Pitching Hypothesis-Driven Data Investigations

This document asks essential questions to plan data experiments. Revisit these questions throughout your reporting, and use them to communicate your intentions and limitations with your editor. It will help determine if a story is worth pursuing by giving an estimation of time, complexity, and impact.

As a side benefit, these questions form the backbone of a methodology to get reviewed by experts, as well as the target of your investigation.

What is the hypothesis of the story?
This is a testable claim that can be proven or disproven. The hypothesis is revised and honed in the early stages of an investigation.

Who is being harmed and at what scale?
Give a sense of who is being harmed, and whether outcomes are proportionate?

Who is causing the harm and what is the accountability angle?
What frameworks exist– legal, company policies, sworn testimonies, to hold them accountable?

What is the evidence (anecdotal or otherwise) you've gathered that leads you to think you have a viable hypothesis?
If there is existing work on this topic (journalistic, academic, lived experience) how will you build off that work?

What is a viability study you can perform?
How can you begin to collect and analyze data as a proof of concept?

What data will you need to run an analysis? How will you gather the data?
Web scraping, public records requests, using open data? How will you decide what makes up your sample? Specificity is your friend.

How complicated is the data collection?
Will you be able to do this alone? Do you need to use proxies, cloud instances? Do you need to collect data over time?

Will you need to clean and filter that data?
What records will you throw out and why?

And what are the limitations of the data set(s) you are proposing to use? How will you test its accuracy?
Every dataset is imperfect and carries inherent assumptions. How will you assure what you're writing about is aligned with your dataset and bulletproof that dataset?

Do you need to classify the data for your experiment? If so, please describe how you propose doing that. Are there outside classifications or experts you can lean on? What are the limitations of your classification method?
Seldom is the outcome you want to measure already encoded in a column. Instead, you need to make that assessment.

How will you analyze the data? What statistical tests will run? Please list any limitations to your proposed method and any alternatives.
Start with simplicity.

What specific sentences will you be able to write based on your findings? What's the headline? What's the lede?
Use this space to brainstorm how you'll frame your findings. This helps make your findings tangible. I use TK's at this step.

Can you imagine the charts or other visualizations this data will produce?
How will you communicate your findings visually?