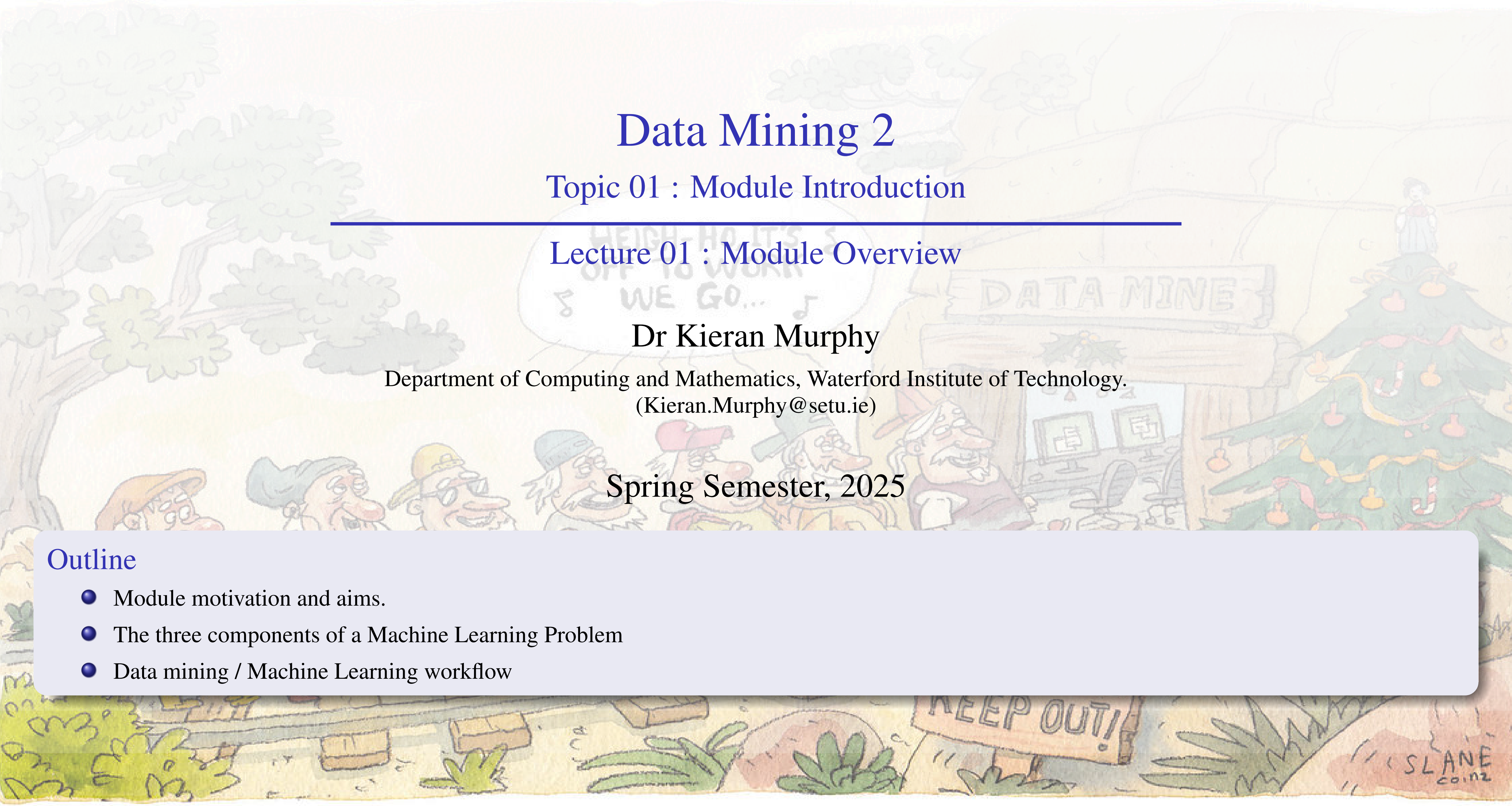



# Data Mining 2

## Topic 01 : Module Introduction

## Lecture 01 : Module Overview

Dr Kieran Murphy

Department of Computing and Mathematics, Waterford Institute of Technology.
(Kieran.Murphy@setu.ie)

Spring Semester, 2025

### Outline

- Module motivation and aims.
- The three components of a Machine Learning Problem
- Data mining / Machine Learning workflow

# What is Data Mining ?

We are drowning in data but starving for knowledge!

Necessity is the mother of invention $\Rightarrow$ Data Mining $\approx$ Automated analysis of massive data sets.

**Definition 1 (Data Mining)**

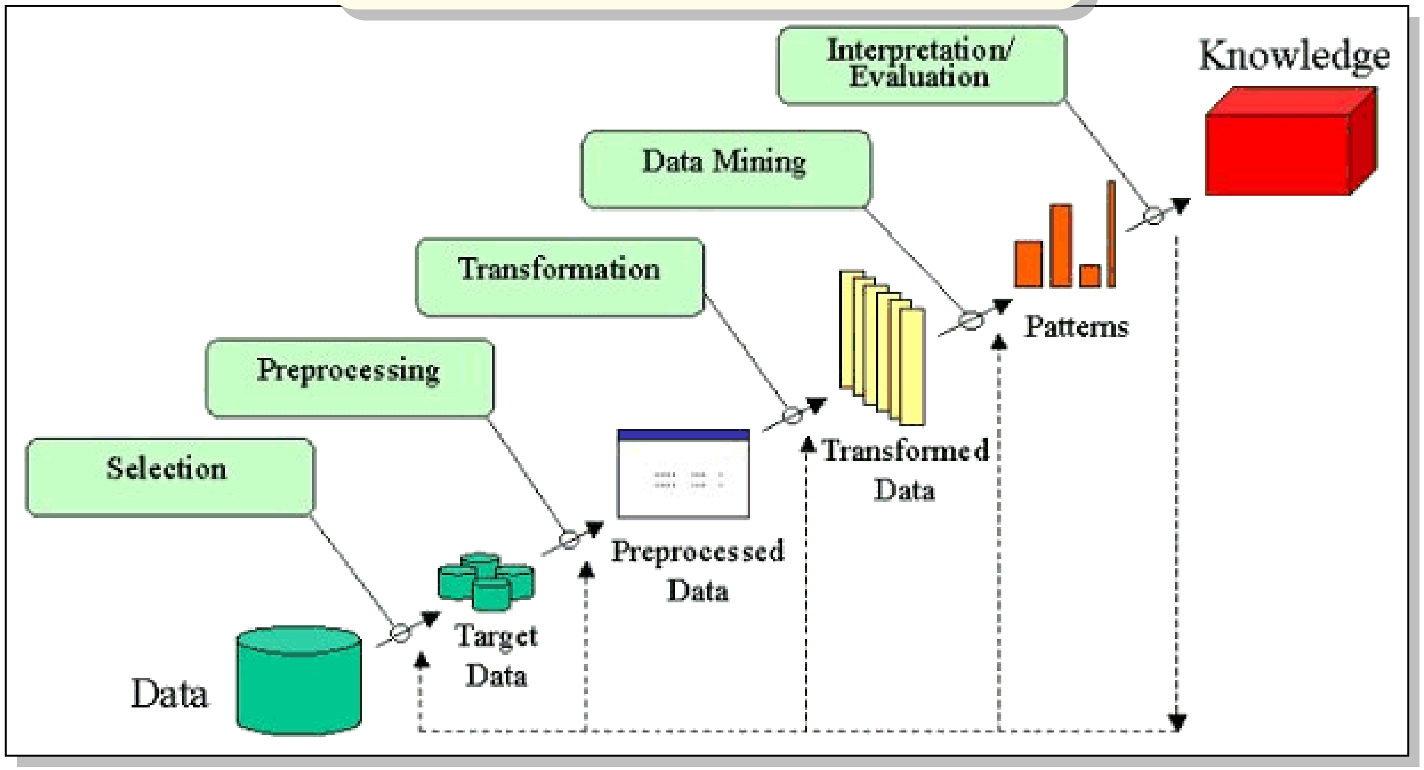
The non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data in large data repositories

| | |
|---|---|
| non trivial | — obvious knowledge is not useful (we already know it) |
| implicit | — hidden difficult to observe knowledge |
| previous unknown | — if known then, why go to this effort? |
| potentially useful | — actionable easy to understand |

# Data Mining vs Knowledge Discovery in Data (KDD)

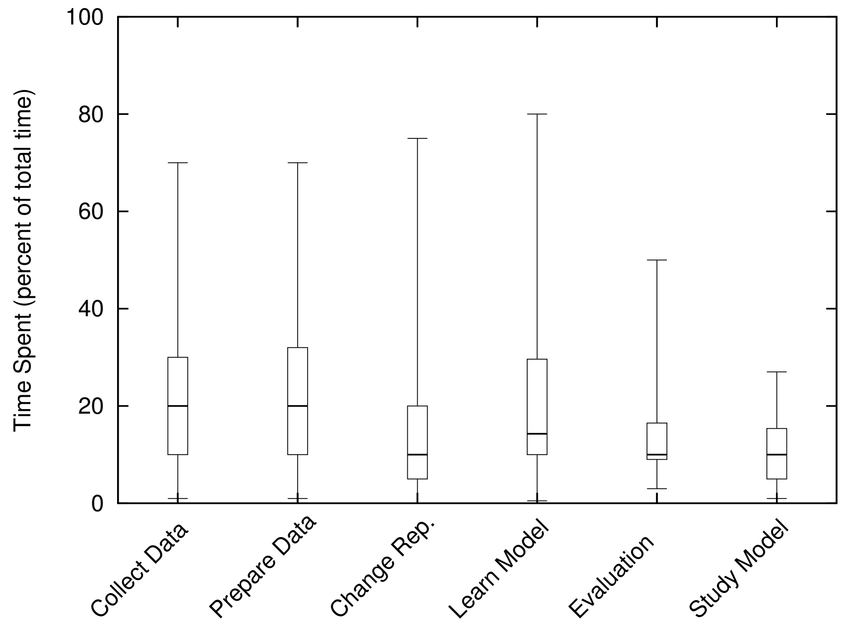
- Data mining and KDD are often used interchangeably.
- Actually data mining is only a part of the KDD process.

