Data Selection and Procurement
(and a Little Big Data)

Carl F. Mela

1Department of Marketing
Fuqua School of Business, Duke University

2013 Workshop on Quantitative Marketing And Structural Economics

Hal Varian on Data (2009)

“I keep saying the sexy job in the next ten years will be statisticians. The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.”
Surely Google Can Help Me ...

Maybe There’s a Book on This ...
Determine The Data You Need
Start With the Research Question

Why use a structural model?
- Infer missing primitives (Nevo 2001)
- Test competing theories (Draganska, Klapper and Vilas-Boas 2010)
- Policy simulation (Duflo, Hanna and Ryan 2012)

Better data and/or better model?
- Chintagunta, Dube and Singh (2003) observe wholesale costs - instrument in demand estimation
- Fong, Simester and Anderson (2011) use price experiments to deal with price endogeneity
- Albuquerque and Bronnenberg (2009) micromoments to infer preference

Determine the Data You Need
Typology of Data

There are several sources of data:
- Proprietary firm data (free, flexible, slow)
  - e.g., I married for it (Mela, Gupta and Lehmann 1997)
  - My emphasis today
- Public data and records (free, less flexible, medium)
  - e.g., Study of Adolescent Health (Mele 2013)
  - scrape the Internet (Roos, Mela and Shachar 2013)
- Private data Nielsen; IMS and JD Power (expensive, rigid, fast)
  - e.g., Impact RX physician prescriptions (Narayanan and Manchanda 2009)
  - Grants or library can help with cost (TNS Media data)
Determine the Data You Need

Contents of Data

All data should contain:

- Dependent variables of interest you seek to explain
- Covariates that drive the effect of interest
- A set of instruments (natural experiment) that identifies the covariates of interest
  - Can all variables be endogenous?
- Sometimes these overlap

You also need institutional knowledge (e.g., assumptions, capacity constraints, contracts, etc.)

- Gordon and Hartmann (2013) use data from Nielsen Campaign Media Analysis Group
  - 4 calls to learn advertisers must offer lowest spot rate, though candidates can outbid at published rates
  - Understand the data and how they are collected!!

Determine the Data You Need

Proprietary Data Example

How does marketing affect social networking? What information do you need?

Which interactions available and of interest (DV)?

- Frequency, intensity, and direction of interactions, syntax of interactions
- What drives the social interactions (IV)?
  - Marketing, past behaviors, etc.
- Plausible instruments to identify effect of interest?
  - Wind instruments for blogging on friending (Shriver, Nair and Hofstetter 2013)

How much data can you handle (agents, periods, states, & variables)?

- Consider simulating ideal data set and model prior to searching for the data!
Determine the Data You Need

Proprietary Data Example

How does marketing affect social networking? What information do you need?

Which interactions available and of interest (DV)?
  ▶ Frequency, intensity, and direction of interactions, syntax of interactions
  ▶ What drives the social interactions (IV)?
    ▶ Marketing, past behaviors, etc.
  ▶ Plausible instruments to identify effect of interest?
    ▶ Wind instruments for blogging on friending (Shriver, Nair and Hofstetter 2013)

How much data can you handle (agents, periods, states, & variables)?
  ▶ Consider simulating ideal data set and model prior to searching for the data!
Determine the Data You Need

Public Data Example

Effort and Success of Studies Using Personal Data Collection versus Company-Supplied Data Bases

Gerard J. Tellis

<table>
<thead>
<tr>
<th>Paper</th>
<th>Topic</th>
<th>Time</th>
<th>Effort</th>
<th>Sources</th>
<th>Published in</th>
<th>Award</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golder &amp; Tellis</td>
<td>Pioneering Advantage</td>
<td>3 years</td>
<td>500 man-hours</td>
<td>Numerous magazines and 120 books</td>
<td>MktScience</td>
<td>William O'Coke</td>
<td>&gt; 400</td>
</tr>
<tr>
<td>Golder &amp; Tellis</td>
<td>Incumbent’s Curse</td>
<td>4 years</td>
<td>4 research assistants, numerous hours</td>
<td>Over 200 books, over 500 articles in periodicals</td>
<td>MIT</td>
<td>Harold Maynard</td>
<td>&gt; 300</td>
</tr>
<tr>
<td>Golder &amp; Tellis</td>
<td>New Product Takeoff</td>
<td>6 months</td>
<td>400 man-hours</td>
<td>Mechanics, publications, and books</td>
<td>MJASA</td>
<td>Frank M. Bass</td>
<td>&gt; 150</td>
</tr>
<tr>
<td>Golder &amp; Tellis</td>
<td>Success of Entry in China and India</td>
<td>Entrenchment over 9 years</td>
<td>10 people, numerous hours</td>
<td>McDougall, Leslie Nuth, Na-Tech, and others</td>
<td>FMASA</td>
<td>Paul Kos/AMAS</td>
<td>&gt; 60</td>
</tr>
</tbody>
</table>

