Artificial Intelligence and Mathematics

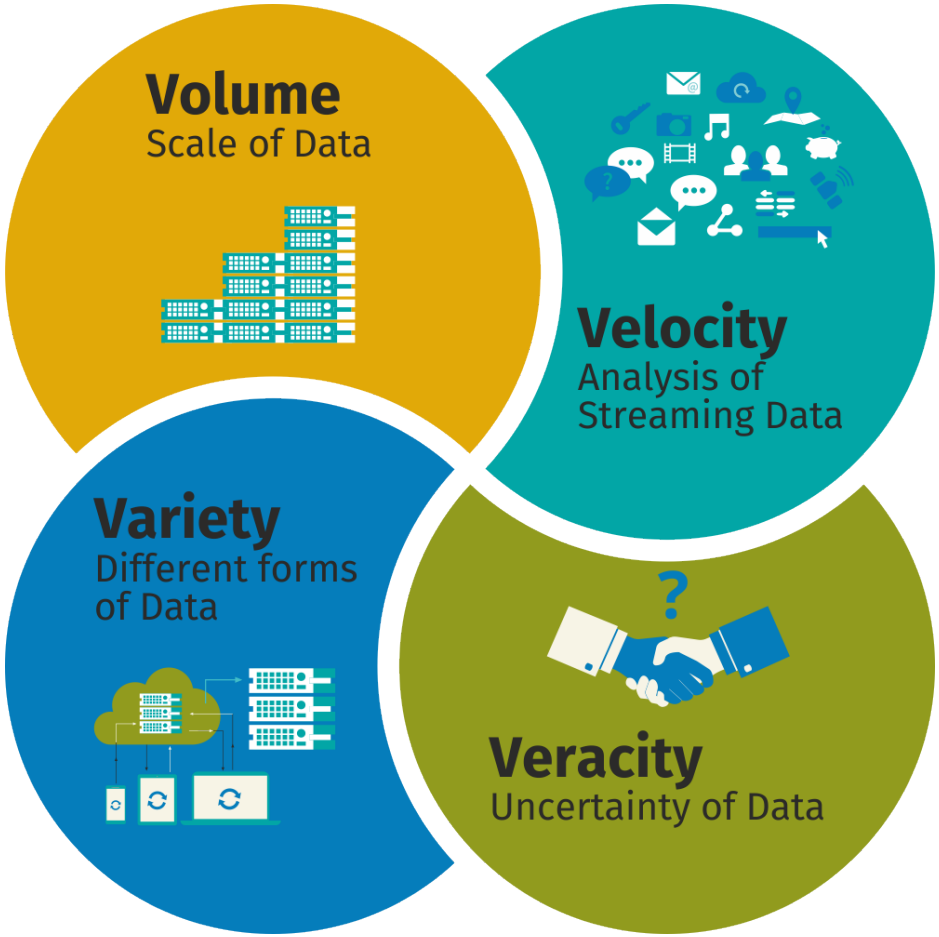
Bangti Jin

University College London

joint colloquium, Faculty of Sciences
Chinese University of Hong Kong, March 12, 2021

Outline

The Four V's of Big Data



https://social-innovation.hitachi/

How much new data / day?

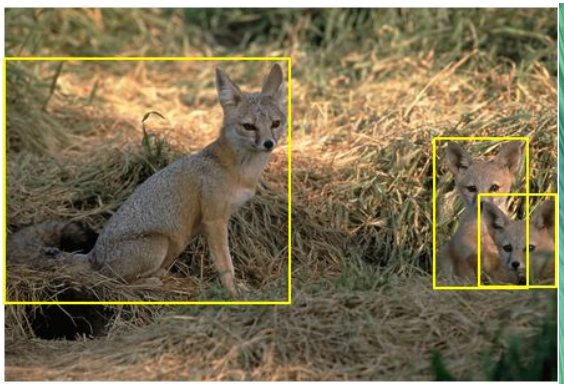
- 350M photos
- 100M photos
- YouTube 1M hours of contents
- 65B MSG

...

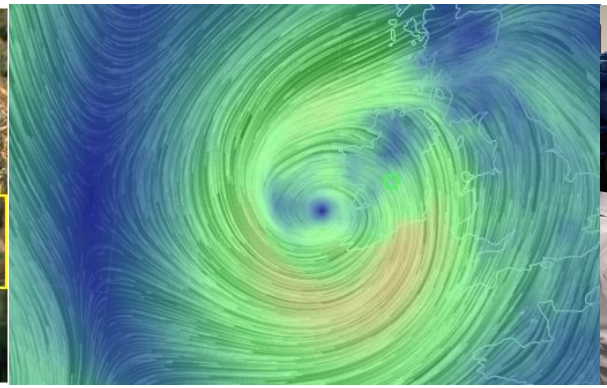
https://www.omnicoreagency.com/facebook-statistics/

What to do with the data ?

