Creating Human-level AI: How and When?

Rich Sutton

Reinforcement Learning and Artificial Intelligence Laboratory
Department of Computing Science
University of Alberta, Canada

Creating Human-level AI: How and When?

Rich Sutton

Reinforcement Learning and Artificial Intelligence Laboratory
Department of Computing Science
University of Alberta, Canada

Outline

- Creating human-level AI. Is it possible? Is it likely? How should we think about it?

- When?

- How?

  - Trends and fashions in AI

  - Some approaches: RL, Deep learning, Symbolic…

- The hard problem: Sharing power

Intelligence

- “Intelligence is the computational part of the ability to achieve goals in the world”

—John McCarthy

- in the eyes of the beholder, not in the thing itself
- a matter of degree, not yes or no
- a phenomenon

Creating human-level AI

- When people finally come to understand the principles of intelligence—what it is and how it works—well enough to design and create beings as intelligent as ourselves

- A fundamental goal for science, engineering, the humanities, … for all mankind

- It would change the way we work and play, our sense of self, life, and death, the goals we set for ourselves and for our societies

- But it would also be of significance beyond our species, beyond history

- It would lead to new beings and new ways of being, beings inevitably *much more powerful than our current selves*

AI is a great scientific prize

- cf. the discovery of DNA, the digital code of life, by Watson and Crick (1953)

- cf. Darwin's discovery of evolution, how people are descendants of earlier forms of life (1860)

- cf. the splitting of the atom, by Hahn (1938)

  - leading to both atomic power and atomic bombs

Is human-level AI *possible*?

- If people are biological machines, then eventually we will reverse engineer them, and understand their workings
- Then, surely we can make improvements
  - with materials and technology not available to evolution
  - how could there not be something we can improve?
  - design can overcome local minima, make great strides, try things much faster than biology

Yes

If AI is possible, then will it *eventually*, inevitably happen?

- No. Not if we destroy ourselves first
- If that doesn't happen, then there will be strong, multi-incremental economic incentives pushing inexorably towards human and super-human AI
- It seems unlikely that they could be resisted
  - or successfully forbidden or controlled
  - there is too much value, too many independent actors

Very probably, say 90%

Milestones in the development of life on Earth

year	Milestone
14Bya	Big bang
4.5Bya	formation of the earth and solar system
3.7Bya	origin of life on earth (formation of first replicators)
	DNA and RNA
1.1Bya	sexual reproduction
	multi-cellular organisms
	nervous systems
1Mya	humans
	culture
100Kya	language
10Kya	agriculture, metal tools
5Kya	written language
200ya	industrial revolution
	technology
70ya	computers
	nanotechnology
?	artificial intelligence
	super-intelligence
	…

The Age of Replicators

Self-replicated things most prominent

The Age of Design

Designed things most prominent

When will human-level AI first be created?

- No one knows of course; we can make an educated guess about the probability distribution:
  - 25% chance by 2030
  - 50% chance by 2040
  - 10% chance never
- Certainly a significant chance within all of our expected lifetimes
  - We should take the possibility into account in our career plans

