

Introduction to Machine Learning

(material adapted from previous courses)

What is Machine Learning?



What do you see?

We need to classify the pictures with a dog and the ones with the muffin.



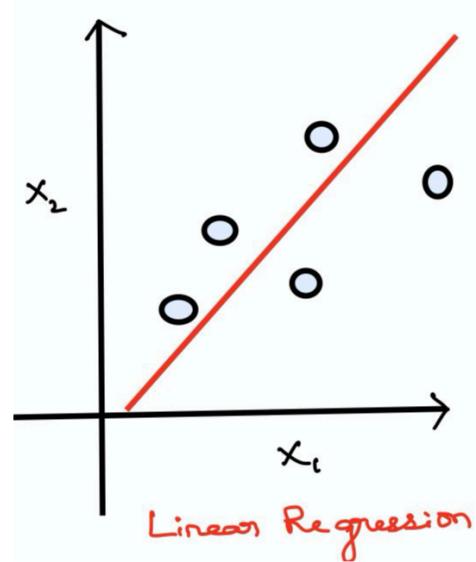
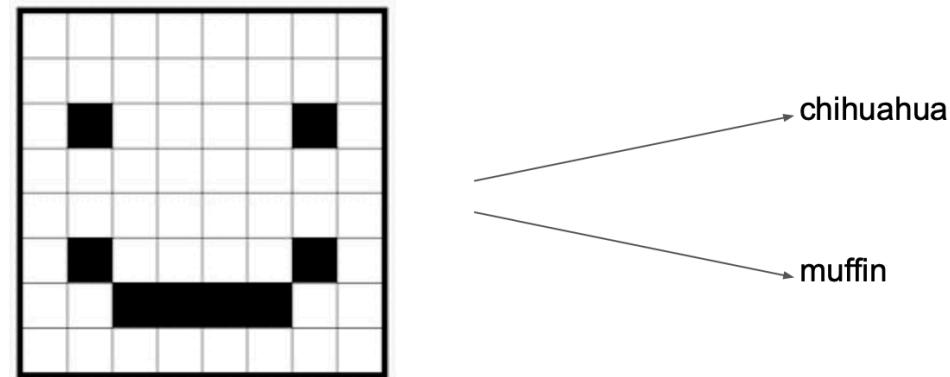
Or do we need to partition the images and manually annotate some of them?

What is Machine Learning?

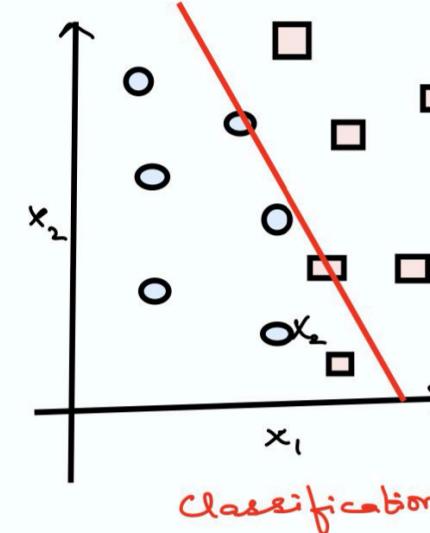


New York Times reported in 1958 that the invention was the beginning of a computer that would “be able to walk, talk, see, write, reproduce itself and be conscious of its existence”

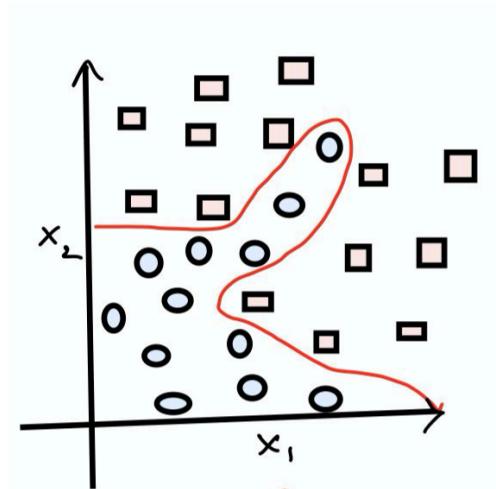
What is Machine Learning? Behind the hood



Linear Regression



Classification



Neural Network

What is Machine Learning?

Machine Learning (ML) is generating abstract hypotheses (models) based on data which can be later used predict results on new data.

Types of ML:

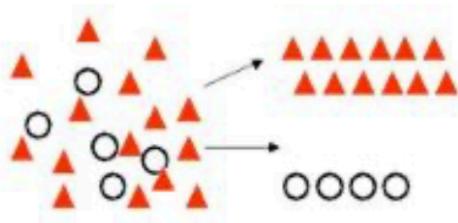
- [1] Supervised
- [2] Unsupervised
- [3] Semi-supervised
- [4] Reinforcement learning



Common models used in Machine Learning

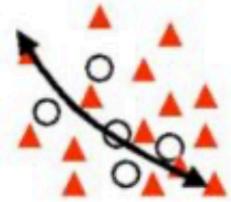
Techniques

Classification



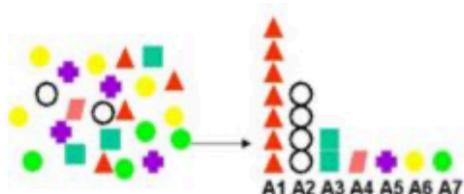
Separation of data point into classes.
Typical example: yes/no answers

Regression



Continuous numerical outcome.
Prediction of future values based on observed ones

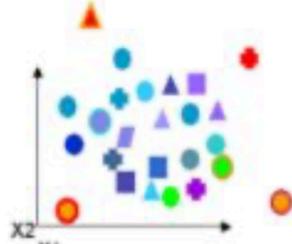
Attribute Ranking



Identification of important/ discriminative
features based on their relationship with the
target attribute.

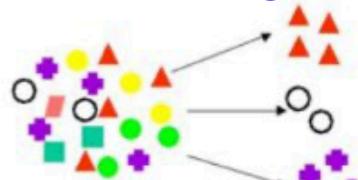
Common models used in Machine Learning

Anomaly detection



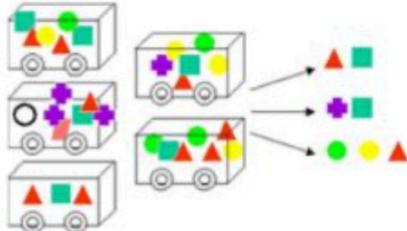
Identification of outliers

Clustering



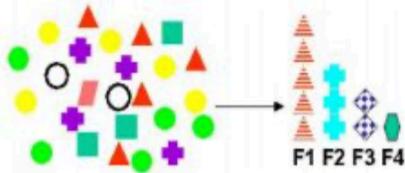
Finding groups/ structure in the data

Association



Finding rules associated with naturally co-occurring terms

Feature selection



Selection of best features to describe a target attribute

Different ML models

[1] Supervised learning

Nearest Neighbors

Support Vector Machines (SVMs)

Decision Trees and Random Forest(s)

