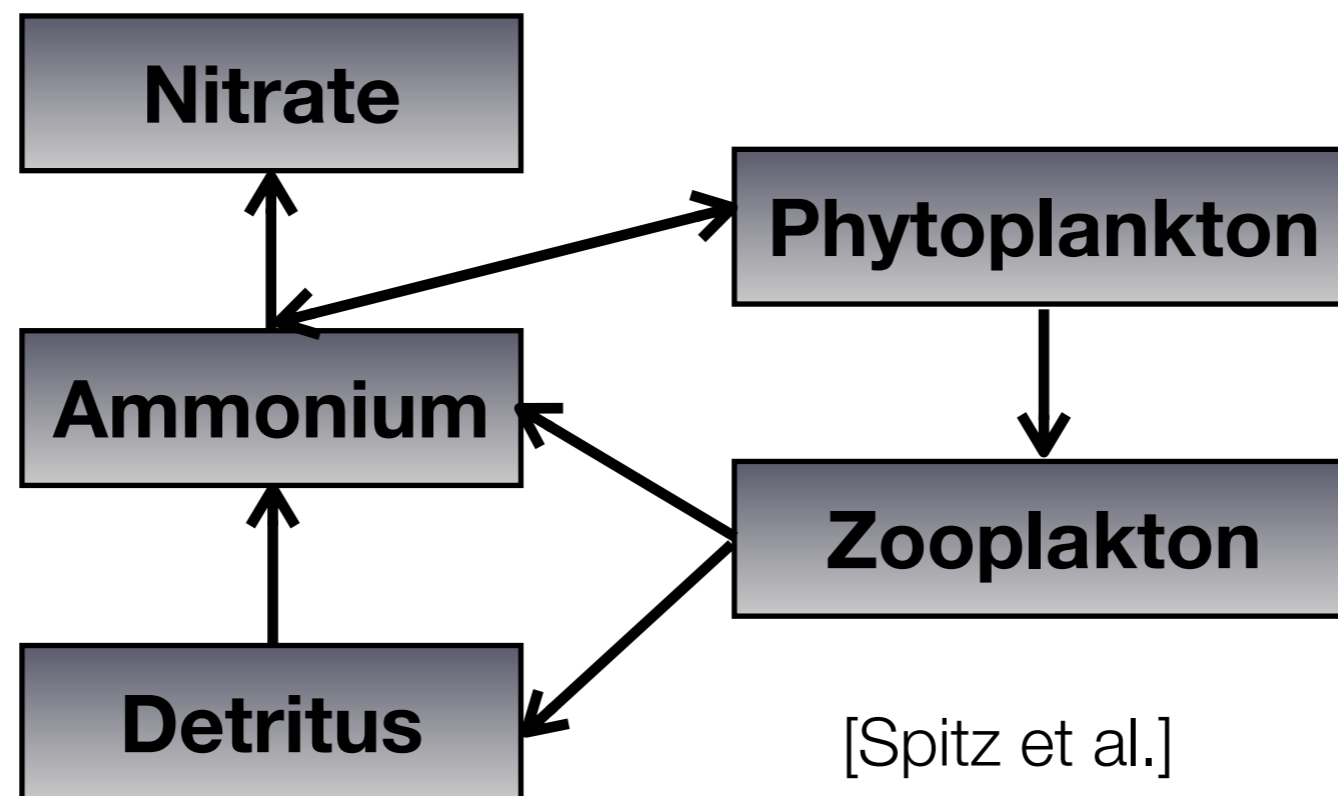


Provenance-Enabled Data Exploration and Visualization with VisTrails

Juliana Freire, Emanuele Santos and Cláudio Silva
SCI Institute
University of Utah

Data Exploration and Visualization: Today

- Scientists need to make sense of increasingly large volumes of data
- Oceanographers: understand nature, environment and their interactions

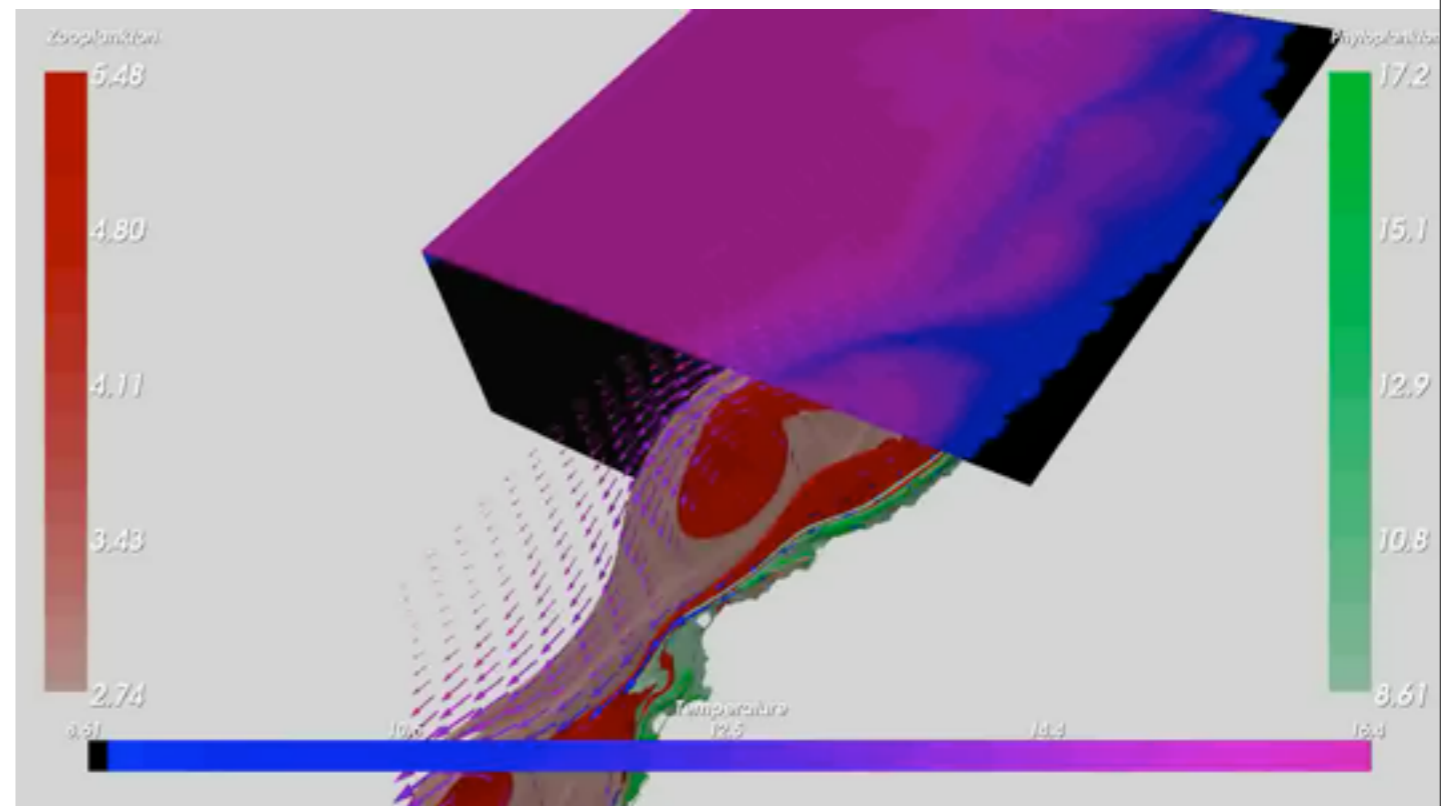


Data Exploration and Visualization: Today

- Scientists need to make sense of increasingly large volumes of data
- Oceanographers: understand nature, environment and their interactions
- Visualization is instrumental to understanding simulation results and obtain insights
 - Identify interesting features
 - Correlate information from multiple models
 - Interactively explore data

Data Exploration and Visualization: Today

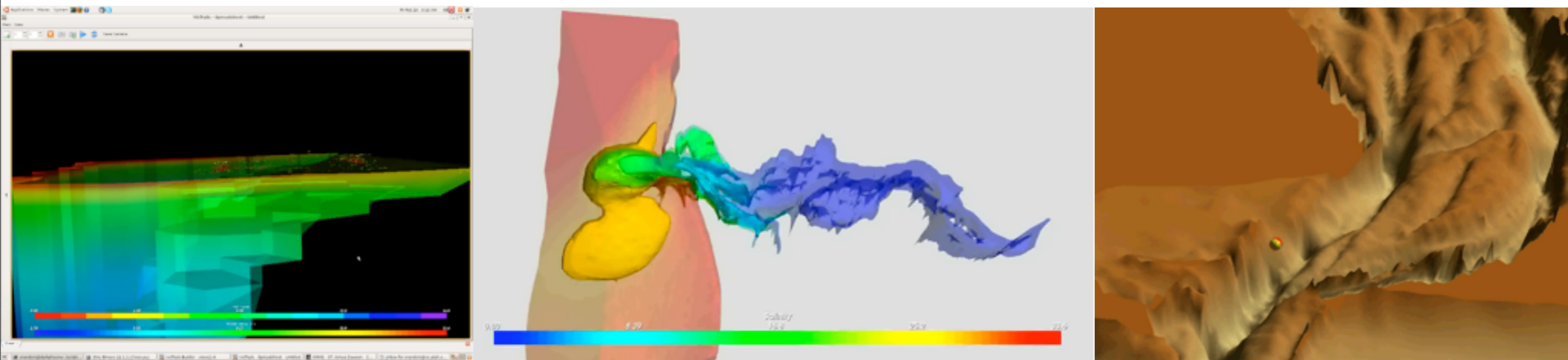
- Scientists need to make sense of increasingly large volumes of data
- Oceanographers: understand nature, environment and their interactions
- Visualization is instrumental to understanding simulation results and obtain insights
 - Identify interesting features
 - Correlate information from multiple models
 - Interactively explore data



Challenges: Volume and Complexity

- Observation and modeling of multiple systems at multiple scales
 - Scientists need to collaborate: biologists, chemists, oceanographers, computer scientists
- Very large number of data products, sensor measurements, and results from numerical models
- Cover more than 10 years of experiments: occupy over 30 TB of storage

[CMOP Simulations - OHSU, OSU, CROOS]

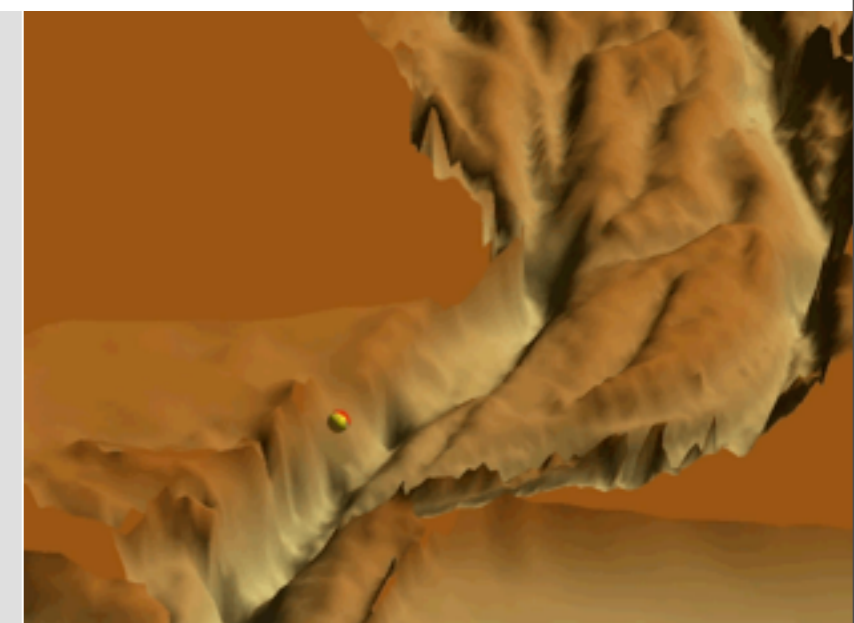
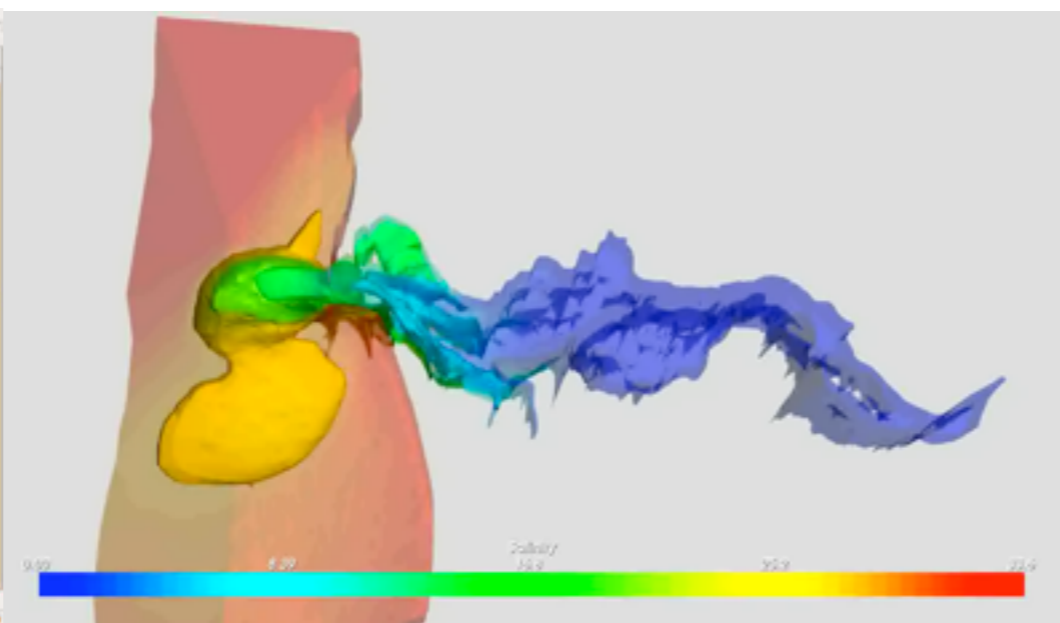
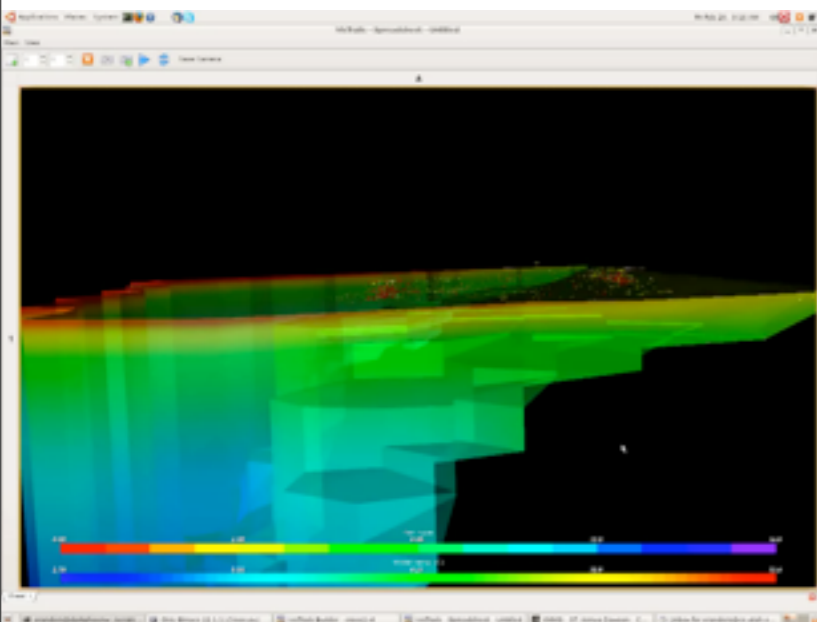


Challenges: Volume and Complexity

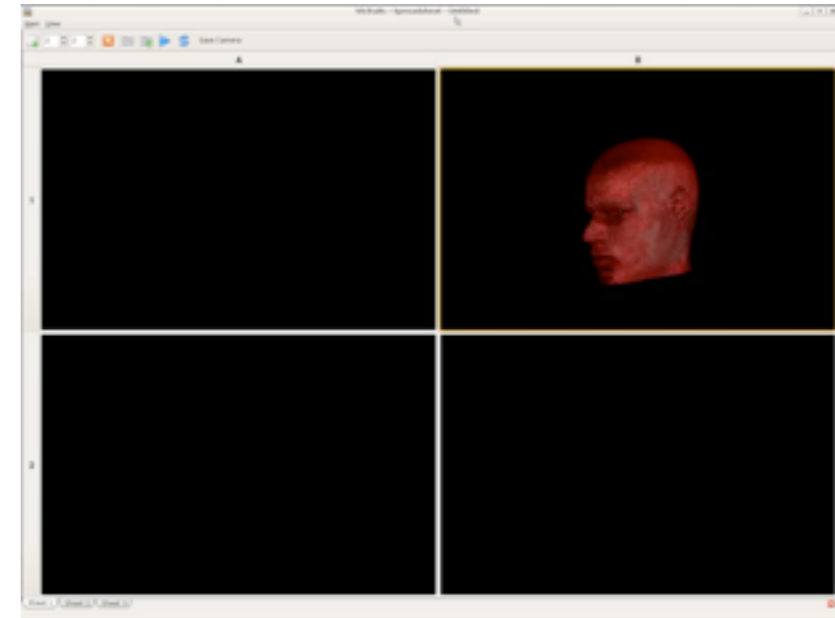
- Observation and modeling of multiple systems at multiple scales
 - Scientists need to collaborate: biologists, chemists, oceanographers, computer scientists

- Very large data sets
and
Need tools to support the data exploration and visualization

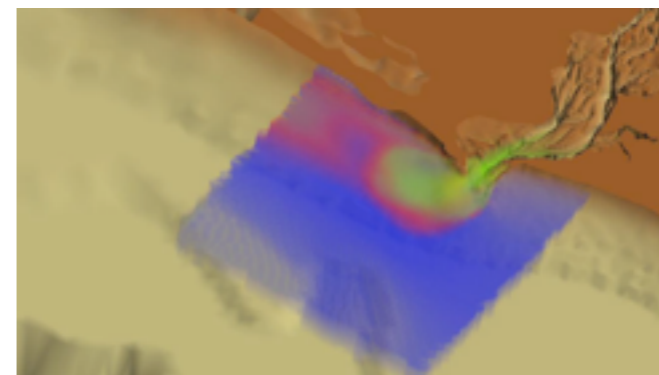
- Complexity of data
of
Need data and process management



