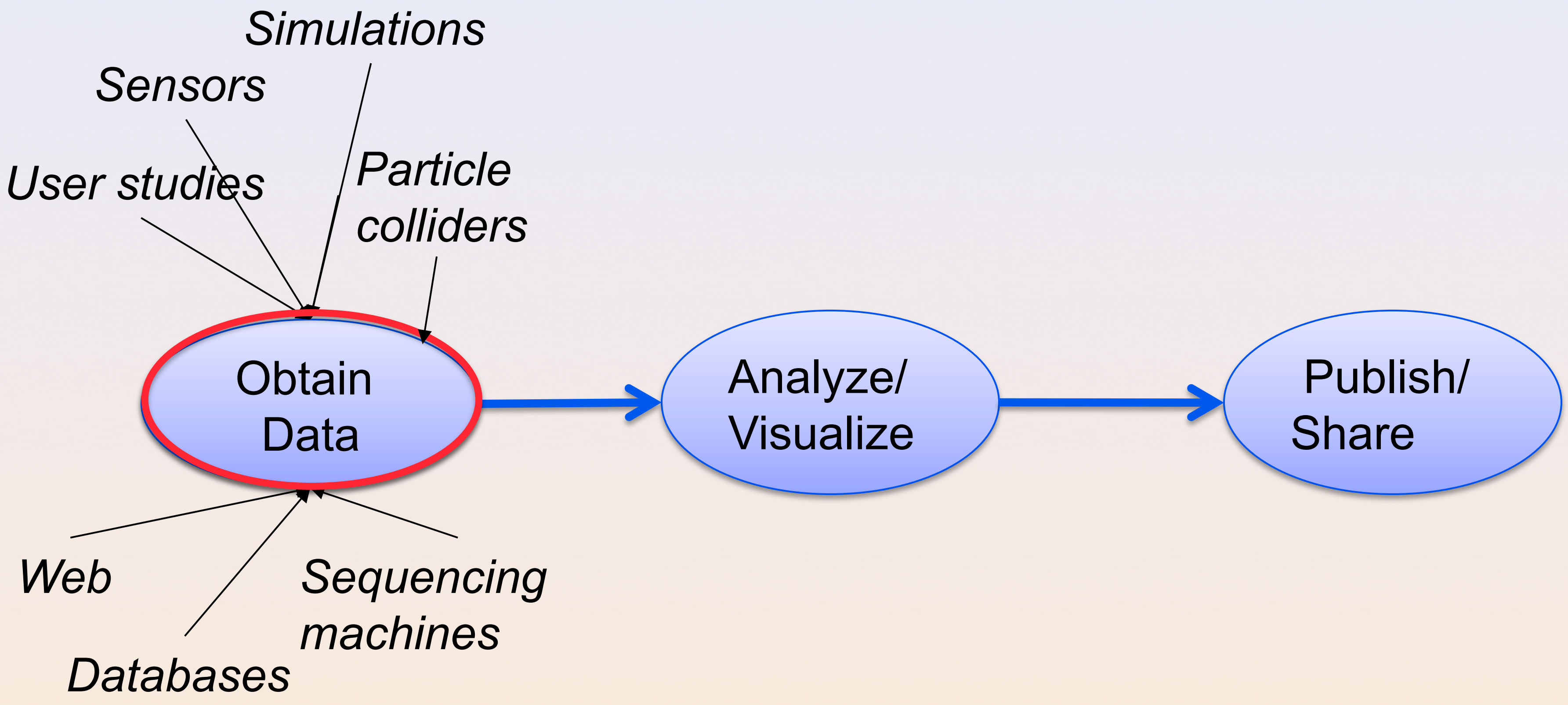
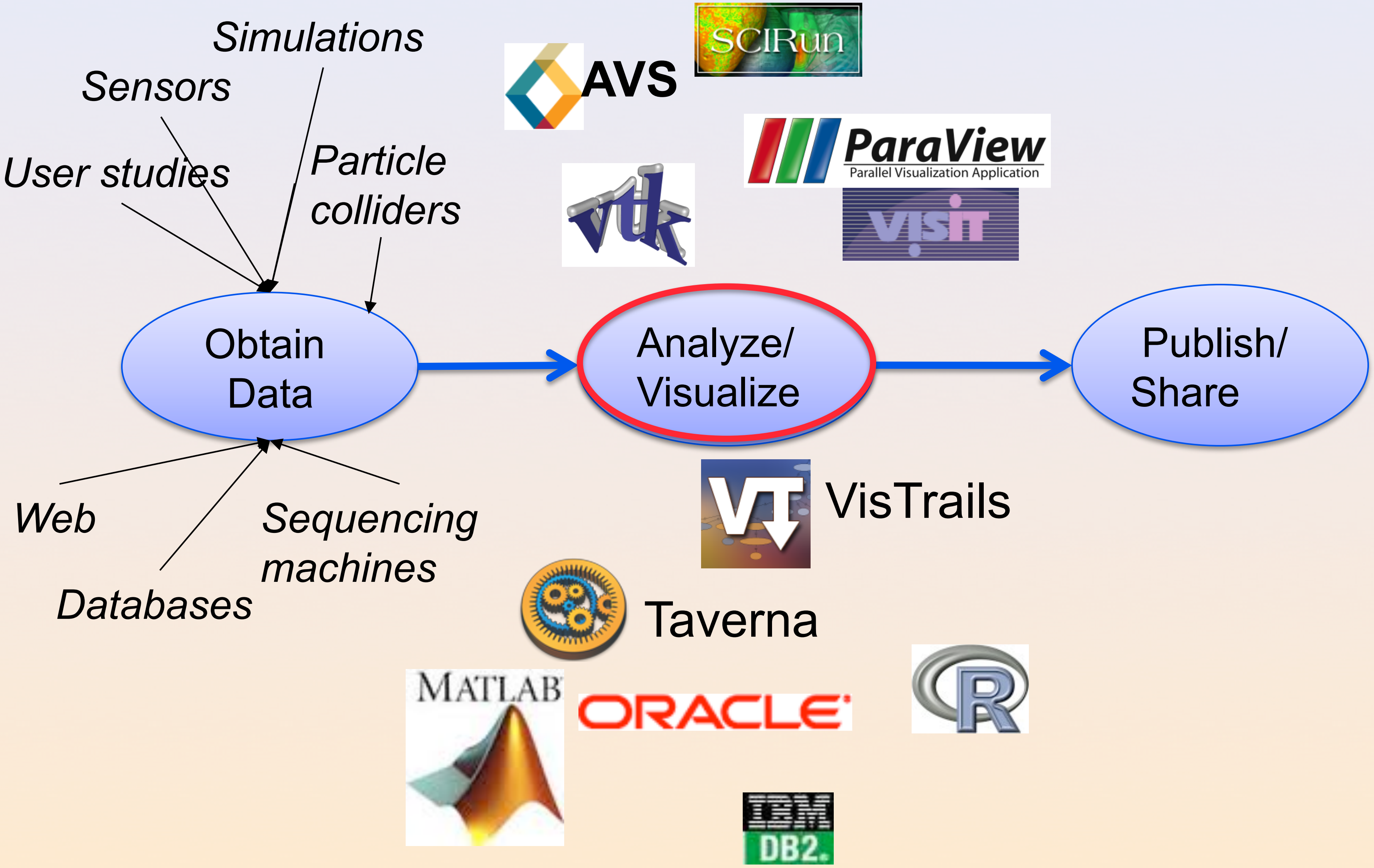
A PROVENANCE-BASED INFRASTRUCTURE FOR CREATING REPRODUCIBLE PAPERS

