BE/Bi 205: Lecture 1

David Van Valen MD, PhD

04/04/2023

Organizational Details

- Title: Deep learning for biological data
- Instructors: David Van Valen, Morgan Schwartz, and Rohit Dilip
- TAs: Morgan Schwartz and Rohit Dilip
- Times: Tuesday/Thursday 1-2:30 pm
- Office Hours: To be scheduled

Grading

- Grading is pass/fail
- 60% Assignments
- 40% Class project
  - 10% Project presentation
  - 30% Final project report

Programming knowledge

- This is a programming intensive course
- Working knowledge of Python is necessary for you to get value from this course
- Familiarity with similar languages (e.g., MATLAB, Julia, etc.) might be sufficient, but our ability to help you transition to Python is limited
- Prior machine learning knowledge will also help you get value from this course

Programming knowledge

- The class is now deep learning framework agnostic!
- The instructors (collectively) have experience with all the commonly used deep learning frameworks (TensorFlow, PyTorch, and JAX)
- PyTorch will be used in recitations
- Assignments and projects can be done in the deep learning framework of choice
- The instructors will try our best to support students if (reasonable) issues with the different frameworks arise during the course

Guest lectures

- We have an exciting lineup of guest lectures during the course
- Please make sure to attend!
- We'll be opening up these lectures to the broader Caltech community

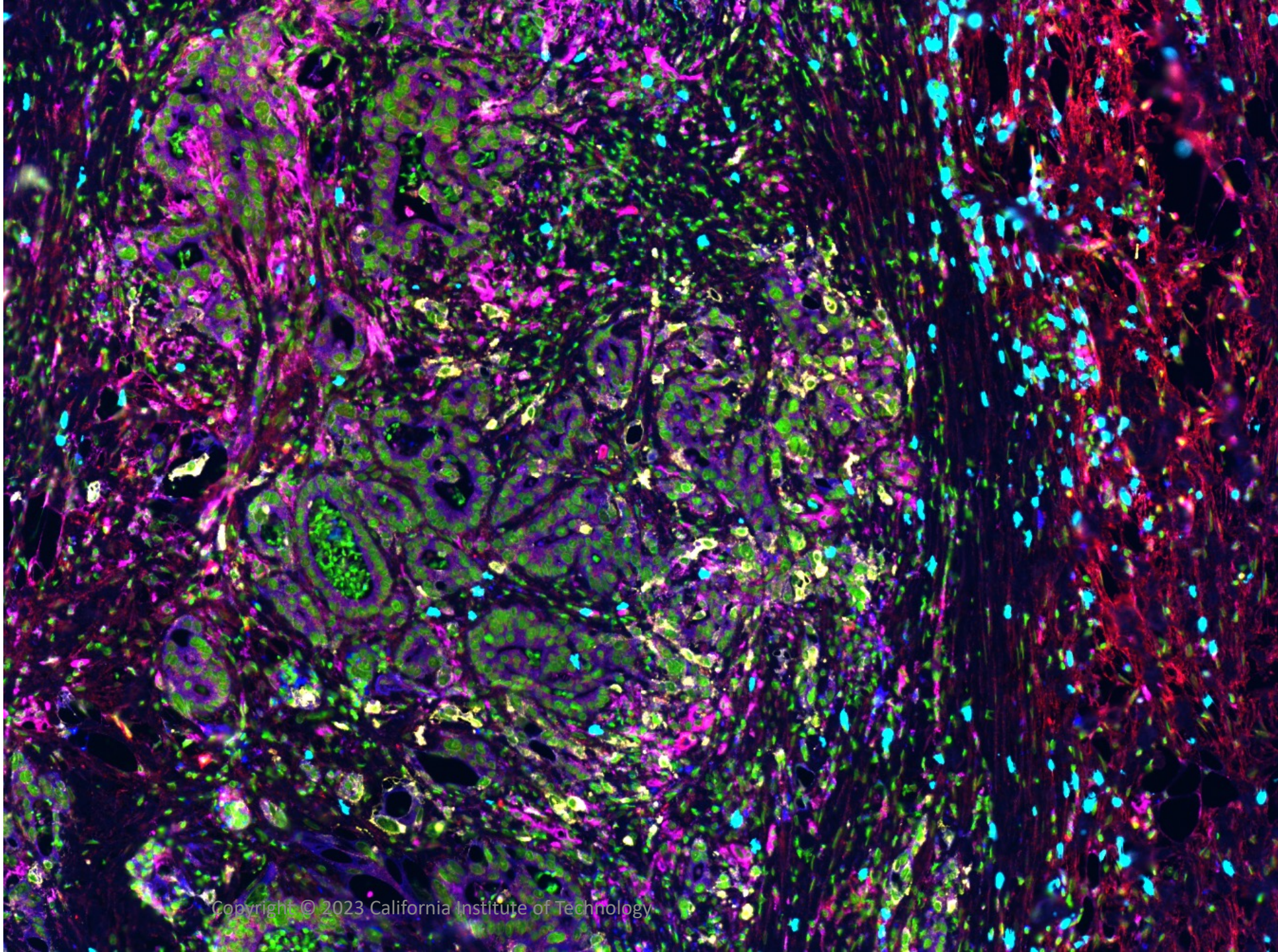
Course Philosophy

- This course is meant to be a *practical* introduction to deep learning methods for biological data
- Our goal is to provide you with the tools necessary so you can apply these methods to your data
- But this space moves very fast, and we want you to know the latest methods, how they work, and what they can do.
- We also want to give you conceptual knowledge so you can navigate what to do when things don't work (which is often most of the time)
- Meeting all these needs requires striking a balance of concepts/practical advice