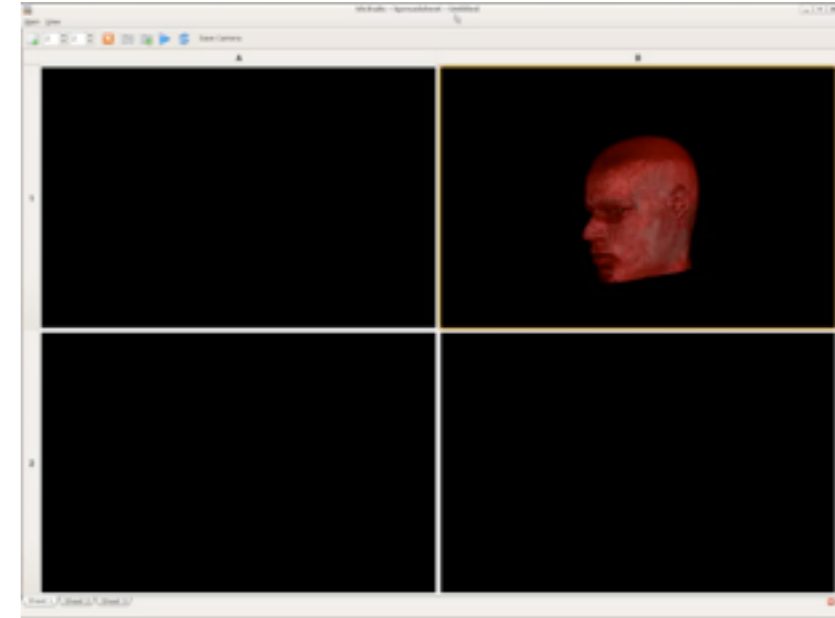
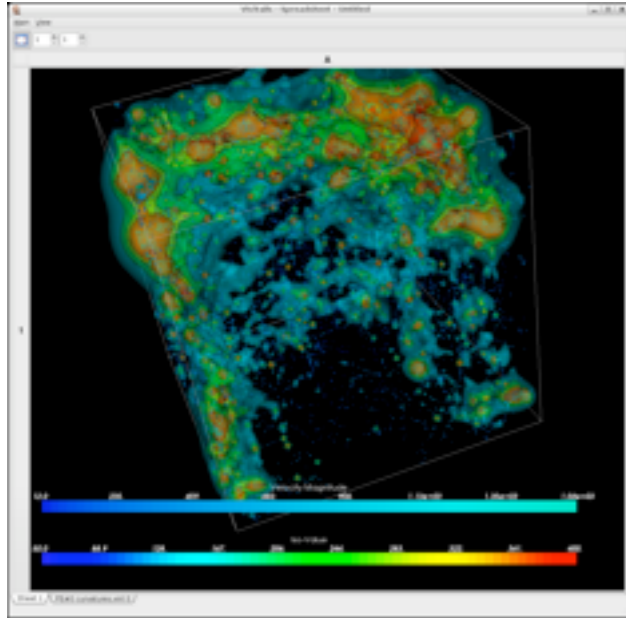
The need for provenance



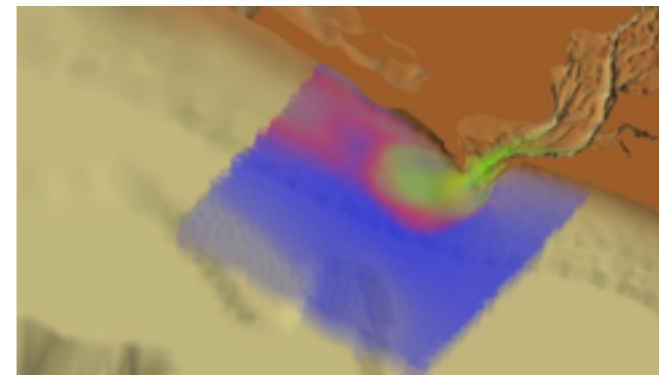
Emp	Dep	Mgr
John	D01	Mary
Susan	D02	Ken



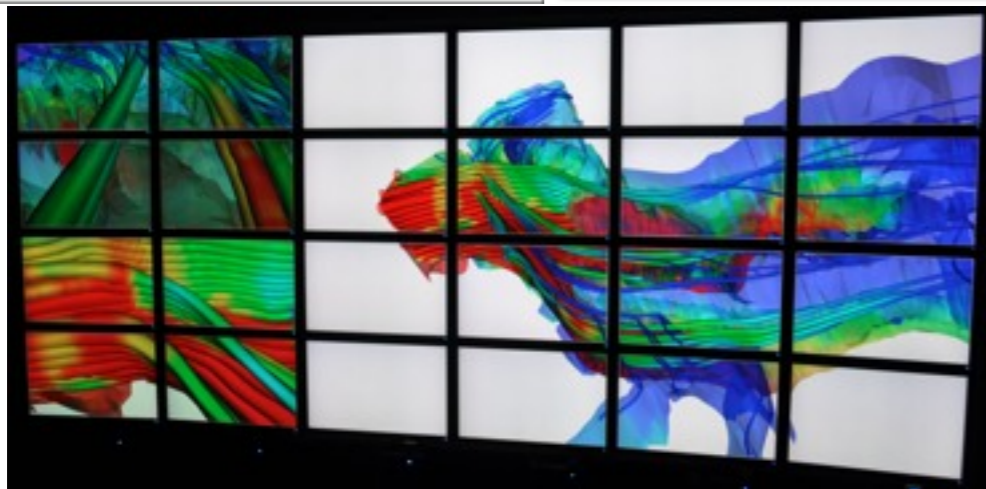
The need for provenance



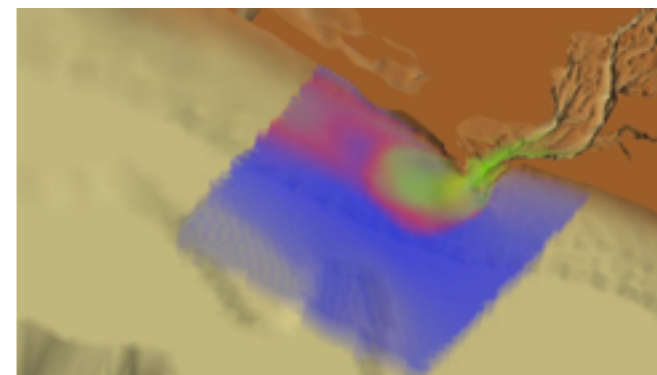
Emp	Dep	Mgr
John	D01	Mary
Susan	D02	Ken



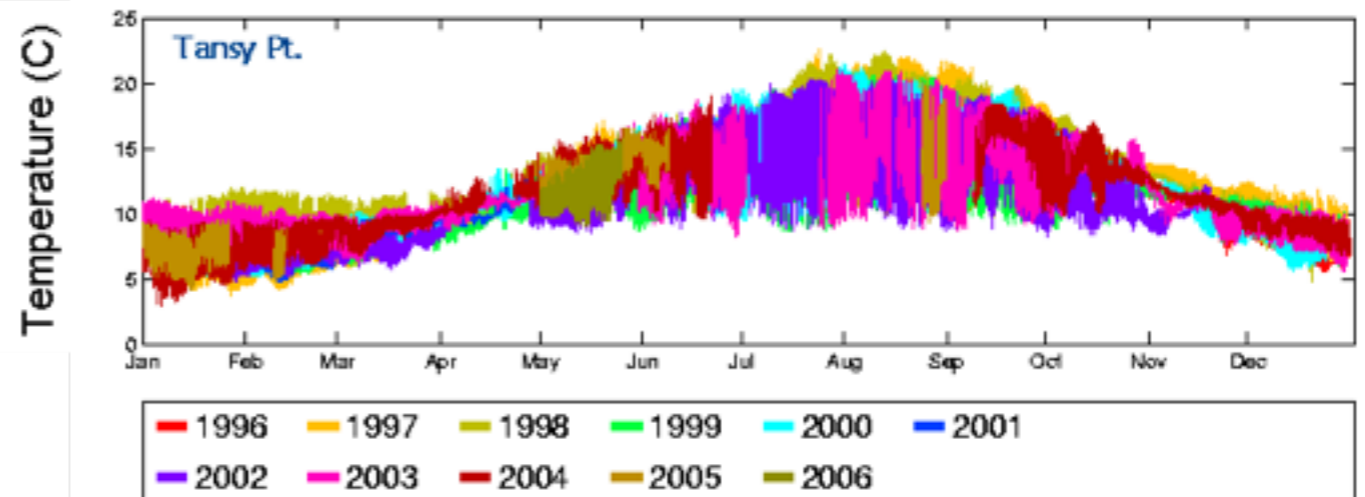
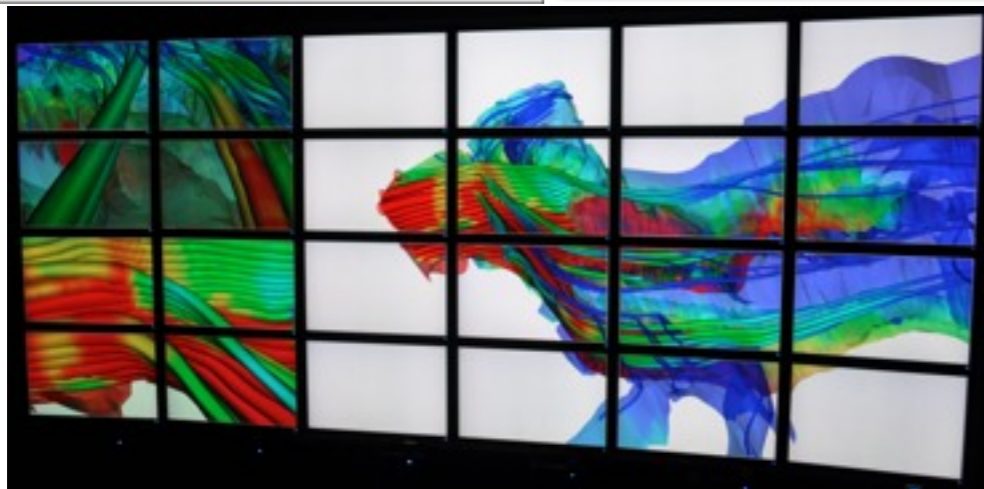
The need for provenance



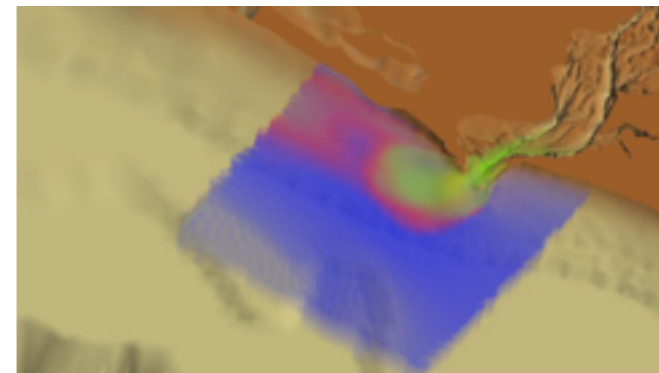
Emp	Dep	Mgr
John	D01	Mary
Susan	D02	Ken



The need for provenance



Emp	Dep	Mgr
John	D01	Mary
Susan	D02	Ken

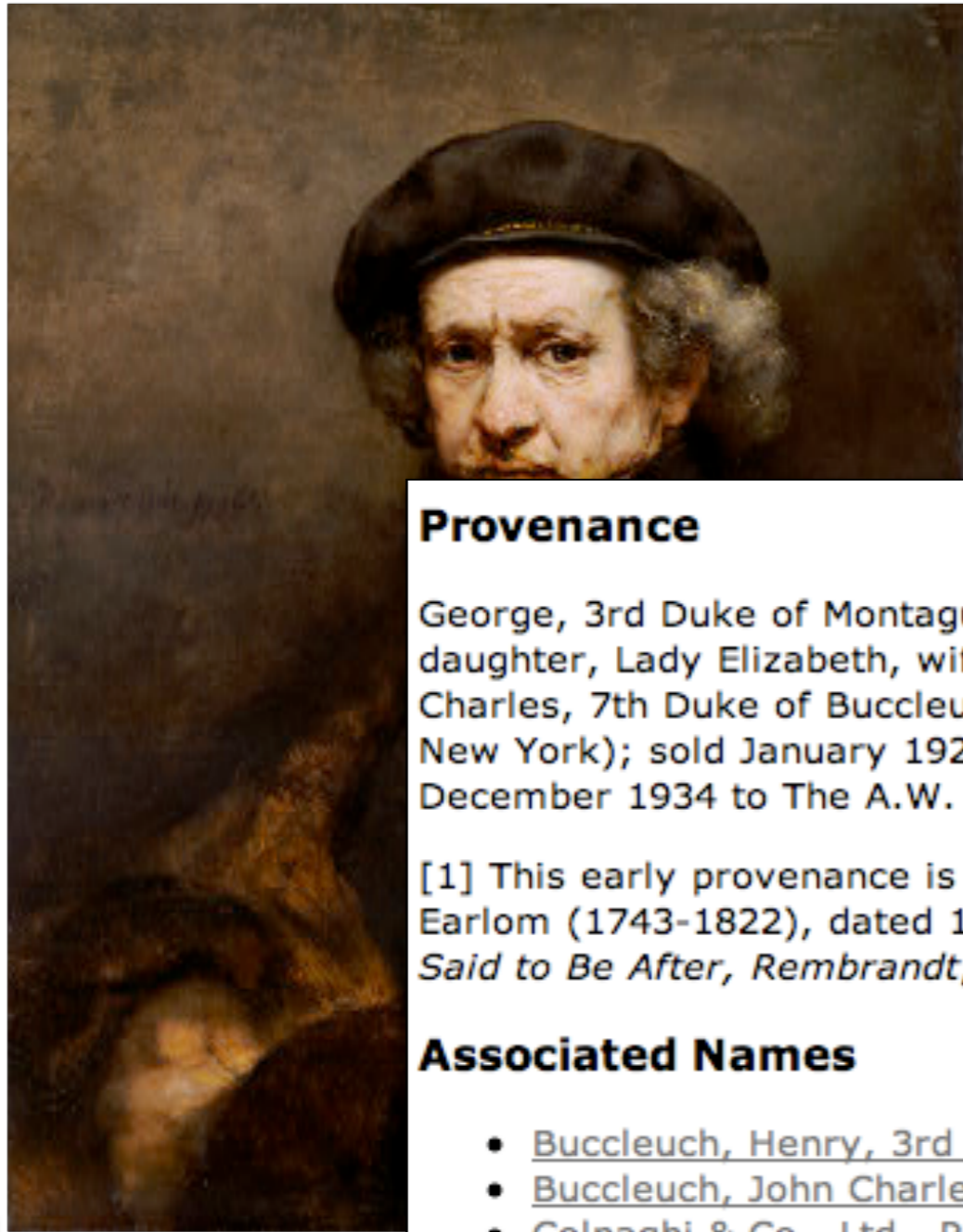


Provenance in Art



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

Provenance in Art



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

Provenance

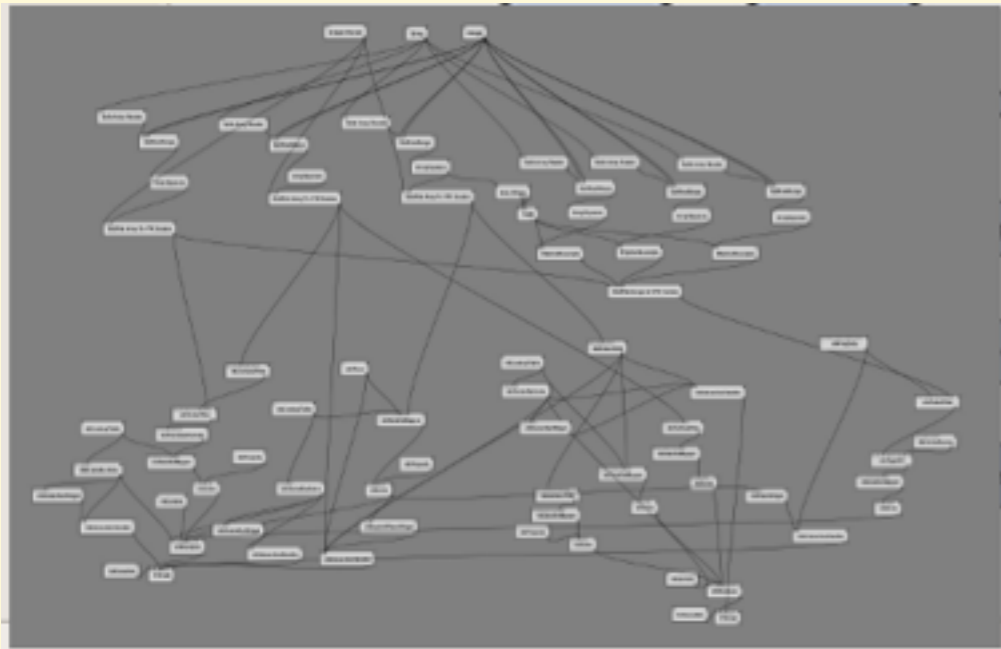
George, 3rd Duke of Montagu and 4th Earl of Cardigan [d. 1790], by 1767; [1] by inheritance to his daughter, Lady Elizabeth, wife of Henry, 3rd Duke of Buccleuch of Montagu House, London; John Charles, 7th Duke of Buccleuch; (P. & D. Colnaghi & Co., New York, 1928); (M. Knoedler & Co., New York); sold 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.

