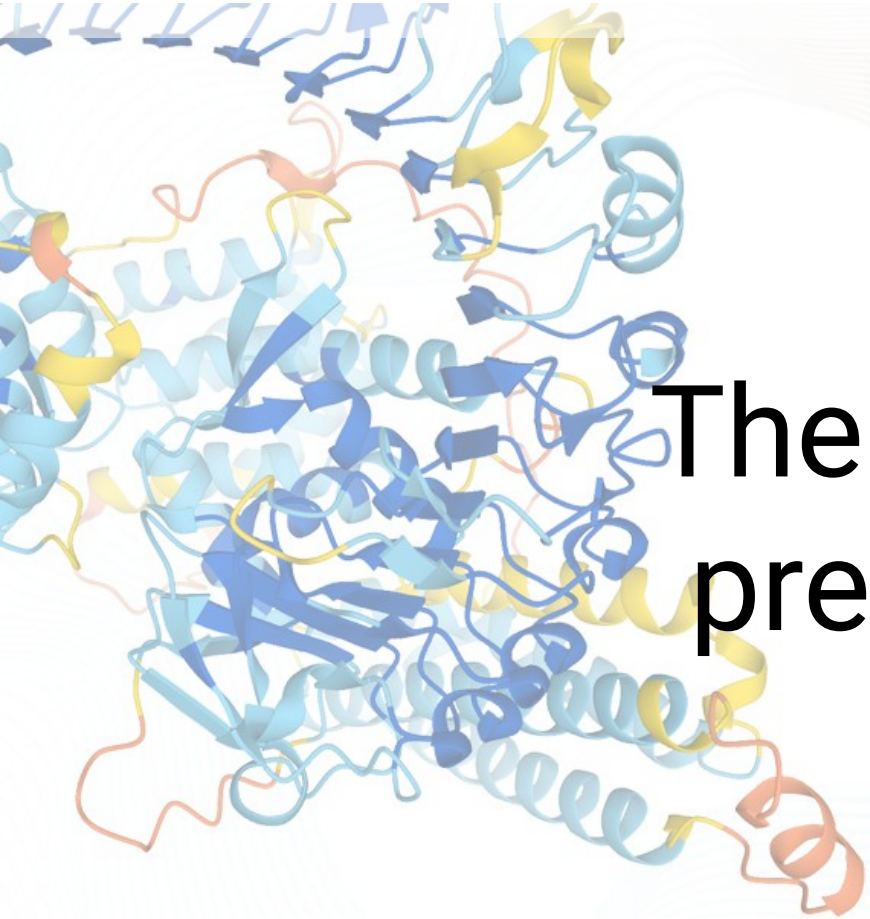
Class core values

1. Be **respect**ful to yourself and others
2. Be **confident** and believe in yourself
3. Always do your **best**
4. Be **cooperative**
5. Be **creative**
6. Have **fun**
7. Be **patient** with yourself while you learn
8. Don’t be shy to **ask “stupid” questions**



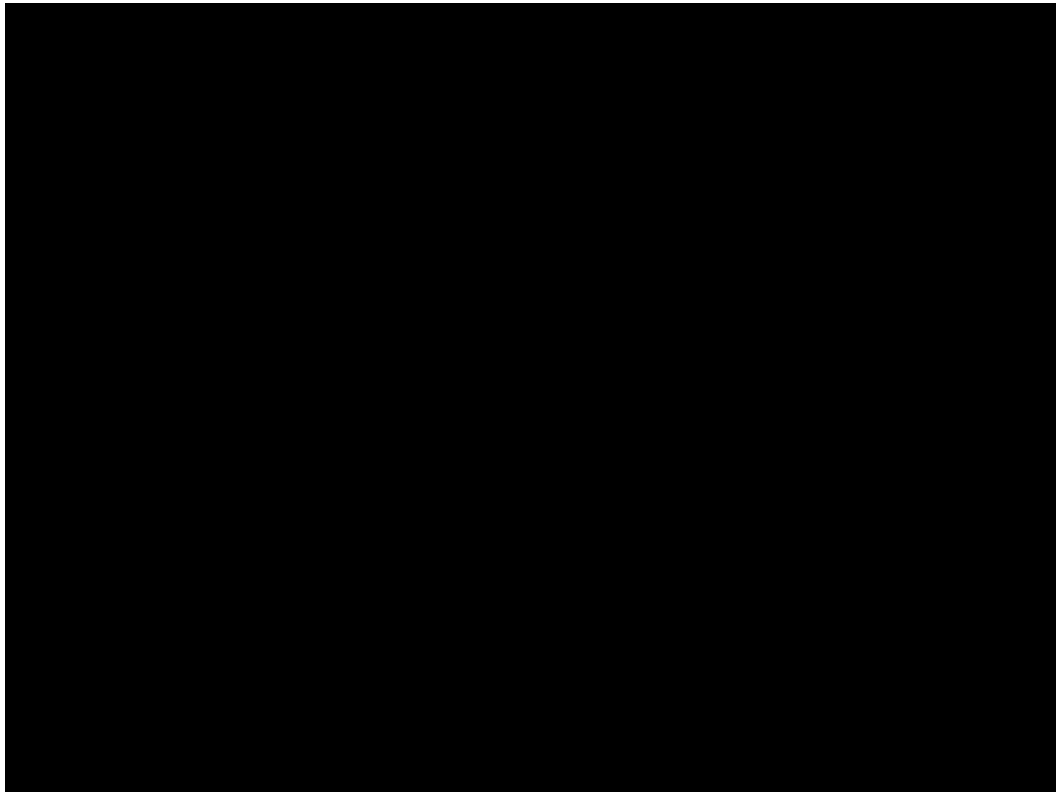
Week 7, Lecture 2

The protein structure prediction challenge

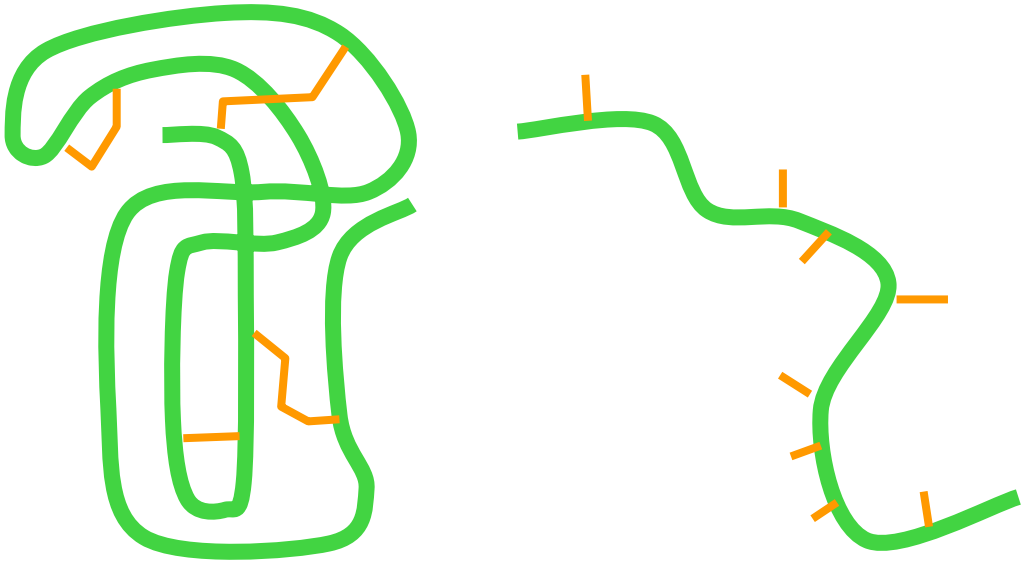
Learning Objectives

1. Describe CASP and protein structure prediction challenge
2. Identify the advances made by machine learning in structure prediction
3. Critically evaluate models generated through structure prediction
4. Identify next challenges in protein structure prediction
5. Describe basics of machine learning modules

The protein structure prediction challenge



The protein structure prediction challenge started with Anfinsen’s experiment



The protein structure prediction challenge started with Anfinsen's experiment

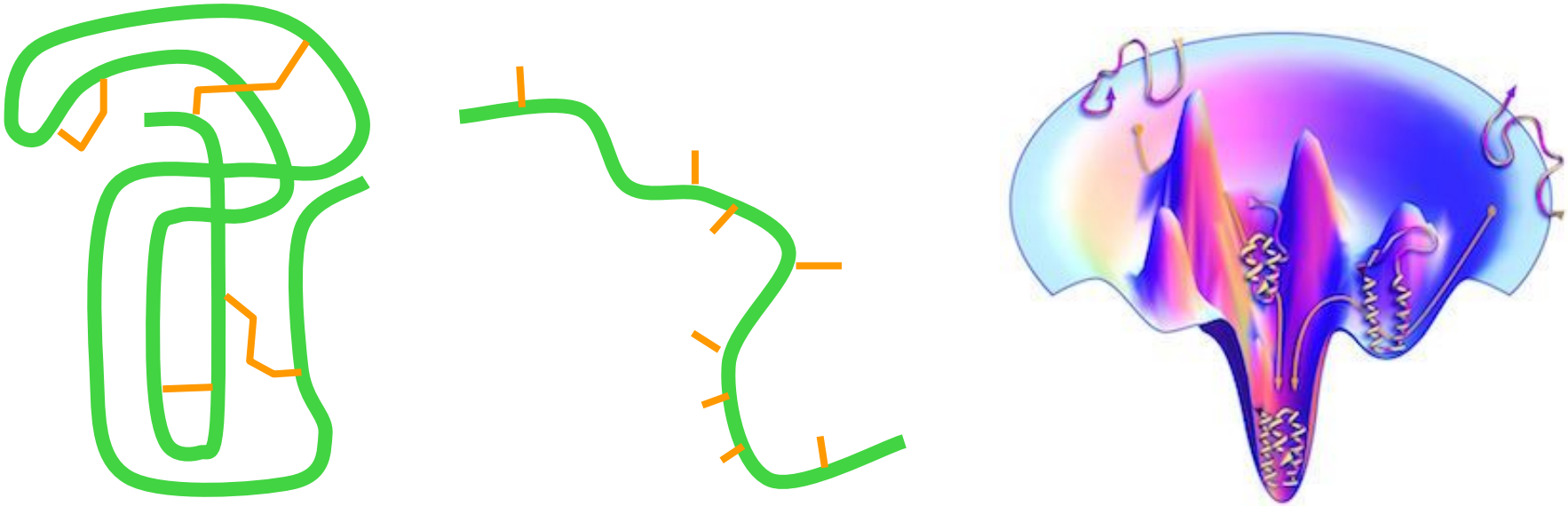


Image from: *Science* **2012** 338 (6110) pp 1042-1046

Protein structure prediction as a computational problem

Describable

Solvable

Tractable

Testable

Non-trivial

Protein structure prediction as a computational problem

Describable

Solvable

Tractable

Testable

Non-trivial

Protein structure prediction as a computational problem

Describable

Solvable

Tractable

Testable

Non-trivial

Protein structure prediction as a computational problem

Describable

Solvable

Tractable

Testable

Non-trivial

Protein structure prediction as a computational problem

Describable

Solvable

Tractable

Testable

Non-trivial

The need for a centralized systematic competition in prediction fields

CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

Biannual experiment to assess protein modeling methods

Key element:

- A Blind experiment on protein structures that are not published

Participants:

- Groups of researchers, have several weeks to predict
- Servers: have 48 hours to predict

CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

Tertiary structure prediction:

1. Comparative modeling based on clear sequence relationship

CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

Tertiary structure prediction:

1. Comparative modeling based on clear sequence relationship
2. Modeling based on more distance evolutionary connections

CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

Tertiary structure prediction:

1. Comparative modeling based on clear sequence relationship
2. Modeling based on more distance evolutionary connections
3. Modeling based on non-homologous fold relationships

CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

Tertiary structure prediction:

1. Comparative modeling based on clear sequence relationship
2. Modeling based on more distance evolutionary connections
3. Modeling based on non-homologous fold relationships
4. Template-free modeling

CASP:

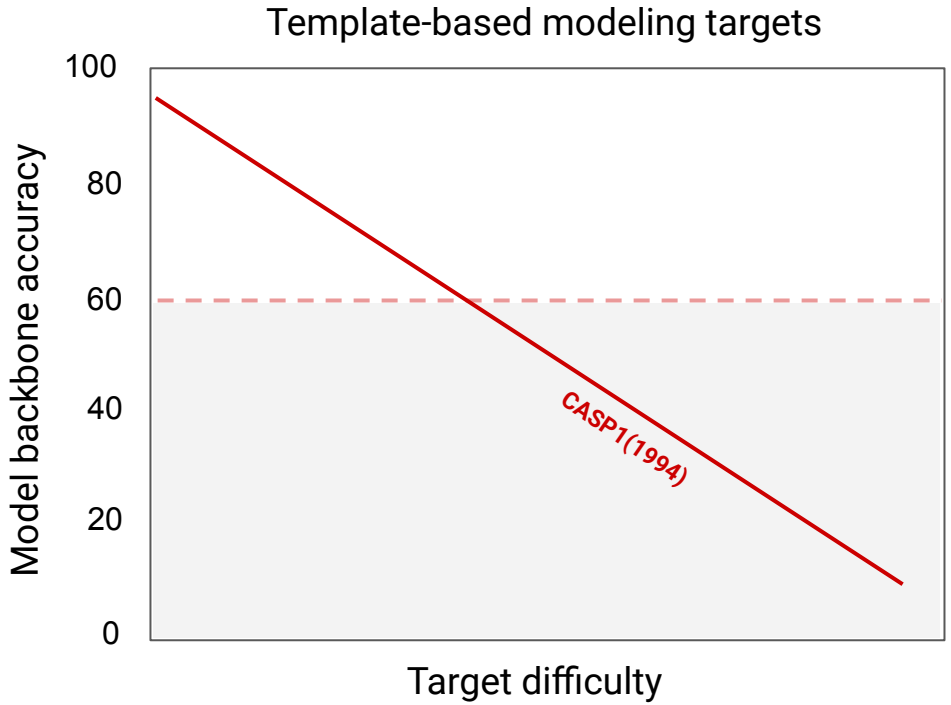
Critical **A**ssessment of protein **S**tructure **P**rediction

Tertiary structure prediction
Secondary structure prediction

1994
CASP1

CASP:

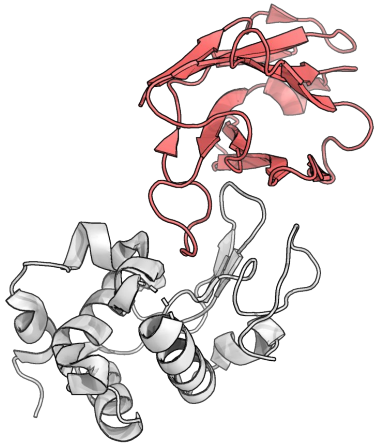
Critical **A**ssessment of protein **S**tructure **P**rediction



CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

Protein complexes → CAPRI

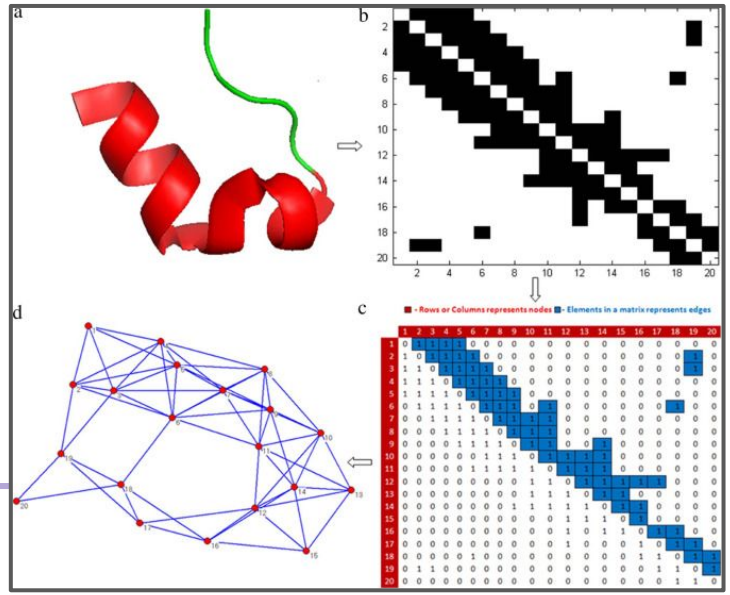


1994
CASP1

CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

Residue-residue contact prediction

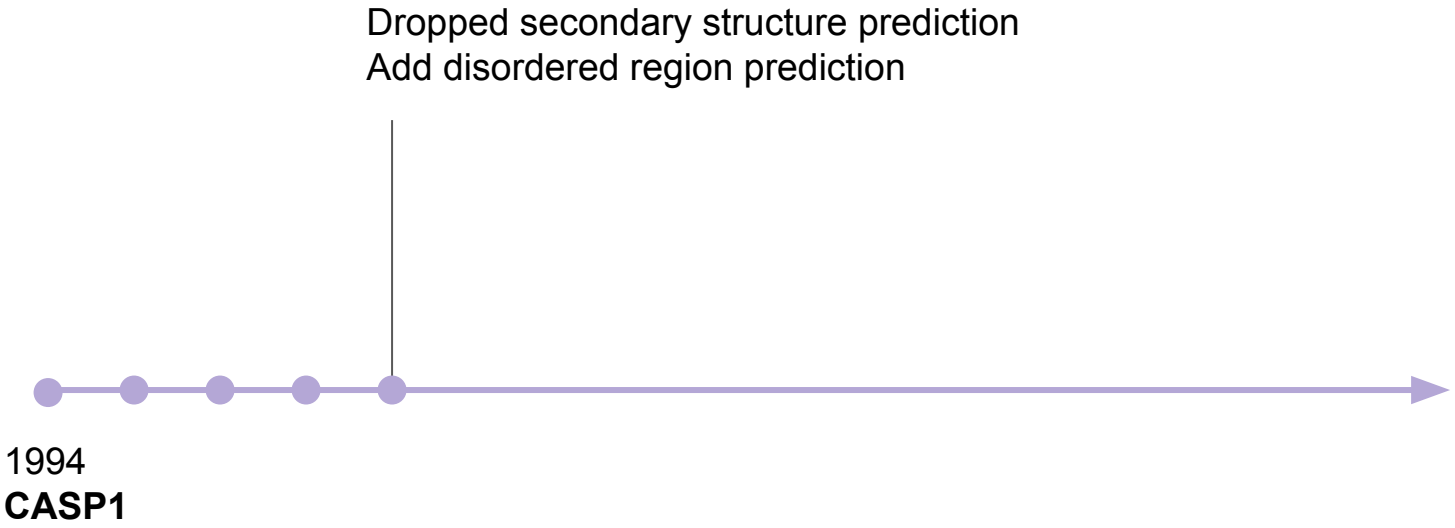


1994
CASP1

Image from: *Physica A* **2017** 465, 782-791

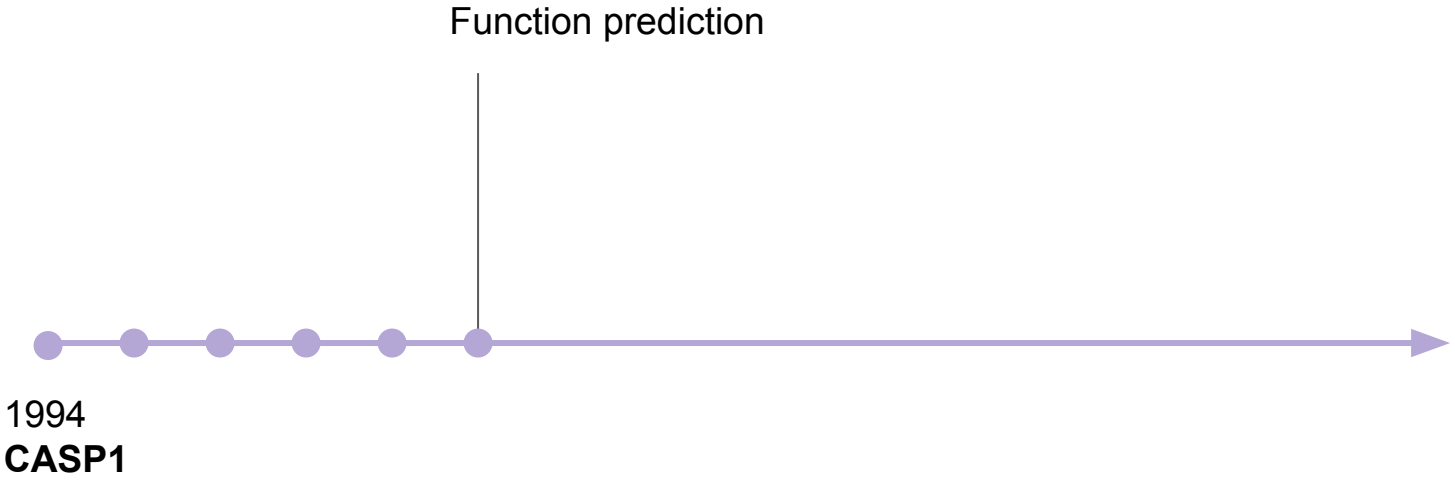
CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction



CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

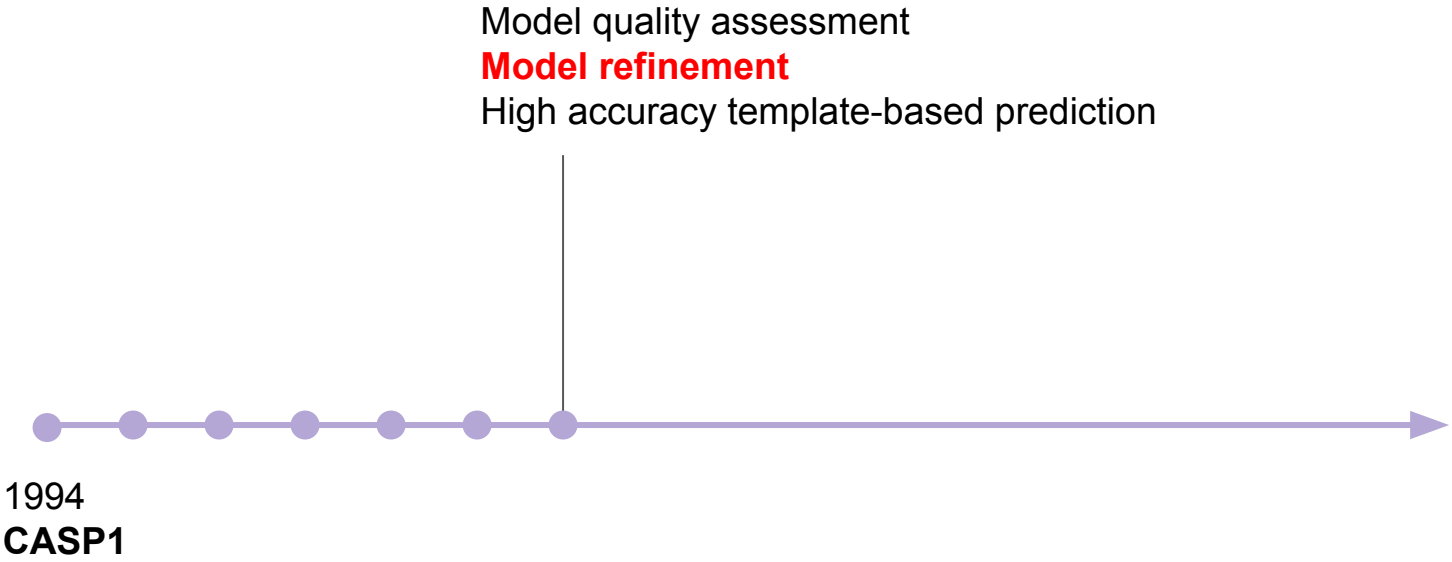


CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

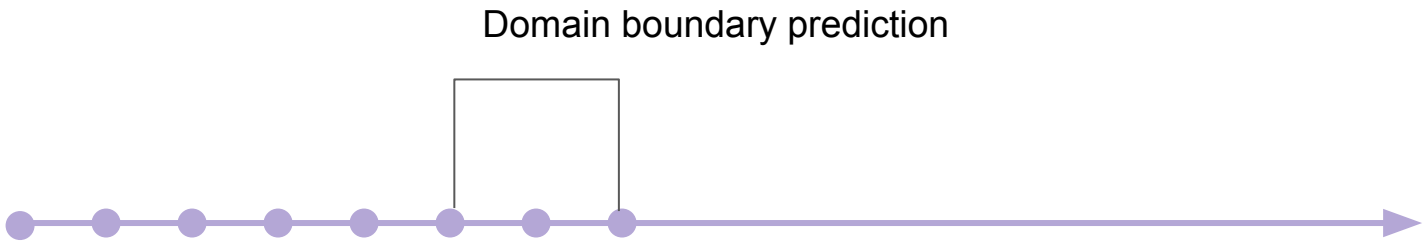
CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction



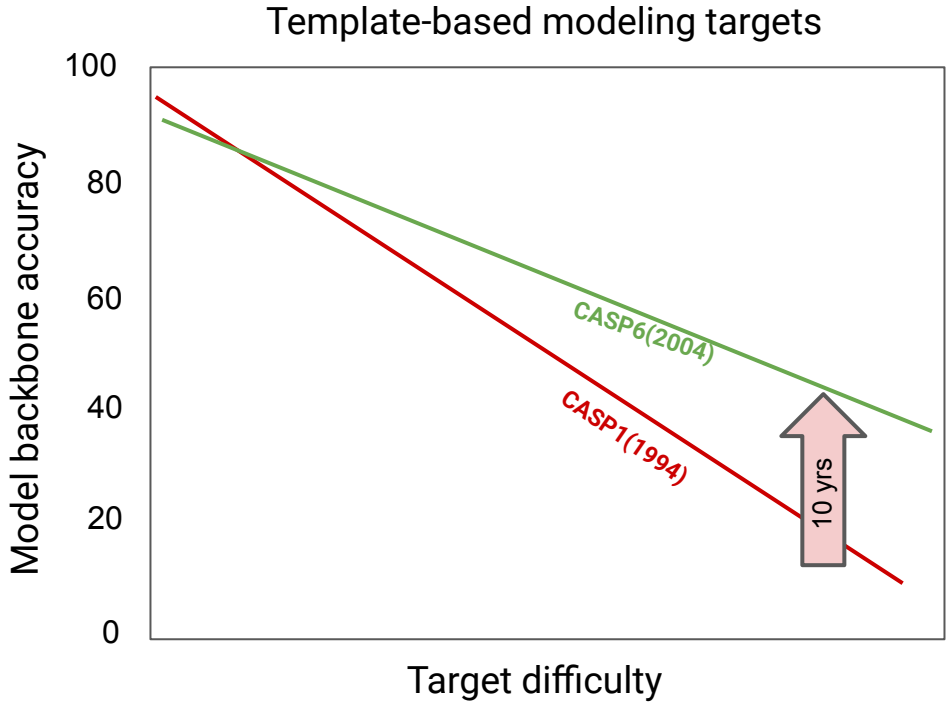
CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction



CASP:

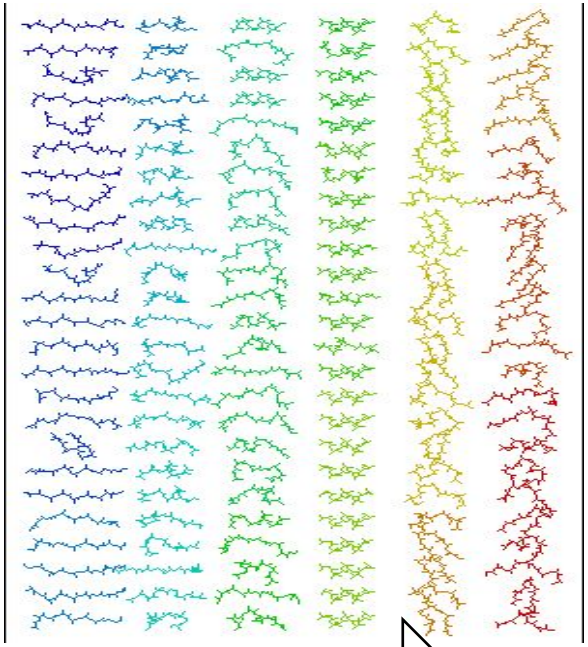
Critical **A**ssessment of protein **S**tructure **P**rediction



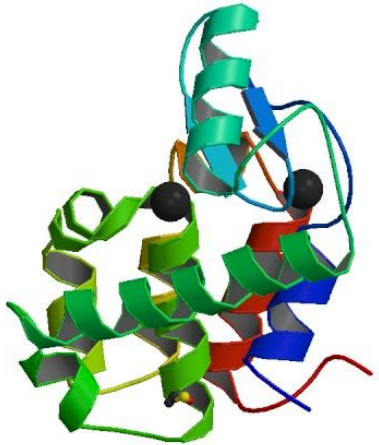
Fragment assembly proved to be one of the most successful methods for structure prediction

MNIFEMLRIDEGLRLKIYKDTE
GYYTIGIGHLLTKSPSLNASKS
ELDKAIGRNTNGVITKDEAEKL
FNQDVDAAVRGILRNAKLKPVY
DSLDAVRRAALINMVFQMGETG
VAGFTNSLRMLQQKRWDEAAVN
LAKSRWYNQTPNRAKRVITTFR
TGTWDAYKNL

Primary Sequence



***de novo* folding**



Tertiary Structure

PDB doesn’t cover the entire potential fragments

- For a 9 residue fragment there are **20^9** possible sequence combinations.
- In the PDB there are only **3 million** characters.
- Only **~0.001%** of the sequence space covered by the PDB.

To account for this, we take a look at fragment similarities

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Slides fro Dr. Sergey Ovchinnikov resources

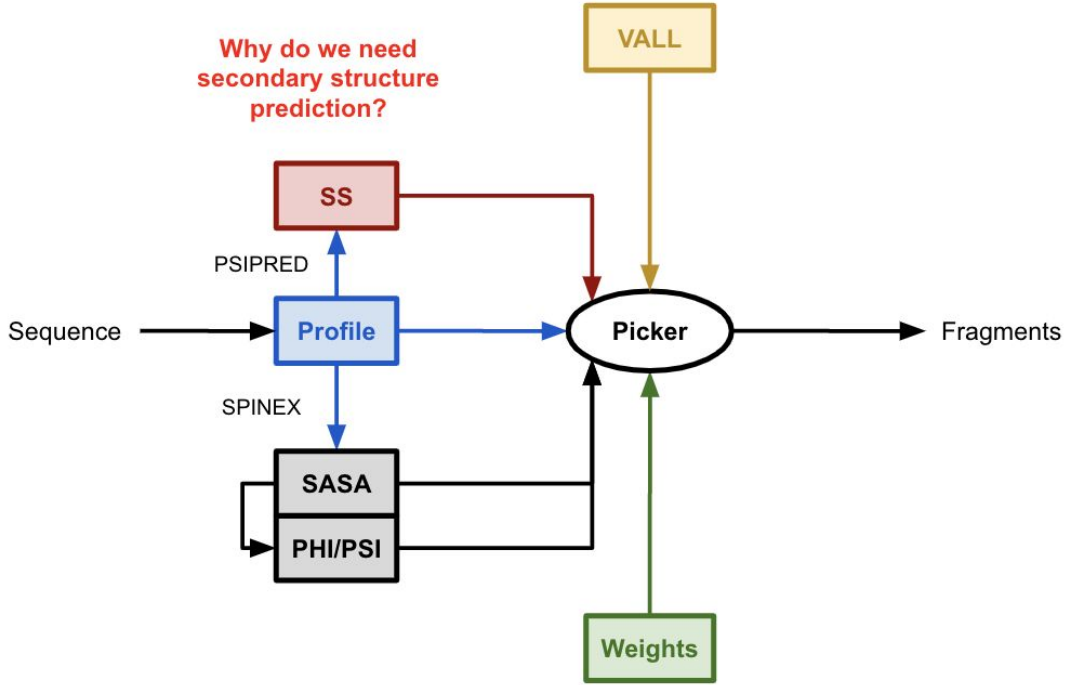
Proc Natl Acad Sci. 1992 Nov 15;89(22):10915-9.

These similarities along with other features are used for picking fragments

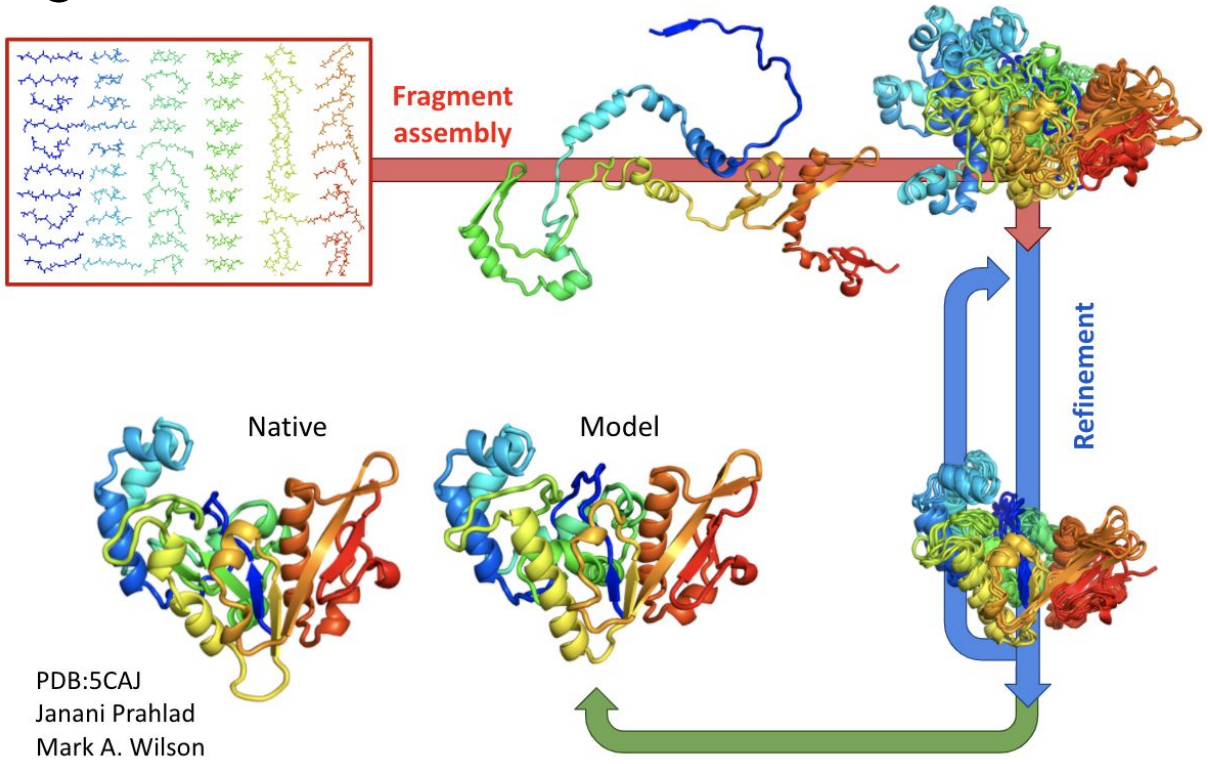
VALL (Very Awesome Looking Loops) = Database of the entire PDB, contains the following:

- PDB, chain, position etc.
- Secondary Structure
- Dihedrals: Phi/Psi/Omega
- Coordinates: Ca/Cb/CEN
- Solvent Accessible Surface Area (SASA)
- Sequence Profiles (PSSM)
  - Constructed using MSA+BLOSUM62
- Structural Profiles (PSSM)
  - Constructed using RMSD+DEPTH matches.

