

# Supporting Provenance-Rich Science with VisTrails

Juliana Freire

Web and Databases Lab

Scientific Computing and Imaging Institute

School of Computing

University of Utah



# The Need for Provenance Management



United States Patent and Trademark Office  
An Agency of the Department of Commerce

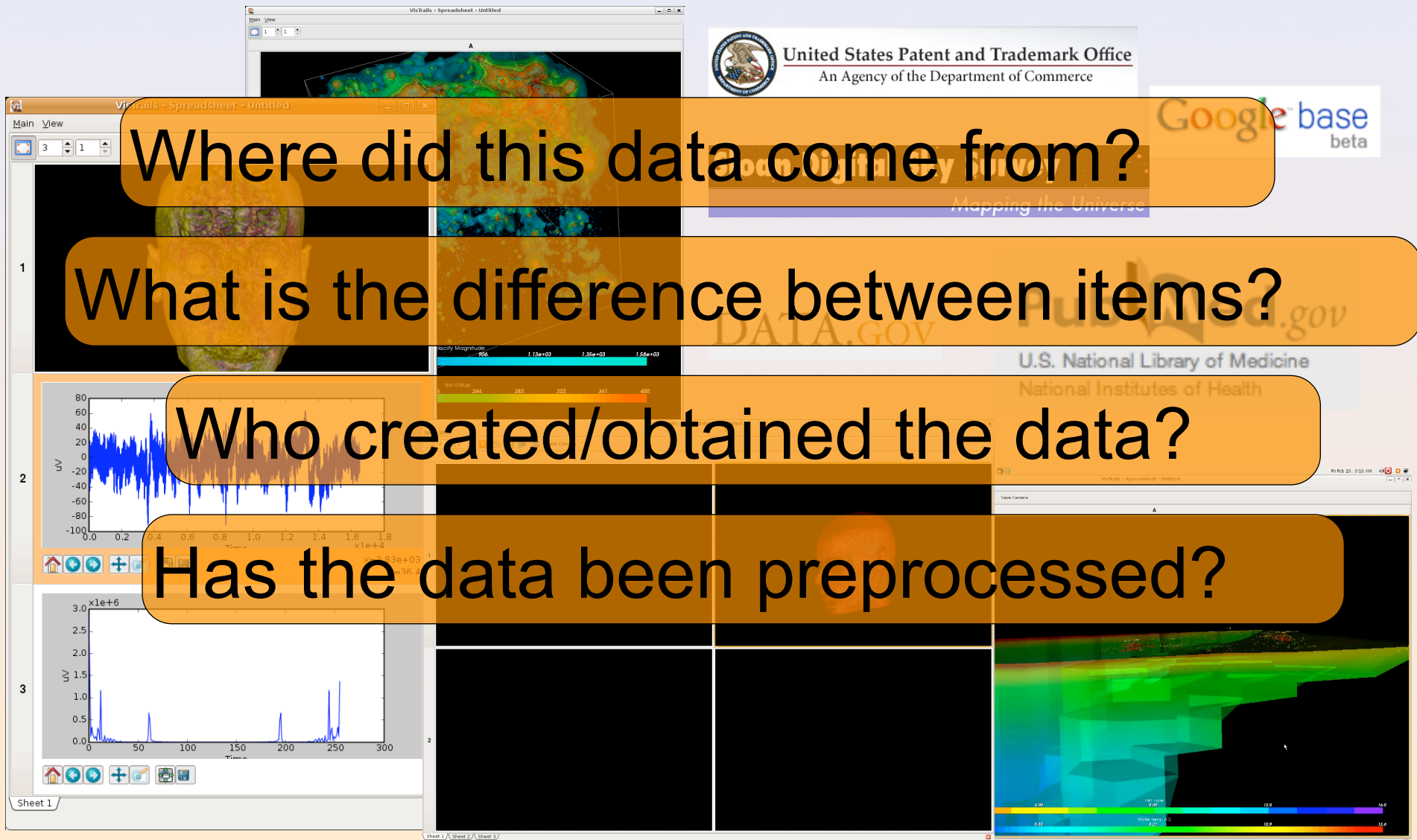


Where did this data come from?

What is the difference between items?

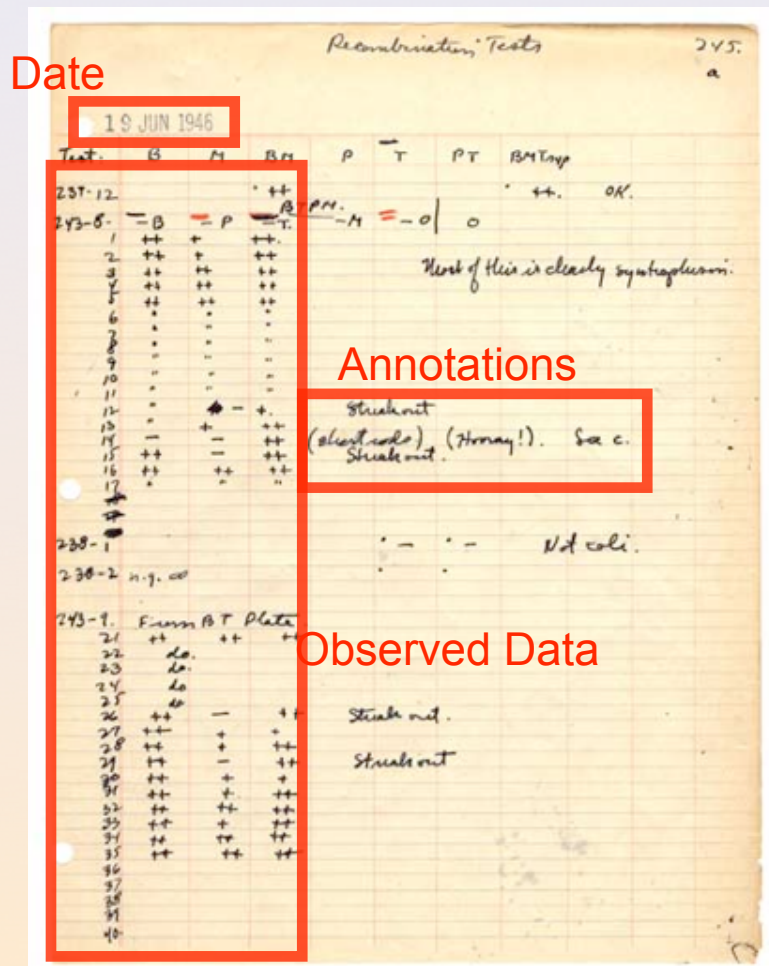
Who created/obtained the data?

Has the data been preprocessed?



# Provenance in Science

- ◆ Provenance is as (or more) important as the result!
- ◆ Old solution:
  - Lab notebooks
- ◆ New problems:
  - Large volumes of data
  - Complex analyses
  - Writing notes doesn't scale
- ◆ New solution:
  - Automated provenance capture with user-defined annotations



# Provenance in Computational Science

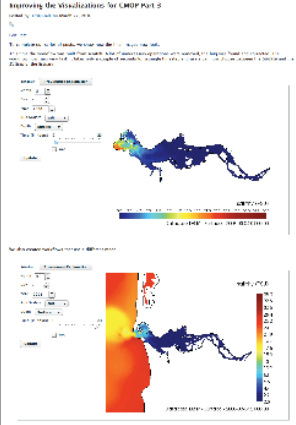


Fig. 7: Using the blog to document processes: A visualization expert created a series of blog posts to explain the problems found when generating the visualizations for CMOP.

**ACKNOWLEDGMENTS**

Our research has been funded by the National Science Foundation (grants IIS-0905385, IIS-0746550, ATM-0825821, IIS-0844546, CNS-0751152, IIS-0713637, OCE-0424602, IIS-0534628, CNS-0514485, IIS-0513692, CNS-0524096, CCF-0401498, OISE-0405402, CCF-0528201, CNS-0551724), the Department of Energy SciDAC (VACET and SDM centers), and IBM Faculty Awards (2005, 2006, 2007, and 2008). E. Santos is partially supported by a CAPES/Fulbright fellowship.

**REFERENCES**

- [1] L. Bavoi, S. Callahan, P. Crosato, J. Freire, C. Scheidegger, C. Silva, and H. Vo. ViTrails: Enabling Interactive Multiple-View Visualizations. In *IEEE Visualization 2005*, pages 135-142, 2005.
- [2] S. P. Callahan, J. Freire, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Towards provenance-enabling panview, pages 120-127, 2008.
- [3] Chemical Space. <http://vtrails.com/chemspace.html>
- [4] NSF Center for Coastal Margin Observation and Prediction (CMOP). <http://www.stcmop.org>
- [5] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of SIGMOD*, pages 1345-1350, 2008.
- [6] R. T. Fickling. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2003.
- [7] S. Foual and J. Claretout. Guest editors' introduction: Reproducible research. *Computing in Science Engineering*, 11(1):5-7, jan-feb 2009.

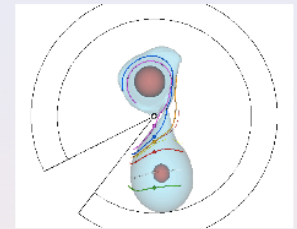
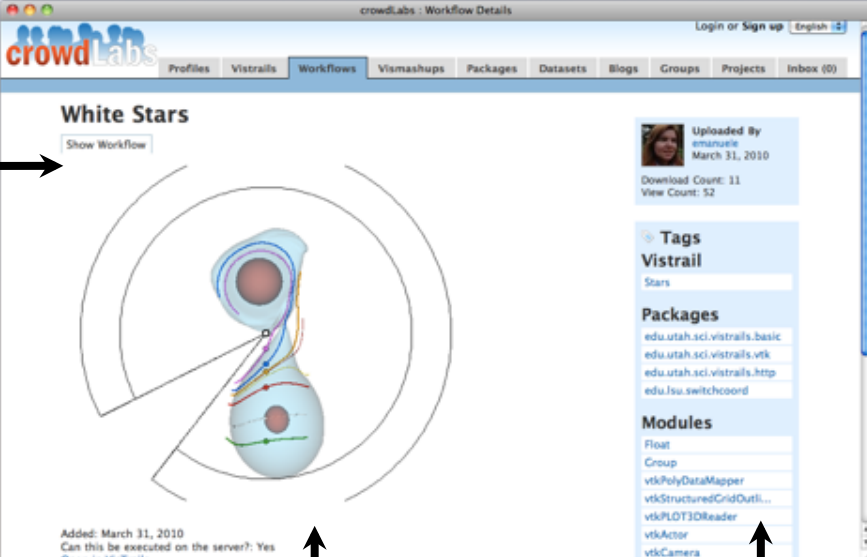


Fig. 8: Visualizing a binary star system simulation. This is an image that was generated by embedding a workflow directly in the text. The original workflow is available at <http://www.crowdlabs.org/vistrails/workflow/details/119/>.

- [8] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11-21, May-June 2008.
- [9] J. Freire, C. Silva, S. Callahan, E. Santos, C. Scheidegger, and H. Vo. Managing rapidly-evolving scientific workflows. In *International Provenance and Annotation Workshop (IPAW)*, LNCS 4145, pages 10-18. Springer Verlag, 2006.
- [10] R. Hoffmann. A wiki for the life sciences where authorship matters. *Nature Genetics*, 40(9):1047-1051, 2008.
- [11] IBM. OpenDX. <http://www.research.ibm.com/dx/>
- [12] Kivware. Paraview. <http://www.paraview.org>
- [13] Kivware. The visualization toolkit. <http://www.vtk.org>
- [14] Many Eyes Wikified. <http://wikified.research.ibm.com>
- [15] M. McKern. Harnessing the Web Information Ecosystem with Wiki-based Visualization Dashboards. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1081-1088, 2009.
- [16] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conkles, and C. Ewles. WikiPathways: Pathway editing for the people. *PLoS Biology*, 6(7), 2008.
- [17] D. D. Roure, C. Goble, and R. Stevens. The design and realization of the virtual research environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5):561-567, 2009.
- [18] E. Santos, L. Lars, J. Ahrens, J. Freire, and C. Silva. Vismashup: Streamlining the creation of custom visualization applications. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1539-1546, 2009.
- [19] Swireel. <http://www.swireel.com>
- [20] J. Tobline and E. Santos. Visualizing a Journal that Serves the Computational Sciences Community. *Computing in Science & Engineering*, 12(3), 2010. To appear.
- [21] J. E. Tobline. Scientific Visualization: A Necessary Choice. *Computing in Science & Engineering*, 9(6):76-81, 2007.
- [22] C. Upoon, J. Thomas Faulhaber, D. Karmis, D. H. Laidlaw, D. Schlegel, J. Vroom, K. Giaratz, and A. van Dam. The Application Visualization System: A Computational Environment for Scientific Visualization. *IEEE Computer Graphics and Applications*, 9(4):30-42, 1989.
- [23] F. B. Viega, M. Waterberg, F. van Ham, J. Kivim, and M. McKern. ManyEyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121-1128, 2007.
- [24] Vistrails Visualization Tool. <https://www.fnl.gov.br/odes/vist/>
- [25] The ViTrails Project. <http://www.vistrails.org>



crowdlabs : Workflow Details

crowdlabs Profiles Vistrails Workflows Vismashups Packages Datasets Blogs Groups Projects Inbox (0)

White Stars

Show Workflow

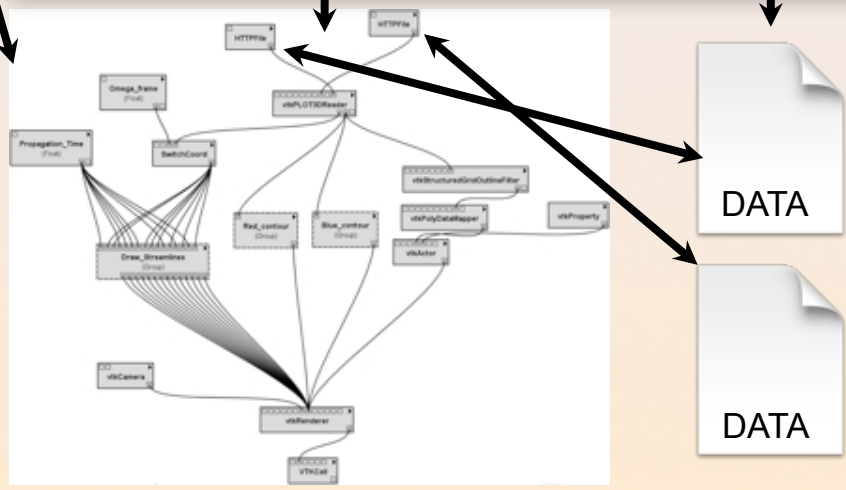
Added: March 31, 2010  
Can this be executed on the server?: Yes  
[View in ViTrails](#)

Uploaded By: emmanuel March 31, 2010  
Download Count: 11  
View Count: 52

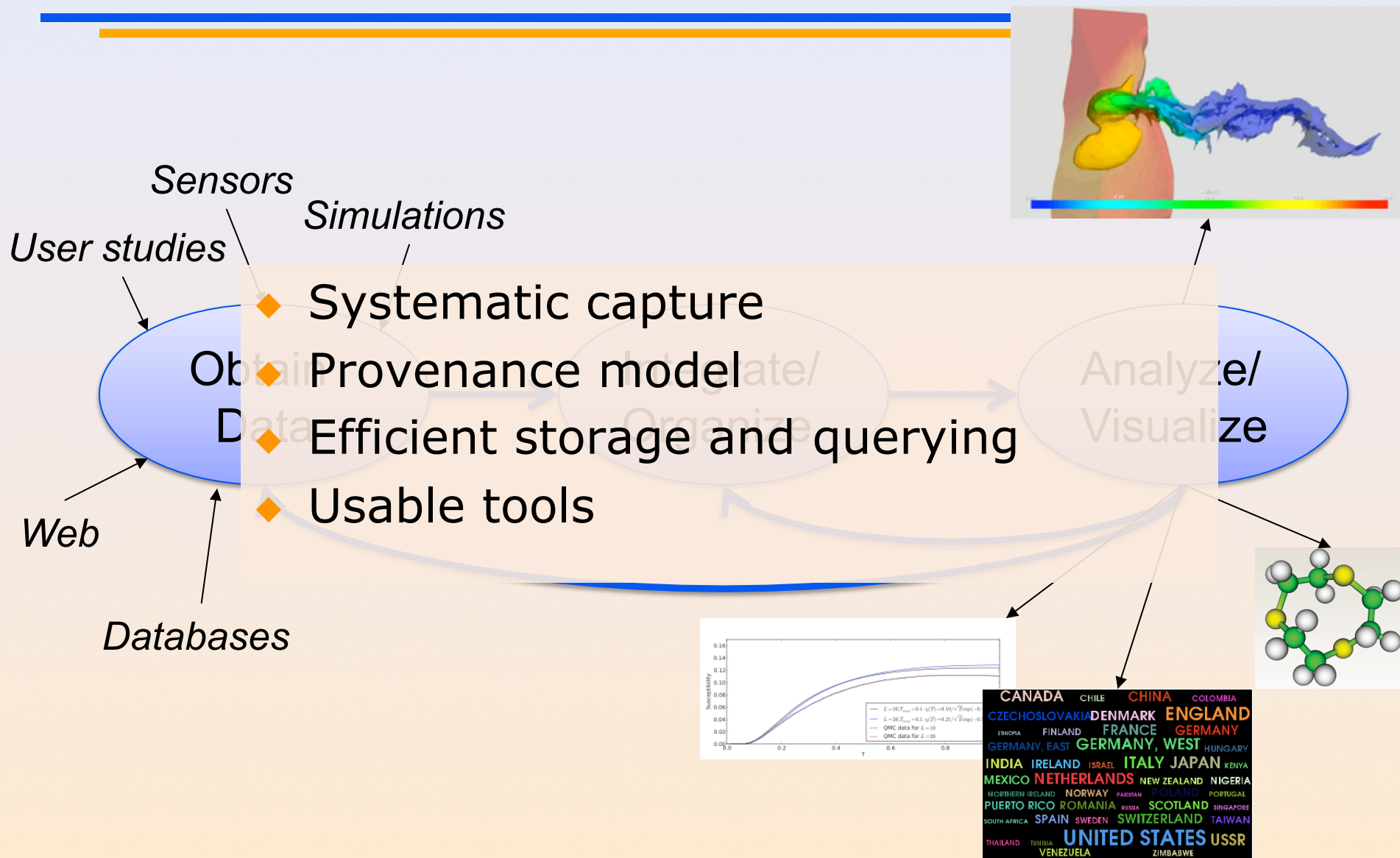
Tags: Vistrail, Stars

Packages: edu.utah.sci.vistrails.basic, edu.utah.sci.vistrails.vtk, edu.utah.sci.vistrails.http, edu.lsu.switchcoord

