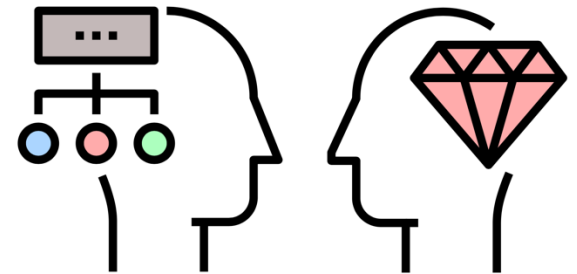


Machine Learning for Materials

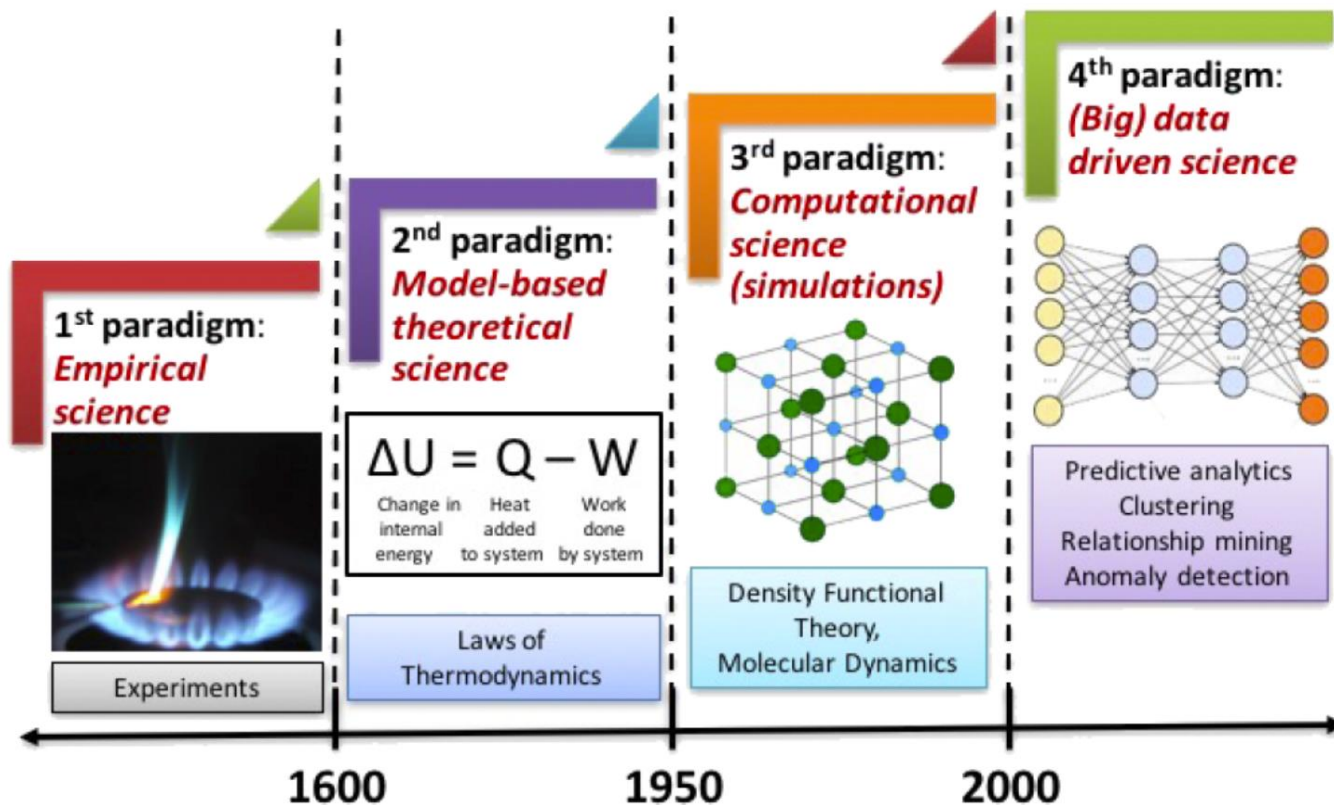
Aron Walsh

Department of Materials
Centre for Processable Electronics



New Era of Materials Research

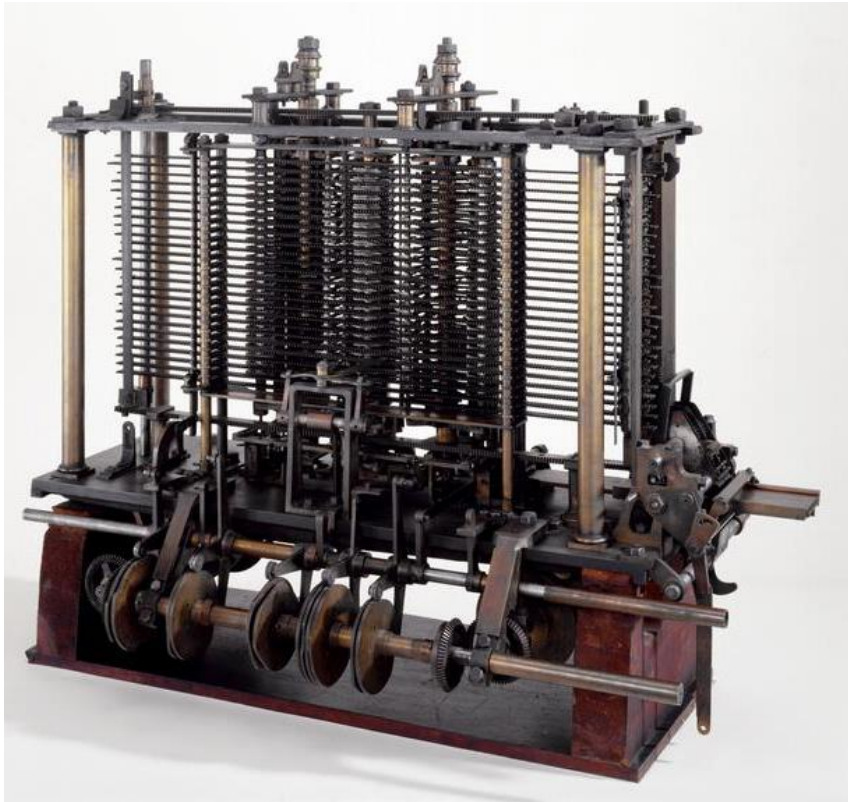
The research toolkit for materials science now includes powerful data-driven statistical models



