Practical Implications of Theoretical Deep Learning

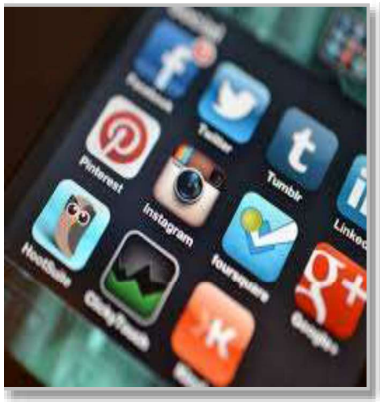
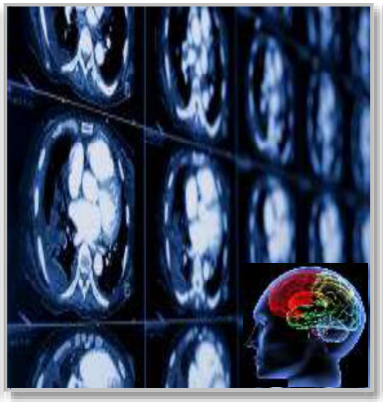
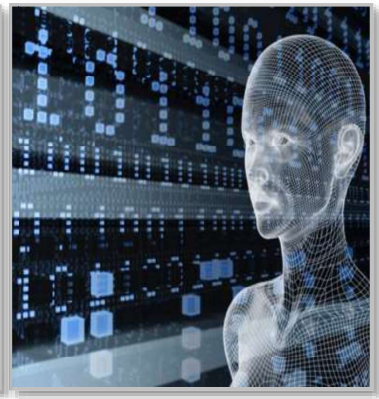
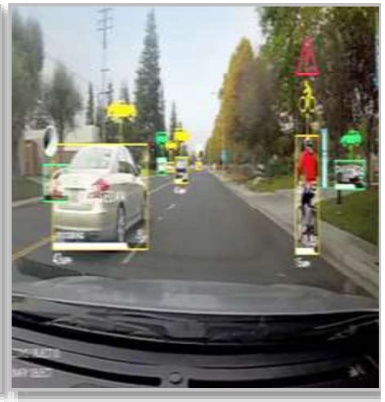
Nadav Cohen

Israel Machine Vision Conference (IMVC) 2020

Outline

Deep Learning

EVERY INDUSTRY WANTS DEEP LEARNING

Cloud Service Provider	Medicine	Media & Entertainment	Security & Defense	Autonomous Machines
				
➢ Image/Video classification ➢ Speech recognition ➢ Natural language processing	➢ Cancer cell detection ➢ Diabetic grading ➢ Drug discovery	➢ Video captioning ➢ Content based search ➢ Real time translation	➢ Face recognition ➢ Video surveillance ➢ Cyber security	➢ Pedestrian detection ➢ Lane tracking ➢ Recognize traffic sign

NVIDIA

Source

NVIDIA (`www.slideshare.net/openomics/the-revolution-of-deep-learning`)

Beats World Champion at Go

Converses With Humans

Drives Cars

Controls Manufacturing Plants

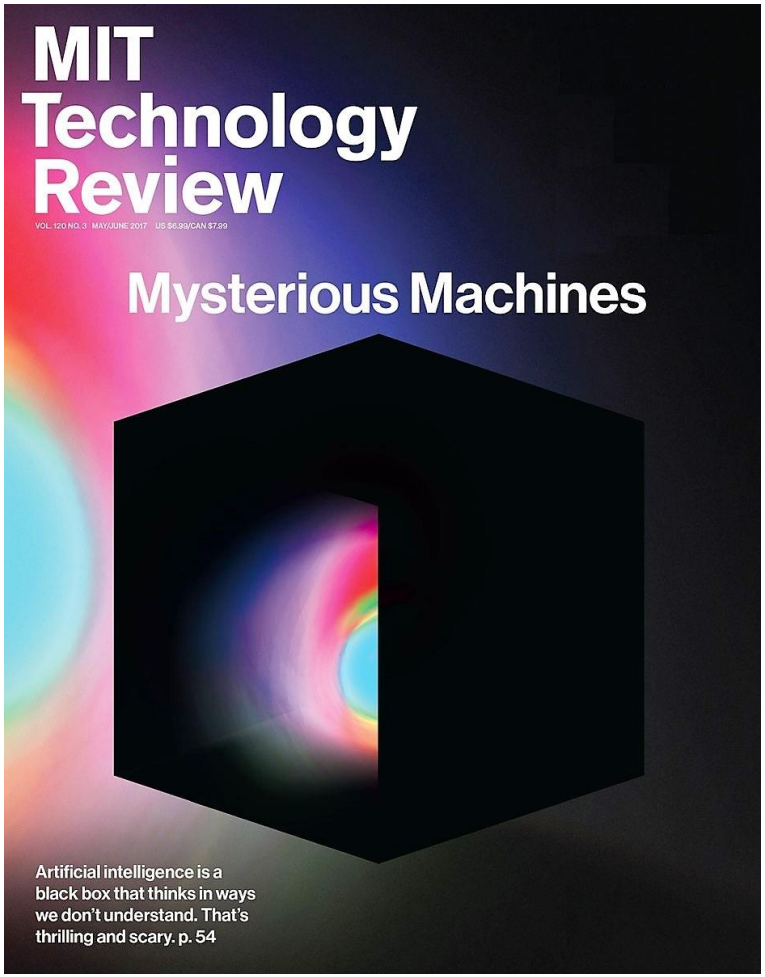
IMUBIT

Solving the industry's hardest problems

At Imubit, we're driven by a mission to tackle and solve the toughest challenges at chemical plants and refineries. We help hydrocarbon

"We make more money on days we run Imubit than days we don't. We now have 24×7 optimized units."

But Not Well Understood



Intelligent Machines

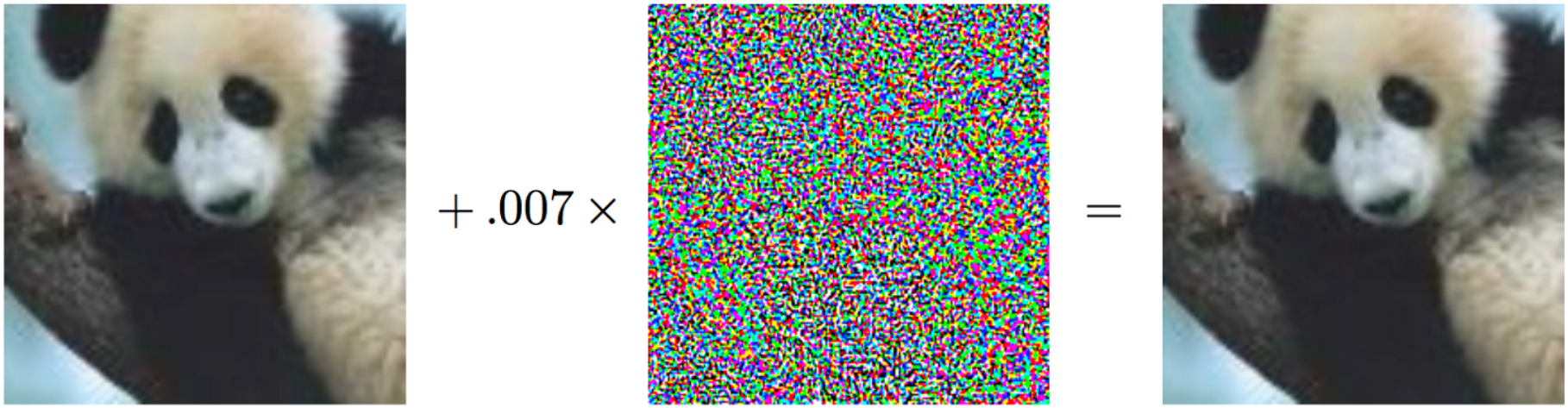
The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

by Will Knight April 11, 2017

Last year, a strange self-driving car was released onto the quiet roads of Monmouth County, New Jersey. The experimental vehicle, developed by researchers at the chip maker Nvidia, didn't look different from other autonomous cars, but it was unlike anything demonstrated by Google, Tesla, or General Motors, and it showed the rising power of artificial intelligence. The car didn't follow a single instruction provided by an engineer or programmer. Instead, it relied entirely on an algorithm that had taught itself to drive by watching a human do it.

Susceptible to Adversarial Attacks



"panda"

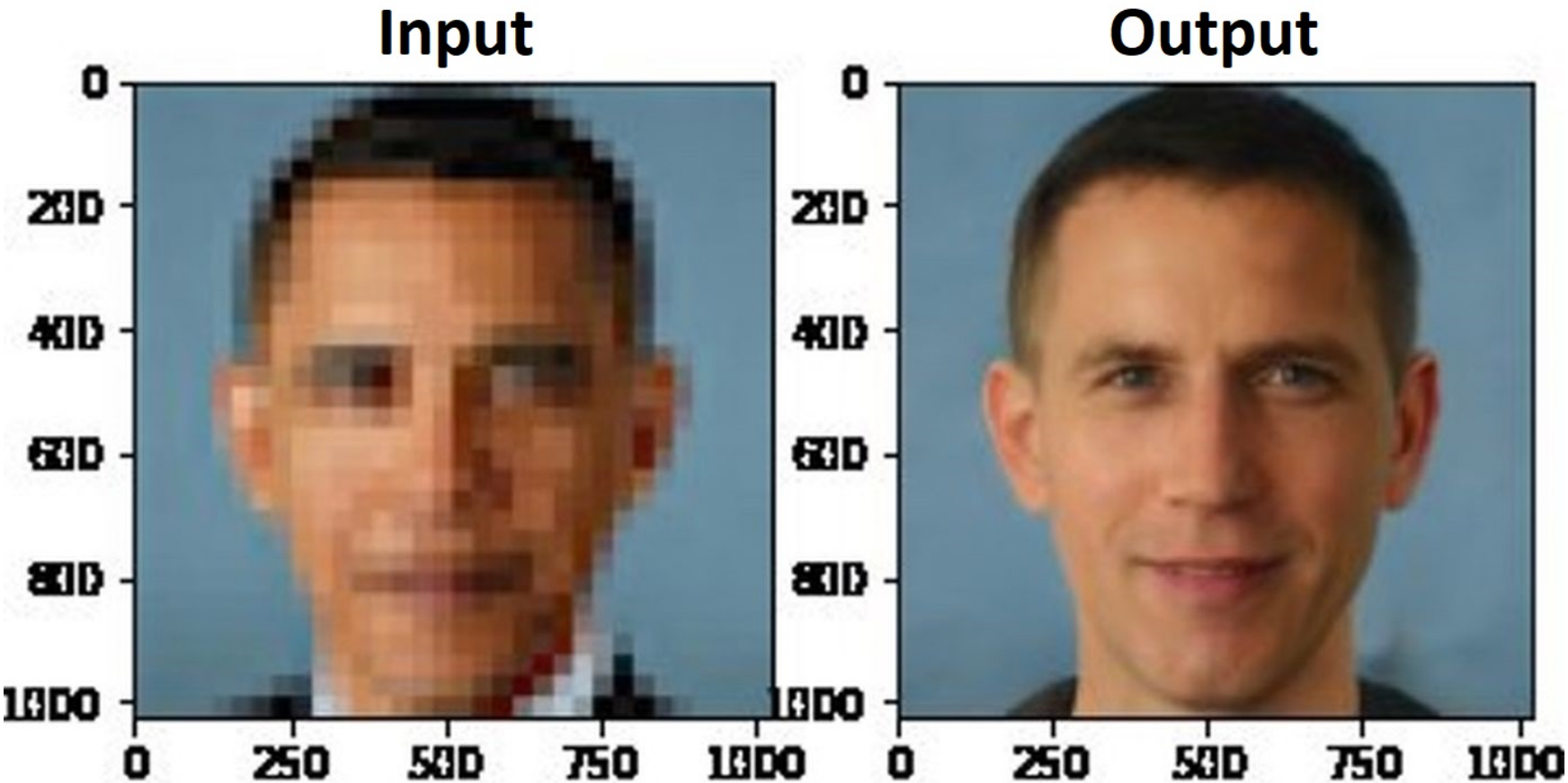
57.7% confidence

noise

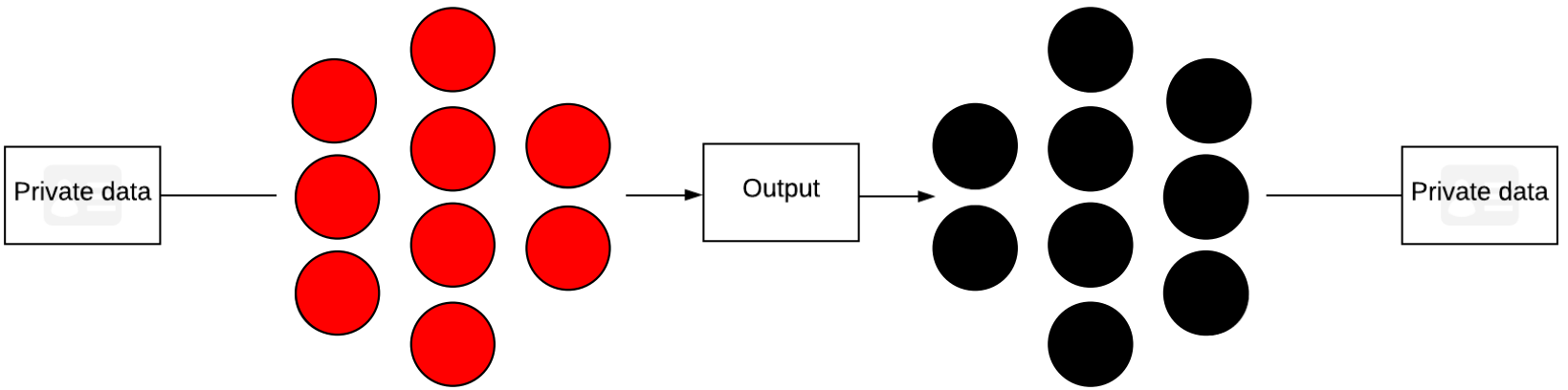
"gibbon"

