AI/ML in life sciences

David Dai - April 9, 2021

Overview

Laying the foundation - what is AI/ML?

Bridging the gap - biology to algorithms

Case Study - AlphaFold

Resources

What is AI?

- “AI” is used very loosely nowadays, and seems to mean many different things
- I will refer to “AI” not as artificial general intelligence (ie. consciousness, ability to reason), but simply as artificially-intelligent systems that can do non-trivial tasks
- At times, I may interchange AI and ML (machine learning); ML is a subset of AI, and has recently been the most successful, hence will be the main focus

Artificial intelligence

What (I think) it is currently

- Current AI systems use a mix of the following strategies for task-solving:
    - **Expert-derived** rules-based systems
    - **Data-driven** pattern-matching systems

Automated cell culture media exchanger

Hypothetical AI Product

- A company develops a machine that automatically refreshes cell culture media, tailored to the specific cell type
- The data science team propose two methods to automate this process:
    - R&D team provides specific optimal conditions for all on-market cell types (eg. mice cells ~24h, rat cells ~30h, …) -> **Expert-derived**
    - DS team applies natural language processing methods to research literature and cell product brochures to identify optimal cell types vs. conditions -> **Data-driven**

Limitations of these approaches

- Expert systems: requires subject expertise, rules change, edge cases
- Data-driven systems: data collection can be time consuming and expensive, data may not be available

Despite limitations, AI has (and will be) very impactful

- Sentient machines are probably not imminent
- Cumulative (and maybe individually small) effect of AI (specifically ML) will bring about impactful change in our research, work, and every day lives [1]
- AI will soon be “just another tool in the toolbox”

[1] Moore's Law for Everything

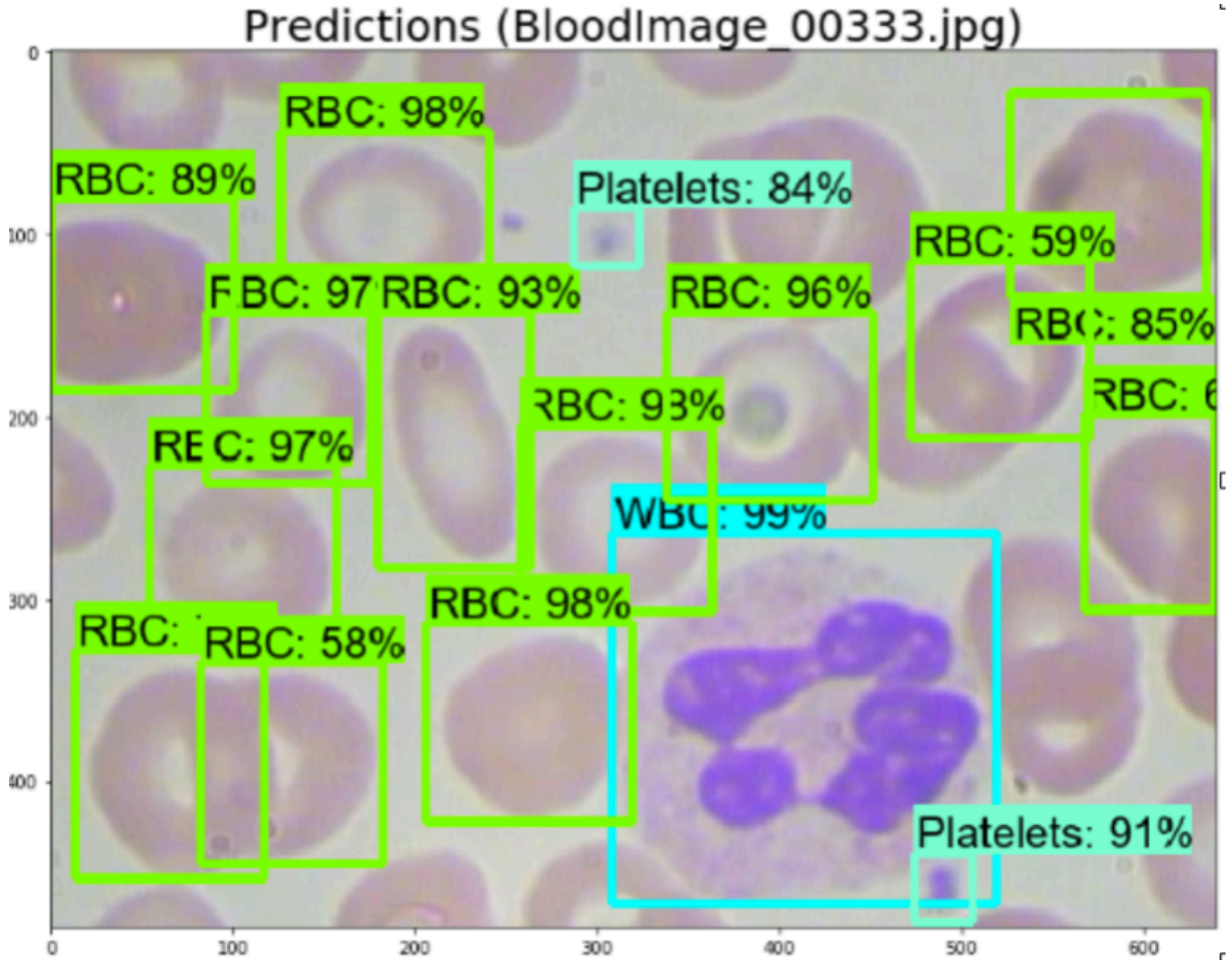
Machine learning

Teaching computers to learn from data

- Machine learning (ML) is a subset of AI
- ML allows us to build models that “learn” patterns from historical data so that it can be applied to future data
- Three classical branches of machine learning
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

Supervised learning

- Learn the relationship between input data and output
- Suitable for well-defined tasks
- Eg. classifying the location and type of blood cells from an image
  - Input data: image
  - Output: cell type & location



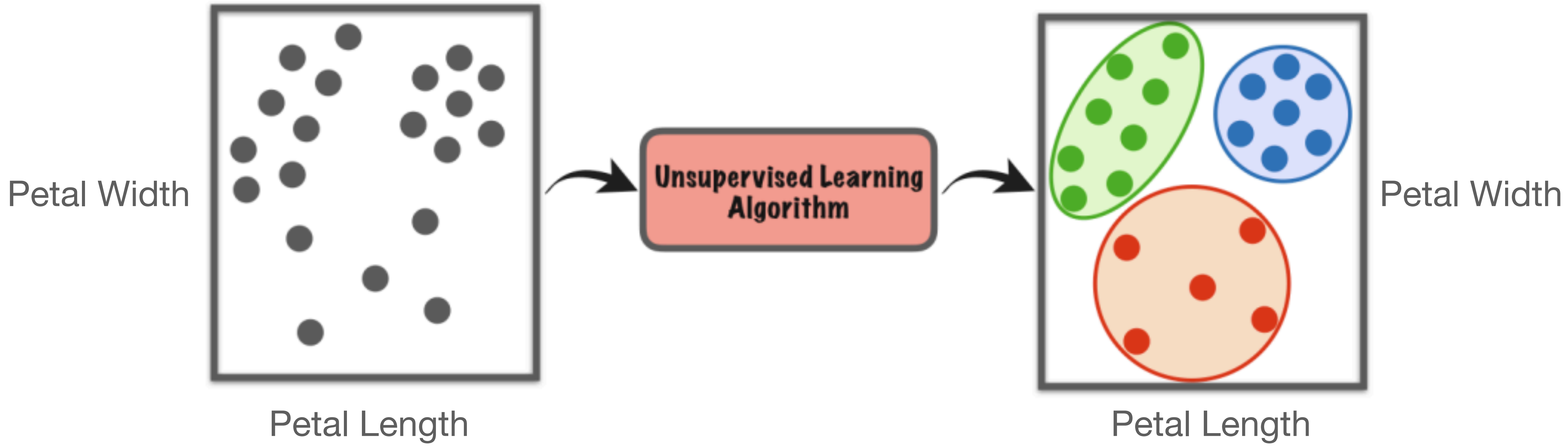
https://towardsdatascience.com/detection-and-classification-of-blood-cells-with-deep-learning-part-2-training-and-evaluation-53381dbbc565

