Lecture 1 – Introduction to Data Science and Machine Learning

ME494 – Data Science and Machine Learning for Mech Engg

Instructor – Subramanian Sankaranarayanan

Teaching Assistants

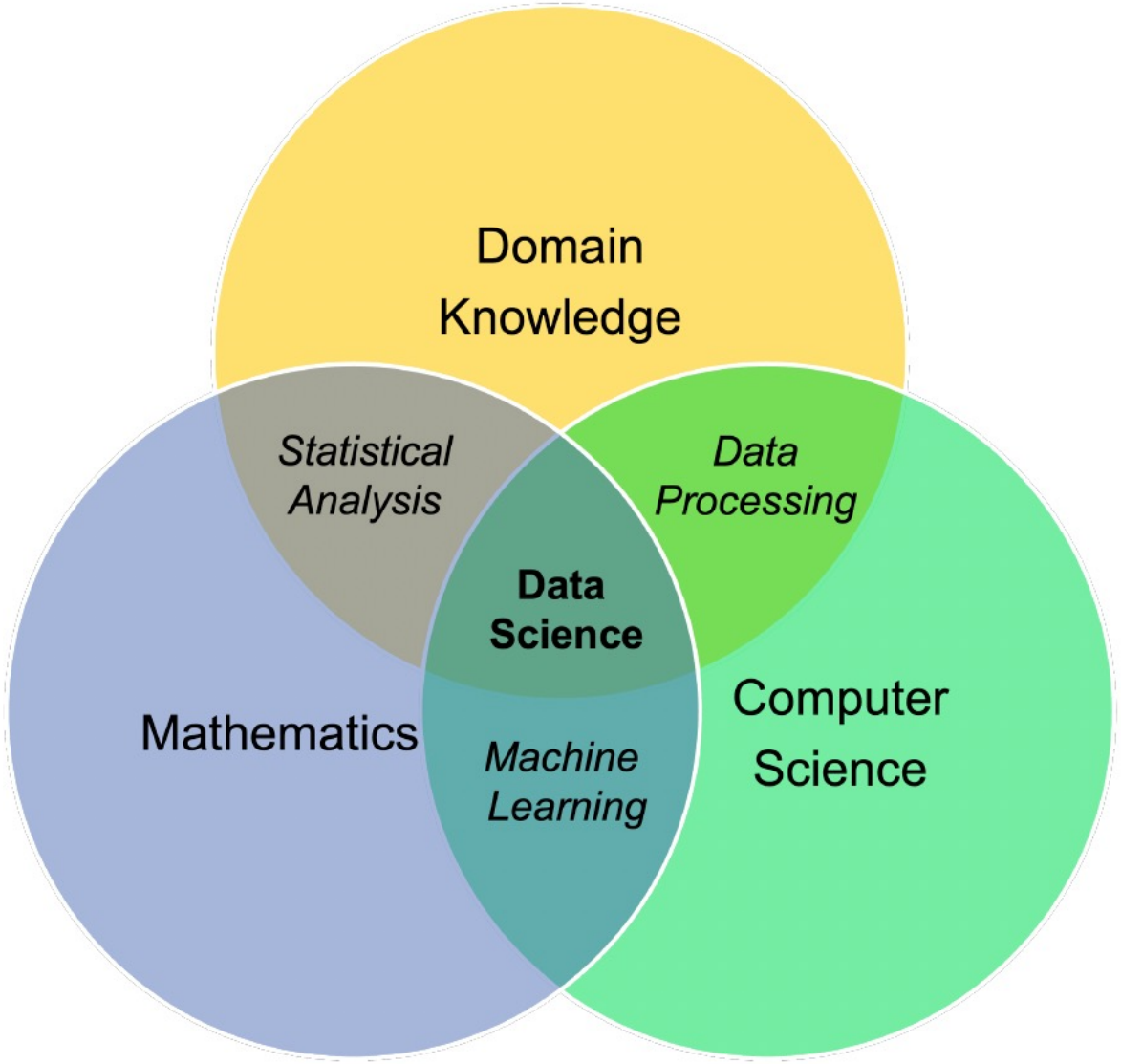
Aditya Koneru (akoner3@uic.edu)

Suvo Banik (sbanik2@uic.edu)

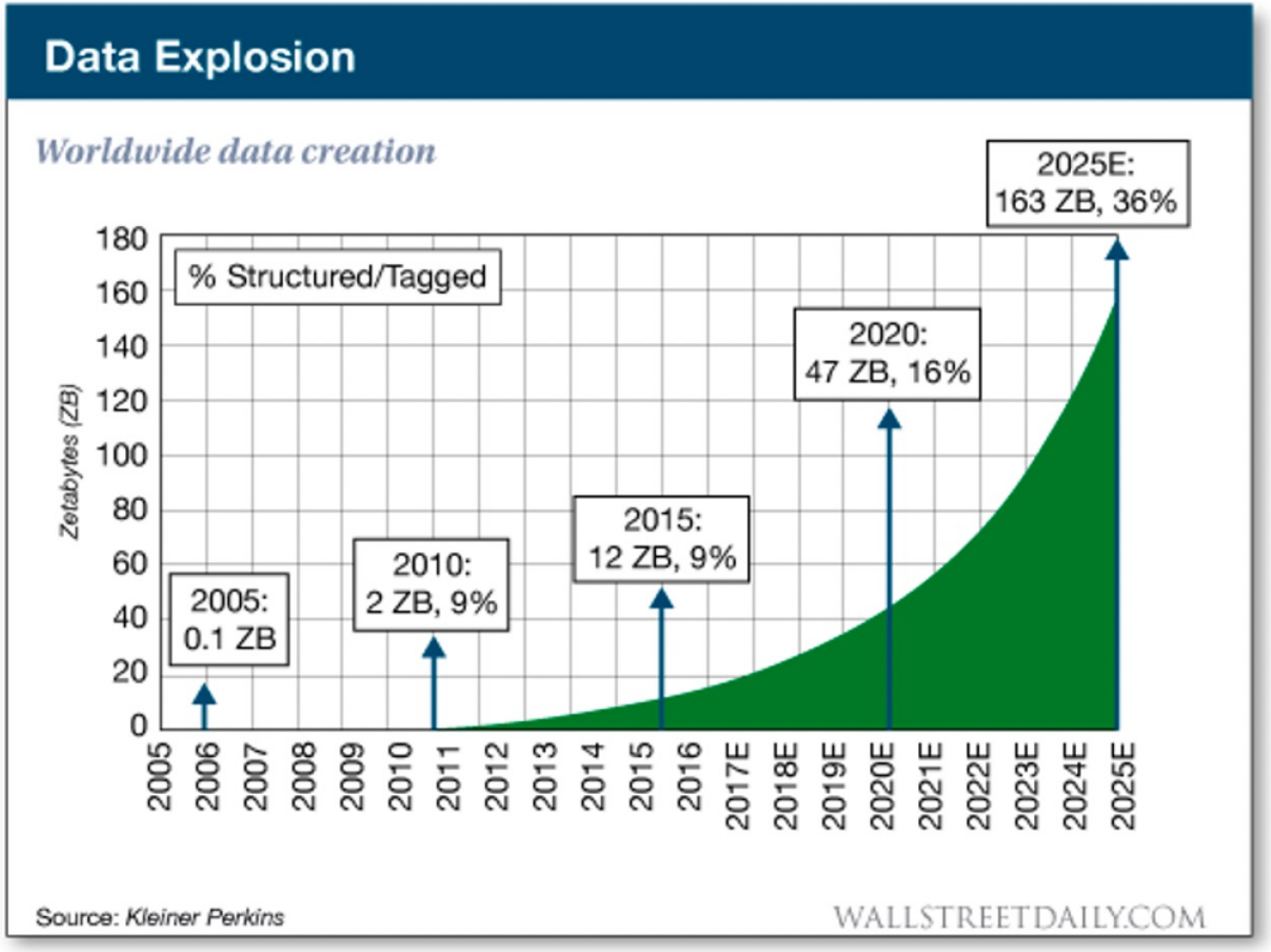
What is Data Science?

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

What is Data Science?



Data is the new gold



Types of materials data

Qualitative data

- Nominal measurement.
- E.g., Metal/Insulator, Stable/Unstable.
- No rank or order.