Juliana Freire
Juliana.freire@nyu.edu
VisTrails Group & Web and Databases Lab
NYU Poly

Science Today: Data Intensive



Science Today: Data + Computing Intensive



Simulations

Sensors

User studies

Particle colliders

Obtain Data

Analyze/ Visualize

Publish/ Share

Web

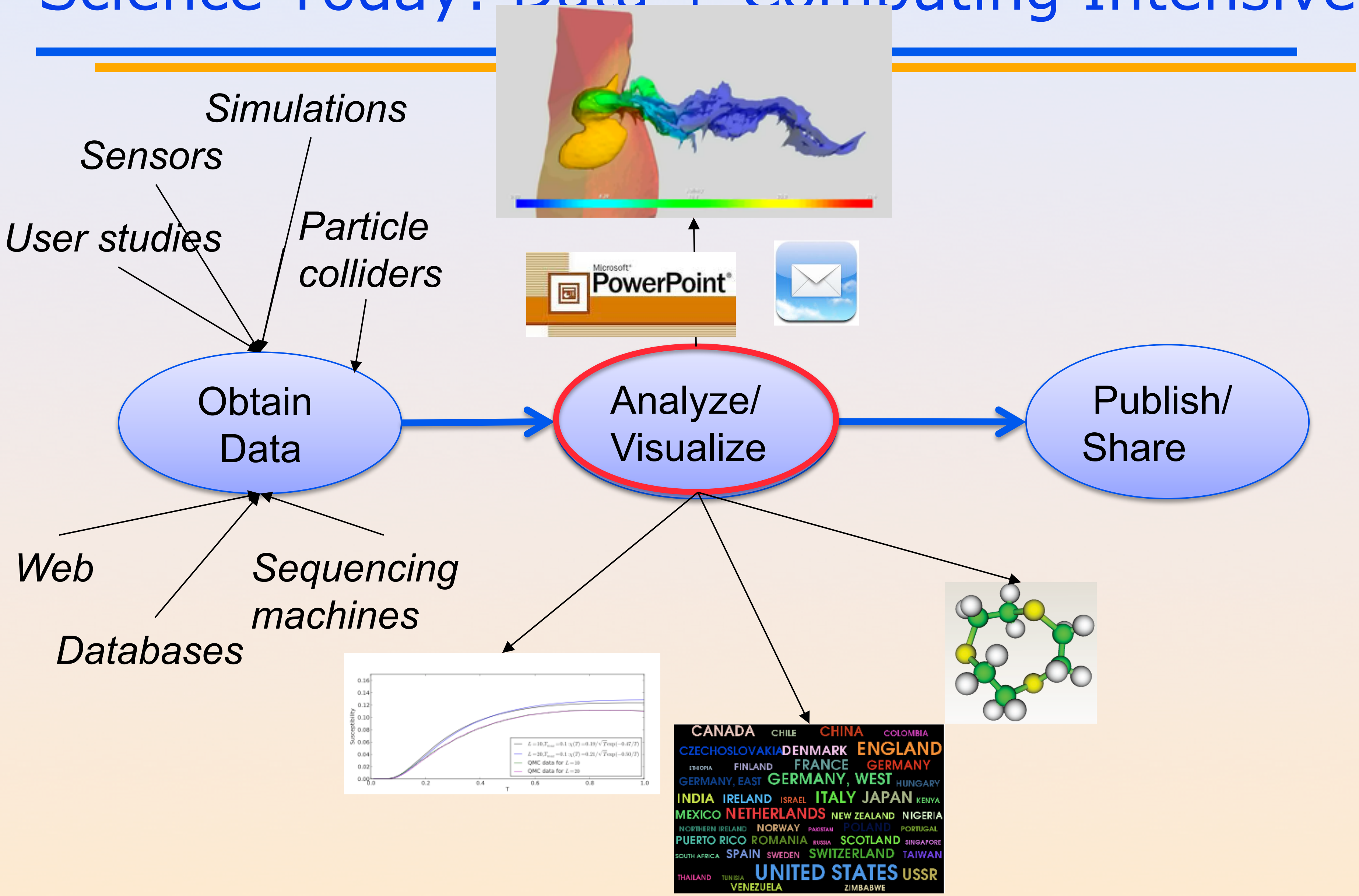
Databases

Sequencing machines

VisTrails

Taverna

Science Today: Data + Computing Intensive



Simulations

Sensors

User studies

Particle colliders

Web

Databases

Sequencing machines

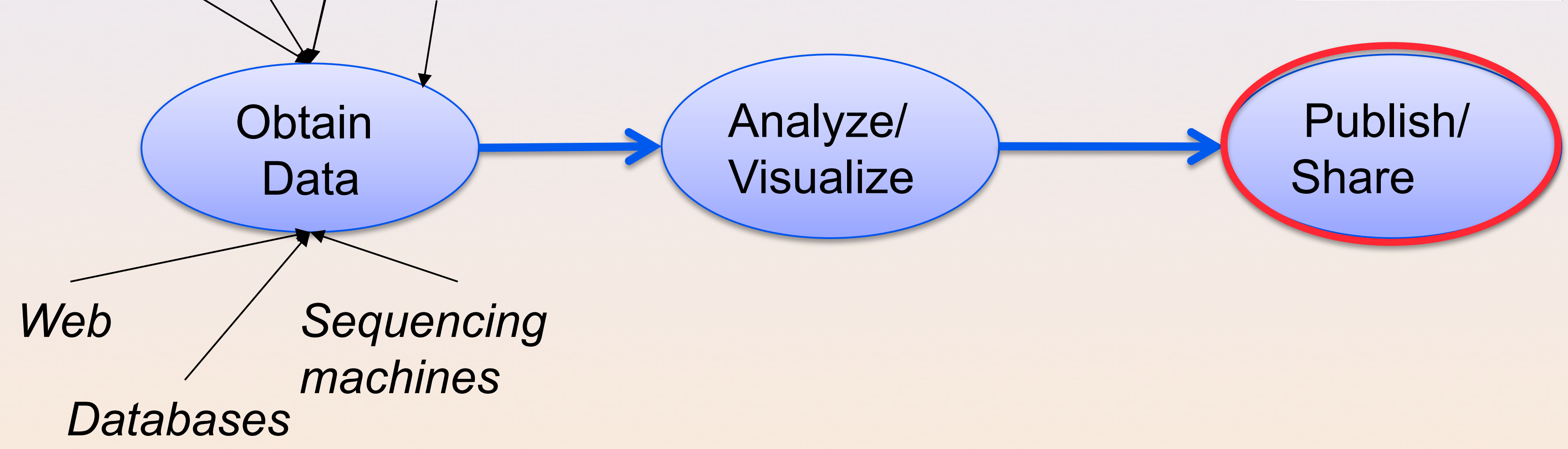
Obtain Data

Analyze/ Visualize

Publish/ Share

Microsoft® PowerPoint®
Susceptibility
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.0
0.2
0.4
0.6
0.8
1.0
T
$L=10, T_{max}=0.1: \chi(T)=0.19/\sqrt{T}\exp(-0.47/T)$
$L=20, T_{max}=0.1: \chi(T)=0.21/\sqrt{T}\exp(-0.50/T)$
QMC data for $L=10$
QMC data for $L=20$
CANADA CHILE CHINA COLOMBIA
CZECHOSLOVAKIA DENMARK ENGLAND
ETHIOPIA FINLAND FRANCE GERMANY
GERMANY, EAST GERMANY, WEST HUNGARY
INDIA IRELAND ISRAEL ITALY JAPAN KENYA
MEXICO NETHERLANDS NEW ZEALAND NIGERIA
NORTHERN IRELAND NORWAY PAKISTAN POLAND PORTUGAL
PUERTO RICO ROMANIA RUSSIA SCOTLAND SINGAPORE
SOUTH AFRICA SPAIN SWEDEN SWITZERLAND TAIWAN
THAILAND TUNISIA UNITED STATES USSR
VENEZUELA ZIMBABWE

Science Today: Data + Computing Inte



Simulations

Sensors

User studies

Particle colliders

Web

Sequencing machines

Databases

CrowdLabs: Social Analysis and Visualization for the Sciences

Emanuele Santos, Phillip Mates, Juliana Freire, and Cláudio T. Silva, Senior Member, IEEE

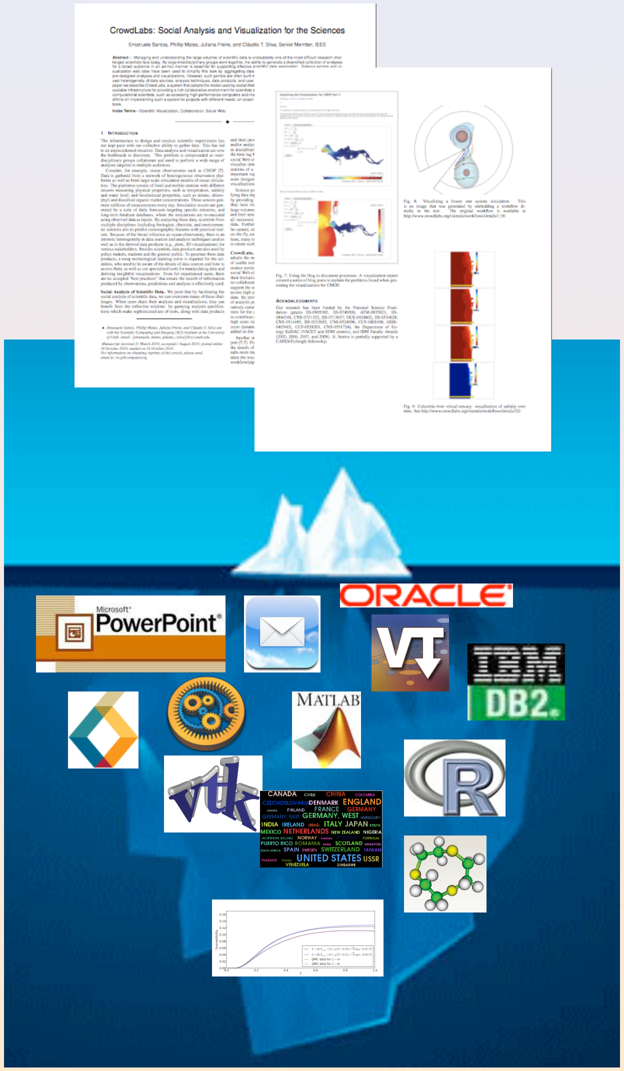
Abstract— Managing and understanding the large volumes of scientific data is undoubtedly one of the most difficult research challenges scientists face today. As large interdisciplinary groups work together, the ability to generate a diversified collection of analyses for a broad audience in an ad-hoc manner is essential for supporting effective scientific data exploration. Science portals and visualization web sites have been used to simplify this task by aggregating data from different sources and by providing a set of pre-designed analyses and visualizations. However, such portals are often built manually, and are not flexible enough to support the vast heterogeneity of data sources, analysis techniques, data products, and user communities that need to access this data. In this paper we describe CrowdLabs, a system that adopts the model used by social Web sites and that combines a set of usable tools and a scalable infrastructure for providing a rich collaborative environment for scientists and that also takes into account the requirements of computational scientists, such as accessing high-performance computers and manipulating large amounts of data. We describe our efforts on implementing such a system for projects with different needs: an ocean observatory, and an online interactive astrophysics book.

Index Terms—Scientific Visualization, Collaboration, Social Web

1 Introduction

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


CrowdLabs: Social Analysis and Visualization for the Sciences
Microsoft PowerPoint
ORACLE
VT
IBM DB2
MATLAB
vtk
R