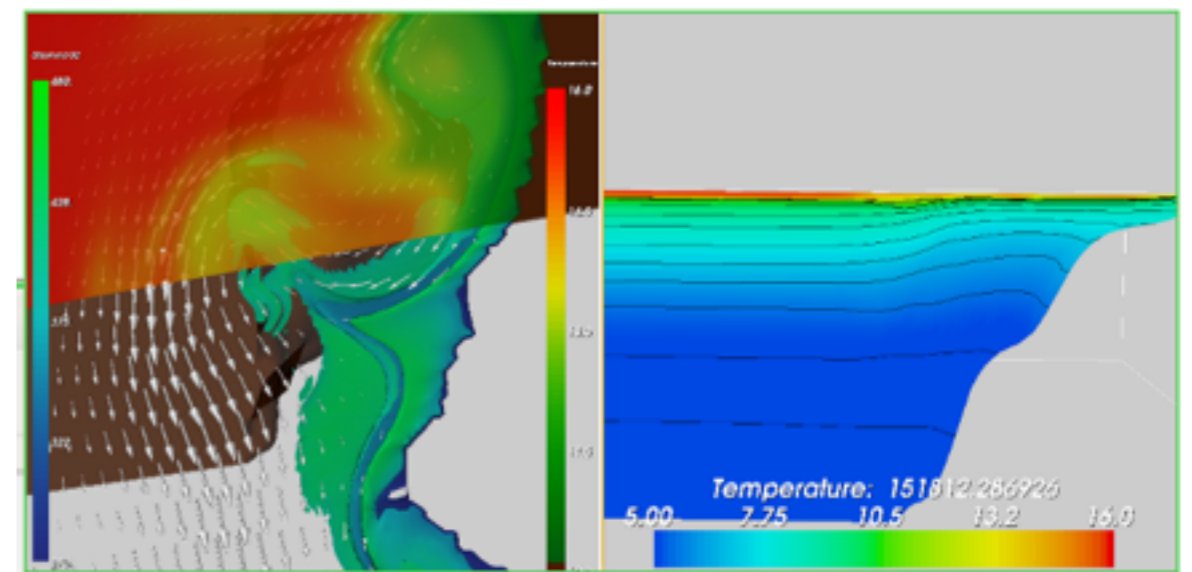
Associated Names

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

Provenance of Digital Data



derived

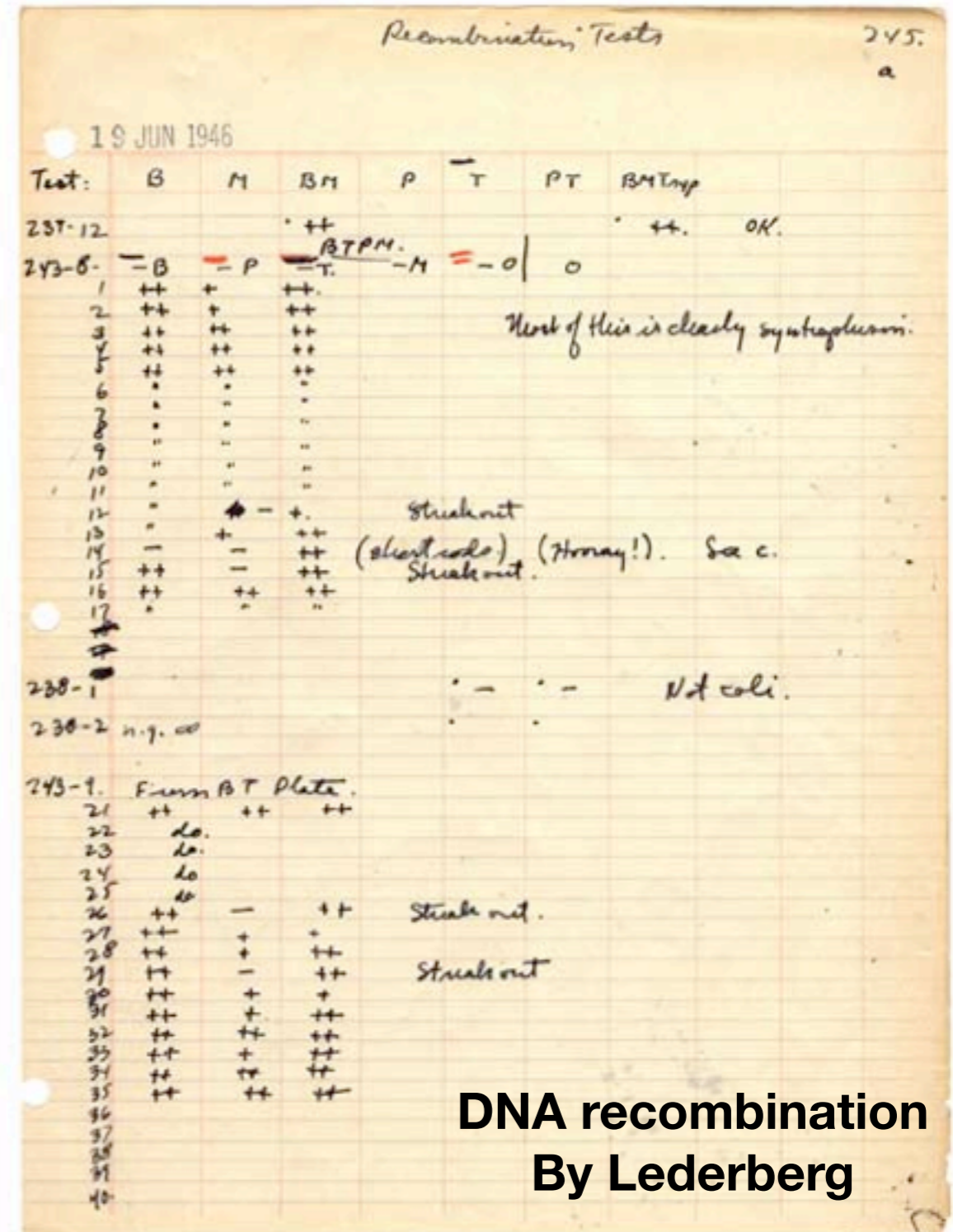


Simulation code
version, visualization
pipeline, parameters,
creator, date,
machines used, ...

Provenance

Provenance in Science

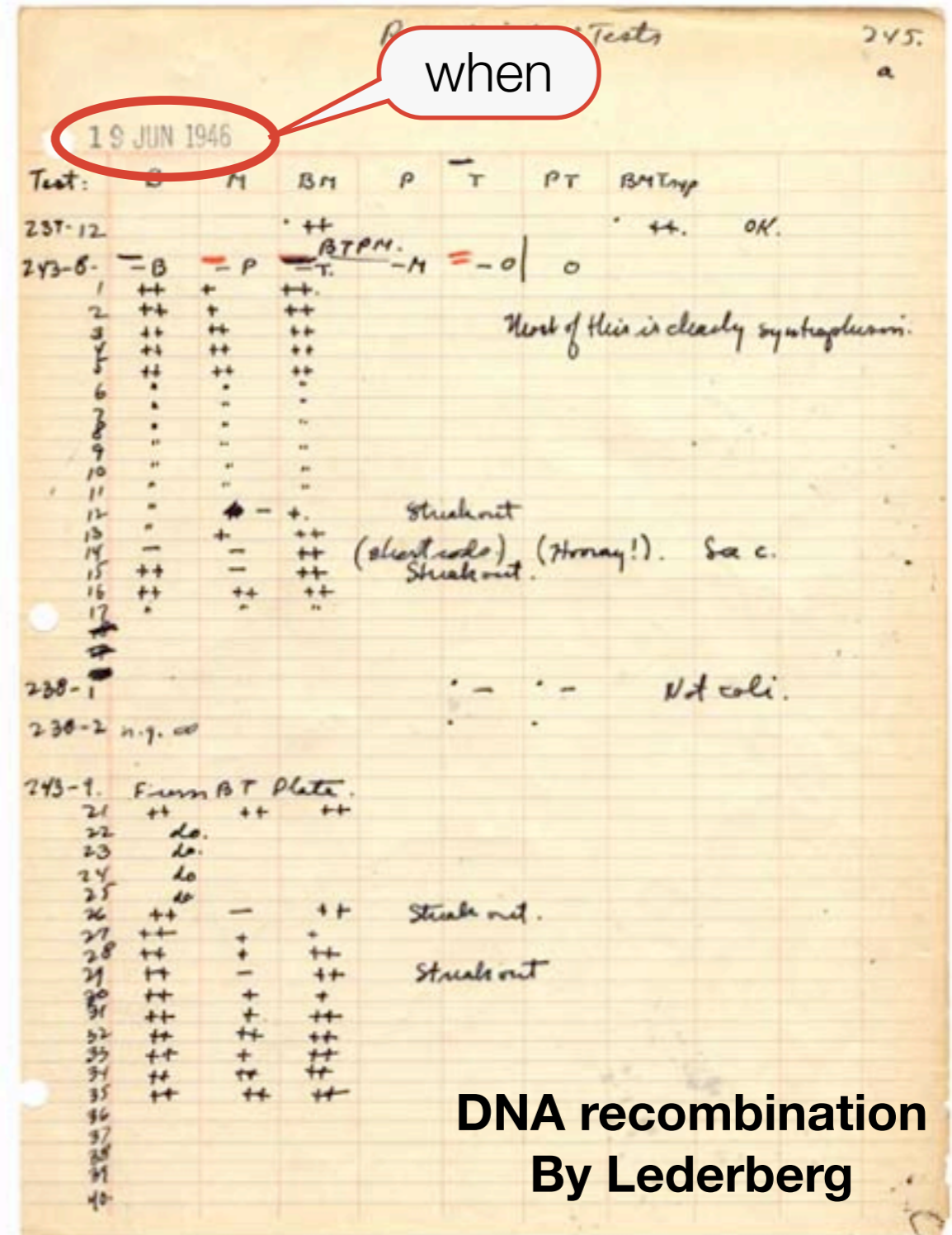
- Provenance is as (or more!) important as the result
- Not a new issue
- Lab notebooks have been used for a long time
- What is new?
 - Large volumes of data
 - Complex analyses - computational processes
- Writing notes is no longer an option: need systematic provenance capture



**DNA recombination
By Lederberg**

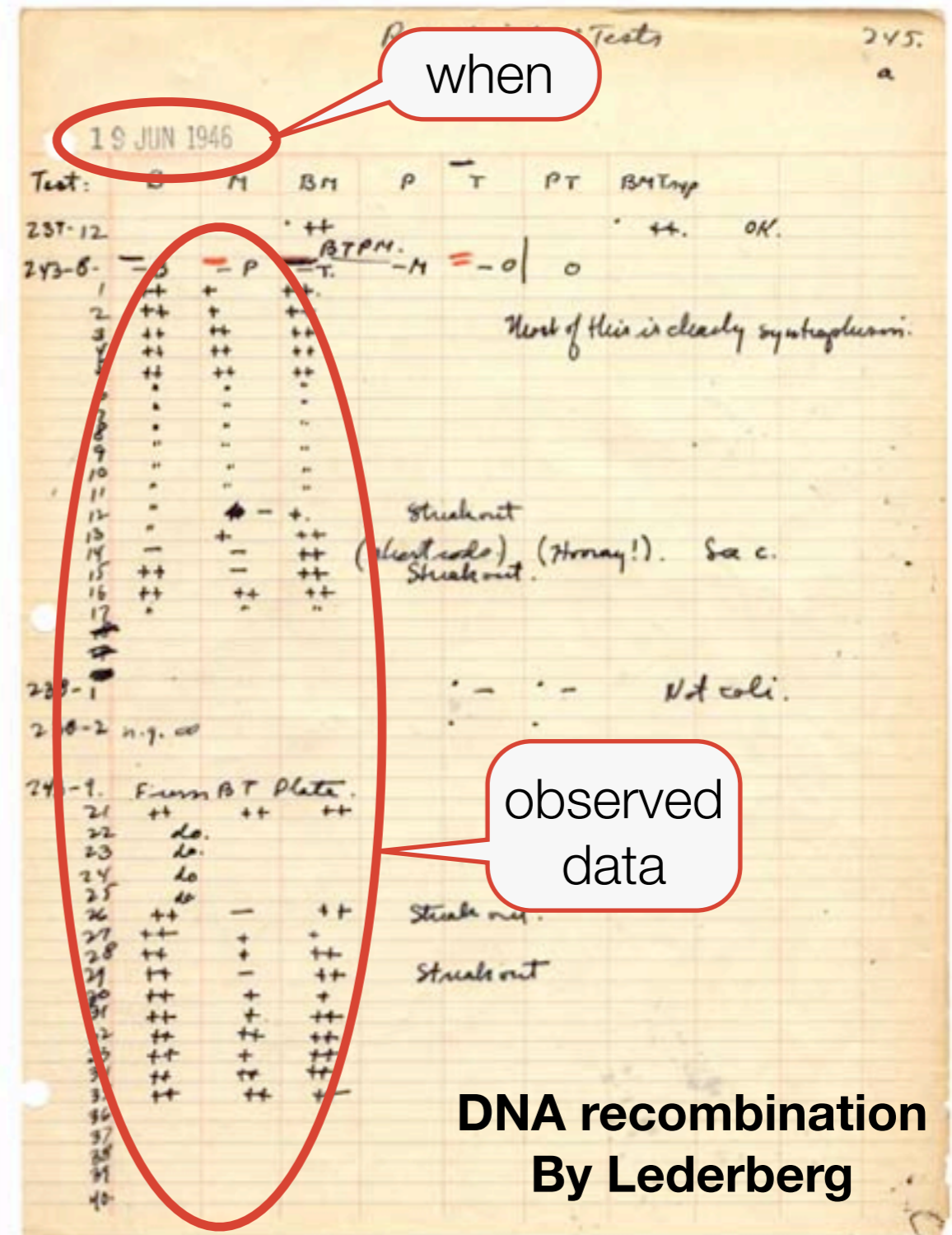
Provenance in Science

- Provenance is as (or more!) important as the result
- 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: need systematic provenance capture



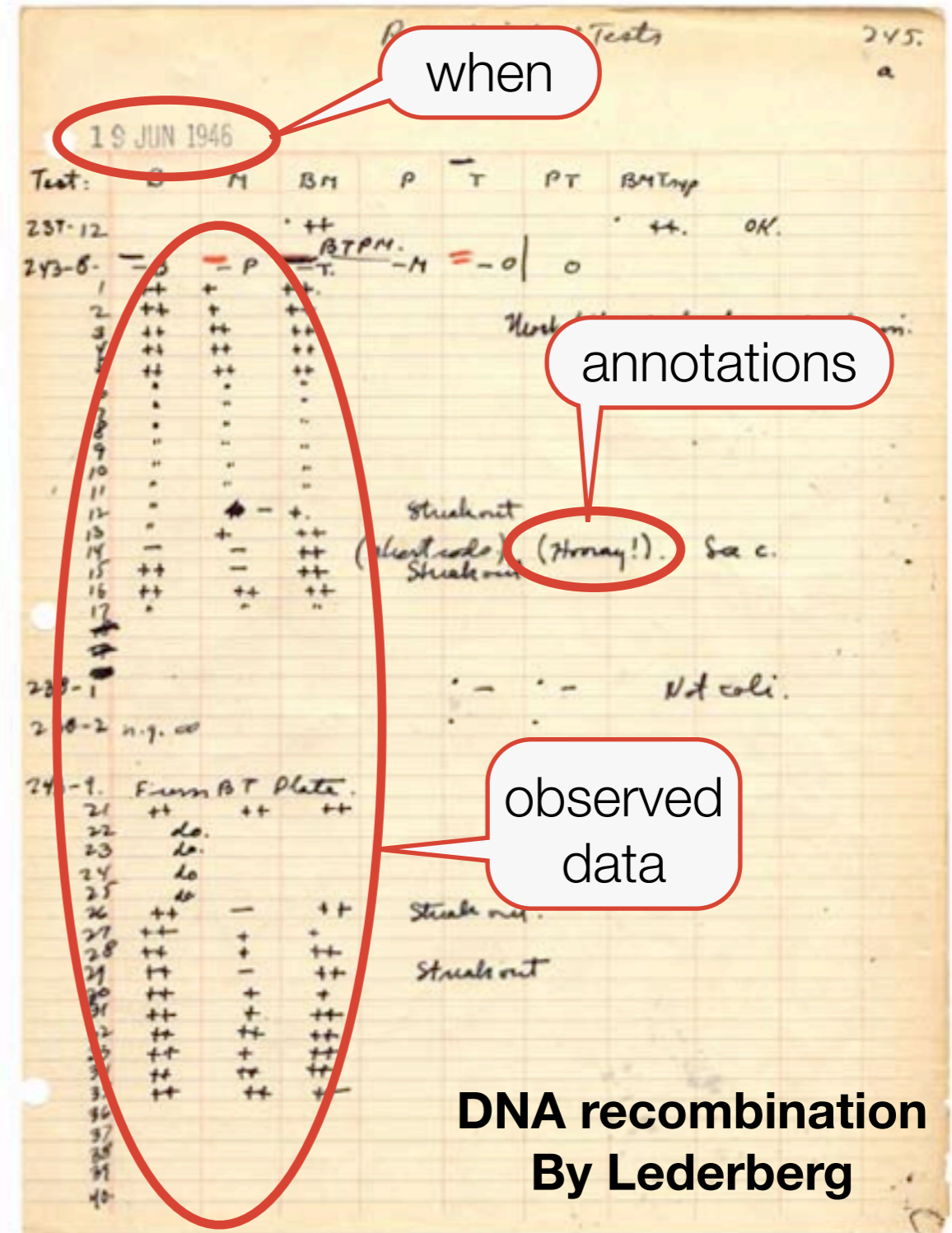
Provenance in Science

- Provenance is as (or more!) important as the result
- 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: need systematic provenance capture



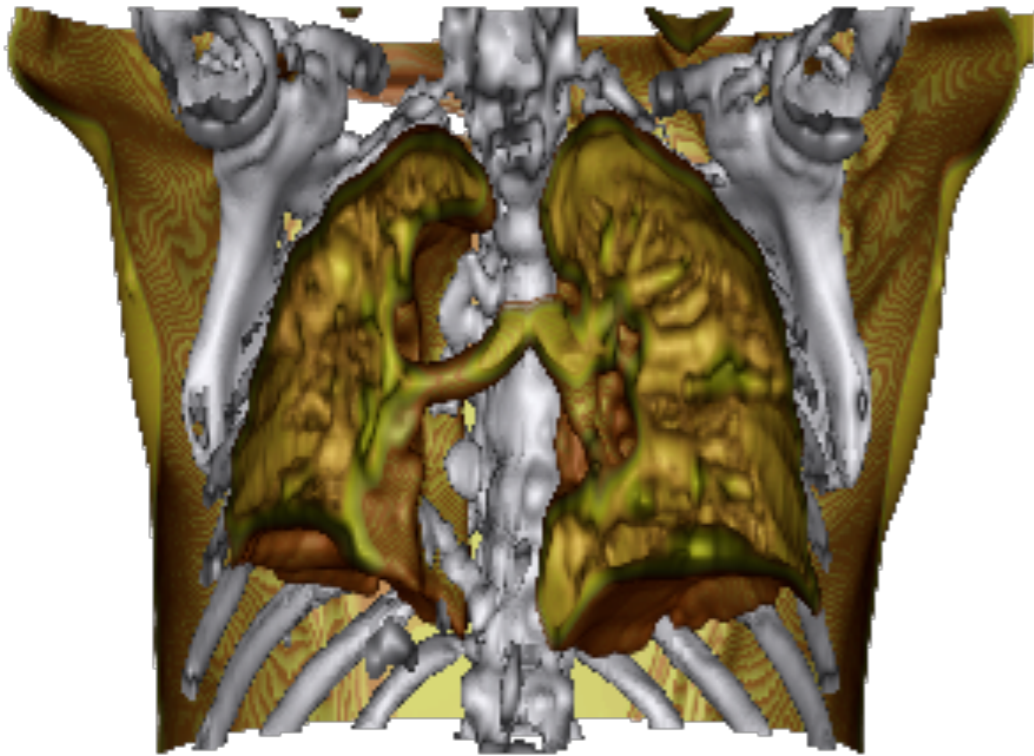
Provenance in Science

- Provenance is as (or more!) important as the result
- 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: need systematic provenance capture

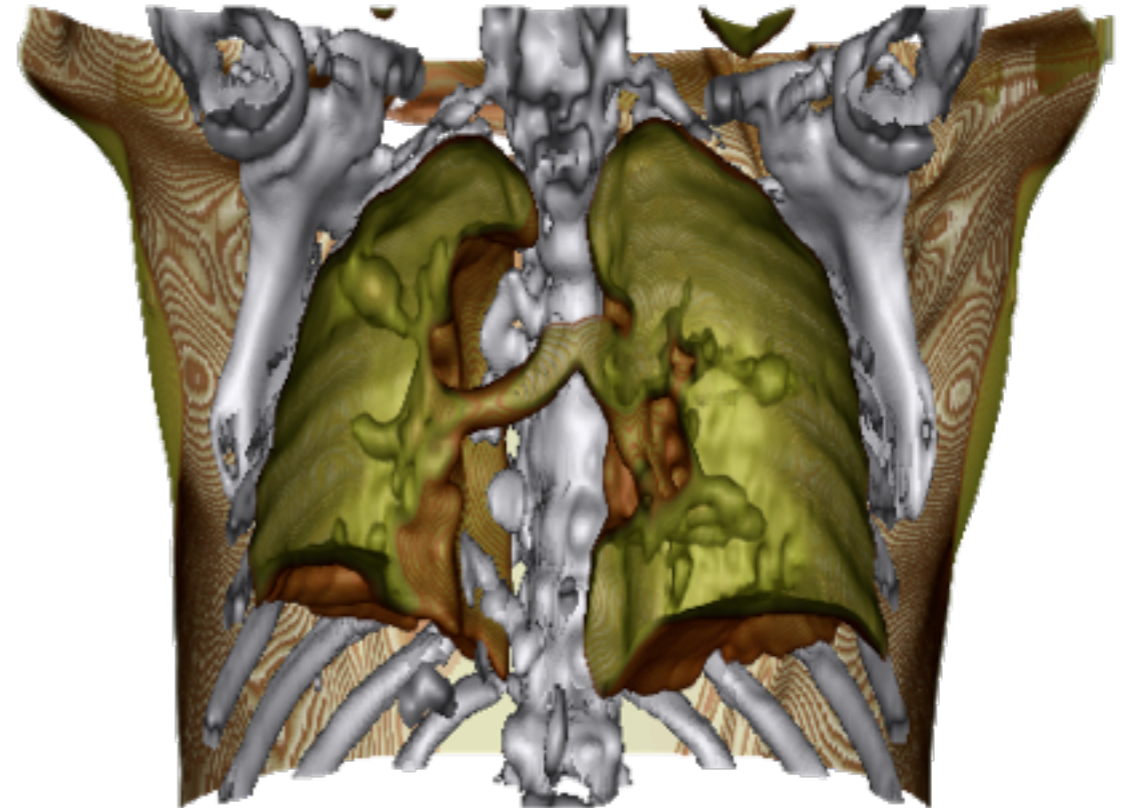


Uses of Computational Provenance

anon4877_zspace_20060331.jpg



anon4877_lesion_20060401.jpg



How were these images created?

