Discovery in a Research Portfolio

Tools for Structuring, Analyzing, Visualizing and Interacting with

Proposal and Award Portfolios

November 2010

Final Report on Recommendations from

National Science Foundation CISE and SBE AC Subcommittee[1]

[1] The Committee would like to thank the efforts of the many people who gave unstintingly of their time with this challenging effort. These include NSF staff Jean Feldman, Jannele Gosey, Maxine Hynson, Leslie Jensen, Paul Markovitz, Matt Powell, Karen Santoro, Amanda Shea, Greg Smith, Jean Thomas, Jenn Thornhill and NORC staff Tim Mulcahy and Johannes Fernandes-Huessy. Many NSF program officers and science assistants attended the meetings and provided input, as well as representatives from other agencies. Ned Talley from NIH provided extraordinarily useful input and suggestions throughout the process.

Executive Summary

A core part of NSF's mission is to keep the United States at the leading edge of discovery. It does this by funding research in traditional academic areas that has both high intellectual merit and broader impact. It also funds transformative and interdisciplinary research. In order to effectively do the former, program managers need to identify the appropriate reviewers and panelists to ensure the best possible peer review of proposals, and manage their portfolio of awards. In order to effectively do the latter, program managers need to identify and describe emerging areas and research topics in collections of proposals. Beyond the duties of individual program managers, NSF also has substantial reporting requirements that require providing adequate descriptions of its scientific portfolio and its outcomes, at the program, division, directorate and agency level, to key interested parties. Although NSF staff still rely on traditional, highly manual, methods to do their jobs, such methods are becoming less practical given the rapidly changing nature of science, the increased recognition of the importance of funding interdisciplinary and potentially transformative research, and the significant increase in the number of proposals submitted.

At the same time, computational methods including machine learning, visualization, human-computer interaction, and link structure analysis—areas of effort funded by core NSF programs over the last several decades--hold opportunities for providing NSF leadership with new insights and "optics" into its investments and portfolio, and the relationship of its portfolio to the status and dynamics of scholarship and innovation.

This report describes the efforts and recommendations of an Advisory Subcommittee of the Advisory Committees for the Social, Behavioral and Economic (SBE) Sciences and the Computer and Information Science and Engineering (CISE) Directorates. The members of the Subcommittee included experts in machine learning, data mining, information visualization, human-centered computing, science policy, and visual analytics. The SBE/CISE Advisory Subcommittee was co-chaired by members of the CISE (Stu Feldman, Eric Horvitz, and Vijay Raghavan), and SBE (Ira Harkavy and Jeff MacKie-Mason) Advisory Committees, and guided under the leadership of NSF program managers (Julia Lane and Mary Lou Maher).

The Subcommittee participants were charged with identifying and demonstrating techniques and tools that could characterize a specific set of proposal and award portfolios. In addition, the subcommittee was asked to provide recommendations on ways that NSF could better structure existing data, make use of existing machine learning, analysis, and visualization techniques to complement human expertise and better characterize its programmatic data.

As a part of this work, ten research teams, each directed by a member of the subcommittee, participated in an experimental effort in which NSF proposals were analyzed in a secure environment, using the latest techniques for information analysis and visual analytics.   The teams interacted with program managers from NSF and from other parts of the federal government in order to produce their demonstration analyses.

The key findings of the subcommittee are:

1. Proposals should be re-structured to allow automatic extraction of key data.
2. There is great potential to combine data from multiple internal and external sources that could be used to describe and benchmark NSF investments. This includes the potential to use topic modeling to describe current, emerging and potentially transformative areas of research.
3. There is great potential to identify "nodes" of science – peoples, topics and documents – and their interrelationships. The nodes can be used to better characterize, categorize, and cluster proposals, and hence facilitate the identification of reviewers as well as conflicts of interest. The interrelationships can be used, inter alia, to identify funding gaps, unexpected missing links, and interdisciplinary opportunities.
4. Visualization techniques that graphically represent data clusters and linkages have the potential to assist in recognizing data patterns and providing rapid access to data.
5. Any system should be user focused in its design, allowing usage minimal training, and be engineered to complement human expertise.

The subcommittee developed two sets of recommendations: the first set for structuring, extracting and searching portfolio/proposal data and the second set for deriving and making sense of knowledge. These are:

1. NSF should change the ways in which it structures its existing database of proposals by using data extraction techniques. NSF should consider developing techniques for capturing proposal data in a more structured format in the future.

2. NSF should develop approaches to deriving a knowledge level representation of the proposal data and a set of tools that users can use to make sense of this knowledge for open ended tasks and queries.

Motivation

A core part of NSF's mission is to maintain the United States' position at the leading edge of discovery by funding research in traditional academic areas, identifying broader impacts, and by supporting "high-risk, high pay-off ideas." This requires describing research portfolios along a variety of dimensions: by scientific disciplines, geography, award size, diversity, and broader impact.

At the programmatic level, program managers need to stay abreast of emerging areas and research topics in their fields and in collections of proposals, identify the appropriate reviewers and panelists to ensure the best possible peer review of proposals, and manage their portfolio of awards. At the agency level, NSF has substantial reporting requirements to key interested parties that require providing adequate descriptions of its scientific portfolio and its outcomes, based on inputs from programs, divisions, and directorates.

Traditional methods for accomplishing these objectives are becoming less sufficient given the rapidly changing nature of science, the increased recognition of the importance of funding interdisciplinary and potentially transformative research, and the significant increase in the number of proposals submitted. Manual approaches have been labor intensive and error prone.

Advances in information technology, notably in data management, text mining, topic extraction, and visualization offer substantial opportunities to do better. Improvements should be possible in two areas. The first of these is deriving knowledge of interest from selected data, such as clusters of research topics; emerging areas; geographic distribution of research investment; discipline distribution of research investment and impact; diversity of principal investigators, graduate students, post doctoral fellows; multi-disciplinary collaboration; and broader impact. The second is making sense of that knowledge by, for example, assessing the amount of interdisciplinary research; conveying the information in a consistent manner to program managers; and conveying the information in a consistent manner to external audiences.

Approach

A joint CISE- SBE subcommittee was formed to develop recommendations to improve the way NSF staff work with its proposal and award portfolio. The subcommittee was composed of one or two members each from the CISE and SBE Advisory Committees, plus researchers in machine learning, data mining, information visualization, human-centered computing, science policy, and visual analytics. Thus the goal was to leverage the CISE and SBE relationships with members of their respective scientific communities, who have interest and expertise in relevant technologies and tools for information extraction, social network, and scientometric analysis. The teams interacted with both NSF program managers and program managers in other parts of the federal government.

