USC University of Southern California

TRGN 515:
Advanced Human Genomic Analysis Methods

Week 6:

Machine learning for genomic data analysis

Bilgenur Baloglu, Ph.D.
- Clinical Instructor of Translational Genomics,
USC Keck School of Medicine
- Bioinformatics scientist III,
Thermo Fisher Scientific
Email: baloglu@usc.edu

Some of the slides adapted from:
Mitch Guttman (Caltech)
Manolis Kellis (MIT)
Andrew Ng (Coursera)

On tap today!

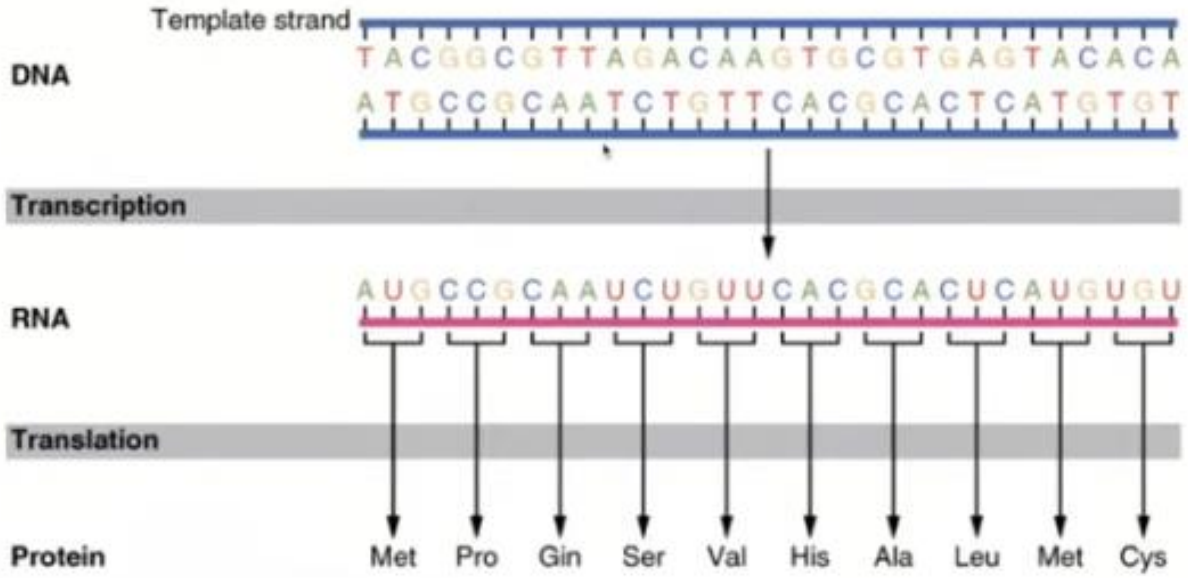
- Genetics, genomics and everything in between
- How do we handle large scale sequencing data?
- What is bioinformatics and how do we use it for genomics data?
- What is machine learning?
- Introduction to machine learning terminology and the models

What we focus on in this class: Machine Learning, not Deep Learning

- Find articles in the resources/articles directory
- This week's reading material:
  - Yang et al 2020, Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA
  - Monaco et al 2021, A primer on machine learning techniques for genomic applications
  - Wan et al 2021, Beyond sequencing: machine learning algorithms extract biology hidden in Nanopore signal data

A bit of genetics, mostly genomics and bioinformatics

The genetic code and protein



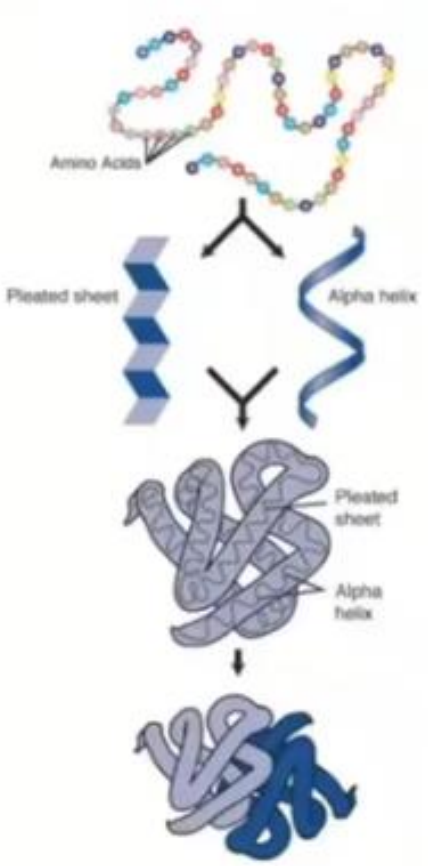
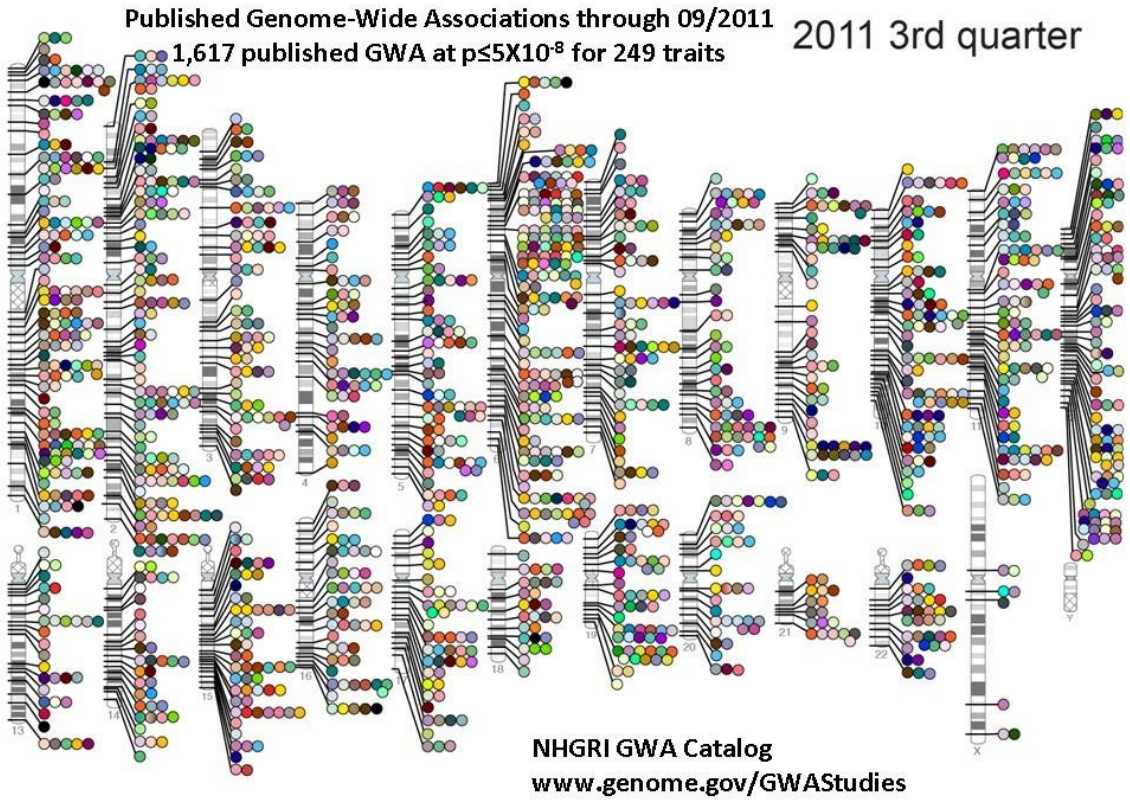


Image source: Bill Chen

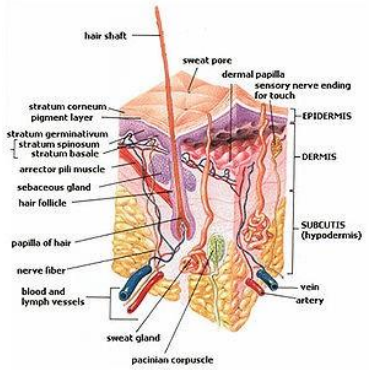
Brief intro to human genetics

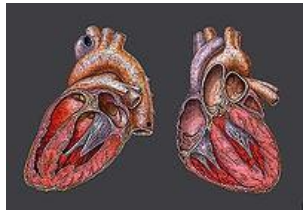
- **Human genome**: 3.2B letters, 2 copies, 23 chromosomes, 20-25k genes, ~3M common SNPs, ~500k haplotype blocks

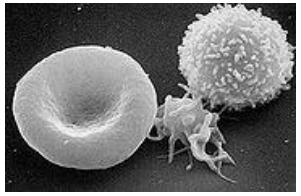


One genome ➡ Many cell types

ACCAGTTACGACGGTCA
GGGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
TACCAGTA







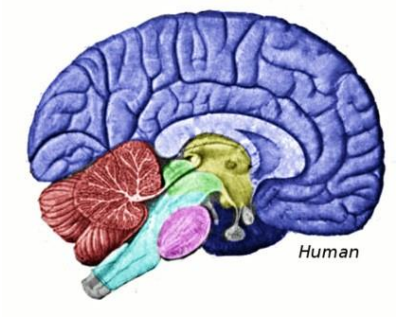


Image Source wikipedia

Control elements and genes define cell types/states

Active control elements
Active control elements
Active genes
Repressed elements

- ~25,000 genes
- ~2 million novel putative control elements!

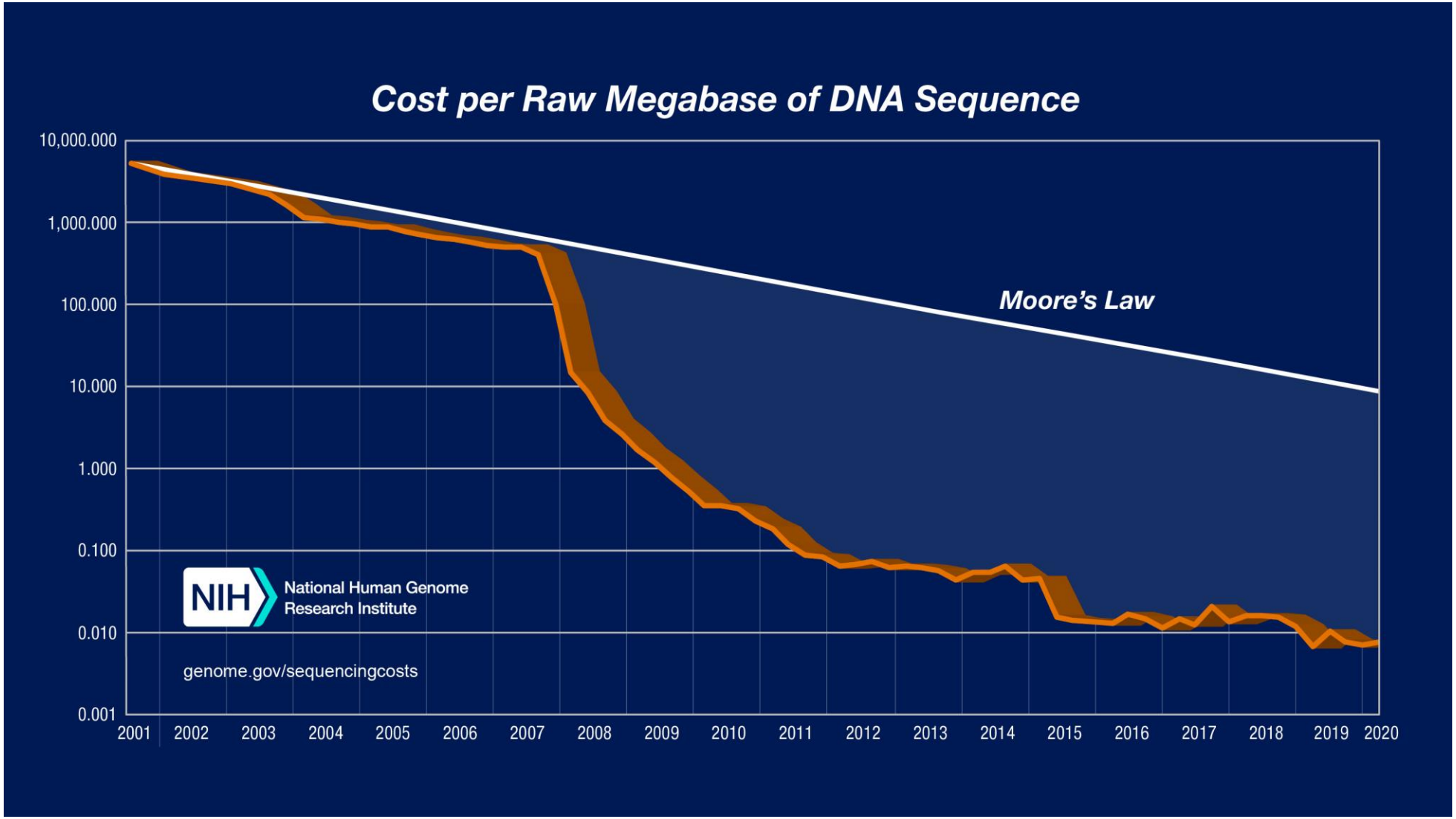
- Predict regulatory element location
- Map RE on gene
- Predict gene expression from RE states.