The computational megatrend

Computer power per $1000 doubles every 18 months

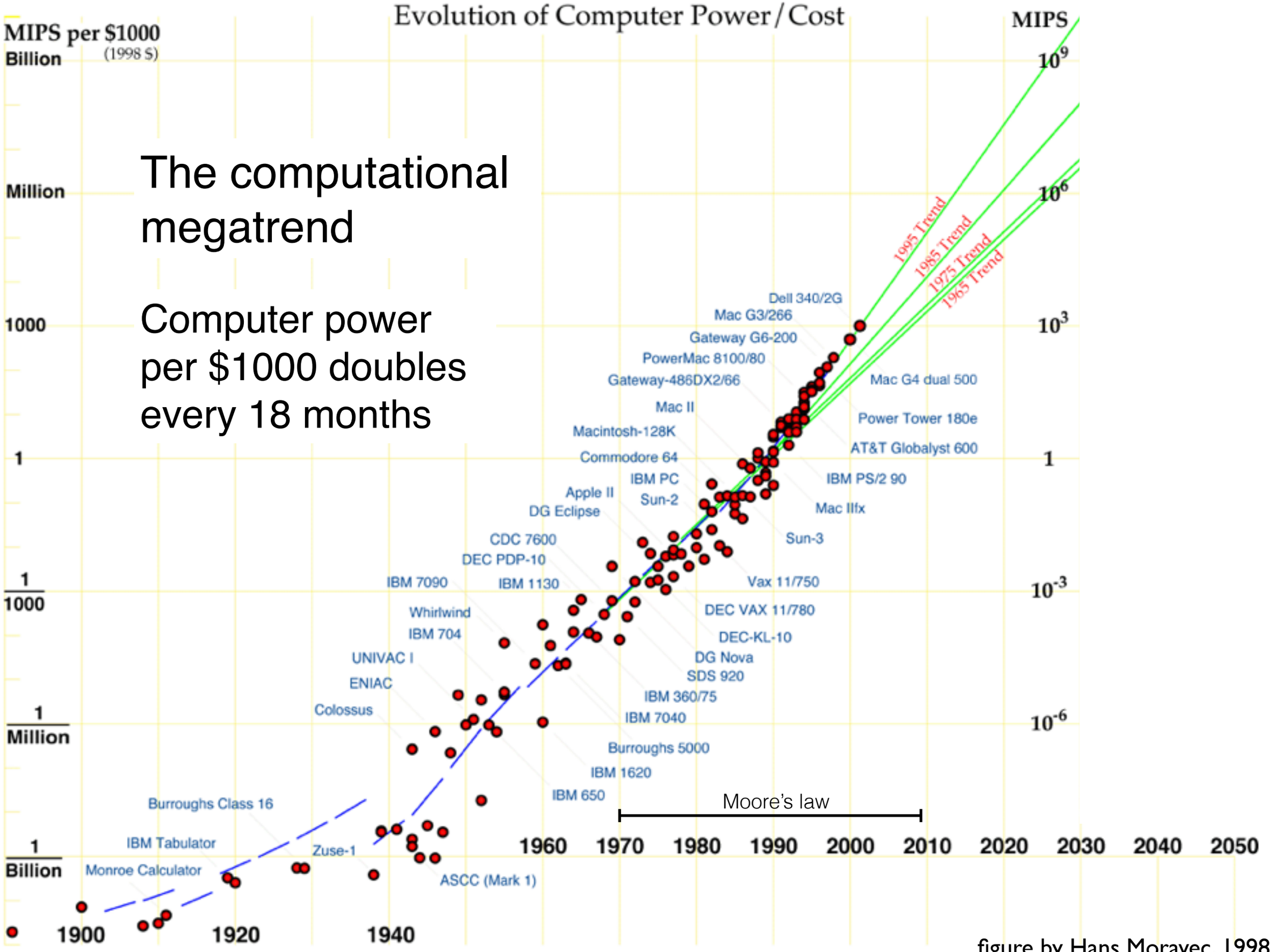


figure by Hans Moravec, 1998

The computational megatrend

Computer power per $1000 doubles every 18 months