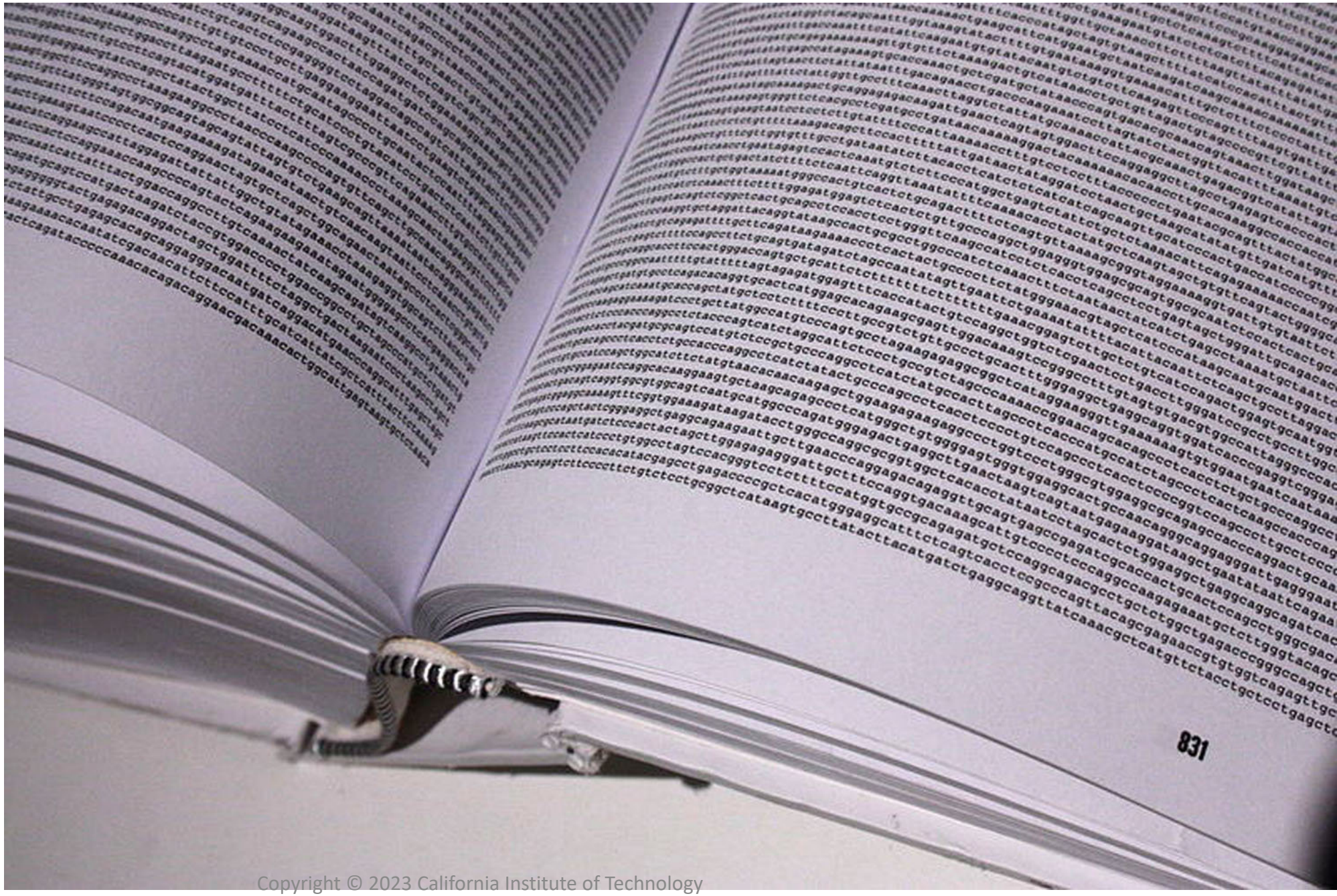
Course Philosophy

- The intersection of deep learning and the life sciences is a fast-moving space – this course will also move fast.
  - New this year: Advances in deep learning for sequence and structural data!
- Lectures – Higher level and conceptual. Lectures will be hybrid, but it is better if you attend live.
  - Life events and missing some lectures we can accommodate.
- Notebooks – During the course, we will release several Jupyter notebooks that provide guided exercises to introduce some of the concepts covered in class
  - These are not graded – they exist mainly for your educational benefit

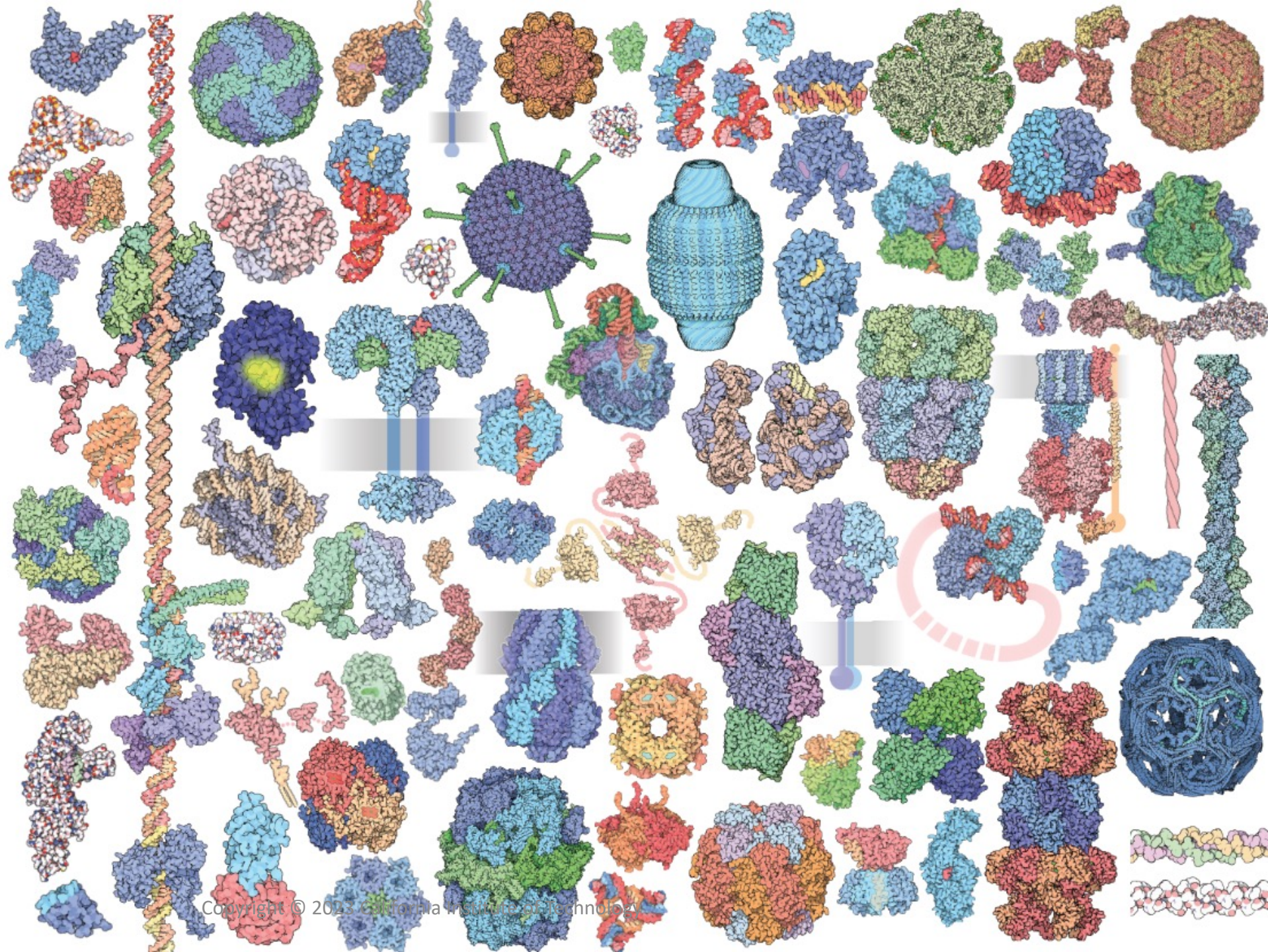
Common datatypes in biology: Images



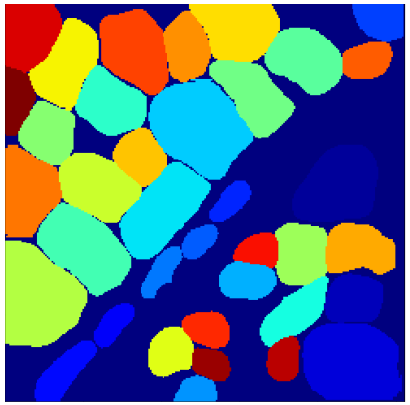
Common datatypes in biology: Sequences



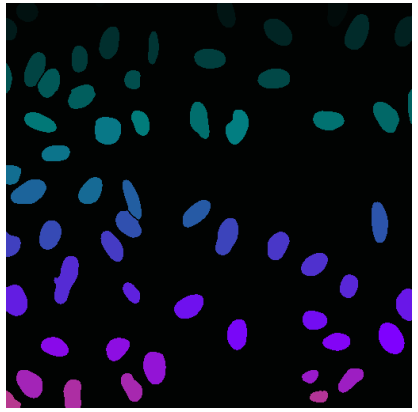
Common datatypes in biology: Structures



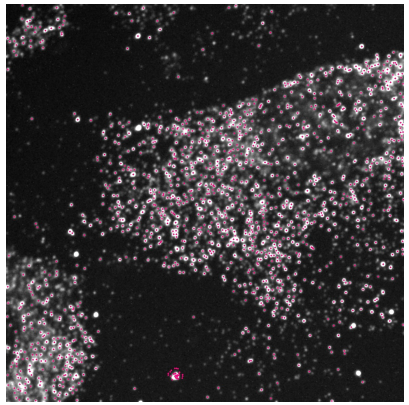
Deep learning is changing our relationship with biological data



