Data Selection and Procurement

Carl F. Mela

November 16, 2010

1T. Austin Finch Foundation Professor of Business Administration, Fuqua School of Business Administration, Duke University, 100 Fuqua Drive, Durham, NC 27708-0120. Email: mela@duke.edu. Phone: 919-660-7767. The author would like to thank Peter Fader, Avi Goldfarb, Mike Hanssens, John Hauser, Wagner Kamakura, Sanjog Misra, Andres Musalem, Peter Rossi, Gerard Tellis, Catherine Tucker and Ken Wilbur for helping to shape the ideas in this paper. I would also like to thank the conference organizers, Brett Gordon, Wes Hartmann, Rick Staelin, and Raphael Thomadsen for their feedback and to the attendees of the 2010 Structural Modeling Workshop at Duke University.
Abstract: Data Selection and Procurement

In this note I overview the data selection and procurement process for structural models. Data selection for structural models presents unique challenges because it is imperative to consider the information that identifies causal effects of interest and because the institutional context has substantial ramification for the structure of the model.

I discuss three types of field data upon which to build empirical models; data that are proprietary to firms and public data that can come from the public domain or be purchased from private research firms and the benefits and limits of each. I then detail a process for obtaining proprietary data and the potential pitfalls inherent in the process.
“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.” Varian (2009)

1 Overview

Equipped with nothing but a camera and a BLP model in hand, Thomadsen (2005) photographed fast food menu pricing and inferred how prices affect demand in geographically differentiated industries; even in the absence of demand data. This example illustrate the interplay between structural models and the data used to estimate them. While structure is not an ideal substitute for data, structural models require thoughtful consideration of the phenomenon modeled and how the information used to estimate the model is sufficient to uncover the theoretical parameters of interest.

In this article, I will outline the data selection and procurement process for structural models. This consists of six steps. First, one must determine the ideal data to address the research problem. Equipped with the ideal hypothetical data, the second step involves finding the right source for that data. To the degree this involves proprietary data sources, the third step is pitching the research topic to agents who have the data the research needs. Assuming the data source agrees, the fourth step involves negotiating the appropriate disclosure agreement; without permission to publish, years of effort can be wasted. The fifth step involves the data transfer and checking. Last, should the data be complete, the sixth step involves keeping the communication lines open with the firm while the research project proceeds and reporting/implementing the results when done. Though each step is not unique to structural models, its careful emphasis on theory and causality require additional consideration when it comes to specifying data. Structure forces researchers to consider which aspects of the research problem are evident in the data and which require structure or additional information to impute; this tradeoff often guides the data choices. In the next section, I address
these concerns and other aspects of specifying the appropriate data.

2 Determine the Data You Need

Determining the data you need involves two steps; choosing an impactful topic to guide the choice of data and ensuring the data that is chosen can address the issue of interest. The research question is primal. Without a well defined research question, data selection and procurement become moot. It is beyond the scope of this paper to address what constitutes an interesting research question; an excellent resource to address that issue is Varian (1997). Suffice to say, novel topics typically require novel data. Accordingly, there exists a strong incentive to pursue novel data sources and refrain from limiting one’s research by relying upon data that are readily accessible, amplifying the data hurdles in structural modeling research.

2.1 Data Components

Data contain four components; i) a dependent variable of interest one seeks to explain (the interesting research problem); ii) covariates that drive the effect of interest (treatment effects; or the marketing decisions available to the firm); iii) a set of instruments (or natural experiment) that identifies the effects of the covariates of interest and iv) a set of institutional factors that underpin the data generating process. Regarding points i) - iii), it is important that the treatment effects are randomized to the outcomes. When this is the case, the covariates and instruments are the same. However, in many instances, this is not the case.

In an illustrative example, Hofstetter et al. (2010) consider the effect of content generation on social ties on a windsurfing website. Website users who create content might receive more friend invitations. To the extent firms can induce its users to post (especially if friending also induces posting), the interaction can enhance website engagement. Yet unobservable factors can influence both outcomes making it hard to assess the causal relationship between
the two; for example, friending and posting might both increase with Internet penetration. Accordingly, Hofstetter et al. (2010) collect exogenous instruments that vary with posting, but not friending. In particular, they consider wind speed, reasoning that an increase in wind will lead to more posting, but that its effect on friending is negligible because most friends are established in an online environment. The instrument is ideal in many regards, because it is hard to argue wind is endogenous (unless one has Aeolus as a friend).

