LOSCHMIDT
LABORATORIES

12. Artificial Intelligence in Life Sciences

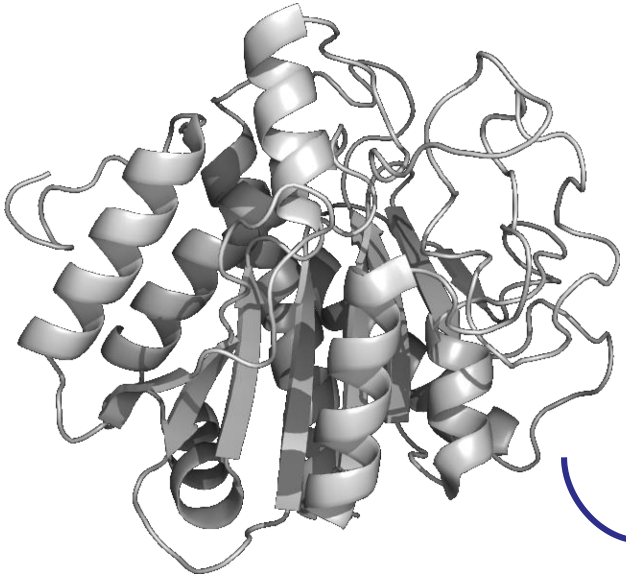
Outline

- □ Motivation
- □ Introduction to AI and ML
- □ Modern challenges in Bioengineering
- □ Basics of ML
- □ Recent applications

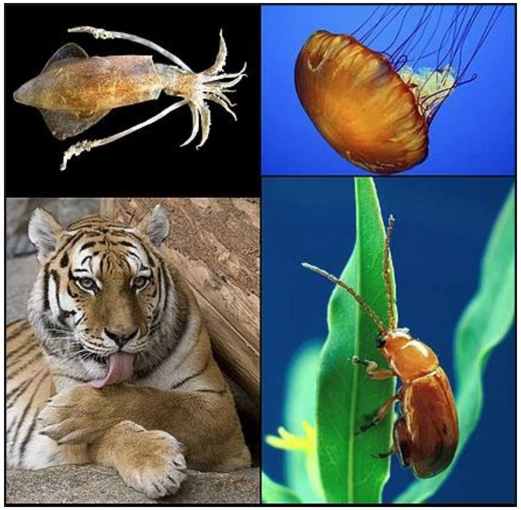
Motivation

Motivation

MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHC
AGLGRLIACDLIGMGDSDKLDPSGPERYAYAEHRDYLDALWEALDLGDRVV
LVVHDWGSALGFDWARRHRERVQGIAYMEAIAMPIEWADFPEQDRDLFQ
AFRSQAGEELVLQD



Function

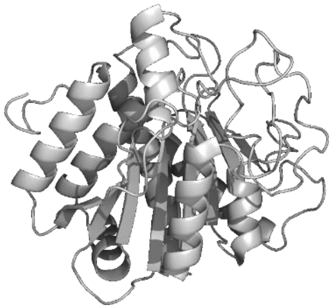


Motivation

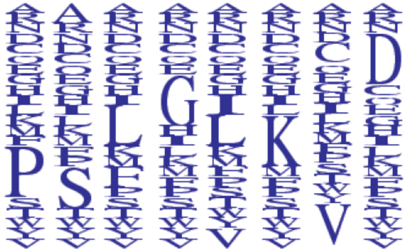
Sequence

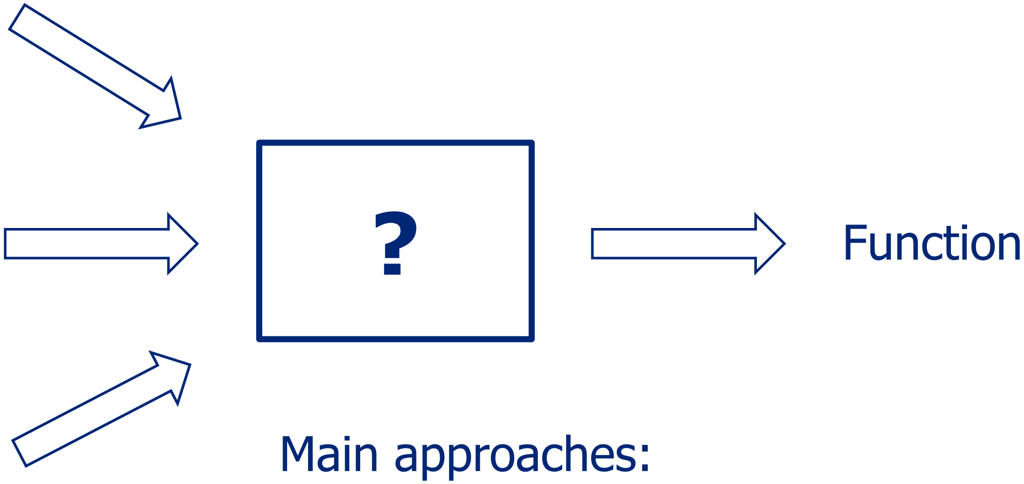
MSLGAKPFGEKKFIEIKGRRMAYIDEGTG
DPILFQHGNPTSSYLWRNIMPHCAGLGR
LIACDLIGMGDSDKLDPSGPERYAYAEHR
DYLDALWEALDLGDRVVLVVHDWGSAL
GFDWARRHRERVQGIAYMEAIAMPIEW
ADFPEQDRDLFQAFRSQAGEELVLQD

Structure



Evolution





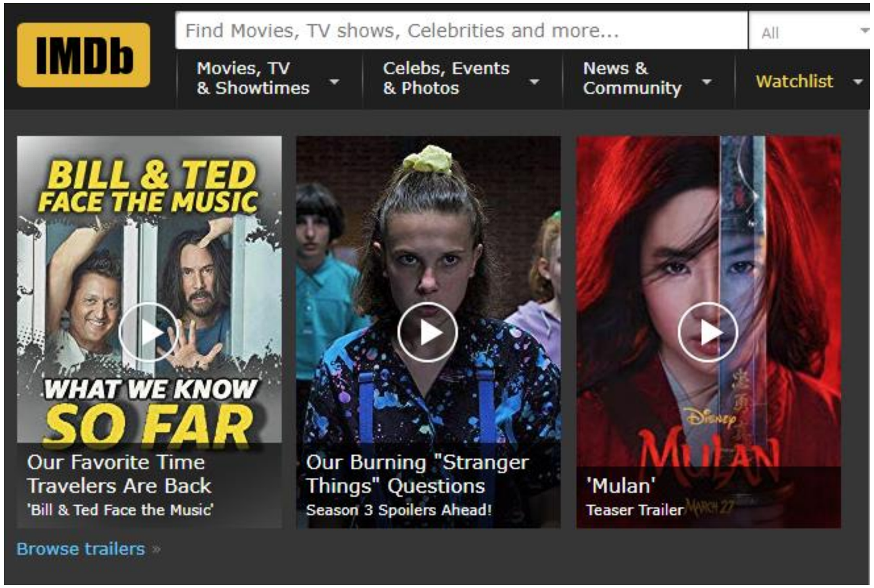
Main approaches:

- Experimental
- Rule-based
- Machine learning

Introduction to AI and ML

Introduction to AI and ML

- Recommendation engines
- Gaming
- Image & speech recognition
- Anomaly detection
- Natural language processing
- Data mining
- …



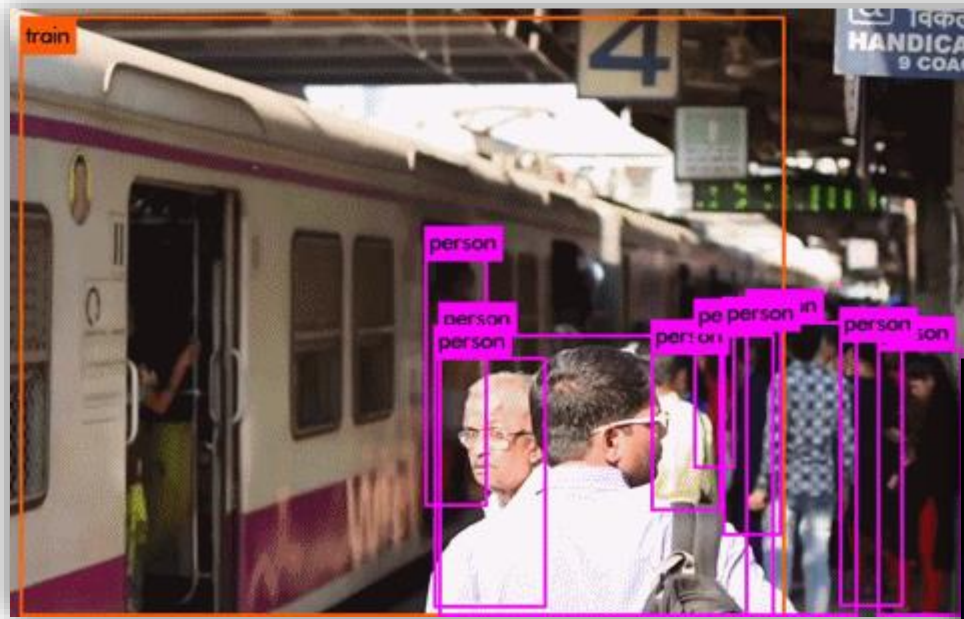
Our Favorite Time Travelers Are Back
'Bill & Ted Face the Music'

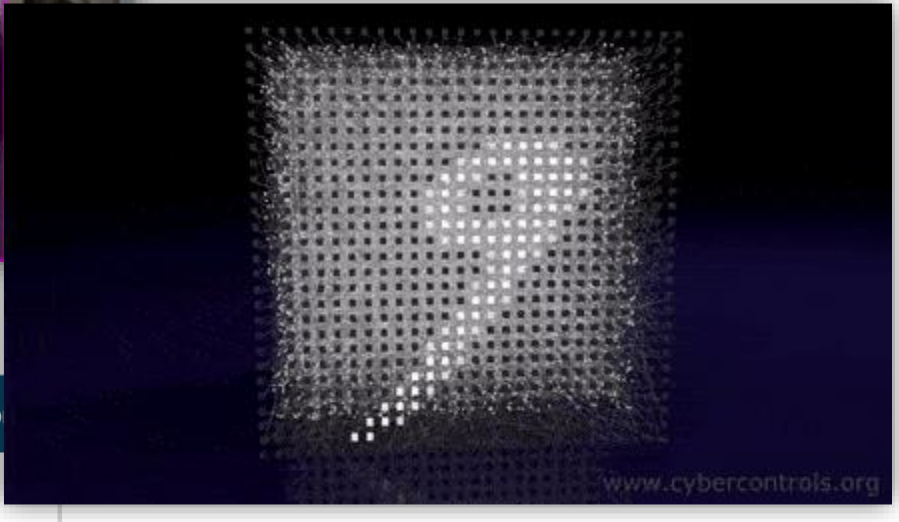
Our Burning "Stranger Things" Questions
Season 3 Spoilers Ahead!

