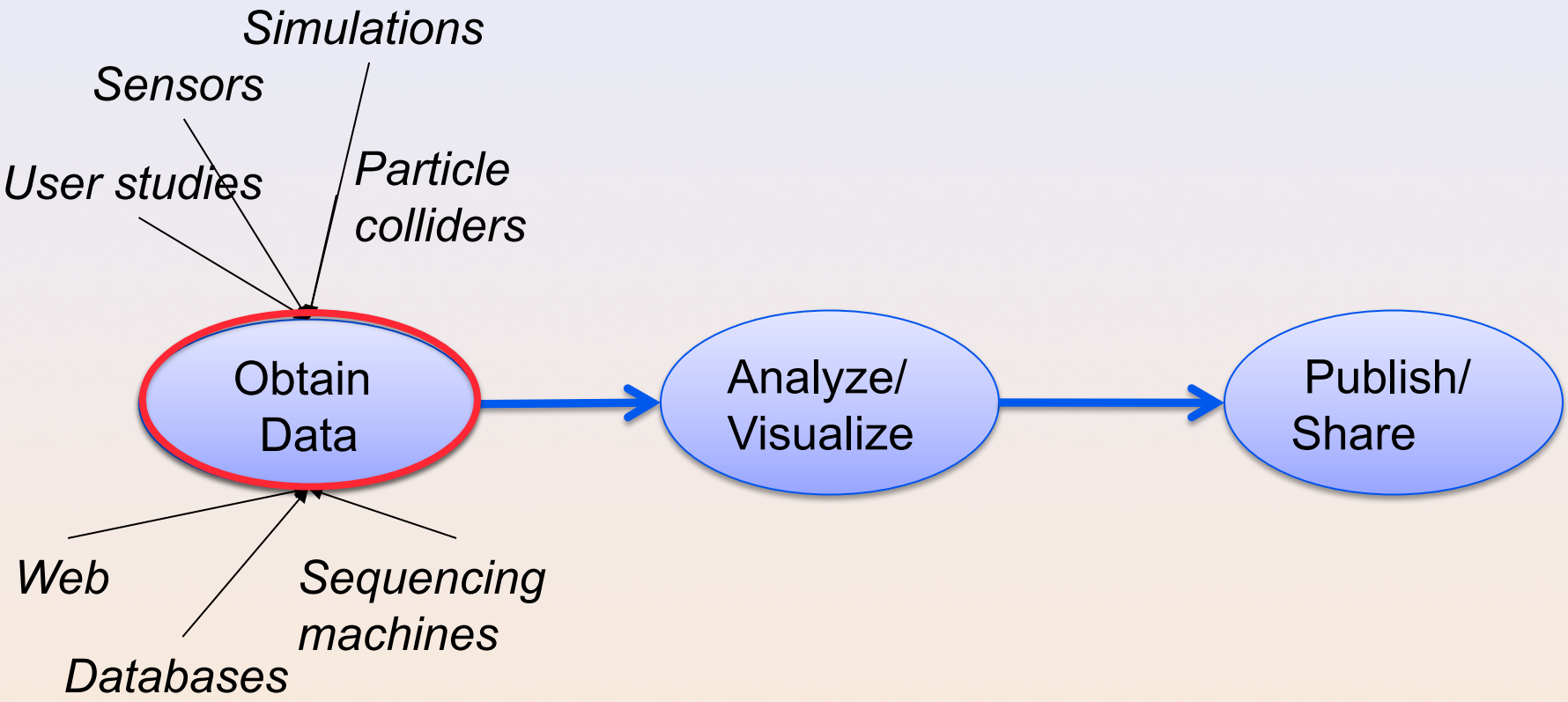
