Discovery Informatics Workshop (DIW 2012)

February 2-3, 2012
Arlington, VA

Funded by NSF with grant IIS-1151951

http://diw.isi.edu/2012
### Workshop Participants

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>Field Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cecilia Aragon</td>
<td>U. Washington</td>
<td>Interaction and visualization</td>
</tr>
<tr>
<td>Phil Bourne</td>
<td>UC San Diego</td>
<td>Biology, future scientific publications</td>
</tr>
<tr>
<td>Elizabeth Bradley</td>
<td>U. Colorado</td>
<td>Qualitative reasoning</td>
</tr>
<tr>
<td>Will Bridewell</td>
<td>Stanford U.</td>
<td>Machine learning and discovery</td>
</tr>
<tr>
<td>Paolo Ciccarese</td>
<td>Harvard U.</td>
<td>Ontologies and semantic web</td>
</tr>
<tr>
<td>Susan Davidson</td>
<td>U. Pennsylvania</td>
<td>Databases and provenance</td>
</tr>
<tr>
<td>Helena Deus</td>
<td>Digital Enterprise Research Institute Ireland</td>
<td>Semantic web</td>
</tr>
<tr>
<td>Yolanda Gil</td>
<td>U. Southern California</td>
<td>Workflows and semantic web</td>
</tr>
<tr>
<td>Clark Glymour</td>
<td>Carnegie Mellon U.</td>
<td>Philosophy of science, causality</td>
</tr>
<tr>
<td>Carla Gomes</td>
<td>Cornell U.</td>
<td>Constraint reasoning and sustainability</td>
</tr>
<tr>
<td>Alexander Gray</td>
<td>Georgia Institute of Technology</td>
<td>Data mining and astrophysics</td>
</tr>
<tr>
<td>Haym Hirsh</td>
<td>Rutgers U.</td>
<td>Social computing</td>
</tr>
<tr>
<td>Larry Hunter</td>
<td>U. Colorado Denver</td>
<td>Natural language and biology</td>
</tr>
<tr>
<td>David Jensen</td>
<td>U. Massachusetts Amherst</td>
<td>Machine learning</td>
</tr>
<tr>
<td>Kerstin Kleese van Dam</td>
<td>Pacific Northwest National Laboratory</td>
<td>Semantic scientific data management</td>
</tr>
<tr>
<td>Vipin Kumar</td>
<td>U. Minnesota</td>
<td>Machine learning and climate</td>
</tr>
<tr>
<td>Pat Langley</td>
<td>Arizona State U.</td>
<td>Computational scientific discovery</td>
</tr>
<tr>
<td>Hod Lipson</td>
<td>Cornell U.</td>
<td>Robotics</td>
</tr>
<tr>
<td>Huan Liu</td>
<td>Arizona State U.</td>
<td>Social computing</td>
</tr>
<tr>
<td>Yan Liu</td>
<td>U. Southern California</td>
<td>Machine learning and biology</td>
</tr>
<tr>
<td>Miriah Meyer</td>
<td>U. Utah</td>
<td>Scientific visualization</td>
</tr>
<tr>
<td>Andrey Rzhetsky</td>
<td>U. Chicago</td>
<td>Genetics</td>
</tr>
<tr>
<td>Steve Sawyer</td>
<td>Syracuse U.</td>
<td>Social computing</td>
</tr>
<tr>
<td>Alex Schliep</td>
<td>Rutgers U.</td>
<td>Bioinformatics</td>
</tr>
<tr>
<td>Christian Schunn</td>
<td>U. Pittsburgh</td>
<td>Cognitive science and discovery</td>
</tr>
<tr>
<td>Nigam Shah</td>
<td>Stanford U.</td>
<td>Ontologies and semantic web</td>
</tr>
<tr>
<td>Karsten Steinhaeuser</td>
<td>U. Minnesota</td>
<td>Data mining and climate</td>
</tr>
<tr>
<td>Alex Szalay</td>
<td>The Johns Hopkins U.</td>
<td>Astrophysics and citizen science</td>
</tr>
<tr>
<td>Loren Terveen</td>
<td>U. Minnesota</td>
<td>Interaction and social computing</td>
</tr>
<tr>
<td>Raul E. Valdes-Perez</td>
<td>Vivisimo Inc.</td>
<td>Commercialization, knowledge-based discovery</td>
</tr>
<tr>
<td>Evelyne Viegas</td>
<td>Microsoft Research</td>
<td>Semantic computing</td>
</tr>
</tbody>
</table>
Outline

