Using Deep Learning for DNA classification tasks

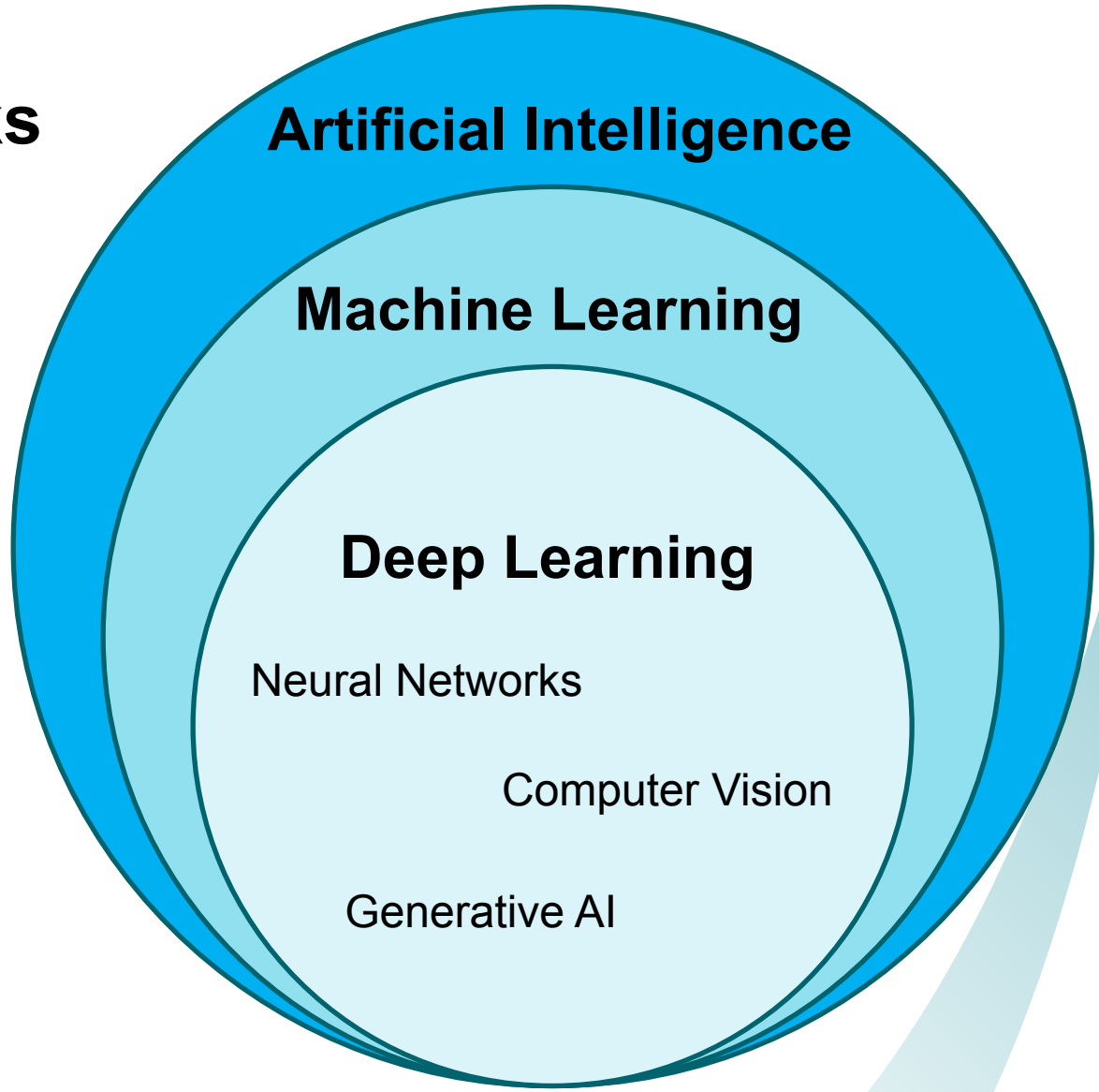
Chris Powers

SEDNA Office Hours

20 November 2025

Machine learning and deep learning as statistical frameworks

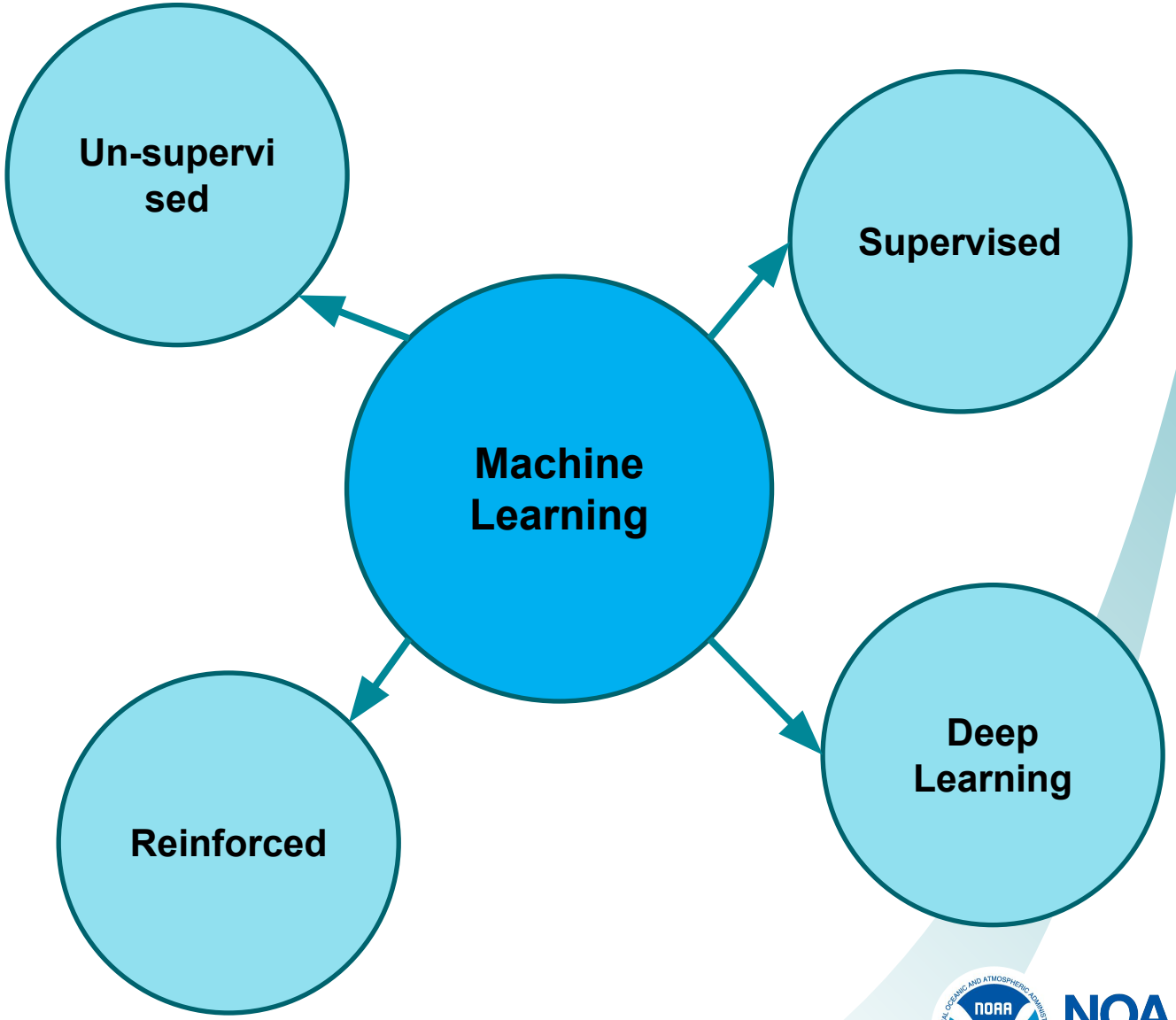
- **Artificial intelligence (AI)** is the broadest framework, defined as creating so-called “intelligent machines”
- **Machine learning (ML)**, a subset of AI, allows intelligent machines to identify patterns and trends without explicit direction
- **Deep learning (DL)**, a subset of ML, allows for the handling of complex tasks and unstructured data



NOAA FISHERIES

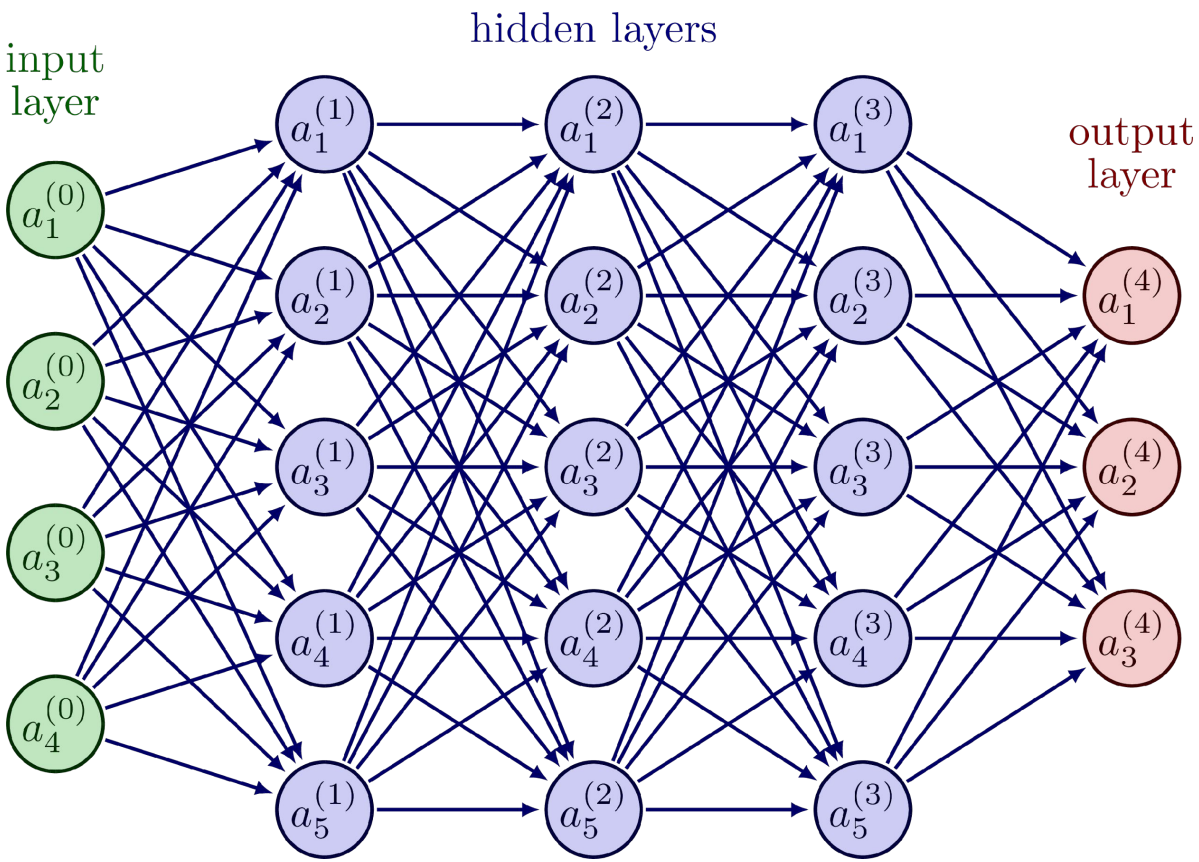
ML may seem inaccessible, but most of us use it in science without realizing

- ML is broadly considered to be highly complex, but this is not necessarily the case
- Simple examples of ML are part of every biologists statistical toolkit:
  - Simple linear regression
  - Generalized additive models
  - Classification trees
  - Maximum likelihood analysis
  - Clustering
- Today, deep-learning is the focus
  - Convolutional neural networks



