Provenance for Data Analysis and Visualization

Claudio Silva, Steve Parker, Ayla Khan, Emanuele Santos, Erik Anderson, and others

Scientific Computing and Imaging Institute

School of Computing

University of Utah

Outline

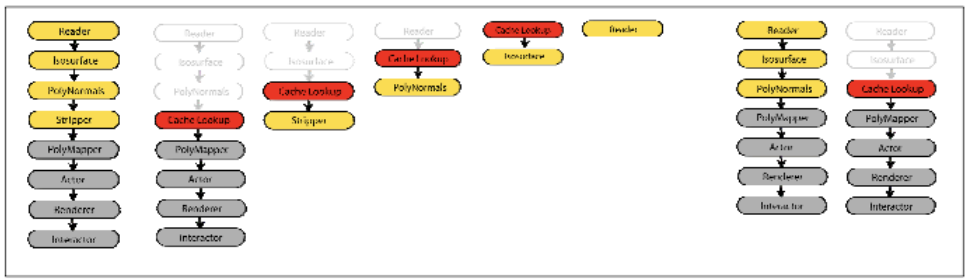
- Past
- Present
- Future

Past: Provenance in Kepler

- **Defining what was the "important" provenance -- large effort: almost everyone at SPA involved…**
  - "Provenance Schema", Meiyappan Nagappan, David Koop, Daniel Crawl, currently in version 7. This schema thus stores information that can be used to view the 4 kinds of provenance:
    - Data Provenance - execution data
    - Process Provenance - workflow execution and system information
    - Workflow Provenance - workflow specification and evolution
