

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, 2026

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!

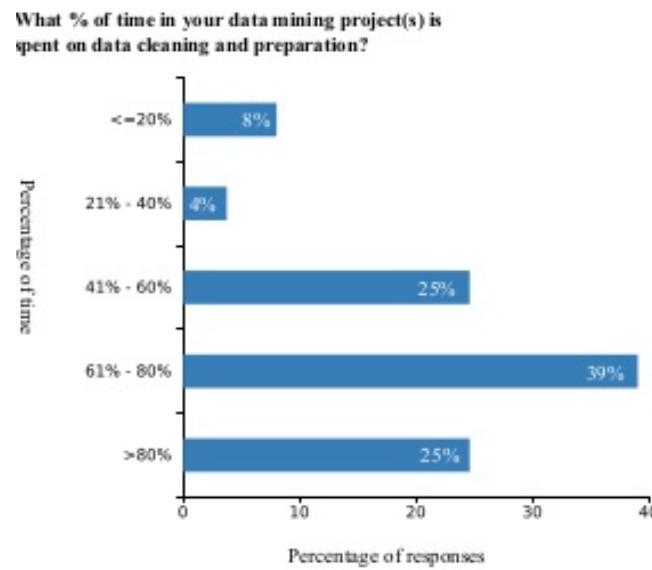
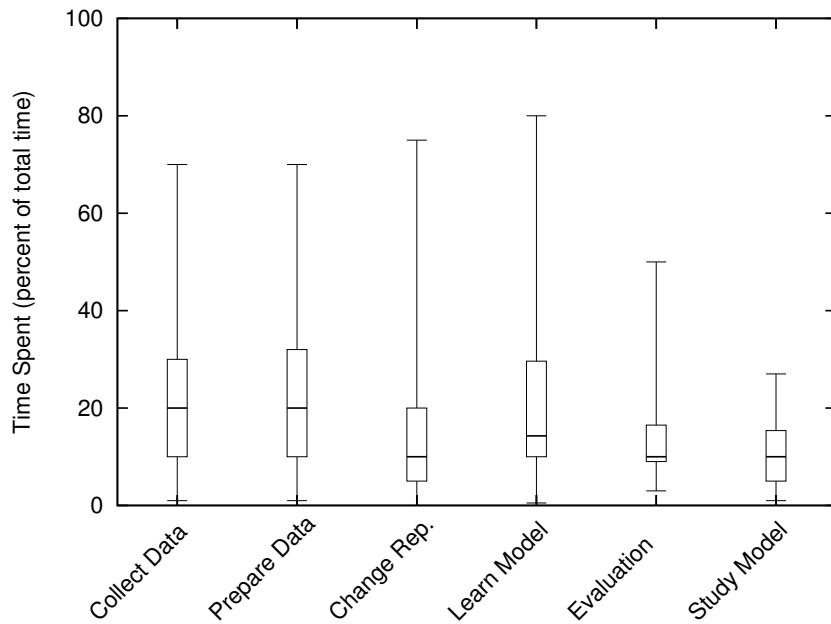
Necessity is the mother of invention \Rightarrow Data Mining \approx Automated analysis of massive data sets.

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 (Model Building) is less than half of Data Mining

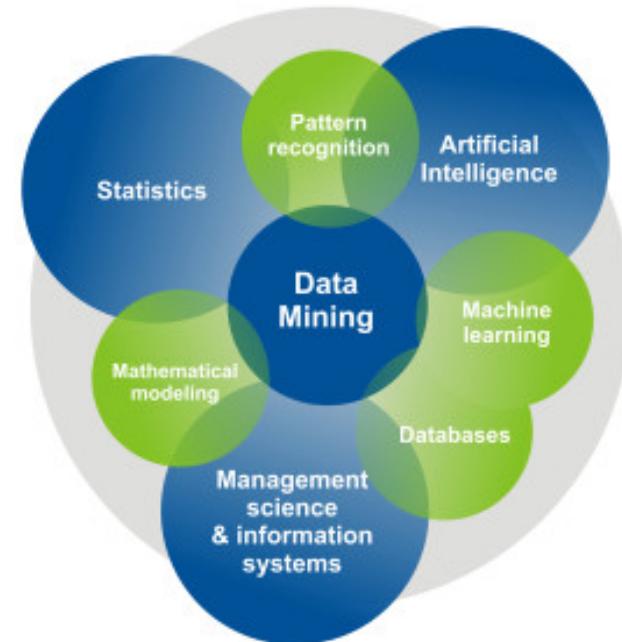
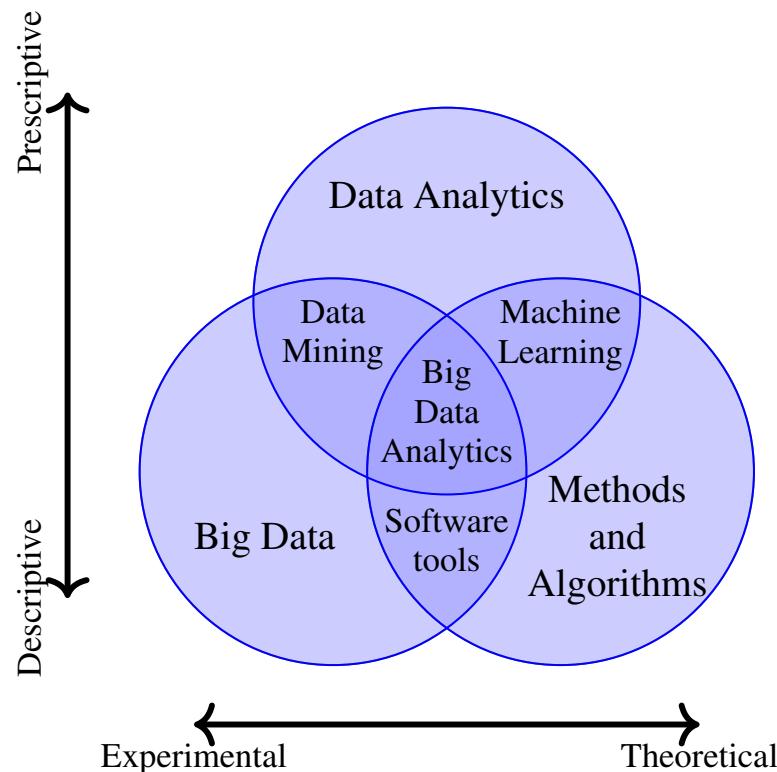


- Boxplots: median is 20% on collecting data, 20% on preparing data, and 10% on changing data representation — all before starting on model.
- Bar chart — data cleaning and preparation consumes at least 80% of project time for 25% of the participants, and 61% to 80% for another 39%.