These similarities along with other features are used for picking fragments

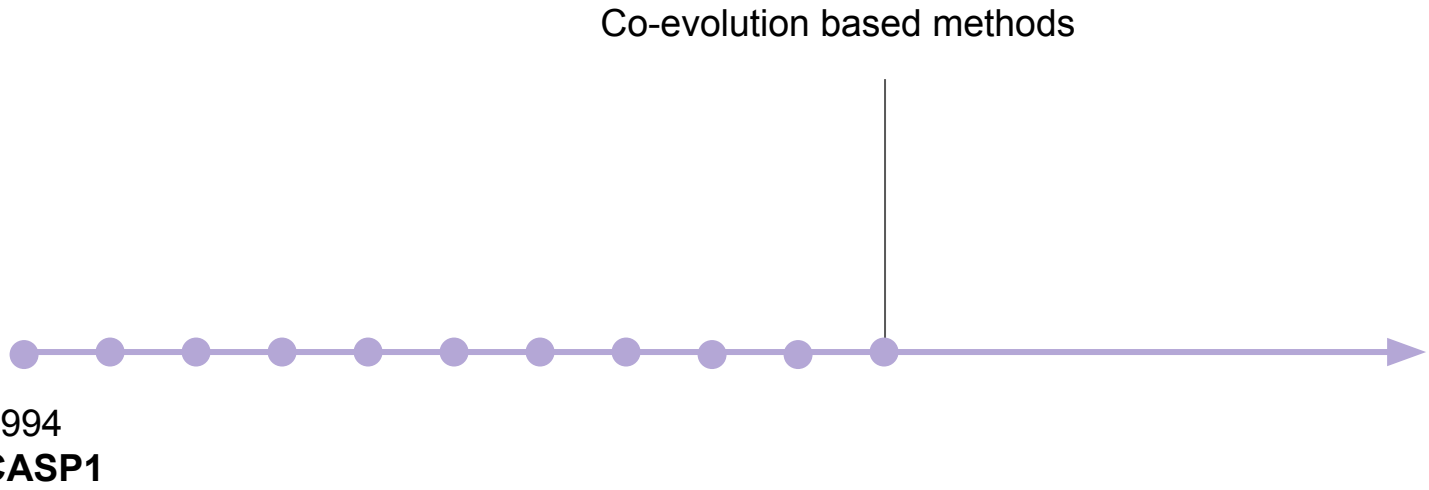


Model refinement focuses on taking the final step of perfecting a structure



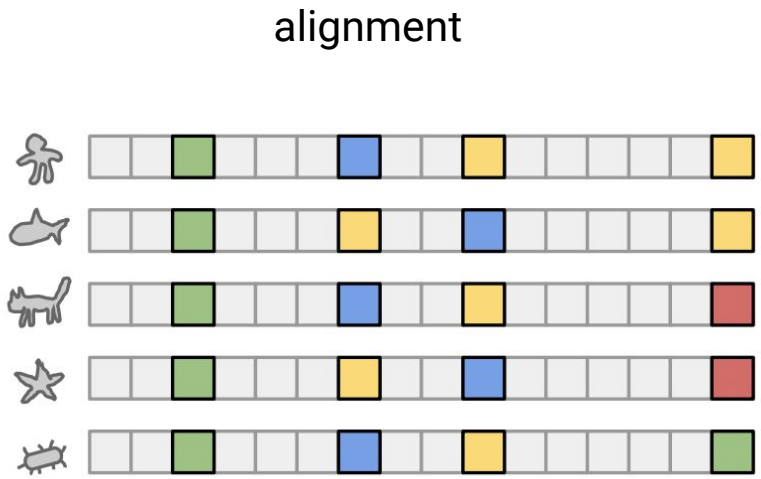
CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction

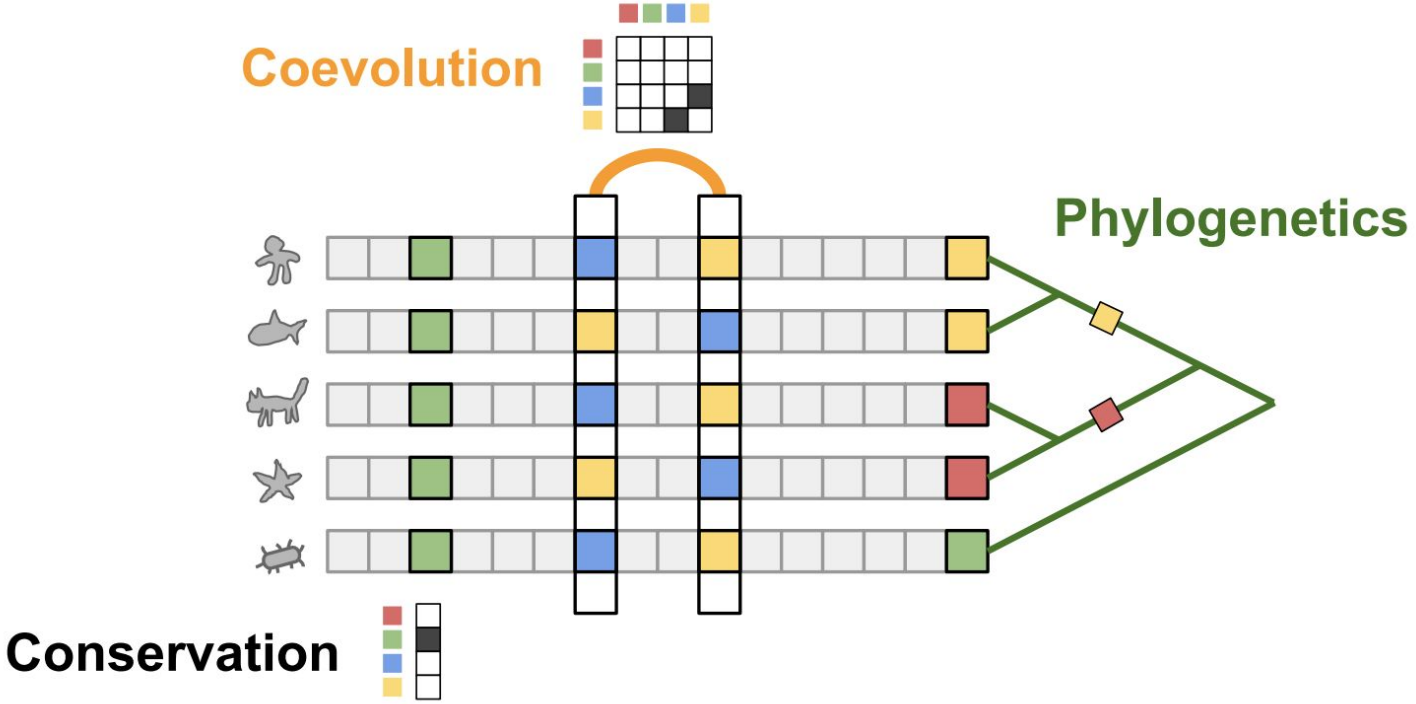


Using coevolution data revolutionized the structure prediction field

Using coevolution data revolutionized the structure prediction field



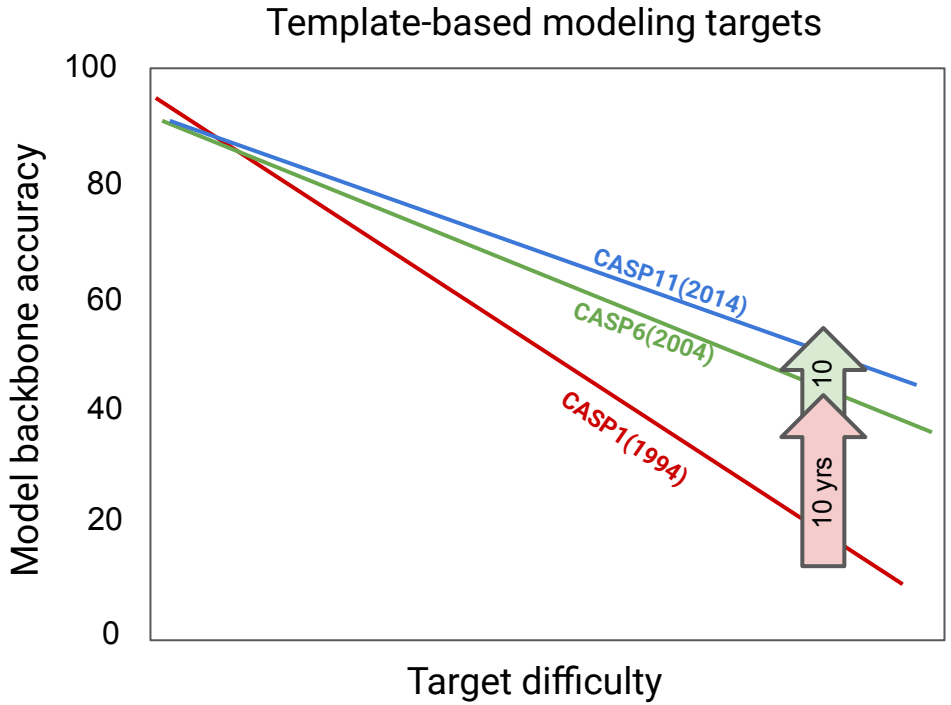
Using coevolution data revolutionized the structure prediction field



Figures by Dr. Sergey Ovchinnikov

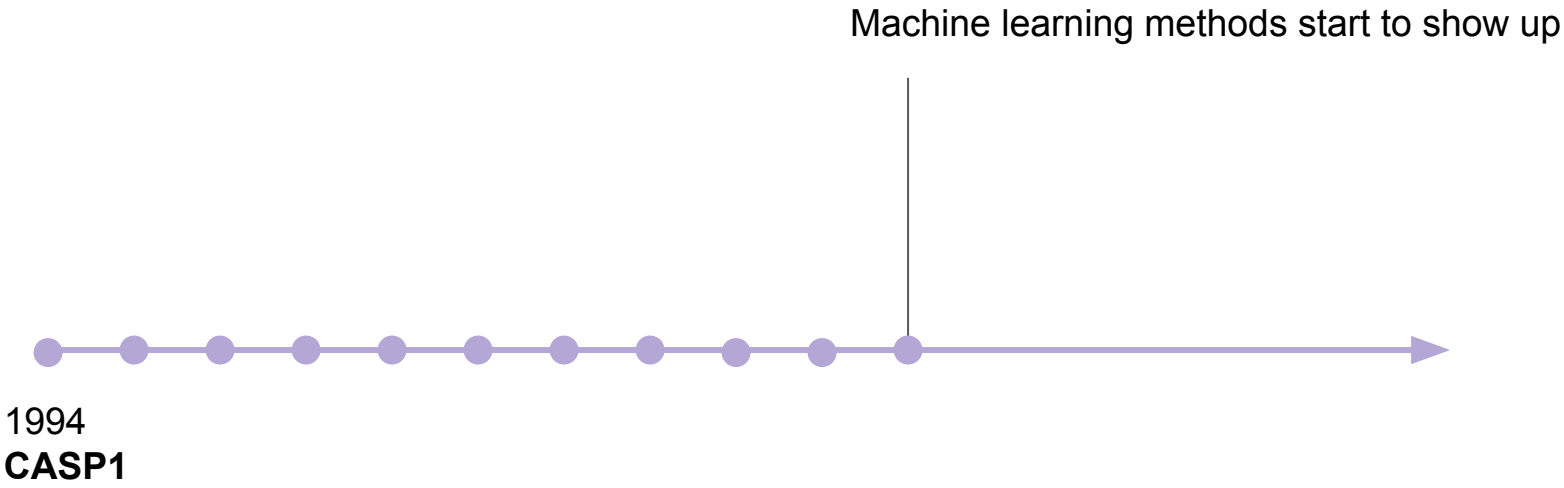
CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction



CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction



A gentle introduction to machine learning

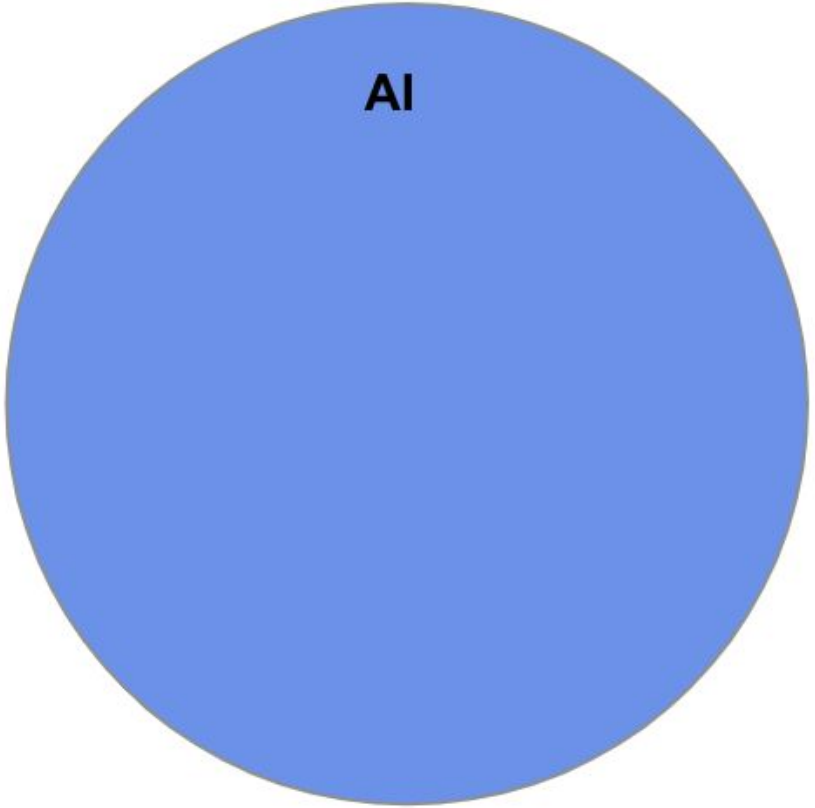
Machine learning for Protein Prediction and Design

