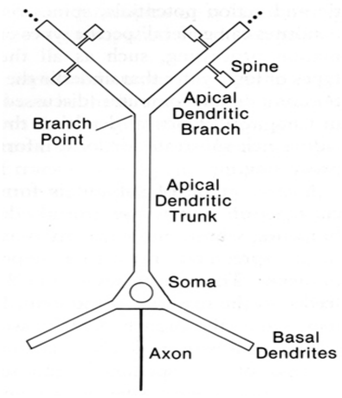
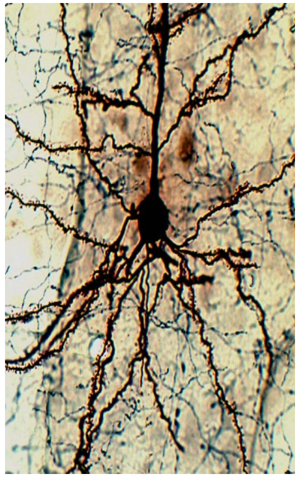
Biologically Plausible or Implausible?

The (Easy) Case to Make against Backprop

Operationa/Structural Differences Brains vs. Computers

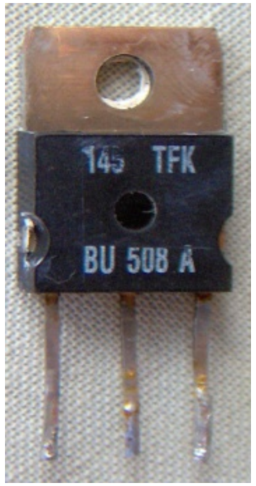
Neuron/Synapses

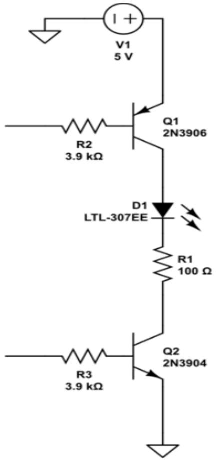




VS.

Transistor/Gates





OPERATING/COMPUTING CHARACTERISTICS

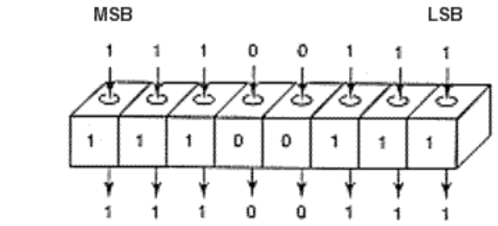
~10^3 Hz	Clock Rate	**10^9** Hz
~10^2 m/s	Signal Velocity	**10^8** m/s
~1	Signal-to-Noise	**10^6**
~10^4	Parallel Connections	**~1**

Organizational/Conceptual Differences: Brains vs. Computers

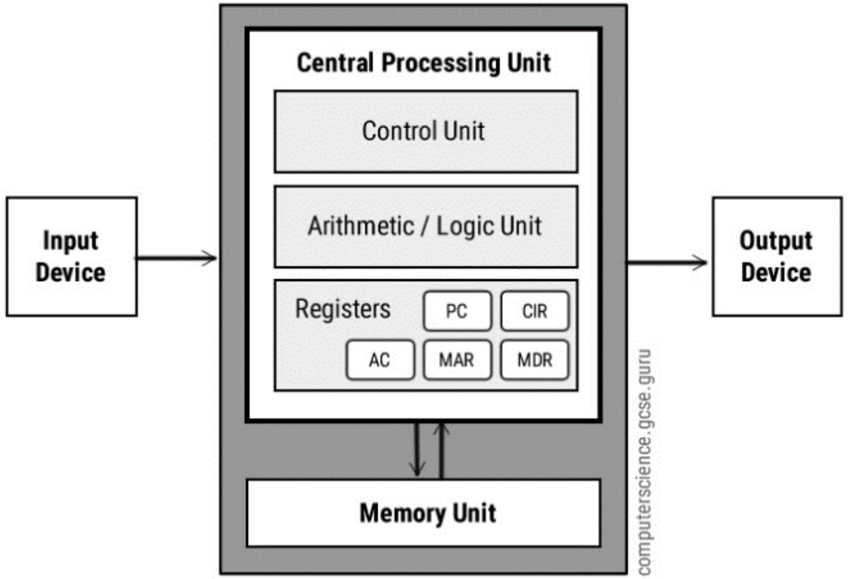
MEMORY STORAGE/RETRIEVAL

Neuron/Synapses		Transistor/Gates
Distributed Circuits	Medium	**Address Registers**
Content Addressable	Mechanism	**Instruction Addressable**

MEMORY REGISTER



Von Neumann Architecture: Memory Separate from Computation.