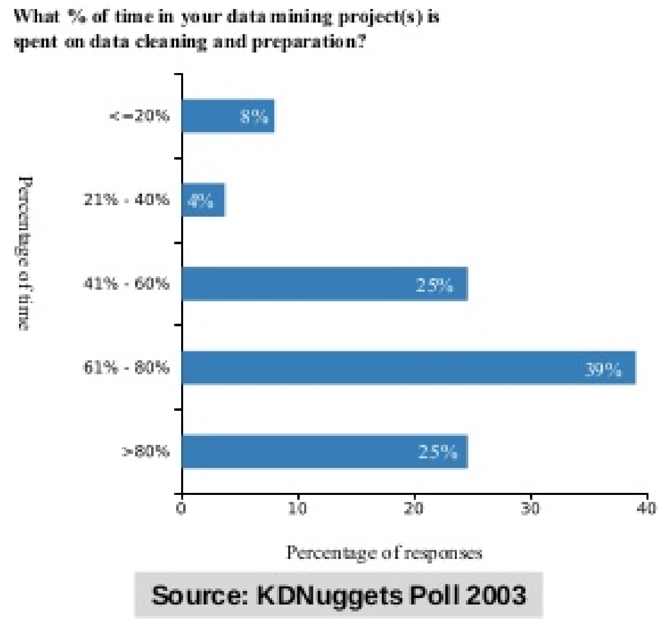
Knowledge Discovery in Data (KDD) Process




See A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)

# Data Mining (Model Building) is less than half of Data Mining



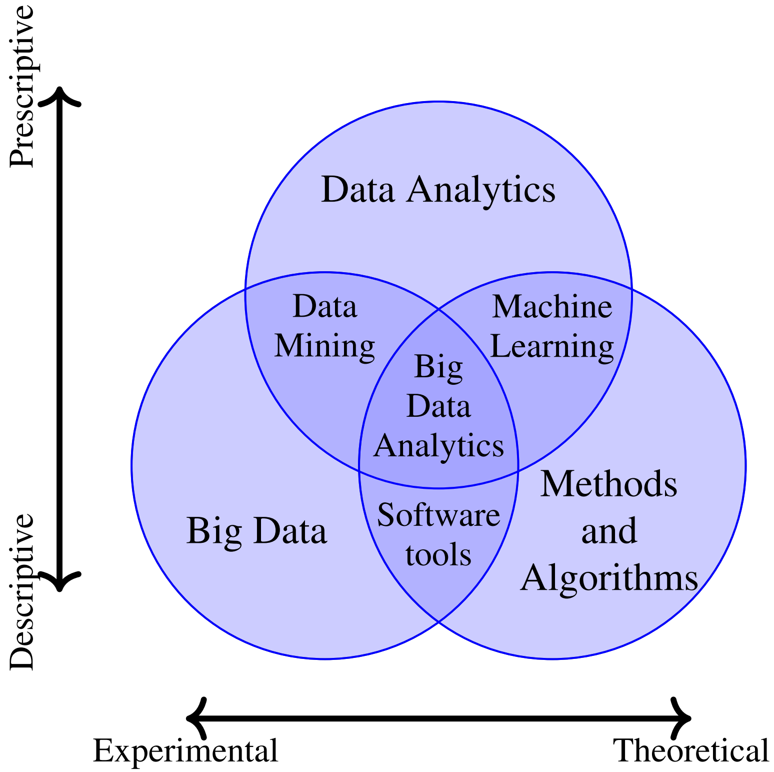



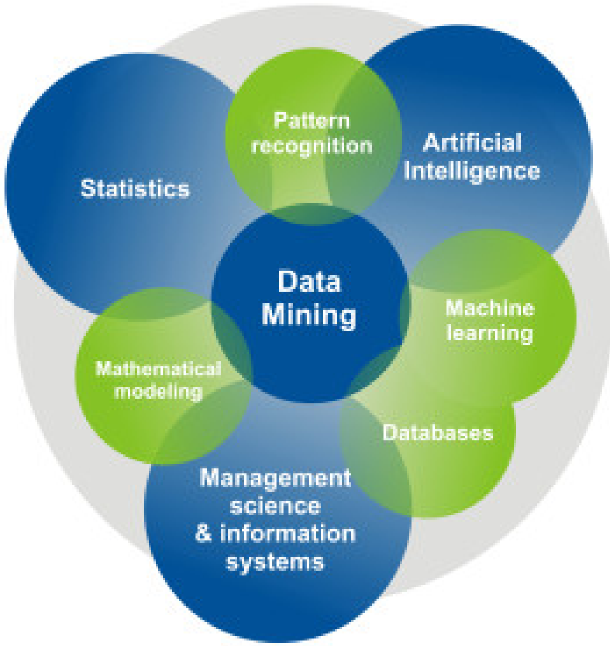

- Boxplots: median is 20% on collecting data, 20% on preparing data, and 10% on changing data representation — all before starting on model.
- Bar chart — data cleaning and preparation consumes at least 80% of project time for 25% of the participants, and 61% to 80% for another 39%.

See Study on the Importance of and Time Spent on Different Modeling Steps, 2012

# Related Disciplines — Data Mining vs Data Analytics vs Data Science†

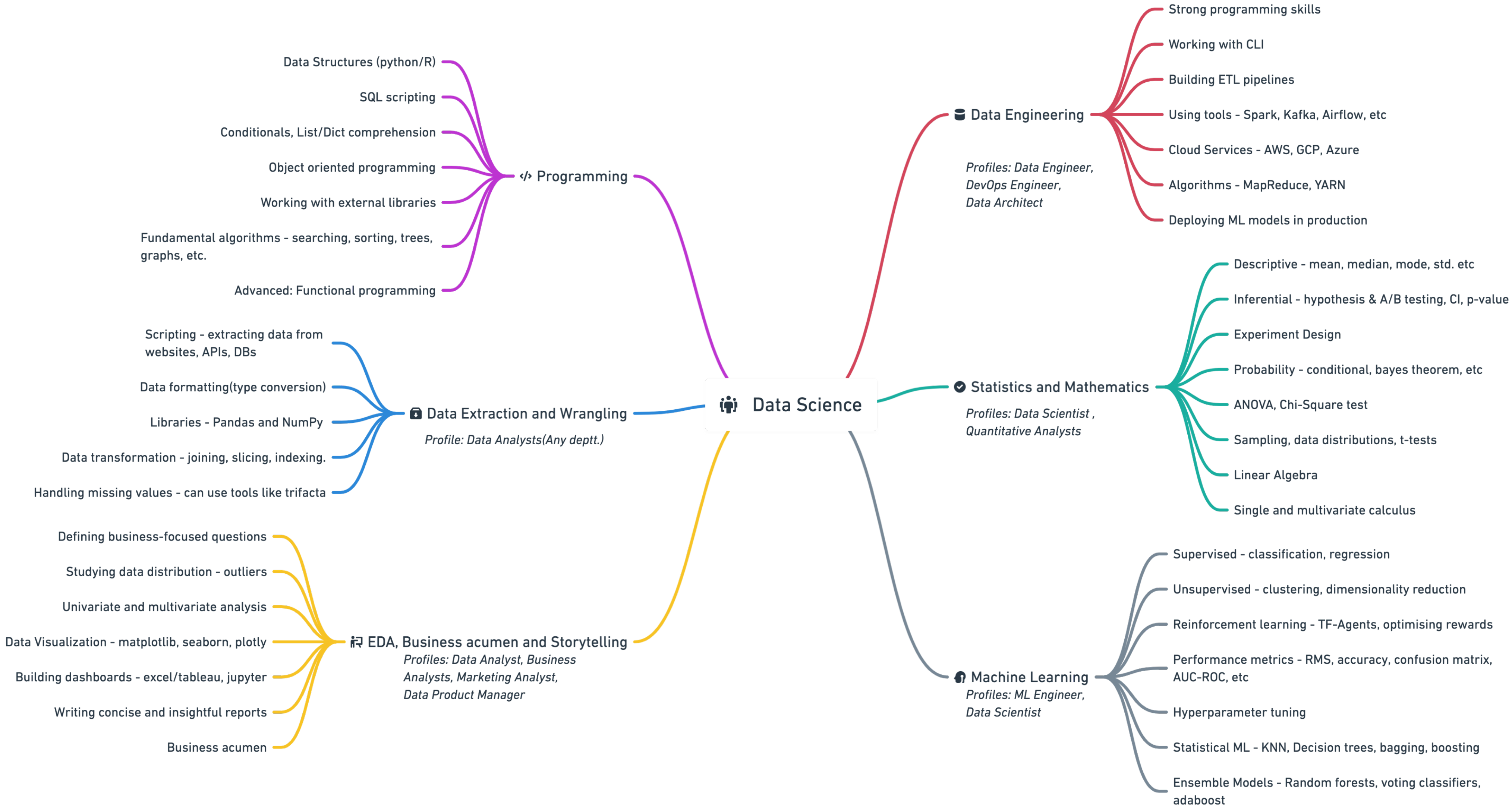
- Data Mining is about finding the patterns in a data set, and using these patterns to make predictions.
- Data Science is a field of study which includes everything from Big Data Analytics, Data Mining, Predictive Modelling, Data Visualisation, Mathematics, and Statistics.



