Applied Data Science - Notes

Dom Hutchinson

February 5, 2021

Contents

1 Introduction **2**

2 Data Ingress & Pre-Processing **3**

2.1 Data Structures 3

2.2 Data Formats 3

2.3 Web-Scraping & APIs 5

3 Privacy **6**

3.1 IRL Examples 6

3.2 Security & Statistical Database 7

3.3 k-Anonymity 9

3.4 Other Approaches 10

4 Ethics **11**

5 Storage & Management **11**

6 Transformation & Integration **11**

7 Exploration & Visualisation **11**

8 Deployment **11**

9 Sharing & Privacy **11**

1 Introduction

Remark 1.1 - *Types of Data*
Data comes in many forms including, but not limited to, the following

- Dense & Sparse data.
- Structured/Relational Data.
- Numerical; Categorical; Ordinal; or Boolean.
- Test (Emails, Tweets, Articles).
- Records (User-Level Data, Timestamped Event Data, Log Files).
- Geo-Based Location Data.
- Data-Time Data.
- Network Data.
- Sensor Data.
- Images and Video.
- Audio and Music.

Remark 1.2 - *Big & Small Data*
Whether a dataset is big or small depends on the computational-resources available, and thus will vary over time. Here are some ways to evaluate this

	Big Data	**Small Data**
Data Condition	Always unstructured, not read for analysis, many relational database tables that need to be merged.	Ready for analysis, flat file, no need to merge tables.
Location	Cloud, offshore, SQL server etc.	Database, local PC.
Data Size	Over 50k variables, over 50k individuals, random samples, unstructured	File that is in a spreadsheet, that can be viewed on a few sheets of paper.
Data Purpose	No intended purpose.	Intended purpose for data collection.

Remark 1.3 - *What is Data Science?*
Data-Driven Science. An interdisciplinary field about scientific processes and systems to extract knowledge or insights from data in various forms.

Data science incorporates fields from: Mathematics, Computer Science; &, Domain Expertise.

Remark 1.4 - *Motivating Applications*
Data science is motivated by its applications. Here are some examples of such applications

- *Amazon* use recommender systems to suggest products to customers.
- *Energy Companies* use data science to try and predict future usage of customers, so that resources can be applied efficiently.

- *Agriculture* use sensors in fields to collect data in order to monitor crops and predict weather.
- *Healthcare* use sensors in homes to monitor the health of people over long periods of time (especially when the person cannot go to the hospital).

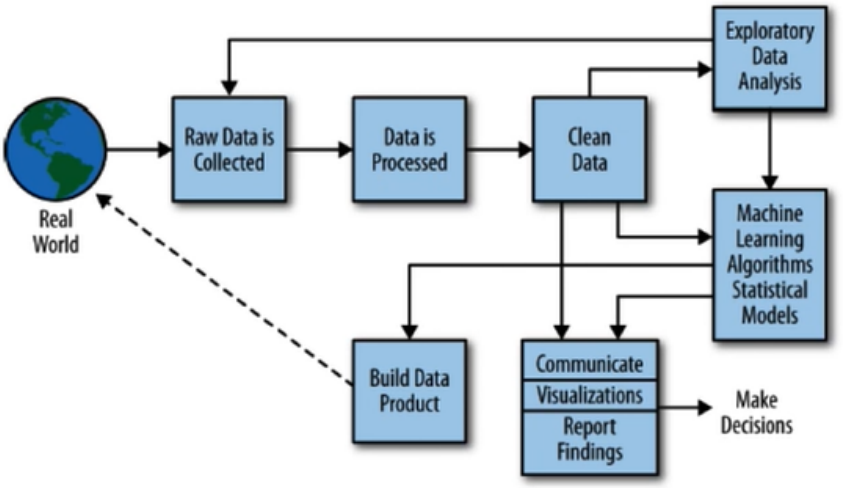

Figure 1: The pipeline for approaching problems in data science.

Proposition 1.1 - *Data Science Pipeline*
See `Figure 1` for a pipeline for approaching data science problems.

2 Data Ingress & Pre-Processing

2.1 Data Structures

Proposition 2.1 - *Native Python Data Structures*
Here are some data structures which are native to python and are popular in data science

- `list` - List of elements of varying types.
- `set` - List of unique elements of varying types.
- `dict` - Key-Value pairings.

Proposition 2.2 - *Non-Native Python Data Structures*
Here are some data structures which are not native to python but are popular in data science.

- `np.array`.
- `pandas.DataFrame`.

2.2 Data Formats

Definition 2.1 - *Object Persistence*
Object Persistence is the process of ensuring that the objects which are created are kept through multiple sessions. This comes in two stages

i). *Serialisation* - Translating data structures or objects from memory into a format which can be stored.

ii). *Deserialisation* - The inverse. Translating data structures which have been stored, into memory.

