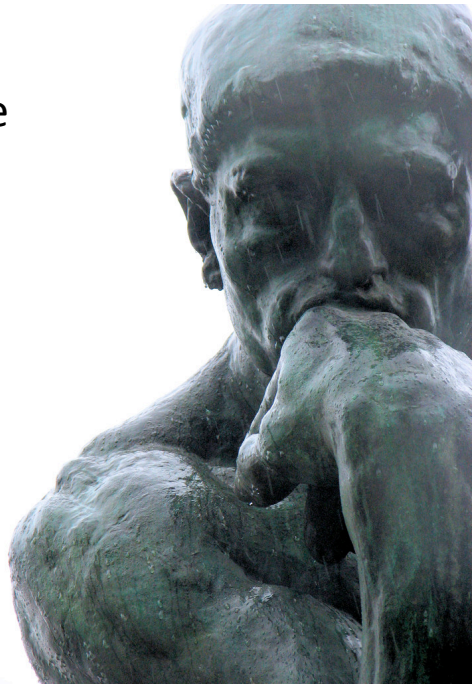
On the emergence of the self through survival of the fittest

Souvik Das
April 9, 2015

The problem of addressing the problem

The study of consciousness is *notorious* for vague definitions. We all sense the mystery but are not sure what the useful questions to probe the fundamentals are. Here,

- I develop my own **definitions**. Some convergence with accepted jargon and with layman English.
- I formulate what I think are the fundamental **questions** in terms of these definitions.
- I try to **answer** one of the questions explicitly with a computer program that demonstrates proof of concept of what a full answer might look like.



The field is also *notorious* for losing people to new age voodoo thinking, getting stuck on "philosophical traditions", misinterpretations of quantum mechanics and Deepak Chopra. My intellectual principles in trying to break new ground:

- There exists natural explanations for all things accessible by observation and reason.
- Rely on reasonable induction to some degree if inference or deduction is not possible.
- Abstract reasoning into the core of physics may be warranted, but if not, don't go there
- The Feynman principle: "*What I cannot make, I do not understand*."

Definitions and concepts

- **Consciousness**: The broadest word to describe the phenomenon. Something we all have but cannot be sure other organizations of matter like frog or plant have. Reasonable assumption: animal life forms experience this on a continuum.
- **Self**: The central point of observation, the sense of **I**, the **ego**, that which is not the rest of the universe, the local **identity**.
- **Uniqueness of self**. **Continuity of self**. Both rely on memory. Imagine memory -> 0.
- **Experience**: The subjective, deeply private, "taste" of what is being perceived. **Qualia**.
- **Cognition**: The processing of sensory information, from low to high levels. High level processing is called **intelligence**.
- **Agency**: The sense of **volition**, **free will**, the ability to make a decision and move physical objects independent of history. Automatons don't have agency.

Questions

- **What is the self? How does it work? How did it come about?**

Thesis

- The self is a pattern of information, a configuration of neural network (NN-ware) in the brain of most animals that delineates "I" from "not I". (Can be damaged with other faculties more or less intact.)
- This NN-ware emerged as a survival mechanism through natural selection in animals with nervous systems. ← ***Thesis of talk***
- Conjecture: Higher primates like us have a recursive awareness of self that occurs due to layering of new NN-ware layers over old ones through an arms race for survival strategies in evolution.

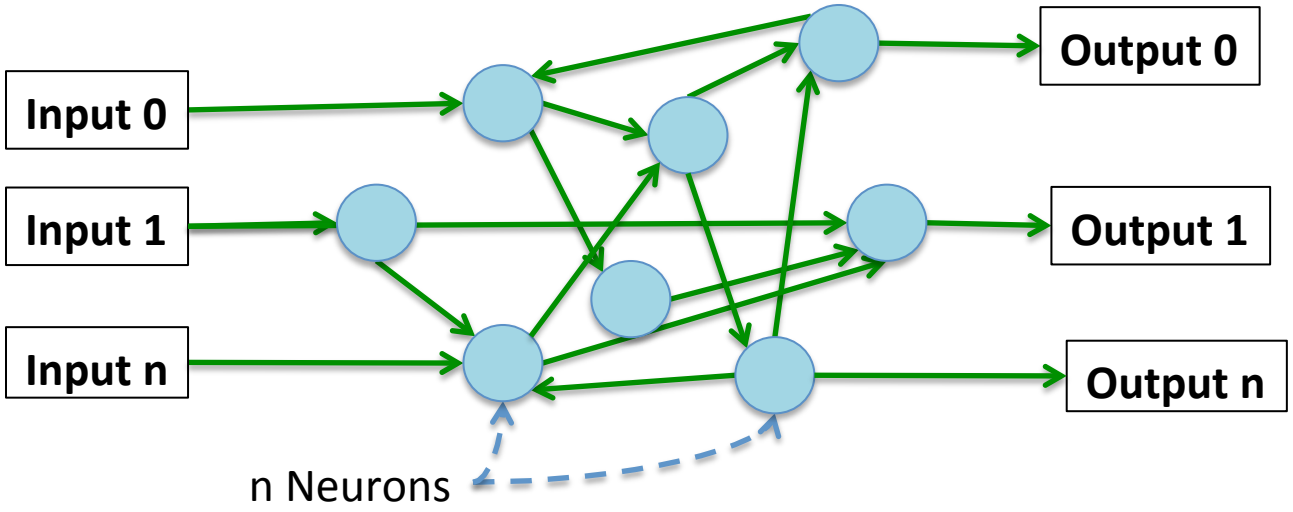
- **How does qualia work?**

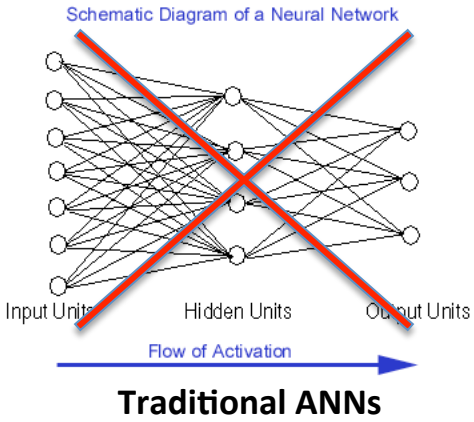
"The sensation of color cannot be accounted for by the physicist's objective picture of light-waves. Could the physiologist account for it, if he had fuller knowledge than he has of the processes in the retina and the nervous processes set up by them in the optical nerve bundles and in the brain? I do not think so." – Schrodinger.

I think he's wrong. A different talk.

Self from no-self in an Artificial Neural Network

- I create an artificial neural network with **completely random connections** between neurons and having no a-priori structure. Brains initialized with 30 neurons.





