Next-Generation Machine Learning for Biological Networks

Authors: Diogo M. Camacho, Katherine M. Collins, Rani K. Powers, James C. Costello, and James J. Collins

Presented by: Avirup Guha Neogi

Outline

- Introduction
- Machine Learning
- Types of Machine Learning
- Deep Learning
- Applications of Machine / Deep Learning in Biological Domain
- Conclusion

Why Machine Learning

- A dramatic increase in the number of large, highly complex datasets being generated from biological experiments.
- Example : The Cancer Genome Atlas has sampled multiple -omics measurements from over 30,000 patients across dozens of different cancer types, totaling over 2.5 petabytes of raw data.
- Machine Learning addresses this complexity, providing next-level analyses that allow one to take new perspectives and generate novel hypotheses about living systems.

Machine Learning

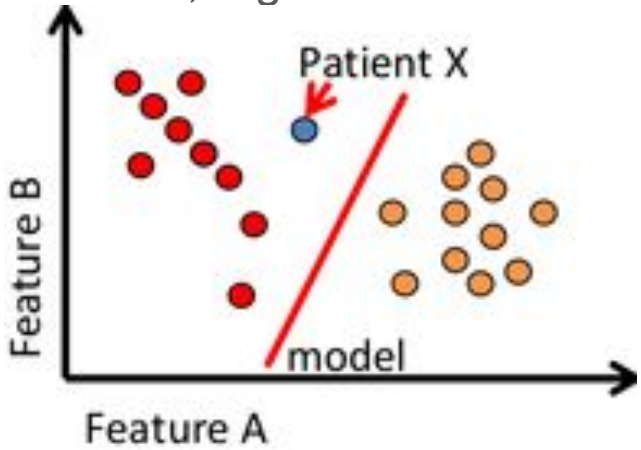
- Learn a generalizable pattern from data through experience
- Sub-branch of Artificial Intelligence.
- Often #variables >> #samples (high dimensional data).
- Aims to generate predictive models based on an underlying algorithm and a given dataset.
- Finds applications in disease biology, drug discovery, microbiome research, and synthetic biology.

Supervised Learning (Predictive Modeling)

- Learning from fully labeled samples.
- Goal: making predictions.

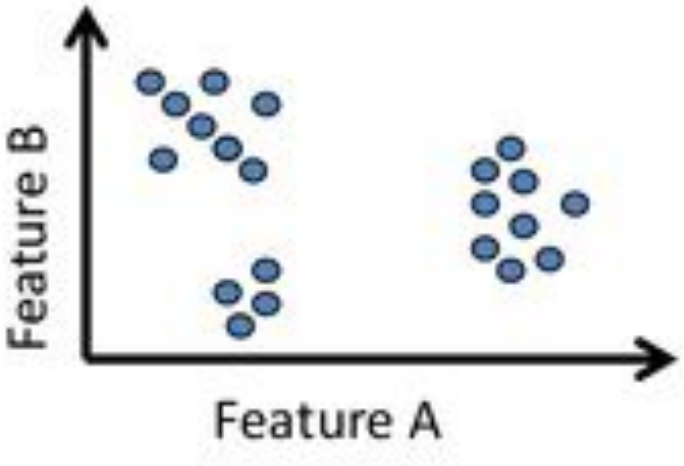
- Primary goal is neither understanding of models nor gaining new knowledge.

- Main examples: Classification, regression.



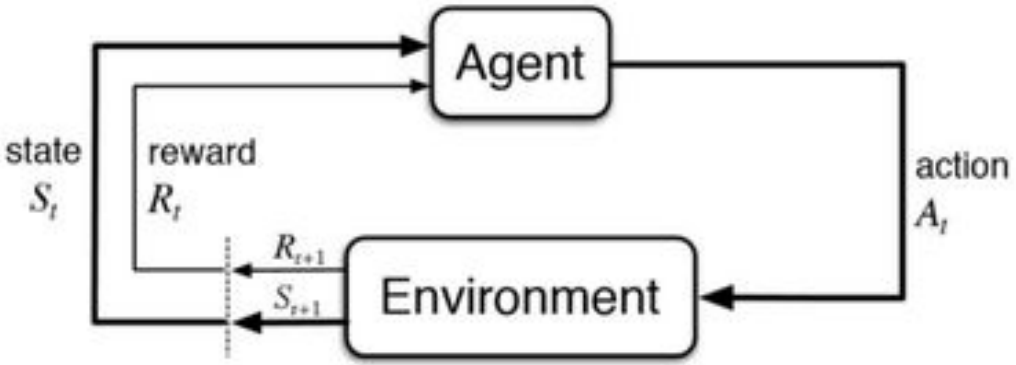
Unsupervised Learning

- Learning without labels.
- Often used in Data Mining
- Example: clustering, PCA



