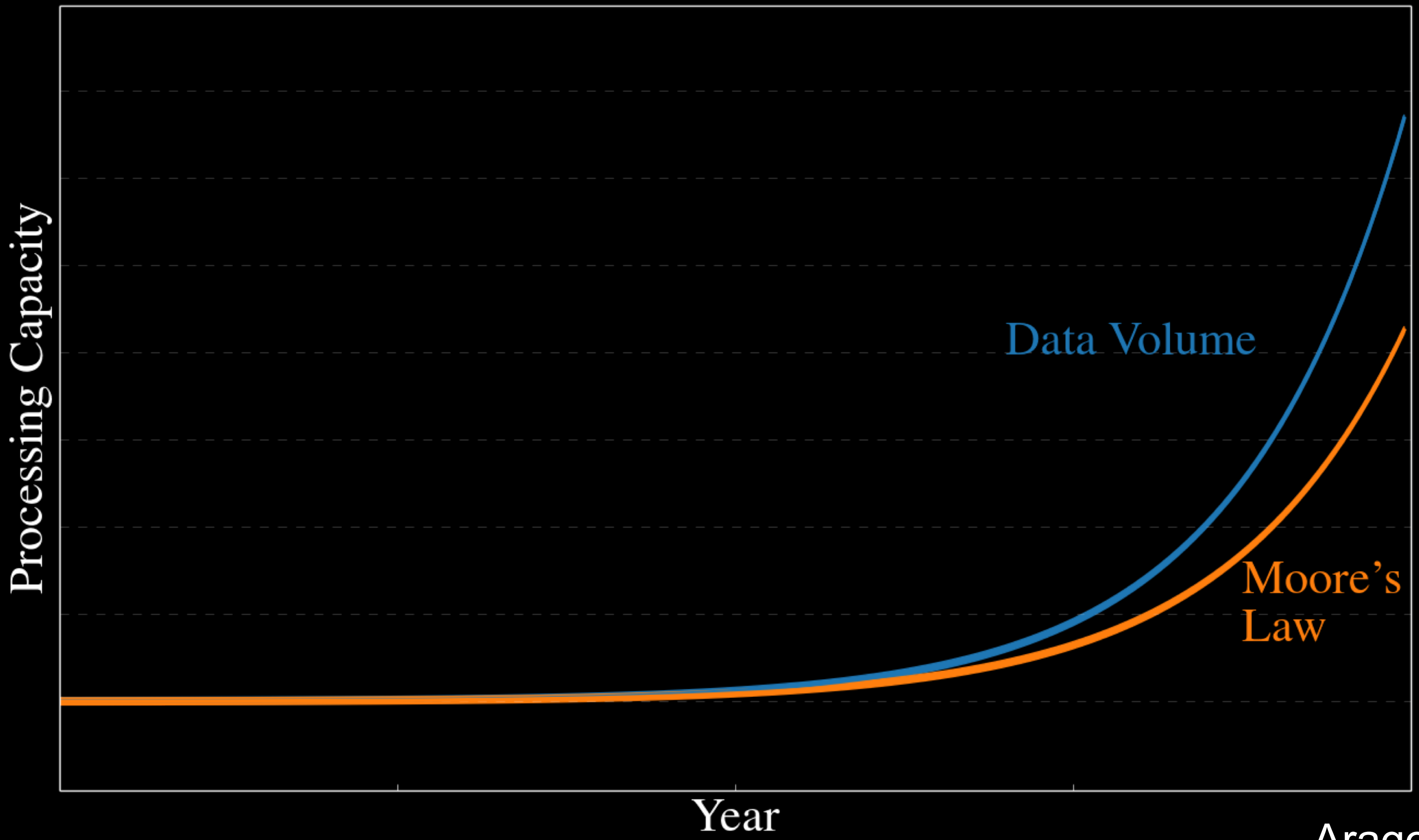
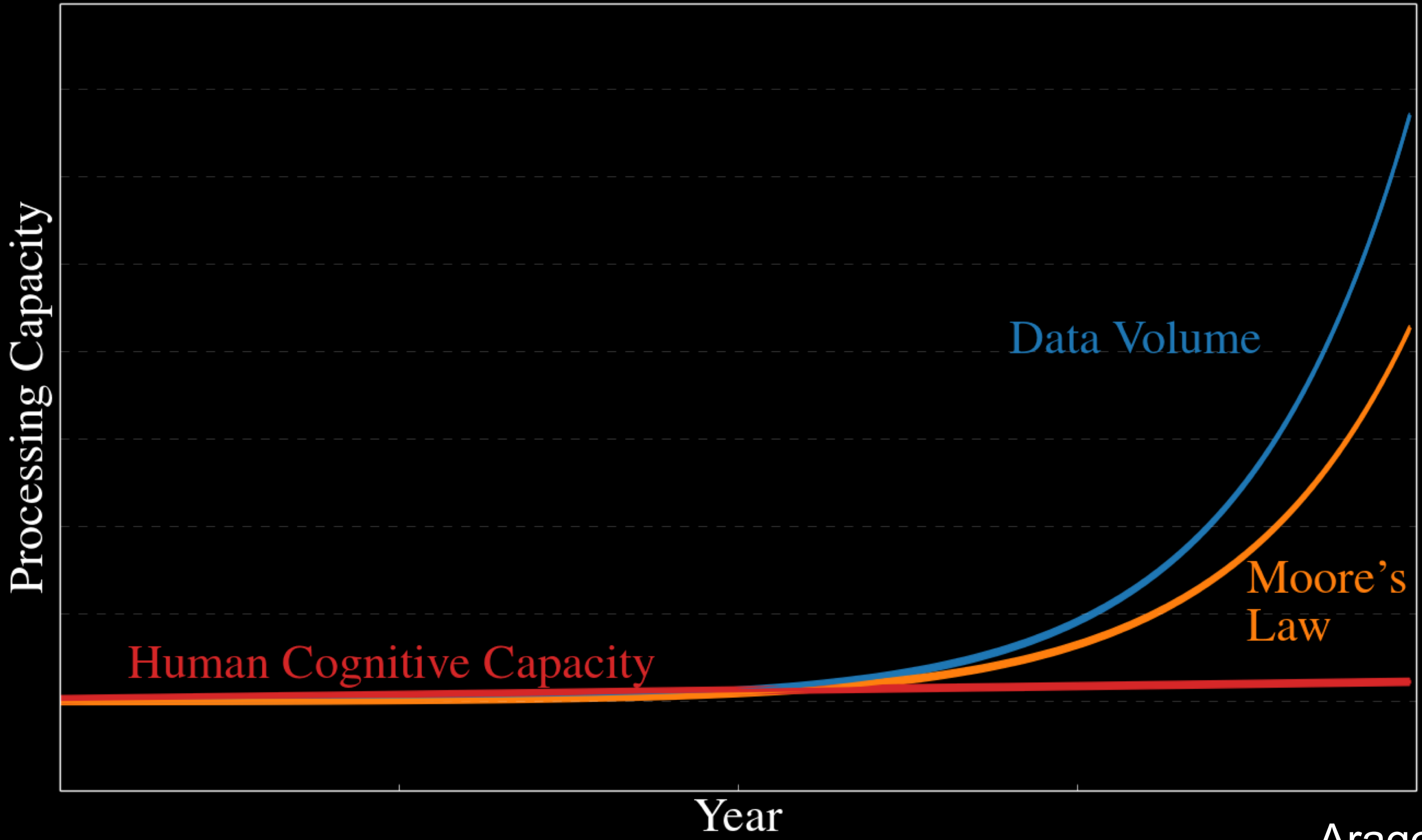

The Rise of the Machines…
Andrew Connolly
Director, DIRAC Institute
University of Washington
Johan Aurik

Why Machine Learning?