The subcommittee was charged with identifying and demonstrating techniques and tools that characterize a specific set of programmatic portfolios, including proposals, award abstracts, and research and education outcomes. The subcommittee was asked to identify tools and approaches that are most effective in deriving knowledge from the data provided, i.e., most robust in terms of permitting program officers to visualize, interact, and understand the knowledge derived from the data.

The subcommittee began this process with a planning meeting in September 2009 during which alternative tools and techniques were discussed, and advice given on principles for selecting the data to be studied. In order to provide the best possible advice to NSF, the researchers on the subcommittee requested access to a dataset comprising specific parts of a large collection of proposal jackets. NSF provided the project description, summary, cover page and references from all proposals available in electronic form from 2006 to 2010.

The subcommittee members were given access to the data on a secure data enclave. The legal authorization for this access was determined after repeated consultations with the Office of the General Counsel (OGC), the Division of Information Systems (DIS) and the Division of Grants and Agreements (DGA). This resulted in the following determinations:

1. The use of the subcommittee was reasonable for the purpose described, since they are working on behalf of NSF. The system of records notice provided sufficient notice to PIs that NSF may use proposals for the evaluation of NSF procedures.
2. Graduate students could have access to Privacy Act protected information data for the purpose of doing analysis that provides advice to NSF. Contracts with the graduate students will include the standard Privacy Act FAR clauses applying Privacy Act section 552a(m).
3. DIS required that confidential data are protected by NIST medium level security standards. The approach taken went far beyond the minimum protections necessary to protect confidentiality. It employed a portfolio approach, using multiple methods to protect the data. The conceptual framework provided a safe setting, safe data, safe people and safe outputs to protect confidentiality. That meant that the following protections were in place:
    a. Technical – the environment met the NIST standards adhered to by NSF for the protection of confidential microdata
    b. Legal – it enforced NSF confidentiality standards
    c. Educational – there was training of all users of the data
    d. Operational – Only authorized researchers on approved projects had access. They were only allowed to do statistical analysis. There were operationally defined safeguards to protect the handling of the data
    e. Statistical – NSF deidentified the data as required. All output was disclosure proofed before being disseminated

4. NSF developed subcontracts with the subcommittee members. Those contracts included the following features
   a. Confidentiality agreements (for researchers and graduate students)
   b. Ability to publish results
   c. Funding to support graduate students
   d. Deliverables – a report that was provided to the subcommittee

Three virtual meetings of the subcommittee and NSF program officers were held; one in June and two in September of 2010, to report results of the analysis of publicly available award abstract data. A final meeting was held on October 12 and 13, in which members described the results of the analysis of proposal data to the Subcommittee leadership group, and formulated recommendations for NSF.

Committee Work and Findings

The committee's work is summarized in the white papers (Appendix 1). There were five major findings.

1. ***Proposals should be re-structured to allow automatic extraction of key data.***

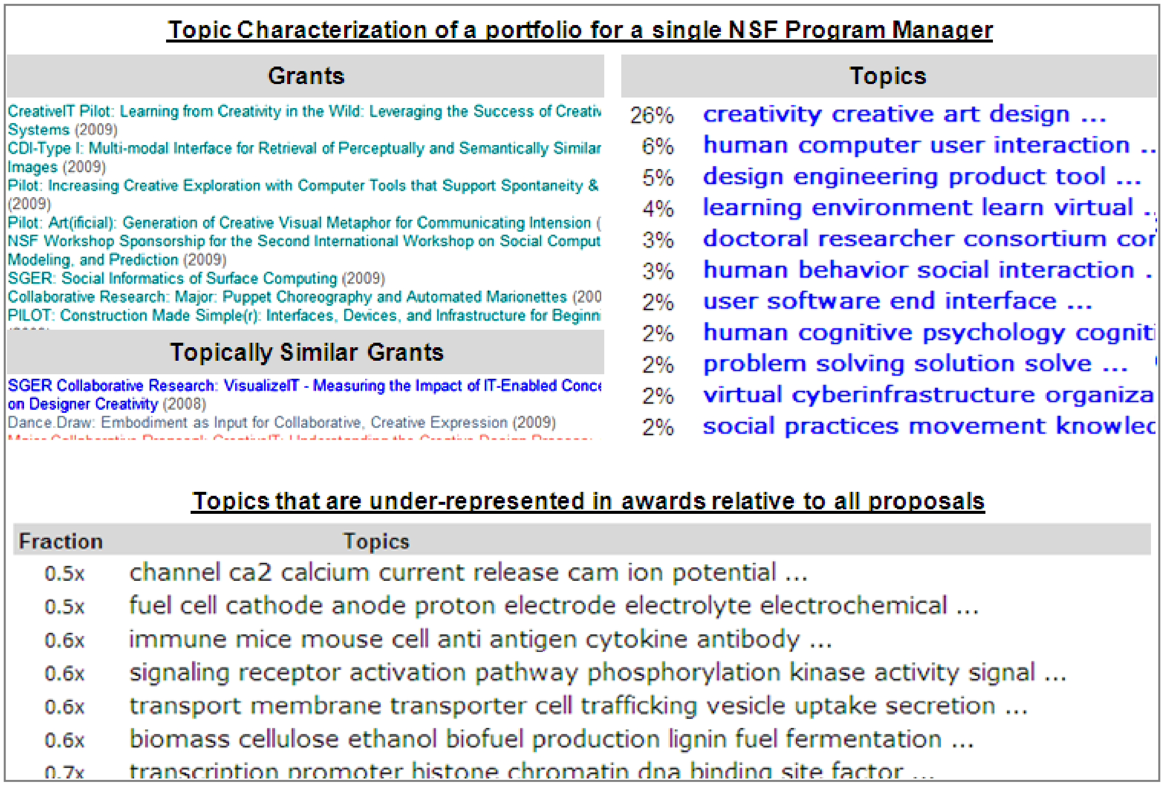
Relevant information is currently distributed across different formats and systems. Proposal data are currently stored in both structured (database) and unstructured (pdf) formats. Program managers currently access proposal data through ejacket for proposal content and review information, and EIS for statistical data derived from the proposal data.

