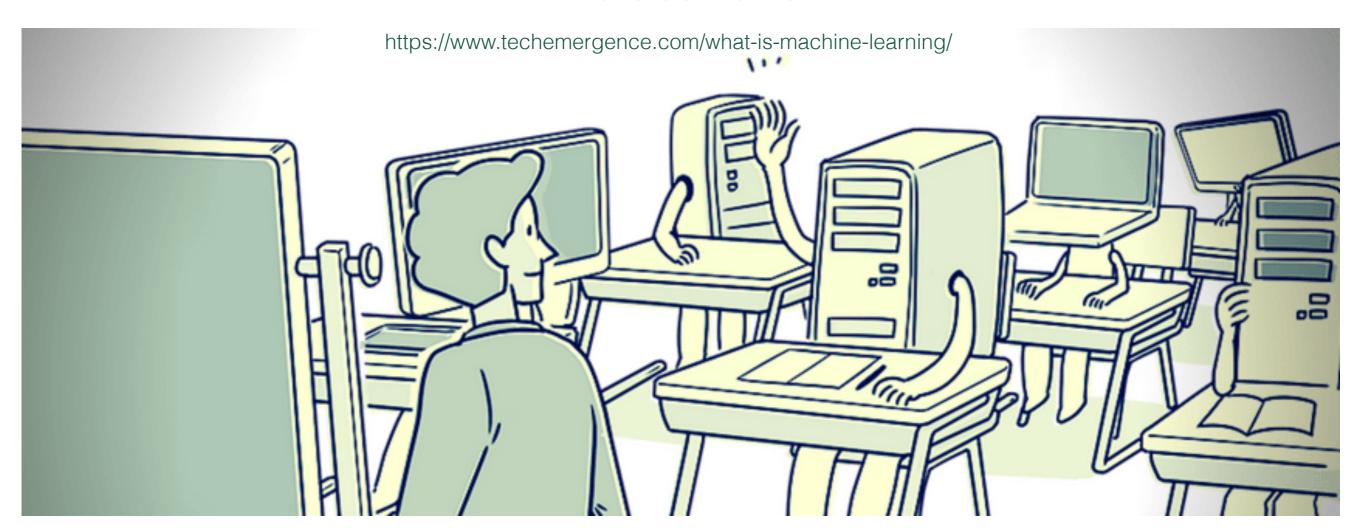
Introduction to Machine Learning

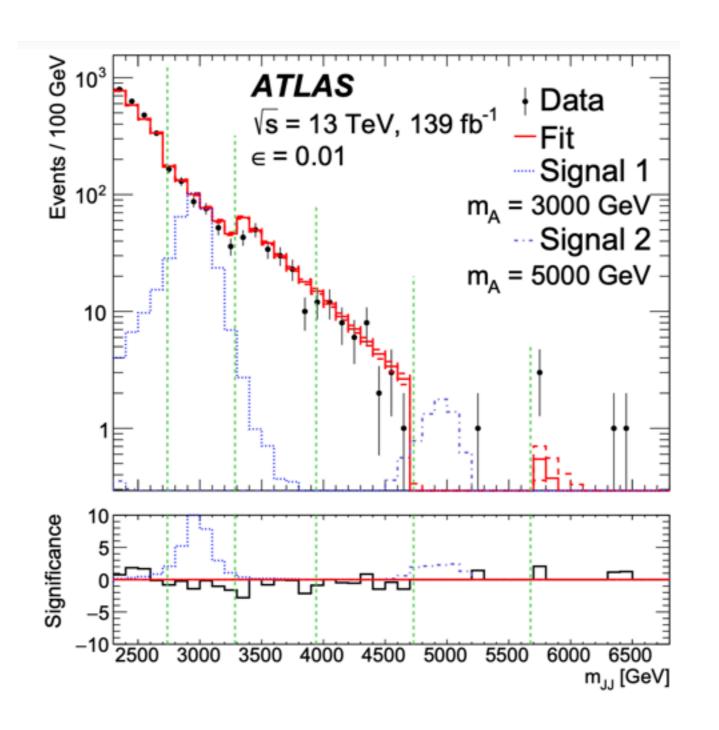
Lecture 3



QUC Winter School 2021 KIAS December 2021

Questions from last lecture?

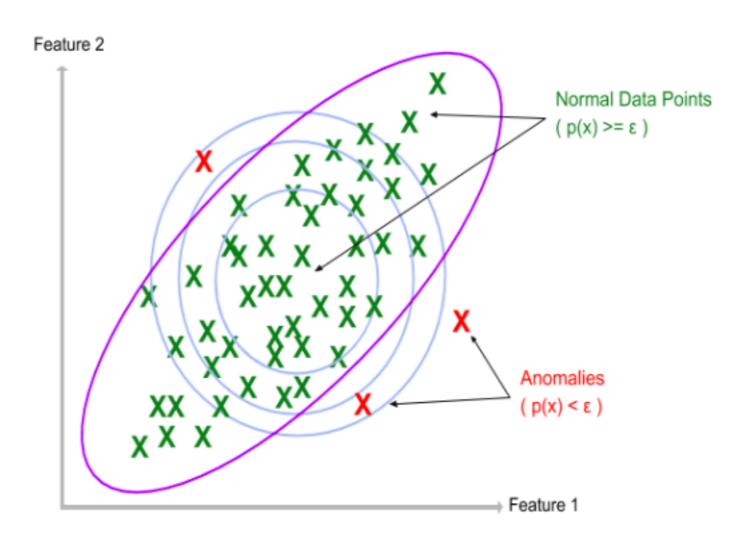
ATLAS search using weak supervision [2005.02983]



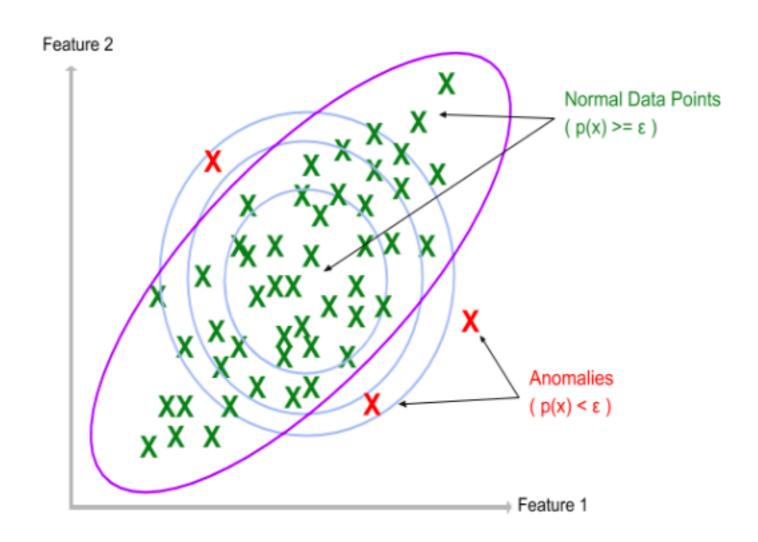
Start of lecture 3

- Anomaly Detection
- Generative Networks
- Flows for sampling/inference

What is Anomaly Detection?

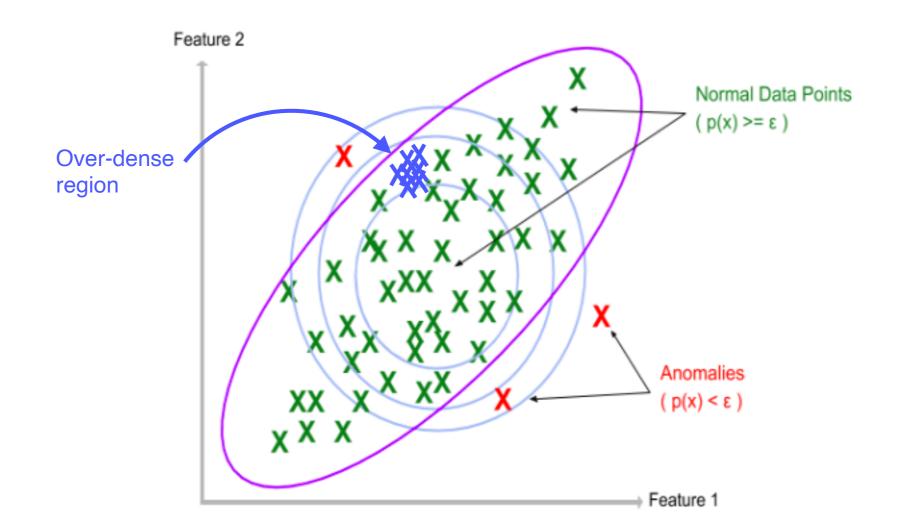


What is Anomaly Detection?



1) Events that occur with low probability (look different in some feature space)

What is Anomaly Detection?