Aragon 2008

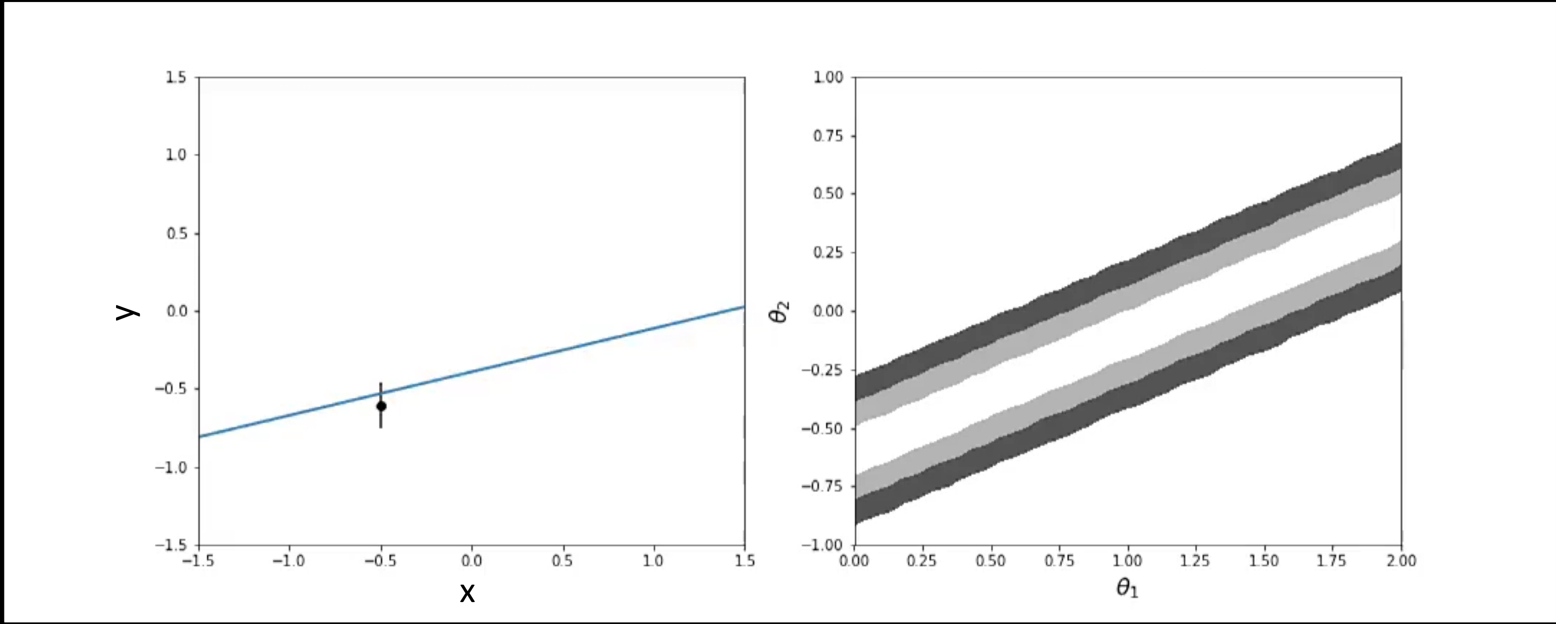
Why Machine Learning?



Aragon 2008

A somewhat singular opinion

$$y = \theta_1 x + \theta_2$$



Outline

- A brief history of the emergence of machine learning
- Impact of machine learning in astronomy
- What makes a technique successful
- The emergence of deep learning
- What next?

The emergence of ~~AI Machine Learning~~ AI

- <1980s: Artificial Intelligence
  - Development of ontologies
  - Focus on representing and reasoning and expert systems
- 1980s - 1990s: Machine Learning
  - Neural networks (back propagation)
  - Data rich problems and relaxed optimization algorithms
- 2000 – 2015: Machine Learning and Deep Learning
  - Easy access to ML libraries (GMM, SVM, Decision Trees)
  - Convolutional neural networks
- 2015 – Artificial intelligence

See Russell and Norvig

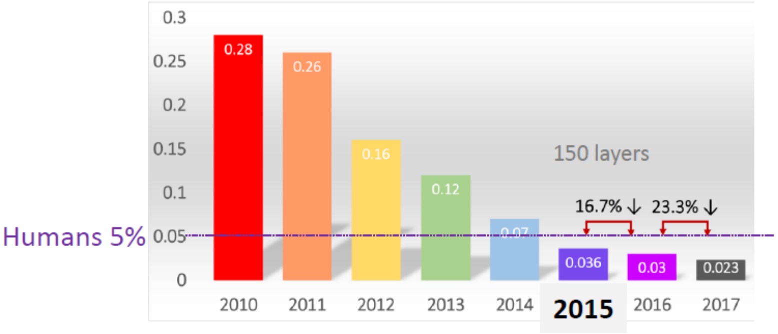
O. Russakovsky et al, arXiv:1409.0575; K. He, X. Zhang, S. Ren, J. Sunar, arXiv:1512.03385
WMW Jie Hu, Li Shen (Oxford), Gang Sun, 2017

