AIML430/COMP309: ML Tools and Techniques
Lecture 6: Data Mining and Machine Learning

Ali Knott
School of Engineering and Computer Science, VUW

Preamble: Machine learning (ML)

Machine learning systems are the kind of systems we have reviewed so far on this course.

- Supervised learning algorithms
  - Classification algorithms
  - Regression algorithms
- Unsupervised learning algorithms
  - Clustering algorithms
  - Dimensionality reduction algorithms.

What's data mining? How is it different from ML?

Data mining is the use of data analysis techniques for business applications, and other practical applications.

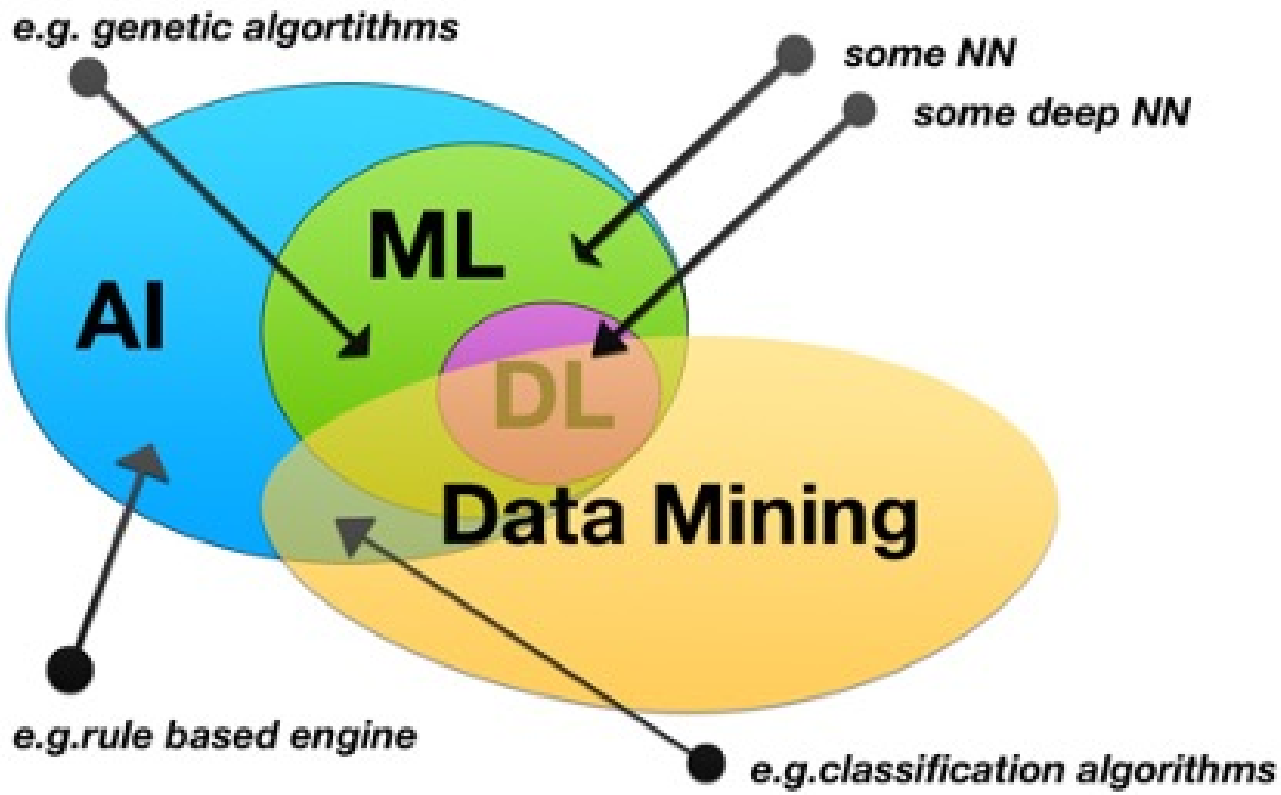
- Data analysis techniques *include* machine learning (ML) techniques—but include other methods too.
- 'Business applications' and 'practical applications': I'll give some examples.
  - In practice, the domain of application is pretty broad.

Not all systems that use machine learning are data mining systems!

- ChatGPT incorporates ML systems, but it's not a data mining tool.
- An AI system for diagnosing skin cancers from images incorporates a ML system, but it's not a data mining tool.

AI, ML, Deep Learning, Data Mining

The main point to take in: data mining overlaps with all parts of AI/ML, but also includes things that are *not* AI/ML.



So what's data mining?