Modules: Float, Group, vtkPolyDataMapper, vtkStructuredGridOutli..., vtkPLOT3DReader, vtkActor, vtkCamera



# Provenance Management: Desiderata



# The VisTrails System

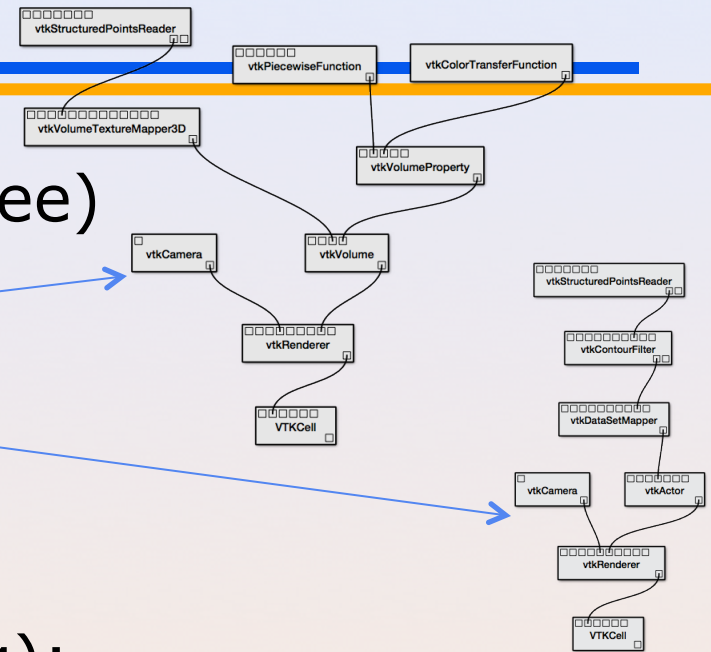
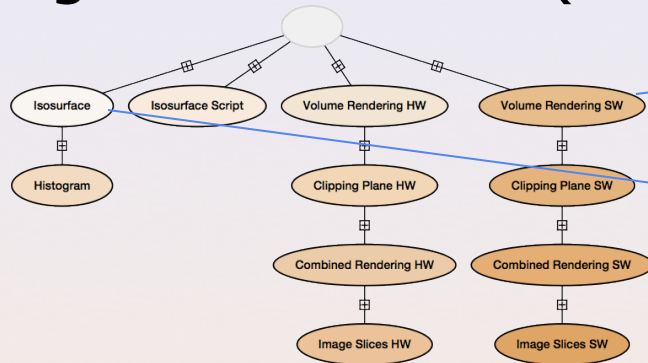


- ◆ Comprehensive *provenance infrastructure* for computational tasks
- ◆ Focus on exploratory tasks such as simulation, visualization, and data analysis
  - Workflows + Visualization
- ◆ *Transparently* tracks provenance of the discovery process---from data acquisition to visualization
  - The *trail* followed as users generate and test hypotheses
- ◆ *Leverage provenance to streamline exploration*
  - Query and mine provenance
- ◆ Focus on usability—build tools for
- ◆ Featured as an NSF discovery



# Provenance in VisTrails

## ◆ Design Provenance (Version Tree)



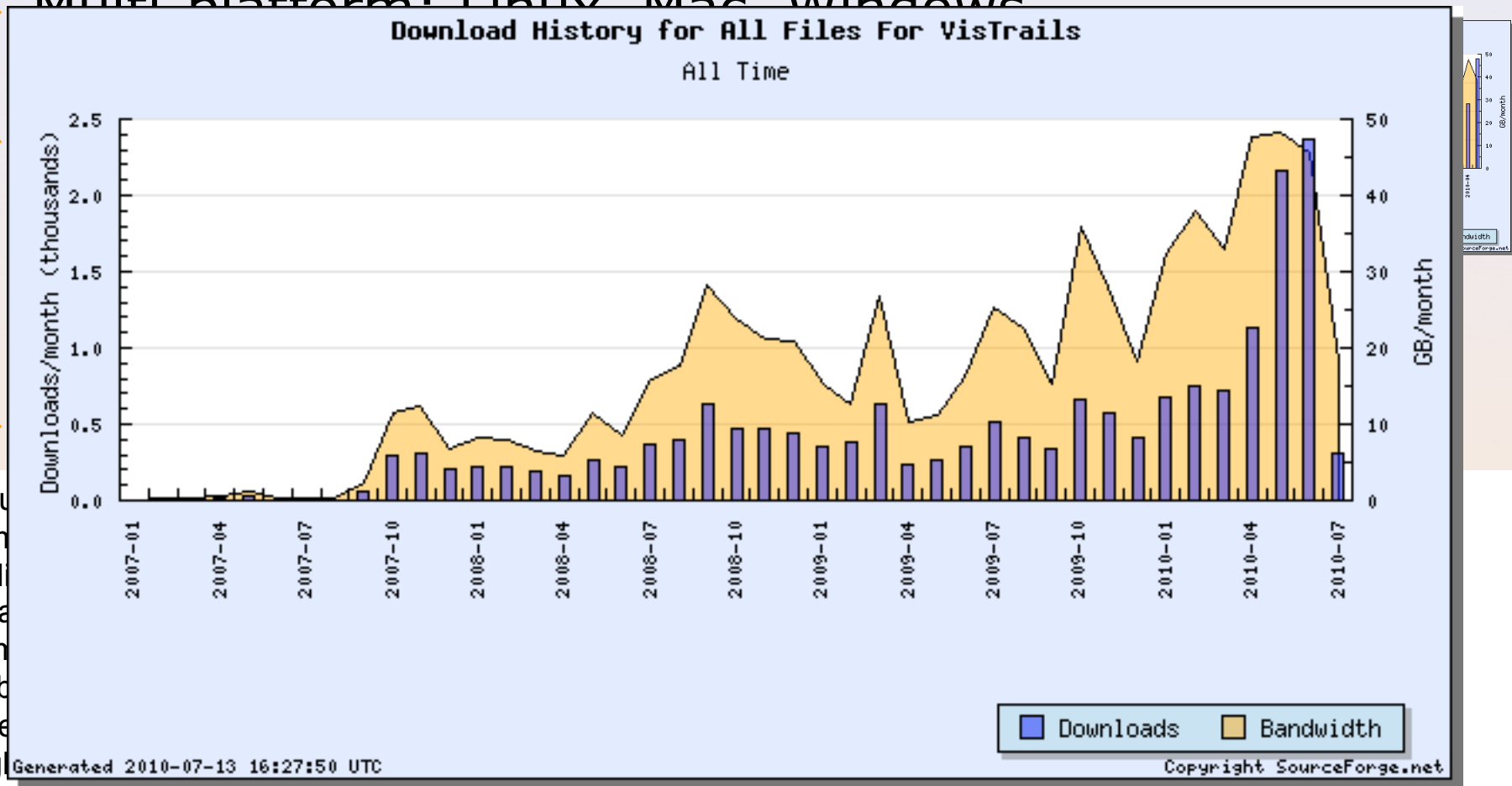
## ◆ Execution Provenance (Logging):

```
<module id="12" name="vtkDataSetReader" start_time="2010-02-19 11:01:05"
  end_time="2010-02-19 11:01:07"> <annotation key="hash"
value="c54bea63cb7d912a43ce"/></module><module id="13"
name="vtkContourFilter" start_time="2010-02-19 11:01:07"
  end_time="2010-02-19 11:01:08"/><module id="15"
name="vtkDataSetMapper" start_time="2010-02-19 11:01:09"
  end_time="2010-02-19 11:01:12"/>
```



# The VisTrails System

- ◆ VisTrails is open source: <http://www.vistrails.org>
- ◆ Multi platform: Linux, Mac, Windows



- ◆ VisTrails
- ◆ Simulation (Galaxy, Climate, etc.)
- ◆ Quality assurance
- ◆ Clinical trials
- ◆ Habitat modeling
- ◆ Open source
- ◆ High performance computing
- ◆ Cosmology simulations (LANL)

- University of North Carolina, Chapel Hill
- UTEP



# Climate Data Analysis

FILE VARIABLES

Directory: /home/e...\_le\_data

Computer

cdtest10.xml

cdtest13.xml

File: /home/emanuele/src/cdat\_build/sample\_data/clt.nc

Variable: clt (120, 46, 72) [Total clou] Define

DEFINED VARIABLES

Plot: Isofill

T-time: 1982-1-1 0:0:0

1988-12-1 0:0:0

X-latitude: 90.0

90.0

Y-longitude: 180.0

175.0

VisTrails VCDAT

History Tree

Total Cloudiness 1979

Total Cloudiness 1980

Total Cloudiness 1981

Total Cloudiness 1982

open

\_call\_

quickplot

VisTrails Shell

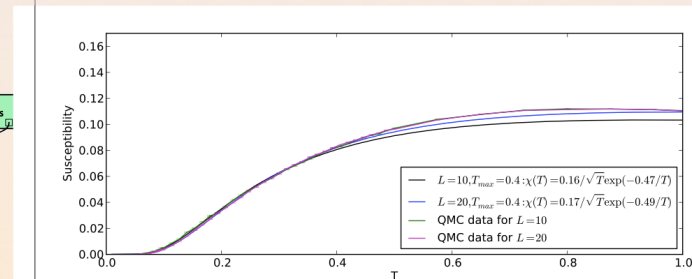
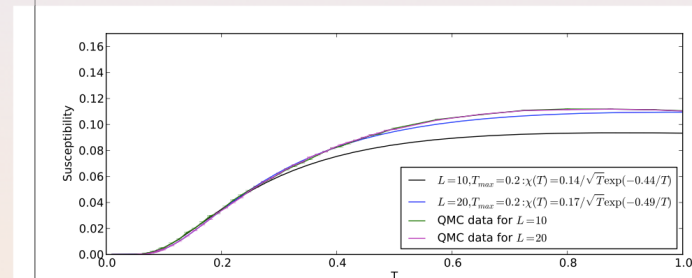
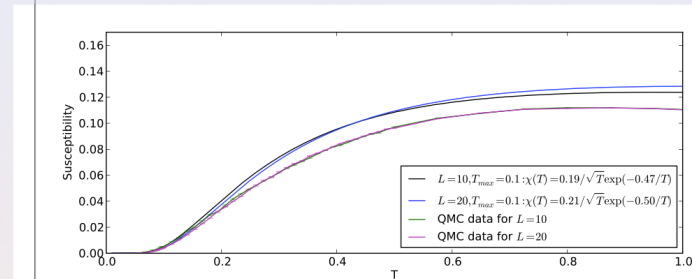
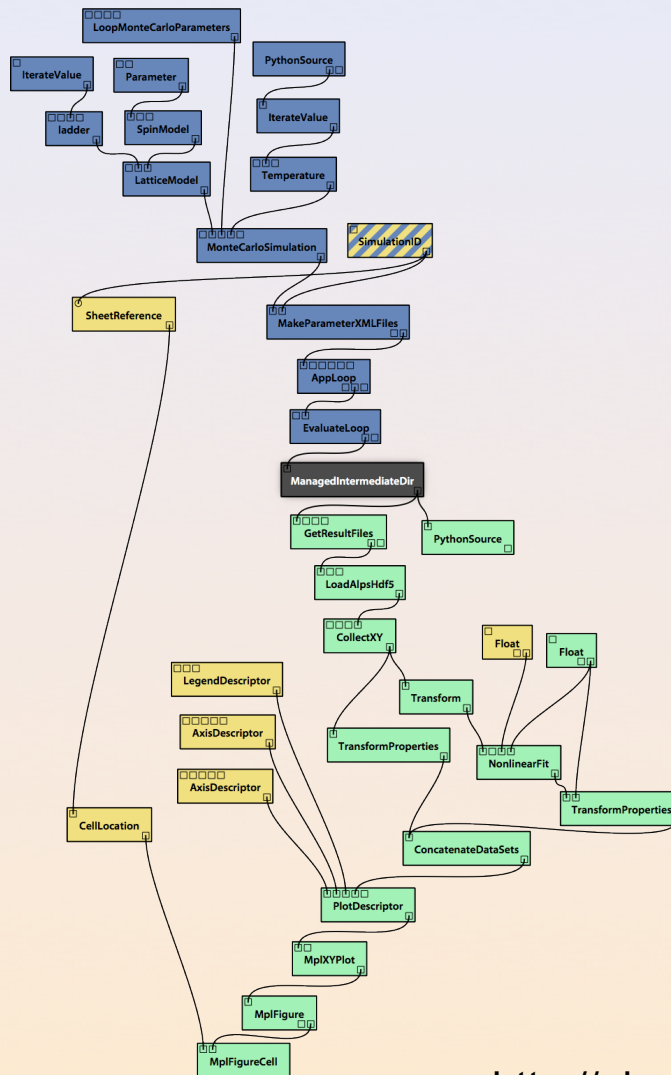
```
VisTrails shell running Python 2.6.4 (r264:75821M, Oct 27 2009, 19:48:32)
[GCC 4.0.1 (Apple Inc. build 5493)] on darwin.
Type "copyright", "credits" or "license" for more information on Python.
>>> import vcs, cdms2
>>> cdat = load_package('CDAT')
>>> cdmsfile = cdms2.open('/home/emanuele/src/cdat_bin/sample_data/clt.nc')
>>> data = cdmsfile('clt')
>>> q = cdat.quickplot()
>>> q.dataset = data
>>> run()
```

Visualization Spreadsheet

A B C D

[CDAT Project, Lawrence Livermore National Lab]