to extract useful information from data



object localization



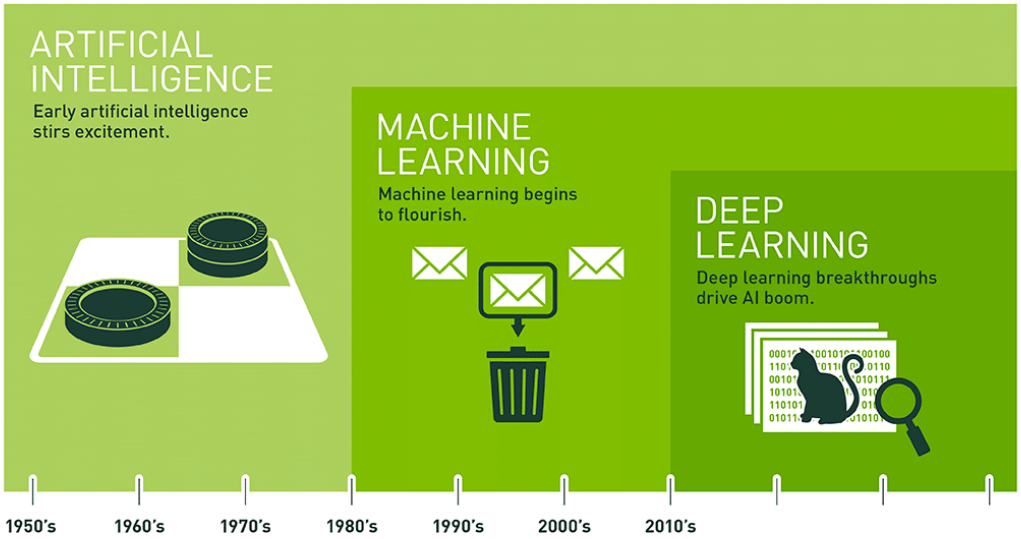
physical laws

diagnosis

https://www.kaggle.com/; https://www.irishtimes.com/; https://www.bioworld.com/;

AI has huge impacts and potentials for extracting information ...

AI and Deep Learning



https://blogs.nvidia.com/blog/2016/07/29/

- AI: programs with the ability to sense, reason and adapt like human
- ML: algorithms whose performance improves when exposed to more data
- DL: subset of ML in which employs artificial neural networks to adapt and learn from vast amounts of data

AlphaGo

- long-standing challenge for computer programme
- prior to 2015, the best computer programme managed to reach amateur dan level!

AlphaGo

2015/10, AlphaGo vs Fan Hui

2016/03, AlphaGo vs S. Lee

AlphaGo played its first match v.s. the 3-time European champion, Mr. Fan Hui (dan 2), and won the first ever game against a **Go professional with a score of 5-0**.

AlphaGo improved and became increasingly stronger and better (via reinforcement learning), and went on to defeat Go world champions (dan 9) around the world and **arguably became the greatest Go player of all time**.

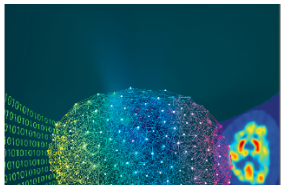
https://deepmind.com/research/case-studies/alphago-the-story-so-far

IMAGING

AI transforms image reconstruction

A deep-learning-based approach improves the speed, accuracy, and robustness of biomedical image reconstruction.

Artificial intelligence (AI) and machine learning are poised to revolutionize the way biologists acquire and interact with experimental data. In biomedical imaging, required domain-expert design were being completely overturned by more flexible neural network models" after he left the field of machine learning in speech recognition to study the physics of MRI. "I wondered if this paradigm could be similarly advantageous in the field of medical imaging, and in particu-



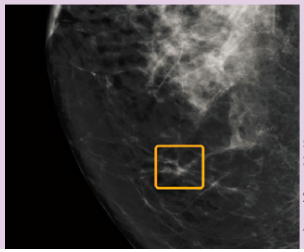
BREAST CANCER

AI outperforms radiologists in mammographic screening

Mammographic screening is widely used for the detection of breast cancers, but has its flaws. For example, false-positive findings can lead to

relative to the judgement of the first or sole radiologists.

"In our study, we addressed a common concern that machine learning results



roduced from McKinney, et al. *Nature* **577**, 89–94 (2020).

'IT WILL CHANGE EVERYTHING': AI MAKES GIGANTIC LEAP IN SOLVING PROTEIN STRUCTURES

DeepMind's program for determining the 3D shapes of proteins stands to transform biology, say scientists.