Society produces huge amounts of data.

- This comes from all kinds of sources: e-commerce, retail, social media platforms, transportation and logistics, healthcare, economics, geography, environment, sports (. . .)

Data is *potentially valuable* to us—but *raw data is useless*.

- We need techniques to automatically *access* and *collate* our data, *organise* it, and *extract information* from it.

The 'extraction of information' part involves a lot of ML. (Supervised and unsupervised.)

- ML extracts information from data. (Raw data is not 'information'.)

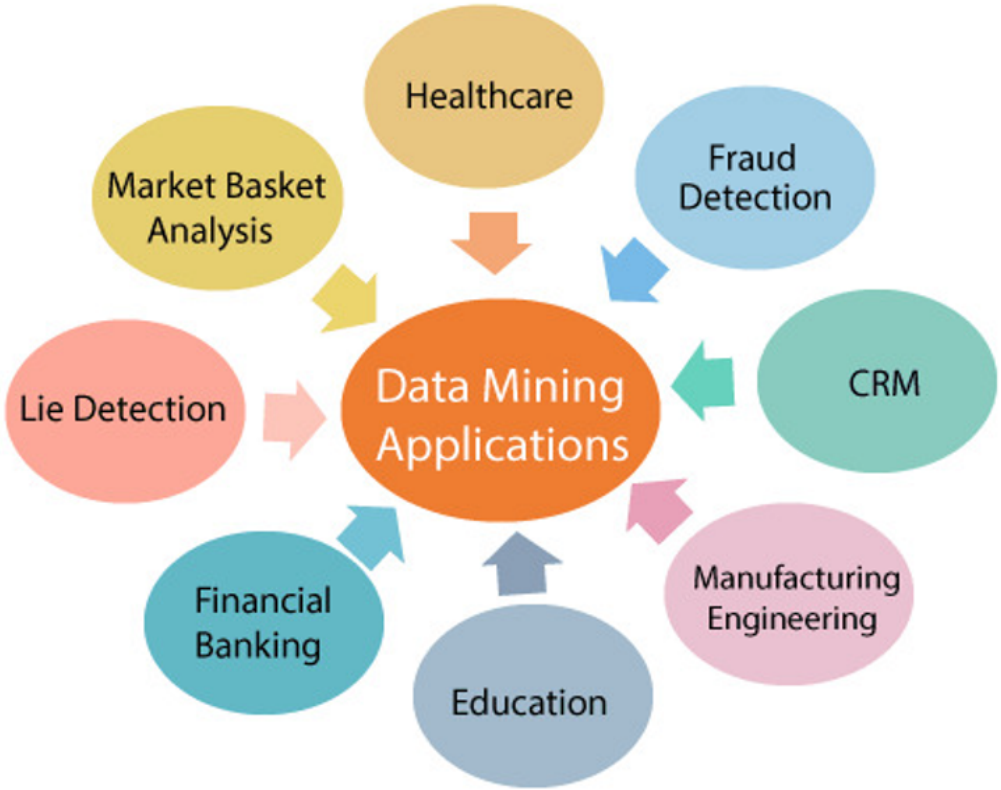
The earlier parts (accessing, collating and organising of data) is data mining, but involves less ML.

A bit of a definition

For Fayad et al., 1996, 'knowledge discovery in databases' is 'the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases'.

- 'Process': data mining comprises many iterative steps
- 'Nontrivial': some experimentation, or search, or inference, is involved
- 'Valid': the discovered patterns should hold true on new data
- 'Novel': the patterns should be ones that weren't previously known
- 'Potentially useful': the discoveries should plausibly lead to some benefit for the user
- 'Understandable': the discoveries should be communicable to stakeholders.

Some typical application areas for data mining



To notice here:

- These are *practical* application areas, in the public and private sectors.
- We're not focussing on basic science—though there are applications there too.