NOAA FISHERIES

Deep learning models often rely on a neural network structure

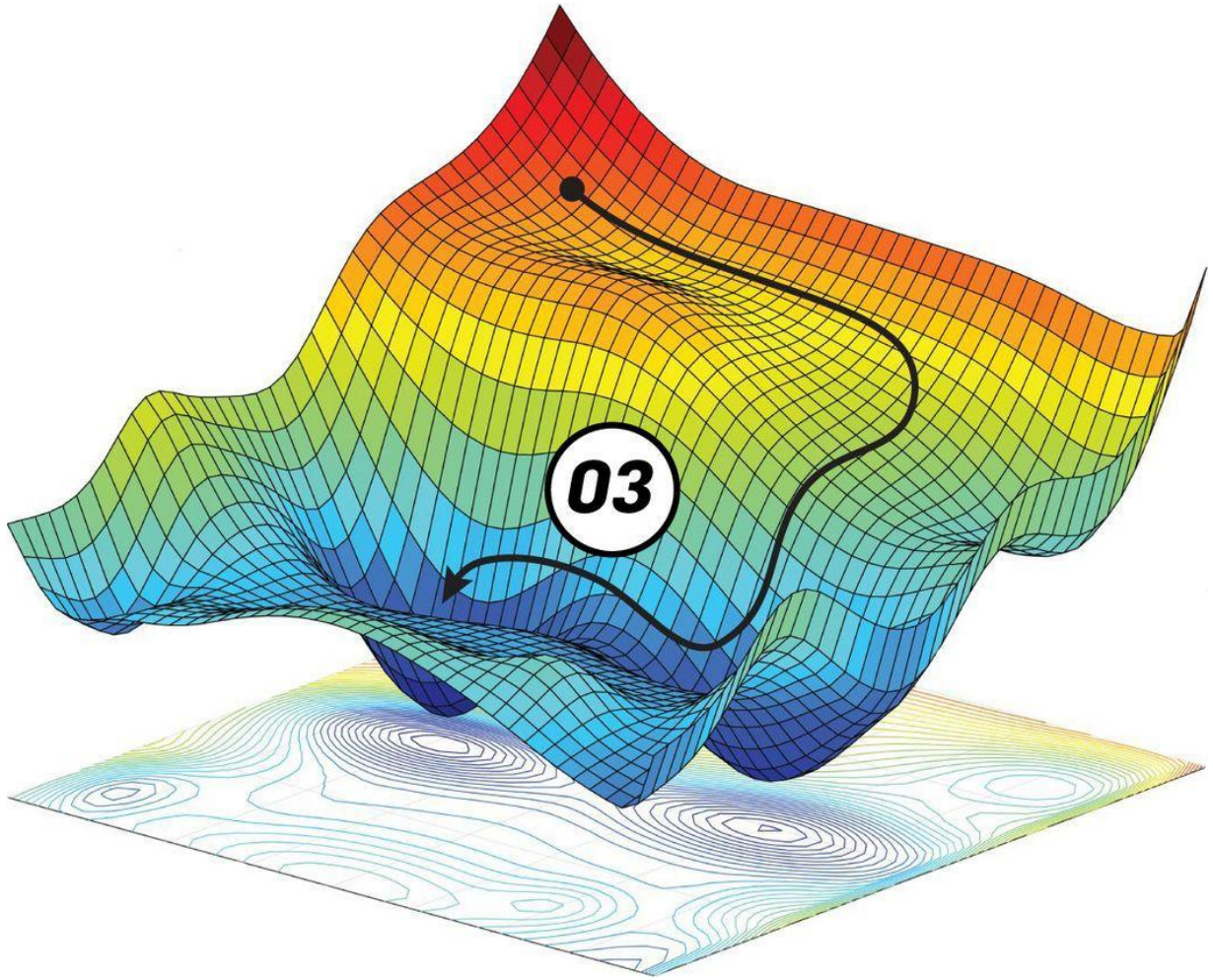


https://tikz.net/neural_networks/

- Unstructured input data in some tensor format
- Tunable parameters in “hidden layers”
- Output layer relatable to question asked of model (e.g. number of classification labels)
- Number of hidden layers is proportional to capacity to learn
- Iteratively **tune** parameters over many **epochs**, while monitoring accuracy and confidence

NOAA FISHERIES

Deep-learning models iterate to find local minima



https://blog.paperspace.com/part-3-generic-python-implementation-of-gradient-descent-for-nn-optimization/

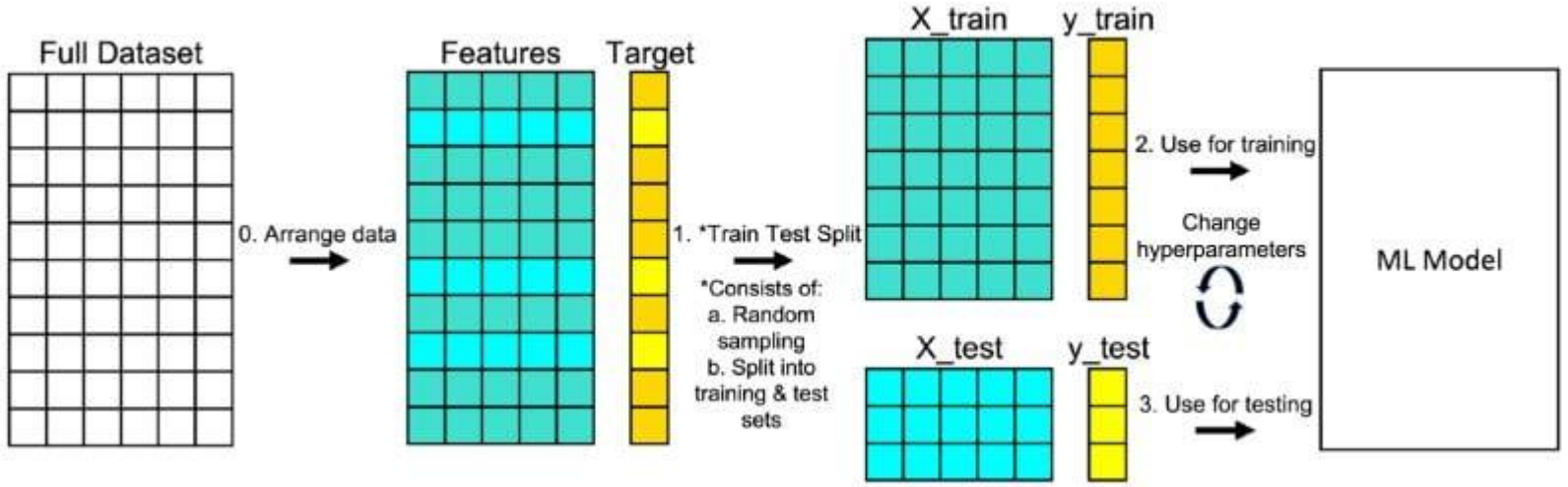
- Apply gradient descent (or something like it) to minimize errors
  Note, actual framework is an nth dimensional space
- Every step represents iteratively updated weights on neurons in a network
- Once the model stops improving, or reaches a minima, the model is done training

NOAA FISHERIES

Deep-learning models require best practices in data analysis, including training/validation splits

- Deep learning models have an enormous capacity to learn!
- Correlates with a large capacity to **memorize**
- The perfect end in a gradient descent would be memorization of the dataset
- Ultimate goal: learning trends in underlying data to **generalize**
- How do we achieve this?

  Assess performance on unseen data, not training data



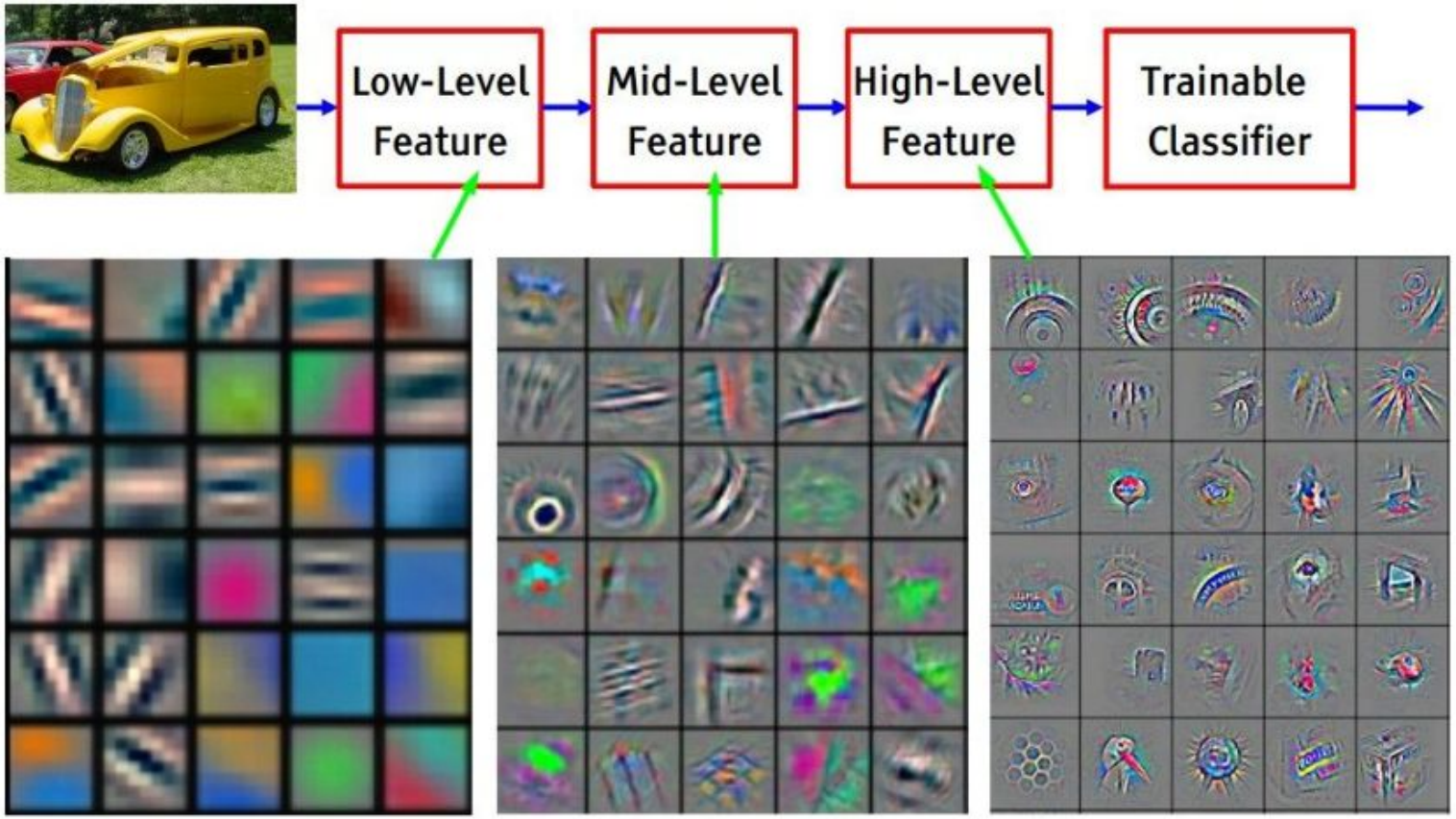
NOAA FISHERIES

NOAA FISHERIES

Convolutional Neural Networks

Convolutional neural networks (CNNs) extract abstract features that are useful for classification (Computer vision)

A convolutional neural network is an older type of deep learning model (1990s), designed for processing grid-like data such as images, that learns features by applying a set of filters or kernels to the input data

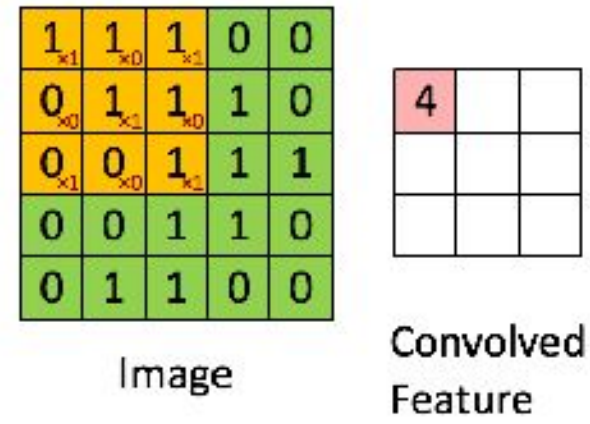


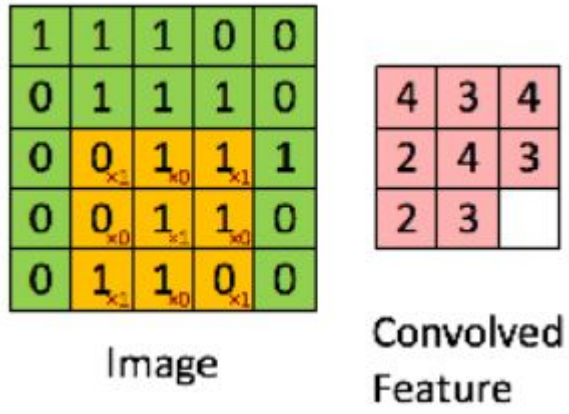
Example of image classification convolution
From: https://i.sstatic.net/bN2iA.png

NOAA FISHERIES

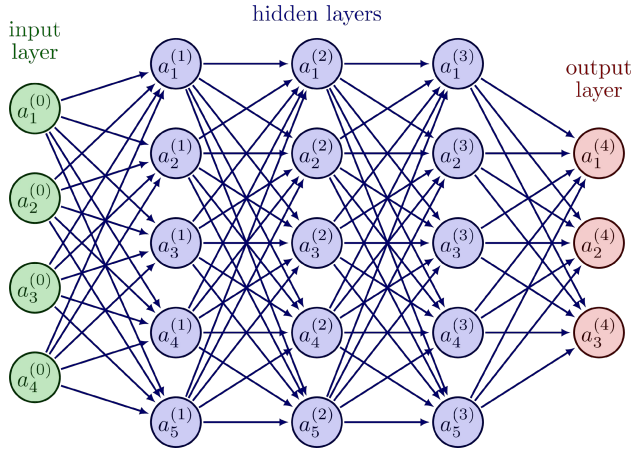
Convolution functions apply a sliding window approach to transform the data