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



What? Why? and How? What?

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

Not all are believers...

The screenshot shows the homepage of mindmatters.ai. The header features the text "MIND MATTERS" in large red letters. Below it is a large image of a brain composed of circuit boards, set against a dark blue background with glowing circuit lines. A prominent headline reads "AI: STILL JUST CURVE FITTING, NOT FINDING A THEORY OF EVERYTHING". Below this, a quote from the New York Times is displayed: "The AI Feynman algorithm is impressive, as the New York Times notes, but it doesn't devise any laws of physics". At the bottom, there's a section about Judea Pearl's Turing Award win and a quote from him. The footer includes social sharing icons and a copyright notice.

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

The screenshot shows the booste.io website. It features a red header with the text "Machine Learning, Without The Code" and a subtext "Add custom machine learning models to your project, while hardly lifting a finger." Below this is a "Get Started" button. To the right, there are two windows demonstrating the platform's interface. The top window shows a "Build Your Custom Model" interface with options for "Model Type" (YoloV3, VGG, BERT, CRF-2, NMT, Fast NST) and "Custom Classes" (Google Images, Contractor, Upload). The bottom window shows a flowchart of the ML pipeline: Data Acc → Labeling → Training → Deployment, with a specific "Deploying YoloV3 Model To Endpoint..." step highlighted.

We handle the entire ML pipeline.

- Data Collection
- Data Annotation
- Model Training
- Model Deployment

... MLOps

The screenshot shows the mlflow.org website. The header features the text "mlflow" and "A platform for the machine learning lifecycle". Below this is a large blue banner with a wavy network pattern. The main content area has a dark blue background with white text. It features a "Latest News" section with links to recent releases (MLflow 1.13.1, 1.13.0, 1.12.1) and an announcement about PyTorch integration. At the bottom, there are three sections: "WORKS WITH ANY ML LIBRARY, LANGUAGE & EXISTING CODE", "RUNS THE SAME WAY IN ANY CLOUD", and "DESIGNED TO SCALE FROM 1 USER TO LARGE ORGS", each accompanied by an icon and a brief description.

What? Why? and How? What?

Data Science in 2022 — Generative AI and LLM

Generating code ...

A screenshot of a web browser window showing the GitHub Copilot interface. The URL is `github.com/T+L Project/GitHub Copilot - Your AI pair program...`. The page features a large heading "Your AI pair programmer" and a subtext explaining that GitHub Copilot uses the OpenAI Codex to suggest code and entire functions in real-time. Below this are two buttons: "Get Copilot >" and "Explore docs". On the right, there's a code editor window displaying a file named `sentiments.ts` with the following code:

```
1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch('http://text-processing.com/api/sentiment/');
9   method: "POST",
10   body: `text=${text}`,
11   headers: {
12     "Content-Type": "application/x-www-form-urlencoded",
13   },
14});
```

... images from text ...

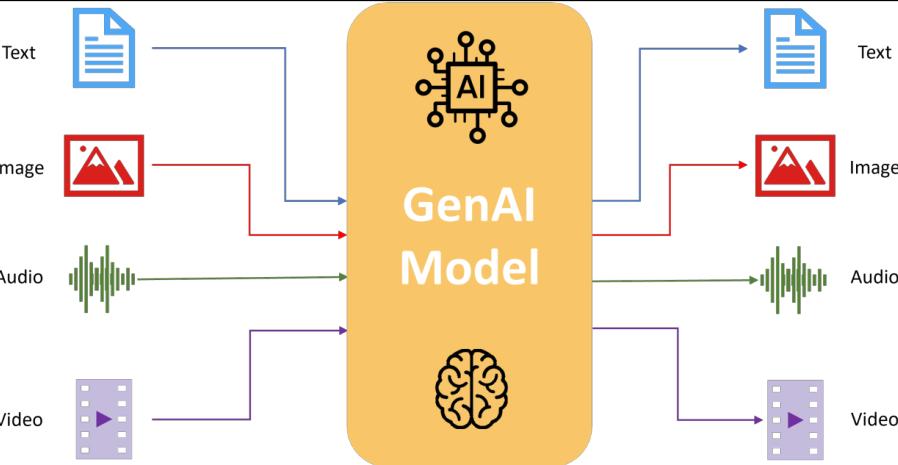
A screenshot of a web browser window showing the DALL-E 2 interface. The URL is `openai.com/DALL-E 2`. The page has a large title "DALL·E 2" and a subtext explaining that it is a new AI system that can create realistic images and art from a description in natural language. It features a sidebar with links like "SIGN UP", "FOLLOW ON INSTAGRAM", "VIEW API DOCS", and "VIEW RESEARCH". Below the sidebar are several generated images, including a cartoon cat, a sailboat on water, and a colorful hot air balloon.

... any text (using chatGPT)

