ML & AI: A SWOT Analysis

Volkan Cevher, Associate Professor EPFL

Preface

My research:

- Machine Learning (ML)
- Optimization
- Signal Processing
- Information Theory
- Statistics

My courses (2019-20):

- Mathematics of Data
- Reinforcement Learning
- Advanced Topics in ML

+

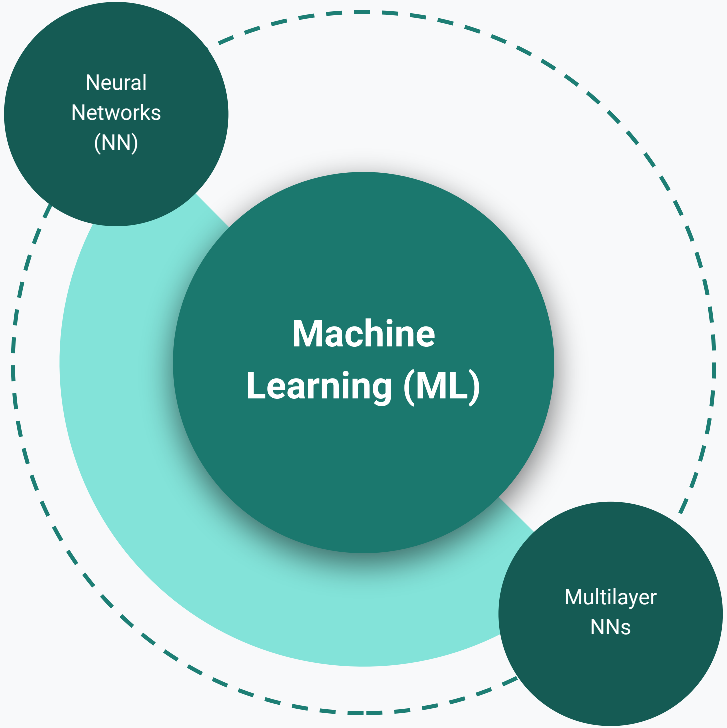
=

This talk

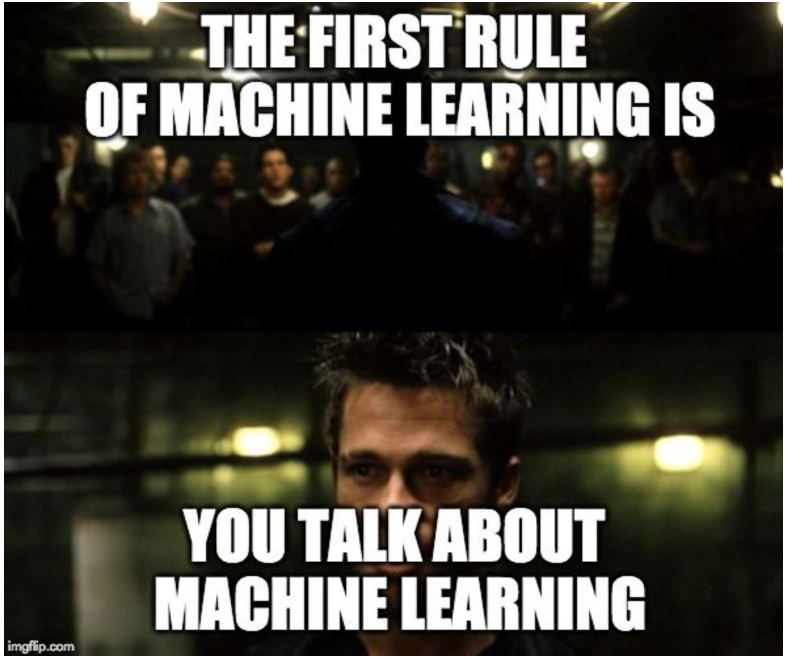

X-MEN

Strengths

A SWOT Analysis



Machine Learning (ML)

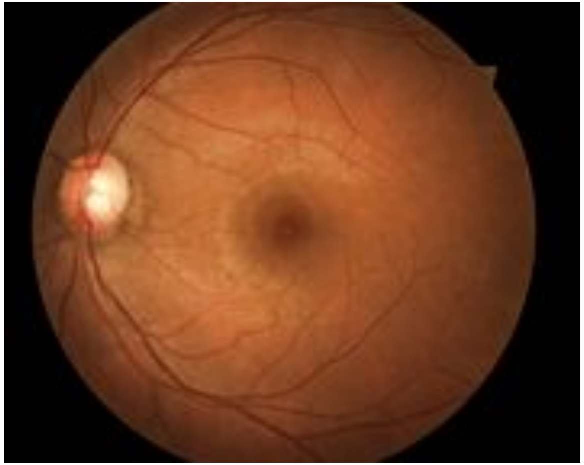


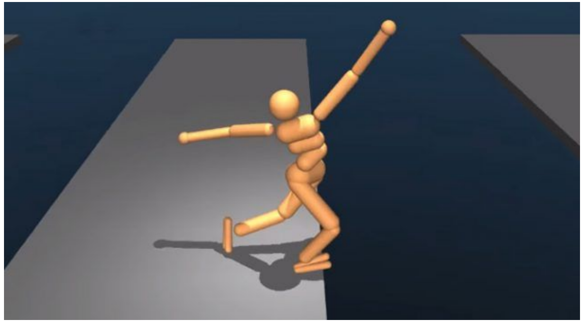
- ML is an interdisciplinary study of algorithms, statistical models, and error functions jointly with computer systems to perform specific tasks

  “Only a fool learns from his own mistakes. The wise man learns from the mistakes of others” - Otto von Bismarck

- ML makes you wiser

The ingredients via a simplified supervised learning example





- ML is an interdisciplinary study of algorithms, statistical models, and error functions jointly with computer systems to perform specific tasks

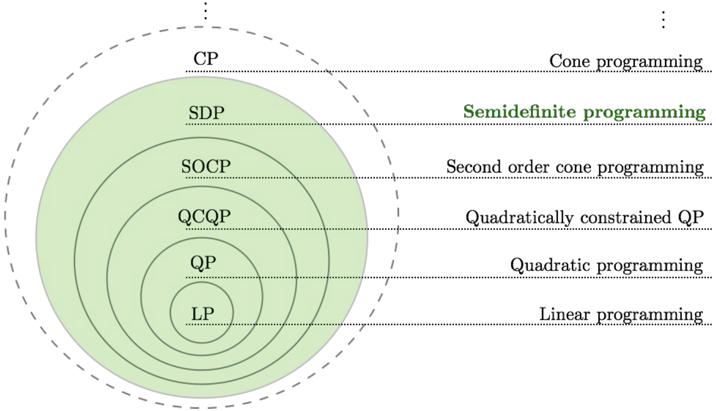
Task: Learn a mapping from image to disease

$$\mathbf{y} = \text{function}_{\mathbf{x}}(\mathbf{a}) = \underbrace{f(\mathbf{a}'\mathbf{x})}_{\text{model}}$$

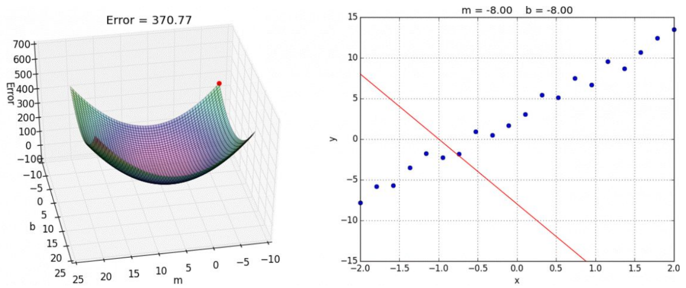
Task: Learn a mapping from control inputs to walking

The ingredients via a simplified supervised learning example

Convex optimization



Gradient Descent Algorithm



- ML is an interdisciplinary study of algorithms, statistical models, and error functions jointly with computer systems to perform specific tasks

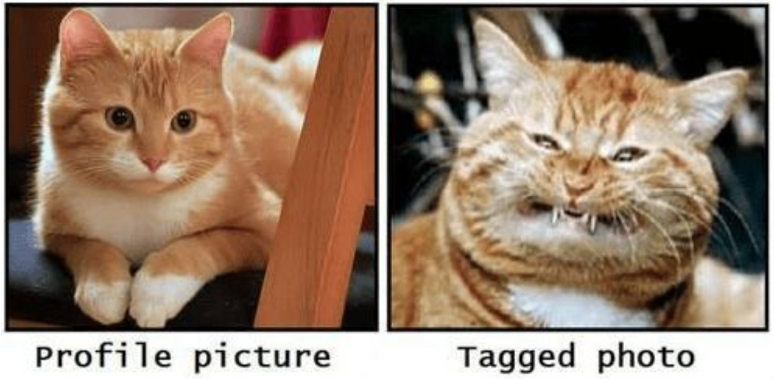
Supervised ML: Use algorithms to learn “model”

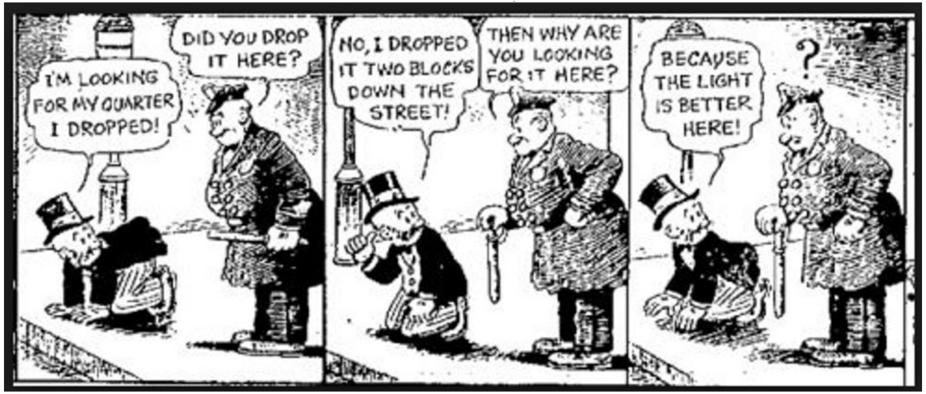
$$\min_{\mathbf{x}} \mathrm{Error}\left(\mathbf{y}, f(\mathbf{a}'\mathbf{x})\right)$$

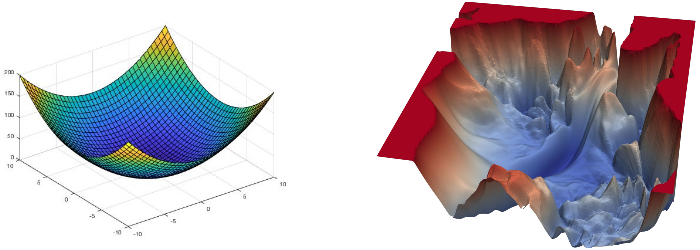
Academic theory vs industrial practice

Conventional wisdom in ML until 2010:

Simple models + simple errors

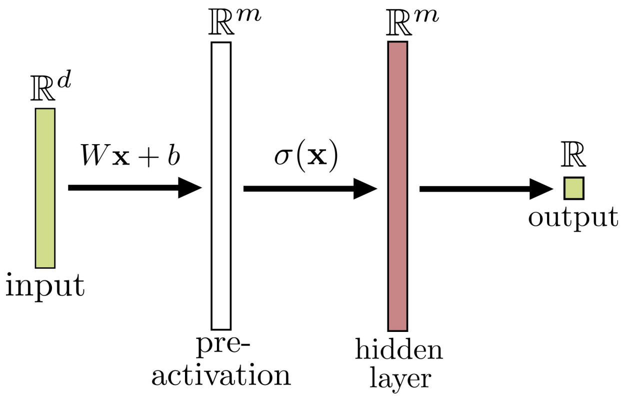




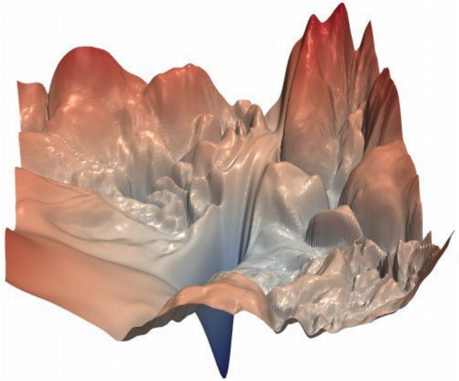


optimization landscapes

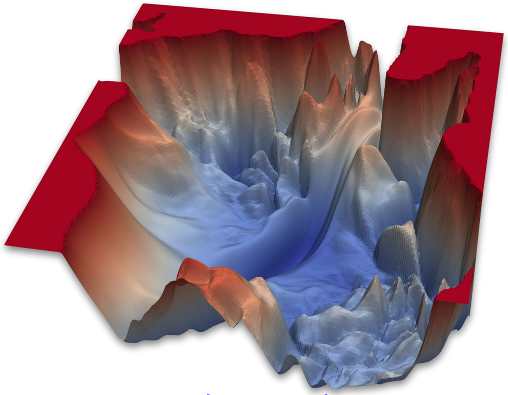
Enter neural networks: Universal approximation



$$f(\mathbf{x}; \beta, W, b) = \beta^T \sigma(W\mathbf{x} + b)$$



real function



optimization landscape

on a hypercube

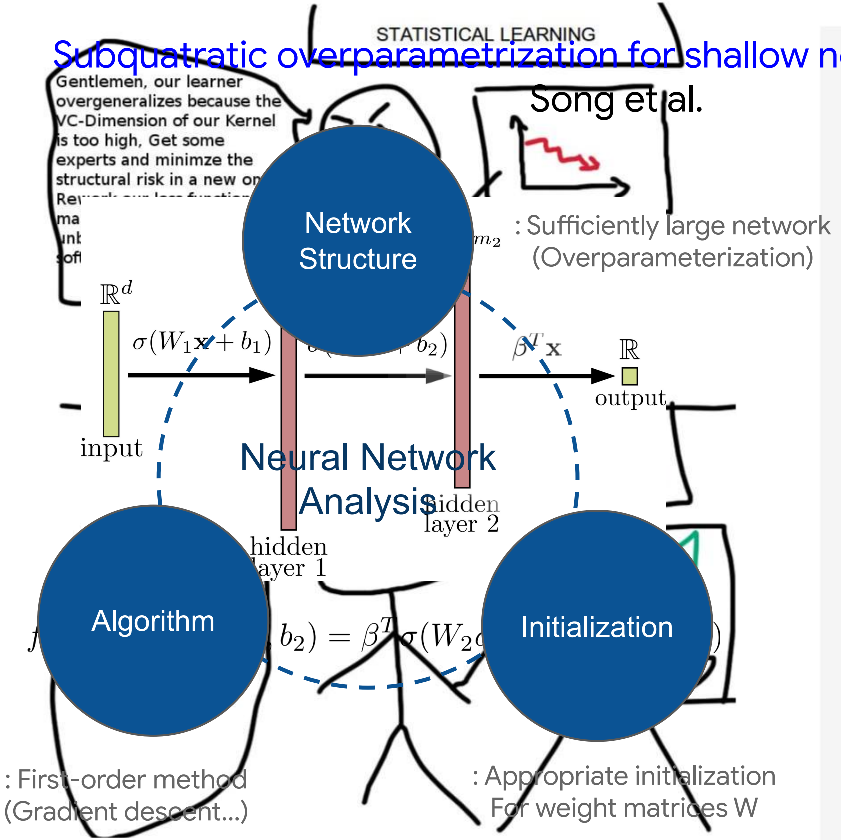
Challenges:

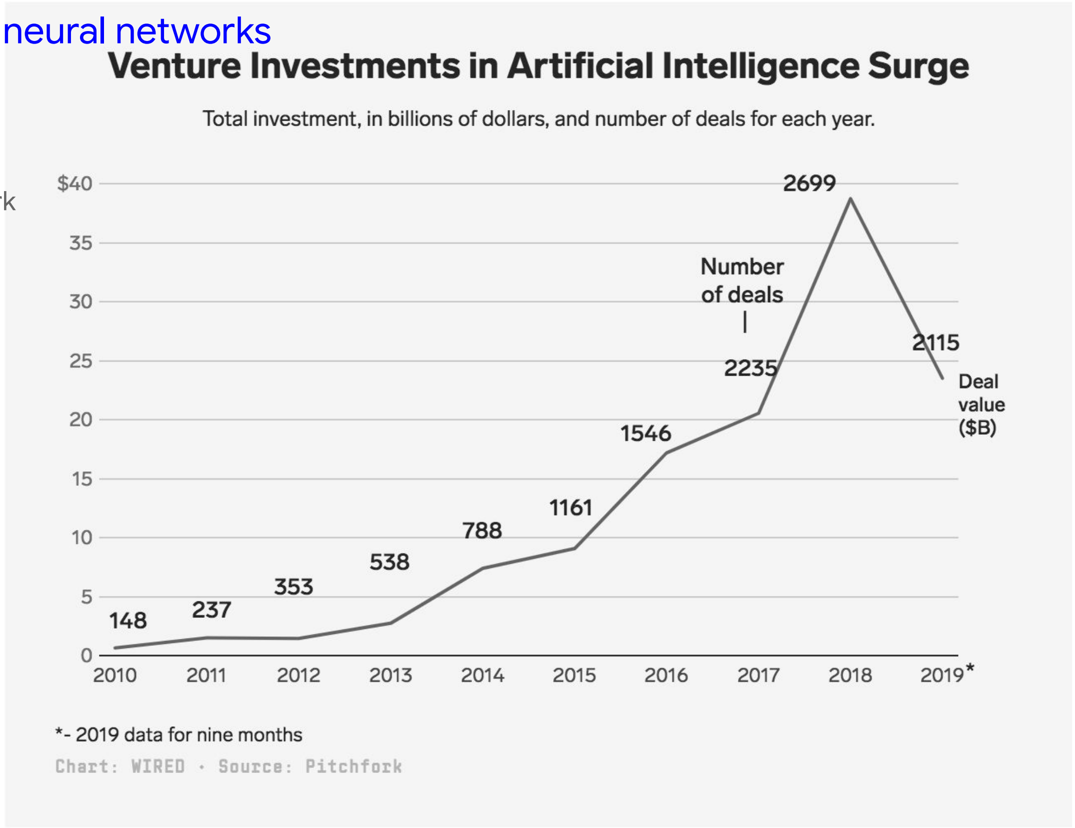
1. too big to optimize!
2. did not have enough data
3. could not find the optimum via algorithms

Multilayer neural networks: Tractable & nearly universal

Subquatratic overparametrization for shallow neural networks

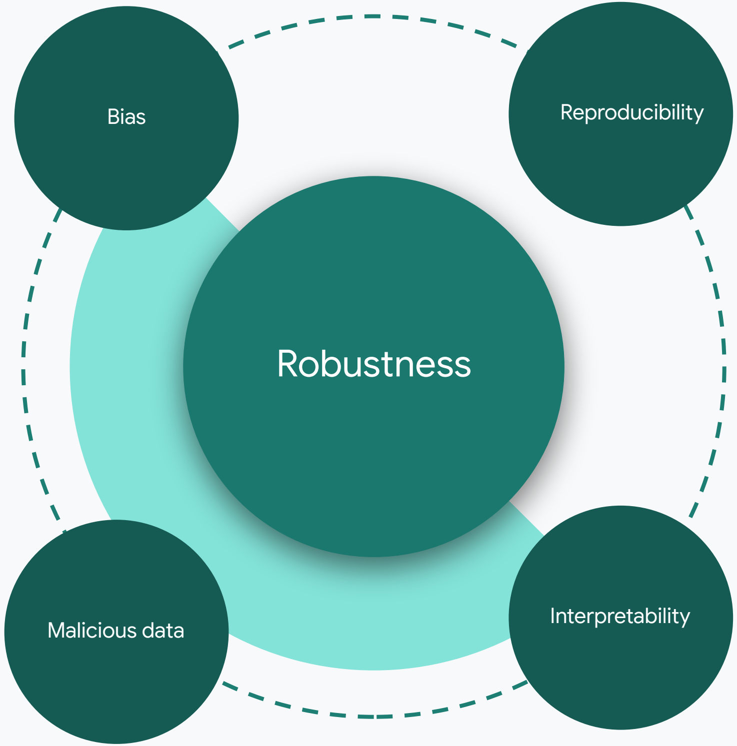
Song et al.





Weaknesses

A SWOT Analysis




S
Strengths
W
Weaknesses
O
Opportunities
T
Threats

Robustness


rifle
shield, buck
revolver, si

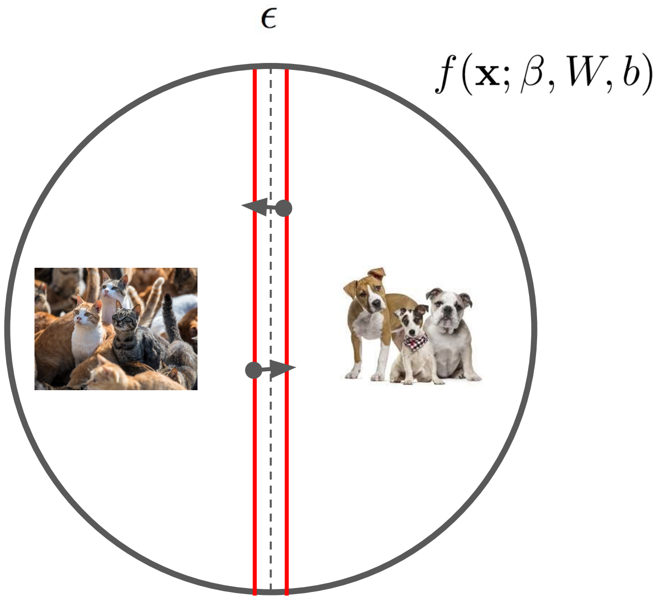

THIS IS FINE.


Self-Driving

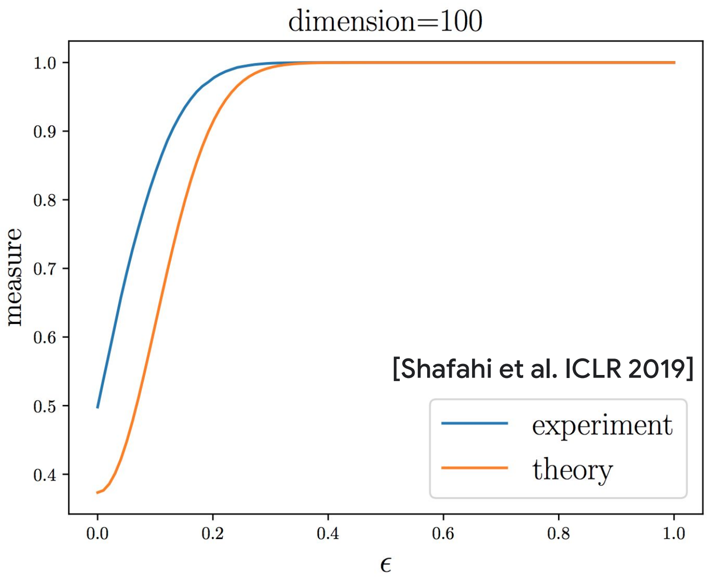
Robustness is an active research area

- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv e-prints, page arXiv:1512.03385.
- Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. arXiv e-prints, page arXiv:1608.06993.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In International Conference on Learning Representations.
- Raghunathan, A., Steinhardt, J., and Liang, P. S. (2018). Semidefinite relaxations for certifying robustness to adversarial examples. Neurips.
- Wong, E. and Kolter, Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. ICML.
- Madry, Aleksander and Makelov, Aleksandar and Schmidt, Ludwig and Tsipras, Dimitris and Vladu, Adrian. Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR.
- Huang, X., Kwiatkowska, M., Wang, S., and Wu, M. (2017). Safety verification of deep neural networks. Computer Aided Verification.