    - System Provenance - system information
- **Implementation (On-going)**
  - Crawl leads the Kepler implementation effort
  - Many others working on different parts as necessary, e.g. Nagappan working on "data provenance"; Norbert adding workflow-specific provenance; Santos working on "analysis provenance"; many more…

Past: Provenance in Kepler

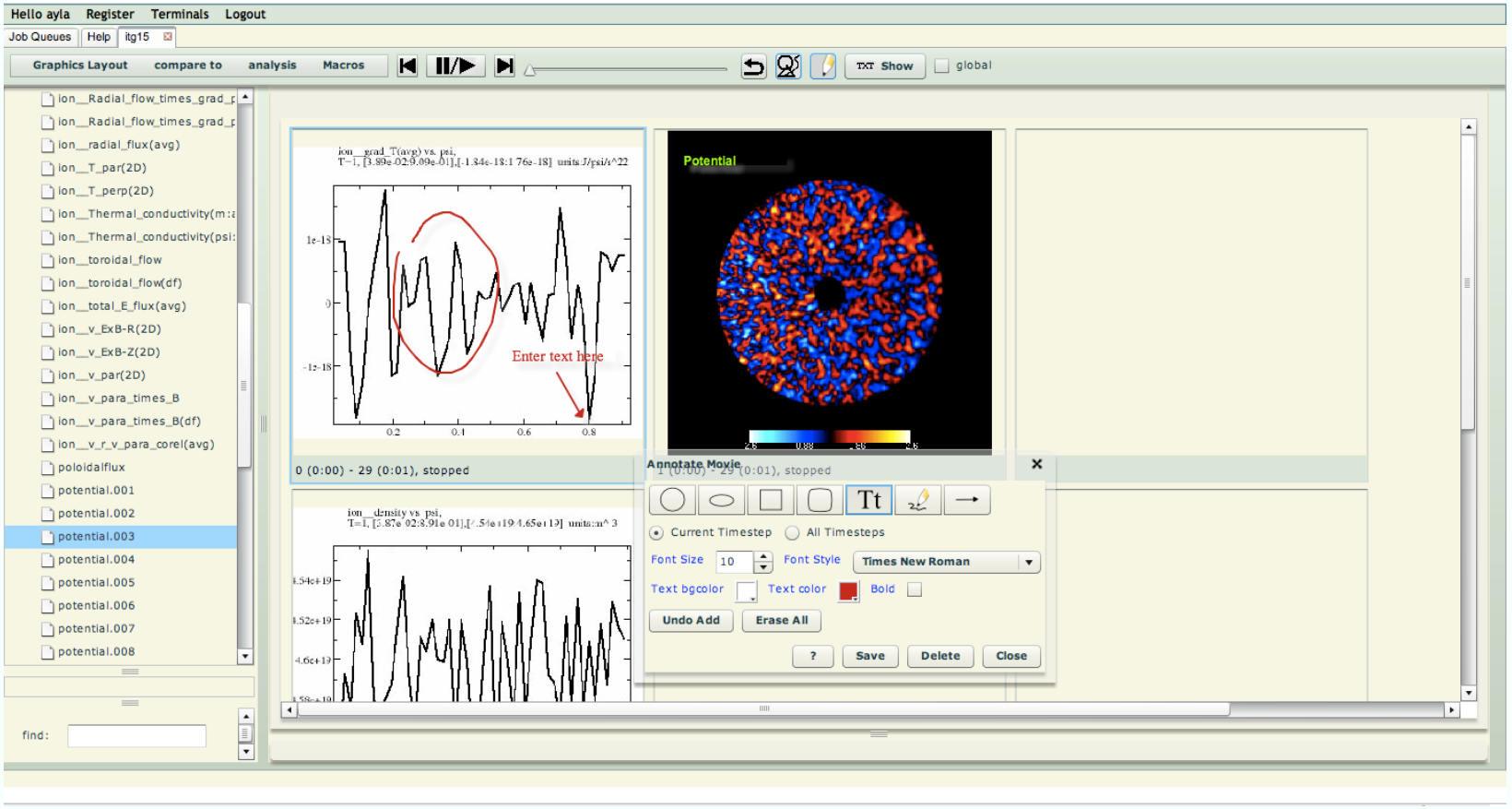
- **Kepler "Smart Re-Runs"**
  - **Provenance Collection Support in the Kepler Scientific Workflow System, I.Altintas, O. Barney, E. Jaeger-Frank, IPAW 2006, Chicago, Illinois, May 2006.**
  - **VisTrails: Enabling Interactive Multiple-View Visualizations, L. Bavoil, S. Callahan, P. Crossno, J. Freire, C. Scheidegger, C. Silva, and H. Vo, IEEE Visualization 2005.**



