123B S20 Module 7:
Future Directions of Bioinformatics

Today's plan

- A little context
- Deep Learning
- Projects:
  - Poriferal Vision
  - Data mining GenBank with simulated eyes
  - Coral Vision
  - Adverb

Machine Learning: The big idea

- What kind of machine?
- What kind of learning?

Machine Learning: The big idea

- What kind of machine?
  - Software classifiers (and similar)
  - With lots of parameters
  - Examples:
    - Position weight matrix
      - Parameters = column scores, threshold
    - HMM
      - Parameters = initial/transition/emission probabilities

- What kind of learning?
  - Training set of trusted examples + algorithms + statistics
  - Optimize the parameters

Neural networks: The big idea

- Analogy to our brains.
- Maybe our brains do complicated algorithms, but it doesn't *feel* like they do.
- Our collective intuition about learning, based on tens of thousands of years of experience:
  - Somehow we learn stuff.
  - Somehow we access what we learned.
- The most algorithmic things in our brains: individual neurons.
- If we model enough neurons algorithmically, will they act like brains?
  - How do we model a single neuron?
  - How many is enough?

A real neuron

- Binary: output is firing or not firing, no in-between.
- Decision is based on inputs, which are connected to senses or outputs of other neurons.

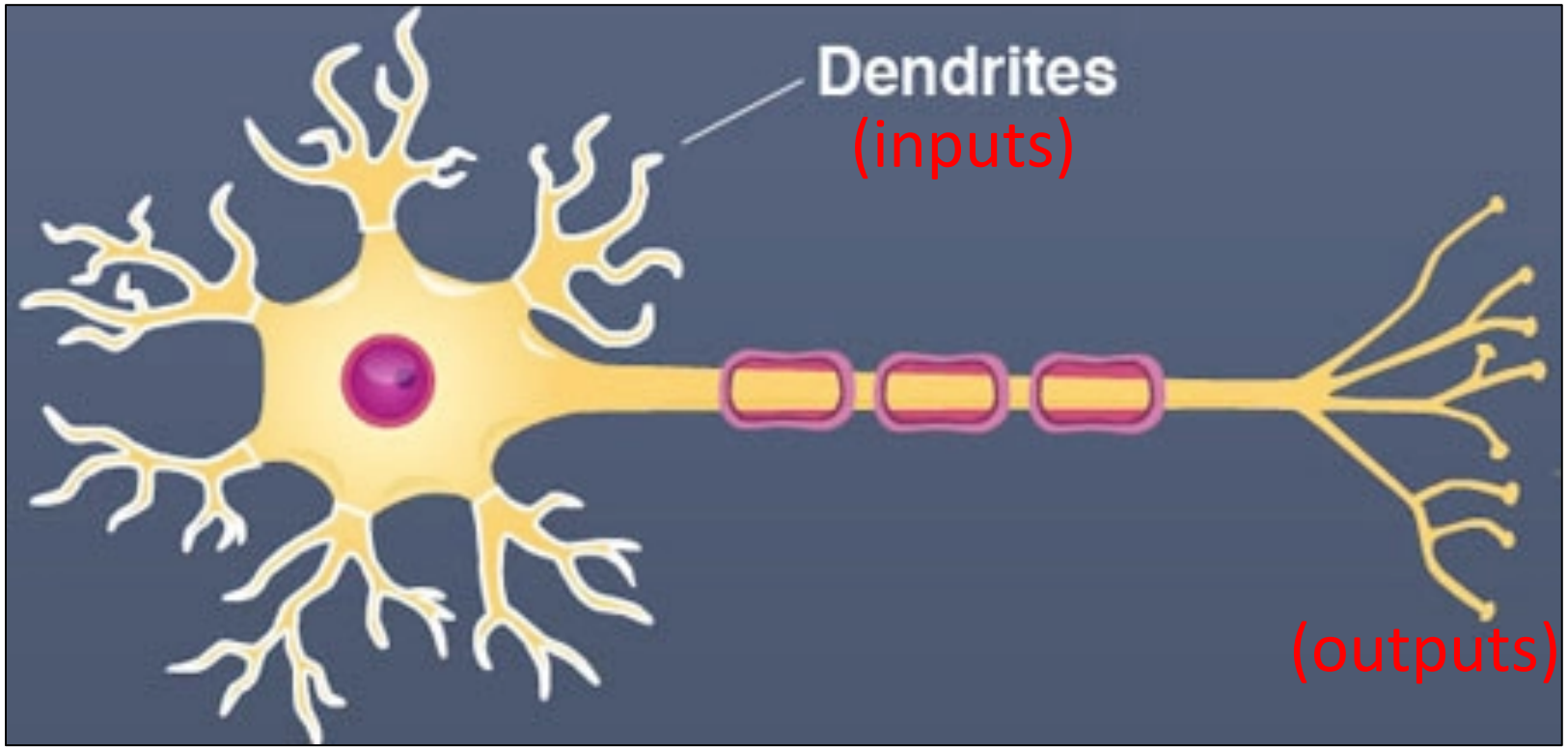
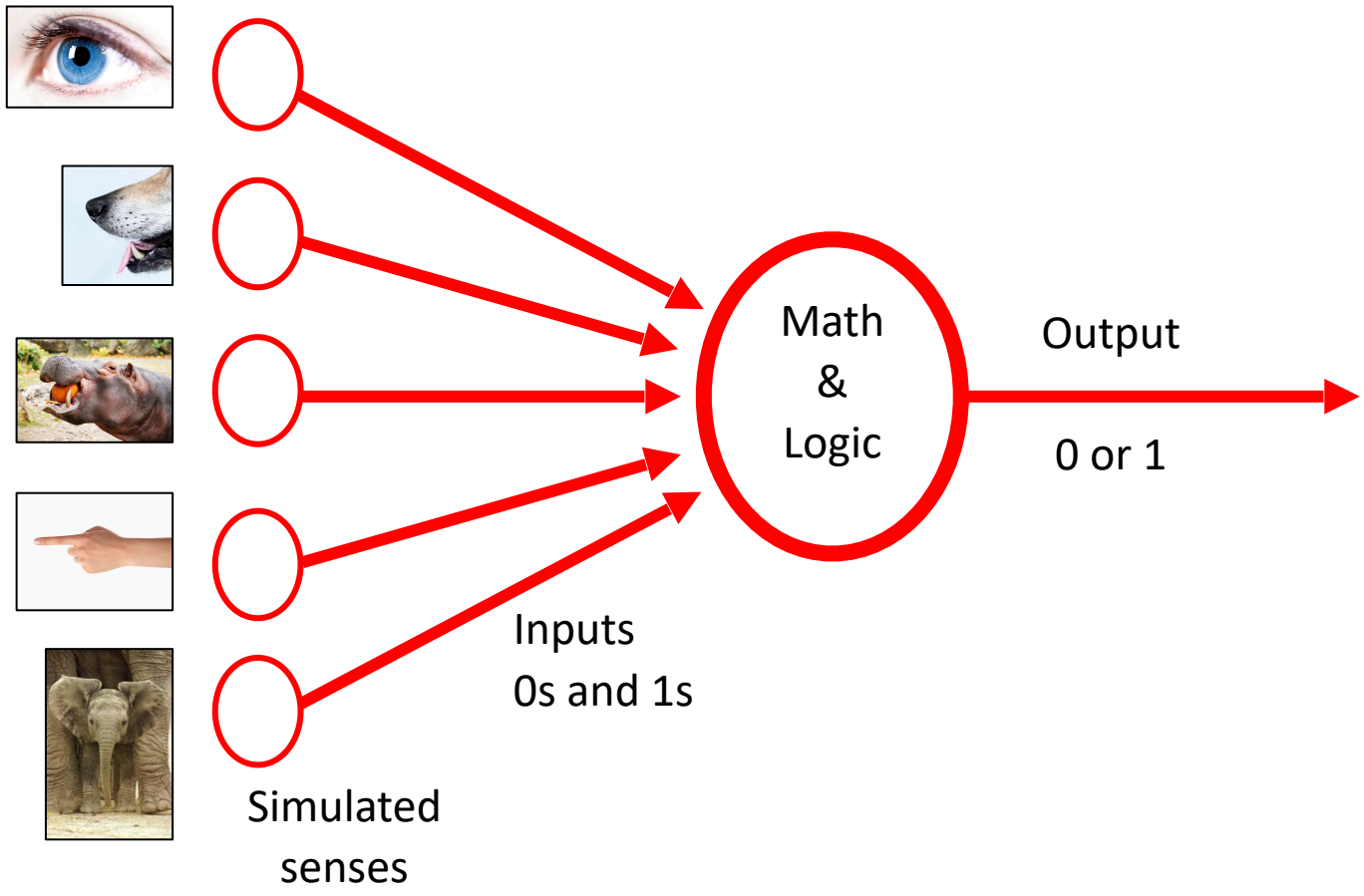
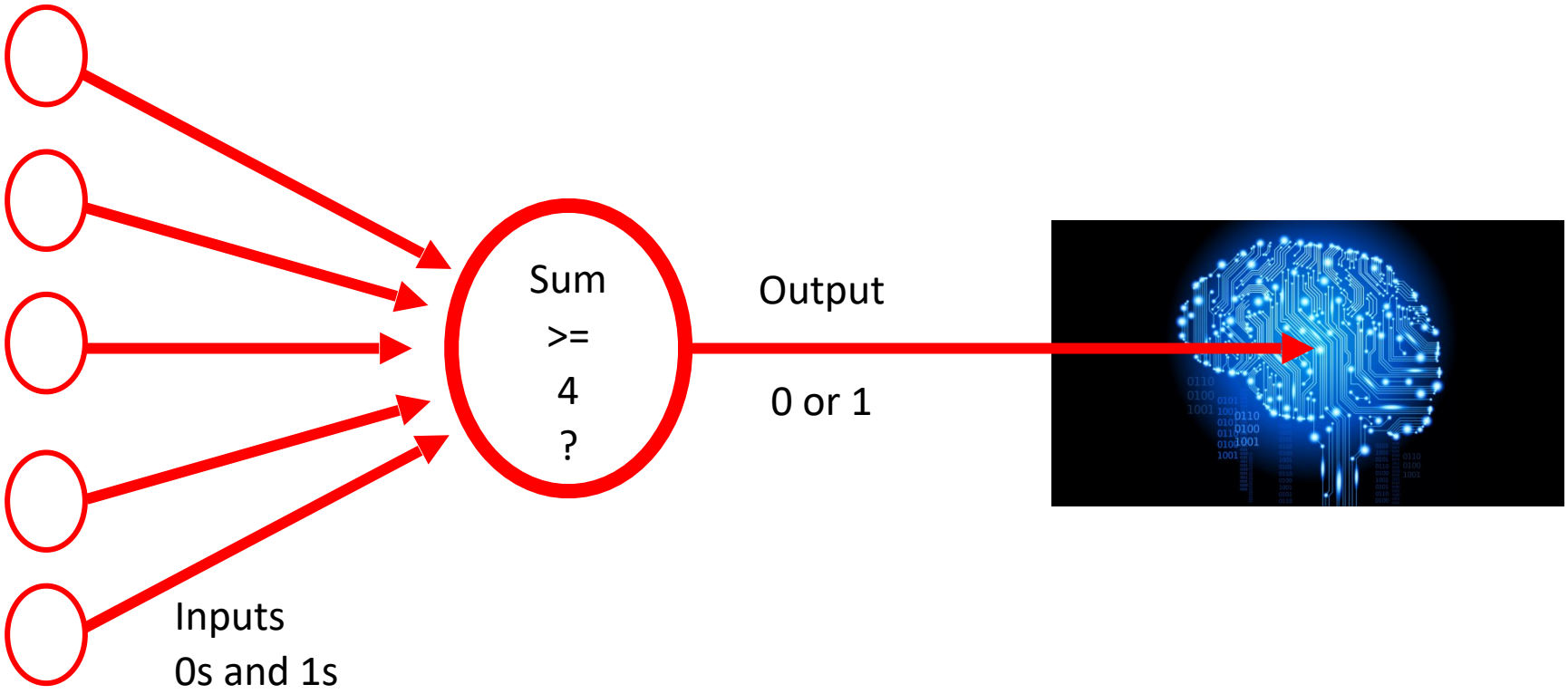