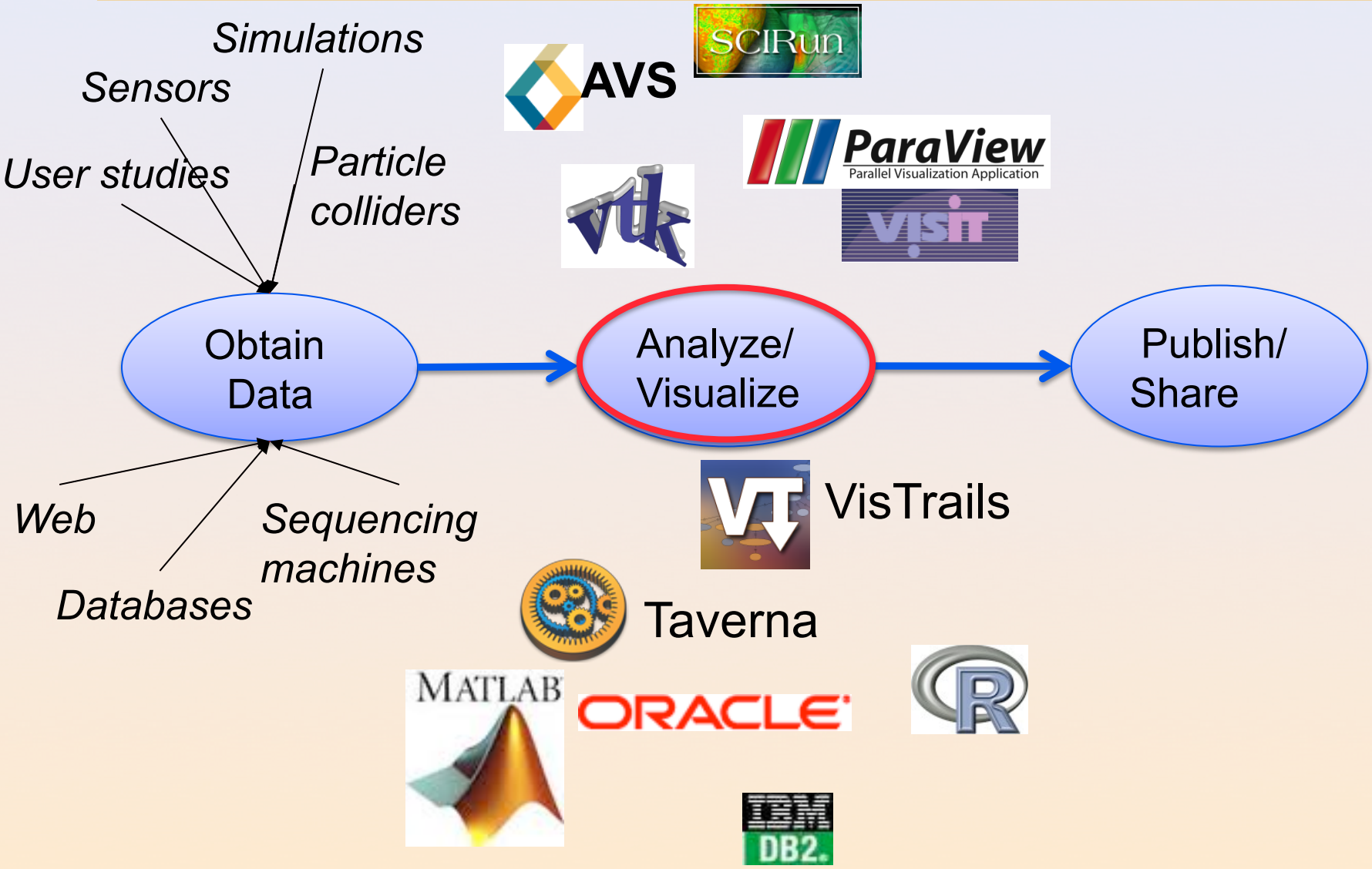
Provenance-Rich Publications

