Feature Engineering for PHM Applications

From **Feature Engineering** to **Feature Learning**

Weizhong Yan, PhD, PE

Principal Scientist
Machine Learning Lab
GE Global Research Center

imagination at work

phmsociety

What is Feature Engineering?

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.
-Jason Brownlee, Machine Learning Mastery

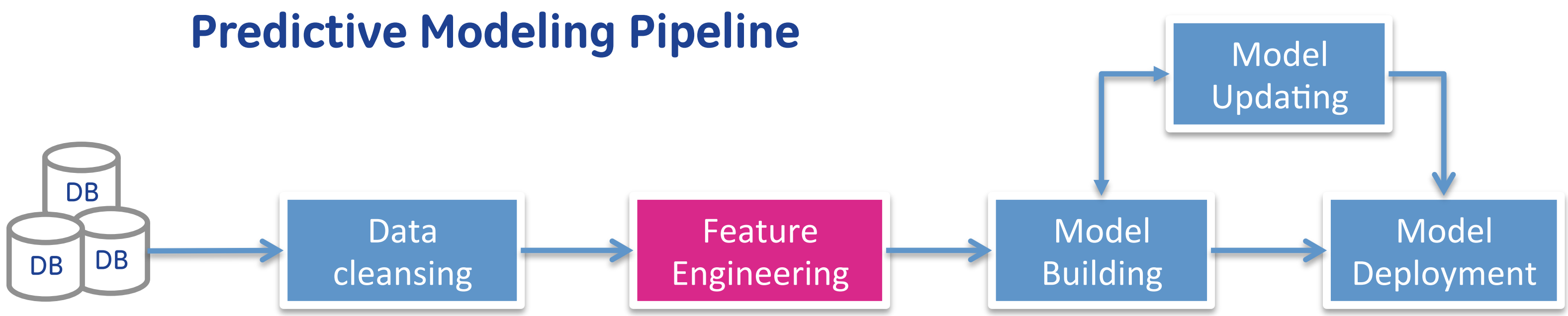
Feature engineering is manually designing what the input x's should be.
- Tomasz Malisiewicz, vision.ai Co-founder

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work better
- Wikipedia

Feature engineering is the act to inject knowledge into a machine learning model
- Anonymous

imagination at work

What is Feature Engineering?



The FE process includes:

- Remove unnecessary and/or redundant variables
- Modify variable data types, e.g., from categorical to numeric
- Combine some of existing variables
- Create new features
- Transform features
- …

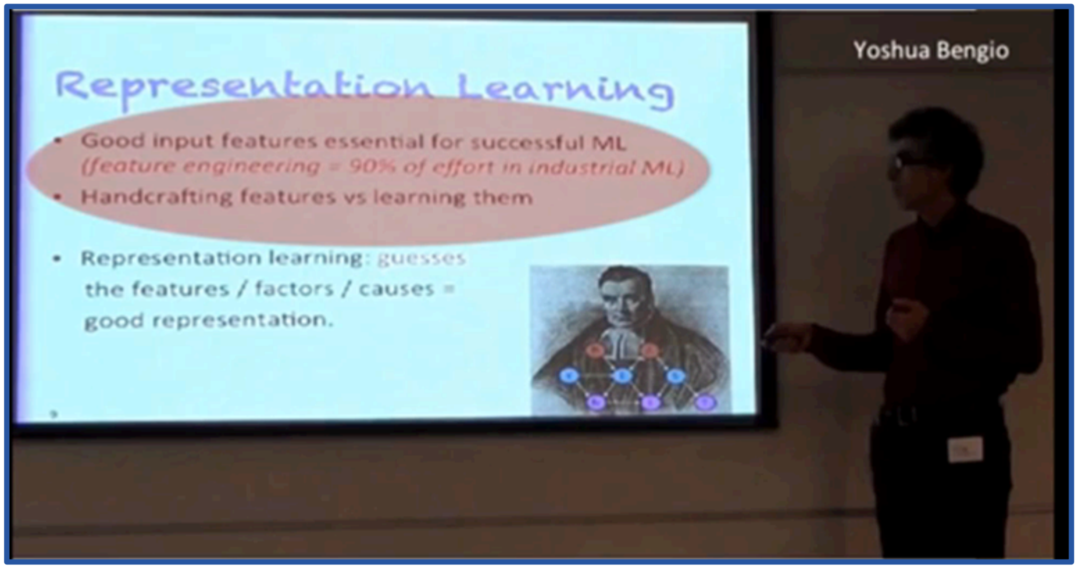
imagination at work

Feature engineering is important …

"Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering."
—Andrew Ng, Stanford University

"At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used."
- Pedro Domingos, University of Washington

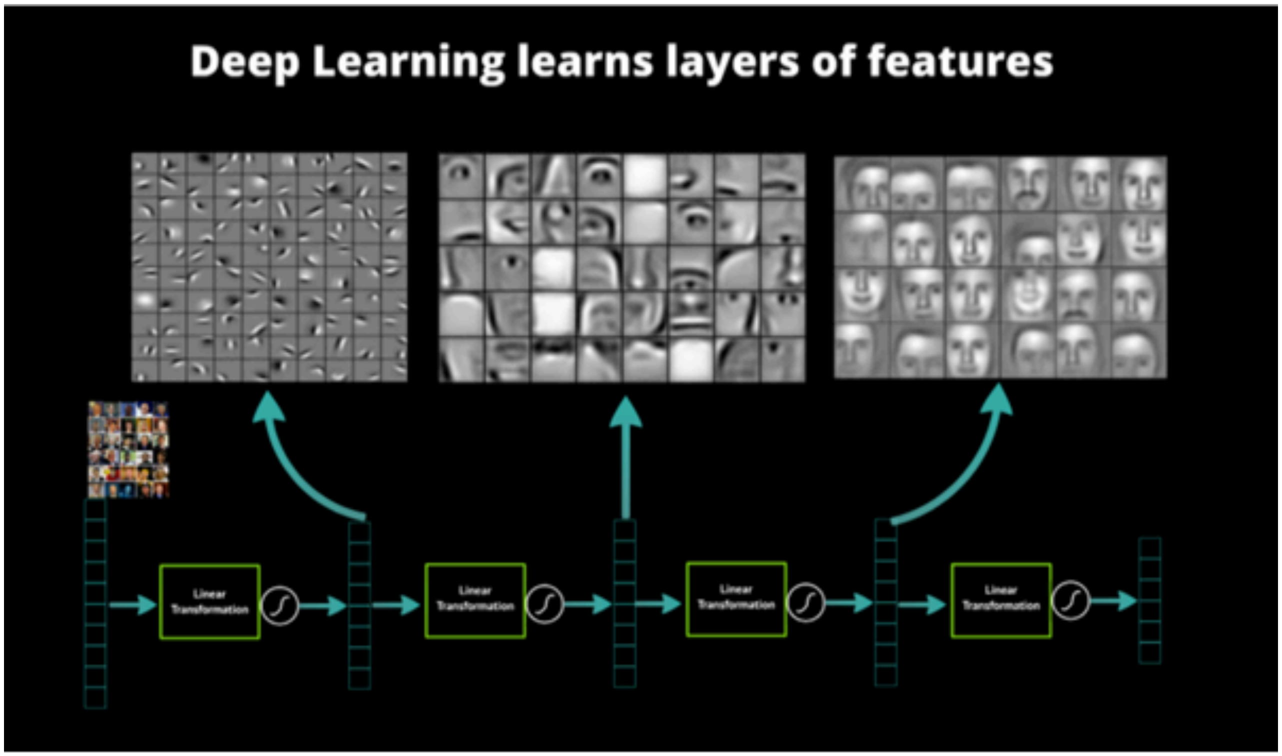
Feature engineering is hard and time-consuming …



"Good input features are essential for successful machine learning. Feature engineering ≈ **90%** of effort in industrial machine learning"

–Yoshua Bengio, University of Montreal

Feature learning alleviates some difficulties of feature engineering …



Source: dataRobot.com

… but finding a set of good features is still an unsolved problem

imagination at work

Outline

- Big picture
- Feature engineering
- (Shallow) Feature learning
- Deep feature learning

Big picture

Feature Engineering

Feature extraction

Feature dim. reduction
- Feature selection
- Feature low-dim projection

Many ways to categorize the methods

- **Knowledge based**
- Manual, labor intensive
- Domain/problem specific
- Not scalable

Feature Learning

Shallow feature learning

Supervised
- Multiple kernel learning
- Neural networks
- Transfer learning

Unsupervised
- Clustering
- Nonlinear embedding
- Matrix factorization
- SOM
- Genetic programming
- Sparse coding

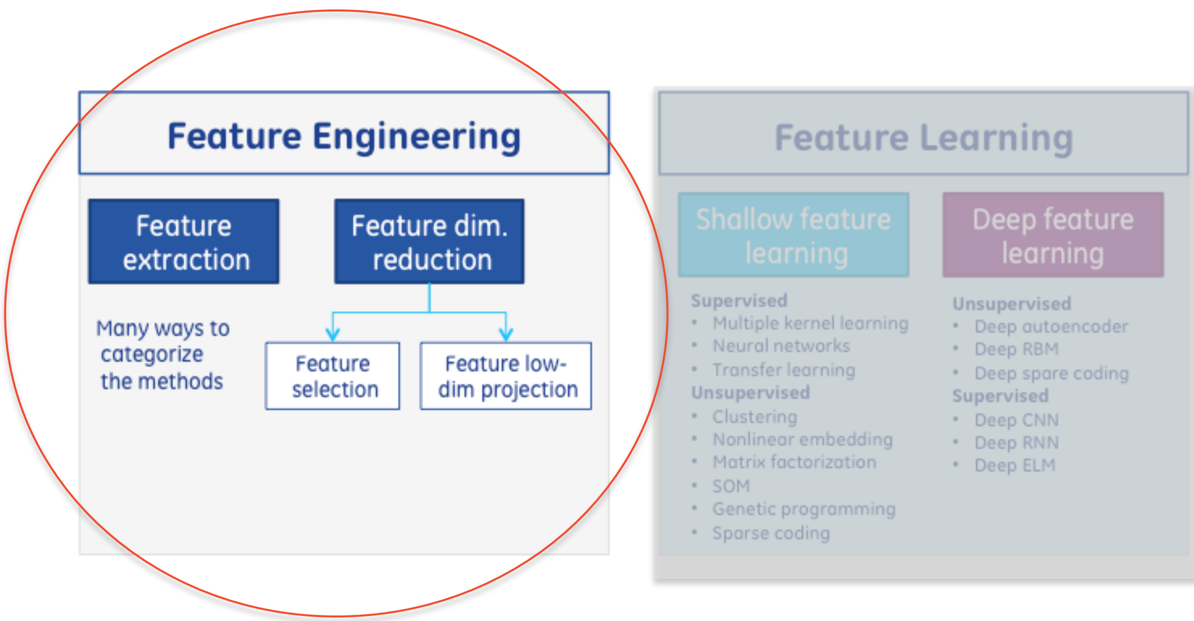
Deep feature learning

Unsupervised
- Deep autoencoder
- Deep RBM
- Deep spare coding

Supervised
- Deep CNN
- Deep RNN
- Deep ELM

✓ **Data driven**
✓ Automated
✓ Generic
✓ Scalable

Feature Engineering (FE) (knowledge based)



Characteristics of FE

- Manual, ad hoc
- Time-consuming
- Domain/application specific (as supposed to data specific in feature learning)
- Not optimal
- Not scalable

Domain specific: features in one domain do not generalize to other domains

Domains:
- PHM
- Computer vision
- Speech recognition
- Text analytics
- Business analytics
- ....

