

An inspirational forward-
looking generative & well-
seasoned perspective on the
denoised adjacent possible

Tobias Golling

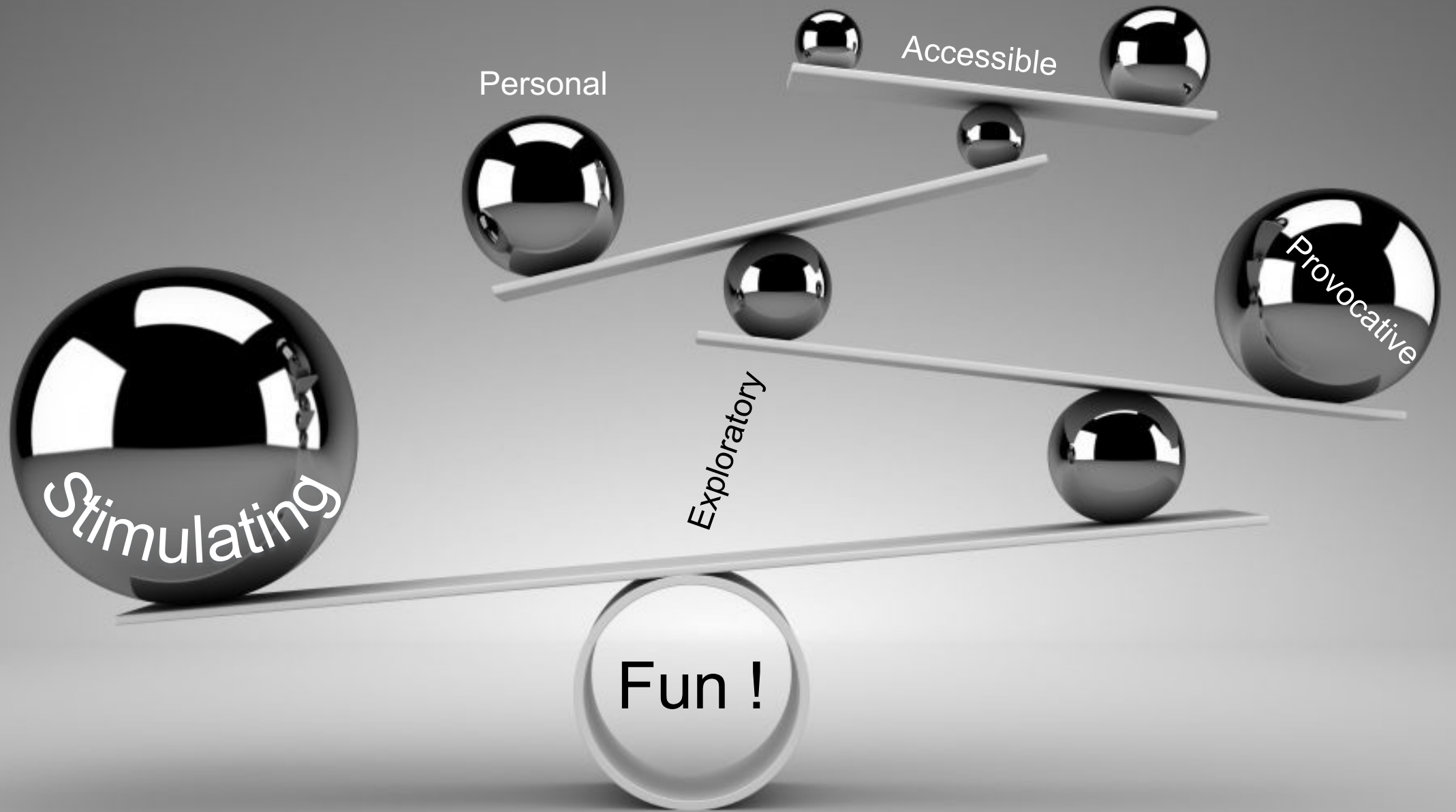


UNIVERSITÉ
DE GENÈVE



Focus on Big Picture

Thanks to organizers who gave me a *blank check*
ML...LHC...unique perspective



Snapshot

Biased

ATLAS-centric

Too long, but incomplete
isn't good
the cattle steal



Outline

- Look back & ahead
- The Challenges
- Gedankenexperimente
- Look to the Data
- The role of Theory
- The Strategy
- The role of Experiment
- Concluding remarks

Pat on the back

A photograph of a vast field of wheat under a bright, hazy sunset sky. The sun is a large, glowing orb in the upper right, casting a golden light over the entire scene. The wheat stalks are in the foreground, some in sharp focus, while others blur into the distance. The overall mood is warm and serene.

ML@HEP = widespread impact

Contributions well beyond our field

A look ahead

Model	ℓ, γ	Jets [†]	E_T^{miss}	$\int \mathcal{L} dt [\text{fb}^{-1}]$	Limit	Reference
ADD $G_{KK} + g/q$	$0 e, \mu$	1-4 j	Yes	36.1	M_D 7.7 TeV	$n = 2$ 1711.03301
ADD non-resonant $\gamma\gamma$	2γ	-	-	36.7	M_S 8.6 TeV	$n = 3$ HLZ NLO 1707.04147
ADD OBH	-	2 j	-	37.0	M_{th} 8.9 TeV	$n = 6$ 1703.09217
ADD BH high $\sum p_T$	$\geq 1 e, \mu$	$\geq 2 j$	-	3.2	M_{th} 8.2 TeV	$n = 6, M_D = 3 \text{ TeV}$, rot BH 1606.02265
ADD BH multijet	-	$\geq 3 j$	-	3.6	M_{th} 9.55 TeV	$n = 6, M_D = 3 \text{ TeV}$, rot BH 1512.02586
ADD $G_{KK} \rightarrow \gamma\gamma$	2γ	-	-	36.7	G_{KK} mass 4.1 TeV	$k/\bar{M}_{Pl} = 0.1$ 1707.04147
ADD $G_{KK} \rightarrow WW/ZZ$	multi-channel	-	-	36.1	G_{KK} mass 2.3 TeV	$k/\bar{M}_{Pl} = 1.0$ CERN-EP-2018-179
ADD RS $g_{KK} \rightarrow tt$	$1 e, \mu$	$\geq 1 b, \geq 1J/2j$	Yes	36.1	g_{KK} mass 3.8 TeV	$\Gamma/m = 15\%$ 1804.10823
ADD / RPP	$1 e, \mu$	$\geq 2 b, \geq 3 j$	Yes	36.1	KK mass 1.8 TeV	Tier (1,1), $\mathcal{B}(A^{(1,1)} \rightarrow tt) = 1$ 1803.09678
BSM $Z' \rightarrow \ell\ell$	$2 e, \mu$	-	-	36.1	Z' mass 4.5 TeV	$\Gamma/m = 1\%$ 1707.02424
BSM $Z' \rightarrow \tau\tau$	2τ	-	-	36.1	Z' mass 2.42 TeV	1709.07242
BSM leptophobic $Z' \rightarrow bb$	-	2 b	-	36.1	Z' mass 2.1 TeV	1805.09299
BSM leptophobic $Z' \rightarrow tt$	$1 e, \mu$	$\geq 1 b, \geq 1J/2j$	Yes	36.1	Z' mass 3.0 TeV	1804.10823
BSM $W' \rightarrow \ell\nu$	$1 e, \mu$	-	Yes	79.8	W' mass 5.6 TeV	ATLAS-CONF-2018-017
BSM $W' \rightarrow \tau\nu$	1τ	-	Yes	36.1	W' mass 3.7 TeV	1801.06992
BSM $V' \rightarrow WV \rightarrow qq\bar{q}\bar{q}$ model B	$0 e, \mu$	2 J	-	79.8	V' mass 4.15 TeV	$g_V = 3$ ATLAS-CONF-2018-016
BSM $V' \rightarrow WH/ZH$ model B	multi-channel	-	-	36.1	V' mass 2.93 TeV	$g_V = 3$ 1712.06518
BSM $W_2' \rightarrow tb$	multi-channel	-	-	36.1	W' mass 3.25 TeV	CERN-EP-2018-142
BSM Λ	$2 e, \mu$	2 j	-	37.0	Λ 21.8 TeV	η_{LL} 1703.09217
BSM Λ	$2 e, \mu$	-	-	36.1	Λ 40.0 TeV	η_{LL} 1707.02424
BSM Λ	$\geq 1 e, \mu$	$\geq 1 b, \geq 1 j$	Yes	36.1	Λ 2.57 TeV	$ C_{4l} = 4\pi$ CERN-EP-2018-174
BSM vector mediator (Dirac DM)	$0 e, \mu$	1-4 j	Yes	36.1	m_{med} 1.55 TeV	$g_q = 0.25, g_l = 1.0, m(\chi) = 1 \text{ GeV}$ 1711.03301
BSM colored scalar mediator (Dirac DM)	$0 e, \mu$	1-4 j	Yes	36.1	m_{med} 1.67 TeV	$g = 1.0, m(\chi) = 1 \text{ GeV}$ 1711.03301
BSM V_{Y1} EFT (Dirac DM)	$0 e, \mu$	1 J, $\leq 1 j$	Yes	3.2	M_* 700 GeV	$m(\chi) < 150 \text{ GeV}$ 1608.02372
BSM LQ 1^{st} gen	$2 e$	$\geq 2 j$	-	3.2	LQ mass 1.1 TeV	$\beta = 1$ 1605.06035
BSM LQ 2^{nd} gen	2μ	$\geq 2 j$	-	3.2	LQ mass 1.05 TeV	$\beta = 1$ 1605.06035
BSM LQ 3^{rd} gen	$1 e, \mu$	$\geq 1 b, \geq 3 j$	Yes	20.3	LQ mass 640 GeV	$\beta = 0$ 1508.04735
BSM $T \rightarrow Ht/Zt/Wb + X$	multi-channel	-	-	36.1	T mass 1.37 TeV	SU(2) doublet ATLAS-CONF-2018-XXX
BSM $B \rightarrow Wt/Zb + X$	multi-channel	-	-	36.1	B mass 1.34 TeV	SU(2) doublet ATLAS-CONF-2018-XXX
BSM $T_{5/3} T_{5/3} T_{5/3} \rightarrow Wt + X$	$2(SS) \geq 3 e, \mu$	$\geq 1 b, \geq 1 j$	Yes	36.1	$T_{5/3}$ mass 1.64 TeV	$\mathcal{B}(T_{5/3} \rightarrow Wt) = 1, c(T_{5/3} Wt) = 1$ CERN-EP-2018-171
BSM $Y \rightarrow Wb + X$	$1 e, \mu$	$\geq 1 b, \geq 1 j$	Yes	3.2	Y mass 1.44 TeV	$\mathcal{B}(Y \rightarrow Wb) = 1, c(YWb) = 1/\sqrt{2}$ ATLAS-CONF-2016-072
BSM $B \rightarrow Hb + X$	$0 e, \mu, 2 \gamma$	$\geq 1 b, \geq 1 j$	Yes	79.8	B mass 1.21 TeV	$\kappa_B = 0.5$ ATLAS-CONF-2018-XXX
BSM $Q \rightarrow Hb + X$	$1 e, \mu$	$\geq 4 j$	Yes	20.3	Q mass 690 GeV	1509.04261
BSM excited quark $q^* \rightarrow qg$	-	2 j	-	37.0	q^* mass 6.0 TeV	only u^* and d^* , $\Lambda = m(q^*)$ 1703.09127
BSM excited quark $q^* \rightarrow q\gamma$	1γ	1 j	-	36.7	q^* mass 5.3 TeV	only u^* and d^* , $\Lambda = m(q^*)$ 1709.10440
BSM excited quark $b^* \rightarrow b\gamma$	-	1 b, 1 j	-	36.1	b^* mass 2.6 TeV	1805.09299
BSM excited lepton ℓ^*	$3 e, \mu$	-	-	20.3	ℓ^* mass 3.0 TeV	$\Lambda = 3.0 \text{ TeV}$ 1411.2921
BSM excited lepton ν^*	$3 e, \mu, \tau$	-	-	20.3	ν^* mass 1.6 TeV	$\Lambda = 1.6 \text{ TeV}$ 1411.2921
BSM Type III Seesaw	$1 e, \mu$	$\geq 2 j$	Yes	79.8	N^0 mass 560 GeV	$m(W_R) = 2.4 \text{ TeV}$, no mixing ATLAS-CONF-2018-020
BSM Majorana ν	$2 e, \mu$	2 j	-	20.3	N^0 mass 2.0 TeV	1506.06020
BSM Higgs triplet $H^{\pm\pm} \rightarrow \ell\ell$	$2, 3, 4 e, \mu$ (SS)	-	-	36.1	$H^{\pm\pm}$ mass 870 GeV	DY production 1710.09748
BSM Higgs triplet $H^{\pm\pm} \rightarrow \ell\tau$	$3 e, \mu, \tau$	-	-	20.3	$H^{\pm\pm}$ mass 400 GeV	DY production, $\mathcal{B}(H_L^{\pm\pm} \rightarrow \ell\tau) = 1$ 1411.2921
BSM Monopole (non-res prod)	$1 e, \mu$	1 b	Yes	20.3	spin-1 invisible particle mass 657 GeV	$a_{\text{non-res}} = 0.2$ 1410.5404
BSM Multi-charged particles	-	-	-	20.3	multi-charged particle mass 785 GeV	DY production, $ q = 5e$ 1504.04188
BSM Magnetic monopoles	-	-	-	7.0	monopole mass 1.34 TeV	DY production, $ g = 1g_D$, spin 1/2 1509.08059

$\sqrt{s} = 8 \text{ TeV}$ $\sqrt{s} = 13 \text{ TeV}$

Mass scale [TeV]

* Selection of the available mass limits on new states or phenomena is shown.
† Large-radius jets are denoted by the letter j (J).

