AIML430/COMP309: ML Tools and Techniques
Lecture 7: Data Mining 2—CRISP-DM

Ali Knott
School of Engineering and Computer Science, VUW

CRISP-DM

CRISP-DM is a widely used *process* for data mining.

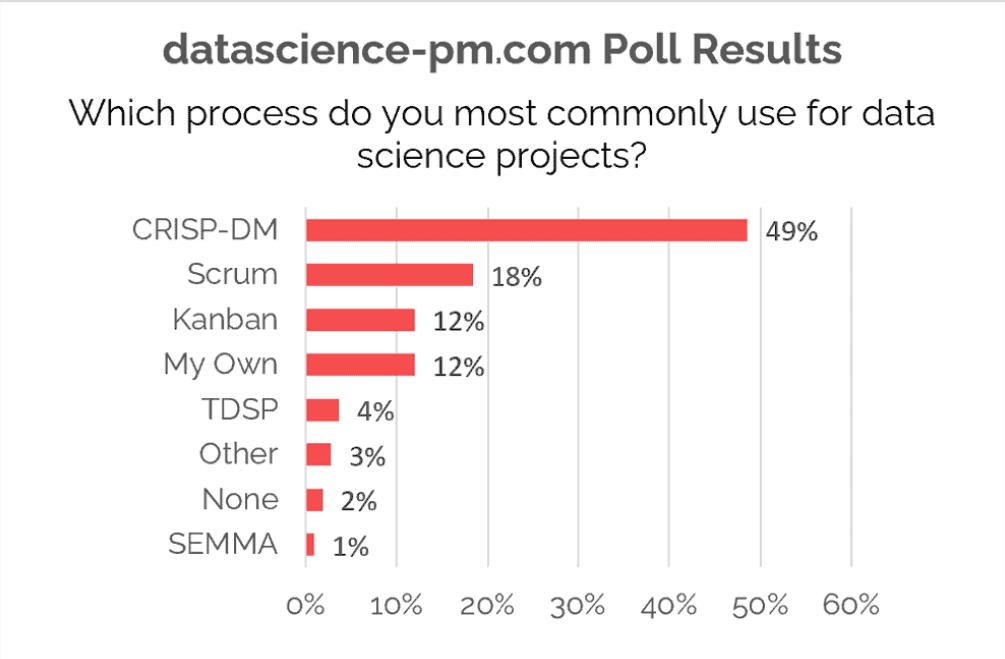
- It stands for 'Cross Industry Standard Process for Data Mining'.
- It was developed in 1996, by a consortium of companies, including Daimler, NCR andTeradata (as part of an EU research project).

CRISP-DM is an open standard, which is a topic in itself.

- All industries have a range of bodies that set technical standards.
- For computing, some bodies are convened by governments (e.g. the ISO, the IEC, the EU's CENELEC), and others just represent 'communities' (e.g. the IEEE, the W3C).
- Standards are written documents: some of these have to be paid for, but *open standards* are free!

CRISP-DM is still widely used. . .

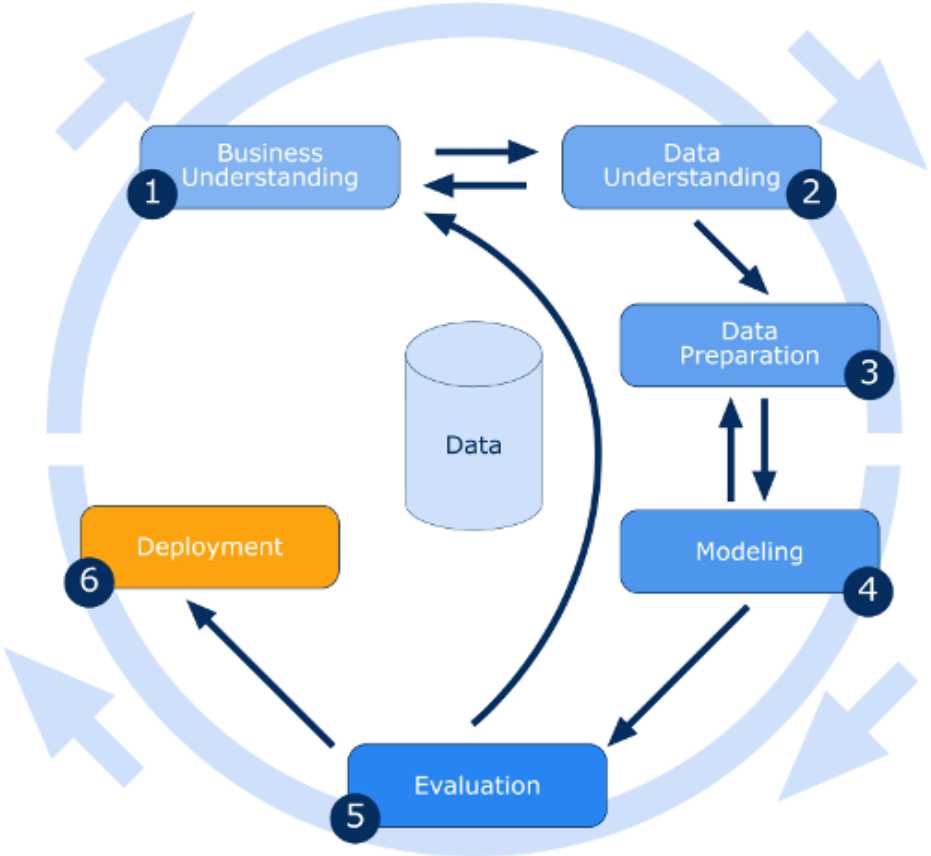
A poll earlier this year by the Data Science Process Alliance found CRISP-DM is still way out in front. . .