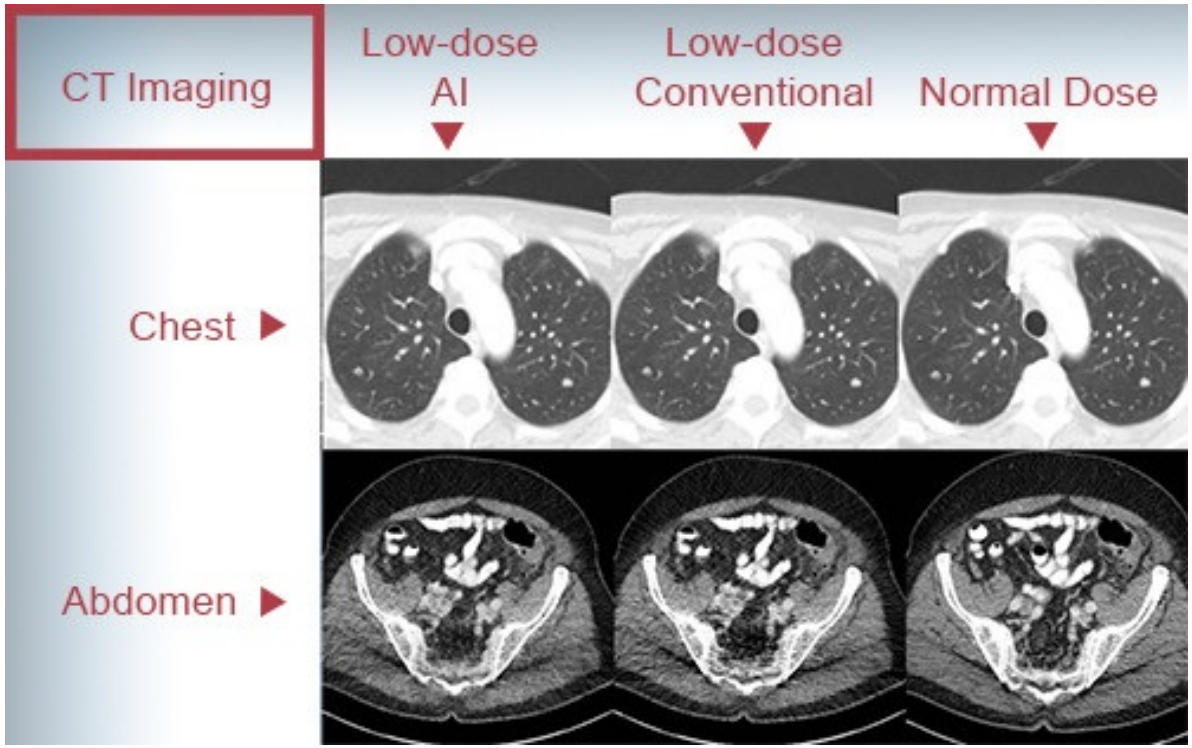
CHEMISTRY

AI designs organic syntheses

Software that devises effective schemes for synthetic chemistry has depended

editorials highlighting scientific breakthroughs from **Nature**

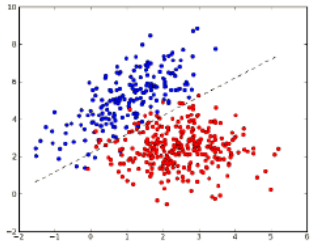
low-dose CT reconstruction

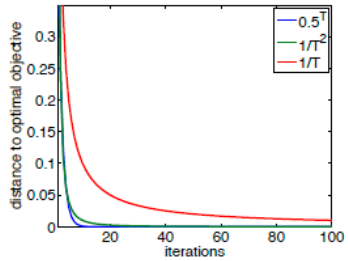


Shan HM, Padole A, ..., Wang G **Nature Mach. Intel.** 2019

- favorable / comparable reconstructions for low-dose CT
- orders of magnitude faster