2. ***There is great potential to combine data from multiple internal and external sources that could be used to describe and benchmark NSF investments. This includes the potential to use topic modeling to describe current, emerging and potentially transformative areas of research.***

The various teams identified multiple approaches for utilizing the text and citations extracted from proposals. They demonstrated applications that derived different types of information and they showed how this information could be used in various contexts relevant both to day-to-day NSF operations and strategic assessments of broader NSF goals.

The teams found that the text of proposals could be distilled using a combination of key-phrase extraction and topic modeling. The value of the topic modeling approach lies in reducing an intractably complex set of many thousands of concepts into much more manageable groupings. Since these topics are derived from the proposals themselves they reflect the discourses that investigators use to describe their research, and hence can be used to analyze and understand the full research portfolio.

The topic modeling tool from David Newman compiled the grants in a searchable interface in which program managers can assess the topic content of proposals in their respective programs, and access related proposals that are highly relevant but not in their program (Figure 1). In addition, Newman co-analyzed the text from NIH and NSF grants to provide information on their areas of overlap.   The Leskovec team used topic modeling to assess the historical relationship between awards (using public abstracts) and published literature, for a preliminary assessment of the areas in which NSF has been a trend leader, vs. areas in which NSF has provided support for trends that are already in emergence. Using a key-phrase approach, the Chen team applied text segmentation to pinpoint core information in proposals (Figure 2).



Topic Characterization of a portfolio for a single NSF Program Manager

Grants	Topics	
CreativeIT Pilot: Learning from Creativity in the Wild: Leveraging the Success of Creativ Systems (2009)	26%	creativity creative art design ...
CDI-Type I: Multi-modal Interface for Retrieval of Perceptually and Semantically Similar Images (2009)	6%	human computer user interaction ..
Pilot: Increasing Creative Exploration with Computer Tools that Support Spontaneity & (2009)	5%	design engineering product tool ...
Pilot: Art(ificial): Generation of Creative Visual Metaphor for Communicating Intension (	4%	learning environment learn virtual ..
NSF Workshop Sponsorship for the Second International Workshop on Social Comput Modeling, and Prediction (2009)	3%	doctoral researcher consortium cor
SGER: Social Informatics of Surface Computing (2009)	3%	human behavior social interaction .
Collaborative Research: Major: Puppet Choreography and Automated Marionettes (200	2%	user software end interface ...
PILOT: Construction Made Simple(r): Interfaces, Devices, and Infrastructure for Beginn	2%	human cognitive psychology cogniti

Topically Similar Grants

SGER Collaborative Research: VisualizeIT - Measuring the Impact of IT-Enabled Conce on Designer Creativity (2008)	2%	problem solving solution solve ...
Dance.Draw: Embodiment as Input for Collaborative, Creative Expression (2009)	2%	virtual cyberinfrastructure organiza
	2%	social practices movement knowlec

Topics that are under-represented in awards relative to all proposals

Fraction	Topics
0.5x	channel ca2 calcium current release cam ion potential ...
0.5x	fuel cell cathode anode proton electrode electrolyte electrochemical ...
0.6x	immune mice mouse cell anti antigen cytokine antibody ...
0.6x	signaling receptor activation pathway phosphorylation kinase activity signal ...
0.6x	transport membrane transporter cell trafficking vesicle uptake secretion ...
0.6x	biomass cellulose ethanol biofuel production lignin fuel fermentation ...
0.7x	transcription promoter histone chromatin dna binding site factor ...

Figure 1: Topic Modeling to Characterize the Entire Set of NSF Proposals

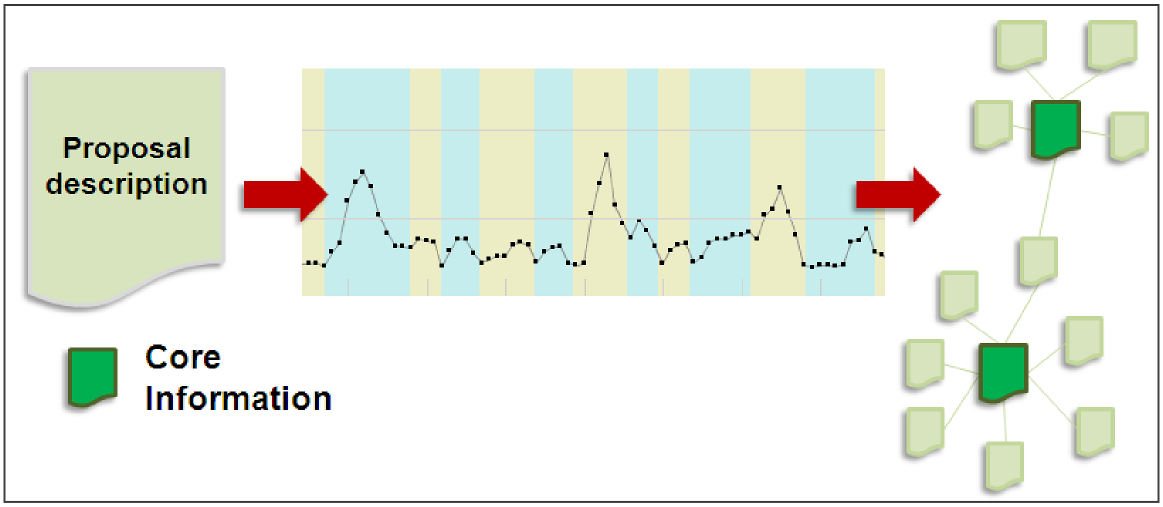


Figure 2: Text Segmentation to Pinpoint Core Information in a Proposal

3. ***There is great potential to identify "nodes" of science – peoples, topics and documents – and their interrelationships. The nodes can be used to better characterize, categorize and cluster proposals, and hence facilitate the identification of reviewers as well as conflicts of interest. The interrelationships can be used, inter alia, to identify funding gaps, missing links and interdisciplinary opportunities.***

Various teams made use of information from proposals and the publications they cite to place investigators, concepts and proposals within larger networks. The resulting interrelationships can be used to produce a large and densely connected dataset in which gaps, links and interdisciplinarity can be identified. For example, the Contractor team presented a tool for visualizing specific links between people, documents and concepts (Figure 3).

