



DEEP LEARNING

Ian Goodfellow, Yoshua Bengio,
and Aaron Courville

A DEEP LEARNING JOURNEY

YOSHUA BENGIO

New in ML NeurIPS 22 mentorship workshop
November 28th, 2022



IVADO

CIFAR

Falling in Love with a Research Direction

1985-1986:

- finishing my undergrad and wondering what I would do next
- reading lots of papers in various areas
- zoomed in on the neural networks research
- read papers by Geoff Hinton, David Rumelhart and other early connectionists

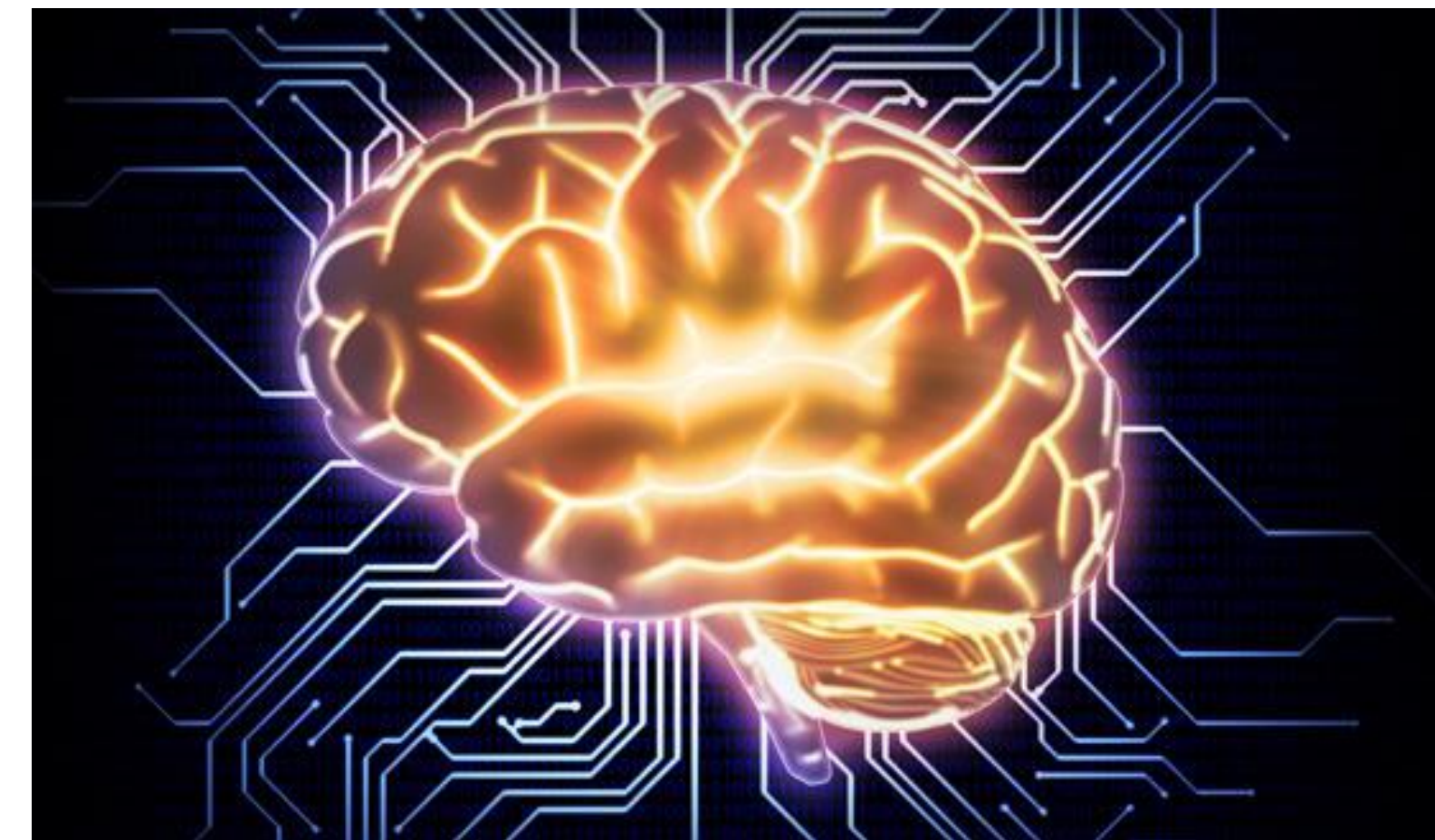
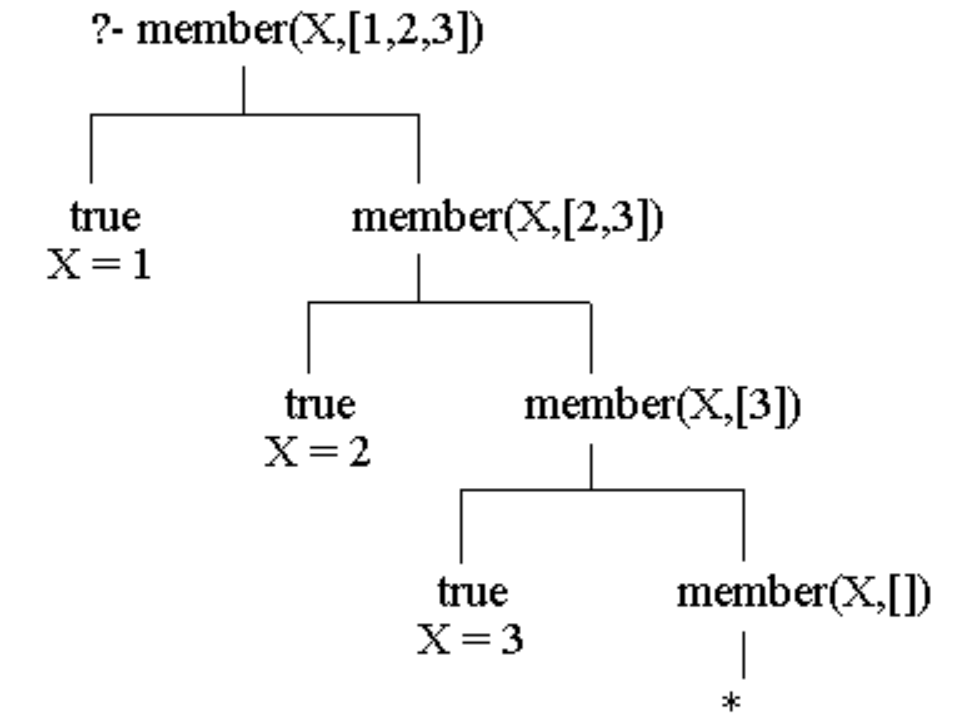
1986-1988:

- read about Boltzmann machine, implemented it for phoneme classification
- MSc thesis on Boltzmann machines for speech recognition
- read about backpropagation, got excited about it, started working with it
- went to 1988 Connectionist Summer School and met many other passionate graduate students and researchers

1988-1991: PhD thesis on neural nets (RNNs and ConvNets) and HMMs hybrids

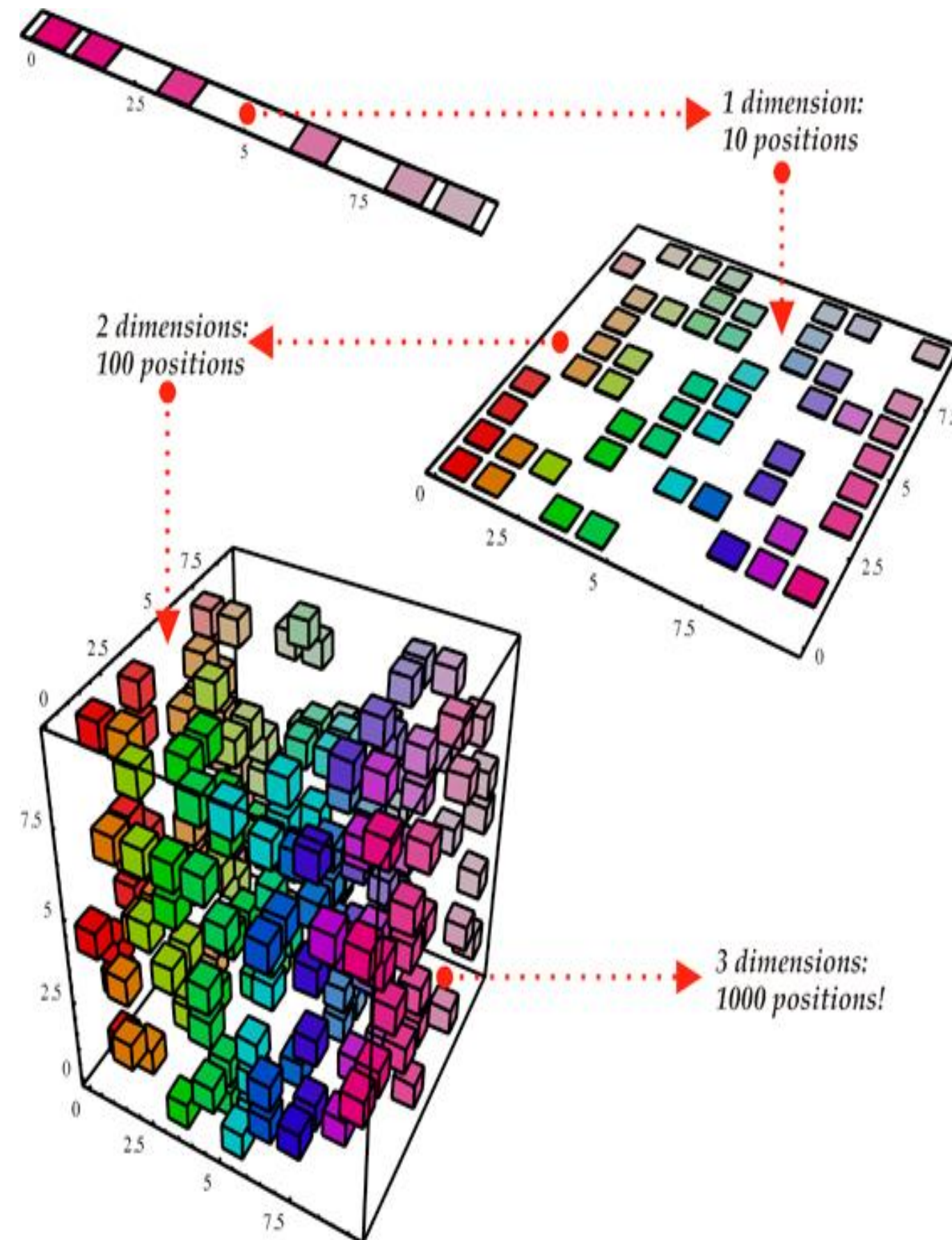
The Machine Learning approach to AI

- **Classical AI, rule-based, symbolic**
 - knowledge is provided by humans
 - but intuitive knowledge (e.g. much of common sense) not communicable
 - machines only do inference (using designed alg.)
 - no strong learning, adaptation
 - insufficient handling of uncertainty
 - not grounded in low-level perception and action
- **Machine learning tries to fix these problems**
 - succeeded to a great extent
 - higher-level (conscious) cognition not achieved yet



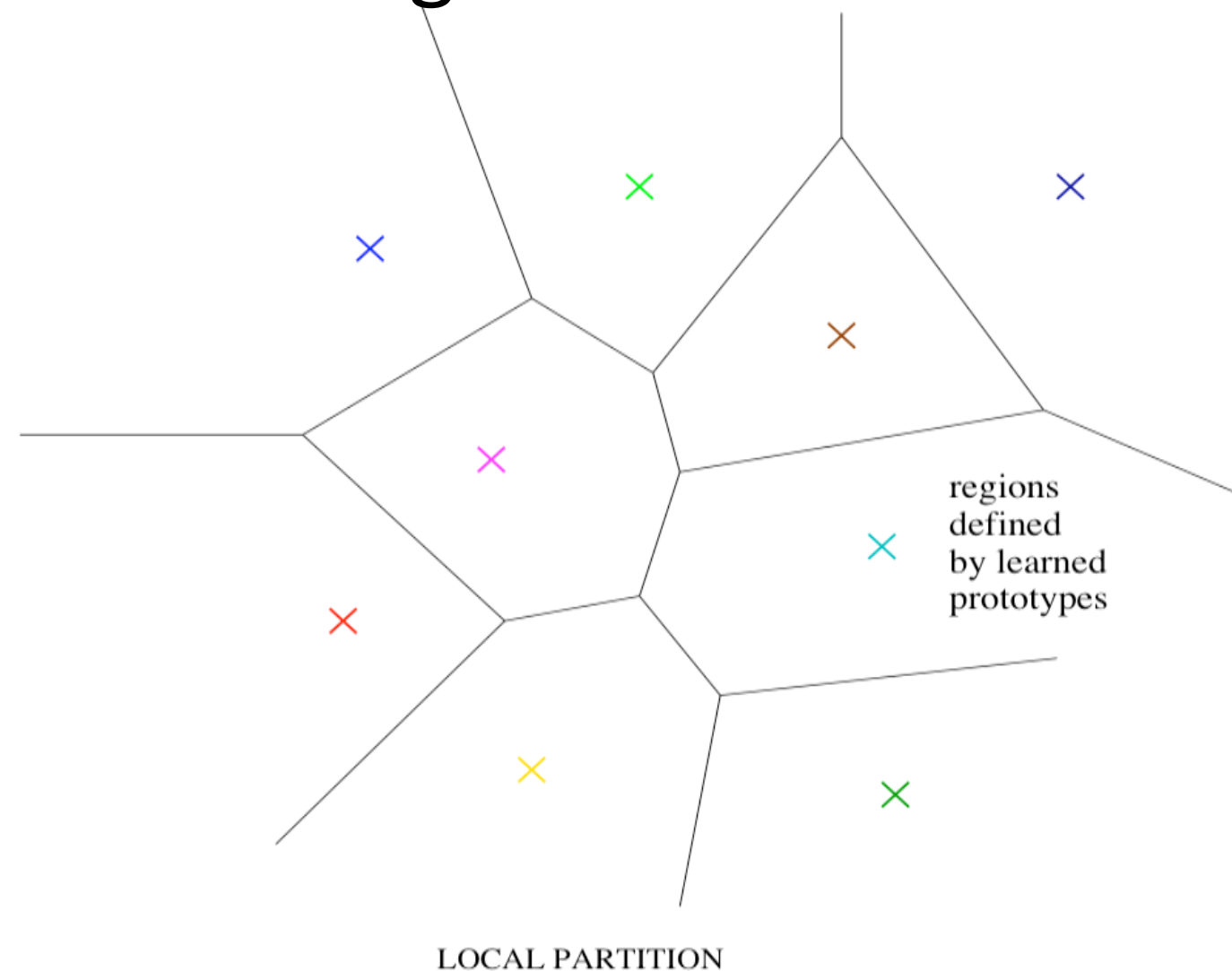
ML 101. What We Are Fighting Against: The Curse of Dimensionality