→

PHM applications:
- Vibration analysis
- SHM
- Turbine machines
- Electrical systems
- Electronic devices
- Batteries
- ....

→

Vibration analysis
- Bearings
- Gears
- ....

imagination at work

FE - Feature extraction

Different Technologies

- Statistical analysis
- Signal processing
- Image processing
- Time-series analysis
- Control theory
- Information theory

Different PHM applications

- Vibration analysis
- Turbine machines
- Electrical systems
- Electronic devices
- Batteries
- SHM

Different data types

- Continuous
- Categorical
- Binary
- …

Time dependency

- Time independent (stationary)
- Time dependent (non-stationary)

Univariate vs. multivariate

Different data sampling rate

…

Example: Feature extraction for vibration analysis

Stationary signals

Time domain

- <u>Statistical-based</u>
  - RMS
  - Variation
  - Skewness
  - Kurtosis
  - crest factor
- <u>Model-based</u>
  - AR model
  - HMM model
- <u>Signal processing</u>
  - TSA
  - Correlation
  - Convolution
  - Fractal analysis
  - Correlation dimension

Frequency domain

- Spectral analysis
- Envelope analysis
- Cepstrum analysis
- Higher order spectrum

Non-stationary signals

Time-frequency

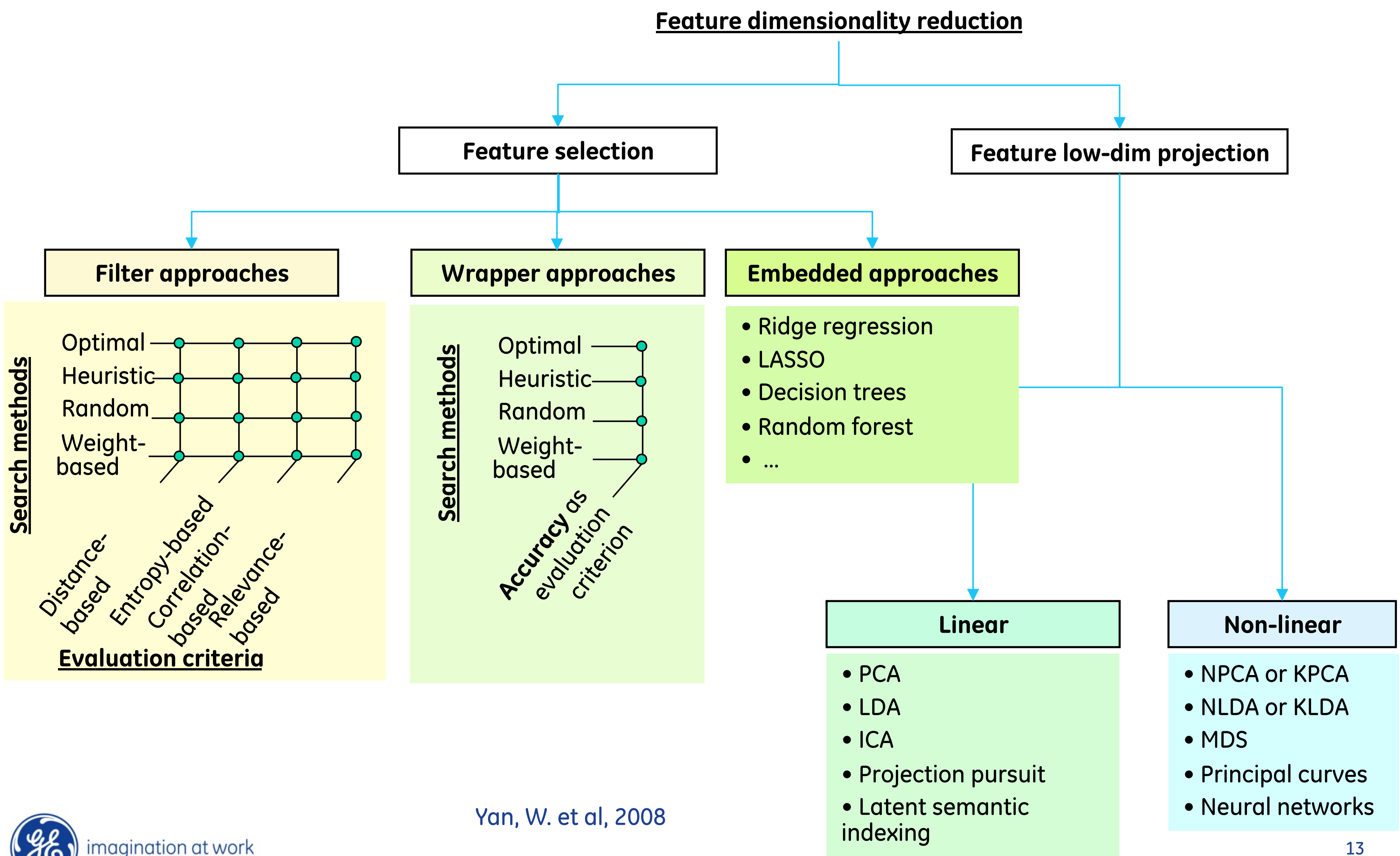
- Short-time Fourier Transform (STFT)
- Wigner-Ville distribution (WVD)
- Empirical mode decomposition (EMD)
- Basis pursuit
- Spectral kurtosis
- Cyclostationary analysis

Wavelets

- Continuous wavelet transform (CWT)
- Discrete wavelet transform (DWT)
- Wavelet packet transform
- Morlet wavelet
- Hilbert-Huang transform

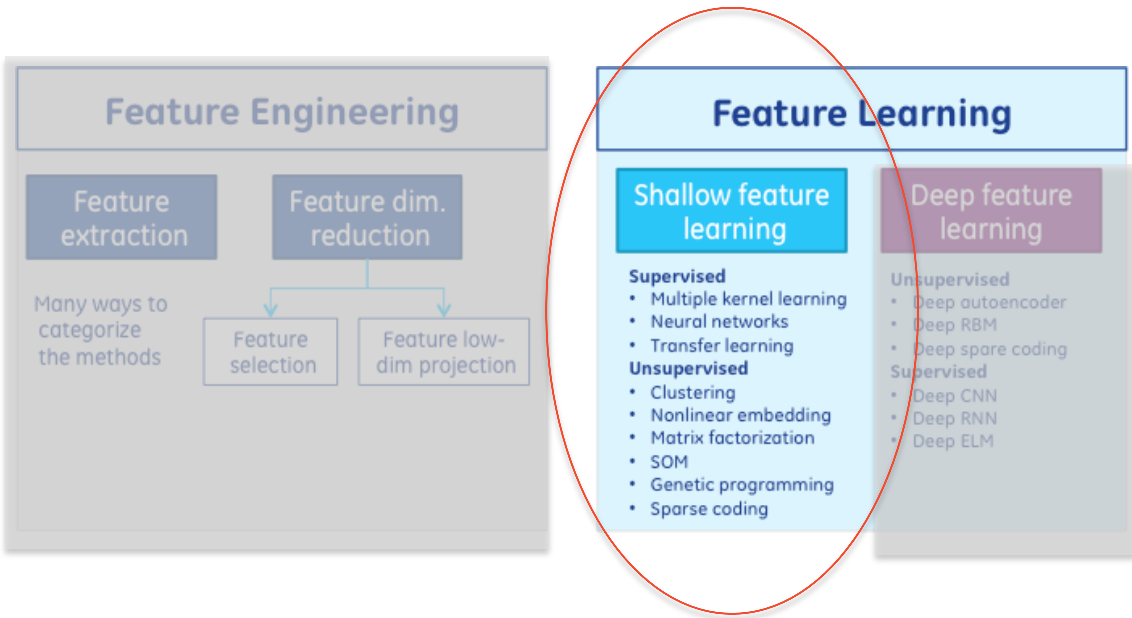
Yan, W. et al, 2008

FE - Feature dim. reduction



Yan, W. et al, 2008

(Shallow) Feature Learning (FL) (data driven)

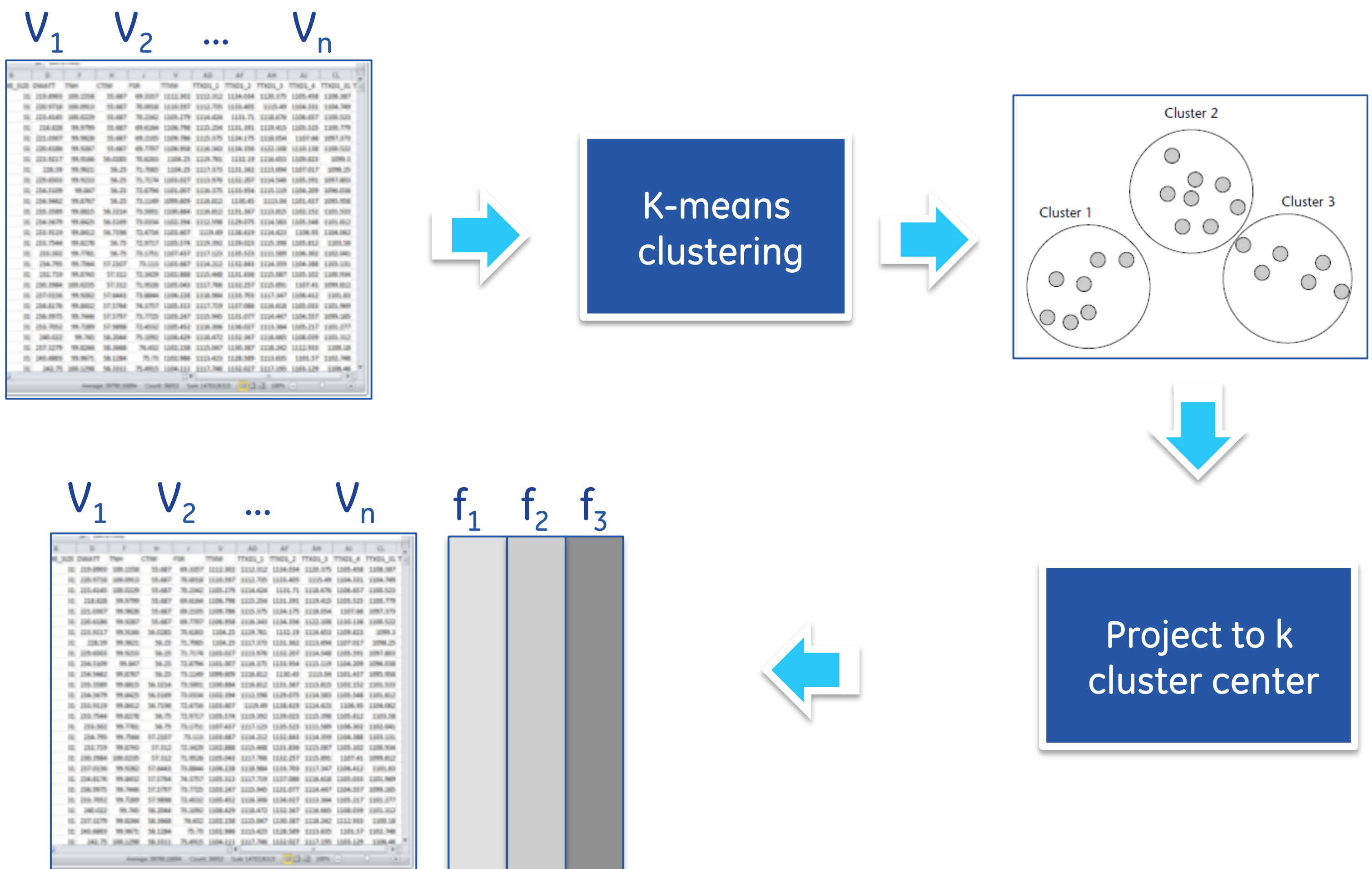


Shallow feature learning

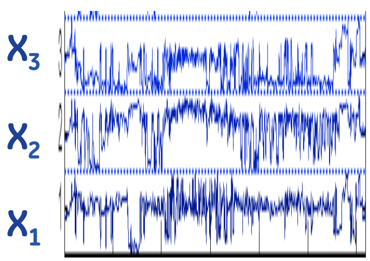
Including many unsupervised learning, manifold learning, and low-dim projection algorithms