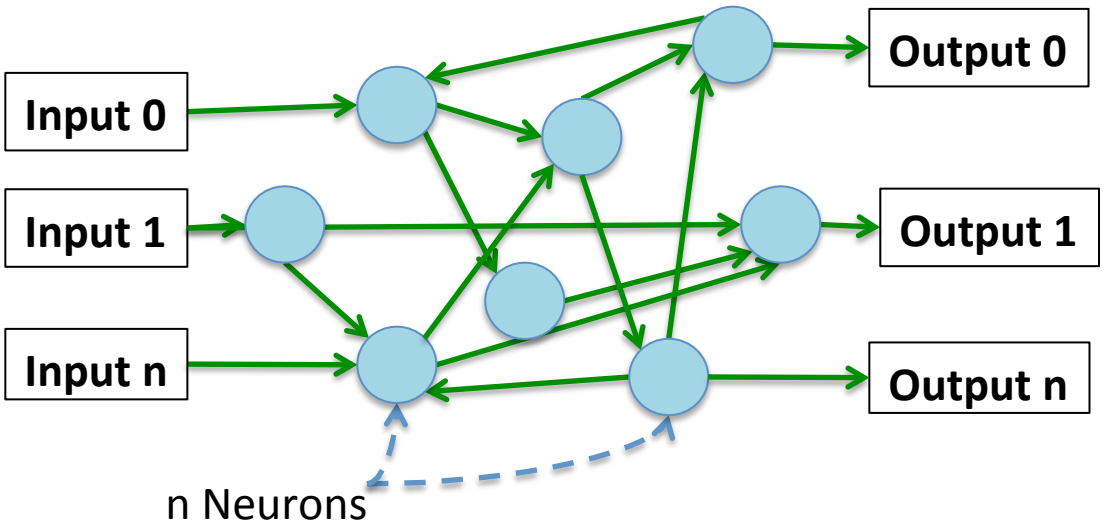
Traditional ANNs

- If two neurons are connected, they have:
  - a **distance parameter** between them (0, 1)
  - a **synaptic weight** that is directional (0, 1)

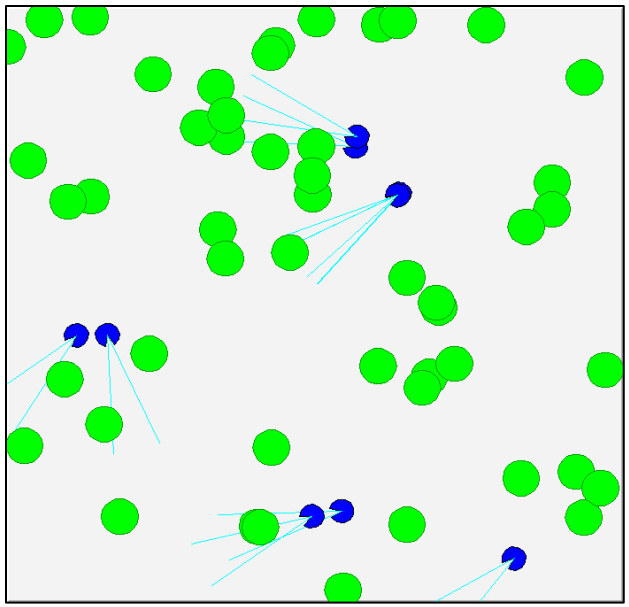
- The brain is hooked up to an on-screen bot whose field of vision feeds into input neurons 0 – 11. Neurons 12, 13, 14, 15 when fired make the bot take one step forward, 0.1 radians left, 0.1 radians right, and one step back, respectively.

Self from no-self in an Artificial Neural Network

The Brain



The World

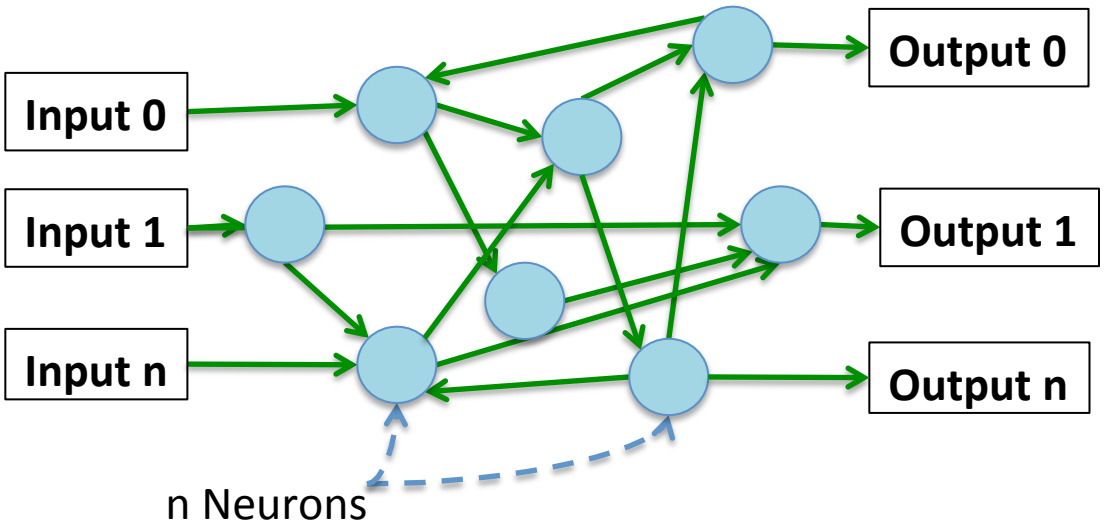


- Several bots (blue) are made, each with a 30 neuron “mush” brain. No correlation between what it sees and how it moves initially.

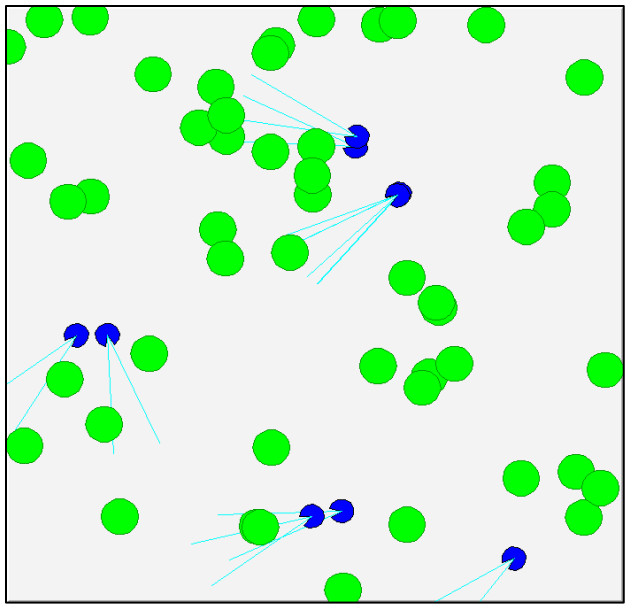
- If bot bumps into food (green):
  - a copy of the bot is made with small random mutation in the distance parameters
  - **no incentive given to that bot**. Bot is not aware of daughter bot. **No training.**
  - oldest bot in group has to die in order to maintain population, not kill CPU.