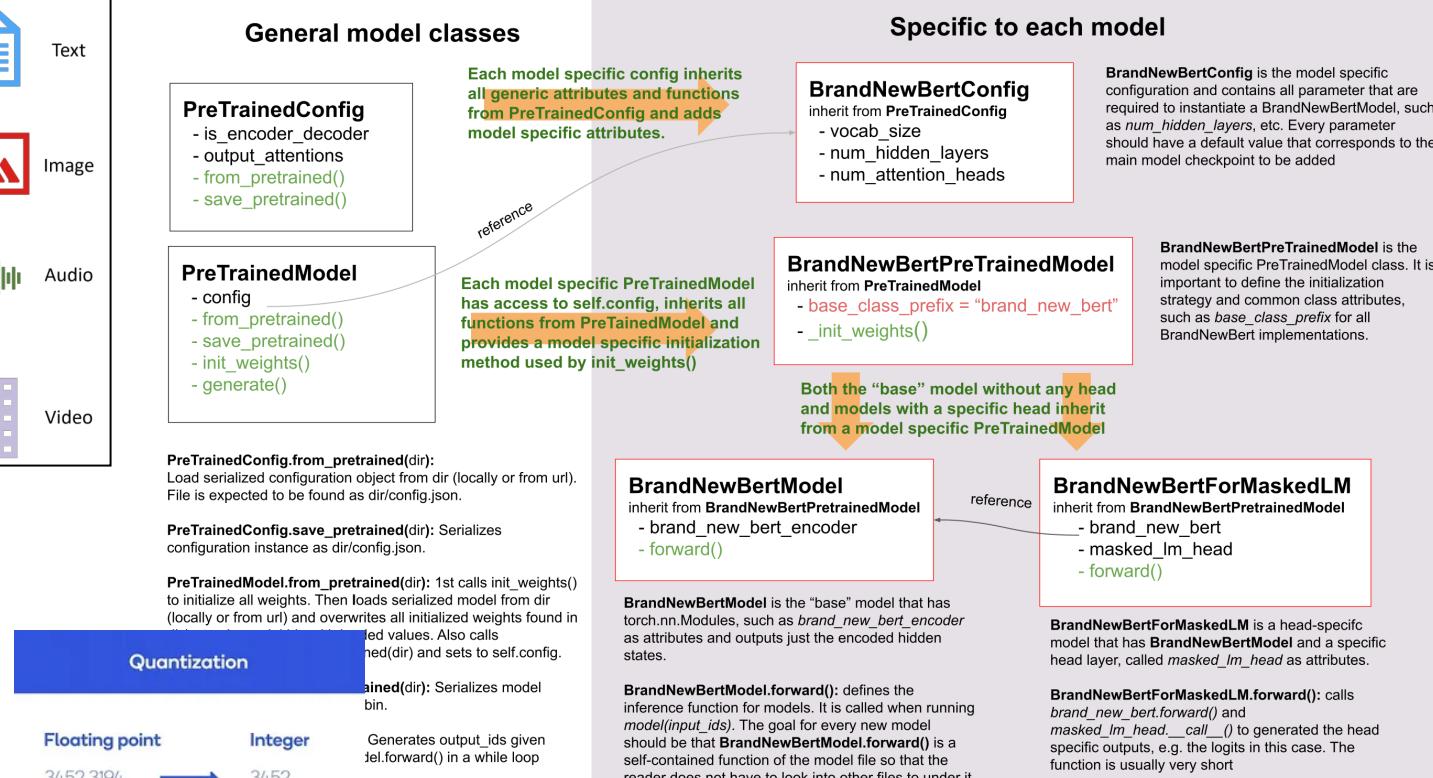
A screenshot of a web browser window showing the ChatGPT 3.5 interface. The URL is `chat.openai.com/ChatGPT 3.5`. The page features a large "How can I help you today?" button with a circular icon. Below it are several AI-generated tasks: "Tell me a fun fact about the Roman Empire", "Plan a trip to explore the nightlife scene in Bangkok", "Create a content calendar for a TikTok account", and "Help me pick a birthday gift for my mom who likes garde...". At the bottom, there's a message input field "Message ChatGPT..." and a note: "ChatGPT can make mistakes. Consider checking important information."

Data Science in 2023 — Generative AI going Mainstream

Multimodel models



Sharing and reusing of models (Hugging Face)



Open sourcing of models

Data Science in 2024 — AI Agents and Enterprise RAG

Retrieval-Augmented Generation (RAG): Grounding LLMs in private enterprise data to reduce hallucinations.

K2VIEW BLOG

May 27, 2025

Enterprise RAG: Beware of connecting LLMs directly to data sources



Hod Rotem, Chief Product Evangelist

Table of Contents

Please select

▶ Enterprise R... 7:43 f in x

When deploying enterprise RAG, you may want to give your LLM's agents and functions direct access your operational systems. But that's not a great idea.

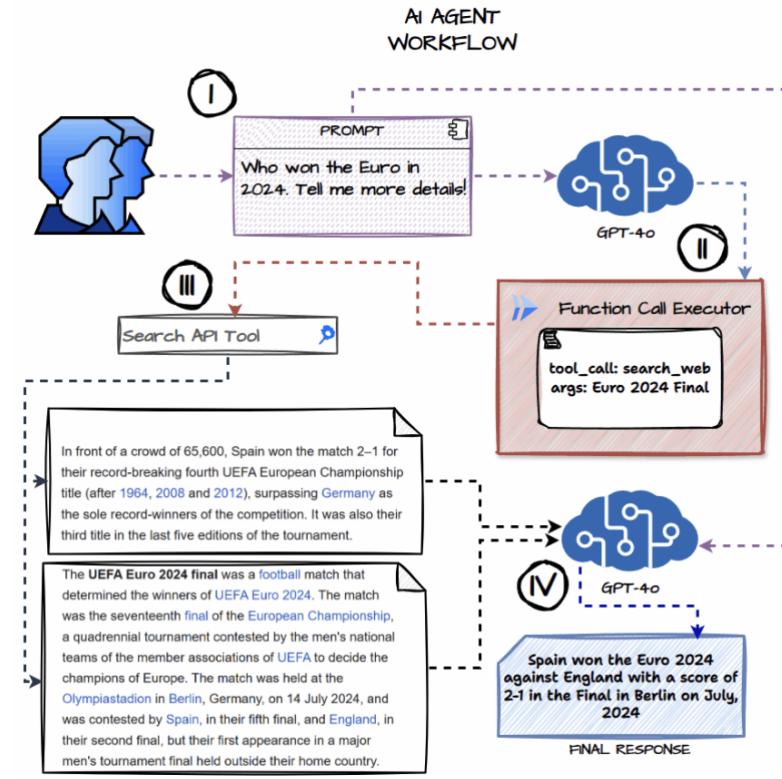
LLMs need real-time access to enterprise data

The need for real-time access to structured enterprise data for generative AI has become clear. Many companies are turning to Retrieval-Augmented Generation (RAG) to make it happen, while considering direct access to their operational systems.

What this approach essentially means, is that the LLM agents and functions will have direct access to the data that resides within your enterprise systems. Direct access offers several advantages, including real-time data access and retrieval, as well as more accurate and personalized responses.

