Applying AI in the Wild

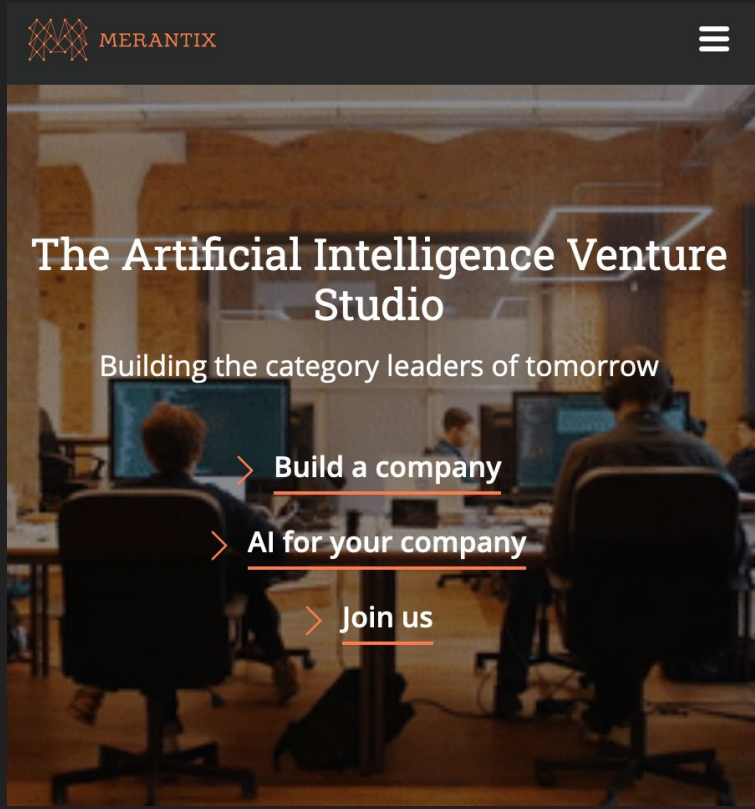
Dr. Johannes Otterbach
VP of ML Research @ Merantix Labs

Personal Introduction

- PhD in Physics (Kaiserslautern, 2011)
- PostDoc (Harvard) until 2014
- Since then in the Tech Industry
  - Palantir
  - Lendup
  - Rigetti
  - OpenAI
- Starting April 1st:
  VP of ML Research @ Merantix Labs

Merantix

- Founded 2016; located in Berlin
- Venture Studio for AI-based tech companies
- Consulting & research division: Merantix Labs
- We’re hiring!
  - SWEs
  - Data Engineers and Scientists
  - ML Engineers and Researchers
  - ...



What is AI

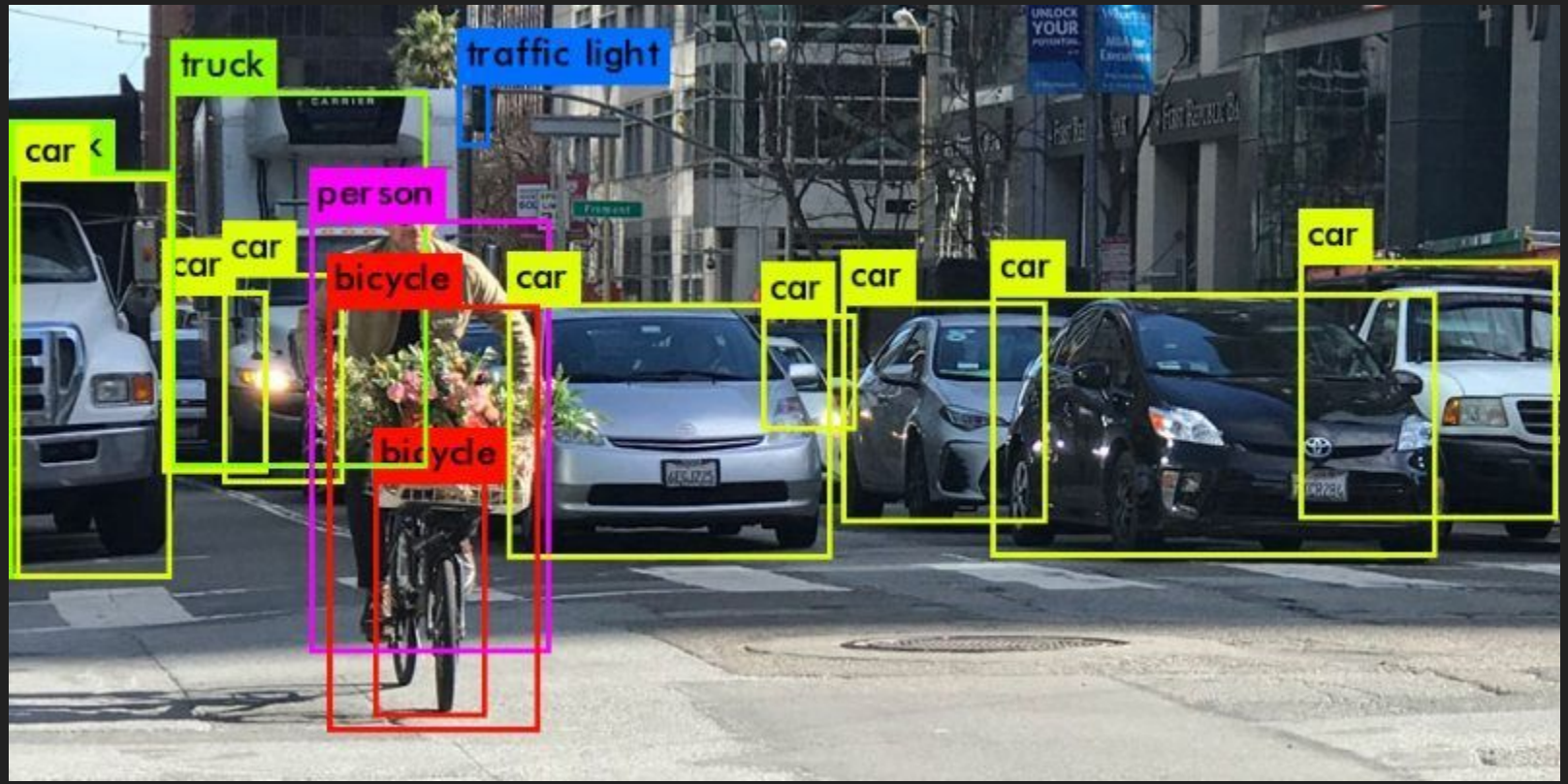
- 1956: Dartmouth workshop
  - Considered the founding event of modern AI

> *We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that* ***every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it****. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.*

What is AI

- Led to big breakthroughs in systems using Rule-Based Decision Making
- Ruled-based decision making (GOFAI) is challenged by complex situations
  - Generalization is impossible
  - Cannot deal with uncertainty or new information
  - …
- Humans can clearly do it: hence we need new approach
  - Neural Networks started “working” in the 90’s
  - Big breakthrough in the 2008-2012 time period

AI breakthroughs



AI breakthroughs

Microsoft's new breakthrough: AI that's as good as humans at listening... on the phone

Microsoft's new speech-recognition record means professional transcribers could be among the first to lose their jobs to artificial intelligence.

By Liam Tung | October 19, 2016 -- 10:10 GMT (03:10 PDT) | Topic: Innovation

AI breakthroughs

ThisPersonDoesNotExist.com uses AI to generate endless fake faces

49

Hit refresh to lock eyes with another imaginary stranger

By James Vincent | Feb 15, 2019, 7:38am EST

Listen to this article

SHARE

AI breakthroughs

Home / Technology / Computer Sciences

SEPTEMBER 1, 2008

Stanford's 'autonomous' helicopters teach themselves to fly

AI breakthroughs

COMPUTING

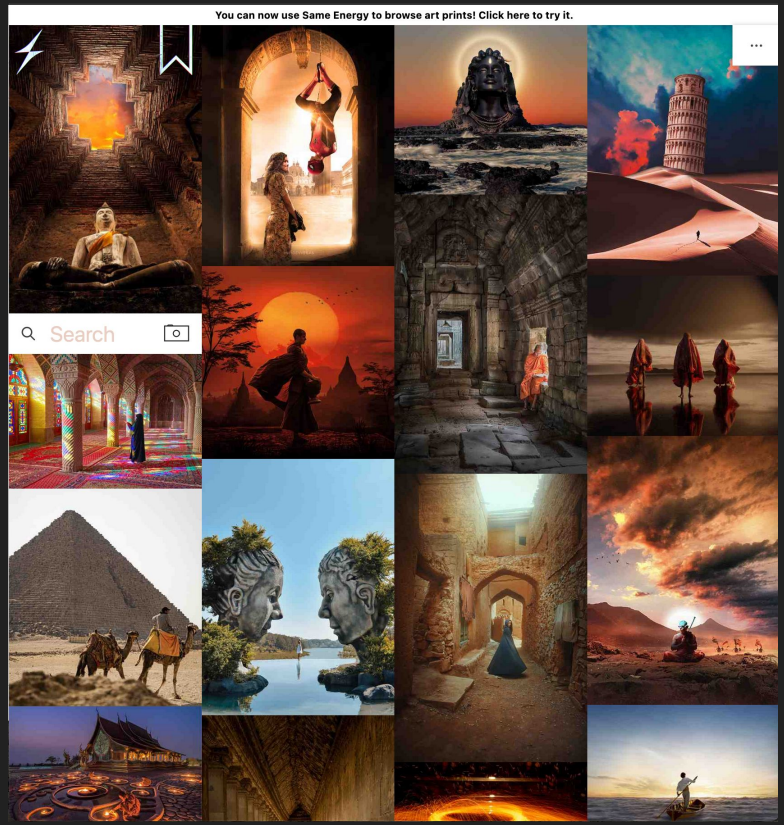
AI versus AI: Self-Taught AlphaGo Zero Vanquishes Its Predecessor

DeepMind's Go game-playing AI—which dominated its human competition—just got better

By Larry Greenemeier on October 18, 2017

AI breakthroughs

- **Visual** instead of word-based search
- website: same.energy



Comment | Open Access | Published: 17 August 2020

Machine learning for chemical discovery

Alexandre Tkatchenko

Nature Communications **11**, Article number: 4125 (2020) | Cite this article

11k Accesses | **7** Citations | **122** Altmetric | Metrics

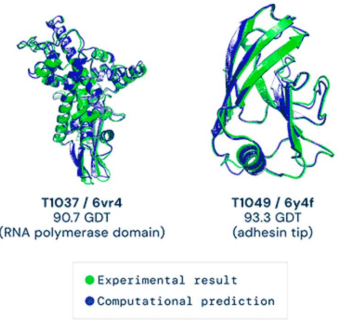
Discovering chemicals with desired attributes is a long and painstaking process. Curated datasets containing reliable quantum-mechanical properties for millions of molecules are becoming increasingly available. The development of novel machine learning tools to obtain chemical knowledge from these datasets has the potential to revolutionize the process of chemical discovery. Here, I comment on recent breakthroughs in this emerging field and discuss the challenges for the years to come.

07 Dec 2020 | 20:06 GMT

AlphaFold Proves That AI Can Crack Fundamental Scientific Problems

DeepMind's breakthrough demonstrates deep learning's potential to dramatically accelerate scientific discovery

By **Payal Dhar**



Artificial Intelligence Helps Hunt Down Superconductors

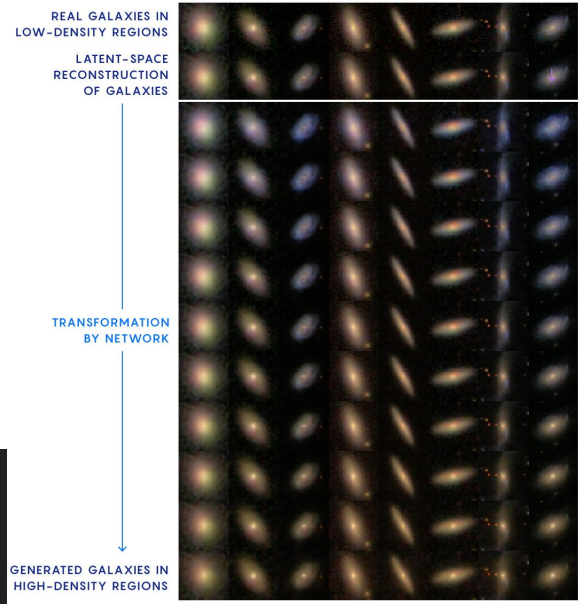
Researchers use machine learning to speed up the trial-and-error search for new materials that can conduct electricity without resistance.

Media credits: *maxuser via Shutterstock*

PHYSICS

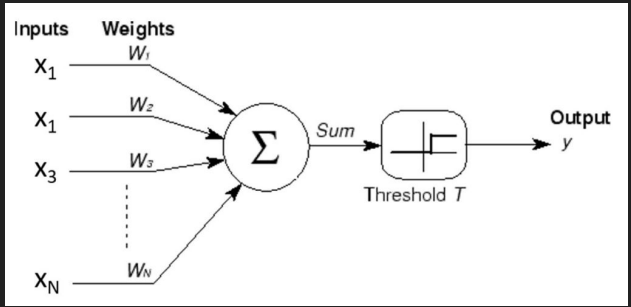
Thursday, March 7, 2019

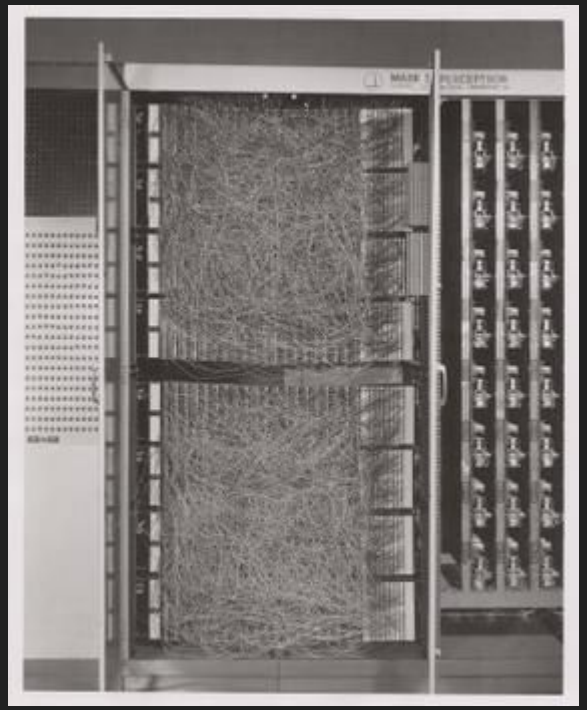
Yuen Yiu, Staff Writer

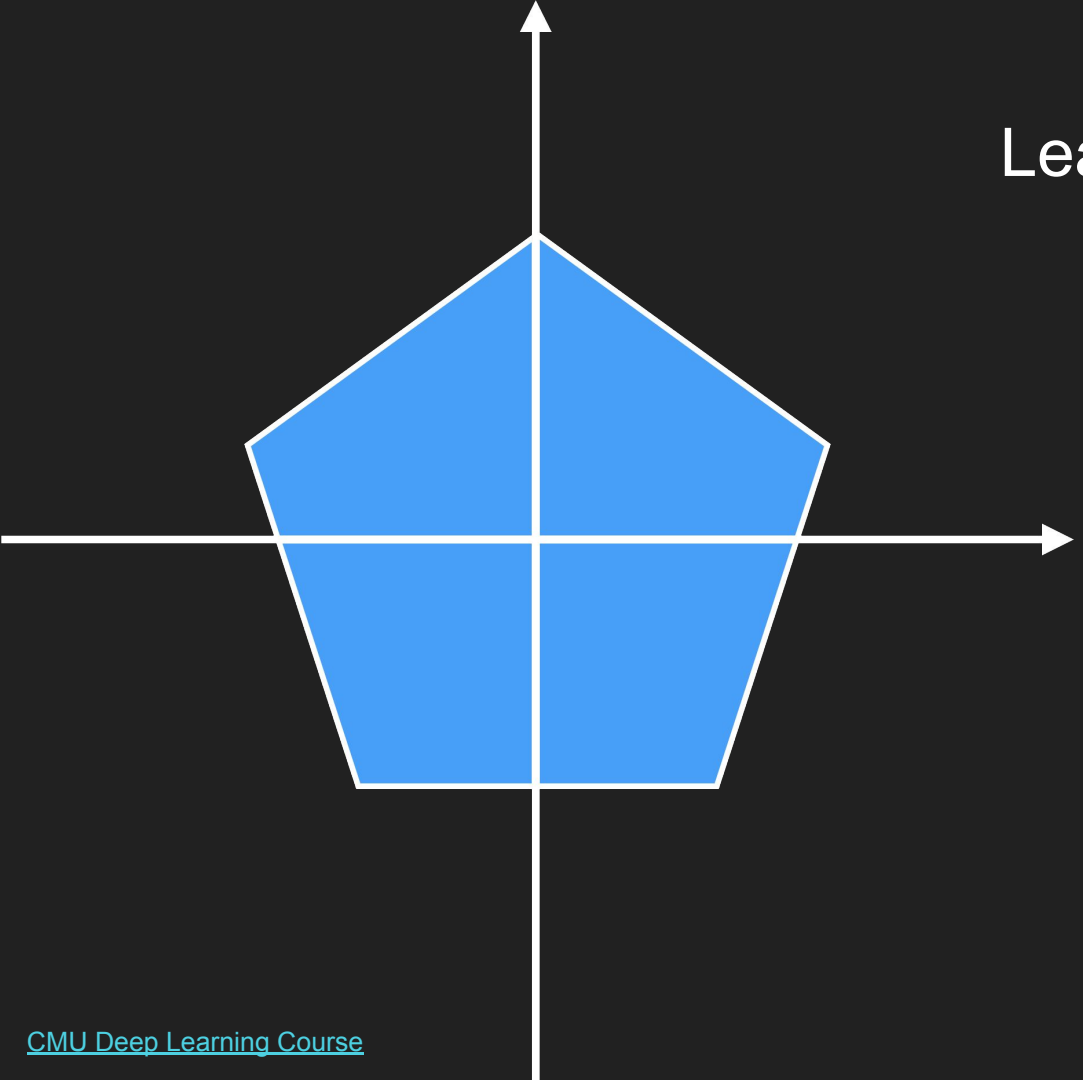


How does it work? Rosenblatt and the Perceptron

- Mark 1 is first physical implementation of the perceptron algorithm (1958)
- Inspired by Neuron architecture of the brain
- Simplified version of the perceptron algorithm can be used to learn boolean functions