Self from no-self in an Artificial Neural Network

The Brain



The World

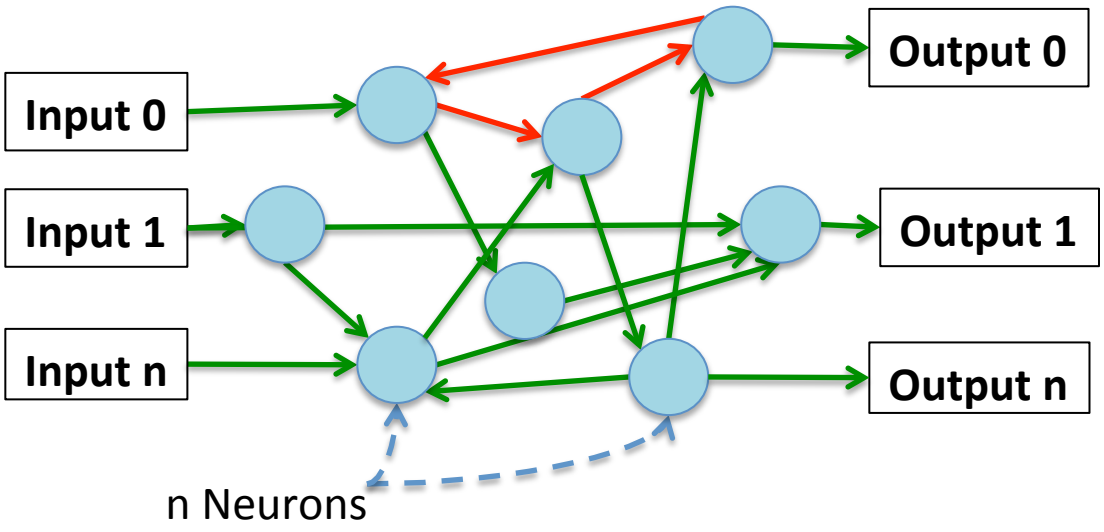


- Synaptic weights are reinforced or decayed exponentially within a lifetime based on use of connection

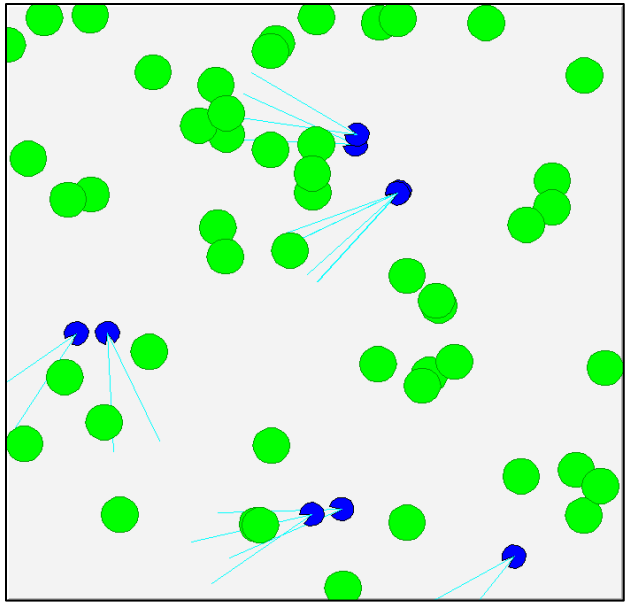
- Impulses in the brain are simulated in time-steps. Neuron fires if sum of inputs integrated over time crosses a threshold. After firing, the sum goes to 0. Synaptic weight and distance multiply the impulse being communicated to a downstream neuron

Self from no-self in an Artificial Neural Network

The Brain



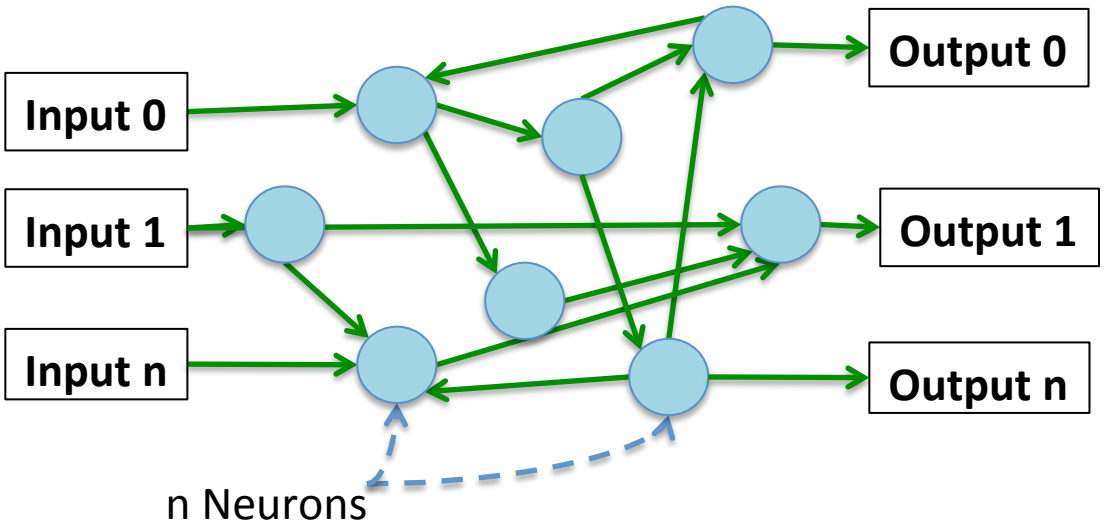
The World



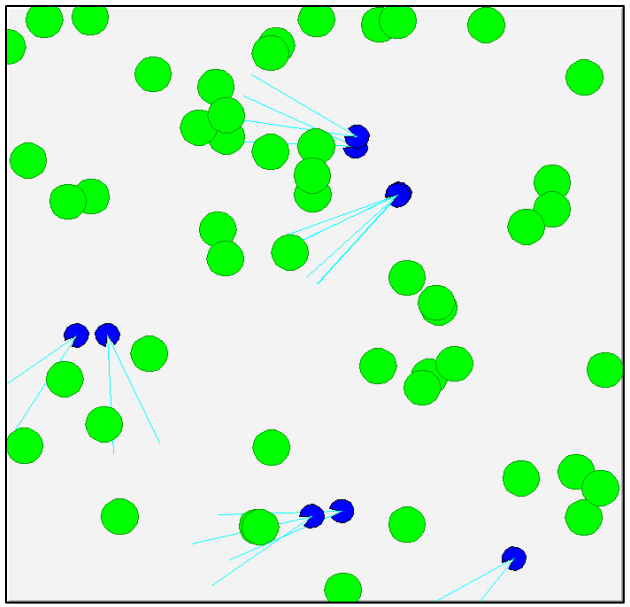
- Time-stepped brain allows time-signature inputs and sophisticated motions.
- Feedback loops possible that will self-reinforce synaptic weights if pulsed, creating some sort of memory.
- Sophisticated NN-ware can emerge in this brain. Some ancient versions of the self?

The Mutation Rates – Should they Mutate?