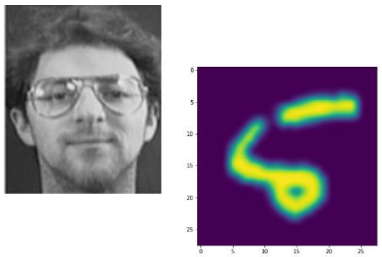
99.3% confidence

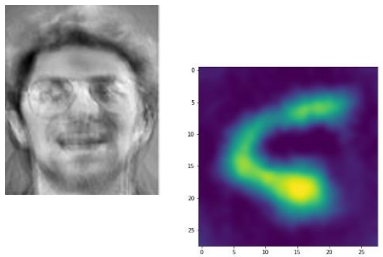
Exhibits Undesired Biases



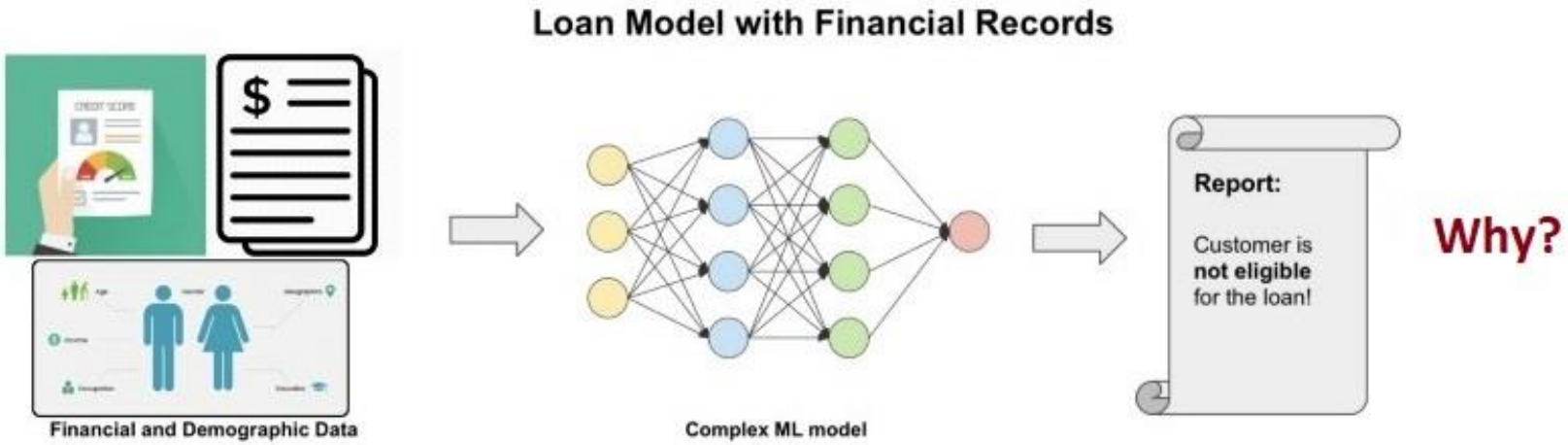
Leaks Private Information

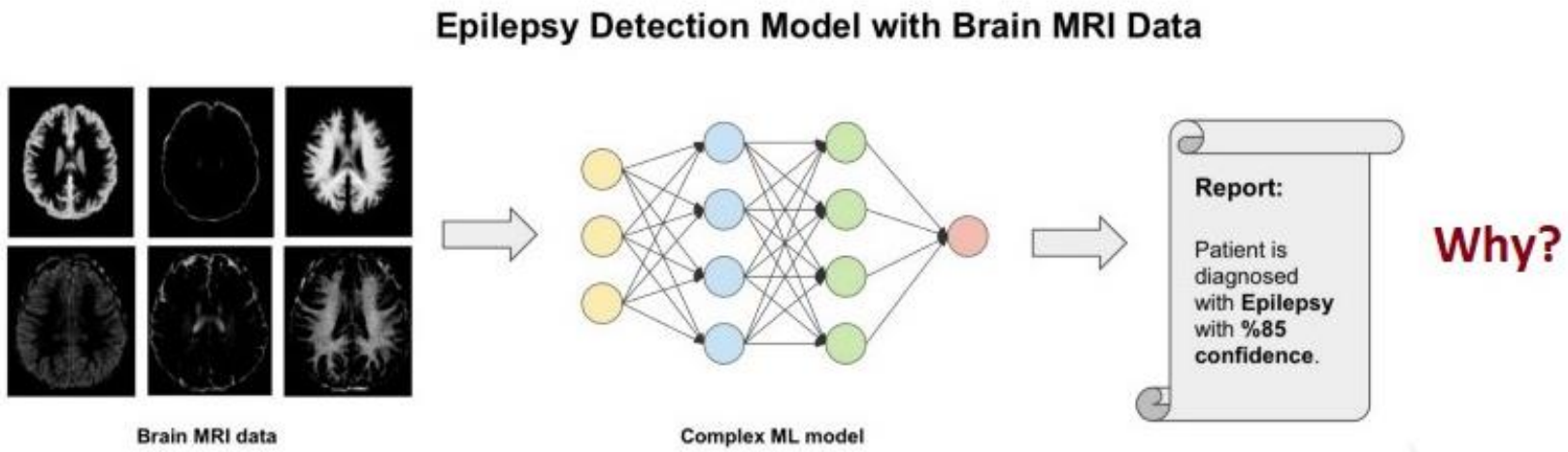




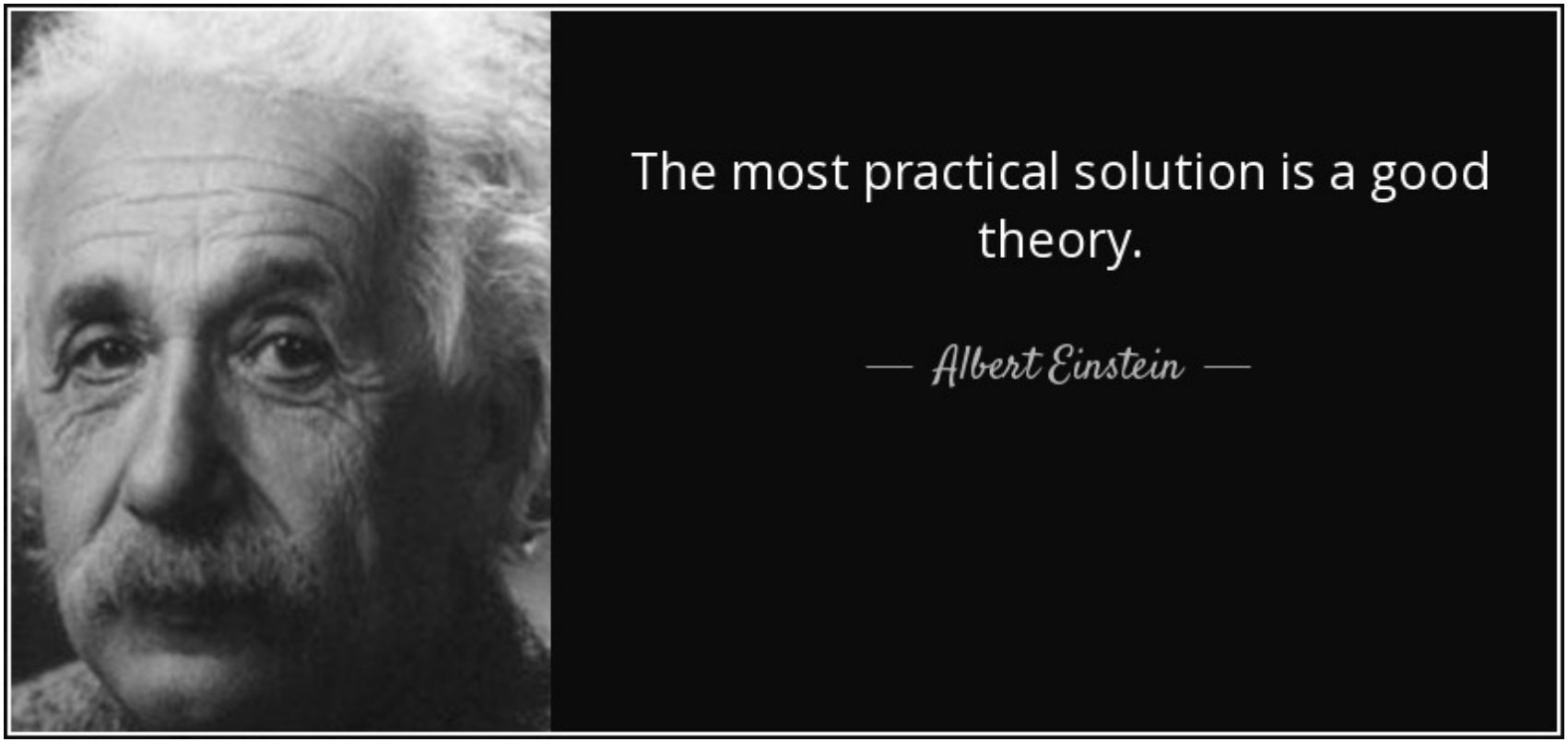


Lacks Interpretability





Theory May Help Alleviate These Shortcomings



Outline

Statistical Learning Setup

Statistical Learning Setup

$\mathcal{X}$ — instance space (e.g. $\mathbb{R}^{100\times 100}$ for 100-by-100 grayscale images)

Statistical Learning Setup

$\mathcal{X}$ — instance space (e.g. $\mathbb{R}^{100\times 100}$ for 100-by-100 grayscale images)

$\mathcal{Y}$ — label space (e.g. $\mathbb{R}$ for regression or $\{1, \ldots, k\}$ for classification)

Statistical Learning Setup

$\mathcal{X}$ — instance space (e.g. $\mathbb{R}^{100\times 100}$ for 100-by-100 grayscale images)

$\mathcal{Y}$ — label space (e.g. $\mathbb{R}$ for regression or $\{1, \ldots, k\}$ for classification)

$\mathcal{D}$ — distribution over $\mathcal{X} \times \mathcal{Y}$ (unknown)

Statistical Learning Setup

$\mathcal{X}$ — instance space (e.g. $\mathbb{R}^{100\times 100}$ for 100-by-100 grayscale images)

$\mathcal{Y}$ — label space (e.g. $\mathbb{R}$ for regression or $\{1, \ldots, k\}$ for classification)

$\mathcal{D}$ — distribution over $\mathcal{X} \times \mathcal{Y}$ (unknown)

$\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_{\geq 0}$ — loss func (e.g. $\ell(y, \hat{y}) = (y - \hat{y})^2$ for $\mathcal{Y} = \mathbb{R}$)

Statistical Learning Setup

$\mathcal{X}$ — instance space (e.g. $\mathbb{R}^{100\times 100}$ for 100-by-100 grayscale images)

$\mathcal{Y}$ — label space (e.g. $\mathbb{R}$ for regression or $\{1, \ldots, k\}$ for classification)

$\mathcal{D}$ — distribution over $\mathcal{X} \times \mathcal{Y}$ (unknown)

$\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_{\geq 0}$ — loss func (e.g. $\ell(y, \hat{y}) = (y - \hat{y})^2$ for $\mathcal{Y} = \mathbb{R}$)

Task

Given training set $S = \{(X_i, y_i)\}_{i=1}^m$ drawn i.i.d. from $\mathcal{D}$, return hypothesis (predictor) $h : \mathcal{X} \to \mathcal{Y}$ that minimizes population loss:

$$L_{\mathcal{D}}(h) := \mathbb{E}_{(X,y)\sim\mathcal{D}}[\ell(y, h(X))]$$

Statistical Learning Setup

$\mathcal{X}$ — instance space (e.g. $\mathbb{R}^{100\times100}$ for 100-by-100 grayscale images)

$\mathcal{Y}$ — label space (e.g. $\mathbb{R}$ for regression or $\{1,\ldots,k\}$ for classification)

$\mathcal{D}$ — distribution over $\mathcal{X}\times\mathcal{Y}$ (unknown)

$\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_{\geq 0}$ — loss func (e.g. $\ell(y,\hat{y}) = (y-\hat{y})^2$ for $\mathcal{Y}=\mathbb{R}$)

Task

Given training set $S = \{(X_i, y_i)\}_{i=1}^m$ drawn i.i.d. from $\mathcal{D}$, return hypothesis (predictor) $h : \mathcal{X} \to \mathcal{Y}$ that minimizes population loss:

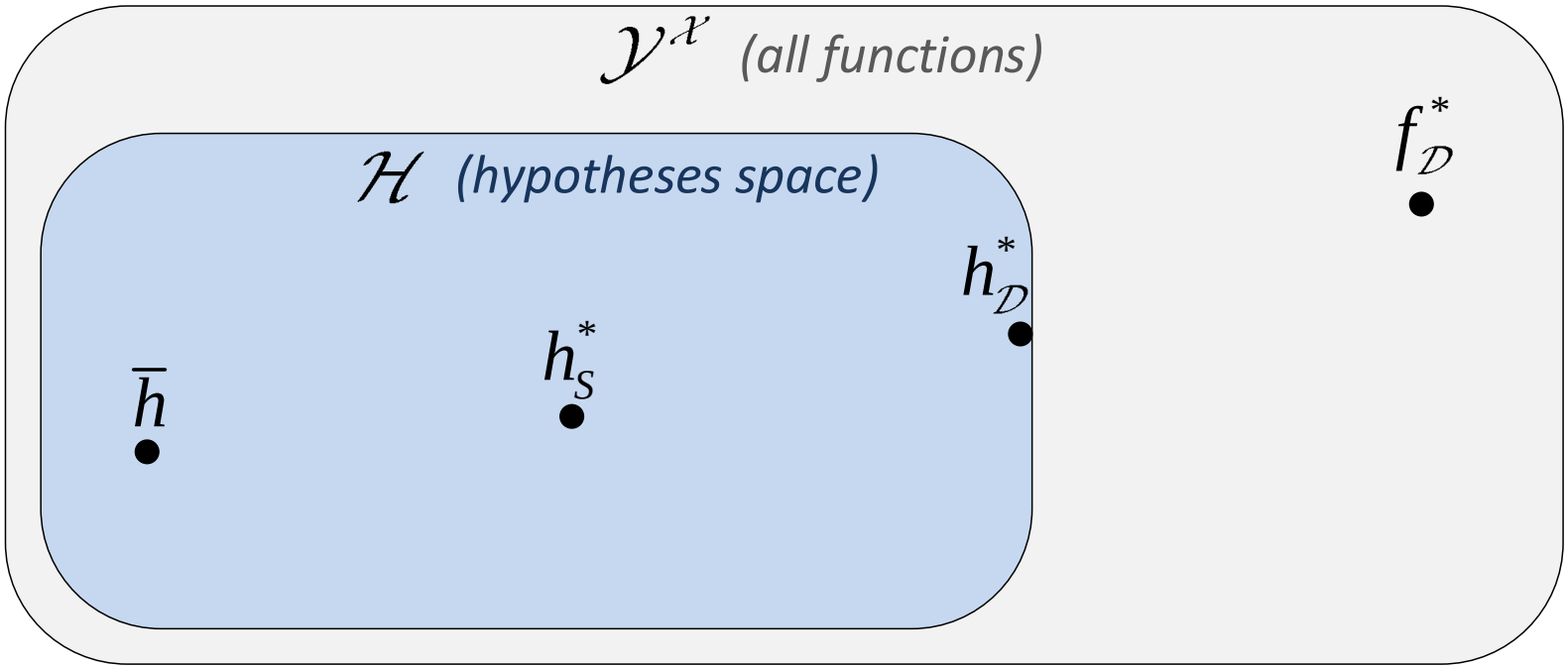
$$L_{\mathcal{D}}(h) := \mathbb{E}_{(X,y)\sim\mathcal{D}}[\ell(y, h(X))]$$

Approach

Predetermine hypotheses space $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, and return hypothesis $h \in \mathcal{H}$ that minimizes empirical loss:

$$L_S(h) := \frac{1}{m}\sum_{i=1}^{m} \ell(y_i, h(X_i))$$

Three Pillars of Statistical Learning Theory: Expressiveness, Generalization and Optimization



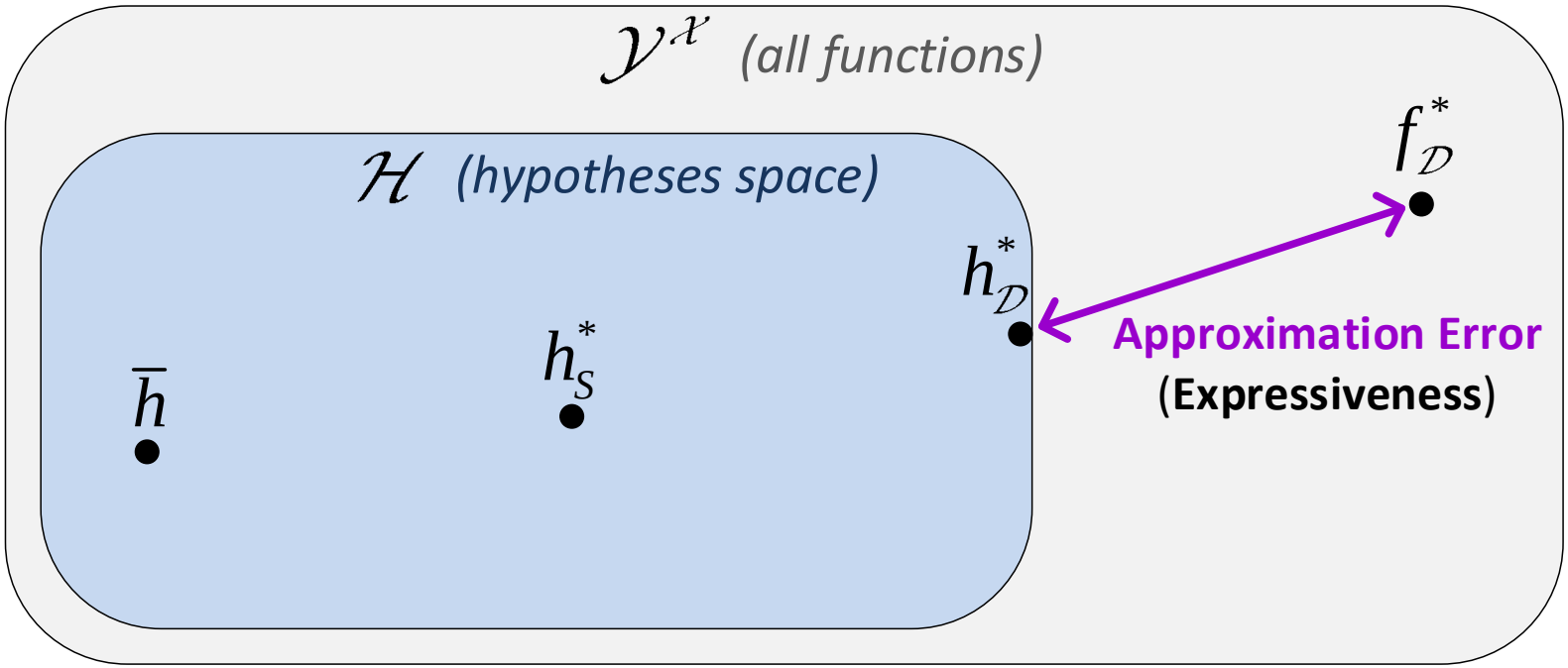
$f^*_{\mathcal{D}}$ — ground truth (minimizer of population loss over $\mathcal{Y}^{\mathcal{X}}$)