- Important for scalability
- The Cache Manager determines pipeline sharing
- Each module is broken into a series of subnetworks
- Each subnetwork receives a unique ID, comprising its modules, connectivity and parameters
- Results are linked to the ID, and only computed if missing in the cache

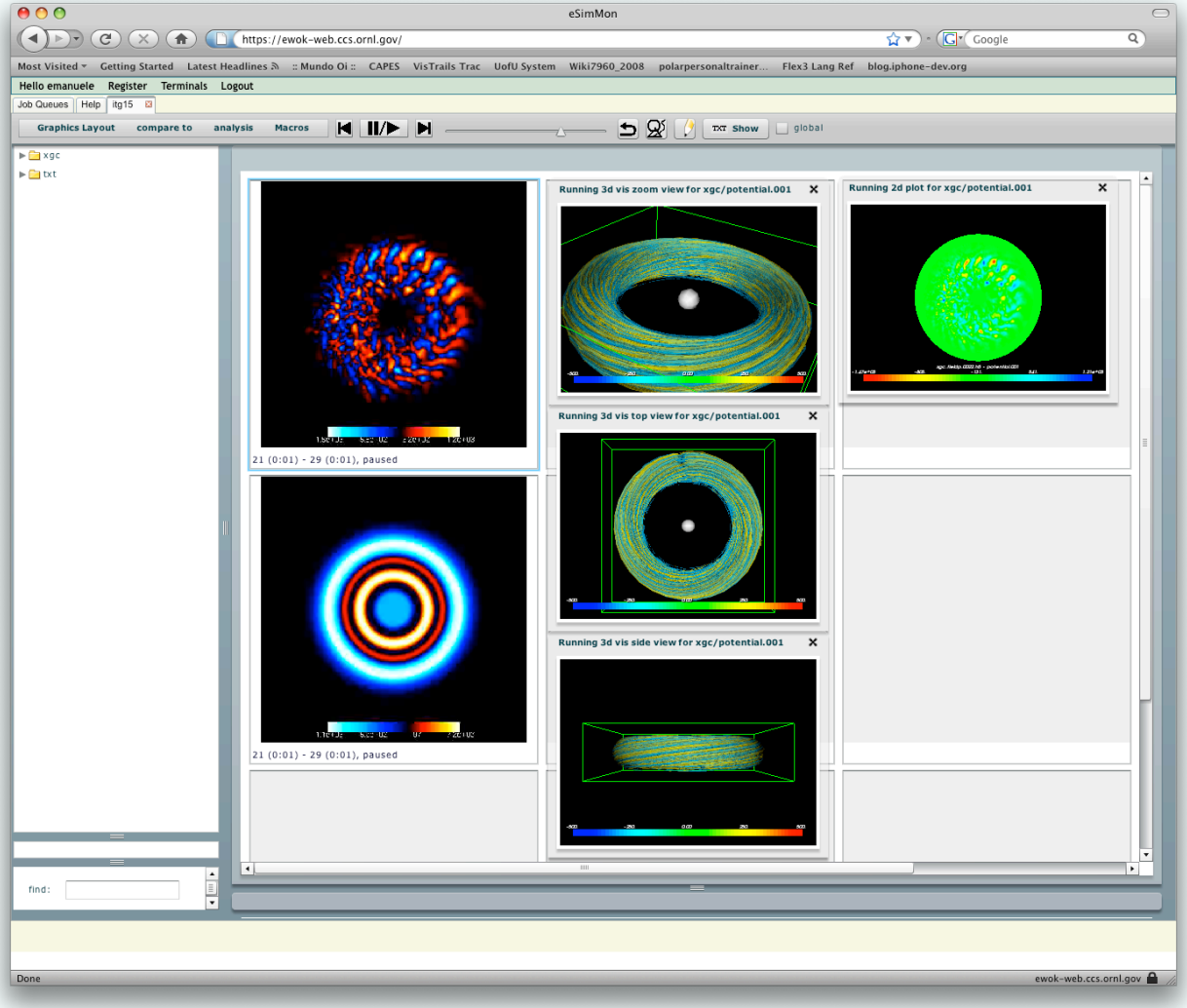
See Bavoil et al, IEEE Visualization, 2005