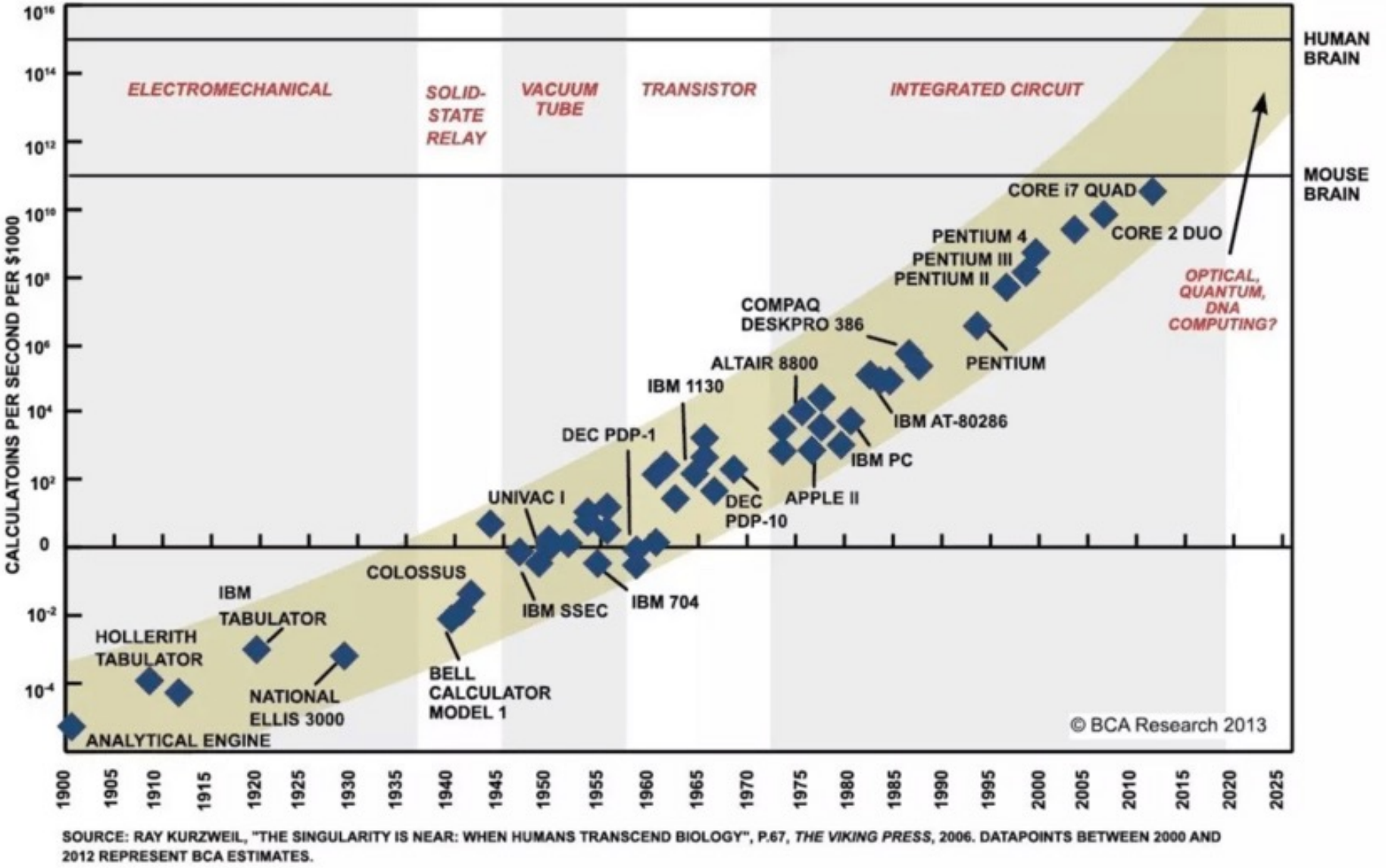
Ranked data

- Ordinal measurement (ordered).
- E.g., Insulator/ semiconductor/ conductor.
- Does not indicate distance between ranks.

Quantitative Data

- Interval/ratio measurement (equal intervals and true 0).
- E.g., melting point, elastic constant, electrical/ionic conductivity.
- Considerable information and permits meaningful arithmetic operations.

What about Computing?



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

Exascale computing

What is machine learning?

"Learning is any process by which a system improves performance from experience."

- Herbert Simon

Economist and Nobel Laureate

Definition by Tom Mitchell (1998):

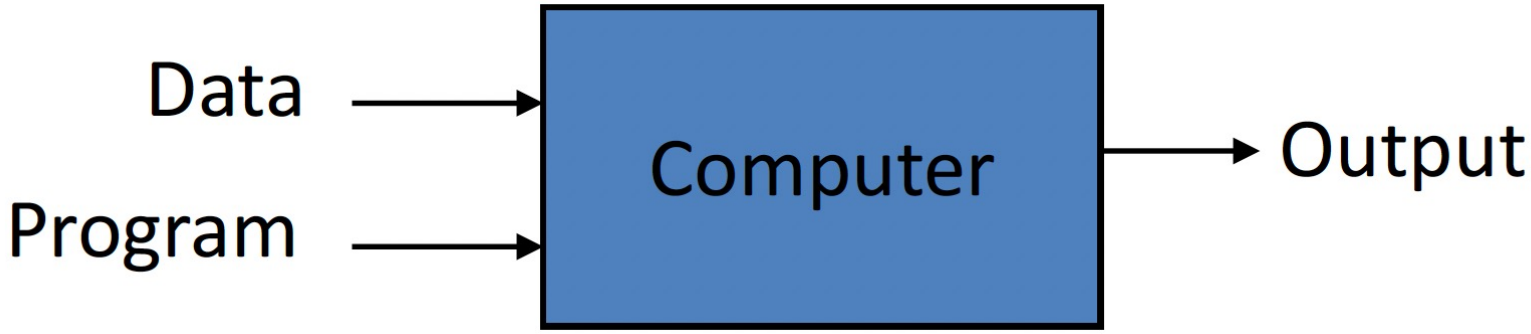
Machine Learning is the study of algorithms that

- improve their performance P
- at some task T
- with experience E.

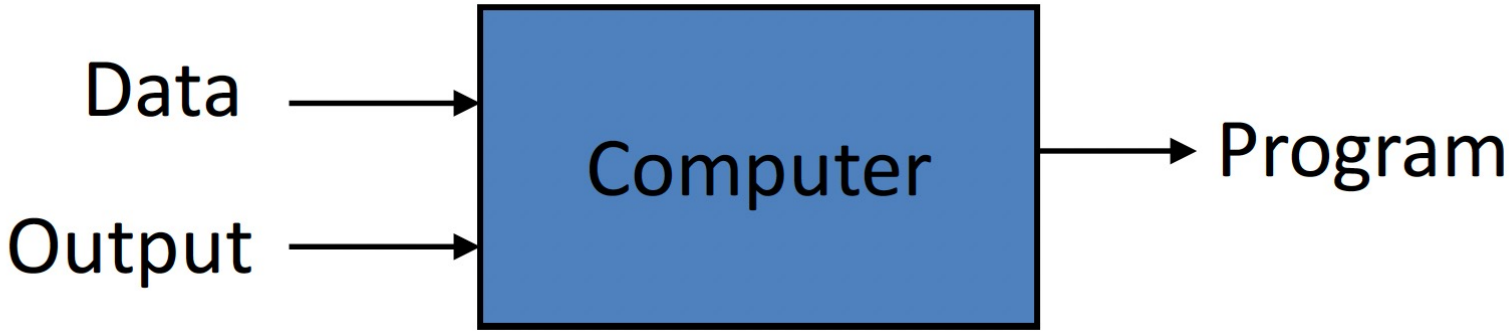
A well-defined learning task is given by <P, T, E>.

Source - **Machine Learning, Tom Mitchell, McGraw Hill, 1997**

Traditional Programming



Machine Learning

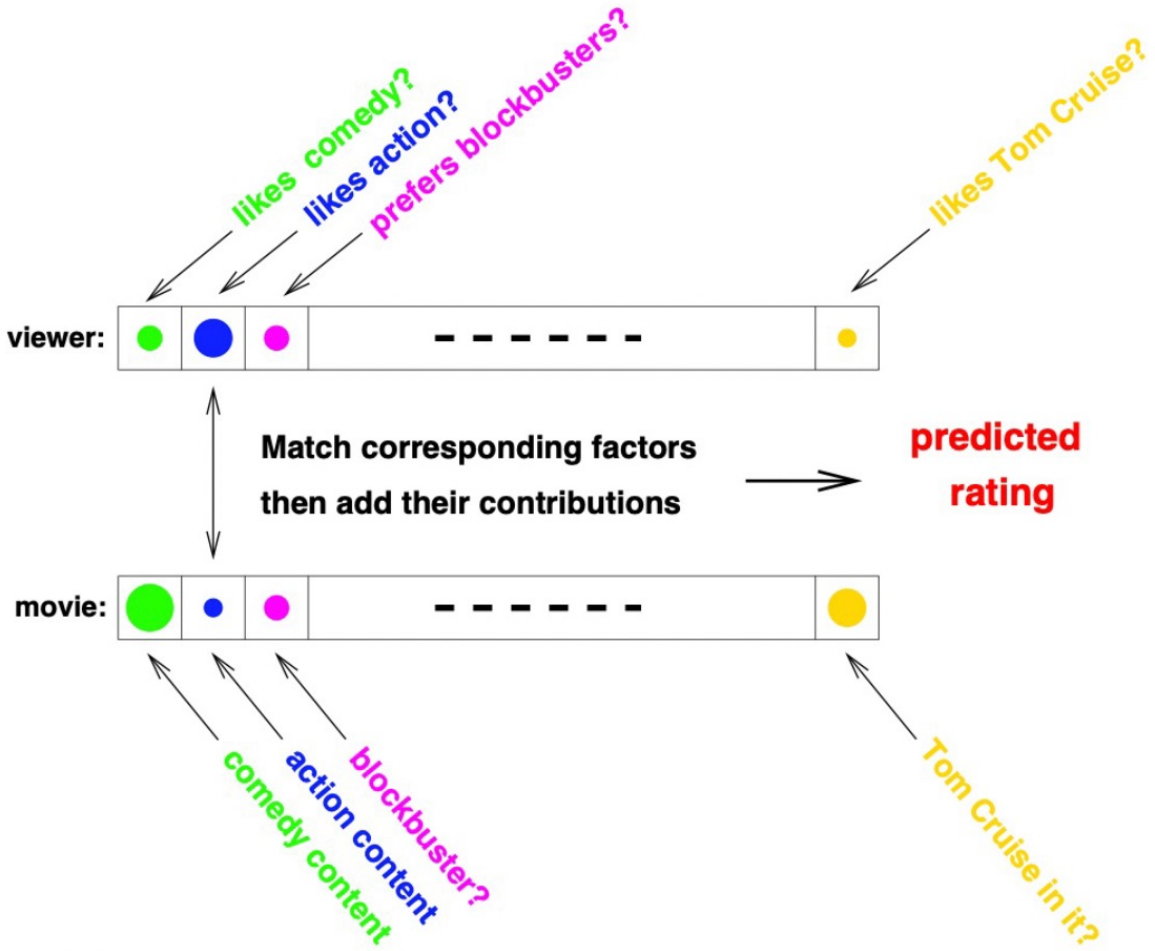


Some more examples of tasks that are best solved by using a learning algorithm

- Recognizing patterns:
  - Facial identities or facial expressions
  - Handwritten or spoken words
  - Medical images
- Generating patterns:
  - Generating images or motion sequences
- Recognizing anomalies:
  - Unusual credit card transactions
  - Unusual patterns of sensor readings in a nuclear power plant
- Prediction:
  - Future stock prices or currency exchange rates

A real life example

Movie Recommendation – The Netflix Problem



Courtesy: Learning from data, Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin

Use data on past user activity to **learn** to identify their preferences

Another Example – Loan Approval

age	32 years
gender	male
salary	40,000
debt	26,000
years in job	1 year
years at home	3 years
...	...

