Name/Link	Pros	Cons	Type	Notes/Description
Amadeus	Includes five years of exportable financial statement data as well as basic company and management information. Amadeus can be used to screen for companies by numerous criteria such as geography and industry. Advanced reporting features allow the user to create customized reports for one or more companies.		General Corporate	A database of public and private companies in European countries, including much of Eastern Europe.
BEA Economic Areas	Appropriate level of geography for economic activity, novel (vs states), relatively fine grained	Few ready control variables, often have to construct your own using county data aggregated up	General Corporate	
BEA Input-output tables	Determine related industries and how related	Requires some heavy lifting at the 4-digit SIC level	General Corporate	
BEA data on US Investment Abroad by Country/Industry			General Corporate	The International Investment Division of the Bureau of Economic Analysis publishes annual data on aggregated US direct investment abroad and foreign owned FDI in the US. These publications are based on surveys that US parent companies are required by law to report to the BEA and include both financial and operating activities of US MNCs. This is an annual publication, with more information collected every five years in what are called benchmark surveys (1982, 1989, 1994, 1999, 2004). There is a lot of data at the industry and country level of analysis, including things like investment amounts, R&D expenditures, total sales, employees, royalties, etc.
BEA data on US Investment Abroad by Firm/Subsidiary	These data contain the most comprehensive information on the foreign operations of US multinationals.	Can only be accessed at the BEA in Washington DC. There is no off-site access. The data themselves have not been organized in a panel and each researcher who is approved to	General Corporate	The International Investment and Trade in Services Survey Act requires U.S. MNCs to report detailed information on the financial and operating activities of both U.S. parent companies and their foreign affiliates, as well as information on the value of transactions between the parents and affiliates. The BEA has collected these data on

		access these data must spend a fair amount of time organizing the yearly surveys into a usable panel. Must be a US citizen to even apply to access these data.		an annual basis since the early 1980s. These data are made available to academic researchers if submitted proposals are approved.
Census of Manufacturers data from the U.S. Census Bureau	best source for actual production numbers	hard to get access and not as complete as one would hope	General Corporate	Basic data obtained for all establishments include kind of business, geographic location, type of ownership, total revenue, annual and first quarter payroll, and employees in the pay period including March 12. Establishments receiving a long form provide added detail and added data (including inventories, capital expenditures, identification of some 1,000 materials consumed, cost of materials, energy consumed, and quantity and value of shipments for some 11,000 products). Short forms request much less data detail and no identification of materials consumed.
Chinese listed firm data from WIND	The WIND data is generally good quality.	We have found odd cases where is contains contradictory entries to other China data sets (such as CSMAR) and have had to go back to company reports to check the numbers in the original by hand.	General Corporate	
COMPUSTAT	Broad and longitudinal with many accounting variables, good industry level and firm level controls; comprehensive financials; Both databases provide wide, longitudinal coverage of the performance of major companies. Worldscope used to have better international coverage, but the gap appears to have closed recently. Both databases provide large samples in standard formats, with good documentation and support.	poor user interface; I'm often frustrated by lack of data on private firms (not sure how to fix this problem though); but limited to public firms and many records are unusable; The weaknesses of the databases are well known: public companies only, subject to the vagaries of and changes in public accounting standards, subject to the weaknesses of the standard industry classification schemes, etc.	General Corporate	Description: Standard & Poor's Compustat, Global Vantage, Research Insight, and Emerging Markets Database (EMDB). Standard & Poor's Compustat (U.S.) and Global Vantage (international) databases contain fundamental financial and price data for both active and inactive publicly traded companies. Compustat goes back annually to 1950, and Global Vantage goes back to 1993. Standard & Poor's EMDB (Emerging Markets Database) includes market performance, indicators, market capitalization, monthly value traded, local stock market price indexes, and prices, etc., in local currency and U.S. dollars, back to 1995. See: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=871169 for a comparison with WorldScope. -.There may be an updated critique, but McElreath, Robert B., and C. Donald Wiggins. "Using the Compustat Tapes in

