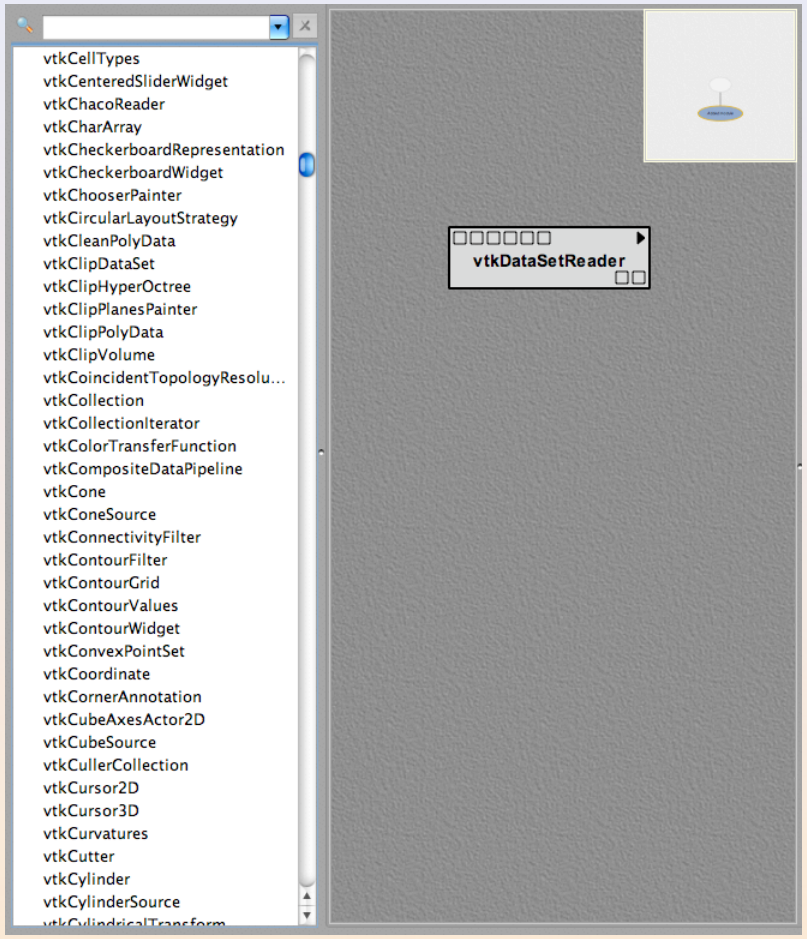
Provenance Analytics and Visualization

Juliana Freire
VisTrails Group & Web and Databases Lab

Provenance Analytics: Opportunities

- Provenance beyond reproducibility
- Opportunity for knowledge discovery, sharing and re-use
- Query information
  - Understand processes and data dependencies
  - Find *useful* workflows, e.g., given a piece of data or task, which workflow should we run?
- Mine information
  - Discover *interesting* patterns (e.g., common workflow patterns) → *recommendation system, discover analogies*
  - Identify homogeneous workflow groups by clustering → *organize collections* [Santos et al., IPAW 2008]
  - *Infer workflow specification* from execution log [Aalst et al., TKDE 2004]