Unsupervised learning

- Learn from implicit structure of the input data

- Allows us to quantify the similarity/dissimilarity between data points

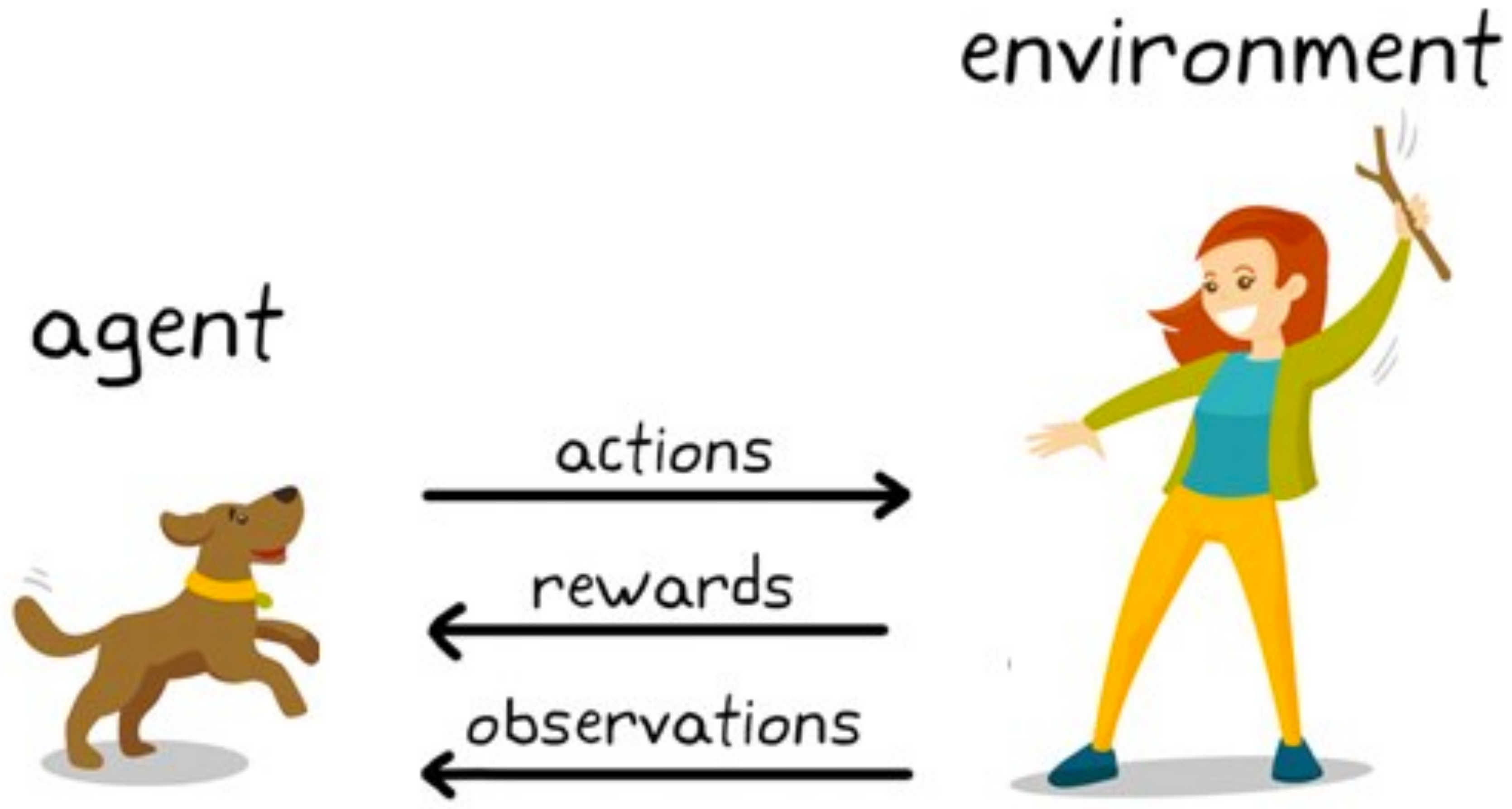
- Suitable for exploratory/poorly-defined tasks, or if labelled data is not available



https://towardsdatascience.com/supervised-vs-unsupervised-learning-in-2-minutes-72dad148f242

Reinforcement learning

- Training an agent to learn by interacting with its environment
- Agent learns by iterating through a cycle:
  - Observe environment
  - Take action
  - Receive reward
  - Update belief



https://www.kdnuggets.com/2019/10/mathworks-reinforcement-learning.html

How does a model “learn”?

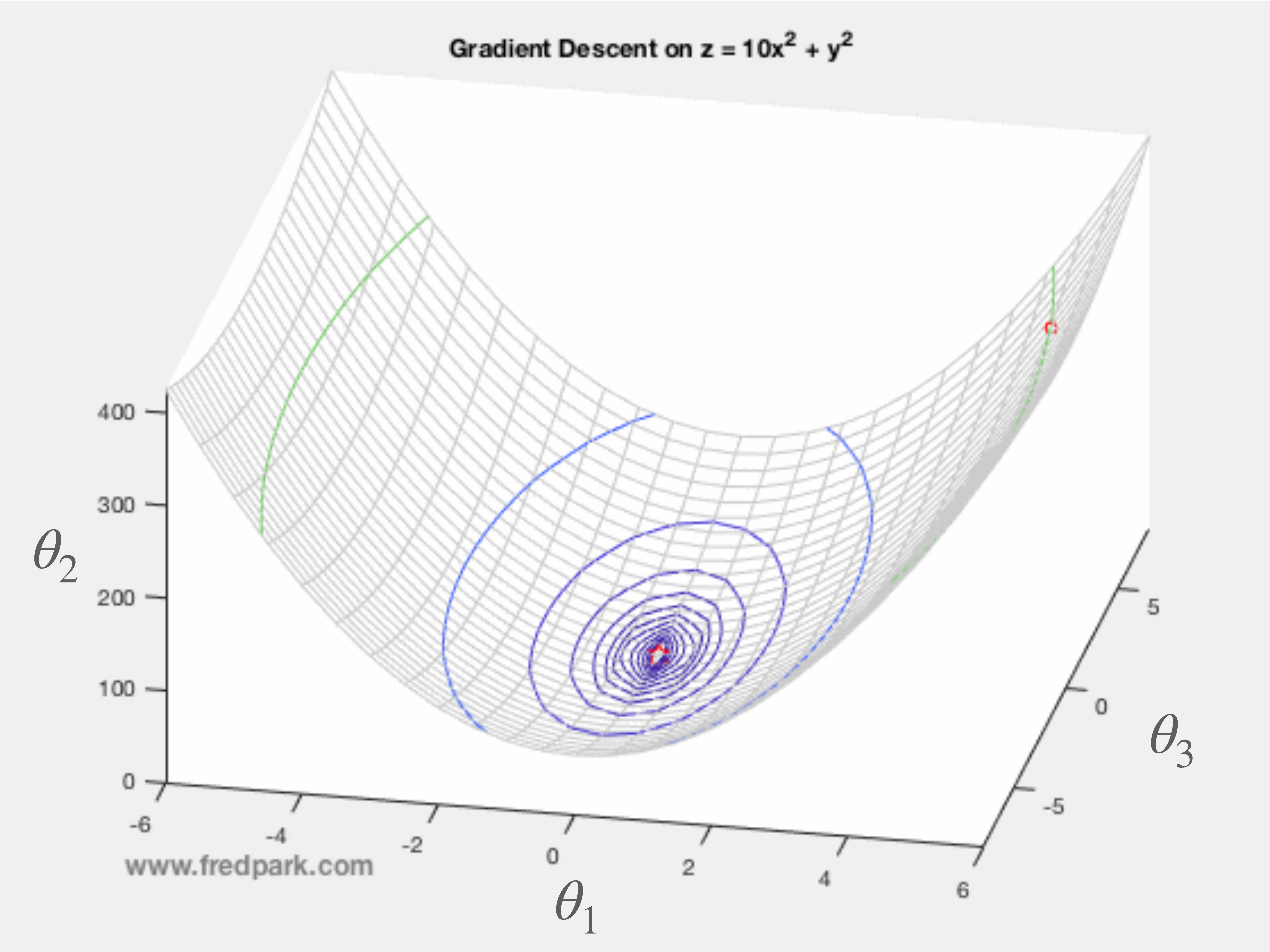
- A model that “learns” from data can be viewed as an optimization process
- The “learning” occurs as the model optimizes its parameters to find a combination of parameters that produces a function that best fits the data
- “Best fit” is determined by a scoring criteria (ie. objective) that compares the model to the observed data, which is designed for a specific task
    - Example objective functions: Classification accuracy (supervised learning), distance to cluster centroid (unsupervised learning), reward function (reinforcement learning)