- How can we possibly generalize to settings we have never seen data for?
- Looking at similar examples is not sufficient when dealing with high-dimensional data, like images

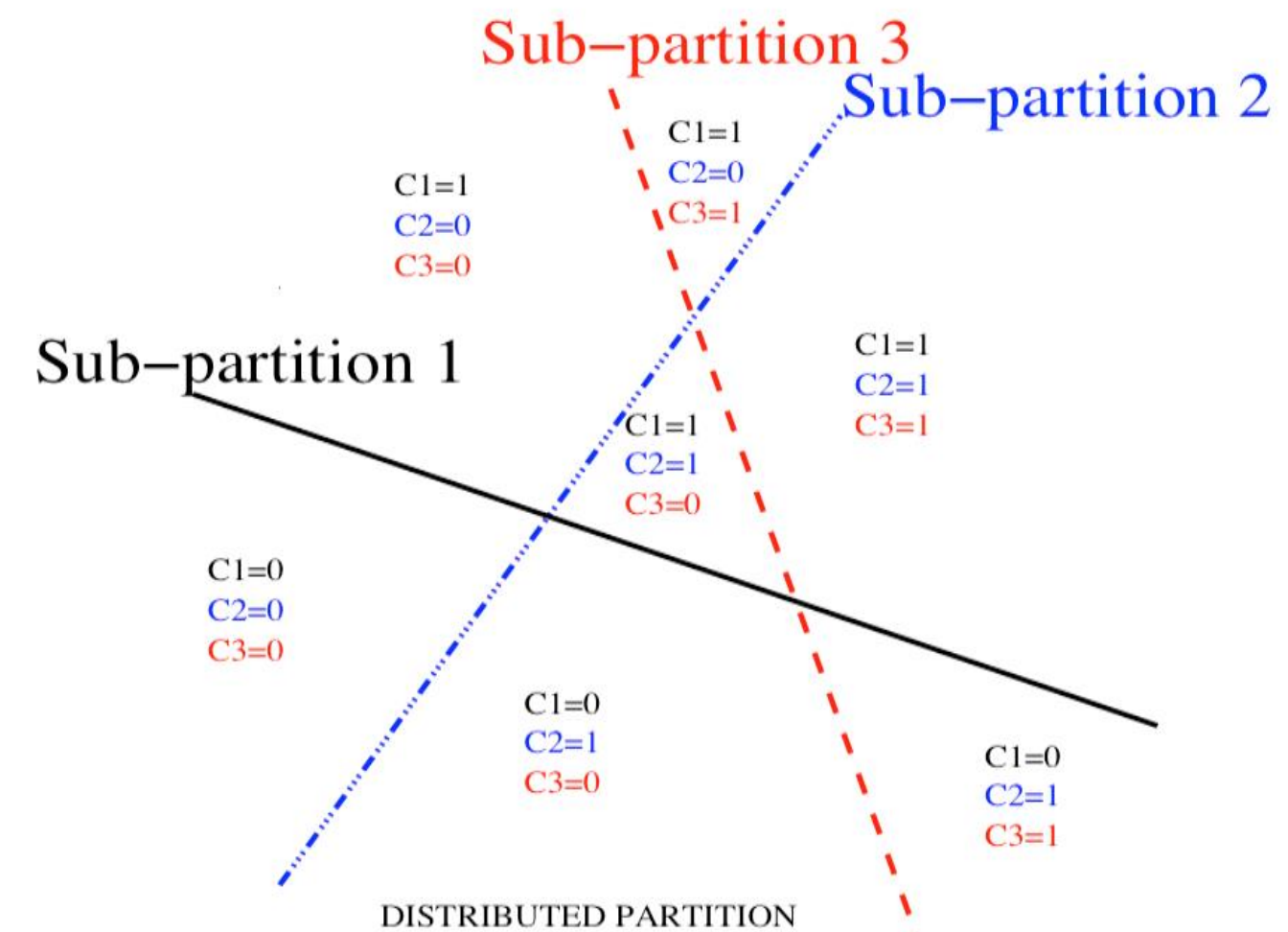


Distributed Representations: The Power of Compositionality

- Distributed (possibly sparse) representations, learned from data, can capture the **meaning** of the data and state
- Parallel composition of features: can be exponentially advantageous



Not Distributed



Distributed

Machine Learning & AI → Deep Learning

- Five key ingredients for ML towards AI
 1. Lots & lots of data
 2. Very flexible models
 3. Enough computing power
 4. Computationally efficient inference
 5. **Powerful assumptions that can defeat the curse of dimensionality and achieve strong generalization to new cases**

The Neural Net Approach to AI

- **Brain-inspired**
- Synergy of a large number of simple adaptive computational units
- Focus on **distributed representations**
 - **E.g. word representations** (Bengio et al NIPS'2000)
- View intelligence as arising of combining
 - an objective or reward function
 - an approximate optimizer (learning rule)
 - an initial architecture / parametrization
- End-to-end learning (all the pieces of the puzzle adapt to help each other)



Recap: Machine Learning 101

- Family of functions
- Tunable parameters
- Examples sampled from unknown data generating distribution
- A measure of the error made by the trained function
- Approximate minimization algorithm to search for good choice of parameters, iteratively reducing the average training error

Learning Long-Term Dependencies with Gradient Descent is Difficult

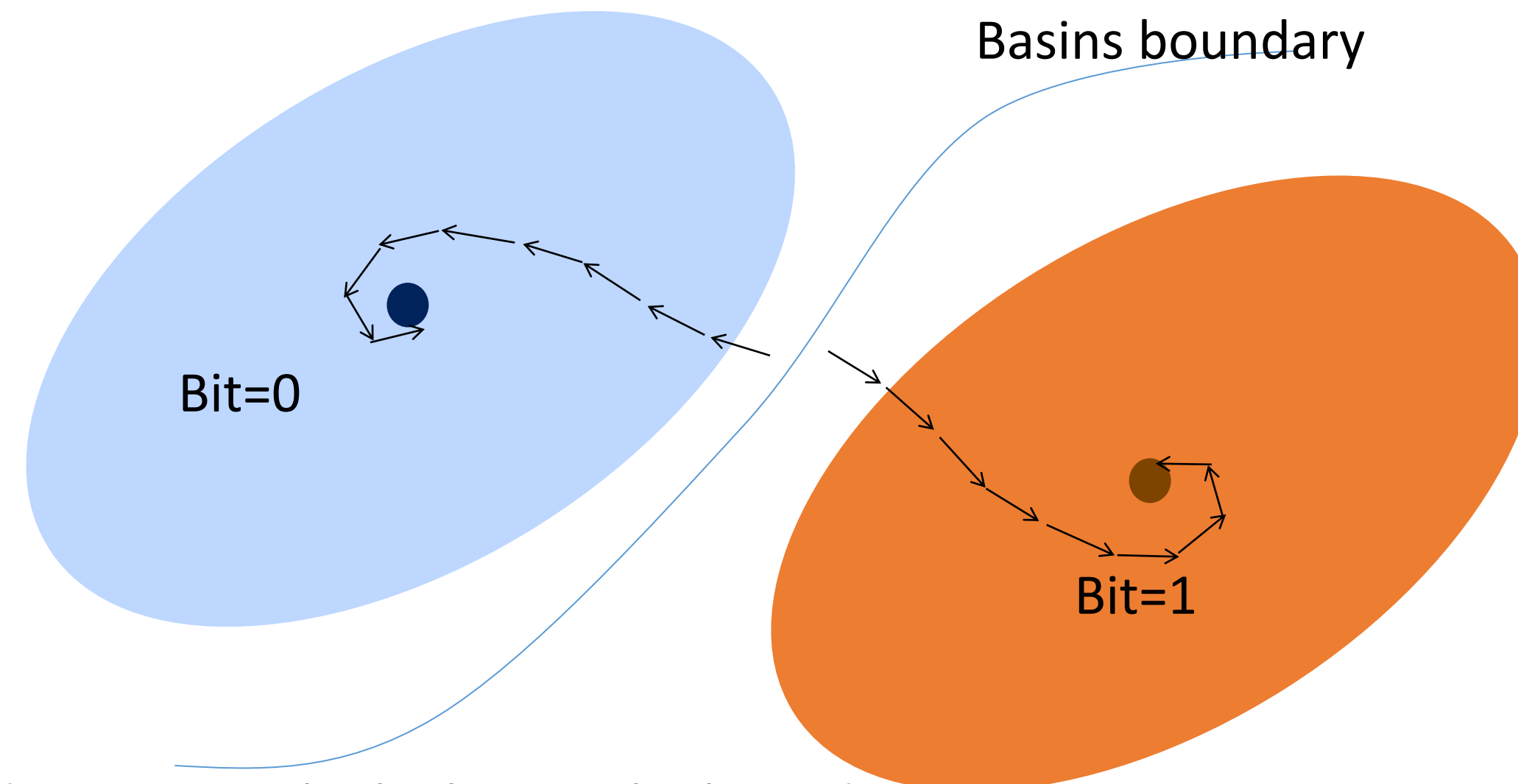


Y. Bengio, P. Simard & P. Frasconi, IEEE Trans. Neural Nets, **1994**

Lesson: negative results can be very important, teach us something that drives much downstream research → **2014 work on self-attention**

How to store 1 bit? Dynamics with multiple basins of attraction in some dimensions

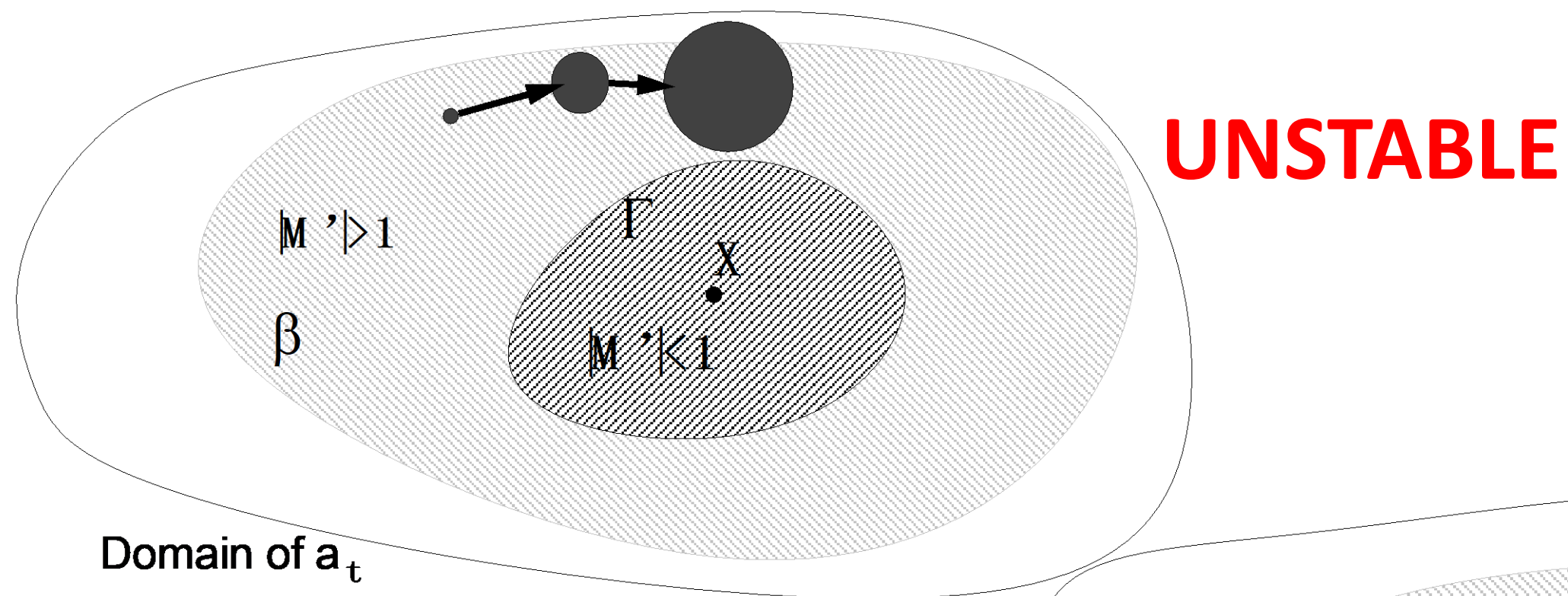
- Some subspace of the state can store 1 or more bits of information if the dynamical system has multiple basins of attraction in some dimensions



