Sampling strategies & rare events

Tackling the problem that most people do NOT visit the sites we are interested in

Damian Trilling

28-10-2018

ccs.amsterdam

The ideal scenario

1. We survey people about their interests, attitudes, …
2. We track what they *really* do
3. (optionally): We survey them again

The problem

Even if we start out with a massive dataset of http requests/URLs visited/..., we end up with a tiny fraction being relevant

- People use multiple devices (and apps!) – we miss out on a lot of (news) exposure
- Most people do *not* visit *the same* news sites (long tail)
- Even if they read about the same event, they do not necessarily do so via the same URL

An example

From the Personalised Communication project

- Around 500 persons installing the plugin
- Only $\approx$ 150 actually generating data (possible reason: using different computer, browser update, plugin uninstalled, technical problem on our end?)
- Out of them, maybe 40 read news at a somewhat regular basis
- In any given week, probably not a single article read by more than one person

(numbers are rough indicators)

Solutions?

- Recruit a massive sample (how?)
- Let go of idea of representative sample; sample on sub-group we want to study (e.g., regular users of news sites)
- Let go of idea of fine-grained analysis; focus on broad categories (in survey *and* tracking data)
- ...

The big question

What’s possible — and what’s not even worth trying?