- Motivation for Discovery Informatics
  - Why now
- Possible Grand Challenges in Discovery Informatics
- Research challenges
- Vision scenarios for several domain sciences
  - Biology, social sciences, geosciences, astronomy
- Moving Ahead
Science Has a Never-Ending Thirst for Technology

- Computing is a substrate for science innovation
- Ongoing investments in cyberinfrastructure have a tremendous impact in scientific discoveries, e.g.,
  - Shared high end instruments
  - High performance computing
  - Distributed services
  - Data management
  - Virtual organizations
- These investments are extremely valuable for science, but do not address important aspects of modern science

http://diw.isi.edu/2012
Hallmarks of 21st Century Science

- **Discovery processes** are increasingly complex
  - Often involve many diverse elements (data, software, expertise, etc)
  - Processes remain largely human-driven
  - Human cognitive limitations become a bottleneck
  - Need new approaches to address this complexity

- **Data** is central to advances in many realms of science
  - Scientific knowledge is about models that predict and explain data
  - Need to increase our ability to connect knowledge/models to data

- Discovery is a **social** endeavor
  - Increasingly requiring amalgamation of diverse expertise and effort
  - Need technologies that can synthesize human abilities in all forms
Discovery Informatics: Emerging Themes

1. Computational support of the discovery process
2. Data and models
3. Social computing for discovery

http://diw.isi.edu/2012
Discovery Informatics: Why Now

- Discovery informatics will help **address the human bottleneck**
  - Cognitive limitations
  - Process efficiency
- Discovery informatics as “**multiplicative science**”: Investments in this area can be leveraged across science and engineering
  - Address current redundancy in {bio|geo|eco|…}-informatics
- Discovery informatics will both **empower and leverage the public**
  - “Personal data” will give rise to “personal science”
    - I study my genes, my medical condition, my backyard’s ecosystem
  - Harness the efforts of massive numbers and diverse types of individuals
    - Citizens as sensors, human labelers of data, human processors, etc
    - Games with a purpose
- Discovery informatics will **enable lifelong learning and training of future workforce**
  - Focus on usable tools that encapsulate, automate, and disseminate important aspects of state-of-the-art scientific practice

http://diw.isi.edu/2012
Outline

- Motivation for Discovery Informatics
  - Why now

- Possible Grand Challenges in Discovery Informatics

- Research challenges

- Vision scenarios for several domain sciences
  - Biology, social sciences, geosciences, astronomy

- Moving Ahead
Possible Grand Challenges for Discovery Informatics

1) a “Google” for scientists

- Top results are most relevant
- Search engine goes all over diverse open sites
- Each result is “hyperlinked” to data, models, processes, scientists, etc.
  - Highlights contradictions
- One stop shop across sciences, when drilling down on results the specialized tools come up

Cyclin E

Carbon rates Lake Mendota

Networks with abnormal Katz centrality

http://diw.isi.edu/2012
Possible Grand Challenges for Discovery Informatics

2) Assembling the “A-Team”

- Each movie is done by assembling a team from scratch
  - Many different skills are needed
- Social tools that take goals, find resources/expertise, shepherd subactivities
  - Dynamically assembled, as if we were producing a movie
- Reputation comes from the quality of work/tools/capabilities
- Computing support for big to medium to small science efforts
  - Big studio productions
  - Indie movies
  - Home movies

Director
Barbara Jones

Executive producer
Sandeep Jain

Producers
Matthew Baines and Li Cheng

Crane engineer
...

Casting
...

Make up
...

Director’s assistant
...

Special effects crew
...

Casting
...

http://diw.isi.edu/2012
Possible Grand Challenges for Discovery Informatics

3) The Scientist’s Butler

- Watches your work
  - What you did yesterday/last month
- Is aware of what others do
- Makes connections
- Suggests to you:
  - “I brought you an article that contradicts your results”
  - “I run your experiment with another dataset I found and result supports your theory”
  - “Would you want to try a method that was published last week in PNAS and is applicable to your data?”

http://diw.isi.edu/2012
Outline

- Motivation for Discovery Informatics
  - Why now

- Possible Grand Challenges in Discovery Informatics
  - Research challenges

- Vision scenarios for several domain sciences
  - Biology, social sciences, geosciences, astronomy

- Moving Ahead
Discovery Informatics: Emerging Themes

1. Computational support of the discovery process
2. Data and models
3. Social computing for discovery

http://diw.isi.edu/2012
THEME 1: Computational Support of the Discovery Process

- Unprecedented complexity of scientific enterprise
  - Is science stymied by the human bottleneck?