Note: gradients MUST be high near the boundary

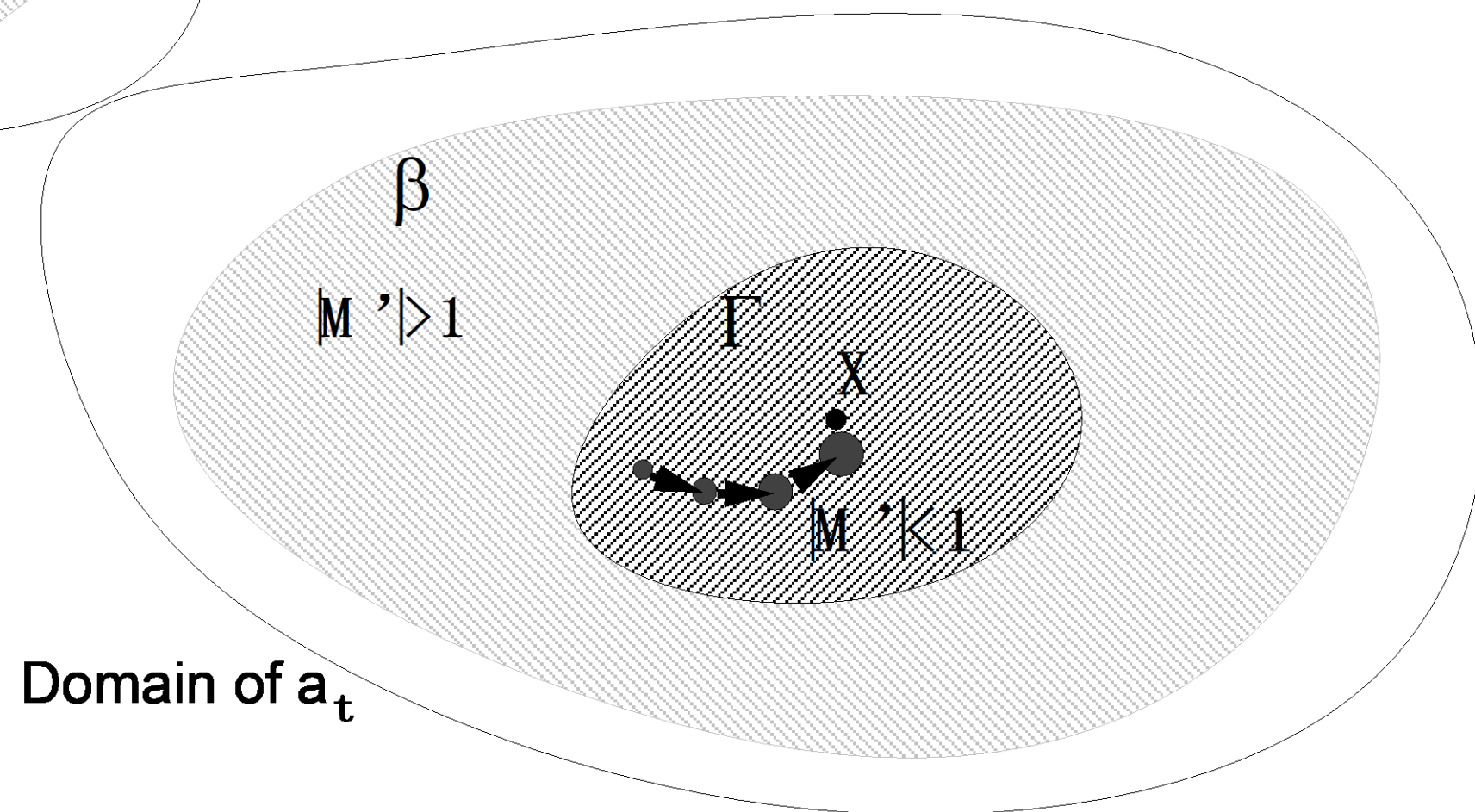
Robustly storing 1 bit in the presence of bounded noise

- With spectral radius > 1 , noise can kick state out of attractor



- Not so with radius < 1

CONTRACTIVE
→ STABLE



Storing Reliably \rightarrow Vanishing gradients

- Reliably storing bits of information requires spectral radius < 1
- The product of T matrices whose spectral radius is < 1 is a matrix whose spectral radius converges to 0 at exponential rate in T

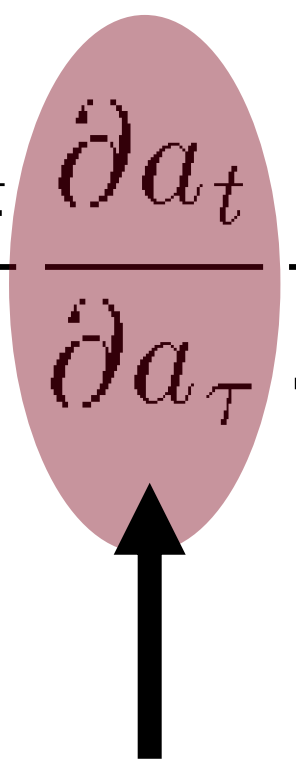
$$L = L(s_T(s_{T-1}(\dots s_{t+1}(s_t, \dots))))$$

$$\frac{\partial L}{\partial s_t} = \frac{\partial L}{\partial s_T} \frac{\partial s_T}{\partial s_{T-1}} \dots \frac{\partial s_{t+1}}{\partial s_t}$$

- If spectral radius of Jacobian is $< 1 \rightarrow$ propagated gradients vanish

Why it hurts gradient-based learning

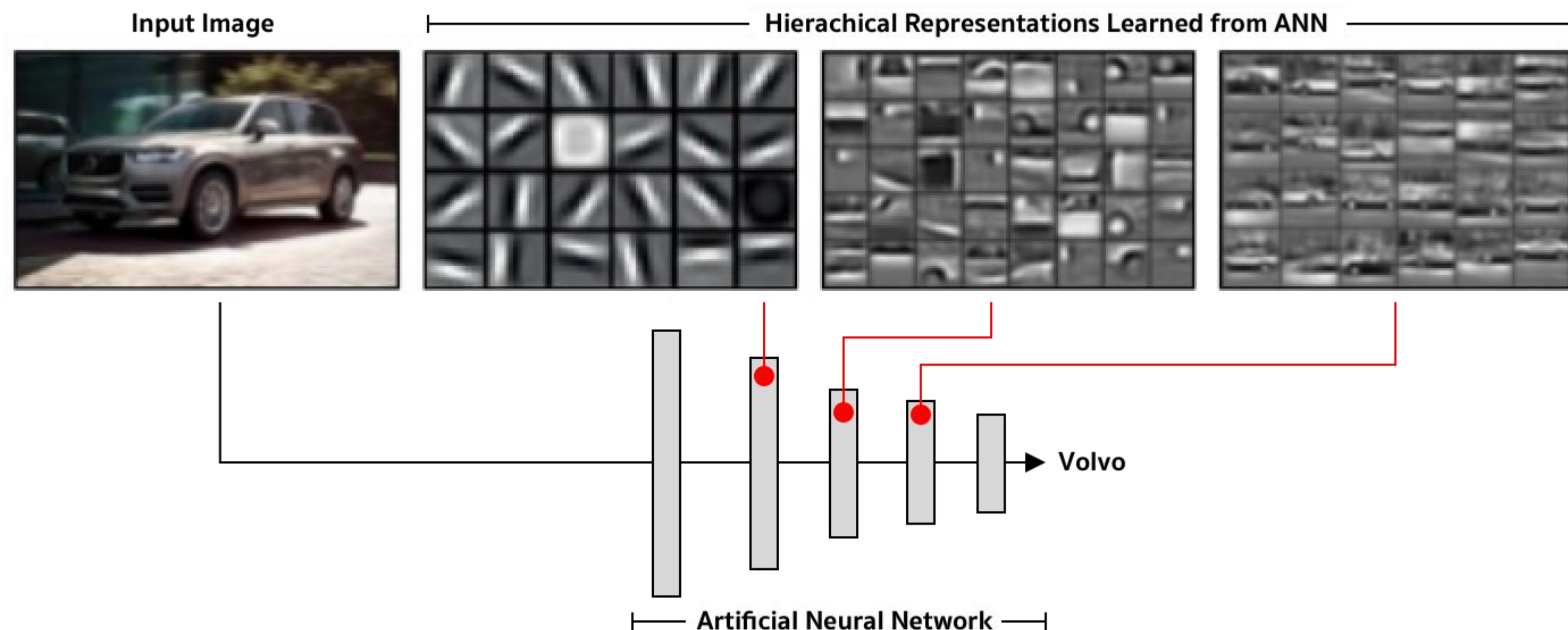
- Long-term dependencies get a weight that is exponentially smaller (in T) compared to short-term dependencies

$$\frac{\partial C_t}{\partial W} = \sum_{\tau \leq t} \frac{\partial C_t}{\partial a_\tau} \frac{\partial a_\tau}{\partial W} = \sum_{\tau \leq t} \frac{\partial C_t}{\partial a_t} \frac{\partial a_t}{\partial a_\tau} \frac{\partial a_\tau}{\partial W}$$


Becomes exponentially smaller for longer time differences, when spectral radius < 1

Deep Learning: Learning an Internal Representation

- Unlike other ML methods with either
 - no intermediate representation (linear)
 - or fixed (generally very high-dimensional) intermediate representations (SVMs, kernel machines)
- What is a good representation? Makes other or downstream tasks easier.



nature

Explore content ▾ About the journal ▾ Publish

[nature](#) > [review articles](#) > [article](#)

Published: 27 May 2015

Deep learning

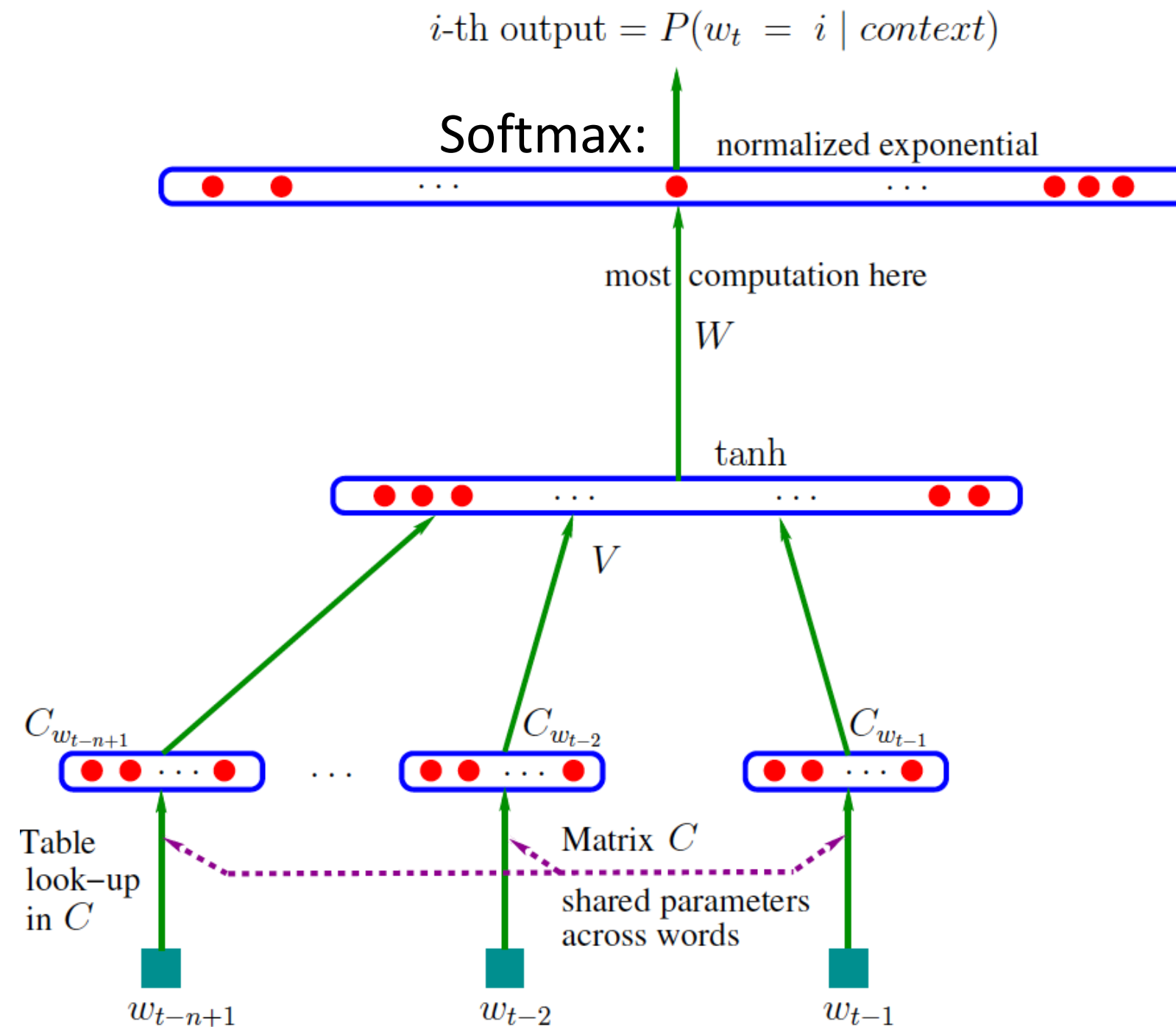
[Yann LeCun](#) , [Yoshua Bengio](#) & [Geoffrey Hinton](#)

