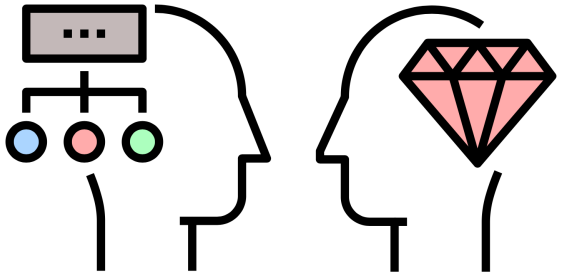
Imperial College London

Machine Learning for Materials

9. Research Challenge

Aron Walsh

Department of Materials
Centre for Processable Electronics



Course Contents

1. Course Introduction
2. Materials Modelling
3. Machine Learning Basics
4. Materials Data and Representations
5. Classical Learning
6. Artificial Neural Networks
7. Building a Model from Scratch
8. Recent Advances in AI
9. and 10. Research Challenge

Course Assessment

Aim for working knowledge of ML with practical sessions and coursework

Computational exercises (40%)

Completed - well done!

Research challenge (60%)

Individual assignment

(details today)

Class Outline

Research Challenge

A. Ethics

B. Assignment

C. Approach

Ethics of Machine Learning

How do these translate to the materials context?

Bias and Fairness

Influence on decision making processes

Transparency and Explainability

Interpretation of model predictions

Privacy and Data Protection

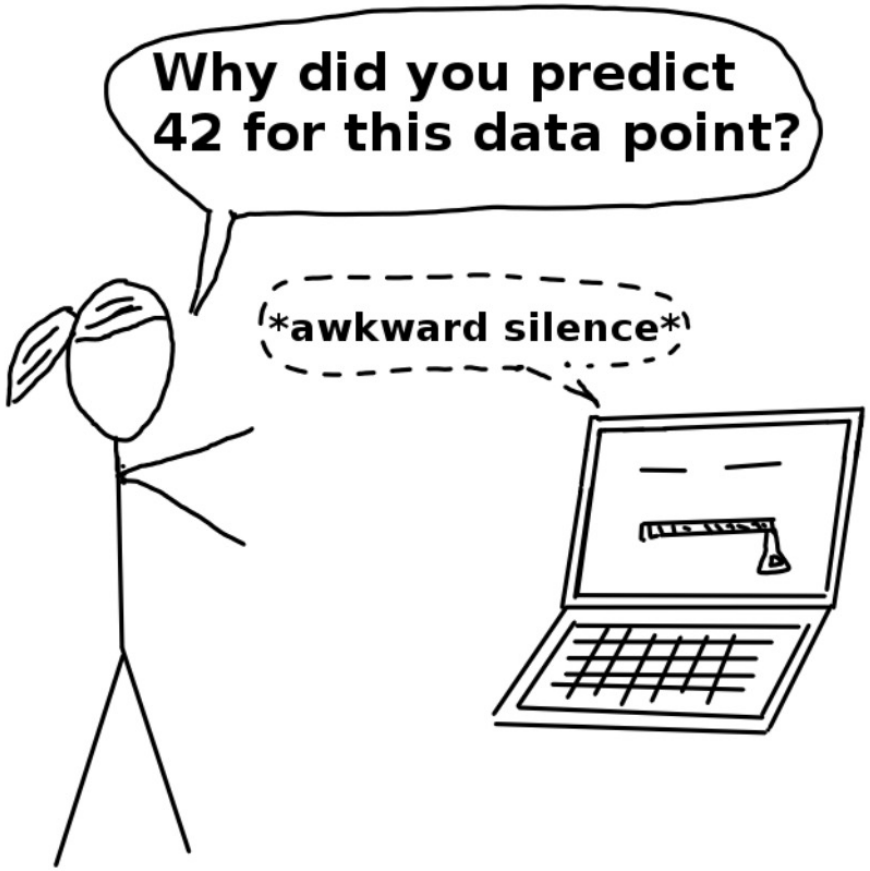
Collection, storage and using sensitive data

