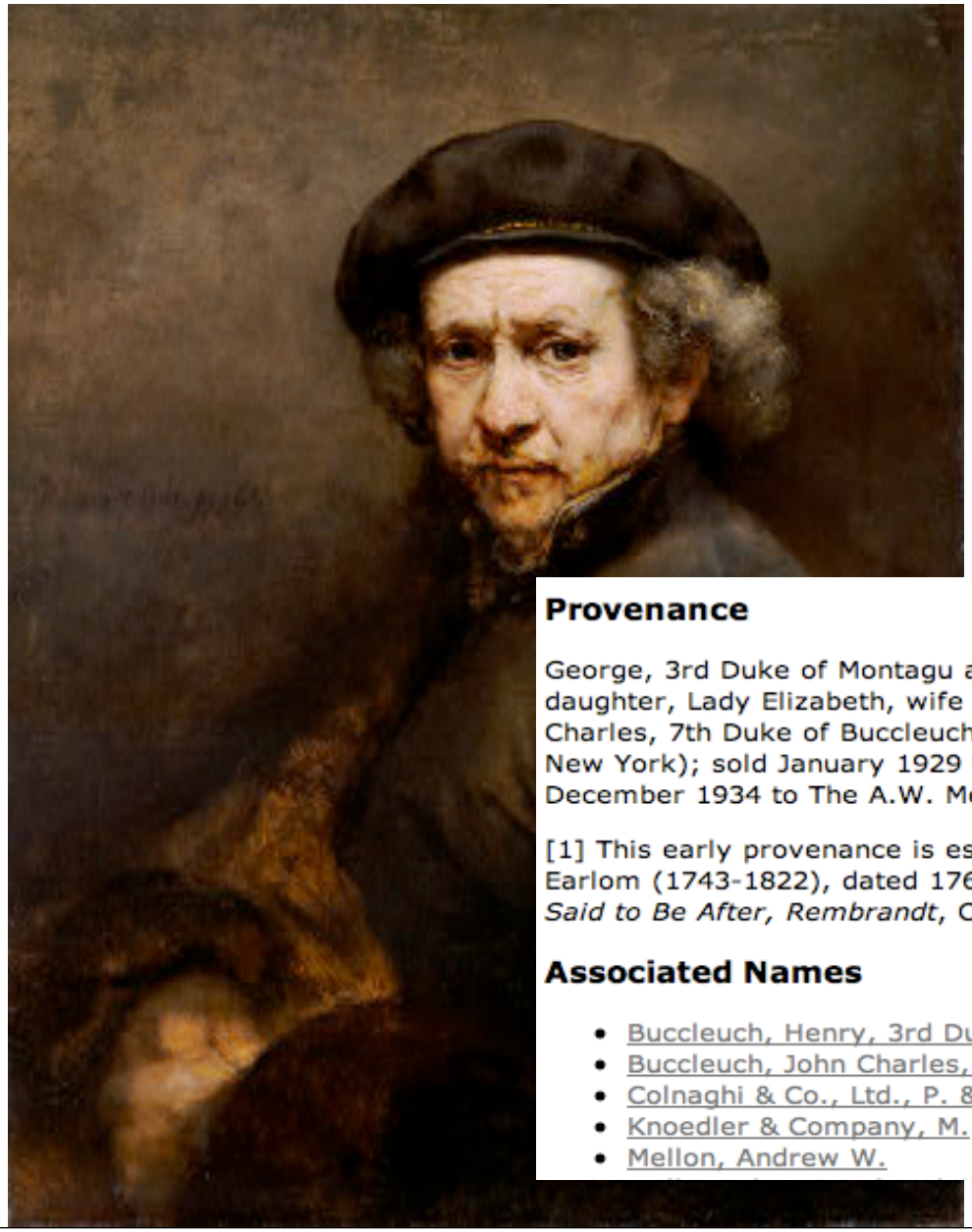
Provenance-Rich Visualization and Data Analysis

Claudio T. Silva
Visualization and Geometric Computing (VGC)
Scientific Computing and Imaging Institute
School of Computing
University of Utah

Special thanks to Juliana Freire and the VisTrails Group.

Funded by NSF, DOE, Sandia, ExxonMobil, NIH, IBM

Provenance in Art



Rembrandt van Rijn
Self-Portrait, 1659
Andrew W. Mellon Collection
1937.1.72

Provenance

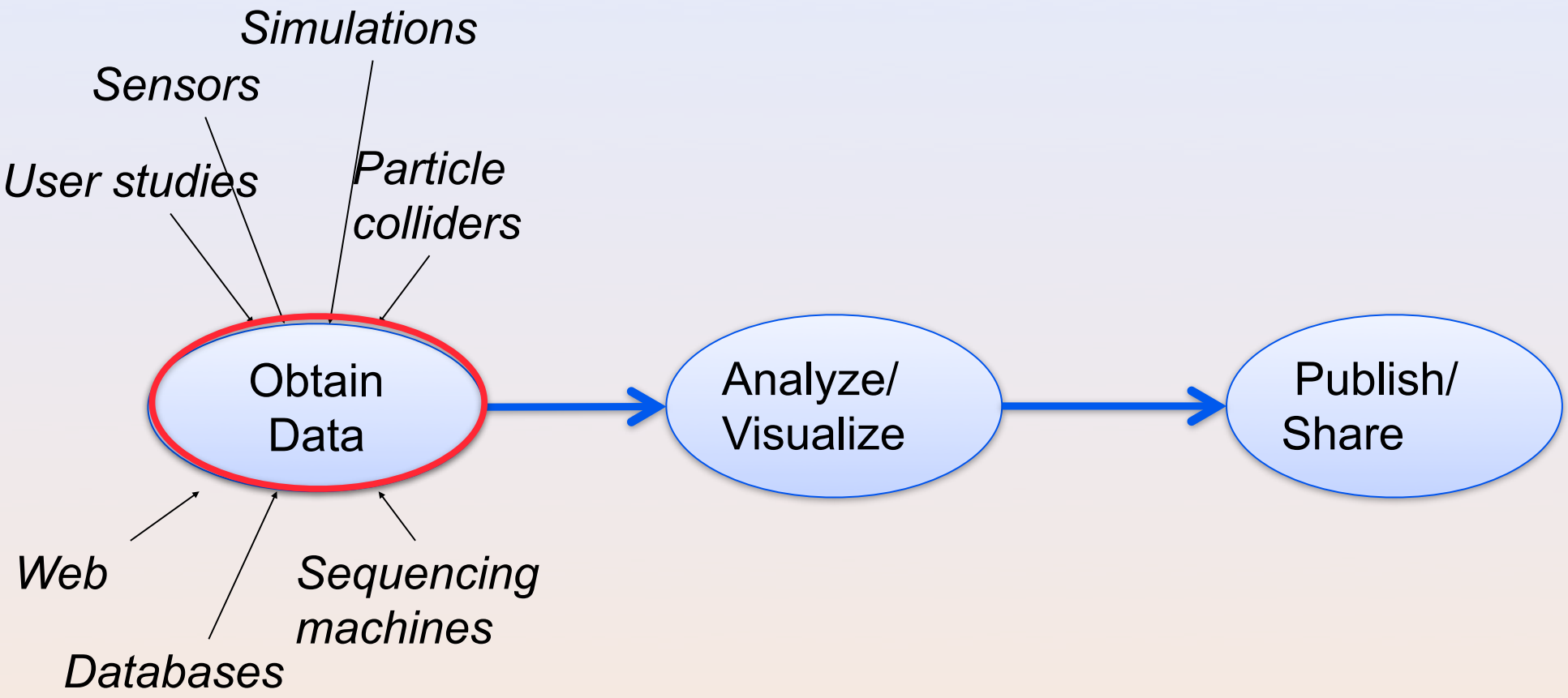
George, 3rd Duke of Montagu and 4th Earl of Cardigan [d. 1790], by 1767;[1] by inheritance to his daughter, Lady Elizabeth, wife of Henry, 3rd Duke of Buccleuch of Montagu House, London; John Charles, 7th Duke of Buccleuch; (P. & D. Colnaghi & Co., New York, 1928); (M. Knoedler & Co., New York); sold January 1929 to Andrew W. Mellon, Pittsburgh and Washington, D.C.; deeded 28 December 1934 to The A.W. Mellon Educational and Charitable Trust, Pittsburgh; gift 1937 to NGA.

[1] This early provenance is established by presence of a mezzotint after the portrait by R. Earlom (1743-1822), dated 1767. See John Charrington, *A Catalogue of the Mezzotints After, or Said to Be After, Rembrandt*, Cambridge, 1923, no. 49.

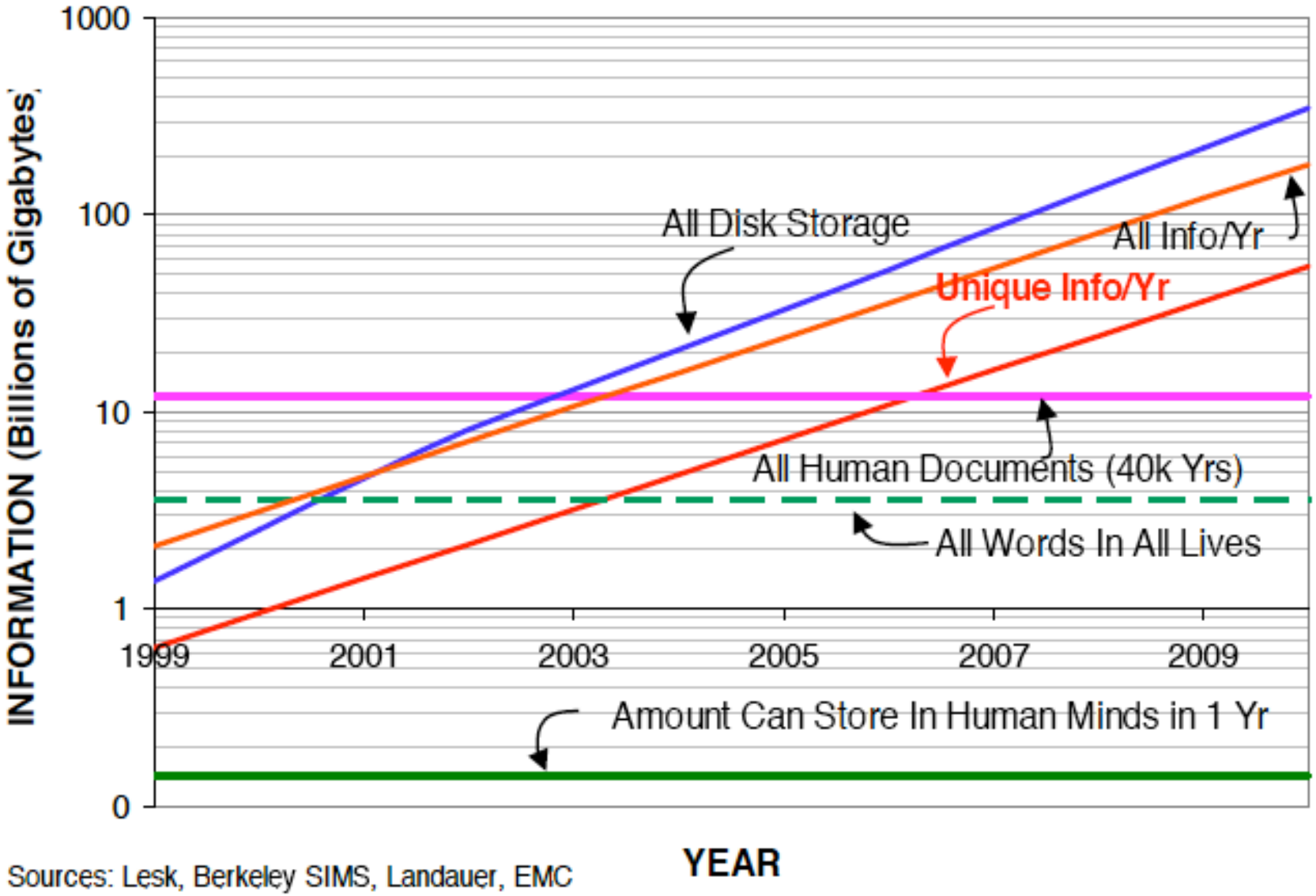
Associated Names

- Buccleuch, Henry, 3rd Duke of
- Buccleuch, John Charles, 7th Duke of
- Colnaghi & Co., Ltd., P. & D.
- Knoedler & Company, M.
- Mellon, Andrew W.