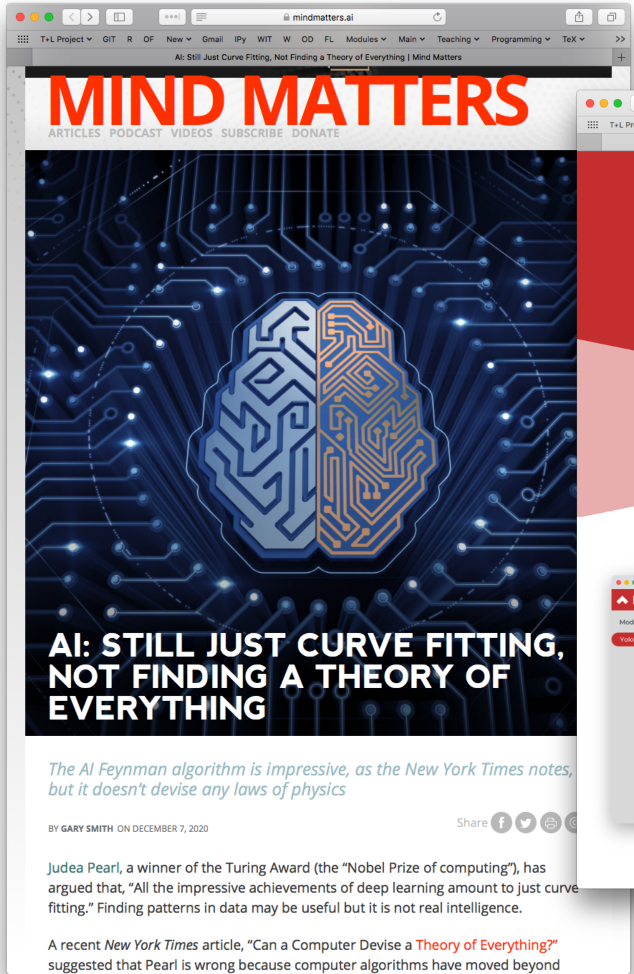

*In other words, have we titled this module correctly? Probably not, and it should be called Data Analytics 2 or Data Science 2

# Data Science Mind Map

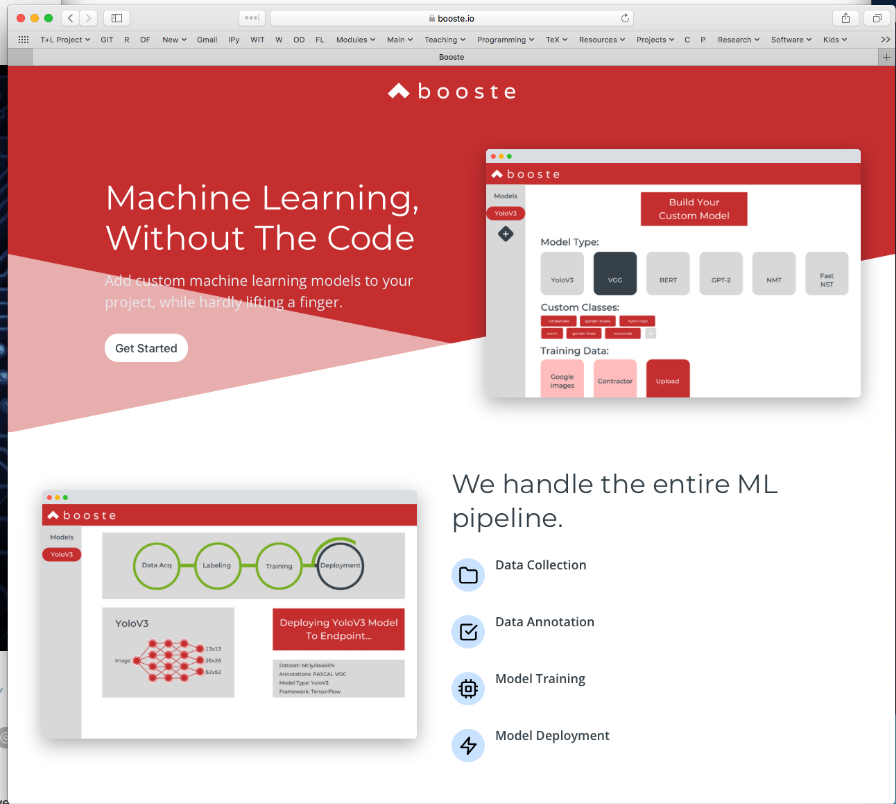


# Data Science in 2021 — ML Models as assets, ML Deployment Services

Not all are believers. . .



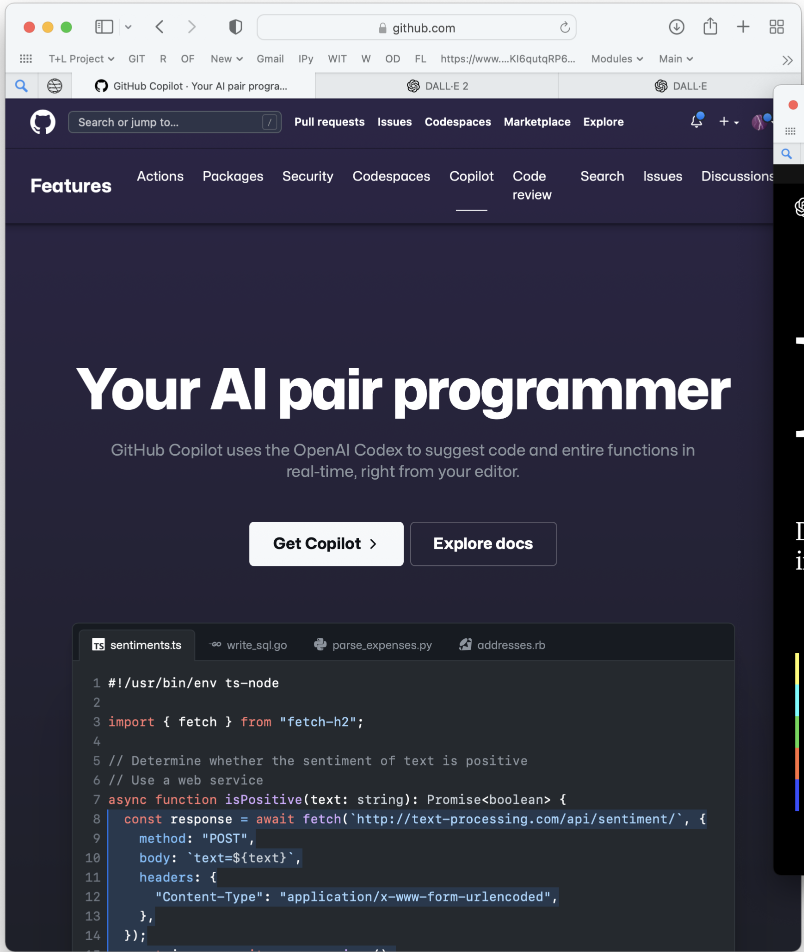

. . . lower barriers and models as assets . . .