In addition to the data itself, structural models require a delineation of the agents, their objective functions (e.g., maximize revenues or profits), their information states (e.g., how much they know about competitors decisions), and the rules of the game (e.g., are decisions made jointly or sequentially). Organizations involved in the game are an integral source of contextual data. Counterfactual analyses regarding strategy, for example, are more persuasive when one can document the agent is actually in an off equilibrium state. Working with a firm to obtain data is sometimes necessary to the appropriate specification of these models. Whenever possible, it is sensible to contact agents in the industry prior to modeling it. While this insight is not data, per se, the institutional data is no less critical for structural model estimation. Consider Gordon and Hartmann (2010), who explore the role of political advertising. Prevailing advertising rates are readily available from Nielsen’s Campaign Media Analysis Group (CMAG). However, instead of taking these rates as given the authors interviewed CMAG four times, eventually clarifying rates with its president. They learned advertisers are obligated to charge lowest advertising spot rates for campaigns, and these rates often deviate from published market rates. Moreover, they learned that market rates are more prevalent for large campaigns. The key point is that, without consulting the agents in the game, they could have used the wrong rates and made the wrong inferences regarding advertising campaigns in political campaigns.

A final consideration is that structure is often a poor substitute for data, meaning that it becomes especially desirable to obtain as many primitives from the data as are feasible to improve the reliability of the other primitives to be estimated or inferences regarding
conduct; for example, there is little need to estimate a firm’s marginal costs as in Nevo (2001) if one can obtain data about them. Only when these data are infeasible to obtain does it make sense to use structure to infer them. An added benefit of better data is often that fewer assumptions need be invoked.¹

2.2 Data Types

With a research question of interest in hand, and a sense of what exogenous variables might identify the effects of interest, several data sources exist to address the research question, including i) firm proprietary data, ii) free public data, and iii) commercially available market research data.

Proprietary data are typically sourced directly from firms. Because these data are developed in conjunction with a firm to address a research problem, they are free, flexible, and often tailored directly to the research problem. Moreover, because firms are engaged with customers, proprietary data are often emblematic of important research problems. Further, by working directly with an agent involved in decision making, one can often obtain the necessary insights into the rules of the game, information states and agent objective functions needed to properly specify the structural model. However, the time to procure such data are often lengthy to collect because of the resources a firm needs to commit to obtain them. An example of these type of data is Yao and Mela (2008), who worked with an auction house to obtain data on the bidding history of bidders and the listing history of sellers. The research explored how auction house pricing affected revenue. By obtaining data directly from the firm, they were able to observe all bidders’ bids across all auctions; in contrast, auction

¹Sometimes data used in model estimation can augmented with supplementary information to validate model assumptions. For instance, Wilbur (2008) infers television show advertisements estimated from a structural model of two-sided networks in the television industry. He then validates the model by taping a sample of television shows to ensure the imputed advertisements are consistent with the estimated advertisements – and finds the correlation to be about 90%. Duan and Mela (2009) develop a supply side model of apartment rental pricing to infer the marginal costs of apartment units. Using self-reported marginal costs of $204 obtained from a survey of apartment managers, the authors find their estimated costs of $233 to be quite close. Musalem et al. (2010) collect data on out of stocks to show that models ignoring this problem generate downwardly biased estimates of demand (i.e., models incorrectly assume zero sales reflects lack of demand, not availability).
papers often collect data by parsing the website for a single auction, meaning that only the leading bids are observed. Hence, Yao and Mela (2008) could obtain a more complete sense of the distribution of customer valuations and control for unobserved heterogeneity from the repeated observations.

Public data, like websites, are free and can often be collected more expediently than proprietary data. Yet these data are harder to customize to the research application. Illustrative of public domain data research, Forman et al. (2009) explore competition between local and electronic markets by collecting public records pertaining to local bookstore openings and scraping Internet web sites to determine Amazon book sales in that region. Another example is afforded by Kim et al. (2009), who explore how consumers search across alternatives. Ideally, this would involve collecting the specific set of goods searched by a consumer. However, these data are often not available from Internet retailers owing to privacy concerns. So Kim et al. (2009) parsed the Amazon.com website to scrape the set of alternative products viewed by the collection of visitors to a particular product page as well as the rank ordering of the viewership overlap. Together with the site’s rules for determining these ordering, the authors were able to obtain estimates of the precise number of joint visits to competing products’ pages. This enabled them to form inferences about search. Facilitating the task of scraping Internet data, tools such as Mozenda (www.Mozenda.com) have become commercially available. Government websites are another major source of publicly available data; Du and Kamakura (2008) develop a model of consumers seeking to maximize their utility over their lifecycle by selecting various goods ranging from education to housing. Collecting such a long stream of data would be a formidable task, but they were able to exploit the National Bureau of Economic Research (NBER) Consumer Expenditure Survey which spans 22 years. Archival research of public data can often be time consuming to the extent it involve finding and integrating data sources. Dataferret (dataferret.census.gov), is one tool for navigating and extracting U.S. census data and similar products exist for other public and international data; Swedish consumer prices and household expenditures across a wide range
of categories are available at www.scb.se. Of course, not all records are available from the Internet making data collection more challenging. Bronnenberg et al. (2009) explore order of entry effects in spatial market shares; in order to obtain product entry timing, the authors had to visit the National Museum of American History in Washington DC.