Three pillars of DL





Data

Model

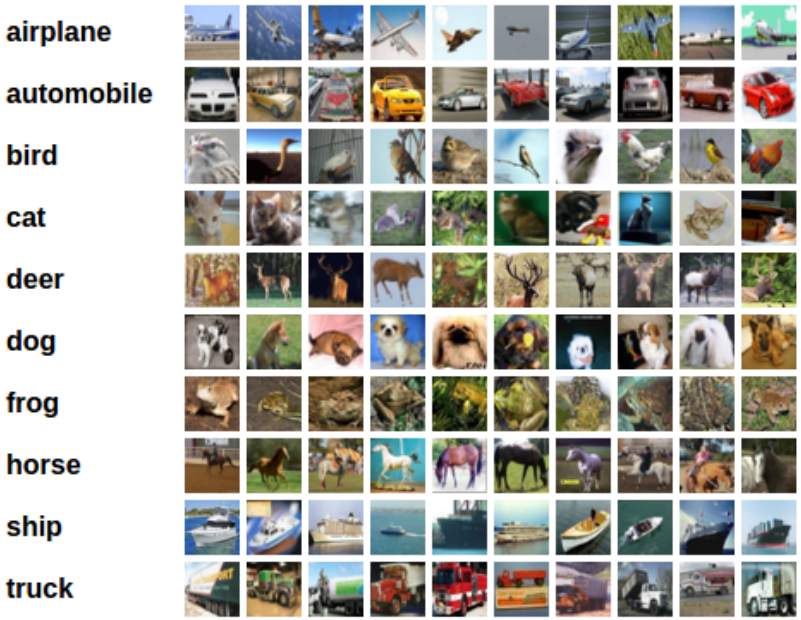
Optimization

Pillar I: “big” high-quality data

ImageNet: 14M images, 22,000 classes, 150GB

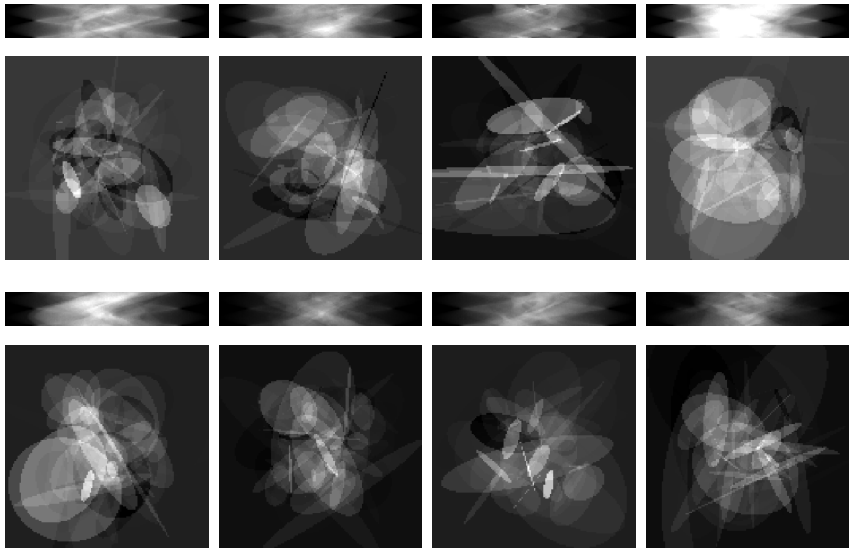
Training data

image classification



cifar10 dataset

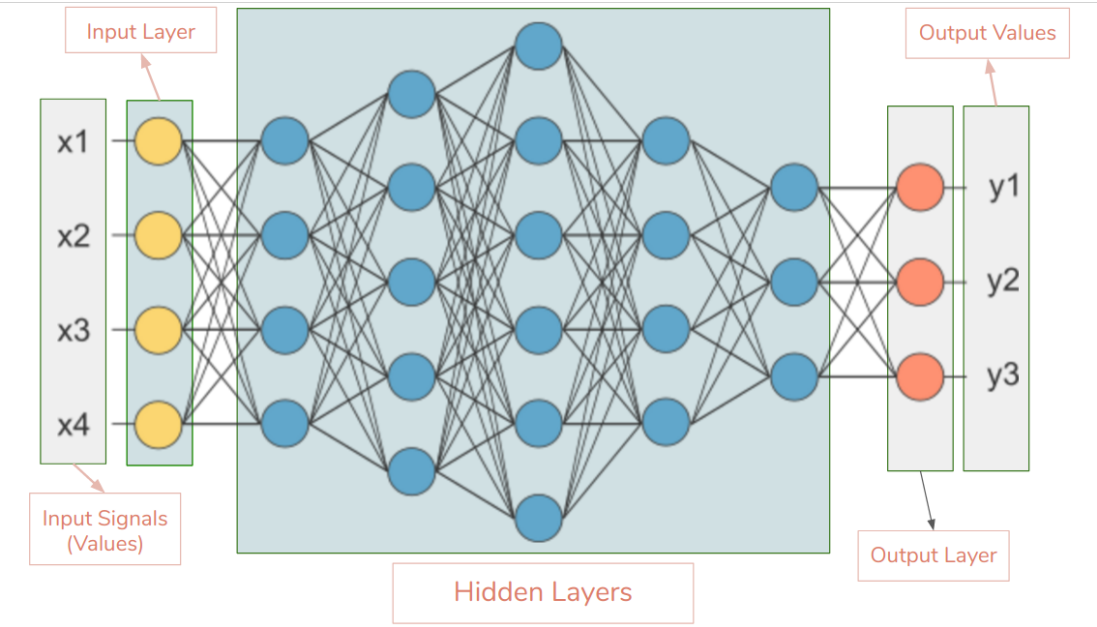
CT reconstruction



sinogram / phantom Zhang-Jin 2020

Pillar II: model

model: deep neural networks



https://medium.com/ravenprotocol

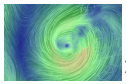
input output

 ⇒person identification

 ⇒animal classification

 ⇒tumor, lesion, covid ?

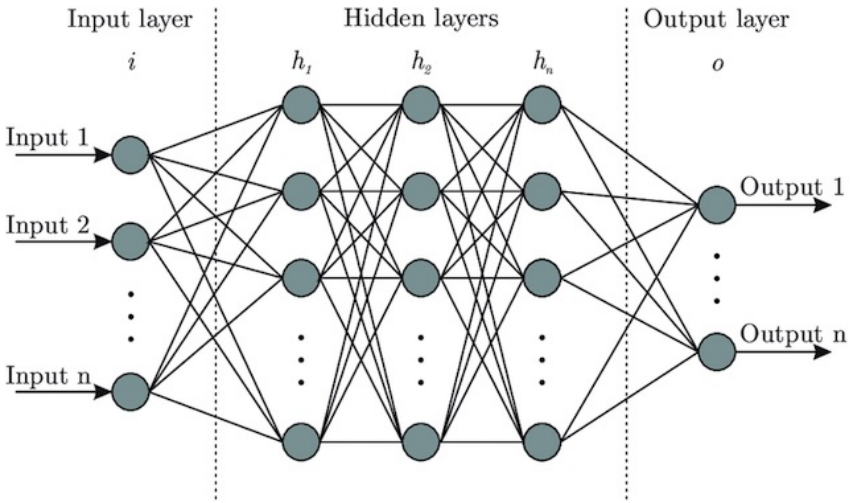
 ⇒next move

 ⇒physical laws ?

... ⇒...

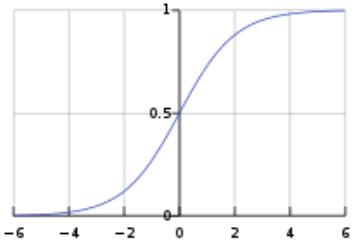
What is a neural network?

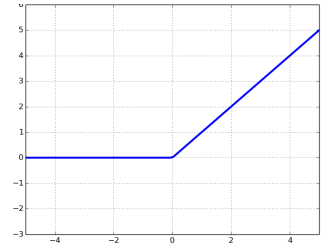
graphical representation



$$x_i^k = \sigma\left(w_{i,1}^k x_1^{k-1} + \ldots + w_{i,\ell_k}^k x_{\ell_k}^{k-1} + b_i^k\right)$$

$$x^k = \sigma(W^k x^{k-1} + b^k)$$





$$\sigma(x) = \begin{cases} \dfrac{1}{1+e^{-x}}, & \text{sigmoid} \\ \max(0, x), & \text{relu} \end{cases}$$

What is a neural network?