Science Today: Data Overload

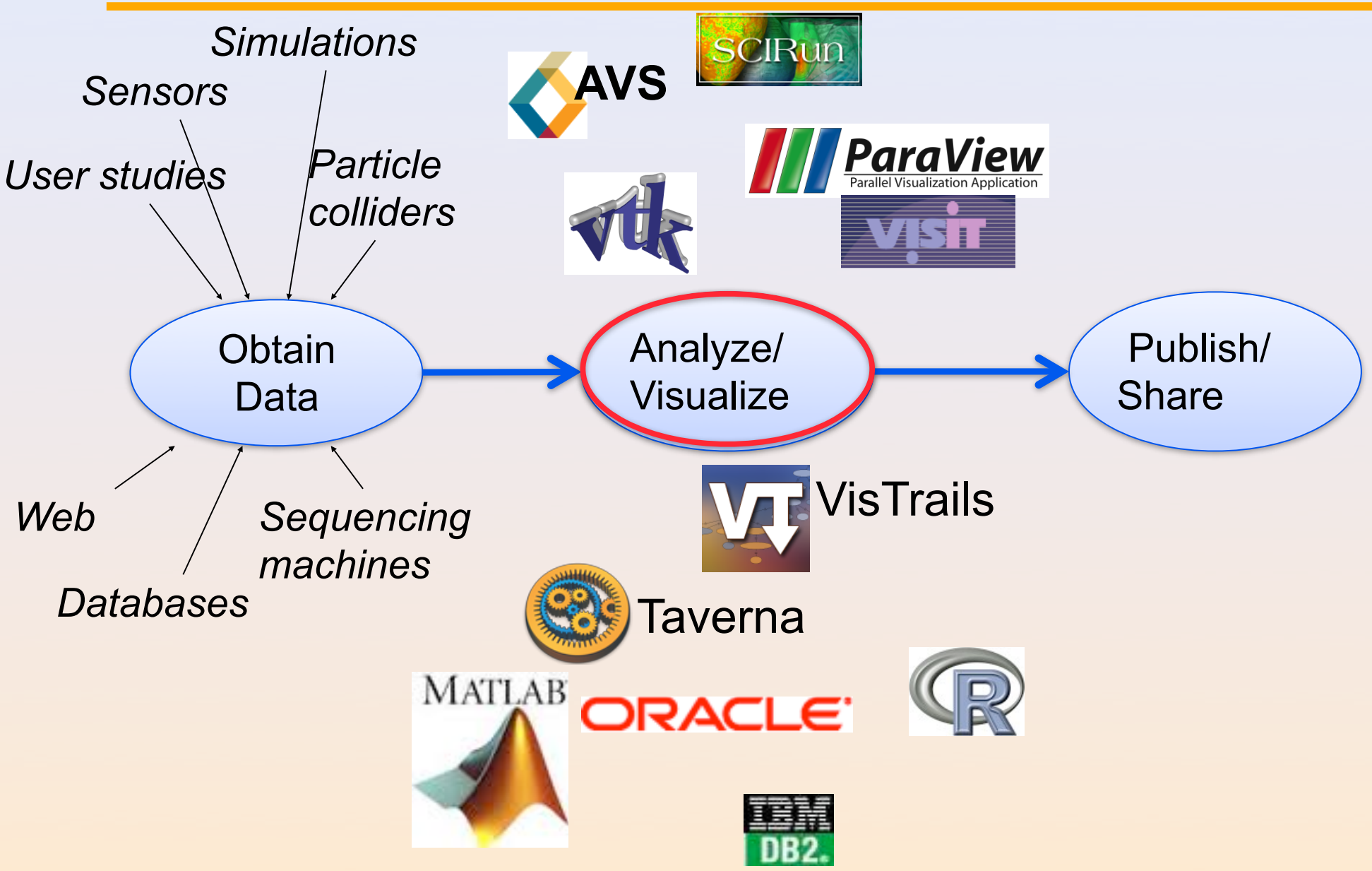


Information Big Bang

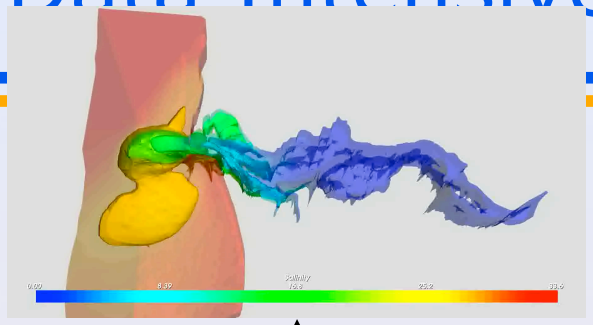


Sources: Lesk, Berkeley SIMS, Landauer, EMC

Science Today: Data Intensive



Science Today: Data Intensive



Simulations

Sensors

User studies

Particle colliders

Microsoft PowerPoint

Obtain Data

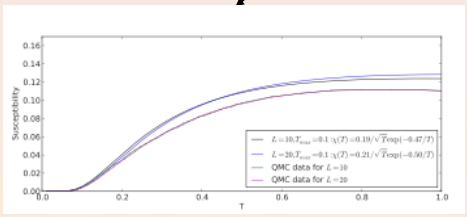
Analyze/ Visualize

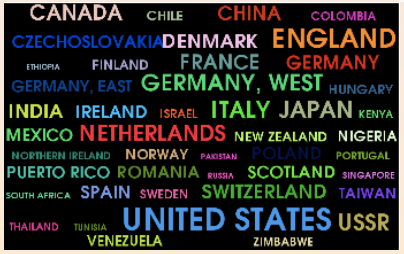
Publish/ Share

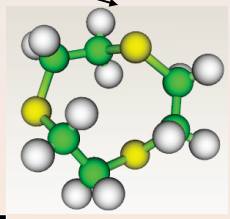
Web

Sequencing machines

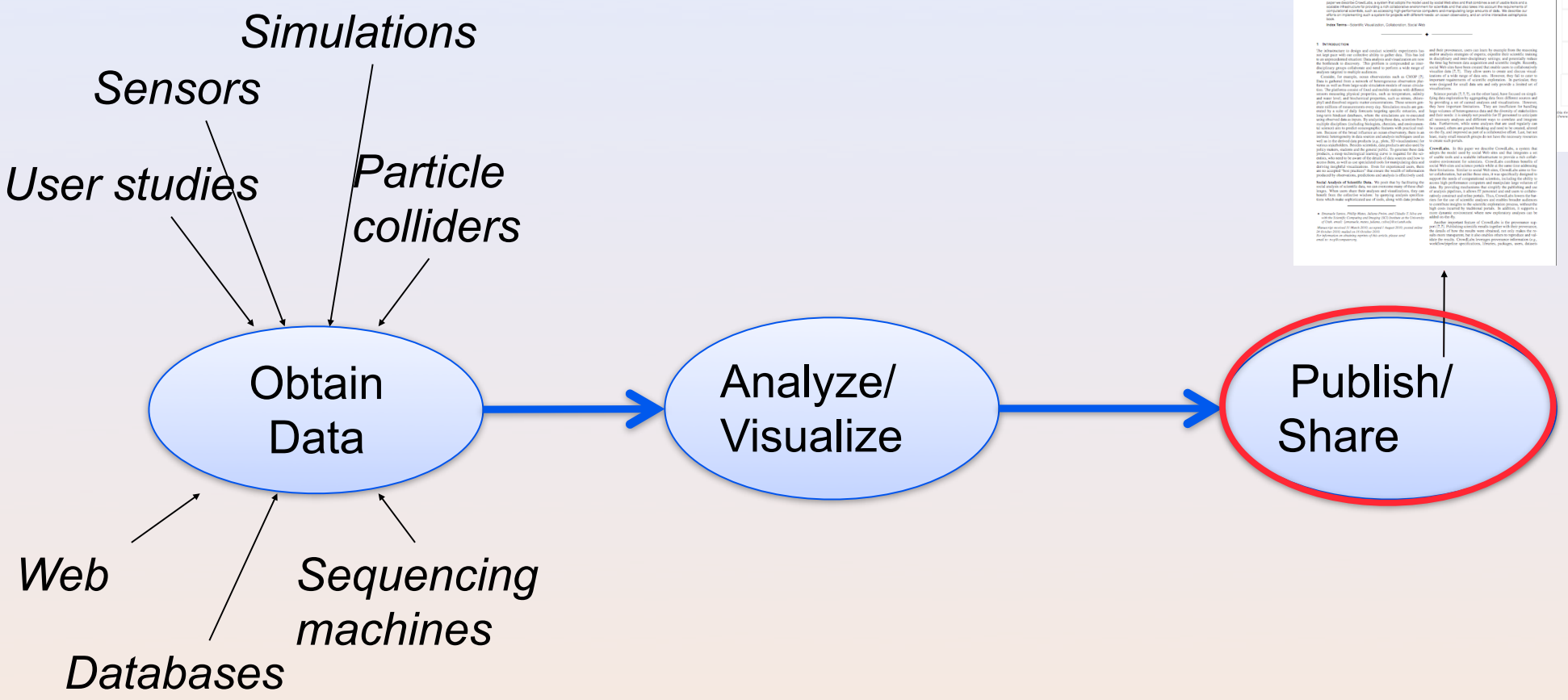
Databases







Science Today: Data Intensive




CrowdLabs: Social Analysis and Visualization for the Sciences

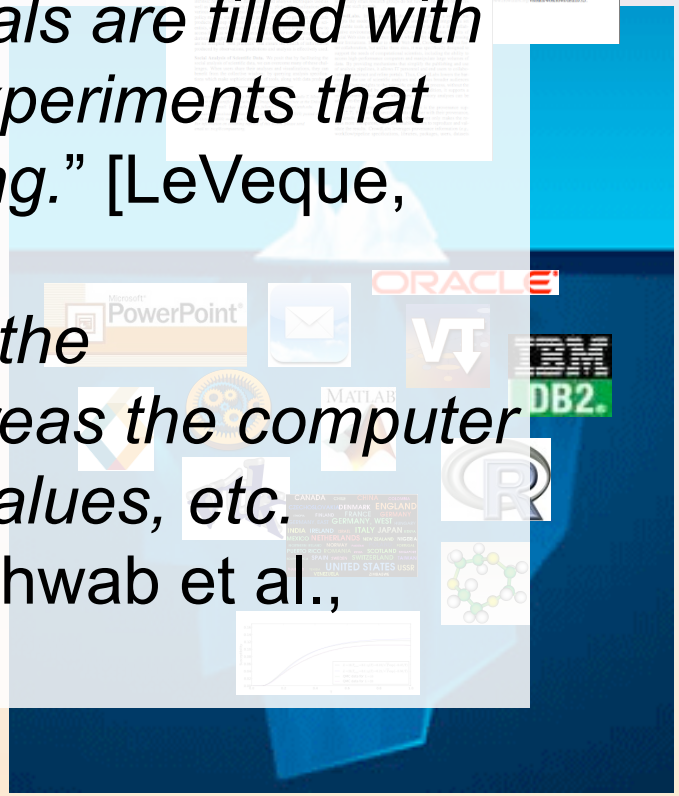
Science Today: Incomplete Publications

- 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]



Visualization algorithms

On Histograms and Isosurface Statistics

Hamish Carr* Brian Duffy Barry Denby

University College Dublin

Abstract

In this paper, we show that histograms represent spatial function distributions with a nearest neighbour interpolation, resulting in systematic underrepresentation of transitional features of the data, and that isosurface statistics, which use higher quality interpolation, give better representations of the function distribution. We also use our experimentally collected isosurface statistics to resolve some questions as to the formal complexity of isosurfaces.

Keywords: histograms, isosurfaces, isosurface statistics

1 Introduction

Scientific and medical visualization commonly represents physical phenomena as continuous functions sampled over a spatially defined domain, then reconstructed by interpolation, either using a filter kernel or a geometric mesh. In either case, the interpolation method implicitly applies knowledge derived from the spatial relationships between the sample points to determine the function value at previously unsampled points. This interpolation method is then applied when visualizing the data.

Independent of the method is chosen for visualizing spatial data, the fundamental task is often to identify the important function values. For this task, information about the importance of individual isovalues is crucial, as isovalues are one of the most significant inputs to rendering methods. In turn, importance is usually measured using statistics of the sample points, most commonly the well-known *histogram*. This technique visualizes the distribution of the samples over the function's *range* by binning the samples and displaying a bar graph of the number of samples in each bin.

Histogram computation, however, assumes independent samples with no inherent relationships, while spatial sampling assumes that samples where close spatial relationships imply functional relationships. In Section 3, we show that the histogram is formally equivalent to the *nearest neighbour* interpolant, which is widely recognized as the worst possible interpolant that can be chosen. We believe that this relationship has gone unnoticed principally because of the ubiquity of the histogram as a statistical tool, and because the histogram was developed in contexts such as population statistics, where no meaningful spatial interpolant exists.

In Section 4, we show how to remedy the defects of the histogram by substituting isosurface statistics, which give substantially better representation at little additional cost. We demonstrate these defects, and the superior quality of isosurface statistics, by comparing histograms with isosurface statistics for a variety of real data sets made available on the internet.

In Section 7, we use our isosurface statistics to confirm some existing estimates of isosurface complexity. In particular, we show that Itoh & Koyamada's estimate [8] of $O(N^{2/3})$ is an underestimate, and that the observed relationship is closer to $O(N^{0.82})$. We also confirm a previous observation [4] that Marching Cubes generate an average of 2.05 triangles per cube.

*e-mail: hamish.carr@ucd.ie

2 Previous Work

Histograms are one of the oldest techniques known for displaying statistics [7]: the term was coined in 1892, but the technique may have been used earlier. Fundamentally, the histogram is a bar graph used to represent the distribution of function values in a population. In this bar graph, the independent variable represents the possible values of a set of observations, and the dependent variable represents the number of observations with a given value.

In computer graphics, image processing and visualization, histograms are often computed for the pixel (or voxel) values in a sampled data set. This information is then used for histogram equalization [6], manual transfer function construction [6, 9] and automated detection of significant isovalues [16, 17].

As we will see in the next section, histograms assume nearest neighbour interpolation, and it is interesting to observe that some authors have computed statistics in this way, while others have computed isosurface statistics based on specific interpolants. None, however, appears to have considered the relationship between the interpolant assumed for visualization, and the interpolant assumed for assessing function distributions.

Bajaj, Pascucci & Schikore [1] displayed isosurface statistics and topology in their *contour spectrum* to give users cues to interesting isovalues. These isosurface statistics were computed precisely in piecewise polynomial form based on an assumed linear barycentric interpolant over simplicial cells.

Carr, Snoeyink & van de Panne [3] extended these statistical computations to individual contours but used discrete approximations of their statistics by counting and summing individual sample values: this can be viewed as using nearest neighbour interpolation for their statistics, i.e. using histograms.

Pekar, Wiemker & Hempel [16] suggested using discrete isosurface statistics to supplement histograms for detecting significant isovalues. In addition, these authors showed how to compute the Laplacian-weighted histogram to find isovalues at which significant boundary effects occurred. Given the discrete nature of the Laplacian computation described, the effect of this is to convolve the image with a small interpolation kernel, then compute the histogram. Again, this is performed on individual voxels, and neither assumes nor employs an interpolant when generating statistics.

Tenginakai, Lee & Machiraju [17] extended histograms to multi-dimensional histograms based on discrete computations of local higher-order moments, but used single isovalue bins without spatial interpolation.

Thus, although both isosurface statistics and histograms have been used to identify significant isovalues and other properties of data sets, individual authors have tended to use one or the other but not both, and the relationship between these two approaches has gone unexamined. In this paper, our principal aim is to examine this relationship, and to argue that isosurface statistics are a better representation of function distribution than histograms.

Isosurface statistics also arise in the context of complexity analysis of isosurface extraction algorithms, which typically exploit the fact that k, the output cost of rendering, is significantly less

Revisiting Histograms and Isosurface Statistics

Carlos E. Scheidegger John Schreiner Brian Duffy Hamish Carr Cláudio T. Silva

Abstract—Recent results have shown a link between geometric properties of isosurfaces and statistical properties of the underlying sampled data. However, this has two defects: not all of the properties described converge to the same solution, and the statistics computed are not always invariant under isosurface-preserving transformations. We apply Federer's Coarea Formula from geometric measure theory to explain these discrepancies. We describe an improved substitute for histograms based on weighting with the inverse gradient magnitude, develop a statistical model that is invariant under isosurface-preserving transformations, and argue that this provides a consistent method for algorithm evaluation across multiple datasets based on histogram equalization. We use our corrected formulation to reevaluate recent results on average isosurface complexity, and show evidence that noise is one cause of the discrepancy between the expected figure and the observed one.

1 Introduction

In scientific and medical visualization, we commonly represent physical quantities as continuous functions defined over a continuous domain. These are constructed by resampling observed points using some reconstruction kernel defined on the underlying space or on the connectivity of a geometric grid. When visualizing data, we work directly on the continuous function, which is mathematically and computationally convenient.

Since humans are not good at assimilating large quantities of numerical data, visualization seeks to map numerical properties of this continuous function to visual properties such as colour, brightness and saturation or to geometric properties such as boundaries and edges. Thus, one of the first steps in visualization is to define a mapping from the function to visual properties. Defining this mapping often depends on understanding the frequency and possibly the spatial distribution of the numerical values.

Historically, function distributions have been computed with *histograms*, which simply count the number of samples with each function value. However, recent work by Carr et al. [3] has shown that there are serious problems in using histograms as representations of function distributions. In particular, histogram computation assumes that the reconstruction uses a *box filter* (or *nearest neighbor interpolation*).

Using this observation, Carr et al. [3] proposes several alternative statistics that converge faster than histograms. These are based on interpreting the *isosurface areas* as measurements of higher-order interpolants on the domain. In addition, the authors related their statistics (and, implicitly, the histogram) to the algorithmic complexity of isosurface rendering, demonstrating a larger ($O(N^{0.82})$) experimental result than the $O(N^{2/3})$ previously predicted [7]. While the proposed statistics certainly converge faster than histograms, two problems can be identified. First, these converge to a slightly different result than histograms. Second, the mathematics suggest some counter-intuitive results about the average complexity of isosurfaces in the domain. Here, we use the term "convergence" to mean that as we use increasingly finer grids, the computed functions approach some limit function.

In this paper, we address these issues by revisiting the development of the isosurface statistics using a celebrated result in geometric measure theory: Federer's Coarea Formula [12]. This formula allows one to relate integrals over level sets of a function to integrals over the domain on which the function is defined. Crucially, it has the effect of normalizing isosurface statistics to the packing density of the isosurfaces, allowing us to correct the problems identified and propose improved solutions. Moreover, this formula gives us additional results. In particular, we show that in a particular case, histogram equalization can be used as a normalization to get the $O(N^{2/3})$ result originally predicted. Throughout the paper, we use the term "level set" mostly interchangeably with isosurface, but will choose the term "level set" when we want to disregard the particular isovalue associated to the set of points in the preimage.

Our contributions are as follows. We build on the work of Carr et al. [3], and introduce a more mathematically grounded approach based on Federer's Coarea Formula (FCF). It clarifies the subtle relation between histograms and isosurface areas, and, crucially, shows the role of gradient magnitudes in that relationship. It also provides a well-founded way to compute expectations over all isosurfaces in a volume. Practically, we suggest using the cell span as an approximation for gradient magnitude, and provide experimental evidence for the validity of this approximation. Finally, we revisit Carr et al.'s study on average isosurface complexity and study the effect of volume noise in those estimates.

This paper is organized as follows. Section 2 briefly reviews relevant previous work, while Section 3 summarizes the contents of the previous result [3]. We then introduce the FCF in Section 4, and use this mathematical tool to develop an improved formulation of isosurface statistics in Section 5, based on dividing the statistics by the local gradient magnitude of the continuous function. In Section 6, we extend the application of the FCF to the computation of average isosurface complexity, and show that there exist transformations that preserve level sets, but change the original isosurface statistics (we will make this notion precise). We also show how to compute the average complexity in a way that implicitly accounts for any such transformations, and explain the discrepancy in the original results. In Section 7 we confirm that the average isosurface complexity of a function, when sampled increasingly finely in a domain, is $O(N^{2/3})$. We also revisit the experiments of Carr et al. [3] in light of the corrected integral formulation. In our experiments, we find an even higher average complexity, of $O(N^{0.96})$. Finally, we show that noise seems to largely explain this high figure, by performing a set of experiments on synthetic data. We then summarize our results in Section 8 and speculate on future directions of research in Section 9.

- *Carlos E. Scheidegger, John Schreiner, and Cláudio T. Silva are with the Scientific Computing and Imaging Institute, University of Utah. E-mail: {cscheid, jmschrei, csilva}@sci.utah.edu.*
- *Brian Duffy and Hamish Carr are with the UCD School of Computer Science & Informatics, University College Dublin, Ireland. E-mail: {B.Duffy, hamish.carr}@ucd.ie.*

Manuscript received 31 March 2008; accepted 1 August 2008; posted online 19 October 2008; mailed on 13 October 2008.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

2 Previous Work

Histograms are ubiquitous in plotting, and also in computer graphics and visualization. It is one of the oldest techniques available for displaying data [6], and is often used as the basis of techniques such as histogram equalization [5], which defines a non-linear transfer func-

Visualization algorithms

Visualization Viewpoints
Editor: Theresa-Marie Rhyne

The Need for Verifiable Visualization

Robert M. Kirby and Cláudio T. Silva
University of Utah

The use of simulation science as a means of scientific inquiry is increasing at a tremendous rate. The process of mathematically modeling physical phenomena, experimentally estimating important key modeling parameters, numerically approximating the solution of the mathematical model, and computationally solving the resulting algorithm has inundated the scientific and engineering worlds. As increasingly more science and engineering practitioners advocate the use of computer simulation for analyzing and predicting physical and biological phenomena, the computational science and engineering community has started asking introspective questions, such as[1]

- Can computer-based predictions be used as a reliable basis for making crucial decisions?
- How can you assess a computer-based prediction's accuracy or validity?
- What confidence (or error measures) can be assigned to a computer-based prediction of a complex event?

Those researchers outside traditional computational engineering and science areas (traditional areas such as computational fluid dynamics [CFD] and computational solid mechanics [CSM]) are sometimes shocked to hear these questions being asked, as they often have assumed that these types of issues had been settled long ago—at the inception of computing and computational modeling. A study of the computational science and engineering (CS&E) literature from the past 40 years clearly shows that these questions have not been ignored.

Scientists who employ computing for solving problems have always been concerned with accuracy, reliability, and robustness. It was not until the past 10 years, however, that the CS&E community has joined together in an attempt to generate a unified perspective from which to evaluate these questions. The consequence of these efforts has led to what some are calling a new CS&E discipline—validation and verification, or V&V, which seeks to articulate processes by which we can obtain answers to these questions. Let us take a closer look.

Validation and verification

Figure 1 shows the common simulation science pipeline consisting of the physical phenomena of interest, mathematical modeling of the phenomena, simulation, and evaluation (often through a combination of postprocessing and visualization). It also identifies where validation and verification fit into this process (we will further explain these terms later in the article).