Nature **521**, 436–444 (2015) | [Cite this article](#)

738k Accesses | 34119 Citations | 1179 Altmetric

Neural Language Models

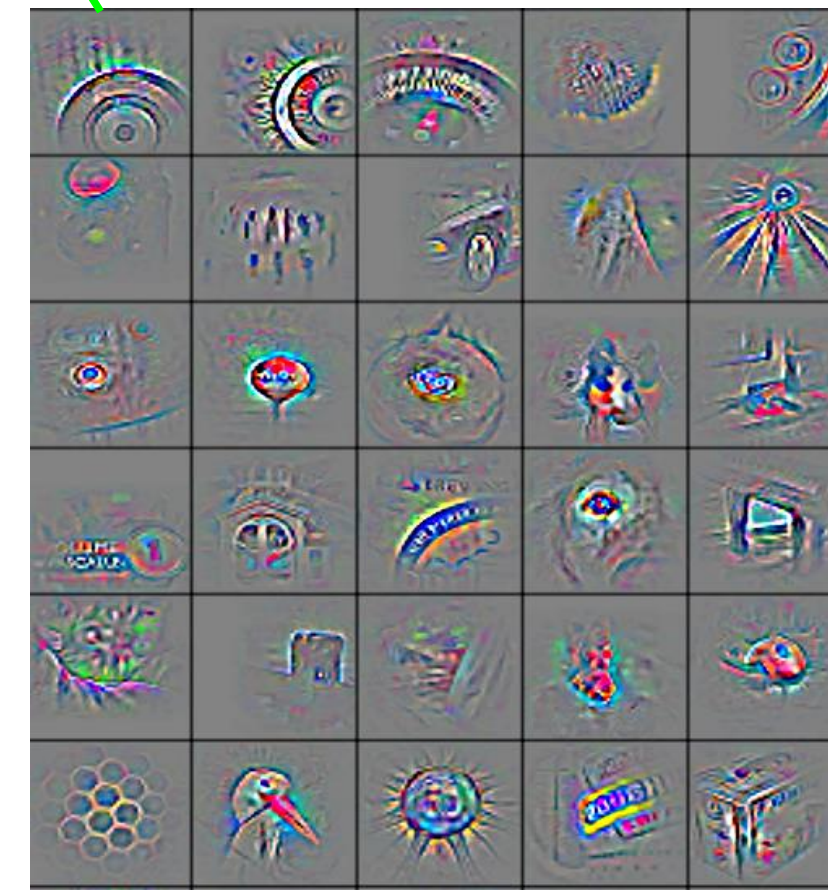
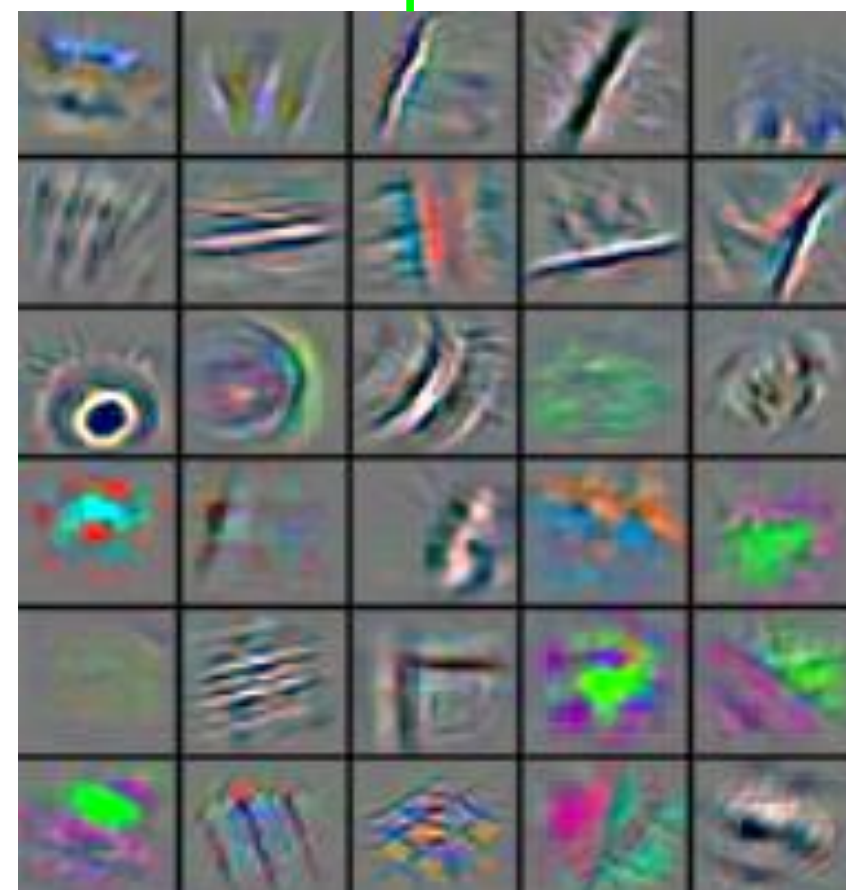
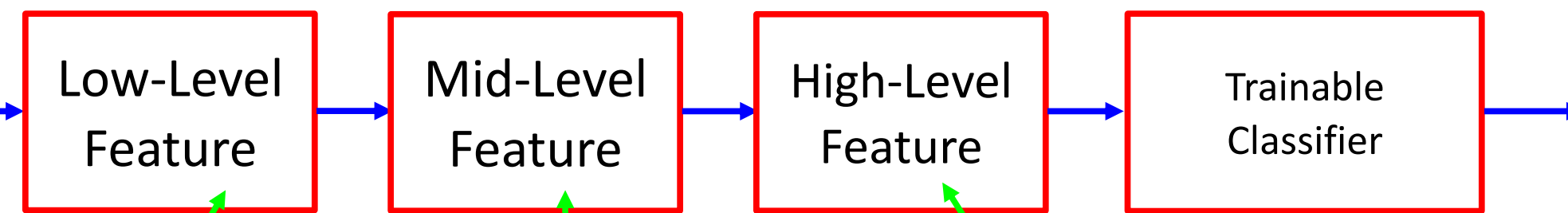
- *Bengio et al NIPS'2000 and JMLR 2003 "A Neural Probabilistic Language Model"*
 - Each word represented by a distributed continuous-valued code vector = **embedding**
 - Shared across n-grams (tuples of words)
 - Generalizes to sequences of words that are **semantically similar** to training sequences



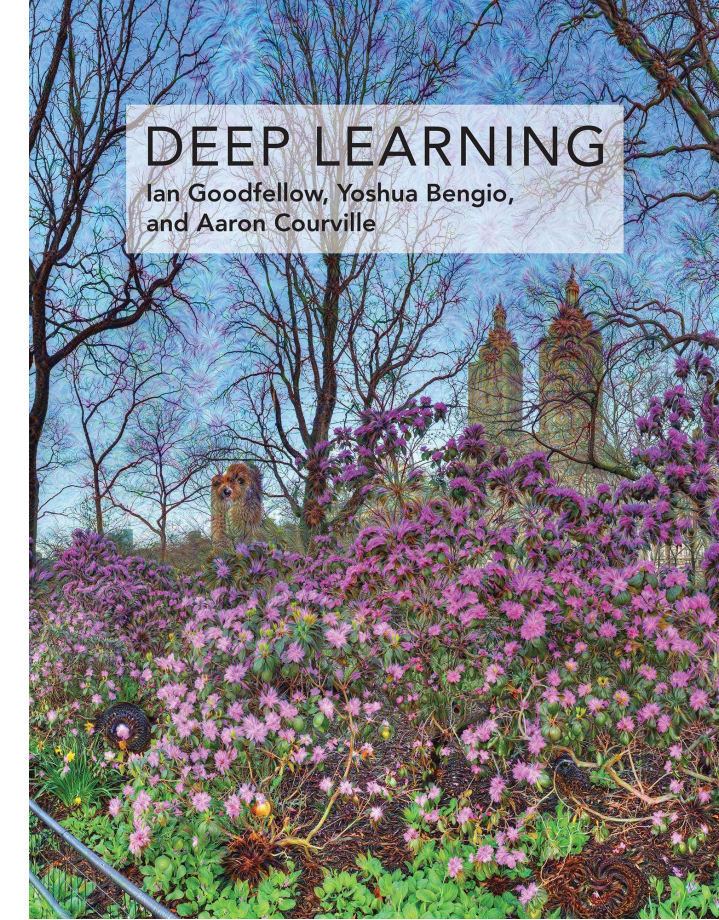
$$P(w_1, w_2, w_3, \dots, w_T) = \prod_t P(w_t | w_{t-1}, w_{t-2}, \dots, w_1)$$

Why Multiple Layers? The World is Compositional

- Hierarchy of representations with increasing level of abstraction
- Each stage is a kind of trainable feature transform
- **Image recognition**: Pixel \rightarrow edge \rightarrow texton \rightarrow motif \rightarrow part \rightarrow object
- **Text**: Character \rightarrow word \rightarrow word group \rightarrow clause \rightarrow sentence \rightarrow story
- **Speech**: Sample \rightarrow spectral band \rightarrow sound \rightarrow ... \rightarrow phone \rightarrow phoneme \rightarrow word