Computer Revolution

Analytical Engine

Automated calculations



Charles Babbage (1837)

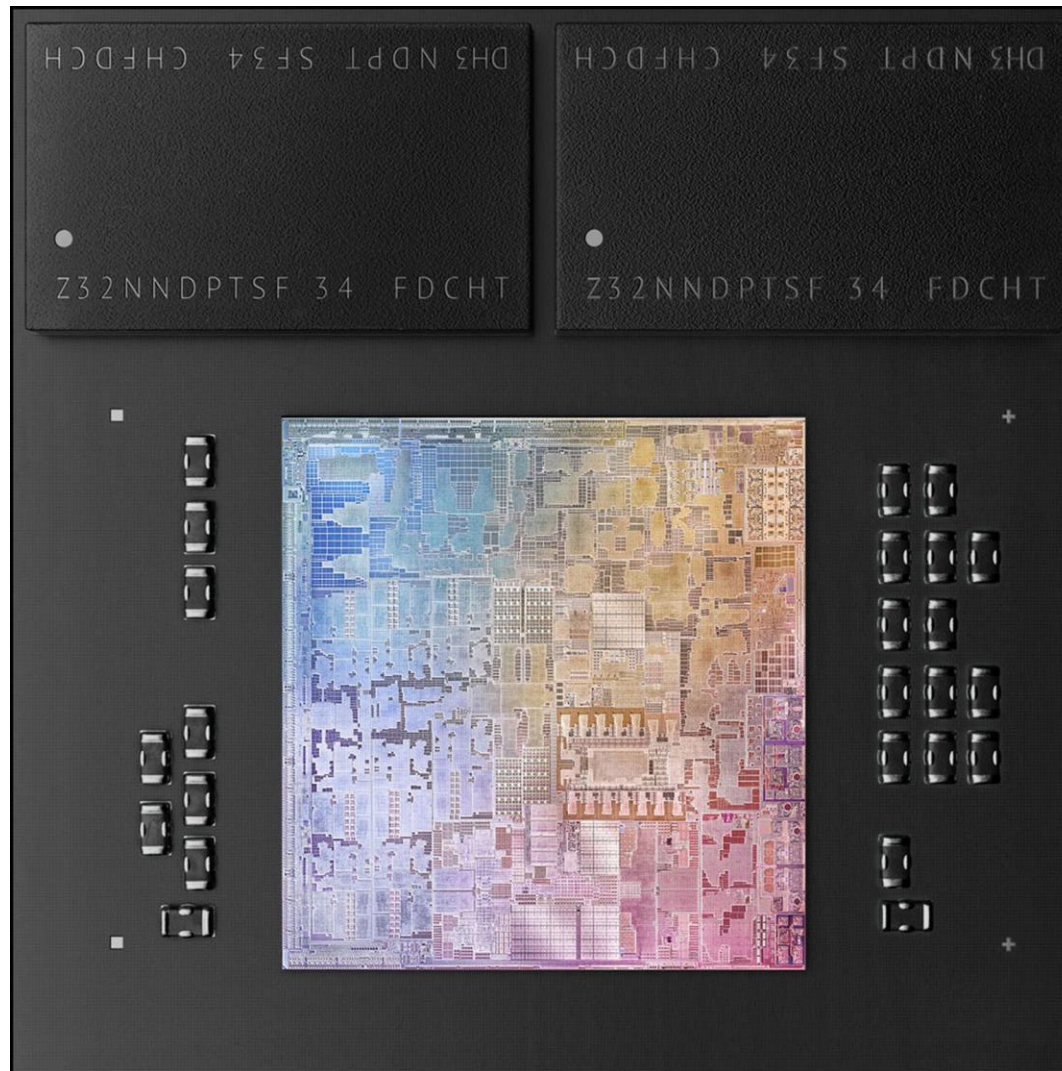
“The science of operations has its own truth and value”



Ada Lovelace (1840)

Multiple two 20 digit numbers in ~3 minutes

Computer Revolution

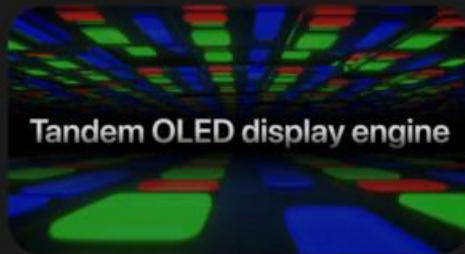


“System on a chip” microprocessor from <https://www.apple.com>

Computer Revolution

120GB/s

Unified memory bandwidth



Tandem OLED display engine

**Dynamic
Caching**

**Hardware-accelerated
mesh shading**

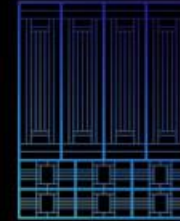


**Hardware-accelerated
ray tracing**

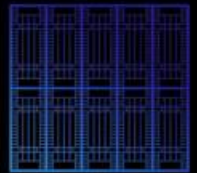


Apple M4

Up to



**10-core
CPU**



**10-core
GPU**

Up to

50%

faster CPU than M2

Up to

4x

faster GPU than M2

ProRes

AV1

Over

28 billion transistors

Second-generation

3 nm technology

Neural Engine with

38 trillion ops/sec


“System on a chip” microprocessor from <https://www.apple.com>

Exascale Supercomputing

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	El Capitan - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	Aurora - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	Eagle - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
5	HPC6 - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461

Exascale computing refers to 10^{18} floating point operations per second; <https://top500.org>