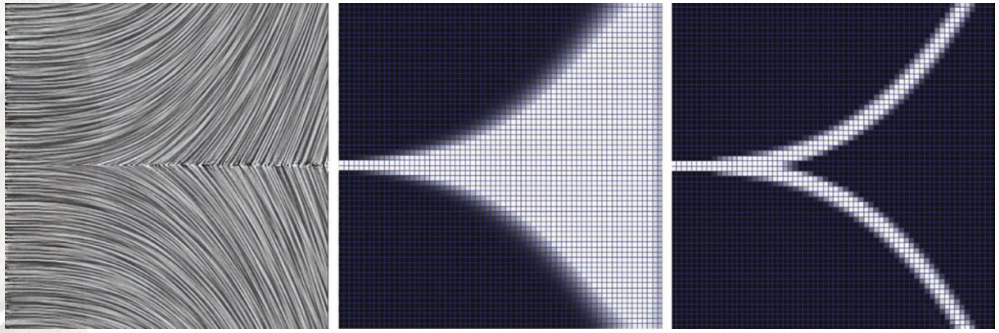
Scientists frequently use visualization techniques to help them assess their simulation results. Visualization is the lens through which scientists often view modeling and discretization interactions—hence, visualization itself must be explicitly considered as part of the V&V process. Simulation researchers often create their own visualization tools, claiming that they "don't trust" visualization techniques that they themselves have not implemented. CFD researchers creating visualizations of their own data joke that they are experts in the presentation of their own brand of CFD: colorful faulty dynamics. Such a statement can only be truly understood in the light of the V&V process; it is the means by which simulation scientists gain confidence in their algorithms and implementations as well as those by others within their community. To gain the simulation community's confidence, the visualization process must come under this process's umbrella.

Visualization techniques have lagged behind in error and uncertainty analysis of the methodology as a component of a larger scientific pipeline. Little systematic research efforts have addressed quantifying and minimizing the visualization error budget (a concept we will discuss later in the article). Furthermore, there is a real need to look at this visualization error budget in the context of the error that the rest of the computational pipeline generated and how it impacts visualization algorithms (note that this is distinct from the area of "error and uncertainty visualization," which is concerned with visualizing errors and uncertainties).

 Published by the IEEE Computer Society 0272-1716/08/$25.00 © 2008 IEEE

"The Need for Verifiable Visualization"

Example 1:



Example 2:

"Recent work details a case in which a misleading volume visualization led to unnecessary surgery for a patient."

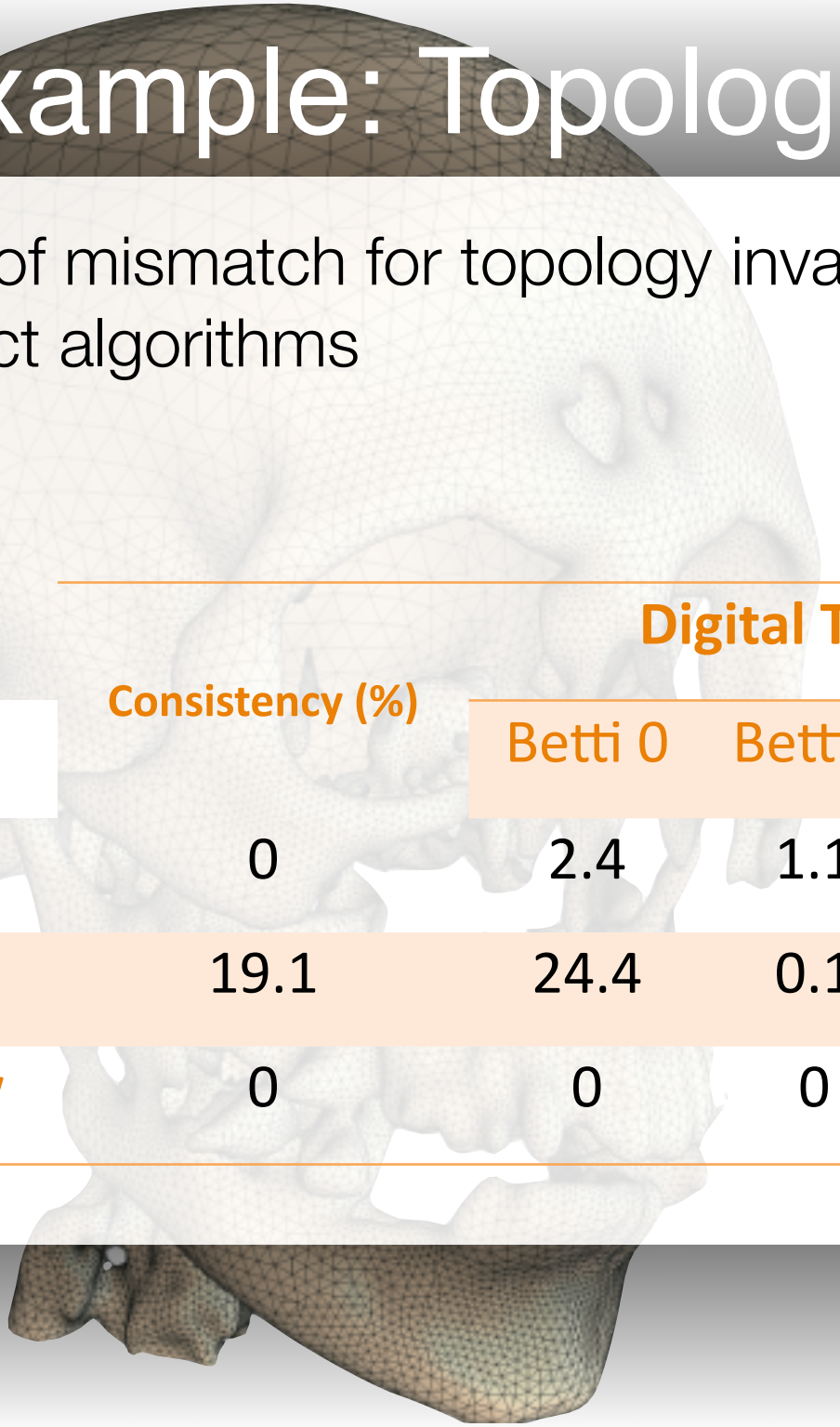
"Uncertainty Visualization in Medical Volume Rendering Using Probabilistic Animation," IEEE Transactions on Visualzation and Computer Graphics. 13(6). 1648-1655, Nov-Dec. 2007.

SCI INSTITUTE

Example: Topological Accuracy

- Rate of mismatch for topology invariants for topologically correct algorithms

	Consistency (%)	Digital Topology (%)				SMT (%)
		Betti 0	Betti 1	Betti 2	Euler	Euler
MC33	0	2.4	1.1	2.4	3.4	5.4
DelIso	19.1	24.4	0.1	20	37.2	33.2
MCFlow	0	0	0	0	0	0

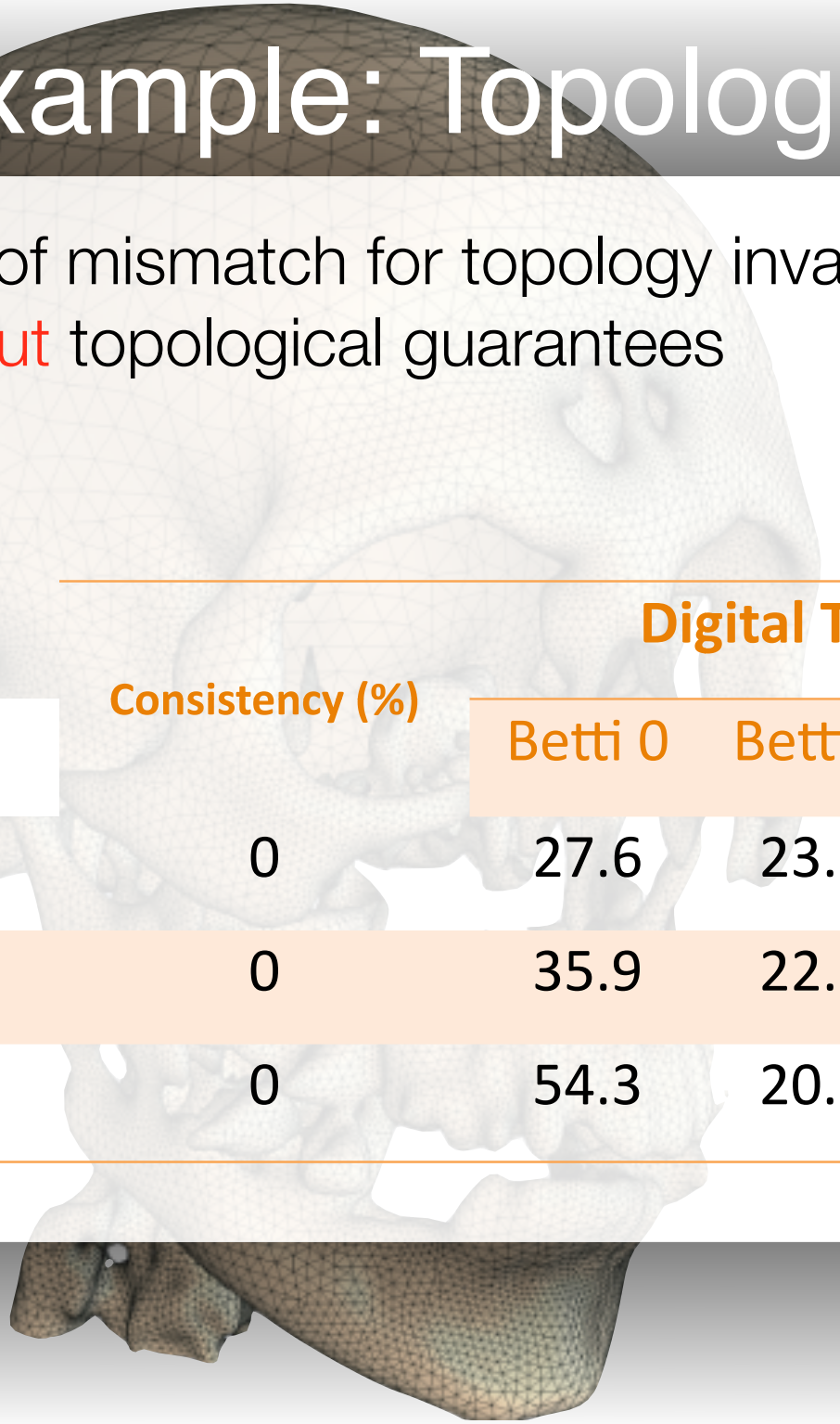


SCI INSTITUTE

Example: Topological Accuracy

- Rate of mismatch for topology invariants for algorithms without topological guarantees

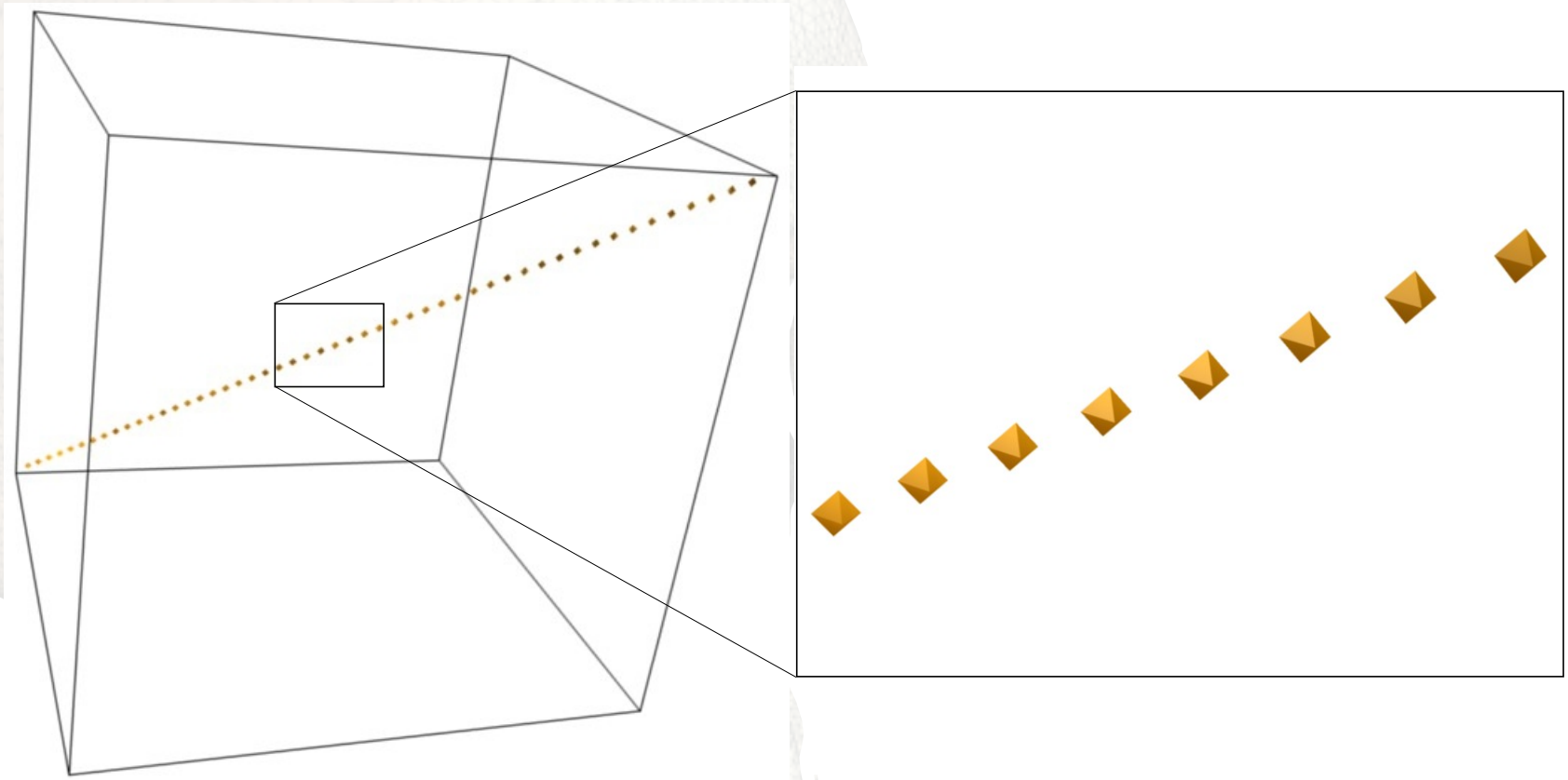
	Consistency (%)	Digital Topology (%)				SMT (%)
		Betti 0	Betti 1	Betti 2	Euler	Euler
VTKMC	0	27.6	23.2	27.6	43.5	70.7
Afront	0	35.9	22.8	35.9	47.5	25.5
Macet	0	54.3	20.9	54.3	64.0	100.0