Main purpose of LHC: discover BSM

Hypothesis: subtle signatures

Obligation: no stone unturned

New search portfolio
is **diverse &**
powered by **ML**

Concrete challenges



Computing (simulation, on-the-edge,...)



Modeling (pileup, reco,...)



Human resources*

ATLAS + CMS = $O(1000)$ search papers

~2 years per analysis

Average of 4 people

8000 person years



Let's unpack the challenges

Core Challenge: vast signature space **unexplored**

	e	μ	τ	q/g	b	t	γ	Z/W	H	BSM \rightarrow SM ₁ \times SM ₁				BSM \rightarrow SM ₁ \times SM ₂			BSM \rightarrow complex			
										q/g	γ/π^0 's	b	...	tZ/H	bH	...	$\tau qq'$	eqq'	$\mu qq'$...
e	[37,38]	[39,40]	[39]	\emptyset	\emptyset	\emptyset	[41]	[42]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	[43,44]	\emptyset	
μ		[37,38]	[39]	\emptyset	\emptyset	\emptyset	[41]	[42]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	[43,44]	
τ			[45,46]	\emptyset	[47]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	[48,49]	\emptyset	
q/g				[29,30,50,51]	[52]	\emptyset	[53,54]	[55]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	
b					[29,52,56]	[57]	[54]	[58]	[59]	\emptyset	\emptyset	\emptyset	\emptyset	[60]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	
t						[61]	\emptyset	[62]	[63]	\emptyset	\emptyset	\emptyset	\emptyset	[64]	[60]	\emptyset	\emptyset	\emptyset	\emptyset	
γ							[65,66]	[67-69]	[68,70]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	
Z/W								[71]	[71]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	
H									[72,73]	[74]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	
BSM \rightarrow SM ₁ \times SM ₁	q/g									\emptyset	\emptyset	\emptyset		\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	
	γ/π^0 's										[75]	\emptyset		\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	
	b											[76,77]		\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	
	\vdots													\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	
\vdots																				

Multi-body final states, unconventional signatures... ⚡



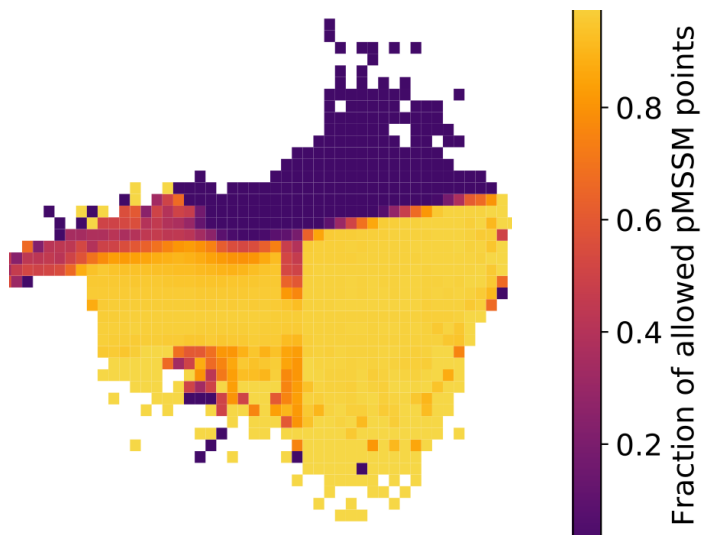
Two complementary approaches

Theory at face value

e.g. MSSM (105) \rightarrow pMSSM (19)

Strong physics bias

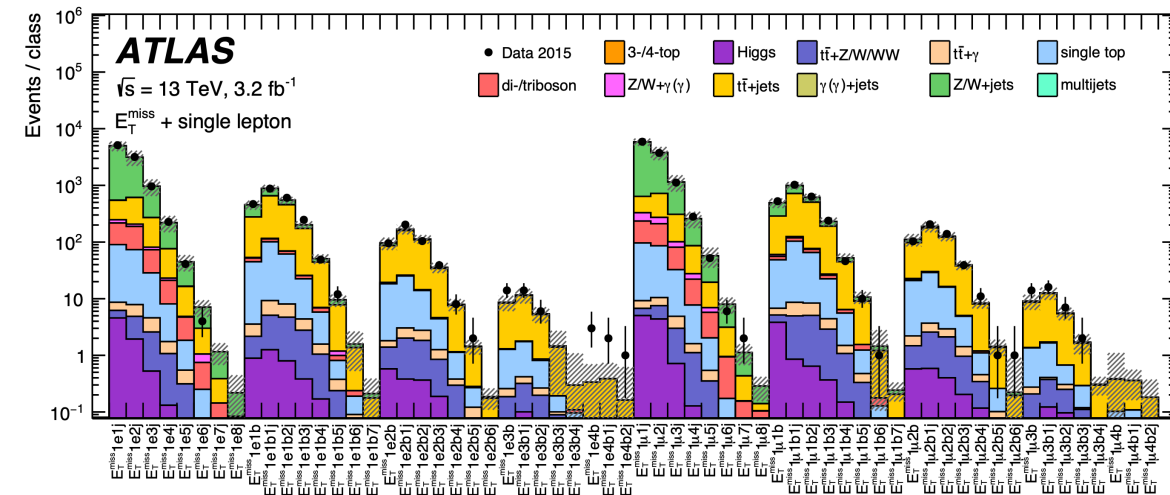
Smart sampling



[2207.05103]

Model-agnostic signature-based

e.g. [1807.07447]



10^5 signal region

Limitations

MC mismodeling

Only cut and count

Look elsewhere effect

Pros & Cons

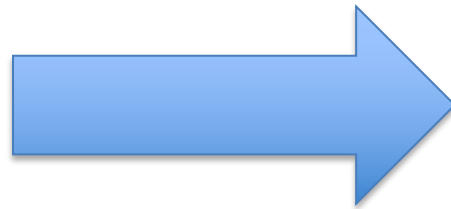
Theory guided: forward modeling

Interpret 

Incomplete signatures space 



[Courtesy of Hitoshi Murayama]



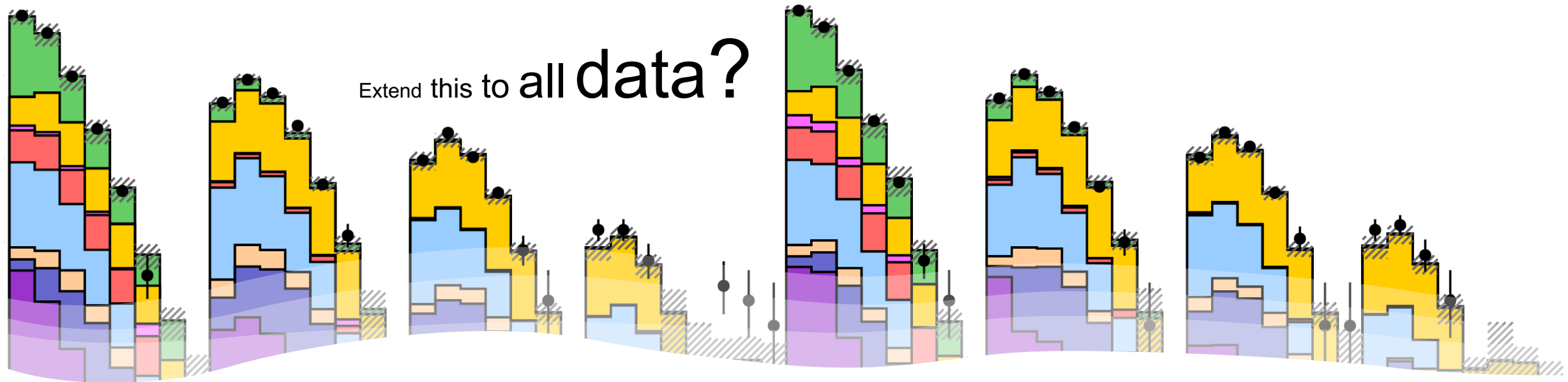
	e	μ	τ	q/g	b	t	γ	Z/W	H	BSM \rightarrow SM ₁ \times SM ₁			BSM \rightarrow SM ₁ \times SM ₂			BSM \rightarrow complex		
										q/g	γ/π^0 's	b	...	tZ/H	bH	...	$\tau q q'$	$e q q'$
e	[37,38]	[39,40]	[39]	\emptyset	\emptyset	\emptyset	[41]	[42]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	[43,44]	\emptyset
μ		[37,38]	[39]	\emptyset	\emptyset	\emptyset	[41]	[42]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	[43,44]
τ			[45,46]	\emptyset	[47]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	[48,49]	\emptyset
q/g				[29,30,50,51]	[52]	\emptyset	[53,54]	[55]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
b					[29,52,56]	[57]	[54]	[58]	[59]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
t						[61]	\emptyset	[62]	[63]	\emptyset	\emptyset	\emptyset	\emptyset	[64]	[60]	\emptyset	\emptyset	\emptyset
γ							[65,66]	[67-69]	[68,70]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
Z/W								[71]	[71]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
H									[72,73]	[74]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
BSM \rightarrow SM ₁ \times SM ₁	q/g									\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
	γ/π^0 's	[75]								\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
	b										[76,77]	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
	...											\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
	...											\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset

Model-agnostic:

Interpret 

Scan *all* signature space \rightarrow *data-driven* 

What if we had perfect modeling



Data slicing?

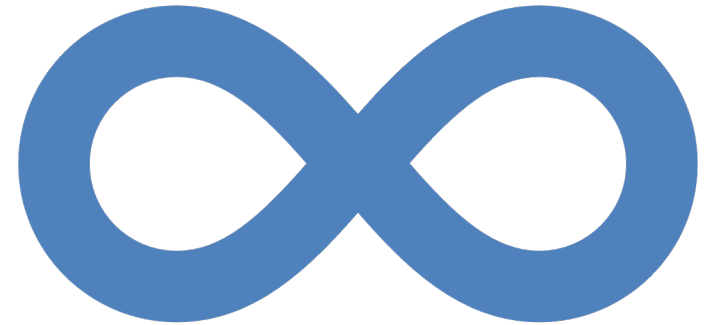
Does this scale?