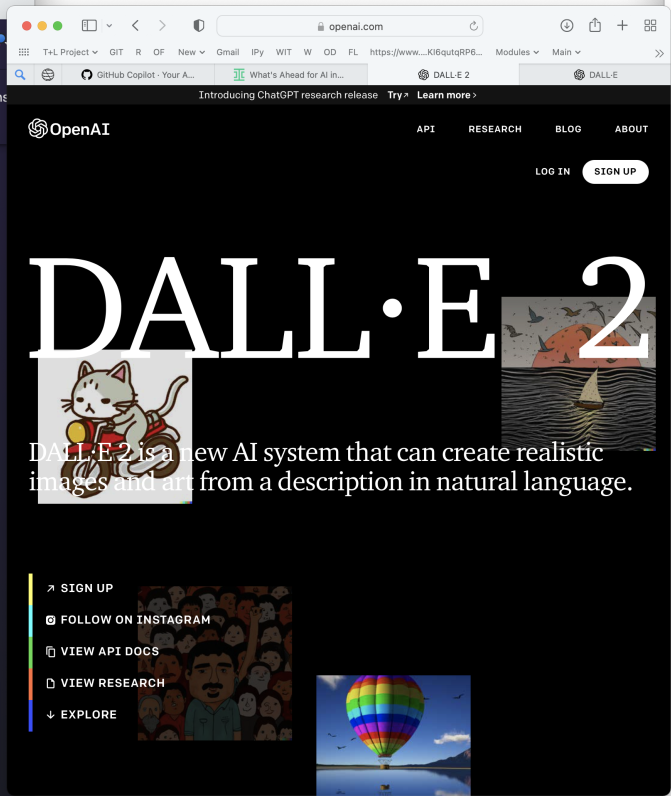

. . . MLOps

# Data Science in 2022 — Generative AI and LLM

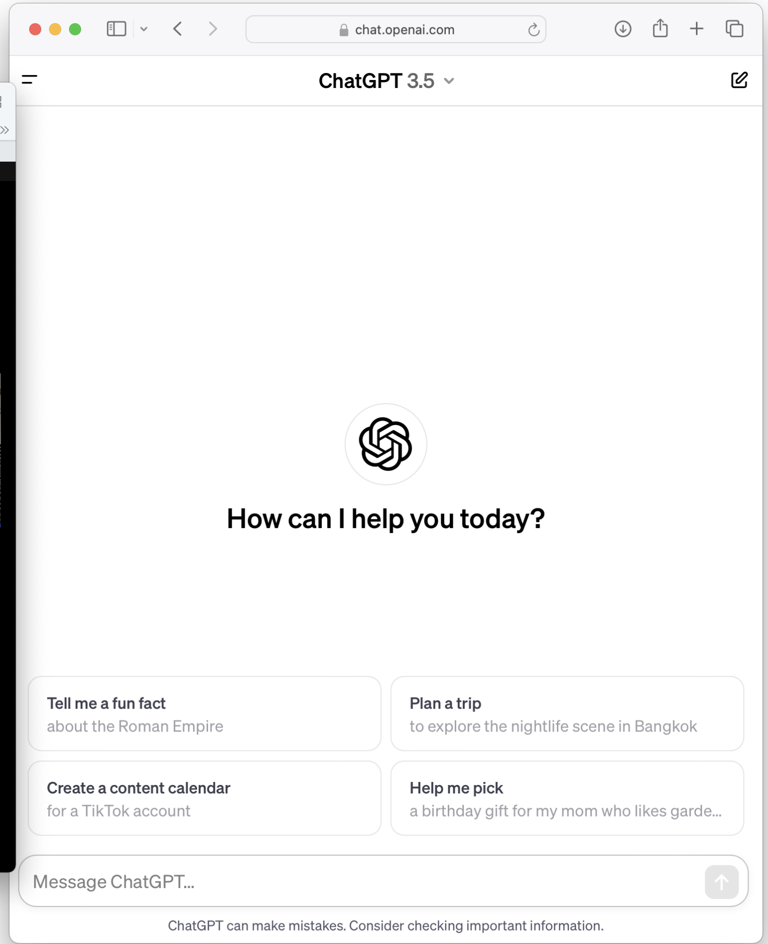
Generating code . . .




. . . images from text . . .

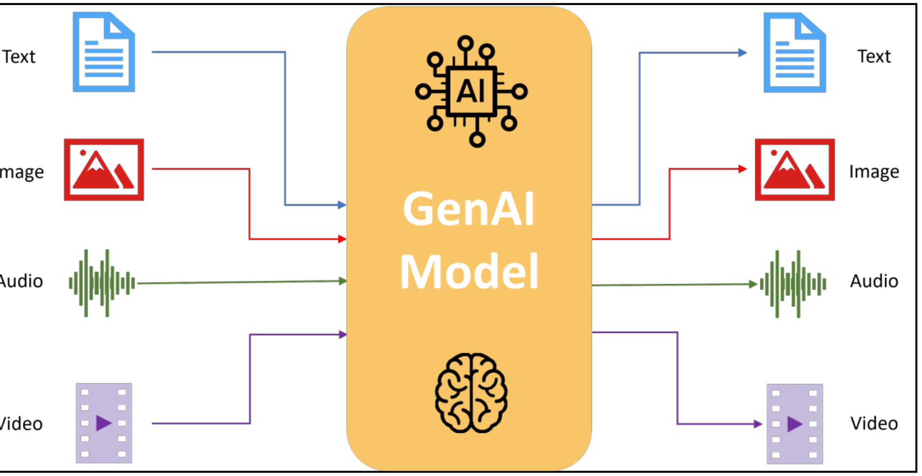



. . . any text (using chatGPT)

