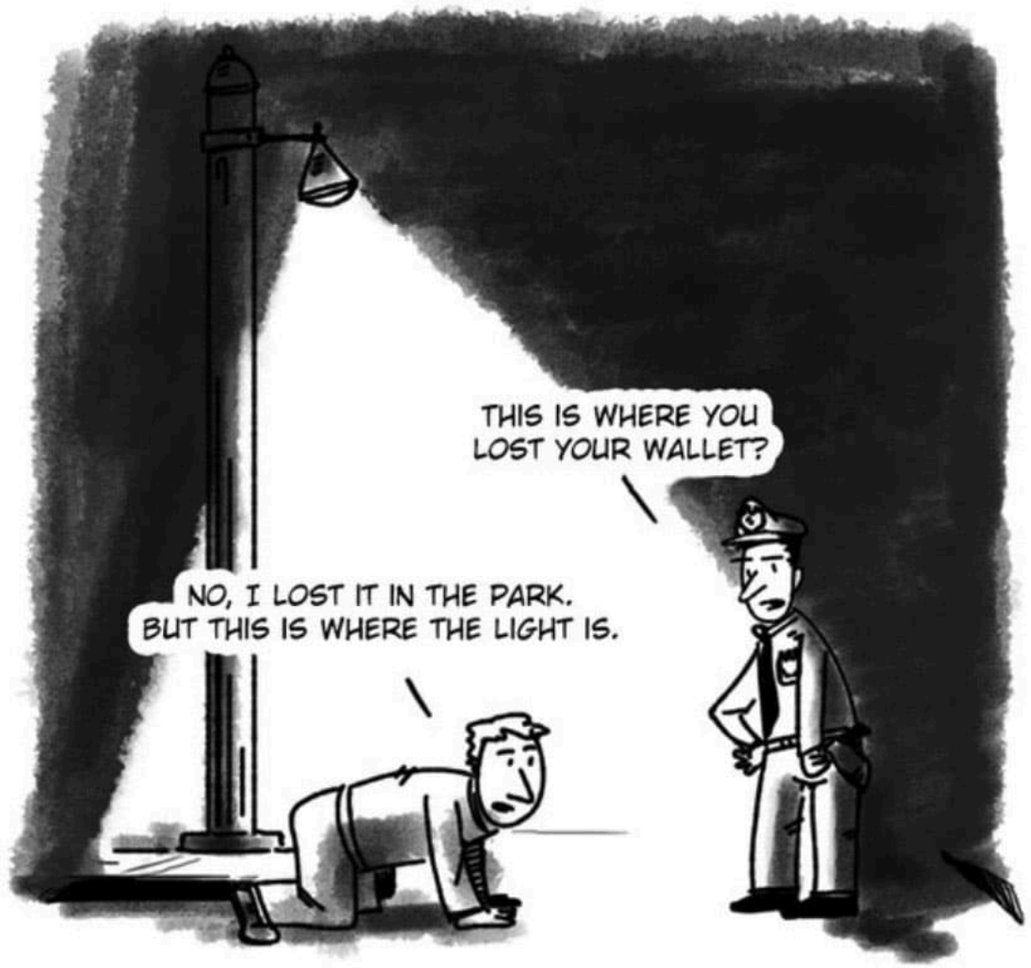
New Paradigms for New Physics Searches with Machine Learning

David Shih

January 13, 2021

Theory Mini-workshop
HKUST IAS Program on HEP

Where is the new physics??

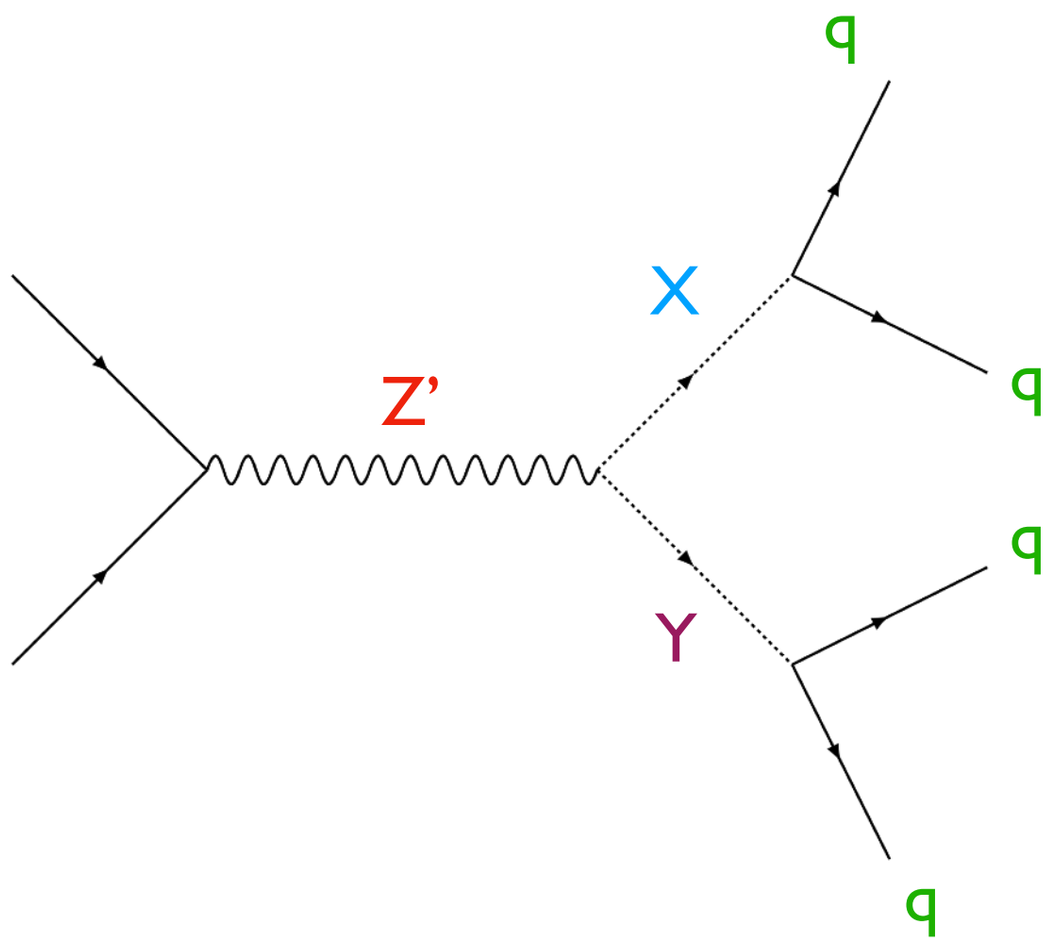


Despite thousands of searches for new physics at the LHC, nothing but limits and null results so far.

What if new physics is hiding in the data but we haven't looked in the right places yet?

A Benchmark Example

LHC Olympics 2020 R&D Dataset
https://doi.org/10.5281/zenodo.2629072



No explicit search at the LHC for this scenario!

Could be hiding in the dijet resonance search at >5sigma significance!!

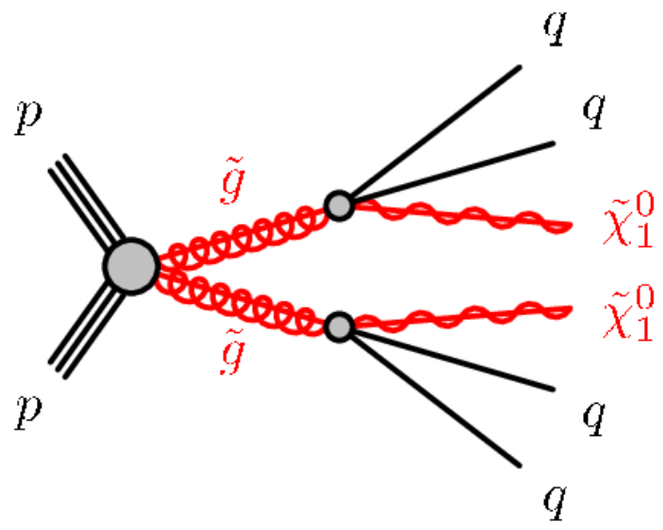
The most common approach

Model specific searches

Most NP searches at the LHC are heavily optimized with specific signals in mind (SUSY, extra dimensions, …)

ATLAS jets+MET 2010.14293

	BDT-GGd1	BDT-GGd2	BDT-GGd3	BDT-GGd4
N_{j}	≥ 4			
$\Delta\phi(j_{1,2,(3)}, \boldsymbol{p}_{\mathrm{T}}^{\mathrm{miss}})_{\mathrm{min}}$	> 0.4			
$\Delta\phi(j_{i>3}, \boldsymbol{p}_{\mathrm{T}}^{\mathrm{miss}})_{\mathrm{min}}$	> 0.4			
$E_{\mathrm{T}}^{\mathrm{miss}}/m_{\mathrm{eff}}(N_{\mathrm{j}})$	> 0.2			
m_{eff} [GeV]	> 1400		> 800	
BDT score	> 0.97	> 0.94	> 0.94	> 0.87
$\Delta m(\tilde{g}, \tilde{\chi}_1^0)$ [GeV]	1600–1900	1000–1400	600–1000	200–600



Kinematic cuts (or BDTs) optimized using simulations of signal AND background.

The most common approach

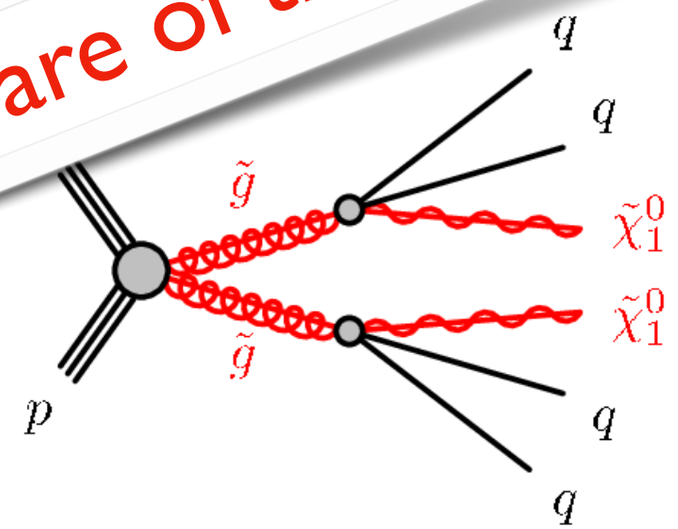
Model specific searches

Most NP searches at the LHC are heavily optimized with sp… signals in mind (SUSY, extra dimensions, …)

> 99% of searches at the LHC are of this type

ATLAS jets+MET 2010.14293

	BDT-GGd1	BDT-GGd2	BDT-GGd3	
N_{j}		≥ 4		
$\Delta\phi(j_{1,2,(3)}, \boldsymbol{p}_{\mathrm{T}}^{\mathrm{miss}})$ min				
$\Delta\phi(j_{i>3}, \boldsymbol{p}_{\mathrm{T}}^{\mathrm{miss}})$ min				
$E_{\mathrm{T}}^{\mathrm{miss}}/m_{\mathrm{eff}}(N_{\mathrm{j}})$				
m_{eff} [GeV]			> 800	
BDT sc…		> 0.94	> 0.94	> 0.87
$\Delta m(\tilde{g}, \tilde{\chi}_1$…	…00–1900	1000–1400	600–1000	200–600



Kinematic cuts (or BDTs) optimized using simulations of signal AND background.

Of course, we should continue to perform these searches, because NP could always be right around the corner…

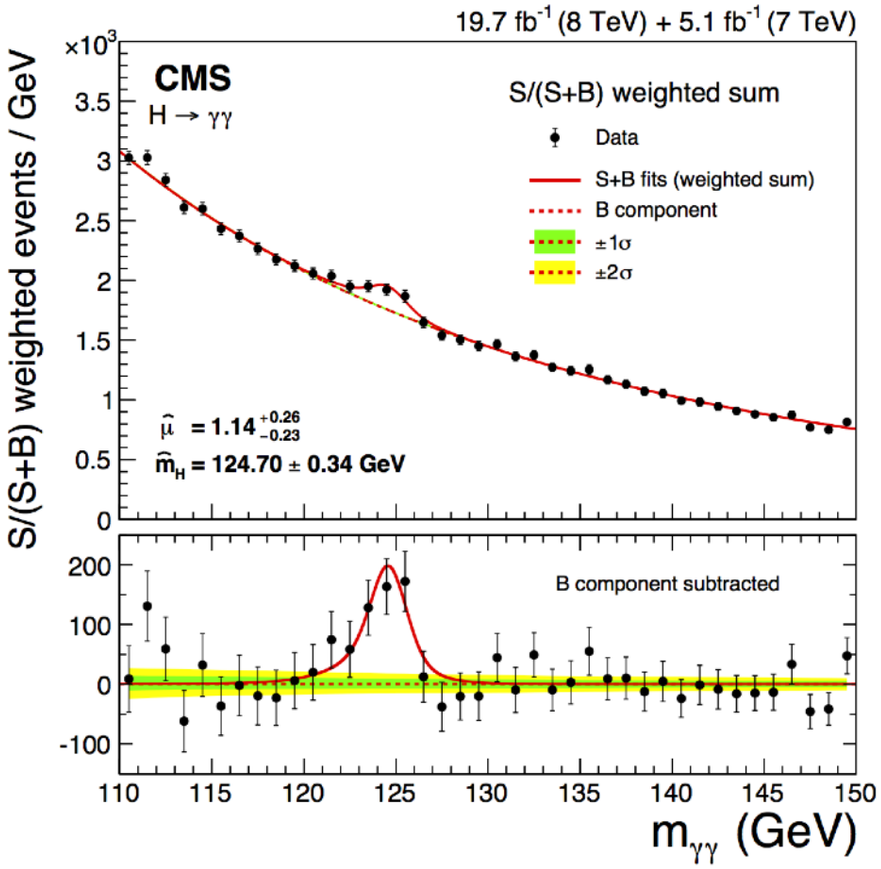
But we should also recognize that perhaps new physics is on a different street, or in a different building altogether…

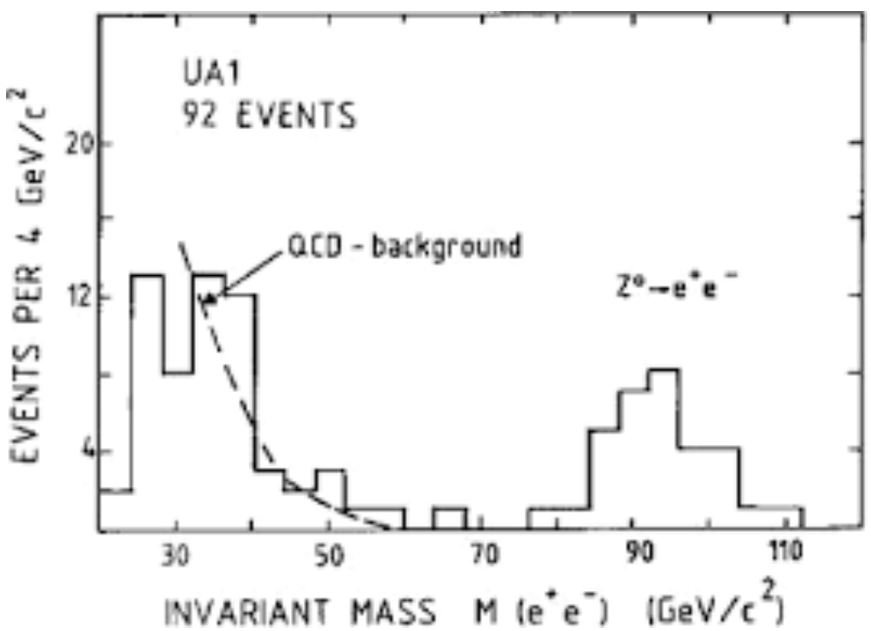
Existing model-independent approaches

"the bump hunt"

Idea: assume signal is localized in some feature (usually invariant mass) while background is smooth.

Interpolate from **sidebands** into **signal region**, search for an excess.