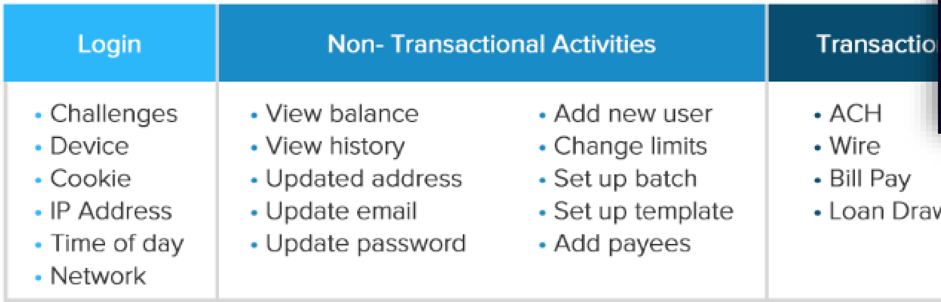
'Mulan'
Teaser Trailer

Browse trailers »

Introduction to AI and ML

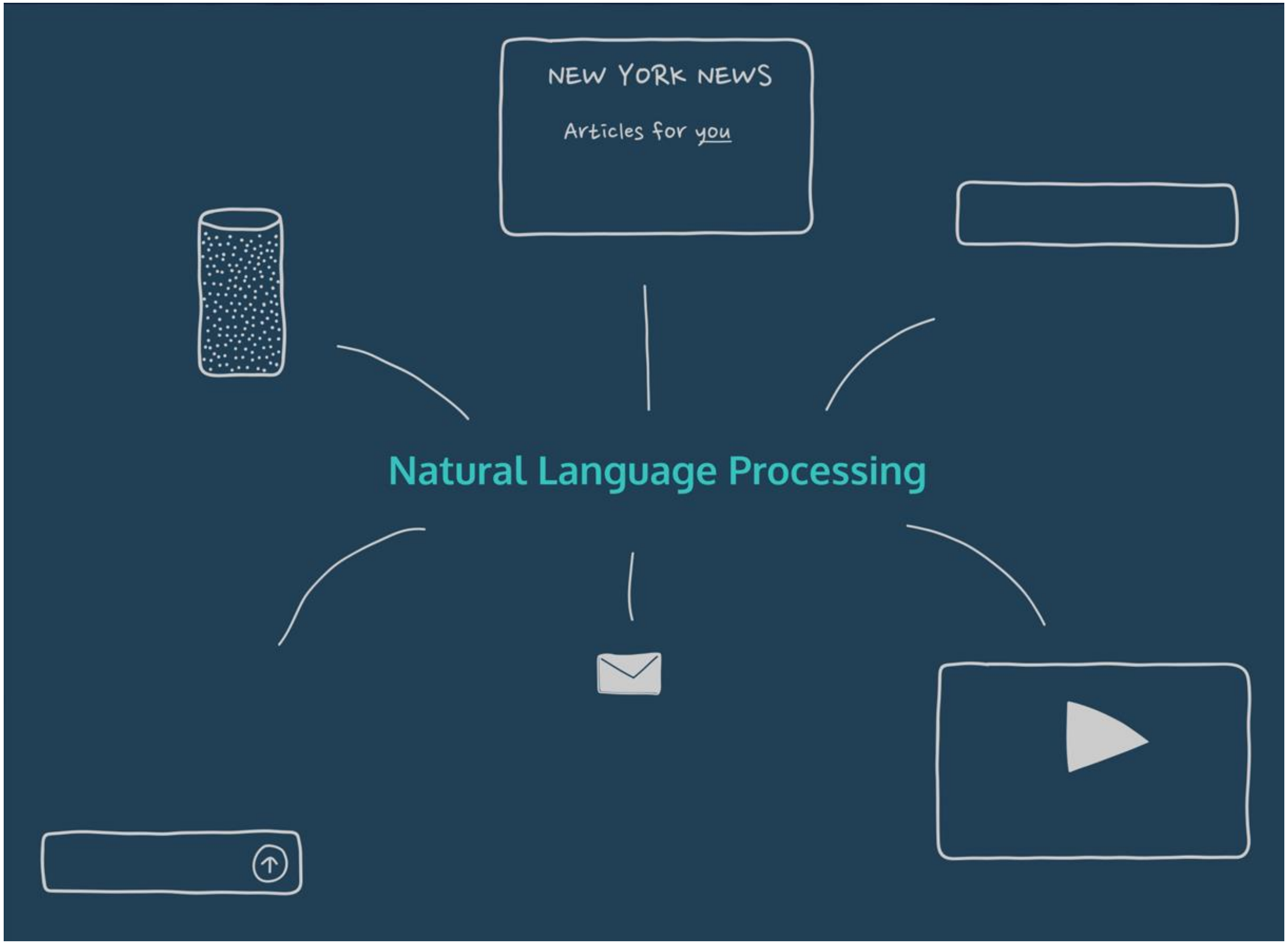






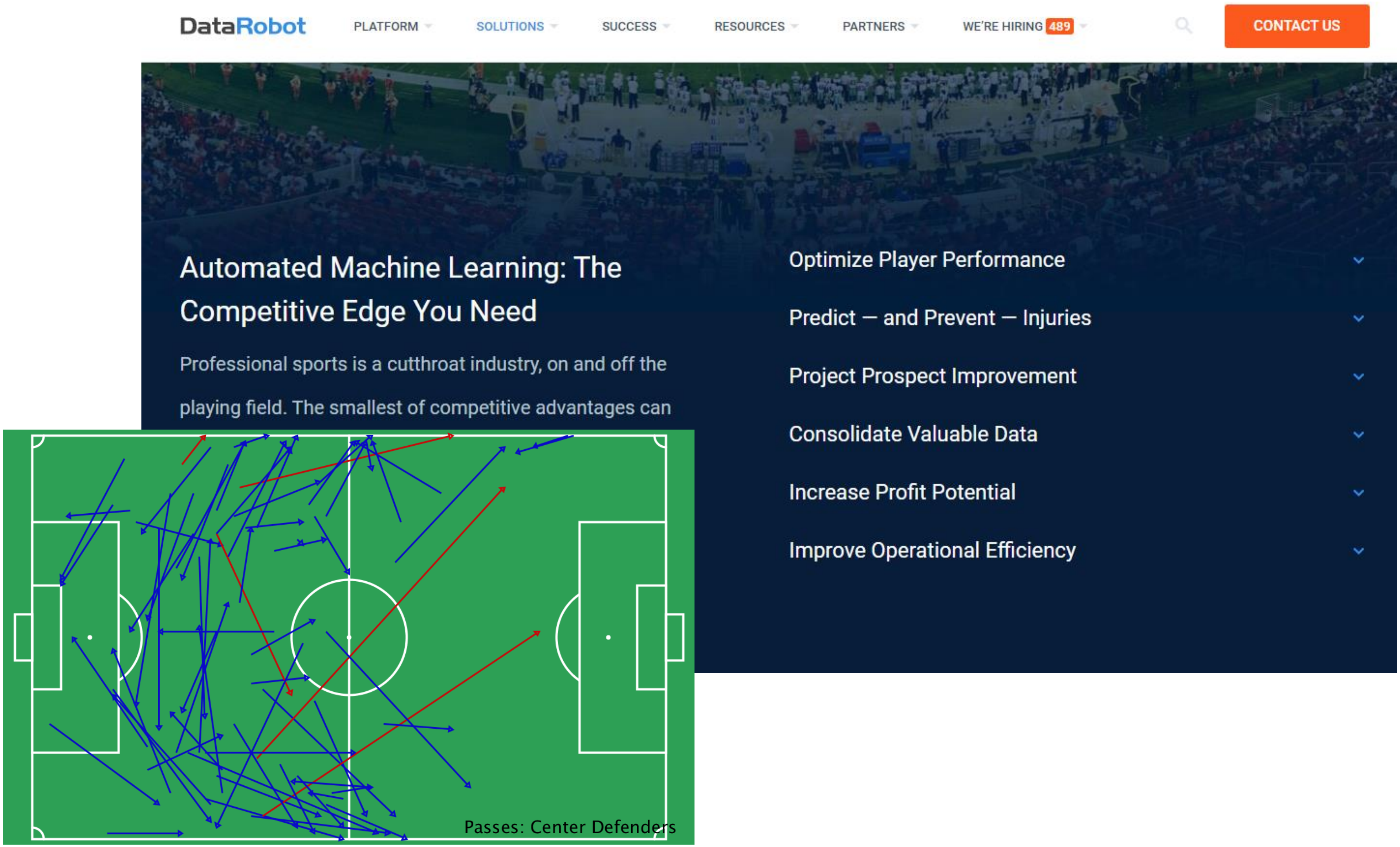
Login	Non- Transactional Activities		Transactio
• Challenges	• View balance	• Add new user	• ACH
• Device	• View history	• Change limits	• Wire
• Cookie	• Updated address	• Set up batch	• Bill Pay
• IP Address	• Update email	• Set up template	• Loan Draw
• Time of day	• Update password	• Add payees	
• Network			

