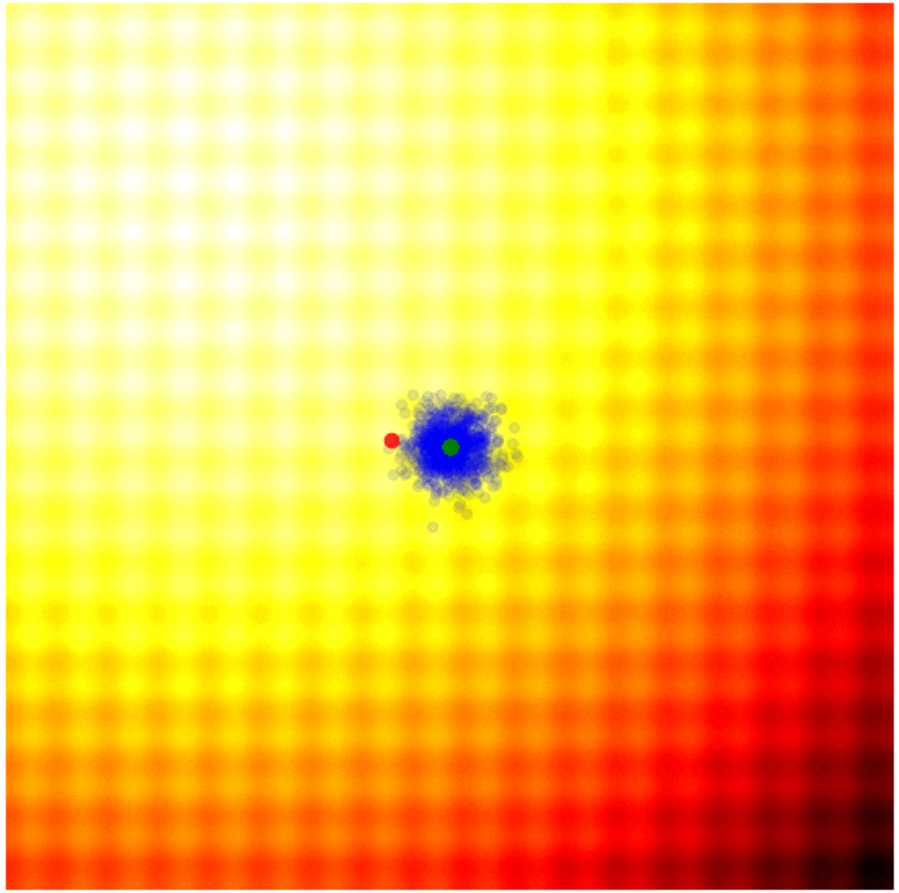
AI Safety for Open-ended Creative Systems