Juliana Freire
VisTrails Group

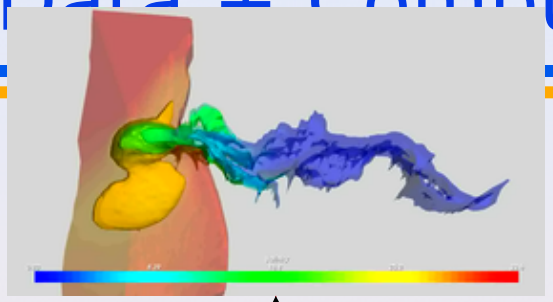
Science Today: Data Intensive



Science Today: Data + Computing Intensive



Science Today: Data + Computing Intensive



Simulations

Sensors

User studies

Particle colliders

Microsoft PowerPoint

Obtain Data

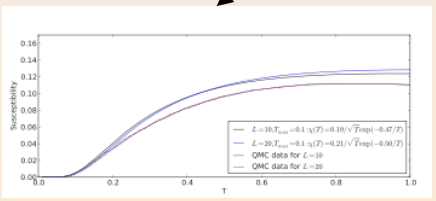
Analyze/ Visualize

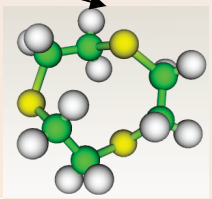
Publish/ Share

Web

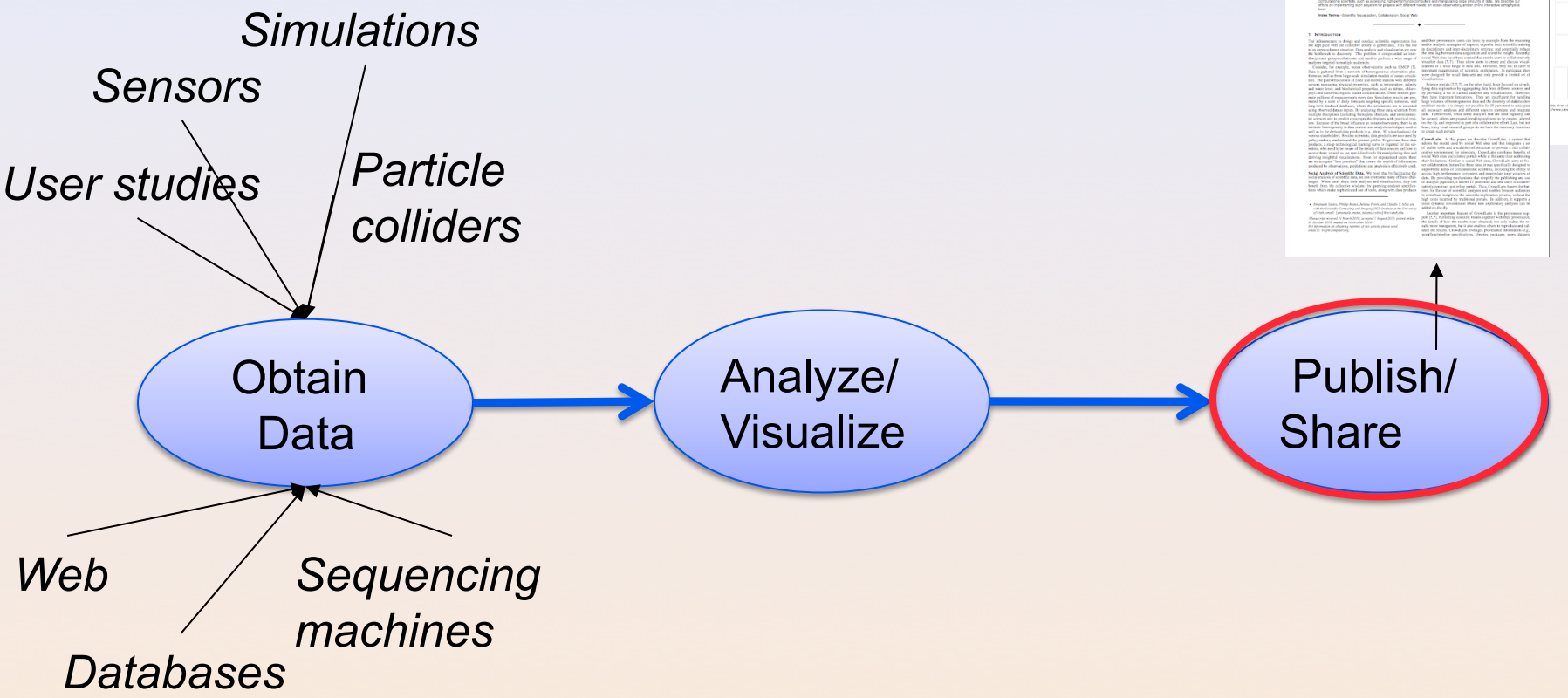
Sequencing machines

Databases





Science Today: Data + Computing Inte



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

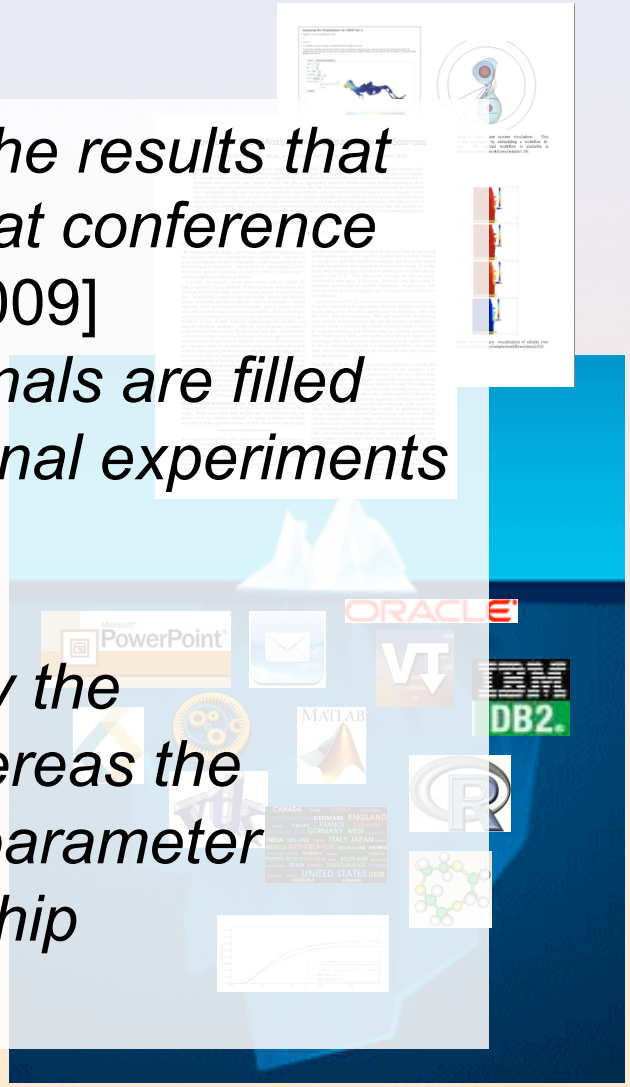
Science Today: Incomplete Publications

- 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]