The learning process

Components of a trainable model

- *x* - input data (eg. an image containing blood cells, flower petal width/length)
- *y* - target data (eg. cell type classification, centroid of the flower clusters)
- θ - the learned parameters that characterize the model, *f*
- *L* - an objective function

$$\underset{\theta}{\operatorname{argmin}}\, L\Big(y, f(x, \theta)\Big)$$



http://www.fredpark.com/blog/gradient-descent

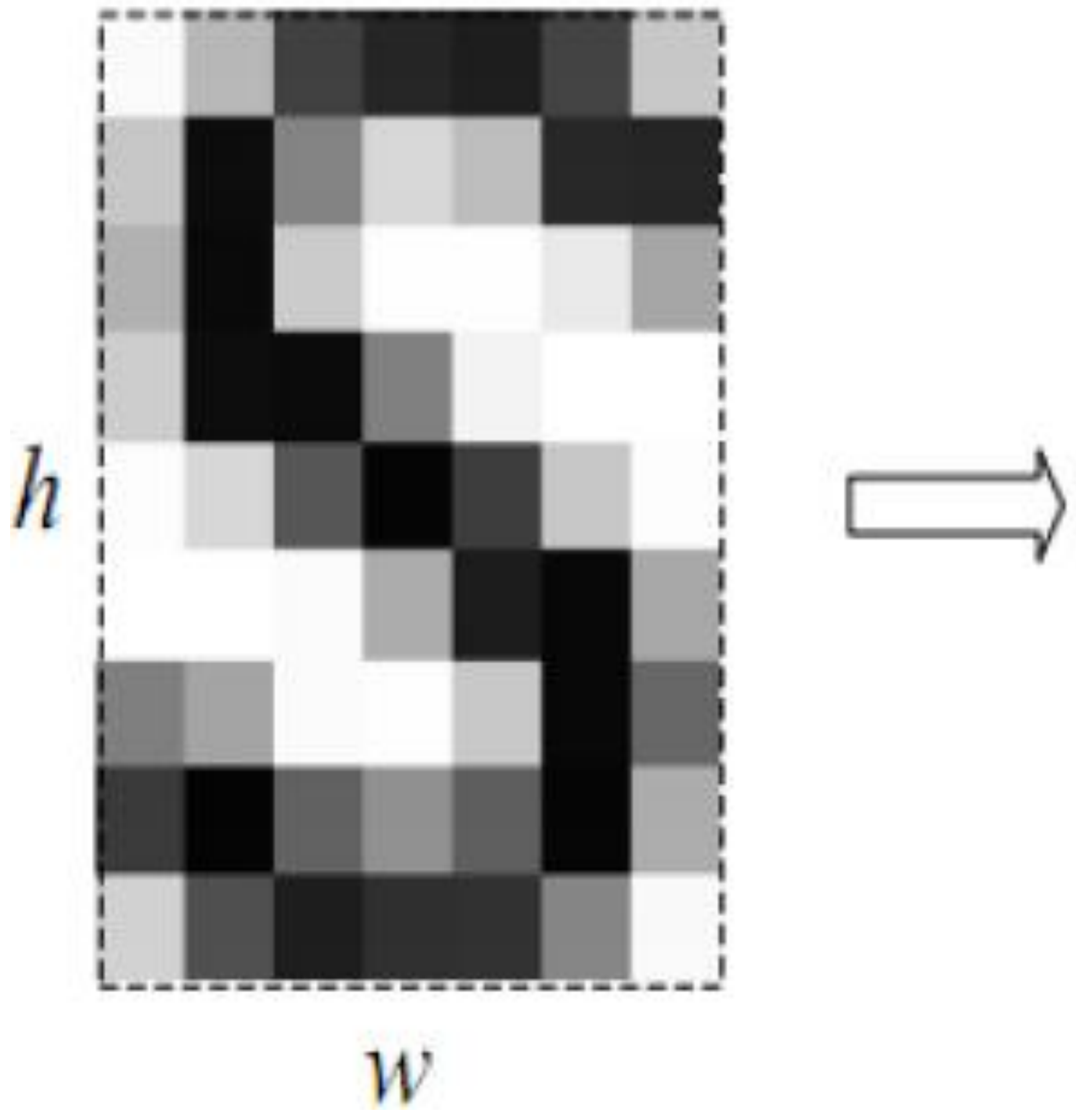
Biology to ML models

- ML models rely on numeric data -> how do we translate biological constructs to numbers?
- Key is transforming unstructured data into information-preserving, numerical representations
    - Image data representation
    - Language data representation

Image data

Representing greyscale images

- Greyscale images can be represented as a 2-dimensional array of pixel (ie. light) intensity



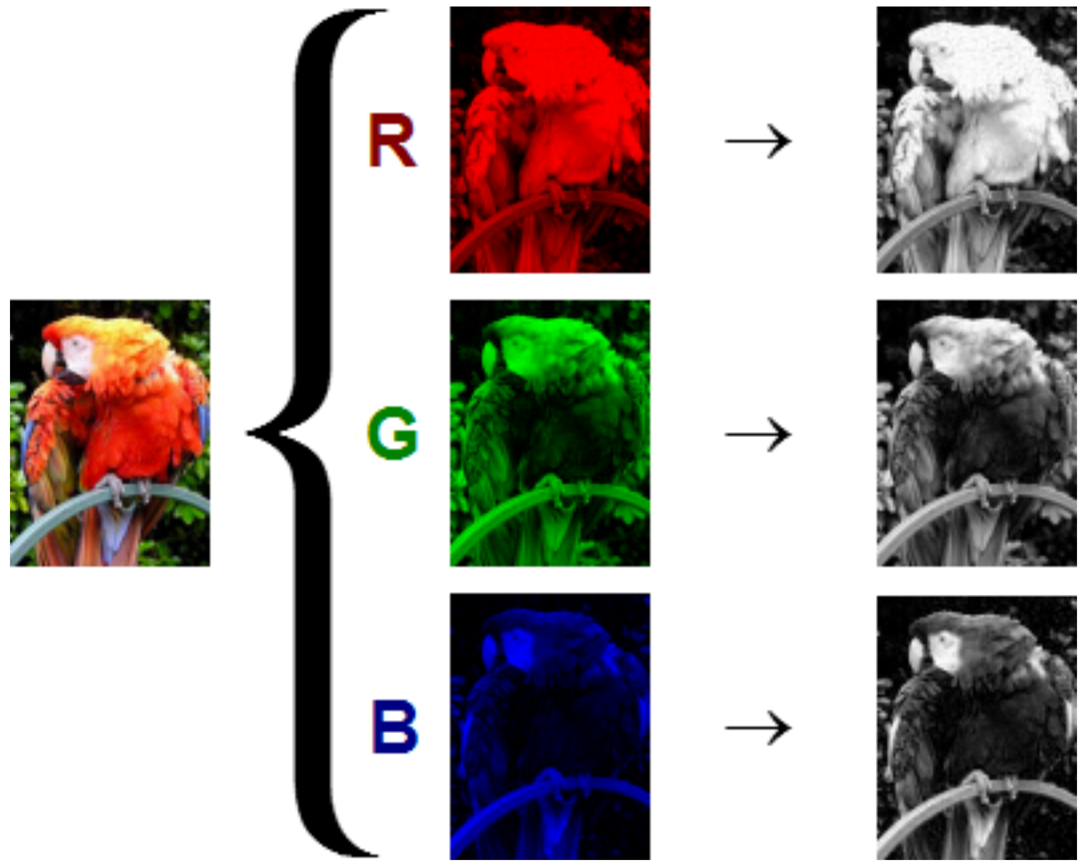
$$\mathbf{x} = \left(\begin{array}{l} 251, 181, 068, 041, 032, 071, 197, \\ 196, 014, 132, 213, 187, 043, 041, \\ 174, 011, 200, 254, 254, 232, 164, \\ 202, 014, 012, 128, 242, 255, 255, \\ 253, 212, 089, 005, 064, 196, 253, \\ 255, 255, 251, 196, 030, 009, 165, \\ 127, 162, 251, 254, 197, 009, 105, \\ 062, 005, 100, 144, 097, 006, 170, \\ 207, 083, 032, 051, 053, 134, 250 \end{array} \right)$$

https://www.researchgate.net/figure/The-pixel-matrix-feature-extraction-method_fig2_284003940

Image data

Representing coloured images

- Similarly, coloured images can be represented by decomposing the Red/Green/Blue light intensity
  - Extract the RGB channel light intensity as three individual 2-dimensional array
  - Concatenate the three arrays into a 3D dimensional array, with the third dimension representing the RGB channels

