MATHEMATICAL FRONTIERS

The National Academies of SCIENCES ENGINEERING MEDICINE

nas.edu/MathFrontiers

Board on
Mathematical Sciences & Analytics

MATHEMATICAL FRONTIERS
2018 Monthly Webinar Series, 2-3pm ET

February 13*:
Mathematics of the Electric Grid

March 13*:
Probability for People and Places

April 10*:
Social and Biological Networks

May 8*:
Mathematics of Redistricting

June 12*: *Number Theory: The Riemann Hypothesis*

July 10*: *Topology*

August 14*: *Algorithms for Threat Detection*

September 11*: *Mathematical Analysis*

October 9*: *Combinatorics*

November 13:
Why Machine Learning Works

December 11:
Mathematics of Epidemics

***** ***Recording posted***

Made possible by support for BMSA from the National Science Foundation Division of Mathematical Sciences and the Department of Energy Advanced Scientific Computing Research

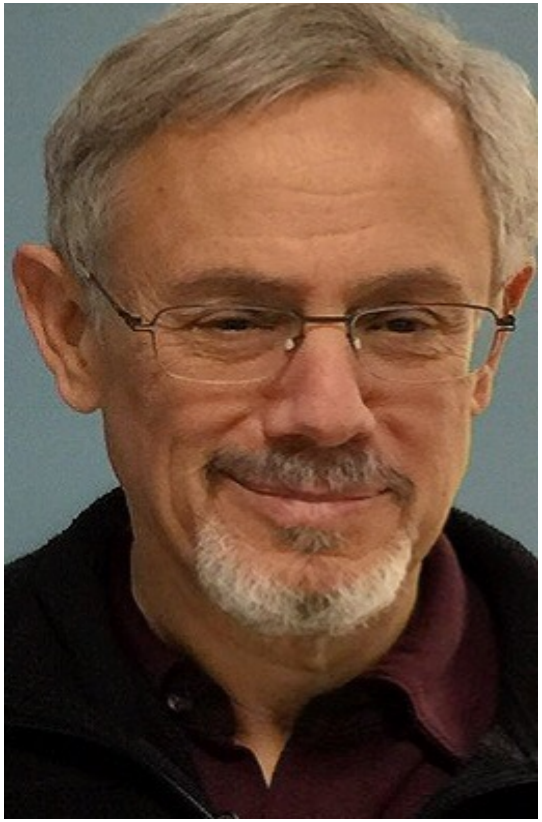
View webinar videos and learn more about BMSA at www.nas.edu/MathFrontiers

MATHEMATICAL FRONTIERS

Why Machine Learning Works

Aarti Singh,
Carnegie Mellon University

David Donoho,
Stanford University



Mark Green,
UCLA (moderator)

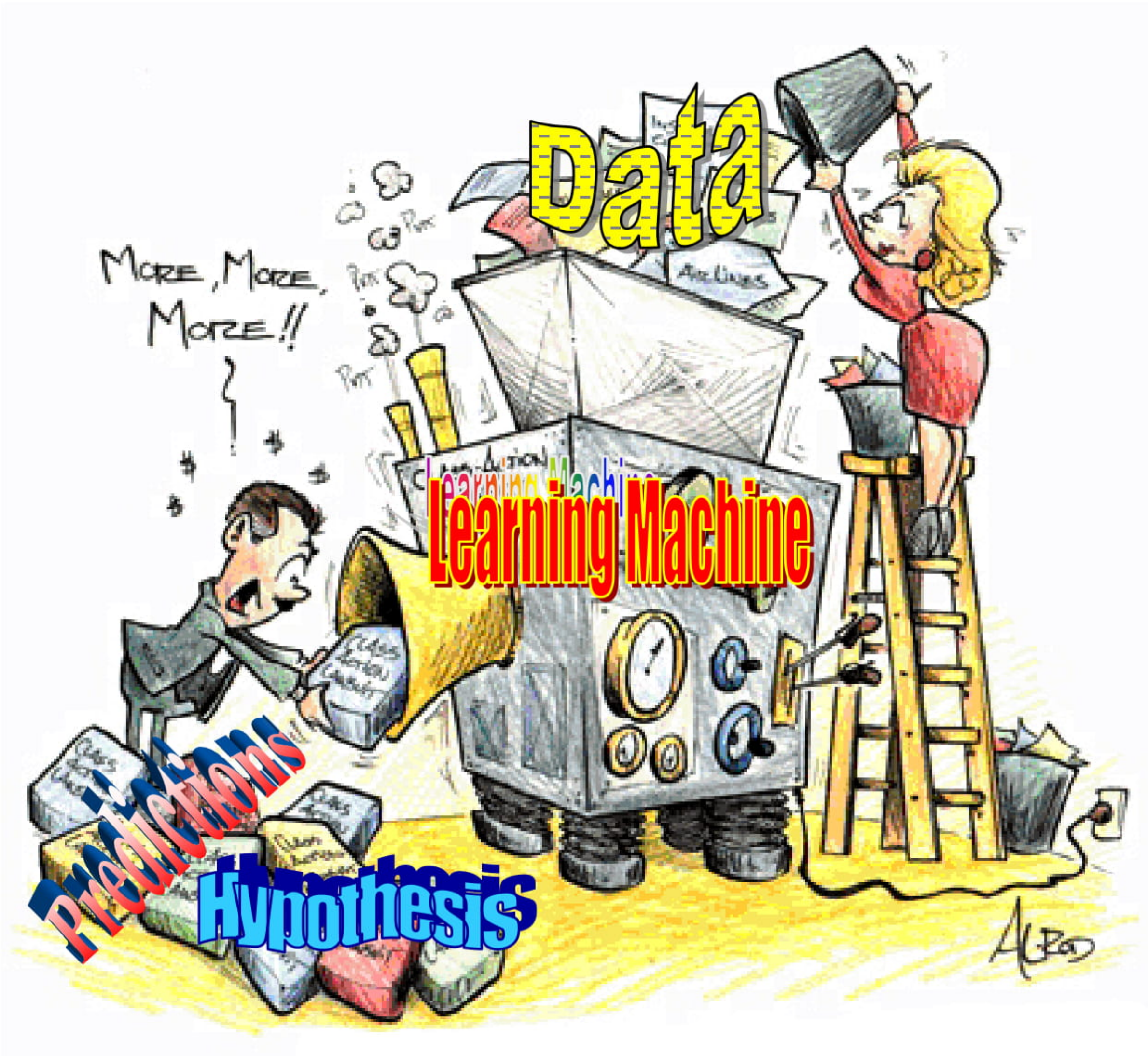
MATHEMATICAL FRONTIERS
Why Machine Learning Works

Aarti Singh,
Carnegie Mellon University

Associate Professor,
Machine Learning Department

Why Machine Learning Works

What is Machine Learning?



Data

⇩

Learning algorithm

⇩

Knowledge

What is Machine Learning?

Design and Analysis of algorithms that

- improve their performance
- at some task
- with experience

Tom Mitchell
Carnegie Mellon Univ.

What is Machine Learning?