				Financial Research: Problems and Solutions." *Financial Analysts Journal* 40, no. 1 (1984): 71-76 cover the strengths and weaknesses in depth.
CRSP			General Corporate	Security Prices for all stocks traded on all major stock exchanges
Danish Census data				
Directory of Corporate Affiliates	Most comprehensive information on corporate hierarchies and ownership There are both hard copy versions and an electronic version that is available through university libraries. I know the hard copy annual versions have private firms. The electronic version goes back to 1993 I believe and it has the CUSIPs and CNUMs, so it is actually quite easy to link this up with Compustat now.	- Expensive - Data is sold for the most current year only. Thus, unless data is gathered year by year, it can be difficult to track ownership changes over time. (CONTESTED, SEE STRENGTHS) - Also, data is by firm name and an identifier unique to the directory, so matching to other datasets has to be by name (CONTESTED, SEE STRENGTHS).	General Corporate	subsidiaries/ divisions owned by a large set of firms
Dun & Bradstreet Reference Book			General Corporate	
Euromonitor reports	The Euromonitor reports are helpful introductions to the strategic issues faced by firms in different industries/countries, good for background information.		General Corporate	major firms in consumer products manufacture, and of a particular industry in a given country; Market research reports about various consumer products and industries worldwide. Includes economic indicators, demographic data, foreign trade statistics, and sources of information.
Edgar	Contains all firm SEC filings in electronic format. These SEC filings contain actual contracts for these firms, including license, alliance and other collaborative contracts.	To find contracts, one is looking for a needle in the haystack. You need to have some software or macros for searching for desired agreements.	General Corporate	
Eventus	easy-to-use program to calculate stock price reaction to announcements of events		General Corporate	Eventus performs event studies using data read directly from CRSP stock databases or pre-extracted from any source. The Eventus system includes utility programs to convert calendar dates to CRSP trading day numbers, convert CUSIP identifiers to CRSP permanent

				identification numbers, and extract event study cumulative or compounded abnormal returns for cross-sectional analysis. Software, not dataset.
ExecuComp			General Corporate	Compensation data from SEC filings of top-five highest paid executives in US publicly-traded firms (1992-present) Tracks executive compensation in S&P 1000 firms. Top executives' salary, bonus, and stock option data since 1992 is available. The database also has company-specific financial statement information to supplement the compensation data. In WRDS, click on COMPUSTAT North America, then on Executive Compensation.
FIVE Project on Firm and Industry Evolution and Entrepreneurship			General Corporate	This is a repository for data on the evolution of specific industries. Each data set contains firm-level data, often on an annual basis beginning at the inception of the industry. Documentation and data are available for free in easy-to-use formats on the website. Currently, there are two data sets available, and we have commitments for additional data in the future. The two current data sets are the Sorenson Workstation FIVE Data and the Henderson Photolithography FIVE Data.
Fundacion Empresa Publica in Spain (now called SEPI).	This is amazingly rich data that is similar to the census of manufacturers in the United States. It is available for research use, but it is now for sale, in addition to requiring permission to access from the Spanish government. The Fundacion Empresa Publica data is fantastic – very rich, very detailed. Also, because it comes from government sources, it is generally reliable.	The downside is that you cannot identify each firm by name (only by number). Also, some of the variables have changed year over year, their definitions are inconsistent across years, and there is some noise in the measures (although not atypical).	General Corporate	Spanish census
Encuesta Sobre Estrategias Empresariales (Fundación Sepi).	Strength: long panel of firms with data on outputs, inputs, assets, investments, innovation, pricing, etc. Allows calculating productivity and relating this to firm level decisions.	Weakness: Access is given on a project per project basis.	General Corporate	Representative Data on Spanish industry (panel since 1990). Data is available (at price) for specific projects.
Japanese				Similar to Compustat, but contains information on

Development Bank (JDB) Database			Japanese publicly traded firms. This is a private source that must be bought. The Japan Company Handbook: available in many libraries. The entire series is available at the Library of Congress. In addition to financial data, this bi-annual publication contains a description of product lines of each Japanese firms and information on who owns the firm. Toyo Keizai Shinposha Directory - a Japanese language directory that contains information on the foreign subsidiaries of both public and private Japanese firms. I only know of the hard copy version of this source. The Ivey Business school has spent years creating a panel from the annual publications. It makes sense to try to co-author with an Ivey person if one does not speak Japanese, have access to these publications or have quite a bit of time to create their own panel. I found the subsidiary information to be uneven -depending on the parent firm. These data are richer than what is available in the Directory of Corporate Affiliates for US firms.
Texas data	Strengths: 1) Entire business population of a large state.	Weakness: 1) Quite a bit of clean-up required. 2) Only last 5 years directly available at this point in time (I have been collecting it for a decade).	General Corporate	There is also a number to call about older data.
TRINET		Primary weaknesses: (1) limited to every odd-numbered year 1979-1989; (2) currently only quasi-public - people who can get an OK from one of two of the professors who paid for	General Corporate	This data source was compared to COMPUSTAT in Davis, Rachel, and Irene M. Duhaime. "Diversification, Vertical Integration, and Industry Analysis: New Perspectives and Measurement." *Strategic Management Journal* 13, no. 7 (1992): 511-24. Couldn't find any current links to this database.

