Visualization and Explainability |

Swati Mishra

Applications of Machine Learning (4AL3)

Fall 2024

McMaster University
ENGINEERING

Review

- Principal Component Analysis
- Dimensionality Reduction view
- Linear Dimensionality Reduction Technique
- Computing PCA components

Visualization

- Visualization in Machine Learning is the most critical and challenging part of the process.

- We need visualization for
  - Understanding training data
  - Inspecting the model
  - Communicating model results
  - Dealing with high dimensional data

Understanding Training Data

Feature order | ☐ Reverse order | Feature search (regex enabled)

Features: ☑ int(6) ☑ string(9)

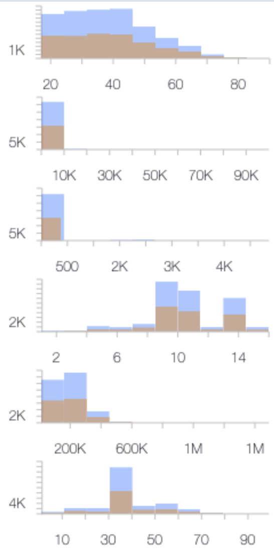
train test

Numeric Features (6)

Chart to show: Standard

☐log ☐expand ☐percentages

	count	missing	mean	std dev	zeros	min	median	max
Age								
train	32.6k	0%	38.58	13.64	0%	17	37	90
test	16.3k	0%	38.77	13.85	0%	17	37	90
Capital Gain								
train	32.6k	0%	1,077.65	7,385.29	91.67%	0	0	100k
test	16.3k	0%	1,081.91	7,583.94	91.87%	0	0	100k
Capital Loss								
train	32.6k	0%	87.3	402.96	95.33%	0	0	4,356
test	16.3k	0%	87.9	403.11	95.31%	0	0	3,770
Education-Num								
train	32.6k	0%	10.08	2.57	0%	1	10	16
test	16.3k	0%	10.07	2.57	0%	1	10	16
fnlwgt								
train	32.6k	0%	190k	106k	0%	12.3k	178k	1.48M
test	16.3k	0%	189k	106k	0%	13.5k	178k	1.49M
Hours per week								
train	32.6k	0%	40.44	12.35	0%	1	40	99
test	16.3k	0%	40.39	12.48	0%	1	40	99