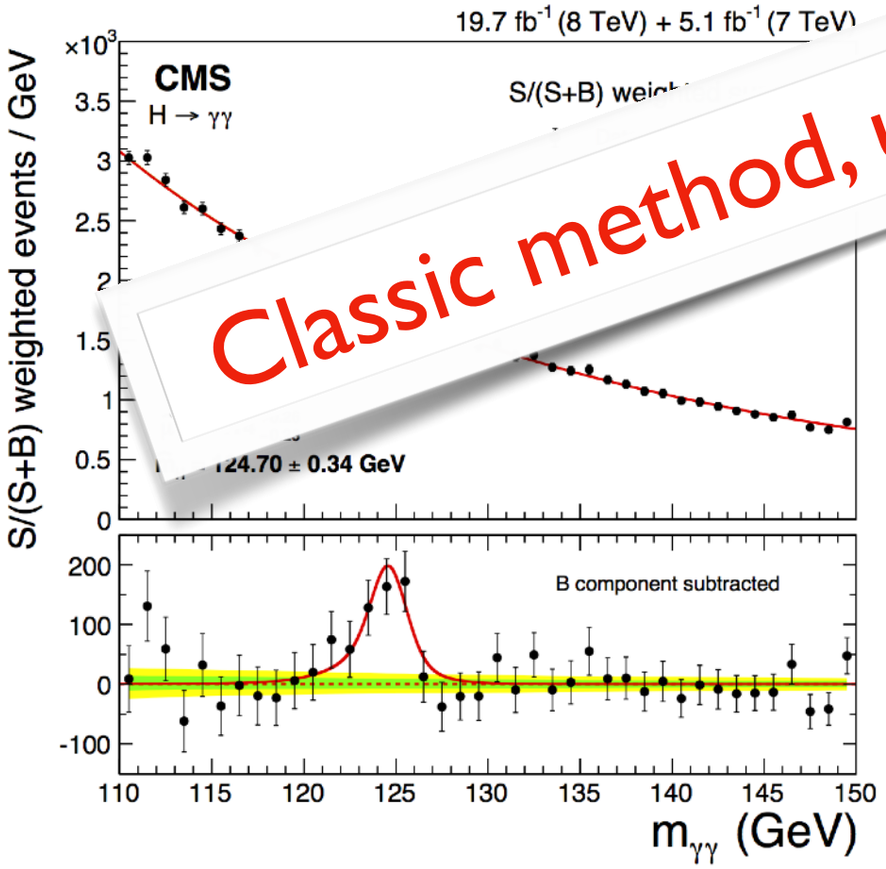


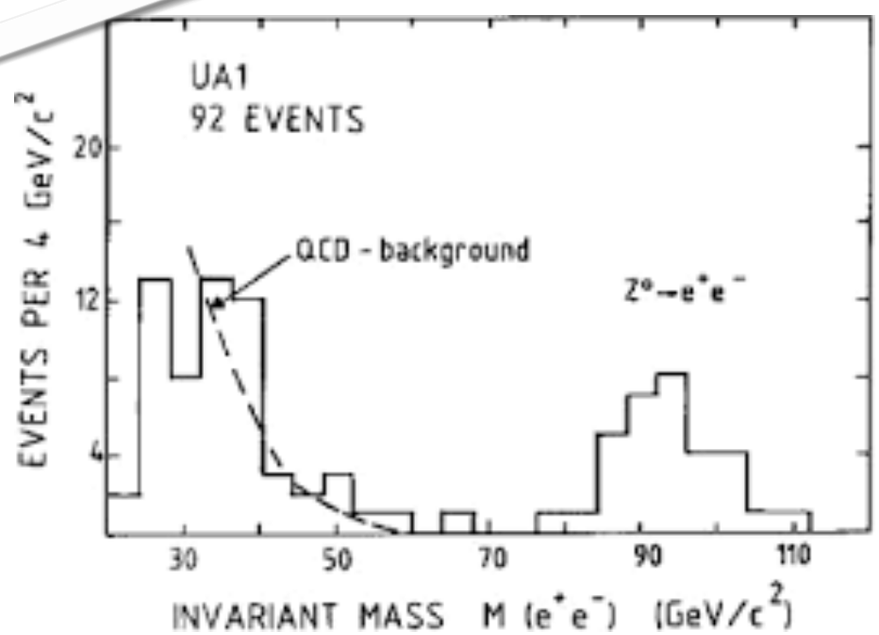
Existing model-independent approaches

"the bump hunt"

Idea: assume signal is localized in some feature (usually invariant mass) while background is smooth.

Interpolate from **sidebands** into **signal re** excess.



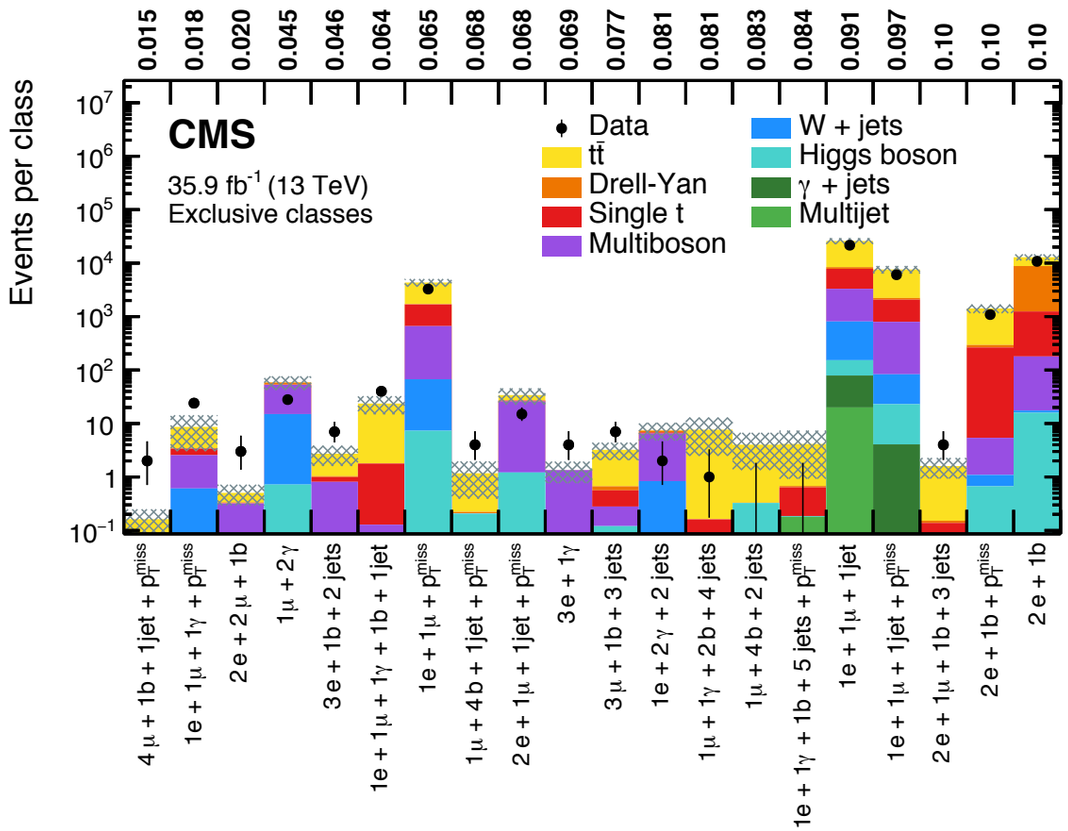


Classic method, used in many discoveries.

Existing model-independent searches

"the general search"

Idea: divide the phase space up into thousands of bins, compare **data to SM simulation** in each one



CMS

"MUSIC"

CMS-PAS-EXO-14-016

ATLAS

"Model independent general search"

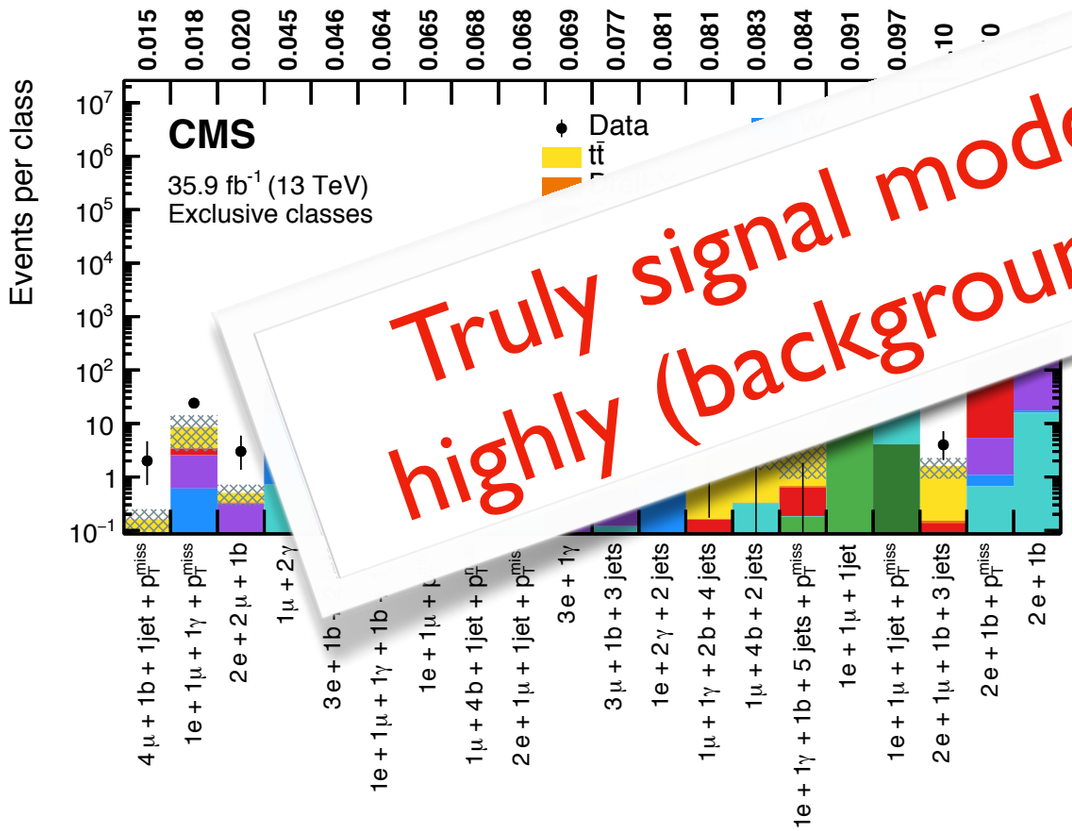
1807.07447
EPJC 79:120 (2019)

See also proposals by D'Agnolo, Wulzer et al (1806.02350, 1912.12155): train DNN on full phase space to distinguish data from background

Existing model-independent searches

"the general search"

Idea: divide the phase space up into thousands of bins, com
data to SM simulation in each one



Truly signal model independent, but still highly (background) simulation dependent

ATLAS

"MUSIC"

CMS-PAS-EXO-14-016

"Model independent general search"

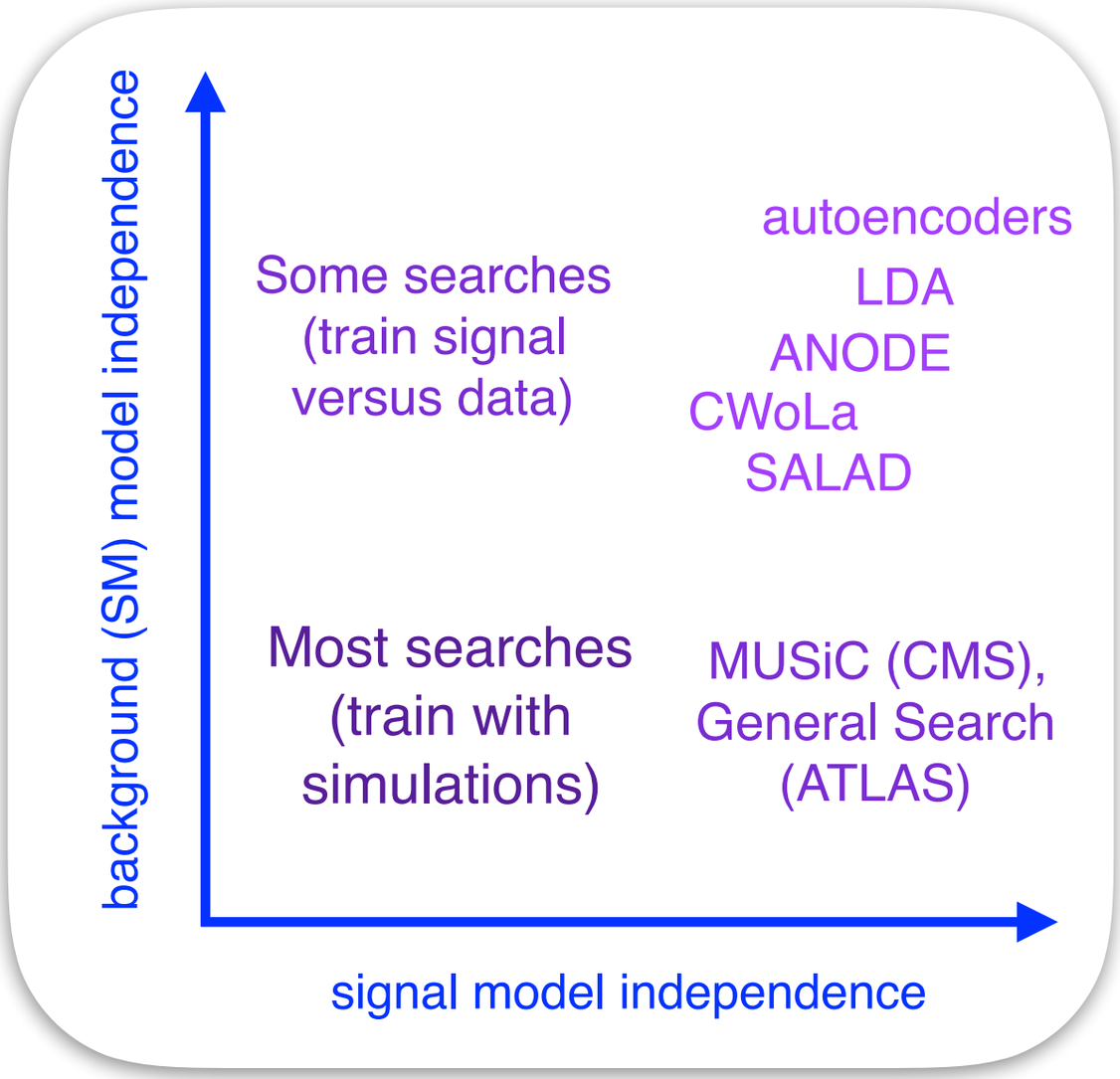
1807.07447
EPJC 79:120 (2019)

See also proposals by D'Agnolo, Wulzer et al (1806.02350, 1912.12155): train DNN on full phase space to distinguish data from background

New paradigms for model-agnostic searches

Can advances in machine learning open up new avenues for model-independent searches?

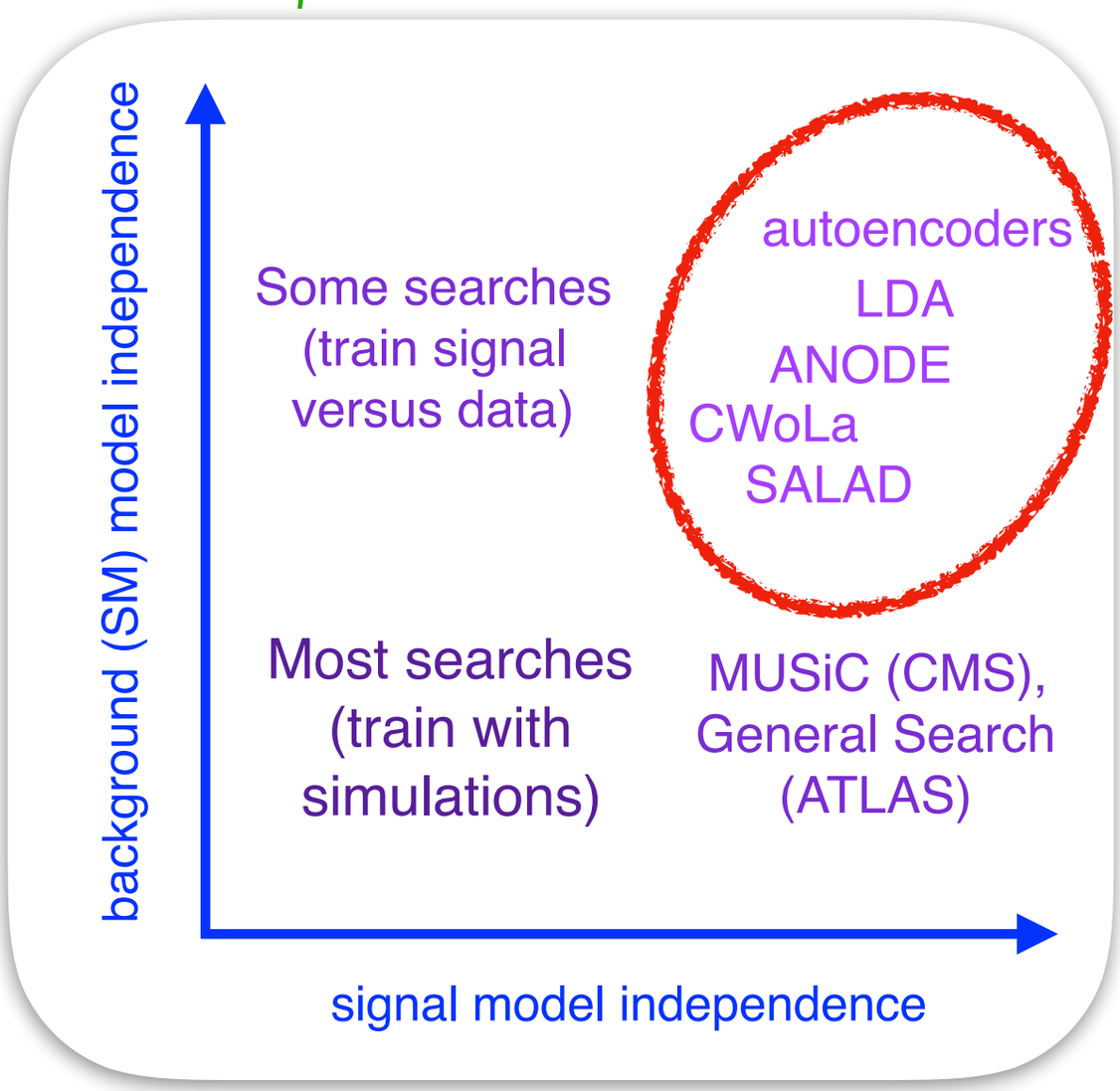
from Nachman & DS 2001.04990



New paradigms for model-agnostic searches

Can advances in machine learning open up new avenues for model-independent searches?

from Nachman & DS 2001.04990



Many new ideas recently!