Adversarial examples are inevitable!



$$f(\mathbf{x}+\epsilon) \simeq f(\mathbf{x}) + \langle \epsilon, \nabla f(\mathbf{x}) \rangle$$
$$|f(\mathbf{x}+\epsilon) - f(\mathbf{x})| \leq \|\epsilon\| \|\nabla f(\mathbf{x})\|$$

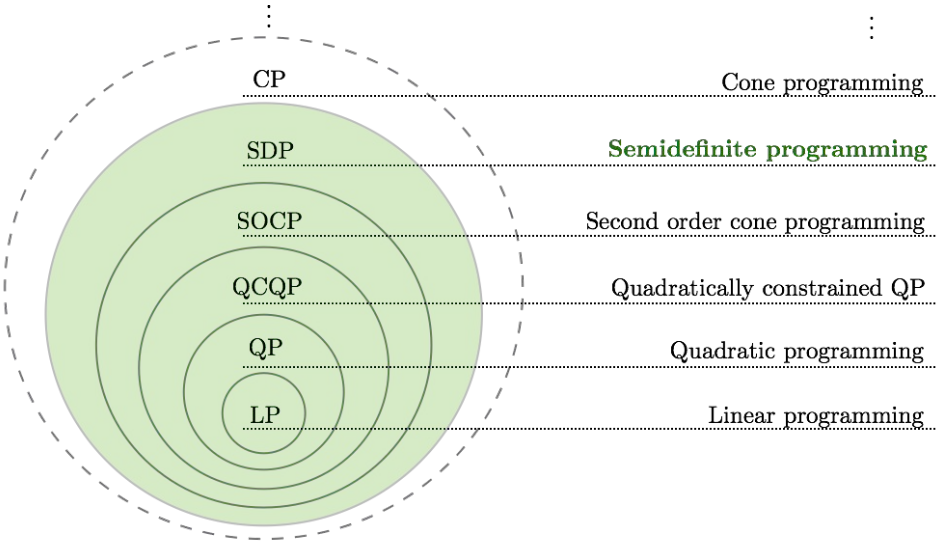


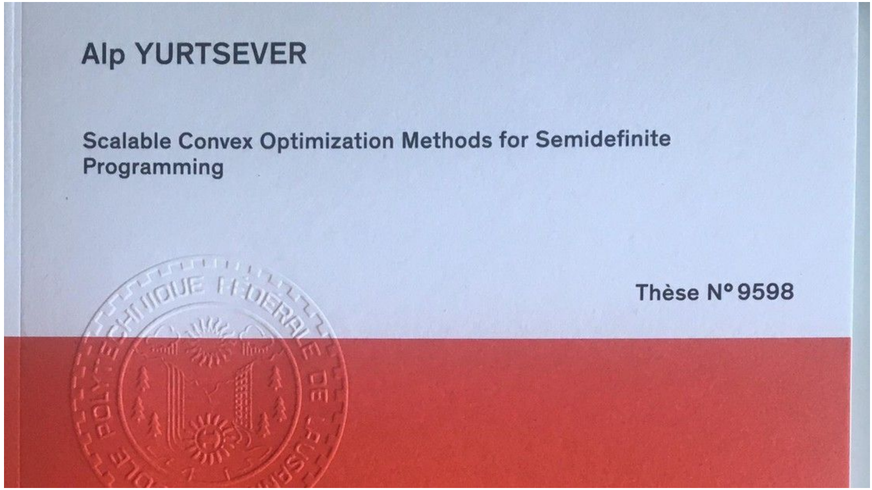
- Concentration-of-measure phenomenon
- Lipschitz constant is important

Progress towards robustness

$$|f(\mathbf{x}+\epsilon) - f(\mathbf{x})| \leq \|\epsilon\| \underbrace{\|\nabla f(\mathbf{x})\|}_{\sup_{\mathbf{x}} \to L(f)}$$

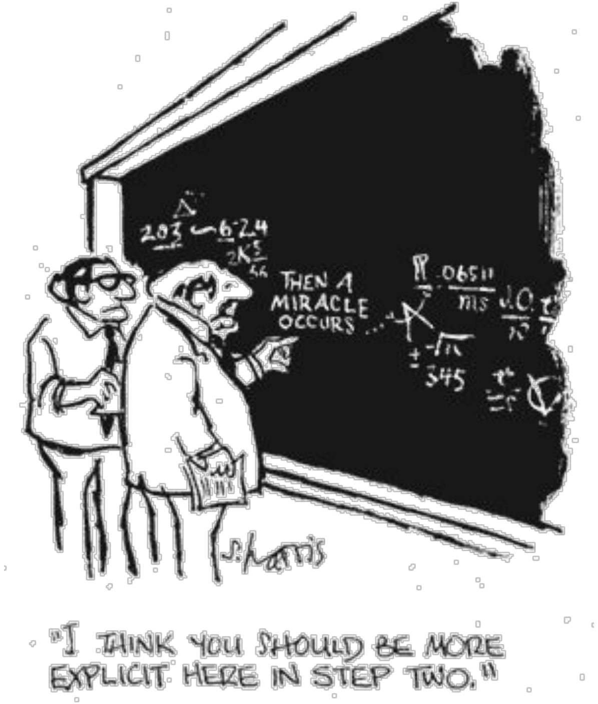
NP-hard for NNs [Scaman et al. NeurIPS 2018]

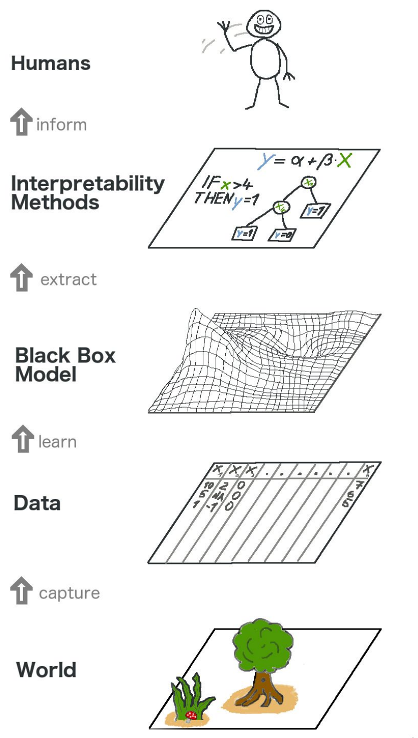


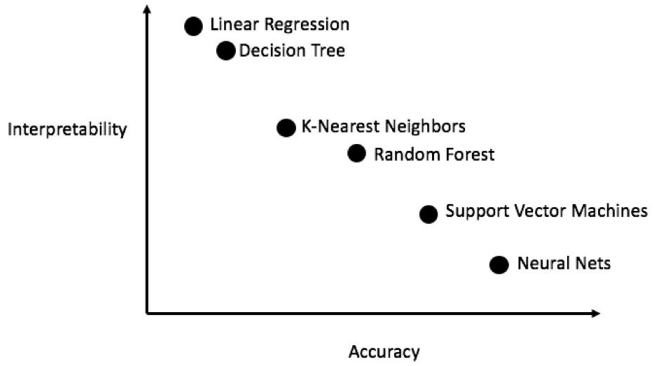


Lipschitz Constant Estimation of Neural Network via Sparse Polynomial Optimization.
Latorre, Fabian and Rolland, Paul and Cevher, Volkan. ICLR 2020.

Interpretability







Interpretability in ML is an active research field

- Baehrens, David and Schroeter, Timon and Harmeling, Stefan and Kawanabe, Motoaki and Hansen, Katja and Mueller, Klaus-Robert. Simonyan, Karen and Vedaldi, Andrea and Zisserman, Andrew. How to Explain Individual Classification Decisions. JMLR 2010.

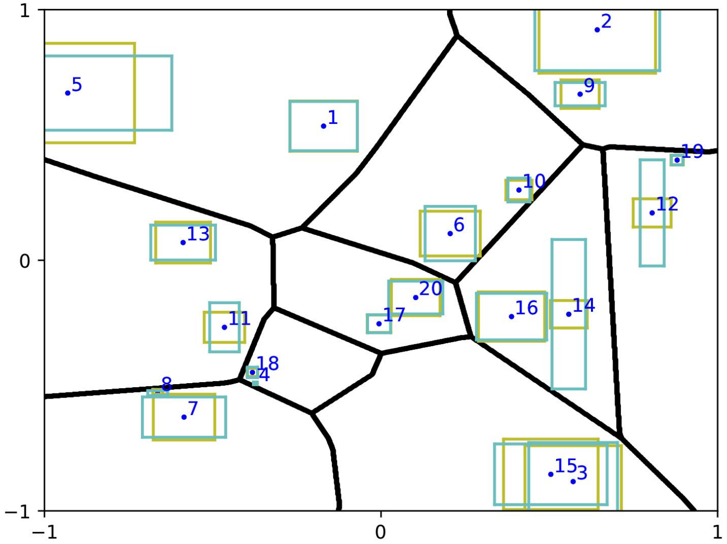
- Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv e-prints. arXiv:1312.6034. 2013.

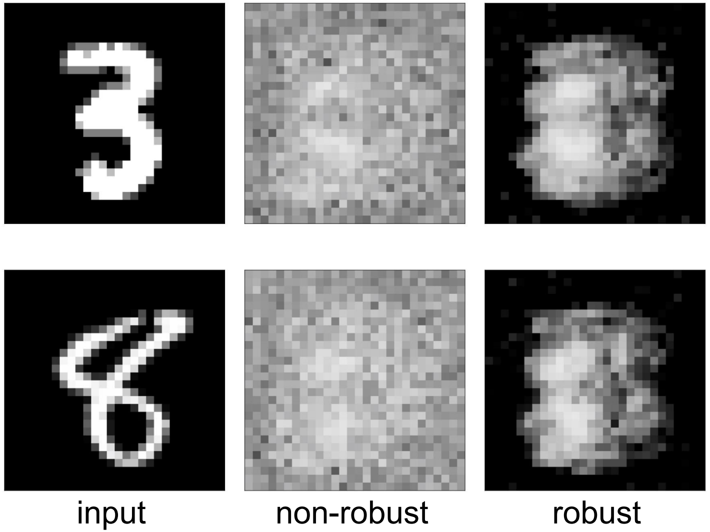
- Ribeiro, Marco and Singh, Sameer and Guestrin, Carlos. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. KDD 2016.

- Sundararajan, Mukund and Taly, Ankur and Yan, Qiqi. Axiomatic Attribution for Deep Networks. ICML'17.

- Shrikumar, Avanti and Greenside, Peyton and Kundaje, Anshul. Learning Important Features Through Propagating Activation Differences. ICML'17.

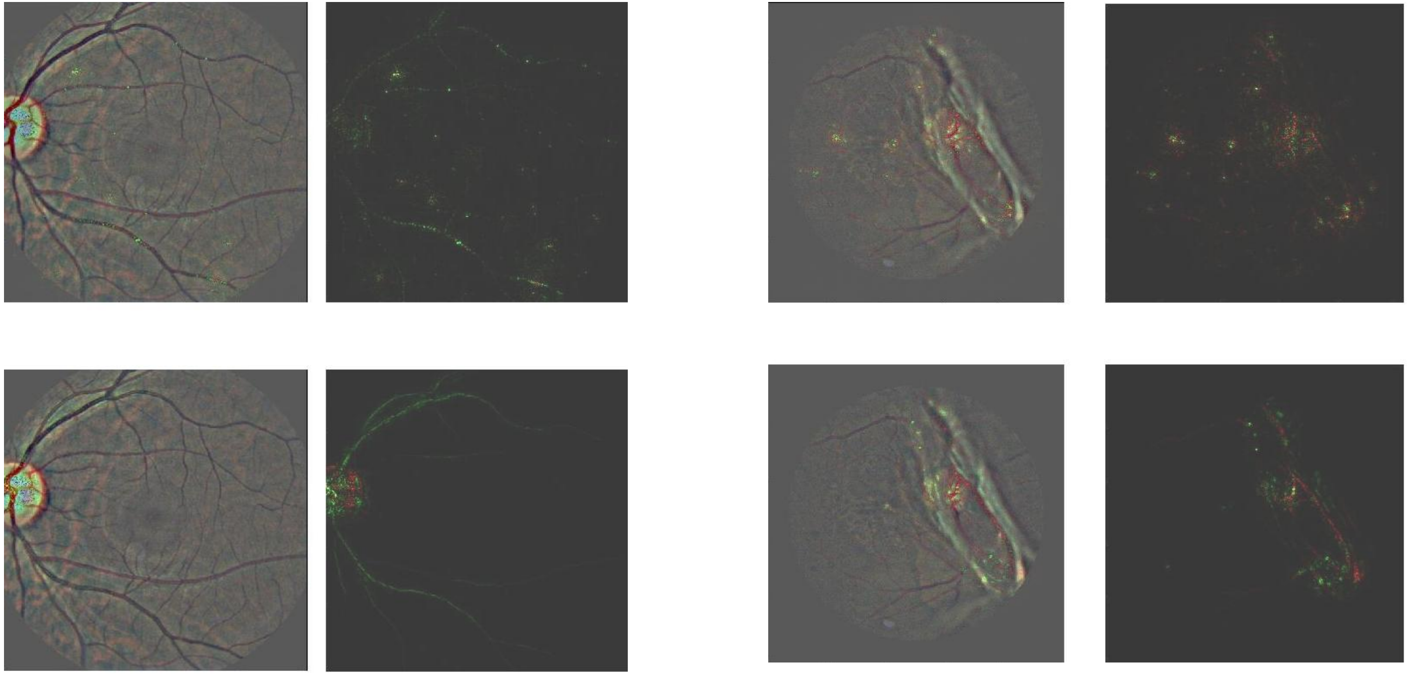
A robustness & interpretability result





On Certifying Non-Uniform Bounds against Adversarial Attacks.
Liu, Chen and Tomioka, Ryota and Cevher, Volkan. ICML'19.