Linear and Non-linear regression

Artificial Neural Networks

Deep Learning

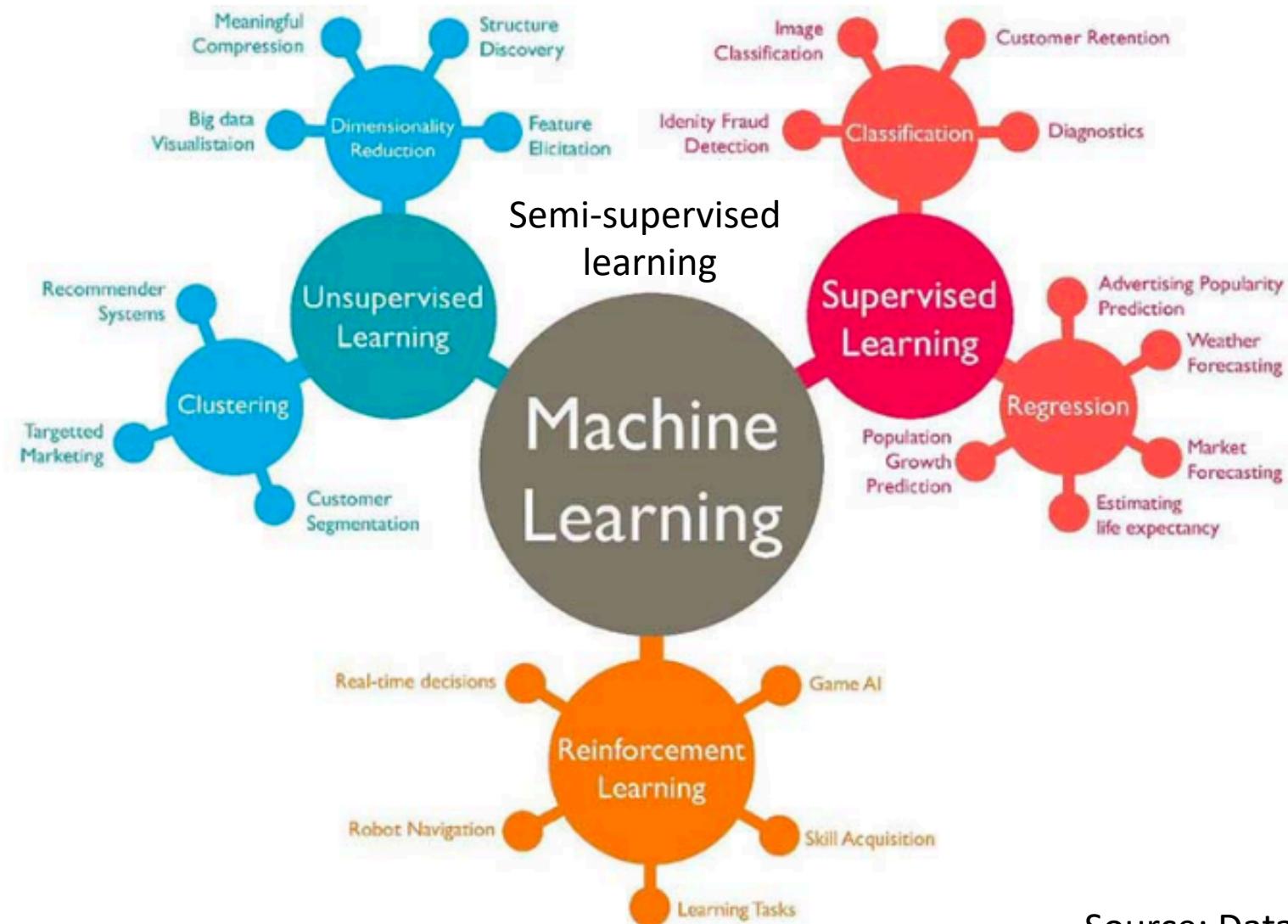
[2] Unsupervised Learning

Dimensionality reduction

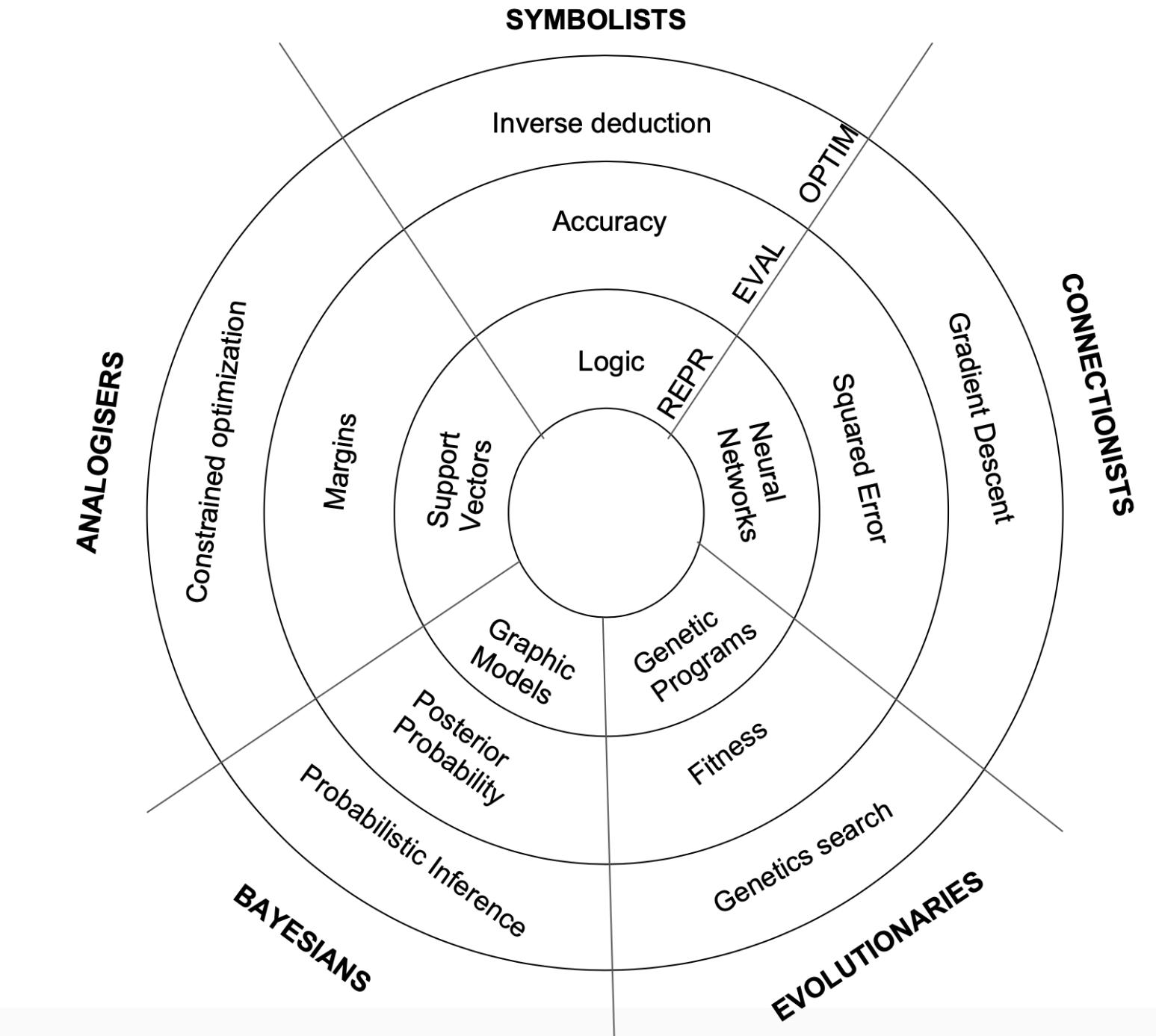
Clustering

Self Organizing Maps (Kohonen Maps)

Types of Machine Learning Algorithms



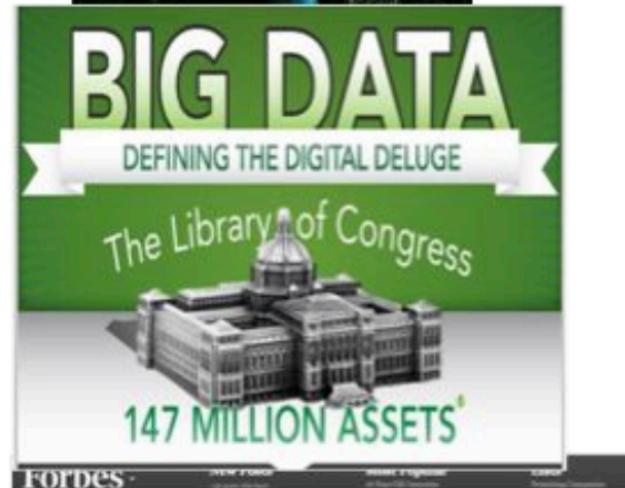
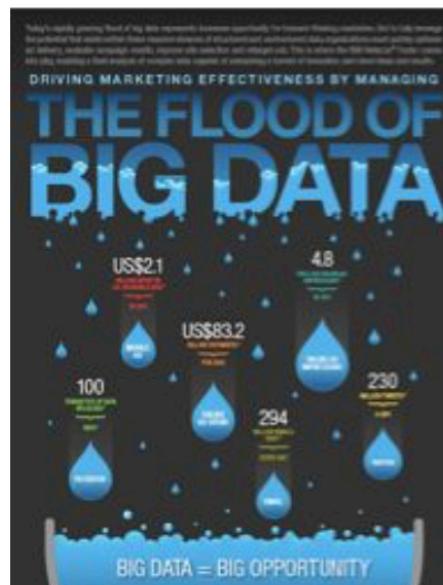
Source: Data Science Central