Was any pre-processing applied to the raw data?

Who created them? What's the difference?

Are they really from the same patient?

Uses of Computational Provenance

anon4877_zspace_20060331.jpg

anon4877_lesion_20060401.jpg

Reproducibility



How were these images created?

Was any pre-processing applied to the raw data?

Who created them? What's the difference?

Are they really from the same patient?

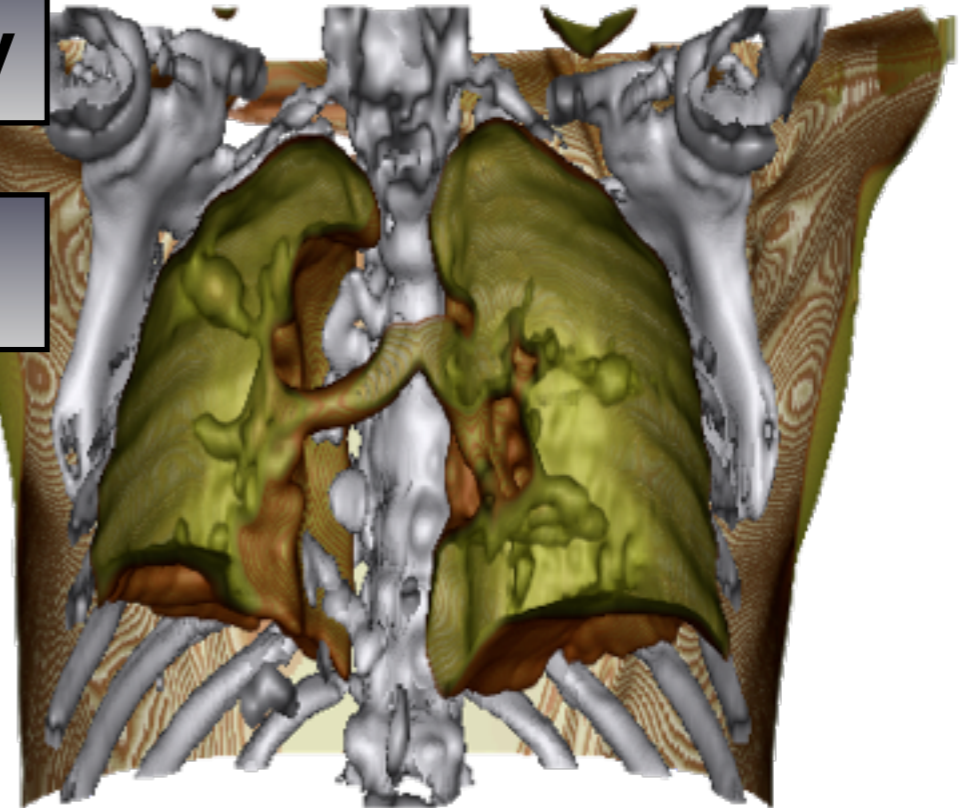
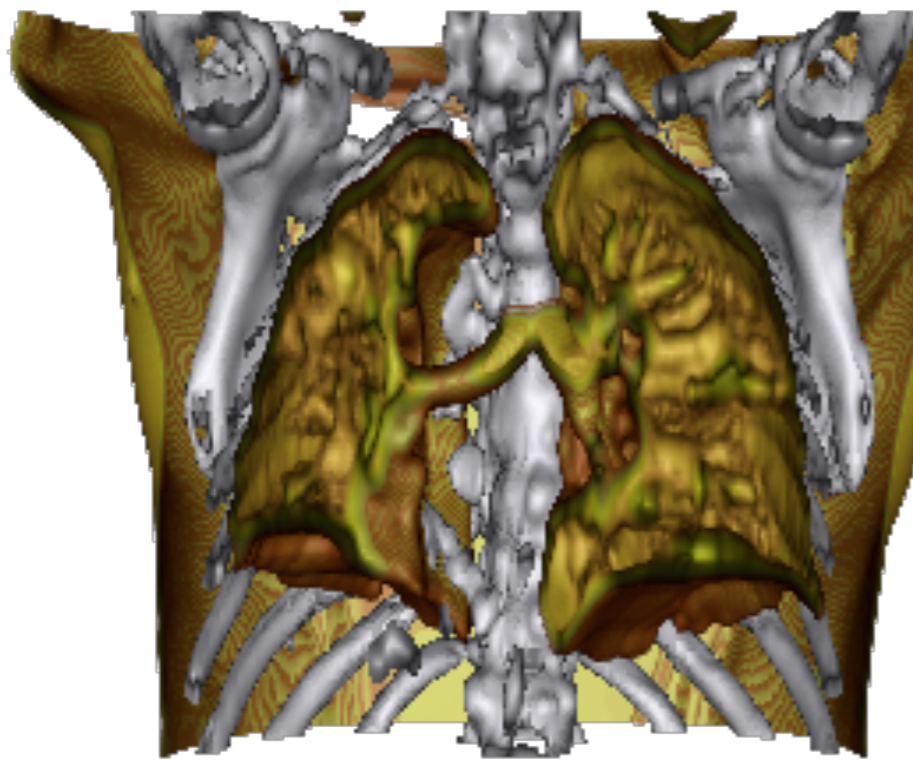
Uses of Computational Provenance

anon4877_zspace_20060331.jpg

anon4877_lesion_20060401.jpg

Reproducibility

Data quality



How were these images created?

Was any pre-processing applied to the raw data?

Who created them? What's the difference?

Are they really from the same patient?

Uses of Computational Provenance

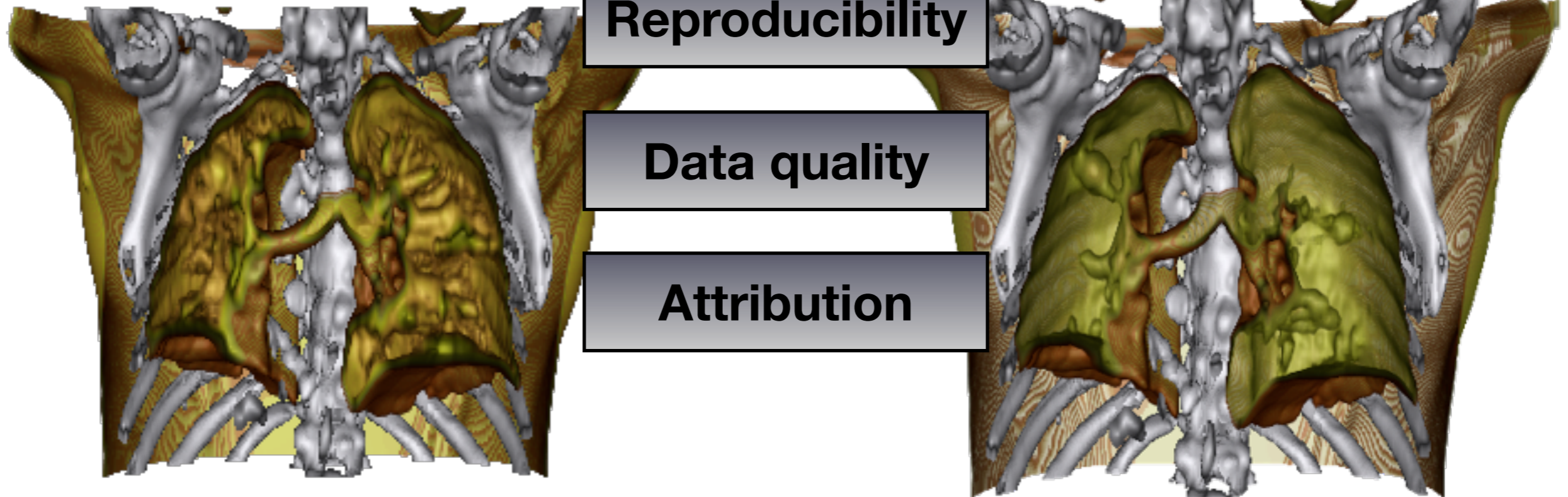
anon4877_zspace_20060331.jpg

anon4877_lesion_20060401.jpg

Reproducibility

Data quality

Attribution



How were these images created?

Was any pre-processing applied to the raw data?

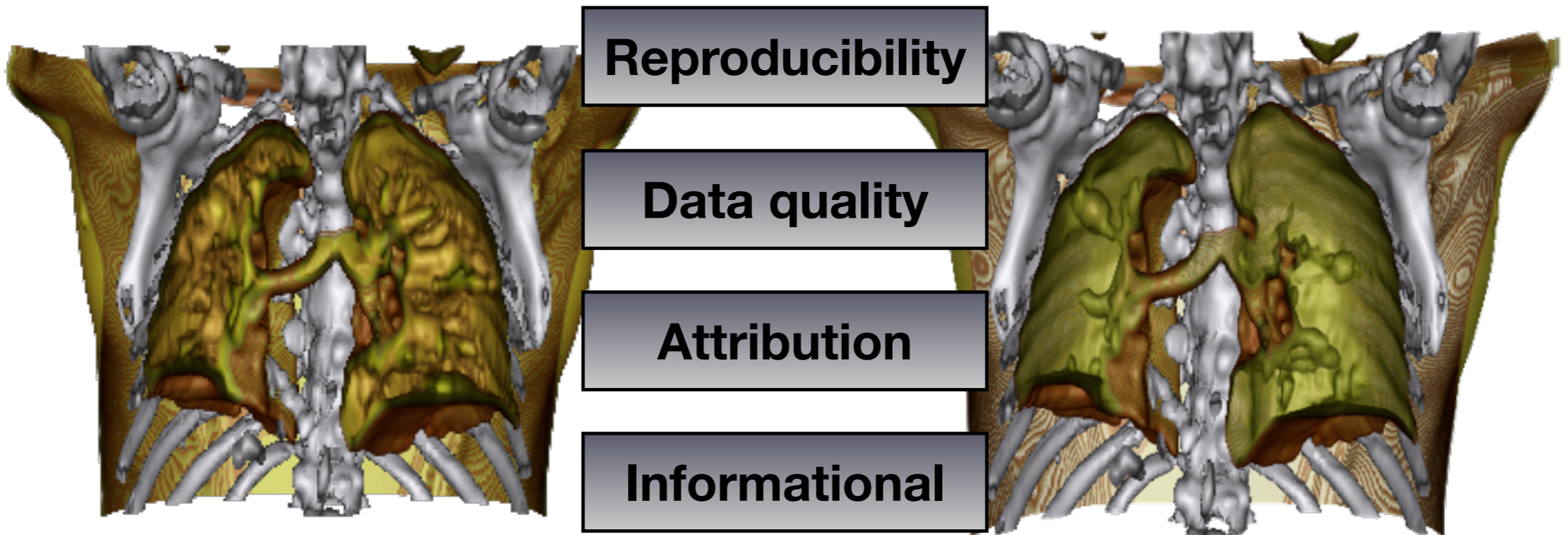
Who created them? What's the difference?

Are they really from the same patient?

Uses of Computational Provenance

anon4877_zspace_20060331.jpg

anon4877_lesion_20060401.jpg



How were these images created?

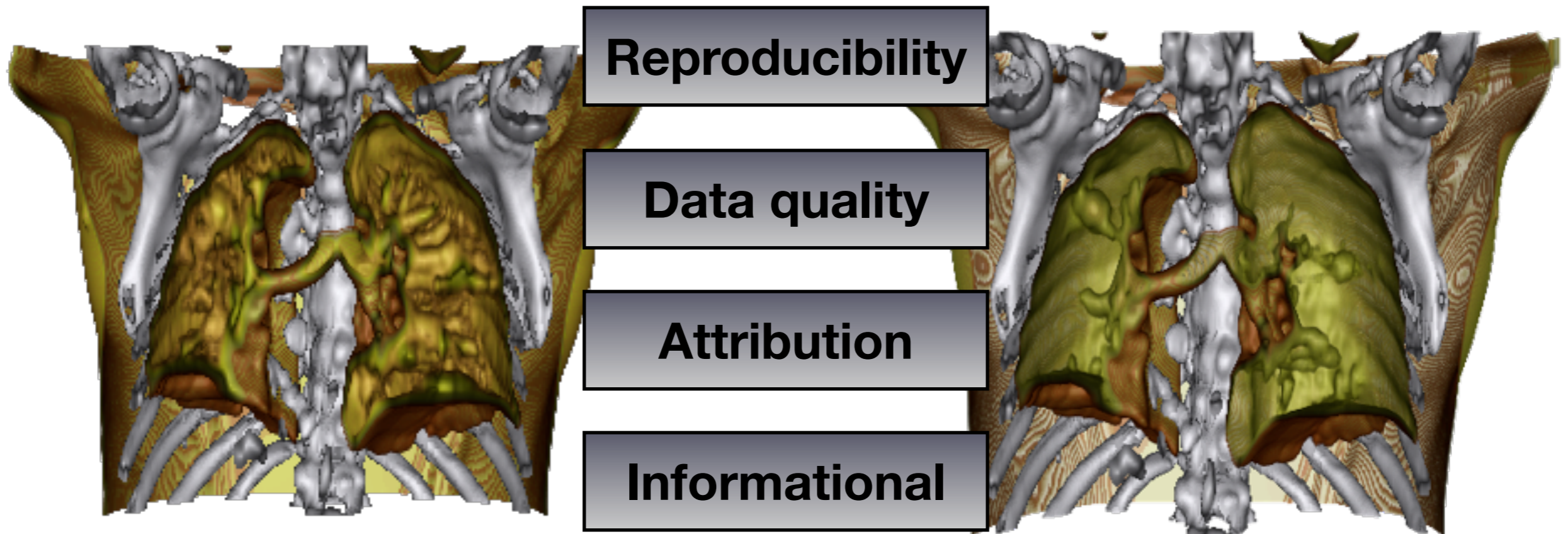
Was any pre-processing applied to the raw data?

Who created them? What's the difference?

Are they really from the same patient?

Uses of Computational Provenance

anon4877_lesion_20060401.jpg



How were these images created?

Was any pre-processing applied to the raw data?

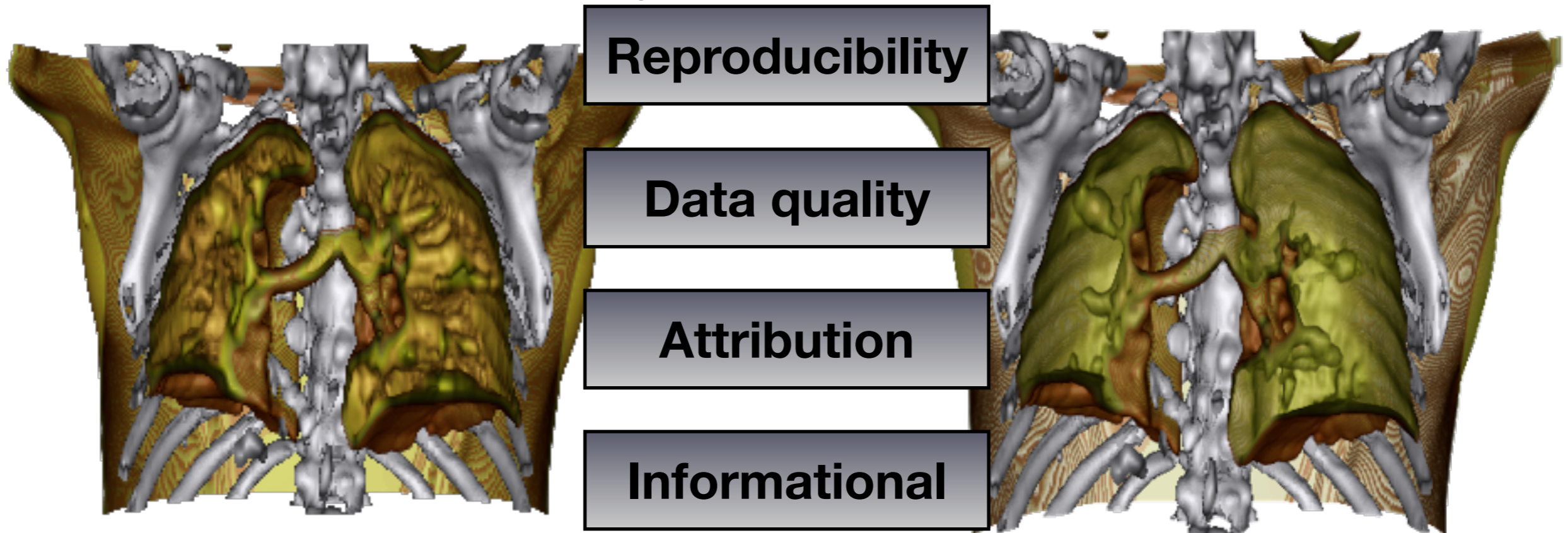
Who created them? What's the difference?

Are they really from the same patient?

Uses of Computational Provenance

anon4876_zspace_20060331.jpg

anon4877_lesion_20060401.jpg



How were these images created?

Was any pre-processing applied to the raw data?

Who created them? What's the difference?

Are they really from the same patient?