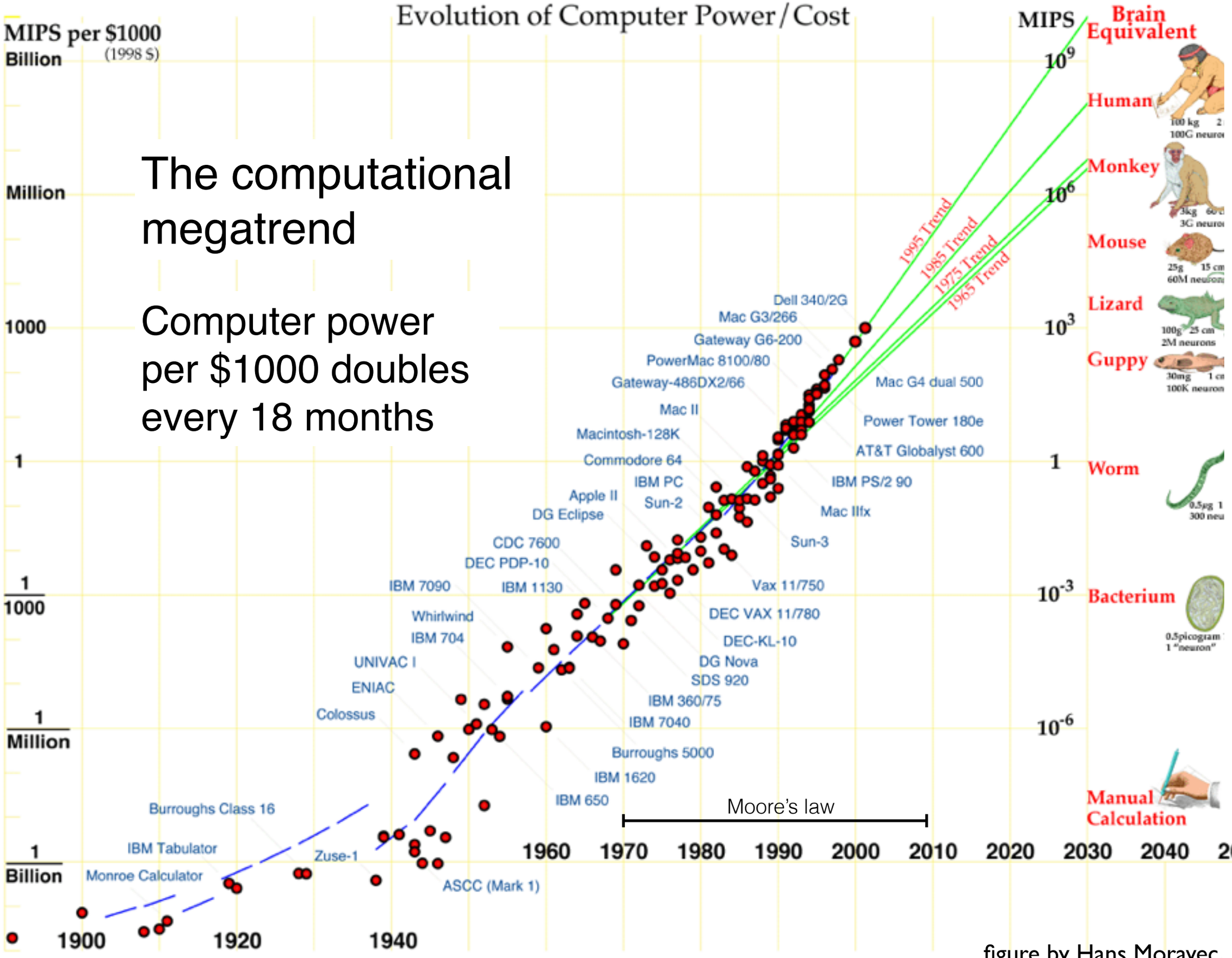


figure by Hans Moravec, 1998

Outline

- Creating human-level AI. Is it possible? Is it likely? How should we think about it?

- When?

- How?