Social Impacts

From productivity increases to job displacements

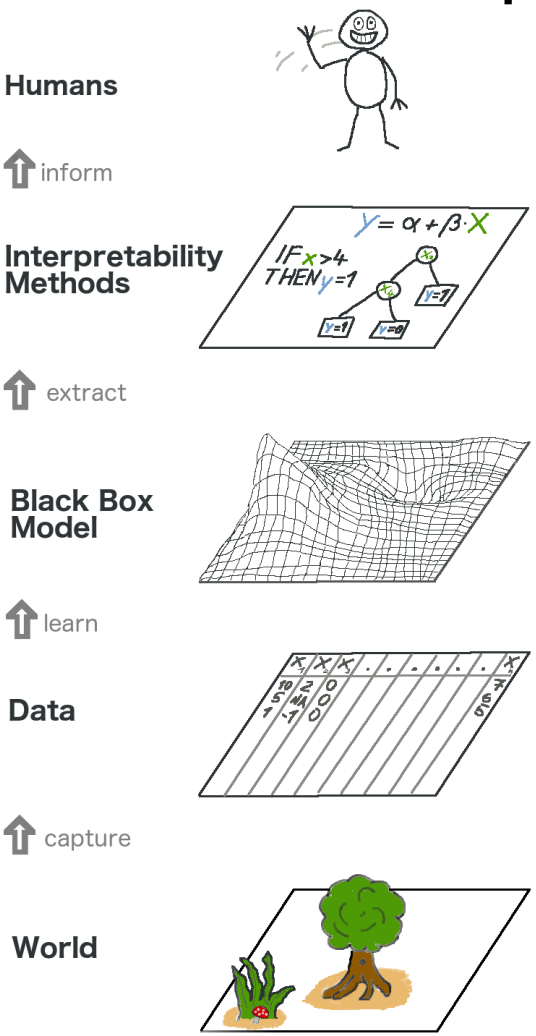
Ethics of Machine Learning

Importance of interpretable and explainable models



Ethics of Machine Learning

Importance of interpretable and explainable models



Some interpretability methods

- Feature Importance
- SHAP (SHapley Additive exPlanations)
- Interpretable Surrogate Models
- Counterfactual Explanations

Ethics of Machine Learning

"Developers must show that their models are safe, transparent and explainable to users"

NEWS EXPLAINER | 16 February 2024

What the EU's tough AI law means for research and ChatGPT

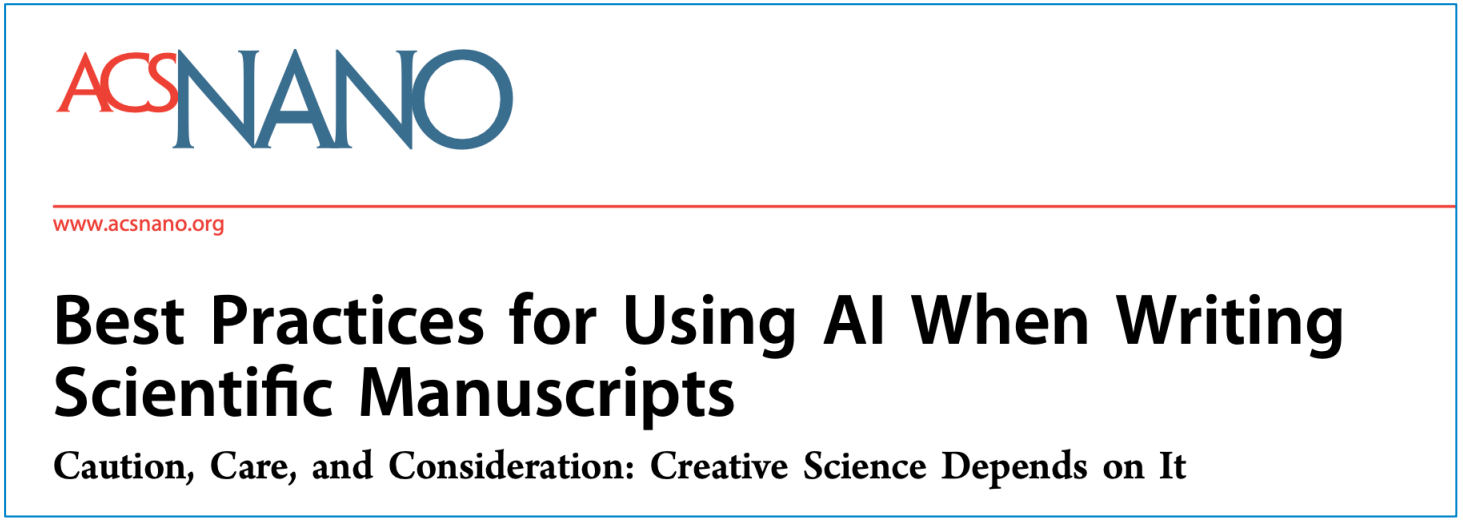
The EU AI Act is the world's first major legislation on artificial intelligence and strictly regulates general-purpose models.

