Wut 2 Do Wif Data?

Data science is not about data. Data science is about insight – the knowledge and suggestions that you can glean by inspecting and using data. And that insight usually starts with a set of questions. Here are some examples, hopefully making you think a bit more about your own questions (which in Emily's case is the correlation between cuteness, cuddles and the amount of Meow Mix in her dish).

You don't always know what the good questions are, but you usually know (or pick) the framework that you're asking them in. This is how I usually approach this:

- Look at context - ask question (or get question from user)
- Get data
- Phrase question in way that data can answer
- Write down issues with data
- Clean data
- Investigate question
- Check conclusions and possible issues with conclusions
- Describe possible further investigations / data gathering
- Which might mean improving on the data that you obtained this time

Here are two examples, to help you think about your own questions. One example analyses text, the other numbers; both are simple but raise many difficult questions.

Example 1: Starting with a question

Look at context- ask question

I've somehow been spending a lot of time lately thinking about poo… erm… sanitation, open defecation and farm slurry. Some of this stemmed from a question I asked about a UN 'fact' that was quoted without a data provenance – that more people have access to a mobile phone than to a toilet. My question was simple: "is this true?".

Get data

Now at this point, I had no data. So I looked at the resources I had available (me and an internet full of open data) and the value of the result (me satisfying my curiosity), and scoped out the size of the project: I'd look for open data (i.e. not ping any of my contacts for data, set up surveys or anything that involved other peoples' goodwill – that's a valuable resource), and use that to determine whether the question could be answered. I'm spoiling the surprise, but this is something that happens a lot with development data: you start out with a clear question, find that the data isn't there to answer it, then adapt either the data (by gathering more) or the question (by

reducing its scope, or changing it to a set of also-valuable questions that the data can help you with).

So data. I searched all the usual suspects (see opencrisis.com for a list), but couldn't find any dataset of surveys that included both access to toilets and mobile phones. There's probably been one or more of these done, they could probably be dug up with a lot of phone calls, but they weren't easily visible online. The datasets that I did find were one on sanitation from WSSinfo and another on mobile phone densities from ITU. And these have issues:

- The datasets were hard to find.
- I looked at the last 5 years (anything older than that in development isn't that useful), but there was no data after 2010 in these datasets.
- The datasets were unrelated
- The dataset formats were hard to machine-read (they included merged cells, explanations etc).
- It was difficult to track provenance – e.g. what decisions did the people creating these datasets make? What assumptions?
- There were data issues: numbers were rounded up, data was at country level, countrynames didn't match between the two datasets, there were multiple charactersets in the files (e.g. Å, A, ?).

Phrase question in way the data can answer

So onto the question. Taking the question "more people have access to mobile phones than toilets" as a start point, we can rephrase this as: number of people with mobiles > number of people with toilets

or (mobile% - toilet%)*population > 0

or (mobile% - (100-opendef%)) > 0

Where mobile% is the percentage of people with mobile phones, toilet% is the percentage of people with access to a toilet (not, note, owning a toilet - or I'd be looking through the sanitaryware import and latrine digging figures for each country), opendef% is the number of people open defecating (pooing outside). And we can answer this question using with the datasets.

Write down issues and clean data

And even once the numbers for open defecation (a polite phrase for "has no toilet and has to poo outside") and telephones were compared, that comparison only created a bunch more questions. Most of these questions exist because of the idea of statistical independence – if you gather two datasets independently of each other, it's only possible to compare them under some

really tight statistical conditions. Some of these questions were:

- Is there actually a correlation between the two datasets? Phone densities are quoted as the number of phones per hundred people, and are often over 100 (I think I have 4 phones at home, but I've lost count now). Most of the countries with phones > toilets are in the developing world: don't some people in the developing world have more than one phone? In some cities (e.g. Benin City) I've visited, phone signal availability is so variable that people have up to 5 simcards each, on different carriers. Were the results uniform - the datasets were listed by country - what if the cities have lots of phones and toilets, and the rural areas don't? What does that do to the numbers?
- And how do you count up people without toilets? Are these percentages estimates or survey results? If they're surveys, how big were the surveys, and were they demographically and geographically representative (e.g. were city and country people surveyed proportionately, and how was this done - on paper or by phone?). We're talking about people here - how likely were they to be truthful about toilets - having to poo outside could be deeply embarassing, and perhaps hard to admit.
- Where does my composting toilet fit in this? If I have an 'unusual' outdoor toilet, does that count as a toilet or open defecation?
- What do we do with a zero value in the datasets? What do we do with values over 100 per 100 people (I truncated these to 100, so extra phones had less of an effect, but I felt uneasy doing that).
- Did we just list the people who, with the right tools, can campaign for more toilets?
- Etc…