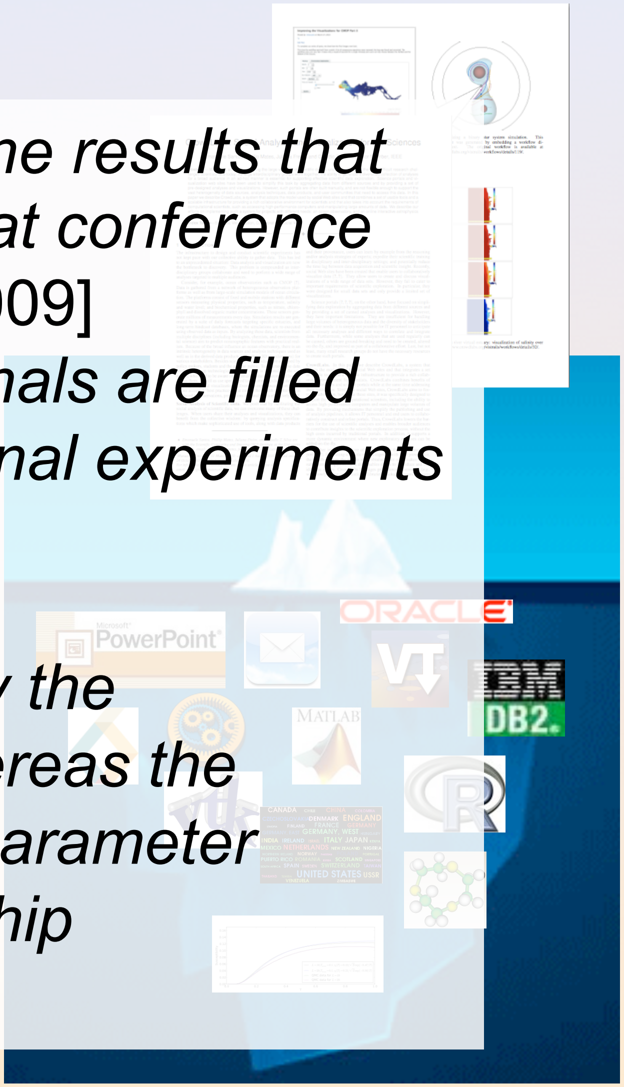
Science Today: Incomplete Publications

- 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

"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

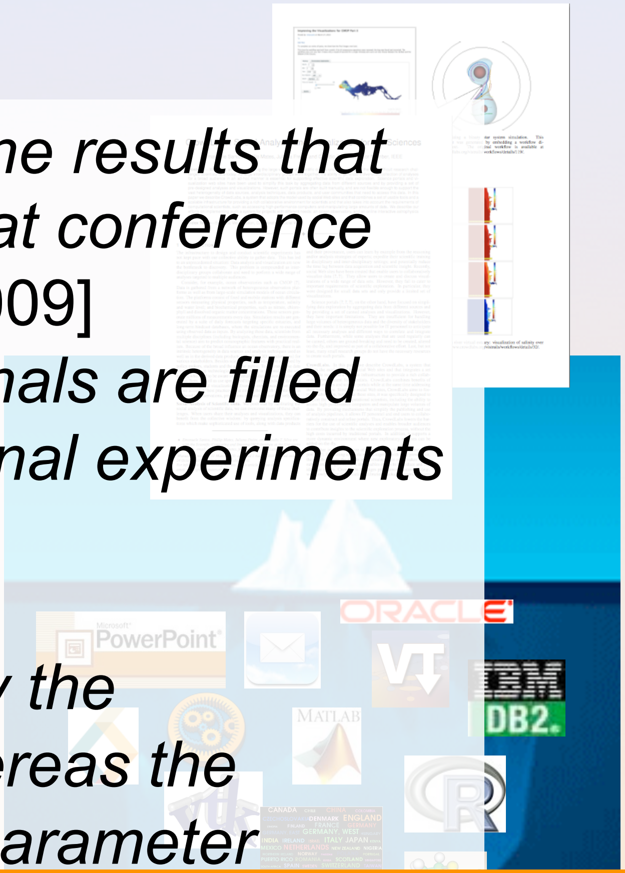
- Publications are just the tip of the iceberg
  - Scientific record is incomplete--- to large to fit in a paper
  - Large volumes of data
  - Complex processes
- Can't (easily) reproduce results

“*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 value*

itself.”

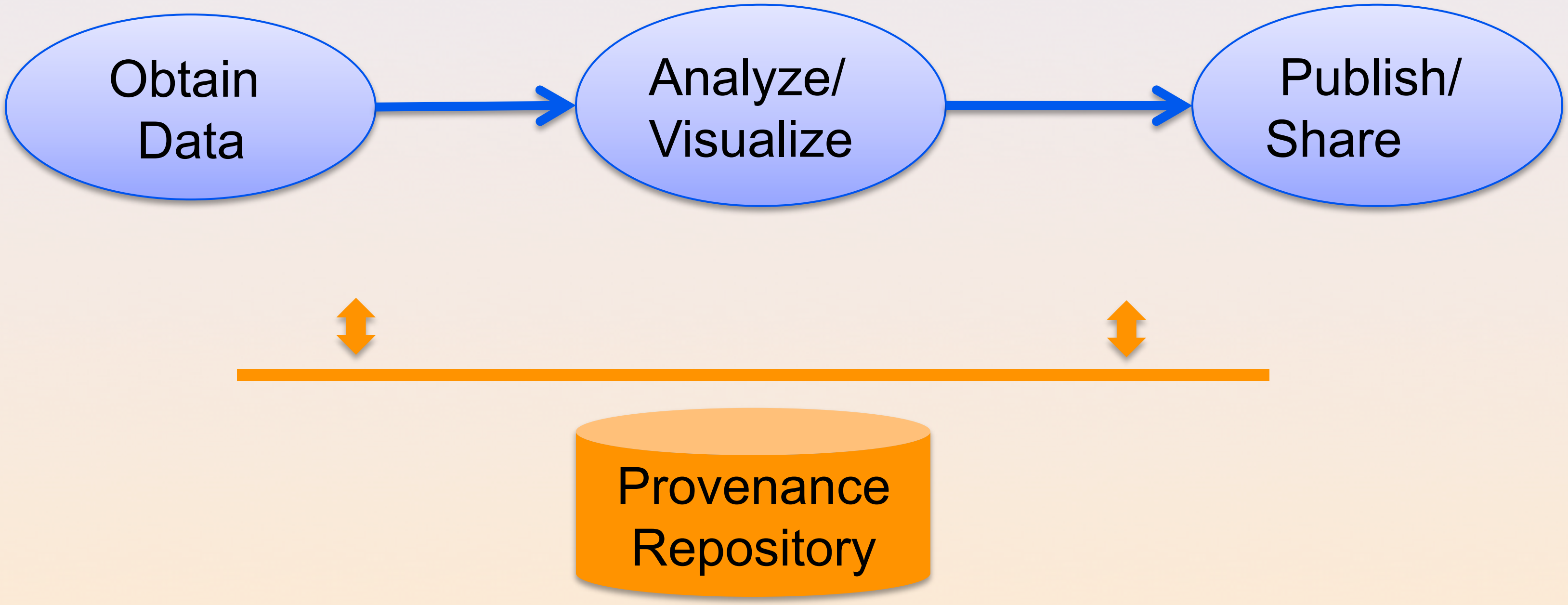


http://en.wikipedia.org/wiki/Scientific_misconduct

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

Nobel Laureate Retracts Two Papers, NYTimes 09/24/2010

Vision: Provenance-Rich Science



Vision: Provenance-Rich Science

Provenance is the scientific record

Provenance Repository

Analy

a

Proven

Repos

CrowdLabs: Social Analysis and Visualization for the Sciences
Microsoft PowerPoint
ORACLE
VT
IBM DB2
MATLAB
vtk
R

lish/

re

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
- 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