Spectacular success

Image recognition challenge

ImageNet: 1000 categories, 1.2 million images

Classification error rate



Deep learning errors < humans

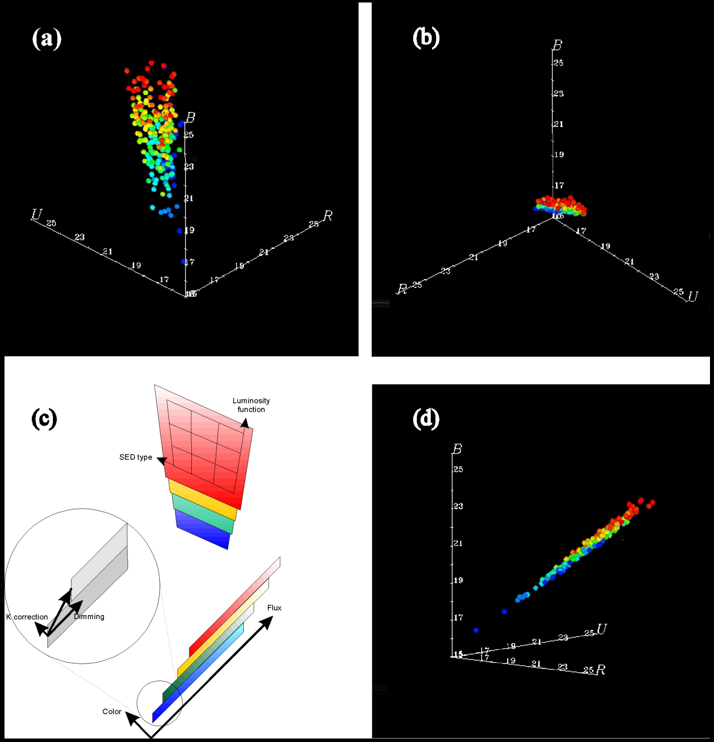
Three C’s of machine learning in Astronomy

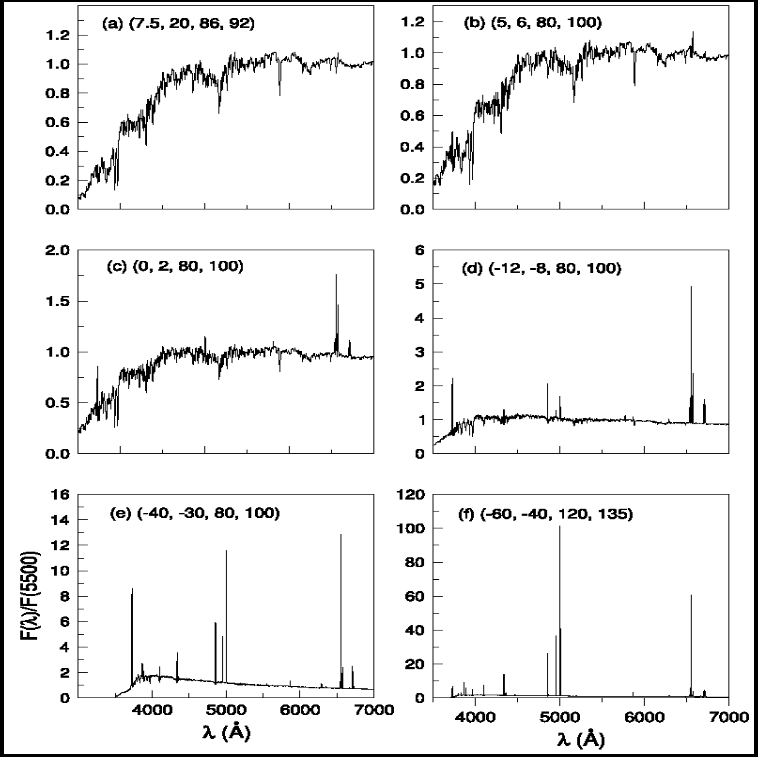
Compression

Classification

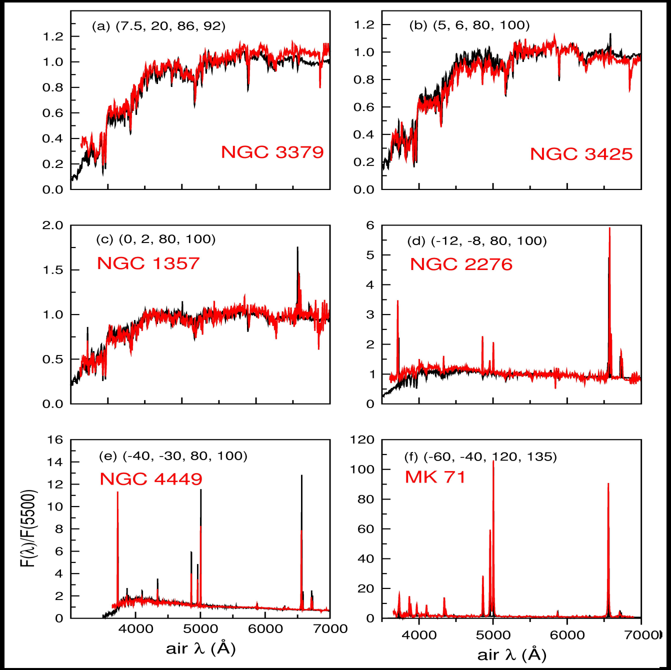
cSelection of features

Changing the representation

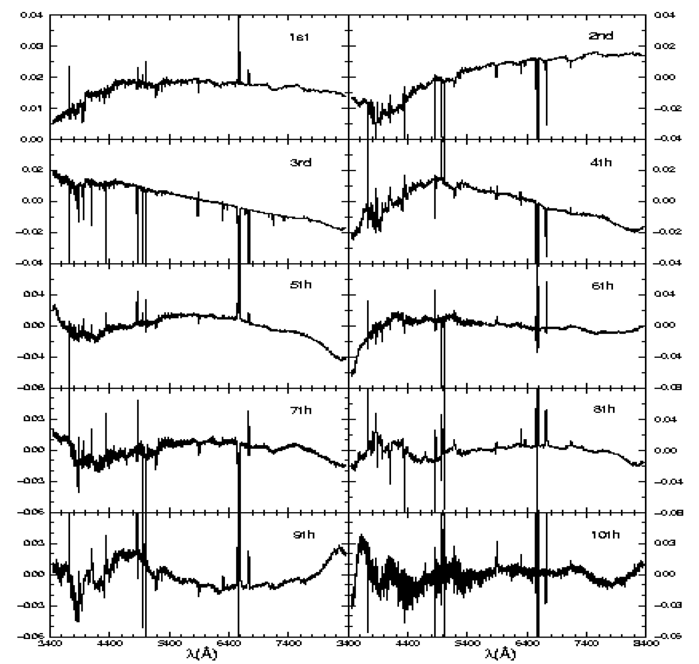




Compression or dimensionality reduction



$$f_{\lambda_k} = \sum_{i=1}^{M} a_i e_{i\lambda_k},$$



Orthogonal basis functions (PCA)