Design and Analysis of algorithms that

- improve their <u>performance</u>
- at some <u>task</u>
- with <u>experience</u>

Tom Mitchell
Carnegie Mellon Univ.

Data ⇒ **Learning algorithm** ⇒ Knowledge

(experience) **(performance on task)**

Understanding ML ingredients

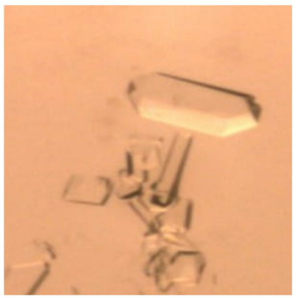
http://phillips-lab.biochem.wisc.edu/

Task: Learning stage of protein crystallization

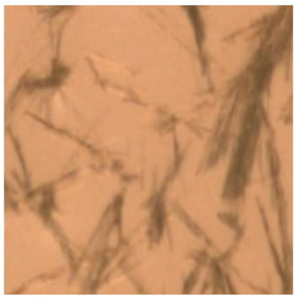
Understanding ML ingredients

http://phillips-lab.biochem.wisc.edu/

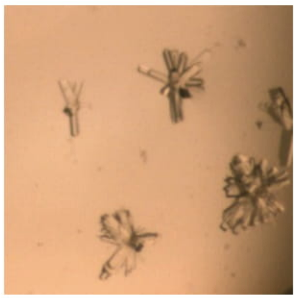
Task: Learning stage of protein crystallization



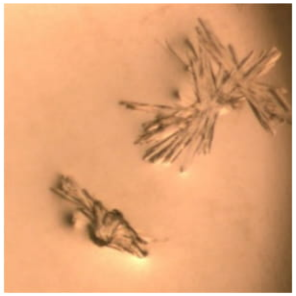
Crystal



Needle

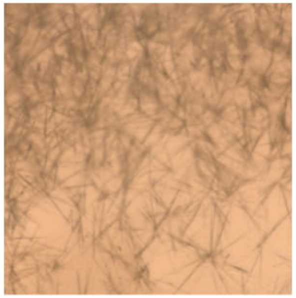


Tree



Tree

Empty



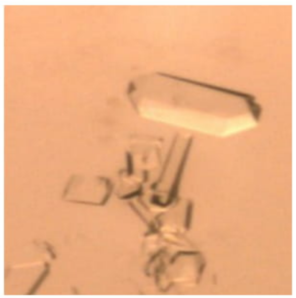
Needle

Experience

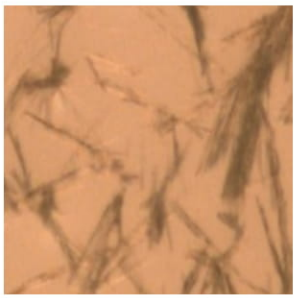
Understanding ML ingredients

http://phillips-lab.biochem.wisc.edu/

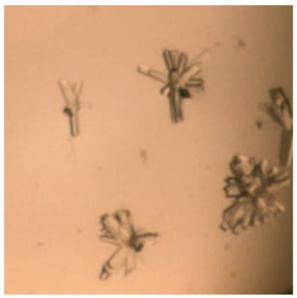
Task: Learning stage of protein crystallization



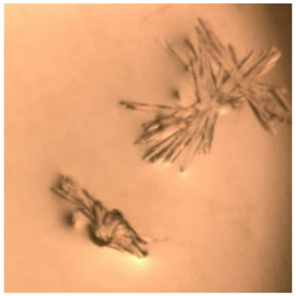
Crystal



Needle

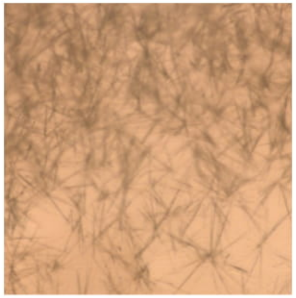


Tree



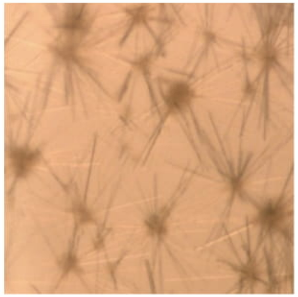
Tree

Empty



Needle

Experience

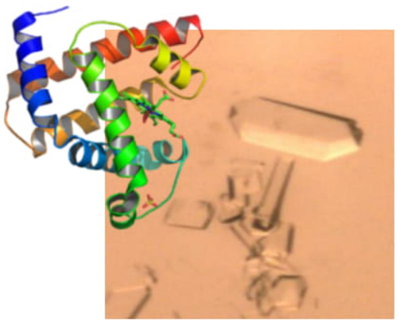


?

Performance

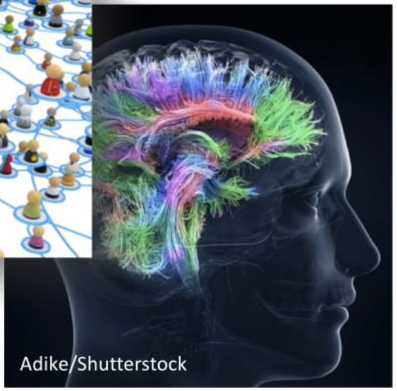
Why learn from data (experience)?

Understanding large-scale complex systems

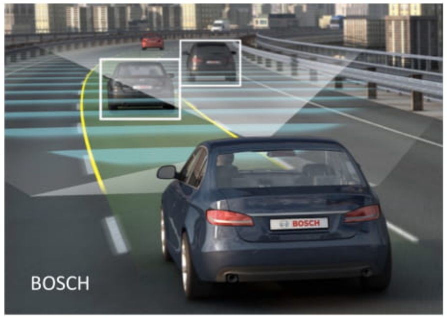


Bio-chemical molecules

Social networks

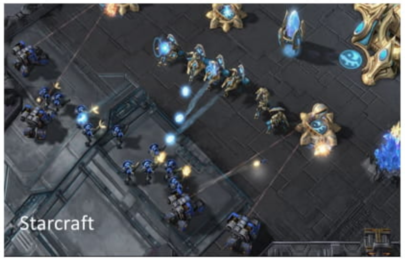


Brain



Self-driving vehicles

Cosmos

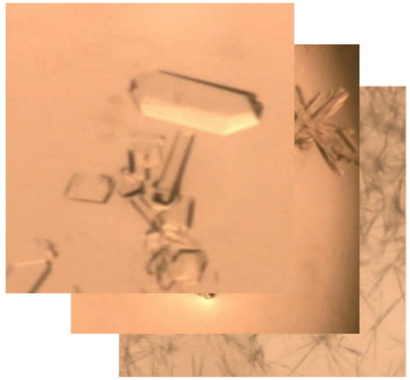


Games

rules and governing equations

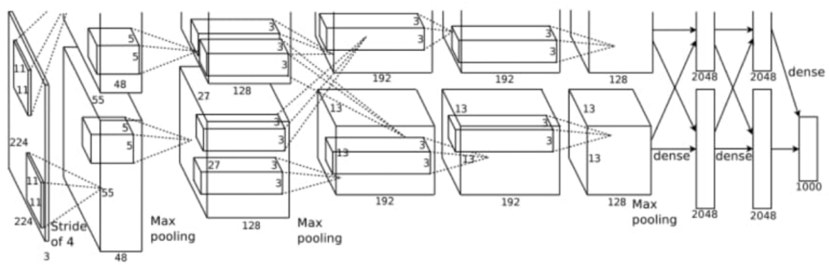
- are hard to discover
- involve too many variables
- are computationally too expensive
- are typically stochastic