Present: Dashboard



your name here

Present: Dashboard



Present/Future: Analysis Workflows

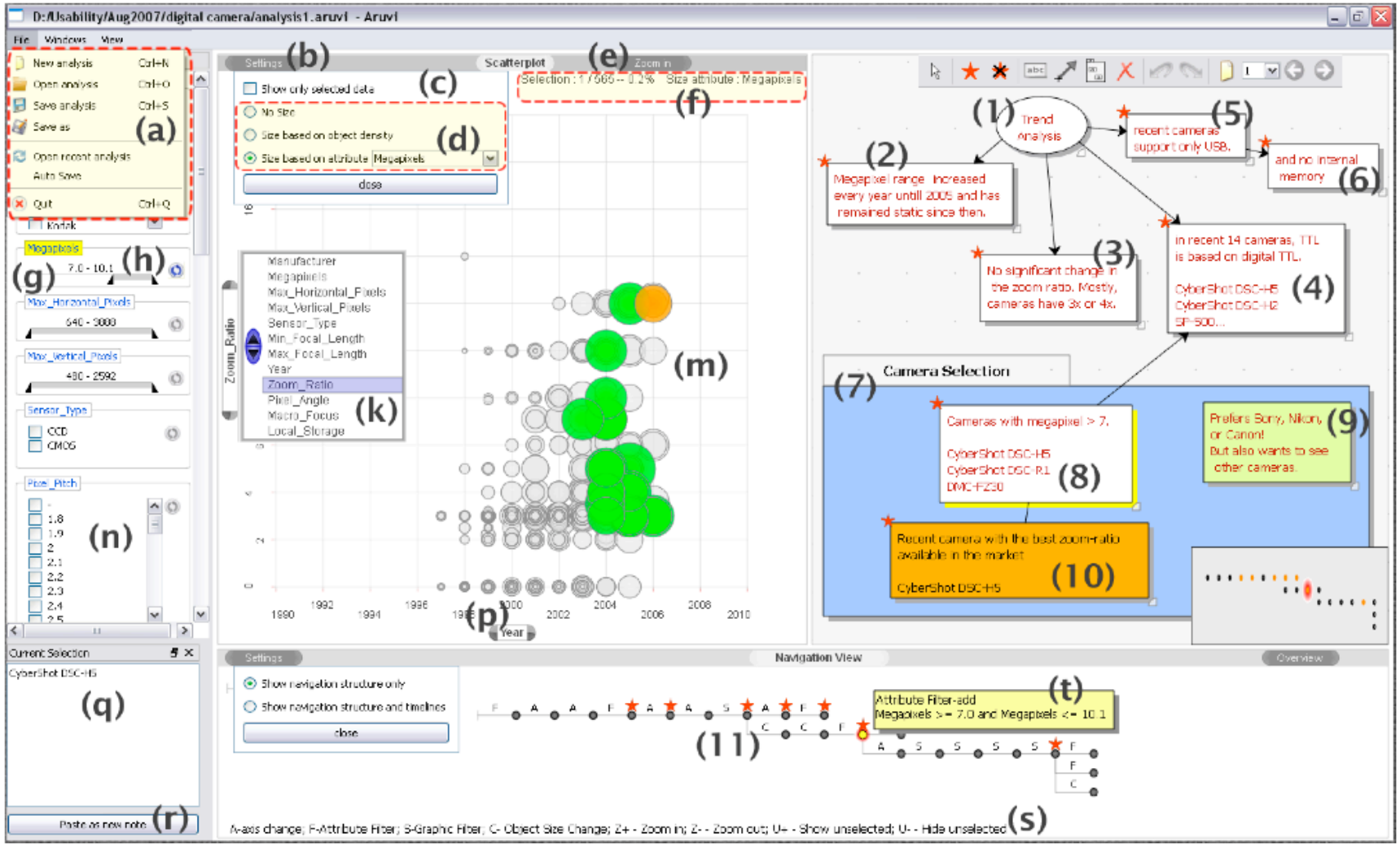
Jackie Chen

- Ayla Khan & Brad Grimm (new hire)

- **Maintenance/Updates**
  - Ayla Khan (Dashboard)
  - Emanuele Santos (Medleys)
  - Erik Anderson (3-D visualizations)

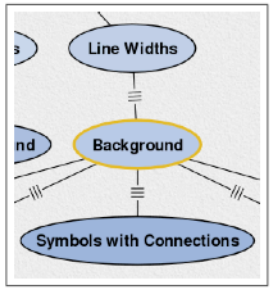
Future: Provenance for Analysis and more

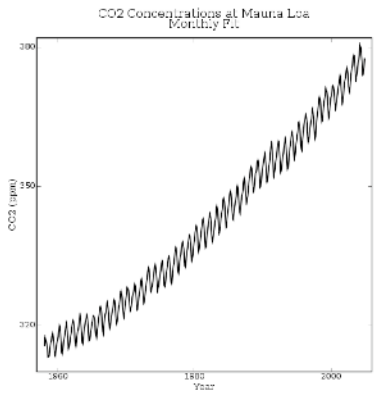
Aruvi (Shrinivasan & van Wijk, 2008)



Provenance and Teaching

- **Leverage provenance to improve the way we teach CS and Science**
  - **www.vistrails.org/index.php/SciVisFall2007 (also 2008)**
- **Lecture provenance: student can reproduce results**