Look elsewhere

→ Data-driven

What if we had infinite
compute & storage

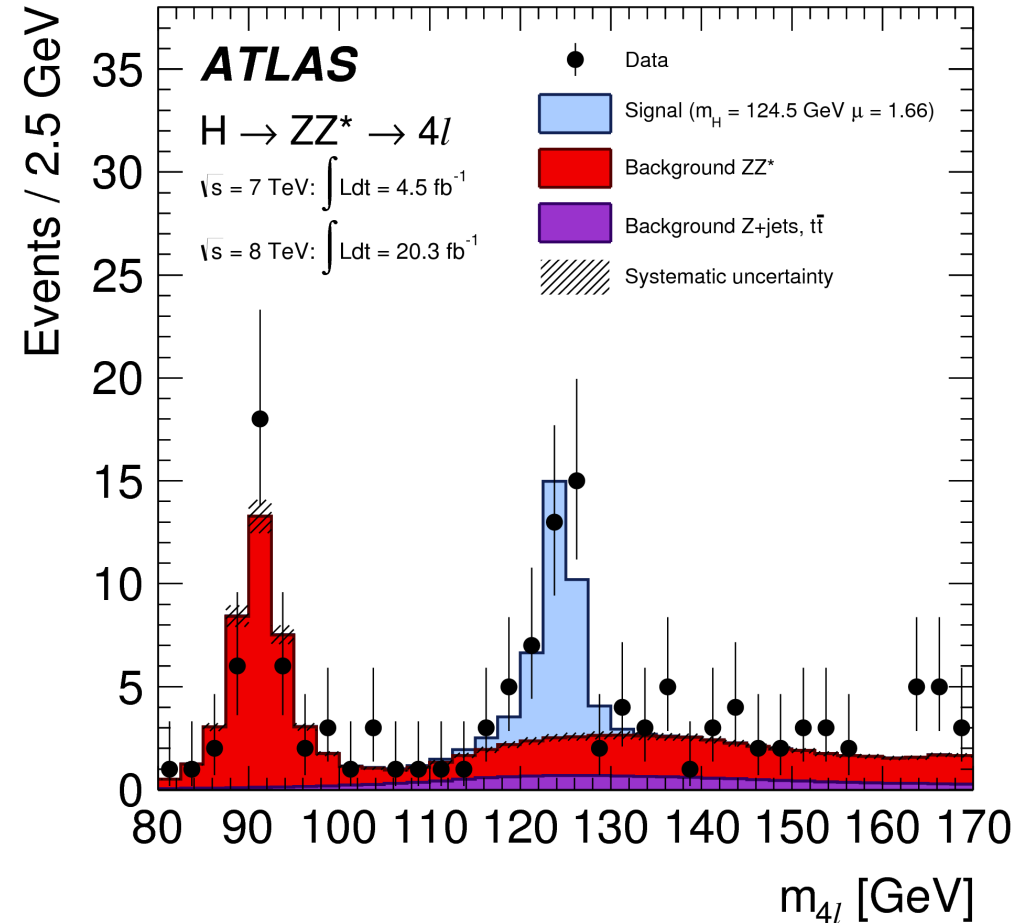
The human factor



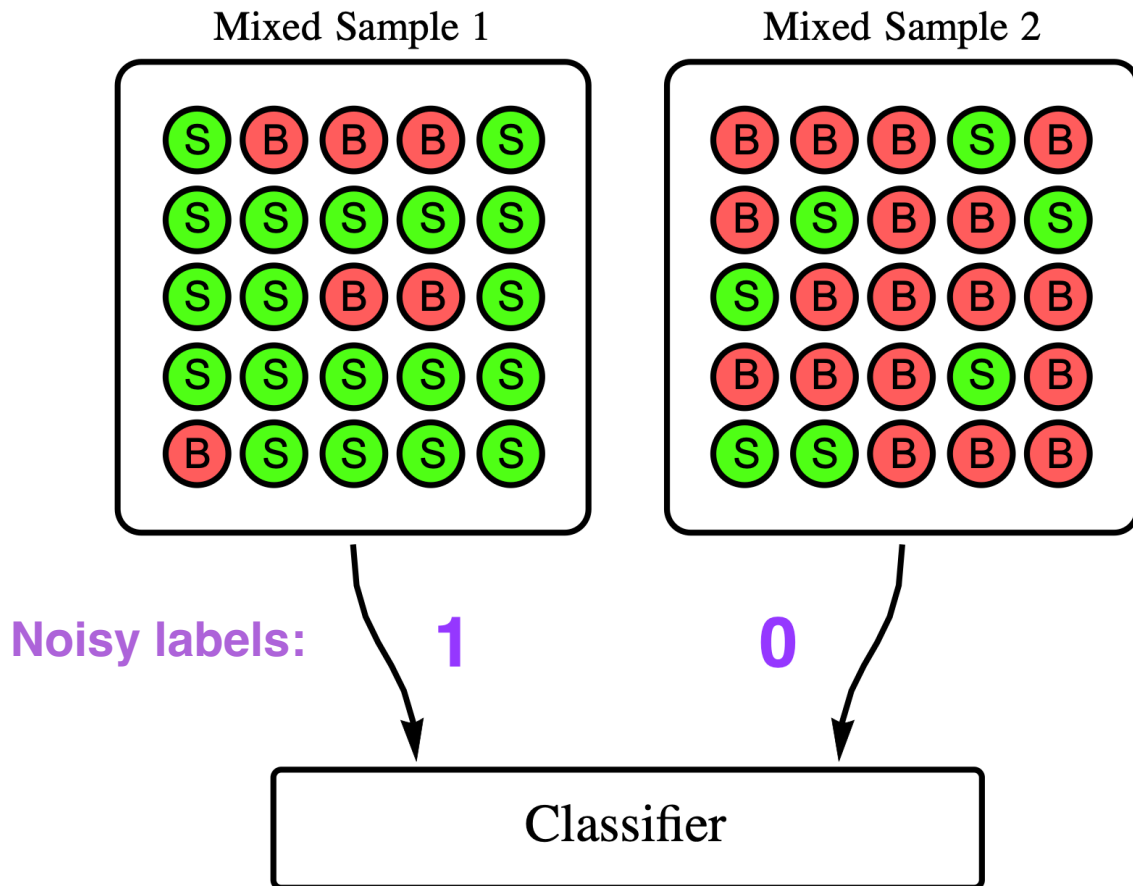
A word on *labels*

No notion of *event label*

Only *probability* to be S or BG



Classification without labeling (CWoLa)



Feature space = prior

Bump hunt [[1902.02634](#)]

ATLAS analysis [[2005.02983](#)]

Beyond resonances

e.g. symmetries [[2203.07529](#)]

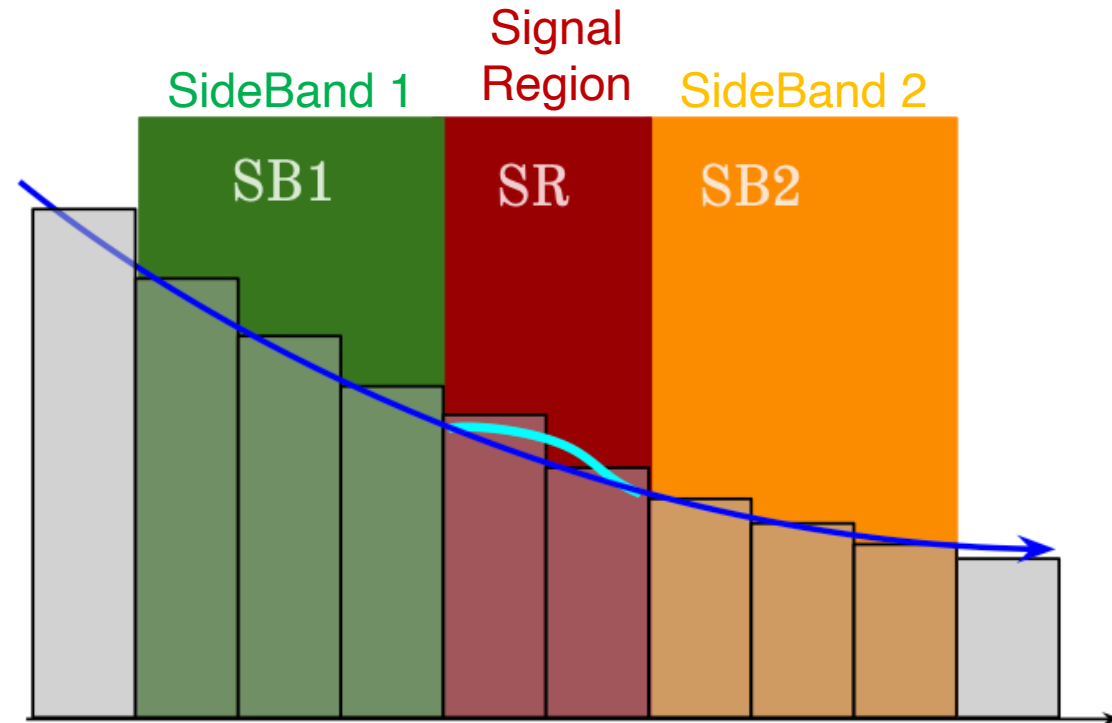
[[1702.00414](#), [1708.02949](#)]

The CWoLa spirit: *all in data*

Main challenge:
calibration

Background calibration in bump hunt

Interpolate BG *template*
from SB's to SR



Calibration methods

Matrix method: no signal enhancement possible ⚡

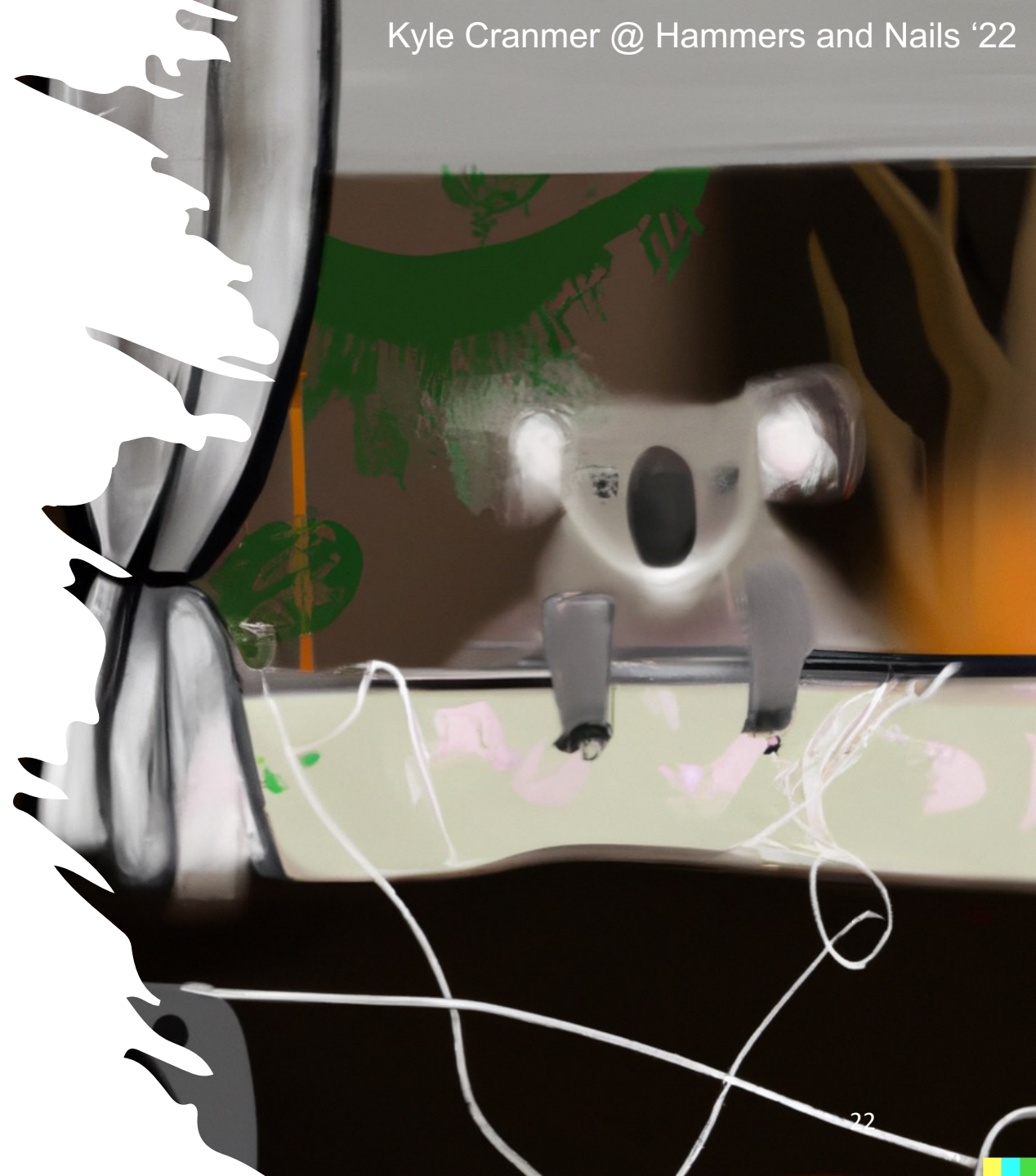
Traditional interpolation: high feature space ⚡

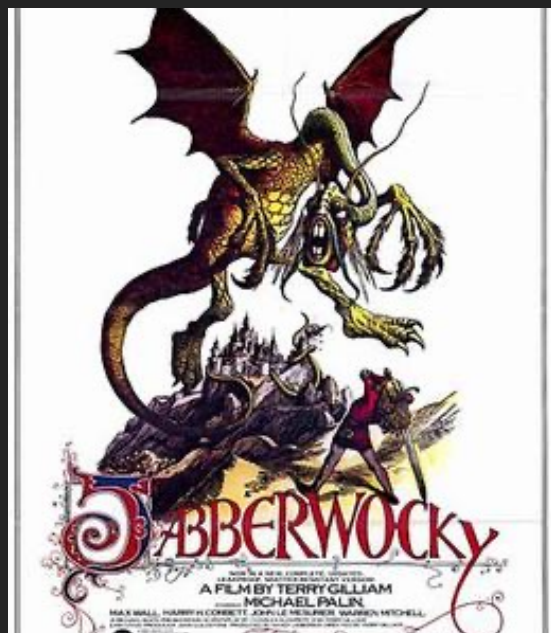
Density estimates: ANODE, [CATHODE](#), [LaCATHODE](#)*