		assembly of the data (Rumelt and Teece).		
UK Data Archives			General Corporate	
US Census Bureau			General Corporate	
US Census Bureau longitudinal research database (LRD) LRD	Information on larger firms is excellent	The information for private firms is limited (employment and location). For smaller firms a statistical sampling approach is used. Access to the data is only for approved researchers.	General Corporate	The LRD is a survey put out every year by the census for a sample of US firms.
US Census Bureau longitudinal business database (LBD)	The main advantage and disadvantage are common to the two of them, and probably to other Census datasets: Both are very rich datasets that have been used by relatively few researchers (main advantages)	but the Census has made it very costly to access them in practice by requiring that you go through a proposal approval process that can take many months, even more than a year (main disadvantage, which has also contributed to the advantage that few people have used them)	General Corporate	US Census Bureau longitudinal business database (LBD) LBD is establishment level data for all firms in the US, public and private. However, Again, access is provided only to approved researchers. See discussion paper CES-WP-02-17. US Census Bureau: the Business Information Tracking Series (BITS) and the Longitudinal Research Database (LRD). - BITS covered all US establishments (manufacturing--i.e. plants-- or non-manufacturing) with paid employees from 1989-1996. I used it to study the diversification discount. The Census discontinued this database, but it has been superseded by a new one with similar variables but much longer coverage: The Longitudinal Business Database (LBD). This database covers over 24 million establishments from 1976 to the present (the "present" for the Census means a few years behind; I think the latest available now is 2005). Very few scholars have used this data source; the only paper I am aware of is one by Davis, Haltiwanger, Jarmin, Lerner, and Miranda, about the effect of private equity on employment. Main advantage is its large coverage, main disadvantage is that there are very few variables; no performance measures and only measure of size is number of employees. - LRD covers manufacturing establishments (i.e. plants only), but has many more variables than LBD that allow to compute detailed productivity measures. More researchers have used this database, such as Antoinette Schoar, and

				Maksimovic and Philips in Finance, and many others in Economics. More info about both databases and how they can be accessed can be found at: http://www.ces.census.gov.
US census RDC data			General Corporate	The Center for Economic Studies (CES) and its network of Census Research Data Centers (RDCs) support and encourage research activity using Census Bureau microdata to improve Census Bureau programs. CES and the RDCs provide access to researchers, federal agencies, and other institutions to non-publicly available Census Bureau data files. (from http://www.ces.census.gov/index.php/ces/cmshome).
U.S. EPA's Toxic Release Inventory			General Corporate	detailed data at facility level, a potential gem for strategy researchers even beyond environmental scholars
Worldscope	provides large samples in standard formats, with good documentation and support	public companies only, subject to the vagaries of and changes in public accounting standards, subject to the weaknesses of the standard industry classification schemes, etc.	General Corporate	Financial information for non-U.S. public companies, updated monthly. Coverage dates back more than ten years. See (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=871169 for a comparison with COMPUSTAT Both databases provide wide, longitudinal coverage of the performance of major companies. Worldscope used to have better international coverage, but the gap appears to have closed recently.
CorpTech		This is very poor quality, especially for the price ($5000 or so). It has very little consistent firm information beyond location, and in the one sector I examined, there were many errors in classification such as labeling a public firm as private, manufacturing facilities as independent firms, etc.	Entrepreneurship	firm-level data on US establishments
Business plan archive	The strength is a unique sample of dot-com era firms,	the weakness is that this is a convenience sample, so a lot of effort is needed to make sure inference is reasonable.	Entrepreneurship	The dataset is run by David Kirsch at the University of Maryland, and consists of a sample of several thousand business planning documents.

VenturExpert by Thomson Financial	– by no means perfect, but the best source on entrepreneurial ventures - has become the 'standard' Venture Capital database. - widely used, so audiences understand it, but there is also a certain amount of VE fatigue out there. Still it is a rich dataset. I think the primary weakness is that information on portfolio companies and VC firms is not longitudinal. - Strengths: It is longitudinal, provides pretty comprehensive data on venture capital investments and their outcomes, especially in the US.	Weaknesses: (1) The data on startups' IPOs and acquisitions are available, though I found that they were incomplete, and needed to be cross-checked with other data sources. (2) Data are collected through surveys, which makes them especially suspect in early years and outside the US. (3) One should not expect a lot of detailed information on the composition of the venture capital firms, or characteristics of startups. This database is really only about investments. Data on VCs and startups often need to be filled in from other sources. (4) There are always questions about accuracy and comprehensiveness. There are several studies that examine these issues, and demonstrate some biases one should be aware of.	Entrepreneur ship	- For strengths/weaknesses see http://faculty.chicagobooth.edu/steven.kaplan/research/kss1.pdf - For concordance between VentureXpert industry classification scheme (VEIC), and the more common SIC system see http://www-management.wharton.upenn.edu/dushnitsky/ - Kaplan, Sensoy & Stromberg (2002). How well do venture capital databases reflect actual investments? Working paper, available through SSRN. - Gompers and Lerner (2000), *The Venture Capital Cycle,* Chapter 16 - provides data on venture capital-backed startups around the world. Provides information about the start-up company (e.g., investors, rounds of funding, names of executives, some valuation data). Data collected via surveys to the firms (so is self-reported). Pretty standard source of information for VC-backed firms, and many universities have a subscription to the database, so fairly easily accessible. Pretty good domestic coverage. Not sure how good the international coverage is. - I use this through a library subscription, but not extensively. I have not always been able to find funded firms that should be included. - amazing data on portfolio company officers and board members but only current data (ie., not the history of officers and directors; just current ones).
VentureSource		I do not believe this database sells institutional subscriptions so is less easily accessible (more expensive). I usually try to triangulate between the two databases when I can.	Entrepreneur ship	The following paper discusses some of the pros and cons of VenturExpert vs. VentureSource: Steven N. Kaplan, Berk A. Sensoy, and Per Strömberg, "How well do venture capital databases reflect actual investments?"
CORI	- Selection is limited (but growing) although it provides a great source of publicly available SEC data on material contracts by large public firms.	CORI allows researchers to search over 600,000 SEC contract documents; however not all material contracts are yet available in the database, so potential biases may exist based	Contracts and Alliances	created by the Contracting and Organizations Research Institute at the University of Missouri-Columbia.