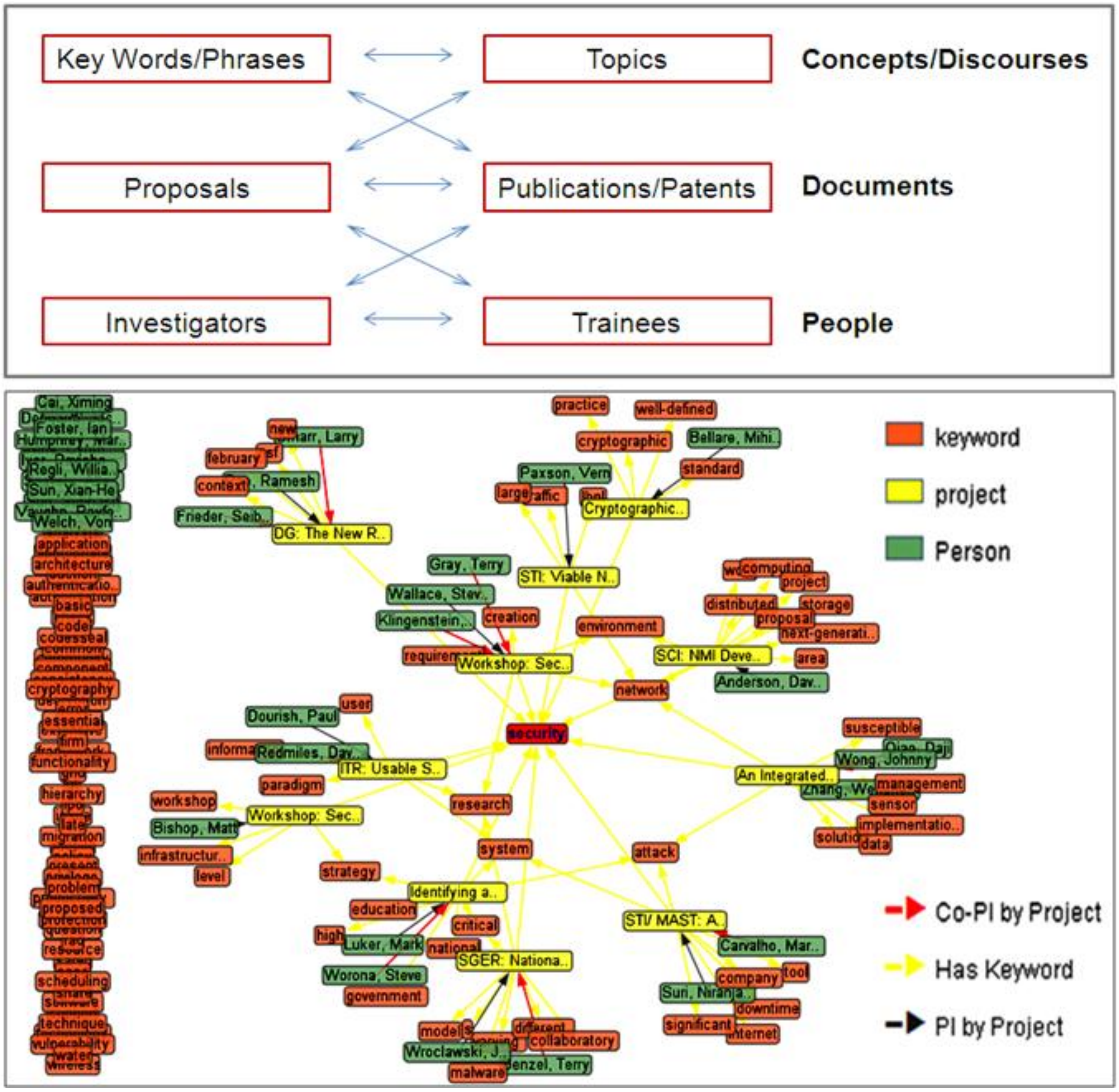


Figure 3: Links between people, documents and concepts

4. ***Visualization techniques that graphically represent data clusters and linkages have the potential to assist in recognizing data patterns and providing rapid access to data.***

The teams recommended developing the capability for interactive and automated generation and visualization of concept and document clusters. For example, the Raghavan team showed tree map based concept hierarchies to permit an analysis of portfolio structures (Figure 4). The Ribarsky team combined automated and interactive document clustering that would allow program managers to dynamically interact with their portfolios (Figure 5). The North team demonstrated the potential for document clustering allows users to determine the relevant dimensions by moving documents in an interactive spatially defined process.

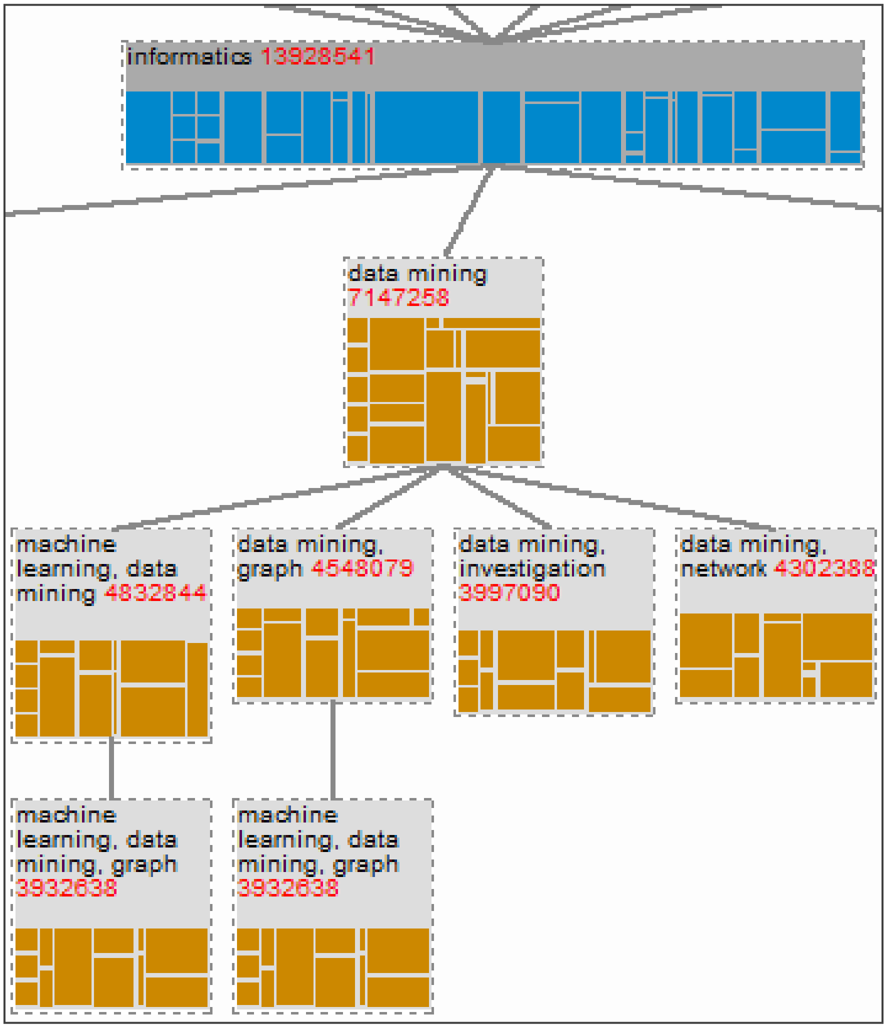

informatics 13928541
data mining 7147258
machine learning, data mining 4832844
data mining, graph 4548079
data mining, investigation 3997090
data mining, network 4302388
machine learning, data mining, graph 3932638
machine learning, data mining, graph 3932638