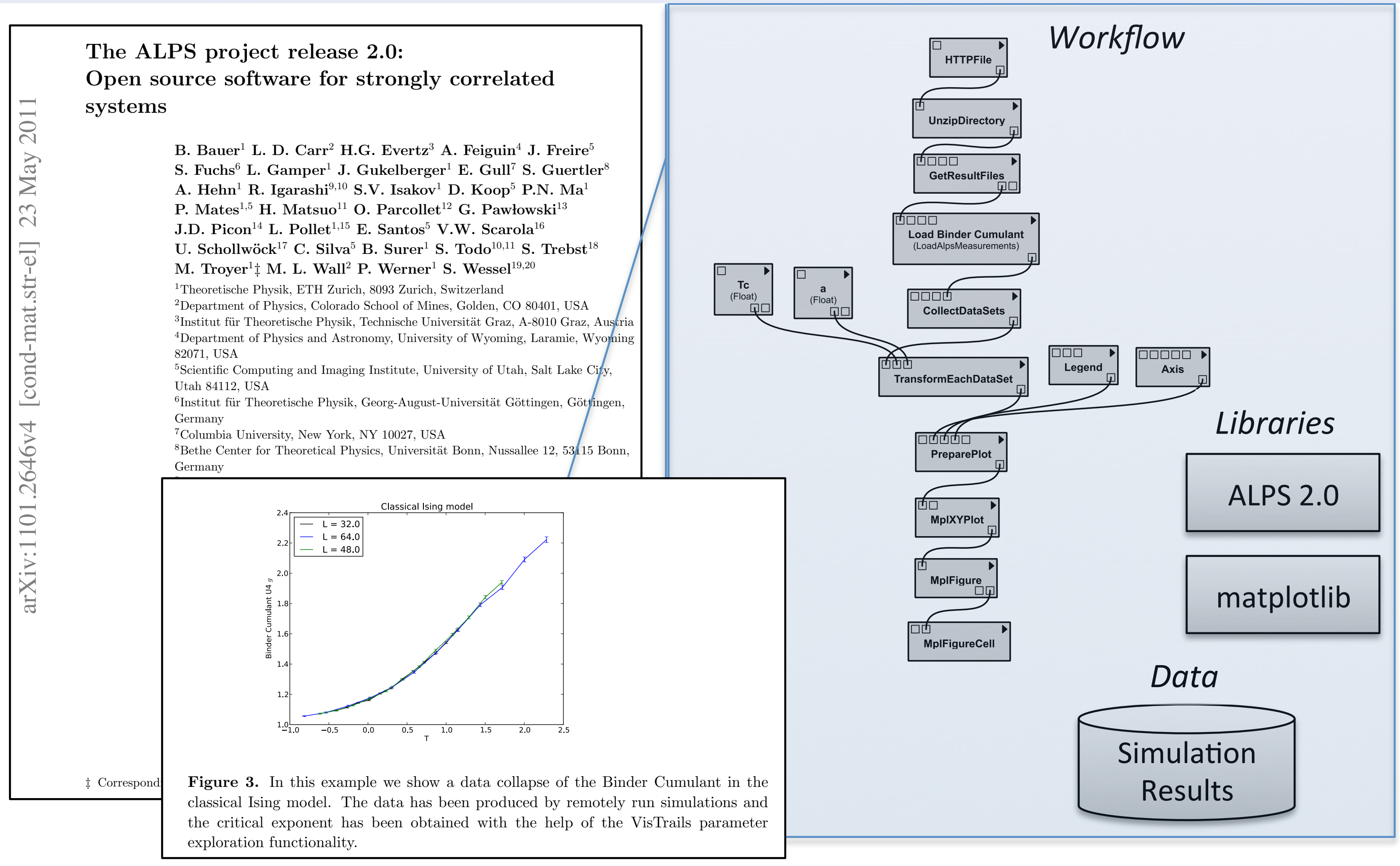
Our Approach: An Infrastructure to Support Provenance-Rich Papers [Koop et al., ICCS 2011]

- 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?

An *Provenance-Rich* Paper: ALPS2.0



arXiv:1101.2646v4 [cond-mat.str-el] 23 May 2011

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

B. Bauer[1] L. D. Carr[2] H.G. Evertz[3] A. Feiguin[4] J. Freire[5] S. Fuchs[6] L. Gamper[1] J. Gukelberger[1] E. Gull[7] S. Guertler[8] A. Hehn[1] R. Igarashi[9,10] S.V. Isakov[1] D. Koop[5] P.N. Ma[1] P. Mates[1,5] H. Matsuo[11] O. Parcollet[12] G. Pawłowski[13] J.D. Picon[14] L. Pollet[1,15] E. Santos[5] V.W. Scarola[16] U. Schollwöck[17] C. Silva[5] B. Surer[1] S. Todo[10,11] S. Trebst[18] M. Troyer[1]‡ M. L. Wall[2] P. Werner[1] S. Wessel[19,20]

[1] Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland
[2] Department of Physics, Colorado School of Mines, Golden, CO 80401, USA
[3] Institut für Theoretische Physik, Technische Universität Graz, A-8010 Graz, Austria
[4] Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA
[5] Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA
[6] Institut für Theoretische Physik, Georg-August-Universität Göttingen, Göttingen, Germany
[7] Columbia University, New York, NY 10027, USA
[8] Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany

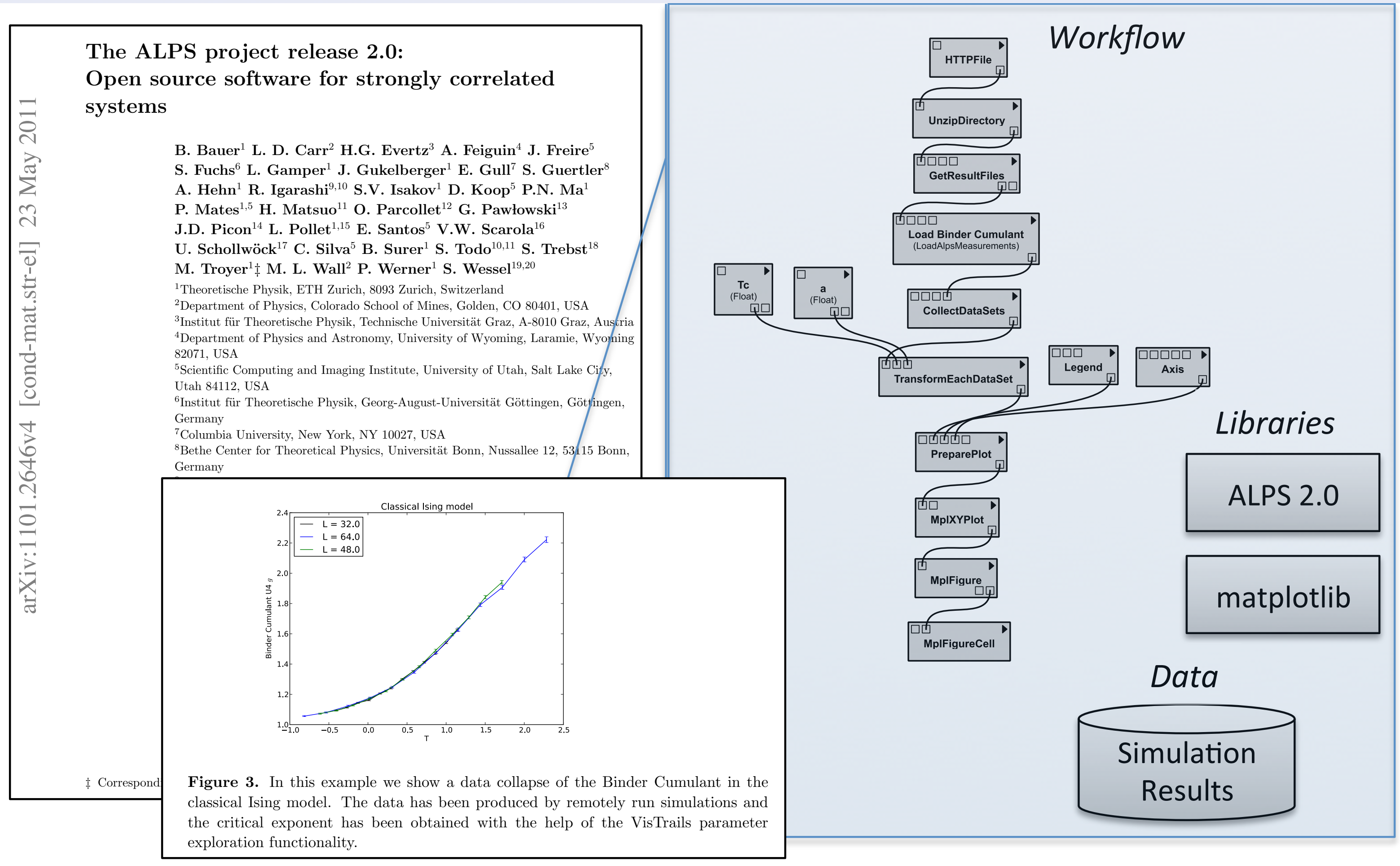
‡ Correspond

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]

