

Report for Preliminary Study

Patomporn (Jab)
CMS DQM-ML4DC
8 July 2019

Outline

- Datasets
- Model
- Results
- Let find the cutoff (Application)
- Extended Investigation

Datasets

- JetHT
- 2016 Datasets
 - 259 Features
- Lumisection certification granularity
 - Good LS defined in Golden JSON
 - Rest of LS are bad

Datasets

- **Preprocessing**
 - **MinMaxScalar** Transformation
 - Consider Lumisection i and Feature j

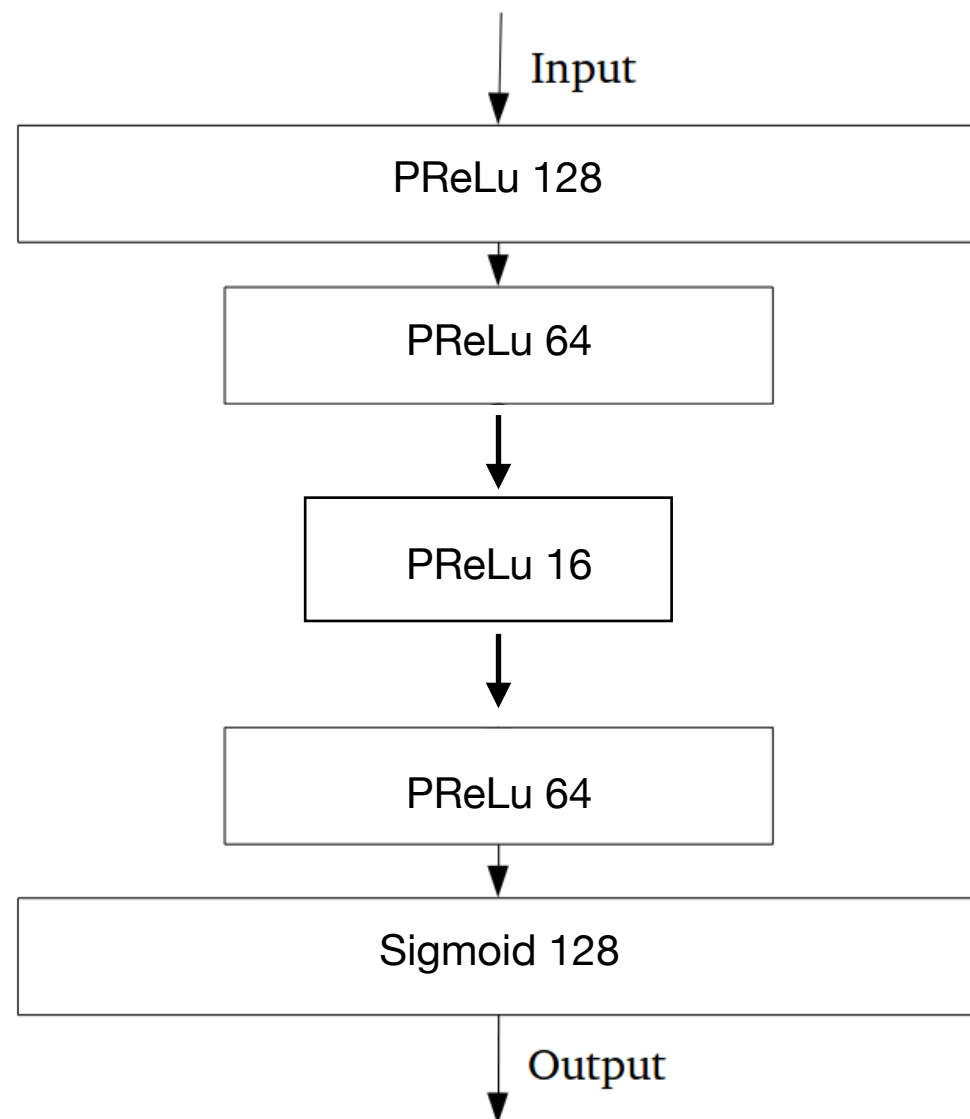
$$x'_{ij} \leftarrow \frac{x_{ij} - \min_{\forall i \in S_{\text{train}}} \{x_{ij}\}}{\max_{\forall i \in S_{\text{train}}} \{x_{ij}\} - \min_{\forall i \in S_{\text{train}}} \{x_{ij}\}}$$

- Then our datapoint should be in range $[0, 1]$

Model

- 4 Flavours of Autoencoder
 - **Vanilla**
 - Sparse
 - Contractive
 - Variational
- No Neural-net
 - Isolation Forest
 - **Schölkopf's One-Class SVM**

My Vanilla AE



- Sigmoid fn. in output should bound between zero and one
- PReLU for the rest of activ. fn.

Training Dependency

- Truncated normal variable initializer
- Batch size 256
- EPOCHS 1200

Schölkopf's One-Class SVM

- Minimize (Soft margin)