It's interesting that 'data mining' blends with 'data science', 'data analytics'.

So what is CRISP-DM??

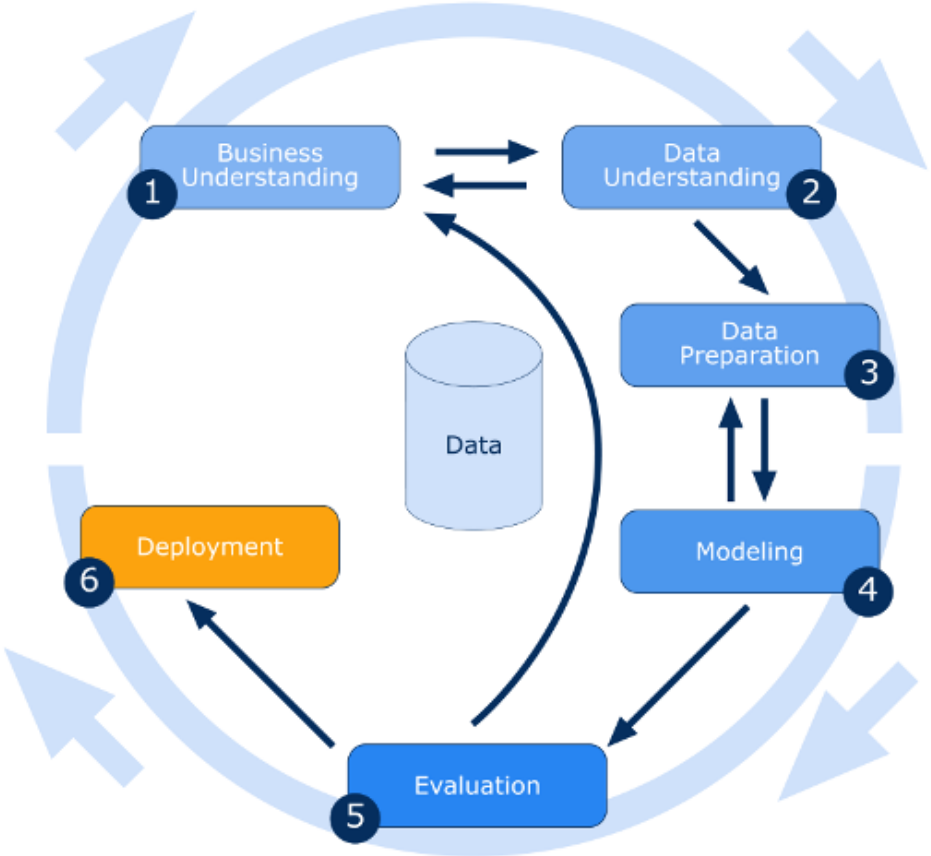
It's a 'widely used, non-proprietary, and industry agnostic methodology and procedures for best practices in data mining'.



- Non-proprietary means 'open standard'
- Industry-agnostic means 'works in many domains'
- A methodology is just a well thought-out, explicit *process*.

So what is CRISP-DM??

It's a 'widely used, non-proprietary, and industry agnostic methodology and procedures for best practices in data mining'.



There are 6 steps—

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Evaluation
- Deployment.

We'll look at those one-by-one!

1. Business understanding

You have to work out a few important things—

- *Who should you talk to* in the organisation?
  - Who's in charge? Who really knows things?
- What does the business want to achieve from data mining?
  - What do the directors want?
  - Is this sensible/possible? Should it be revised?
- What resources are available to you?
  - Where is the relevant data?
  - How much time/money do you have? Who can help?

When you have all that, you can make a project plan.