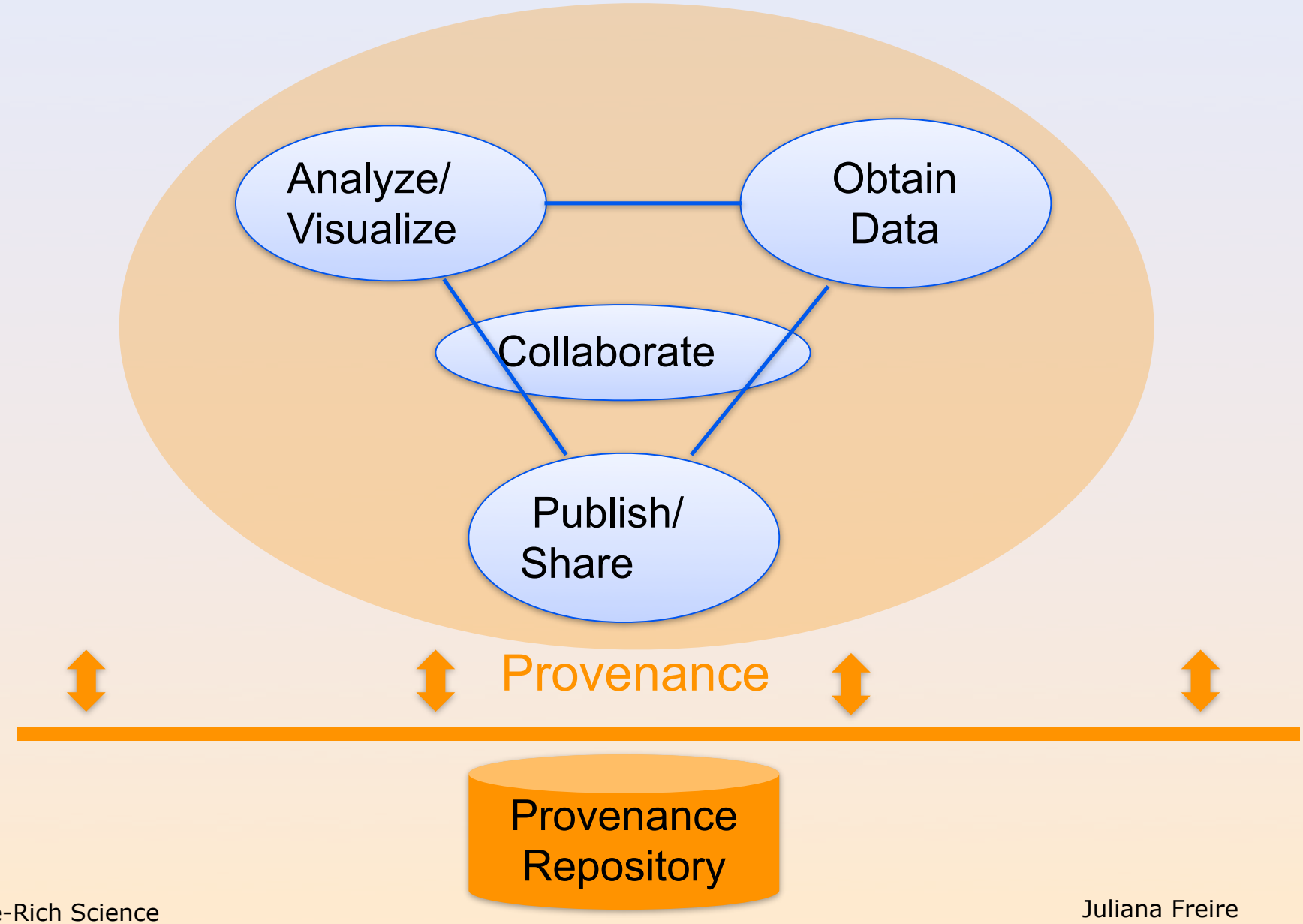
SCI INSTITUTE

Example: Topological Accuracy

- VTKMC

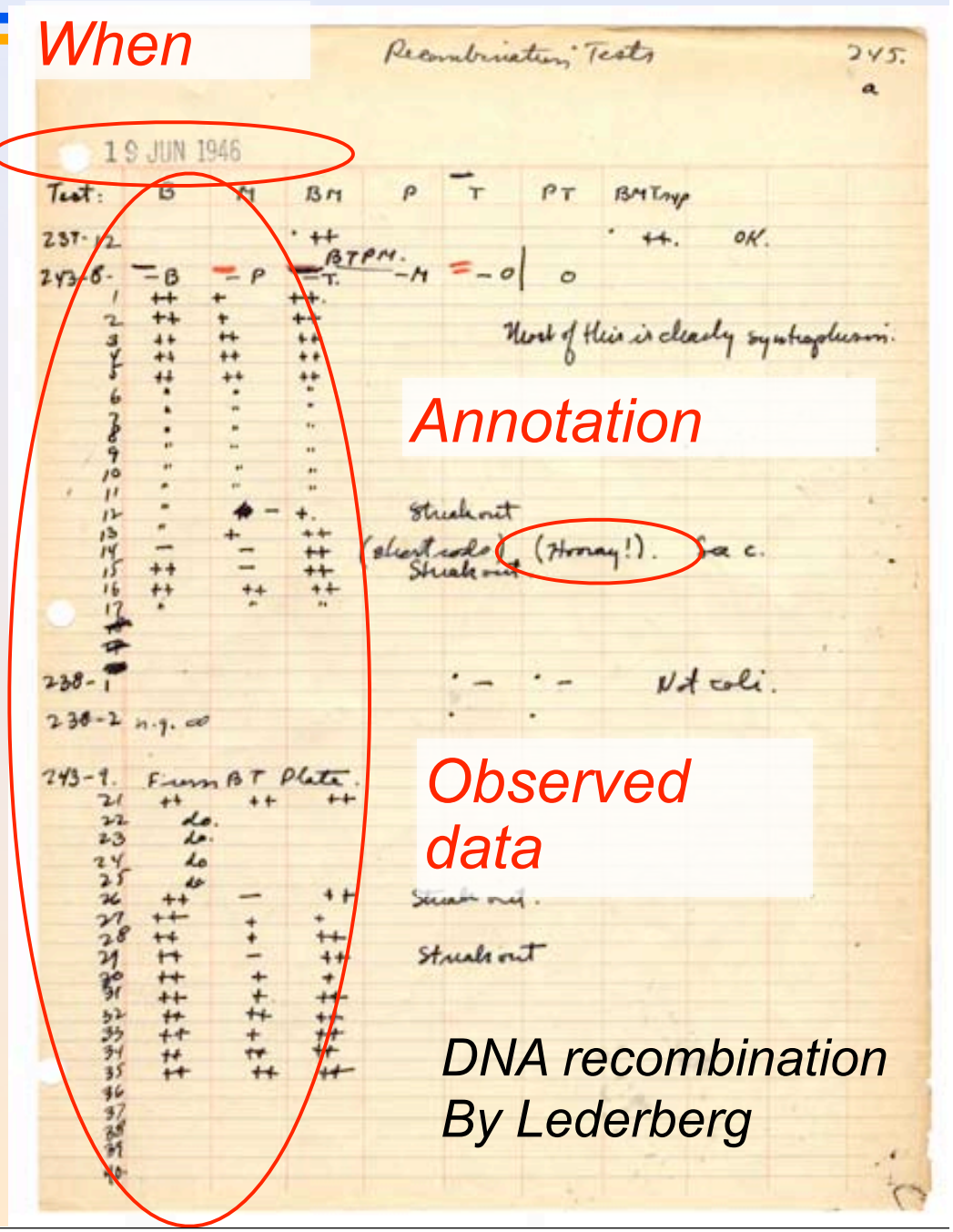


Vision: Provenance-Rich Science



Provenance in Science

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

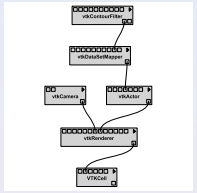


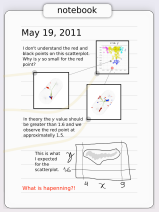
Provenance in Science

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

Provenance is as (or more!) important as the results

Vision: Provenance-Rich Science





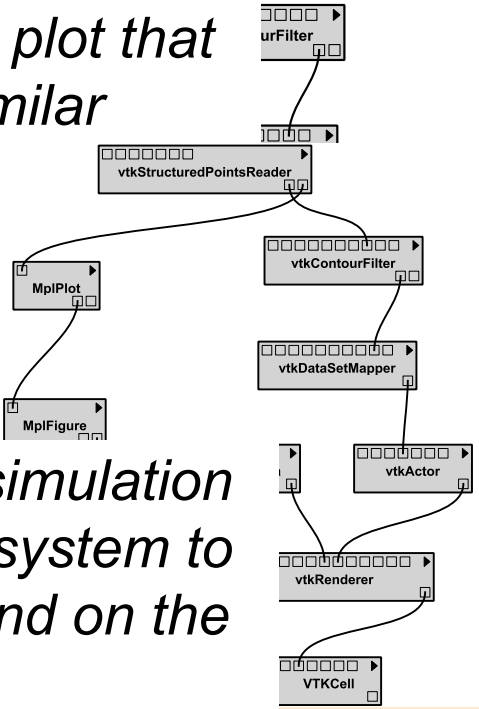
Provenance Repository

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

post-doctoral researcher

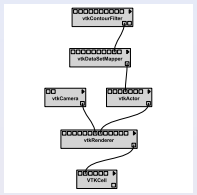
While she plots distributions from the results of a new workflow, the system analyzes previous workflows and suggests quantities to plot that were frequently used in similar

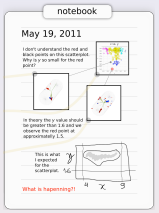
Data mining algorithms ider commonalities and differen between her analysis and tl other group members, and



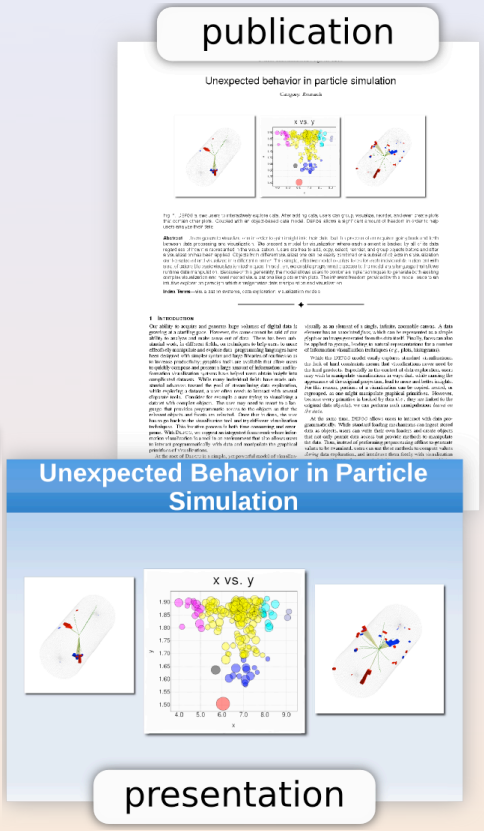
She identifies an error in the simulation code, updates it and tells the system to re-run her analyses that depend on the simulation results

Vision: Provenance-Rich Science





Provenance Repository

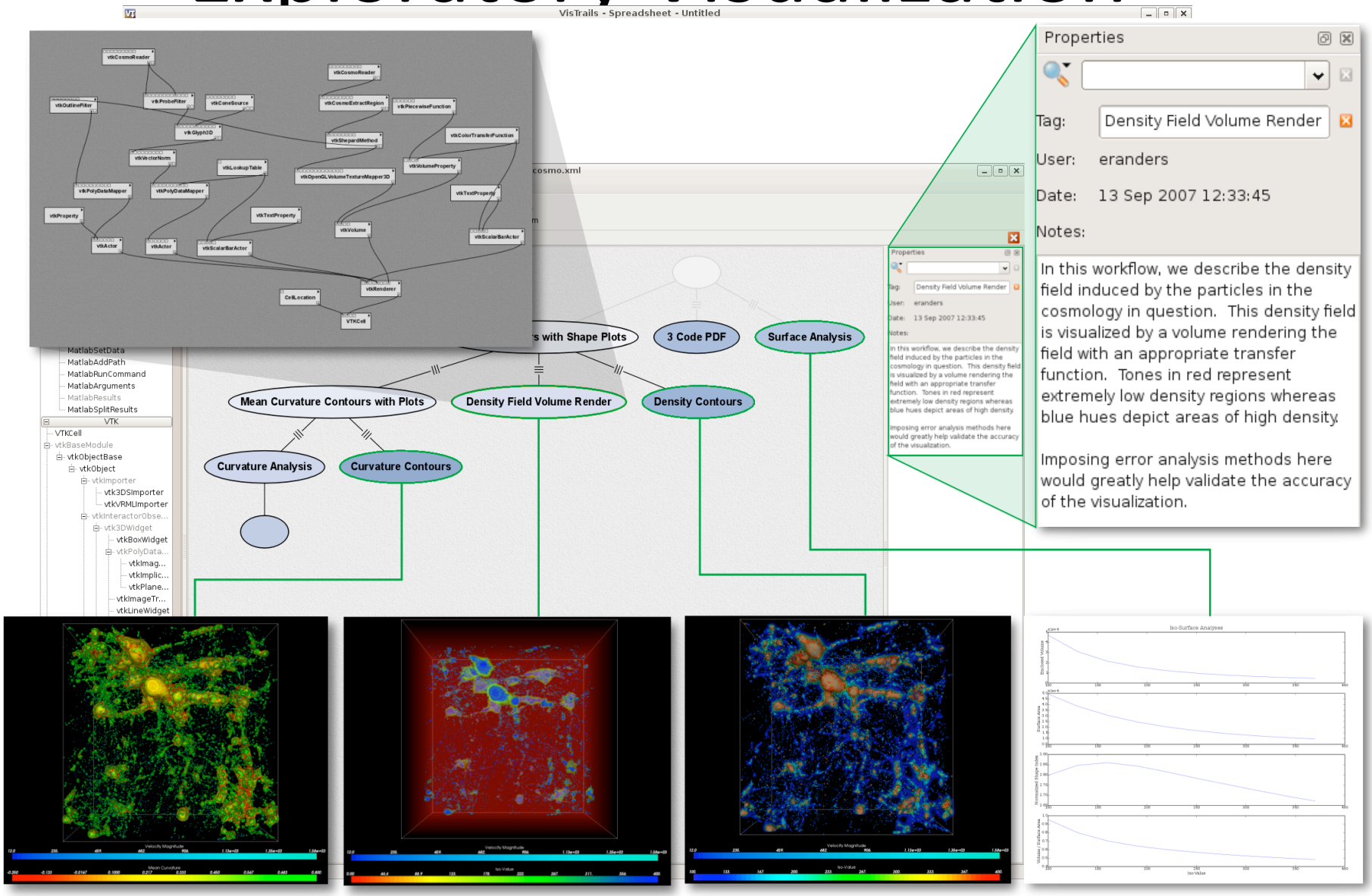


post-doctoral researcher

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

Provenance and Data Exploration

Exploratory Visualization



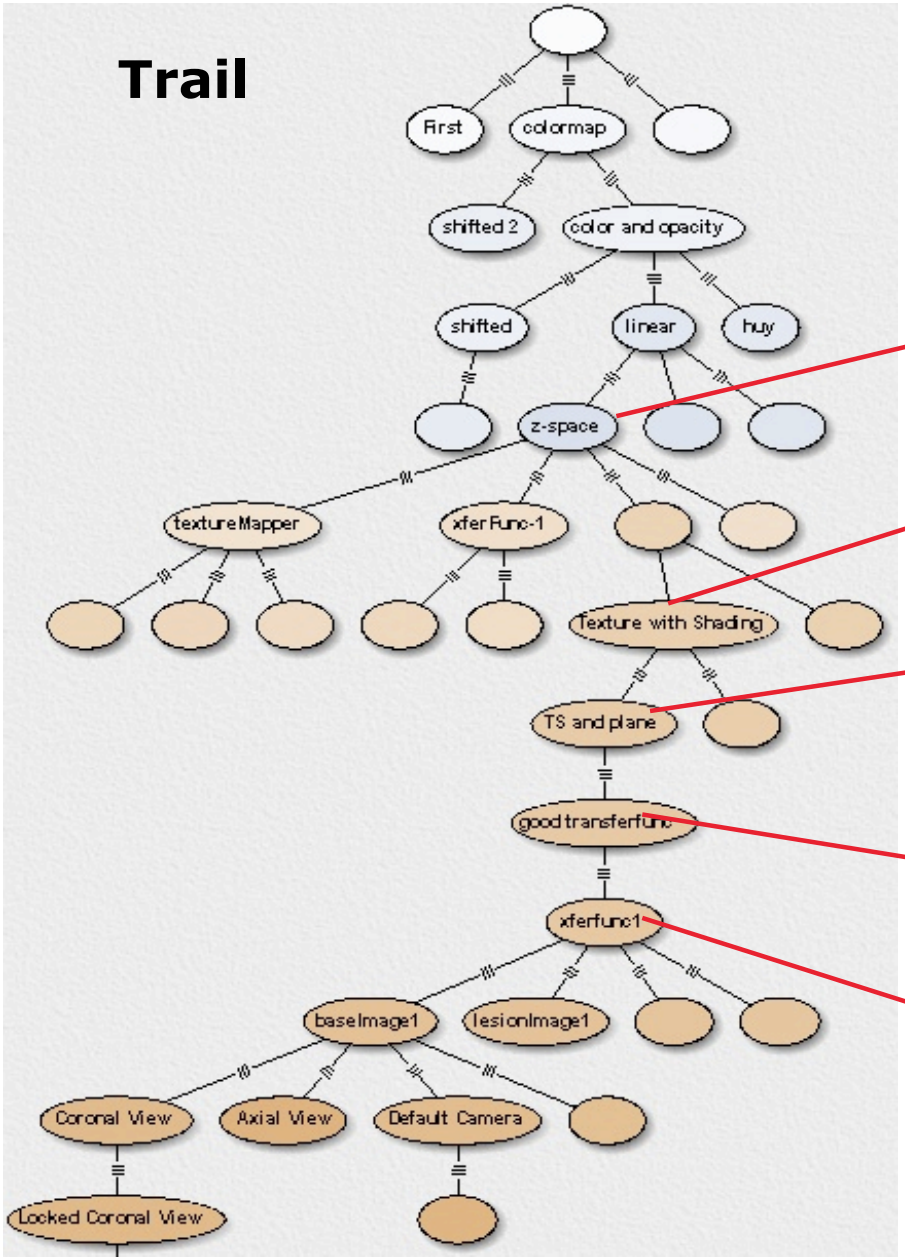
"The Cosmic Code Comparison Project," Ahrens, Anderson, Heitmann, Habib, et al

VisTrails: Managing Exploration

- Comprehensive *provenance infrastructure* for computational tasks
  - Data + workflow provenance
  - *Treat workflow as a 1st-class data product*
- Support for *exploratory* tasks such as visualization and data mining
  - Task specification iteratively refined as users generate and test hypotheses
- VisTrails manages the data, metadata and the exploration process, scientists can focus on *science!*
- Not a replacement for visualization or scientific workflow systems: infrastructure that can be combined with and enhance these systems

