Event Analytics for Innovation Trajectories: Understanding Inputs and Outcomes for Entrepreneurial Success

Submission Type: Article

Authors:

C. Scott Dempwolf, Department of Urban Studies and Planning, National Center for Smart Growth, University of Maryland, College Park

Ben Shneiderman, Department of Computer Science, University of Maryland Institute for Advanced Computer Science, University of Maryland, College Park

Short Title: Event Analytics for Innovation Trajectories

Corresponding Author: C. Scott Dempwolf
Assistant Research Professor
School of Architecture, Planning and Preservation
3835 Campus Dr.
College Park, MD 20742
(301) 405-6307 (voice)
(301) 314-9583 (fax)
dempy@umd.edu

Acknowledgement: This research was supported in part by the National Science Foundation, Award #1551041.

The authors declare no conflicts of interest.
Event Analytics for Innovation Trajectories:
Understanding Inputs and Outcomes for Entrepreneurial Success

Abstract: New analysis tools are expanding the options for innovation researchers. While previous researchers often speculated on the relationship between inputs, such as patents or funding, and outcomes such as product releases or IPOs, new software tools enable researchers to analyze innovation event data more efficiently. Tools such as EventFlow make it possible to rapidly scan visual displays, algorithmically search for patterns, and study an aggregated view that shows common and rare patterns. This paper presents initial examples of how event analytic software tools such as EventFlow could be applied to innovation research, using data from 34,331 drugs or medical devices.

Keywords: Event analytics; EventFlow; Innovation metrics; STI, Visualization

Introduction: STI Systems and Processes

Worldwide interest in promoting economic growth through innovation has grown dramatically. As a result, there is increased effort by researchers in Science Policy and Scientometrics to study and measure Science, Technology and Innovation (STI) to help understand the basis for success or failure. They are concerned with understanding, describing, measuring and visualizing the scope, organization and structure of human knowledge as a dynamic collection of concepts (Scharnhorst, Börner, and Besselaar, 2012).

Such concepts are connected to and acted upon by a network of scholars and inventors engaged in the discovery and creation of new knowledge and technologies. These discoverers and inventors in turn engage with networks of institutions, agencies, organizations, intermediaries, entrepreneurs and investors who sponsor their activities and help translate the results into new products and services in the marketplace. Taken together, these networks along with their embedded knowledge and resources comprise what we have recently recognized as innovation...
ecosystems. Understanding, modeling and measuring these dynamic and complex adaptive systems has become an important priority within science policy and scientometrics (Börner, 2016).

Our modeling of research and development activities enriches the prevailing network approach with event analytics by focusing on time-stamped point events (such as getting a patent) or interval events (such as the funding period covered by a grant or contract).

We see STI processes as comprised of sequences of point and interval events that together result in the translation of knowledge and research into new products and services in the marketplace. Point events are associated with a single date / time, for example the date of a patent application. Interval events are associated with start and end dates / times. Research projects or research grants with start and end dates are examples of interval events. These events generally fall into one of several categories including research, invention, proof, and several types of commercialization events. Each event is associated with a document or record that describes the event, the key people and organizations involved and what roles they played, when and where the event occurred, along with other attributes. The information from these records, especially dates, may be used to model event networks of people, organizations, places and documents.

Events that contribute to the development of specific products and services may be associated with each other, creating product and service event sequences or trajectories. The trajectories may be connected through the networks of the people, organizations, places and documents.

---

1 The phrases "products" or "products in the marketplace" are construed broadly throughout this paper to include all types of innovation and all types of "marketplaces" including public domain.

2 The order of activities here generally follows the linear model of innovation. This ordering is primarily a matter of convenience and should not be construed as proffering any particular model or theory of STI processes.
involved, and through their contributions to specific product and service event sequences. Conceptually, this dual modeling structure (innovation networks / innovation event trajectories) provides a linkage between STI as complex adaptive systems and STI as complex processes.