The Brain



The World



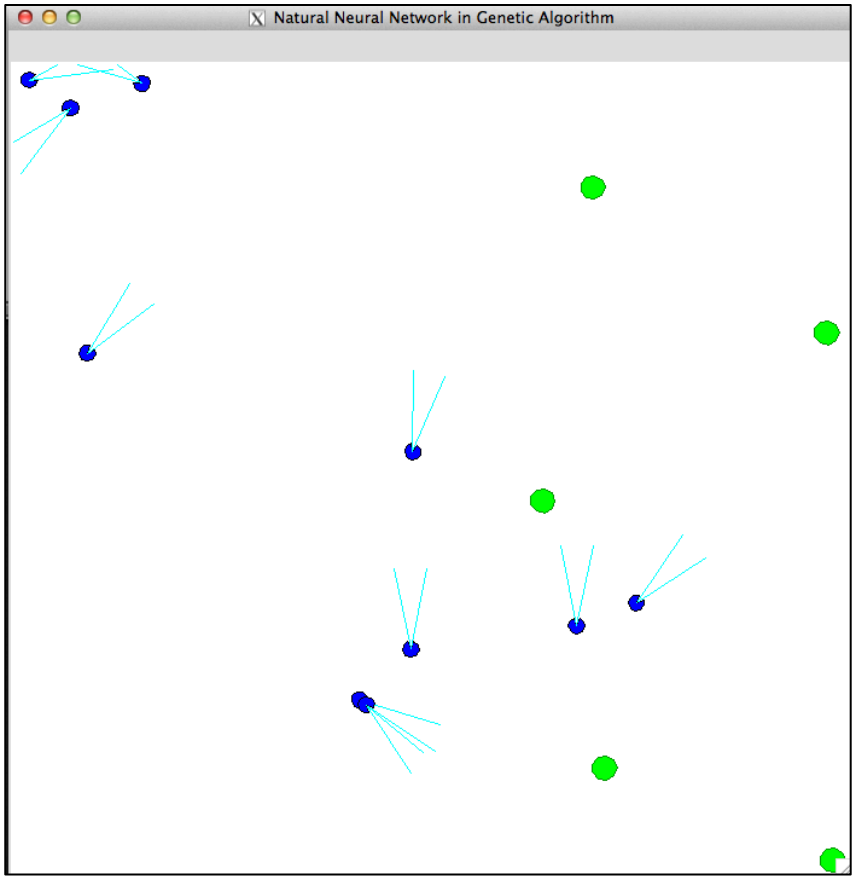
- When new bot & brain is made, the following mutations are allowed:

1. A neuron is added or subtracted from the brain with an equal probability of $\mu_{newNeuron} << 1$
2. A new connection between neurons is created or deleted with equal probability of $\mu_{newConnection} << 1$
3. An existing connection distance is modified with a flat smear of $\pm\mu_{modConnection} << \pm 1$

Live Demo

- Watch live demo:
  ./BrainInWorld -debug 1 -timeStep 10 -worldSize 200 -nBots 10 -nFoods 5 -nPredators 0

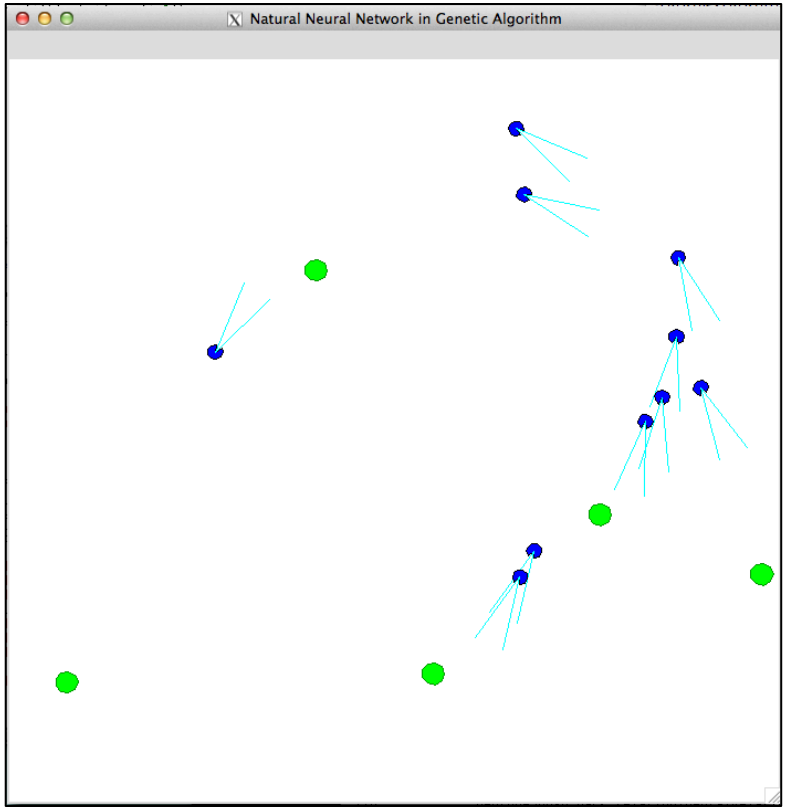
- Convince yourself they're quite dumb. Take 1 step forward, 2 steps back, etc.



Live Demo

- Speed it up:
  ./BrainInWorld -debug 1 -timeStep 100 -worldSize 200 -nBots 10 -nFoods 5 -nPredators 0

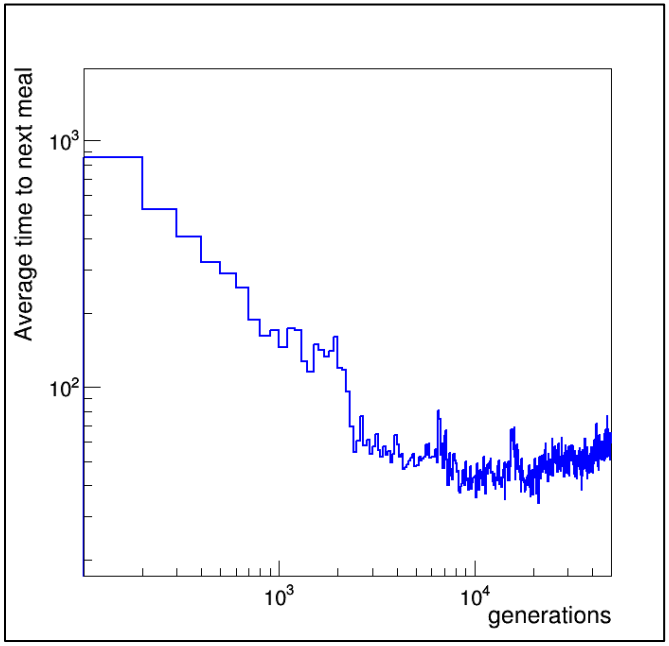
- Competitive behavior emerges. **No incentive exists from the point of view of an individual bot.** It doesn't even know it reproduced or that eating food is a "good thing". How does this emerge?
  - *Hint*: A daughter bot exists only because its parent bot had some tendency to hit food, i.e. some wiring in the brain correlated food in field of vision to firing motor neurons so as to reach it.

