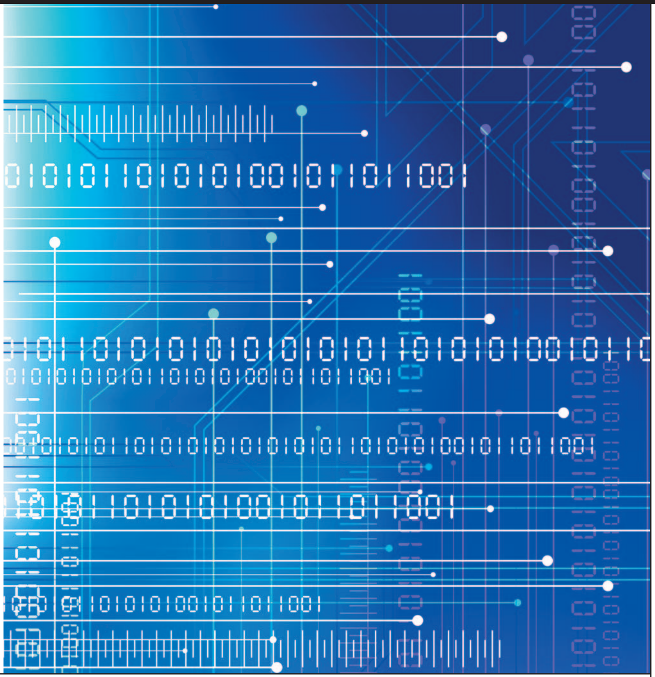


Portfolio Mining

Krishna P.C. Madhavan, Mihaela Vorvoreanu, and Niklas Elmqvist, *Purdue University*

Aditya Johri, Naren Ramakrishnan, and G. Alan Wang, *Virginia Tech*

Ann McKenna, *Arizona State University*

Portfolio mining facilitates the creation of actionable knowledge, catalyzes innovations, and sustains research communities.

Governmental investments in science, engineering, and education are key enablers to social and economic well-being and global competitiveness. In the US, key grant-making bodies such as the National Science Foundation (NSF), the National Institutes of Health (NIH), and the National Institute of Standards and Technology (NIST) meet this need.

Members of these organizations, in collaboration with members of other agencies and of the scientific community, make decisions about who and what to fund. In recent years, federal investments in R&D have exceeded US$140 billion annually (www.nsf.gov/statistics/infbrief/nsf10327). The NSF alone receives more than 50,000 proposals and makes more than 10,000 research awards each year (www.nsf.gov/about/budget/fy2013/pdf/04_fy2013.pdf).

Given this vast enterprise, it becomes extremely difficult for an organization, let alone an individual, to derive any meaningful insights into the nature of these investments. Furthermore, because of the unintended consequences of investments and the nonlinear path of scientific research, the impact of the original investments that enable future discoveries is hard to ascertain. Given the difficulty of quantifying the benefits of scientific investments, criticism of research investments is common (Y. Doz, J. Santos, and P. Williamson, *From Global to Metanational: How Companies Win in the Knowledge Economy*, Harvard Business School Press, 2001).

Enabling federal agencies to understand their R&D investments in meaningful ways is not only critical to setting policy for scientific progress, it also helps reassure an electorate about the proper stewardship of their tax dollars. More fundamentally, there is a need to understand federal investment portfolios—that is, to treat the entire discovery enterprise and associated work as a single data ecosystem for the advancement of science and learning.

PORTFOLIO MINING

Portfolio managers at federal institutions receive a large number of proposals on a daily basis that they must review and then decide to fund or decline. Simultaneously, they receive a massive influx of reports from the projects that they manage, usually containing information about potentially new and exciting research activities and results. In addition, portfolio managers are tasked with presenting complex, detailed analyses of their work to management and government oversight bodies that can be extremely time-consuming to conduct.