Transform BG template: [CURTAINS](#) & update*

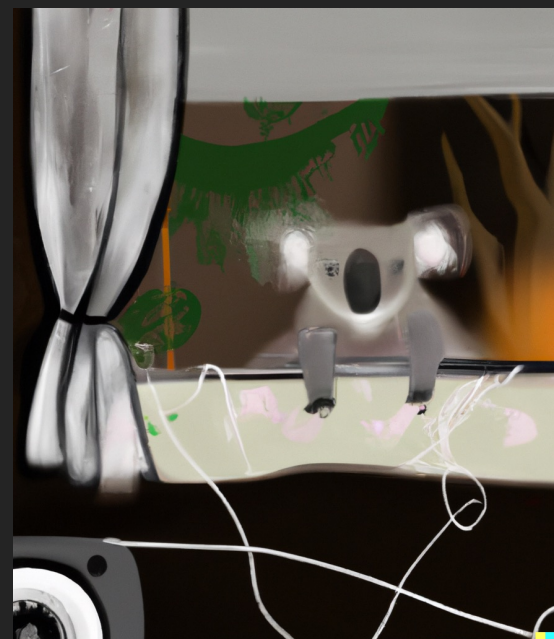
Hybrid BG in MC morphed to data: [SALAD](#) & variation*

La
CWo-CUR-CATH-LAD





≠



Mix & match

Room to optimize & extend

Complementarity → Robustness

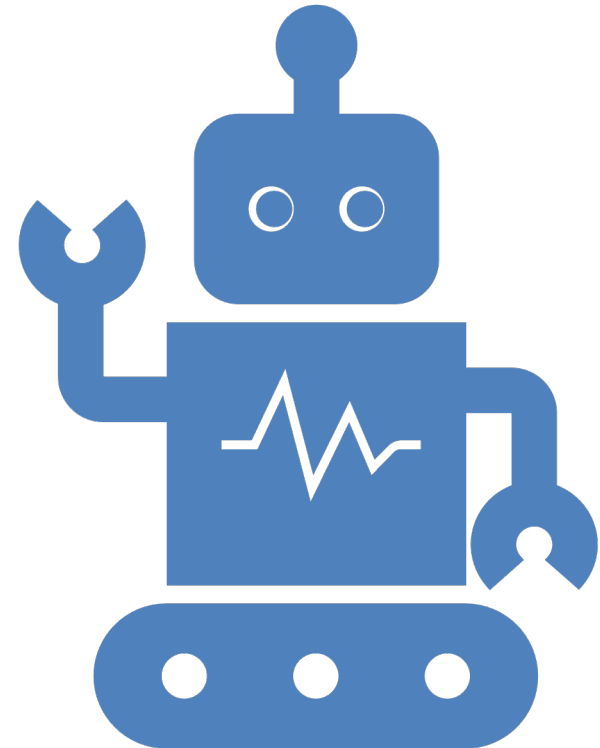


The edge of data-derived

Ticks all boxes: “**compute, modeling, human**”

Simplified systematics treatment

Automation



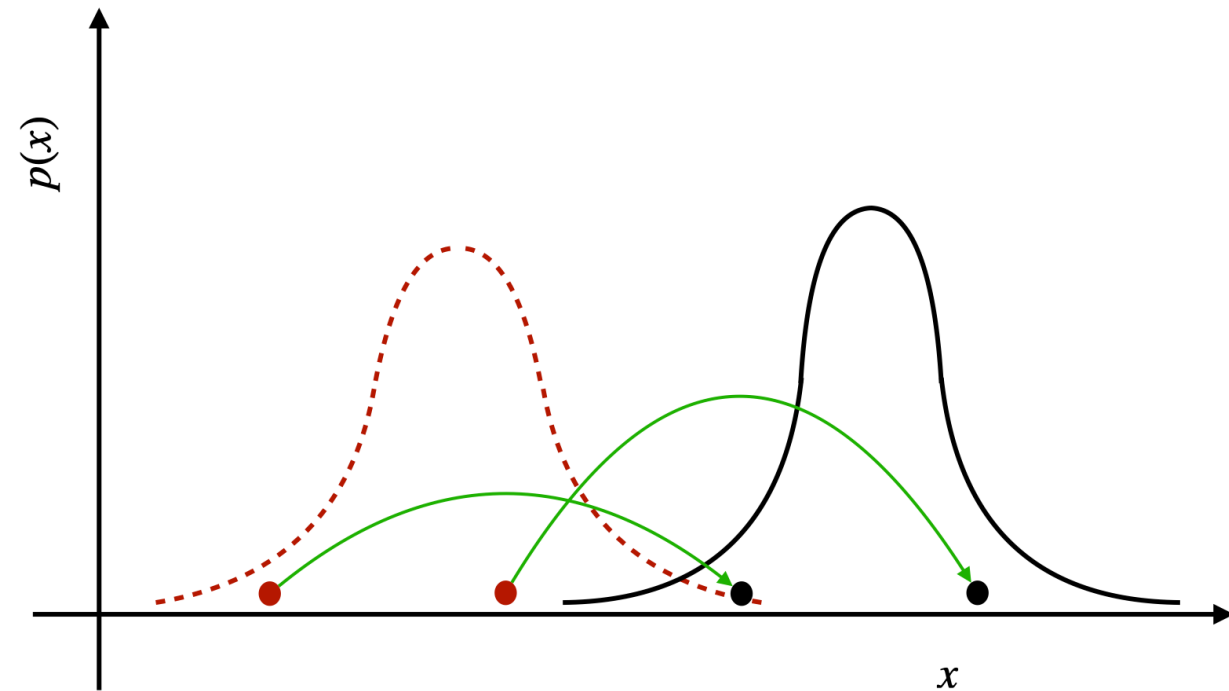
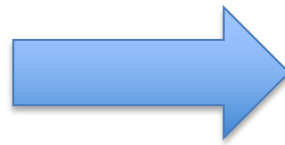
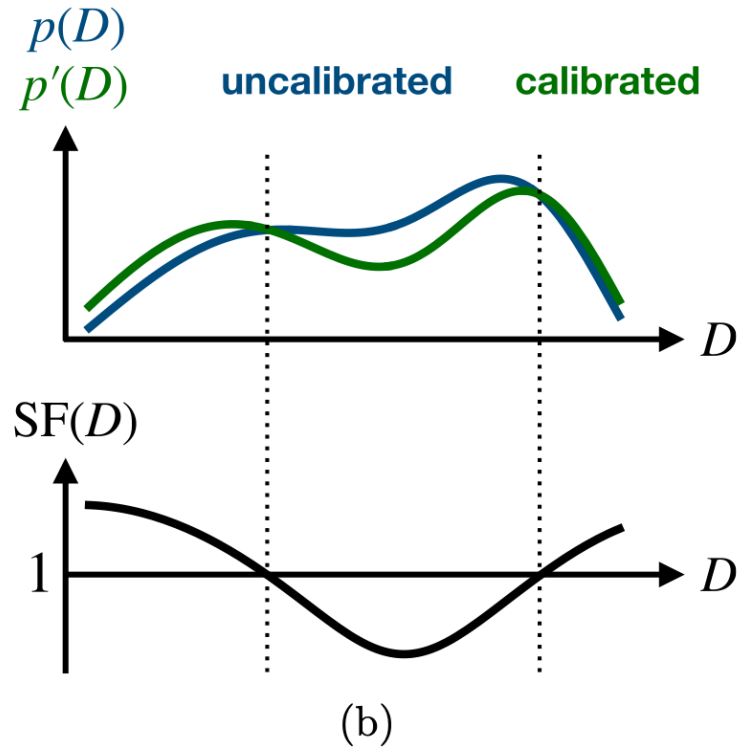


The next level of MC corrections

Feed smart corrections back to our
theory models

Transport your problems away

Calibrating stochastic simulations with optimal transport



The 5th paradigm of scientific discovery

- Simulators: costly, non-differentiable
- (*Local/modular*) surrogate simulators*
 - Fast & differentiable
 - Incorporate inductive bias
 - Anomaly detection, optimisation, design, inference,...

Looks like a

point cloud

GNN
Transformers
Foundation models: backbone*

[PFN](#)

[ParticleNet](#)

[Particle Transformer](#)

[ATL-PHYS-PUB-2022-039](#)

[ATL-PHYS-PUB-2022-040](#)

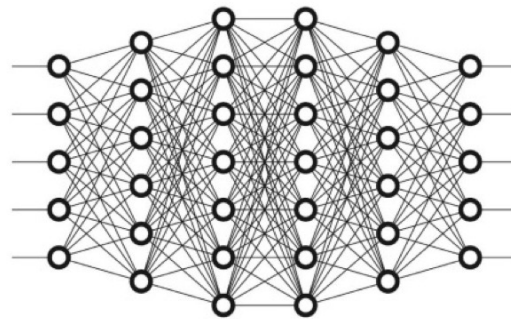
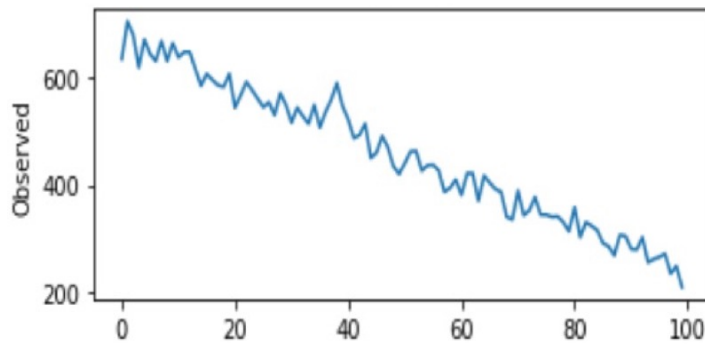
*Lukas Heinrich, H&N's '22

Run: 427394
Event: 3038977 29
2022-07-05 17:02:31 CEST

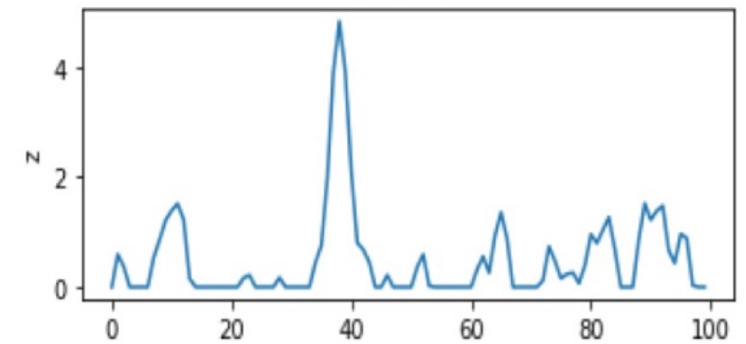
Variation on *typical* bump hunt

Regression:

Mass distribution

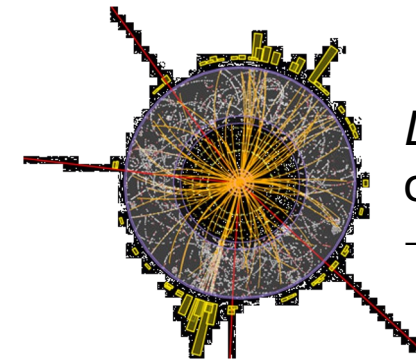
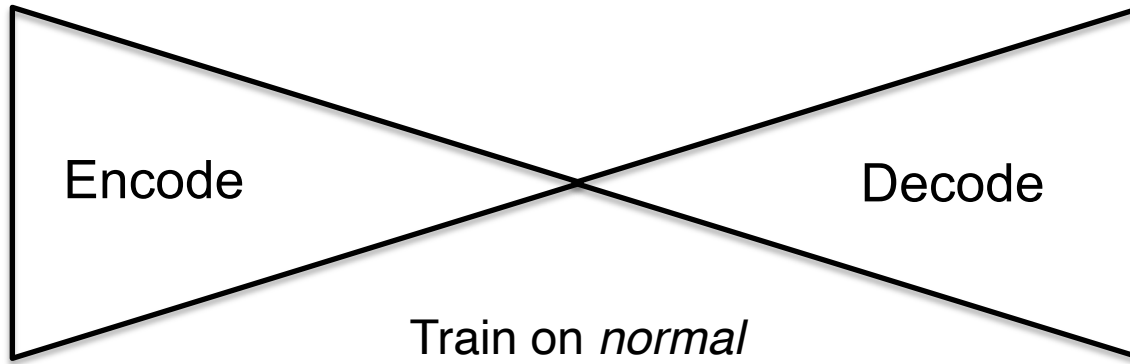
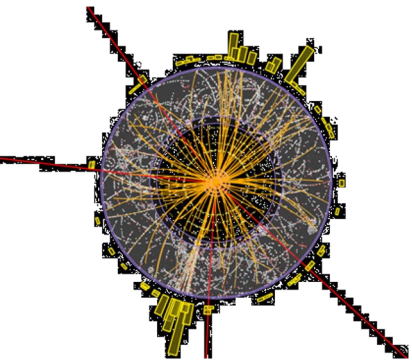


p-value distribution



Challenges: training, validation, out of domain etc.