- Clustering, e.g., k-means, GMM
- Matrix factorization, e.g., PCA, ICA, NMF, sparse coding
- Nonlinear embedding, e.g., isomap, LLE, Laplacian eigenmaps, etc., – manifold learning
- Neural networks, e.g., SOM, autoencoder
- Genetic programming
- Sparse coding / dictionary learning
- …

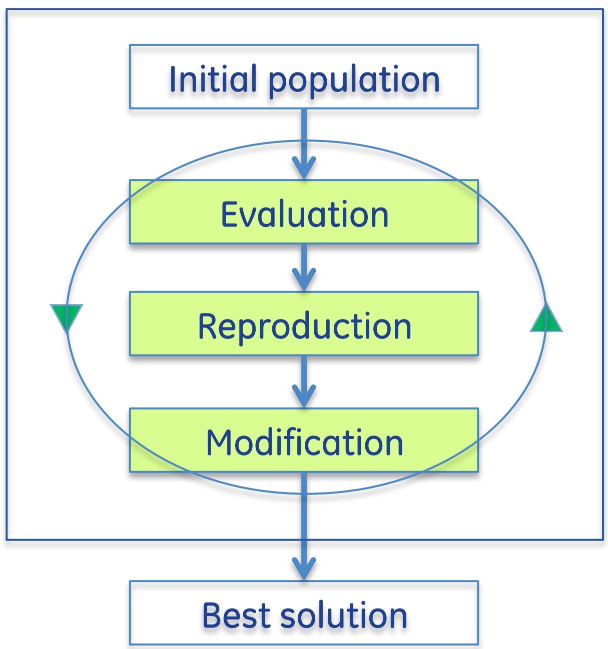
Shallow feature learning
- k-means clustering

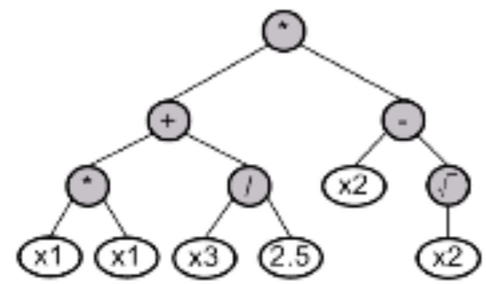


Shallow feature learning
- genetic programming (GP)



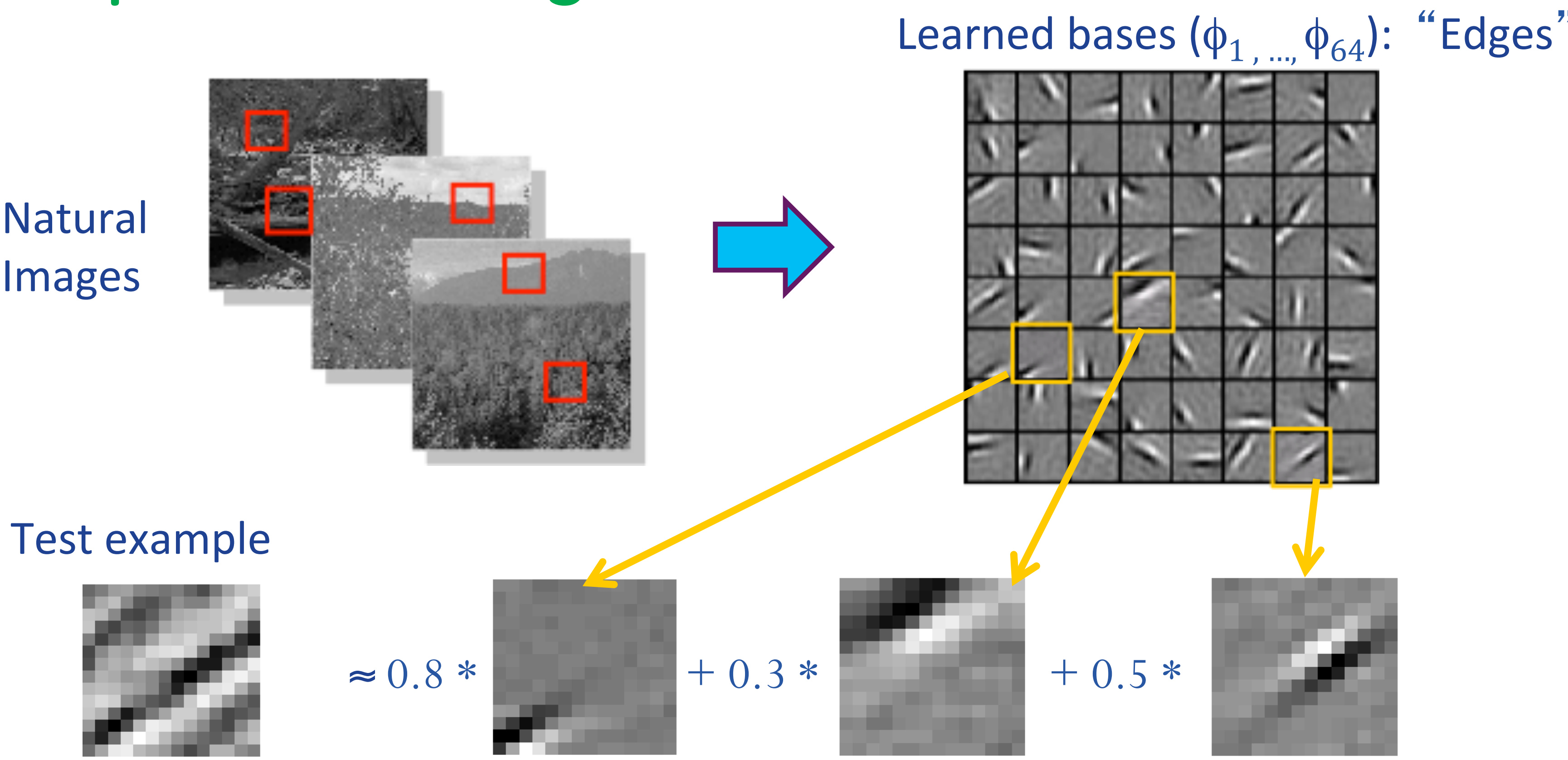
GP algorithm





$$F = (x_1^2 + (\frac{x_3}{2.5})) * (x_2 - \sqrt{x_2})$$

Shallow feature learning
- sparse coding



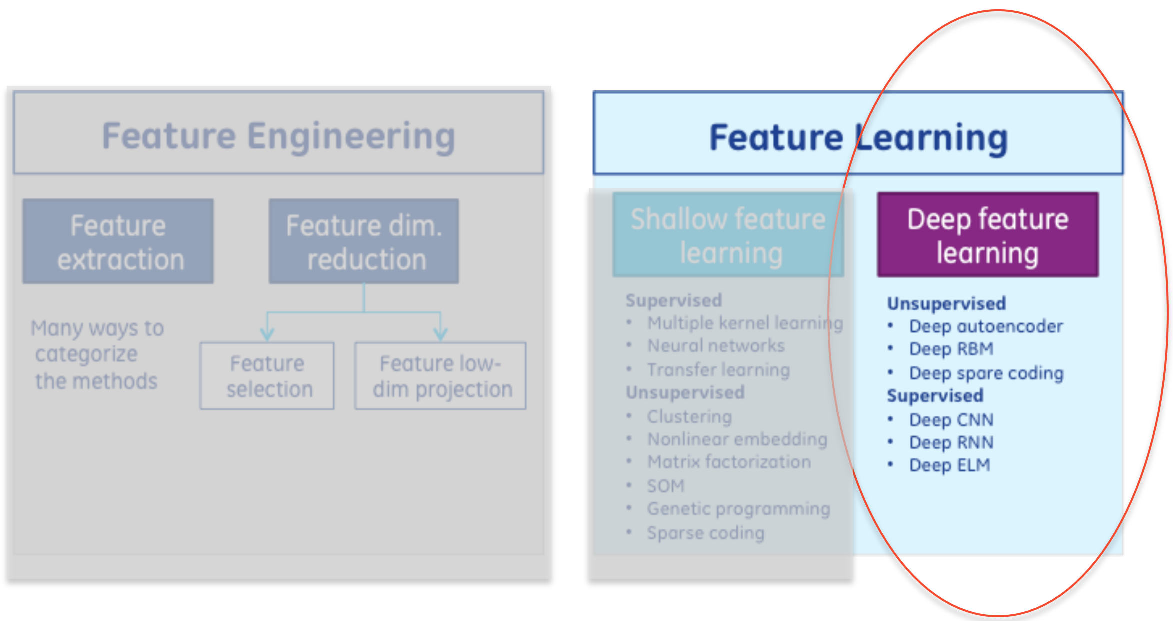
$$x \approx 0.8 * \phi_{36} + 0.3 * \phi_{42} + 0.5 * \phi_{63}$$

[0, 0, …, 0, **0.8**, 0, …, 0, **0.3**, 0, …, 0, **0.5**, …] feature representation

imagination at work

Lee, et al, NIPS 2007

Deep Feature Learning (FL) (data driven)



imagination at work

What is Deep Learning?

Deep learning is a part of broader family of machine learning methods that involve learning multiple levels of representations of data

Deep learning $\approx$ representation learning

All deep learning is representation learning, but

Not all representation learning is deep learning

Deep learning $\neq$ unsupervised learning

Not all unsupervised learning is deep learning

Not all deep learning is unsupervised learning

Deep learning in the news

MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

Archive > Volume 521 > Issue 7553 > Insights > Reviews > Article

Take part in Nature Publishing Group's annual reader survey here for the chance to win a Macbook Air.

