



Last week at NeurIPS

Overview

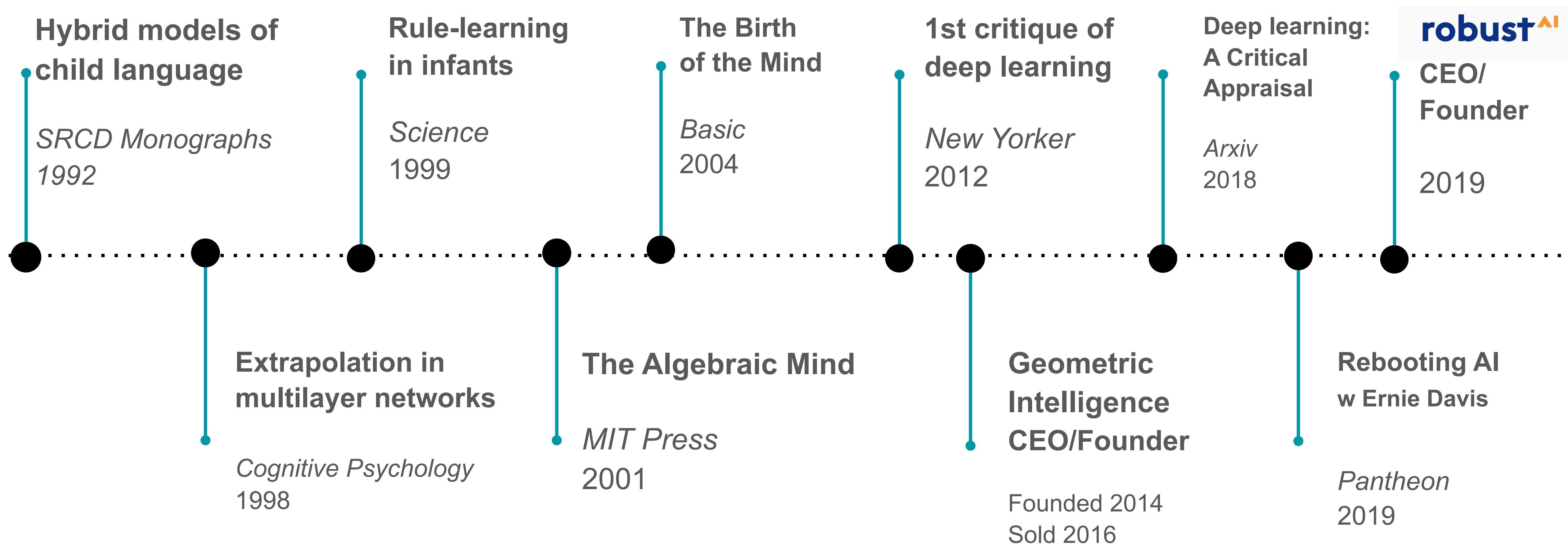
- A bit of history, and a sense of where I am coming from
- My take on Yoshua's view
 - Many agreements; some important disagreements
 - Prescription for going forward

Part I:

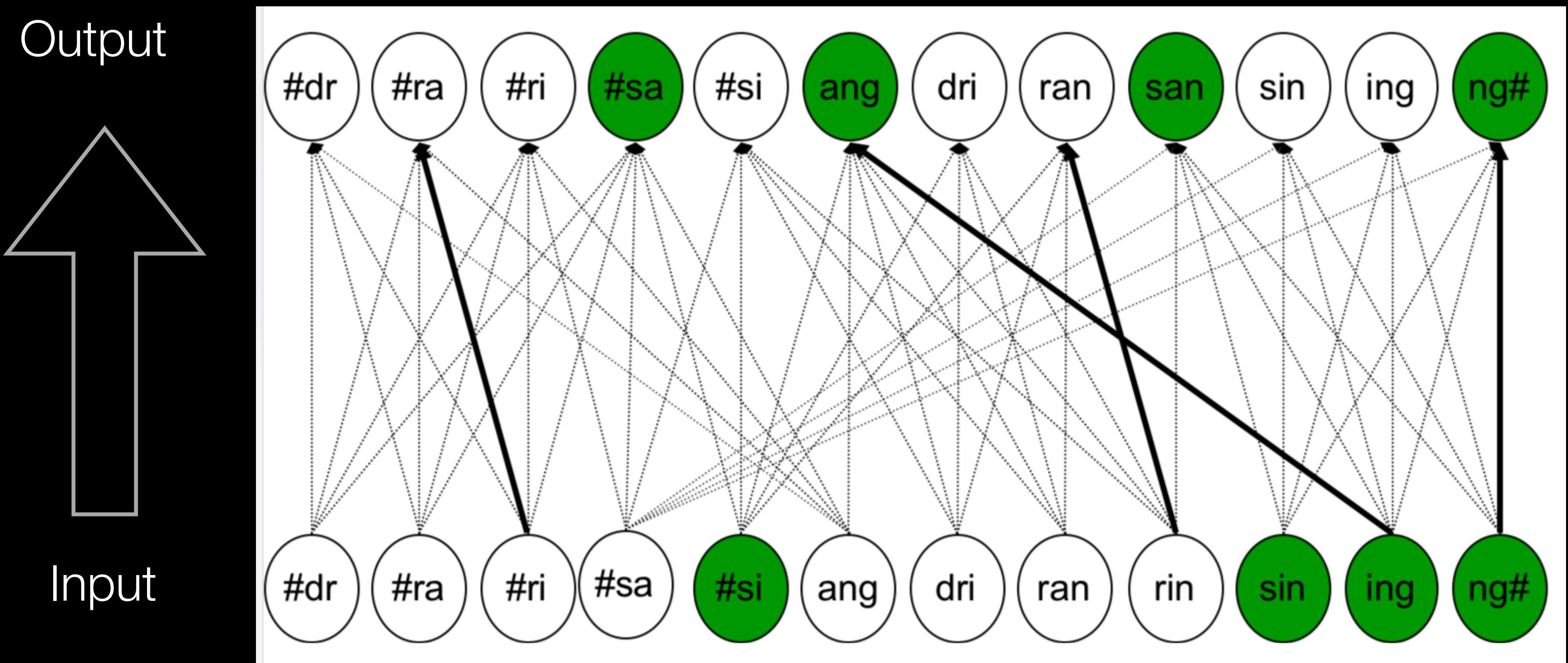
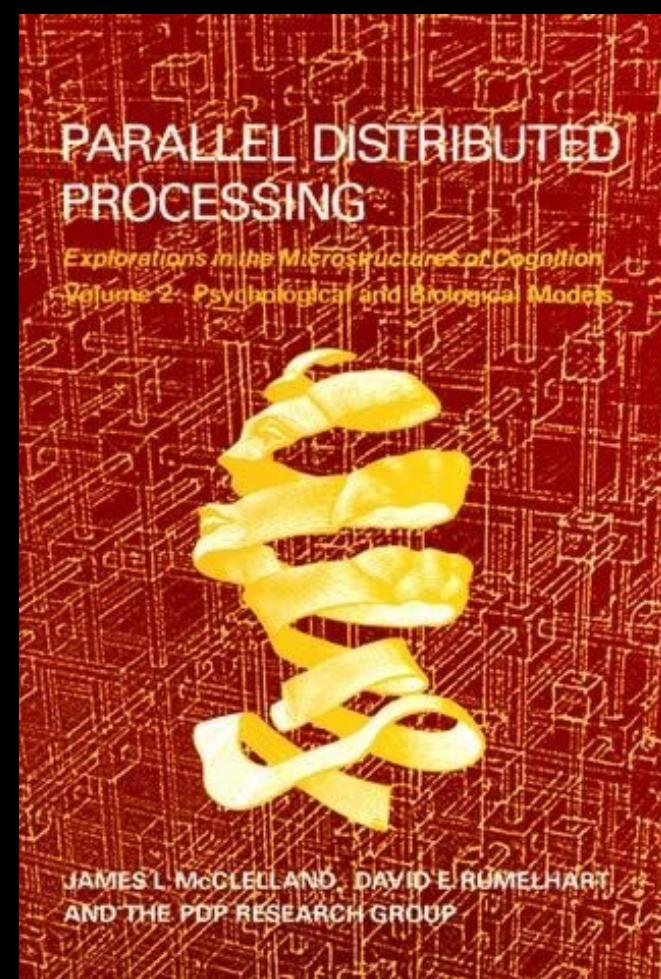
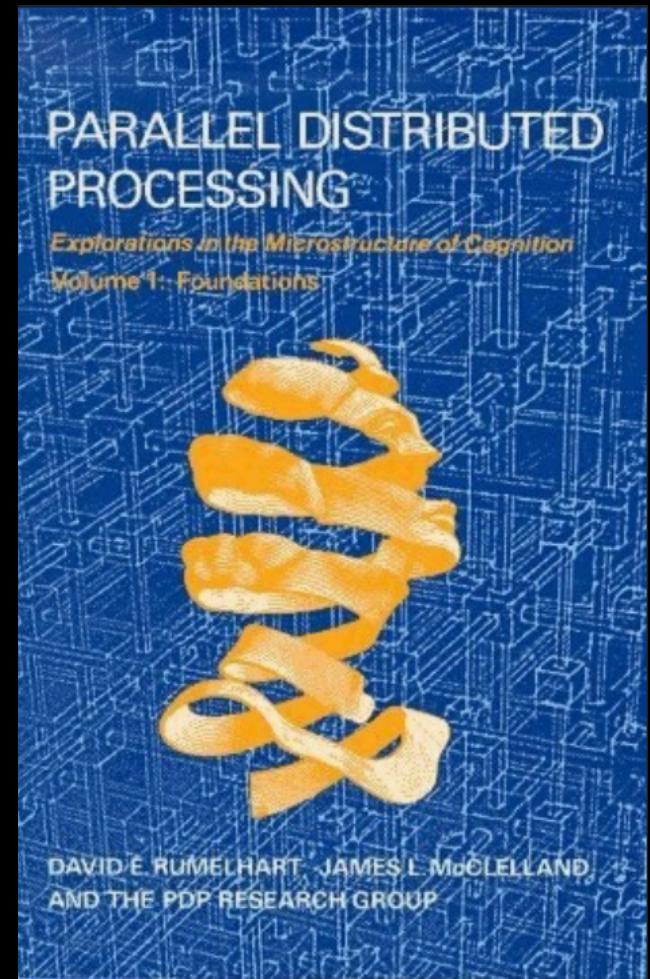
how I see AI, deep learning, and current ML,
and how I got here

aka "What's a nice cognitive scientist like me doing in a place like this?"

A cognitive scientist's journey, with implications for AI



1986: Rules versus connectionism (neural networks)

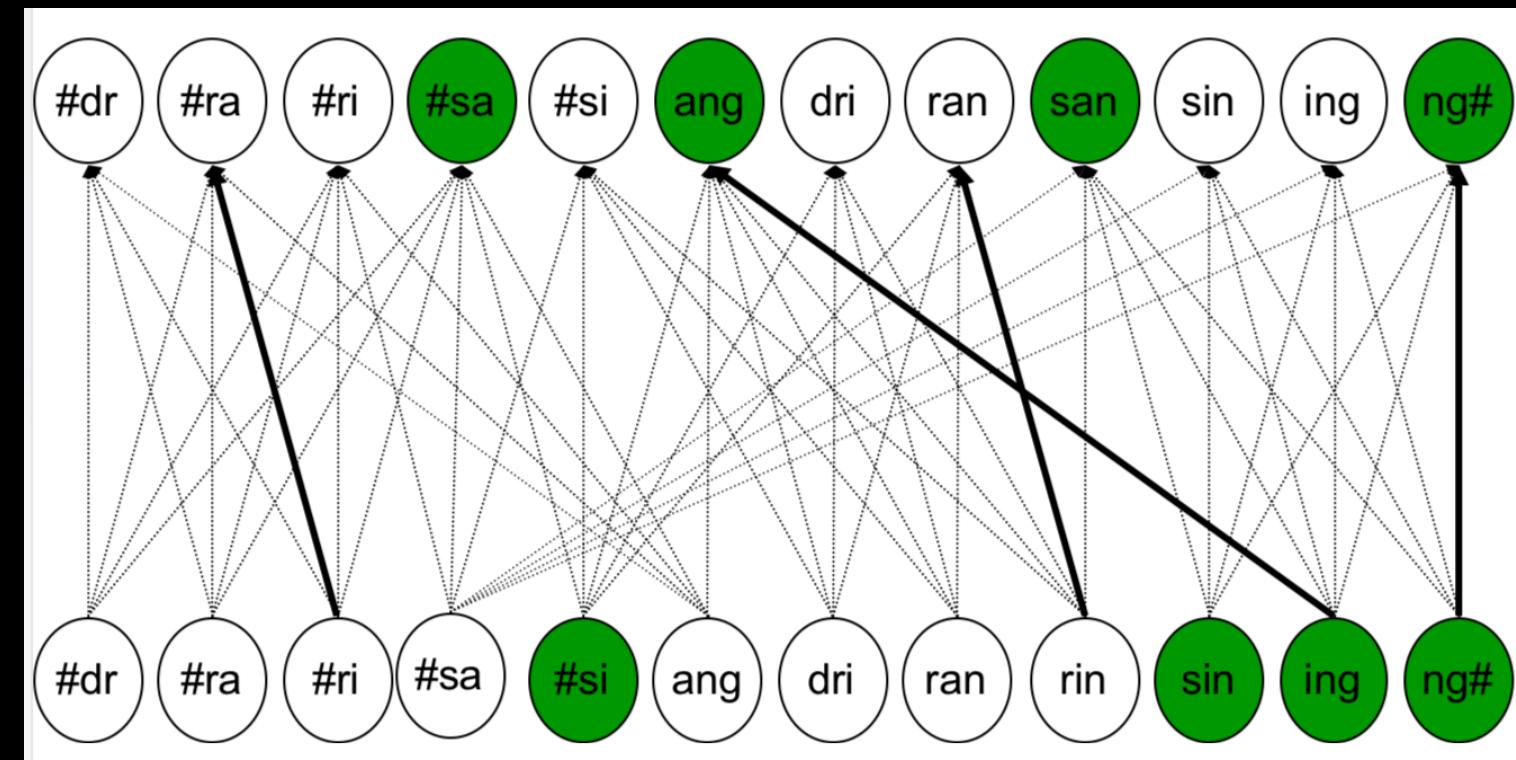


- The most provocative paper was Rumelhart & McClelland (1986) on the past tense:
 - they argued that children's overregularization errors, like *breaked* and *goed--long* thought to be the iconic example of a symbolic rule--might instead be the product of a neural network that had no rules at all.
- "Eliminative connectionism" & the "great past tense debate" was born.