What aspects of the process could be improved
Computational Support of the Discovery Process

Many Opportunities for Improvement

- Make assumptions through background knowledge (combination of existing knowledge) via
  - Literature
  - Data
  - Collaboration
- Internalization -> idea(s)
- Consider the importance/novelty/feasibility/cost/risk of the idea(s)
- Formulate testable hypothesis(s)
- Make consistent/validate with/against existing knowledge

- Design the experiment (or study)
  - Identify controls
  - Inventory materials/equipment
  - Protocols
  - Statistics, comp tools
- Execute the experiment (or study)
  - Get funding
  - Adaptive /real time experimentation
  - Integrative interpretation
- Analyze/explore/validate the data
- Interpreting the results
  - Collaborative analysis
- Putting the results in context
- Communicating and
- Prioritizing the next thing

http://diw.isi.edu/2012
Computational Support of the Discovery Process

State of the Art

- Knowledge bases created from publications
  - Ontological annotations of articles including claims and evidence
  - Text mining to extract assertions to create knowledge bases
  - Reasoning with knowledge bases to suggest or check hypotheses

- Workflow systems to dynamically configure data analysis
  - Make process explicit and reproducible
  - Shared repositories of reusable workflows
  - Augmenting scientific publications with workflows

- Emerging provenance standards (OPM, W3C’s PROV)
  - Record relations among process steps, sources, data, agents

- Visualization
  - 3 separate fields: scientific visualization, information visualization, and visual analytics
  - “design studies”
  - Combining visualizations with other data

http://diw.isi.edu/2012
Computational Support of the Discovery Process

Research Challenges

- Developers and consumers must both be engaged in the process.
  - Make processes explicit -> manage, disseminate
- Define tools in terms of their role in processes
- Tension between targeted tools and generalized tools
- Develop methodology and education for tool design and usability
  - What has worked, and what has not worked
  - Understand adoption: when is a new tool worth the effort
- Pervasive and cheap reproducibility
  - Automated and scalable provenance management
- Improve flow of data from tool to tool in a process
- Appropriate metadata and provenance at all stages of a process
  - Formal representations of knowledge linked to supporting data and associated metadata
- Improved methods for abductive inference
- User-centered design
- Combining visualizations with other data, with models, with processes

http://diw.isi.edu/2012
THEME 2: Data and Models

- Complexity of models and complexity of data analysis
- Data analysis activities placed in a larger context

Interplay of models and data
Interplay of Models and Data

- One of the central processes of science is the interplay between models and data
  - Data informs model generation and selection
  - Models inform data collection and interpretation from both observations and experimentation
  - An iterative feedback loop exists between these two

- Improving this process would:
  - Increase the speed and accuracy of scientific research
  - Support development of more comprehensive models that cover larger datasets
  - Allow the effective study of more complex phenomena
  - Systematically transfer knowledge and best practices between scientific groups and fields
  - Broaden participation in science
Data & Models

State of the Art

Some individual scientific projects have the tools to iterate between data and models effectively and automatically, but…

- Few, if any, scientific fields have model formalisms and algorithms for this
- Requires high degree of hand-holding and does not generalize

Representations of data and models vary widely across different sciences, but typically…

- Scientists have far richer conceptions of data and models than currently expressed; they lack context, metadata
- Researchers must choose between lack of expressiveness and onerous complexity

Methodologies vary widely across different sciences, but typically…

- Not formalized in ways that support computation
- Limited in scalability to data and model space

Tend to focus on data -> models, not completing the feedback loop

http://diw.isi.edu/2012
Data & Models

Research Challenges

- Identify equivalence classes of scientific modeling domains (generality without compromising usefulness)
- Increase expressiveness of data and model representations
- Design scalable methods (datasets, hypothesis spaces)
- Enable reproducibility and model reusability
- Define principles of, design, and build interactive environments that support scientific tasks, e.g., model construction, design of data collection, data analysis
- Develop evaluation methods for discovery systems and scientific conclusions drawn from data and models

http://diw.isi.edu/2012
THEME 3: Social Computing for Science

- Multiplicative gains through broadening participation
- Some challenges require it, others can significantly benefit