Image source: Bill Chen

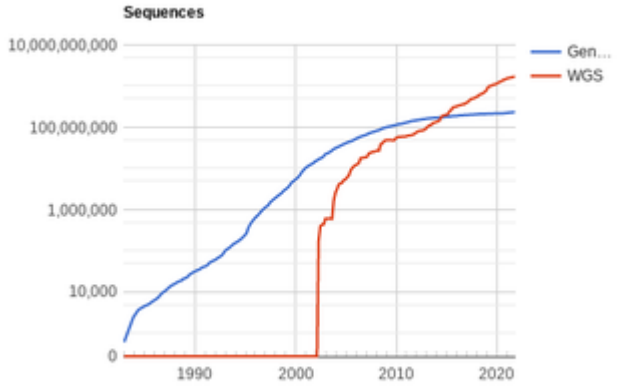
The four modalities of biological research

- Observation driven
  - Collection of observations in nature, i.e., collecting insects
- Theory driven
  - Various ideas that explain the observations
- Hypothesis / experiment driven
  - Suggest ideas of how things work & design an experiment to prove
  - What genes explain a certain trait in a given insect? Design several experiments to find a gene
- **Large scale molecular data accumulation**
  - **No longer hypothesis driven. Efficient but an exhaustive way of generating large scale, high throughput data**
  - **Let's find all genes, all promoters in the genome**

First and foremost: DNA sequence data

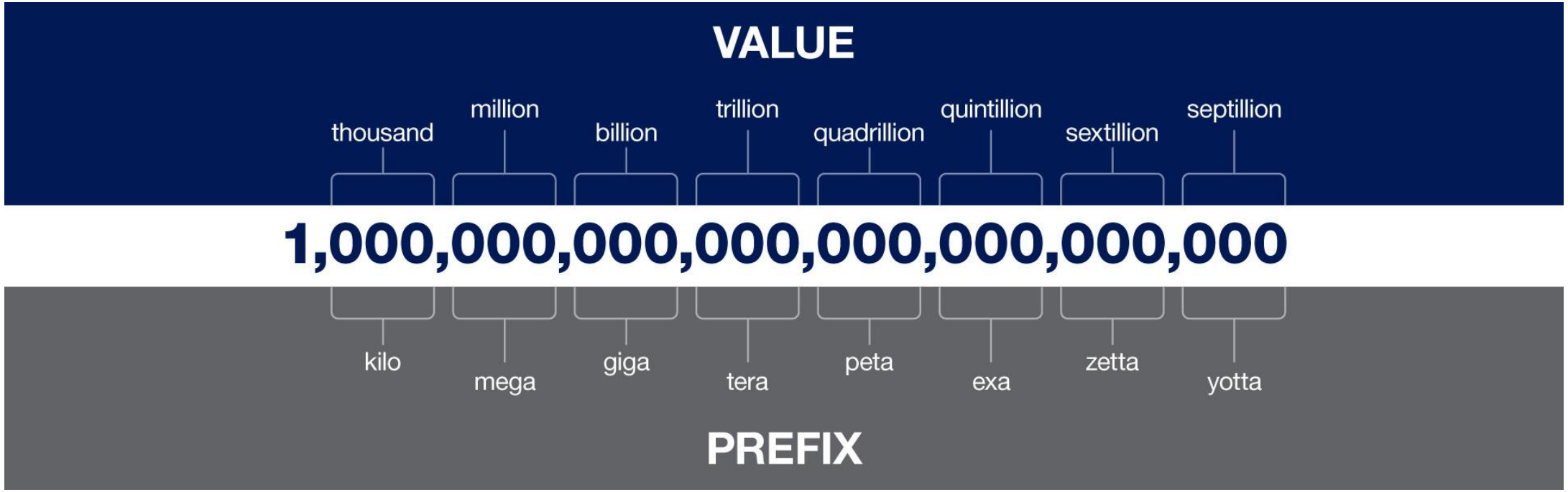


The scale of genomic data



NCBI

- 20 years since human genome project got started
- Between 2 and 40 exabytes of data may be generated in the next decade for genomics research alone



Sequencing data

- The number and complexity of datasets only increasing
- Big and heterogeneous
- Single omics approaches may not be able to handle the scale of the NGS data
- We need tools that can handle, extract and interpret the valuable information hidden within this large trove of data

What is hiding in the high-throughput biological data?

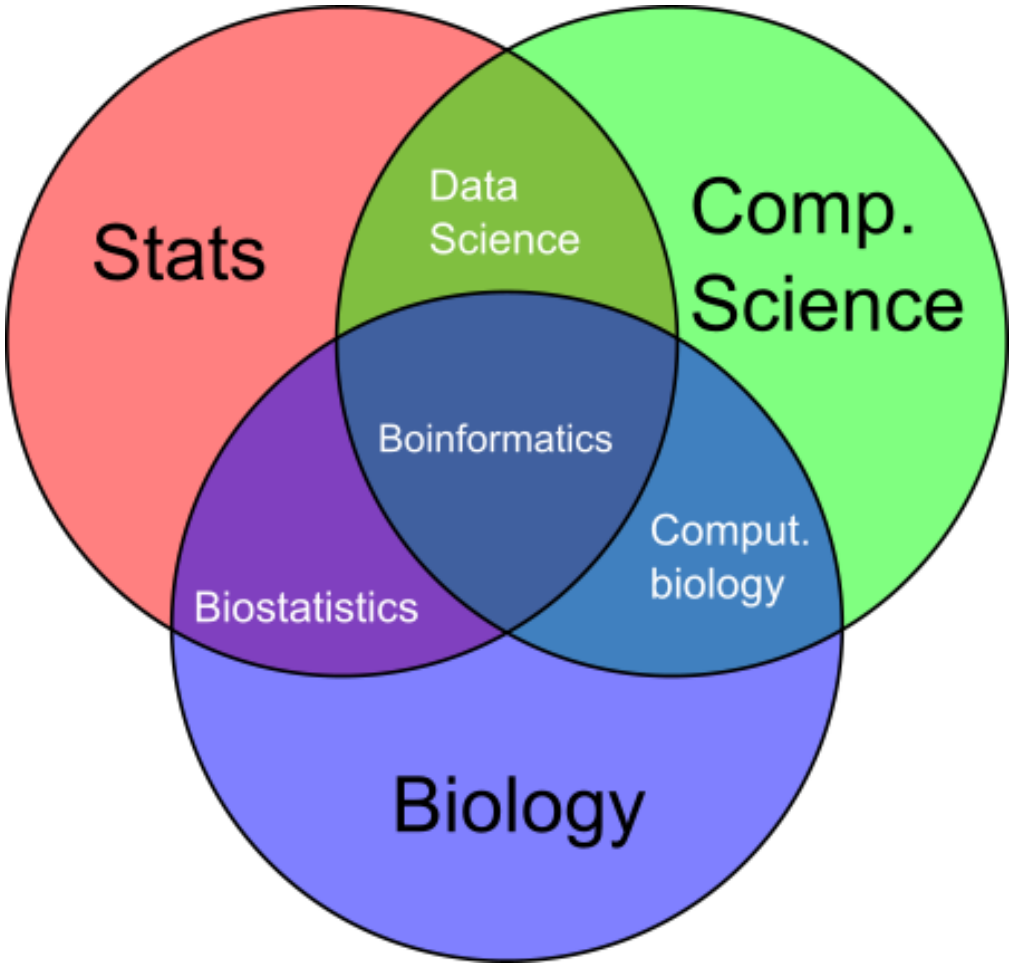
- Hidden in all these data classes is information that reflects
  - existence, organization, activity, functionality …… of biological machineries at different levels in living organisms

Most effectively utilizing and analyzing this information computationally is essential for Bioinformatics

When you have large scale data

- You have to organize it to learn something from it
- Many of the key ideas in bioinformatics predate genomics but genomics is what makes bioinformatics important
- In our class, our focus is on **genomic sequence data and methods**

Bioinformatics



Computational Goals of Bioinformatics

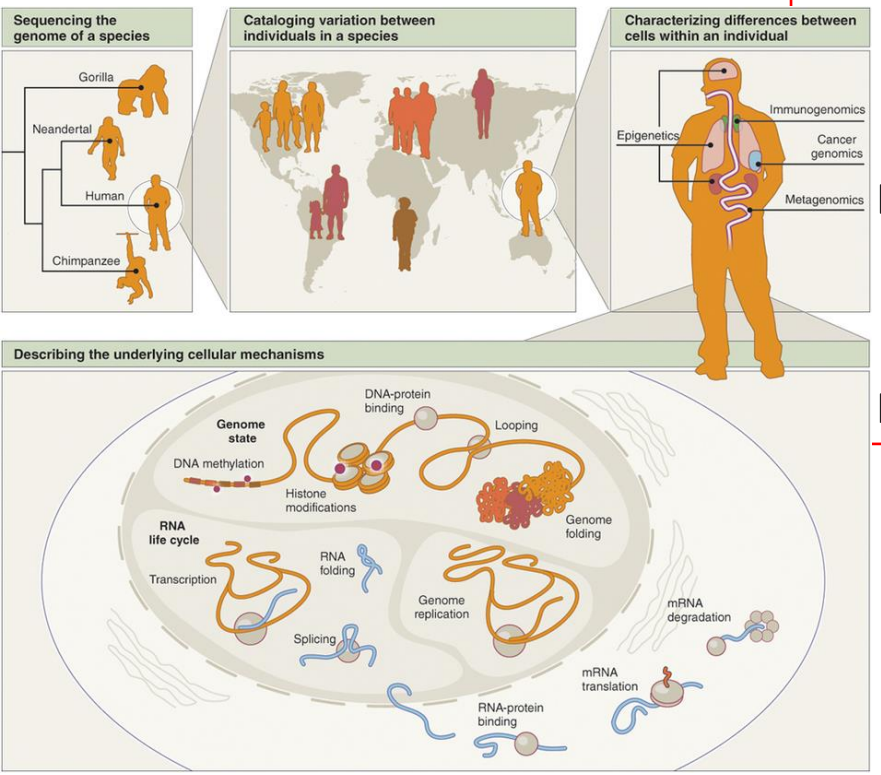
- *Learn & Generalize*: Discover conserved patterns (models) of sequences, structures, interactions, and metabolism from well-studied examples, see Yang et al. 2020
- *Predict*: Infer function or structure of newly sequenced genes, genomes, proteins or proteomes from these generalizations: list of gene prediction software
- *Organize*: Develop a systematic and genomic approach to molecular interactions, metabolism, cell signaling, gene expression, see Anshul Kundaje's research
- *Simulate*: Model gene expression, gene regulation, protein folding, protein-protein interaction, protein-ligand binding… Read up on Deep Mind's AlphaFold
- *Engineer*: Construct novel organisms or novel functions or novel regulation of genes and proteins, see Frances Arnold's research
- *Gene Therapy*: Target specific genes, or mutations, RNAi to change a disease phenotype, see Dyno Therapeutics' ML assisted vector design

What large scale and heterogeneous genomic data you can think of?

What kind of applications of genomic data you can think of?

What patterns can we extract from genomic data?