$h^*_{\mathcal{D}}$ — optimal hypothesis (minimizer of population loss over $\mathcal{H}$)

h^*_S — empirically optimal hypothesis (minimizer of empirical loss over $\mathcal{H}$)

$\bar{h}$ — returned hypothesis

Three Pillars of Statistical Learning Theory: Expressiveness, Generalization and Optimization



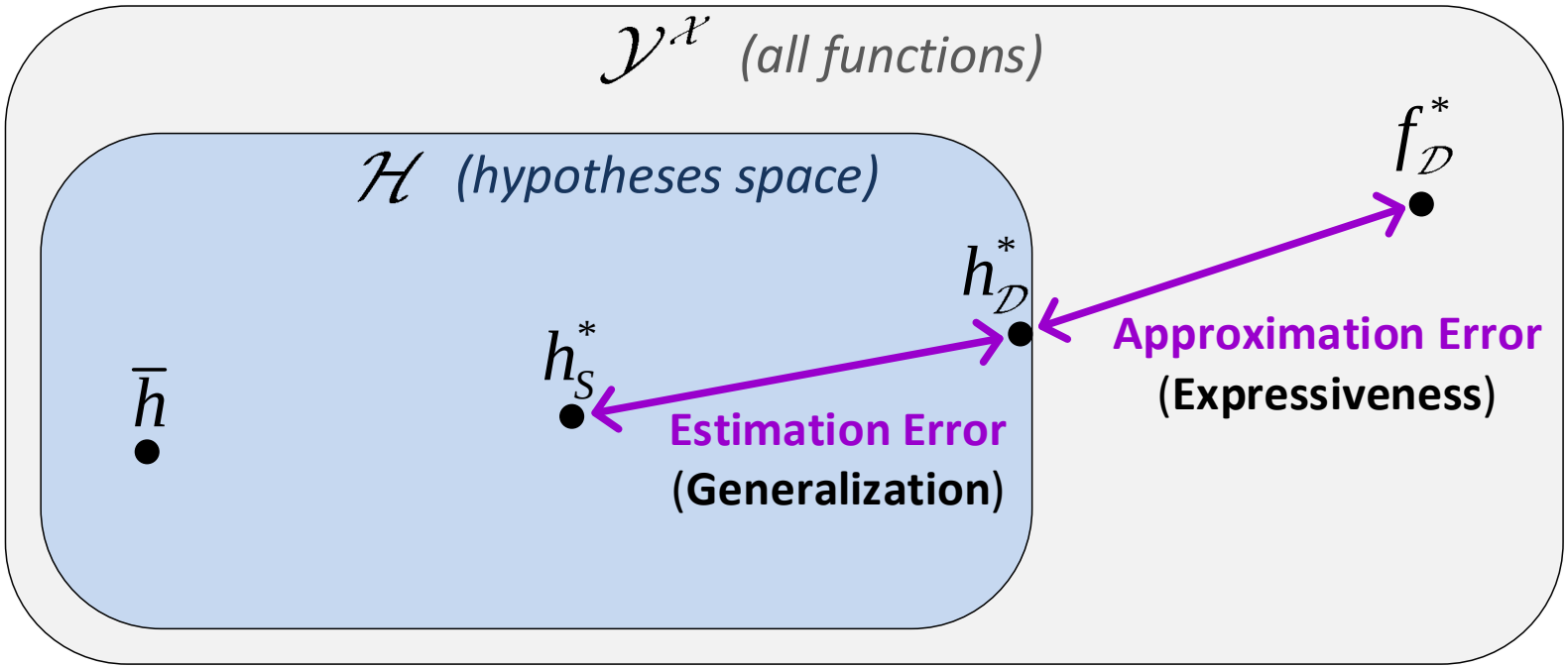
$f^*_{\mathcal{D}}$ — ground truth (minimizer of population loss over $\mathcal{Y}^{\mathcal{X}}$)

$h^*_{\mathcal{D}}$ — optimal hypothesis (minimizer of population loss over $\mathcal{H}$)

h^*_S — empirically optimal hypothesis (minimizer of empirical loss over $\mathcal{H}$)

$\bar{h}$ — returned hypothesis

Three Pillars of Statistical Learning Theory: Expressiveness, Generalization and Optimization



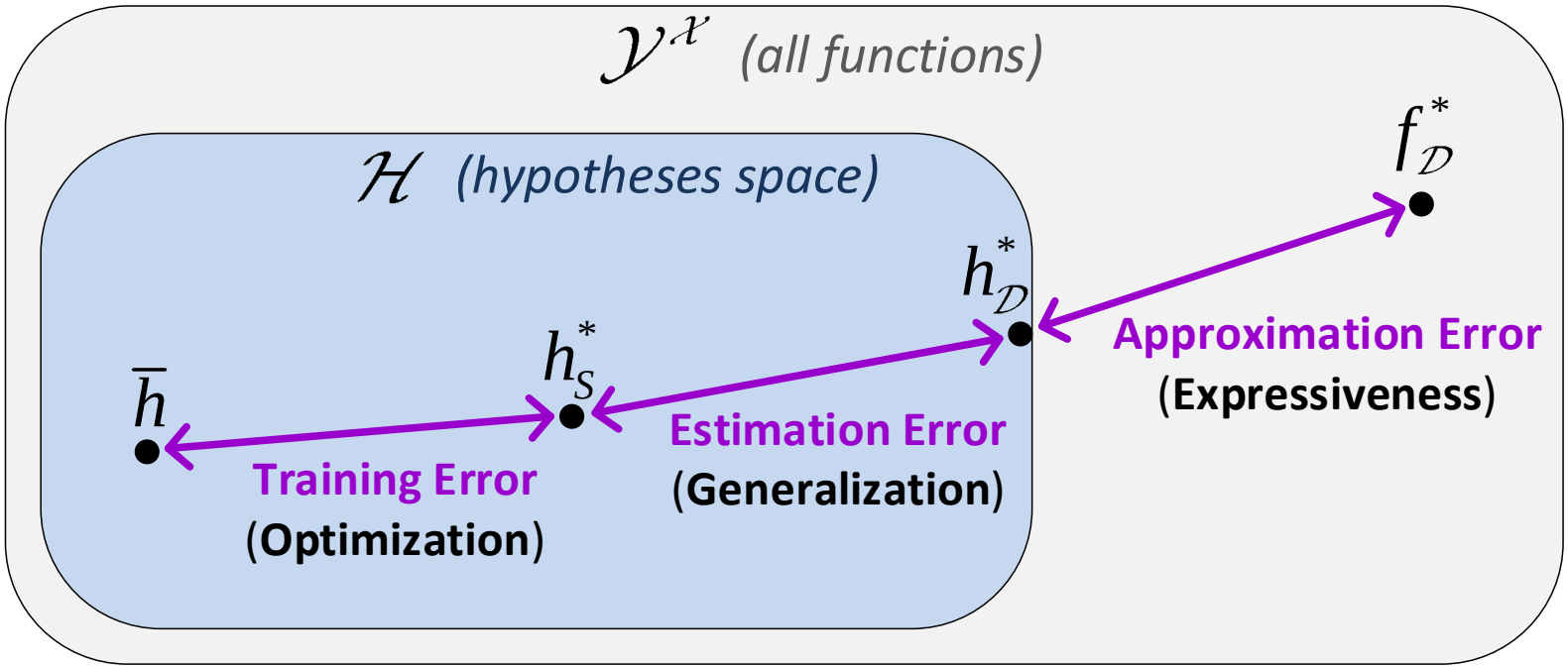
$f^*_{\mathcal{D}}$ — ground truth (minimizer of population loss over $\mathcal{Y}^{\mathcal{X}}$)

$h^*_{\mathcal{D}}$ — optimal hypothesis (minimizer of population loss over $\mathcal{H}$)

h^*_S — empirically optimal hypothesis (minimizer of empirical loss over $\mathcal{H}$)

$\bar{h}$ — returned hypothesis

Three Pillars of Statistical Learning Theory: Expressiveness, Generalization and Optimization



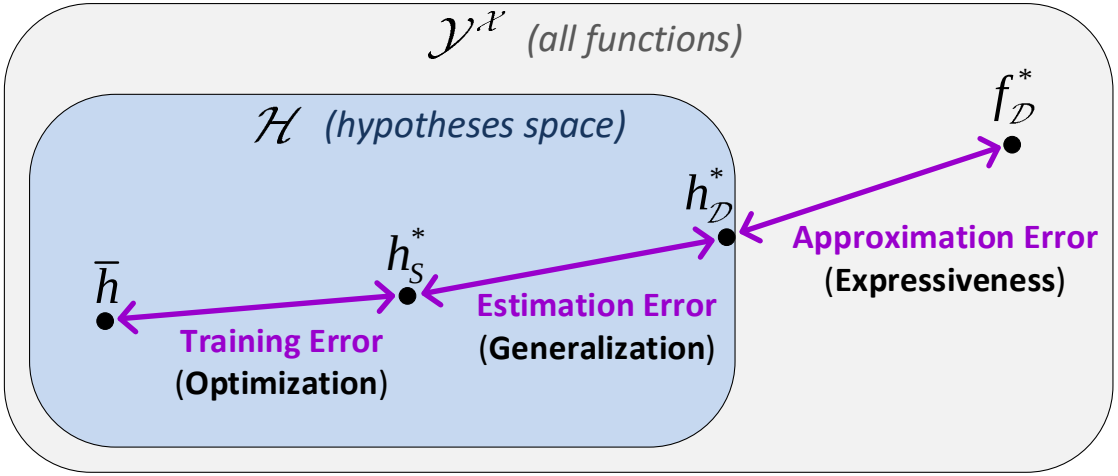
$f^*_{\mathcal{D}}$ — ground truth (minimizer of population loss over $\mathcal{Y}^{\mathcal{X}}$)

$h^*_{\mathcal{D}}$ — optimal hypothesis (minimizer of population loss over $\mathcal{H}$)

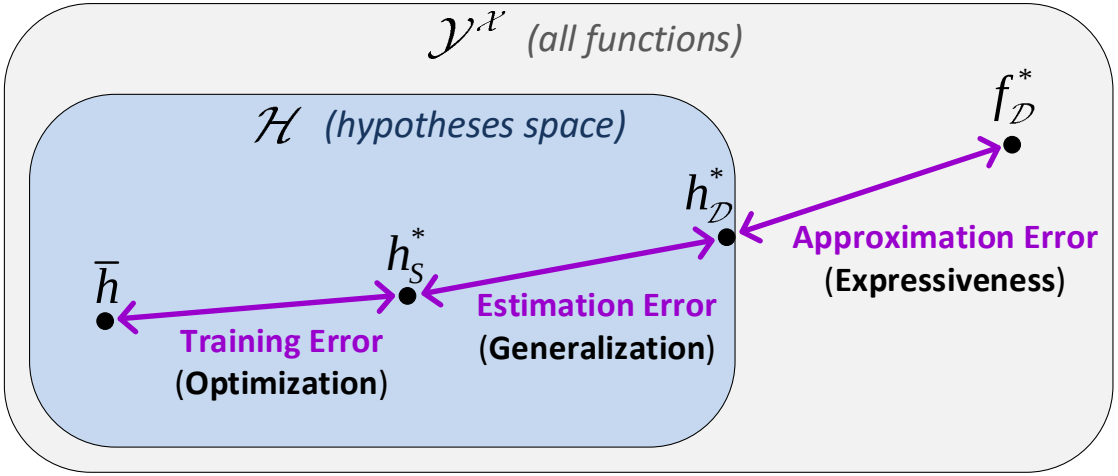
h^*_S — empirically optimal hypothesis (minimizer of empirical loss over $\mathcal{H}$)

$\bar{h}$ — returned hypothesis

Classical Machine Learning



Classical Machine Learning

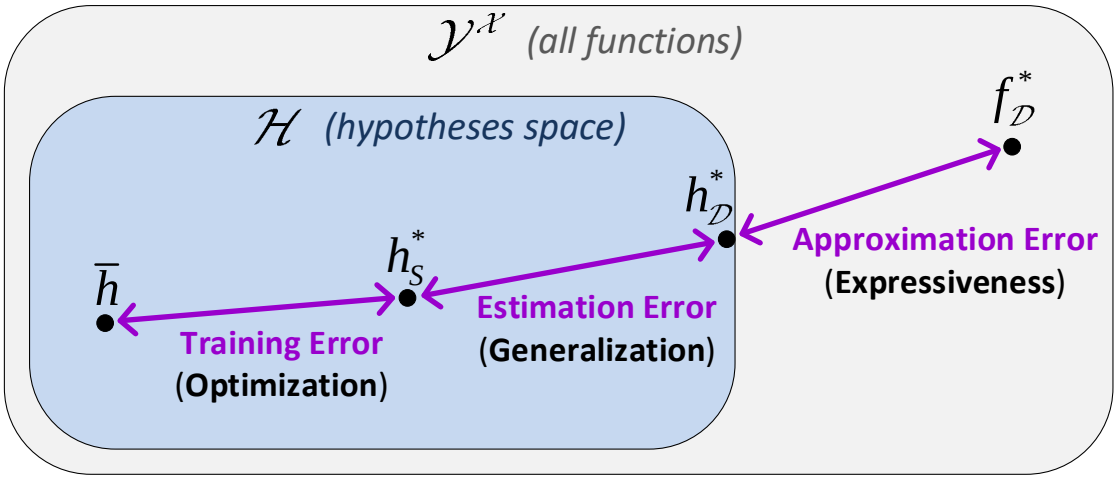


Optimization

Empirical loss minimization is a convex program:

$$\bar{h} \approx h^*_S \quad (\text{ training err} \approx 0 \text{ })$$

Classical Machine Learning



Optimization

Empirical loss minimization is a convex program:

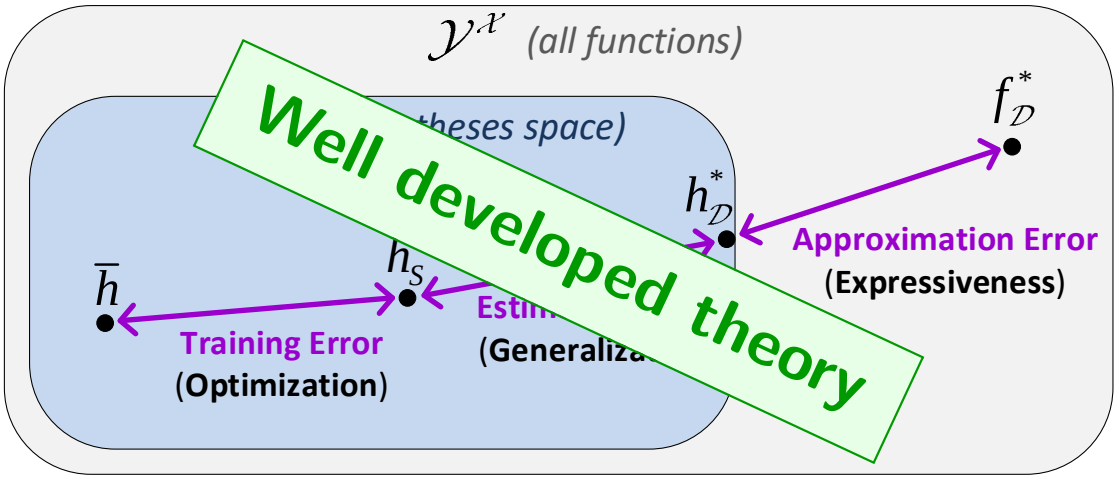
$$\bar{h} \approx h^*_S \quad (\text{ training err} \approx 0 \text{ })$$

Expressiveness & Generalization

Bias-variance trade-off:

$\mathcal{H}$	approximation err	estimation err
expands	↘	↗
shrinks	↗	↘

Classical Machine Learning



Optimization

Empirical loss minimization is a convex program:

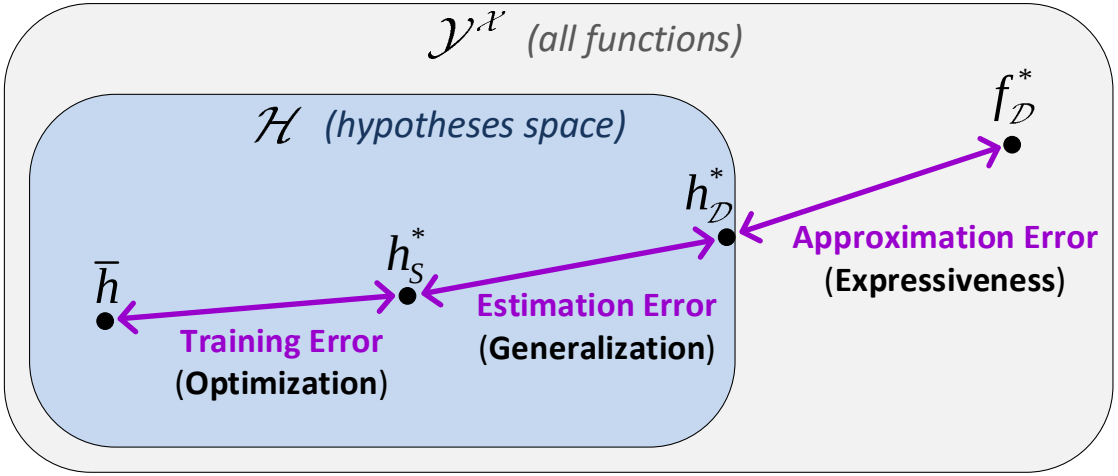
$$\bar{h} \approx h^*_S \quad (\text{ training err} \approx 0 \text{ })$$

Expressiveness & Generalization

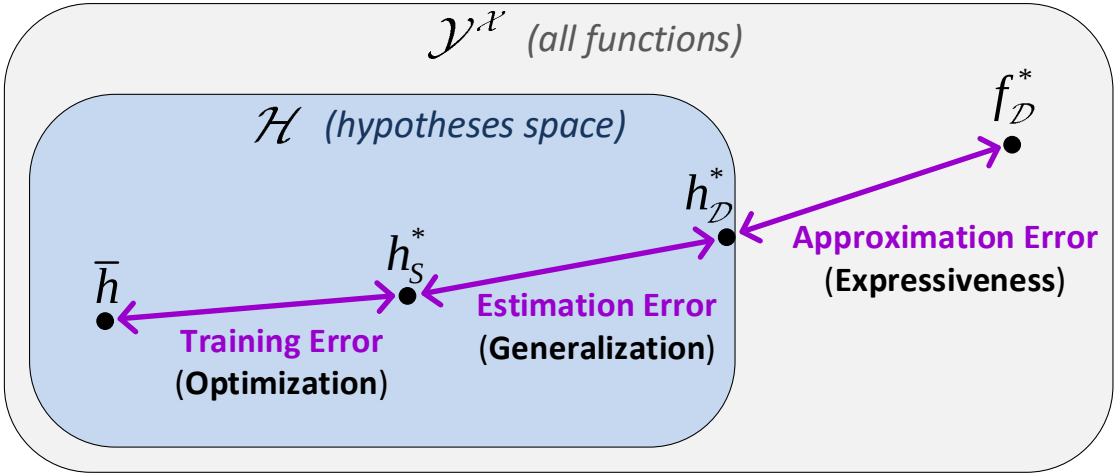
Bias-variance trade-off:

$\mathcal{H}$	approximation err	estimation err
expands	$\searrow$	$\nearrow$
shrinks	$\nearrow$	$\searrow$

Deep Learning



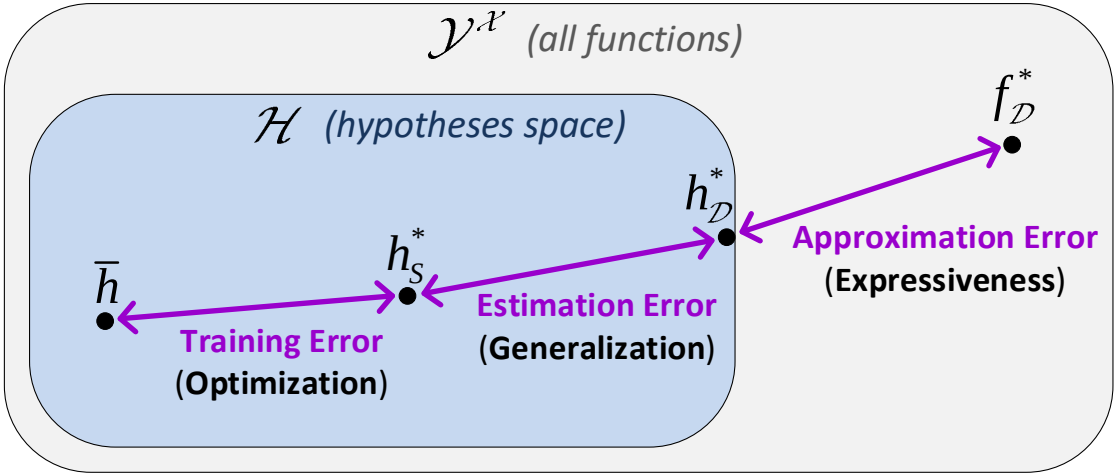
Deep Learning



Optimization

Empirical loss minimization is a non-convex program:

Deep Learning

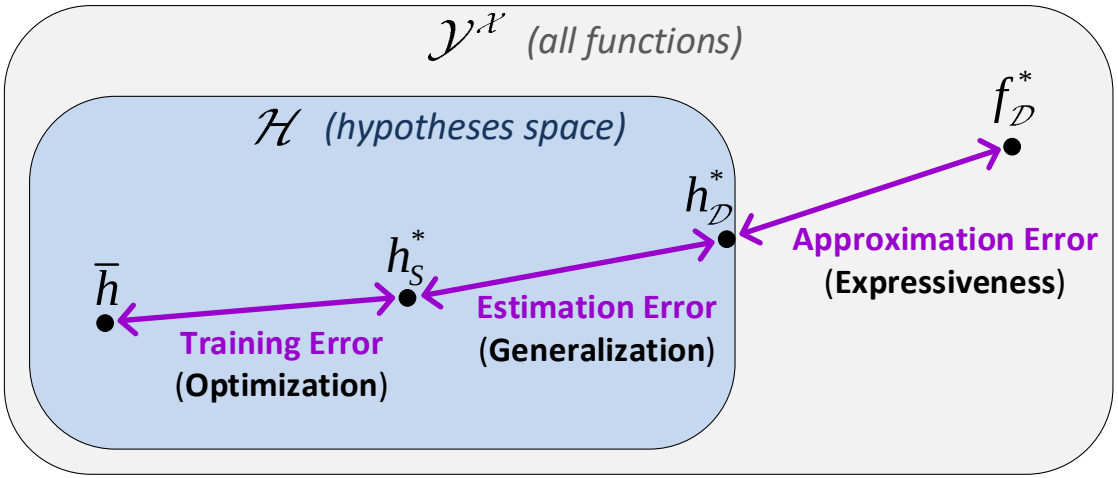


Optimization

Empirical loss minimization is a non-convex program:

- h^*_S is not unique — many hypotheses have low training err

Deep Learning

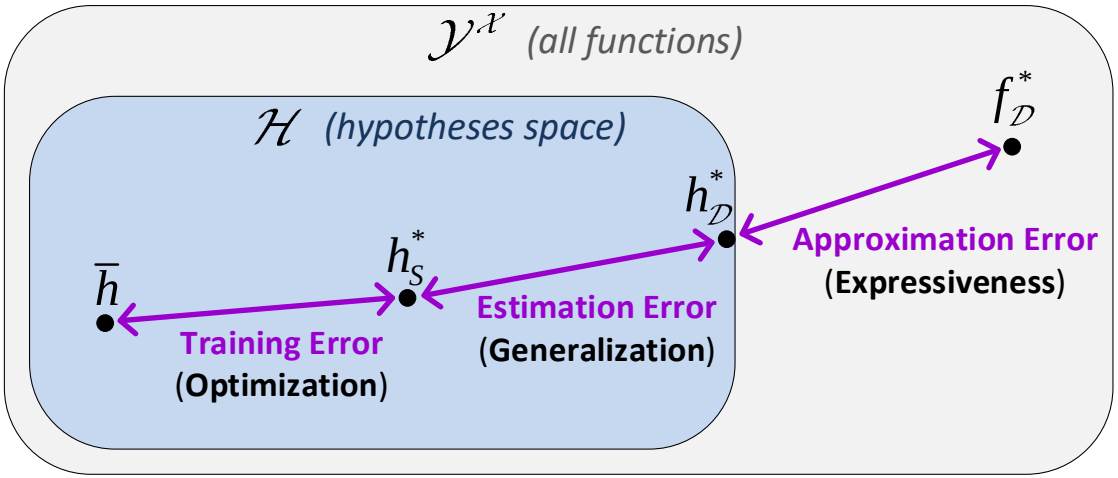


Optimization

Empirical loss minimization is a non-convex program:

- h^*_S is not unique — many hypotheses have low training err
- Gradient descent (GD) somehow reaches one of these

Deep Learning



Optimization

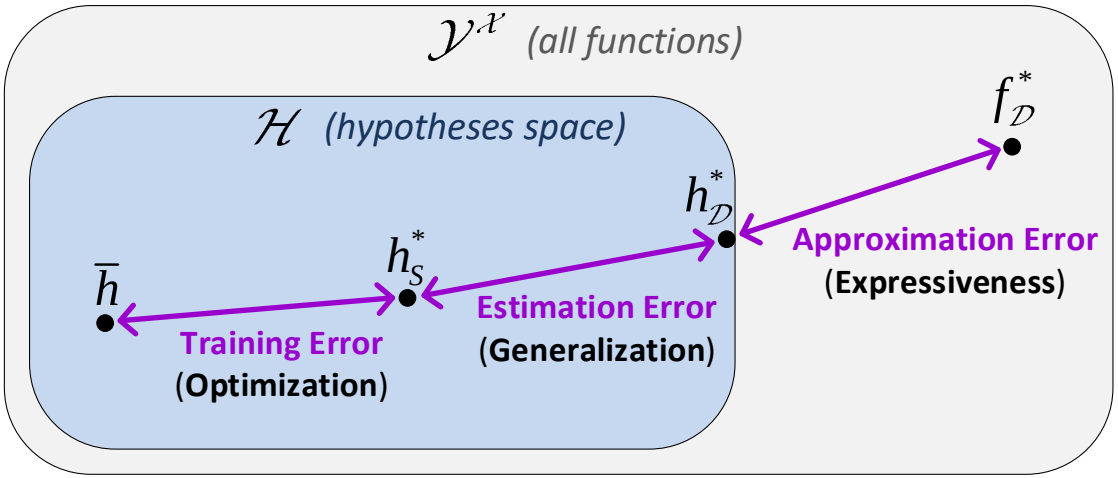
Empirical loss minimization is a non-convex program:

- h^*_S is not unique — many hypotheses have low training err
- Gradient descent (GD) somehow reaches one of these

Expressiveness & Generalization

Vast difference from classical ML:

Deep Learning



Optimization

Empirical loss minimization is a non-convex program:

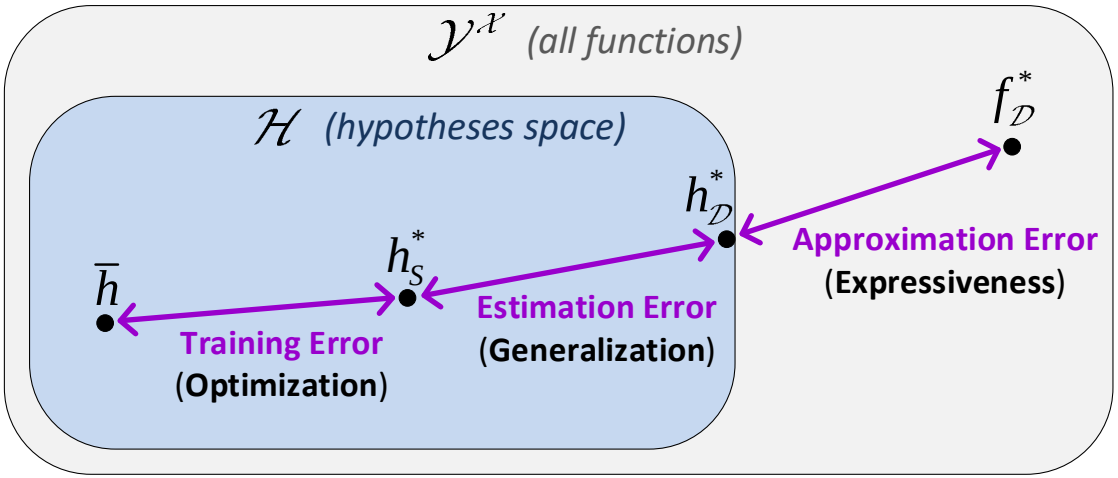
- h^*_S is not unique — many hypotheses have low training err
- Gradient descent (GD) somehow reaches one of these

Expressiveness & Generalization

Vast difference from classical ML:

- Some low training err hypotheses generalize well, others don't

Deep Learning



Optimization

Empirical loss minimization is a non-convex program:

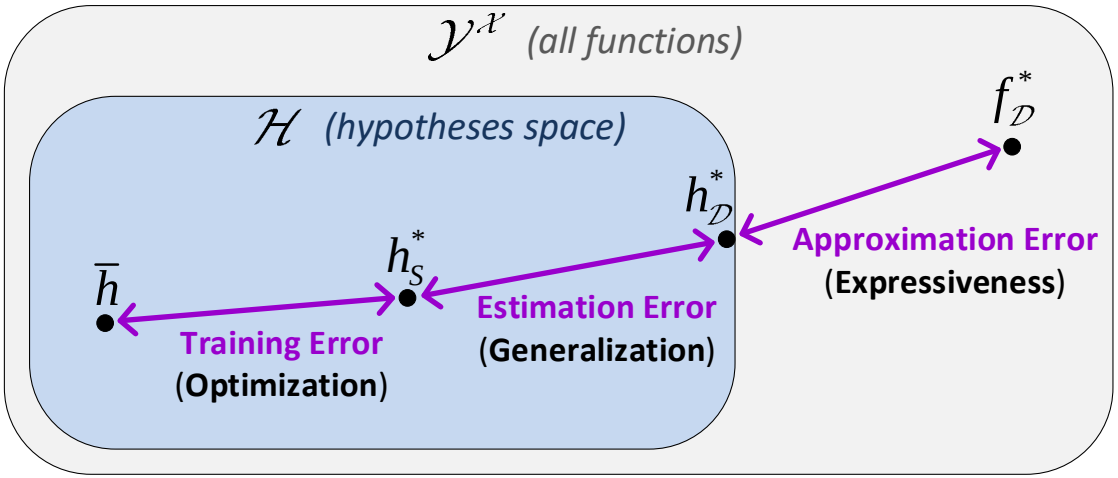
- h^*_S is not unique — many hypotheses have low training err
- Gradient descent (GD) somehow reaches one of these