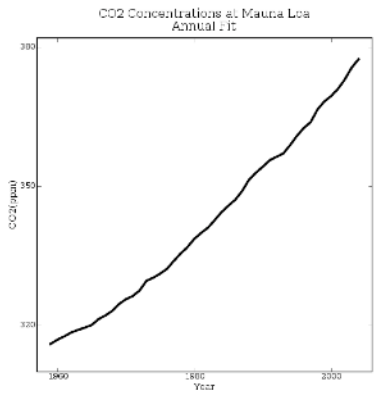
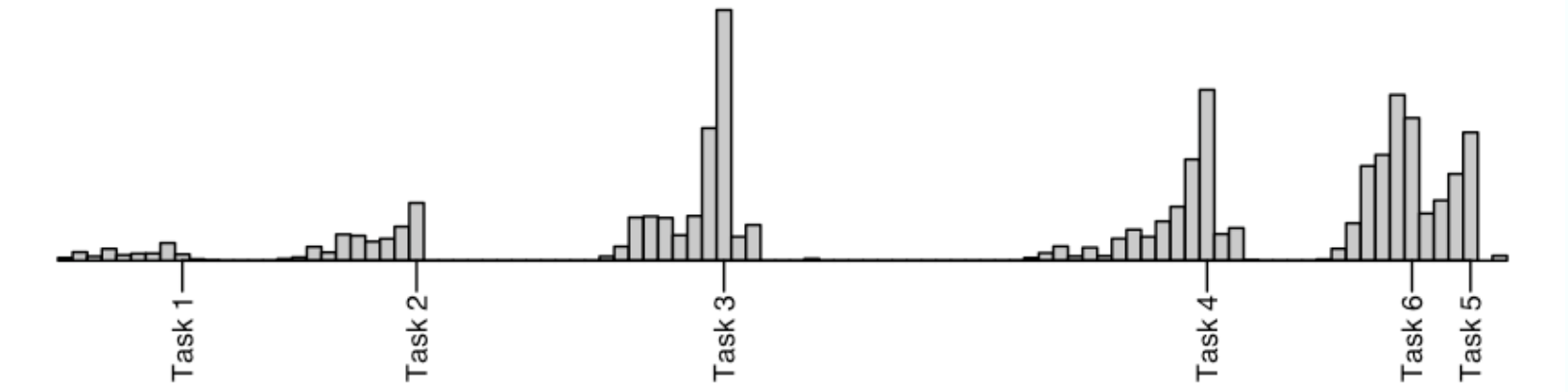
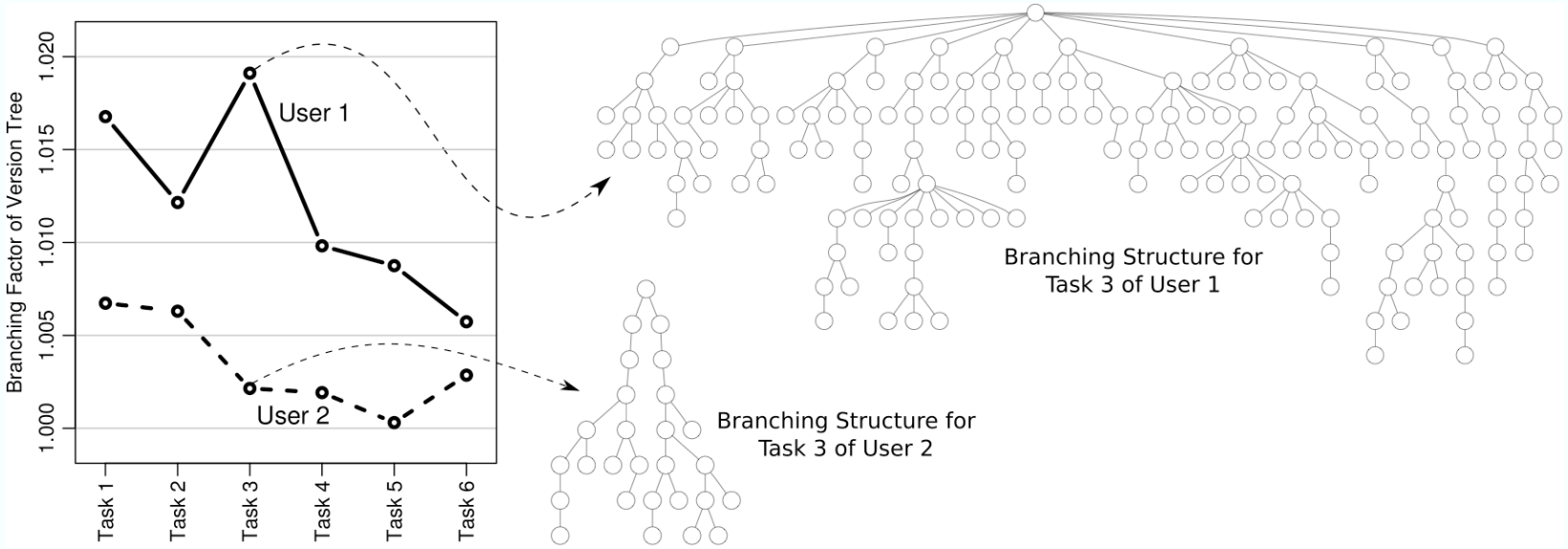


Figure 5.2: Plots of the Mauna Loa data set showing monthly measurements (left) with the yearly trend (right) using the principles for improving vision. The plot on the right is the same that was shown previously in Figure 5.1.