Remark 2.1 - *Bespoke Serialisation & Deserialisation*
Bespoke serialisation and deserialisation methods can be crafted manually. (e.g. Instantiating an output file; Writing each element of a list to a different line in the file; Closing the file.)

However, there are limitations to bespoke methods:

- Methods are specific to each use case (not standardised).
- Methods may not be robust.
- Methods require testing against many test cases.
- Object metadata is not encoded.

These limitations are rarely a problem in very controlled environments.

Definition 2.2 - *Comma-Separated Values (CSV)*
Comma-Separated Values (CSV) files are well suited to tabular data (inc. matrices). Each line of a *CSV File* stores one row of the table, and each element in a row is separated by a comma.

CSV Files are generally very readable, as well as time- and space-efficient for tabular data.

Remark 2.2 - *Using CSV Files*
As *CSV Files* are very popular, there are many library methods which can interact with them. Including reading & writing to and from memory (e.g. `pandas.read_csv`).

Definition 2.3 - *JavaScript Object Notation (JSON)*
JavaScript Object Notation[1] (JSON) is a standardised syntax for storing & exchanging data, used for serialisation. JSON uses text files which are human-readable and `dict`-like (i.e. have value-key pairs). JSON is designed to be simple so is a very robust language & suits many purposes.[2]

Limitations of JSON are that a specific conversion process may be required to convert a JSON file into objects in memory, and JSON files can become large due to key repetition.

Definition 2.4 - *Hierarchical Data Format (HDF5)*
Hierarchical Data Format (HDF5) is a standardised format for serialisation. HDF5 is a binary format and tries to mimic file system-like access. HDF5 files have the following three components

i). *Datasets* - Array-like collections of data. Thousands of datasets can be stored in a single file, and can be categorised and tagged however you want.

ii). *Groups* - Folder-like structures which groups datasets & other groups.

iii). *Metadata* - Information which pertains to all the datasets. (e.g. author, edit data & version).

[1] JSON is not just compatible with JavaScript and is common for many languages & APIs
[2] `https://jsonlint.com/` is a useful site for JSON validation.

HDF5 files are ideal for large numerical data sets, and can easily be manipulated by `numpy`. HDF5 files support a variety of transparent storage features (inc. compression, error-detection and chunked I/O), which `numpy.array` do not.[3]

Proposition 2.3 - *HDF5 files vs Traditional File Systems*
There are a few key differences between *HDF5 Files* and *Traditional File Systems*.

i). *HDF5 Files* are portable, as the entire structure is contained in the file independent of the underlying file system.[4]

ii). Datasets in *HDF5 Files* are all homogeneous hyper-rectangular numerical arrays, whereas files in traditional file system can be anything.

iii). Metadata can be added to groups in *HDF5 Files*. This is not possible in traditional file systems.

Remark 2.3 - *Other Standardised Serialisation Method*
XML, *Protocol Buffers* & YAML are other popular standardised serialisation methods.

2.3 Web-Scraping & APIs

Remark 2.4 - *Terms of Use*
Web Scraping should be done within the website's terms of use.

Definition 2.5 - *Web Scraping*
Web Scraping is the practice of collecting data from websites. This can take many forms, but typically involves taking a raw webpage and parsing the desired data.

Proposition 2.4 - *Approaching Web Scraping*
To perform *Web Scraping* successfully, you need a good idea of how a webpage is structure. Typically webpages are based around a `html` file, which are well structured[5]. Identifying combinations of tags, classes & ids in the `html` file can help locate the desired data.

It is harder, sometimes impossible, to navigate poorly designed websites as they are less structured and inconsistent.

Remark 2.5 - *Tools for Web Scraping*
Web Scraping technologies need to be tolerant to several artefacts of real-world data (known as *"wrangling"*) as-well-as errors in the website.

Some popular tools for *Web Scraping* are

- `BeautifulSoup` - A python library for parsing `XML` & `HTML`.
- `scrapy` - A python library. Generally faster than `BeautifulSoup`.
- `Selenium` - A web-browser plugin, generally used to test web services.

[3] `http://docs.h5py.org/en/latest/quick.html` provides a quick-start guide to using HDF5 files in python.
[4] However, it does depend on the HDF5 library.
[5] Webpages are often interpreted to have a tree structure, with each tag being a node

Definition 2.6 - *Web APIs*
Web APIs greatly simplify *Web Scraping* by provide a portal for explicit data acquisition, and are generally less prone to the issues which arise when *Web Scraping.*[6]

- The code running *Web APIs* is optimised for data requesting & retrieval. It does not waste time on visualisation or aesthetics. This means the bandwidth required for an *API* request is much lower than for a similiar *Web Scraping* process (As images etc. do not need to be loaded).
- *Web API* querying is robust, reliable, well maintained and documented with a static schema (`HTML`-based *Web Scraping* is not).
- *Web APIs* use standardised *Serialisation Tools* (e.g. JSON).
- *Web APIs* have already extracted and organised the desired data, however this does mean the user can only access what the operator will allow. This is much better than `HTML`-based *Web Scraping* where you rely upon fickle naming conventions of tags.