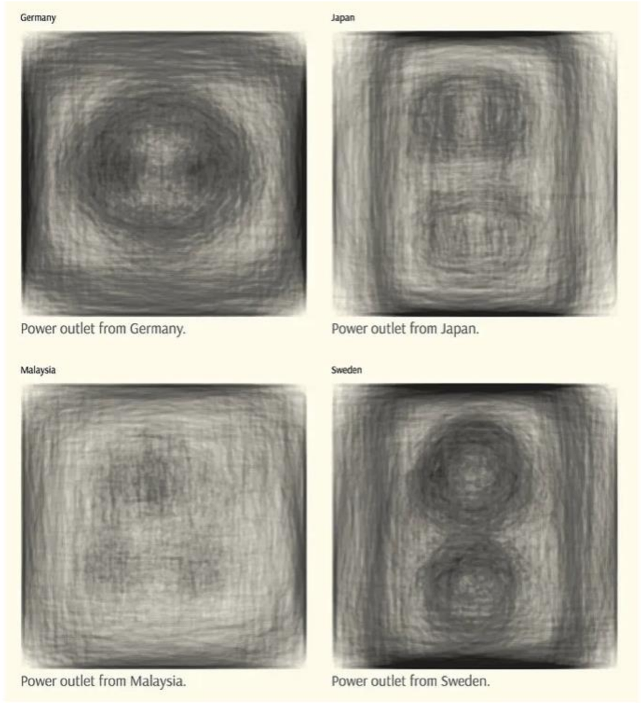


https://pair-code.github.io/facets/

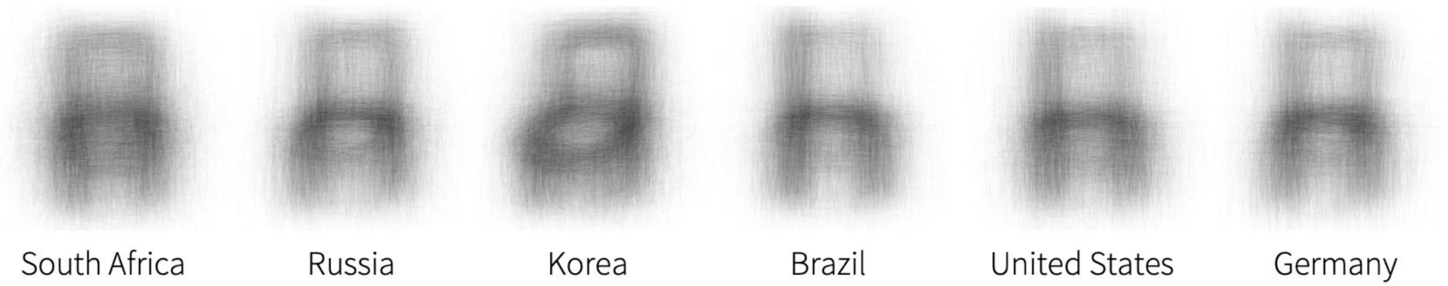
Understanding Training Data

https://quickdraw.withgoogle.com

Understanding Training Data

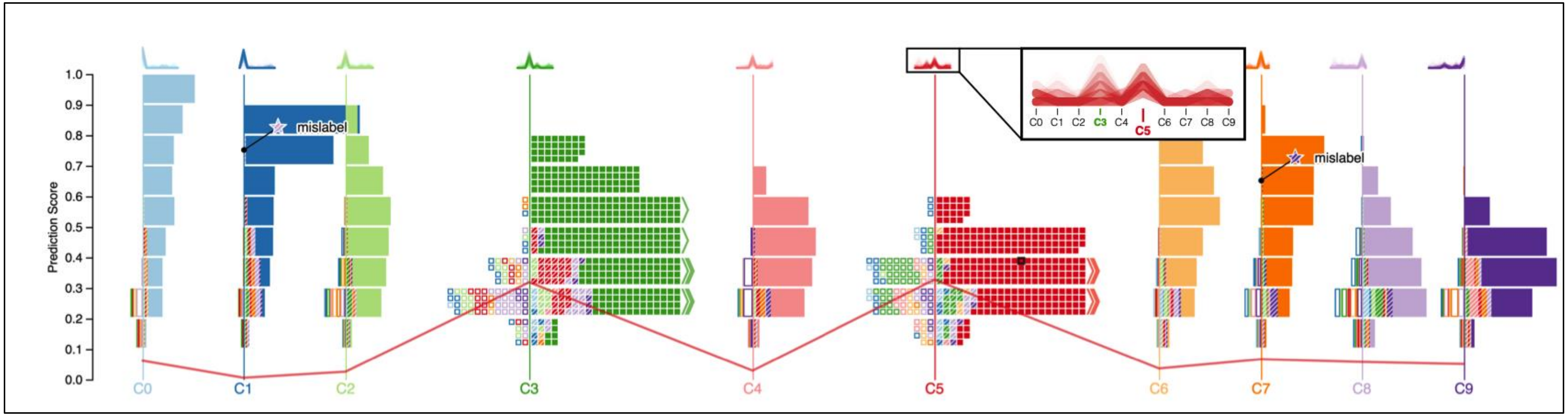


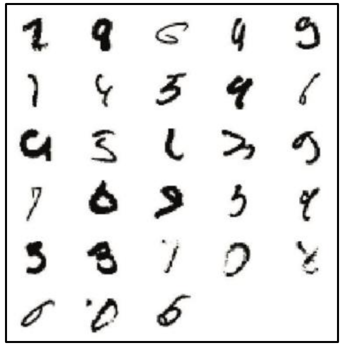
oh crap I forgot my converter



https://medium.com/@enjalot/machine-learning-for-visualization-927a9dff1cab

Inspecting the model





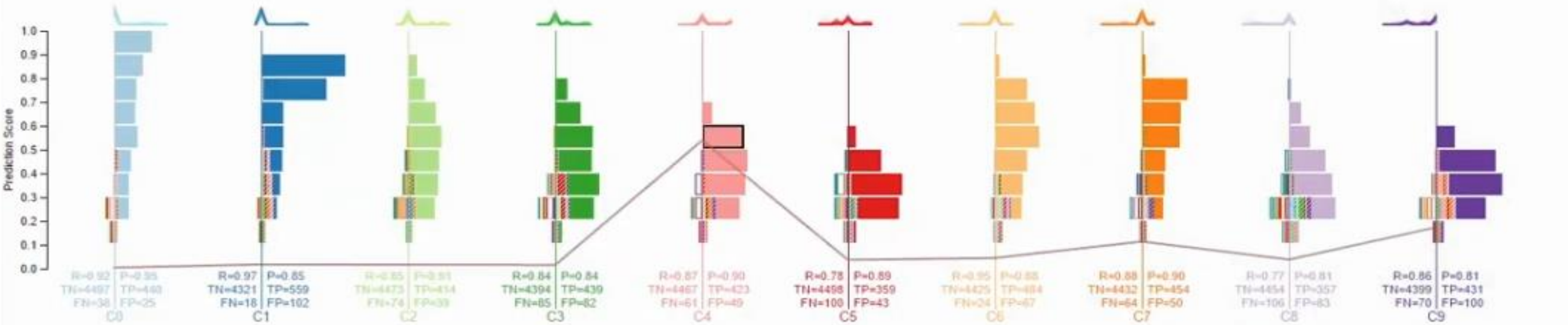
https://www.microsoft.com/en-us/research/video/squares-supporting-interactive-performance-analysis-multiclass-classifiers-2/

McMaster University

squares

https://squares.azurewebsites.net

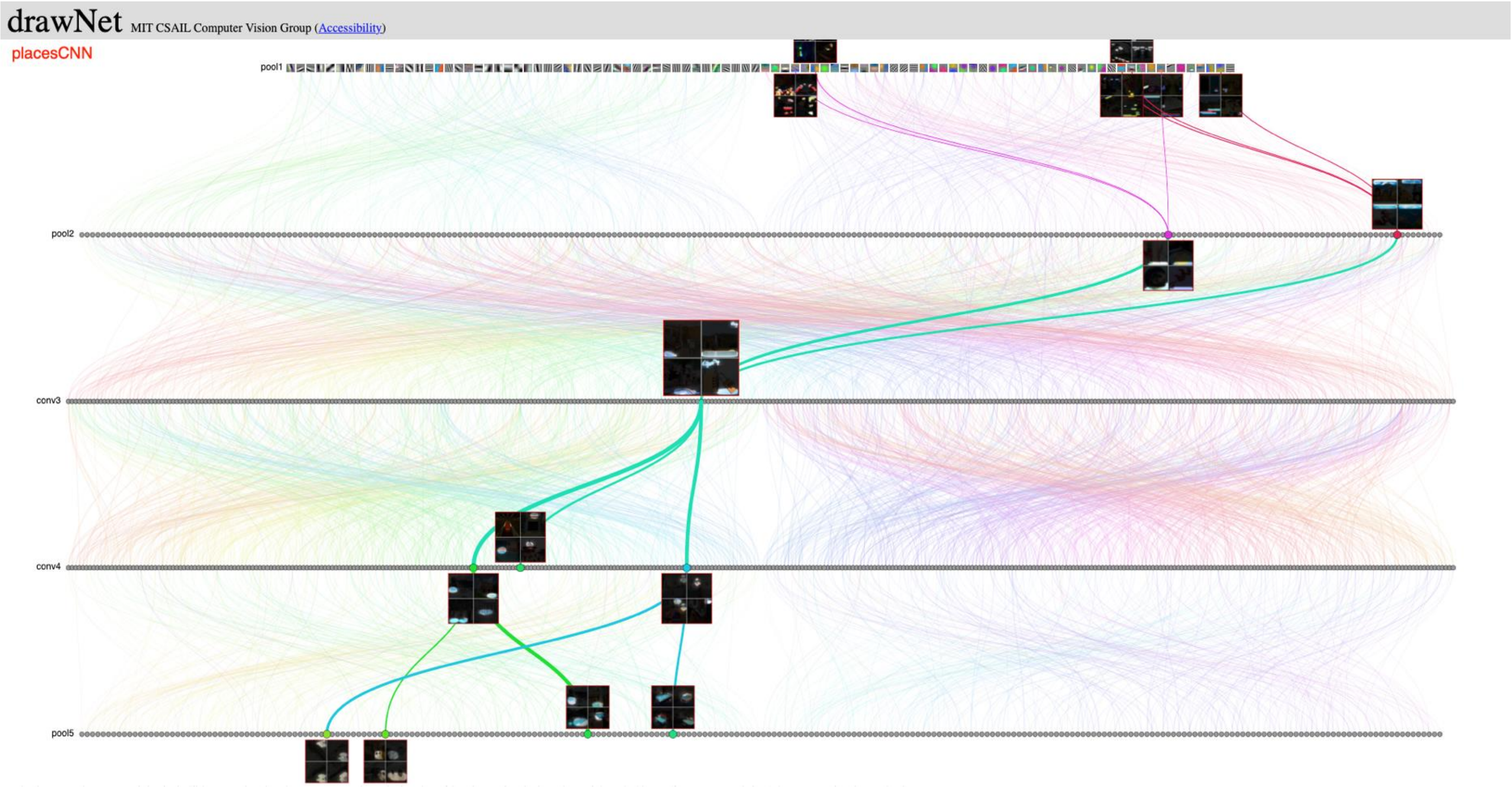
File View Help



Dataset: mnist_randomforest.csv | 10 Classes, 5000 Instances, 5000 Shown | Acc.: 0.87, Prec.: 0.87, Recall: 0.87, TP: 436.0, FP: 64.0, TN: 4436.0, FN: 64.0 | 1 Hovered, 0 Selected