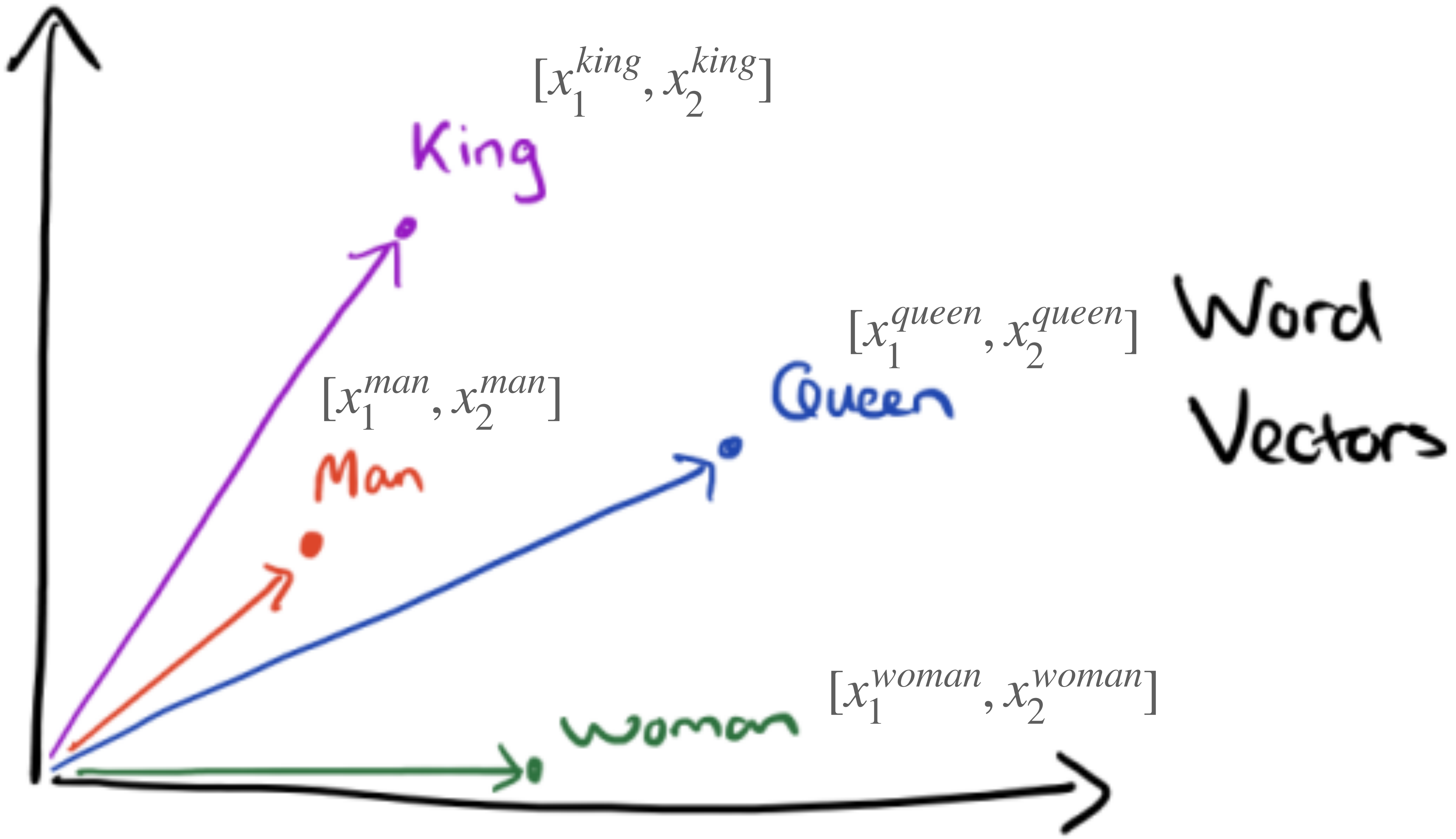


https://upload.wikimedia.org/wikipedia/commons/5/56/RGB_channels_separation.png

Language data

Representing words

- Words can be represented as *n*-dimensional vectors (eg. 2-dimensional vector of $[x_1, x_2]$)
- Vectors preserve “word-to-word” relationships
- $v^{king} - v^{queen} \approx v^{man} - v^{woman}$

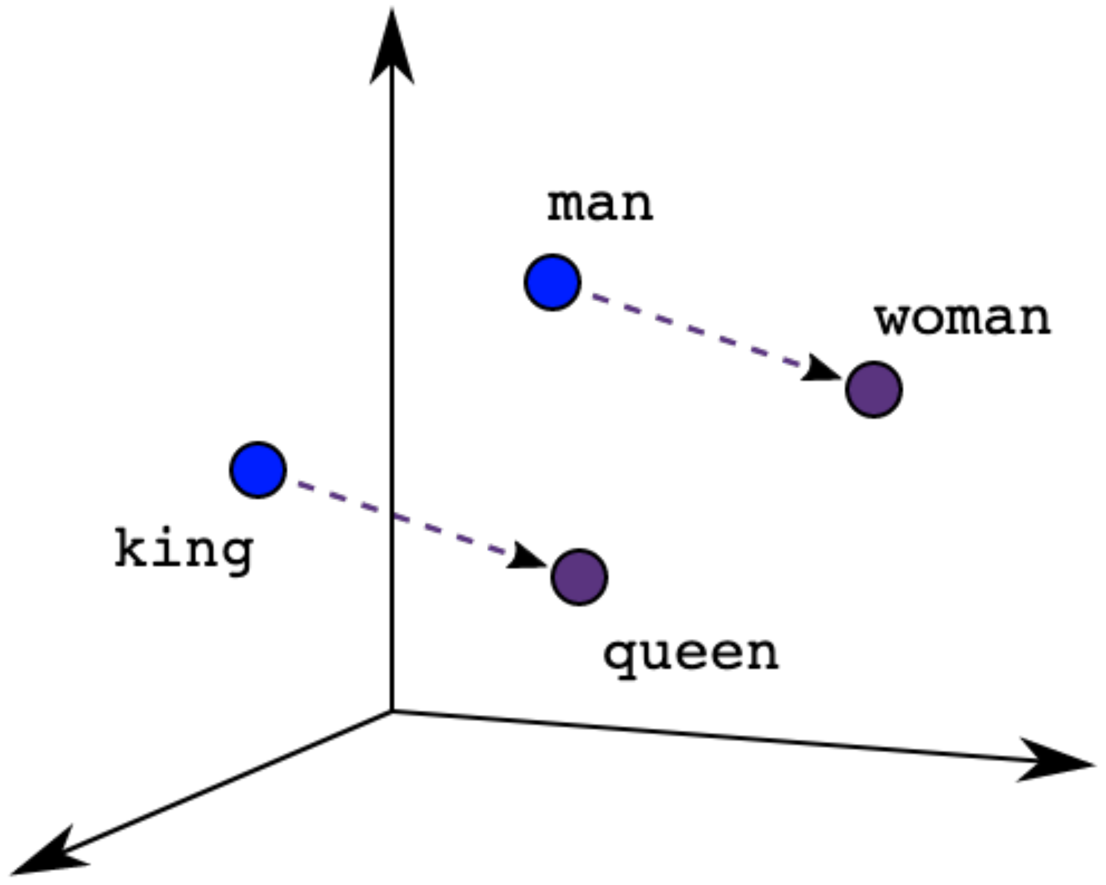


$[x_1^{king}, x_2^{king}]$
King
$[x_1^{queen}, x_2^{queen}]$
Word Vectors
$[x_1^{man}, x_2^{man}]$
Queen
Man
Woman
$[x_1^{woman}, x_2^{woman}]$