- 1) Events that occur with low probability (look different in some feature space)
- 2) Events that don't look very different, but occur at a higher rate than expected

ATLAS Exotics Searches* - 95% CL Upper Exclusion Limits

	AILAS Preliminary
$\int \mathcal{L} dt = (3.2 - 139) \text{ fb}^{-1}$	$\sqrt{s} = 8$, 13 TeV
	Reference

63.004	tus: May 2020					J.C.o.	$t = (3.2 - 139) \text{ fb}^{-1}$	\sqrt{s} = 8, 13 TeV
	Model	ℓ, γ	Jets†	E _{miss}	∫£ dt[fb			Reference
Extra dimensions	ADD $G_{WX}+g/q$ ADD non-resonant $\gamma\gamma$ ADD OBH ADD BH high $\sum \rho\gamma$ ADD BH multiet RS1 $G_{WX}\to\gamma\gamma$ Bulk RS $G_{WX}\to WW/ZZ$ Bulk RS $G_{WX}\to WV\to \ell \gamma qq$ Ouk RS $g_{WX}\to \ell t$ SUED / RPP		1-4] - 2j 2 2] 3 3 j - 8 2 j / 1 J 2 1 6, 2 1 J 2 2 h, 2 3	(2) Yes	36.1 36.7 37.0 3.2 3.6 36.7 36.1 139 36.1 36.1	M, 8.0 M, 8.2	TeV n = 3 HL2 ML0 TeV n = 6	1711.03901 1707.04147 1703.09127 1606.02366 1512.02586 1707.04147 1800.02380 2004.14636 1894.10329 1800.09678
Gauge bosons	$\begin{array}{l} \operatorname{SSM} Z' \to \ell\ell \\ \operatorname{SSM} Z' \to \tau\tau \\ \operatorname{Leptophobic} Z' \to bb \\ \operatorname{Leptophobic} Z' \to tt \\ \operatorname{SSM} W' \to \ell\tau \\ \operatorname{SSM} W' \to \tau\tau \\ \operatorname{HYT} W' \to WZ \to \ell \nu \operatorname{reg} \operatorname{model} \\ \operatorname{HYT} V' \to WV \to \operatorname{qqq} \operatorname{model} \\ \operatorname{HYT} V' \to WH/ZH \operatorname{model} B \\ \operatorname{HYT} W' \to WH \operatorname{model} B \\ \operatorname{HYT} W' \to WH \operatorname{model} B \\ \operatorname{LPSM} W_R \to tb \\ \operatorname{LPSM} W_R \to tb \\ \operatorname{LPSM} W_R \to \mu N_R \end{array}$	1 e,μ 1 τ 8 1 e,μ 8 0 e,μ multi-channe	$\geq 1b_i \geq 2$	Yes Yes Yes	159 36.1 36.1 139 138 36.1 138 36.1 138 36.1 138	Z' mass 5.1 TeV Z' mass 2.42 TeV Z' mass 2.1 TeV Z' mass 4.1 TeV W' mass 6.0 TeV W' mass 3.7 TeV W' mass 4.3 TeV Y' mass 3.8 TeV W' mass 2.93 TeV W' mass 3.2 TeV W ₁ mass 3.2 TeV W ₁ mass 5.0 TeV	$\Gamma/m = 1.2\%$ $g_V = 3$ $m(N_E) = 0.5 \text{ TeV, } g_1 = g_0$	1908.08248 1709.07242 1805.09299 2006.05138 1906.05938 1801.09932 2004.16936 1906.08569 1712.06518 CERN-EP-2020-073 1807.10473 1904.12678
6	Cl qqqq Cl ffqq Cl terr	– 2 c.µ ≥1 c.µ	2j - 21b, 21j	- Yes	37.0 139 36.1	Λ Λ 2.57 TeV	21.8 TeV v _{i1} 35.8 TeV v _{i2} C _{el} = 4a	1703.09127 CERW-EP-2000-066 1811.02905
MO	Axial-vector mediator (Dirac DM) Colored scalar mediator (Dirac D VV_{KX} EFT (Dirac DM) Scalar reson. $\phi \to \zeta_K$ (Dirac DM)	0 e,μ 0 e,μ	1-4j 1-4j 1J, 51j 1b, 0-1J		36.1 36.1 3.2 36.1	m _{red} 1.55 TeV m _{red} 1.67 TeV ML 700 GeV m _b 3.4 TeV	g_0 =0.25, g_0 =1.0, $m(y) = 1.0$ eV g=1.0, $m(y) = 1.0$ eV m(y) < 190.0eV y = 0.4, $i = 0.2$, $m(y) = 10.0$ eV	1711.03301 1711.03301 1608.02372 1812.09743
9	Scalar LO 1 st gen Scalar LO 2 st gen Scalar LO 3 st gen Scalar LO 3 st gen	1,2 e 1,2 μ 2 τ 0-1 e.μ	≥ 2 j ≥ 2 j 2 b 2 b	Yes Yes - Yes	36.1 36.1 36.1 36.1	LOTTAGE 1.4 TeV LOTTAGE 1.56 TeV LOTTAGE 1.03 TeV LOTTAGE 970 CeV	$\beta = 1$ $\beta = 1$ $S(LQ_i^0 \rightarrow br) = 1$ $S(LQ_i^0 \rightarrow tr) = 0$	1902,00377 1902,00377 1902,00103 1902,00103
Heavy quarks	VLQ $TT \rightarrow Ht/Zt/Wb + X$ VLQ $BS \rightarrow Wt/Zb + X$ VLQ $T_{S/3}T_{S/3} T_{S/3} \rightarrow Wt + X$ VLQ $Y \rightarrow Wb + X$ VLQ $B \rightarrow Hb + X$ VLQ $QQ \rightarrow WqWq$	$1e,\mu$	el .	Yes	36.1 36.1 36.1 36.1 79.8 20.3	T mare 1.37 TeV Bmass 1.34 TeV T symbos 1.64 TeV Y mass 1.85 TeV Bmass 1.21 TeV O nass 890 GeV	SU(2) doublet SU(2) doublet $S(T_{S,2} \to W(2) = 1, c(T_{S,2}W(2) = 1, S(Y \to W(2) = 1, c_N(W(2)) = 1, c_N(W(2)) = 1, c_N(W(2)) = 1$	1808.03949 1808.03949 1807.11859 1812.07849 ATLAS-CONF-2018-024 1509.04251
Excited fermions	Excited quark $q^* \rightarrow qg$ Excited quark $q^* \rightarrow qg$ Excited quark $b^* \rightarrow bg$ Excited lepton c^* Excited lepton v^*	- 1 γ - 3 e.μ 3 e.μ, τ	2j 1j 1b,1j - -	-	139 36.7 36.1 20.3 20.3	g* mass 6.7 TeV g* mass 5.3 TeV b* mass 2.6 TeV r* mass 3.0 TeV r* mass 1.6 TeV	only of and d^* , $h = m(q^*)$ only of and d^* , $h = m(q^*)$ h = 3.0 TeV h = 1.6 TeV	1910.08447 1709.10440 1805.08298 1411.2921 1411.2521
Other		1 e.μ 2 μ 2,3,4 e.μ (88 3 e.μ.τ - - - - 13 TeV ertial data	≥ 2 j 2 j 5) - - - - √s = 10 full d		79.8 36.1 36.1 20.3 36.1 34.4	N° mass 580 GeV Ny mass 3.2 TeV H ²² mass 670 GeV H ⁴⁴ mass 1,22 TeV monopole mass 1,22 TeV 10 ⁻¹ 1	$m(W_N) = 4.1 \text{ TeV, } g_k = g_N$ DY production DY production, $g(H_k^{an} \rightarrow 0\tau) = 1$ DY production, $ g = 5e$ DY production, $ g = 1g_0$, spin 1/2 10 Mass scale [TeV]	ATLAS-CONF-2018-020 1899,11105 1710,09748 1411,2591 1612,03673 1505,10130

^{*}Only a selection of the available mass limits on new states or phenomena is shown.

[†]Small-radius (large-radius) jets are denoted by the letter j (J).