Powerful Statistical Techniques

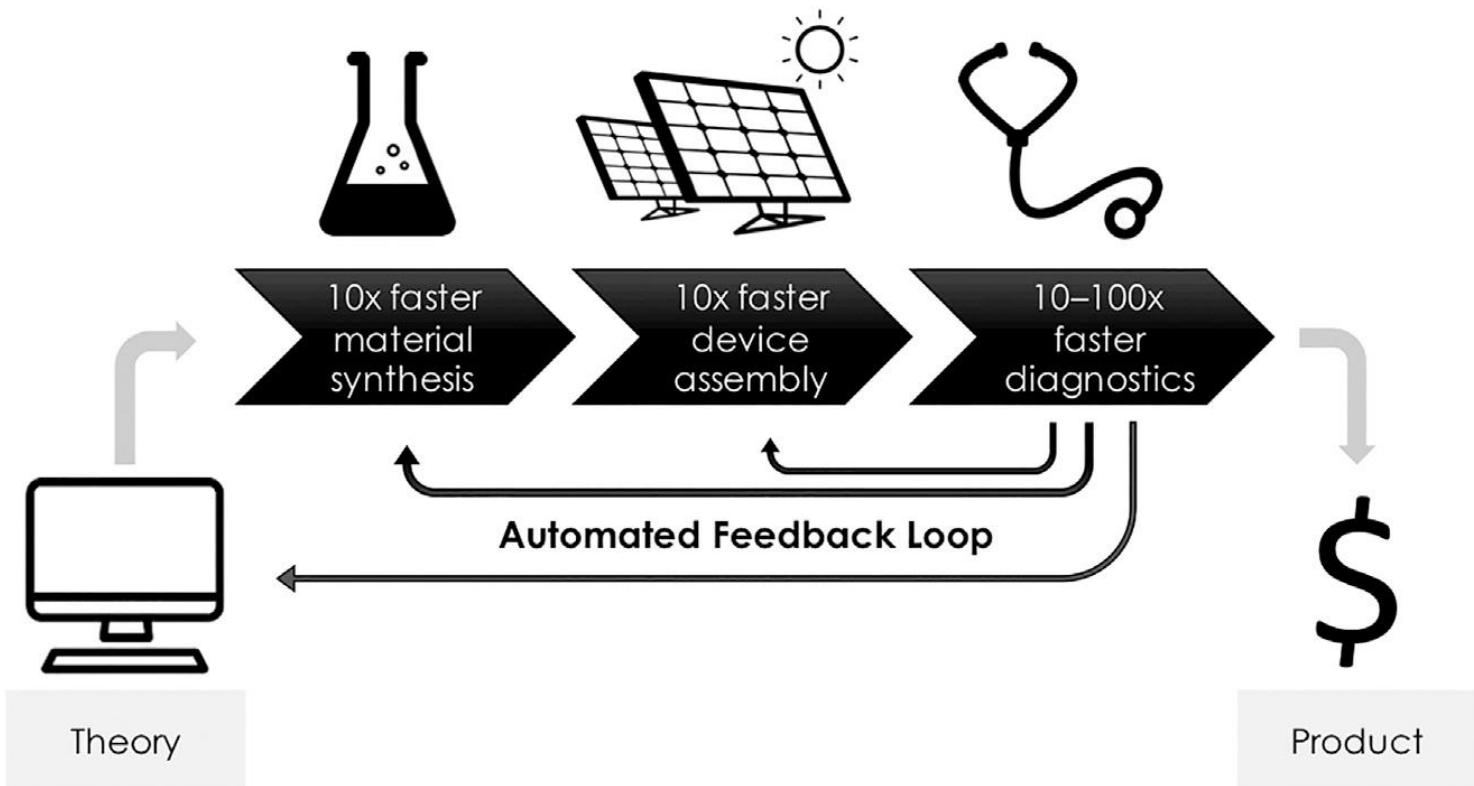


```
from langchain import OpenAI
model = OpenAI( temperature=0.9)
text = "Suggest three novel cathodes for Al ion batteries"
print(model(text))
```

Answers provided included transition metal oxides (V_2O_5), Chevrel phases (Mo_6S_8), Prussian blues ($Fe_4[Fe(CN)_6]_3$)

Efficient Research Workflows

Integration of computational techniques to accelerate discovery & development cycles



Module Contents

1. Introduction
2. Machine Learning Basics
3. Materials Data
4. Crystal Representations
5. Classical Learning
6. Artificial Neural Networks
7. Building a Model from Scratch
8. Accelerated Discovery
9. Generative Artificial Intelligence
10. Recent Advances

Class Outline

Course Introduction

A. Overview

B. Expectations

C. Assessments

What is Machine Learning (ML)?

Statistical algorithms that learn from training data and build a model to make predictions