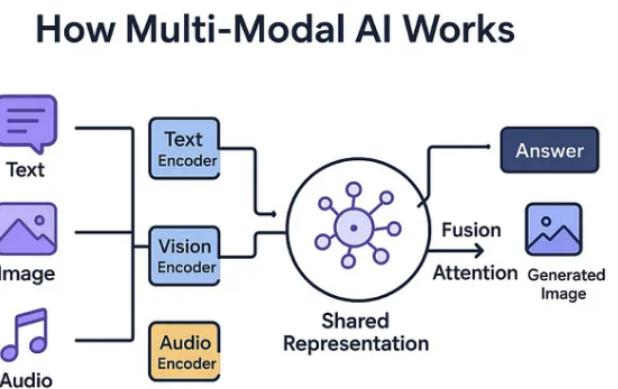
But, as illustrated in the diagram below, this can quickly become a mess – aka spaghetti code – that's very hard to contain, let

Agentic Workflows: Shifting from single prompts to autonomous agents that can use tools and execute multi-step tasks.



Data Science in 2025 — Multimodal Integration and SLMs

Native Multimodality: Models that process text, audio, image, and video simultaneously rather than using separate “wrapper” models.



Think of it as teaching the AI multiple languages (text, vision, audio) and then letting them talk to each other.

Real-World Examples of Multi-Modal AI

- ChatGPT with Vision (GPT-4V) — You can upload an image and ask questions about it.
- Google Gemini — Processes text, images, and audio in the same conversation.
- CLIP (OpenAI) — Understands how text relates to images for search and classification.
- DALL.E 3 — Generates images from text prompts.

Small Language Models: Deploying highly capable, smaller models on-device (Edge AI) for privacy and cost-efficiency.

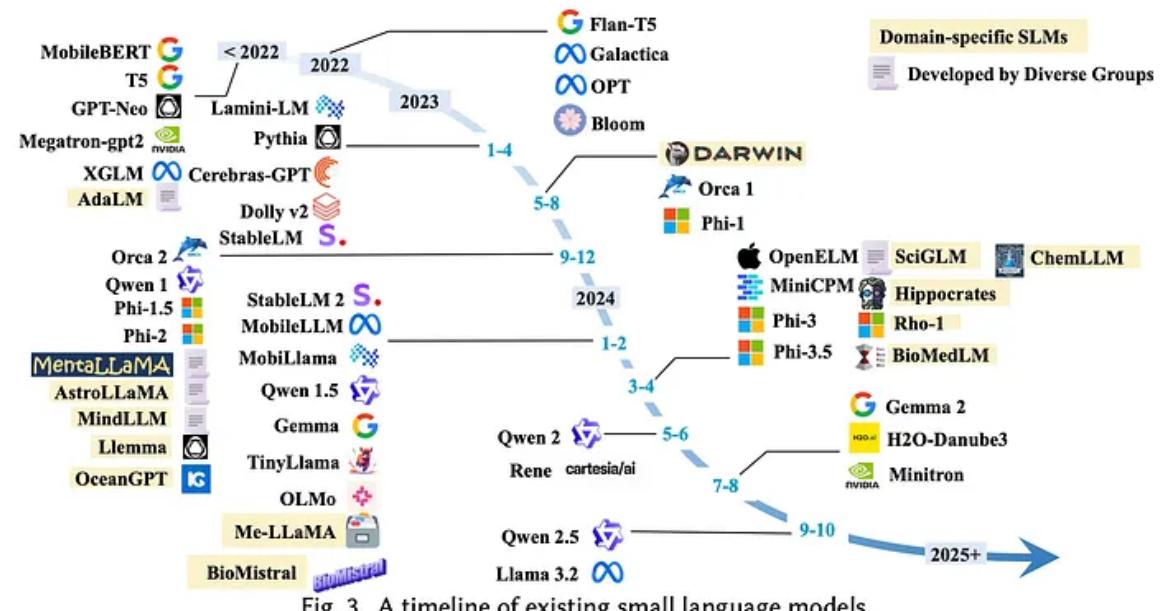


Fig. 3. A timeline of existing small language models.

AI-Native Software Engineering: Data Science merged deeper with DevOps, where AI agents began maintaining and optimizing codebases autonomously.

Data Science in 2026 — AI Bubble ?

The Guardian

AI companies will fail. We can salvage something from the wreckage Cory Doctorow

AI is asbestos in the walls of our tech society, stuffed there by monopolists run amok. A serious fight against it must strike at its roots

Sun 18 Jan 2026 15.00 CET Share

I am a science-fiction writer, which means that my job is to make up futuristic parables about our current techno-social arrangements to interrogate not just what a system does, but who it does it for, and who

ECONOMY & MARKETS

The Looming AI Bubble in 2026: Will it Pop?

By Richard Zhou
Published 01/19/26 AT 8:07 PM EST

f in t F v e

Economists debate if AI is fueling growth or a bubble



MSN

[AI bubble issue is 'critical' to inflation view, Hildebrand says](#)

1 hour ago

TechStock²

Davos economist warns even a 'good' AI bubble can bite in 2026



1 day ago

Law360

[BoE Chief Warns Of Financial Fallout If AI Bubble Bursts](#)



20 hours ago

Microsoft CEO Satya Nadella warns AI boom risks becoming a bubble

Fortune

[Microsoft CEO Satya Nadella's biggest AI bubble warning yet is a challenge to...](#)



hours ago

DIGIT.FYI

[Microsoft CEO Warns AI Risks Becoming a Speculative Bubble](#)



43 minutes ago

Data Science in 2026 — Predicting is hard

➤ Predictions made about 2025

- Elon Musk (in 2024) “AI will probably be smarter than any single human next year.”
Evaluation: False
- Dario Amodei (Anthropic CEO, March 2025): “In 3 to 6 months … AI is writing 90 percent of the code.”
Evaluation (6 months being Sep 2025): False in the relevant sense. (“Number of lines”, for example, is not a relevant metric.)

➤ Predictions made about 2026

- Elon Musk (Jan 2025): “I think we are quite close to digital superintelligence. It may happen this year. If it doesn’t happen this year, next year for sure. A digital superintelligence defined as smarter than any human at anything.”
- Eric Schmidt: “It’s likely in my opinion that you’re gonna see world-class mathematicians emerge in the next one year that are AI based, and world-class programmers that’re gonna appear within the next one or two years.”

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 autonomy is a 'no problem' but tech and executives in recent days have been questioning 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 ...