Image source: Quizlet

An Artificial Neuron

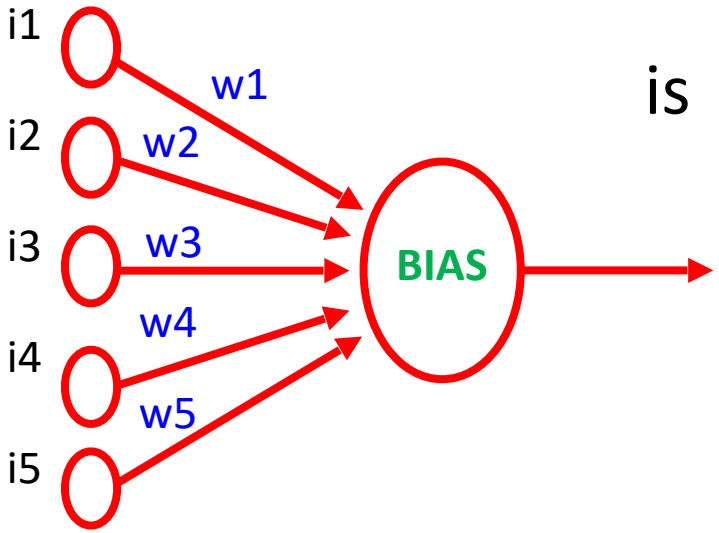


Realistic? Imagine a primitive metazoan with an early eye



Looking closer at an artificial neuron

- Output (= 0 or 1) is a function of "learned" parameters called *bias* and *weights*
  - 1 weight per input.
  - Each input is multiplied by its weight.
  - Output = 1 if sum-of-weights + bias > 0, otherwise output = 0.
  - "Learning" these parameters is the secret sauce.



Neural Networks: History

- 1940s, 1950s:
  - Brains are made of neurons
  - Brains think
  - Therefore if we model neurons we can model thought
- 1960s:
  - Computers have been invented
  - Interesting results with even a single software neuron
  - Theory about multiple neurons, but computers aren't powerful enough
  - 1969: Minsky & Pappert "Perceptrons: An introduction to computational geometry"
- 1980s:
  - Moderate commercial success, e.g. handwriting recognition
  - AI boom and bust
  - "The AI Winter" – through end of century
- 21st century:
  - Computers catch up
  - Success in some application domains, including image recognition
  - Google releases TensorFlow

The MARK 1 Perceptron

- 1958
- Filled a room
- $2,600,000 (1958)
- $22M today
- Could “distinguish left from right”

Frank Rosenblatt doing brain surgery with a wire cutter

Neural Networks: History

- 1940s, 1950s:
  - Brains are made of neurons
  - Brains think
  - Therefore if we model neurons we can model thought
- 1960s:
  - Computers have been invented
  - Interesting results with even a single software neuron
  - Theory about multiple neurons, but computers aren't powerful enough
  - 1969: Minsky & Pappert "Perceptrons: An introduction to computational geometry"
- 1980s:
  - Moderate commercial success, e.g. handwriting recognition
  - AI boom and bust
  - "The AI Winter" – through end of century
- 21st century:
  - Computers catch up
  - Success in some application domains, including image recognition
  - Google releases TensorFlow

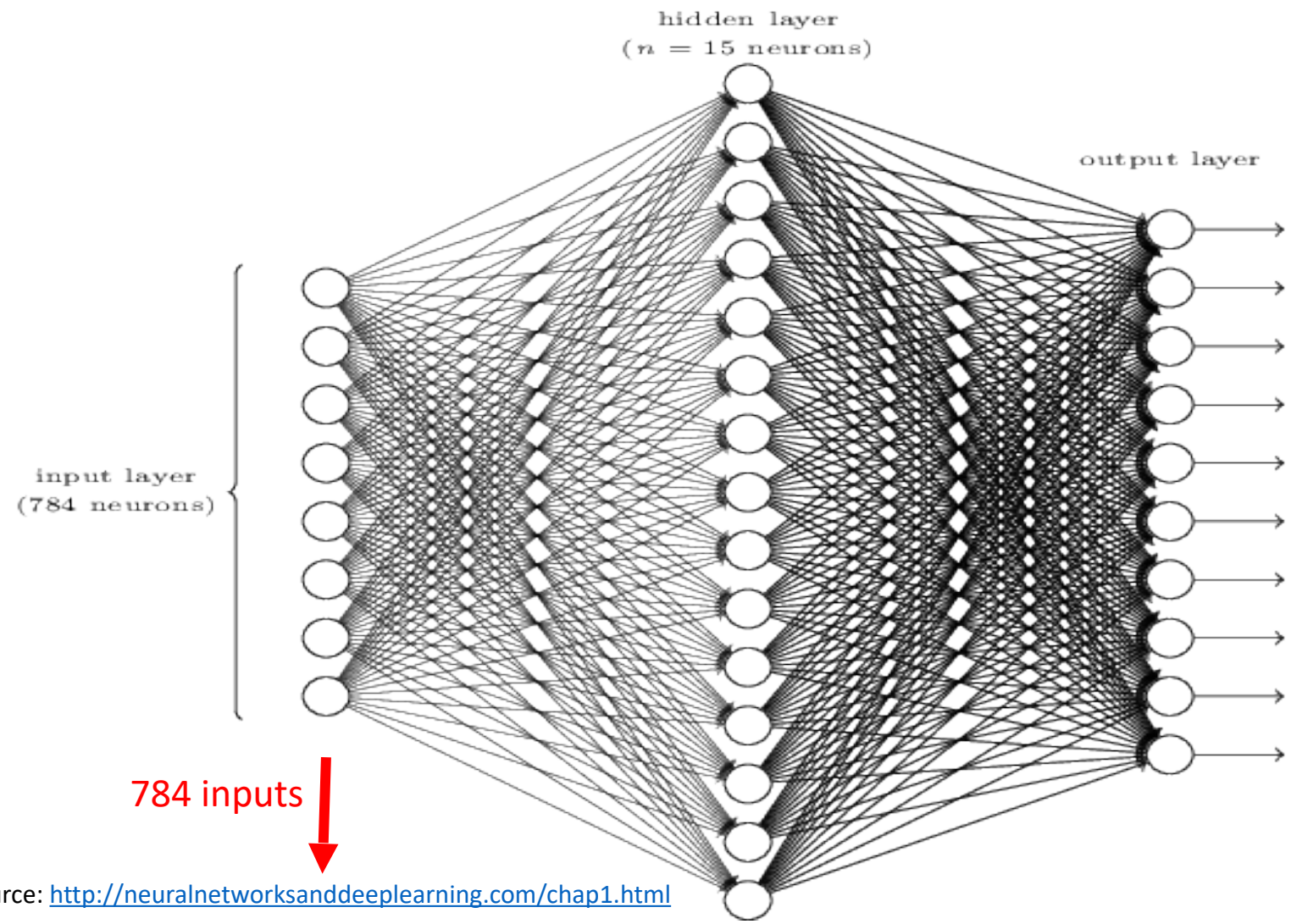
Neural Networks: History

- 1940s, 1950s:
  - Brains are made of neurons
  - Brains think
  - Therefore if we model neurons we can model thought
- 1960s:
  - Computers have been invented
  - Interesting results with even a single software neuron
  - Theory about multiple neurons, but computers aren't powerful enough
  - 1969: Minsky & Pappert "Perceptrons: An introduction to computational geometry"
- 1980s:
  - Moderate commercial success, e.g. handwriting recognition
  - AI boom and bust
  - "The AI Winter" – through end of century
- 21st century:
  - Computers catch up
  - Success in some application domains, including image recognition
  - Google releases TensorFlow

Neural Networks: History

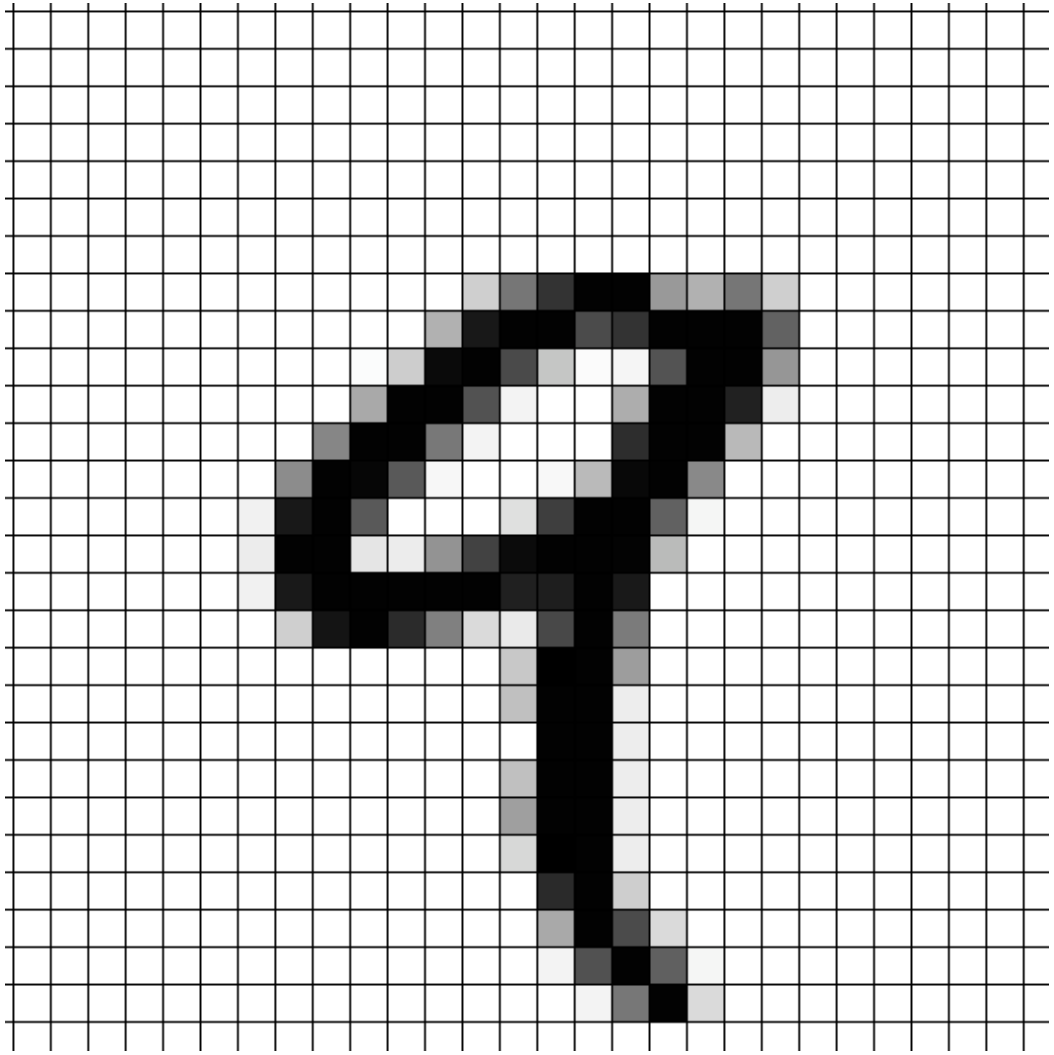
- 1940s, 1950s:
  - Brains are made of neurons
  - Brains think
  - Therefore if we model neurons we can model thought
- 1960s:
  - Computers have been invented
  - Interesting results with even a single software neuron
  - Theory about multiple neurons, but computers aren't powerful enough
  - 1969: Minsky & Pappert "Perceptrons: An introduction to computational geometry"
- 1980s:
  - Moderate commercial success, e.g. handwriting recognition
  - AI boom and bust
  - "The AI Winter" – through end of century
- 21st century:
  - Computers catch up
  - Success in some application domains, including image recognition
  - Google releases TensorFlow

