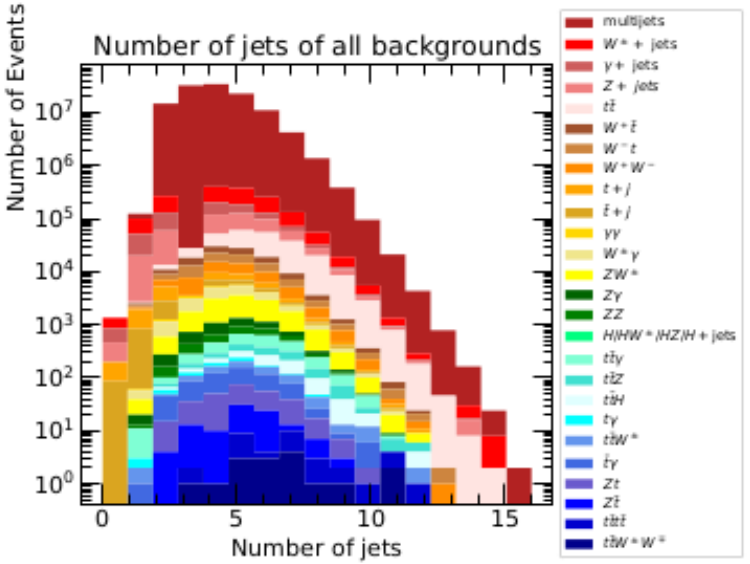
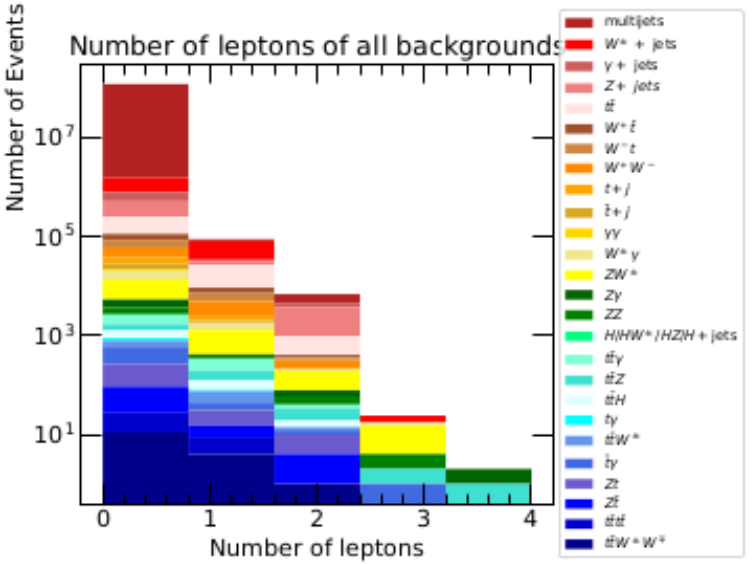




The Dark Machines Anomaly Score Challenge I

Based on arXiv: 2105.14027

https://github.com/bostdiek/DarkMachines-UnsupervisedChallenge

Joe Davies
Queen Mary University of London

Paper Authors

The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider

T. Aarrestad[CERN] M. van Beekveld[Ox] M. Bona[QMUL] A. Boveia[OSU]
S. Caron[HEF, Nikhef] J. Davies[QMUL] A. De Simone[SISSA, INFN] C. Doglioni[Lund]
J.M. Duarte[UCSD] A. Farbin[UnivArlington] H. Gupta[GSoC] L. Hendriks[HEF, Nikhef]
L. Heinrich[CERN] J. Howarth[Glasgow] P. Jawahar[WPI, CERN] A. Jueid[UnivKonkuk]
J. Lastow[Lund] A. Leinweber[UnivAdelaide] J. Mamuzic[IFIC] E. Merényi[UnivRice]
A. Morandini[RWTH] P. Moskvitina[HEF, Nikhef] C. Nellist[HEF, Nikhef]
J. Ngadiuba[FNAL, Caltech] B. Ostdiek[Harvard, AIFI] M. Pierini[CERN] B. Ravina[Glasgow]
R. Ruiz de Austri[IFIC] S. Sekmen[KNU] M. Touranakou[NKUA, CERN]
M. Vaškevičiūte[Glasgow] R. Vilalta[UnivHouston] J.-R. Vlimant[Caltech] R. Verheyen[UCL]
M. White[UnivAdelaide] E. Wulff[Lund] E. Wallin[Lund] K.A. Wozniak[UniVie, CERN]
Z. Zhang[HEF, Nikhef]