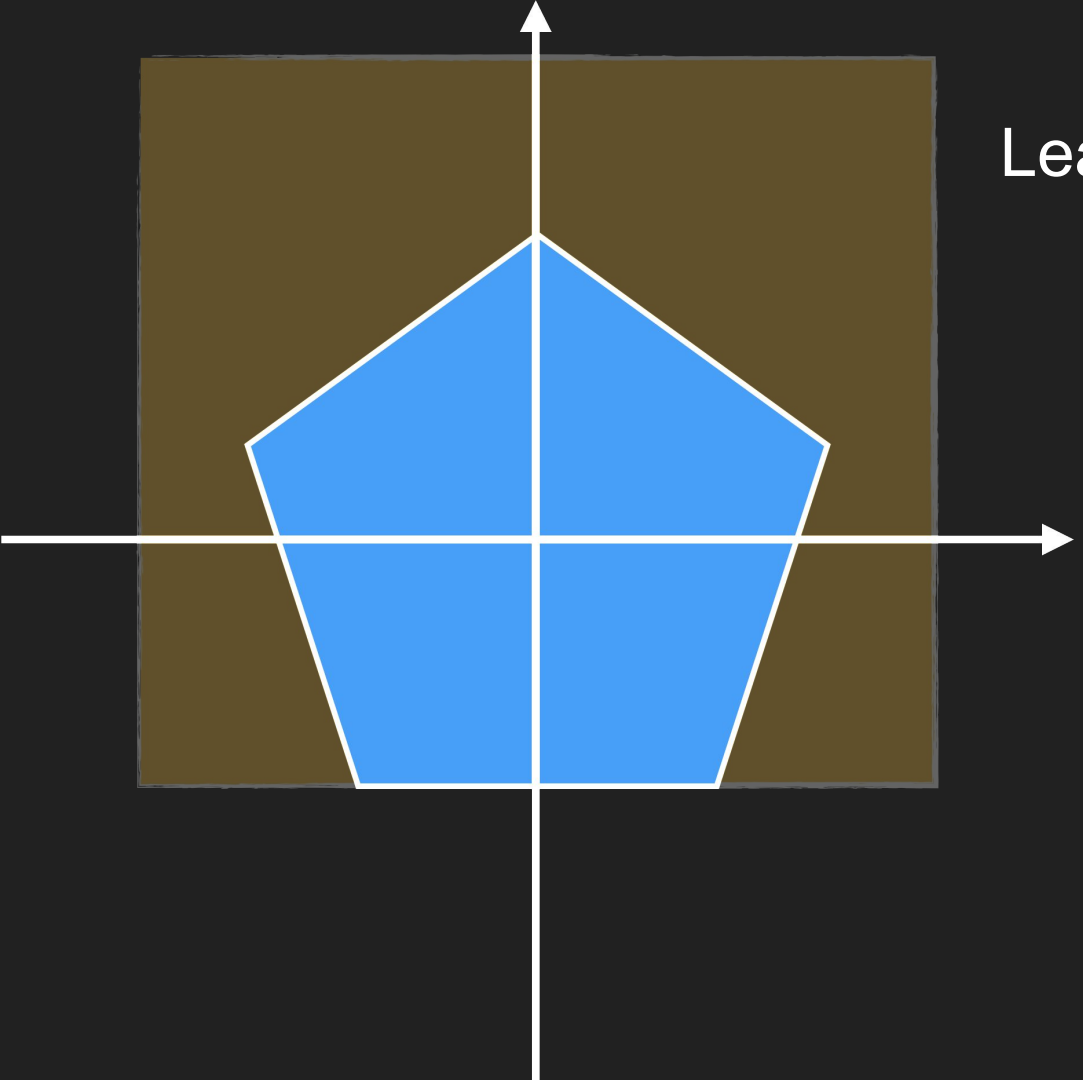


Learning Boolean Functions

Want to learn boolean formula for blue pentagon

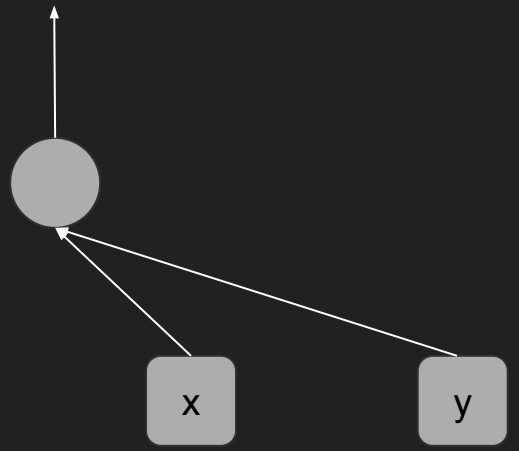
x

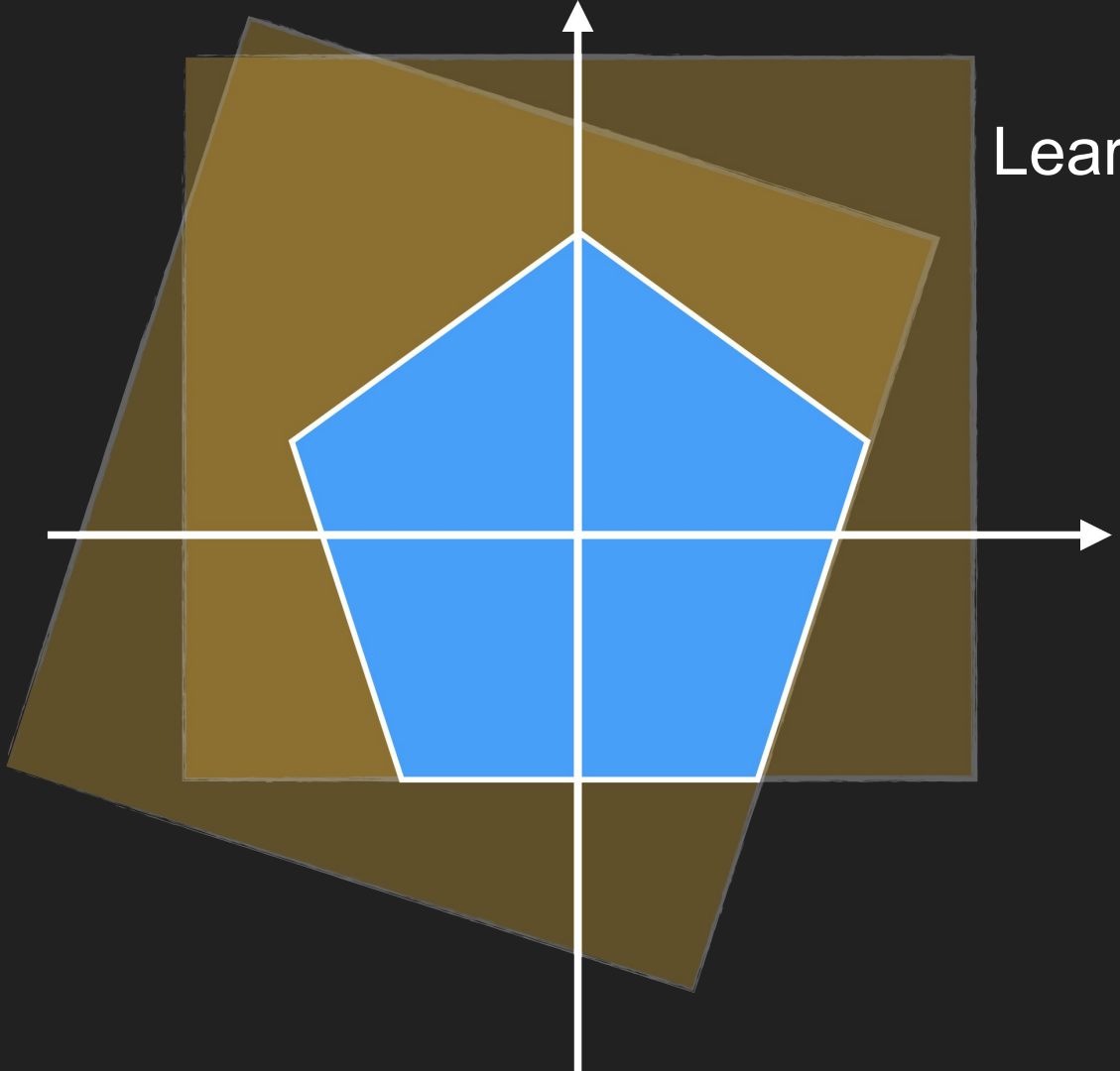
y



Learning Boolean Functions

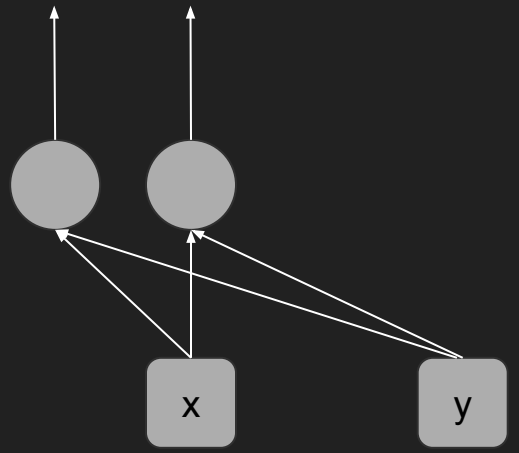
Single edge is given by a half-space and assign 1 if half space condition is true, otherwise 0

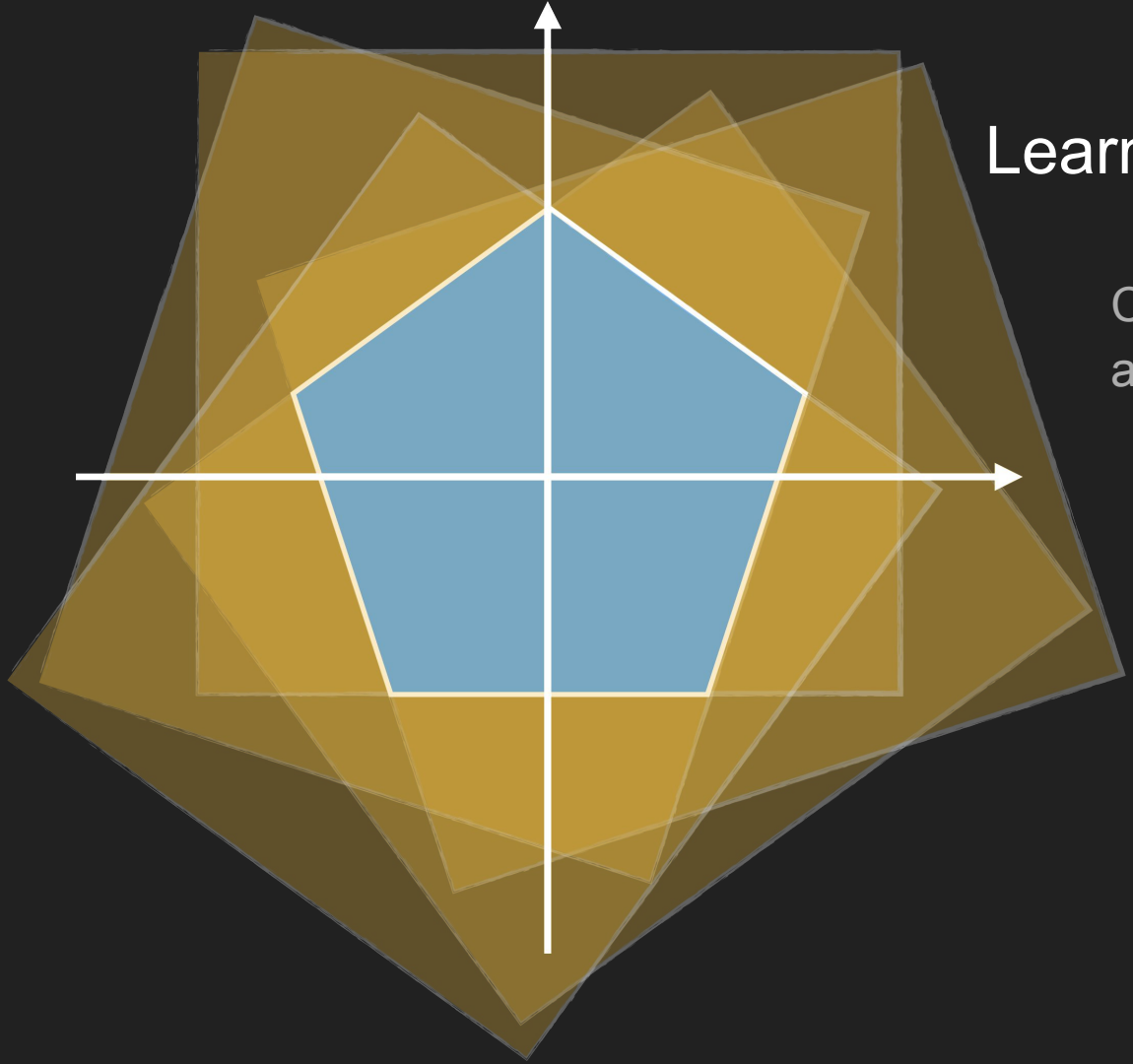




Learning Boolean Functions

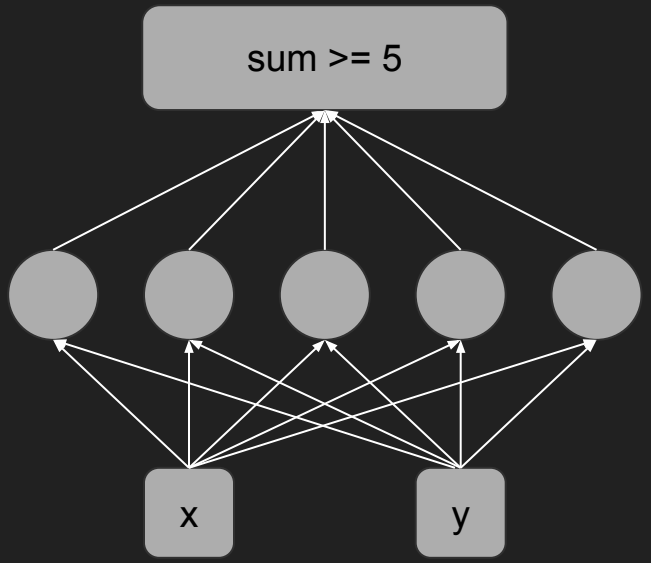
Keep constructing half-spaces iteratively



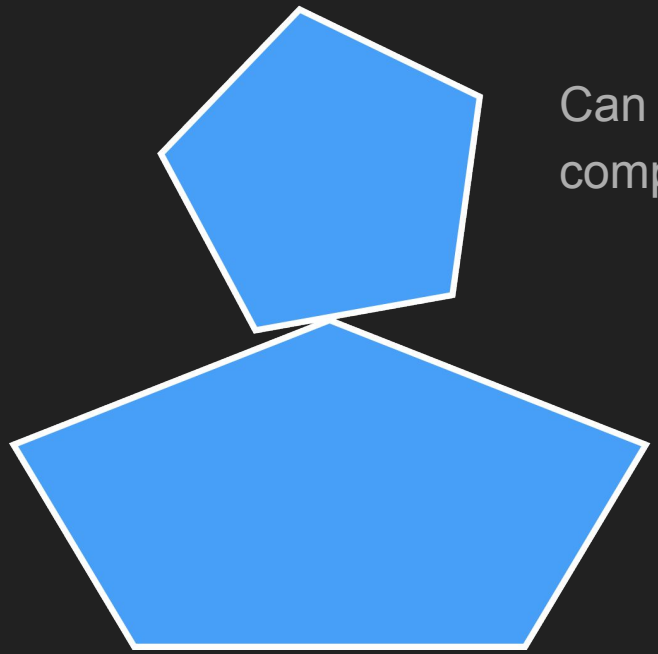


Learning Boolean Functions

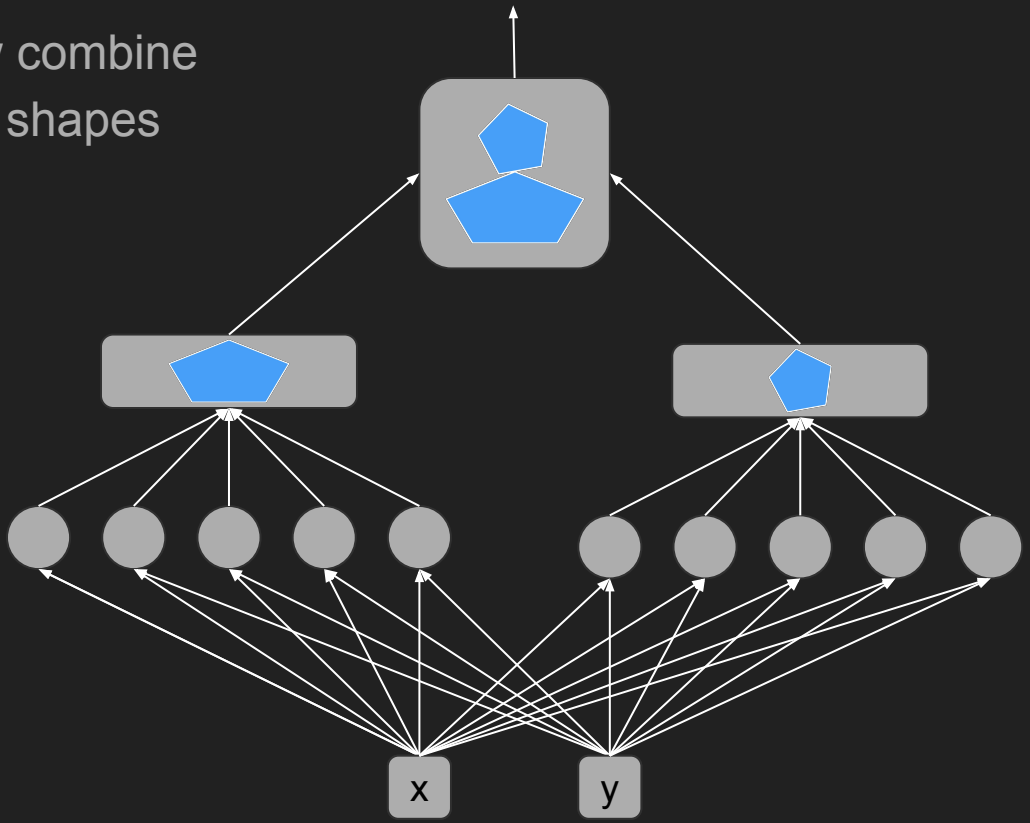
Combine the half-space outputs in a summation