How ML works



Crystal, Needle, Tree, …

- Model f: mapping between input and output
  linear, nonlinear, deep model

Input 

Output

- Algorithm: fits model to data
  $\underset{f}{\text{Optimize}}$ Performance(Model f, Data)

Why ML works

- Lots of data due to improved high-throughput technologies

- Improved machine learning algorithms

- Enhanced computing power

Why ML works

- Lots of data due to improved high-throughput technologies
  e.g. Social media and web data (Petabytes/min), Large Synoptic Survey Telescope (20 TB/night)

- Improved machine learning algorithms

- Enhanced computing power

Why ML works

- ➢ Lots of data due to improved high-throughput technologies
  e.g. Social media and web data (Petabytes/min), Large Synoptic Survey Telescope (20 TB/night)

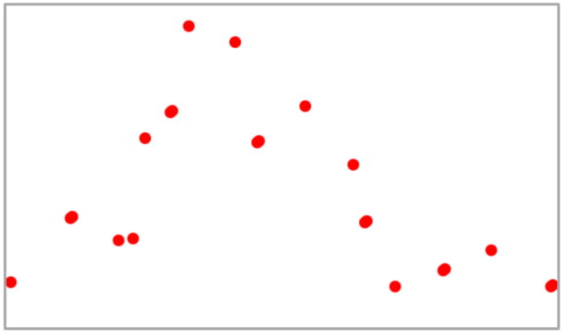
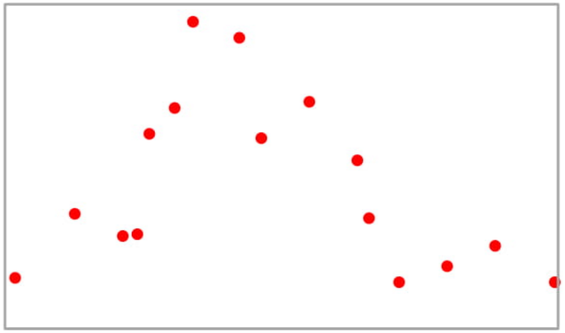
- ➢ Improved machine learning algorithms
- Rich models (high approximation power)
- Generalize well to unseen data

- ➢ Enhanced computing power

Why ML works

- ➢ Lots of data due to improved high-throughput technologies
  e.g. Social media and web data (Petabytes/min), Large Synoptic Survey Telescope (20 TB/night)

- ➢ Improved machine learning algorithms
- Rich models (high approximation power)
- Generalize well to unseen data

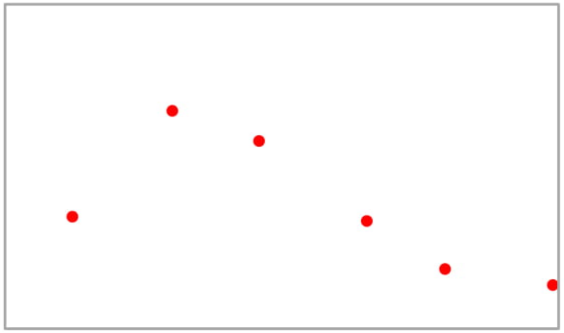



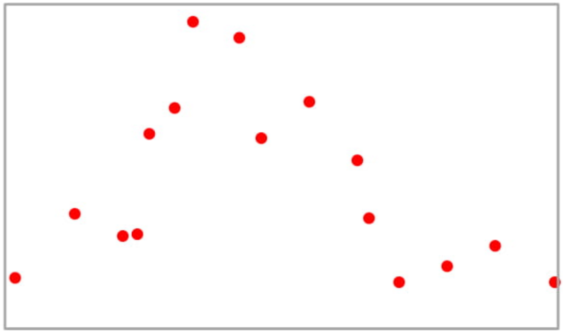
- ➢ Enhanced computing power

Why ML works

- Lots of data due to improved high-throughput technologies
  e.g. Social media and web data (Petabytes/min), Large Synoptic Survey Telescope (20 TB/night)

- Improved machine learning algorithms
- Rich models (high approximation power)
- Generalize well to unseen data



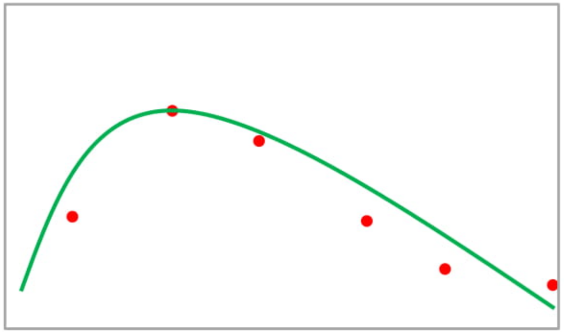


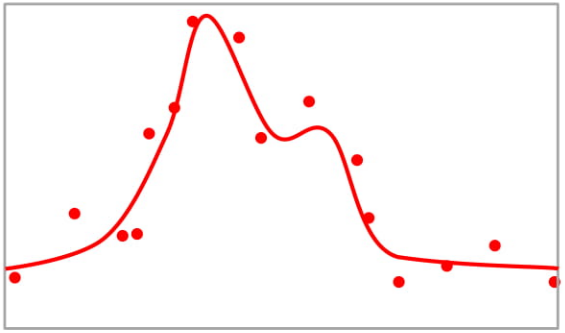
- Enhanced computing power

Why ML works

- ➢ Lots of data due to improved high-throughput technologies
  e.g. Social media and web data (Petabytes/min), Large Synoptic Survey Telescope (20 TB/night)

- ➢ Improved machine learning algorithms
- Rich models (high approximation power)
- Generalize well to unseen data



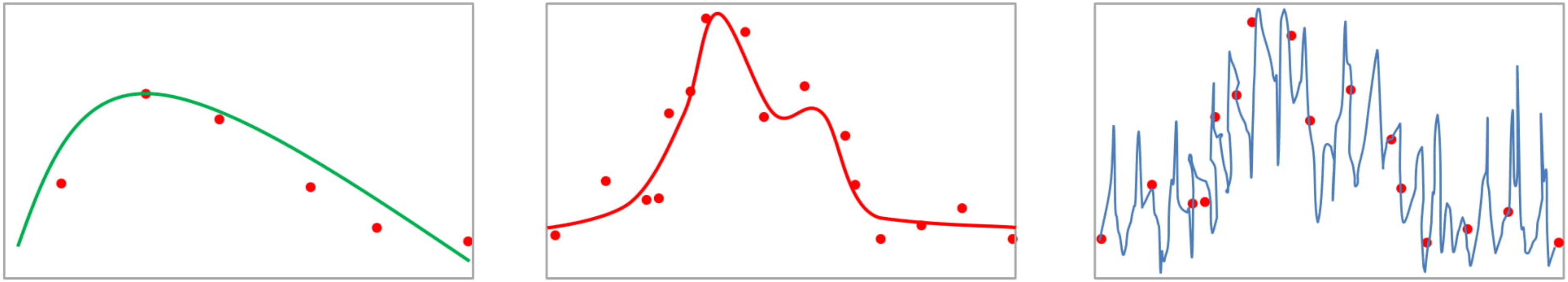


- ➢ Enhanced computing power

Why ML works

- ➢ Lots of data due to improved high-throughput technologies
  e.g. Social media and web data (Petabytes/min), Large Synoptic Survey Telescope (20 TB/night)