A Deep Learning neural network for recognizing handwritten digits



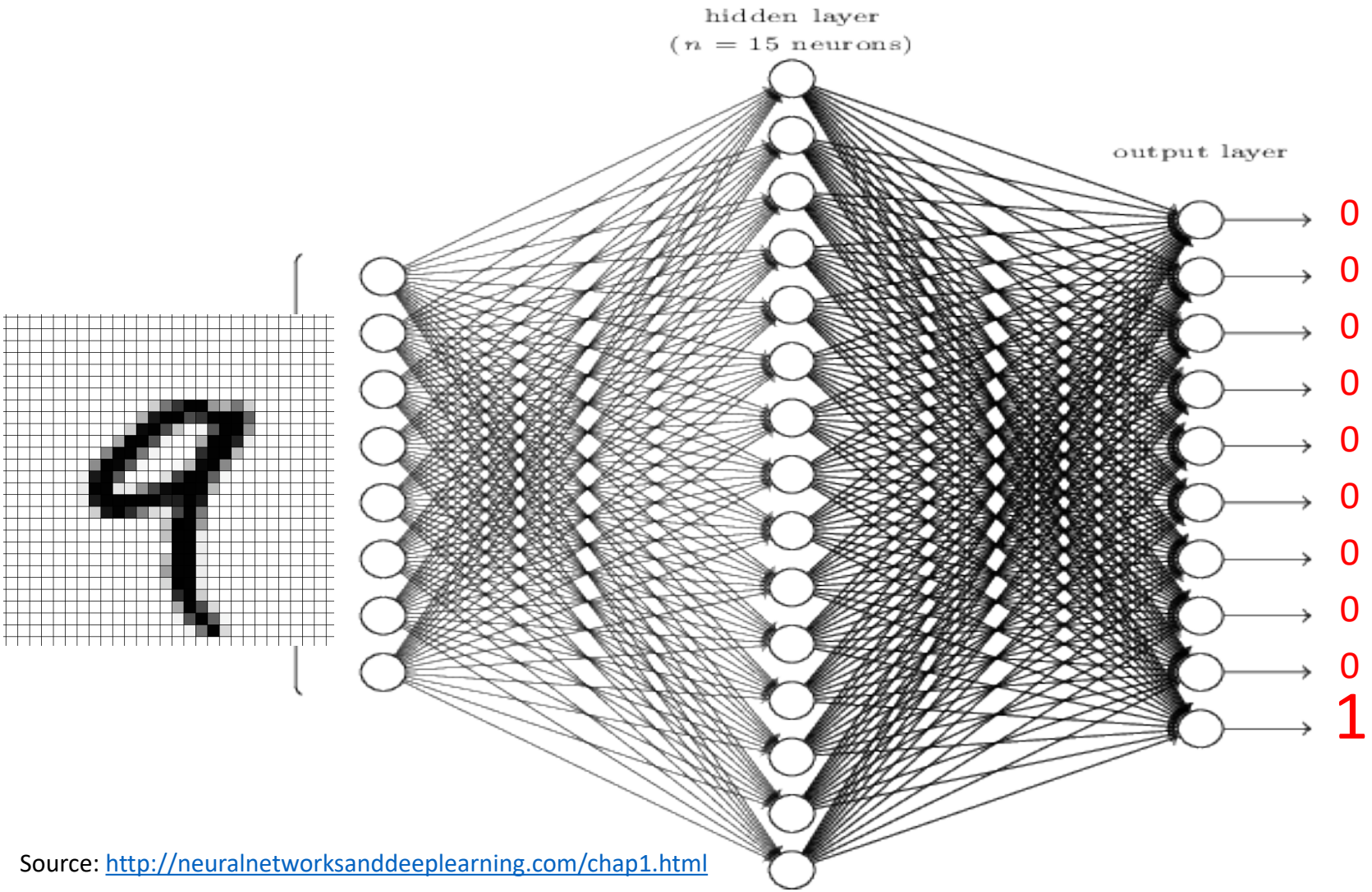
Source: http://neuralnetworksanddeeplearning.com/chap1.html

784 inputs = 28 x 28 pixels



- Convert each pixel to 0 (white) or 1 (black)
- These are the inputs to the 1st layer of neurons

What we want:



Source: http://neuralnetworksanddeeplearning.com/chap1.html

Now all we need is ...

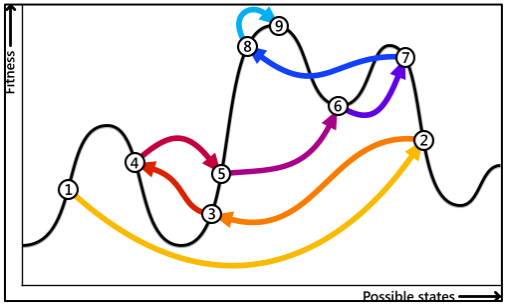
- A bias for each of 25 non-input neurons
- A weight for each of 784 inputs to each of 15 “hidden” neurons = 784 x 15 = 11,760 numbers
- A weight for each of 15 inputs to each of 10 output neurons = 150 numbers
- → Compute the optimal value for each of 25 + 11,760 + 150 = 11,935 numbers

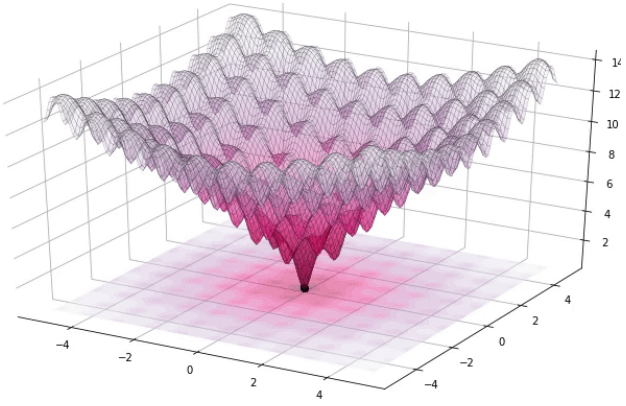
It reminds me of…

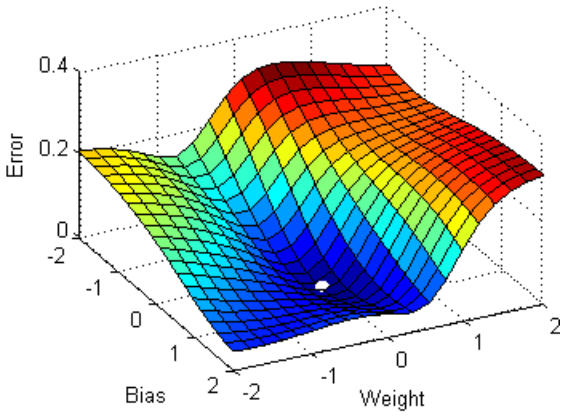
Copyright Matt Groening

But that's good! We have ways to handle this kind of problem

- Simulated annealing
- Genetic algorithms
- Stochastic gradient descent







How to train a dog to smell lung cancer byproducts in someone's breath

Dogs Can Smell Lung Cancer on Your Breath, Even If You've Just Had Lunch

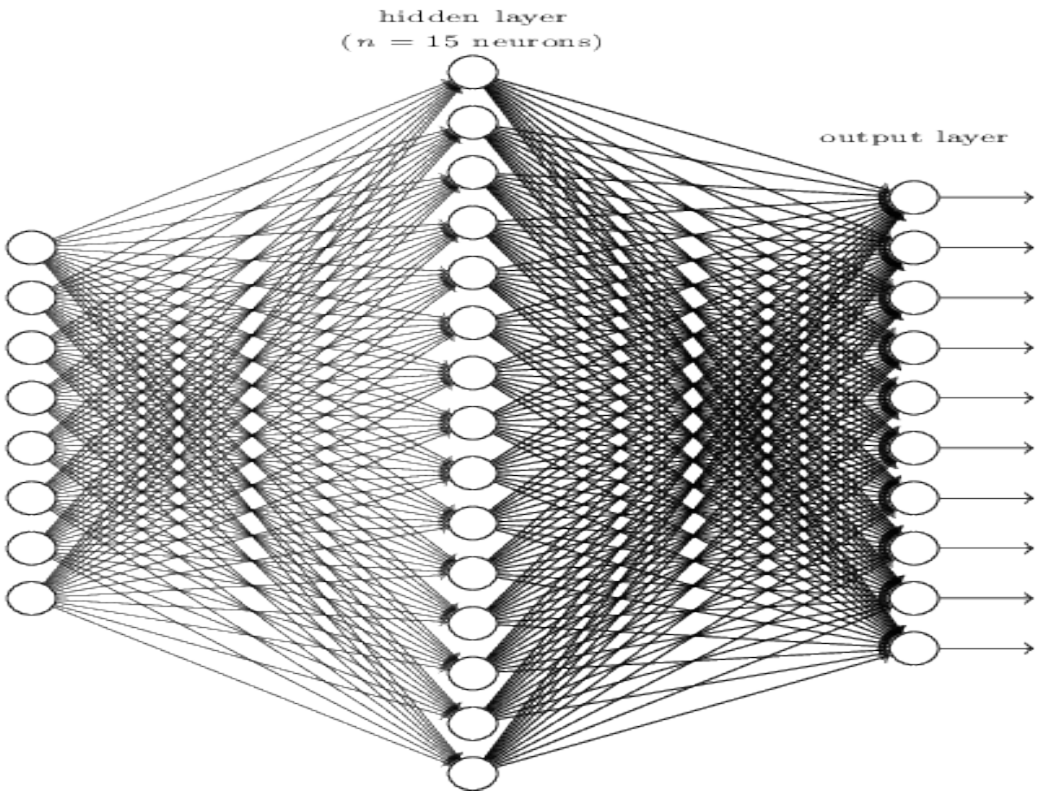
By Veronique Greenwood | August 22, 2011 12:24 pm

44

Specially trained sniffer dogs can smell something on the breath of lung cancer patients.

