The State of the Art in Machine Learning

(and how to survive the coming robot apocalypse)

Agenda

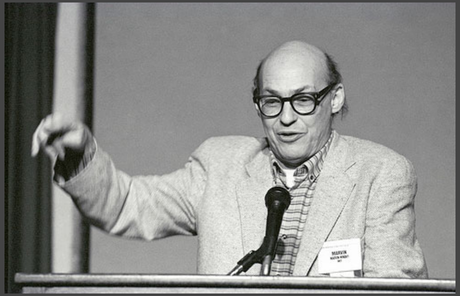
- A little theory and history
- The map is not the territory: a somewhat artificial taxonomy of ML
- What's new in machine learning?
  - A little more history and distinctions
- How it impacts the company
- Existential threats and smart machines: how it impacts everything.
- Demo: Statistica ML capabilities
- Q/A

A. Kolmogorov

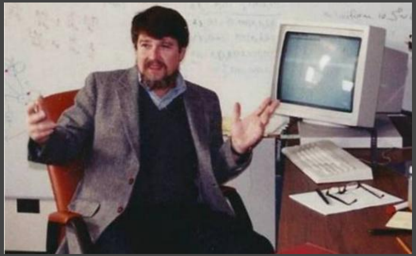
A. Samuel

R. Solomonoff

V. Vapnik



M. Minksy



D. Rumelhart

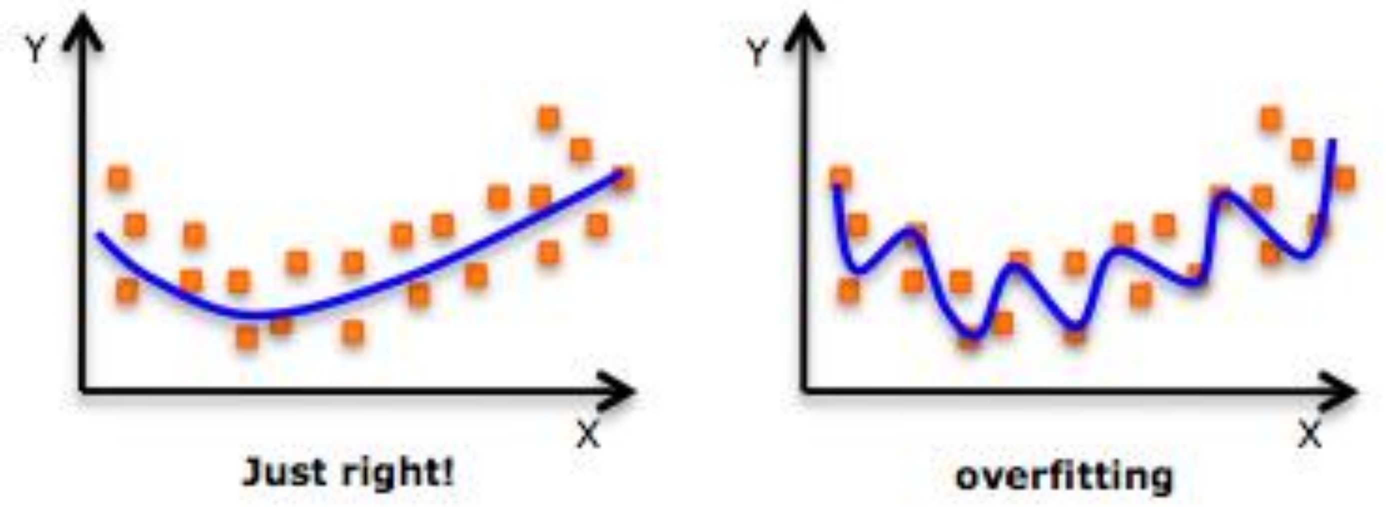
G. Hinton

Some Theory

- Start with Kolmogorov or Algorithmic Information Theory (AIT): given a sequence of bits, the probability of a given sequence gets small fast, $p(s) = 2^{-m}$
- The algorithmic complexity (AC) of a bit sequence is the length of the smallest generator algorithm on a universal Turing machine for the sequence
- AC is uncomputable
- The smallest generator algorithm makes the fewest assumptions about the data:
  - Formalization of Occam's Razor
  - Given a bit sequence of size m generated by an underlying function of size (n<m), discovering that function eliminates all risk for misgenerating extensions of the sequence
  - Inductive ideal
- **But the ideal inductive algorithm is uncomputable.**

Relevance of Theory

Compression == Truth



Note: Other related theory includes VC dimension and Valiant PAC theory.
Also see Johnson-Lindenstrauss lemma and Random Indexing.