Fabulous idea: VAE-based outlier detection



Distance
original – reconstructed
→ anomaly metric

Is it working?

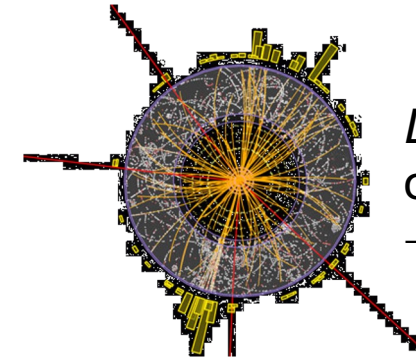
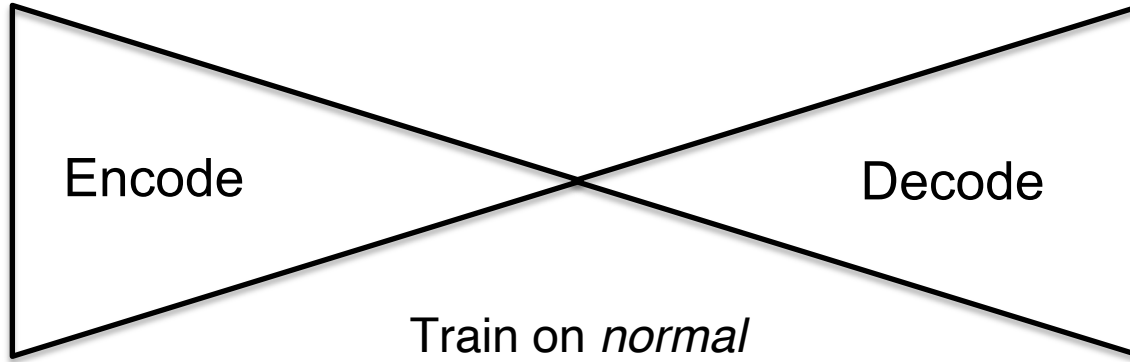
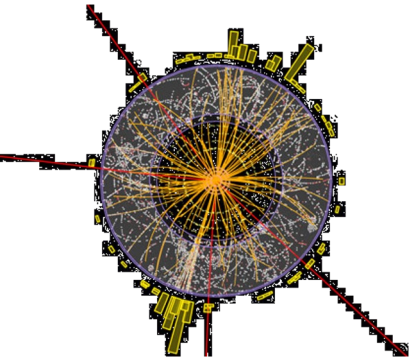
HOW DEEP IS

THE LEARNING PIT



[many references... extend to trigger]

Fabulous idea: VAE-based outlier detection



Distance
original – reconstructed
→ anomaly metric

Is it working?

HOW DEEP IS

THE LEARNING PIT



[NAE]

[many references... extend to trigger]

MODEL:
JAMES NOTTINGHAM
challenginglearning.com

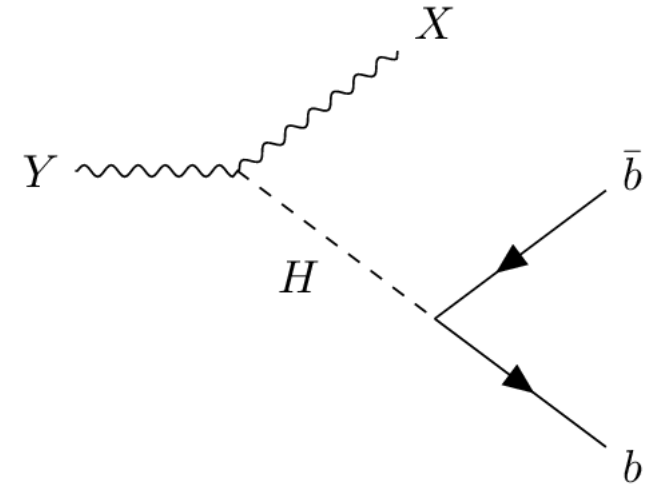
sketchplanations

The *add-on* anomaly search

For any standard search

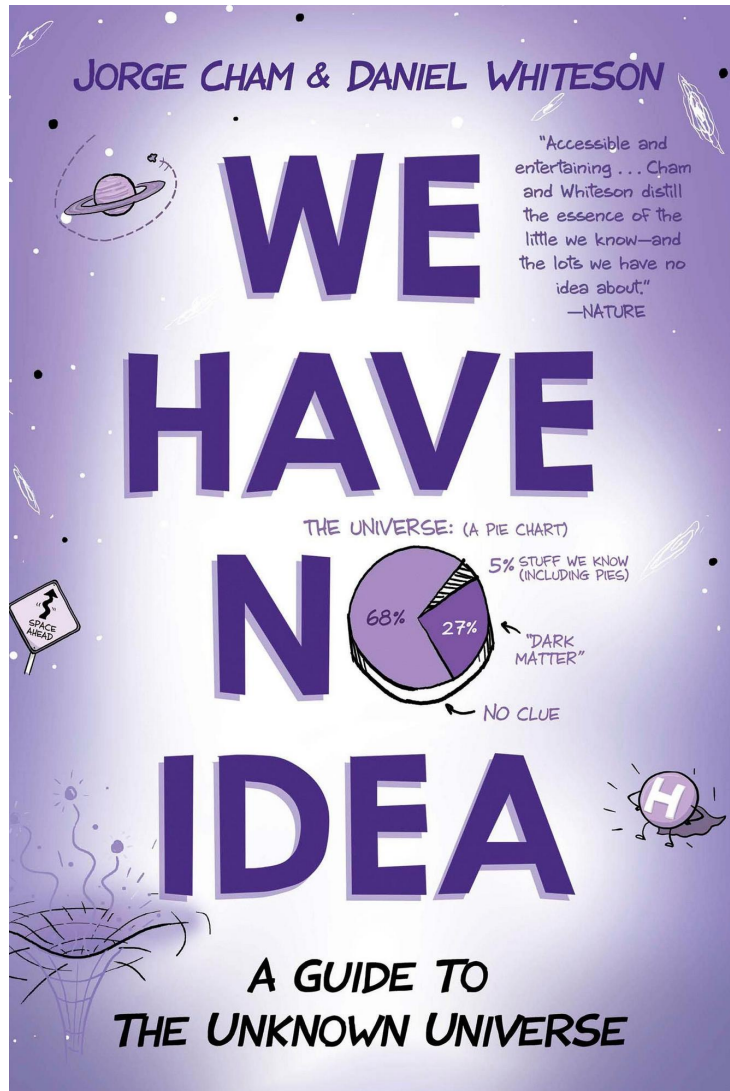
Add *anomaly* signal region*

Comes *for free*



[[ATLAS-CONF-2022-045](#)]

Imagine we abandon theory guidance

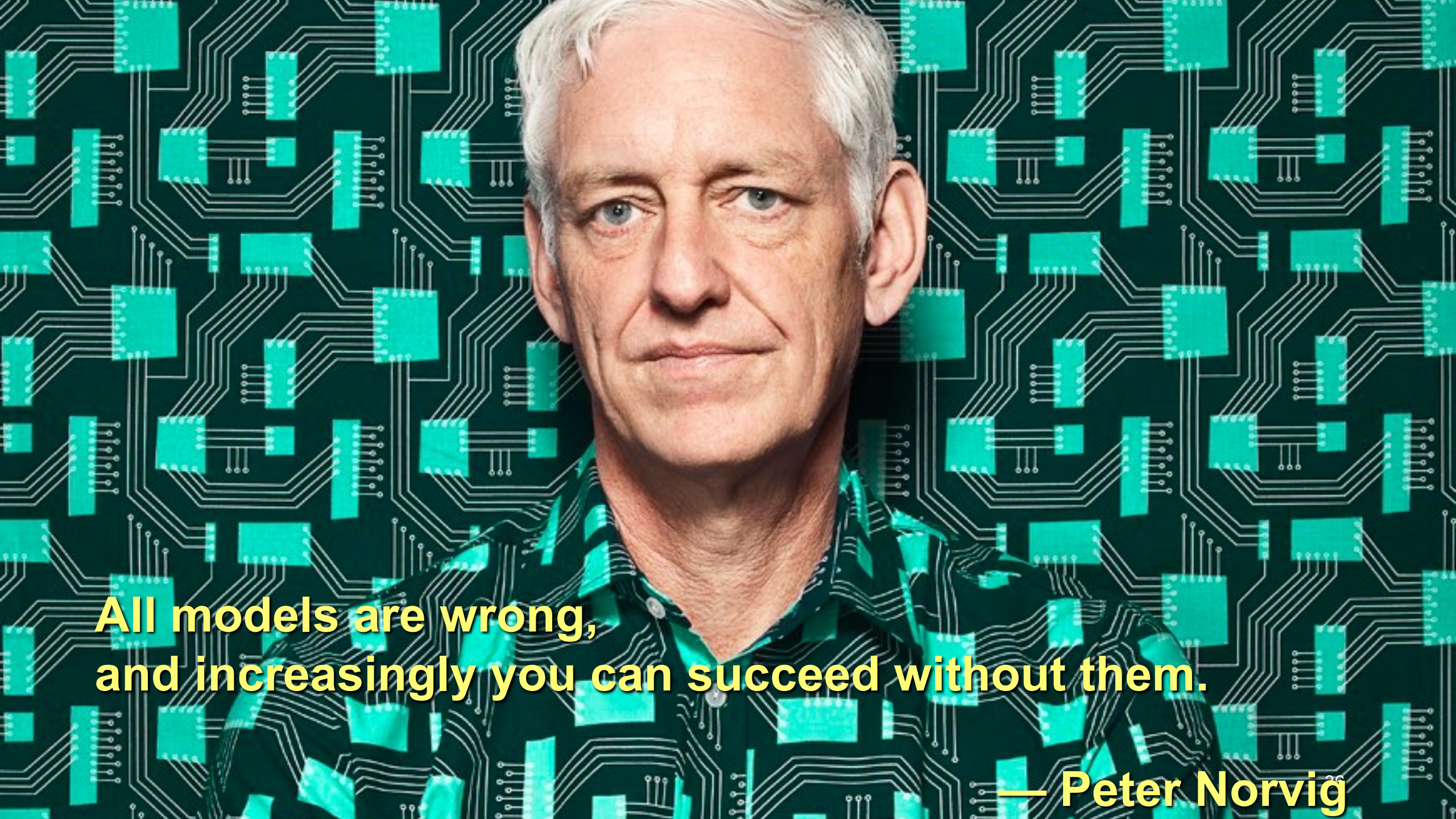


Instead we have an AI oracle

Would we be satisfied?

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

Philosophy of Science

What is scientific understanding?



All models are wrong, but some are are useful.