In November 2010, a subcommittee of the NSF Computer and Information Science and Engineering (CISE)/Social, Behavioral, and Economic Sciences (SBE) Advisory Committee released a report titled *Discovery in a Research Portfolio: Tools for Structuring, Analyzing, Visualizing, and Interacting with Proposal and Award Portfolios*. This report pointed to the tremendous manual work performed by NSF program officers in managing proposals and awards, and highlighted the need for advanced data mining and visualization techniques to synthesize

0018-9162/12/$31.00 © 2012 IEEE

and organize knowledge in ways that increase productivity and identify innovations early.

Portfolio mining allows for a comprehensive understanding of how funding decisions have resulted in the current research ecology.

Numerous national and international projects seek to undertake portfolio mining on a massive scale. For example, Science and Technology for America's Reinvestment (STAR) Metrics (www.starmetrics.nih.gov) is a collaborative effort of multiple US federal agencies and academic institutions to "build a scientific data infrastructure that brings together inputs, outputs, and outcomes from a variety of sources in an open as fashion as possible" (J. Lane and S. Bertuzzi, "Measuring the Results of Science Investments," *Science*, 11 Feb. 2011, pp. 678-680).

However, managers will not use a central portfolio mining resource just because it is available (J. Brazelton and G.A. Gorry, "Creating a Knowledge-Sharing Community: If You Build It, Will They Come?," *Comm. ACM*, Feb. 2003, pp. 23-25). Large-scale, national efforts such as STAR Metrics are needed, but they are also inherently slow. As these projects ramp up, the data deluge will feed a vicious cycle of data complexity and bloat.

While STAR Metrics and other similar efforts use data mining and visualization tools to shed light on scientific investments at a national level, the sheer scope of such efforts makes it difficult to obtain immediately relevant results. Program managers need tools that optimize the process of deriving insights about prior investments and expedite the diffusion of key findings and new ideas to a large community of researchers and observers.

Portfolio mining platforms and tools also have a critical community building and capacity expansion aspect. Much of the knowledge generated by communities such as those fostered through any of the NSF programs is contained within their networks and social structures. Visually understanding and characterizing these connections helps to make the hidden knowledge structures useful and beneficial to the community.

DIA2

In October 2011, a joint collaboration of Purdue University, Virginia Tech, Arizona State University, and Stanford University initiated a new project—Deep Insights Anytime, Anywhere (DIA2)—that offers a potential model for advancing the state of the art of portfolio management and the rapid prototyping of associated tools.

Portfolio mining platforms and tools have a critical community building and capacity expansion aspect.

Funded by the NSF's Education and Human Resources (EHR) directorate, this Web-based knowledge mining and interactive visualization platform is designed to be a central resource for the science, technology, engineering, and mathematics (STEM) education community. Although DIA2's main focus is on the set of projects funded by the Transforming Undergraduate Education in STEM (TUES) and predecessor programs, its approaches are generalizable to the entire NSF. The project's guiding vision is to fundamentally change the way NSF portfolio managers access and utilize the results of years of research investments.

Traditionally, portfolio management platforms have emphasized the computational aspects of the work rather than the requirements of actual users. What makes DIA2 unique is its emphasis on the stakeholders that would use such a system on a daily basis—namely, program assistants, program officers, and administrators.

To inform its design decisions, the DIA2 team is actively engaging its user base. In the project's first year, 31 EHR program officers, science assistants, administrators, and technical staff participated in seven focus groups. Eight hours of one-on-one interviews supplemented these sessions. The DIA2 team also distributed diaries to the participants to capture data as they were working. Based on these user studies and analyses of the captured data, the DIA2 team prepared a set of typical user personas and mapped specific design requirements to their needs.

In collaboration with the IT staff, the DIA2 team designed an innovative, low-footprint, high-density hardware stack and deployed it within the NSF network. Due to privacy and legal constraints, the team was not permitted to access the actual datasets that would underlie the portfolio mining system. Instead, using provided metadata, it created a database schema and populated it with a combination of publicly available data and dummy datasets.