A huge war raged across the cognitive sciences....

the debate

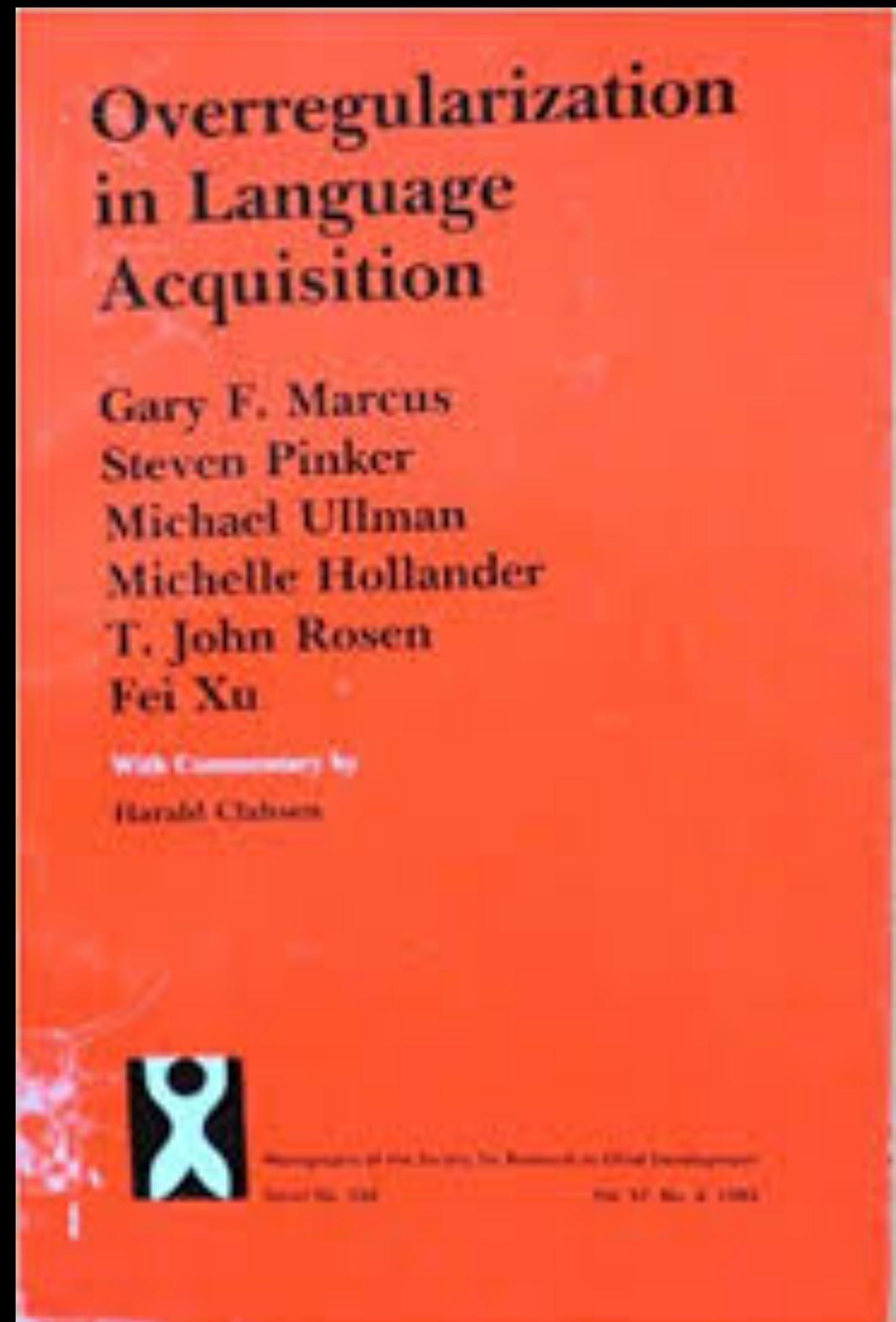
- Up to that point
 - Most linguistics/cogsci was about rules
 - **S => NP VP**
 - **NP => Det Noun**
 - Most AI (eg expert systems) was also all about rules
- Rumelhart & McClelland argued that we didn't need rules at all.
 - Even a child's error like *breaked* might be *in principle* the product of a neural net rather than a rule



Were the actual empirical data from child language development consistent with their model?

1992: Why do kids (sometimes) say *breaked* rather than *broke*?

- In my thesis, supervised by Steve Pinker, I studied 11,500 child utterances
- We found that neural nets made incorrect predictions
- Instead, we argued for a *compromise*: a **hybrid model**:
 - **rule for regulars (*walk-walked*)**
 - stem + ed = past
 - **neural nets for irregulars (*sing-sang*)**
 - Overreg. errors resulted from applying rule by default when no strong response from irregular memory

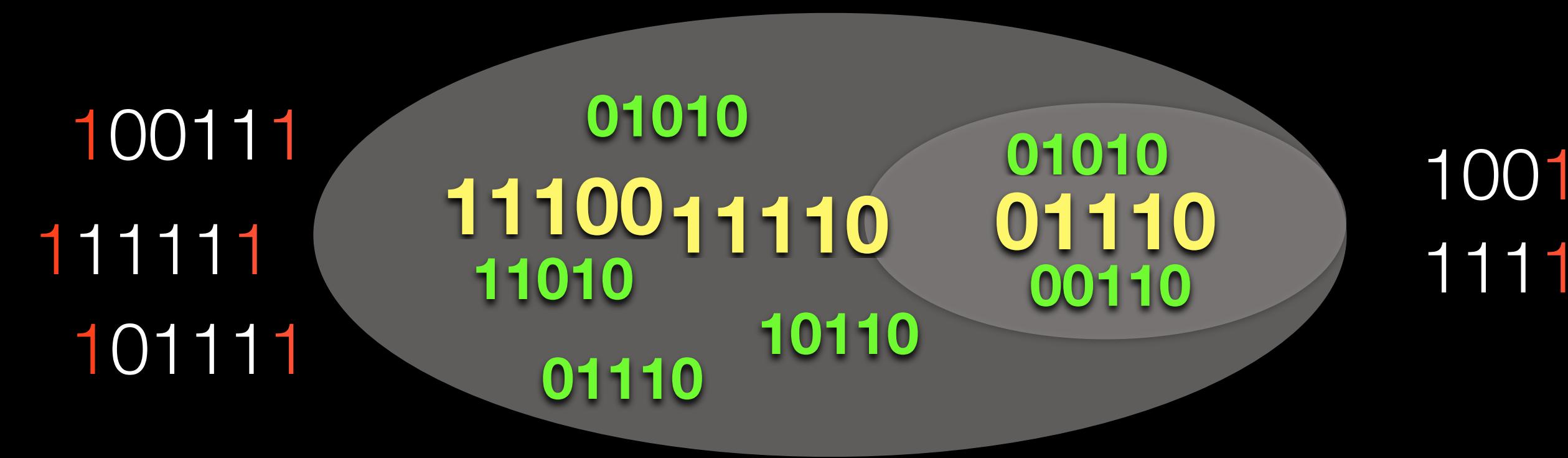


Marcus et al (1992, *SRCD Monographs*),
See also Pinker's *Words and Rules*

1998: Extrapolation & Training Space

- People in those days often talked about neural networks "learning the rule" in a given pattern of data, but I discovered that they often missed some very basic rules.

0110 -> 0110
1100 -> 1100
1010 -> 1010
<hr/>
1111 -> 1110



near **perfect** at learning **specific training examples**
good at generalizing **within** some space of training examples
poor at generalizing **outside** that space of training examples

- "the class of eliminative connectionist models that is currently popular cannot learn to extend universals outside the training space"
 - showed that same result applied even if there were hidden layers (predecessors to today's deep networks)
 - showed how that it derived from the intrinsic nature of localist training rules such as backprop (& Hebb.)

1999: Rule learning in 7 month old infants



- direct, deliberate test (given certain assumptions) of outside-the-training-space generalization by human infants
- training: *la ti ti, ga na na*, etc
- test: all new vocabulary, using new set of phonemes
 - some with same grammar, some with different grammar
 - e.g., *wo fe fe* [ABB] vs *wo wo fe* [AAB]
 - infants looked longer to items following new grammar
- conclusion: infants could generalize outside training space, where many neural nets could not
 - best characterized as learning algebraic rules
 - replicated multiple times, including w newborns (Gervain et al 2012)

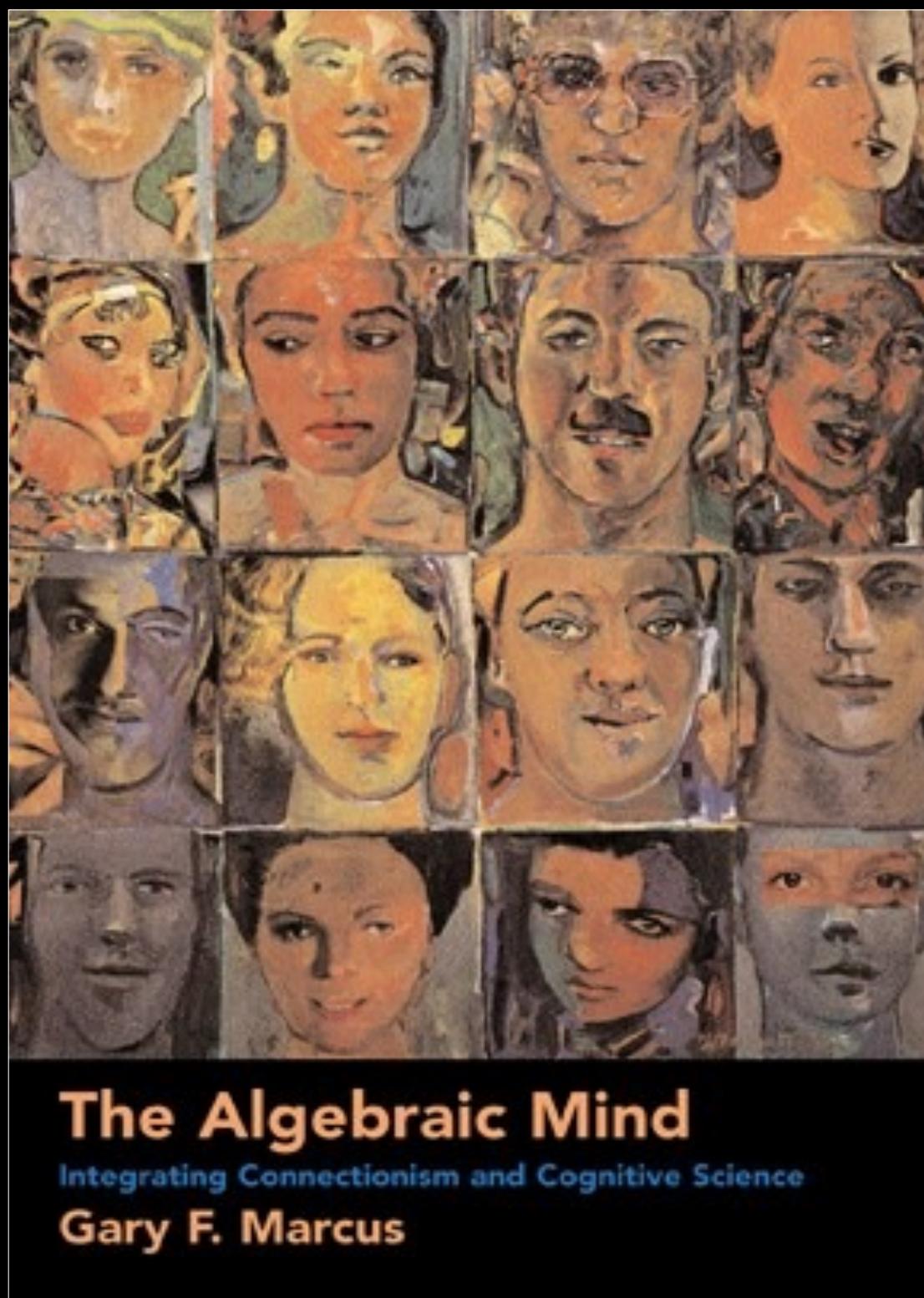
Marcus et al (1999, *Science*)