Finally, there exist a litany of commercial market research firms who specialize in collecting data and selling this information to marketers and researchers. These include Impact RX in pharmaceuticals, Nielsen and IRI for Internet and shopping panels, JD Power for automotive, NPD for point of sale data on technical goods and so forth. These commercial data are immediate to obtain and often expensive, though grants often defray the cost. Sometimes, the university library will purchase the data so it will be available to all researchers; for example, Duke subscribes to TNS’s Ad Spender data, and both economists and marketers have used the data. Another advantage of commercial market research data is that they are quick to obtain and are formatted in a fashion that makes them easy to use. A key limitation is that they are limited in the sense that there is no flexibility in the information they collect. Illustrative of this style of research, Dong et al. (2009) explore the effect of a firm’s strategic targeting of detailing on estimates of physician response to detailing. Using data from Impact RX, they show that failure to account for endogeneity arising from the supply side leads to an underestimation of the benefits of targeting. Dube et al. (2010) obtained data from NPD Techworld regarding the price and sales of video game consoles as well as game availability. This enabled them to infer the role of network effects in how a market tips to a standard. Goldfarb and Xiao (2008) integrate two types of data sources, commercial researcher data and public data; to explore the role of managerial ability to iterate to equilibrium in an entry game, they purchased entry data from a reseller, used government data to ascertain local entry conditions, and an Internet search to ascertain managerial experience of the considered firms.

In the remainder of the paper, I will focus upon a systematic approach to data acquisition for proprietary data. Acquisition of public data (from commercial market research firms and
public sources) is more formulaic though it can sometimes be time consuming. Because proprietary data involves the consent of a counterparty, there are additional considerations in obtaining data that I address next.

3 Find the Right Contact

Partnering with a firm often involves working with a visionary and prominent agent within the firm. Individuals with the willingness to take risks on academic projects, allocate time to the endeavor and have the initiative to sell the potential to the organization are rare. Several domains exist in which to find such research partners; these include corporate engagements, personal interactions, collegial (academic) connections.

Corporate contacts involve interactions made in the process of professional engagements. For example, consulting relationships can lead to access to data. Though rarely will firms allow consulting data from a current project to be used for research, they will often allow academics to use the data after a period of a few years or guide researchers to other data in the firm. Similarly, advisory board positions can be used to obtain data. One of my key reasons for participating on the analytic advisory board of IRI was to develop an academic data set (see Bronnenberg et al. (2008)). Consulting and advisory board opportunities are a function of the relevance of the work; suggesting the potential for a feedback loop between data and research; more relevant data lead to more relevant research which, in turn, yields more data. The key theme is to work on inculcating corporate contacts through active research. Former work colleagues are often instrumental in supporting research. The data for Yao and Mela (2008) arose through my interactions with a former colleague.

In addition to working on applied problems, promoting relevant research plays a role in facilitating corporate interactions. Conferences such as the Advertising Research Foundation (ARF) annual conference, the American Marketing Association Advanced Research Techniques Forum, the Institute of International Research conferences, the Direct Marketing
Education Foundation annual meeting and the Word of Mouth Marketing Association annual conference all represent singular opportunities to be involved. Kempe and Wilbur (2010) use cable TV set top box viewership data from a major cable company that was obtained after the second author presented related research at the ARF. Another colleague of mine, Wagner Kamakura presents a tutorial at the ART Forum, which has opened up a number of data opportunities. Organizing research presentations at and with local companies is another approach that is quick and can yield high dividends. Another means for increasing the managerial exposure of one’s research is to publish a summary of the work in a managerial outlet such as the *Harvard Business Review*. Writing books or creating software for general distribution also opens doors. The key point is that your research must be highly relevant and must be strongly promoted to open doors for new data.