Learning types

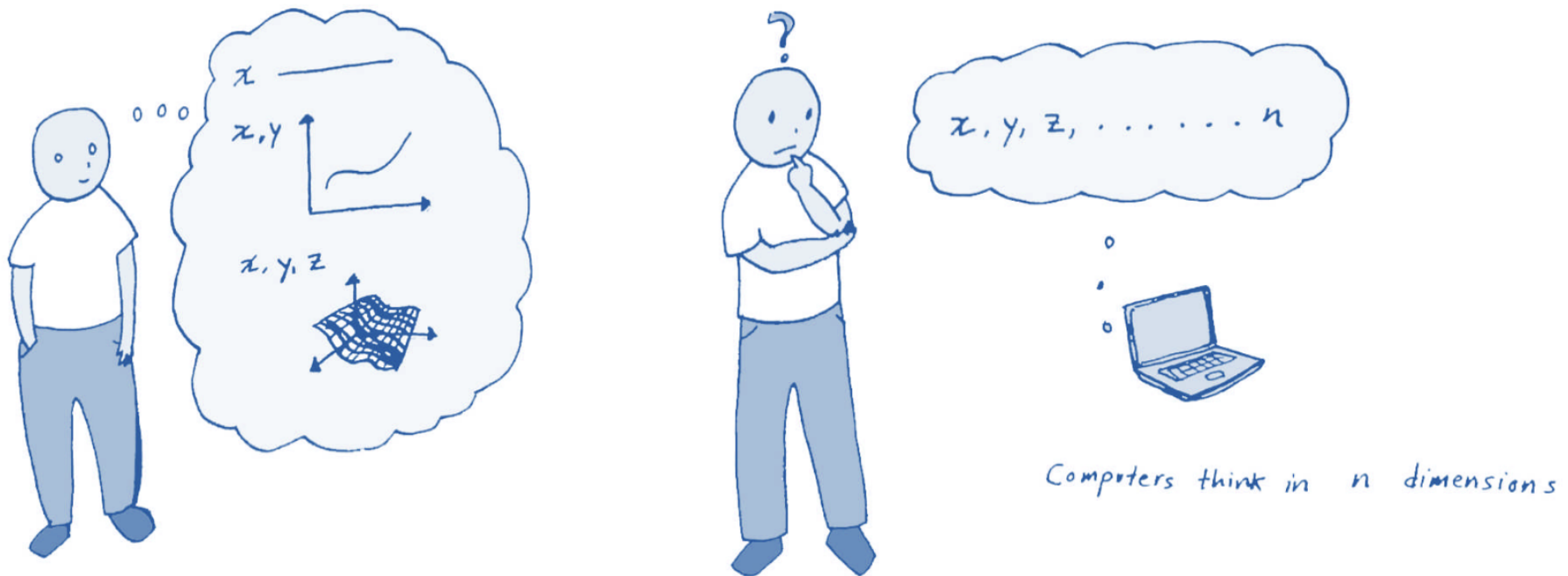
Unsupervised (**identify patterns**), supervised (**use patterns**), reinforcement (**maximise reward**)

Data types

Materials features can be binary (e.g. **stability**), categorical (e.g. **symmetry**), integer (e.g. **stoichiometry**), continuous (e.g. **rate**)

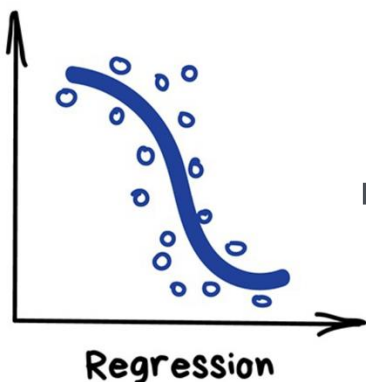
What is Machine Learning (ML)?

Statistical algorithms that identify and use patterns in multi-dimensional datasets

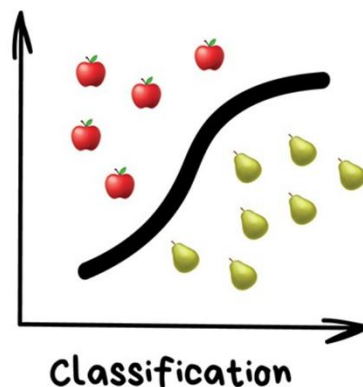


What is Machine Learning (ML)?

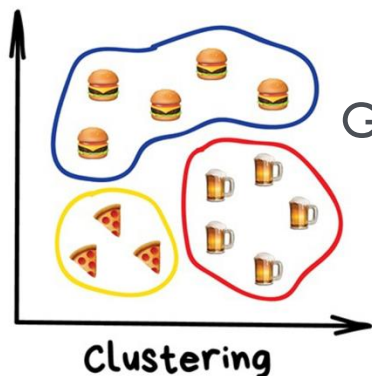
Statistical algorithms that identify and use patterns in multi-dimensional datasets



Predict a value, e.g.
regression to extract a
reaction rate



Predict a category, e.g.
decision trees to predict
reaction outcome



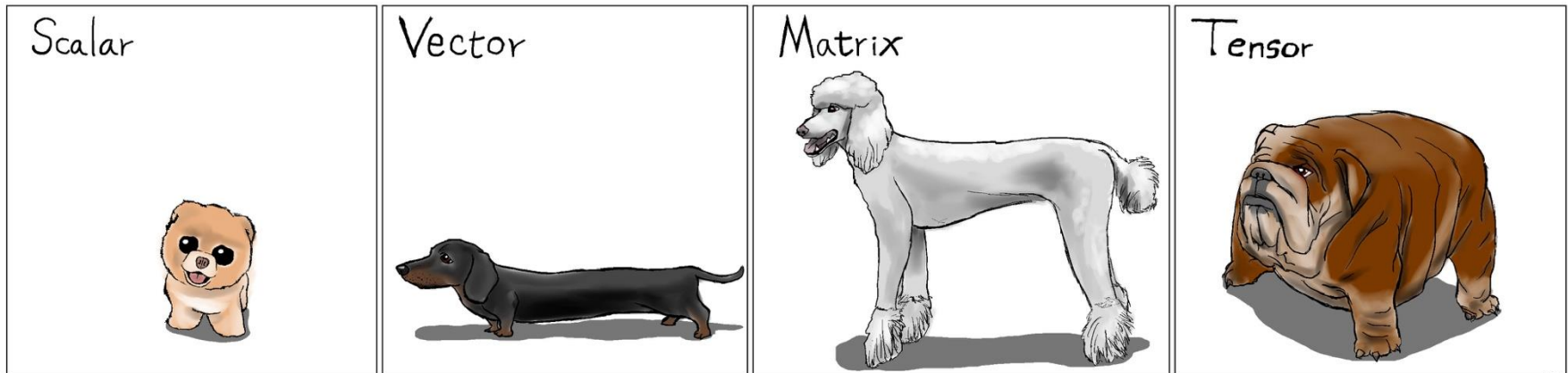
Group by similarity, e.g.
high-throughput
crystallography



Maximise reward, e.g.
reaction conditions to
optimise yield

What is Machine Learning (ML)?