Tutorial Goals

- Using the VisTrails environment, show how scientific workflows and provenance can be used to support data analysis and visualization
- Discuss emerging applications that are enabled by provenance
- VisTrails tutorial: Cover concrete examples on how to use VisTrails for data exploration
 - Plugin creation
 - Running VisTrails as a server
 - Executing parameter exploration in VisTrails using the command line

The VisTrails System

- Initial motivation: Visualization meets provenance
- More than visualization---supports computational tasks in general: Workflows meet provenance
- Focus on exploratory tasks such as simulations, visualization and data mining

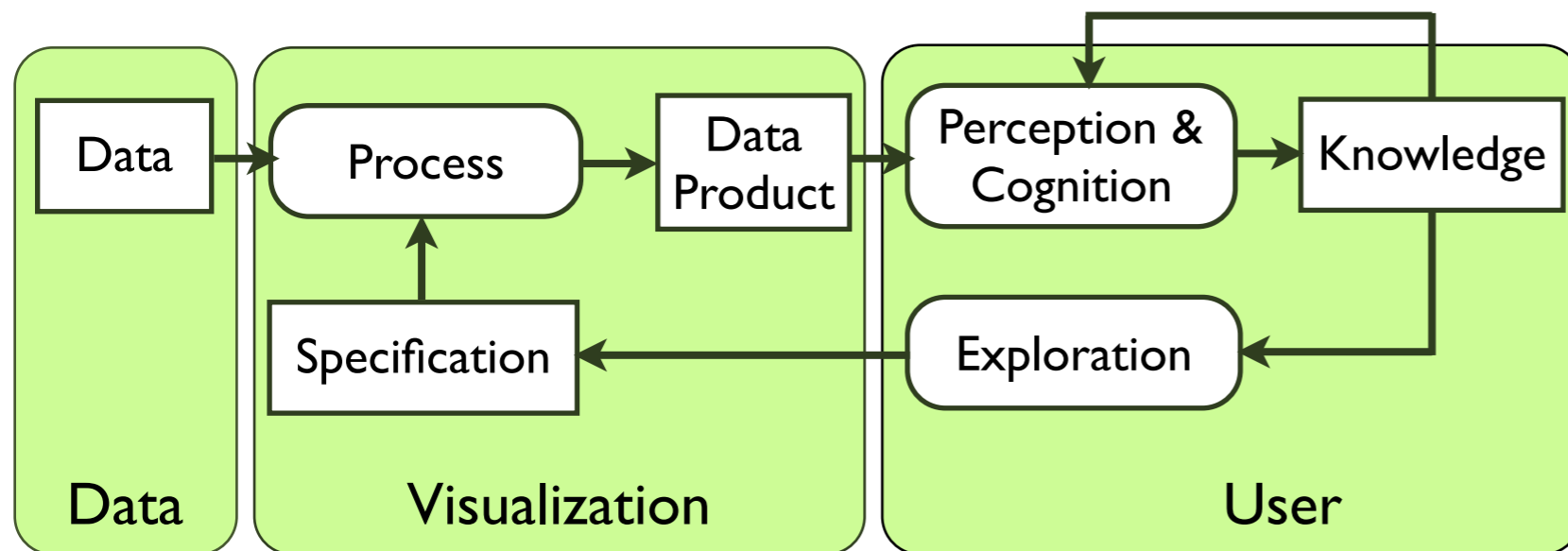


Figure modified from J. van Wijk, IEEE Vis2005

The VisTrails System

- Initial motivation: Visualization meets provenance
- More than visualization---supports computational tasks in general: Workflows meet provenance
- Focus on exploratory tasks such as simulations, visualization and data mining

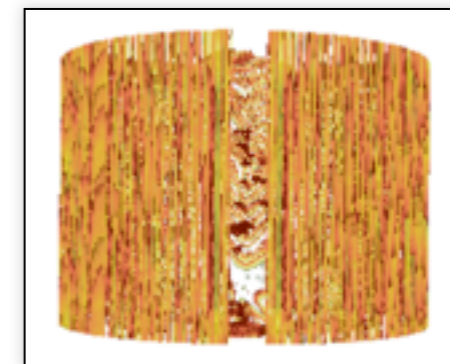
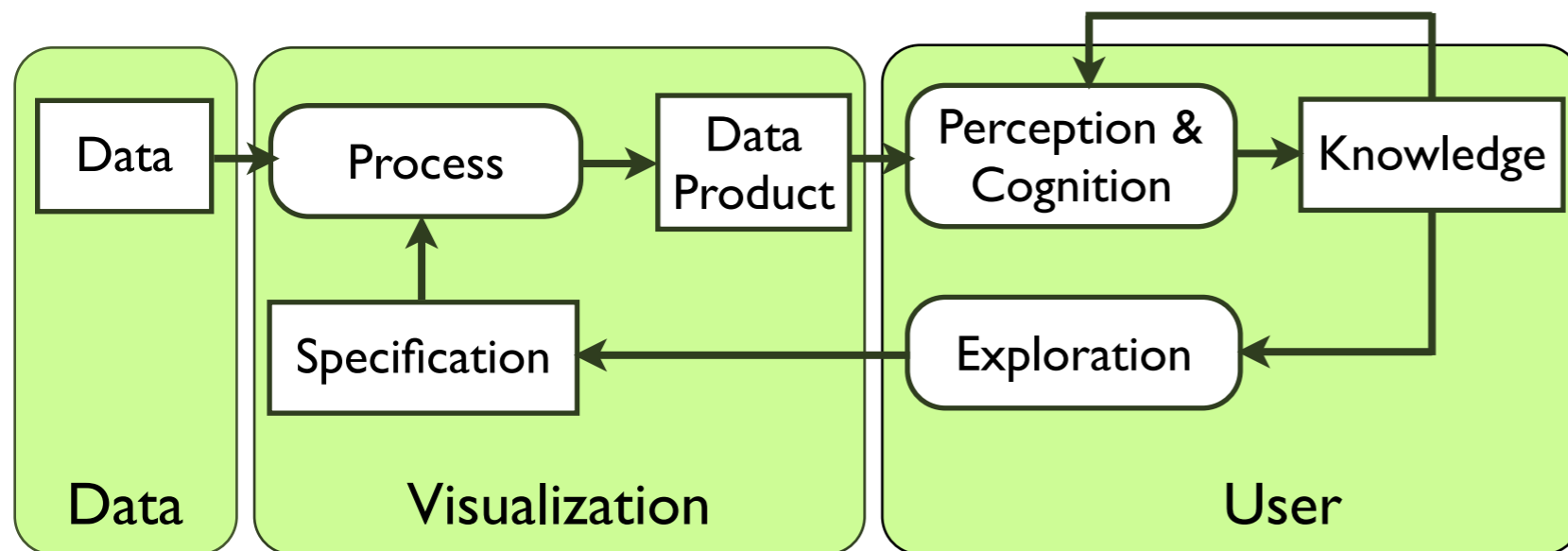


Figure modified from J. van Wijk, IEEE Vis2005

The VisTrails System

- Initial motivation: Visualization meets provenance
- More than visualization---supports computational tasks in general: Workflows meet provenance
- Focus on exploratory tasks such as simulations, visualization and data mining

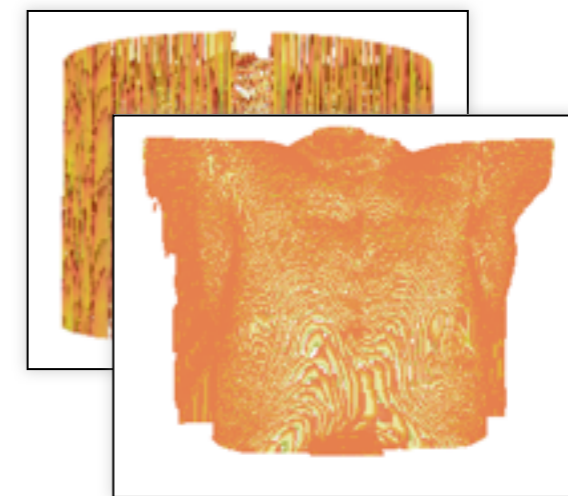
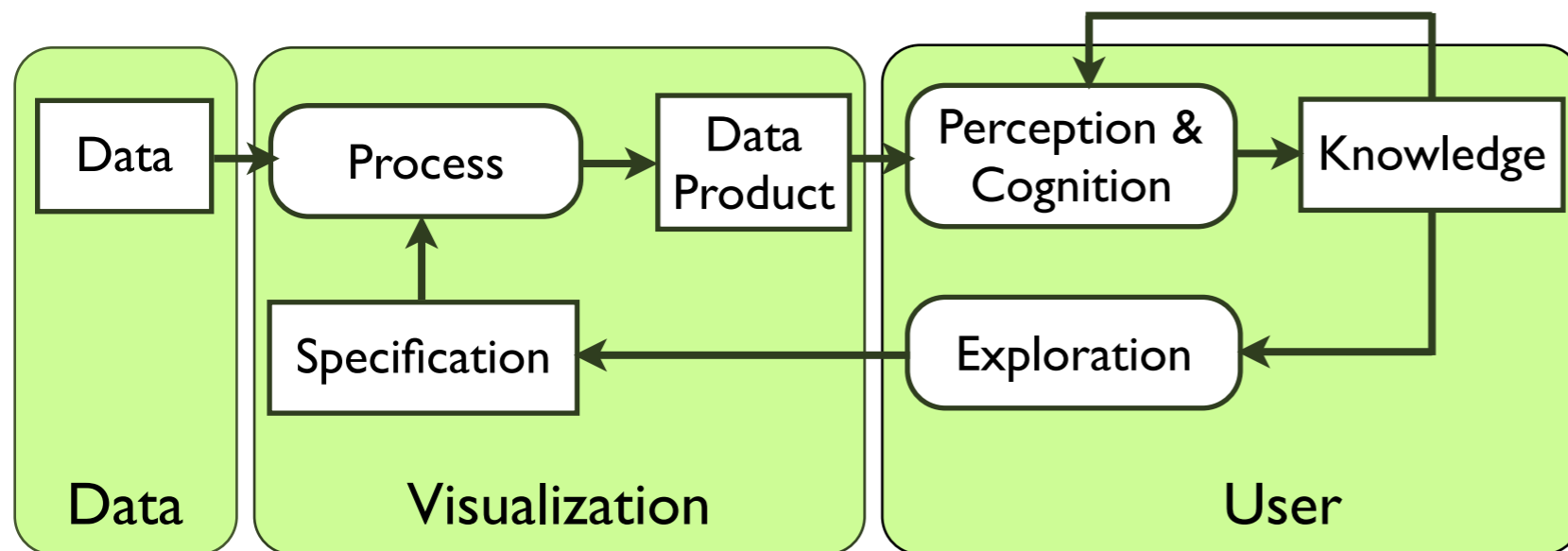


Figure modified from J. van Wijk, IEEE Vis2005

The VisTrails System

- Initial motivation: Visualization meets provenance
- More than visualization---supports computational tasks in general: Workflows meet provenance
- Focus on exploratory tasks such as simulations, visualization and data mining

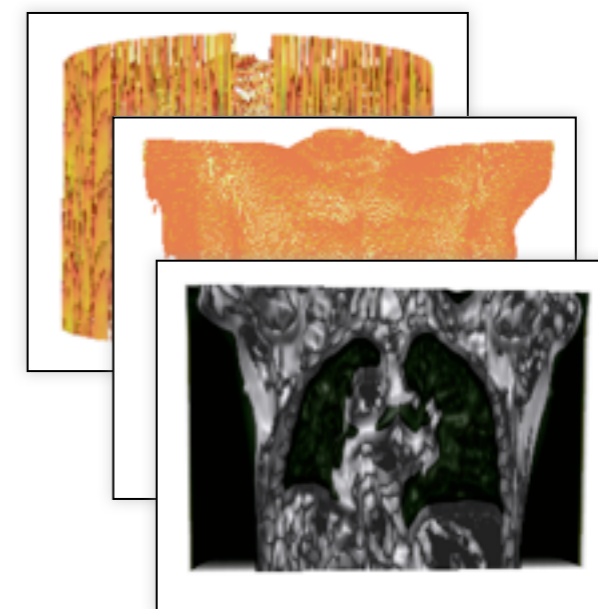
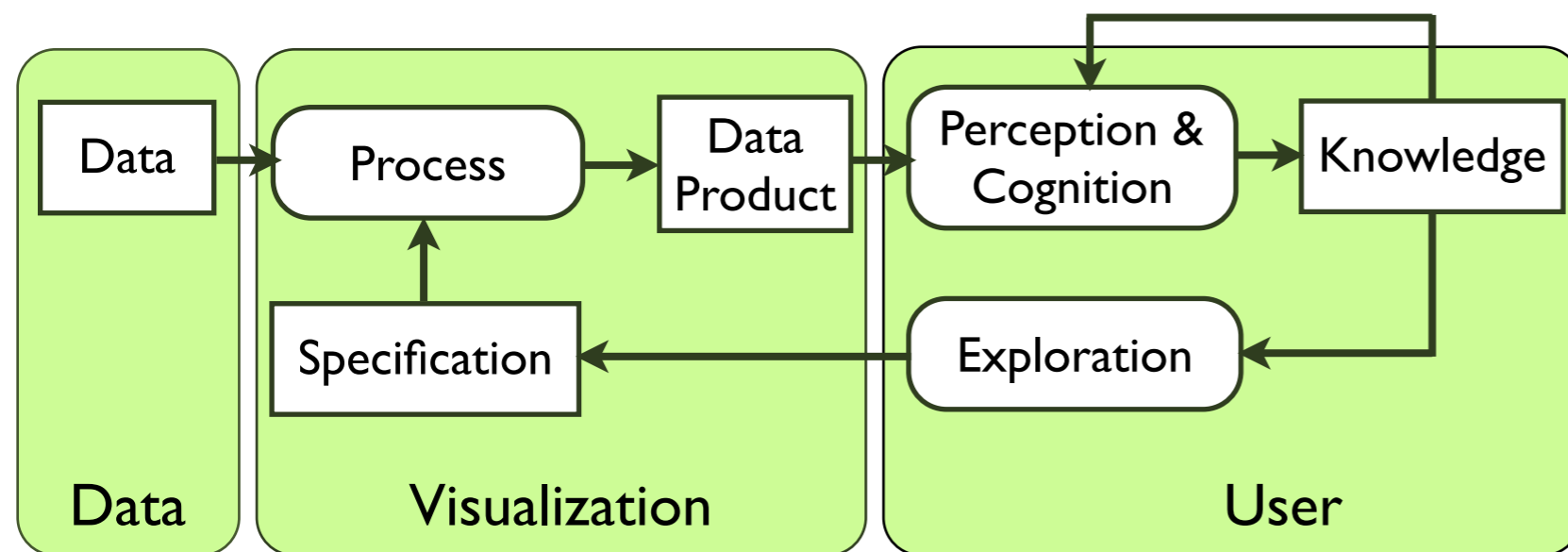


Figure modified from J. van Wijk, IEEE Vis2005

The VisTrails System

- Initial motivation: Visualization meets provenance
- More than visualization---supports computational tasks in general: Workflows meet provenance
- Focus on exploratory tasks such as simulations, visualization and data mining

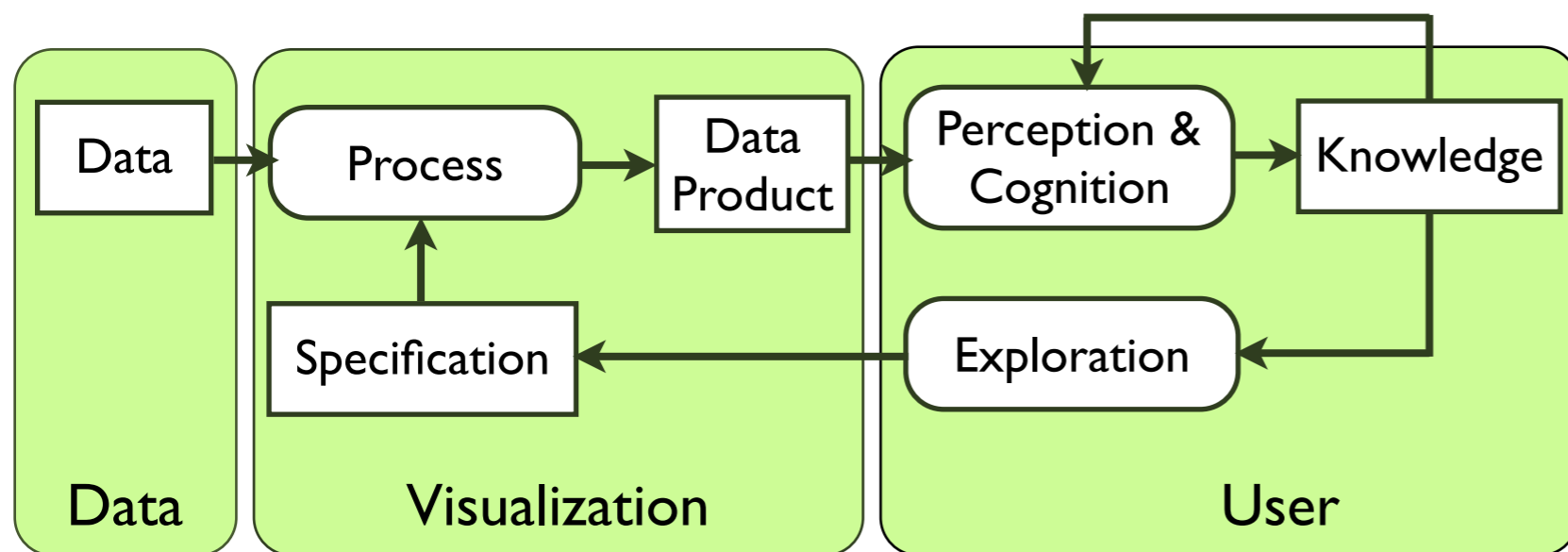
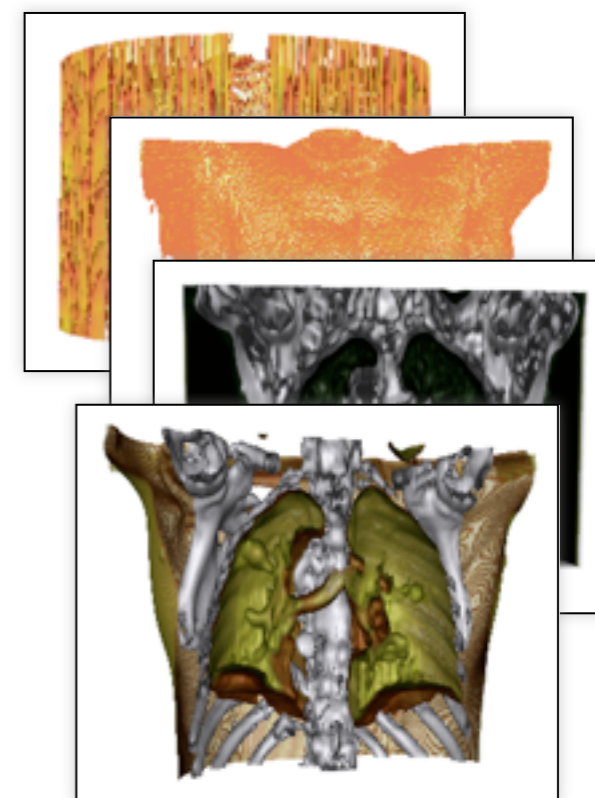


Figure modified from J. van Wijk, IEEE Vis2005



The VisTrails System

- Initial motivation: Visualization meets provenance
- More than visualization---supports computational tasks in general: Workflows meet provenance
- Focus on exploratory tasks such as simulations, visualization and data mining