Expressiveness & Generalization

Vast difference from classical ML:

- Some low training err hypotheses generalize well, others don't
- W/typical data, solution returned by GD often generalizes well

Deep Learning



Optimization

Empirical loss minimization is a non-convex program:

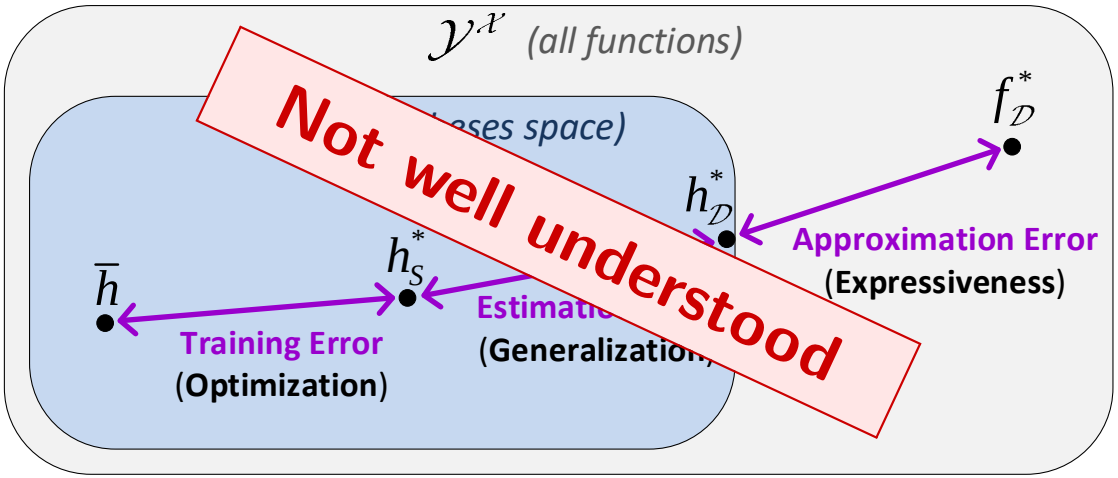
- h^*_S is not unique — many hypotheses have low training err
- Gradient descent (GD) somehow reaches one of these

Expressiveness & Generalization

Vast difference from classical ML:

- Some low training err hypotheses generalize well, others don't
- W/typical data, solution returned by GD often generalizes well
- Expanding $\mathcal{H}$ reduces approximation err, but also estimation err!

Deep Learning



Optimization

Empirical loss minimization is a non-convex program:

- h^*_S is not unique — many hypotheses have low training err
- Gradient descent (GD) somehow reaches one of these

Expressiveness & Generalization

Vast difference from classical ML:

- Some low training err hypotheses generalize well, others don't
- W/typical data, solution returned by GD often generalizes well
- Expanding $\mathcal{H}$ reduces approximation err, but also estimation err!

Outline

Outline

Sources

Deep SimNets
C + Sharir + Shashua
Computer Vision and Pattern Recognition (CVPR) 2016

On the Expressive Power of Deep Learning: A Tensor Analysis
C + Sharir + Shashua
Conference on Learning Theory (COLT) 2016

Convolutional Rectifier Networks as Generalized Tensor Decompositions
C + Shashua
International Conference on Machine Learning (ICML) 2016

Inductive Bias of Deep Convolutional Networks through Pooling Geometry
C + Shashua
International Conference on Learning Representations (ICLR) 2017

Boosting Dilated Convolutional Networks with Mixed Tensor Decompositions
C + Tamari + Shashua
International Conference on Learning Representations (ICLR) 2018

Deep Learning and Quantum Entanglement:
Fundamental Connections with Implications to Network Design
Levine + Yakira + *C* + Shashua
International Conference on Learning Representations (ICLR) 2018

Bridging Many-Body Quantum Physics and Deep Learning via Tensor Networks
Levine + Sharir + *C* + Shashua
Physical Review Letters (PRL) 2019

Collaborators

Or Sharir

Amnon Shashua

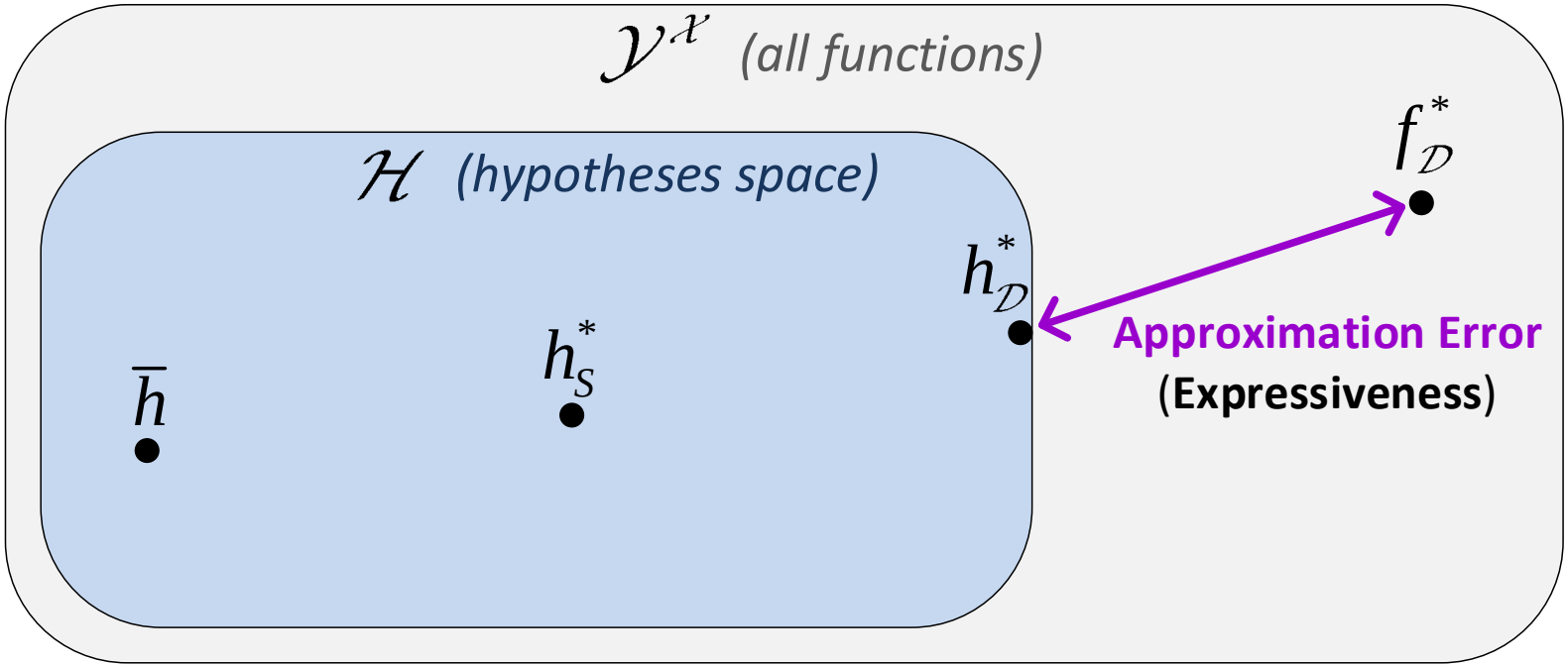
Yoav Levine

Ronen Tamari

THE HEBREW UNIVERSITY OF JERUSALEM

David Yakira

Expressiveness



$f^*_{\mathcal{D}}$ — ground truth

$h^*_{\mathcal{D}}$ — optimal hypothesis

h^*_S — empirically optimal hypothesis

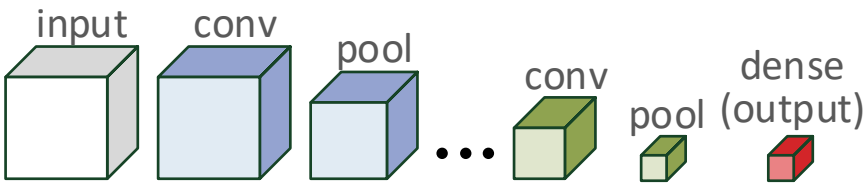
$\bar{h}$ — returned hypothesis

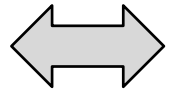
Tensor Analysis for Convolutional Neural Networks

Tensor Analysis for Convolutional Neural Networks

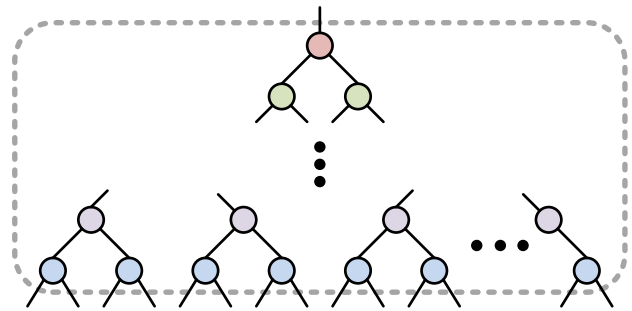
We derived an equivalence:

Convolutional Neural Networks (CNN)





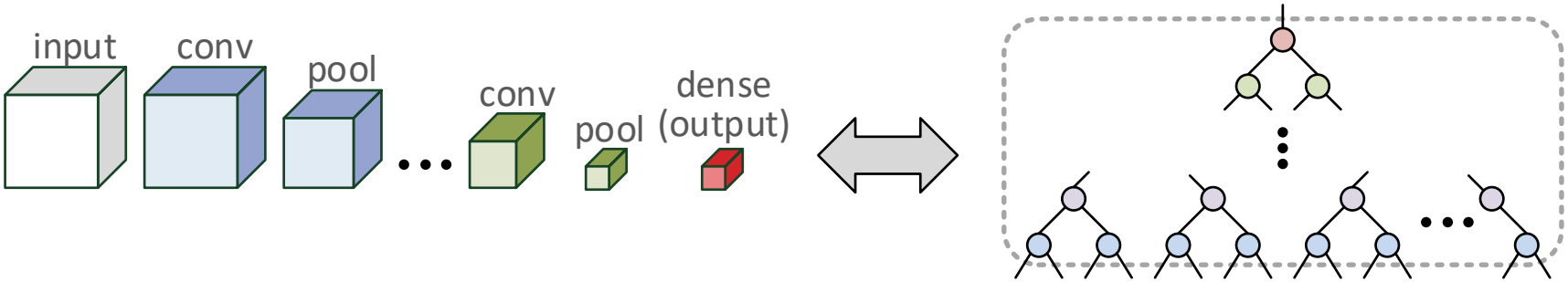
Hierarchical Tensor Factorizations (HTF)



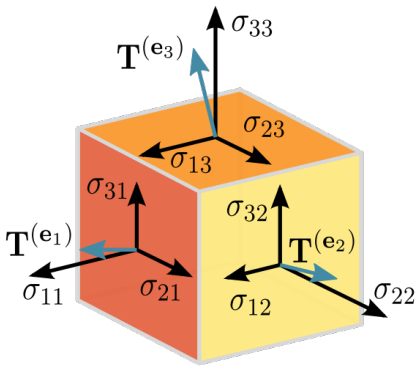
Tensor Analysis for Convolutional Neural Networks

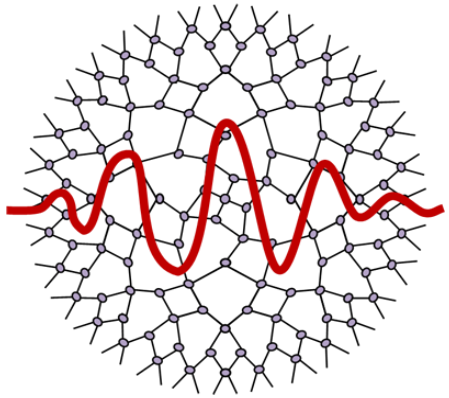
We derived an equivalence:

Convolutional Neural Networks (CNN) ⟺ **Hierarchical Tensor Factorizations** (HTF)



HTF are widely used in Applied Math and Quantum Physics

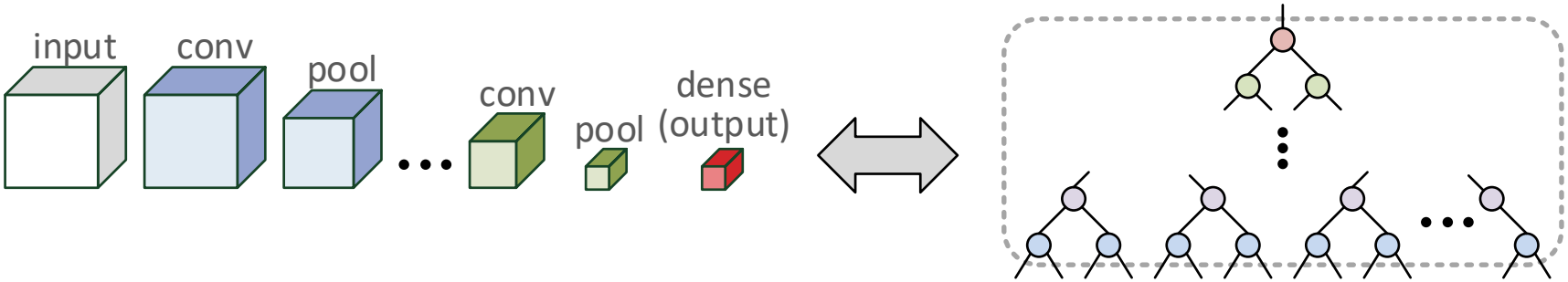




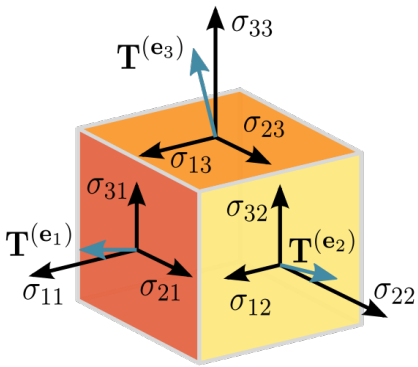
Tensor Analysis for Convolutional Neural Networks

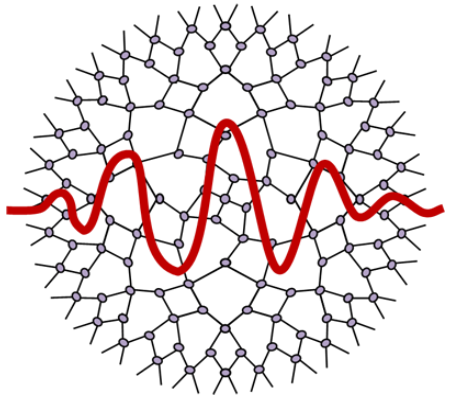
We derived an equivalence:

Convolutional Neural Networks (CNN) ⟺ **Hierarchical Tensor Factorizations** (HTF)