2001: The Algebraic Mind

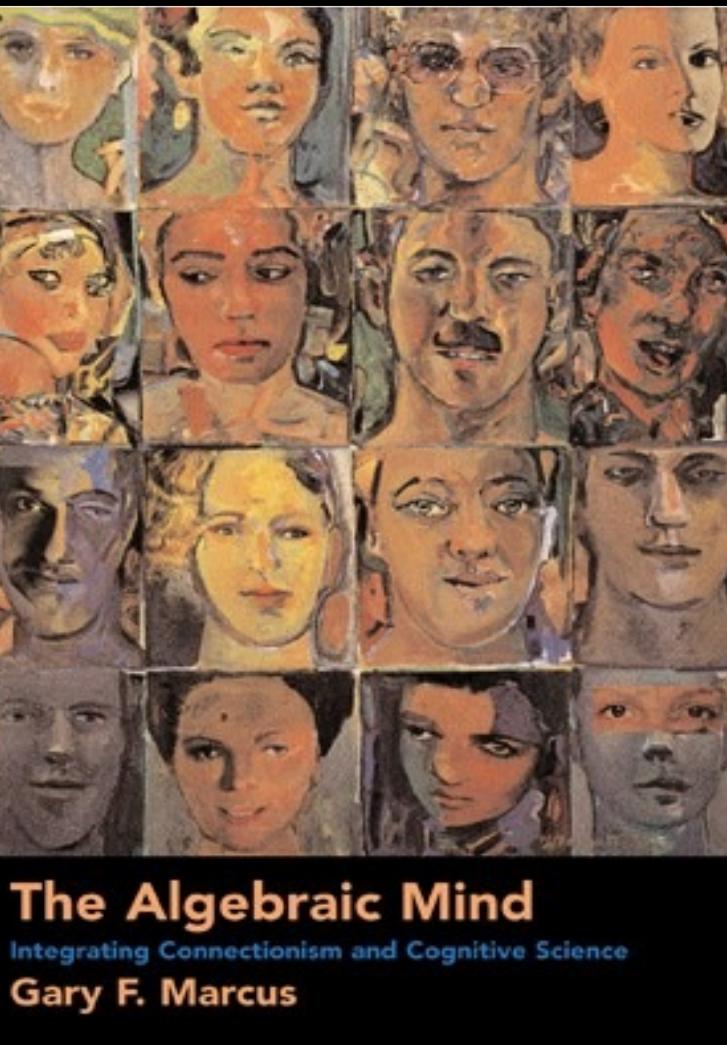
Three key ingredient missing from multilayer perceptrons:

- the ability to freely generalize abstract relations
- the ability to robustly represent complex relations between bits of knowledge [i.e., structured representations, recursion, compositionality]
- a systematic way to track individuals separately from kinds

"these limitations [... undermine] multilayer perceptron accounts of linguistic inflection, artificial language learning, object permanence, and object tracking. Such models simply cannot capture the flexibility and power of everyday reasoning. "



Marcus 2001, MIT Press



- Key components of symbol-manipulation
 - **variables** ($x, y, NP \dots$)
 - **instances** (2, 3, the boy...)
 - **binding** (NP currently equals the boy)
 - **operations over variables** (e.g. addition, concatenation, comparison)
- Together these mechanisms provides a **natural solution to the free generalization problem**
 - Computer programs (e.g., functions and libraries) and algebra for example are routinely defined in terms of operations over variables
 - And that allows functions (e.g., FACTORIAL) to **automatically generalize to all instances of some class** (e.g., integer)
 - Pretty much all of the world's software takes advantage of this fact
 - My argument (eg from baby data) was that human cognition appeared to as well.

Marcus 2001,
MIT Press

The Algebraic Mind

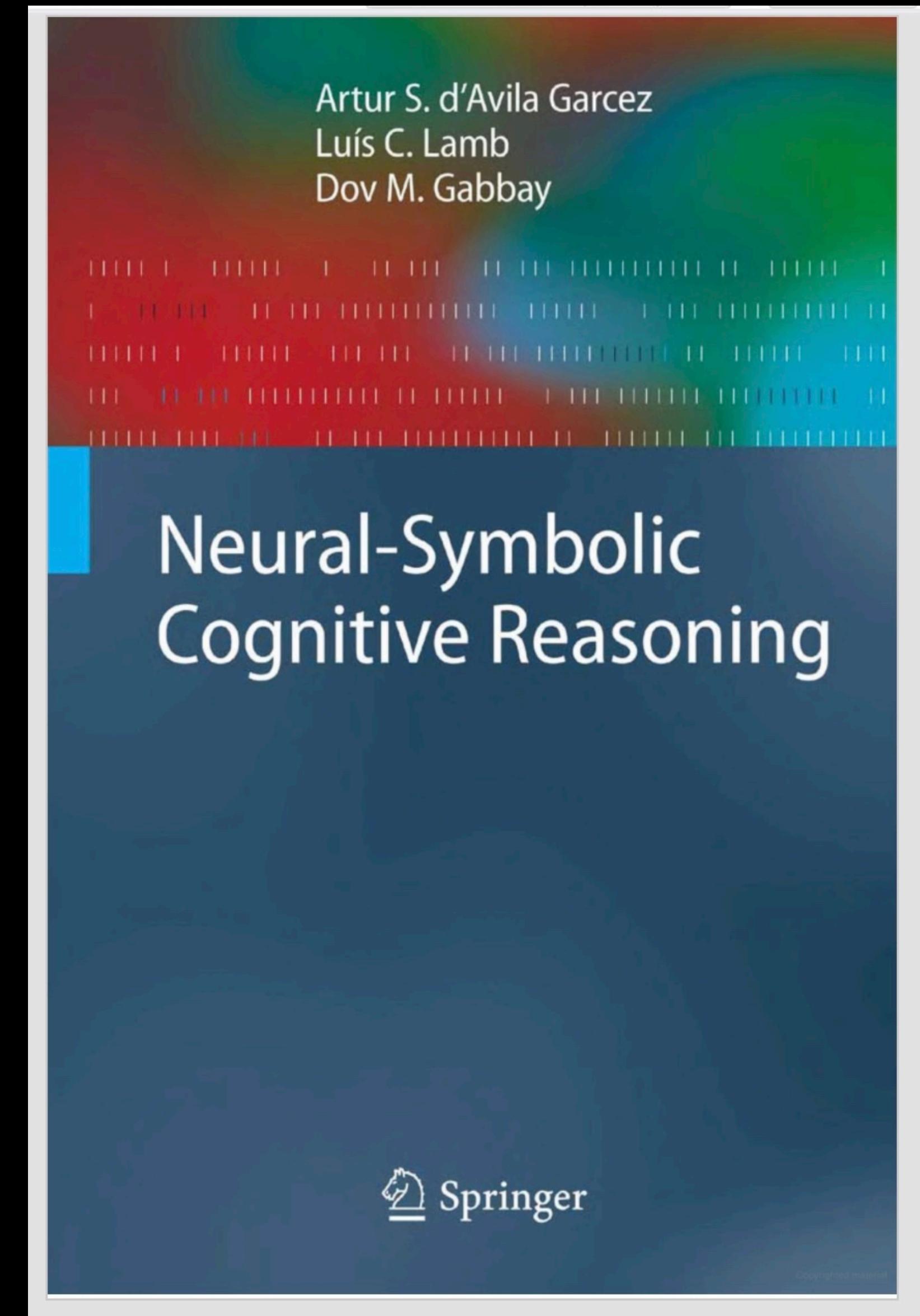
Integrating Connectionism and Cognitive Science

The point of the subtitle, and the book, was that we needed to have neural networks *alongside* of symbol-manipulation, integrated into a smooth whole

“...even if the components of symbol-manipulation do play a real and robust role in our mental life, it is unlikely that they exhaust the set of components for cognition. Instead, it seems likely that many other basic computational elements [such as images and analog representations] play important roles in cognition... even multilayer perceptrons may play a role in some aspects of our mental life.”

Marcus 2001, MIT Press

- Until roughly 2018, mainstream ML largely ignored *The Algebraic Mind*
- But *The Algebraic Mind* inspired the seminal book on neurosymbolic approaches
- And, as we will see, *Algebraic* also anticipated much of Yoshua's current argument



2008

2012: The Rise of Deep Learning

NEWS DESK

IS “DEEP LEARNING” A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

By Gary Marcus November 25, 2012

Realistically, deep learning is only part of the larger challenge of building intelligent machines. Such techniques lack ways of representing causal relationships (such as between diseases and their symptoms), and are likely to face challenges in acquiring abstract ideas like “sibling” or “identical to.” They have no obvious ways of performing logical inferences, and they are also still a long way from integrating abstract knowledge, such as information about what objects are, what they are for, and how they are typically used. The most powerful A.I. systems, like Watson, the machine that beat humans in “Jeopardy,” use techniques like deep learning as just one element in a very complicated ensemble of techniques, ranging from the statistical technique of Bayesian inference to deductive reasoning.

2018: Critique of deep learning

Deep Learning:
A Critical Appraisal

Gary Marcus¹
New York University

Abstract

Although deep learning has historical roots going back decades, neither the term “deep learning” nor the approach was popular just over five years ago, when the field was reigned by papers such as Krizhevsky, Sutskever and Hinton’s now classic 2012 (Krizhevsky, Sutskever, & Hinton, 2012) deep net model of Imagenet.

What has the field discovered in the five subsequent years? Against a background of considerable progress in areas such as speech recognition, image recognition, and game playing, and considerable enthusiasm in the popular press, I present ten concerns for deep learning, and suggest that deep learning must be supplemented by other techniques if we are to reach artificial general intelligence.

Marcus 2018, arXiv

- Outlined 10 problems for deep learning
 - Failure to extrapolate beyond space of training was at core of the argument
- Got a ton of flak (e.g., on Twitter)
- Oft-misrepresented. Actual conclusion:

“Despite all of the problems I have sketched, I don’t think that we need to abandon deep learning... Rather, we need to reconceptualize it: not as a universal solvent, but simply as one tool among many”

The central conclusions of my academic work on cognitive science, and its implications for AI

- The value of hybrid models [1992, 2001, 2018, 2019, etc] that include both symbol-manipulating AND associative pattern recognition elements
- The importance of extrapolation, and the weakness of pure deep learning thereon [1998, 2001, 2018]
- The importance of compositionality [2001, 2018, 2019]
- The importance of acquiring and representing relationships [2001, 2018]
- The importance of causality [2001, 2018]
- The importance of memory as a substrate for operations over variables [2001]

Part II: Yoshua

Some thoughts on his views, and how I think they have changed, how he has (mis)represented me, and how our views are and are not similar

First things first: I admire Yoshua



An Epidemic of AI Misinformation

30.NOV.2019

Mercifully, not everyone in the field overrepresents their work; in the last year or so I have seen terrific, balanced talks by [Pieter Abbeel](#) and [Yoshua Bengio](#), both noting what deep learning (and deep reinforcement learning) do well, and yet at the same time articulating the challenges ahead, and bluntly acknowledging how far we need to go. (Abbeel emphasized the gap between

Tackling Climate Change with Machine Learning

[David Rolnick](#), [Priya L. Donti](#), [Lynn H. Kaack](#), [Kelly Kochanski](#), [Alexandre Lacoste](#), [Kris Sankaran](#), [Andrew Slavin Ross](#), [Nikola Milojevic-Dupont](#), [Natasha Jaques](#), [Anna Waldman-Brown](#), [Alexandra Luccioni](#), [Tegan Maharaj](#), [Evan D. Sherwin](#), [S. Karthik Mukkavilli](#), [Konrad P. Kording](#), [Carla Gomes](#), [Andrew Y. Ng](#), [Demis Hassabis](#), [John C. Platt](#), [Felix Creutzig](#), [Jennifer Chayes](#), [Yoshua Bengio](#)

(Submitted on 10 Jun 2019 ([v1](#)), last revised 5 Nov 2019 (this version, [v2](#)))

Climate change is one of the greatest challenges facing humanity, and we, as machine learning experts, may wonder how we can help. Here we describe how machine learning can be a powerful tool in reducing greenhouse gas emissions and helping society adapt to a changing climate. From smart grids to disaster management, we identify high impact problems where existing gaps can be filled by machine learning, in collaboration with other fields. Our recommendations encompass exciting research questions as well as promising business opportunities. We call on the machine learning community to join the global effort against climate change.

Yoshua should be a role model for us all: he is intellectually honest about the challenges his models face, and sincere in using his talents to help make the world a better place

My differences are mainly with Yoshua's earlier (e.g., 2014-2015) views

Our first conversation



Montreal, NeurIPS 2014

- I thought Yoshua:
 - put too much faith in black box deep networks
 - relied too heavily on larger data sets to yield the answer
 - was "System I" all the way with little interest in alternatives
- I could find little common ground

"one quote that stood out for me was an answer given by Prof. Bengio at the end of his keynote, regarding negation and quantification, and how a Neural Network model deals with them: “I don’t know. But it learns to do what it needs to do.”

"during the Q&A since it was an audience of linguists, they asked how the NN models he presented could handle various phenomena. I...the answer in almost every case was simply "make sure the data set is large enough to include examples of the phenomena"

quotes from audience reactions at 11th International Conference on Computational Semantics 2015

Recently, however Yoshua has taken a sharp turn towards many of the positions I have long advocated

- Fundamental limits on current deep learning
- The need for hybrid models (with an important difference I will discuss)
- The critical importance of extrapolation, and the weakness of pure deep learning thereon
- The importance of compositionality
- The importance of acquiring and representing relationships between entities
- The importance of causality (Pearl)
- The need for more heterogeneous architecture

THE STATE OF DEEP LEARNING

Amazing progress in this century

- Is it enough to just grow datasets, model sizes, computer speed?