Working closely with this metadata-based schema and guided by the results of its in-depth user studies, the DIA2 set about creating an interactive, dashboard-based framework that is accessible only by approved NSF staff. This easy-to-use framework is designed to let program managers drag and drop widgets and customize the views of their portfolio.

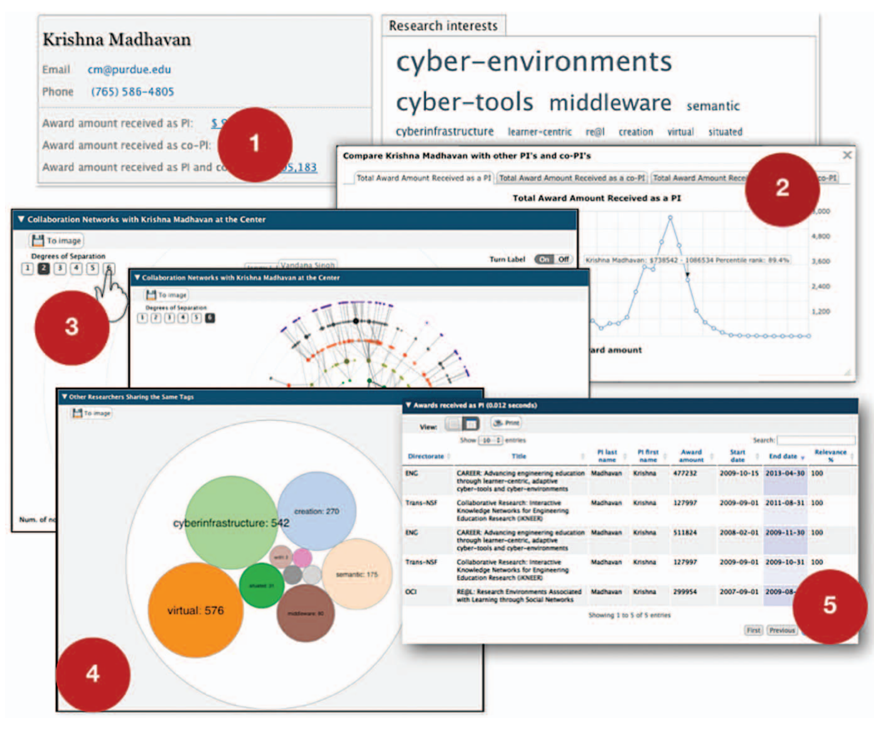
PORTFOLIO ANALYTICS

As DIA2 evolves, the development team envisions four broad classes of analytics capabilities: search and visualization, community and collaboration analysis, temporal modeling, and investment analysis.

Search and visualization

DIA2 provides basic, on-demand search and visualization capability to NSF program officers (or any interested user with access privileges).

As Figure 1 shows, a search of publicly available data about a principal investigator (PI) reveals basic aggregated as well as detailed information: aggregate awards received by the PI (window 1), award amounts (2), collaboration networks (3), tags denoting research areas (4), and a complete list of currently active and completed awards (5). The user can dive as deep into the data as desired, controlling the entire exploration process along the way.


Krishna Madhavan

Email cm@purdue.edu

Phone (765) 586-4805

Award amount received as PI: $ 9

Award amount received as co-PI:

Award amount received as PI and co 5,183

▼ Awards received as PI (0.012 seconds)