Provenance-Rich Publications

- Bridge the gap between the scientific process and publications
  - The scientific record needs to be *complete and trustworthy*
  - Papers with *deep* captions
- Show me the proof: results that can be reproduced and validated
  - Encouraged by ACM SIGMOD, HPDC, a number of journals, funding agencies, academic institutions (e.g., 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: Tools and Strategies for Scientific Computing

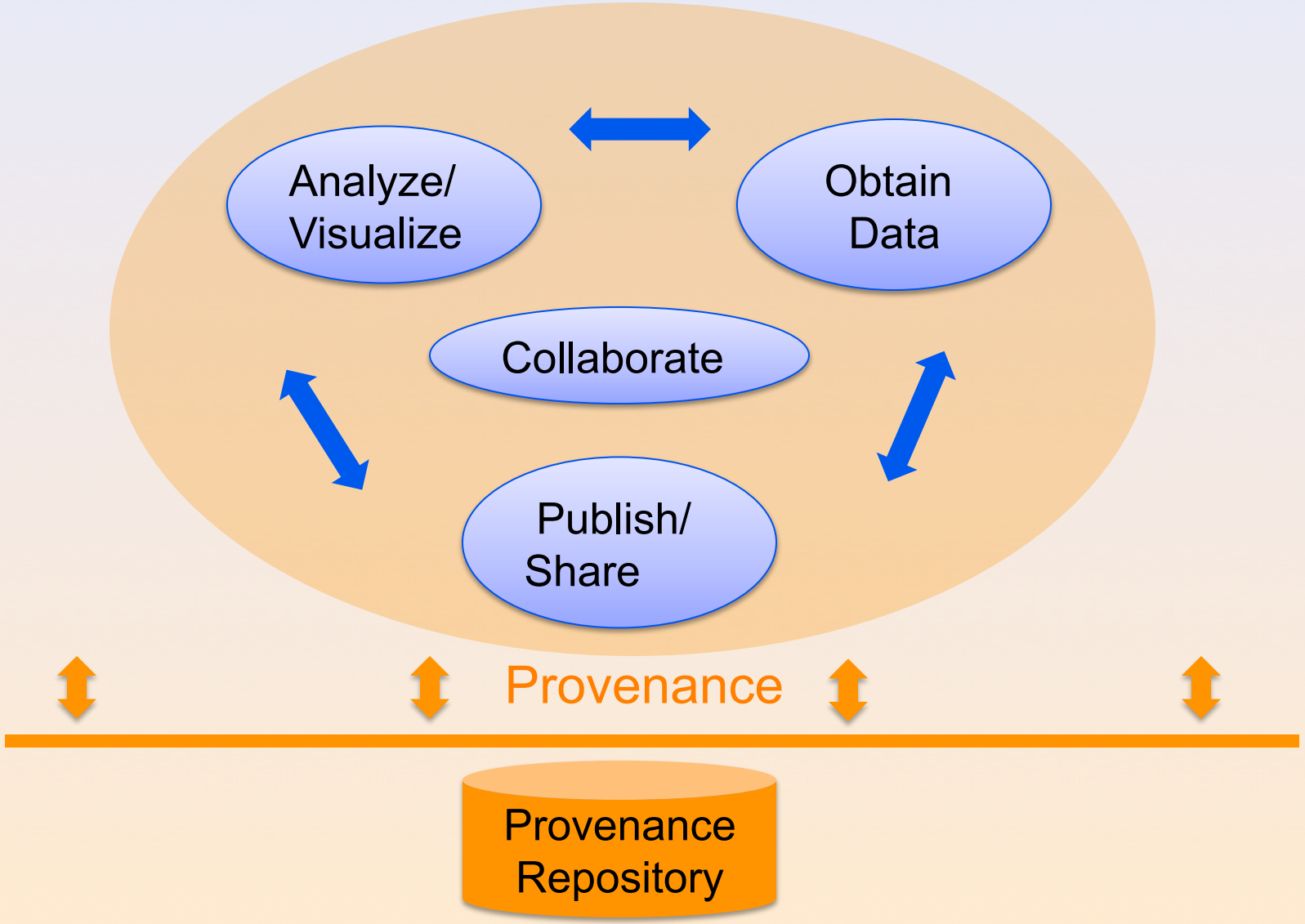
Provenance-Rich Publications: Benefits

- Produce more knowledge---not just text
- 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 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…
- How to represent and query experiments?
  - Workflows, source code, data
- Need to simplify the process of sharing, reviewing and re-using scientific experiments and results

Need Provenance-Rich Science



Our Approach: An Infrastructure to Support Provenance-Rich Papers

- Tools for *authors* to create *workflows* that encode the computational processes, package the results, and link from publications
  - Support different approaches to packaging workflows/data/ environment for publication
- Tools for testers to repeat and validate results
  - How to generate experiments that are most informative given a time/resource limit?
- 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?