Statistical algorithms that operate on
multi-dimensional arrays of numerical data



x

1

x_i

[7 8 3]

x_{ij}

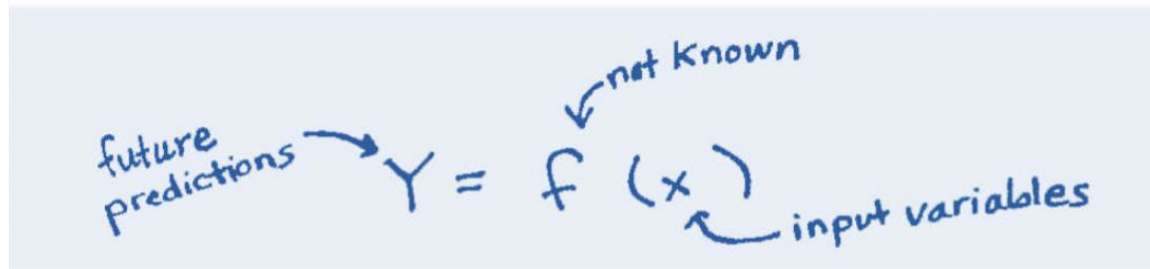
$$\begin{bmatrix} 7 & 2 & 3 \\ 4 & 8 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

x_{ijk}

$$\begin{bmatrix} [1 \ 7] & \dots & [6 \ 4] \\ \vdots & \ddots & \vdots \\ [5 \ 6] & \dots & [2 \ 8] \end{bmatrix}$$

What is Machine Learning (ML)?

Statistical algorithms that operate on
multi-dimensional arrays of numerical data



A handwritten diagram on a light blue background showing the equation $Y = f(x)$. An arrow points from the text "future predictions" to the variable Y . Another arrow points from the text "input variables" to the variable x . A third arrow points from the text "not known" to the function f .

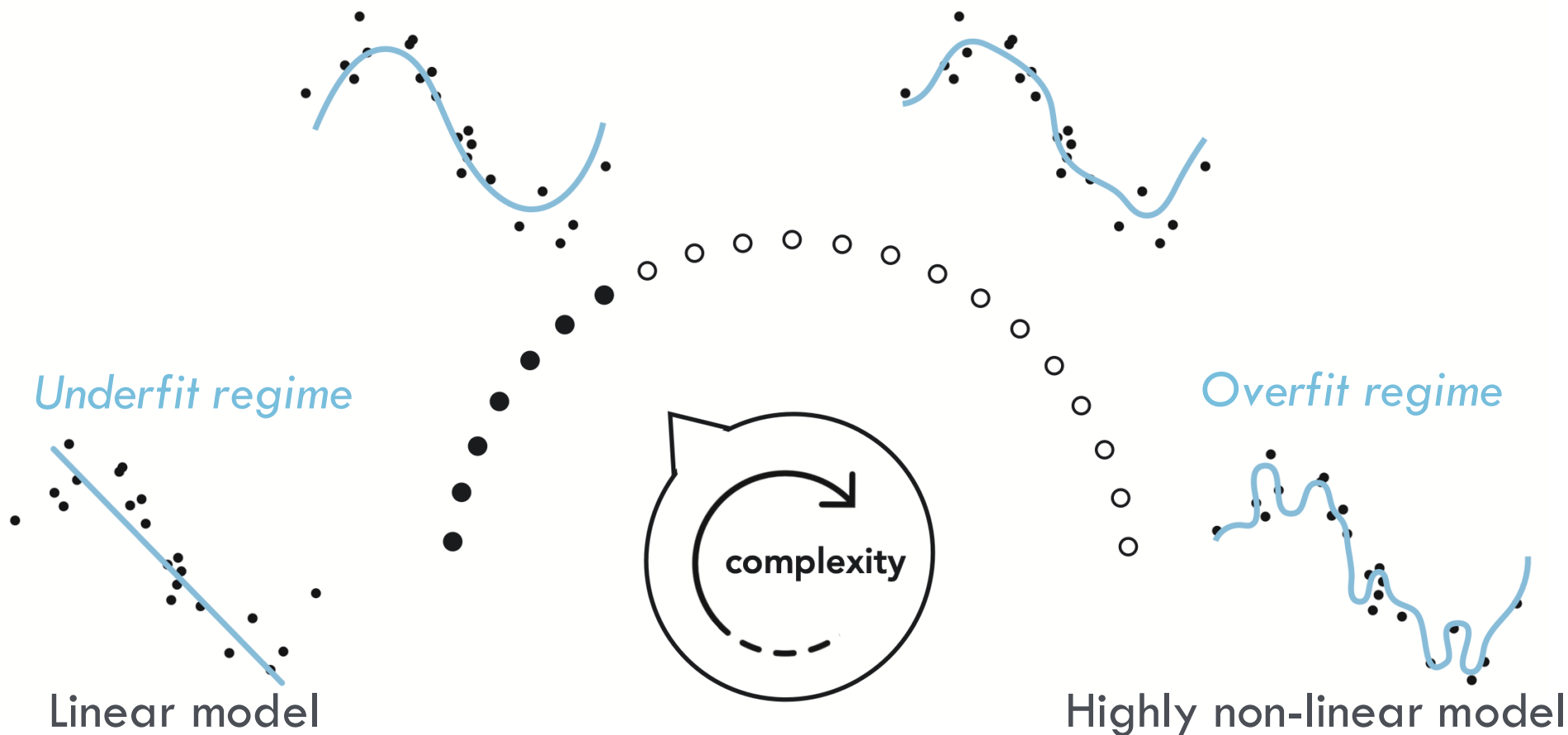
LEARN FROM X (RELEVANT DATA)
TO MAKE Y (ACCURATE PREDICTIONS)

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \end{bmatrix}$$

3×1 matrix 3×5 matrix 5×1 matrix

ML ~ Function Approximation

Model selection, training, and testing tunes a “complexity dial” for your problem of interest