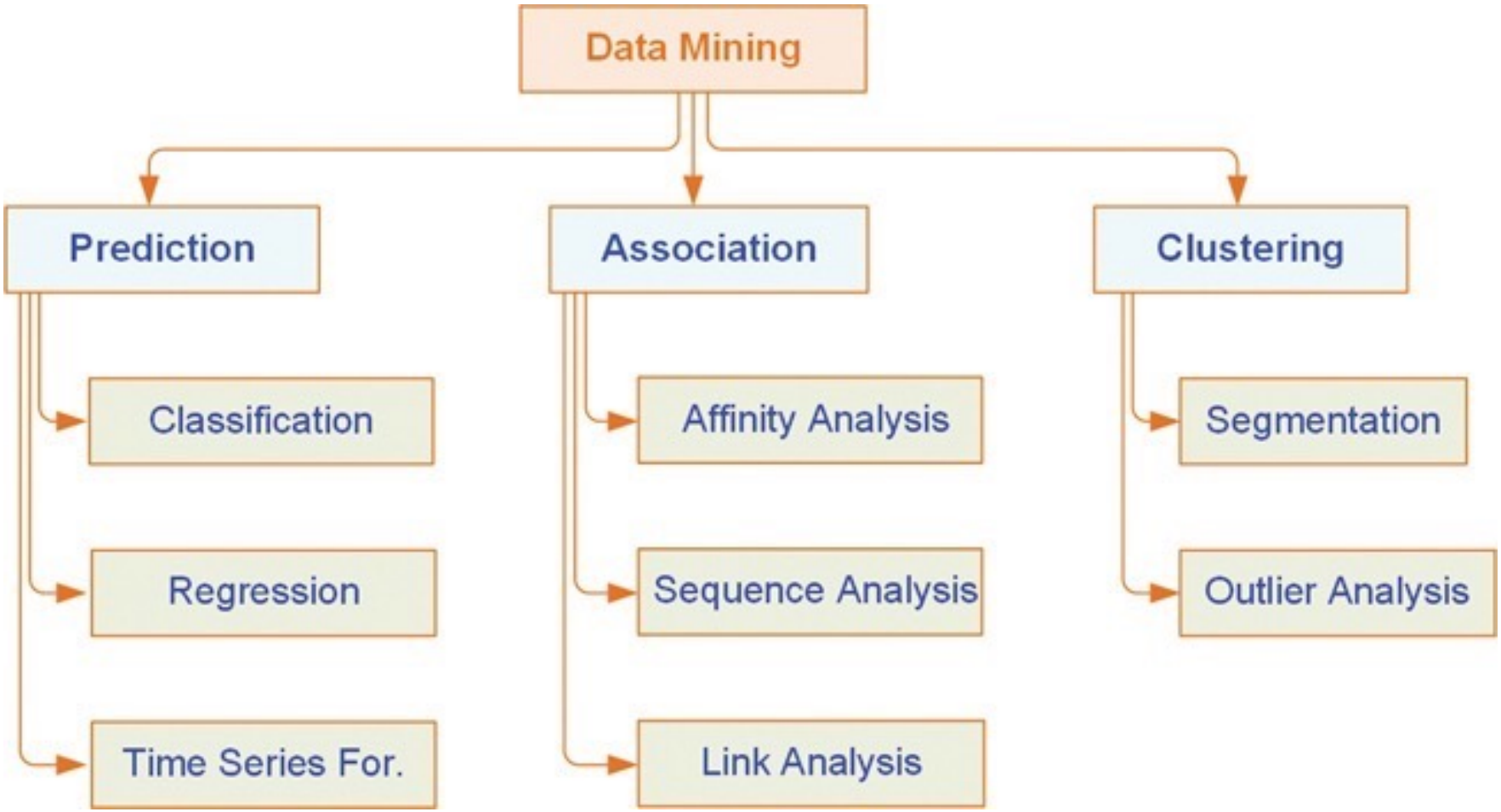
Data mining's ML component: 'pattern disovery'

ML models of various kinds can learn useful patterns in the identified data.

- Those patterns may be *explanatory*, and *teach* stakeholders something about the data...
- They may be *predictive*, and let stakeholders *make use* of the data in new ways...
- Ideally, they will be both! But prediction and explanation don't always go hand-in-hand.

Data mining's ML component: 'pattern disovery'

Data mining's taxonomy of 'useful patterns' cuts a little differently through the space of ML models, compared to our AI taxonomy.



‘Prediction’ and ‘Clustering’ tasks

‘Prediction’ processes mostly deploy supervised ML models.

- Time series is an instance of supervised ML that we haven’t looked at yet.

‘Clustering’ processes use (one type of) unsupervised ML model.

- ‘Outlier analysis’ would be well handled by algorithms like Gaussian Mixture models, or DBSCAN.
- ‘Segmentation’ is just another word for clustering, I think. (‘Customer segmentation’ gets used a lot.)

‘Association’ is a newer category.

I’ll discuss the three types of association analysis separately.

- Affinity analysis
- Link analysis
- Sequence analysis

Affinity analysis (market basket analysis)

The raw data here are transactions.

- Each transaction is a group of items that were bought together, on some particular occasion.