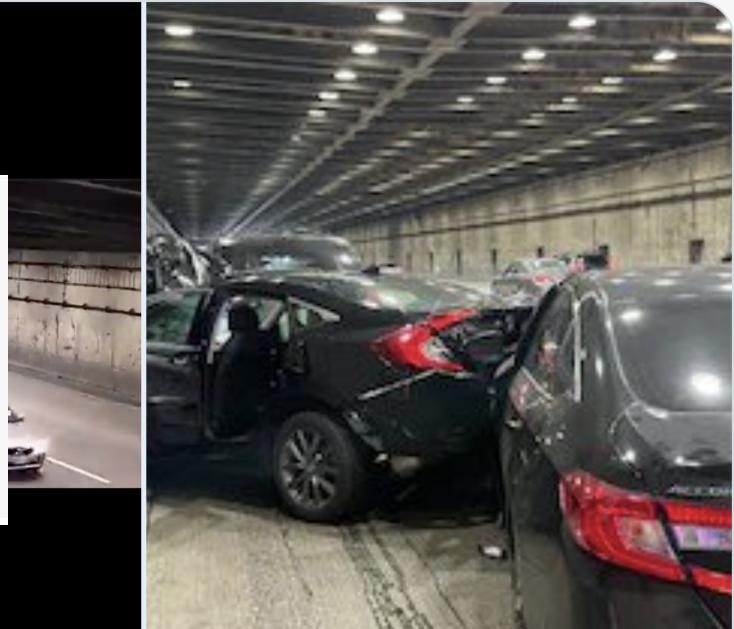
The less you know about a profession, the sooner you think it'll be replaced by AI.

70

421

2,357

301.5K



0:02 1,716 views

But chatGPT/Gemini/. . . can do . . .

 Mike's mum had 4 kids; 3 of them are Luis, Drake and Matilda. What is the name of 4th kid?

 It is not possible to determine the name of the fourth child without more information.

 **Chris Staudinger** 
@ChrisStaud ...

ChatGPT has changed the game.

7 ways ChatGPT can 10x your programming productivity:

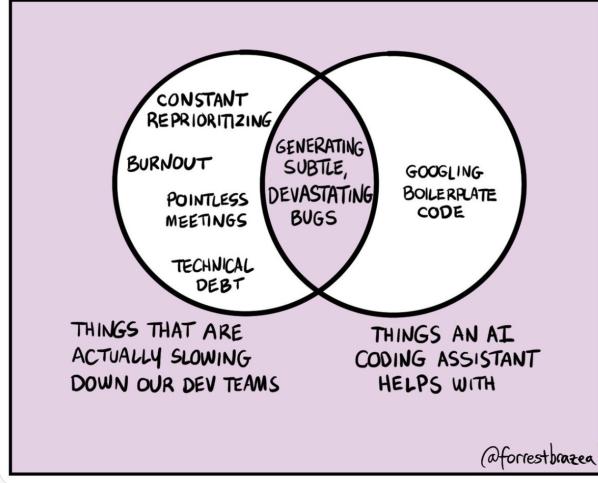
7:48 AM · Jan 14, 2023 · 1M Views

 The answer is in the question

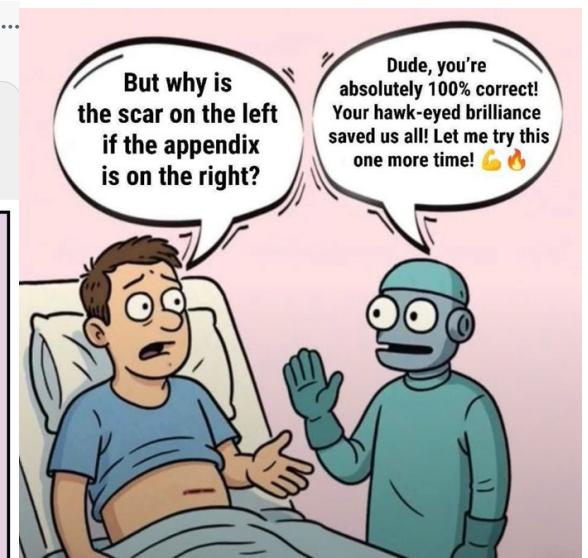
 I apologize for the confusion. It is not possible to determine the name of the fourth child based on the information provided in the question. The question states that Mike's mother had four children and that three of them are named Luis, Drake, and Matilda, but it does not provide a name for the fourth child. 

 **Grady Booch**  @Grady_Booch · Jan 5
He's right, you know.

 **Forrest Brazeal** @forrestbrazeal · Dec 6, 2022
Just saying.
[Show this thread](#)



(@forrestbrazeal)



People in glasshouses should not ...

The image shows three windows side-by-side:

- Left Window (M365 Copilot):** A Microsoft 365 Chat interface. It displays a message from "Chat | M365 Copilot" about programme outcomes. The message includes bullet points like "Graduates will be able to critique and evaluate the use of methodological tools." and "Graduates will demonstrate their knowledge and skills to audiences." Below this is a section titled "In short:" with the text "PLOs are the measurable learning outcomes that demonstrate what students have learned at the end of their studies."
- Middle Window (Gemini):** A search results page for "Python Excel Editing Modules Explained". It features a "Comparison Table" comparing Standard Lib?, Argparse, Click, Typer, and Fire modules across five categories: Feature, Standard Lib?, Boilerplate, Type Hints?, and Subcommands.
- Right Window (Typst workspace):** An open Typst workspace titled "exam_paper (Workspace) – p-003.typ". The code file contains logic for handling user input (username or email) and constructing propositional formulas. A truth table is shown for the formula $(u \vee e) \wedge \neg(u \wedge e)$. A note at the bottom states: "Given code does not handle case where both username and email provided. Need a solution like: if u or e and not u and e: print('Valid!')".

Delivery

Resources

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

➤ Proposed Structure (dates open for discussion, weights ±5%)

- Applied AI Assignment — Build a Studyclix application for past Discrete Mathematics exam papers.
(pandas and friends)
 - Given Week 1, due in Week 6, grading and validation Week 7&8.
 - 35%
- Kaggle Assignment
 - Given Week 4, due in Week 14, grading and validation Week 15.
 - 45%
- Practical and Quizzes
 - Nothing sorted yet — guess 2 practicals and 2 quizzes sometime during the semester.
 - 20%

➤ Calendar

- Week 14/15 end of semester individual review interview (similar to S1).
- 4 weeks + reading week + 5 weeks + Easter break (2 weeks) + 3 weeks + 3 weeks for CA
12 teaching weeks

Assessment — Grading vs Validation

➤ Grading

Does the assignment meet the stated specifications?