IBM estimates that 90% of world's data has been created in the last two years



Commercial
World Data:
Financial &
Retail Data



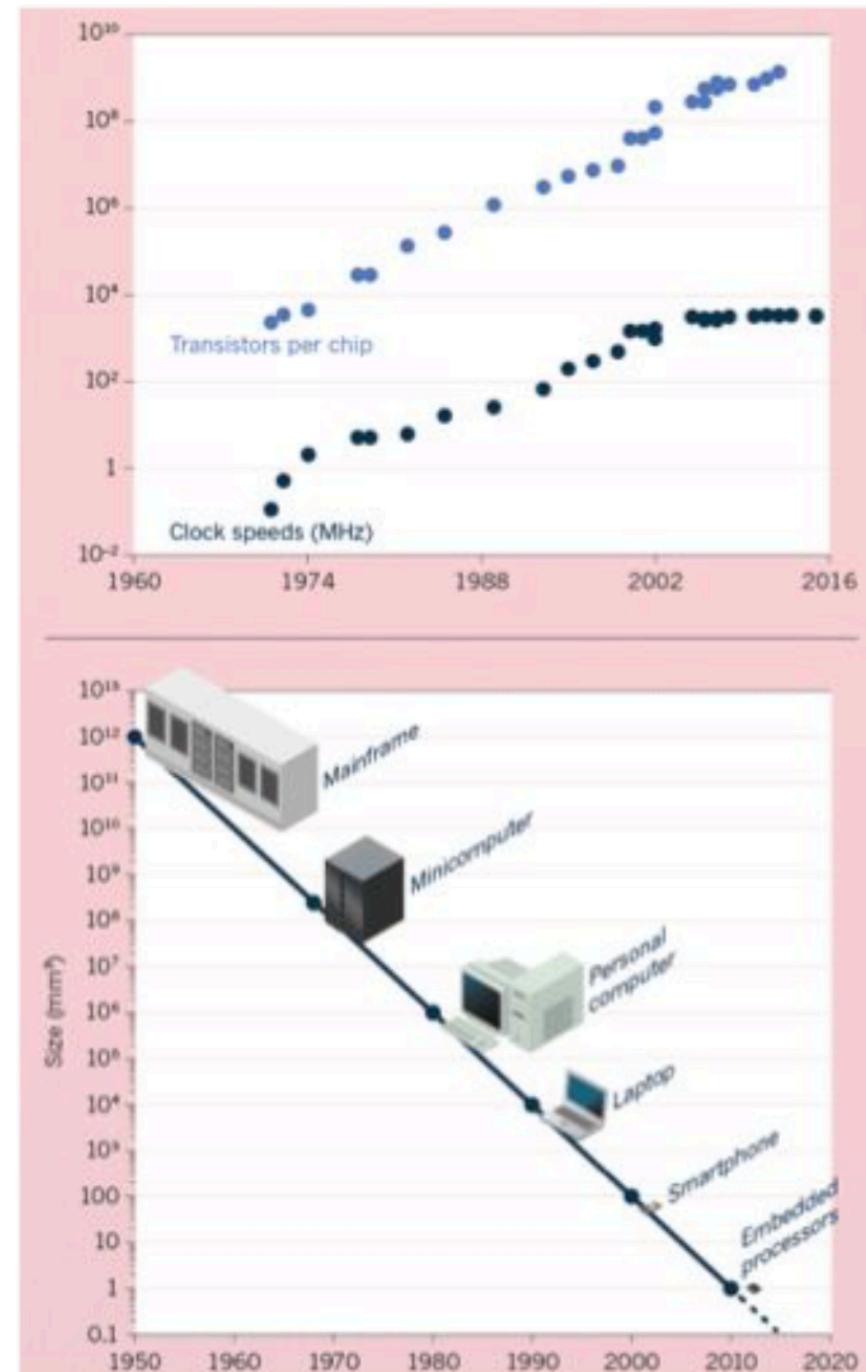
*

Human brain	2.5 PB
Spotify	10 PB
Ebay	90 PB
Facebook	300 PB
Google	15000 PB

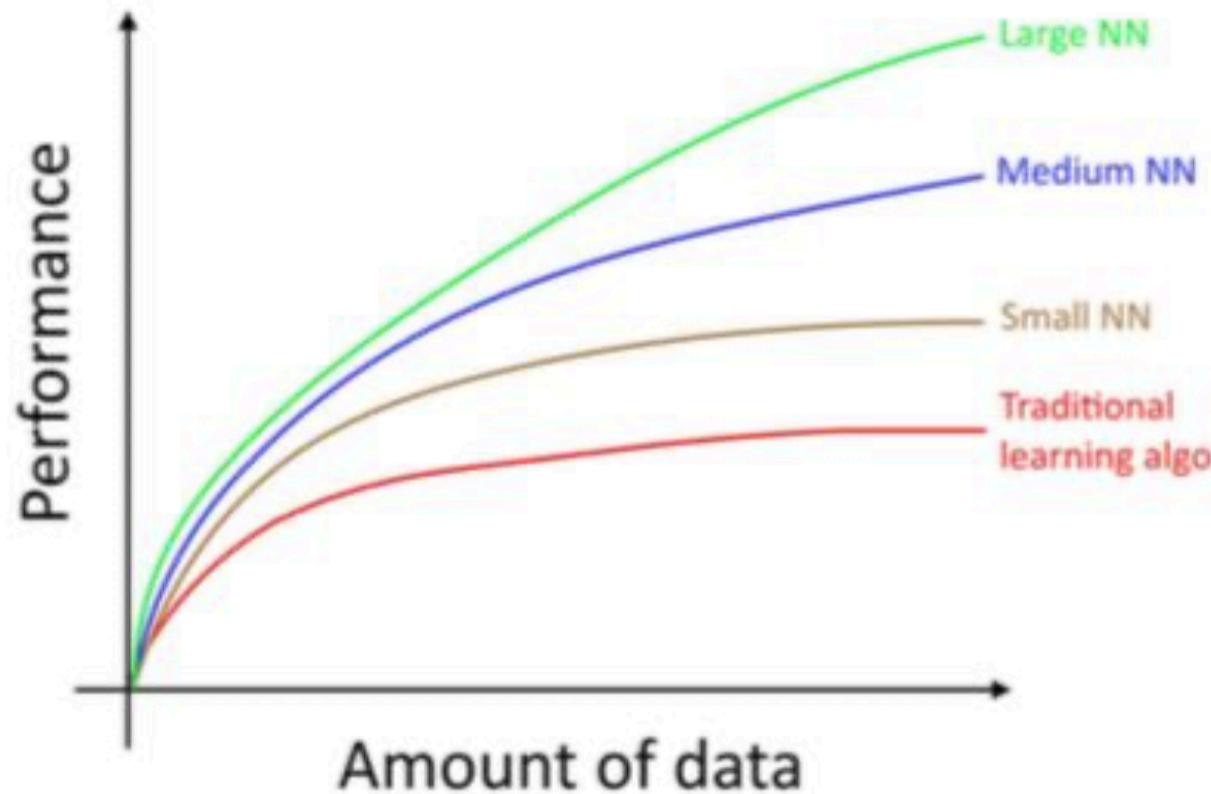
*Royal Society website

Moore's Law: Exponential Scaling of Computer Technology

- Exponential increase in the number of transistors per chip.
- Led to improvements in speed and miniaturization.
- Drove widespread adoption and novel applications of computer technology.



More data ... more detailed results?

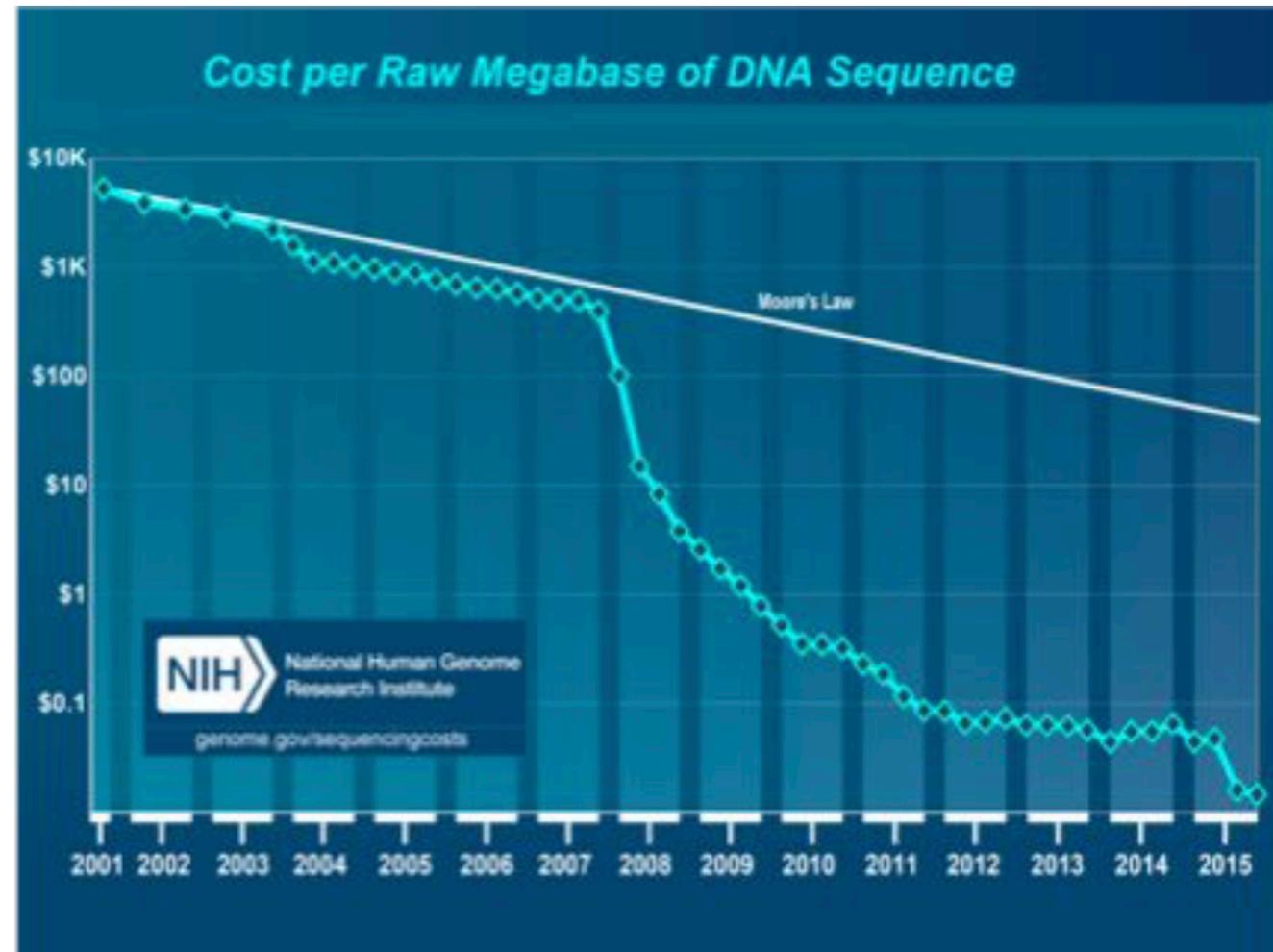


Source: Andrew Ng's book

Machine Learning – biological examples

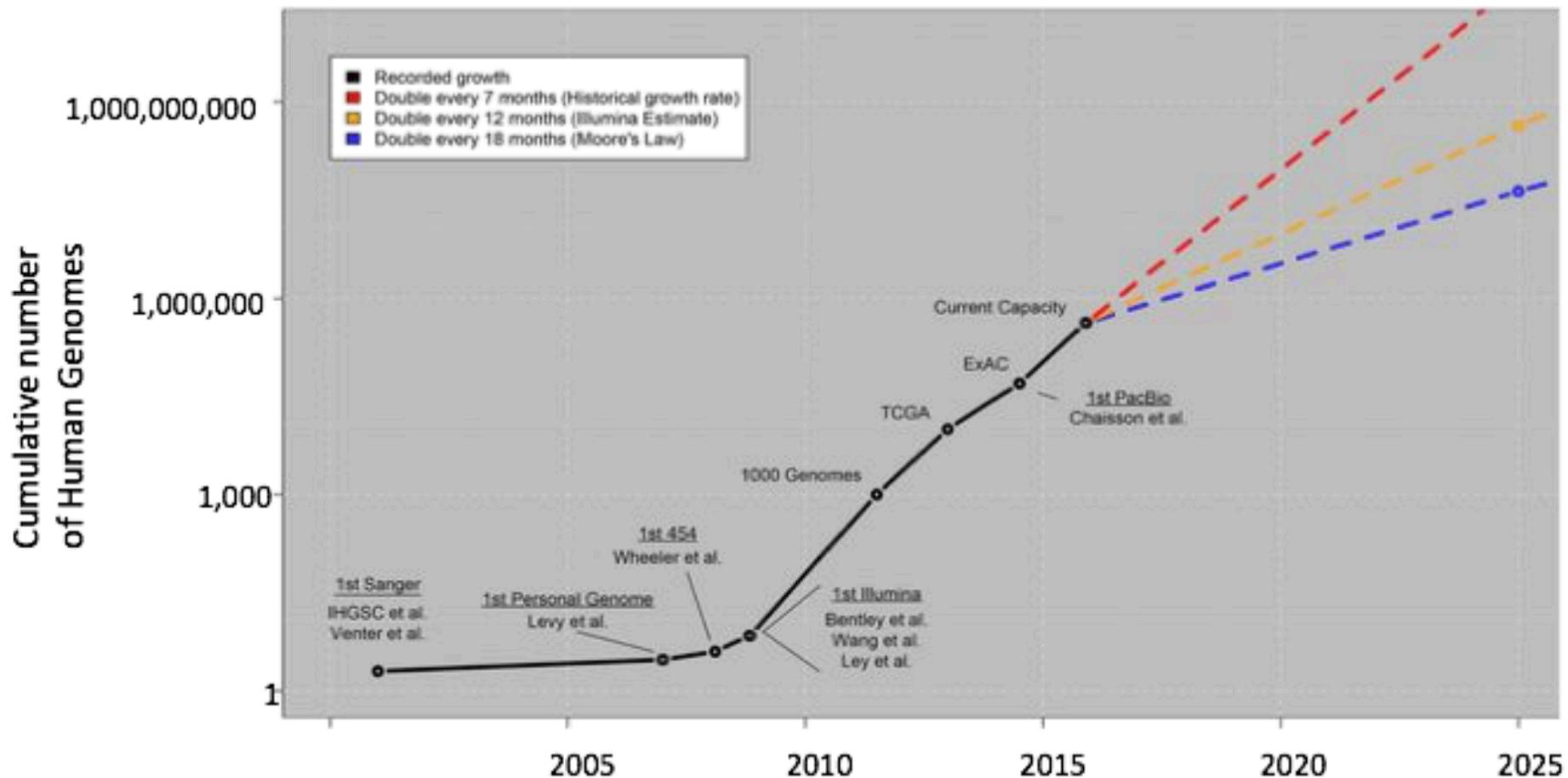
DNA sequencing has gone through technological S-curves

- In the early 2000's, improvements in Sanger sequencing produced a scaling pattern similar to Moore's law.
- The advent of NGS was a shift to a new technology with dramatic decrease in cost).