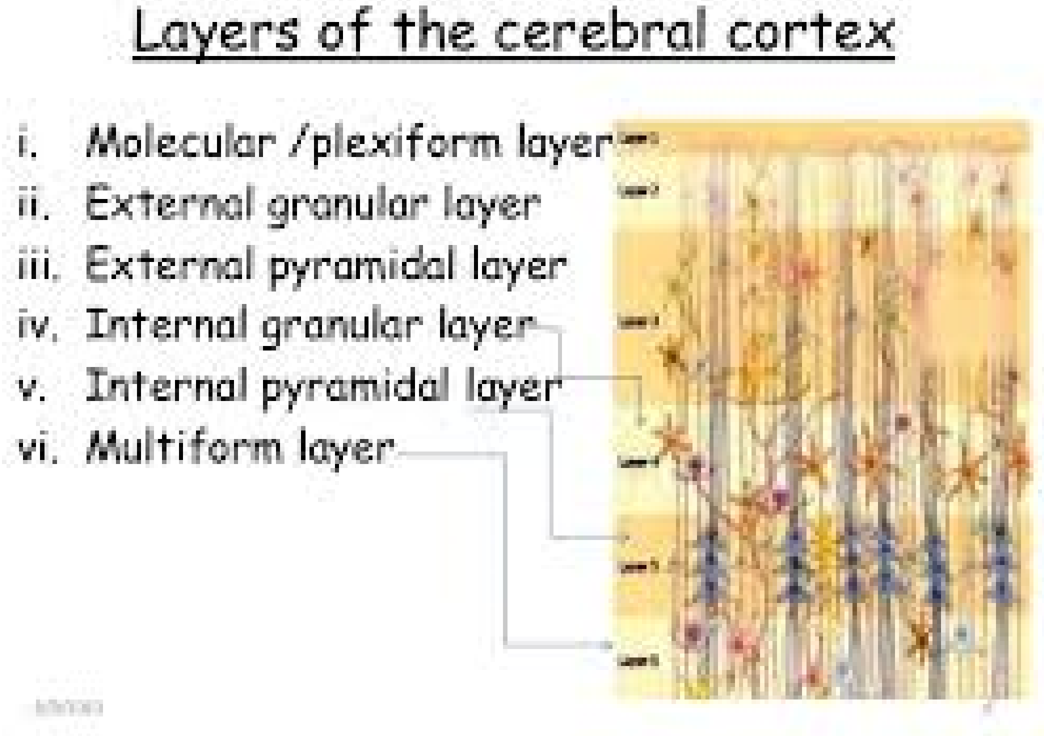
Biological Implausibility

Algorithmic

1) The learning is only possible in supervised mode
2) Where is the error signal?
3) Non-local: error gradients propagated through entire network
4) Learning requires clock to sync forward/backward passes
5) Learning requires computation of derivatives
6) Learning requires symmetric weights
7) No top-down feedback or recurrence
8) Not tolerant to operational noise

Biological Implausibility: Another Reason

Most processing in real cortical networks is not long-range but short-range



Cerebral cortex is the **LEAST** likely place in the brain for supervised learning

In cerebellum, climbing fibers look like teaching signal; nothing similar in cortex

