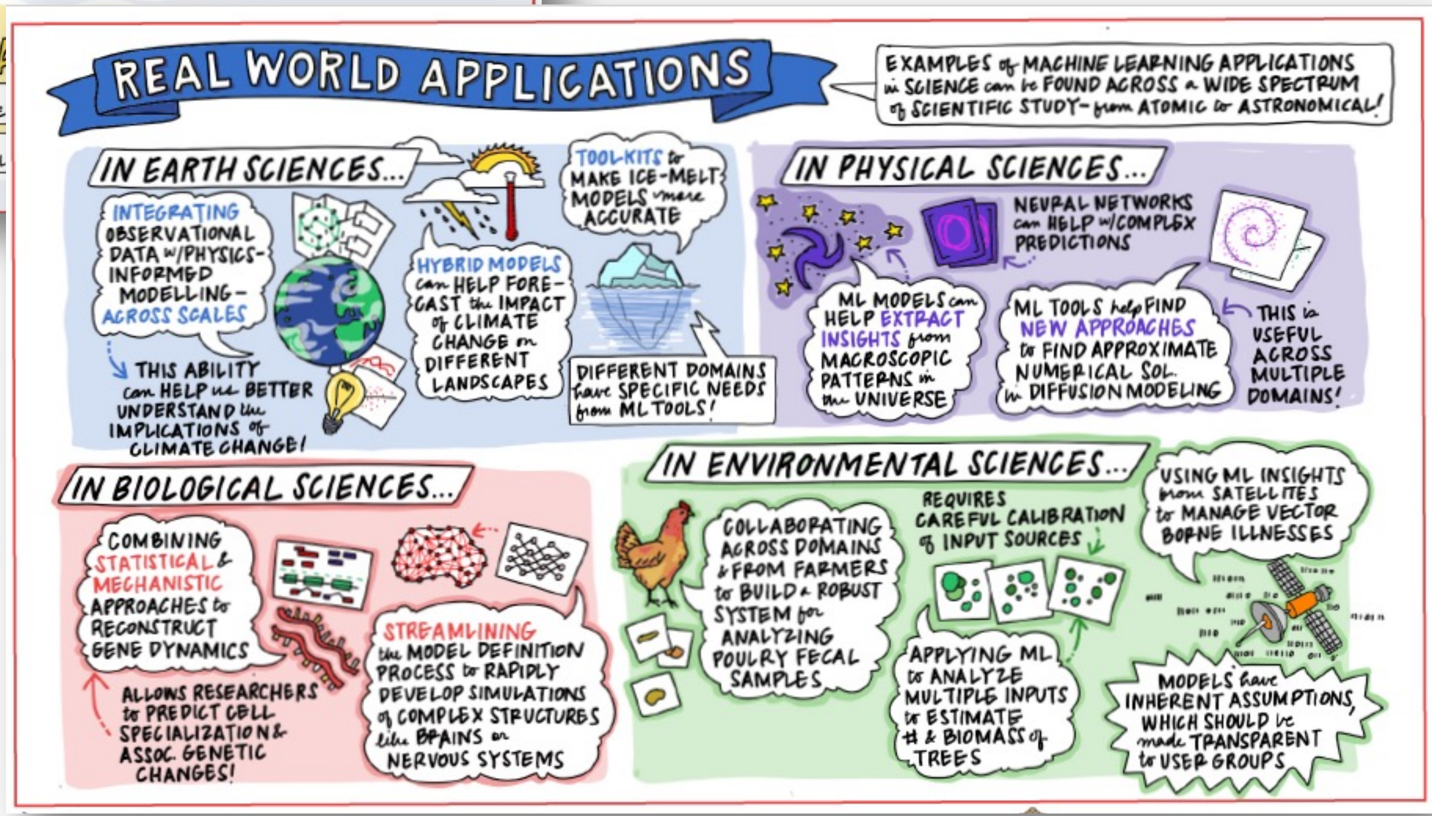


Optimal design = exciting frontier in ML@HEP

COMMUNITY EFFORTS

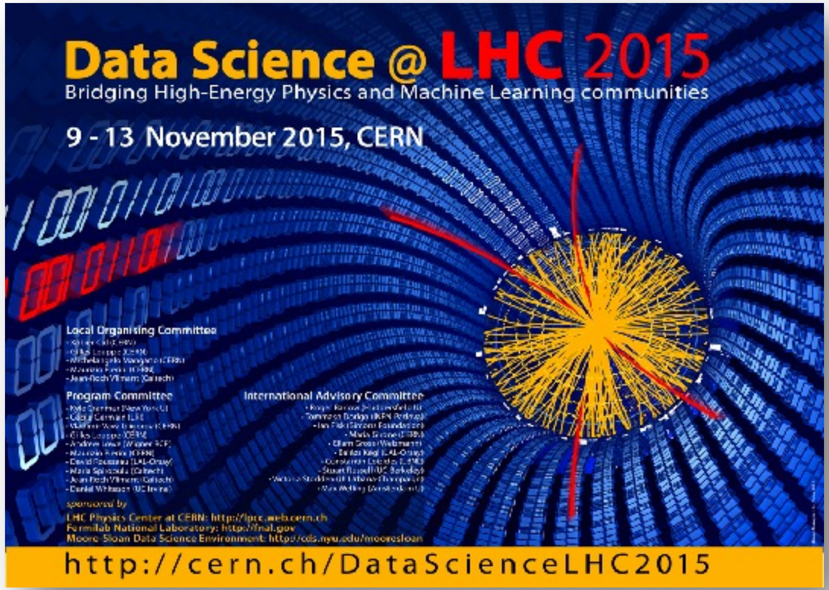
Get organized

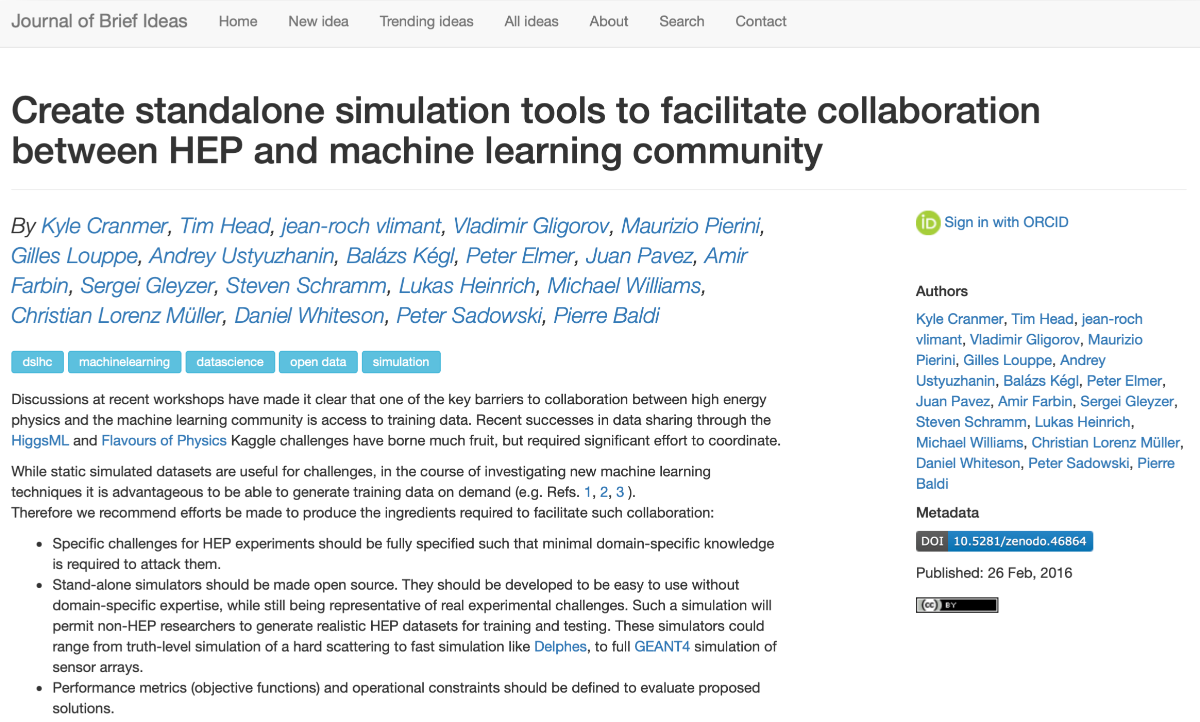

MACHINE LEARNING in SCIENCE
BRIDGING DATA-DRIVEN & MECHANISTIC MODELLING APPROACHES
SNAPSHOTS of AI in SCIENCE
BUILDING EFFECTIVE SIMULATIONS
CONNECTING DATA to CAUSALITY
ENCODING DOMAIN KNOWLEDGE
an EMERGING RESEARCH AGENDA
AI for SCIENCE
the ML for SCIENCE ECOSYSTEM
THIS CONFERENCE WAS HELD at SCHLOSS DAGSTUHL 19-23 SEPTEMBER, 2022.
TO DO:
AI for SCIENCE ROADMAP
PARTICIPA
LEARNING from RECENT EXPERIENCE
CONNECTING DATA to CAUSAL


REAL WORLD APPLICATIONS
EXAMPLES of MACHINE LEARNING APPLICATIONS in SCIENCE can be FOUND ACROSS a WIDE SPECTRUM of SCIENTIFIC STUDY - from ATOMIC to ASTRONOMICAL!