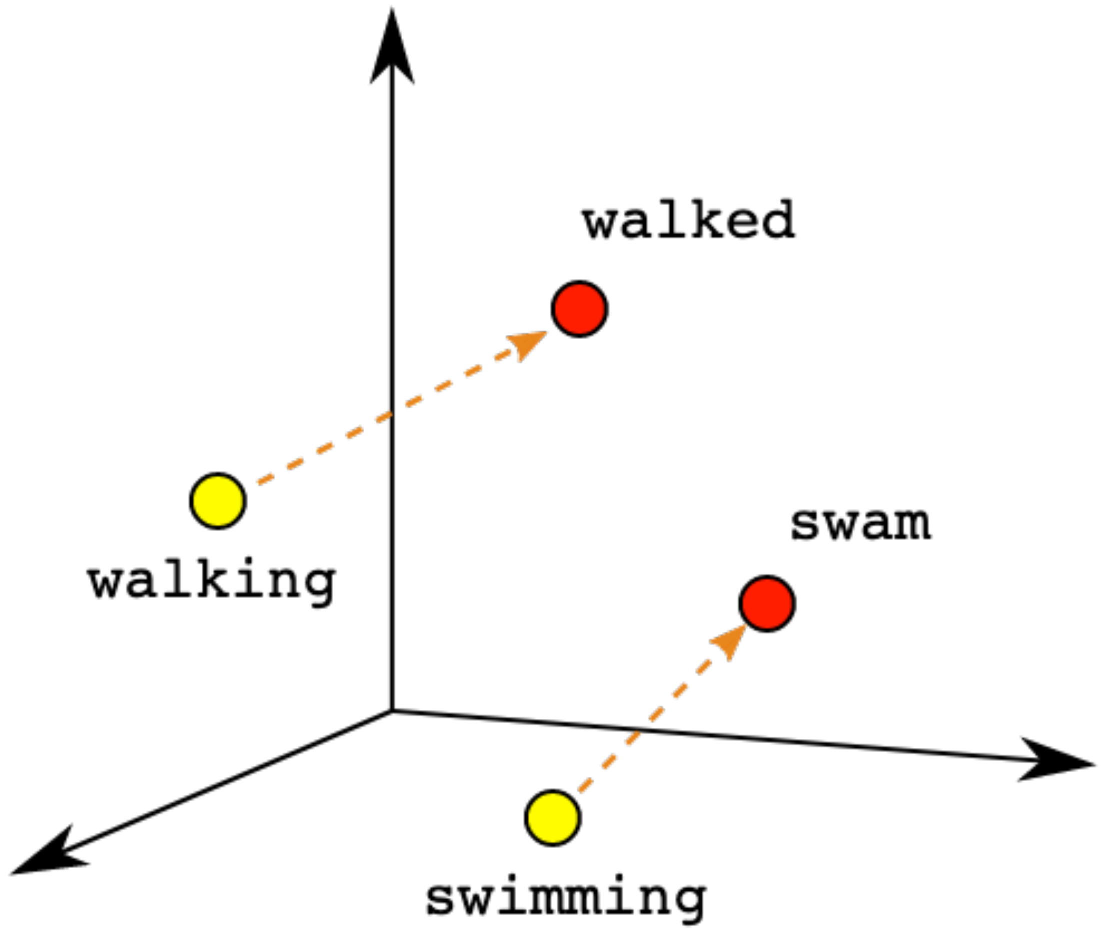
https://www.depends-on-the-definition.com/guide-to-word-vectors-with-gensim-and-keras/

Language data

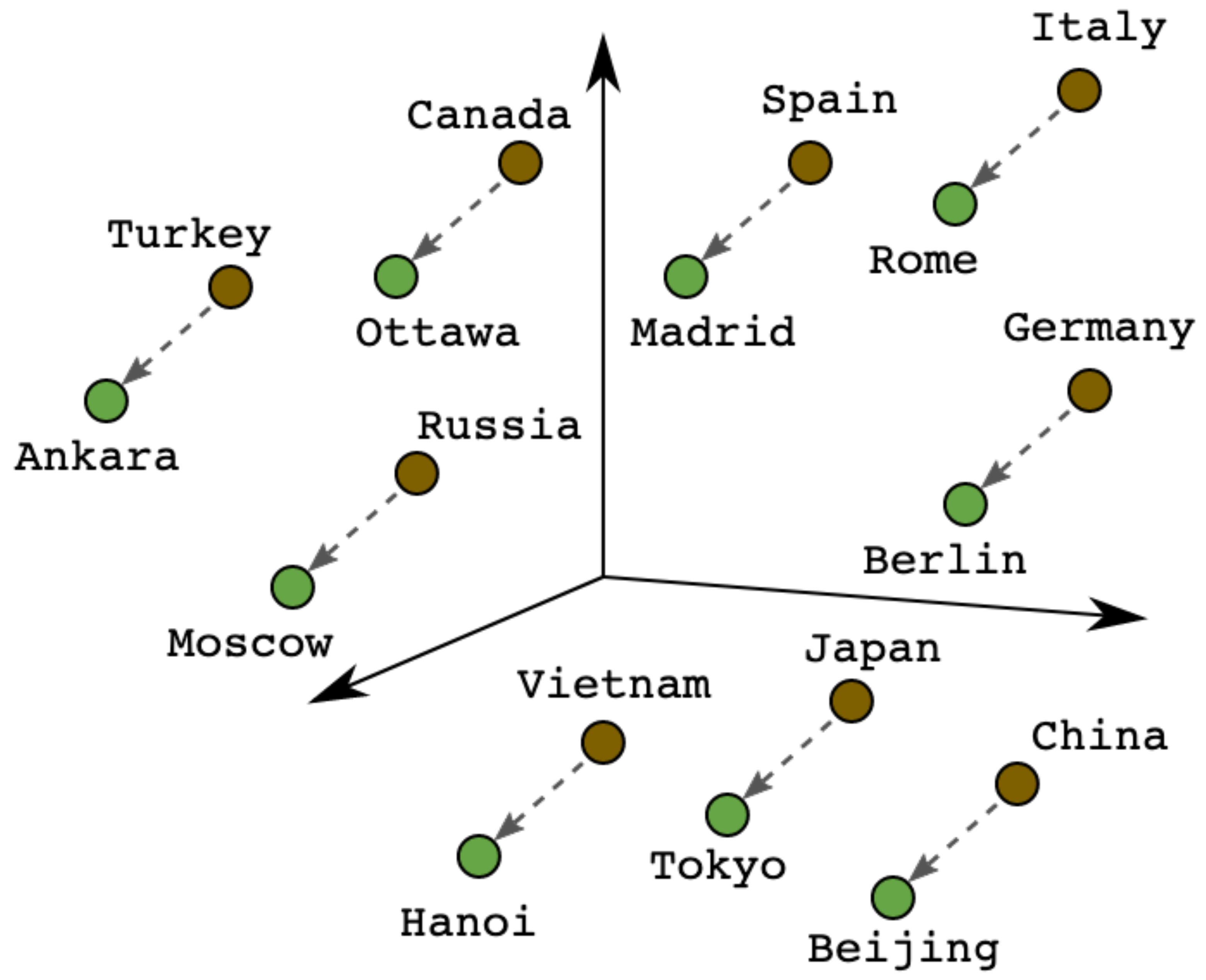
Inter-word relationships



Male-Female



Verb Tense



Country-Capital

https://medium.com/@h_bushroh/text-similarity-with-fasttext-word-embeddings-c765d97df682

Language data

Representing sentences

- Word vectors can be concatenated to represent phrases/sentences
- These concatenated vectors encode the phrase by preserving the context of individual words and the temporal correlation of words
- Extends to other types of sequence data (eg. DNA/RNA, amino acid sequences)

$$\begin{bmatrix}100\\2\\\vdots\\240\end{bmatrix} \quad \begin{bmatrix}20\\804\\\vdots\\102\end{bmatrix} \quad \begin{bmatrix}1\\2\\\vdots\\12\end{bmatrix} \quad \begin{bmatrix}490\\29\\\vdots\\300\end{bmatrix} \begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}$$