Segmentation



Cell Tracking



smFISH Analysis

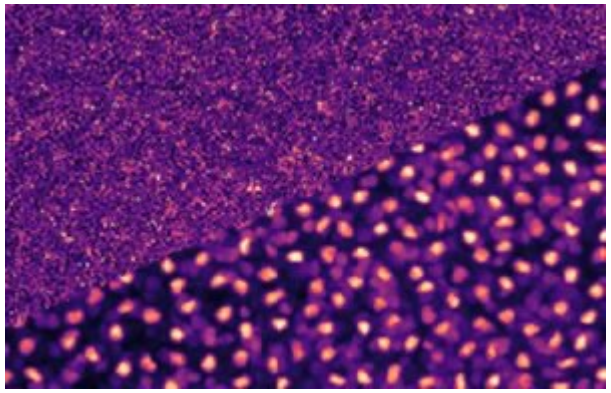
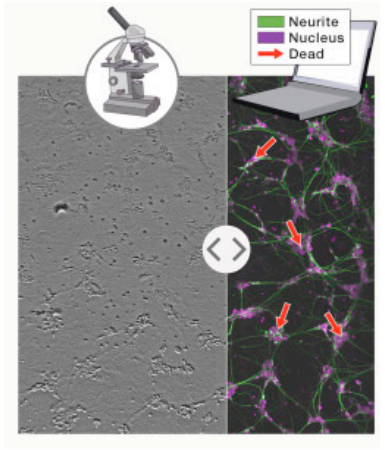


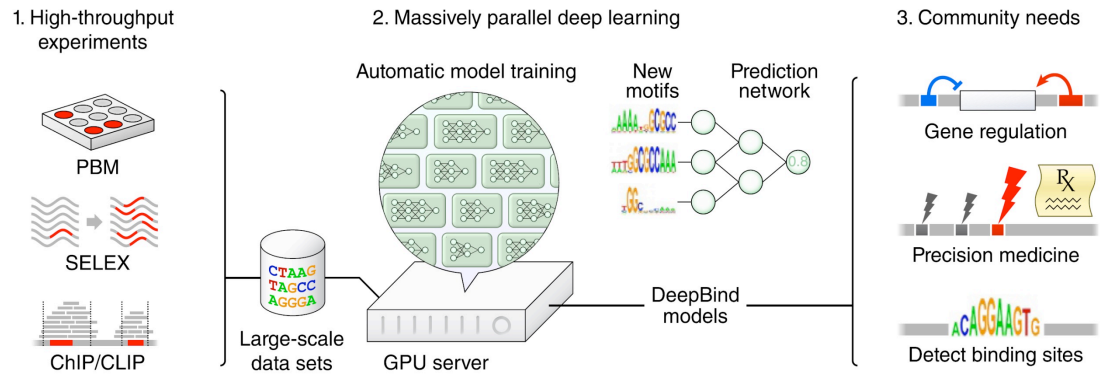
Image Restoration



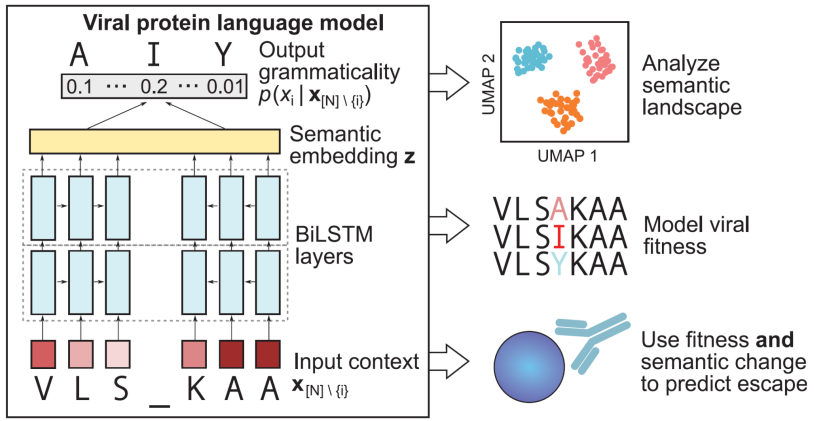
Augmented Microscopy

Weigert et al Nature Methods 2018
Johnson et al Nature Methods 2018
Moen et al biorxiv 2019
Moen et al Nature Methods 2019

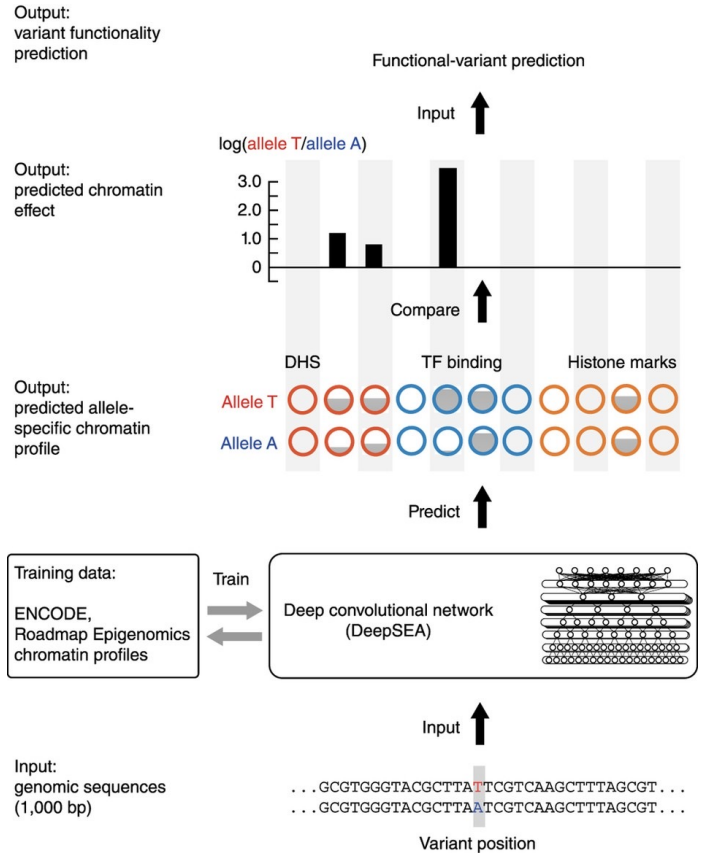
Deep learning is changing our relationship with biological data



