Provenance-Rich Science

Juliana Freire
VisTrails Group & Web and Databases Lab
Science Today: Data Overload

Obtain Data → Analyze/Visualize → Publish/Share

- Simulations
- Sensors
- User studies
- Particle colliders
- Web
- Databases
- Sequencing machines
- Sequencing machines
Science Today: Data Intensive

Obtain Data

- Simulations
- Sensors
- User studies
- Web
- Databases
- Sequencing machines
- Particle colliders

Analyze/Visualize

- AVS
- ParaView
- VTK
- VI
- VisIt

Publish/Share

- Taverna
- VisTrails
- VTK
- MATLAB
- Oracle
- R
- IBM
- Excel
- DB2
Science Today: Data Intensive

Obtain Data

- Simulations
- Sensors
- User studies
- Particle colliders
- Web
- Databases
- Sequencing machines

Analyze/Visualize

Publish/Share

- Sequencing machines
- Particle colliders
- Web
- Databases
- User studies
- Sensors
- Simulations

Provenance-Rich Science
Crete, 2011
Juliana Freire
Science Today: Data Intensive

Obtain Data → Analyze/Visualize → Publish/Share

- Simulations
- Sensors
- User studies
- Particle colliders
- Web
- Sequencing machines
- Databases
- Sequencing machines
- Databases

Provenance-Rich Science
Crete, 2011
Juliana Freire
Science Today: Incomplete Publications

- Publications are just the tip of the iceberg.
  "It’s impossible to verify most of the results that computational scientists present at conference and in papers." [Donoho et al., 2009]
  "Scientific and mathematical journals are filled with pretty pictures of computational experiments that the reader has no hope of repeating." [LeVeque, 2009]
  "Published documents are merely the advertisement of scholarship whereas the computer programs, input data, parameter values, etc. embody the scholarship itself." [Schwab et al., 2007]
Vision: Provenance-Rich Science

- Obtain Data
- Collaborate
- Publish/Share
- Analyze/Visualize

Provenance Repository
Provenance in Science

- Not a new issue!
- Lab notebooks have been used for a long time
- What is new?
  - Large volumes of data
  - Complex analyses—computational processes
- Writing notes is no longer an option
Provenance in Science

- Interpret and reproduce results
- Understand the experiment and chain of reasoning that was used in the production of a result
- Verify that an experiment was performed according to acceptable procedures
- Identify the inputs to an experiment were and where they came from
- Assess data quality
- Track who performed an experiment and who is responsible for its results

*Provenance is as (or more!) important as the results*
Vision: Provenance-Rich Science

Get template from the electronic notebook system adopted by her group, which manages the metadata catalogs, version control systems, workflows, and analysis tools.

While she plots distributions from the results of a new workflow, the system analyzes previous workflows and suggests quantities to plot that were frequently used in similar analyses.

Data mining algorithms identify commonalities and differences between her analysis and those of other group members, and suggests additional workflows she can try.

She identifies an error in the simulation code, updates it and tells the system to re-run her analyses that depend on the simulation results.
Vision: Provenance-Rich Science

She prepares a presentation to her group and later journal submission where she includes the results and their provenance.
Outline

- Provenance and data exploration
  - The Change-Based Provenance model
- Sharing and publishing results
- Querying and mining provenance
  - Opportunities and Challenges
  - Ongoing work
Provenance and Data Exploration
Exploring and Visualizing Data

- Data analysis and visualization are key to scientific discovery --- “A picture is worth a thousand words”
Exploring and Visualizing Data

- Data analysis and visualization are key to scientific discovery --- “A picture is worth a thousand words”
- But their creation has become a bottleneck...
- The design of effective visualizations and analyses requires a complex process and a deep understanding of existing techniques and how they relate to human cognition.
Data Exploration and Complexity

Hypothesis

Visualization/Analysis
Ideas

too much time here!

Visualization/Results

Python + R

Vistrails Screenshots +
Adobe Illustrator

Python + R

GraphViz +
Inkscape

Python + R

SSDBM 2008
Data Exploration and Workflows

- Workflows have been traditionally used to automate repetitive tasks
- In exploratory tasks, *change is the norm!*
  - Data analysis and exploration are iterative processes
  - You need several workflows...