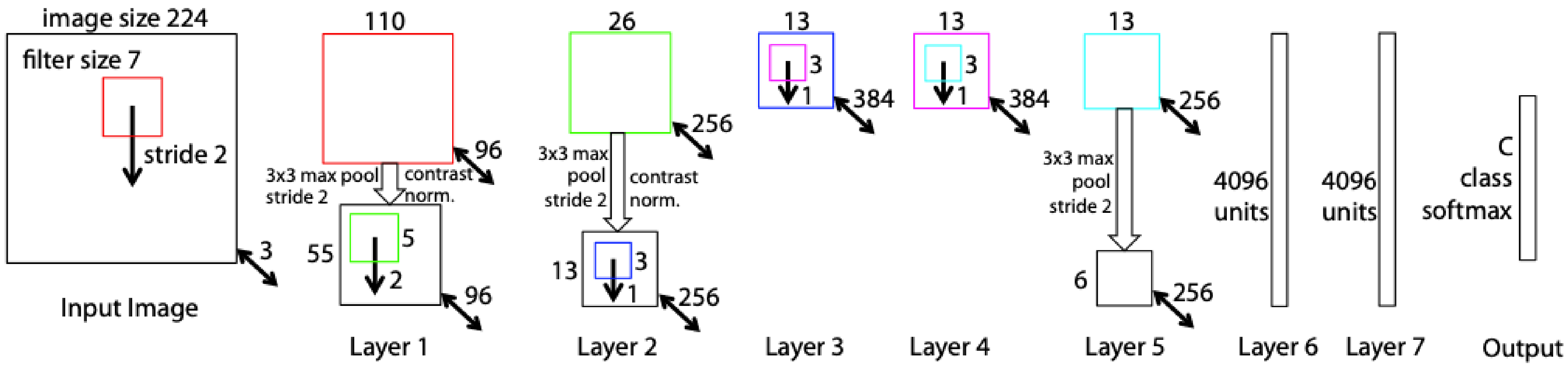
★	Image	TRUE	Assigned	Correct	Prediction Score	C0	C1	C2	C3	C4	C5	C6	C7
☆		C7	C7	1	0.5849	0.0103	0.0226	0.0646	0.0330	0.0621	0.0279	0.0383	0.5849
☆		C1	C1	1	0.3576	0.0111	0.3576	0.1180	0.0431	0.0450	0.0447	0.0589	0.0235
☆		C7	C7	1	0.6777	0.0185	0.0310	0.0485	0.0324	0.0328	0.0491	0.0152	0.6777
☆		C9	C9	1	0.2158	0.0146	0.0634	0.0233	0.1524	0.1435	0.1895	0.0541	0.0745
☆		C5	C5	1	0.2964	0.0346	0.2051	0.1039	0.1367	0.0286	0.2964	0.0606	0.0314
☆		C3	C3	1	0.2365	0.0121	0.0946	0.0367	0.2365	0.0617	0.1952	0.0599	0.1379
☆		C2	C2	1	0.3628	0.0069	0.1414	0.3628	0.1552	0.0153	0.0526	0.1246	0.0190
☆		C1	C1	1	0.8319	0.0017	0.8319	0.0342	0.0166	0.0087	0.0201	0.0123	0.0258
☆		C2	C2	1	0.7116	0.0192	0.0188	0.7116	0.0230	0.0390	0.0157	0.0597	0.0198
☆		C1	C1	1	0.7070	0.0014	0.7070	0.0384	0.0223	0.0234	0.0203	0.0194	0.0637
☆		C7	C7	1	0.7910	0.0081	0.0135	0.0240	0.0252	0.0417	0.0272	0.0079	0.7910
☆		C1	C1	1	0.8477	0.0014	0.8477	0.0253	0.0172	0.0068	0.0230	0.0165	0.0215
☆		C6	C6	1	0.3483	0.0364	0.0256	0.0358	0.0728	0.1373	0.1377	0.3483	0.0257
☆		C8	C8	1	0.3779	0.0115	0.1461	0.0314	0.0668	0.0879	0.0825	0.0701	0.0486
☆		C2	C2	1	0.3196	0.0638	0.0291	0.3196	0.1187	0.0546	0.0869	0.1316	0.0617
☆		C0	C0	1	0.9639	0.9639	0.0003	0.0020	0.0026	0.0008	0.0198	0.0053	0.0016
☆		C4	C4	1	0.5391	0.0044	0.0179	0.0161	0.0156	0.5391	0.0369	0.0449	0.1140

Communicating Model Results:



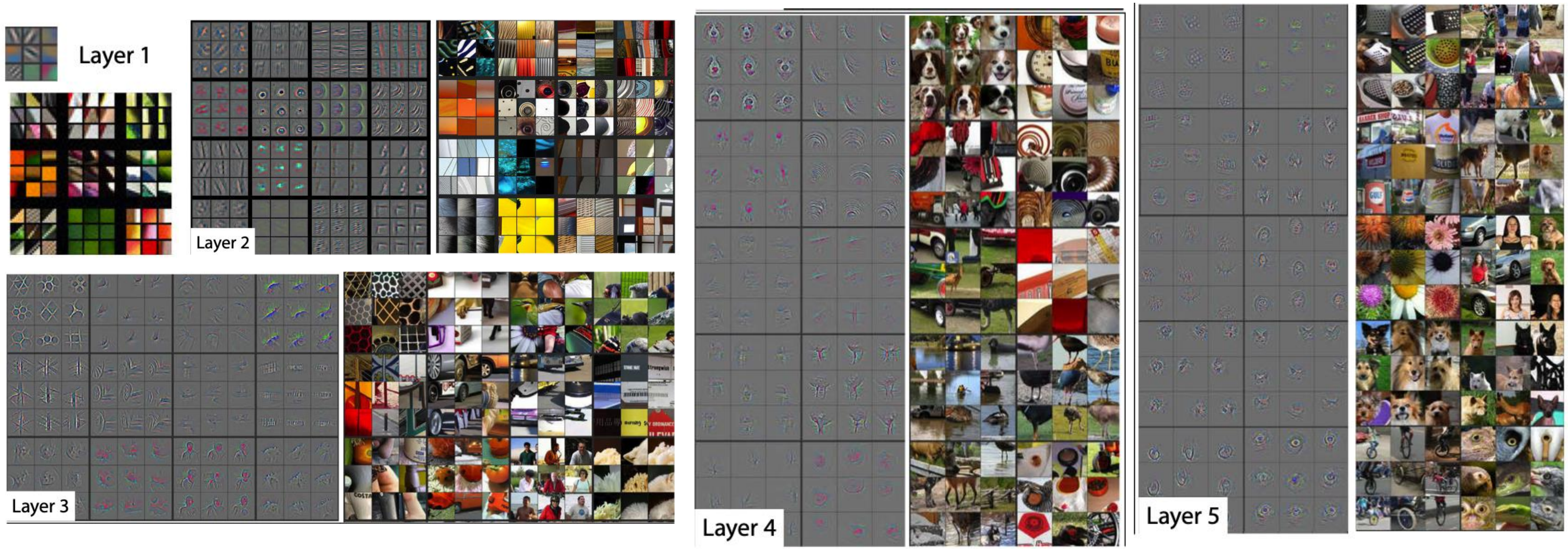
Source: https://people.csail.mit.edu/torralba/research/drawCNN/drawNet.html

Dealing with High Dimensionality:



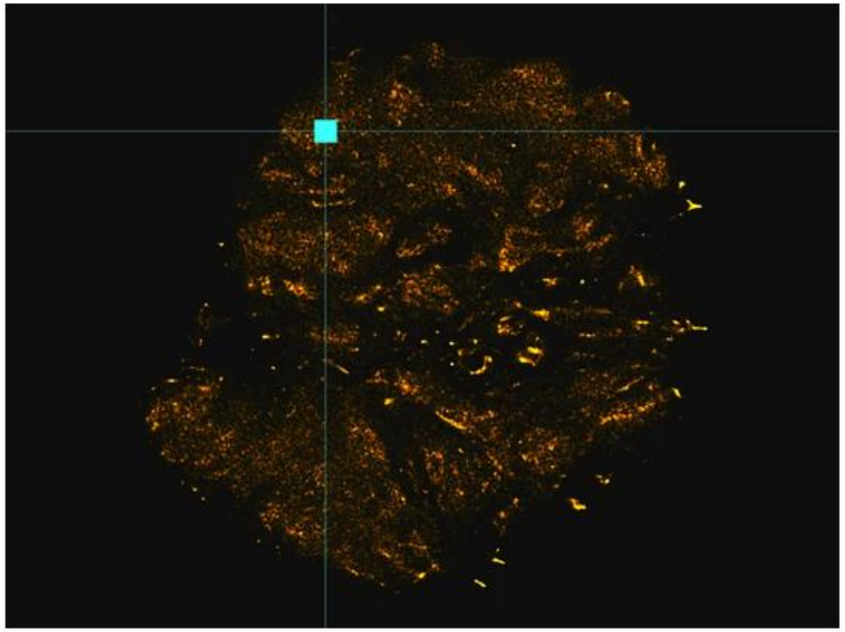
McMaster University

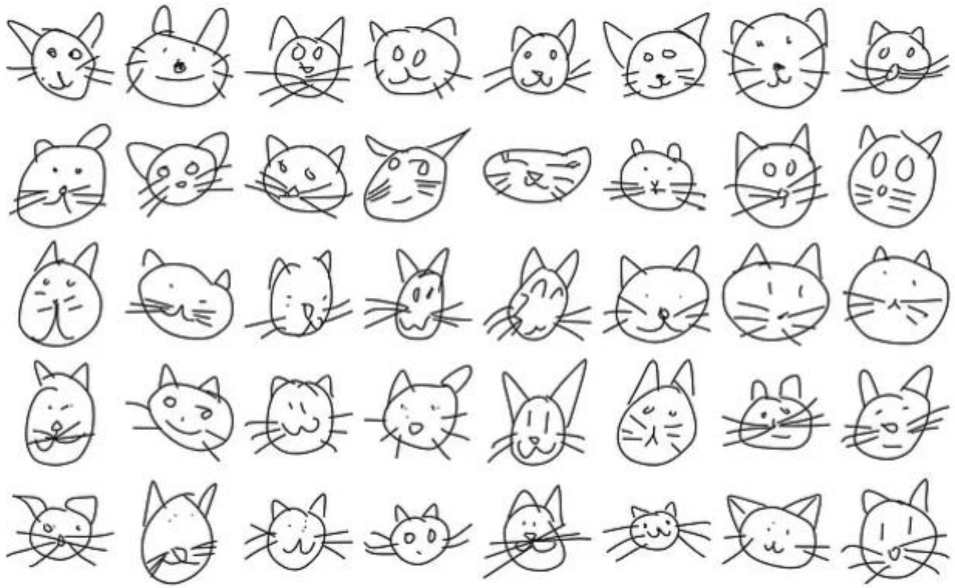
Dealing with High Dimensionality:



Dealing with High Dimensionality:

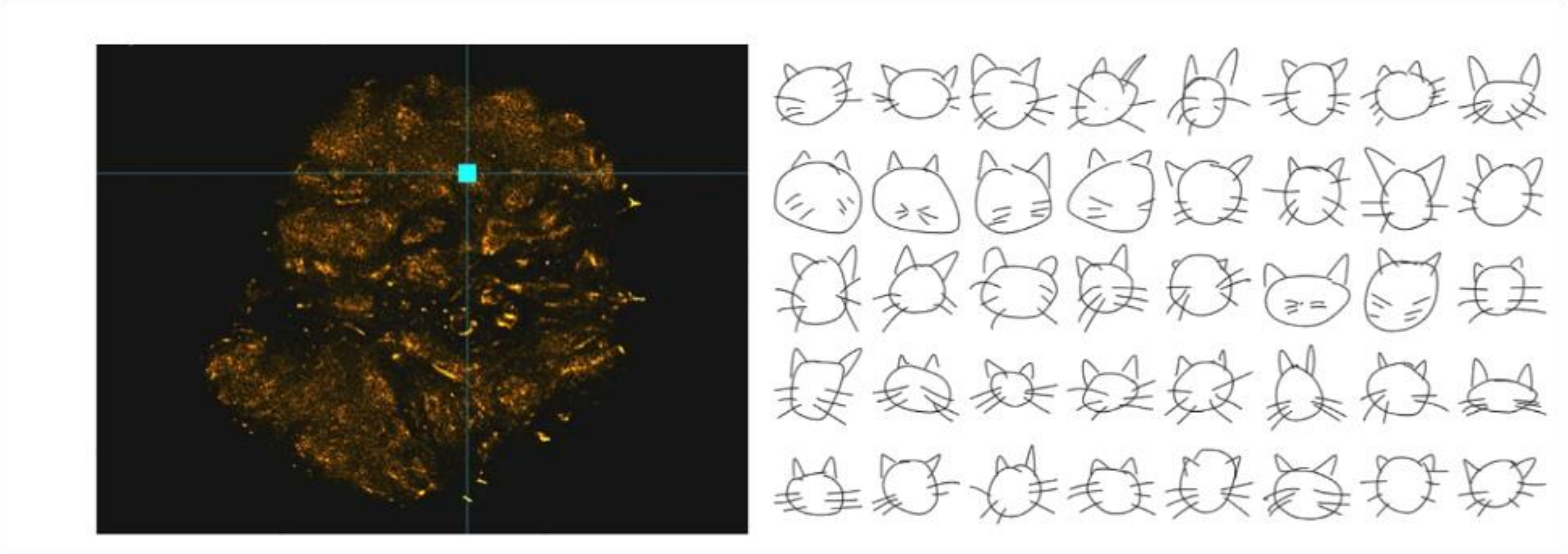
T-SNE is a popular technique





Dealing with High Dimensionality:

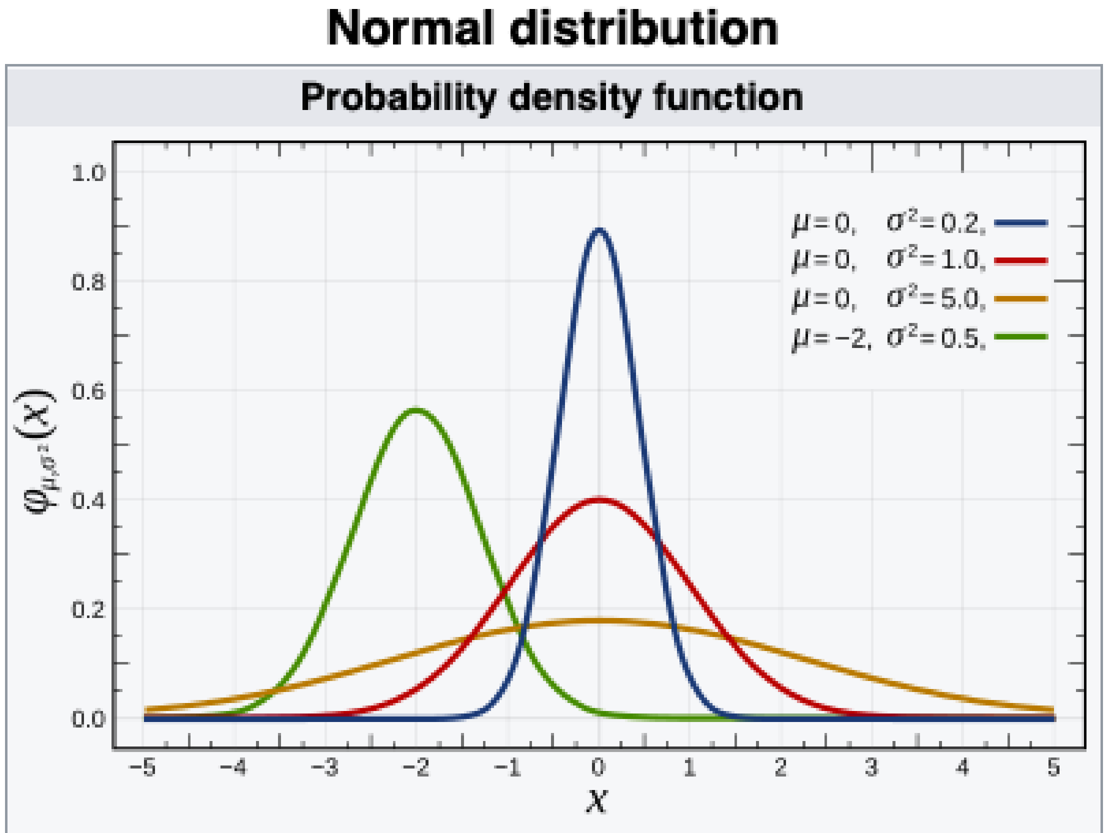
T-SNE is a popular technique



t-SNE

- Gaussian Distribution:
  - a normal distribution or Gaussian distribution is a type of continuous probability distribution for a real-valued random variable.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



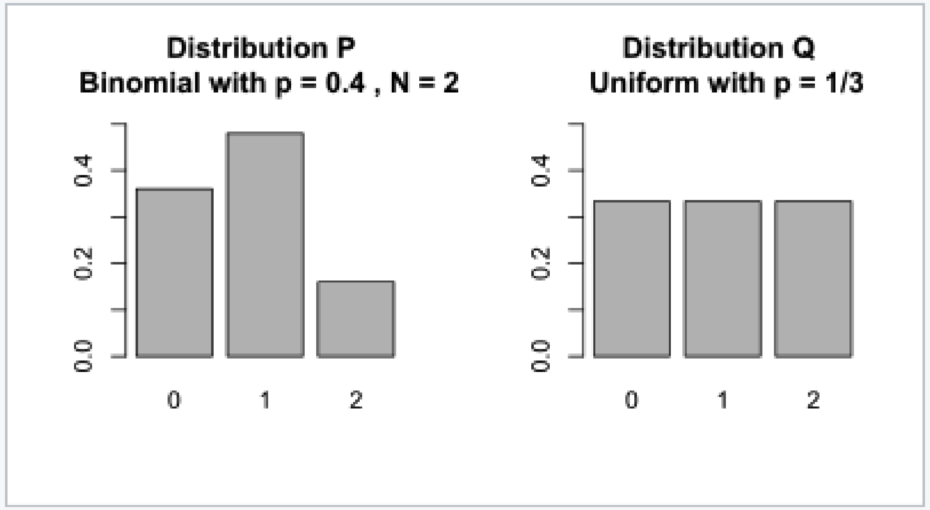
The red curve is the *standard normal distribution*.