		on the selection criteria. The data is pulled from SEC filings and thus it focuses on large, public companies.		
CATI		Both are based on third-party reports of alliance announcements. As a result neither is exhaustive and there are definitely errors re dates, alliances announced but never initiated, etc., and there is a great deal of missing data, but SDC is arguably the best data available for alliance announcement data covering a broad range of industries. Uses of the data that are least problematic are for cross-section studies that do not require a complete population or industry network, and that focus on alliance initiation rather than evolution or termination. It is not possible to construct a reliable panel dataset from either SDC or CATI.	Contracts and Alliances	I believe CATI is still available from John Hagedoorn at Maastricht, but I'm not sure. (Note: NOT at http://www.cati.com/) I owned an early version of the data but have not used it in recent years as it was somewhat superceded by SDC. - Both SDC and CATI aim to provide info on inter-firm alliances established since around the late 1980s. CAIT focuses specifically on technology alliances (CATI stands for cooperative agreements and technology indicators) and is rather Euro-centric although it claims global coverage; SDC is broader-based in terms of types of agreements covered and also claims global coverage, but is still spotty on Asian agreements.
Frandata	Company has lots of data on franchising for sale, but only the most recent franchise agreements		Contracts and Alliances	database of franchisor executive addresses.
Recap for biotechnology alliance data. /	- Reasonably priced, - pretty good for biotech/pharma/univ alliance data; - concentrates on ventures in the biotechnology sector. Provides much more in-depth coverage of biotech firms (e.g., product development status) relative to data found at VenturExpert or VentureSource.	but difficult to match to other sources of pharma information. Spotty information on contract terms.	Contracts and Alliances	I have started to use this database very recently, and have little wisdom to share. I found the below article helpful in comparing alternative sources for alliance data: -Schilling, Melissa A. "Understanding the Alliance Data." *Strategic Management Journal* 30, no. 3 (2009): 233-60. A discussion of the Recap database, its strengths and weaknesses can be found in Josh Lerner & Robert P. Merges "The Control of Technology Alliances: An Empirical Analysis of the

				Biotechnology Industry", Journal of Industrial Economics, 46 (June 1998) 125-156.
Technology Licensing data		both these data are limited in that they are idiosyncratic samples:	Contracts and Alliances	a. MIT technology licensing data: These data were collected by Scott Shane, and I used them as a co-author. The data included all MIT licensed patents from 1980-1996. b. STanford Technology licensing data: These data were collected by myself, and I am using them in a project currently. The data from Stanford are not nearly as clean as MIT, but we have much richer data into licensing contracts. I do not see how the data will be made public without approval from Stanford's OTL.
SDC (Securities Data Corporation): tracks mergers, acquisitions, joint ventures Thomson Financial SDC Platinum (includes database on alliances & joint ventures, formerly "SDC Database on Alliances and Joint Ventures) http://thomsonreuters.com/products_services/financial/sdc#overview	Strengths: largest dataset of m&a data,- Most comprehensive coverage of alliance activity with some coding on nature of alliance activities	Weaknesses: missing data, big & bulky if downloading significant queries Based on news reports, press releases, SEC filings, etc., therefore, will not include all alliances for all firms. This makes it problematic where used to construct a firm's network, since the information is incomplete. -There is likely a size and nationality bias, given that the data is constructed from such news reports etc. We will more likely see reporting for large than small firms and more US than foreign firms. -Pre-1990 data is very patchy and not very reliable. - If used for event studies, dates must be externally validated. - Wonderful longitudinal deal-level data, but the data has issues. The joint venture and M&A databases unsurprisingly	Contracts and Alliances	- Both SDC and CATI aim to provide info on inter-firm alliances established since around the late 1980s. CAIT focuses specifically on technology alliances (CATI stands for cooperative agreements and technology indicators) and is rather Euro-centric although it claims global coverage; SDC is broader-based in terms of types of agreements covered and also claims global coverage, but is still spotty on Asian agreements. Both are based on third-party reports of alliance announcements. As a result neither is exhaustive and there are definitely errors re dates, alliances announced but never initiated, etc., and there is a great deal of missing data, but SDC is arguably the best data available for alliance announcement data covering a broad range of industries. Uses of the data that are least problematic are for cross-section studies that do not require a complete population or industry network, and that focus on alliance initiation rather than evolution or termination. It is not possible to construct a reliable panel dataset from either SDC or CATI.