![Diagram](Figure modified from J. van Wijk, IEEE Vis 2005)
Exploration and Creativity Support

- Reflective reasoning is key in the exploratory processes
  
  "Reflective reasoning requires the ability to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward, sometimes backtracking when a promising line of thought proves to be unfruitful. ...the process is slow and laborious"
  
  Donald A. Norman

- Need external aids—tools to facilitate this process
  - Creativity support tools [Shneiderman, CACM 2002]

- Need aid from people—collaboration
The VisTrails System

- Combines features of visualization, data analysis, and scientific workflow systems
  - Integrate multiple libraries (e.g., VTK, R, matplotlib, *)
- Transparently tracks provenance of the discovery process---from data acquisition to visualization
  - The trail followed as users generate and test hypotheses
- Leverage provenance to streamline exploration
  - Support for reflective reasoning and collaboration
  - Query and mine pipelines and provenance
  - Auto-completion for workflows
  - ...
- Focus on usability—build tools for scientists
- The system is open source: http://www.vistrails.org
  - Multi-platform: Linux, Mac, Windows
  - Written in Python + Qt
Demo
Change-Based Provenance

- Treat workflow as a first-class data item
- Provenance = changes to computational tasks
  - Add a module, add a connection, change a parameter value
Change-Based Provenance

- Treat workflow as a first-class data item
- Provenance = changes to computational tasks
  - Add a module, add a connection, change a parameter value
Change-Based Provenance

- Treat workflow as a first-class data item
- Provenance = changes to computational tasks
  - Add a module, add a connection, change a parameter value
- A *vistrail* node $v_t$ corresponds to the workflow that is constructed by the sequence of actions from the root to $v_t$
  
$\mathit{v}_t = x_n \circ x_{n-1} \circ \ldots \circ x_1 \circ \emptyset$

- Extensible change algebra

[Freire et al, IPAW 2006]
Provenance Beyond Reproducibility

- Support for reflective reasoning
- Ability to compare data products

[Freire et al., IPAW 2006]
Computing Workflow Differences

- No need to compute subgraph isomorphism!
- A vistrail is a rooted tree: all nodes have a common ancestor—diffs are well-defined and *simple to compute*

\[
\begin{align*}
    v_{t_1} &= x_i \circ x_{i-1} \circ \ldots \circ x_1 \circ \emptyset \\
    v_{t_2} &= x_j \circ x_{j-1} \circ \ldots \circ x_1 \circ \emptyset \\
    v_{t_1} - v_{t_2} &= \{x_i, x_{i-1}, \ldots, x_1, \emptyset\} - \\
                        &\quad \{x_j, x_{j-1}, \ldots, x_1, \emptyset\}
\end{align*}
\]

- Different semantics:
  - Exact, based on ids
  - Approximate, based on module signatures
Provenance Beyond Reproducibility

- Support for reflective reasoning
- Ability to compare data products
- Explore parameter spaces and compare results

[Freire et al., IPAW 2006]
Exploring the Change Space

- Scripting workflows: Parameter explorations are simple to specify and apply
- Exploration of parameter space for a workflow $\mathbf{v}_t$
  
  $\text{setParameter}(\text{id}_n, \text{value}_n) \circ \ldots \circ (\text{setParameter}(\text{id}_1, \text{value}_1) \circ \mathbf{v}_t)$

- Exploration of multiple workflow specifications
  
  $\text{addModule}(\text{id}_i, \ldots) \circ (\text{deleteModule}(\text{id}_i) \circ \mathbf{v}_1)$
  
  $\ldots$

  $\text{addModule}(\text{id}_i, \ldots) \circ (\text{deleteModule}(\text{id}_i) \circ \mathbf{v}_n)$

- Results can be conveniently compared in the VisTrails spreadsheet
- Can create animations too!
- Caching to avoid redundant computations [Bavoil et al., IEEE Vis 2005]
A program is to a workflow what an unstructured document is to a (structured) database.
A program is to a workflow what an unstructured document is to a (structured) database.
Change-Based Provenance: Summary

- General: Works with any system that has undo/redo!
Provenance Enabling 3rd-Party Tools