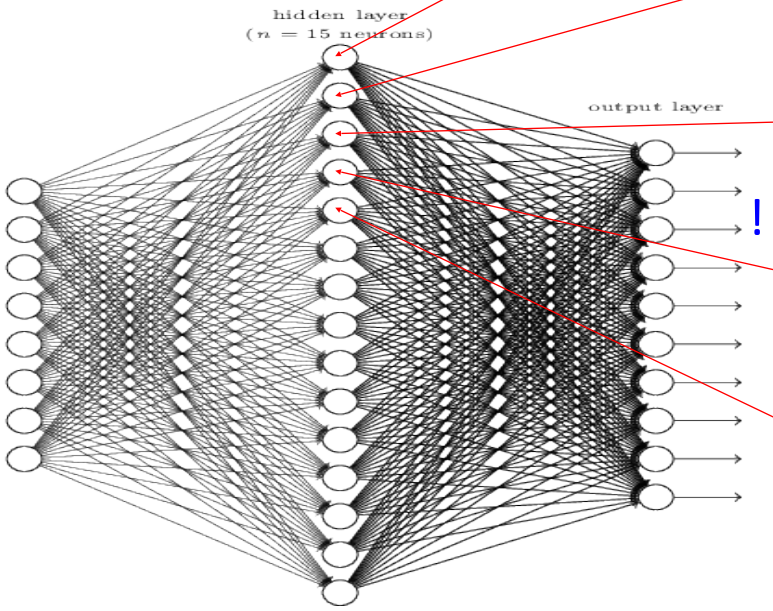
- "Discover" magazine, Aug 22, 2011
- Dog sniffs many "training samples", eventually is good at detecting that odor elsewhere

How to train a neural network to recognize handwriting

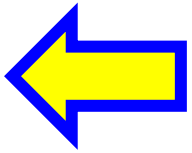
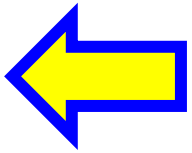


- Initialize params to random
- Expose to lots of 0s (“These are zeroes”), 1s (“These are ones”), etc
- Adjust params based on response to training examples

Just 1 of many possible training outcomes



Today's plan

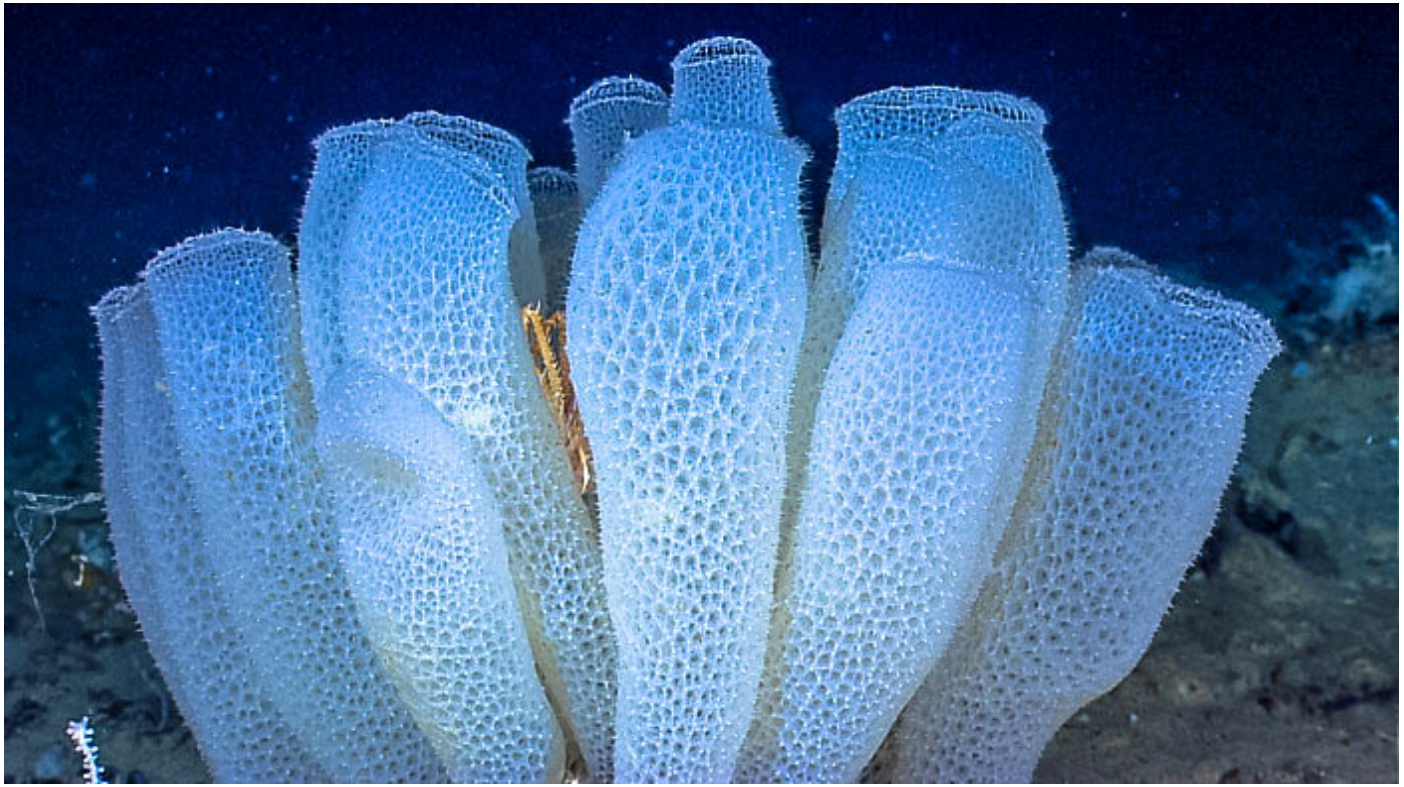
- A little context
- Deep Learning 
- Projects: 
  - Poriferal Vision
  - Data mining GenBank with simulated eyes
  - Coral Vision
  - Adverb

Today's plan

- A little context
- Deep Learning
- Projects:
  - Poriferal Vision
  - Data mining GenBank with simulated eyes
  - Coral Vision
  - Adverb

Poriferal Vision

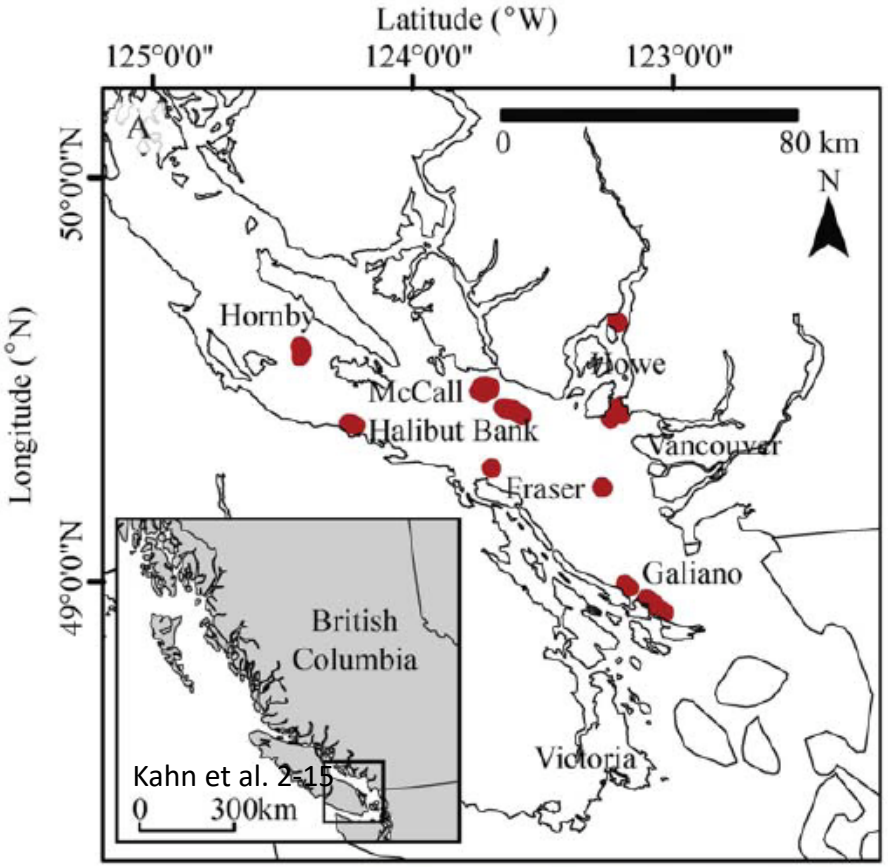
- Glass sponges: Phylum Porifera, Class Hexactinellida
  - Ecologically important
  - Especially during climate change