Spring 2022
BioE410/510

parisah ‘at’ uoregon.edu

A gentle introduction to machine learning

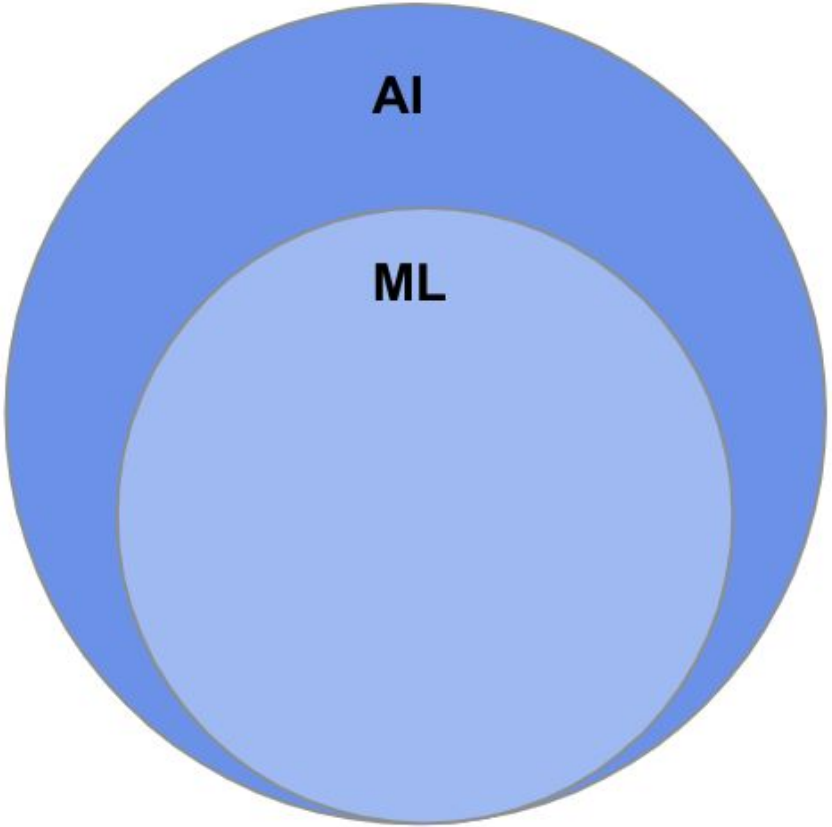
Artificial Intelligence



A gentle introduction to machine learning

Artificial Intelligence

Machine Learning

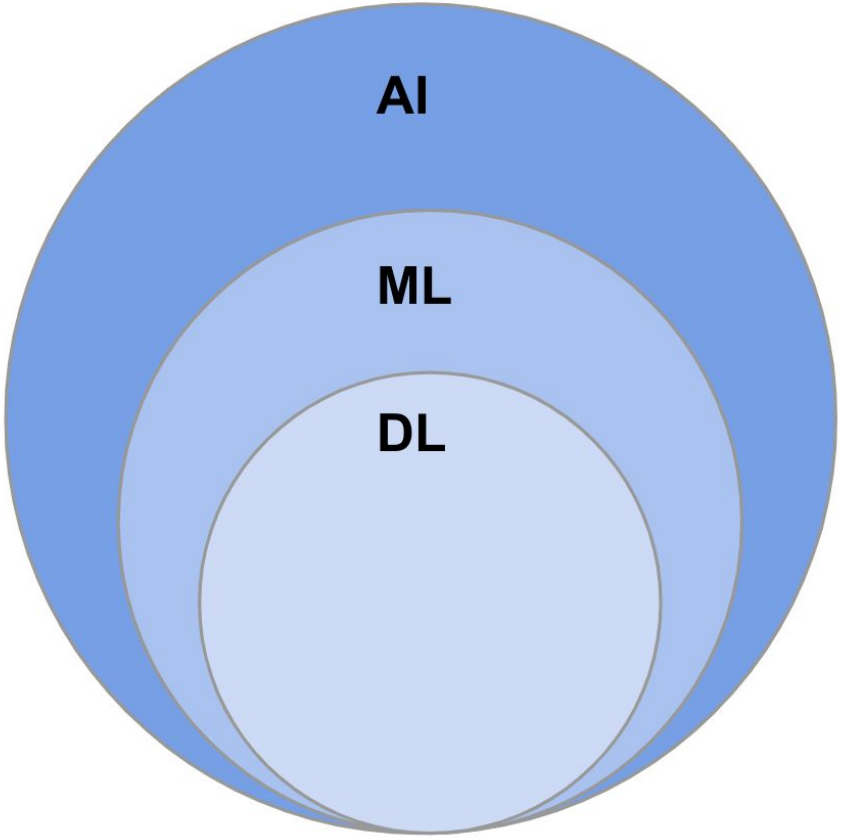


A gentle introduction to machine learning

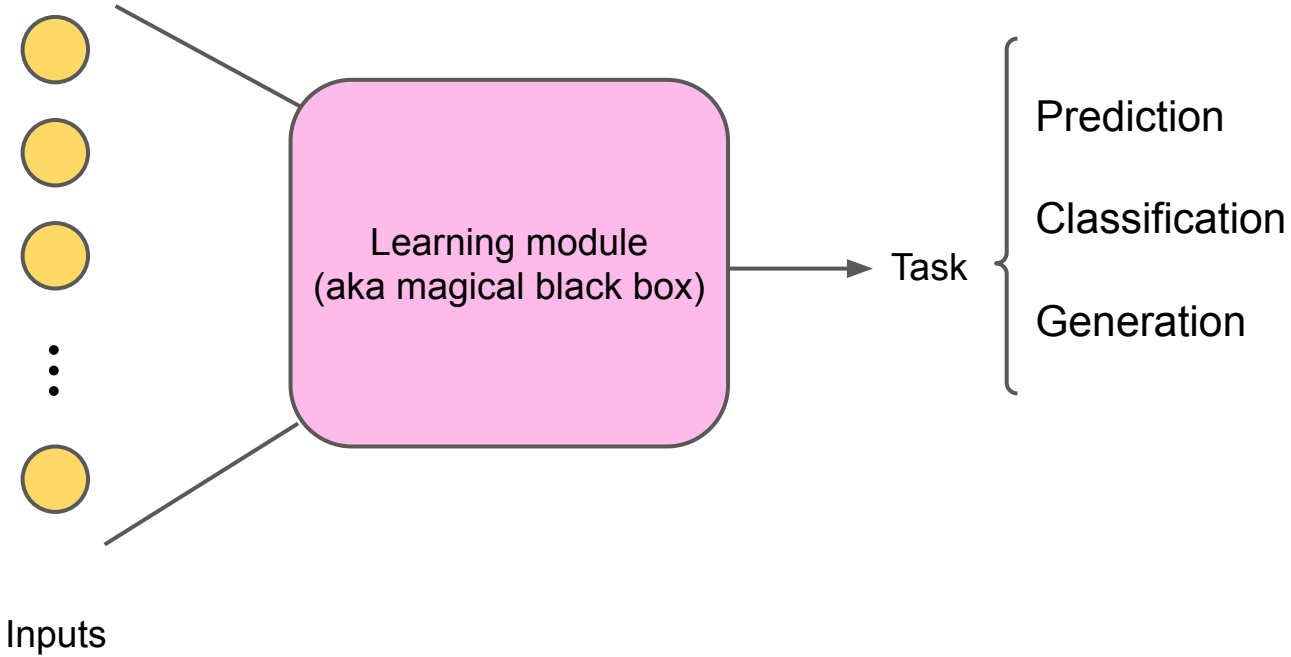
Artificial Intelligence

Machine Learning

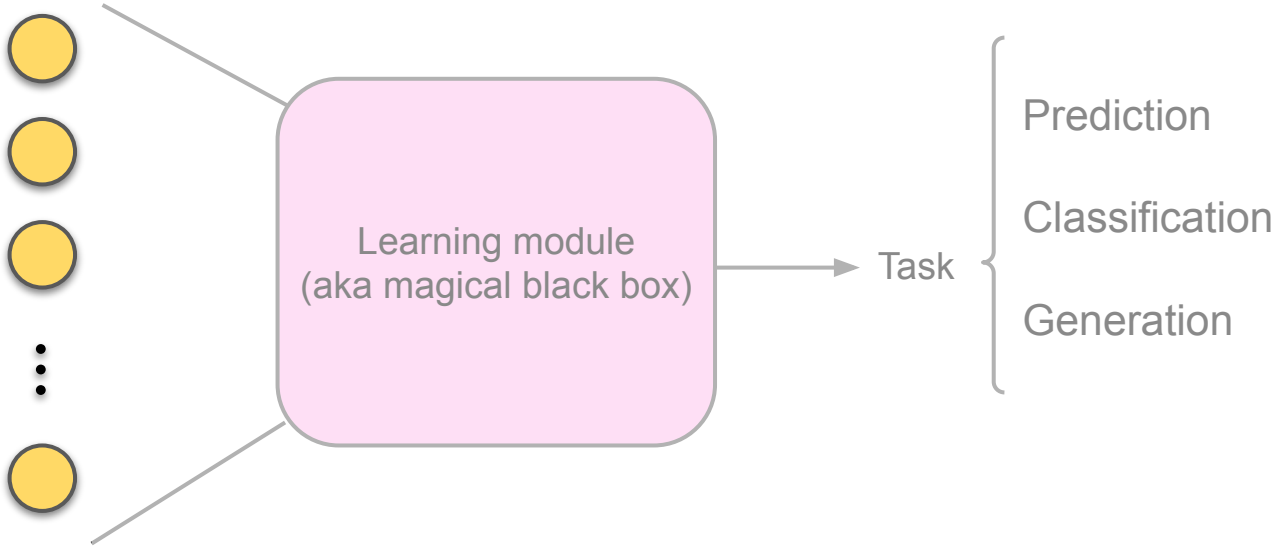
Deep learning



The basic components of a learning system

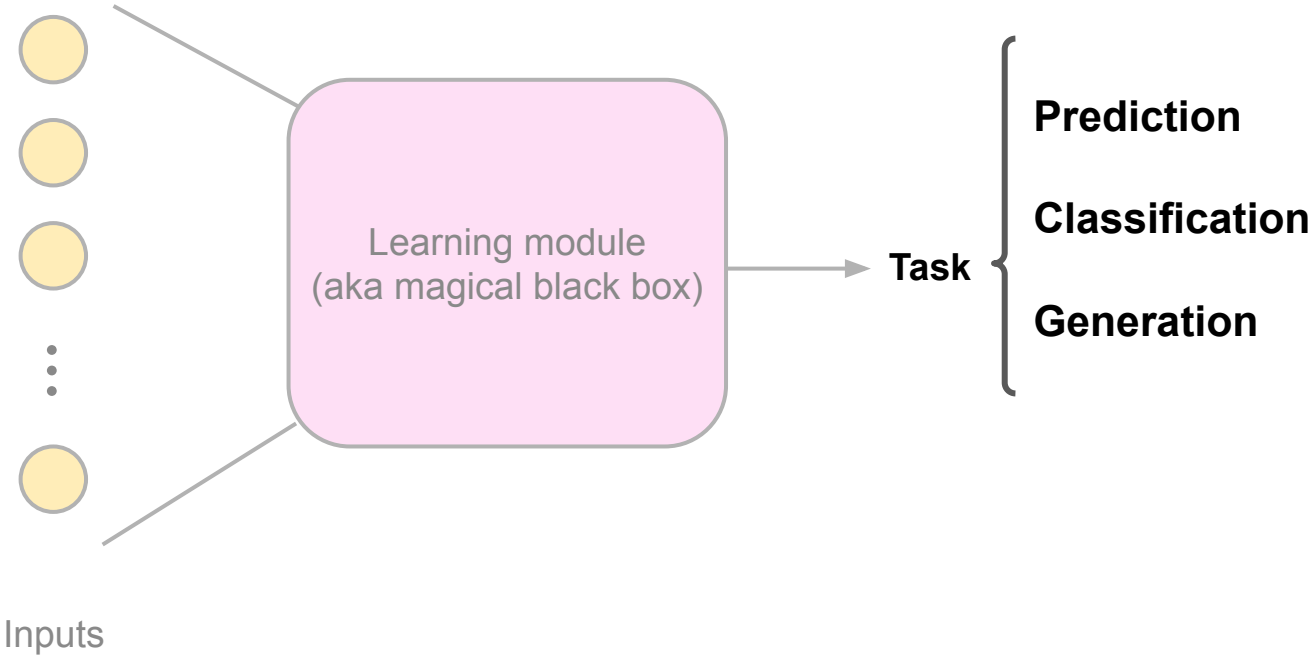


The basic components of a learning system

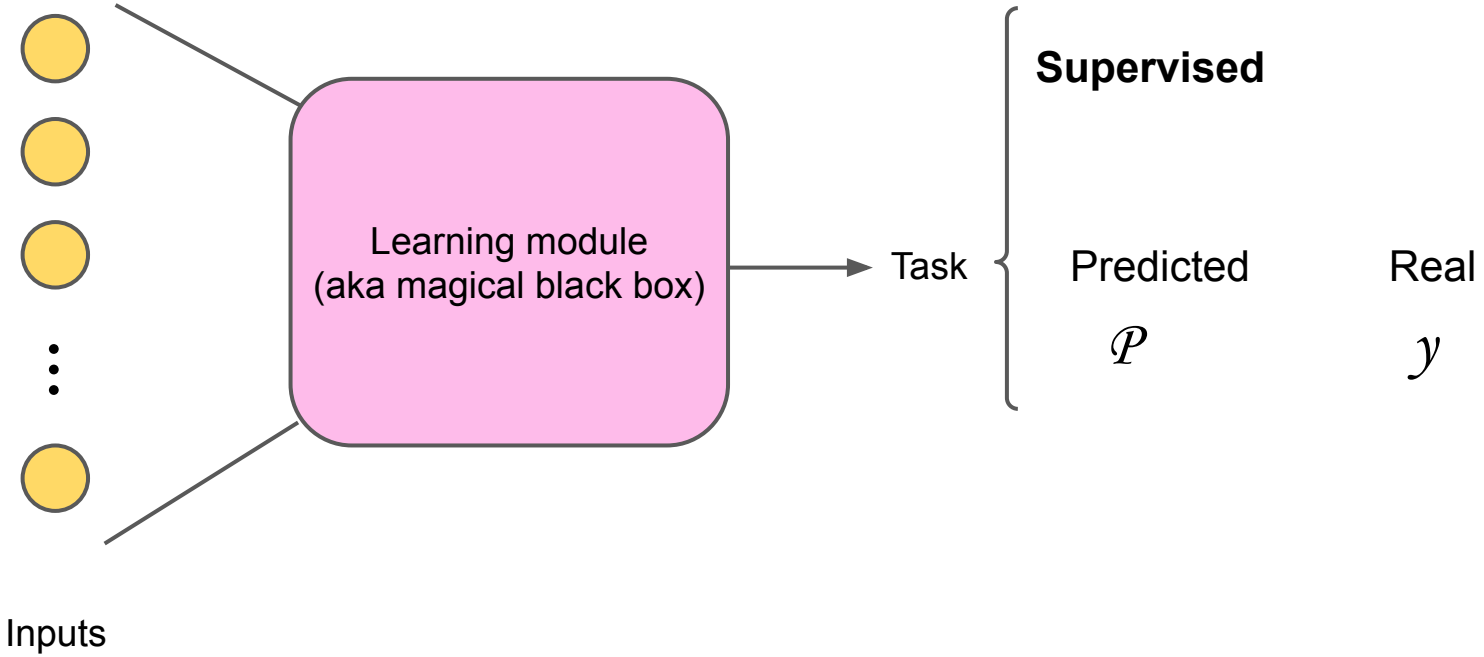


Inputs

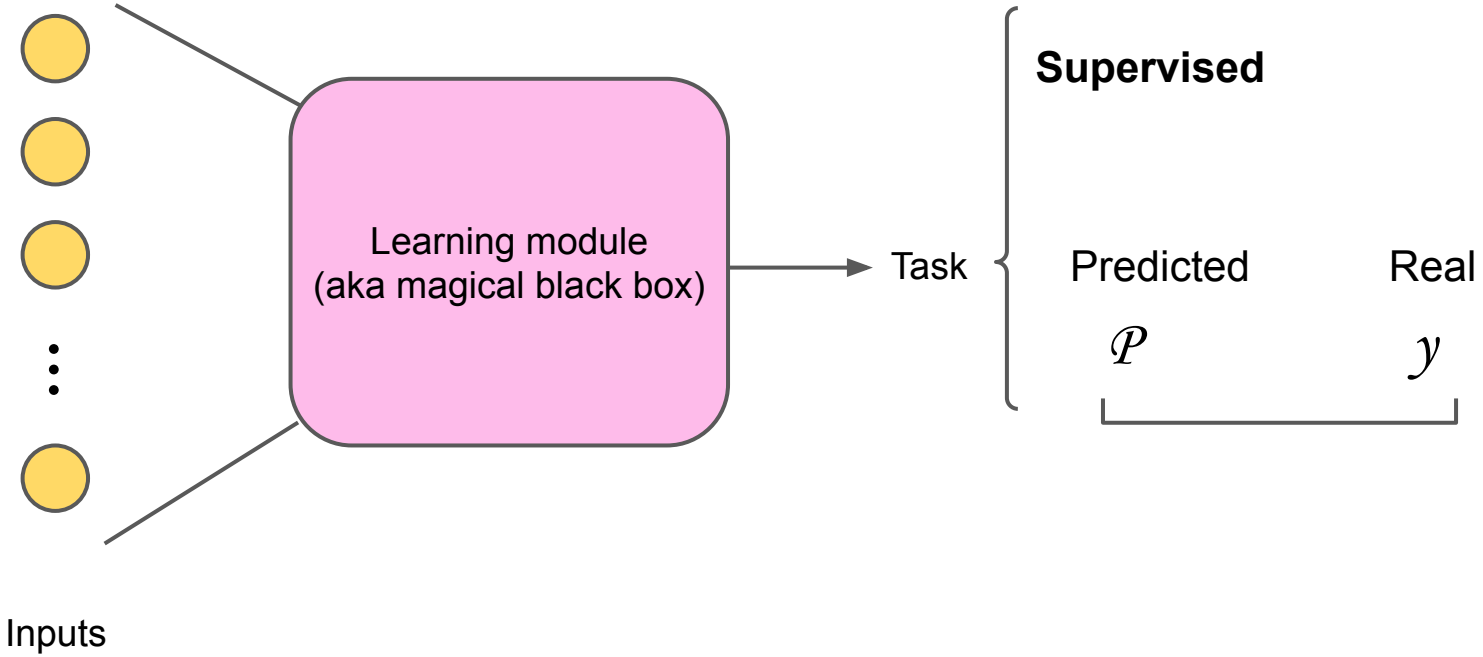
The basic components of a learning system



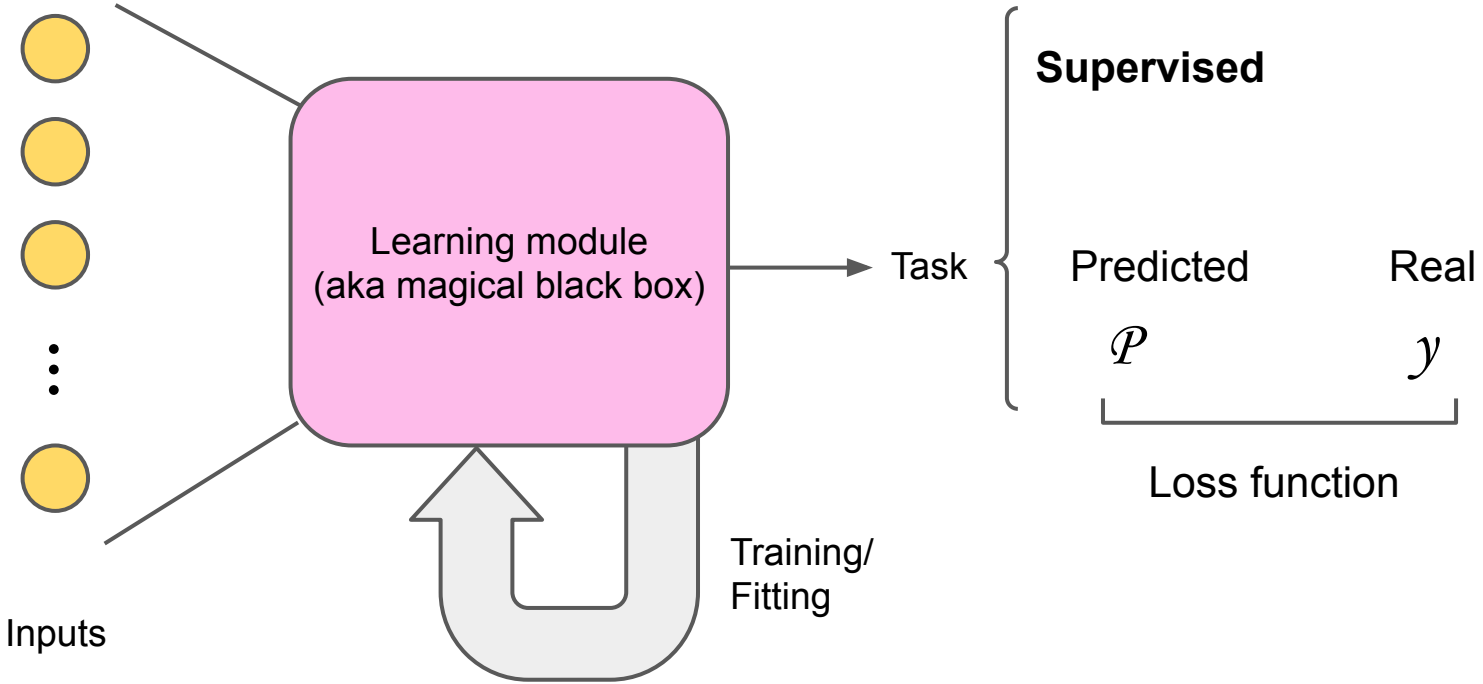
Supervised tasks use labeled data



Supervised tasks use labeled data



Supervised tasks use labeled data



Performance of supervised tasks can be determined by a number of metrics

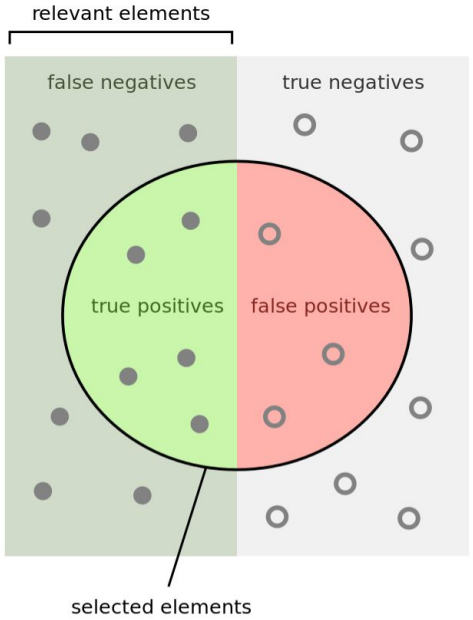
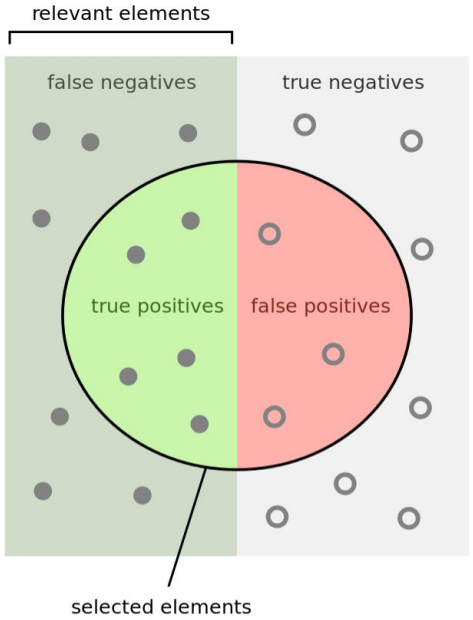