Investigate question, check conclusions, describe possible future investigations

So, having found run the question against the data, here are the numbers for 2010:

country	population	opendefecation	not opendefecation	phones	phones minus loos	people affected
India	1.22E+09	51.09471	48.90529	61.4226	12.51732	153288799
Indonesia	2.4E+08	26.25828	73.74172	88.08497	14.34325	34405290
Brazil	1.95E+08	3.694356	96.30564	100	3.694356	7202000
Morocco	31951000	15.86805	84.13195	100	15.86805	5070000
South Africa	50133000	7.745397	92.2546	100	7.745397	3882999
Viet Nam	87848000	4.177671	95.82233	100	4.177671	3669999
Benin	8850000	56.39548	43.60452	79.94351	36.33899	3216000
Cambodia	14138000	60.53897	39.46103	57.65042	18.1894	2571616
Peru	29077000	7.232521	92.76748	100	7.232521	2102999
Colombia	46295000	6.486662	93.51334	96.07475	2.561412	1185805

Mauritania	3460000	53.64162	46.35838	80.23792	33.87954	1172232
Guatemala	14389000	6.046285	93.95371	100	6.046285	870000
Namibia	2283000	51.86159	48.13841	85.50451	37.36609	853067
Ecuador	14465000	4.638783	95.36122	100	4.638783	670999
Honduras	7601000	8.748849	91.25115	100	8.748849	664999
Niger	15512000	78.85508	21.14492	24.53329	3.388367	525603
El Salvador	6193000	5.926046	94.07395	100	5.926046	367000
Botswana	2007000	15.39611	84.60389	100	15.39611	309000
Mongolia	2756000	11.71988	88.28012	91.09104	2.810925	77469
Suriname	525000	6.095238	93.90476	100	6.095238	32000

Reading the whole table, the bottom line is that 200 million or so people have phones but not toilets, if you use the ITU and Wssinfo data, and ignore statistical independence (that's an enormous ignore). That's out of 7 billion people worldwide. So yes, it's potentially an issue, but it's more interesting to think about where, and what that means. For instance, there are 200 million people with phones who, if they get the right SMS apps or information, can lobby for governments and NGOs to build toilets in their areas, or for the plans, materials, money or labour to do this for themselves. If anyone wants to start a "givemealoo" site with an SMS connection and publicity through SMS and local radio, they now know where to start…

Example 2: Starting with a dataset

Sometimes you start with a dataset, and the question "what can you glean from this?". For instance, my partner had a set of job descriptions that he liked, and wanted to find more like them. The long answer would be to do some supervised learning with these and other descriptions, and build a jobsite scraper that classified each description into "interesting" or "not interesting". The short answer was to look for patterns, features and possibly clusters in the dataset.

The data was from a mix of different websites, all with a different structure (and different headings for 'experience' etc.), so I treated each page as unstructured text (e.g. I ignored labels and punctuation and treated each page as a huge collection of words). I started by building a histogram of the words used: a list of the top 30 words I found across all the documents, with how many times each one appeared. This list contained a lot of stopwords – common words that don't add anything useful to the histogram, like "and", "the", "of", "to" and "in", that I then removed from the list, to give a list of terms that might be useful to Dan.

Removing stopwords is a common thing in text processing - normally I'd use a standard list of stopwords (e.g. Porter) for this, but I didn't want to miss any industry-specific terms that might be on those lists, so I built my own stopword list. For development data, you'll probably do this a lot too, e.g. "crisis" isn't a really useful term to find when you're working on crisis information. So I

built a histogram (minus stopwords): the top 10 words in it were: estate (26), real (26), development (12), design (11), manage (11), planning (11), sales (11), investment (10), senior (8), portfolio (8).

I showed this to Dan and he said “great – but what about pairs of words”. .. something that might have been triggered by the top 2 words on that list (“real” and “estate”). So I modified the code to produce a histogram of adjacent words, and got: real estate (26), new york (5), trade marketing (4), job description (4), estate portfolio (4), senior strategist (4), city area (3), estate investment (3), funding approvals (3), area job (3).

I could have continued this – looking for chains of words, e.g. “real estate” linked to “estate portfolio” etc., and linked it to a jobsite scraper to automatically alert Dan to jobs that were similar to his “interesting” ones (you’ve probably worked out by now that he’s a real estate architectural designer), but the lists enough were enough for him: he got search terms that he hadn’t thought of, and is happily sifting through sites with them. Which is another lesson to learn: sometimes a seemingly simple thing will have enough of an effect to make a user happy, without needing complex analysis. Unless you're playing with a dataset out of curiosity, that's often a good place to stop.