**Why Innovation is Hard to Measure and how Event Analytics Can Help**

A streamlined definition of innovation is the process of working on marketplace problems, which elicit innovators to transform ideas and scientific knowledge into new products (broadly defined to include services). The innovation process connects marketplace problems with research events, however each product follows a unique path involving different types of activities including research, publication, invention, prototyping, ‘proof’, and several commercialization events culminating in a new product launch. The trajectory a product takes may involve multiple events within any stage, and may involve revisiting a prior stage if remedial work is required. Thus the first difficulty in measuring innovation is the unique and variable nature of the innovation trajectory or sequence of events for each product.

A second difficulty is that early stage research events are often undertaken for the purposes of knowledge creation and publication. In fact, the explicit innovation goal of a new product may not yet exist. There is a temptation to define the distinctions between science, technology and innovation more rigidly, but this creates as many problems as it solves. The creative moment when the product is first envisioned involves a specific set of conditions that are a function of the sequence and characteristics of events up to that point. It is as if the innovation path suddenly appears midway through the journey.

Mathematically this describes a Markov chain or Bayesian network model in which each event in the sequence is influenced by the cumulative effect of everything that has happened up to that
point. Neither the final destination nor the intermediate events can be known with certainty.
They may however be estimated based on certain probability distributions.

Modeling and analyzing innovation event trajectories for successful products a posteriori establishes the basis for estimating those baseline probability distributions. This in turn allows the formulation and testing of more sophisticated hypotheses. It may also allow the development of predictive models, or facilitate machine learning and the development of related big data applications. Finally, the goal would be prescriptive modeling that would enable policy makers at funding agencies, investors, and entrepreneurs to make decisions that lead to more successful outcomes.

**Current Innovation Metrics and the need for New Measures of Innovation**

In 2011 the Committee on National Statistics and the Board on Science, Technology, and Economic Policy of the National Research Council convened the Panel on Developing Science, Technology, and Innovation Indicators for the Future and charged the members with assessing the current state of innovation metrics and preparing recommendations for future measures of STI. The panel’s 2014 report was detailed and extensive in both areas, drawing on both U.S. and international research (National Research Council, 2014). The report is intended to provide guidance to the National Center for Science and Engineering Statistics (NCSES) at the U.S. National Science Foundation (NSF), the study’s sponsors.

NCSES currently produces many statistical measures of innovation inputs, outputs and long-term outcomes including metrics of: Research and Development R&D; National R&D expenditures and performance (by type of industry and source of funds); Commercial Outputs and Outcomes;
Knowledge Outputs; Science, Technology, Engineering, and Mathematics (STEM) Education; STEM Workforce/Talent; and Organizations/Institutions (National Research Council, 2014 Box 3-1, pp 38-39).

Traditionally, NCSES and its predecessors have used surveys including BRDIS to trace the inputs and outputs of the innovation system. More recently, alternative data sources including administrative and electronic transaction records for example, are increasingly available (National Research Council, 2014, p56). Along with these new data sources, widespread and low cost computing power has made the use of new analytic methods possible. These methods include network and temporal analysis, for example. The availability of new tools including NodeXL\(^3\) for network analysis and EventFlow\(^4\) for temporal analysis, for example, can help innovation researchers develop new innovation metrics.

The panel was unequivocal on its recommendation that NCSES should develop new metrics of innovation, particularly innovation outputs. These metrics are needed, the panel concluded, “to assess the impact of federal, state, and local innovation policies, such as the amount and direction of federal R&D funding, support for STEM education at the graduate level, and regulation of new products and services. In addition, having good measures of innovation output facilitates comparison of the United States with other countries in a key area that promotes economic growth” (National Research Council, 2014, p43). The report also listed a selection of real and relevant policy questions for which new metrics are required to formulate appropriate answers.