Yip et al 2003

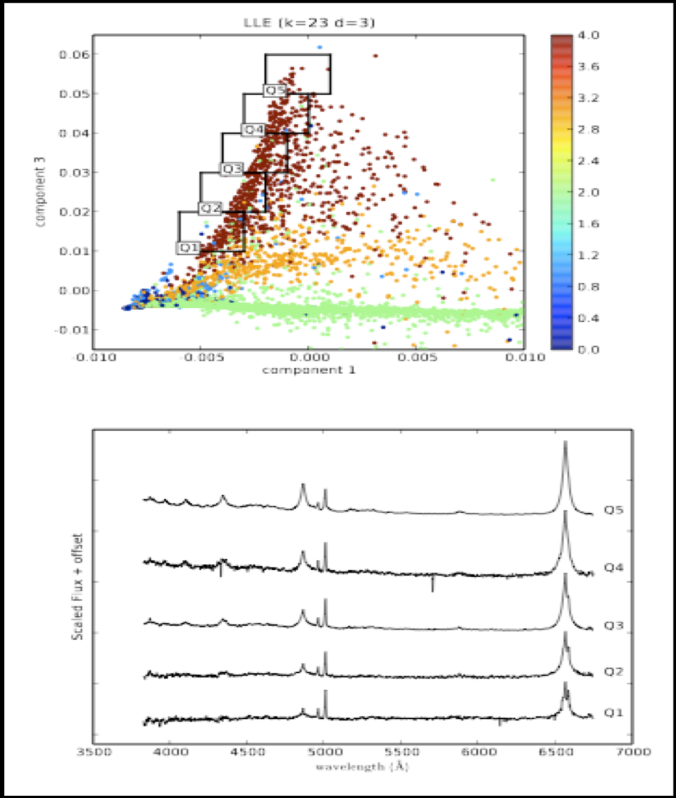
Complex compression: manifold learning

- Better compression (LLE, ICA, Diffusion Maps)
  - LLE: Identifies local weights. Projects onto a (defined) subspace preserving weights

$$\mathcal{E}_1(W) = \sum_{i=1}^{N} \left| \mathbf{x_i} - \sum_{j=1}^{N} W_{ij}\mathbf{x_j} \right|^2 .$$

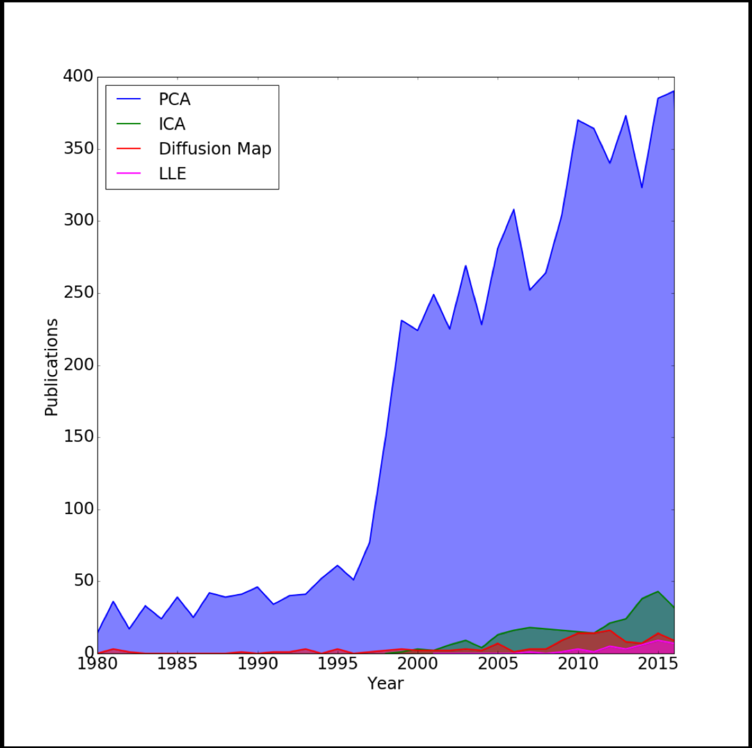
$$\mathcal{E}_2(Y) = |Y - WY|^2 ,$$

  - Diffusion maps: random walk on the data, walking to a nearby data-point is more likely than walking to another that is far away
  - More compact than PCA (15 vs 3 dimensions)



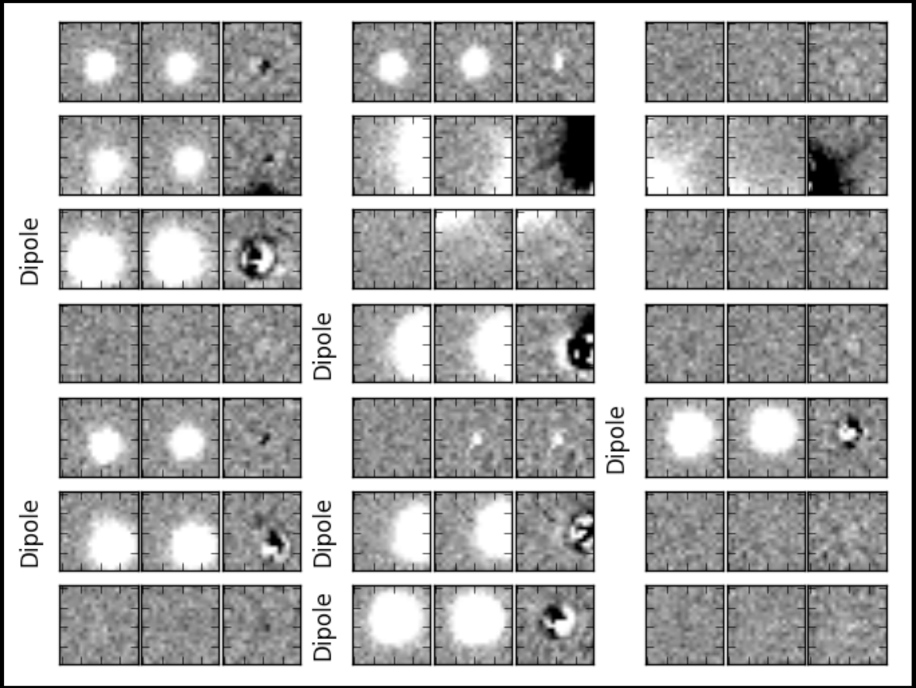
Vanderplas et al 2011

Why don't "better" techniques always "win"



- Interpretability
  - What drives the classification
- Extrapolation
  - Changing instrumentation/data
  - Basis functions vs archetypes
- Noise
  - Missing and incomplete data
- Speed?