Unsupervised Learning

Clustering Algorithms

- Hierarchical
- Agglomerative
- Splitting

Autoencoder Neural Networks

SOMs (self-organizing maps)

Singular Value Decomposition

Principal Components Analysis

Sequence Induction (Sequitur)

EM methods like Latent Dirichlet Allocation

Supervised Learning

Generalized Linear Models

- Linear Regression
- Logistic Regression

Categorizers

- Support Vector Machines
- K Nearest Neighbor
- Naïve Bayes Algorithms
- Tree Algorithms
- C&RT
- Random Forests
- Artificial Neural Networks/Deep Learning
- Extreme machines/Reservoir

Other/Related

Semi-supervised variants

Complicated stuff

- Image and scene understanding

Evolutionary optimization

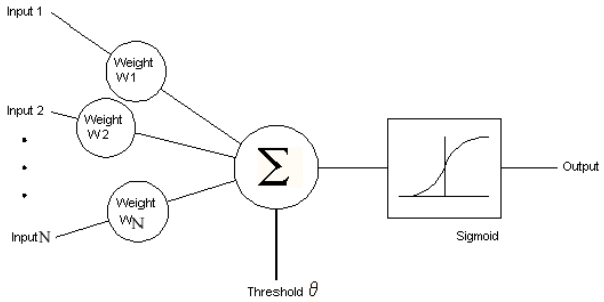
NLP-specific stuff

- Continuous Automatic Speech Recognition
- Grammar induction (Supervised or Unsupervised)
- Sequence prediction

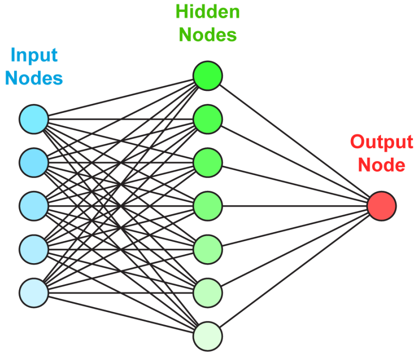
What's New in Machine Learning?

- What's old is new again!
  - Ensemble methods use known methods in groups or ensembles to reduce overfitting:
    - Bagging: Overlapping sub-training sets
    - Boosting: Boost some learners that are successful on misclassified data to improve the overall performance
    - Random Forests: use a bunch of classifier trees trained on bagged data subsets and take the mode for classification
    - Netflix Challenge
  - Neural Networks via "Deep Learning":
    - Pre-train via autoencoding
    - Then backpropagation train for the categories

Artificial Neural Networks get Real Deep



Perceptron
(Analysis by Minsky and Papert, 1969)

