Introduction to Machine Learning

(material adapted from previous courses)

What is Machine Learning?

What do you see?

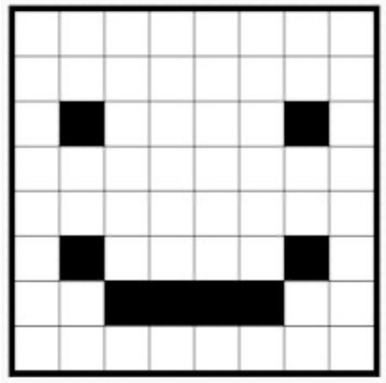
We need to classify the pictures with a dog and the ones with the muffin.

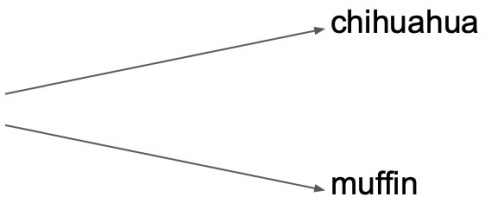
Or do we need to partition the images and manually annotate some of them?

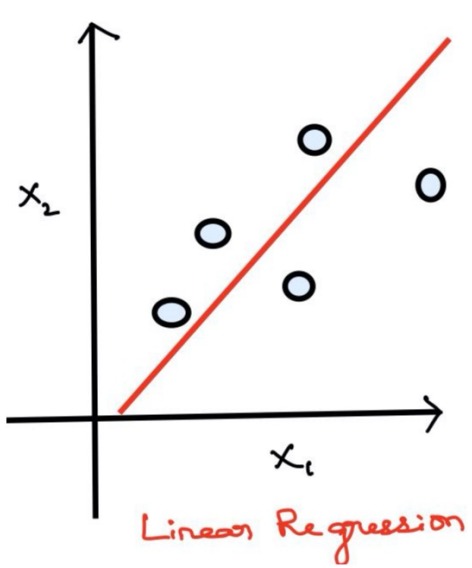
What is Machine Learning?

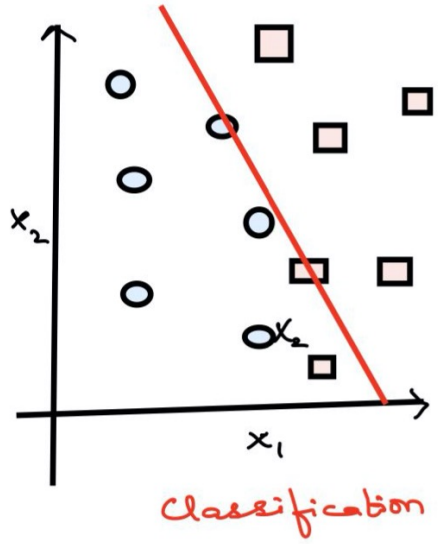
New York Times reported in 1958 that the invention was the beginning of a computer that would “be able to walk, talk, see, write, reproduce itself and be conscious of its existence”

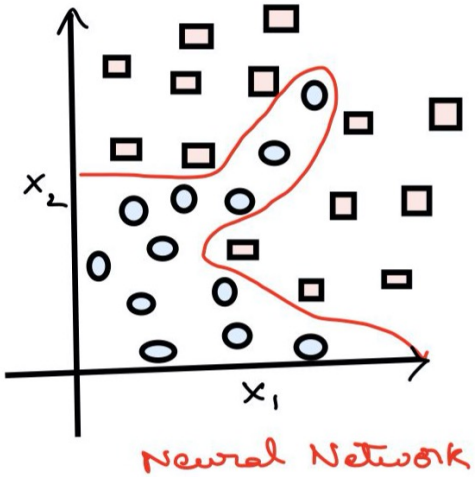
What is Machine Learning? Behind the hood











What is Machine Learning?

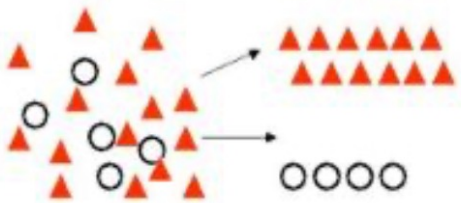
Machine Learning (ML) is generating abstract hypotheses (models) based on data which can be later used predict results on new data.

Types of ML:
[1] Supervised
[2] Unsupervised
[3] Semi-supervised
[4] Reinforcement learning

Common models used in Machine Learning

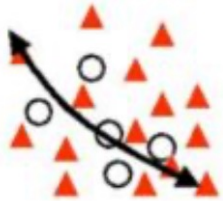
Techniques

Classification



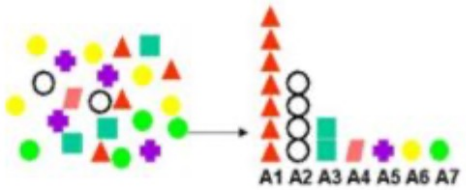
Separation of data point into classes.
Typical example: yes/no answers

Regression



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

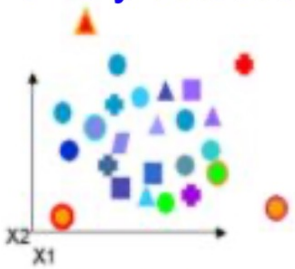
Attribute Ranking



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

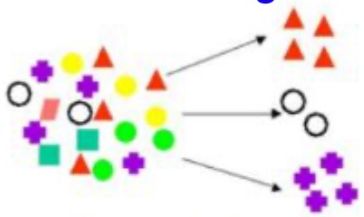
Common models used in Machine Learning

Anomaly detection



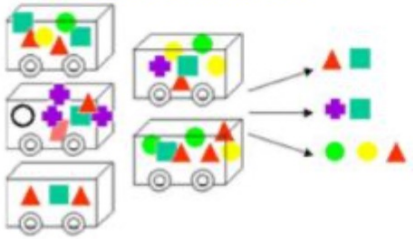
Identification of outliers

Clustering



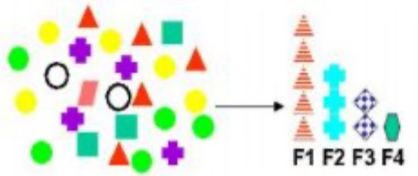
Finding groups/ structure in the data

Association



Finding rules associated with naturally co-occurring terms

Feature selection



Selection of best features to describe a target attribute

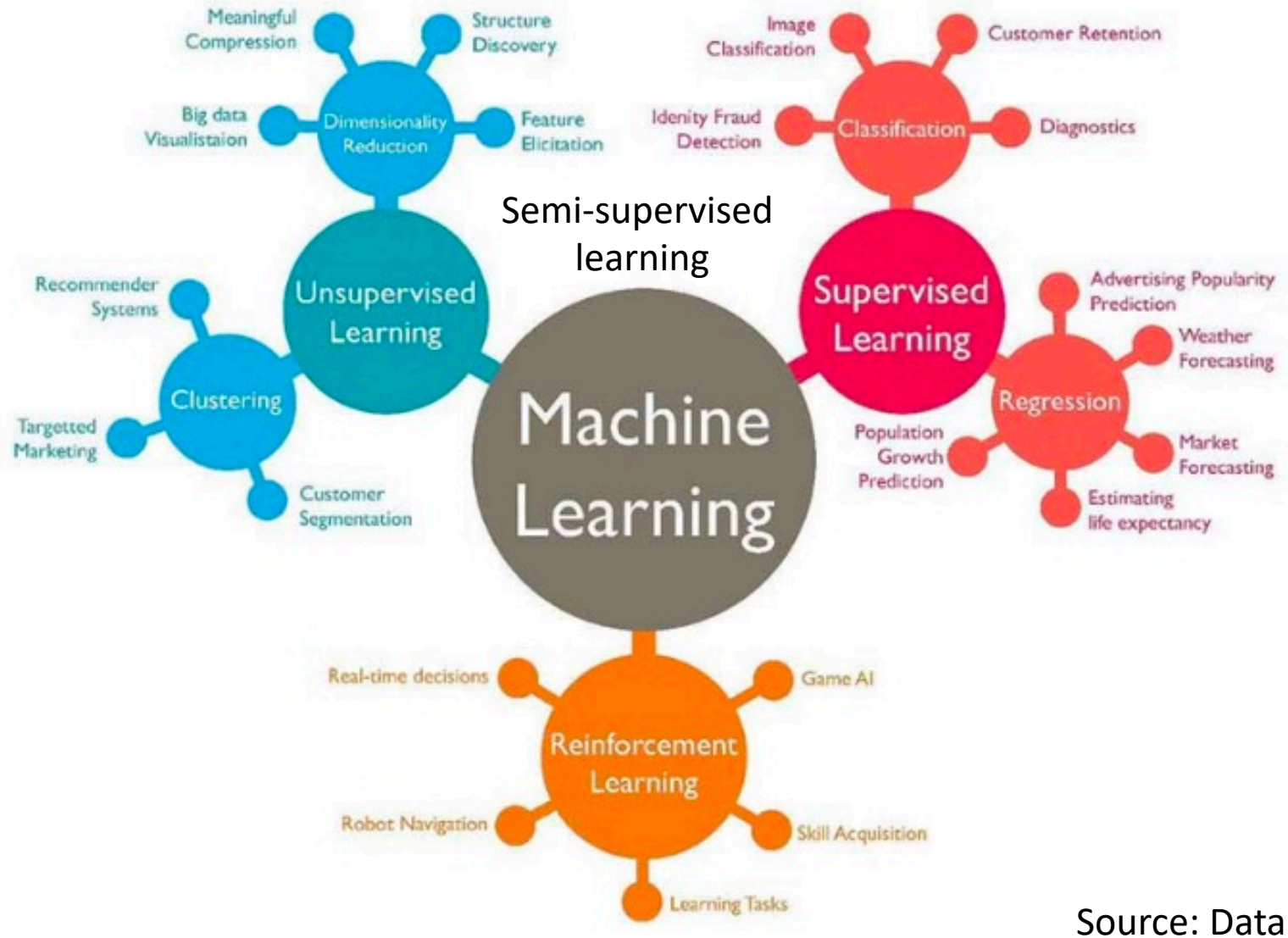
Different ML models

[1] Unsupervised Learning
- Dimensionality reduction
- Clustering
- Self Organizing Maps (Kohonnen Maps)