Learning Boolean Functions

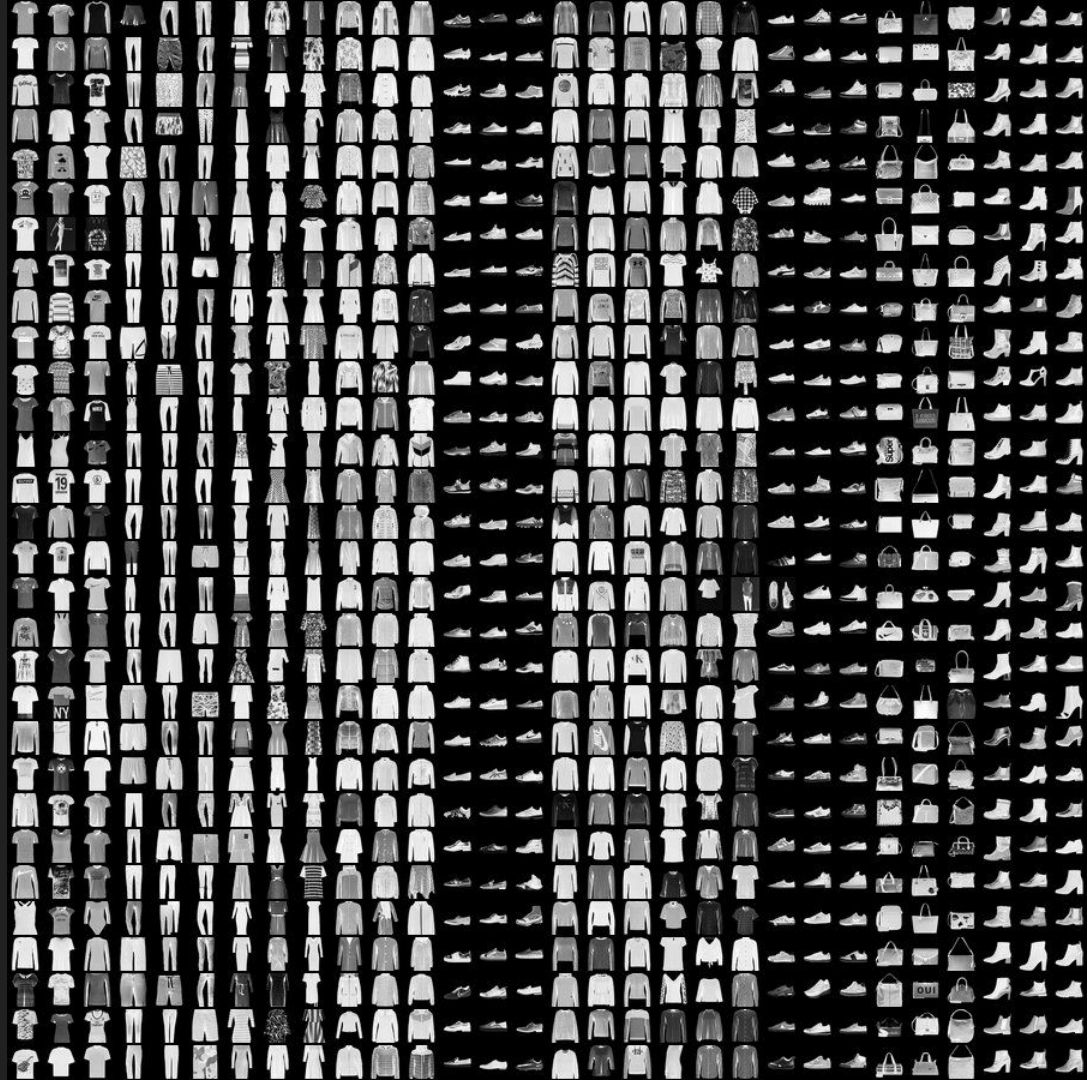


Can now combine
complex shapes



Learning vs Designing

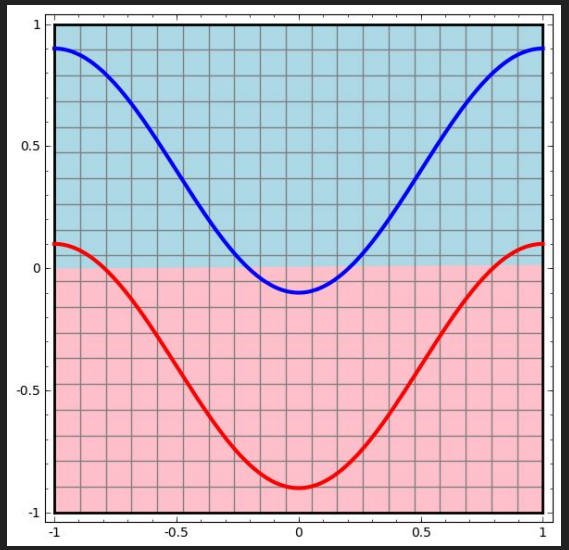
- Imagine designing rules for all the different fashion items
- This is still an “easy problem”
- Let the algorithm learn the rules, a.k.a. supervised learning

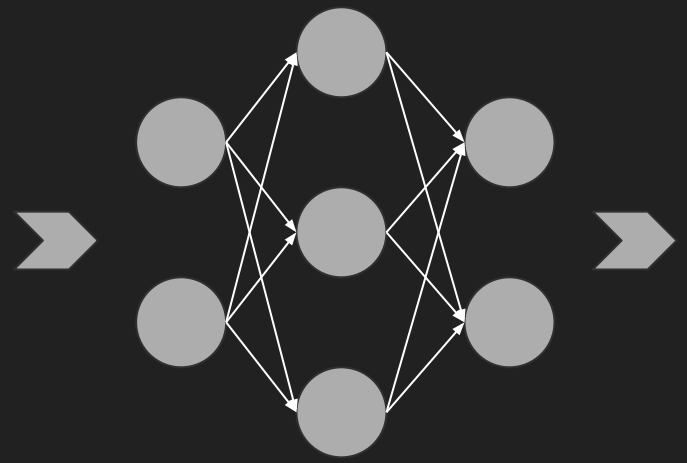


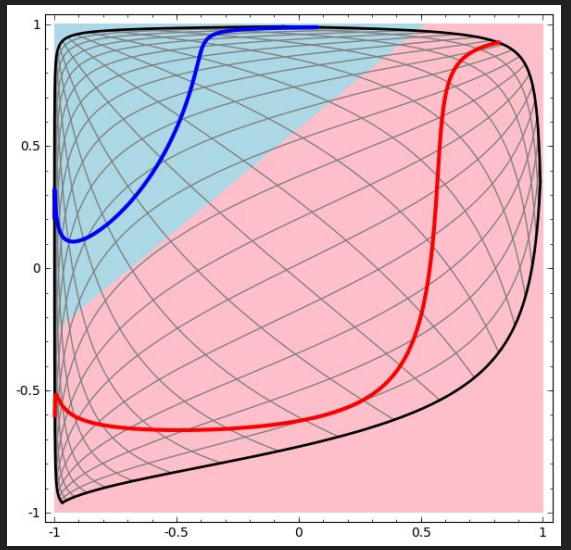
Feature Extractors

Intuition:

NNs learn features that allow linear separability in an appropriate space

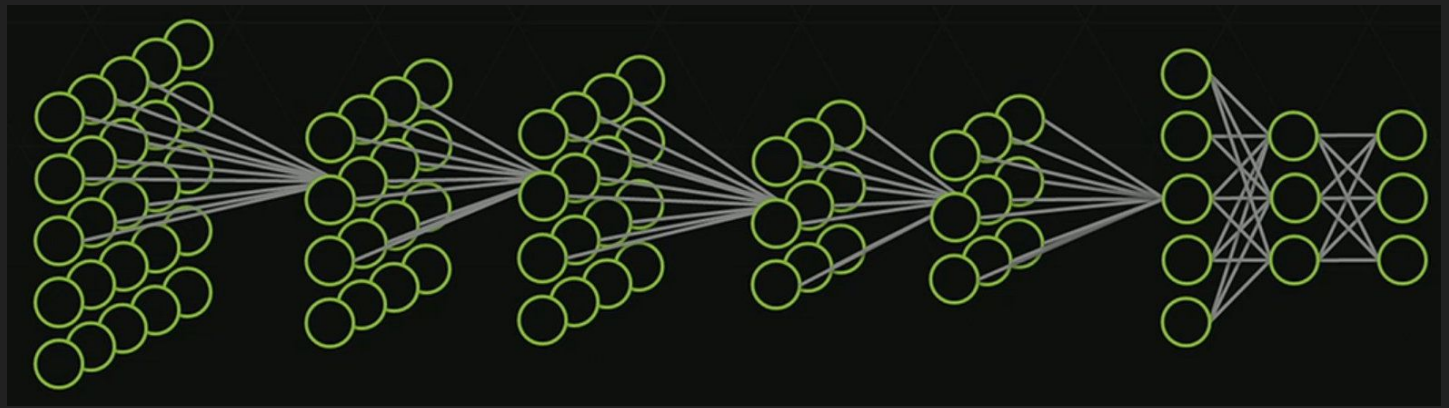






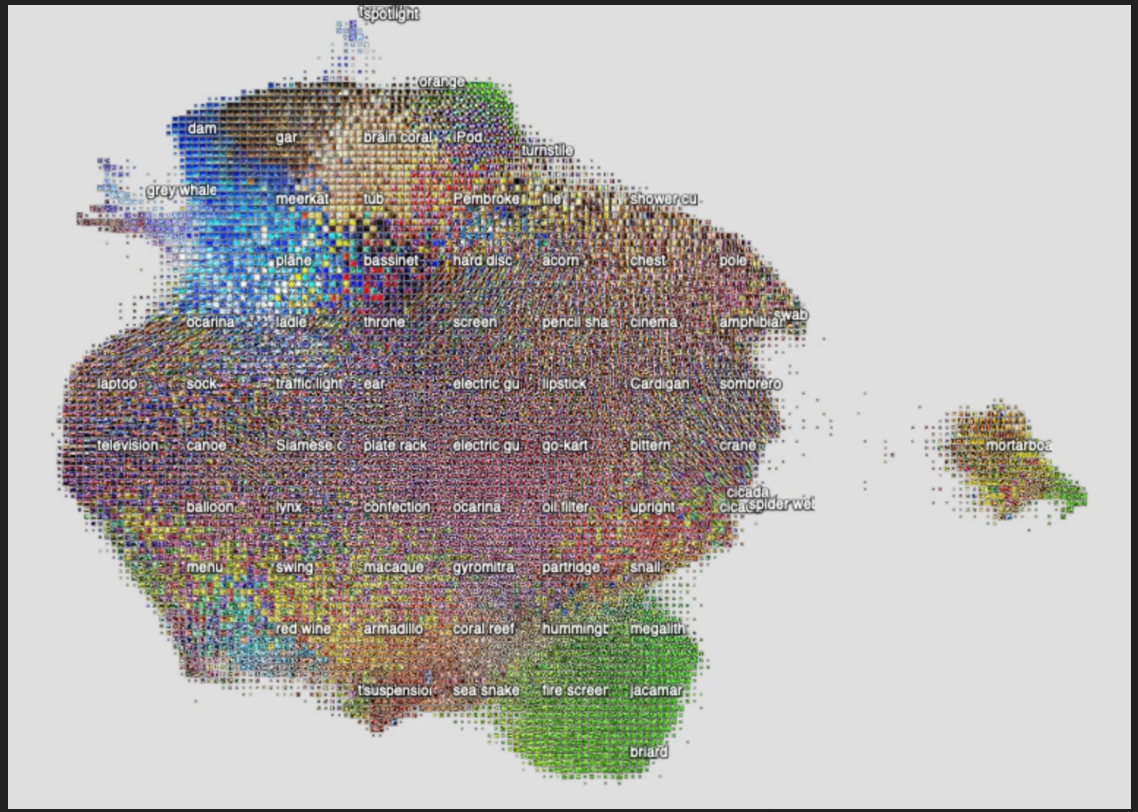
Chris Olah: Neural Networks, Manifolds, and Topology

Filter Banks and Compositionality



H. Lee et al, Doi:10.1145/2001269.2001295

Activation Atlases



- Can inspect whole NNs
- Usable for importance assessment of certain filters
- distill.pub as reference

(one) goal of AI

- Goal: Learn the rules of the world by observing the surrounding environment

- NNs are good at extraction “features” from observational data
  - much more scalable than designing custom IF-ELSE rules

- So what are the challenges?

Yann Lecun's Cherry Cake

- "Pure" Reinforcement learning (cherry)
  - A few bits per example
- Supervised learning (icing)
  - human supplied data (very costly)
  - 10 - 10000 bits per example
  - models conditional distributions p(y|x)
- Un-/self-supervised learning (THE CAKE):
  - Millions of bits per example
  - models full distribution p(y, x)
  - often complex data characteristics

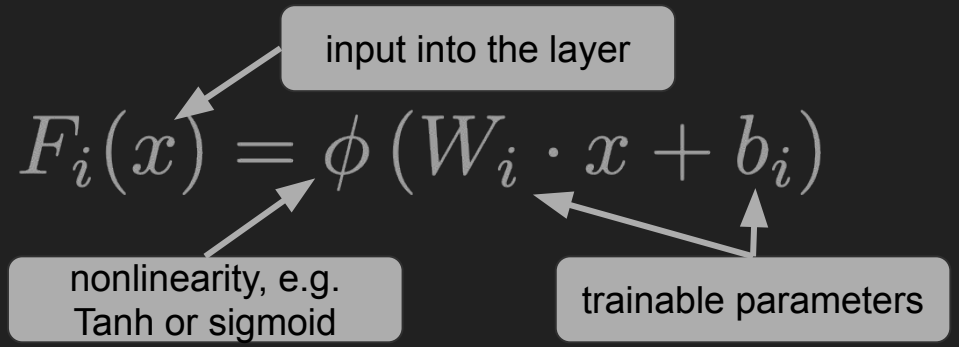
Questions so far?

Rough Outline of what’s to come

- Training Neural Networks 101
- Overview over the CNN and Transformer architectures
- Phase Transitions & Training Dynamics
- Scaling Laws of Deep Learning
- Some Applications

Backprop + SGD — Workhorses for training NNs

- A single layer:



- Full NN is stacking several of these layers on top of each other

$$\mathrm{NN}(x) = F_L \circ F_{L-1} \circ \cdots \circ F_1(x)$$

Backprop + SGD — Workhorses for training NNs

- Loss function / objective
  - supervised learning: predict outcome y based on independent variable vector x: **p(y|x)**
- Example binary classification

$$p(y|x) = \prod_{i=1}^{N} p(y_i|x_i)$$
$$= \prod_{i=1}^{N} [(1 - y_i)(1 - p(y_i = 1|x)) + y_i p(y_i = 1|x)]$$

- Log-space is often easier to deal with

$$\log(p(y|x)) = \sum_{i=1}^{N} [(1 - y_i) \log(1 - p(y_i = 1|x)) + y_i \log(p(y_i = 1|x))]$$

Backprop + SGD — Workhorses for training NNs

- log(p) also called log-likelihood

- Fitting objective: Maximize the likelihood, a.k.a. **MLE**

$$\min_{W,b} \left[\mathcal{L}(x, y; \{W_i, b_i\}_{i \in \{1,\ldots,L\}})\right] = \min_{W,b} \left[-\log(p(y|x))\right]$$

- Optimization using Gradient descent

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(x, y; \theta_t)$$

- Gradient is computed through iterative application of chain rule
  - Can be very cumbersome, hence proliferation of AD libraries, e.g. PyTorch, Jax, Tensorflow, Keras, etc.