Endless Attempts to Neurally Justify Back-prop

Trends in Cognitive Sciences March, 2019

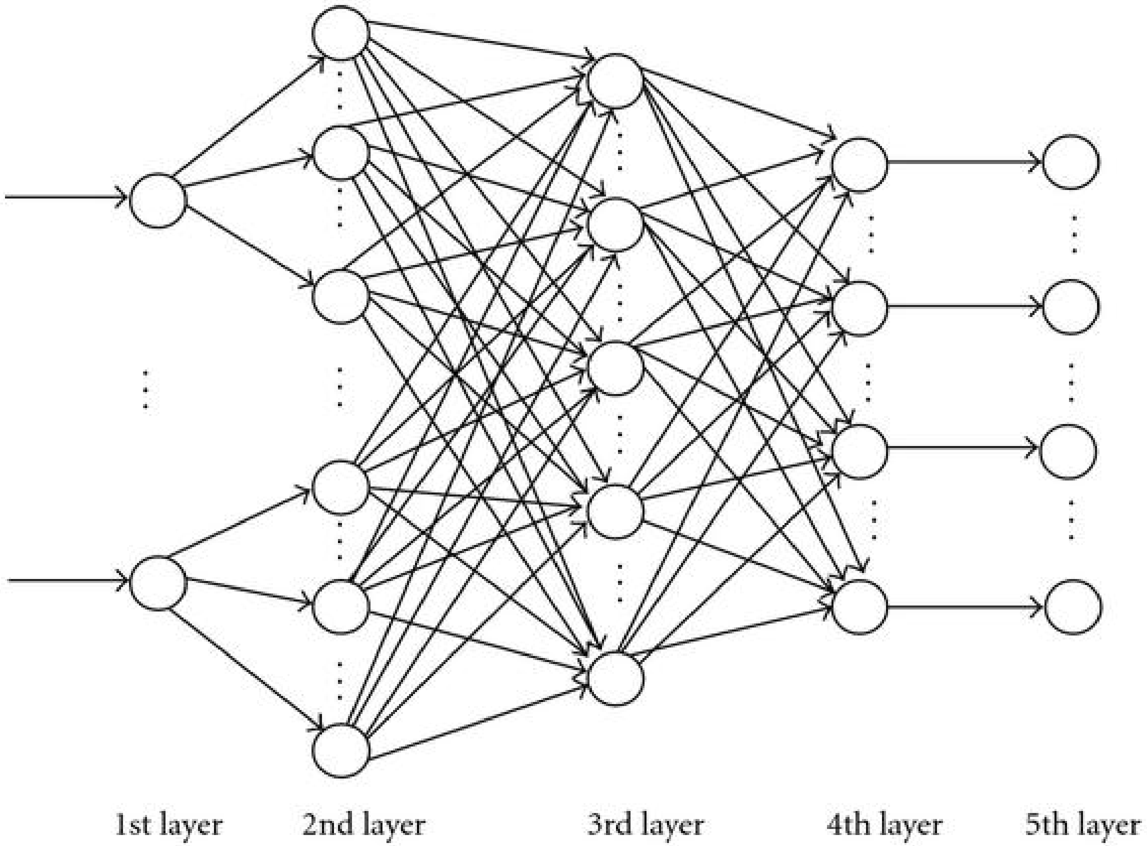
Theories of Error Back-Propagation in the Brain

This review article summarises recently proposed theories on how neural circuits in the brain could approximate the error back-propagation algorithm used by artificial neural networks.

Stability-Plasticity Dilemma

Major Performance Limitation

Pertinent to any distributed information processing system (Grossberg).



Distributed Systems: Blessing and a Curse

Why such networks in the first place?

PDP Group (1986)

- Graceful decay
- Robustness to noise
- Generalization

A Few Words about Generalization

- Crucial to any intelligent system
- The network embodies one big continuous function.
- Gradually changes parameters to shape function per experience
- It stores nothing – everything is shared

Stability-Plasticity Defined

The Downside to Sharing: Overwriting

In a highly distributed and massively interconnected system, how can the learning of one item (through synaptic change) not impact the learning of a second item, when the synapses and nodes are shared? The system needs to be sufficiently plastic to rapidly accommodate new information without overwriting old information.

Can be seen as *Generalization-Interference* Tradeoff

The system needs to be sufficiently plastic to rapidly accommodate new information without overwriting old information.

How is it Solved in Today's Deep Nets?

It isn't – you live with it.

- Data is always presented Interleaved
- A huge amount of iterations
- Network is frozen at the end

Why is it not studied more?