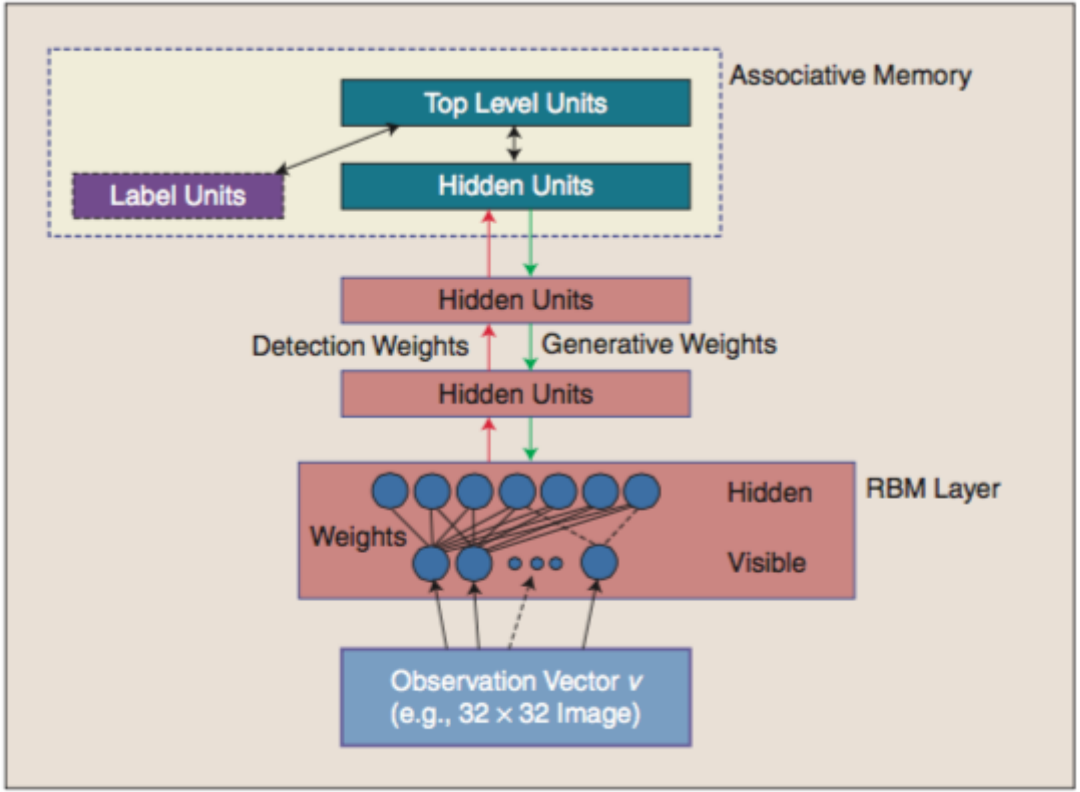


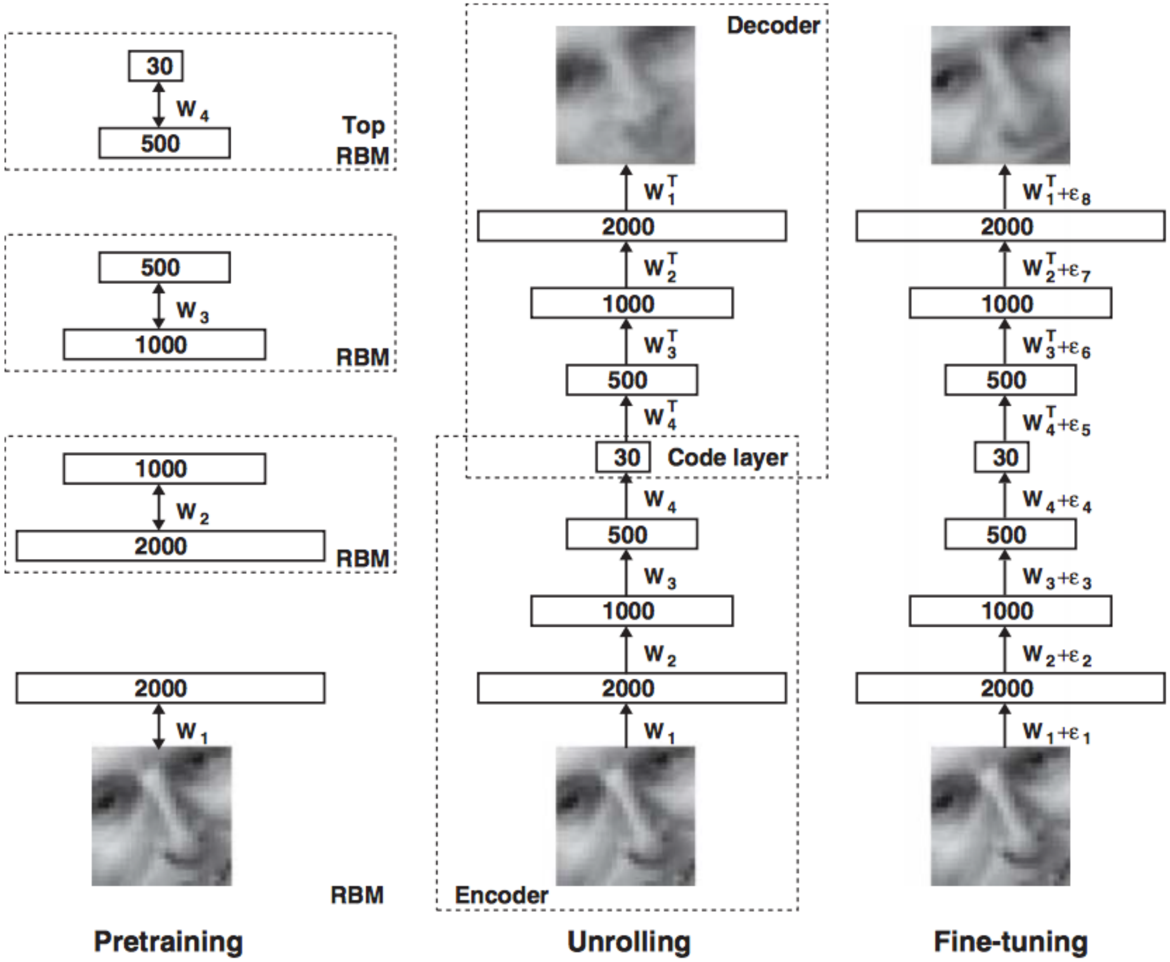
Connectionism
Backpropagation
(PDP books, late 80s)