Image from medium

Performance of supervised tasks can be determined by a number of metrics



$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

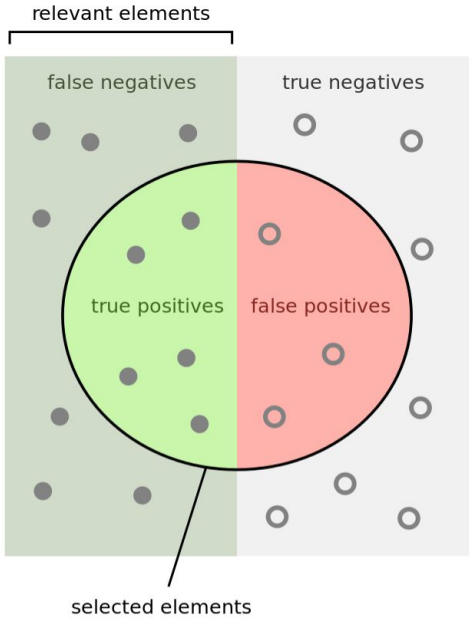
$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

Image from medium

Performance of supervised tasks can be determined by a number of metrics



$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

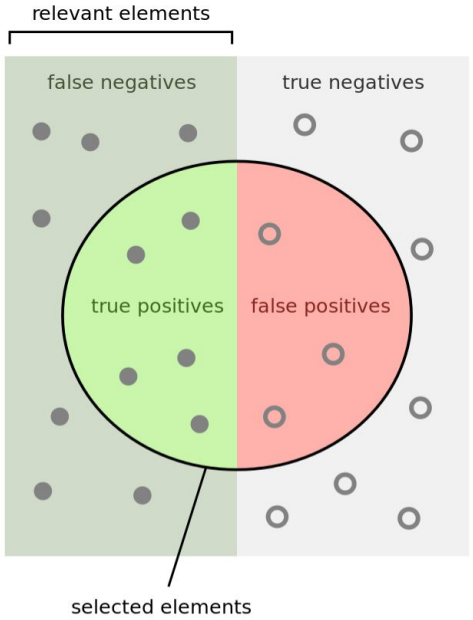
$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

Image from medium

Performance of supervised tasks can be determined by a number of metrics



$$precision = \frac{TP}{TP + FP}$$

$$\mathbf{recall} = \frac{TP}{TP + FN}$$

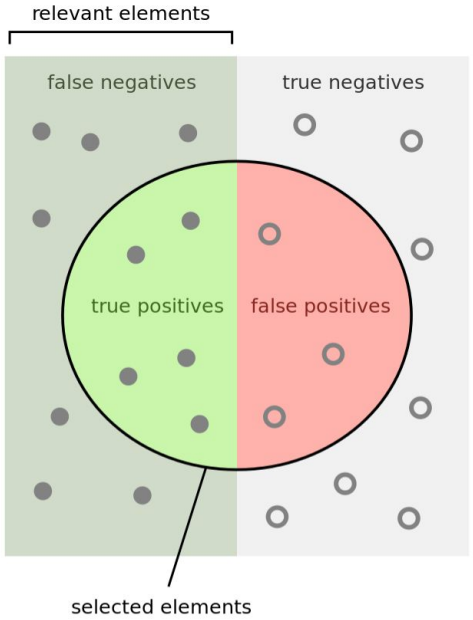
$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

Image from medium

Performance of supervised tasks can be determined by a number of metrics



$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

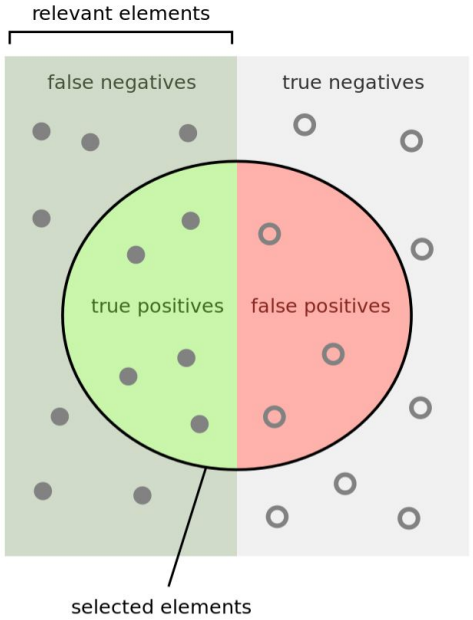
$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

Image from medium

Performance of supervised tasks can be determined by a number of metrics



$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

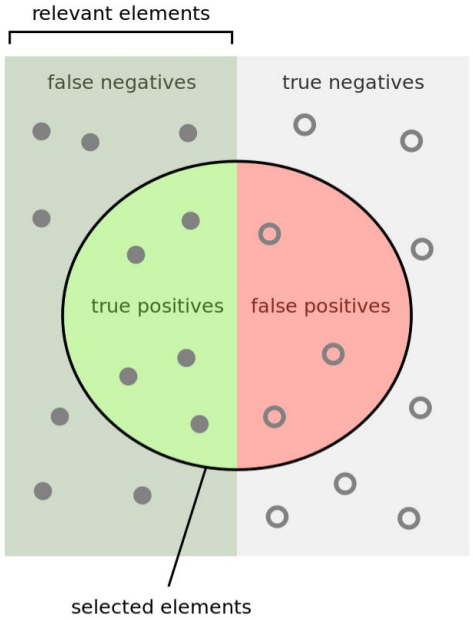
$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

Image from medium

Performance of supervised tasks can be determined by a number of metrics



$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

Image from medium

In class activity 1:

Recall, precision, accuracy

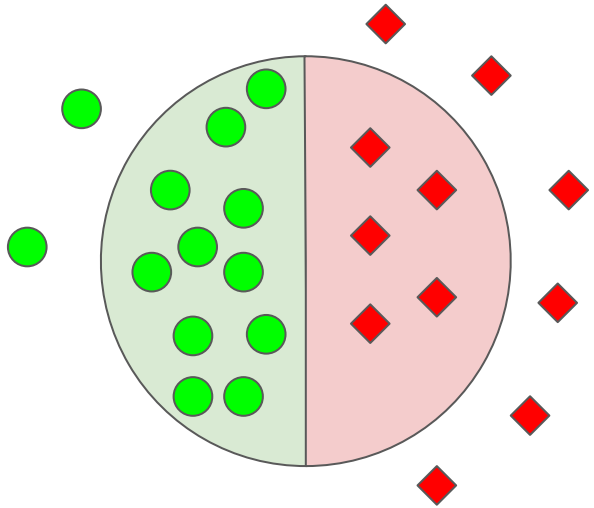
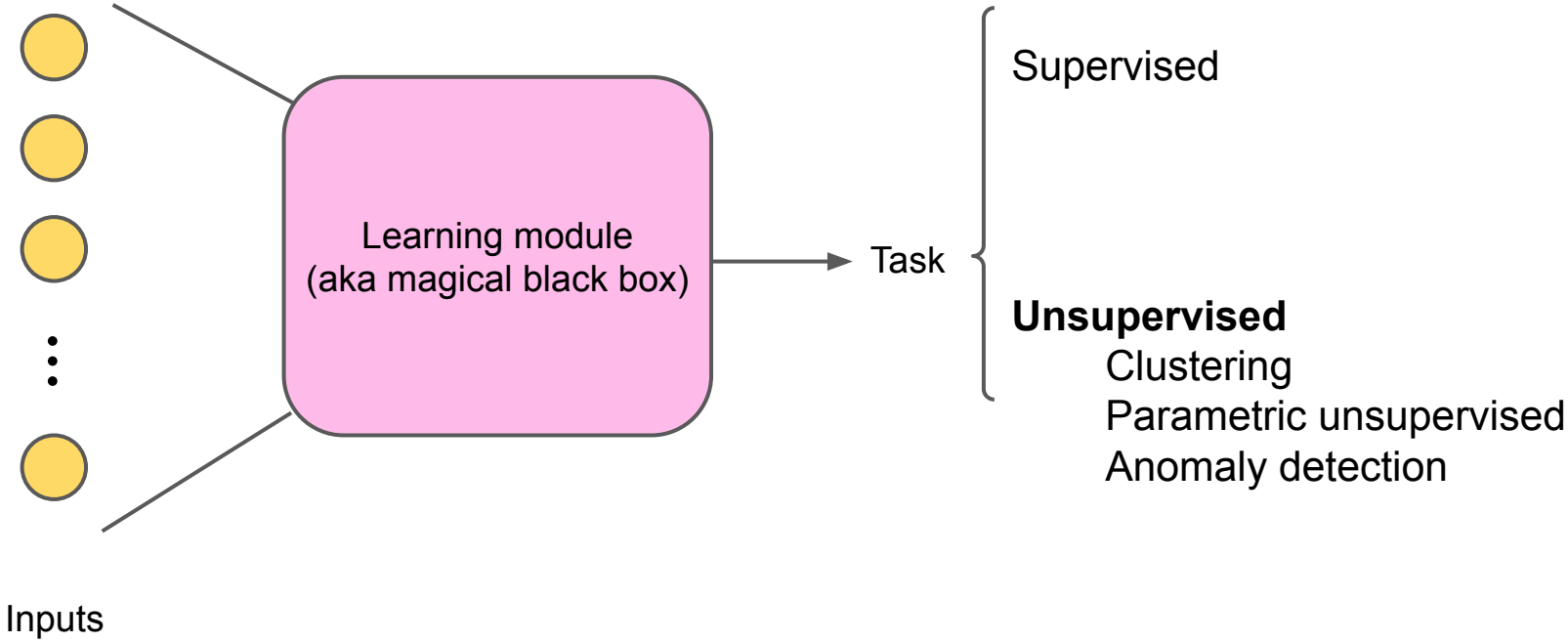
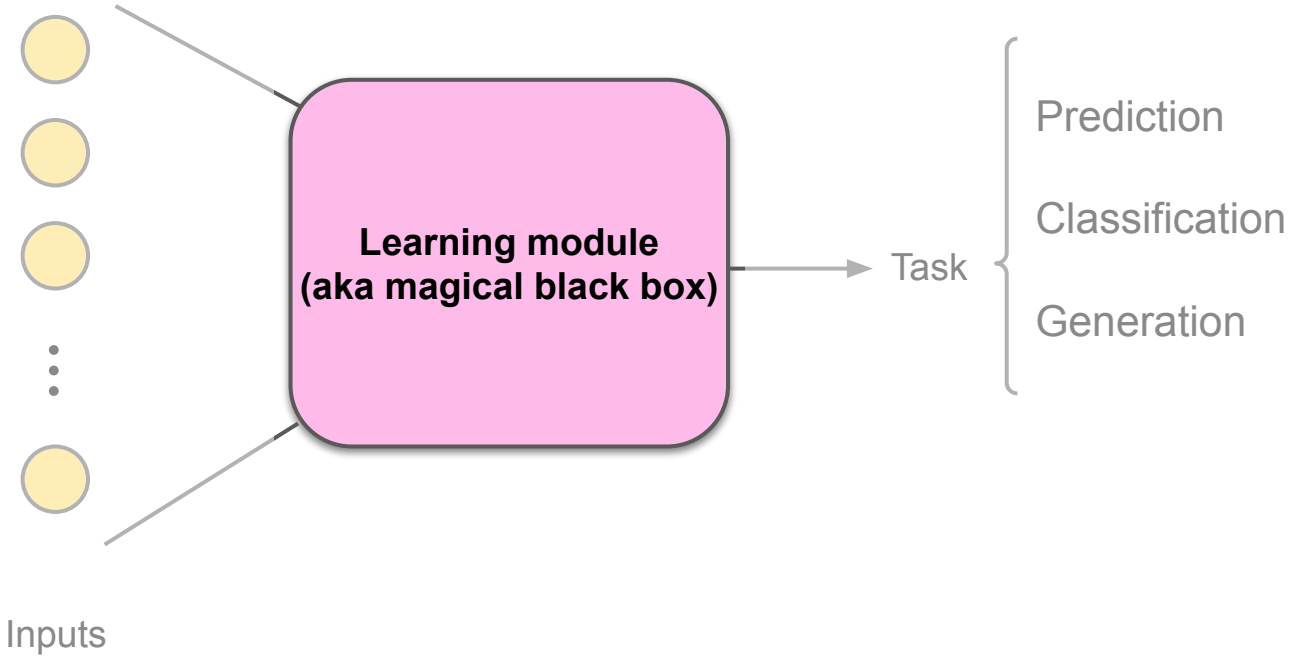


Image from Thermo Fisher Scientific: Machine learning, a primary to laboratory applications

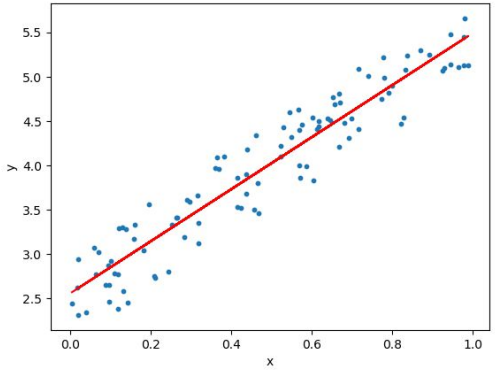
Unsupervised tasks do not need labeled data



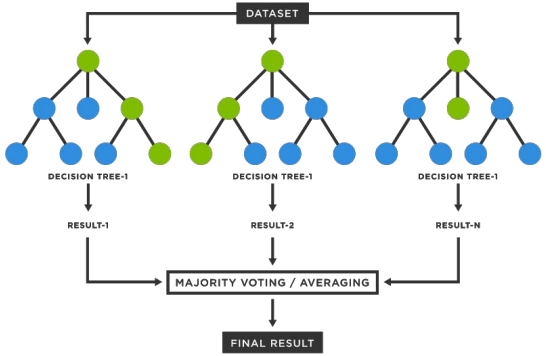
The basic components of a learning system



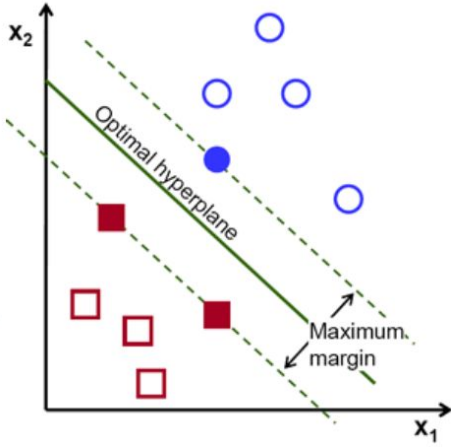
Simple models are very powerful at making predictions/clustering



Regression



Random forest



Support Vector Machine

Neural nets gained traction due to enhanced power of computers and increase in data

Neural nets gained traction due to enhanced power of computers and increase in data