ATLAS Exotics Searches* - 95% CL Upper Exclusion Limits

						Opper Exerus					- Fremininary
St	atus: May 2020							∫£ dt	t = (3.2 - 1)	139) fb ⁻¹	\sqrt{s} = 8, 13 TeV
	Model	t, y	Jets†	E _{miss}	∫£ dt[ft		Limit				Reference
Extra dimensions	ADD $G_{WN}+g/q$ ADD non-resonant $\gamma\gamma$ ADD CBH ADD BH high $\sum \rho \gamma$ ADD BH multipl RS1 $G_{WN} \rightarrow \gamma\gamma$ Bulk RS $G_{WN} \rightarrow WW/ZZ$ Bulk RS $G_{WN} \rightarrow WV \rightarrow \ell \nu n n n$ Bulk RS $G_{WN} \rightarrow t n n$ Bulk RS $G_{WN} \rightarrow t n$ SUED / RPP		1=4j - 2j 2 2 j 2 3 j - ol 2 j / 1 J 2 1 b, 2 1 J 2 2 b, 2 3	/2] Yes	36.1 36.7 37.0 8.2 3.6 36.7 36.1 139 36.1	Mo Mo Mo Mo Mo Gen noss Gen noss Gen noss Ext noss Ext noss Ext noss Ext noss		7.7 Te 8.6 T 8.9 Te	feV $n = 3 \text{ M}$ n = 6 N $n = 6$, $Nn = 6$, N	= 1.0 = 1.0	1711.03901 1707.04147 1703.09127 1606.02365 1512.02566 1707.04147 1606.02080 2004.14636 1804.10823 1600.04676
Gauge bosons	SSM $Z' \rightarrow \ell\ell$ SSM $Z' \rightarrow \tau\tau$ Leptophobic $Z' \rightarrow bb$ Leptophobic $Z' \rightarrow t\ell$ SSM $W' \rightarrow \tau\tau$ HYT $W' \rightarrow WZ \rightarrow \ell rqq$ model HYT $V' \rightarrow WV \rightarrow qqqq$ model HYT $V' \rightarrow WH/ZH$ model B HYT $W' \rightarrow WH$ model B LPSM $W_H \rightarrow tb$ LPSM $W_H \rightarrow tN_R$	B 0 e,μ multi-channe	≥ 1 b, ≥ 2	Yes Yes Yes	189 36.1 36.1 189 138 36.1 138 36.1 138 36.1 189 36.1	Z' mass Z' mass Z' mass Z' mass W' mass W' mass W' mass V' mass		5.1 TeV 2.42 TeV 2.1 TeV 4.1 TeV 6.0 TeV 3.7 TeV 4.3 TeV 3.8 TeV 2.93 TeV 3.2 TeV 3.2 TeV 5.0 TeV	$F/w = 3$ $g_V = 3$ $g_V = 3$ $g_V = 3$ $g_V = 3$		1903.09248 1709.07242 1805.09289 2006.05138 1806.05088 1801.09992 2004.19636 1906.08589 1712.08518 CERN-EP-2020-073 1807.10473 1804.12678
70	Cl qqqq Cl (fqq Cl tert		2j ≥1 b, ≥1j	-	37.0 139 36.1	Λ Λ Λ		2.57 TeV		35.8 TeV (η_L	1708.09127 CEPN-EP-2000-005 1811.02905
MO	Axial-vector mediator (Dirac DM) Colored scalar mediator (Dirac D VV_{XX} EFT (Dirac DM) Scalar reson, $\phi \to \xi_X$ (Dirac DM)	DM) Θ e,μ Θ e,μ	1-4j 1-4j 1J,≤1j 1b,0-1J		36.1 36.1 3.2 36.1	M _{red} Mare M. m _o		.55 TeV 1.67 TeV 3.4 TeV	$g=1.0, x$ $m(\chi) <$	$g_{\rm g}$ =1.0, $m(p) = 1$ GeV m(p) = 1 GeV 150 GeV $g_{\rm g} = 10$ GeV	1711.03301 1711.03301 1608.02372 1812.09743
93	Sosiar LO 1 st gen Sosiar LO 2 st gen Sosiar LO 3 st gen Sosiar LO 3 st gen	1,2 e 1,2 µ 2 r 0-1 e.µ	2 2 j 2 2 j 2 b 2 b	Yes Yes - Yes	36.1 36.1 36.1 36.1	LOTTERS LOTTERS LOTTERS LOTTERS LOTTERS				$-b_T = 1$ $-b_T = 0$	1902.00377 1902.00377 1902.00103 1902.00103
Heavy	$\begin{array}{l} VLQ \ TT \rightarrow Ht/Zt/Wb + X \\ VLO \ BS \rightarrow Wt/Zb + X \\ VLQ \ T_{S/3}T_{S/3} T_{S/3} \rightarrow Wt + X \\ VLQ \ Y \rightarrow Wb + X \\ VLQ \ B \rightarrow Hb + X \\ VLQ \ QQ \rightarrow WqWq \end{array}$	1 e.u	el	. Yes	36.1 36.1 36.1 36.1 79.8 20.3	Timuse Bimass Tissimass Yimass Bimass Oinass	1.34	7 TeV 6 TeV 1.84 TeV 1.85 TeV TeV		subset $\rightarrow W(t) = 1, c(T_{5/2}W(t) = 1$ $W(t) = 1, c_N(W(t)) = 1$	1808.02343 1808.02343 1807.11853 1812.07343 ATLAS-CONF-0018-084 1509.04261
Excited fermions	Excited quark $\sigma^* \to \sigma g$ Excited quark $\sigma^* \to \sigma g$ Excited quark $b^* \to b g$ Excited lepton c^* Excited lepton ν^*	- 1 γ - 3 e.μ 3 e,μ,τ	2j 1j 1b,1j - -	=	139 36.7 36.1 20.3 20.3	q" mass q" mass b" mass i" mass in" mass		5.7 TeV 5.3 TeV 2.6 TeV 3.0 TeV			1910.08447 1709.10440 1805.08298 1411.2921 1411.2521
Other		1 e.µ 2 p 2,3,4 e.p (88 3 e.p. τ - - x = 13 TeV artial data	≥ 2 j 2 j 3) - - - - - √s = 10 full d		79.8 36.1 36.1 20.3 36.1 34.4	Nº mass Nº mass Hºº mass Hºº mass mati-derged particle mass monopole mass	560 GeV 670 GeV 1.22 T	2.37 TeV	DY prod DY prod DY prod DY prod	= 4.1 TeV, $g_k = g_V$ better better, $\mathcal{Z}(H_k^{os} \to \ell r) = 1$ better, $ g = 5e$ better, $ g = 1g_G$, spin 1/2 ass scale [TeV]	ATLAS-CONF-2018-020 1809-11105 1710-02748 1411-2921 1612-03673 1905-10130

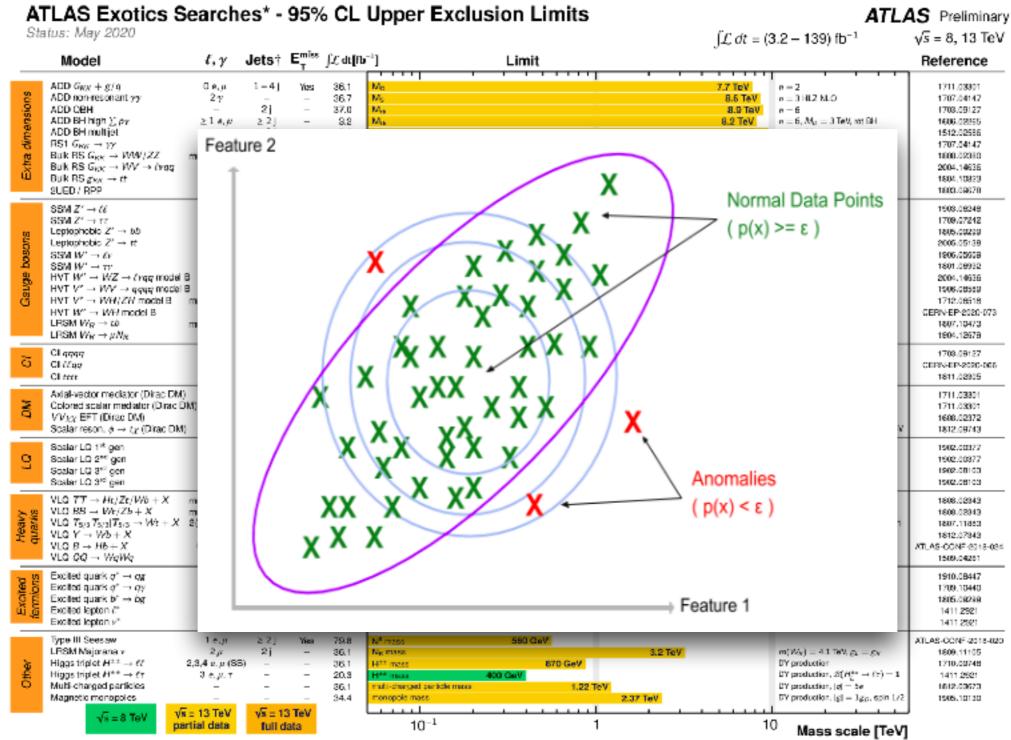
[&]quot;Only a selection of the available mass limits on new states or phenomena is shown.

Higher rates, but hard to distinguish from SM

Look more different, but more low rates

ATLAS Preliminary

[†]Small-radius (large-radius) jets are denoted by the letter j (J).