Figure 4: Treemap View of Concept Hierarchies

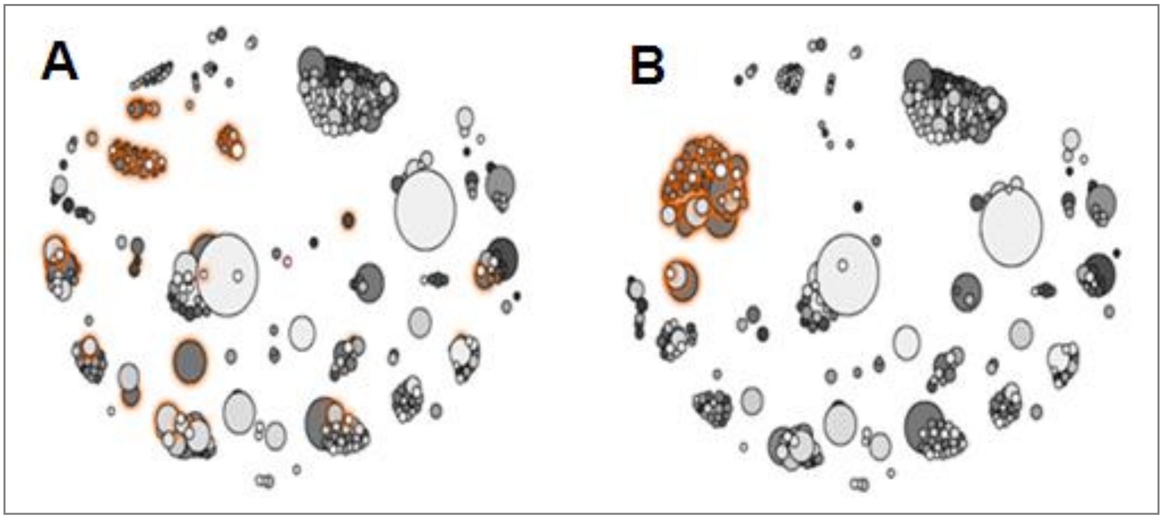
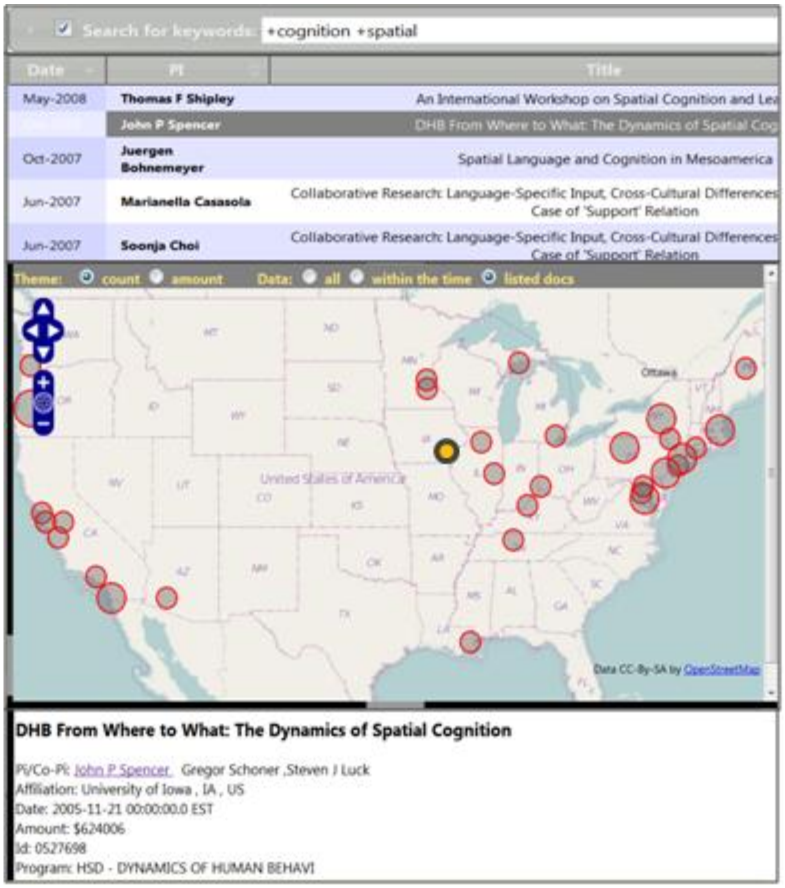


Figure 5: Combined Automated and Interactive Document Clustering
A: Documents are clustered in a two-dimensional graphed output for visualization. The user selects a theme of interest, causing relevant documents to be highlighted with gold halos. B: The user increases the importance of this theme for document clustering, which results in highlighted proposals moving close to each other in the resulting organizational framework.

The teams also recommended visualizing portfolios both geospatially and temporally For example, the MacEachren team showed the value of geospatial representations of researchers/institutions (Figure 6). Interactive frameworks for temporal analyses were demonstrated by both the Ribarsky and Chen teams.