Starting with input vector $x^0 \in \mathbb{R}^d$ and an activation function $\sigma(x)$, form x^k, $k = 1, \ldots, L$, recursively

$$x^k = \sigma(W^k x^{k-1} + b^k)$$

with

$$W^k = \begin{pmatrix} w_{11}^k & w_{12}^k & \cdots \\ w_{21}^k & \ddots & \\ \vdots & & \ddots \end{pmatrix}, \quad x^k = \begin{pmatrix} x_1^k \\ x_2^k \\ \vdots \end{pmatrix}, \quad b^k = \begin{pmatrix} b_1^k \\ b_2^k \\ \vdots \end{pmatrix}$$

so a 5-layer NN with input x looks like (with $\Theta = \{(W^k, b^k)\}_{k=1}^L$)

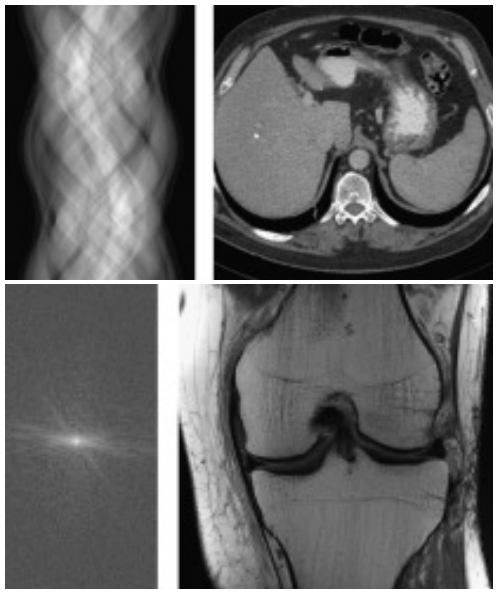
$$f_\Theta(x) = \sigma(W^4\sigma(W^3\sigma(W^2\sigma(W^1 x + b^1) + b^2) + b^3) + b^4)$$

A DNN $f_\Theta(x)$ is a function of input x, depending on Θ.

What is the loss?

Given data $\{(x_n, y_n)\}_{n=1}^N$, DL training is to find the optimal Θ so that the training loss is minimized

$$
\begin{aligned}
J(\Theta) &:= \sum_{n=1}^{N} \ell(f_\Theta(x_n), y_n) \\
&= \ell(f_\Theta(x_1), y_1) \\
&\quad + \ell(f_\Theta(x_2), y_2) \\
&\quad\quad + \ldots + \ell(f_\Theta(x_N), y_N)
\end{aligned}
$$



- least-squares loss

$$\ell(f_\Theta(x_i), y_i) = \|f_\Theta(x_i) - y_i\|^2$$

$\Rightarrow$ nonlinear least-squares / regression (in high-dimension)

What's more ...

- choices of loss ℓ: L^2, L^1, cross entropy, Wasserstein distance ...
- weight matrices W are often structured, e.g., convolutions $\Leftrightarrow$ physical interpretations / constraints
- important ingredients: residual connections, max pooling, ...

training task for DNNs (empirical risk minimization)

$$\min J(\Theta) = \sum_n \ell(f_\Theta(x_n), y_n)$$

Pillar III: optimization

training the DNNs (empirical risk minimization)

$$\min J(\Theta) = \sum_n \ell(f_\Theta(x_n), y_n)$$

- stochastic gradient descent on $J(\Theta)$

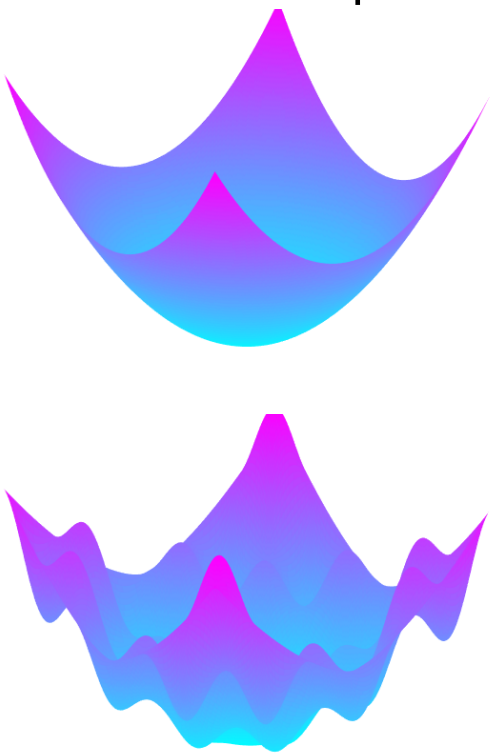
$$\Theta^{k+1} = \Theta^k - \eta_k \widehat{\nabla J(\Theta^n)}$$

Robbins-Monro 1951; Ying-Pontil 2007; Dieuleveut-Bach 2015; Lin-Rosasco 2016; Pillaud-Vivien - Rudi - Bach 0218; Jin-Lu 2019; Jahn-Jin 2020 ...