Picture Source: Wikipedia

t-SNE

- Kullback–Leibler (KL) divergence :
  - A measure of how one reference probability distribution P is different from a second probability distribution Q.
  - Also called relative entropy and I-divergence

$$D_{\text{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \, \log\left(\frac{P(x)}{Q(x)}\right)$$



Two distributions to illustrate relative entropy

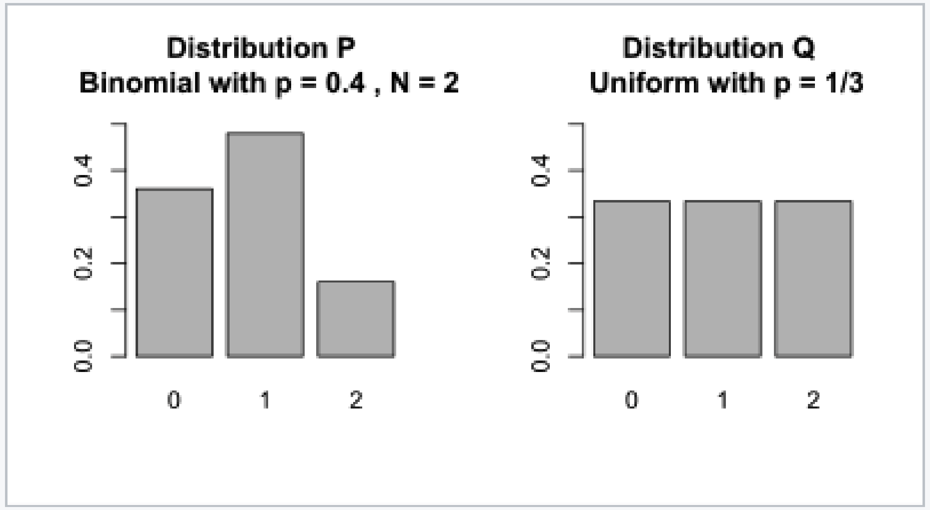
Picture Source: Wikipedia

t-SNE

- Kullback–Leibler (KL) divergence :
  - A measure of how one reference probability distribution P is different from a second probability distribution Q.
  - Also called relative entropy and I-divergence

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x)\ \log\left(\frac{P(x)}{Q(x)}\right)$$

$$DKL\ (P||Q) = \frac{9}{25} ln\left(\frac{\frac{9}{25}}{\frac{1}{3}}\right) + \frac{12}{25} ln\left(\frac{\frac{12}{25}}{\frac{1}{3}}\right) + \frac{4}{25} ln(\frac{\frac{4}{25}}{\frac{1}{3}})$$



Two distributions to illustrate relative entropy

x	0	1	2
Distribution $P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
Distribution $Q(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

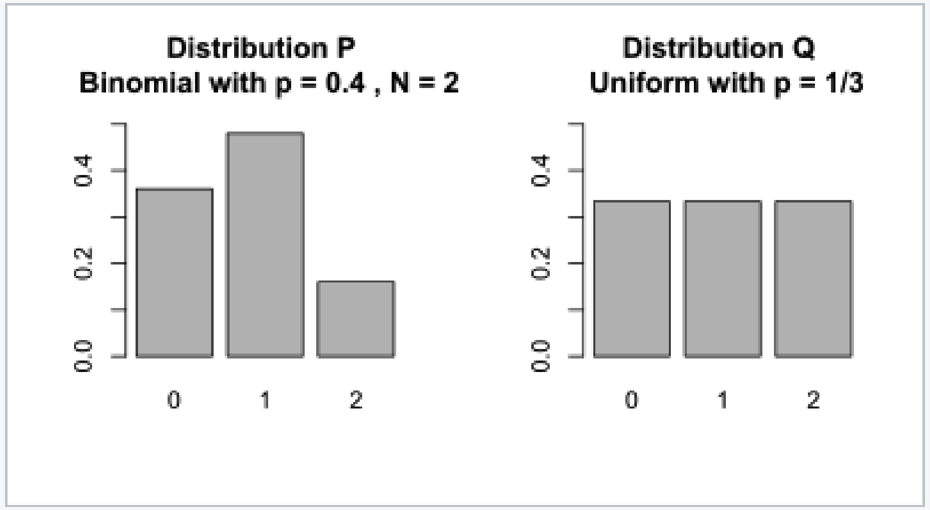
Picture Source: Wikipedia

t-SNE

- Kullback–Leibler (KL) divergence :
  - A measure of how one reference probability distribution P is different from a second probability distribution Q.
  - Also called relative entropy and I-divergence

$$D_{\text{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \, \log\left(\frac{P(x)}{Q(x)}\right)$$

$$DKL\,(P||Q) = \frac{9}{25} ln\left(\frac{\frac{9}{25}}{\frac{1}{3}}\right) + \frac{12}{25} ln\left(\frac{\frac{12}{25}}{\frac{1}{3}}\right) + \frac{4}{25} ln(\frac{\frac{4}{25}}{\frac{1}{3}})$$



Two distributions to illustrate relative entropy

x	0	1	2
Distribution $P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
Distribution $Q(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Symmetry is an issue!

Picture Source: Wikipedia

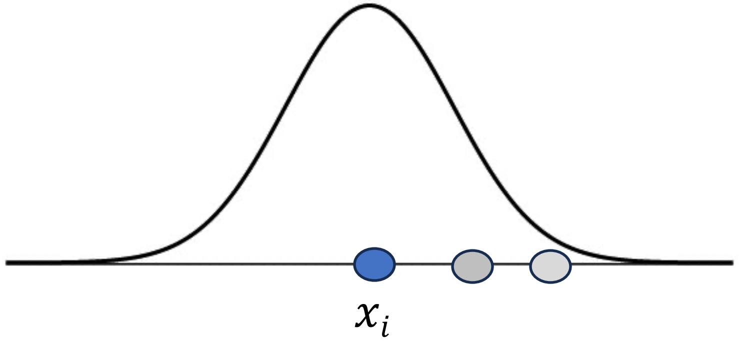
t-SNE

- For a N high-dimensionality points, Stochastic Neighbor Embedding starts by converting the high-dimensional Euclidean distances between datapoints into conditional probabilities that represent similarities.

- The similarity of datapoint x_j to datapoint x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i .

- For nearby points, $p_{j|i}$ is relatively high.

- For widely separated points , $p_{j|i}$ is infitesimal high.



Original Paper: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

t-SNE

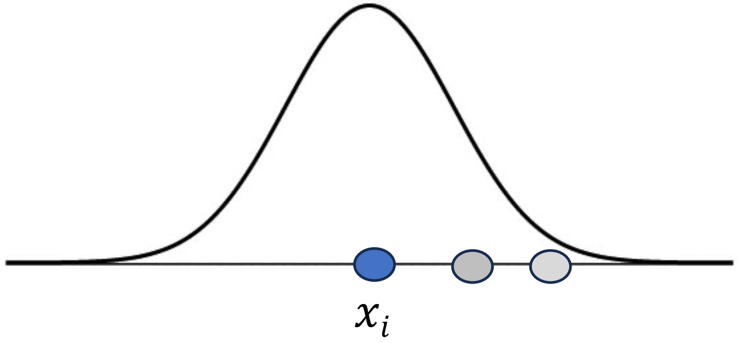
- For a N high-dimensionality points, Stochastic Neighbor Embedding starts by converting the high-dimensional Euclidean distances between datapoints into conditional probabilities that represent similarities.