Visualization as a Tool for Exploration and Understanding

---

\(^3\) NodeXL: Network Overview, Discovery and Exploration for Excel. (https://nodexl.codeplex.com/)

\(^4\) EventFlow: Visual Analysis of Temporal Event Sequences (http://hcil.umd.edu/eventflow/)
Innovation researchers have used diverse visualizations to explore data, derive insights and present results. Traditional visualizations include these data types with example applications from innovation research:

1. Choropleth maps to show intensity of innovation activity by county, state, etc.
2. Scatterplots and heat maps
3. Timelines and hierarchies to show intensity of innovation activity in patent taxonomies
4. Networks to show connections among university or venture capital firms and start-up companies

[figure 1 about here]

The emergence of tools for new data types offers fresh opportunities for innovation researchers to understand event patterns that could guide interventions to increase the success of innovation efforts. Current interest in event analytics has been triggered by the growth of electronic health records, which now provide online access to tens of millions of patient histories. These histories reveal patterns of medication compliance, links between treatments and side effects, and the relationship between interventions and outcomes (see for example Carter, Burd, Monroe, Plaisant and Shneiderman, 2013; Onukwugha, Plaisant, and Shneiderman, 2016).

Increasing availability of innovation histories could produce similar benefits by allowing researchers for the first time to study the relationships between events in start-up companies and their eventual success or failure. Event analytics is a new and growing topic within visual analytics that combined interactive exploration with statistical tools to find expected common trajectories and unexpected anomalies. Patterns may be as simple as seeing how often patents
lead to start-up companies getting founded or venture capital investments lead to acquisition of
start-up companies, or they may be more complex.

Temporal event sequences consist of thousands or millions of events, which include the record
ID (company name, ID#, etc.), a date-time-stamp (could be by the year or day or to the second,
e.g. 2016-2-25), and an event category (patent, company launched, IPO, etc.). This information
about single point events can be assembled into records with a dozen or a thousand events.
(Table 1)

[Table 1 about here]

Temporal event sequences also include interval events, such as a one-year SBIR grant, a research
project or clinical trial, in which case the event will have a start and an end date-time-stamp.
(Table 2).

[Table 2 about here]

Initial efforts are usually to clean the data, which often contains incorrect, incomplete, redundant,
miss-labeled, or surprising inputs. Typical errors include blank fields, erroneous record ID,
misspelled event category, incorrect date-time-stamp, or a start date that is later than an end date.
Visual displays amplify human abilities to spot errors such as outliers in a scatterplot, surprising
spikes in a timeline, or missing links in a network diagram.

The second data challenge involves record matching and disambiguation across data sources.
For example, this project involves matching data from FDA approvals, clinical trials, patents,
research grants and other sources where EventFlow records correspond to individual products.
While products are named in the FDA databases and often in clinical trial data, those names
often do not appear in patent or research grant data. Federal agencies including the National
Institutes of Health (NIH) and the Food and Drug Administration (FDA) have produced some ad-hoc databases that help with some of this matching - allowing us to present some preliminary results in this paper - but much of this work remains to be done.

Once data has been cleaned and matched, standard algorithms for identifying volatile or stable periods in time lines can be used to speed analyses. The combination of visual displays and statistical methods brings great power to analysts.

**How Long Does Innovation Take?**

Innovation trajectories\(^5\) describe the sequences of innovation activities that translate initial and intermediate inputs into intermediate outputs and final outcomes. Like physical trajectories, innovation trajectories are functions of *innovation inputs* as well as *time*.

Innovation inputs include knowledge, talent and a product idea; intellectual property (IP; proof-of-concept / proof-of-relevance; entrepreneurship; and capital, for example. Each event in an innovation sequence uses innovation inputs and produces outputs or outcomes that in turn become intermediate inputs in later activities. *Entrepreneurial success* is the desired *outcome* and is defined herein as successful commercialization of a product resulting in the launch of a new product in the marketplace.

A useful empirical question is *how long do these innovation trajectories take?* The answer to this question has implications for public and private investment in innovation as well as public policy. For example, one open policy question is: *do innovation accelerators actually accelerate innovation and if so, by how much?* Policymakers considering the investment of public funds in

---

\(^5\) A *trajectory* in the context of measuring innovation is a path, progression or line of development resembling a physical trajectory - the curve along which a physical body moves through space. - Merriam-Webster
programs to accelerate want to know if such programs are effective before committing public funds (Dempwolf, Auer and Dippolito, 2014).