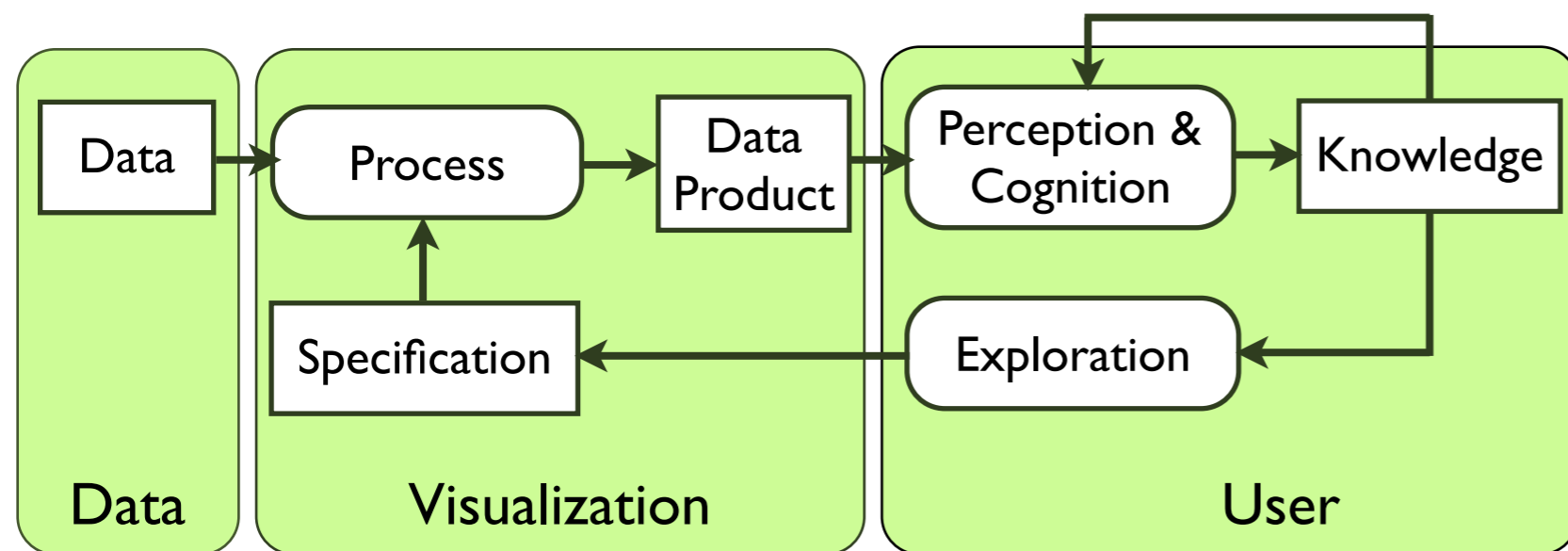
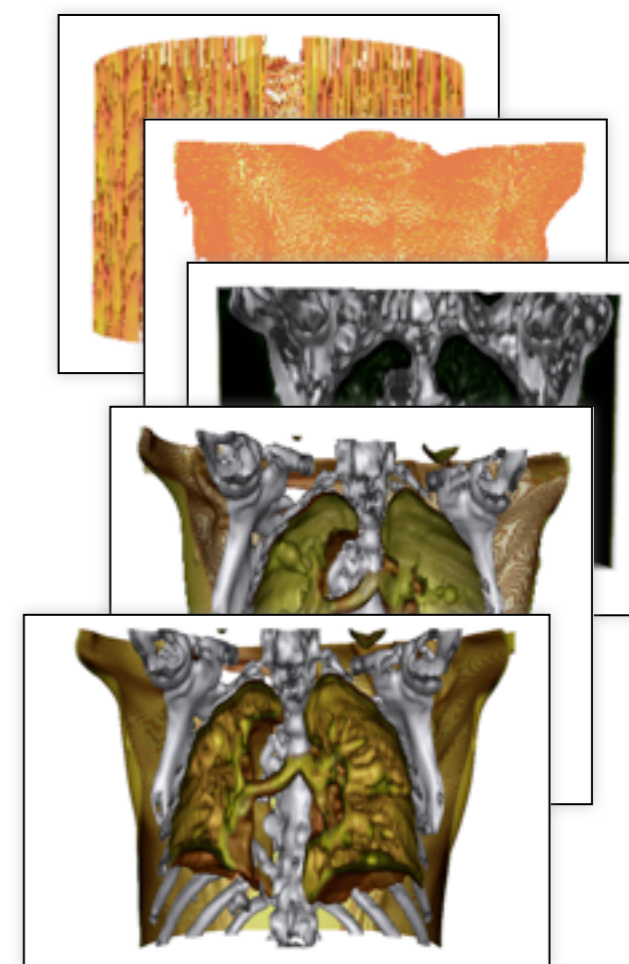
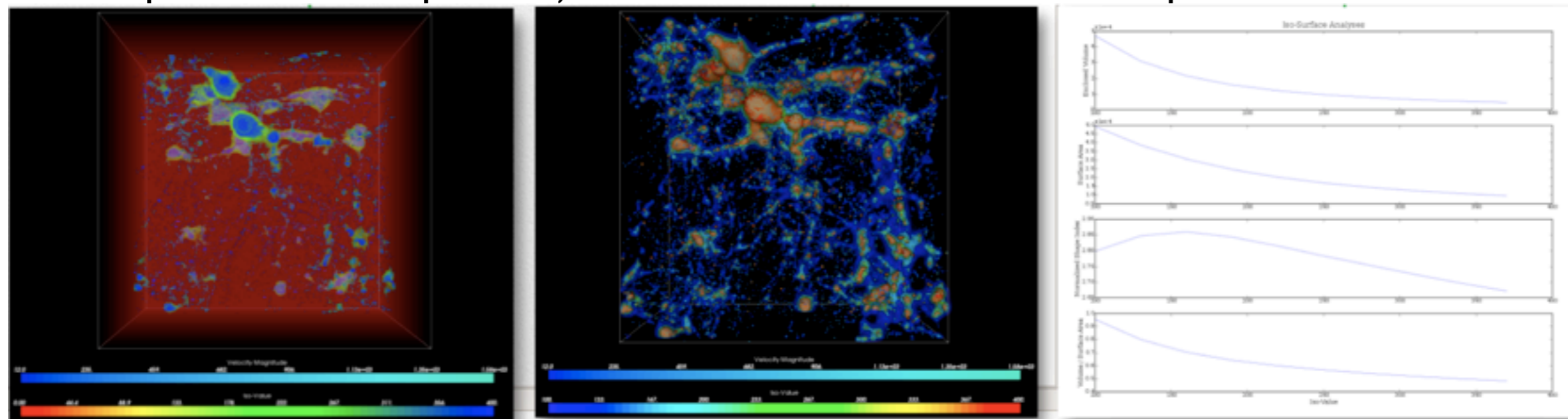


Figure modified from J. van Wijk, IEEE Vis2005



The VisTrails System

- Visualization meets provenance
- More than visualization---supports computational tasks in general
 - Workflows meet provenance
- Focus on exploratory tasks such as simulations, visualization and data mining
 - Help users explore, interact with and compare results



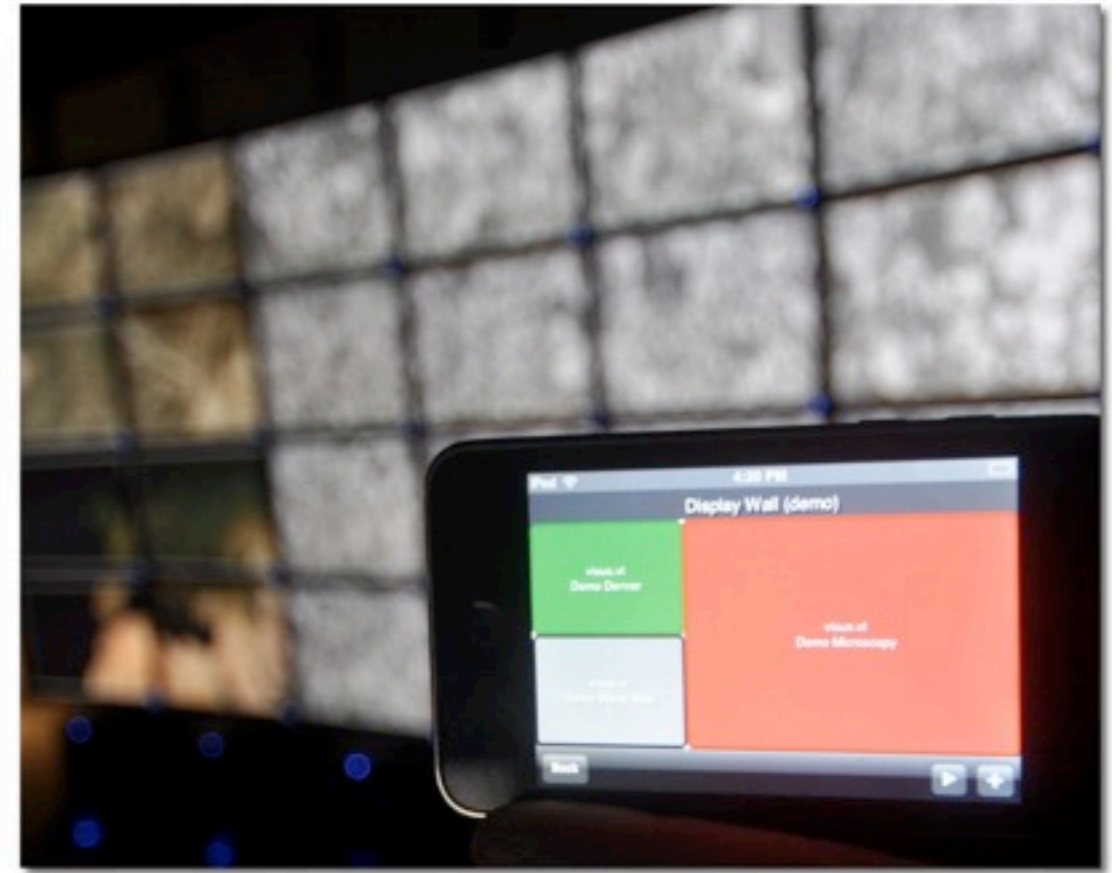
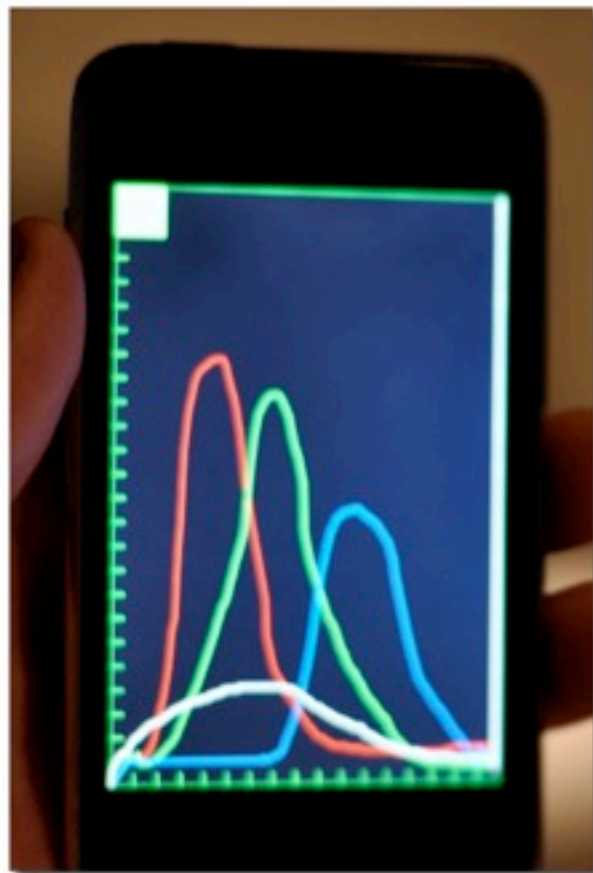
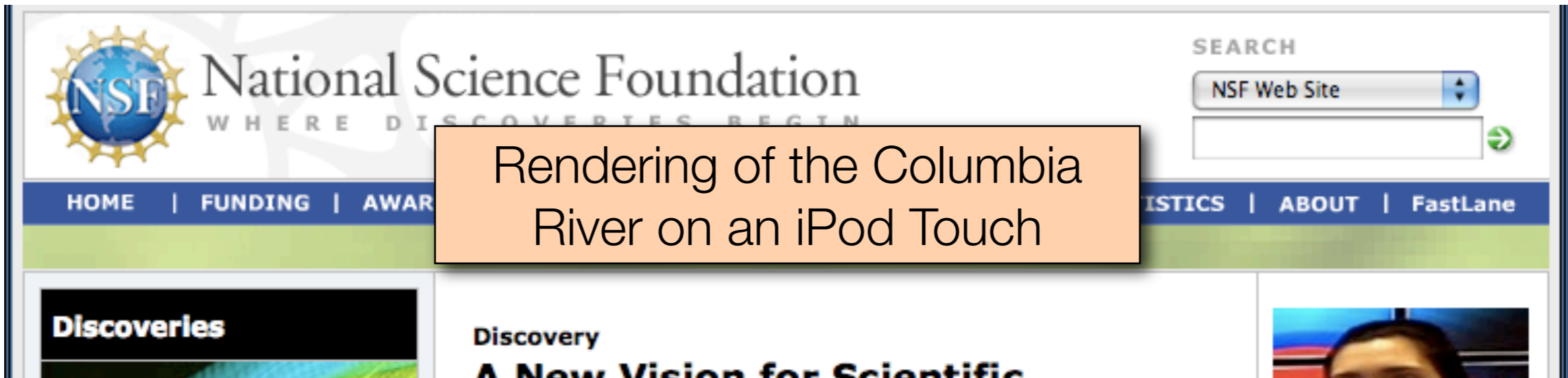
Cosmology Simulations - LANL

VisTrails: At Large and on the go

High-resolution rendering of the Columbia river estuary at a display wall

The screenshot shows the NSF Discoveries website. At the top left is the NSF logo with the tagline "WHERE DISCOVERIES". A navigation bar includes links for HOME, FUNDING, AWARDS, DISCOVERIES (highlighted), NEWS, PUBLICATIONS, STATISTICS, ABOUT, and FastLane. The main content area features a "Discoveries" sidebar on the left with a menu of research areas. The central article is titled "A New Vision for Scientific Visualizations" and describes how new technologies allow researchers to create high-quality visualizations from large data sets. A large image shows a person interacting with a curved display wall showing a complex 3D visualization of a river estuary. To the right, there is a video interview with Juliana Freire from the University of Utah, with a "View Video" link and a description of the interview. Below the video is a smaller image of a globe with data points.

VisTrails: At Large and on the go



The VisTrails System

- Visualization meets provenance
- More than visualization---supports computational tasks in general
 - Workflows meet provenance
- Focus on exploratory tasks such as simulations, visualization and data mining
- Explore and compare results
- Python + Qt = multi-platform
 - Linux, Mac, Windows
- VisTrails is open source: <http://www.vistrails.org>

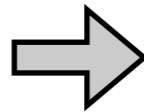
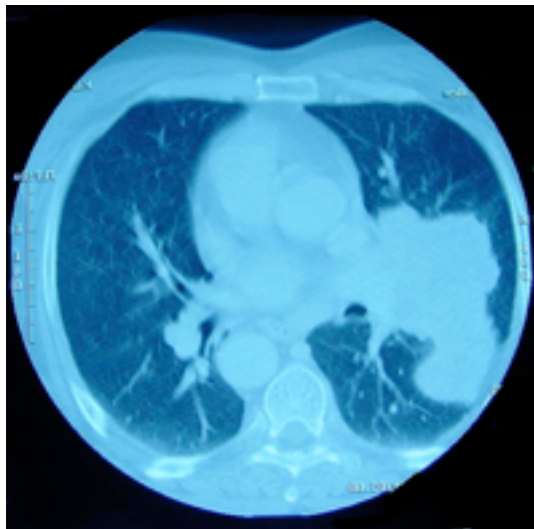
VisTrails: Managing Provenance

- Transparently tracks provenance of the discovery process
 - The trail followed as users generate and test hypotheses
- Use provenance to streamline exploration
- Focus on usability—build tools for scientists
 - VisTrails manages the data, metadata and the exploration process, scientists can focus on science!
- Infrastructure can be combined with and enhance scientific workflow and visualization systems

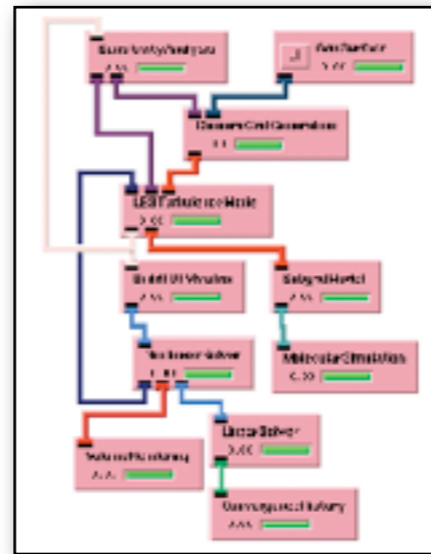
Data Exploration and Workflows

Data Exploration and Workflows

raw data:CT scan

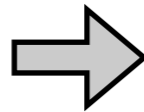
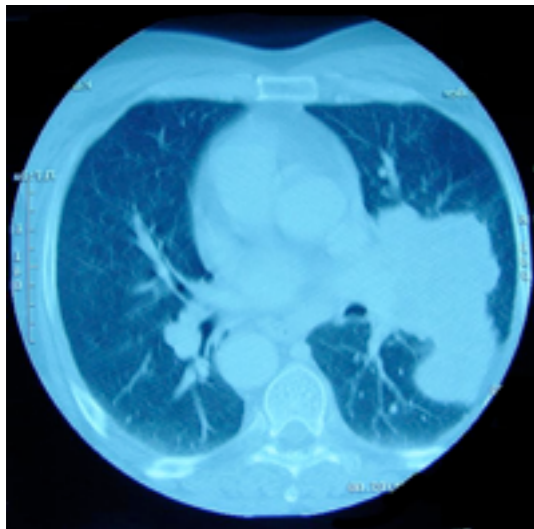


workflow

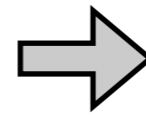
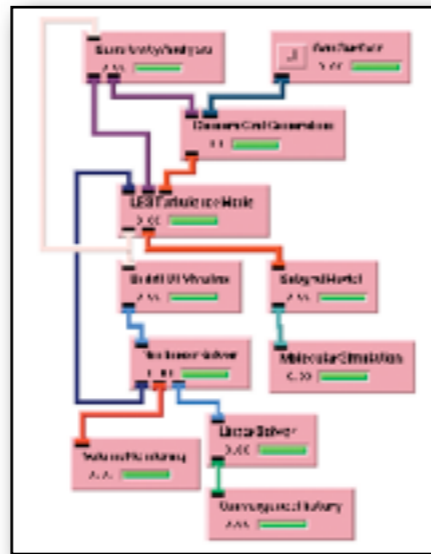


Data Exploration and Workflows

raw data:CT scan

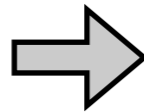
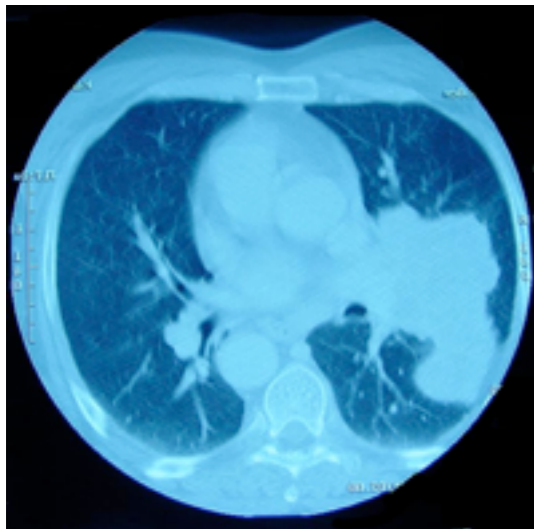


workflow

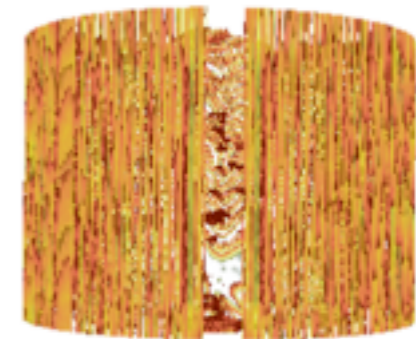
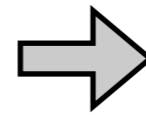
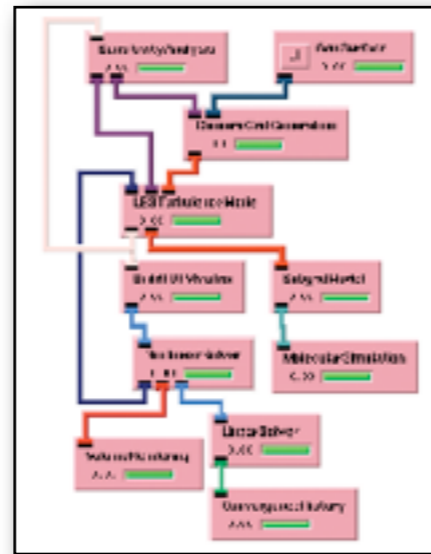