Many new approaches inspired by (or at least demonstrated on) the **LHC Olympics 2020 Data Challenge**

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 Yunus[18]

to appear on the arXiv next week…

Many new approaches inspired by (or at least demonstrated on) the **LHC Olympics 2020 Data Challenge**

LHC Olympics 2020: Black Boxes

https://doi.org/10.5281/zenodo.3547721

In 2019, Gregor Kasieczka, Ben Nachman and I initiated the LHC Olympics 2020 Challenge.

It consisted of three "black boxes" of simulated data (bg dominated!):

- 1 million events each
- 4-vectors of every reconstructed particle (all hadronic) in the event
- Particle ID, charge, etc not included
- Single R=1 jet trigger pT>1.2 TeV

The goal of the challenge was for participants to analyze each box and

1. Decide whether or not it contains new physics
2. Characterize the new physics, if it's there

LHC Olympics 2020: Black Boxes

https://doi.org/10.5281/zenodo.3547721

In 2019, Gregor Kasieczka, Ben Nachman and I initiated the LHC Olympics 2020 Challenge.

It consisted of three "black boxes" of simulated data (h

- 1 million events each
- 4-vectors of

ger pT>1.2 TeV

Stay tuned for Ben's talk to see what was in the boxes and the outcome of the challenge

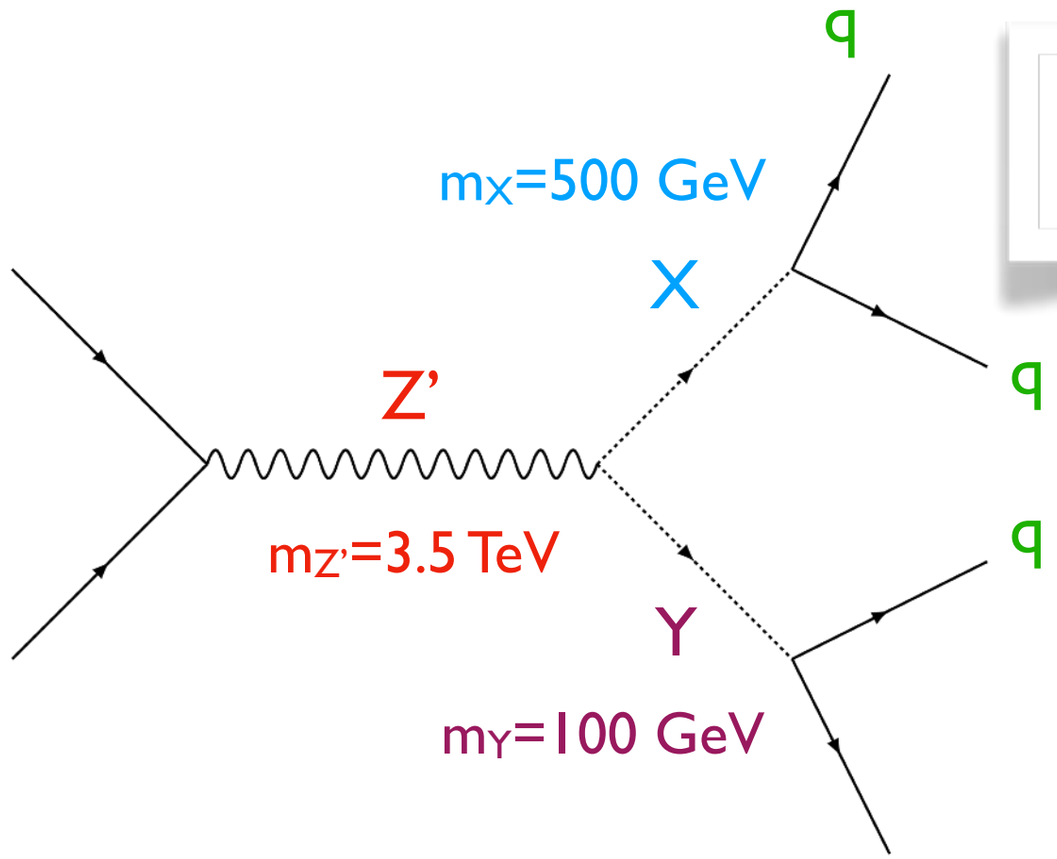
The goal of the challenge was for participants to analyze each box and

1. Decide whether or not it contains new physics
2. Characterize the new physics, if it's there

LHC Olympics 2020: R&D Dataset

https://doi.org/10.5281/zenodo.2629072

Prior to the challenge, we also released a labeled R&D dataset consisting of 1M QCD dijet events and 100k signal events

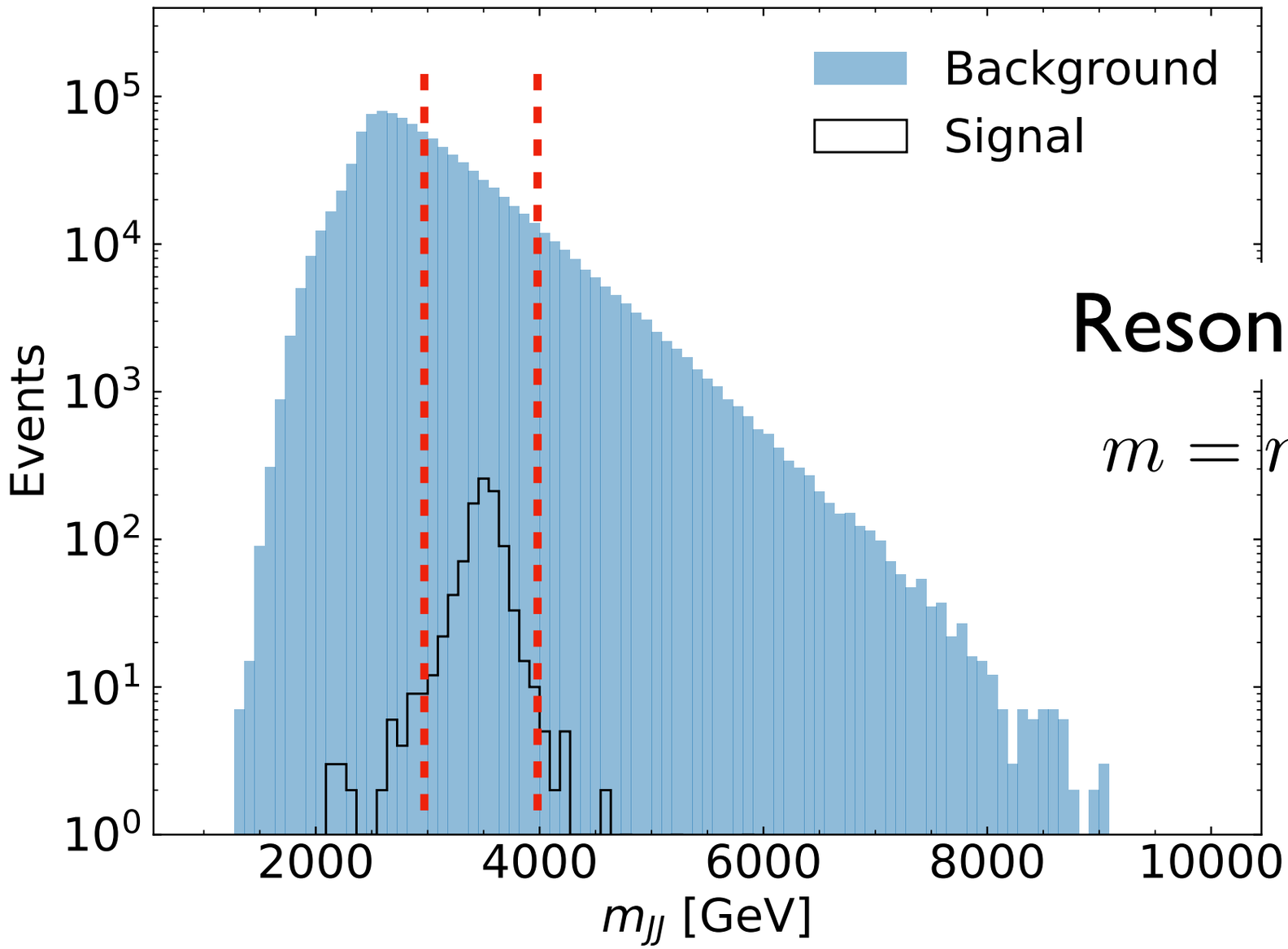


No explicit search at the LHC for this scenario!

Unofficial theme of LHCO2020:

"enhancing the bump hunt"

LHC Olympics 2020: R&D Dataset



Resonant feature

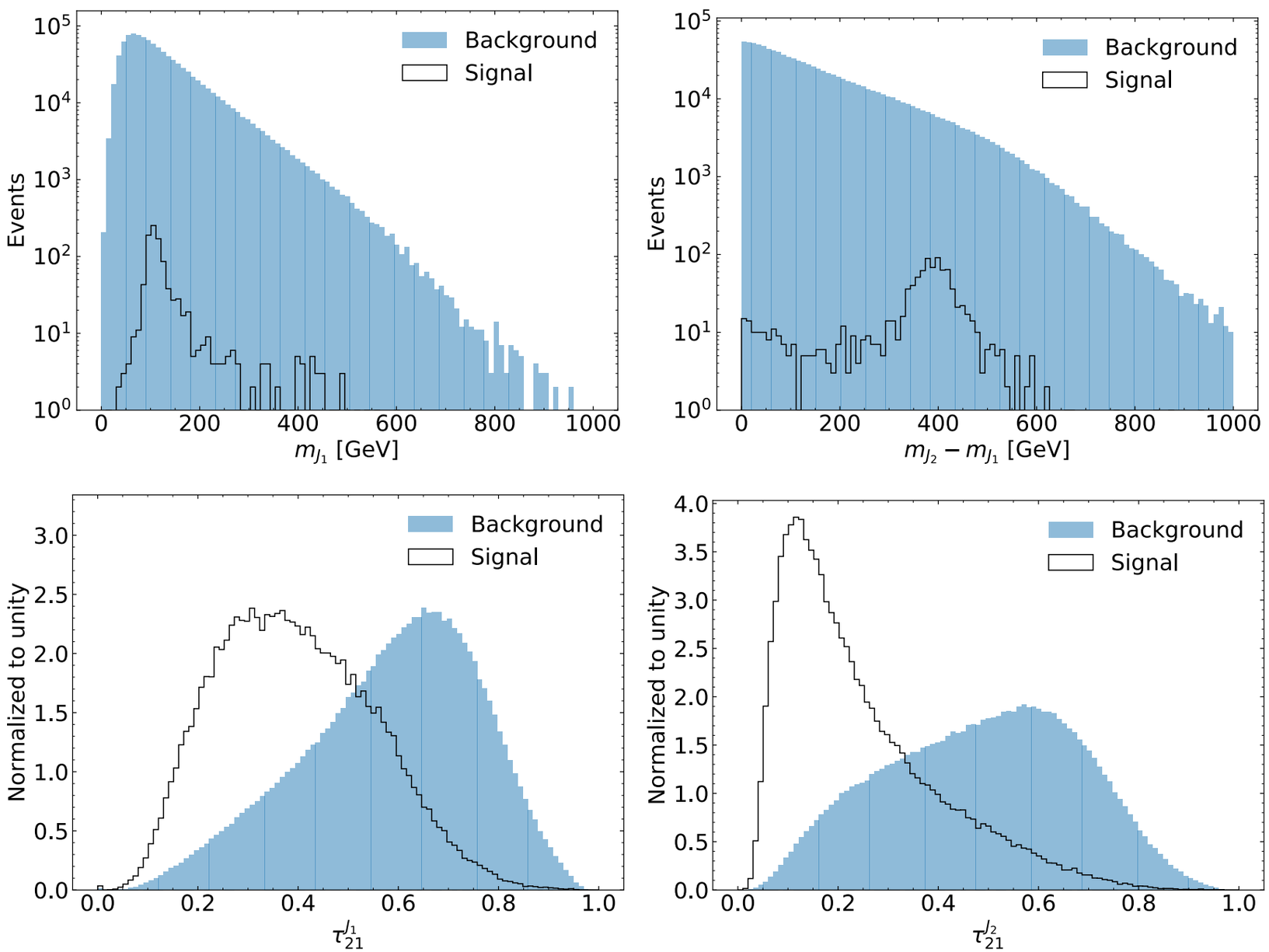
$$m = m_{Z'} = m_{JJ}$$

Benchmark signal strength:

S=500, B=500,000, B_{SR}=61,000

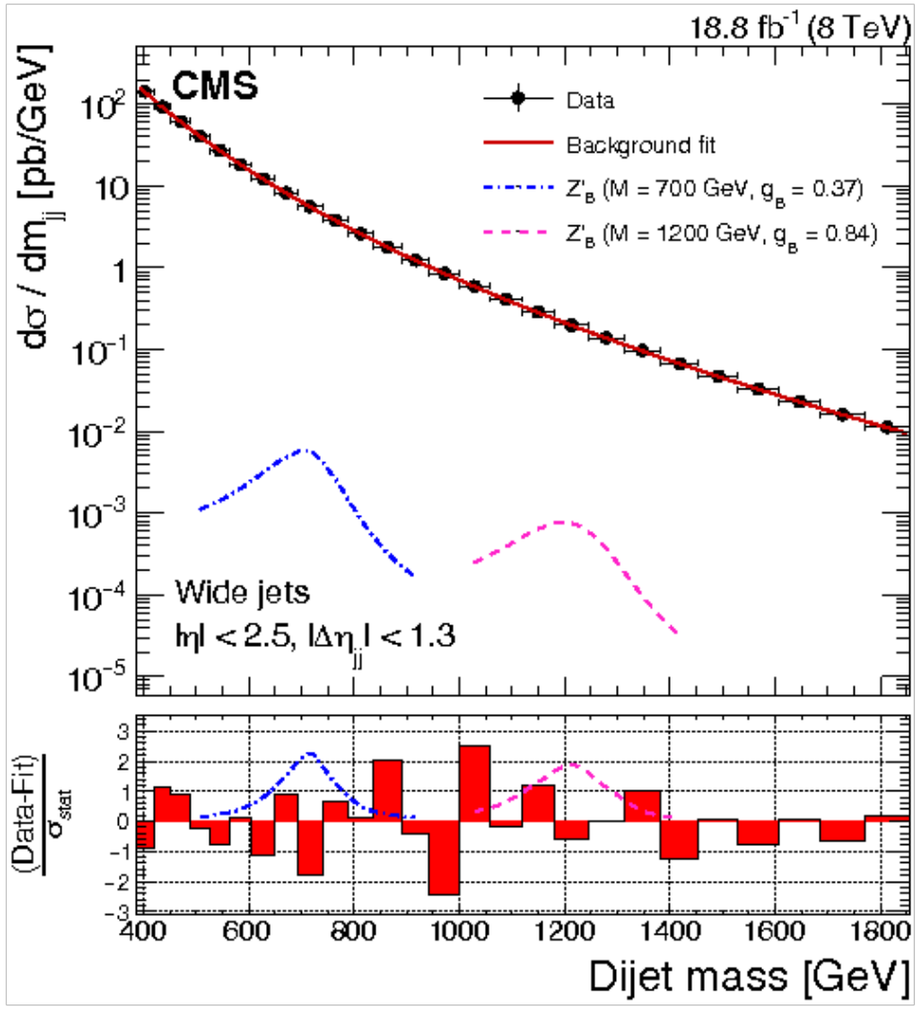
$S/B_{SR} \sim 6\times10^{-3}$, $S/\sqrt{B_{SR}} \sim 1.5$