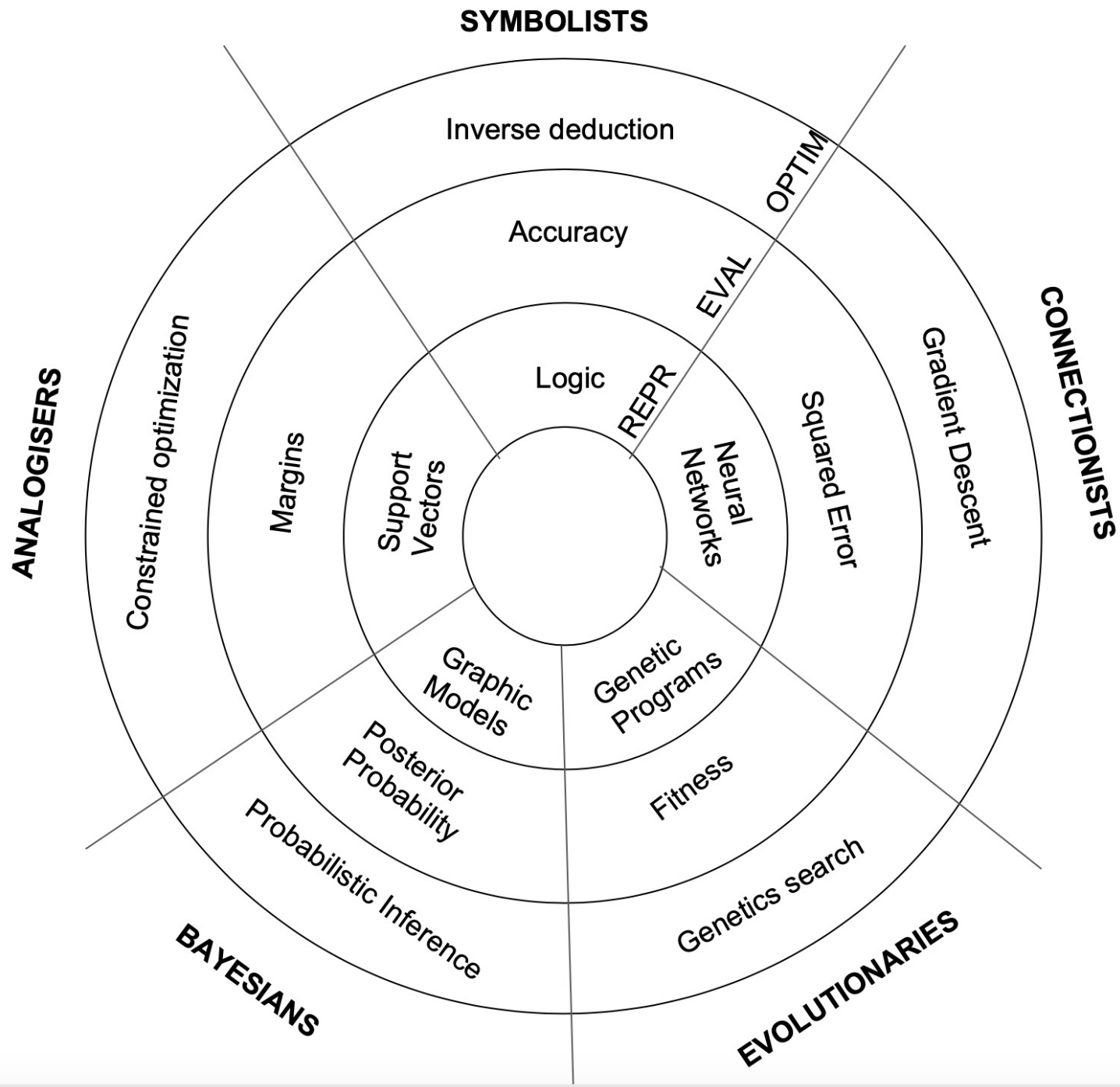
[2] Supervised learning
- Nearest Neighbors
- Support Vector Machines (SVMs)
- Decision Trees and Random Forest(s)

- Linear and Non-linear regression
- Artificial Neural Networks
- Deep Learning

Types of Machine Learning Algorithms

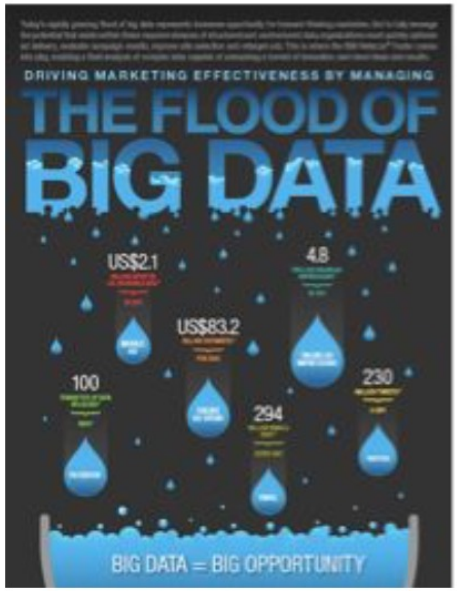


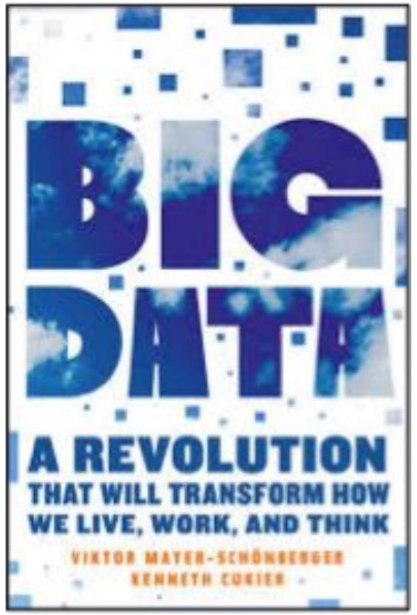
Source: Data Science Central


SYMBOLISTS
Inverse deduction
OPTIM
Accuracy
EVAL
CONNECTIONISTS
Logic
REPR
Gradient Descent
Squared Error
Neural Networks
Support Vectors
Margins
ANALOGISERS
Constrained optimization
Graphic Models
Genetic Programs
Posterior Probability
Fitness
Probabilistic Inference
Genetics search
BAYESIANS
EVOLUTIONARIES