☑ Search for keywords: +cognition +spatial

Date	PI	Title
May-2008	Thomas F Shipley	An International Workshop on Spatial Cognition and Lea
	John P Spencer	DHB From Where to What: The Dynamics of Spatial Cog
Oct-2007	Juergen Bohnemeyer	Spatial Language and Cognition in Mesoamerica
Jun-2007	Marianella Casasola	Collaborative Research: Language-Specific Input, Cross-Cultural Differences Case of 'Support' Relation
Jun-2007	Soonja Choi	Collaborative Research: Language-Specific Input, Cross-Cultural Differences Case of 'Support' Relation

Theme: ● count ○ amount Data: ○ all ○ within the time ● listed docs

DHB From Where to What: The Dynamics of Spatial Cognition

PI/Co-PI: John P Spencer , Gregor Schoner ,Steven J Luck
Affiliation: University of Iowa , IA , US
Date: 2005-11-21 00:00:00.0 EST
Amount: $624006
Id: 0527698
Program: HSD - DYNAMICS OF HUMAN BEHAVI

Figure 6: Geographic Distribution of Awards

5. ***Any system should be user focused in its design, with minimal training.***

The white paper by the Madhavadan team highlighted the importance of a user focus in system design, and proposed features that any system should aspire to. Such features include interfaces that do not require manuals and training, and that make sense to users by employing familiar metaphors for commonly operations. In addition, the team recommended that proprietary technology and third party installations be avoided, and that systems be mobile compatible and future ready, by using such technologies as HTML5 and CSS.

Recommendations

The subcommittee developed two sets of recommendations: the first set for structuring, extracting and searching portfolio/proposal data and the second set for deriving and making sense of knowledge

1. NSF should change the ways in which it structures its existing database of proposals by using data extraction techniques. NSF should consider developing techniques for capturing proposal data in a more structured format in the future.

2. NSF should develop approaches to deriving a knowledge level representation of the proposal data and a set of tools that users can use to make sense of this knowledge for open ended tasks and queries.

Detailed recommendations for structuring, extracting and searching portfolio/proposal data

Recommendation S1: NSF should develop strategies to structure proposal data beyond the cover sheet, and beyond pdf formats for collecting and analyzing proposal data.

Recommendation S2: NSF should provide better database access to its portfolio data. Specific recommendations are:

- NSF should provide direct, simple search tools for database access to raw, extracted and derived data.
- NSF should provide API access to the raw data to allow users to customize their search tasks.
- NSF should consider an IMAP-like interface to data, XML-DB formats, RDF, and open source DBMS.
- NSF should provide a common, unified data repository to avoid proliferation of duplicated repositories and tools to reduce data quality errors: fix it in one place, fixed for everyone.
- NSF should consider providing an easily available public data source in order to encourage researchers to invent better methods/tools that can be used by NSF in the future.

Recommendation S3: NSF should implement data conversion and data extraction tools. Specific recommendations are:

NSF should implement more capable tools for converting pdf to text that are specialized for proposal data and include tokenization, indexing, domain specific data capture using OCR for mathematical and chemical formulae.

NSF should consider the following sources when extracting data:

- Proposal data: Cover page, Project Summary, Project Description, References
- Annual reports
- Biographies (including "Conflict of interest")
- External data: relevant publication data

NSF should give special consideration to the following issues related to data extraction:

- Show confidence in automated extraction output because there will be errors.
- Some additional fields are important on a per-domain basis.
- Chemical formula, mathematical equations, biological processes may require special techniques.

Recommendation S4: NSF should reconsider the proposal data format in the future. NSF should develop a proposal mark up language that identifies key aspects of proposal information in a structured form including ideas, methods, people, key relationships in the data, and allow meta-data formats for special data such as chemical equations, graphs, and citations.

NSF should make changes to the proposal submission process to structure data more easily. Proposals can be submitted in a more structured format, with separate sections such as:

- General fields: people, institutions, publication venues, proposal data, project description
- Proposal subfields: program, title, PIs, summary, $ amount, dates
- People subfields: first name, last name, title, email, current and past institutions
- Project description subfields: objectives, methodology, equipment, citations, tables, images, diagrams

Recommendation S5: NSF should allow Program Managers and other staff to add meta-data to the proposal data, such as user "tags", tag clouds, comments, and other semantic information.

Recommendations for deriving and making sense of knowledge

Recommendation D1: NSF should develop structural and temporal visualization tools that can be used to explore and examine multiple types of entities and their interrelationships such as investigators, research topics, and publications.

This will allow NSF to better:

- Understand proposals across topics and people
- Analyze PIs’ areas of research
- Reveal and highlight missing links across people and topics
- Organize and make use of lessons learned
- Find opportunities that have not been addressed in the topics and networks
- Maintain independence between derived knowledge and the structured data

Recommendation D2: NSF should develop ways to combine data external to the proposal system with proposal data. Some suggestions of how to facilitate this include:

- Constructing and maintaining social/collaboration networks among NSF PIs to explore the value and outcome of the proposals..
- Building a comprehensive researcher id and information system to better locate potential reviewers.

- Building NSF keyword set and semantic network to produce a more relevant view of the content of proposals, the scope of programs, and the changes of research trends: The keyword analysis and the structure of keyword networks highly depend on the set of keywords included. Recent developments in Web 2.0 technologies as well as the Semantic Web could be leveraged in this effort.
- Mapping to existing patent, publication and citation data through such programs as STAR METRICS to help identify emerging fields.
- Use the co-authorship, citation, cocitation, citation similarity, and collaboration networks to integrate comprehensive information and accurately characterize the interactions among the researchers. This information should help to answer many questions (such as COIs and level of collaboration) from a social network perspective.

Recommendation D3: NSF should incorporate a tool box of techniques that allows staff to better search and retrieve proposal content. Several tools and techniques are demonstrated in the white papers attached to this report. A subset of the categories of these techniques include:

- Topic modeling tools to identify the research topics.
- Social network tools that can analyze interdependencies between people and projects.
- Citation analysis tools to assess the impact of various research projects.

Recommendation D4: NSF should develop interactive tools to enable users to better perform tasks that are exploratory in nature. Some suggestions are:

- Direct manipulation and propagation should be facilitated with easy to use sliders that span sensitive variables.
- Maintain logs of user interaction, use data from the user to fine tune the tool for that user, and beware of problems with using logs of user data.
- Include information about level of confidence from perspective of the tool.
- Allow users to trace back from original topics, and to access information on data provenance.
- Build a knowledge base of queries and results, users and their user experiences, with a time stamp, finding repeating questions, and learn from the past.
- Identify interactions that are local to a user and interactions that are relevant to the Foundation.
- Encourage and enable collaboration among program managers.