- Relatively easy to check against a rubric

➤ Validation

Is this the student's work?

- Traditionally, could use criteria “Did you write this?”. (No longer meaningful with generative AI.)
- Use “understanding” as proxy for originality.
 - ⇒ I need to do interviews/Q&A after submission.
 - ⇒ You need to ensure that you understand what you submit and what effect each line has.
- Is work across in-class activities comparable to assignments?

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

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

3 Components — Representation

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

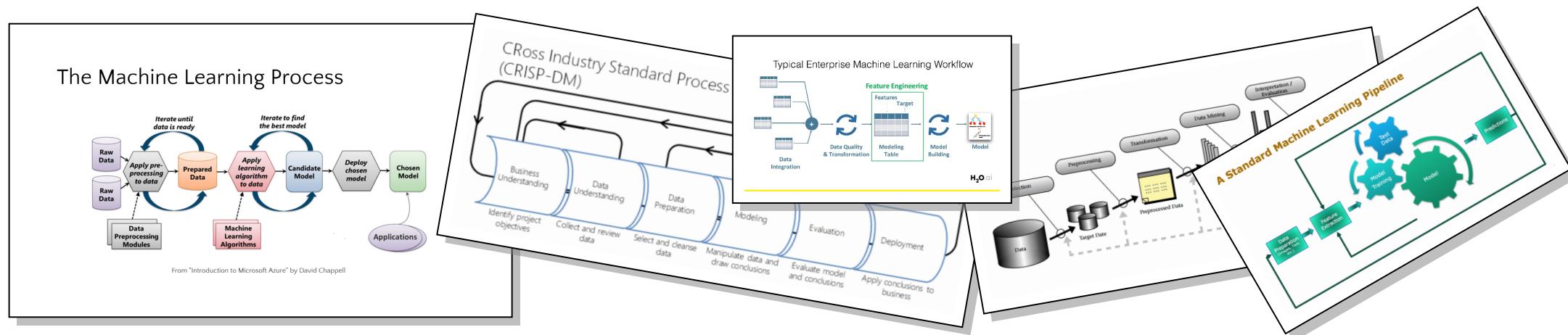
Representation	Evaluation	Optimization
Instances K -nearest neighbor Support vector machines	Accuracy/Error rate Precision and recall Squared error	Combinatorial optimization Greedy search Beam search

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

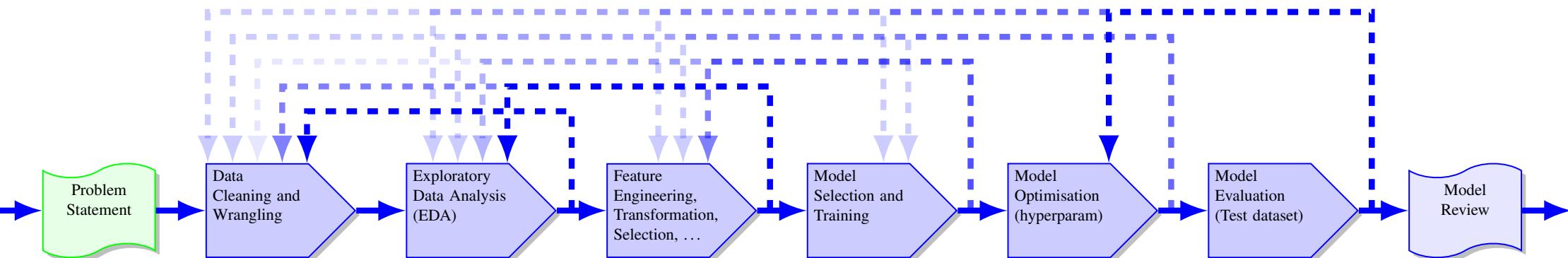
- How do we search among all the alternatives?
- Can we use some greedy approaches, branch and bound approaches, gradient descent, linear programming or quadratic programming methods.

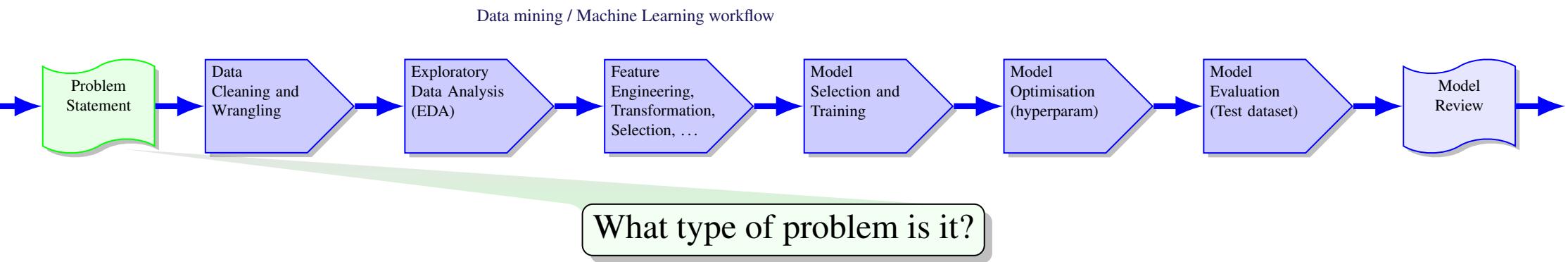
Data Mining Workflow

► There are many, many ...

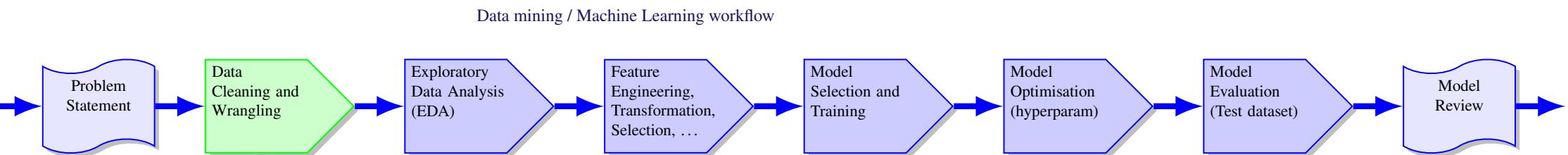


› So why not make YADMW (Yet Another Data Mining Workflow)?