It perhaps goes without saying that family or college contacts can be instrumental in research. In my case, one example of a paper that materialized through a friend of a family connection is Yao and Mela (2010), who develop a dynamic structural model of advertiser bidding and consumer search in the context of key word search.

The Academy also facilitates research connections on a number of levels. One can invite speakers to class that serve to both enhance the relevance of the class and also open discussions for new data. An illustration of this approach is Bronnenberg et al. (2010), who contrast the purchase behavior of households who received a DVR to those who did not and find no effect of DVRs on the sales of advertised goods. The idea for the research was conceived in conjunction with a visit from IRI and TiVo to my MBA class and the data followed from additional discussions with the speakers. Second, alumni and students can also source data and provide research insights. Chintagunta et al. (2010) assess the effect of user reviews on sales; the data for this project came from a student of the third author. Along these lines, it is helpful to read the resumes of students and to learn to use the alumni database systems; many will have deep connections into firms that capture data to address research problems of interest to you. Third, the University itself sometimes has
data. Grubb and Osborne (2010) model plan choice and usage in the cell phone industry wherein consumers learn about their usage needs with time. A distinguishing feature of their data that enable them to identify learning effects are the call detail records at the user level from the adoption of a plan. These were obtained from a university that proffered these plans. Fourth, colleagues can also be a rich source of data. Hartmann and Nair (2007), in a dynamic model of demand, consider the problem of tied goods. The model is calibrated using razors and blades; they augmented their original data with advance data from the IRI data set provided by Bronnenberg et al. (2008). We had to obtain special permission to provide it, but many colleagues will make a concerted effort to help peers in this regard. Fifth, more organized exemplars of this cooperative collegial data effort include the Marketing Science Institute (MSI) (www.msi.org), the Wharton Interactive Media Initiative (WIMI) (http://www.whartoninteractive.com) and the University of Chicago Booth School’s Kilts Center for Marketing (http://research.chicagobooth.edu/marketing/index.aspx). MSI has a roster of blue chip marketing firms and will make introductions to facilitate data acquisition and research grants. WIMI has been especially active in generating interesting data; recently they have partnered with Expedia to offer data from the hospitality industry. Iyengar et al. (2007) consider a structural model of consumer learning about service quality; the research stems from data that were obtained from the Teradata Center for Customer Relationship Management at Duke. Sixth, several journal now have databases available on line. A good example is the online database website at Marketing Science (http://www.informs.org/Journal/MktSci/Online-Databases).

Finally, it is worth considering the level of the organization with which to work. Ideally, it is sufficiently senior that there is budgetary discretion to authorize the research and obtain approval for the data, but not so senior that the project is of modest relevance.
4 Make the Pitch

Once one has established contact with a firm whose data are aligned with the research proposition, the next step is to pitch the project. Two points are critical. First, the pitch should address what the firm is to gain from the work. Second, it should address the costs.

Regarding the former, one needs to address specifically how the research will affect which decisions made by whom and the resulting impact on profit or revenues. You should also indicate why the firm or contact needs you. Another way of thinking about these questions is whether the work will help the contact in his/her annual review and frame your pitch from the perspective of why it would help your contact and the firm. The firm cares only about whether your work is an asset to them, else, they have no incentive to provide data. If the research considers pricing, for example, it is imperative to consider how prices are currently set, how and why you would set them differently, and the implications for the firm’s revenues. Yao and Mela (2008) consider pricing in online auctions. One approach to pricing might be to regress demand against price and use this demand curve to establish the optimal price. Aside from endogeneity concerns, however, this approach is not feasible in online auctions because price changes are infrequent and the growth rates are so substantial that older data points are not so relevant. The authors instead imputed sellers’ reservation values for listing based on past listing behavior and used these reservation values to generate a demand curve. Thus, the research enabled the firm to consider how to price in the absence of a regression based approach and was of key interest to the firm in setting pricing policy. The bottom line, of course, is that the firm cares little about publications, sophisticated modeling or tenure aspirations. They would like to know the return on their investment from sponsoring your work, so this had better be stated clearly in the pitch.

Costs are integral to firms’ willingness to share data. It is sensible to determine the most “raw” form of the data that can be pulled from a firm’s servers with the minimal level of effort on the part of the firm, and then do the data cleaning and aggregating one’s self. Likewise, to the extent you can develop your code and analysis in a way that a firm can easily port
your analysis to other years or products, they are more likely to embrace the research.