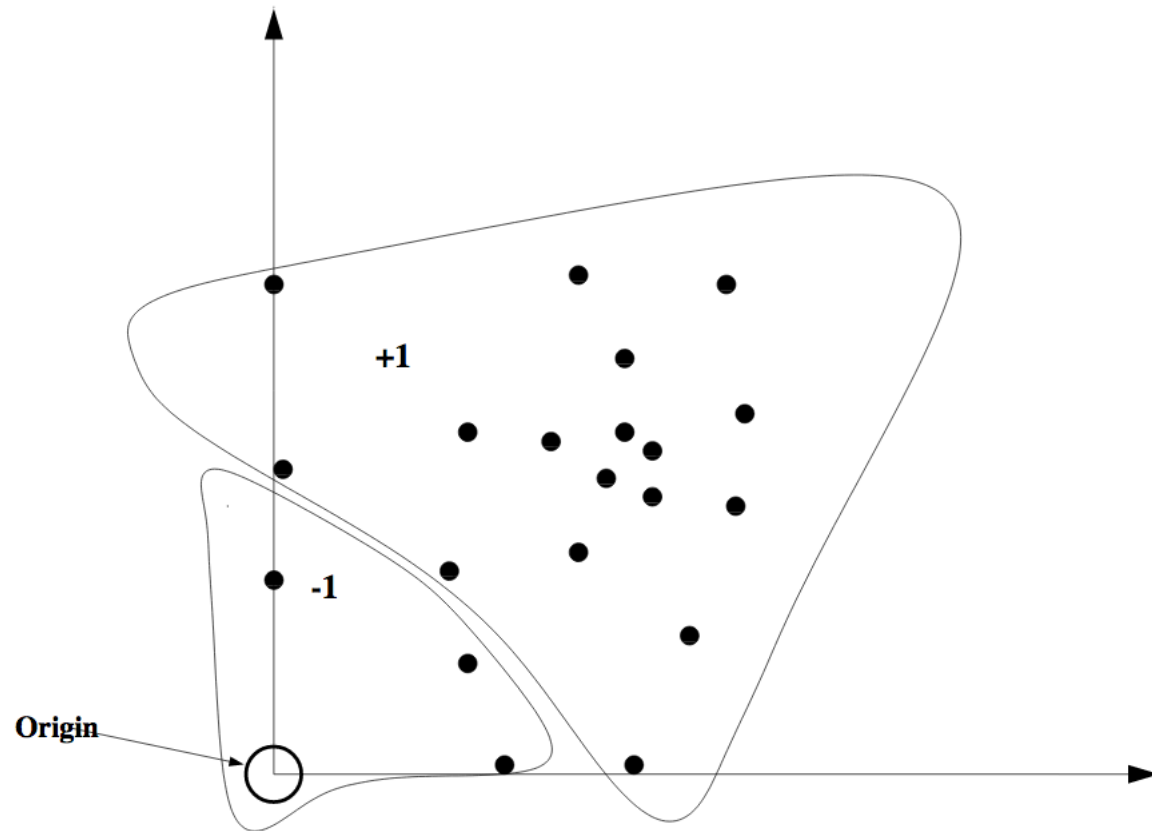
$$\frac{||w||^2}{2} + \frac{1}{\nu l} \sum_{i=1}^l \xi_i - \rho$$

- Under

$$w \cdot \Phi(x_i) \geq \rho - \xi_i; \xi_i \geq 0$$

- **Kernel:** Gaussian Base Radial function (GBF)