NOTE: the tank parable

Deep Learning

- Hinton's Contrastive Divergence
- Restricted Boltzmann Machines
- Train to autoencode starting with the observation layer
- Training is Gibbs sampling to reduce error at outcome of each layer by modifying weights
- Other advances: dropout (leave some nodes out to avoid overfitting)





30
W 4
500
Top RBM
500
W 3
1000
RBM
1000
W 2
2000
RBM
2000
W 1
RBM
Pretraining
Decoder
W 1 T
2000
W 2 T
1000
W 3 T
500
W 4 T
30
Code layer
W 4
500
W 3
1000
W 2
2000
W 1
Encoder
Unrolling
W 1 T +ε 8
2000
W 2 T +ε 7
1000
W 3 T +ε 6
500
W 4 T +ε 5
30
W 4 +ε 4
500
W 3 +ε 3
1000
W 2 +ε 2
2000
W 1 +ε 1
Fine-tuning

General Notes of Performa

- Yang, Y. and X. Liu, A re-examination of text categorization methods. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99, pp 42--49), 1999.

Table 1: Performance summary of classifiers

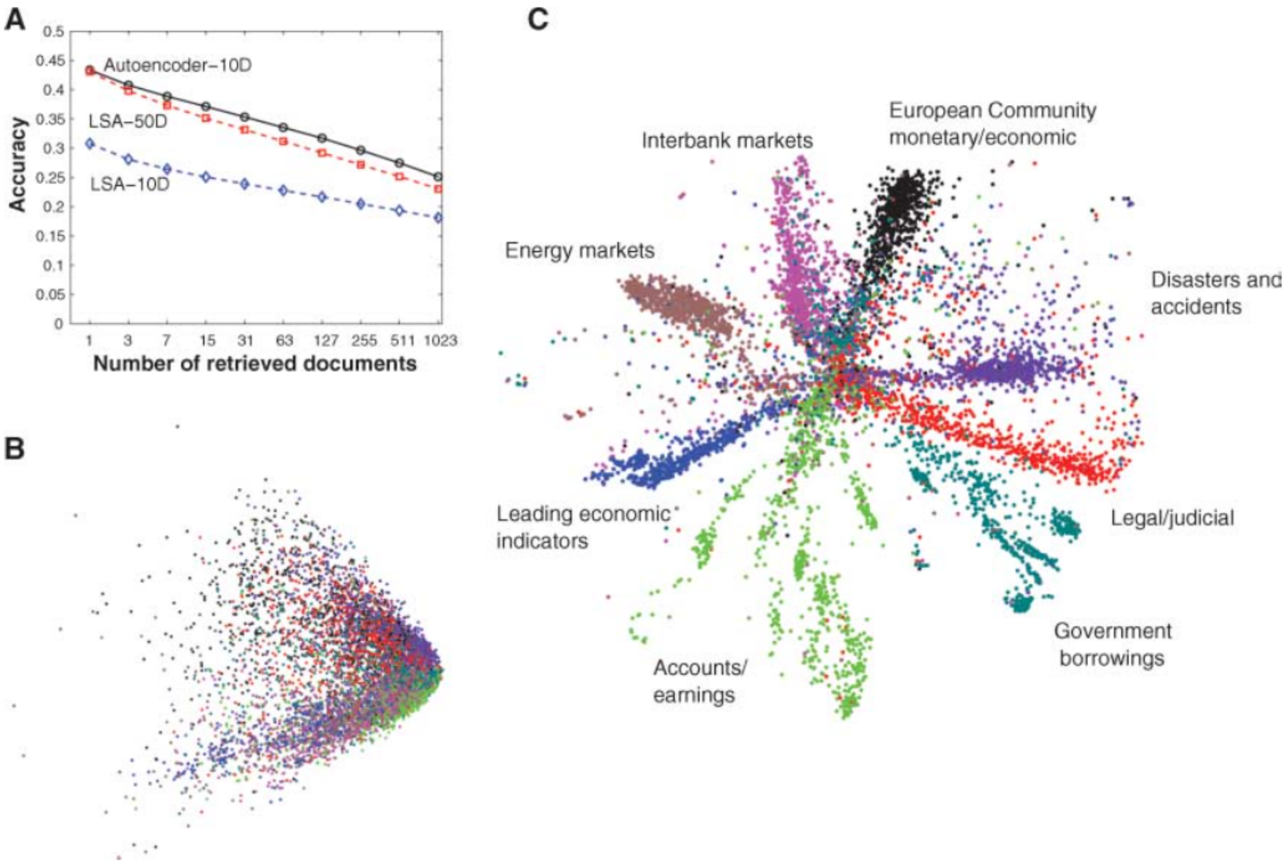
method	miR	miP	miF1	maF1	error
SVM	.8120	.9137	.8599	.5251	.00365
KNN	.8339	.8807	.8567	.5242	.00385
LSF	.8507	.8489	.8498	.5008	.00414
NNet	.7842	.8785	.8287	.3765	.00447
NB	.7688	.8245	.7956	.3886	.00544