For how much amount should the bank approve the loan?

Courtesy: Learning from data, Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin

Loan Approval as an ML problem

Convert “learning” to a function finding problem

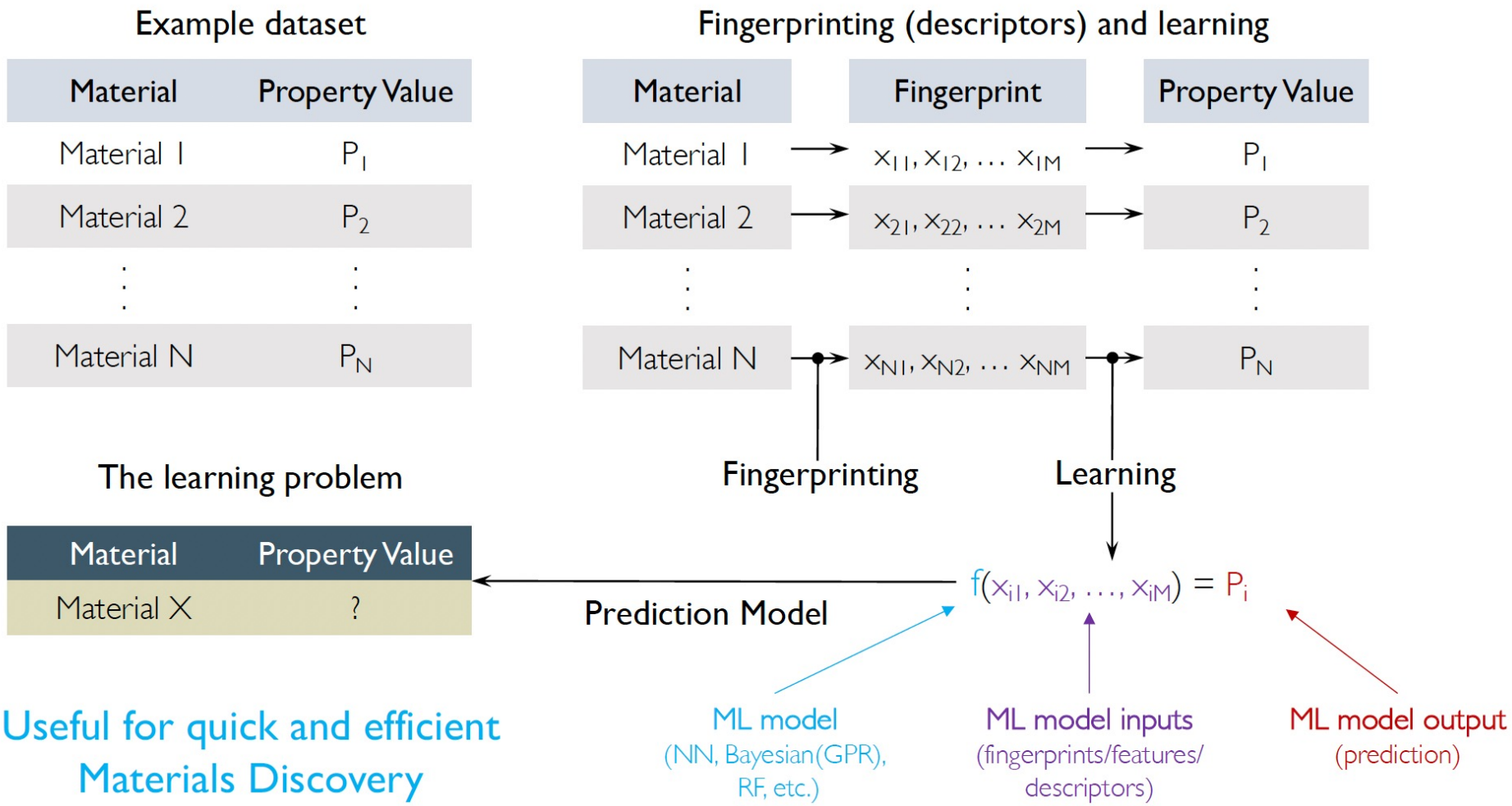
	x_1 (age)	x_2 (gender)	x_3 (salary)	x_4 (debt)	...	$y = f(x_1, x_2, ...)$
Historical data: Person 1	...	...	...	...	...	...
Person 2	...	...	...	...	...	...
...	...	...	...	...	...	...
Person N	...	...	...	...	...	...
Prediction: Person N+1	32	male	40,000	26,000	...	?

Model (hypothesis)

Loan amount for Person N+1 = y_{N+1} = $f(x_{N+1}, x_{N+2}, ...)$

Output (prediction) Input (customer application)

Machine Learning in Materials Science



Useful for quick and efficient Materials Discovery

Types of Learning

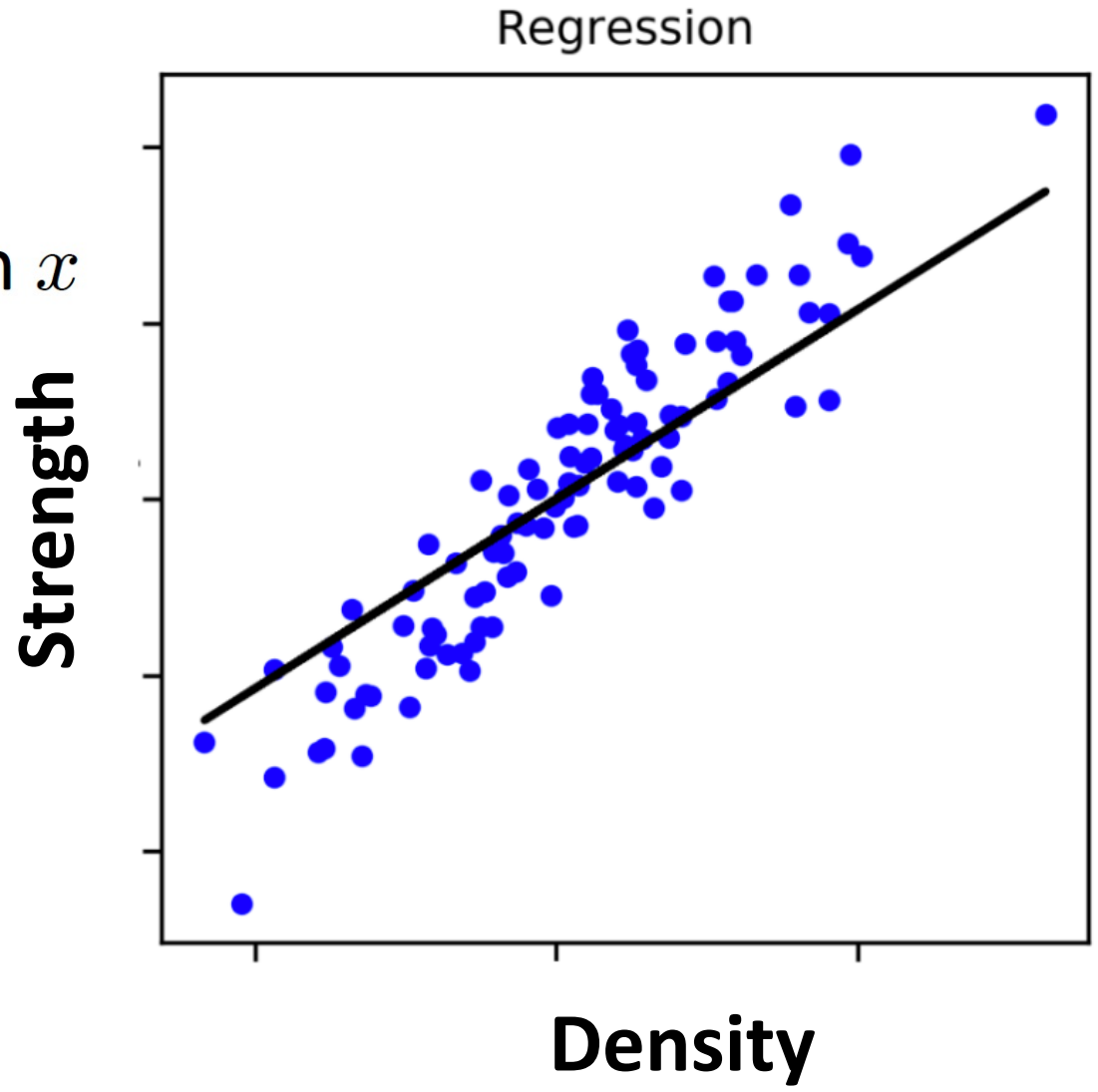
- **Supervised (inductive) learning**
  - Given: training data + desired outputs (labels)
- **Unsupervised learning**
  - Given: training data (without desired outputs)
- **Semi-supervised learning**
  - Given: training data + a few desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions

Supervised Learning: Regression

- Given (x_1, y_1), (x_2, y_2), …, (x_n, y_n)
- Learn a function $f(x)$ to predict y given x
  - y is real-valued == regression

x is the density of the material

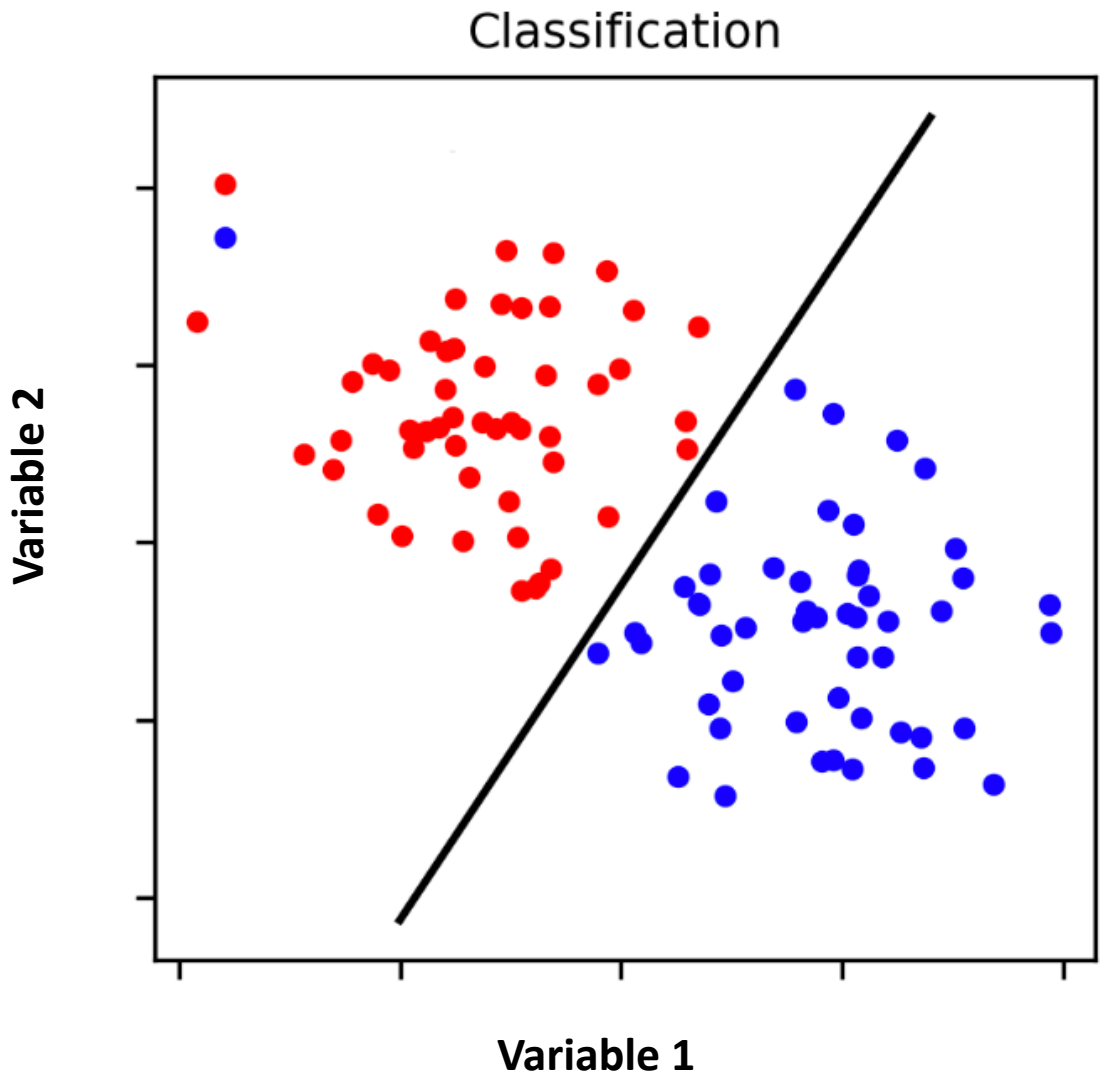
f(x) tells us how the strength varies as a function of density



Supervised Learning: Classification

- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification

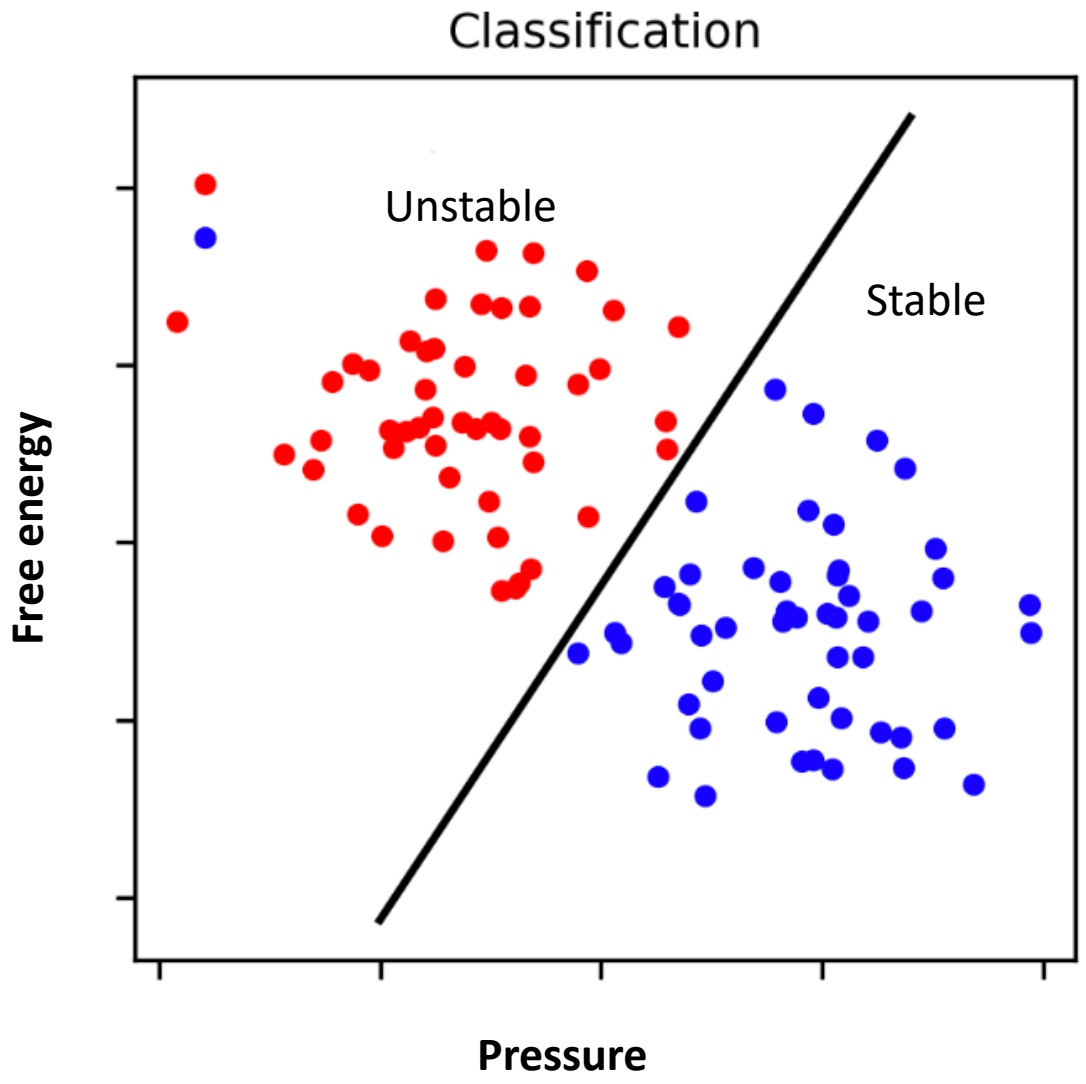
Aim is to predict function f(x) that allows us to classify or separate out dataset into individual classes.



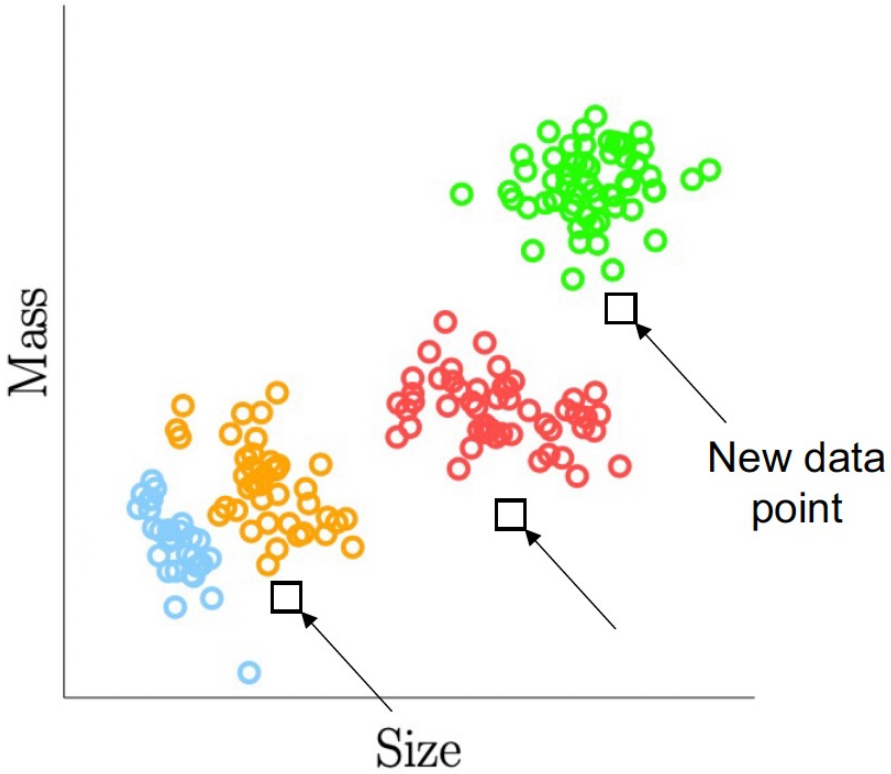
Supervised Learning: Classification

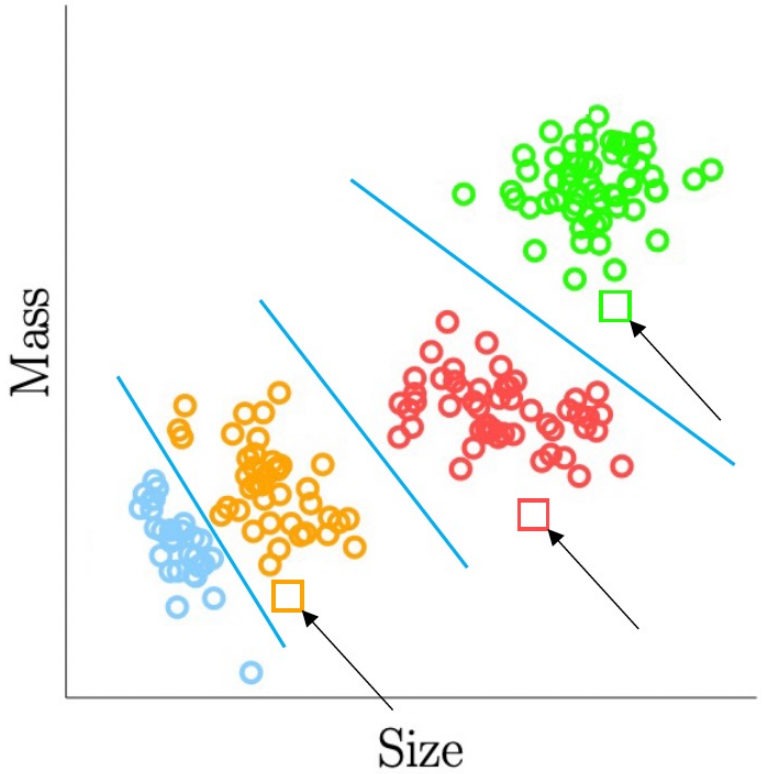
- Given (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification

Phase diagrams of materials, for example



Supervised Learning



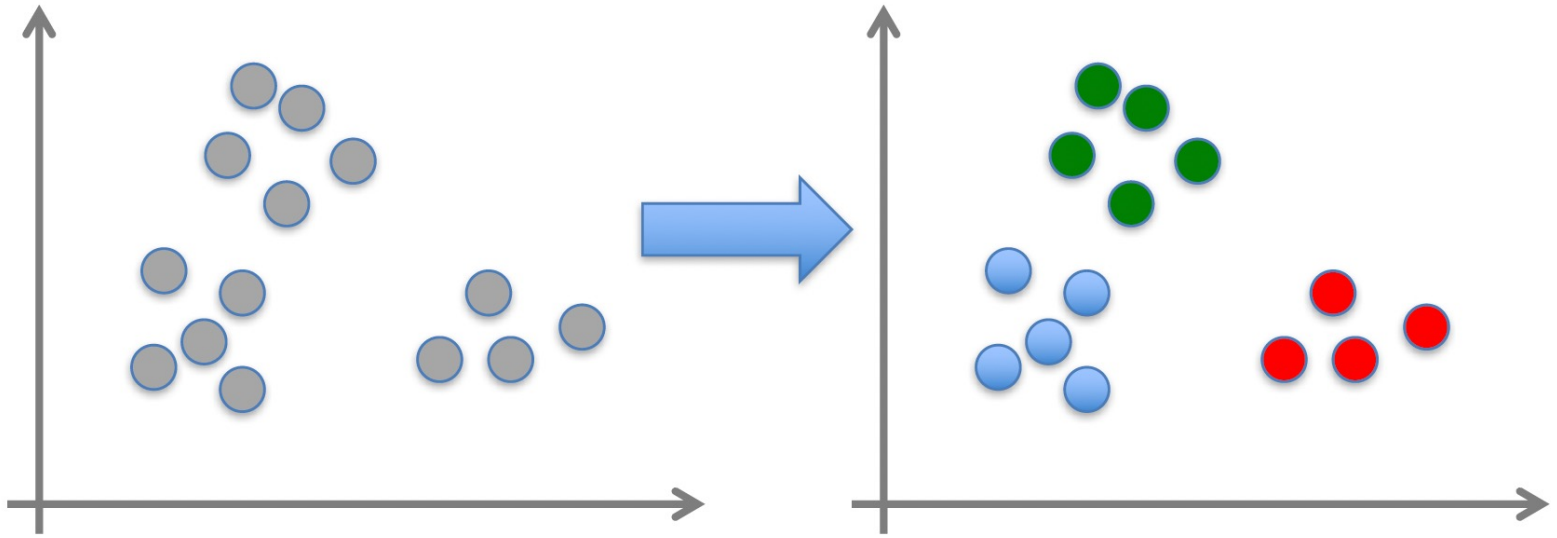


Model(Feature) = Prediction

$f(x_{i1}, x_{i2}, \ldots, x_{iM}) = P_i$

Unsupervised Learning

- Given $x_1, x_2, \ldots, x_n$ (without labels)
- Output hidden structure behind the x's
  - E.g., clustering



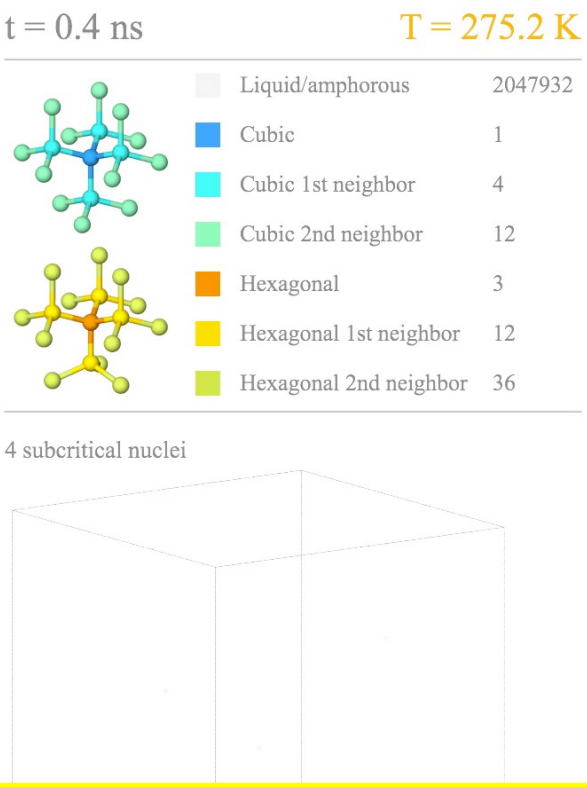
Data Mining Large Scale Simulations using Unsupervised ML

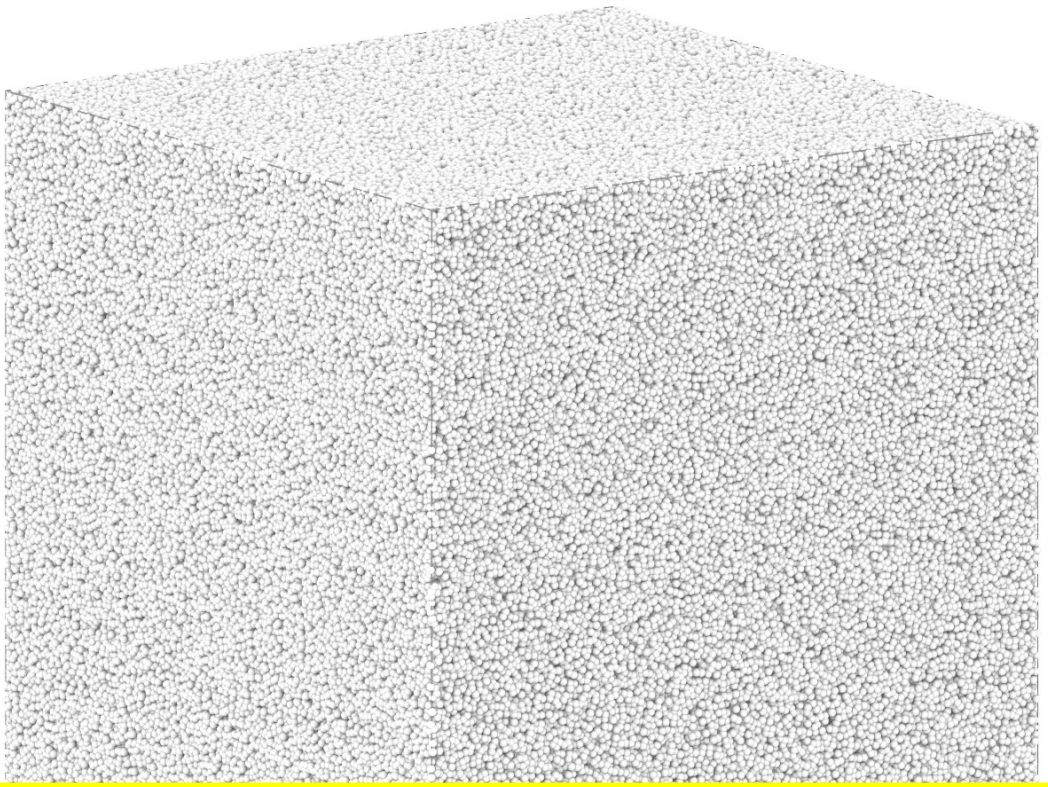
H. Chan, M. Cherukara, B. Narayanan and S. K.R S. Sankaranarayanan

Method to Identify grains in 3-D polycrystalline materials

Patent (15/672,168)