5 Negotiate the NDA

Assuming one is sufficiently fortunate that the firm has promised to send the ideal data set one’s way, the firm will often ask you to sign a non-disclosure agreement (NDA). The purpose of this agreement is to protect the firm from potentially adverse outcomes. This might occur, for example, if you were to forecast the firm’s growth is to slow and the analysts reading this information begin to issue sell recommendations against the firm. Firms can also be concerned about the release of private information that may advantage a competitor or create concerns with regulators (such as their cost data). Alternatively, one might inadvertently reveal sale force compensations leading to discontentment on the part of those sales agents who fell undercompensated. An example of the problems firms face when sharing data is afforded by Netflix, who sought to leverage the academic community by providing disguised account level rental data to researchers; the goal was to determine which movies should be recommend to whom. The research resulting from the $1 million prize improved the Netflix’s recommendation and forecasting systems and generated considerable positive press. However, Narayanan and Shmatikov (2008) integrated the Netflix data with the Internet Movie Database (www.IMDB.com) and imputed the identities of the disguised panelists as well as their rental histories. Netflix was subsequently sued by the Federal Trade Commission and faced a class action lawsuit; they also canceled their sequel prize competition.

The firm has little interest in seeing your work published given their primary focus on revenue generation and fear that competitors might benefit as much or more from your work. Hence, they will seek to protect all data, often asking you to sign the same NDA as its consultants. Were you to do this and still publish your work, you can be sued for breach of contract and damages. To add insult to injury, the agreements will often ask you to pay for the legal costs. It is preferable to abort the research project than sign an NDA that
precludes publication or mandates sponsor review prior to publication. Two years into your
research is too late to discover you are not allowed to publish you work; that time can not
be recovered. Another standard practice is for firms to indicate that any work you do with
them becomes there property. This means you can not ever use the tools you develop in the
future with other organizations.

Ideally, you would like no restrictions whatsoever on what you publish. In practice,
this is rarely amenable and you can agree to disguise the firm’s name and/or aggregate
the data (meaning you can publish aggregate statistics or parameter value, but not release
individual-level observations). Subject to these caveats, most firms will be amenable to
signing an agreement.

If the firm amends the NDA to your liking, it is prudent to have the university lawyers
review the document before you sign it. They are often experienced in this domain. An
email to your faculty asking for prototypes of NDAs is also a sensible exercise because others
have worked through this issue. Note that an NDA protects not only the firm, but your right
to publish. Hence, it is inadvisable to proceed without one. You should also check with your
human subjects committee; though not experimental data, if you obtain person specific data
that are not commercially available, it is sometimes necessary to obtain university approval
for these data.

It is also worth consideration to add a clause regarding your ability to share data with
others. In a cooperative game wherein all researchers share data, all researchers are better
off. Unfortunately, most firms will not allow this. Hence, be creative in terms of the NDA.
For example, it might be amenable to the firm to share aggregated data, older data, or
certain components of the data. One can make the case to the firm that the returns to
sharing data are amplified when others work on it.

On a final note, reputational effects become paramount. Section 3 details approaches to
increase interactions with the management community. With time and with applied research,
one can develop a reputation for utility and discretion that can facilitate the NDA process
6 Data Delivery and Data Checking

Though a signed NDA is a key hurdle, additional pitfalls await in the data hand off. Generally, there is a short window over which data are collected and transferred. A bad time to learn about incomplete data is two years into the research when reviewers ask for a robustness check and you find a relevant field has not been populated. It is not uncommon at that stage for your research contact to have moved to another position or for the firm’s information system to render the possibility of supplementing the data impossible.

Data are often massive, so it is wise to start with a smaller set of metadata (for example, first receive only a few households of data over a short period). This can be used to check basic statistics, the degree of missing observations, and odd values in the data. Simple regressions and correlations can be used to consider the relationships between key variables of interest to ensure there is sufficient information to explore these. Check to see if the variables in the data let you address the considered questions, as noted in Section 2.1. Outliers observed at this stage can be informative; when outliers exist they should be explained. In grocery data, for example, I have observed unusually high weekly sales as a result of institutional, non-consumer purchases. Such purchases should not be included in a model of consumer demand.