- Determine by tangent distant from data point to hyperplane

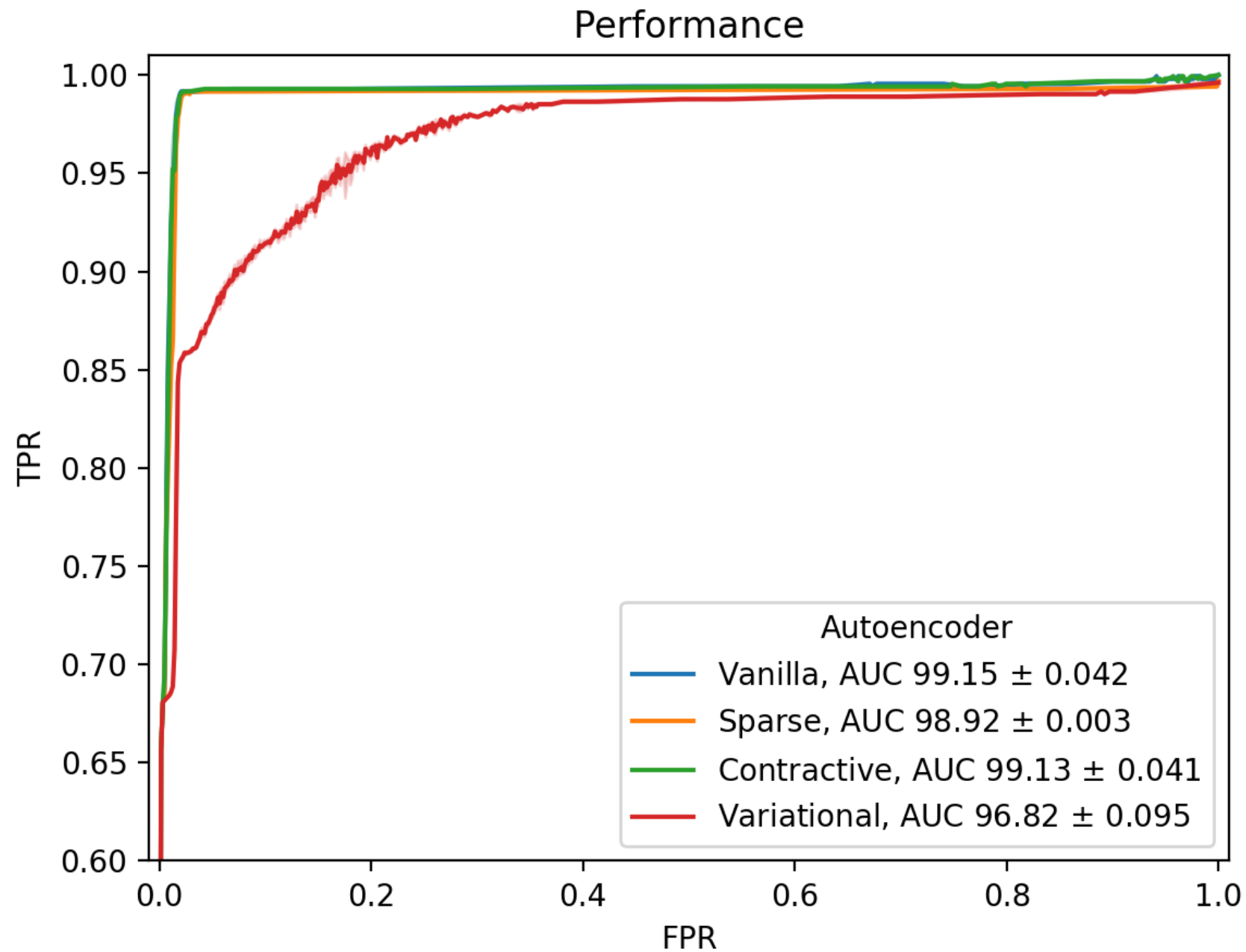


[1] <http://www.jmlr.org/papers/volume2/manevitz01a/manevitz01a.pdf>

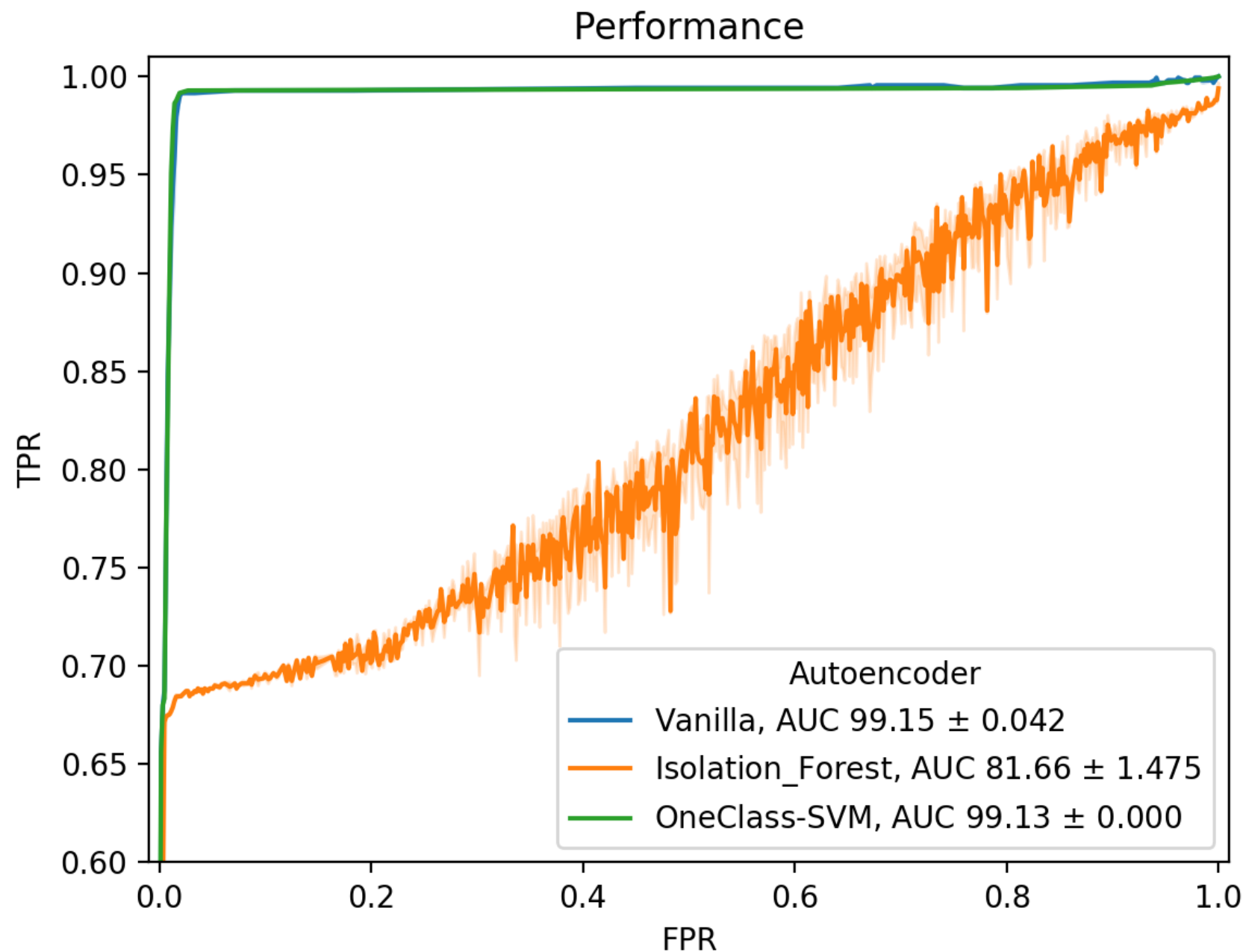
[2] <http://papers.nips.cc/paper/1723-support-vector-method-for-novelty-detection.pdf>

Results

AE Performance



Vanilla vs no-NN Performance



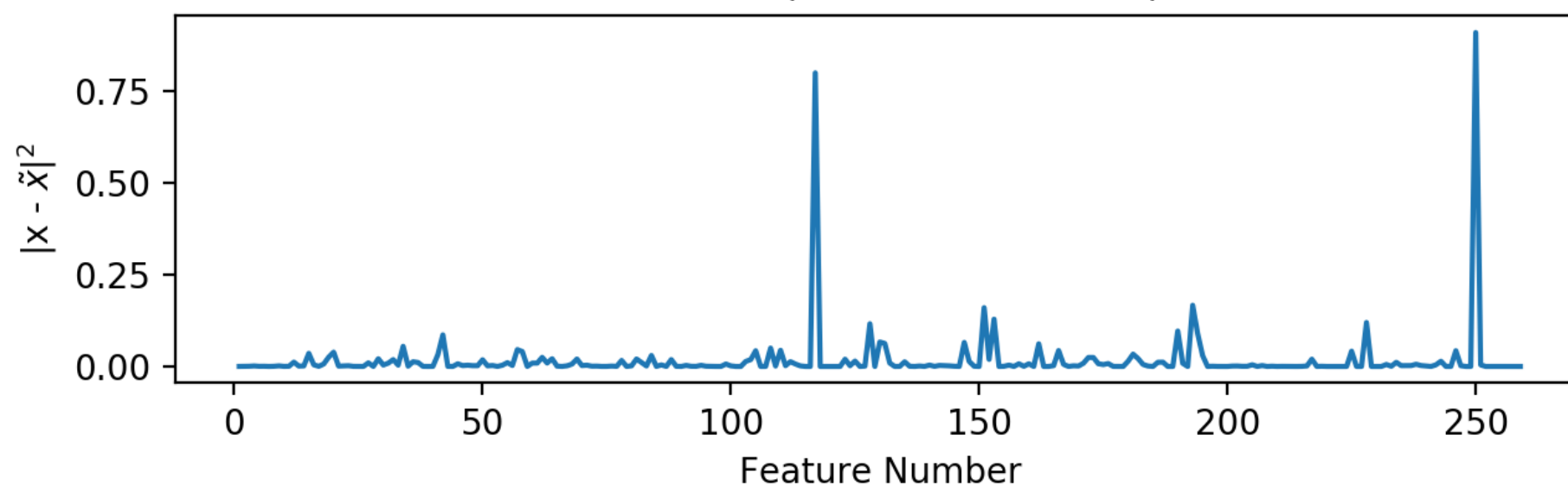
Under configuration

- Isolation Forest: tree = 200, sampling_size = 512
- OneClass-SVM: nu=0.1, gamma=0.1(inverse gaussian width)