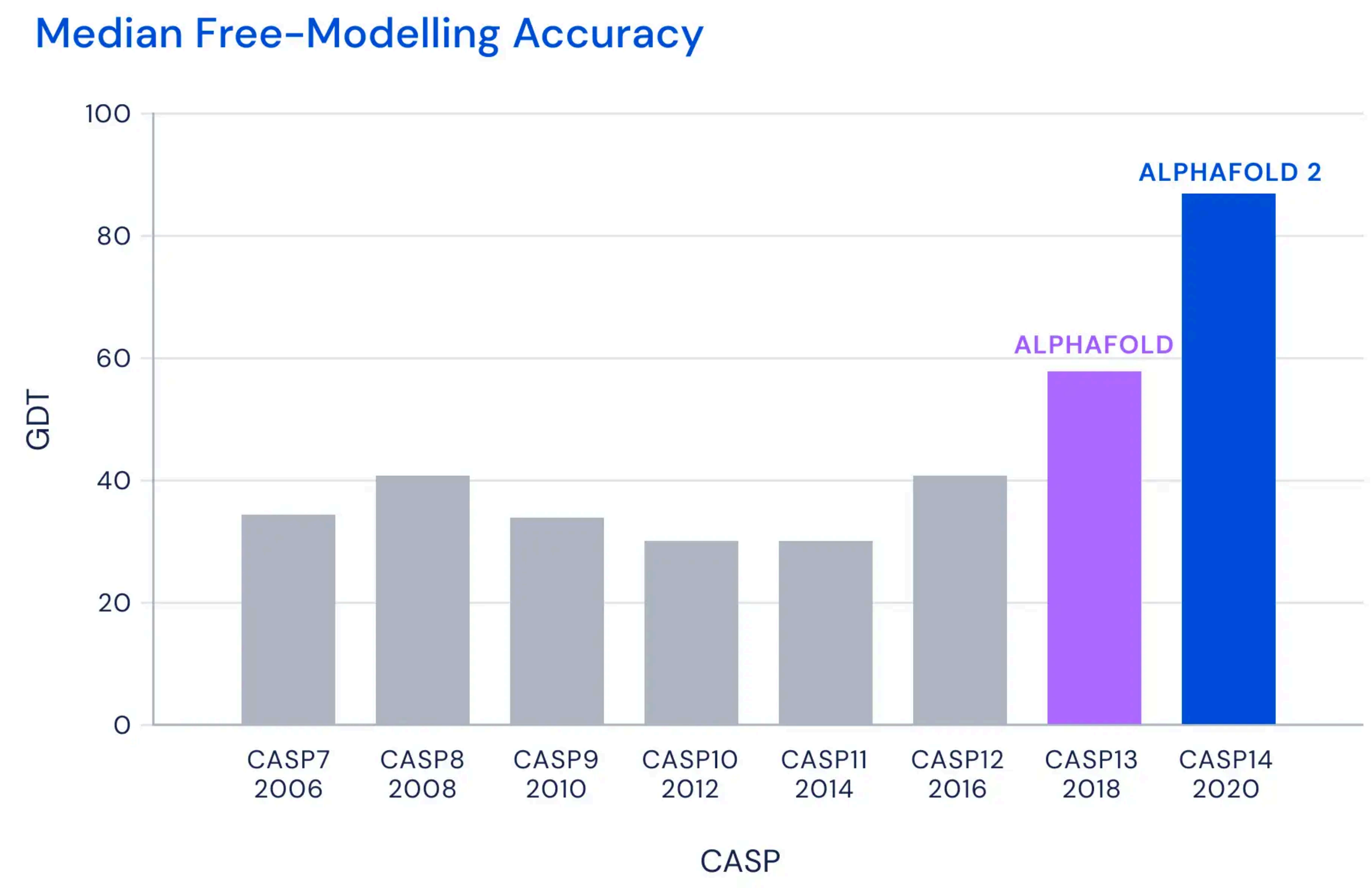
My name is David.

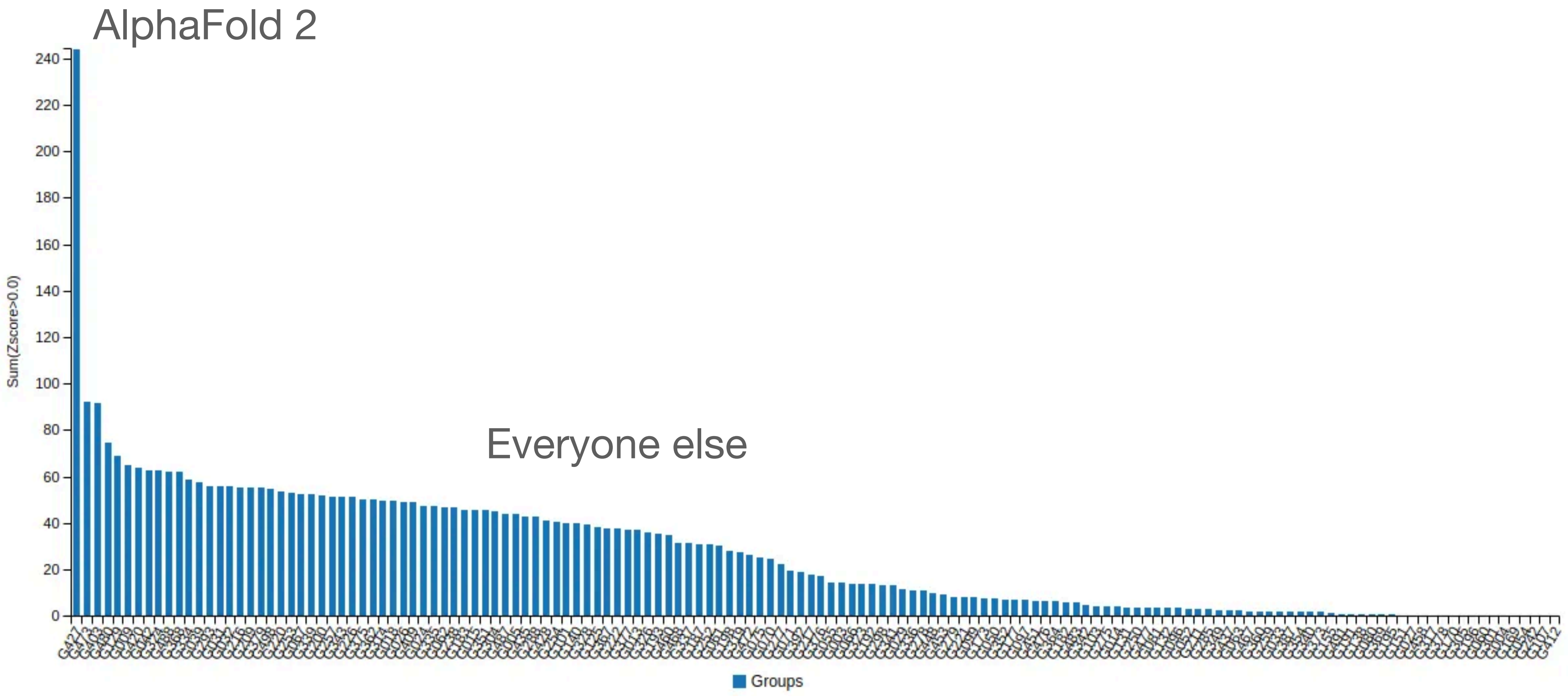
$$\begin{bmatrix}100 & 20 & 1 & 490 & 0\\2 & 804 & 2 & 29 & 0\\\vdots & \vdots & \vdots & \vdots & \vdots\\240 & 102 & 12 & 300 & 0\end{bmatrix}$$

Case Study: The Protein Folding Problem and the AlphaFold System

ML for protein folding

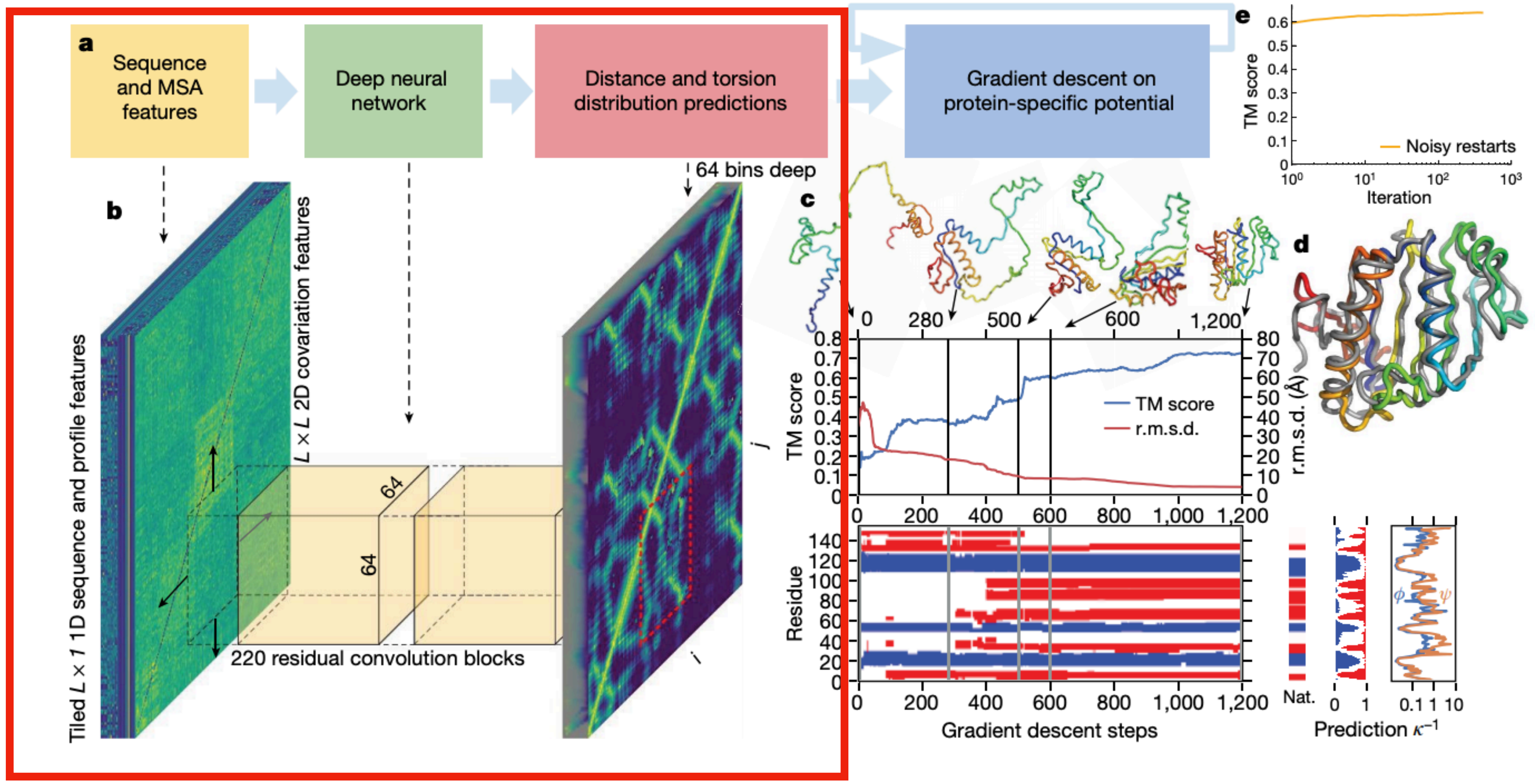
- Critical Assessment of protein Structure Prediction (CASP) bi-annual competition
- Predicting protein structures from amino acid sequences
- In the past two competitions, Google DeepMind used machine learning in their AlphaFold system to great success