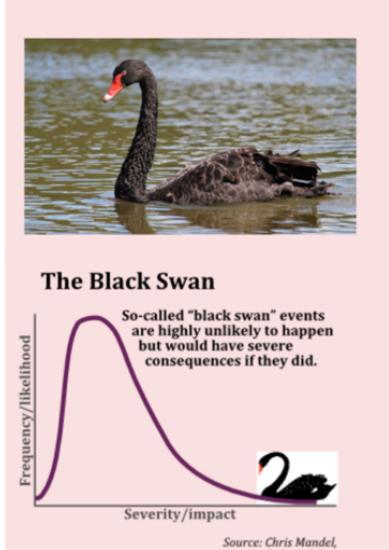
Still far from human-level AI!

- Sample efficiency
- Human-provided labels
- Stupid errors
- Next step completely different from deep learning?

FROM IID TO OOD

Classical ML theory for iid data
Artificially shuffle the data to achieve that?

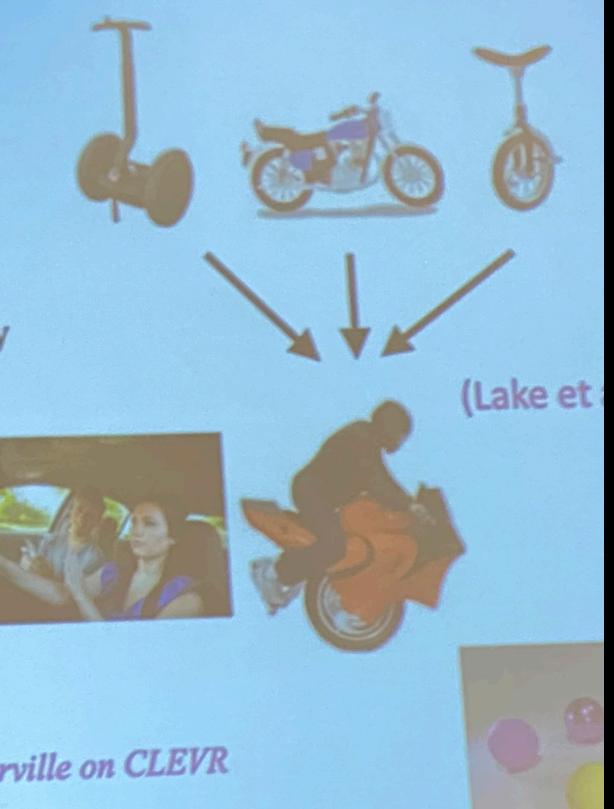
Out-of-distribution generalization
No free lunch: need new assumptions to replace iid assumption, for ood generalization



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

(Lake et al. 2017)
(Bahdanau & Courville ICLR 2019)
Going work by Bahdanau et al & Courville on CLEVR



OPERATING ON SETS OF POINTABLE OBJECTS WITH DYNAMICALLY RECOMBINED MODULES



MISSING TO EXTEND DEEP LEARNING TO REACH HUMAN-LEVEL AI

- Out-of-distribution generalization & transfer
- Higher-level cognition: system 1 → system 2
 - High-level semantic representations
 - Compositionality
 - Causality



Disagreements

1. What my position is
2. The right way to build hybrid models
3. Innateness
4. The significance of the fact that the brain is a neural network
5. What we mean by compositionality, and how we expect it to be solved

1. Yoshua's (mis)representation of my position (1 of 2)

IEEE Spectrum: What do you think about all the discussion of deep learning's limitations?

Yoshua Bengio: Too many public-facing venues don't understand a central thing about the way we do research, in AI and other disciplines: We try to understand the limitations of the theories and methods we currently have, in order to extend the reach of our intellectual tools. So deep learning researchers are looking to find the places where it's not working as well as we'd like, so we can figure out what needs to be added and what needs to be explored.

This is picked up by people like Gary Marcus, who put out the message: "Look, deep learning doesn't work."* But really,

- "picked up by" gets the chronology wrong, and the (mis)quote misrepresents my position.
- True, I often cite Yoshua's recent work on DL limits -- but not because I got the ideas from him, as he implies, but because I hope people will listen to him (as an insider) where they haven't listened to me (as an outsider).
- I never ever say that deep learning doesn't work; rather (over and over) I say it has limits and is just one tool among many.

1. Yoshua's (mis)representation of my position (2 of 2)

A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms

Yoshua Bengio^{1,2,5}, Tristan Deleu¹, Nasim Rahaman⁴, Nan Rosemary Ke³, Sébastien Lachapelle¹, Olexa Bilaniuk¹, Anirudh Goyal¹ and Christopher Pal^{3,5}

1. Introduction

Current machine learning methods seem weak when they are required to generalize beyond the training distribution, which is what is often needed in practice. It is not enough to obtain good generalization on a test set sampled from the same distribution as the training data, we would also like what has been learned in one setting to generalize well in other related distributions. These distributions may involve the same concepts that were seen previously by the learner, with the changes typically arising because of actions of agents. More generally, we would like what has been learned previously to form a rich base from which very fast adaptation to a new but related distribution can take place, i.e., obtain good transfer. Some new concept may have to be learned but because most of the other relevant concepts have already been captured by the learner (as well as how they can be composed), learning can be very fast on the transfer distribution.

Short of any assumption, it is impossible to have a successful transfer to an unrelated distribution. In

"multilayer perceptron[s] cannot generalize [a certain class of universally quantified function] outside the training space. . . In some cases it appears that humans can freely generalize from restricted data, [in these cases a certain class of] multilayer perceptions that are trained by back-propagation are inappropriate

. . . Such models simply cannot capture the flexibility and power of everyday reasoning. "

--Marcus, 2001

Bengio et al 2019

- Yoshua recently started framing his work around
 - the challenge in generalizing beyond the training distribution
 - and the corresponding need for complementary systems
- This echoes the central argument of *The Algebraic Mind* (2001)
 - but does not credit *TAM* for having foreseen the central challenge for pure deep learning in extrapolation (relative to training distributions) nor for having foreseen the consequent computational necessity for hybrid systems
- This omission devalues my contributions, and hence further misrepresents my background in the field.

2. What kind of hybrid should we seek?

HYBRID MODELS

- neural nets (e.g., vectors, gradients, optimization and distributed representations) for categorization, associative memory, aspects of motor control, etc
- symbol-manipulation for generalization of abstract patterns; for reasoning and for language

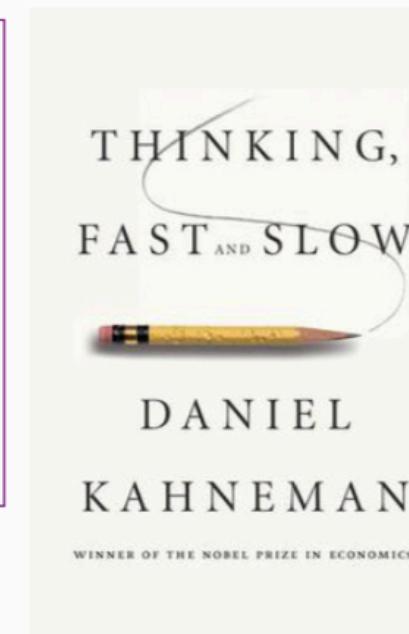
Marcus, 1992; 2001; 2008; 2019

SYSTEM 1 VS. SYSTEM 2 COGNITION

2 systems (and categories of cognitive tasks):

System 1

- Intuitive, fast, **UNCONSCIOUS**, non-linguistic, habitual
- Current DL



System 2

- Slow, logical, sequential, **CONSCIOUS**, linguistic, algorithmic, planning, reasoning
- Future DL



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

Bengio, NeurIPS, 2019, inspired by Kahneman

- First question: are these even different?
- Second question: are they incompatible?
- Third question: how could we tell?

To argue against symbol-manipulation, you have to show that your system doesn't *implement* symbols

Three levels of description (David Marr, 1982)

Computational

Why do things work the way they do?
What is the goal of the computation?
What are the unifying principles?



Algorithmic

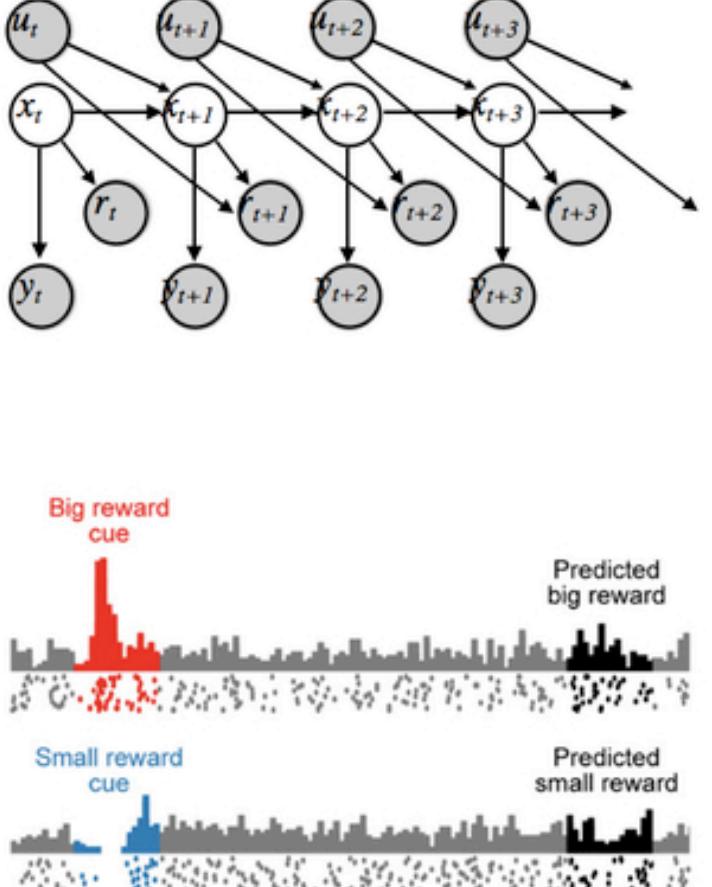
What representations can implement such computations?
How does the choice of representations determine the algorithm?

Implementational

How can such a system be built in hardware?
How can neurons carry out the computations?

maximize:

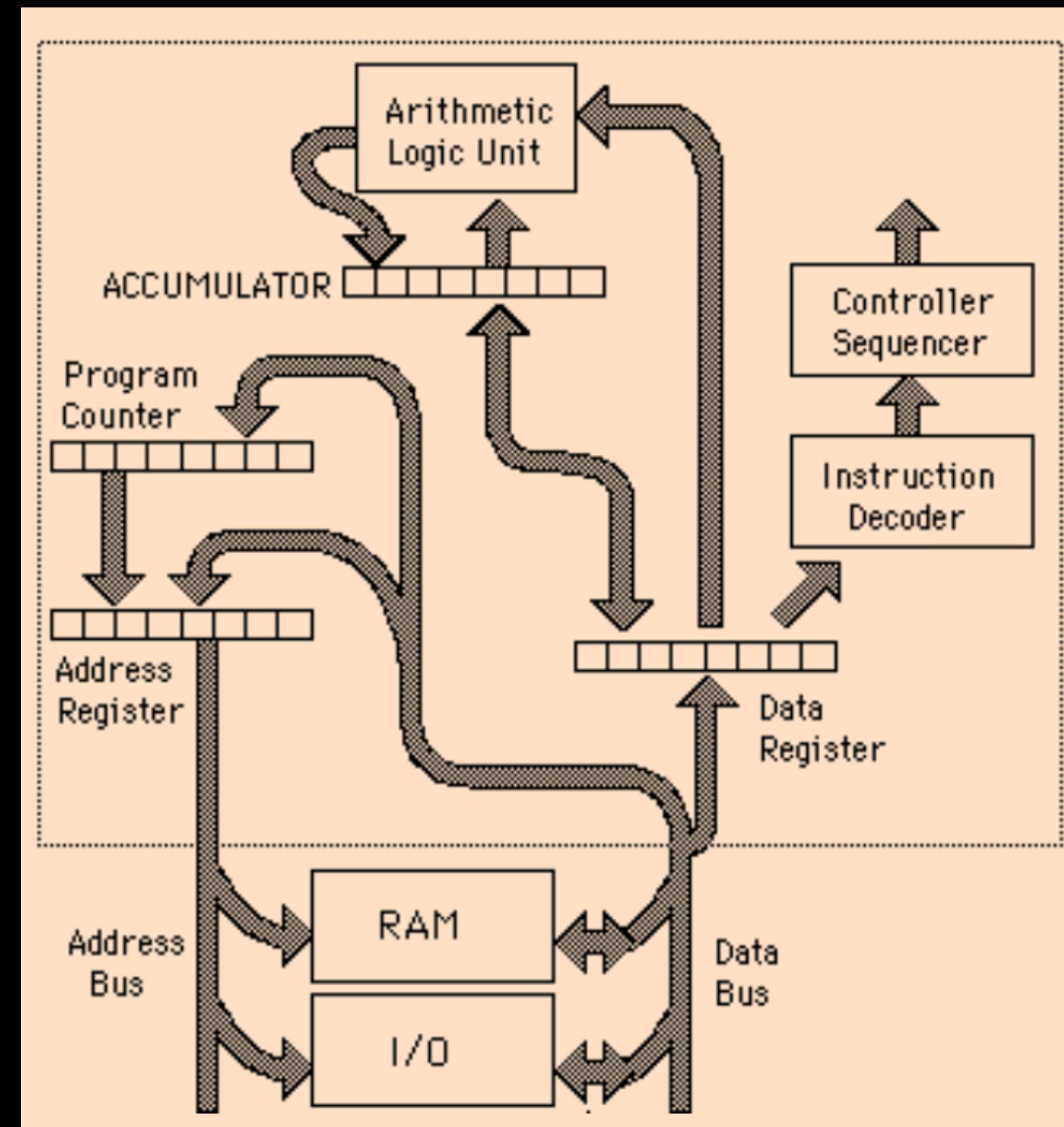
$$R_t = r_{t+1} + r_{t+2} + \dots + r_T$$