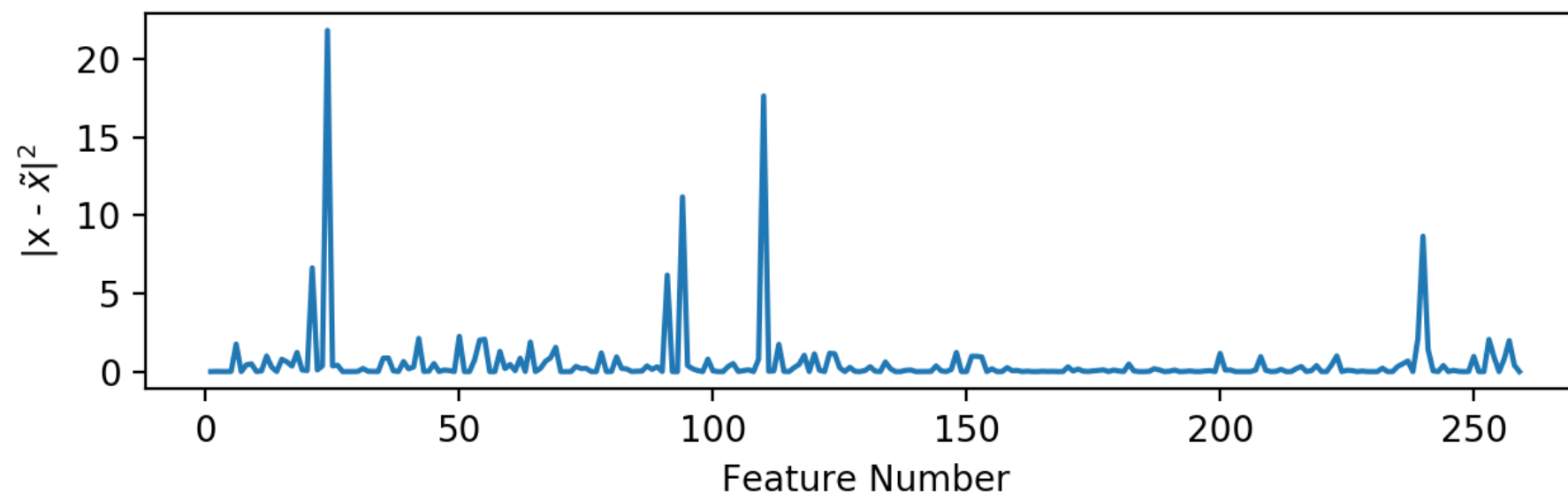
Example from Vanilla

Example of Good and Bad LS

Good LS (Total SE = 4.39)

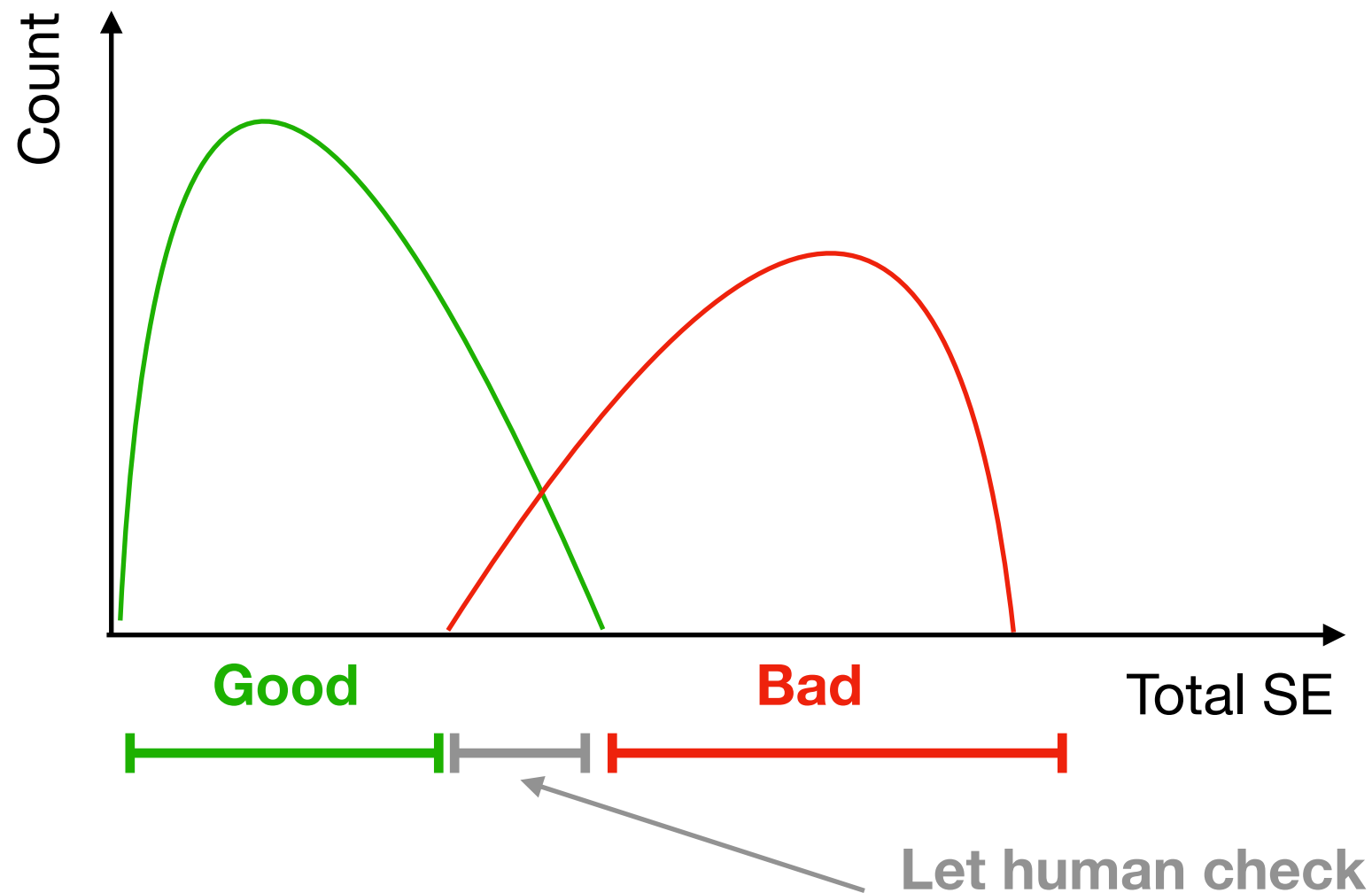


Bad LS (Total SE = 143.65)