Classification

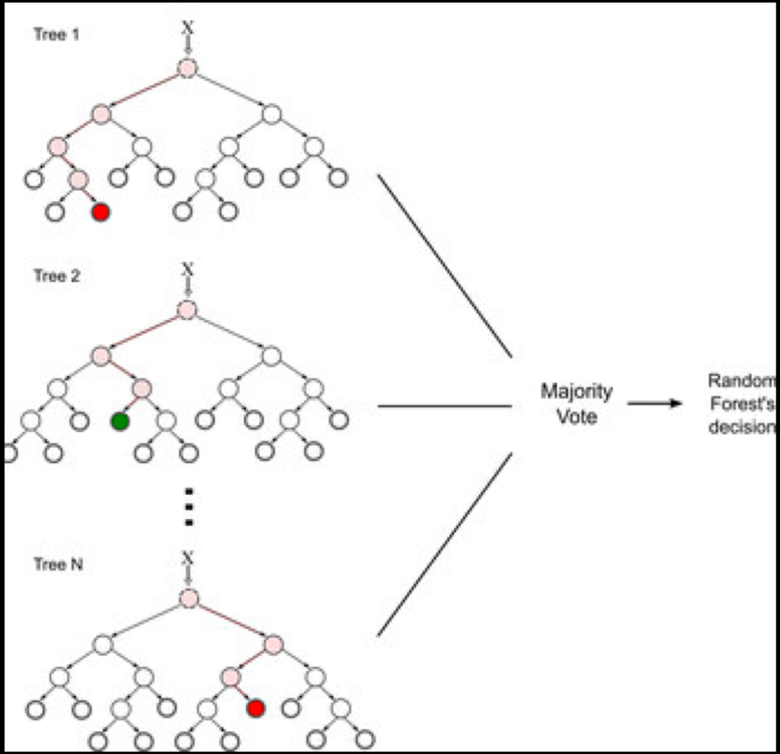


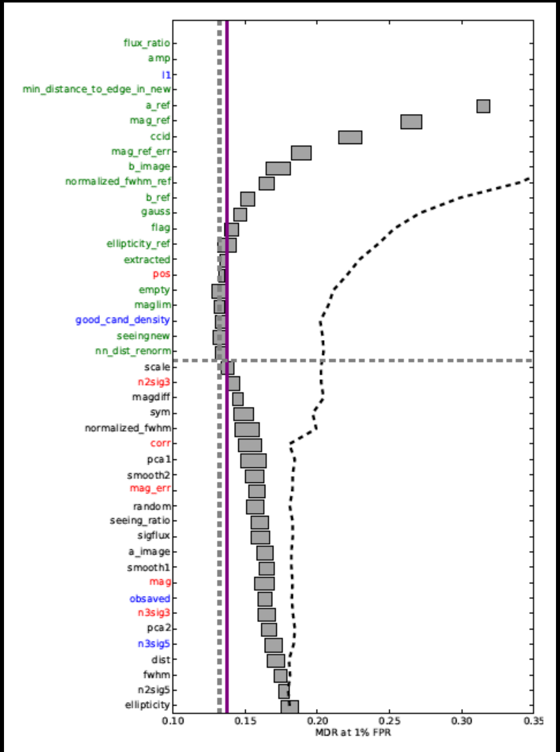
$$n(>\nu) = \frac{1}{2^{5/2}\pi^{3/2}}\nu e^{-\nu^2/2}$$

At 5σ we expect to find ~5 false positives in a 4K x 4K image (<<1%)

