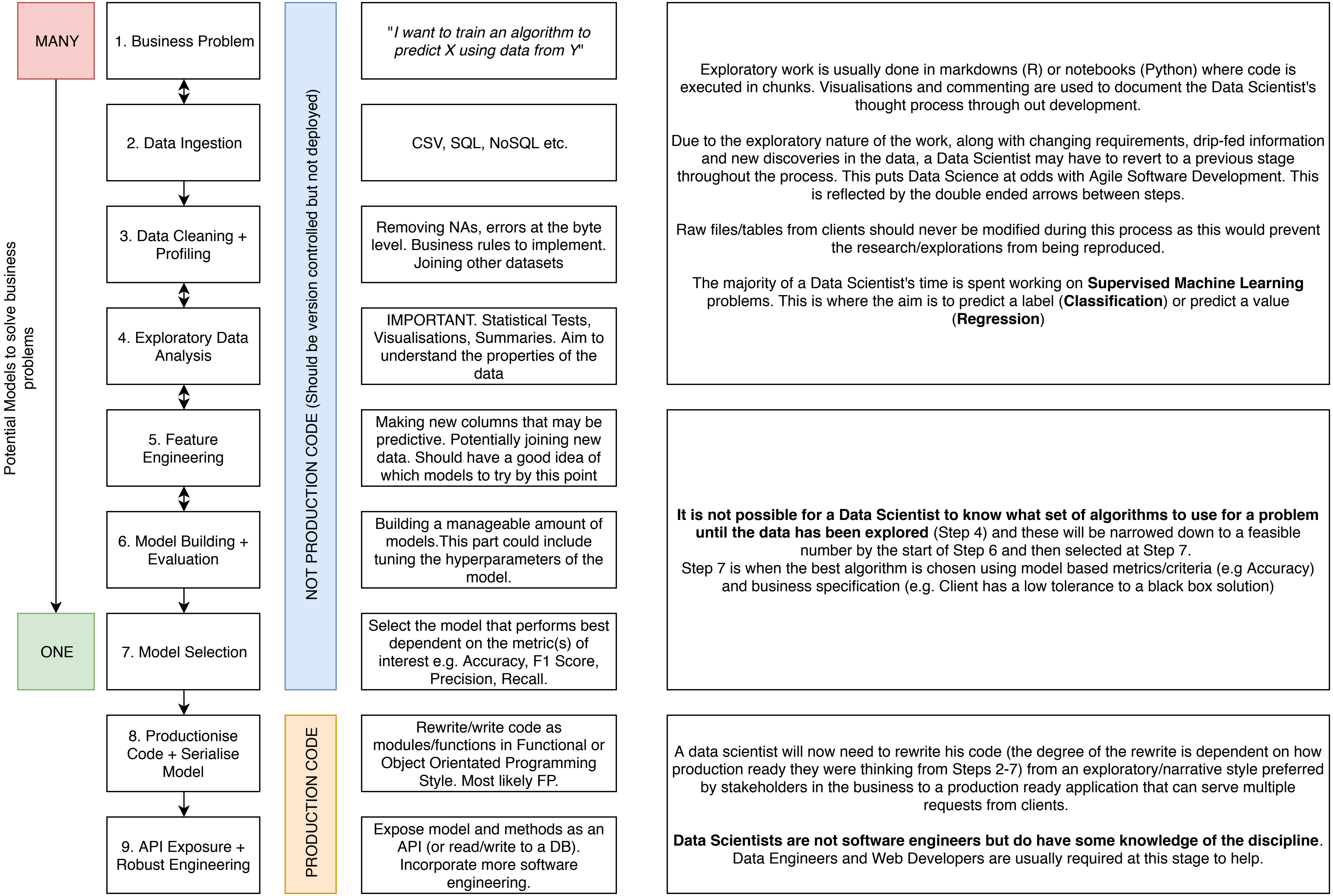
# Data Science Workflow



MANY

Potential Models to solve business problems

ONE

NOT PRODUCTION CODE (Should be version controlled but not deployed)

PRODUCTION CODE

| | |
|---|---|
| 1. Business Problem | "*I want to train an algorithm to predict X using data from Y*" |
| 2. Data Ingestion | CSV, SQL, NoSQL etc. |
| 3. Data Cleaning + Profiling | Removing NAs, errors at the byte level. Business rules to implement. Joining other datasets |
| 4. Exploratory Data Analysis | IMPORTANT. Statistical Tests, Visualisations, Summaries. Aim to understand the properties of the data |
| 5. Feature Engineering | Making new columns that may be predictive. Potentially joining new data. Should have a good idea of which models to try by this point |
| 6. Model Building + Evaluation | Building a manageable amount of models.This part could include tuning the hyperparameters of the model. |
| 7. Model Selection | Select the model that performs best dependent on the metric(s) of interest e.g. Accuracy, F1 Score, Precision, Recall. |
| 8. Productionise Code + Serialise Model | Rewrite/write code as modules/functions in Functional or Object Orientated Programming Style. Most likely FP. |
| 9. API Exposure + Robust Engineering | Expose model and methods as an API (or read/write to a DB). Incorporate more software engineering. |

Exploratory work is usually done in markdowns (R) or notebooks (Python) where code is executed in chunks. Visualisations and commenting are used to document the Data Scientist's thought process through out development.

Due to the exploratory nature of the work, along with changing requirements, drip-fed information and new discoveries in the data, a Data Scientist may have to revert to a previous stage throughout the process. This puts Data Science at odds with Agile Software Development. This is reflected by the double ended arrows between steps.

Raw files/tables from clients should never be modified during this process as this would prevent the research/explorations from being reproduced.

The majority of a Data Scientist's time is spent working on **Supervised Machine Learning** problems. This is where the aim is to predict a label (**Classification**) or predict a value (**Regression**)

**It is not possible for a Data Scientist to know what set of algorithms to use for a problem until the data has been explored** (Step 4) and these will be narrowed down to a feasible number by the start of Step 6 and then selected at Step 7.
Step 7 is when the best algorithm is chosen using model based metrics/criteria (e.g Accuracy) and business specification (e.g. Client has a low tolerance to a black box solution)

A data scientist will now need to rewrite his code (the degree of the rewrite is dependent on how production ready they were thinking from Steps 2-7) from an exploratory/narrative style preferred by stakeholders in the business to a production ready application that can serve multiple requests from clients.

**Data Scientists are not software engineers but do have some knowledge of the discipline**. Data Engineers and Web Developers are usually required at this stage to help.

**Alistair Rogers**
**22 Oct 2018**