We can learn a lot from DNA sequence data



Shendure and Aiden 2012

- Profiling
  - microRNAs
  - Immunogenomics
  - Transcriptomics

Epigenomics

- Map histone modifications
- Map DNA methylation

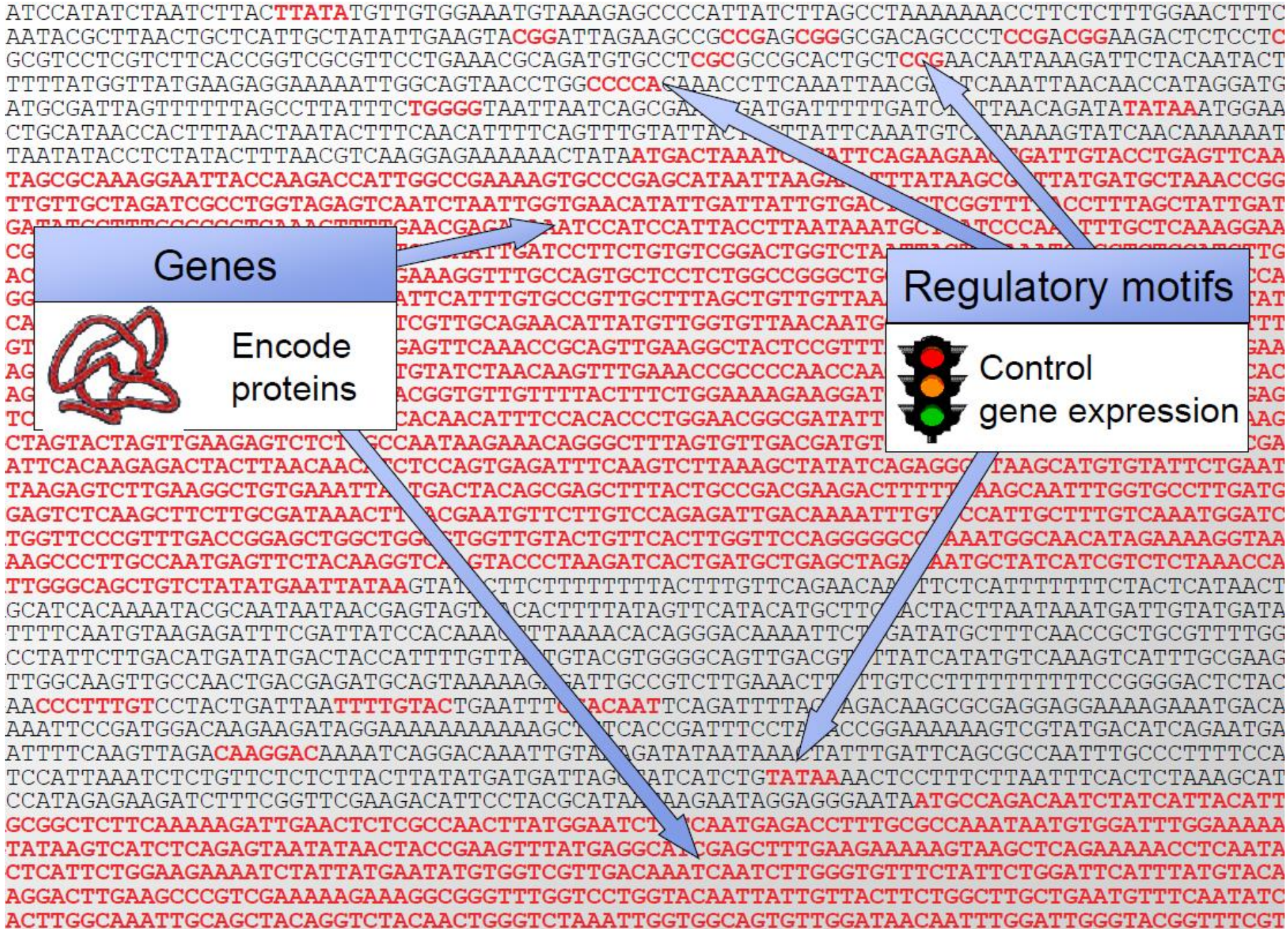
Nucleic acid Interactions

- Cancer genomics
  - Map translocations, CNVs, structural changes
  - Profile somatic mutations
- Genome assembly
- Ancient DNA (Neanderthal)
- Pathogen discovery
- Metagenomics

- Polymorphism/mutation discovery
  - Exon (and other target) sequencing
  - Disease gene sequencing

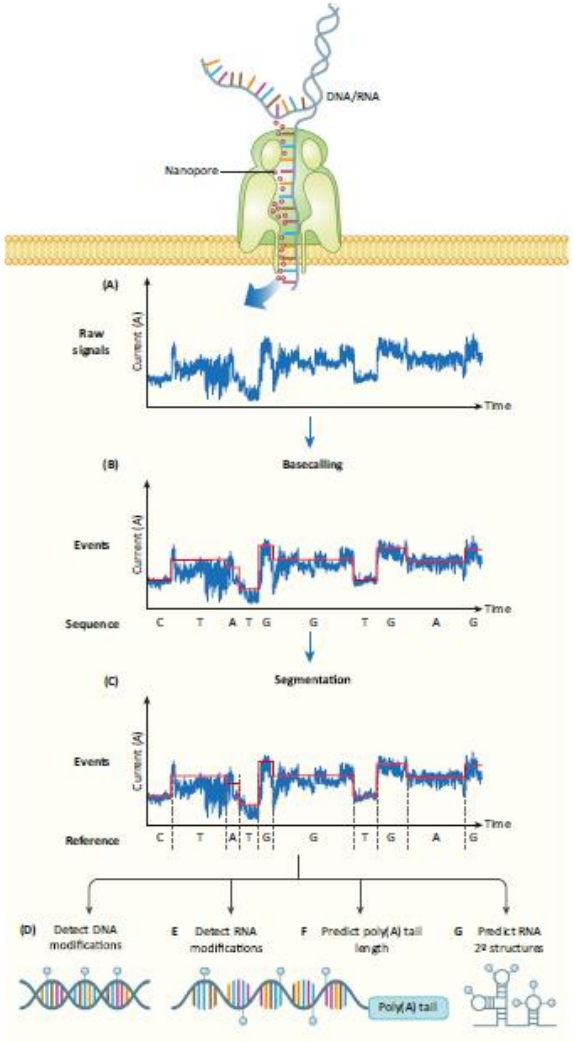
- Variation and association studies
- Genetics and gene discovery

TATTGAATTTTCAAAAATTCTTACTTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTG
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAAACCATGAA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCAC
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAG
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAG
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAACT
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGG
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAG
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTTCCGGGGACTCTAC
AACCCTTTGTCCTACTGATTAATTTTGTACTGAATTTGGACAATTCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGACA
AAATTCCGATGGACAAGAAGATAGGAAAAAAAAAAAGCTTTCACCGATTTCCTAGACCGGAAAAAAGTCGTATGACATCAGAATGA
ATTTTCAAGTTAGACAAGGACAAAATCAGGACAAATTGTAAAGATATAATAAACTATTTGATTCAGCGCCAATTTGCCCTTTTCCA
TCCATTAAATCTCTGTTCTCTCTTACTTATATGATGATTAGGTATCATCTGTATAAAACTCCTTTCTTAATTTCACTCTAAAGCAT
CCATAGAGAAGATCTTTCGGTTCGAAGACATTCCTACGCATAATAAGAATAGGAGGGAATAATGCCAGACAATCTATCATTACATT
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAA
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACA
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATATC
ACTTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCGT


Genes
Encode proteins
Regulatory motifs
Control gene expression

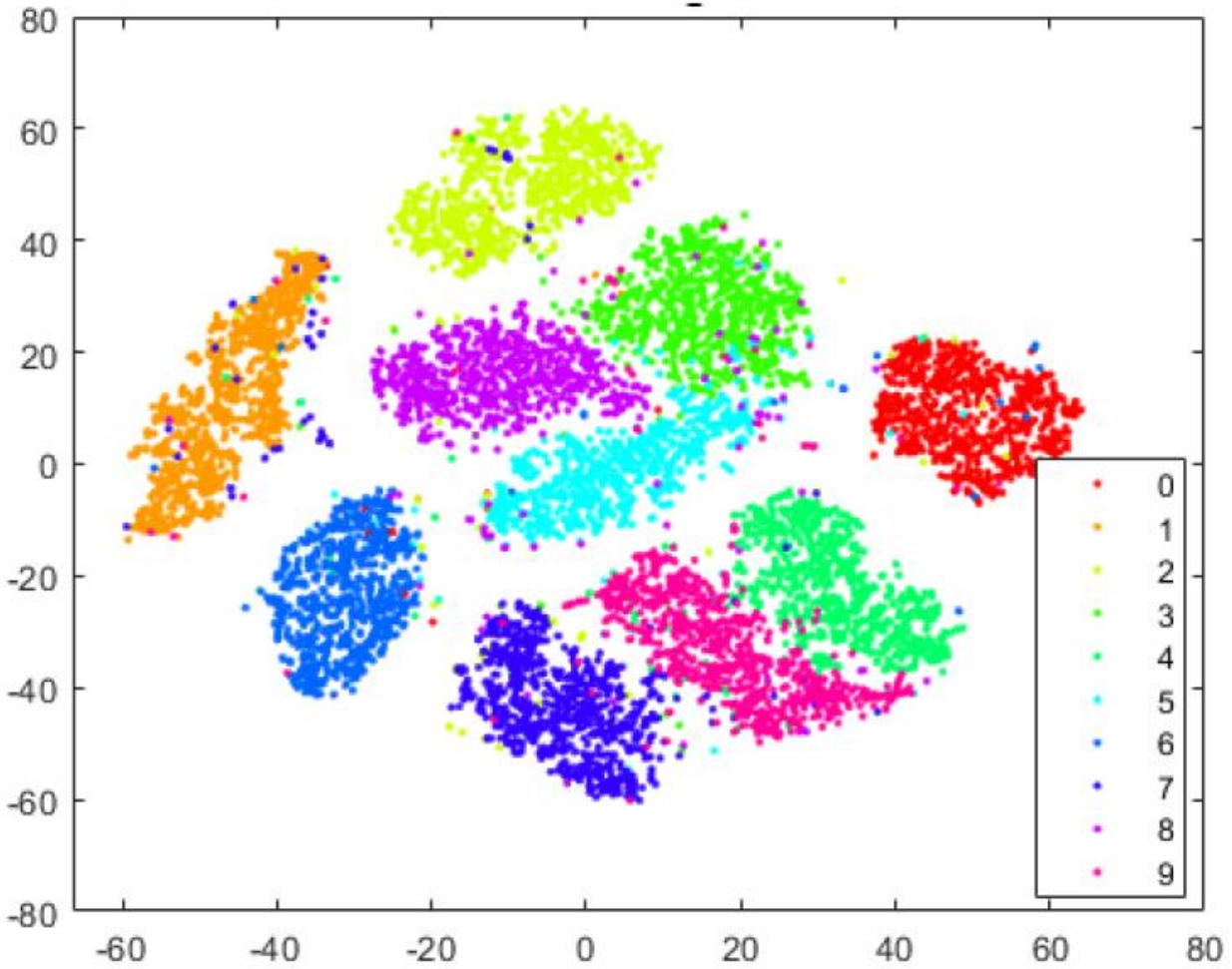
Extracting signal from noise

Extracting biology hidden in Nanopore signal data



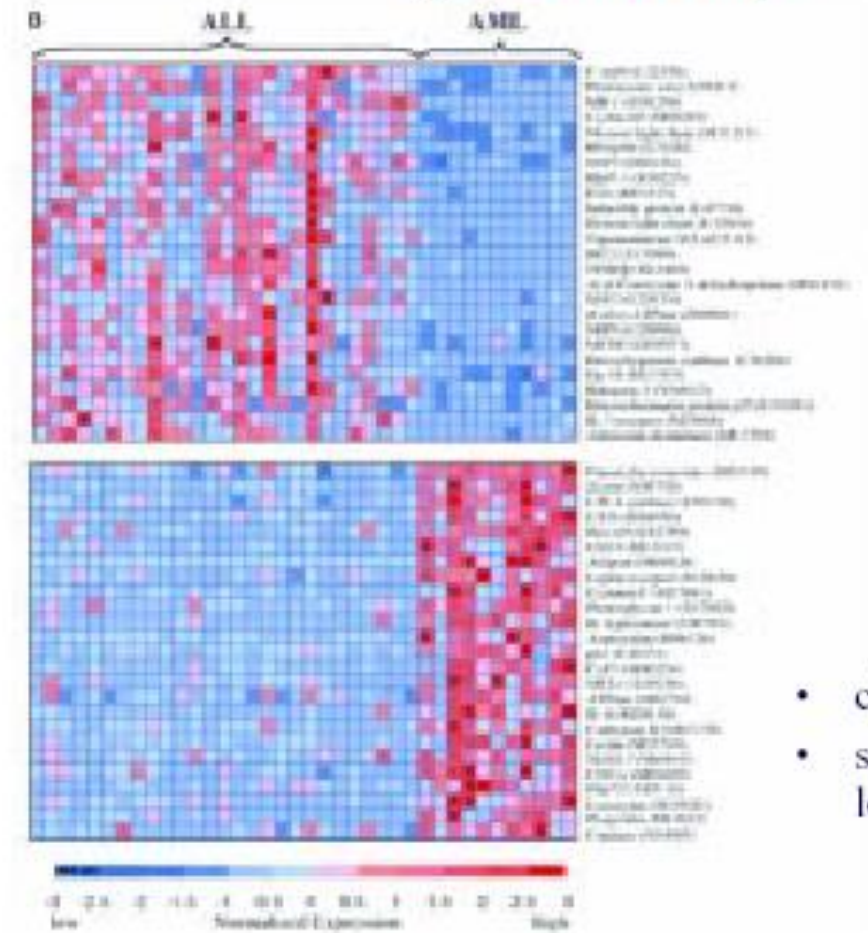
Wan, et al., 2021

Parametric tSNE for single cell RNA-seq data



Expression data

Can diseases be characterized by patterns of gene activity?