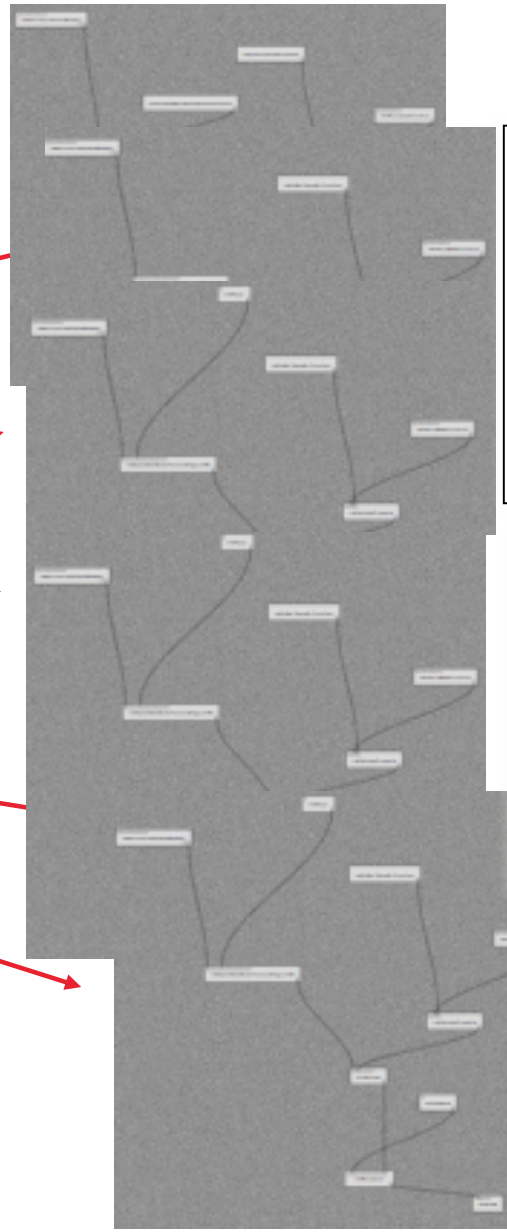
http://www.vistrails.org

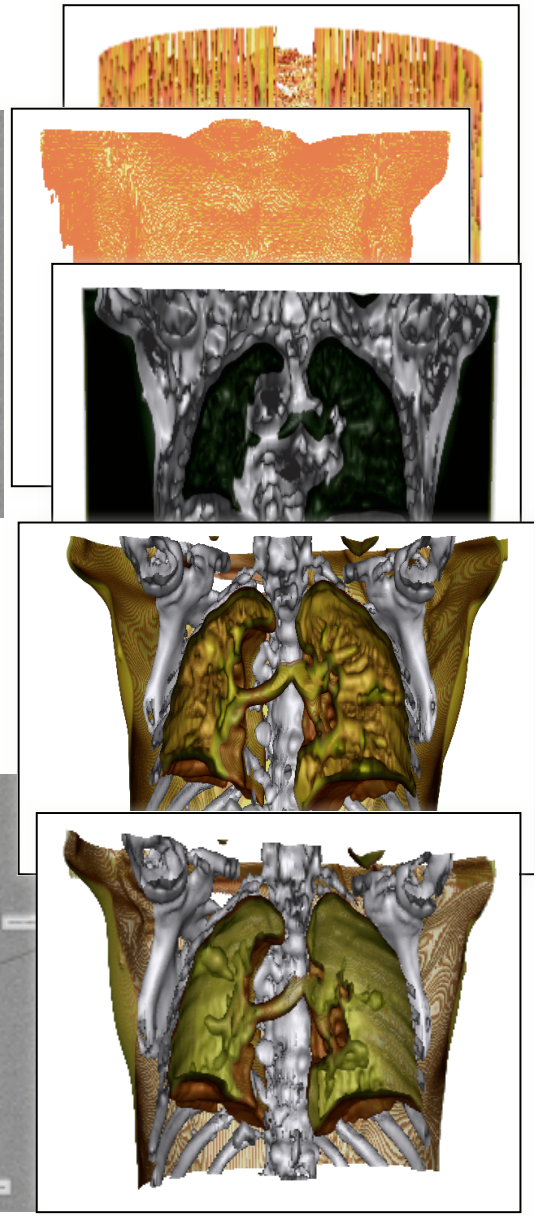
Keeping Exploration Trails



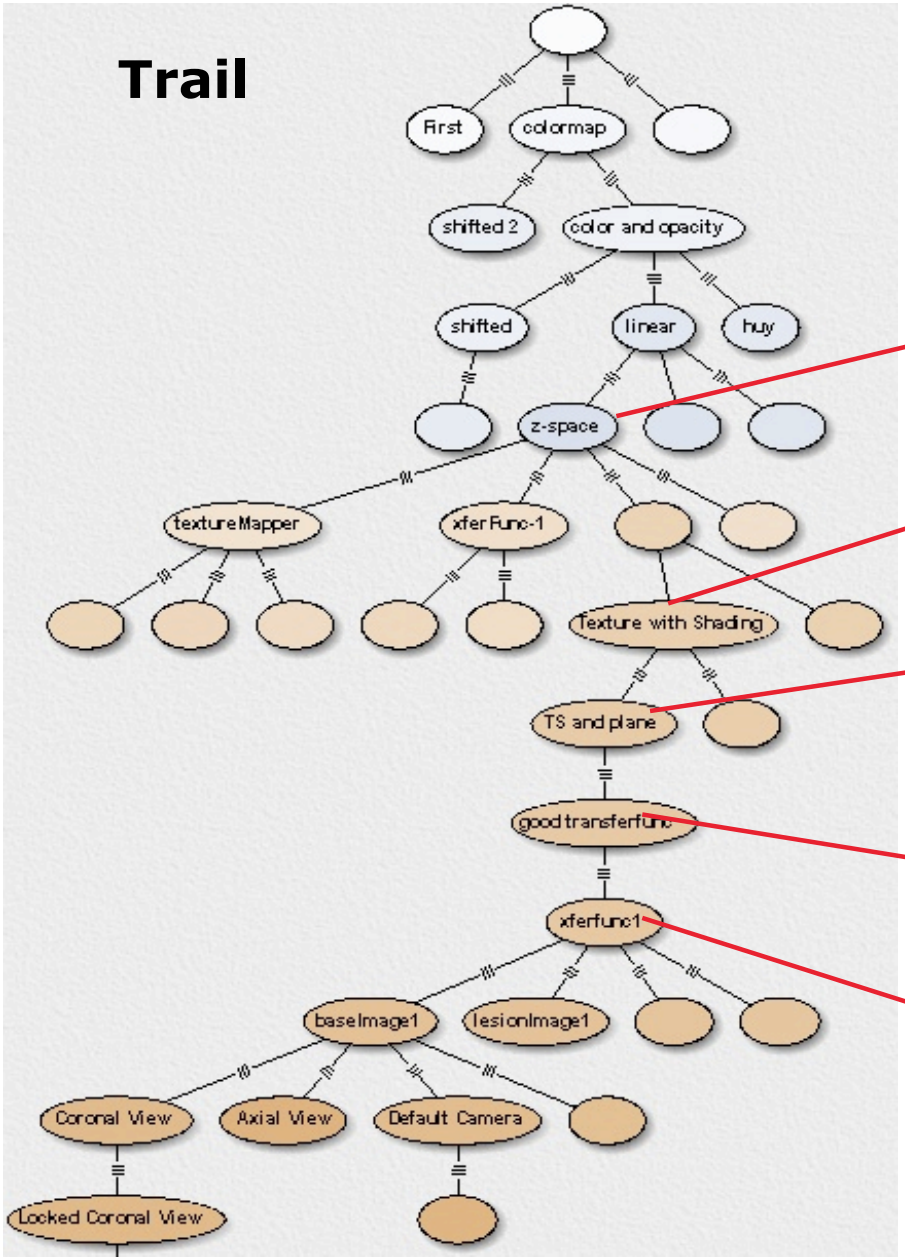
Workflows

Data Products





Keeping Exploration Trails



Notes	User
Initial visualization with z-scaling corrected	*juliana*
Added texture and shading	*eranders*
Added plane to visualize internal structure	*eranders*
Found good transfer function	*eranders*
Identified lesion tissue	*stevec*

Change-Based Provenance

- Records actions
- Provenance = changes to computational tasks
  - Add a module, add a connection, change a parameter value
- Extensible *change* algebra

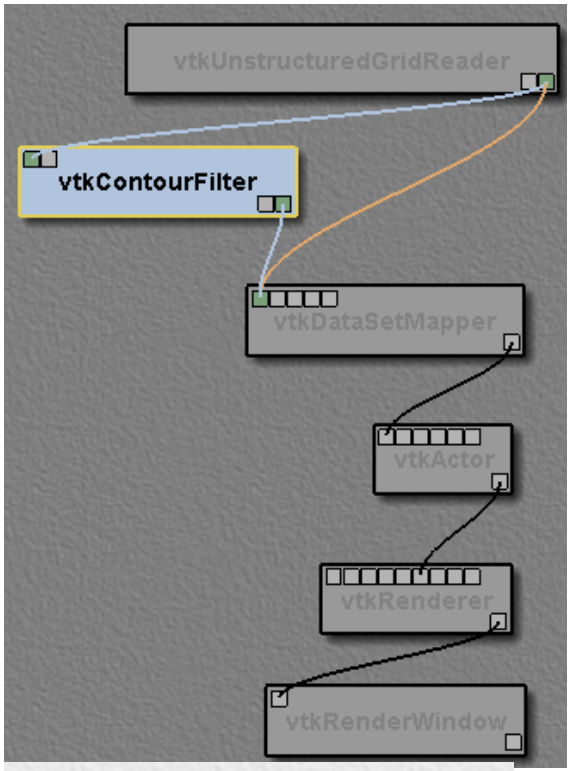
addModule

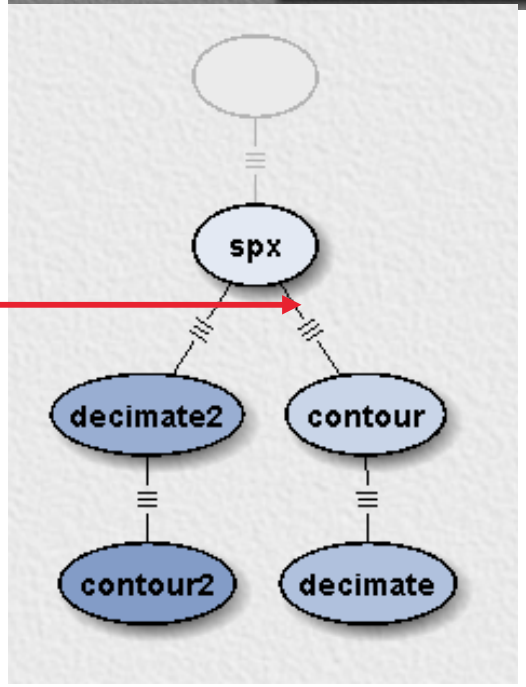
deleteConnection

addConnection

addConnection

setParameter





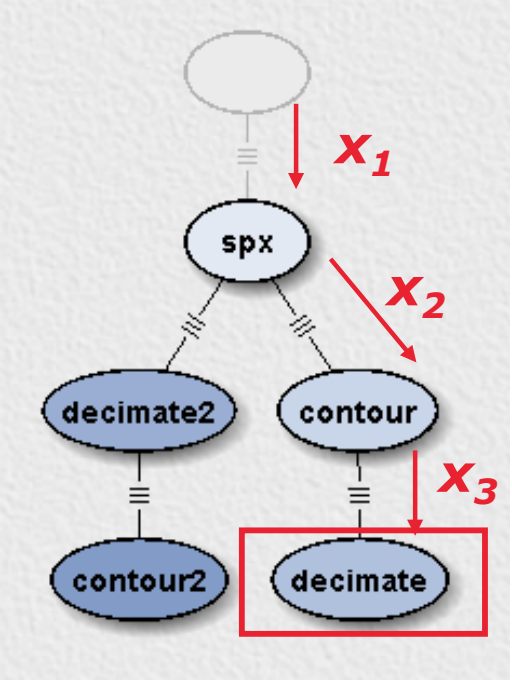
Change-Based Provenance

- Records actions
- Provenance = changes to computational tasks
  - Add a module, add a connection, change a parameter value
- Extensible *change* algebra
- A *vistrail* node v_t corresponds to the workflow that is constructed by the sequence of actions from the root to v_t

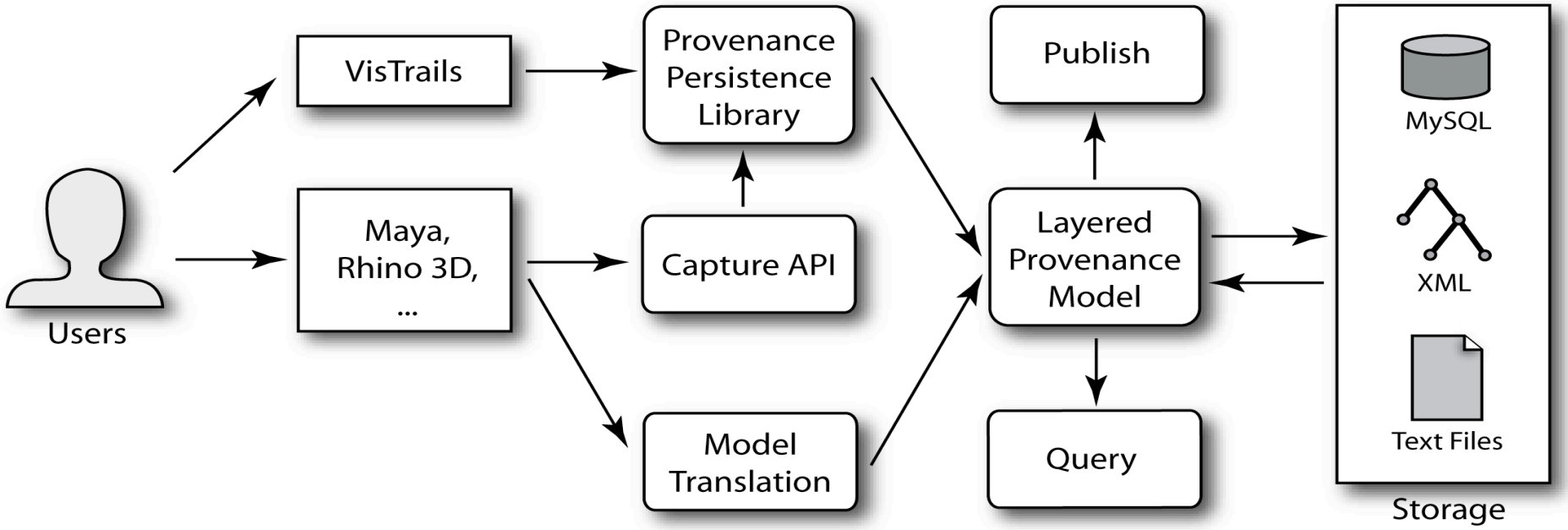
$$v_t = x_n \circ x_{n-1} \circ \ldots \circ x_1 \circ \emptyset$$

[Freire et al, IPAW 2006]