Transcription factor binding



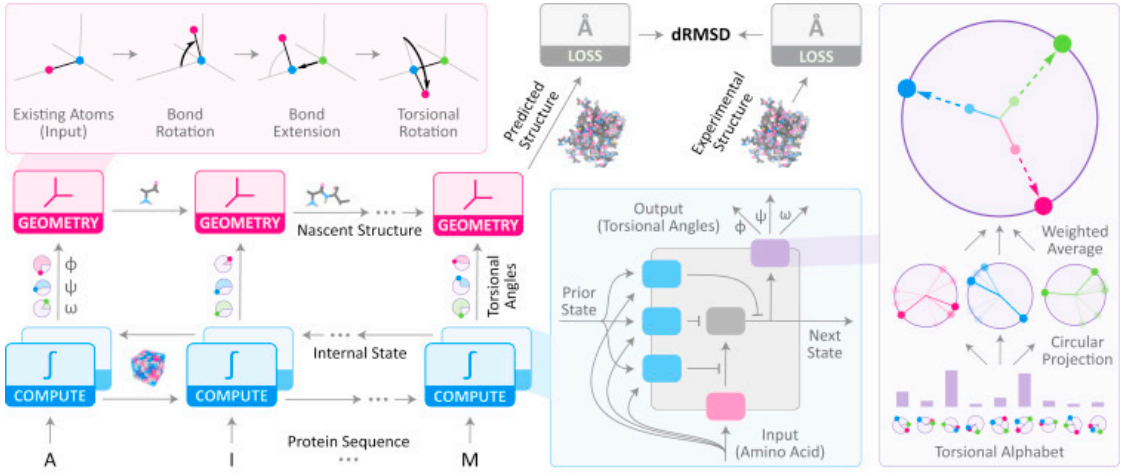
Viral escape

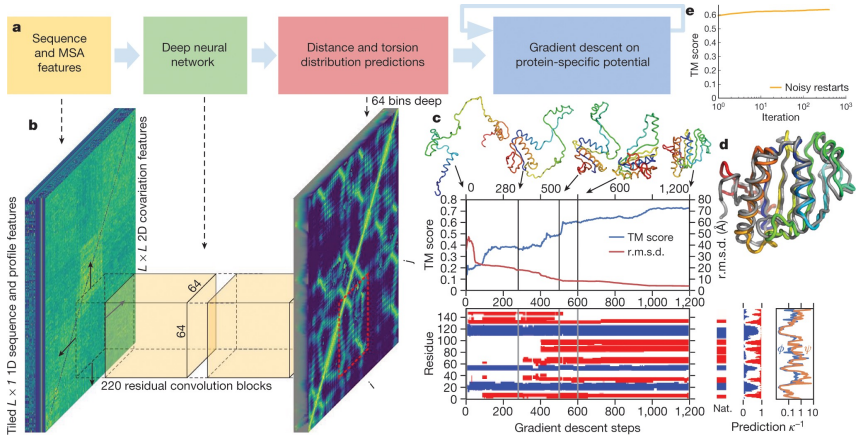


Genetic variant prediction

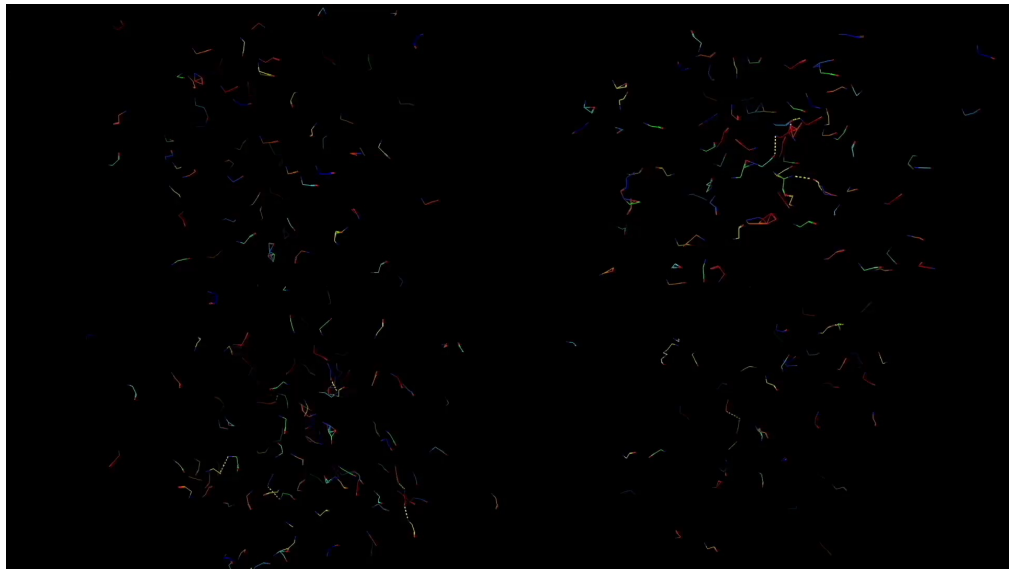
Alipanahi et al Nature Biotech 2015
Zhou et al Nature Methods 2015
Hie et al Science 2020

Deep learning is changing our relationship with biological data





Protein folding



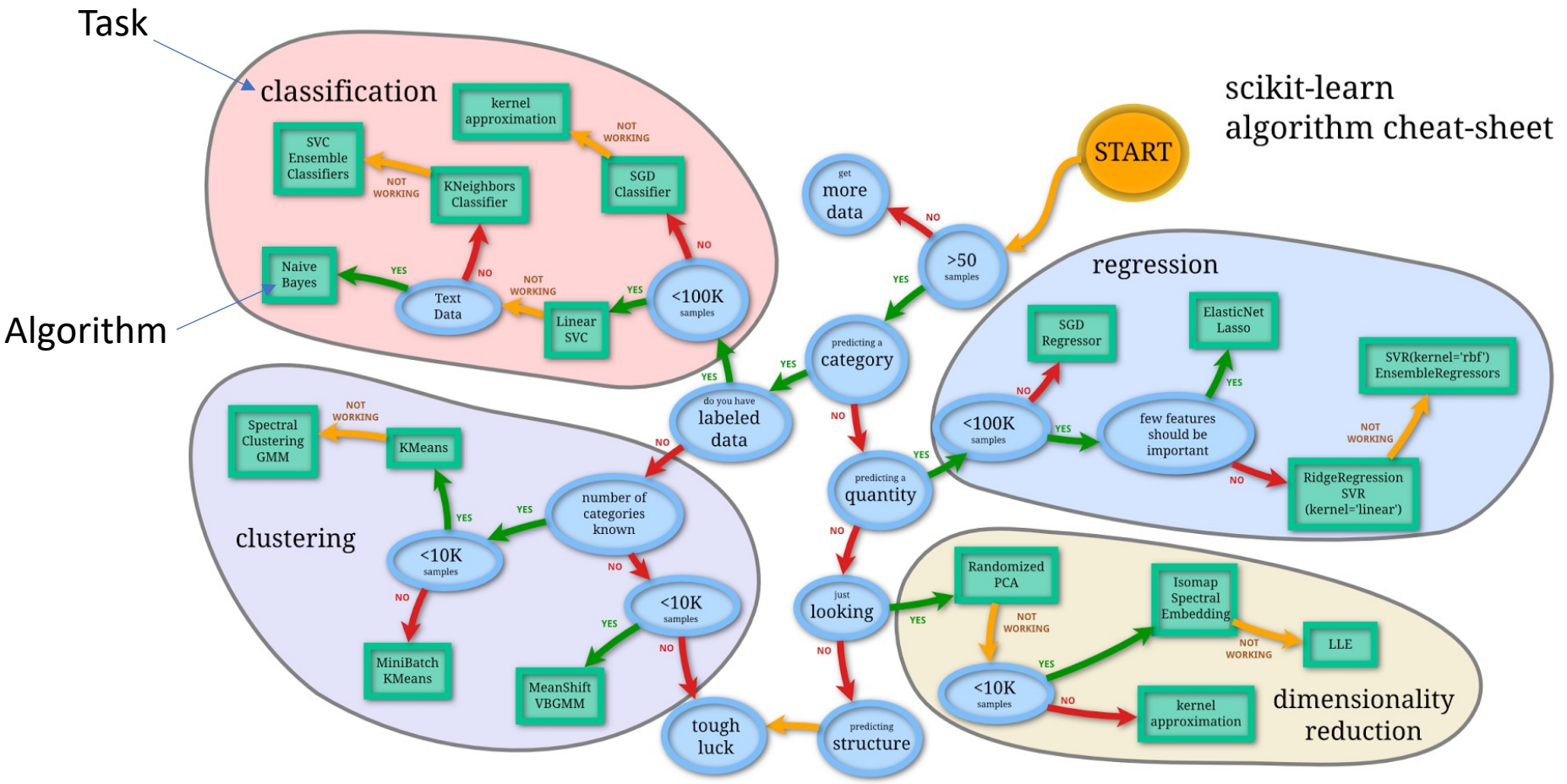
Protein design

Al-Quraishi, Cell Systems 2019
Alley et al Nature Methods 2020
https://nanand2.github.io/proteins/