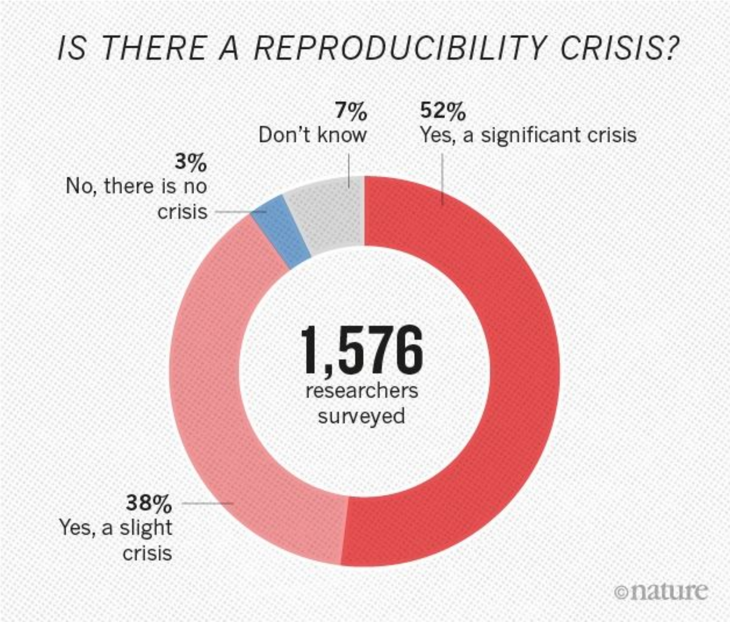
Further evidence: Robust training <> interpretability



Robust fundus classification & dataset bootstrapping via interpretable features
Krawczuk et al.

S Strengths | W Weaknesses | O Opportunities | T Threats

Reproducibility



Grad student descent

Posted on 2014/01/25 by sciencedryad

SORT BY **BEST**

Brudaks 224 points · 1 year ago

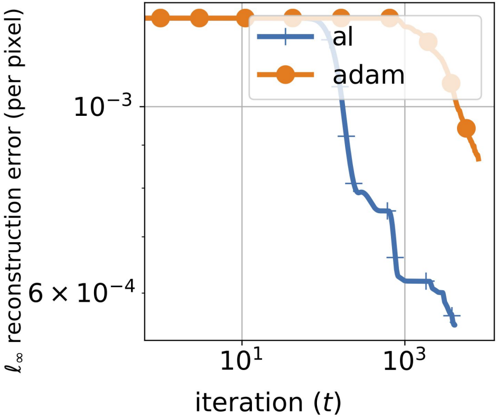
A popular method for designing deep learning architectures is GDGS (gradient descent by grad student).

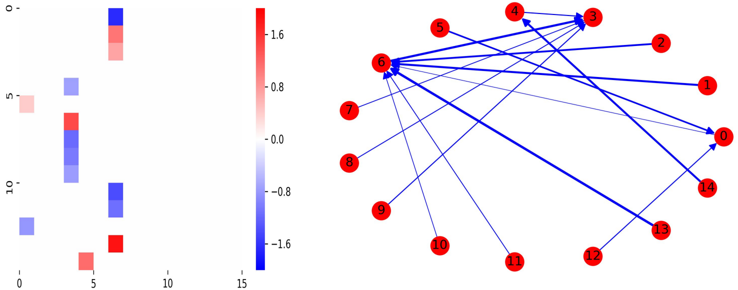
This is an iterative approach, where you start with a straightforward baseline architecture (or possibly an earlier SOTA), measure its effectiveness; apply various modifications (e.g. add a highway connection here or there), see what works and what does not (i.e. where the gradient is pointing) and iterate further on from there in that direction until you reach a (local?) optimum.

Share Report Save

Reproducibility challenge: Non-convexity

Lagrangian perspective: New theory for nonlinear optimization with nonlinear constraints





iAL

Sahin M. F. et. al. [NeurIPS 2019]

ADMM

Latorre F. et. al. [NeurIPS 2019]

AL^2

Eftekhari A. et. al. [Under review]

Extending reproducibility via universality in convex optimization

One algorithm to rule them all!

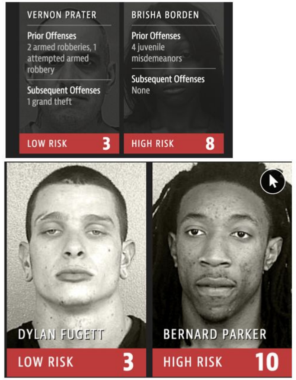
Smooth	$\mathcal{O}\left(1/k^2\right)$	✓	✓	
Stochastic	$\mathcal{O}\left(1/\sqrt{k}\right)$		✓	
Nonsmooth	$\mathcal{O}\left(1/\sqrt{k}\right)$	✓	✓	✓
Strongly convex	$\mathcal{O}\left(\rho^k\right), \rho < 1$			✓

k is the iteration counter.

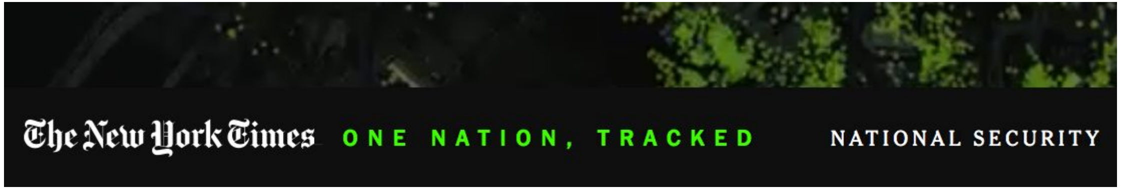
✓ Universal primal-dual, Yurtsever et al.
✓ UniXGrad, Kavis et al. Accelegrad, Levy et al.
✓ Random extrapolation, Alacaoglu et al.