Introduction to AI and ML



From towardsdatascience.com

Introduction to AI and ML



DataRobot PLATFORM SOLUTIONS SUCCESS RESOURCES PARTNERS WE'RE HIRING 489 CONTACT US

Automated Machine Learning: The Competitive Edge You Need

Professional sports is a cutthroat industry, on and off the playing field. The smallest of competitive advantages can

Optimize Player Performance

Predict — and Prevent — Injuries

Project Prospect Improvement

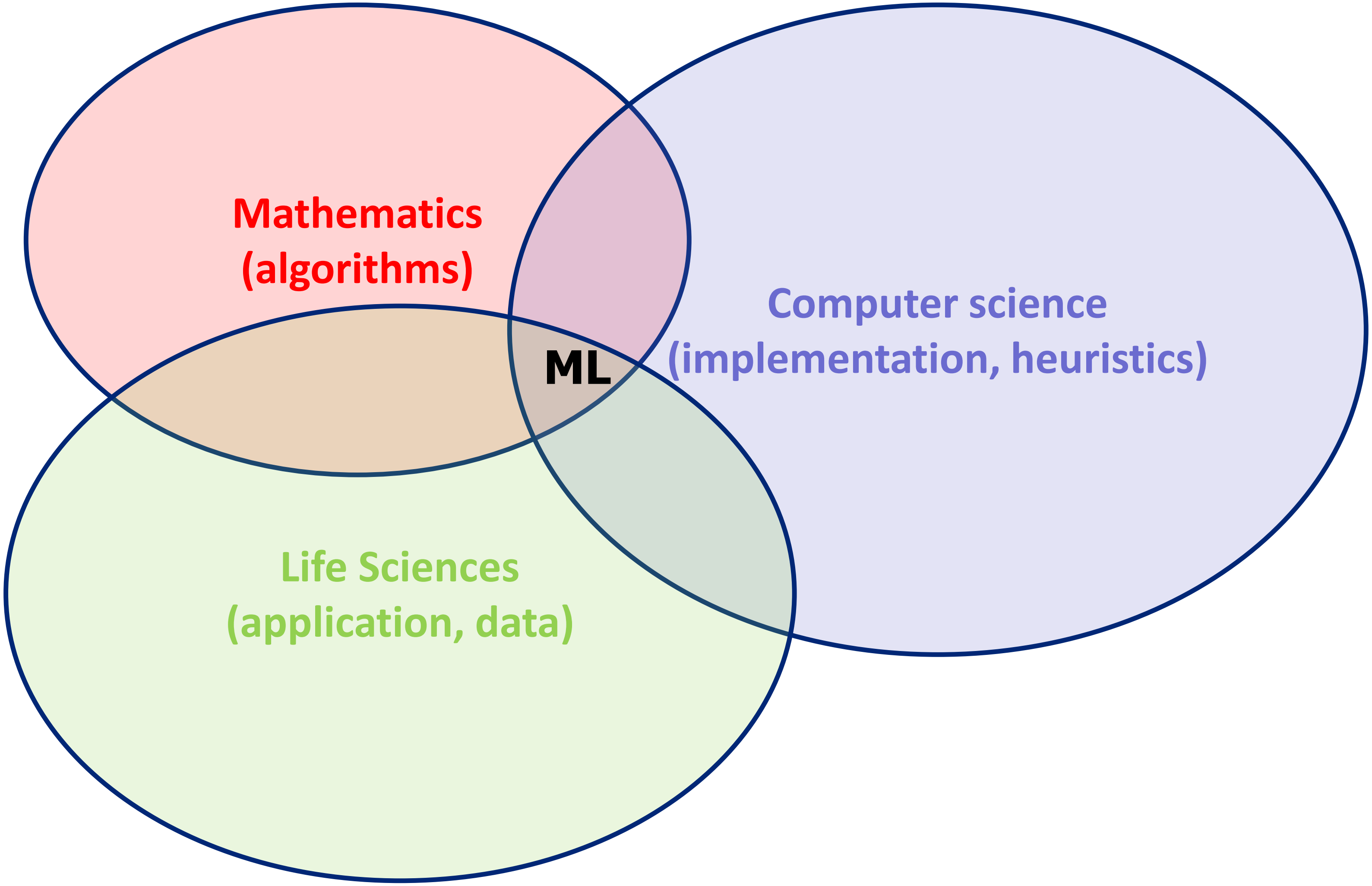
Consolidate Valuable Data

Increase Profit Potential

Improve Operational Efficiency

From www.analyticsindiamag.com/data-mining-strategy-development-football/

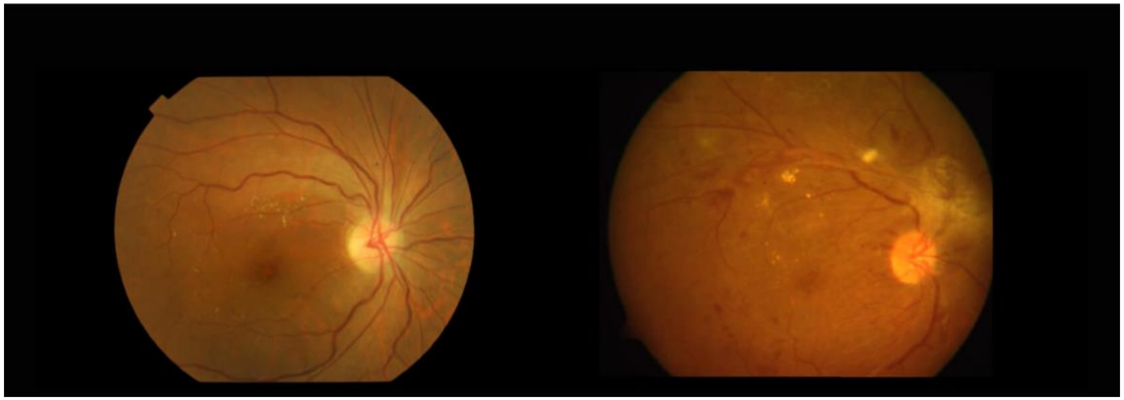
Introduction to AI and ML

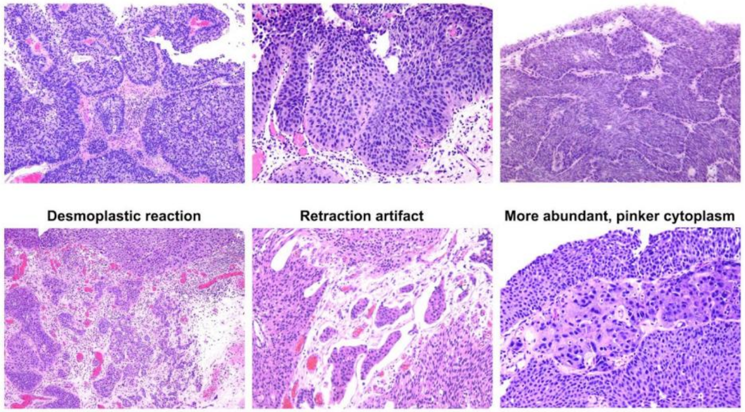


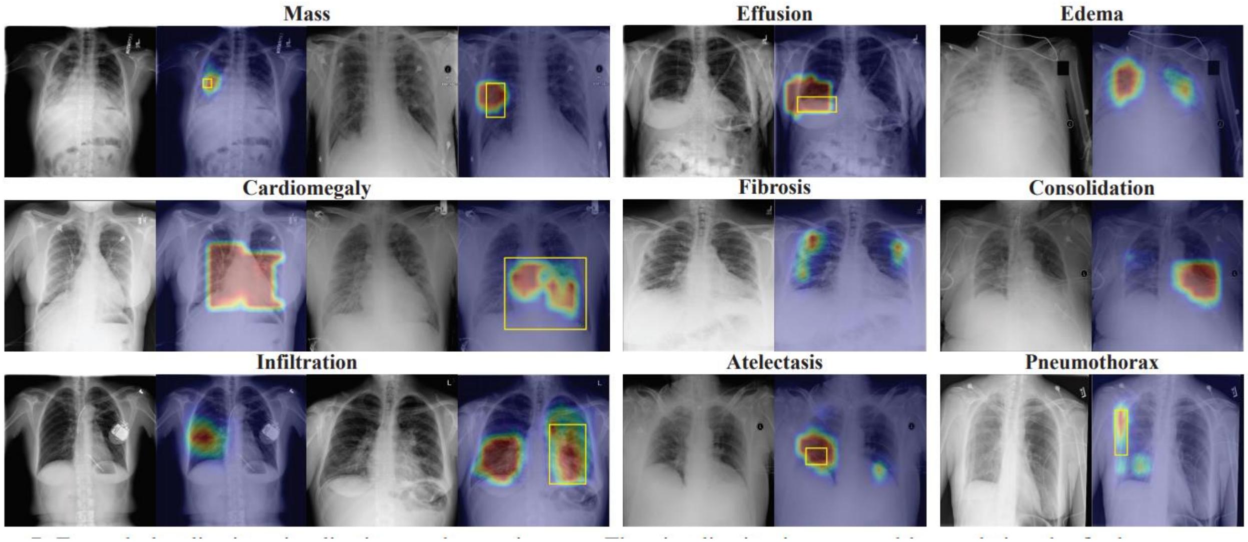
Modern challenges

in Bioengineering

Modern Challenges

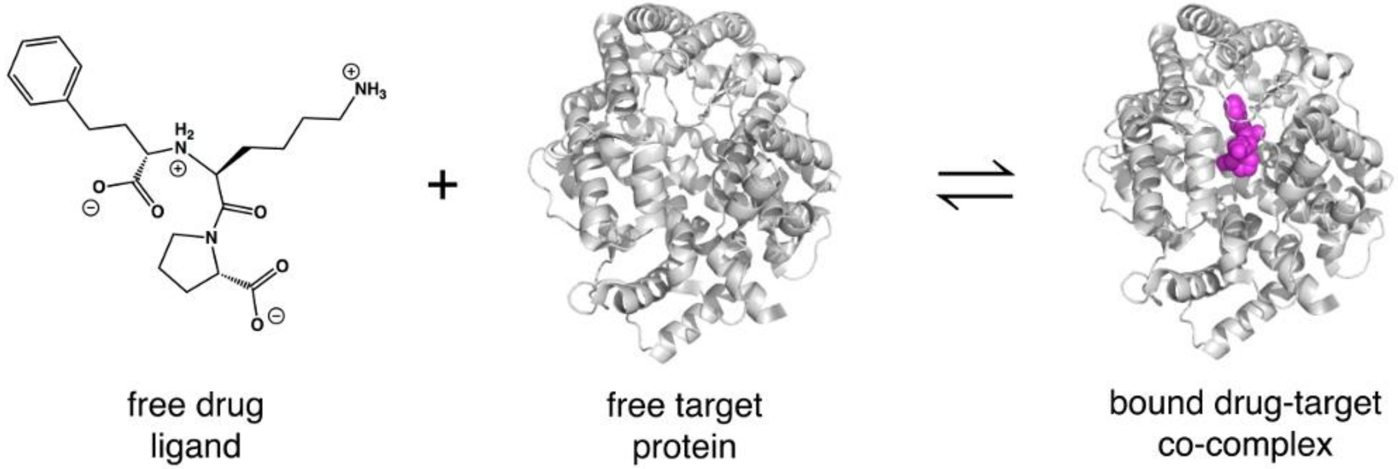


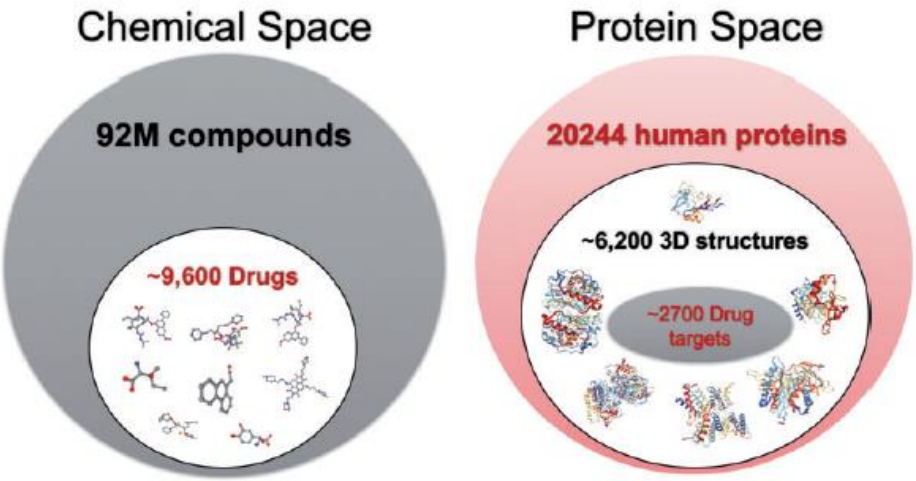




From ai.google/healthcare/

Modern Challenges

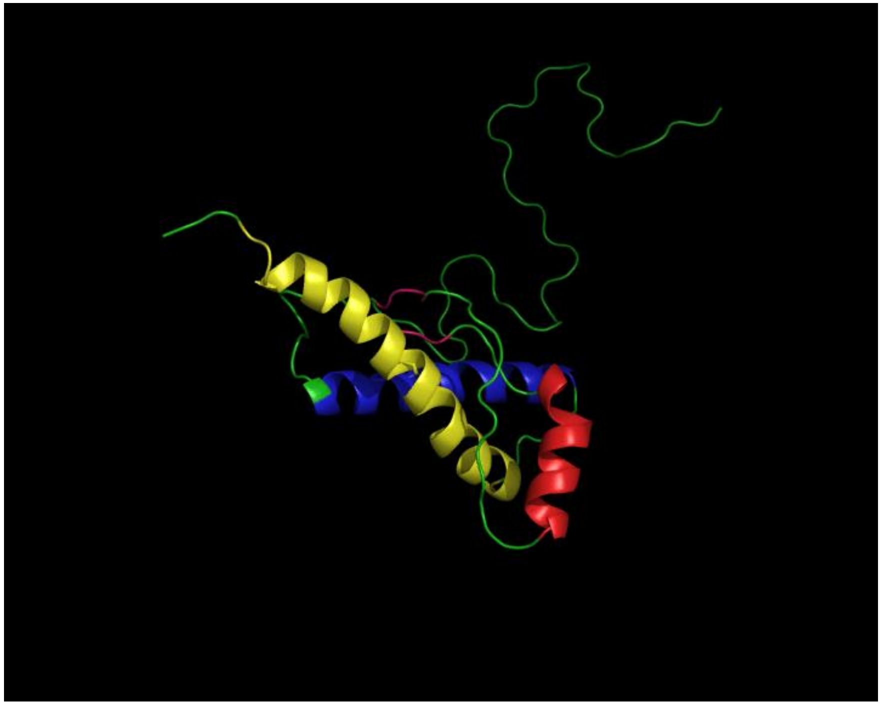




From C4X Discovery

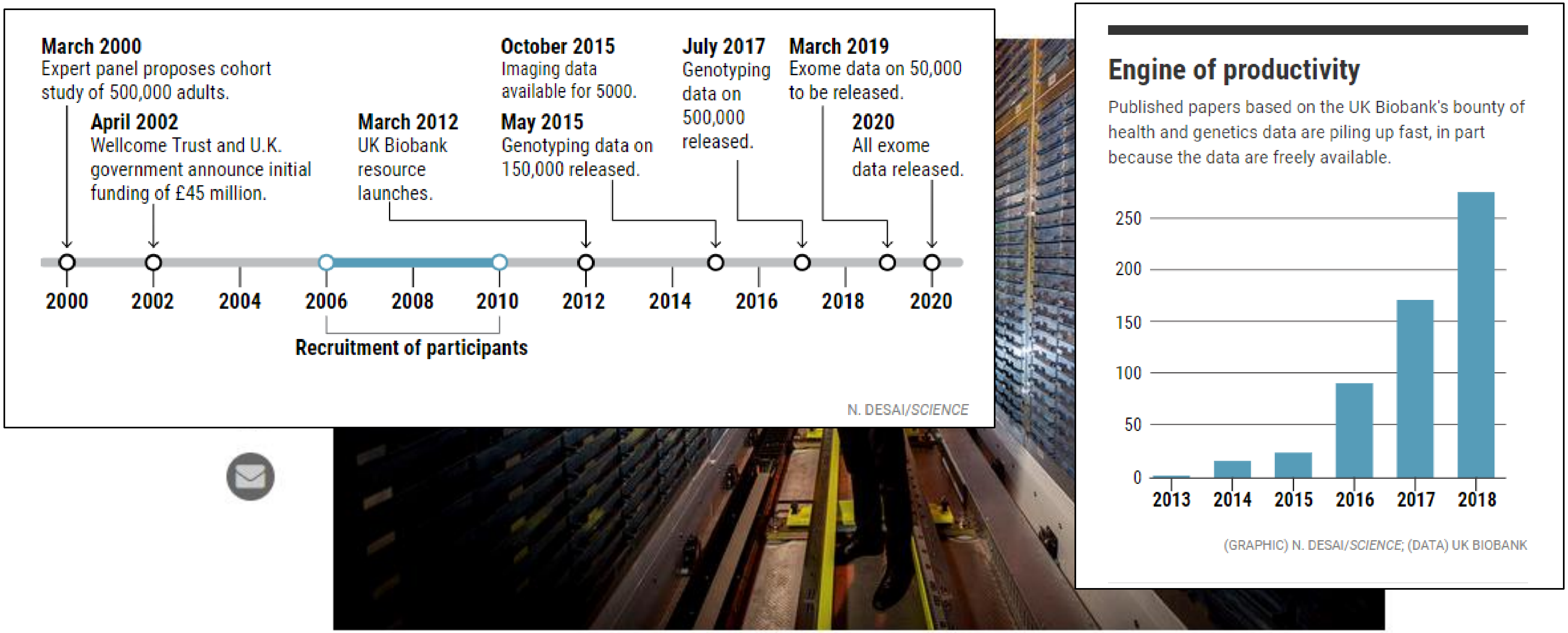
Modern Challenges

- **How to annotate:**
  - **protein structure**
  - **protein function**
  - **protein interactions**
  - **…**
- **How to increase:**
  - **enzymatic activity**
  - **stability & solubility**
  - **substrate specificity**
  - **enantioselectivity**