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

Commercial World Data: Financial & Retail Data





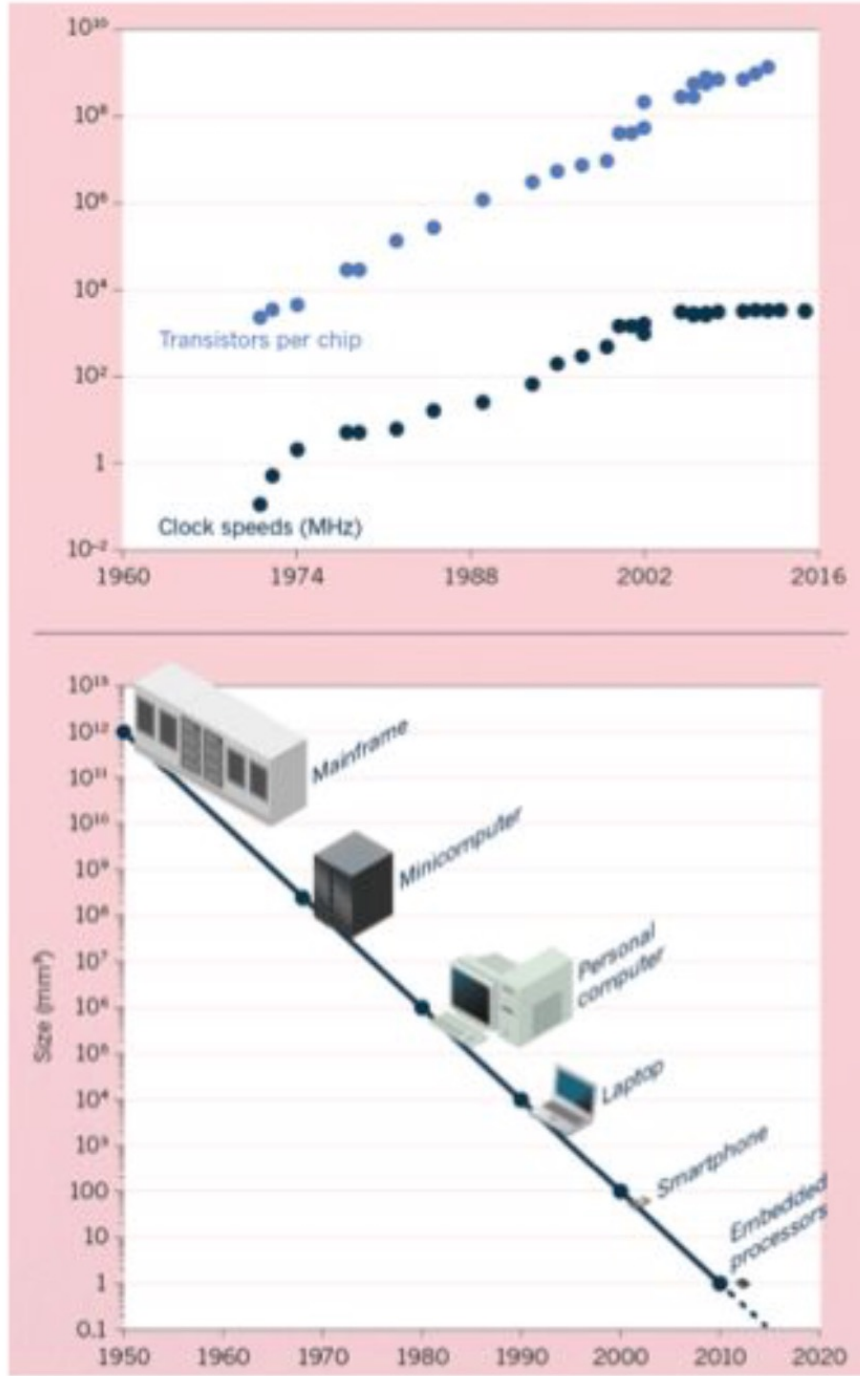
*

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

***Royal Society website**

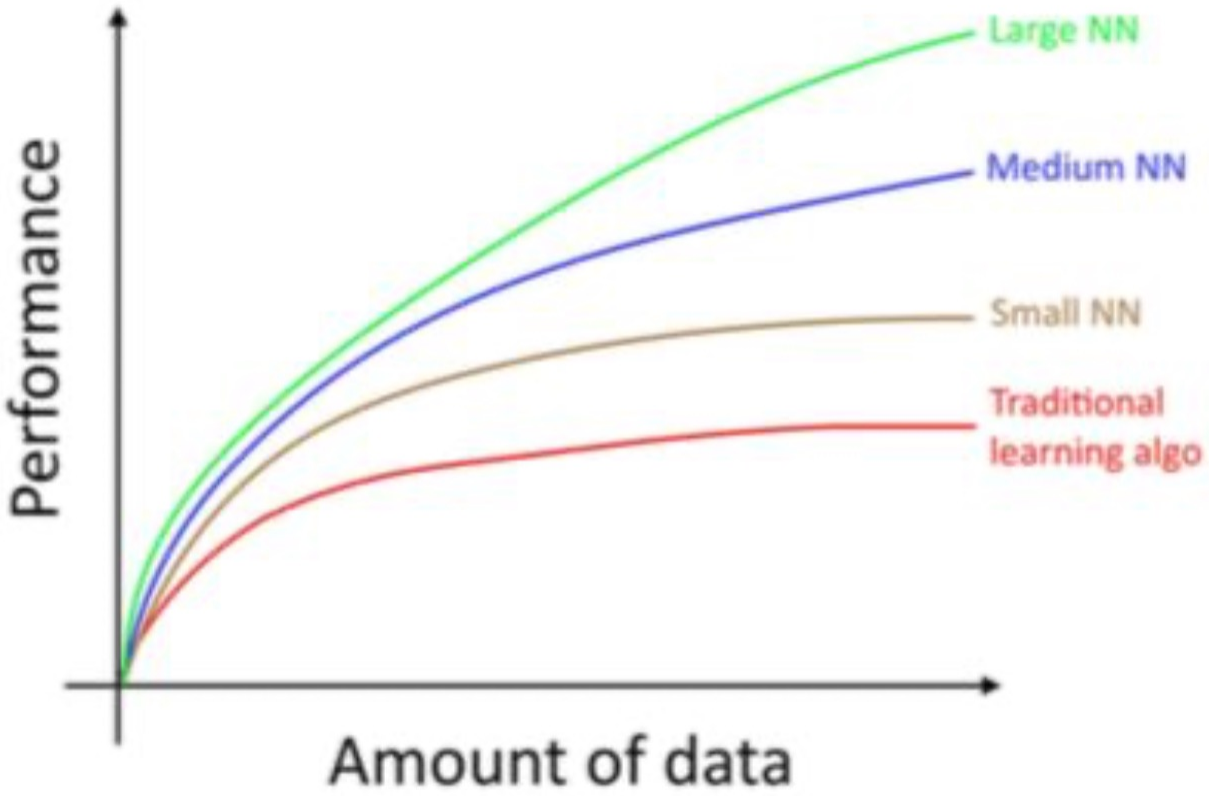
Moore's Law: Exponential Scaling of Computer Technology

- 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.



[Waldrop ('15) Nature]

More data … more detailed results?



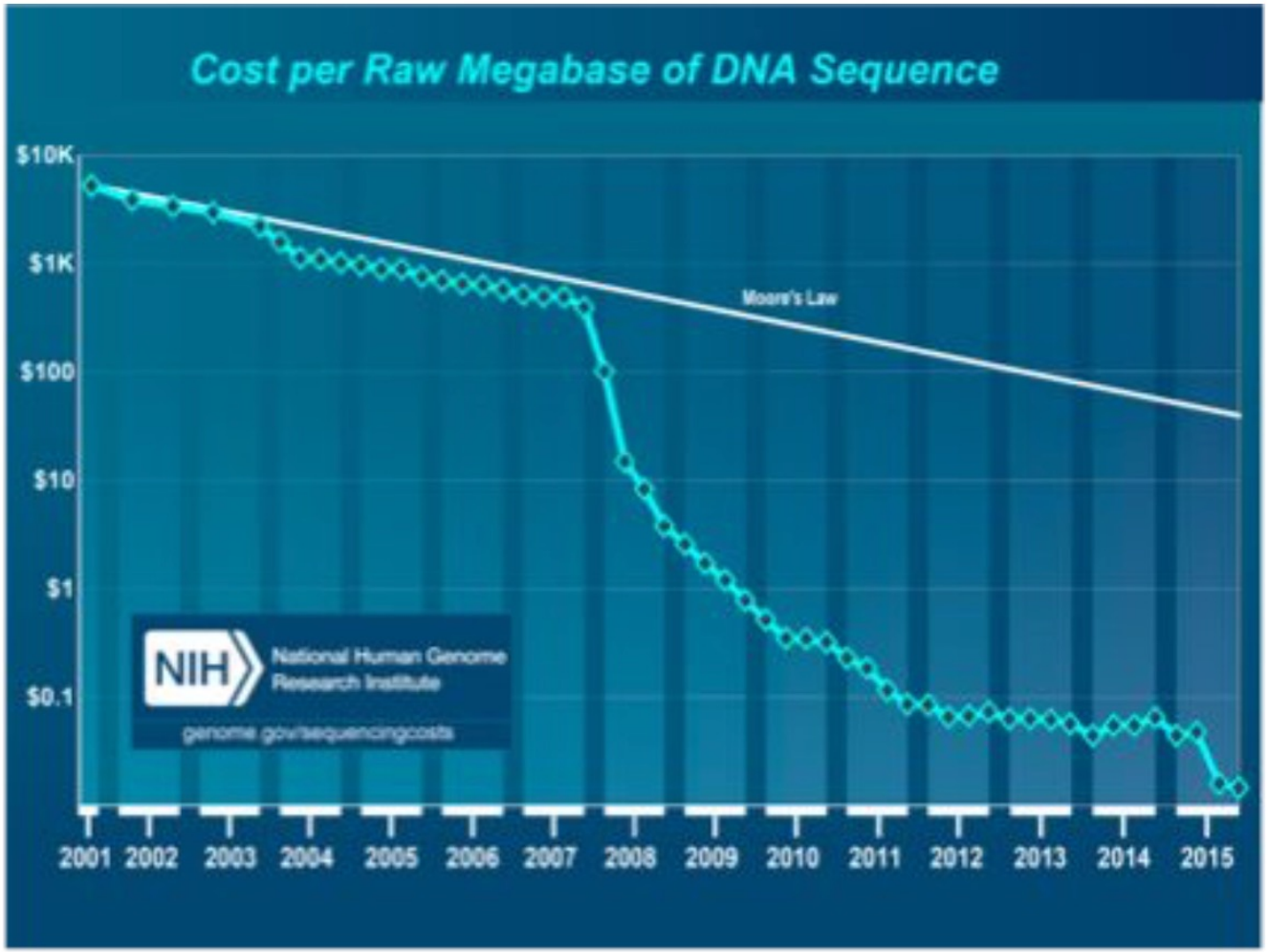
Source: Andrew Ng's book

Machine Learning – biological examples

DNA sequencing has gone through technological S-curves

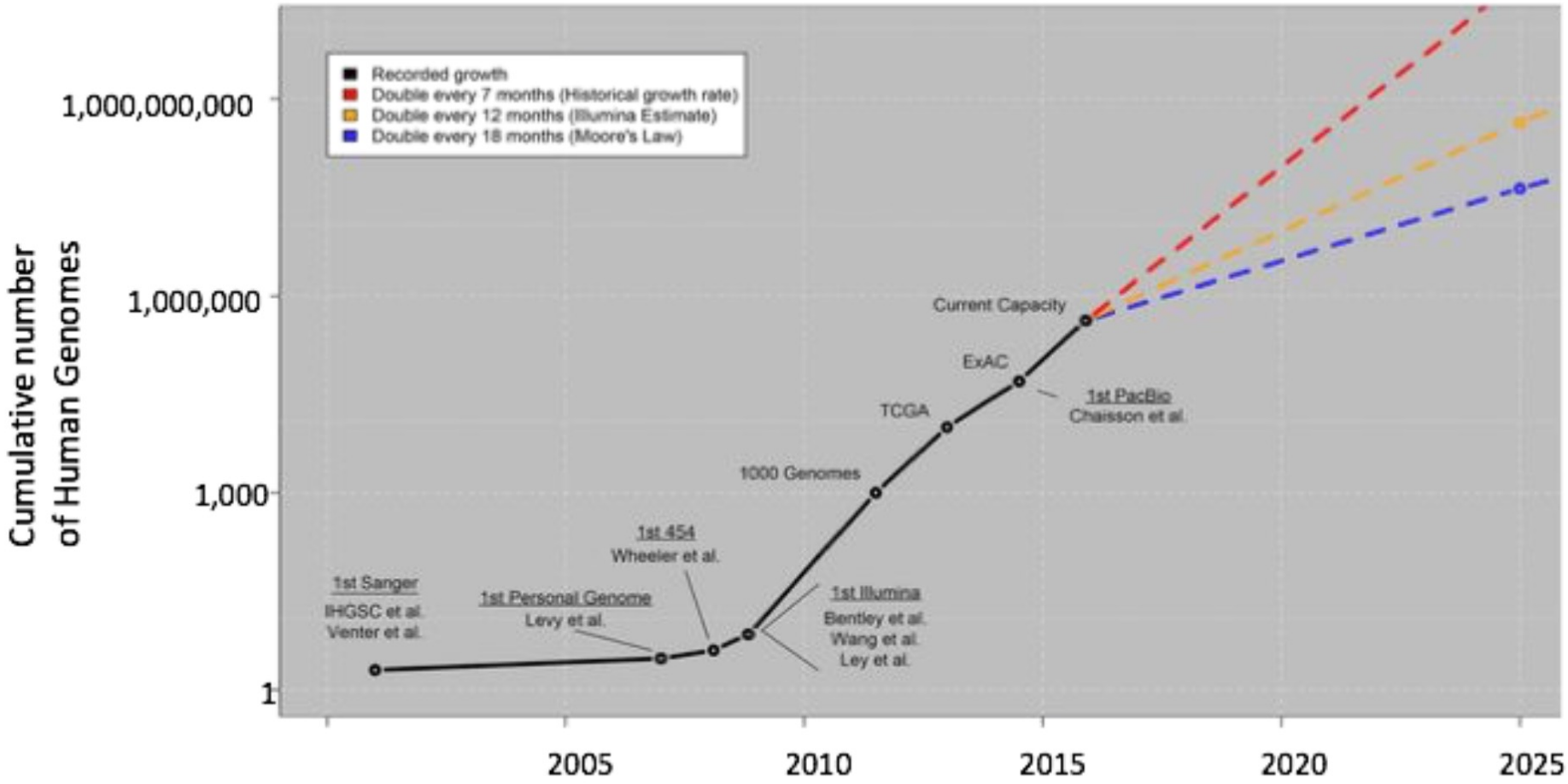
- In the early 2000's, improvements in Sanger sequencing produced a scaling pattern similar to Moore's law.