[Callahan et al., IPAW 2008]
Provenance Plugin for ParaView

Provenance Explorer Plug-in for Kitware's ParaView 3.0

VisTrails Inc.

http://www.cs.utah.edu/~juliana/videos/paraview_plugin.avi
Change-Based Provenance: Summary

- General: Works with any system that has undo/redo!
  - Provenance SDK to be released soon!
- Concise representation
- Uniformly captures data and workflow provenance
  - Where does the data product come from?
  - How has workflow changed over time?
- Results can be reproduced
- Detailed information about the exploration process
- **Provenance beyond reproducibility:**
  - Scientists can return to any point in the exploration space
  - Compare multiple results side-by-side
  - Support for collaboration
  - Understand problem-solving strategies
  - Knowledge re-use
Sharing Results
Users have to go through many (complicated) steps to reproduce and validate results:

- install software
- download libraries
- download example files
- learn how to use the software
- manipulate workflows

Science portals and customized apps are too expensive to develop
VisMashup

- Leverages provenance to simplify the creation, maintenance, deployment and use of *customized applications* (mashups)
- Uses dataflows as the underlying model
- Keeps detailed provenance information of the application development process and use

[Santos et al., IEEE TVCG 2009]
Acquire and Analyze Workflows → Create Views (Simplify Workflows) → Combine Views → App generation and deployment

Mine workflow provenance
Acquire and Analyze Workflows → Create Views (Simplify Workflows) → Combine Views → App generation and deployment
**VisMashup**

- Deploy visualization applications on different configurations

---

**Acquire and Analyze Pipelines**

**Create Views (Simplify Pipelines)**

**Combine Views**

**App generation and deployment**

---

Provenance-Rich Science

Crete, 2011

Juliana Freire
Publishing Scientific Results
**Improved muscular efficiency displayed as Tour de France champion matures**

Edward F. Coyle

Human Performance Laboratory, Department of Kinesiology and Health Education, The University of Texas at Austin, Austin, TX 78712 (E-mail: coyle@mail.utexas.edu).

Mont Hall 222, Dept. of Kinesiology and Health Education, The Univ. of Texas at Austin, Austin, TX 78712.

**MUCH HAS BEEN LEARNED**

This case describes the physiological maturation from ages 21 to 28 yr of a professional cyclist who has won numerous stage races and even the prestigious Tour de France at ages 27–32 yr. Maximal oxygen uptake (2193) in the trained state remained at 6.1 l/min, and his blood lactate threshold occurred at a V̇O₂ equivalent to that at age 25 to 28 yr. It is noteworthy that age 25 to 28 yr of this champion developed advanced cancer, requiring surgeries and chemotherapy. During the months leading up to each of the Tour de France victories, he maintained body weight and body fat (by 5–7 kg (e.g., 75%), whereas body weight remained at 81 kg (28). Laboratory measures of the anthropometry (e.g., 5.0 l/min or 0.1 kg), and the following tests were performed after informed consent was obtained, with procedures approved by the Internal Review Board of The University of Texas at Austin. Mechanical efficiency and the blood lactate threshold (LT) were assessed in the subject bicycling at 50, 60, 70, 80, and 90% V̇O₂ max. After a 10- to 20-min period of active recovery, V̇O₂ max when cycling was measured. Therefore, body composition was determined by hydrostatic weighing and/or analysis of skin-fold thickness (34, 35).
"raw data from the January 1993 test that revealed several additional deviations from the published methodology. Coyle used a 20-min ergometer protocol (not 25 min), including 2- and 3-min stages where respiratory exchange ratios (RER) exceeded 1.00. An RER > 1.00 invalidates use of the Lusk equations (5) to estimate energy expenditure.”