View: Print

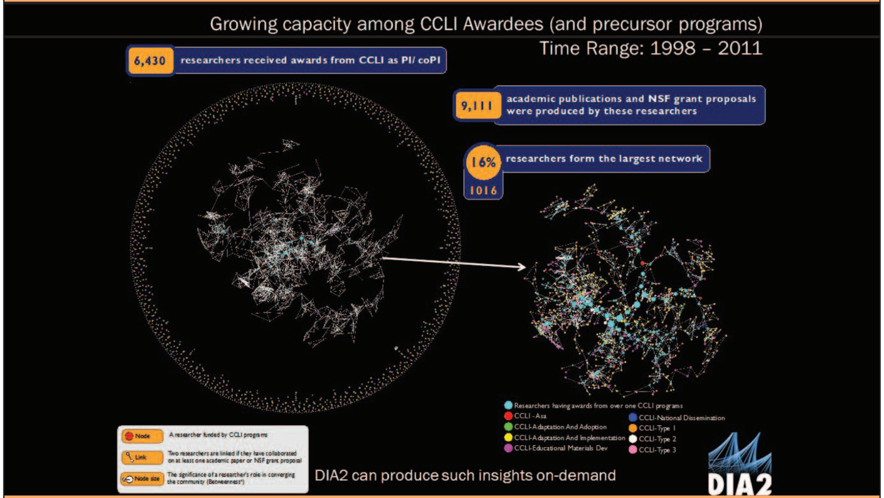
Show 10 entries Search:

Directorate	Title	PI last name	PI first name	Award amount	Start date	End date	Relevance %
ENG	CAREER: Advancing engineering education through learner-centric, adaptive cyber-tools and cyber-environments	Madhavan	Krishna	477232	2009-10-15	2013-04-30	100
Trans-NSF	Collaborative Research: Interactive Knowledge Networks for Engineering Education Research (iKNEER)	Madhavan	Krishna	127997	2009-09-01	2011-08-31	100
ENG	CAREER: Advancing engineering education through learner-centric, adaptive cyber-tools and cyber-environments	Madhavan	Krishna	511824	2008-02-01	2009-11-30	100
Trans-NSF	Collaborative Research: Interactive Knowledge Networks for Engineering Education Research (iKNEER)	Madhavan	Krishna	127997	2009-09-01	2009-10-31	100
OCI	RE@L: Research Environments Associated with Learning through Social Networks	Madhavan	Krishna	299954	2007-09-01	2009-08-	

Showing 1 to 5 of 5 entries

First Previous

Figure 1. Search and visualization in DIA2. A search of publicly available data about a principal investigator reveals numerous insights. The user can dive as deep into the data as desired, controlling the entire exploration process along the way.


Growing capacity among CCLI Awardees (and precursor programs)

Time Range: 1998 – 2011

6,430 researchers received awards from CCLI as PI/ coPI

9,111 academic publications and NSF grant proposals were produced by these researchers

16% researchers form the largest network

1016

Node: A researcher funded by CCLI programs

Link: Two researchers are linked if they have collaborated on at least one academic paper or NSF grant proposal

Node size: The significance of a researcher's role in converging the community (Betweenness*)

DIA2 can produce such insights on-demand

Researchers having awards from over one CCLI programs
CCLI - Asa
CCLI-Adaptation And Adoption
CCLI-Adaptation And Implementation
CCLI-Educational Materials Dev
CCLI-National Dissemination
CCLI-Type 1
CCLI-Type 2
CCLI-Type 3
DIA2

Figure 2. Community and collaboration analysis in DIA2. Example of interactive, on-demand analysis of the collaboration network among recipients of awards made by the NSF's CCLI (Course, Curriculum, and Laboratory Improvement) program.

Community and collaboration analysis

DIA2 enables NSF program officers and other staff to obtain a systems-level view of the collaboration that occurs within their portfolio. Figure 2 illustrates the type of interactive, on-demand analyses that the system provides—in this case, an analysis of awards made by the NSF's CCLI (Course, Curriculum, and Laboratory Improvement) program.

By changing the resolution of such collaboration graphs—that is, by adjusting the scope and range of the data—users can gain new insights into the often complex hidden structures that exist within a knowledge community.