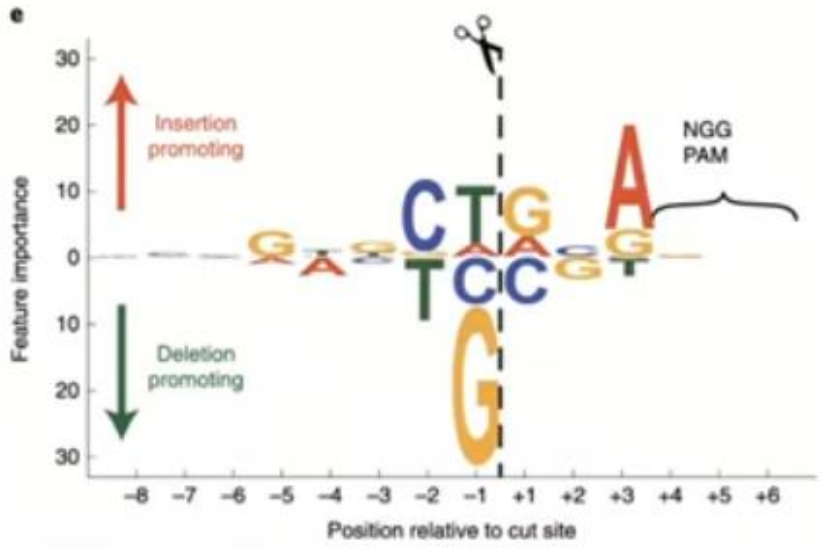


- clustering
- supervised machine learning

Predicting CRISPR repair sites

SPROUT: CRISPR outcome prediction

Spacer DNA is a region of non-coding DNA between genes.



Using machine learning model (gradient boosting decision trees) to predict CRISPR repair outcomes at a given target site.

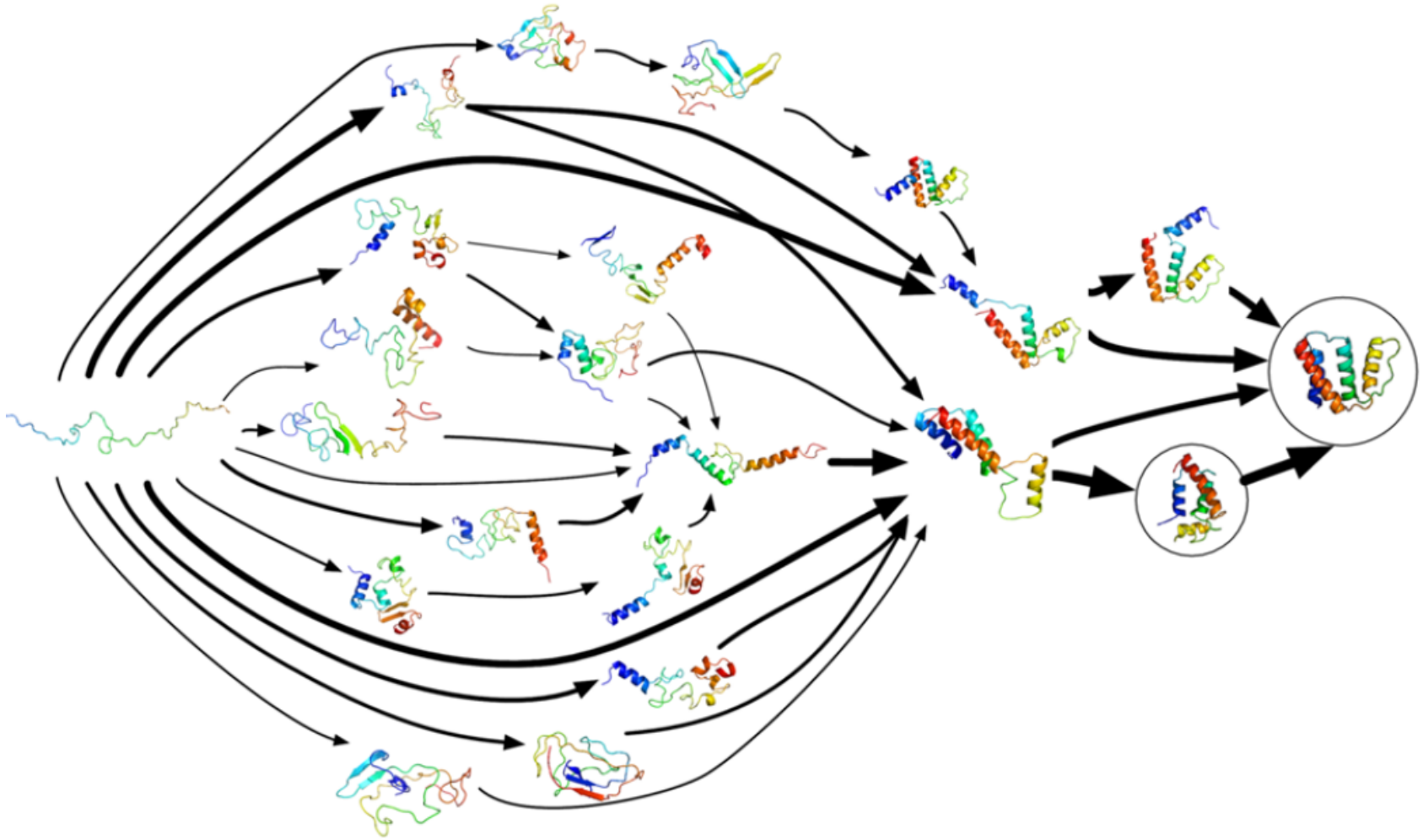
1656 unique target locations within 559 genes, with an average of 98 discrete repair outcomes per target site.

Input:
Local DNA sequence (20NT) of the spacer sequence plus the protospacer adjacent motif (PAM).

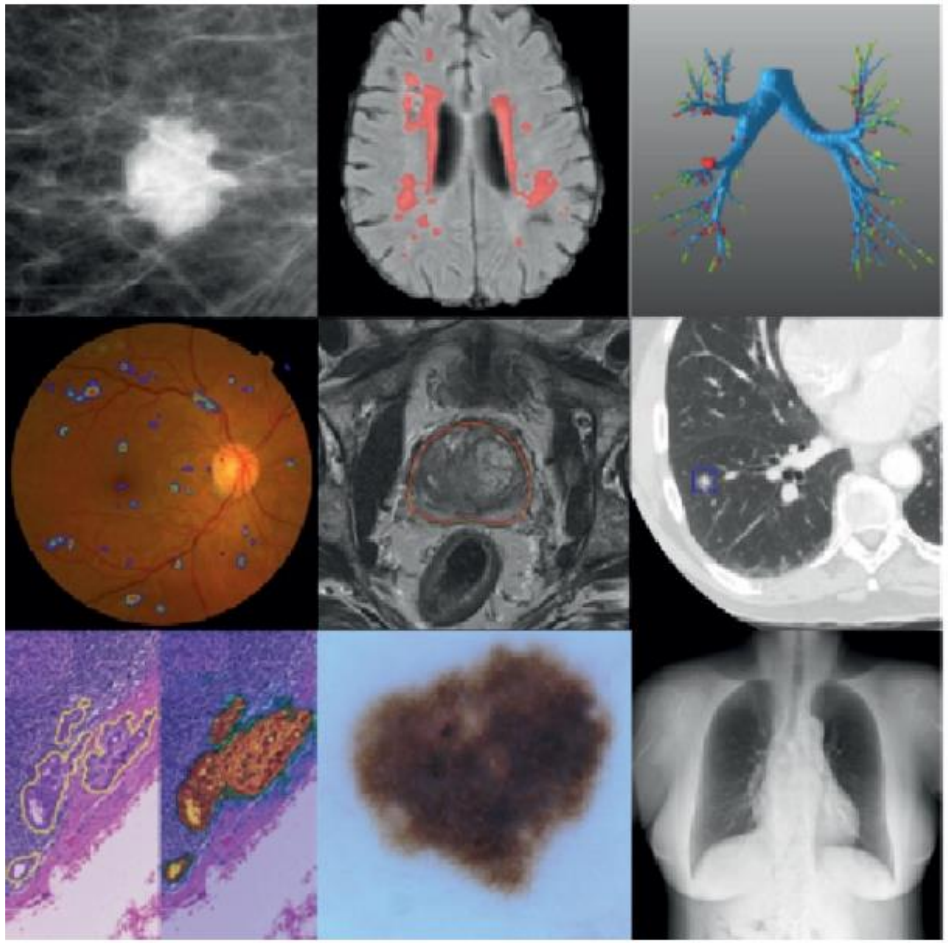
Output:
Predict the length, probability and sequence of nucleotide insertions and deletions.

Leenay et al., 2019

Beyond sequencing: Protein folding and drug design



Beyond sequencing: Medical image analysis



Collage of some medical imaging applications in which deep learning has achieved state-of-the-art results.

From top-left to bottom-right:

1. mammographic mass classification
2. segmentation of lesions in the brain,
3. leak detection in airway tree segmentation,
4. diabetic retinopathy classification
5. prostate segmentation,
6. nodule classification,
7. breast cancer metastases detection,
8. skin lesion classification
9. bone suppression

Some genomics problems to think about in bioinformatics/machine learning scope

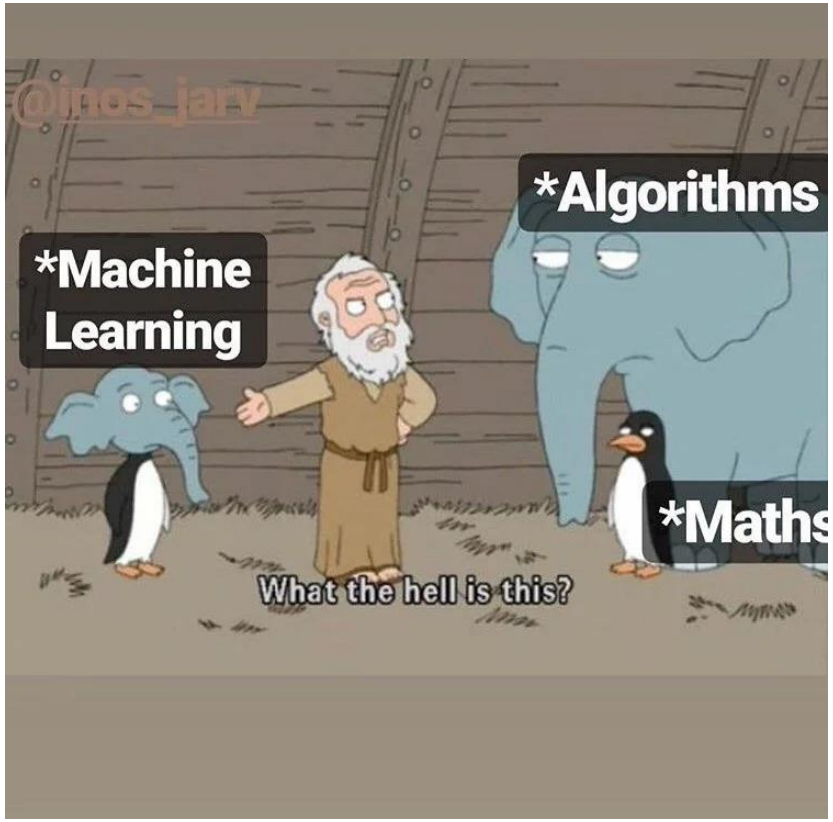
- Identifying protein binding sites
- Sequencing the genomes of different organisms in evolutionary studies
- Profiling the genomes of individuals in medical studies for the purpose of discovering variations/variants
- Sequence characteristics from a stretch of DNA, such as the frequency of some nucleotides or presence of certain motifs
- The binding of proteins to DNA and RNA
- RNA secondary structures

Machine learning

So, what exactly is machine learning and how does it work?