New temporal metrics for innovation will help future researchers answer many policy questions including those identified in the National Research Council’s 2014 report. Indeed, baseline measures may hold the key to developing a better class of metrics for innovation and its economic impacts. Realistic estimates of confidence intervals for the duration of innovation sequences could reduce certain types of investment risk, thus making more capital available for prototyping and commercialization activities.

Billions of dollars are invested in the commercialization of new products, however most of that money increasingly favors later-stage investments where there is greater certainty about the product’s potential success and how long investment capital will be tied up. The question of how to shrink the so-called Valley of Death and get more investment capital flowing into earlier stage investments has remained unanswered in business, economic development and public policy circles for many years. Event analytics may help shed some light on this problem, catalyzing significant economic impacts in the process.

**Focusing on Drugs and Medical Devices**

This paper demonstrates our analytic methods using drugs and medical devices, which is an important topic for which data is readily available because they are regulated products. We model innovation trajectories as sequences of events leading to the launch of a new products, which is the desired outcome for entrepreneurial success. Clinical trials and FDA approvals offer useful proxies for the commercialization process where available data is often limited. Certain FDA approvals may also provide useful proxies for product launch dates.
Event Analytics for Innovation Trajectories

EventFlow produces several event analytics and different visualizations that can help users understand innovation trajectories in new ways. By grouping similar event sequence patterns together, EventFlow provides users with descriptive statistics and visualizations for groups of records with the same sequence pattern. These have several uses:

**Descriptive Statistics (Metrics or Measures):** For most research projects the production of descriptive statistics is not cause for much excitement. However, in the case of innovation there are no clear metrics on how long innovation processes take.

**Visualization and Exploration of Sequence Patterns:** Understanding the compositions and frequencies of different sequence patterns may also yield new insights and frame better hypotheses. EventFlow provides tools for visually simplifying event sequences to reveal common and rare patterns (Monroe et al., 2013; Du et al., 2016).

**Theory Formation (Modeling):** A key goal for researchers is to develop and test theories so as to guide future activities. The well-established linear model of innovation (basic research leads to applied research, then product development, culminating in commercialization) has its followers, as well as many critics. Comparisons with alternative models such as the ABC principle (applied and basic combined) could advance understanding of what leads to more frequently successful outcomes (Shneiderman, 2016). It is fairly common practice in articles and presentations to show the linear model because of its simplicity, and then immediately state that in practice innovation rarely follows the linear model. The popular understanding of innovation might be improved by documenting the prevalence of the linear model and its alternatives.
**Hypothesis Testing:** Event analytics can be as simple as seeing if event type A occurs more frequently before or after event type B, for example do patents precede or follow founding of companies. Another simple question is: how soon after founding a company do companies release a product? A refined version of this question is to see the distribution of times between founding a company and releasing a product.

There are more sophisticated questions that can be posed in event analytic tools, such as Do companies with three or more patents before product launches have more successful outcomes than companies with fewer patents?

**Modeling and Measuring Innovation Trajectories: Data and Examples**

The following examples are based on a dataset comprised of 34,331 records each representing a specific drug or medical device. Each record contains the events – research, patents, clinical trials and FDA approvals – associated with that product. In total the model includes 85,690 events. The list of event types and the count of each type is shown at the bottom of the left EventFlow panel shown in figure 2.

As a practical matter, answering the question *how long does innovation take* requires identifying start and end points. In our first example we take the date of first patent application as the starting point and a reasonable proxy for the date that the initial product idea was first conceived. Limiting our analysis to drugs and medical devices, we take the date of final FDA approval as the end date and a reasonable proxy for product launch date. Neither the dates that commercial ideas were originally conceived, nor the actual product launch dates are reliably recorded or made publicly available, thus the need for proxies.
The datasets available for modeling STI processes (see table 3) have several current limitations, and much of the work yet to be done under this study involves cleaning, matching, transforming and linking existing datasets. We present two preliminary examples that demonstrate some of event analytics capabilities of EventFlow (www.cs.umd.edu/hcil/eventflow), and which suggest the methods and kinds of final results we might expect when all of the data cleaning and matching is completed.