0.2 ns

Modern Challenges



March 2000
Expert panel proposes cohort study of 500,000 adults.

April 2002
Wellcome Trust and U.K. government announce initial funding of £45 million.

March 2012
UK Biobank resource launches.

May 2015
Genotyping data on 150,000 released.

October 2015
Imaging data available for 5000.

July 2017
Genotyping data on 500,000 released.

March 2019
Exome data on 50,000 to be released.

2020
All exome data released.

N. DESAI/*SCIENCE*

Engine of productivity

Published papers based on the UK Biobank's bounty of health and genetics data are piling up fast, in part because the data are freely available.

(GRAPHIC) N. DESAI/*SCIENCE*; (DATA) UK BIOBANK

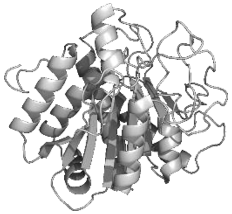
UK Biobank Principal Investigator Rory Collins stands amid stored biospecimens from the project's half-million participants. NIGEL HILIER

Huge trove of British biodata is unlocking secrets of depression, sexual orientation, and more

By Jocelyn Kaiser, Ann Gibbons | Jan. 3, 2019 , 1:20 PM

From sciencemag.org

Modern Challenges



1. A large number of relevant features;
   - From bio-physico-chemical to textual;
   - Heterogeneous data (e.g. clinical, imaging, and genomic data);

2. Complex tasks and large parameter space;
   - A single 300-amino-acid-long protein will have 300·19=5700 single-point variants!