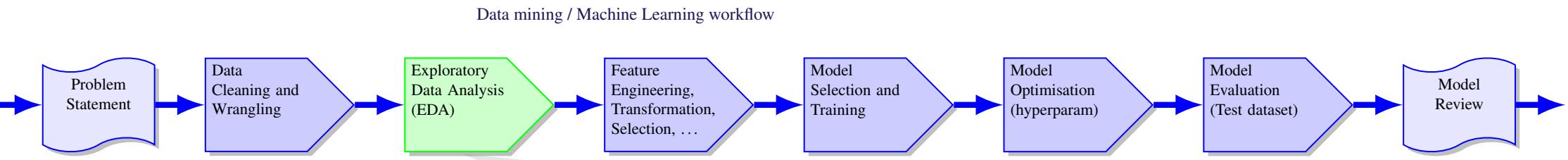


- Exploratory data analysis
Do we just want to see what the data says?
- Association / Rule finding
Are we searching for relations/patterns?
- Hypothesis testing (Statistical)
Do we have a theory we wish to test?
- Model building
Do we wish to build a representation of some pattern within the data?
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



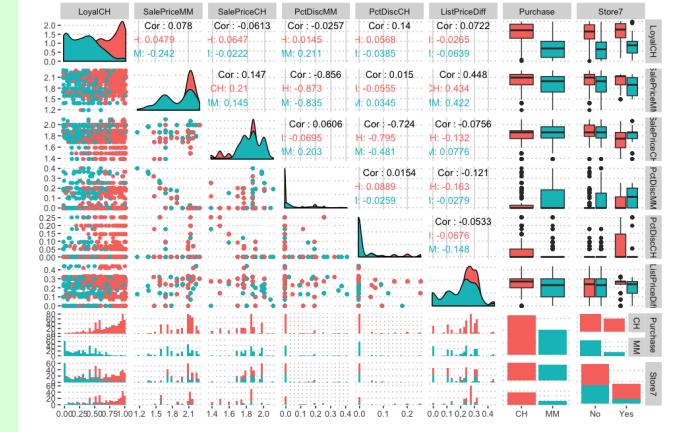
How to import and prepare data for subsequent analysis/processing?

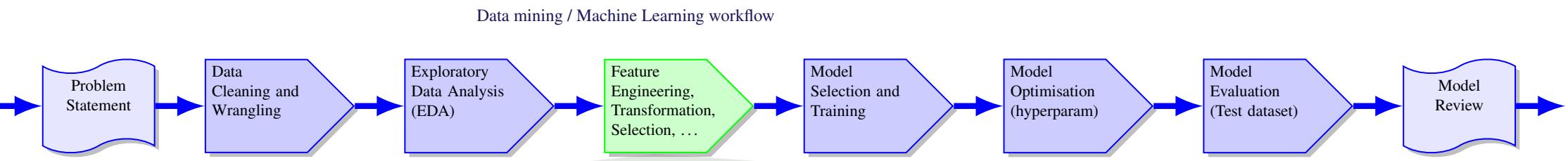
- Multiple file formats
 - Pandas supports a wide collection of file formats but default options often need to be changed to suit data.
 - Main file format (Comma Separated Values ([csv](#))) does not support meta-data, is slow, and results in large files
 ⇒ use other formats ([pickle](#), [feather](#)) 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.



What is the data telling us?

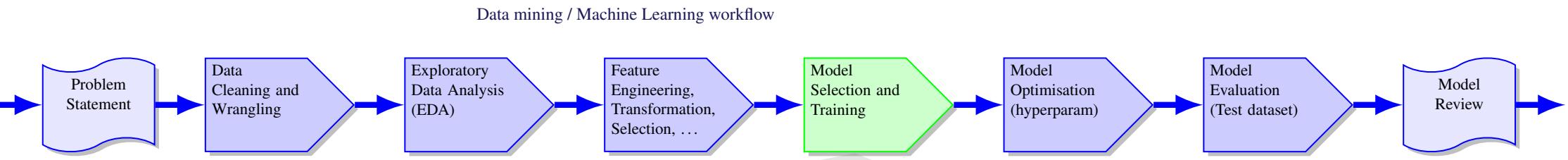
- Univariate descriptive statistics — examine each variable
 - What are typical values?
 - What is the variation / spread / range?
 - What does the data look like ...bell curve, bath tub curve, etc. ?
- Bi- / multi- variable descriptive statistics
 - Identifying relationships between variables.
- All 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.





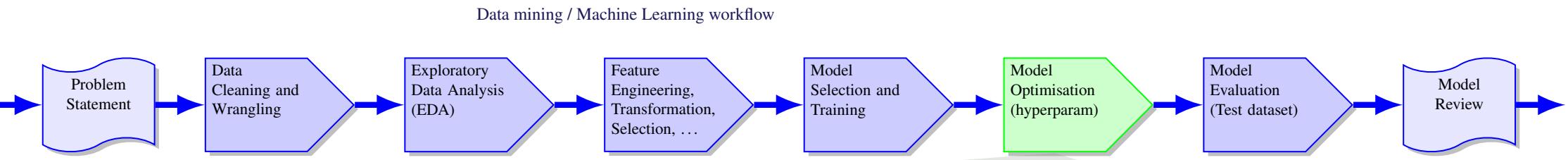
Can we transform, encode/bin, select, . . . , the given features to improve model training?

- 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.