NATURE | INSIGHT | REVIEW

Deep learning

Yann LeCun, Yoshua Bengio & Geoffrey Hinton

Affiliations | Corresponding author

Nature **521**, 436–444 (28 May 2015) | doi:10.1038/nature14539
Received 25 February 2015 | Accepted 01 May 2015 | Published online 27 May 2015

Deep learning in the news

☰ SECTIONS HOME SEARCH The New York Times

SCIENCE

Scientists See Promise in Deep-Learning Programs

By JOHN MARKOFF NOV. 23, 2012

TECH 2/19/2015 @ 1:06PM | 6,601 views

Microsoft's Deep Learning Project Outperforms Humans In Image Recognition

Big Data

IBM acquires AlchemyAPI to bring deep learning to Watson

Published On: Sun, Jun 21st, 2015 Technology / Technology & StartUps / World | By **News Brief**

Facebook's Newest Deep Learning System Makes Images That Humans Think Are Real 40% Of The Time

Pin it

June 16, 2015

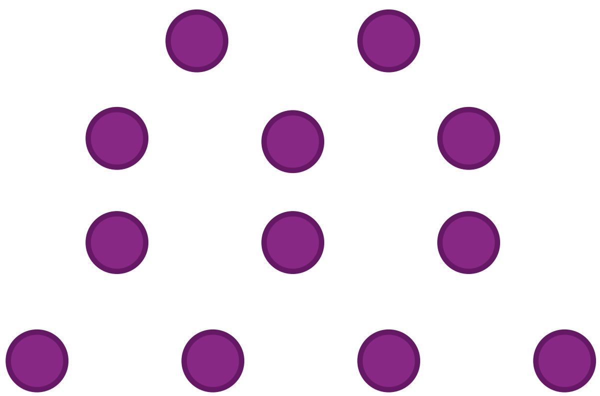
Deep Learning Machine Beats Humans in IQ Test and performs between bachelor and masters degree level

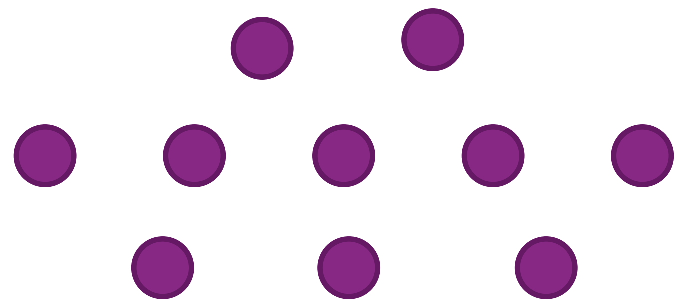
artificial intelligence, china, deep learning, future, intelligence, pre-singularity, science, singularity

TECH 3/24/2015 @ 10:14AM | 8,105 views

NVIDIA GTC: NVIDIA Bets Big On Deep Learning

Deep vs. shallow neural networks





Two-layer (plus input layer) neural networks are an universal approximator

Why deep?

Given the same number of non-linear (neural network) units, a deep architecture is more expressive than a shallow one (Bishop 1995)

Some functions compactly represented with k layers may require exponential size with 2 layers

… However, deep networks have challenges

- Needs labeled data (most data is not labeled)
- Scalability – does not scale well over multiple layers
  - Very slow to converge
  - “Vanishing gradients problem” : errors shrink exponentially with the number of layers
- For more: “Understanding the Difficulty of Training Deep Feed Forward Neural Networks”: http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS2010_GlorotB10.pdf

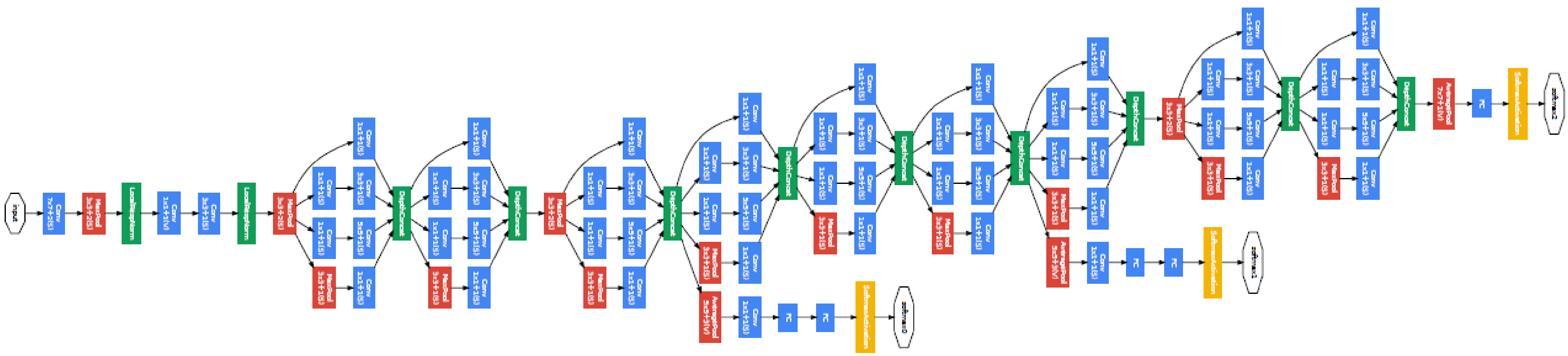
The deep breakthroughs

- Hinton, et al, 2006, "Reducing the dimensionality of data with neural networks", Science, 2006
- Bengio, et al, 2006 "Greedy layer-wise training of deep networks", NIPS 2006
- LeCun, et al, 2006, "Efficient learning of sparse representation with an energy based model", NIPS 2006
    - Stacked RBMs or AE
    - Layer-wise training with unlabeled data (unsupervised learning)
    - Fine tuning with labeled data

Going deep

googleNet (2014 imageNet competition)

of layers = 27
Overall # of layers (independent building blocks) = 100
Total # of tunable parameters = 5MM+



Source: "Going deeper with convolutions", Szegedy, et al., CVPR 2015

Going deeper and deeper…

- 11.2 billion parameters by Google
- 15 billion parameters by Lawrence Livermore National Lab
- 160 billion parameters by Digital Reasoning
- ???

imagination at work

Deep learning has achieved state-of-the-art performance in different areas

Speech recognition

deep learning results

task	hours of training data	DNN-HMM	GMM-HMM with same data
Switchboard (test set 1)	309	18.5	27.4
Switchboard (test set 2)	309	16.1	23.6
English Broadcast News	50	17.5	18.8
Bing Voice Search (Sentence error rates)	24	30.4	36.2
Google Voice Input	5,870	12.3	
Youtube	1,400	47.6	52.3

ImageNet competition

Rank	Name	Error rate	Description
1	**U. Toronto**	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

Deep learning won all competitions

1. IJCNN Traffic Sign Recognition Competition, 2011
2. ISBI Brain Image Segmentation Contest, 2012
3. ICDAR Chinese hand-writing recognition, 2011
4. MICCAI Mitosis detection grand challenge, 2013

imagination at work

Deep learning applications (products)

- IBM Watson
- Google self-driving cars
- Google Glasses
- Facebook Face recognition
- Facebook user modeling
- Microsoft natural language processing
- Apple Siri

Deep learning has not been used for PHM applications

imagination at work

Unsupervised vs. supervised

- **Unsupervised**
  - Deep auto-encoder and its variants (AE, DAE, SAE)
  - Deep Restricted Boltzmann machines (RBM)
  - Deep sparse coding (DSC)

Explicit feature learning

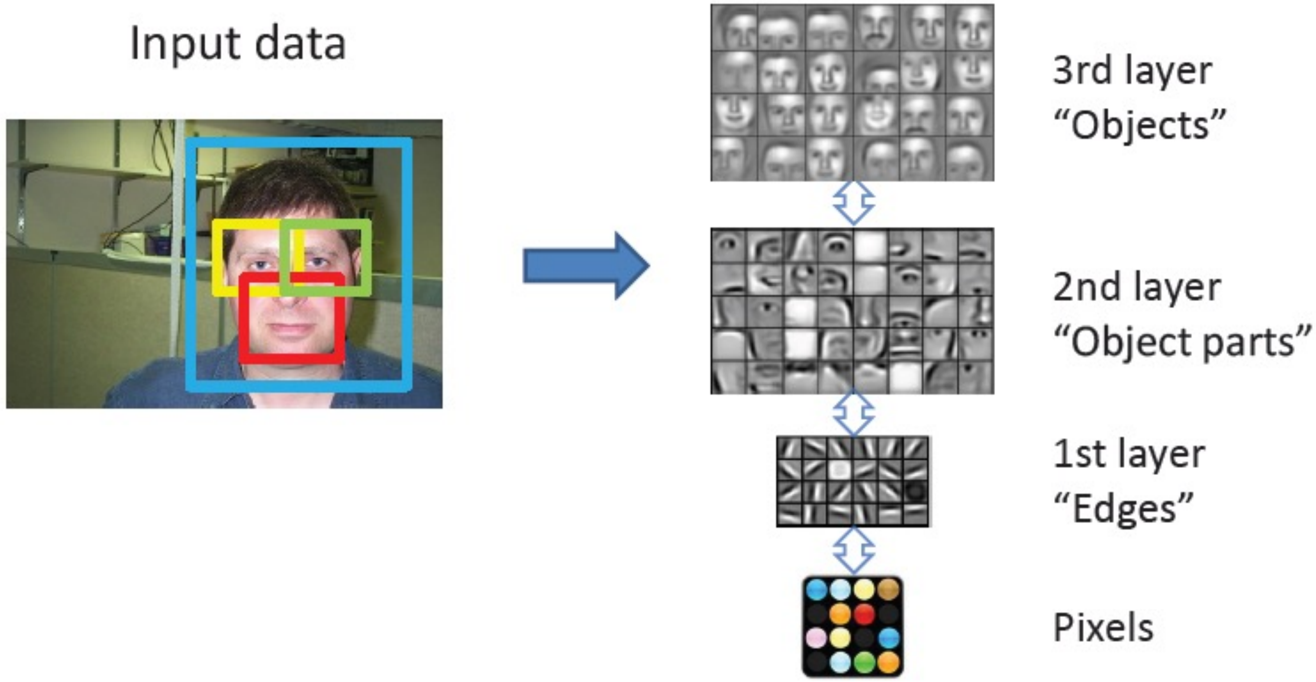
- **Supervised**
  - Convolutional neural networks (CNN)
  - Deep recurrent neural networks (RNN)
  - Deep extreme learning machines (ELM)

Implicit feature learning

Hybrid: Unsupervised pre-training + supervised fine tuning

Unsupervised deep feature learning is interesting and useful…

In most real-world applications, PHM included, labeled data is sparse (difficult to obtain), while unlabeled data is abundantly available



H. Lee (2010)

Unsupervised feature learning did well

Audio

TIMIT Phone classification	Accuracy
Prior art (Clarkson et al.,1999)	79.6%
Stanford Feature learning	**80.3%**

TIMIT Speaker identification	Accuracy
Prior art (Reynolds, 1995)	99.7%
Stanford Feature learning	**100.0%**

Images

CIFAR Object classification	Accuracy
Prior art (Krizhevsky, 2010)	78.9%
Stanford Feature learning	**81.5%**

NORB Object classification	Accuracy
Prior art (Ranzato et al., 2009)	94.4%
Stanford Feature learning	**97.3%**

Video

Hollywood2 Classification	Accuracy
Prior art (Laptev et al., 2004)	48%
Stanford Feature learning	**53%**

YouTube	Accuracy
Prior art (Liu et al., 2009)	71.2%
Stanford Feature learning	**75.8%**

KTH	Accuracy
Prior art (Wang et al., 2010)	92.1%
Stanford Feature learning	**93.9%**

UCF	Accuracy
Prior art (Wang et al., 2010)	85.6%
Stanford Feature learning	**86.5%**