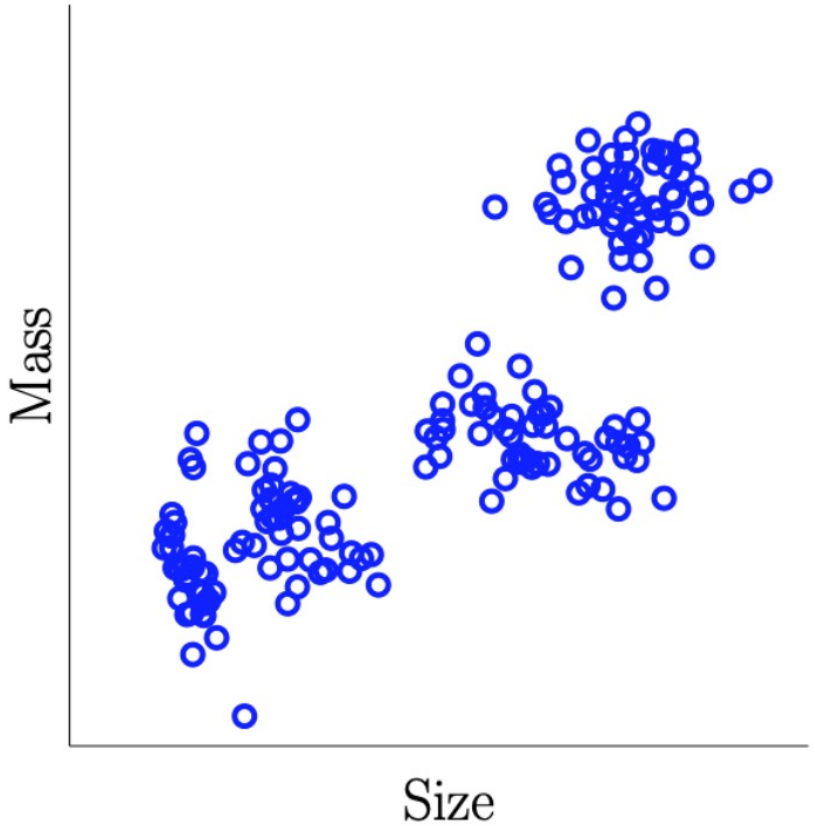
Invention number: IN-16-126

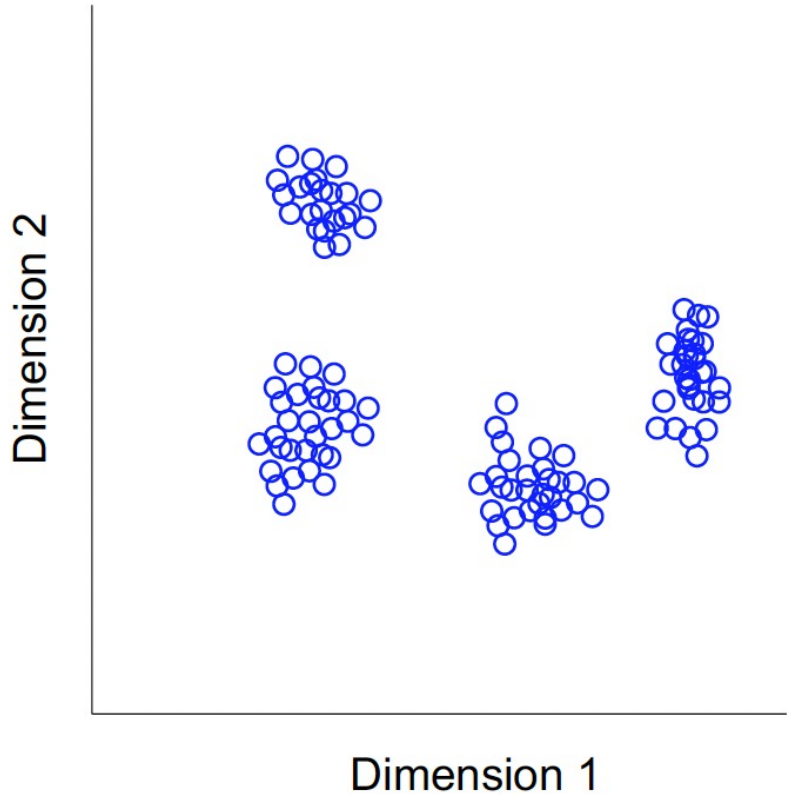




Rapid analysis allows In situ visualization

Unsupervised Learning





Instead of Model(Feature) = Prediction; we have Model(Feature) = NewFeature

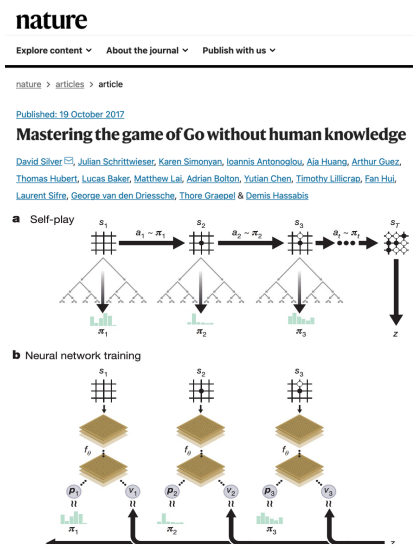
$$f(x_{i1}, x_{i2}, \ldots, x_{iM}) = x'_i$$

Reinforcement Learning

- Given a sequence of states and actions with (delayed) rewards, output a policy
  - Policy is a mapping from states → actions that tells you what to do in a given state
- Examples:
  - Credit assignment problem
  - Game playing
  - Robot in a maze
  - Balance a pole on your hand

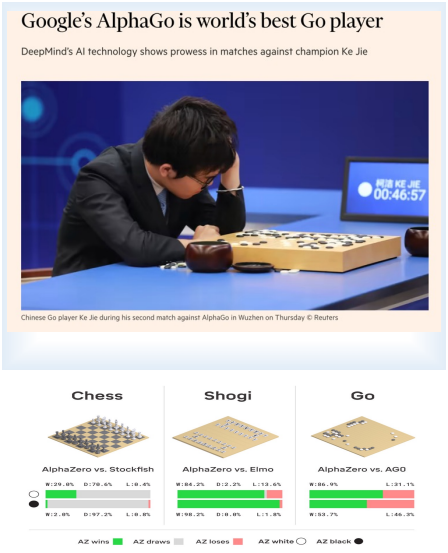
Reinforcement Learning

AlphaGo

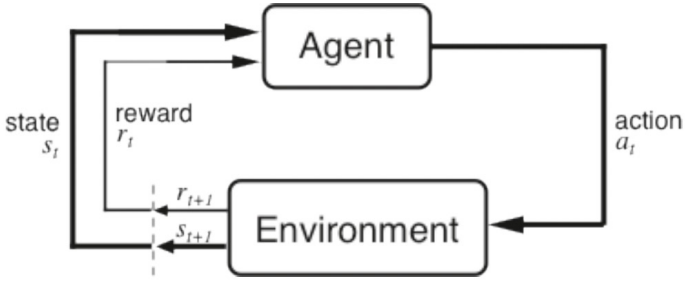


Board game Go

Discrete space
Policy driven



The Agent-Environment Interface



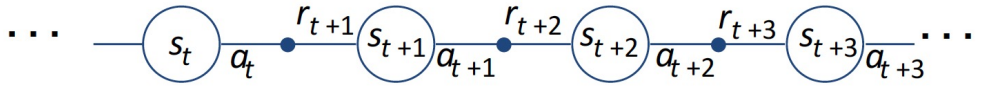
Agent and environment interact at discrete time steps : $t = 0, 1, 2, \mathrm{K}$

Agent observes state at step t : $s_t \in S$

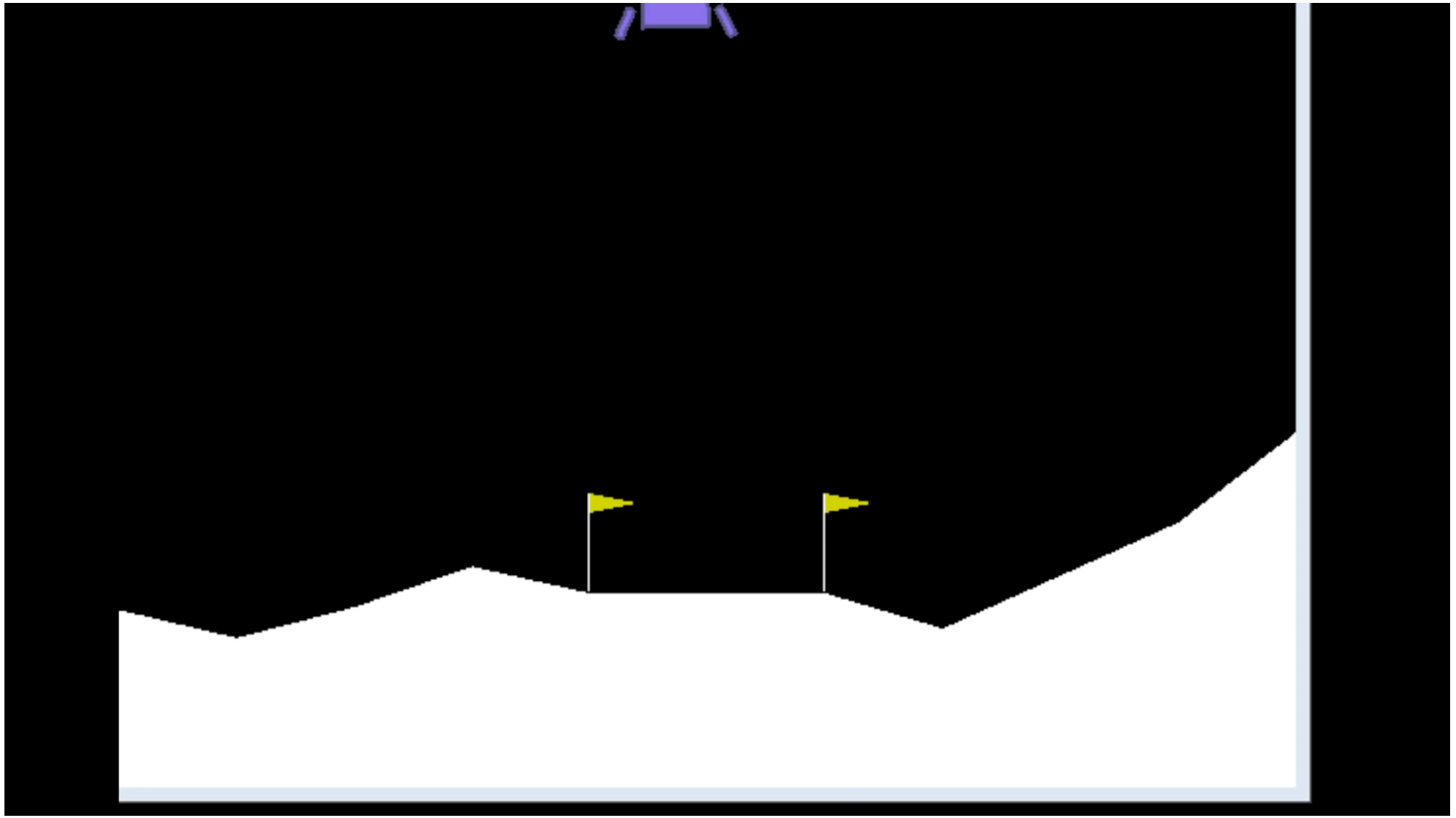
produces action at step t : $a_t \in A(s_t)$

gets resulting reward : $r_{t+1} \in \Re$

and resulting next state : s_{t+1}



RL Training of Games

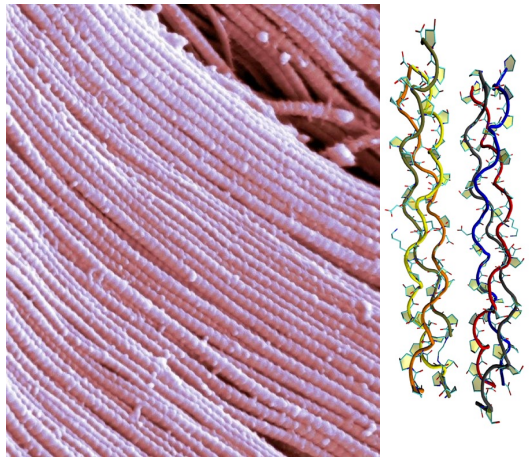


Nature Communications **volume 13**, Article number: 368 (2022)