HTF are widely used in Applied Math and Quantum Physics



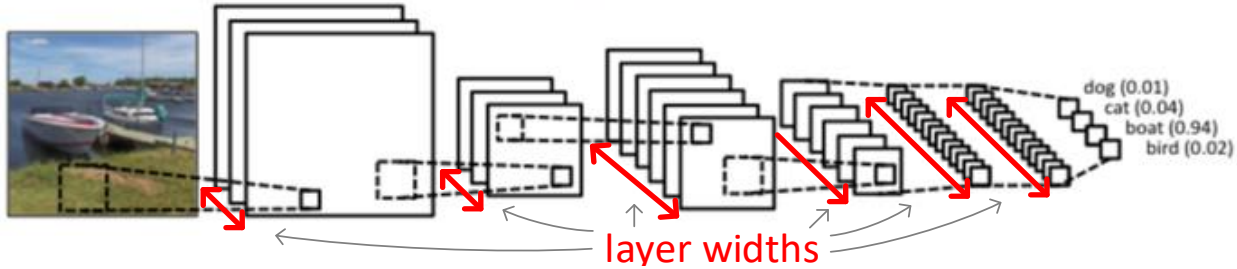


We adopted tools from these domains to analyze expressiveness of CNN

Result: Guideline for Choosing Layer Widths

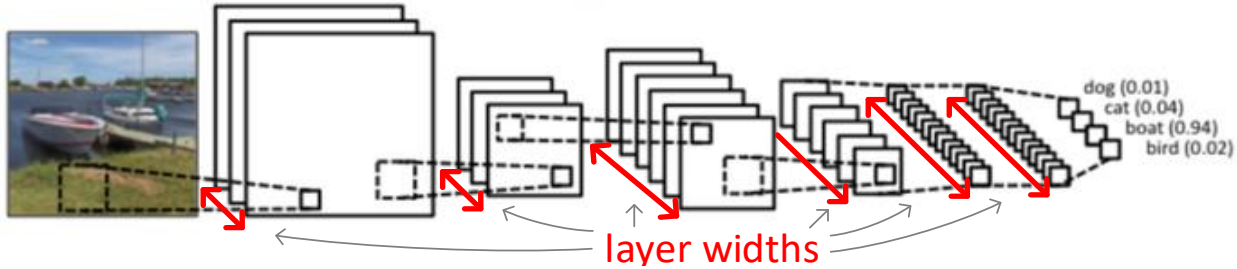
Result: Guideline for Choosing Layer Widths

Currently no principle for choosing widths (# of channels) of CNN layers



Result: Guideline for Choosing Layer Widths

Currently no principle for choosing widths (# of channels) of CNN layers

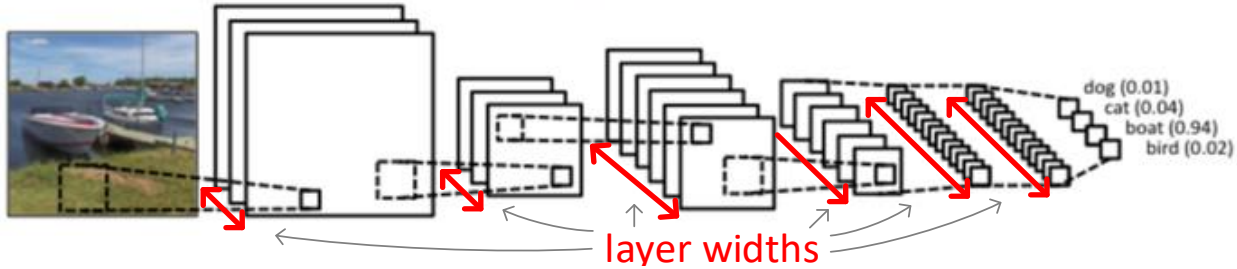


Theorem

Deep (early) layer widths needed to express long (short)-range correlations

Result: Guideline for Choosing Layer Widths

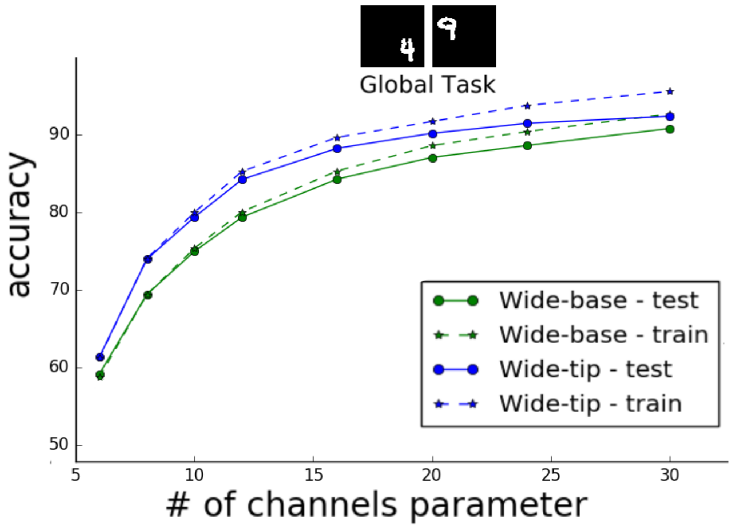
Currently no principle for choosing widths (# of channels) of CNN layers

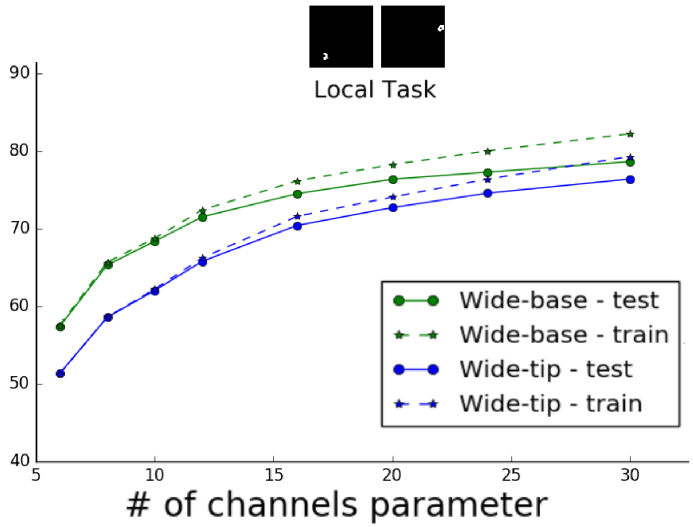


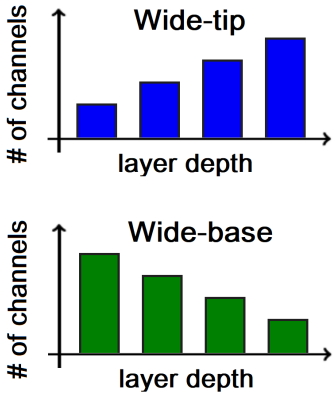
Theorem

Deep (early) layer widths needed to express long (short)-range correlations

Experiment



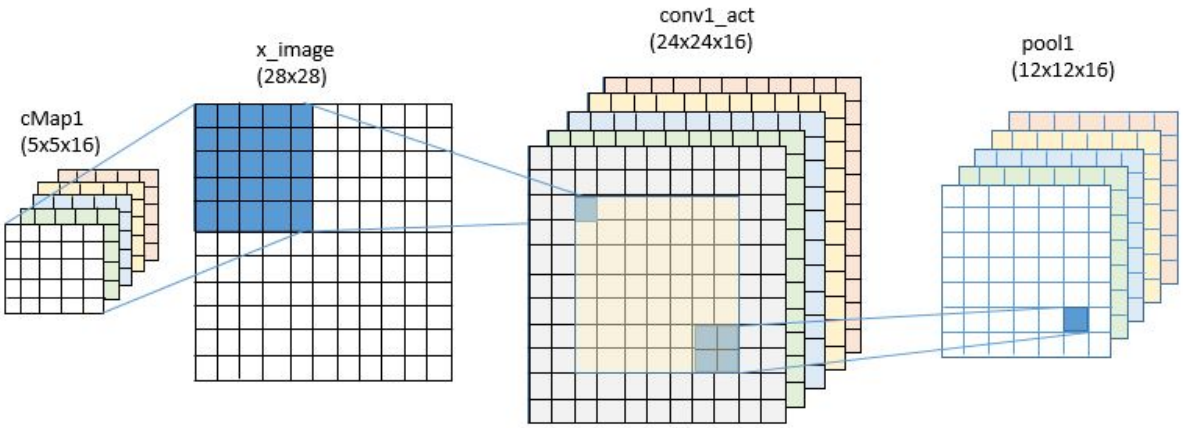




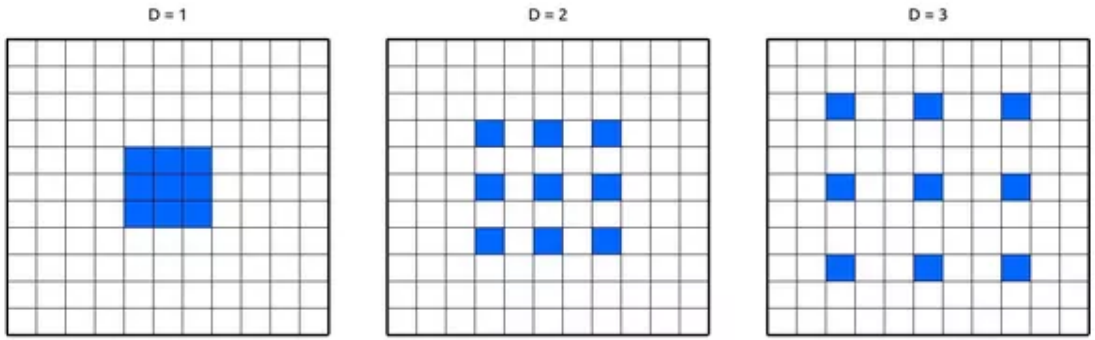
Result: Guideline for Choosing Pooling Geometry

Result: Guideline for Choosing Pooling Geometry

CNN typically employ square conv/pool windows



Recently, dilated windows have also become popular



Currently no principle for choosing window geometries

Result: Guideline for Choosing Pooling Geometry (cont'd)

Theorem

Input elements pooled together early have stronger correlation

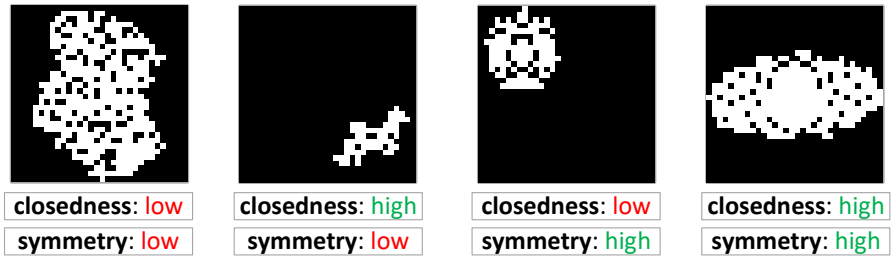
Result: Guideline for Choosing Pooling Geometry (cont'd)

Theorem

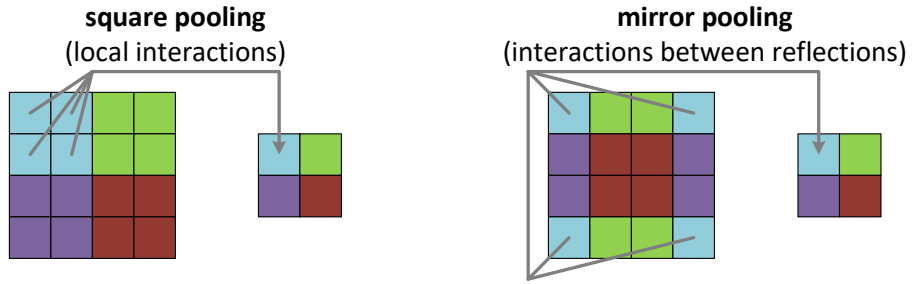
Input elements pooled together early have stronger correlation

Experiment

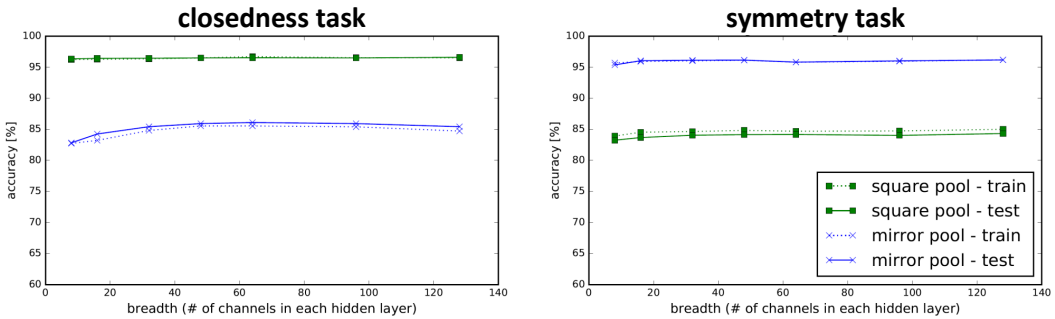
data



archs



results



Outline

Sources

On the Optimization of Deep Networks:
Implicit Acceleration by Overparameterization

Arora + ***C*** + Hazan (alphabetical order)

International Conference on Machine Learning (ICML) 2018

A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks

Arora + ***C*** + Golowich + Hu (alphabetical order)

International Conference on Learning Representations (ICLR) 2019

Implicit Regularization in Deep Matrix Factorization

Arora + ***C*** + Hu + Luo (alphabetical order)

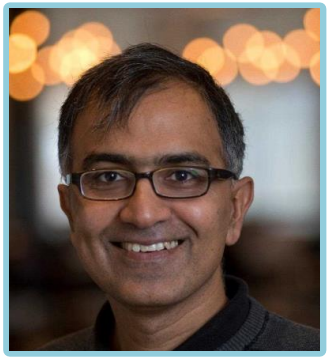
Conference on Neural Information Processing Systems (NeurIPS) 2019