Reinforcement Learning

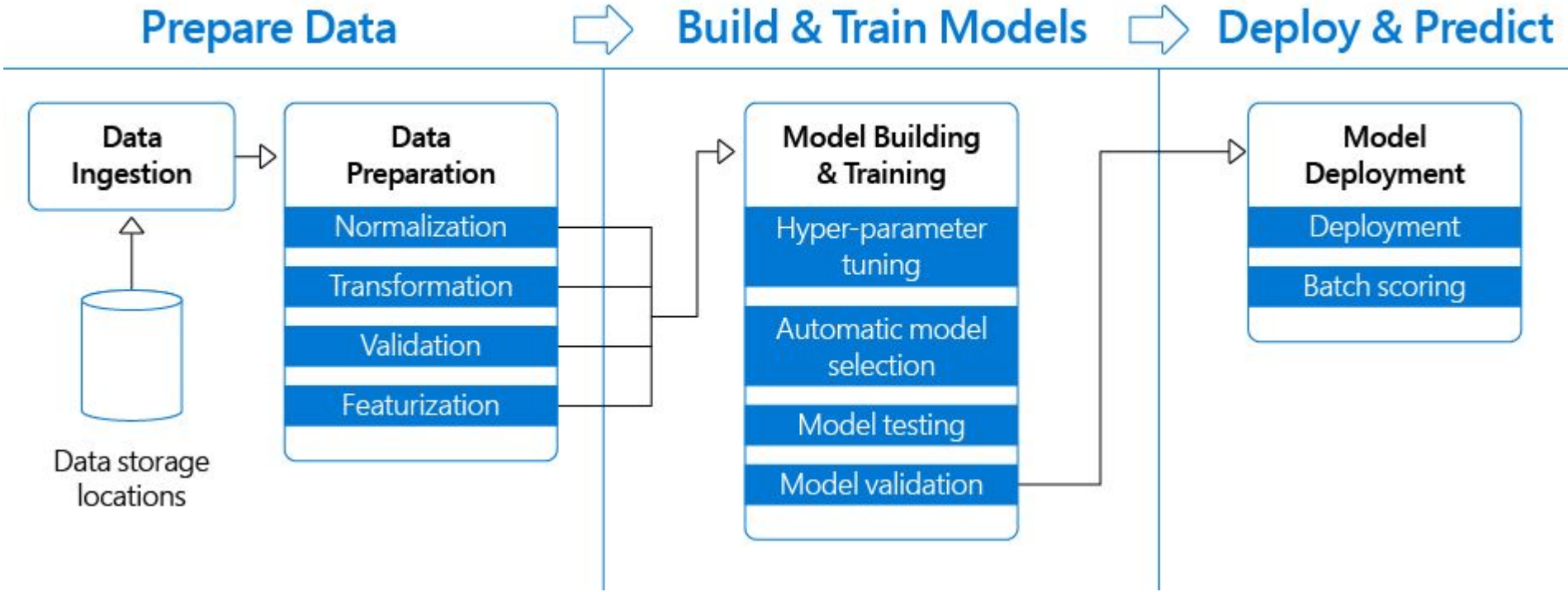
- Learning from criticism (<<good>>, <<bad>>)
- **Goal**: Same as in Supervised Learning
- Often used in Robotics
- Other example: Optimize clinical decision making



Semi-Supervised Learning

- Learning from labels as well as unlabelled samples
- **Goal:** Most often same as supervised learning. Sometimes only labelling of unlabelled training samples(transductive Learning)