- The advent of NGS was a shift to a new technology with dramatic decrease in cost).



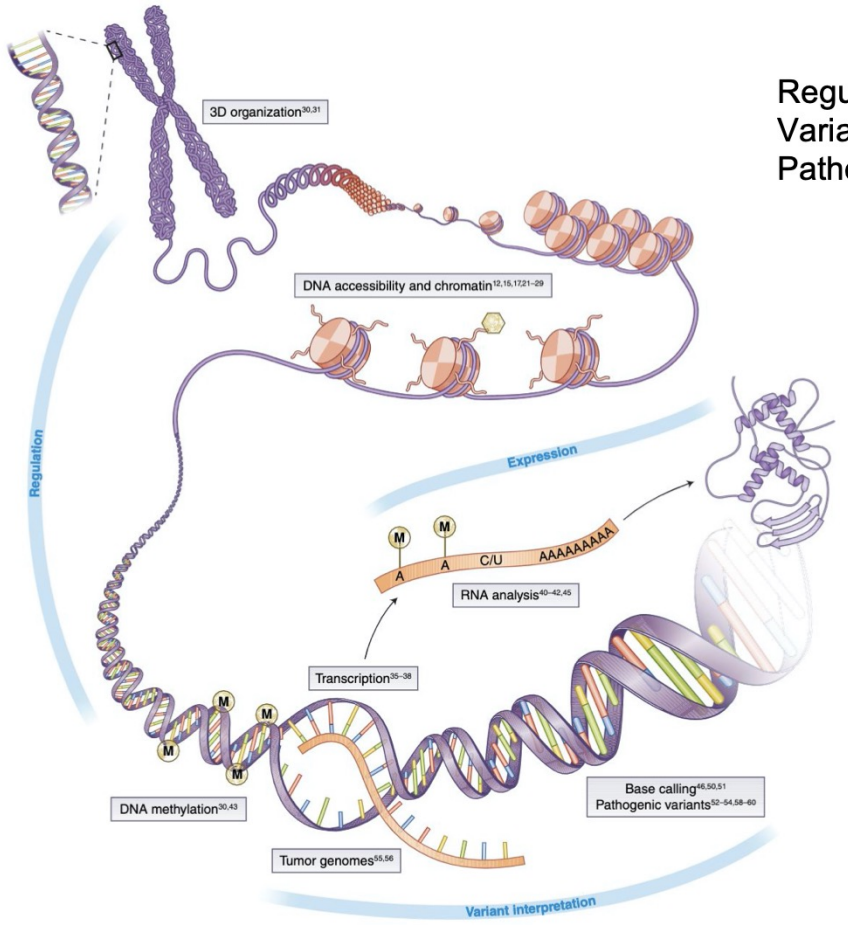
Moore's law

Available data. Where are we heading?



By 2019 total NGS data to exceed 2 Exabytes

Applications of ML in genomics

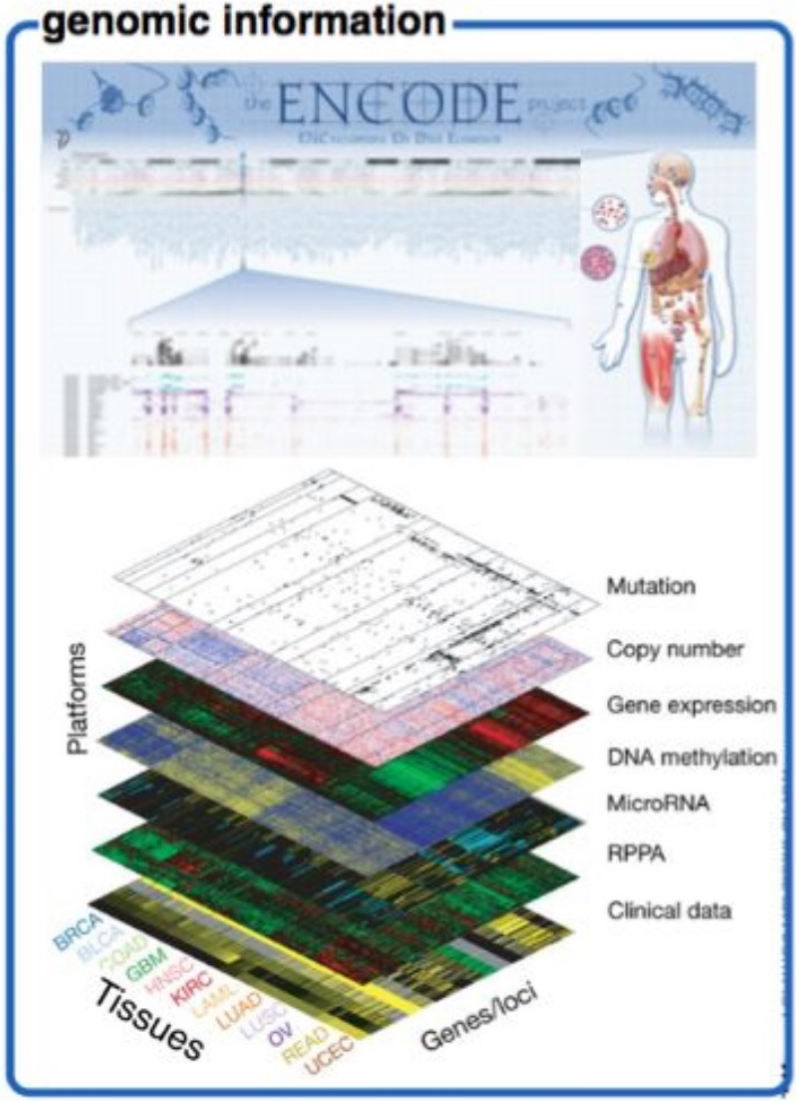


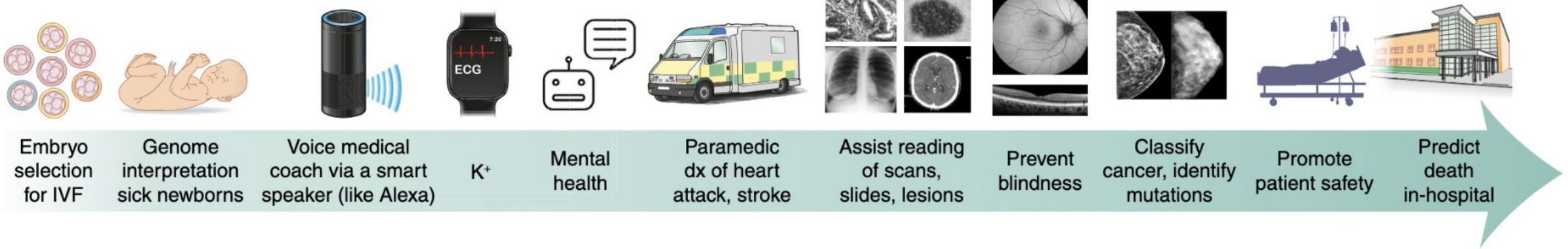
Regulatory genomics,
Variant calling and
Pathogenicity scores.

Zou, J. et al., 2018. A primer on deep learning in genomics. *Nature genetics*. Available at: http://dx.doi.org/10.1038/s41588-018-0295-5.

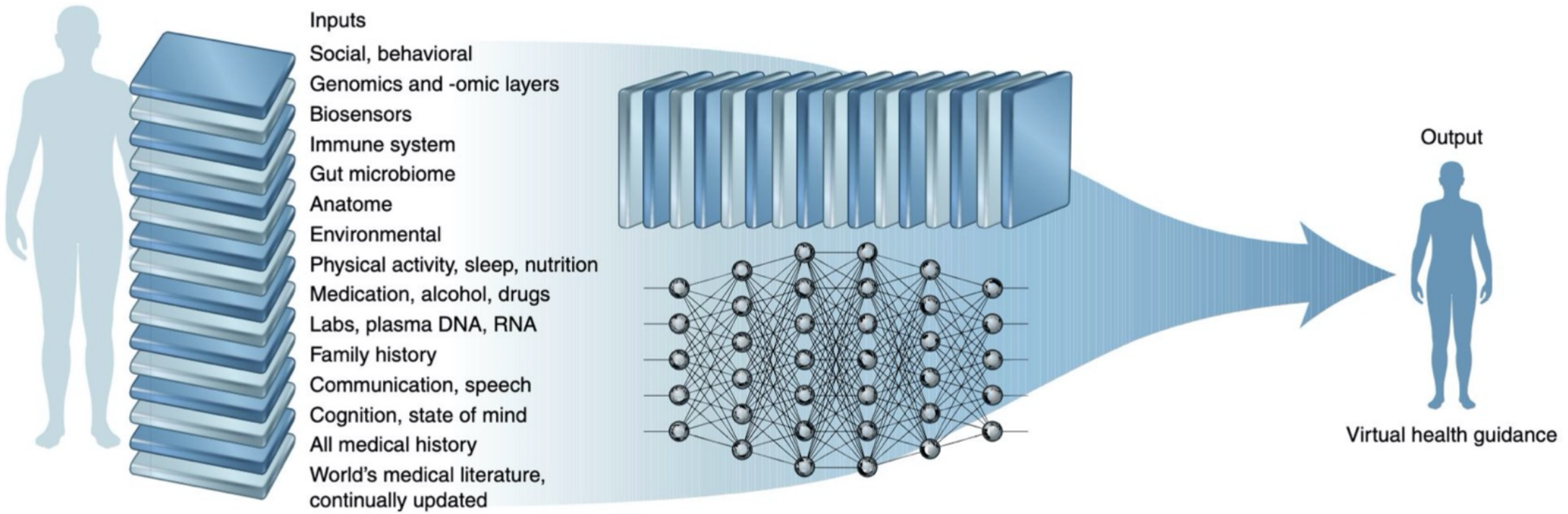
Applications of ML in genomics. More data

- Large-scale organisation providing an overview of the genome
- Integration of heterogeneous data