- Example: Slide a kernel (or filter) over an array of data and perform a dot product multiplication with the kernel
  Extracts unique features of the kernel
- Isolates a local signal across an image in the hidden layer part of the network
- This can train the neural network on emergent trends





2-Dimensional convolution. From: https://ujjwalkarn.me/wp-content/uploads/2016/07/convolution_schematic.gif?w=268&h=196

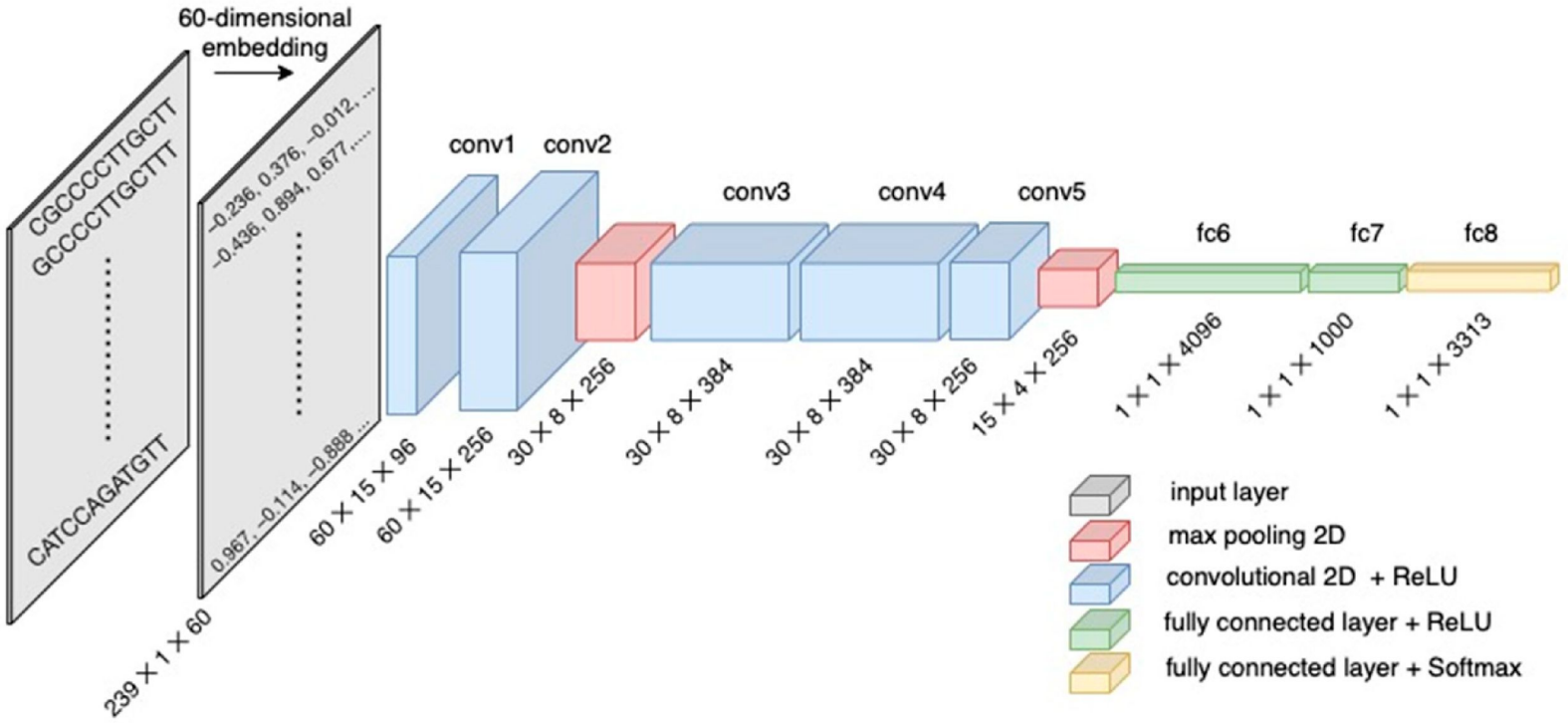


https://tikz.net/neural_networks/

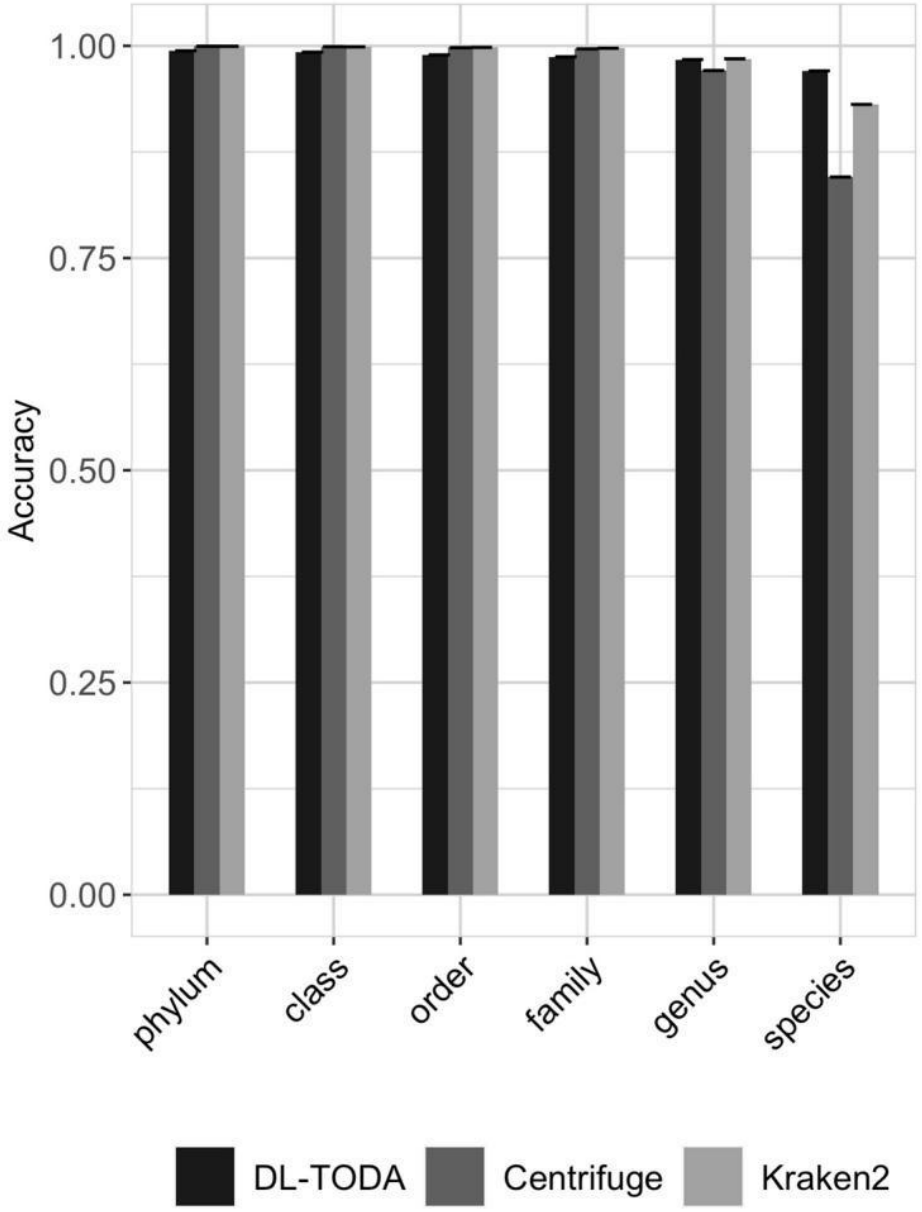
NOAA FISHERIES

Why use an image classification technique for DNA?

- CNNs first used on DNA in 2015
- DNA sequences contain a wide variety of conserved motifs and domains that encode functional and taxonomic information
- The convolutional feature extraction process isolates motifs and domains



CNN architecture of a DNA model, DL-TODA
From: Cres *et. al.* 2023, *Biomolecules*



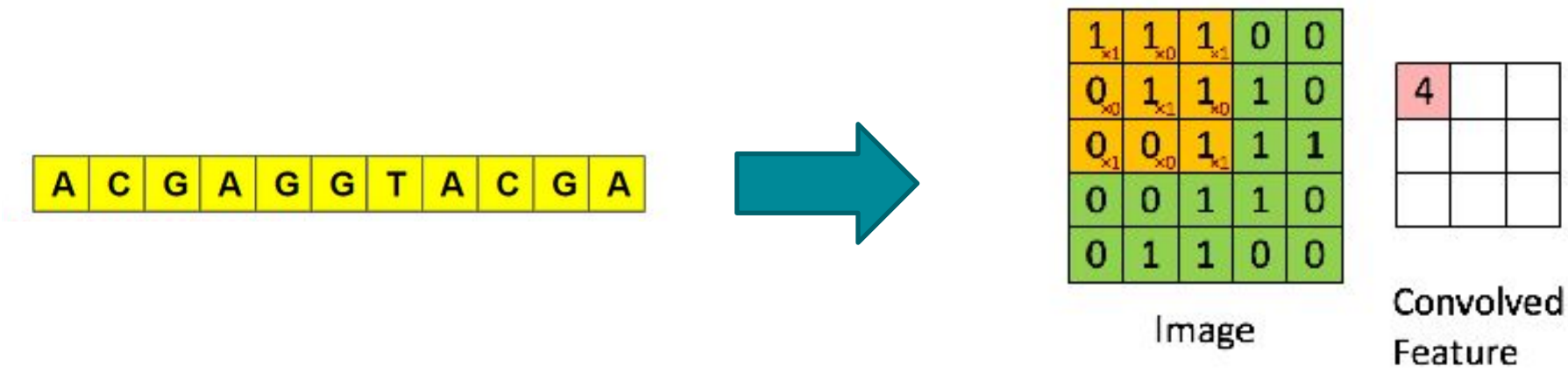
Convolutional Neural network approach out-performs gold-standard tools for taxonomic assignment at species level

- The convolutions enable the model to pick up on obscure, hidden signals

CNN architecture of a DNA model, DL-TODA
From: Cres *et. al.* 2023, *Biomolecules*

To prep DNA for classification, it needs to be embedded into a convolvable tensor

- DNA sequences are not inherently readable by convolutional neural networks
  Convert the data to a machine-readable tensor
- Common techniques include:
  One-hot encoding
  K-mer tokenization