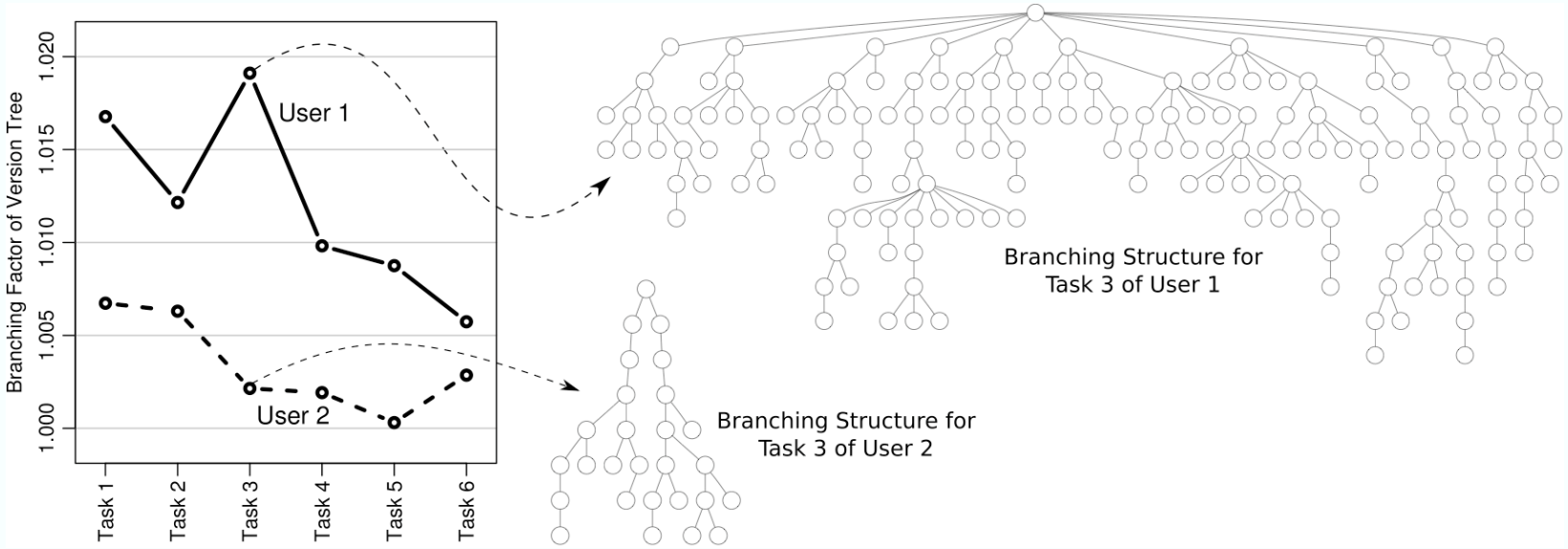
- **Workflow evolution provenance provides insights regarding**
  - **Task complexity and nature: number of actions; structural vs. parameter changes; task duration**
  - **Student confusion: large branching factor=lots of trial and error steps**