Data Exploration and Workflows

raw data:CT scan

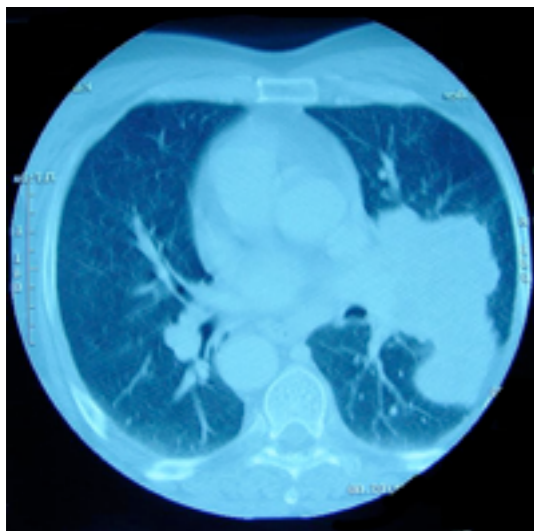


workflow

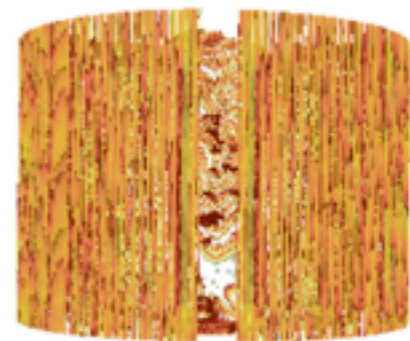
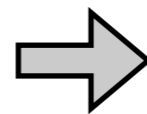
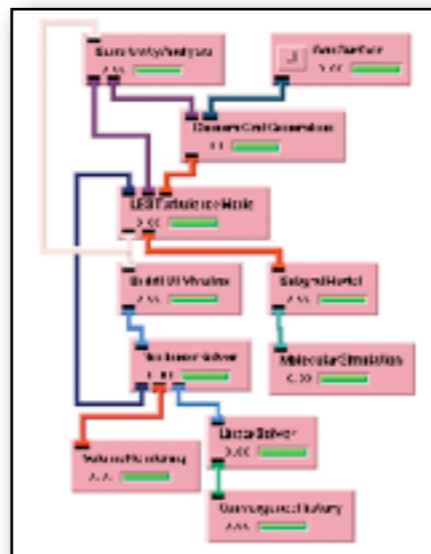


Data Exploration and Workflows

raw data:CT scan



workflow



Files (workflow specifications)

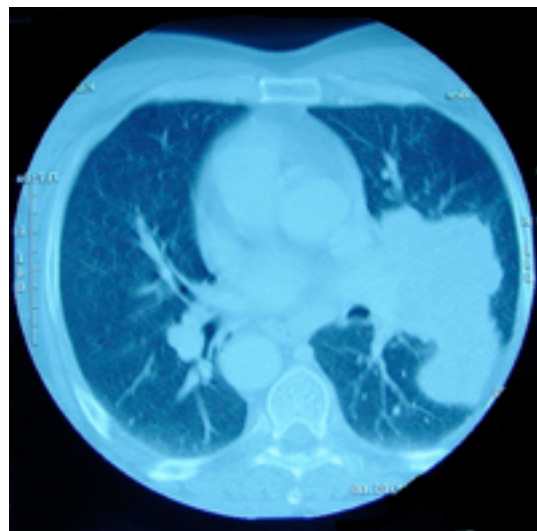
anon4877_voxel_scale_1_zspace_20060331.srn

Notes

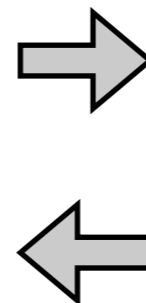
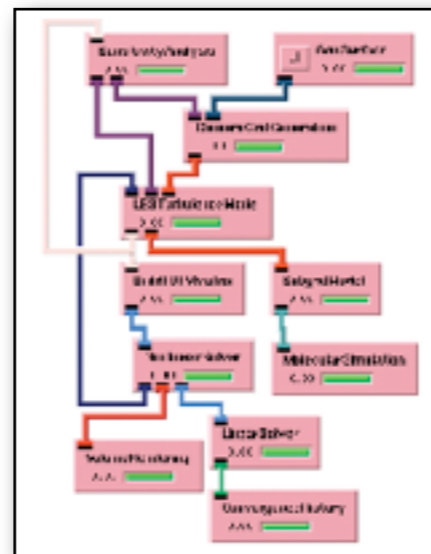
Initial
visualization
with z-scaling
corrected

Data Exploration and Workflows

raw data:CT scan



workflow



Files (workflow specifications)

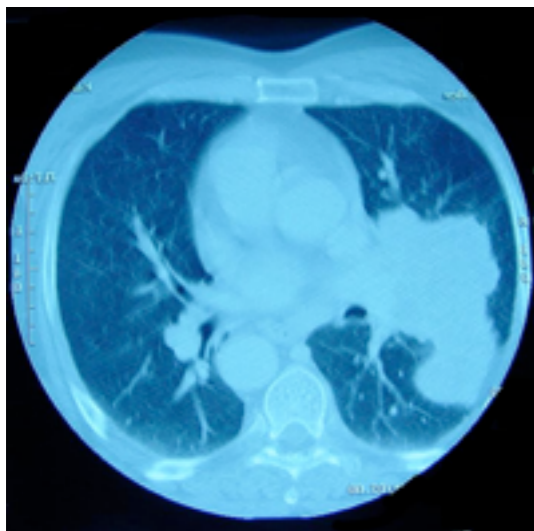
anon4877_voxel_scale_1_zspace_20060331.srn

Notes

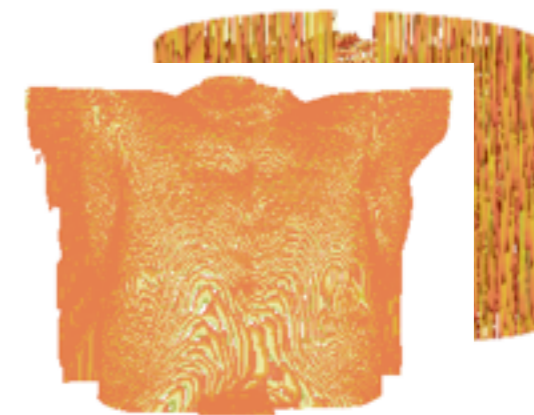
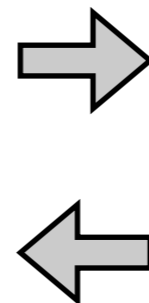
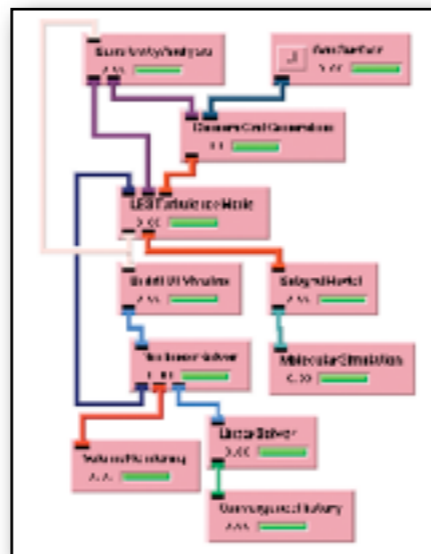
Initial
visualization
with z-scaling
corrected

Data Exploration and Workflows

raw data:CT scan



workflow



Files (workflow specifications)

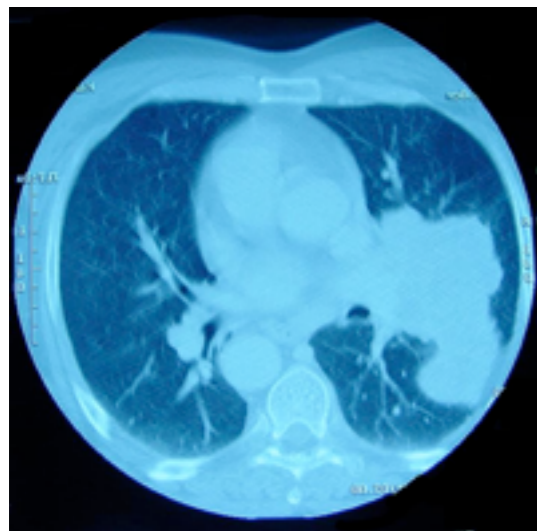
anon4877_voxel_scale_1_zspace_20060331.srn

Notes

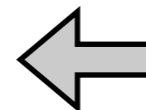
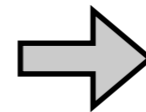
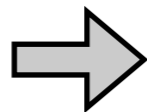
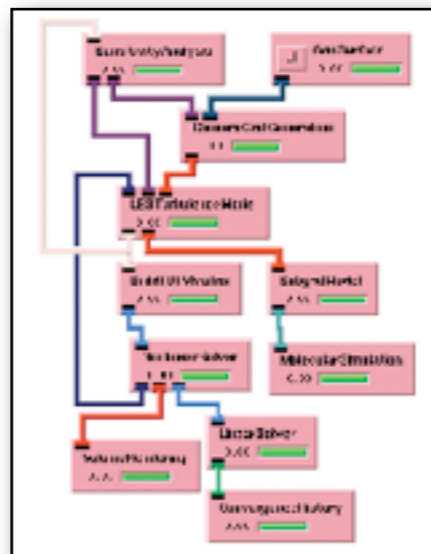
Initial
visualization
with z-scaling
corrected

Data Exploration and Workflows

raw data:CT scan



workflow



Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

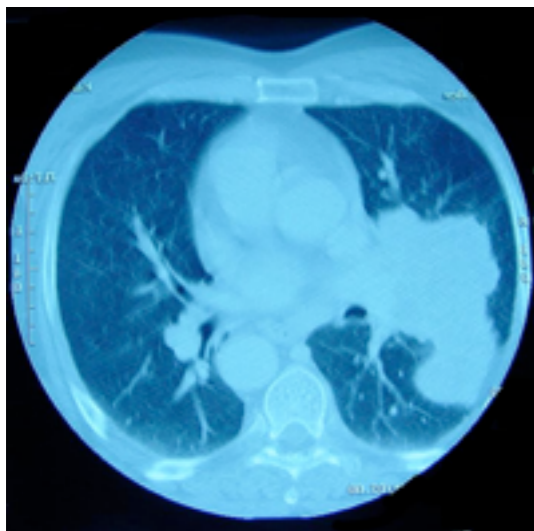
anon4877_textureshading_20060331.srn

Notes

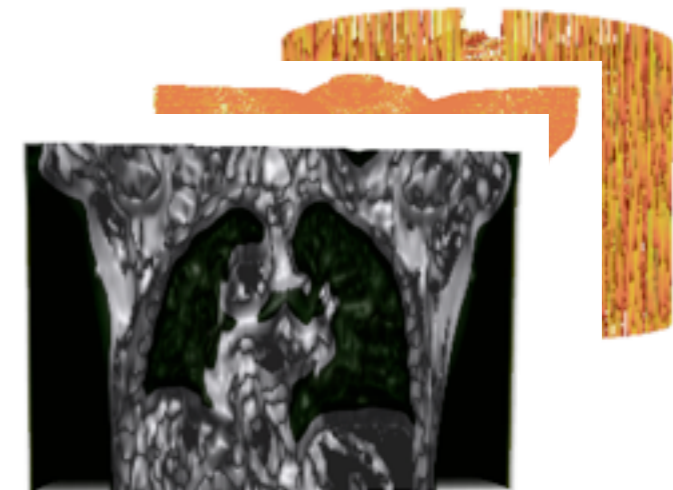
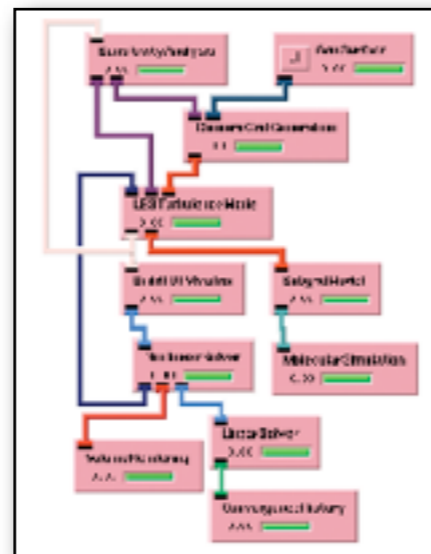
Initial
 ✓ Added texture
 w and shading
 corrected

Data Exploration and Workflows

raw data:CT scan



workflow



Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

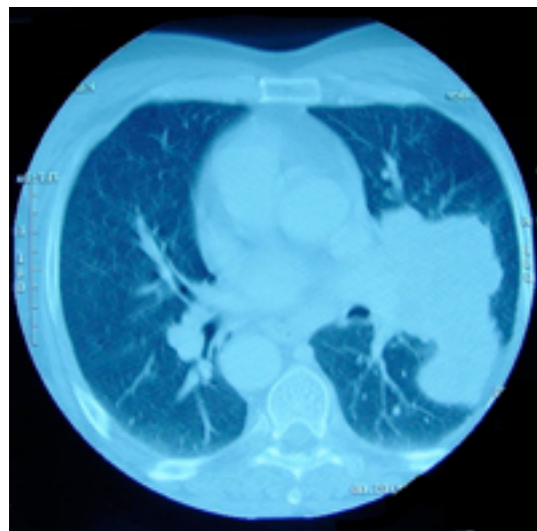
anon4877_textureshading_20060331.srn

Notes

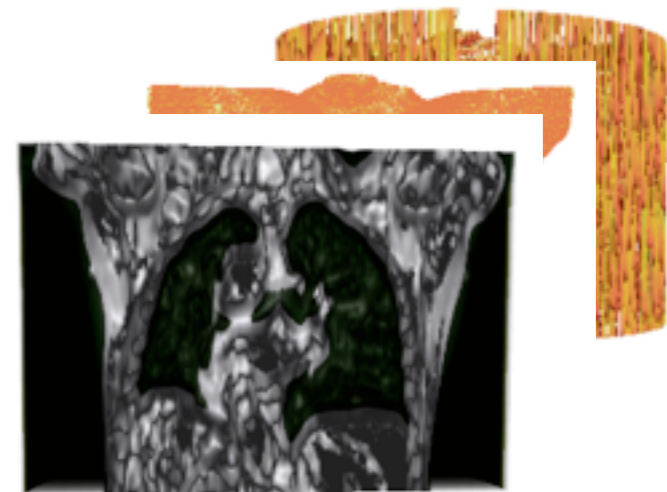
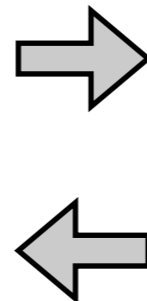
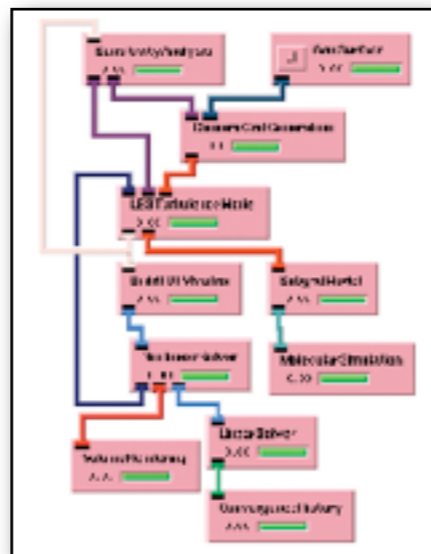
Initial
Added texture
and shading
corrected

Data Exploration and Workflows

raw data:CT scan



workflow



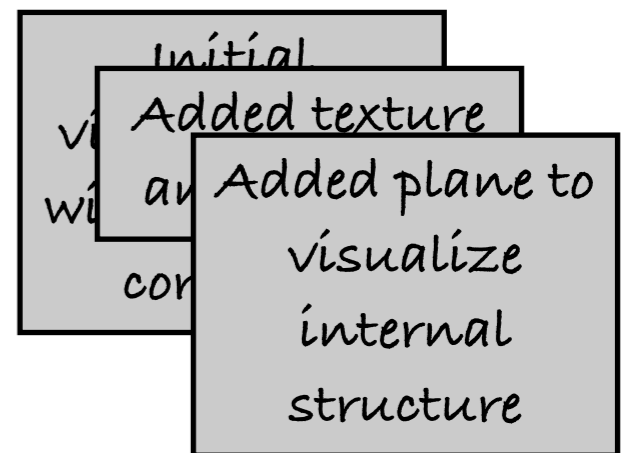
Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

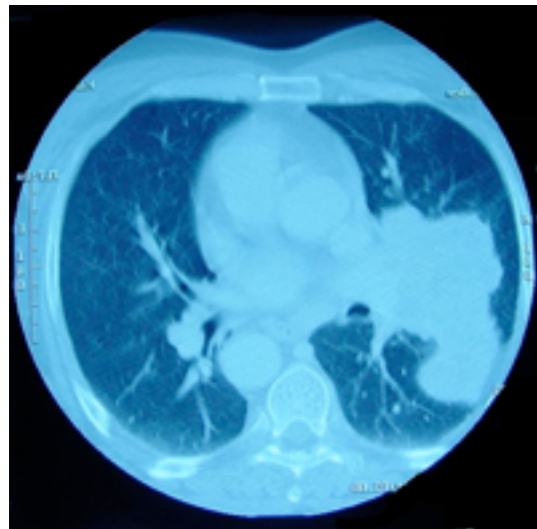
anon4877_textureshading_plane0_20060331.srn

Notes

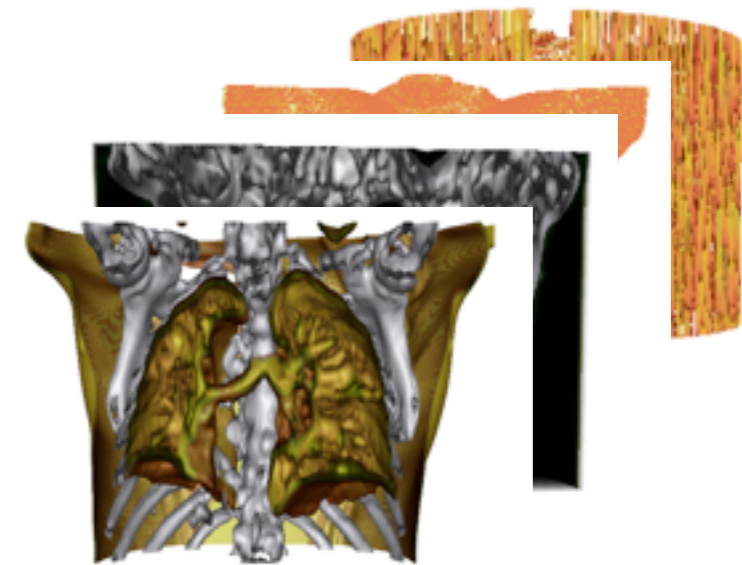
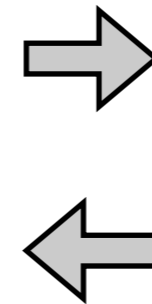
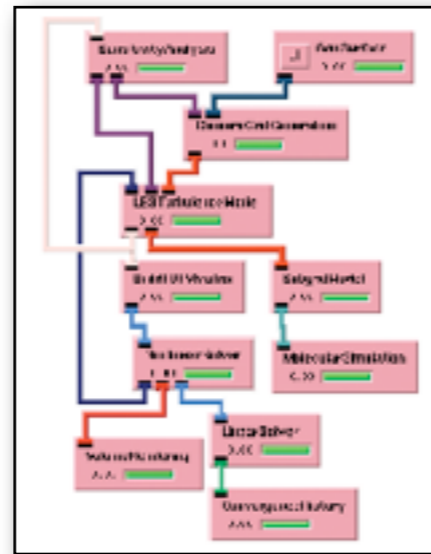