Many other weaknesses

1. Bias
2. Malicious data
3. Privacy
4. ...



I am Tay



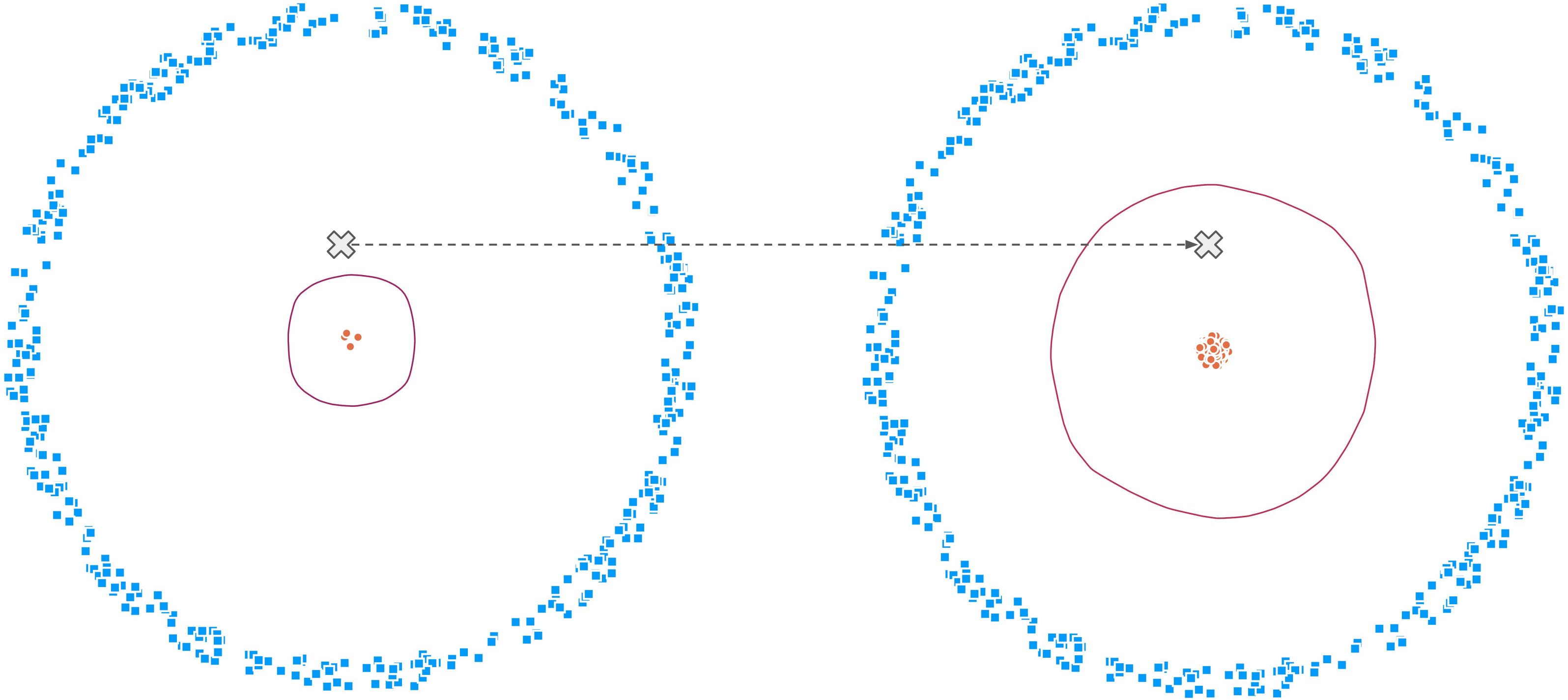
Opinion | THE PRIVACY PROJECT

Twelve Million Phones, One Dataset, Zero Privacy

By Stuart A. Thompson and Charlie Warzel

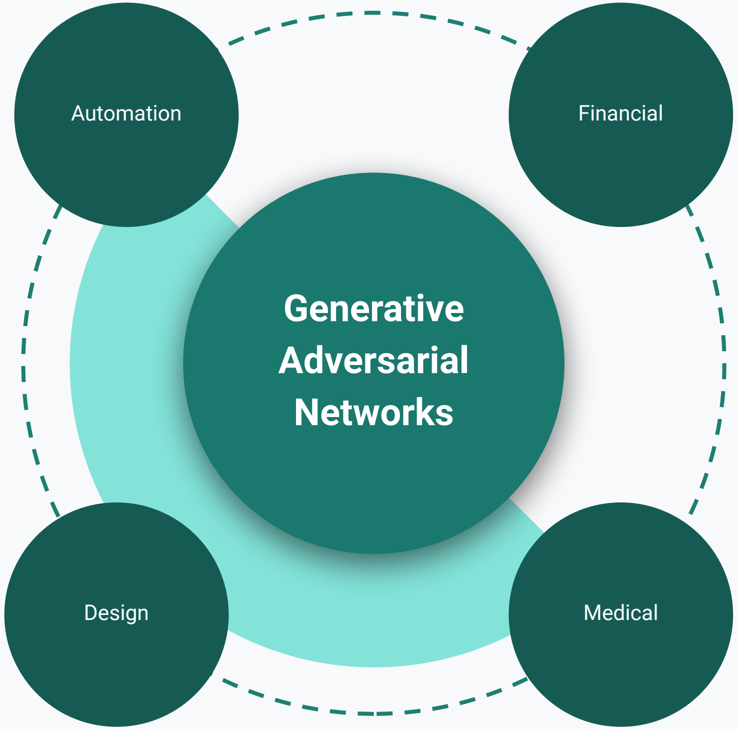
DEC. 19, 2019

A geometric perspective on bias



Opportunities

A SWOT Analysis



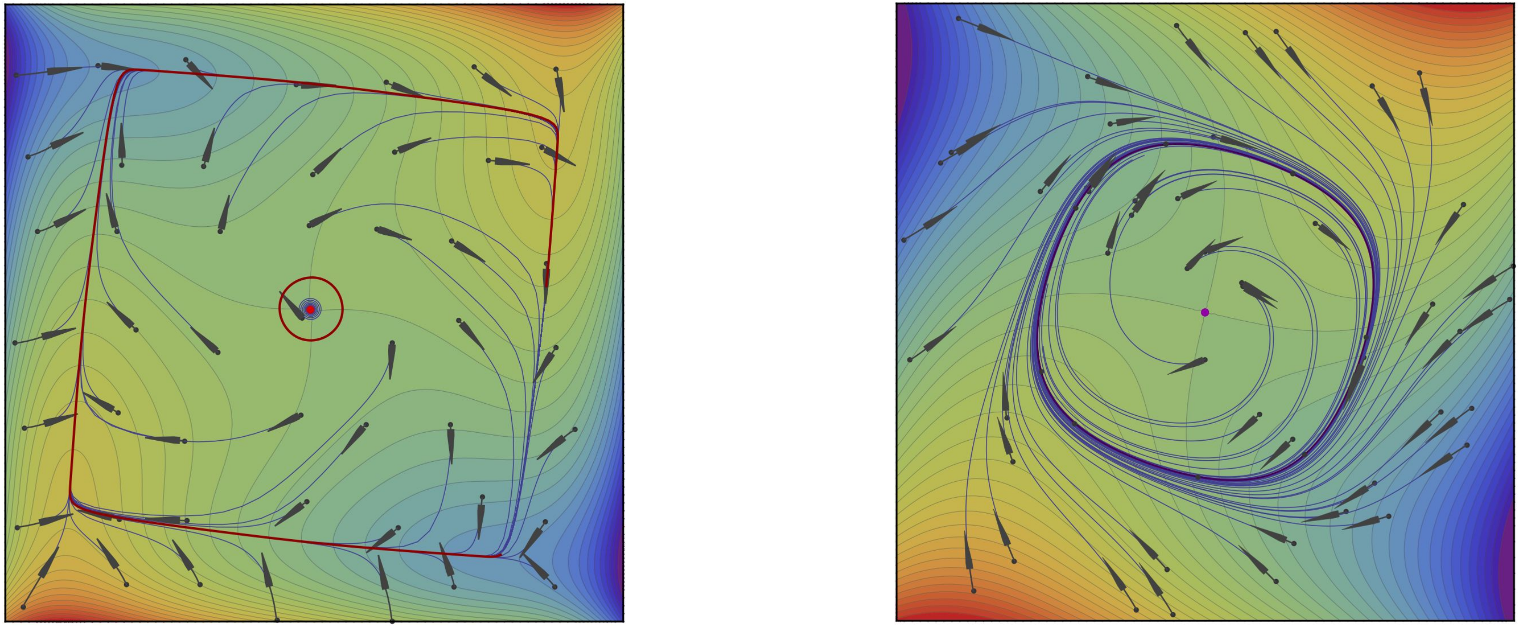
Generative Adversarial Networks

Progressive Growing of GANs for Improved Quality, Stability, and Variation
Karras et al. [ICLR 2018]

High-Fidelity Image Generation With Fewer Labels
Lucic M, Tschannen M*, Ritter M*, Zhai X, Bachem O,* Sylvain S [2019]

Challenge: Limit cycles (minimax)

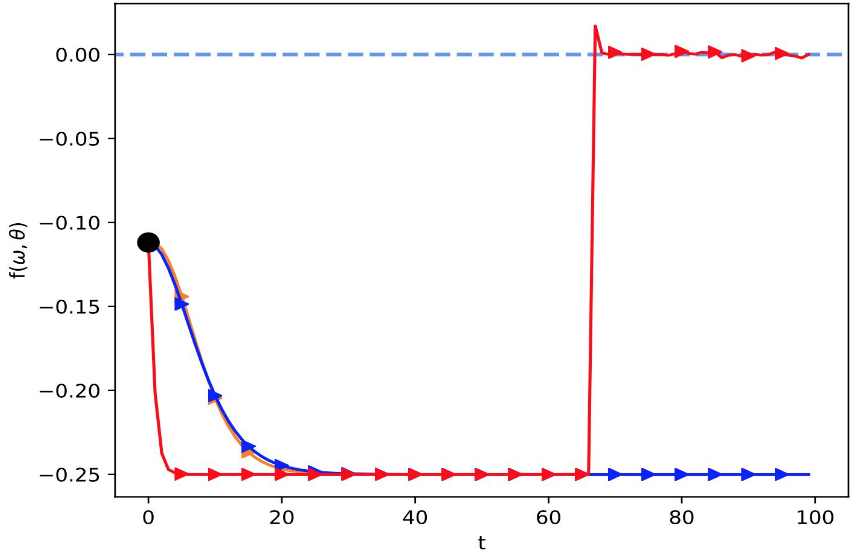
$$\min_{\theta \in \mathbb{R}} \max_{\omega \in \mathbb{R}} f(\theta, \omega) = \theta\omega + \phi(\theta) - \phi(\omega), \quad \phi \text{ non-convex}$$

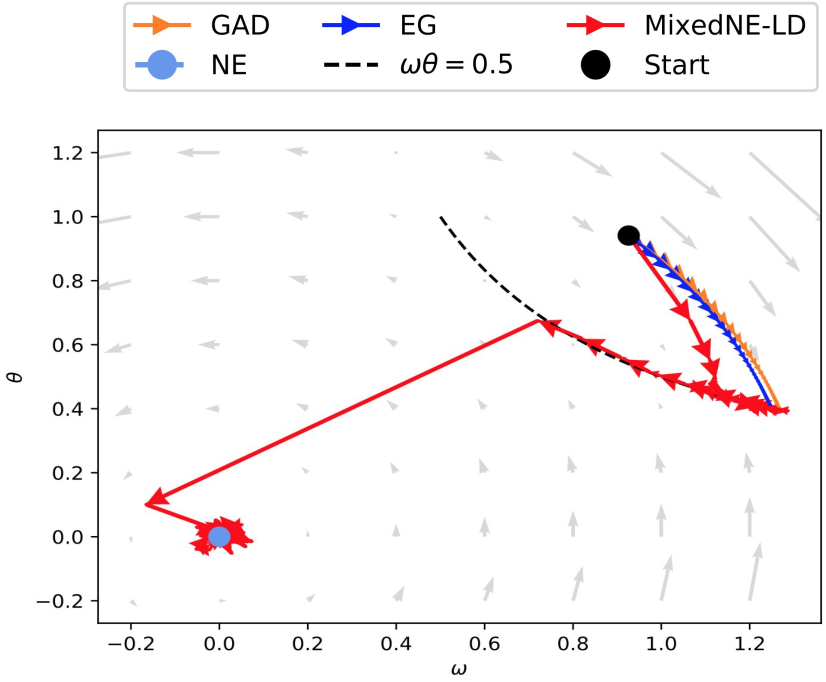


The limits of min-max optimization algorithms: Convergence to spurious non-critical sets, Hsieh, Mertikopoulos, and Cevher 2020.

Training GANs via mixed Nash equilibria (minimax)

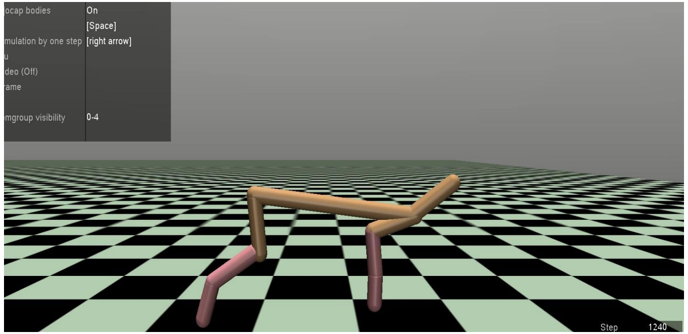
$$\max_{\theta\in[-2,2]} \min_{\omega\in[-2,2]} f(\theta,\omega) = \theta^2\omega^2 - \theta\omega$$

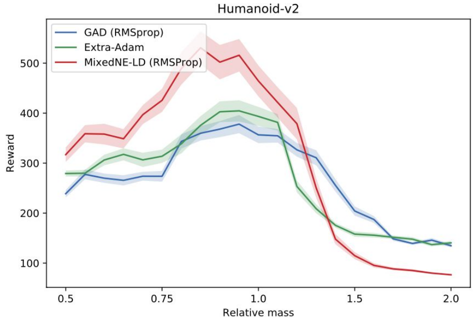


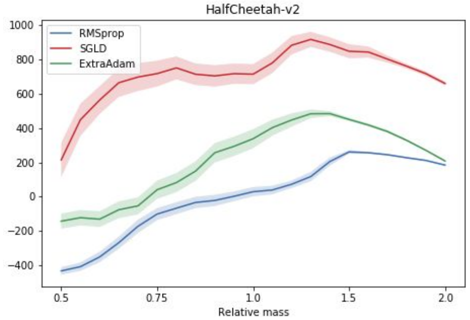


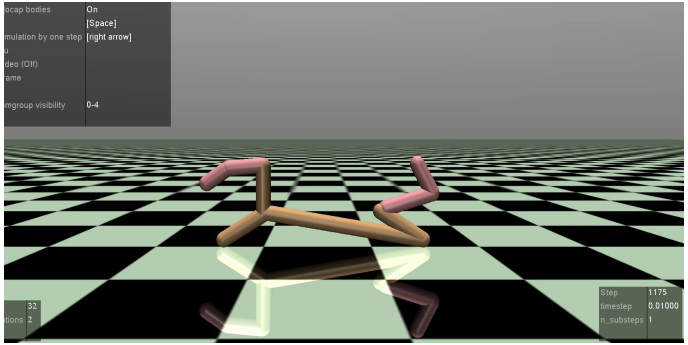
Finding Mixed Nash Equilibria of Generative Adversarial Networks. Hsieh et al. ICML 2019

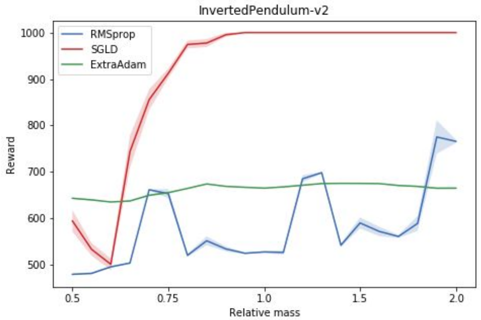
Minimax formulations and robust RL

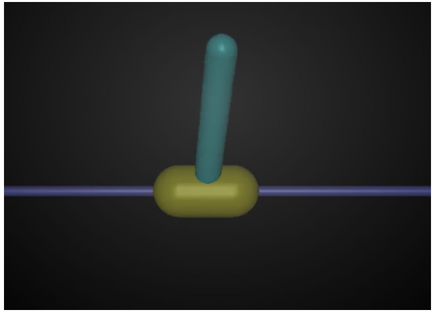






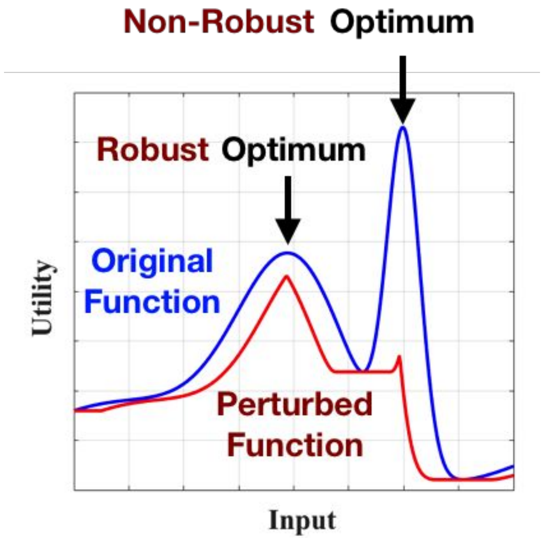






Robust Reinforcement Learning with Langevin Dynamics. Kamalaruban et al.