Temporal modeling

Temporal modeling aims to understand the evolution of topics (in award abstracts) over time and visualize them longitudinally for important trends. It is likely that these trends are not independent but rather correlated—for example, funding for one area could potentially be at the expense of another. Toward this end, the DIA2 team is exploring correlated topic models but, rather than using predefined structures, is characterizing the correlation structures implicit in the data and designing algorithms that model these structures.

One interesting aspect of topic evolution is how topics subdivide or coalesce over time. There could be multiple explanations for such topic regroupings in the NSF—for example, programs become specialized, alter their focus due to funding changes, or become absorbed by new programs that cover the same area. The goal is to capture such trends computationally so that an NSF portfolio manager would be able to inspect the results and reason whether they are consistent with programmatic changes or represent a shift in the research marketplace.

The DIA2 team is also approaching topic evolution from the perspec-

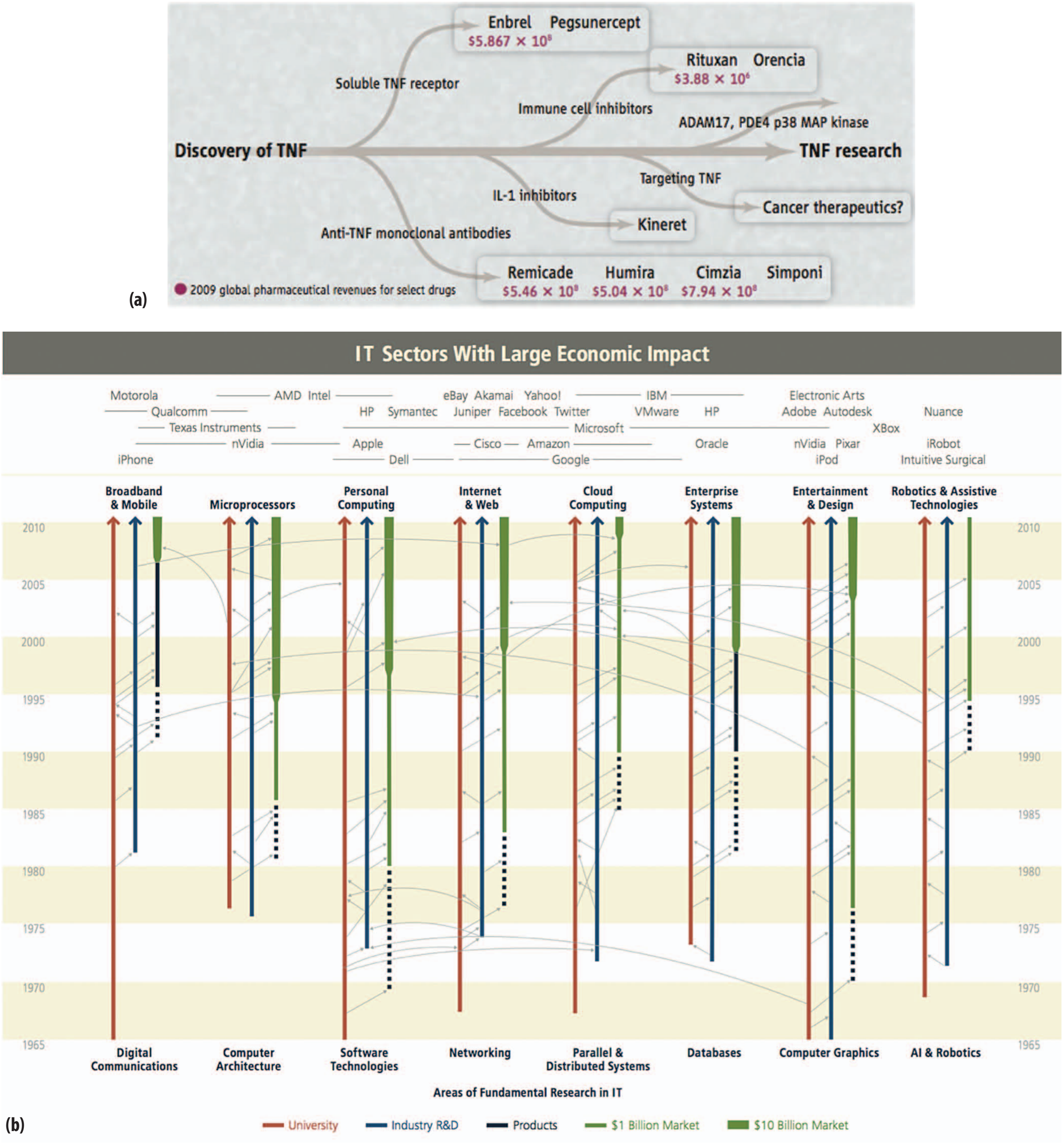


Figure 3 Data mining for investment analysis. (a) Evaluating the impact of the discovery of the tumor necrosis factor (TNF) on the drug industry. (Source: J. Lane and S. Bertuzzi, "Measuring the Results of Science Investments," *Science*, 11 Feb. 2011, pp. 678-680. Reprinted with permission.) (b) "Tire tracks" model of university-industry R&D-product reinforcements. (Source: National Research Council (NRC) Computer Science and Telecommunications Board, *Continuing Innovation in Information Technology*, National Academies Press, 2012. Reprinted with permission.)

tive of time-series segmentation. For instance, given a 10-year time course of projects, which years were the important transition points? Do these points correspond to the creation of a new program, the community gathering around a research methodology, or important changes in administration? Typical time-series segmentation analyses focus on a single time series, whereas program managers are interested in analyzing a broad collection of abstracts and other information over time.

Investment analysis