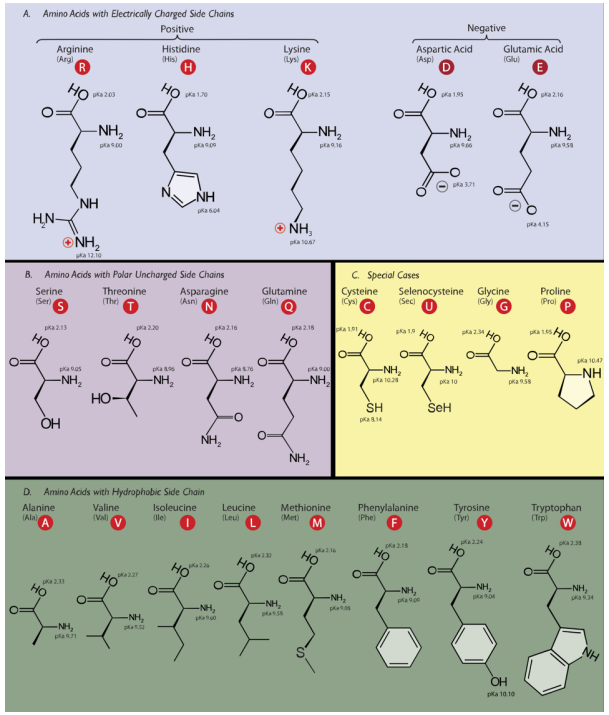
RL for peptide discovery

Sequence problem in peptides

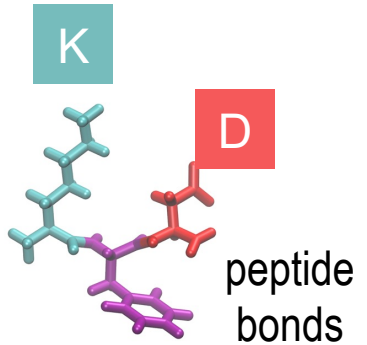
Examples

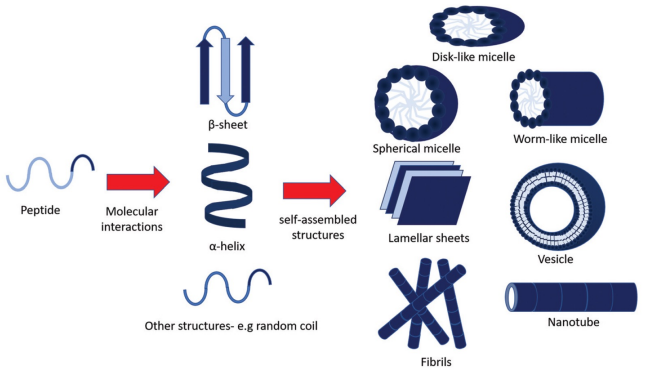


20 Amino Acids



E.g. Tripeptide KFD





- Useful for catalysis, sensing, tissue engineering
- Easy tunable properties
- Huge diversity

Org. Biomol. Chem., 15, 5867 (2017)

Traditional Approaches to Peptide Design

Sequence Challenge in Peptide Discovery

# of possibilities: 20^n

Sequence length (n)	# of candidates	Simulation time
3	8000	3 weeks
5	3.2 M	20 years
8	25.6 B	Many years

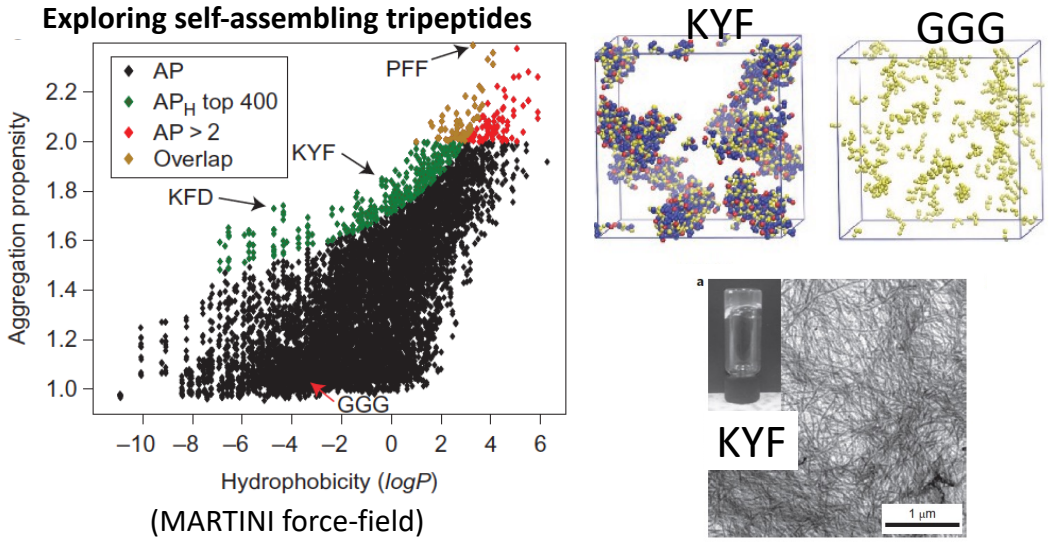
Empirical Rational Design

- Patterning (npnpn)
- Hydrophobicity scales
- Human bias; based on limited data

Computational Design

- MD simulation to estimate peptide properties
- Hydrophobicity scales
- Brute-force search; non-scalable

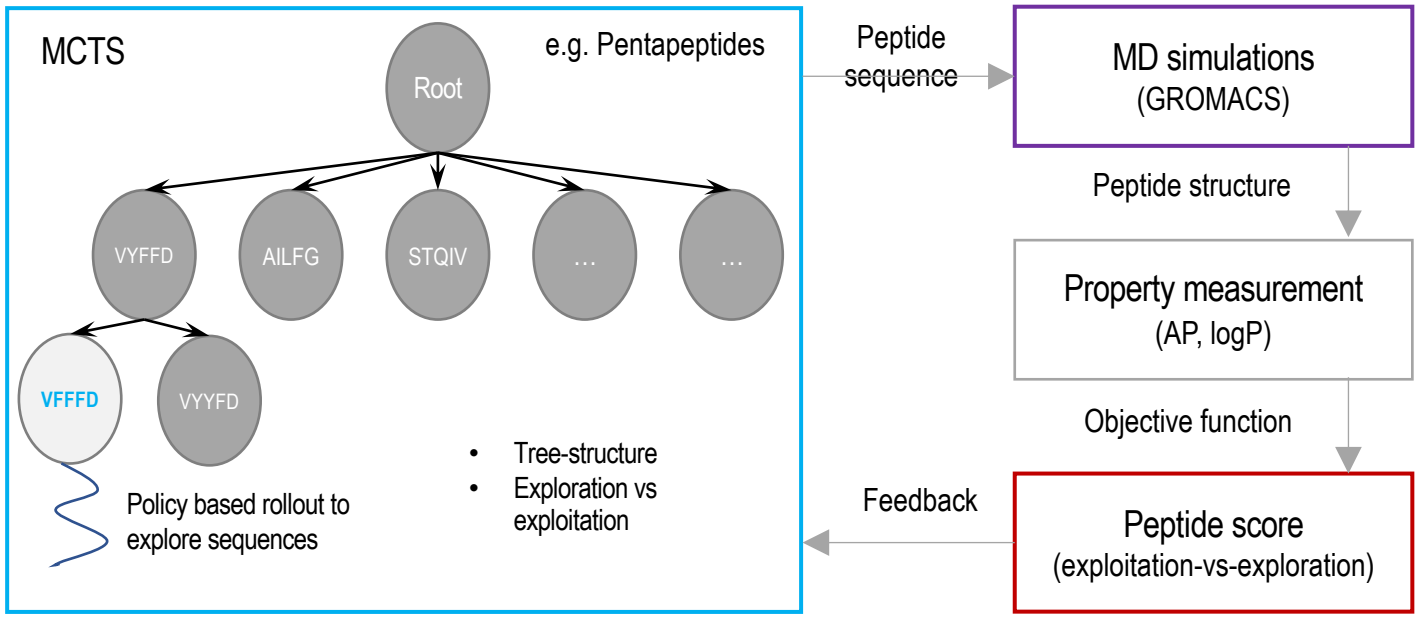
Can we do better?