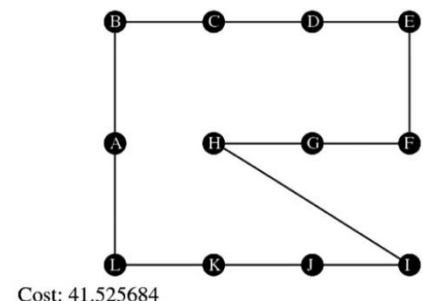
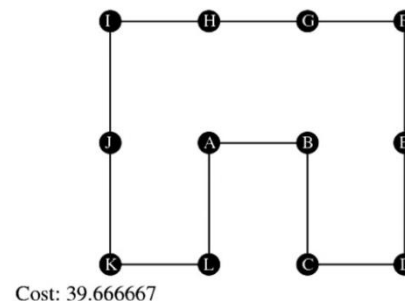
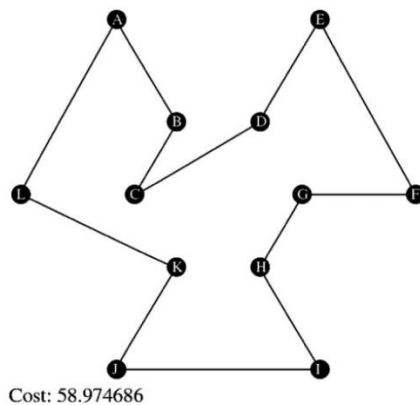
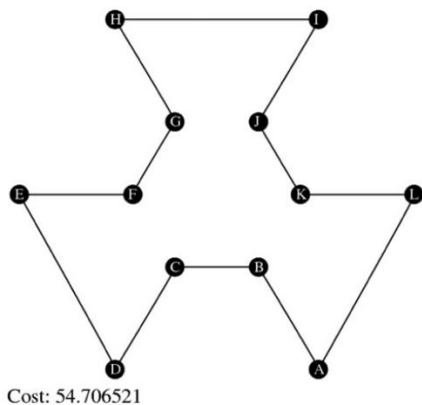


Why Machine Learning (ML)?

Many problems are difficult to solve using standard techniques, e.g. combinational expansions

Non-deterministic polynomial hard (NP-hard)

Challenging class of computational problems, where finding an efficient solution remains an open and difficult task



Travelling salesman: find the shortest route that visits each city once and returns home

Why Machine Learning (ML)?

Many problems are difficult to solve using standard techniques, e.g. combinational expansions

Relevant challenges in materials science

Reaction engineering

Navigate configurational space of reactants & products

Crystal structure prediction

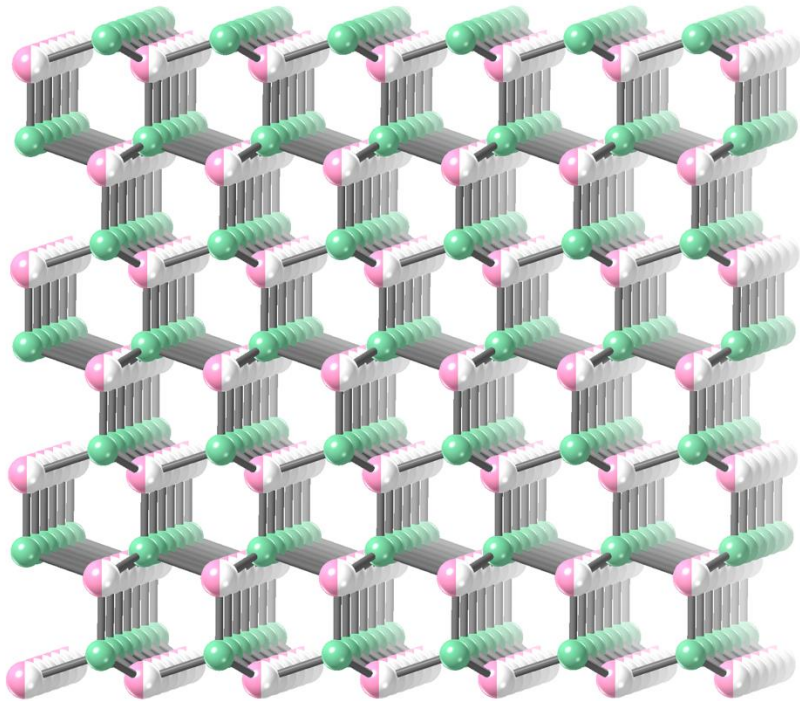
Find the optimal 3D structure(s) for a given composition

Materials design

Achieve target functionality within chemical constraints

Why Machine Learning (ML)?

Solid-solutions are used to control structure and properties, e.g. $(1-x)\text{ZnO} + (x)\text{ZnS} \rightarrow \text{ZnO}_{1-x}\text{S}_x$