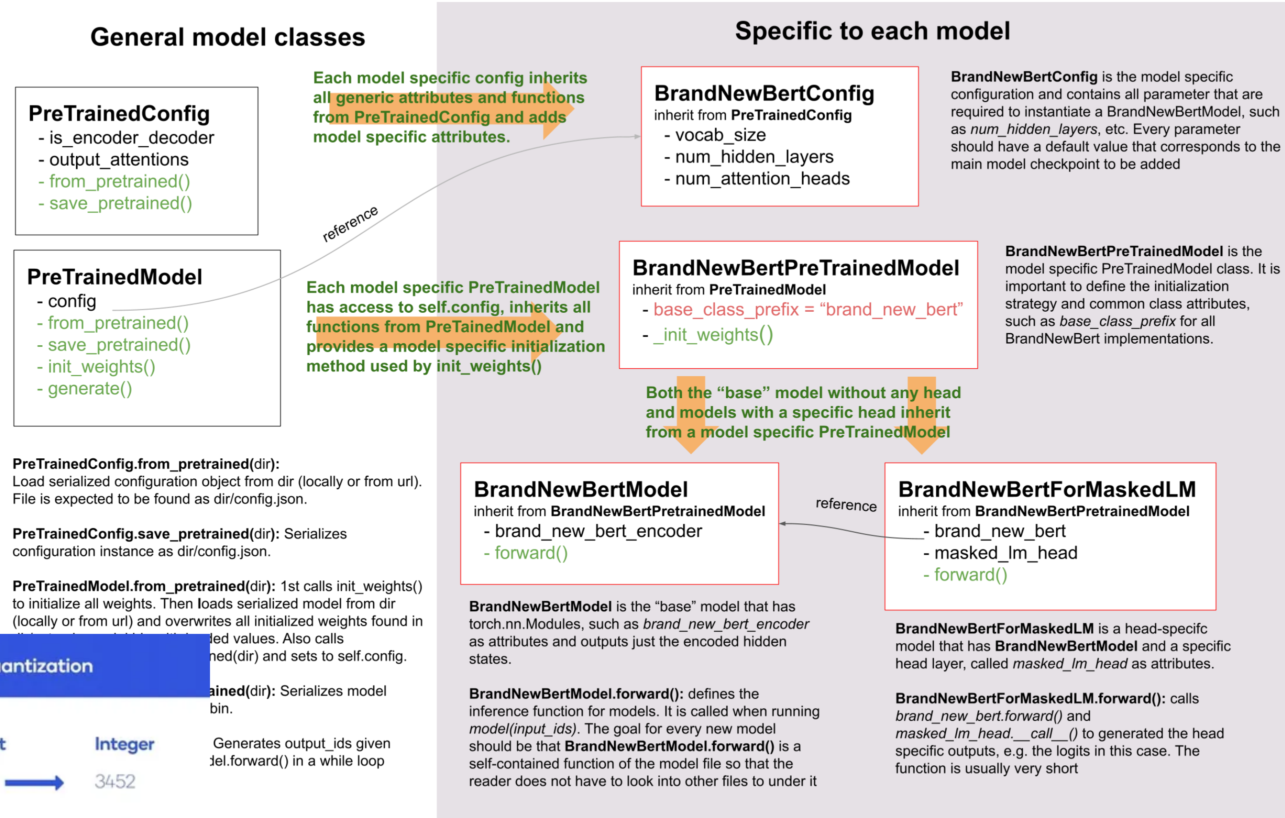


# Data Science in 2023 — Generative AI going Mainstream

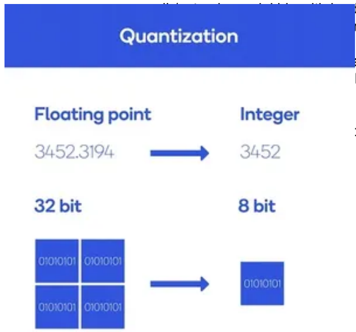
## Multimodel models




## Sharing and reusing of models (Hugging Face)

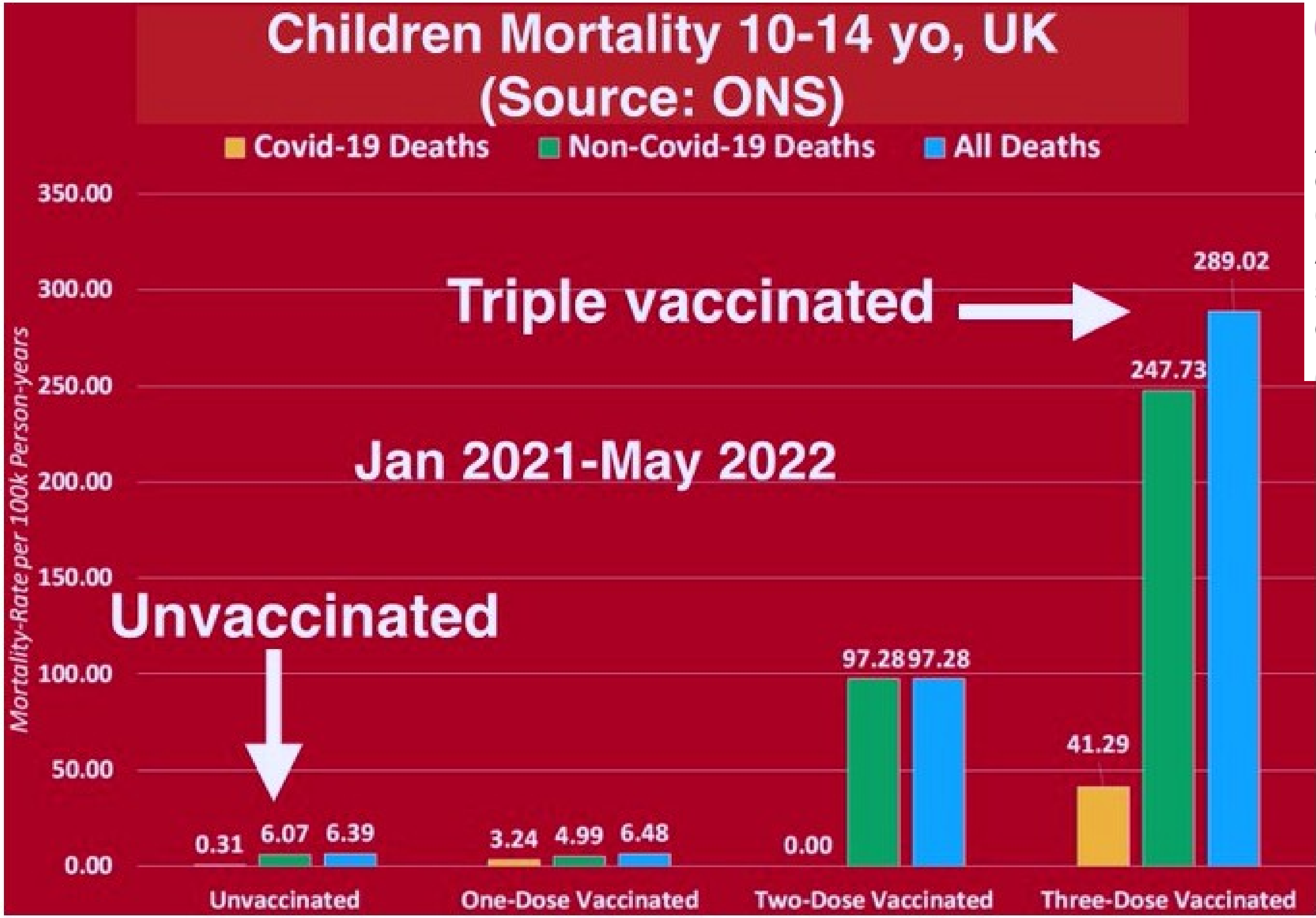


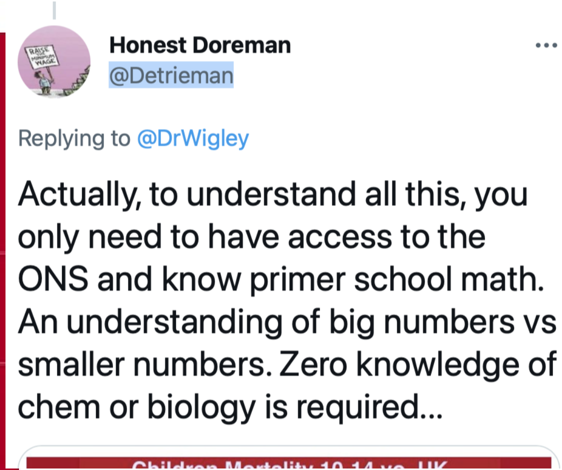





Open sourcing of models

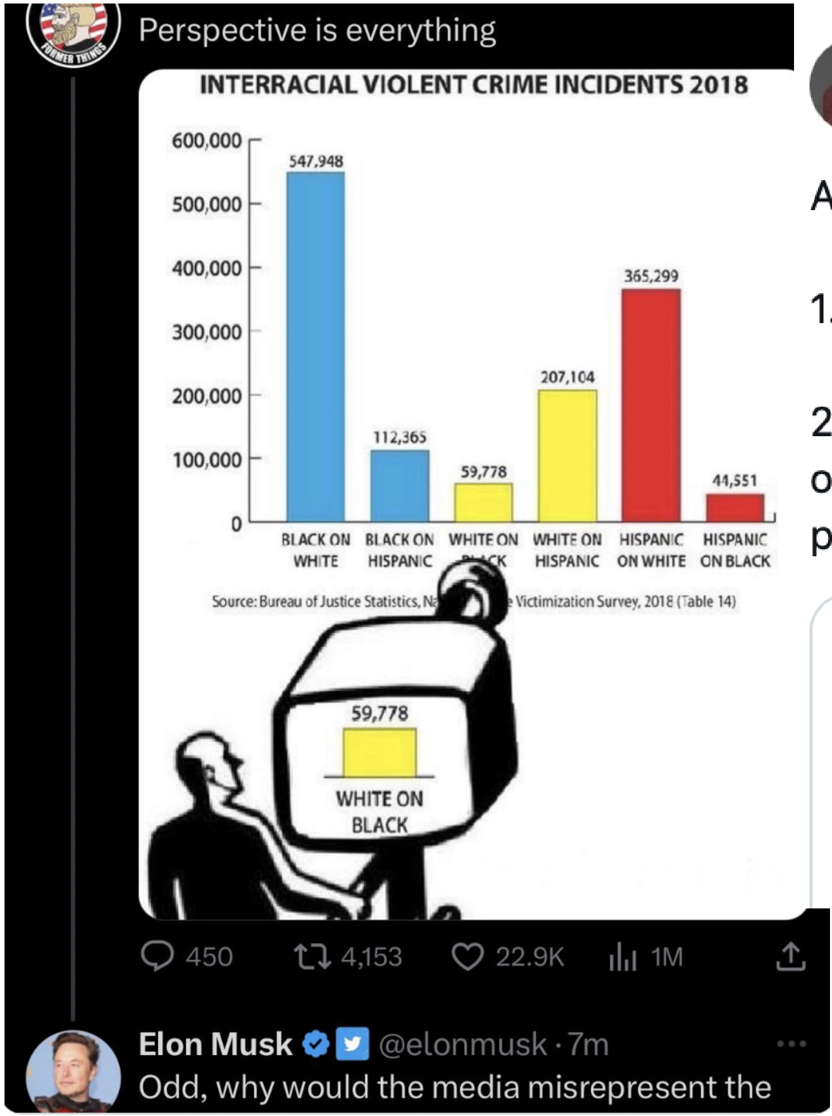
# Lies, Dammed Lies and Statistics




— Jan 13, 2023

# Lies, Dammed Lies and Statistics




**Keith Boykin**
@keithboykin

A factual rebuttal to Elon Musk's crime meme.

1. White offenders cause more violent crime than every other group.

2. The rates at which white victims experience violent crime from Black offenders is similar to the overall percentage of Black people in the population (14.6%).

> 🔥**Kareem Carr | Statistician** 🔥 @kareem_carr · May 8, 2023
>
> I know a lot of you wanted a technical breakdown of this meme so here it is!
>
> I don't think you will find this level of detail anywhere else so keep reading if you don't want to miss out.

— May 9, 2023

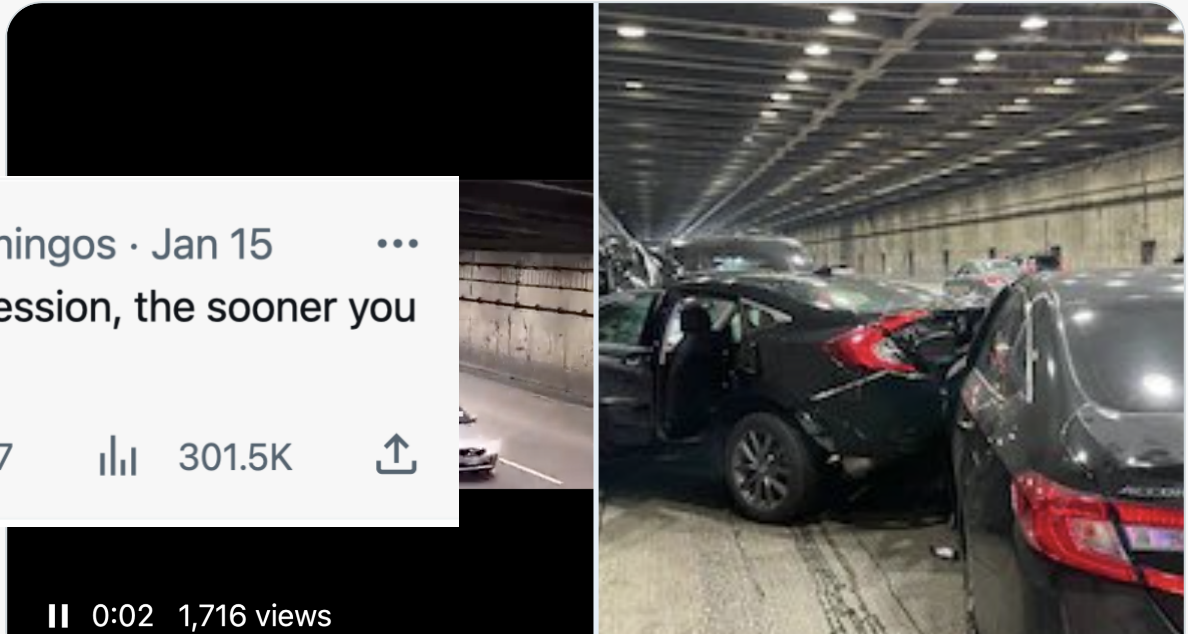
# Hype ? Again ?