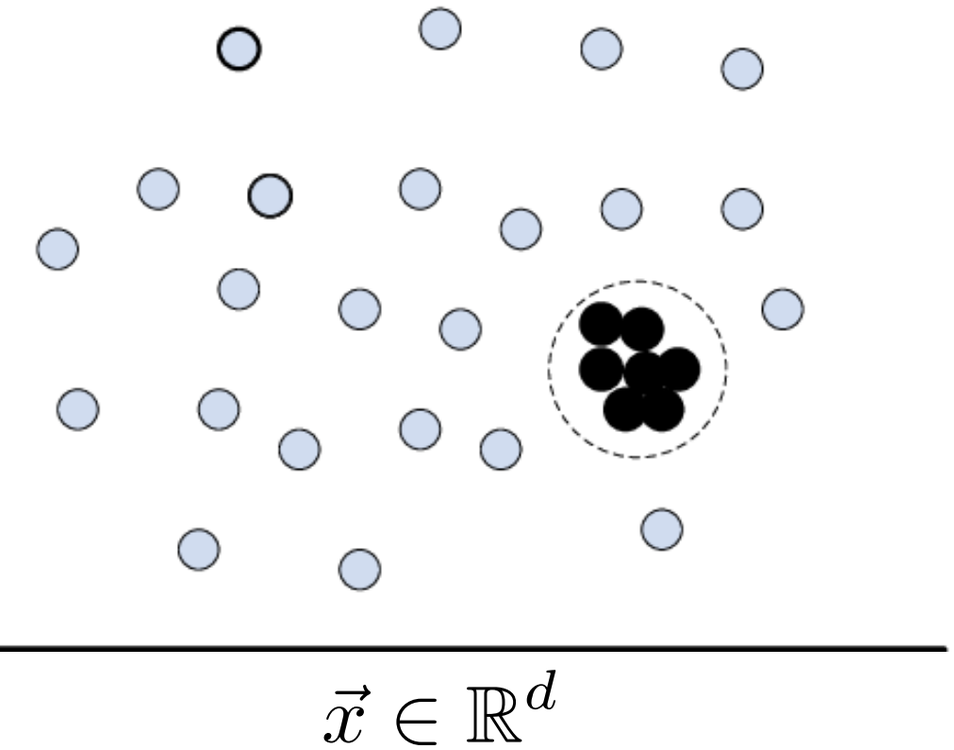
LHC Olympics 2020: R&D Dataset



Additional features: $x = (m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2})$

Enhancing the bump hunt



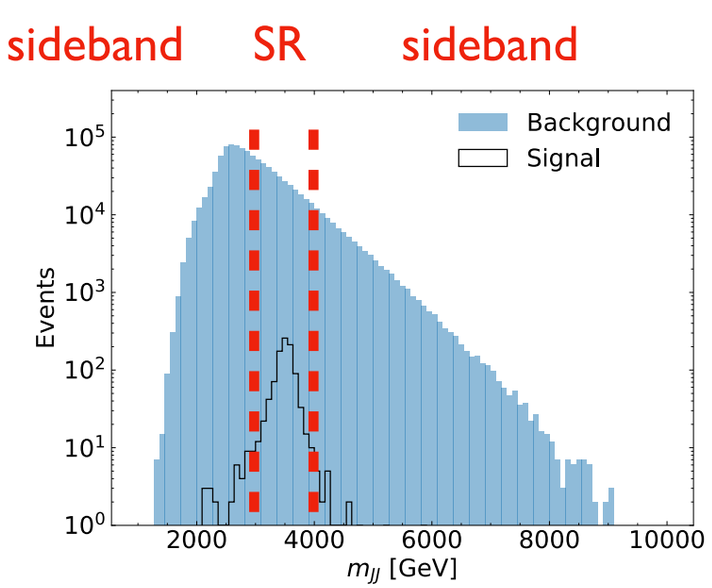


primary resonant feature (mJJ)

additional features

If the signal is localized in additional features, can we find it in a model-independent way?

Enhancing the bump hunt

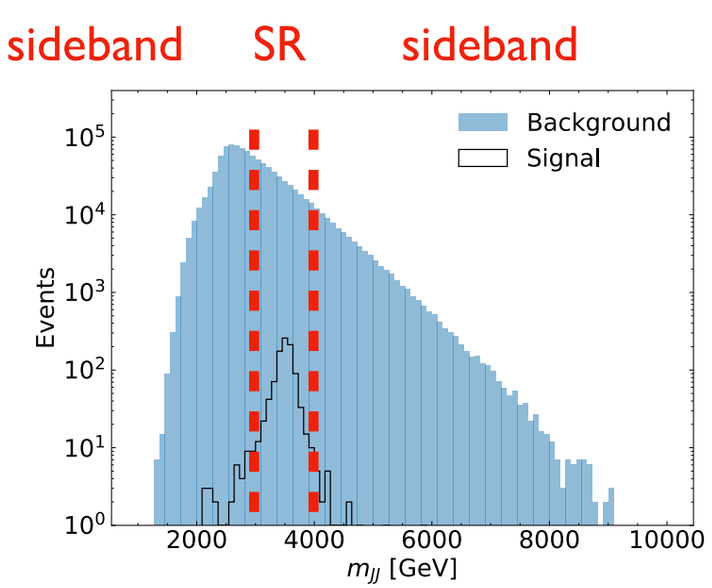


The optimal model-agnostic discriminant would be

$$R(x) = \frac{\mathcal{P}(x|data)}{\mathcal{P}(x|bg)}$$

Key idea: use sidebands in mJJ to model P(x|bg)

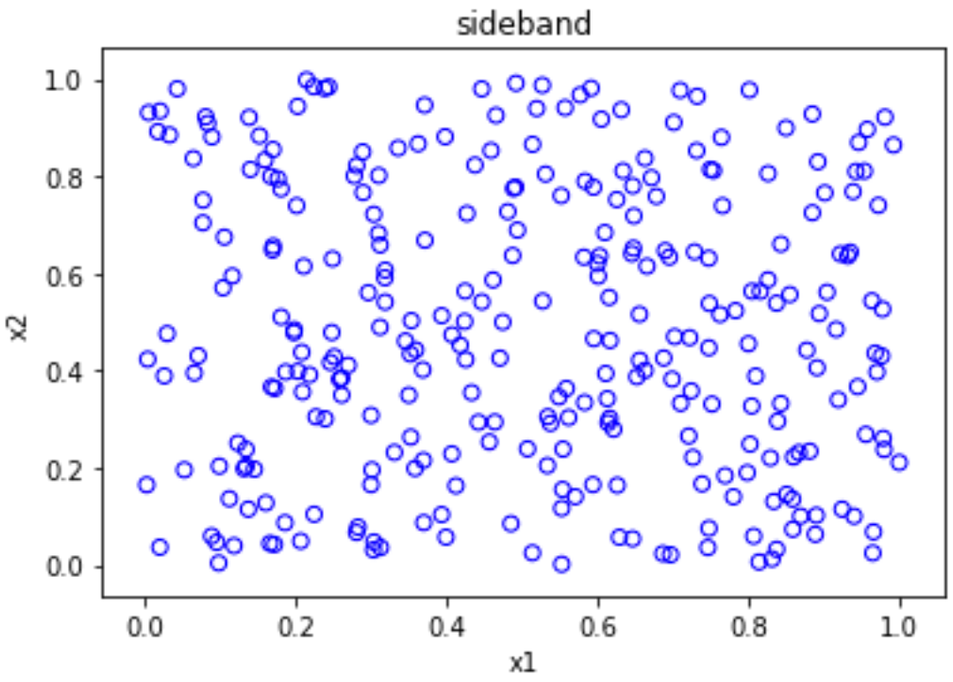
Enhancing the bump hunt

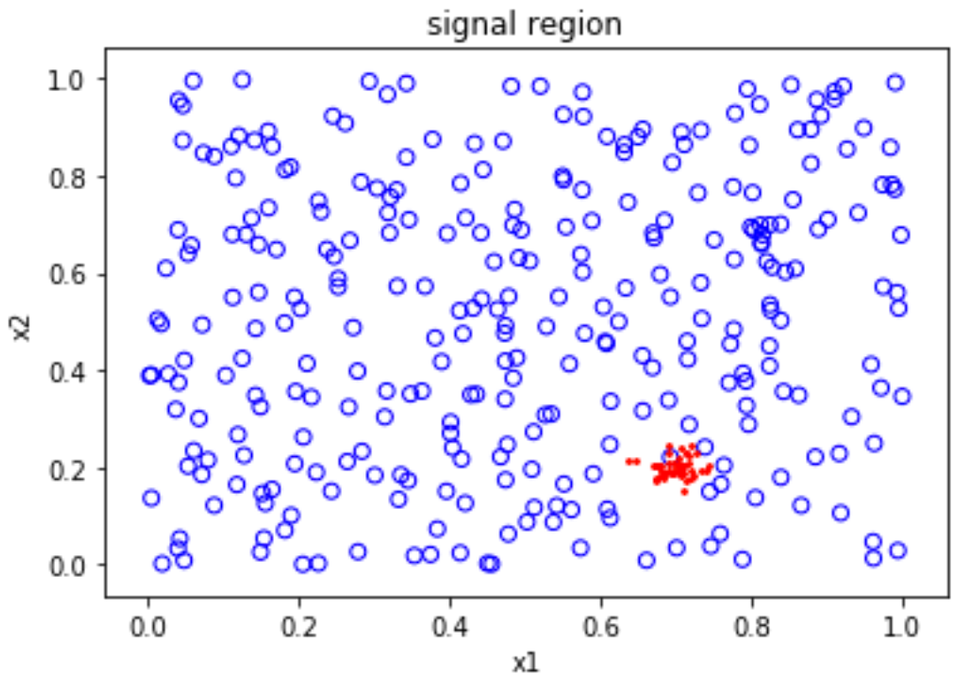


The optimal model-agnostic discriminant would be

$$R(x) = \frac{\mathcal{P}(x|data)}{\mathcal{P}(x|bg)}$$

Key idea: use sidebands in mJJ to model P(x|bg)





ANODE: Anomaly Detection with Density Estimation

Nachman & DS 2001.04990

Example of a new approach inspired by LHCO2020.

(See Ben's talk for additional new approaches!)

Use **neural density estimation** to directly learn the conditional probability densities from the data

$$\mathcal{P}(x|data;\ m_{JJ} \in SR) \qquad \mathcal{P}(x|data;\ m_{JJ} \notin SR) = \mathcal{P}(x|bg;\ m_{JJ} \notin SR)$$

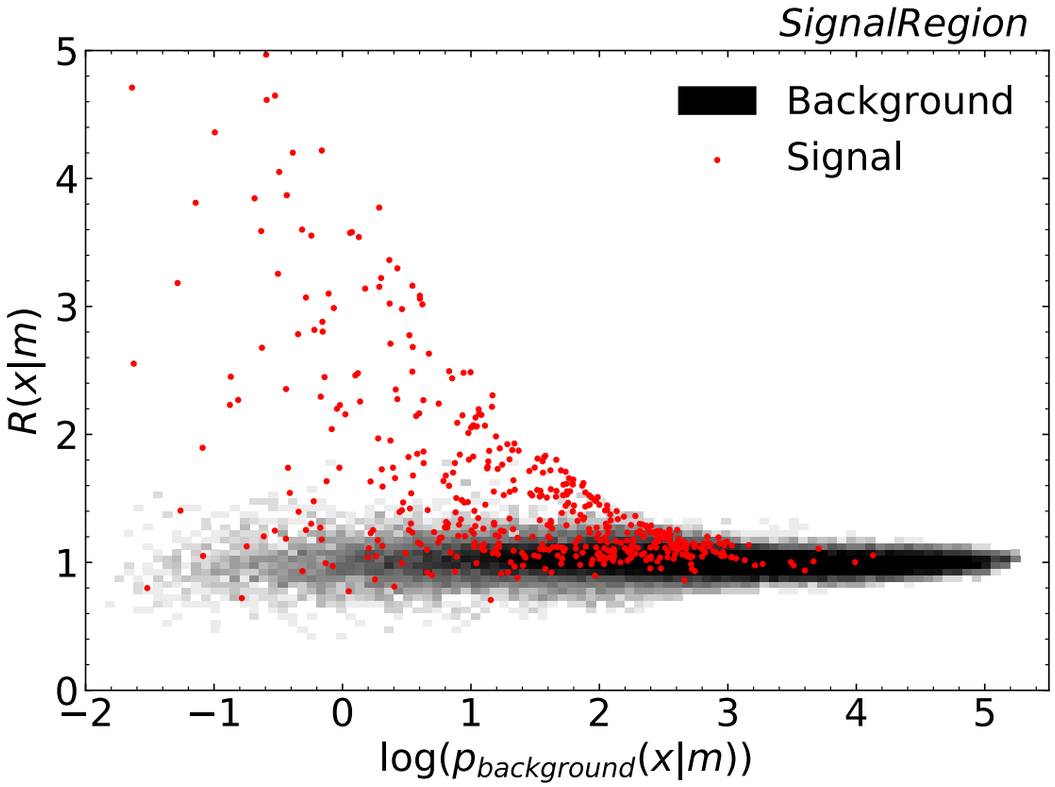
interpolate in (x,mJJ)

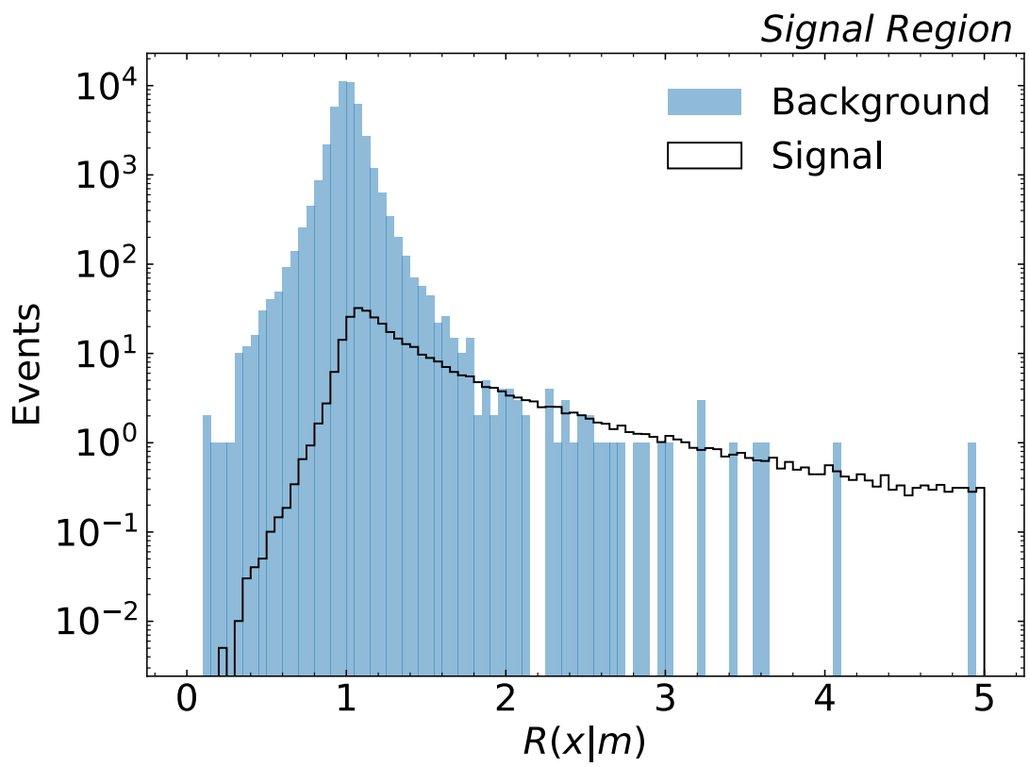
$$\mathcal{P}(x|bg;\ m_{JJ} \in SR)$$

Construct the likelihood ratio: $$R(x) = \frac{\mathcal{P}(x|data;\ m_{JJ} \in SR)}{\mathcal{P}(x|bg;\ m_{JJ} \in SR)}$$