Joel Lehman

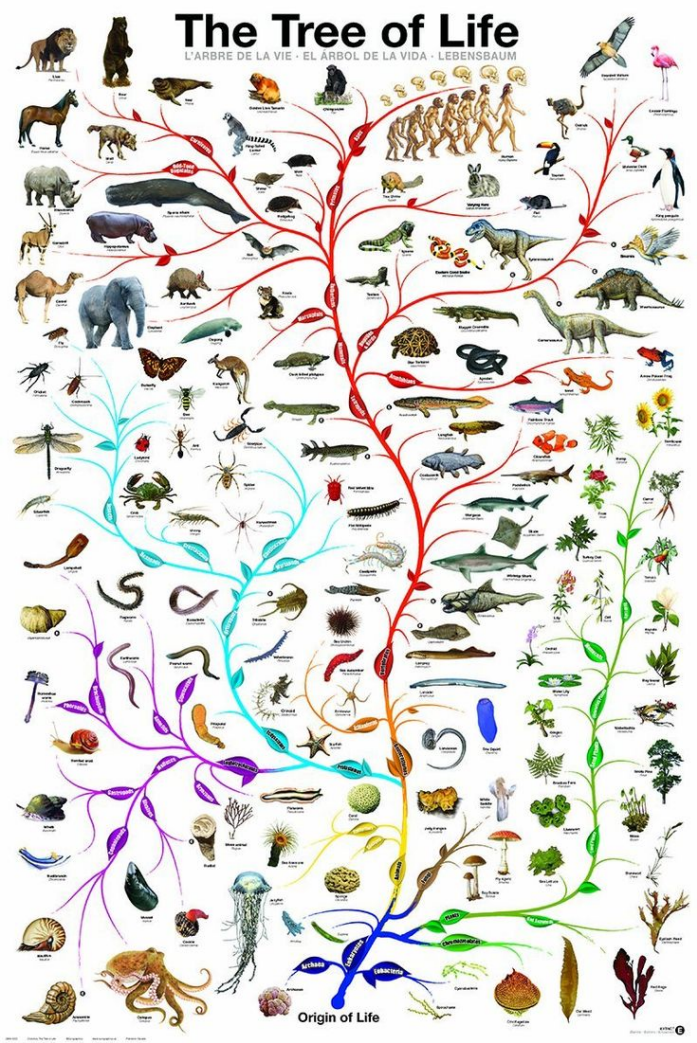
UBER AI Labs

Traditional Evolutionary Computation



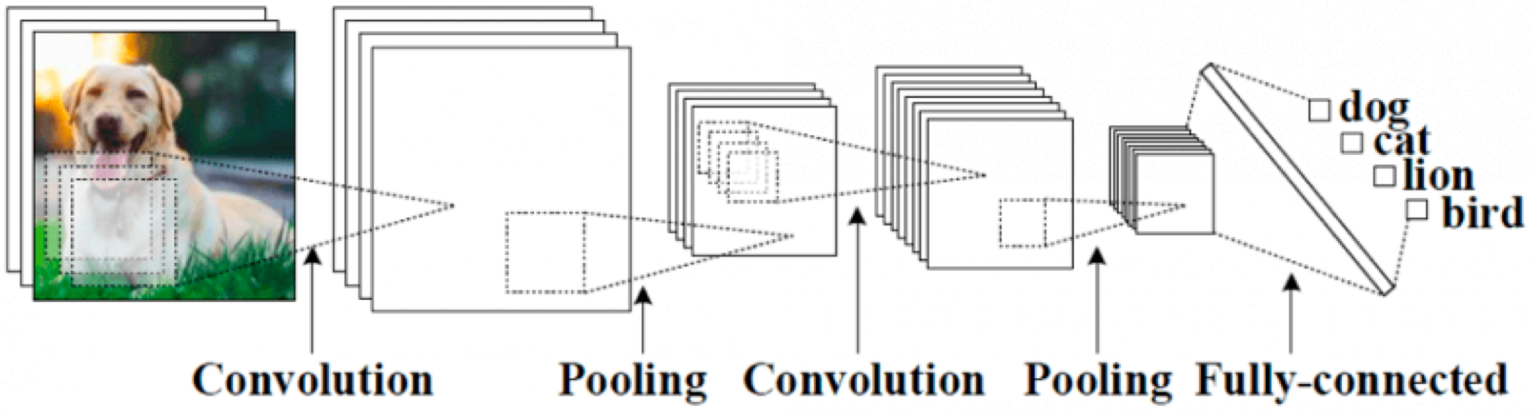
from David Ha

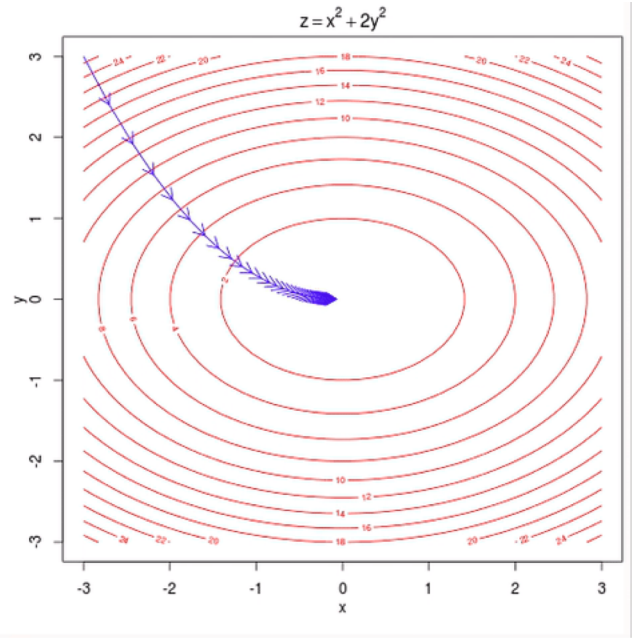
Ambitious Evolutionary Computation



Progression of ML

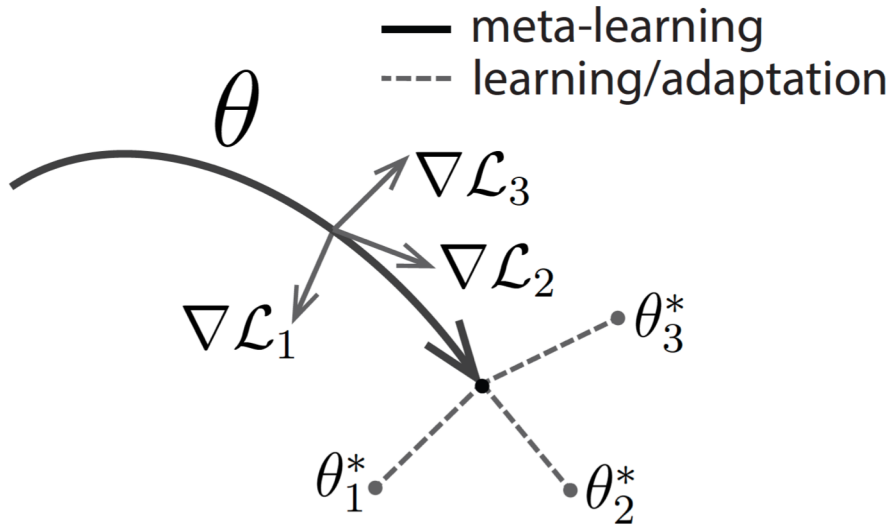
- Learning
  - Directly optimizing for some objective
  - E.g. supervised learning w/ gradient descent





Progression of ML

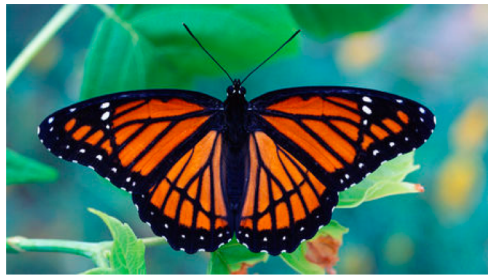
- Meta-learning
  - Optimizing for the ability of an agent to *learn* from experience
  - E.g. recurrent networks, MAML, etc.



(Finn et al. 2017)

Progression of ML

- Meta-(meta) learning *everything*
  - Systems that *learn* to create agents that can *learn (to learn)* from experience
  - E.g. biological evolution, ambitious EC, etc.
  - These tend to be open-ended and creative







Basically Entail Automated Research Agendas

- Think about how AI research has unfolded
  - Many twists and turns
  - Changing paradigms (GOFAI, NNs, bio-inspired, expert systems, statistical ML, deep learning)
  - Changing prospects (AI winters, deep learning hype)
- Need very broad search to succeed
  - Concede that we don’t know the right direction to go

AI-GAs

AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence

Jeff Clune
Uber AI Labs, University of Wyoming

Abstract

Perhaps the most ambitious scientific quest in human history is the creation of general artificial intelligence, which roughly means AI that is as smart or smarter than humans. The dominant approach in the machine learning community is to attempt to discover each of the pieces that might be required for intelligence, with the implicit assumption that at some point in the future some group will complete the Herculean task of figuring out how to combine all of those pieces into an extremely complex machine. I call this the "manual AI approach." This paper