Find the cutoff

- **Expect** the total square error (SE) distribution to be like

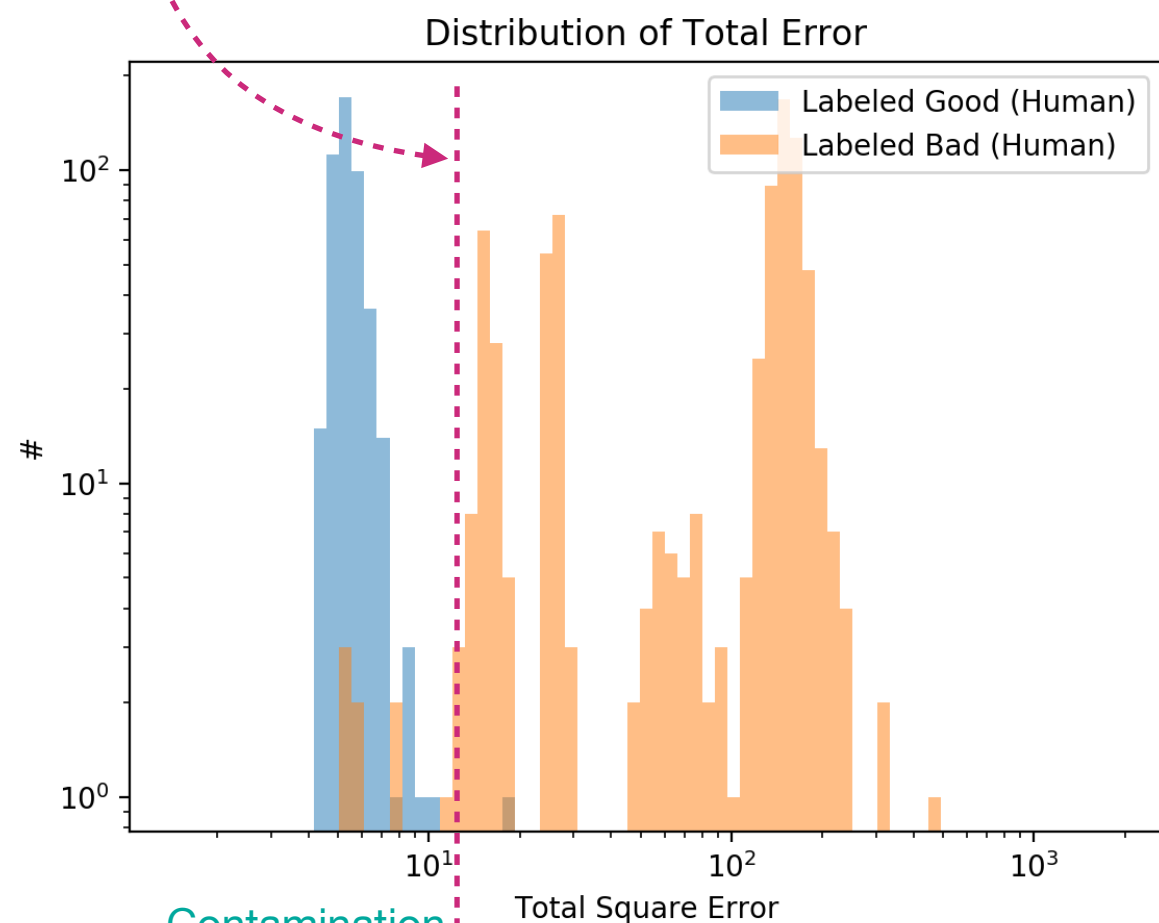


- Next slide is realistic..

Find the cutoff

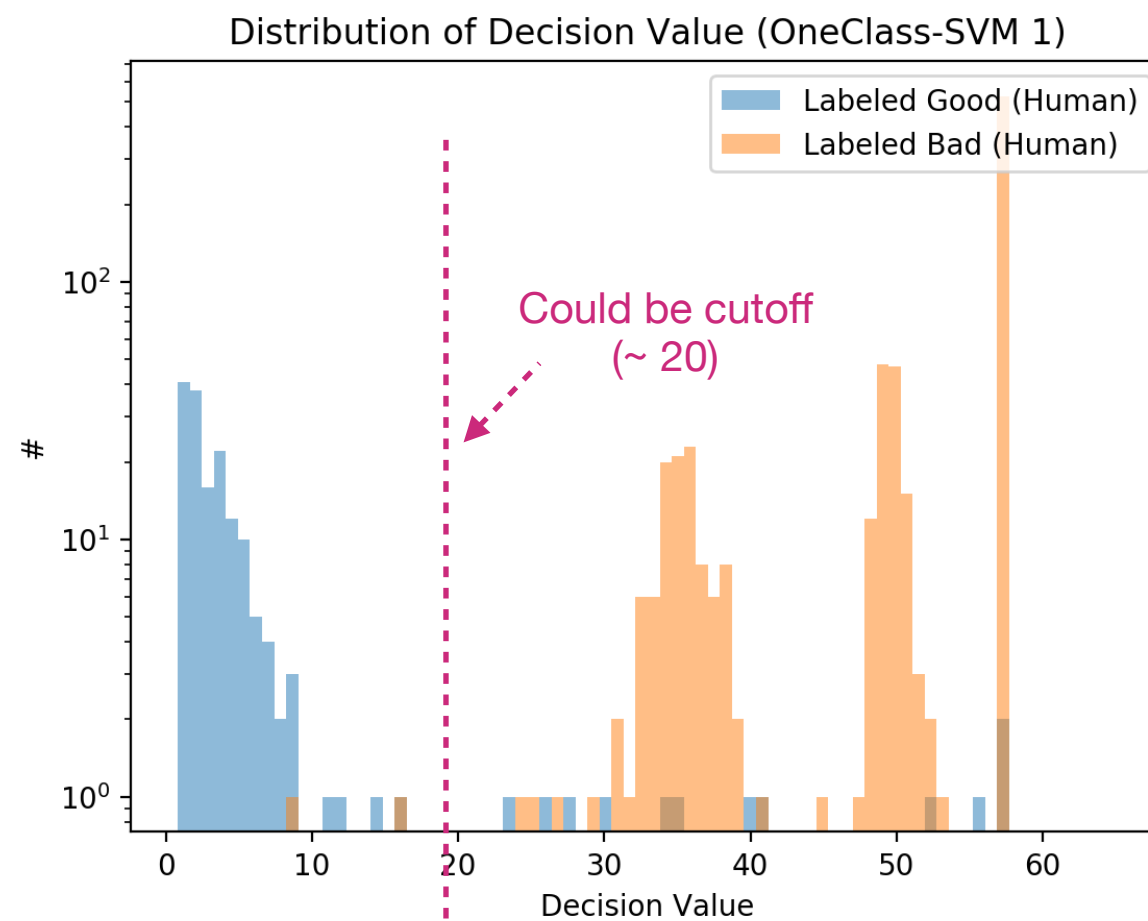
Vanilla AE

Could be cutoff
(~ 10)



