A PROVENANCE-BASED INFRASTRUCTURE FOR CREATING REPRODUCIBLE PAPERS

Juliana Freire
Juliana.freire@nyu.edu
VisTrails Group & Web and Databases Lab
NYU Poly
Science Today: Data Intensive

Obtain Data

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

Analyze/Visualize

Publish/Share
Science Today: Data + Computing Intensive

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

Analyze/Visualize
- AVS
- ParaView
- VTK
- VisIt

Publish/Share
- Taverna
- VisTrails
- MATLAB
- Oracle
- R
- IBM
- DB2

Reproducible Research '11
UBC, Vancouver
Juliana Freire
Science Today: Data + Computing Intensive

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

Analyze/Visualize
- Boltzmann's constant
- Special Relativity
- Monte Carlo simulations
- Numerical methods
- Data analysis
- Visualization techniques

Publish/Share
- Sequencing machines
- Particle colliders
- User studies
- Databases
- Web
- Simulations
- Sensors
- Scientific papers
- Conference presentations
- Open data sharing platforms

Reproducible Research '11
UBC, Vancouver
Juliana Freire
Science Today: Data + Computing Intensive

Obtain Data

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

Analyze/Visualize

Publish/Share

- Sequencing machines
- Particle colliders
- User studies
- Databases
- Web
- Sequencing machines
- Particle colliders
- User studies
Science Today: Incomplete Publications

- Publications are just the tip of the iceberg
  - Scientific record is incomplete---to large to fit in a paper
  - Large volumes of data
  - Complex processes
- Can’t (easily) reproduce results
Science Today: Incomplete Publications

- Publications are just the tip of the iceberg.
  - Scientific record is incomplete—too large to fit in a paper.
  - Large volumes of data.
  - Complex processes.

- Can’t (easily) reproduce results.

  “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]
Science Today: Incomplete Publications

- Publications are just the tip of the iceberg.
  - Scientific record is incomplete—too large to fit in a paper.
  - Large volumes of data.
  - Complex processes.

- Can’t (easily) reproduce results.

  “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.”

http://ori.dhhs.gov/misconduct/cases/
Nobel Laureate Retracts Two Papers, NYTimes 09/24/2010
Vision: Provenance-Rich Science

Obtain Data → Analyze/Visualize → Publish/Share

Provenance Repository
Vision: Provenance-Rich Science

Provenance is the scientific record
Provenance-Rich Publications

- Bridge the gap between the scientific process and publications
  - Papers with *deep* captions and a *complete and trustworthy* scientific record
- Show me the proof: results that can be reproduced and validated
- Encouraged by ACM SIGMOD, a number of journals, funding agencies, academic institutions
  - E.g., ETH
    http://www.vpf.ethz.ch/services/researchethics/Broschure
- Several workshops, different communities
  - Beyond The PDF, SIAM Symposium on Reproducible Research, AMP Workshop on Reproducible Research, Workshop on Archiving Experiments
Provenance-Rich Publications: Benefits

- Produce more knowledge---not just text
- Allow scientists to stand on the shoulders of giants
- Science can move faster
  - [http://www.nytimes.com/2011/06/26/opinion/sunday/26ideas.html?_r=1](http://www.nytimes.com/2011/06/26/opinion/sunday/26ideas.html?_r=1)
- Allow scientists to stand on their own shoulders!
- 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 scientific community to different techniques and tools: expedite their training; and potentially reduce their time to insight
Provenance-Rich Publications: Challenges

- It is too hard, time-consuming for authors to prepare compendia of reproducible results
  - Data, computations, parameter settings, environment, etc.
- It is too hard for reviewers (and readers) to install, compile, and reproduce experiments
  - Different OSes, library versions, hardware, large data, incompatible data formats...
- Need to simplify the process of sharing, reviewing and re-using scientific experiments and results
Our Approach: An Infrastructure to Support Provenance-Rich Papers [Koop et al., ICCS 2011]

- Tools for authors to create reproducible papers
  - Specifications that encode the computational processes
  - Package the results
  - Link from publications
  Support different approaches

- Tools for testers to repeat and validate results
  - Explore different parameters, data sets, algorithms

- 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?
  - How to describe experiments?
The ALPS project release 2.0:
Open source software for strongly correlated systems