miR = micro-avg recall; miP = micro-avg prec.;
miF1 = micro-avg F1; maF1 = macro-avg F1.

Table 2: Statistical significance test results

sysA	sysB	s-test	S-test	T-test	T'-test
SVM	kNN	$>$	$\sim$	$\sim$	$\sim$
SVM	LLSF	$\gg$	$\sim$	$\sim$	$\sim$
kNN	LLSF	$\gg$	$\sim$	$\sim$	$\sim$
SVM	NNet	$\gg$	$\gg$	$\gg$	$\gg$
kNN	NNet	$\gg$	$\gg$	$\gg$	$\gg$
LLSF	NNet	$\sim$	$\gg$	$\gg$	$\gg$
NB	kNN	$\ll$	$\ll$	$\ll$	$\ll$
NB	LLSF	$\ll$	$\ll$	$\ll$	$\ll$
NB	SVM	$\ll$	$\ll$	$\ll$	$\ll$
NB	NNet	$\ll$	$\sim$	$\sim$	$\sim$

"$\gg$" or "$\ll$" means P-value ≤ 0.01;
"$>$" or "$<$" means $0.01 <$ P-value ≤ 0.05;
"$\sim$" means P-value > 0.05.



28 JULY 2006 VOL 313 SCIENCE www.sciencemag.org

Deep Learner Performance

IMAGENET **Large Scale Visual Recognition Challenge 2014 (ILSVRC2014)**

Results of ILSVRC2014

Object detection Classification+Localization **Team information** Per-class results

Legend:
Yellow background = winner in this task according to this metric; authors are willing to reveal the method
White background = authors are willing to reveal the method
Grey background = authors chose not to reveal the method
Italics = authors requested entry not participate in competition