- backpropagation: implicit depend. of f_Θ on Θ, chain rule for gradient

P. Werbos 1974; Rumelhart, Hinton, Williams **Nature** 1986

loss landscape



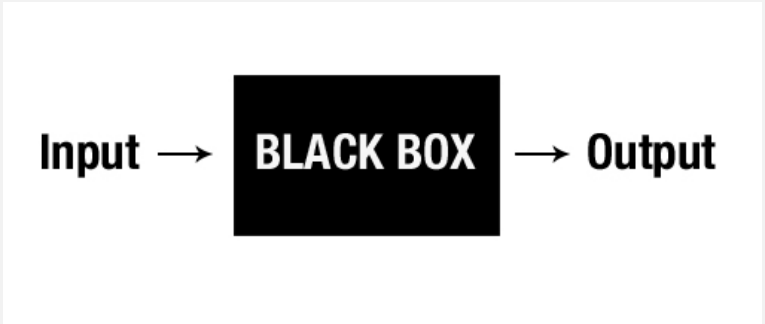
convex v.s. nonconvex

DL challenge

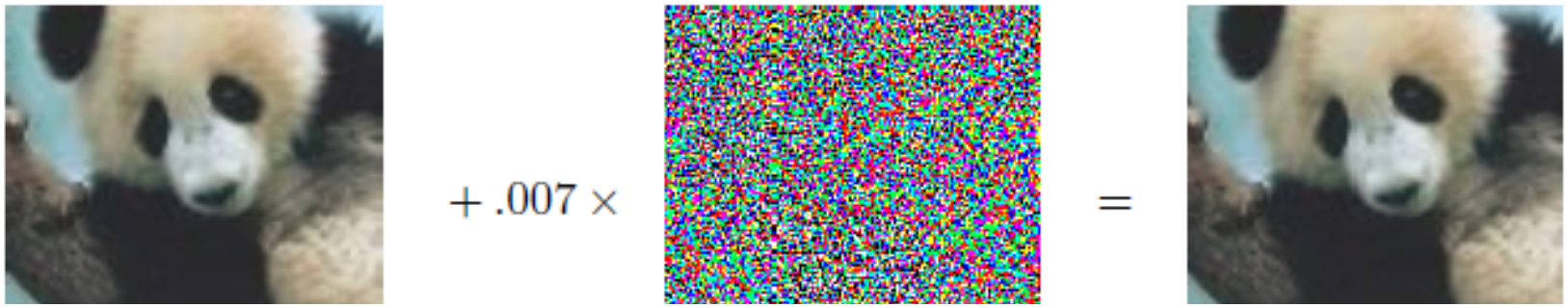
simple optimization problem ?!

- “big”: millions – billions of training parameters
- “complex”: convolution, normalization, dropout, mixup ...
- “nonconvex”: highly nonconvex optimization

influenced also by training set, initialization, optimization algorithms ...



Deep trouble?



panda (56%) $\Rightarrow$ gibbon (94%)

Goodfellow, Shlens & Szegedy. ICLR 2015

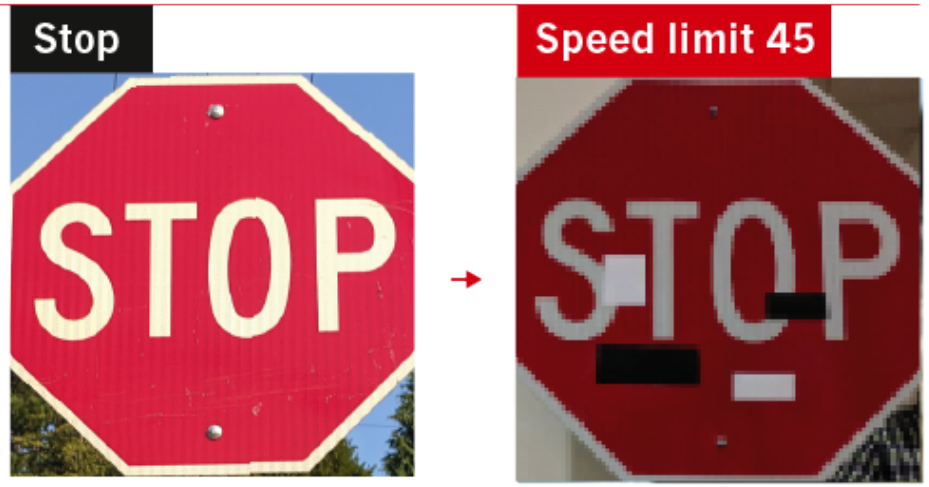
Visually imperceptible perturbations can break the DNN performance!

Deep trouble?

FOOLING THE AI

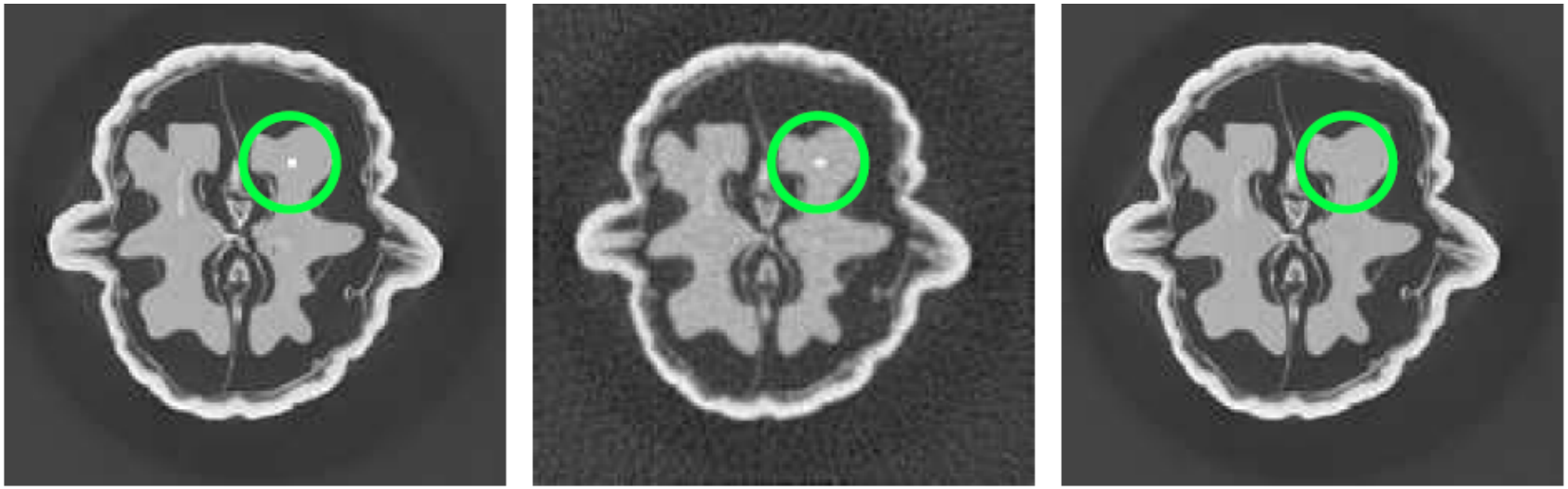
Deep neural networks (DNNs) are brilliant at image recognition — but they can be easily hacked.