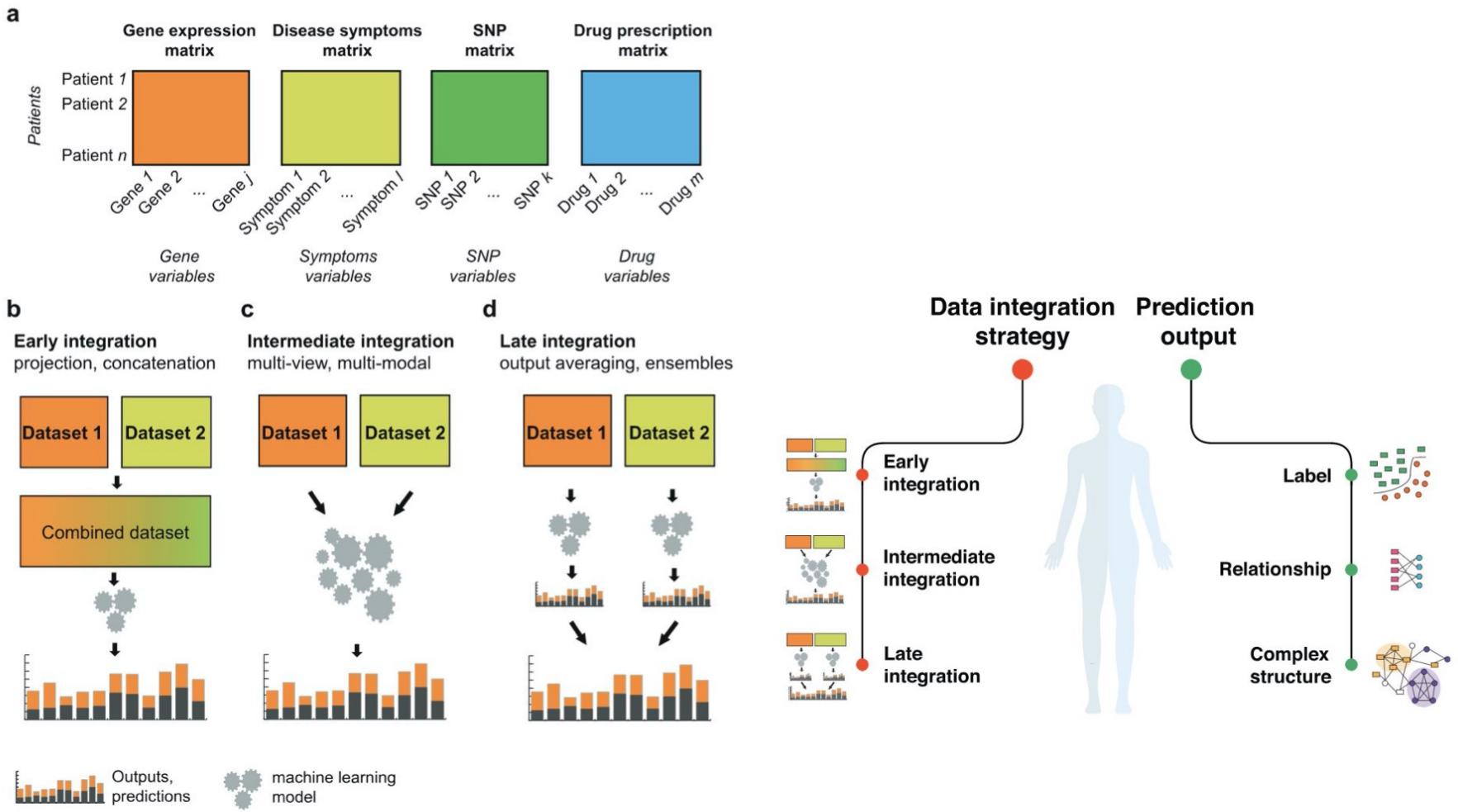
Multi-modal data inputs & algorithms to provide individualized guidance



Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), pp.44–56.

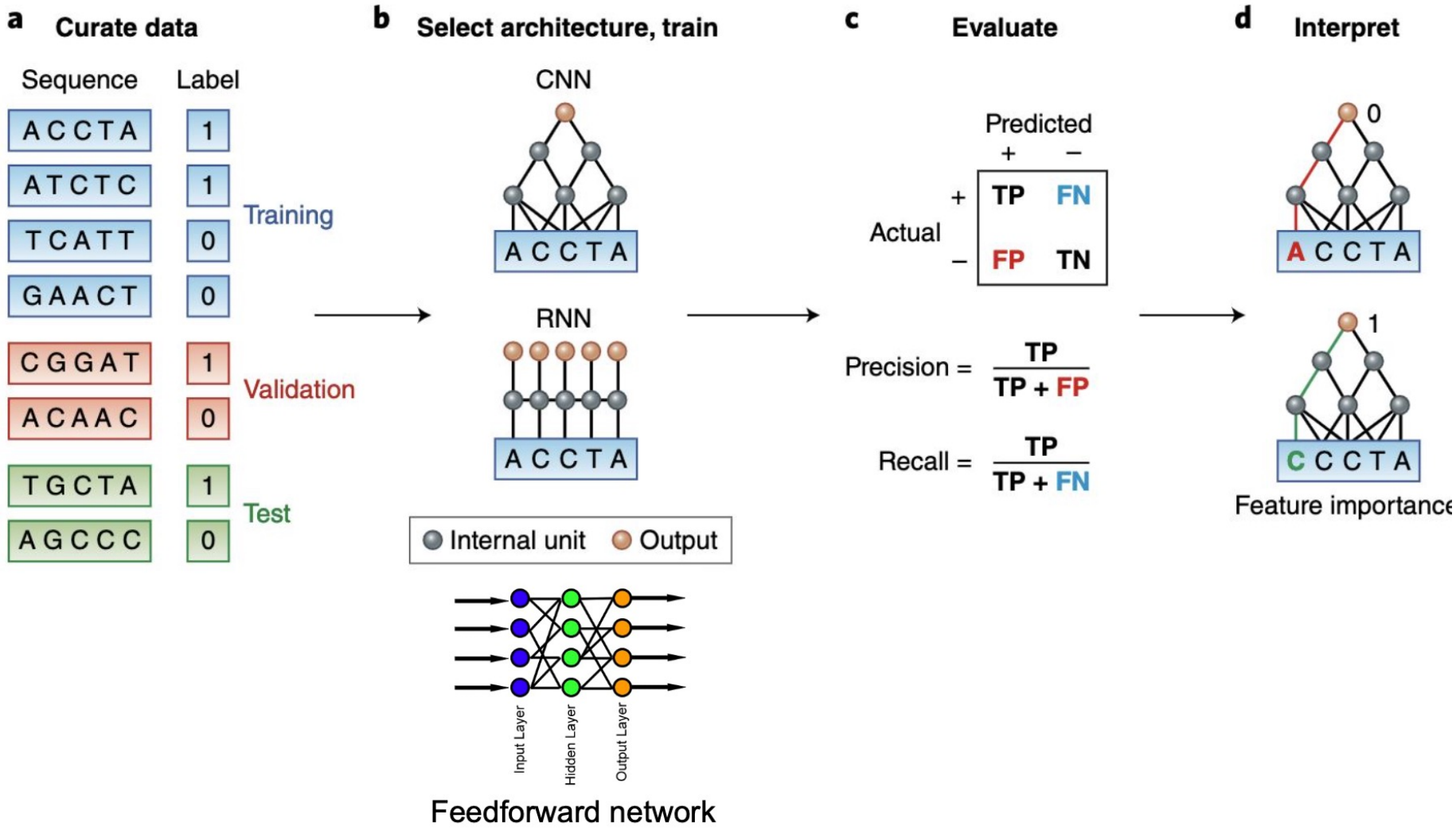
Data integration. Can ML help?

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



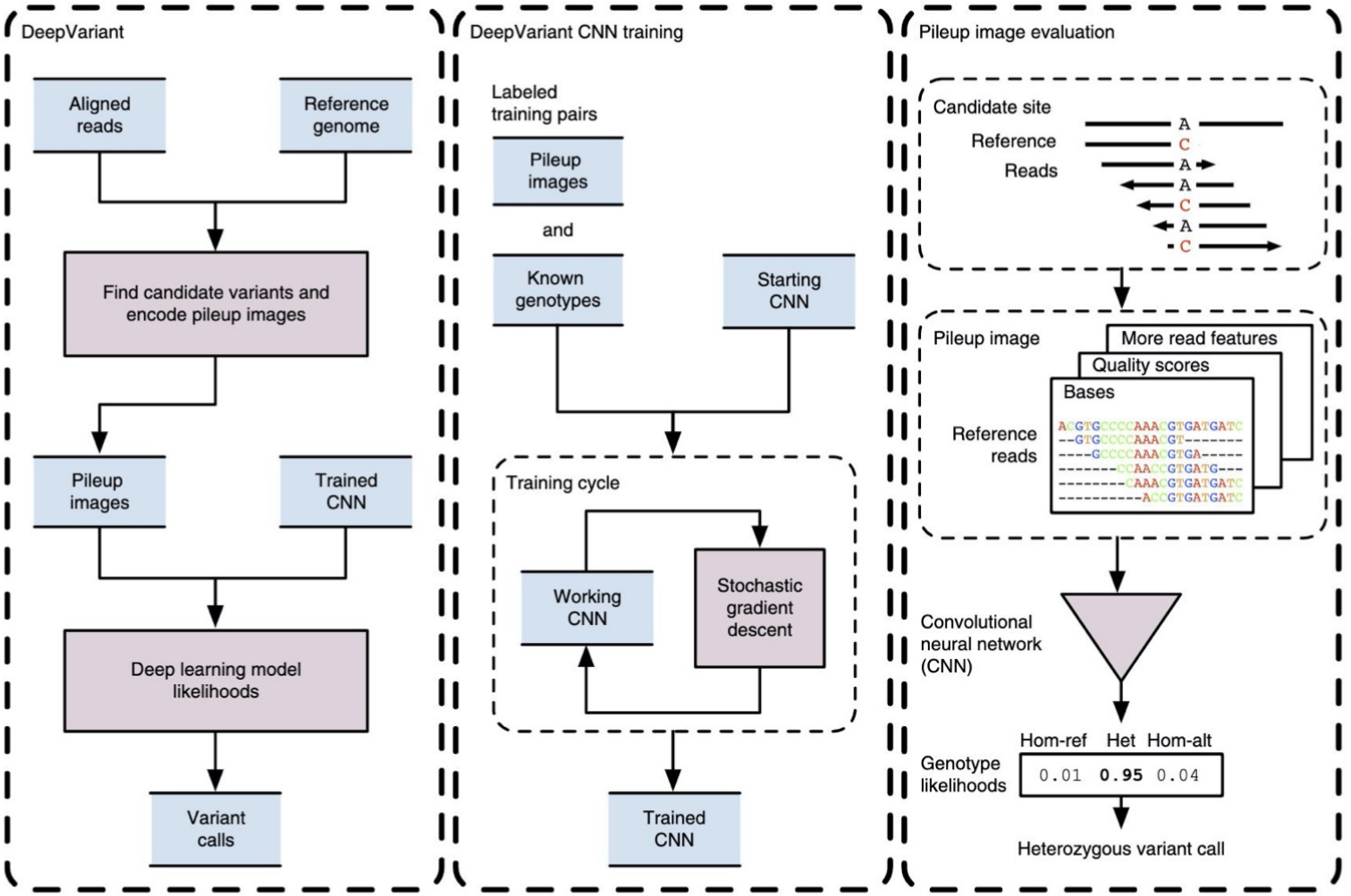
Zitnik, M. et al., 2019. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. An international journal on information fusion, 50, pp.71–91.