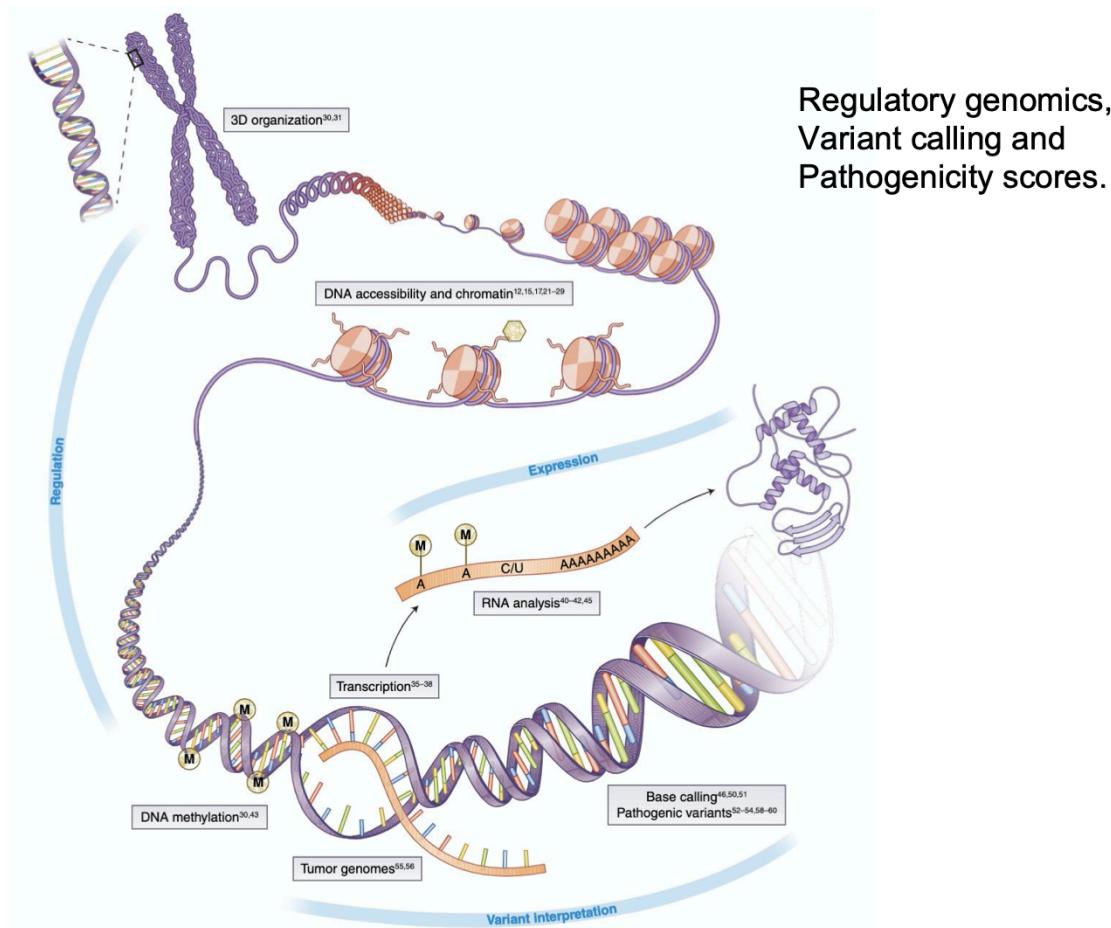
Moore's law

Available data. Where are we heading?

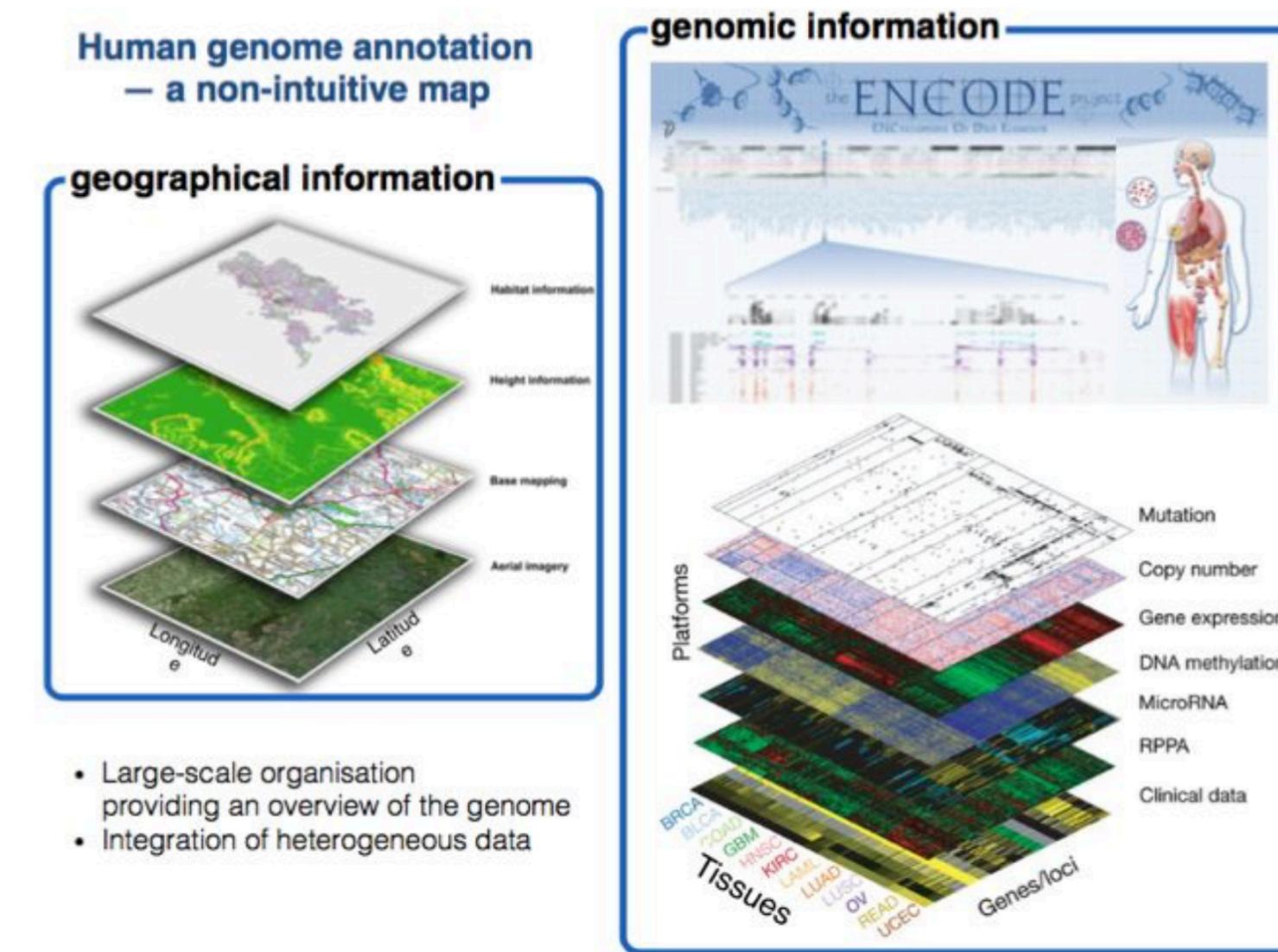


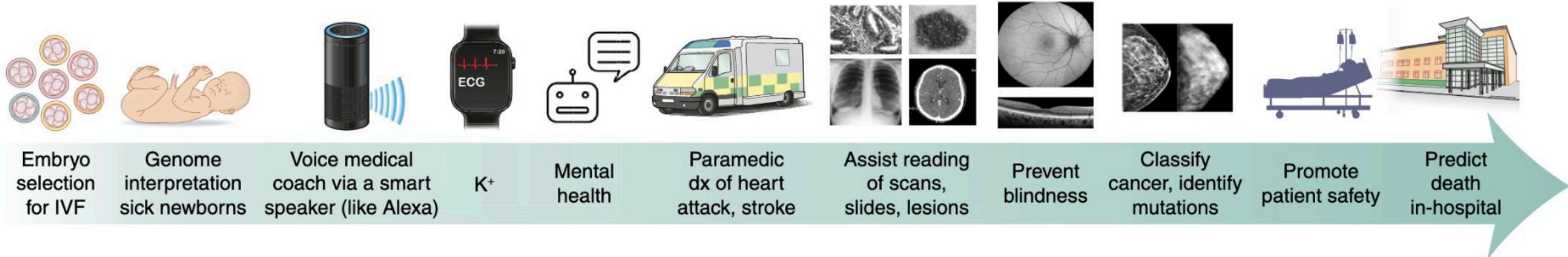
By 2019 total NGS data to exceed 2 Exabytes