vistrail



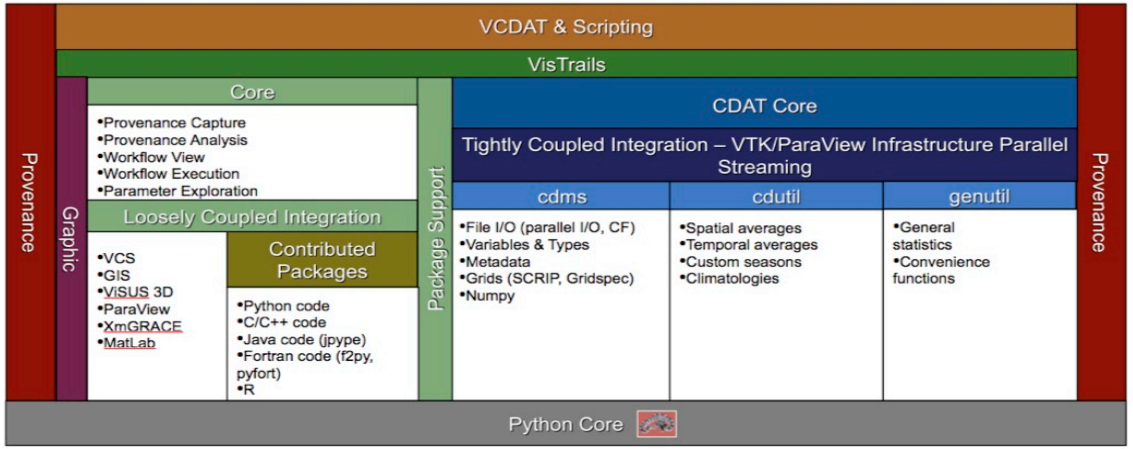
Provenance API



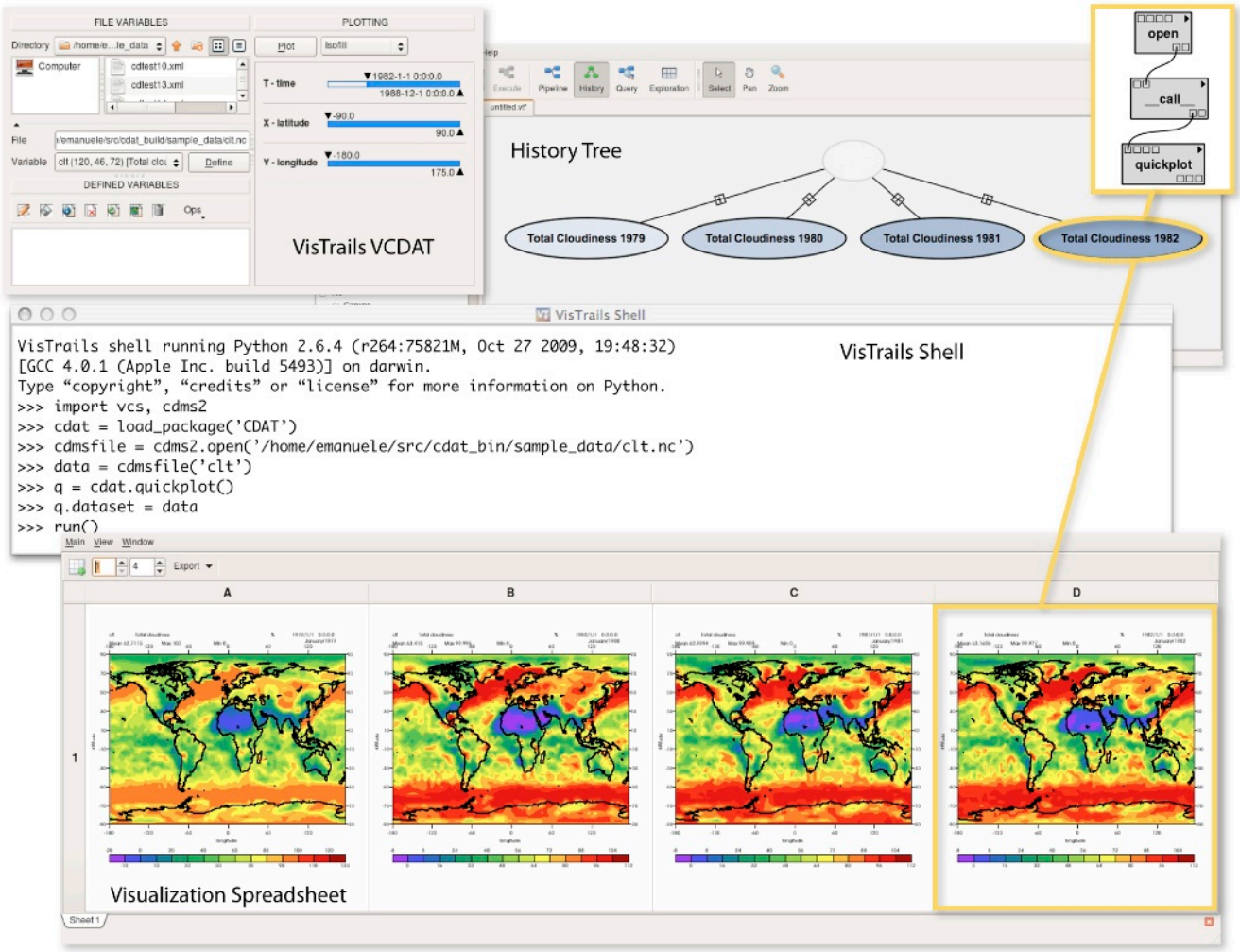
UV-CDAT

- Ultra-scale Visualization Climate Data Analysis Tools

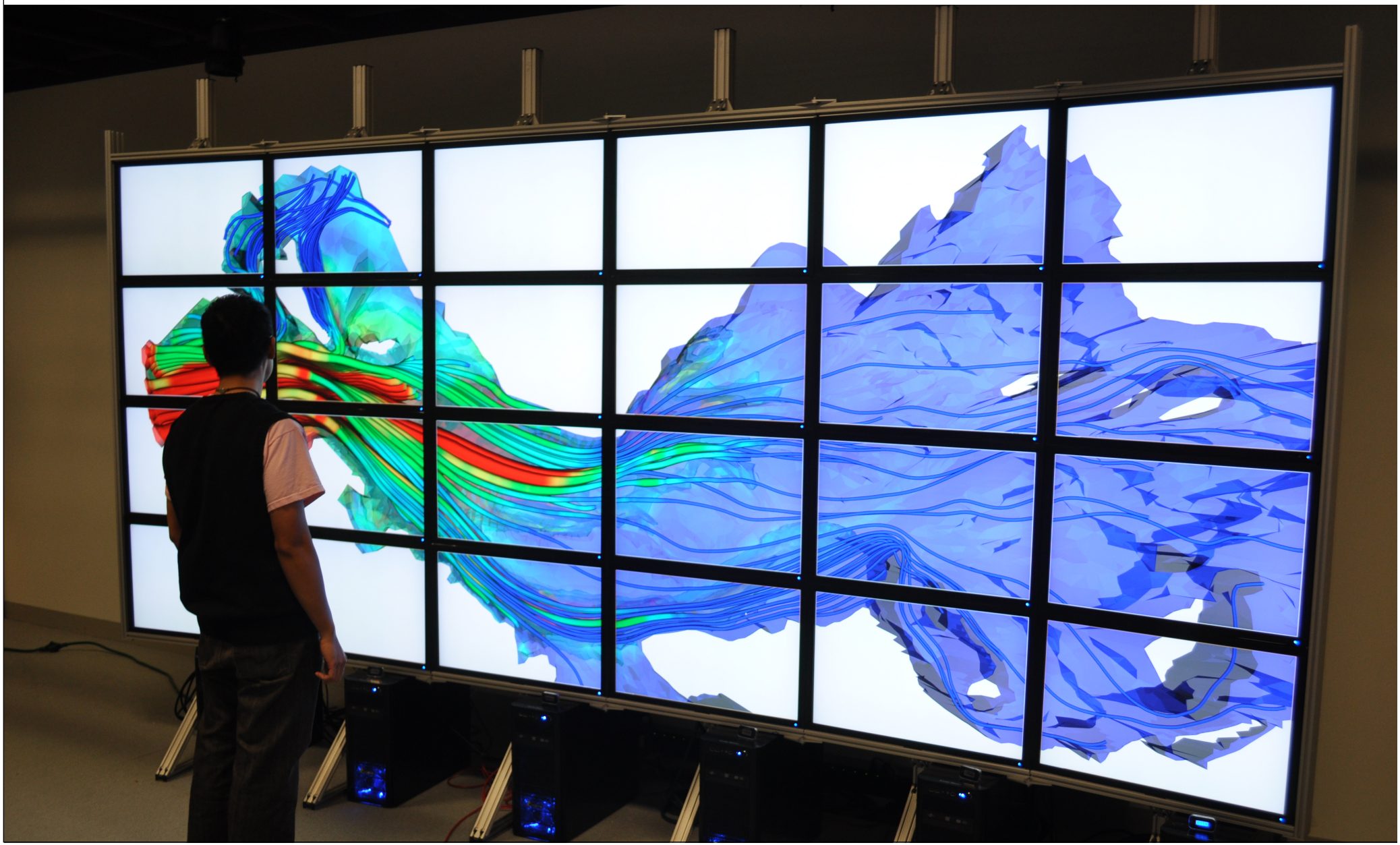
Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT) Architectural Layers

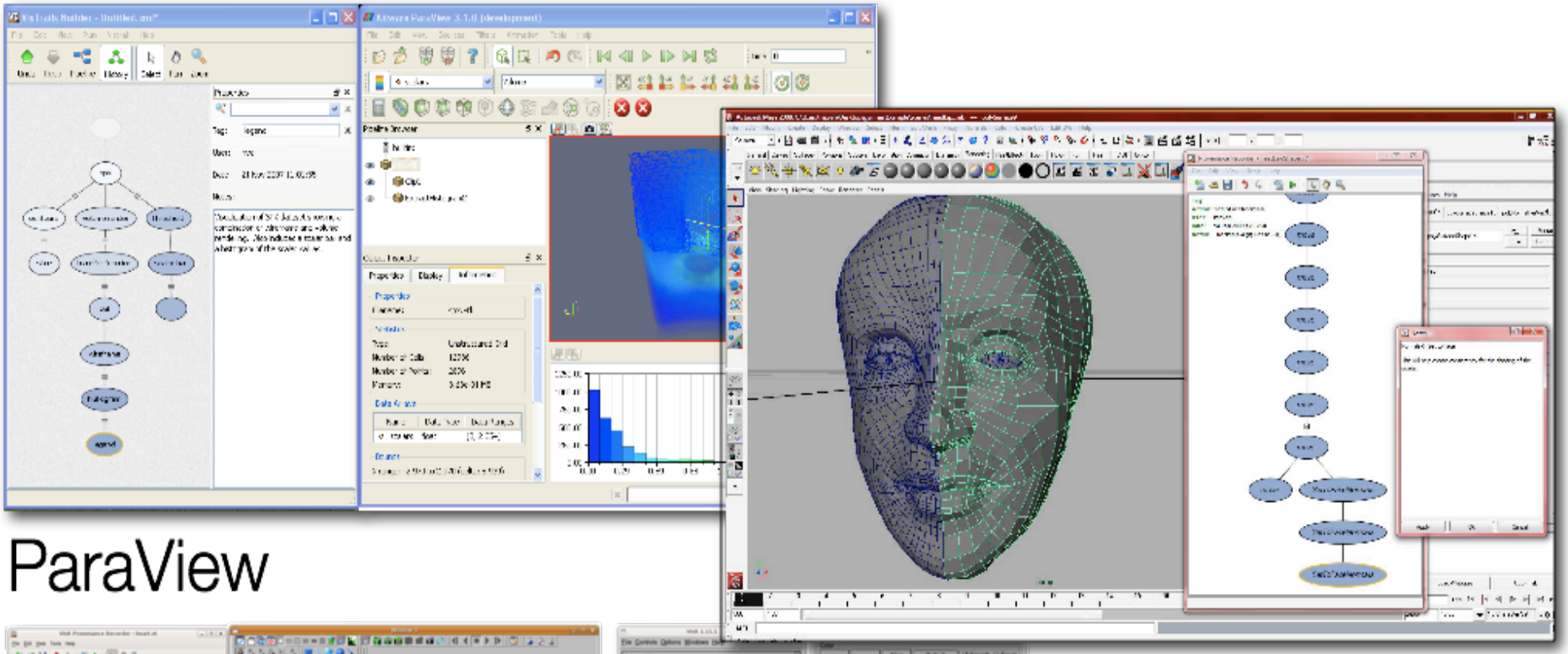


UV-CDAT (under development)



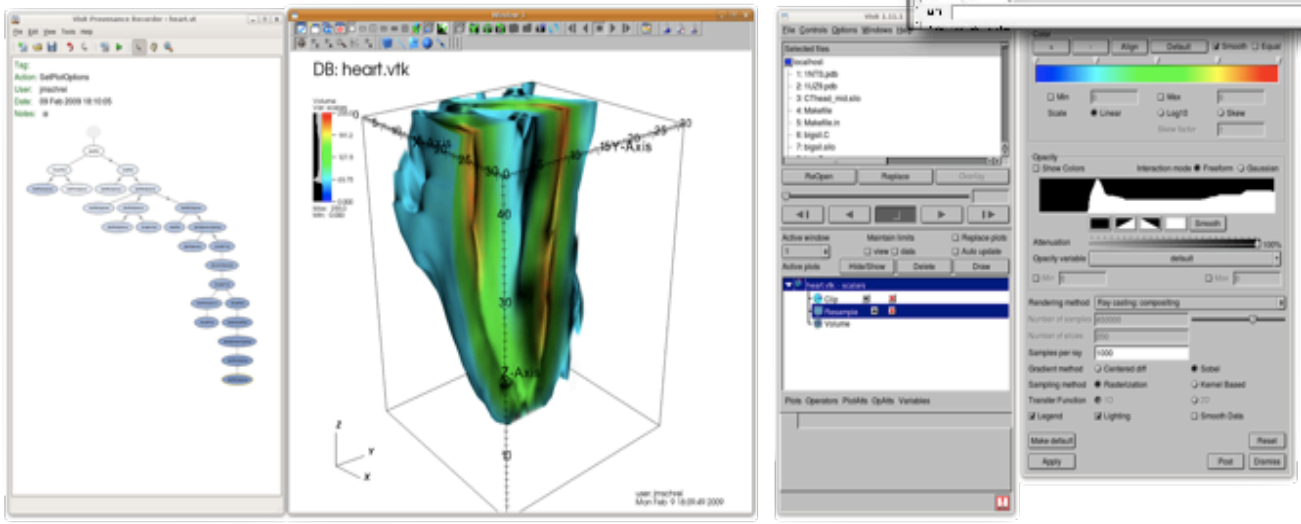
Display Wall Support





ParaView

Maya

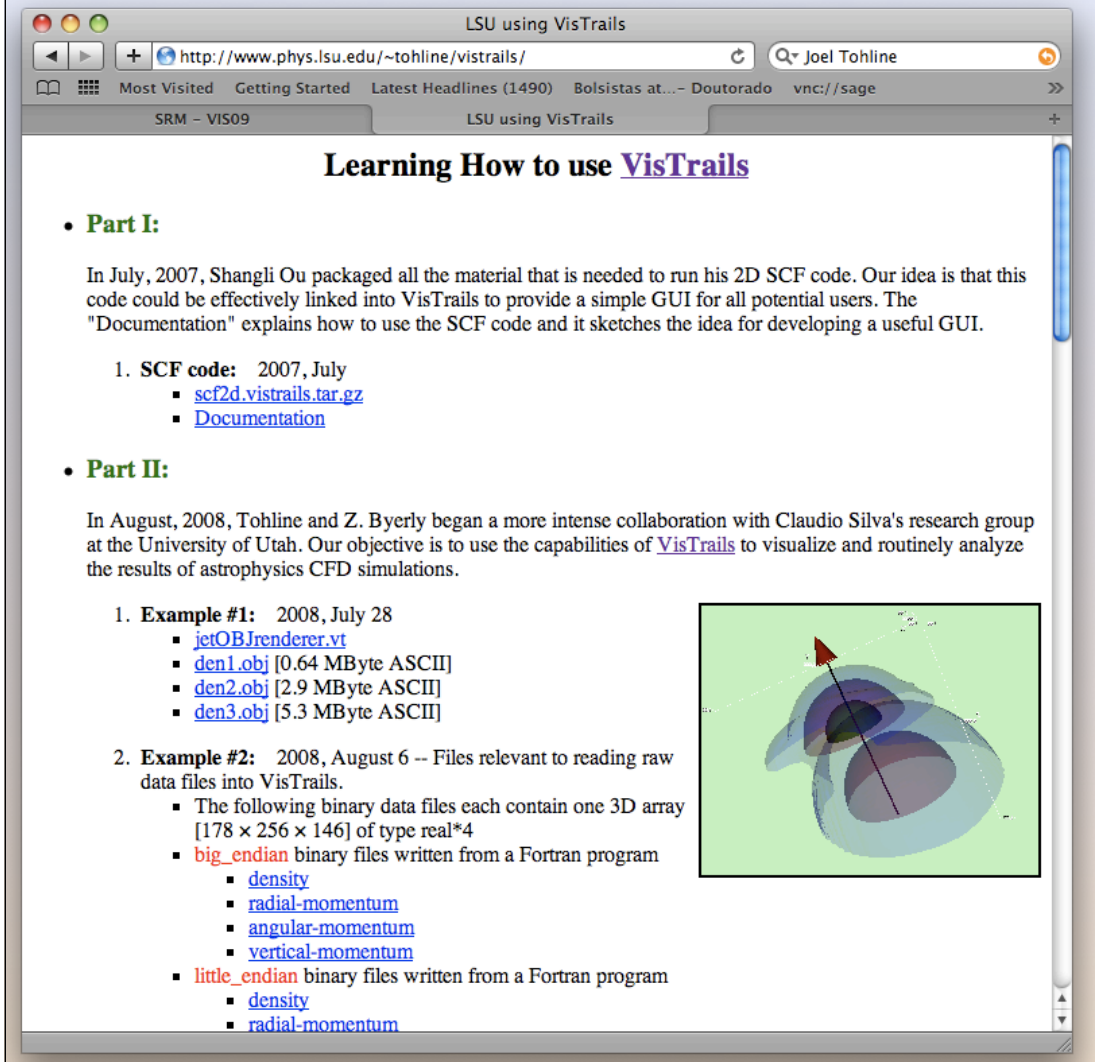


VisIt

VisTrails Plugin for ParaView

Sharing Results

Sharing Analyses and Collaborating

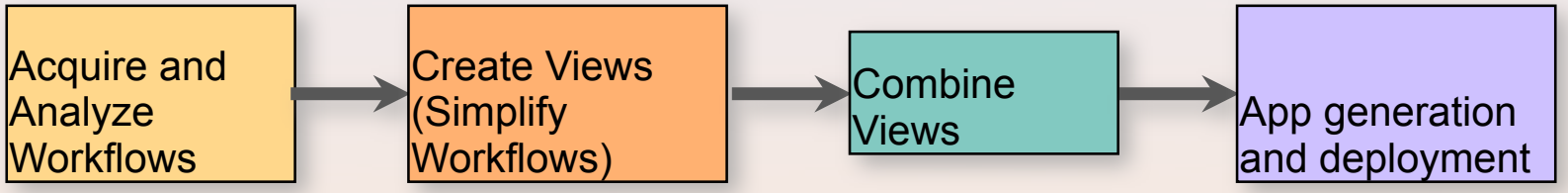


Web

- Users have to go through many (complicated) steps to reproduce and validate results:
  - install software
  - download libraries
  - download example files
  - learn how to use the software
  - manipulate workflows
- Science portals and customized apps are too expensive to develop

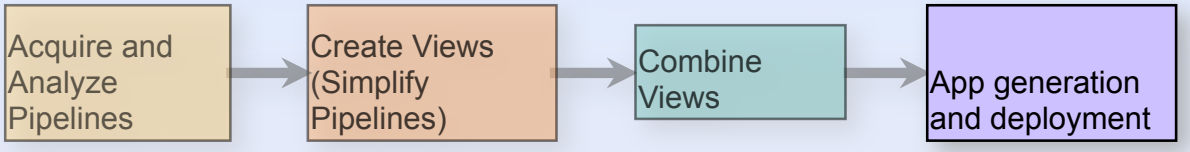
VisMashup

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



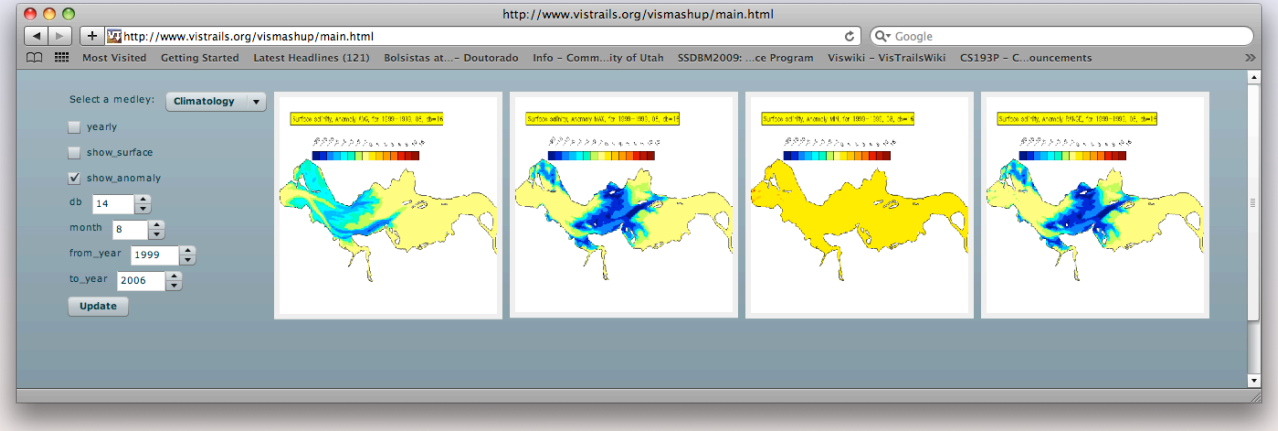
[Santos et al., IEEE TVCG 2009]