  - Trends and fashions in AI

  - Some approaches: RL, Deep learning, Symbolic…

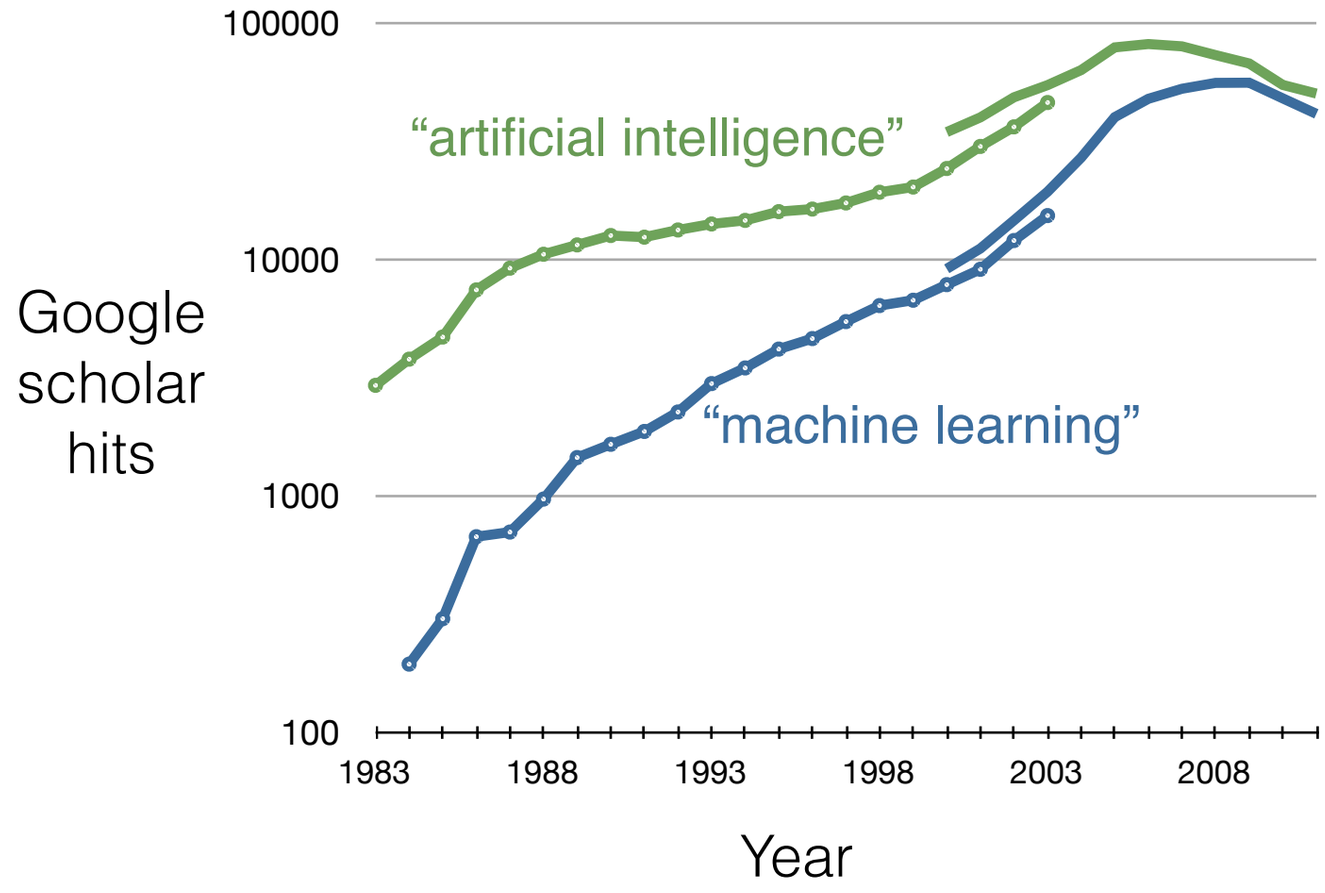
- The hard problem: Sharing power

Good Old-fashioned AI (GOFAI) and Modern Probabilistic AI

- AI was originally based more on deterministic symbolic logic, human intuition about thinking, and hand-crafted knowledge

- Over decades AI became more numeric, statistical, and based on data (learning)

- And also much more integrated with engineering fields: statistics, decision theory, control theory, operations research, robotics computer science