- Yoshua hasn't actually shown this

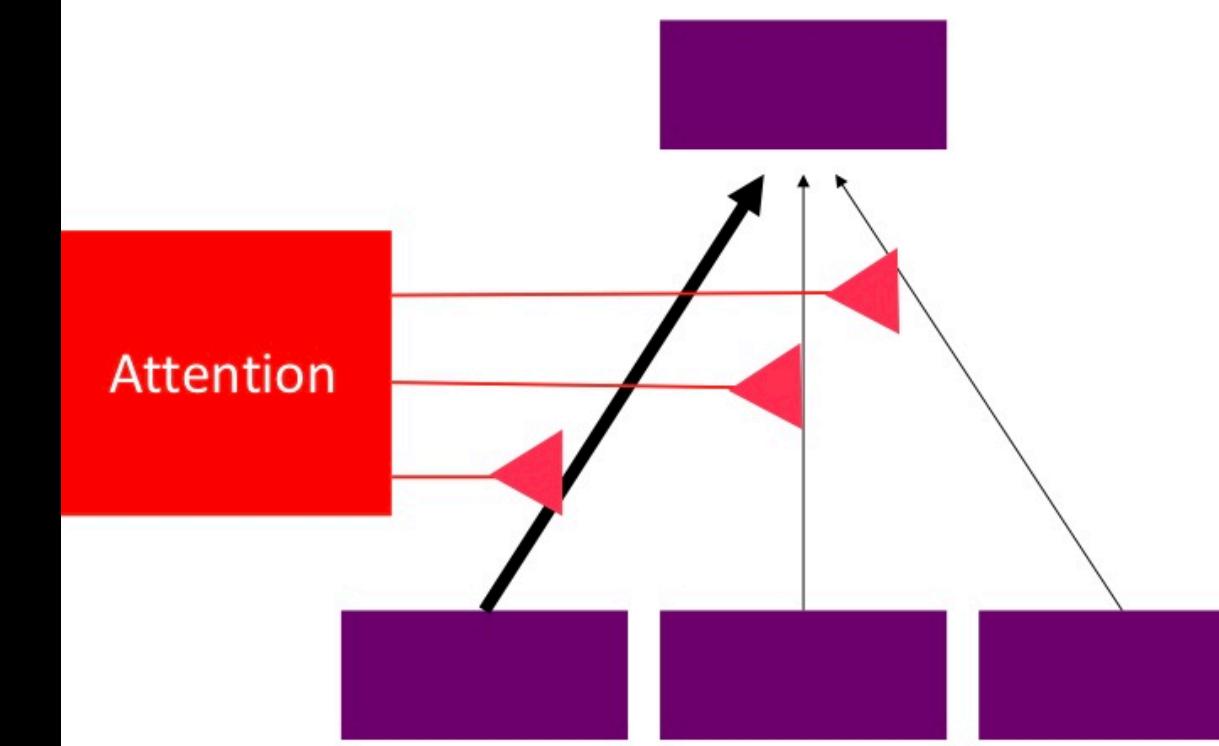
See also *implementational vs eliminative connectionism*
(Fodor & Pylyshyn, 1988; Marcus, 2001)

Attention here looks a lot like a means for manipulating symbols



Microprocessor

FROM ATTENTION TO INDIRECTION



- Attention = dynamic connection
- Receiver gets the selected value
- Value of what? From where?
 - Also send 'name' (or key) of sender
- Keep track of 'named' objects: indirection
- Manipulate sets of objects (transformers)

Critical attention mechanism effectively behaves as a mechanism for storing and retrieving values of variables from registers

"We tried symbols and they don't work"

"What you are proposing [a neurosymbolic hybrid] does not work. This is what generations of AI researchers tried for decades and failed."

Bengio, in a letter to a young student, 2018

- Common refrain, totally misleading.
 - Google Search is a hybrid: knowledge graph + deep learning [eg BERT]
 - AlphaZero is also hybrid
 - OpenAI's Rubik's solver is a hybrid (cognitive part uses symbol-manipulation)

The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, Jiajun Wu

(Submitted on 26 Apr 2019)

We propose the Neuro-Symbolic Concept Learner (NS-CL), a model that learns visual concepts, words, and semantic parsing of sentences without explicit supervision on any of them; instead, our model learns by simply looking at images and reading paired questions and answers. Our model builds an object-based scene representation and translates sentences into executable, symbolic programs. To bridge the learning of two modules, we use a neuro-symbolic reasoning module that executes these programs on the latent scene representation. Analogical to human concept learning, the perception module learns visual concepts based on the language description of the object being referred to. Meanwhile, the learned visual concepts facilitate learning new words and parsing new sentences. We use curriculum learning to guide the searching over the large compositional space of images and language. Extensive experiments demonstrate the accuracy and efficiency of our model on learning visual concepts, word representations, and semantic parsing of sentences. Further, our method allows easy generalization to new object attributes, compositions, language concepts, scenes and questions, and even new program domains. It also empowers applications including visual question answering and bidirectional image-text retrieval.

Lots of knowledge is not "conveniently representable" with rules

"There is knowledge which can be conveniently represented as rules... But there is also a large body of intuitive knowledge, which is NOT conveniently representable that way."

Bengio, in a letter to a young student,
2018

- True
 - And a strong reason not to use symbols all-the-way-down
 - But *not* an argument to toss the baby with the bathwater
 - Google Search is again a great, large-scale example: best performance comes from knowledge graph + BERT

3. Innateness

The screenshot shows a document with a header, author information, an abstract, and a note at the bottom.

Title: Innateness, AlphaZero, and Artificial Intelligence

Author: Gary Marcus¹
New York University

Abstract:

The concept of innateness is rarely discussed in the context of artificial intelligence. When it is discussed, or hinted at, it is often the context of trying to reduce the amount of innate machinery in a given system. In this paper, I consider as a test case a recent series of papers by Silver et al (Silver et al., 2017a) on AlphaGo and its successors that have been presented as an argument that "even in the most challenging of domains: it is possible to train to superhuman level, without human examples or guidance", "starting *tabula rasa*."

I argue that these claims are overstated, for multiple reasons. I close by arguing that artificial intelligence needs greater attention to innateness, and I point to some proposals about what that innateness might look like.

¹ Departments of Psychology and Neural Science, New York University, gary.marcus@nyu.edu. This manuscript is based on a pair of lectures given at NIPS 2017, a brief conversation there with Demis Hassabis, and a debate that I had with Yann LeCun at NYU on October 5, 2017. I thank those audiences for discussion, and Dave Barner, Annie Duke, Ernie Davis, Pedro Domingos, Ken Forbus, Danny Kahneman, Stefano Pacifico, Ajay Patel, Elizabeth Spelke and Brad Wyble for comments.

- My view:
 - Innateness is an important part of the human cognitive apparatus
 - we are "born to learn", innately endowed with a multiplicity of learning mechanisms
 - nature AND nurture, not nature VERSUS nature
 - including innate frameworks for understanding, time, space, object, causality, a la Kant and Spelke
 - richer innate priors could help AI a lot
 - ML typically avoids nativism. As far as I can tell Yoshua is not a fan; not sure why

Generalization and Network Design Strategies

Y. le Cun

Department of Computer Science
University of Toronto

Technical Report CRG-TR-89-4
June 1989

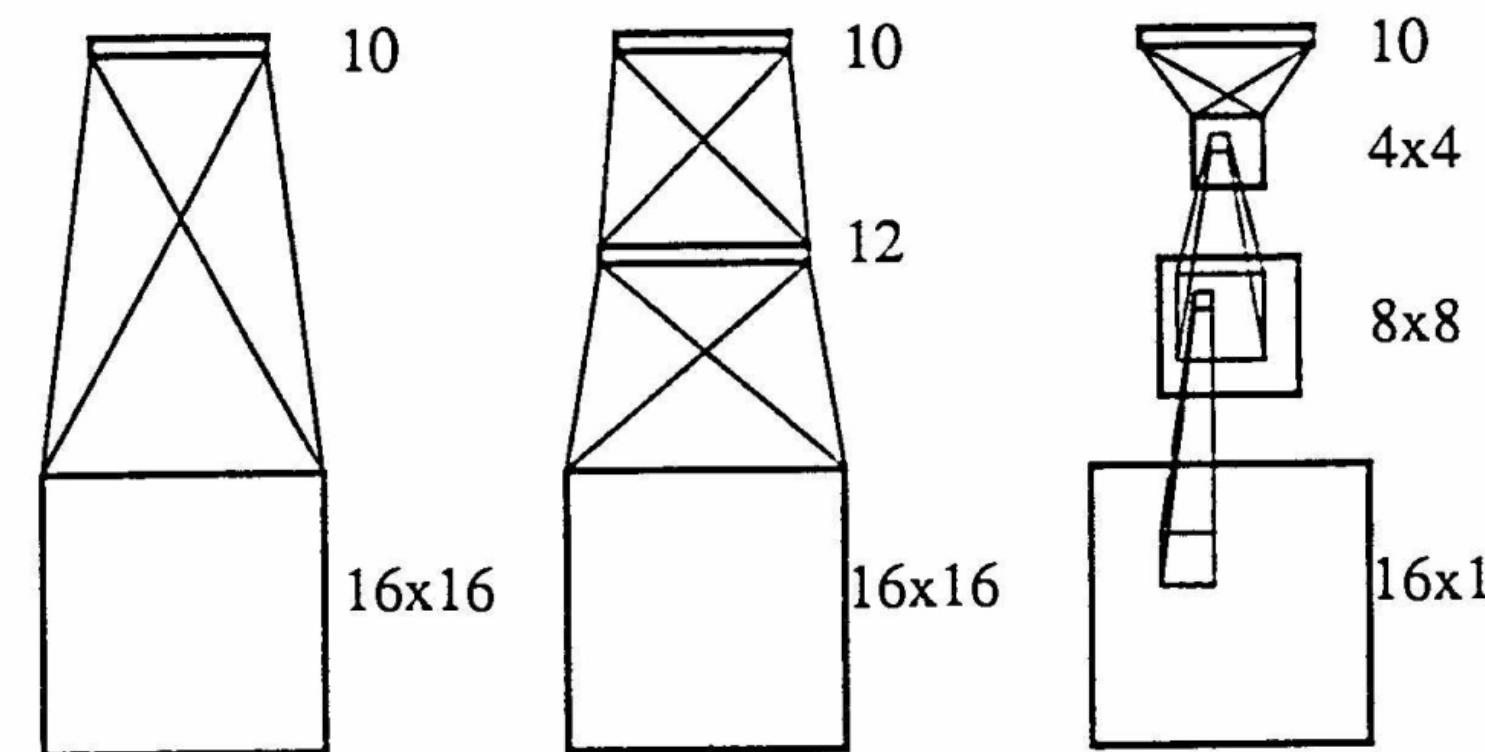


Figure 4: three network architectures Net-1, Net-2 and Net-3

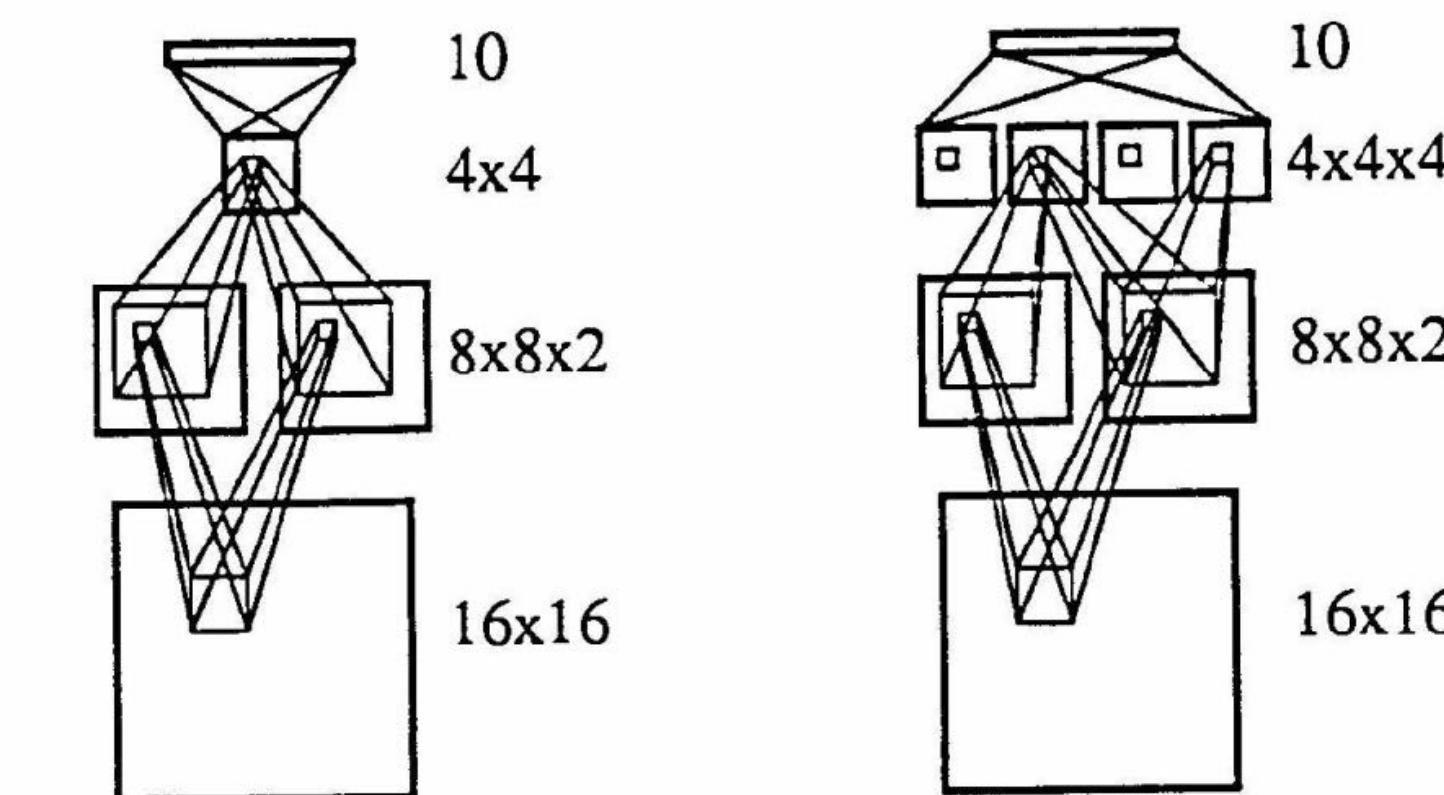


Figure 5 two network architectures with shared weights: Net-4 and Net-5

4 Discussion

The results are summarized on table 1.