Machine Learning Pipeline



Families of Learning Algorithms

Generative / parametric model

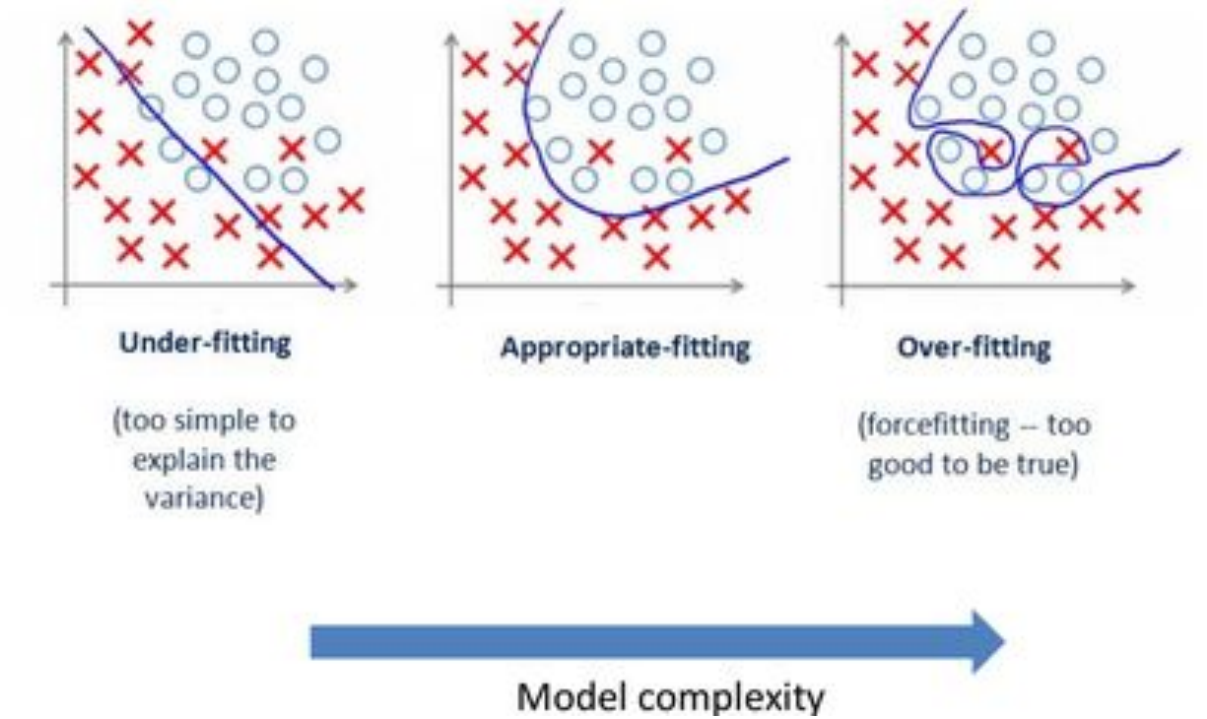
• Mixture models • Hidden Markov Models • Bayesian Networks

Non-parametric models

• k nearest neighbors • Boosting • Decision trees • Support Vector Machines

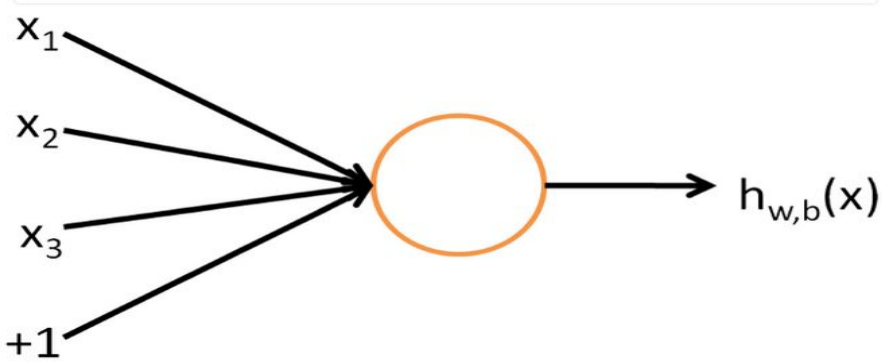
• Neural Networks

Classification Functions of different complexities



Deep Learning

- Neural Networks is one of the most popular machine learning algorithms at present.
- Variants are CNN (Convolutional Neural Networks), RNN(Recurrent Neural Networks), AutoEncoders.
- Neural networks outperform other algorithms in accuracy and speed.
- Its most basic building block is an artificial neuron.