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

A *Reproducible* Paper: ALPS2.0



arXiv:1101.2646v4 [cond-mat.str-el] 23 May 2011

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

B. Bauer[1] L. D. Carr[2] H.G. Evertz[3] A. Feiguin[4] J. Freire[5] S. Fuchs[6] L. Gamper[1] J. Gukelberger[1] E. Gull[7] S. Guertler[8] A. Hehn[1] R. Igarashi[9,10] S.V. Isakov[1] D. Koop[5] P.N. Ma[1] P. Mates[1,5] H. Matsuo[11] O. Parcollet[12] G. Pawłowski[13] J.D. Picon[14] L. Pollet[1,15] E. Santos[5] V.W. Scarola[16] U. Schollwöck[17] C. Silva[5] B. Surer[1] S. Todo[10,11] S. Trebst[18] M. Troyer[1]‡ M. L. Wall[2] P. Werner[1] S. Wessel[19,20]

[1]Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland
[2]Department of Physics, Colorado School of Mines, Golden, CO 80401, USA
[3]Institut für Theoretische Physik, Technische Universität Graz, A-8010 Graz, Austria
[4]Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA
[5]Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA
[6]Institut für Theoretische Physik, Georg-August-Universität Göttingen, Göttingen, Germany
[7]Columbia University, New York, NY 10027, USA
[8]Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany

‡ Correspond

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]

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

Some Videos

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

Reproducible Papers

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

The ALPS 2.0 paper
http://adsabs.harvard.edu/abs/2011arXiv1101.2646B

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*
- *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
- 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*
- *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 different levels of granularity facilitates the understanding of the computations and results

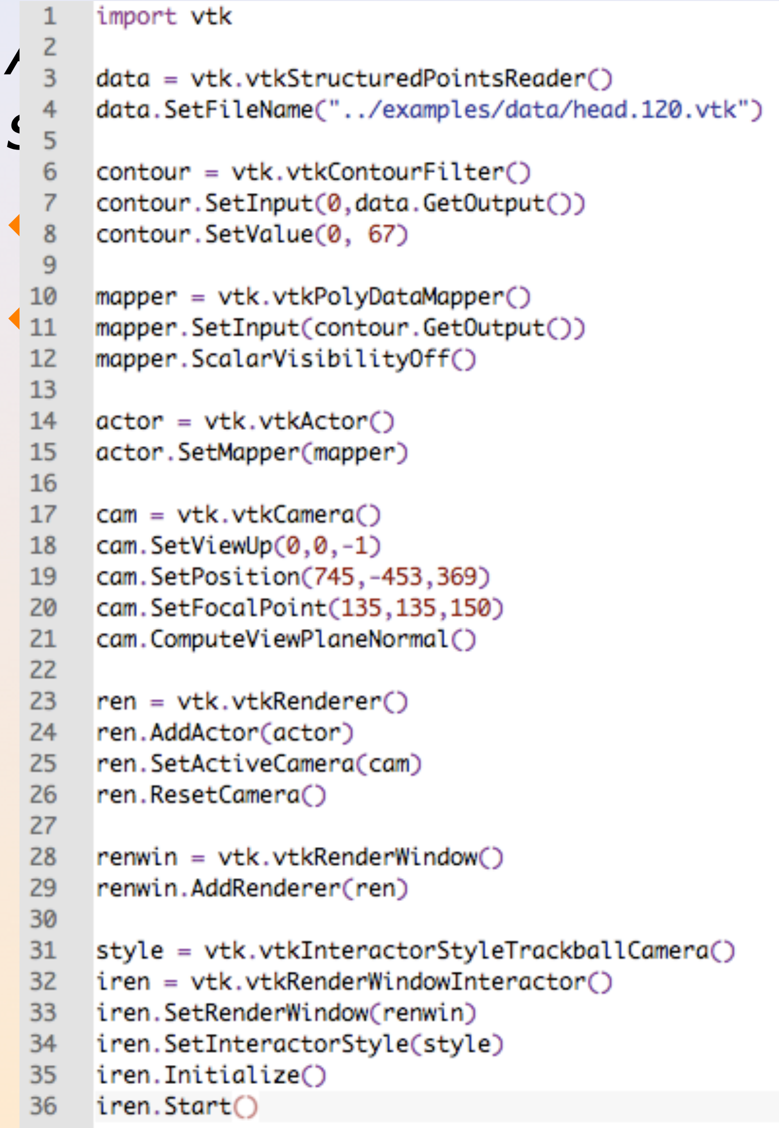
[Koop et al., ICCS 2011]

Writing & Development

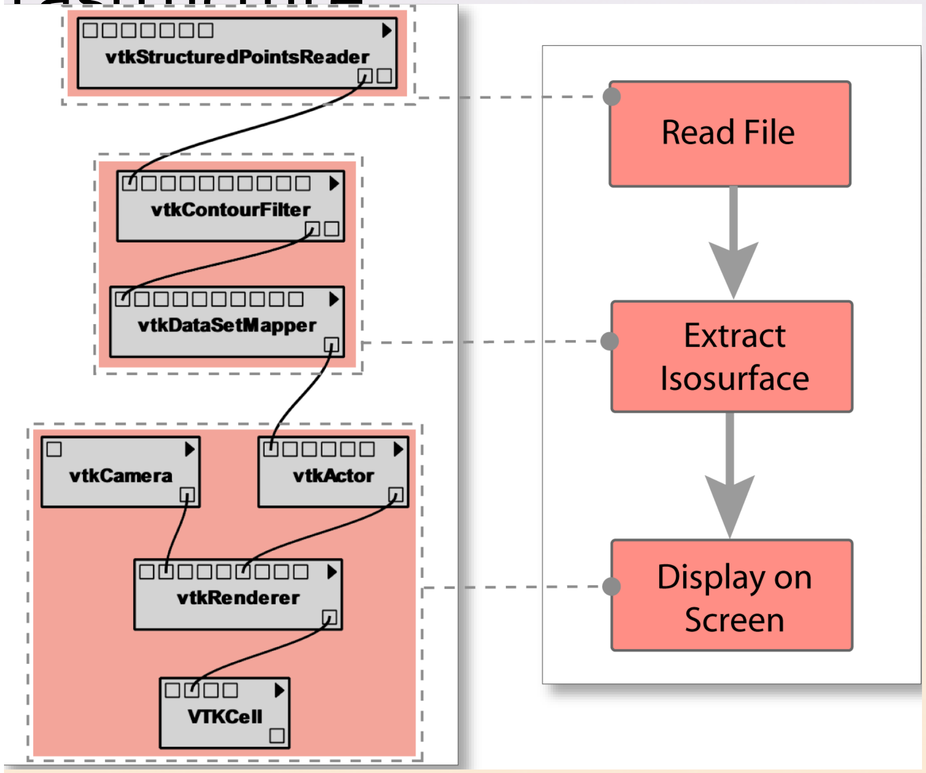
working in an environment that

f an executable paper

frastructure



```
import vtk

data = vtk.vtkStructuredPointsReader()
data.SetFileName("../examples/data/head.120.vtk")

contour = vtk.vtkContourFilter()
contour.SetInput(0,data.GetOutput())
contour.SetValue(0, 67)

mapper = vtk.vtkPolyDataMapper()
mapper.SetInput(contour.GetOutput())
mapper.ScalarVisibilityOff()

actor = vtk.vtkActor()
actor.SetMapper(mapper)

cam = vtk.vtkCamera()
cam.SetViewUp(0,0,-1)
cam.SetPosition(745,-453,369)
cam.SetFocalPoint(135,135,150)
cam.ComputeViewPlaneNormal()

ren = vtk.vtkRenderer()
ren.AddActor(actor)
ren.SetActiveCamera(cam)
ren.ResetCamera()

renwin = vtk.vtkRenderWindow()
renwin.AddRenderer(ren)

style = vtk.vtkInteractorStyleTrackballCamera()
iren = vtk.vtkRenderWindowInteractor()
iren.SetRenderWindow(renwin)
iren.SetInteractorStyle(style)
iren.Initialize()
iren.Start()
```