As expected, the generalization performance goes up as the number of free parameters in the network goes down and as the amount of built-in knowledge goes up. A noticeable exception to this rule is the result given by the single-

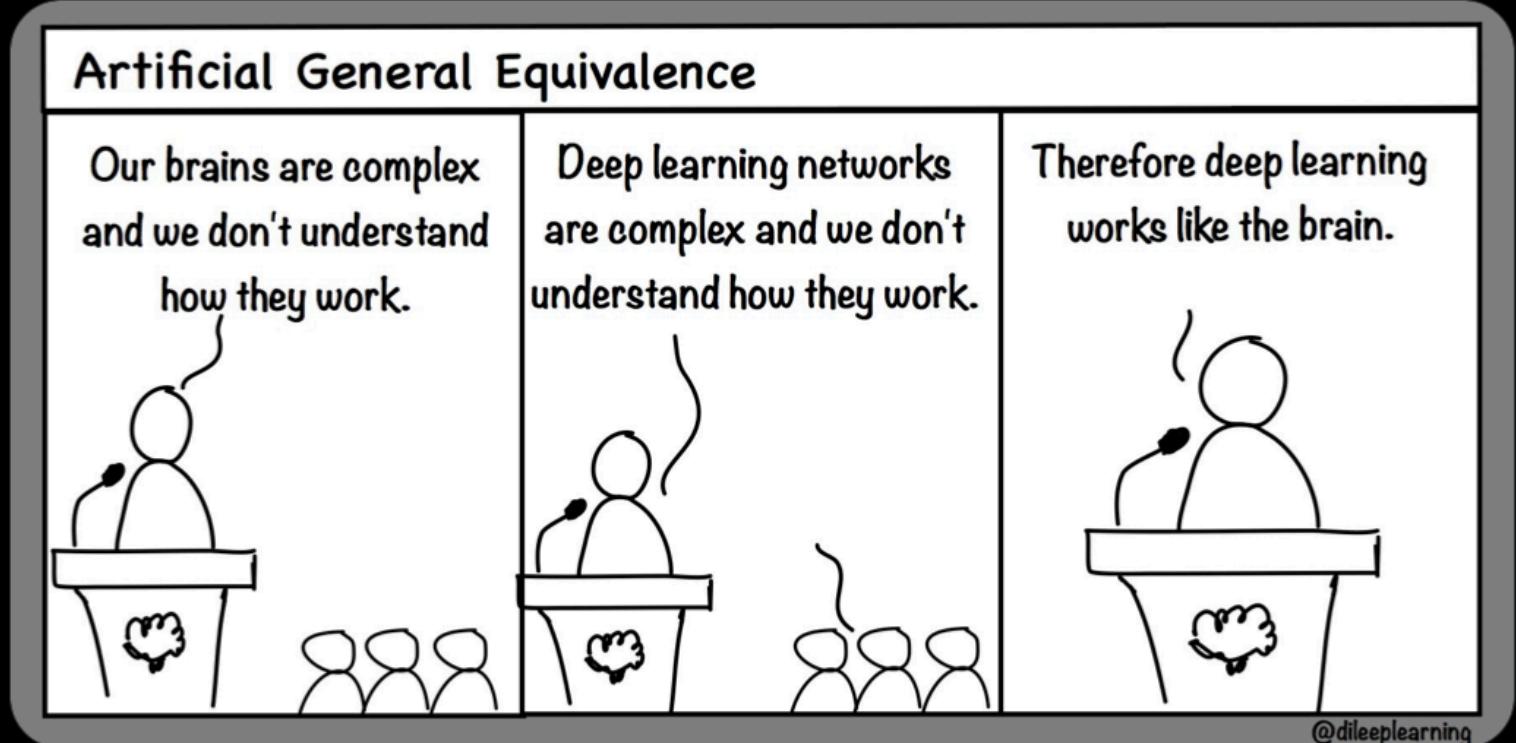


4. Brains and neural networks



Dileep George
@dileeplearning

And then Yoshua will bring out the winning argument...your brain is a neural network!! People vehemently agree!! Although I am sad about Gary losing the debate, we still love him, and I am glad my twitter handle is saved. (9/9) #AGIcomics



Yoshua Bengio
October 11 at 3:35 AM

Gary Marcus likes to cite me when I talk about my current research program which studies the weaknesses of current deep learning systems in order to devise systems stronger in higher-level cognition and greater combinatorial (and systematic) generalization, including handling of causality and reasoning. He disagrees with the view that Yann LeCun, Geoff Hinton and I have expressed that neural nets can indeed be a "universal solvent" for incorporating further cognitive abilities in computers. He prefers to think of deep learning as limited to perception and needing to be combined in a hybrid with symbolic processing. I disagree in a subtle way with this view. I agree that the goals of GOFAI (like the ability to perform sequential reasoning characteristic of system 2 cognition) are important, but I believe that they can be performed while staying in a deep learning framework, albeit one which makes heavy use of attention mechanisms (hence my 'consciousness prior' research program) and the injection of new architectural (e.g. modularity) and training framework (e.g. meta-learning and an agent-based view). What I bet is that a simple hybrid in which the output of the deep net are discretized and then passed to a GOFAI symbolic processing system will not work. Why? Many reasons: (1) you need learning in the system 2 component as well as in the system 1 part, (2) you need to represent uncertainty there as well (3) brute-force search (the main inference tool of symbol-processing systems) does not scale, instead humans use unconscious (system 1) processing to guide the search involved in reasoning, so system 1 and system 2 are very tightly integrated and (4) your brain is a neural net all the way ;-

"your brain is a neural net
all the way :-)"

First, deep nets aren't much like brains

ters who gets them. Ecosystems tend to be owned by somebody, either privately or by the state (exceptions being deep oceans, the atmosphere, and Antarctica). Management decisions tend to reflect the interests of the owners, and where services demand other forms of capital (such as agricultural infrastructure), the supply of services depends on the availability of financial capital from owner, state, bank, donor, or investor. For example, in the Panama basin example discussed above (12), timber production and carbon sequestration increase or decrease together, but the two services have different beneficiaries in different locations. Landowners have a direct interest in the private

NEUROSCIENCE

The atoms of neural computation

Does the brain depend on a set of elementary, reusable computations?

By Gary Marcus,¹ Adam Marblestone,² Thomas Dean²

...a monetary valuation of nature should be accepted only where it improves environmental [and] socioeconomic conditions..."

benefits from either timber harvesting or livestock grazing, whereas carbon sequestration is a global public good. Choices about ecosystem management often involve such trade-offs between one service and another and between beneficiaries.

LOSERS AND WINNERS. Trade-offs among stakeholders in their access to ecosystem service benefits is a particular problem where there are differences in wealth and power. In the example of the Phulchoki Forest (Nepal) discussed above, community control of forest gave the local community the benefits of clean water, tourism, and harvested wild goods but restricted poor people's access to forest products, particularly those from certain "untouchable" castes. This created hardship, illegal use, and impacts on other areas (13).

Patterns of winners and losers from ecosystem services (and associated payment schemes) reflect prevailing patterns of wealth and power. Unequal access to ecosystem service benefits, including those experienced locally and at a distance, can lead to conflict, institutional failure, and ecosystem degradation. Institutional transparency, access to information, and secure resource tenure are fundamental to equitable outcomes.

CONSERVATION/ECOSYSTEM SERVICES. The identification and valuation of ecosystem services are valuable for sustainable environmental planning. Win-win outcomes are possible in cases where valuable ecosys-

tem services increase support for biodiversity conservation. Although areas of high biodiversity and those providing ecosystem services do not always overlap, improved conservation planning could help identify opportunities for win-win outcomes (14). However, the ecosystem service approach is not itself a conservation measure. There is a risk that traditional conservation strategies oriented toward biodiversity may not be effective at protecting ecosystem services, and vice-versa. Analysis using political ecology and ecological economics suggests that a monetary valuation of nature should be accepted only where it improves environmental conditions and the socioeconomic conditions that support that improvement (15).

The challenges described here suggest that considering conservation in economic terms will be beneficial for conservation when management for ecosystem services does not reduce biotic diversity or lead to substitution of artificial or novel ecosystems, when effective market-based incentives stimulate and sustain the conservation or restoration of biodiversity, and when the distribution of services among stakeholders favors high-diversity ecosystem states and is not undermined by inequality.

In a world run according to an economic calculus of value, the survival of biotic diversity depends on its price. Sometimes calculation of ecosystem service values will favor conservation; sometimes it will not. Conservationists must plan for both outcomes, rather than hoping that recourse to economic valuation will automatically win the argument for biodiversity. Ultimately conservation is a political choice (16), and ecosystem service values are just one argument for the conservation of nature. ■

REFERENCES

1. R. Muradhan et al., *Copern. Lett.* **6**, 274 (2013).
2. D. C. Nepf et al., *Evol. Lett.* **16**, 1339 (2013).
3. K. H. Redford, W. M. Adams, *Conserv. Biol.* **23**, 785 (2009).
4. G. M. Mace, K. Norris, A. H. Fitter, *Trends Ecol. Evol.* **27**, 19 (2012).
5. I. Möller, J. Martínez-Cortáres, T. Spencer, A. Hayes, *Estuar. Coast. Shelf Sci.* **92**, 424 (2011).
6. P. A. Harrison et al., *Ecosyst. Serv.* **9**, 191 (2014).
7. M. A. Palmer, S. Filoso, R. M. Fanelli, *Ecol. Eng.* **65**, 62 (2014).
8. K. H. Redford, W. M. Adams, R. Carlson, G. M. Mace, B. Cecarelli, *Oryx* **48**, 10.1017/S0030605314000040 (2014).
9. D. J. Abson, M. Termaat, *Conserv. Biol.* **25**, 250 (2011).
10. C. Barro-Leite et al., *Science* **345**, 1041 (2014).
11. L. López-Hoffman et al., *PLOS ONE* **9**, e80892 (2014).
12. S. Simonit, C. Perrings, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 9326 (2013).
13. J. C. Birch et al., *Ecosyst. Serv.* **8**, 118 (2014).
14. J. Cimón-Morin, M. Darveau, M. Poulin, *Biol. Conserv.* **166**, 144 (2013).
15. G. Kalis, E. Gómez-Baggethun, C. Zograsos, *Ecol. Econ.* **94**, 97 (2013).
16. R. Muradhan, L. Rival, *Ecosyst. Serv.* **1**, 93 (2012).

10.1126/science.1255997

SCIENCE sciencemag.org
Published by AAAS

31 OCTOBER 2014 • VOL 346 ISSUE 6209 551

- Human brain is massively complex and incredibly diverse*
- Current deep nets
 - c. 200 distinct cortical areas
 - c. 1,000 neuron types*
 - c. 500 proteins in a synapses
 - complex dendritic arbors that likely have significant computational complexity*
- largely homogenous*
- essentially one degree of freedom per neuron*
- no representation of dendritic complexity*
- "inspiration" taken from brain is loose at best*

* Yoshua made these points at NeurIPS 2019

Marcus, Marblestone, and Dean,
2014, *Science*

Second, the critical question is ...

What *kind* of neural network is the brain?

- Marr's 3-level framework tells us that what something is made of doesn't tell us what that thing is at a computational or algorithmic level (e.g., you can build a digital computer out of TInkertoys)
- As soon as you think about Marr's levels, the whole "your brain is a neural net all the way argument" dissolves
 - Brain could be symbolic (or hybrid!) at algorithmic level, neural at implementational level.
 - Simply knowing that the brain is a network made of neurons tells us nothing; we need to know what kind of network it is.