An *Provenance-Rich* Paper: ALPS2.0

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

B. Bauer[1] L. D. Carr[2] A. Feiguin[3] J. Freire[4] S. Fuchs[5]
L. Gamper[1] J. Gukelberger[1] E. Gull[6] S. Guertler[7] A. Hehn[1]
R. Igarashi[8,9] S.V. Isakov[1] D. Koop[4] P.N. Ma[1] P. Mates[1,4]
H. Matsuo[10] O. Parcollet[11] G. Pawłowski[12] J.D. Picon[13]
L. Pollet[1,14] E. Santos[4] V.W. Scarola[15] U. Schollwöck[16] C. Silva[4]
B. Surer[1] S. Todo[9,10] S. Trebst[17] M. Troyer[1]‡ M.L. Wall[2]
P. Werner[1] S. Wessel[18,19]

[1]Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland
[2]Department of Physics, Colorado School of Mines, Golden, CO 80401, USA
[3]Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA
[4]Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA
[5]Institut für Theoretische Physik, Georg-August-Universität Göttingen, Göttingen, Germany
[6]Columbia University, New York, NY 10027, USA
[7]Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany
[8]Center for Computational Science & e-Systems, Japan Atomic Energy Agency, 110-0015 Tokyo, Japan
[9]Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, 332-0012 Kawaguchi, Japan
[10]Department of Applied Physics, University of Tokyo, 113-8656 Tokyo, Japan
[11]Institut de Physique Théorique, CEA/DSM/IPhT-CNRS/URA 2306, CEA-Saclay, F-91191 Gif-sur-Yvette, France
[12]Faculty of Physics, A. Mickiewicz University, Umultowska 85, 61-614 Poznań, Poland
[13]Institute of Theoretical Physics, EPF Lausanne, CH-1015 Lausanne, Switzerland
[14]Physics Department, Harvard University, Cambridge 02138, Massachusetts, USA
[15]Department of Physics, Virginia Tech, Blacksburg, Virginia 24061, USA
[16]Department for Physics, Arnold Sommerfeld Center for Theoretical Physics and Center for NanoScience, University of Munich, 80333 Munich, Germany
[17]Microsoft Research, Station Q, University of California, Santa Barbara, CA 93106, USA
[18]Institute for Solid State Theory, RWTH Aachen University, 52056 Aachen, Germany
[19]Institut für Theoretische Physik III, Universität Stuttgart, Pfaffenwaldring 57, 70550 Stuttgart, Germany

‡ Corresponding author: troyer@comp-phys.org

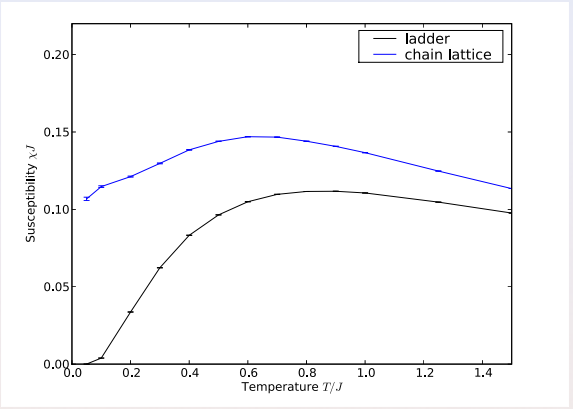
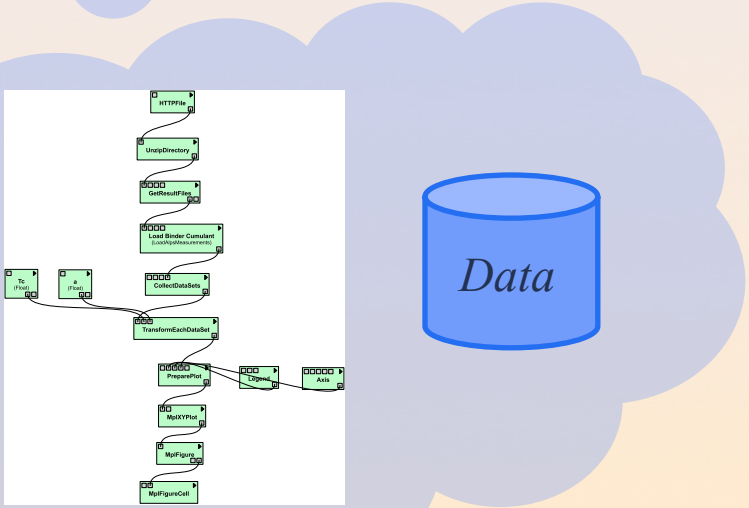


Figure 1. A figure produced by an ALPS VisTrails workflow: the uniform susceptibility of the Heisenberg chain and ladder. Clicking the figure retrieves the workflow used to create it. Opening that workflow on a machine with VisTrails and ALPS installed lets the reader execute the full calculation.