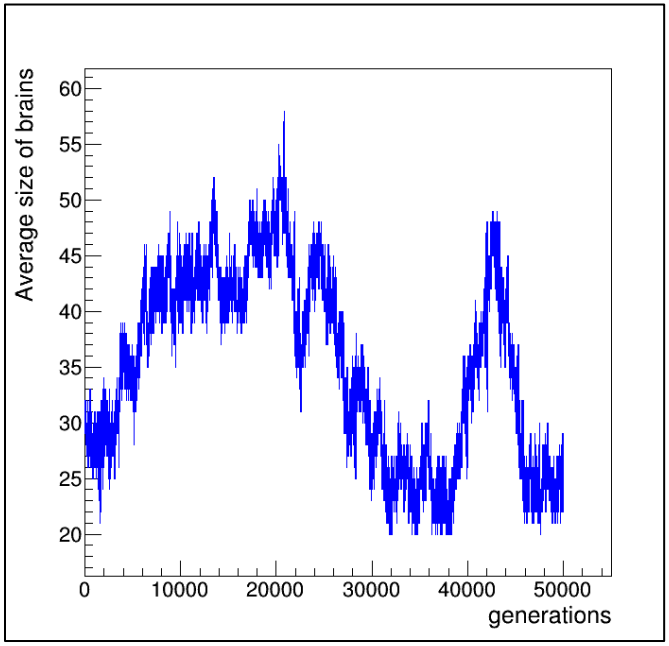


- Can you see "**common ancesto**r" behavior? Run with -debug = 1 to see ancestral names.

- Behavior **suggestive of self-preservation** emerges as bots compete with their offspring

The Blind Watchmaker





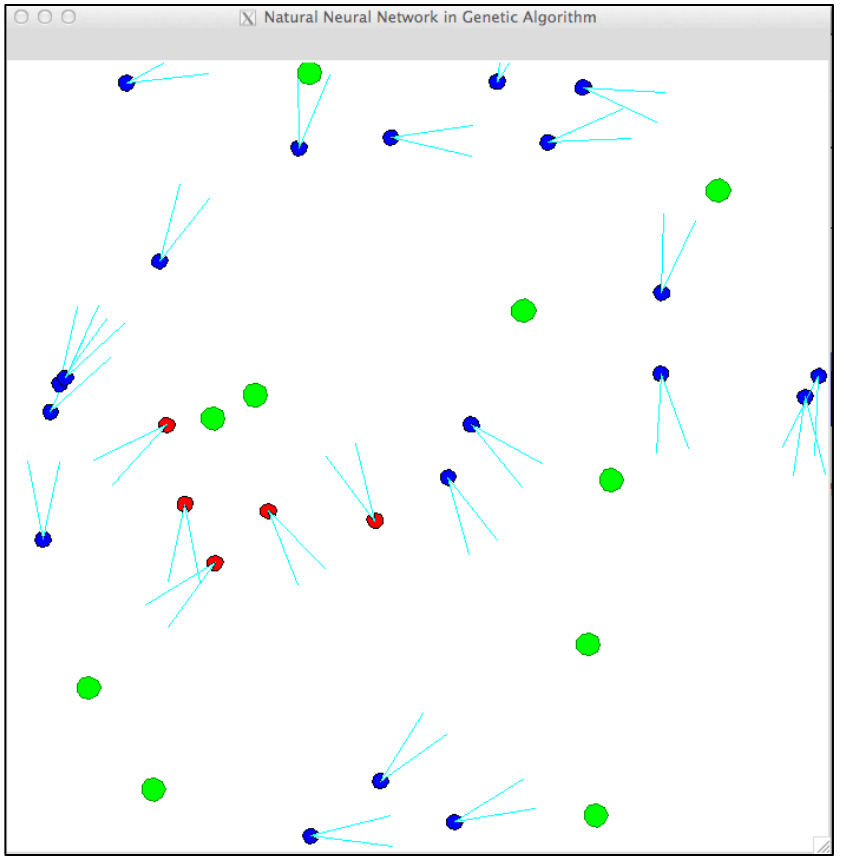
- Evolution is blind. Won't easily correct statistical fluctuations into dumbness. Use -nBots = 20 at least for meaningful plots.
- Average time for bot to reach food falls rapidly as 1/generation and then asymptotes
- Average number of neurons fluctuates as new efficient **strategies are discovered**. Less brain cells imply faster reactions.

The Blind Watchmaker has **created NN-ware** out of randomness.

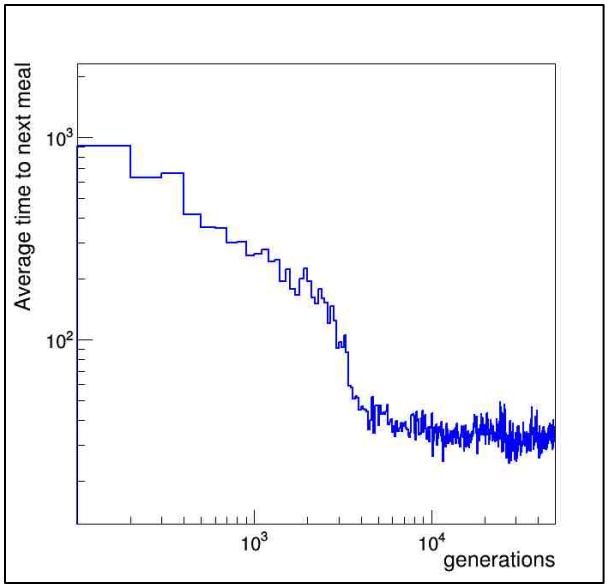
Introducing Predators

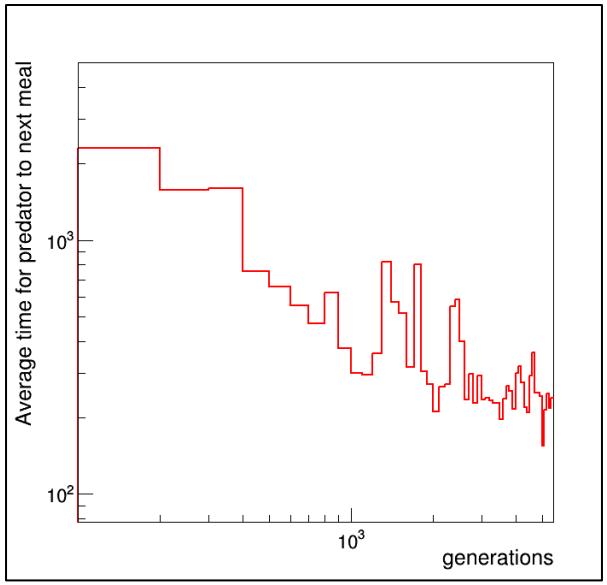
- Predators have the same kind of brain as the bots, but eat bots instead of food.
- When predator eats bot,
  - A copy of that predator is created modulo mutations
  - The oldest bot in the group is copied modulo mutations
- Predators can see bots and food. Bots can see predators and food.
- Live demo:

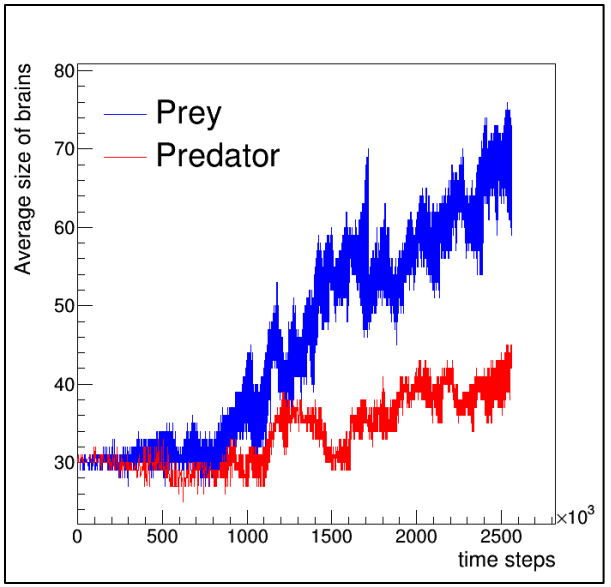
./BrainInWorld -debug 1 -timeStep 100 -worldSize 200 -nBots 20 -nFoods 10 -nPredators 5