The first example models and analyzes the trajectories starting with clinical trials and ending with last FDA approval for 2,402 medical devices. Clinical trial success is typically a necessary input for final FDA approval. In certain cases, successful results in early stage trials may be sufficient for provisional, temporary approval, allowing the drug or device to be deployed prior to completion of the full set of clinical trials. The preliminary results of this second analysis demonstrates EventFlow’s ability to simplify the visualization of the dataset in ways that suggest overarching patterns in the data and allow researchers to pose clear, simple questions for further investigation. In this case, the visualization shows two distinct groups in the data: one in which the FDA approval is received after clinical trials are completed; and one in which FDA approval is received during the clinical trials (see Figures 3 and 4). The visualizations suggest several additional research questions, demonstrating EventFlow’s usefulness as a tool for data exploration.

The second example analyzes drug innovation trajectories from first patent to last FDA approval for 884 drugs resulting in mean, median and standard deviation metrics for these trajectories (see Figure 5 and 6).

Data gathering for Innovation Trajectories
We use the EventFlow software to model innovation trajectories in drugs and medical devices from multiple datasets:

[Table 3 about here]

**A Brief Introduction to EventFlow**

Based on work with 40+ case study projects, we find that point and interval events provide sufficient richness to describe the records in most applications. Point events have a Record ID, a categorical event name, and a time stamp. Interval events have a Record ID, a categorical event name, a start time, and an end time. Each point or interval event can have attributes, such as a patent category or a clinical trial director’s name. Eventflow constructs a record by collecting all the events that have the same Record ID. When a dataset is loaded, EventFlow shows the records in a timeline view, with time moving from left to right. The records are shown in a scrolling Timeline window (rightmost panel in Figure 2) sorted by Record ID with the lowest value at the top. Within each record, point events are shown as triangles with a distinct color for each point event type. Interval events are shown as colored horizontal lines with bars on the ends.

The center panel aggregates individual records into a summary overview showing patterns in how events are related to one another within records. The aggregated display starts with the most common first event at the top left. The records with the higher frequency of an event name will be grouped first and shown as a vertical bar whose height indicates the number of records with that sequence. The grouping by common event names continues from left to right till all events are shown. Point events are shown as a vertical bar, where the distance to adjoining vertical bars shows the average time between the events. Interval events are shown by a rectangular region, whose width is the average duration of the intervals. Complete explanations are in the user
manual, which includes many videos demonstrating the use of EventFlow

[Figure 2 about here]

Example 1: Medical Device Clinical Trials and FDA Approvals

Figure 3 shows Clinical Trials → FDA Approval for 2,325 medical devices. The EventFlow overview panel reveals two common patterns. For just over half the records, FDA approval was received on average 2 years 8 months AFTER the end of clinical trials (upper cohort). In just under half the records, FDA approval was received DURING clinical trials. Several EventFlow tools were used to clean up and simplify the visualization without altering the underlying data model.

[Figure 3 about here]

Figure 4 shows Clinical Trials → FDA Approval for 2,325 medical devices. With the same underlying model as depicted in figure 3, this image shows the exploration of event distributions for two non-adjacent time points – the start of clinical trials to final FDA approval. While the overall duration for the upper cohort averages 6 years and 10 months, we can quickly see from the time scale bar that the duration from the start of clinical trials to FDA approval in the lower cohort is about two years shorter. Overall duration of clinical trials is considerably longer in the lower cohort. These simple graphics immediately provoke and/or frame several research questions. Our intent here is not to answer or even ask those questions, but rather to demonstrate the power of event analytics in facilitating that process.

[Figure 4 about here]

Example 2: From First Patent Applications to Final FDA Approval
Figure 5 shows events from First Patent → FDA Approval for 688 drugs. The overview panel reveals that there are 6 main sequence patterns between these two events. The predominant pattern covering nearly half the records involves a period of patenting for several years followed by a gap, followed by FDA approval. Presumably clinical trials and other activities are taking place as well between first patent and final FDA approval. However, three-way data matching across FDA, Clinical Trials and Patent databases has yet to be done.