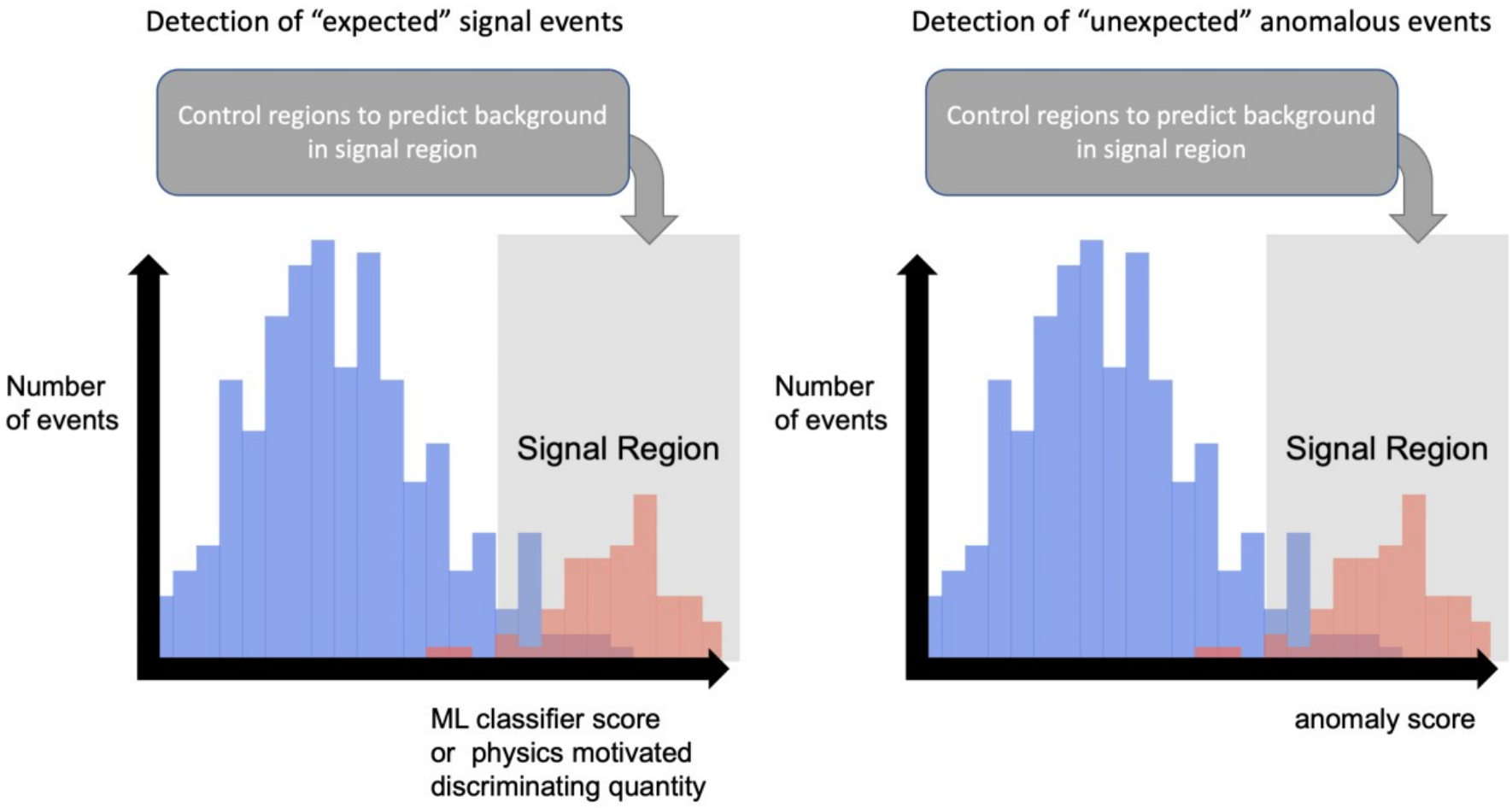
Challenge Justification and Goals

- Goal is to perform model-agnostic searches
- General searches (without explicit BSM signal assumption) already performed by
  - DØ Collaboration at Tevatron using SLEUTH
  - H1 Collaboration at HERA using 1-D signal detection algorithm
  - CDF Collaboration at Tevatron (using similar to above)
- ‘Bump’ hunting searches for localized excesses in events often used in these searches
  - Machine Learning can perform these hunts using anomaly detection techniques that have become more sophisticated

Challenge Outline

- Dataset of > 1 Billion SM Events used to train ML models
  - https://zenodo.org/record/3685861
- Hackathon Dataset: (https://zenodo.org/record/3961917)