3. Large datasets available and new data are collected.

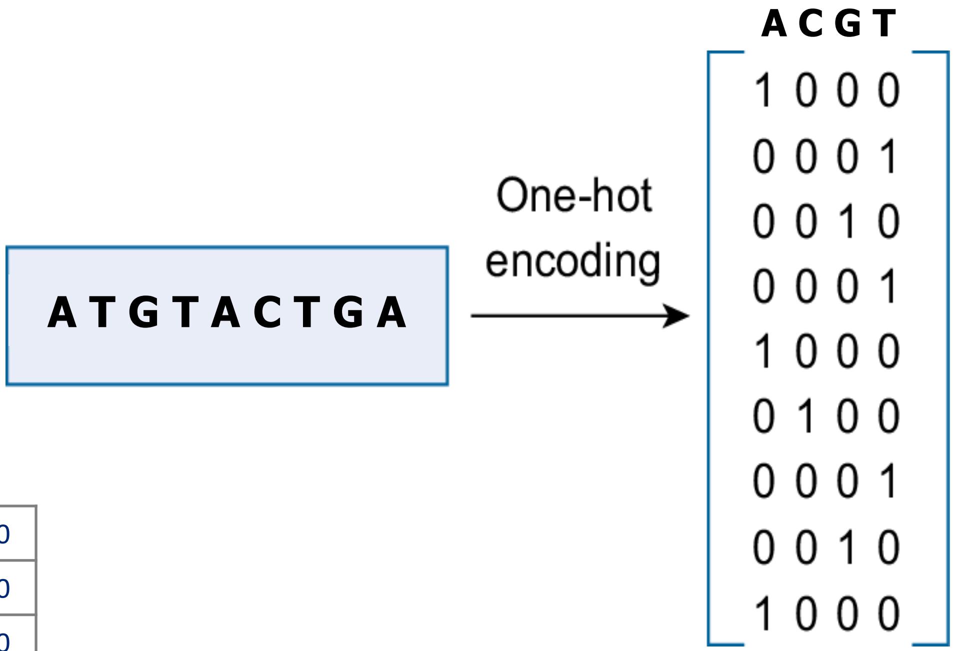
Ideal ML setup = complex tasks + relevant features + abundant data

Basics of ML

Basics of ML: features

- **Feature vector:** $x = (x_1, x_2, \dots, x_n)$
- **All features must be converted to numbers:**

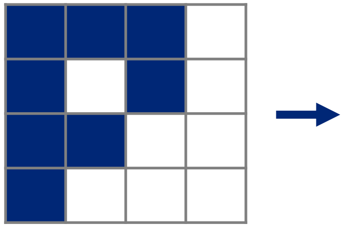
yes =1, no = 0



ATGTACTGA

One-hot encoding

$$\begin{array}{cccc} A & C & G & T \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{array}$$



1	1	1	0
1	0	1	0
1	1	0	0
1	0	0	0

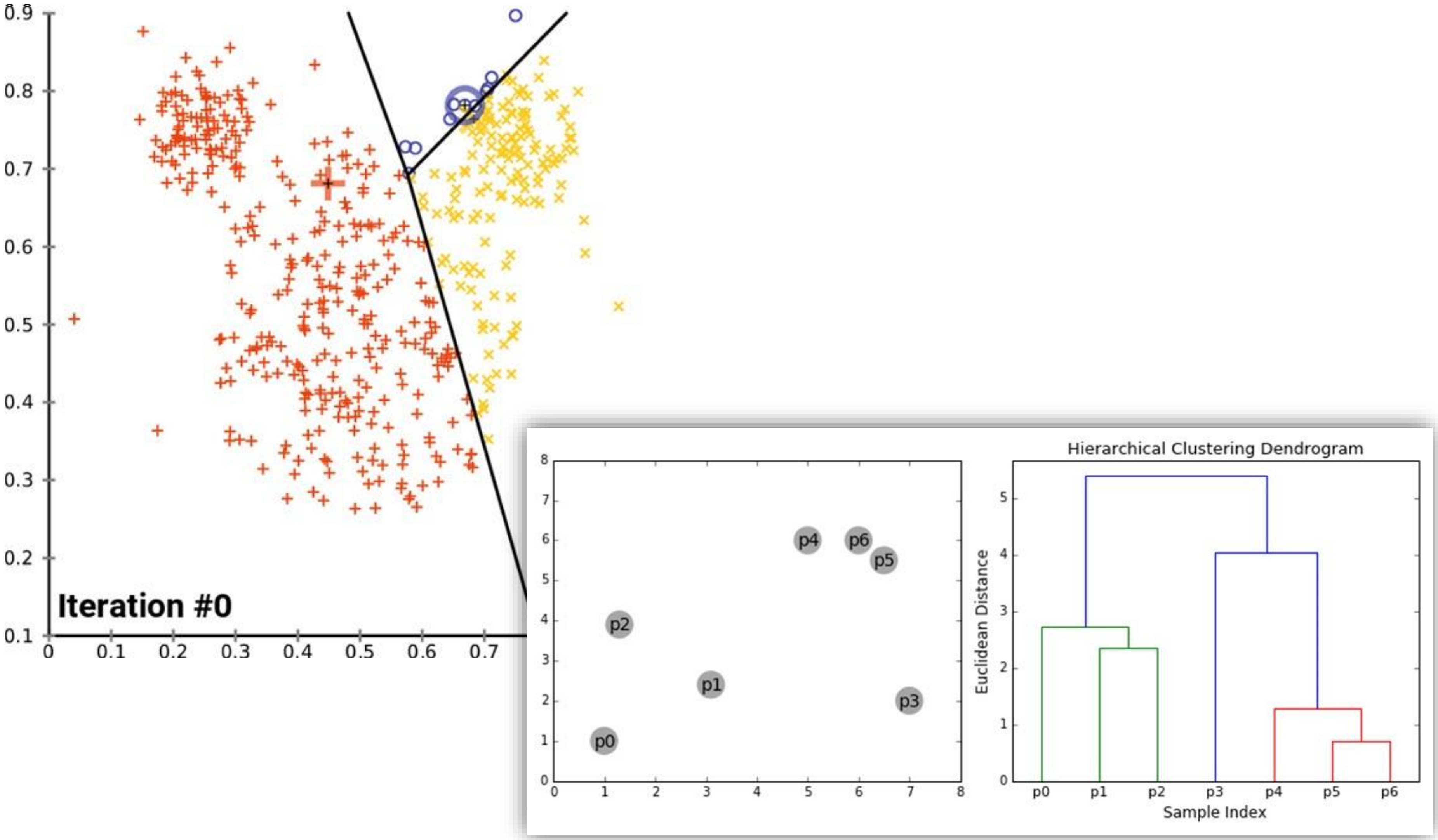
Basics of ML: types

- Unsupervised learning
  - only features are available;
  - goal: cluster the data or reduce their dimensionality;

- Supervised learning
  - features and labels are available;
  - goal: learn to predict the label based on the features;

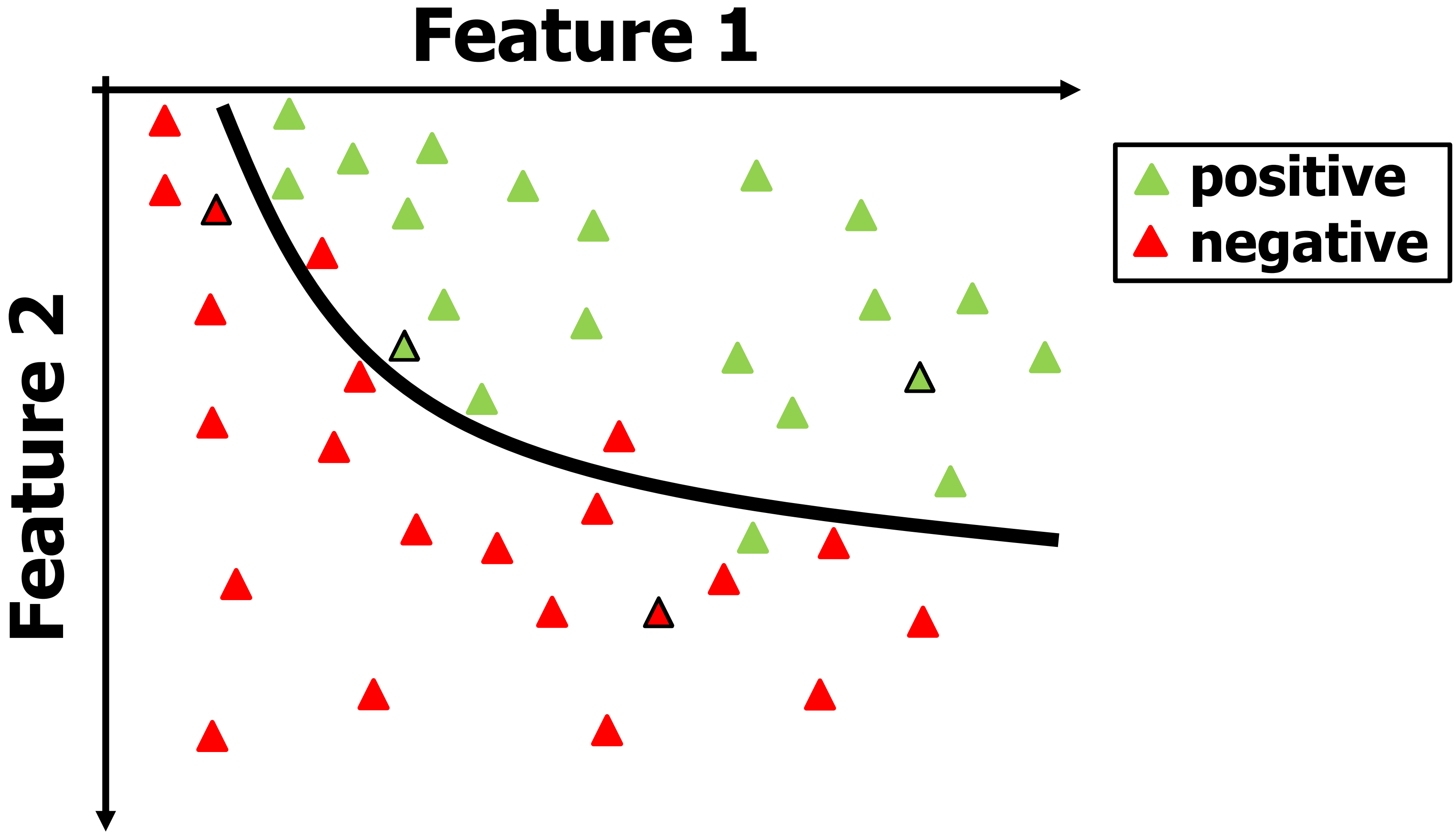
- Reinforcement learning, feature learning, anomaly detection, etc…