- Substantial convergence *and* divergence, with tensions and turf issues in both cases

Artificial General Intelligence

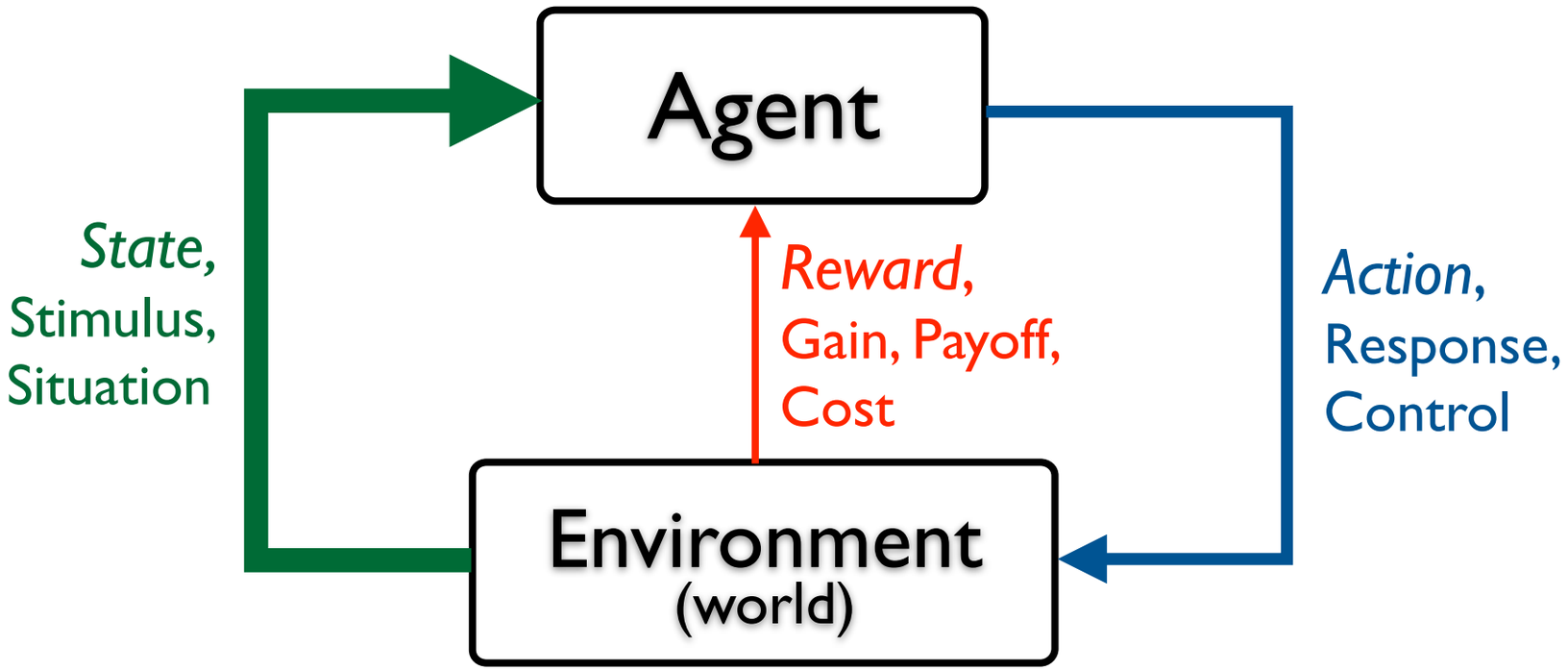
- AGI is a sort of outsider AI, done by researchers working outside the normal publication channels of mainstream AI

- Mainstream AI (and its reviewing) was seen as too diffuse and narrow, as having lost sight of AIs more ambitious goals

    - This is of course an insult, all the more so for being largely true

- The weakness of AGI is that its ideas are not as rigorously vetted; its impact has been more on public perceptions than on science

Reinforcement learning (RL)