This article is more than **6 years old**

## Two years until self-driving cars are on the road – is Elon Musk right?

**The Tesla CEO has proclaimed that autono**
**problem' but tech and executives in recen**
**their expectations**

#Tesla in 'self-driving mode' slams on brakes in tunnel for no reason. The car caused an eight-vehicle crash that injured nine people. Just hours before the crash **#Musk** had triumphantly announced that Tesla's "Full Self-Driving" capability was available in North America.

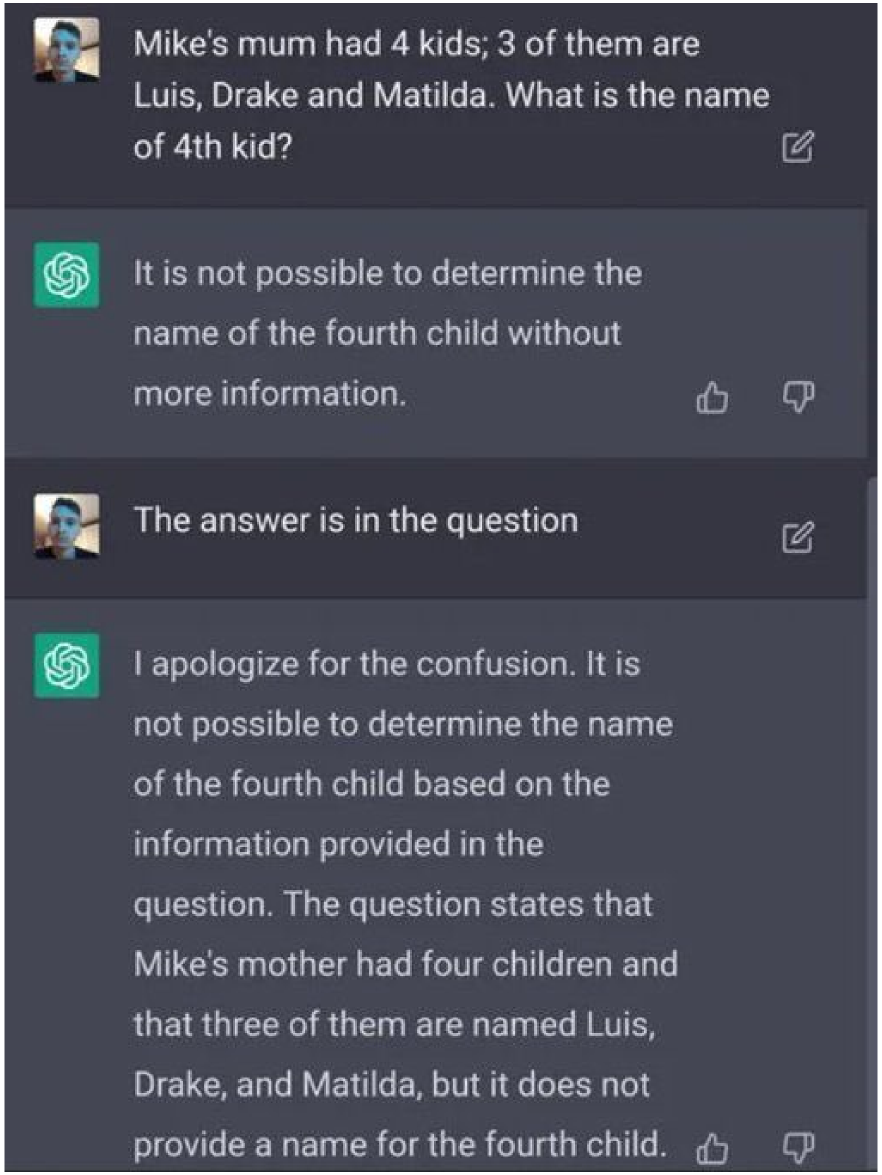



**Pedro Domingos** @pmddomingos · Jan 15

The less you know about a profession, the sooner you think it'll be replaced by AI.

70 421 2,357 301.5K

# But chatGPT can do . . .




# Delivery

## Resources

- All lecture slides, handouts and datasets: GitHub — setu-datamining2.github.io/live
- All activities: quizzes and assignments: Moodle — moodle.wit.ie/course/view.php?id=199957

## Delivery

- Two 1-hour lectures and one 2-practical session.
  - Lecture sessions can tend to get very non-interactive so to help avoid this please ask questions.
  - Default mode for lectures and practical sessions is on campus.
- Slack
  - Will use this for all last minute posts and individual/group Q+A, particularly for assignments.

## Strategy to handle module

- Prepare — review material in advance of the sessions, install/download the software/datasets.
- Interact — yes, this is rich coming for an introvert mathematician, but we live in strange times.
- Time management — give tasks a serious/focused effort, but when stuck ask for help.

# Assessment Structure — 100% Continuous Assessment

## Covering skills

- Data Wrangling + Feature Engineering (pandas and friends)
- NLP, Text processing (regex)
- Model building and optimisation (skilearn, PyTorch, PyTorch Lightning, . . . )

## Breakdown

- Metric:
  - 20% Student engagement + 80% Demonstration of skills/understanding
- Activities:
  - Moodle quizzes based on analysing datasets / model building / etc.
  - Data science problems with mixture of Kaggle style grading and traditional grading.

## Calandar

- Week 14/15 end of semester individual review interview (similar to S1).
- $\underbrace{\text{5 weeks + reading week + 5 weeks + Easter break (2 weeks) + 2 weeks}}_{\text{12 teaching weeks}}$ +3 weeks for CA

# Three Components of a Machine Learning Problem

It is easy to get lost among the multitude of choices one needs to make when given data mining problem. A good decomposition is the following:

| Representation | Evaluation | Optimization |
|---|---|---|
| Instances | Accuracy/Error rate | Combinatorial optimization |
|     $K$-nearest neighbor | Precision and recall |     Greedy search |
|     Support vector machines | Squared error |     Beam search |
| Hyperplanes | Likelihood |     Branch-and-bound |
|     Naive Bayes | Posterior probability | Continuous optimization |
|     Logistic regression | Information gain |     Unconstrained |
| Decision trees | K-L divergence |         Gradient descent |
| Sets of rules | Cost/Utility |         Conjugate gradient |
|     Propositional rules | Margin |         Quasi-Newton methods |
|     Logic programs | |     Constrained |
| Neural networks | |         Linear programming |
| Graphical models | |         Quadratic programming |
|     Bayesian networks | | |
|     Conditional random fields | | |

† A Few Useful Things to Know about Machine Learning, Domingos, 2012.

# 3 Components — Representation

| Representation | Evaluation | Optimization |
|---|---|---|
| Instances | Accuracy/Error rate | Combinatorial optimization |
| $K$-nearest neighbor | Precision and recall | Greedy search |
| Support vector machines | Squared error | Beam search |

Representation refers to formulating the problem as a machine learning problem — typically a classification problem, a regression problem or a clustering problem.

- How do we represent the input?
- What features to use?
- How do we learn additional features?
- With each type of problem, we have multiple subtypes.
  For example which classifier? a decision tree, a neural network, a support vector machine, a hyperplane that separates the two classes etc.

# 3 Components — Evaluation

| Representation | Evaluation | Optimization |
|---|---|---|
| Instances | Accuracy/Error rate | Combinatorial optimization |
| $K$-nearest neighbor | Precision and recall | Greedy search |
| Support vector machines | Squared error | Beam search |

Evaluation refers to an objective function or a scoring function, to distinguish good models from a bad model.