ANODE: Results on LHCO R&D Dataset

Nachman & DS 2001.04990

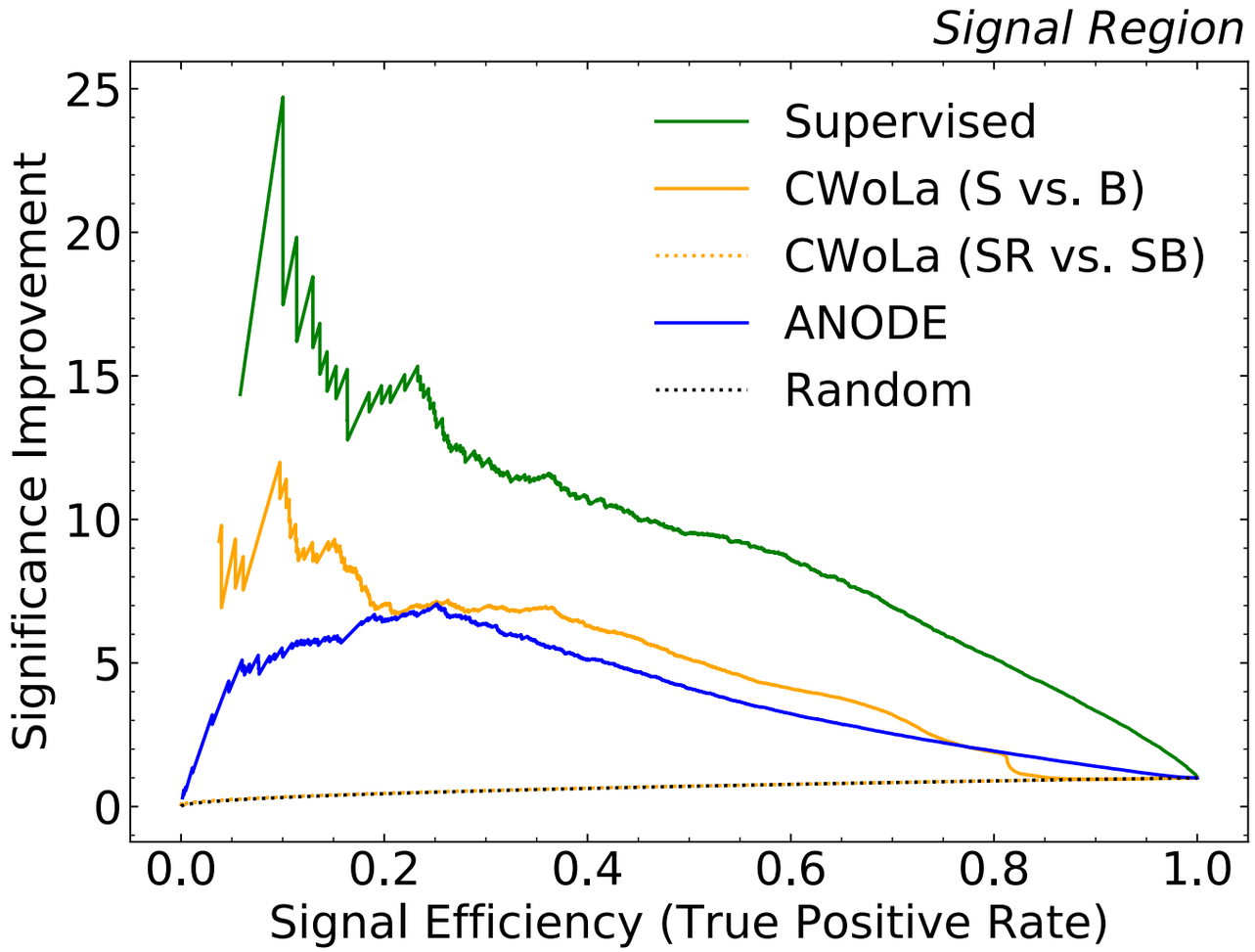




The method works! ANODE is sensitive to the signal!

ANODE: Results on LHCO R&D Dataset

Nachman & DS 2001.04990



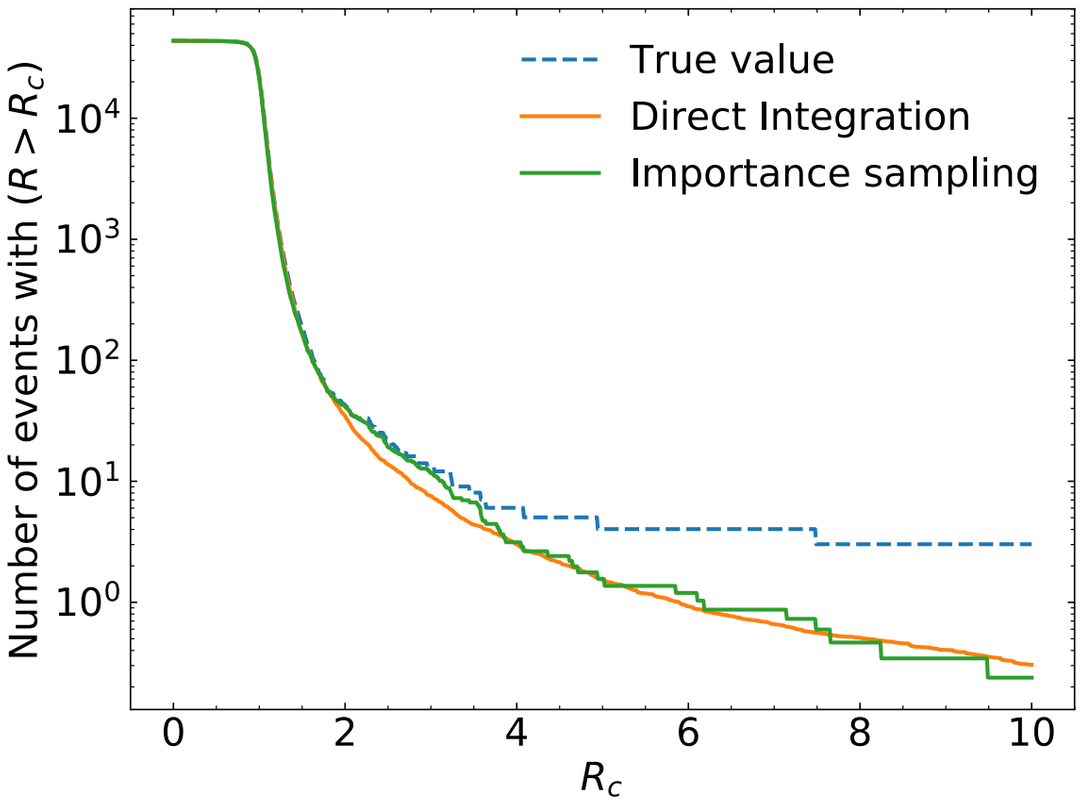
Can enhance the significance of the bump hunt by a factor of up to 7!

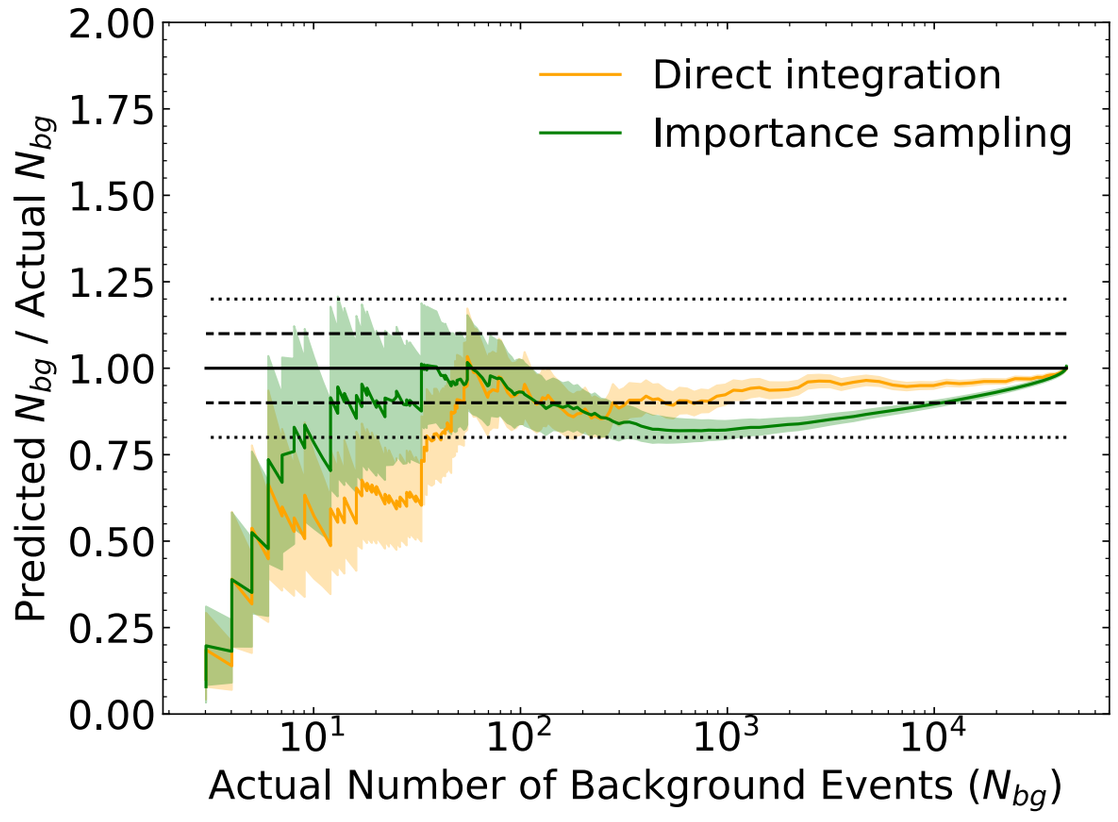
1.5σ (dijet bump hunt) => 10σ (ANODE+dijet bump hunt)

ANODE: Results on LHCO R&D Dataset

Nachman & DS 2001.04990

Novel aspect of ANODE: can estimate backgrounds directly with P(x|bg; m∈SR)



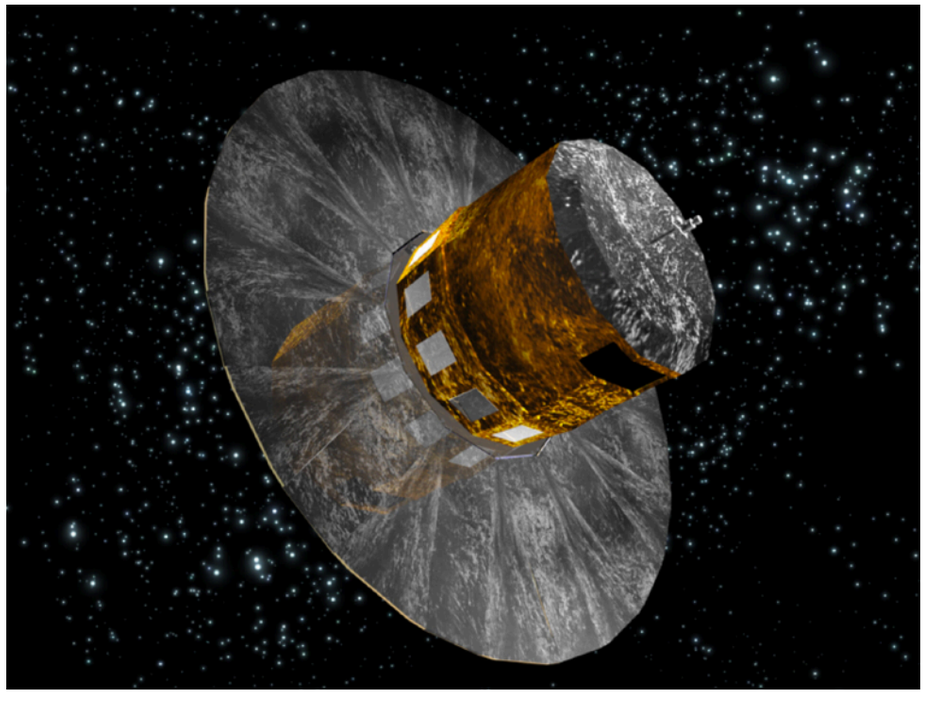


Via Machinae: ANODE IN SPACE

DS, Matt Buckley, Lina Necib & John Tamanas, in preparation

ANODE is a completely general method for identifying multivariate overdensities in data using sidebands. It could have many diverse applications.

We are currently using it to find **stellar streams** in Gaia data.

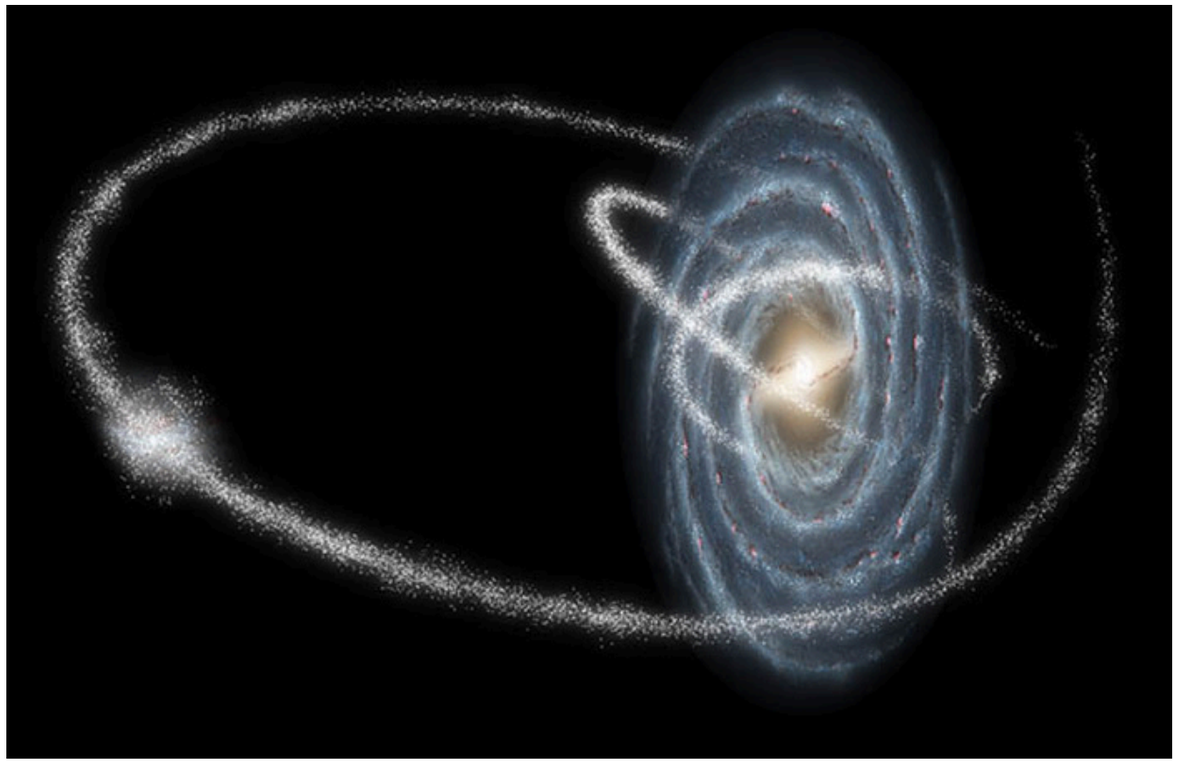


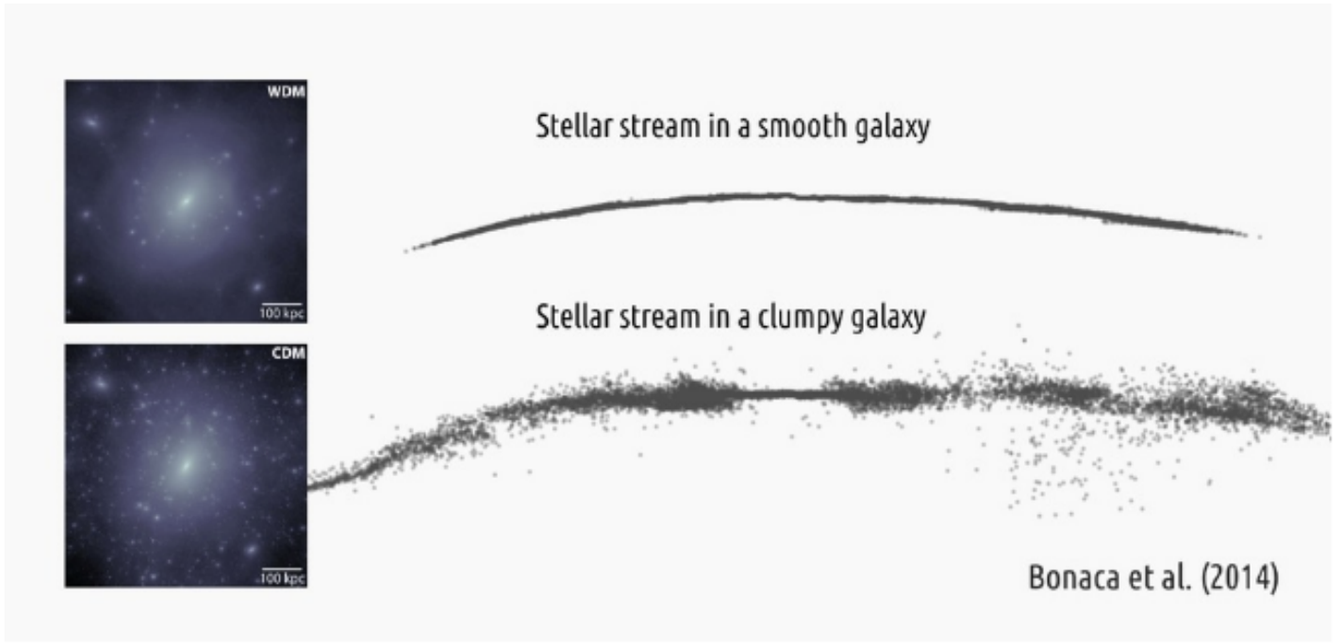
Gaia satellite:

- Launched in 2013; extended to 2025
- Mission: map out the full 6d phase space of the stars in our galaxy
- Angular positions, velocities, color and magnitude of over 1 billion stars in our galaxy
- radial positions and velocities for a smaller subset of nearby stars (not used in this work)

Stellar streams

Cold stellar streams are tidally-stripped remnants of globular clusters and dwarf galaxies, falling into and orbiting our galaxy.