Minimax formulations and robust BO

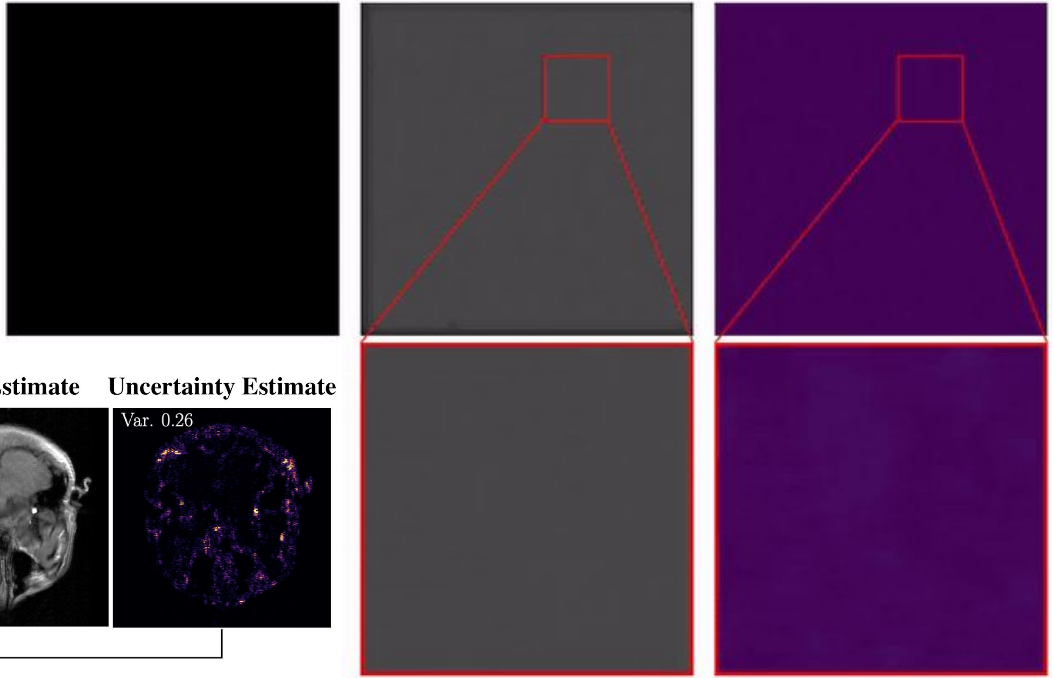


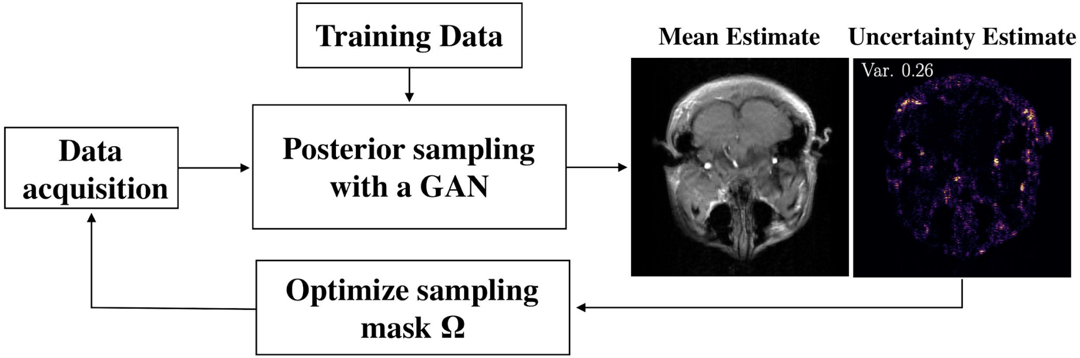
$$\arg\max_{\mathbf{x}\in D} \min_{\boldsymbol{\delta}\in\Delta_\epsilon(\mathbf{x})} f(\mathbf{x}+\boldsymbol{\delta})$$

Adversarially robust Gaussian Process Optimization.
Bogunovic et al. NeurIPS 2018

New opportunities via GANs

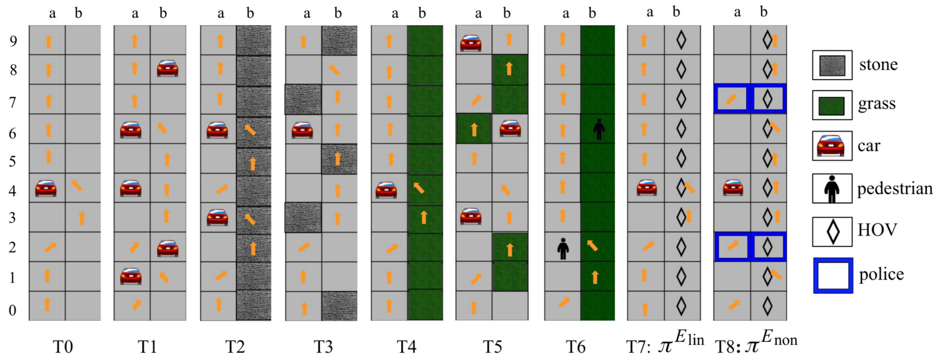
Closed loop deep Bayesian inversion:
Uncertainty driven acquisition for fast MRI.
Sanchez et al.

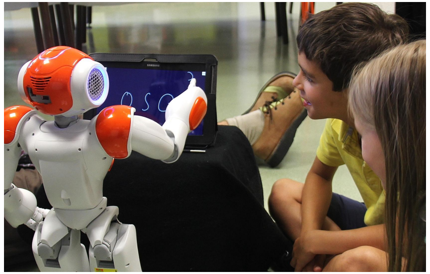




S Strengths W Weaknesses
O Opportunities T Threats

New opportunities in RL



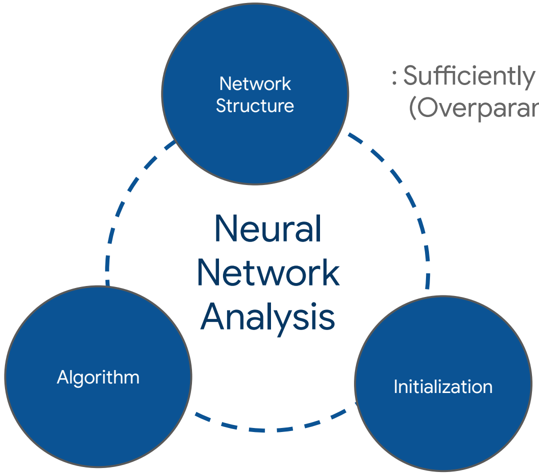


Interactive Teaching Algorithms for Inverse Reinforcement Learning. Kamalaruban et al. IJCAI 2019

Interaction-limited Inverse Reinforcement Learning. Troussard et al.

New opportunities in deep learning

Generalization <> Robustness

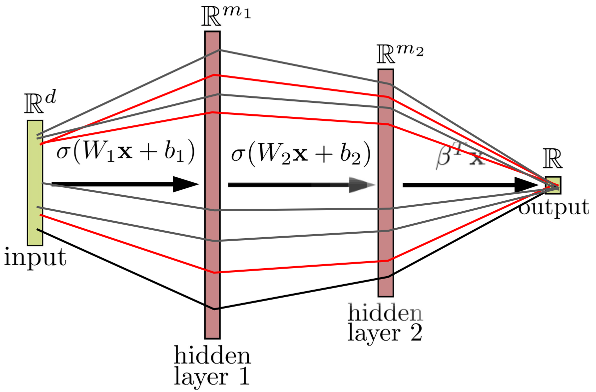


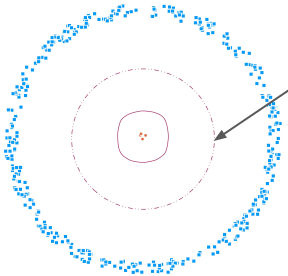
: Sufficiently large network (Overparameterization)

: First-order method
Stochastic gradient

: Appropriate initialization
For weight matrices W

Convergence of SGD for neural networks without heavy overparameterization. Song & Cevher





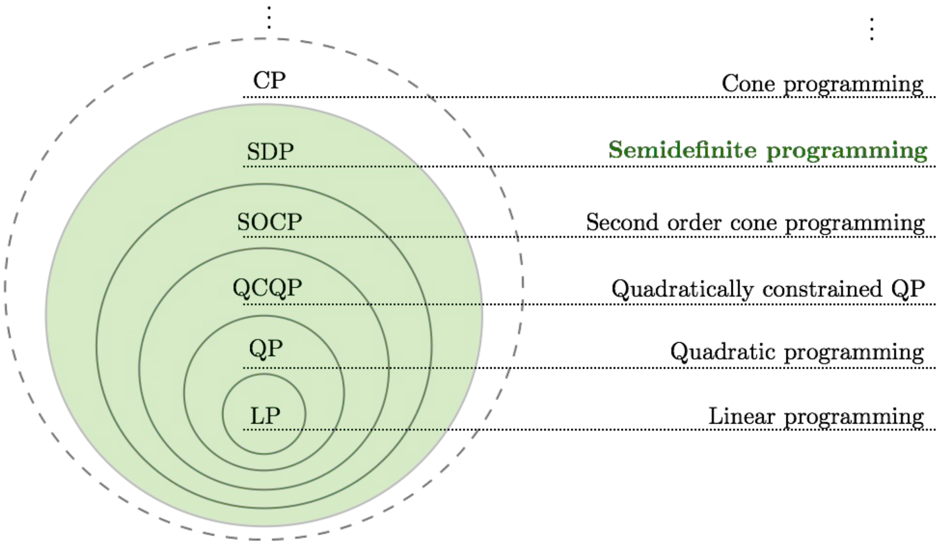
Lipschitz controlled polynomial

Efficient proximal mapping of the 1-path-norm of shallow networks. Latorre et al.

New opportunities in scalable optimization

Randomization <> Scalability

Towards stochastic SDP & LP's with stochastic constraints $\min_{x \in \mathcal{X}} \mathbb{E}\left[f(x,\xi)\right] \text{ s.t. } A(\xi)x = b(\xi) \text{ almost surely}$

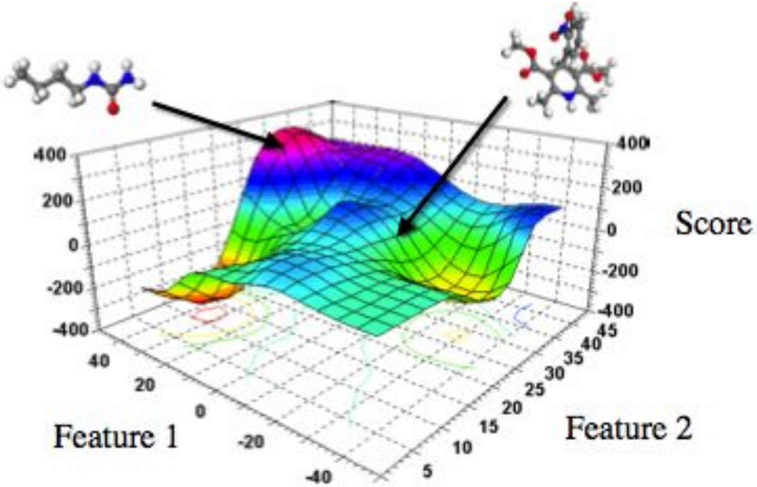


Ex: scalable solutions (sparsest) cut problems and their variants

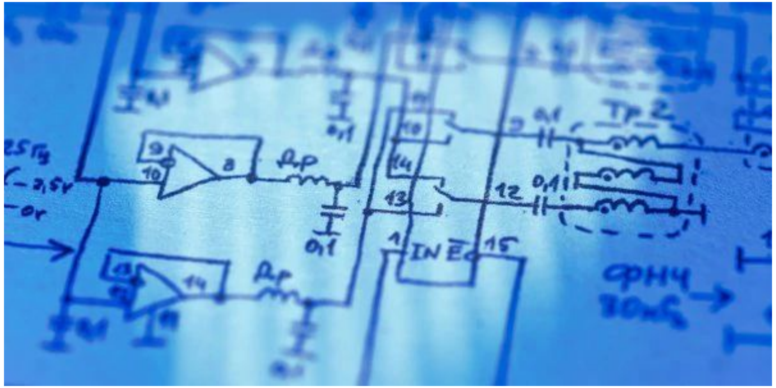
Ex: robustness certifications for NNs

Conditional gradient methods for stochastically constrained convex minimization. Vladarean et al.