		suffer the greatest problems, whereas data on underwriting appears to be generally reliable.		
State of Illinois Franchise Bureau database of franchise agreements	Good collection.		Contracts and Alliances	
State of California Department of Corporations franchise agreement database.	Good collection.		Contracts and Alliances	The actual link to the database may be here (Corporations California Electronic Access to Securities Information (Cal-EASI) Website).
Derwent Innovations Index (from Thomson Scientific)	– Also a really good database, in my opinion. In it, librarians at Thomson Scientific screen patents from the top 40 jurisdictions and classify them by the innovations they represent. This means that when two patents represent the same innovation, they are filed under the same record (i.e., the innovation record). These innovations are then categorized by codes, and the codes are available in a catalogue. Although codes are not uniform, sometimes representing a class of products and sometimes a class of technologies, the categorization can prove extremely useful for many different projects.		Technology	
Derwent World Patents Index® (DWPI)			Technology	
Eurostat Community Innovation Survey (available with micro-adjustment to hide the identity of the individual	Strength: Interesting firm level database with qualitative information on many innovation related measures in all EU countries. The type of information is rather unique.	Weakness: To match the data with firm level data (balance sheet/income statement) one needs the name of the firm which is not available. This means you actually need access	Technology	

firms).		to the data through the individual country that organized the survey for Eurostat, complicating access and interesting cross-country comparisons. As a result the more interesting papers with this data set are based on proprietary country level data.		
NBER patent citation datafile	– easy to use, would like to see it updated more frequently (like annually)	- Becoming overused? - Was not updated through 2007 last I checked.	Technology	These data comprise detail information on almost 3 million U.S. patents granted between January 1963 and December 1999, all citations made to these patents between 1975 and 1999 (over 16 million), and a reasonably broad match of patents to Compustat (the data set of all firms traded in the U.S. stock market). These data are described in detail in Hall, B. H., A. B. Jaffe, and M. Tratjenberg (2001). "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498. UPDATES The NBER is working on a major NSF-funded update and extension of this data. A new release of these files, bringing existing data up to date through December 2004, is anticipated for 2010 or 2011. A variety of additional fields and indexes will also be provided. These are anticipated to include "link-out" tables connecting patent numbers to geographic entities (e.g. SMSAs), and a codification of inventor names. The PI for this project is Iain Cockburn. Please contact him if you have questions or comments, or would like to contribute to this project. Updates through 2002 are available on Bronwyn H. Hall's website.

NBER Cassis	- The Cassis product is the USPTO's latest updates to the patent data file. The data are not very clean - but this is useful if one needs fields not in the NBER data file. Also, Cassis provides a dvd-rom of all reassignments.		Technology	Cassis Optical Disc Products are available worldwide at Patent and Trademark Depository Libraries and intellectual property offices and through direct sales to the public. These products deliver patent and trademark information in various configurations including full-text facsimile images and searchable text records.
Analytical Business Enterprise Research and Development database (ANBERD)			Technology	ANBERD data on R&D spending by country/industry, data on gross production by industry. The Analytical Business Enterprise Research and Development database (ANBERD) was developed to provide a consistent data set that overcomes the problems of international comparability and breaks in the time series of official business enterprise R&D data (OFFBERD) provided to the OECD by its member countries through the joint OECD/Eurostat R&D survey. Through the use of established estimation techniques, the OECD Secretariat has created a database designed to provide analysts with comprehensive and internationally comparable time-series on industrial R&D expenditures. Since many of the data points are estimated, ANBERD data may differ from data published nationally.
Micropatent (from Thomson Scientific)	the main use I have found for this database is extracting additional information once I have a list of patents to work with (sampled through other means). Information such as forward citations, cited prior art, etc. can be found for many jurisdictions (not only USA). Many other uses are also available.		Technology	MicroPatent is the world's leading source for online patent and trademark information. Combining advanced technology with the most comprehensive, up-to-date IP information, MicroPatent delivers a complete intellectual property solution—whether it's online searching, document delivery, patent analysis, file histories, or professional search assistance. (from website)
USPTO Patent Assignment and Bibliographic data		I have purchased DVDs (which are $100 or so each) and downloaded information from the web in XML format. The reassignment data has a number of problems with misspellings	Technology	Announcing beta.uspto.gov The USPTO is pleased to announce the beta test release of its new Web site. The new site has been redesigned to improve the look and feel, as well as to enhance the user experience with improved navigation. The USPTO's goal is to make the Web site technologically up-to-date, user-