B. Bauer1 L. D. Carr2 H.G. Evertz2 A. Feiguin3 J. Freire3
S. Fuchs4 L. Gamper1 J. Gukelberger1 E. Gull5 S. Guertler5
A. Hehn1 R. Igarashi6,7,8 S.V. Isakov7 D. Koop7 P.N. Ma7
P. Mates1,7 H. Matsuo11 O. Parcollet12 G. Pawlewicz13
J.D. Picon14 L. Pollet15 E. Santos1 V.W. Scarola16
U. Schollwöck17 C. Silva1 B. Surer1 S. Todo16,17 S. Trebst18
M. Troyer19‡ M. L. Wall2 P. Werner2,10 S. Wessel19

1Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland
2Department of Physics, Colorado School of Mines, Golden, CO 80401, USA
3Institut für Theoretische Physik, Technische Universität Graz, A-8010 Graz, Austria
4Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming, 82071, USA
5Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA
6Institut für Theoretische Physik, Georg-August-Universität Gottingen, Gottingen, Germany
7Institut für Theoretische Physik, Georg-August-Universität Gottingen, Gottingen, Germany
8Columbia University, New York, NY 10027, USA
9Bethe Center for Theoretical Physics, University Bonn, Nussallee 22, 53115 Bonn, Germany
10Department for Physics, Arnold Sommerfeld Zentrum für Theoretische Physik und Astrophysik, \"TH Munich, 85748 Garching, Germany
11Department of Physics, University of Tokyo, 3-8-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
12Department of Physics, University of Tokyo, 3-8-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
13Department of Physics, University of Tokyo, 3-8-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
14Department of Physics, University of Tokyo, 3-8-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
15Department of Physics, University of Tokyo, 3-8-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
16Department of Physics, University of Tokyo, 3-8-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
17Department of Physics, University of Tokyo, 3-8-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
18Department of Physics, University of Tokyo, 3-8-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
19Department of Physics, University of Tokyo, 3-8-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

‡Corresponding author: troyer@comp-physics.org

Figure 3. In this example we show a data collapse of the Binder Cumulant in the classical Ising model. The data has been produced by remotely run simulations and the critical exponent has been obtained with the help of the VisTrails parameter exploration functionality.

[Bauer et al., JSTAT 2011]


Reproducible Research ’11 UBC, Vancouver

Juliana Freire 15
A Reproducible Paper: ALPS2.0

The ALPS project release 2.0:
Open source software for strongly correlated systems

B. Bauer¹ L. D. Carr⁵ H.G. Evertz⁶ A. Feiguin⁷ J. Freire⁸
S. Fuchs⁵ L. Gamper⁵ J. Gukelberger⁷ E. Gull⁸ S. Guertler⁹
A. Hehn⁵ R. Igarashi⁵ 10 S.V. Isakov⁷ D. Koop⁷ P.N. Ma⁹
P. Matos⁷ H. Matsuo¹¹ O. Parcollet¹¹ G. Pawlowski¹¹
J.D. Picon¹¹ L. Pellet¹¹ E. Santos¹¹ V.W. Scarola¹¹
U. Schollwöck¹² C. Silva¹⁰ B. Surer¹ S. Todo¹⁰ 11 S. Trebst¹⁸
M. Troyer¹ M. L. Wall⁵ P. Werner⁸ S. Wessel¹⁹ 20
¹Theoretische Physik, ETB Zurich, 8093 Zurich, Switzerland
²Department of Physics, Colorado School of Mines, Golden, CO 80401, USA
³Institut für Theoretische Physik, Technische Universität Graz, A-8010 Graz, Austria
⁴Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA
⁵Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA
⁶Institut für Theoretische Physik, Georg-August-Universität Göttingen, Göttingen, Germany
⁷Department of Physics, Columbia University, New York, NY 10027, USA
⁸Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 24, 53115 Bonn, Germany
⁹Department of Physics, Harvard University, Cambridge, MA 02138, USA
¹⁰Bethe Center for Theoretical Physics, Universitat Bonn, Nussallee 24, 53115 Bonn, Germany
¹¹Heisenberg Institute and University of Bremen, 28359 Bremen, Germany
¹²Institute for Solid State Theory, RWTH Aachen University, 52056 Aachen, Germany
¹³Department of Physics, Virginia Tech, Blacksburg, Virginia 24061, USA
¹⁴Physics Department, Harvard University, Cambridge, Massachusetts 02138, USA
¹⁵Institute of Theoretical Physics, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
¹⁶Faculty of Physics, Jagiellonian University, 30-348 Krakow, Poland
¹⁷Institute for Solid State Theory, RWTH Aachen University, 52056 Aachen, Germany
¹⁸Faculty of Physics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
¹⁹Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
²⁰Department of Physics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Figure 3. In this example we show a data collapse of the Binder Cumulant in the classical Ising model. The data has been produced by remotely run simulations and the critical exponent has been obtained with the help of the VisTrails parameter exploration functionality.