IN EARTH SCIENCES...
INTEGRATING OBSERVATIONAL DATA w/ PHYSICS-INFORMED MODELLING - ACROSS SCALES
THIS ABILITY can HELP us BETTER UNDERSTAND the IMPLICATIONS of CLIMATE CHANGE!
HYBRID MODELS can HELP FORECAST the IMPACT of CLIMATE CHANGE on DIFFERENT LANDSCAPES
TOOL-KITS to MAKE ICE-MELT MODELS more ACCURATE
DIFFERENT DOMAINS have SPECIFIC NEEDS from ML TOOLS!
IN PHYSICAL SCIENCES...
NEURAL NETWORKS can HELP w/ COMPLEX PREDICTIONS
ML MODELS can HELP EXTRACT INSIGHTS from MACROSCOPIC PATTERNS in the UNIVERSE
ML TOOLS help FIND NEW APPROACHES to FIND APPROXIMATE NUMERICAL SOL. in DIFFUSION MODELING
THIS is USEFUL ACROSS MULTIPLE DOMAINS!
IN BIOLOGICAL SCIENCES...
COMBINING STATISTICAL & MECHANISTIC APPROACHES to RECONSTRUCT GENE DYNAMICS
ALLOWS RESEARCHERS to PREDICT CELL SPECIALIZATION & ASSOC. GENETIC CHANGES!
STREAMLINING the MODEL DEFINITION PROCESS to RAPIDLY DEVELOP SIMULATIONS of COMPLEX STRUCTURES like BRAINS or NERVOUS SYSTEMS
IN ENVIRONMENTAL SCIENCES...
COLLABORATING ACROSS DOMAINS & FROM FARMERS to BUILD a ROBUST SYSTEM for ANALYZING POULRY FECAL SAMPLES
REQUIRES CAREFUL CALIBRATION of INPUT SOURCES
APPLYING ML to ANALYZE MULTIPLE INPUTS to ESTIMATE # & BIOMASS of TREES
USING ML INSIGHTS from SATELLITES to MANAGE VECTOR BORNE ILLNESSES
MODELS have INHERENT ASSUMPTIONS, WHICH SHOULD be made TRANSPARENT to USER GROUPS

[Slides]

The HEP-AI ecosystem





Journal of Brief Ideas | Home | New idea | Trending ideas | All ideas | About | Search | Contact

Create standalone simulation tools to facilitate collaboration between HEP and machine learning community