Lots of complicated math
Lots of data
Statistical tools

- Automating automation
- Getting computers to program themselves
- Let the data do the work instead!



What is machine learning?

What is Machine Learning?

[Shalev-Shwartz and Ben-David, 2014]:

"Learning is the process of converting experience into expertise or knowledge."

[Mohri et al., 2012]:

"Machine learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions."

[Murphy, 2012]:

"The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest."

[Hastie et al., 2001]:

"[...] state the learning task as follows: given the value of an input vector $\boldsymbol{x}$ *, make a good prediction of the output* $\boldsymbol{y}$ *, denoted by* $\hat{\boldsymbol{y}}$*"*

What is machine learning?

What is Machine Learning?

A computer program is said to learn from **experience E**

with respect to some **class of tasks T**

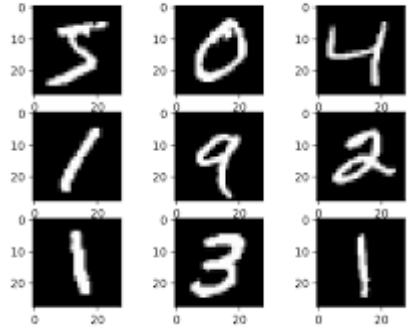
and **performance measure P,**

if its performance at tasks in T, as measured by P, improves with experience E.

[Mitchell, 1997]

- experience E: training set of images of handwritten digits with labels (training set)
- task T: classifying handwritten digits within new images (test set)
- performance measure P: percent of test set digits correctly classified in new images (test set)

What is machine learning?



$$f : X \to Y$$
$$f(x;\theta) = \hat{y}$$

θ:

- **weights** and **biases** (intercepts)
- coefficients β
- parameters

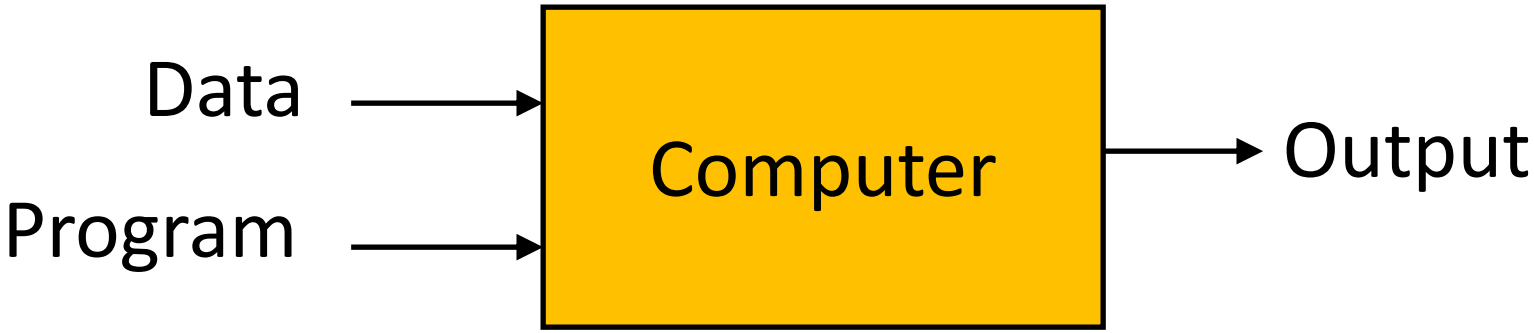
f:

- model
- hypothesis h
- classifier

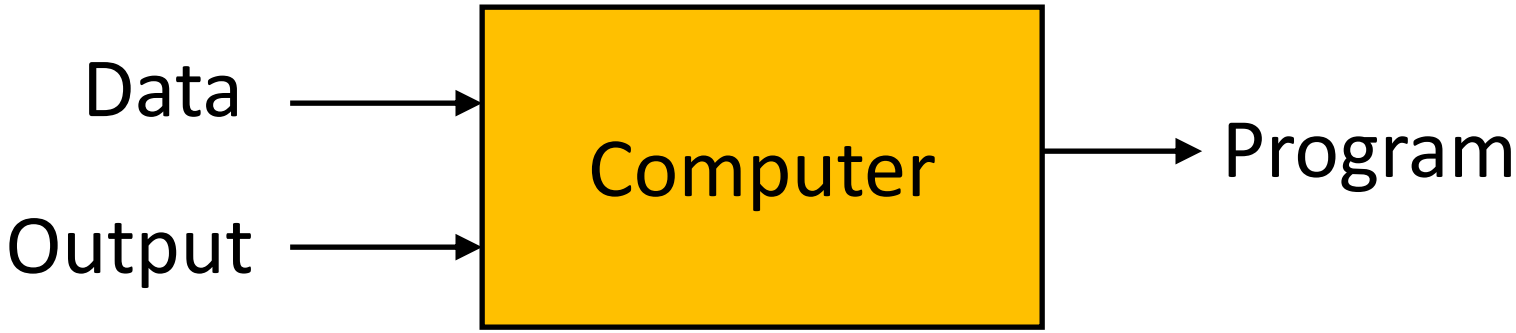
Problem Set 1

$$x \in [0,1]^{784}$$
$$\hat{y} \in [0,1]^{10}$$
$$W \in \mathbb{R}^{784\times 10}$$
$$b \in \mathbb{R}^{10}$$
$$f(x;W,b) = \varphi_{\text{softmax}}(W^{\mathsf{R}}x + b)$$

Traditional Programming



Machine Learning

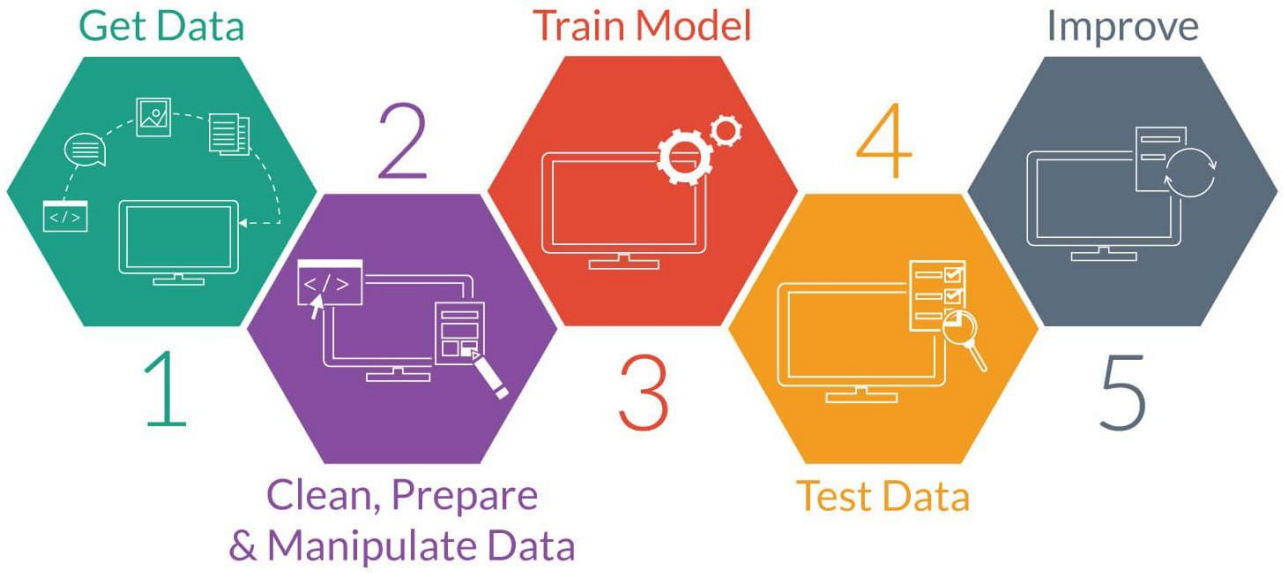


Generally speaking, this is how ML models work

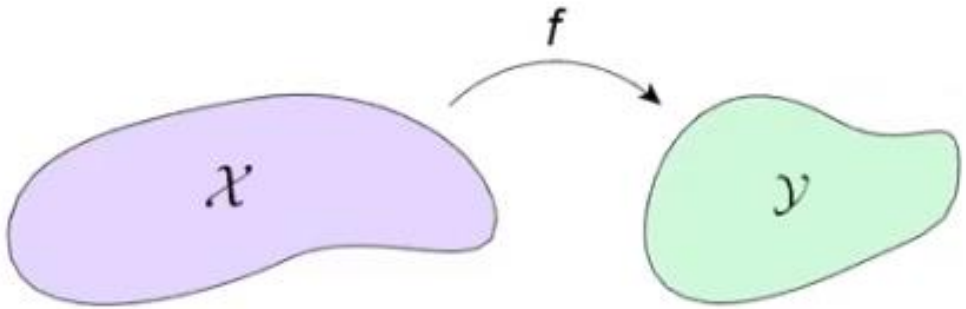
- All ML methods need to learn from input data
- Majority of them require a training set (we will talk about it)
- After training, another data set (usually a part of the original database not used for training) is used to validate and select the best-fit ML model

Basics of machine learning

- Get data
- Formulate an objective, clean up data
- Choose algorithm
- Train (loss)
- Validate results (metrics)



Some terminology in machine learning we will cover



Input $\boldsymbol{X} \in \mathrm{X}$:

- **features** (in machine learning)
- predictors (in statistics)
- independent variables (in statistics)
- regressors (in regression models)
- input variables
- covariates

Output $\boldsymbol{y} \in \mathrm{Y}$:

- **labels** (in machine learning)
- responses (in statistics)
- dependent variables (in statistics)
- regressand (in regression models)
- target variables

Training set $S_{\text{training}} = \{(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{N} \in \{\mathrm{X}, \mathrm{Y}\}^N$, where N is number of training examples

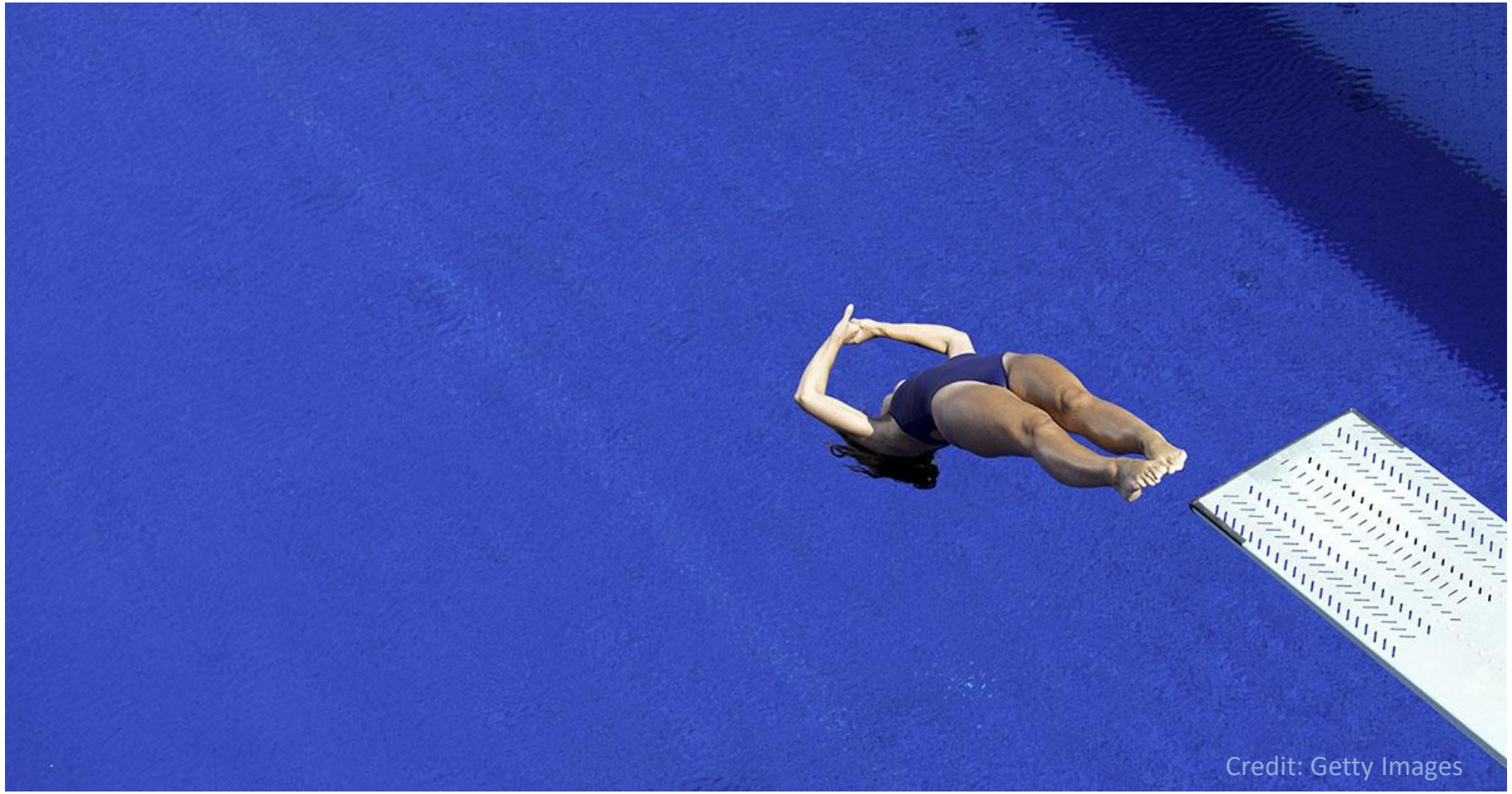
An example is a collection of features (and an associated label)