Steps for a ML-based pipeline



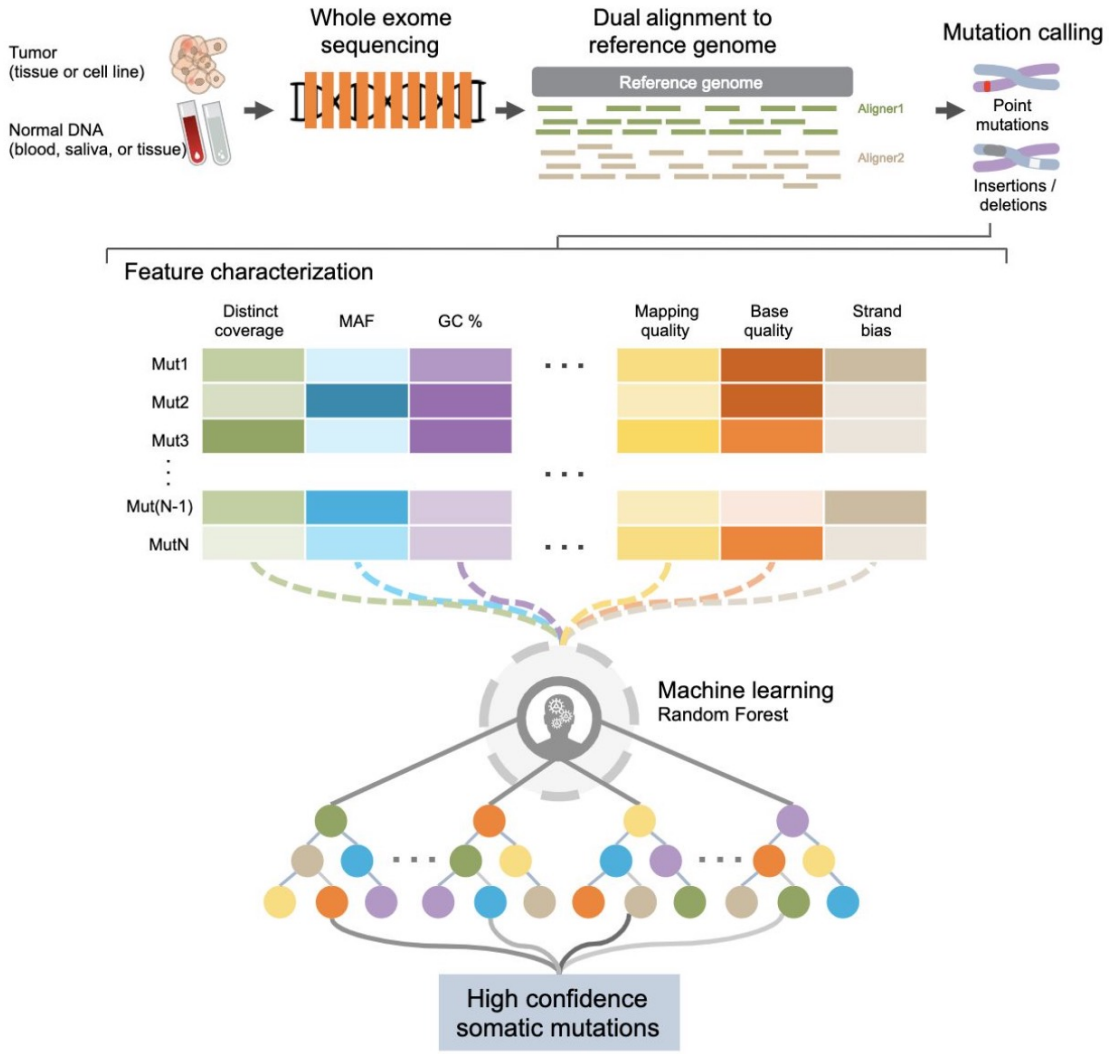
Zou, J. et al., 2018. A primer on deep learning in genomics. *Nature genetics*. Available at: http://dx.doi.org/10.1038/s41588-018-0295-5.

Example 1. Variant caller using NNs



Poplin, R. et al., 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10), pp.983–987.

Example 2. Somatic mutation discovery using RFs



Wood, D.E. et al., 2018. A machine learning approach for somatic mutation discovery. *Science translational medicine*, 10(457), http://dx.doi.org/10.1126/scitranslmed.aar7939.

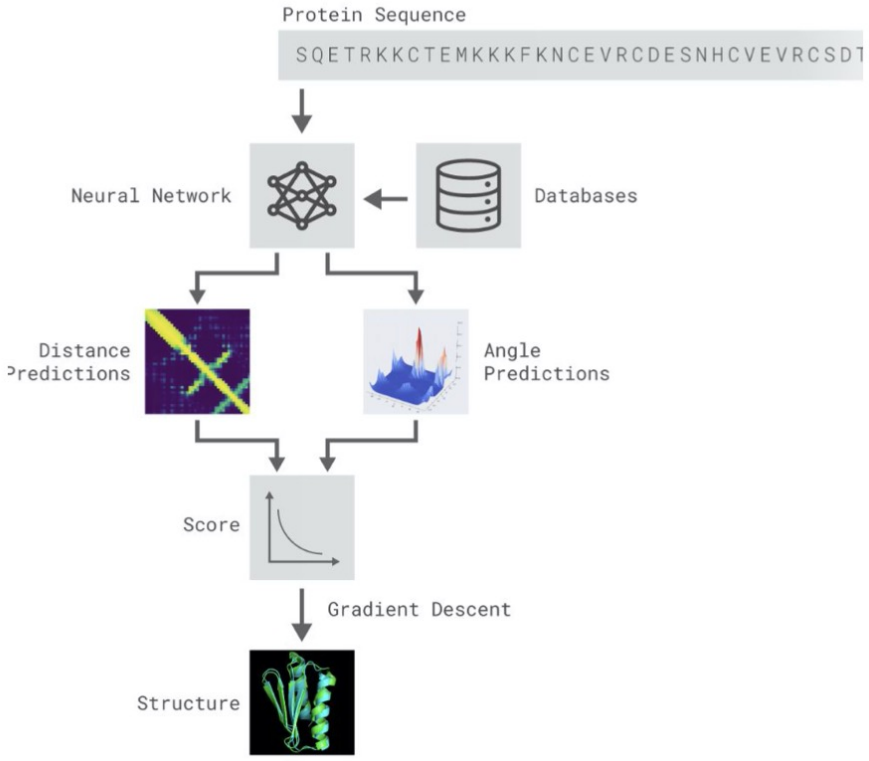
Example 3. DeepMind AlphaFold

AlphaFold: Using AI for scientific discovery

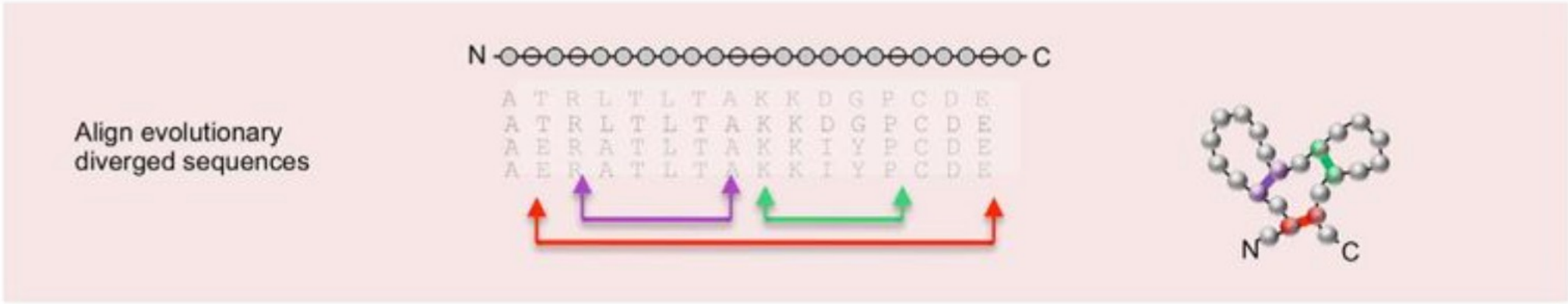
Today we're excited to share DeepMind's first significant milestone in demonstrating how artificial intelligence research can drive and accelerate new scientific discoveries. With a strongly interdisciplinary approach to our work, DeepMind has brought together experts from the fields of structural biology, 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



Calculate covariance matrix for each pair of sequence positions for all pairs of amino acids (A,B)

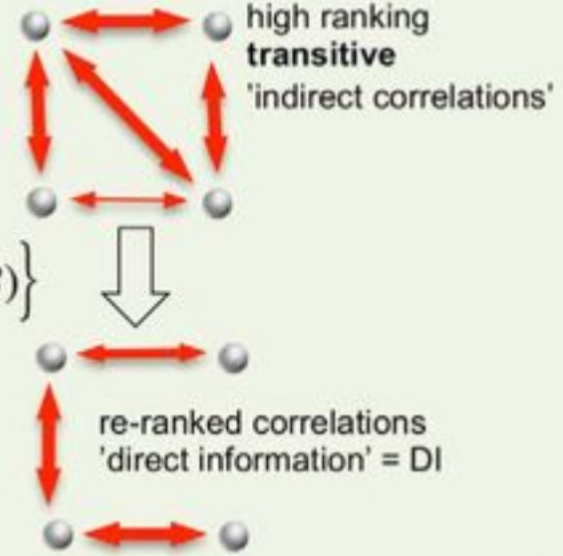
$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)P_j(B)$$

$$C_{ij}^{-1}(A,B) = -e_{ij}(A,B)_{i \neq j}$$

Identify maximally informative pair couplings using **statistical model** of entire protein to infer residue-residue co-evolution