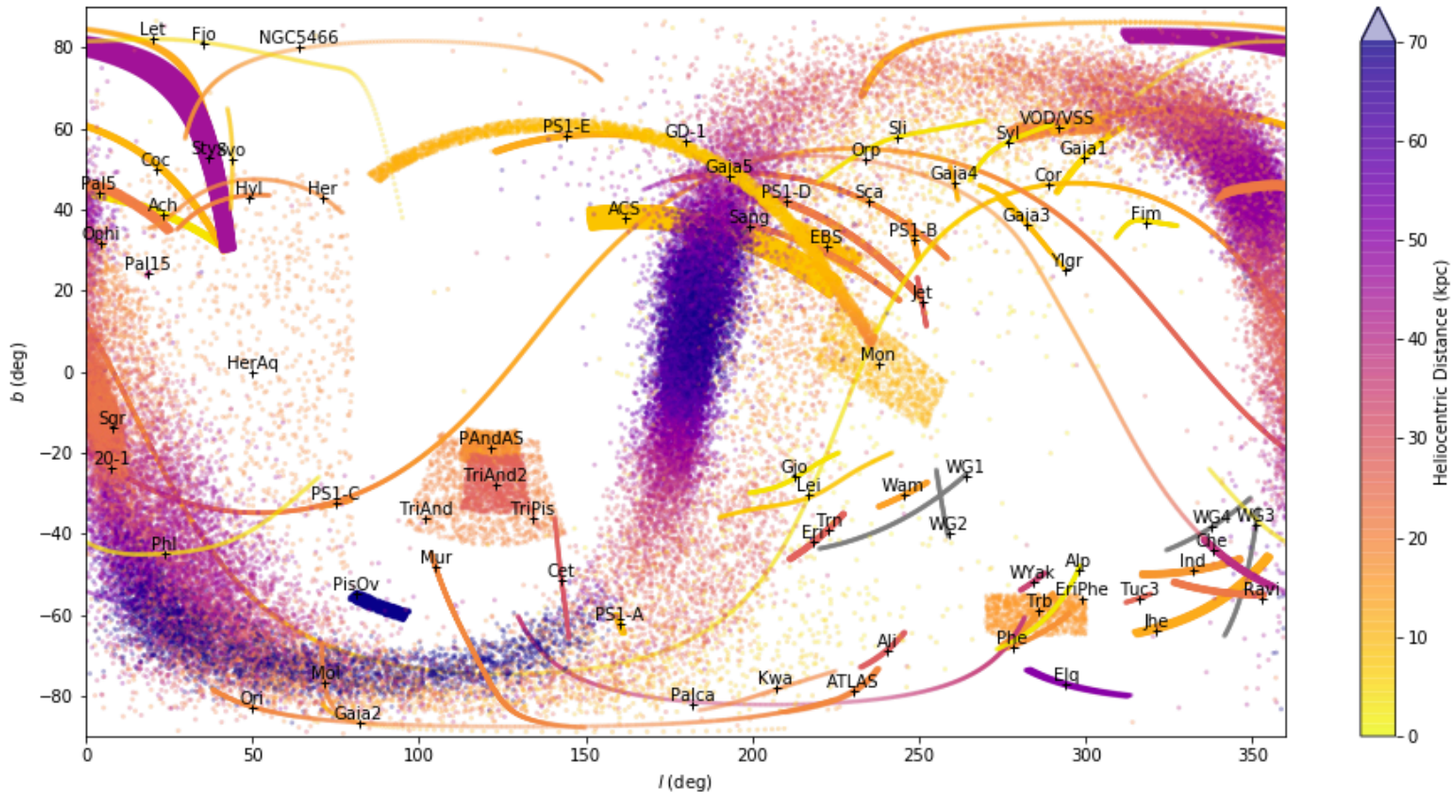


They are very interesting objects of study for astrophysicists and particle physicists.

In particular, they could be unique probes into dark matter substructure.

Stream finding

https://github.com/cmateu/galstreams

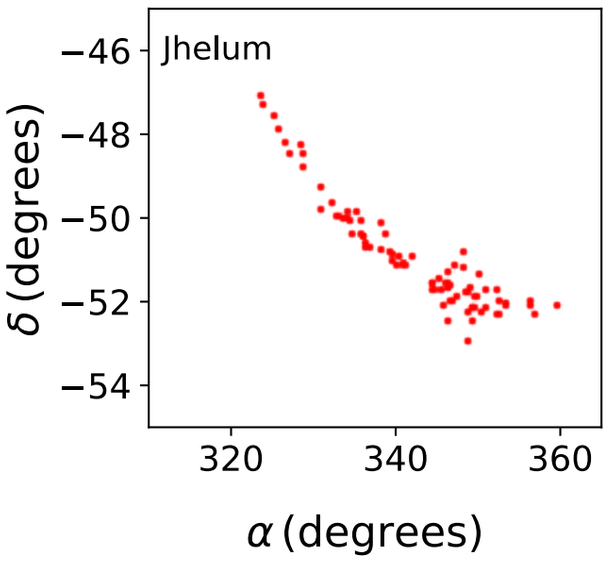


Existing methods (eg STREAMFINDER, Malhan & Ibata 2018) have found many new streams in the Gaia data, but they make a number of model-dependent assumptions (form of the galactic potential, orbits, isochrones, …).

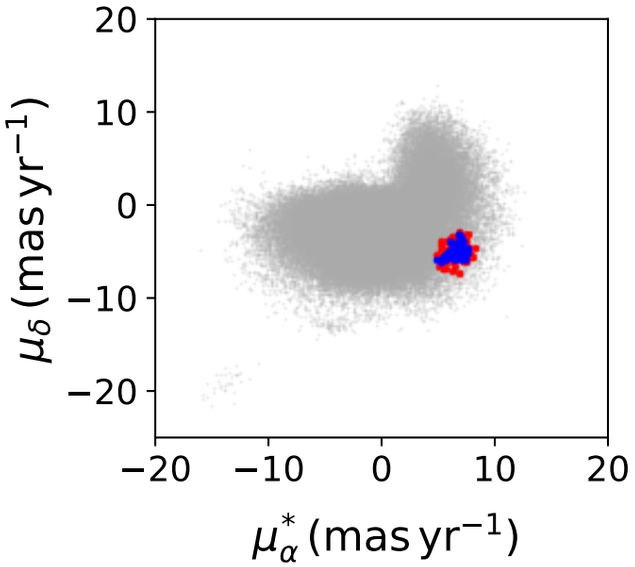
We were interested in whether unsupervised ML could be used to find streams more model-independently.

Stream finding

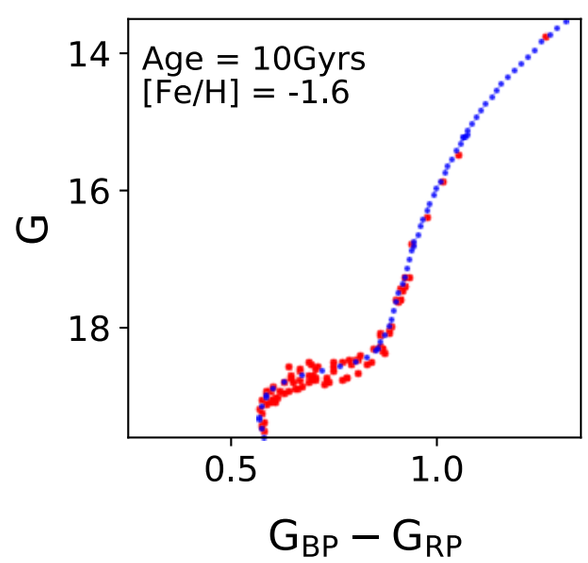
from Malhan et al 2018



position



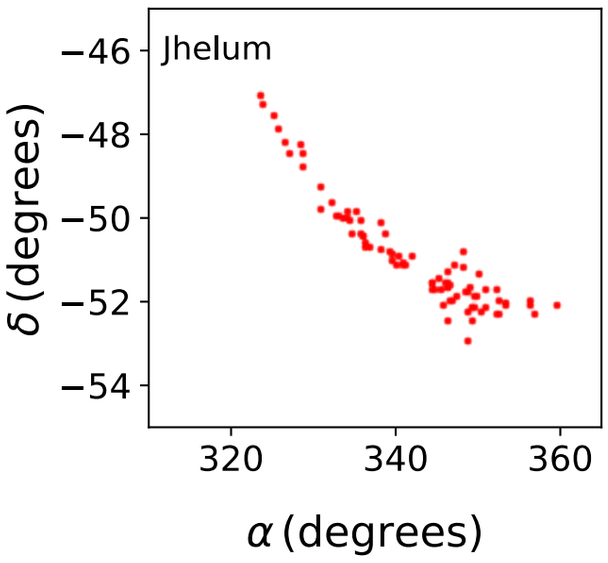
velocity



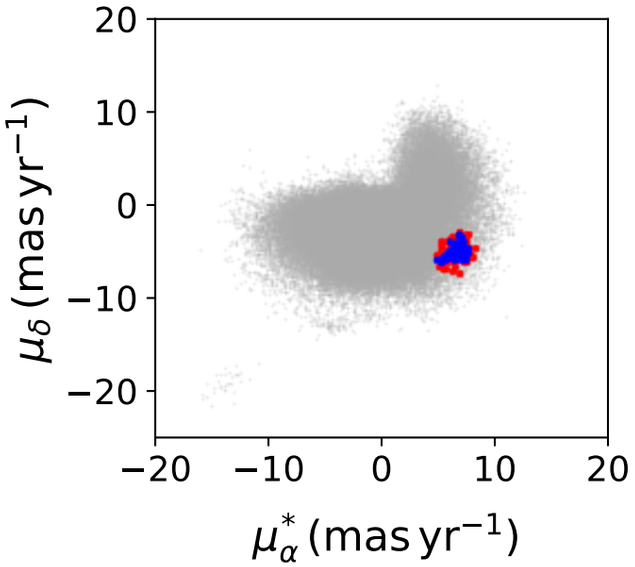
photometry

Stream finding

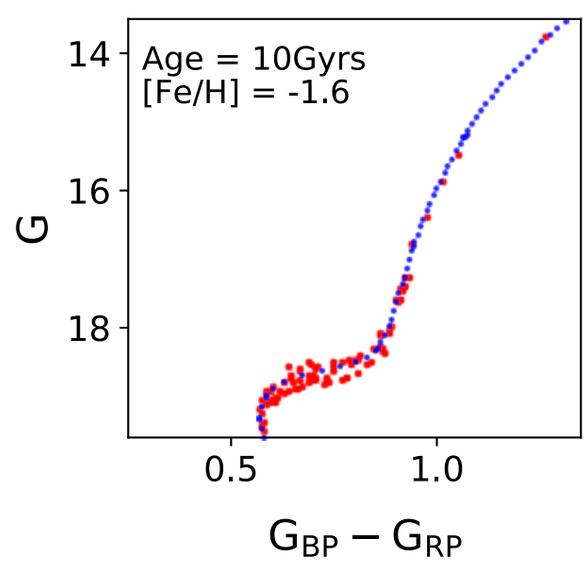
from Malhan et al 2018



position



velocity

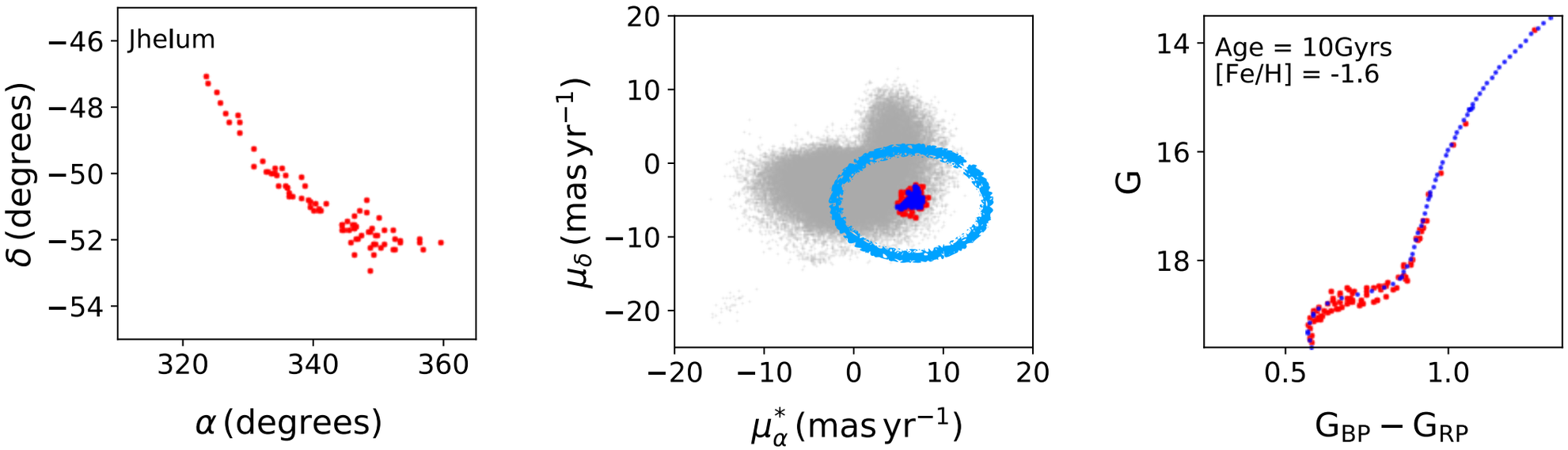


photometry

Idea: streams are local overdensities in position, velocity and photometric space.

Stream finding

from Malhan et al 2018



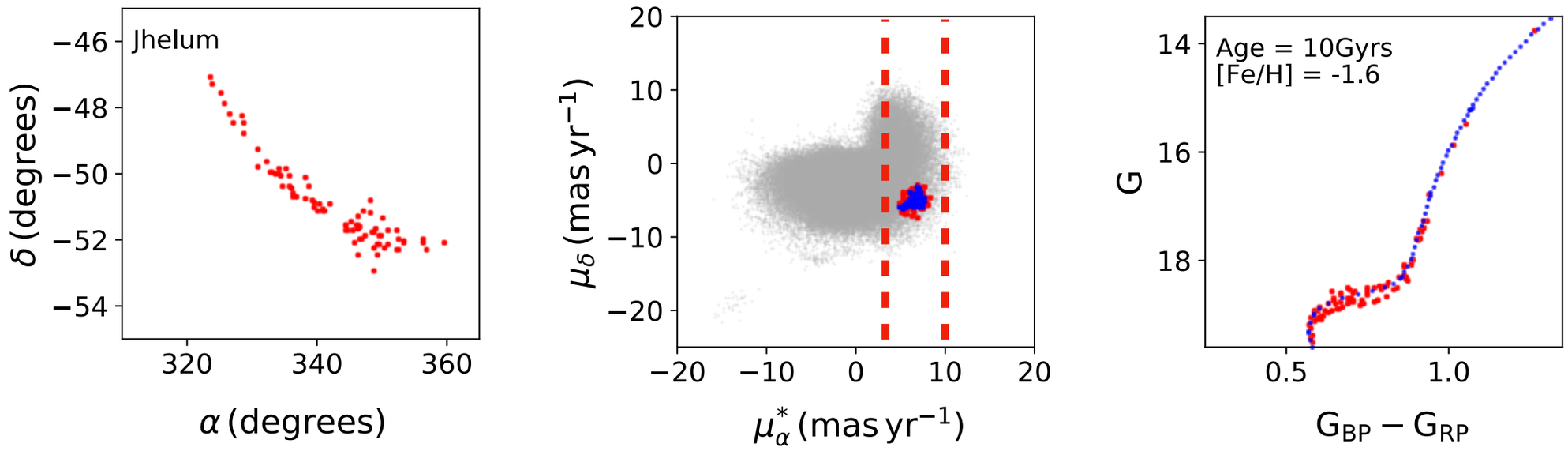
position velocity photometry

Idea: streams are local overdensities in position, velocity and photometric space.