		and other inconsistencies. One has to re-do all the work that went into creating the NBER patent data (defining technology fields, standardizing firm names, assigning CUSIPs, etc.).		friendly, and responsive to customer feedback. Please keep in mind this is the initial release of the beta site and the content is not currently up to date. The Office of the Chief Information Officer (OCIO) is still in the process of migrating and reorganizing the content. We invite you to explore the beta web site and give us your constructive feedback. Your comments and suggestions will help us to build a site that better serves your needs. Please submit your feedback either through moderator, or send your comments to betafeedback@uspto.gov.
Andy Rose's website			Location (i.e., country, state, county, …)	
Bureau of Economic Analysis	Great background longitudinal data that can be used as controls such as: prices; industry output levels.	Unfortunately, it has very limited data beyond the US.	Location (i.e., country, state, county, …)	
BLS data for data on CPI and cross-country labor data			Location (i.e., country, state, county, …)	
Eugene	Wonderful source for dyadic country data and some national political and macroeconomic data		Location (i.e., country, state, county, …)	EUGene is a program designed primarily for political scientists. It has 2 purposes. First, EUGene generates data for variables used to test Bruce Bueno de Mesquita and colleagues' version of an expected utility theory of war and dispute initiation (Bueno de Mesquita, 1981, 1985; Bueno de Mesquita and Lalman, 1992). Second, EUGene serves as a data management tool for creating data sets for use in the quantitative analysis of international relations; with the country-year, directed-dyad-year, non-directed-dyad-year, and directed-dispute-dyad-year as the unit of analysis. Until now, these data have been unavailable, and these data management tasks have frequently been cumbersome and difficult. A paper exploring EUGene's capabilities and

				rationale is available here. EUGene is an acronym for Expected Utility Generation and data management program. EUGene is freeware, but is copyrighted. A full download (see the download page) contains the program, expected utility data, documentation, and source code.
Getty Thesaurus on Geographic Names for latitude/longitude data for particular world cities			Location (i.e., country, state, county, …)	
Hofstede's culture data			Location (i.e., country, state, county, …)	
NBER trade databases, via Feenstra,	- The Feenstra site is a very thorough and helpful resource.		Location (i.e., country, state, county, …)	
Penn World Tables (for GDP data),			Location (i.e., country, state, county, …)	The Penn World Table provides purchasing power parity and national income accounts converted to international prices for 188 countries for some or all of the years 1950-2004. The European Union or the OECD provide more detailed purchasing power and real product estimates for their countries and the World Bank makes current price estimates for most PWT countries at the GDP level (from website)
POLCON data set			Location (i.e., country, state, county, …)	Scroll down website to find: The Political Constraint Index (POLCON) Dataset (in zipped MS Access or MS Excel formats with codebook) Log-in page asks for e-mail address, etc. · Download POLCON Database * 2006 release (with data to 2004) now available *
Polity			Location	The Polity IV Project continues the Polity research

			(i.e., country, state, county, …)	tradition of coding the authority characteristics of states in the world system for purposes of comparative, quantitative analysis. (from website)
Shalom Schwartz's measures on culture			Location (i.e., country, state, county, …)	Schwartz Value Survey, 1998-2005
World Development Indicators: Through World Bank.	This database is as good as it gets for longitudinal data on many measures for a large number of countries.	Weaknesses: We are never quite sure how well or standardized the data collection is, especially in developing countries. Many interesting series have a lot of missing data, and developed countries are better represented.	Location (i.e., country, state, county, …)	
WHO Mortality data		The files available here do not constitute a user-friendly data collection which the average user can download and access. These are the basic underlying raw data files, together with the necessary instructions, file structures, code reference tables, etc. which can be used by institutions and organizations which need access at this level of detail AND have available the required information technology (IT) resources to use this information. (from website)	Location (i.e., country, state, county, …)	Cross-country data on deaths by disease and year, starting in the 1950s. Disease coding changes over time and at different points in time for different countries, and there are many missing observations. Gary King has written (and published on his website) a book discussing how to use it.
Audit Bureau of Circulation	Great longitudinal data on magazines.	The most serious issues are with sample selection since magazines pay to be included and only paying magazines are included.	Industry- or field-specific	
Chilean TV data:			Industry- or	

			field-specific	
proprietary to tv network there, hour by hour data for different channels and over years			field-specific	
Hall's reports	Detailed advertising data on magazines, much of it collected since the 1930s.	As with ABC, Hall's Reports cover only a small number of the thousands of magazines selected primarily based on size.	Industry- or field-specific	
Global cement industry statistics, available for purchase from Cembureau	- The cement data are published documents, and very thorough in that they cover all countries of the world.	Some of the older data are slightly suspect.	Industry- or field-specific	
IMS R&D Focus and Pharmaprojects: drug development data		Both allow exporting of information into Excel or other formats, but require substantial work to convert into panel datasets. They keep failed projects, which is of key importance, but relabel projects post-merger.	Industry- or field-specific	- Both are proprietary and cost $5000-$15000 per annual subscription, depending on how one negotiates an academic discount and the number of users.
IMS Midas data: drug price information	- This is essentially the only source of price and share data at the package source across countries.	It is not easy to work with, and the information for certain countries is known to be incorrect (the US data does not include rebates, e.g.). It is also very expensive.	Industry- or field-specific	Covers 45 of the world's leading markets, Accounts for 87% of the Total world pharmaceutical market (from website)
IMS US Prescription Audit and marketing data	- Same as Midas data but only for the US, and with much more detail on marketing.	Very expensive.	Industry- or field-specific	Can't find something with this specific name from the IMS website. Try here.
Integrated Circuit Engineering or ICE	- The primary strength of the data lies in its coverage of corporate objectives, facility level product and process technology investment, and alliance level product and process technology investment. The corporate		Industry- or field-specific	This data provides global information at the corporate-, top management team-, manufacturing plant-, and alliance-levels of analysis. The providers indicate that they cover upwards of 70% of global competitors. I have been able to confirm that the ICE data is consistent with other sources in terms of factors such as mean annual revenue, variance in annual revenue, and distribution of small, medium, and