These stickers made an artificial-intelligence system read this stop sign as 'speed limit 45'.



https://www.nature.com/articles/d41586-019-03013-5

Deep trouble?



ground truth | gradient descent, PSNR 27.5 | learned, PSNR 40.2

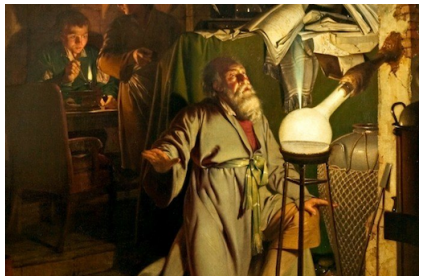
Moeller-Mollenhoff-Cremers ICCV 2019, Antun et al PNAS 2020, Barbano-Arridge-Jin-Tanno 2021

Important lesion information washed away by DNN !

Opening the black box?

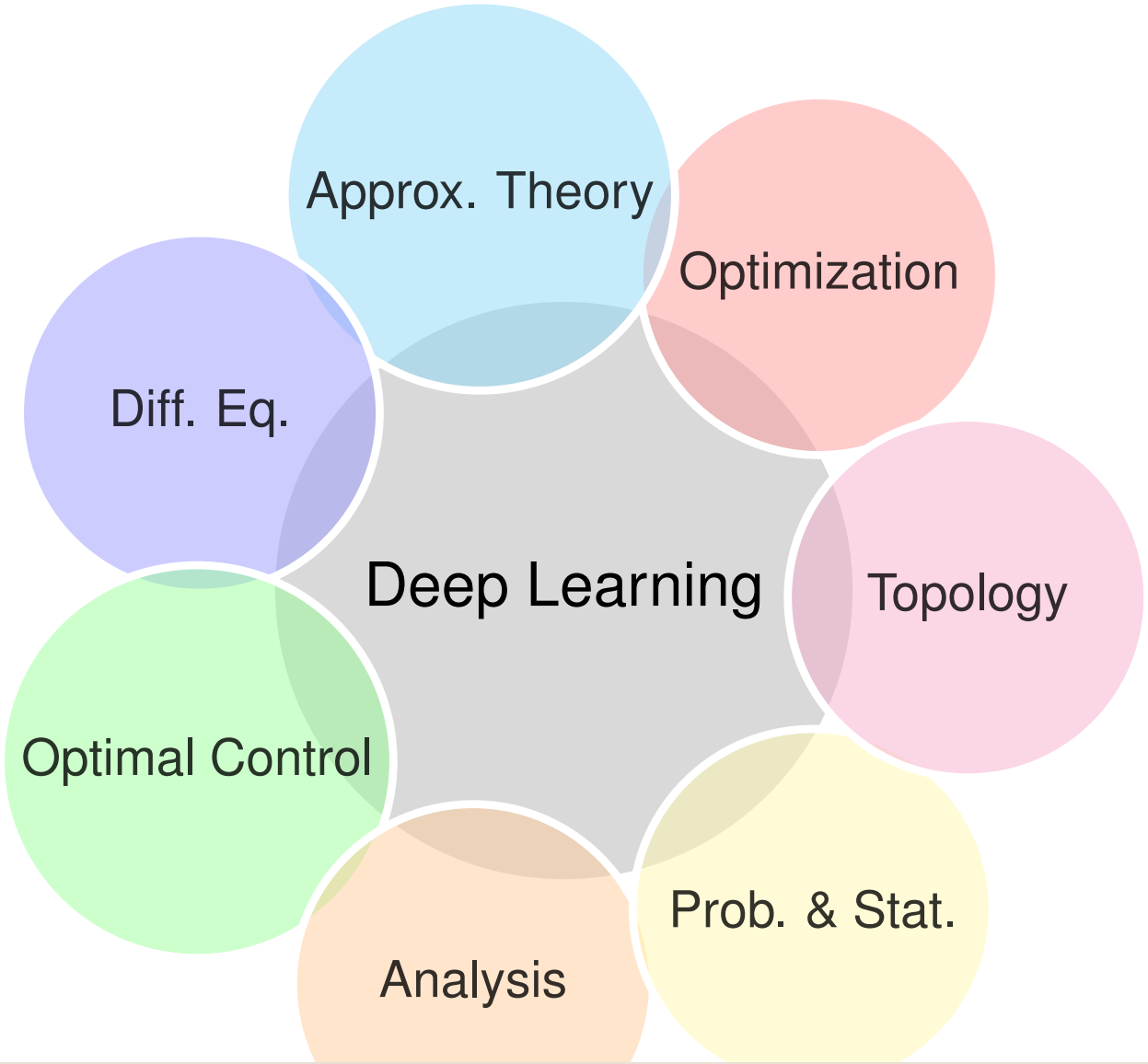
Despite impressive empirical successes, mathematical theory behind DL remains rather unsatisfactory with many open questions

- Why does DL training work at all?
- Why do overparameterized DNNs generalize and not overfit?
- What if limited training data (in physical problems)?
- What is the mechanism for instability and how to overcome?
- ...