  - 4 different channels (selection cuts)
  - 11 different BSM signals
  - 19 total mass spectra
  - 34 unique signal/channel combinations
- Train each method 4 times (once per channel) using SM
- Select ML methods which perform best to apply to blinded Secret Dataset: https://zenodo.org/record/4443151

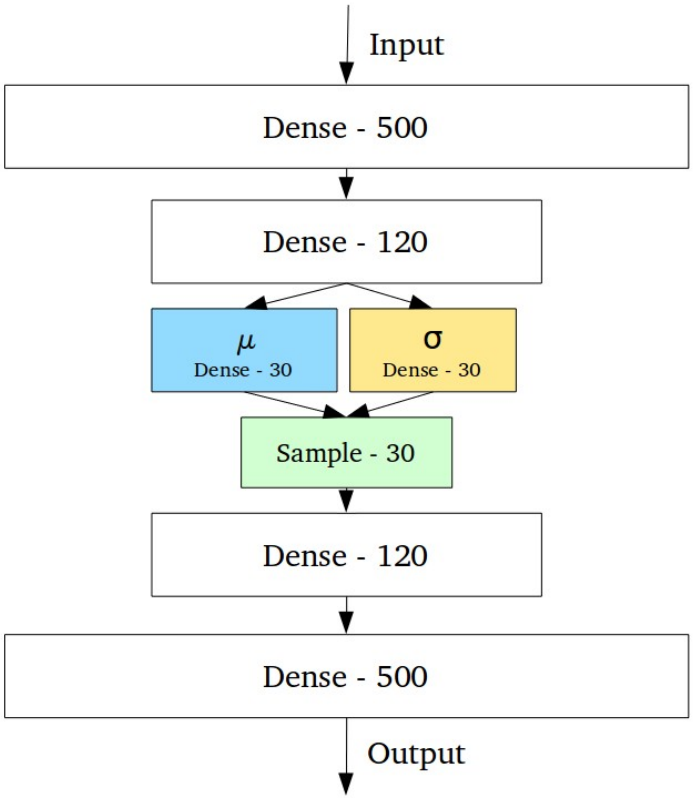
General Strategy



Use fixed cuts for background rejections of 10^{-2}, 10^{-3}, and 10^{-4}

Variational Autoencoder

- **Same structure as an AE except the latent space is continuous by design**
- **Sampling can be done on latent vectors to produce a continuous set of outputs**
- **(Generally) MSE + Kullback-Liebler divergence used as error**