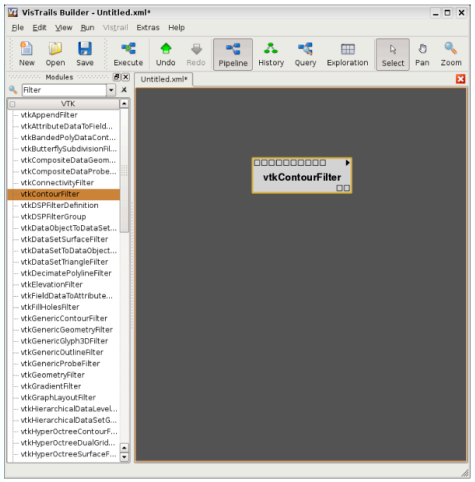
Provenance and Teaching



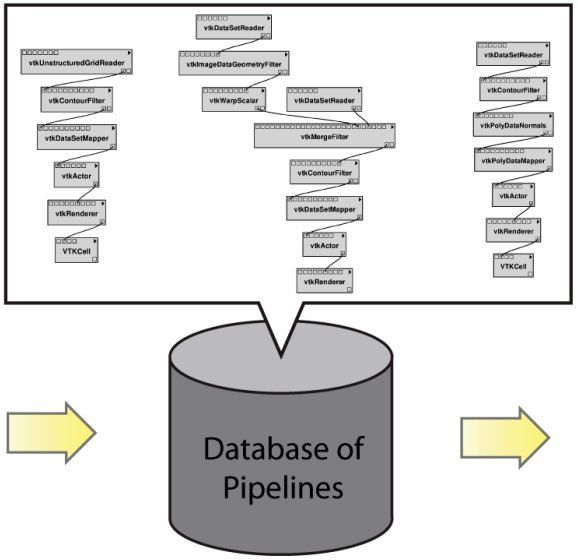
Provenance and Teaching



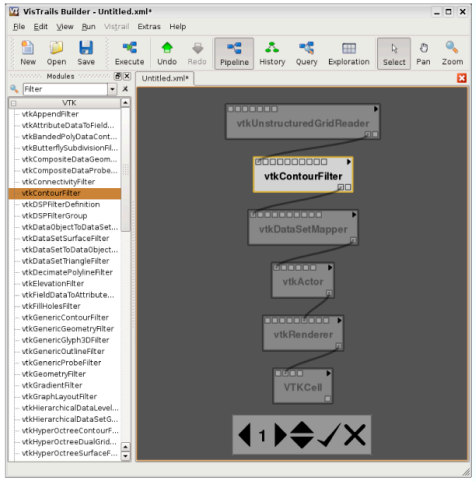
VisComplete: A Workflow Recommendation System



(a)

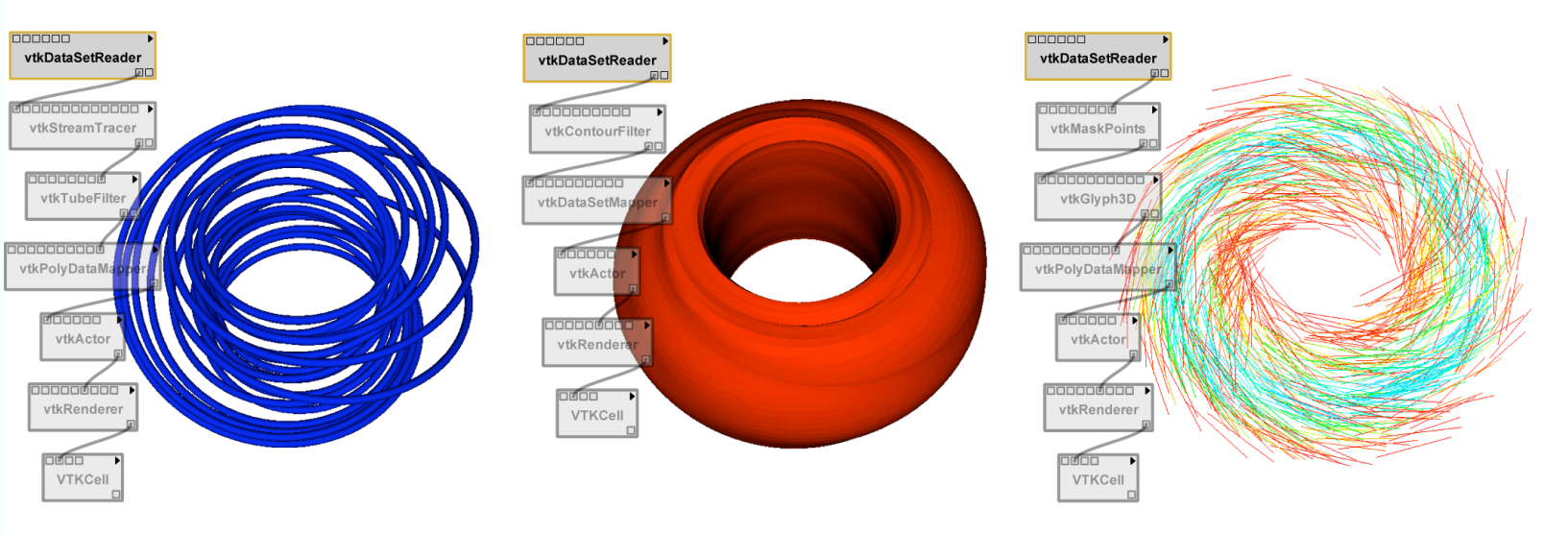


(b)