A perceptron

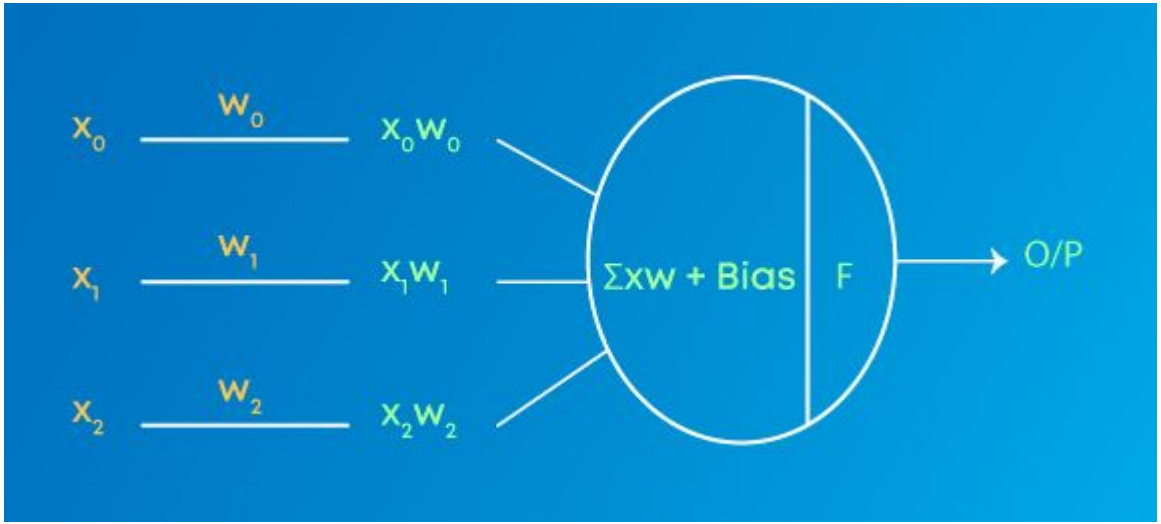


Image from MyGreatLearning

Deep neural nets are data hungry

A dense neural net

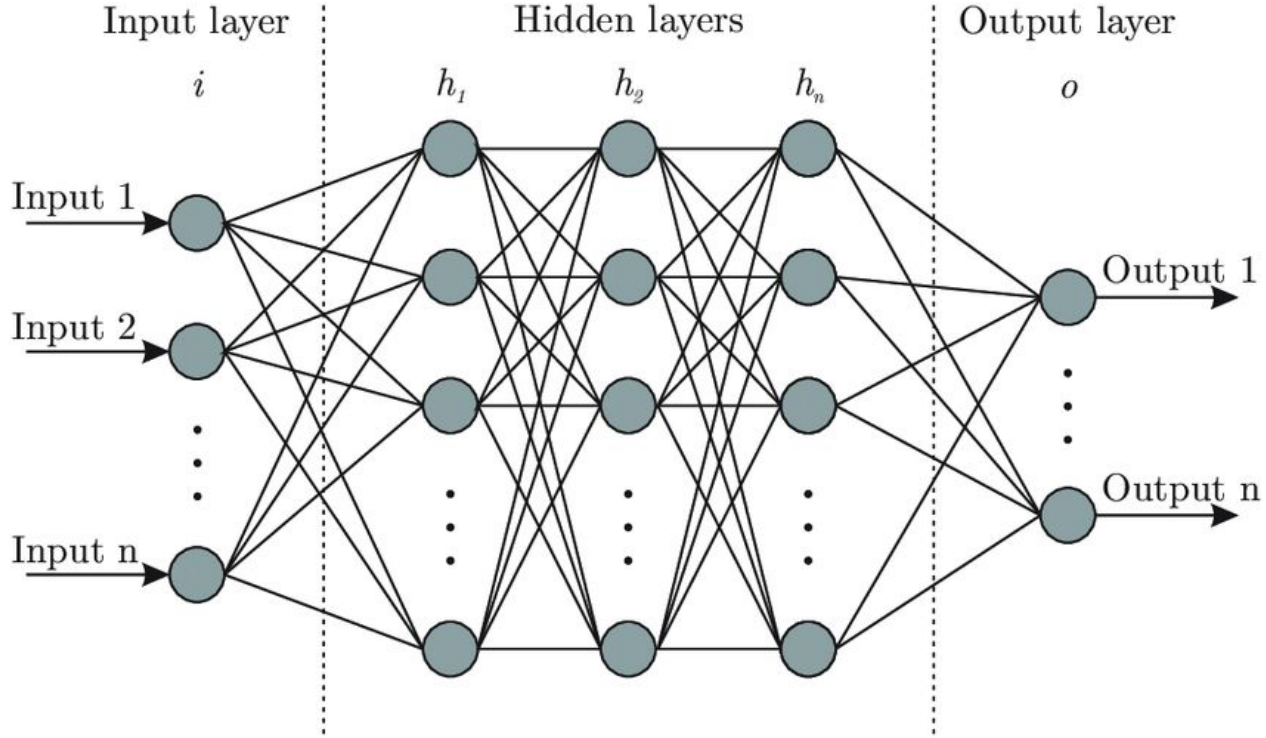


Image from Kdnuggets

In class activity:

Examples of simple learning algorithms

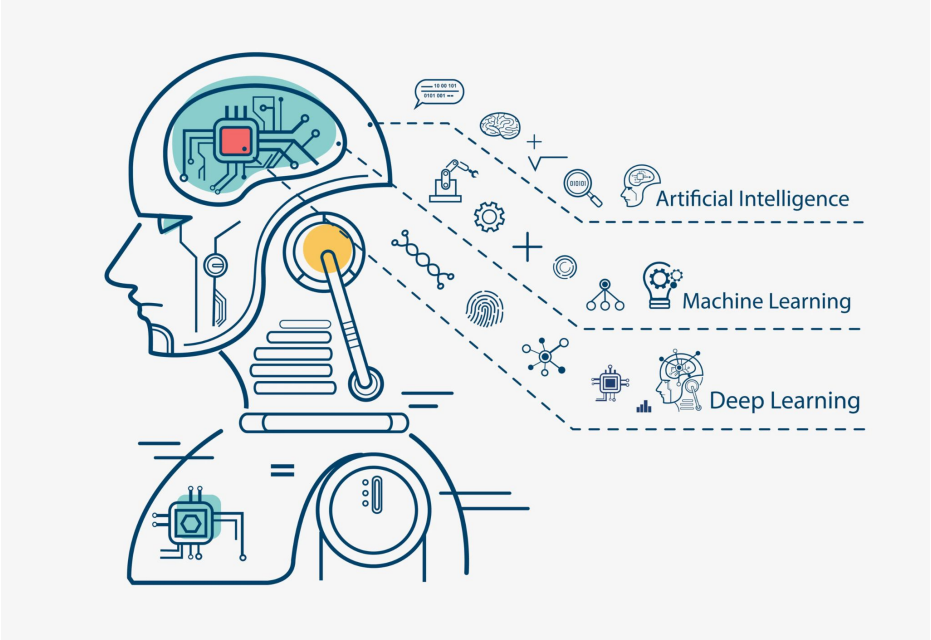


Image from Thermo Fisher Scientific: Machine learning, a primary to laboratory applications

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 1:

Train:Test split

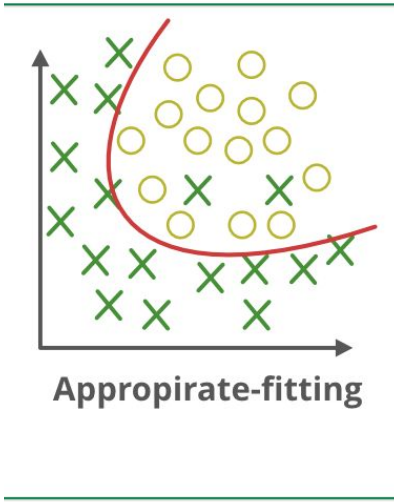


Image from geeks for geeks

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 1:

Train:Test split

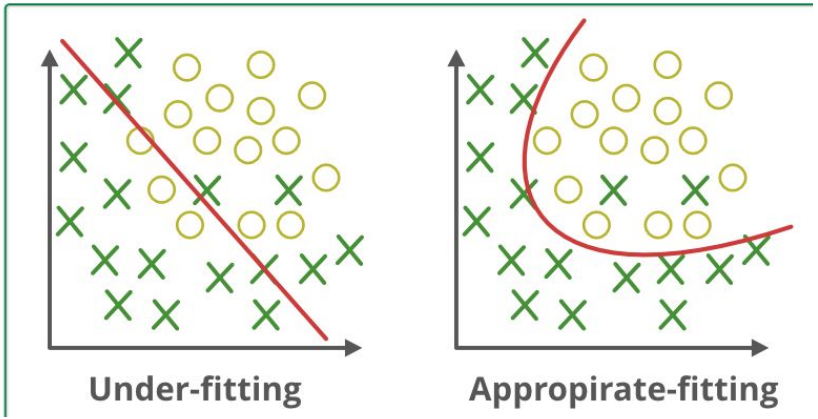


Image from geeks for geeks

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 1:

Train:Test split

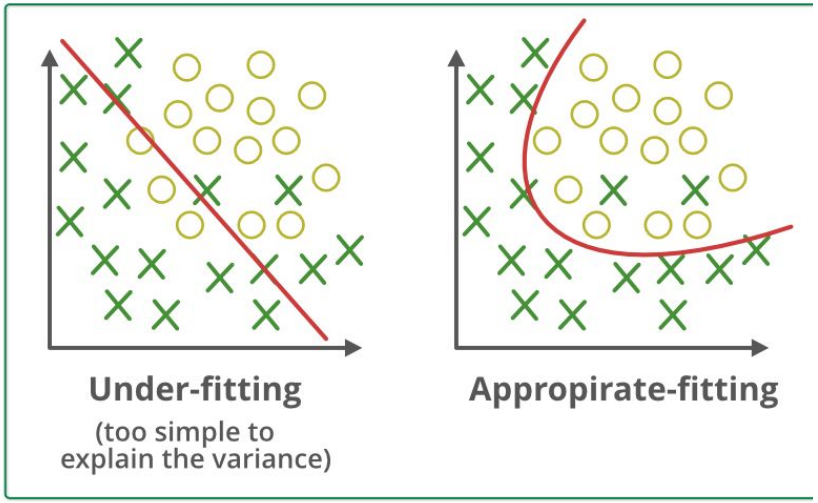


Image from geeks for geeks

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 1:

Train:Test split

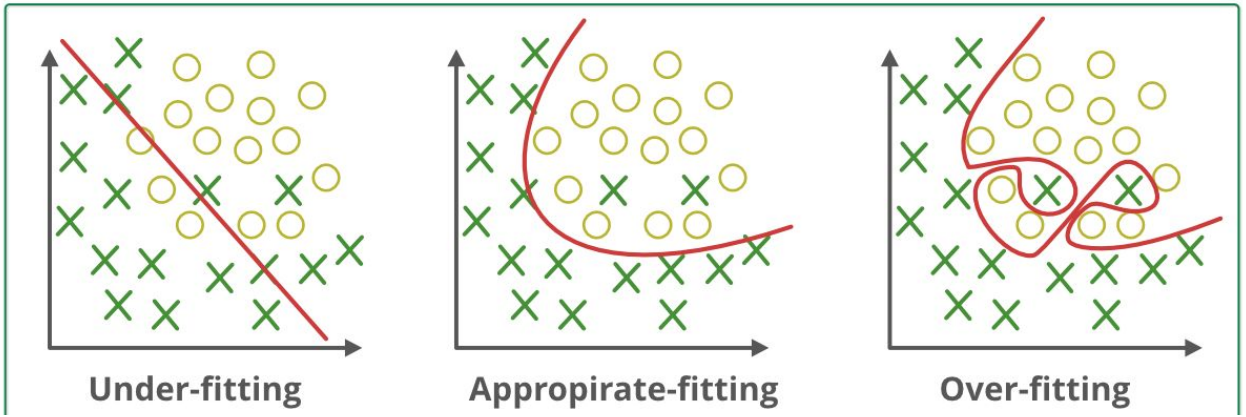


Image from geeks for geeks

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 1:

Train:Test split

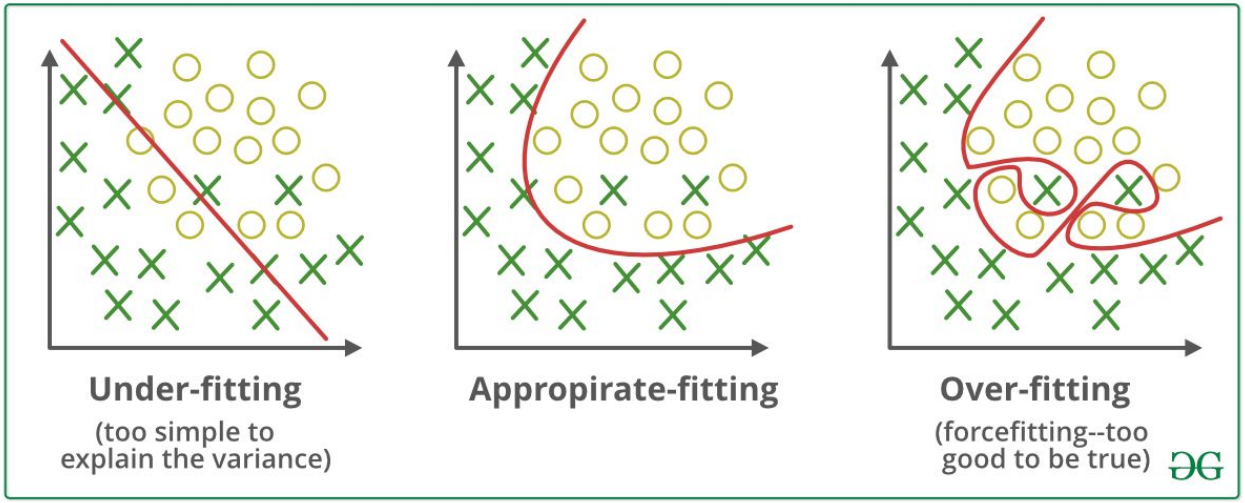


Image from geeks for geeks

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 2:

Regularization

Image from Medium

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 3:

Number of hidden layers and number of neurons in each

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 4:

Learning rate

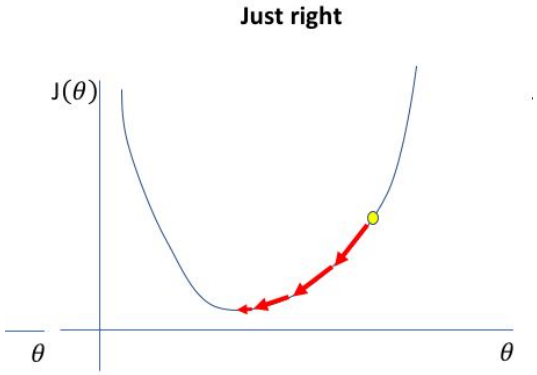


Image from Jeremy Jordan blog

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 4:

Learning rate

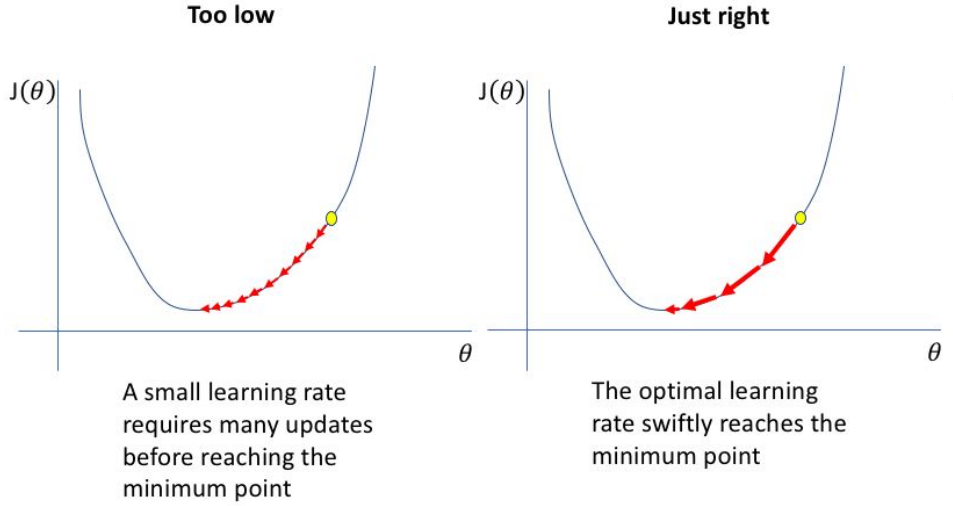


Image from Jeremy Jordan blog

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 4:

Learning rate

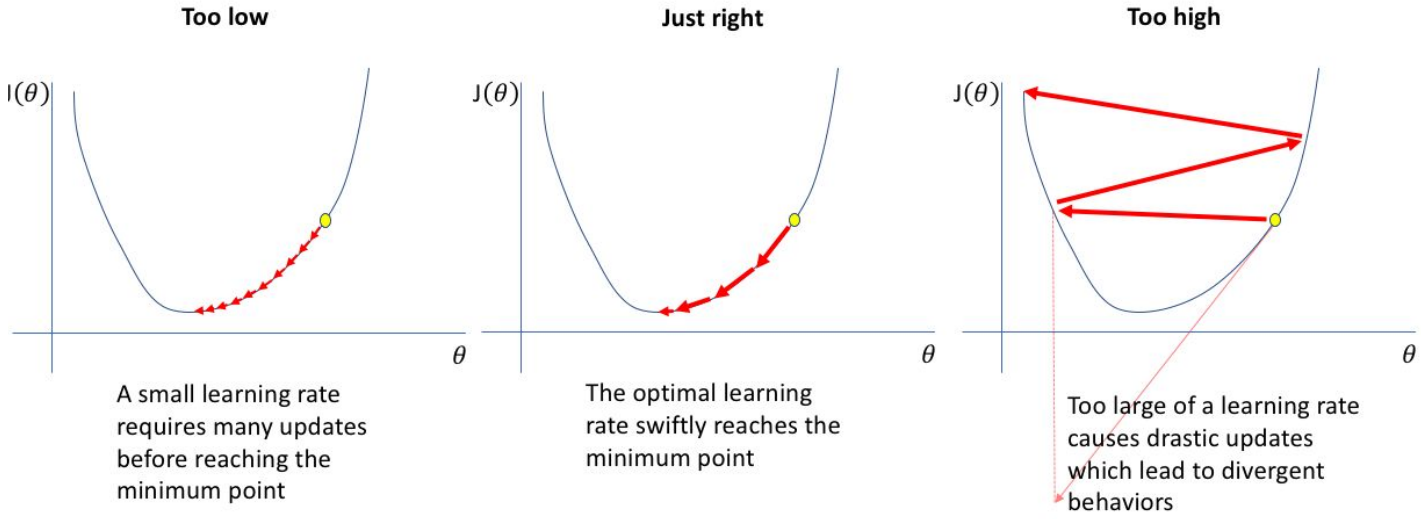


Image from Jeremy Jordan blog

In addition to internal weights, deep neural net require hyper-parameter tuning

Consideration 5:

Activation function

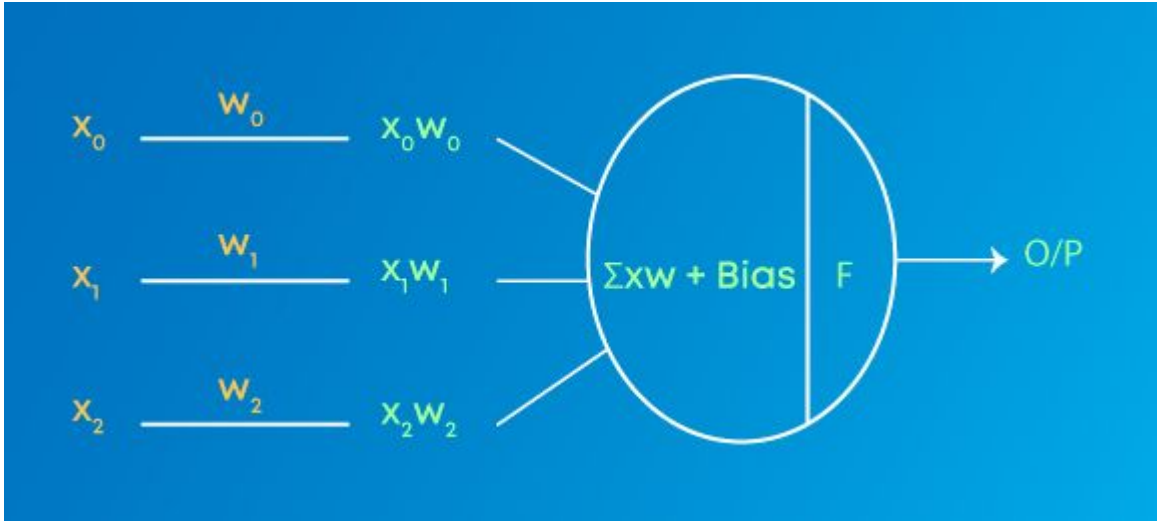


Image from MyGreatLearning

In addition to internal weights, deep neural net require hyper-parameter tuning

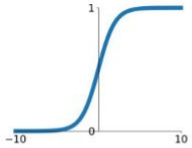
Consideration 5:

Activation function

Activation Functions

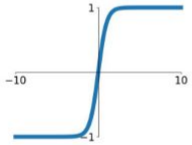
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



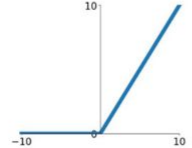
tanh

$$\tanh(x)$$



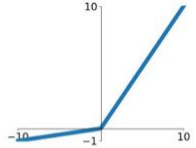
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$



Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$

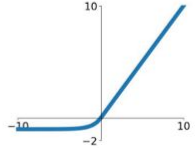
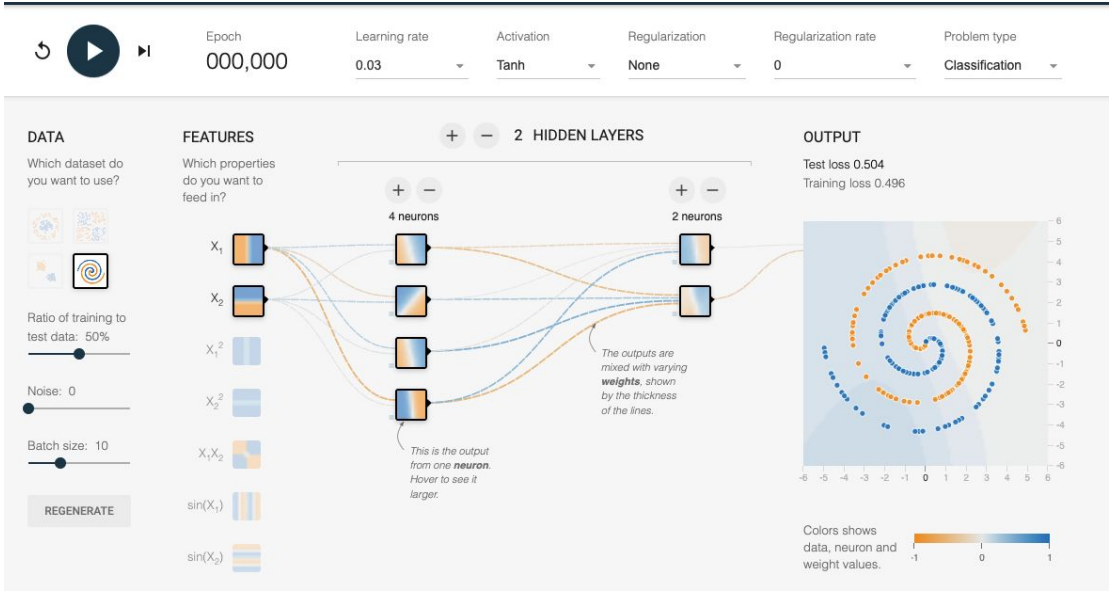


Image from Medium

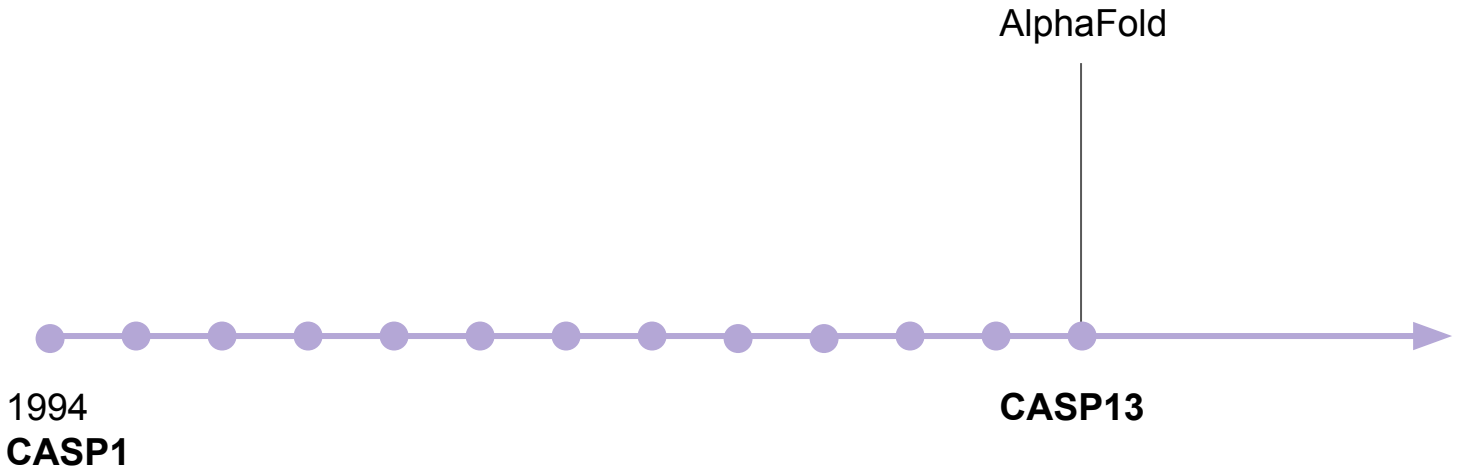
In class activity:

Hyper-parameter tuning

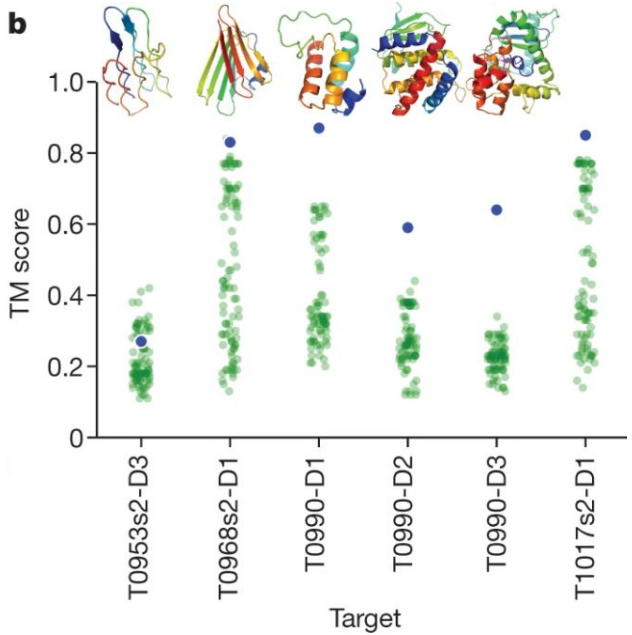


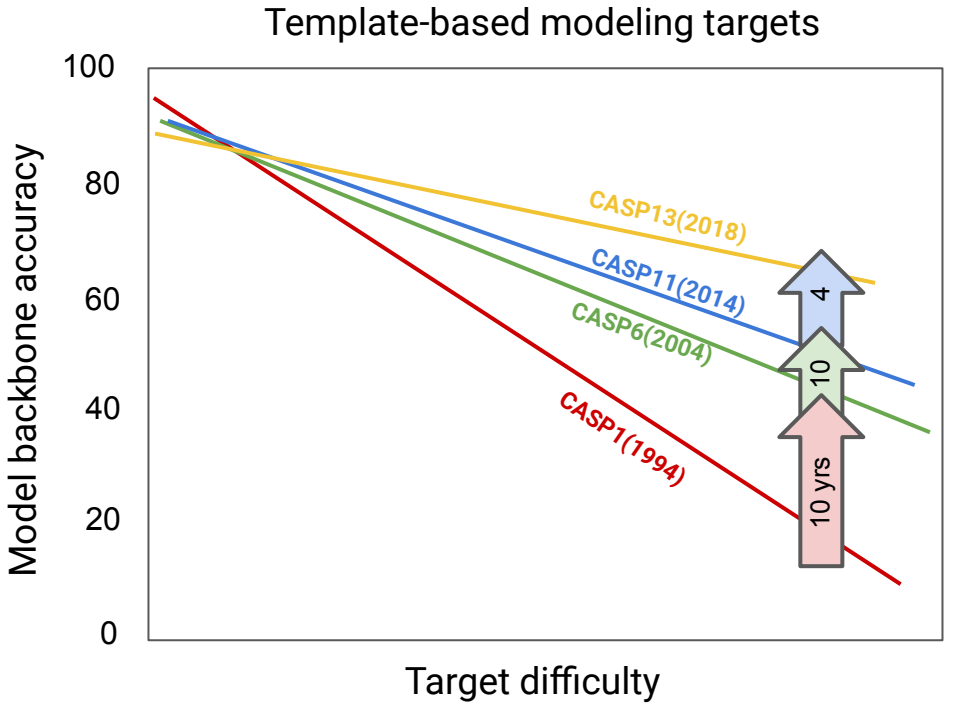
CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction



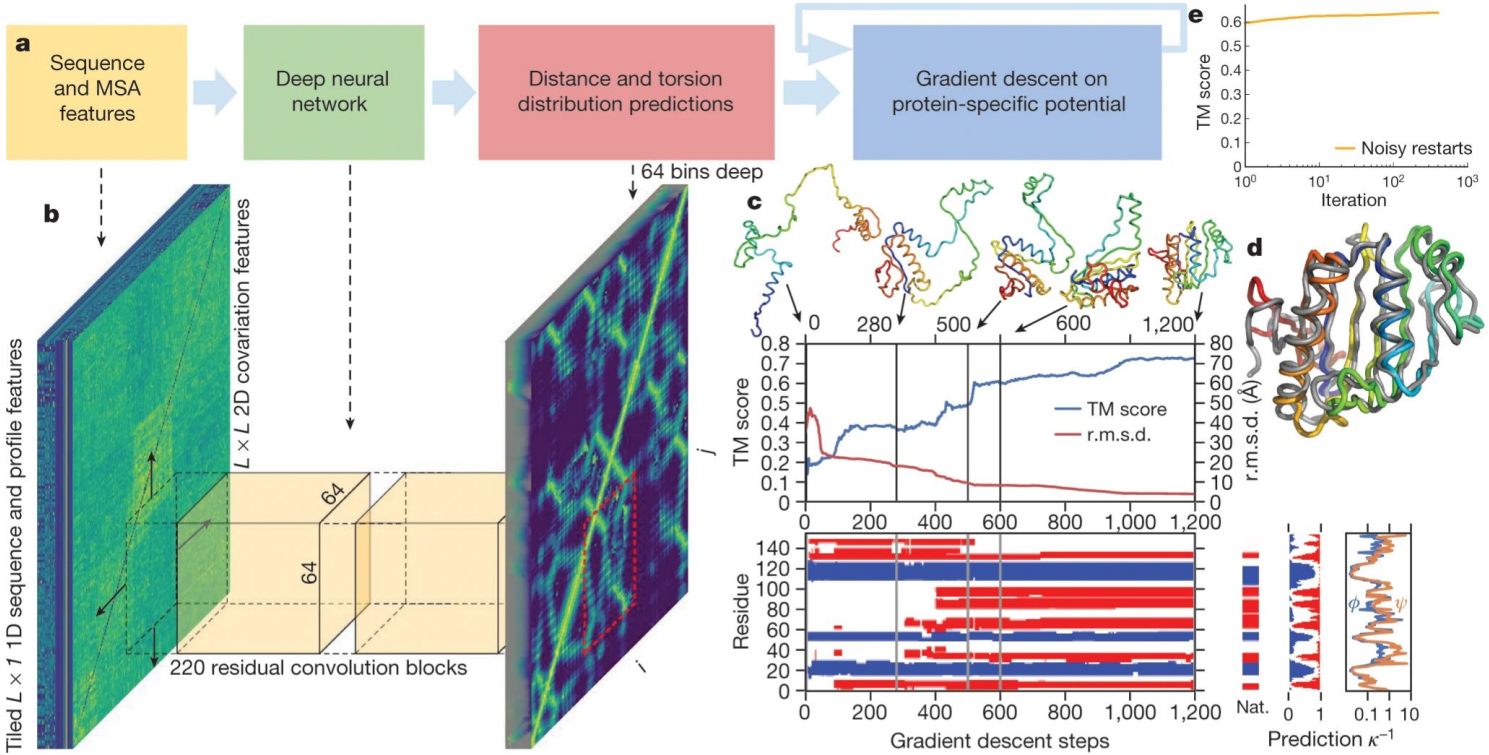
AlphaFold caused a paradigm shift the field of structure prediction





Senior et al, *Nature* **2020**

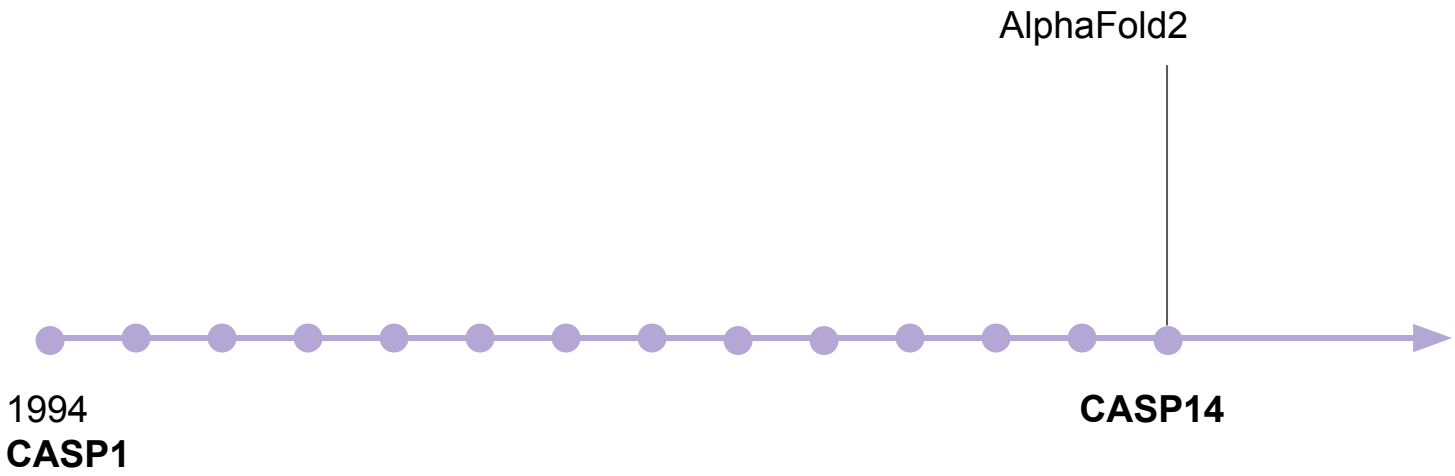
AlphaFold caused a paradigm shift the field of structure prediction