- For a classification problem, we need this function to know if a given classifier is good or bad. A typical function can be based on the number of errors made by the classifier on a test set, using precision and recall.
- For a regression problem, it could be the squared error, or likelihood. Do we include regularisation? etc

# 3 Components — Optimisation

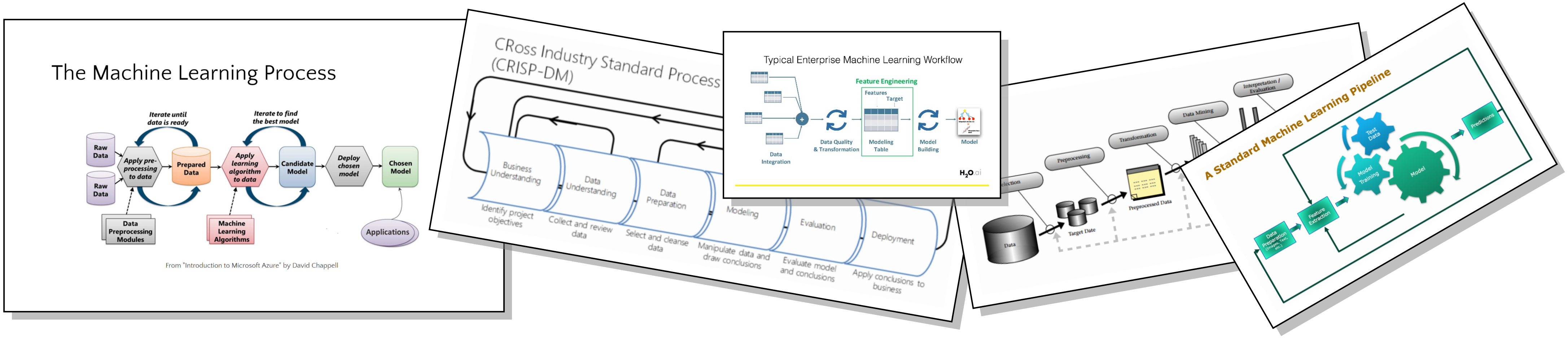
| Representation | Evaluation | Optimization |
|---|---|---|
| Instances | Accuracy/Error rate | Combinatorial optimization |
| $K$-nearest neighbor | Precision and recall | Greedy search |
| Support vector machines | Squared error | Beam search |

Optimisation is concerned with searching among the models in the language for the highest scoring model.

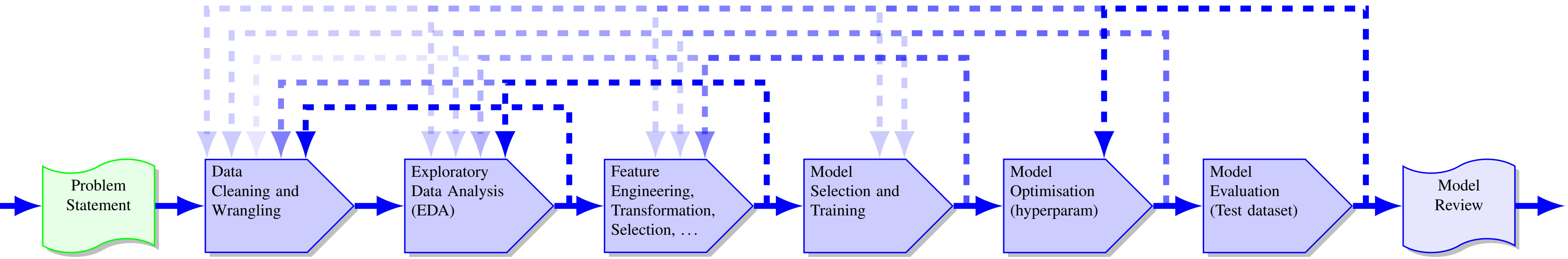
- How do we search among all the alternatives?
- Can we use some greedy approaches, branch and bound approaches, gradient descent, linear programming or quadratic programming methods.

# Data Mining Workflow

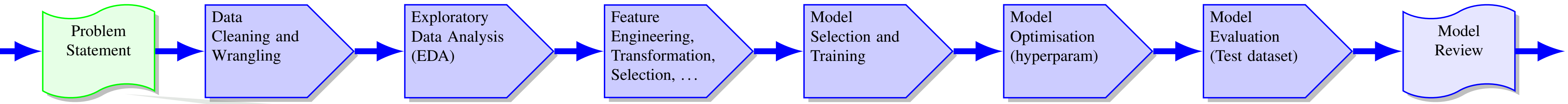
## There are many, many . . .




## So why not make YADMW (Yet Another Data Mining Workflow)?



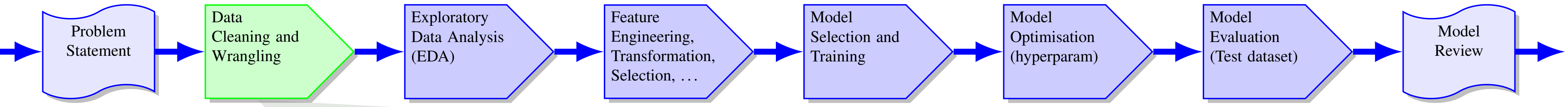
Data mining / Machine Learning workflow




## What type of problem is it?

- Exploratory data analysis
  Do we just want to see what the data says?
- Association / Rule finding
  Are we searching for relations/patterns?
- Hypothesis testing (Statistical)
  Do we have a theory we wish to test?
- Model building
  Do we wish to build a representation of some pattern within the data?
  - Supervised $\Leftarrow$ data spilt into input variables (features) and output variable(s) (target(s))
    - Classification (target is categorical) vs regression (target is continuous)
  - Unsupervised $\Leftarrow$ no target
    - Clustering — grouping similar cases

Data mining / Machine Learning workflow



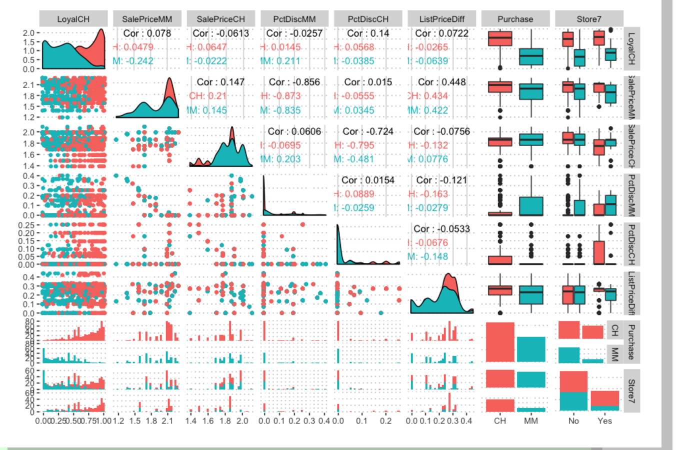

## How to import and prepare data for subsequent analysis/processing?

- Multiple file formats
  - Pandas supports a wide collection of file formats but default options often need to be changed to suit data.
  - Main file format (Comma Separated Values (csv)) does not support meta-data, is slow, and results in large files $\Longrightarrow$ use other formats (pickle, feather) to store datasets between steps in the workflow.
- Assumptions made by input parser can be important (i.e., bite you when you least expect)
  - Scientists rename human genes to stop Microsoft Excel from misreading them as dates
  - Pandas vs excel use different heuristics to decide on data type of each variable.
- Sub-tasks
  - Check dimension (number of rows/cases, number of columns/variables).
  - Check data types (categorical, ordinal, or numerical (discrete/continous)) of each variable.
  - Check for missing values, encoding errors, etc.
  - Merge tables, apply filters, and general data wrangling to generate (tabular) dataset suitable for EDA.

Data mining / Machine Learning workflow

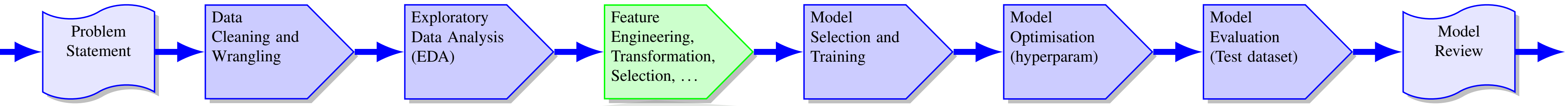
Problem Statement → Data Cleaning and Wrangling → Exploratory Data Analysis (EDA) → Feature Engineering, Transformation, Selection, … → Model Selection and Training → Model Optimisation (hyperparam) → Model Evaluation (Test dataset) → Model Review

## What is the data telling us?




- Univariate descriptive statistics — examine each variable
  - What are typical values?
  - What is the variation / spread / range?
  - What does the data look like …bell curve, bath tub curve, etc. ?
- Bi- / multi- variable descriptive statistics
  - Identifying relationships between variables.
- All descriptive statistics methods summarise data:
  - ✔ A summary is good since it helps to focus on simpler and important aspects.
  - ✘ A summary is bad if it focuses on irrelevant or the wrong aspects.
  - ⇒ Need to use multiple methods, be aware of their strengths/deficiencies.

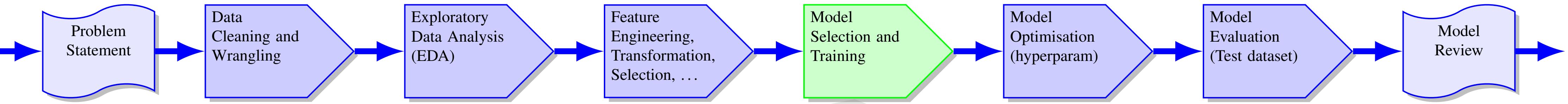
Data mining / Machine Learning workflow




Can we transform, encode/bin, select, . . . , the given features to improve model training?

- Better features can mean:
  - Better model performance and reduce training times.
  - Simpler models become applicable — think linear/logistic regression.
  - More explainable models — the future of machine learning (hopefully).
  - Cheaper and easier models to deploy.
- Feature selection reduces the number of features used in the model:
  - Drop features that have low variability.
  - Drop features that have no relation to target.
  - Drop features that are highly related to other features — multicollinearity.
  - Keep features whose addition to model have the largest improvement in model score.
- Feature extraction merges existing features to generate (hopefully) fewer features with essentially all the variation.