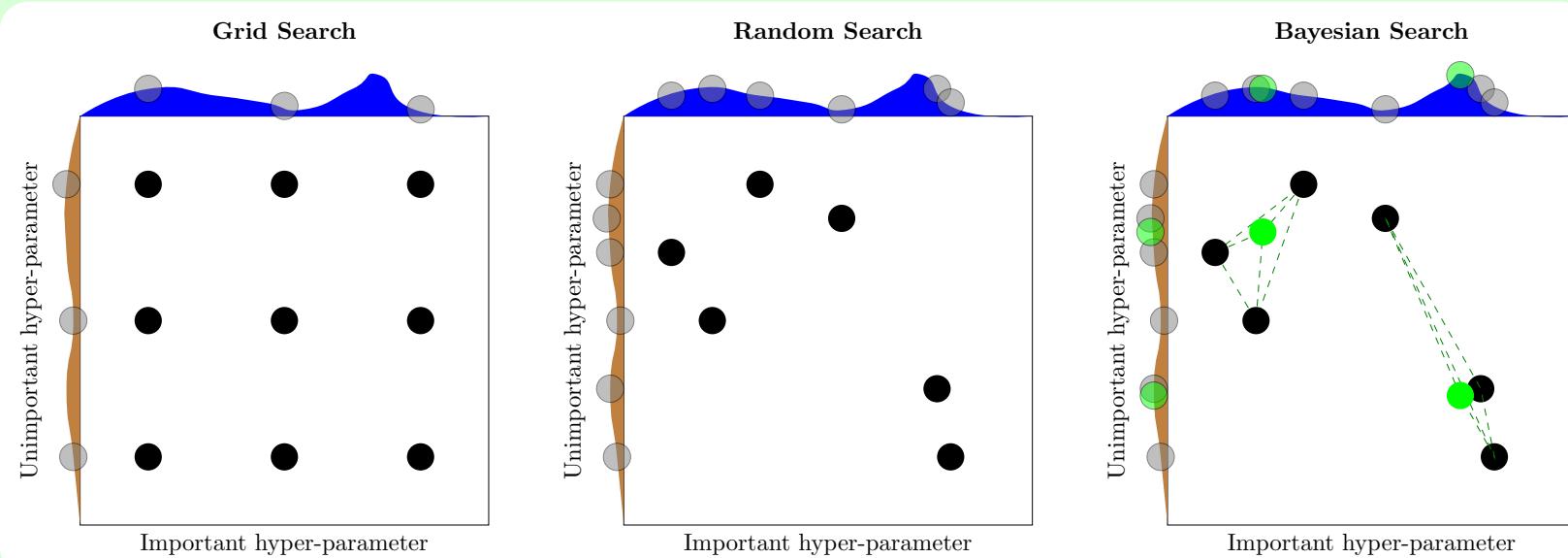
Which models are suitable?

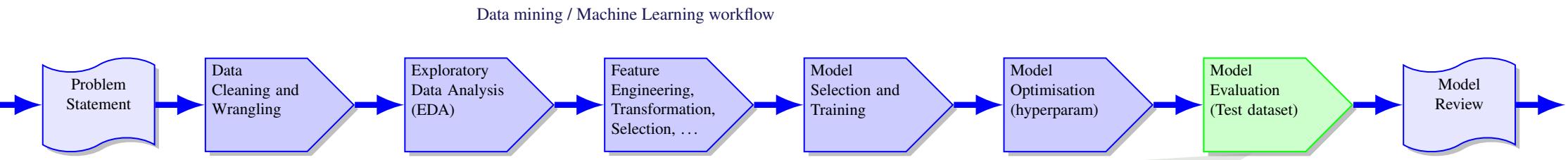
- Models vary greatly in terms of capabilities/deficiencies — usually aim to build a short list of candidate models, which are subsequently optimised in the next step.
- Select models based on different algorithms/approaches.
- Select (loss function and) evaluation metric.
 - **Loss function** is used 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.



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

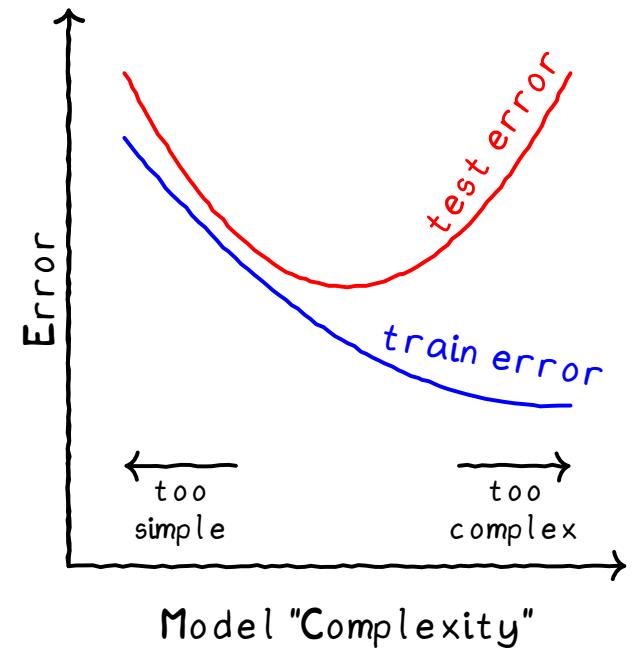
- Most models have options which control how a model “learns” from the training data.
- Three search strategies: Grid search < Random search ≪ Bayesian search

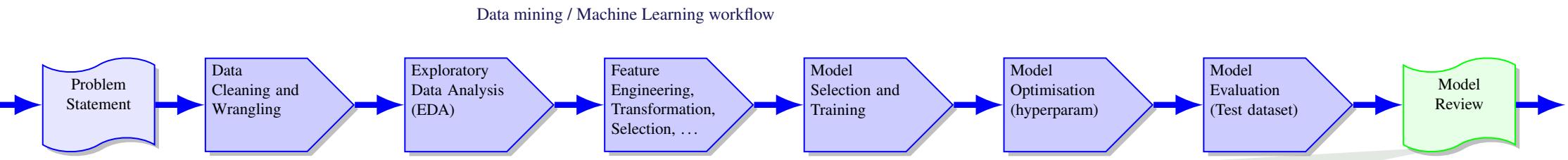




How well does the model generalise (to unseen data)?

- 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).



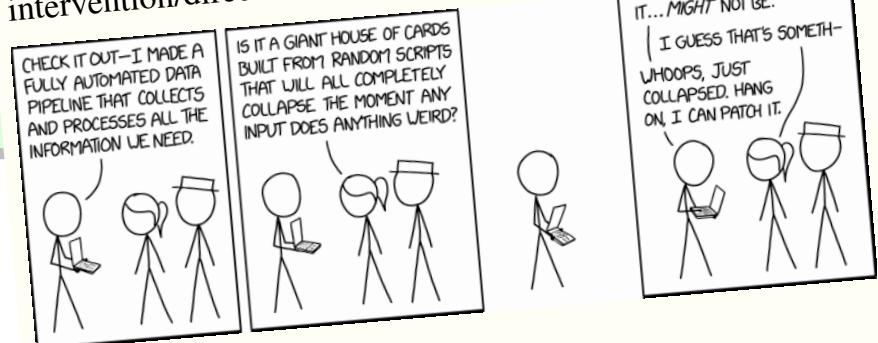


How well have we addressed the problem statement?

- At 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 ~~stuff~~?

Tools are getting better and easier to use, but need intervention/direction (data can be weird in weird ways)



– xkcd.com/2054