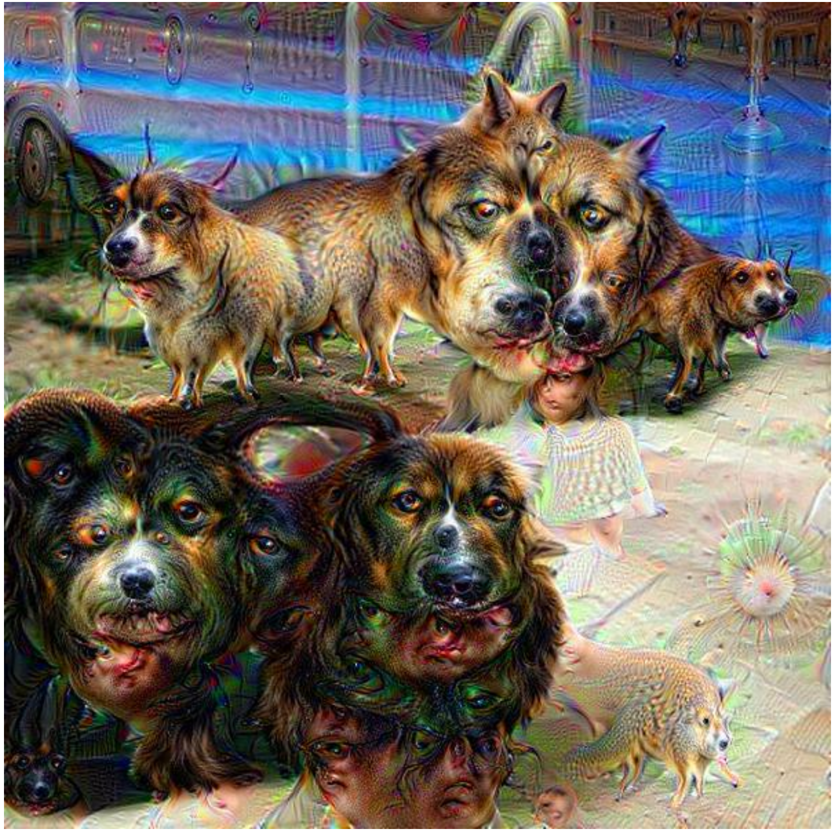
Object detection

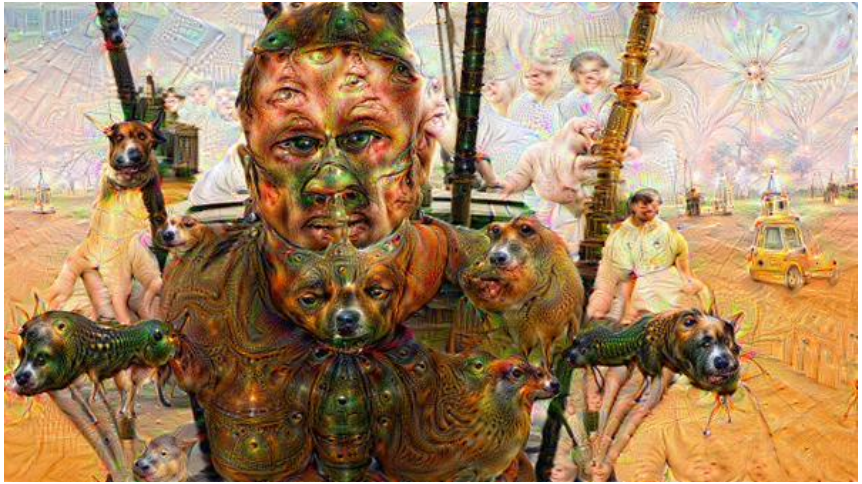
Task 1a: Object detection with provided training data

Object detection with provided training data: Ordered by number of categories won

Team name	Entry description	Number of object categories won	mean AP
NUS	Multiple Model Fusion with Context Rescoring	106	0.37212
MSRA Visual Computing	A combination of multiple SPP-net-based models (no outside data)	45	0.351103
UvA-Euvision	Deep learning with provided data	21	0.320253
1-HKUST	run 2	18	0.288669
Southeast-CASIA	CNN-based proposal classification with proposal filtration and model combination	4	0.304022
1-HKUST	run 4	4	0.285616
Southeast-CASIA	CNN-based proposal classification with proposal filtration and sample balance	2	0.304783
1-HKUST	run 2	0	0.288669
CASIA_CRIPAC_2	CNN-based proposal classification with part classification and object regression	0	0.286158
1-HKUST	run 3	0	0.284595
1-HKUST	run 1	0	0.261543
MSRA Visual Computing	*A single SPP-net model for detection (no outside data)*	---	*0.318403*