Applications of ML in genomics

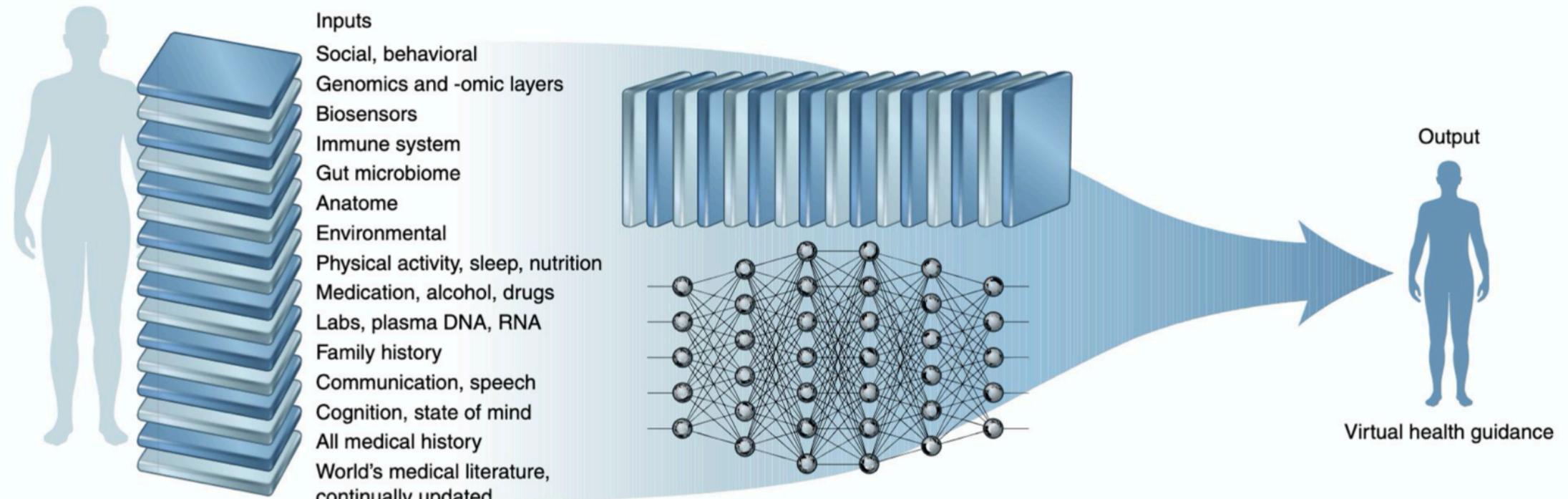


Applications of ML in genomics. More data



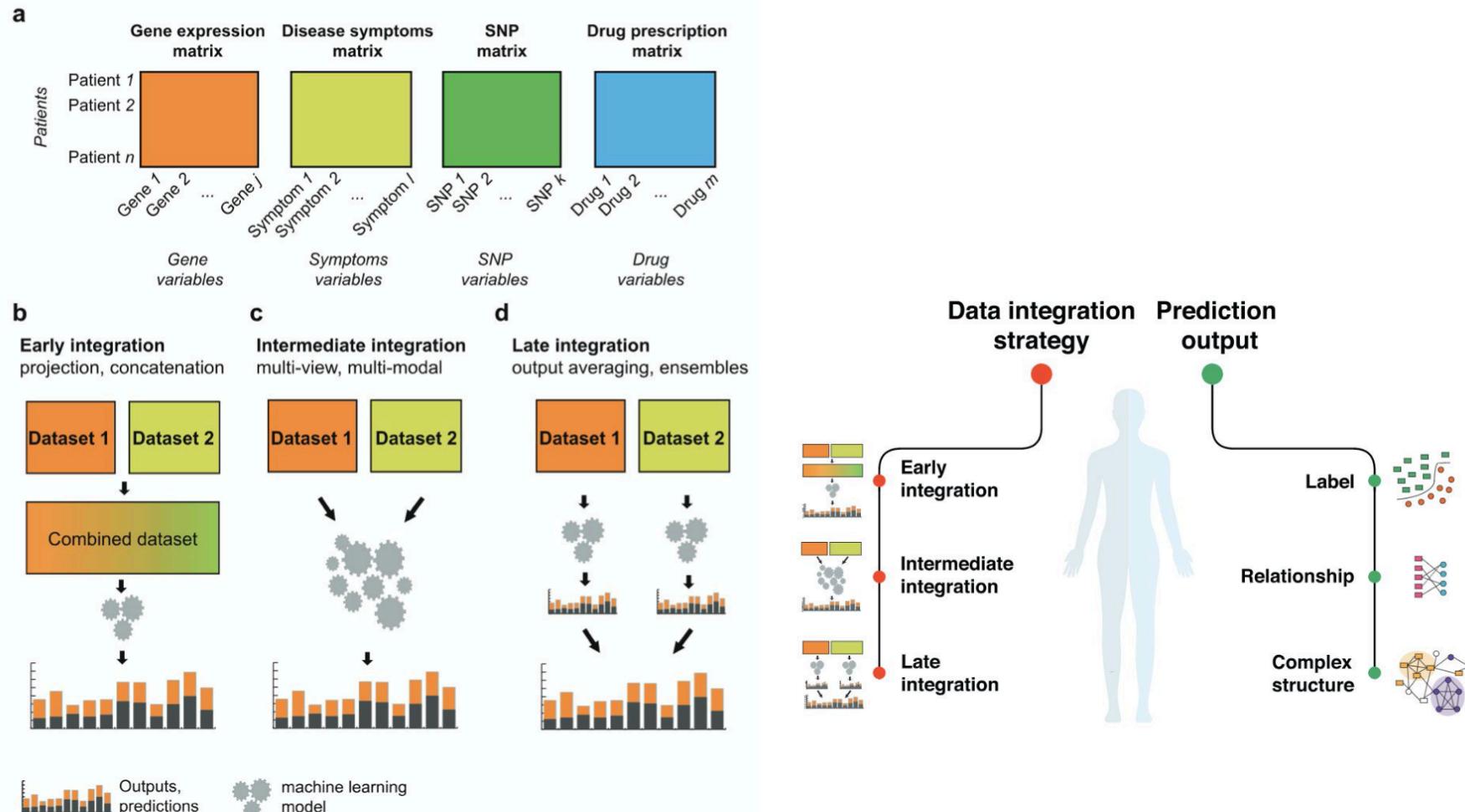


Multi-modal data inputs & algorithms to provide individualized guidance

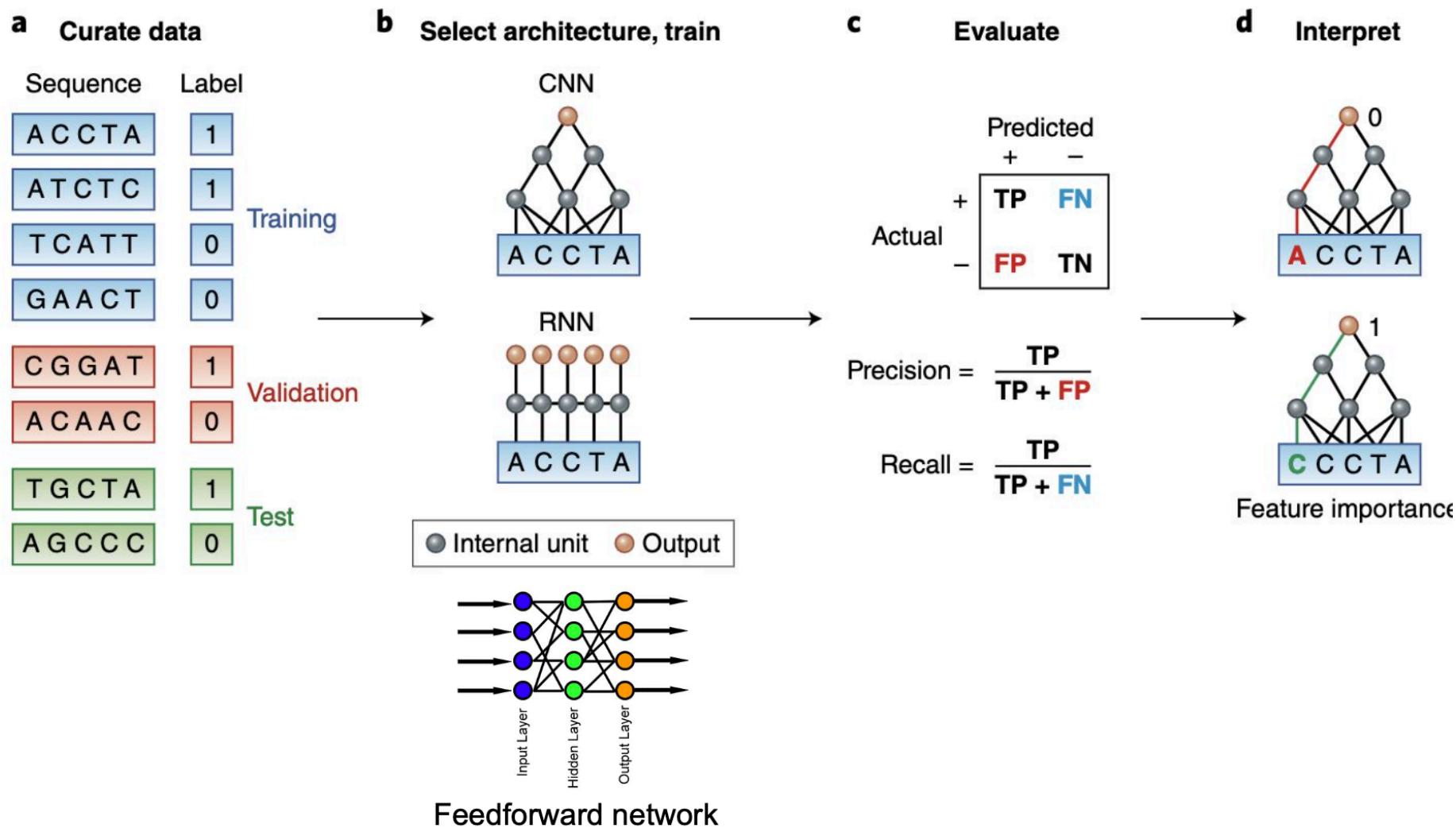


Data integration. Can ML help?

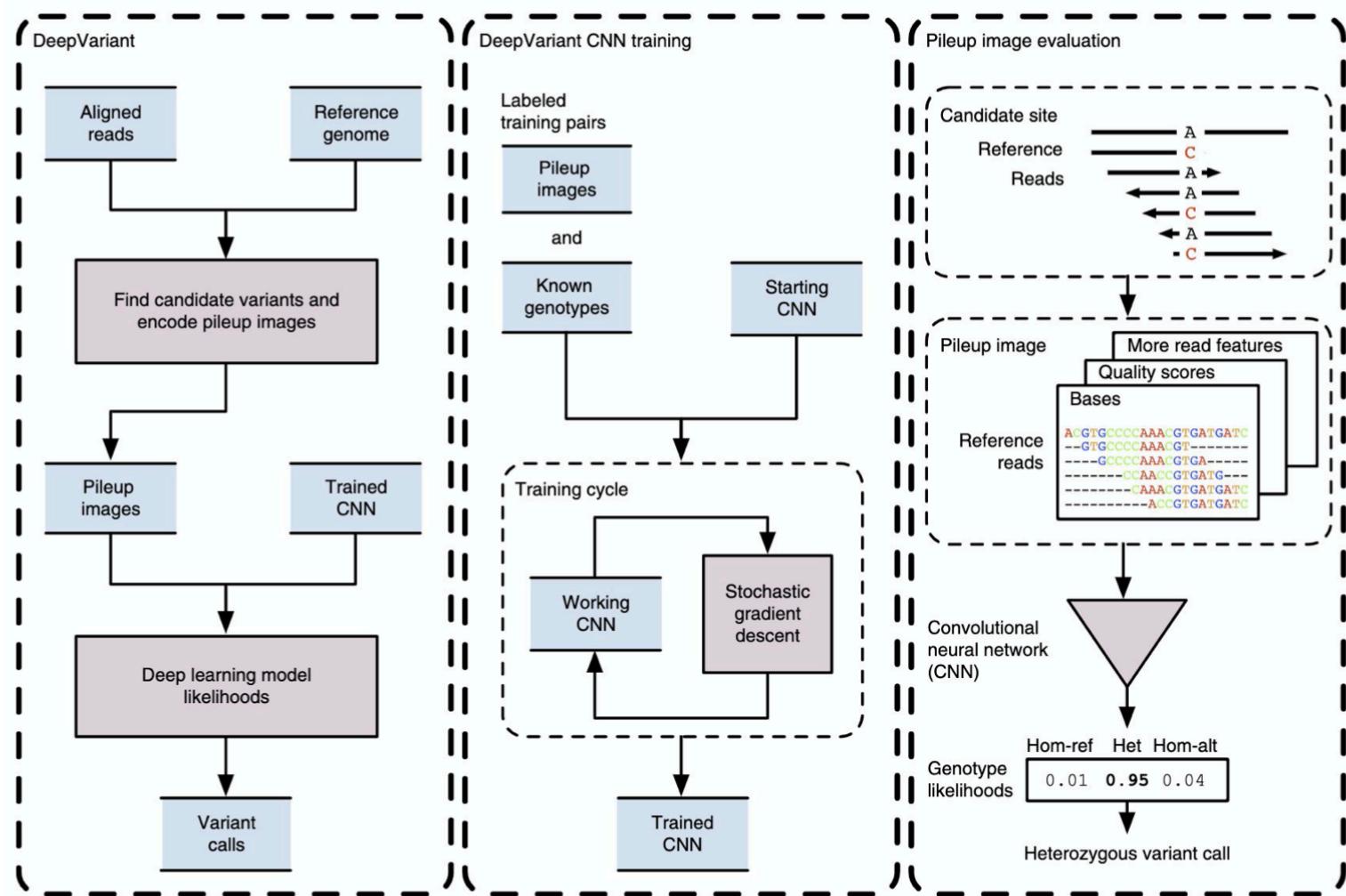
high-dimensional, incomplete, biased, heterogeneous, dynamic, and noisy.