Guidance in Workflow Design

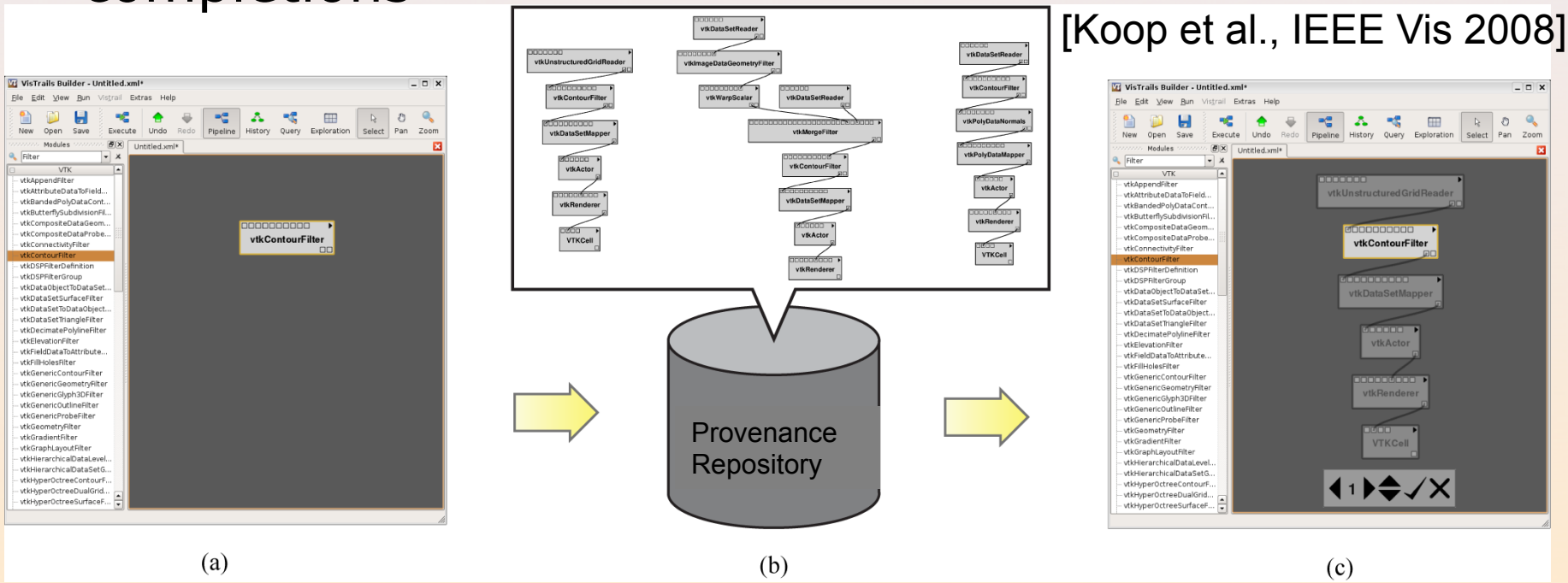


Guidance in Workflow Design

VisComplete: A Workflow Recommendation System

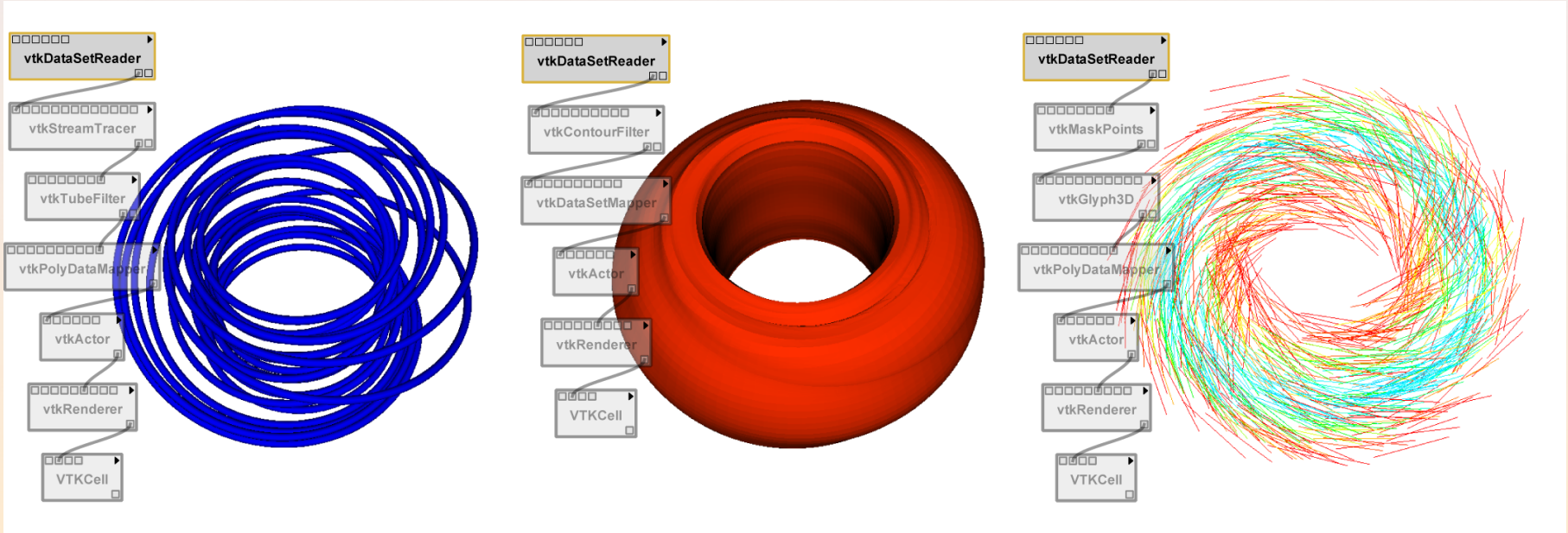
- *Mine* graph fragments that co-occur in a provenance collection
- Predict sets of likely workflow additions to a given partial workflow
- Similar to a Web browser suggesting URL completions