Contamination
~ 1.5% in
Good LS

One-Class SVM



Contamination
~ 1.1% in
Good LS

Contamination
~ 1.5% in
Bad LS

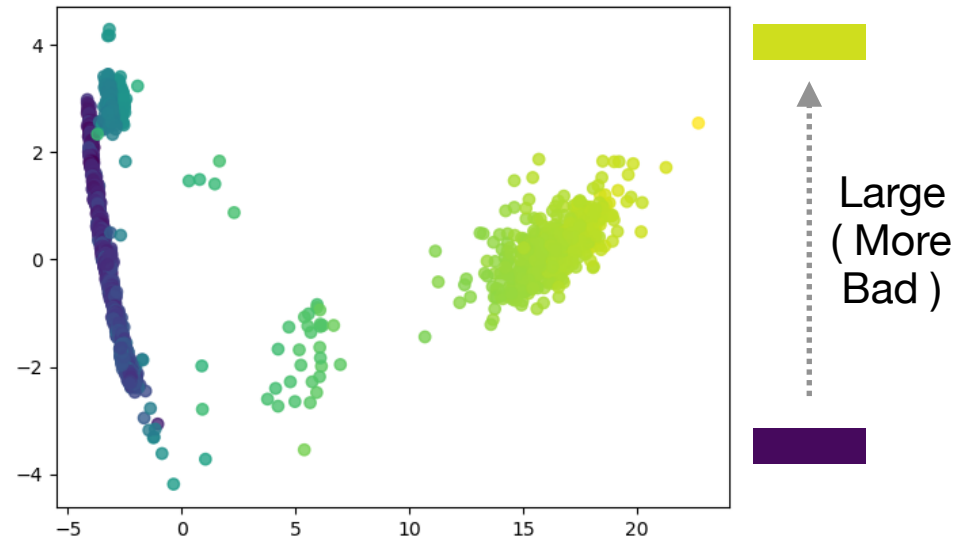
Spot the same Bad => Good LS

Extended Investigation

Loss value from Vanilla AE

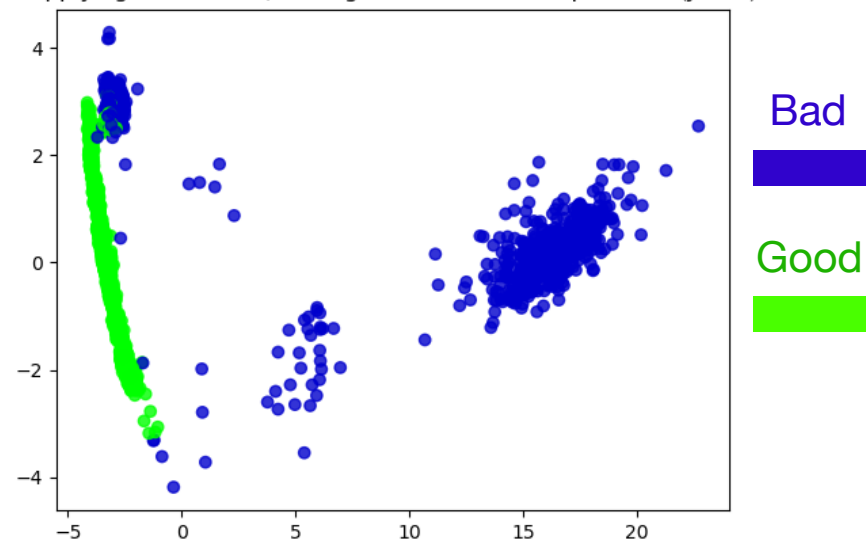
Loss as color shading

Loss from AE data, testing set visual in Principal Basis (JetHT)