Definition 2.7 - *RESTful APIs*
Representational State Transfer APIs (REST/RESTful) are a popular form of *Web API* and generally require an *API Key* to access data.

Request to *RESTful APIs* generally involves constructing a URL which contains your keys and the parameters of your query.

Remark 2.6 - *Regular Expression (RegEx)*
Regular Expression (ReGex) queries are useful for extracting data, either while web scraping or from API requests.[7] Python has the `re` library for *RegEx* queries, two popular methods from this library are

i). `re.match` - Attempts to match a *RegEx* pattern to the whole string.

ii). `re.search` - Searches for the first occurrence of a *RegEx* pattern in a string.

3 Privacy

3.1 IRL Examples

Remark 3.1 - *AOL Data Incident, 2006*
In 2006, *AOL Research* released to the public, for the purpose of research, a text file containing 20mn search terms from over 650k users, over a 3-month period.

The users were anonymised in the in the data, but personally identifiable information was present in many of the search terms.

AOL Research retracted the document shortly after publishing.

Remark 3.2 - *Netflix Prize, 2007*
In 2007, *Netflix* ran a competition where $1mn was offered to anyone who was able to produce a system which outperformed their existing recommender system by at least 10%.

The training data offered by *Netflix* contained ¡user,movie,date,grade¿ and customer ids were anonymised.

[6] See `https://github.com/public-apis/public-apis` for a categorised list of public APIs.
[7] `http://pythex.org/` is a website which can perform *RegEx*.

It was later found that some users had their *Netflix* and *IMDb* accounts linked (so reviews were posted on both). This meant the training data could be partially de-anonymised. This led to lawsuits and *Netflix* cancelling future competitions.

Remark 3.3 - *Pay Attention*
The moral of these examples is that you have to be really careful and diligent when anonymising data as there may be non-obvious ways for people to de-anonymise it. (Often through little fault of your own).

3.2 Security & Statistical Database

Definition 3.1 - *Statistical Database*
A *Statistical Database* is a database for statistics. They can come in several forms

- *Tabular Data* - Tables with counts or magnitudes.
- *Queryable Databases* - On-line databases which accept statistical queries (e.g. min, max, sum etc.)
- *Microdata* - Files where each record contains information about an individual.

Remark 3.4 - *Privacy Trade-Off*
Statistical Databases have to trade-off privacy and functionality. Generally, more privacy means less functionality.

Definition 3.2 - *Identifiers*
Identifiers are attributes that unambiguously identify someone (e.g. passport number, NI number, name etc.)

Definition 3.3 - *Quasi-Identifiers*
Quasi-Identifiers are attributes which identify someone, but there is some ambiguity. Often, having multiple *Quasi-Identifiers* are enough to confidently identify someone. (e.g. address, age, gender).

Definition 3.4 - *Confidential Attributes*
Confidential Attributes are attributes which contain sensitive information about an individual (e.g. salary, religion, medical conditions etc.)

Definition 3.5 - *Non-Confidential Attributes*
Non-Confidential Attributes are attributes which do not contain sensitive information about an individual.

Remark 3.5 - *Anonymisation, Identifiers & Quasi-Identifiers*
When anonymising a data set:

- *Identifiers* need to be suppressed from the data sets (i.e. removed).
- *Quasi-Identifiers* hold disclosure risk as they can often not be suppress due their high analytical value, but can often be linked with other non-anonymised datasets to de-anonymise your dataset.

Definition 3.6 - *Attribute Disclosure*
Attribute Disclousure is when the value of a confidential attribute is made available, and in doing so the individual is easier to identifier (compared to the value not being disclosed).

Definition 3.7 - *Identity Disclosure*
Identity Disclosure is when an entire record in an anonymised data set can be linked with that individual's identity.

Definition 3.8 - *Membership Disclosure*
Membership Disclosure is whether or not data about an individual is contained in a dataset.

Definition 3.9 - *External Attack on a Table*
An *External Attack* on a table occurs when another dataset is released which contains data which comprises more table, or can be combined with your table to comprise others.

Definition 3.10 - *Internal Attack on a Table*
An *Internal Attack* on a table occurs when those within the table (i.e. with access to it and know which data represents them) are able to deduce which data belongs to other people by a process of elimination.

Only really possible when the table has few respondents.