[Koop et al., ICCS 2011]

Writing & Development

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

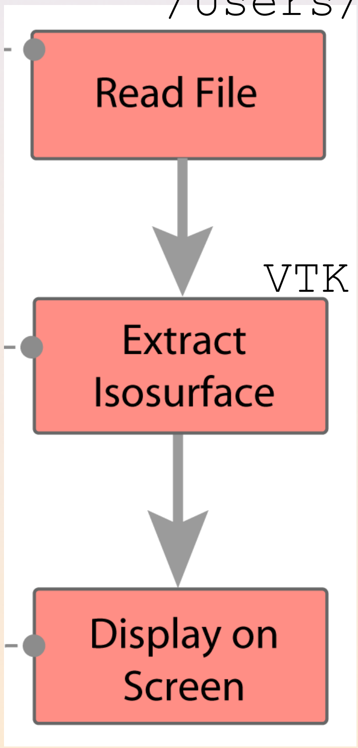
- Provenance of data and computations: workflow provenance is not sufficient

[Koop et al., ICCS 2011]

Sharing an Experiment

- Juliana creates an experiment

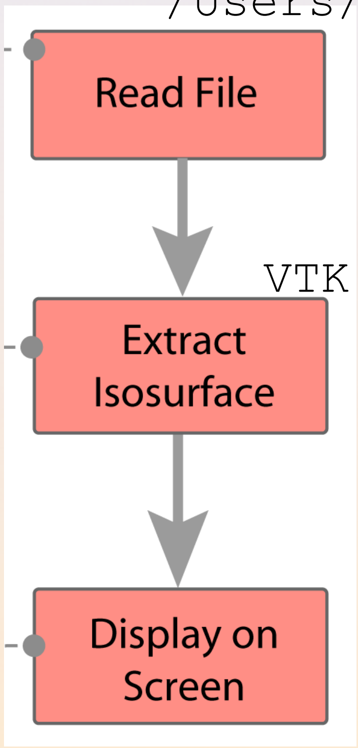
`/Users/juliana/head.vtk`



`VTK 1.2`

- Ian tries to run Juliana’s experiment

`/Users/juliana/head.vtk`



File not found!

`VTK 1.2`

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

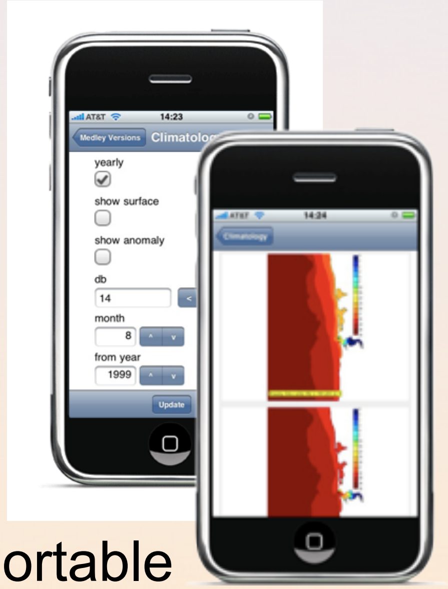
Improve the quality of reviews: reviewers have the ability to explore and validate conclusions

- 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]

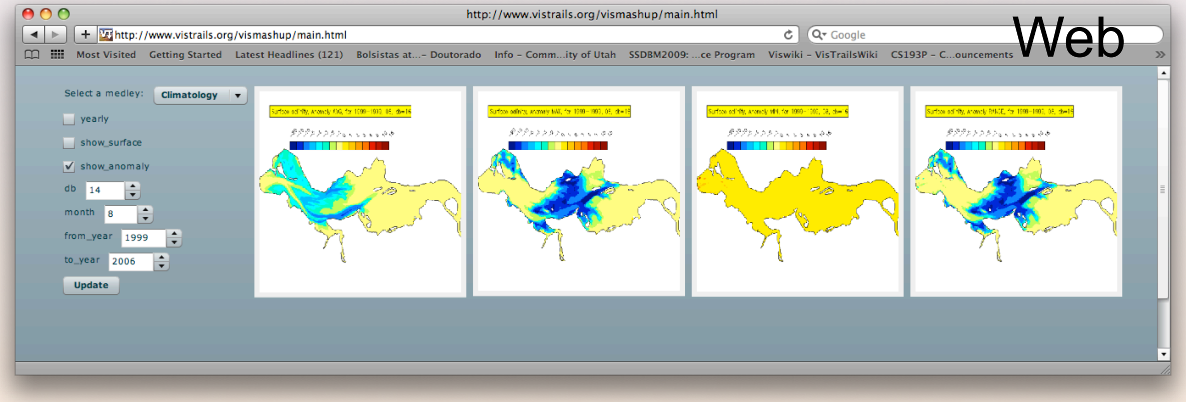
Publishing, Maintenance, & Re-Use

- Simplify interaction: the VisMashup system [Santos@TVCG2009]
- Publish using different media, not just *documents*



yearly ☑
show surface □
show anomaly □
db 14
month 8
from year 1999
Update

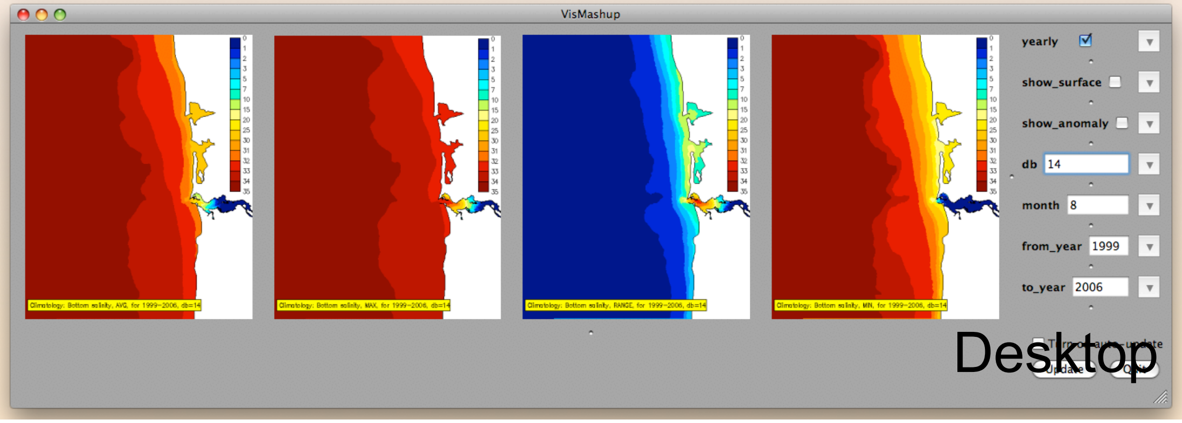
Portable Devices



http://www.vistrails.org/vismashup/main.html

Select a medley: Climatology
□ yearly
□ show_surface
☑ show_anomaly
db 14
month 8
from_year 1999
to_year 2006
Update