[Figure 5 about here]

Figure 6 shows First Patent → FDA Approval for 688 drugs. The question of how long it takes to get a new drug to market is most often answered by rules of thumb or anecdotal evidence. This image is among the first to actually show statistics and a distribution, with average duration of 9 years 4 months for two prevalent event sequence patterns. These results are preliminary. Additional cleaning and matching of the data along with the augmentation of record attributes may allow for useful confidence intervals to be generated by, for example, segmenting the sample according to drug class or other attributes.

[Figure 6 about here]

Discussion and Future Directions

This paper presents a new tool and novel approach for temporal analysis of innovation trajectories using examples and data from drug and medical device activities. While significant data processing work remains to match events from multiple datasets to product records, the brief examples shown in this paper suggest that temporal analysis of innovation trajectories with EventFlow can yield valuable information about the structure of innovation processes and new statistical metrics of how long these activities and processes take.
Innovation processes have social, spatial, technological and temporal characteristics. Quantitative analyses using geospatial and social network methods have yielded many useful insights and a variety of quantitative methods have been applied to understanding and visualizing the technological dimension of innovation. However most temporal analyses have been less robust. The development of a new statistical temporal baseline and metrics helps solve this problem and facilitates many new types of analyses.

As the clinical trial → FDA approval example suggested, innovation processes where FDA approval is obtained during clinical trials appear to shorten time-to-market by about two years\(^6\). That same analysis raises obvious questions about the two types of processes. Why is there a two- to three-year lag in the upper group between completion of the clinical trials and FDA approval? Are the FDA approvals in the lower group qualitatively different from those in the upper group? For example, are they “preliminary” or “fast-track” approvals? Are the devices in the upper group qualitatively different from those in the lower group? What are the implications for science and regulatory policy? Expanding product-based temporal analyses beyond drugs and medical devices will allow exploration of questions regarding how differences in the sequences of activities impact innovation outcomes across a range of different technologies.

Other seemingly simple questions where the metrics developed using EventFlow could help include:

- Do innovation accelerators actually accelerate innovation? That is, do they shorten the duration of the innovation process from idea to market?
- Do regions with higher innovation network density innovate faster? What network structures are associated with faster innovation?

\(^6\) Results are preliminary. Additional data validation work is in progress.
Both are active research questions for the authors. Regarding accelerators, a 2014 study of innovation accelerators for the U.S. Small Business Administration found no good metrics in the literature that answered the question of whether accelerators did indeed accelerate innovation (Dempwolf, Auer, and Dippolitto, 2014). A subsequent network analysis comparing outcomes between 77 accelerator-affiliated startups and 77 non-accelerator-affiliated startups receiving angel funding using found that the accelerator subnetwork was 8.5 times larger than the unaffiliated angel network and exhibited more opportunity for brokerage. Accelerators invested 33% less per startup in angel funding ($100K vs $150K) and 50% less overall ($1.3B vs 2.6B) than unaffiliated angels. Combined their startups raised an additional $41B in subsequent funding rounds and acquisitions (Dempwolf, 2014). While these results suggest that accelerator-affiliated startups may be more efficient, they do not answer the question of whether the accelerator-related startups achieved those results faster than non-accelerator startups. A pending EventFlow offers the potential to answer that question using the same dataset (CrunchBase) as the 2014 study.

The question of whether regions with higher network density innovate faster was recently embedded in a successful funding application for the National Institute for Innovation in Manufacturing Biopharmaceuticals (NIIMBL) under the National Institute of Standards and Technology (NIST). The authors will use EventFlow and NodeXL to model the network structure and innovation outcomes of NIIMBL partners and others in multiple regions throughout the U.S. over the next 5 years to answer this and other related questions.

Current Data Limitations
As promising as the preliminary results are, several data limitations are hindering broader application of this temporal analysis technology to understanding and measuring innovation processes.