Since they are **cold**, the stars in the stream are clustered in velocity.

Stream finding

from Malhan et al 2018



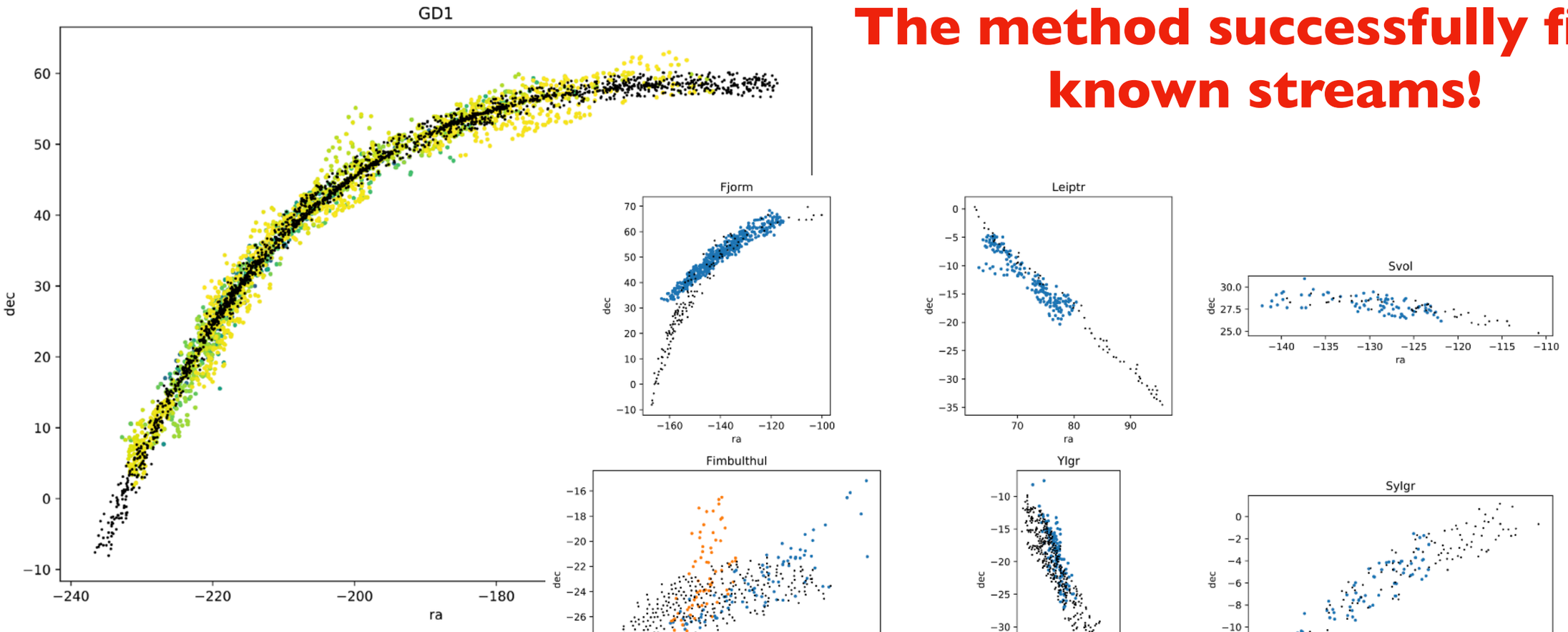
position velocity photometry

Idea: streams are local overdensities in position, velocity and photometric space.

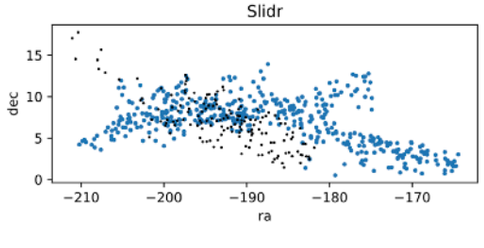
Since they are **cold**, the stars in the stream are clustered in velocity.

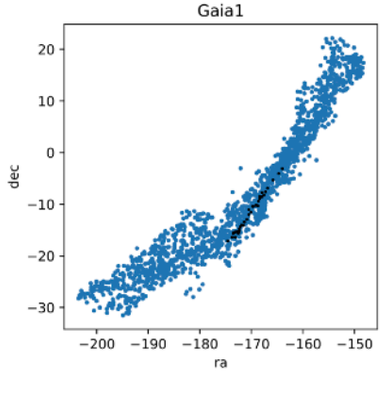
Can sideband in one of the velocities and use ANODE to look for local overdensities!

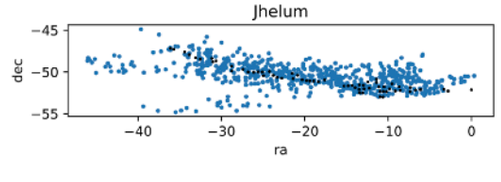
Via Machinae: preliminary results



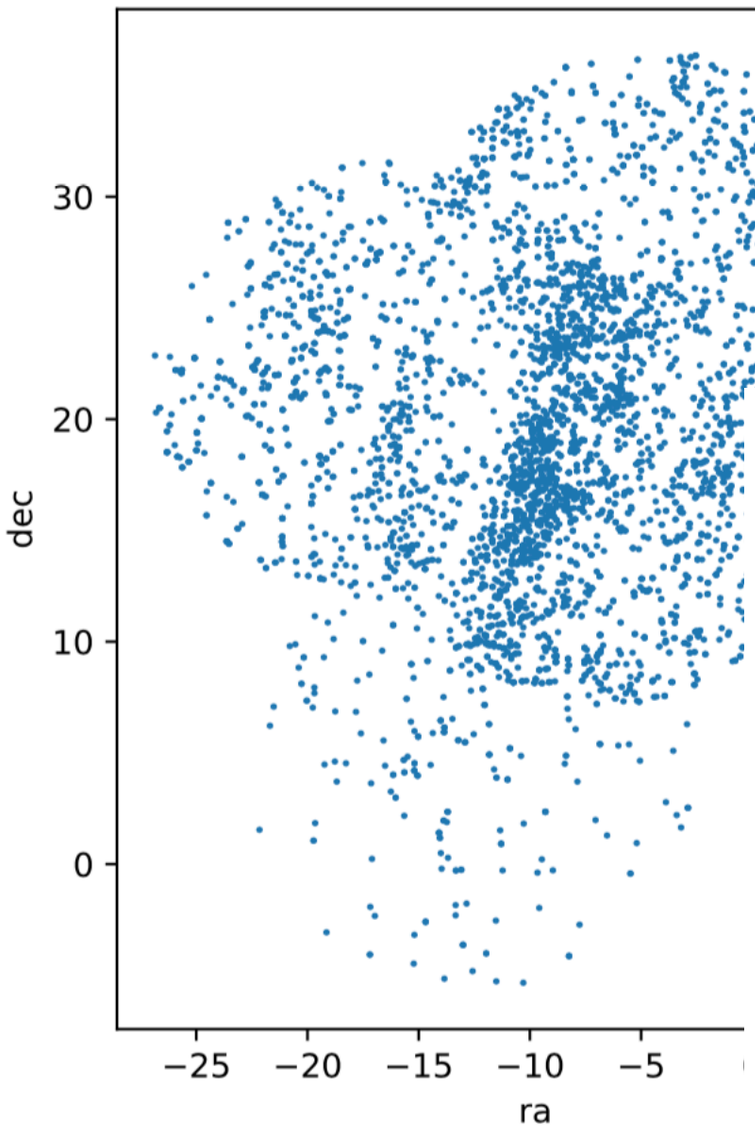
The method successfully finds known streams!



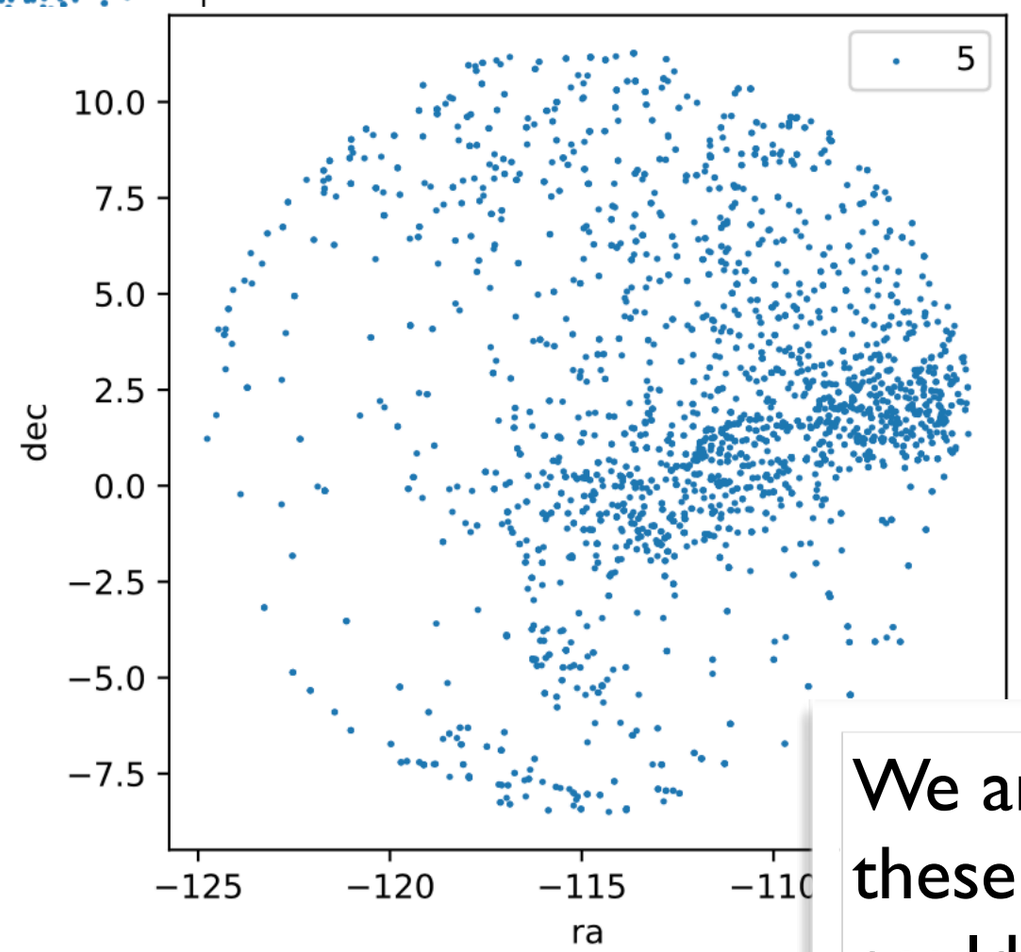


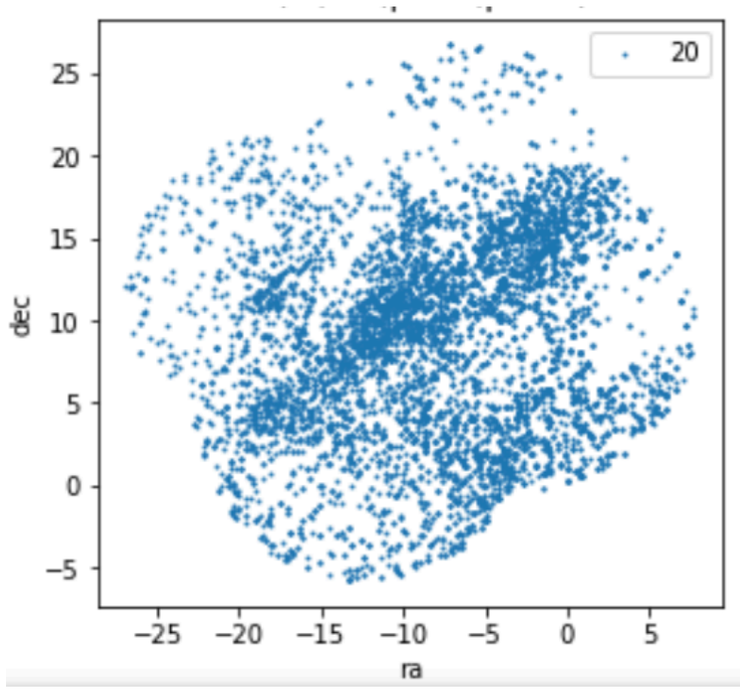


Via Machinae: preliminary results



The method also finds ~600 other stream candidates





We are currently investigating these further to see if they could be real. Stay tuned!

Summary and Outlook

- Advances in machine learning are opening up new and exciting avenues for model independent new physics searches at the LHC.

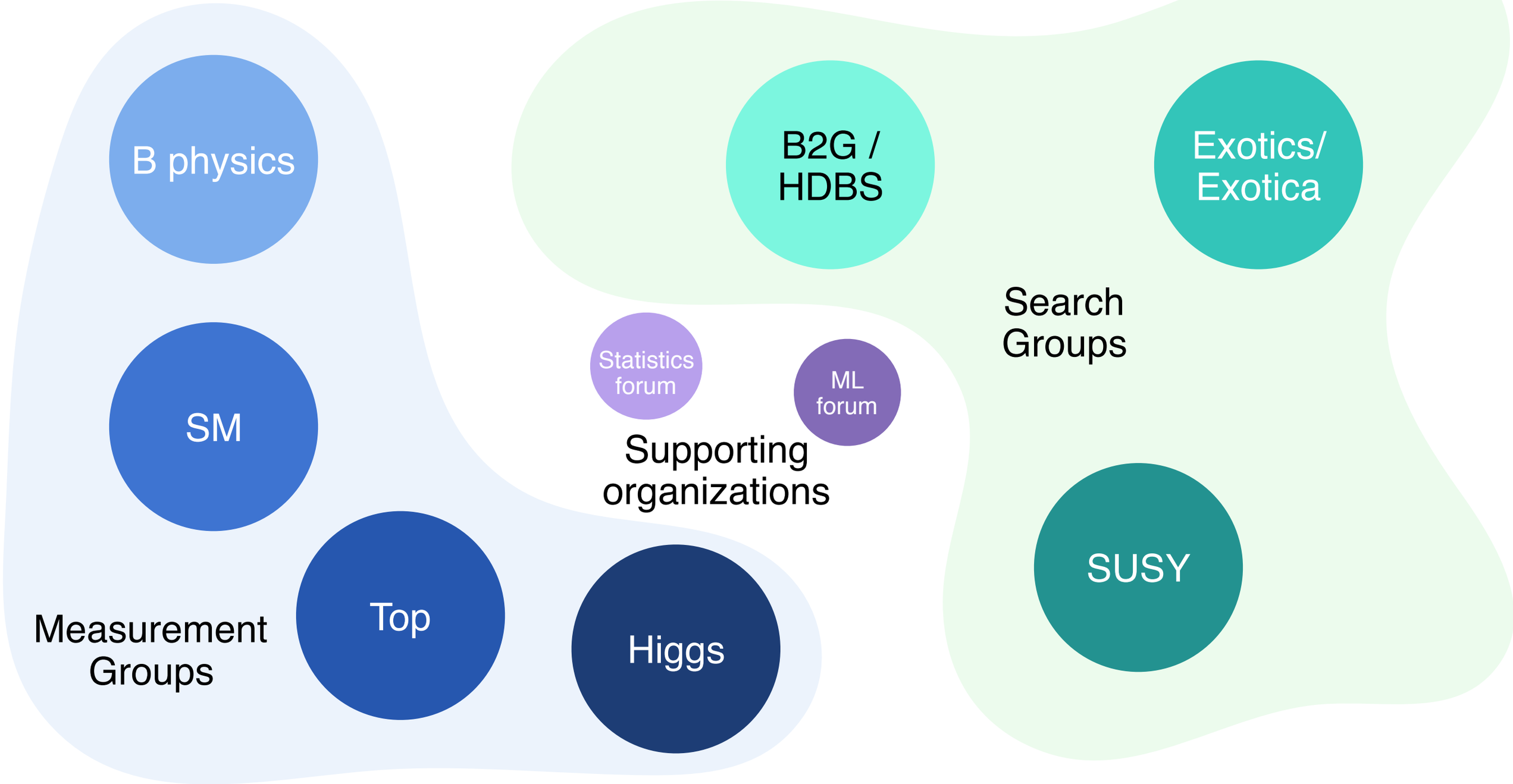
- Ideas and methods inspired by LHC problems are being applied successfully to other fields such as astronomy and astrophysics.

- The LHC Olympics 2020 provided a very useful testing ground for the development and common benchmarking of new approaches.

- Much work remains to be done in order to port these ideas over to ATLAS and CMS and implement them as actual analyses on real data.

- We need more ideas for model-independent searches at the LHC. This is just the beginning!