Training: use S_{training} to learn functional relationship $f : \mathrm{X} \rightarrow \mathrm{Y}$

Some terminology in machine learning we will cover

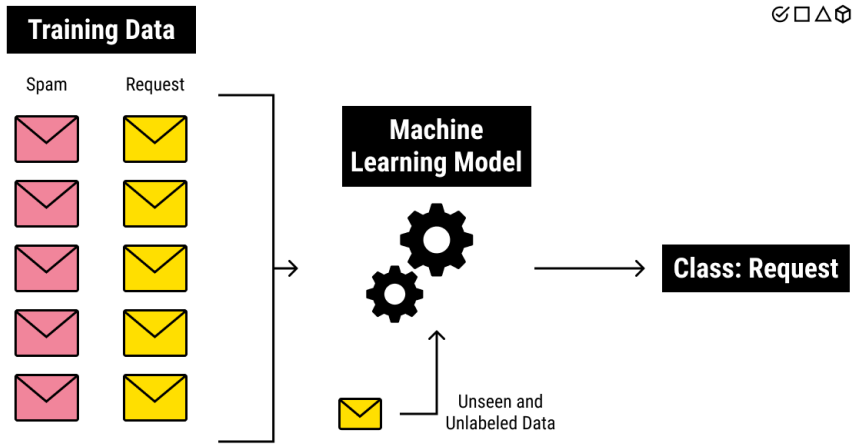
- Train, test and validation datasets
- Model bias and variance
- Underfitting
- Overfitting
- Cost function
- Loss functions
- Hyperparameter and hyperparameter tuning
- Optimization algorithms
- Cross validation
- Evaluation metrics: Accuracy, precision, recall

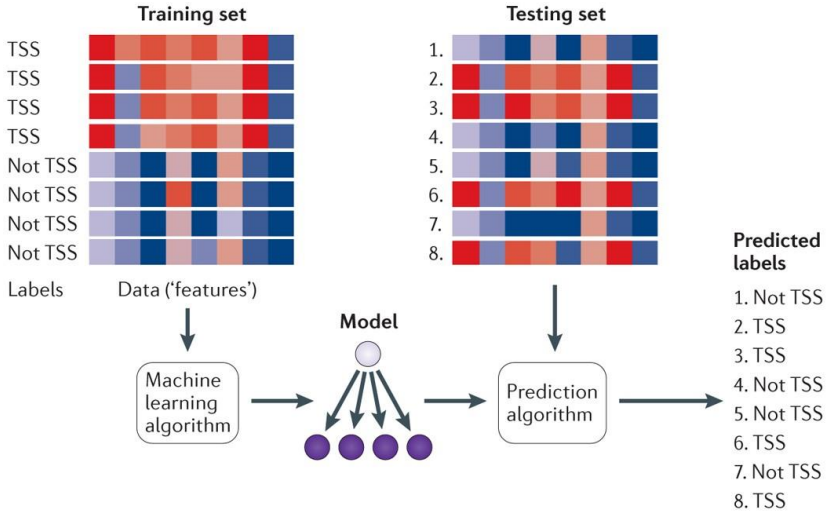
Let’s dive in!



Training set

- Data sample employed to fit the model (i.e., to find the parameters that can best describe the full dataset) and the performance of an ML algorithm significantly depends on it. **It teaches the model what the expected output looks like**
  - If too small
    - may lead to *underfitting*
  - If it has too much data
    - may lead to *overfitting*

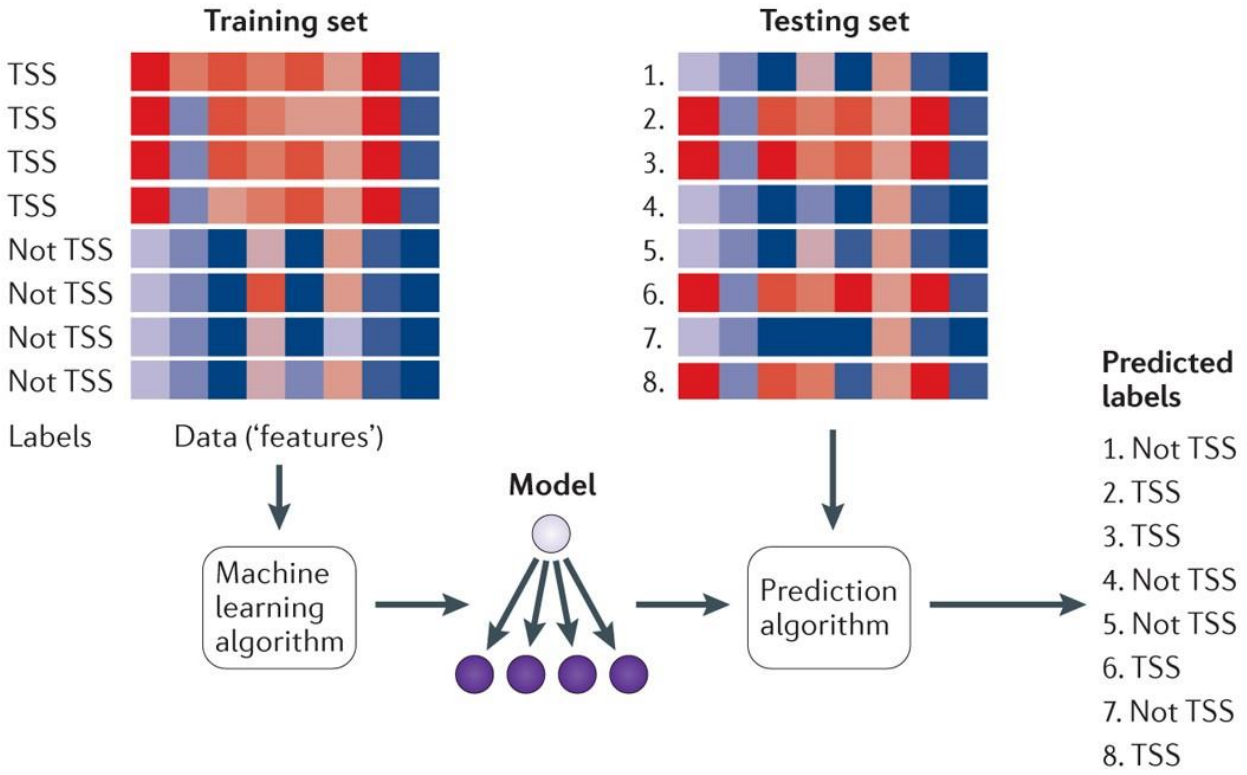




Libbrecht and Noble, 2015

Test set

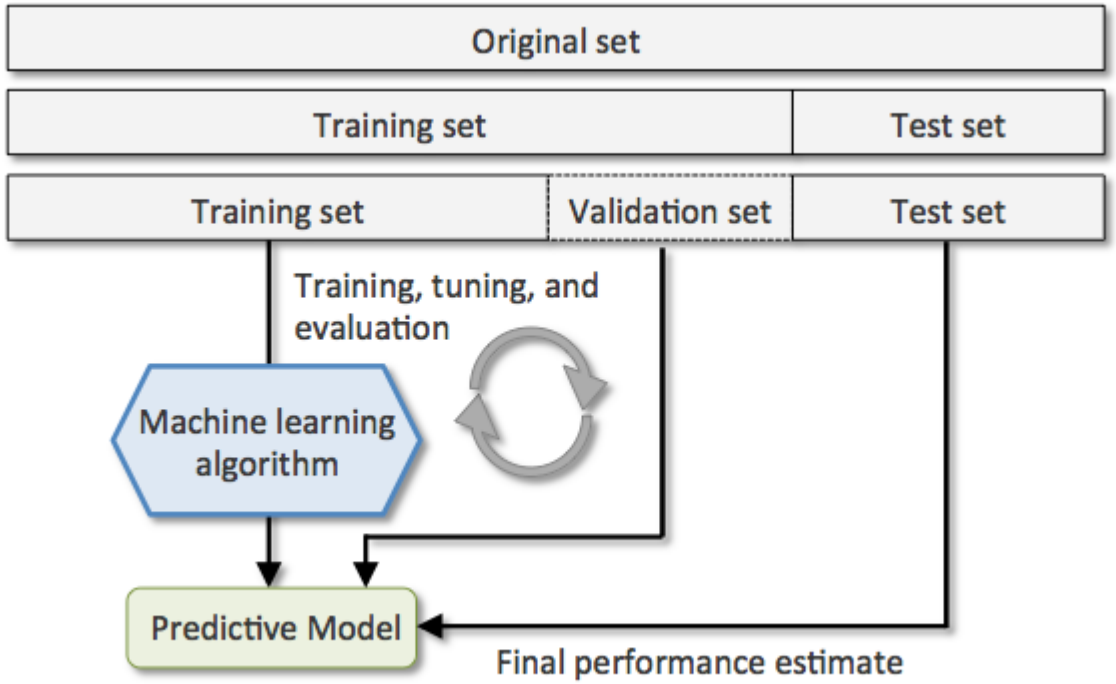
- Provides an unbiased estimation of a final *model fit* on the training dataset. It returns the actual performance of the model and is only used once the model has been fully trained.



