Chapter 11 Data Science: Investigating Big Data

TL;DR Collect Raw Data

Identify available relevant datasets

- Existing collections (NB their collection isn't your collection: be aware of biases, data gaps etc)
- Social media feeds: searches, APIs, and spreadsheets

Extract data

- Get data onto your machine / into the team's datastores
- In a format you can use: pdfs aren't usable data, and a spreadsheet isn’t a MISP object…
- check that your data is clean

Data science

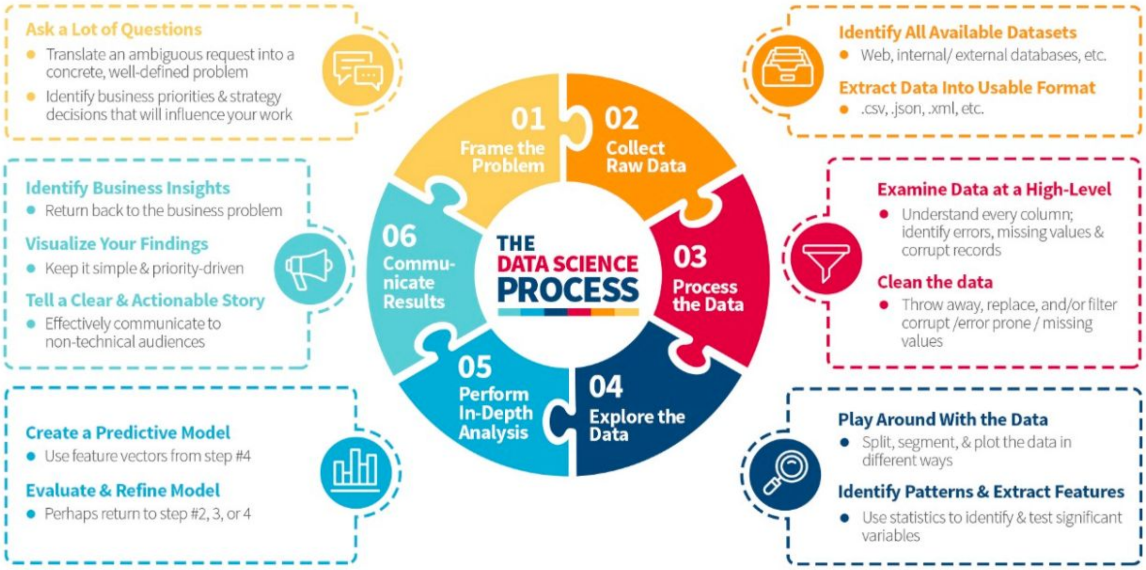
THE DATA SCIENCE PROCESS
01 Frame the Problem
02 Collect Raw Data
03 Process the Data
04 Explore the Data
05 Perform In-Depth Analysis
06 Communicate Results
Ask a Lot of Questions
Translate an ambiguous request into a concrete, well-defined problem
Identify business priorities & strategy decisions that will influence your work
Identify All Available Datasets
Web, internal/ external databases, etc.
Extract Data Into Usable Format
.csv, .json, .xml, etc.
Examine Data at a High-Level
Understand every column; identify errors, missing values & corrupt records
Clean the data
Throw away, replace, and/or filter corrupt /error prone / missing values
Play Around With the Data
Split, segment, & plot the data in different ways
Identify Patterns & Extract Features
Use statistics to identify & test significant variables
Create a Predictive Model
Use feature vectors from step #4
Evaluate & Refine Model
Perhaps return to step #2, 3, or 4
Identify Business Insights
Return back to the business problem
Visualize Your Findings
Keep it simple & priority-driven
Tell a Clear & Actionable Story
Effectively communicate to non-technical audiences

image from https://medium.com/the-mission/deconstructing-data-science-breaking-the-complex-craft-into-its-simplest-parts-15b15420df21

Data science is a process. There are many versions of this process, but it basically goes from identifying and asking a set of business questions, to attempting to answer them by finding and cleaning datasets, building models based on them, and using those models to understand part of the world and/or predict what might happen next in it, then explain that the people who need to make decisions based on your findings.

The start and end of data science is all about people: we need to think in terms of end-users, questions and problems.