Implicit Regularization in Deep Learning May Not Be Explainable by Norms

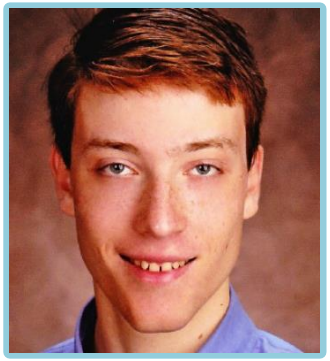
Razin + ***C***

Conference on Neural Information Processing Systems (NeurIPS) 2020

Collaborators



Sanjeev Arora



Noah Golowich

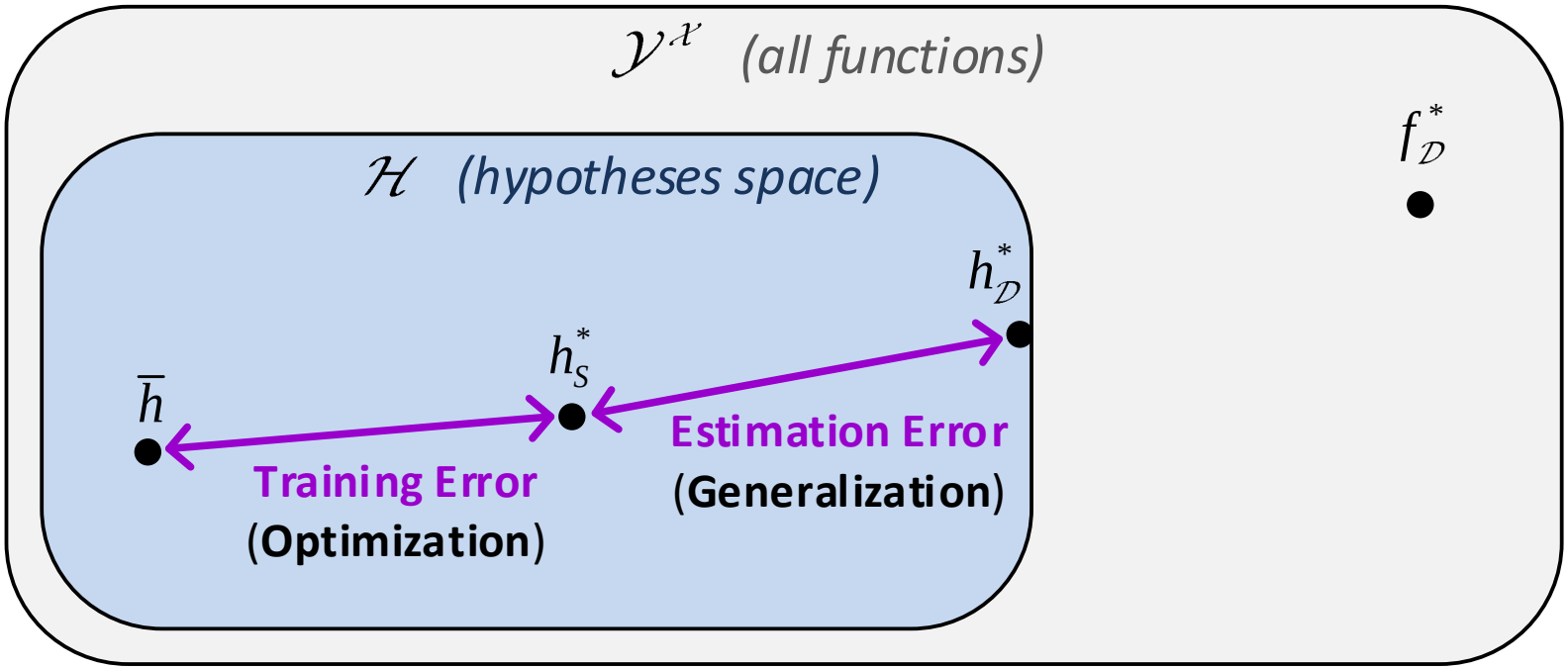
Elad Hazan

Wei Hu

Yuping Luo

Noam Razin

Optimization & Generalization



$f^*_{\mathcal{D}}$ — ground truth

$h^*_{\mathcal{D}}$ — optimal hypothesis

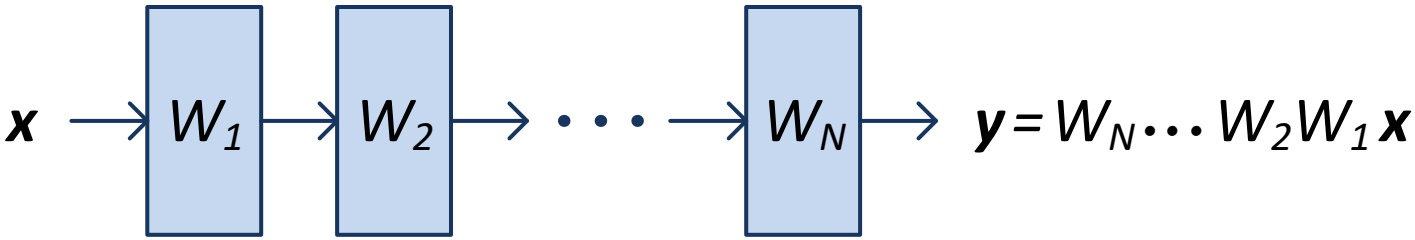
h^*_S — empirically optimal hypothesis

$\bar{h}$ — returned hypothesis

Dynamical Analysis for Linear Neural Networks

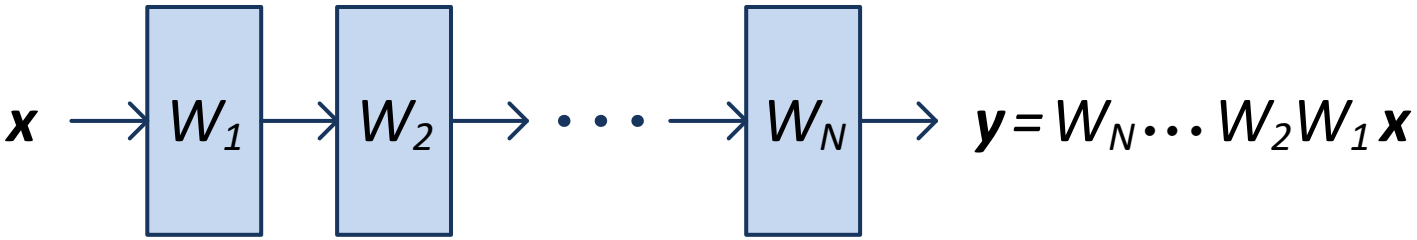
Dynamical Analysis for Linear Neural Networks

Linear Neural Networks (LNN) are neural networks with no activation



Dynamical Analysis for Linear Neural Networks

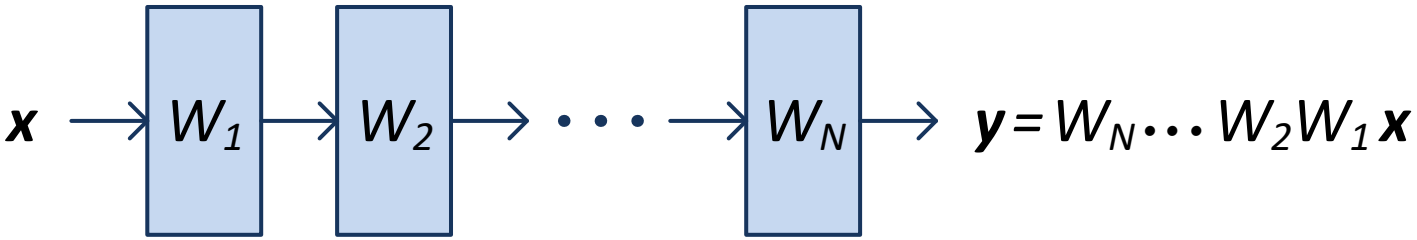
Linear Neural Networks (LNN) are neural networks with no activation



Expressiveness of LNN is trivial, but optimization & generalization are not!

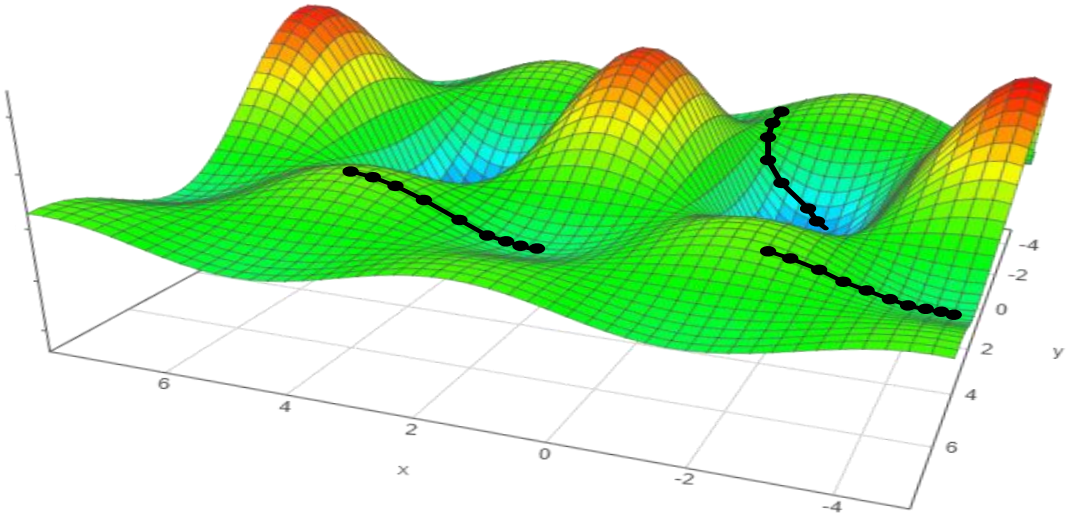
Dynamical Analysis for Linear Neural Networks

Linear Neural Networks (LNN) are neural networks with no activation



Expressiveness of LNN is trivial, but optimization & generalization are not!

We study them by analyzing the dynamics (trajectories) of GD



Result: Depth Can Accelerate Optimization

Result: Depth Can Accelerate Optimization

Theorem

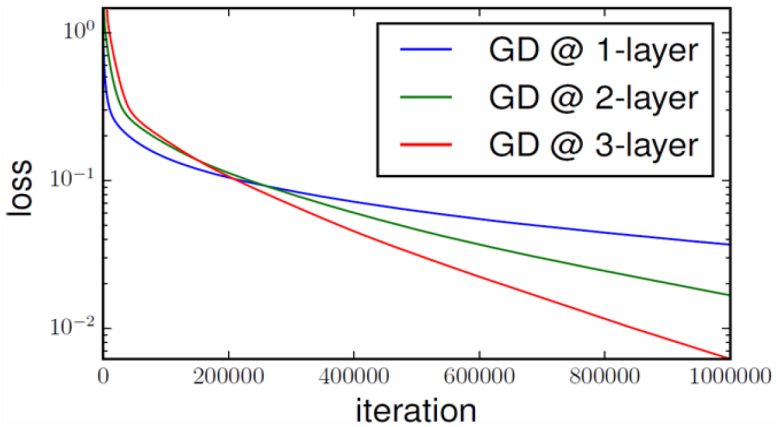
GD over LNN can converge arbitrarily faster than GD over linear model

Result: Depth Can Accelerate Optimization

Theorem

GD over LNN can converge arbitrarily faster than GD over linear model

Experiment

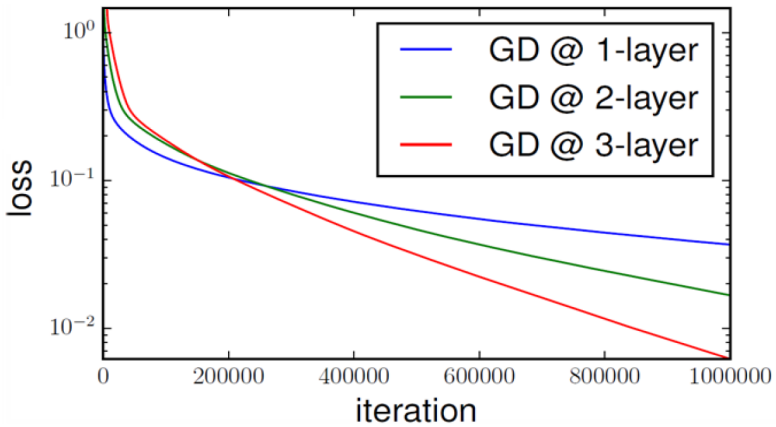


Result: Depth Can Accelerate Optimization

Theorem

GD over LNN can converge arbitrarily faster than GD over linear model

Experiment



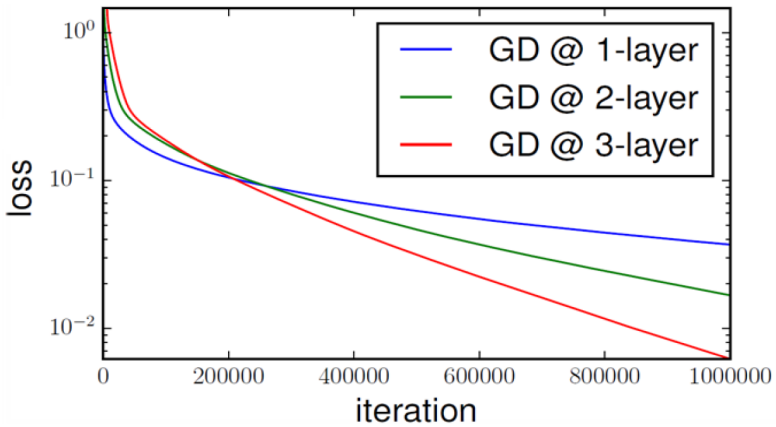
Depth can speed-up GD, even without any gain in expressiveness, and despite introducing non-convexity!

Result: Depth Can Accelerate Optimization

Theorem

GD over LNN can converge arbitrarily faster than GD over linear model

Experiment



Depth can speed-up GD, even without any gain in expressiveness, and despite introducing non-convexity!

Practical Application

Blind Super-Resolution Kernel Estimation using an Internal-GAN

Sefi Bell-Kligler **Assaf Shocher** **Michal Irani**
Dept. of Computer Science and Applied Math
The Weizmann Institute of Science, Israel

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

"...we found that using a deep linear network is dramatically superior to a single-strided one. This is consistent with recent findings in theoretical deep-learning..."

Result: Depth Implicitly Minimizes Rank

Result: Depth Implicitly Minimizes Rank

Theorem

GD over LNN finds solutions with sparse spectrum (low rank)

Result: Depth Implicitly Minimizes Rank

Theorem

GD over LNN finds solutions with sparse spectrum (low rank)

↑

Leads to generalization for matrix completion

Result: Depth Implicitly Minimizes Rank

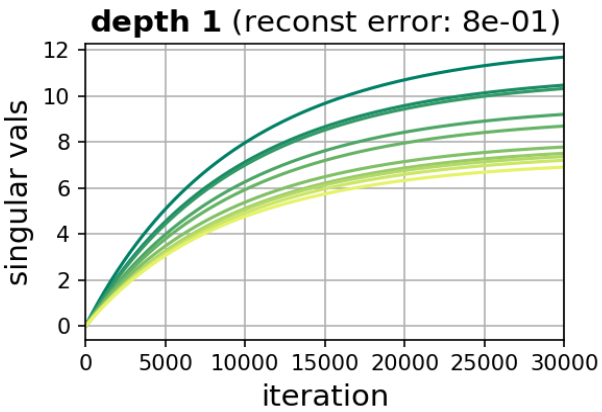
Theorem

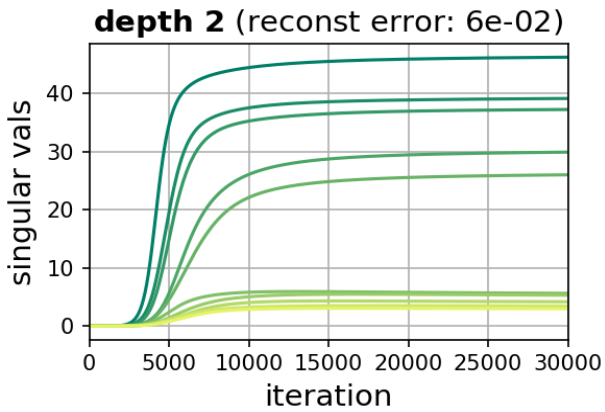
GD over LNN finds solutions with sparse spectrum (low rank)

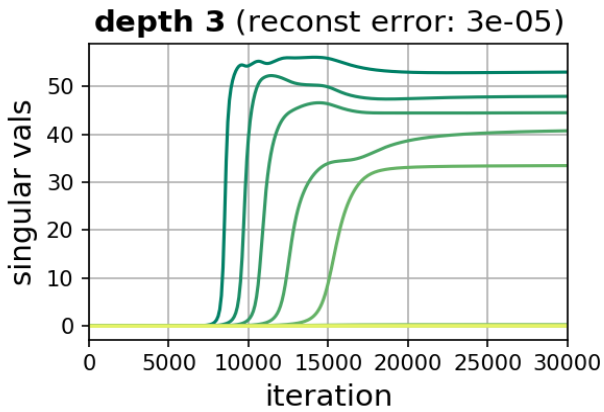
↑

Leads to generalization for matrix completion

Experiment







Result: Depth Implicitly Minimizes Rank

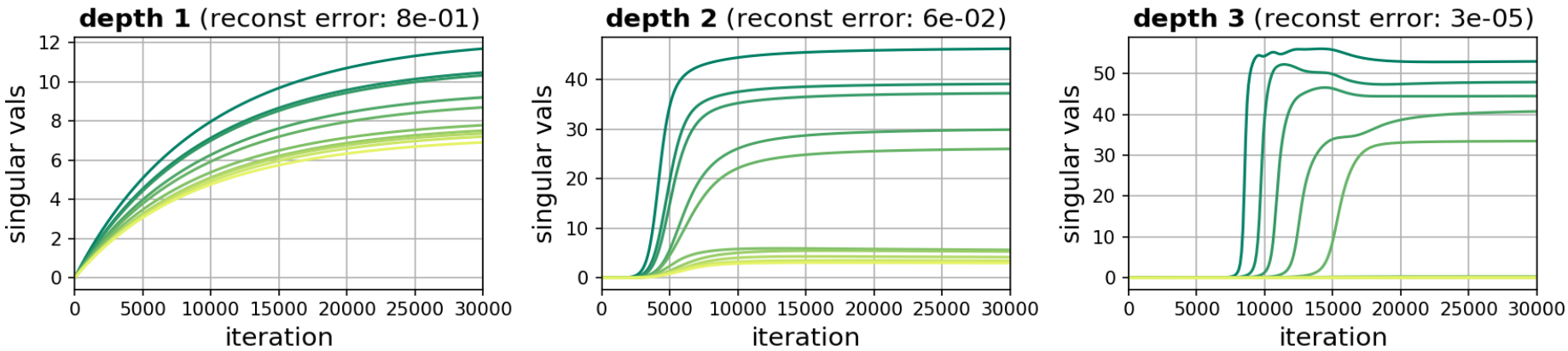
Theorem

GD over LNN finds solutions with sparse spectrum (low rank)

↑

Leads to generalization for matrix completion

Experiment



Practical Application

Implicit Rank-Minimizing Autoencoder

Li Jing
Facebook AI Research
New York
ljng@fb.com

Jure Zbontar
Facebook AI Research
New York
jzb@fb.com

Yann LeCun
Facebook AI Research
New York
yann@fb.com

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

"In this work, the rank of ... codes is implicitly minimized by relying on the fact that gradient descent learning in multi-layer linear networks leads to minimum-rank solutions."