http://adsabs.harvard.edu/abs/2011arXiv1101.2646B

An *Executable* Paper: ALPS2.0

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

B. Bauer[1] L. D. Carr[2] A. Feiguin[3] J. Freire[4] S. Fuchs[5]
L. Gamper[1] J. Gukelberger[1] E. Gull[6] S. Guertler[7] A. Hehn[1]
R. Igarashi[8,9] S.V. Isakov[1] D. Koop[4] P.N. Ma[1] P. Mates[1,4]
H. Matsuo[10] O. Parcollet[11] G. Pawłowski[12] J.D. Picon[13]
L. Pollet[1,14] E. Santos[4] V.W. Scarola[15] U. Schollwöck[16] C. Silva[4]
B. Surer[1] S. Todo[9,10] S. Trebst[17] M. Troyer[1]‡ M.L. Wall[2]
P. Werner[1] S. Wessel[18,19]

[1]Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland
[2]Department of Physics, Colorado School of Mines, Golden, CO 80401, USA
[3]Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA
[4]Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA
[5]Institut für Theoretische Physik, Georg-August-Universität Göttingen, Göttingen, Germany
[6]Columbia University, New York, NY 10027, USA
[7]Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany
[8]Center for Computational Science & e-Systems, Japan Atomic Energy Agency, 110-0015 Tokyo, Japan
[9]Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, 332-0012 Kawaguchi, Japan
[10]Department of Applied Physics, University of Tokyo, 113-8656 Tokyo, Japan
[11]Institut de Physique Théorique, CEA/DSM/IPhT-CNRS/URA 2306, CEA-Saclay, F-91191 Gif-sur-Yvette, France
[12]Faculty of Physics, A. Mickiewicz University, Umultowska 85, 61-614 Poznań, Poland
[13]Institute of Theoretical Physics, EPF Lausanne, CH-1015 Lausanne, Switzerland
[14]Physics Department, Harvard University, Cambridge 02138, Massachusetts, USA
[15]Department of Physics, Virginia Tech, Blacksburg, Virginia 24061, USA
[16]Department for Physics, Arnold Sommerfeld Center for Theoretical Physics and Center for NanoScience, University of Munich, 80333 Munich, Germany
[17]Microsoft Research, Station Q, University of California, Santa Barbara, CA 93106, USA
[18]Institute for Solid State Theory, RWTH Aachen University, 52056 Aachen, Germany
[19]Institut für Theoretische Physik III, Universität Stuttgart, Pfaffenwaldring 57, 70550 Stuttgart, Germany

‡ Corresponding author: troyer@comp-phys.org

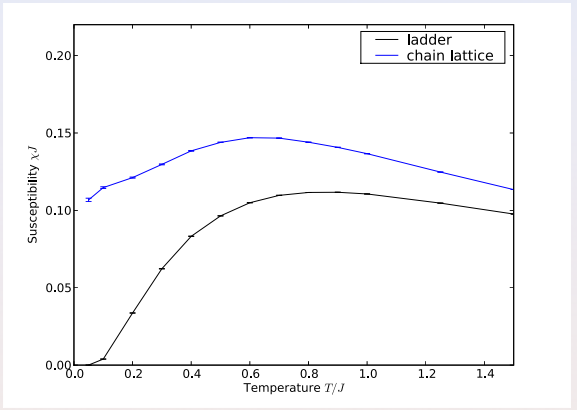
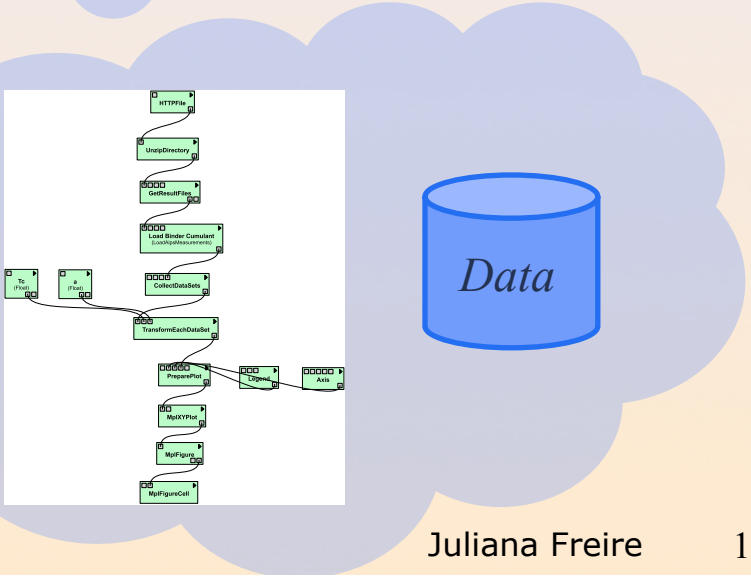