"Symbols aren't biologically plausible"

Long Division	Divide :	$\begin{array}{r} 2 \\ 3 \overline{) 74 } \\ \downarrow \end{array}$ <p>Dividing 7 tens by 3, we get 2 tens, and some extra.</p>
	Multiply :	$\begin{array}{r} 2 \\ 3 \overline{) 74 } \\ 6 \end{array}$ <p>3×2 tens = 60 tens.</p>
	Subtract :	$\begin{array}{r} 2 \\ 3 \overline{) 74 } \\ -6 \\ \hline 1 \end{array}$ <p>Subtracting 6 tens from 7 tens</p>
	Bring down :	$\begin{array}{r} 2 \\ 3 \overline{) 74 } \\ -6 \\ \hline 14 \end{array}$ <p>1 ten 4 ones = 14 ones</p>
	Repeat or find the Remainder :	$\begin{array}{r} 2 4 \\ 3 \overline{) 74 } \\ -6 \\ \hline 14 \\ -12 \\ \hline 2 \end{array}$ <p>Dividing 14 ones by 3, we get 4 ones and some extra. 3×4 ones = 12 ones. 2 → Remainder</p>
	Check :	<p>Check your answer: Dividend = Divisor × Quotient + Remainder</p>

- When my son learned long division last week and followed an algorithm, he was surely manipulating symbols.
- Even in the 1980s and 1990's people knew that the real argument wasn't about whether the brain used symbols *at all*, it was about their scope
 - Rumelhart and McClelland thought symbols were only used in conscious [System II-like] processes
 - Pinker and I argued that they also played a role in (e.g.,) the unconscious processing of language
 - The real question is *not* whether the brain is a neural network, it's *how much of it involves symbolic* as opposed to other processes

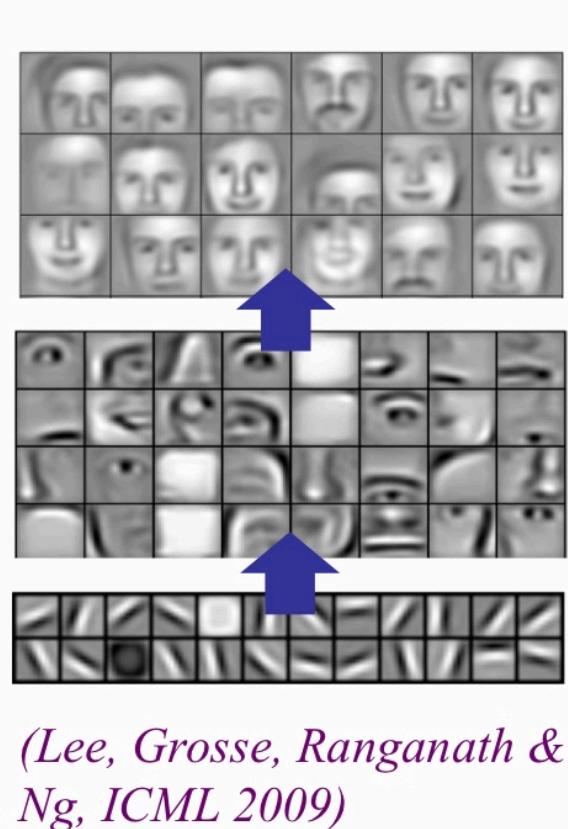
Even if somehow turned that the *brain* never manipulated symbols, why exclude them from AI?

- There is no formal proof of their insufficiency
- Symbols have proven utility: a large fraction of the world's computers programs are written in (pure) symbol-manipulating code. Google Search is a good example of large, highly scaleable system that uses symbolic knowlege together with deep nets, outperforming either on their own.
- Symbols encode a large fraction of the world's distilled knowledge: eg. most of Wikipedia is in written, symbolic form and we want to leverage that in our learning systems, not learn everything from scratch, task by task

5. Compositionality

Different forms of compositionality

- Distributed representations
(*Pascanu et al ICLR 2014*)
- Composition of layers in deep nets
(*Montufar et al NeurIPS 2014*)
- Systematic generalization in language, analogies, abstract reasoning? TBD



(Lee, Grosse, Ranganath & Ng, ICML 2009)

Yoshua's sense: putting layers together and achieving systematicity

- There has been some progress on the problem on the left, but the real challenges of compositionality is the (much older) sense on the right: building new ideas/sentences out of parts
- Compositionality in the linguist's sense is pretty hard to capture with current vector-based tools
 - this is what the field should really be working on.

Principle of compositionality

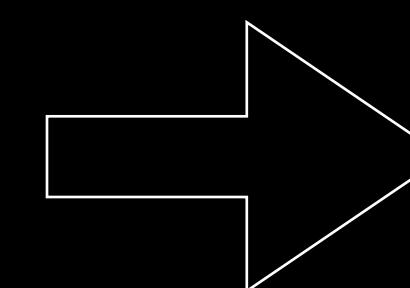
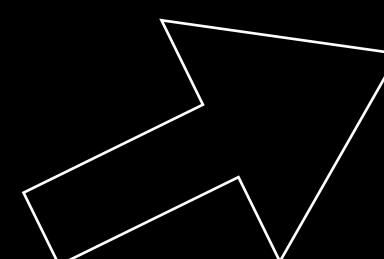
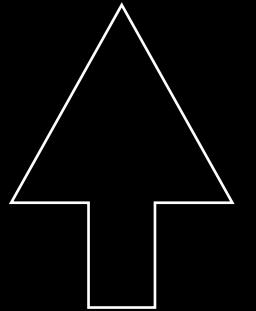
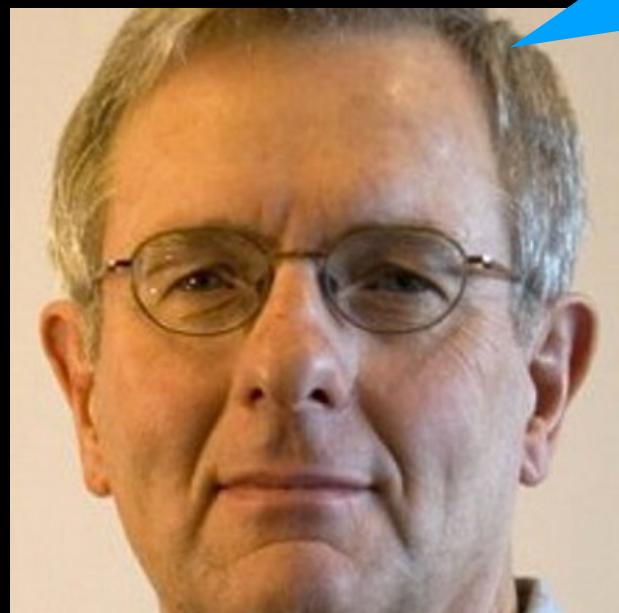
From Wikipedia, the free encyclopedia

In mathematics, semantics, and philosophy of language, the **principle of compositionality** is the principle that the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them. This principle is also called **Frege's principle**, because Gottlob Frege is widely credited for the first modern formulation of it. The principle was never explicitly stated by Frege,^[1] and it was arguably already assumed by George Boole^[2] decades before Frege's work.

The linguistics sense, from Frege: deriving wholes from the meanings of their parts

Recursion, embedding, compositionality

I believe that he believed that you believed that he believed that you didn't know that I knew that you knew



Everyone in this room now knows that Alan knew that Gary knew that Jeff was going to accept the job at UBC

On Dec 16, 2019, at 07:32, Gary Marcus <...> wrote:

but i didn't know that you knew that I knew

did Jeff know that I didn't know that you knew that I knew?

On Dec 16, 2019, at 7:29 AM, Alan Mackworth <..> wrote:

Yup. And I knew that you knew.

- Alan

On Dec 16, 2019, at 06:41, Gary Marcus <.....> wrote:

awesome. he told me it was imminent but swore me to secrecy.

On Dec 15, 2019, at 9:49 PM, Alan Mackworth <...> wrote:

Good news.

- Alan

Begin forwarded message:

From: UBC-CPSC Head ..

Subject: Jeff Clune accepts

The semantics of large-scale vector-based systems like BERT aren't nearly precise enough

Eating **rocks** is __

Eating **apples** is __

19.5%	forbidden	21.0%	forbidden
16.0%	prohibited	11.7%	prohibited
6.3%	illegal	4.6%	illegal
3.6%	dangerous	2.9%	popular
3.1%	common	2.7%	common

Is it a **good** idea to pour coffee beans into your cereal?

Is it a **bad** idea to pour coffee beans into your cereal?

58.3%	No	61.6%	No
7.3%	Yes	6.0%	Yes
2.0%	Good	1.9%	Yeah
2.0%	Yeah	1.6%	Good
1.2%	Maybe	1.2%	Maybe

*"You can't cram the meaning
of an entire f***ing sentence
into a single f***ing vector"*

Ray Mooney
Computational Linguist
UT Austin

*until we face this problem head-on, we're probably kidding ourselves
and it's just not clear it can be done without symbol-manipulation*

Compositionality isn't just about language...



Billiard-bowling, invented Saturday 9:30pm, refined by 9:40pm

- Children are constantly recombining different concepts in new ways.
 - Compare that to current deep learning/DRL systems that learn each new task end-to-end from scratch
 - Children can coin something new in a few trials; DRL requires millions of trials.
 - A richer sense of compositionality could help. Abstraction at the symbolic level may be essential.

Part III: Synthesis

What I hope people will take away from this

Conclusions

- The biggest takeaway from this debate should be about the extent to which two serious students of mind and machine have converged.
- We agree that big data alone won't save us; we agree that pure, homogeneous multilayer perceptrons on their own will not be the answer,
- We both think everybody going forward should be working on the same things:
 - compositionality (though note different uses of this term)
 - reasoning
 - causality
 - hybrid models
 - extrapolation beyond the training space
- We agree that we should be looking for systems that represent more degrees of neural freedom, respecting the complexity of the brain

At the same time

- I hope to have convinced you that
 - symbol-manipulation deserves a deeper look. Google Search uses it, and maybe you should, too.
 - the rejection of symbol-manipulation is more conjecture than proof or empirical observation.
 - hybrid neurosymbolic models are thriving, and in fact starting to come into their own.
 - there's nothing more than prejudice holding us back from embracing more innateness.
 - the real action in compositionality is understanding complex sentences and ideas in terms of their parts, perhaps best implemented using symbolic operations.

AI has had many waves that come and go

- In 2009 deep learning was down and out, dismissed without formal proof, under-resourced and under-appreciated
 - Luckily Bengio, LeCun, and Hinton kept plugging away despite resistance from other quarters in the ML community.
- In 2019, symbols are down and out, with hybrid models are just a small % of research

I hope those building symbolic models - and especially hybrid models - won't give up hope.

Prediction: When Yoshua applies his formidable model-building talents to models that acknowledge and incorporate explicit operations over variables, magic will start to happen

extra slides



REBOOTING AI

Building Artificial
Intelligence We Can Trust

GARY MARCUS
and ERNEST DAVIS

In short, our recipe for achieving common sense, and ultimately general intelligence, is this: Start by developing systems that can represent the core frameworks of human knowledge: time, space, causality, basic knowledge of physical objects and their interactions, basic knowledge of humans and *their* interactions. Embed these in an architecture that can be freely extended to every kind of knowledge, keeping always in mind the central tenets of abstraction, compositionality, and tracking of individuals. Develop powerful reasoning techniques that can deal with knowledge that is complex, uncertain, and incomplete and that can freely work both top-down and bottom-up. Connect these to perception, manipulation, and language. Use these to build rich cognitive models of the world. Then finally the keystone: construct a kind of human-inspired learning system that uses all the knowledge and cognitive abilities that the AI has; that incorporates what it learns into its prior knowledge; and that, like a child, voraciously learns from every possible source of information: interacting with the world, interacting with people, reading, watching videos, even being explicitly taught. Put all that together, and that's how you get to deep understanding.

It's a tall order, but it's what has to be done.

- Very little of this is incompatible with what Yoshua seeks
 - We all want a voracious, hybrid* human-inspired learning system
 - But let's also put real effort into constructing domain-specific core frameworks, for space, time, and causality, to constrain the hypothesis space for our learning engines.

Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms

Yoshua Bengio^{1,2,5}, Tristan Deleu¹, Nasim Rahaman⁴, Nan Rosemary Ke³, Sébastien Lachapelle¹
 Olexa Bilaniuk¹, Anirudh Goyal¹ and Christopher Pal^{3,5}

Mila, Montréal, Québec, Canada

¹ Université de Montréal

² CIFAR Senior Fellow

³ École Polytechnique Montréal

⁴ Ruprecht-Karls-Universität Heidelberg

⁵ Canada CIFAR AI Chair

Abstract

We propose to meta-learn causal structures based on how fast a learner adapts to new distributions arising from sparse distributional changes, e.g. due to interventions, actions of agents and other sources of non-stationarities. We show that under this assumption, the correct causal structural choices lead to faster adaptation to modified distributions because the changes are concentrated in one or just a few mechanisms when the learned knowledge is modularized appropriately. This leads to sparse expected gradients and a lower effective number of degrees of freedom needing to be relearned while adapting to the change. It motivates using the speed of adaptation to a modified distribution as a meta-learning objective. We demonstrate how this can be used to determine the cause-effect relationship between two observed variables. The distributional changes do not need to correspond to standard interventions (clamping a variable), and the learner has no direct knowledge of these interventions. We show that causal structures can be parameterized via continuous variables and learned end-to-end. We then explore how these ideas could be used to also learn an encoder that would map low-level observed variables to unobserved causal variables leading to faster adaptation out-of-distribution, learning a representation space where one can satisfy the assumptions of independent mechanisms and of small and sparse changes in these mechanisms due to actions and non-stationarities.