- The similarity of datapoint x_j to datapoint x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i .

- For nearby points, $p_{j|i}$ is relatively high.

- For widely separated points , $p_{j|i}$ is extremely small.

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$



t-SNE

- Since we are only interested in modelling pair wise similarities, $p\ i|i$ is set to zero.
- Consider the low dimensionality counterparts (y_i , y_j) of high dimensionality data points (x_i , x_j)
- The similar the conditional probability, $q_{j|i}$ can be computed.
- Here variance of the Gaussian (σ) is set to $\sqrt{\frac{1}{2}}$

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)} \qquad q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq i}\exp\left(-\|y_i - y_k\|^2\right)}$$

t-SNE

- SNE minimizes the sum of Kullback-Leibler (KL) divergences over all datapoints using a gradient descent method where,

- P_i represents the conditional probability distribution over all other datapoints given data- point x.

- Q_i represents the conditional probability distribution over all other map points given map point yi

$$\boxed{C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}}$$

Since symmetry is a problem we set:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{N}$$

t-SNE

- SNE minimizes the sum of Kullback-Leibler (KL) divergences over all datapoints using a gradient descent method where,
- P_i represents the conditional probability distribution over all other datapoints given data- point x.
- Q_i represents the conditional probability distribution over all other map points given map point yi

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

Gradient updates

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$$

Without symmetric adjustments

t-SNE

- SNE minimizes the sum of Kullback-Leibler (KL) divergences over all datapoints using a gradient descent method where,
- P_i represents the conditional probability distribution over all other datapoints given data- point x.
- Q_i represents the conditional probability distribution over all other map points given map point yi

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)$$

Gradient updates

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t) \left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)} \right)$$

With symmetric adjustments

McMaster University

t-SNE

- Parameters to be selected:
  - Variance of the Gaussian that is centered over each high-dimensional datapoint

- SNE performs a binary search for the value of σ_i that produces a P_i with a fixed perplexity that is specified by the user.

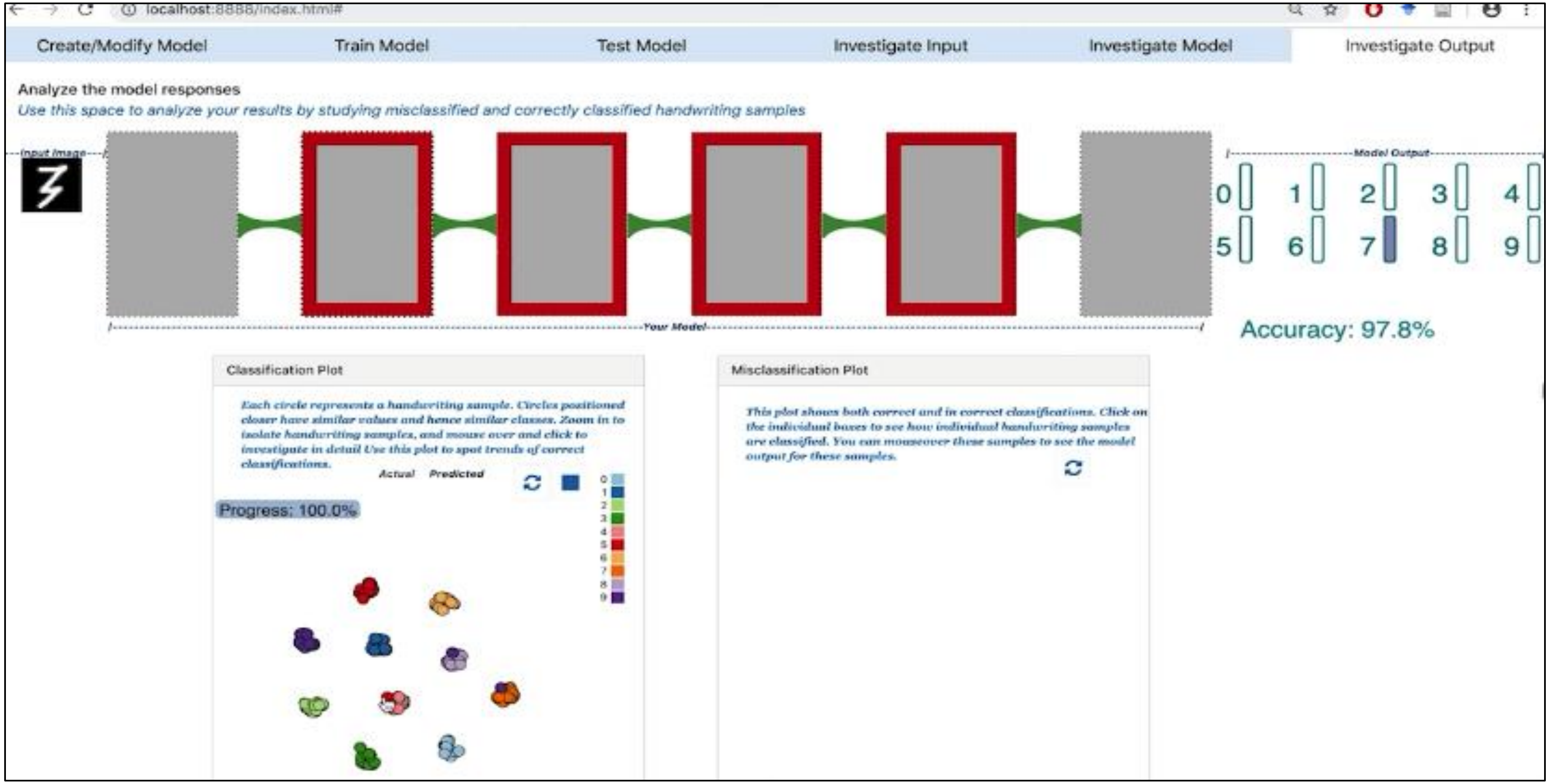
$$Perp(P_i) = 2^{H(P_i)}$$

Where, Shannon entropy of Pi measured in bits:

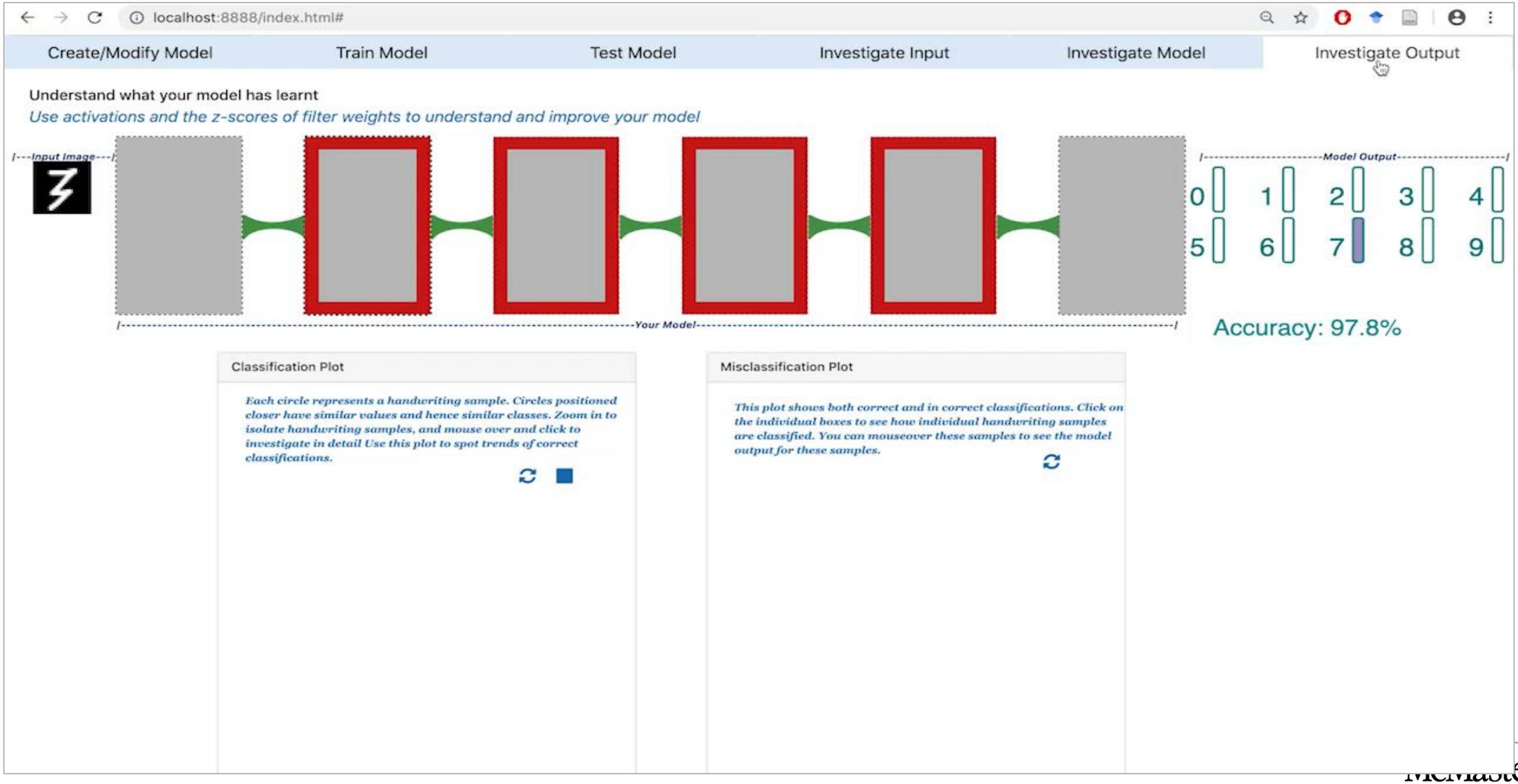
$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

- Step size for gradient descent.

t-SNE



localhost:8888/index.html#
Create/Modify Model
Train Model
Test Model
Investigate Input
Investigate Model
Investigate Output
Analyze the model responses
Use this space to analyze your results by studying misclassified and correctly classified handwriting samples
Input Image
Your Model
Model Output
0 1 2 3 4
5 6 7 8 9
Accuracy: 97.8%
Classification Plot
Each circle represents a handwriting sample. Circles positioned closer have similar values and hence similar classes. Zoom in to isolate handwriting samples, and mouse over and click to investigate in detail Use this plot to spot trends of correct classifications.
Actual
Predicted
Progress: 100.0%
0 1 2 3 4 5 6 7 8 9
Misclassification Plot
This plot shows both correct and in correct classifications. Click on the individual boxes to see how individual handwriting samples are classified. You can mouseover these samples to see the model output for these samples.

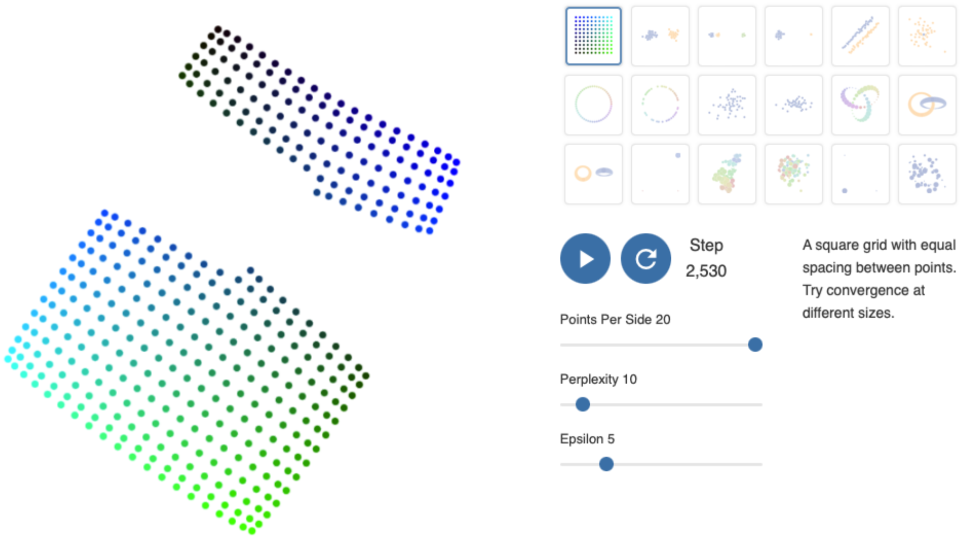


McMaster University

Interpreting t-SNE

How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.



MARTIN WATTENBERG Google Brain
FERNANDA VIÉGAS Google Brain
IAN JOHNSON Google Cloud

Oct. 13 2016
Citation: Wattenberg, et al., 2016

Source: https://distill.pub/2016/misread-tsne/

Readings

Reference Material:

Links included in the slides.

Thank You