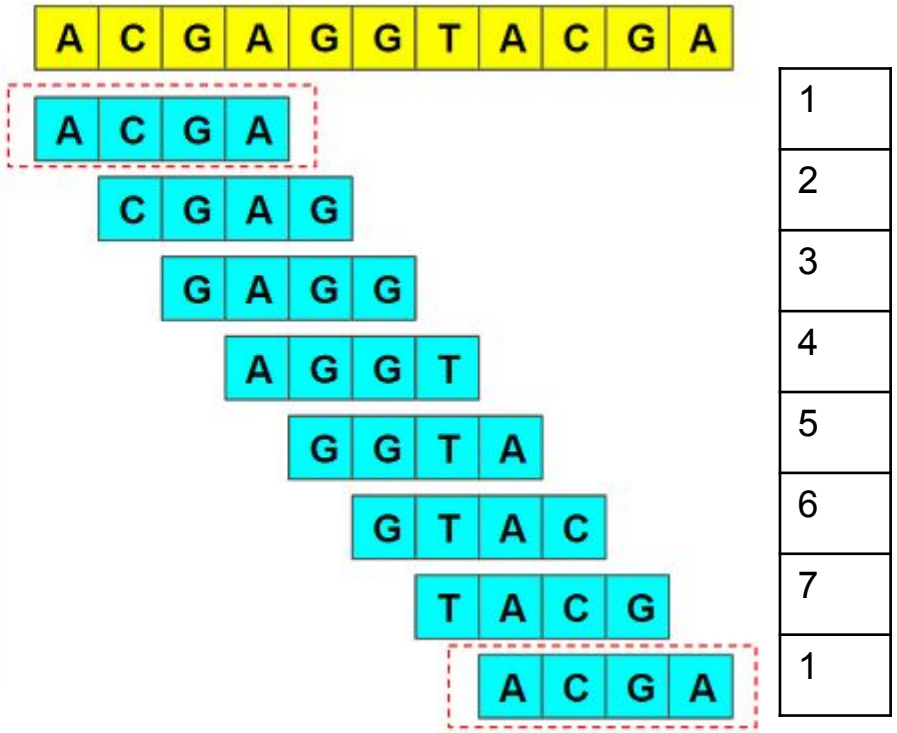


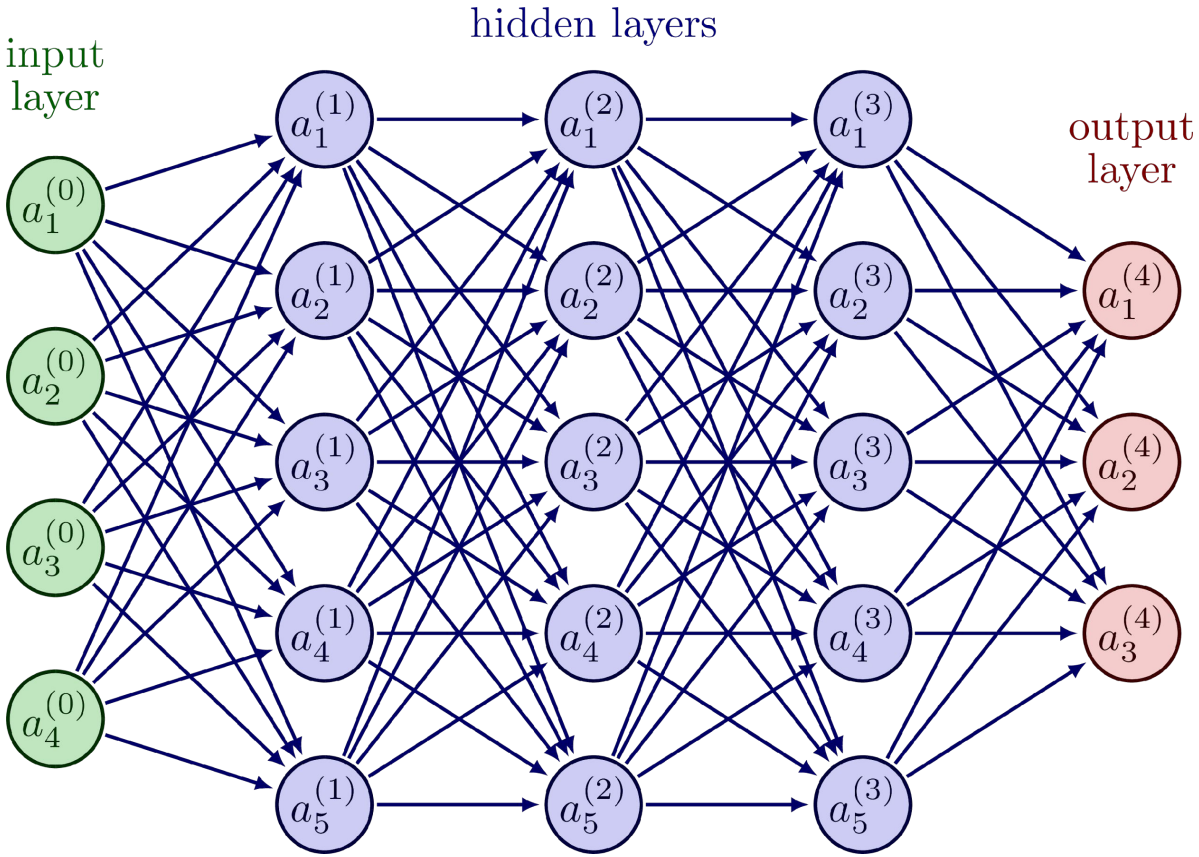
2-Dimensional convolution
From:
https://ujjwalkarn.me/wp-content/uploads/2016/07/convolution_schematic.gif?w=268&h=196

To prep DNA for classification, it needs to be embedded into a convolvable vector

- DNA sequences are not inherently readable by convolutional neural networks
  Convert the data to a machine-readable vector
- Common techniques include:
  One-hot encoding
  K-mer tokenization

A	T	G	C	-	A
1	0	0	0	0	1
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0





https://tikz.net/neural_networks/

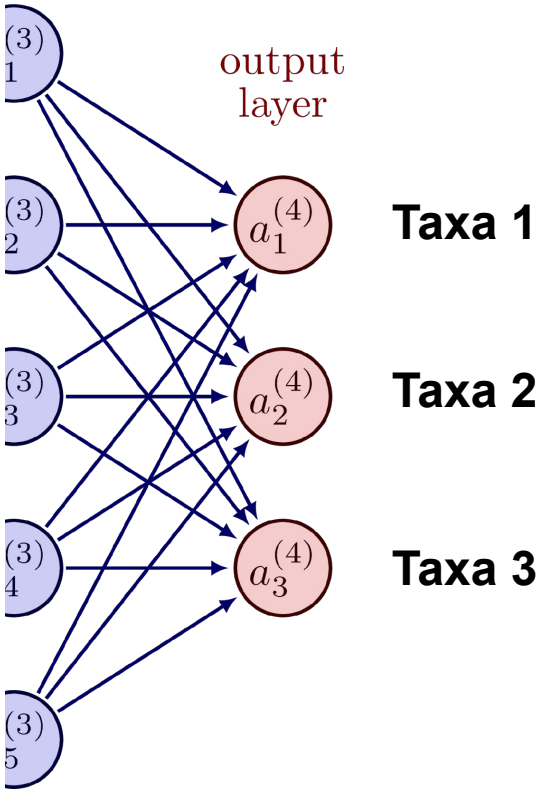
- The input vector is defined as the green input layer
- The hidden layers are defined by the size of the model
  The convolutions happen in hidden layers
- The output layers are defined by the **task**

NOAA FISHERIES

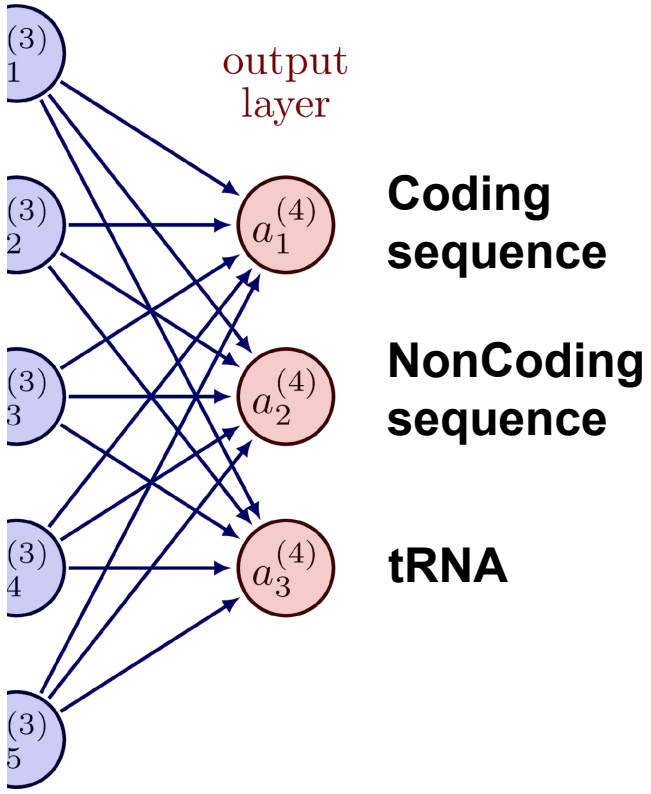
Defining a meaningful ML task for the CNN

- The output layer will be a **vector of probabilities** associated with a label

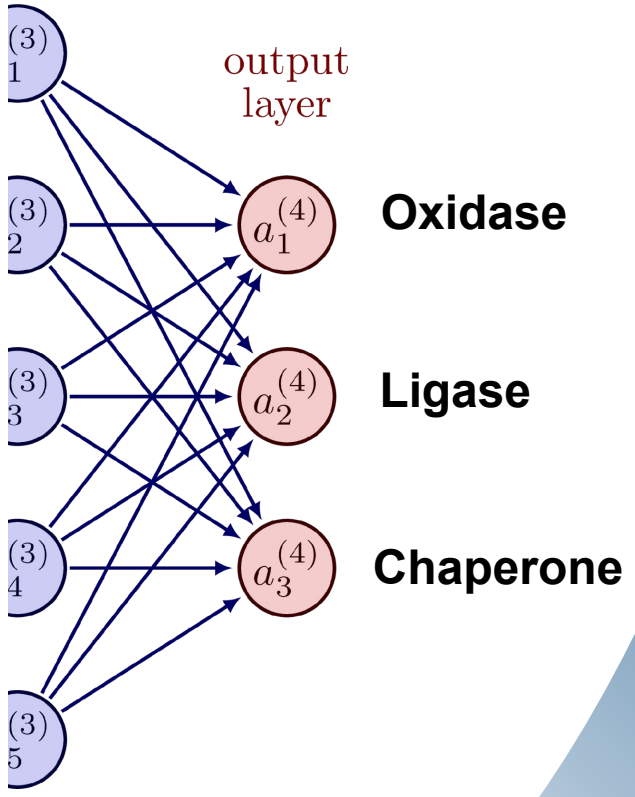
Taxonomic Prediction



Gene Calling



Protein Function Prediction



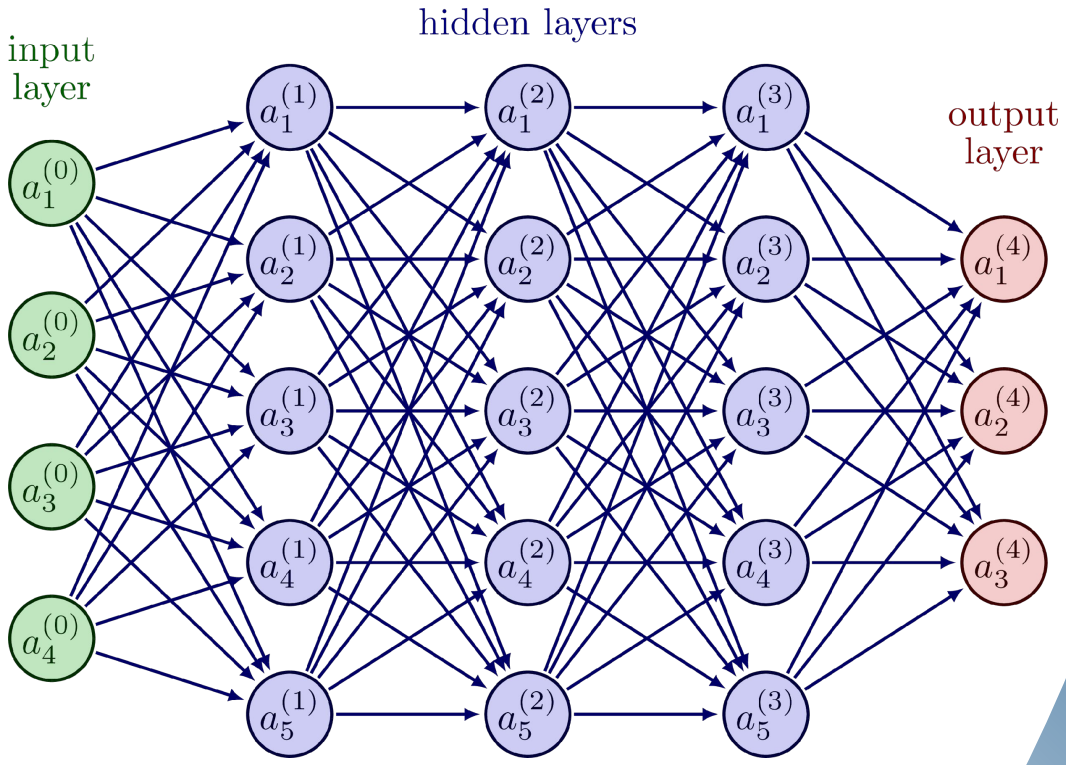
The model trains iteratively to maximize accuracy (minimize loss) over *n* epochs

NOAA FISHERIES

Overall: CNNs offer a lot of flexibility to deeply analyze trends in DNA

- Older technique, but reliable
- Visualize hidden trends through convolutions
- Classification output can be highly flexible
- Complexity of the tasks correlates with the size of the network correlates with capacity to learn