	objectives session provides a seemingly valid qualitative snapshot of a company's overall strategy similar to what one might distill from a company website and annual report. The facility level data provides seemingly accurate information on plant level production capacity, products, and process technology. At times it is difficult to compare plants across time. The alliance level data is similar to what we find in SDC—it may be slightly richer in terms of content but provides information on only the "key agreements."			large revenue firms. Can't find a website that I think this is relating to.
International Fabs on Disk	I think the semiconductor data is wonderful data, in general. It is amazingly detailed at a disaggregated level.	The only complaint that I have about the data is that it is difficult to track performance and sometimes has some inconsistencies.	Industry- or field-specific	(available from the SEMI – semiconductors equipment manufacturers industry association – www.semi.org) – if I remember correctly however, the dataset now has a new name but I am not sure what it is since I haven't used it in awhile. Link at far left is to: Worldwide Fab Database Reports at SEMI.
The Internet Movie Data Base (IMDB)	- IMDB is great for that subset of strategy interested in the film industry. The amount of data is phenomenal (literally millions of career histories), and it contains unique identifiers for individuals, film titles, and companies.	One drawback is that all the data is based on user contribution, so the quality and extent of information varies from title to title. Another drawback is that the data is difficult to work with - it is not a flat file ready to load in to Stata, so a bit of programming experience is required.	Industry- or field-specific	
Medical Healthcare and Marketplace Guide	strength: full population in US medical sector; rich data on structure, product markets, executives.	Weakness: most of it is not digitized, in hardcopy format	Industry- or field-specific	Annual subscription is $900 and only $400 for each additional user.
MOTOR CARRIER ANNUAL	Strengths - provides income statement, balance sheet, and other	Weaknesses - only covers firms with sales exceeding $1	Industry- or field-	

REPORTS	data for U.S. interstate motor carriers whether public firms or private firms; covers both regulated and deregulated period; spans many years.	million/year; some data fields are sparse in some years; rumor has it that firms began to stop submitting the data in the mid-1990s, so it gets spotty then.	specific	
Nielsen	terrific for consumer-level information,	very expensive however, panel	Industry- or field-specific	
NLM's clinicaltrials		Better coverage in recent years, but registration of trials is not enforced and anything on the site is an essentially voluntary disclosure by firms.	Industry- or field-specific	Registry of clinical trials, which can be downloaded in XML format. Contains information about geographic scope, trial participants, trial sponsors, etc.
Oil and Gas Journal	Detailed project-level investment data worldwide on the construction of oil and gas productive facilities. The data is in excel format year by year.	Massive data cleaning and matching required both programming and manual work. Before 1996 the data is in a very poor format (and is not available in excel only pdf or hard copies). Missing data is sometimes a problem for the earlier years. Otherwise generally reliable.	Industry- or field-specific	
Pharmaprojects (from PJB)	This is a really good database that tracks drugs entering clinical trials. It contains technical as well as market information and it is that combination that makes it adequate for research in technology strategy.	That said, its main weakness is incompleteness: it is impossible to get a full roster of all drugs that all companies are introducing to clinical trials worldwide, so this database represents those drugs that firms are willing to disclose.	Industry- or field-specific	
semiconductor industry association			Industry- or field-specific	including billing and shipping reports
Simmons	reasonably expensive but very good because cross-media consumer level purchase behavior, nice information on consumer demographics, attitudes, and beliefs, time-series of cross-section		Industry- or field-specific	