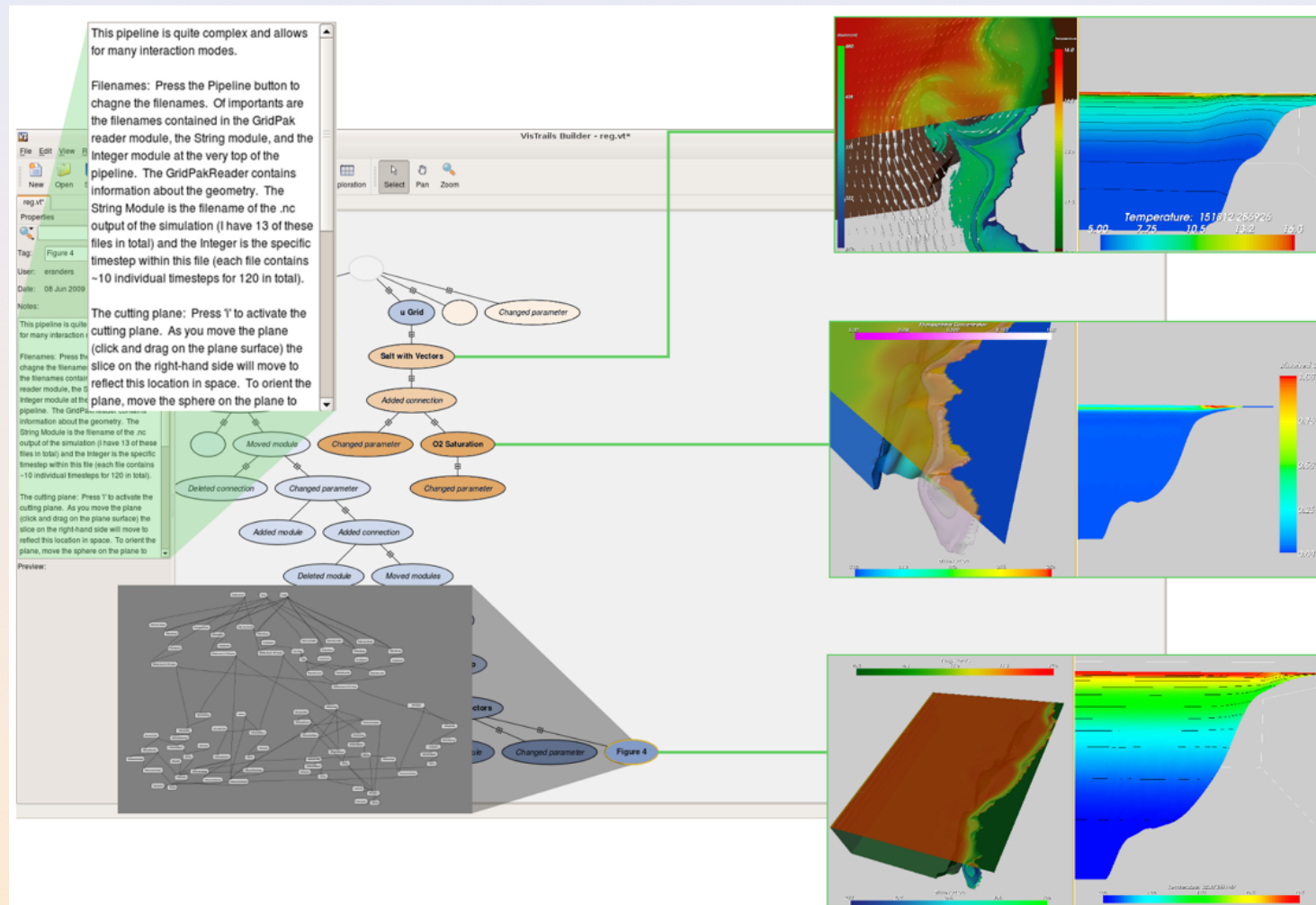
# Quantum Lattice Models



[ALPS Project, ETH-Zurich]

[http://alps.comp-phys.org/mediawiki/index.php/Main\\_Page](http://alps.comp-phys.org/mediawiki/index.php/Main_Page)

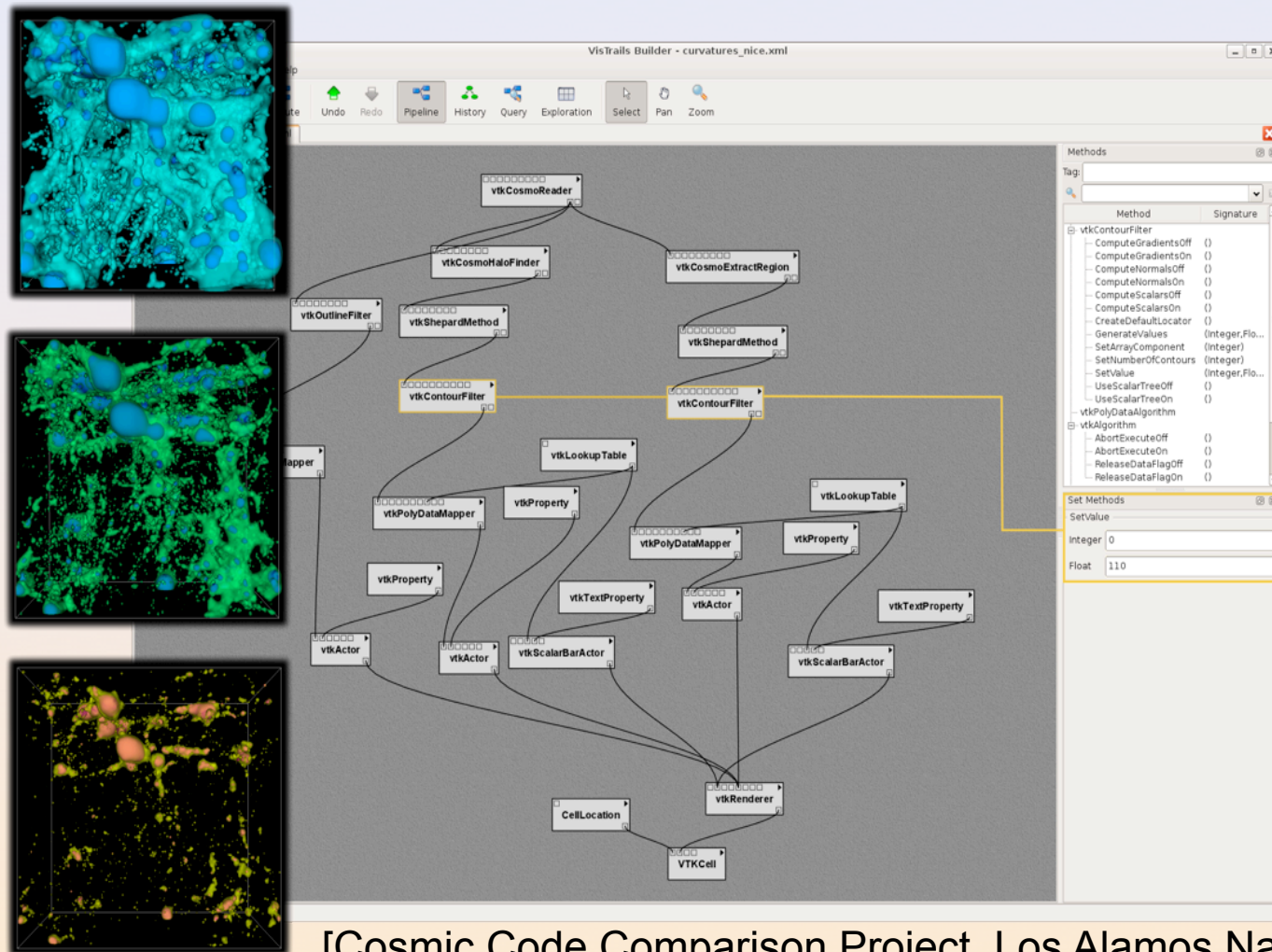
# Coastal Margin Observation & Prediction



[NSF Science & Technology Center for Coastal Margin Observation & Prediction]

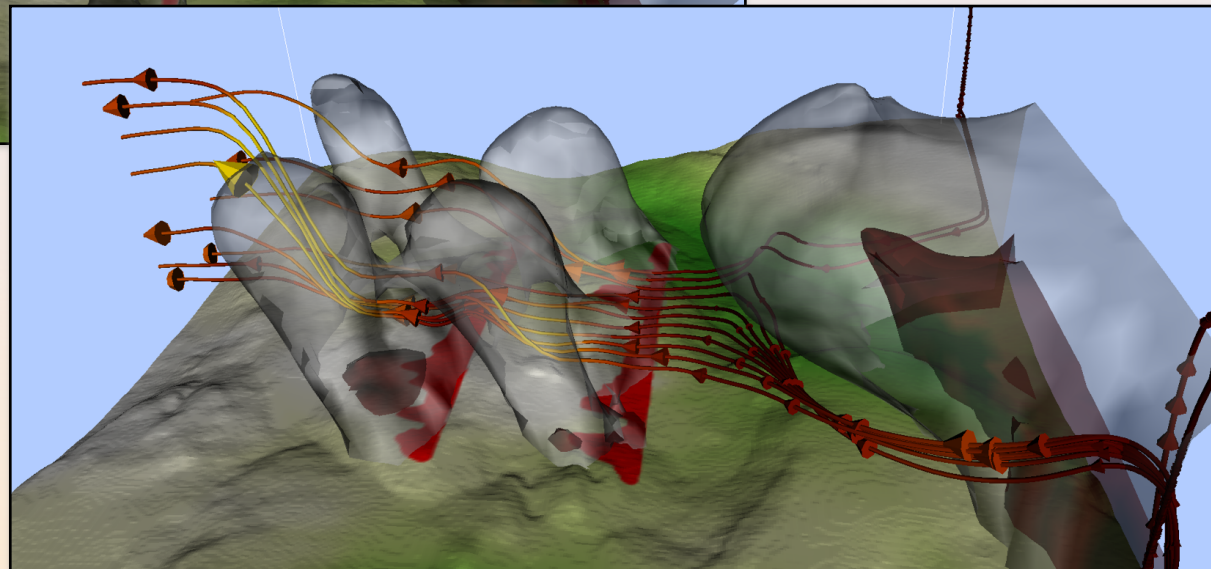
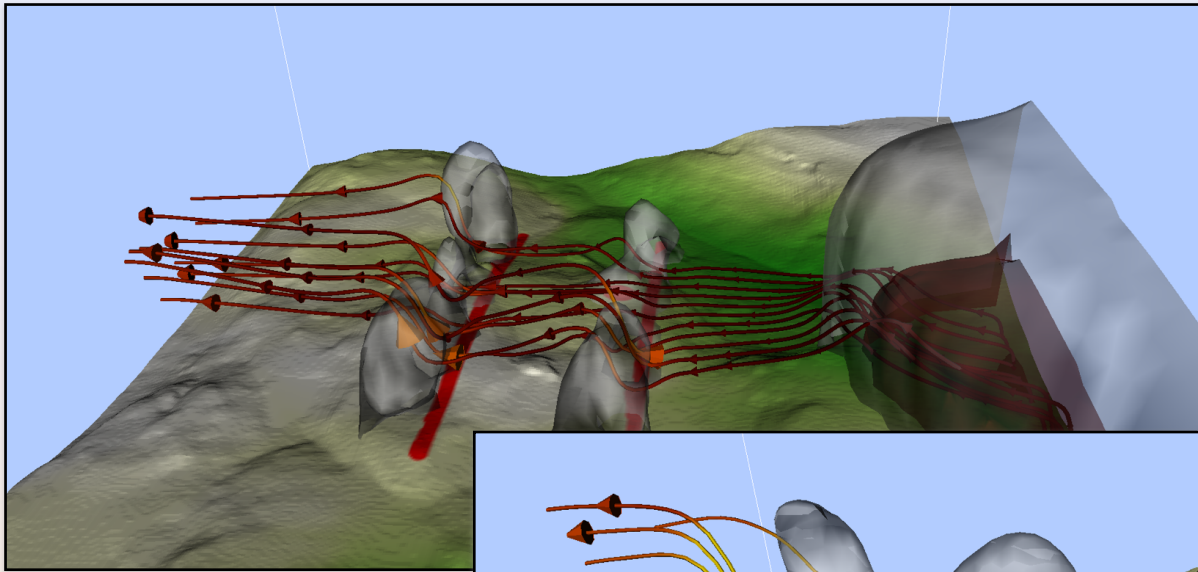
<http://www.stccmop.org/>

# Comparing Cosmological Simulations



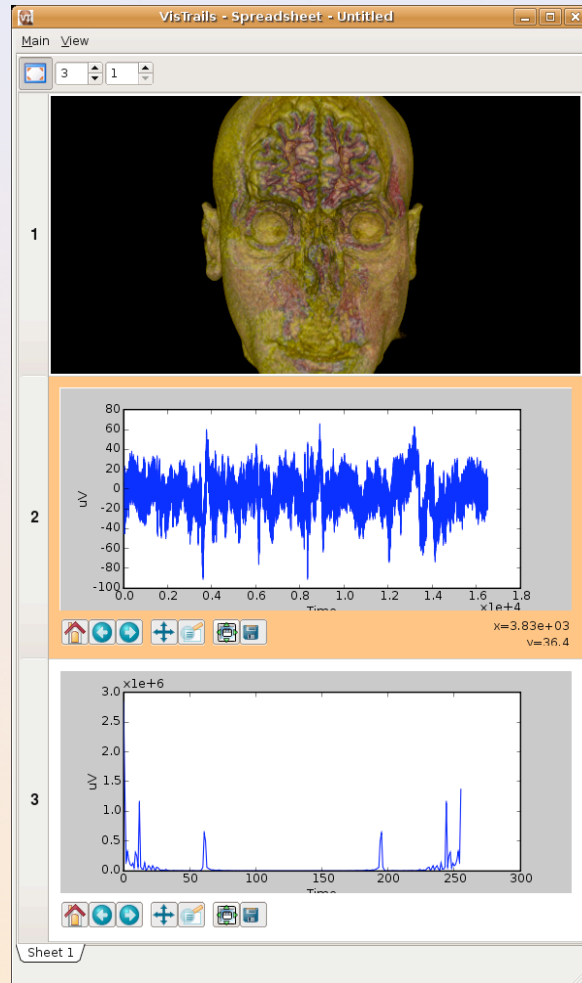
[Cosmic Code Comparison Project, Los Alamos National Lab]

# Wildfire Prediction



[NSF CDI WRF-Fire Project, University of Colorado-Denver]

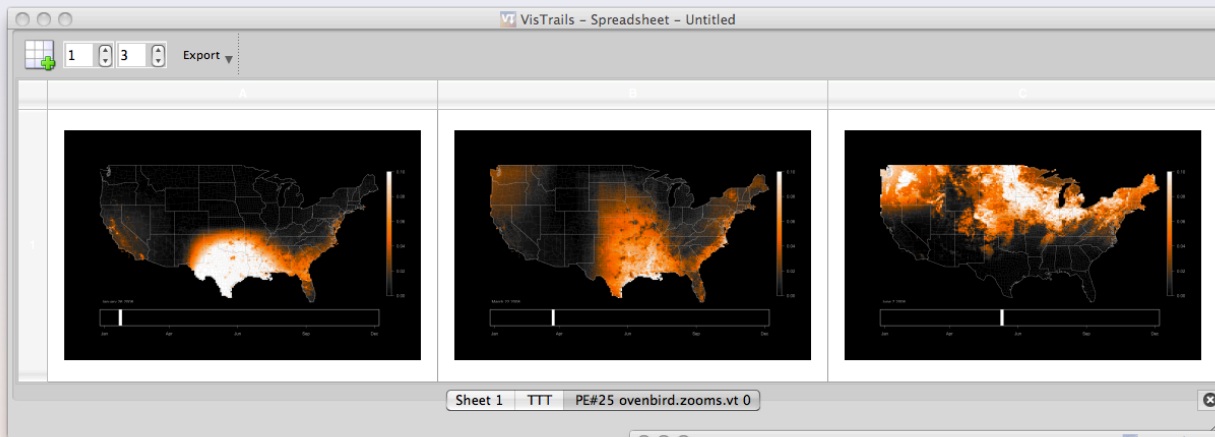
# Studying the Effects of TMS on Memory



[Psychiatry-U of Utah]

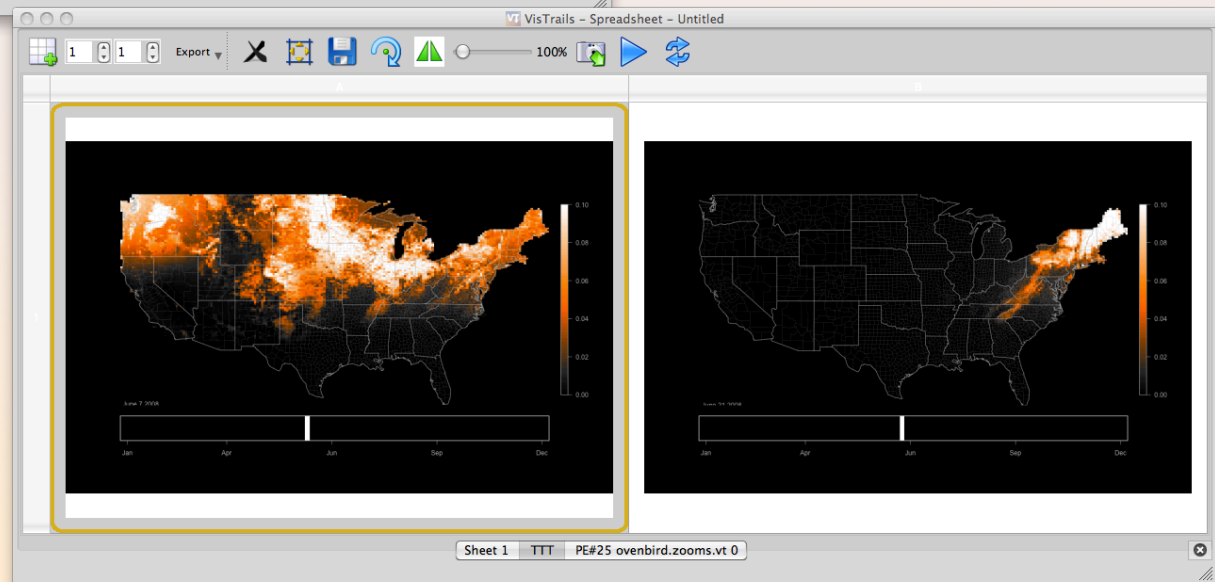
# Identifying large-scale threats to biodiversity

[EVA Group, NSF DataONE]



Breeding season distribution for Savannah Sparrow

Breeding distribution of the Savannah Sparrow on the left and the Black-throated Blue Warbler on the right