An AI Arms Race







Results from: *./BrainInWorld -debug 0 -worldSize 1000 -nBots 50 -nFoods 50 -nPredators 20*

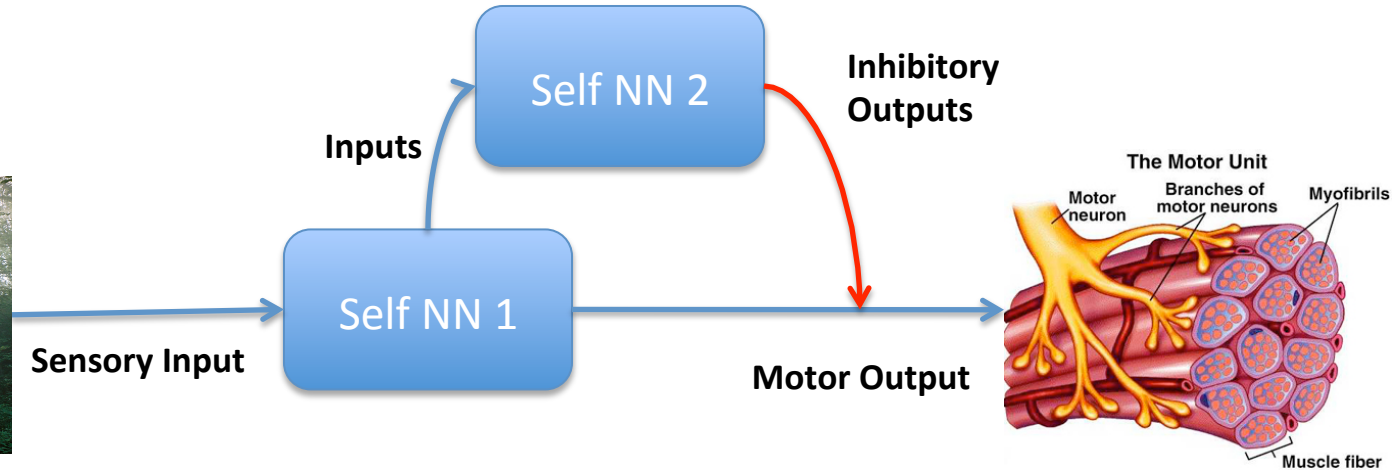
- Bots appear to avoid predators in its search for food
- Predators learn to stalk the food of prey
- Predators learn to swirl in opposite direction. Strategies oscillate.
- The time it takes for bots to eat food and predators to eat bots drops rapidly before a strategic arms race for brain cells takes off.

Looking closely at the behavior of the bots, behavior suggestive of self-preservation that **implies an understanding of the existence of the self** and its purported importance is apparent.

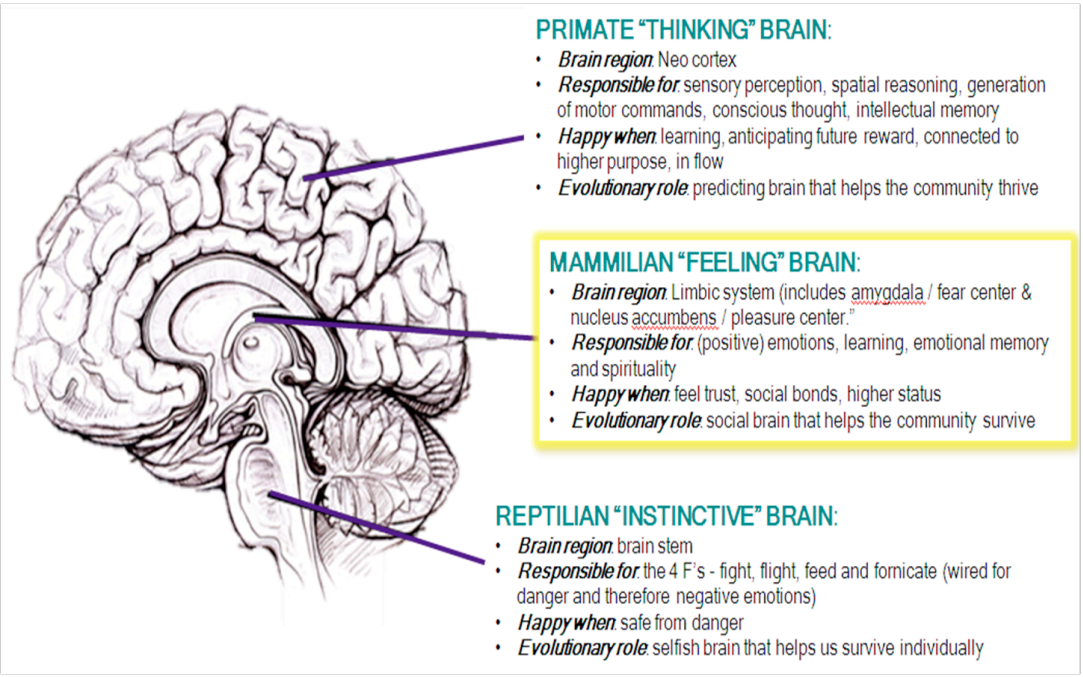
What about self-awareness I know and love?

- The natural selection environment is very simple so we can understand the basic principles of the system. More complicated game worlds can yield sophisticated survival strategies. Add predators and other pitfalls, a layered self can emerge.