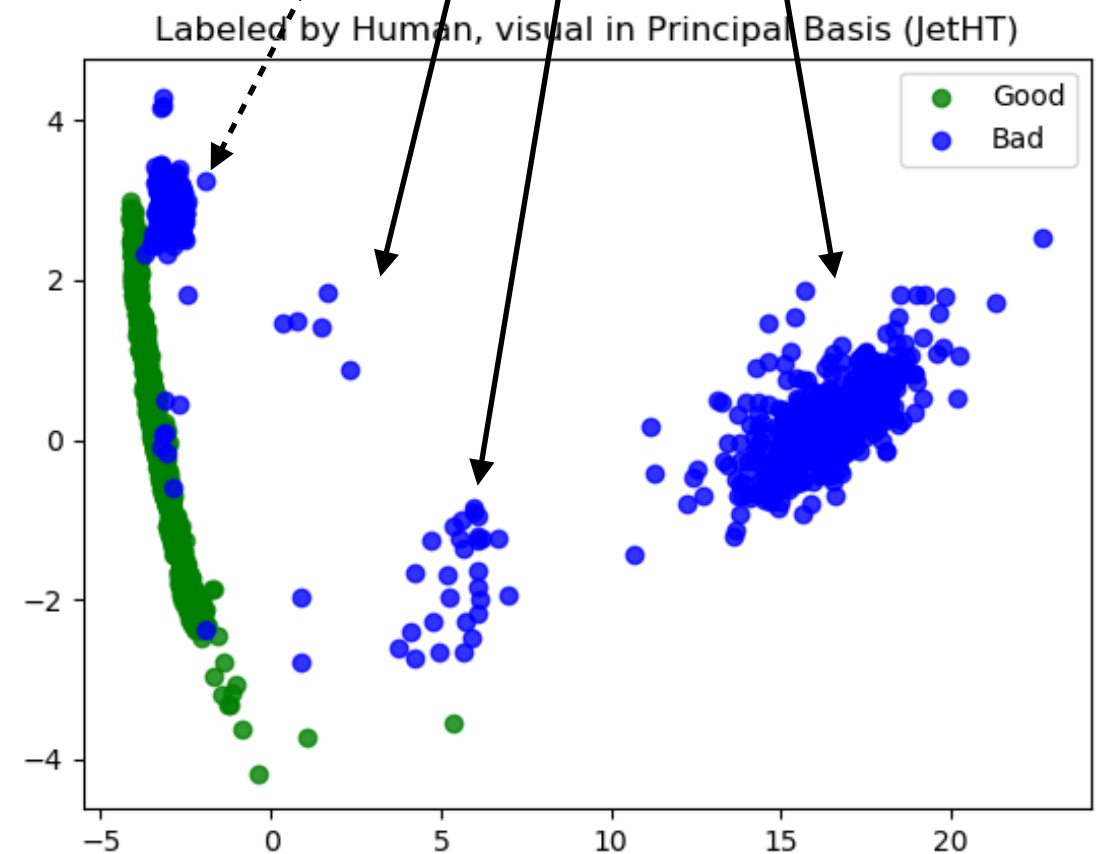
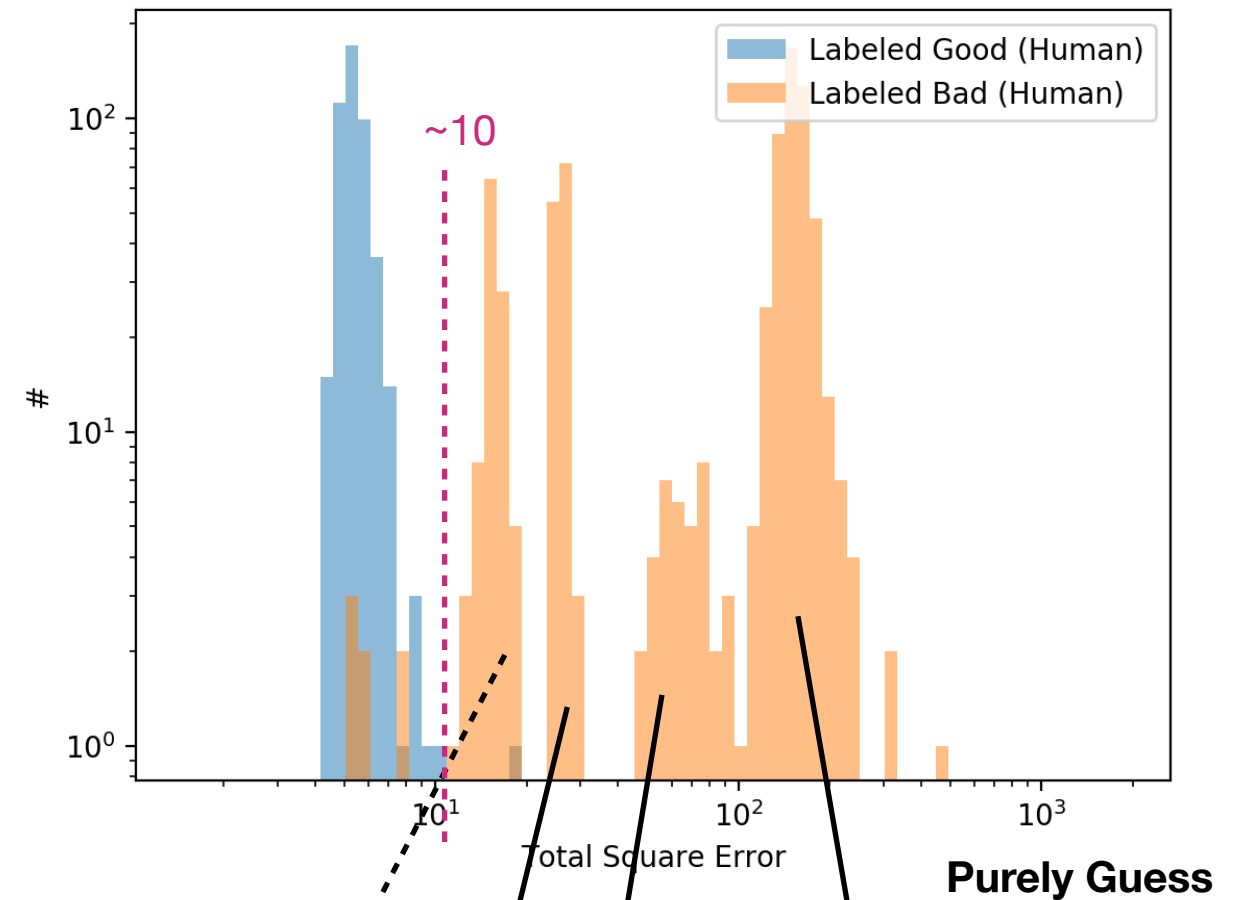


Applying cutoff (MSE > 10.0 is bad)

Applying cutoff in AE, testing set visual in Principal Basis (JetHT)



Distribution of Total Error



Suspicious spot for 2018 datasets

- ~100k good LS and only ~1k in bad LS
- Found ~300 NaN value in some columns from new datasets
- Would be perfect if one who might use new datasets participate to help each other to inspect and discuss (not only for the model)

Backup Slide

Sparse Model

- Unsupervised
- Similar to Vanilla Autoencoder
- Tweak by **L1 Regularization (Prevent overfitting)**

$$\mathcal{L}_{\text{tot}} \equiv \frac{1}{N} \sum_i^N |x_i - \tilde{x}_i|^2 + \lambda \sum_j ||w_j||$$

Set $\lambda = 1e - 4$

Contractive Model

- Unsupervised
- Similar to Vanilla Autoencoder
- Tweak by **Jacobi Matrix (Prevent variation in data)**

$$\mathcal{L} = \frac{1}{N} \sum_i^N |x_i - \tilde{x}_i|^2 + \boxed{\lambda ||J_h(x)||^2}$$

$$\text{Set } \lambda = 1e - 4$$

$$\text{Where } ||J_h(x)||^2 \equiv \sum_{ij} \left(\frac{\partial h_j}{\partial x_i} \right)^2$$

Contractive Model

- Case PReLU activ. fn.