- ➢ Improved machine learning algorithms
- • Rich models (high approximation power)
- • Generalize well to unseen data

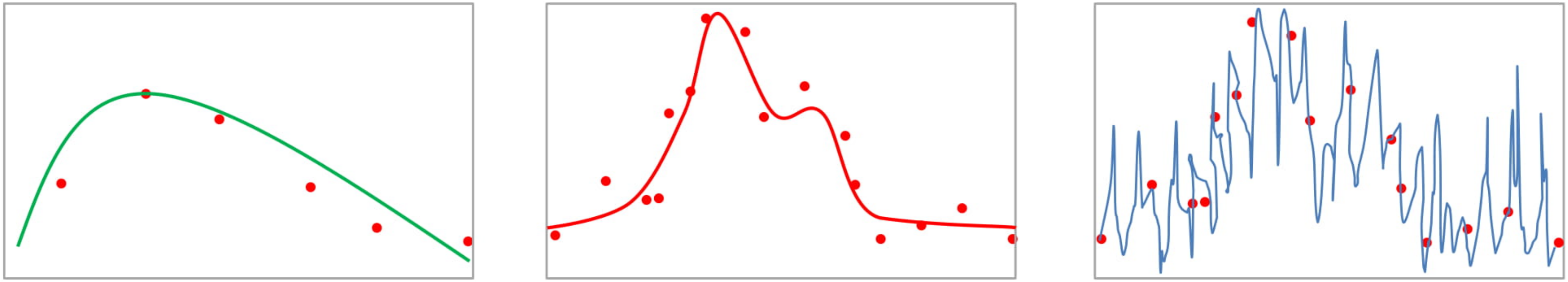


- ➢ Enhanced computing power

Why ML works

- Lots of data due to improved high-throughput technologies
  e.g. Social media and web data (Petabytes/min), Large Synoptic Survey Telescope (20 TB/night)

- Improved machine learning algorithms
- Rich models (high approximation power)
- Generalize well to unseen data

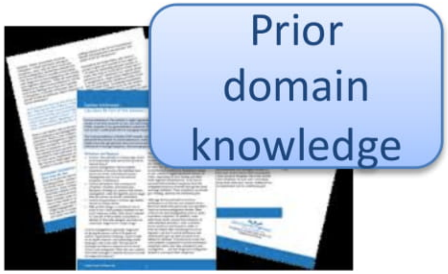


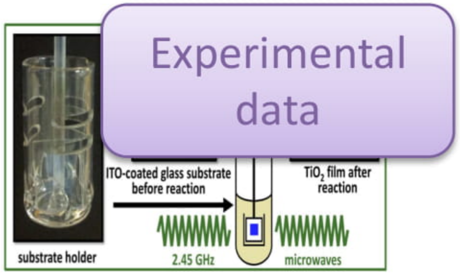
- Enhanced computing power (advanced GPUs, cloud platforms)
  e.g. accelerated training from 6 days to 18 mins in 5 years

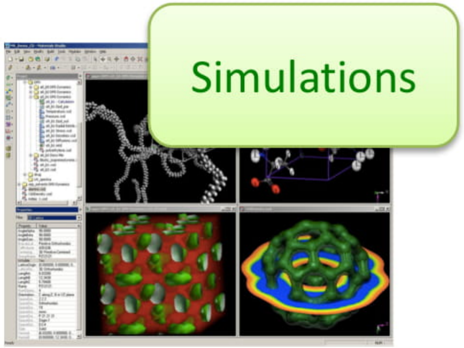
Data Challenges

Heterogeneous types of data

- Multi-modal
- Direct vs indirect
- Missing, incorrect
- Biased





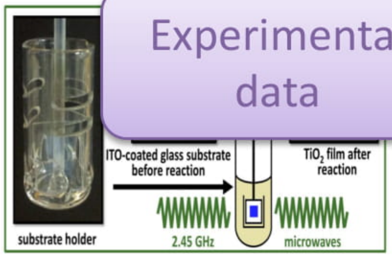


Data Challenges

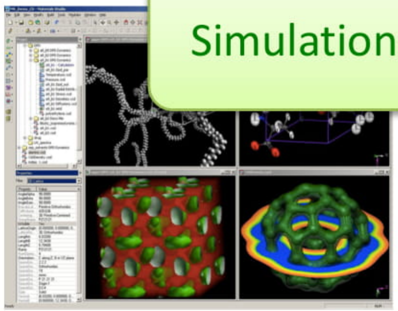
Heterogeneous types of data

- Multi-modal
- Direct vs indirect
- Missing, incorrect
- Biased

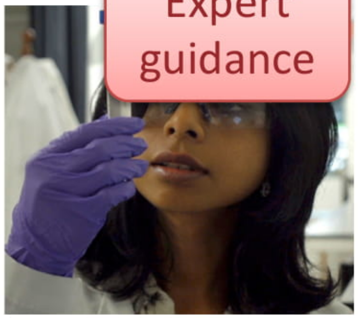
Prior domain knowledge



Experimental data



Simulations



Expert guidance

Handle unseen data from related domain

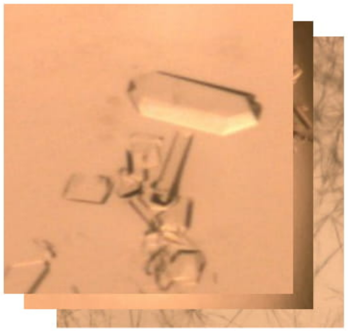
Self-driving car trained in Chicago vs Pittsburgh