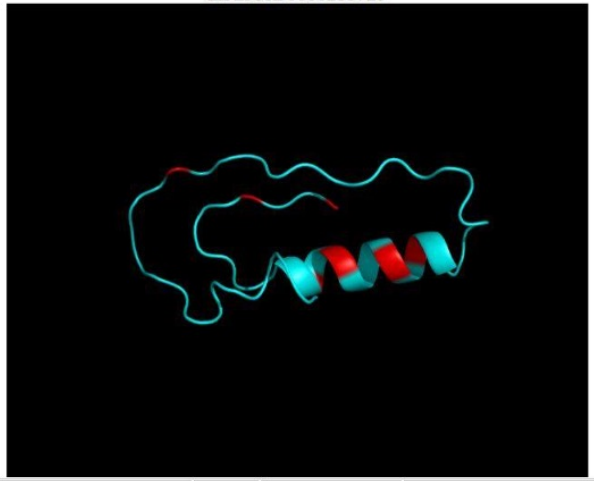
$$P_{ij}^{Dir}(A,B) = \frac{1}{Z}\exp\left\{e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B)\right\}$$

$$DI_{ij} = \sum_{A,B=1}^{q} P_{ij}^{Dir}(A,B)\ln\frac{P_{ij}^{Dir}(A,B)}{f_i(A)f_j(B)}$$



Example 3. DeepMind AlphaFold

mPLsORF0000239729



Mutation ID(s)	Change	Effect on sORF	sORF secondary structure
COSM3816594-AA=p.W105* COSM3816595-AA=p.W105*	G>A	W16>* (PDB W17)	H
COSM4056367-AA=p.Y91H COSM4056368-AA=p.Y91H	T>C	Y2>H (PDB Y3)	E
COSM5632701-AA=p.L109V COSM5632702-AA=p.L109V	C>G	L20>V (PDB L21)	H
COSM6324258-AA=p.A94V COSM6324259-AA=p.A94V	C>T	A5>V (PDB A6)	C
COSM6761144-AA=p.E103D COSM6761145-AA=p.E103D	G>T	E14>D (PDB E15)	H
COSM964154-AA=p.D111Y	G>T	D22>Y (PDB D23)	H
COSM964155-AA=p.R129H	G>A	R40>H (PDB R41)	C

mPLsORF0000424584

Mutation ID(s)	Change	Effect on sORF	sORF secondary structure
COSN1260962 COSN5770834	A>G	Y10>C (PDB Y10)	H
COSN18723188	C>T	P32>L (PDB P32)	C
COSN20080080	C>T	R13>C (PDB R13)	C
COSN20121983	C>T	I28>I (PDB I28)	C
COSN20648546 COSN21264986 COSN23079780	G>A	W50>* (PDB R50)	H
COSN20650206 COSN21677358 COSN22583798 COSN23084029	C>T	G42>G (PDB G42)	E
COSN20650207 COSN22687956 COSN23081188	T>C	W50>R (PDB R50)	H
COSN21264985 COSN22557586 COSN23085391	A>G	K49>K (PDB K49)	H
COSN21675097	T>C	Y43>Y (PDB Y43)	H
COSN23082255	G>A	K34>K (PDB K34)	H
COSN23087874	A>G	K26>E (PDB E26)	H
COSN24397730	G>A	R23>H (PDB R23)	H
COSN26666338	A>T	E67>V	H
COSN27001349 COSN8016762	G>A	T20>T (PDB T20)	C
COSN27001446	C>T	T20>M (PDB T20)	C
COSN27001936	G>T	R70>M	C
COSN27003013	G>A	G54>D (PDB G54)	C

Example 1. Chemistry

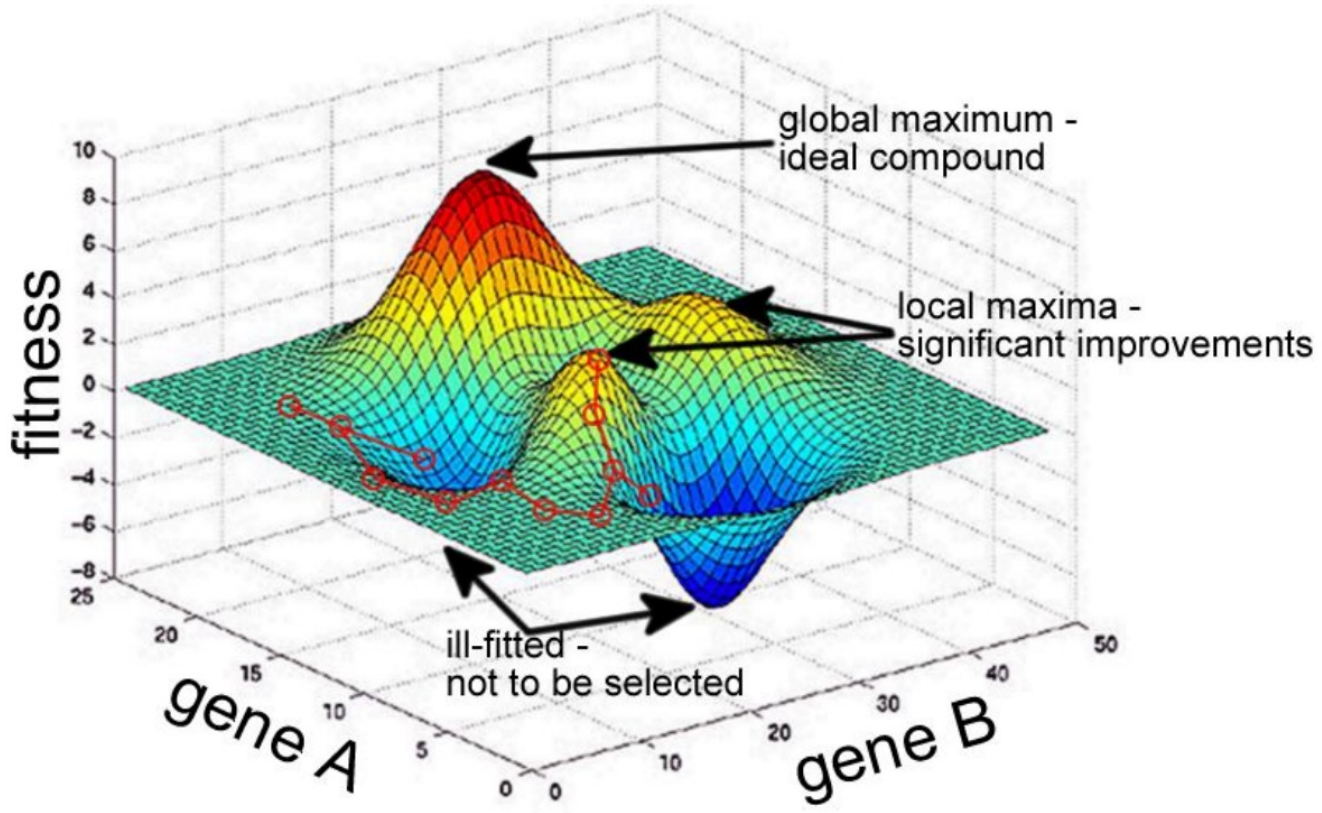


Figure 3 - A fitness landscape considering two hypothetical genes as, for example, block size and processing temperature of a polymer synthesis. Fitness could be hardness, for instance. The landscape contains local maxima, a global maximum, and a global minimum. In red, the path of a genetic algorithm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

Source: Rodrigues et al, A survey of Big Data and Machine Learning for Chemistry

Example 2. Chemistry

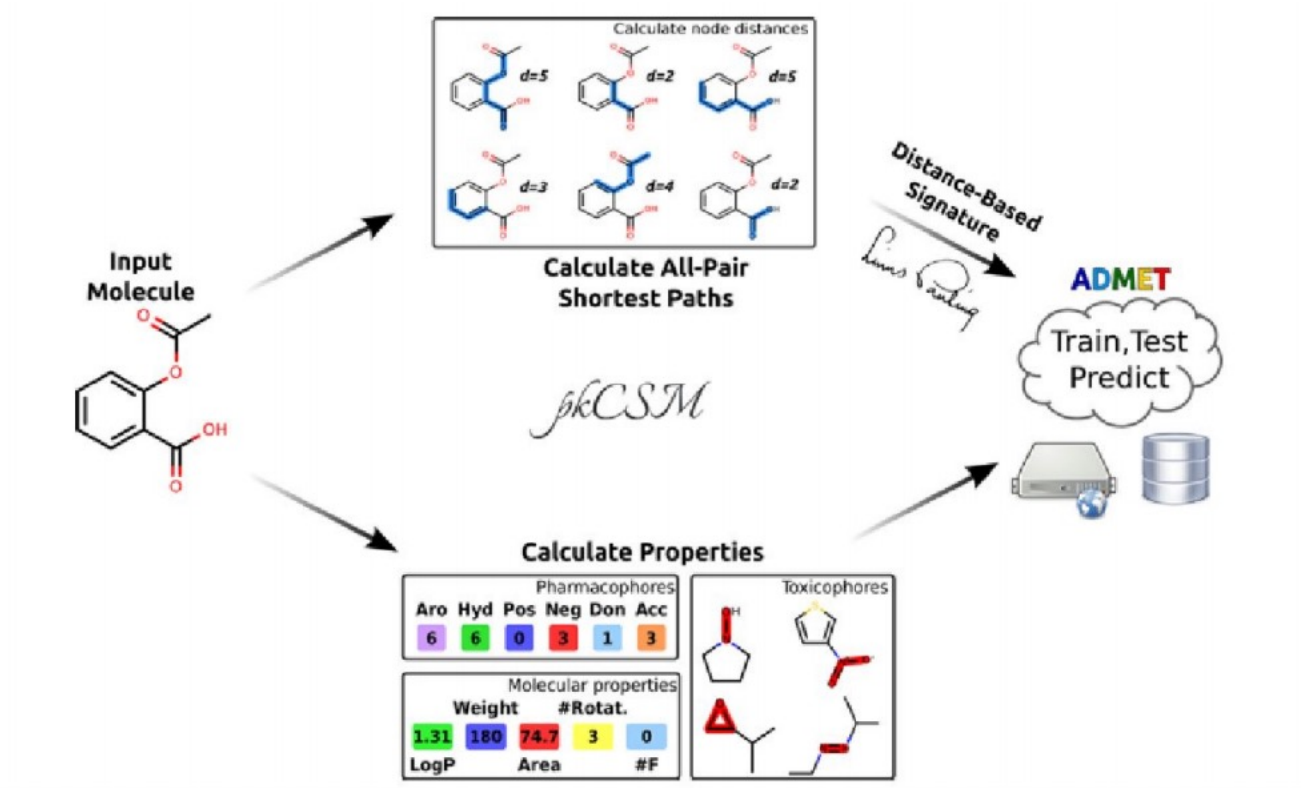


Figure 6 – The workflow of pkCSM is represented by the two main sources of information, namely the calculated molecular properties and shortest paths, for an input molecule. With these pieces of information, the ML system is trained to predict ADMET properties. Reproduced from [64].