Deep learning is changing our relationship with biological data

- We've seen drastic progress over the last five years on previously "impossible" problems related to different biological data types
  - Images: Image segmentation, object tracking
  - Sequences: Transcription factor binding, variant prediction
  - Structures: Protein folding, protein design
- Why has deep learning allowed us to progress so quickly?
- What is deep learning anyway?

What is machine learning?



https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

What is machine learning?

- A set of computational methods from different but overlapping fields (computer science, mathematics, statistics, signal processing, etc.) that extract insight from large scale datasets.

- A key difference between machine learning methods and other computational methods is the dependence on **data**.
- Machine learning methods are powered by data – expansive, diverse datasets are often required to achieve “state-of-the-art” performance

Data is at the heart of the machine learning revolution

https://www.edge.org/response-detail/26587

Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level spontaneous speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games – aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the first Jeopardy! Champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (2010)	Mixture-of-Experts algorithm (1991)
2014	Google's GoogLeNet image classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories	Convolutional neural network algorithm (1989)
2015	Google's DeepMind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough:		3 years	18 years

Data is at the heart of the machine learning revolution

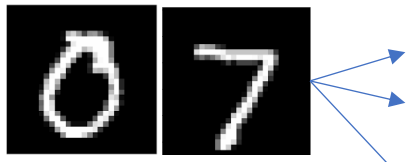
- Advances in machine learning methods in biology are intimately related to advances in generating biological data
  - Imaging: Microscopes and reporters
  - Sequences: Genomics
  - Structures: X-Ray Crystallography and Cryo-EM
- Machine learning methods are succeeding because the data powering them are of sufficient **size, diversity, and information content** to lead to performant models

Data is at the heart of the machine learning revolution

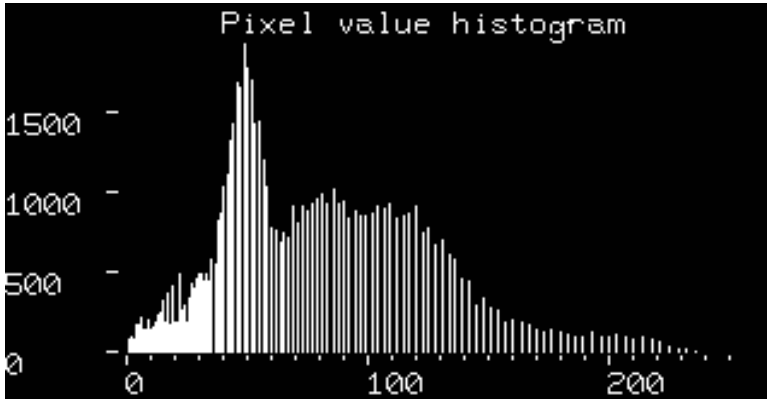
- Developing datasets is an **integral** part of developing machine learning methods
- Dave's personal view – the data is the software!
  - The best applications have defined goals and interpretable metrics, so you know what is an improvement and what is not
  - If you want to solve a problem with machine learning, you need to develop the data and the algorithm jointly

Machine learning terminology

- Features/Representations: Measurable properties or characteristics of a dataset/phenomenon being observed
- Example:

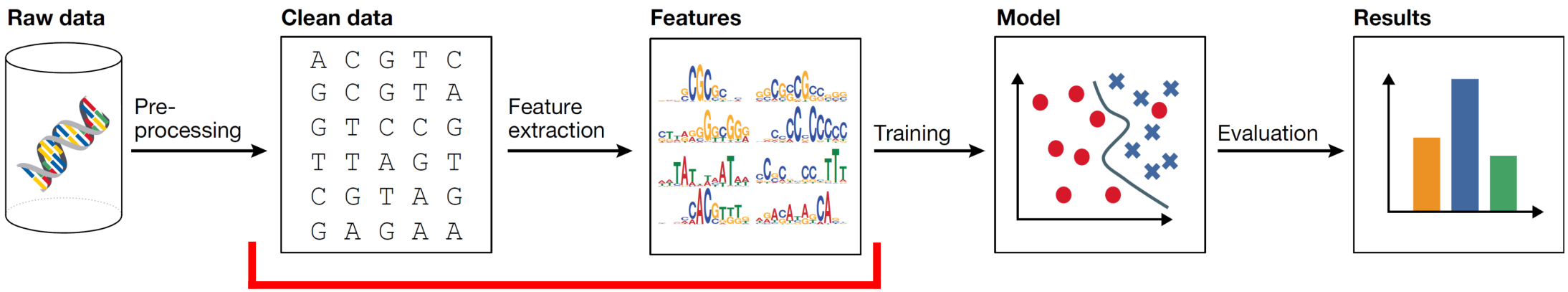


Pixel histogram
Total intensity
Number of edge pixels
Number of corner pixels
Holes
…



Machine learning terminology

- Features/Representations: Measurable properties or characteristics of a dataset/phenomenon being observed
- Example:

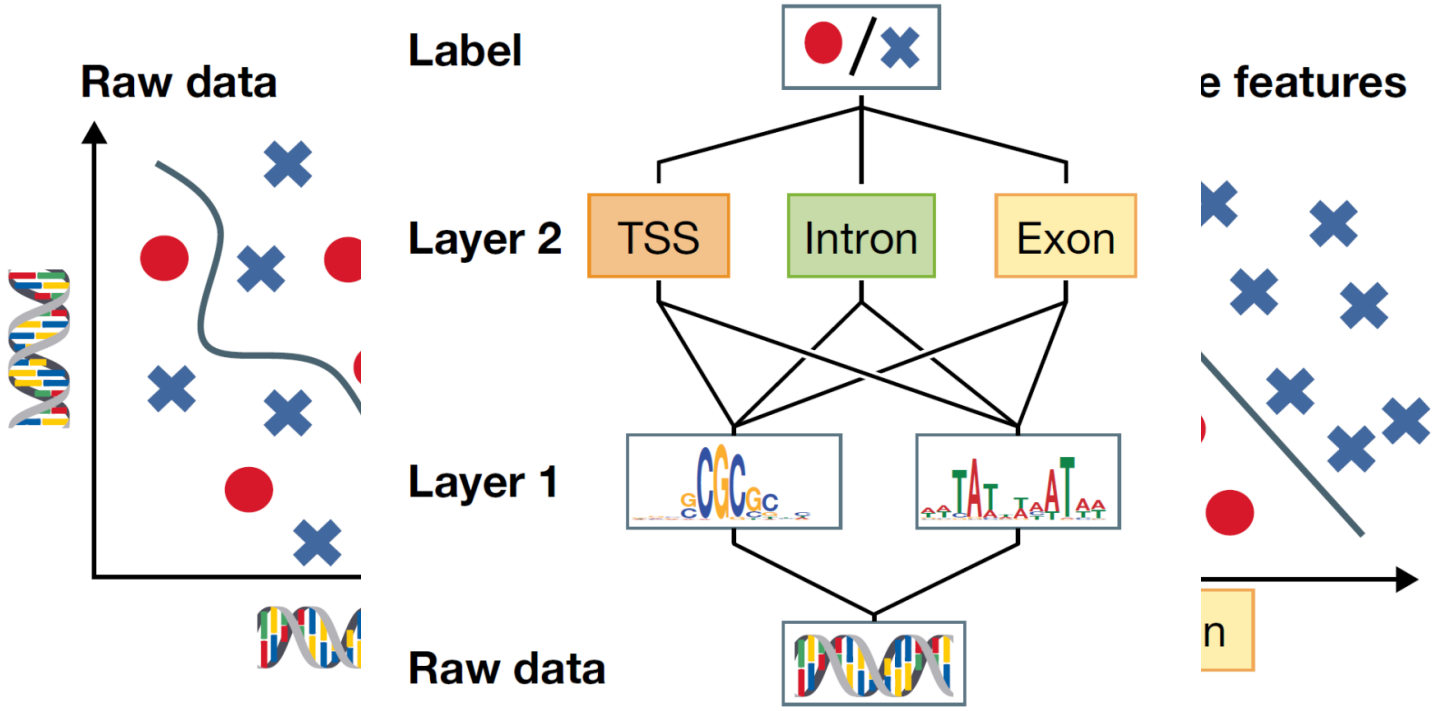


Machine learning terminology

- Features/Representations: Measurable properties or characteristics of a dataset/phenomenon being observed
- Features can be designed manually to take advantage of human intelligence and insight
- Manually crafted features have the advantage of interpretability
- Manually crafting robust and informative features is hard

Machine learning terminology

- Representation learning: A set of methods that discovers representations directly from data that lead to high performance on a given task
- Example:



Angermueller et al, Mol Sys Biol 2016

Machine learning terminology

- Representation learning: A set of methods that discovers representations directly from data that lead to high performance on a given task
- Representation learning removes the need to manually create features

Machine learning terminology

- Labels: Annotations of datasets that contain meaning or insight
- Examples:

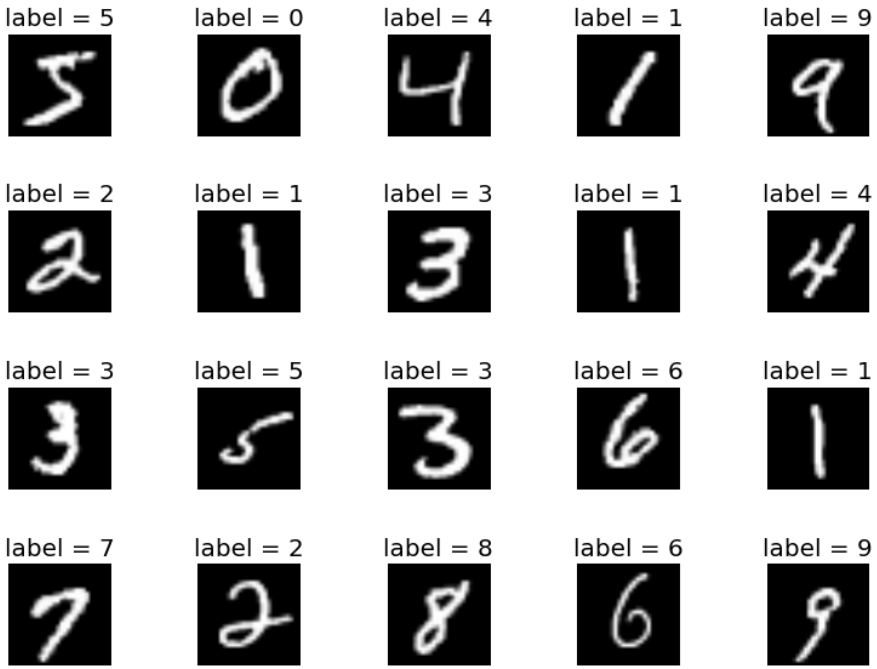
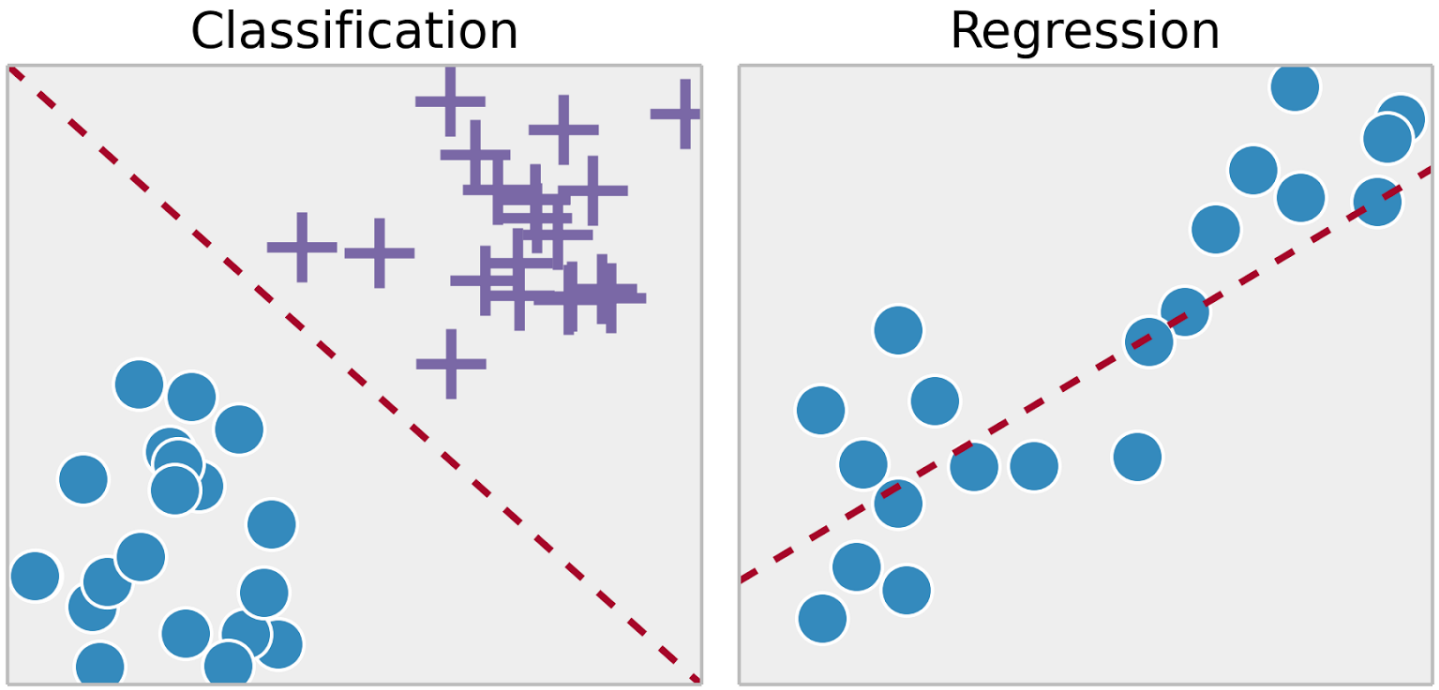