"...all of the published delta efficiency values are wrong. ... there exists no credible evidence to support Coyle's conclusion that Armstrong's muscle efficiency improved."

http://jap.physiology.org/cgi/content/full/105/3/1020


http://ori.dhhs.gov/misconduct/cases/

Nobel Laureate Retracts Two Papers, NYTimes 09/24/2010
Provenance-Rich Publications

- Bridge the gap between the scientific process and publications
  - The scientific record needs to be complete and trustworthy
- Show me the proof: Publish results that can be reproduced and validated
  - Papers with deep captions
  - Encouraged by ACM SIGMOD, a number of journals and funding agencies
Provenance-Rich Publications: Benefits

- Produce and share more knowledge
- Allow scientists to stand on the shoulders of giants (and their own...)
- Science can move faster!
- Higher-quality publications
  - Authors will be more careful
  - Many eyes to check results
- Describe more of the discovery process: people only describe successes, can we learn from mistakes?
- Expose users to different techniques and tools: expedite their training; and potentially reduce their time to insight
Provenance-Rich Documents

CrowdLabs: Social Analysis and Visualization for the Sciences
Emanuele Santos, Philip Moses, Juliana Freire, and Claudio T. Silva, Senior Member, IEEE

Abstract—Managing and understanding the large volumes of scientific data is undoubtedly one of the most difficult research challenges scientists face today. As large interdisciplinary research teams work together, they are able to generate a diverse array of analyses and publications. However, such products are often built manually, and are not flexible or efficient enough to support the vast heterogeneity of data sources, analysis techniques, data products, and user communities that need to access these data. In this paper we describe CrowdLabs, a system that adopts the models used by social Web sites and that combines a set of usable tools and a scalable infrastructure to provide a rich collaborative environment for scientific visualization. CrowdLabs offers a suite of visualizations of a wide range of data sets. However, it is suited to different requirements of scientific exploration. In particular, they were designed for small data sets and only provide a limited set of visualizations.

Fig. 8: Visualizing a binary star system simulation. This is an image that was produced by submitting a workflow diagram to the system. The original workflow is available at http://www.crowdlabs.org/vistrails/workflows/details/32/.

http://www.crowdlabs.org/vistrails/workflows/details/32/

Provenance-Rich Science Crete, 2011 Juliana Freire 48
\begin{figure}[t]
\centering
\vistrail[wfid=32,buildalways=false]{width=0.5\linewidth}

\caption{Columbia river virtual estuary: visualization of salinity over time. See http://www.crowdlabs.org/vistrails/workflows/details/32.}
\label{fig:cmop}
\end{figure}
Provenance-Rich Publications: Challenges

- Representing computational processes
  - Use workflows: one per result reported in the paper
  - Semantics
- Packaging software
  - Software versions, environment
  - Virtual machines
- Shipping data
- Computational infrastructure
  - Support multiple OSes, multiple versions of tools, ...
- Who will pay for it?
  - Publishers? Libraries? Authors?
- Longevity
Repeatability Central

- Goal: Develop a community experiment platform where scientists can share and generalize reproducible results
- *Tools for authors* to create workflows that encode the computational processes and package the results for publication
- *Tools for testers* to repeat and validate results
- Interfaces for searching, comparing and analyzing experiments and results
  - Can we discover better approaches to a given problem?
  - Or discover relationships among workflows and the problems?

- *Pilot for ACM SIGMOD 2011: Joint work with Dennis Shasha and Philippe Bonnet*
Provenance Analytics: Querying, Mining and Re-Using Provenance
Provenance Analytics: Opportunities

- Opportunity for knowledge discovery, sharing and re-use
- Query information
  - Understand processes and data dependencies
  - Find *useful* workflows, e.g., given a piece of data or task, which workflow should we run?
- Mine information
  - Discover *interesting* patterns (e.g., common workflow patterns) → *recommendation system, discover analogies*
  - Identify homogeneous workflow groups by clustering → *organize collections* [Santos et al., IPAW 2008]
  - *Infer workflow specification* from execution log [Aalst et al., TKDE 2004]
The Need for Guidance in Workflow Design
VisComplete: A Workflow Recommendation System

- **Mine** provenance collection: Identify graph fragments that co-occur in a collection of workflows
- Predict sets of likely workflow additions to a given partial workflow
- Similar to a Web browser suggesting URL completions

![Diagram of the VisComplete system](image)

[Koop et al., IEEE Vis 2008]
VisComplete: A Workflow Recommendation System

- Mine provenance collection: Identify graph fragments that co-occur in a collection of workflows
- Predict sets of likely workflow additions to a given partial workflow
- Similar to a Web browser suggesting URL completions
VisComplete: Demo

VisComplete: Data-driven Suggestions for Visualization Systems
Querying Provenance