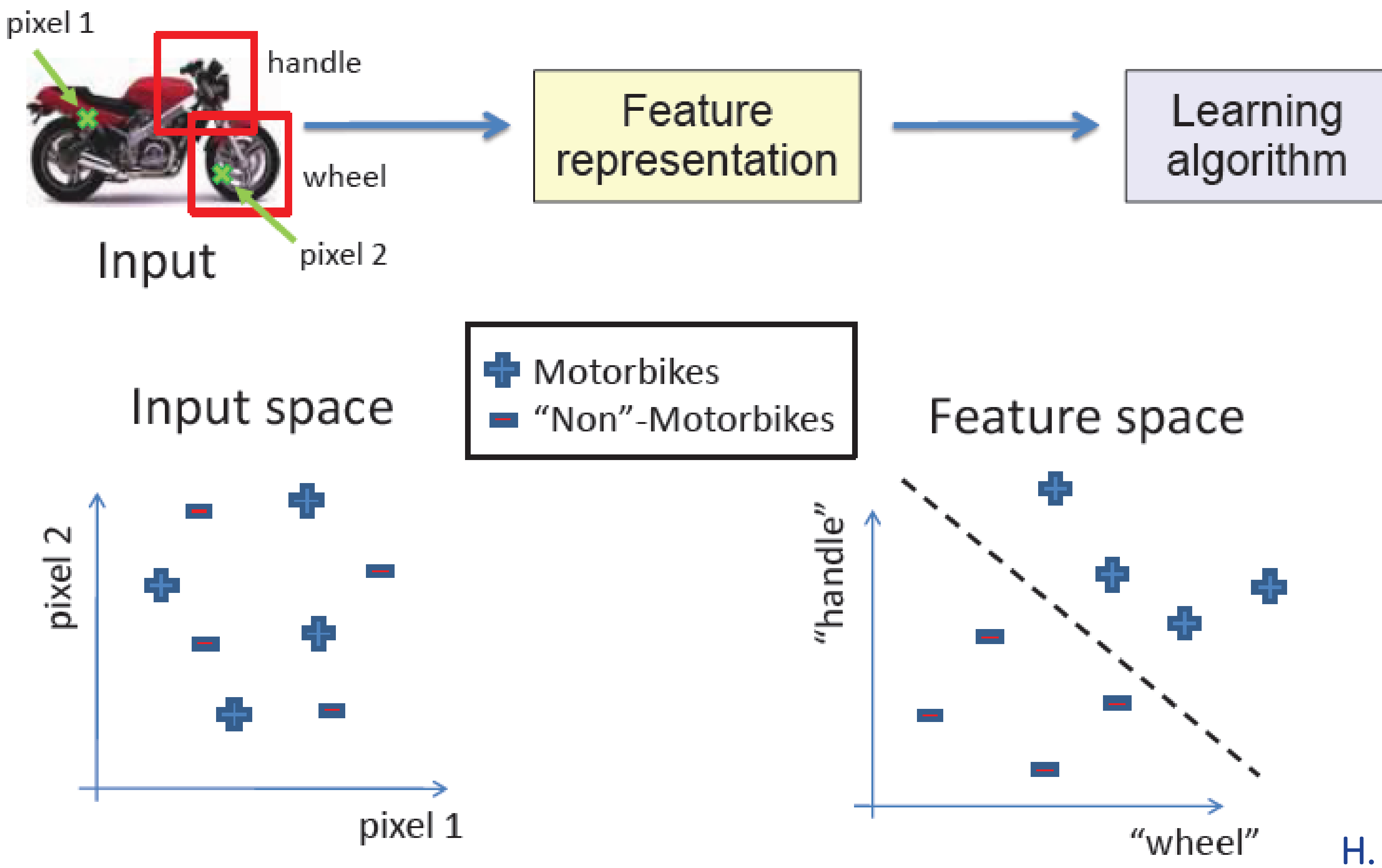
Multimodal (audio/video)

AVLetters Lip reading	Accuracy
Prior art (Zhao et al., 2009)	58.9%
Stanford Feature learning	**65.8%**

Other unsupervised feature learning records:
Pedestrian detection (Yann LeCun)
Different phone recognition task (Geoff Hinton)
PASCAL VOC object classification (Kai Yu)

Andrew Ng., ICML 2011

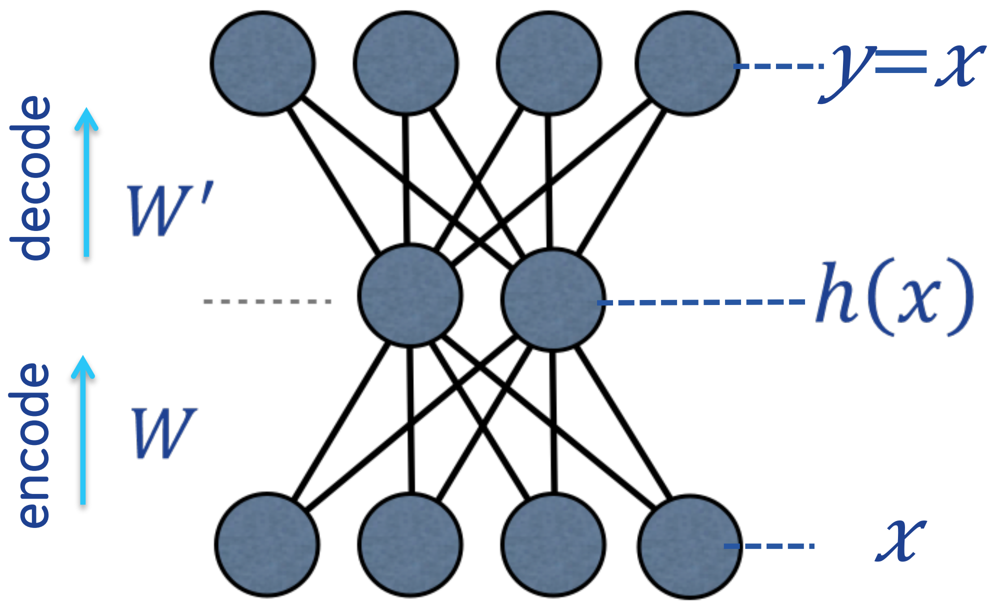
Why unsupervised feature learning works – a simple explanation



H. Lee (2010)

Auto-encoder – one of the popular DL building blocks

AE: a MLP with output being equal to input

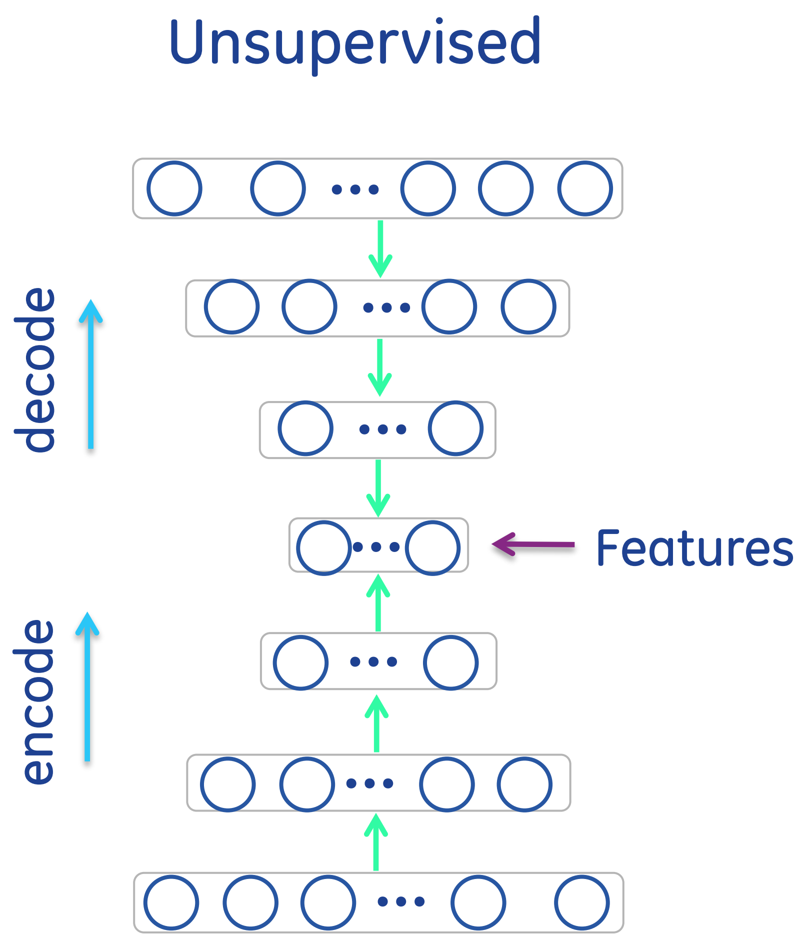


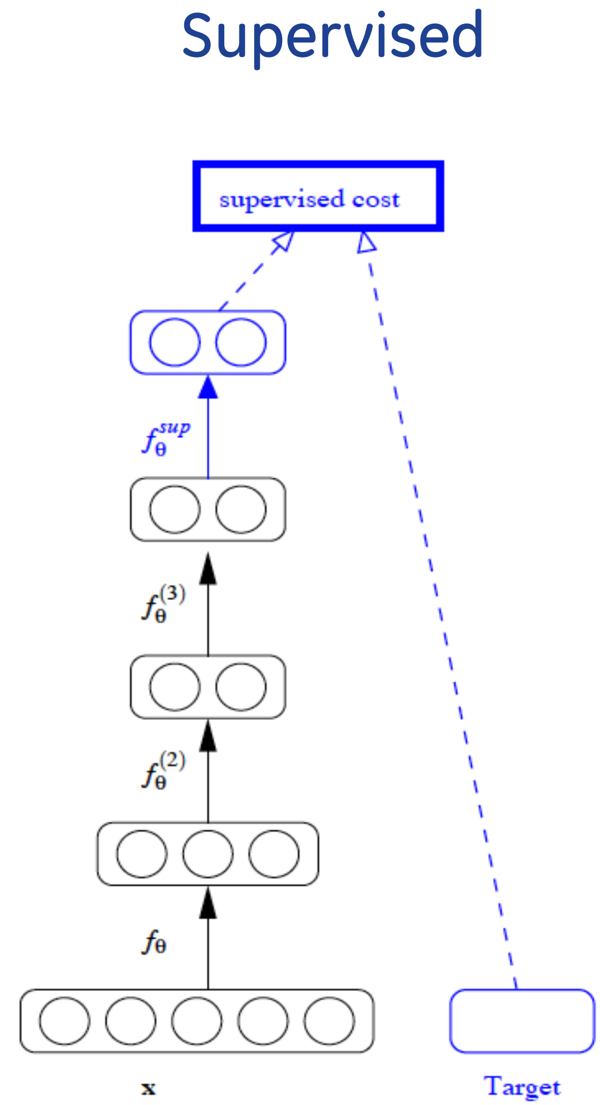
$$y = s_g(W'h + b_y)$$

$$h(x) = s_f(Wx + b_h)$$

$$L(x, y) = -\sum (x_i - y_i)^2$$ OR $$L(x, y) = -\sum x_i log(y_i) + (1 - x_i) log(1 - y_i)$$

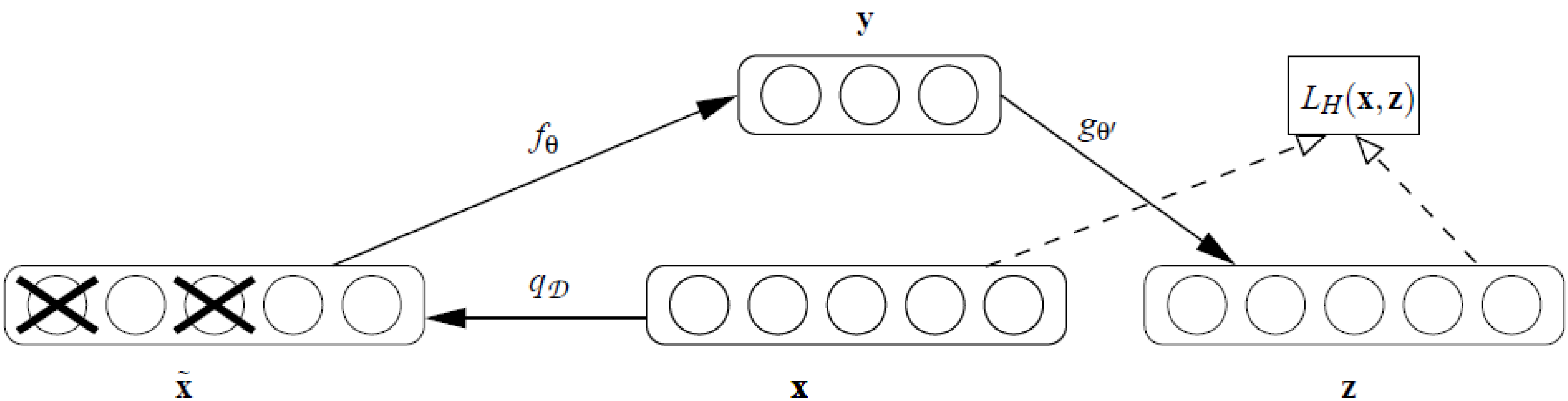
Deep AE





Vincent et al. (2010)