AlphaFold 1 System

- The AlphaFold 1 System [2] consists of multiple components
- Specifically, the ML sub-component contains a supervised learning task
  - Input: amino acid sequence
  - Output: a matrix of inter-residue distances and torsion angles


a
Sequence and MSA features
Deep neural network
Distance and torsion distribution predictions
Gradient descent on protein-specific potential
b
Tiled L × 1 1D sequence and profile features
L × L 2D covariation features
64
64
220 residual convolution blocks
64 bins deep
i
j
c
0
280
500
600
1,200
TM score
r.m.s.d. (Å)
TM score
r.m.s.d.
Residue
Gradient descent steps
Nat.
Prediction κ⁻¹
φ
ψ
d
e
TM score
Noisy restarts
Iteration

Machine learning portion of the AlphaFold 1 System

[2] Improved protein structure prediction using potentials from deep learning

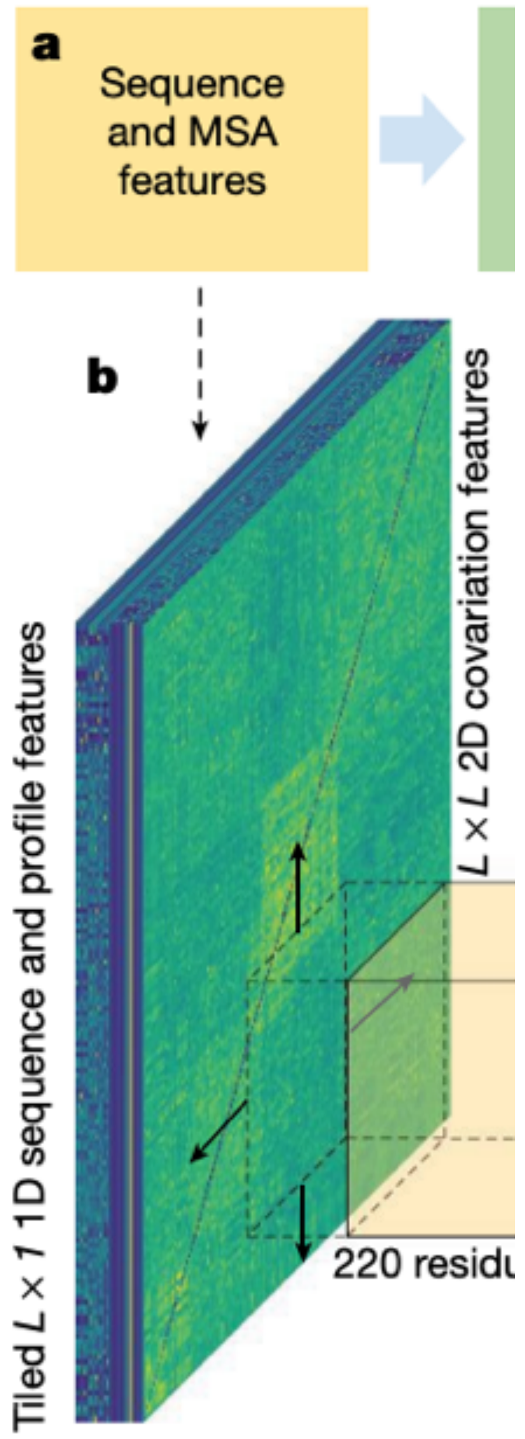
Feature engineering from AA sequence

- Using domain knowledge to create numeric representations of the protein sequence

- These representations contain information that may indicate inter-residue distances

- Eg. from known protein structures, how often does residue A come in contact with residue B

QTKCEKKKCVCENCERSTYL
SERKTMKFNERDSHVVCDKTC


a
Sequence and MSA features
b
Tiled $L \times 1$ 1D sequence and profile features
$L \times L$ 2D covariation features
220 residu

Unsupervised learning of AA sequences

- Multiple sequence alignment [3] is an algorithm that uses many sample sequences of related proteins to infer residue contact
- Residue pairs that are consistent across sequences indicate that those residues may be in close contact (evolutionary covariation)
- Conversely, residues pairs that are uncorrelated are unlikely to be in contact

