Greg Morrison

A birds-eye view of SWITCH big data project

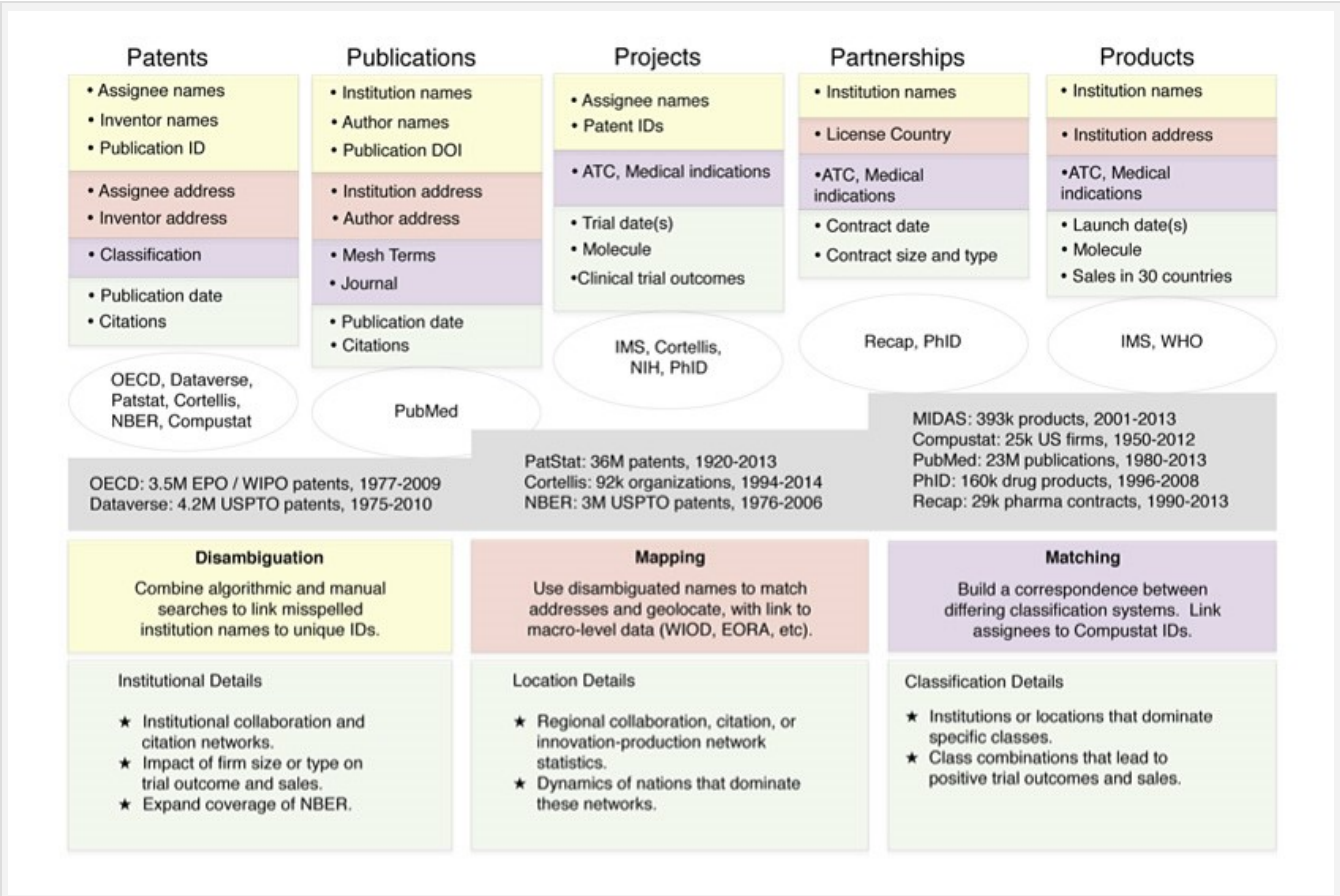

Figure 1: A summary of available data sources and the topics they cover (top row), ongoing projects to link the data sources (middle row), and applications of the resulting data merge (bottom row).