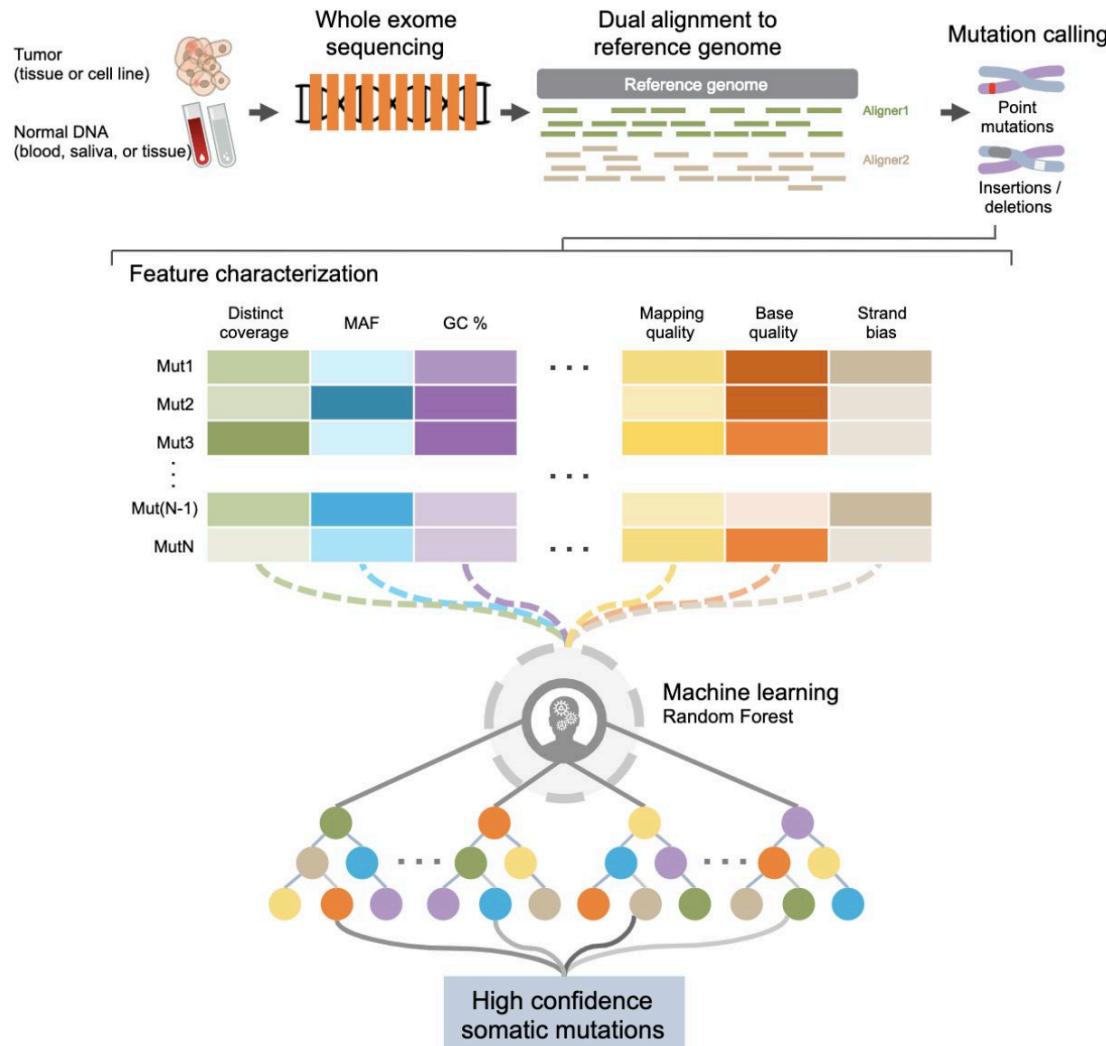
Steps for a ML-based pipeline



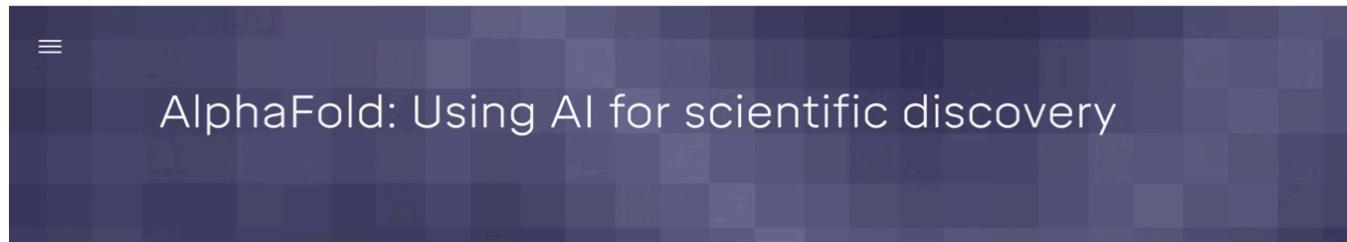
Example 1. Variant caller using NNs



Example 2. Somatic mutation discovery using RFs



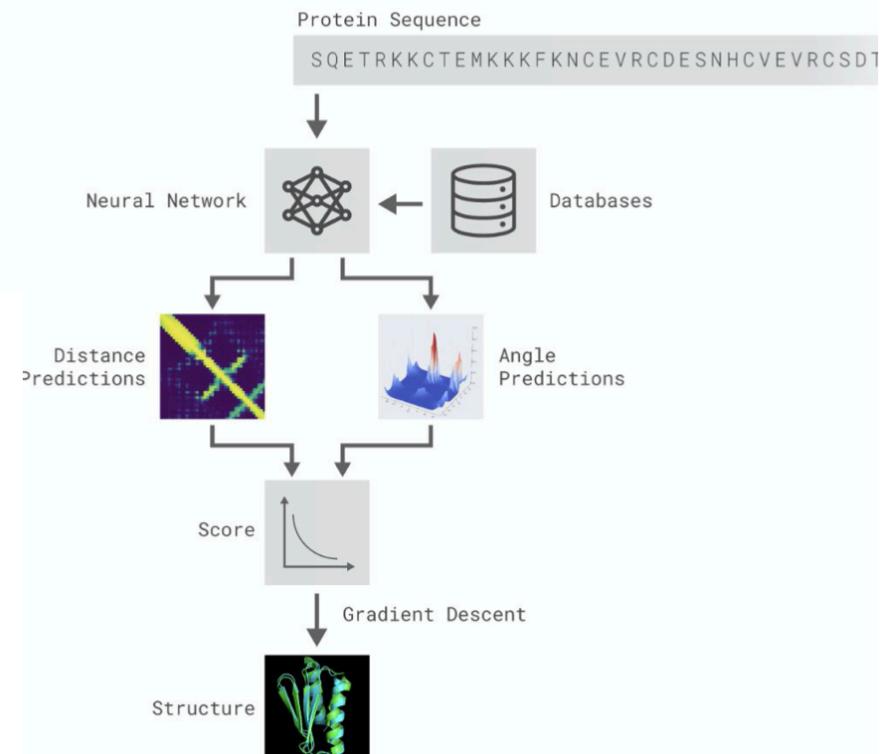
Example 3. DeepMind AlphaFold



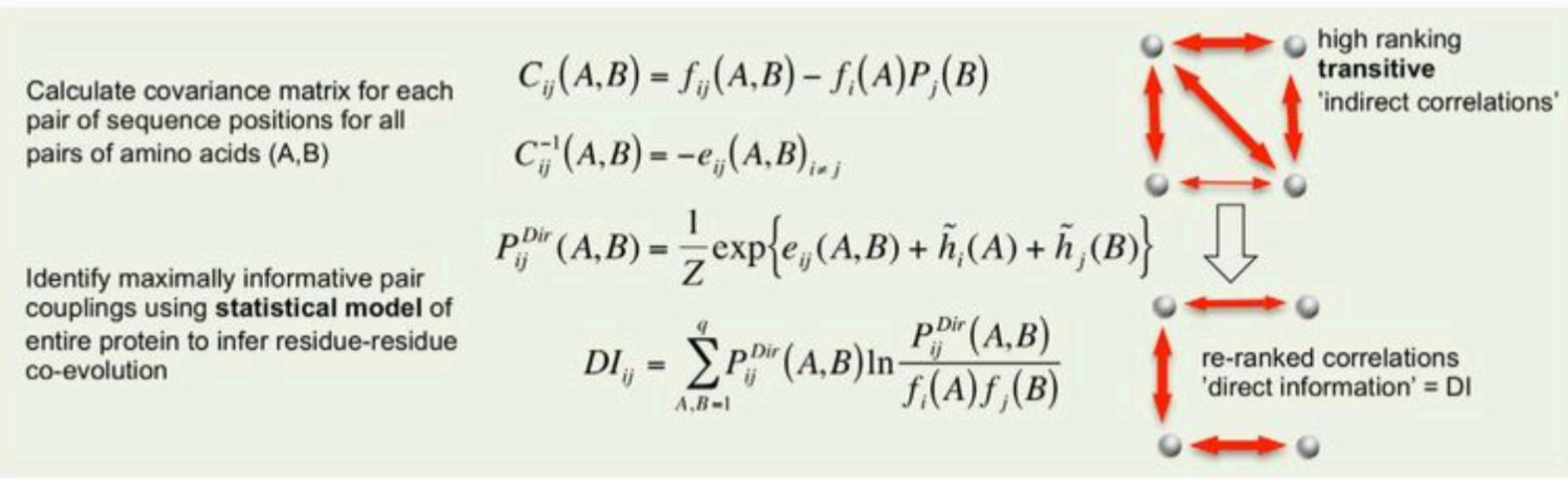
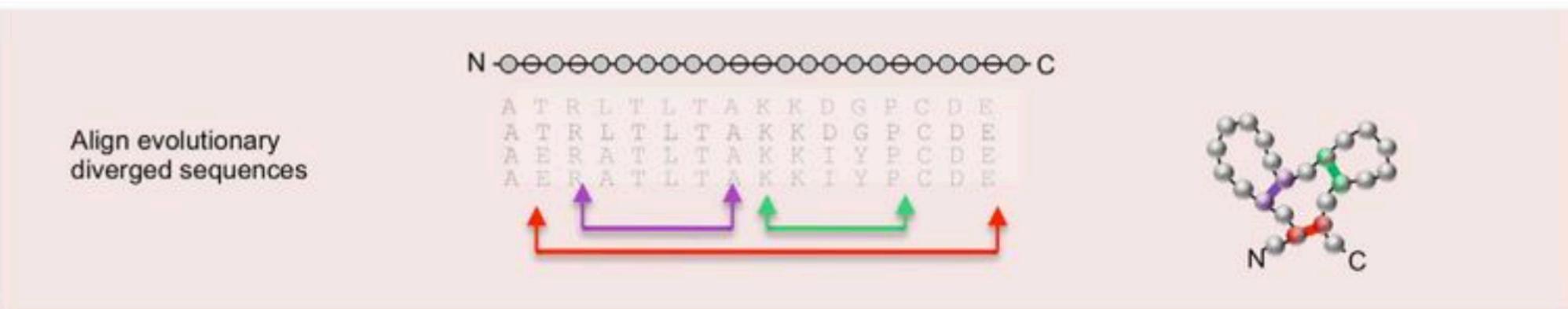
Today we're excited to share DeepMind's first significant milestone in demonstrating how artificial intelligence research can drive and accelerate new scientific discoveries. With a strongly interdisciplinary approach to our work, DeepMind has brought together experts from the fields of structural biology, physics, and machine learning to apply cutting-edge techniques to predict the 3D structure of a protein based solely on its genetic sequence.