# Provenance and Data Exploration



# 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

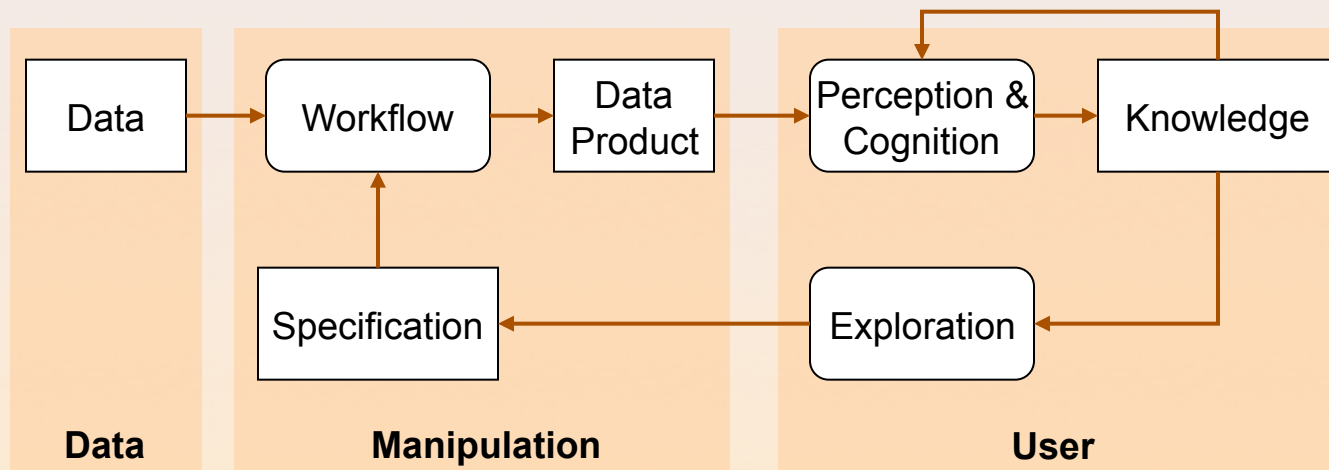


Figure modified from J. van Wijk, IEEE Vis 2005

# 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

# Change-Based Provenance

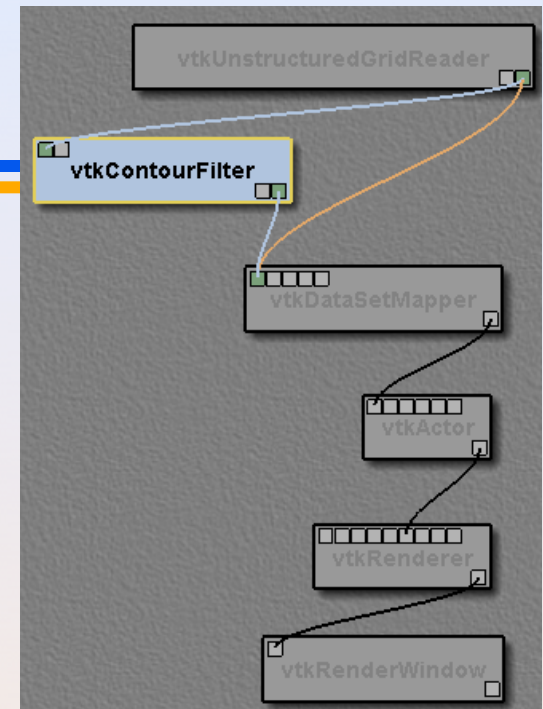
---

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

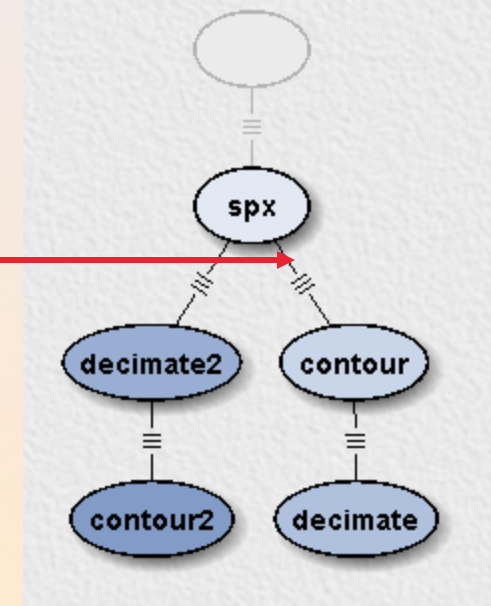
[Freire et al., IPAW 2006]

# 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



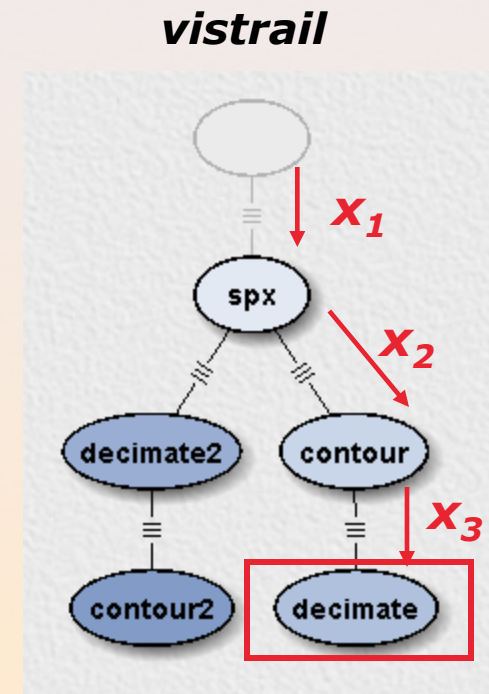
addModule  
deleteConnection  
addConnection  
addConnection  
setParameter



# Change-Based Provenance

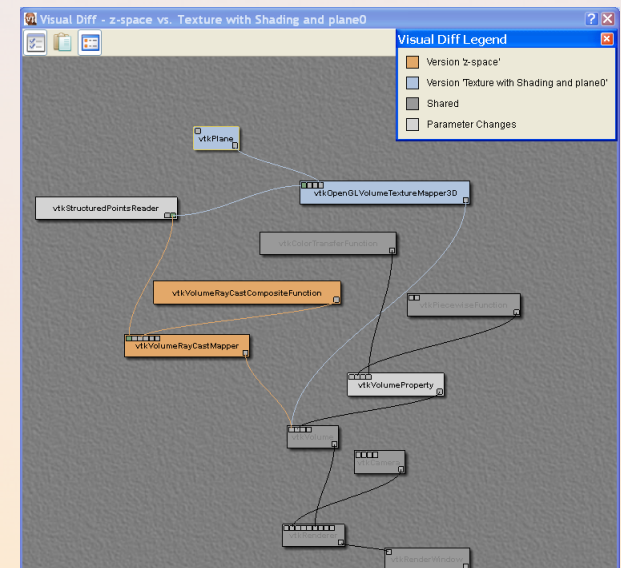
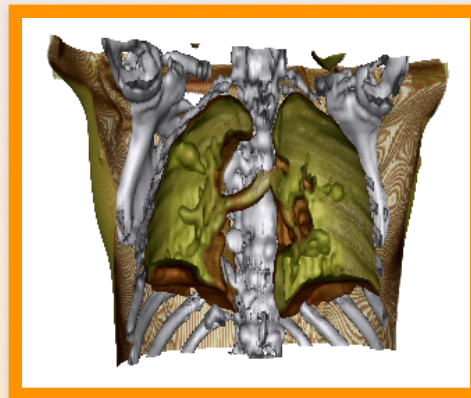
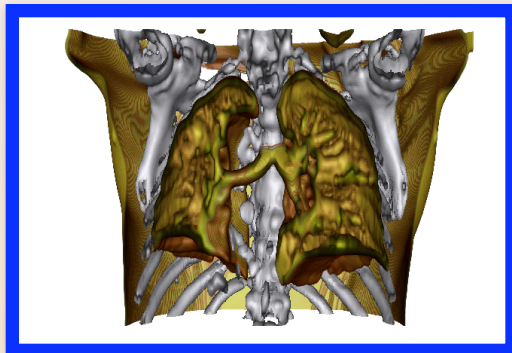
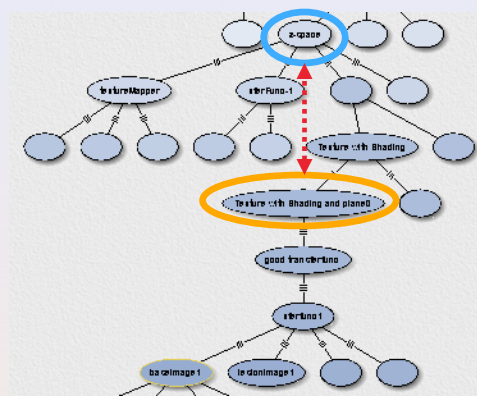
- ◆ Treat workflow as a first-class data item
- ◆ Provenance = changes to computational tasks
  - Add a module, add a connection, change a parameter value
- ◆ A *vistrail* node  $v_t$  corresponds to the workflow that is constructed by the sequence of actions from the root to  $v_t$ 
$$V_t = X_n \circ X_{n-1} \circ \dots \circ X_1 \circ \emptyset$$
- ◆ Extensible *change* algebra

[Freire et al, IPAW 2006]



# Provenance Beyond Reproducibility

- ◆ Support for reflective reasoning
- ◆ Ability to compare data results



[Freire et al., IPAW 2006]

# Computing Workflow Differences

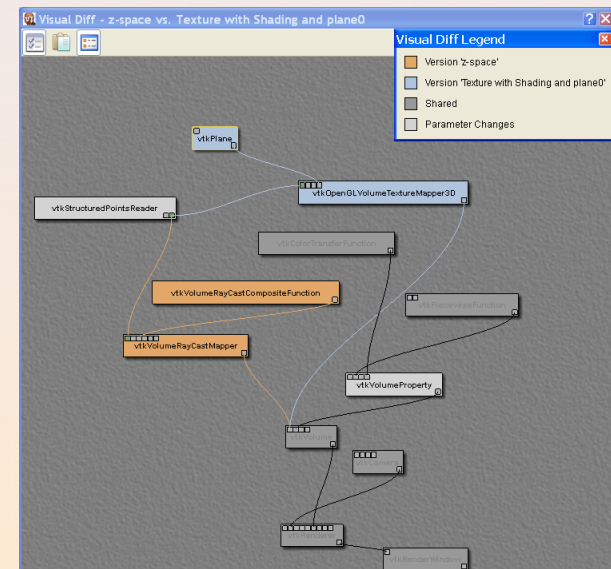
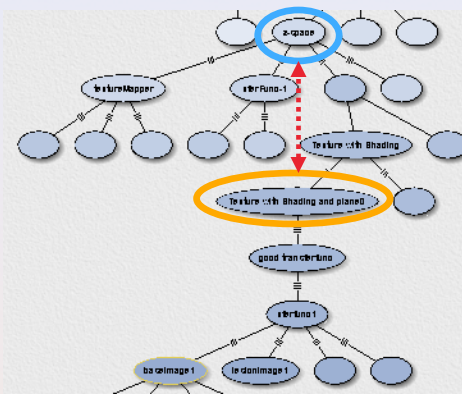
- ◆ No need to compute subgraph isomorphism!
- ◆ A vistrail is a rooted tree: all nodes have a common ancestor—diffs are well-defined and *simple to compute*

$$vt_1 = x_i \circ x_{i-1} \circ \dots \circ x_1 \circ \emptyset$$

$$vt_2 = x_j \circ x_{j-1} \circ \dots \circ x_1 \circ \emptyset$$

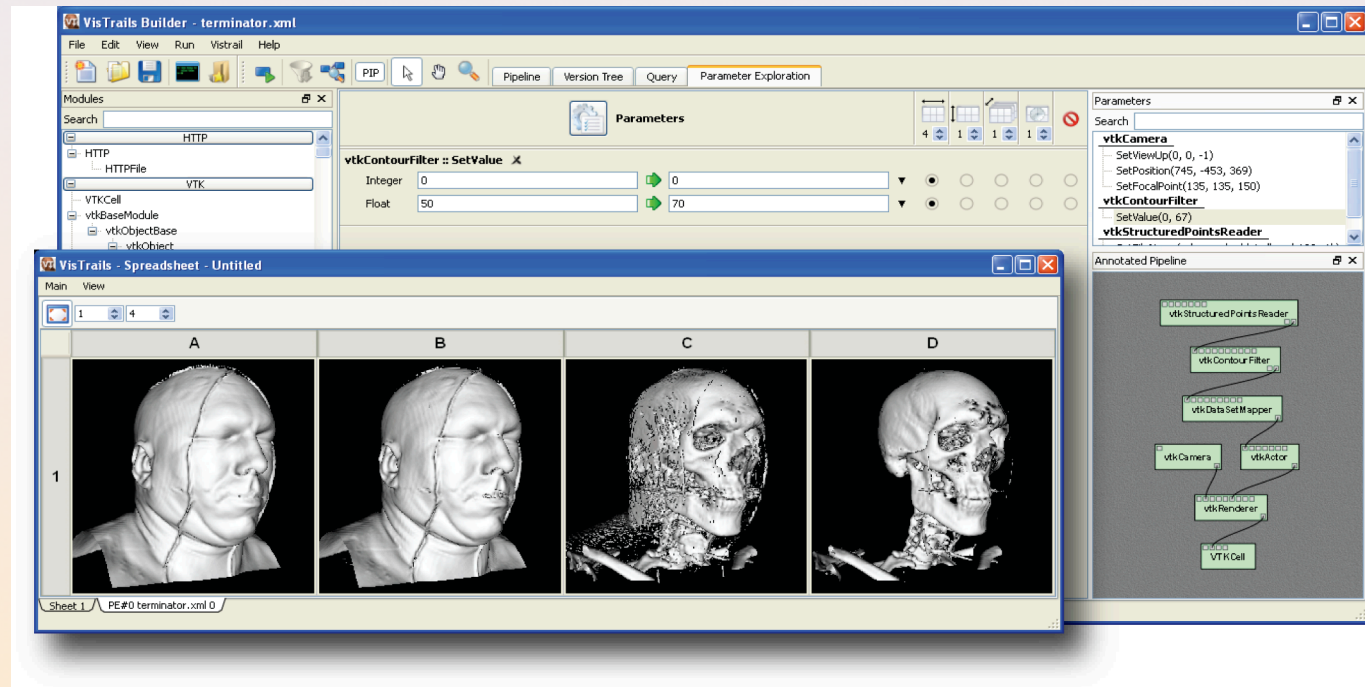
$$vt_1 - vt_2 = \{x_i, x_{i-1}, \dots, x_1, \emptyset\} - \{x_j, x_{j-1}, \dots, x_1, \emptyset\}$$

- ◆ 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]

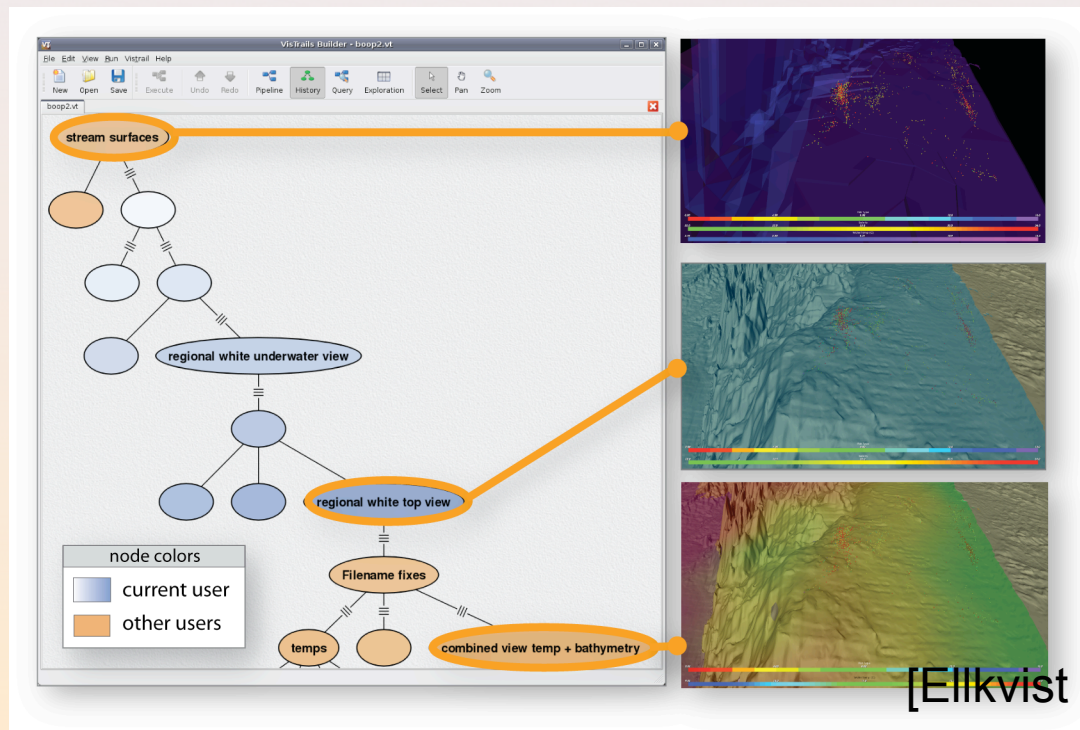


# Exploring the Change Space

- ◆ Scripting workflows: Parameter explorations are simple to specify and apply
- ◆ Exploration of parameter space for a workflow  $\mathbf{v}_t$   
( $setParameter(id_n, value_n) \circ \dots \circ (setParameter(id_1, value_1) \circ \mathbf{v}_t)$ )
- ◆ Exploration of multiple workflow specifications  
( $addModule(id_i, \dots) \circ (deleteModule(id_j) \circ \mathbf{v}_1)$   
...  
( $addModule(id_i, \dots) \circ (deleteModule(id_j) \circ \mathbf{v}_n)$ )
- ◆ Results can be conveniently compared in the VisTrails spreadsheet
- ◆ Can create animations too!
- ◆ Caching to avoid redundant computations [Bavoil et al., IEEE Vis 2005]