- **Graph traversal**
  - Derivation lineage
  - Data dependencies

  *Find the process that led to resampled-head-vis.png*

  *Which data sets contributed to resampled-head-vis.png*

- **Graph patterns**

  *Find all invocations of vtkContourFilter with isosurface value = 57 that are preceded by resampling*
Querying Provenance: Challenges

- Provenance is a graph
- How to specify queries? [Scheidegger et al., TVCG 2007]
Querying Provenance by Example

- Hard to specify queries using text
- Querying workflows by example [Scheidegger et al., TVCG 2007; Beeri et al., VLDB 2006; Beeri et al. VLDB 2007]
  - WYSIWYQ -- What You See Is What You Query
  - Interface to create workflow is same as to query
Querying Provenance: Challenges

- Provenance is a graph
- How to specify queries? [Scheidegger et al., TVCG 2007]
- Recursive queries are expensive
  - Re-use work on deductive databases (?)
- Structural queries are expensive: need to check graph containment
  - Re-use existing techniques for graph indexing [Yan et al., SIGMOD 2004, Zhao et al., VLDB 2007]
- Exploratory queries not supported
Querying Provenance: Challenges

- Provenance is a graph
- How to specify queries? [Scheidegger et al., TVCG 2007]
- Recursive queries are expensive
- Structural queries are expensive: need to check graph containment
  - Re-use existing techniques for graph indexing [Yan et al., SIGMOD 2004, Zhao et al., VLDB 2007]
- Exploratory queries not supported
  - Queries are disconnected
  - Potentially larger number of answers
- **Our approach**: [Koop et al., Submitted 2010]
  - Extend existing techniques to support disconnected queries
  - Create summary graphs
  - Great reduction in the number of isomorphism checks
Querying Provenance: Challenges

- Provenance is a graph
- How to specify queries? [Scheidegger et al., TVCG 2007]
- Recursive queries are expensive
- Structural queries are expensive: need to check graph containment
- Exploratory queries not supported
- How to display results? [Ellkvist et al., KEYS 2009]

Summarize collection by clustering

Generate descriptive snippets
Mining Provenance: Challenges

- Provenance is a graph
- Mining is expensive
- Structure is more complex
  - Modules with parameters+values
  - Typed connections
- How to model provenance?
  - For clustering, a vector-space based representation produced results correlated to results obtained using a more expensive structural representation [Santos et al., IPAW 2008]
- Which notions of distance and metrics make sense for different applications and data sets?
- Which algorithms are effective and efficient?
  [Lauro Lins, Nivan Ferreira. Work in progress]
Conclusions and Future Work

- Provenance management is essential for data-intensive science
  - It is a requirement and an enabler!
- Sharing provenance creates new opportunities
- We should all leverage it!
- But it also creates many challenges
  - Integrate provenance from multiple sources
  - Efficient storage, querying
  - Trust and security
  - Packaging provenance for publication and reproducibility
  - Longevity …
- Great opportunity for computer scientists!
Acknowledgments

- Claudio Silva and the VisTrails team
- This work is partially supported by the National Science Foundation grants IIS 1050422, IIS 0905385, IIS 0844572, IIS 0746500, CNS 0751152,; the Department of Energy, an IBM Faculty Award, and a University of Utah Seed Grant.
Thank you
Obrigada
Social Analysis of Scientific Data
Social Analysis of Data

- Sharing data can have important implications to science
  - *Sharing of Data Leads to Progress on Alzheimer’s*, NYTimes, Aug 12, 2010

- Should also share analyses, visualizations, tools!

- ManyEyes, Swivel, Tableau Public: collaborative visualization
  - Limitations: small, tabular data; fixed set of visualization techniques; no provenance

- myExperiment: focus on bioinformatics-related Web services, share workflows
  - Limitations: can’t run workflows; no provenance
Social Analysis of **Scientific** Data

- **New requirements:**
  - Support for large data
  - User-defined visualizations and analyses
  - Execute workflows close to the data
  - Provenance

- **Preliminary work:** [http://www.crowdlabs.org](http://www.crowdlabs.org)

- Benefit from the collective wisdom: by *querying* analysis specifications which make sophisticated use of tools, along with data products and their provenance, users can learn by example from the reasoning and/or analysis strategies of experts