[Bauer et al., JSTAT 2011]


Reproducible Research ’11 UBC, Vancouver Juliana Freire
Some Videos

Editing an executable paper written using LaTeX and VisTrails

Exploring a Web-hosted paper using server-based computation

An interactive paper on a Wiki*
http://www.vistrails.org/index.php/User:Tohline/CPM/Levels2and3

Reproducible Papers

An interactive paper on a Wiki*
http://www.vistrails.org/index.php/User:Tohline/CPM/Levels2and3

The ALPS 2.0 paper
Writing & Development

An author benefits from working in an environment that simplifies the creation of an executable paper

- First prototype: Leverage VisTrails’ infrastructure

[Koop et al., ICCS 2011]
The VisTrails System

- Workflow-based system for data analysis and visualization
- Comprehensive *provenance infrastructure*
- *Transparency*ly 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 provenance
- 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
The VisTrails System

- Workflow-based system for data analysis and visualization
- Comprehensive *provenance infrastructure*
- *Transparency*ly 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 provenance

- Visualizing environmental simulations (CMOP STC)
- Simulation for solid, fluid and structural mechanics (Galileo Network, UFRJ Brazil)
- Quantum physics simulations (ALPS, ETH Switzerland)
- Climate analysis (CDAT)
- Habitat modeling (USGS)
- Open Wildland Fire Modeling (U. Colorado, NCAR)
- High-energy physics (LEPP, Cornell)
- Cosmology simulations (LANL)
- Study on the use of tms for improving memory (Psychiatry, U. Utah)
- eBird (Cornell, NSF DataONE)
- Astrophysical Systems (Tohline, LSU)
- NIH NBCR (UCSD)
- Pervasive Technology Labs (Heiland, Indiana University)
- Linköping University (Sweden)
- University of North Carolina, Chapel Hill
- UTEP
Writing & Development

An author benefits from working in an environment that simplifies the creation of an executable paper

- Leverage VisTrails’ infrastructure
- Computations specified as workflows
  - Ability to combine tools
  - Support different levels of granularity facilitates the understanding of the computations and results

[Koop et al., ICCS 2011]
An author benefits from working in an environment that simplifies the creation of an executable paper.

- **Leverage VisTrails**
  - Computations specified as workflows
  - Ability to combine tools
  - Support different levels of granularity facilitates the understanding of the computations and results

[Koop et al., ICCS 2011]
An author benefits from working in an environment that simplifies the writing of an executable paper.

- Provenance of data and computations: workflow provenance is not sufficient.

[Koop et al., ICCS 2011]
Sharing an Experiment

Juliana creates an experiment

Fedora 1.2

Extract Isosurface

Display on Screen

Ian tries to run Juliana’s experiment

File not found!

Cannot execute
Writing & Development

An author benefits from working in an environment that simplifies the writing of an executable paper

- Provenance of data and computations: workflow is not sufficient
- Need ‘more’ information: computational environment (OS, library versions, etc.)
  - Also use virtual machines, CDEPack
- Need better file management
  - Designed support for strong links between data and their provenance [Koop@SSDBM2010]
  - Use versioning servers (e.g., GIT, SVN, Oracle DBFS)
- Connect results to their provenance
  - Support LaTeX, Word, Powerpoint, HTML, wiki [Koop et al., ICCS 2011]
Review & Validation

Improve the quality of reviews: reviewers have the ability to explore and validate conclusions

- Execution environment
  - Use provenance, virtual machines, CDEPack to deal with software dependencies
  - Support local, remote, and mixed execution: alternatives to handle proprietary code and data, special hardware

- Testing and validating computations and their results
  - Reproduce
  - Workability: explore parameters and configurations the authors might not have described in the paper
  - VisTrails’ data exploration infrastructure comes in handy here

[Koop et al., ICCS 2011]
Publishing, Maintenance, & Re-Use

- Simplify interaction: the VisMashup system [Santos@TVCG2009]
- Publish using different media, not just documents
Publishing, Maintenance, & Re-Use