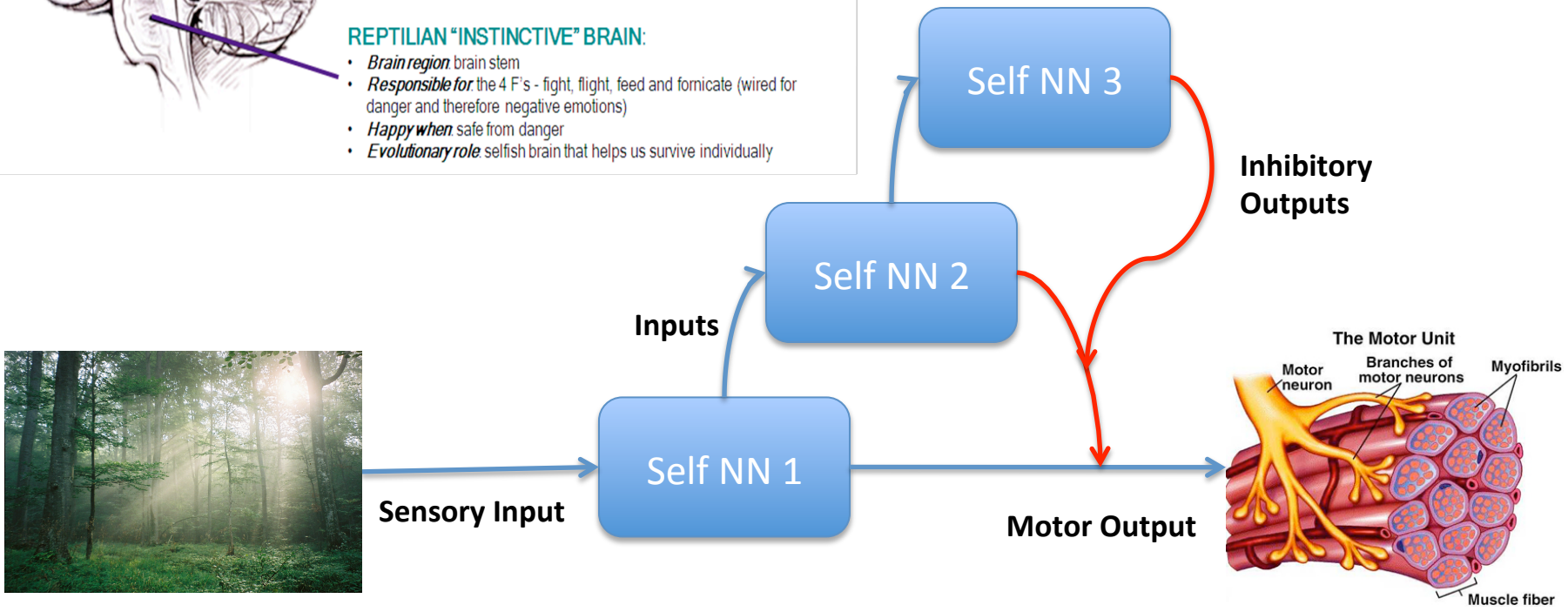
- Inhibitory neurons have not been simulated. I think they are essential to layered self-preservation circuits, which would bring it closer to the self-awareness we experience.



What about self-awareness I know and love?



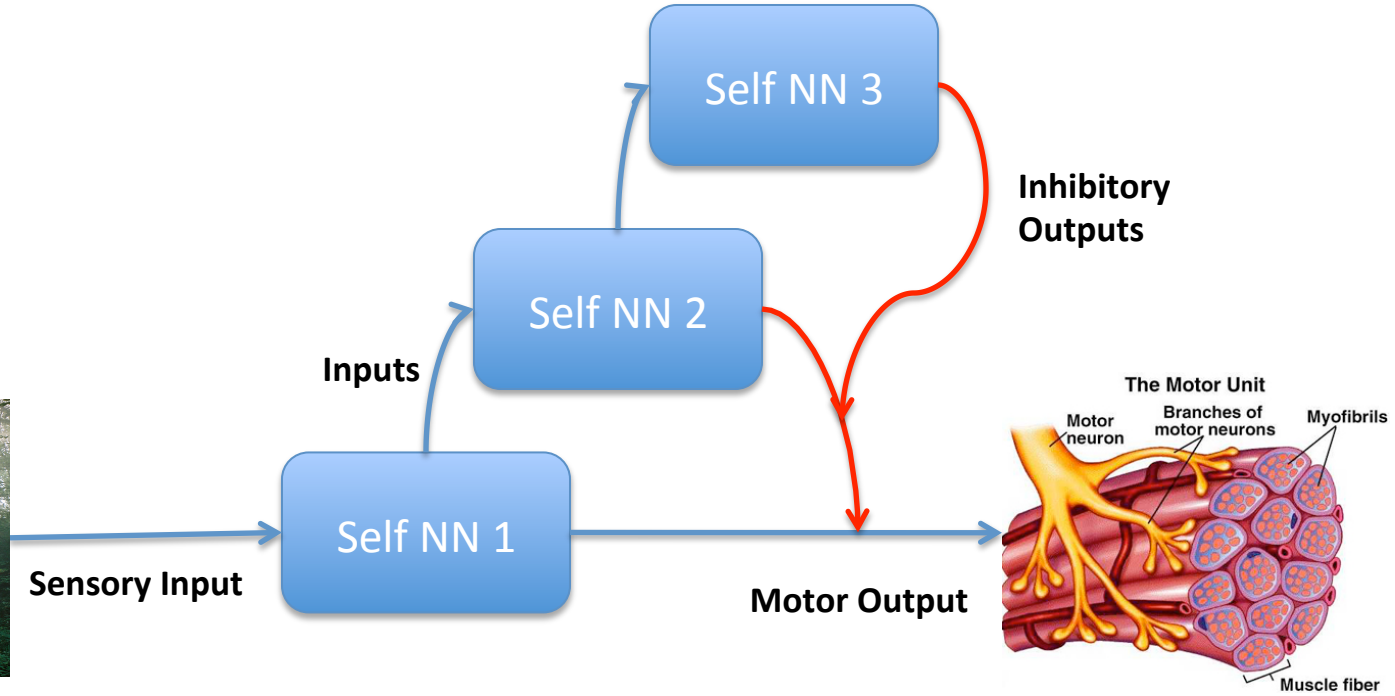
- This hunch is structurally mirrored by evolution.
- We identify with the highest layers of the self NN-ware as the conscious self.



What about self-awareness I know and love?

Cerebral cortex in rats' brains is set up like the Internet. USC Research

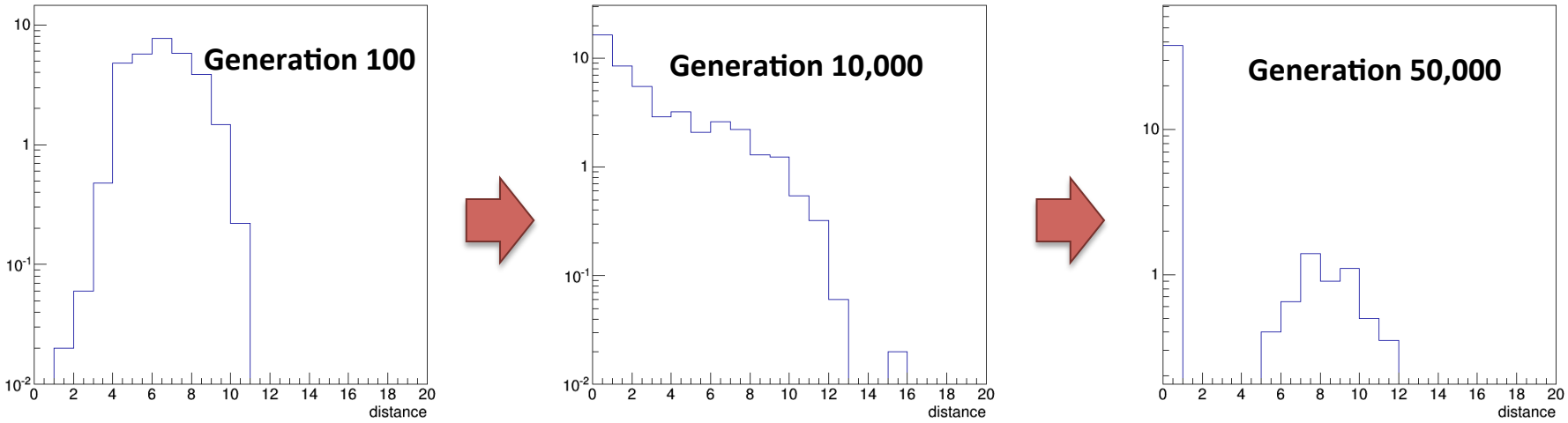
"Now, with a more comprehensive picture of how neurons connect to one another, they've discovered local networks of neurons layered like the shells in a Russian nesting doll. … Two local networks — one governing vision and learning, and another tapped into bodily concerns like muscle and organ function — make up the inner shell of the rat's cerebral cortex. Two others — one governing smell, and another that assembles and makes sense of the information from the other three networks — make up the outer shell."