VisMashup

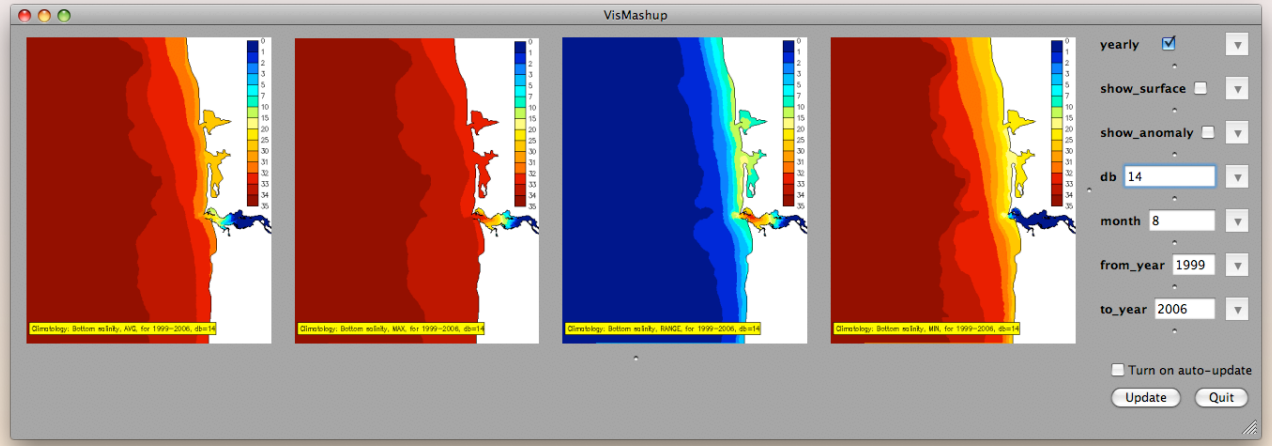


- Deploy visualization applications on different configurations

Web



Portable Devices



Desktop

[Santos et al., Submitted 2010]

Social Analysis of Scientific Data

Social Analysis of Data

- Sharing data can have important implications to science
  - *Sharing of Data Leads to Progress on Alzheimer's*, NYTimes, Aug 12, 2010
- Should also share analyses, visualizations, tools!
- ManyEyes, Swivel, Tableau Public: collaborative visualization
  - Limitations: small, tabular data; fixed set of visualization techniques; no provenance
- myExperiment: focus on bioinformatics-related Web services, share workflows
  - Limitations: can't run workflows; no provenance
- nanoHub: focus on nanotechnology, share tools, Web-based access to HPC resources (grid, cloud)

Social Analysis of **Scientific** Data

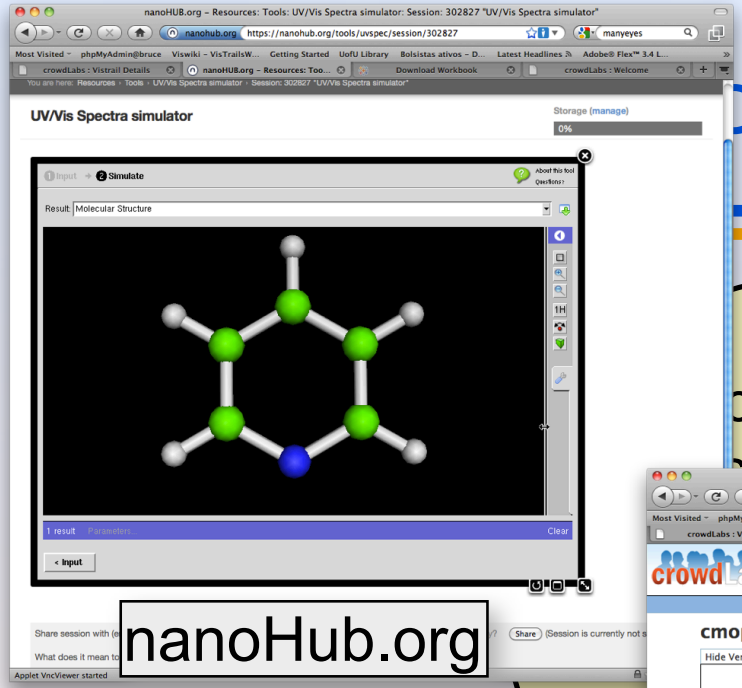
- New requirements:
  - Support for large data
  - User-defined visualizations and analyses
  - Execute workflows close to the data
  - Provenance

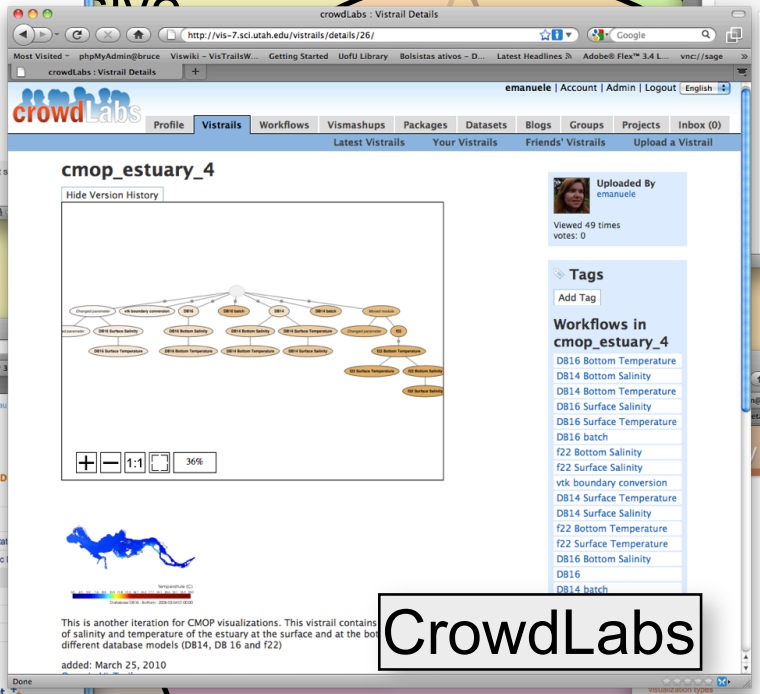
- Preliminary work: http://www.crowdlabs.org

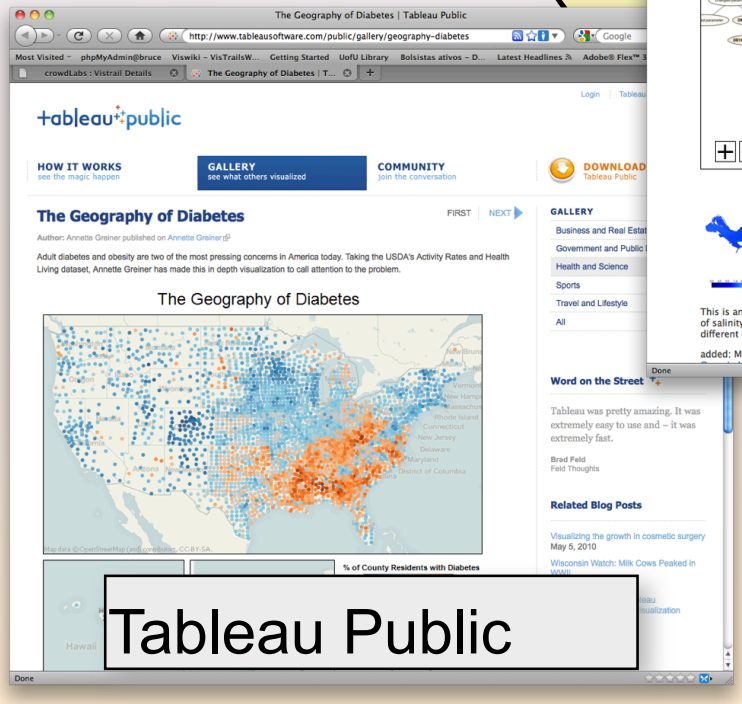
- Benefit from the collective wisdom: by querying analysis specifications which make sophisticated use of tools, along with data products and their provenance, users can learn by example from the reasoning and/or analysis strategies of experts

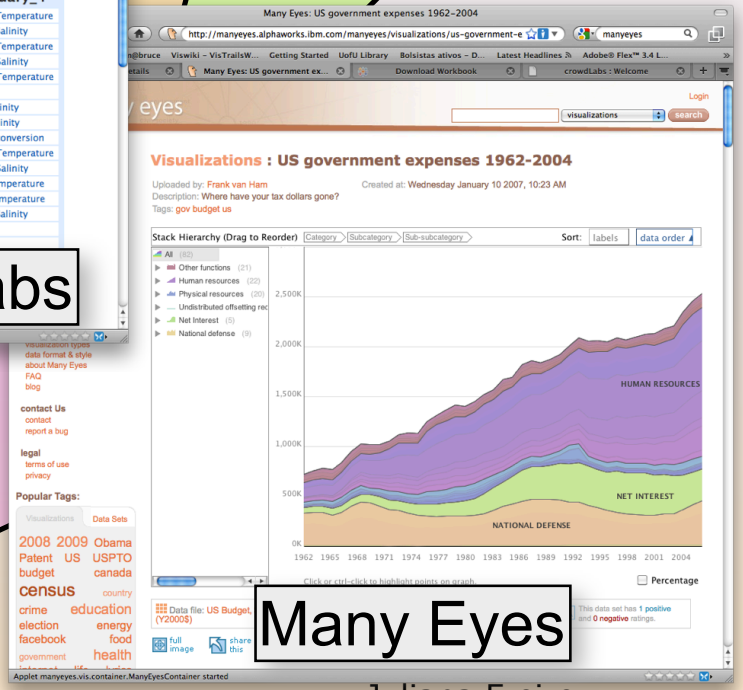
cial Data Anal

compute-









More soon…

Publishing Scientific Results

Scientific Publications and Provenance

J Appl Physiol 98: 2191–2196, 2005.
First published March 17, 2005; doi:10.1152/japplphysiol.00216.2005.

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/japplphysiol.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 ~6 l/min, lean body weight remained at ~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., ~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 intensely 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 y. 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 (±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 ~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, Belmont Hall 222, Dept. of Kinesiology and Health Education, The Univ. of Texas at Austin, Austin, TX 78712 (E-mail: coyle@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. Section 1734 solely to indicate this fact.

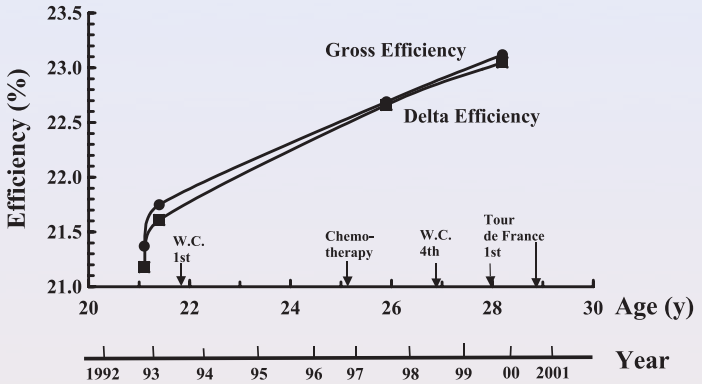
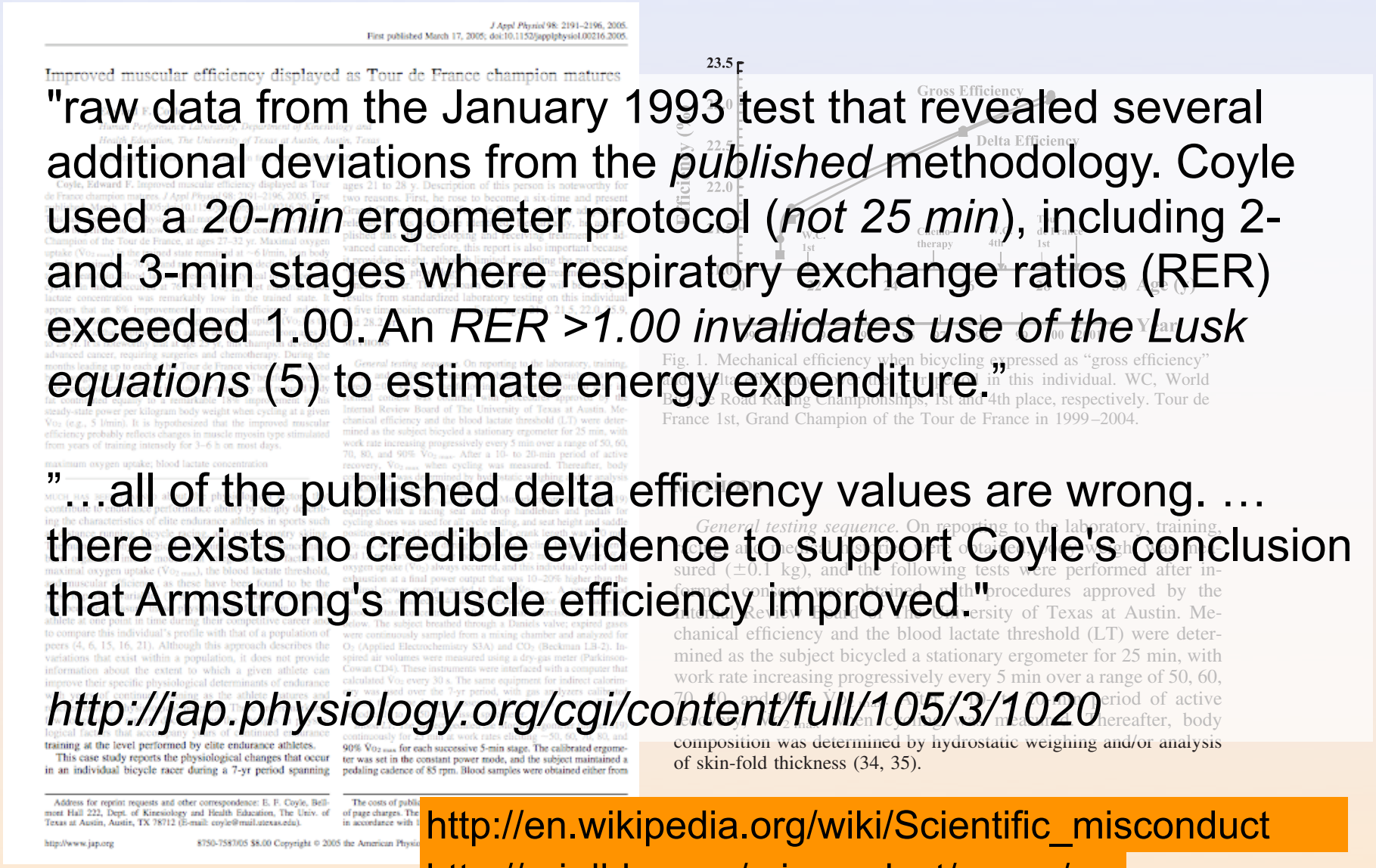


Fig. 1. Mechanical efficiency when bicycling expressed as "gross efficiency" and "delta efficiency" over the 7-yr period in this individual. WC, World Bicycle Road Racing Championships, 1st and 4th place, respectively. Tour de France 1st, Grand Champion of the Tour de France in 1999–2004.

METHODS

General testing sequence. On reporting to the laboratory, training, racing, and medical histories were obtained, body weight was measured (±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



"raw data from the January 1993 test that revealed several additional deviations from the *published* methodology. Coyle used a *20-min* ergometer protocol (*not 25 min*), including 2- and 3-min stages where respiratory exchange ratios (RER) exceeded 1.00. An *RER >1.00 invalidates use of the Lusk equations* (5) to estimate energy expenditure."

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

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

http://en.wikipedia.org/wiki/Scientific_misconduct
http://ori.dhhs.gov/misconduct/cases/

Provenance-Rich Publications

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

Provenance-Rich Publications: Benefits

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

Provenance-Rich Documents

CrowdLabs: Social Analysis and Visualization for the Sciences

Emanuele Santos, 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

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 inter-disciplinary 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 science) 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

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

Manuscript received 31 March 2010; accepted 1 August 2010; posted online 24 October 2010; mailed on 16 October 2010.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

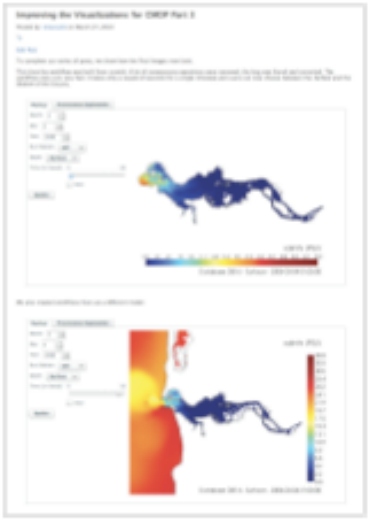


Fig. 7: Using the blog to document processes: A visualization expert created a series of blog posts to explain the problems found when generating the visualizations for CMOP.

Acknowledgments

Our research has been funded by the National Science Foundation (grants IIS-0905385, IIS-0746500, ATM-0835821, 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.

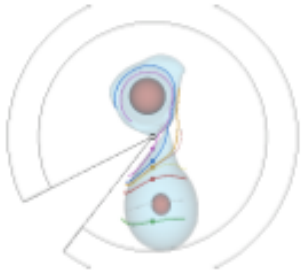


Fig. 8: Visualizing a binary star system simulation. This is an image that was generated by embedding a workflow directly in the text. The original workflow is available at http://www.crowdlabs.org/vistrails/workflows/details/1190.