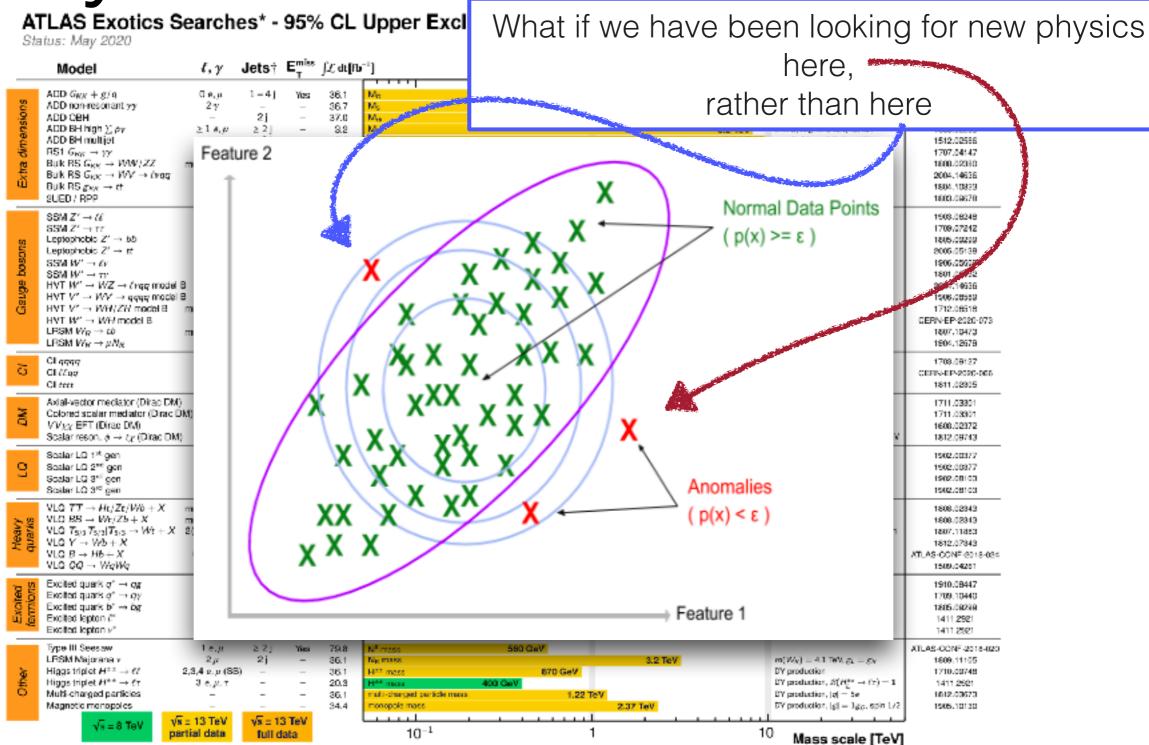


^{*}Only a selection of the available mass limits on new states or phenomena is shown.

+Small-radius (large-radius) jets are denoted by the letter j (J).

Higher rates, but hard to distinguish from SM

Look more different, but more low rates



^{*}Only a selection of the available mass limits on new states or phenomena is shown.

†Small-radius (large-radius) jets are denoted by the letter j (J). Higher rates, but hard to distinguish from SM

Look more different, but more low rates

- 1) Model Agnostic: We are good at looking for models we know of, but what if we don't know what we should be looking for?
- Simulation Independent: With no signal model, it is possible to use methods directly on data from the LHC without Monte Carlo simulations.

How to detect anomalies?

Over 50 papers in HEP anomaly detection https://iml-wg.github.io/HEPML-LivingReview/

LHC Olympics [2101.08320] focuses on finding over densities in all-hadronic events

- Black Box 1: Similar to example data: 4 methods found resonance
- Black Box 2: SM only. 4 methods claimed a resonance, 1 claimed lack-of-resonance
- Black Box 3: Correct resonance not detected by any group

Dark Machines Challenge [2105.14027] focuses on finding individual events which look different

- Train on SM-only events, apply to many different signals
- Find methods which work best for most signals
- No method finds every new physics signal

How to detect anomalies?

PHYSICAL REVIEW D 101, 075021 (2020)

Searching for new physics with deep autoencoders

Marco Farina, ^{1,2} Yuichiro Nakai, ² and David Shih²

¹C.N.Yang Institute for Theoretical Physics, Stony Brook, New York 11794, USA

²NHETC, Dept. of Physics and Astronomy Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA

(Received 10 January 2019; accepted 27 March 2020; published 13 April 2020)

We introduce a potentially powerful new method of searching for new physics at the LHC, using autoencoders and unsupervised deep learning. The key idea of the autoencoder is that it learns to map "normal" events back to themselves, but fails to reconstruct "anomalous" events that it has never encountered before. The reconstruction error can then be used as an anomaly threshold. We demonstrate the effectiveness of this idea using QCD jets as background and boosted top jets and R-parity violating (RPV) gluino jets as signal. We show that a deep autoencoder can significantly improve signal over background when trained on backgrounds only, or even directly on data which contain a small admixture of signal. Finally, we examine the correlation of the autoencoders with jet mass and show how the jet mass distribution can be stable against cuts in reconstruction loss. This may be important for estimating QCD backgrounds from data. As a test case, we show how one could plausibly discover 400 GeV RPV gluinos using an autoencoder combined with a bump hunt in jet mass. This opens up the exciting possibility of training directly on actual data to discover new physics with no prior expectations or theory prejudice.

DOI: 10.1103/PhysRevD.101.075021

Sci Post

SciPost Phys. 6, 030 (2019)

QCD or what?

Theo Heimel¹, Gregor Kasieczka², Tilman Plehn^{1*} and Jennifer M. Thompson¹

Institut für Theoretische Physik, Universität Heidelberg, Germany
 Institut für Experimentalphysik, Universität Hamburg, Germany

⋆ plehn@uni-heidelberg.de

Abstract

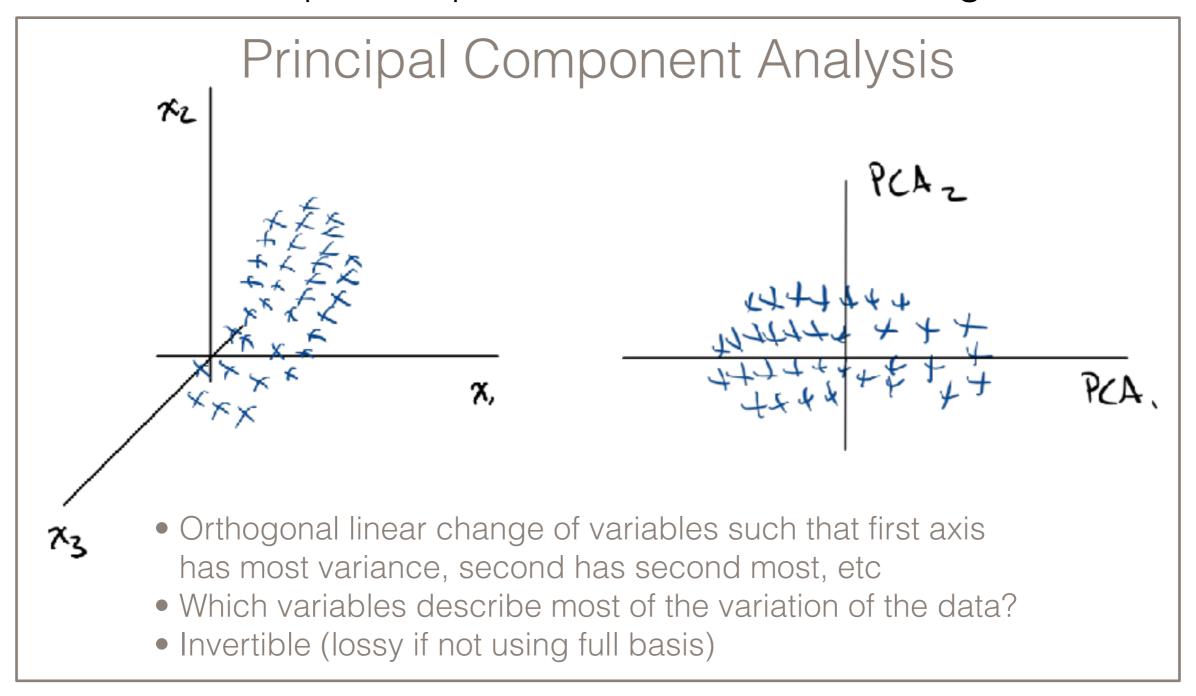
Autoencoder networks, trained only on QCD jets, can be used to search for anomalies in jet-substructure. We show how, based either on images or on 4-vectors, they identify jets from decays of arbitrary heavy resonances. To control the backgrounds and the underlying systematics we can de-correlate the jet mass using an adversarial network. Such an adversarial autoencoder allows for a general and at the same time easily controllable search for new physics. Ideally, it can be trained and applied to data in the same phase space region, allowing us to efficiently search for new physics using un-supervised learning.