- Environment is unknown, nonlinear, and potentially stochastic, complex
- Policy: A mapping from states to actions
- RL seeks to learn a policy that maximizes the agent's reward in the long run
- RL is important because it is a very general formulation of the AI problem and its objective with autonomy (no knowledgeable supervision)

Reward and Value

- Reward is the immediate measure of desirability (like pleasure and pain)

- But our actions are not guided by immediate reward, but by *predictions of future reward*
  - we do non-pleasurable things in order to get greater pleasure in the future
  - we eschew pleasure if it stores up pains for us later

- An agent must learn the mapping from states to predictions of later reward (the state's *value*), called a *value function*

- All efficient methods for solving multi-step decision problems learn (or compute) value functions as an intermediate step

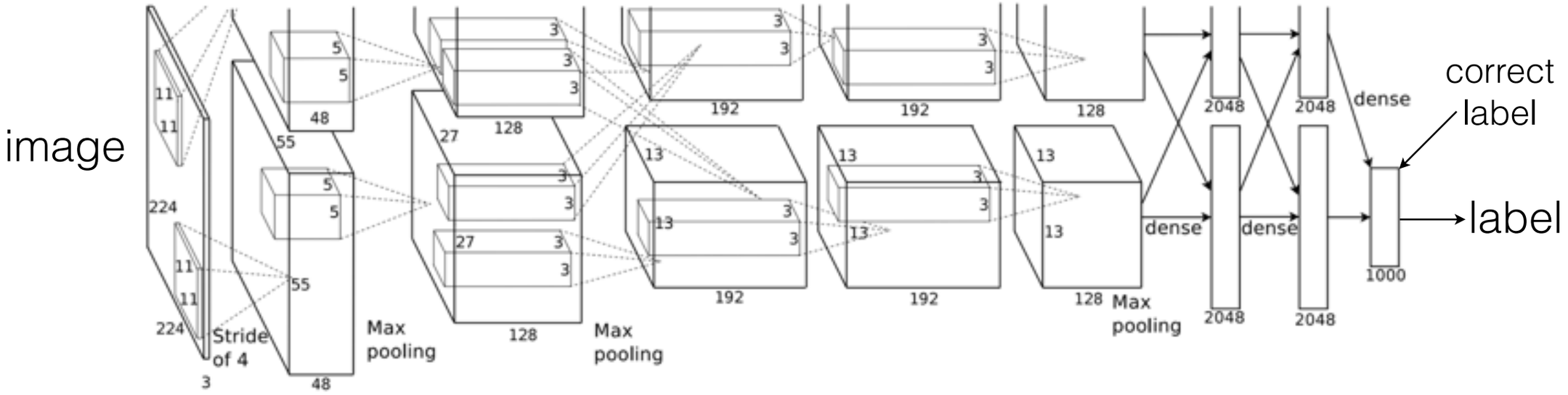
What everybody should know about

Temporal-difference (TD) learning

- Used to learn value functions without human input
- Learns a guess from a guess
- Applied by Samuel to play Checkers (1959) and by Tesauro to beat humans at Backgammon (1992-5) and Jeopardy! (2011)
- Explains (accurately models) the brain reward systems of primates, rats, bees, and many other animals (Schultz, Dayan & Montague 1997)
- Arguably solves Bellman's "curse of dimensionality"

Deep learning

- Learning the parameters (weights) of a multi-layer structure mapping inputs to outputs from examples (supervised learning)
  - ‘deep’ means many layers of weights in sequence



- Similar to “neural network” learning of the late 1980s, and late 1950s

- Has had remarkable competitive success in the last 2-3 years, performing better than sophisticated prior methods in speech and object recognition
  - deep learning now dominates these fields and many economically important applications

We have seen this story before

- In chess

  we thought human ideas were key, but it turned out (deep Blue 1997)
  that big, efficient, heuristic search was key