Ultimately, one of the greatest benefits of data mining is providing "big picture" summaries of portfolios and their impact as a function of investments made over time. In Figure 3a, for example, mining patent data can yield insights into the impact of the discovery of the tumor necrosis factor (TNF), a class of cell signaling molecules responsible for programmed cell death (apoptosos), on the drug industry. Figure 3b illustrates the popular "tire tracks" model, used to capture mutually reinforcing developments between university R&D, industrial R&D, and industrial growth.

Automatically extracting such summaries is a promising area of research that DIA2 hopes to leverage in the near future.

By enabling principal investigators, program managers, and administrators to interactively synthesize, mine, and visualize data, portfolio mining facilitates the creation of actionable knowledge, catalyzes innovations, and sustains research communities. The DIA2 team envisions the framework eventually becoming a central resource for the entire STEM education community. Work is already underway to systematically bring to light knowledge artifacts extracted from data and enable users to comment on and critique them through Web-based tools. C

Krishna P.C. Madhavan *is an assistant professor in the School of Engineering Education at Purdue University. Contact him at cm@purdue.edu.*

Mihaela Vorvoreanu *is an assistant professor of computer graphics technology and technology leadership and innovation at Purdue University. Contact her at mihaela@purdue.edu.*

Niklas Elmqvist *is an assistant professor in the School of Electrical and Computer Engineering at Purdue University. Contact him at elm@purdue.edu.*

Aditya Johri *is an assistant professor in the Department of Engineering Education, College of Engineering, Virginia Tech. Contact him at ajohri@vt.edu.*

Naren Ramakrishnan*, Discovery Analytics column editor, is the Thomas L. Phillips Professor of Engineering at Virginia Tech and director of its Discovery Analytics Center. Contact him at naren@cs.vt.edu.*

G. Alan Wang *is an assistant professor in the Department of Business Information Technology, Pamplin College of Business, at Virginia Tech. Contact him at alanwang@vt.edu.*

Ann McKenna *is an associate professor in the Department of Engineering, College of Technology and Innovation, Arizona State University. Contact her at ann.mckenna@asu.edu.*

The work reported here was partially supported by National Science Foundation awards TUES-1123108, TUES-1122609, and TUES-1123340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.