Image level label

Pixel level label

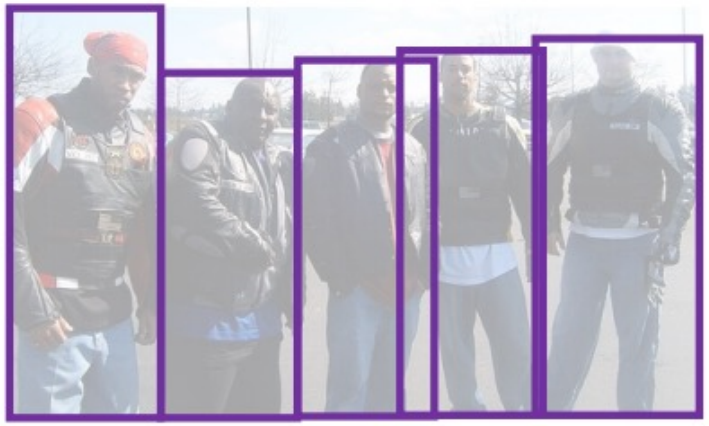
Machine learning terminology

- Tasks: The type of prediction being performed
- Example: Classification vs regression

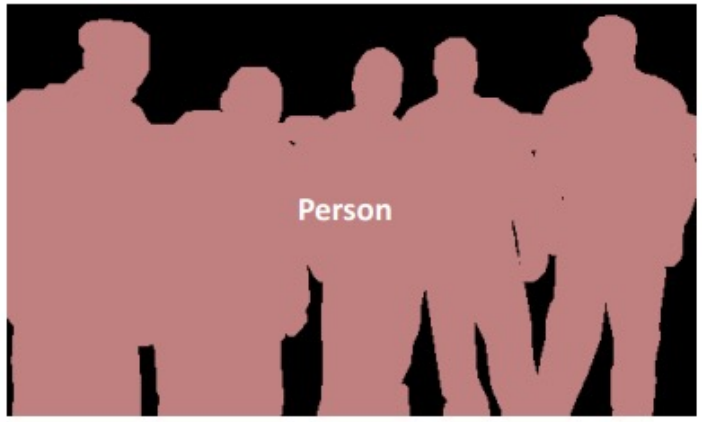


Machine learning terminology

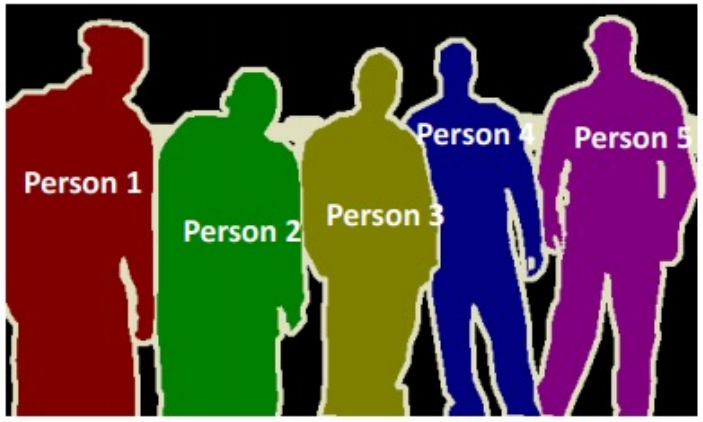
- Tasks: The type of prediction being performed
- Example: Object Detection vs Semantic Segmentation vs Instance Segmentation



Object Detection



Semantic Segmentation

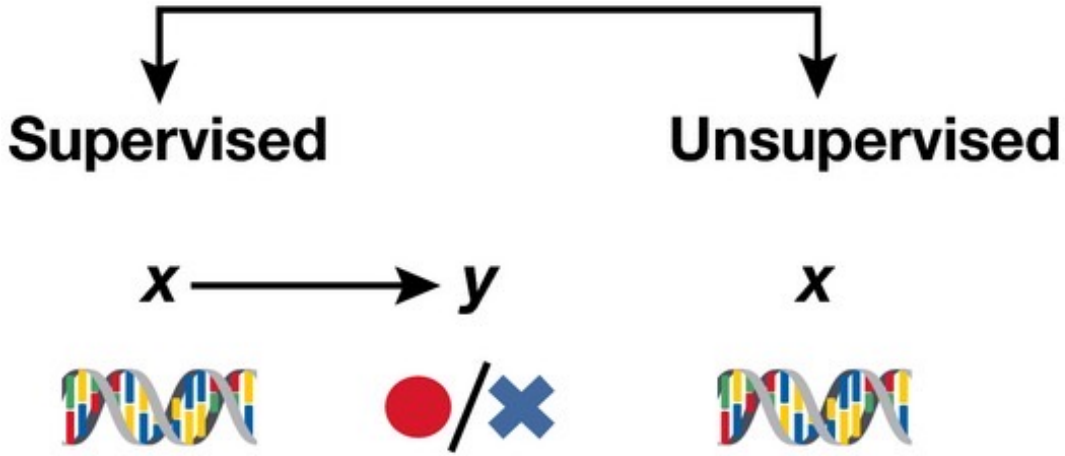


Instance Segmentation

Machine learning terminology

- Supervised Machine Learning: Learning to perform tasks using **human generated labels**
  - Weakly Supervised Machine Learning: Learning to perform tasks with **noisy, limited, or imprecise labels**
- Unsupervised Machine Learning: Learning to perform tasks **without human generated labels**

Machine learning terminology

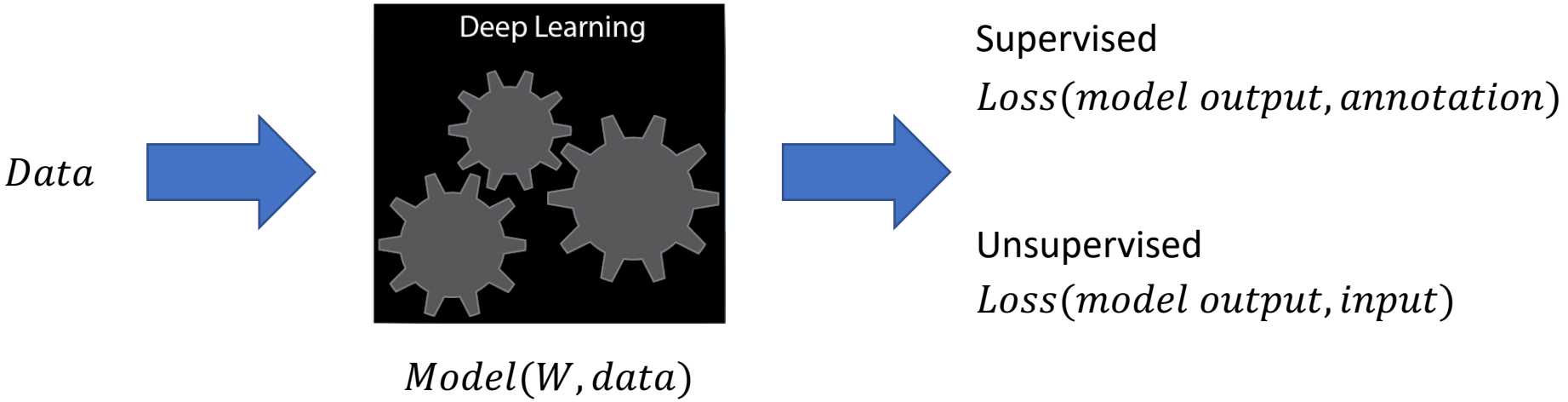


- Linear regression
- Logistic regression
- Random Forest
- SVM
- ...