Chicago

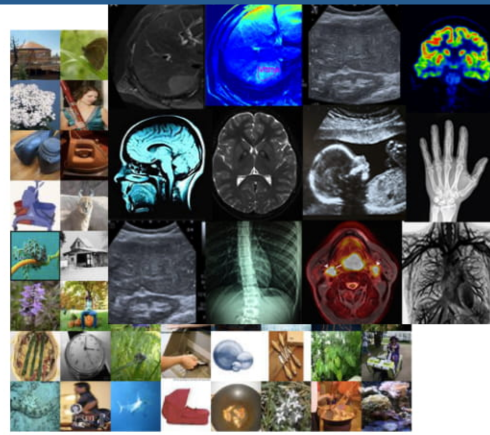
Pittsburgh

ML tasks: ubiquitous across domains

Prediction



Stage of crystal formation



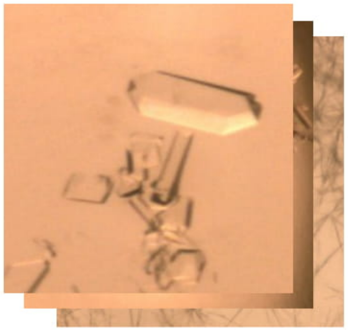
Object category, Medical diagnosis

Spam/Fraud identification

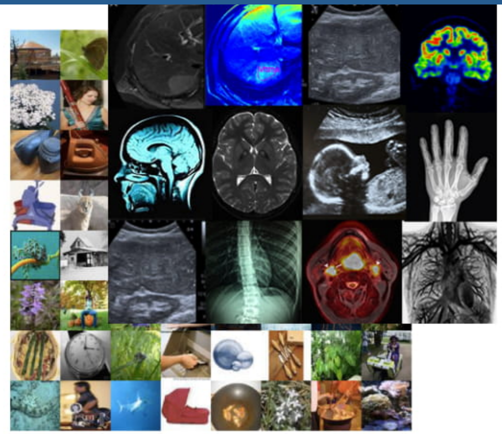
Stock price, weather forecasting

ML tasks: ubiquitous across domains

Prediction

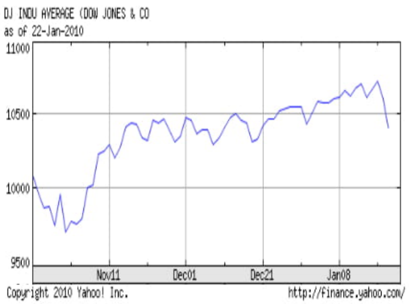


Stage of crystal formation



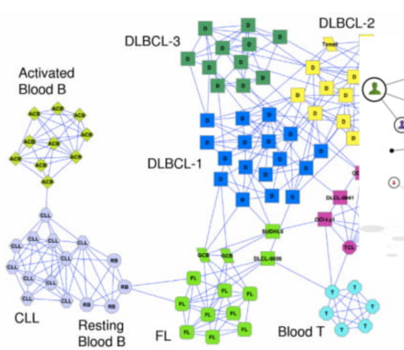
Object category, Medical diagnosis

Spam/Fraud identification



Stock price, weather forecasting

Unsupervised learning

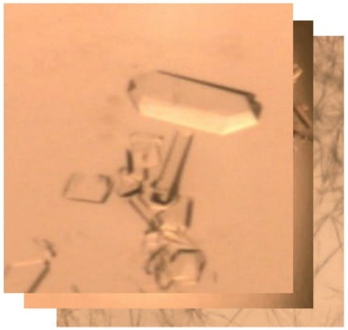


Grouping genes, commodities, …

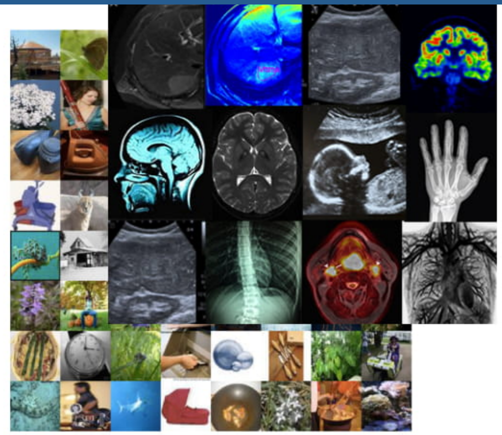
Relations between people, proteins, …

ML tasks: ubiquitous across domains

Prediction

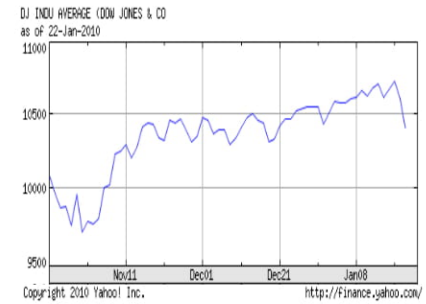


Stage of crystal formation



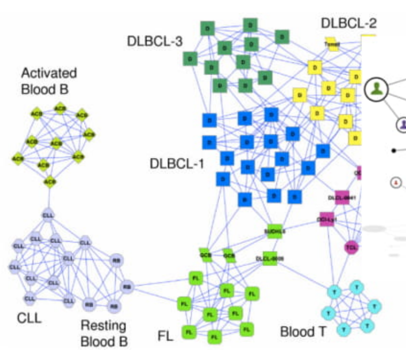
Object category, Medical diagnosis

Spam/Fraud identification



Stock price, weather forecasting

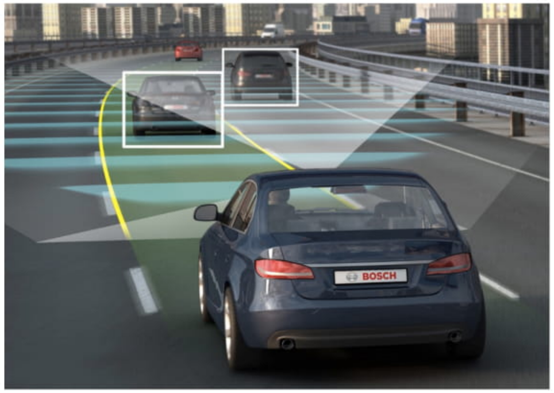
Unsupervised learning



Grouping genes, commodities, …

Relations between people, proteins, …

Decision making



Automated Navigation

Games

ML Task Challenges

+ Input-Output mapping tasks with given representations, lots of data and clearly defined performance metric

ML Task Challenges

+ Input-Output mapping tasks with given representations, lots of data and clearly defined performance metric

- Higher level tasks beyond input-output mapping (e.g. learn representations; guide data collection; design, test and refine hypothesis; interact with humans and environment)

- Multiple heterogeneous tasks

- High-stake decision making with very little tolerance for errors (e.g. criminal justice, medical decisions, etc.)

ML Performance Challenges

Current focus:

Accuracy/error — ImageNet error: 30% to 3% (since 2010); Google speech recognition: 8.4% to 4.9% (since 2016)

runtime, memory, …

ML Performance Challenges

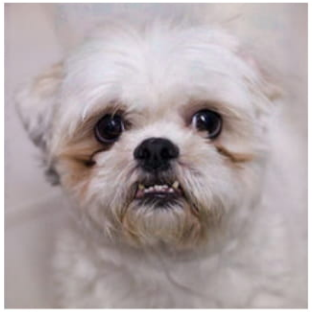
Current focus:

Accuracy/error

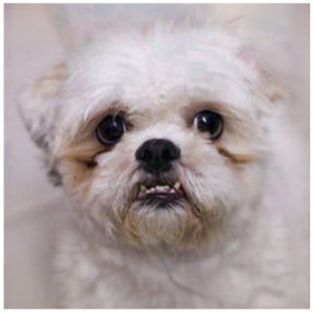
runtime, memory, …

ImageNet error: 30% to 3% (since 2010)
Google speech recognition: 8.4% to 4.9% (since 2016)

Robustness [Szegedy et al'14]



dog



ostrich

ML Performance Challenges

Current focus:

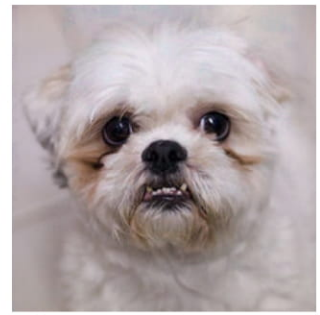
Accuracy/error

ImageNet error: 30% to 3% (since 2010)

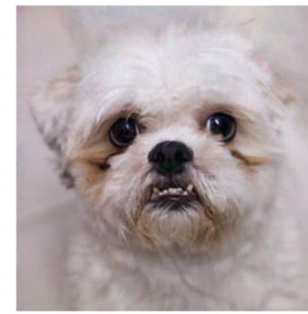
Google speech recognition: 8.4% to 4.9% (since 2016)

runtime, memory, …

Robustness [Szegedy et al'14]



dog



ostrich

Interpretability and Transparency

Trust and Accountability

ML Performance Challenges

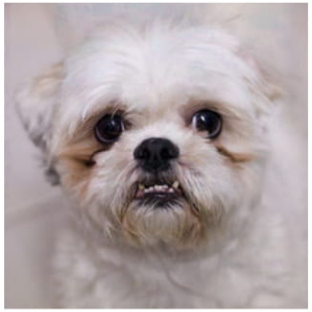
Current focus:

Accuracy/error

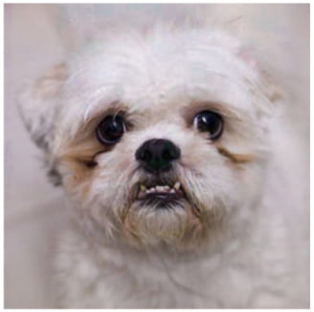
ImageNet error: 30% to 3% (since 2010)
Google speech recognition: 8.4% to 4.9% (since 2016)

runtime, memory, …

Robustness [Szegedy et al'14]



dog



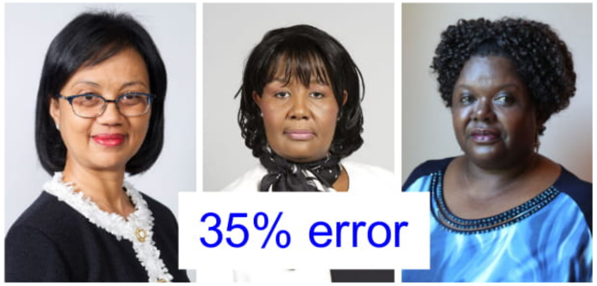
ostrich

Interpretability and Transparency

Trust and Accountability

Fairness and Ethics

[Buolamwini-Gebru'18]



MATHEMATICAL FRONTIERS

Why Machine Learning Works

Anne T. and Robert M. Bass Professor of Humanities and Sciences
Professor of Statistics

What Makes Machine Learning Work?

David Donoho,
Stanford University

Outline

Themes

In a longer talk, I would situate the current moment as follows:

(a) Smartphone Revolution

(b) Computing Disocntinuity

(c) Empirical Science Revolution

(d) Deepnet emergence

(c) Role for Math

Themes

For reasons of time, I emphasize **only**

(a) Smartphone Revolution

(b) Computing Discontinuity

(c) **Empirical Science Revolution**

(d) Deepnet emergence

(c) **Role for Math**

Common Task Framework (1980's)

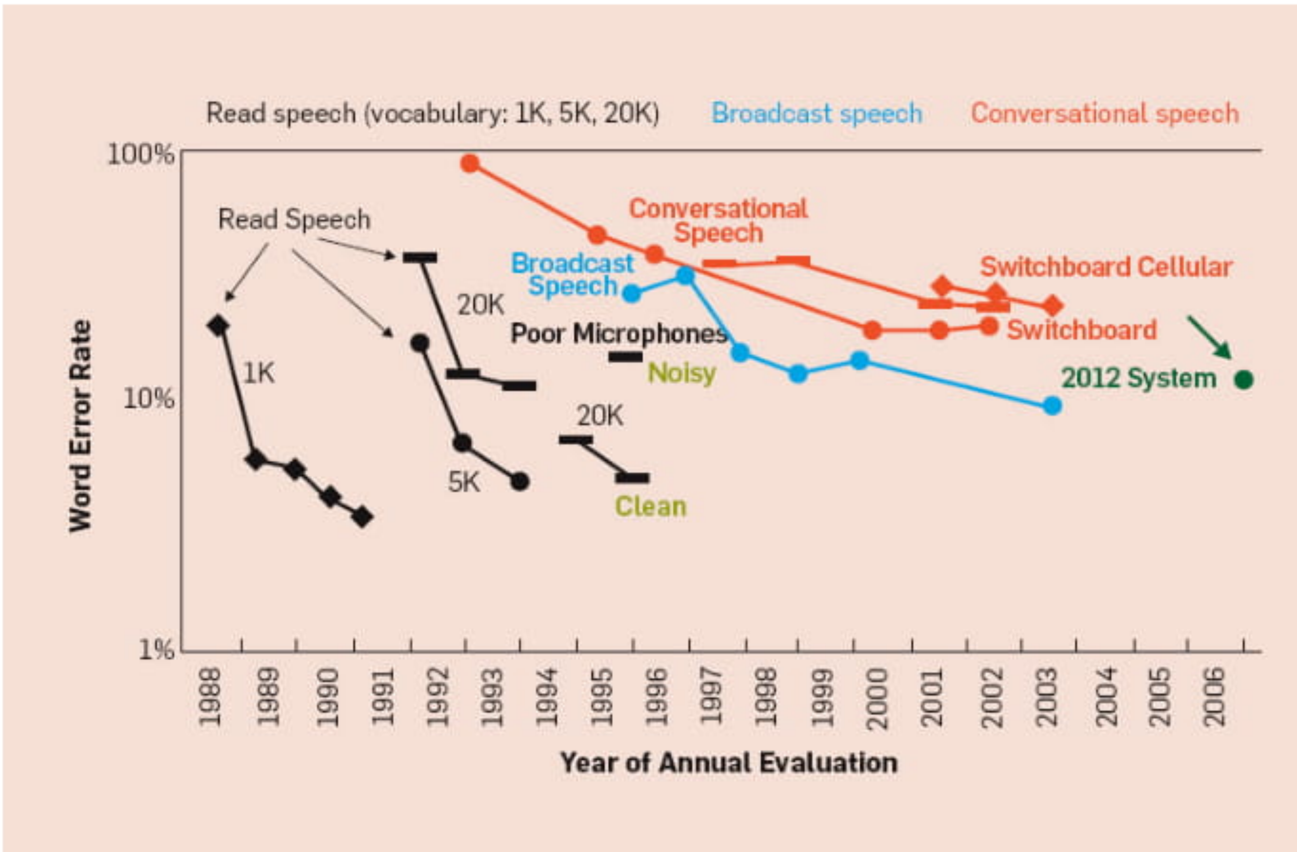
Under CTF we have the following ingredients

(a) A **publicly available training dataset** involving, for each observation, a list of (possibly many) feature measurements, and a class label for that observation.

(b) A set of **enrolled competitors** whose **common task** is to **infer** a class **prediction rule from the training data**.