The number of false positives in previous surveys are 100:1 through to 10:1

Random Forests: dealing with high dimensional data

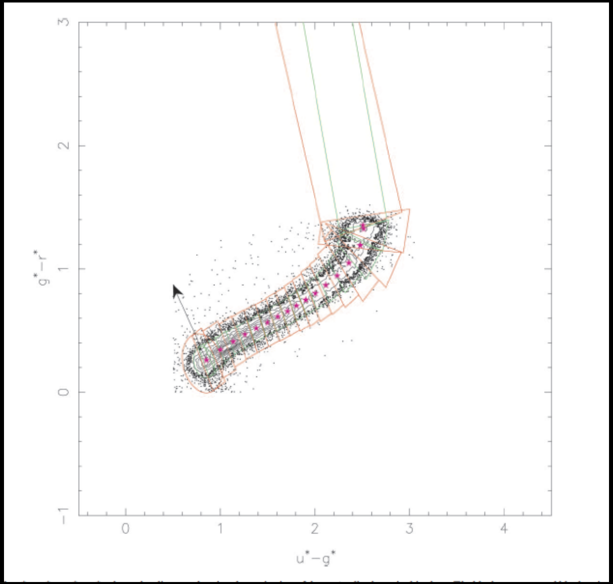




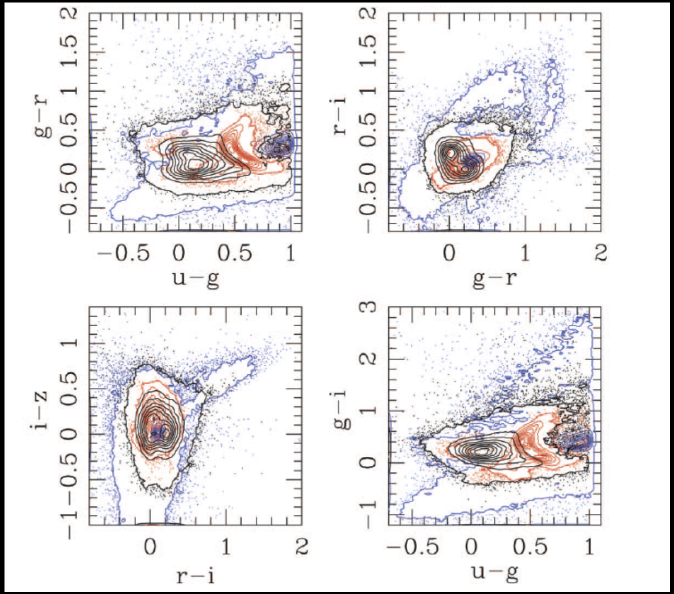
Richards et al

Random forests reduced the number of false positives from 100:1 to 2:1

Classification: supervised, semi-supervised

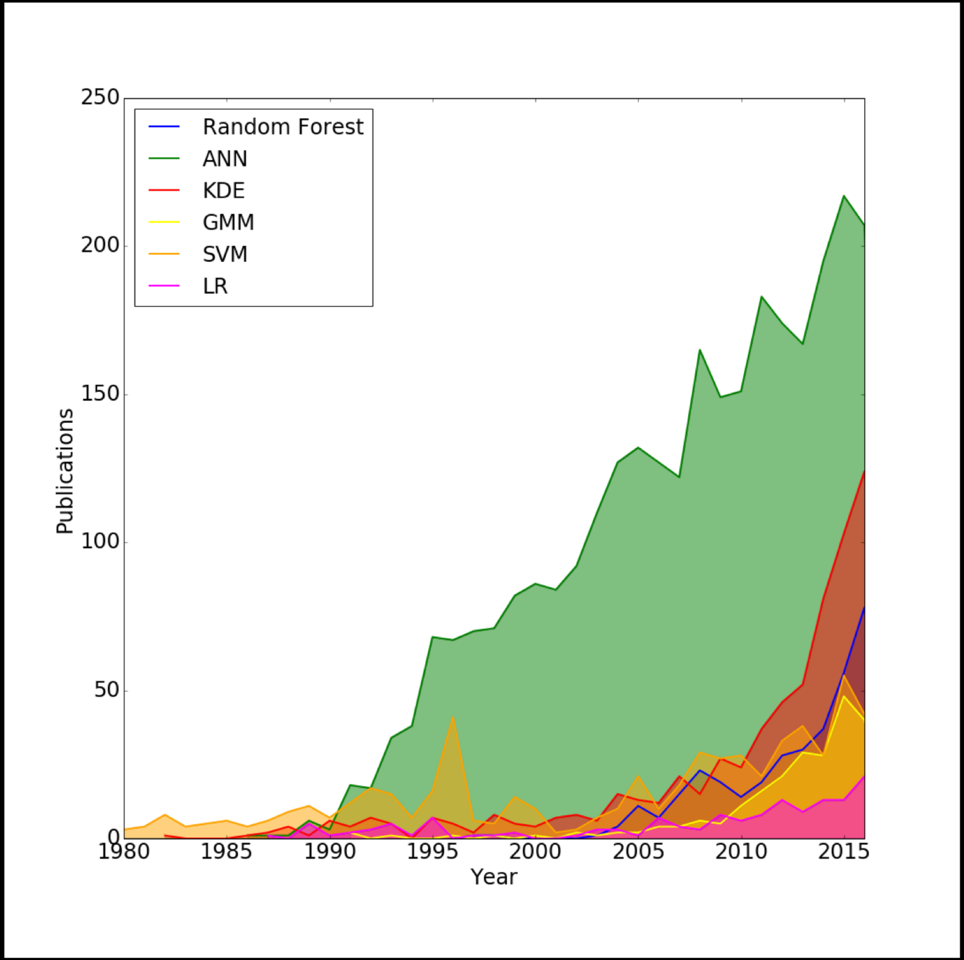


Newberg et al: 60% purity in the QSO samples

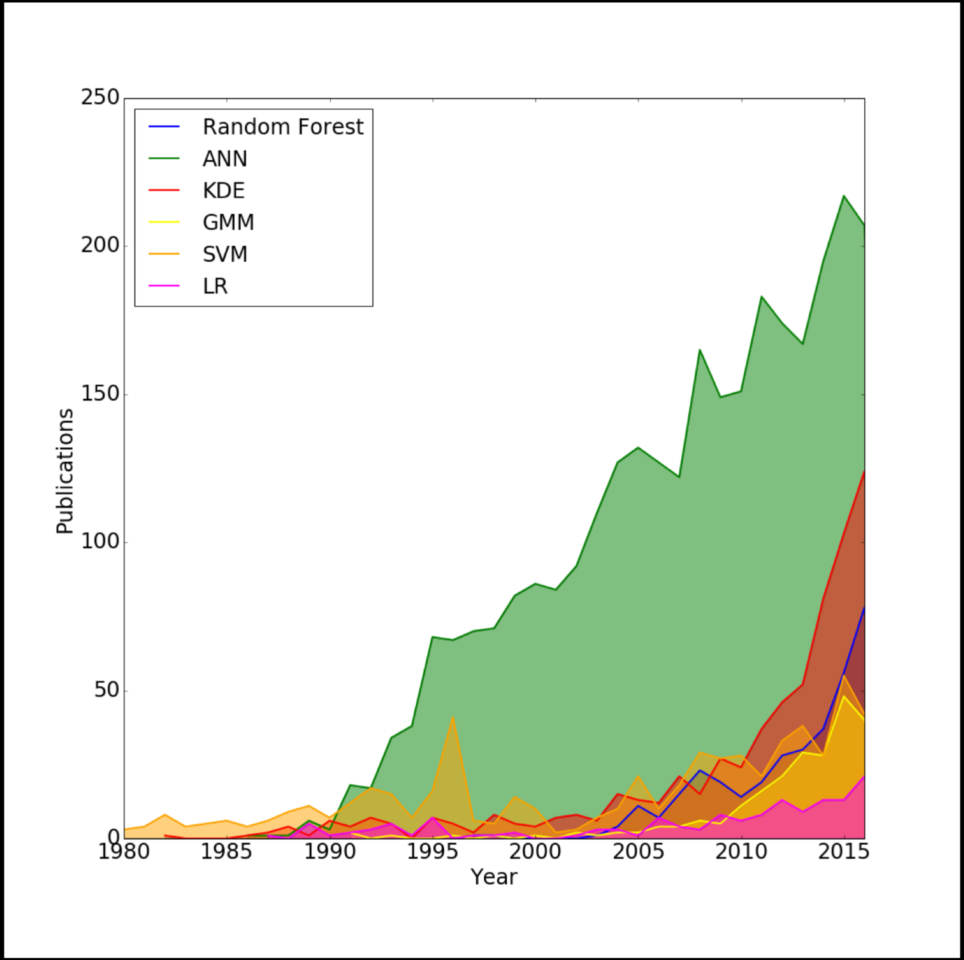


Richards et al: >90% purity in the QSO samples using a density estimation approach

$$P(C_1|x) = \frac{p(x|C_1)P(C_1)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)}.$$

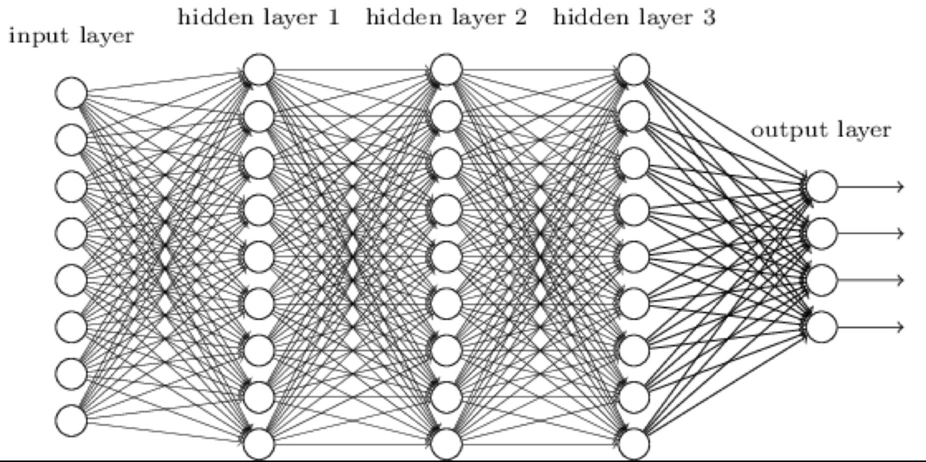

Random Forest
ANN
KDE
GMM
SVM
LR
Publications
Year
0
50
100
150
200
250
1980
1985
1990
1995
2000
2005
2010
2015

Experiments/data can drive adoption



cSelection of features: what is next?