# Provenance Beyond Reproducibility

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



[Elkvist et al., IPAW 2008]

# Collaborative Exploration

---

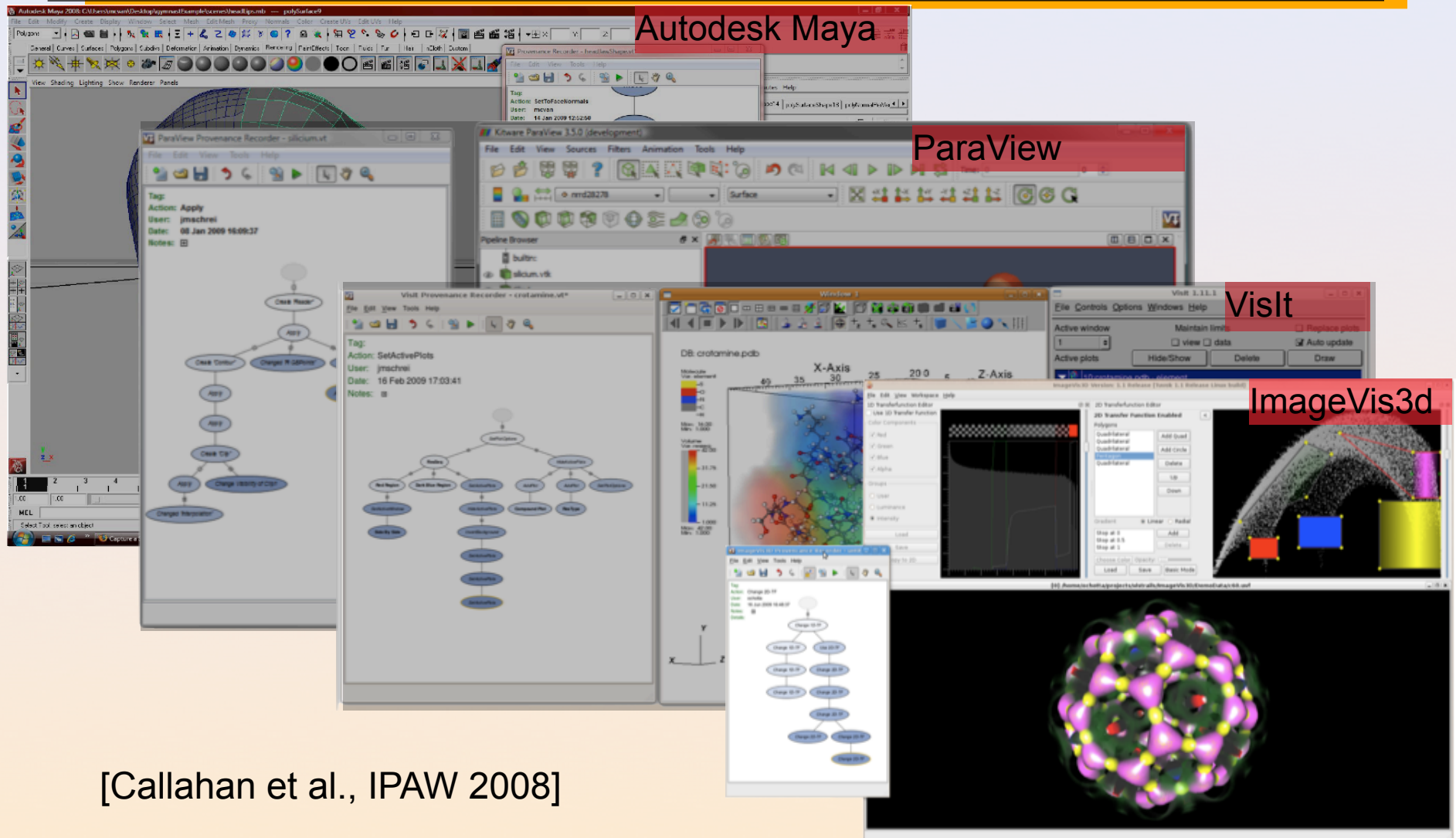
- ◆ Collaboration is key to data exploration
  - Translational, integrative approaches to science
- ◆ Store provenance information in a database
- ◆ Synchronize concurrent updates through locking
  - Real-time collaboration [Ellkvist et al., IPAW 2008]
- ◆ Asynchronous access: similar to version control systems
  - Check out, work offline, synchronize
  - Users exchange patches
- ◆ No need for a central repository—support for distributed collaboration
  - For details see Callahan et al., SCI Institute Technical Report, No. UUSCI-2006-016 2006

# 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](http://www.cs.utah.edu/~juliana/videos/paraview_plugin.avi)

# Change-Based Provenance: Summary

---

- ◆ General: Works with any system that has undo/redo!
- ◆ Concise representation
- ◆ Uniformly captures data and workflow provenance
  - Data provenance: where does a specific data product come from?
  - Workflow evolution: how has workflow structure 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
  - Scalable exploration of the parameter space—results can be compared side-by-side in the spreadsheet
  - Support for collaboration
  - Understand problem-solving strategies—knowledge re-use

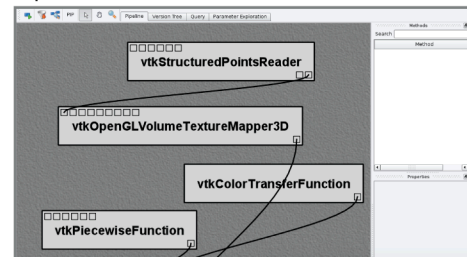
# Querying and Re-Using Provenance



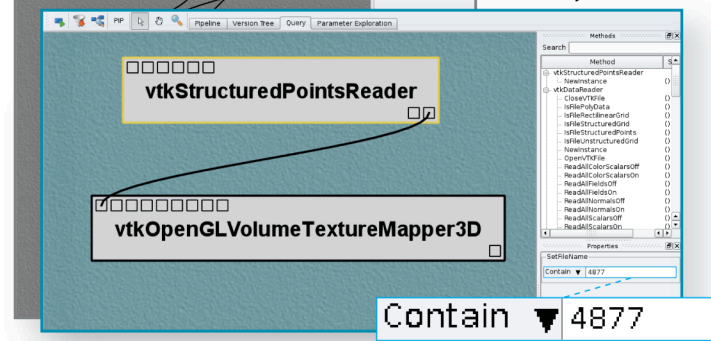
# Querying Provenance by Example

- ◆ Provenance is represented as graphs: 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

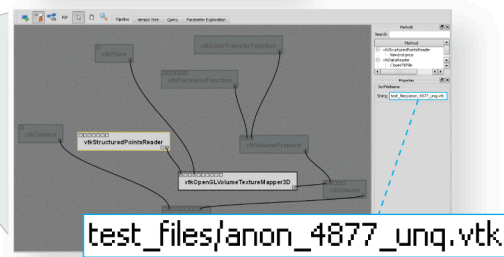
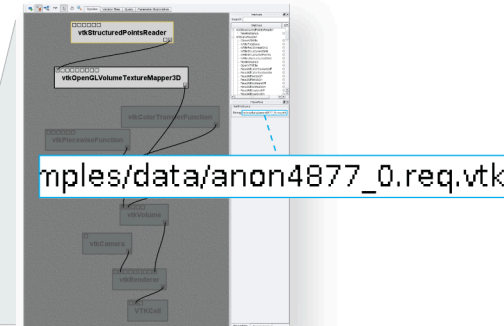
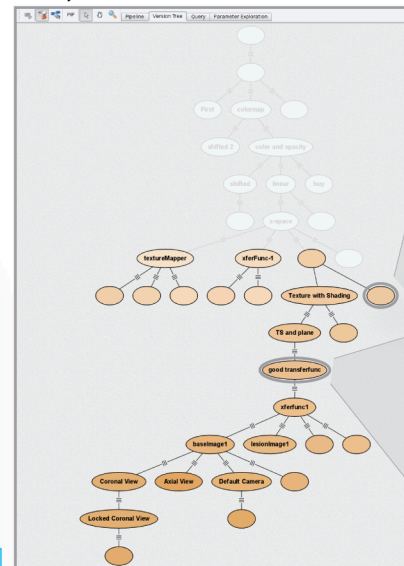
Pipeline Interface



Query Interface



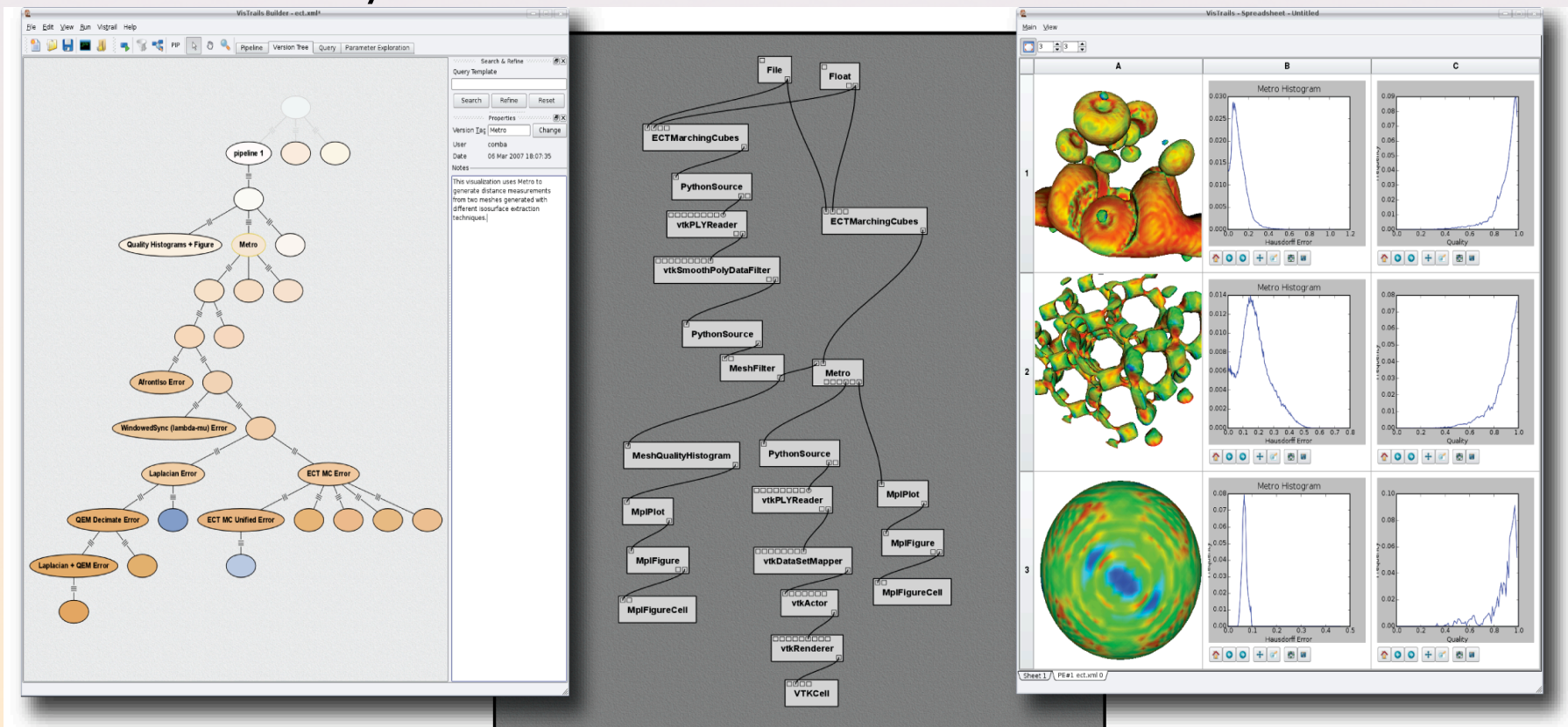
Query Result



# Creating Workflows

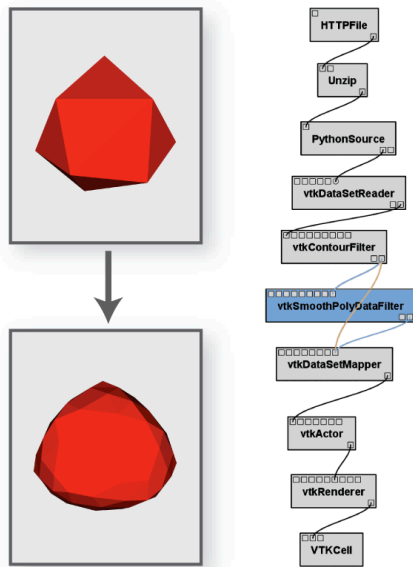
- ◆ Complex workflows are hard to assemble
  - Programming expertise
  - Domain knowledge
  - Familiarity with different tools

*Steep learning curve*

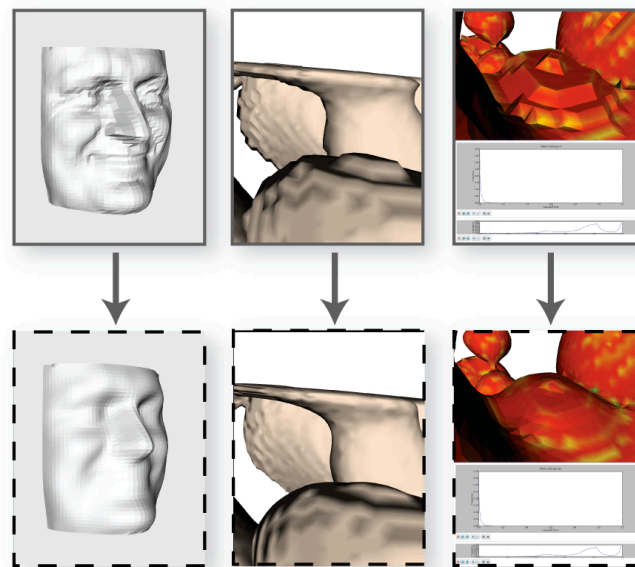


# Refining Analyses by Analogy

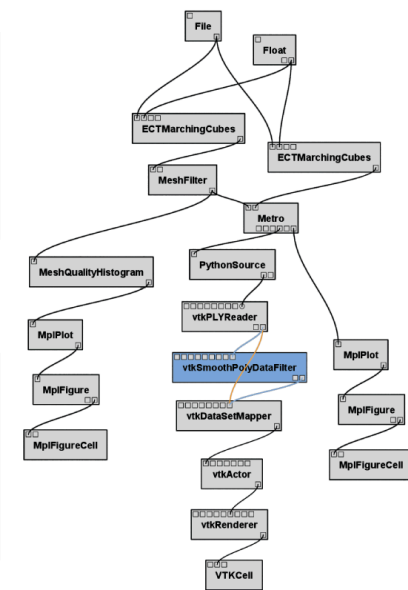
- ◆ Leverage the wisdom of the crowds in *shared provenance*
  - Some refinements are common, e.g., change the rendering technique, publish image on the Web
- ◆ Apply refinements by analogy, automatically