References

Determine the Data You Need
Marked Data Examples

Goldfarb and Xiao (2011)

▶ Do managers think strategically about entry?
  ▶ telephone entry data $5000 (DV)
  ▶ census.gov and data.gov and dataferret for local market conditions (IV)
  ▶ Internet search (phone directories, who’s who, linked-in, etc.) for managerial experience (IV)
Find the Right Contact

Sources

- Corporate
- Personal
- Collegial
- Public

Find the Right Contact

Corporate - 1

Past colleagues from work (Yao and Mela 2009)
Consulting

- free consulting for free data
- old data from old consulting

Cold calls (they work for Catherine!)
Find the Right Contact

Corporate - 2

Relevant work leads to relevant data - get known!
▶ Advisory boards (Bronnenberg Kruger and Mela 2008)
▶ Industry Conferences and Seminars
  ▶ ARF, ART, IIR, DMEF, WOMMA, etc.
  ▶ Wilbur and DVR/set top data
  ▶ Kamakura, Fader ART Forum Tutorials
▶ Local lunches
▶ Publish in HBR & other managerial outlets
▶ Publish books and software
▶ PR your work

Find the Right Contact

Personal

Family (Yao and Mela 2010; Ansari and Mela 2003)

Friends (College buddies)
Find the Right Contact

Collegial
Library (e.g., TNS Ad Spending Data)

Colleagues (Hartmann and Nair 2009; Ansari et al. 2008)

Class Speakers (Bronnenberg, Dube and Mela 2010)

University Admin, Alumni and Students (Grubb and Osborne 2012; Bronnenberg Kim and Mela 2013)

Academic cooperatives
  ▶ MSI: corporate contacts
  ▶ WCAI: Loyalty programs, Hertz satisfaction and transactions, PG&E smart meters, 4Ps multi channel multi brand retailer, CPG likes and buys, Red Cross disaster donors, Sirius acquisition and retention, Stub Hub relationship marketing, Expedia hotel search, etc.

---

Find the Right Contact

Public

Data.gov or Census.gov

Nielsen, IMS and JD Power

Scrape the Internet (e.g., Python, SiteScraper)
Make the Pitch

What does the firm gain (incentive compatible)?

- What decisions will they make better?
- What is impact on profits or revenue?
- How will your contact get promoted for helping you?

What does it cost the firm?

- Dollar and time costs
- Do your own data cleaning if it’s easiest for firm
- “Executives rank providing data to academics right there with root canals and colonoscopies. “Have cool tool” is not effective”
  - W. Kamakura

Make the Pitch

Peter Rossi’s advice: “Go to the absolute top people in the org. It is almost impossible to get data from mid level people. Then get down on your knees and beg.”

Your pitch should be concise

- What will the firm gain, what will it cost, what is the time line, and why do they need you?
- Most do not care at all about publications

Role play: persuade Google exec to provide you data on key word search ...
Negotiate the Non Disclosure Agreement (NDA) I Firm

The firm’s goal is to gain actionable insights not shared by other firms nor interpreted by analysts

The firm also seeks to keep its clients confidential

Your goal is to publish

Disguising data and/or aggregating results achieves these aims

▶ Be careful - in IRI data set some markets had only a few mass merch – even if disguised one could infer

▶ Netflix prize cancellation

Trust is critical
Negotiate the Non Disclosure Agreement (NDA) I
Firm

The firm’s goal is to gain actionable insights not shared by other firms nor interpreted by analysts.