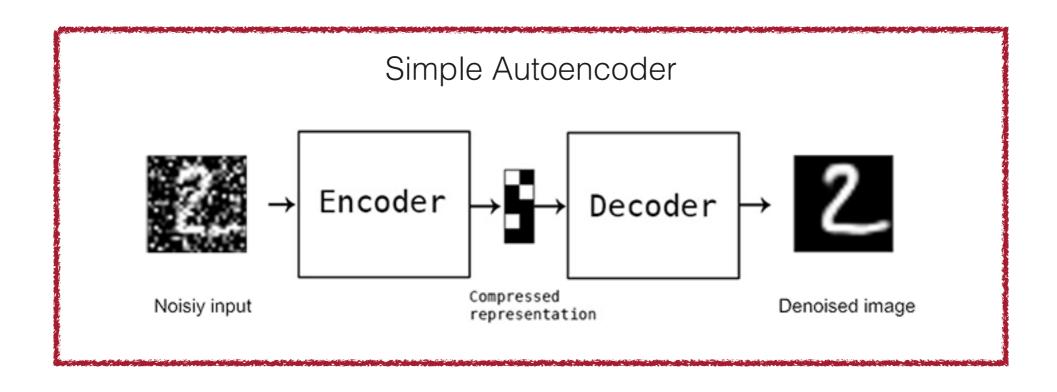
[<u>1808.08992</u>]

[<u>1808.08979</u>]

- 1. Use *Autoecoders* for anomaly detection
- 2. Emphasize training directly on data, how much signal can be in training and still work?

- Method for dimensionality reduction
- Focuses on important pieces of information and ignores noise





- Encoding and Decoding can be non-linear
- Encoder learns what is important in the data and what is not
- Size of compressed representation chosen before training
- Compressed representation changed each training of the networks

- Unlike PCA, autoencoders need to be trained
- Need input data, output predictions, and a target
 - Target data is the same as input data

- Unlike PCA, autoencoders need to be trained
- Need input data, output predictions, and a target
 - Target data is the same as input data
- How to compare output predictions and target?
 - Use Mean Squared Error

$$d_{\text{ME},2} = \text{MSE} \equiv \frac{1}{n} \sum_{i=1}^{n} \left| D(E(x))_i - y_i \right|^2$$

- Unlike PCA, autoencoders need to be trained
- Need input data, output predictions, and a target
 - Target data is the same as input data
- How to compare output predictions and target?
 - Use Mean Squared Error

$$d_{\mathrm{ME},2} = \mathrm{MSE} \equiv \frac{1}{n} \sum_{i=1}^{n} \left| D\big(\mathrm{E}(\mathrm{x}) \big)_i - \mathrm{y_i} \right|^2$$

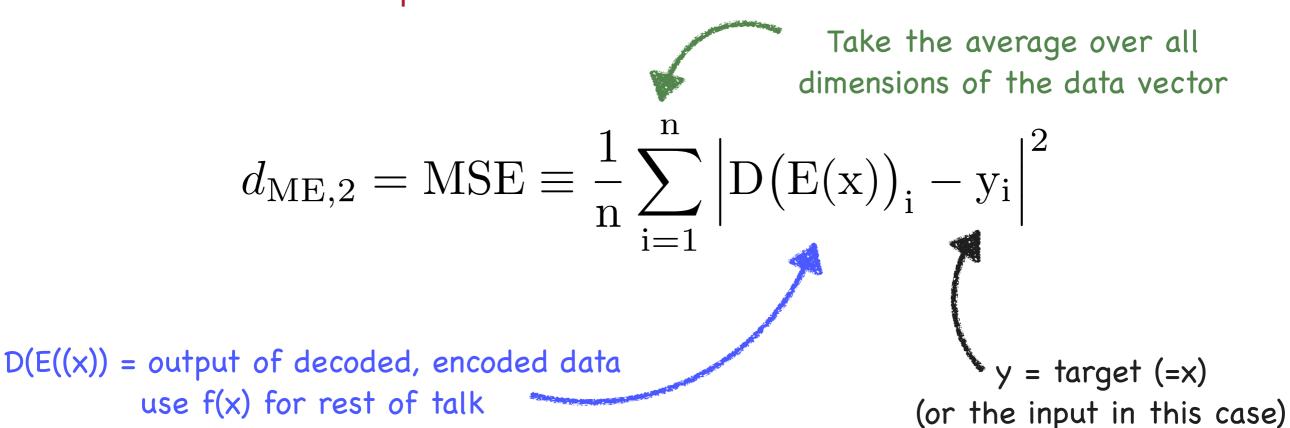
$$y = \text{target (=x)}$$
(or the input in this case)

- Unlike PCA, autoencoders need to be trained
- Need input data, output predictions, and a target
 - Target data is the same as input data
- How to compare output predictions and target?
 - Use Mean Squared Error

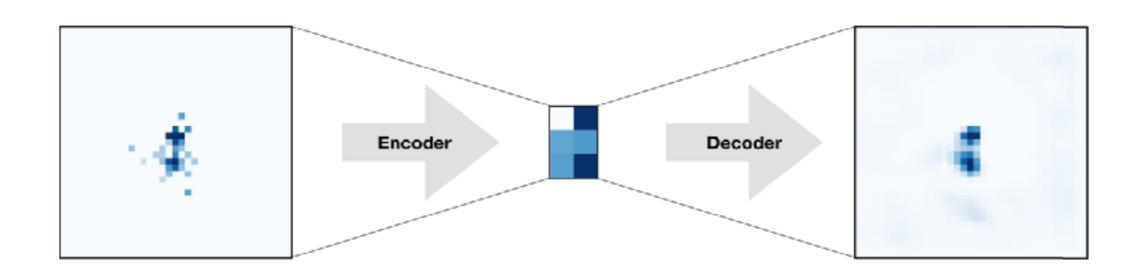
$$d_{\rm ME,2}={\rm MSE}\equiv\frac{1}{n}\sum_{\rm i=1}^{\rm n}\left|{\rm D}\big({\rm E}({\rm x})\big)_{\rm i}-{\rm y_i}\right|^2$$
 D(E((x)) = output of decoded, encoded data use f(x) for rest of talk (or the input in th

(or the input in this case)

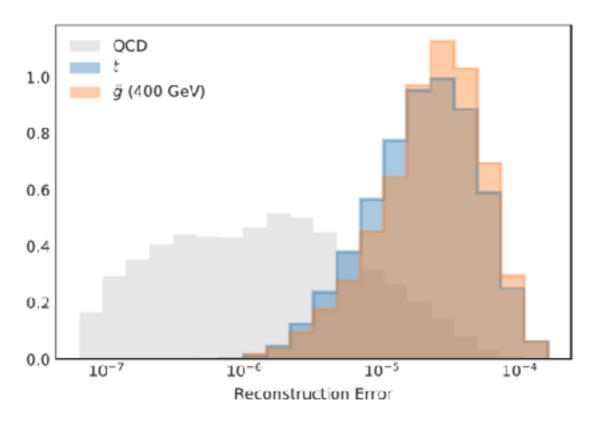
- Unlike PCA, autoencoders need to be trained
- Need input data, output predictions, and a target
 - Target data is the same as input data
- How to compare output predictions and target?
 - Use Mean Squared Error



Autoencoders for Anomaly Detection

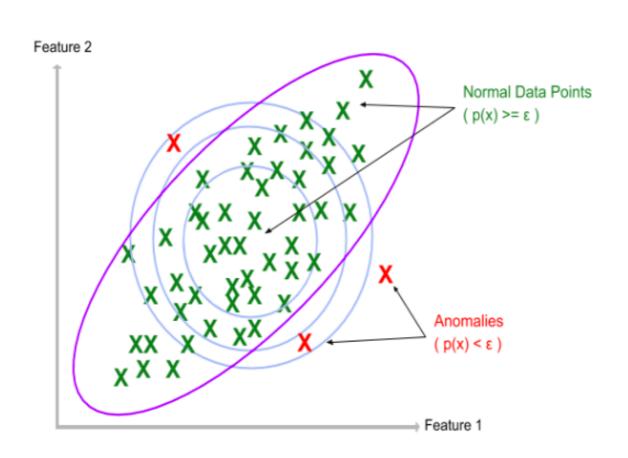


- Networks are trained to minimize the reconstruction error of events from SM background
- The encoding-decoding of BSM events will have larger reconstruction errors



[<u>1808.08992</u>]