Disinformation Modelling at Scale

There's a lot of academic work on modelling disinformation at scale. Some of the models used include:

- automated fact checking
- cascade and time-based models
- social network analysis \(Pablo is keen on scale-free networks for this\)

Tactical data science

Working from the data we have instead is instructive and can teach us things about the disinformation environment, but that’s not tactical data science. Most of the work so far hasn’t been tactical. At speed, this becomes a threat intelligence nerd fight, that looks very similar to the other threat intelligence nerd fights: disinformation creators vs disinformation defenders.

A lot of what we do is detective work, where the algorithms and tools are there to assist us. This has a lot in common with data forensics, threat intelligence work and OSINT.

Where and how to look for examples

“tactical data science” is the work you do in the moment, chasing disinformation incidents and campaigns as they happen. There’s a lot of literature out there on disinformation algorithm design, which is nice, useful in some circumstances \(e.g. as a dayjob\), but not helpful to people faced with “social media is happening, work out how to reduce harm”. There’s a lot there of the “there’s a dataset, let’s see what we can do with it” persuasion.

Places to look for ideas in this field include:

- Trained amateurs - sites like towards data science \(lots of student projects\), github, medium.
- Academics - known research groups, paper repositories, conference outputs
- Adjacent groups
- Student projects - yes, it’s students, but they’re usually supervised in latest techniques, keen to try them out online, and willing to write up their code.

Good search terms include “computational propaganda”, “misinformation”, “disinformation”.

This is where the data science comes in...

Where does Disinformation Data Come From?

The cynical amongst us would say that we're drowning in disinformation data. Mainstream news has many stories about disinformation incidents and takedowns, political groups are quick to decry "fake news", and almost everyone working on disinformation has a favourite fake cure or conspiracy theory.

In practice, if we're looking for disinformation in our specific areas of interest \(e.g. the CTI League currently works on Covid19 related disinformation\) in time to make a difference to its effects, we need to do some groundwork and build out connections, information feeds and catalogues of good places to look.

Types of data

We need to think about data. Mostly we're dealing with data that's moving, at rest and static.

- Moving data: A lot of research places have social media listening - downloading all the social media messages etc around topics, hashtags etc of interest.
- Data at rest: this is the data we've grabbed during investigations, usually as part of finding more of a network and its effects. We're often actively analysing it, working out how we can affect the environment it's in.
- Static data: this data isn't going to change. Some of it is moving data that we've stored, and the environment it was in has been overtaken by events. It's of interest because it contains patterns to be mined, and could contain clues to later behaviours. Other static data is used to support investigations.

Data inputs: Alerts and Canaries

We receive alerts about possible disinformation incidents from members of the disinformation team, and from other teams connected to us. Typically we get alerts around an artefact or theme, e.g.

- A new narrative emerging online, either in general social media or known conspiracy / extremist / target etc groups
- A local or world event that might spark a disinformation incident
- Anomalous or significant-sized online activity that might be associated with a disinformation incident
- Command signals from known disinformation groups (e.g. qanon)

The types of artefact that we typically receive include:

- Images
- Messages, e.g. tweets, facebook posts, SMS or Messenger/Telegram etc messages

- URLs

The processes for investigating these are discussed in more depth in the next chapter.

Several accounts and groups are either known producers or early adopters of many disinformation campaigns. We've dubbed these "canaries", as in the entities that give the first signals that something is happening \(canary, as in "canary in a coal mine"\).

Data sources: disinformation data streams

When we get our first data inputs, it's a good idea to check them against other disinformation and related data collections, to see if they've been picked up by other researchers, or those researchers have already collected data related to these inputs that can be of use to our investigation. The data feeds are continually updated, so are a good source for breaking data; the static data collections are good for finding history on data, source, narratives etc.

Narratives

- EuVsDisinfo database https://euvsdisinfo.eu/disinformation-cases/
- Ryerson Claimwatch dashboard
- Indiana Hoaxy (twitter, articles)

Data

- Botsentinel: lists "trollbots" (bot-like and troll-like accounts) and the themes they're promoting https://botsentinel.com/ (not just Covid19)
- Hamilton68 - live feed from accounts attributable to Russia or China (may or might not contain propaganda; useful for seeing current themes). Public version is live feeds from official Russian sites (embassies, RT etc), not trolls. Academics can ask for a more detailed feed. https://securingdemocracy.gmfus.org/hamilton-dashboard/ (not just Covid19)
- Indiana University OSOME Decahose