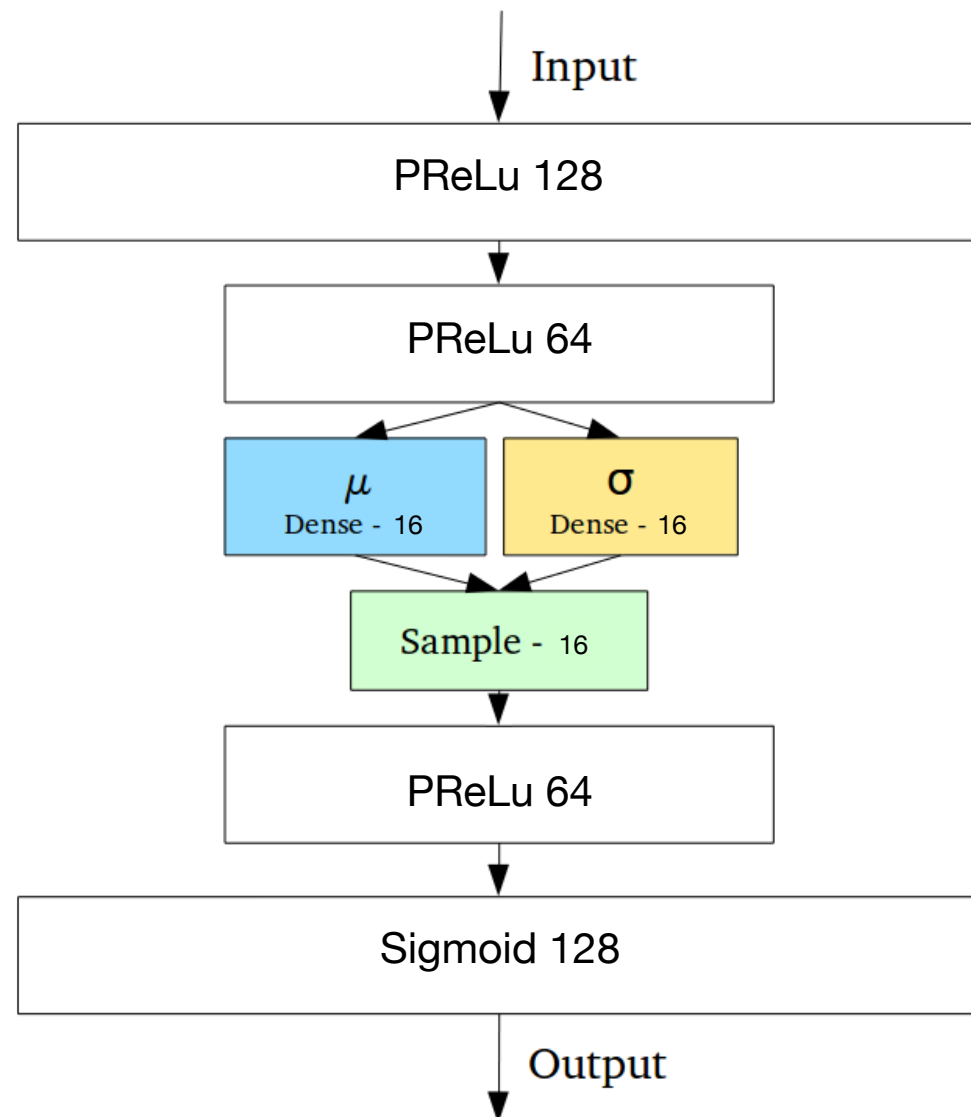
$$||J_h(x)||^2 \equiv \sum_j [\alpha_j H(-(w_{ji}x^i + b_j)) + H(w_{ji}x^i + b_j)] \sum_i (w_{ji})^2$$

- Case Sigmoid activ. fn.

$$||J_h(x)||^2 \equiv \sum_j [h_j * (I - h_j)] \sum_i (w_{ji})^2$$

Variational Model

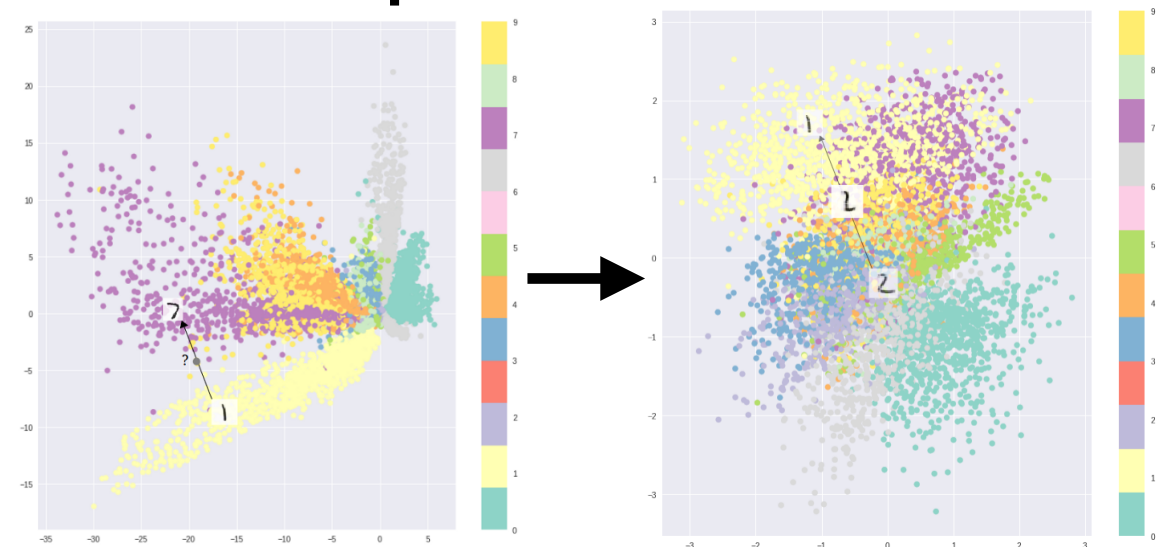
- Tweak by **Random Sampling in Encoding vector**
(**Remove discontinuity in Latent Space**)



$$\mathcal{Z}_i \equiv \mathcal{N}(\mu_i, \sigma_i)$$

"Random new sampling by gaussian"

Ex: Latent space in MNIST



Variational Model

- Kullback–Leibler divergence

$$\mathcal{D}_{\text{KL}}(p|q) \equiv \langle \log p - \log q \rangle$$

- where p is observed value, and q is approx. fn.
- Since our q is gaussian, then

$$\mathcal{L}_{\text{tot}} \equiv \mathcal{L}_{\text{MSE}} + \frac{1}{2} \sum_i (\mu_i^2 + \sigma_i^2 - \log(\sigma_i^2) - 1)$$

Isolation Forest

- Ensemble Forest from tree by subsampling (Ψ)
 - Iteratively picking up features and random value to contract the node (equivalent to step fn.)
- Anomaly score likely to be average depth of the instance over forest

$$s(x, \Psi) \equiv e^{-\langle h(x) \rangle / c(\Psi)}$$

- Where
 - $h(x)$ is the depth in tree h
 - $c(\Psi)$ normalization factor growing as $\log_2(\Psi)$ from branching