Backprop + SGD — Workhorses for training NNs

- Full-batch vs stochastic gradient descent

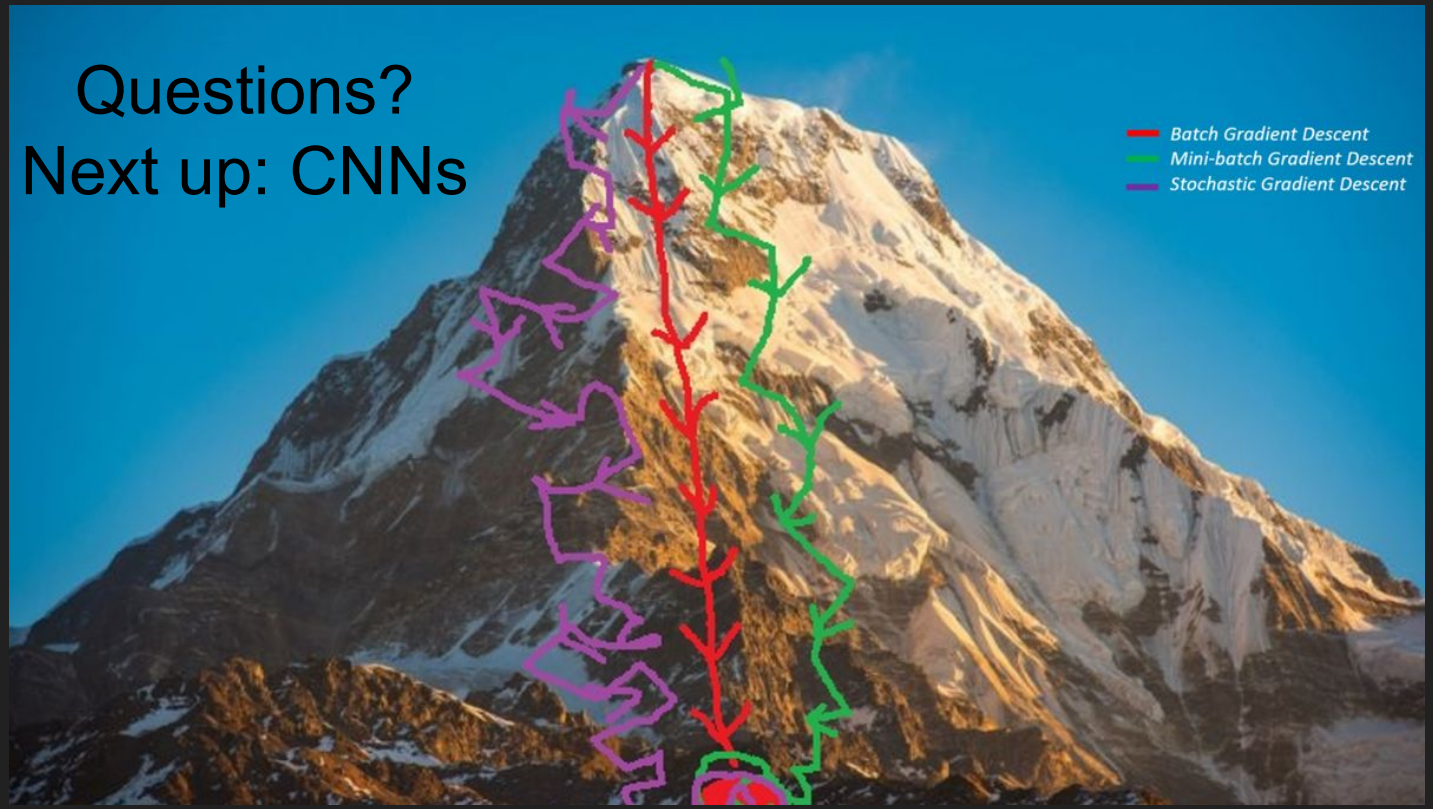
$$\nabla_\theta \mathcal{L}(x, y; \theta) = \sum_{i=1}^{N} \nabla_\theta \ell(x_i, y_i; \theta) \approx \sum_{i=1}^{B} \nabla_\theta \ell(x_i, y_i; \theta)$$

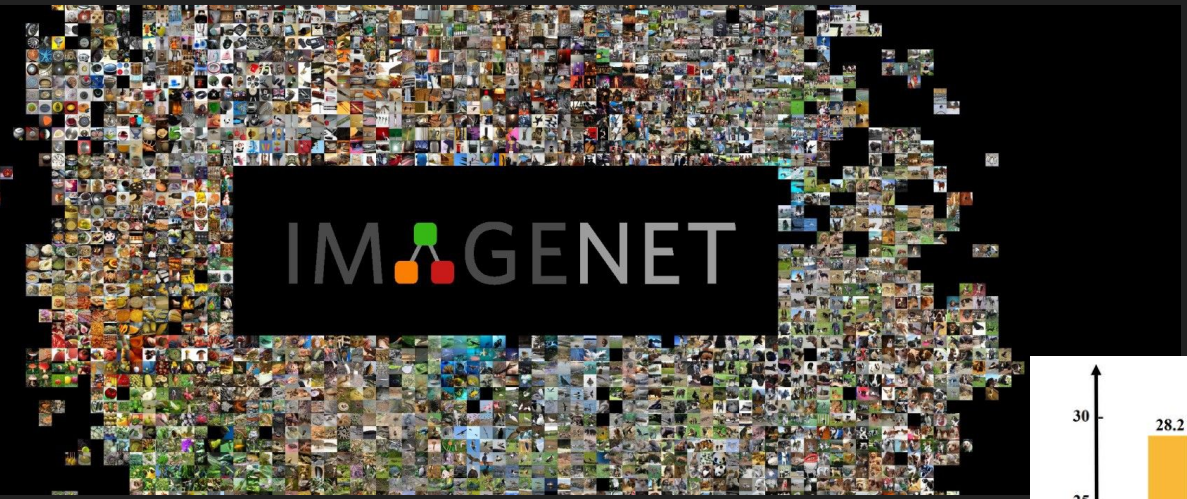
- Properties of SGD
  - Often all data-points don't fit into memory
  - Seems to find better optima (they seem to generalize better)
  - Connected to Bayesian Inference

Challenges?

- Loss landscapes are non-convex
  - somehow not a problem in practice
- Convergence can be really slow
  - Lots of saddle points and flat areas exist
- Design of NN affects the optimization
  - Vanishing & Exploding gradients, e.g. with sigmoid nonlinearity
  - Normalization and regularization interferes with convergence
  - Initial guess for starting point can be decisive
  - …
- Lots of research to improve SGD
  - Momentum, Adam, Adamax, …

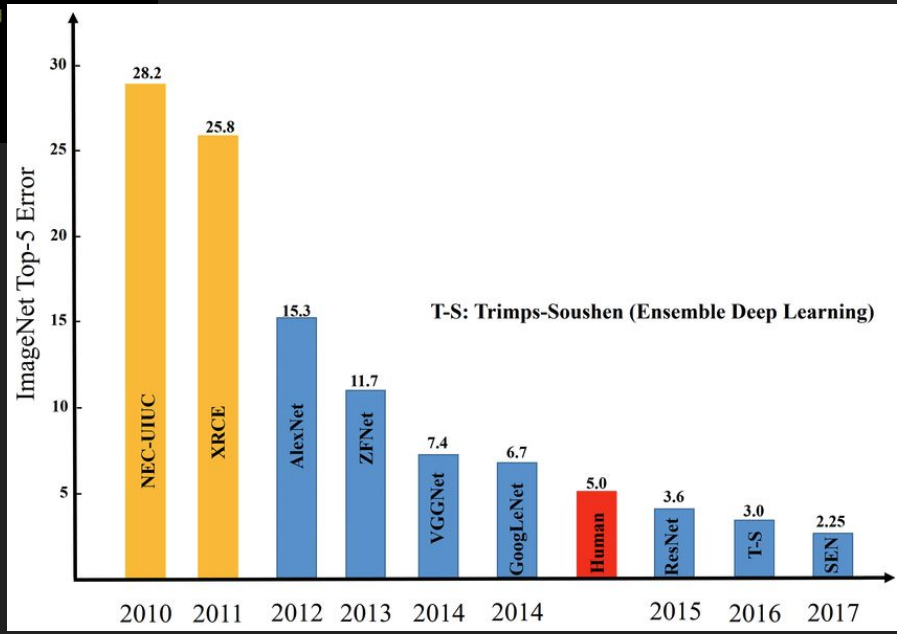
Questions?
Next up: CNNs





Imagenet 2012

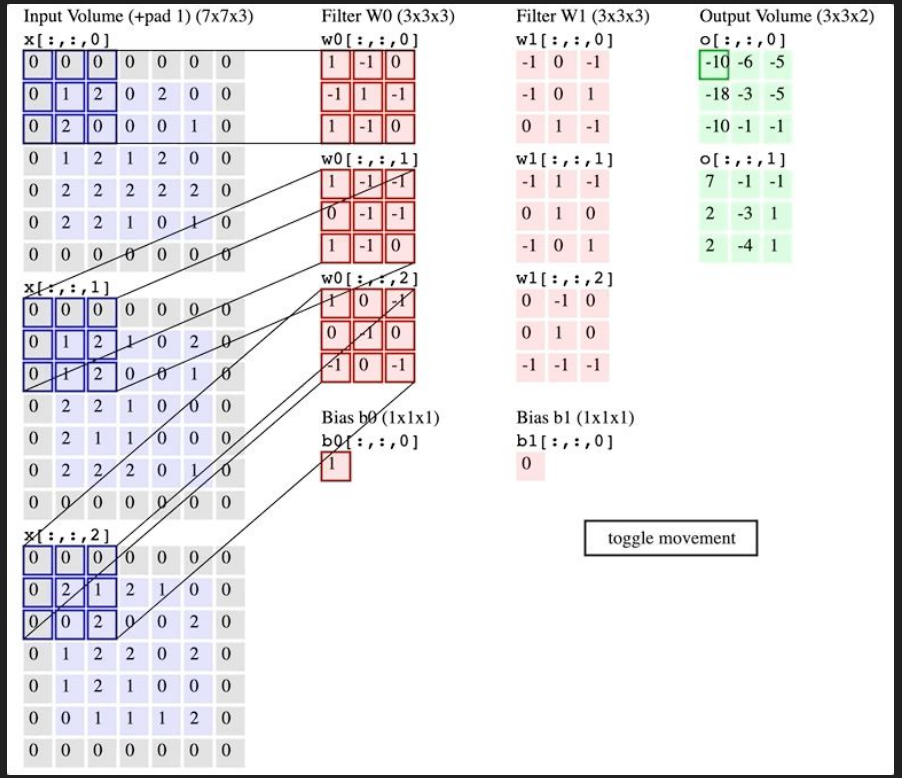
- Transition from hand-engineered features to Neural Networks
- CPU to GPU



[1] Alyafeai et al. (2019), A Fully-Automated Deep Learning Pipeline for Cervical Cancer Classification, 10.1016/j.eswa.2019.112951

Convolutional NNs

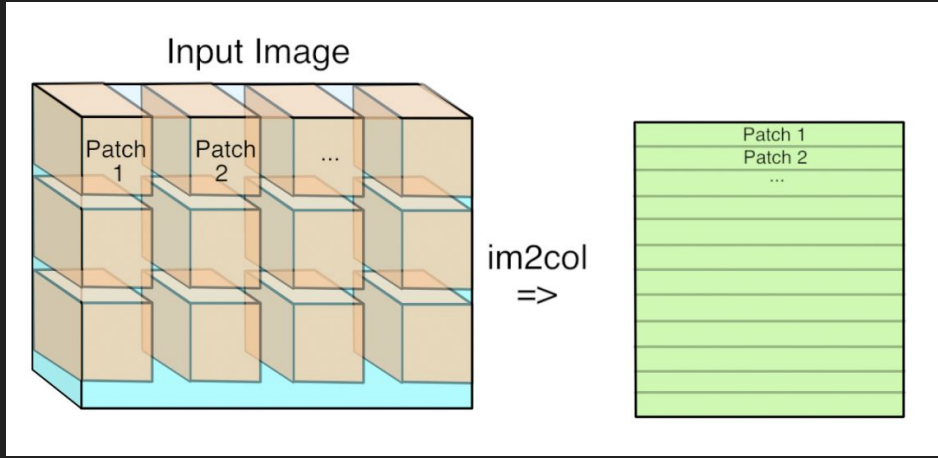
- Convolution Layers and DFTs are closely related
- Images are tensors with shape H x W x C
- Think of C as RGB but can be others depending on application
- Filters will be learned during optimization

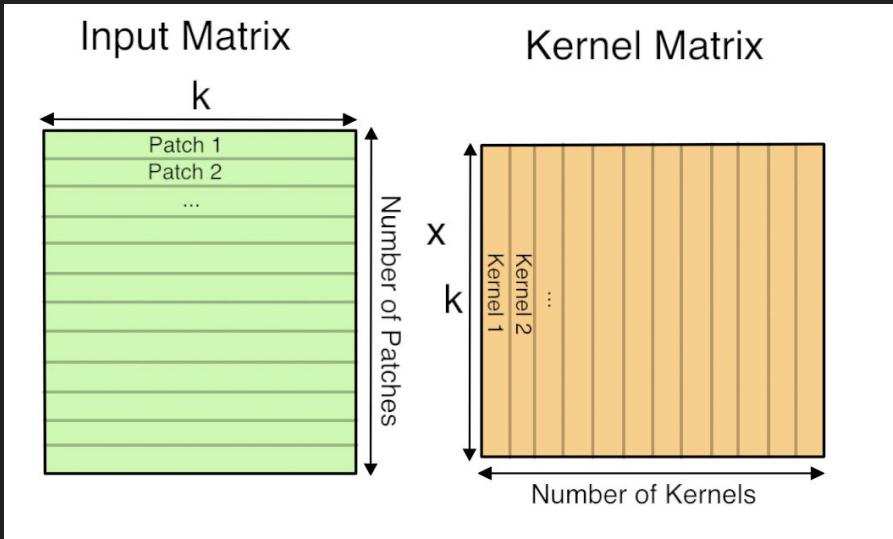


CNNs for Vision Recognition

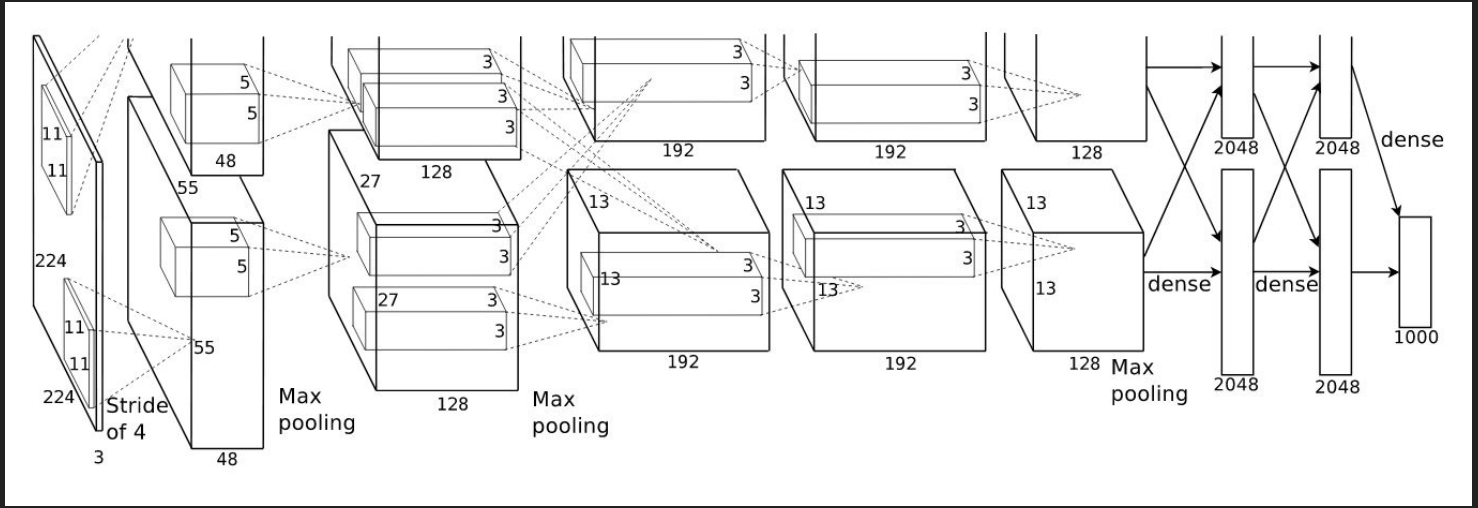
Convolutions as MatMuls

- Patch extraction with sliding window
- Multiplication by Kernel matrix
- Convolutions have typically only a single kernel → weight sharing for efficiency
- Reshape into image-like tensor





AlexNet — Well designed Convolutional NNs



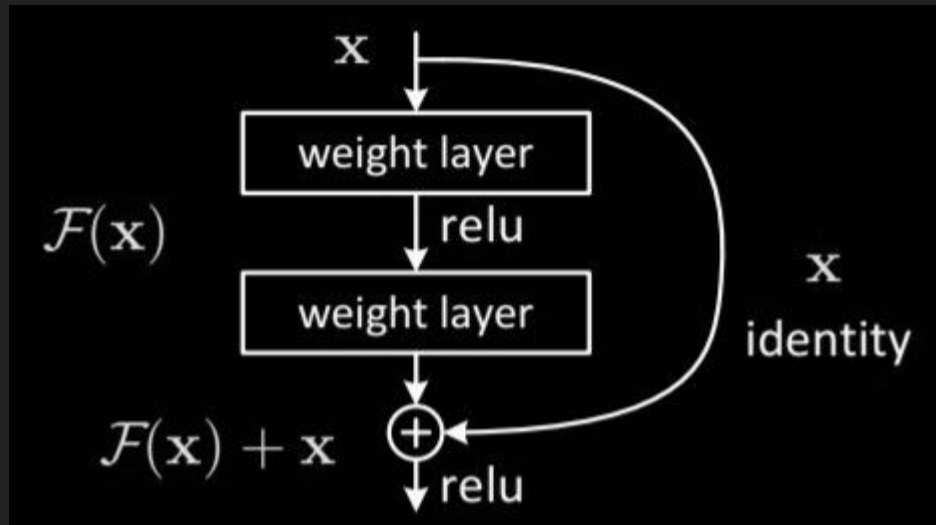
- Stacked Convolutional Layers
- ReLU nonlinearity instead of Tanh
  - Avoids (some) vanishing gradients
  - Numerically much faster!
- Dropout regularization
- Data Augmentation
- Distributed GPU training

Krishevsky et al. 2012, ImageNet Calssification with Deep CNNs

Evolution of CNNs

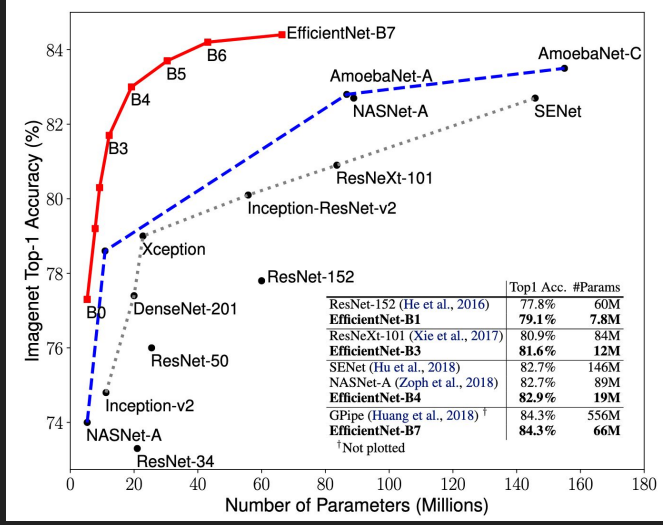
- Increasing depth and width:
  - VGG-16, GoogLeNet
- ResNets
  - He et al., arXiv:1512.03385
  - Replace simple convolutional layer with layers of residual blocks
  - Avoiding of vanishing gradient problems, e.g. with sigmoid non-linearities