- **Questions on CNNs?**



https://tikz.net/neural_networks/

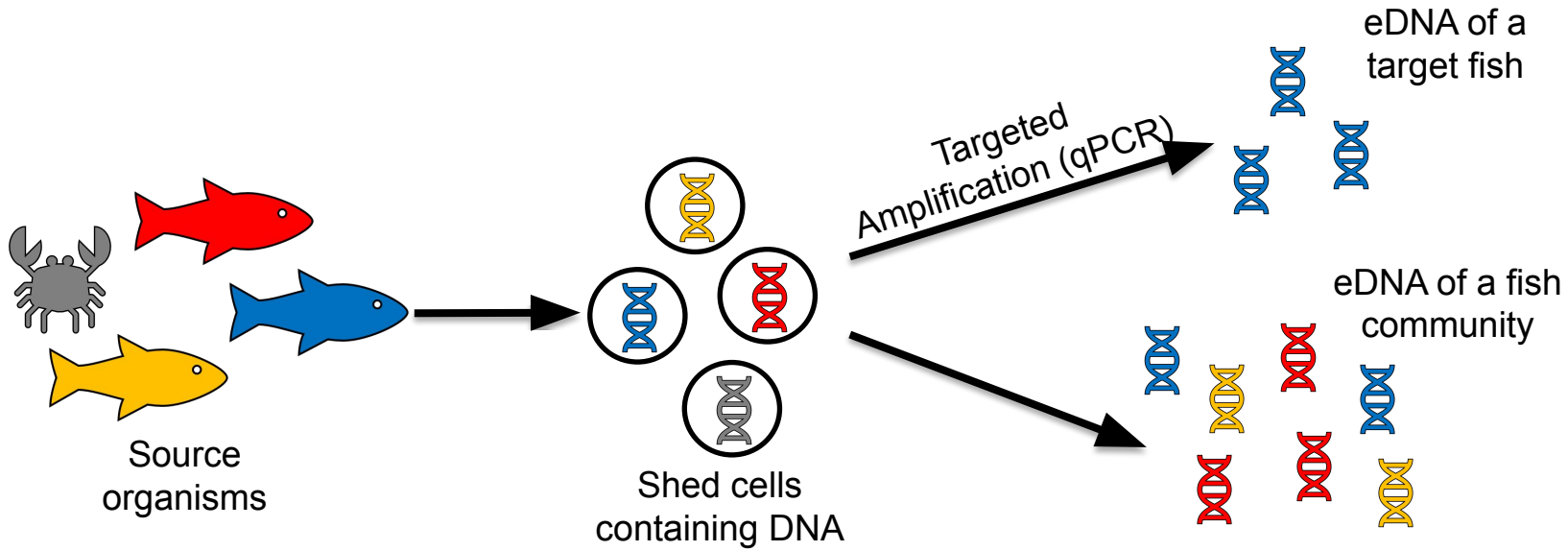
NOAA FISHERIES

NOAA FISHERIES

Case study: Using deep-learning to improve DNA classification

Deep learning and fish eDNA metabarcoding

- Environmental DNA (eDNA) can be use to target a specific taxonomic group for sequencing based on shed DNA
- Fish eDNA is commonly sequenced using two primers in the 12S region
- Many species share the same sequence, limiting taxonomic resolution
- Current sequence classification approaches operate near theoretical maximum accuracy
- Other information may be used to identify these sequences



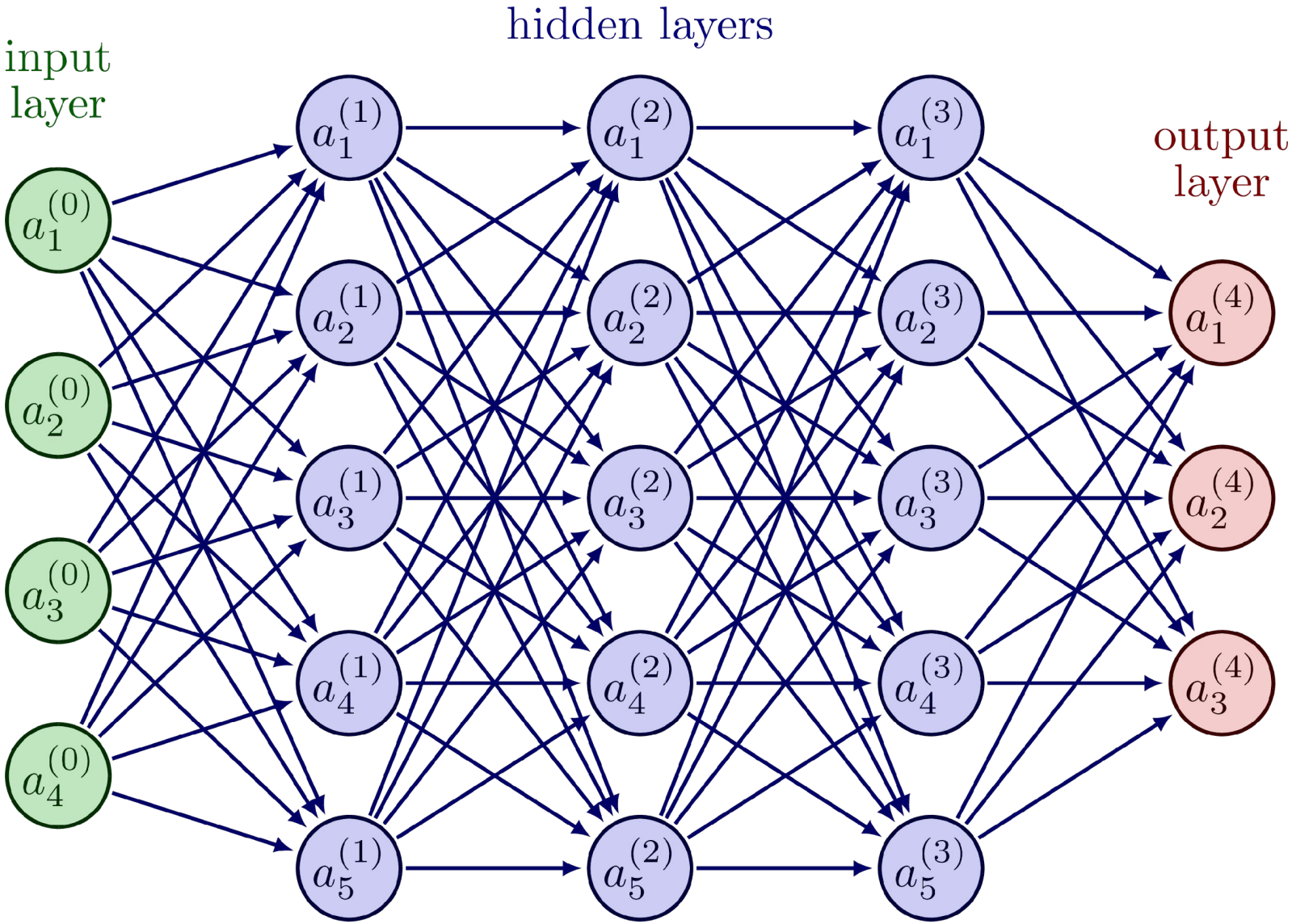
ePlacer: Using Deep-Learning to augment taxonomic assignment.

- Use CNN to set up a DNA classification network
- Add a geographic location tensor
  From OBIS
- Combine the geographic and DNA vectors
- Resolve fish with exact sequence overlap, but distinct geography
- Here, models were trained and tested for 2 fish eDNA marker genes

A	T	G	C	-	A
1	0	0	0	0	1
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0.1	0.5	0.1	0	0	0	0
0	0	0	0.5	1	0.5	0	0	0	0
0	0	0	0.1	0.5	0.1	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

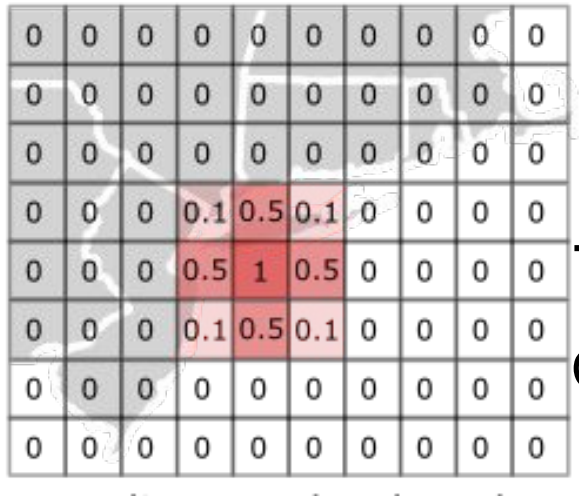
Network

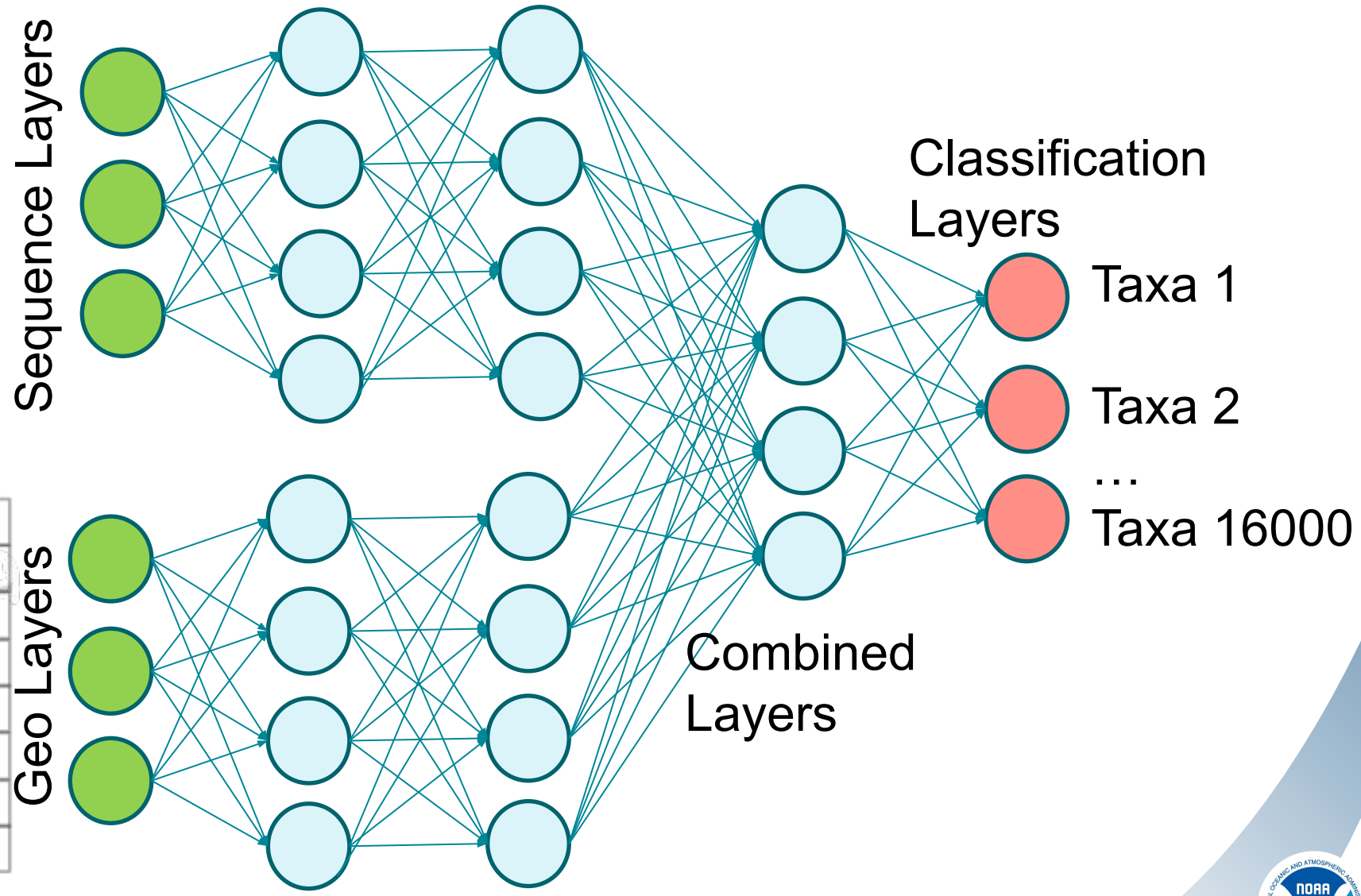


https://tikz.net/neural_networks/

ePlacer: Using Deep-Learning to augment taxonomic assignment.