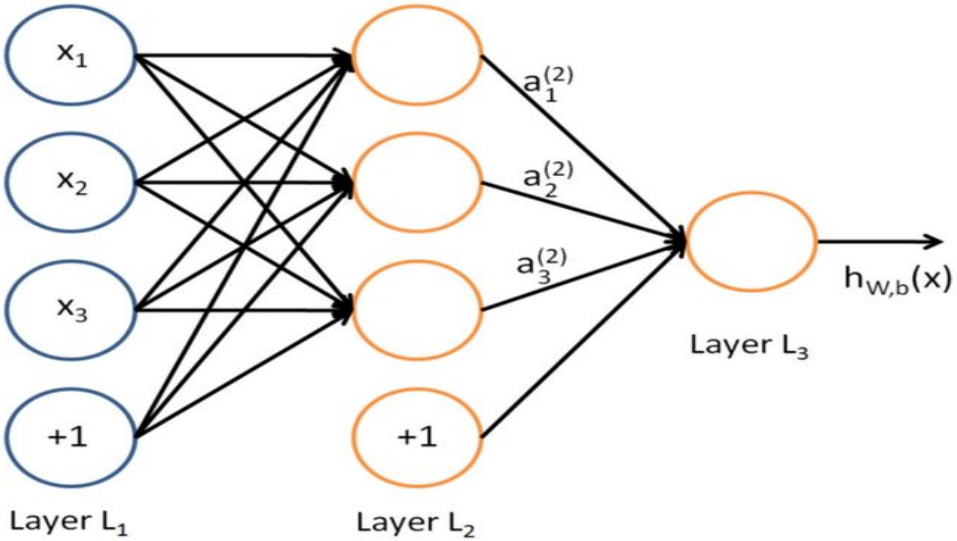


Activation Functions

- Step Function
- Sigmoid Function
- Tanh Function
- ReLU Function
- Leaky ReLU Function

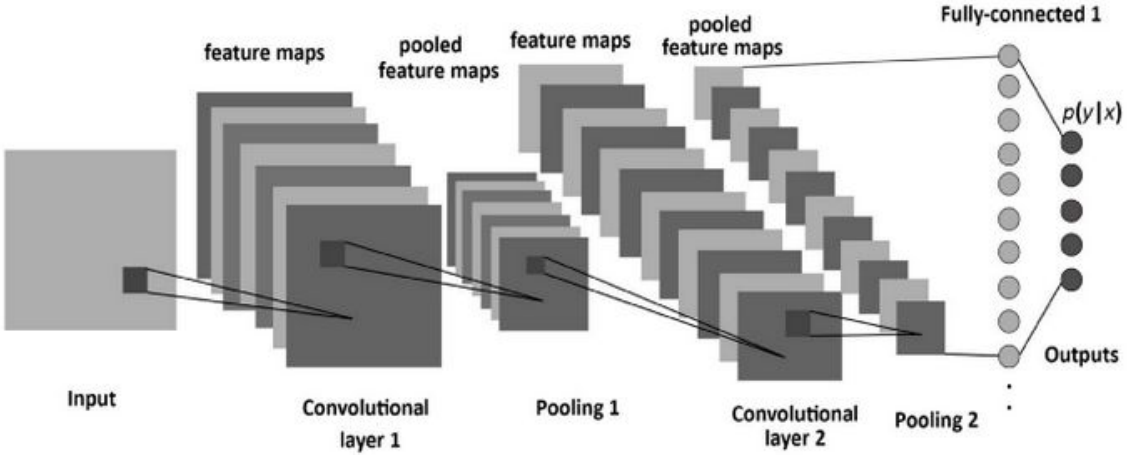
Neural Network

- The Leftmost layer is called the input layer.
- The Rightmost layer is called the output layer.
- The middle layer is called the hidden layer.



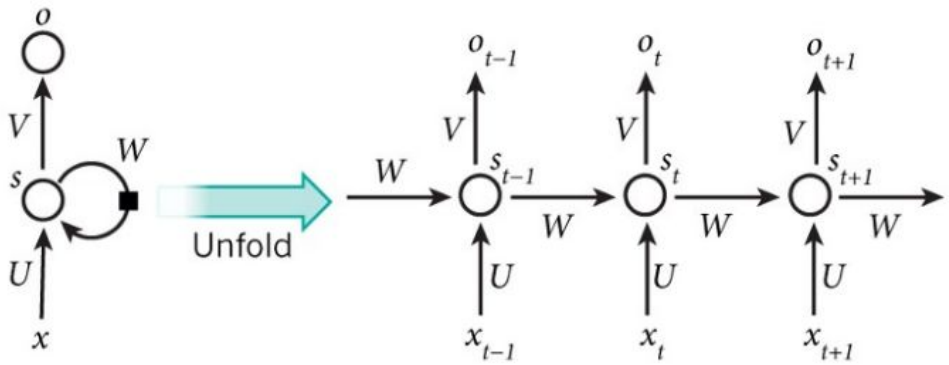
Convolution Neural Network

- Variants of ANN widely used in computer vision.
- The hidden layers of a CNN typically consist of convolutional layers, pooling layers, fully connected layers, and normalization layers.