$$\sum_{i=1}^{N} \frac{1}{2}(t_i - y_i)^2$$

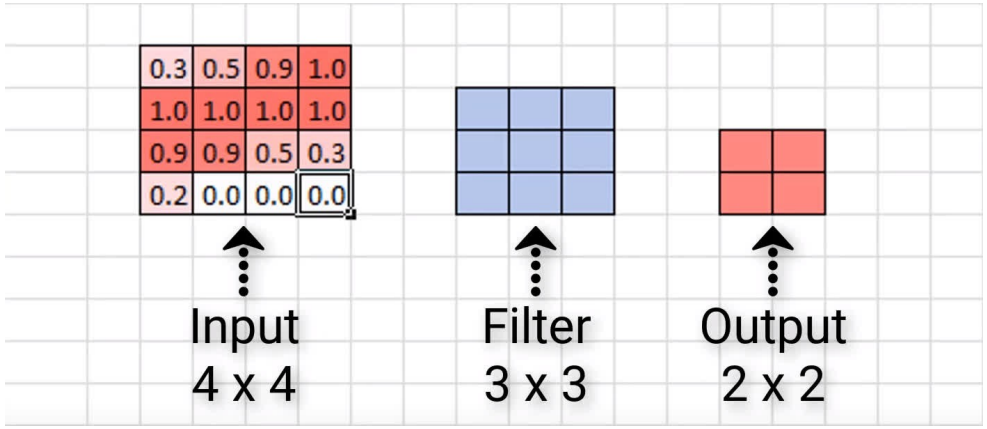
Typical MSE Error

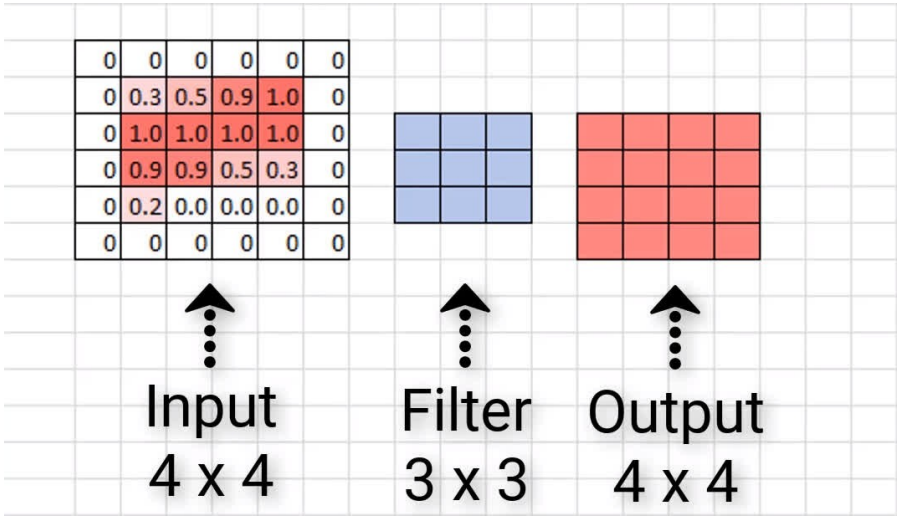
$$\sum_{i=1}^{n} \sigma_i^2 + \mu_i^2 - \log(\sigma_i) - 1$$

KL-Divergence

Challenges with the VAE

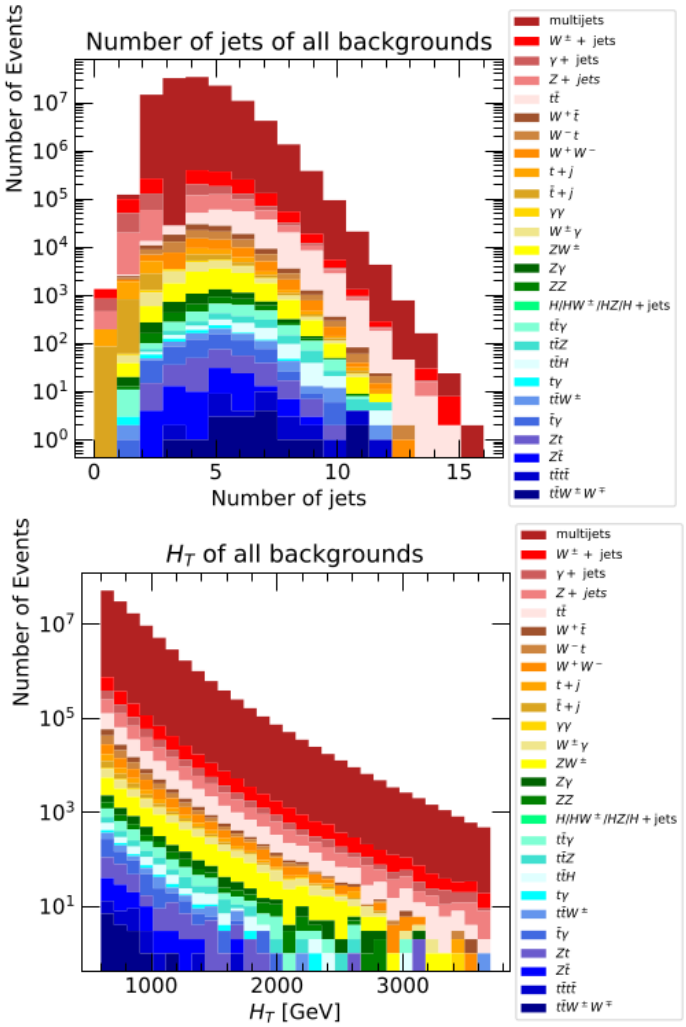
- Should the events be zero padded?
- Should we take a smaller number of objects?
- Which anomaly score to use:
  - Just one or the other of reconstruction or KL
  - Radius in the latent space
  - Beta parameters (and how to tweak them)