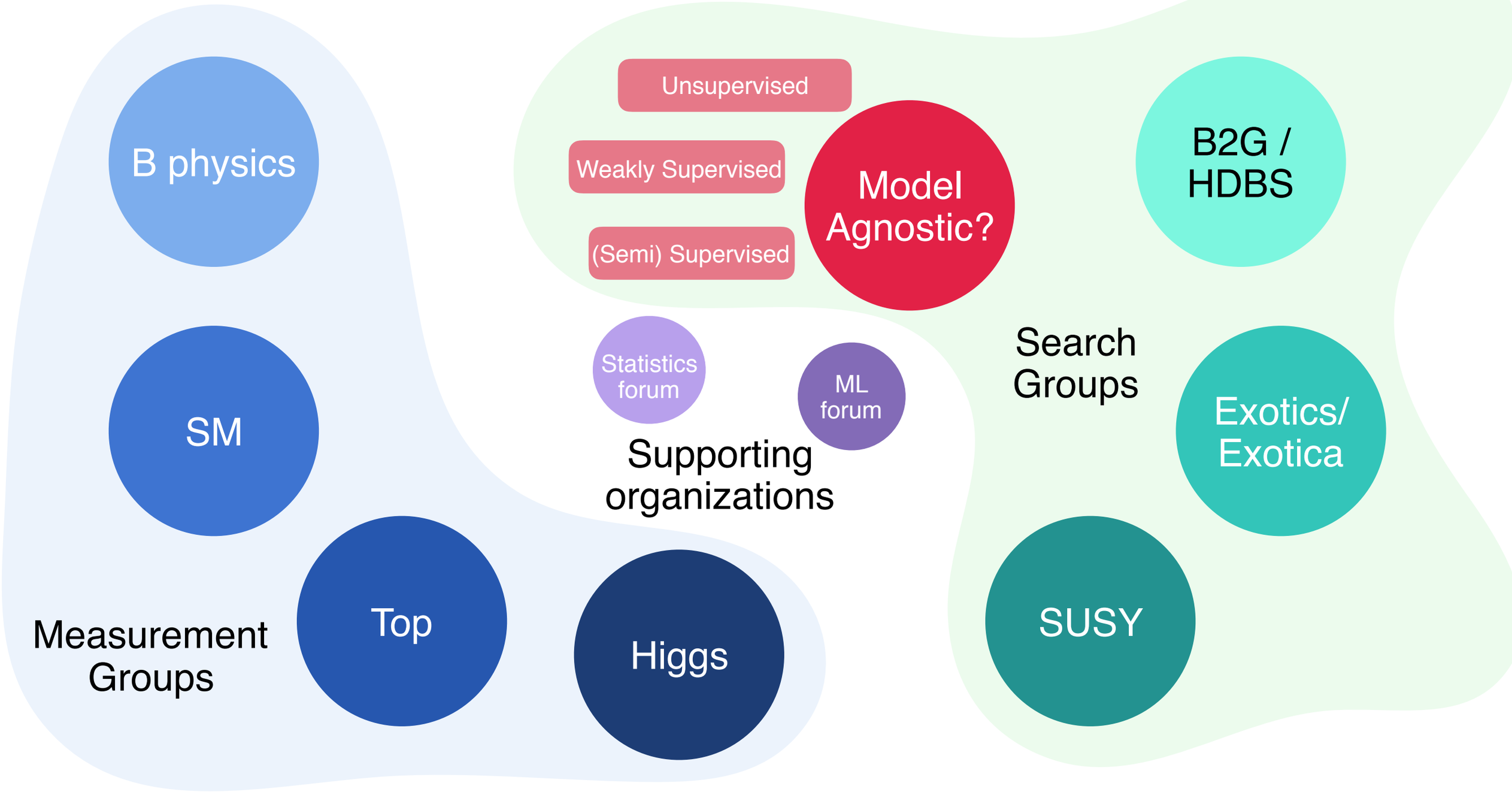
Current Organization of Physics Analysis Groups at the LHC



Q: Why is there no model independent search group???

A vision for the future...

Future Organization of Physics Analysis Groups at the LHC??



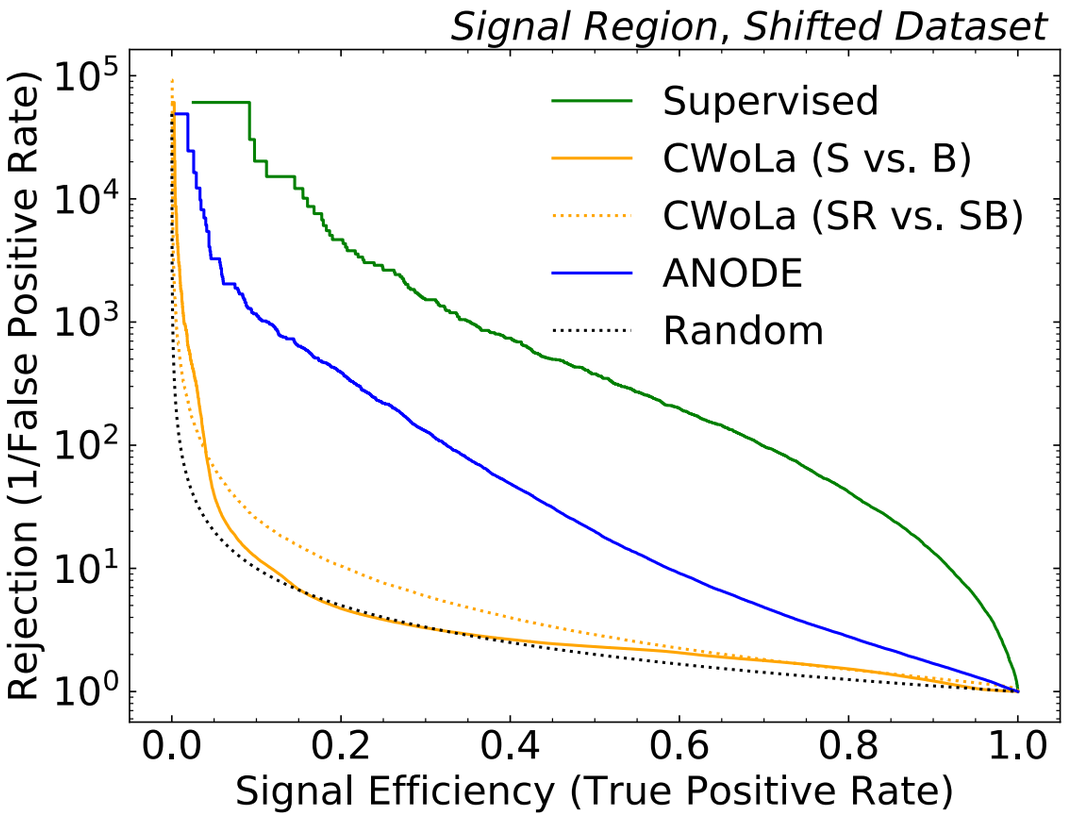
from G. Kasieczka, B. Nachman, DS (eds), et al 2101.nextweek

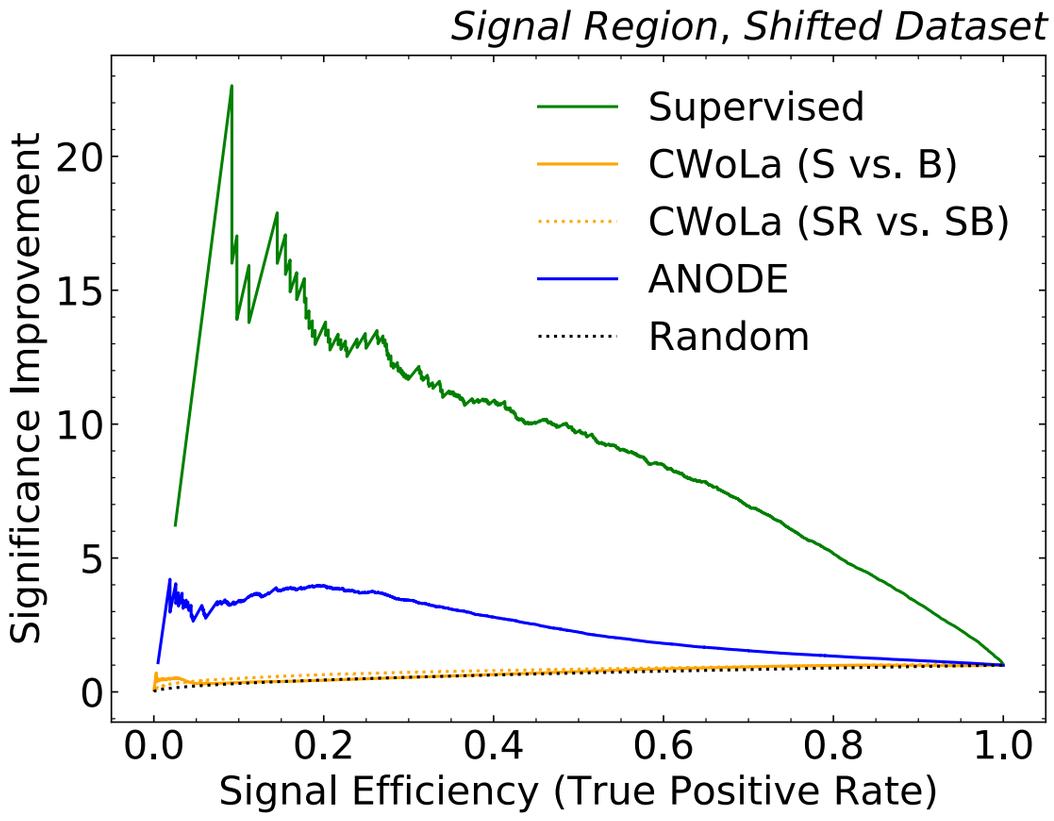
Thanks for your attention!

ANODE: Results on LHCO R&D Dataset

Ben Nachman & DS 2001.04990

Can also consider performance on a feature set which is not independent of m. We introduced artificial correlations just as proof of concept: $m_{J_{1,2}} \to m_{J_{1,2}} + c\, m_{JJ}$

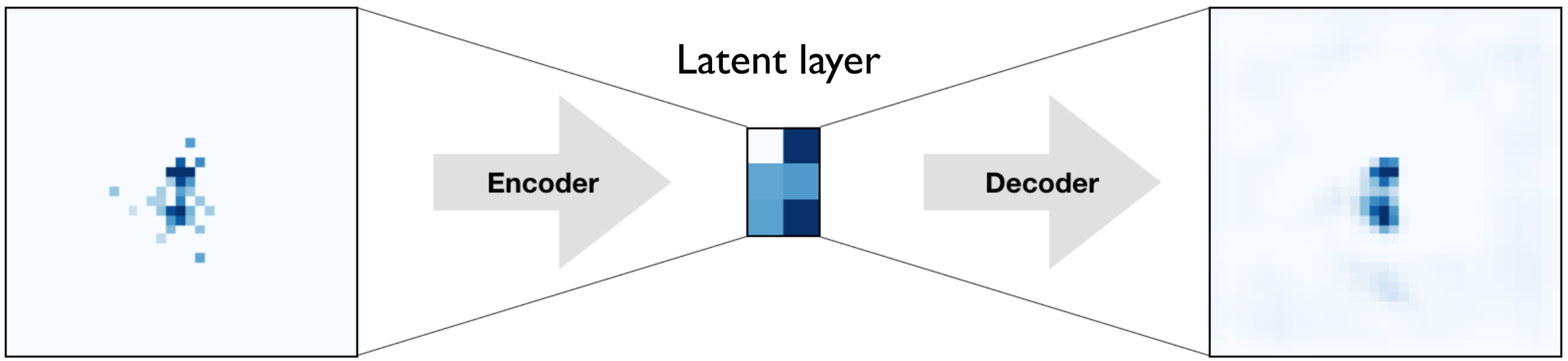




ANODE is robust while CWoLa completely fails!

Searching for NP with deep autoencoders

Heimel et al 1808.08979; Farina, Nakai & DS 1808.08992



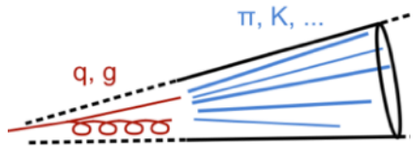
An autoencoder maps an input into a reduced "latent representation" and then attempts to reconstruct the original input from it.

Can use reconstruction error as an anomaly threshold!

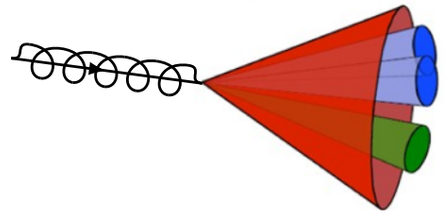
See also:
Hajer et al "Novelty Detection Meets Collider Physics" 1807.10261
Cerri et al "Variational Autoencoders for New Physics Mining at the Large Hadron Collider" 1811.10276

Performance

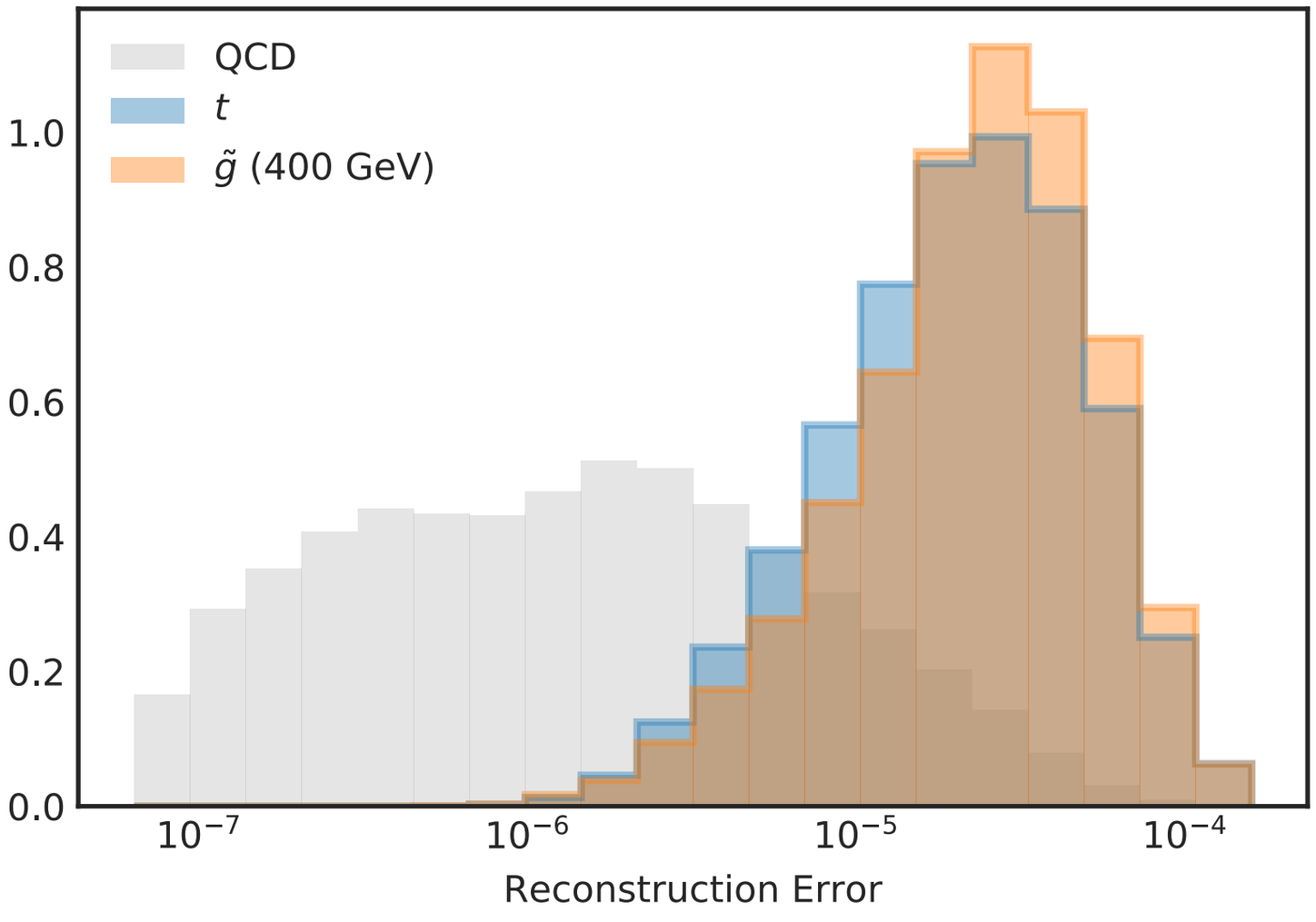
background (QCD)



signal (tops and 400 GeV RPV gluinos)



It works as an anomaly detector!



Robust against contamination with signal — can use in fully unsupervised mode