(MARTINI force-field)

P. W. J. M. Frederix et al., Nature Chemistry, 7, 30 (2015)

AI-expert for Peptide Discovery



Key components

A. Monte Carlo tree search (MCTS) → Promising sequence generation
B. MD simulations → Model structure
C. Scoring function → Sequence evaluation

Search acceleration: Random forest ML model for efficient rollouts **(MCTS+RF scheme)**

Scoring function
(AP)*(logP)

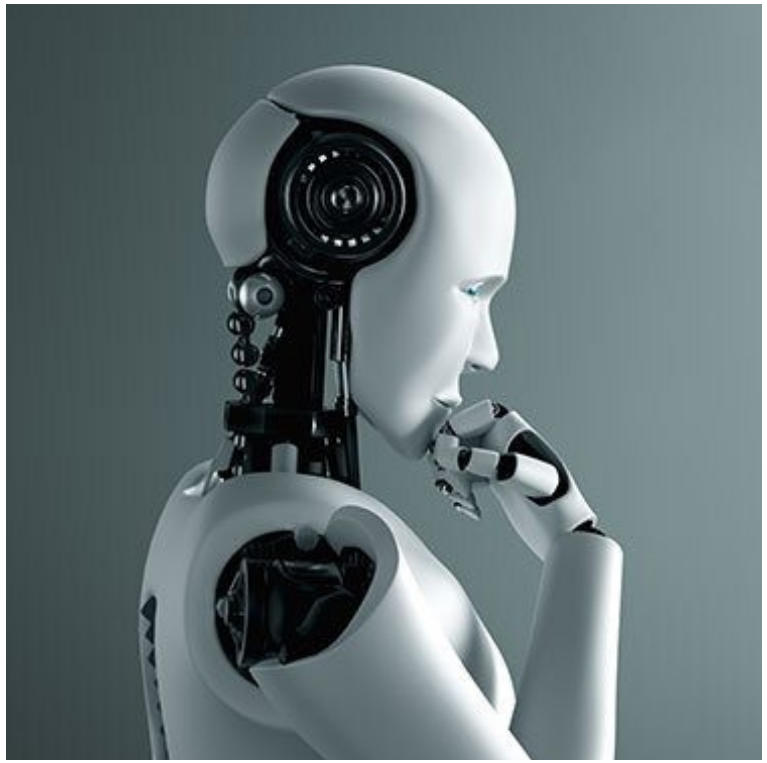
Aggregation propensity

Hydrophilicity

Batra, R, et al. "Self-assembling Peptide Discovery: Overcoming Human Bias With Machine Learning." Nature Chem (2022) **https://doi.org/10.1038/s41557-022-01055-3**

What about mechanical engineers?

Robotics + AI/ML/Data Science + Domain Expertise



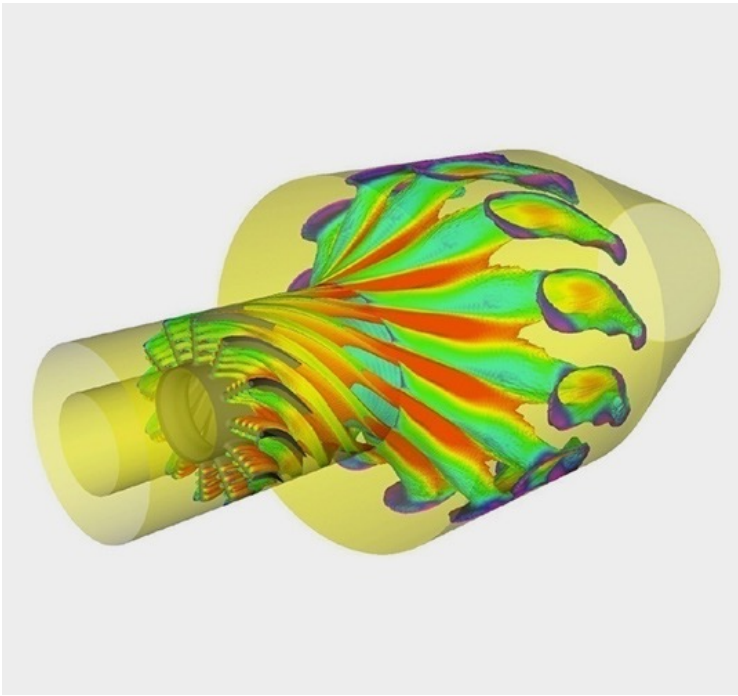
Source image - https://www.thebigredgroup.com/

+

jupyter NumPy SciPy pandas

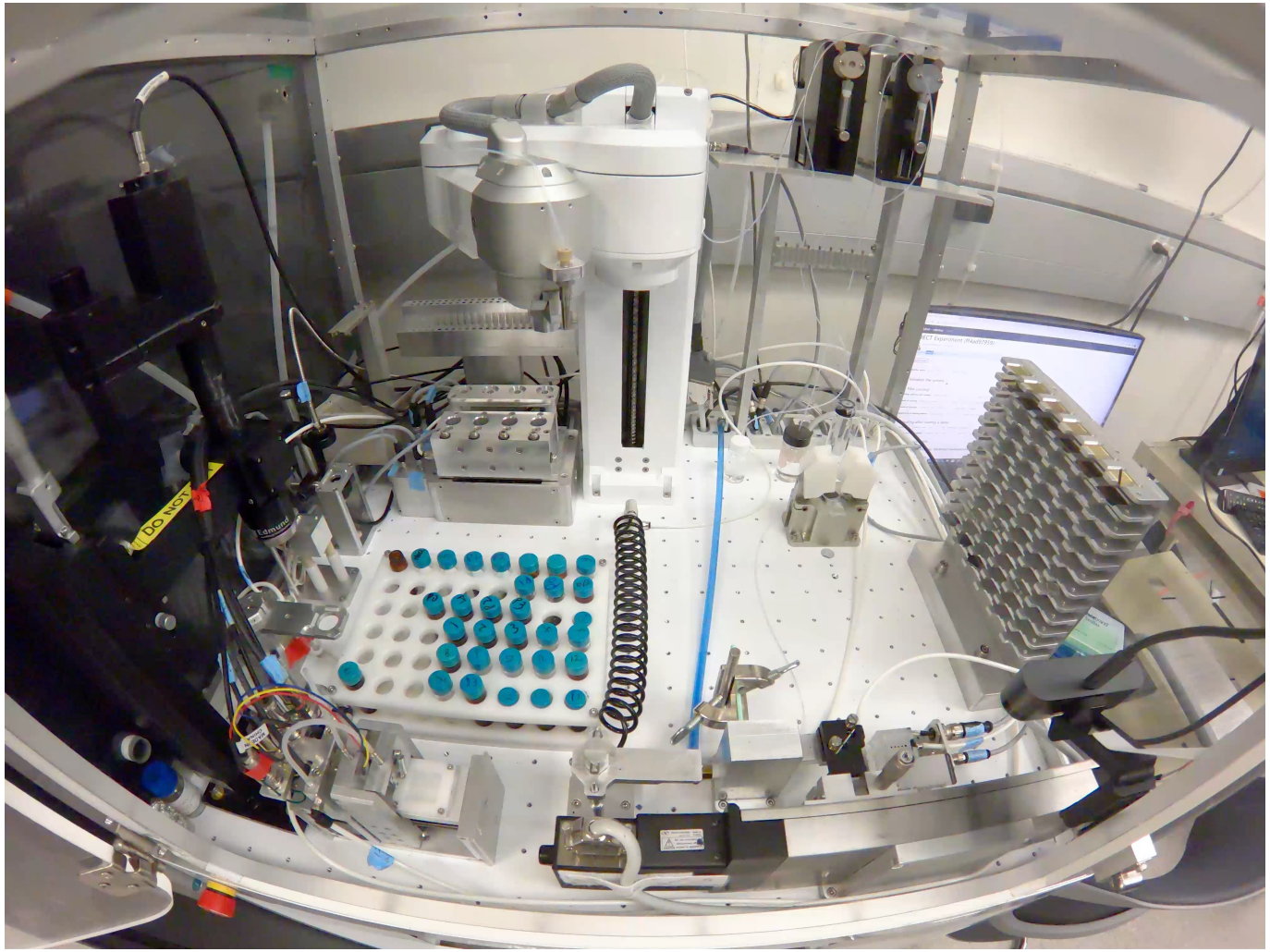
scikit learn TensorFlow

+



Source image - https://altair.com/fluids-thermal-applications

Robotics and Data Science ➔ Autonomous Discovery



(https://www.anl.gov/cnm/polybot)

What We'll Cover in this Course

Week (Tentative)	Topic
1	Introduction to Data Science
1	Basics of python programming (Hands on)
2	Basics of python programming (Hands on)
2	Exploratory data analysis and visualization (Hands on)
3	Linear regression
4	Gradient descent
5	Datasets for ML (Training, Testing, k-fold cross-validation)
6	Databases (Materials Project, AFLOW, Jarvis, polymer genome etc)
7	Fingerprinting
8	Unsupervised learning – Clustering
9	Principal Components Analysis

Week (Tentative)	Topic
10	Genetic Algorithms (single vs multi-objective)
11	Decision Trees (Random Forest)
12	Reinforcement Learning using Monte Carlo Tree Search
13	Neural networks in materials science and mech engg
14	Neural networks in materials science and mech engg
15	Intro to computer vision and image classification
16	Project presentation

References and Reading Material

- **Sobester, András, Alexander Forrester, and Andy Keane. Engineering design *via* surrogate modelling: a practical guide. John Wiley & Sons, 2008, ISBN: 9780470060681.**

- **Krishna Rajan; Informatics for Materials Science and Engineering. Butterworth-Heinemann, 2013, ISBN 978-0-12-394399-6.**

- **Rasmussen, C. E., & Williams, C. K. I. Gaussian processes for machine learning. MIT Press, 2005, ISBN: 9780262182539.**

- **Python Data Science Handbook: Essential Tools for Working with Data** by Jake VanderPlas

References and Reading Material

- **Mueller, Tim, Aaron Gilad Kusne, and Rampi Ramprasad. Machine learning in materials science: Recent progress and emerging applications. Reviews in computational chemistry 29 (2016): 186-273.**

- **Andrew White, Deep Learning for Molecules and Materials, https://whitead.github.io/dmol-book/intro.html**

- **Ian J. Goodfellow, Yoshua Bengio and Aaron Courville. Deep Learning, MIT Press, 2016, ISBN-13: 978-0262035613**