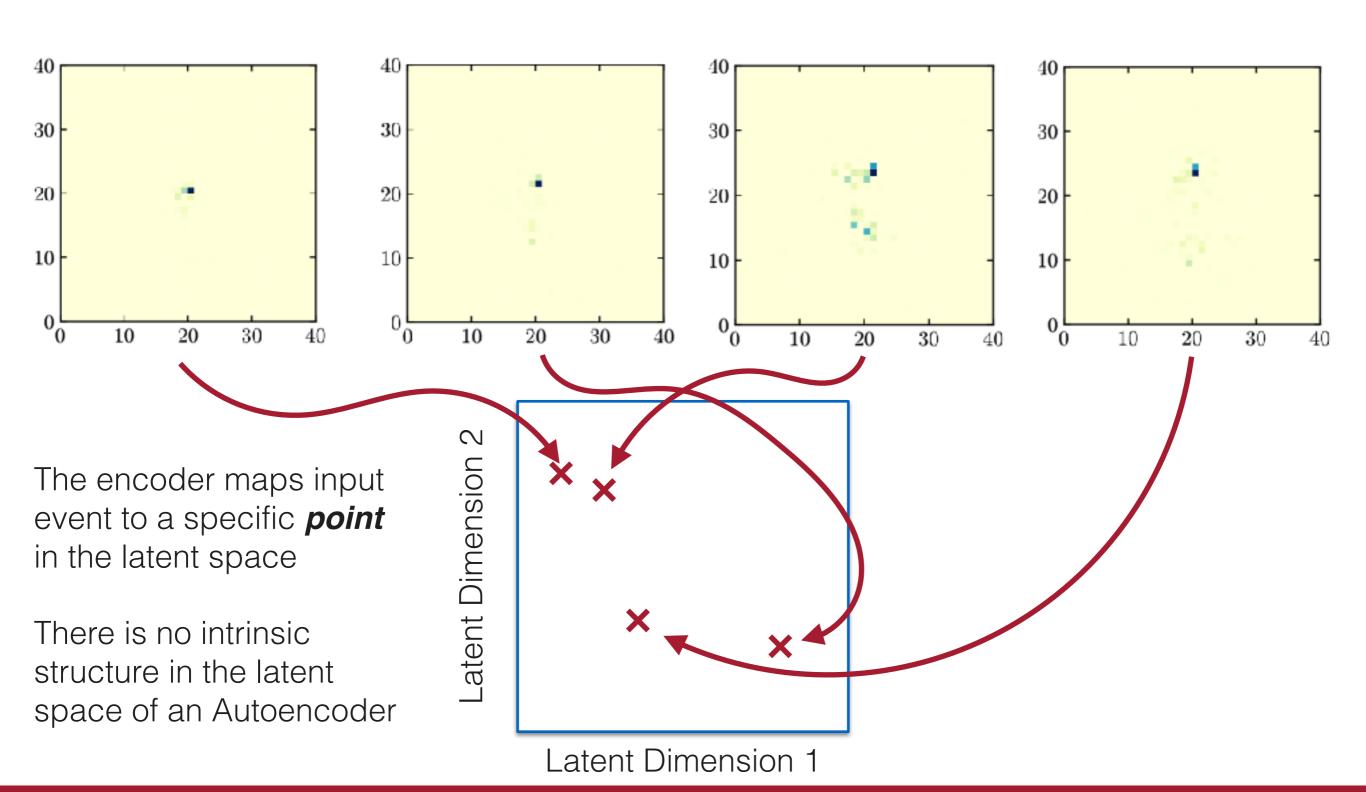
Challenges for Autoencoders

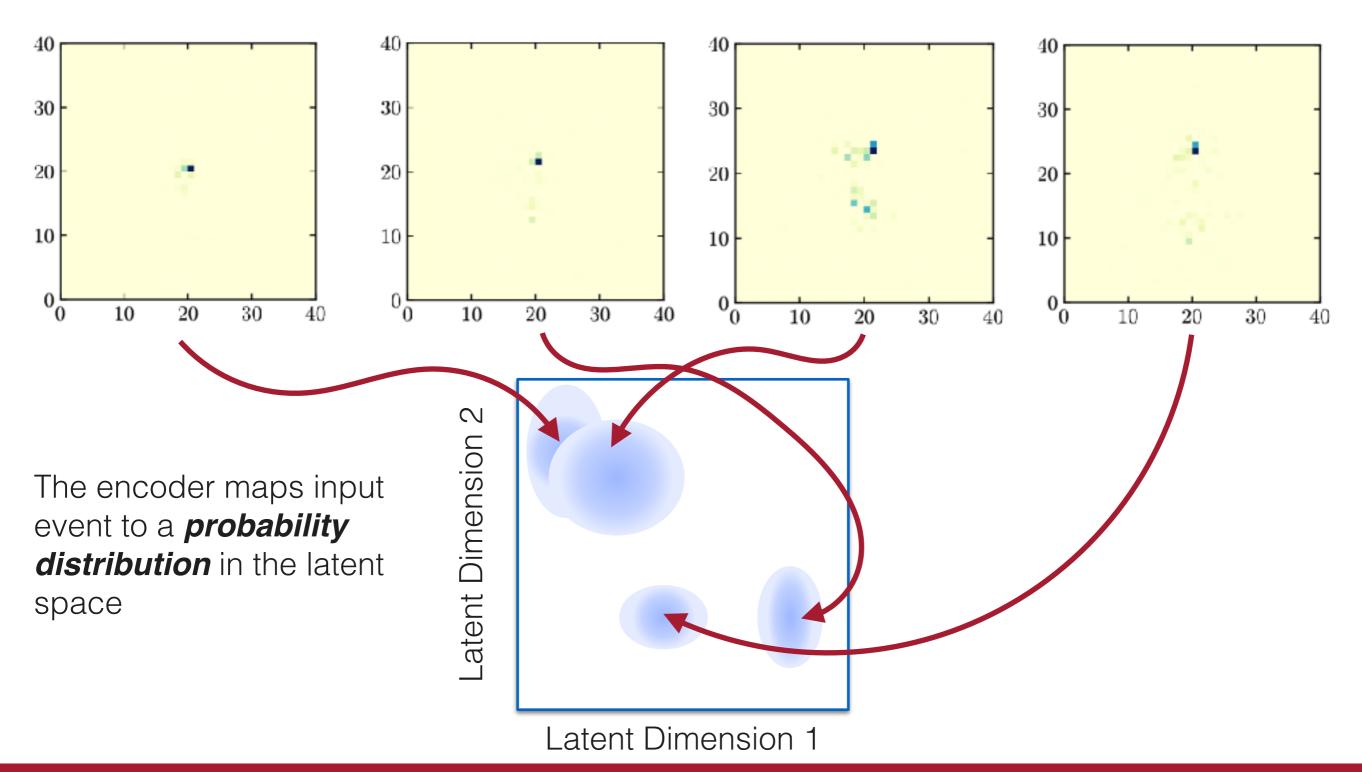


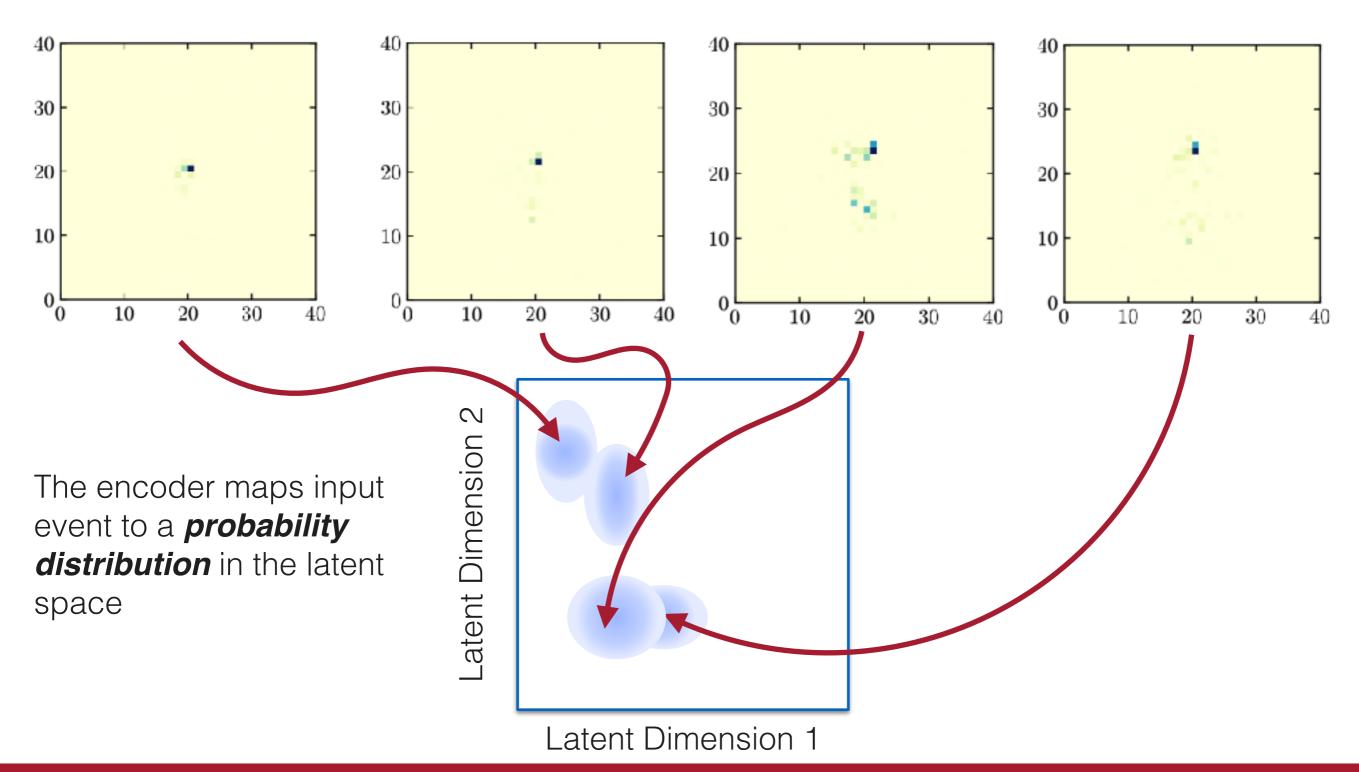
Do autoencoders fit with this picture of anomaly detection?