Appendix A: Subcommittee members

CISE AC members:

Stu Feldman, Google
Eric Horvitz, Microsoft Research
Jeff MacKie-Mason, University of Michigan
Vijay Raghavan, University of Louisiana at Lafayette

SBE AC members:

Ira Harkavy, University of Pennsylvania
Jeff MacKie-Mason, University of Michigan

Researchers with a computing perspective:

Lada Adamic, University of Michigan
James Allen, University of Rochester and Florida Institute of Human & Machine Cognition,
David Blei, Princeton University
Katy Börner, Indiana University
Chaomei Chen, Drexel University
Hsinchun Chen, University of Arizona
Noshir Contractor, Northwestern University
C. Lee Giles, Pennsylvania State University
Jure Leskovec, Stanford University
Andrew MacCallum, University of Massachusetts (Amherst)
Alan MacEachren, Pennsylvania State University
Krishna P. C. Madhavan, Purdue University
Albert Mons, Concept Web Alliance
David Newman, University of California Irvine
Chris North, Virginia Polytechnic Institute
Peter Pirolli, Palo Alto Research Center
Ben Schneiderman, University of Maryland
Jim Thomas, Pacific National Laboratory (passed away)

Researchers with a science policy perspective:

Izja Lederhendler, National Institutes of Health
Chuck Lynch, National Institutes of Health
Tim Hays, National Institutes of Health
Dorothy Miller, Environmental Protection Agency
Edmund Talley, National Institutes of Health
Bill Valdez, Department of Energy

Appendix B: Summary of White Papers

Lee Giles, Pennsylvania State University

This team developed tools to extract machine-readable structured metadata from various types of unstructured data sources such as a PDF collection of proposals and publications, so that further visualization and analysis can be performed.

The team focused on extracting three types of information from the collection of proposals:

1. Cover Page Metadata. This places all metadata in the cover page in appropriate XML fields.
2. Citations and References. While parsing these reference strings at the end of a document is often straightforward for human readers, the sheer diversity of different standards used by different communities, coupled with inadvertent errors on the part of proposers, makes this process difficult to automate.
3. Key phrase Extraction. Automatically identifying important concepts in a body of text will enhance the effectiveness in visualizing and understanding of data.

David Newman, University of California Irvine

This team applied topic modeling (LDA) to create a unified topic basis for a wide variety of analyses.

The team found that topic representation provides an immediate structure to compare, contrast and combine proposals. Furthermore, topics are a convenient basis for both querying and reporting. Finally, topics are a useful basis for visualizations, both in terms of computing relations between proposals, and annotating and color coding visualizations. In summary, the benefits of topic modeling are:

- Topic modeling automatically learns categories that describe research ideas
- Topics are learned directly from words used in text (title, abstract, full text proposal)
- Topic modeling does not need dictionaries, thesauri, ontologies, or other categorization schemes
- Learned topics are usually meaningful, intuitive and coherent
- Topics convert words/text into real-valued measure (so can measure proportions, aggregates, trends, etc.)
- Topic modeling is mature and well-researched
- Topic modeling is highly scalable (can topic model millions of documents in minutes)
- Topics can be used to create useful reports
- Topics can be the analytic basis for structuring, organizing and understanding sets of NSF grants

Katy Börner and Angela Zoss, Indiana University

This team explored different temporal, topical, and network analysis methods to identify evolving and emerging populations and topics. The team applied a subset of the 180 algorithm plugins in the Science of Science Tool (http://sci2.slis.indiana.edu) on the NSF awards and proposals data. The RefMapper plugin was used to analyze the interdisciplinarity of cited prior work, i.e., journals. Results are communicated via a science map data overlay to:

- Explore the interdisciplinarity of proposal sets as an indicator of emerging areas and to understand how various science fields are interlinked.
- Examine the range of topics in a given award portfolio.
- Assess the amount of interdisciplinary research.
- Identify areas that are funded by multiple directorates or divisions, and
- Identify emerging areas for future funding solicitations.

This team also considered temporal analysis to identify emerging trends based on topic bursts, i.e., sudden increases in the usage frequencies of terms or phrases, using Kleinberg's algorithm. This information could be used 'seed' activity via workshop money or pilot grants or consult key experts in these emerging areas when compiling new solicitations.

Network analysis algorithms were applied to study evolving collaboration structures at different levels of analysis, e.g., evolving co-author networks, co-investigator networks but also bimodal networks of authors and their institutions. Existing (successful) collaboration structures could be indicative of the future success of investigators and their proposals. The techniques applied by this team made use of Giles's parser and Newman's topic modeling results.

Chaomei Chen, Drexel University

The white paper prepared by this team included a set of questions that NSF may want to ask of its portfolio, and then focused on three main tasks:

- How to identify the core information and extract high-quality terms.
- How to differentiate awarded and declined proposals with survival analysis of the immediacy and persistency of hot topics in proposals.
- How to quantify the transformative potentials of proposal.

The team described 2 steps to identify core segments of text:

1. The first step is to divide a full-length project description into a series of passages of text. The internal cohesiveness of text within a passage is higher than between passages. The process is known as text segmentation.
2. The second step is to select the most representative segment(s) as the core information of a proposal. Once the text segments are identified in step 1, many existing techniques from the information retrieval community and machine-learning community in particular are available to compute the similarity between any two segments, including vector space models, latent semantic indexing, probabilistic models, and topic models.

The team described a process to find hot topics:

1. Use Part of Speech (PoS) tagging algorithms.
2. Detect frequency of specific noun phrases.
3. Do a burst analysis (Kleinberg) to detect hot topics.

The team explored ideas for identifying transformative (novel) research. They show how the nature of transformative research should be detectable along two of the computationally observable dimensions: synthesis distance and structural divergence.

Jure Leskovic, Stanford University

This team studied the lead/lag of topics in computer science between two corpora. They considered the funded grant abstracts of NSF and the ISI Web of Knowledge publications between 1991 and 2008. The main contribution of this work is that they propose an approach based on topic modeling (Blei's LDA algorithm) and time series analysis to compute the topic-specific lead/lag across corpora based on purely textual and time-stamp information. An additional complexity in this dataset is that each document can discuss multiple topics, and therefore one needs to decompose each document into its topics before analyzing them.

Noshir Contractor, Northwestern University

This team demonstrated the use of a visual-analytic tool that helps NSF Program Managers to:

1. Explore potential reviewers / COIs.
2. Build synergies among funded projects / PIs.
3. Identify new avenues of research / constellation of keywords.

They described a tool where program managers could simply provide the content of a proposal and the names of the PIs and you automatically receive the names of potential reviewers. The tool could be used to highlight potential conflicts of interests, such as previous collaboration on papers or funded proposals.

In order to accomplish the goal of the project this team employed a three-layer model: the person layer (PI / Co-PI), the artifact layer (proposal / publication) and the concept layer (keywords / subject categories). Using data from various sources, they build relation networks in each layer and between layers to reveal the connections among different PIs, proposals, and keywords.