It is also important to understand the data structure completely at this phase; a complete data dictionary should list, for each file, the variable definition the unit of observation and, ideally, the variables that intersect data files and enable you to join them. The Teradata cell phone data set (Iyengar et al. (2007)) included djsatge files for call detail records, stores, handsets, demographics, promotional targeting and billing statements. It is imperative to note, for example, that there is a handset indicator in the call detail records than enables one to splice the calls made with the handsets owned. There are additional look up tables that
link numeric codes for promotions to the details of the promotions. These relational data structures can be enormously complex and should be completely mapped out. Likewise, obtain definitions for each variable. Would you know, for example, what that variable “user_behavior_id_skc_url” means?

Assuming the data are in great shape and you are ready to consider the data transfer, the next step is to determine how much data can you handle. Yao and Mela (2008) focused on coin auctions because there was relatively little cross bidding behavior with other categories, especially for antique coins. While it was conceivable to obtain information beyond coins, the total data would have been difficult to store and process. Hence, it was not feasible to obtain information regarding all auctioned goods. Once the scope of the data are defined, data transfer is becoming increasingly easy as the price of external hard drives falls. A company can easily record a terabyte of data onto a hard drive and ship it; assuming one has the processing power to analyze it. Sometimes it can be sensible to visit the company and meet with the IT department to help expedite the process.

7 Project Management

Academics move at a glacial pace. In contrast, businesses are driven by quarterly returns. This mismatch in expectations creates problems if mismanaged. There exist four project phases to consider; initiating the project, the analysis phase, project completion and project implementation.

In the earliest phase, it is imperative to re-check the full data to ensure it matches the meta-data. Often, data are collected going forward; this was the case with Yao and Mela (2010) where the sponsoring firm collected a couple of months of log file data around the project’s data requirements. This is an active phase of collaboration where the data suggest more questions about the institutional context of the project and possibly the need for more data. In the middle phase, model estimation, validation and paper writing are ongoing.
In this phase, checking in with your sponsoring firm every couple of months is especially important to ensure continued interest of the firm.

Once the project is complete, an onsite or phone conference is required to a) ensure insights from the project are realized and b) to present research to the firm to ensure the project is not missing any key institutional details. In conjunction with your primary contact, it is desirable to invite as many persons as possible from the firm; this will maximize feedback and possibly open the door for new projects.

The presentation should (unless to a modeling group in the firm) remain conceptual with few equations. I recall one example where I computed derivatives on a profit function to recommend an optimal price; the firm to whom I presented the analysis thought I was out of touch. On the next occasion, I simply plotted the profit surface and showed its maximum and the response was far more positive. In addition to overviewing your project in readily accessible terms, be clear about what the firm should do differently as a result of your research. This affords an opportunity to learn more about implementation concerns.

Finally, should this meeting go well, the occasion sometimes presents itself to implement the research. A superlative example of this potential is presented in Misra and Nair (2009) who study salesforce compensation plans. Using an agency theoretic model to underpin a dynamic structural model, Misra and Nair (2009) are able to infer sales person effort functions. Conditional upon these functions, they explore counterfactual compensation schemes to improve firm profits. Interactions with the firm quickly ruled out some that were infeasible, but the authors persuaded the firm to try an alternative approach. Early results suggest their research resulted in a $1,000,000 per month (9%) improvement to the firm’s bottom line. The paper exhibits the potential for structural models and data to interact in a fashion that can generate impressive improvements in business practice.
8 Conclusion

Structural models, like all empirical research, are predicated upon finding the right data. In many regards, the hurdles for this task are more challenging in the context of structural models because of its emphasis on causal attribution. Hence, in addition to an outcome and some variables that drive that outcome, it is imperative to carefully consider i) a set of instruments to help make causal attributions and ii) institutional details about the agents, rules of the game and information states.

In this paper, therefore, I detailed a process for data selection and procurement by advancing the desiderata of good data and an approach for obtaining it. Though publicly available data, and data from market research firms are often sufficient, the paths to obtain these data are more incumbent upon researcher initiative than that of a private market partner. Hence, I outlined a series of steps in obtaining proprietary data, including making a contact with the firm, pitching the project, negotiating an NDA, transferring the data, managing the project and reporting the results. It is my hope these tips will be useful to students and others seeking to estimate structural models. Finally, it is also my hope that, to the extent these endeavors are successful, researchers will take steps to share these data in a collective effort to advance the field.
References

Bronnenberg, Bart J., Jean-Pierre H. Dube, Carl F. Mela. 2010. Do DVRs Influence Sales? 
*Journal of Marketing Research* **47**(6) 1–.