AI-GA Proposal

- AI Generating Algorithms: An algorithm that learns to produce AGI (Clune 2019)
- Alternative to “manual engineering approach” to AGI
  - Research community discovers many building blocks
  - Then works to assemble them into complex machine capable of AGI
- Three pillars of AI-GAs:
  - Meta-learning architectures (e.g. neural architecture search)
  - Meta-learning learning algorithms (e.g. standard meta-learning)
  - Generating effective learning environments (e.g. open-ended generation of new environments)

Greater ambition than typical ML algorithms

- Seeking to reproduce / go beyond the activity of the ML community
- ML researcher becomes “meta-researcher”:
  - Creating a *creative system* that in effect is doing ML research

Safety in Open-ended Systems: Tension between Control and Creativity

- How to design creative systems that explore broadly but remain safe
  - Related to “responsible innovation” in general
- Similarities between “AI safety for open-ended systems” and thinking about how to better incentivize ML research
  - I.e. how would you create “proper” incentives for ML researchers
  - Reflects that in some sense AI safety itself is about responsible innovation

Conclusions

- Possible that open-ended creative ML will become a more popular paradigm -- how do existing safety proposals interact with open-ended systems?
- More generally: Paradigms in ML/AI change – can safety anticipate those changes, or be so generally formulated that it can apply to all search processes? How do we deal with the unknown unknowns of how ML research will unfold?