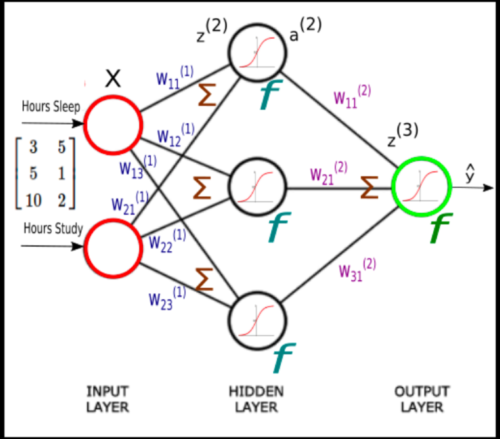
Deep Learning



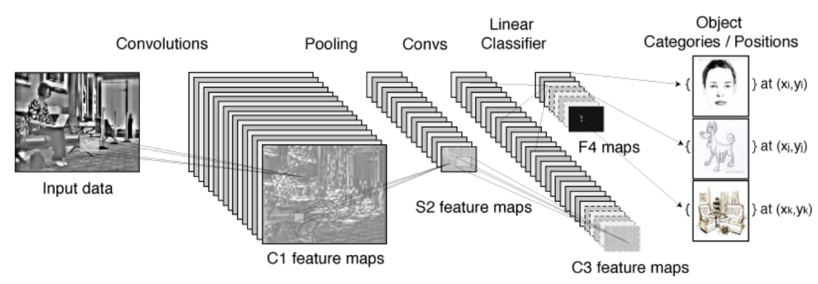
Deep neural network

http://neuralnetworksanddeeplearning.com/

$f(\Sigma X_i w_i + b_j)$



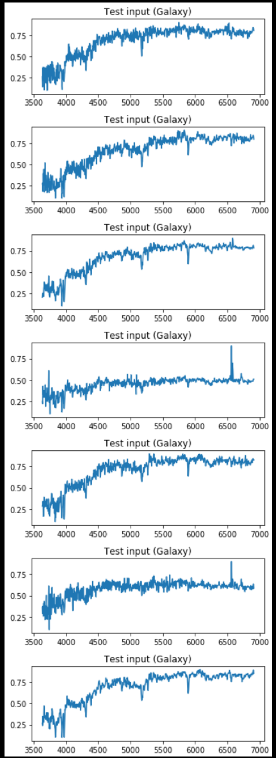
K Hong

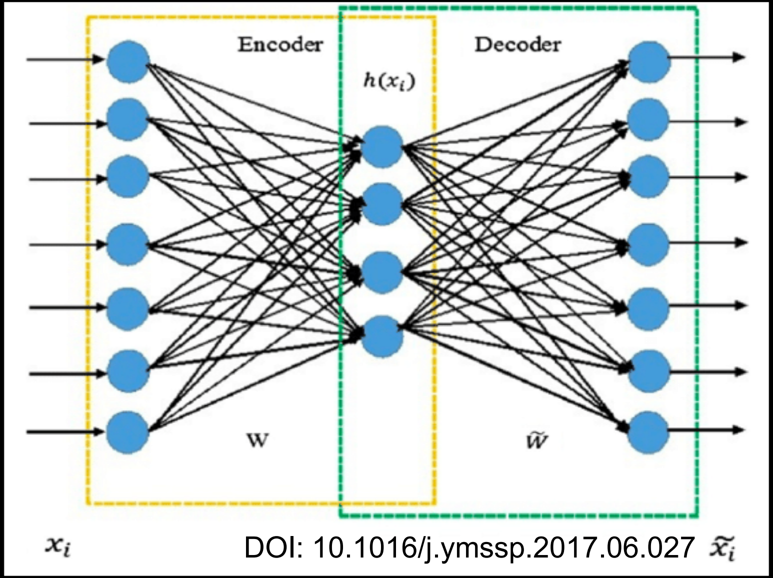


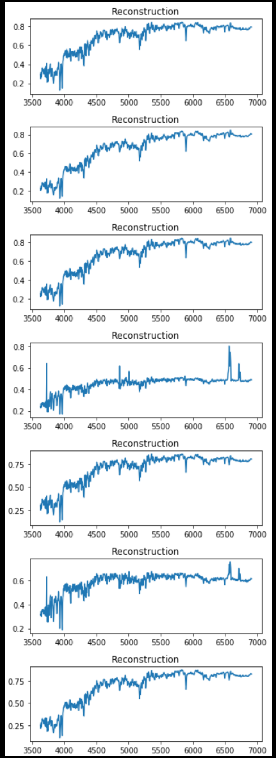
Convolution neural network

Torch's textbook

Autoencoding: non-linear dimensionality reduction



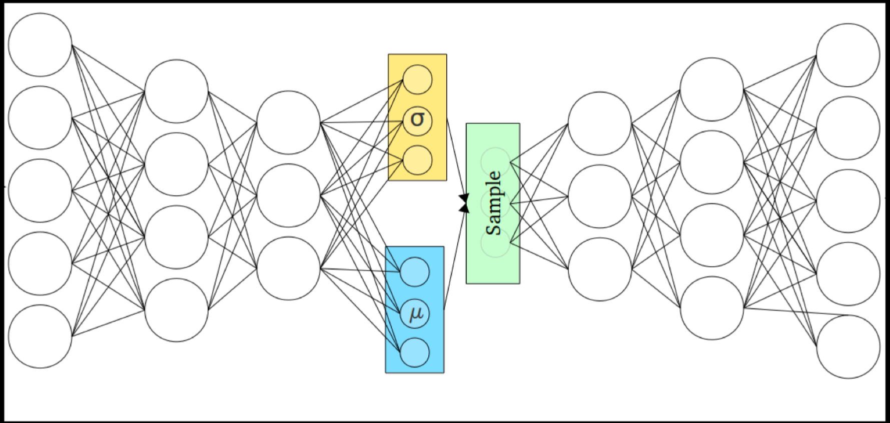




$$\mathcal{C} = -E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]$$

Denoising, Inpainting (interpolation), compression of high dimensional space

Variational Autoencoders



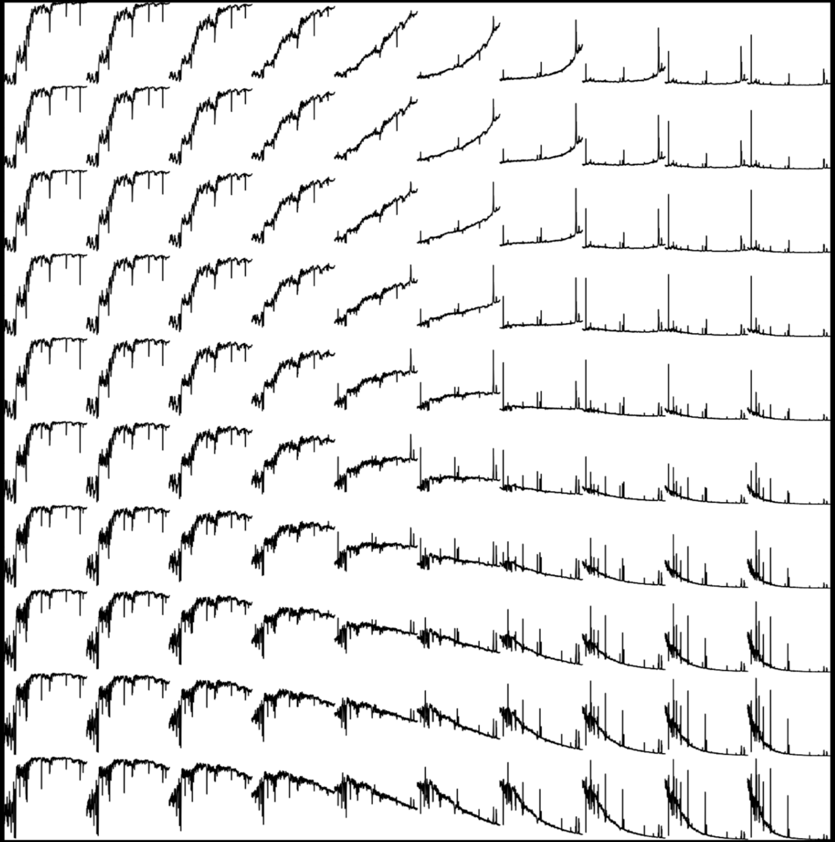
Irhum Shafkat Medium

$$\mathcal{C} = -E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] + \sum_{i=1}^{D} KL\Big(q_\phi(z_i|\mathbf{x}) \parallel p(z_i)\Big)$$

- Latent space is not always continuous or easily interpolatable. This makes it hard for generative models.
- Instead we map the input to a distribution (replace the bottleneck layer with mean and standard distribution)
- Vector for the decoded network is sampled from the distribution

Keep an eye out for disentangled VAE (forces neurons to be uncorrelated – reduces the number of activated neurons)

Encoding the spectra



4000 element spectrum to 2 components

VAE

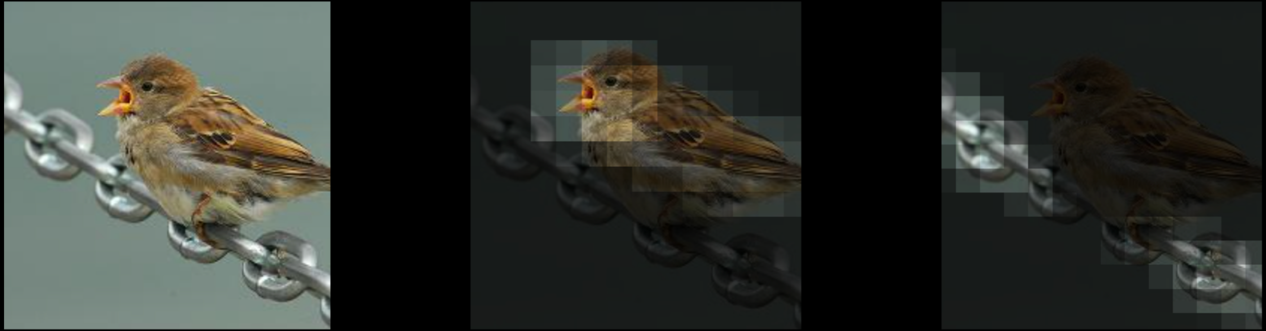
- Encoder: 2 layer (900 - 500)
- Decoder: 2 layer (500 - 900)