[Koop et al., IEEE Vis 2008]



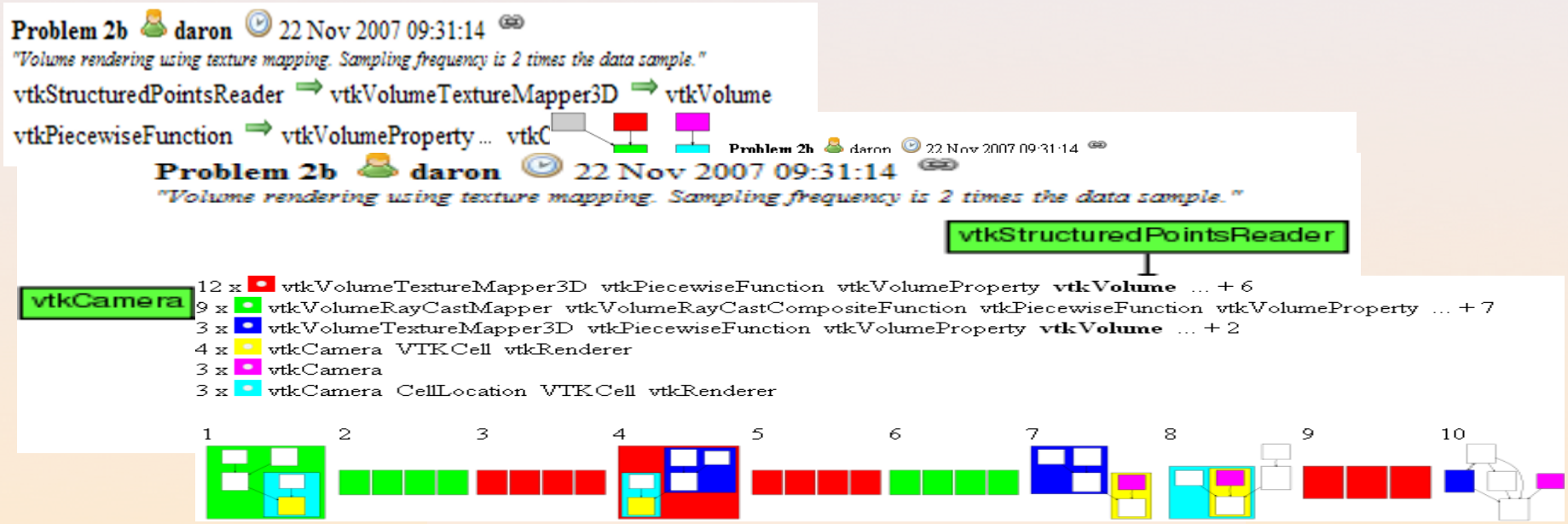
VisComplete: A Workflow Recommendation System

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



Querying Provenance

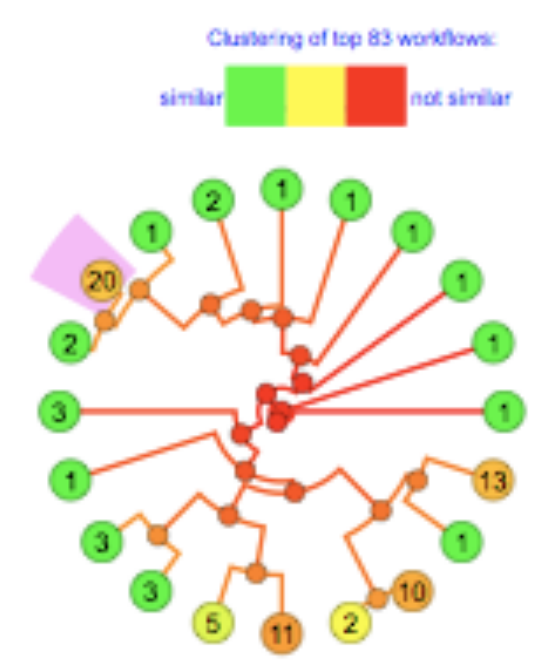
- Provenance is a graph
- Visual interfaces to specify queries [Beeri et al., VLDB 2006, Scheidegger et al., TVCG 2007]
  - WYSIWYQ -- What You See Is What You Query
- Visual interfaces to explore the results [Ellkvist et al., KEYS 2009]