Libbrecht and Noble, 2015

Validation set

- Dataset used to evaluate the model fit on the training dataset and is employed to fine-tune the *model hyperparameters*



Training, validation, and test sets

Training set ($S_{training}$):

- set of examples used for learning
- usually 60 - 80 % of the data

Validation set ($S_{validation}$):

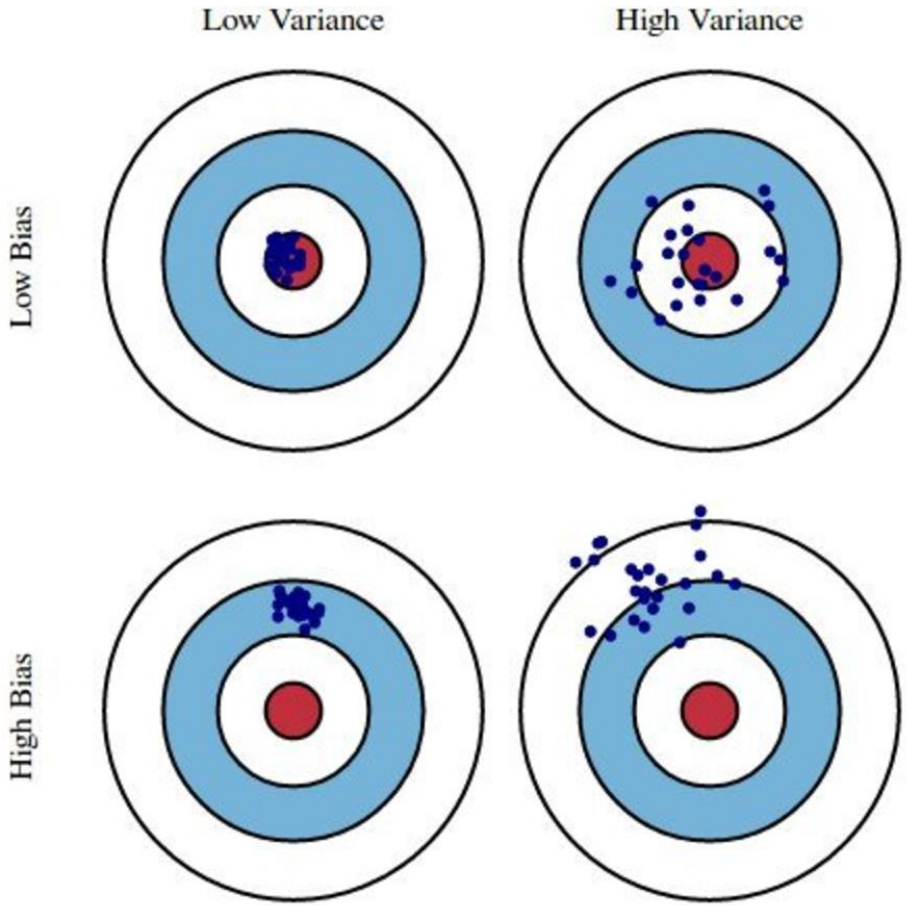
- set of examples used to tune the model hyperparameters
- usually 10 - 20 % of the data

Test set (S_{test}):

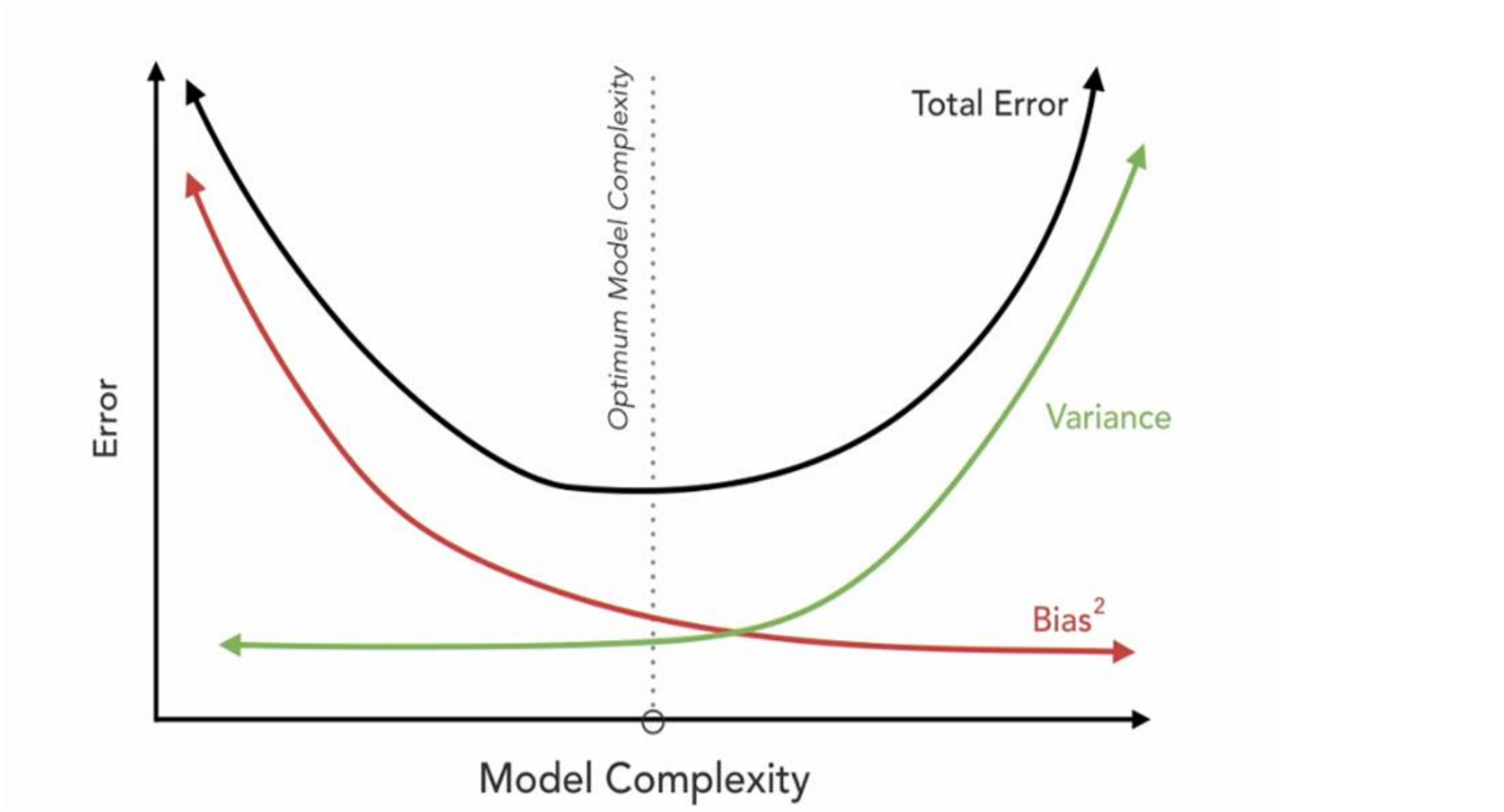
- set of examples used only to assess the performance of fully-trained model
- after assessing test set performance, model must not be tuned further
- usually 10 - 30 % of the data

Bias and variance in ML

- Bias: The difference between the average prediction of the model and the expected value we are trying to predict
  - A model with high bias makes wrong assumptions about the data
- Variance: The variability of the model's predictions. Variance indicates how sensitive the model is to the randomness of the data in the training set

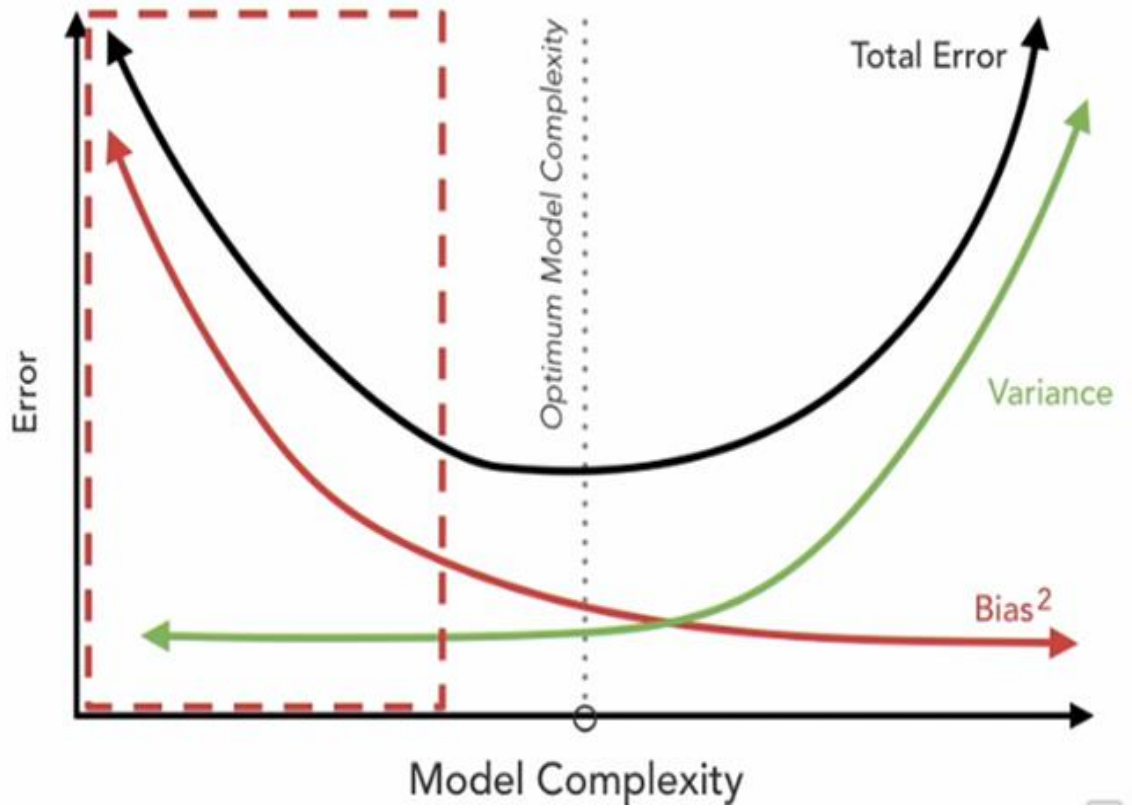


Total error = Bias + Variance + (Irreducible error)

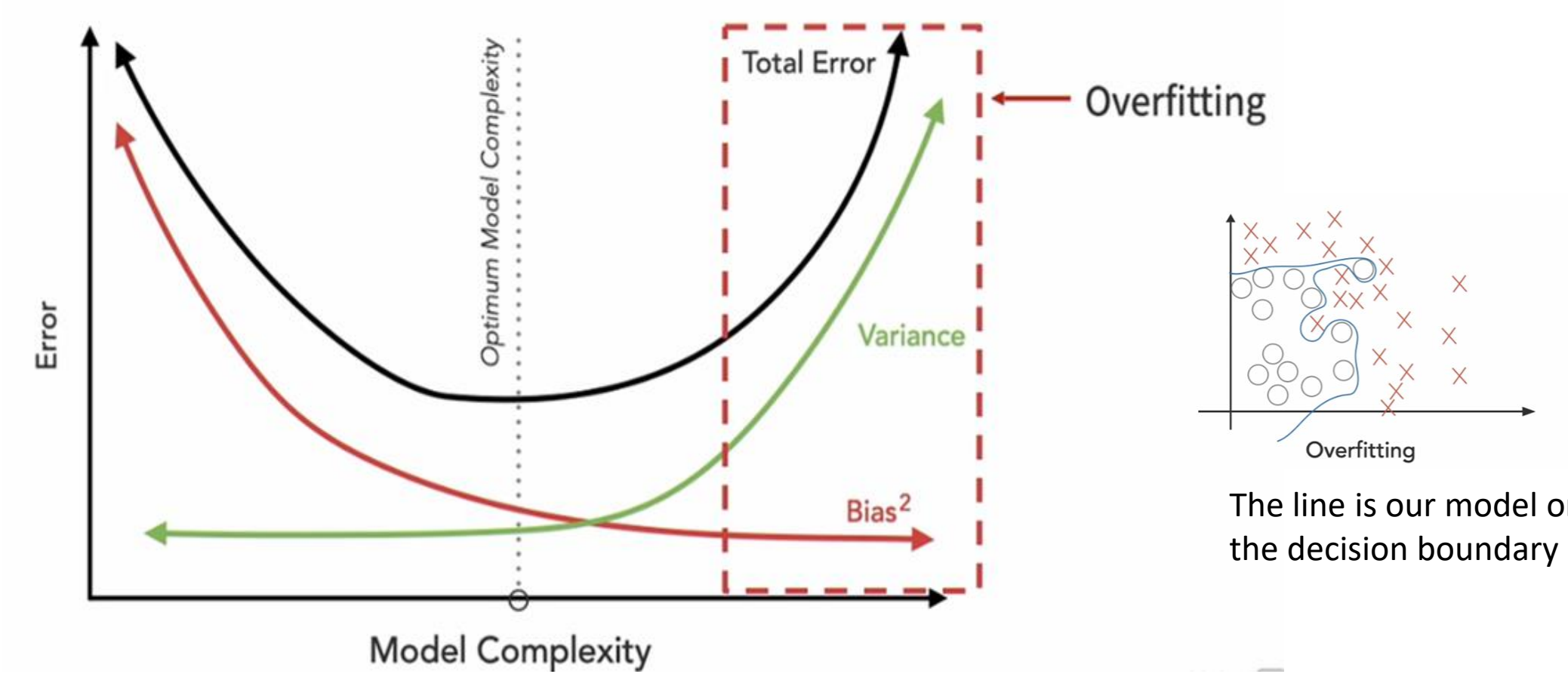


Underfitting = High bias + low variance

Underfitting occurs when an algorithm cannot capture the underlying trend of the data.