For the concept layer, the team used CRAWDAD tool to extract the keywords from proposals' cover pages and summaries. Additionally, the team derived information such as keyword similarity based on co-occurrence in the same proposal and citation index (using UCINET tool).

All the information described above was uploaded into the C-IKNOW (Cyberinfrastructure for Inquiring Knowledge Networks On the Web), which is an interactive web-based software tool, developed with support from prior NSF-funded projects, for understanding and enabling knowledge networks.

Krishna Madhavan, Purdue University ; Aditya Johri, Virginia Polytechnic Institute

This team's goal was to research, design, and develop tools and services that allow program officers and other members of the stakeholder group to utilize sophisticated data mining tools while being shielded from the complexity of the underlying data structures.

These tools provide extremely interactive visualizations that allow users to dive as deep into the data as they want. This team's assumes a very simple, yet powerful user-centered, design philosophy summed up in the tagline: Deep Insights: No Manuals, No Training. This work leverages prior NSF-investments through the Interactive Knowledge Networks for Engineering Education Research (iKNEER) project. This team demonstrates:

- The ability to locate and retrieve any piece of information from the database system.
- Interactive, on-demand analyses of PIs, their funding profiles, and collaboration networks.
- Interactive analyses of programs and program officers.
- The users are easily able to control and interact with data at varying granularity without having to worry about the underlying data structures.

The team explores the tools required to answer the following questions:

- What are the dominant paradigms of broader impacts seen within the NSF portfolio?
- What are the dominant methodologies described in the proposals? When do they come into dominant use?
- What major projects funded by NSF are influential and prominently featured in other proposals?

Alan MacEachren and Jin Chen, Pennsylvania State University

This team describes GrantsForager, a prototype that illustrates the potential of being able to access geographic information associated with awards, to support cluster-seeded query and filtering with flexible user modification, and to support a process of information foraging enabled by visual feedback.

This supports quick filtering of NSF Awards Abstracts to find those that meet thematic, temporal, directorate, geographic, and other criteria. In addition, the methods and tools have the potential to support exploration of research themes as they vary geographically, differ by program, and change over time. The team specifies a list of questions that can be answered by Grants Forager:

Theme related:

- Given a new submission, who are potential reviewers?
- What awards and research topics are related to a particular theme (represented by some keywords), and in a user-specified time range?
- How do the topics and geographic distribution of awards vary by NSF program(s) or directorate?

Geographically related:

- How are NSF awards allocated across congressional districts?
- How are NSF awards focusing on particular research themes allocated across congressional districts (or states or other user-specified regions)?
- What are the awards and research topics in a particular congressional district (or state or other user-specified region)?
- How do these awards compare between congressional districts (or states or other user-specified regions)?

- What are the primary geographic places in which place-°©based NSF awards are focused?

Time related:

- What are the overall research topics for a user-specific time period?
- How do the topics vary for two user-specific time ranges, or between NSF program(s) for a particular time?

Vijay Raghavan, University of Louisiana at Lafayette; Ying Xie, Kennesaw State University; Tom Johnsten, University of South Alabama

The Proposal Information Management System developed by this team is named the Concept Map-based Organizer for Research Portfolios (C-MORE). C-MORE is designed to structure and manage the NSF proposal repository as an enterprise-wide resource. C-MORE consists of three main software layers that correspond to tools and technologies, respectively, for concept extraction, for concept organization that can support the application of a variety of automated analysis techniques, and for visualization of results in terms of relationships between concepts, proposals and researchers. Since text analysis technologies are rapidly evolving and user requirements for the types of analyses to be performed change over time, our strategy of studying these layers in an integrated way and the deployment of the proposed knowledge warehouse is expected to provide the flexibility required to accommodate such new demands and emerging technologies.

The C-MORE system features both top-down, analysis-driven navigation and query-based information exploration. It is able to provide decision support at both managerial and strategic levels. The functionality of C-MORE is defined in terms of three visualization constructs: concept cluster map, concept hierarchical map, and concept mesh map. A concept cluster map shows at the highest level the proposal distribution over a variety of research areas (or, topics). Research areas are extracted automatically from a targeted subset of proposals and each topic is represented by an atomic concept (a single term or phase). Users of the C-MORE system can zoom into a research area in the concept cluster map and drill-down to a variety of research subareas in a hierarchical manner. The drill-down results are presented in the form of a concept hierarchical map. A particular proposal can belong to multiple research subareas, thus it can be related from the top-level via multiple drill-down paths. A concept mesh map shows how concepts, representing topics, are related to each other within the target proposal set. It provides the user with information on what other concepts have strong linkages, the amount of funding allocated, and the number of funded / non-funded proposals as related to the selected concept.

In addition, operators, such as aggregation, difference and intersection have been developed to analyze concept mesh maps at different levels of the concept hierarchy. Comparison of concept mesh maps has the potential to provide valuable information to facilitate various decision makings tasks. For instance, an NSF staff member can generate an aggregated concept mesh map of proposals funded by the CISE program last year and an aggregated concept mesh map of proposals funded by the Biological Science program last year, and then apply the intersection operation on the two aggregated maps. The resulting concept mesh map will provide staff members with information on funded research projects in the interdisciplinary fields of computing and biology.

Jing Yang and William Ribarsky, University of North Caroline Charlotte; Remco Chang, Tufts University

This team developed a suite of tools that can be used effectively separately or, even more powerfully, together. The tools are built around a shared database structure so that results can be communicated among them for deeper analysis and understanding. Topic modeling, based on the work of Blei et al., is used in all the tools. However, this work goes beyond usual topic modeling capabilities in that it involves interactive visualization to make sense of, refine, and relate topics and then use the improved topics for deeper analysis. It also provides rich capabilities to study the development and emergence of topics over time.

The suite begins with an overview tool, Parallel Topics, for looking at an overview of topics for a group of programs. The tool clearly shows relations among multiple topics and permits study of their structure in detail. The user can then employ TopicGarden and STREAMIT to get more detailed understanding. These tools can be launched with their topic model results or with enhanced topic results and filtered sub-collections of proposals, project descriptions, or papers provided by Parallel Topics. TopicGarden provides a multiple topic view of the collection that provides an understanding of topics clusters that goes beyond the usual keywords. STREAMIT reveals temporal trends and semantic relationships within and among the clusters. Both tools provide uniquely powerful browsing and retrieval methods.