	rather than panel			
US Hotel industry data	The hotels data is very good, very reliable, but pretty expensive. Also, they are reluctant to release too much hotel specific information such as occupancy rates due to confidentiality concerns.		Industry- or field-specific	Available for purchase from Smith Travel Research
Brobeck Archive	These data are marvelously rich, but access is cumbersome and time consuming.	Accessing the data has and will continue to require on-site access at the Robert H. Smith School at the University of Maryland, and hitherto has been done only on an experimental basis. The archive should be open to approved researchers in the coming months.	Non-market or Lobbying	This is a database of legal documents of the failed law firm Brobeck, Phleger and Harrison.
Charles Stewart (MIT) dataset on congressional committees			Non-market or Lobbying	
Derwent litigation data			Non-market or Lobbying	
Follow the money (National Institute on Money in State Politics) for state-level data	Strength: micro state-level campaign contribution data on who gave what to whom and when.	Weaknesses: spotty coverage of states and election races before 2000. Does not include lobbying data.	Non-market or Lobbying	I visited the institute that gathers the state-level data with Rick Vanden Bergh a few years ago and was impressed by the rigor they use in compiling and checking the data, which comes to them from a multitude of state election agencies.
Datasets on the ideology and voting patterns of judges			Non-market or Lobbying	- Segal and Spaeth Supreme Court Database (also available here) - Donald Songer Appeals Court Database - Zuk, Gryski Dataset on the Characteristics of Judges
U.S. federal-level lobbying data			Non-market or Lobbying	
Gompers, Ishii and Metrick corporate governance data			Governance	Raw data from IRRC Institute More info here. See also RiskMetrics listed below, it's not clear where the original data set ended up.

Institute for Industrial Research		The Institute has a number of publications on R&D. R&D structure data do not identify firms, but I have done some matching and am glad to share the results.	Governance	data on R&D organization structure.
KLD Research and Analytics: Social and Governance factors	- Comments: unique variables not available elsewhere (i.e., measures of "employee" friendliness), however, number of firms covered is limited. - The KLD data are very good in that they are fairly comprehensive – they track the social performance of all the Russell 3000 firms.	Unfortunately, the KLD measures themselves are sometimes not that great. KLD sometimes tracks information that is easily observable rather than focusing on measuring the "right" things.	Governance	KLD Indexes constructs indexes for investors who integrate environmental, social and governance factors into their investment decisions. KLD's indexes are designed to be transparent, representative and investable.
law and finance variables. -Andrei Shleifer's dataset -Spamann's law and finance variables			Governance	In the law and finance field, there are now a number of competing variables meant to capture the quality of governance institutions (such as Spamann's), and that has been helpful for the growth of the field.
RiskMetrics (formerly IRRC): data on governance measures (G-index), Directors and Shareholder proposals	Strengths: tracks many governance provisions (24+) for a sample (~1500) of publicly-traded firms (data behind Gompers, Ishii, Metrick G-Index)	Weaknesses: survey not conducted for all years (skips years)	Governance	RiskMetrics (through ISS Governance Services) is a leader in corporate governance data. They deliver to WRDS four datasets identified as: RiskMetrics Group Historical Governance, Historical Directors data, Voting Results data, and Shareholder Proposal data.
Social Investment Forum			Governance	data that tracks socially responsible investing
ISI for massive amounts of bibliometric data			Citations and Networks	
PubMed	an absolutely amazing resource that is underutilized by social scientists		Citations and Networks	
Science Citation	This is relatively easy to deal with,		Citations	

Index (now part of ISI)	except for the problem of matching authors to institutions. It is cleaner than the patent data.		and Networks	
Social Register			Citations and Networks	
S&P's Register			Citations and Networks	
Who's Who			Citations and Networks	
CEPII			Yet to Classify /Other	
Concordance Between SIC and US Patent Classification (USPC).	Pros: invaluable for linking patents-to-industry. Easy to use alternate to Silverman concordance.	Cons: Not as fine grained - only mix of 2/3 digit industry level.	Yet to Classify /Other	
PATENT-SIC CONCORDANCE	Strengths - useful way to connect patent classes to SICs at the four-digit SIC level.	Weaknesses - based on a data set that ended in 1994, so the concordance is not likely to be accurate for too many years since then.	Yet to Classify /Other	
Concordance between VentureXpert industry classification scheme (VEIC), and the more common SIC system:			Yet to Classify /Other	
Hoppenstedt			Yet to Classify /Other	
Internet Archive	- The internet archive is a website that provides monthly snapshots of the web back to 1996. I have used the site to date entry and exit of		Yet to Classify /Other	

	internet firms.			
Mergent			Yet to Classify /Other	
Piranhaweb (now Thompson One Analytics)			Yet to Classify /Other	Review here; more info on renaming here
Proprietary hand-coded data.	On the plus side (a big plus), hand collected and proprietary data often allow researchers to say much more about strategy issues. This is one of the best things about our field—we've got data no one else has bothered to collect, since some other fields rely heavily on large, well-trod databases, which can promote a lot of incremental studies (which strategy doesn't suffer from as much, although sometimes I worry that the NBER patent data are starting to be overused).	Hand collecting data and cleaning it up takes a long time. Proprietary data generally are confidential, so they can't be made available to other researchers. Proprietary data also often don't identify the firms by name, so it isn't possible to add publicly available data to the proprietary data on a firm level basis.	Yet to Classify /Other	
the publicly available archives of many technical standards bodies (e.g. http://www.ietf.org)			Yet to Classify /Other	See also http://fisher.osu.edu/~leiblein_1/Strategy%20Resource%20Links2.htm