Dali would be proud





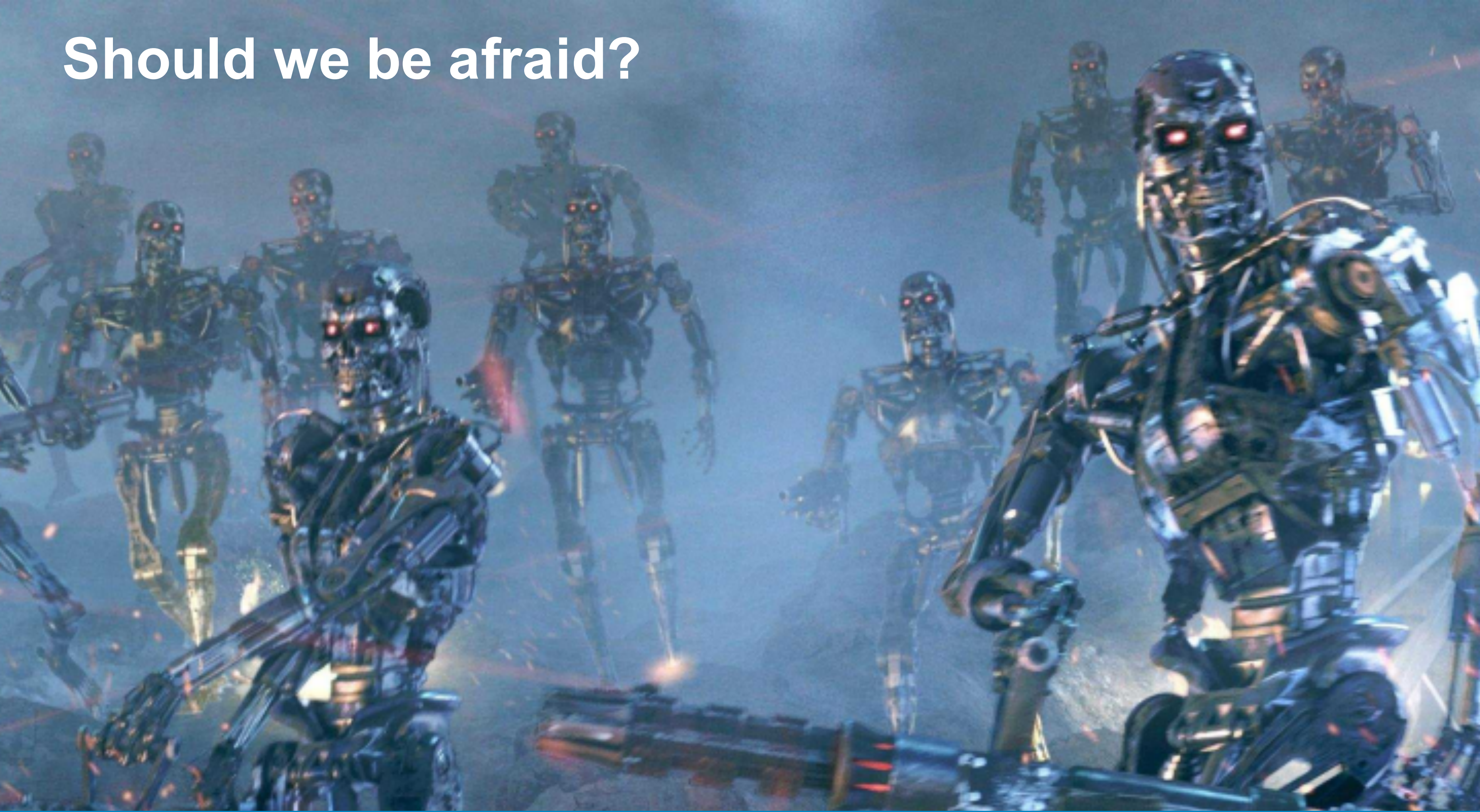
Related Methods

- Convolutional Neural Networks (CNNs)
- Reservoir Computing: related to Boltzmann machines and Hopfield networks
- Extreme Learning: random initial shallow networks and similar training algorithms

Summary of Impacts

- Incremental advances in performance to near-human level for some tasks
  - Good success in image labeling (MNIST at almost human levels)
  - Some progress with Natural Language Processing
  - Lift improvements of single digit/low double digit percentages
- GPU/FPGA/custom silicon training performance improvements: potentially orders of magnitude
- Advanced analytics and ML will continue to become a part of our lives
  - Optimizing business processes
  - Predicting opportunities and threats
  - Prescribing solutions

Should we be afraid?



I visualize a time when we will be to robots what dogs are to humans, and I'm rooting for the machines.

Claude Shannon in OMNI Magazine, 1987

Should we be afraid?

- Recent fearmongering by Stephen Hawking, Elon Musk, Bill Gates, et. al.
- Formation of Future of Life Institute
- Machine Intelligence Research Institute (MIRI): Values Learning Research
- Nick Bostrom's Existential Threats:
  - Goals of self-modifying AI may not be our goals.
  - Are human values learnable/programmable in a way that is not modifiable?

So, again, should we be afraid?

- No general theory of “universal AI”. Well, OK, very abstract ones:
  - Marcus Hutter
  - Jurgen Schmidhuber
- No productive theory of “self-modifying universal AI”
- No examples of high degrees of structural complexity leading to high degrees of surprise…exactly the opposite actually.

Don’t be afraid…trust me.