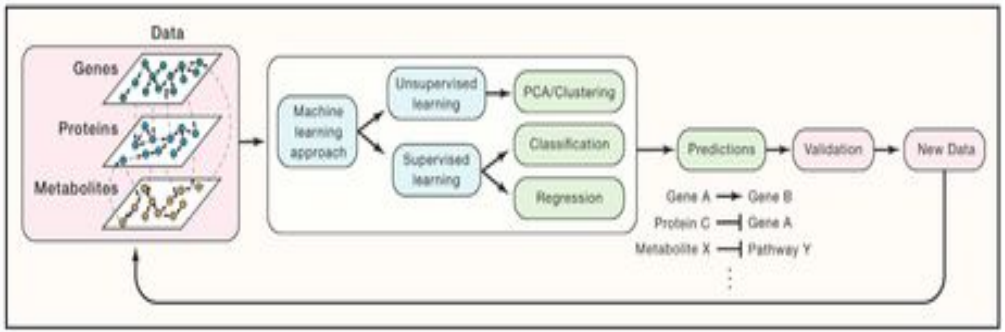
Recurrent Neural Network

- Another variant of ANN heavily usd in Natural Language Processing.
- RNNs are called *recurrent* because they perform the same task for every element of a sequence, with the output being depended on the previous computations.



Applying Machine Learning in Biological Context

- A machine-learning model trained on one dataset may not generalize well to other datasets.
- The new data should also be processed using the same pipeline as the training data.



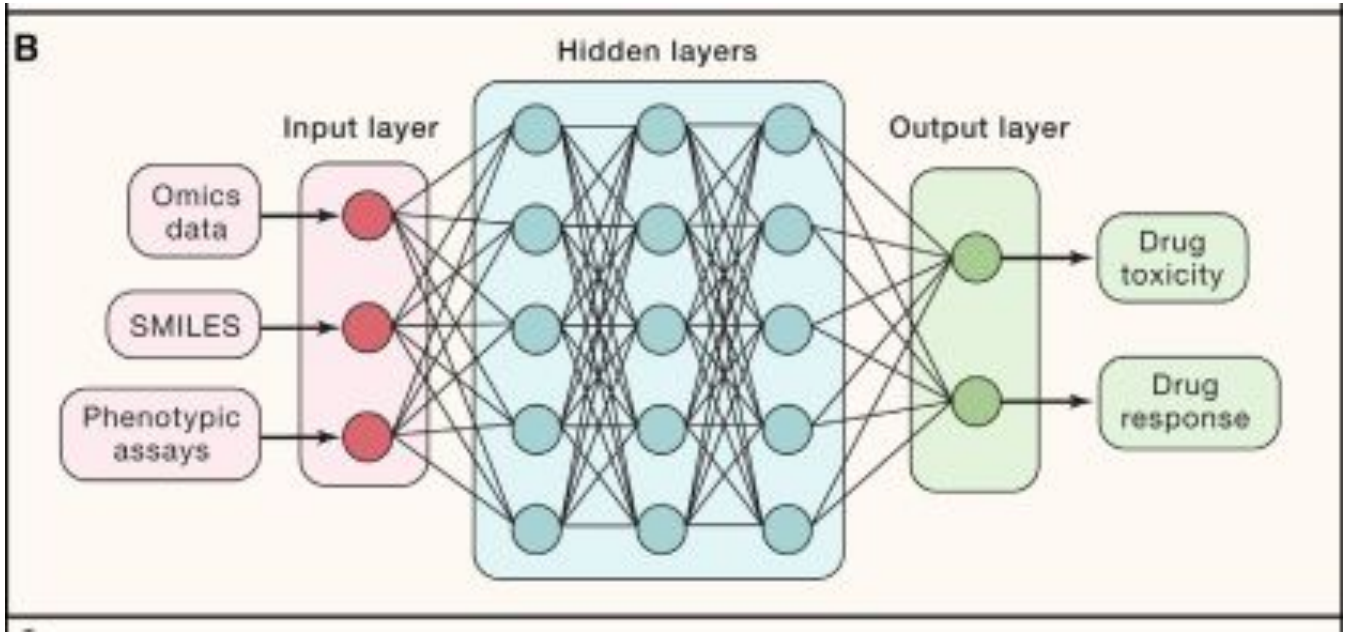
Dialogue for Reverse Engineering Assessment and Methodology (DREAM)