- It's a really hard problem.
- Can work around it.
- Requires a different kind of thinking (complex systems vs, reverse engineering)

Bottom Line: A Big Opportunity

Second AI Winter: Early 2000's

Two Reasons

1. Hard to implement on large problems
    - Data hungry algorithms
    - Vanishing gradient problem on multi-layer networks

2. Support Vector Machines

What Happened in 2012?

1) Convolutional deep neural networks Krishevsky et al. (2012)
2) Big Data and powerful computing
3) Important domain application
4) Social media companies with a lot of money

Mostly Moore's Law!!!

“It’s Déjà Vu All Over Again…Only Worse”

Modified Yogi Berra Quote

Deeper nets means more knobs to tweak

Example: recurrent NN’s or backprop through time

“Any cyclic graph can be sufficiently well-approximated by an a-cyclic graph.”

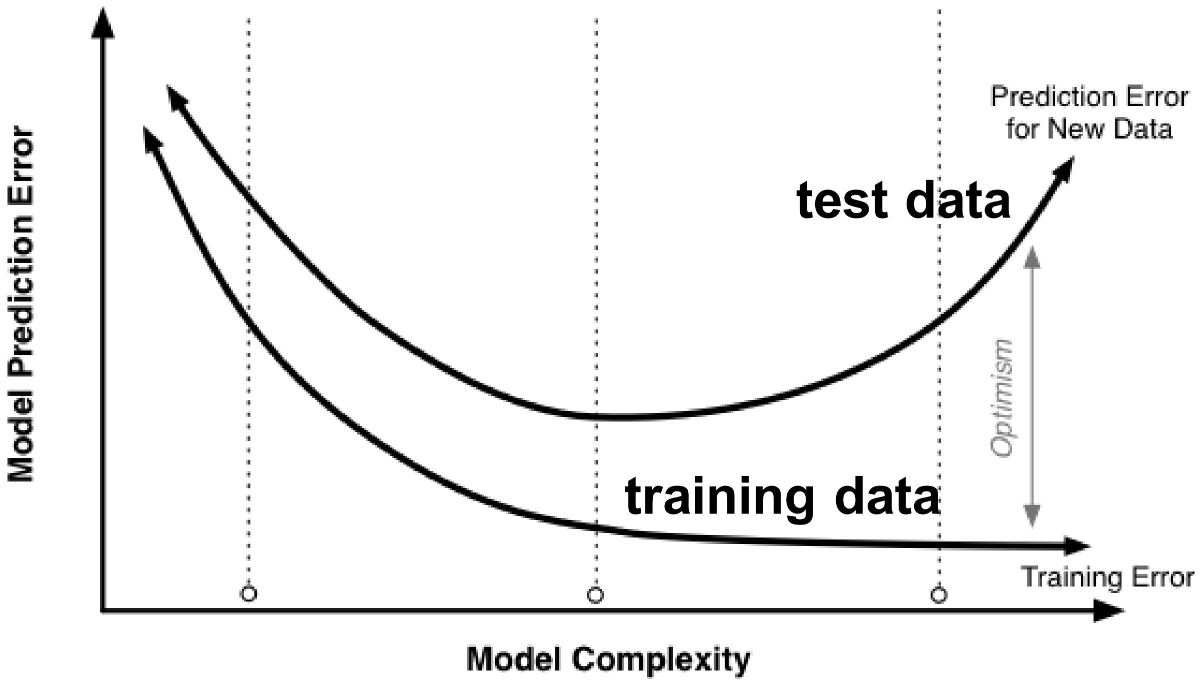
Technology empowers/encourages careless thinking

“There isn’t really any math to deep learning other than the concept of a derivative which is taught in high school calculus… Deep learning is broadly an experimental science, which in many ways is the opposite of math as traditionally envisioned, in which great insights follow deductively from prior great insights. If you ask a basic question like ‘why should use 4 layers instead of 3?’ there is no answer other than ‘4 works better’.”

Why All the Fuss?

Miraculously Manages to “Generalize”. i.e. No Overfitting

Possibilities of underfitting and overfitting (Bias-Variance tradeoff)



How do NN’s have so many parameters yet generalize so well?