The firm also seeks to keep its clients confidential.

Your goal is to publish.

Disguising data and/or aggregating results achieves these aims:

- Be careful - in IRI data set some markets had only a few mass merch – even if disguised one could infer
- Netflix prize cancellation

Trust is critical.

Negotiating the NDA II
University

The University’s goal is not to be sued.

Your goal is not to be sued. Use university lawyers!

Email your faculty; several examples they’ll redact.

Check with human subjects as well (Institutional Review Board) for schools and governments regulations.
Negotiating the NDA II

University

The University’s goal is not to be sued

Your goal is not to be sued. Use university lawyers!

Email your faculty; several examples they’ll redact

Check with human subjects as well (Institutional Review Board) for schools and governments regulations

Data Delivery

Large USB drives work great.

“Get physically in the building with an external hard drive having made an appointment with the IT guy and stand over him while he downloads the data.” - Catherine Tucker

Use metadata and dictionaries prior to entire data set!
Data Checking

Check variable names and columns
  ▶ WTF is “user_action_id skc_url”?  

Are the variable values feasible?
  ▶ Frequencies, means, regression, correlations
  ▶ Which variable let you address research question?

Are the fields well populated?

Pivot variables (relational databases e.g., Sprint)

Do dictionaries make complete sense?

Project Management - I

Early Phase
  ▶ After data checking, follow up quickly
  ▶ Some data collected going forward after detailing needs
    ▶ Yao and Mela (2010)

Mid Phase
  ▶ Bimonthly or monthly phone or email updates
Project Management - II

Completion Phase

▶ Phone or on site conference
▶ Contact ensures all relevant players are there
  ▶ Forget the method (beyond concept)
  ▶ Focus on what they should do differently and profit/sales outcomes

Implementation Phase (Misra and Nair 2011)

▶ Frequent meetings, institutional constraints, and measurement

Sharing Data

Sharing makes us all better off

Most NDAs will not allow you to share

Some firms insist data returned when project done

Negotiate to share data after finite period
  ▶ Old data not so relevant to firm
Big Data

What Is It?

What is Big Data (Wikipedia)?

- Moving target: Don Lehmann inverted matrices by pen
- “Capture, curation, storage, search, sharing, transfer, analysis, and visualization.”

What is the point of Big Data?

- Determine needs and collect data as noted above
- Why archive everything “just in case?”

Big Data

Data Structure Issues (Too Much and Too Little)

Often a Relational Database Structure e.g. SQL

Complete but sparse

- Low response rates and rare events
  - Non-parametrics for rare events (Naik and Tsai 2001)
  - Data fusion for missing data (Gilula et al. 2006; 2011)

Large N and Large P

- 100s of covariates (P)
  - Space reduction methods (Naik and Tsai 2004); variable selection (Fan and Li 2001), etc.
- Many millions of observations (N)
  - Sampling with macromoments/machine learning?
Big Data

Other Issues

Data Visualization (even for small $n$ and $p$; Mela et al. 2013)

- Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky. “A tour through the visualization zoo”.

Algorithms and Computation (Miranda and Fackler 2002; Judd 1998)

- Efficiency in computation
- Use efficiency for more data or more complexity?

Big Data

Quoted Advice from the Gurus

- Start with an existing capability for data management and analytics
- Add some unstructured, large-volume data
- Throw some product/service innovation into the mix
- Add a dash of Hadoop
- Cook up some data in a high-heat convection oven
- Embed this dish into a well-balanced meal of processes and systems
- Promote the chef to Chief Analytics Office

http://digitalcommunity.mit.edu/community/featured_content/big-data/blog/2015/07/14/tm=davenport-reports-on-big-data-in-big-companies
In Conclusion

Data
- Types of data (proprietary, public, and for sale) and how to get it (steps on right)
- Big data becomes data - so see above

Acknowledgments
- Thanks for your interest
- Thanks to Peter Fader, Avi Goldfarb, Mike Hanssens, John Hauser, Wagner Kamakura, Peter Rossi, Gerry Tellis, Catherine Tucker for ideas and information
- Thanks to Brett Gordon, Rick Staelin, and Raph Thomadsen for the conference
- Thanks to all the firms who coughed up data for me