Denoising autoencoder (DAE)

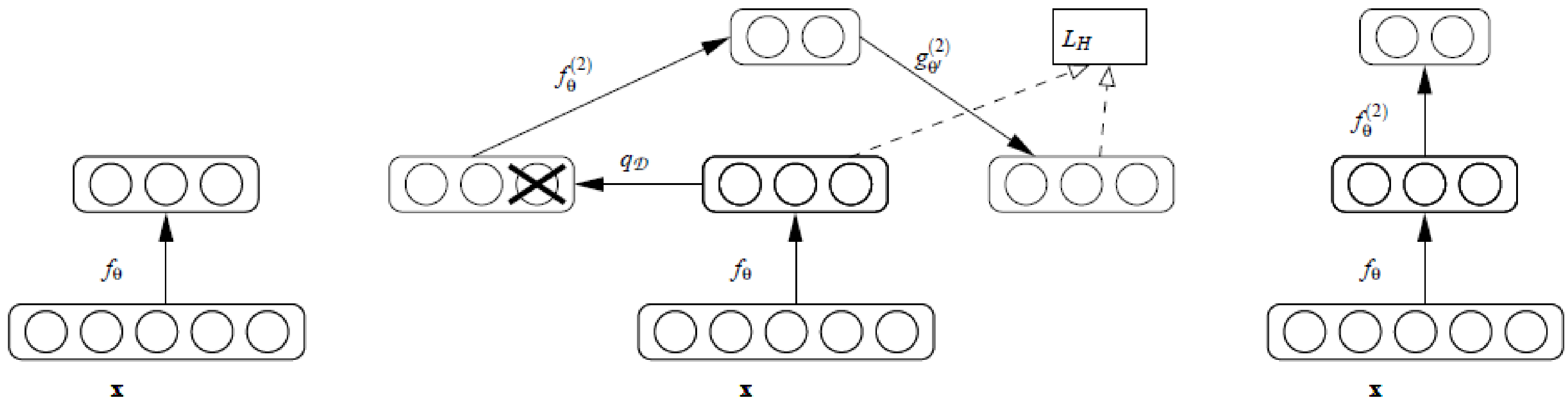


Vincent et al. (2010)

3 different corruption processes:

1. Gaussian noise
2. Masking noise
3. Salt-and-pepper noise

Stacked DAE



Vincent et al. (2010)

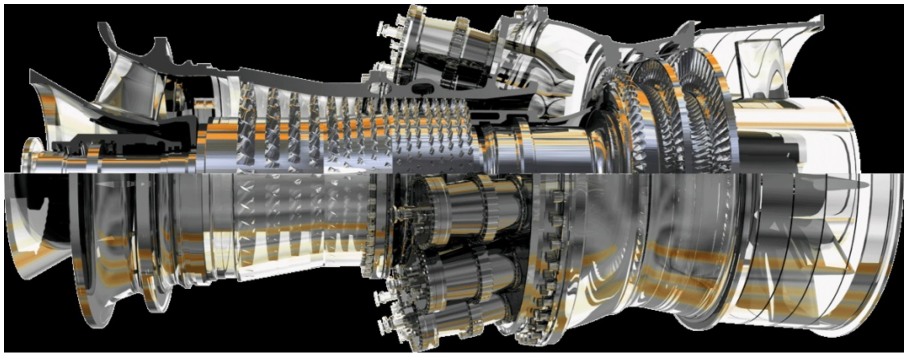
2 design settings:

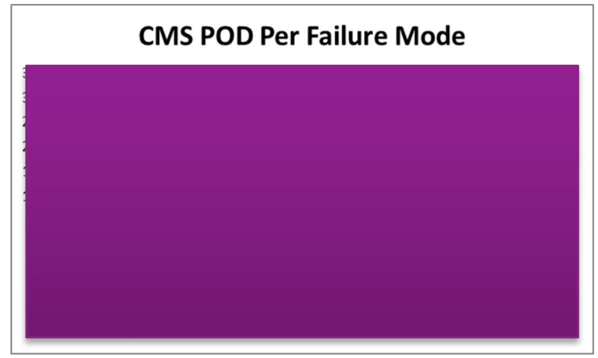
1. Unsupervised feature learning + standalone supervised learning
2. Deep neural network: add logistic regression on top of encoder and supervised fine tune all parameters

A deep feature learning example: Combustor anomaly detection

imagination at work

Gas Turbine Combustor Anomaly Detection





The business pain points

- Current rule-based engine has an insufficient detection rate (*)
- Finding a good set of features (Feature Engineering) takes significant amount of effort
- Labeled data, especially faulty data, is extremely sparse and difficult to get

(*) Source: Reliability combustion events 2008-2010, with M&D data, covering 7&9 E & F class with full-load condition.

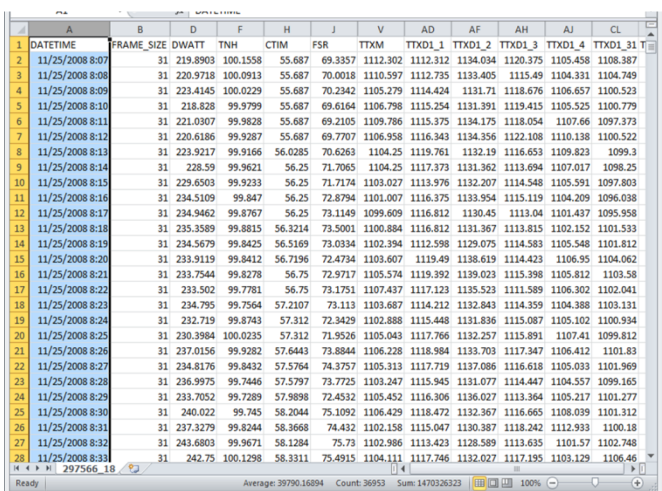
The Data

- Single turbine (TSNxxxxxx)
- Normal (event-free) data: 3 months of data (once per minute)
- POD events: 10 events occurred over 4-month window
- 27 sensor measurements (TC readings)
- Data matrices:

  13,791 x 27 - normal data for feature learning

  300 x 27 - POD events(*)   (*) For POD cases, take 30 points before the POD events

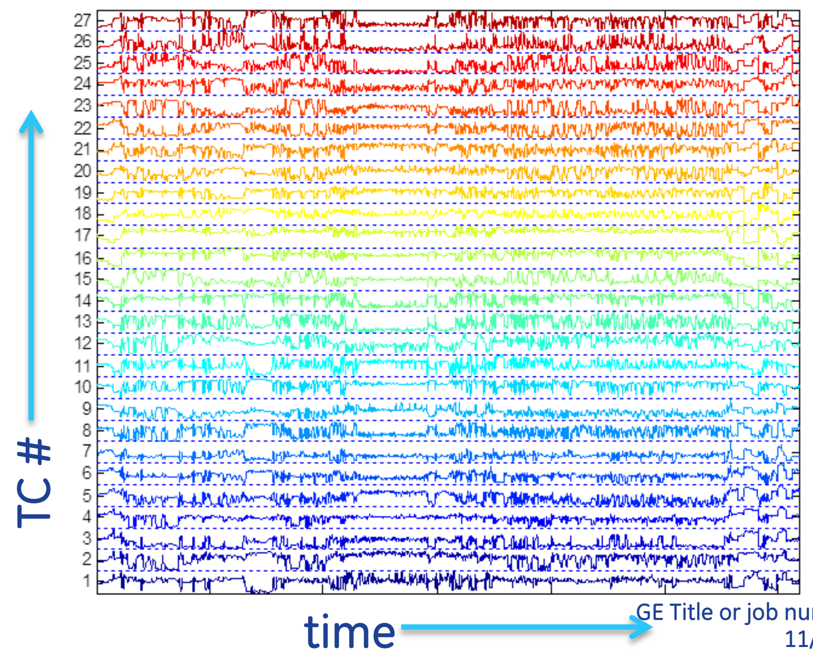
  47,575 x 27 - normal data for model building & validation



	A	B	D	F	H	J	V	AD	AF	AH	AJ	CL
1	DATETIME	FRAME_SIZE	DWATT	TNH	CTIM	FSR	TTXM	TTXD1_1	TTXD1_2	TTXD1_3	TTXD1_4	TTXD1_31
2	11/25/2008 8:07	31	219.8903	100.1558	55.687	69.3357	1112.302	1112.312	1134.034	1120.375	1105.458	1108.387
3	11/25/2008 8:08	31	220.9718	100.0913	55.687	70.0018	1110.597	1112.735	1133.405	1115.49	1104.331	1104.749
4	11/25/2008 8:09	31	223.4145	100.0229	55.687	70.2342	1105.279	1114.424	1131.71	1118.676	1106.657	1100.523
5	11/25/2008 8:10	31	218.828	99.9799	55.687	69.6164	1106.798	1115.254	1131.391	1119.415	1105.525	1100.779
6	11/25/2008 8:11	31	221.0307	99.9828	55.687	69.2105	1109.786	1115.375	1134.175	1118.054	1107.66	1097.373
7	11/25/2008 8:12	31	220.6186	99.9287	55.687	69.7707	1106.958	1116.343	1134.356	1122.108	1110.138	1100.522
8	11/25/2008 8:13	31	223.9217	99.9166	56.0285	70.6263	1104.25	1119.761	1132.19	1116.653	1109.823	1099.3
9	11/25/2008 8:14	31	228.59	99.9621	56.25	71.7065	1104.25	1117.373	1131.362	1113.694	1107.017	1098.25
10	11/25/2008 8:15	31	229.6503	99.9233	56.25	71.7174	1103.027	1113.976	1132.207	1114.548	1105.591	1097.803
11	11/25/2008 8:16	31	234.5109	99.847	56.25	72.8794	1101.007	1116.375	1133.954	1115.119	1104.209	1096.038
12	11/25/2008 8:17	31	234.9462	99.8767	56.25	73.1149	1099.609	1116.812	1130.45	1113.04	1101.437	1095.958
13	11/25/2008 8:18	31	235.3589	99.8815	56.3214	73.5001	1100.884	1116.812	1131.367	1113.815	1102.152	1101.533
14	11/25/2008 8:19	31	234.5679	99.8425	56.5169	73.0334	1102.394	1112.598	1129.075	1114.583	1105.548	1101.812
15	11/25/2008 8:20	31	233.9119	99.8412	56.7196	72.4734	1103.607	1119.49	1138.619	1114.423	1106.95	1104.062
16	11/25/2008 8:21	31	233.7544	99.8278	56.75	72.9717	1105.574	1119.392	1139.023	1115.398	1105.812	1103.58
17	11/25/2008 8:22	31	233.502	99.7781	56.75	73.1751	1107.437	1117.123	1135.523	1111.589	1106.302	1102.041
18	11/25/2008 8:23	31	234.795	99.7564	57.2107	73.113	1103.687	1114.212	1132.843	1114.359	1104.388	1103.131
19	11/25/2008 8:24	31	232.719	99.8743	57.312	72.3429	1102.888	1115.448	1131.836	1115.087	1105.102	1100.934
20	11/25/2008 8:25	31	230.3984	100.0235	57.312	71.9526	1105.043	1117.766	1132.257	1115.891	1107.41	1099.812
21	11/25/2008 8:26	31	237.0156	99.9282	57.6443	73.8844	1106.228	1118.984	1133.703	1117.347	1106.412	1101.83
22	11/25/2008 8:27	31	234.8176	99.8432	57.5764	74.3757	1105.313	1117.719	1137.086	1116.618	1105.033	1101.969
23	11/25/2008 8:28	31	236.9975	99.7446	57.5797	73.7725	1103.247	1115.945	1131.077	1114.447	1104.557	1099.165
24	11/25/2008 8:29	31	233.7052	99.7289	57.9898	72.4532	1105.452	1116.306	1136.027	1113.364	1105.217	1101.277
25	11/25/2008 8:30	31	240.022	99.745	58.2044	75.1092	1106.429	1118.472	1132.367	1116.665	1108.039	1101.312
26	11/25/2008 8:31	31	237.3279	99.8244	58.3668	74.432	1102.158	1115.047	1130.387	1118.242	1112.933	1100.18
27	11/25/2008 8:32	31	243.6803	99.9671	58.1284	75.73	1102.986	1113.423	1128.589	1113.635	1101.57	1102.748
28	11/25/2008 8:33	31	242.75	100.1298	58.3311	75.4915	1104.111	1117.746	1132.027	1117.195	1103.129	1106.46