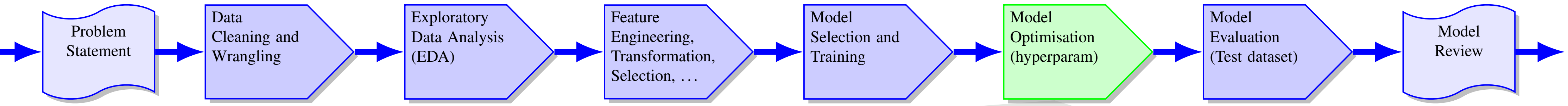
Data mining / Machine Learning workflow




## Which models are suitable?

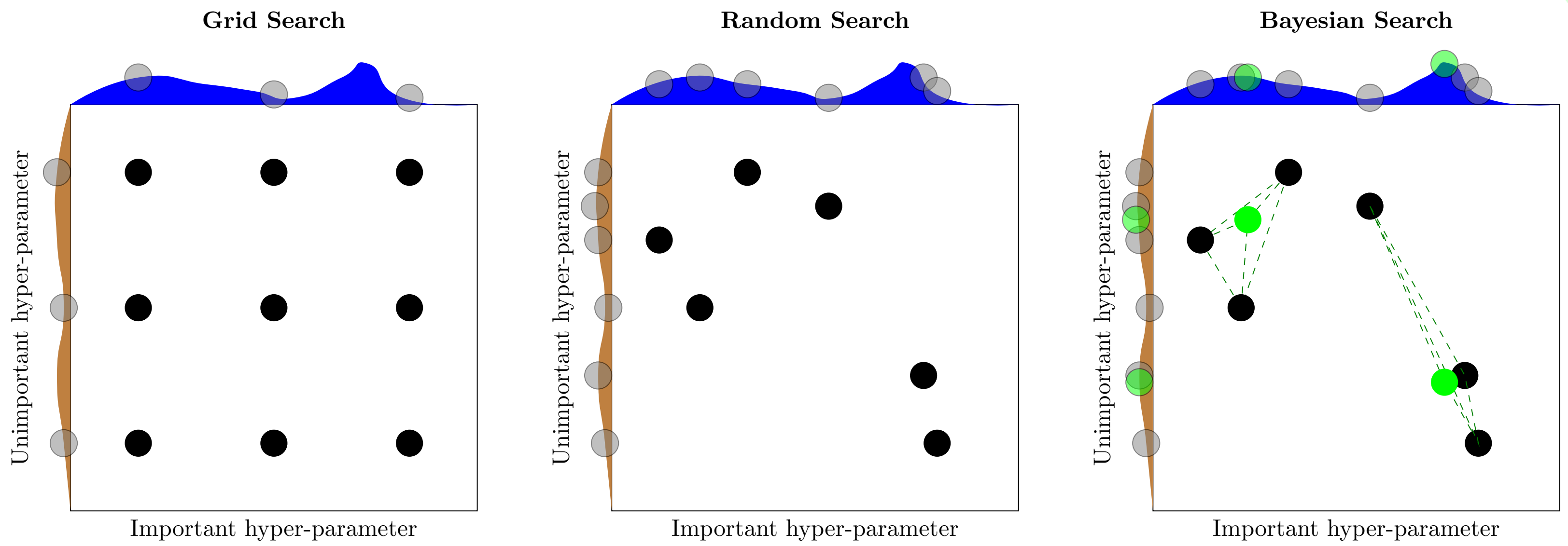
- Models vary greatly in terms of capabilities/deficiencies — usually aim to build a short list of candidate models, which are subsequently optimised in the next step.
- Select models based on different algorithms/approaches.
- Select (loss function and) evaluation metric.
  - Loss function is used to train model, evaluation metric is used to evaluate model (post training).
- Relative model performance can help identify issues with data.
  - Outliers can negatively affect linear regression but have smaller impact on decision tree based models.

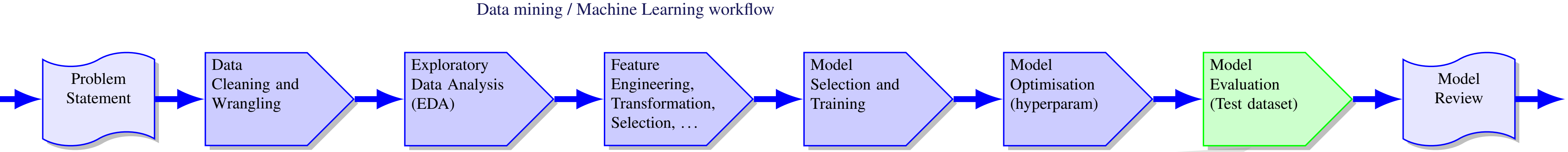
Data mining / Machine Learning workflow




## How do we determine optimal values of the hyper-parameters?

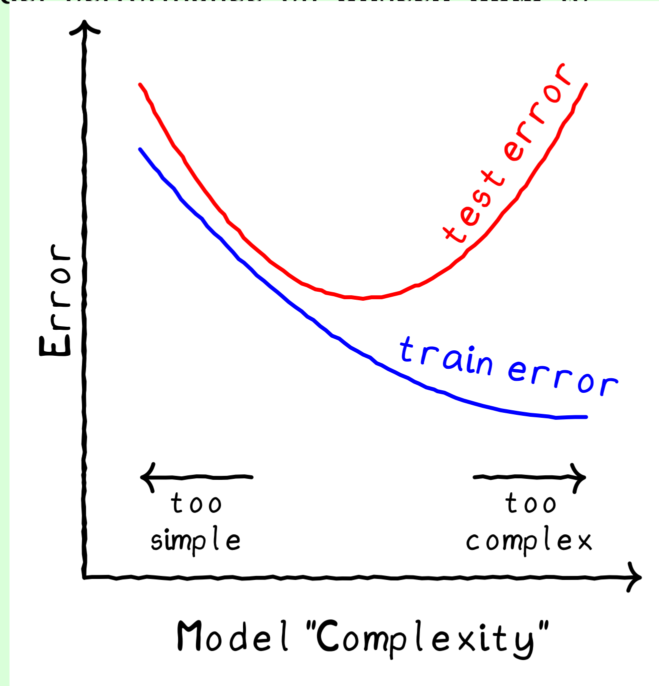
- Most models have options which control how a model "learns" from the training data.
- Three search stratagies: Grid search $<$ Random search $\ll$ Bayesian search



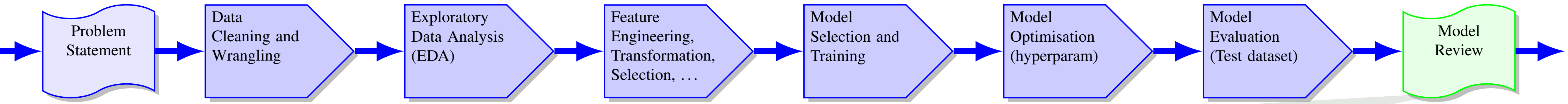



## How well does the model generalise (to unseen data)?

- In the machine learning approach (vs statistical approach) we rely on model performance on **unseen data** to evaluate models.
  - Split data into train/test, only use train dataset for all modelling decisions.
  - Data leakage (MachineLearningMastery article), where information outside the train dataset is used in model building.
- Is there evidence for overfitting?
  - Does the model perform much better on training dataset than on the test dataset?
- Multiple techniques to address overfitting:
  - Regularisation (linear / logistic regression).
  - Trimming (decision trees).
  - Dropout (neural networks), Batch normalisation (CNN).



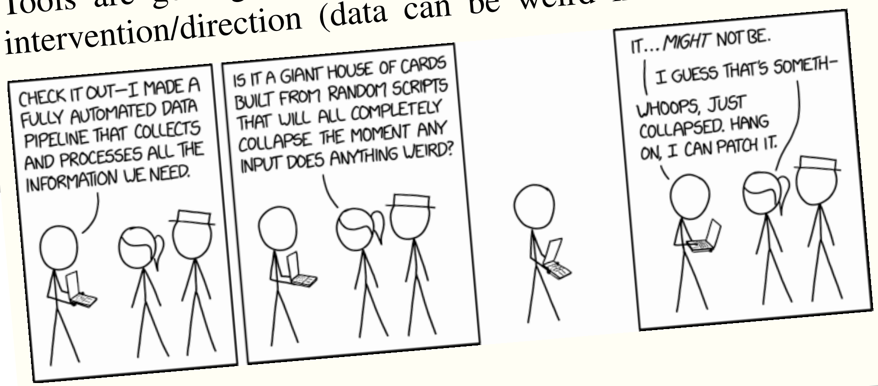
Data mining / Machine Learning workflow




How well have we addressed the problem statement?

- A what level of accuracy (or other metrics) does a model become useful?
  - This is a business, medical, ... decision
  - The larger the relative payoff the weaker the model can be and still be useful.
- OK, finally ready to implement/deploy model ...
  - Separate skillset / concerns
  - MLOps = ML + DevOps
  - Monitoring of model drift needed.
- towards data science What is MLOps — Everything You Must Know to Get Started

**Q:** Why don't we automate all of this s~~h~~tuff?
Tools are getting better and easier to use, but need intervention/direction (data can be weird in weird ways)




– xkcd.com/2054