Prying open the AI brains

In the works

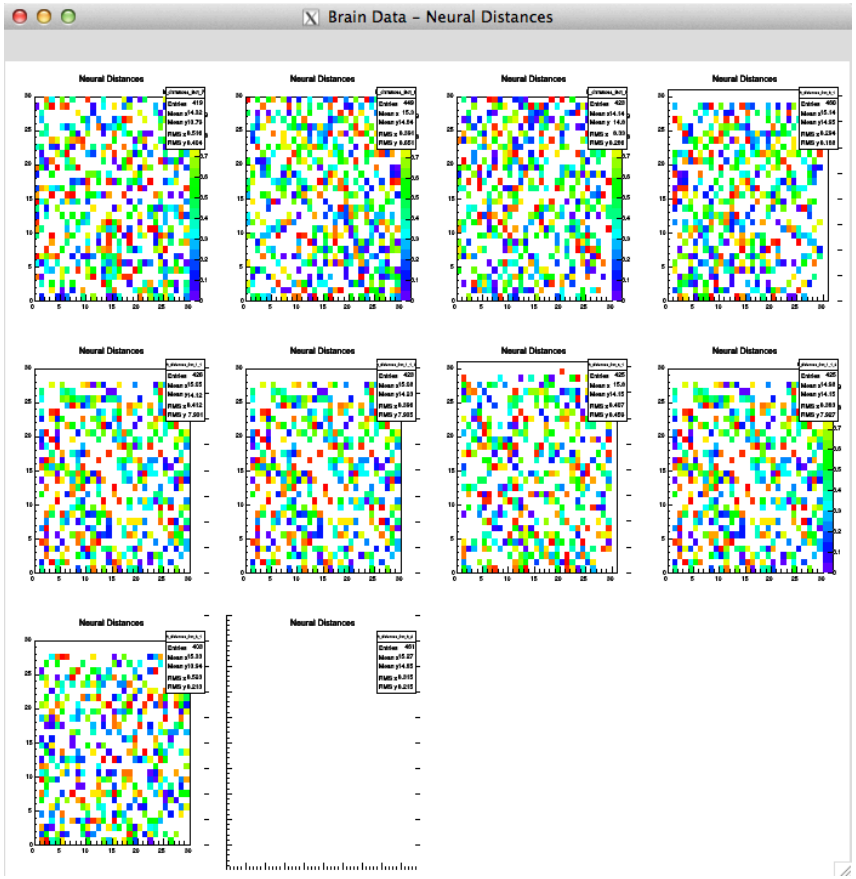
- What measures of network structure can you think of?
  - How can I detect loops mathematically?
    - How can I detect layering if any?



The sum of distance parameters for all neurons in a brain is plotted.
It seems to prune itself such that there are “important neurons” with 5 – 12 connections.

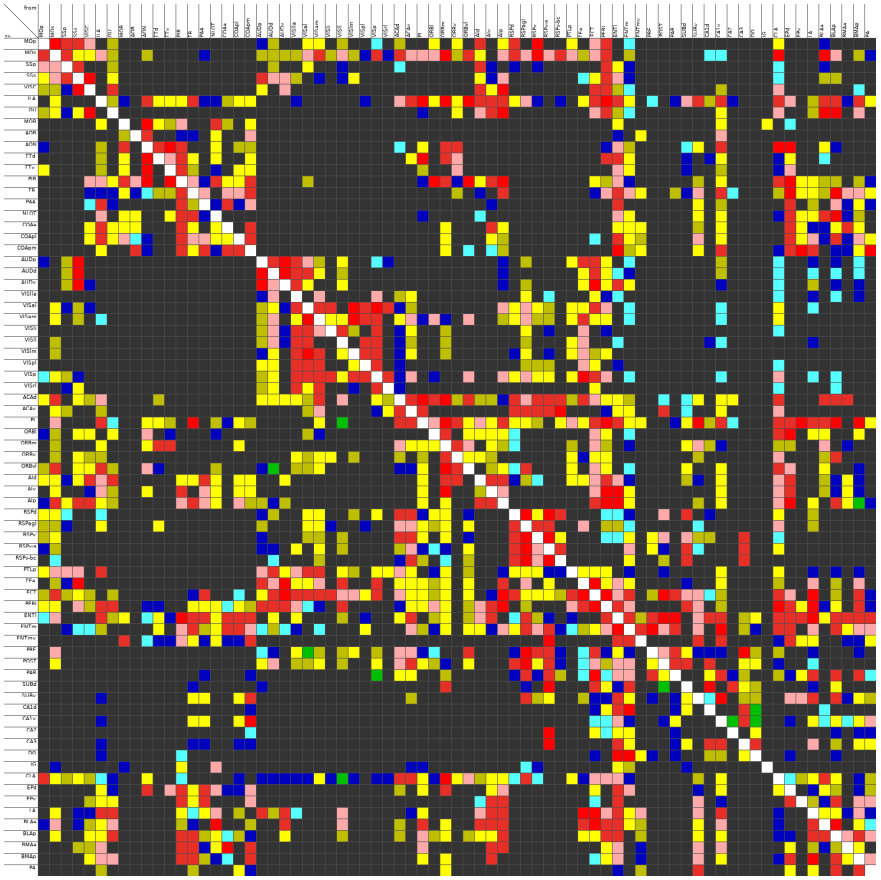
Has to be studied.

Prying open the AI brains



The distance matrices of my brains

Has to be studied.



The distance matrix of a
73 neurons x 73 neurons slice
of a rat's brain.

http://goo.gl/rJihMY

After such knowledge – a talk of its own

Ethical Implications

If the self is nothing more than an information system and its elementary, unique and continuous nature no more than an illusion put in place by evolution, what does it mean for what is *ought*?

- Academic understanding is cool. But internalizing this can be unnerving since our ambitions, hopes and entire lives are centered around it. A loss of center, a humiliation of the ego comparable to heliocentricity and evolutionary biology.
- A radically *new form of empathy* arises.
- With the self seen through, you relax as the universe plays through you.

Alex Grey, 1988 – Adam and Eve

With the ego seen through but not denied, the separation from the world undoes itself

Conclusion

The study of consciousness is where classical mechanics was when “F”, “p”, “m” were being rigorously defined and their relationships understood. A conjecture I think is on solid ground:

Any part of the universe that experiences a self or claims to experience a self has undergone selection through survival of the fittest.