A wurtzite crystal with a partially occupied anion site

Number of configurations for $\text{ZnO}_{0.5}\text{S}_{0.5}$

$$\frac{N!}{\left(\frac{N!}{2}\right)\left(\frac{N!}{2}\right)}$$

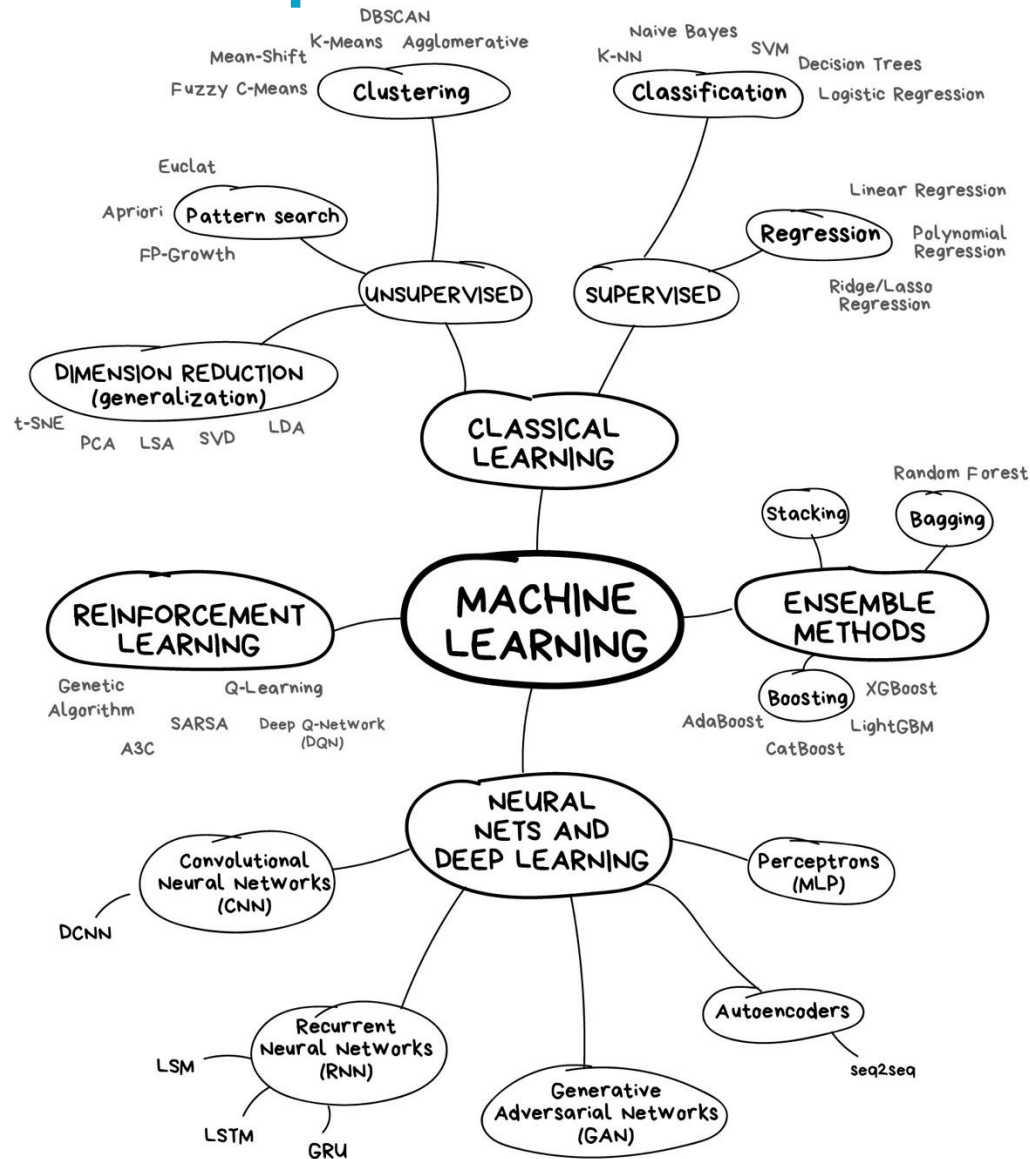
Mixed sites in a supercell model

$$N = 16: 12,870$$

$$N = 32: 6 \times 10^8$$

$$N = 64: 1.8 \times 10^{18}$$

ML Model Map

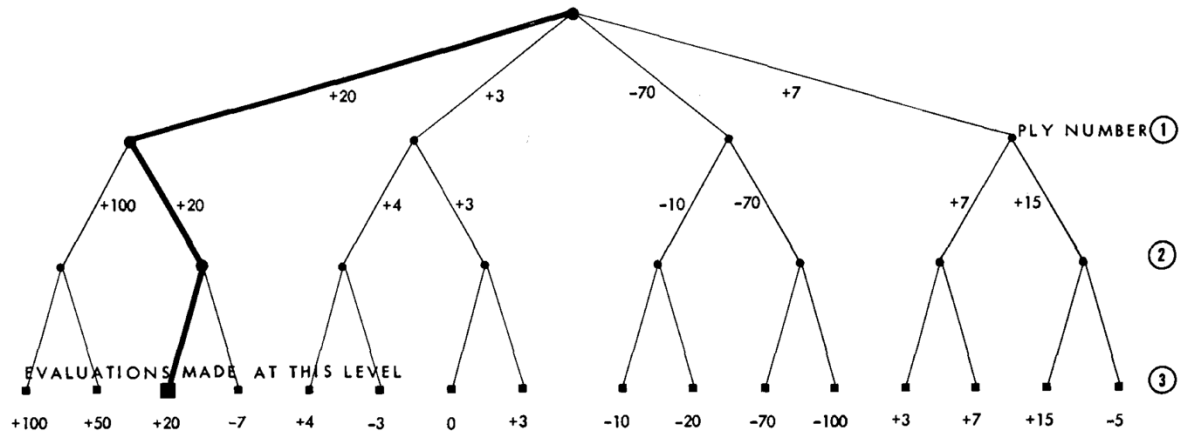


Brief History of ML

Term coined by Arthur Samuel in 1959

Some Studies in Machine Learning Using the Game of Checkers

- ① MACHINE CHOOSES BRANCH WITH LARGEST SCORE
- ② OPPONENT EXPECTED TO CHOOSE BRANCH WITH SMALLEST SCORE
- ③ MACHINE CHOOSES BRANCH WITH MOST POSITIVE SCORE



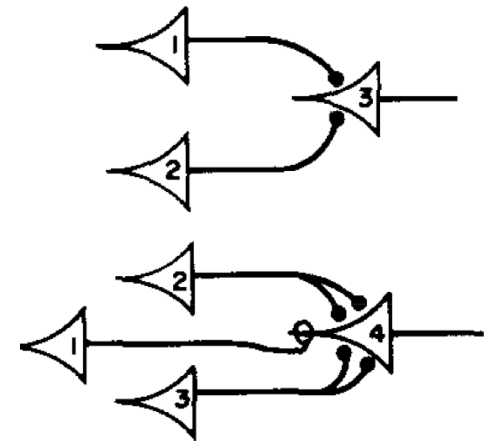
“It is now possible to devise learning schemes which will greatly outperform an average person and that such learning schemes may eventually be economically feasible”

Brief History of ML

An artificial neuron had been proposed in 1943

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*

- (1) The activity of the neuron is an “all-or-none” process.
- (2) A certain fixed number of synapses must be excited within the period of latent addition in order to excite a neuron at any time, and this number is independent of previous activity and position on the neuron.
- (3) The only significant delay within the nervous system is synaptic delay.
- (4) The activity of any inhibitory synapse absolutely prevents excitation of the neuron at that time.
- (5) The structure of the net does not change with time.