Managing human contributions
Social Computing for Science

Opportunities

- Human computation has beaten best of breed algorithms
- Public interest in participating in scientific activity
- Mixed-initiative processes – humans exceed machine in many areas, so we need to assimilate them for the things that they do better
- Community assessment of models, knowledge, etc.
- Social agreement accelerates data sharing
- Social computing as facilitator of ad-hoc collaboration and unanticipated uses of data

http://diw.isi.edu/2012
Social Computing for Science

State of the Art

- Very different manifestations:
  - Collecting data (e.g., pictures of birds)
  - Labeling data (e.g., Galaxy Zoo)
  - Computations (e.g., Foldit)
  - Elaborate human processes (e.g., theorem proving)
  - Bringing people and computing together in complementary ways
Social Computing for Science

Research Challenges

- Create more effective ‘augmented human-computer teams’
  - Developing a taxonomy of approaches
    - Human computation
    - Collaborative knowledge creation
    - Partnering human creativity and brute force computation
  - Develop a design science
    - Track / understand goals, beliefs of people and systems
    - Participant roles and types of contributions
    - Develop catalog of incentives that motivate people to participate in various circumstances
    - Effective communication among the team members
    - Norms of behavior

- Expand the use of social computing methods to include new ways of producing, communicating, and ‘reviewing’ scientific results

http://diw.isi.edu/2012
Outline

- Motivation for Discovery Informatics
  - Why now
- Possible Grand Challenges in Discovery Informatics
- Research challenges
  - Vision scenarios for several domain sciences
    - Biology, social sciences, geosciences, astronomy
- Moving Ahead
Vision Scenario for Biological Sciences (I)

- Track the implications of results from other aspects of biology.
- Make sense of mass phenotyping datasets
- Address the paradox: price of gathering data is plummeting, the price of analyzing it is either flat or increasing.

http://diw.isi.edu/2012
Vision Scenario for Biological Sciences (II): How DI Advances Would Help

Improving process:

- Give me *interesting* different disciplines) based on what I am working on (my model, my model fragment, entities that are being worked on in my lab).

- *In silico* hypothesis testing / comparison against the broad, integrated knowledge.
  - If we solve the knowledge representation and “upload” problem, we can increase the quality and impact of biologists’ work

- Make tools that support a new generation of “systems” scientists who are more integrative and quantitative

http://diw.isi.edu/2012
Vision Scenario for Biological Sciences (III): How DI Advances Would Help

Data and models:

- Tools for evaluation of models against existing knowledge
- Discovering things that matter to individuals
  - Identify asthma attack risk based on garbage pickup schedule
  - City’s poorest, who relied disproportionately on emergency room visits, faced the most expensive health care costs while receiving the worst care.
- Tools for “in your garage” synthetic biology; facilitate the growth of homebrew systems, and also perhaps provide early warning of dangers.
Vision Scenario for Social Science (I)

*Education for better science, better citizens and better communities*

Easy to imagine:
- Shift from data poverty to data wealth
- Ability to ask both big questions – those of societal-level importance – AND pursue deep exploration of specific issues
- Opportunities to discover

For many, current approaches fail to advance their knowledge

For some, current approaches fall short of challenging them

It's wicked expensive

Need a more coherent view of life-long learning

We know education linked to to economy, community, participation
Vision Scenario for Social Science (II)

Current barriers to discovery

- Data unrepresentative and incomplete (poor data quality, segmented data sets, and questionable curation)
- Intrinsic tension between what can be learned from analysis and real issues of privacy and identity
- Models and analytic techniques constrain scientists and decision-makers
- Analysis and findings segmented across different intellectual communities
- Very little insight into long-term effects of educational approaches and choices

Statements true beyond education …
Vision Scenario for Social Science (III)

How DI Advances Would Help

Make data better:
- Improve and expand data collection (e.g., social computing), advance ability to integrate data
- Improve data representation (w/r/t: quality, incompleteness, meta-data on context, provenance)

Respect privacy and regulatory constrains while making use of the data
- Model (formally) and enforce these in use

Advance model development/use and analytic capabilities:
- Reasoning while accounting for all the new features this data provides
- Allowing analysis across varying data types and sources
- Enabling more ‘for whom and under which conditions’ analysis
- Building more robust models (and sharing them)

Synthesize literature across intellectual communities
- Support for bibliometric connection and pattern-finding across papers.