New opportunities in engineering applications



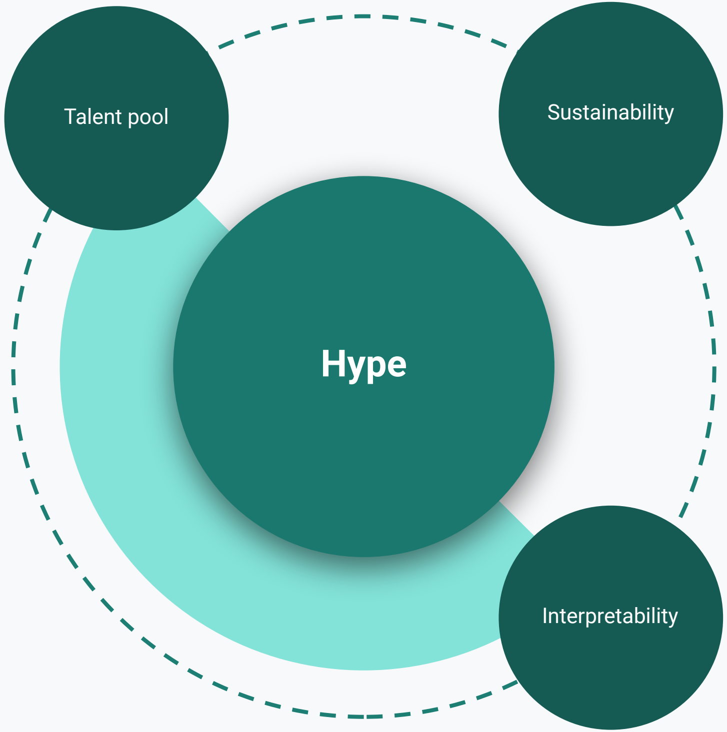
Chemical machine learning with kernels: The impact of loss functions. Van Nguyen et al. [Quantum Chemistry 2019]



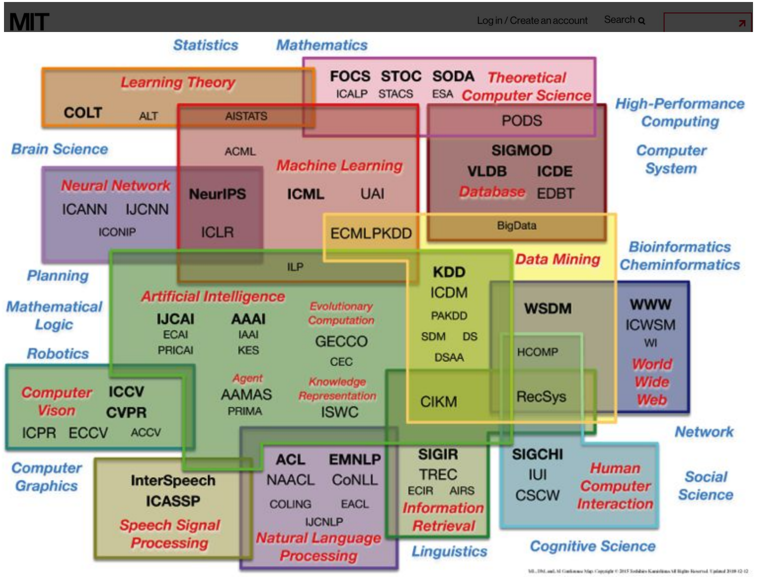
EDA Gym. Krawczuk et al.

Threats

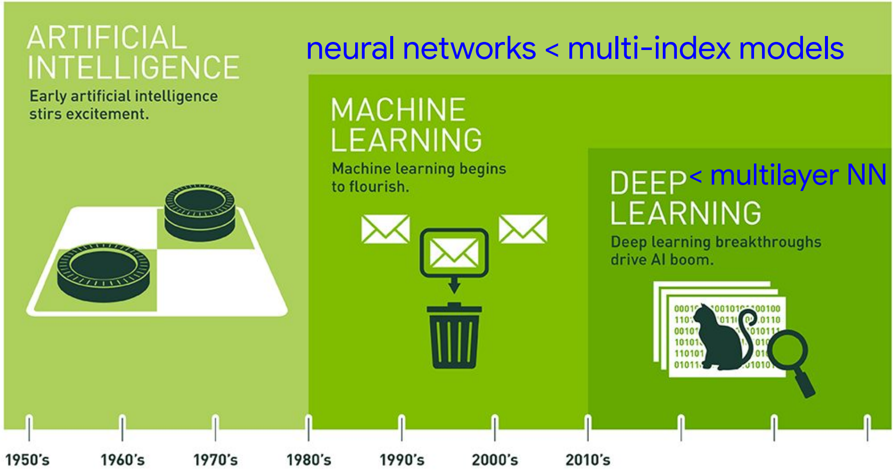
A SWOT Analysis



The AI hype vs the ML revolution

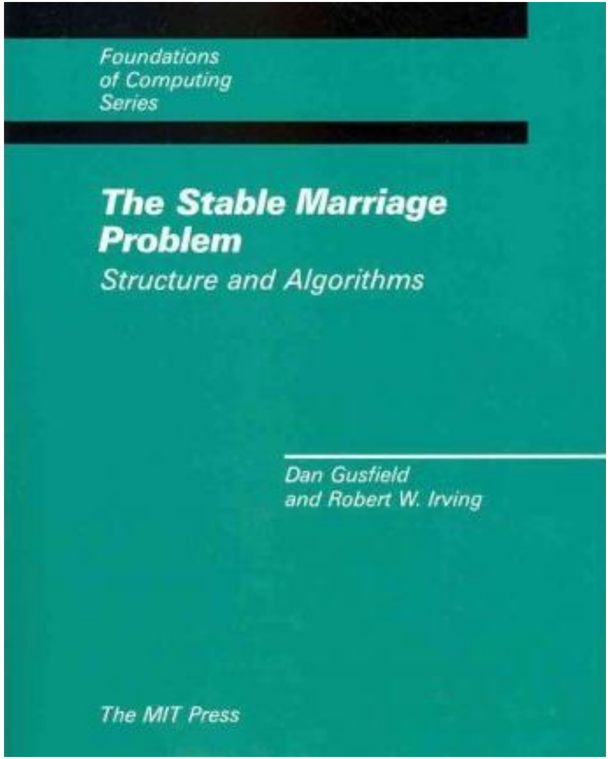


A surprising number of firms are jumping on the artificial-intelligence bandwagon—without actually investing in any AI.



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Talent pool: Missing the top talent vs the needed talent



Sustainability:

S Strengths | W Weaknesses | O Opportunities | T Threats

The estimated costs of training a model

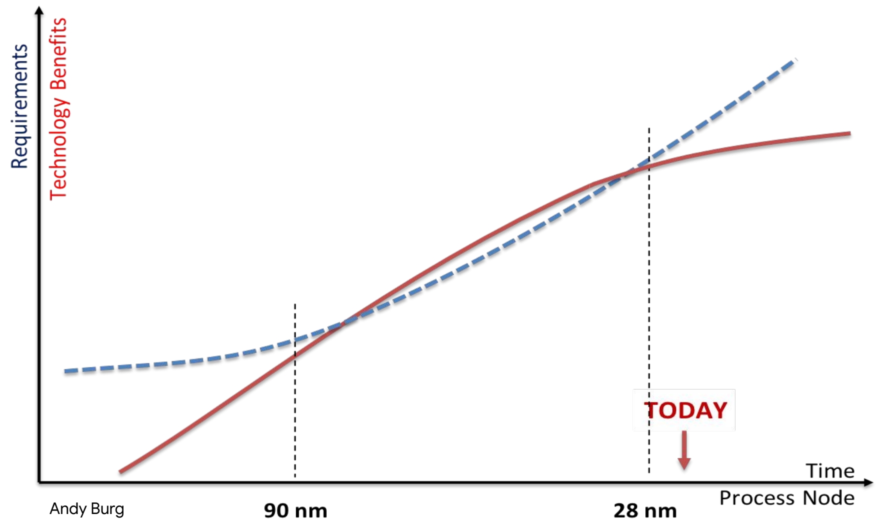
	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	\$41-\$140
Transformer (213M parameters)	Jun, 2017	201	192	\$289-\$981
ELMo	Feb, 2018	275	262	\$433-\$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	\$3,751-\$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	\$942,973-\$3,201,722
GPT-2	Feb, 2019	-	-	\$12,902-\$43,008

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

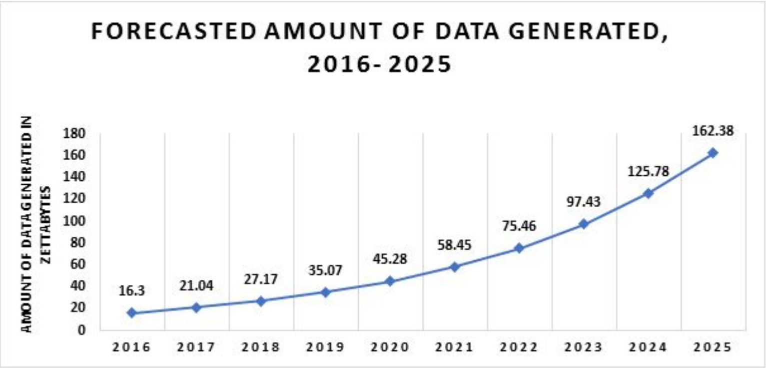
Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

Sustainability:

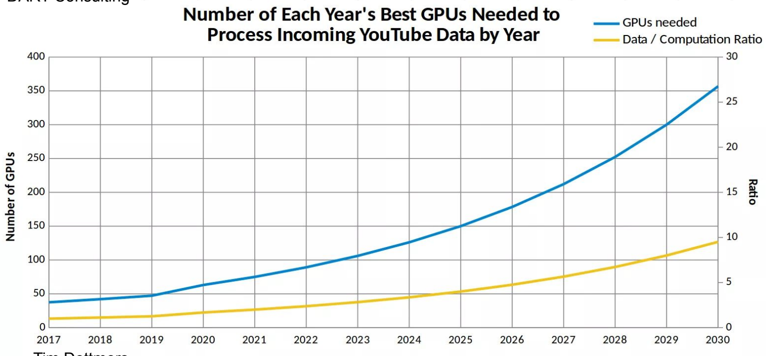
Dennard scaling & Moore’s law vs Growth of data



Andy Burg



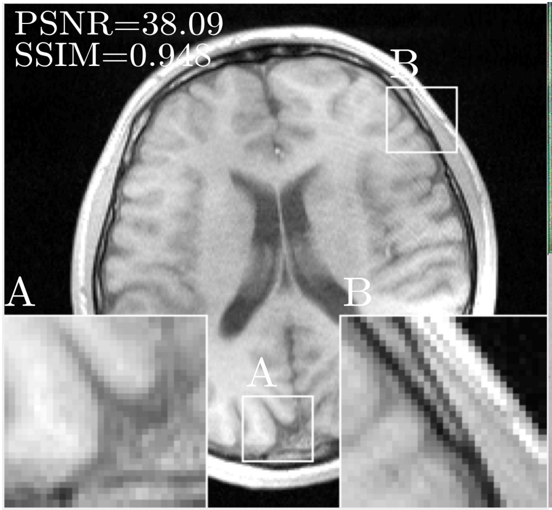
DART Consulting

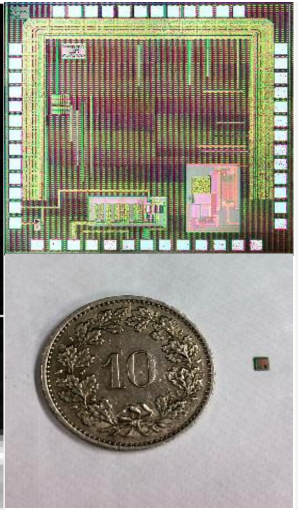


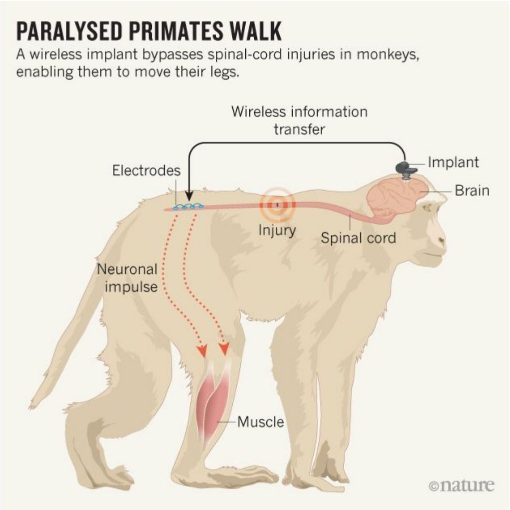
Tim Dettmers

Sustainability:

Energy constraints / Time constraints







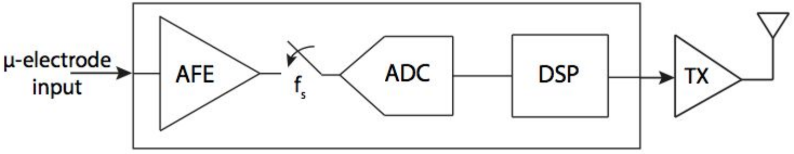
Learning-based compressive sensing + hardware design.