Our system, **AlphaFold**, which we have been working on for the past two years, builds on years of prior research in using vast genomic data to predict protein structure. The 3D models of proteins that AlphaFold generates are far more accurate than any that have come before—making significant progress on one of the core challenges in biology.

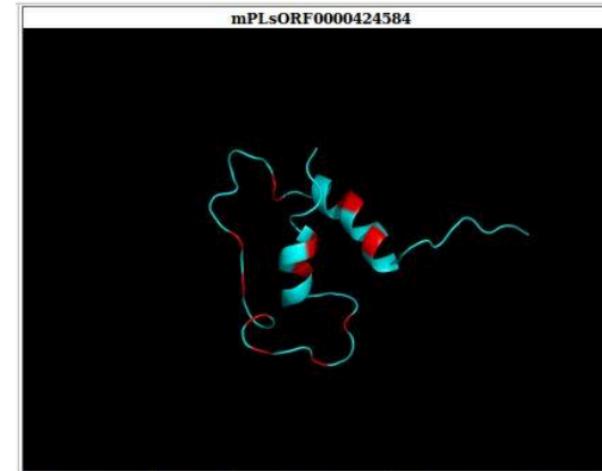
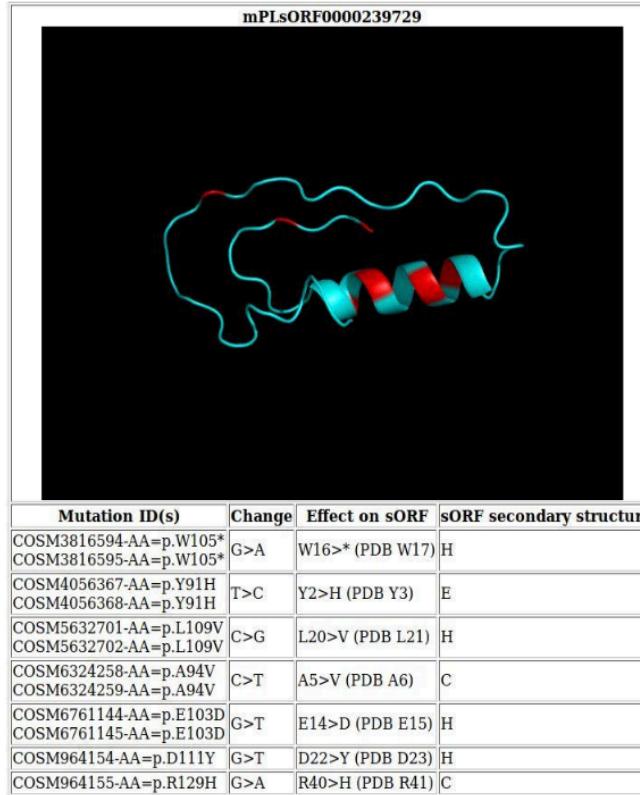
What is the protein folding problem?



Example 3. DeepMind AlphaFold



Example 3. DeepMind AlphaFold



Example 1. Chemistry

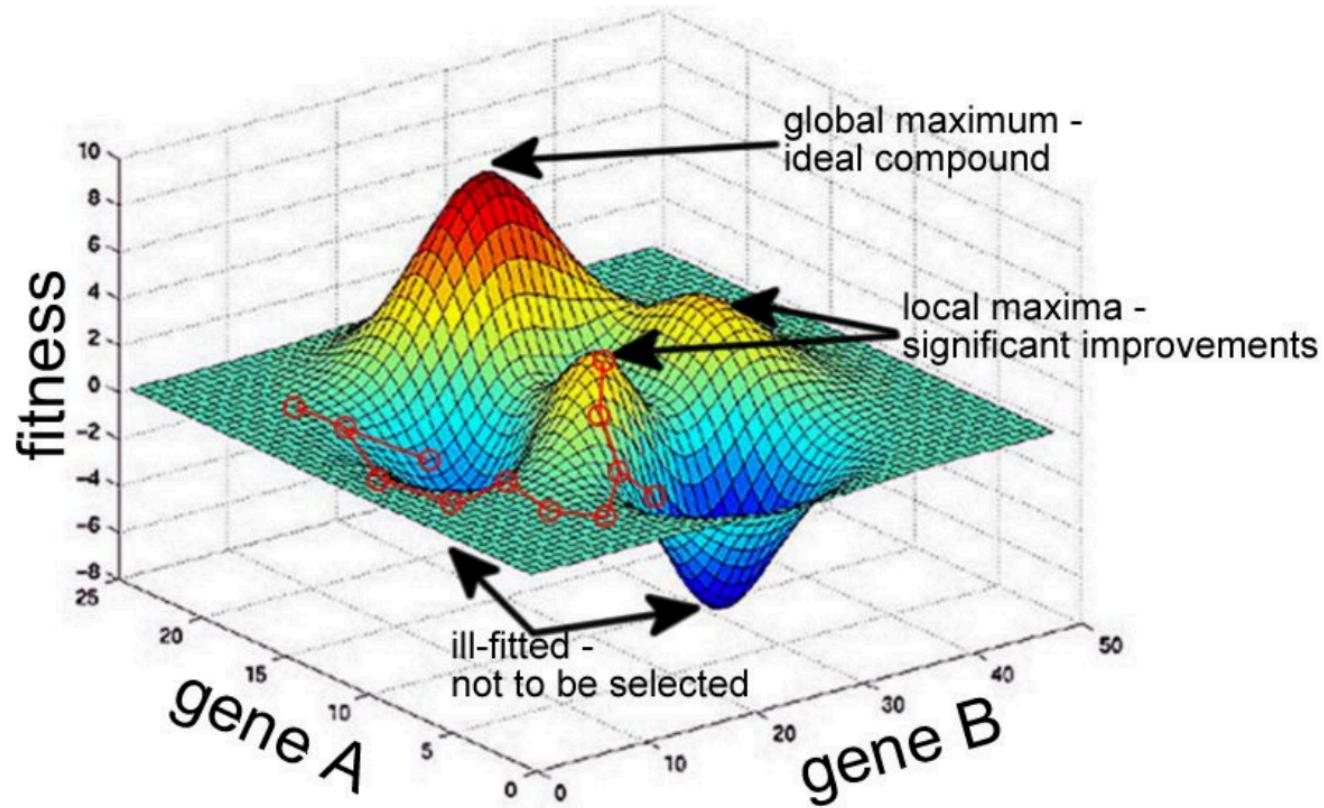


Figure 3 - A fitness landscape considering two hypothetical genes as, for example, block size and processing temperature of a polymer synthesis. Fitness could be hardness, for instance. The landscape contains local maxima, a global maximum, and a global minimum. In red, the path of a genetic algorithm along 11 iterations.

Example 2. Chemistry

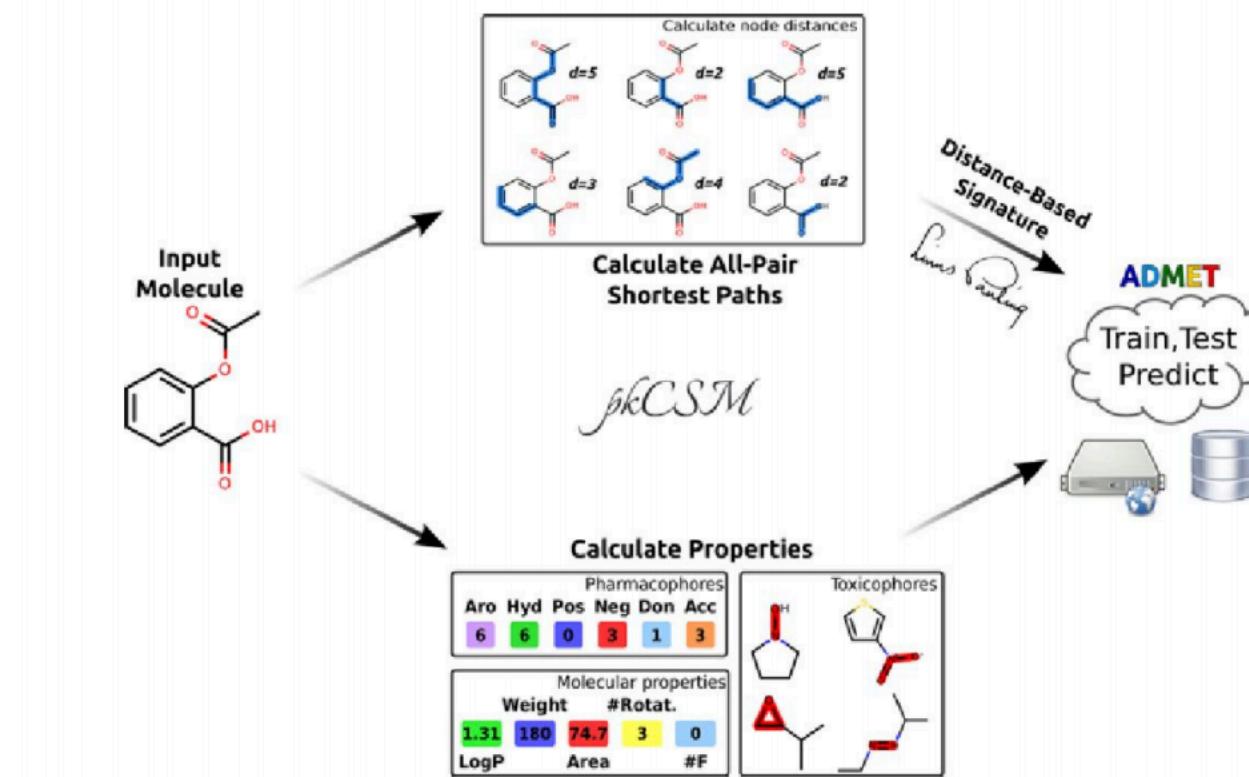


Figure 6 – The workflow of pkCSM is represented by the two main sources of information, namely the calculated molecular properties and shortest paths, for an input molecule. With these pieces of information, the ML system is trained to predict ADMET properties. Reproduced from ⁶⁴.