Web



VisMashup

yearly ☑
show_surface □
show_anomaly □
db 14
month 8
from_year 1999
to_year 2006

Desktop

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

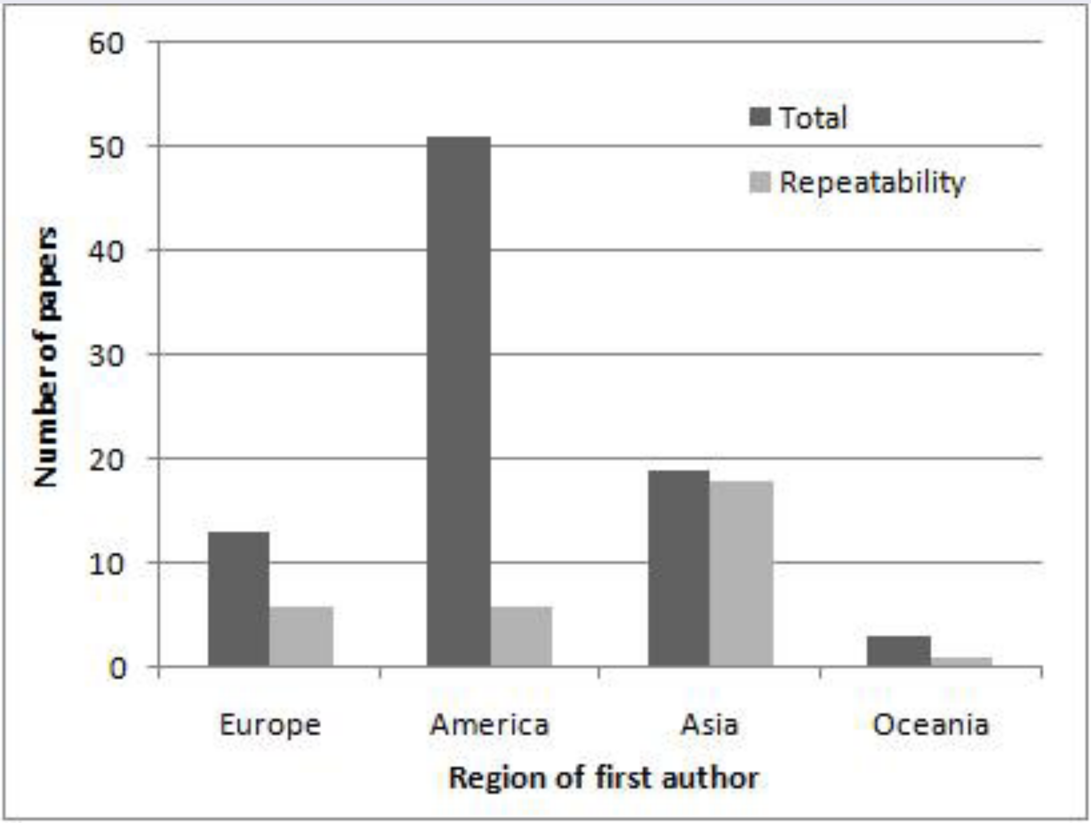
Current Uses and Experiences

- 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 in accepted paper
  - Paper n: 2010-20 subm
  - In 20 and expedite the re
  - Revie
    - » Co lependencies
    - » Ea
  - Reas
    - » Int
    - » Sensitive data
    - » Specific hardware requirements

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

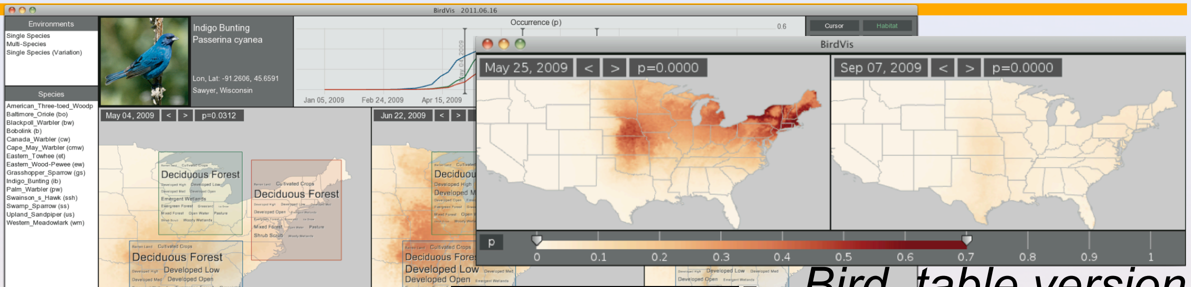
Going Forward

- 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 standard**s** and guidelines for authors and tool developers
- Need provenance support in applications
  - Integrate provenance from different sources, connect the results

Provenance Everywhere



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

predictions

results

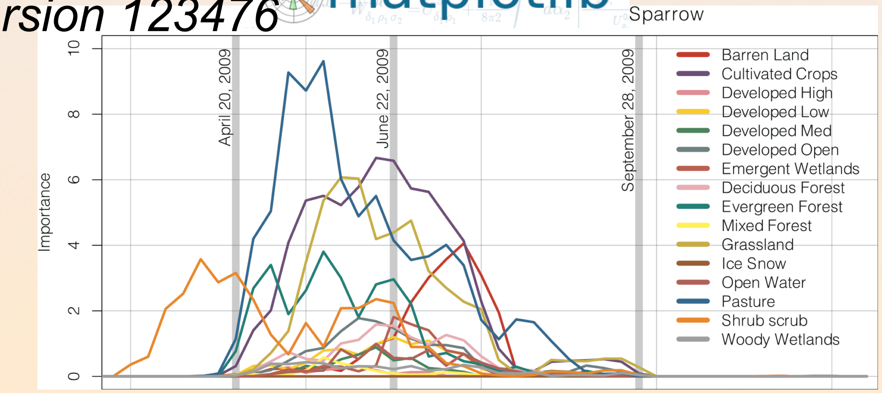
eBird

STEM

predictions

- *matplotlib script+results*

Bird_table version 123476



Provenance-rich presentation

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
- An infrastructure to support the creation, review and re-use of reproducible papers 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.

CMOP
Center for Coastal
Margin Observation
& Prediction

VACET
DOE SciDAC VISUALIZATION AND ANALYTICS
CENTER FOR ENABLING TECHNOLOGIES

THE UNIVERSITY OF UTAH

National Science Foundation
WHERE DISCOVERIES BEGIN

SDM CENTER

IBM

Merci

Ευχαριστω

Thank you

Obrigada