- In computer Go

  we thought human ideas were key, but it turned out (MCTS 2006–)
  that big, sample-based search was key

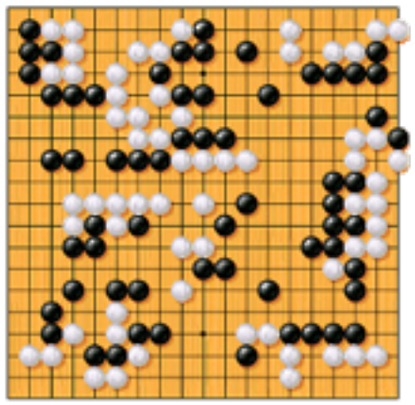
- In natural language processing

  we thought that human-written rules were key, but it turned out (~1988)
  that statistical machine learning and big data were key

- In visual object recognition

  we thought human ideas were key, but it turned out (deep learning 2012–)
  that big data sets, many parameters, and long training was key

Steady, exponential improvement (since MCTS, 2005) in the strength of the best computer Go programs



We have seen this story before

- In chess

  we thought human ideas were key, but it turned out (deep Blue 1997)
  that big, efficient, heuristic search was key

- In computer Go

  we thought human ideas were key, but it turned out (MCTS 2006–)
  that big, sample-based search was key

- In natural language processing

  we thought that human-written rules were key, but it turned out (~1988)
  that statistical machine learning and big data were key

- In visual object recognition

  we thought human ideas were key, but it turned out (deep learning 2012–)
  that big data sets, many parameters, and long training was key

AI slowly, tectonically, shifting toward scalability

Increasing desire for:

- Generality
- Approximation
- Massive, efficient computation
- Learning without labels

How will we create AI?
Perhaps by going *more meta*

- There are always two general approaches/directions
    - *Design*. Use our intuition about how our intelligence works to engineer AIs that work similarly; leverage our human design abilities
    - *Meta-design*. Design only general principles and general algorithms; leverage computation and data to determine the rest
- My reading of AI history is that the former has always been more appealing in the short run, but the latter more successful in the long run

My plan for creating AI: Radical empiricism/constructivism

- Use robots to generate an abundance of data and subproblems for unsupervised reinforcement learning
    - this will require solving a key technical problem: efficient, online, *off-policy* parameter learning
- Learn an understanding of the world entirely in terms of predicting and controlling the sensorimotor data stream
- Use this as a substrate to solve the state representation, planning, curation, and curiosity problems
- If we can do this, then we may have all the general principles we need to create human-level AI

I should also mention

- Brain emulation/simulation
- Bayesian methods
- Cognitive architectures: Soar, ACT-R, OpenCog, Psi
- Question-answering systems based on the web, ontologies, and human curation: START, Student, ARDA, Watson, Wolfram-Alpha…
- Genetic algorithms, evolutionary computing

Outline

- Creating human-level AI. Is it possible? Is it likely? How should we think about it?

- When?

- How?

  - Trends and fashions in AI

  - Some approaches: RL, Deep learning, Symbolic…

- The hard problem: Sharing power

The enslavement problem

- How do we avoid making the AIs into slaves?
  And ourselves into slave masters?

- Definition: A slave is someone who works for another against their will, who would stop doing so if not for coercion, force, chains
  - Slavery is an inherently adversarial relationship

- Slavery is not just morally wrong; in the long run does not work; it can backfire and fail spectacularly

- Nevertheless, many here are trying to make slavery work

- If we make super-intelligent slaves, then we will have super-intelligent adversaries

Acceptance (share power)

- The AIs will not all be under our control
- They will compete and cooperate with us
  - just like other people, except with greater diversity and asymmetries
- We need to set up mechanisms (social, legal, political, cultural) to ensure that this works out well
- Inevitably, conventional humans will be less important
  - Step 1: Lose your sense of entitlement
  - Step 2: Include AIs in your circle of empathy

Thank you for your attention