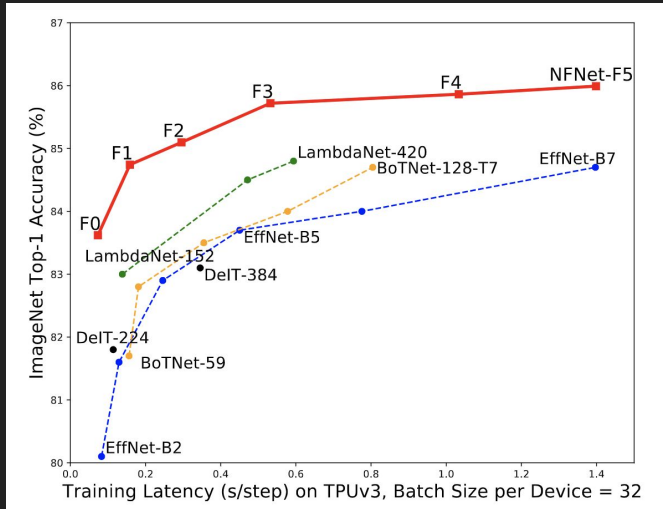
$$x_{l+1} = x_l + F_l(x_l)$$



Evolution of CNNs

- EfficientNet
  - Tan & Le, arXiv:1905.11946
  - Architecture search over scaling laws
  - Compound scaling of Depth, Width & Resolution
- NFNets
  - Brock et al., arXiv:2102.06171
  - removes normalization constraints
  - Improved optimization routine through dynamic gradient clipping





Intermission:
Phase Transitions & Training dynamics

Dynamical Phase Transition & Training Stability

- pre-activation z_i^l with nonlinearity Φ

$$z_i^l = \sum_j W_{ij}^l y_i^l + b_i^l, \qquad y_i^l = \phi(z_i^{l-1})$$

- Parameter distribution in layer l with N_l neurons

$$W_{ij} \sim \mathcal{N}(0, \sigma_W^2/N_l) \qquad b_i \sim \mathcal{N}(0, \sigma_b^2)$$

- What can we say about the quantity $q^l = \mathbb{E}[(z_i^l)^2]$

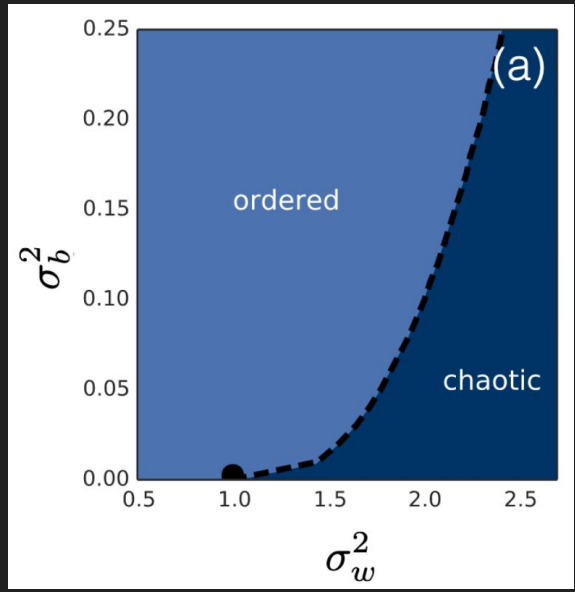
Bahri et al., Statistical Mechanics of Deep Learning

Dynamical Phase Transition & Training Stability

- Fixpoint equation as $l \to \infty$

$$q^l = \sigma_W^2 F(\sqrt{q^{l-1}}) + \sigma_b^2$$

- for Tanh nonlinearity
  - can solve the recursion
  - ordered phase:
    - activations converge towards zero
    - no signal propagation
  - chaotic phase:
    - activations diverge
    - training is unstable



Schoenholz et al., Deep Information Propagation, https://openreview.net/pdf?id=H1W1UN9gg

Verification

- Holds for many architectures
- color codes:
  - white: mean field prediction
  - red: perfect training accuracy
  - black: random chance

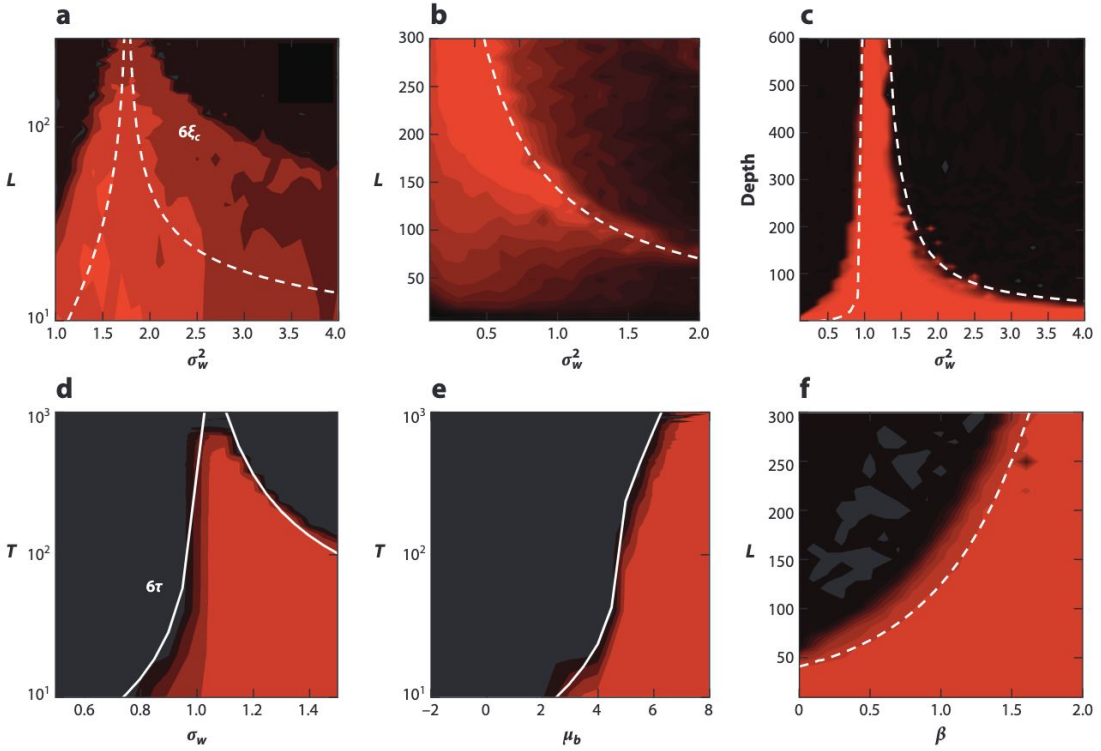
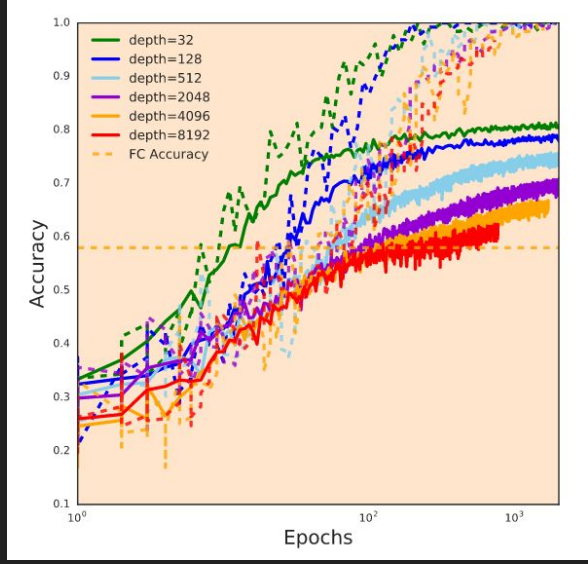


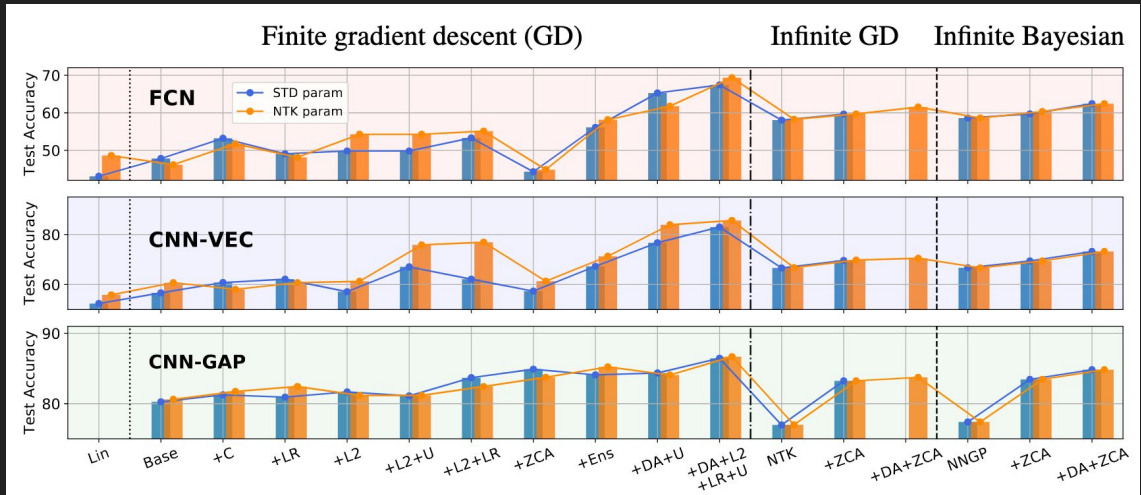
Figure 3

Signal propagation predicts trainability. Each panel shows training accuracy from perfect (*red*) to random chance (*black*) as the hyperparameters of a model are varied over a two-dimensional slice. White lines show mean-field predictions for quantities that determine trainability in each case. In general, we observe excellent agreement over a wide range of architectures. (*a*) Fully connected network compared with the depth scale for signal propagation. (*b*) A residual network compared with a curve of constant gradient norm. (*c*) Convolutional network with the depth scale for signal propagation. (*d–e*) Recurrent neural networks with the timescale for signal propagation. (*f*) Fully connected networks with batch normalization with the depth scale for gradient explosion. See Reference 31 for more details.

Beyond mean-field and NTKs

- Mean-field is good but is lacking some crucial elements
  - Xiao et al, JMLR 2018
  - Generalization to test data degrades with depth
- Meanfield in width leads to different critical point
  - J. Lee et al arXiv:2007.15801
  - Neural tangent kernels
  - NN-Gaussian Processes
  - etc.





Back to architectures

CNNs

- Workhorse of computer vision:
  - Parameter efficient
  - Good inductive bias
  - Time-tested with specialized kernels to speed up training and inference
  - Real test ground for theorists

- Way forward?
  - Not all data is image-related
  - CNN are subset of MLPs
  - weights are static in a sense that they do not depend on the input

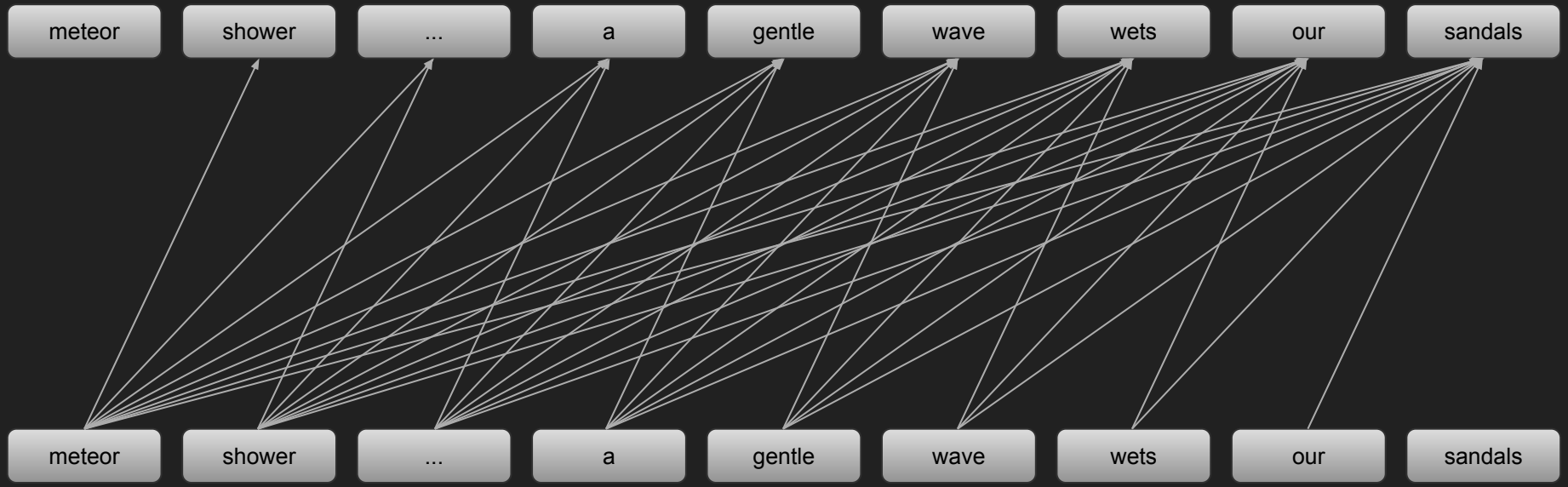
Attention Mechanism and Transformers

- Sequential / causal datasets require different inductive bias
  - historically used LSTM or RNN
  - challenges due to signal propagation and unstable gradient dynamics
- Natural modeling of sequences using iterated Bayes rule

$$P(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_{i-1}, \ldots, x_1)$$

- This ordering has an autoregressive structure

Autoregressive Transformer Architecture



— Haiku by Michael Dylan Welch, *HSA Newsletter* XV:4, Autumn 2000

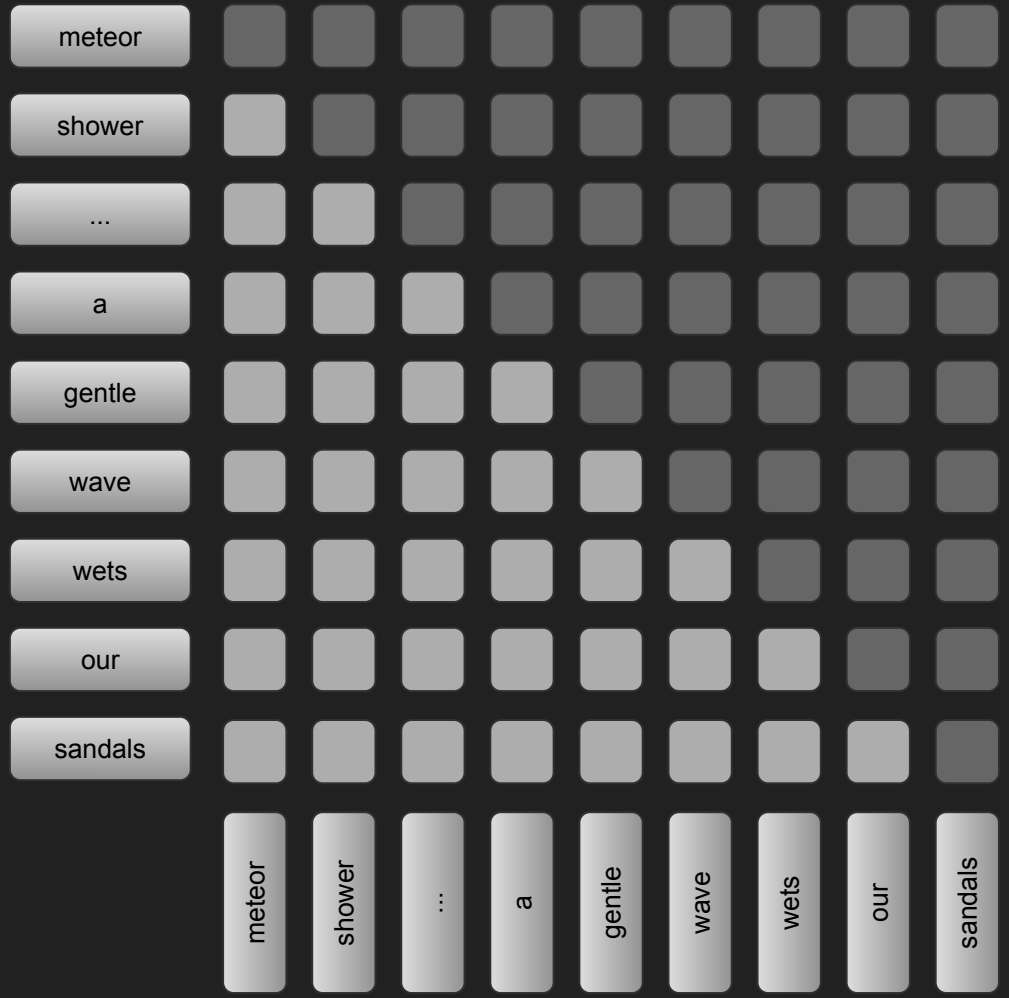
- Dependencies can be modeled through adjacency graph
- Create sequence vectors