1. Data is typically not collected or organized around *products* as the end result of innovation. Product data is available for drugs and medical devices because they are regulated and tested by product name. Otherwise, products are typically not identified in STI data sources. One data source that associates product names with the firms that produce them is the UPC database. The dates associated with UPC records are the date the record was last updated, not the date of product launch, however the source is worth further investigation.

2. STI data resides in multiple unlinked administrative databases and data quality is variable. Data cleaning, matching and disambiguation is a significant, time consuming and ongoing task. Records are not always complete and augmentation may be necessary. Efforts to automate data preparation processes through machine learning and other algorithms are underway but this will still take time.

3. Innovation processes are comprised of many different events and those events may involve different networks of people and organizations. Finding the relationships between events is not always easy.

4. Technology topics have not been standardized across the various types of events, although there have been numerous advances in topical analysis and natural language processing.

5. Data remains incomplete.

6. FDA Drug databases and medical device databases are structured differently and contain different information. For example, medical devices may be linked to clinical trials, but there are no linkages between drugs and clinical trials. Drugs may be linked to patents, but there are no linkages between medical devices and patents.

7. Applying this methodology to other critical industry sectors may be useful. Cleantech and energy, for example, share many similarities with medical devices in terms of inputs, outputs, innovation trajectories, regulations, and challenges. The Lab-to-Market initiative and the Department of Energy's Office of Energy Efficiency and Renewable Energy may offer comparable data to help overcome the identified data challenges.

**Conclusions**

This preliminary exploration of using time stamped event data to understand innovation trajectories shows promising possibilities. Even basic descriptive data reporting can substantially
advance the capacity for evidence-based decisions by policy makers, investors, and entrepreneurs. Key goals include a better understanding of what inputs produce more reliably successful outcomes.

While geospatial, multi-variate, time series, hierarchical, and network data analyses are widely used, event analytics represent a fruitful new path for researchers. As reliable datasets with temporal event sequences become more widely available, these event analytic approaches seem likely to produce valuable results that could speed innovation trajectories and make successful outcomes more common.
References.


<table>
<thead>
<tr>
<th>Record ID</th>
<th>Event Category</th>
<th>Start Date</th>
<th>End Date</th>
<th>Attributes</th>
</tr>
</thead>
</table>
| ALTOPREV  | Patent         | 12/12/1997 |          | docnum="5916595";Organization="Andrx Pharmaceuticals, Inc."
| ALTOPREV  | Patent         | 12/12/1997 |          | docnum="6485748";Organization="Andrx Pharmaceuticals, Inc" |
| ALTOPREV  | FDA Approval   | 6/26/2002  |          | docnum="N21316";Organization="COVIS PHARMA SARL" |
| AMYVID    | Patent         | 8/5/2008   |          | docnum="8506929";Organization="UNIVERSITY OF PENNSYLVANIA" |
| AMYVID    | FDA Approval   | 4/6/2012   |          | docnum="N202008";Organization="AVID RADIOPHARMACEUTICALS" |
Table 2: Sample span events

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Event Category</th>
<th>Start Date</th>
<th>End Date</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMYVID</td>
<td>Research</td>
<td>7/1/2003</td>
<td>2/28/2013</td>
<td><code>docnum=&quot;R01AG022559&quot;;Organization=&quot;UNIVERSITY OF PENNSYLVANIA&quot;</code></td>
</tr>
<tr>
<td>ALTOPREV</td>
<td>Research</td>
<td>4/1/1996</td>
<td>2/29/2000</td>
<td><code>docnum=&quot;R01NS033325&quot;;Organization=&quot;CHILDREN'S HOSPITAL BOSTON&quot;</code></td>
</tr>
<tr>
<td>Drug &amp; medical device data sources</td>
<td>Drugs/Devices</td>
<td>Notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>---------------</td>
<td>------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RePORTER_PATENTS_C_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RePORTER_CLINICAL_STUDIES_C_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTTI AACT Database</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FDA Orange Book</td>
<td>drugs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drugs@FDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Market Approvals (PMA)</td>
<td>med devices</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Pending and potential data sources