Figure 1. A figure produced by an ALPS VisTrails workflow: the uniform susceptibility of the Heisenberg chain and ladder. Clicking the figure retrieves the workflow used to create it. Opening that workflow on a machine with VisTrails and ALPS installed lets the reader execute the full calculation.



http://adsabs.harvard.edu/abs/2011arXiv1101.2646B

Demo

Editing an executable paper written using LaTeX and VisTrails
http://www.vistrails.org/download/download.php?type=MEDIA&id=executable_paper_latex.mov

Exploring a Web-hosted paper using server-based computation
http://www.vistrails.org/download/download.php?type=MEDIA&id=executable_paper_server.mov

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

An Infrastructure to Support Provenance-Rich Papers

- Writing & Development
  - Specifying computations
  - Provenance of data and computations
  - Execution infrastructure
- Review & Validation
  - Local, remote, and mixed execution
  - Interacting, testing and validating computations and their results
- Publishing, Maintenance, & Re-Use
  - Maintenance and longevity
  - Querying and re-using published results

Writing & Development

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

- Leverage VisTrails' infrastructure

The VisTrails System

- Workflow-based system for data analysis and visualization
- Comprehensive *provenance infrastructure*
- *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 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 (Pyschiatry, 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 for different levels of granularity can facilitate the understanding of the computations and results (see Marta's talk)

Writing & Development

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

- Provenance of data and computations
  - Capturing workflow provenance is not sufficient
  - E.g., workflow reads /Users/juliana/a.out, invokes Python 3; and Peter tries to run it on his laptop. Tough luck…
  - Need 'more' provenance, besides parameters and input data, computational environment (OS, library versions, etc.) --- virtual machines, CDEPack
  - Need better file management, e.g., strong links between data and their provenance [Koop@SSDBM2010]
- Connecting results to their provenance
  - LateX, Word, Powerpoint, HTML, wikis

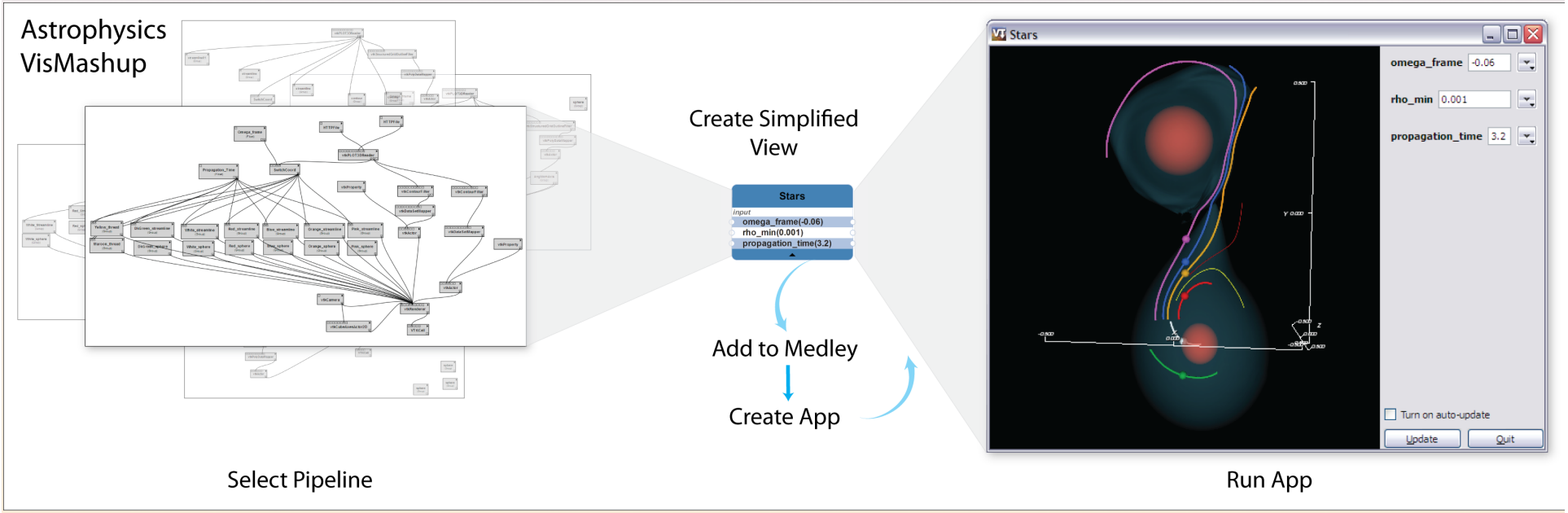
Review & Validation

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

- Execution environment
  - Software dependencies; proprietary code and data; special hardware
  - Virtual machines, CDEpack
  - Local, remote, and mixed execution
- Testing and validating computations and their results
  - Reproduce
  - Workability: explore parameters and configurations the authors might not have described in the paper
  - Obtain insights
  - Data exploration infrastructure

Publishing, Maintenance, & Re-Use

- Simplify interaction: the VisMashup system [Santos@TVCG2009]