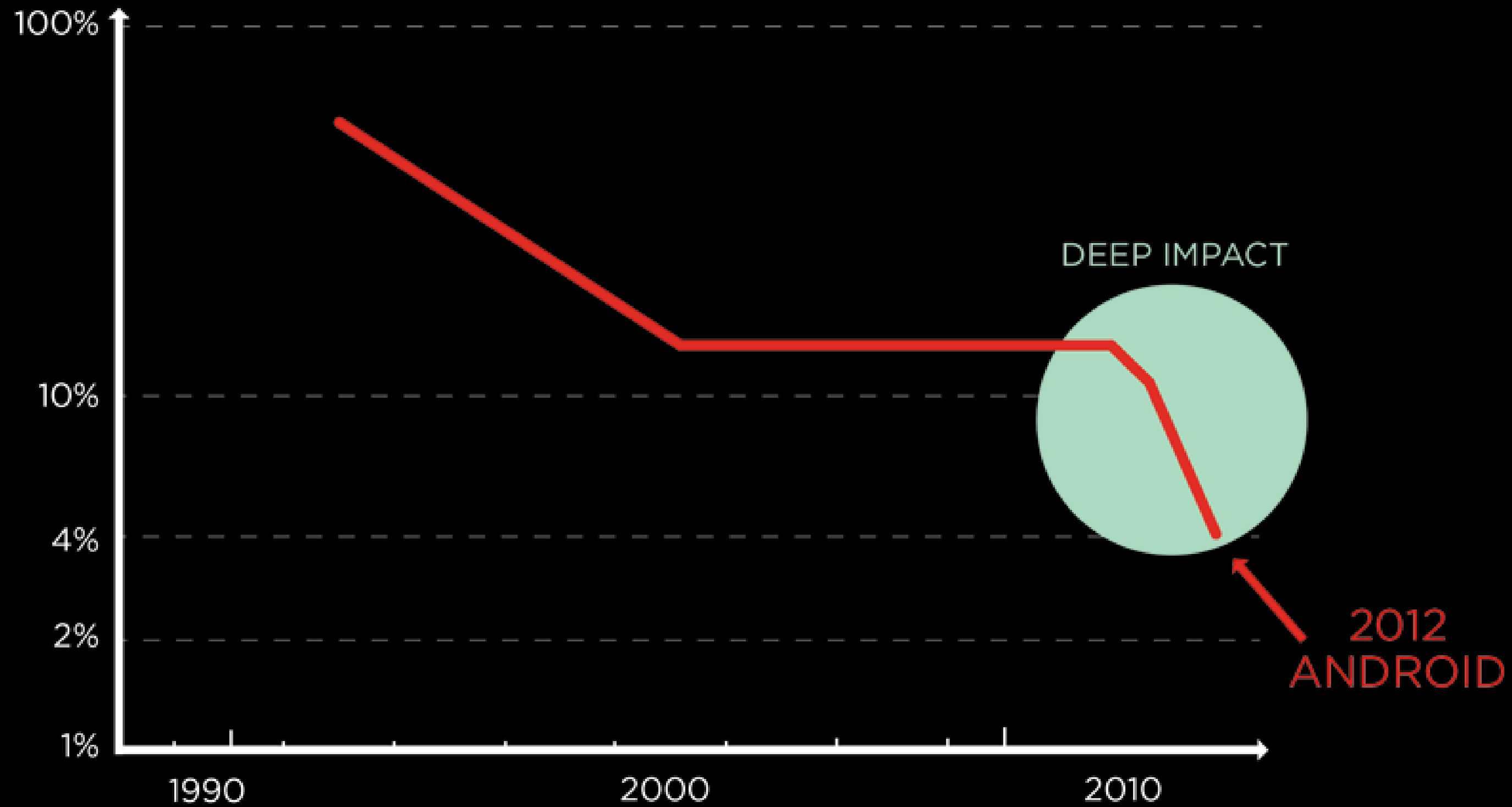


Credits: Yann LeCun

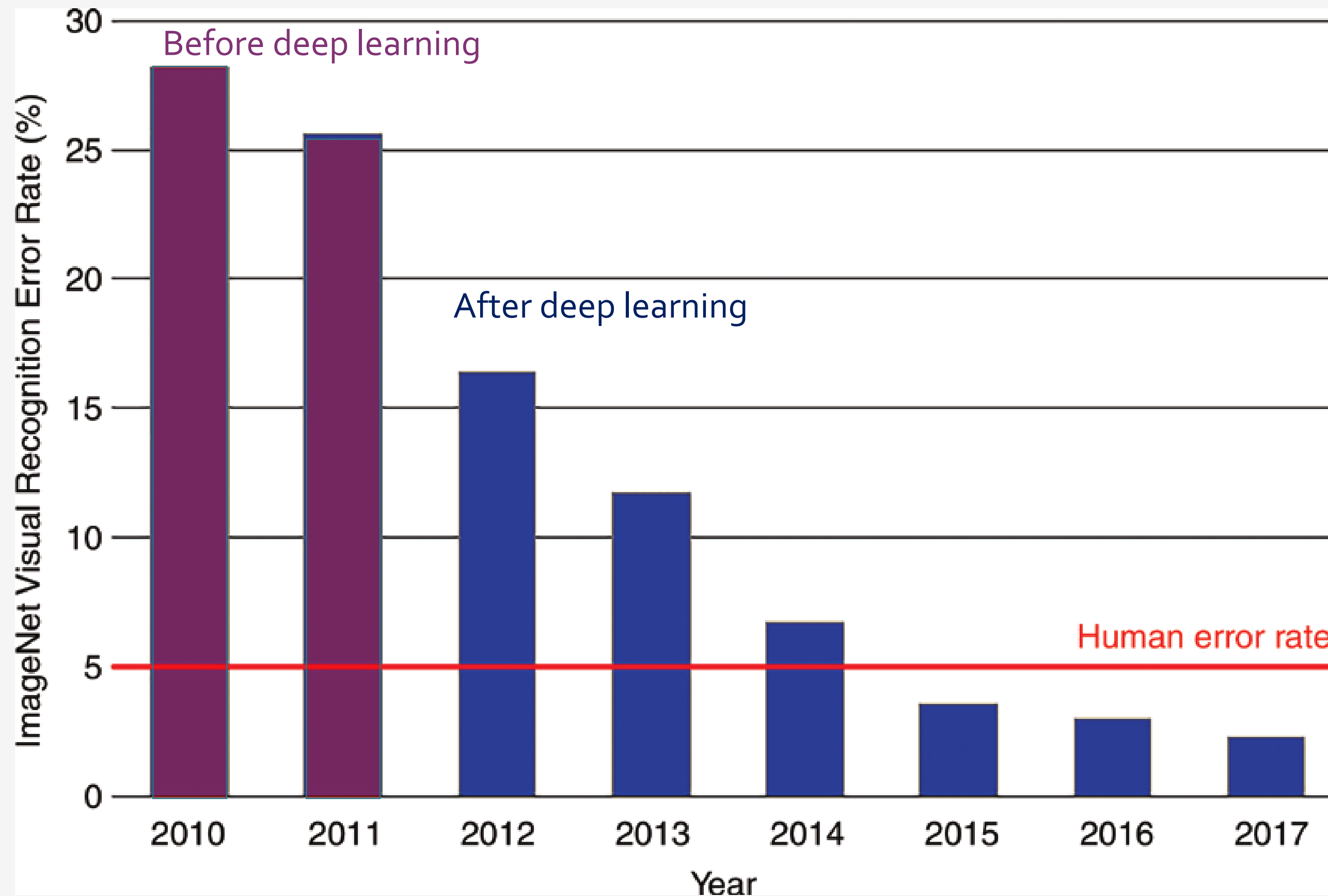


Deep Learning,
MIT Press, 2016

2010-2012: breakthrough in speech recognition



ImageNet Breakthrough in 2012



Persist and Pull the Threads

1996-2012 neural net winter:

- AI research loses its ambition to reach human-level intelligence
- focus on "simpler" (easier to analyze) ML
- difficult to convince grad students to work on neural nets

Persist, but ask the hard questions:

- Following our intuitions
- But try to validate experimentally or mathematically
- Pull the threads to clarify them, ask WHY questions, try to UNDERSTAND
- Importance of support group (CIFAR program)

Generative Adversarial Networks

Goodfellow et al & Bengio NIPS 2014



2014



2015



2016



2017



2018



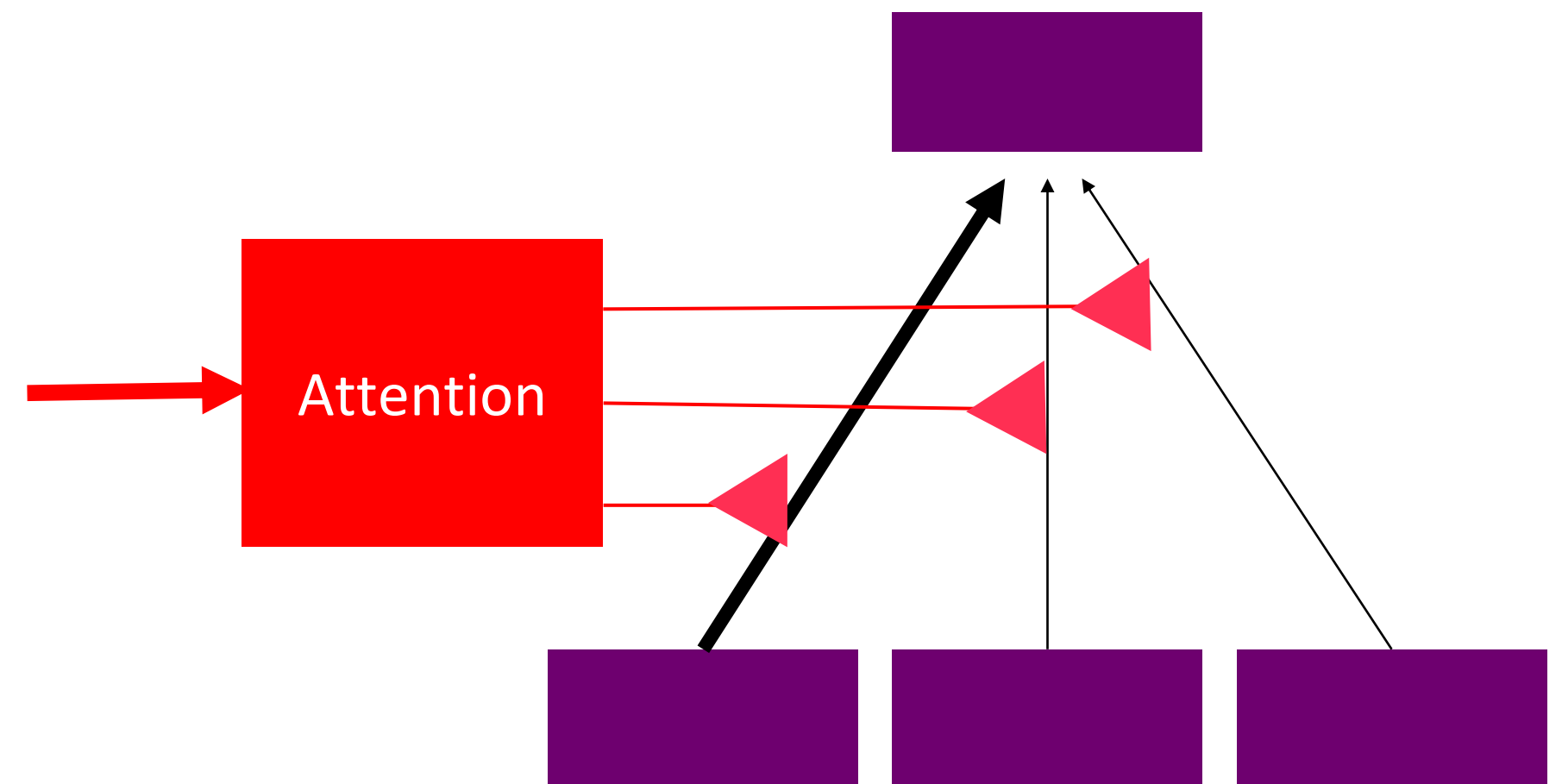
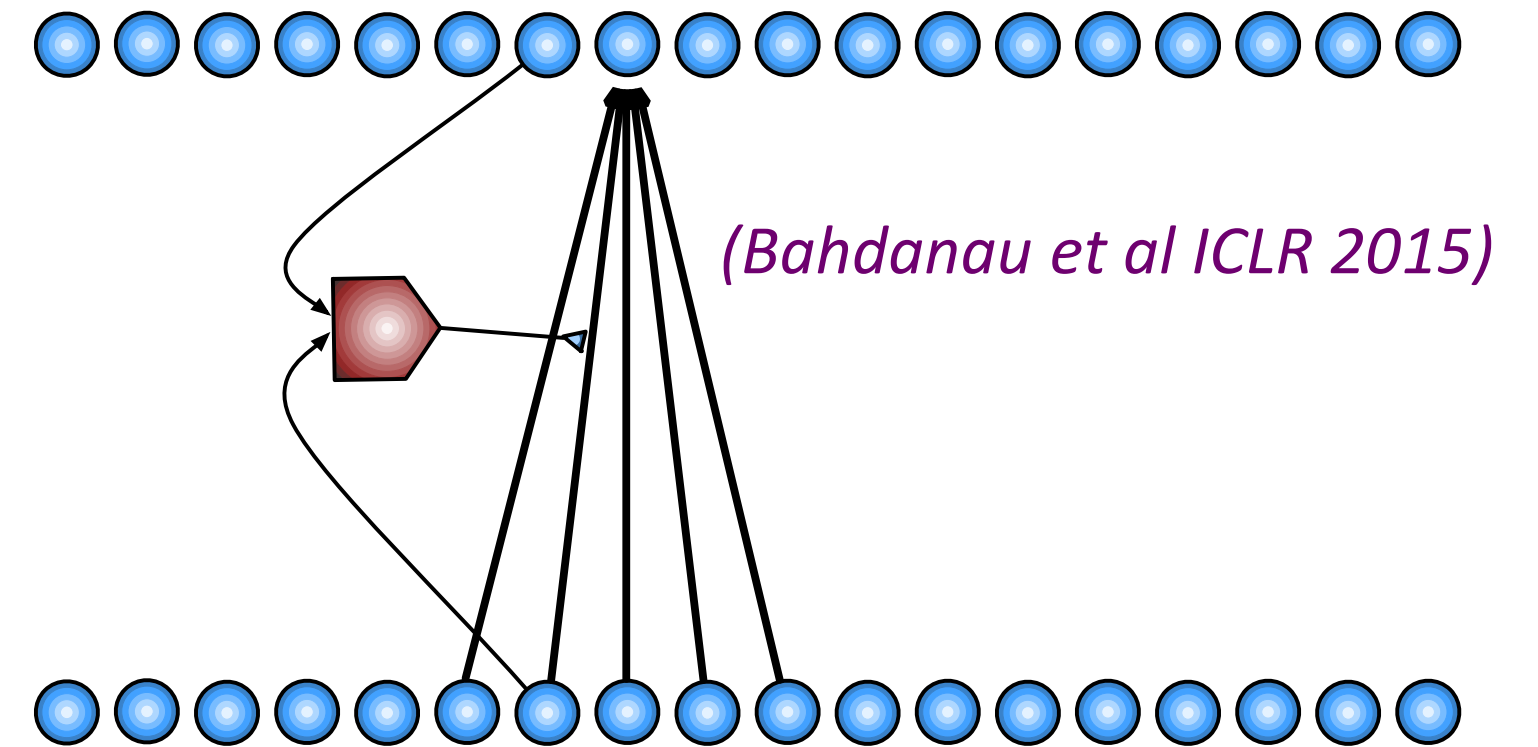
this bird is red with white and has a very short beak