Generate descriptive snippets

Querying Provenance

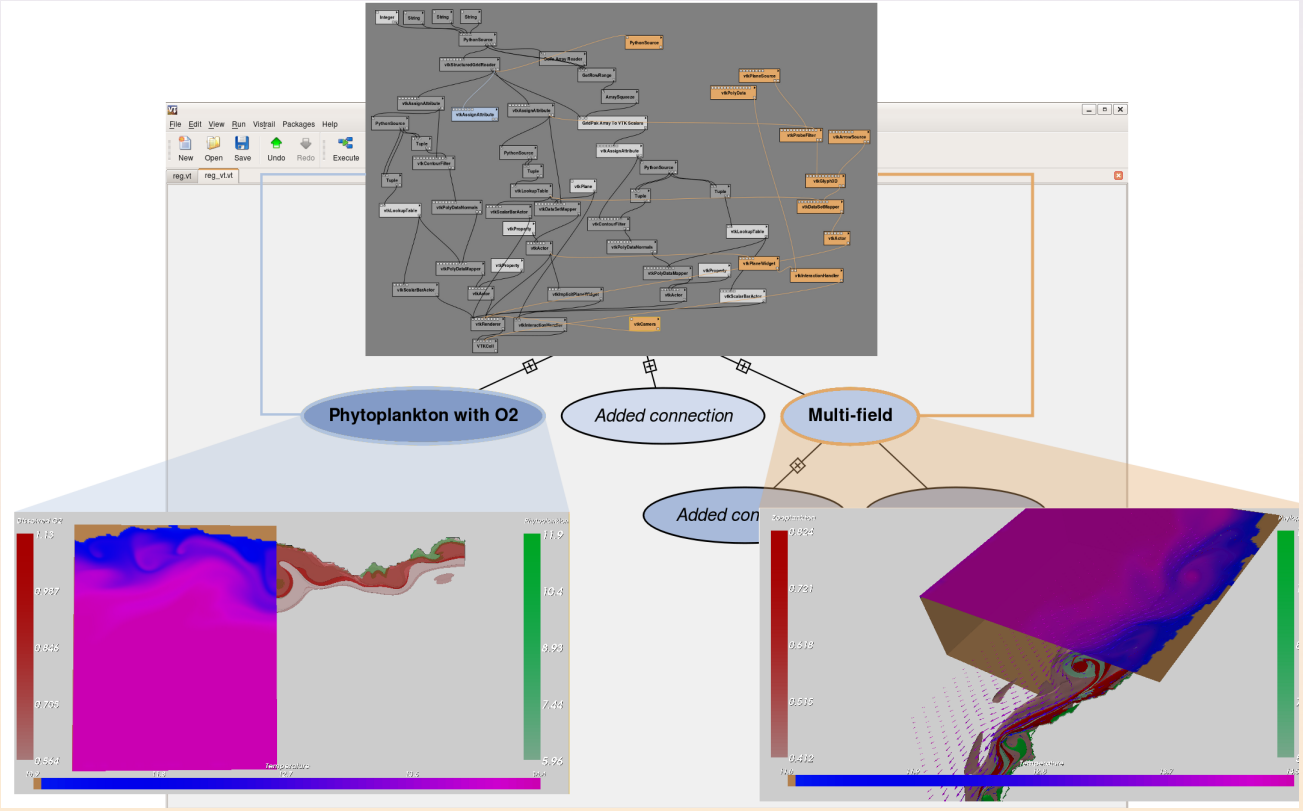
- Provenance is a graph
- Visual interfaces to specify queries [Beeri et al., VLDB 2006, Scheidegger et al., TVCG 2007]
  - WYSIWYQ -- What You See Is What You Query
- Visual interfaces to explore the results [Ellkvist et al., KEYS 2009]



Summarize collection by clustering

Comparing Results

- Ability to compare data products and corresponding workflows



[Freire et al., IPAW 2006]