- This should be high-level, and non-technical, so the directors can understand it too.

An example

KiwiMart is a supermarket chain—they're having problems with their fruit sales, and want to improve, so they bring in a data analytics company.

After some meetings with management, they come up with a goal, and some subsidiary questions:

***Primary goal**: Inventory Optimization, to optimize the inventory management system, have the right amount of fresh fruits in stock to meet customer demand while minimizing wastage due to overstocking.*

***Related questions**:*

- *How customer preferences and purchasing patterns vary across different stores?*
- *Which fruits have the highest and lowest sales volumes in different time periods?*
- *How do promotions and discounts affect sales volumes and customer purchasing behaviour for different fruits?*

An example

The company meets with Kiwimart's database people, and devises two more specific goals, which they approve with management. . .

***Business Goal**: Inventory Optimization*

***Data Mining Goal**: Demand Forecasting, Market basket analysis to discover association rules among products, Customer Segmentation, Develop personalized recommendation models*

***Business Goal**: Sales Improvement*

***Data Mining Goal**: Customer Profiling, Outlier Detection*

They also define some success criteria—

- For Business Goal 1, forecasting accuracy of X%
- For Business Goal 2, sales improvements of X%

2. Data understanding

This phase involves three tasks:

- Collecting relevant data from a range of places
- Verifying the general quality of the data
- Exploring the data, to see what's there

Back to Kiwimart...

The data mining people find various data sources:

- Sales Data: Transaction records containing information on customer purchases
- Customer Data: Information from loyalty program registrations
- Product Data: Details about each fruit

They do some general reality checks:

- Sales: verifies the integrity of the sales records, checks for consistency in data formats, and assesses the completeness of required fields.
- Customer: examines the customer profiles for completeness and consistency
- Product: reviews the product information for accuracy and ensures that essential attributes are present for each fruit.

They run some preliminininary analyses, to understand what they have:

- Sales trend: generates visualizations, to analyse sales trends over time and identify seasonal variations in fruit purchases.
- Customer: uses bar charts and histograms to understand the distribution of customer age groups and locations.

3. Data preparation

There are five subtasks here:

- Data Selection: pick the databases (and fields) that will be useful
- Data Cleaning:
  - Search for duplicate entries! A very common occurrence!
  - Find and fix missing data. . . also very common!
- Data integration:
  - That is, *combine datasets*
- Data Construction:
  - In particular, create *new features*, by processing existing ones
- Data formatting:
  - In particular, *split* assembled dataset into training, validation, and testing sets.

Back to Kiwimart

Some examples of all those things. . .

- Data Selection: some fields of the product database aren't relevant...
- Data Cleaning: remove duplicate records and handle missing values in customer profiles and product information.
- Data Integration: combine sales data, customer data, and product data using common identifiers to create a single dataset for analysis.
- Data Transformation: use Feature Engineering to create some new features (variables) that may provide more meaningful insights. Normalise/scale numerical features to bring them into the same ranges.
- Data Spliting: Dividing the Data into Training and Testing Sets

Step 4: Modelling

Two initial steps:

- *Select the modelling technique* that's most suited to the original business goals.
- *Create a ML experiment design* that's appropriate for these goals.

The next two steps are the heartland of machine learning:

- *Build* your models:
  - Run the selected modelling tool on the prepared dataset to create one or more models.
  - Choose hyperparameter settings.
  - Identify the best models, and describe these.
- *Assess* your final models:
  - Evaluate the models on test sets, using appropriate metrics.
  - Evaluate the results, against your original business goals.

Once more with Kiwimart

The data mining team decide to build two models.

One is a demand forecasting model, to optimise the supermarket's fruit inventory:

- Time Series Forecasting: uses time series forecasting techniques, such as ARIMA (AutoRegressive Integrated Moving Average) or exponential smoothing methods, to predict future demand for specific fruits based on historical sales data

One is a customer segmentation model, that will be helpful for targeted marketing:

- Clustering Algorithms: uses clustering algorithms, like k-means or hierarchical clustering, to segment customers based on their purchasing behaviour and preferences.

Evaluating the Kiwimart models

They evaluate their time series forecasting model with a Mean Absolute Error (MAE) metric, for predictions over 2 months.

- The 2 month timeframe was decided in consultation with people who know about suppliers.
- They used MAE because costs are in direct proportion to errors.

They evaluate their customer segmentation model with a sillhouette score:

- This combines a measure of cluster compactness with a measure of cluster separation.
- Remember, there's no objective way of assessing performance of a clustering algorithm!

Step 5: evaluation

The ‘evaluation’ step of CRISP-DM is a *business* evaluation step.

- How useful is the model produced from data mining *for the business*?
- This is quite separate from quantitative evaluation on the test set!