By Elizabeth Gibney

https://www.nature.com/articles/d41586-024-00497-8

Ethics of Large Language Models

How best to use these models in our research?



"AI language bots are incapable of understanding new information, generating insights, or deep analysis, which would limit the discussion within a scientific paper." …Outdated?

https://pubs.acs.org/doi/10.1021/acsnano.3c01544

Ethics of Large Language Models

How best to use these models in our research?

Theoretical and computational chemistry

Call for Papers: Harnessing the Power of Large Language Model-Based Chatbots for Scientific Discovery

Pavithra Naullage
Sep 21, 2023 • 2 min read

This Virtual Special Issue will foster further discussion on appropriate applications of Chatbots in the chemical, pharmaceutical, material, and biological sciences. Submit your manuscript by August 1, 2024.

https://axial.acs.org/theoretical-and-computational-chemistry

Challenging ML Questions

Models are not unique, different architectures often give similar performance

How to choose the best one?

There is uncertainty in the input data, trained model, and the predicted outputs

How to properly deal with error estimations?

A model may be trained for several systems or across a limited set of conditions

How can I tell if it will extrapolate?

Class Outline

Research Challenge

A. Ethics

B. Assignment

C. Approach

Research Challenge

An opportunity to develop your practical skills. Goals:

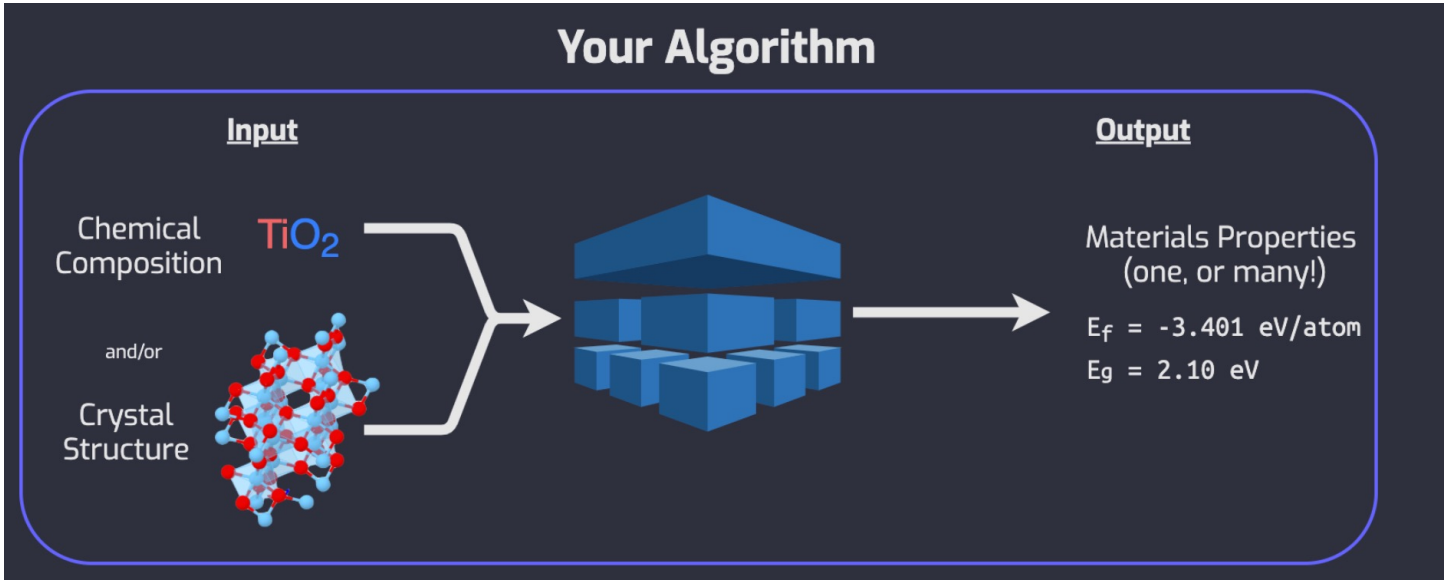
- To apply the ML tools and data skills you have picked up so far
- To extend your knowledge through self-study, exploration, and cohort interactions
- To produce a clearly annotated code with comparison to community benchmarks