SWITCH relies on a wide range of datasets, which can be used to quantitatively understand innovation and competition in pharmaceuticals, and to describe the impact of geographic, technological, institutional, and temporal differences on a variety of processes and domains of investigation.

Main sources of data:

- *Patents*: Multiple databases provide information about patent applications and publications, with different information contained in each list. Many include technological classifications and citations [dataverse, oecd, orbis], with high-resolution geographic information available for applicants [oecd] of European and inventors [dataverse and oecd] of European and US patents. Applicants can also be linked to institution type [nber] and balance sheets [orbis, compustat], and pharma patents can be cross-linked to drug trials with which they are involved [cortellis].

- *Publications*: Pubmed provides information on biomedical publications. Publications can be linked to research institutions, individual researchers, clinical trials, patents and drugs.
- *R&D Projects*: There are multiple dasets that quantify conversion of the theoretical innovation in the Patents and Publications data into drugs that undergo clinical trials. Patents and applicant institutions can be linked to the drug trial outcomes [clinicaltrials.gov, cortellis, ims, PhID], and each drug can be classified based on the Anatomic and Therapeutic Classification (ATC) or related medical indications.
- *Collaborative Alliances*: The licensing of innovation between firms or their direct collaboration on research projects are detailed in a few databases [recap, PhID] that provide information about contracts between firms. In particular, the date, duration, and type of contract are provided.
- *Products*: Once a drug has become available, its impact can be assessed using a variety of data sources. Thanks to the support of IMS International, we can link drug sales data and launch dates to information on R&D projects as well as on division of labor in R&D, tracing the process of drug development from the birth of an idea to the launch of a product and final market dynamics.

To accurately link the inventors, institutions, and products between datasets, we have implemented a process of disambiguation, which aims at recovering a unique name from misspelled or variations on the spelling of a name. To understand the influence of inventor location, we must accurately resolve geographic information from the data provided. With the disambiguation performed we can link the firm names to additional information, such as their balance sheets and other company information. Once the data are fully and accurately linked, a number of relevant issues can be addressed:

- Understanding the structure of institutional collaboration and citation networks. Are successful collaborations typically regional or international? Are collaborations between institutions of different types (Firms, Universities, etc.) more likely to be successful? To what extent do national citation networks (a proxy for innovation spillovers) benefit from international spillovers?
- Understanding the importance of technological distance in comparison to physical distance. Which technological classifications are dominated by a few large institutions and which allow for a number of smaller institutions? Are collaborations between technologically diverse firms more likely than technological focused ones?
- Analyzing clusters and networks of innovators in biopharmaceuticals based on a fully-disambiguated and geo-localized dataset.
- Tracing the sources of innovative drugs in basic research and mapping the diffusion of innovative ideas.
- Investigating the main determinants of firm's decision on where to locate R&D investments and mobility of inventors.

One key goal of our analysis is to develop a comprehensive data set with firm level data. In this effort, we rely on two additional data sets, which we are currently integrating with the above mentioned data.

Compustat North America is a database for US and Canada covering financial, statistical and market information for active and inactive companies that have at least one financial security

listed on global markets. Detailed companies' financial accounts are provided since 1962 for about 25,000 US (active or non-active) firms in both manufacturing and service industries. Quantitative and qualitative information on securities' pricing, geographic and industrial segments of activities by firm are included.

Cortellis is a Thompson-Reuters dataset that links firms, pharmaceutical patents, and drug trials. Pharmaceutical patents include those in the European Patent Office (EPO), the World Intellectual Property Organization (WIPO) under the Patent Cooperation Treaty (PCT), or the United States Patent and Trademark Office (USPTO) from 1995 to 2014. The firm names are disambiguated (corrected for variable or incorrect spellings), and provides links to firm details, contracts between firms, patents, drug products or trials, and a large amount of additional details.