- PCA
- Factor analysis
- Clustering
- Outlier detection
- ...

Machine learning terminology

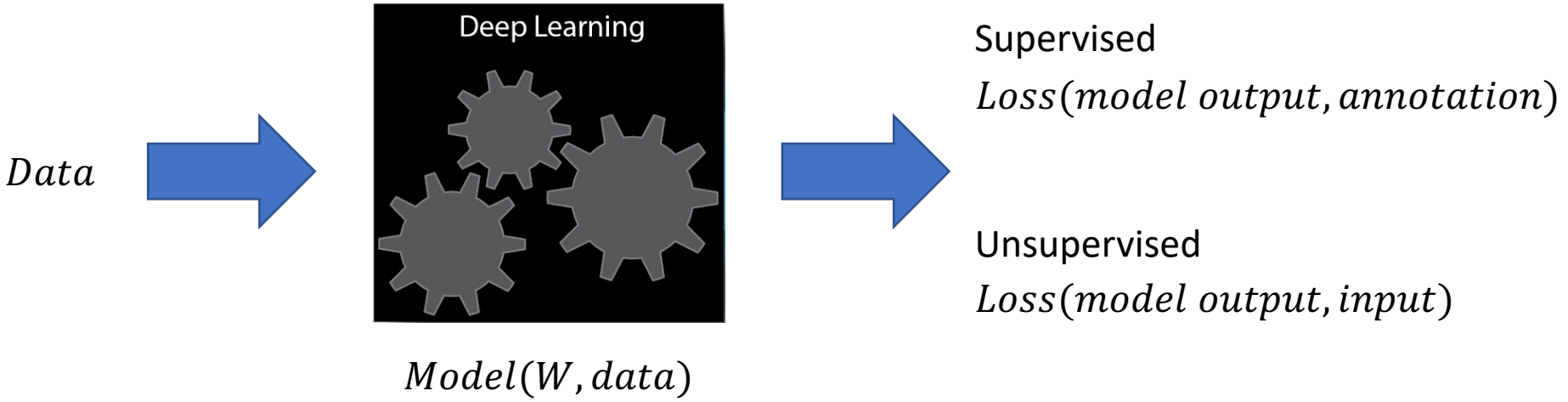
- Loss function: A function that takes in the output of a machine learning model and uses it to infer the model's performance on a task



- Smaller values of a loss function indicate better performance

Machine learning terminology

- Training/Learning: The process of adjusting a model's parameters to improve model performance on a task

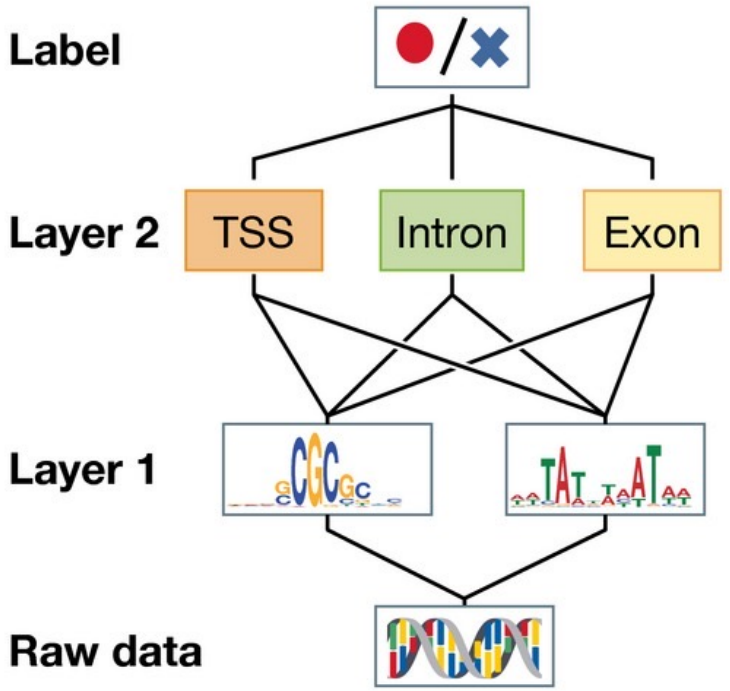


- Example: $Loss = \sum_i (y_i - \hat{y}_i)^2$ Mean Squared Error

Predicted value True value

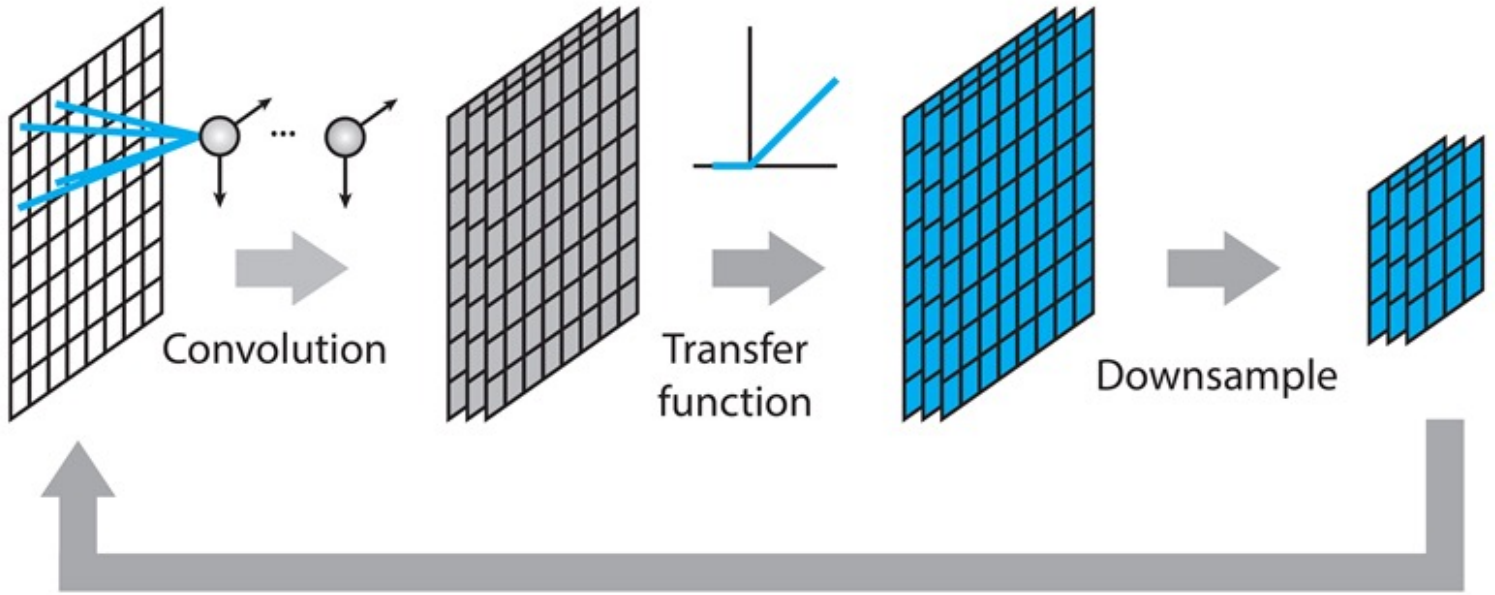
Machine learning terminology

- Deep Learning: A subset of machine learning methods that efficiently learn a hierarchy of representations directly from data.



Machine learning terminology

- Deep Learning: A subset of machine learning methods that efficiently learn a hierarchy of representations directly from data.



Convolutional neural networks