Overfitting = Low bias and high variance



The line is our model or the decision boundary

To summarize

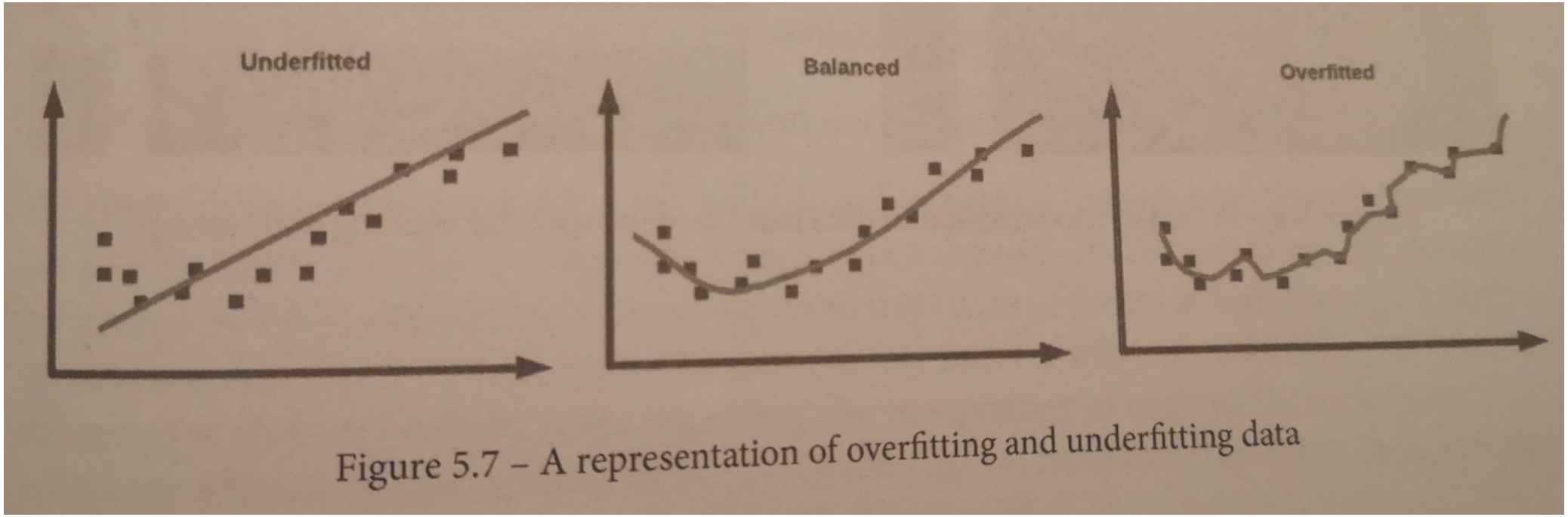
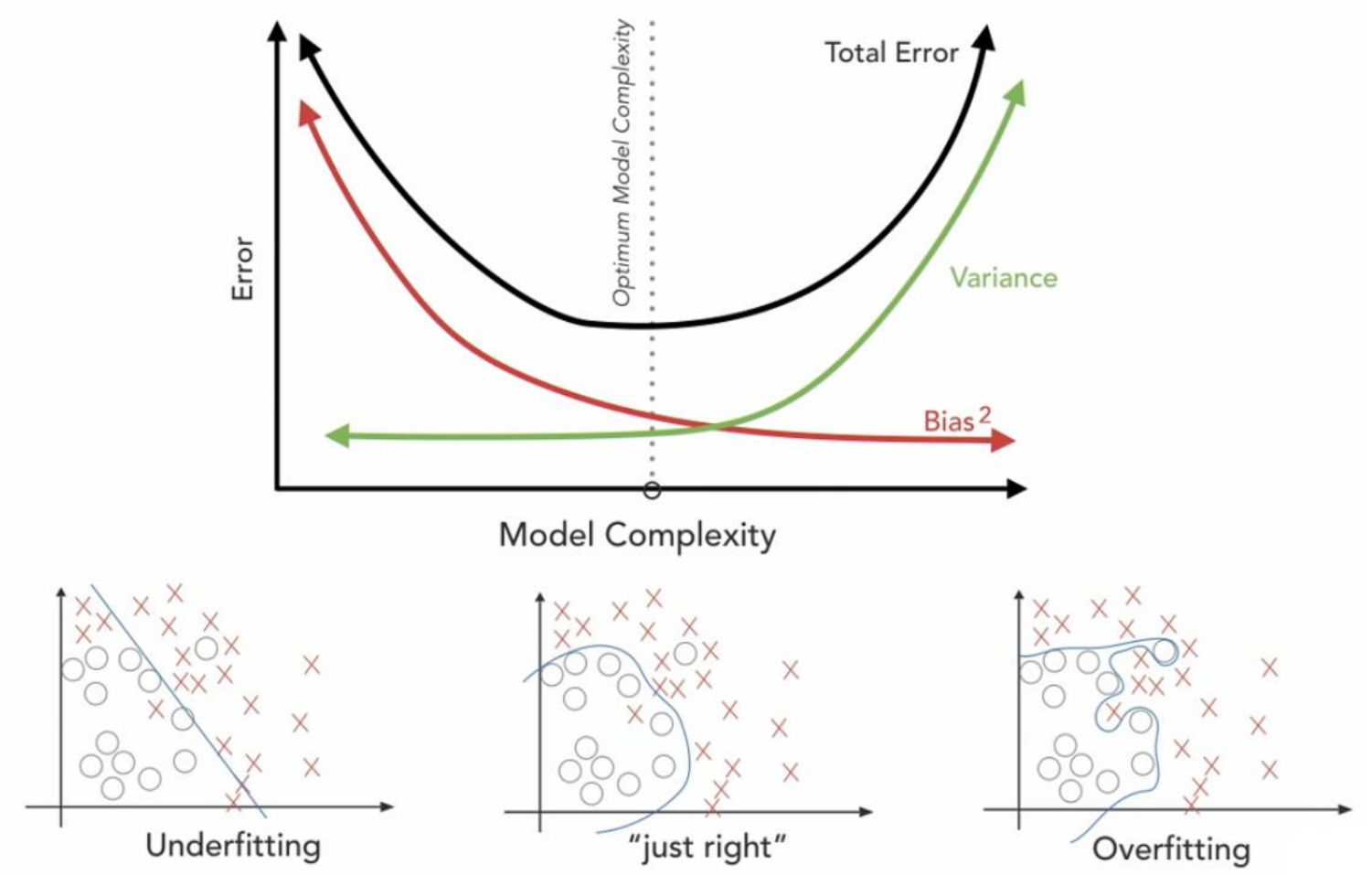


Figure 5.7 – A representation of overfitting and underfitting data

High bias, high training and testing errors, low variance

Low bias, high variance

How do we find the optimal trade off?



Optimal model means

- low bias
- low variance
- low total error

Hyperparameters and hyperparameter tuning

- Choosing a set of optimal hyperparameters for fitting an algorithm. External to the model
  - Train-test split ratio
  - Learning rate in optimization algorithms (e.g. gradient descent)
  - Choice of optimization algorithm (e.g., gradient descent, stochastic gradient descent, or Adam optimizer)
  - Choice of activation function in a neural network (nn) layer (e.g. Sigmoid, ReLU)
  - The choice of cost or loss function the model will use
  - Number of clusters in a clustering task
  - Pooling size
  - Batch size

Parameters

- They are learned or estimated purely from the data during training. Internal to the model
  - The coefficients (or weights) of linear and logistic regression models
  - Weights and biases of a nn
  - The cluster centroids in clustering

Task for Thursday:
Explain the following concepts in high level in class

Concept	Assigned to
Cost function	Ke Wang
Loss function	Cheng-Hsiang Lu
Cross validation	Yutian Liu
Evaluation metrics	Ram Ayyala
Optimization algorithms	Bohan Zhang

Machine learning for genomics data:

What are some ways machine learning is being used in genomics?

Why machine learning in life sciences (more specifically genomics)?

- Enabled by the convergence of three things
  - Inexpensive, high-quality, collection of large data sets (sequencing, imaging, etc.)
  - New machine learning methods
  - High-performance Graphics Processing Unit (GPU) machine learning implementations

Some genomics problems already combined with machine learning tools

- Examining people's faces with facial analysis AI programs to accurately identify genetic disorders
- Using machine learning techniques to identify the primary kind of cancer from a liquid biopsy
- Predicting how a certain kind of cancer will progress in a patient
- Identifying disease-causing genomic variants compared to benign variants using machine learning
- Direct-to-consumer genomics (23andMe, AncestryDNA, etc.)

Most common ML algorithms on NGS single cell sequencing, gene expression and transcriptomics, as of January 2021

Learning algorithm	Example biological applications	#Occurrences
Support Vector Machine	diagnostic classification, intratumoral heterogeneity; tissue-selective genes; gene prediction; gene selection; disease-gene association; gene expression analysis; signatures from gene-pathway; disease gene prioritization; miRNA signatures	157
Random Forest	diagnostic classification, tissue-selective genes; gene prediction; co-acting gene networks; gene selection; mutation-gene-drug relations; gene expression analysis; miRNA biomarkers; drug-induced gene expression; sample-classification	124
Logistic Regression	gene prediction; gene selection; drug-induced gene expression	38
Deep Neural Network	mutation-gene-drug relations; gene expression analysis	36
LASSO	intratumoral heterogeneity; biomarkers selection; gene selection; gene expression analysis	31
Naive Bayes	gene selection; pharmacogenetic prediction	28
K-Nearest Neighbor	gene selection	28
Artificial Neural Network	gene selection; genotype-phenotype analysis; risk classification; transcriptome profiling; variant extraction	25
Autoencoder	gene prediction	24
Principal Component Analysis	single-cell analysis; gene expression	19
Linear Discriminant Analysis	transcriptome profiling; miRNA biomarkers; taxa-condition association	14
Perceptron	gene selection; gene prediction	12
K-means	candidate miRNA targets	8

Monaco et al., 2021

Datasets/repositories

- Have you decided on what questions you find interesting?
- Have you started searching datasets/repositories?
- Potential resources you can consider:
  - Kaggle
  - UCI Machine Learning Repository
  - Google's Datasets Search Engine
  - Amazon and Microsoft Datasets, AWS
  - GTEX Portal
  - National Cancer Institute GDC Data Portal
  - List of datasets for human related ML projects
  - Dataset of human faces with a correctly or incorrectly worn mask
  - Covid-19 open research dataset (more on literature)
  - Medical data for research (MIT's courtesy)
  - Open datasets for ML research (human focused datasets available)

Final Project: Original Research in bioinformatics & ML

- A major aspect of the course is preparing you for original research in computational biology/bioinformatics
    - Framing a biological problem computationally
    - Gathering relevant literature and datasets
    - Solving it using new algorithms, machine learning
    - Interpreting the results biologically
    - Also, ability to present your ideas and research
    - Working in teams of complementary skill sets
    - Receiving feedback and revising your proposal
    - Presenting a research talk to a scientific audience
- Term project experience mirrors this process

References

- Shendure, J. and Aiden, E.L., 2012. The expanding scope of DNA sequencing. *Nature biotechnology*, *30*(11), pp.1084-1094.
- Libbrecht, M.W. and Noble, W.S., 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), pp.321-332.
- Leenay, R.T., Aghazadeh, A., Hiatt, J., Tse, D., Roth, T.L., Apathy, R., Shifrut, E., Hultquist, J.F., Krogan, N., Wu, Z. and Cirolia, G., 2019. Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. *Nature biotechnology*, *37*(9), pp.1034-1037.
- Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y. and Zhang, L., 2020. Review on the application of machine learning algorithms in the sequence data mining of DNA. *Frontiers in Bioengineering and Biotechnology*, p.1032.
- Monaco, A., Pantaleo, E., Amoroso, N., Lacalamita, A., Giudice, C.L., Fonzino, A., Fosso, B., Picardi, E., Tangaro, S., Pesole, G. and Bellotti, R., 2021. A primer on machine learning techniques for genomic applications. *Computational and Structural Biotechnology Journal*, *19*, pp.4345-4359.
- Wan, Y.K., Hendra, C., Pratanwanich, P.N. and Göke, J., 2021. Beyond sequencing: machine learning algorithms extract biology hidden in Nanopore signal data. *Trends in Genetics*.