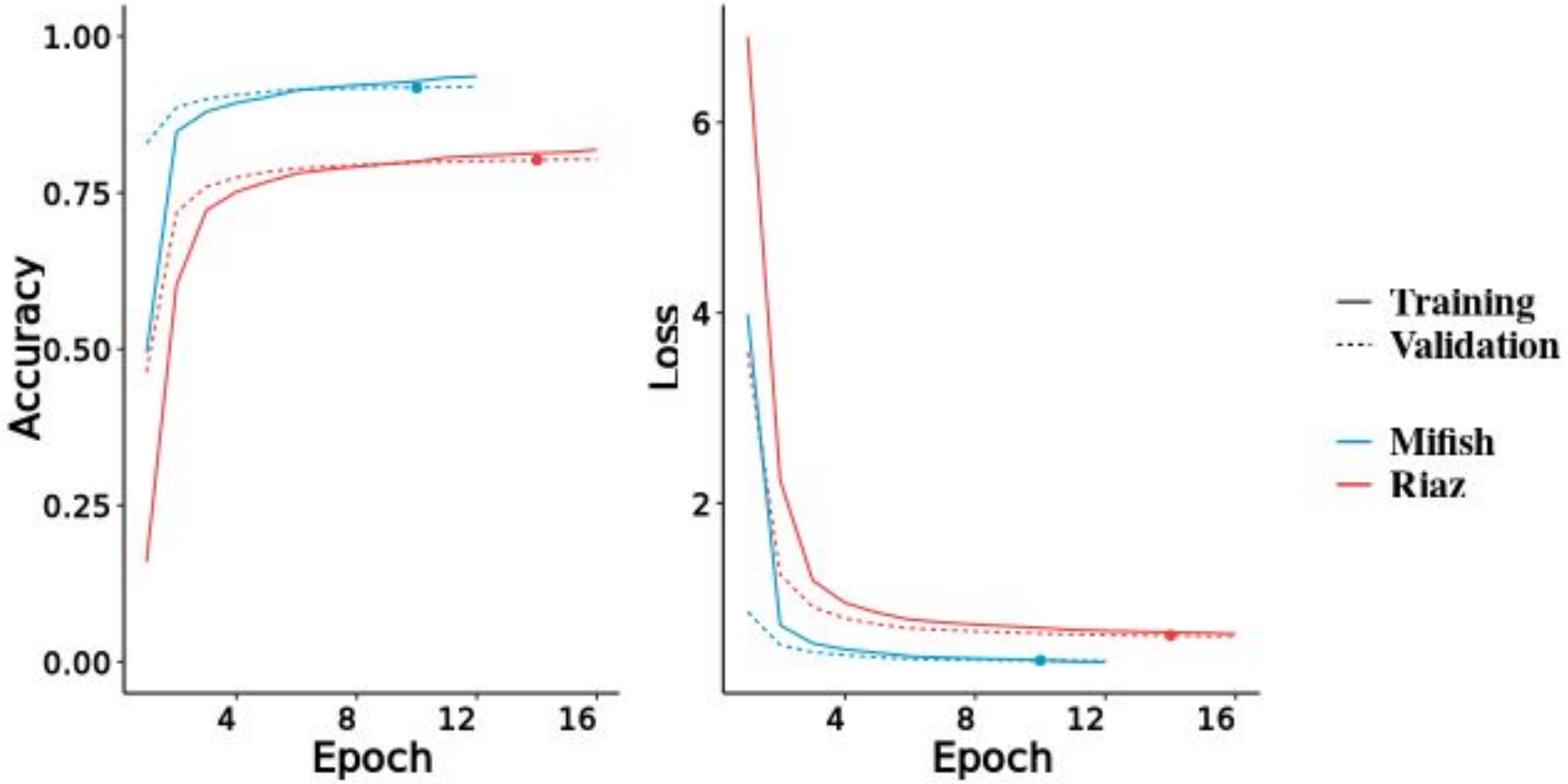
A	T	G	C	-	A
1	0	0	0	0	1
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0



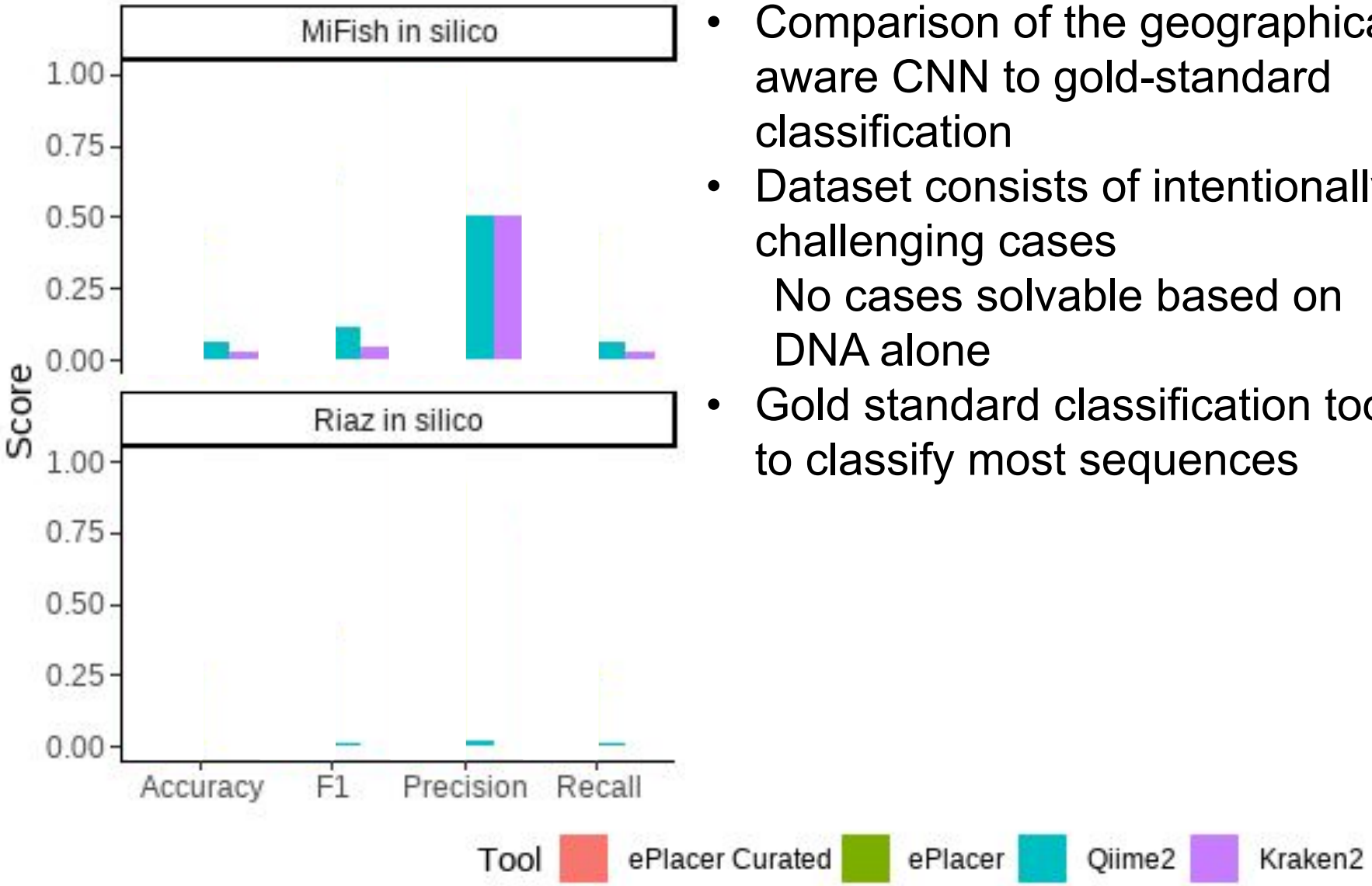


NOAA FISHERIES

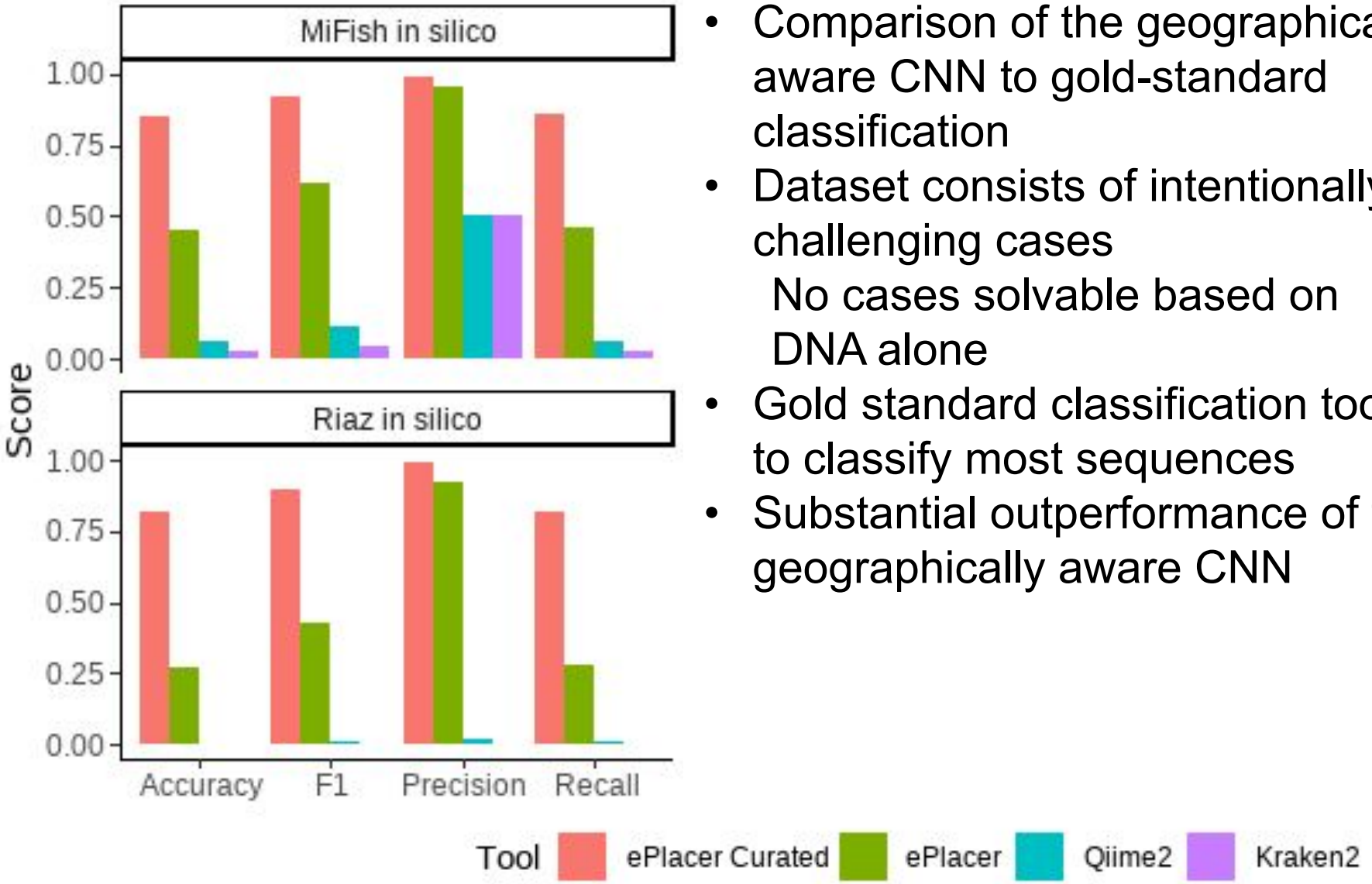
Both marker gene regions exhibited robust learning patterns