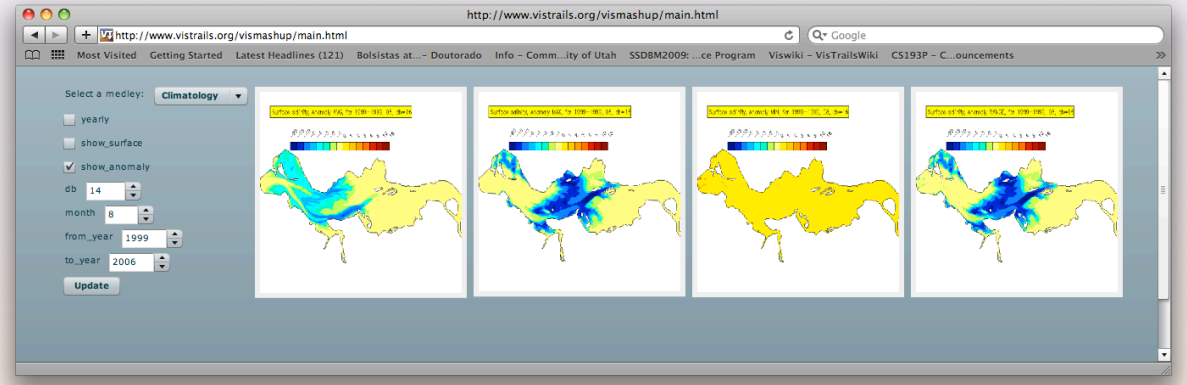


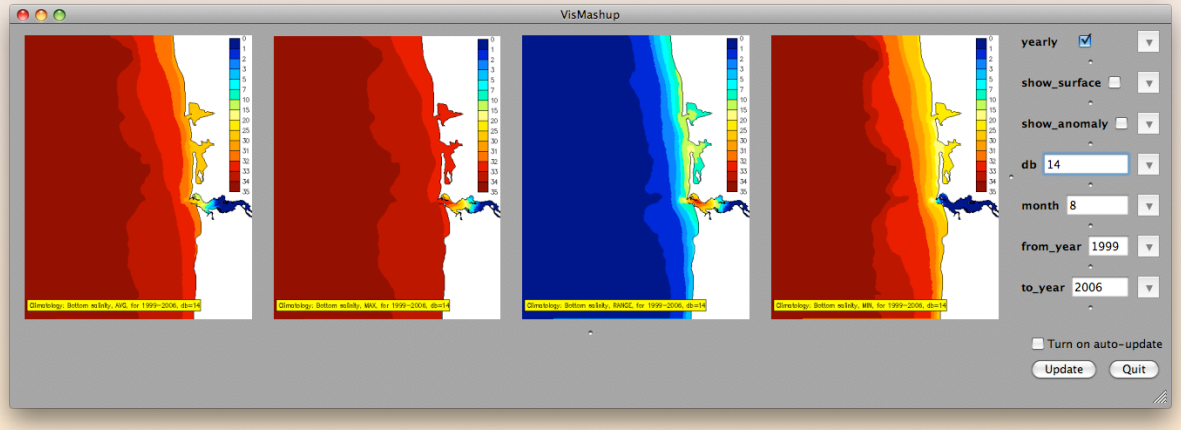
Publishing, Maintenance, & Re-Use

- Simplify interaction: the VisMashup system
- Publish using different media

Web

Portable Devices





Desktop

Publishing, Maintenance, & Re-Use

- Simplify interaction: the VisMashup system
- Publish using different media
- Maintenance and longevity:
  - Software evolves, try new algorithms: need to *upgrade* workflows [Koop@IPAW2010]
- Querying and re-using published results
  - 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

- ALPS community
- Simulations of computational fluid dynamics
- Databases:
  - experiments using distributed database systems, querying Wikipedia
  - http://www.vistrails.org/index.php/RepeatabilityCentral
- ACM SIGMOD repeatability effort
  - Since 2008 verifies the experiments published in accepted papers
  - In 2010, 20% of the papers got the reproducibility stamp!
  - In 2011, lay out a set of guidelines to simplify and expedite the reviewing process
  - http://www.sigmod2011.org/calls_papers_sigmod_research_repeatability.shtml

Conclusions and Future Work

- Provenance is crucial for science and an enabler for *executable* papers
- But we need more than just workflows: Provenance must be at the center of the scientific process!
- As a starting point, we built an end-to-end solution based on VisTrails---currently working on integrating infrastructure with other systems
- Many different requirements: need to mix and match different components, need to support multiple tools

Conclusions and Future Work

- Sharing provenance-rich papers creates new opportunities
  - Expose users to different techniques and tools
  - Users can learn by example; expedite their training; and potentially reduce their time to insight
  - Better science! (e.g., http://www.nytimes.com/2010/08/13/health/research/13alzheimer.html)
- Many challenges and several open computer science and database research questions!

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.

Ευχαριστω
Thank you
Obrigada