Counter-disinformation feeds

- Snopes: (https://www.snopes.com/)

General disinformation datasets

- Twitter IO archive: covers several countries up to a few months ago. Good for getting a sense of the size and ‘feel’ of typical nationstate twtter posts/ networks etc. https://transparency.twitter.com/en/information-operations.html
- Facebook ad library: contains all active ads that a page is running on Facebook products https://www.facebook.com/ads/library/ (About the Ad Library)

Collecting your own data using tools

The datastreams above will help you get a sense of what’s known about the artefact and/or theme that you’re investigating, and sometimes that’s enough to craft a response \(e.g. if there’s a WHO page on a known scam, that might be enough evidence to ask for takedowns etc\). But most of the time, you’ll have to go collect your own data from across social media, and sometimes beyond \(e.g. for paper flyers, we asked people if they’d seen them in their neighbourhoods too\).

Where you collect from, and what you collect will depend some on the artefacts you found, but here are some of the ways.

Twitter data

Twitter data is studied a *lot\precisely because it has a lovely API. Since we use a lot of Python here, let’s talk about Python libraries. If you have twitter API codes, then Tweepy is a good choice. If you don’t want to use the twitter API, try Twint.

Various researchers post twitter data-gathering tools online. Andy Patel’s twitter-gather is good if you’re doing twitter network analysis. We have code based on an early version of this in the github repo. It’s andy_patel.py - call it with “python andy_patel.py name1 name2 name3 etc” where name1 etc are the hashtags, usernames, phrases \(phrases in quotes\) that you want to search Twitter for. Andypatel.py creates a set of files in directory data/twitter/yyyymmddhhmmss_hashtag1 etc with the tweets, most prolific urls, authors, influencers, mentions etc and gephi input data so you can create user-user etc graphs (see the gephi instructions in this BigBook for how to do that\). Data for earlier investigations are in the repo folder data/twitter if you want to see what that looks like.

Useful references on collecting twitter data include https://firstdraftnews.org/latest/how-to-investigate-health-misinformation-and-anything-else-using-twitters-api/

Facebook data

The Facebook API is horrible. Most everyone tracking social media uses a third party like CrowdTangle or scrapes for the data they want. The Crowdtangle chrome extension is available free to anyone, but the full Crowdtangle tool isn't: it's available to news organisations, some academics, and pilot programs, so it's worth checking to see if your team is eligible or has someone with access on it.

Reddit

Reddit data is regularly dumped in an easy to read format. For quick-looks, there are tools like (https://www.redective.com/)

Multi-platform tools

Reaper collects from a set of social media feeds. Trying that out.

Access tokens:

- Facebook: look at list in (https://developers.facebook.com/docs/facebook-login/access-tokens/) - then used (https://developers.facebook.com/tools/explorer/) to check token worked before putting into reaper.
  - “Page Public Metadata Access requires either app secret proof or an app token” - see (https://developers.facebook.com/docs/apps/review/feature#reference-PAGES_ACCESS)

Storing datasets

Social media data can be large, and its value is often in the relationships between objects as well as the objects themselves. Options we've used include collections of CSV and json files held in a DKAN data warehouse, Neo4j and an ELK stack https://www.elastic.co/.

Disinformation Data Science Courses

The CogSecCollab / CTI League team has a data science for disinformation response training series, including:

1. Setting up for disinformation activities: goals, ethics, groups, examples, practice
2. Disinformation basics: creators, outputs, mechanics, effects, feeds
3. Disinformation layers: from strategy/information ops down to tactics/artefacts/TTPs
4. Data collection: OSINT, platforms, user- and network- analysis level tools
5. Handling big data: APIs, cleaning, exploration, storage, automations
6. Communicating results: reporting routes, visualisation, tools, practice
7. Social text analysis: features, techniques, tools, generation methods
8. Image data analysis: tools, techniques, deepfakes/cheapfakes
9. Relationships as data: data-as-networks, features, tools
10. Using machine learning: extending your analysis with ML/AI
11. Acting \(ethically\): counters, coordinations and more ethics
12. Measuring effectiveness: measuring campaigns and counters

This was based on a university course, the aim of which was to take computer science students through the process, learning the basics of disinformation mechanics \(there are many good papers on that\), then walking through the data science processes that are particular and peculiar to tracking large disinformation campaigns across social media at speed.