- Simplify interaction: the VisMashup system [Santos@TVCG2009]
- Publish using different media, not just documents
- Maintenance and longevity
  - Software evolves: need to upgrade experiments [Koop@IPAW2010]
- Querying and re-using published experiments [Freire et al., VLDB 2011]
  - Opportunities for knowledge discovery and re-use
  - A search/query engine for experiments: text + structure [Scheidegger@TVCG2007]
  - Can we discover better approaches to a given problem? Or discover relationships among workflows and problems?
  - Can we combine multiple results?
Current Uses and Experiences

- ALPS community: ETH group has published a number of reproducible papers!
- Simulations of computational fluid dynamics
- Database research:
  - experiments using distributed database systems, querying Wikipedia
  - http://www.vistrails.org/index.php/RepeatabilityCentral
Current Uses and Experiences

- ACM SIGMOD repeatability effort [Bonnet et al., SIGMOD Record 2011 to appear]
  - Since 2008 verifies the experiments published in accepted papers
  - Papers submitted for reproducibility evaluation: 2010-20 submissions; 2011-31 submissions
  - In 2011, lay out a set of guidelines to simplify and expedite the reviewing process; provided tutorials
  - Review was still challenging
    » Common problem: setup failed due to implicit dependencies
    » Easy to solve with a virtual machine...
  - Reasons for not submitting:
    » Intellectual property rights on software
    » Sensitive data
    » Specific hardware requirements

http://www.sigmod2011.org/calls_papers_sigmod_research_repeatability.shtml
Current Uses and Experiences

- ACM SIGMOD repeatability effort [Bonnet et al., SIGMOD Record 2011 to appear]
  - Since 2008 verifies the experiments published in accepted papers.
  - Papers submitted for reproducibility evaluation: 2010-20 submissions; 2011-31 submissions
  - In 2011, lay out a set of guidelines to simplify and expedite the reviewing process; provided tutorials.
  - Review was still challenging
    » Common problem: setup failed due to implicit dependencies
    » Easy to solve with a virtual machine…
  - Reasons for not submitting:
    » Intellectual property rights on software
    » Sensitive data
    » Specific hardware requirements

http://www.sigmod2011.org/calls_papers_sigmod_research_repeatability.shtml
Going Forward

- Need more and better incentives:
  - seal of quality, higher quality software/experiments, easier for newcomers in a project, citations, recognition
- Need a whip(?): Some disciplines require data for publications, should we require computational experiments too? ETH does!
- Need better tools
  - There is no one-size-fits-all solution
  - Many groups building tools---we should join forces and build a Reproducibility Toolkit
- Need standards and guidelines for authors and tool developers
- Need provenance support in applications
  - Integrate provenance from different sources, connect the results
Provenance Everywhere

- Provenance-rich presentation
  - Bird_table version 123476 in Oracle
  - STEM v 1.2 + predictions
  - R Script+results
  - BirdVis vis spec

predictions
results

BirdVis

predictions

matplotlib

Bird_table version 123476

matplotlib

Sparrow

Reproducible Research '11
A Little History and a Challenge

- A long time ago, when I was a PhD student, generating the reference list for papers was very time consuming
  - Find proceedings on the shelf (or walk to library), flip pages to obtain page numbers, type (title, authors, proceedings name, etc.)
- Today
  - Google/Bing author or part of paper title, DBLP, ACM DL, IEEE Explore
  - Copy bib entry in one of many formats (bibtex, EndNote, plain text), paste in paper, voilà!

- Can we do the same for scientific experiments?
Conclusions and Future Work

- Provenance is crucial for science and an enabler for executable papers
- Provenance must be at the center of the scientific process!
- Built an end-to-end solution based on VisTrails---currently working on integrating infrastructure with other systems
  - Provenance-enabling other tools
- Many challenges and several open research questions
- Great opportunity to have impact in science
Additional Information

- The VisTrails System [http://www.vistrails.org](http://www.vistrails.org)
- An infrastructure to support the creation, review and re-use of reproducible papers [http://www.vistrails.org/index.php/ExecutablePapers](http://www.vistrails.org/index.php/ExecutablePapers)
Acknowledgments

- Thanks to: Philippe Bonnet, Philip Mates, Matthias Troyer, Dennis Shasha, Emanuele Santos, Claudio Silva, Joel Tohline, Huy T. Vo, and the VisTrails team.
- This work is partially supported by the National Science Foundation, the Department of Energy, and IBM Faculty Awards.
Merci
Ευχαριστώ
Thank you
Obrigada