By Kyle Cranmer, Tim Head, jean-roch vlimant, Vladimir Gligorov, Maurizio Pierini, Gilles Louppe, Andrey Ustyuzhanin, Balázs Kégl, Peter Elmer, Juan Pavez, Amir Farbin, Sergei Gleyzer, Steven Schramm, Lukas Heinrich, Michael Williams, Christian Lorenz Müller, Daniel Whiteson, Peter Sadowski, Pierre Baldi

dslhc | machinelearning | datascience | open data | simulation

Discussions at recent workshops have made it clear that one of the key barriers to collaboration between high energy physics and the machine learning community is access to training data. Recent successes in data sharing through the HiggsML and Flavours of Physics Kaggle challenges have borne much fruit, but required significant effort to coordinate.

While static simulated datasets are useful for challenges, in the course of investigating new machine learning techniques it is advantageous to be able to generate training data on demand (e.g. Refs. 1, 2, 3). Therefore we recommend efforts be made to produce the ingredients required to facilitate such collaboration:

- Specific challenges for HEP experiments should be fully specified such that minimal domain-specific knowledge is required to attack them.
- Stand-alone simulators should be made open source. They should be developed to be easy to use without domain-specific expertise, while still being representative of real experimental challenges. Such a simulation will permit non-HEP researchers to generate realistic HEP datasets for training and testing. These simulators could range from truth-level simulation of a hard scattering to fast simulation like Delphes, to full GEANT4 simulation of sensor arrays.
- Performance metrics (objective functions) and operational constraints should be defined to evaluate proposed solutions.

Sign in with ORCID

Authors

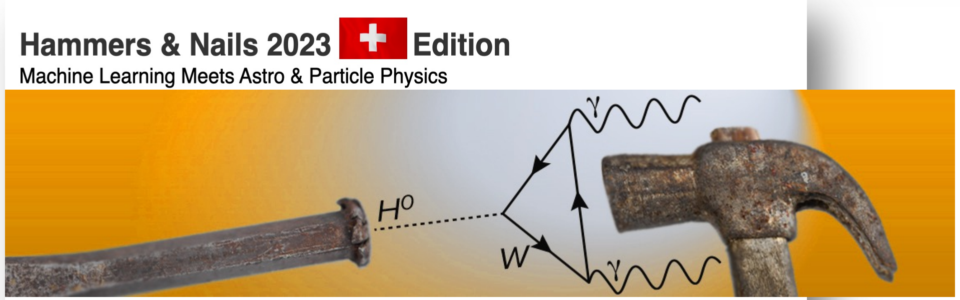
Kyle Cranmer, Tim Head, jean-roch vlimant, Vladimir Gligorov, Maurizio Pierini, Gilles Louppe, Andrey Ustyuzhanin, Balázs Kégl, Peter Elmer, Juan Pavez, Amir Farbin, Sergei Gleyzer, Steven Schramm, Lukas Heinrich, Michael Williams, Christian Lorenz Müller, Daniel Whiteson, Peter Sadowski, Pierre Baldi

Metadata

DOI 10.5281/zenodo.46864

Published: 26 Feb, 2016

EuCAIF



Bottom-up, community consensus, **organized**

You can shape the future of AI in Science

We have a **compelling AI-in-PP story** to tell

- Put PP on the global AI4Science map
- Convince **our** community: AI4Science = future
- Make AI4Science accessible to general public

EuCAIF mission

Facilitate research, provide infrastructure, resources, data, models, connect researchers, define problems & metrics

Topic of interest:

Large-scale foundation models for fundamental physics

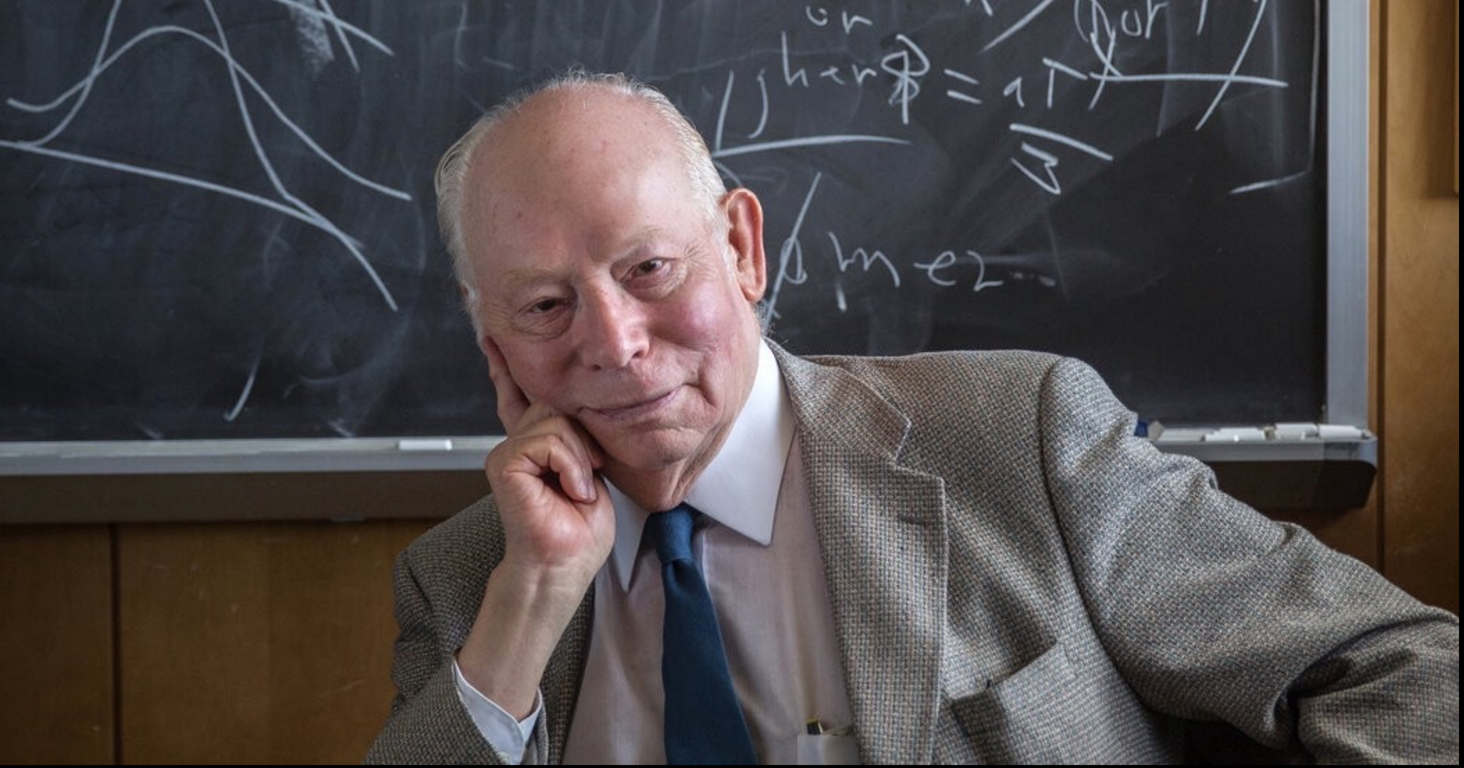
Sign up: https://bit.ly/eucaifcon24-wg1 & provide input

"New directions in science are launched by new tools much more often than by new concepts."

- Freeman Dyson

“Solving intelligence, and then using that to solve everything else.”

- Demis Hassabis, Google DeepMind



"Go for the messes – that's where the action is."

- Steven Weinberg

"Deep Learning today reminiscent of the field of particle physics before the Standard Model: veritable zoo of ~~particles~~ but few unifying principles."