Xu et al 2018, AttnGAN

THE ATTENTION REVOLUTION

- **Focus** on a one or a few elements at a time
- *Learn where to attend, depending on context*
- Operating on unordered SETS
- SOTA in NLP, gave rise to Transformers
- **BYPASSES THE ISSUE OF LEARNING LONG-TERM DEPENDENCIES IN RNNs!!!**



Deep Reinforcement Learning Breakthrough (2016)

- AlphaGo beat world champion Lee Sedol 4-1
- Not expected by AI and Go experts
- Combines deep learning with reinforcement learning



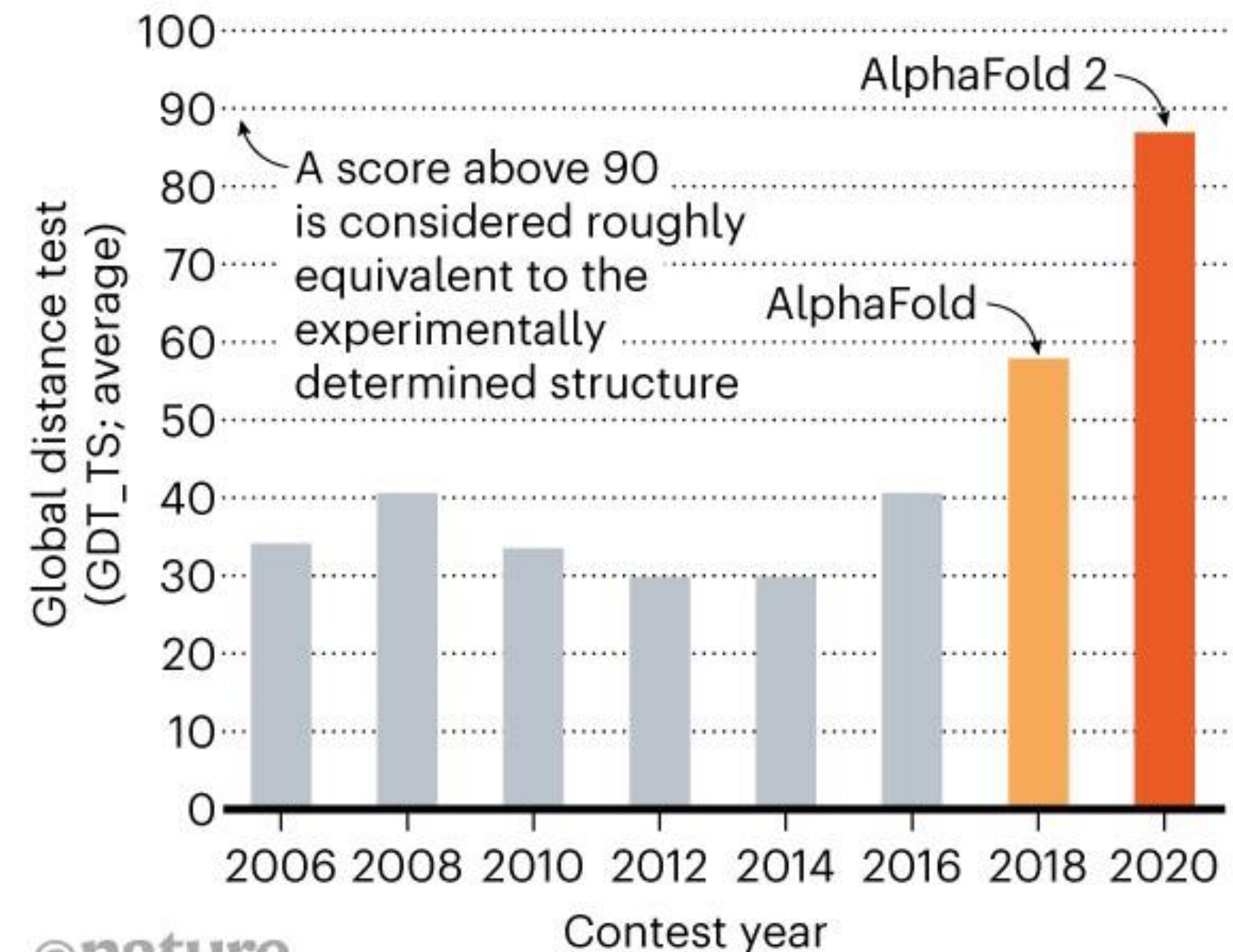
Biological Breakthrough: Protein Structure from Sequence with AlphaFold 2 (2021)



Using attention graph neural networks

STRUCTURE SOLVER

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



Staying Humble

- Better not even think about awards, prizes, recognition: dangerous distractions!
- Ego can blind us, make us overconfident, is the enemy of scientific discoveries
- Hurts our ability to be flexible in our ideas, question what we took for granted, listen to others who disagree with us
- I changed my mind several times: supervised vs unsupervised in 2005, frequentist vs Bayesian in 2022.



2018 Turing Award, announced in the Spring of 2019

Learning Higher Levels of Abstraction

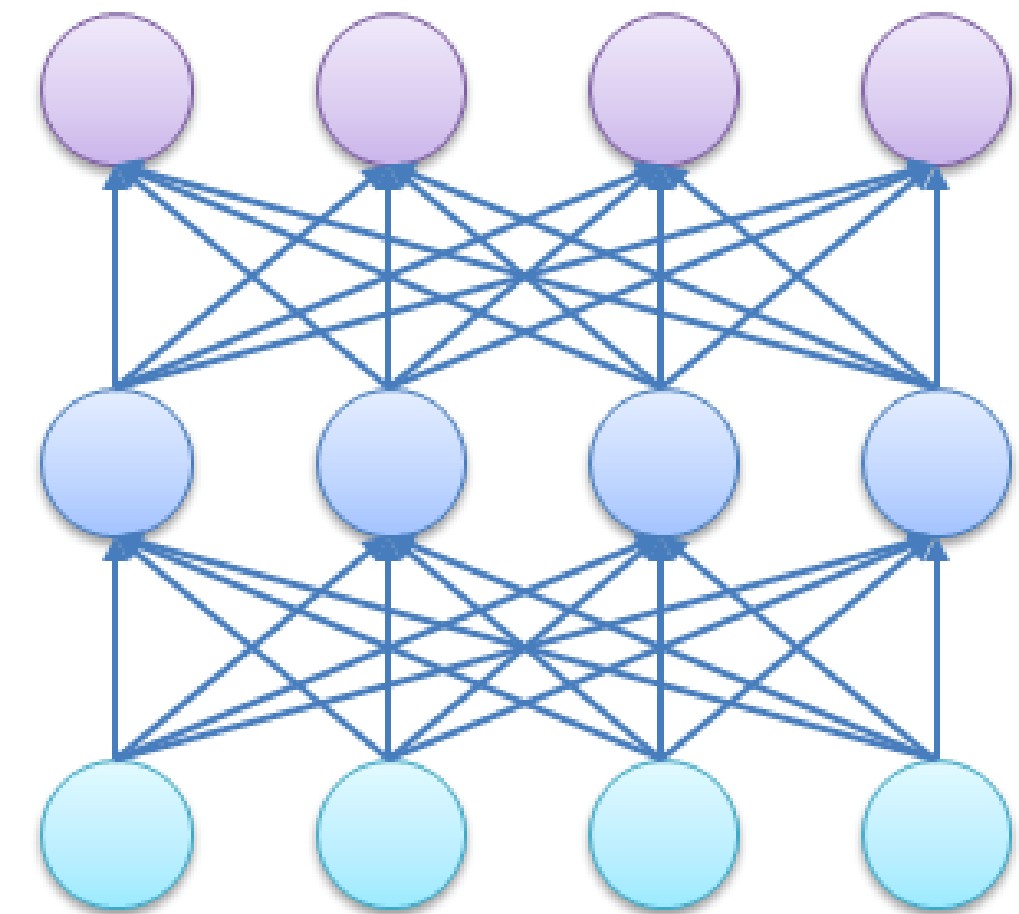
(Bengio & LeCun 2007)

- The big payoff of deep learning is to allow learning higher levels of abstraction
- Higher-level abstractions **disentangle the explanatory variables and their causal mechanisms**, which would allow much easier generalization and transfer to new tasks



How to Discover Good Disentangled Representations

- How to discover abstractions?
- What is a good representation? (*Bengio et al 2013*)
- Need clues (= inductive biases) to help **disentangle** underlying factors **and their dependencies**, e.g.
 - Spatial & temporal scales
 - Simple & sparse dependencies between factors
 - *Consciousness prior*
 - Causal / mechanism independence
 - *Controllable variables = interventions*
 - Multiple spatial and temporal scales
 - *Coarse high-level factors explain lower-level details*



Bypassing the curse of dimensionality

We need to build **compositionality** into our ML models

Just as human languages exploit compositionality to give representations and meanings to complex ideas

Exploiting compositionality: **exponential** gain in representational power

Distributed representations / embeddings: **feature learning**

Current deep architectures: **multiple levels of feature learning**

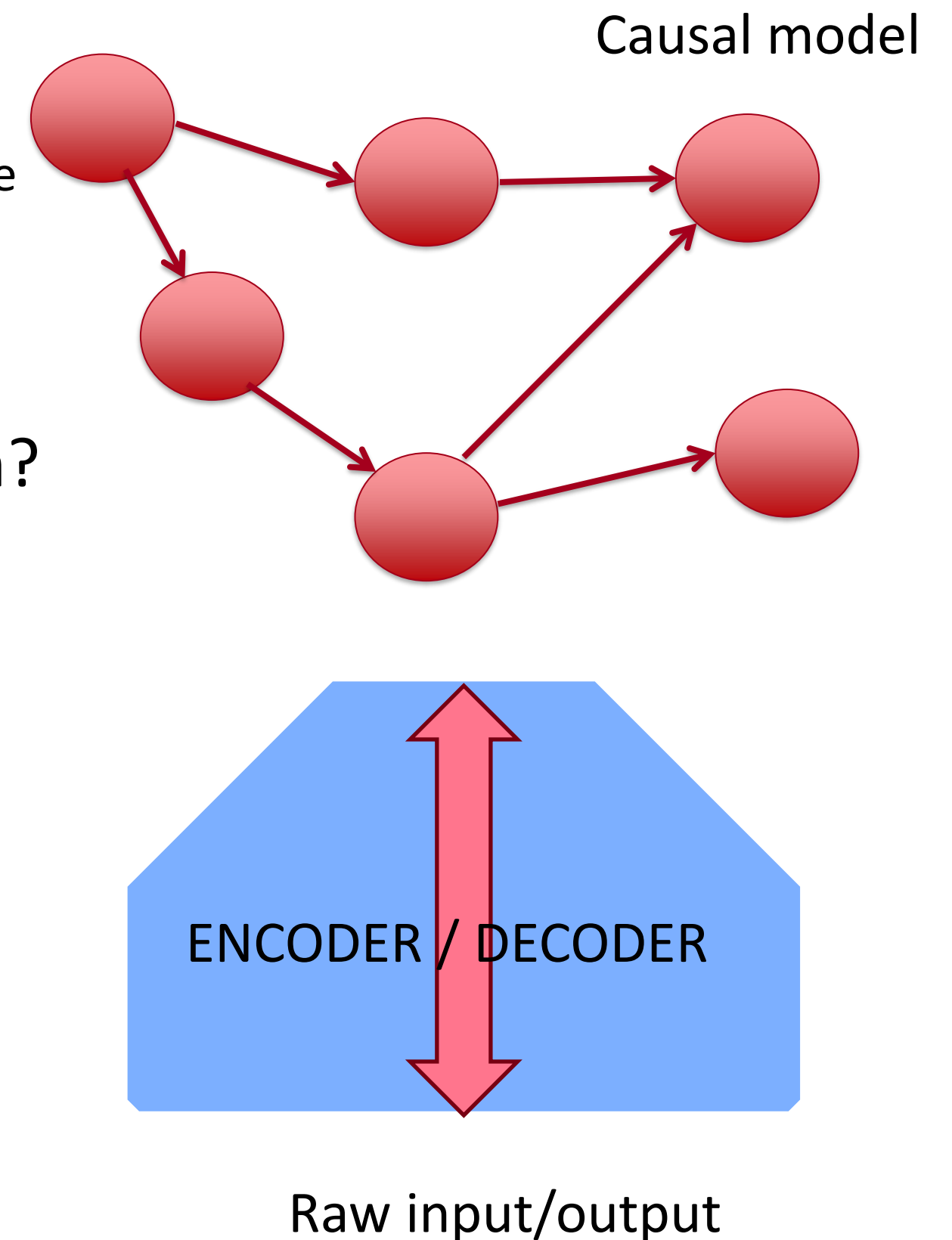
System 2 deep learning: **compose a few concepts at a time**

Prior assumption:

compositionality useful to describe the world around us efficiently

Deep Learning Objective: discover causal representation

- What are the right representations? Causal variables explaining the data
- How to discover them (as a function of observed data)?
- How to discover their causal relationship, the causal graph?
- How are actions corresponding to causal interventions?
- How is raw sensory data related to high-level causal variables and how do high-level causal variables turn into low-level actions and partial observations?
- **Need additional biases: causality is about changes in distribution**



Missing from Current ML: Understanding & Generalization Beyond the Training Distribution

