Working with “big” data from private firms: 7 lessons learned

Prasanna (Sonny) Tambe, NYU Stern

NBER Digitization Winter Conference

March 6th, 2015

(This content has benefited from discussions with John Horton and Andrey Fradkin)

Some of the datasets I've worked on with co-authors

(I am interested in how labor markets for technical skills affect workers, firms, and regions)

careerbuilder®

LinkedIn

monster®

burningglass
CAREERS IN FOCUS

glassdoor

1. To add value quickly, try looking outside the data science teams.

The instinct is often to approach their data science teams.

Data science teams are often worried about:

- Bandwidth
- Limited access to hiring more data scientists
- Improving their data quality
- Being pulled in different directions by their clients PR, Product, etc.

Making another request of data science often adds work to their team. If you can insert yourself between data science and PR, it removes work from TWO teams (+ you can provide academic legitimacy + data science skills + social science training). Everyone's happy.

More broadly, data science is often not the locus of power. Especially, if you want to run experiments, it can often be effective to ally with someone in the organization who is powerful and overworked but needs ideas.

1. Pitch blanket access to the data. Refine the question later.

They bear costs by helping you. Your paper provides limited value to them unless they can learn something. At the same time, you have limited information about their data quality issues or their organizational priorities or KPI's or hot topics.

Pitching very specific research questions and negotiating NDA details on first contact is often counterproductive. The "research equivalent of talking about a pre-nup on the first date".

Bad strategy

- Propose a specific question that they can't possibly be interested in
- Trying to hammer out details of data disclosure

Better strategy

- Propose a broad set of questions which offer value, ask for blanket data access, refine the questions based on the data, and then get back to them about specifics of what you would like to do

3. Plan for academic red tape.

Their legal templates are often designed for IT contractors and onerous (e.g. may stipulate they own all “work product”)

Universities seem to have byzantine approval pathways

- But, usually requires legal approval from a team of university lawyers and the signature of a Dean (e.g. Dean of Faculty Research)
- This can take months and months; so set expectations accordingly

University clock speed is particularly costly due to frequent policy changes at the data provider.

In general, this process has required more extensive management of people and processes than I had anticipated.

4. Be cognizant of the costs of going onsite.

Advantages

- Experimental infrastructure
- Access to data
- Access to data scientists
- Unanticipated insights
- Upgrade human capital
- Lighter paperwork
- It's kind of fun

Disadvantages

- Surprisingly costly to be away from your home institution
- Productivity takes a hit
- "Risky revisions"
- Limited access to academic feedback

See if you can spend some weeks onsite and then VPN in using their hardware. Best of both worlds.

5. Build in significant time for “data forensics”.

There is a significant learning curve with systems and data. Find out if they have a data scientist onboarding process.

The first few months are often dedicated to forensic work – e.g.:

What are the tables? What are the primary keys? foreign keys?
Where is the data high quality? Where is it low quality?

Knowledge about information architecture is often distributed across many teams in different locations, and people constantly leave the firm

e.g. <This UML diagram authored by Scott, March 2006>

For archival projects, the data generating process for a particular data source can be surprisingly time-consuming to uncover, especially for hyper-growth firms

6. Invest in technical skills (but don't over-invest).

- Don't expect much help from their data scientists unless they are incentivized to publish – they are very busy!
- The learning curve is steep, and you don't want to burn up your social capital asking silly questions (what is a Join?)
- It is worth tooling up on SQL, Python, Apache PIG, Hive, Spark etc.
- Outsourcing this to a "technical RA" is hard. The forensic process requires domain knowledge, technical skills, and social science techniques (e.g. understanding what types of measurement error you can live with) that are hard to separate or codify
- You can do a lot of data cleaning, manipulation, and processing with a $4K server (i.e. about 48 GB of RAM), a 64 bit STATA license, and `strfun'. (And you can always sample the data)

7. Push for an archived version of the data that you can keep.

Given the project times in the social sciences, there is significant risk of having a project yanked midway, or having an R&R and be shut out of the data.

Their expectations of what you need in order to publish are different (WWW, KDD)

Especially in the tech industry, your contacts and key decision makers leave for other firms with alarming frequency

Data access policies can change overnight (e.g. after a high-visibility data breach, or a PR kerfuffle, or management turnover)

Try to get a copy of the data you can keep with you.

Using corporate big data for academic research can be profitable, but watch out for falling rocks.

To get data access
Look outside data science.
Pitch broadly, then refine.

To manage risk
Expect delays with paperwork.
Understand costs of going onsite.
Build in time for extensive data forensics.
Invest (but don't overinvest) in technical skills.
Push for an archived version of the data.

Other significant issues include reproducibility & data disclosure, as well as changing journalistic polices on data access and scraping

Thank you!

Email for follow-up questions:
Prasanna (Sonny) Tambe
ptambe@stern.nyu.edu