The Datasets

SM processes			
Physics process	Process ID	σ (pb)	N_{tot} ($N_{10\,\text{fb}^{-1}}$)
$pp \to jj(+2j)$	njets	$19718_{H_T>600\text{GeV}}$	415331302 (197179140)
$pp \to l^{\pm}\nu_l(+2j)$	w_jets	$10537_{H_T>100\text{GeV}}$	135692164 (105366237)
$pp \to \gamma j(+2j)$	gam_jets	$7927_{H_T>100\text{GeV}}$	123709226 (79268824)
$pp \to l^+l^-(+2j)$	z_jets	$3753_{H_T>100\text{GeV}}$	60076409 (37529592)
$pp \to t\bar{t}(+2j)$	ttbar	541	13590811 (5412187)
$pp \to t + \text{jets}(+2j)$	single_top	130	7223883 (1297142)
$pp \to \bar{t} + \text{jets}(+2j)$	single_topbar	112	7179922 (1116396)
$pp \to W^+W^-(+2j)$	ww	82.1	17740278 (821354)
$pp \to W^{\pm}t(+2j)$	wtop	57.8	5252172 (577541)
$pp \to W^{\pm}\bar{t}(+2j)$	wtopbar	57.8	4723206 (577541)
$pp \to \gamma\gamma(+2j)$	2gam	47.1	17464818 (470656)
$pp \to W^{\pm}\gamma(+2j)$	Wgam	45.1	18633683 (450672)
$pp \to ZW^{\pm}(+2j)$	zw	31.6	13847321 (315781)
$pp \to Z\gamma(+2j)$	Zgam	29.9	15909980 (299439)
$pp \to ZZ(+2j)$	zz	9.91	7118820 (99092)
$pp \to h(+2j)$	single_higgs	1.94	2596158 (19383)
$pp \to t\bar{t}\gamma(+2j)$	ttbarGam	1.55	95217 (15471)
$pp \to t\bar{t}Z$	ttbarZ	0.59	300000 (5874)
$pp \to t\bar{t}h(+1j)$	ttbarHiggs	0.46	200476 (4568)
$pp \to \gamma t(+2j)$	atop	0.39	2776166 (3947)
$pp \to t\bar{t}W^{\pm}$	ttbarW	0.35	279365 (3495)
$pp \to \gamma\bar{t}(+2j)$	atopbar	0.27	4770857 (2707)
$pp \to Zt(+2j)$	ztop	0.26	3213475 (2554)
$pp \to Z\bar{t}(+2j)$	ztopbar	0.15	2741276 (1524)
$pp \to t\bar{t}t\bar{t}$	4top	0.0097	399999 (96)
$pp \to t\bar{t}W^+W^-$	ttbarWW	0.0085	150000 (85)



Madgraph+Pythia+Delphes | jets, b-jets, electrons, muons, photons

The Datasets

Channel 1: 214,185 SM Events

- $H_T \geq 600$ GeV
- MET ≥ 200 GeV
- $MET/H_T \geq 0.2$
- At least 4 (b)-jets with $p_T > 50$ GeV
- At least 1 (b)-jets with $p_T > 200$ GeV