- Learning theory only deals with generalization within the same distribution
- Models learn but do not generalize well (or have high sample complexity when adapting) to modified distributions, non-stationarities, etc.
- Poor reuse, poor modularization of knowledge

Generalization Beyond the Training **Distribution**

- **Current industrial-strength ML suffers from robustness issues due to poor performance OOD**
- If not **iid**, need alternative assumptions, otherwise no reason to expect generalization
 - How do distributions change?
 - Humans do a lot better!
 - Inductive biases inspired from brains?
 - *How do humans re-use knowledge?*



SYSTEMATIC GENERALIZATION

- Studied in linguistics
- **Dynamically recombine existing concepts**
- Even when new combinations have 0 probability under training distribution

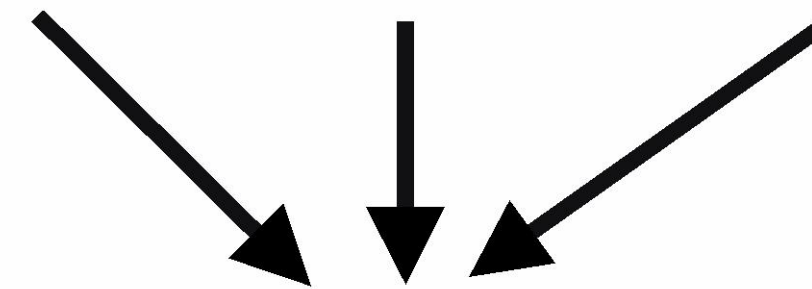
- E.g. Science fiction scenarios
- E.g. Driving in an unknown city

- Not very successful with current DL, which can "overfit" the training **distribution**

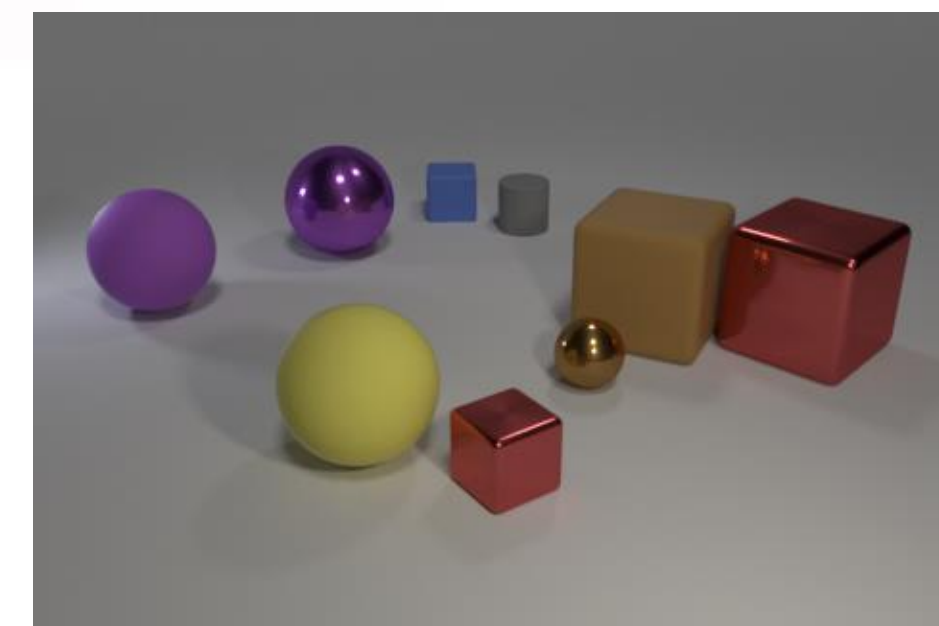
(Lake & Baroni 2017)

(Bahdanau et al & Courville ICLR 2019)

CLOSURE: (Bahdanau et al & Courville arXiv:1912.05783) on CLEVR



(Lake et al 2015)



THE GAP FROM SOTA AI & HUMAN-LEVEL INTELLIGENCE

- Sample complexity: number of examples needed to learn a task
- Out-of-distribution generalization
- Out-of-distribution speed of adaptation (transfer learning)
- Causal discovery and reasoning
- Compositional knowledge representation & inference



UNIQUE CAUSE FOR THE GAP: CONSCIOUS PROCESSING?

Hypothesis:

This gap originates from a type of computation, knowledge representation and inference associated with **conscious processing in humans**, not yet mastered in AI



CONSCIOUS PROCESSING HELPS HUMANS DEAL WITH OOD SETTINGS

Faced with novel or rare situations, humans call upon conscious attention to combine on-the-fly the appropriate pieces of knowledge, to reason with them and imagine solutions.

→ we do not follow our habitual routines, use conscious thinking in novel settings.

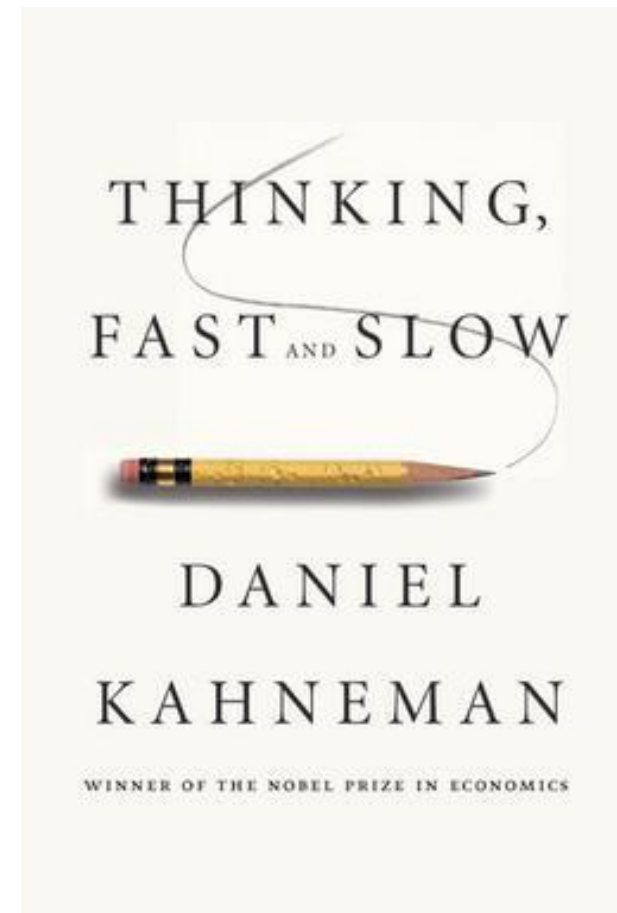


SYSTEM 1 VS. SYSTEM 2 COGNITION

2 systems (and categories of cognitive tasks):

System 1

- Intuitive, fast, **UNCONSCIOUS**, 1-step parallel, non-linguistic, habitual
- Implicit knowledge
- Current DL



System 2

- Slow, logical, **sequential**, **CONSCIOUS**, linguistic, algorithmic, planning, **reasoning**
- Explicit knowledge
- DL 2.0

Manipulates high-level / semantic concepts, which can be recombined combinatorially



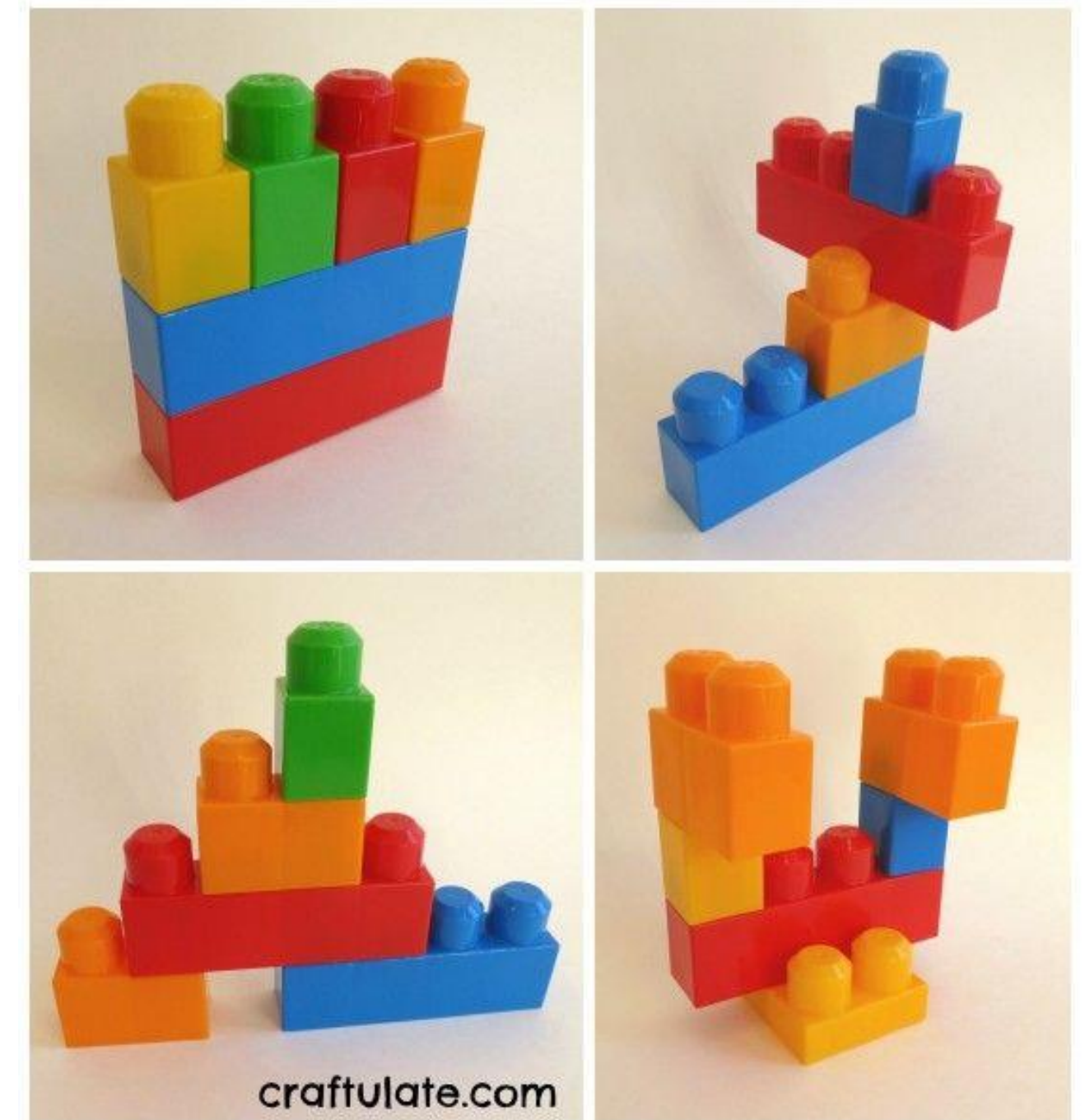
FROM REASONING TO OOD GENERALIZATION?

- **Current industrial-strength ML (including in NLP) suffers from robustness issues due to poor performance OOD**
- Humans use higher-level cognition (system 2) for out-of-distribution generalization
- Why and how does it help?
- How is that related with agency? causality?
- How do we incorporate these principles in deep learning to obtain both system 1 and system 2 deep learning?



FACTORIZING KNOWLEDGE INTO COMPOSABLE PIECES FOR REASONING

- **Current deep learning:** homogenous architectures, **knowledge is not localized**, completely distributed
- **Transfer learning:** reuse just the relevant pieces of knowledge; minimizes interference; maximizes reuse
- **System 2 reasoning selects and combines nameable pieces of knowledge to form thoughts** (imagined futures, counterfactual past, solutions to problems, interpretations of inputs, etc).
- How to **factorize knowledge** into the right recomposable pieces?



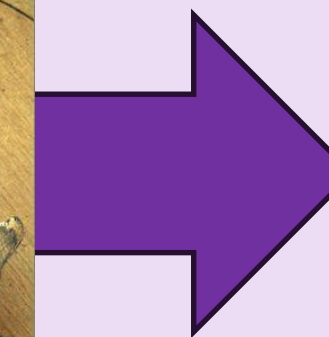
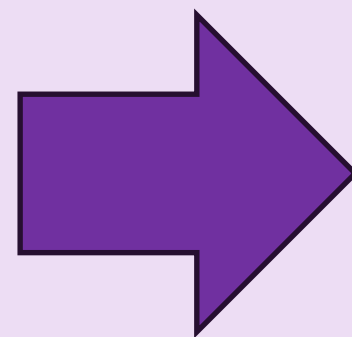
Transfer to modified distribution: Beyond the iid assumption



- iid assumption too strong \rightarrow poor out-of-distribution generalization
- relaxed assumptions: **same causal dynamics**, different state/intervention

Stochastic dynamical system

Initial
conditions



Observed
data

Causality as a framework for OOD generalization, transfer learning, continual learning, etc

- Factor **stationary knowledge** (causal mechanisms) from **non-stationary knowledge** (the values of variables)
- **Intervention** = change in variables that is not just due to the default cause-effect links, but due to an agent
- **Causal model = family of distributions** (which includes tasks)
- The index over these distributions is the choice of interventions (or initial state)
- Stationary knowledge is factorized into **recomposable causal mechanisms**

Why Causal?