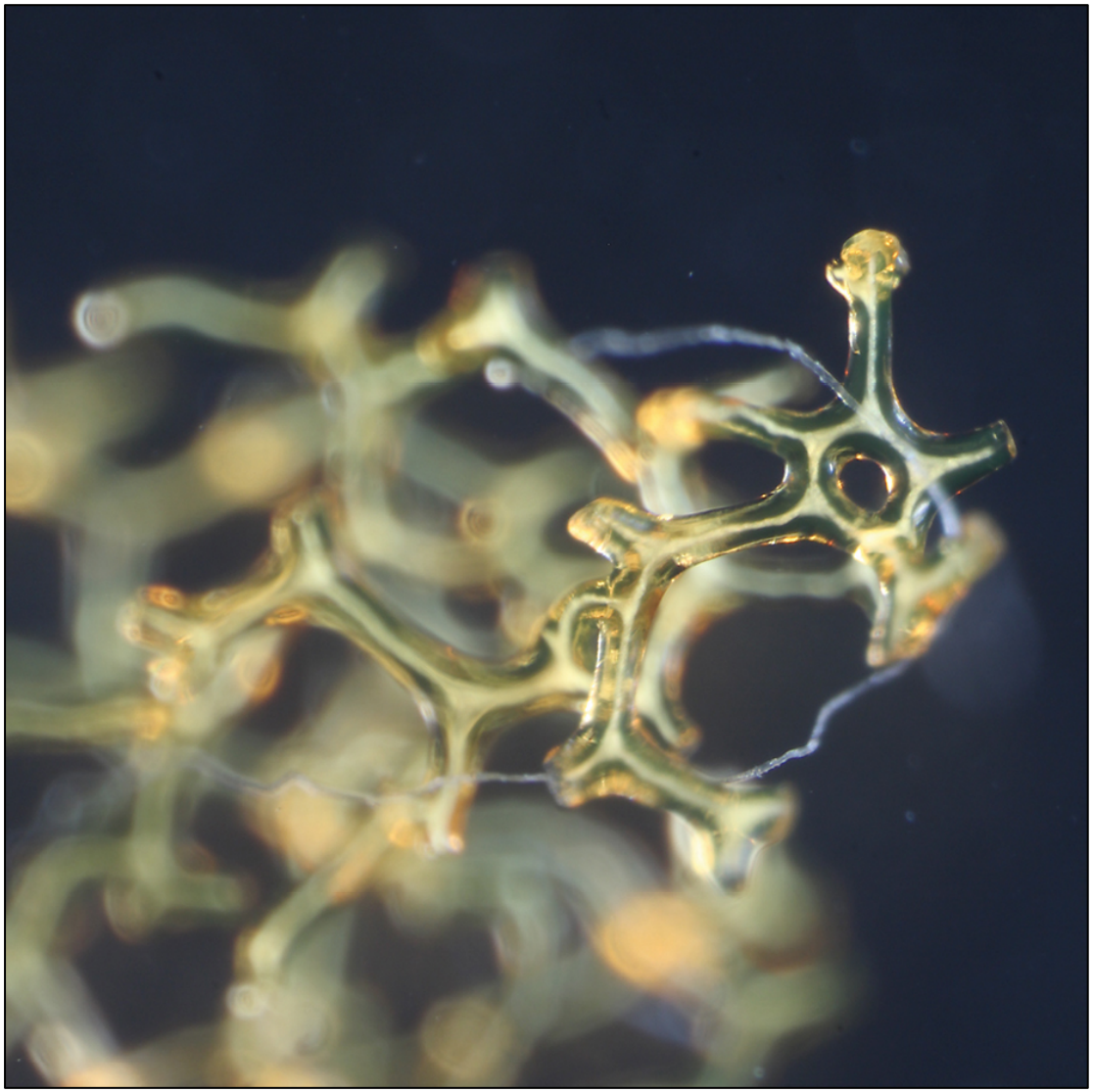
oceanservice.noaa.gov

Ecological importance of glass sponges

- Like corals: ecosystem engineers
- Grazing / water processing rates up to 10x higher than other suspension-feeding communities (Kahn et al. 2015)
- Carbon sink, ammonia source
- Form huge reefs – hundreds of square km – e.g. between Vancouver Island and mainland Canada
- https://www.youtube.com/watch?v=pTZ211cIjX8



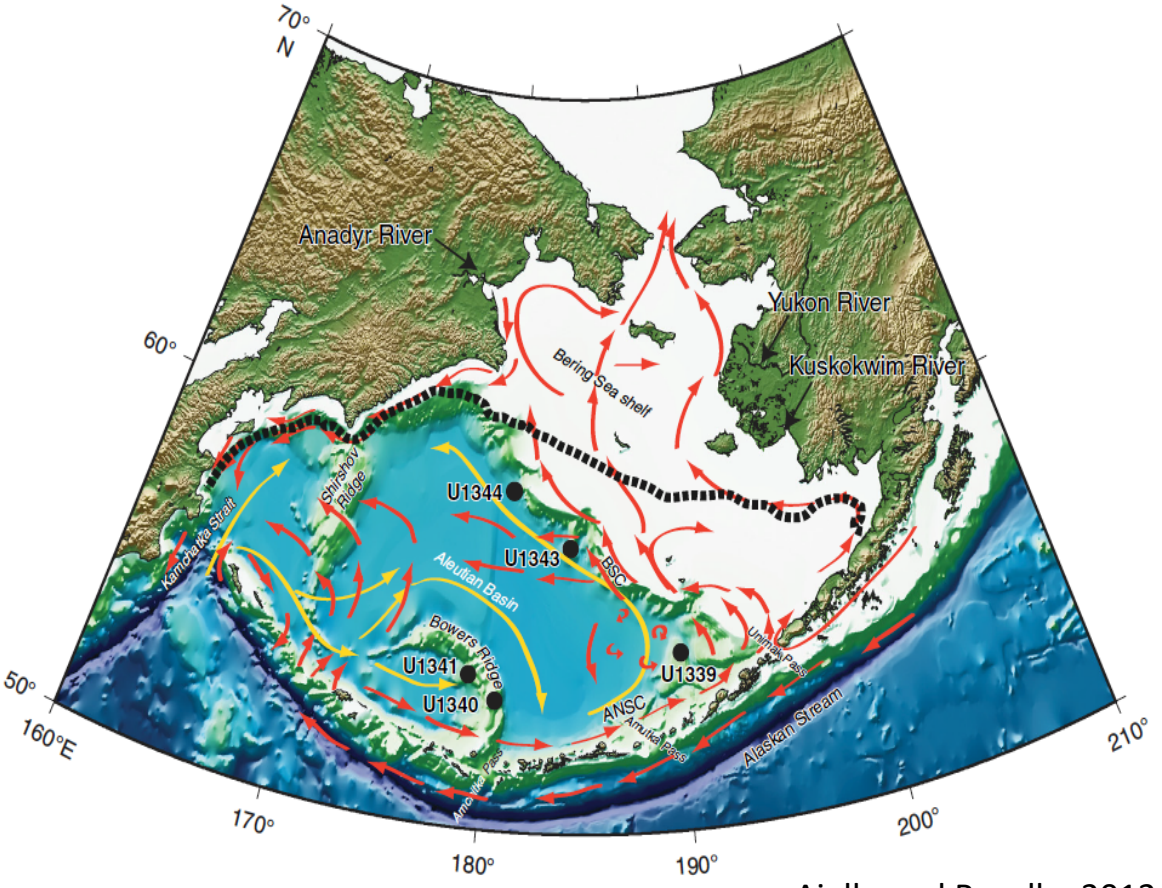
Skeletons: biosynthetic glass spicules



Vanderbilt news

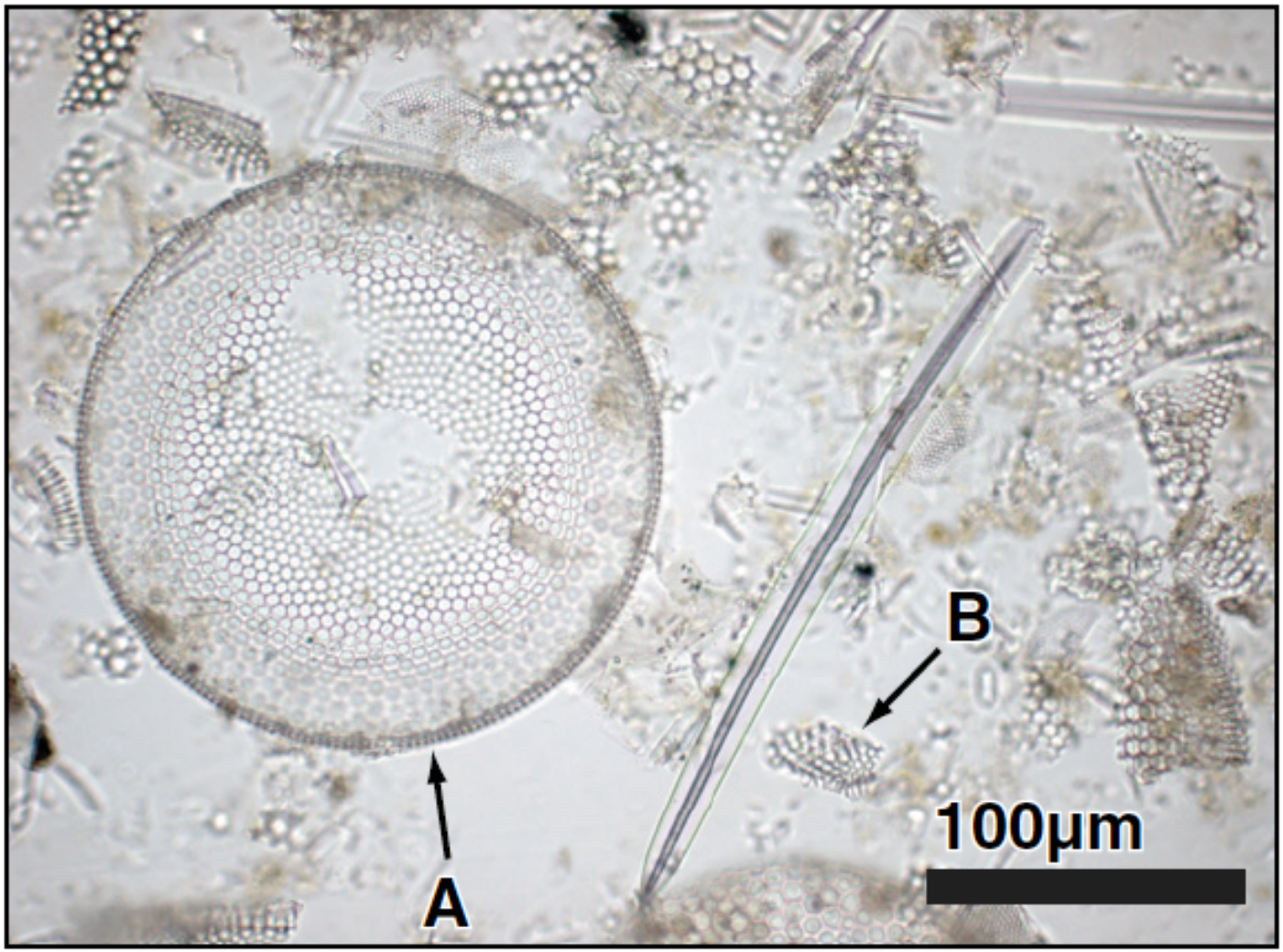
When sea temperatures rise …

- Protect the corals
- Bet on the sponges
- Bering Sea core samples tell us about sponge population change during the Early Pliocene Warm Period (~3mya)
- Need to study many core samples, each containing many spicules and many sediment grains