- DREAM challenges are an open-data, crowd-sourcing effort to find solutions to big-data research questions in network biology and medicine.
- The inference of genome-scale gene regulatory networks.
- The prediction of drug sensitivities and synergies using multi-omic datasets.
- Each DREAM challenge presents the network biology research community with a specific question and the necessary data to address it.
- An evaluation dataset that is hidden from all participants.

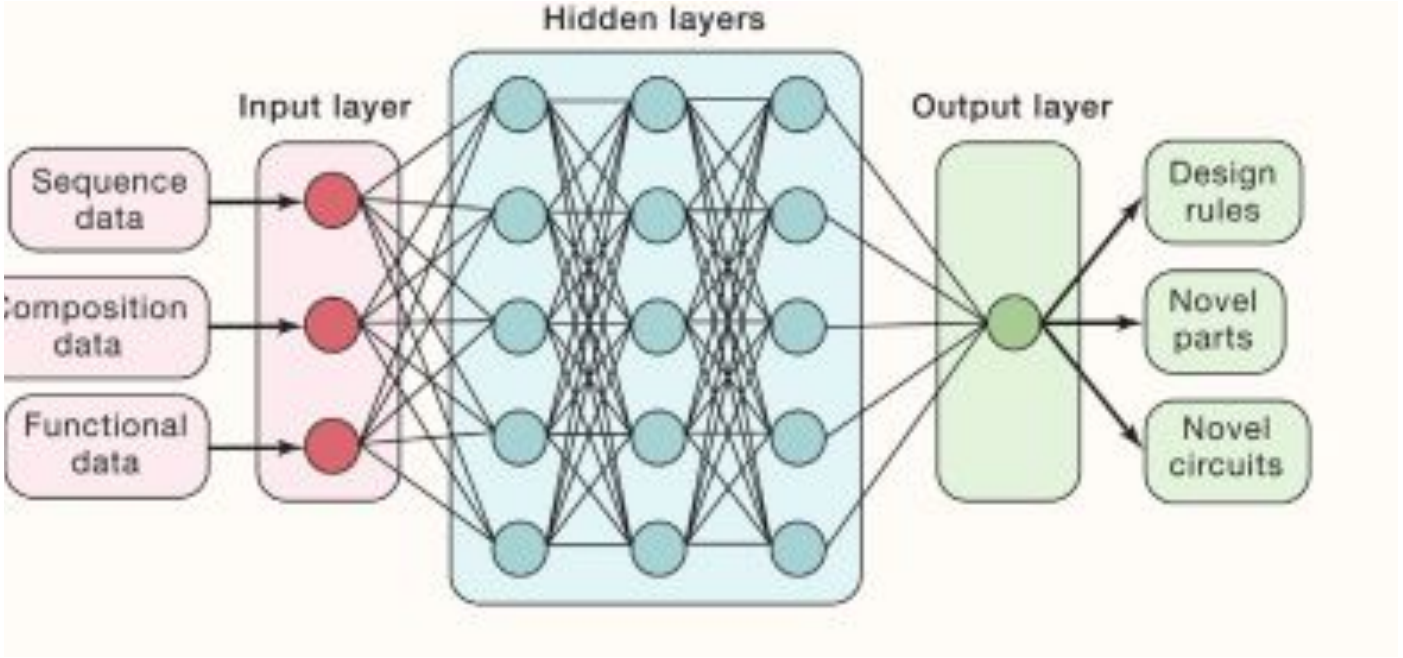
Rules of thumb for building models

- Simple is often better. E.g. linear regression-based models (e.g., elastic nets).
- Prior knowledge improves performance.
- Ensemble models produce robust results.

Application of Deep Learning in Drug Discovery



Application of Deep Learning in Synthetic Biology



Conclusions

- Machine Learning plays an important role in answering biological questions.
- Both Machine Learning and Deep Learning find immense application in the field.
- Appropriate model must be chosen to maximize the performance of the model.
- Deep Learning finds applications in various biological fields like Drug Discovery, Synthetic Biology, etc. and many more.

Thank You