Tutorial Type: Both regular and hands-on lab.
Level: Elementary to advanced.

Abstract—Scientists are now faced with an incredible volume of data to analyze. To explore and understand the data, they need to assemble complex workflows (pipelines) to manipulate the data and create insightful visual representations. Provenance is essential in this process. The provenance of a digital artifact contains information about the process and data used to derive the artifact. This information is essential for preserving the data, for determining the data’s quality and authorship, for both reproducing and validating results – all important elements of the scientific process. Provenance has shown to be particularly useful for enabling comparative visualization and data analysis. This tutorial will inform computational and visualization scientists, users and developers about different approaches to provenance and the trade-offs among them. Using the VisTrails project as a basis, we will cover different approaches to acquiring and reusing provenance, including techniques that attendees can use for provenance-enabling their own tools. The tutorial will also discuss uses of provenance that go beyond the ability to reproduce and share results.

Keywords—visualization system; provenance; comparative visualization;

I. TUTORIAL DESCRIPTION

Computing has been an enormous accelerator to science and has led to an information explosion in many different fields. To analyze and understand scientific data, complex computational processes must be assembled, often requiring the combination of loosely-coupled resources, specialized libraries, and grid and Web services. These processes may generate even more final and intermediate data products, adding to the overflow of information scientists need to deal with. Ad-hoc approaches to data exploration (e.g., Perl scripts) have been widely used in the scientific community, but have serious limitations. In particular, scientists and engineers need to expend substantial effort managing data (e.g., scripts that encode computational tasks, raw data, data products, and notes) and recording provenance information so that basic questions can be answered, such as: Who created this data product and when? When was it modified and by whom? What was the process used to create the data product? Were two data products derived from the same raw data? Not only is the process time-consuming, but also error-prone.

In the visualization domain the situation is not different. Although scientific visualization has matured and now is an indispensable tool for scientists and engineers, they still lack mechanisms to help them not only share visualizations but also share and reuse information about the process used to generate those visualizations. Examples of widely used visualization tools are ParaView [1] and VisIt [2]. ParaView [1] is an open-source, multi-platform application designed to visualize data sets of size varying from small to very large. It is very popular and it has been used by researchers and engineers in both industry and academia. VisIt [2] is another free visualization and graphical analysis tool for viewing scientific data.

Recently, workflow systems [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] have emerged as an alternative to these ad-hoc approaches. Not only do they support the automation of repetitive tasks, but they can also capture complex analysis processes at various levels of detail and systematically capture provenance information for the derived data products [13]. The provenance (also referred to as audit trail, lineage, and pedigree) of a data product contains information about the process and data used to derive the data product. It provides important documentation that is key to preserving the data, to determining the data’s quality and authorship, and to reproduce as well as validate the results. These are all important elements of the scientific process. Provenance in scientific exploration and visualization is thus of paramount and increasing importance, as evidenced by recent specialized workshops [14], [15], [13], [16], [17], [18] and surveys [19], [20], [21], [22]. Provenance has also shown to be particularly useful for enabling comparative visualization and data analysis [23], [24], [25].

Our goal with this tutorial is to inform computational and visualization scientists, users and developers about different approaches to provenance and the trade-offs among them. The instructors will use the open-source VisTrails system1 (Figure 1) and a series of examples from real applications to

1www.vistrails.org
demonstrate different approaches to acquiring and reusing provenance, including techniques that attendees can use for provenance-enabling their own tools. VisTrails is a provenance management and scientific workflow system that was designed to support the scientific discovery process. VisTrails provides unique support for data analysis and visualization, a comprehensive provenance infrastructure, and a user-centered design. The system combines and substantially extends useful features of visualization and scientific workflow systems. A beta version of VisTrails was first released in January 2007. Since then, the system has been downloaded over fifteen thousand times. The VisTrails wiki has had over 339,000 page views, and Google Analytics reports that visitors to the site come from 61 different countries. VisTrails has been adopted in several scientific projects, both nationally and internationally, and in different areas, including Environmental Sciences [26], [27], [28], [29], [30], Psychiatry [31], Astronomy [32], Cosmology [25], High-Energy Physics [24], molecular modeling [33]. VisTrails is being combined with CDAT (Climate Data Analysis Tools) [30], a popular platform for accessing and analyzing climate data available in major portals such as the Earth System Grid. VisTrails has been successfully used as a tool for teaching [34], [35] and it has been adopted at universities in the United States and abroad. VisTrails has recently been featured as an NSF Discovery.

In this tutorial we will also discuss uses of provenance that go beyond the ability to reproduce and share results. We will cover different approaches to capture and reuse provenance. We will show how the availability of provenance supports reflective reasoning as users perform exploratory tasks, including the ability to navigate through versions of analysis and visualization pipelines in an intuitive way, to undo changes but not lose any results, to visually compare different pipelines and their results, and to examine the actions that led to a result. We will also present a series of operations and user interfaces that leverage provenance simplify pipeline design, use and sharing, as well as the publication of documented, reproducible results. Last, but not least, we will discuss techniques and present an API that attendees can use to provenance-enable their own tools.

A. Tutorial History

This tutorial is a modified version of five other tutorials, focusing on specific aspects of provenance that are relevant for visualization scientists, users and developers:


New material in this tutorial includes: interactive demos and hands-on material with a larger focus on data exploration and visualization: provenance-enabling third-party tools and new applications that are enabled by the availability of provenance.

II. CONTENTS AND SCHEDULE

This 1-day tutorial will consist of four blocks, each of about 90 minute length.