Research Challenge

Each group is assigned a dataset on

https://matbench.materialsproject.org

Your job is to produce an original model for the given classification or regression task



Some tasks use chemical composition only, while others use composition and structure

Research Challenge

The starting point is to check the literature.

Read the matbench paper and the models that have been tested

I. Data Preparation

II. Model Choice

III. Training and Testing

https://doi.org/10.1038/s41524-020-00406-3

Class Outline

Research Challenge

A. Ethics

B. Assignment

C. Approach

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://stackexchange.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 give poor suggestions with buggy code based on old versions of Python libraries
- Not a substitute for hands-on coding experience and knowledge of materials

Creative Solutions

Statement to be included in the submitted notebook

Large Language Model (LLM) Usage Declaration

- Did you use an LLM (e.g. GPT-3, Gemini, Co-Pilot)?
- Specify tasks (e.g. summarising research)
- Were any limitations/biases noted?
- How did you ensure ethical use?

Challenge Topics

Challenge	Topic	Type	GTA
A	Phonons (1,265)	Regression	Anthony
B	Bandgap (4,604)	Regression	Irea
C	Perovskites (18,928)	Regression	Xia
D	Glasses (5,680)	Classification	Yifan

One challenge assigned per person. Dataset details in Notebook 9

GTA Assistance

Teaching assistants will be available
in the computer rooms:

(After Class 9) **Tue 20th:** 10-11am

(After Class 10) **Thur 22nd:** 1-2pm

The computer rooms is also booked on 27th and 29th at the same times to facilitate self-study

Submission deadline: 11th March 15:00

Challenge Submission

Two items submitted on Blackboard

1. Jupyter notebook (.ipynb)

and

2. Recorded presentation* (max 5 min) where you introduce your code and your choices on 1. *Data Preparation; 2. Model Choice; 3. Training and Testing*

*Format is flexible. Could be recorded in PowerPoint, screenshare on Zoom, or plain video

Challenge Assessment

	Weight	Guidelines
Data Preparation	20 %	Apply appropriate pre-processing steps
Model Choice	20 %	Justify based on the problem and available data
Training and Testing	20 %	Successfully train, validate and test model
Code	10 %	Clearly organised and annotated
Presentation	30 %	Clarity and conciseness

Lecture 10

Final Class on Thursday at 11am

*Guest lecture on **reinforcement learning***

Dr Zhenzhu Li

Schmidt AI in Science Fellow

Module Feedback

First run of this module, so your feedback is valued & will help to shape it for next year