Channel 2b: 340,268 SM Events

- $H_T \geq 50$ GeV
- MET ≥ 50 GeV
- At least 2 μ/e with $p_T > 15$ GeV

Channel 2a: 20,005 SM Events

- MET ≥ 50 GeV
- At least 3 μ/e with $p_T > 15$ GeV
- At least 1 (b)-jets with $p_T > 200$ GeV
- Few training events, many ML methods struggle

Channel 3: 8,544,111 SM Events

- $H_T \geq 600$ GeV
- MET ≥ 100 GeV
- Large dataset, timed out training on some methods

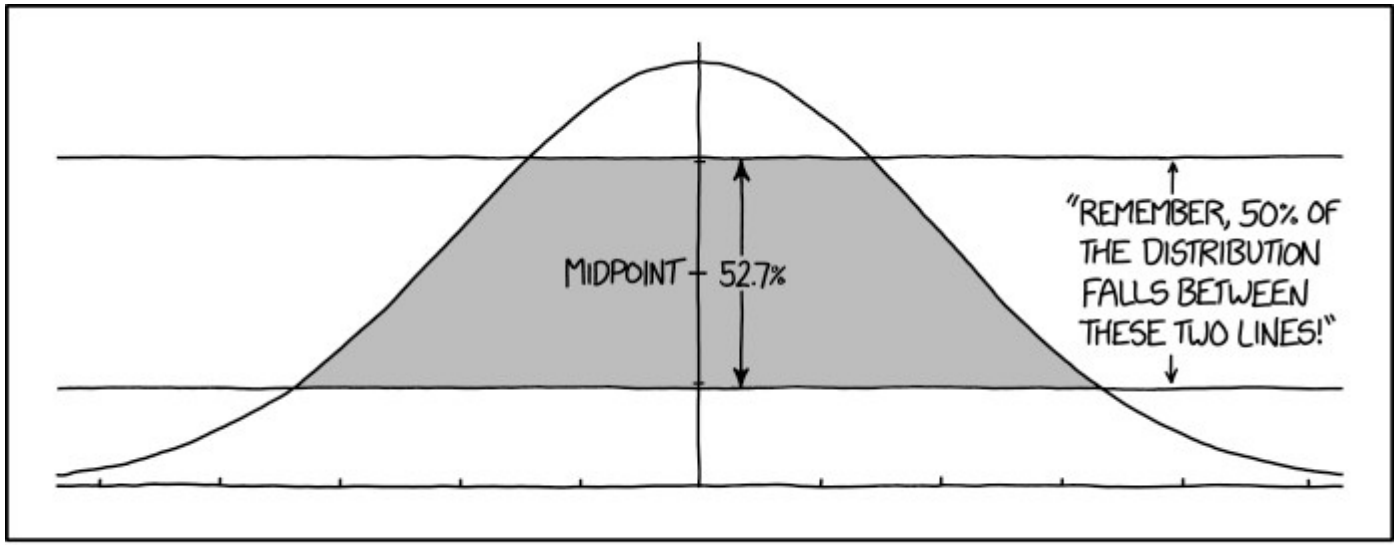
The BSM Physics

BSM process	Channel 1	Channel 2a	Channel 2b	Channel 3
Z' + monojet	×	×		×
$Z' + W/Z$				×
Z' + single top	×			×
Z' in lepton-violating $U(1)_{L_\mu - L_\tau}$		×	×	
$\not{R}$-SUSY stop-stop	×		×	×
$\not{R}$-SUSY squark-squark	×			×
SUSY gluino-gluino	×	×	×	×
SUSY stop-stop	×			×
SUSY squark-squark	×			×
SUSY chargino-neutralino		×	×	
SUSY chargino-chargino			×	

Some processes have different mass spectra or decay modes: 19 signals, 34 Signal-Channel combinations

Conclusion

- **Model-agnostic searches**
- **Primarily use Variational Auto-Encoders**
- **Variety of channels and signals**
- **Stick around to find out about the results!**



HOW TO ANNOY A STATISTICIAN