Writing a problem statement [DS4B Session 1d]

Data work can sometimes seem meaningless. You go through all the training on cool machine learning techniques, find some cool datasets to play with, run a couple of algorithms on them, and then. Nothing. That sinking feeling of "well, that was useless".

I'm not discouraging play. Play is how we learn how to do things, where we can find ideas and connect more deeply to our data. It can be a really useful part of the "explore the data" part of data science, and there are many useful playful design activities that can help with "ask an interesting question". But data preparation and analysis takes time and is full of rabbitholes: interesting but time-consuming things that aren't linked to a positive action or change in the world.

One thing that helps a lot is to have a rough plan: something that can guide you as you work through a data science project. Making a plan for data science has much in common with making plans for Lean and design thinking: you're putting in effort, so you need to be:

- focused on the change you want to make in the world, e.g. solve a problem or change people's minds (there's no point doing analysis if you don't do anything with it),
- pragmatic about the work you need to do (sometimes the answer is a piece of paper, e.g. much simpler than the beautiful concept you had in your mind, but much more likely to be used in the current context) and
- realistic about the problem you're trying to solve and the resources you have around to do that (e.g. use what works for you, not an unachievable 'best' technology).

I like Max Shron's CoNVO for planning small data science projects. In my dayjob, I work a lot with Lean Enterprise and Design Thinking techniques to achieve similar results at scale, but at the least you should have an A4 piece of paper somewhere with this on it, to refer back to:

- **Context**: who needs this work, and what are they doing it for?
- **Needs**: what are you trying to fix?
- **Vision**: what do you expect your final result to look like?
- **Outcome**: how do you get your results to the people who need them? What happens next?

The other question you're going to want to ask is "is it worth me doing this work". It's okay to say "yes, because I learn from it", or "it's fun", but data science is work, and you don't want to feel like you put in a ton of effort and late nights only to realise that effort was wasted. I like the DrivenData competition guidelines for helping with this thinking:

- **Impact:** "… clear win for the organisation in terms of effective planning, resources saved or people served… good story around how they generate social impact…"
- **Challenge:** "… challenging enough for a rich competition…"
- **Feasibility:** "….the right kind of data to answer the question at hand… does it have enough

signal to be useful?..."
- **Privacy:** "… can answer this question while protecting the privacy of individuals in the dataset and the operational privacy of an organisation…"

I haven't yet mentioned the thing that your work will focus on: the "interesting question(s)" that you're trying to answer. There are several contexts you might find yourself in here, from a business team bringing you a well-defined question (at which point you start with the "what is the question you're really trying to answer here" discussion), to having complete freedom over the questions you're asking and the ways they could be turned into action.

One way to get better at asking good questions is to see what other people ask. Look at your subject area: find other projects and questions in it, and see how they're asked (and answered). Look at existing data science projects for inspiration, e.g. Kaggle (and their UseCase list), DrivenData, DataKind and the projects listed in the course reading list, then design your questions. Asking questions about your questions can help here, e.g.:

- Is the question concrete enough? Is it solving a real problem, or just a symptom of that problem? (e.g. "what are the barriers to people engaging with us" vs "how can we get more people to call")
- Can you translate the question into an experiment? g. can you ask something like "I believe people have more phones than toilets" and start proving (or, more generally, disproving) that.
- Is it actionable? And what actions will be taken given the answer?
- What data is needed to do the analysis? At this point, datasets could be anything - tables, images, maps, sensor feeds; anything. Be aware that although data access can limit what you can do, data is just a support here, and focusing on the question can help you think about other ways that you might be able to answer it.

As you do this, you'll find yourself questioning the meaning of many of the components of your original question: I have a longer blogpost on that, but this is where the plan becomes useful: instead of focusing on "what actually counts as a toilet" (yes, that really is a difficult thing to define), go back to your notes about who this is important to, why, and what they could do about it. You'll also find that a seemingly-simple initial question will generate a whole bunch of other questions you'll need to answer too (questions are like bunnies: they breed). Again, use your plan as a guide, and accept that there will usually be several parts to each project. You'll also find that several of these questions could be answered without using available data (e.g. you might be able to get a strong enough 'signal' from surveys that an action is worth further investigation): that too is a useful thing to know.

Plans rarely survive contact with your datasets, users etc. They're not about forcing you to produce things a specific way: they're there to make you think about what you're doing, and stop you from making newbie mistakes like fitting the questions to the data you have available or falling

into data rabbitholes. You might want to go down at least one of those rabbitholes: a fascinating piece of data that you want to explore for fun, or a bunch of other questions that look like really fun things to answer; they're not necessarily bad things, and can be really valuable in themselves, but you do need to be aware that you've done this, and of any impacts it might have on your original goals. Planning might seem a distraction from getting on with the data analysis, but it does help to have a guidestar, something to go back to and think "is this valuable to the people that I'm trying to help here?".

Some short exercises

We ran 3 small exercises in class, to get people thinking about project design. Each of these was time-limited to 3 minutes, to make people concentrate hard on what might be needed, and to hit issues quickly so they could be discussed in class before students tried this at home.

Exercise 1: Ask some interesting questions. Either your own questions, or pick an existing question and think about how it might have been formed.

- Questions that data might help with
- Stories you want to tell with data
- Datasets you'd like to explore (where 'datasets' could be anything - tables, images, maps, sensor feeds, etc)
- Competition questions: Kaggle, DrivenData
- A data science project that interested you

Exercise 2: Get the data. Pick one of your questions:

- List the ideal data you need to answer it
- List the data that's (probably) available

Think about what you'll do if the data you need isn't available:

- What compromises could you make
- Where would you look for more data
- Are there proxies (other datasets that tell you something about your question)
- Are there ways to get more data (surveys, crowdsourcing etc)

Exercise 3: Design your communications. List the types of people you'd want to show your results to.

- How do you want them to change the world? Can they take actions, can they change opinions etc
- Describe the types of outputs that might be persuasive to them - visuals, text, numbers,

stories, art… be as wild with this as you want