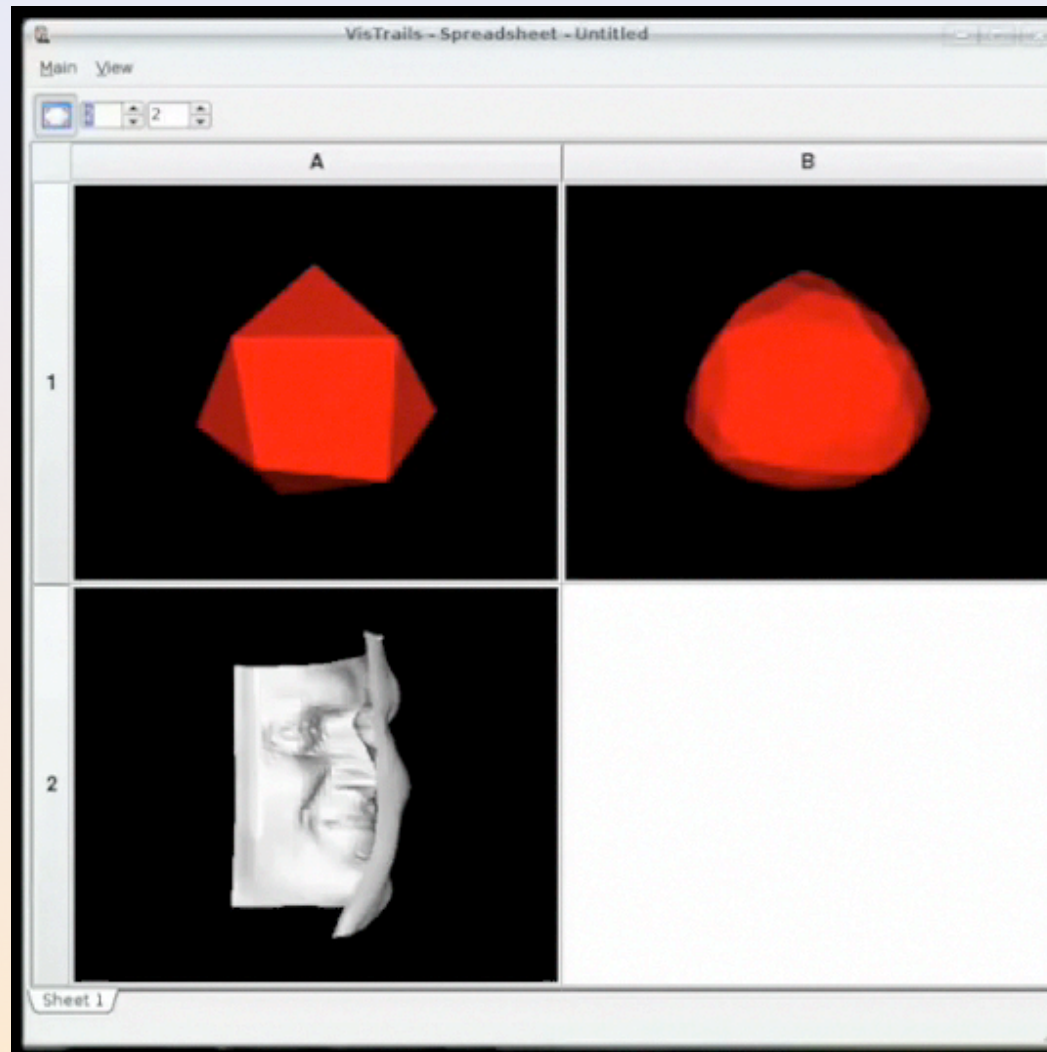
Analogy Template



Automatically constructed visualizations



# Generating Visualizations by Analogy



[Scheidegger et al, IEEE TVCG 2007]

<http://www.cs.utah.edu/~juliana/videos/Analogies.m4v>

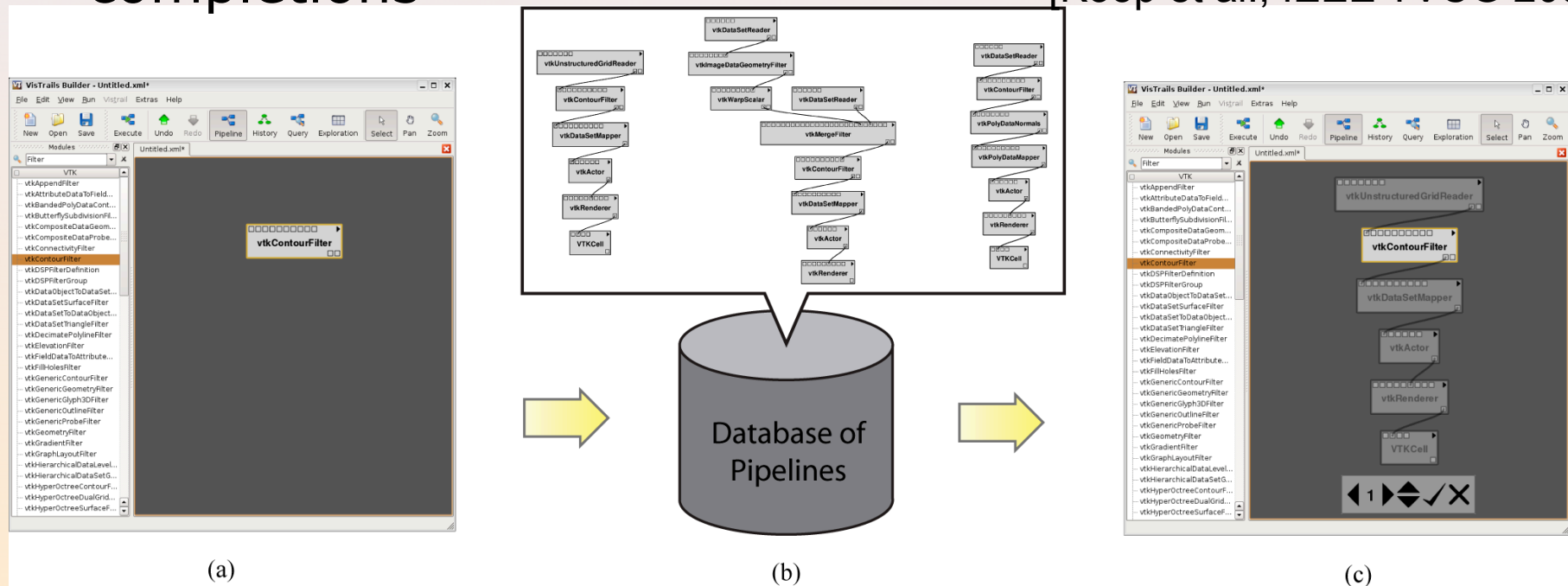
# The Need for Guidance in Workflow Design



# VisComplete: A Workflow Recommendation System

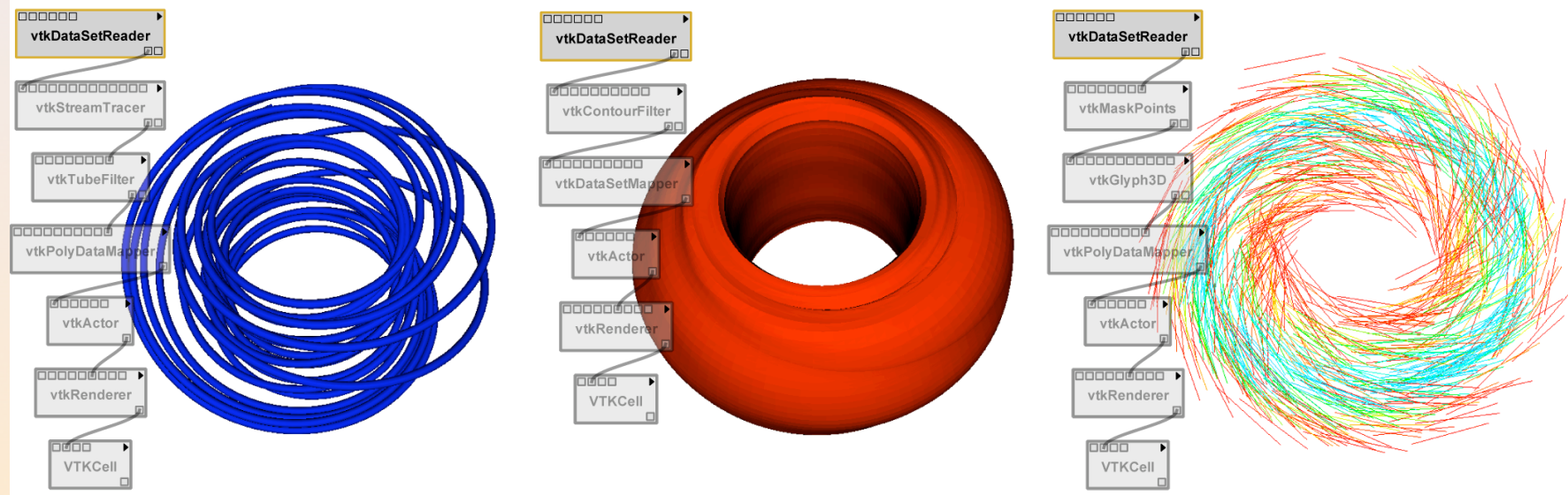
- ◆ *Mine* provenance collection: Identify graph fragments that co-occur in a collection of workflows
- ◆ Predict sets of likely workflow additions to a given partial workflow
- ◆ Similar to a Web browser suggesting URL completions

[Koop et al., IEEE TVCG 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

[http://www.cs.utah.edu/~juliana/videos/viscomplete\\_h\\_264.mov](http://www.cs.utah.edu/~juliana/videos/viscomplete_h_264.mov)



# Publishing and Sharing Results

# Scientific Publications and Provenance

## 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, Texas

Submitted 22 February 2005; accepted in final form 10 March 2005

Coyle, Edward F. Improved muscular efficiency displayed as Tour de France champion matures. *J Appl Physiol* 98: 2191–2196, 2005. First published March 17, 2005; doi:10.1152/jappphysiol.00216.2005.— This case describes the physiological maturation from ages 21 to 28 yr of the bicyclist who has now become the six-time consecutive Grand Champion of the Tour de France, at ages 27–32 yr. Maximal oxygen uptake ( $\dot{V}O_{2\max}$ ) in the trained state remained at  $\sim 6$  l/min, lean body weight remained at  $\sim 70$  kg, and maximal heart rate declined from 207 to 200 beats/min. Blood lactate threshold was typical of competitive cyclists in that it occurred at 76–85%  $\dot{V}O_{2\max}$ , yet maximal blood lactate concentration was remarkably low in the trained state. It appears that an 8% improvement in muscular efficiency and thus power production when cycling at a given oxygen uptake ( $\dot{V}O_2$ ) is the characteristic that improved most as this athlete matured from ages 21 to 28 yr. It is noteworthy that at age 25 yr, this champion developed advanced cancer, requiring surgeries and chemotherapy. During the months leading up to each of his Tour de France victories, he reduced body weight and body fat by 4–7 kg (i.e.,  $\sim 7\%$ ). Therefore, over the 7-yr period, an improvement in muscular efficiency and reduced body fat contributed equally to a remarkable 18% improvement in his steady-state power per kilogram body weight when cycling at a given  $\dot{V}O_2$  (e.g., 5 l/min). It is hypothesized that the improved muscular efficiency probably reflects changes in muscle myosin type stimulated from years of training intensively for 3–6 h on most days.

maximum oxygen uptake; blood lactate concentration

MUCH HAS BEEN LEARNED about the physiological factors that contribute to endurance performance ability by simply describing the characteristics of elite endurance athletes in sports such as distance running, bicycle racing, and cross-country skiing. The numerous physiological determinants of endurance have been organized into a model that integrates such factors as maximal oxygen uptake ( $\dot{V}O_{2\max}$ ), the blood lactate threshold, and muscular efficiency, as these have been found to be the most important variables (7, 8, 15, 21). A common approach has been to measure these physiological factors in a given athlete at one point in time during their competitive career and to compare this individual's profile with that of a population of peers (4, 6, 15, 16, 21). Although this approach describes the variations that exist within a population, it does not provide information about the extent to which a given athlete can improve their specific physiological determinants of endurance with years of continued training as the athlete matures and reaches his/her physiological potential. There are remarkably few longitudinal reports documenting the changes in physiological factors that accompany years of continued endurance training at the level performed by elite endurance athletes.

This case study reports the physiological changes that occur in an individual bicycle racer during a 7-yr period spanning

ages 21 to 28 yr. Description of this person is noteworthy for two reasons. First, he rose to become a six-time and present Grand Champion of the Tour de France, and thus adaptations relevant to this feat were identified. Remarkably, he accomplished this after developing and receiving treatment for advanced cancer. Therefore, this report is also important because it provides insight, although limited, regarding the recovery of "performance physiology" after successful treatment for advanced cancer. The approach of this study will be to report results from standardized laboratory testing on this individual at five time points corresponding to ages 21.1, 21.5, 22.0, 25.9, and 28.2 yr.

### METHODS

**General testing sequence.** On reporting to the laboratory, training, racing, and medical histories were obtained, body weight was measured ( $\pm 0.1$  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 determined as the subject bicycled a stationary ergometer for 25 min, with work rate increasing progressively every 5 min over a range of 50, 60, 70, 80, and 90%  $\dot{V}O_{2\max}$ . After a 10- to 20-min period of active recovery,  $\dot{V}O_{2\max}$  when cycling was measured. Thereafter, body composition was determined by hydrostatic weighing and/or analysis of skin-fold thickness (34, 35).

**Measurement of  $\dot{V}O_{2\max}$ .** The same Monark ergometer (model 819) equipped with a racing seat and drop handlebars and pedals for cycling shoes was used for all cycle testing, and seat height and saddle position were held constant. The pedal's crank length was 170 mm.  $\dot{V}O_{2\max}$  was measured during continuous cycling lasting between 8 and 12 min, with work rate increasing every 2 min. A leveling off of oxygen uptake ( $\dot{V}O_2$ ) always occurred, and this individual cycled until exhaustion at a final power output that was 10–20% higher than the minimal power output needed to elicit  $\dot{V}O_{2\max}$ . A venous blood sample was obtained 3–4 min after exhaustion for determination of blood lactate concentration after maximal exercise, as described below. The subject breathed through a Daniels valve; expired gases were continuously sampled from a mixing chamber and analyzed for  $O_2$  (Applied Electrochemistry S3A) and  $CO_2$  (Beckman LB-2). Inspired air volumes were measured using a dry-gas meter (Parkinson-Cowan CD4). These instruments were interfaced with a computer that calculated  $\dot{V}O_2$  every 30 s. The same equipment for indirect calorimetry was used over the 7-yr period, with gas analyzers calibrated against the same known gasses and the dry-gas meter calibrated periodically to a 350-liter Tissot spirometer.

**Blood LT.** The subject pedaled the Monark ergometer (model 819) continuously for 25 min at work rates eliciting  $\sim 50, 60, 70, 80,$  and  $90\%$   $\dot{V}O_{2\max}$  for each successive 5-min stage. The calibrated ergometer was set in the constant power mode, and the subject maintained a pedaling cadence of 85 rpm. Blood samples were obtained either from

Address for reprint requests and other correspondence: E. F. Coyle, Bellmont Hall 222, Dept. of Kinesiology and Health Education, The Univ. of Texas at Austin, Austin, TX 78712 (E-mail: coy@hmail.utexas.edu).

http://www.jap.org

8750-7581/05 \$8.00 Copyright © 2005 the American Physiological Society

2191

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

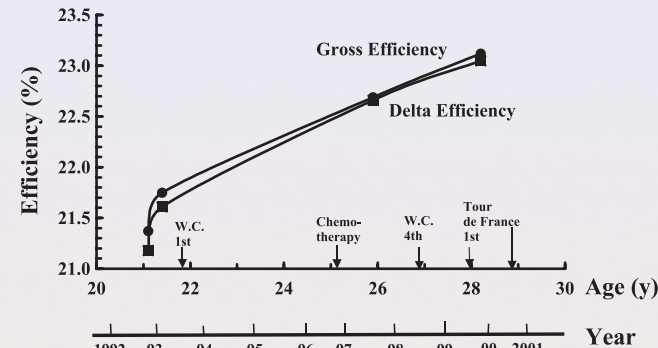


Fig. 1. Mechanical efficiency when bicycling expressed as "gross efficiency" and "delta efficiency" over the 7-yr period in this individual. WC, World Bicycle Road Racing Championships, 1st and 4th place, respectively. Tour de France 1st, Grand Champion of the Tour de France in 1999–2004.

### METHODS

**General testing sequence.** On reporting to the laboratory, training, racing, and medical histories were obtained, body weight was measured ( $\pm 0.1$  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 determined as the subject bicycled a stationary ergometer for 25 min, with work rate increasing progressively every 5 min over a range of 50, 60, 70, 80, and 90%  $\dot{V}O_{2\max}$ . After a 10- to 20-min period of active recovery,  $\dot{V}O_{2\max}$  when cycling was measured. Thereafter, body composition was determined by hydrostatic weighing and/or analysis of skin-fold thickness (34, 35).

# Scientific Publications and Provenance

Improved muscular efficiency displayed as Tour de France champion matures

"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."

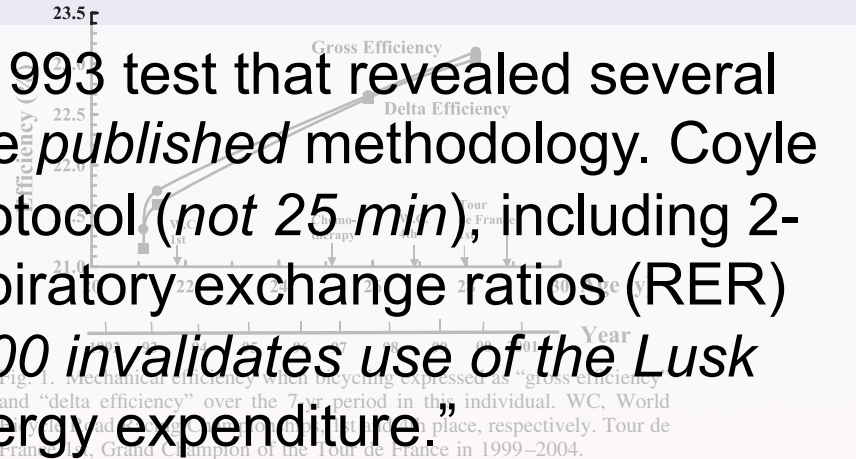
<http://jap.physiology.org/cgi/content/full/105/3/1020>

Address for reprint requests and other correspondence: E. F. Coyle, Bellmont Hall 222, Dept. of Kinesiology and Health Education, The Univ. of Texas at Austin, Austin, TX 78712 (E-mail: coyef@mail.utexas.edu).