The Dataverse patent dataset covers granted patents in the United States Patent and Trademark Office (USPTO) patents from 1975-2010. Dataverse provides a disambiguated list of authors (i.e. that corrects for spelling variations or errors). The NUTS3 regionalization is not available, but the Dataverse dataset does provide Longitude / Latitude coordinates so that inventor locations can be accurately geolocated. The data also includes patent citations and assignee names (without any geolocation), as well as application and grant dates. See: Ronald Lai; Alexander D'Amour; Amy Yu; Ye Sun; Lee Fleming, 2013, "Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database (1975 - 2010)",

The EORA (name of a group of native Australians; the data set is given by this name because it is an Australia-based project) is a multiregional input-output (MRIO) database that covers 187 countries and the period from 1990 to 2011.
EORA has two versions of **MRIO** tables. One uses the heterogeneous classification of industries and, depending on the original source, different countries can have different numbers of industries. The other uses the harmonized classification of industries and each country has the same 25 industries.

Eurostat is the official statistics institute of the European Commission.
It collects general and regional, economic and finance, population, transport, science, environment statistics on the EU countries.
The demographic database includes general population life tables, population age-structure composition, fertility, mortality and migration data from 1960 to 2013.
Eurostat produces also population projections (EUROPOP2010) which are used in the other European Commission reports (Ageing Report, Sustainability report).

The National Bureau of Economic Research (NBER) has patent data in its Patent Data Project (PDP) dataset. The data covers United States Patent and Trademark Office (USPTO) patents from 1976 to 2006. The NBER PDP provides disambiguation for the assignee of various patents (to correct for spelling variations and errors in the assignee field), and provides a link between assignee names and CompuStat Identifiers. Each disambiguated assignee is also labled with an institution type (e.g. US corp., Foreign Corp., University, Hospital, etc.). The PDP also provides information about the dynamic ownership of patents, from assignee to current owner. No information is provided about inventors.

The OECD dataset covers patents filed to the European Patent Office (EPO) or to the World Intellectual Property Organization (WIPO) under the Patent Cooperation Treaty (PCT). The data covers 2.3M EP patents and 2.2M PCT patents, with 1.2M patents covering the same IP filed in both offices. Listed are all patents from 1978-2010, with partial coverage of more recent patents. The data include:

- Detailed assignee and inventor locations on the level of NUTS3 regionalization for all EP and WO patents.
- Citations from all EP and WO patents.
- Triadic families for EP and WO patents, with families covering patents that share at least one priority application in the EPO, Japan Patent Office (JPO) and United States Patent and Trademark Office (USPTO).

Orbis, by Bureau Van Dijk, is a firm-level database with information on more than 113 million companies worldwide. Among others, it includes original information provided by PATSTAT for about 36 mln patents that have been filed by at least a company. In turn,the **PATSTAT** database is established and maintained by the European Patent Office (EPO), containing bibliographical data on the majority of patents currently in force worldwide.
From **Orbis** data, it is possible to match applicant companies and patents with unique identifiers, combining information at the firm-level (financial accounts, ownership) and at the patent-level (inventors lists, publication dates, etc.). The matching process was a result of a mutual agreement between the compilers of Orbis and the OECD.

The Pharmaceutical Industry Database (PhID) dataset covers the full development history of 34k R&D projects, linking them with both the patents and firms that were involved in the creation of the product, the diseases targeted and molecules used, and the sales data in 26 countries.

The World Input-Output Database (WIOD) is a time series of multiregional input-output (MRIO) tables that cover 35 industries for each of the 40 economies (27 EU countries and 13 other major economies) plus the rest of the world and cover the period from 1995 to 2011. The MRIO framework is well suited for the study of the recent phenomenon of trade fragmentation (global value chain) and for the impact analysis of any particular industry on the world economy.
WIOD provides free access to its data and the website is

Recap is a Thompson-Reuters dataset that lists contract deals between pharmaceutical companies (e.g. licensing, research agreements, etc) from 1990 to date. A great deal of data is often available on each deal, including dates, subject areas, therapeutic areas, deal sizes, trial phase, etc. 29k contracts are specified between 3.3k pharmaceutical research firms, universities, and hospitals. While the name spellings are corrected, there is no link immediately provided between the Recap names and the names in other datasets.