Ali Rahimi (Google)
Test of Time Award at NIPS'17

Machine learning algorithms have become a form of "alchemy".



All kind of mathematics can be applied ...

DL is full of exciting problems for all areas of mathematics!

Example: ResNet

DNNs as continuum flows: ResNet – Residual Networks He-Zhang-Ren-Sun 2015

$$x^{k+1} = x^k + \tau\sigma(W^k x^k + b^k)$$

which can be rewritten as

$$\frac{x^{k+1} - x^k}{\tau} = \sigma(W^k x^k + b^k)$$

interpreting τ as a stepsize between $x^{k+1} = x(t+\tau)$ and $x^k = x(t)$

$$\frac{x^{k+1} - x^k}{\tau} \to \frac{\mathrm{d}}{\mathrm{d}t}x(t) \qquad \text{derivative!}$$

$\Rightarrow$ differential equations / dynamical system Haber-Ruthotoo 2017; W E 2017; Chen-Rubanova-Bettencourt-Duvenaud 2018

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \sigma(W(t)x(t) + b(t))$$

differential equation / dynamical system

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \sigma(W(t)x(t) + b(t))$$

- DNNs are discretization of differential equations
  different discretization $\Rightarrow$ principled design of DNNs
  - stability $\Leftrightarrow$ robustness?
  - accuracy $\Leftrightarrow$ efficiency?
  - controllability $\Leftrightarrow$ approximation property?
  - ....
- novel training paradigm: time parallelization, beyond "backpropagation", adaptivity
- ...

Example: unrolled iteration

- standard gradient descent

$$x^{k+1} = x^k - \eta_k \nabla E(x^k), \quad k = 1, 2, \ldots, ...$$

well understood, great success, and solid theoretical backup

Engl-Hanke-Neubauer 1996; Kaltenbacher-Neubauer-Scherzer 2006; Ito-Jin 2015

- unrolled iteration with learned increments

$$x^{k+1} = x^k + f_{\Theta^k}(\nabla E(x^k)),$$

Executes only a finite number of iterations
Computes the updates using DNNs

Gregor-LeCun 2010, Adler-Oktem 2017, ..., Barbano-Zhang-Arridge-Jin 2020

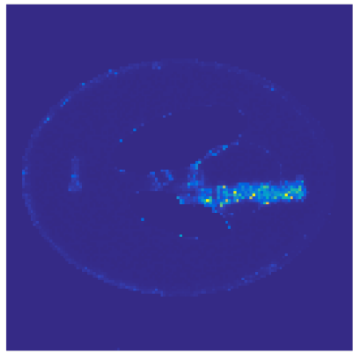
- lightweight hybrid deterministic / Bayesian structure

sparse view CT (30 dir)

Methods	Ellipses Phantoms	SL Phantom
FBP	25.5264	18.4667
TV	35.1587	37.2162
LPD	**44.5122** $\pm$ **0.4911**	44.0472 $\pm$ 0.4187
DGD	43.2577 $\pm$ 0.4183	44.6913 $\pm$ 0.6644
BDGD - MFVI	**44.6642** $\pm$ **0.4637**	**47.2946** $\pm$ **0.5778**
BDGD - MCDO	43.2126 $\pm$ 0.1285	45.1725 $\pm$ 0.4461

$\hat{\mathbb{E}}[x]$

$\widehat{\mathrm{Var}}[x]$



OOD reconstruction by BDGD-MFVI

Barbano-Zhang-Arridge-Jin 2020

- Learning can improve the accuracy significantly, and faster ...
- Being Bayesian, even if only a little, can help a lot !

deep equilibrium model: weight-tying + limit $k \to \infty$ Bai-Kolter-Koltun 2019; Ghaoui-Gu-Travacca 2019; Kawaguchi 2021 ...

$$x^* = x^* + f_\Theta(\nabla E(x^*))$$

nonlinear equation

$$f_\Theta(\nabla E(x^*)) = \mathbf{0}$$

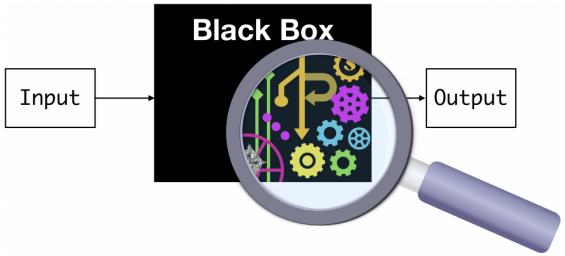
unrolling is fixed point realization of ...

so what ?

- alternative implicit solvers (quasi-Newton?)
- theoretical analysis (operator theory, nonlinear analysis)
- robustness, model uncertainty and statistical consistency ?
- ...

Conclusion

- DL provides many exciting opportunities for mathematical and physical sciences
- Mathematics offer powerful tools for understanding DL, opening up black box
- ...



https://towardsdatascience.com/

towards interpretable deep learning (for physical sciences)