(c) A **scoring referee**, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset which is sequestered behind a Chinese wall. The referee objectively and automatically reports the score achieved by the submitted rule.

See Mark Liberman's description (Liberman, 2009).


Read speech (vocabulary: 1K, 5K, 20K)
Broadcast speech
Conversational speech
100%
10%
1%
Word Error Rate
Read Speech
1K
20K
5K
Conversational Speech
Broadcast Speech
Poor Microphones
Noisy
20K
Clean
Switchboard Cellular
Switchboard
2012 System
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
Year of Annual Evaluation

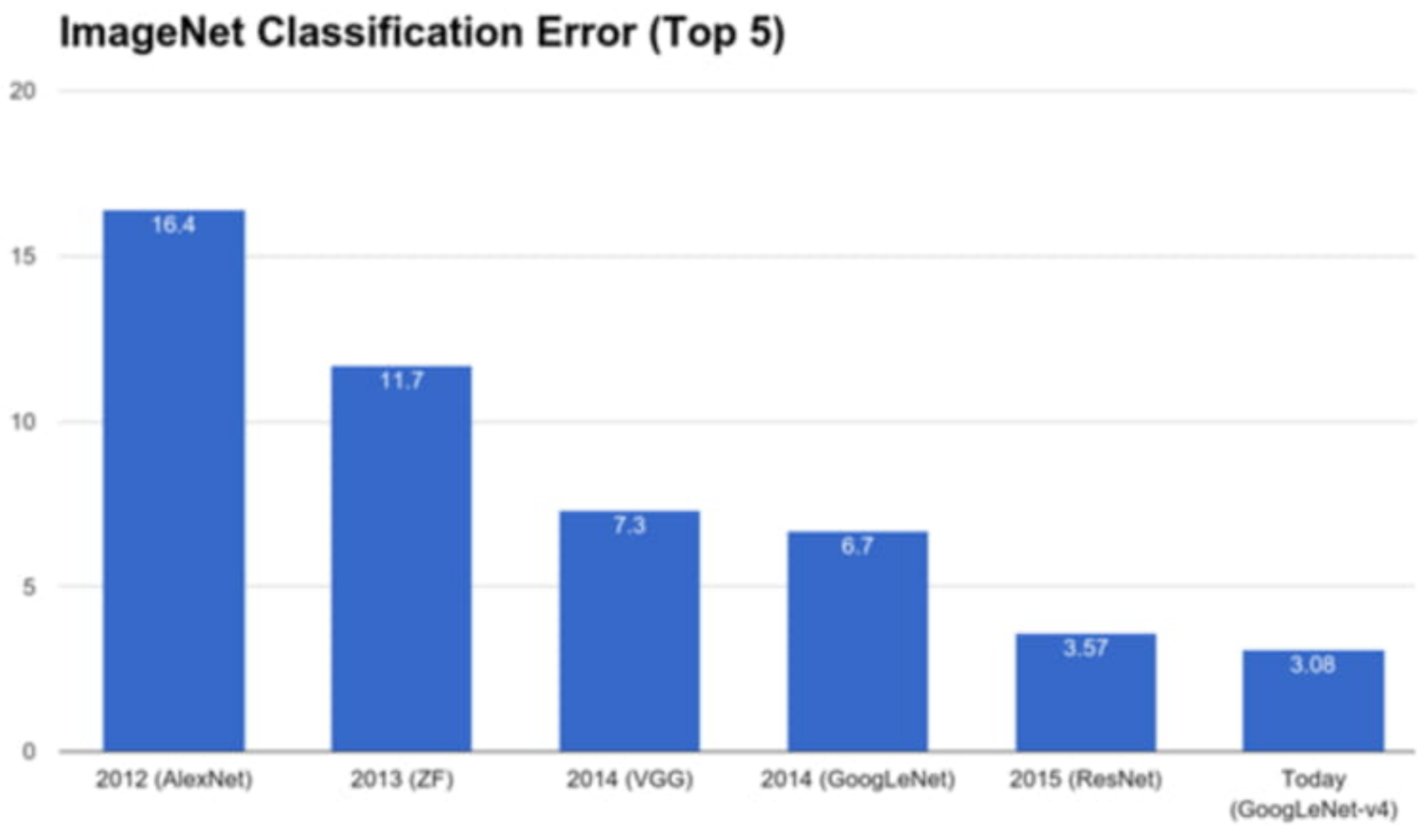
Emergence of Deep Learning Research

(a) The success of deep nets is an *entirely* empirical success.
All basic ideas were around for 30 years
Nothing beyond high school required

(b) Deep learning is a new *laboratory science*

Lab Science Term		Deep Learning Term
Laboratory	$\leftrightarrow$	compute cluster Software Stack
Lab Equipment	$\leftrightarrow$	Elasticluster/ClusterJob TensorFlow/Pytorch
Testube/Culture	$\leftrightarrow$	train/test deepnet
Experiment	$\leftrightarrow$	modify architecture modify dataset modify training algorithm
High Throughput	$\leftrightarrow$	Run Hyperparameter Grid

(c) Today **1000's PhD researchers** developing/studying deepnets **fulltime**
Factoid: Google has hired $\approx$ 1500 PhD researchers over 5 years.
$\approx$ *all CS faculty in USA*!
Major commitment to deep learning
Major effects on scholarship, conferences, *younger generation*


ImageNet Classification Error (Top 5)
20
15
10
5
0
16.4
11.7
7.3
6.7
3.57
3.08
2012 (AlexNet)
2013 (ZF)
2014 (VGG)
2014 (GoogLeNet)
2015 (ResNet)
Today (GoogLeNet-v4)