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734, provided that the data and facts are not misrepresented.

<http://www.jap.org>

8750-7581/05 \$8.00 Copyright © 2005 the American Physiological Society



## METHODS

The same Moxak ergometer (model 819; Monark Ergonomic AB, Sweden) was used in the present study. The subject pedaled the Monark ergometer (model 819; Monark Ergonomic AB, Sweden) at a work rate of 250 W for 25 min, with work rate increasing progressively every 5 min over a range of 50, 60, 70, 80, and 90%  $\dot{V}O_{2\max}$ . After a 10- to 20-min period of active recovery,  $\dot{V}O_{2\max}$  when cycling was measured. Thereafter, body composition was determined by hydrostatic weighing and/or analysis of skin-fold thickness (34, 35).

[http://en.wikipedia.org/wiki/Scientific\\_misconduct](http://en.wikipedia.org/wiki/Scientific_misconduct)  
<http://ori.dhhs.gov/misconduct/cases/>

# Provenance-Rich Publications

---

- ◆ Bridge the gap between the scientific process and publications
- ◆ Results that can be reproduced and validated
  - Papers with *deep* captions
  - Encouraged by ACM SIGMOD and a number of journals
- ◆ Higher-quality publications (?)
- ◆ Describe more of the discovery process: people only describe successes, can we learn from mistakes?
- ◆ Support dynamic and interactive publications
- ◆ Expose users to different techniques and tools
  - Users can learn by example; expedite their training; and potentially reduce their time to insight

# Vision: Provenance-Rich Documents

## CrowdLabs: Social Analysis and Visualization for the Sciences

Emanuel Santos, Philip 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

The infrastructure to design and conduct scientific experiments has not kept pace with our collective ability to gather data. This has led to an unprecedented situation: Data analysis and visualization are now the bottleneck to discovery. This problem is compounded as interdisciplinary groups collaborate and need to perform a wide range of analyses targeted to multiple audiences.

Consider, for example, ocean observatories such as CMOP [7]. Data is gathered from a network of heterogeneous observation platforms as well as from large-scale simulation models of ocean circulation. The platforms consist of fixed and mobile stations with different sensors measuring physical properties, such as temperature, salinity and water level; and biochemical properties, such as nitrate, chlorophyll and dissolved organic matter concentrations. These sensors generate millions of measurements every day. Simulation results are generated by a suite of daily forecasts targeting specific estuaries, and long-term hindcast databases, where the simulations are re-executed using observed data as inputs. By analyzing these data, scientists from multiple disciplines (including biologists, chemists, and environmental scientist) aim to predict oceanographic features with practical realism. Because of the broad influence an ocean observatory, there is an intrinsic heterogeneity in data sources and analysis techniques used as well as in the derived data products (e.g., plots, 3D visualizations) for various stakeholders. Besides scientists, data products are also used by policy makers, students and the general public. To generate these data products, a steep technological learning curve is required for the scientists, who need to be aware of the details of data sources and how to access them, as well as use specialized tools for manipulating data and deriving insightful visualizations. Even for experienced users, there are no accepted “best practices” that ensure the wealth of information produced by observations, predictions and analysis is effectively used.

**Social Analysis of Scientific Data.** We posit that by facilitating the social analysis of scientific data, we can overcome many of these challenges. When users share their analyses and visualizations, they can 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, expedite their scientific training in disciplinary and inter-disciplinary settings; and potentially reduce the time lag between data acquisition and scientific insight. Recently, social Web sites have been created that enable users to collaboratively visualize data [7, 7]. They allow users to create and discuss visualizations of a wide range of data sets. However, they fail to cater to important requirements of scientific exploration. In particular, they were designed for small data sets and only provide a limited set of visualizations.

Science portals [7, 7, 7], on the other hand, have focused on simplifying data exploration by aggregating data from different sources and by providing a set of canned analyses and visualizations. However, they have important limitations. They are insufficient for handling large volumes of heterogeneous data and the diversity of stakeholders and their needs: it is simply not possible for IT personnel to anticipate all necessary analyses and different ways to correlate and integrate data. Furthermore, while some analyses that are used regularly can be canned, others are ground-breaking and need to be created, altered on-the-fly, and improved as part of a collaborative effort. Last, but not least, many small research groups do not have the necessary resources to create such portals.

**CrowdLabs.** In this paper we describe CrowdLabs, a system that adopts the model used by social Web sites and that integrates a set of usable tools and a scalable infrastructure to provide a rich collaborative environment for scientists. CrowdLabs combines benefits of social Web sites and science portals while at the same time addressing their limitations. Similar to social Web sites, CrowdLabs aims to foster collaboration, but unlike these sites, it was specifically designed to support the needs of computational scientists, including the ability to access high-performance computers and manipulate large volumes of data. By providing mechanisms that simplify the publishing and use of analysis pipelines, it allows IT personnel and end users to collaboratively construct and refine portals. Thus, CrowdLabs lowers the barriers for the use of scientific analyses and enables broader audiences to contribute insights to the scientific exploration process, without the high costs incurred by traditional portals. In addition, it supports a more dynamic environment where new exploratory analyses can be added on-the-fly.

Another important feature of CrowdLabs is the provenance support [7, 7]. Publishing scientific results together with their provenance, the details of how the results were obtained, not only makes the results more transparent, but it also enables others to reproduce and validate the results. CrowdLabs leverages provenance information (e.g., workflow/pipeline specifications, libraries, packages, users, datasets

• Emanuel Santos, Philip Mates, Juliana Freire, and Cláudio T. Silva are with the Scientific Computing and Imaging (SCI) Institute at the University of Utah, email: {emmanuel.mates, juliana.freire, csilva}@sci.utah.edu

Manuscript received 17 March 2010; accepted 1 August 2010; posted online 24 October 2010; mailed on 18 October 2010.  
For information on obtaining reprints of this article, please email: reprints@computing.org.

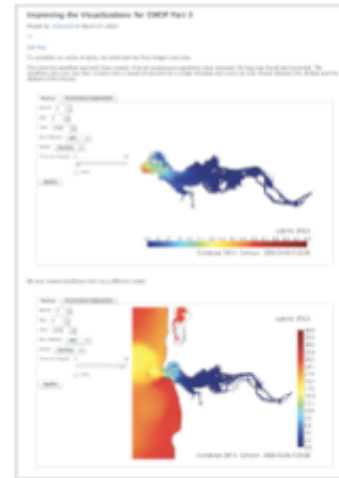


Fig. 7: Using the blog to document processes: A visualization expert created a series of blog posts to explain the problems found when generating the visualizations for CMOP.

### ACKNOWLEDGMENTS

Our research has been funded by the National Science Foundation (grants: IIS-0905385, IIS-0748500, ATM-0838021, IIS-0844548, CNS-0751152, IIS-0713673, OCE-0424602, IIS-0534928, CNS-0514485, IIS-0513692, CNS-0534996, CCF-0401498, OISE-0405402, CCF-0528201, CNS-0551724), the Department of Ecology SciDAC (OACET and SDM centers), and IBM Faculty Awards (2005, 2006, 2007, and 2008). E. Santos is partially supported by a CAPES/Fulbright fellowship.

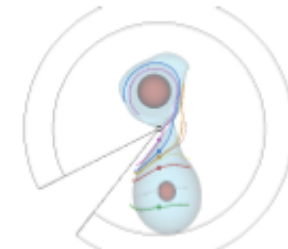


Fig. 8: Visualizing a binary star system simulation. This is an image that was generated by embedding a workflow directly in the text. The original workflow is available at <http://www.crowdlabs.org/vistrails/workflows/details/32/>.

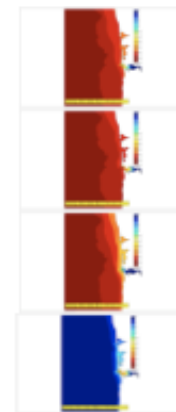
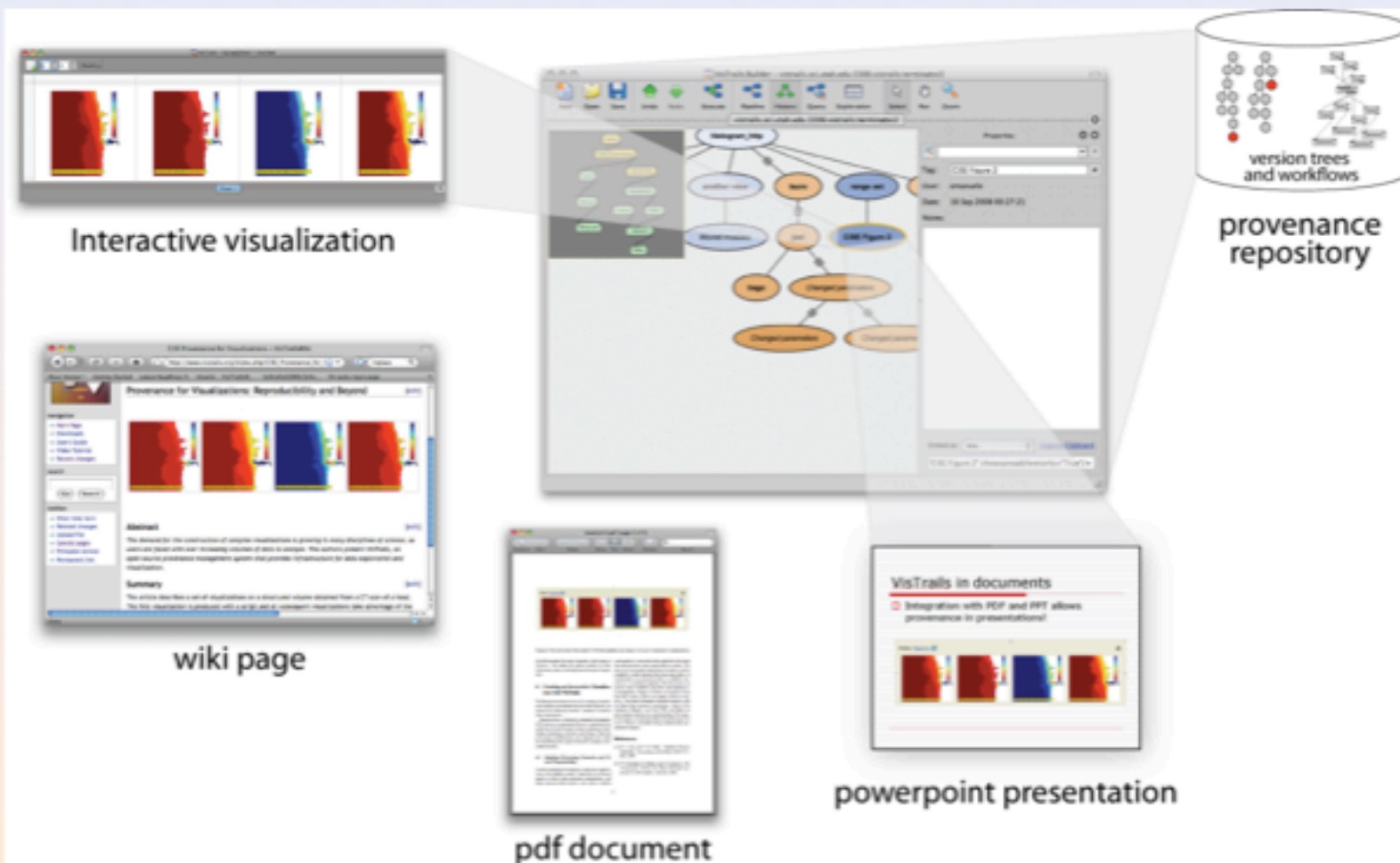


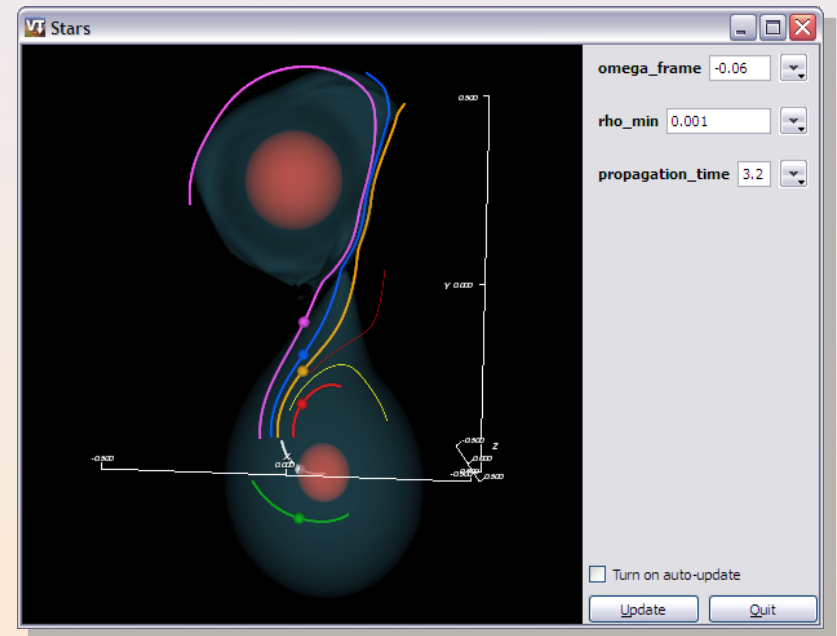
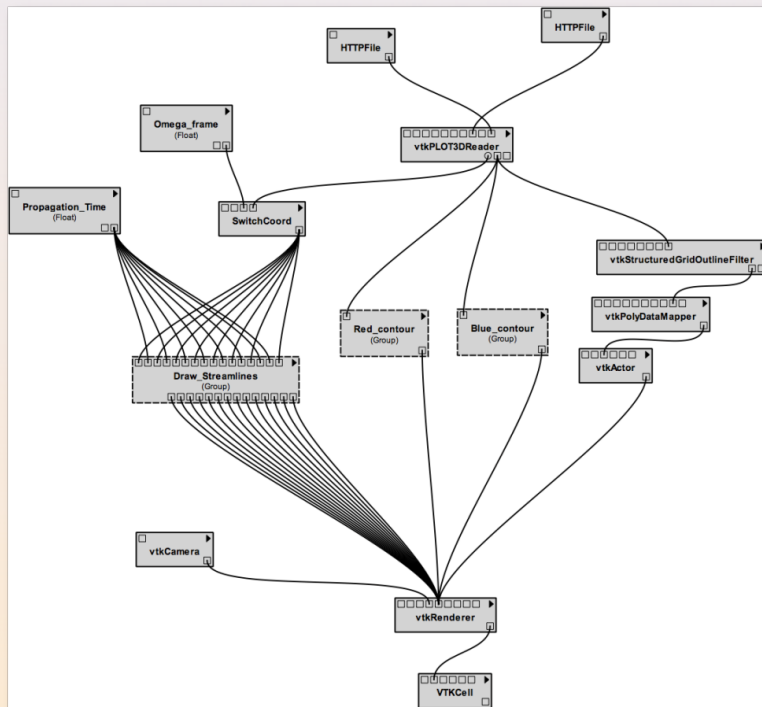
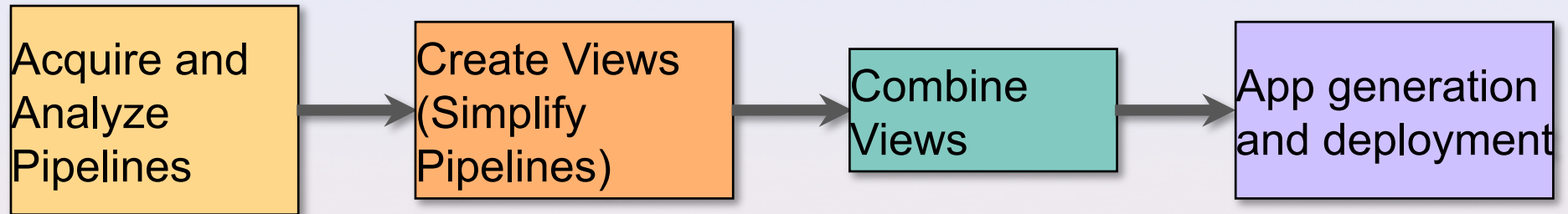
Fig. 9: Columbia river virtual estuary: visualization of salinity over time. See <http://www.crowdlabs.org/vistrails/workflows/details/32/>.