$$q^T = Q \cdot x$$

$$k^T = K \cdot x$$

- Use mask m to construct attention score

$$a = m \odot (q \cdot k^T)$$



Softmax and multi-headed attention

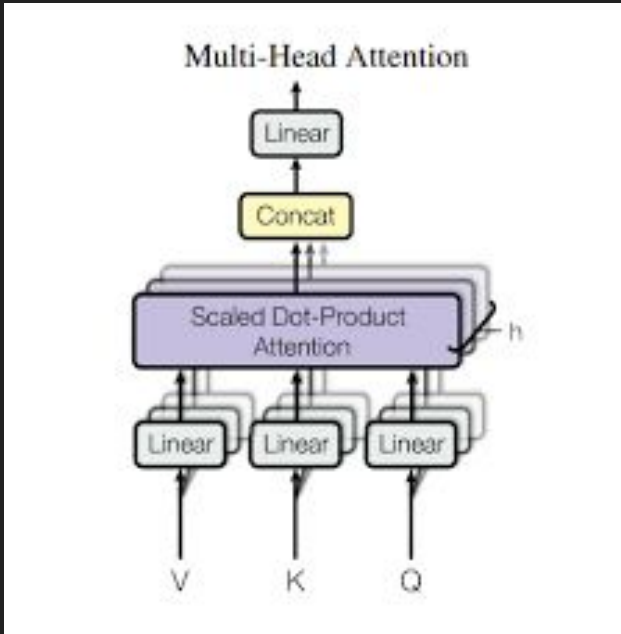
- Use another vector to project back to relevant sequence dimension

$$v^T = V \cdot x$$

- Full attention kernel

$$\text{Attn}[Q, K, V] = \text{softmax}(qk^T) \cdot v$$

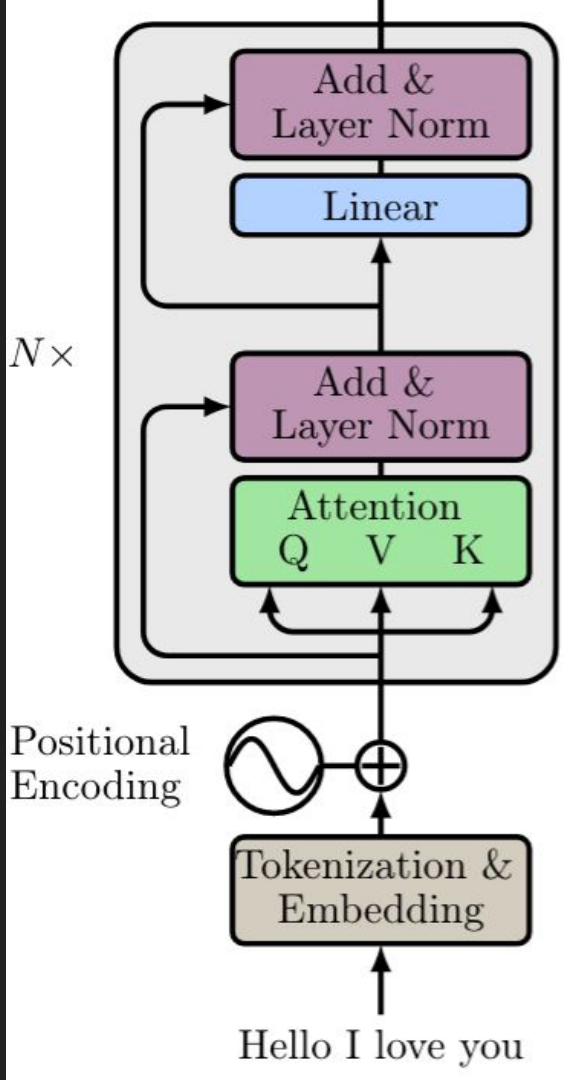
- To optimize parameter use, repeat using several “heads”



Softmax and multi-headed attention

- Transformer Layers are stacked blocks of
  - Multi-headed attention
  - Fully connected feed-forward networks

- Properties
  - $O(T^2D)$ operations for sequence length T and embedding D
  - Due to the QK product the weights for each layer update are dynamic
  - Strictly more expressive than simple MLP architectures.



GPT-2 & NLP

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.

The Nuclear Regulatory Commission did not immediately release any information.

According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

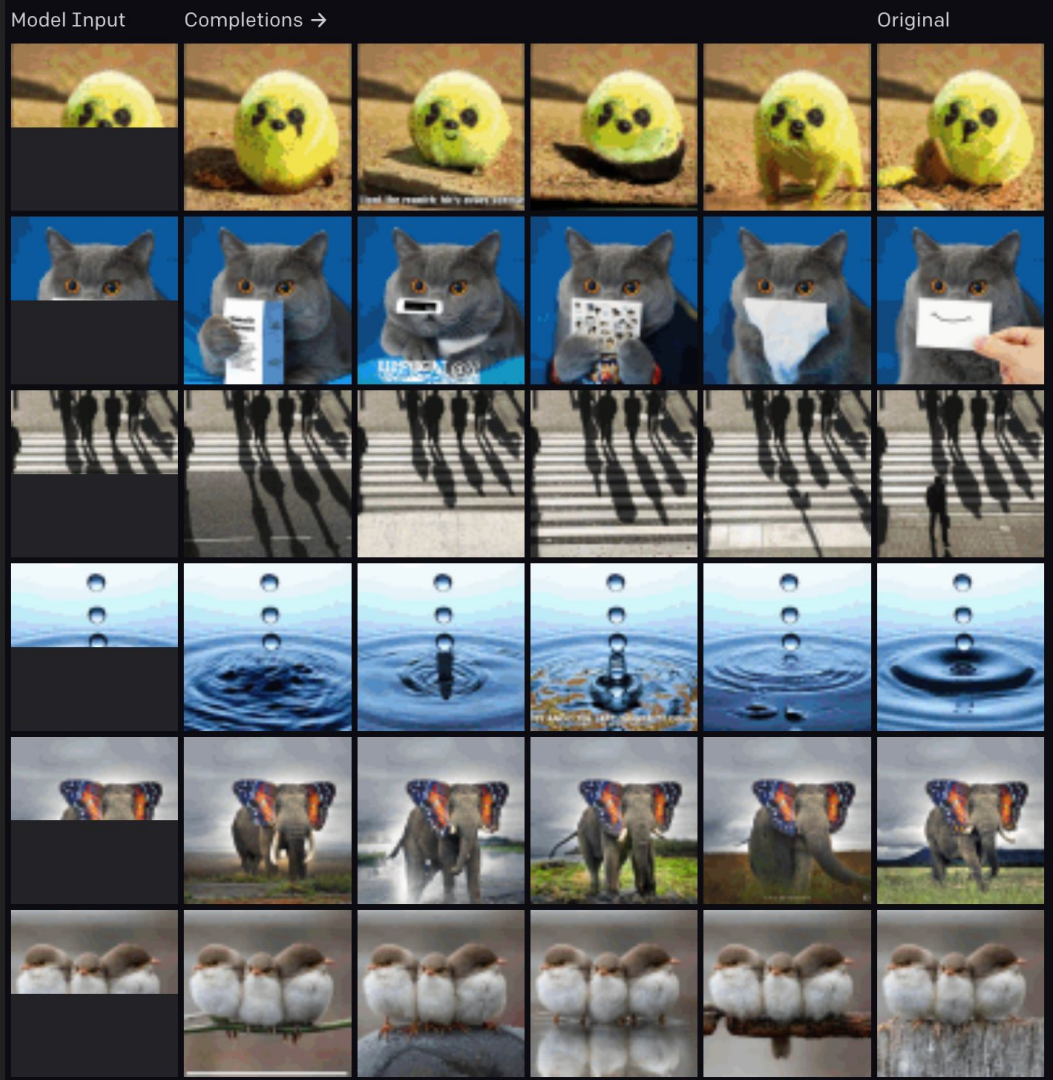
“The safety of people, the environment and the nation’s nuclear stockpile is our highest priority,” Hicks said. “We will get to the bottom of this and make no excuses.

DATASET	METRIC	OUR RESULT	PREVIOUS RECORD	HUMAN
Winograd Schema Challenge	accuracy (+)	**70.70%**	63.7%	92%+
LAMBADA	accuracy (+)	**63.24%**	59.23%	95%+
LAMBADA	perplexity (–)	**8.6**	99	~1-2
Children's Book Test Common Nouns (validation accuracy)	accuracy (+)	**93.30%**	85.7%	96%
Children's Book Test Named Entities (validation accuracy)	accuracy (+)	**89.05%**	82.3%	92%
Penn Tree Bank	perplexity (–)	**35.76**	46.54	unknown
WikiText-2	perplexity (–)	**18.34**	39.14	unknown
enwik8	bits per character (–)	**0.93**	0.99	unknown
text8	bits per character (–)	**0.98**	1.08	unknown
WikiText-103	perplexity (–)	**17.48**	18.3	unknown

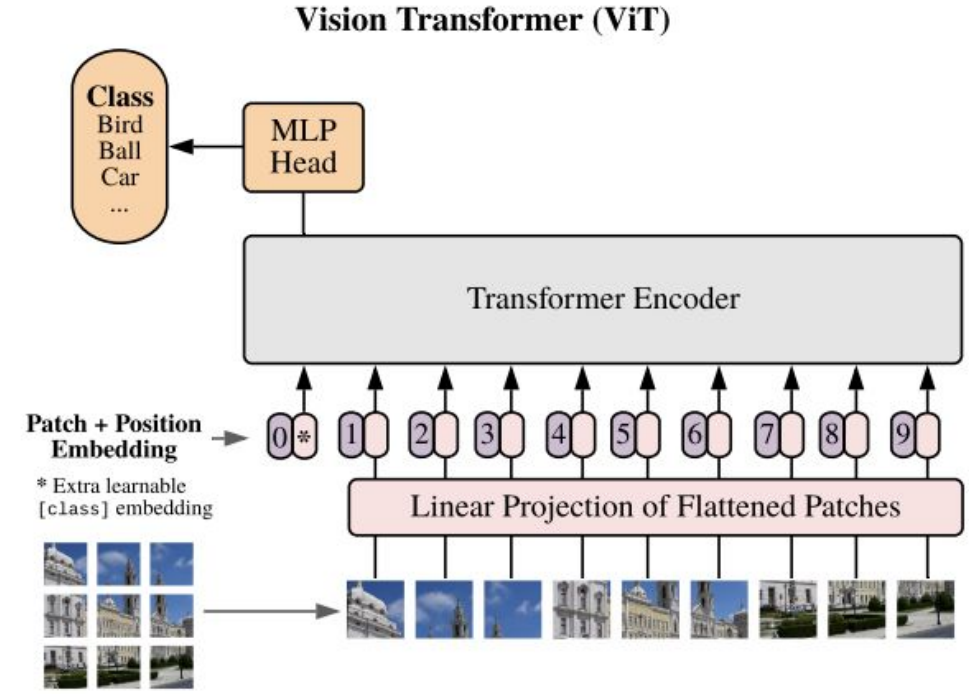
GPT-2 achieves state-of-the-art on Winograd Schema, LAMBADA, and other language modeling tasks.

iGPT

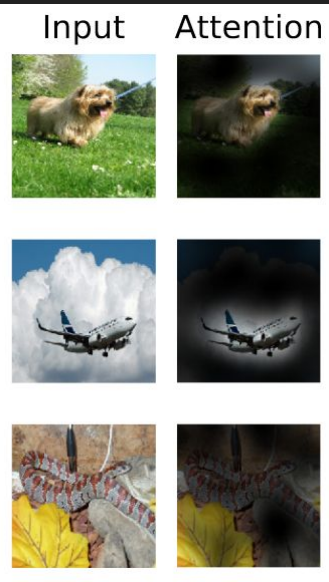
- Model images as sequences
- Realistic completions
- New SotA results for CIFAR10 classification without being trained on CIFAR10



ViT



- Input tokens are image patches
- **Outperforms CNNs on the same data**



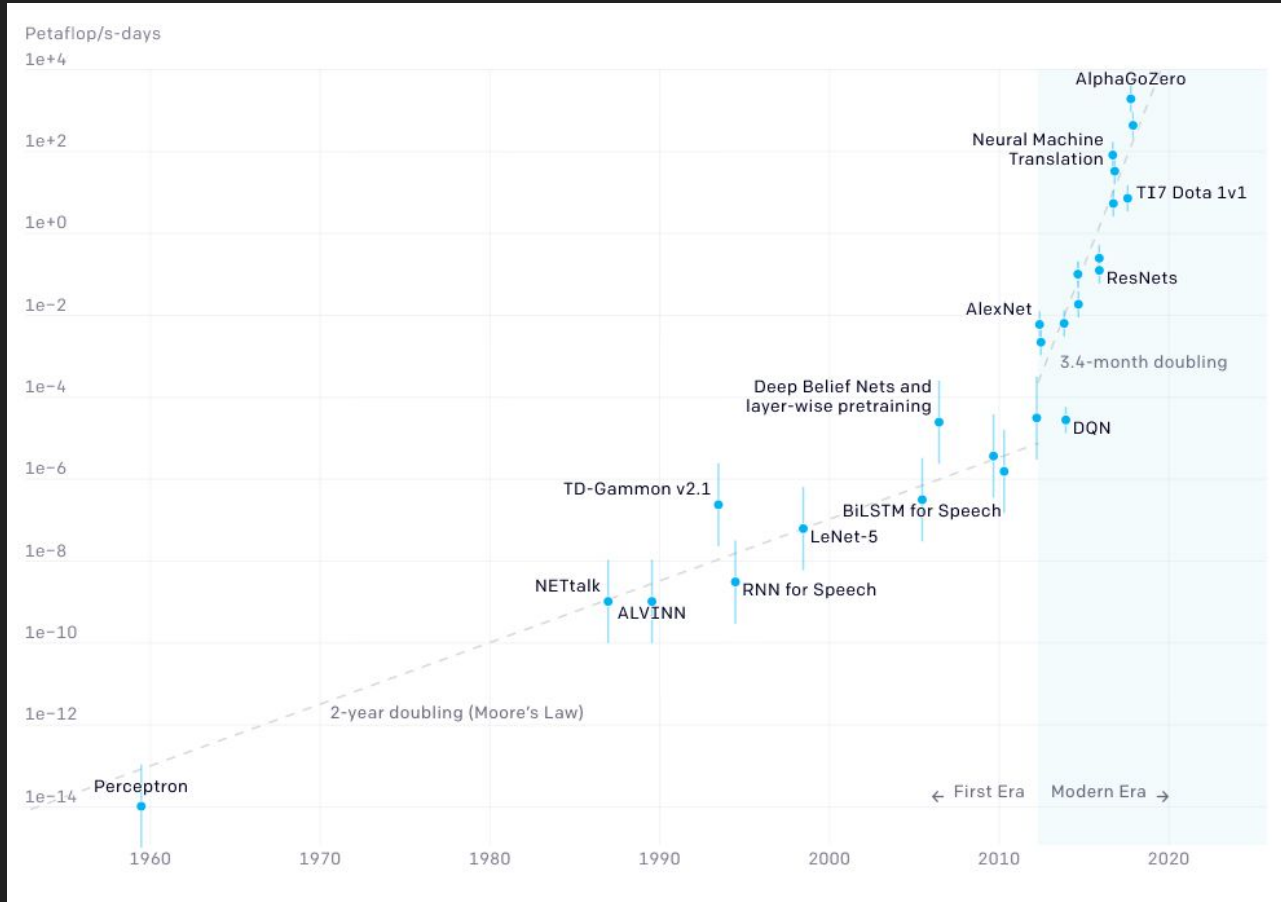
Transformer — Current state

- Sparse, factorized & switch-board attention
- Long context transformers
- Introduce memory states and recurrence
- Speeding up softmax computation using RKHS
- …

Scaling Laws
and
Compute Requirements

Big results require big compute !?