Basics of ML: unsupervised learning



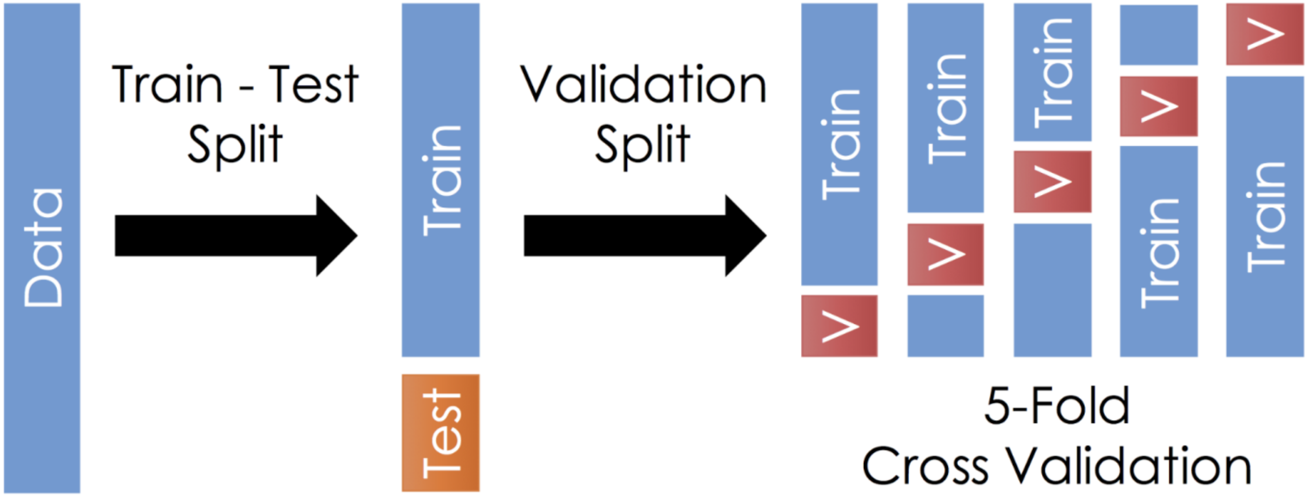
From en.wikipedia.org/wiki/K-means_clustering

Basics of ML: supervised learning



Basics of ML: supervised learning

- We want ML models to be generalizable = good at predicting labels for previously unseen data;
- It is essential to split the data into training set and test sets and use the latter for final evaluation only!
- To fine-tune an algorithm, K-fold cross validation is implemented:



Basics of ML: supervised learning

- Let $\left(x^{(i)}, y^{(i)}\right)$ be our data set, where x are feature values, and y are labels;

- Any ML predictor is fundamentally a function $f(x)$:

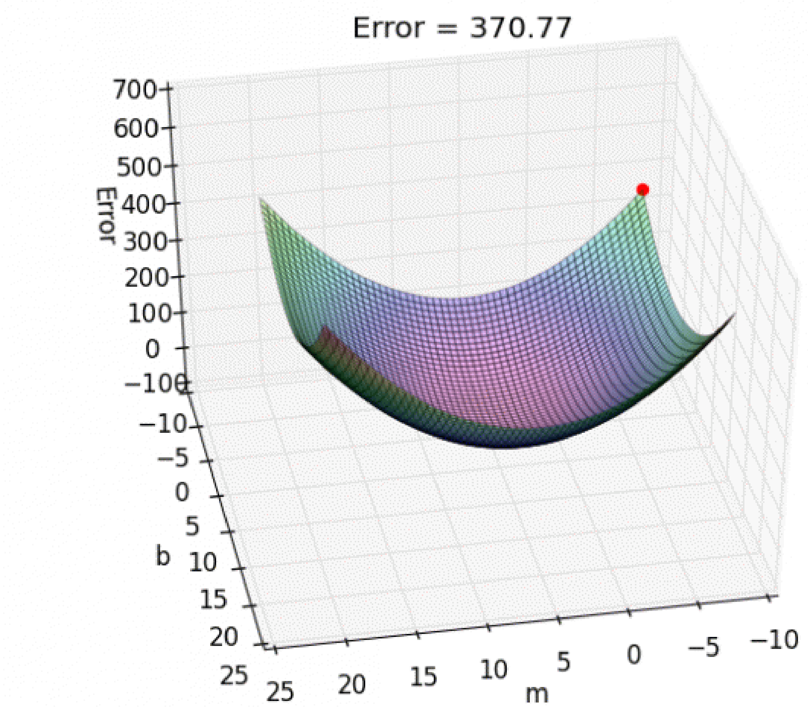
$$f(\text{feature values}) = \text{label}$$

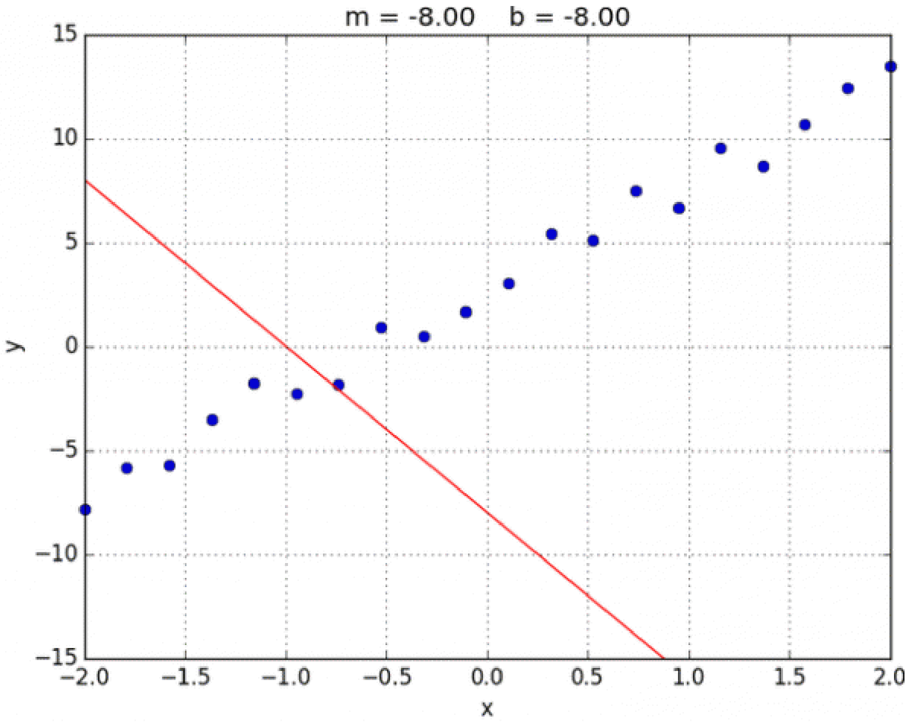
- Usually, a generic group of functions $f(x, \beta)$ is chosen, where β is a set of parameters;

- Then we “train” the ML predictor: pick such β^* that $f\left(x^{(i)}, \beta^*\right)$ is as close to $y^{(i)}$ as possible.

Basics of ML: supervised learning

$$f(x) = m \cdot x + b$$

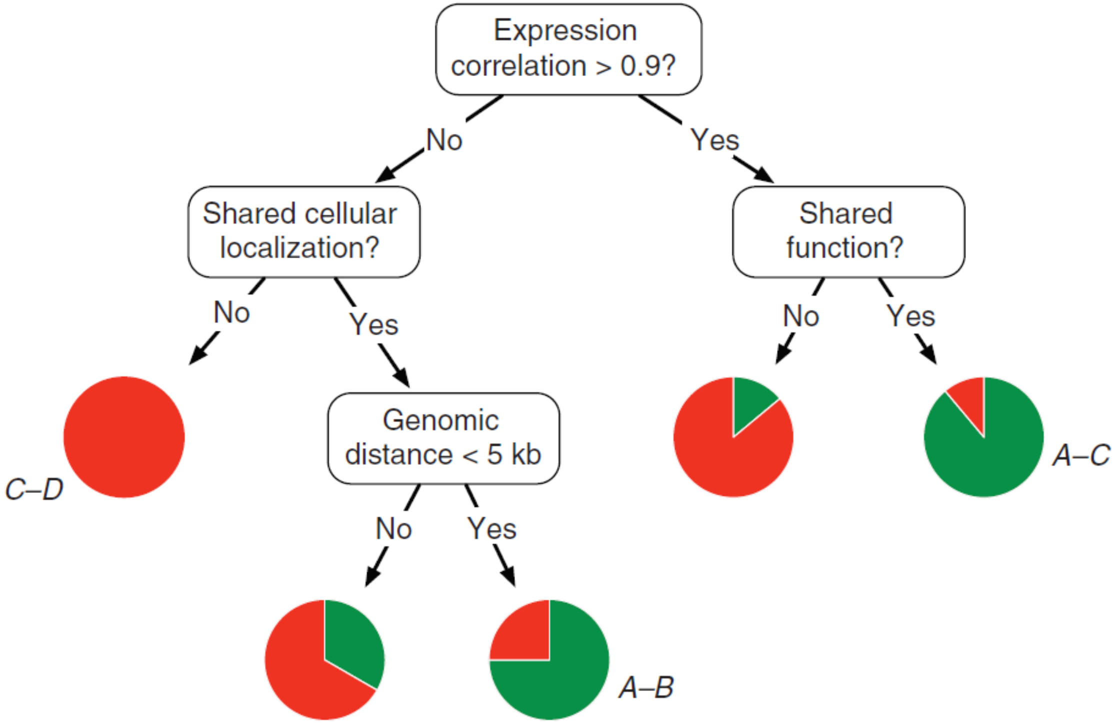




Recent applications

Recent applications: decision tree

Gene Pair	Interact?	Expression correlation	Shared localization?	Shared function?	Genomic distance
A-B	Yes	0.77	Yes	No	1 kb
A-C	Yes	0.91	Yes	Yes	10 kb
C-D	No	0.1	No	No	1 Mb
⋮					



Expression correlation > 0.9?