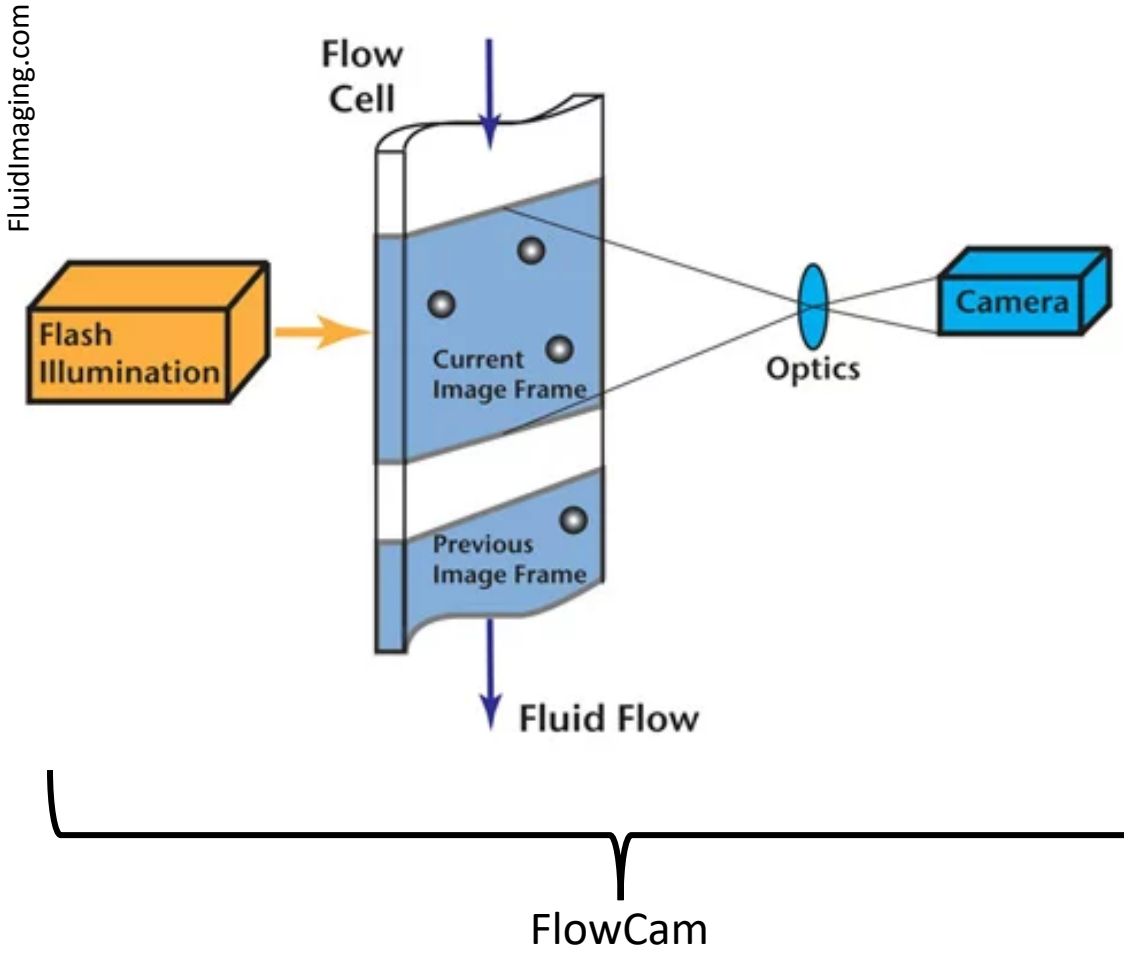
Aiello and Ravello, 2012

Too much information for manual analysis



Aiello and Ravello, 2012

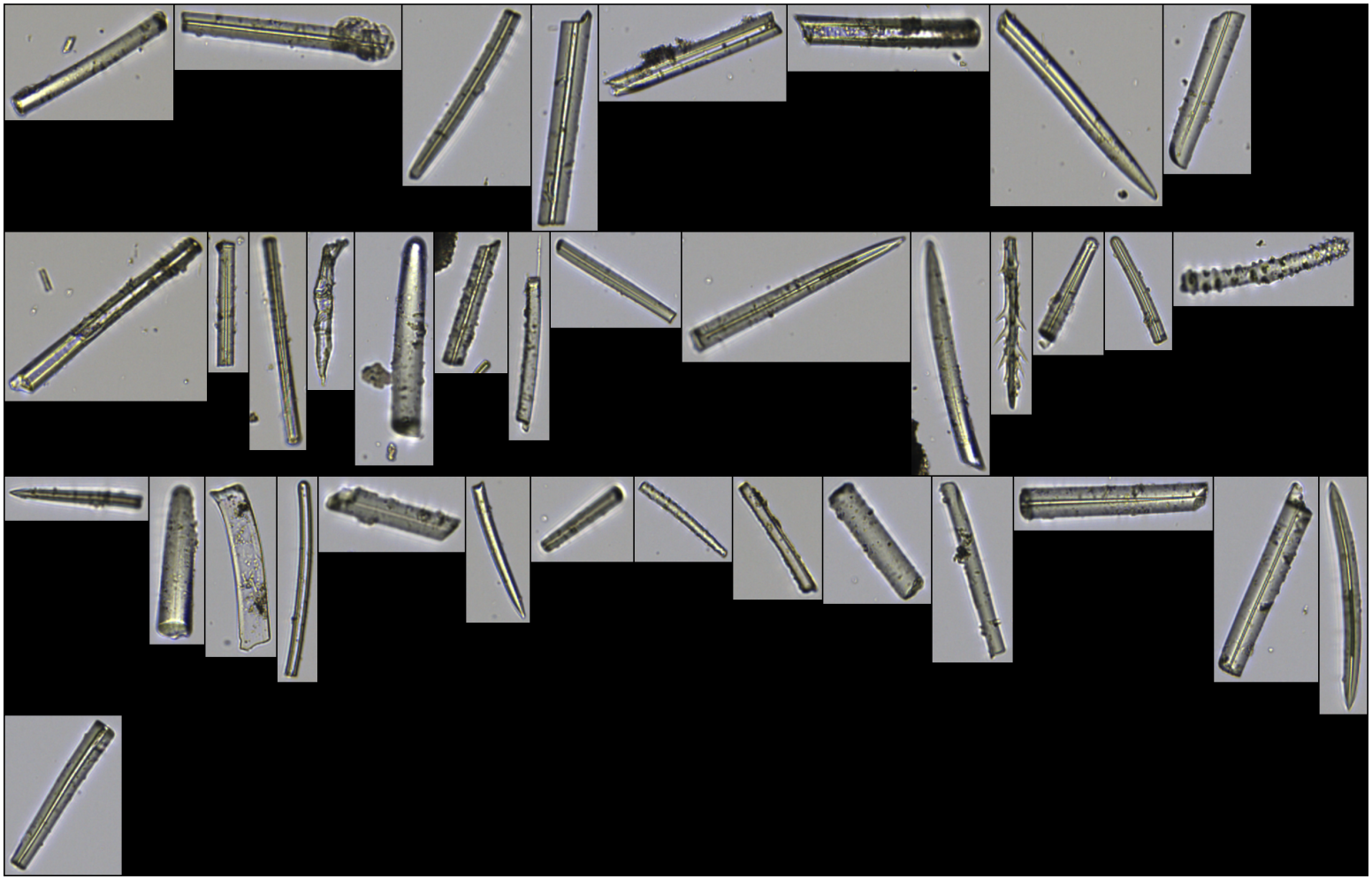
The Plan: Flow Imaging Microscopy + Deep Learning



➔ 1 photo per particle.