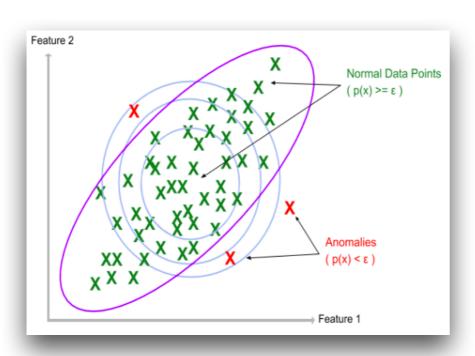
- Delicate balance between reconstructing SM well, but anomalous data poorly
- Latent space doesn't have structure, can't assign probabilistic interpretation
- Is MSE the right metric to use for reconstruction?

Plain Autoencoders









Add a probabilistic interpretation

How likely is a detector event, given a point in latent space?

Model each detector pixel as a Gaussian:

$$p(E|z) \propto \exp\left(\frac{-\left(E - D(z)\right)^2}{2\sigma^2}\right)$$

$$x = \{E_1, E_2, \cdots, E_N\}$$
$$p(x|z) = \prod_{i=1}^{N} p(E_i|z)$$

Probability of the observing an event

$$p(x) = \int p(x|z)p(z)dz \equiv \mathbb{E}_{p(z)}[p(x|z)]$$

What is the latent space to integrate over?

Probability of the observing an event

$$p(x) = \int p(x|z)p(z)dz \equiv \mathbb{E}_{p(z)}[p(x|z)]$$

What is the latent space to integrate over?

Use the encoder to learn/sample the latent space

$$p(x) = \int p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz = \mathbb{E}_{q(z|x)}\left[p(x|z)\frac{p(z)}{q(z|x)}\right]$$

Probability of the observing an event

$$p(x) = \int p(x|z)p(z)dz \equiv \mathbb{E}_{p(z)}[p(x|z)]$$

What is the latent space to integrate over?

Use the encoder to learn/sample the latent space

$$p(x) = \int p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz = \mathbb{E}_{q(z|x)}\left[p(x|z)\frac{p(z)}{q(z|x)}\right]$$

Take log likelihood and use Jensen's inequality to move log inside expectation value

$$\log p(x) \ge \mathbb{E}_{q(z|x)} \left[\log \left(p(x|z) \right) + \log \left(\frac{p(z)}{q(z|x)} \right) \right]$$

Probability of the observing an event

$$p(x) = \int p(x|z)p(z)dz \equiv \mathbb{E}_{p(z)}[p(x|z)]$$

What is the latent space to integrate over?

Use the encoder to learn/sample the latent space

$$p(x) = \int p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz = \mathbb{E}_{q(z|x)}\left[p(x|z)\frac{p(z)}{q(z|x)}\right]$$

Take log likelihood and use Jensen's inequality to move log inside expectation value

$$\log p(x) \ge \mathbb{E}_{q(z|x)} \left[\log \left(p(x|z) \right) + \log \left(\frac{p(z)}{q(z|x)} \right) \right]$$

 MSE of encodeddecoded event

Probability of the observing an event

$$p(x) = \int p(x|z)p(z)dz \equiv \mathbb{E}_{p(z)}[p(x|z)]$$

What is the latent space to integrate over?

Use the encoder to learn/sample the latent space

$$p(x) = \int p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz = \mathbb{E}_{q(z|x)}\left[p(x|z)\frac{p(z)}{q(z|x)}\right]$$

Take log likelihood and use Jensen's inequality to move log inside expectation value

$$\log p(x) \ge \mathbb{E}_{q(z|x)} \left[\log \left(p(x|z) \right) + \log \left(\frac{p(z)}{q(z|x)} \right) \right]$$

 MSE of encodeddecoded event Kullback-Leibler Divergence (KLD) between encoded representation and latent prior

Summary of Autoencoders

Plain autoencoders: no sampling, $L = \mathrm{MSE}$

Variation autoencoders: sampling in latent space, L = MSE + KLD

Sampling adds/forces structure on the latent space.

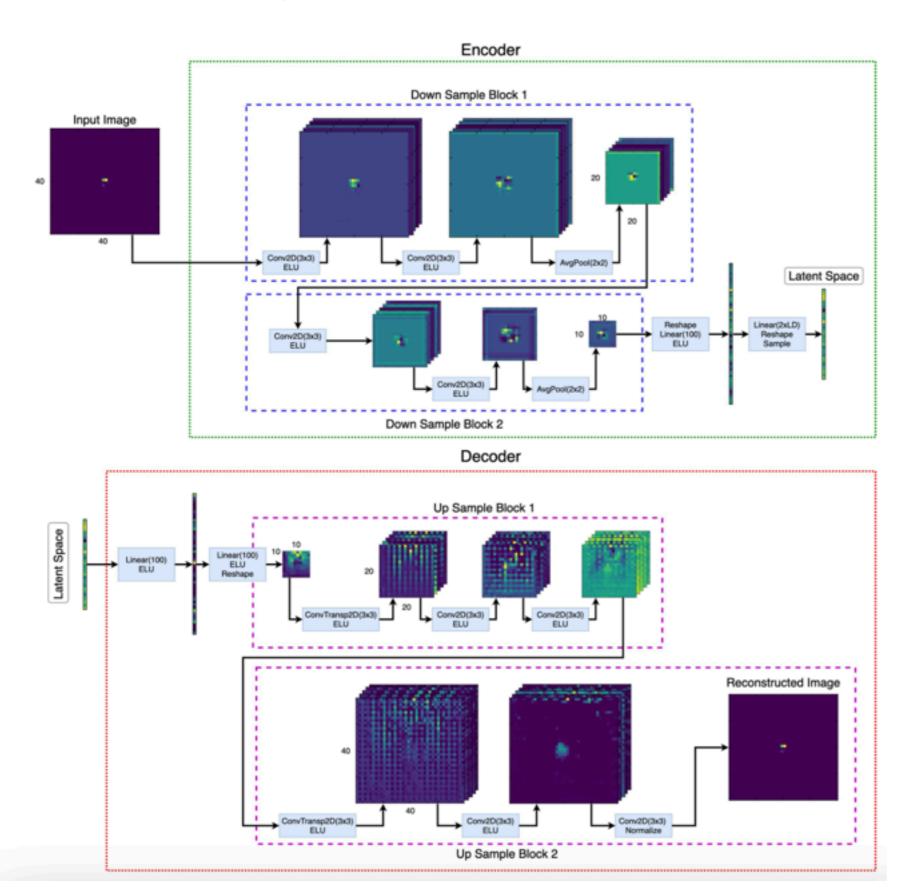
KLD term helps to regularize and adds to

probabilistic interpretation

In practice, these terms may be too far apart, introduce a scaling between them

$$L = (1 - \beta) MSE + \beta KLD$$

Generative Networks



Can randomly sample in the latent space to get "new" data

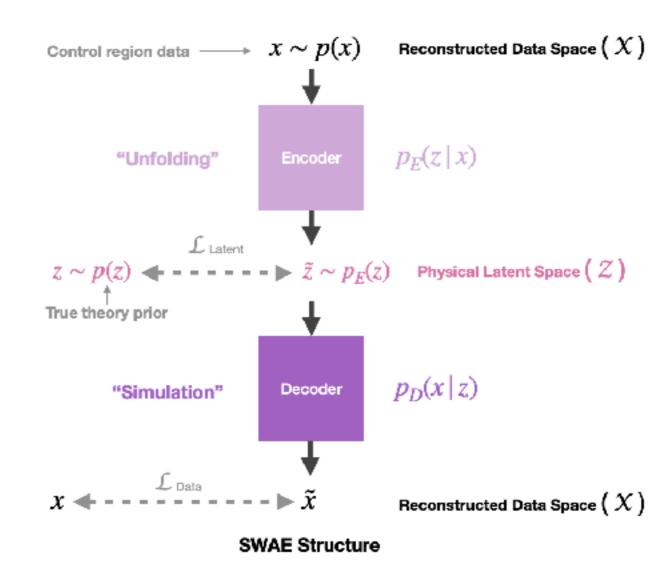
Alter the latent space

- Latent loss (£ Latent)
 - · SW distance for finite samples
- Data loss (£ Data):
 - Mean Squared Error (MSE)
- Total SWAE loss function:

$$\mathcal{L}_{\text{SWAE}} = \mathcal{L}_{\text{Data}} + \lambda \mathcal{L}_{\text{Latent}}$$