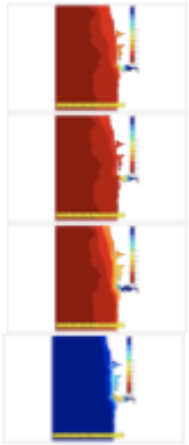
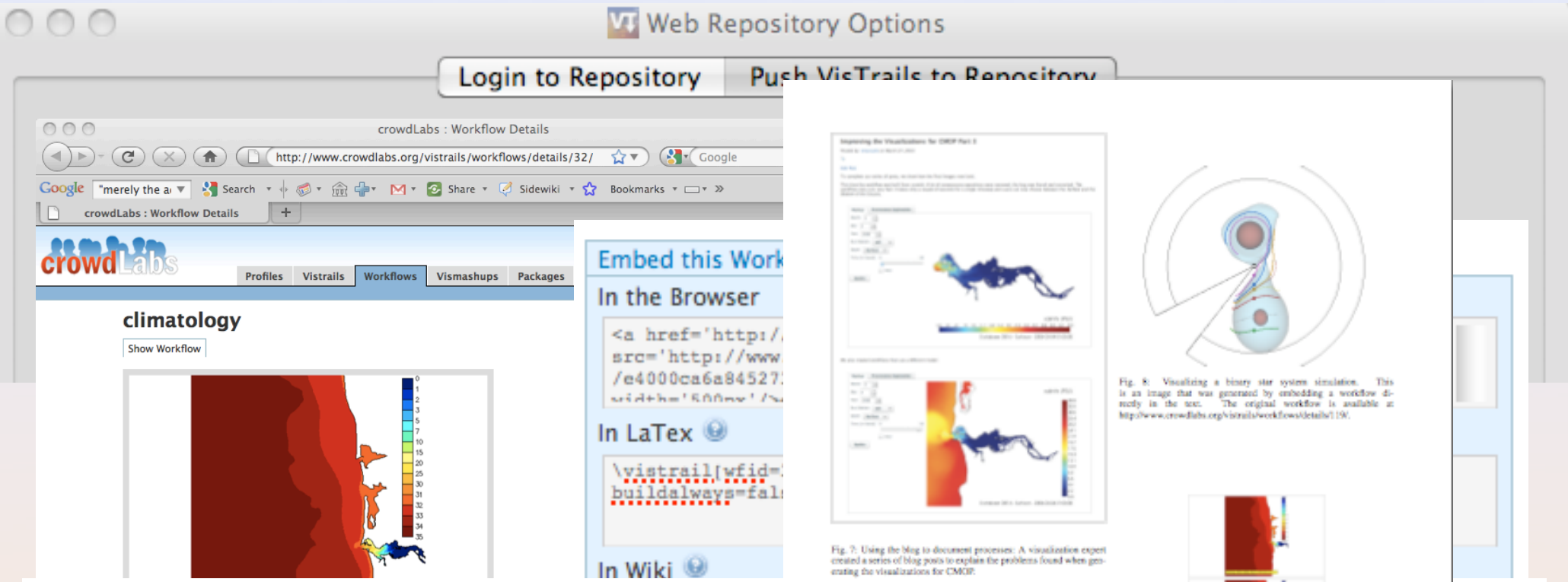


Fig. 9: Columbia river virtual estuary: visualization of salinity over time. See http://www.crowdlabs.org/vistrails/workflows/details/320.

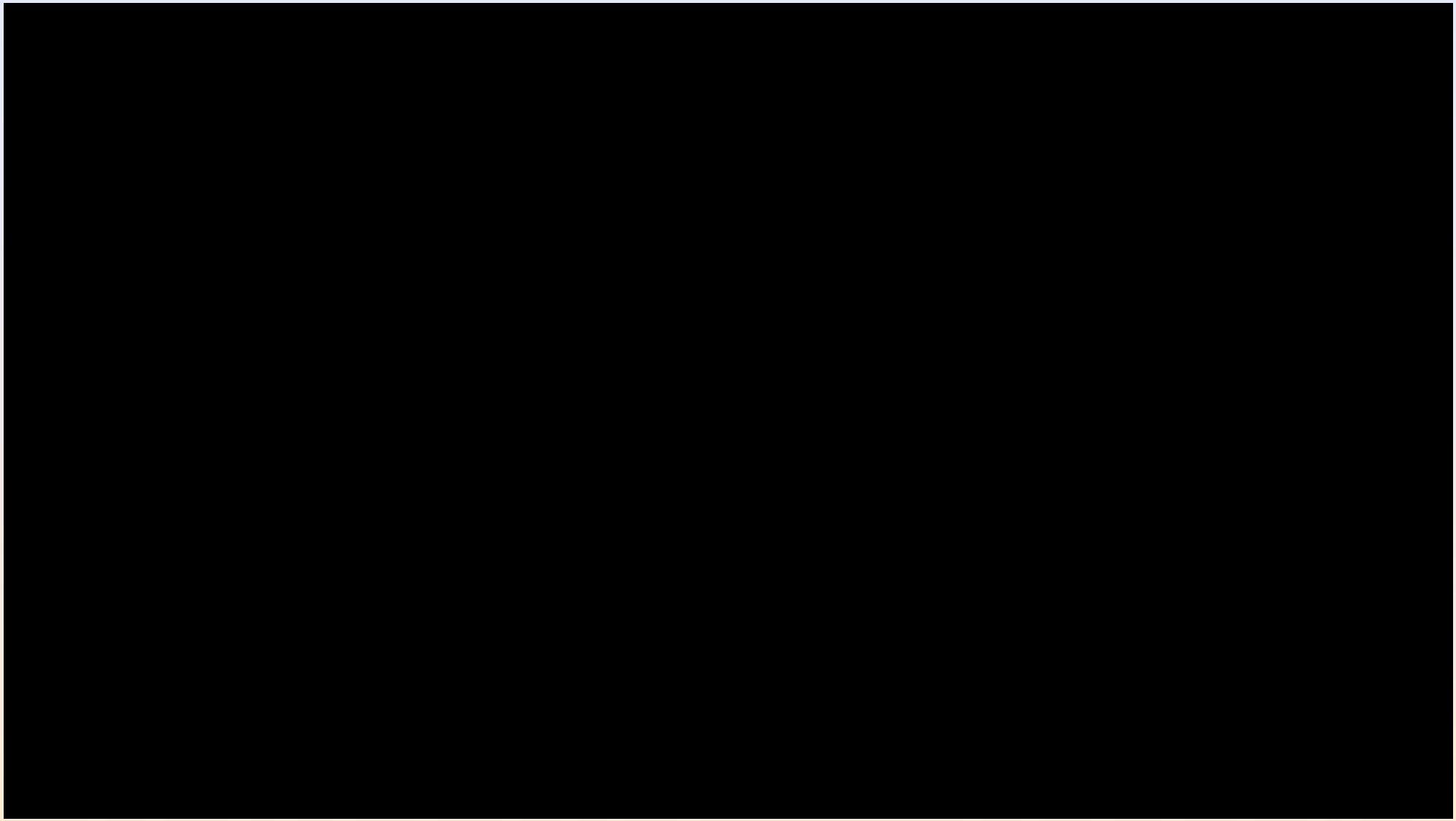
http://www.crowdlabs.org/vistrails/workflows/details/32/

Publishing with VisTrails




```
\begin{figure}[t]
\centering{
\vistrail[wfid=32,buildalways=false]{width=0.5
\caption{Columbia river virtual estuary: visua
www.crowdlabs.org/vistrails/workflows/details/
\label{fig:cmop}
\end{figure}
```

The Provenance-Rich Paper

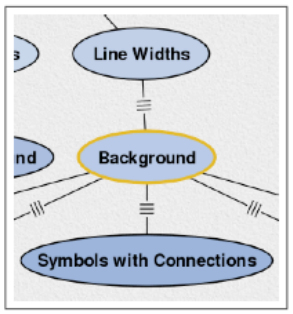


Provenance-Rich Publications: Challenges

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

Provenance and Teaching (1)

- Leverage provenance to improve the way we teach CS and Science
  - http://www.vistrails.org/index.php/SciVisFall2008
  - Lecture proven



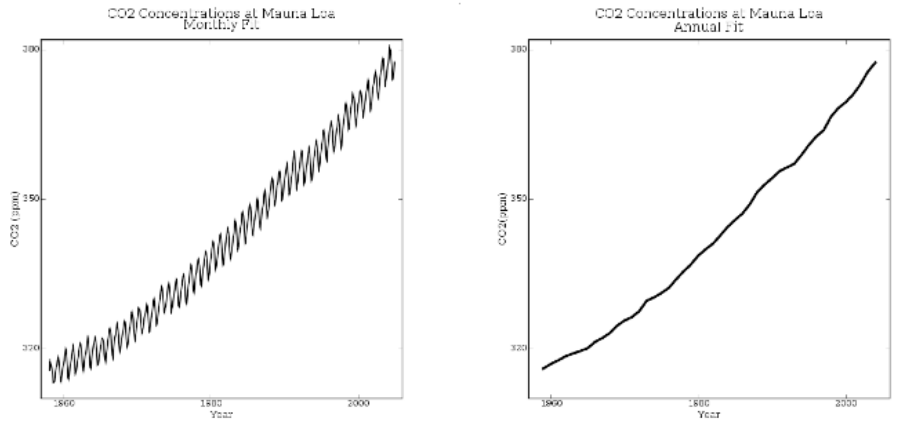
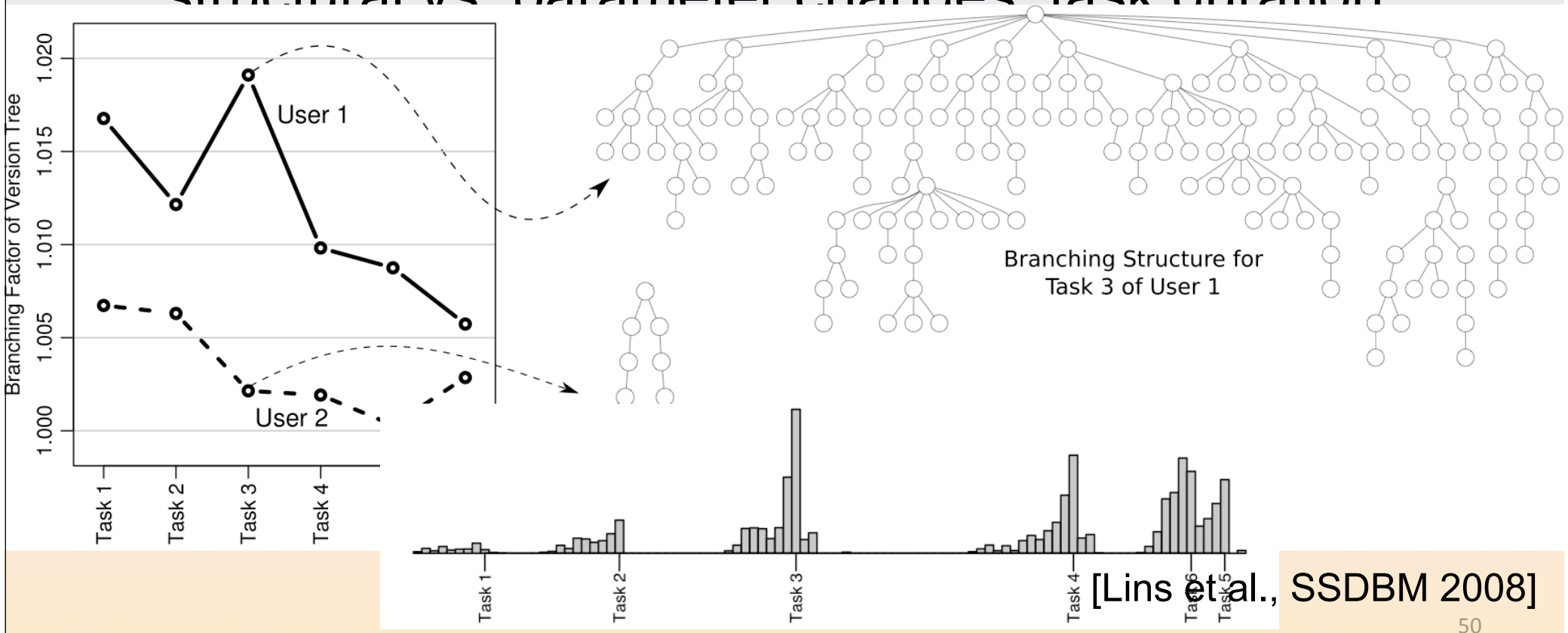


Figure 5.2: Plots of the Mauna Loa data set showing monthly measurements (left) with the yearly trend (right) using the principles for improving vision. The plot on the right is the same that was shown previously in Figure 5.1.

Provenance and Teaching (2)

- Homework provenance provides insights regarding
  - Task complexity and nature: number of actions; structural vs. parameter changes; task duration

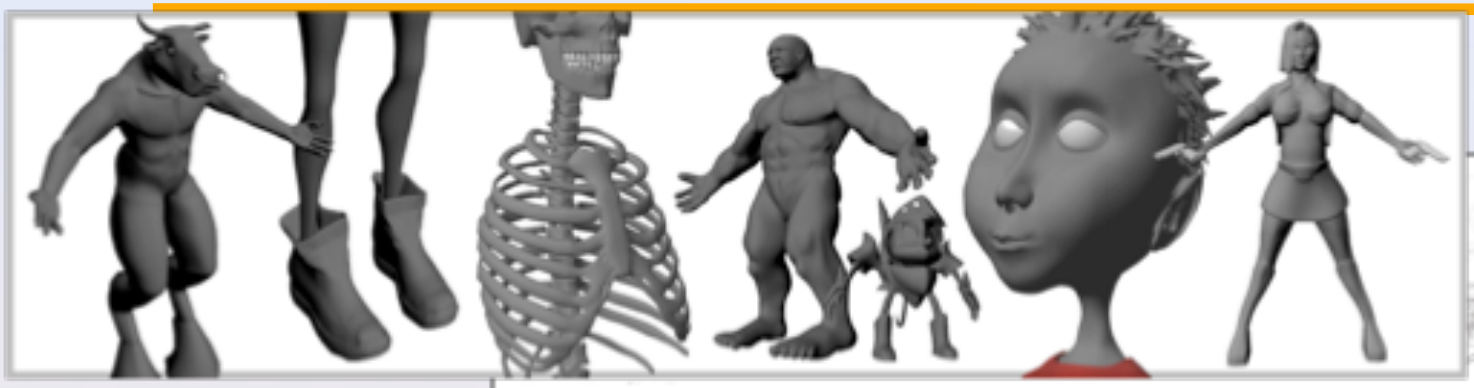


[Lins et al., SSDBM 2008]

Provenance and Teaching (3)

- Homework provenance helps students and instructors to *collaborate*
  - Student is stuck, sends his provenance
  - Instructor understands student's problem, provides hints---student can see what instructor did!
  - They can also collaborate in real time [Ellkvist et al., IPAW 2008]

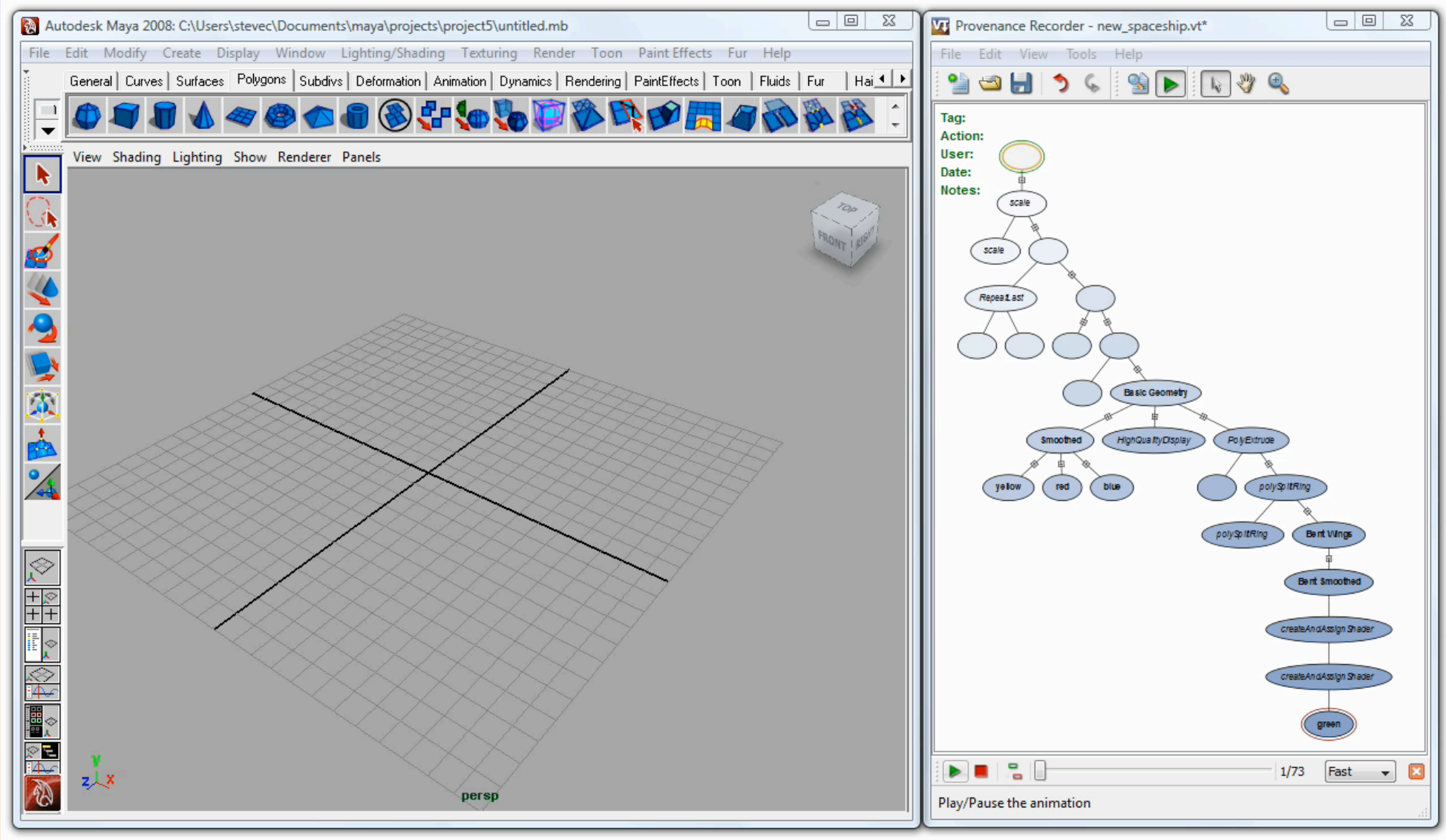
Using Provenance to Teach Electronic Media



[Langefeld and Kessler, Submitted 2009]

- "[...] The students have gotten to the point where they demand the VisTrails files for every demonstration just after I complete [it]"
- "[...] students used [a vistrail instead of a reference model] 62% of the time"
- Students who used provenance produced higher-quality models

Provenance-Based Tutorial for Maya



Conclusions and Future Work

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

Acknowledgments

- Thanks to VGC and VisTrails group
- 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.