Once again with Kiwimart

The data mining team evaluate their time series forecasting model by running a 'ghost' fruit ordering scheme, alongside Kiwimart's actual inventory scheme.

- After 6 months, they can show their 'ghost' scheme would have saved the company $1.5M.

They evaluate their customer segmentation model by running an A/B test, trialling Kiwimart's current marketing promotion alongside a targeted promotion, informed by the model.

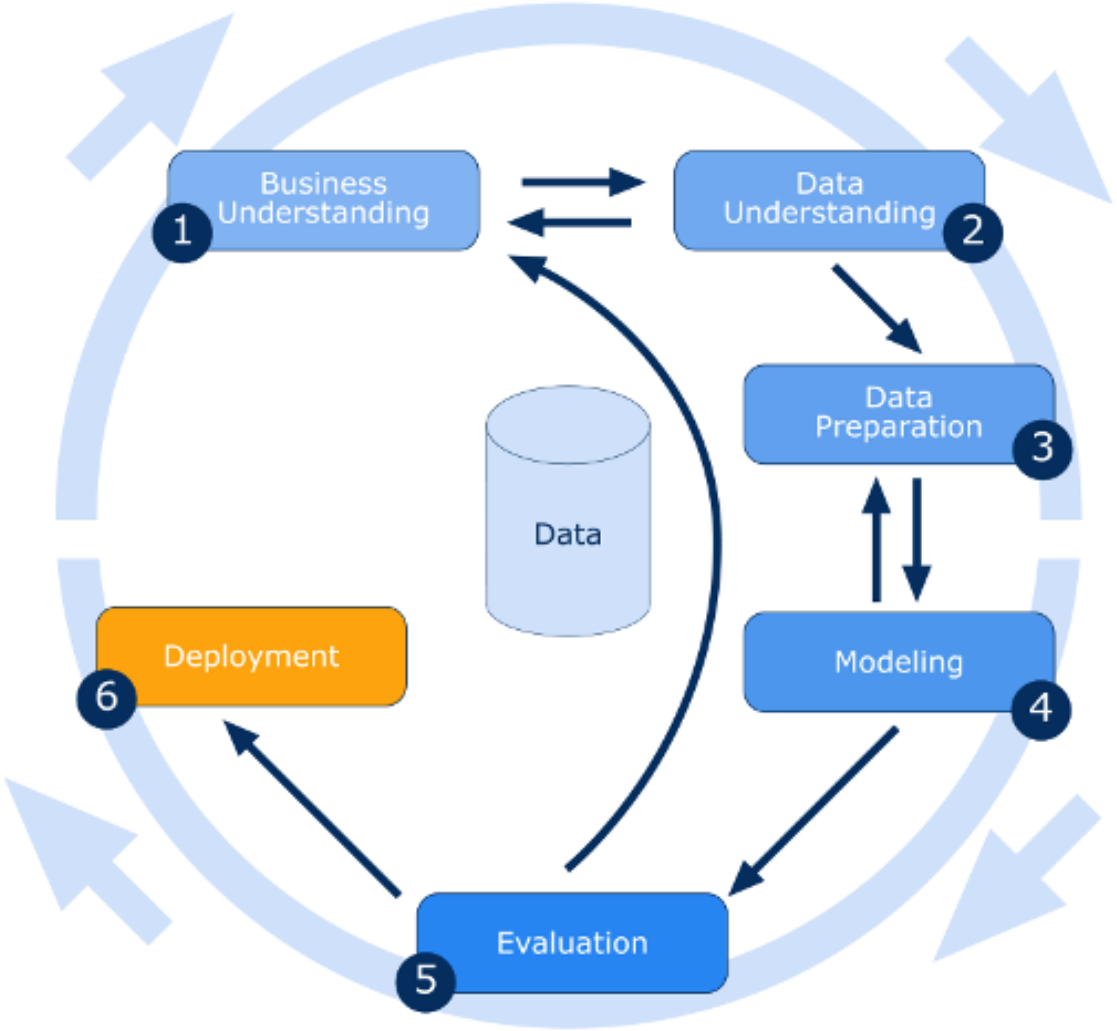
- Two customer groups are randomly selected. . . one promotion scheme is used for each group.
- They show the targeted scheme leads to a boost in sales. (Which is statistically significant.)

Step 6: deployment

If the models evaluate well, they can be deployed.

Iterations in CRISP-DM

CRISP-DM contains various options for iteration.



- Modelling might cause you to rework your data. . .
- Evaluation might bring new business understanding. . .
- Deployment updates how the business runs, and may prompt new iterations of the whole process.

Next lecture. . .

I'll cover

- Some other examples of data mining
- Some tools for data mining.