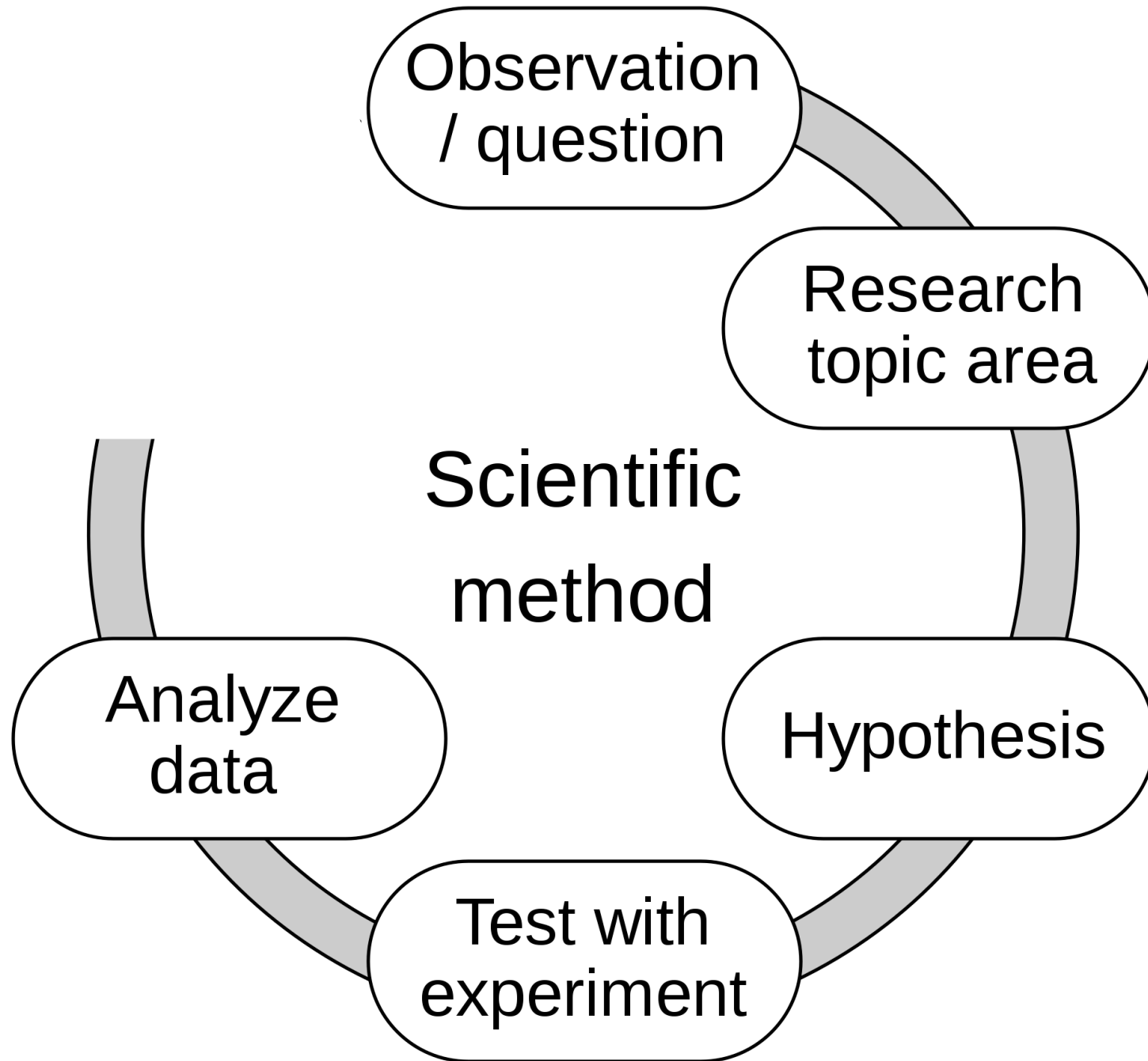
– GEORGE BOX

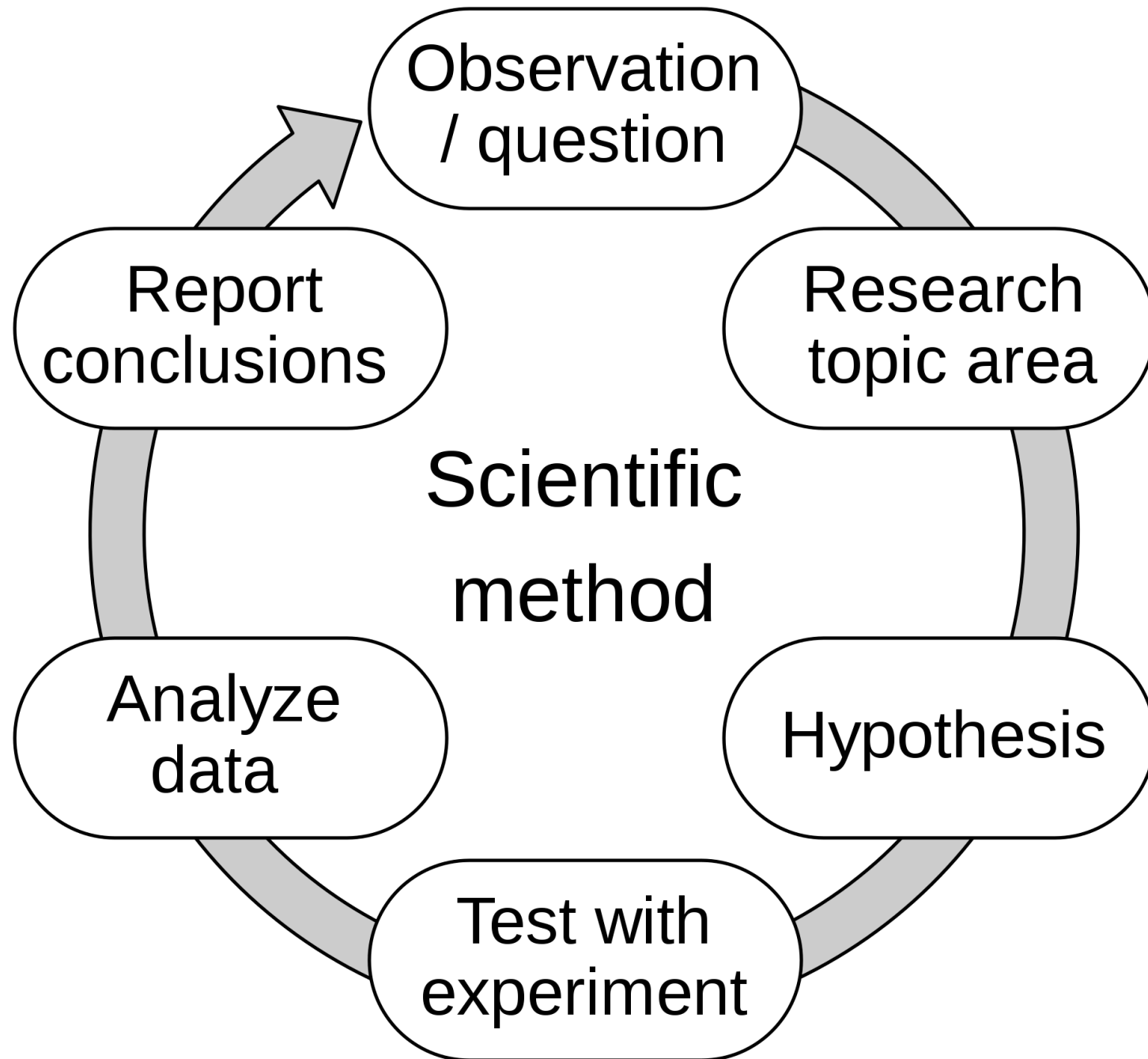
Inject domain knowledge into ML system

- BSM priors:
 - Crowdsource theorists
 - Generative models
 - ...
- Differentiable programming*
- *End-to-end systems*

No. 1 Priority

Maximize
Discovery
Potential

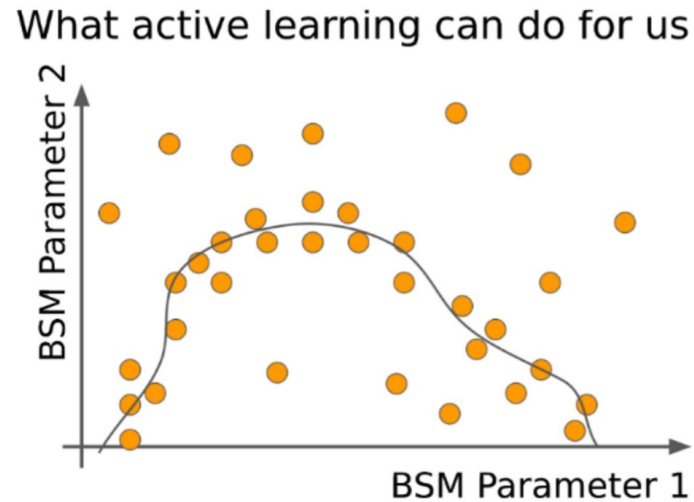
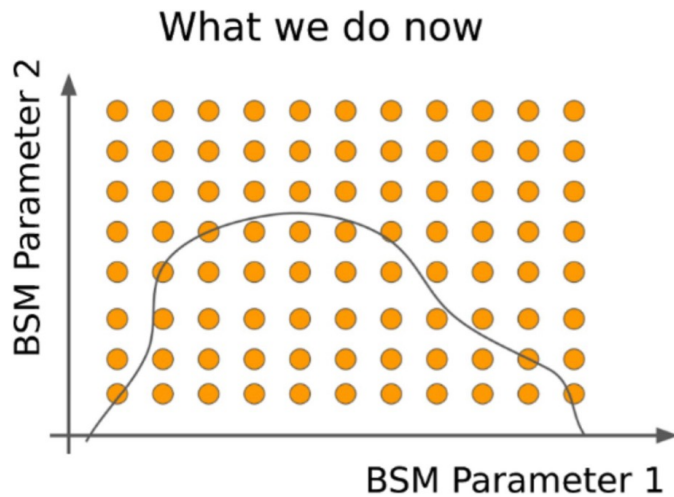




№ 2

(Re)Interpret
Results

Smart sampling with active learning

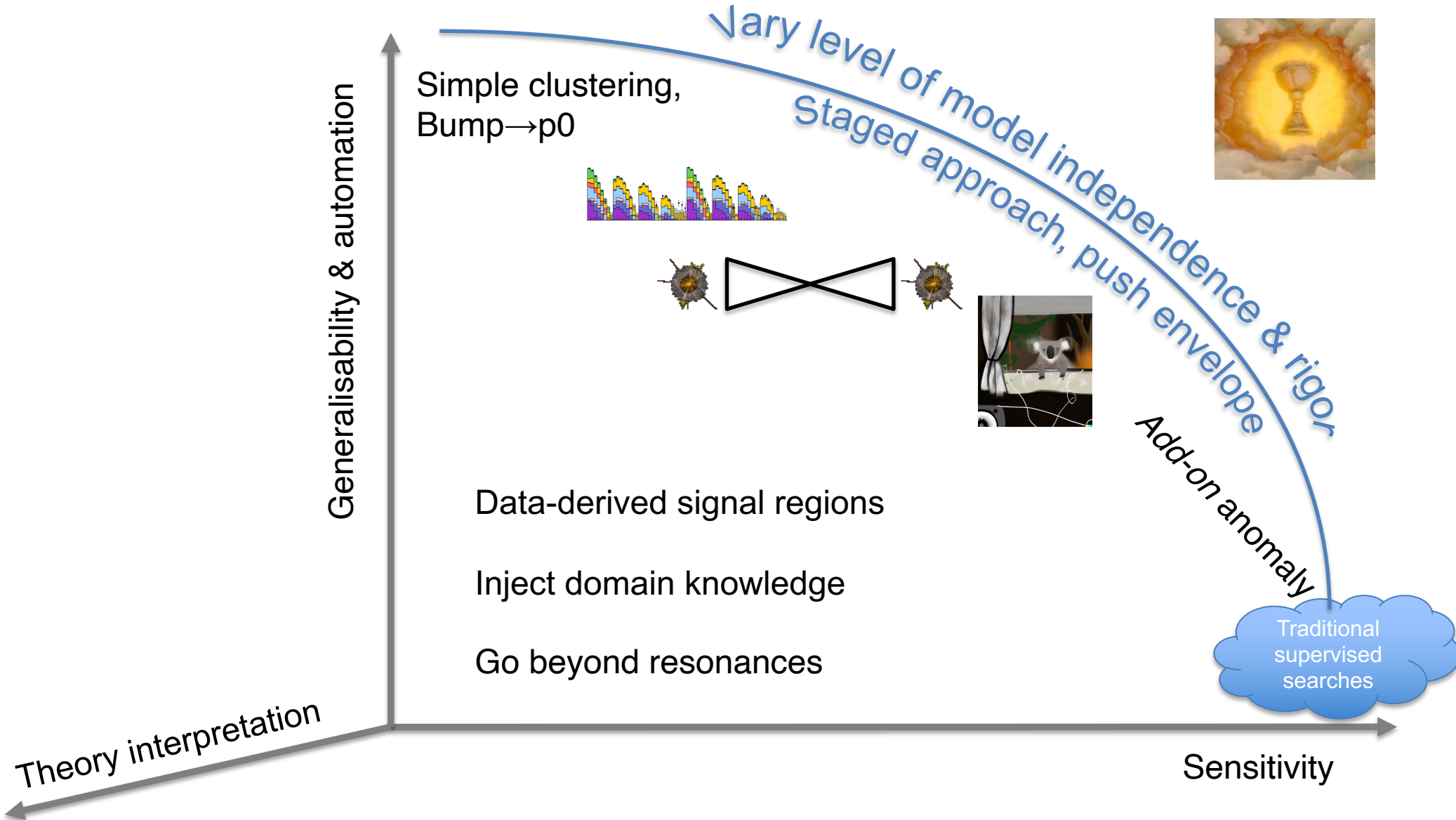


Thrives on
high-dimensional
theory space

→ Simulate on demand

Search Strategy

Diverse



Device
Strategy
in
Experiment

**"There is no
power for change
greater than a
community
discovering what
it cares about."**

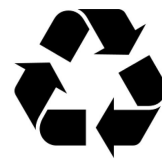
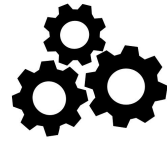
Be

F
findable

A
accessible

I
interoperable

R
reusable



Reusable & differentiable code [e.g. ACTS] → Geant4,...