Advancing predictive models of education on life outcomes (e.g., “what if I go to a community college and then transfer?”)
DI Themes Recurring Across Sciences: Geosciences

Climate Model Intercomparison Project version 5 (CMIP5) expected to reach 2-3PB by 2013, satellites collect observations at high spatial and temporal resolutions

- **Challenges**
  - Automatically identify (potentially constrained, generalized) patterns, causal relationships from large spatio-temporal datasets
  - Simulations and observations – assimilation of data and models
  - Provide interactive, highly responsive visualizations

- **Opportunities**
  - Generate hypotheses for the underlying physical mechanisms
  - Improve prediction and forecasting across temporal scales
    - Early warning for transient events (e.g., hurricanes, tsunamis)
  - Representation of scientific arguments, consensus & controversy

http://diw.isi.edu/2012
DI Themes Recurring Across Sciences: Forensic Paleoclimatology

NOAA Paleoclimatology Archive contains 7K cores up to 3km long, with 13 proxies measured at millimeter intervals

- **Challenges**
  - Determine what happened to a set of unobserved variables over the course of time under the influence of (potentially unknown) processes
  - Reconstruct and align the temporal history of material in core data of different types (glaciers, ocean sediments, trees) at different spatial and temporal scales
  - Handle multiple competing hypotheses, model and data uncertainty

- **Opportunities**
  - Improve reconstruction of past history of the climate
  - Deduce causality and patterns in the global climate system
  - Make better predictions about future climate
  - Evaluating potential interventions

http://diw.isi.edu/2012
DI Themes Recurring Across Sciences: Astronomy

Large Synoptic Sky Telescope (LSST) starts operation in 2018, will collect ~100PB of data within a decade

- Challenges
  - 10’s of TB of data, 70K anomalies per night
  - Tracking and classifying objects and events (possibly unknown)

- Opportunities
  - Go beyond detection, to discovery of general theories/concepts
  - Real-time alerting of discoveries
  - Hybrid (human and automated) control of instruments
  - Coordination of crowd-sourced science
Outline

- Motivation for Discovery Informatics
  - Why now
- Possible Grand Challenges in Discovery Informatics
- Research challenges
- Vision scenarios for several domain sciences
  - Biology, social sciences, geosciences, astronomy
- Moving Ahead
Starting to Articulate a Discipline and Community of Discovery Informatics

- Social computing for discovery
- Computational support of the discovery process
- Data and models

MOVIE CREDITS

Director
Barbara Jones
Executive producer
Sandeep Jain
Producers
...
Crane engineer
...
Casting
...
Make up
...
Director’s assistant
...
Special effects crew
...
Casting
...

http://diw.isi.edu/2012
Forming a Community in Discovery Informatics

- Important pieces of Discovery Informatics are broadly scattered across fields and subfields
  - Domain sciences: {bio/eco/geo/…}-informatics forums
  - Social sciences

- In order for Discovery Informatics to succeed, we need a community and a funding framework that place computer scientists, domain scientists, and social scientists on equal footing

- Characterization of domains and facets that impact current discovery informatics practices is still not understood
  - This cannot be done by domain scientists or computer scientists or social scientists alone
  - What are equivalent classes of domains across sciences

- Methodologies to approach new domains/problems/processes/users do not exist
  - Need to share lessons learned, but they are scattered
  - Failures are important and not well reported

http://diw.isi.edu/2012
Vannewar Bush, “As We May Think”, 1945

“There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers […]. Yet specialization becomes increasingly necessary for progress […]

Professionally our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose.

Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them […] The physician, puzzled by a patient’s reactions, strikes the trail established in studying an earlier similar case […] with side references to the classics for the pertinent anatomy and histology. The chemist, struggling with the synthesis of an organic compound, has all the chemical literature before him in his laboratory, with trails following the analogies of compounds, and side trails to their physical and chemical behavior. The historian, with a vast chronological account of a people, […] can follow at any time contemporary trails which lead him all over civilization at a particular epoch.

There is a new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record. The inheritance from the master becomes, not only his additions to the world's record, but for his disciples the entire scaffolding by which they were erected.”
Herb Simon

“We are still very far from a complete understanding of the whole structure of the psychological processes involved in making scientific discoveries. But our analysis makes more plausible the hypothesis that at the core of this structure is the same kind of selective trial and error search that has been shown to constitute the basis for human problem solving activity.” – 1966

“In an important sense, predicting the future is not really the task that faces us. After all, we, or at least the younger ones among us, are going to be a part of that future. Our task is not to predict the future; our task is to design a future for a sustainable and acceptable world, and then to devote our efforts to bringing that future about. We are not observers of the future; we are actors who, whether we wish to or not, by our actions and our very existence, will determine the future's shape.” -- 2000