- Epoch: 2000

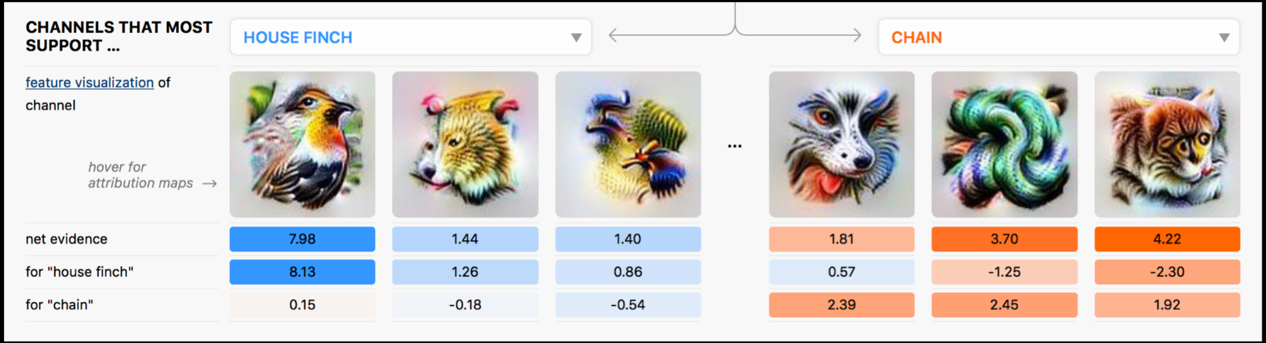
# latent space	VAE	PCA	NMF	AE
1	1.250 ± 0.022	1.626	2.259	1.435 ± 0.067
2	0.857 ± 0.028	0.866	0.999	0.916 ± 0.024
3	0.668 ± 0.021	0.761	0.795	0.936 ± 0.012
5	0.596 ± 0.028	0.658	0.675	0.871 ± 0.030

Vegara et al 2018

Challenges Ahead

- Interpretability: opening the black box





Pair each neuron activation with a visualization and sort them by size of the activation

https://distill.pub/2018/building-blocks/

Challenges Ahead

- Trust: believing the model
  - Ribeiro, Singh, Guestrin, “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”
- Understanding the information content
  - Tishby + Zaslavsky “Deep Learning and the Information Bottleneck Principle”
- Probabilistic modeling
  - TensorFlow Probability: very young and incomplete attempt to develop a generative model and quantify uncertainty
- Transfer learning: small sets of labels
  - Reusing a network or retraining a network with smaller data sets

It is more than Machine Learning

- What is the majority of the data intensive work spent on
  - 90% of the time is data wrangling
  - 10% is the analysis

- Spend time organizing your data and thinking about whether you might use it again. Reproducible science isn't just about "replayable science" it can help with improving your work (git, Jupyter, doc strings, documentation are your friend...)

You will still need to think