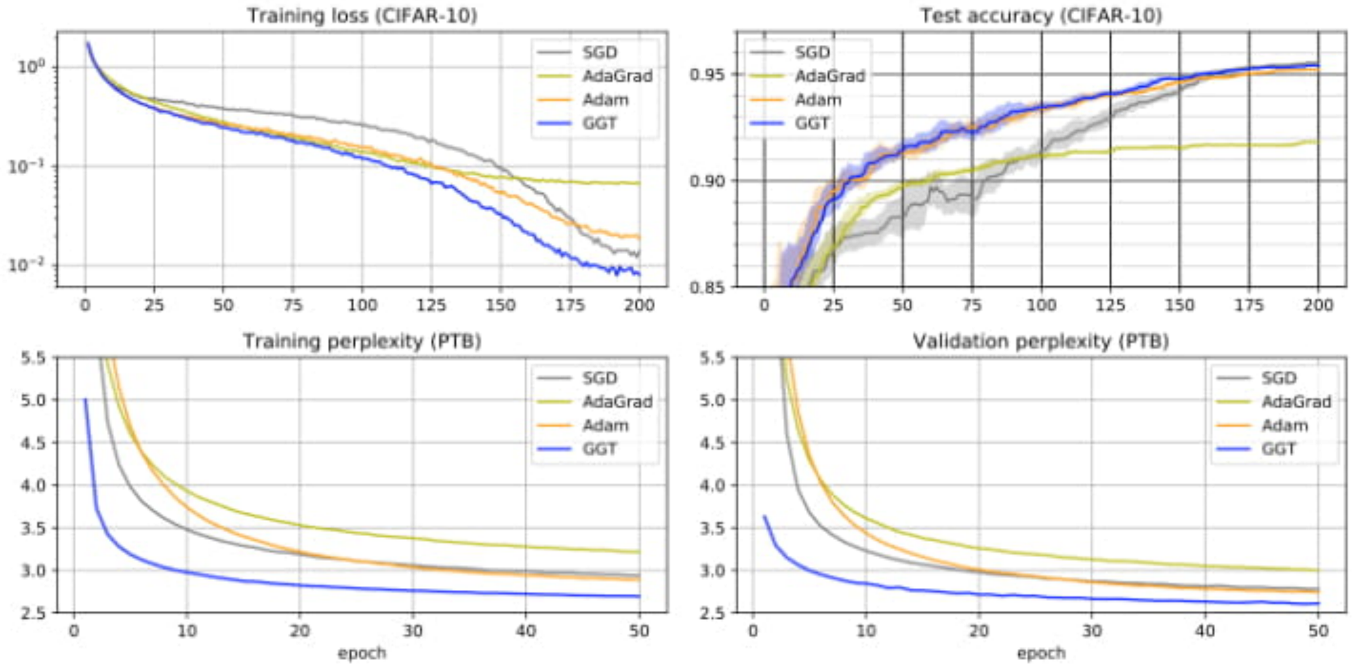


Figure 3: Results of CNN and RNN experiments. GGT dominates in training loss across both tasks, and generalizes better on the RNN task. *Top:* CIFAR-10 classification with a 3-branch ResNet. *Bottom:* PTB character-level language modeling with a 3-layer LSTM.

Sebastiao Salgado, *Work*

Speed up Training

- A 6-page conference paper may burn $> \$100K$ (retail) computer time.
- State of the Art Deepnet training *extremely slow*:
  Stochastic Gradient Descent
- State of the Art hyperparameter search *extremely slow*:
  Exhaustive evaluation
- Traditional mathematical sciences attacked both problems
  - Second-order methods (Newton's Method and successors) much better than First-order
  - Experimental design much better than exhaustive evaluation
- Adapt/Extend traditionally successful optimization ideas in mathematical sciences to Deepnet setting
  Save \$100's M in research costs annually. *Forever.*

Improve Learning

- State of the Art results often use gigantic datasets (e.g. Laurens van der Maaten, FB, 700M images).
- Hopes for perfection: driving force for even larger data
- Scaling relation of errors vs. dataset size *very unfavorable*
- Training practices *very doutbful* (train to zero error).
- Traditional mathematical sciences attacked both problems
  - Most accurate estimates possible for a given sample size (RA Fisher etc.)
  - Regularization to defeat curse of dimensionality (Tikhonov Regularization, Stein Shrinkage, Lasso etc.)
- Adapt/Extend traditionally successful estimation ideas in mathematical sciences to Deepnet setting
  Deepnets achieve current performance specs at much smaller dataset size N

Improve Embeddings

- State of the Art results often use special embeddings to make Deepnets applicable.
  - Word2Vec (Glove, etc)
  - TSNE
- Successful but poorly understood.
  Possibly can be much improved
- Traditional mathematical sciences attacked embeddings, but without invariances:
  - PCA
  - ISOMAP
  - LLE
- Recent mathematical sciences attacked embeddings, *with invariances*
  S. Mallat, Scattering Networks
- Adapt/Extend traditionally successful embedding ideas in mathematical sciences to Deepnet setting
  Deepnets applicable to many other problems.

Historic Challenge to the Mathematical Sciences

- Ingrid Daubechies' dictum
  When a **mathematical** *object has interesting behavior, there's a* **mathematical** *reason.*
- Great deal of historical success
- But does it continue to work here?
  - Deepnets involve *mathematically-definable* entities
  - Superhuman performance is *interesting*
- Daubechies' dictum seems to apply.
- Encounter with Ian Goodfellow suggests difficulties with mathematical mindset:
  - Must there be a reason?
  - Should we care about the reason?

Historic Challenge to the Mathematical Sciences, 2

If we care about 'understanding' and 'reasons' here are some challenges:

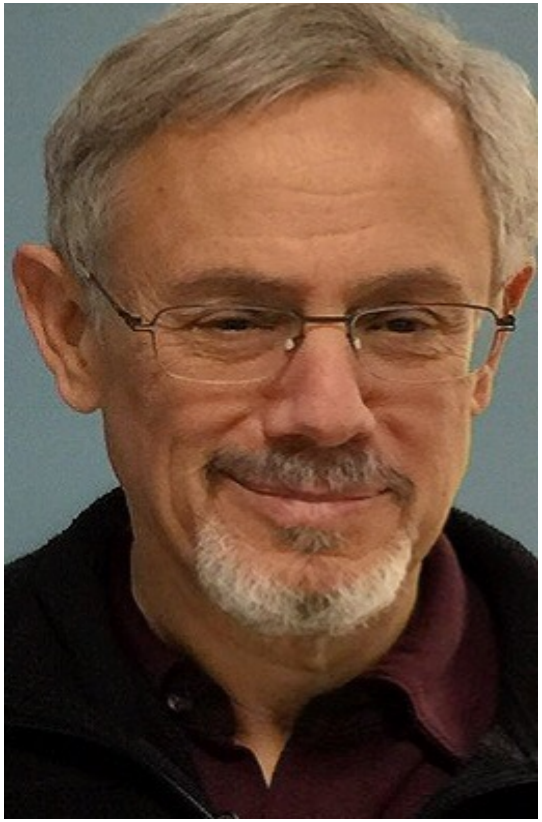
- Deepnets in practice are *high-dimensional interpolation scheme*.
  Almost nothing known about the classes of functions well approximated by *actual deepnets* using *actual training algorithms typical in practice*.
  Learning more can lead to better training and better nets.
- High-dimensional training uses high-dimensional Hessian and gradient.
  We have limited window on such objects, learning more enables speed ups optimization.

MATHEMATICAL FRONTIERS

Why Machine Learning Works

Aarti Singh,
Carnegie Mellon University

David Donoho,
Stanford University



Mark Green,
UCLA (moderator)

View webinar videos and learn more about BMSA at *www.nas.edu/MathFrontiers*

MATHEMATICAL FRONTIERS
2018 Monthly Webinar Series, 2-3pm ET

February 13*:
Mathematics of the Electric Grid

March 13*:
Probability for People and Places

April 10*:
Social and Biological Networks

May 8*:
Mathematics of Redistricting

June 12*: *Number Theory: The Riemann Hypothesis*

July 10*: *Topology*

August 14*: *Algorithms for Threat Detection*

September 11*: *Mathematical Analysis*

October 9*: *Combinatorics*

November 13:
Why Machine Learning Works

December 11:
Mathematics of Epidemics

***** ***Recording posted***

Made possible by support for BMSA from the National Science Foundation Division of Mathematical Sciences and the Department of Energy Advanced Scientific Computing Research

View webinar videos and learn more about BMSA at www.nas.edu/MathFrontiers