297566_18

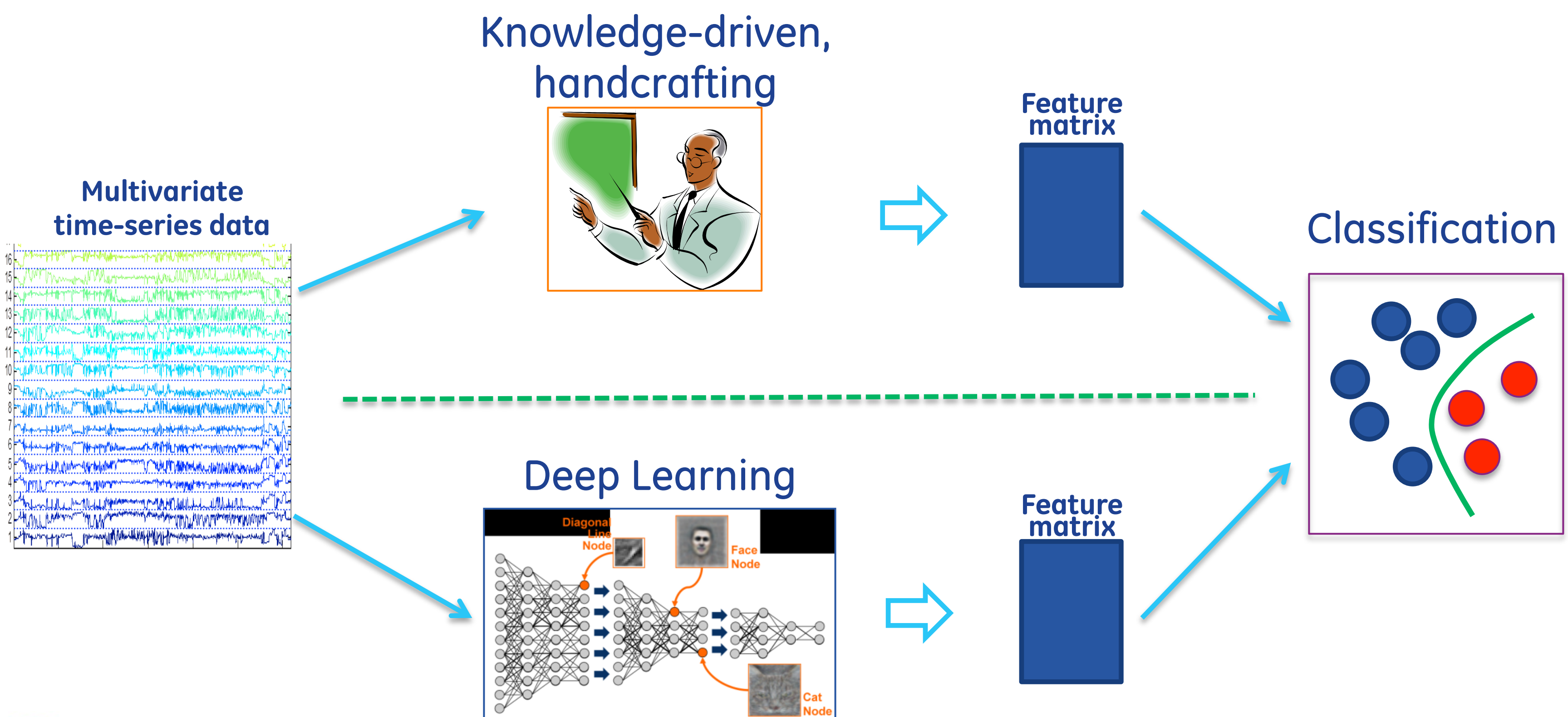
Ready   Average: 39790.16894   Count: 36953   Sum: 1470326323   100%



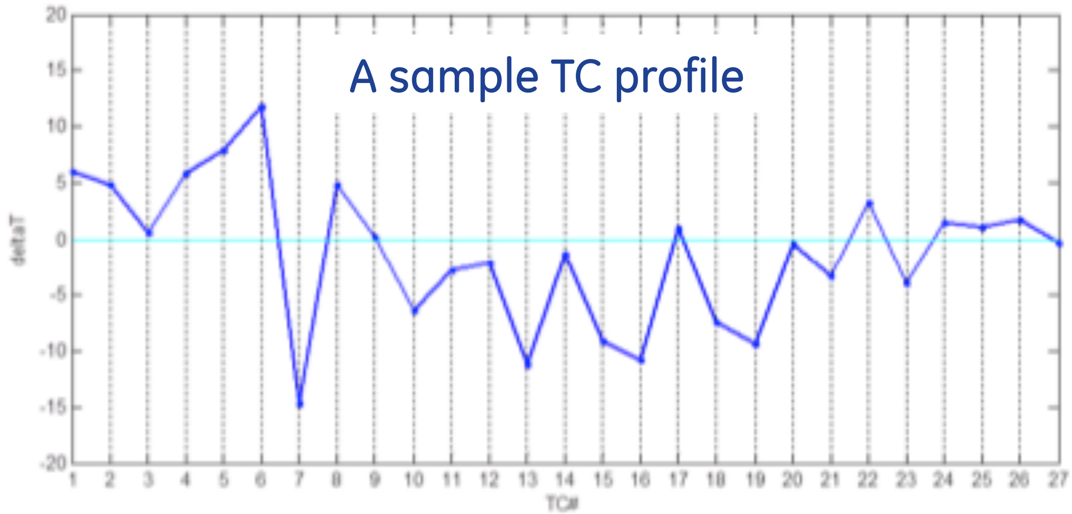
Experiment setup

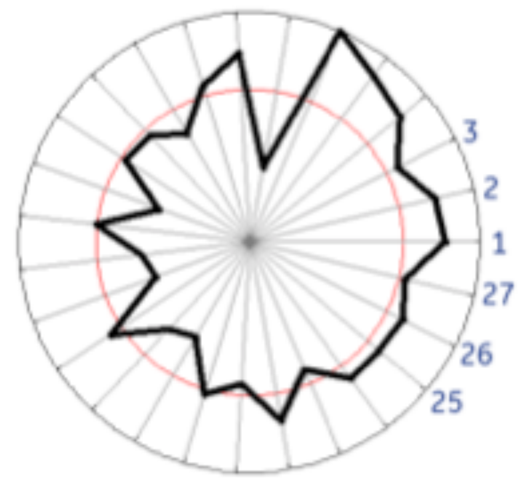
- Unsupervised feature learning

Our goal is to compare learned features against handcrafted features in terms of classification performance