- Sediment
- Spicules
- Diatoms

FlowCam Collage Output

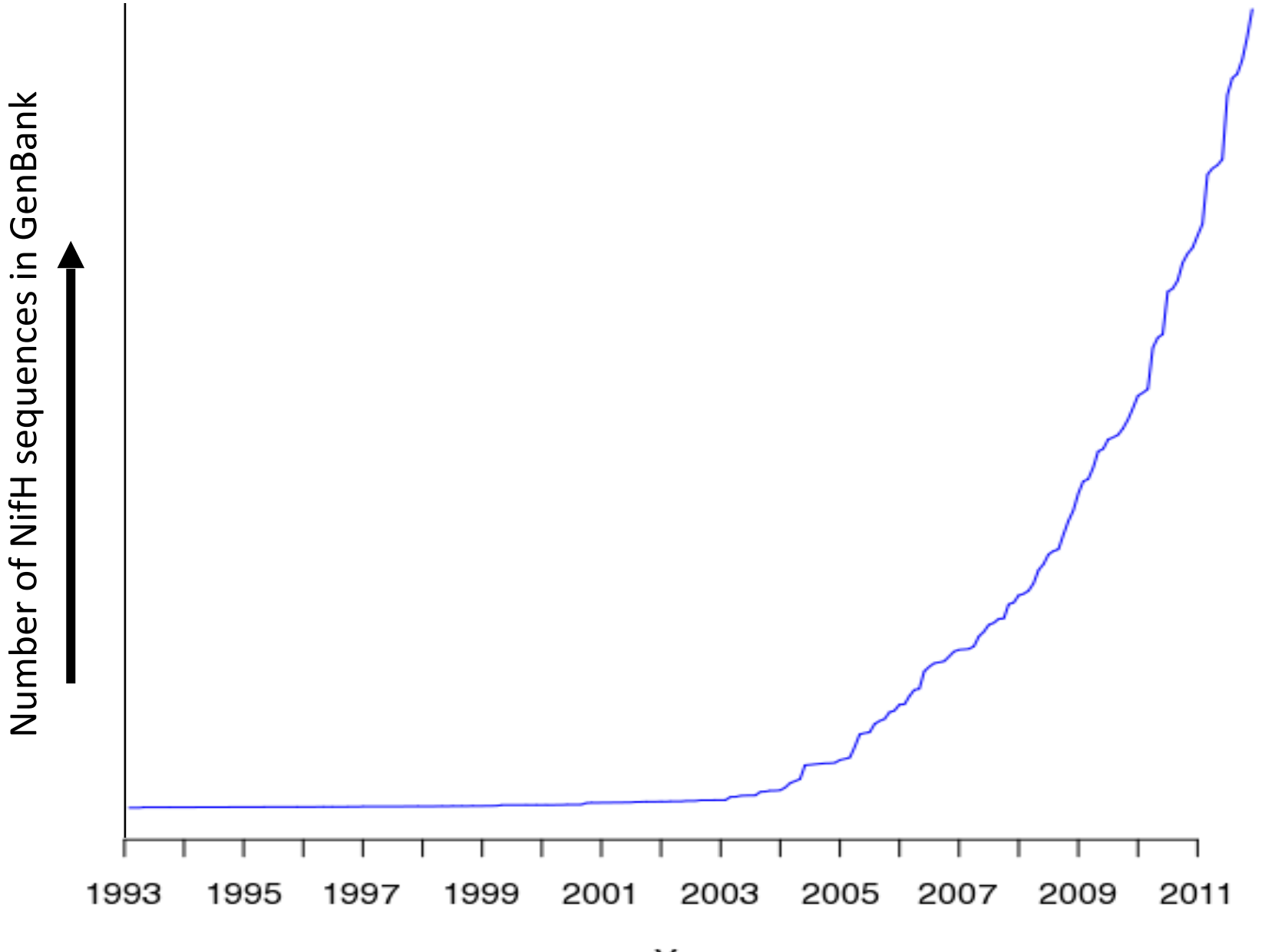


From lab analysis to the open ocean

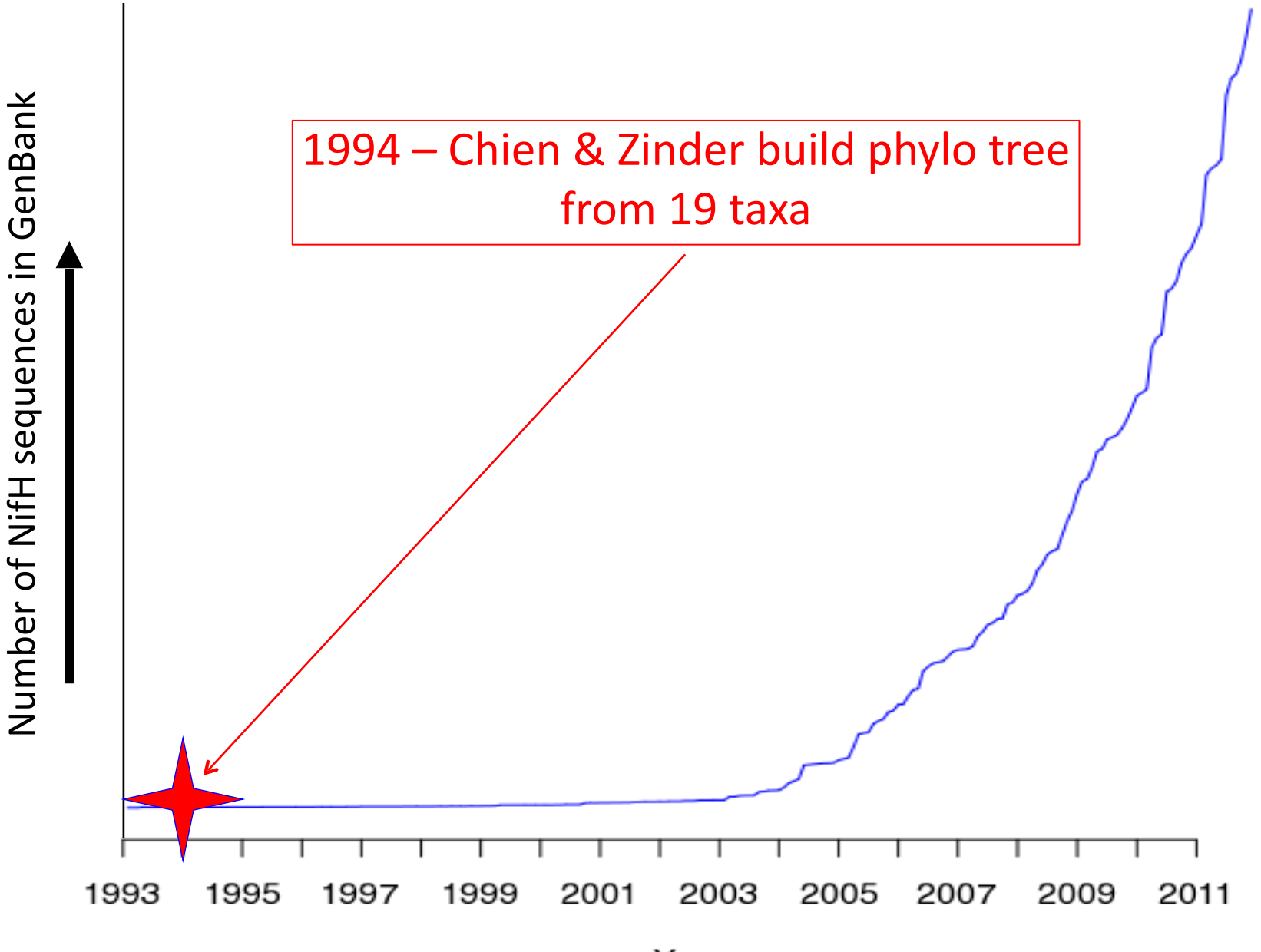
Today's plan

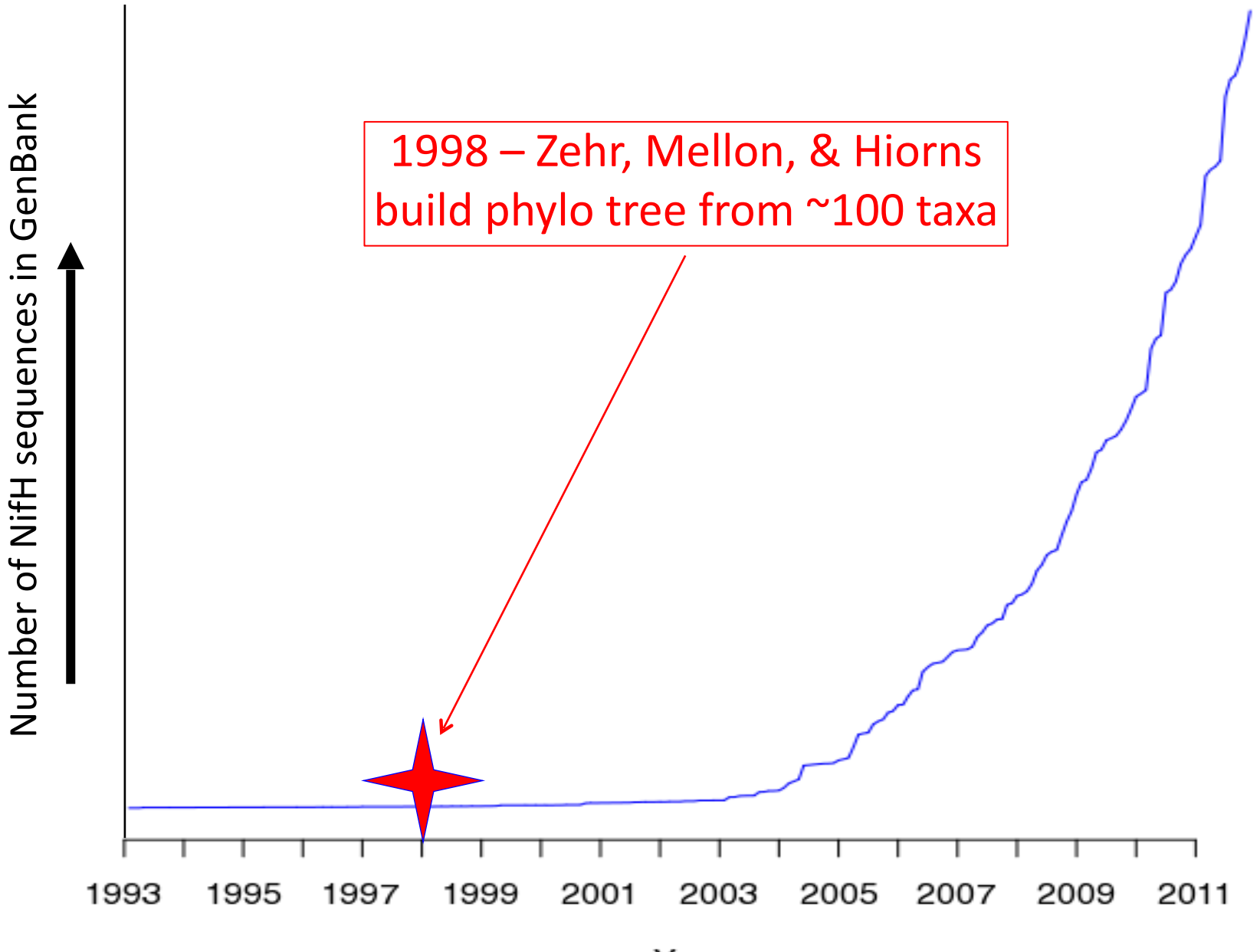
- A little context
- Deep Learning
- Projects:
  - Poriferal Vision
  - Data mining GenBank with simulated eyes
  - Coral Vision
  - Adverb

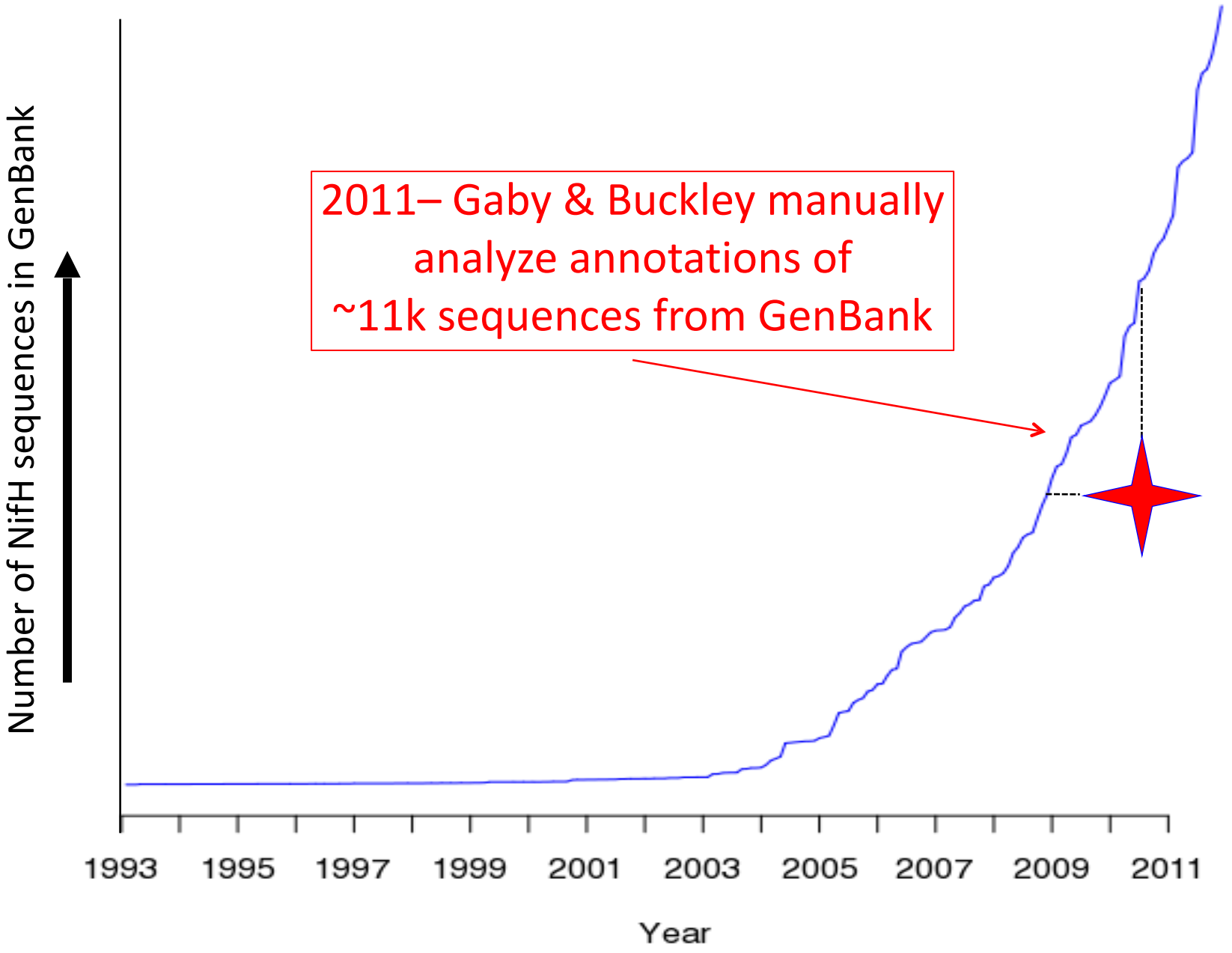
nifH Science: Victim of its own success



NifH Science: Victim of its own success




Number of NifH sequences in GenBank
1998 – Zehr, Mellon, & Hiorns
build phylo tree from ~100 taxa
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011


Number of NifH sequences in GenBank
2011– Gaby & Buckley manually analyze annotations of ~11k sequences from GenBank
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
Year

The *nifH* Catastrophe

- Manual curation no longer tractable
  - Can't keep up with growth rate of nifH sequences
  - Only as good as the annotations
- Automated curation (e.g. Fungene HMMs) wasn't sensitive or specific enough
  - Fungene F+ rate for *nifH* ~= 6%
- Needed a better algorithm to drive automated curation: ARBitrator
  - Only looks at sequence
  - Based on "superiority" of similarity to a *nifH* conserved domain
  - Very low error rates
  - Tuning classifier parameters took ~3 weeks

ARBitrator and CO-ARBitrator

	ARBitrator	CO-ARBitrator
Gene	nifH	COI
# of sequences recovered from GenBank	34,420	1,054,973
Time to tune params	3 weeks	4 months
Error rates (F+, F-)	.033%, undetectable	.0034%, .0018%

☹

☺

Can Deep Learning improve effeciency?