Source: Rodrigues et al, A survey of Big Data and Machine Learning for Chemistry

Example from Physics

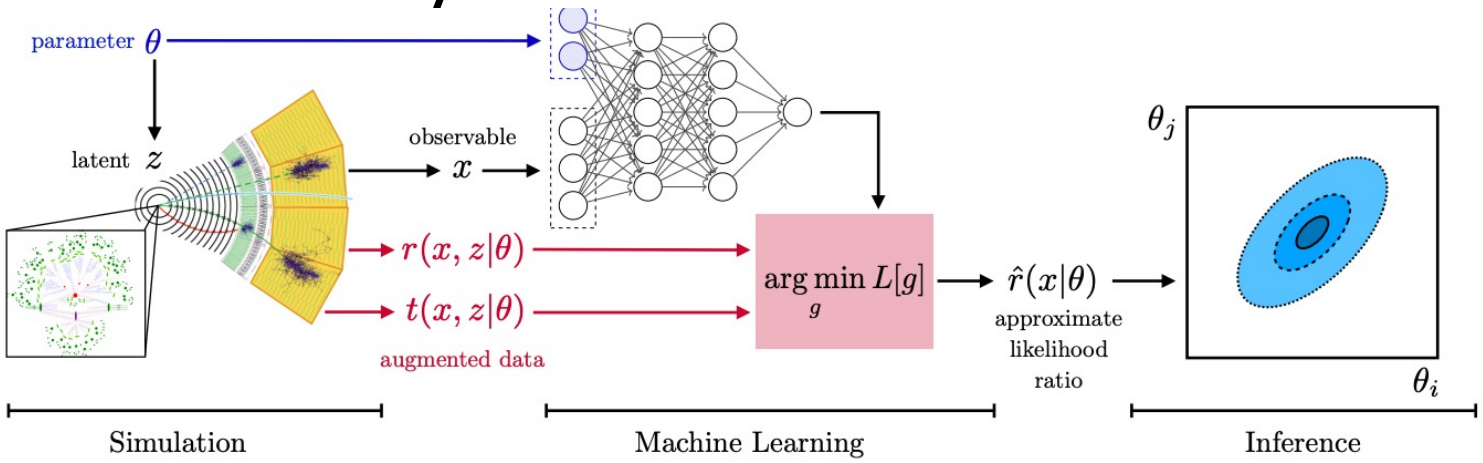


Figure 2 A schematic of machine learning based approaches to likelihood-free inference in which the simulation provides training data for a neural network that is subsequently used as a surrogate for the intractable likelihood during inference. Reproduced from (Brehmer *et al.*, 2018b).

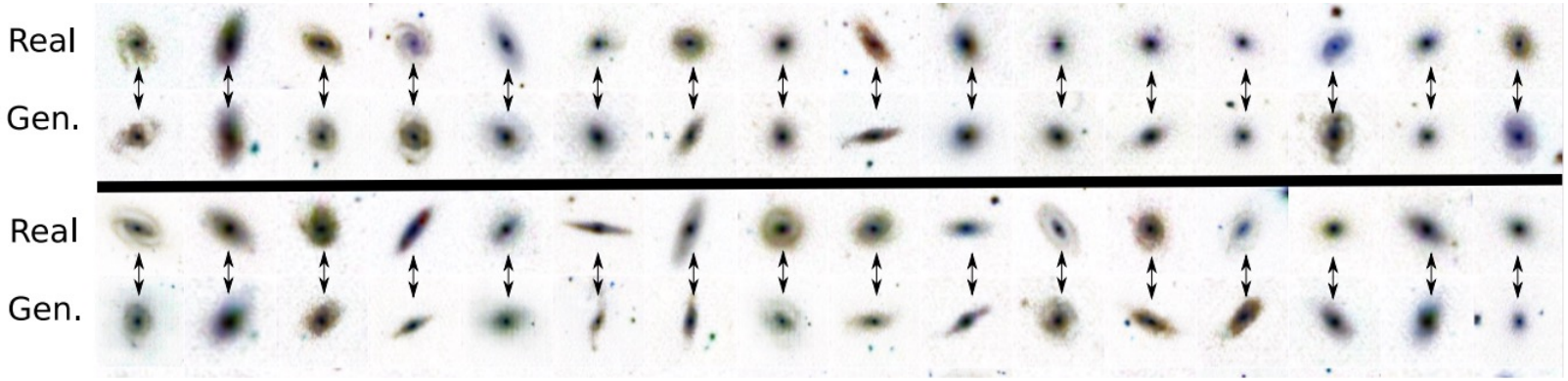
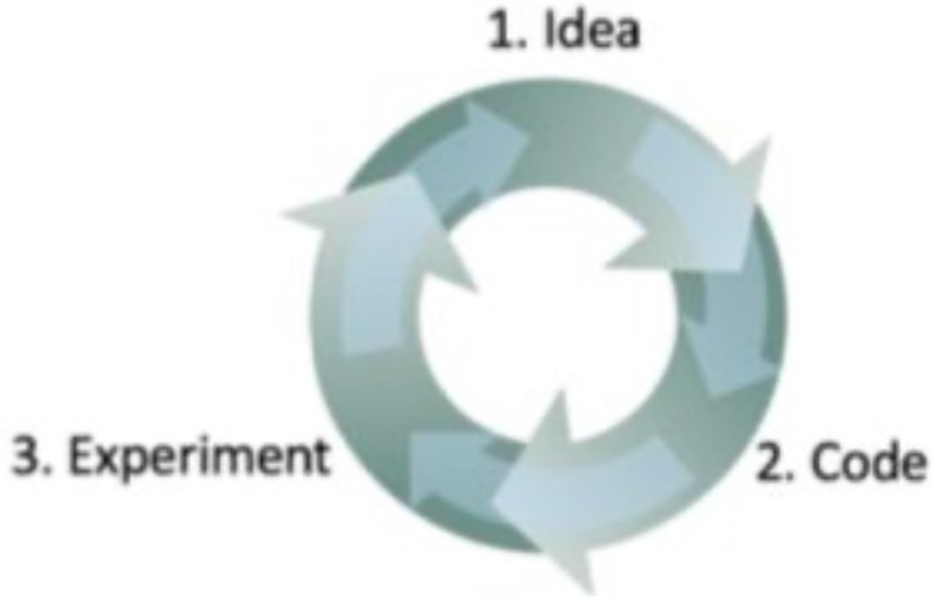


Figure 3 Samples from the GALAXY-ZOO dataset versus generated samples using conditional generative adversarial network. Each synthetic image is a 128×128 colored image (here inverted) produced by conditioning on a set of features $y \in [0, 1]^{37}$. The pair of observed and generated images in each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

Figures from Machine Learning and the physical science, Carleo et al

ML workflow

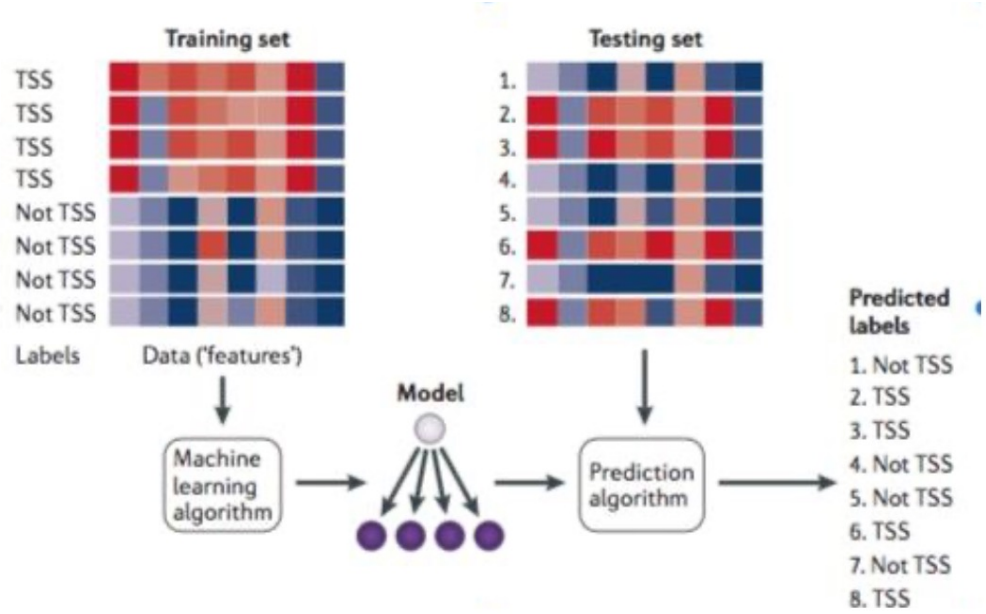


Dev/Test set & Metrics

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

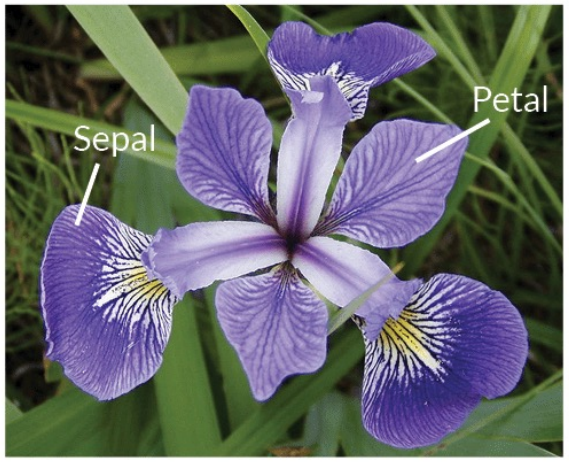


Step 1: develop an algorithm that will lead to successful learning.

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.

Iris dataset



Iris Versicolor

Iris Setosa

Iris Virginica

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Data Set Information:

This is perhaps the best known database to be found in the pattern recognition literature.
Fisher's paper is a classic in the field and is referenced frequently to this day.
The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.
One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Predicted attribute: class of iris plant.