Mining Provenance: Challenges

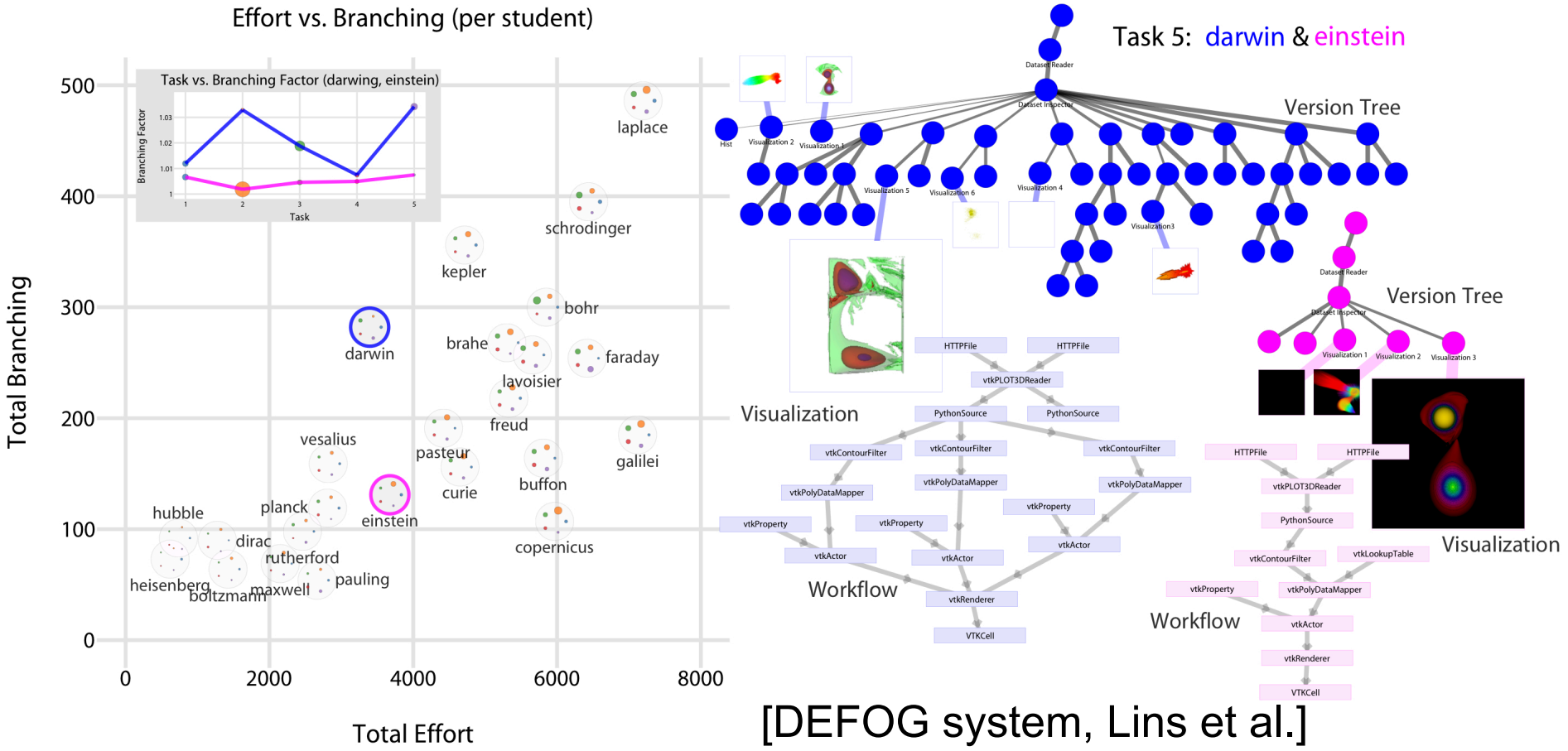
- Provenance is a graph: mining is expensive
- Workflow structure is complex
  - Modules with parameters+values
  - Typed connections
- *How to model provenance?*
  - For clustering, a vector-space based representation produced results correlated to results obtained using a more expensive structural representation [Santos et al., IPAW 2008]
- Which notions of distance and metrics make sense for different applications and data sets?
- Which algorithms are effective and efficient?

[Lauro Lins, Nivan Ferreira. Work in progress]

Mining Provenance: Challenges

- Need analysis/visualization tools

Understanding User Behavior



[DEFOG system, Lins et al.]

Acknowledgments

- This work is partially supported by the National Science Foundation grants IIS 1050422, IIS 0905385, IIS 0844572, IIS 0746500, CNS 0751152,; the Department of Energy, an IBM Faculty Award, and a University of Utah Seed Grant.

Ευχαριστω
Thank you
Obrigada