Yes: Error rate ~= 35%

Can I beat 35% error rate? ;-)

- Very little research on Deep Learning of nucleotide seqs
- Lots of research on images
  - What kinds of layers
  - How many layers
  - Training regimes
- ➔ Let's convert each nucleotide sequence to a jpeg: We'll analyze images of sequences!

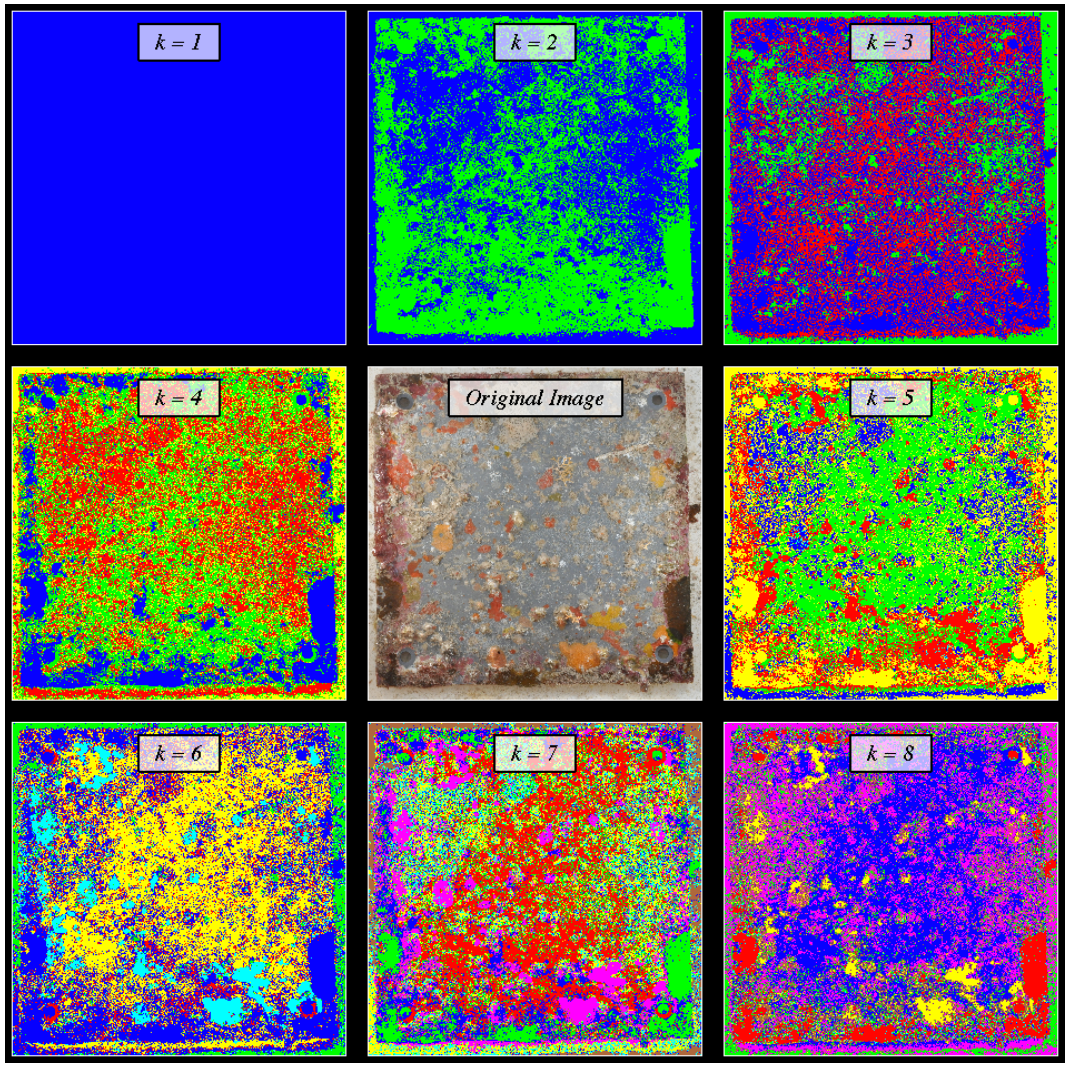
And that's working pretty well

- ~98% accuracy
- 100x worse than ARBitrator, but definitely respectable for a first shot
- ~ 1 day to train the model
  - Eliminates weeks (ARBitrator) or months (CO-ARBitrator) of human ground work
- Article submitted: 2 faculty authors, 2 student authors

Today's plan

- A little context
- Deep Learning
- Projects:
  - Poriferal Vision
  - Data mining GenBank with simulated eyes
  - Coral Vision
  - Adverb

Coral Vision: Applying Deep Learning analysis to ARMS plate photographs

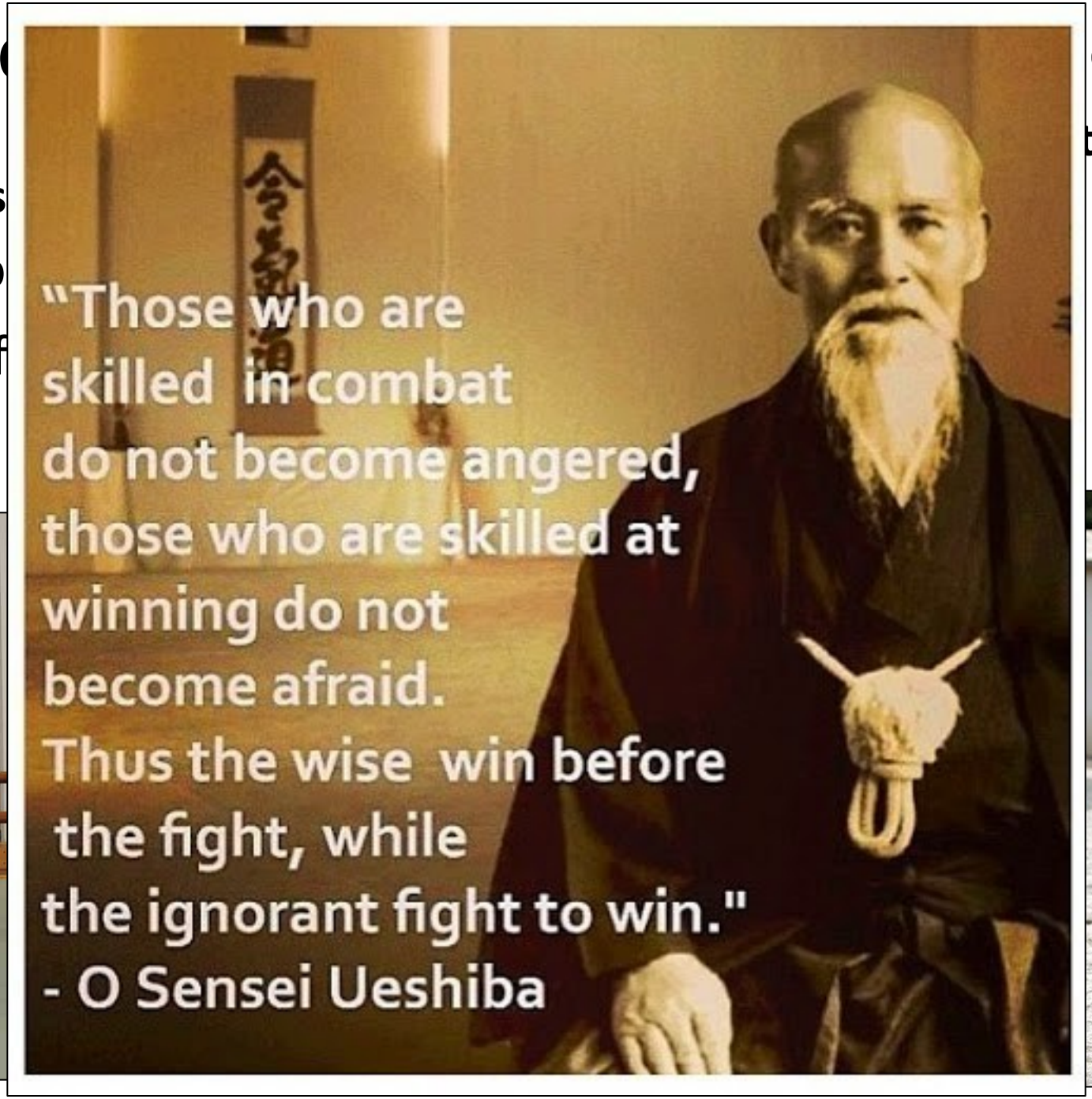


2 kin em

ting

"Clas

- Po
- nif



Coral Vision: What we have

- ~14,000 plate photos
- Metadata
  - Ocean temperature
  - Acidity
  - Depth
  - Longitude, latitude

Coral Vision: The goals

- Segment the images
- Cluster and identify individuals
  - Minimal impact on expert taxonomists
- Compute statistics
  - Population
  - Diversity
- Correlate with metadata, including DNA
- Public website for exploring, analyzing, and contributing data

Today's plan

- A little context
- Deep Learning
- Projects:
  - Poriferal Vision
  - Data mining GenBank with simulated eyes
  - Coral Vision
  - Adverb

Adverb: Ad-hoc Viterbi

- 1990s: computers were just not all that
  - Slower
  - Less memory
  - Viterbi algorithm is O(nstates^2 * seqlen)
    - Time and memory
- Original bioinformatic HMMs were painstakingly designed and trained
  - Intended for heavy re-use
- 2020s: computers are all that
  - Your protein HMM app builds an HMM in ~1 sec
  - Single-use (“ad-hoc”) HMMs are a possibility

COI barcoding

- COI = Cytochrome C oxidase, subunit 1
- All animal species have it
- ”Barcode of animal life”
  - Unique across almost all animal species
- To identify a sample (e.g. blood, hair, tissue, ARMs plate)
  - Extract DNA
  - Amplify COI using primers
  - Sequence and blast
- The BOLD database
  - Vouchered
  - Stringent acceptance criteria

COI barcoding

- If the species of the tissue has previously been identified …
  - It's in your blastable database (BOLD, CO-ARBitrator, or GenBank)
  - You'll get a perfect hit
  - Or maybe slightly imperfect, due to amplification or sequencing error, but there's still exactly 1 lowest E-value hit
  - → reliable identification
- If the species is previously unknown
  - It can't possibly be in your database
  - You can't possibly identify its species
  - Reasonable expectation: genus of best hit will be correct
  - But no!

COI barcoding and novel species

- The barcode concept was intended only for known species
- If query is novel, can we use the taxonomy of the best blast hit in some way?
  - P(correct phylum/class/order) ~= 100%
  - P(correct family) = 67%
  - P(correct genus) = really bad
- Many metagenomic/ARMS studies use COI identification
  - Sample contains both known and unknown species
  - Need an algorithm that can handle both cases

The Adverb Algorithm

- blastn your sequence against BOLD database
- Perfect or near-perfect hit?
  - ➔ previously known, accept species identification
- Imperfect hit?
  - Accept blast's order identification
  - Build an HMM for every family in the order
  - Compute Viterbi score of sequence on every family HMM
  - Highest Viterbi score indicated family identification
  - 90% accuracy
  - No reliable genus identification
- 18 months * 15 nodes * 28 cpus = 630 CPU-years