Common benchmarks, metrics, staged, ceiling analysis,...

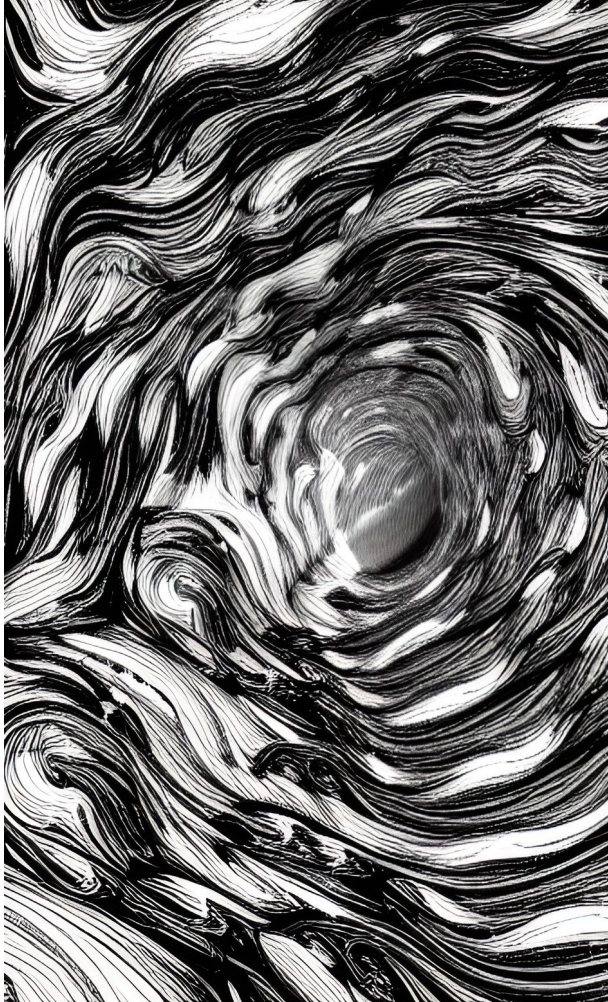
Known knowns

Proof of concepts are *easy*

Production-readiness is hard → Delay



Experiments buy into ML



Brain drain

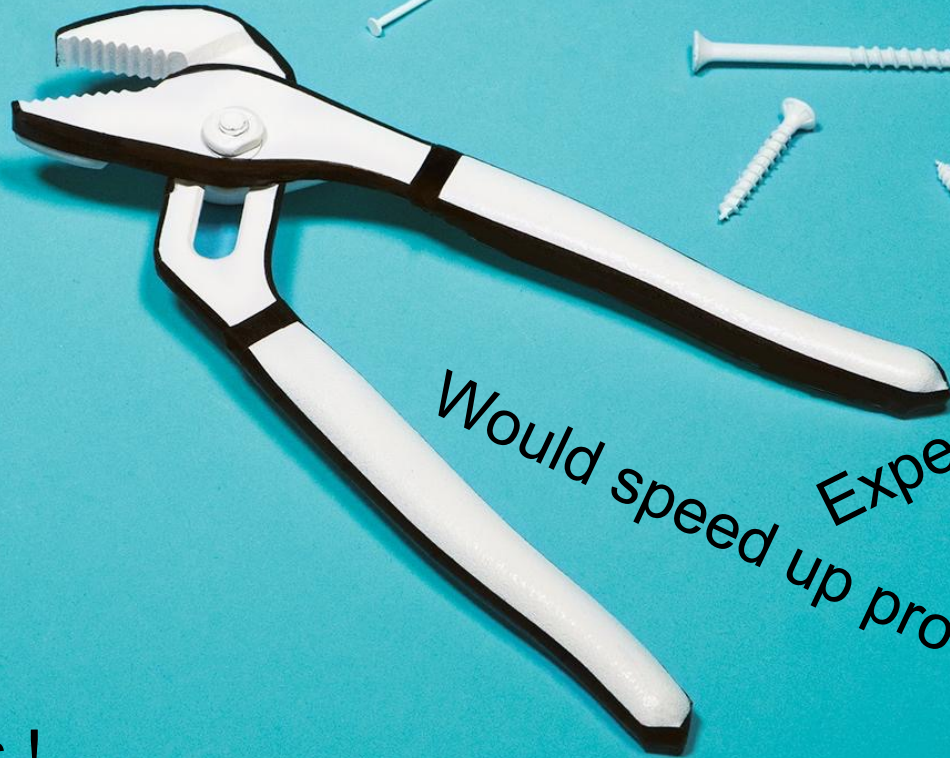
ML@HEP \leftrightarrow *Open Science* @ Experiment

Recognition

Experiment time scale \leftrightarrow Fast-moving ML

Fixes?

Fast-track limited-author method papers



Would speed up production readiness

Experiments could be **the MNISTs**

And of course ML4Jets !

Concluding remarks

HEP evolves

ML is one of our sharpest tools

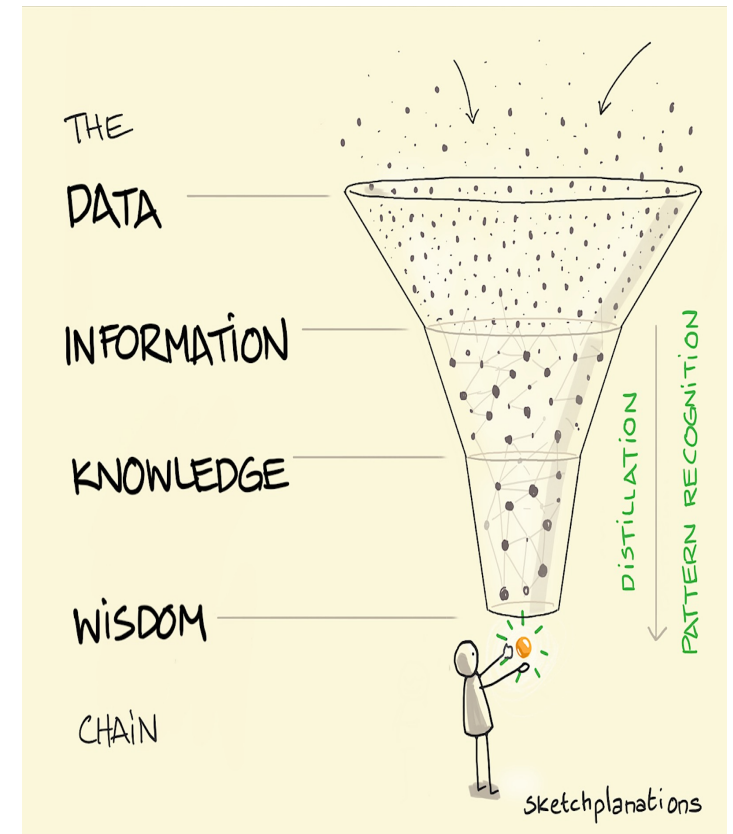
Tackle Big Goals as a Community

Value human resources → automation with ML

Concept → production

New *diverse* search strategy

Need all you bright young ML4Jets minds !



PIs

PhD students

postdocs



TG



Tomke Schröer



Malte Algren



Lukas Ehrke



Matthew Leigh



Debajyoti Sengupta



Sam Klein



Knut Zoch



Manuel Guth



Johnny Raine



This could be you !



Slava Voloshynovskiy



Guillaume Quétant



Mariia Drozdova



Ivan Oleksiyuk



Olga Taran



François Fleuret



Bálint Máté



Atul Kumar Sinha



Daniele Paliotta



PIs

PhD students

postdocs



TG



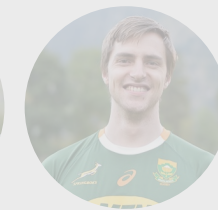
Tomke Schröer



Malte Algren



Lukas Ehrke



Matthew Leigh



Debajyoti Sengupta



Sam Klein



Knut Zoch



Manuel Guth



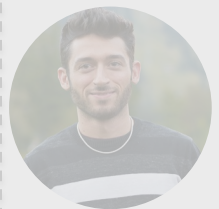
Johnny Raine



This could be you !



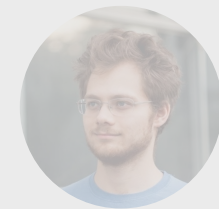
Slava Voloshynovskiy



Guillaume Quétant



Mariia Drozdova



Ivan Oleksiyuk



Olga Taran



François Fleuret



Bálint Máté



Atul Kumar Sinha



Daniele Paliotta

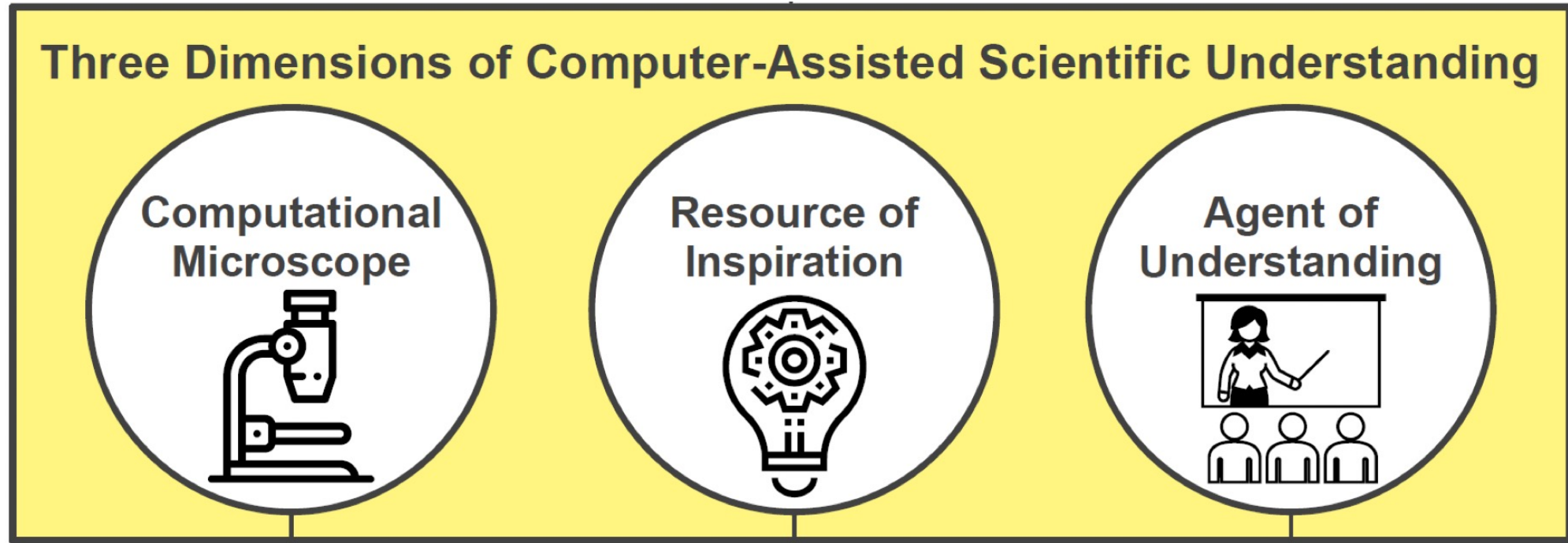
This could be you !