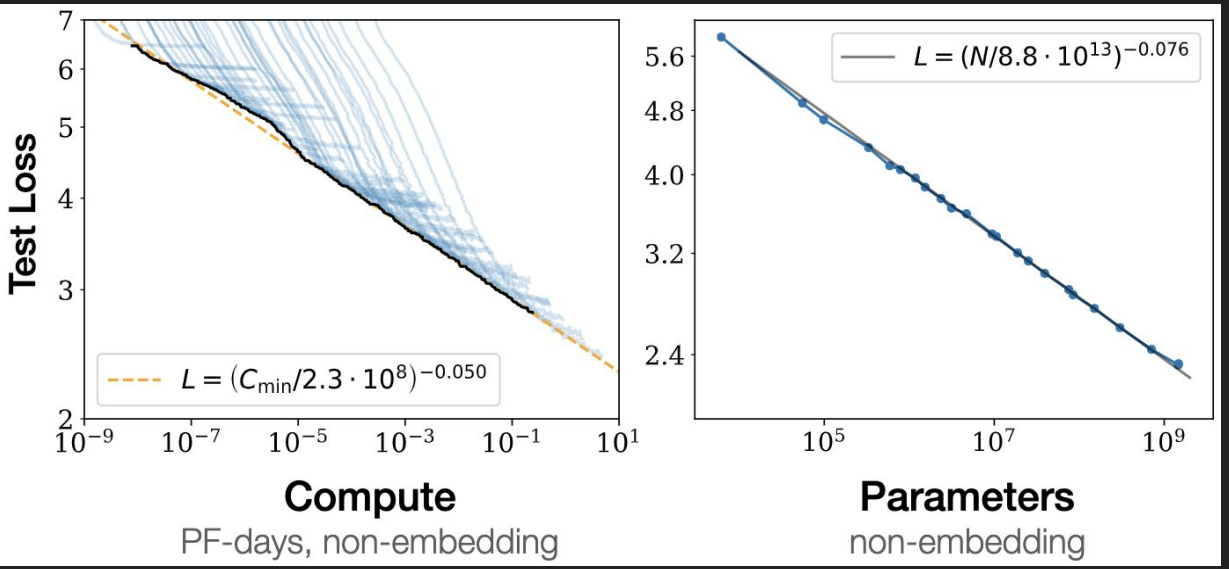
- Moore’s law of AI resources
- before 2012: 2-year doubling period
- since 2012: 3.4-month doubling period



Big results can become more efficient

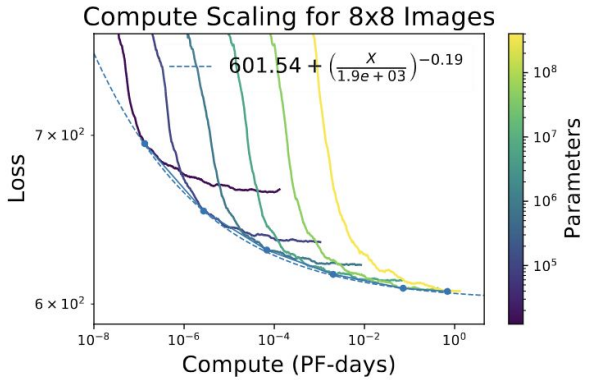
- Moore’s law of AI efficien
- Since 2012:
  16-month halving time

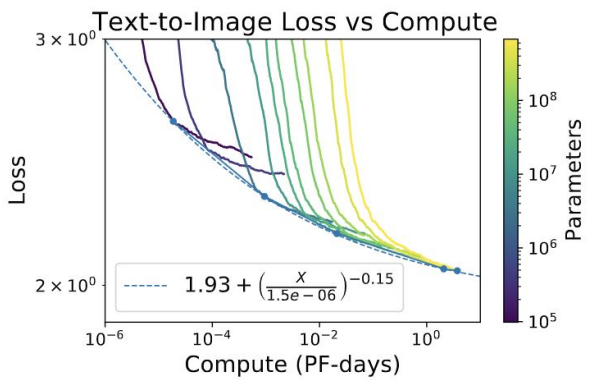
Scaling laws for transformers

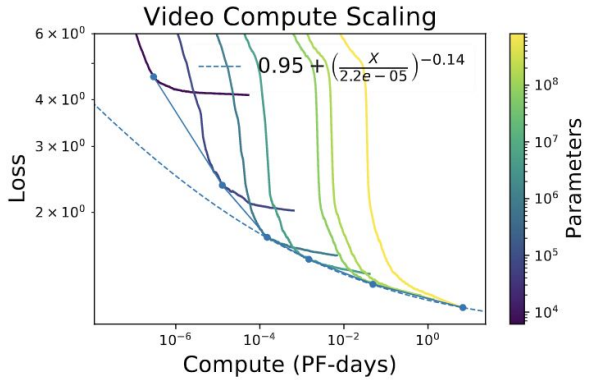


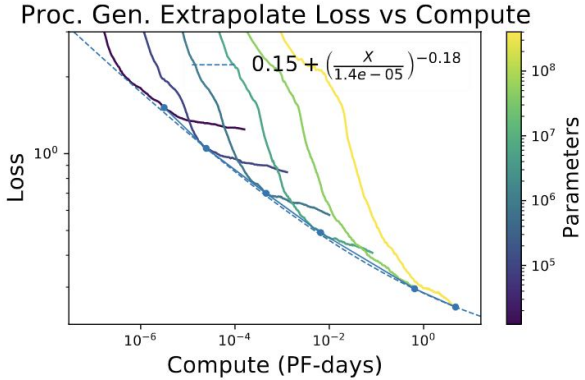
- Pareto frontier of negative log-loss is predictable
- power-law in
  - Compute
  - number of parameters
- Prediction of model performance possible

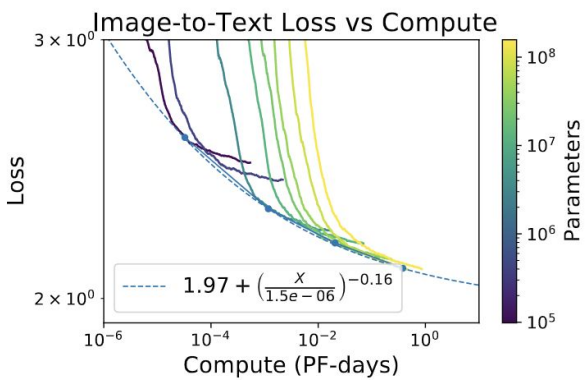
Scaling laws across domains

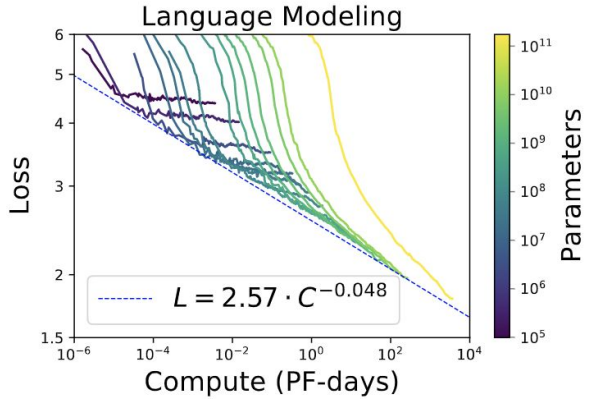






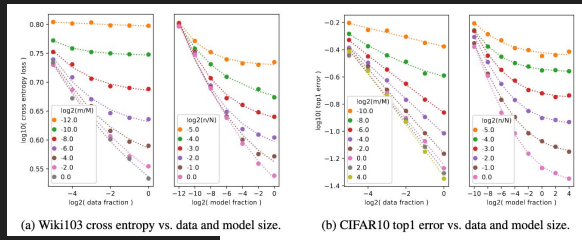




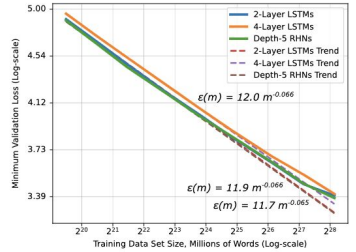


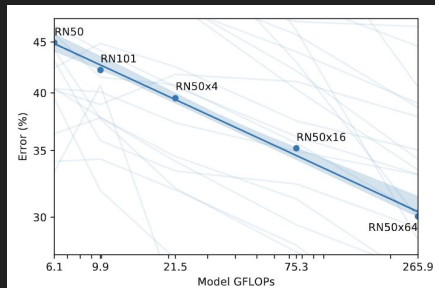
Scaling laws seem to be ubiquitous …

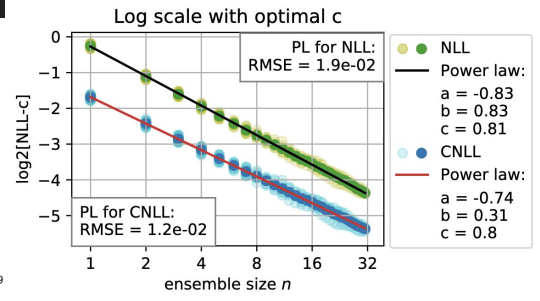
- … and not well understood (yet)
- They appear in:
  - Training loss curves
    - spectrum of the time-dependent Hessian seems important
    - also Glass to Jamming phase transitions occur here
  - compute (= seen samples * model parameter) resources
  - dataset size
  - Ensemble averages of performance indicators for various tasks (see CLIP)
  - model ensembles
  - ...



(a) Wiki103 cross entropy vs. data and model size. (b) CIFAR10 top1 error vs. data and model size.

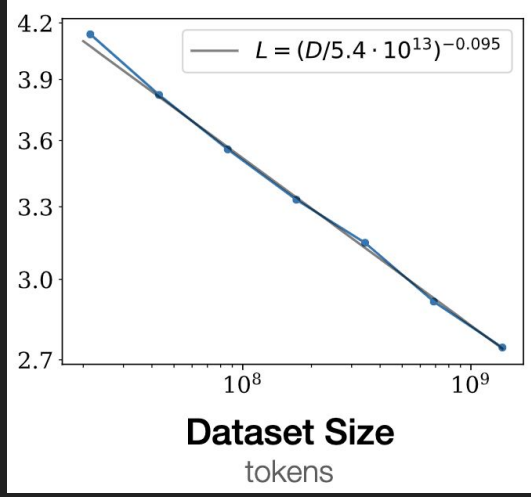


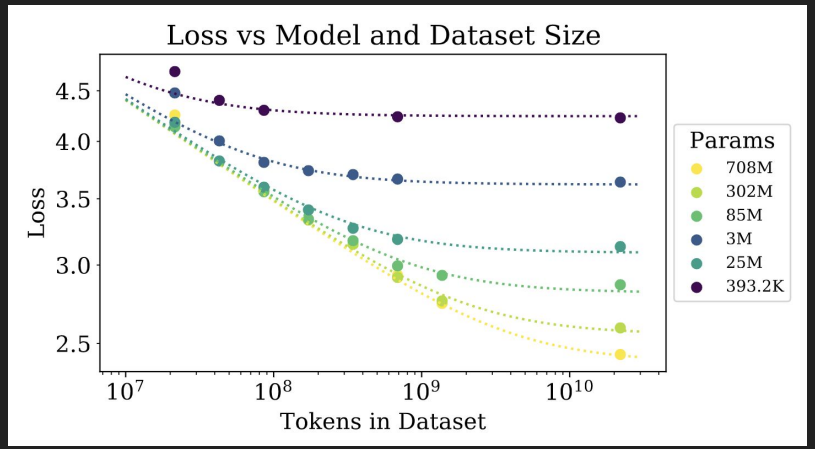


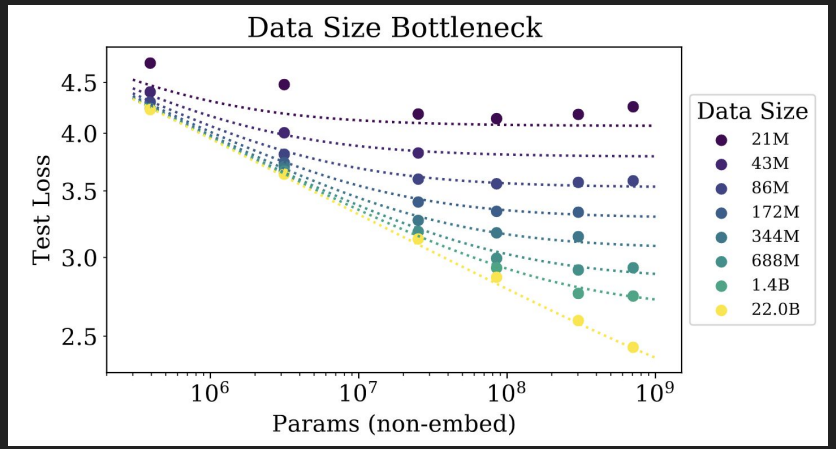


Scaling with dataset size

- Collecting more data is a sure fire way to improve your model
- Dataset and model size are not independent

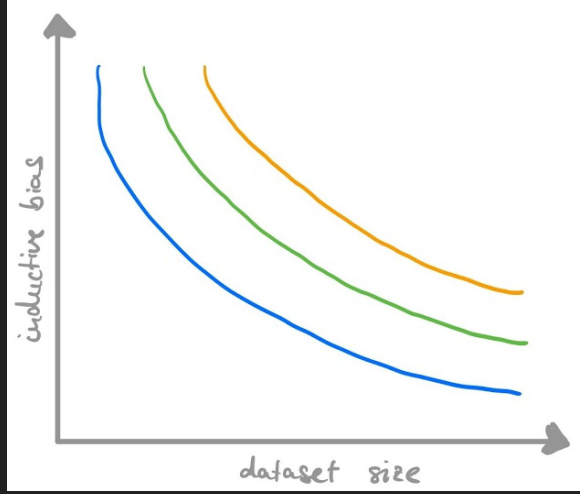


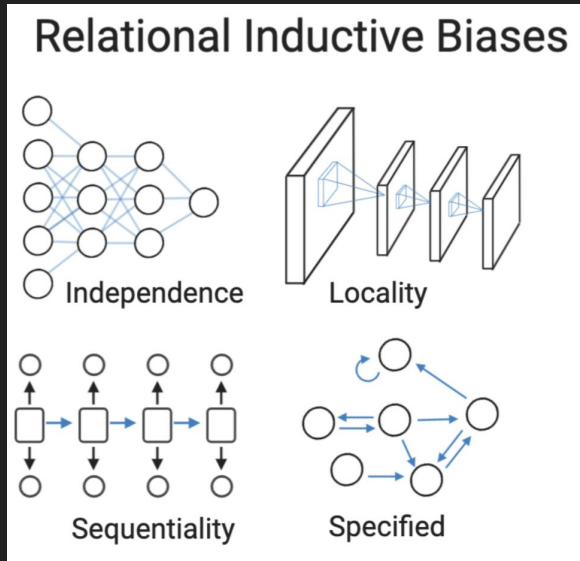




Dataset size and inductive bias

- Inductive biases reduce data requirements
  - Are akin to encoding prior knowledge
  - force parameter evolution onto a manifold changing the effective degrees of freedom
- Remove inductive bias in order to enable structure discovery
  - Needs more data in order to find patterns



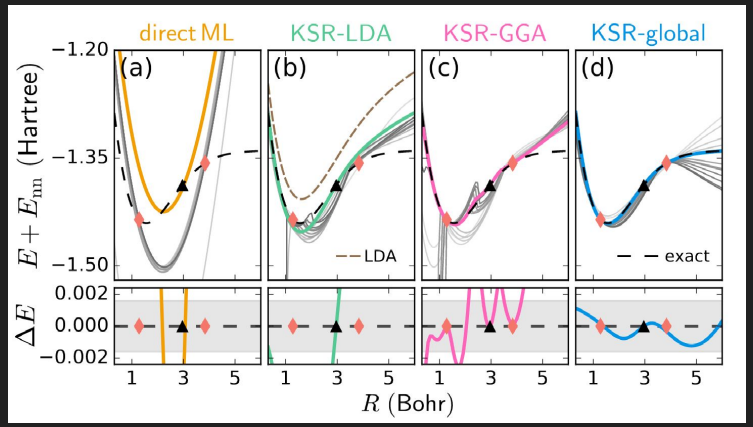


take away

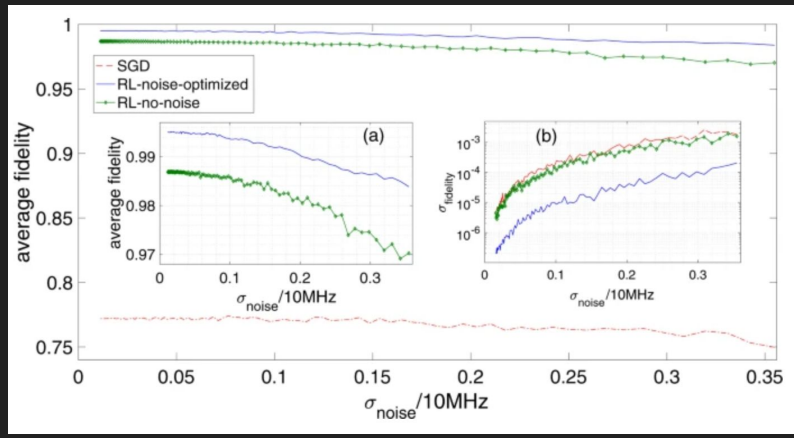
- Loss of neural network is predictable
- Scaling laws in
  - size of dataset
  - number of parameters
  - amount of compute (FLOP)
- Inductive bias can help reduce the amount of any of the above variables
  - trade-off: Model becomes less “general” and more “purpose-tailored”

Some applications

Applications in Science



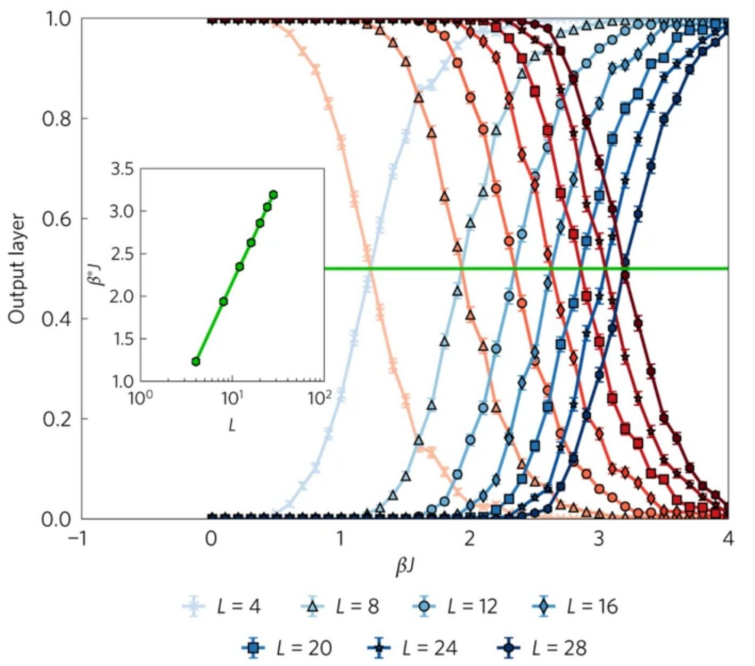
Li et al., arXiv:2009.08551
Backprop through the Kohn-Sham DFT structure equation

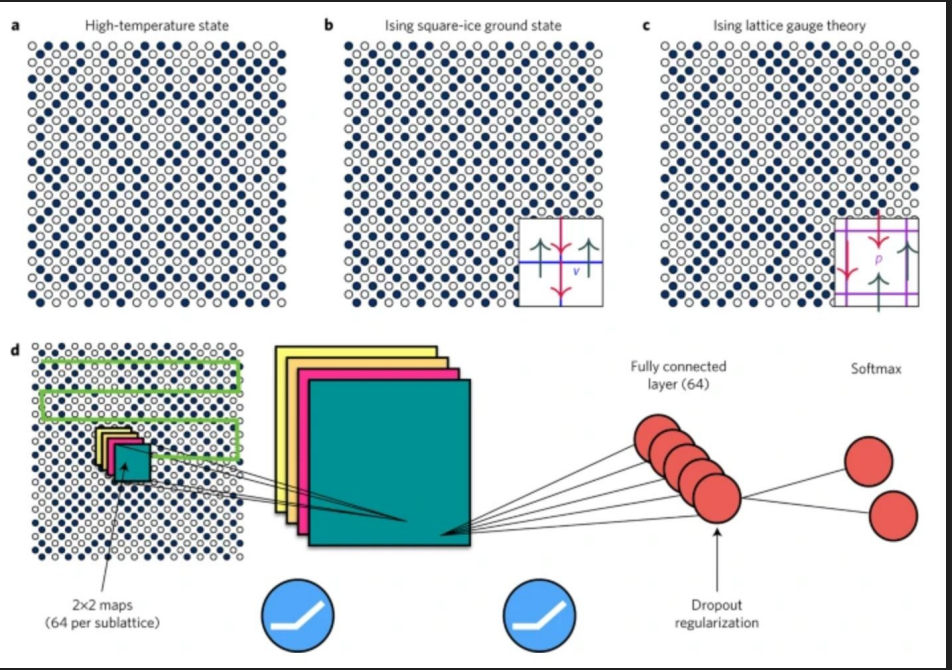


Niu et al, npj Q-Info 5, 33 (2019)
Use RL to tune and control quantum gates for SC qubits