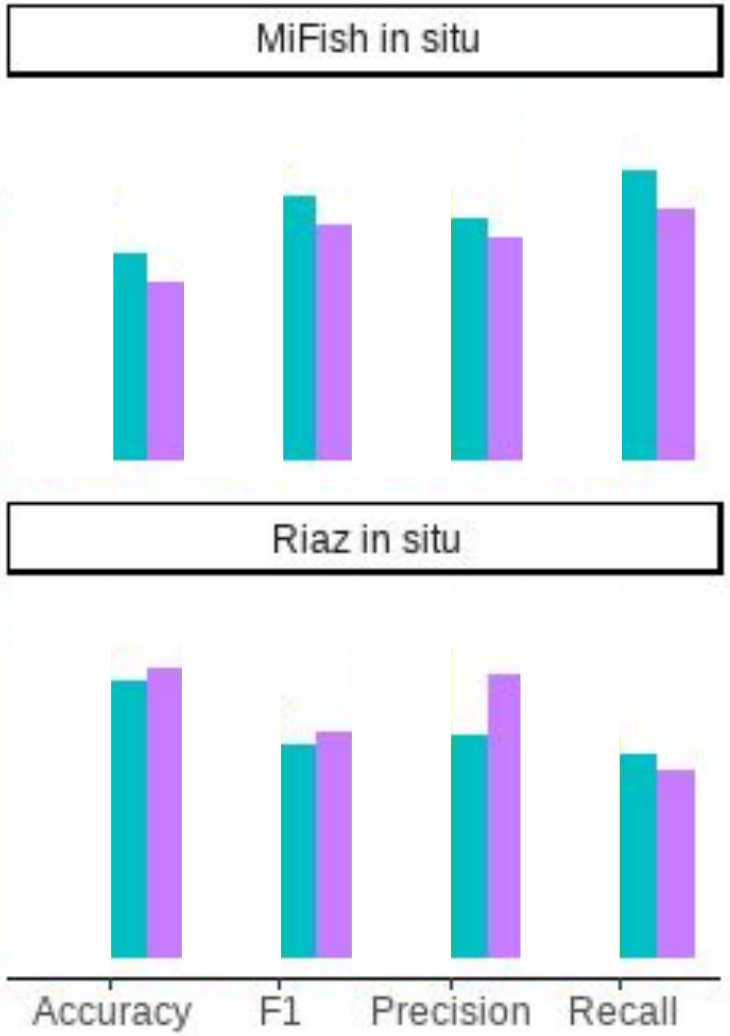
NOAA FISHERIES



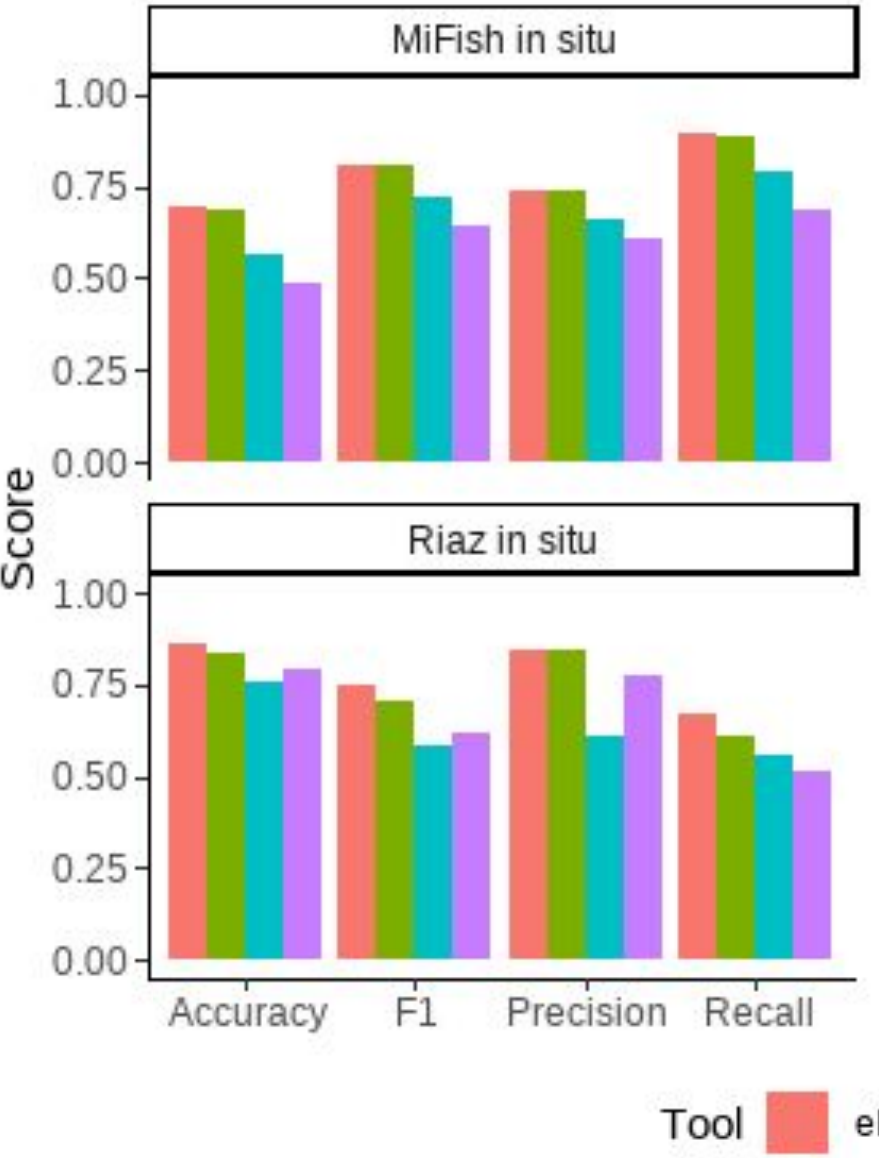
- Comparison of the geographically aware CNN to gold-standard classification
- Dataset consists of intentionally challenging cases
  No cases solvable based on DNA alone
- Gold standard classification tools fail to classify most sequences



- Comparison of the geographically aware CNN to gold-standard classification
- Dataset consists of intentionally challenging cases
  - No cases solvable based on DNA alone
- Gold standard classification tools fail to classify most sequences
- Substantial outperformance of the geographically aware CNN

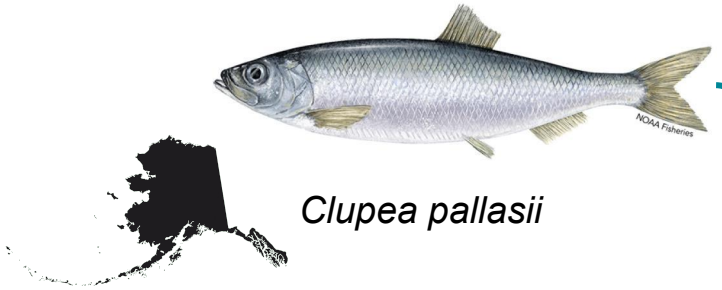


- Comparison of the geographically aware CNN to gold-standard classification
- Dataset consists of real-world data from Stoeckle *et. al.* 2021 and Ledger *et. al.* In Prep.
- ePlacer exhibited over-performance of gold standard, sequence-only tools across all performance metrics



- Comparison of the geographically aware CNN to gold-standard classification
- Dataset consists of real-world data from Stoeckle *et. al.* 2021 and Ledger *et. al.* In Prep.
- ePlacer exhibited over-performance of gold standard, sequence-only tools across all performance metrics

A practical application on MiFish: Examination of the *Clupea*



Clupea pallasii

Clupea harengus

CGGCGTAAAGAGTGGTTATGGAAAACAAGCACTAA
AGCCAAAGAGCCCTCAGGCCGTTATACGCACCCG
GGGCCTCGAACCACTATCACGAAAGTAGCTTTACC
CTCGCCCACCAGAACCCACGAGAGCTGGGACACA

The *Clupea* are difficult to discriminate

Clupea pallasii
Manual curation

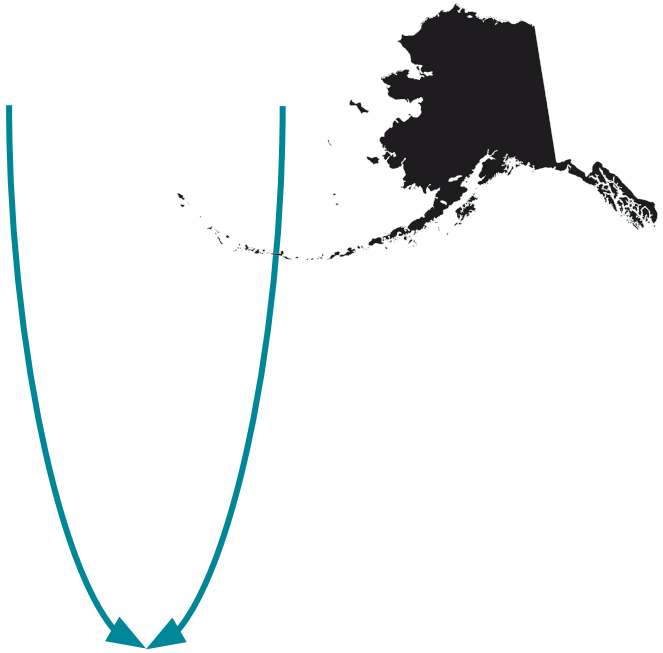
Clupea
Qiime: 0.999 Confidence

NOAA FISHERIES

A practical application on MiFish: Examination of the *Clupea*

CGGCGTAAAGAGTGGTTATGGAA
AACAAGCACTAAAGCCAAAGAGC
CCTCAGGCCGTTATACGCACCCG
GGGCCTCGAACCACTATCACGAA
AGTAGCTTTACCCTCGCCCACCAG
AACCCACGAGAGCTGGGACACA

ePlacer is biogeographically aware, **increasing confidence of assignment** for *Clupea pallasii*



Clupea pallasii
ePlacer: 0.999 Confidence

A practical application on MiFish: Examination of the *Clupea*

CGGCGTAAAGAGTGGTTATGGAA
AACAAGCACTAAAGCCAAAGAGC
CCTCAGGCCGTTATACGCACCCG
GGGCCTCGAACCACTATCACGAA
AGTAGCTTTACCCTCGCCCACCAG
AACCCACGAGAGCTGGGACACA

The ability to use the same tool and databases for taxonomic assignment increases **Interoperability** and **FAIRness** of results

Clupea pallasii
ePlacer: 0.999 Confidence

Clupea harengus
ePlacer: 0.784 Confidence

Augmenting DNA classification using Deep-Learning approaches

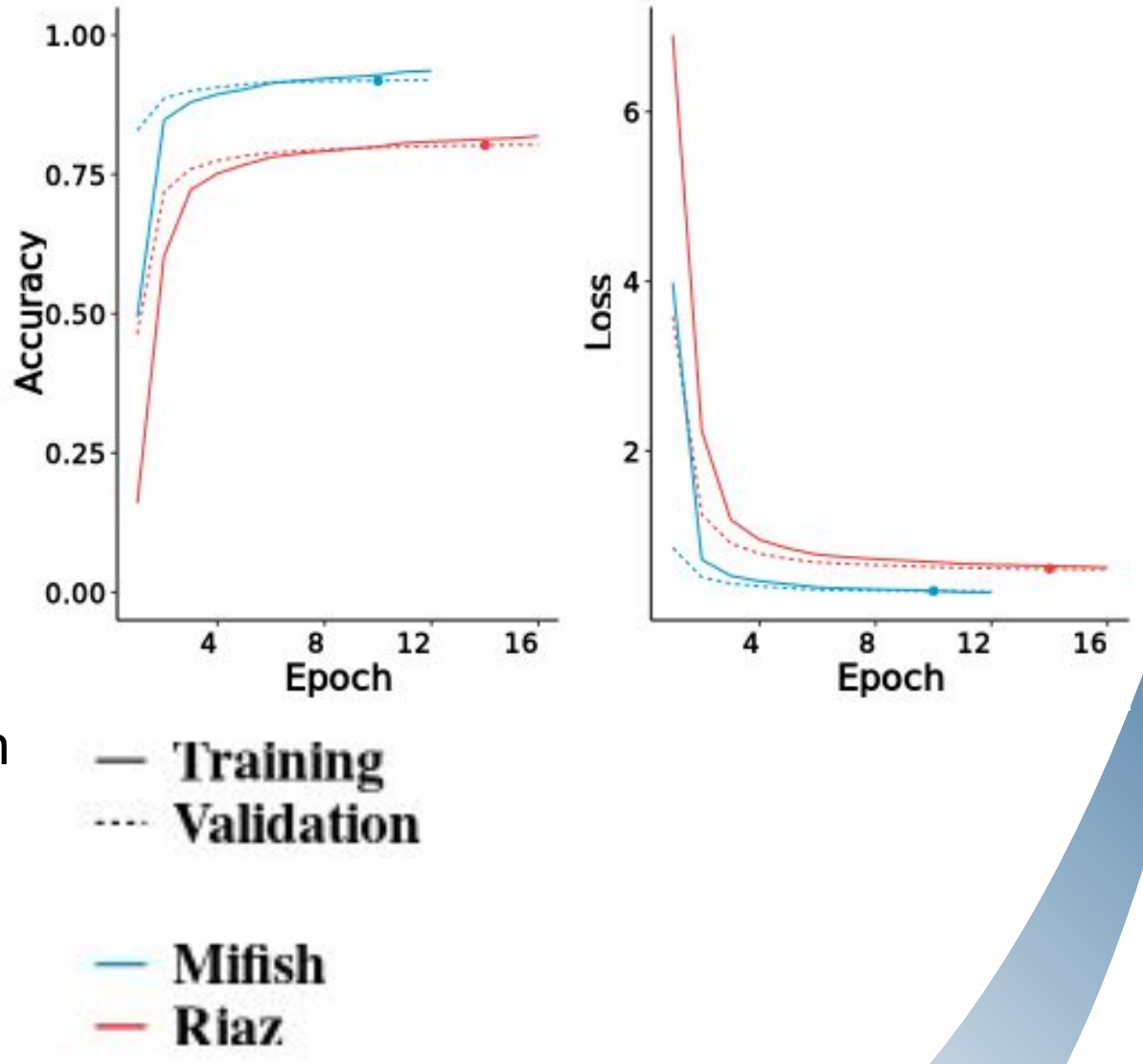
- Deep-learning approaches are powerful when applied to DNA-related problems
- Example: ePlacer substantially increases the accuracy and interoperability of taxonomic assignment

- **Future Directions**
- Incorporate additional data modalities to estimate impacts of ecological frequency in addition to biogeography
- Depth, seasonality, water temperature, distance to coast

- **Questions?**

Monitoring performance of deep-learning: Some key terms

- **Epoch**: An iteration of training. In an epoch, the model has seen the whole training dataset.
- **Training Dataset**: Data used to tune hidden layer parameters. Typically ~ 70%
- **Validation Dataset**: Data used to check performance of model. Typically ~ 30%
- **Accuracy**: Total proportion of correct classifications (correctness)
- **Loss**: A measure of the distance between predicted and actual classification (confidence)