Given a large database of N transactions, we can define:

- An itemset is any tuple of items, $i_1 \ldots i_j$.
- The support for an itemset I is the count of I in the transaction database, divided by N. (The probability of buying all items in the itemset.)
- Confidence (written $i_1 \Rightarrow i_2$) is the conditional probability of i_2 given i_1: $\frac{p(i_1,i_2)}{p(i_1)}$.
- The lift of $i_1 \Rightarrow i_2$: how much more likely i_2 is to be bought when i_1 is also bought.

Affinity analysis (market basket analysis)

This is just straight probability theory—with alternative names.

- $i_1 \Rightarrow i_2$ is called an association rule.

The point of market basket analysis is to help stores select and arrange products.

- Which items should be next to each other on the shelves?
- Which items should shoppers pass first?
- If there are alternative versions of a product, which performs best?

Link analysis

A lot of data is (or can be) structured as a network, with nodes connected by arcs.

Some examples:

- The web: nodes are web pages, arcs are links.
- Social media platforms: nodes are users, arcs are ‘followers’ or ‘friends’.
- Physical cities, countries. . .

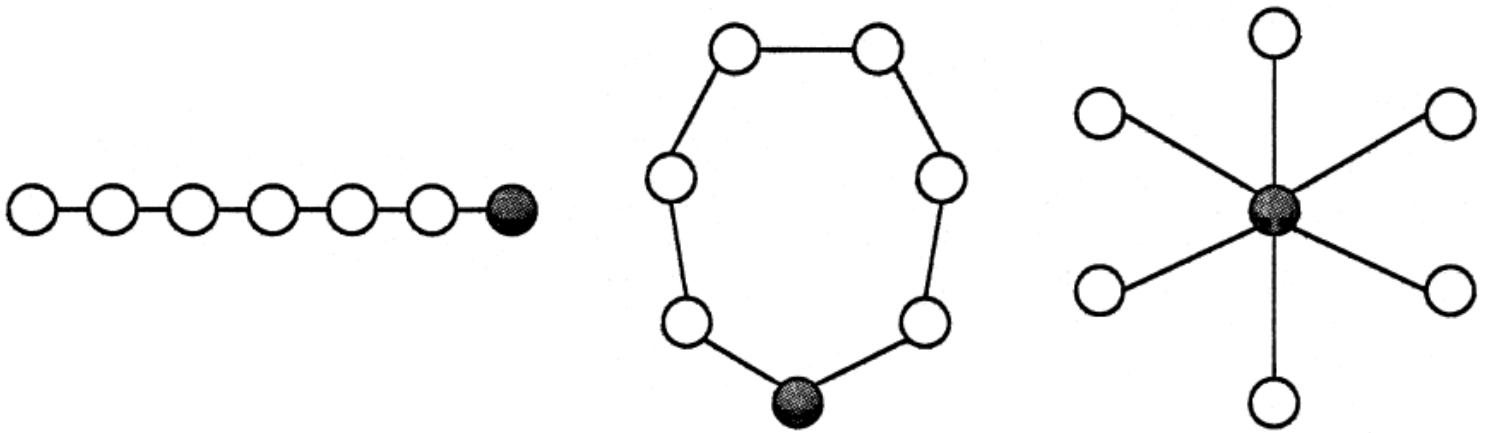
Link analysis is about finding information in networks.

- It involves applications of graph theory.

Some examples of link analysis

One task is to study how information flows in a graph.