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

# Provenance-Rich Publications



# VisMashup: Creating Mashups from Workflows



[Santos et al, IEEE TVCG 2008]

# CrowdLabs

- ◆ Share workflows and analysis in addition to data

The screenshot displays the CrowdLabs web interface. At the top, there is a navigation bar with the CrowdLabs logo and a 'Login or Sign up' button. Below the navigation bar, there are tabs for 'Profiles', 'Vistrails', 'Workflows', 'Vismashups', 'Packages', 'Datasets', 'Blogs', 'Groups', 'Projects', and 'Inbox (0)'. The main content area is titled 'Latest Vismashups' and features two entries: 'Estuary Forecast (f22)' and 'Estuary by Date (DB14)'. The 'Estuary Forecast (f22)' entry includes a small thumbnail image, the text 'Viewed 0 times', '0 Comments', and the date 'March 29, 2010'. Below this, the 'Estuary by Date (DB14)' entry shows a complex network diagram representing a workflow. To the right, a detailed view of the 'Estuary Forecast (f22)' mashup is shown. This view includes a control panel with the following settings: 'Month' set to 3, 'Day' set to 3, 'Year' set to 2010, 'Run Scalars' set to 'salt', 'Depth' set to 'Surface', 'Run Day' set to 1, and 'Time (in hours)' set to 0. A slider for 'Time (in hours)' ranges from 0 to 23, with a 'loop' checkbox checked. An 'Update' button is located below the controls. To the right of the control panel is a map of an estuary system, with a color gradient overlay representing the forecasted data. The map shows a large body of water on the left, connected to a narrower channel that branches into several smaller channels on the right. The color gradient transitions from red on the left to blue on the right, indicating a change in the forecasted variable.



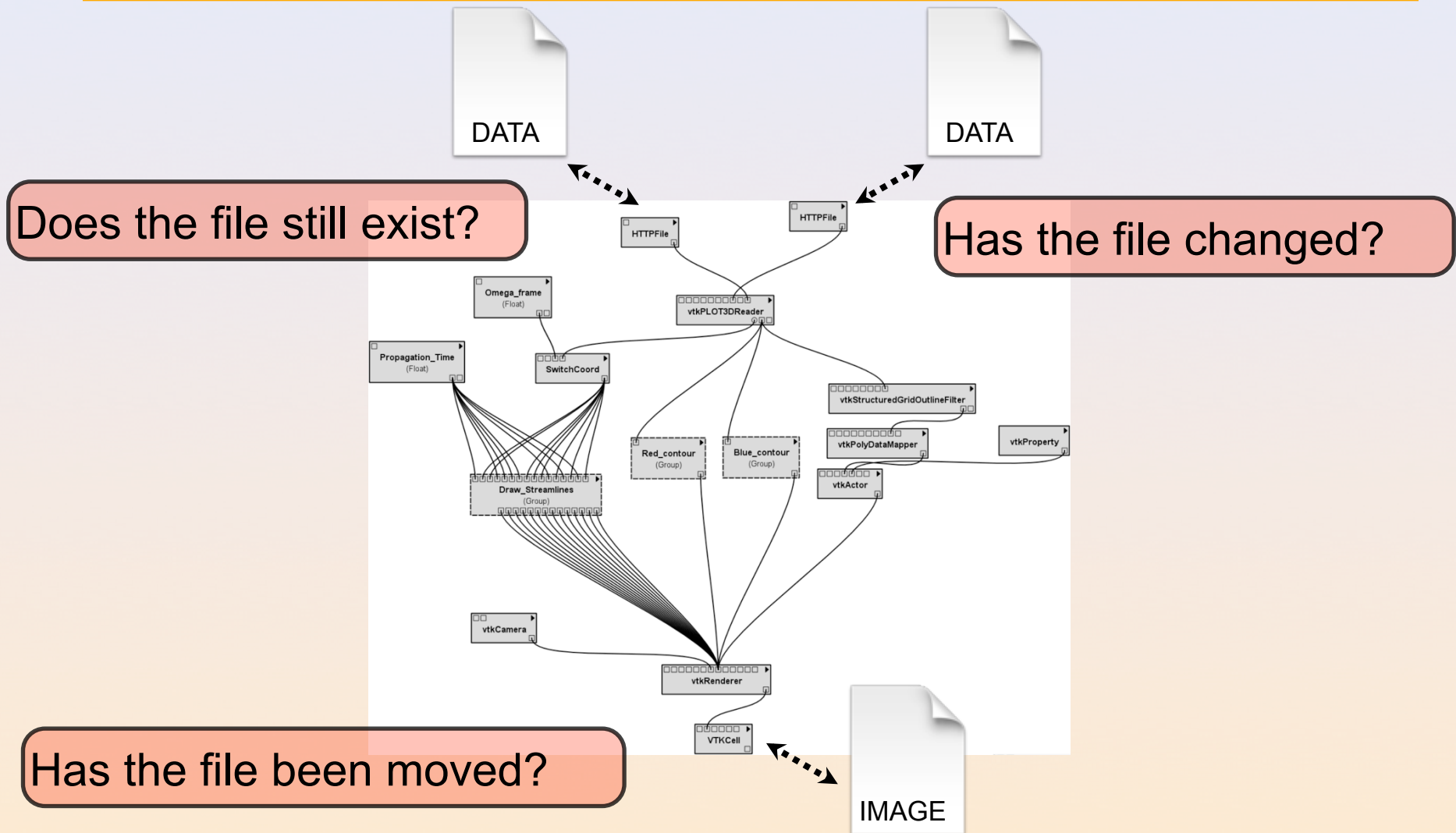
# VisTrails: Some Useful Features

# Extensibility

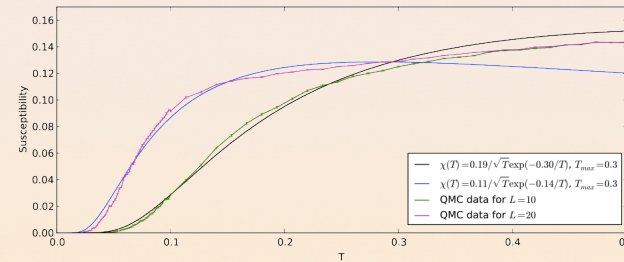
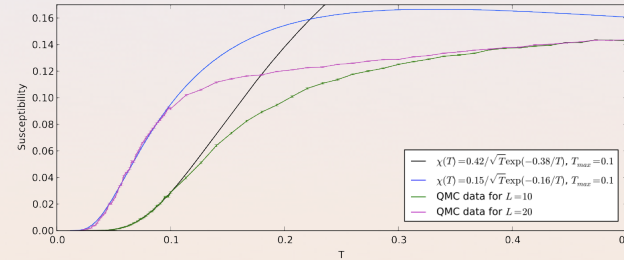
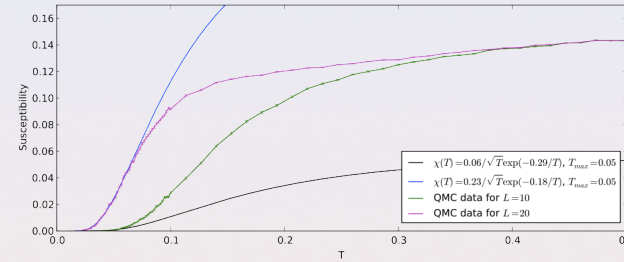
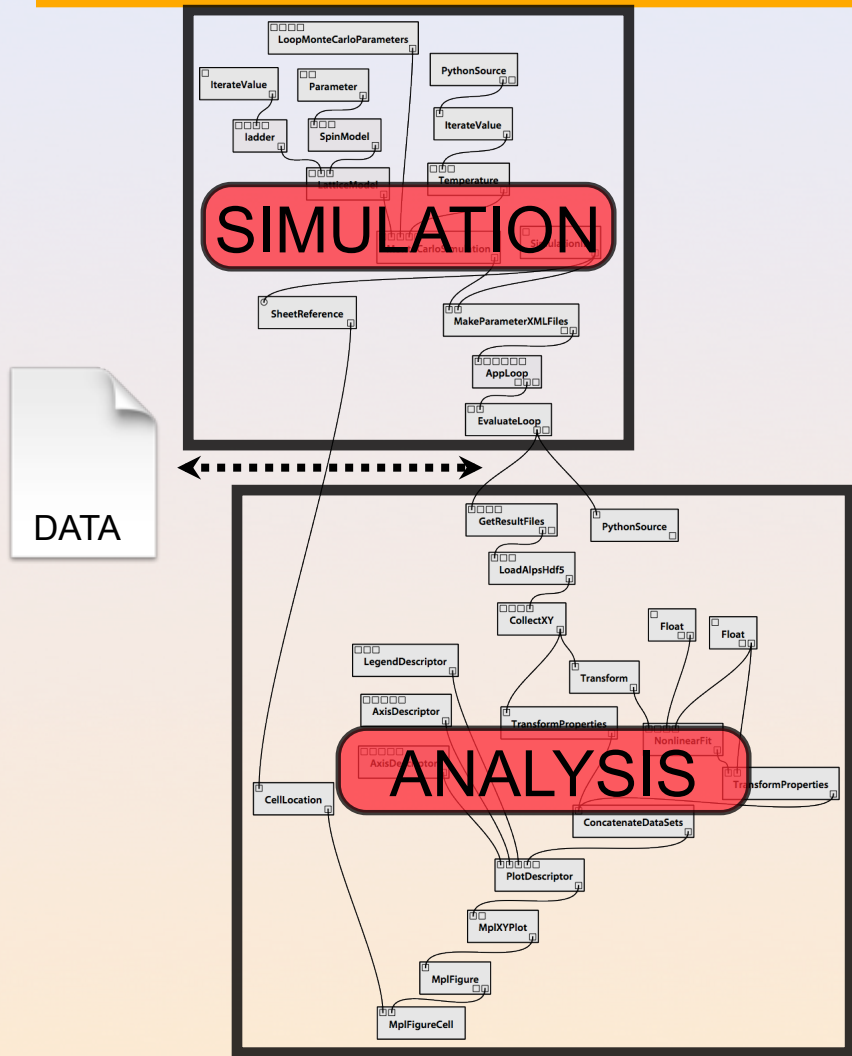
---

- ◆ Package infrastructure
- ◆ Wrap python libraries, command-line calls, or use other interfaces (jpytype, rpy, etc.)
- ◆ Need to specify:
  1. Package identification information
  2. Module structures: input & output ports
  3. Compute method for each module

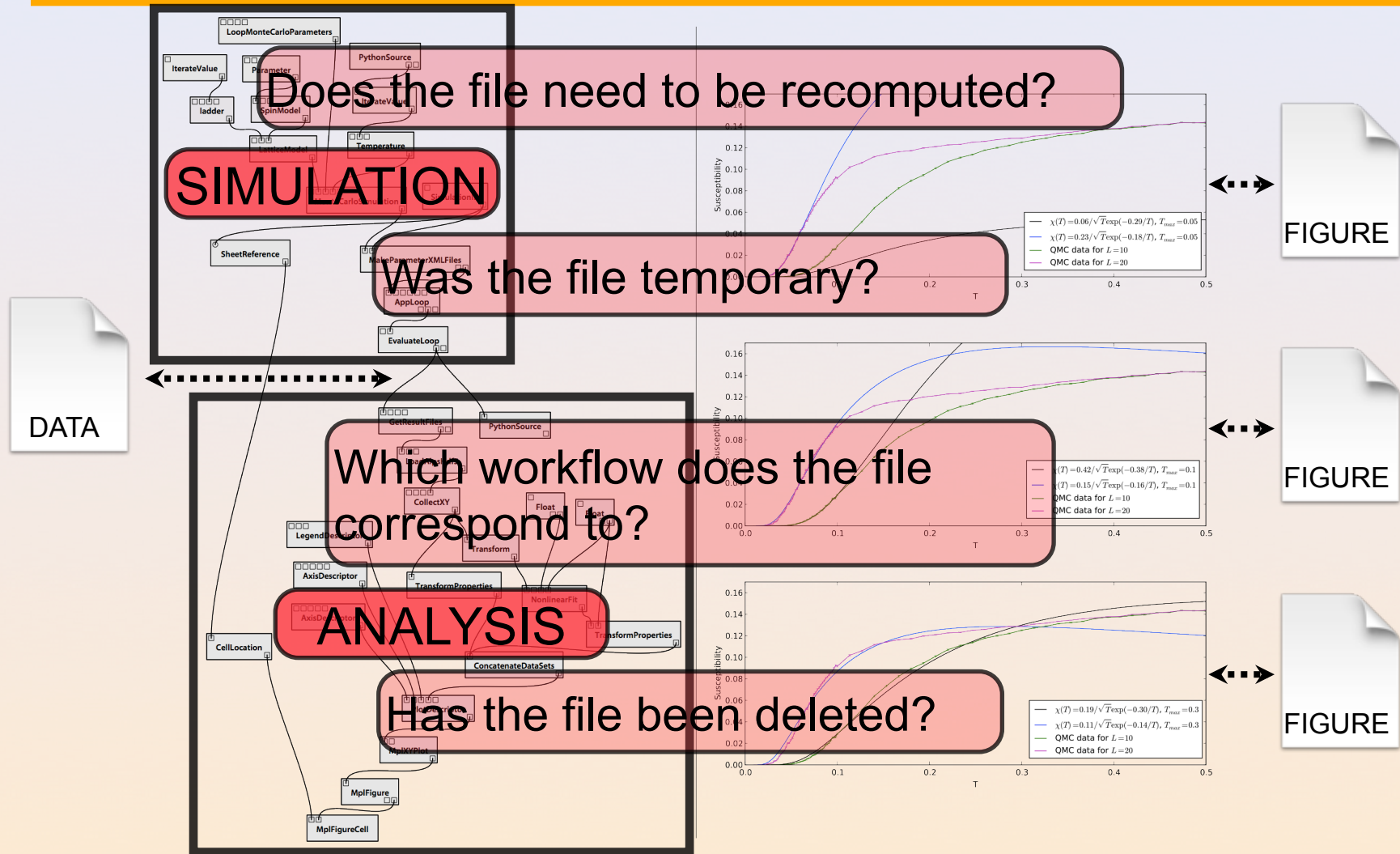
# Provenance Links to Data



# Computationally-Expensive Work



# Provenance Links to Data



# Strong Provenance Links

---

- ◆ Upstream signature
  - Identify workflow outputs by the computational steps and parameters that (may have) had an effect on the output
- ◆ Content hashing
  - Identify files by their content
  - Mirrored by version information from version control systems
- ◆ Universally Unique Identifiers (UUID)
  - Serve as “filenames”
  - Could use some other scheme here

# Strong Provenance Links: Implementation

---

- ◆ Implemented in VisTrails backed by git for file storage and sqlite3 for metadata
- ◆ Files or directories
- ◆ User identifies where inputs, outputs, and intermediates should be; each is represented by a module
- ◆ Modules can be configured:
  - Choice to link to existing data or create new reference
  - Allow naming and tagging (for search)
  - Configure links to local filesystem
- ◆ Caching for free!

# Input Configuration

**ManagedInputFile**

ManagedInputFile Module Configuration

Create New Reference

Path: /Users/dakooop/Code/javascriptTests.html

Name: javascriptTests.html

Tags: javascript test

Use Existing Reference

Name	Tags	User	Date Created	Date Modified	ID	Version	Content
▶ test java...	testing java...				774046ae-...		
▶ javascript...					05044c98-...		
▶ javascript...	javascript test				bfd52f36-1...		
▶					016b8eb6-...		

Keep Local Version

/Users/dakooop/Code/javascriptTests.html

Read From Local Path  Write To Local Path

Cancel OK



# Intermediate and Output Configuration

ManagedIntermediateDir

ManagedIntermediateDir Module Configuration

Always Create New Reference  
 Create New Reference

Name:   
Tags:

Use Existing Reference

Name	Tags	User	Date Created	Date Modified	ID	Version	Content
test output	band code				8d9c77d2-...		
test	band code				8236e45e-...		
					5fcca8a6-1...		
					002d9f32-...		

Keep Local Version

Read From Local Path  Write To Local Path

Cancel OK

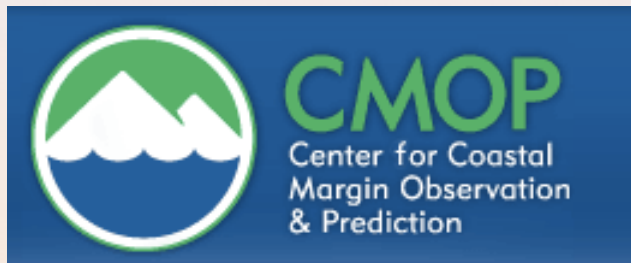
# Conclusions and Future Work

---

- ◆ Provenance management is crucial in many applications
  - VisTrails Project has introduced new algorithms and usable tools for querying, re-using and publishing provenance information
- ◆ VisTrails system is widely used in the scientific community
- ◆ Sharing provenance creates new opportunities [Freire and Silva, CHI SDA, 2008]
  - Expose users to different techniques and tools
  - Users can learn by example; expedite their training; and potentially reduce their time to insight
- ◆ Many challenges and several open computer science questions

# Acknowledgments

- ◆ Thanks to VisTrails and Web&DB groups
- ◆ This work is partially supported by the National Science Foundation, the Department of Energy, an IBM Faculty Award, and a University of Utah Seed Grant.



Thank you