Easy to add additional physically-motivated constraints

$$\mathcal{L} = \mathcal{L}_{\text{SWAE}} + \lambda_i' \mathcal{L}_i$$

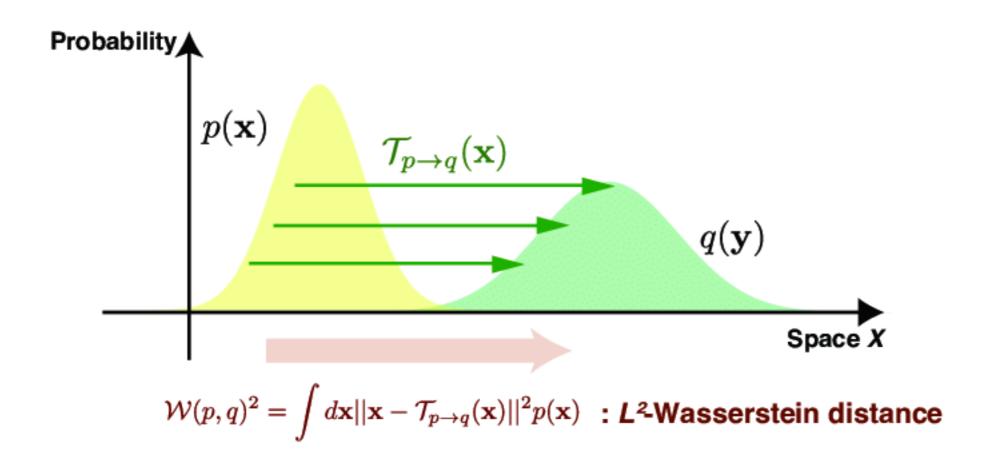


In the previous discussion of VAEs, we used a Gaussian for the prior of the latent space. However, it is also possible to use different/more motivated priors. For instance, use the parton-level distributions.

[2101.08944]

Alter the latent space

Use the Earth mover's distance (Wasserstein distance) to compare the latent prior with the learned latent distributions



Side note: Wasserstein distances are also an active area of research for the other areas of machine learning in HEP, such as classification and anomaly detection.

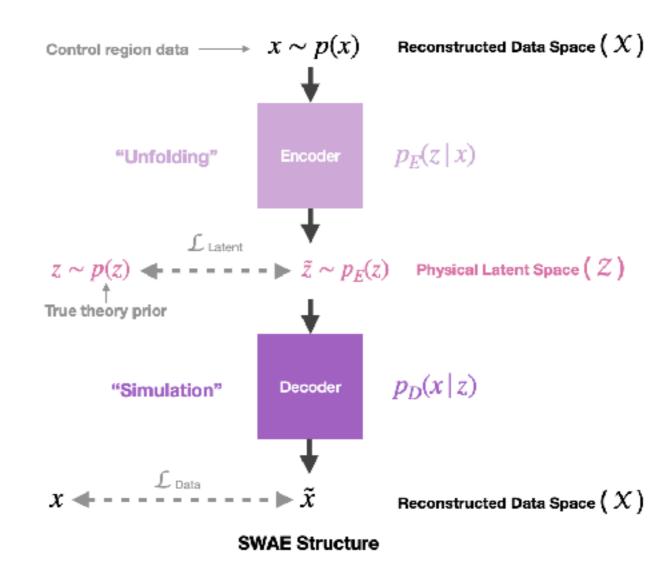
Alter the latent space

- Latent loss (£ Latent)
 - SW distance for finite samples
- Data loss (£ Data):
 - Mean Squared Error (MSE)
- Total SWAE loss function:

$$\mathcal{L}_{\text{SWAE}} = \mathcal{L}_{\text{Data}} + \lambda \mathcal{L}_{\text{Latent}}$$

Easy to add additional physically-motivated constraints

$$\mathcal{L} = \mathcal{L}_{\text{SWAE}} + \lambda_i' \mathcal{L}_i$$



Example in paper: Examine p p > t t~ events with semi-leptonic decays,

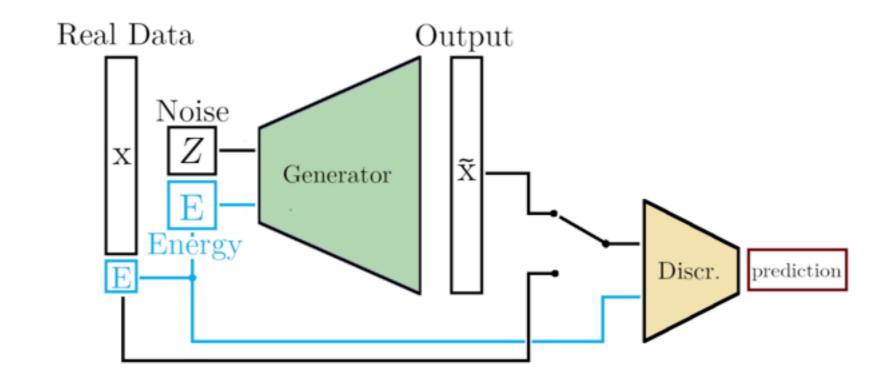
X is the detector level information (lepton 4-vector, MET, and jet four vectors)

Z is the parton level information (lepton, neutrino, and quark four vectors)

Many challenges, active research

[<u>2101.08944</u>]

Other generative models



Generative Adversarial Networks (GAN) work by taking in random noise and generating an event (image, observables, etc). There are two components to this, first is the network which generates the event, second is the discriminator network. This network is trained to tell the difference between real and generated data. The generator is trained to trick the discriminator.

Over 55 papers in the HEP literature using GANs

Other generative models

Ideas seen for generating events:

GANS -> Noise to data

VAES -> Latent representation to data

In both of these, we are trying to capture something about the underlying distributions of the true data, why not try to learn that directly?

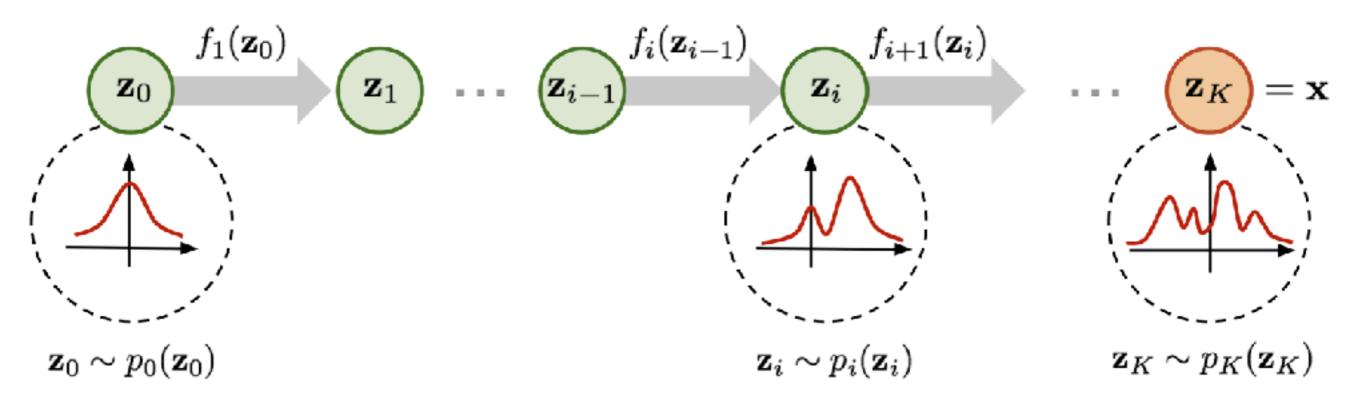
Normalizing Flows

Normalizing Flows

- Normalizing flows implement a change of variables from a simple distribution to the real distribution.
- Often use special layers which are easily invertible and have a simple Jacobian
- Once trained, can draw a random sample from the easy distribution and obtain a sample from a complex distribution
- Invertibility then also allows for taking a sample from the complex distribution and computing the likelihood

Nice review of the methods [1912.02762.pdf] (not HEP)

Normalizing Flows



These models allow for easy density estimation (learn the PDF of the data)

Allows for: Generating events, Generating field configurations for Lattice Gauge Theory, or Anomaly Detection

Conclusions

- There are many types of machine learning which do not simply use labeled data for supervision
- We saw how AEs for anomaly detection do not build in a probabilistic interpretations, but a VAE does
- VAES can also be used to generate data (and the latent space can be made more intuitive)
- There are many other ways to make generative networks
- We can use ML to estimate probabilities which allows for inference as well

Ideas for Project (Tutorial 3)

- 1) Compare raw-4vectors, n-subjettiness basis, and jetimages for top-tagging dataset
 - a) Which does best?
 - b) Are the number of parameters the similar?
 - c) Speed of training?
- 2) Use **anomaly detection** techniques to train only on the background data. Then apply to the test set and see how the performance compares to supervised classification.

References

https://iml-wg.github.io/HEPML-LivingReview/

Particle Data Group has a new review chapter on ML which covers all of these techniques and more https://pdg.lbl.gov/