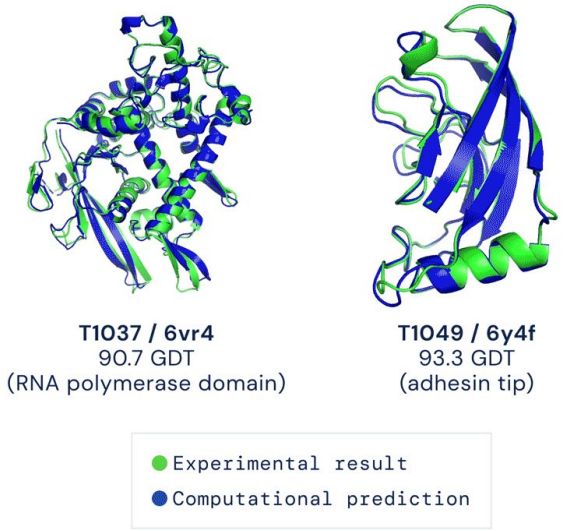
Senior et al, *Nature* **2020**

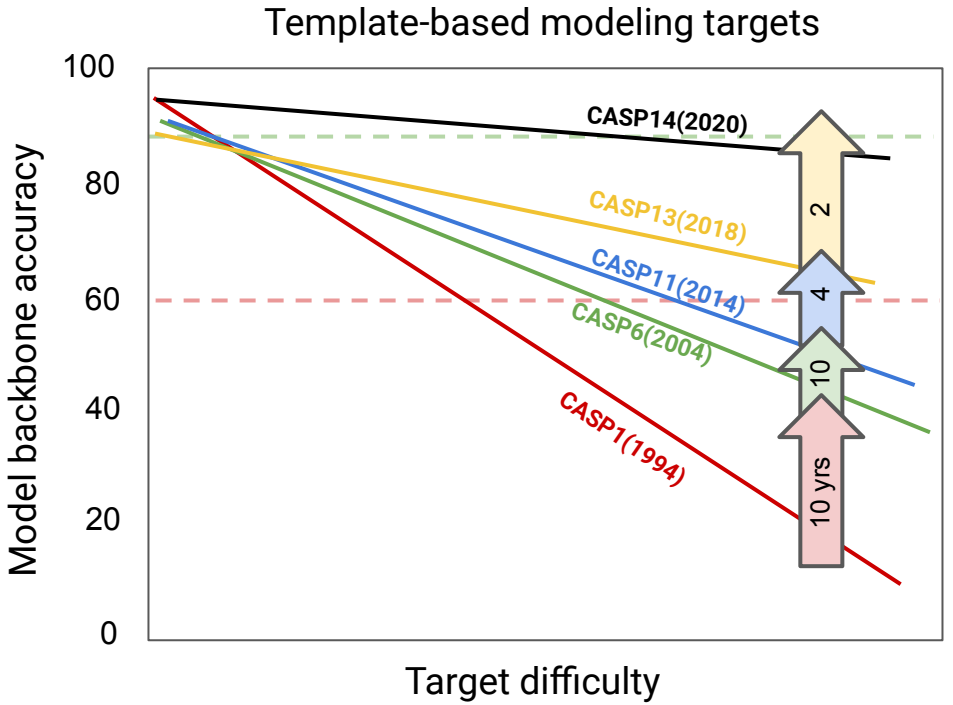
CASP:

Critical **A**ssessment of protein **S**tructure **P**rediction



AlphaFold2 generated highly accurate structures





CASP progress adapted from Sergei Grudinin, AlgoSB 2021
Proteins from Deepmind blog

AlphaFold2 is better than all competitors at ~all tasks

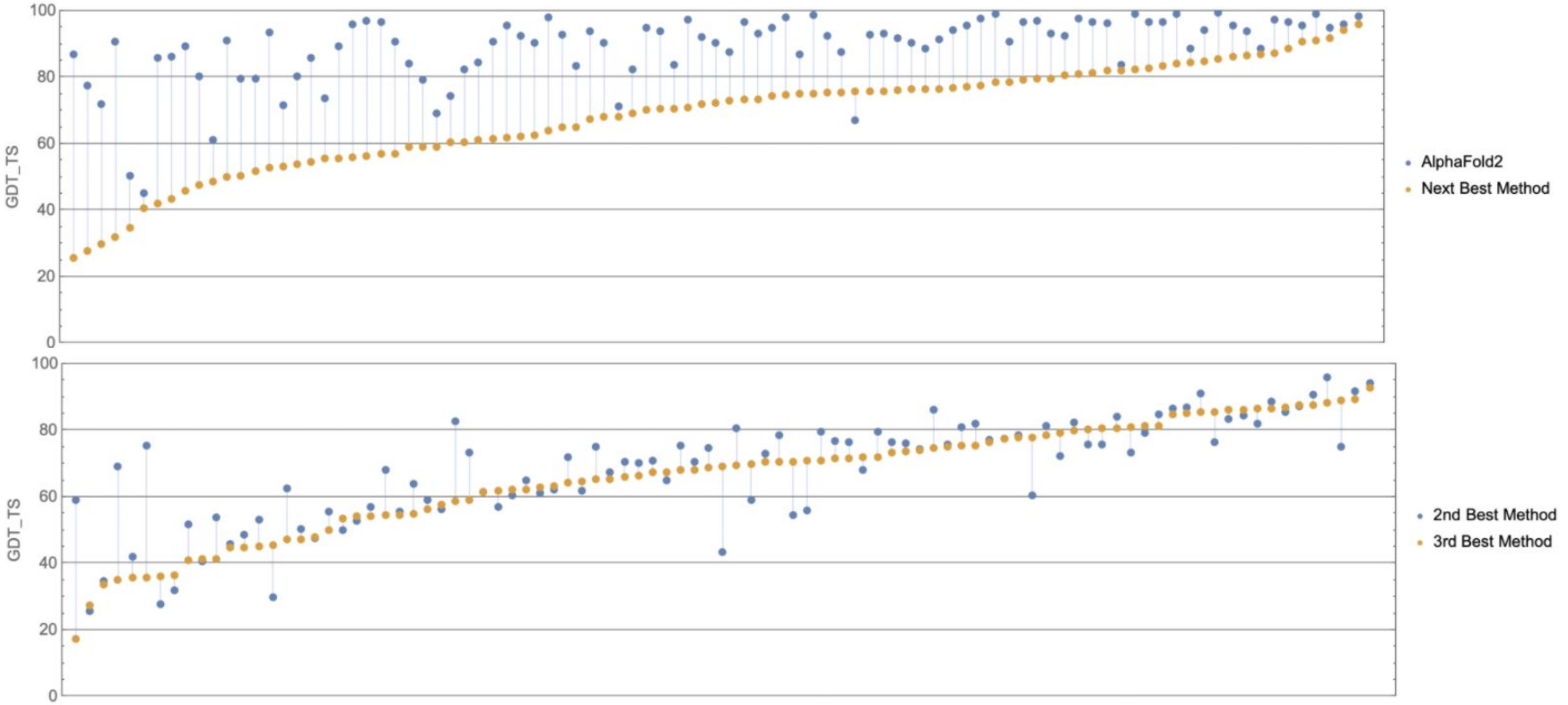
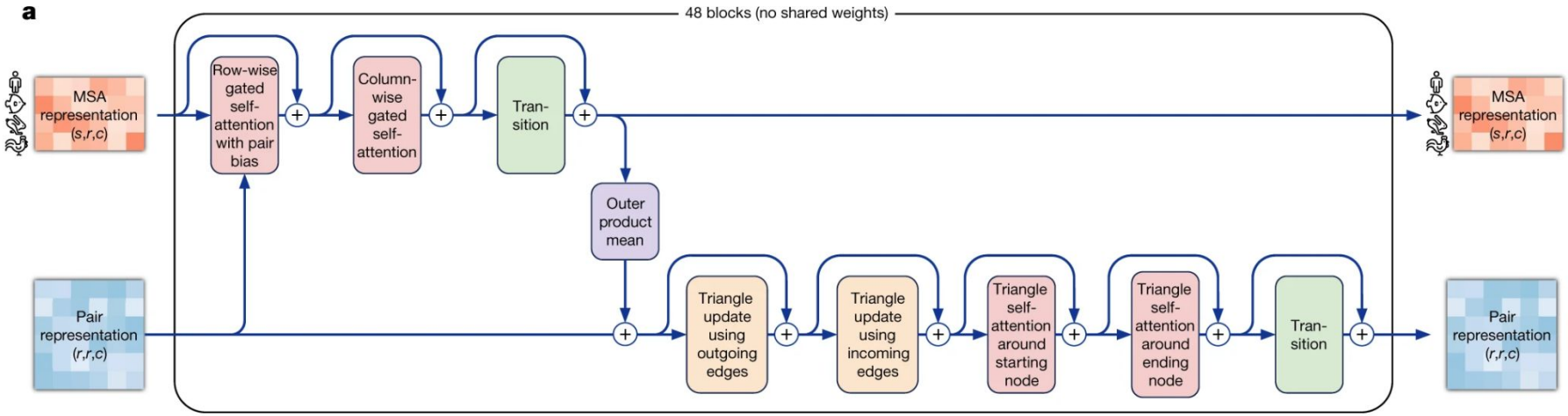


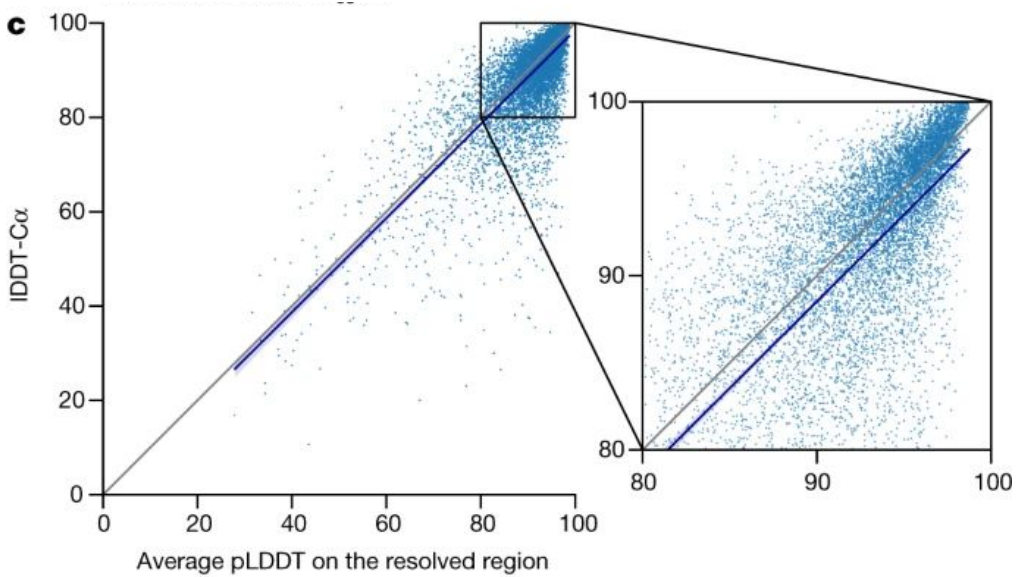
Figure from AlQuraishi's blog

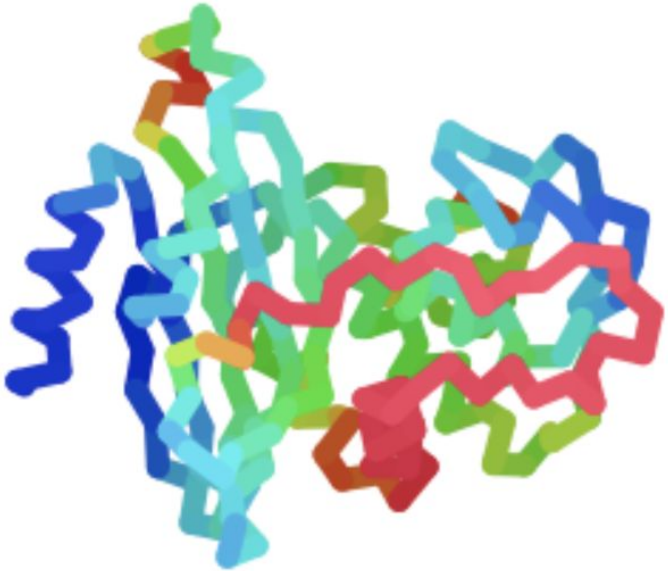
AlphaFold2 uses attention to learn from protein seq



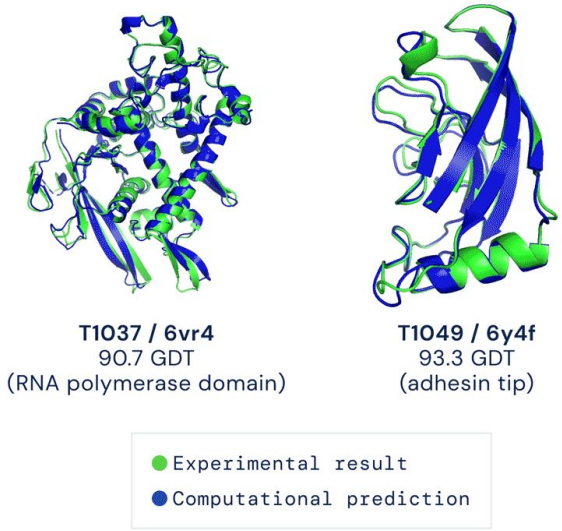
Jumper et al, *Nature* **2021**

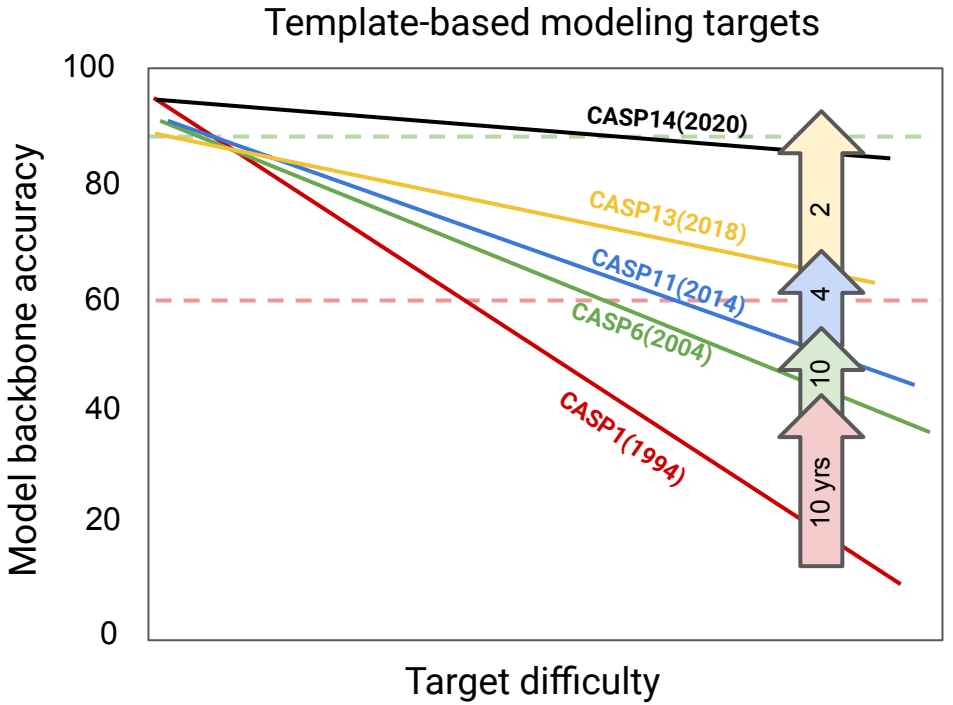
AlphaFold2 can be used for structure prediction





AlphaFold2 finished CASP as we know it

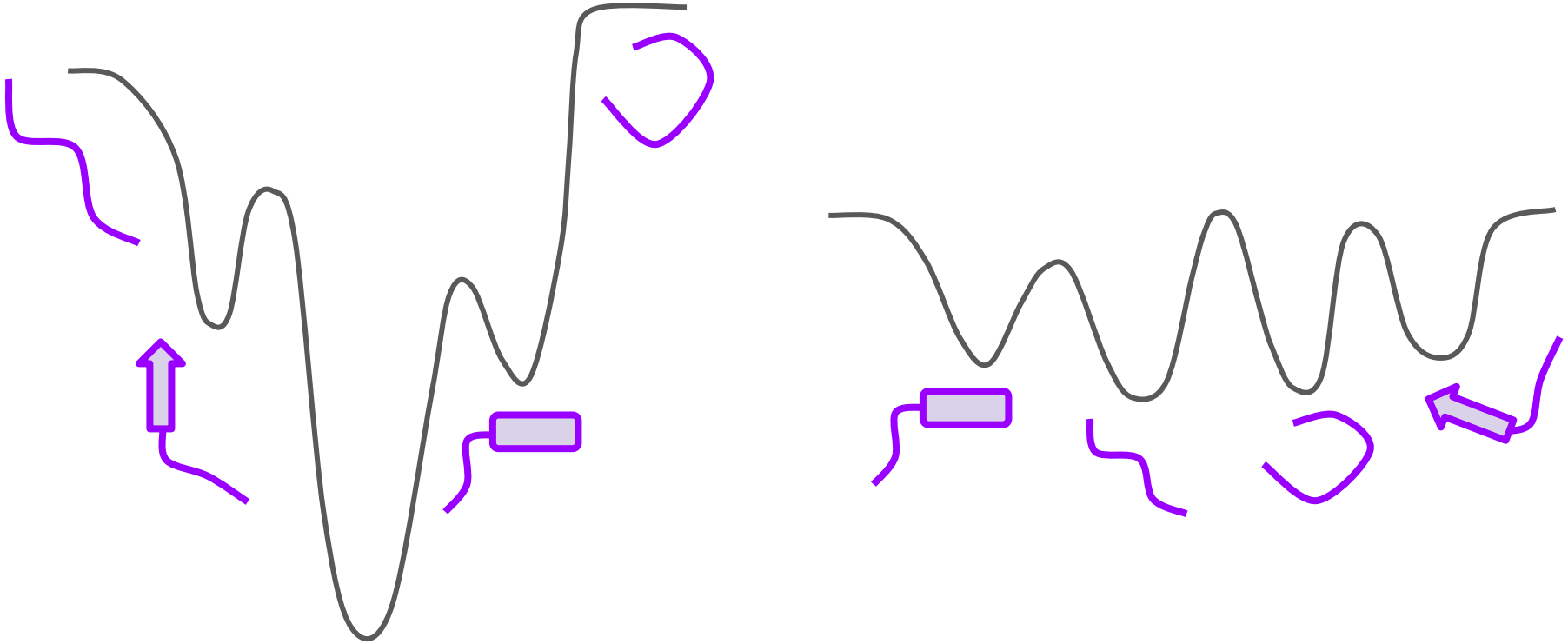




CASP progress adapted from Sergei Grudinin, AlgoSB 2021
Proteins from Deepmind blog

What lies ahead for protein structure-prediction?

What lies ahead for protein structure-prediction?



What lies ahead for protein structure-prediction?

- Prediction of multiple functional conformations
- Prediction of protein-protein interaction
- Disordered protein prediction
- Protein design
- Function prediction

For the next lecture:

1. Read journal for the next week
   a. Moderated by **group IV**

2. Post-class assignment

3. Work on your updated specific aims page

Next lecture:

Machine learning in protein engineering