NOAA FISHERIES

NOAA FISHERIES

Transformer-Based Models

For transformer models, all you need is attention

- Cutting edge technology
- No reliance on transformation/convolution
- Transformer-based models use position-based encodings and attention to learn patterns across the whole sequence

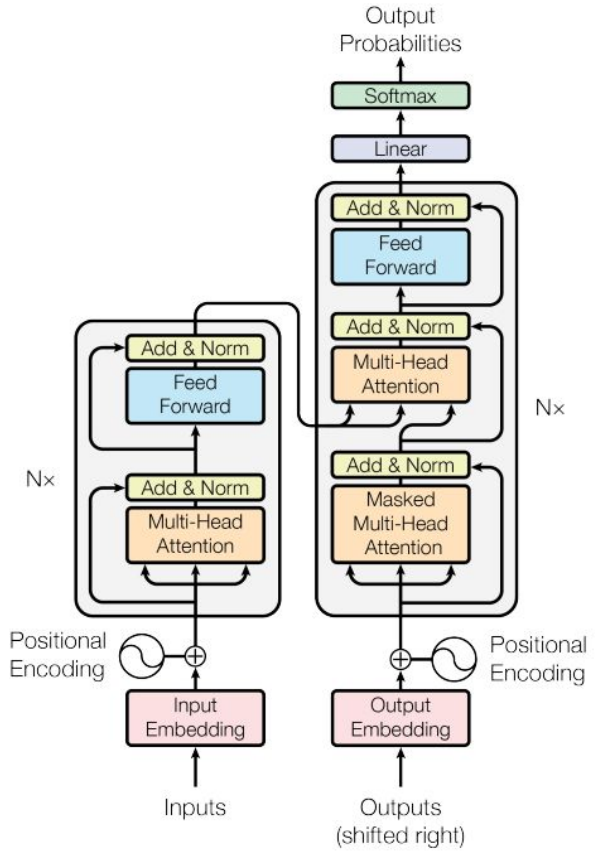
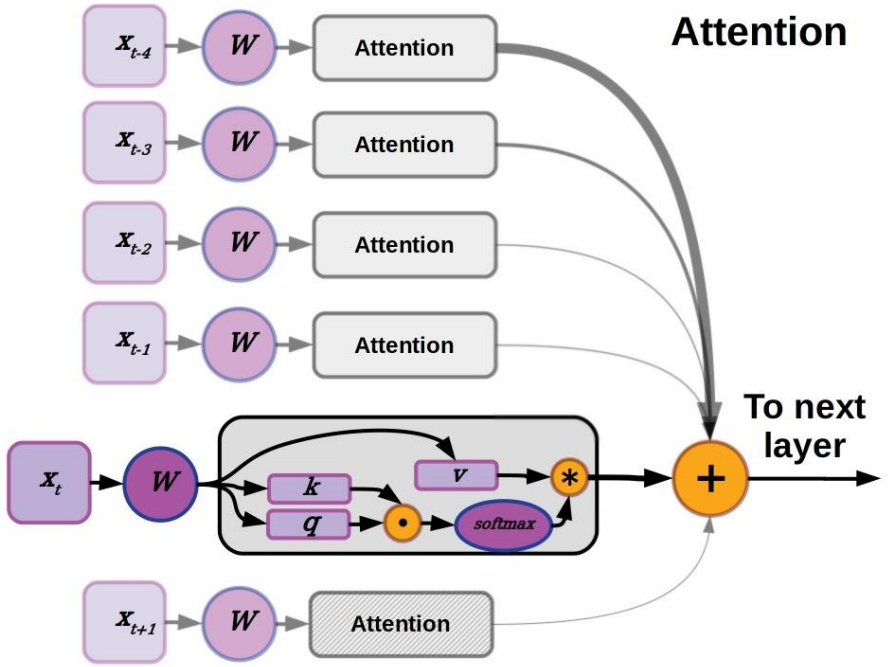


Figure 1: The Transformer - model architecture.

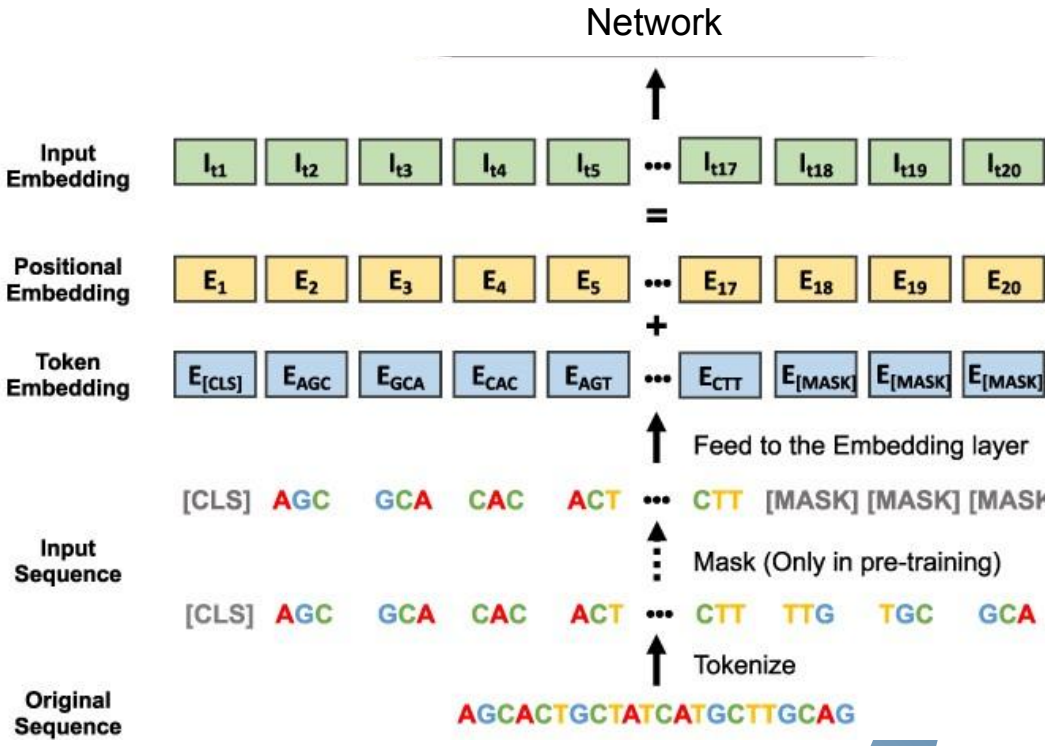
From: Vaswani *et. al.* 2017



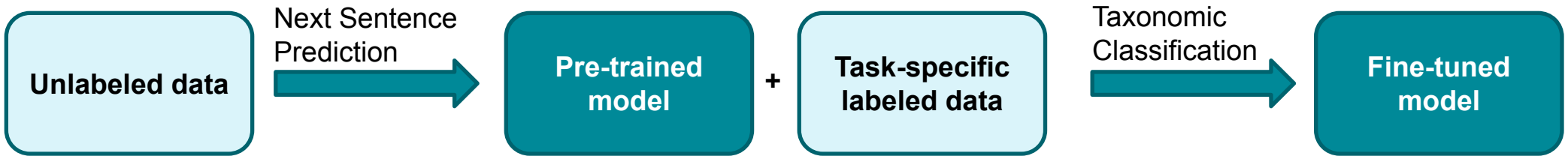
From: https://www.exxactcorp.com/blog/Deep-Learning/a-deep-dive-into-the-transformer-architecture-the-development-of-transformer-models

Why are transformer models useful for DNA?

- Transformer models have been best applied for natural language processing (NLP) tasks
- Current NLP applications are easily adaptable to DNA tasks (see dnaBERT or dnaBERT2)
- DNA encodings share many similar traits to natural language
  words, sentences, paragraphs
  bases, codons, genes
- Generally
  (1) Learn the “DNA language” through pretraining on reference genomes
  (2) fine tune to particular tasks for inference

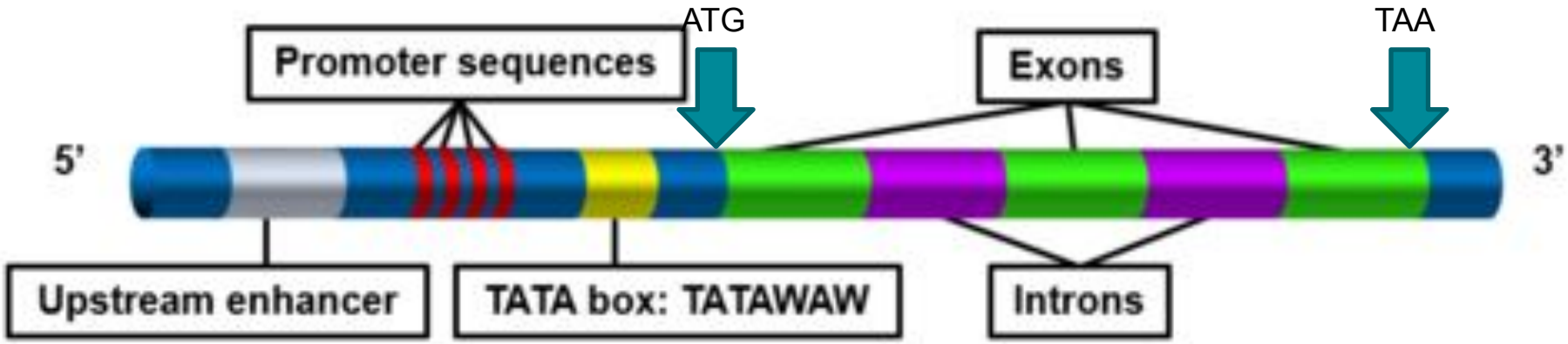


From Ji *et. al.* 2017



Transformer models are exceptionally good at paying attention to the whole sequence
Example: Gene structure

- Because transformers use positional embeddings, it is trivial for the model to begin identifying “spatially important” signals:
  TATA boxes are ~30 bases upstream of ATG
- Increased learning depth could learn protein structure, taxonomic signal etc.



https://en.wikipedia.org/wiki/TATA_box

- Any function for which positional data is important could lend itself to the transformer model approach

Usually, transformer models are trained in two steps

- **Pre-training**
- Large, unlabeled datasets
  - Wikipedia
  - RefSeq
- Self supervised
- Training task: Predict token based on surrounding context
- Learns the “language” (English, DNA, etc)
- Takes a **long** time

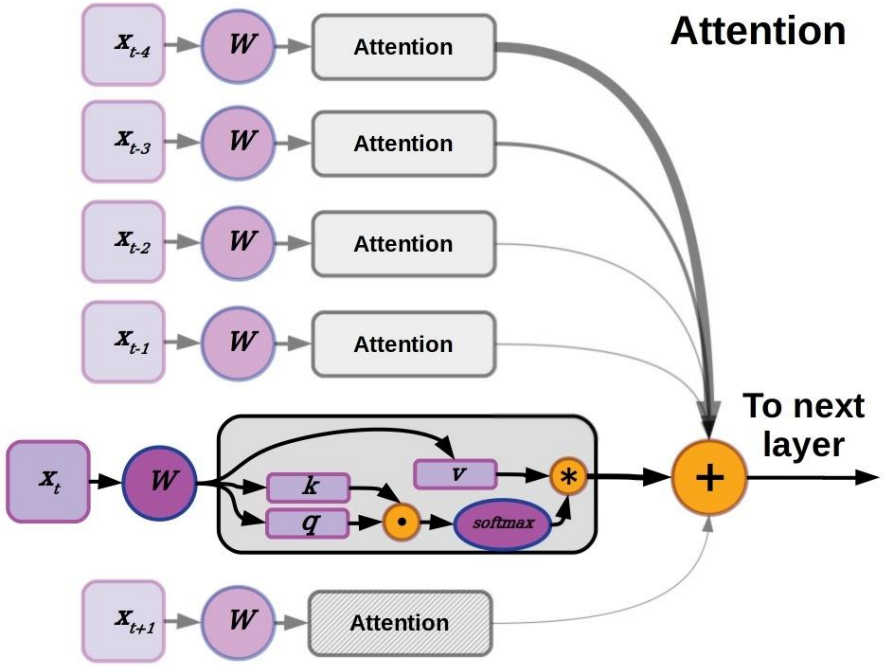
- **Fine-tuning**
- Smaller, labeled dataset
  - Coding sequences
  - Functional annotations
  - Taxonomy
- Model understands the language, so only needs to learn specific knowledge tailored to the task
- Rapid training

- **Inference**
- Using the model for predictive tasks

Overall: Transformers are uniquely able to natively handle DNA without transformation, which makes them powerful

- Cutting edge deep-learning approach
- Positional-awareness allows the model to learn spatial relationships
- Not limited by a kernel size
- Fine-tuning of pre-trained models allow for rapid application of old models to new tasks

- **Questions on Transformers?**



From: https://www.exxactcorp.com/blog/Deep-Learning/a-deep-dive-into-the-transformer-architecture-the-development-of-transformer-models