“Every net, if furnished with a tape, scanners connected to afferents to perform the necessary motor-operations, can compute only such numbers as can a Turing machine”

Brief History of ML

In 1950, Alan Turing proposed a “Learning Machine” that could become intelligent

I.—COMPUTING MACHINERY AND INTELLIGENCE

Q : Please write me a sonnet on the subject of the Forth Bridge.

A : Count me out on this one. I never could write poetry.

Q : Add 34957 to 70764

A : (Pause about 30 seconds and then give as answer) 105621.

Q : Do you play chess?

A : Yes.

“I PROPOSE to consider the question, Can machines think?”

ML in Materials R&D

Growing field combining traditional industry,
large technology companies, and start-ups

Microsoft Research
AI4Science



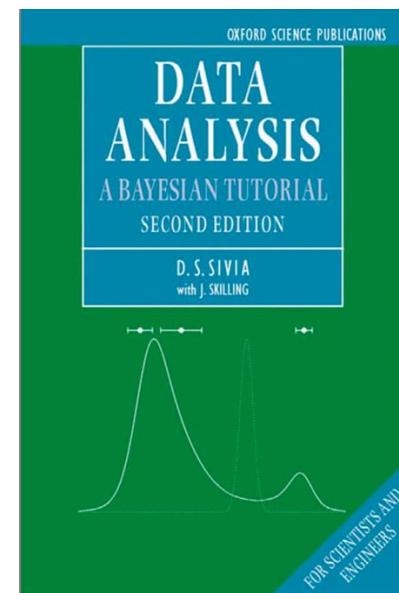
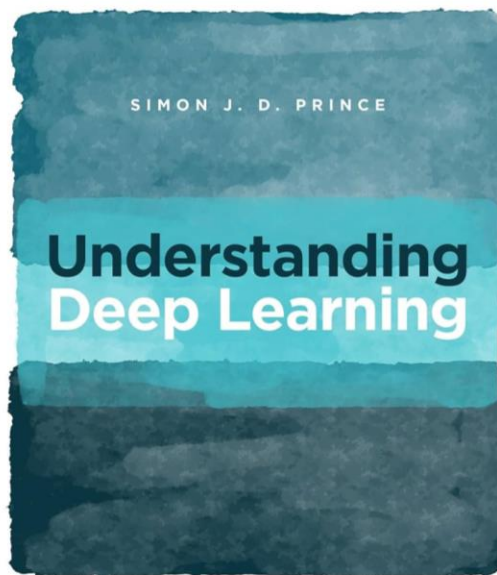
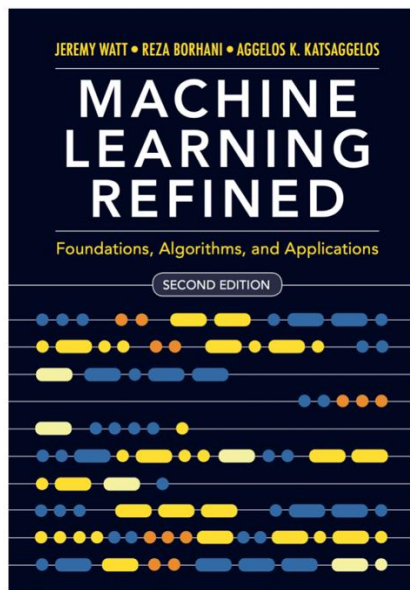
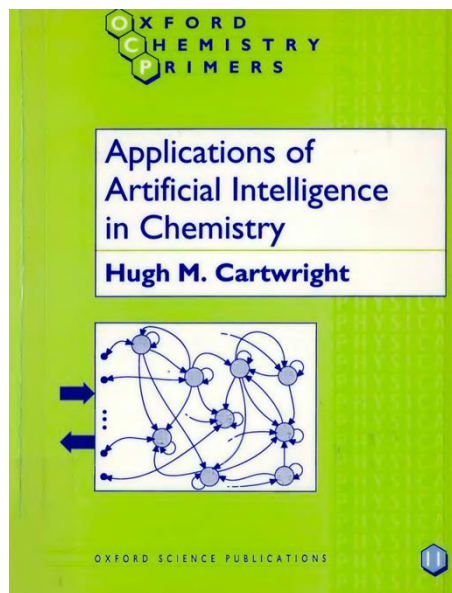
BenevolentAI



Source Material for Course

ML content available from many sources, including blogs, research papers, repositories, and textbooks

These slides are a skeleton, fleshed out with lectures, activities, and reading



General

Specialist

Special thanks to Anthony Onwuli and Zhenzhu Li for assistance

Class Outline

Course Introduction

A. Overview

B. Expectations

C. Assessments

Active Participation

Your engagement is essential. This is a dense course with new concepts, Python coding, and self-study

- Attend all lectures to hear the core content
- Attend all practical sessions for hands-on coding
 - Attempt to solve problems yourself and ask course assistants if you need help

Creative Solutions

There is great flexibility in programming with no unique solution for a given problem

You may be interested in speed or clarity, but ultimately want a robust code

- Check package manuals, e.g. <https://matplotlib.org> & <https://scikit-learn.org>
- Search <https://stackoverflow.com> & <https://github.com> for ideas

Creative Solutions

Many AI assistants for coding exist such as Github Copilot, CodeWhisperer, Codeium, GPT4

- Most helpful when you know the basics first
- Assistants often lack domain expertise and may give poor suggestions with buggy code based on out-of-date libraries
 - Not a substitute for hands-on coding experience and knowledge of materials

2025 Module Assistants

Research Fellow



Dr Zhenzhu Li



Kinga Mastej



Pan D.



Irea
Mosquera-Lois



Xia
Liang



Fintan
Hardy



Yifan
Wu

Class Outline

Course Introduction

A. Overview

B. Expectations

C. Assessments

Module Assessment

Aim for working knowledge of ML with practical sessions and coursework

Computer labs (8×3%)

Notebook submitted on Blackboard
(Due by the end of each session – 16:00)

Research challenge (76%)

Assignment to complete
(details after Lecture 9)

Introductory Quiz

<http://menti.com>

Open on your phone, tablet or laptop