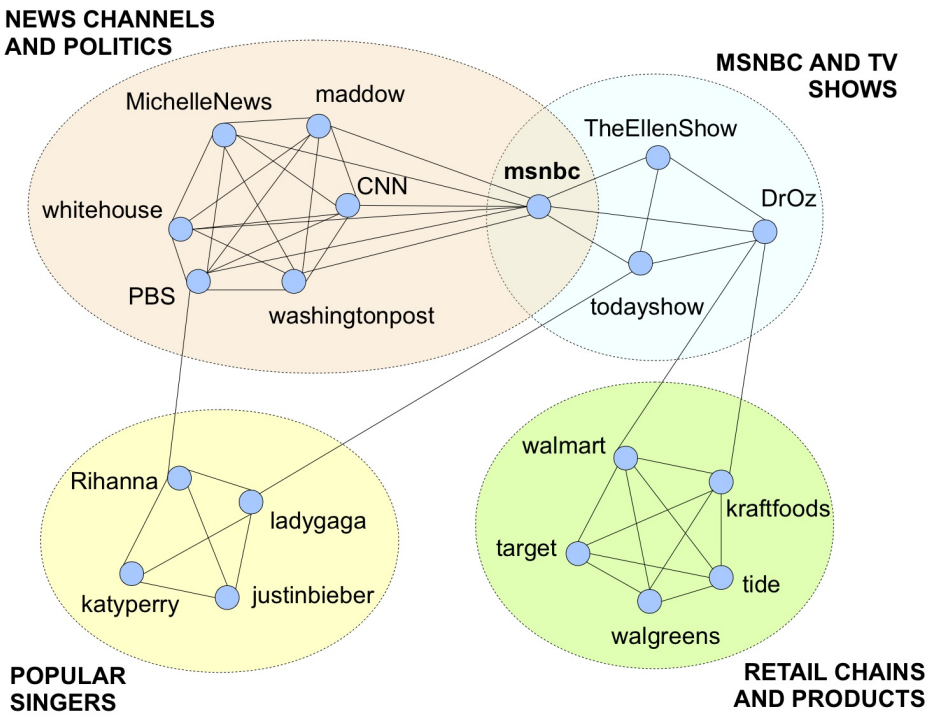
- This is a function of its topology—some nodes are more important than others! (Which are the influential people, or pages?)
- Nodes with high degree are important...
- Nodes which have high closeness to other nodes, or high betweenness, are important...
- Nodes that are cutpoints (would separate bits of graph if removed) are also important...
- Google's PageRank algorithm pays attention to graph topology.



Some examples of link analysis

Another important task is to identify cliques in graphs.

- A clique is a set of nodes that are all connected to one another.
- A community is a set of nodes that are particularly well connected.



Sequence analysis

Here, the data items to be analysed are sequences.

- E.g. clicks on a web page
- E.g. market baskets from a given customer on different dates
- E.g. base pairs in a DNA molecule

We're interested in things like:

- What are common sequential patterns?
  - Which link is most often clicked on first?
  - What is the longest commonly occurring sequence?
  - What are sequences that have 'high value'?
- What item most commonly occurs after item i_1? (Sequential associations)

My guess is that a lot of sequence mining these days is done with neural networks (transformers)...

Domain example: Data mining at Netflix

Netflix is available in over 190 countries, and had 277M subscribers at the start of this year.

- They have a recommender system (RecSys), that suggests to each customer what they might want to watch.
- The RecSys is responsible for 80% of stream time—that's big business!

The RecSys must learn *categories* of user, and of content.

- A common way to do this is with matrix factorisation.
  - Define a big matrix, with all users along the top, and all movies down the side: cells hold *user ratings* of movies.
  - The factorisation challenge: fill in the blanks.
- The Netflix prize offered $1M to anyone predicting ratings better than them, on a training set of 100M ratings. . . (awarded in 2009)

Domain example: Data mining at Facebook

Any company with a recommender system needs to learn user and content categories like this.

Facebook has a particularly sophisticated system...

- 29,000 categories of user! (Data from ProPublica)

Some examples:

- 'Total liquid investible assets $1-$24,999'
- 'Individuals that are frequent transactors at lower cost department or dollar stores'
- 'Receptive to companies offering online auto insurance'
- 'Away from family or hometown'

Domain example: Data mining in healthcare

Many databases, many applications—

- Evidence-based diagnosis: how many patients with symptoms S had disease D?
- Treatment outcome prediction: how many patients with circumstances C and treatment T had outcome O?
- Epidemiology: how fast is an outbreak of disease in region R likely to spread?
- Anomaly detection: look for unusual datapoints, that may need more attention. . .

Data mining tools

There are a few widely used commercial tools:

- IBM: SPSS Modeler
- SAS: Enterprise Miner
- Quest(StatSoft): STATISTICA Data Miner

Some popular open source and/or free tools:

- Weka: from NZ (cs.waikato.ac.nz/ml/weka/)
- RapidMiner (rapidminer.com)
- Orange (orangedatamining.com)
- KNIME Analytics Platform (knime.org)

Code-based tools:

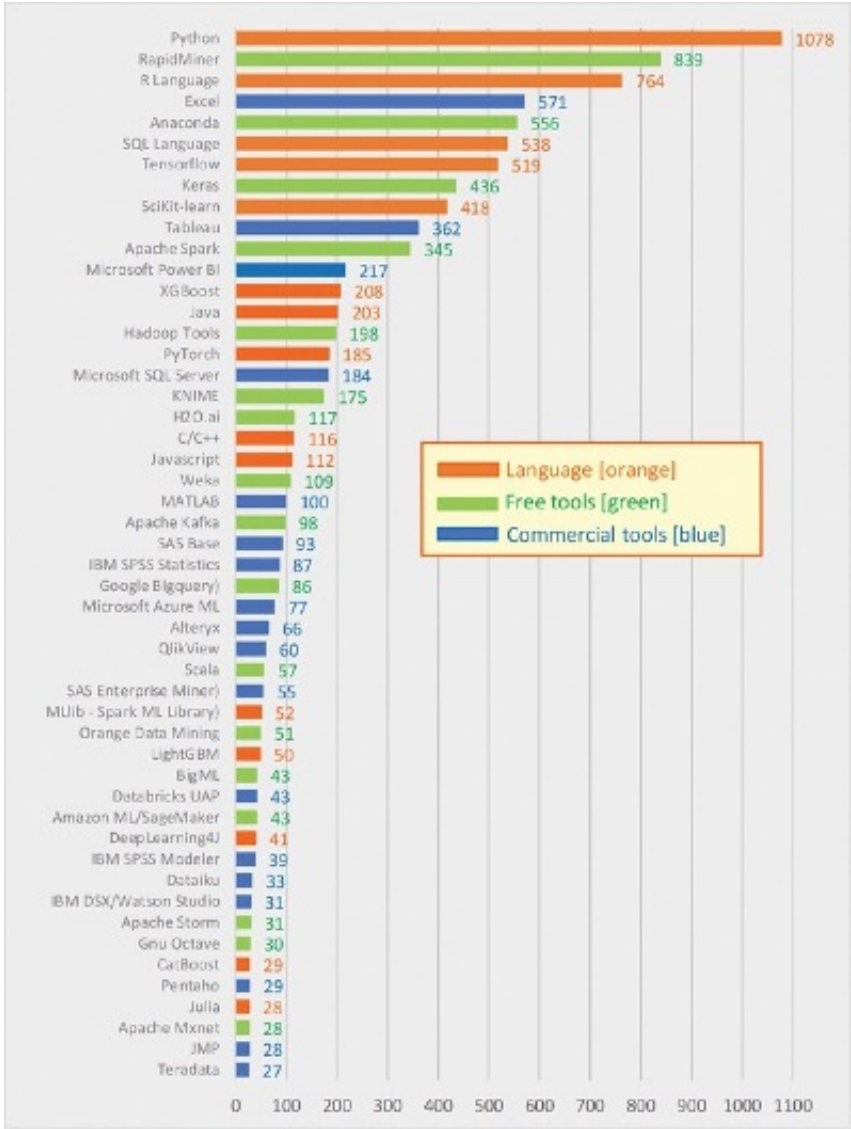
- Python, R and JavaScript

Data mining tools

This is a KDNuggets poll from 2019, asking people what tools they have used in the last three years, ‘for a real project’.

Interesting to see that open-source tools and programming languages are more popular than commercial tools.

Python & scikit-learn (and similar tools) are probably growing in popularity.



The dark side of data mining

Large companies—especially online ones—gather vast amounts of data about users/customers. And often this data can be shared between companies.

- Often, customers don't know how much data they are providing.
  - Privacy legislation (e.g. the EU's GDPR) can be helpful...
  - But there's a limit to what individuals can do.
- There need to be better rules, letting us know what kinds of data mining companies are doing, and how they're using the results.

Again, EU laws are in the vanguard. The EU's Digital Services and Digital Markets Acts are bringing much-needed reform.

- The DMA prohibits 'gatekeeper' companies tracking users off-platform to inform their ad targeting (unless consent is given).

Data mining for the public good?

Big tech companies do data mining to serve their own commercial interests.

- But if *independent researchers* could do data mining on companies' data—with appropriate ethics and oversight—that could be a very good thing.
- Policymakers, and the public, *need to know more* about what's going on in big tech platforms!
  - E.g. how do recommender systems work? What are they optimised for?
  - Are there harmful effects on users that could be avoided, if companies change the settings?
- The EU's DSA will allow researchers and auditors to access company data.
  - The rules for researchers are still being worked out...
  - I'm involved in the conversation! See e.g. here...

A few basic rules for data mining

Access:

- Make sure you are allowed to have the data you are working on!

Anonymity:

- De-identify your dataset before you start! (So you don't know who the individuals are.)

Next lecture. . .

We'll be looking at *processes* for data mining—there are some well-established best practices.

- In particular, CRISP-DM.