NN architectures

- Michael Bronstein on geometric deep learning (freely quoted)

Concluding remarks

Science evolves

ML is one of our sharpest tools

Formulate open-ended questions

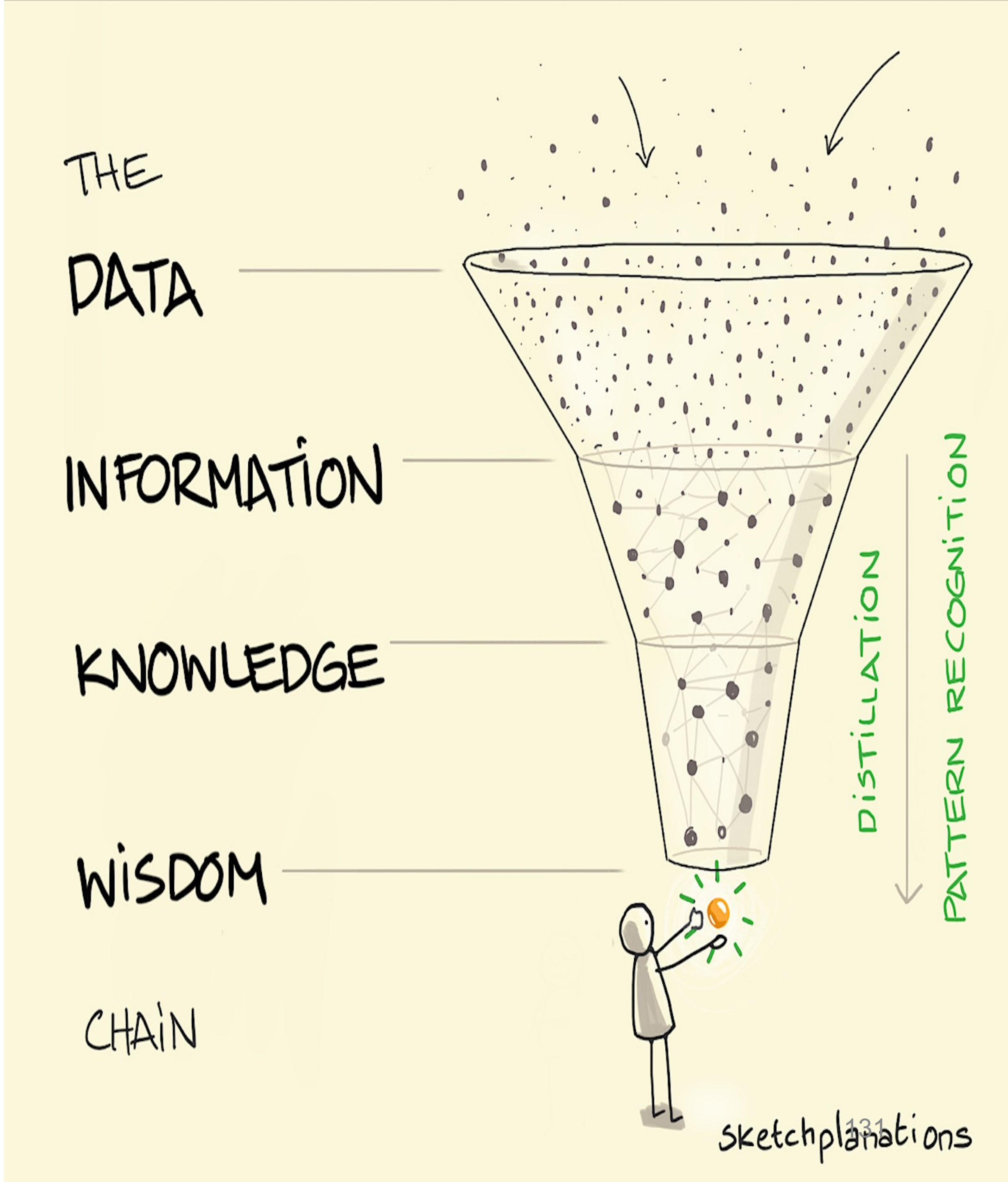
Tackle big goals as a community

Value human resources → automation

Concept → production

AI for scientific discovery

Need all you bright minds !



Thank you !