No

Yes

Shared cellular localization?

Shared function?

No

Yes

No

Yes

C–D

Genomic distance < 5 kb

A–C

No

Yes

A–B

Figure 1 A hypothetical example of how a decision tree might predict protein-protein interactions.

Kingsford, Carl, and Steven L. Salzberg. "What are decision trees?." *Nature biotechnology* 26.9 (2008): 1011.

Recent applications: neural networks

Recent applications: ANNs for drug design

- **QSAR modeling for ligand binding activity;**
- **Structure generation;**
- **Synthetic pathway generation;**

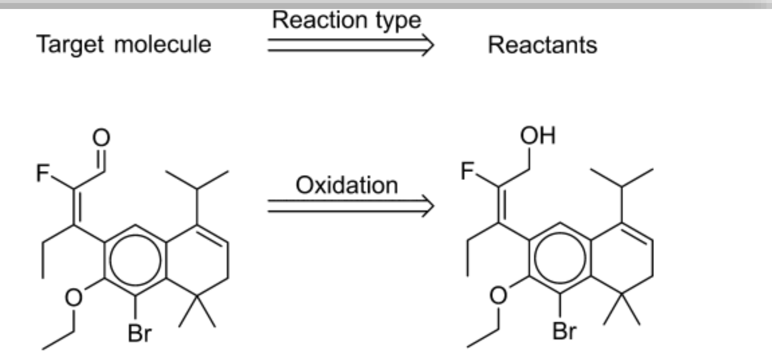


Figure 2. Retrosynthetic reaction prediction task and an example of a possible retrosynthetic disconnection for a target molecule.

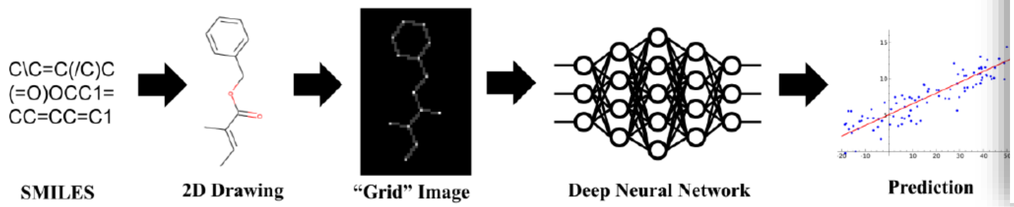
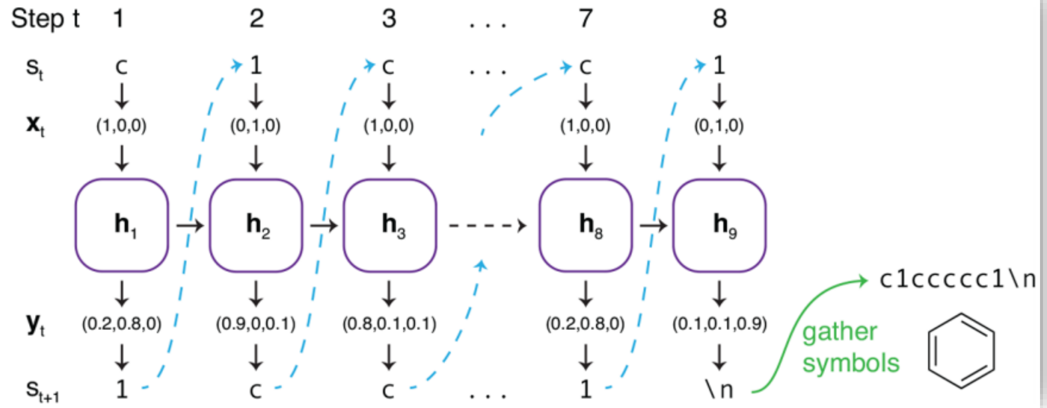
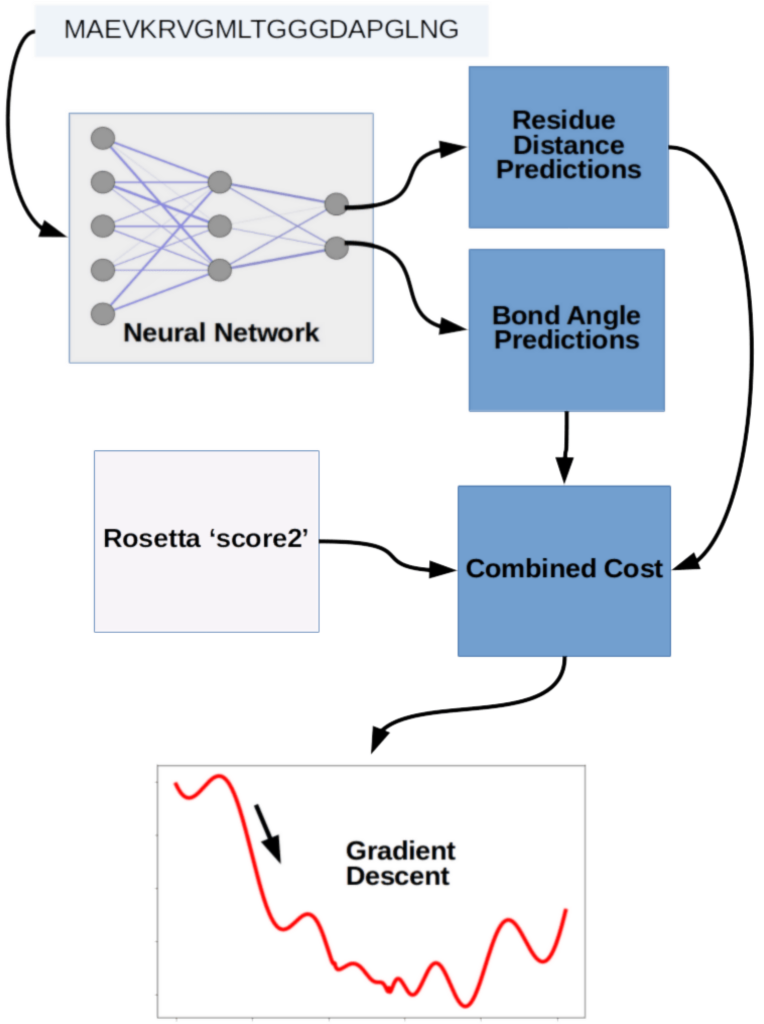


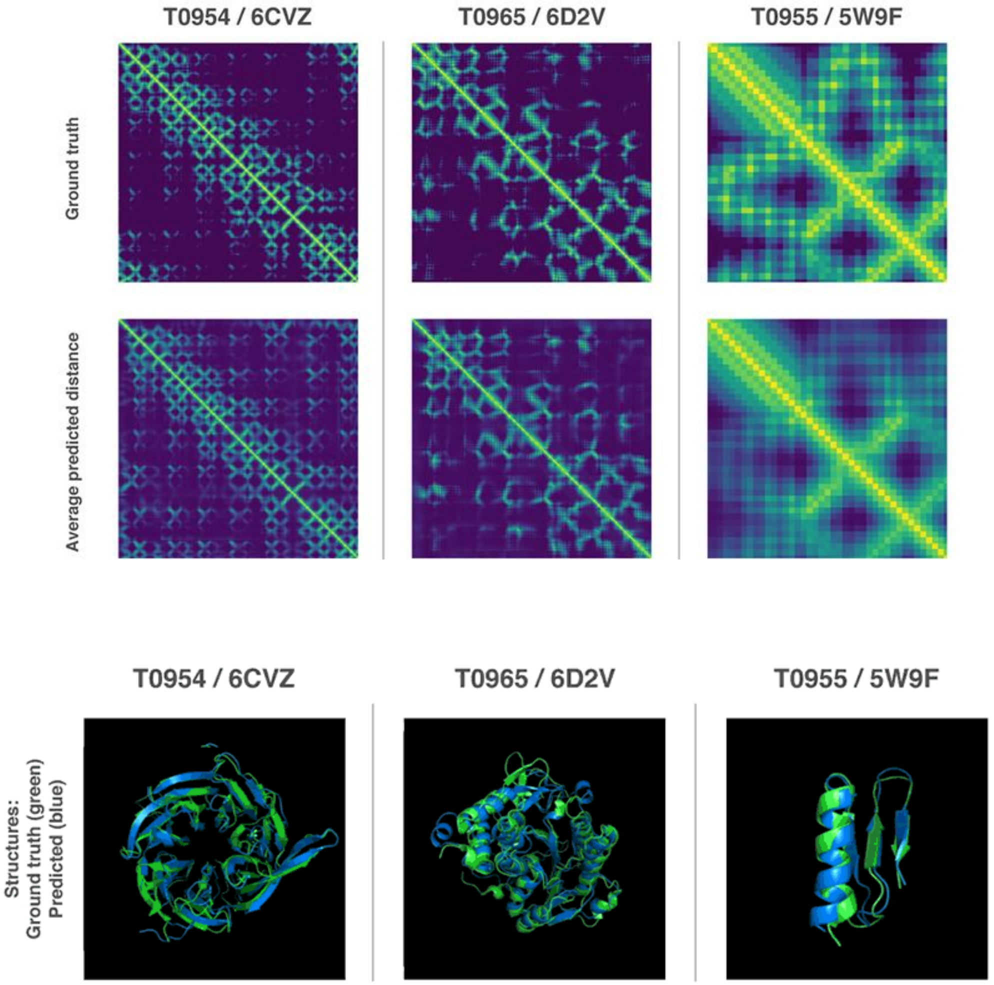
Figure 2: Illustration of the Chemception framework. After a SMILES to structure conversion, the 2D images are mapped onto an 80 x 80 image that serves as the input image data for training a deep neural network to predict toxicity, activity, and solvation properties.