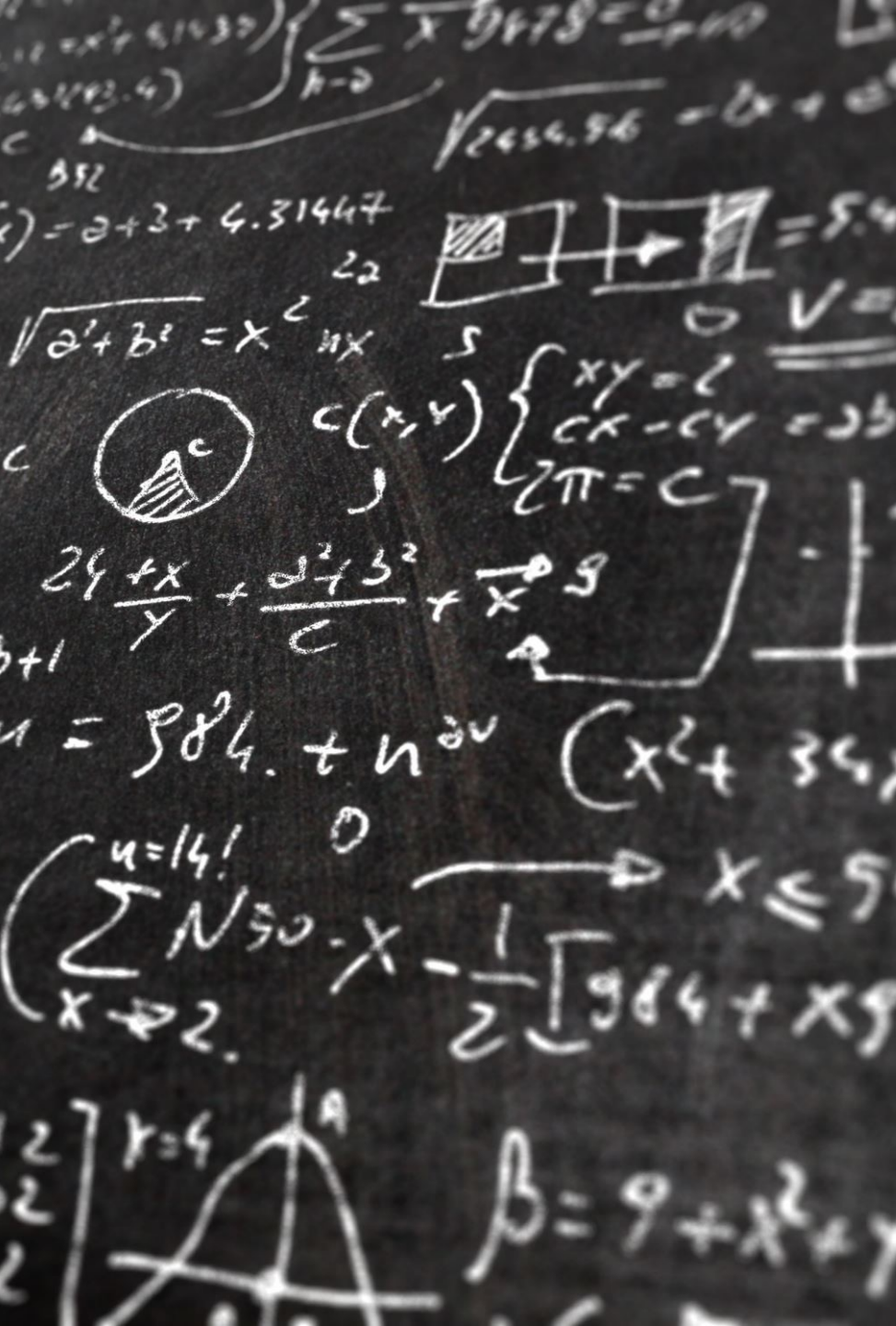
Contact Tobias.Golling@unige.ch



Backup

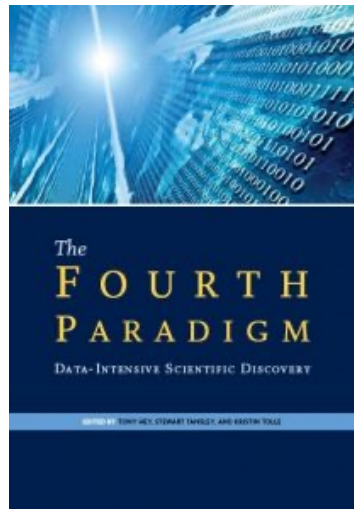
Qualitative change





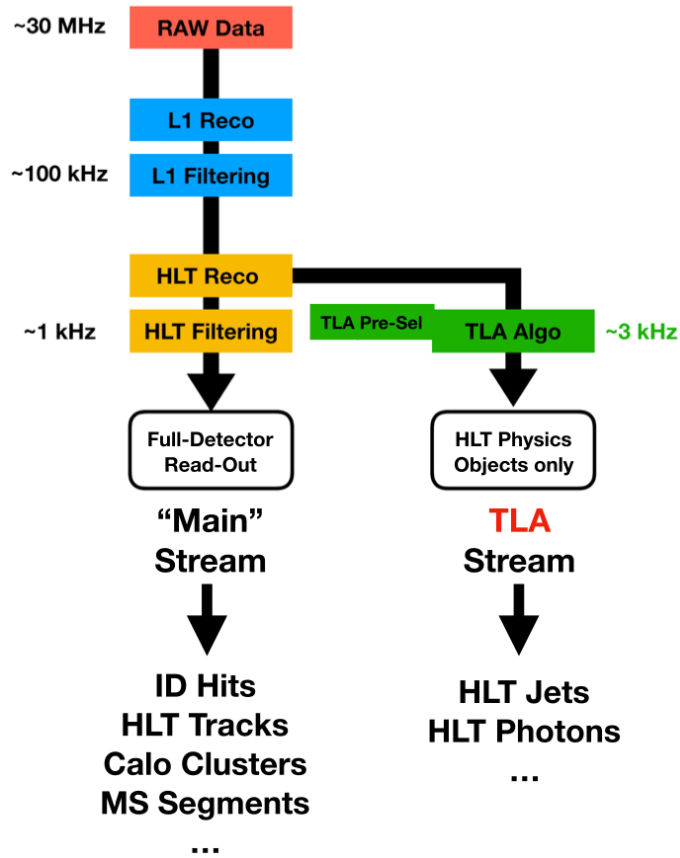
The 4 paradigms of scientific discovery

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



Jim Gray

TLA

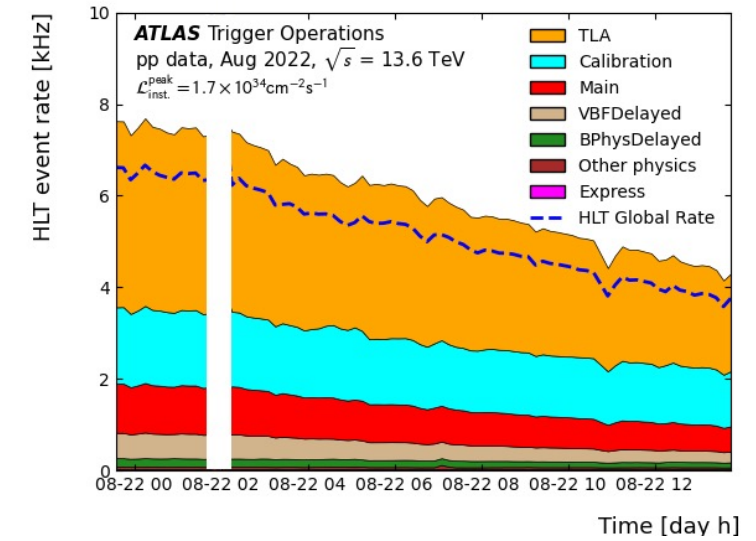


Physics main
 $\sim 1 \text{ kHz} \times 1.5 \text{ MB} \rightarrow 1.5 \text{ GB/s}$

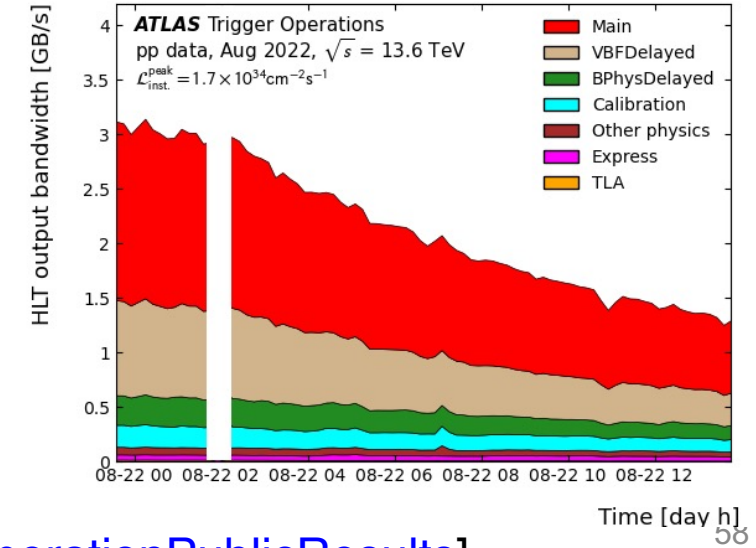
TLA
 $\sim 4 \text{ kHz} \times 5 \text{ kB} \rightarrow 20 \text{ MB/s}$

- Trade event size for rate!
- Save only HLT objects
- Challenge: calibration!

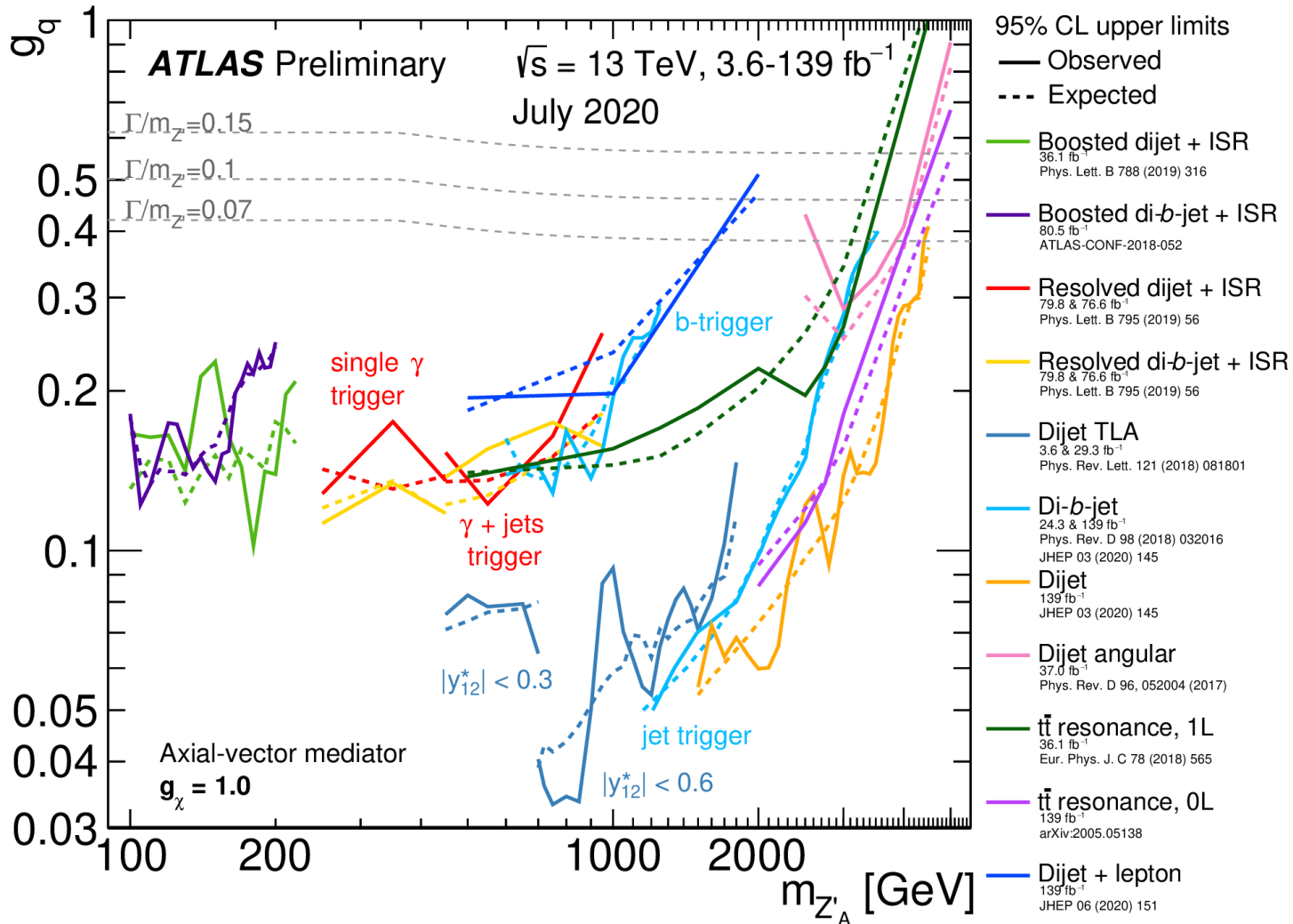
TLA stream has highest rate...



...but lowest bandwidth



BSM@TLA



- Extent search to anomaly@TLA