Domain-driven, handcrafted features



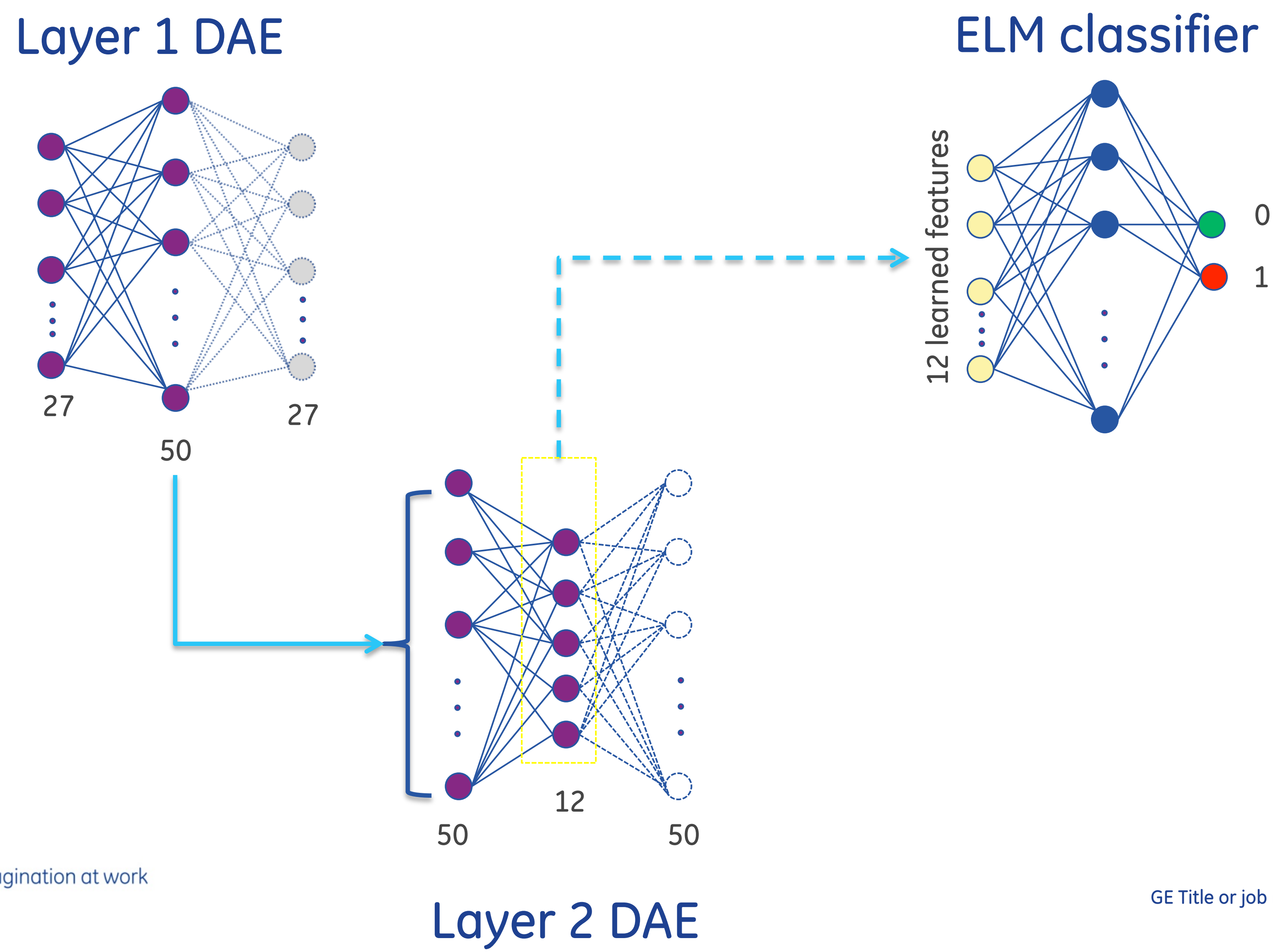


Extracted 12 features

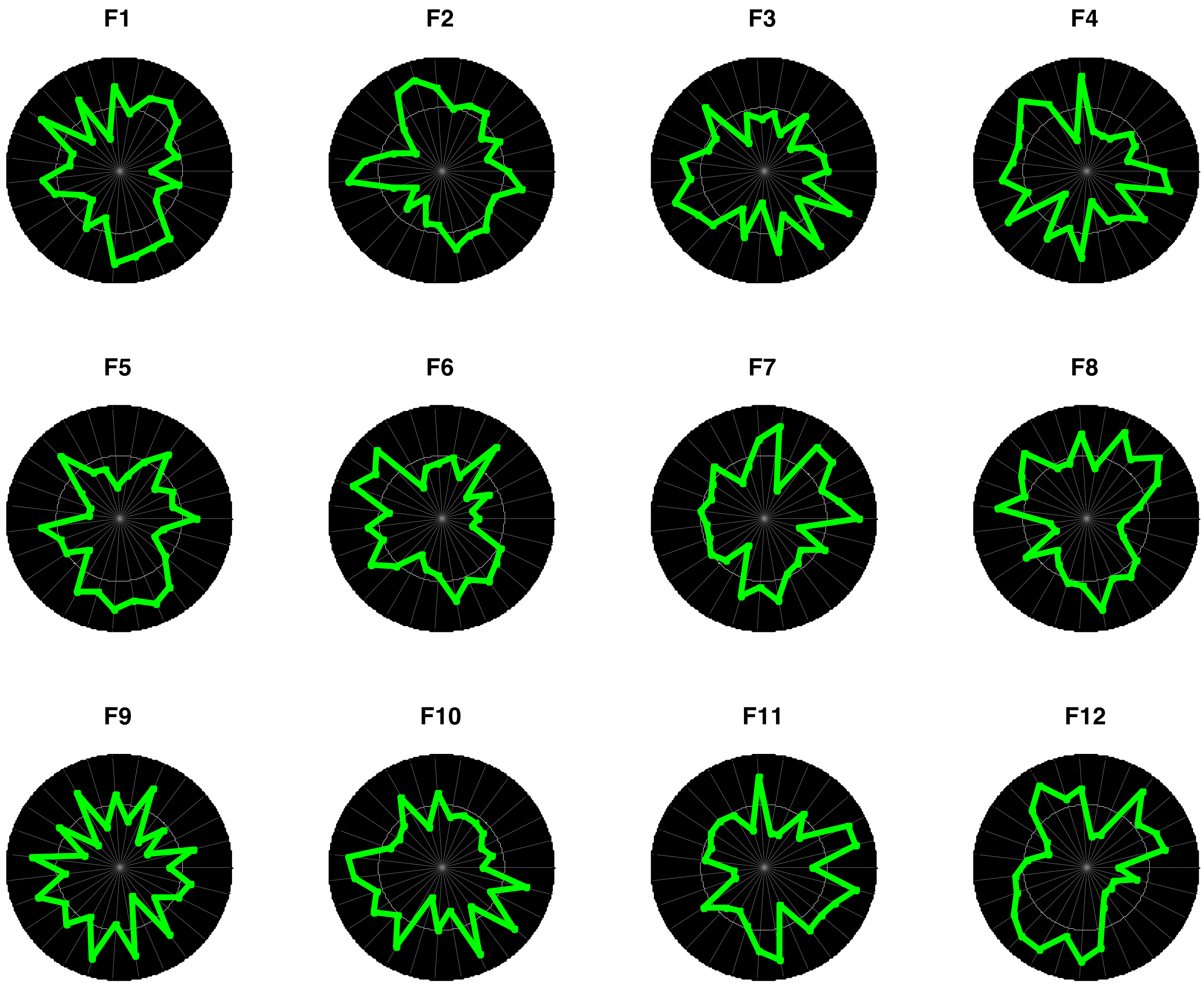
1	DWATT
2	TNH
3	max
4	mean
5	std
6	median
7	# diff b/w positive & negative TCs
8	zero crossing
9	kurtosis
10	skewness
11	max of 3-pt sum
12	max of 3-pt median

imagination at work

Deep feature learning



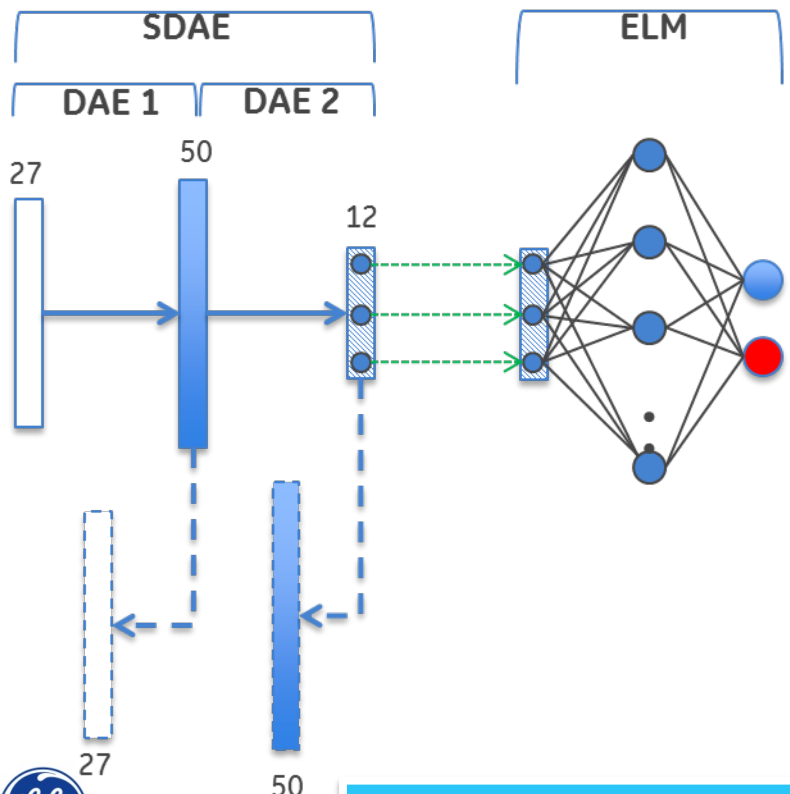
Learned features

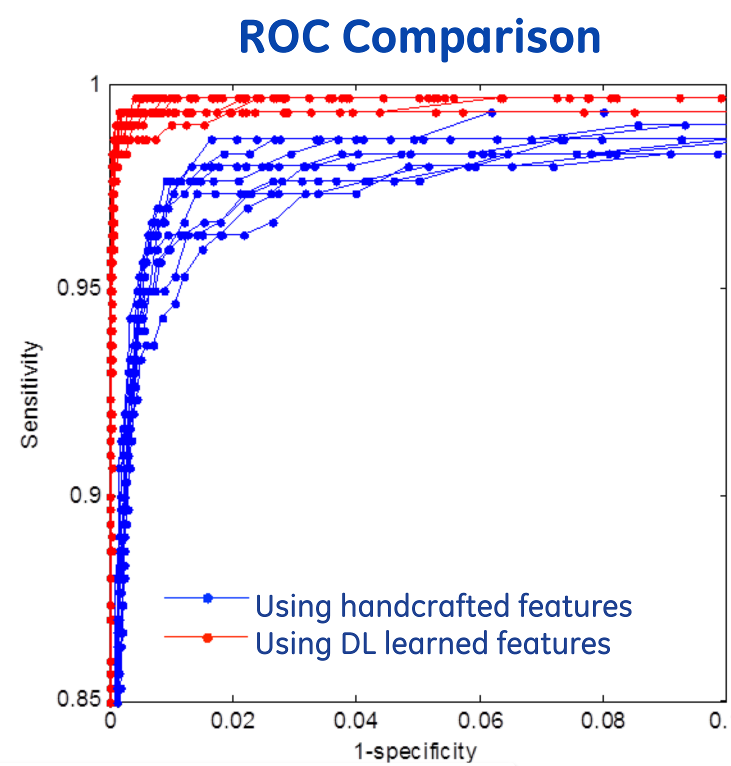


Classification Modeling and Results

Modeling details:

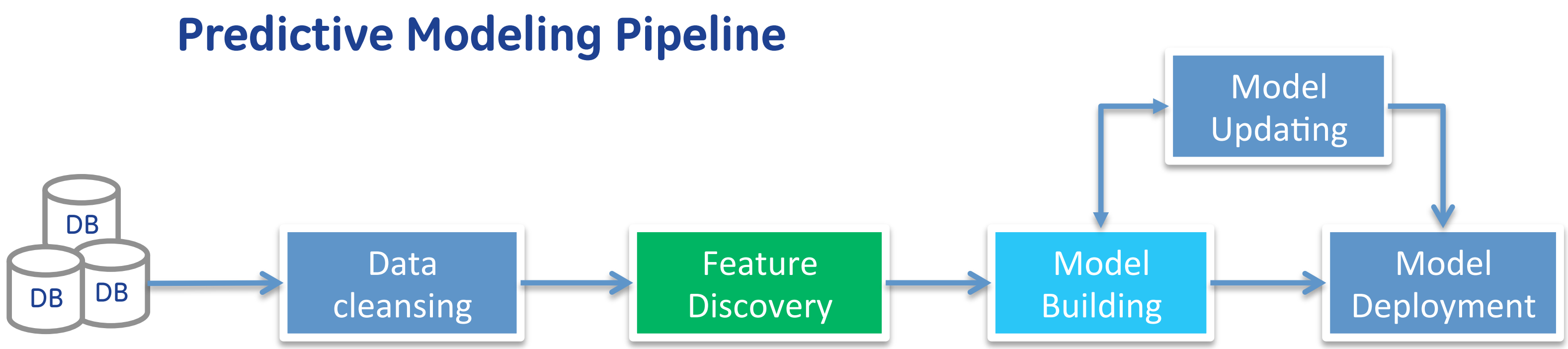
- ELM (a special type of feed-forward Neural network) as the classifier
- Unbalanced data strategy: sample weighting
- Validation method: 5-fold cross-validation (10 times of random runs)





Deep learning can learn features that give much better detection rate than manually-handcrafted features do

Final Remarks -1



- Feature discovery (both FE and FL) is more important than model building, yet it is less well-studied.
- Feature discovery, not model building, can be the differentiator.

imagination at work

Final Remarks - 2

- Traditional knowledge-driven feature engineering is hard and time-consuming, thus is insufficient.
- Feature learning, especially recently-developed deep feature learning, is data-driven, and has some potential in alleviating difficulties faced in FE.

2 directions worth pursuing:

- ✧ Integrating domain knowledge into feature learning **(R)**
- ✧ Tools that can automate feature discovery **(D)**

1 question to be answered:
Is deep learning effective for PHM applications?

Thank You

Questions?

My contact information:
Dr. Weizhong Yan
Principal Scientist
Machine Learning Lab
GE Global Research Center
Email: yan@ge.com

imagination at work