Goh et al. 2017 Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models
Segler et al. 2018 Planning chemical syntheses with deep neural networks and symbolic AI. Nature

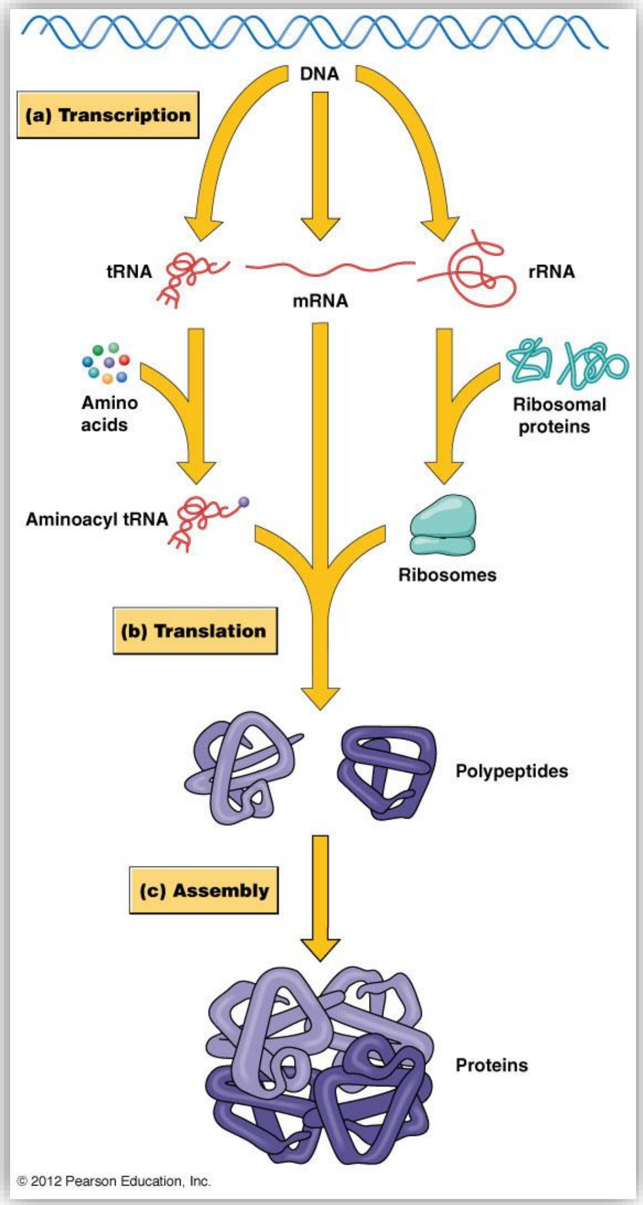
Recent applications: AlphaFoldfor protein folding

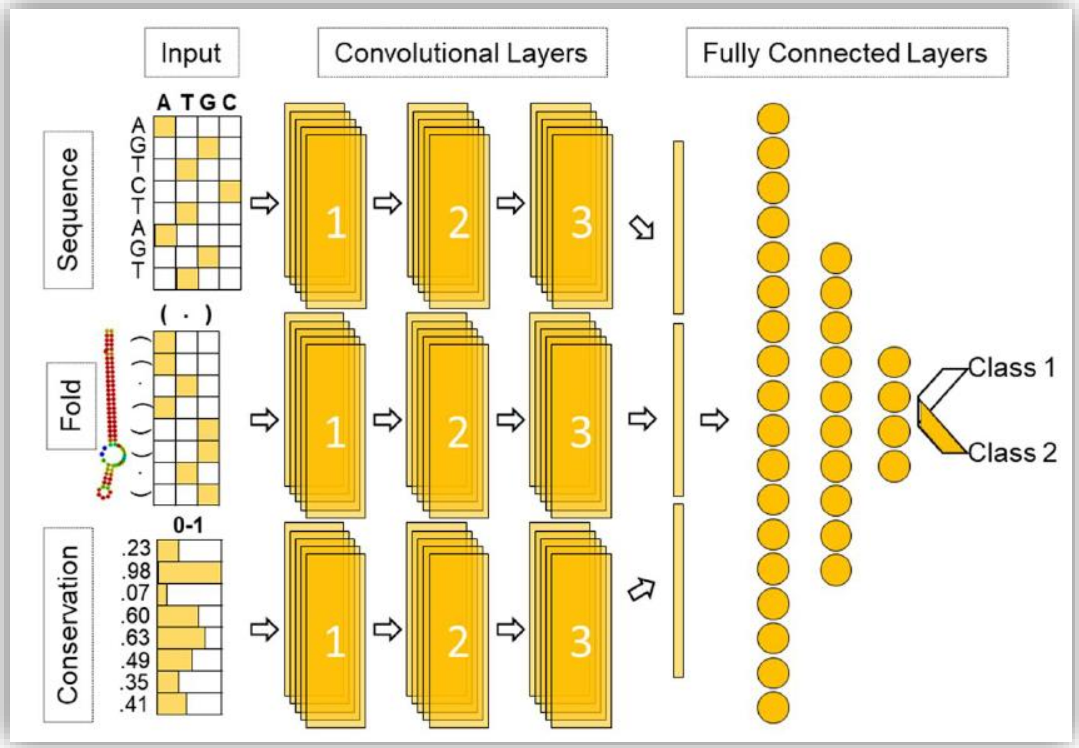




https://deepmind.com/blog/article/alphafold
https://dzone.com/articles/a-summary-of-deepminds-protein-folding-upset-at-ca

Recent applications: MuStARD for genome annotation



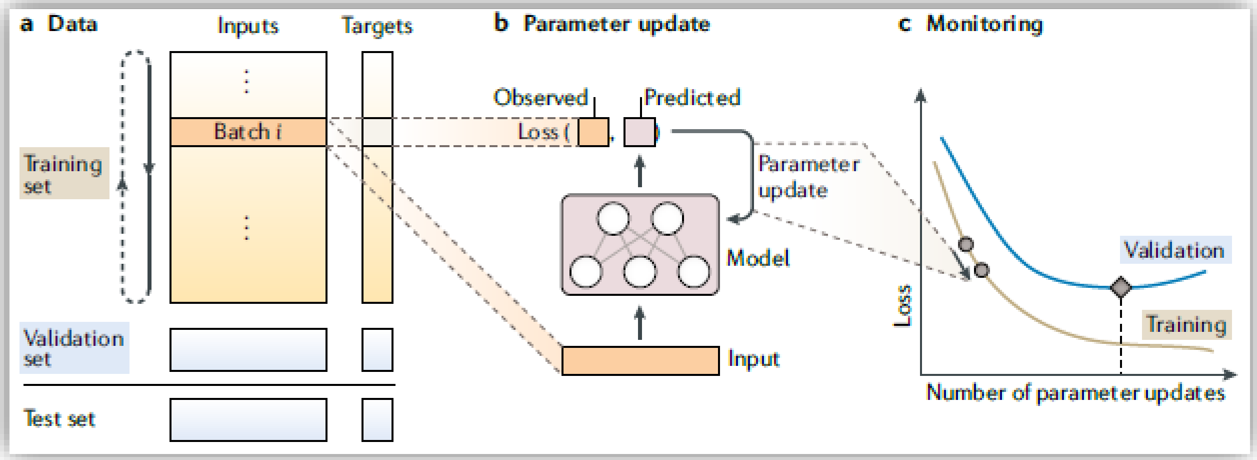


http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-21/CB21.html

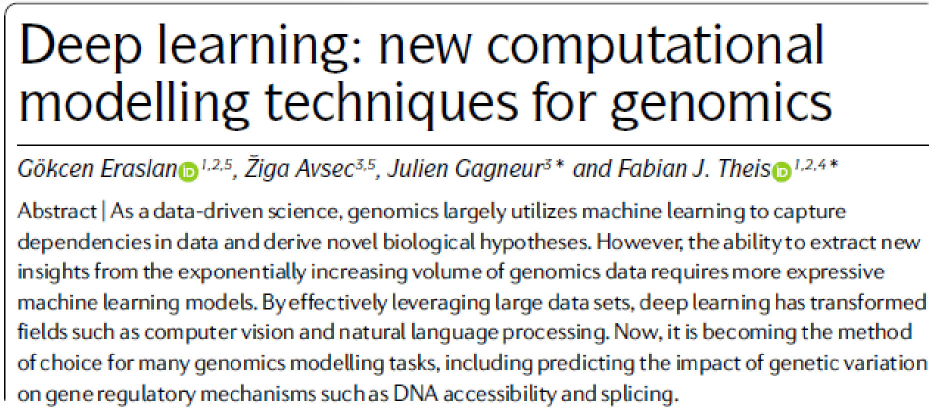
Georgakilas, GK et al. "MuStARD: a Deep Learning method for intra-and inter-species scanning identification of small RNA molecules." bioRxiv (2019)

Reading

- ❑ Eraslan G et al. 2019 "Deep learning: new computational modelling techniques for genomics." Nature Reviews Genetics. (especially pages 2-6)



REVIEWS



Deep learning: new computational modelling techniques for genomics

Gökcen Eraslan[1,2,5], Žiga Avsec[3,5], Julien Gagneur[3]* and Fabian J. Theis[1,2,4]*

Abstract | As a data-driven science, genomics largely utilizes machine learning to capture dependencies in data and derive novel biological hypotheses. However, the ability to extract new insights from the exponentially increasing volume of genomics data requires more expressive machine learning models. By effectively leveraging large data sets, deep learning has transformed fields such as computer vision and natural language processing. Now, it is becoming the method of choice for many genomics modelling tasks, including predicting the impact of genetic variation on gene regulatory mechanisms such as DNA accessibility and splicing.