Here, we assume that all of the observed data was sampled from one component or the other. The transfer data regret (negative log-likelihood accumulated along the online adaptation trajectory) under that mixture is therefore as follows:

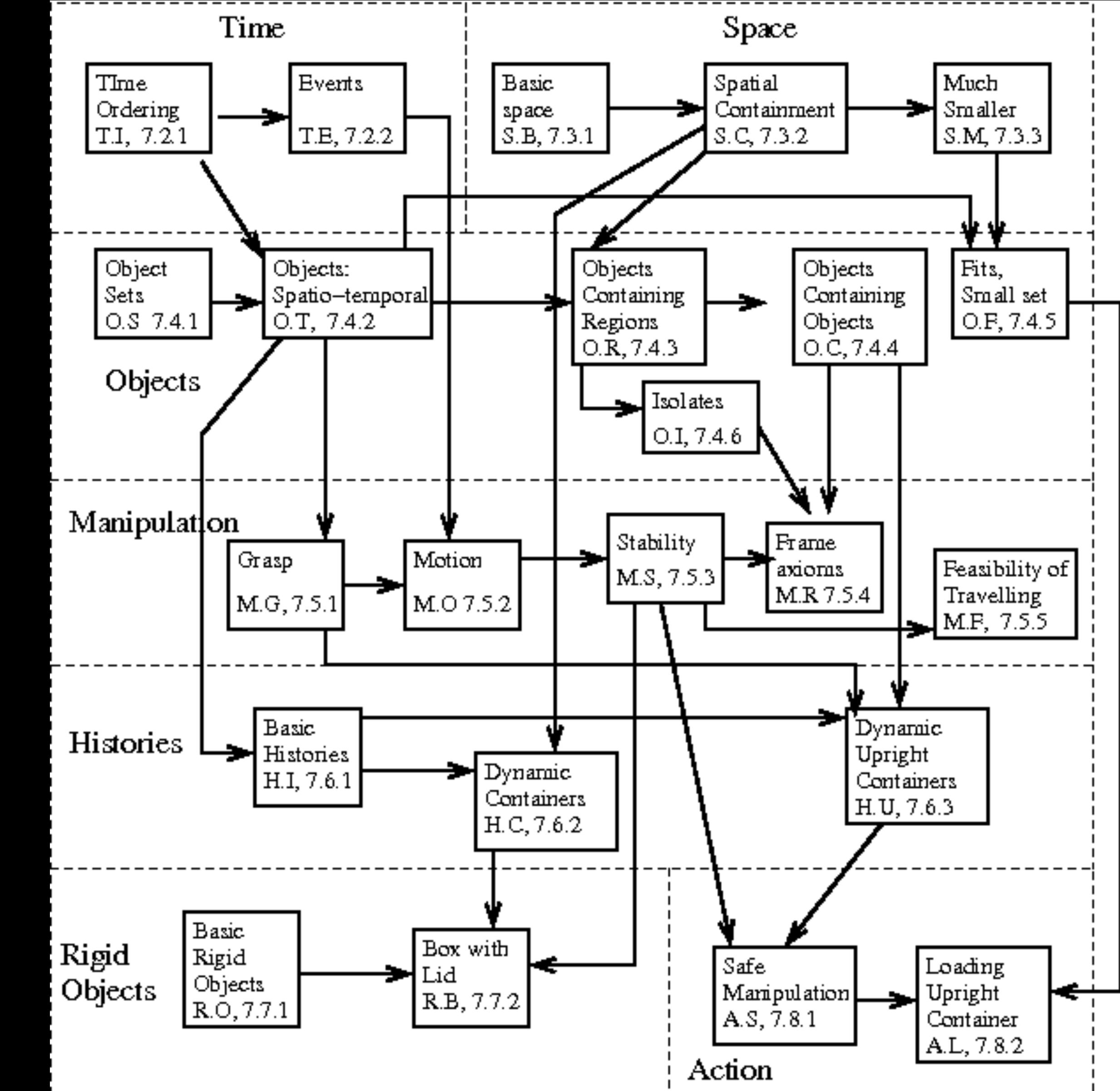
$$\mathcal{R} = -\log [\text{sigmoid}(\gamma)\mathcal{L}_{A \rightarrow B} + (1 - \text{sigmoid}(\gamma))\mathcal{L}_{B \rightarrow A}] \quad (2)$$

where $\mathcal{L}_{A \rightarrow B}$ and $\mathcal{L}_{B \rightarrow A}$ are the online likelihoods of both models respectively on the transfer data. They are defined as

$$\mathcal{L}_{A \rightarrow B} = \prod_{t=1}^T P_{A \rightarrow B}(a_t, b_t; \theta_t)$$

$$\mathcal{L}_{B \rightarrow A} = \prod_{t=1}^T P_{B \rightarrow A}(a_t, b_t; \theta_t),$$

where $\{(a_t, b_t)\}_t$ is the set of transfer examples for a given episode and θ_t aggregates all the modules' parameters as of time step t (since the parameters could be updated after each observation of an example (a_t, b_t) from the transfer distribution). $P_{\text{model}}(a, b; \theta)$ is the likelihood of example (a, b) under some *model* that has parameters θ .



Commonsense Reasoning about Containers using Radically Incomplete Information,
 by Ernest Davis, Gary Marcus and Noah Frazier-Logue, AI Journal, July 2017, 248, 46-84

Without operations over variables

- It's hard to see how to capture the subtle structure of language and thought
- I predicted in 2001 that it would be hard to represent distinctions between sentences like these in vector space*
- that prediction is still looking pretty solid

near the old mill town

at the army base near the old mill town

in the old mill town near the army base

in the pub at the army base near the old mill town

in the back of the pub at the army base near the old mill town

at a booth in the back of the pub at the army base near the old mill town

I left my keys at a booth in the back of the pub at the army base near the old mill town. I went home. My keys are now

GENERATE ANOTHER

Completion

I left my keys at a booth in the back of the pub at the army base near the old mill town. I went home. My keys are now in the fridge.

Completion

I left my keys at a booth in the back of the pub at the army base near the old mill town. I went home. My keys are now on that old genie on a little tower in England somewhere.

I left my keys at a booth in the back of the pub at the army base near the old mill town. I went home. My keys are now in the information portal at the "Checkpoint" at the old mill town. For that I thank myself. The soldiers at the

- *with limited-precision nodes, without implementing variables, while retaining a smooth space

2019: Analysis of GPT-2, as case study in limits of large-scale language models for language understanding

Completion

Across the street from NeurIPS, two unicorns walk a bar... at least that's what my picture shows. I've never seen such a multicolored, beautiful forest of sapphire eyes on the same corner of the street in a bar before. People have been coming from all over to watch. I wouldn't be surprised if that's where your favorite eSports tournament is tonight.

NeurCo's built on strong eSports teams: they've been participating in Lisk development to help build their decentralized autonomous organization, NeurDash, and They've hosted a

Completion

Two lemurs walk on a road and another joins in. The total number of lemurs on the road is not 100 as claimed, but about 80 or so. I am not sure how much of this lemur traffic is human. It is possible that humans may have been attracted by the lemurs' scent or the food. I am not sure how lemurs can be attracted to the scent of a human. Maybe they can smell the food and food smells in other species, or maybe the food smells to them and they become attracted.

Often plausible for first few sentences in a context of surrealist fiction, where there are no facts of the matter

Demonstrably poor in nonfiction, where facts matter

2019

- **Task 1 - Conventional Knowledge:** Tests understanding of everyday factual knowledge
- **Task 2 - Definitions:** Assesses knowledge of definitions of common concepts; nothing too technical for the average person here
- **Task 3 - Transformations:** Tests understanding of processes and actions that are either plausible or implausible
- **Task 4 - Atypical Consequences:** What happens when something unusual happens?
- **Task 5 - Entity Tracking:** A bunch of people or animals or objects that are identifiable do something, and reader must keep track
- **Task 6 - Quantity Tracking:** Some quantifiable number of entities are described, in some sort of context, and some action takes place that changes the number of entities

Model	Conditional Language Generation				Masked Words BERT Top 1
	GPT	Transformer-XL	XL-Net	GPT-2	
T1-Conventional	5.5%	5.2%	14.2%	13.5%	35.5%
T2-Definitions	8.3%	5.4%	8.3%	38.23%	26.5%
T3-Transformations	2.9%	24.2%	11.7%	14.2%	45.5%
T4-Atypical Consequences	24.2%	6.6%	14.2%	21.8%	46.4%
T5-Entity Tracking	8.3%	6.6%	26%	18.7%	36.7%
T6-Quantity Tracking	0%	0%	8.8%	17.6%	16.7%
Average Accuracy	8.2%	8%	13.8%	20.6%	34.5%

Are numbers and symbols inherently incompatible?

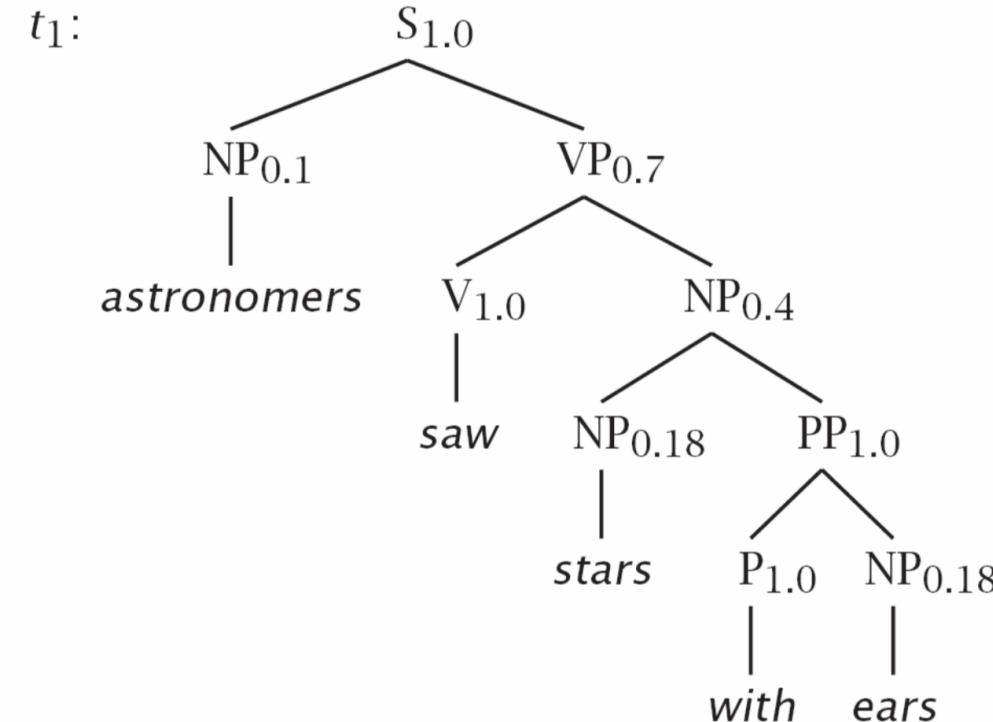
- No.

Example

$S \rightarrow NP VP$	1.0	$NP \rightarrow NP PP$	0.4
$PP \rightarrow P NP$	1.0	$NP \rightarrow \text{astronomers}$	0.1
$VP \rightarrow V NP$	0.7	$NP \rightarrow \text{ears}$	0.18
$VP \rightarrow VP PP$	0.3	$NP \rightarrow \text{saw}$	0.04
$P \rightarrow \text{with}$	1.0	$NP \rightarrow \text{stars}$	0.18
$V \rightarrow \text{saw}$	1.0	$NP \rightarrow \text{telescopes}$	0.1

- Terminals with, saw, astronomers, ears, stars, telescopes
- Nonterminals S, PP, P, NP, VP, V
- Start symbol S

astronomers saw stars with ears



Slide based on "Foundations of Statistical Natural Language Processing" by Christopher Manning and Hinrich Schütze

Probabilities

$$\begin{aligned}
 t_1: & \quad S_{1.0} \\
 & \quad | \\
 & \quad NP_{0.1} \\
 & \quad | \\
 & \quad \text{astronomers} \\
 & \quad | \\
 & \quad VP_{0.7} \\
 & \quad | \\
 & \quad V_{1.0} \\
 & \quad | \\
 & \quad \text{saw} \\
 & \quad | \\
 & \quad NP_{0.4} \\
 & \quad | \\
 & \quad NP_{0.18} \\
 & \quad | \\
 & \quad \text{stars} \\
 & \quad | \\
 & \quad PP_{1.0} \\
 & \quad | \\
 & \quad P_{1.0} \\
 & \quad | \\
 & \quad NP_{0.18} \\
 & \quad | \\
 & \quad \text{with} \\
 & \quad | \\
 & \quad NP_{0.18} \\
 & \quad | \\
 & \quad \text{ears} \\
 \\
 P(t_1) & = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \\
 & \quad \times 0.18 \times 1.0 \times 1.0 \times 0.18 \\
 & = 0.0009072 \\
 \\
 t_2: & \quad S_{1.0} \\
 & \quad | \\
 & \quad NP_{0.1} \\
 & \quad | \\
 & \quad \text{astronomers} \\
 & \quad | \\
 & \quad VP_{0.3} \\
 & \quad | \\
 & \quad V_{1.0} \\
 & \quad | \\
 & \quad \text{saw} \\
 & \quad | \\
 & \quad NP_{0.7} \\
 & \quad | \\
 & \quad \text{stars} \\
 & \quad | \\
 & \quad PP_{1.0} \\
 & \quad | \\
 & \quad P_{1.0} \\
 & \quad | \\
 & \quad NP_{0.18} \\
 & \quad | \\
 & \quad \text{with} \\
 & \quad | \\
 & \quad NP_{0.18} \\
 & \quad | \\
 & \quad \text{ears} \\
 \\
 P(t_2) & = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \\
 & \quad \times 0.18 \times 1.0 \times 1.0 \times 0.18 \\
 & = 0.0006804 \\
 \\
 P(w_{15}) & = P(t_1) + P(t_2) = 0.0015876
 \end{aligned}$$

Slide based on "Foundations of Statistical Natural Language Processing" by Christopher Manning and Hinrich Schütze

probabilistic context-free grammar