Definition 3.11 - *Dominance Attack on a Table*
An *Internal Attack* on a table occurs a (or a few) respondents dominate the contribution in a particular cell of a magnitude table, these *Dominant Respondents* are then able to upper-bound the contributions from the rest due it being easier to differentiate other users from them.

Proposition 3.1 - *Combatting Table Attacks*
There are two main approaches to combatting *Table Attacks*

i). *Non-Perturbative* techniques which do <u>not</u> modify the values in the cells, but they may suppress or recode them (e.g. cell suppression, recoding of categorical attributes).

ii). *Perturbative* techniques do modify the values in the cells (e.g. controlled rounding).

Proposition 3.2 - *Combatting Database Attacks*
There are three main approaches to combatting attacks on databases, all focused around adjusting/controlling queries

i). *Query Input Perturbation* - Perturbation is applied to records on which queries are computed.

ii). *Query Output Perturbation* - Perturbation is applied to the results of a query.

iii). *Query Restriction* - The database refuses certain queries.

iv). *Camoflage* - Returns answers which have been made non-exact in some deterministic fashion.

Remark 3.6 - *Differential Attack*[8]

[8]Not proper name.

Suppose you know that someone is going to be changing doctors soon (and thus will be removed from the doctors database), if you copied the database before and after this person moves you could likely deduce details about the individual by comparing the two versions.

How do we prevent this?

Definition 3.12 - *Differential Privacy*
Consider a learner implements a summary statistic A (e.g. sum of all elements); an adversary proposes two datasets S and S' which differ by only one row; and a test set Q. The learner wants the summary statistic to have the same value if only one row has changed.

Summary statistic A is *ε-Differentially Private* iff

$$\left|\ln\left(\frac{\mathbb{P}(A(S)\in Q)}{\mathbb{P}(A(S')\in Q)}\right)\right| \leq \varepsilon$$

Note that *Differential Privacy* is a condition on the release mechanism of the data A, not on the data S, S' itself.

Remark 3.7 - *Differential Privacy in Practice*
In practice we avoid differential privacy issues by adding noise to our responses (either to all responses, or incorrectly responding to some queries).

3.3 k-Anonymity

Definition 3.13 - *k-Anonymity*
A dataset satisfies *k-Anonymity* if each combination of values of *Quasi-Identifier Attributes* in the table are shared by at least k records.

Measuring *k-Anonymity* requires a brute force approach which compares every pair of records in the table.

Proposition 3.3 - *Achieving k-Anonymity*
There are two main approaches to achieving *k-Anonymity*:

- *Suppression* - Replace the values of attributes with an asterisk.
- *Generalisation* - Values of attributes are replaced with a broader category. (e.g. age-ranges rather than exact age)

Proposition 3.4 - *k-Anonymity Attacks*
k-Anonymity does not protect against attribute disclosure attacks in general if the values of *Confidential Attributes* are very similar in the group of k records who share *Quasi-Identifier* values.

Here are two specific attacks that can happen

- *Homogeneity Attack* - If all records in the set of k identical records have the same *Sensitive Value*, then the value can likely be predicted (and would comprise all k).
- *Background Knowledge Attack* - If associations exist between one or more *Quasi-Identifier Attributes* and a *Sensitive Attribute* then the set of possible values of the *Sensitive Attribute* is reduced.

Definition 3.14 - *l-Diversity*
l-Diversity is an extension of *k-Anonymity*.

l-Diversity requires that the values of all *Confidential Attributes* within any group of k-anonymous records contain at least l clearly distinct values.

Proposition 3.5 - *Limitations of l-Diversity*
Here are two limitations of *l-Diversity*

- Not every value shows equal sensitivity. (A rare positive indicator for a disease may provide more information than a common negative indicator).
- It ensures diversity of sensitive values in each group, but it does not recognise that values may be semantically similar.

Definition 3.15 - *t-Closeness*
t-Closeness requires that the distribution of the confidential attribute within a group of k records is similar to the distribution of the confidential attribute in the entire data set (at most distance t between the two distributions).

Remark 3.8 - *Tools for Anonymisation*
Here are three popular tools for anonymisation

i). *ARX* - Applies k-anonymity, l-diversity and t-clossness in Java.

ii). *Argus* - Creates safe microdata files.

iii). *SdcMicro* - Provides disclosure control methods for anonymisation and risk-estimation.

3.4 Other Approaches

Definition 3.16 - *Privacy-Preserving Data Mining, PPDM*
Privacy-Preserving Data Mining (PPDM) seeks privacy for the data's owner when several owners wish to co-operate without giving away their data to each other

Definition 3.17 - *Private Information Retrieval*
Private Information Retrieval (PIR) seeks user privacy to allow the user of a database to retrieve information without the database knowing which item was retrieved.

4 Storage & Management

5 Transformation & Integration

6 Exploration & Visualisation

7 Deployment

8 Sharing & Privacy