<table>
<thead>
<tr>
<th>Source</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBIR/STTR</td>
<td>Pending</td>
</tr>
<tr>
<td>CrunchBase</td>
<td>Potential (proprietary)</td>
</tr>
<tr>
<td>NSF</td>
<td>Pending / potential</td>
</tr>
</tbody>
</table>

### Supporting and core data sources

<table>
<thead>
<tr>
<th>Source</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NIH RePORTER</td>
<td></td>
</tr>
<tr>
<td>PatentsView</td>
<td></td>
</tr>
<tr>
<td>USASpending</td>
<td></td>
</tr>
<tr>
<td>STARMETRICS</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1 (a) Choropleth map: biomedical – pharmaceutical hot spot analysis by county, 2009. Analysis by Zhi Li, University of Maryland. Data Source: StatsAmerica (http://www.statsamerica.org/); (b) Spatial hot spot analysis of job concentrations in Professional, Scientific and Technical Services in Maryland, 2014. Source: Dempwolf, C. et.al. (2015); (c) and (d) Spatial distribution and concentration of innovative companies in Howard County, MD Source: Analysis and graphics by Cole Greene in Dempwolf, C. et.al. (2015).
Figure 1 (cont.) (e) Time evolution of the community structure of the network of citations between papers published in journals of the American Physical Society (APS). Time is divided into nine decades, from 1927 until 2006. In each decade, the most cited papers were selected (about 3,000). The communities are classified based on the APS journal where the largest relative fraction of papers in the community were published (indicated by the symbols). While links between different decades usually involve consecutive periods, there are five links connecting well-separated scientific ages (thick edges in the figure). From Chen and Redner (2010). Source: Scharnhorst, Börner, and Besselaar, 2012. P274 (prepub copy); (f) Network model of Regenerative Medicine Cluster in Howard County, MD 2010 – 2015. Source: Dempwolf, C. et.al. (2015).
The EventFlow user interface consists of three panels. The Control Panel on the left displays model information along with formatting and processing options. The Timeline Panel on the right displays event timelines for individual records, along with tabs for searching and filtering records based on events and attributes. In the center is the Overview Panel which aggregates records based on event sequence patterns, providing a condensed graphical representation of those event patterns.
Clinical Trials → FDA Approval for 2,325 medical devices. The EventFlow overview panel reveals two common patterns. For just over half the records, FDA approval was received on average 2 years 8 months AFTER the end of clinical trials (upper cohort). In just under half the records, FDA approval was received DURING clinical trials. Several EventFlow tools were used to clean up and simplify the visualization without altering the underlying data model.
Figure 4 Clinical Trials → FDA Approval for 2,325 medical devices. With the same underlying model as depicted in figure 3, this image shows the exploration of event distributions for two non-adjacent time points – the start of clinical trials to final FDA approval. While the overall duration for the upper cohort averages 6 years and 10 months, we can quickly see from the time scale bar that the duration from the start of clinical trials to FDA approval in the lower cohort is about two years shorter, while the overall duration of clinical trials is considerably longer in the lower cohort.
Figure 5 First Patent --> FDA Approval for 688 drugs. The overview panel reveals that there are 6 main sequence patterns between these two events. The predominant pattern covering nearly half the records involves a period of patenting for several years followed by a gap, followed by FDA approval. Presumably clinical trials and other activities are taking place as well between first patent and final FDA approval. However, three-way data matching across FDA, Clinical Trials and Patent databases has yet to be done.
Figure 6  First Patent --> FDA Approval for 688 drugs. The question of how long it takes to get a new drug to market is most often answered by rules of thumb or anecdotal evidence. This image is among the first to actually show statistics and a distribution, with average duration of 9 years 4 months for two prevalent event sequence patterns. These results are preliminary. Additional cleaning and matching of the data along with the augmentation of record attributes may allow for useful confidence intervals to be generated by, for example, segmenting the sample according to drug class or other attributes.