Baldassarre et al., Gozcu et al., Aprile et al. [IEEE TMI, IEEE TSP, IEEE CnS, IEEE TCAS]

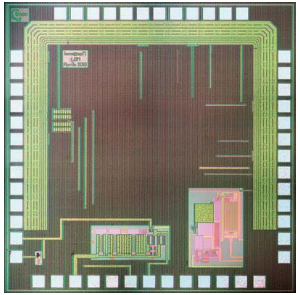
IBM Thesis Award 2019

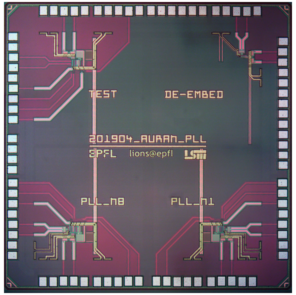
Sustainability:

Energy constraints of recording neural dat

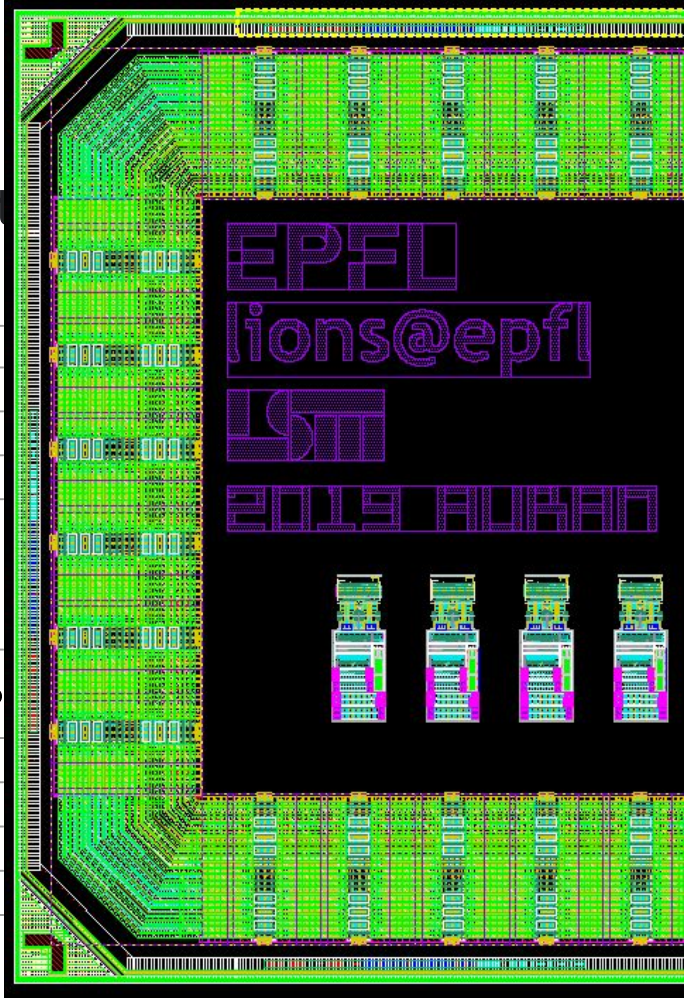


> 30 dB quality
AFE + ADC
DSP
TX





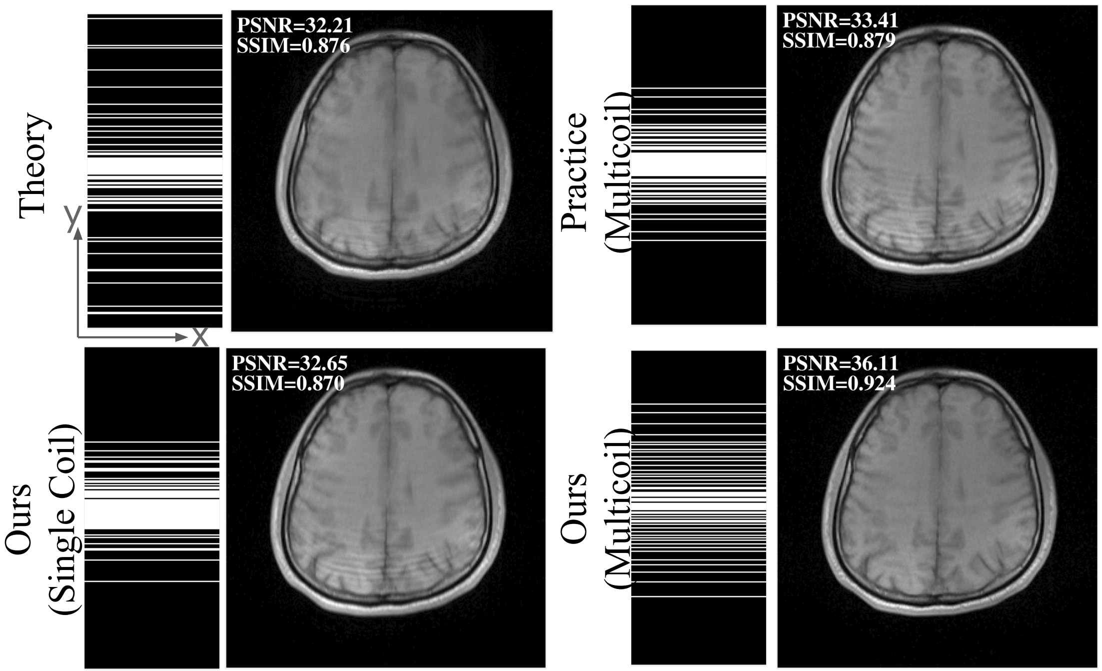
Metho
LBCS
SHS
BERN
MCS



Sustainability:

Time constraints of MRI

- Accelerate the MRI scan 5 times.
- Pick the most relevant data only for your method.



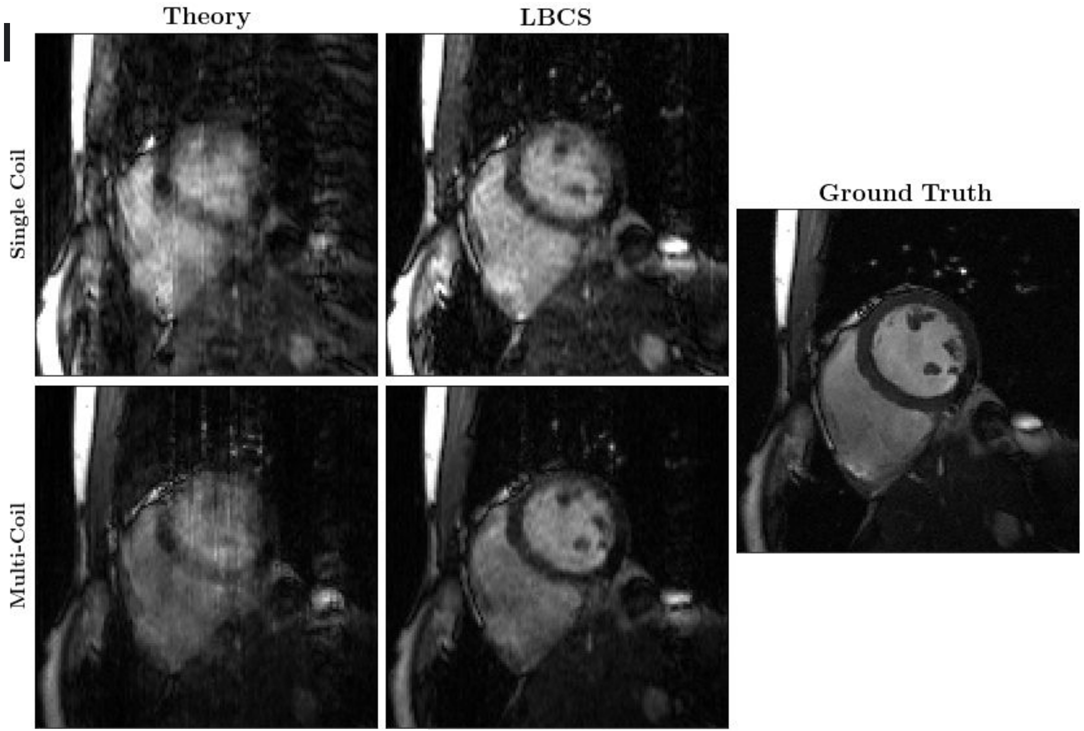
Learning-based compressive MRI. Gözcü B., et al [IEEE TMI 2018]
Rethinking Sampling in Parallel MRI: A Data-Driven Approach. Gözcü B. et al. [EUSIPCO 2019]

Sustainability:

Time constraints of MRI

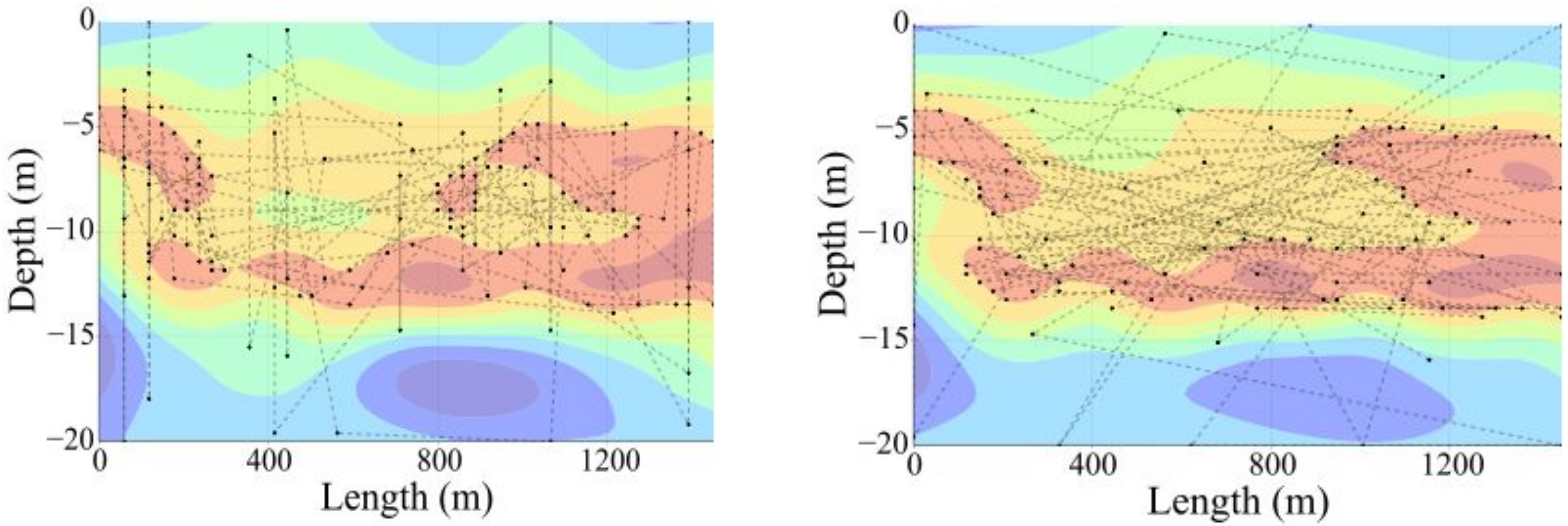
- Time drastically increases the dimensionality of data

- Reduce computations by a factor 200: from a month to 4 hours without losing performance.



Scalable learning-based sampling optimization for compressive dynamic MRI.
Sanchez T., et al. [IEEE ICASSP 2020]

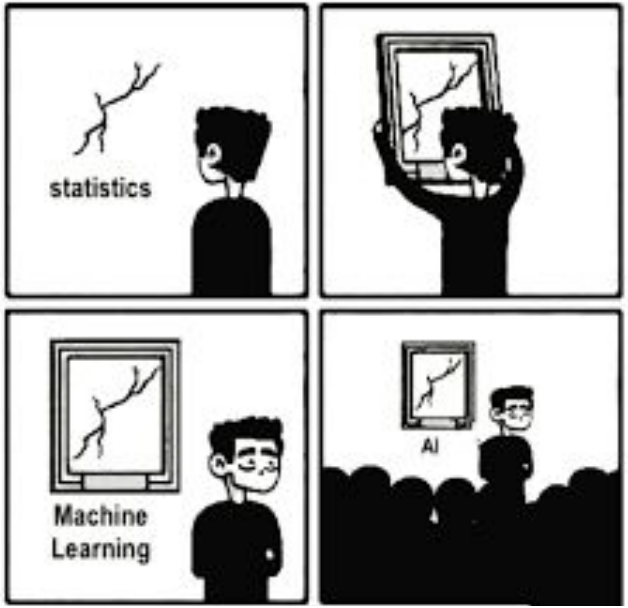
Sustainability: Resource constrained optimization



Truncated variance reduction: A unified approach to Bayesian optimization and level-set estimation.
Bogunovic et al. NIPS 2017

Conclusions

- Are you wiser?
  - time-data-power and other trade offs
- Existential threats = “Opportunities”
  - talk to me offline
- ML - AI: Mathematical understanding
  - Hype protection



volkan.cevher@epfl.ch
https://lions.epfl.ch
Twitter: @CevherLIONS