Correlation
≠
Causation

- Causal model = family of distributions indexed by interventions / environments / initial states etc **with shared parameters (mechanisms)**
- Learner must predict the effect of interventions, needs to generalize **Out-Of-Distribution = to new interventions**
- Intervention = perfectly realized **abstract actions** of an agent
 - more realistic: intentions to achieve a change in abstract variable = **goal**
- Unlike multi-task and meta- learning, instead of learning task- or environment specific parameters, perform inference on interventions

MAIN SYSTEM 2 INDUCTIVE BIASES

Modularity (millions of "experts" in cortex), reusable modules

- dynamic selection of composition of which modules are relevant at any time

Working memory bottleneck (handful of items at a time)

- Past content of WM is available to all expert modules
- Competition between modules and cooperation to select (discrete?) content

Sparsity of high-level dependencies (few variables at a time interact)

- same variable may be involved in several dependencies, multiple modules
- dependencies (energy fn) operate conditionally on correctly typed args

Sequential selection of "thoughts", each involving just 1 or few modules

Causal semantics of variables, localized intervention=high-level intention

RECENT EVIDENCE OF OOD GAINS WITH SYSTEM 2 INDUCTIVE BIASES

- **Recurrent Independent Mechanisms:** *Goyal et al 2019, arXiv:1909.10893, ICLR 2021*
- **Coordination Among Neural Modules Through a Shared Global Workspace:** *Goyal et al 2021, arXiv:2103.01197, ICLR 2022*

- **Neural Production Systems:**

Goyal et al 2021, arXiv:2103.0193, NeurIPS 2021

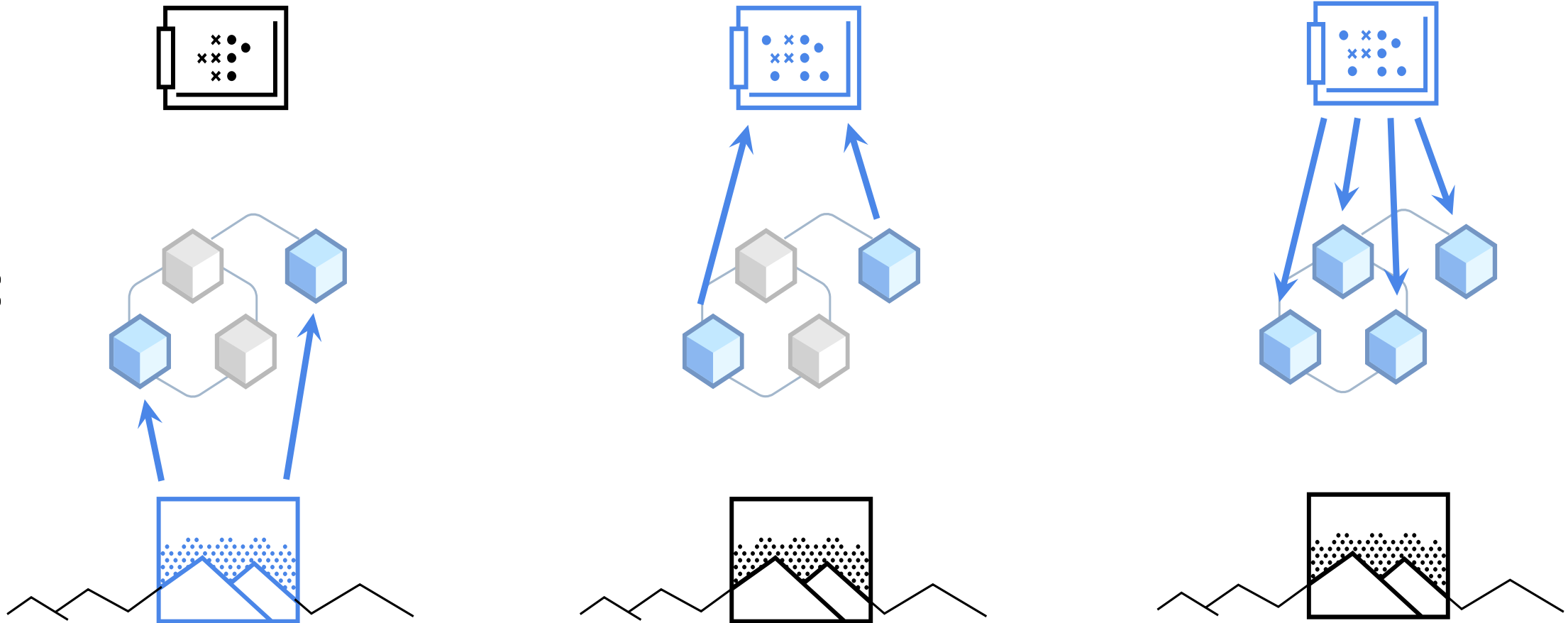
- **Discrete-Valued Neural Communication:**

Liu et al 2021, arXiv:2107.02367, NeurIPS 2021

1. Parallel, competing specialists

2. Write to shared workspace

3. Broadcast workspace contents



Large deep nets as probabilistic inference and probabilistic reasoning machines:

BAYESIAN STRUCTURE LEARNING WITH GENERATIVE FLOW NETWORKS

UAI'2022, arXiv:2202.13903

Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, Yoshua Bengio



GFLOWNET TUTORIAL

<https://tinyurl.com/gflownet-tutorial>

**Papers at NeurIPS 2021, ICML 2022, UAI 2022, NeurIPS 2022,
+ArXiv**

ML going out of labs, into society

- ML is not just a research question anymore
- ML-based products are being designed and deployed
 - **new responsibility for AI scientists, engineers, entrepreneurs, governments**



AI is a Powerful Tool

- AI = **tool**
- Dual-use
- **Wisdom race:** advances in technology vs advances in wisdom
- **How to maximize its beneficial use and minimize its misuse?**



Montreal Declaration (2017)



< >

Montréal Declaration
Responsible AI_

< / >

1- WELL-BEING PRINCIPLE

2- RESPECT FOR AUTONOMY PRINCIPLE

3- PROTECTION OF PRIVACY AND INTIMACY PRINCIPLE

4- SOLIDARITY PRINCIPLE

5- DEMOCRATIC PARTICIPATION PRINCIPLE

6- EQUITY PRINCIPLE

7- DIVERSITY INCLUSION PRINCIPLE

8- PRUDENCE PRINCIPLE

9- RESPONSIBILITY PRINCIPLE

10- SUSTAINABLE DEVELOPMENT PRINCIPLE

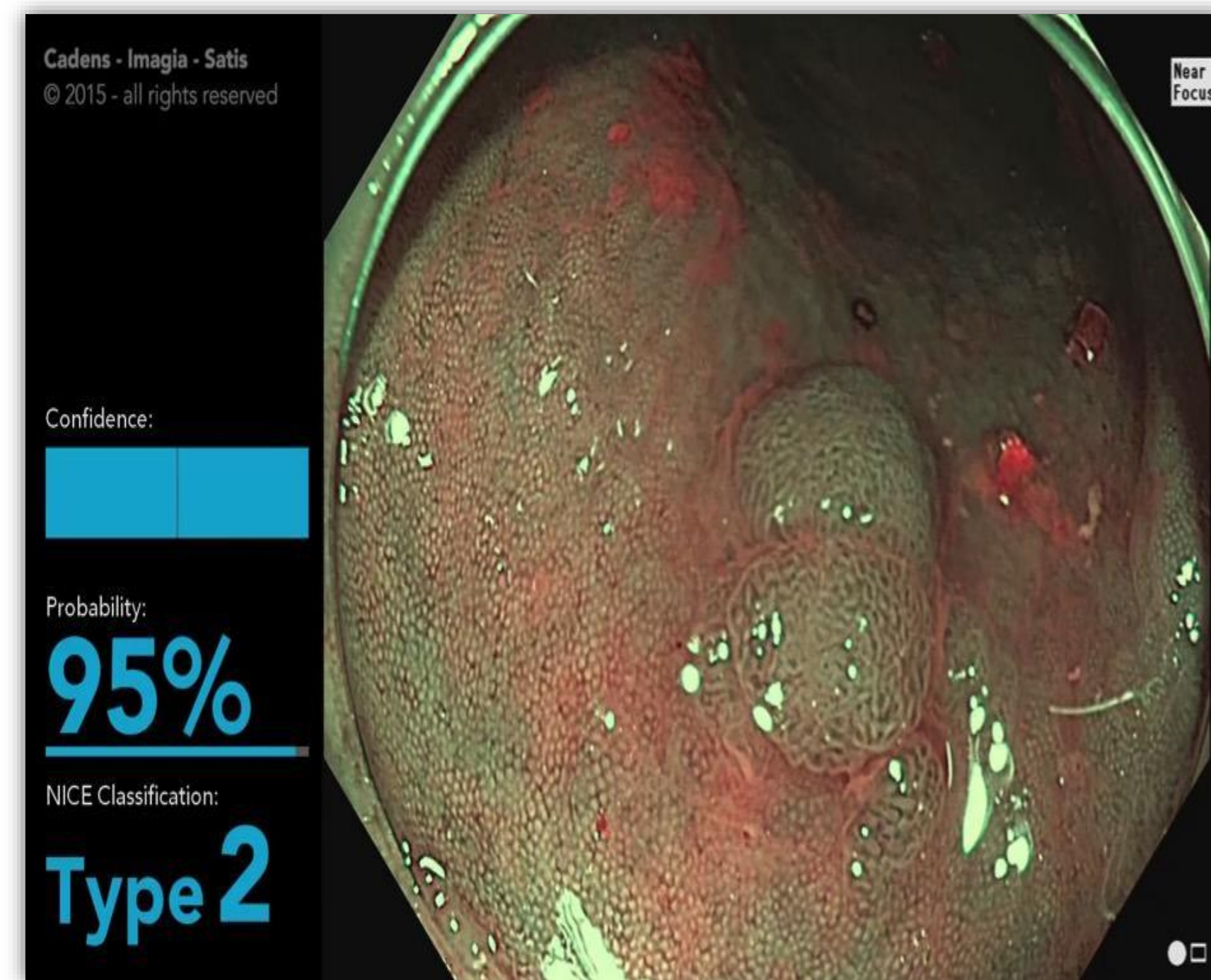
Dangers and Concerns with AI

- Big Brother and killer robots
- Misery for jobless people, at least in transition
- Manipulation from advertising and social media
- Reinforcement of social biases and discrimination
- Increased inequality and power concentration in few people, companies & countries



Beneficial AI Potential

- Medical applications
 - Environmental applications
 - Educational applications
 - Democratizing justice
 - Humanitarian applications
- Etc.



*Imagia detects cancer cells,
helps doctors*

Market Failure: Societal vs Commercial Value of Innovations

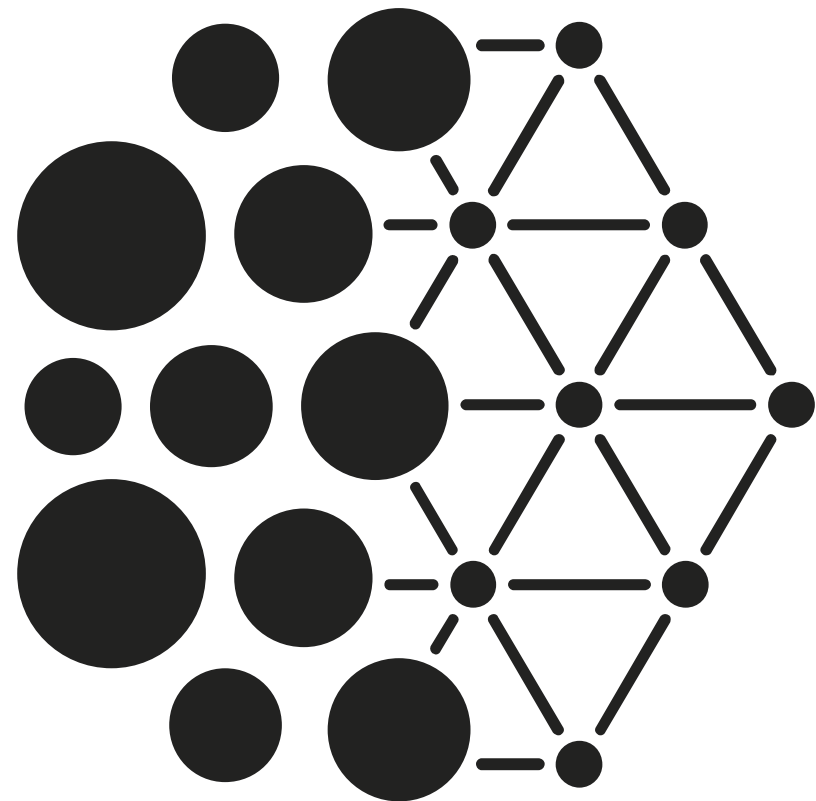
Some innovations would be greatly beneficial to society but do not have enough commercial value (or too much risk) for current markets

e.g. involve **PUBLIC GOODS**

- **AI for discovering new antibiotics: needed to prevent next pandemic**
- AI for discovering new carbon capture materials
- AI to fight climate change, AI to bring better education worldwide

→ Governments should invest \$ and academics should invest efforts to fill the gap

THANK YOU!



Mila

Université 
de Montréal

 McGill