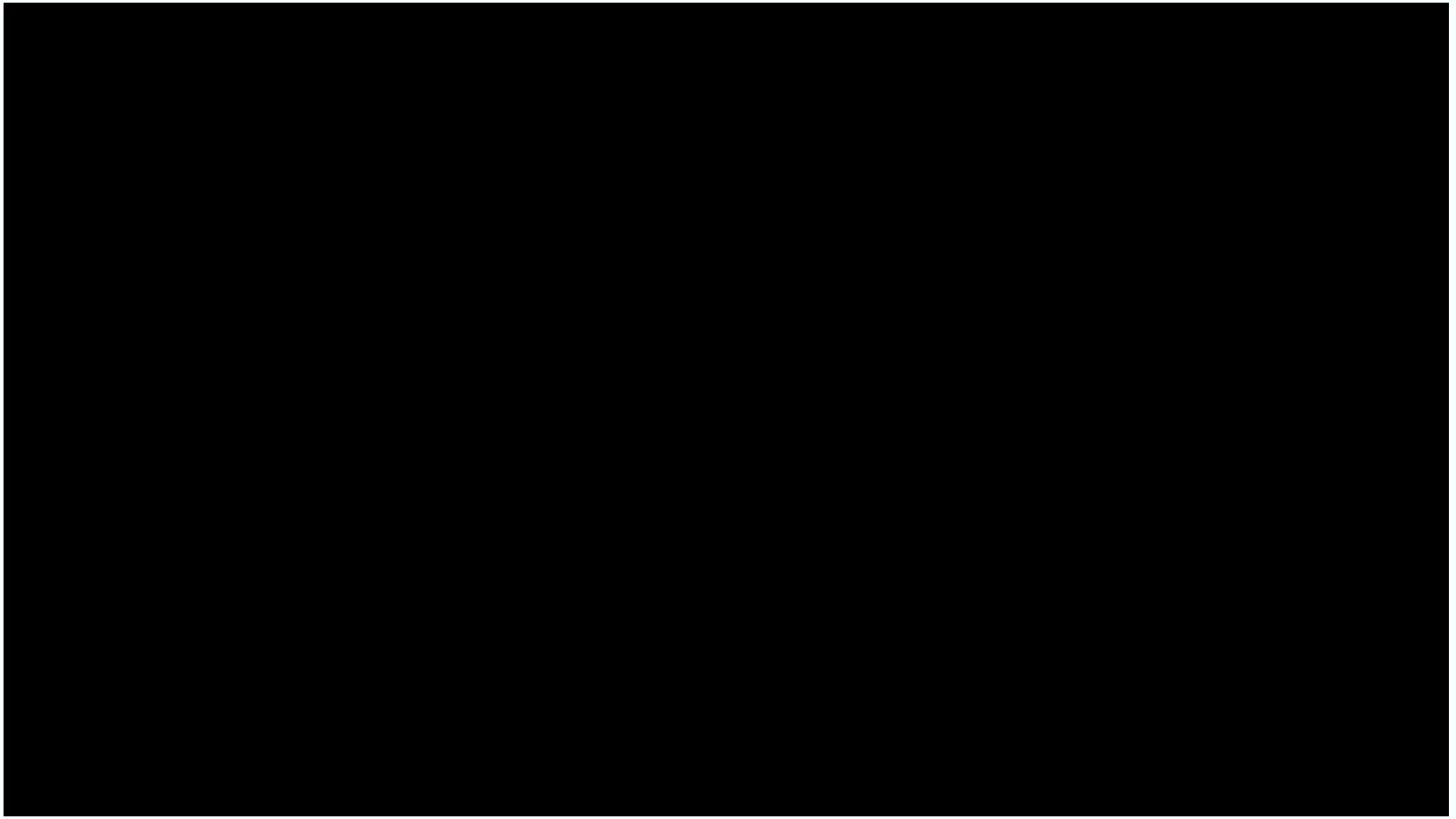
(c)

VisComplete: A Workflow Recommendation System



your name here

The Provenance-Enabled Publication



Collections of Workflows

- **Video**

Acknowledgement

- **Funded by:**

- Level 1
  - Level 2
  - Level 2
  - Level 2
- Level 1
  - Level 2
  - Level 2
  - Level 2