Module I: Motivation and Provenance Basics

• What is Provenance? Why is it important? How can Provenance be leveraged?
• Background: Visualization Systems
• Background: Scientific Workflow Basics
Provenance in Scientific Workflow Systems
Provenance in Databases
The Open Provenance Model [36]

We discuss the motivation for and importance of provenance for computation and visualization tasks. We also present real scenarios in different application domains (including environmental observatories, forecasting systems, and clinical studies) to illustrate the need for and benefits of provenance. We then review work on visualization systems (e.g., [1], [2], [8], [37], [5], [38], [39]) and provenance for data products that are derived by complex computational tasks modeled as workflows (e.g., [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50]), and contrast these to recent work on provenance for structured data [51], [52], [53]. We survey approaches to provenance adopted by scientific workflow systems. We present and compare different proposals for capturing, modeling, storing and querying provenance (e.g., [54], [55], [50], [56], [57], [22], [21]).

Module II: Provenance for Data Exploration and Visualization

Building visualizations
Operations that support exploration: scalable exploration of large parameter spaces; comparison of data products and their corresponding workflows [58], [59].
Operations that support visualization design: pipeline refinement by analogy [54]; a recommendation system for visualization pipelines [60].
Provenance analytics: Techniques to support provenance mining and illustration through examples of useful information that can be extracted from provenance ([61], [62]).
Provenance for collaboration [63].

We show how provenance can be used to simplify exploratory processes. In particular, we give examples that show how to build pipelines and combine visualization with other tasks [32]. We also present mechanisms that allow the flexible re-use of visualization pipelines; scalable exploration of large parameter spaces; and comparison of data products as well as their corresponding pipelines [58], [59]. We also show that useful knowledge is embedded in provenance which can be re-used to simplify the construction of visualization through pipeline refinement by analogy [54] and an auto-completion mechanism for visualization pipelines [60]. Last, but not least, we present techniques to support provenance analytics and illustrate through examples, useful information that can be extracted from provenance (see e.g., [61], [62]).

Module III: Provenance-Enabling Existing Libraries and Tools

Provenance-enabling libraries and command-line tools: how to write VisTrails packages
Provenance-enabling ParaView
Provenance-enabling VisIt

Figure 2. Provenance-enabled VisIt.

Provenance-enabling your own tool

We show how to wrap libraries and command-line tools in VisTrails so they can benefit from VisTrails' provenance mechanism. We also show how existing interactive tools can be (easily) extended to support provenance. We will present case studies on adding provenance to two distinct tools: VisIt [2] (see Figure 2) and ParaView [1] (see Figure 3). We will also describe guidelines on how to add provenance to user-custom tools.

Module IV: New Applications

Science 2.0: collaborative, open science
Provenance-rich publications
Visualization Mashups
Provenance in teaching

We describe applications that are enabled by the availability of provenance. We demonstrate how provenance-rich objects can be included in documents. We also show how the VisTrails infrastructure can be used to create provenance-aware custom visualization applications called visualization mashups [64]. We also explore the benefits of provenance in teaching [34].

III. Presentation Requirements

Equipment and software: a digital projector for the instructors and computers with VisTrails installed and Internet access for the attendees.

IV. Tutorial Resources

This tutorial will provide the slides used on each module, examples and datasets used on the hands-on parts and a survey paper. We will also make available the related articles. We will maintain a Wiki page at http://www.vistrails.org/index.php/Tutorials/Sibgrapi2010 that will contain the most up-to-date information about the tutorial. For an example of slides used in previous versions of this tutorial, please visit the url above.

V. Instructors Background

Cláudio Silva received the BS degree in mathematics from the Federal University of Ceara, Brazil, in 1990, and the PhD degree in computer science from the State University of New York at Stony Brook in 1996. He is a Professor of computer science and a faculty member of the Scientific
Juliana Freire is an Associate Professor at the School of Computing, University of Utah. Before joining the University of Utah, she was member of technical staff at the Database Systems Research Department at Bell Laboratories (Lucent Technologies) and an Assistant Professor at OGI/OHSU. She has co-authored over 100 technical publications and holds 4 U.S. patents. She has chaired or co-chaired several workshops and conference, and she has participated as a program committee member in over 60 events. Dr. Freire’s research has focused on extending traditional database technology and developing techniques to address new data management problems introduced by the Web and scientific applications. Within scientific data management, she is best known for her work on provenance management and for being a co-creator of VisTrails system. In 2008, Dr. Freire received an NSF CAREER Award and an IBM Faculty Award. Her research has been funded by grants from the National Science Foundation, Department of Energy and the University of Utah.

Emanuele Santos is a research assistant and graduate student at the University of Utah. She received M.S. and B.S. degrees in computer science from the Federal University of Ceara in Brazil. Between 2002 and 2005, she lectured college-level computer science courses in Brazil. Her research interests include scientific data management, comparative visualization and provenance-rich publications. She is a Fulbright scholar and one of the main developers of VisTrails, an open-source scientific workflow and provenance management system.

**Erik W. Anderson** is a research assistant and graduate student at the University of Utah. He received a B.S. degree in computer science and a B.S. degree in electrical and computer engineering from Northeastern University. His research interests include scientific visualization, signal processing, computer graphics, and multimodal visualization. He is regularly holding seminars on neuroscience analysis and visualization at the SCI Institute. He is also one of the main developers of VisTrails, an open-source scientific workflow and provenance management system.

**REFERENCES**