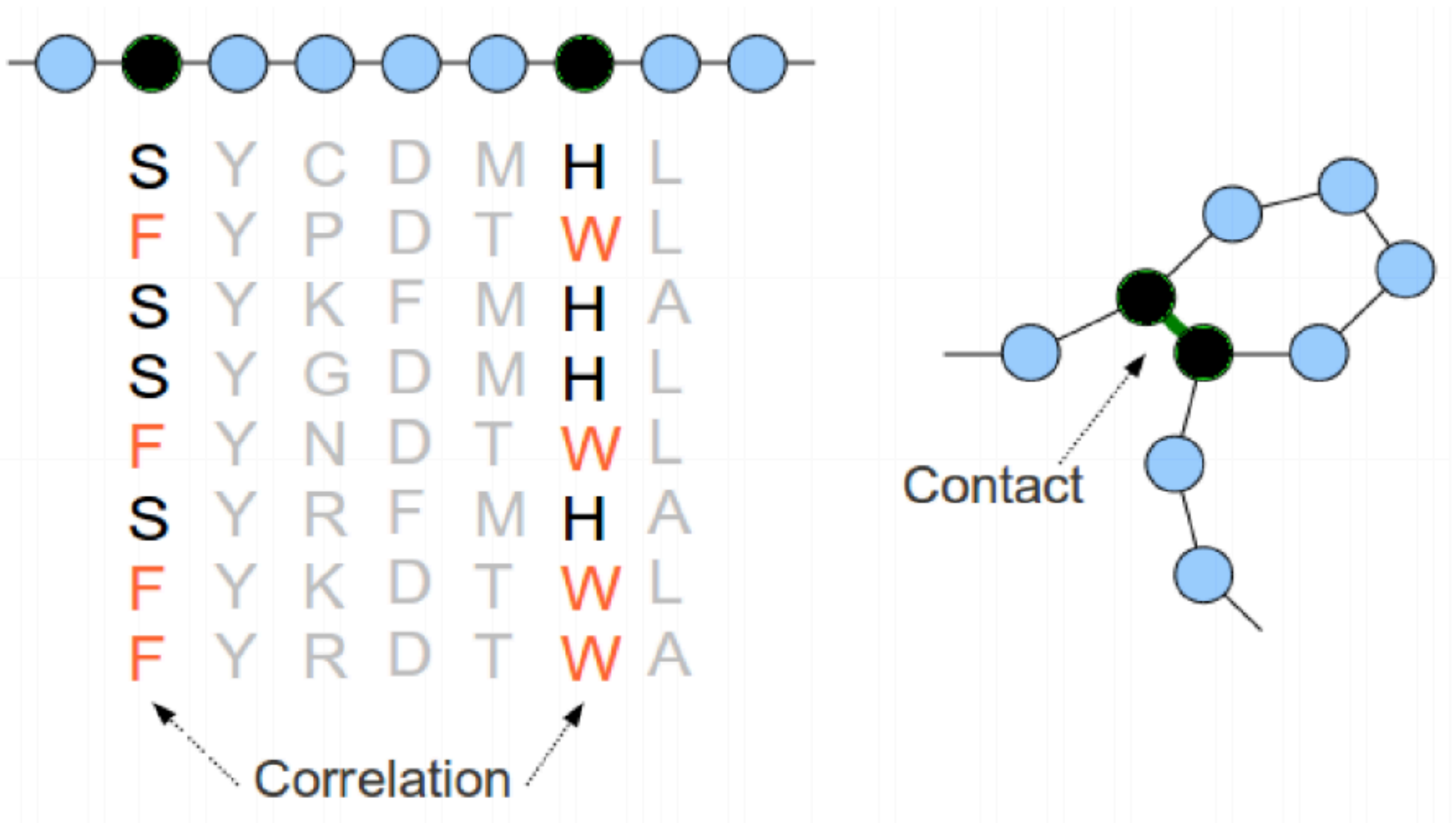
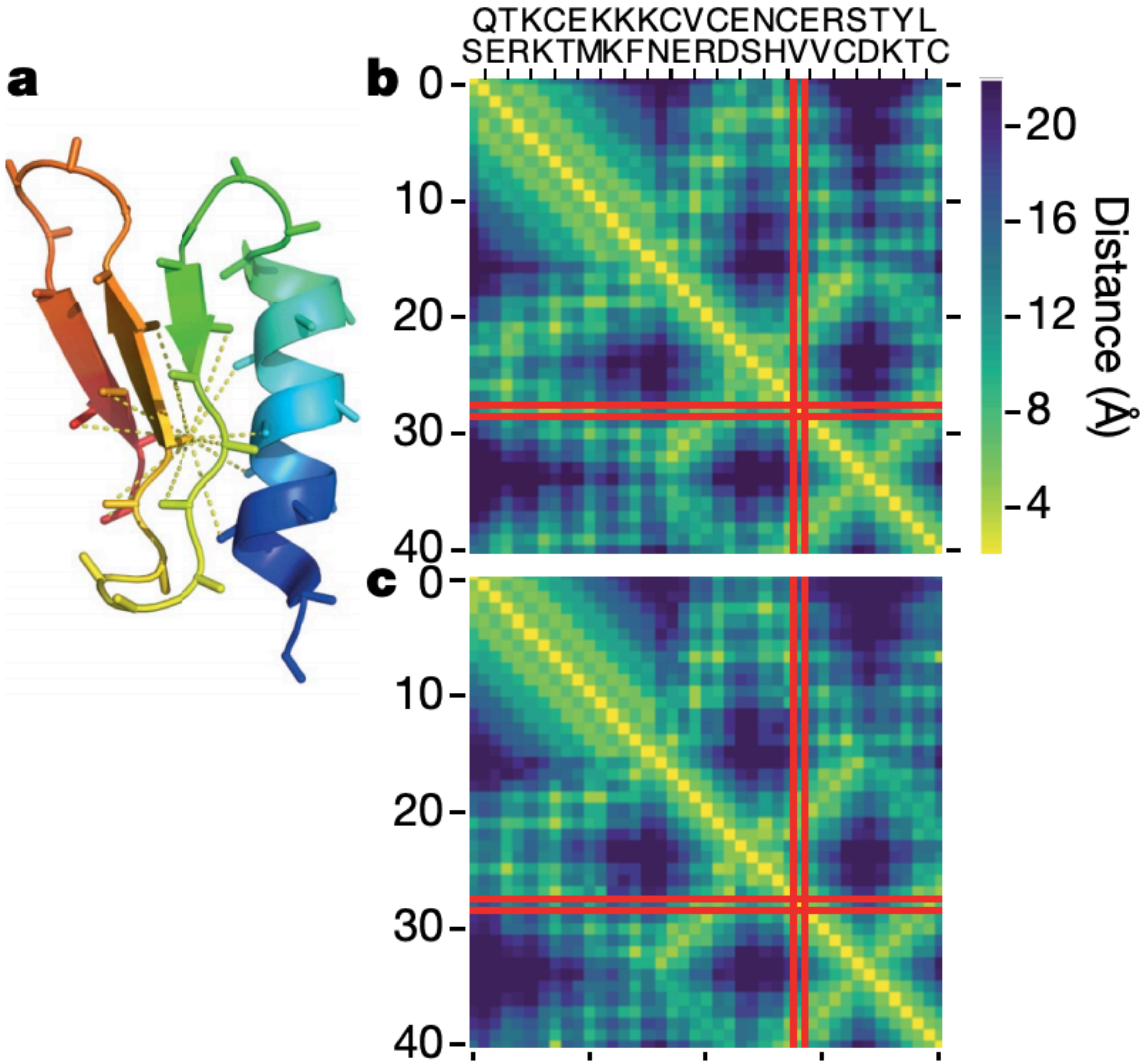


FIG. 1. (Color online) Left panel: small MSA with two positions of correlated amino-acid occupancy. Right panel: hypothetical corresponding spatial conformation, bringing the two correlated positions into direct contact.

[3] Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models

Supervised learning for predicting inter-residue distance

- Use databases of protein sequences with known structures (ie. known inter-residue distances)
- Build a supervised learning model that learns the relationship between amino acid sequences and inter-residue distance
- Predict inter-residue distance for protein sequences with unknown structures



Other examples of ML in basic sciences research

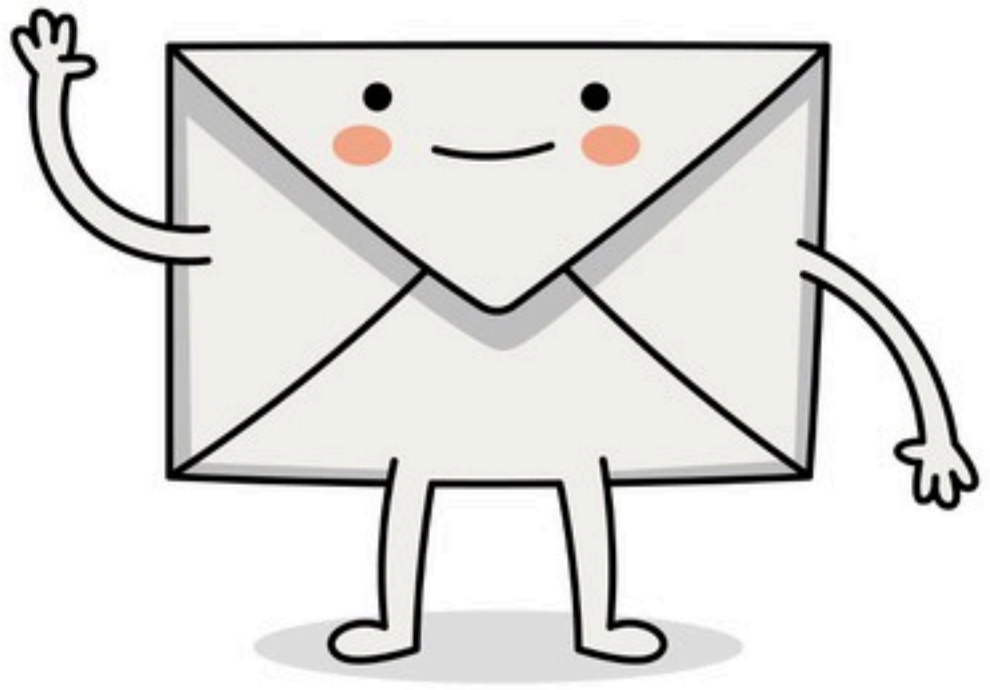
- Genetic engineering attribution: given a sequence of a plasmid, predict the lab that it originated from
- Molecular translation: given a picture of a chemical structure, translate it into its corresponding International Chemical Identifier text string
- A Deep Learning Approach to Antibiotic Discovery - Stokes et. al, Feb 2020, Cell.

Helpful resources for learning ML

- Machine Learning - Stanford (Coursera)
- fast.ai

Get in touch!

- Email: davidwh.dai@gmail.com
- Twitter: @dwhdai
- LinkedIn: https://www.linkedin.com/in/dwhdai/



References

[1] Moore's Law for Everything

[2] Improved protein structure prediction using potentials from deep learning

[3] Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models