Both use various forms of inductive bias

Applications in Science

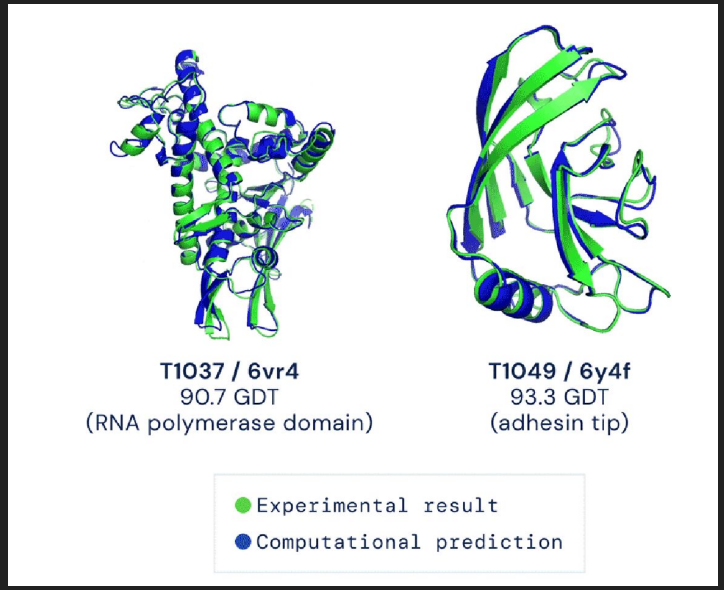


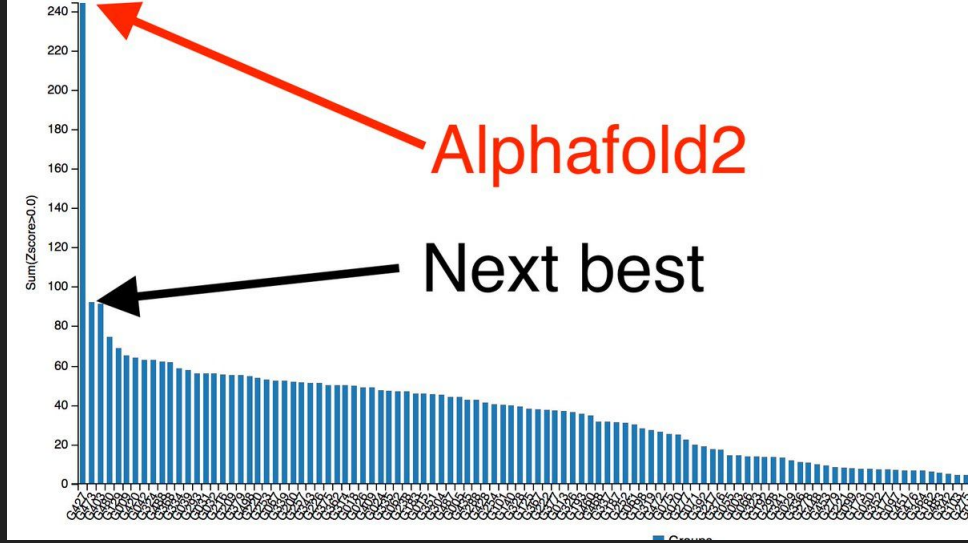


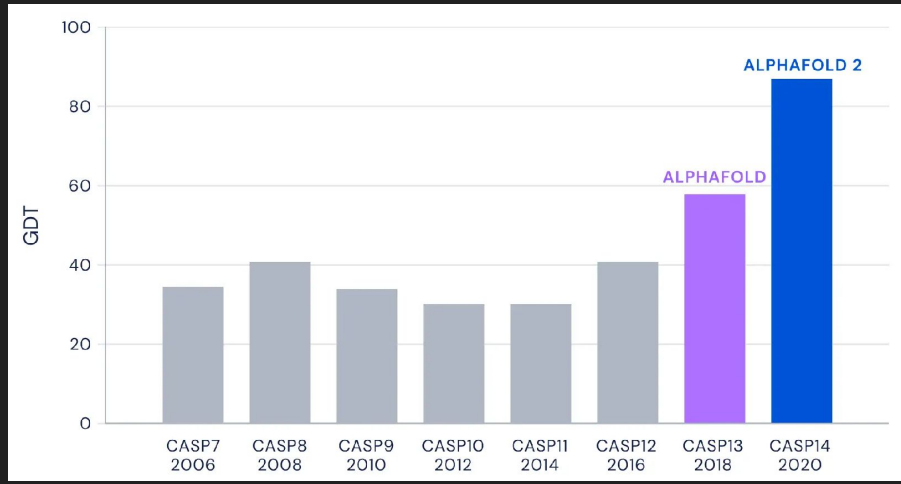
Carrasquilla & Melko, Nature Phys. 13, 431 (2017)

Using CNN to estimate the critical temperature of an Ising model based on Monte Carlo samples.

Applications in Science







ImageNet-like moment:
DeepMind solves the Protein-Folding problem

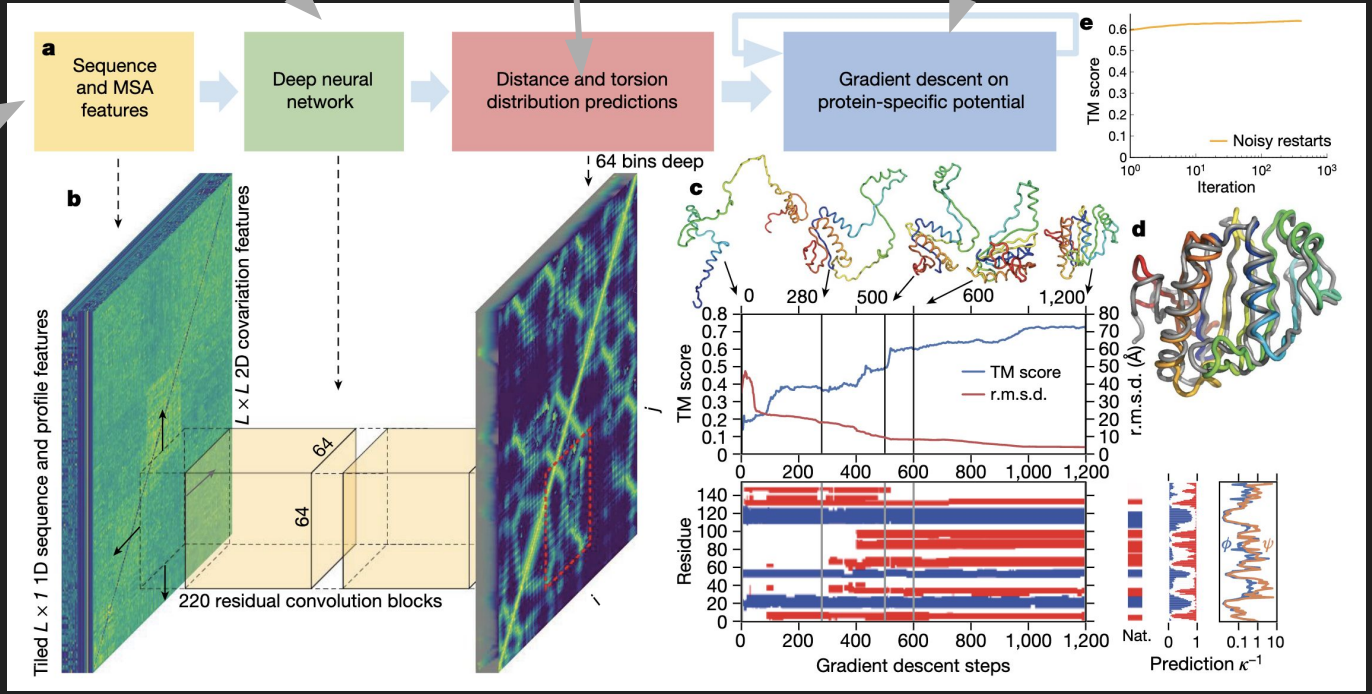
AlphaFold

SotA-style ResNet

Supervised target is **intermediate representation**

Differentiable structure prediction

Feature Engineering as Inductive Bias and Data Augmentation



AlphaFold 2

Feature Engineering as Inductive Bias and Data Augmentation

Transformer (best guess)

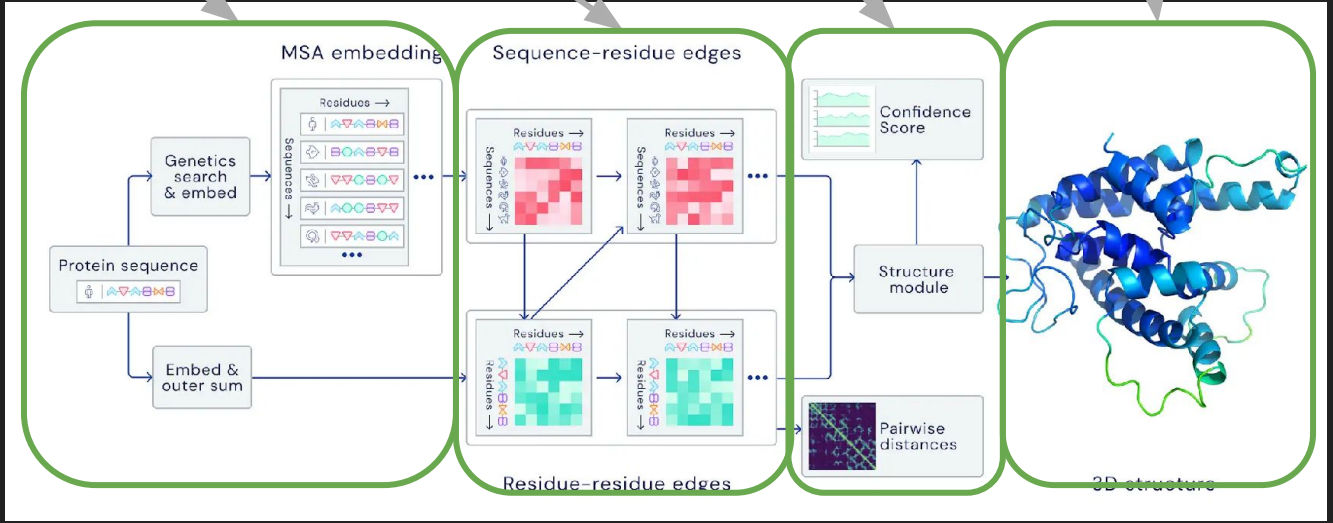
Supervised target is intermediate representation

Differentiable structure prediction (maybe feedback to transformer network)

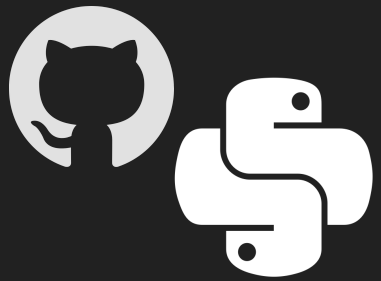
Uses all knobs

- SotA architecture
- Data augmentation and inductive bias
- model and compute scale

- ca 150-200 GPUs for a few weeks



Pushing the Needle in or with AI

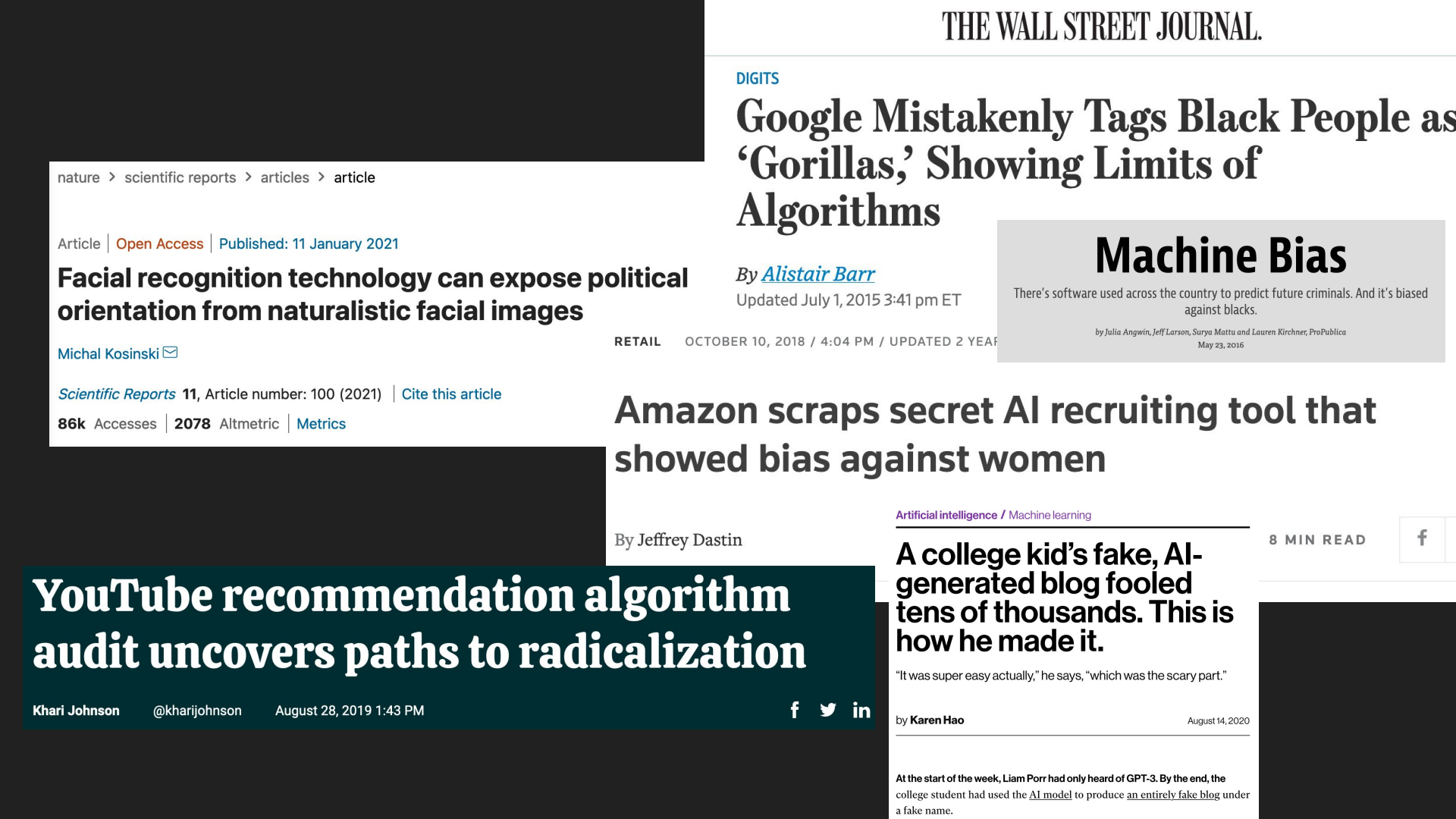


- AI is an experimental science (for now)
  - theory has a lot of catching up to do
  - Consider your experiment design carefully
    - E.g. train, validation, test split
    - data & code hygiene
- Code is a research product and deliverable
  - Your studies need to be **reproducible**
  - Code needs to be **Open Sourced** for others to build upon and improve
  - Consider open-sourcing the data to train and evaluate your model
- Develop Benchmarks as a community
  - To compare approaches or simply measure improvements (see CASP, ImageNet competitions)
  - Luckily Physics has a bunch of them

take-outs ...

- Optimization techniques and improvements
  - learned optimizers and Monte Carlo techniques
  - zeroth-, first-, second-order optimization
  - Initialization techniques
- Other common model structures and architectures
  - (Variational) Auto-Encoders (VAE), Generative Adversarial Networks (GAN), Normalizing Flows, Energy-based model, Non-equilibrium models, LSTMs, Graph NNs, GPNNs, ...
- Other research avenues
  - Active Learning, Robustness and Generalization, Interpolation vs Extrapolation, Reinforcement Learning & Optimal Control

PSA
AI, Society & Ethics


nature > scientific reports > articles > article
Article | Open Access | Published: 11 January 2021
Facial recognition technology can expose political orientation from naturalistic facial images
Michal Kosinski
Scientific Reports 11, Article number: 100 (2021) | Cite this article
86k Accesses | 2078 Altmetric | Metrics
THE WALL STREET JOURNAL.
DIGITS
Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms
By Alistair Barr
Updated July 1, 2015 3:41 pm ET
Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016
RETAIL OCTOBER 10, 2018 / 4:04 PM / UPDATED 2 YEAR
Amazon scraps secret AI recruiting tool that showed bias against women
By Jeffrey Dastin
YouTube recommendation algorithm audit uncovers paths to radicalization
Khari Johnson @kharijohnson August 28, 2019 1:43 PM
Artificial intelligence / Machine learning
8 MIN READ
A college kid's fake, AI-generated blog fooled tens of thousands. This is how he made it.
"It was super easy actually," he says, "which was the scary part."
by Karen Hao August 14, 2020
At the start of the week, Liam Porr had only heard of GPT-3. By the end, the college student had used the AI model to produce an entirely fake blog under a fake name.

- AI is and will become even more powerful
- AI has real-world impact! Be conscientious!
- Think about the subject of your experiment
  - In Big Tech: It's **HUMANS**
- AI is dual use technology
  - Face Recognition: Police State vs. protection of people with Alzheimer's
  - Reinforcement Learning: Alpha Dogfight vs household robot