Outline

Recap

Recap

Theory in deep learning may lead to better practice

Recap

Theory in deep learning may lead to better practice

Fundamental questions:

Expressiveness Optimization Generalization

Recap

Theory in deep learning may lead to better practice

Fundamental questions:

Expressiveness Optimization Generalization

Examples of theories with practical implications:

Recap

Theory in deep learning may lead to better practice

Fundamental questions:

Expressiveness Optimization Generalization

Examples of theories with practical implications:

- Expressiveness via tensor analysis

Recap

Theory in deep learning may lead to better practice

Fundamental questions:

Expressiveness Optimization Generalization

Examples of theories with practical implications:

- Expressiveness via tensor analysis

  $\Longrightarrow$ guidelines for designing CNN per required input correlations

Recap

Theory in deep learning may lead to better practice

Fundamental questions:

Expressiveness Optimization Generalization

Examples of theories with practical implications:

- Expressiveness via tensor analysis

  $\Longrightarrow$ guidelines for designing CNN per required input correlations

- Optimization & generalization via dynamical analysis

Recap

Theory in deep learning may lead to better practice

Fundamental questions:

Expressiveness Optimization Generalization

Examples of theories with practical implications:

- Expressiveness via tensor analysis

  $\Longrightarrow$ guidelines for designing CNN per required input correlations

- Optimization & generalization via dynamical analysis

  $\Longrightarrow$ acceleration and low rank solutions via linear neural networks

Recap

Theory in deep learning may lead to better practice

Fundamental questions:

Expressiveness Optimization Generalization

Examples of theories with practical implications:

- Expressiveness via tensor analysis

  $\Longrightarrow$ guidelines for designing CNN per required input correlations

- Optimization & generalization via dynamical analysis

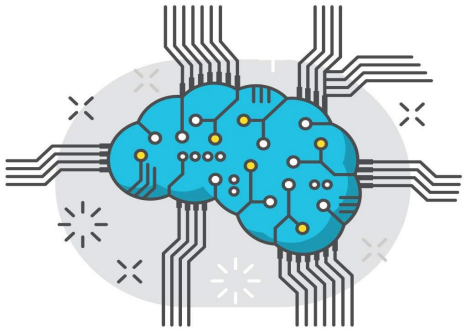
  $\Longrightarrow$ acceleration and low rank solutions via linear neural networks

  $\uparrow$

  (cf. Bell-Kligler, Shocher & Irani (2019); Jing, Zbontar & LeCun (2020))

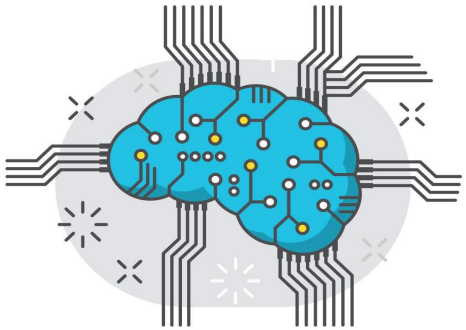
The Road to Strong Artificial Intelligence

Strong Artificial Intelligence



The Road to Strong Artificial Intelligence

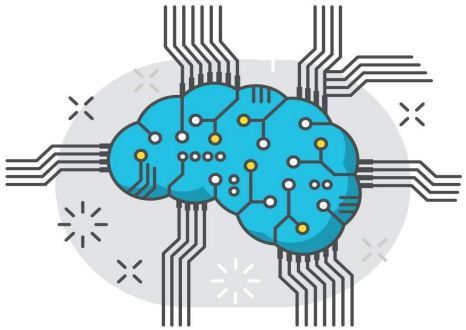
Strong Artificial Intelligence

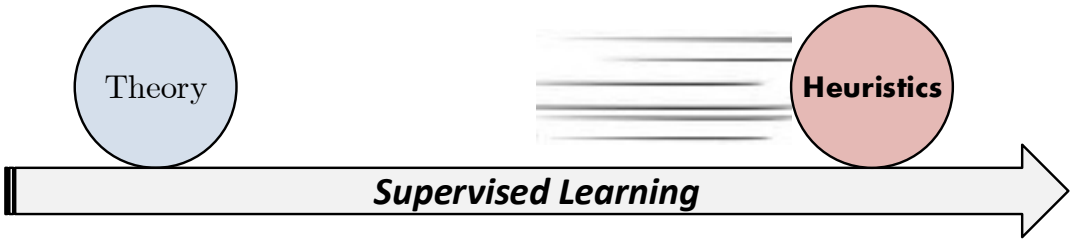


Supervised Learning

The Road to Strong Artificial Intelligence

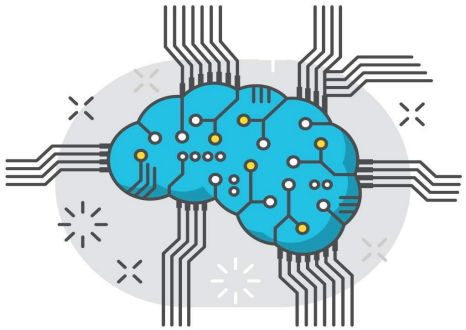
Strong Artificial Intelligence

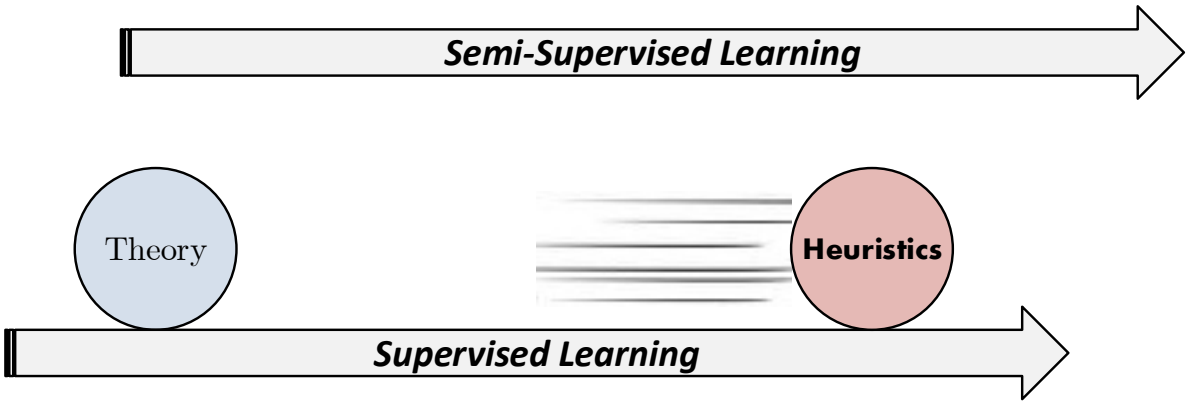




The Road to Strong Artificial Intelligence

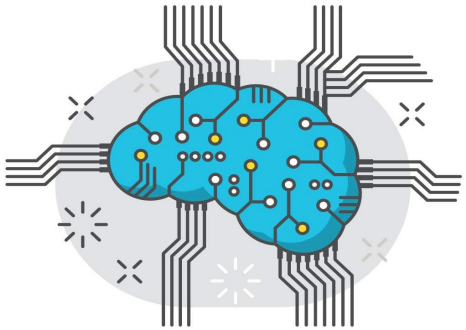
Strong Artificial Intelligence





The Road to Strong Artificial Intelligence

Strong Artificial Intelligence



Reinforcement Learning

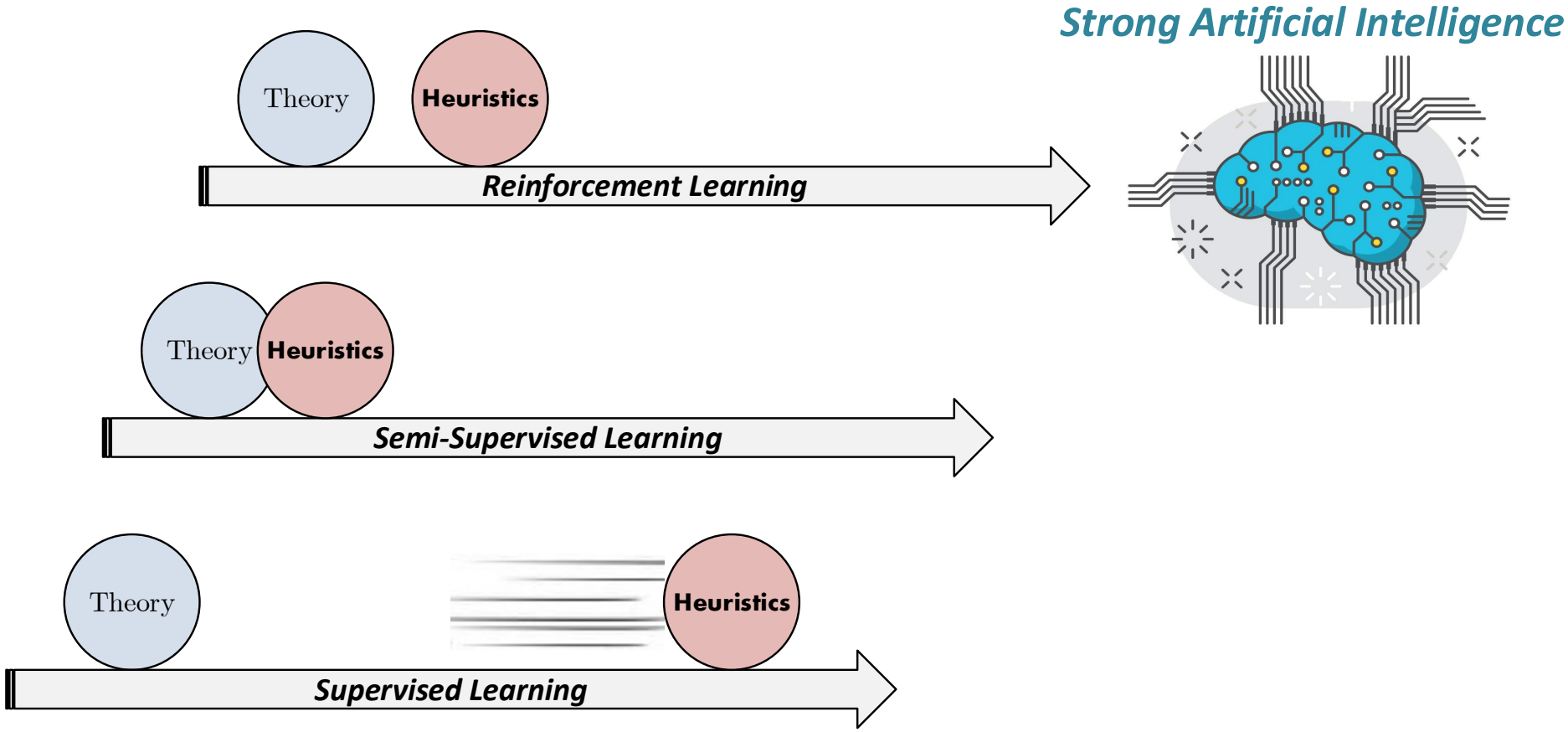
Semi-Supervised Learning

Theory

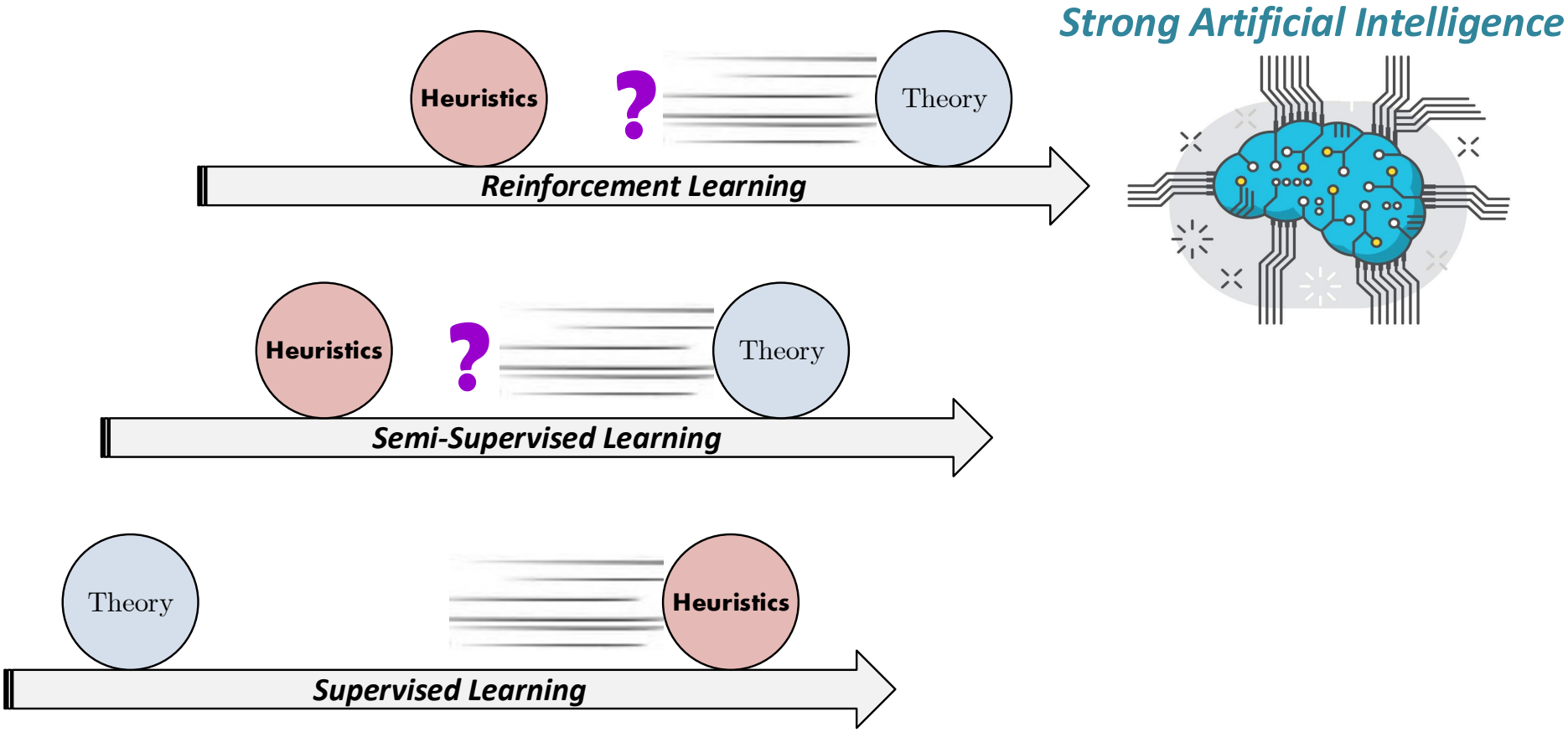
Heuristics

Supervised Learning

The Road to Strong Artificial Intelligence



The Road to Strong Artificial Intelligence



Outline

1. Practice Needs Theory in Deep Learning
2. Fundamental Questions: Expressiveness, Optimization & Generalization
3. Examples of Theories with Practical Implications
   - Expressiveness via Tensor Analysis
   - Optimization & Generalization via Dynamical Analysis
4. Conclusion

Thank You