Example from Physics

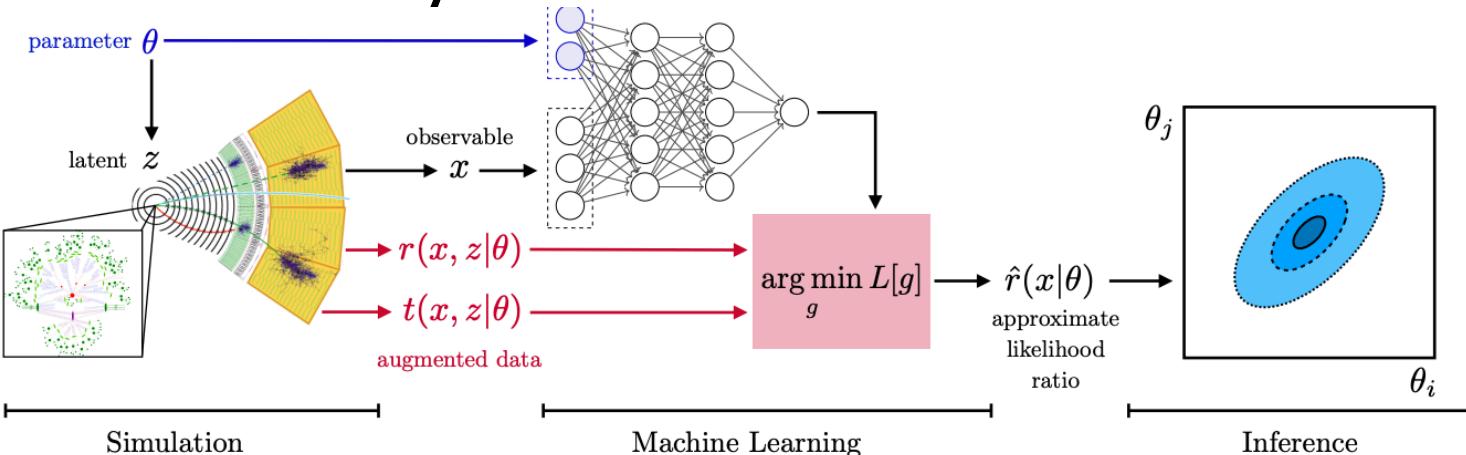


Figure 2 A schematic of machine learning based approaches to likelihood-free inference in which the simulation provides training data for a neural network that is subsequently used as a surrogate for the intractable likelihood during inference. Reproduced from (Brehmer *et al.*, 2018b).

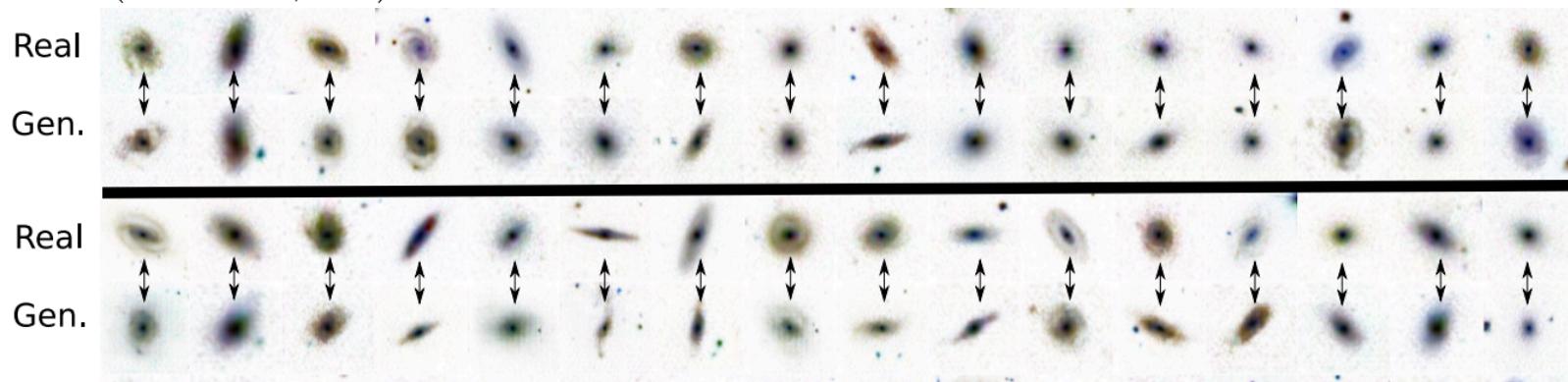
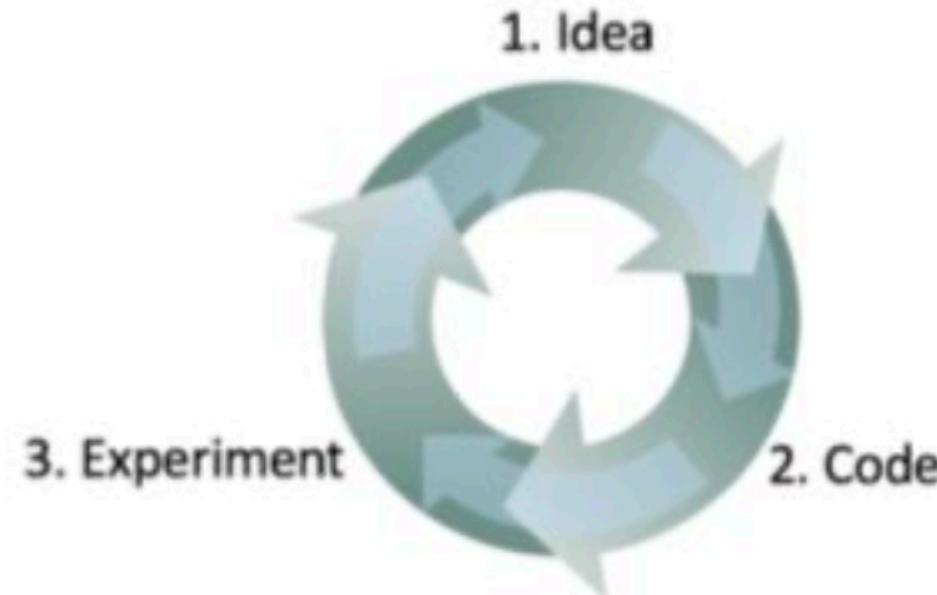


Figure 3 Samples from the GALAXY-ZOO dataset versus generated samples using conditional generative adversarial network. Each synthetic image is a 128×128 colored image (here inverted) produced by conditioning on a set of features $y \in [0, 1]^{37}$. The pair of observed and generated images in each column correspond to the same y value. Reproduced from (Ravanbakhsh *et al.*, 2016).

ML workflow

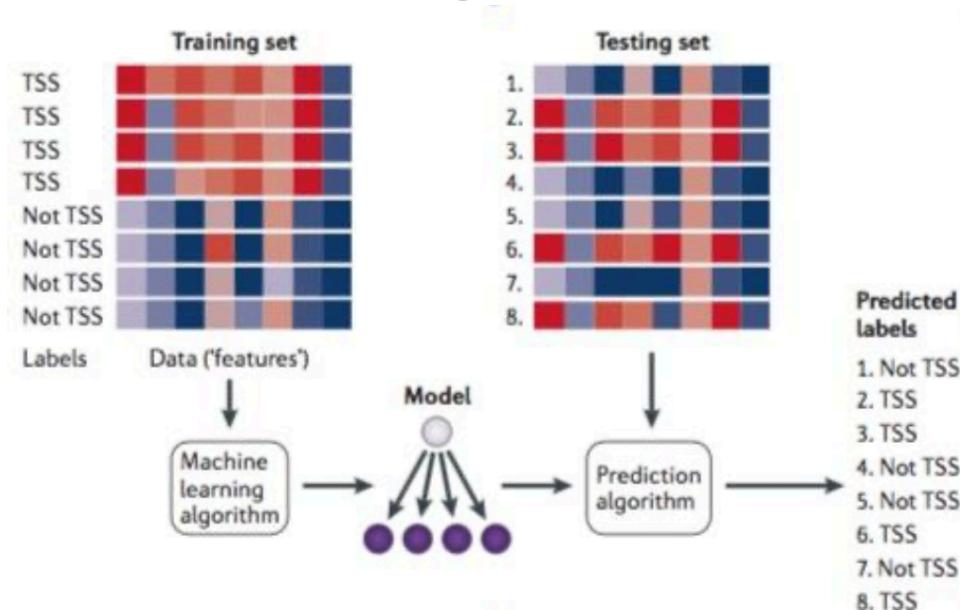


Dev/Test set & Metrics

Keep in mind:

1. The distributions should be identical
2. Overfitting dev set
3. The metric is measuring something else
4. Establish metric and dev/test set soon

ML (supervised learning) workflow



Step 1: develop an algorithm that will lead to successful learning.

Step 2: the algorithm is provided with a large collection of TSS sequences as well as, optionally, a list of sequences that are known not to be TSSs. The annotation indicating whether a sequence is a TSS is known as the label. The algorithm processes these labelled sequences and stores a model.

Step 3: new unlabelled sequences are given to the algorithm, and it uses the model to predict labels (in this case, 'TSS' or 'not TSS') for each sequence. If the learning was successful, then all or most of the predicted labels will be correct.