Data Exploration and Workflows

raw data:CT scan



workflow



Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

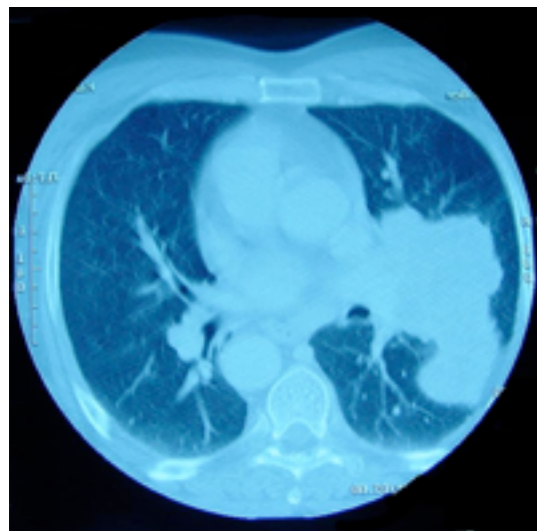
anon4877_textureshading_plane0_20060331.srn

Notes

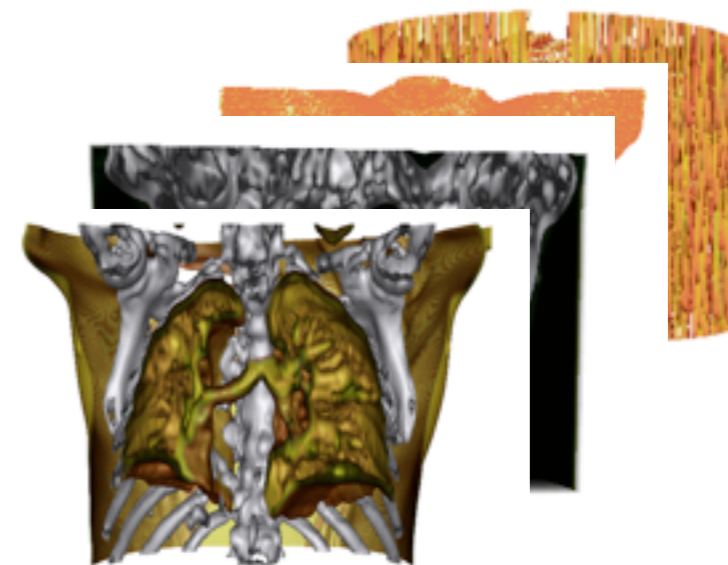
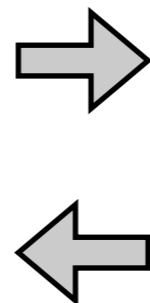
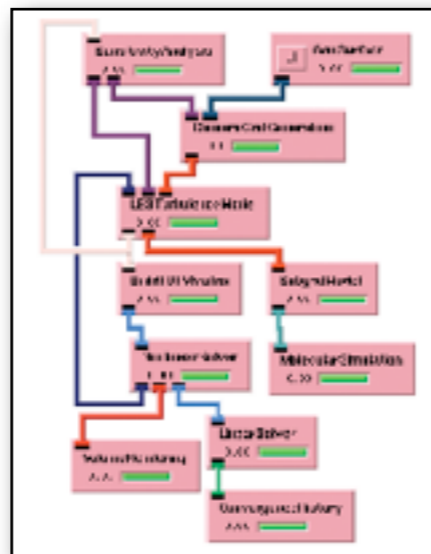
Initial
v Added texture
w av
con Added plane to
visualize
internal
structure

Data Exploration and Workflows

raw data:CT scan



workflow



Files (workflow specifications)

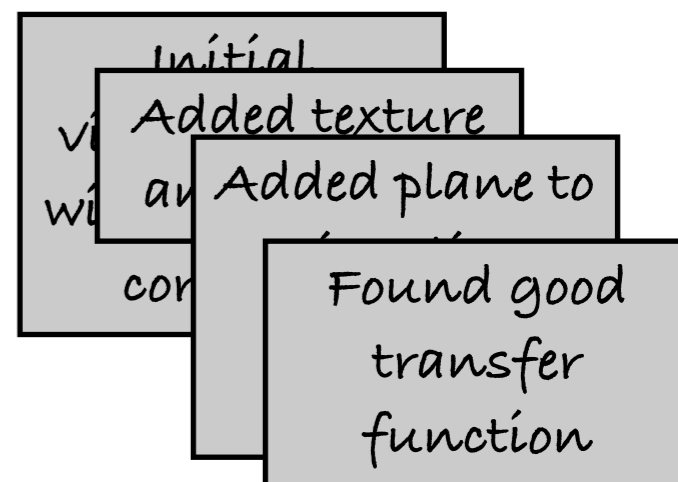
anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

anon4877_textureshading_plane0_20060331.srn

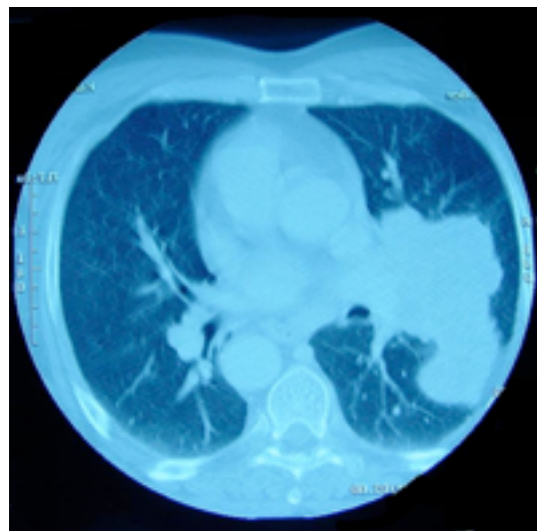
anon4877_goodxferfunction_20060331.srn

Notes

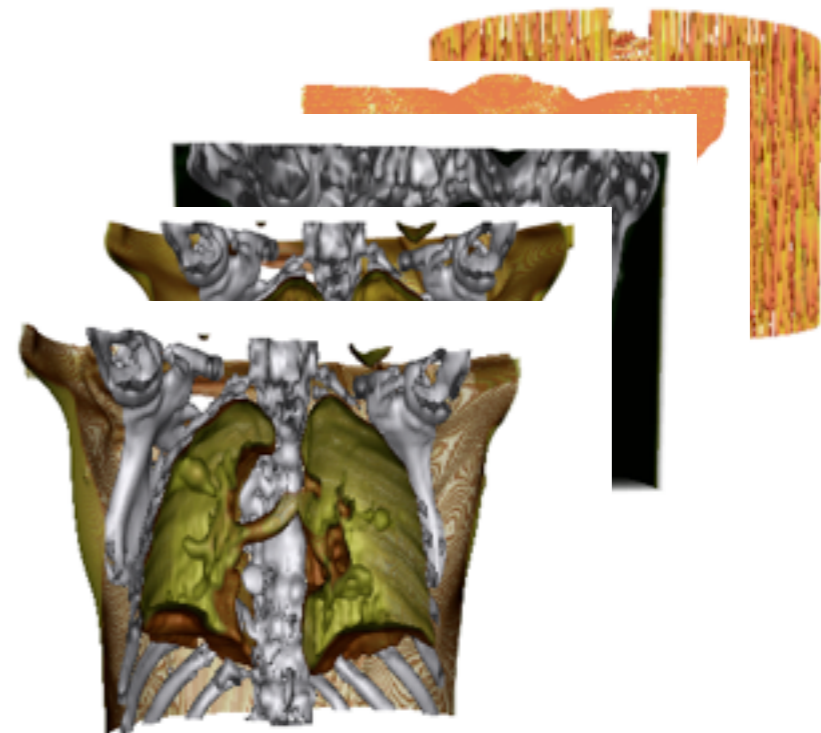
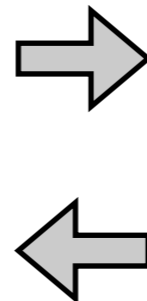
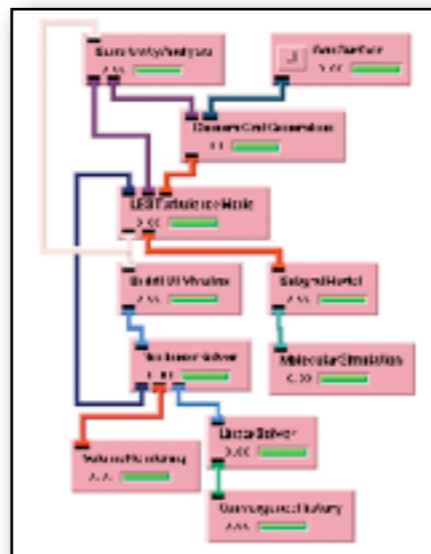


Data Exploration and Workflows

raw data:CT scan



workflow



Files (workflow specifications)

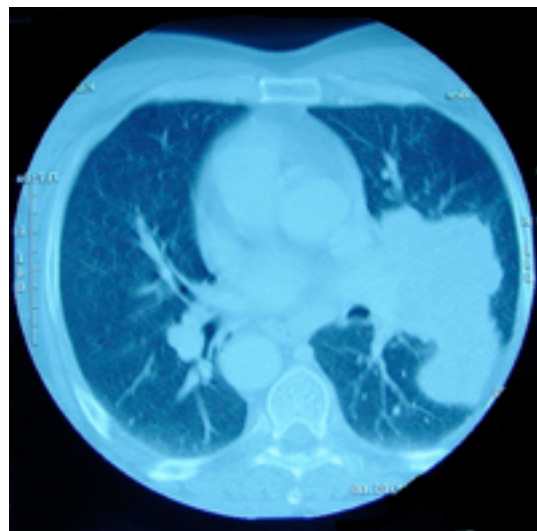
- anon4877_voxel_scale_1_zspace_20060331.srn
- anon4877_textureshading_20060331.srn
- anon4877_textureshading_plane0_20060331.srn
- anon4877_goodxferfunction_20060331.srn

Notes

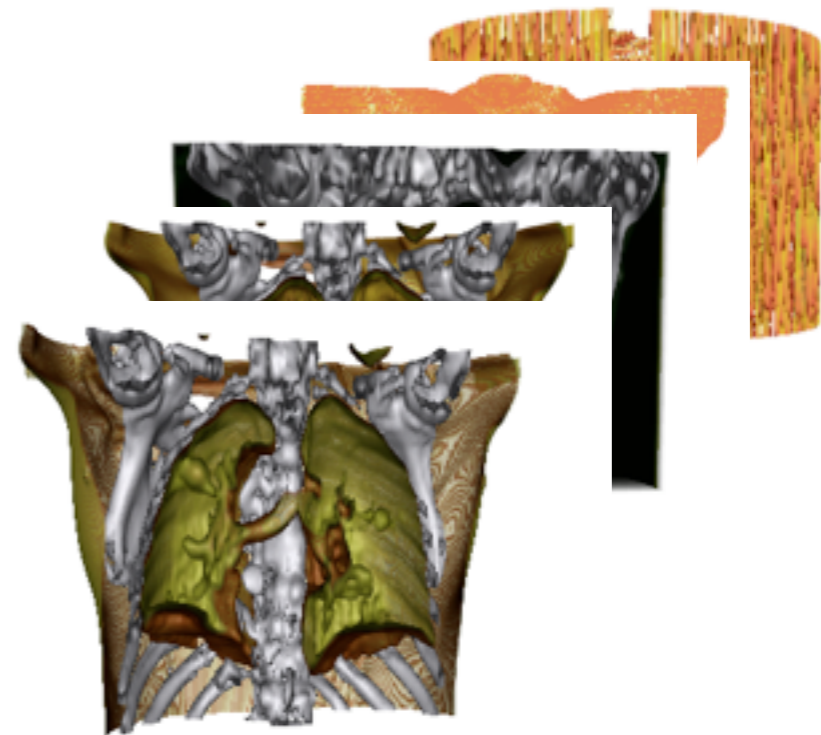
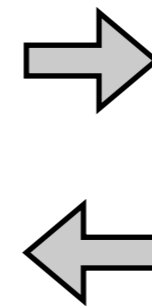
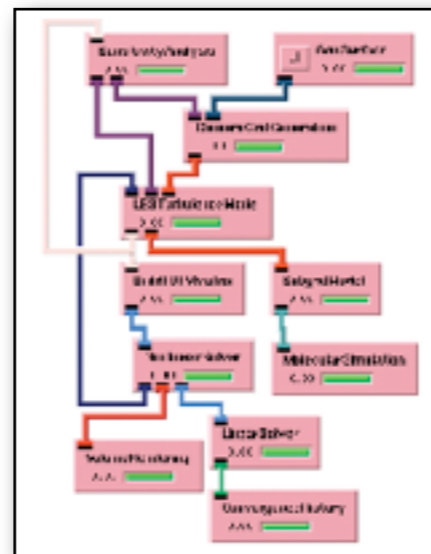
- Initial
- Added texture
- Added plane to
- Found good transfer function

Data Exploration and Workflows

raw data:CT scan



workflow



Files (workflow specifications)

- anon4877_voxel_scale_1_zspace_20060331.srn
- anon4877_textureshading_20060331.srn
- anon4877_textureshading_plane0_20060331.srn
- anon4877_goodxferfunction_20060331.srn
- anon4877_lesion_20060331.srn

Notes

- Initial
- Added texture
- Added plane to
- Found good transfer
- Identified lesion tissue

Data Exploration and Workflows: Issues

- Hard to assemble and refine workflows
 - Need in-depth knowledge to weave tools together
- Data provenance is maintained manually through file-naming conventions and detailed notes
 - A time-consuming process
- Hard to understand the exploratory process and relationships among workflows
 - Hard to further explore the data, e.g., locate relevant data products/workflows and modify them
- Hard to collaborate, and work is likely to be lost if creator leaves

The generation and maintenance of workflows is a major bottleneck in the scientific process

Data Exploration and Workflows: Issues

- Hard to assemble and refine workflows
 - Need in-depth knowledge to weave tools together
- Data provenance is maintained manually through file-naming conventions and detailed notes
 - A time-consuming process
- Hard to understand the exploratory process and relationships among workflows
 - Hard to further explore the data, e.g., locate relevant data products/workflows and modify them
- Hard to collaborate, and work is likely to be lost if creator leaves

Data Exploration and Workflows: Issues

- Hard to assemble and refine workflows
 - Need in-depth knowledge to weave tools together
- Data provenance is maintained manually through file-naming conventions and detailed notes
 - A time-consuming process
- Hard to understand the exploratory process and relationships among workflows
 - Hard to further explore the data, e.g., locate relevant data products/workflows and modify them
- Hard to collaborate, and work is likely to be lost if creator leaves

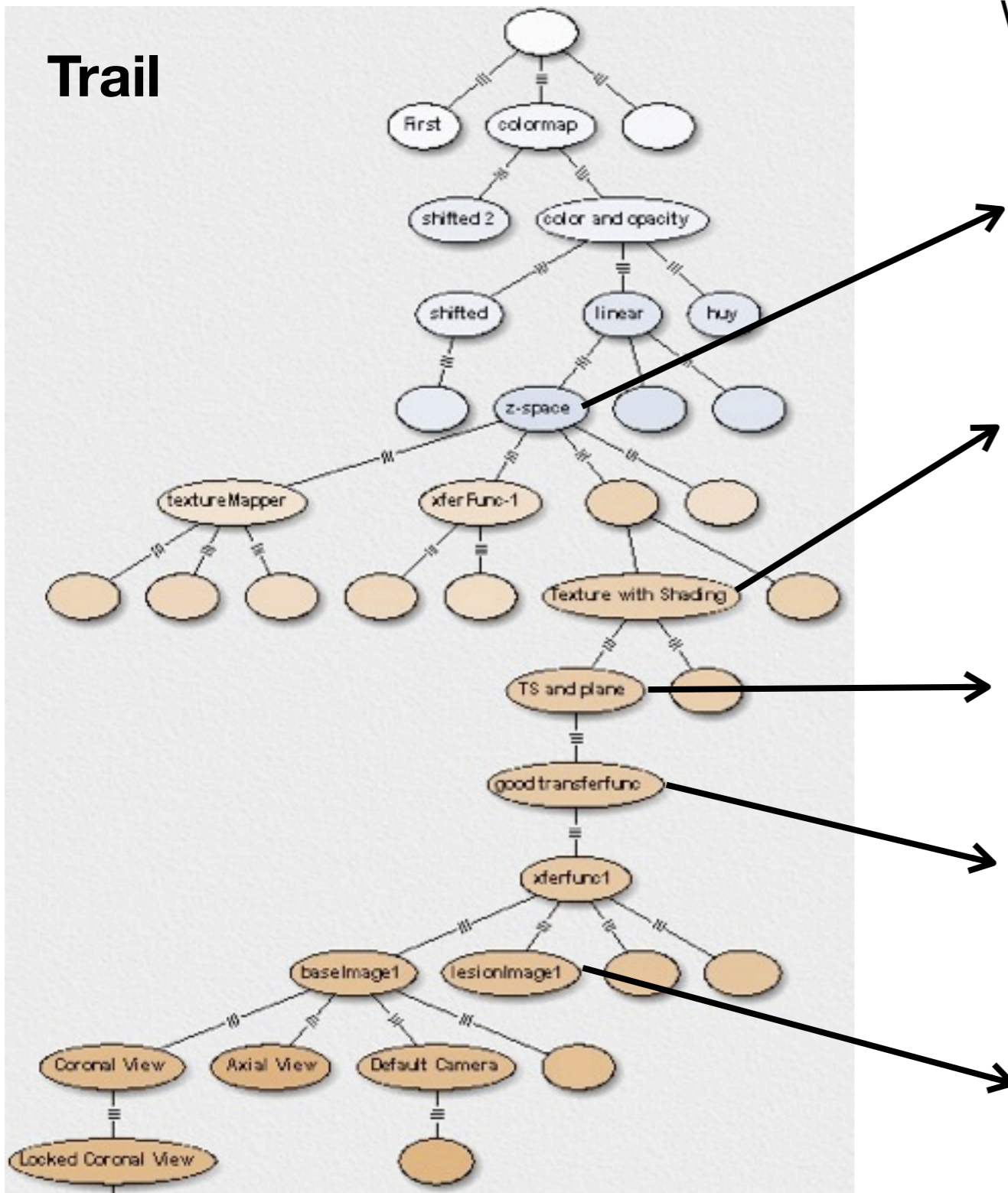
The generation and maintenance of workflows is a major bottleneck in the scientific process

Keeping Exploration Trails

Workflows

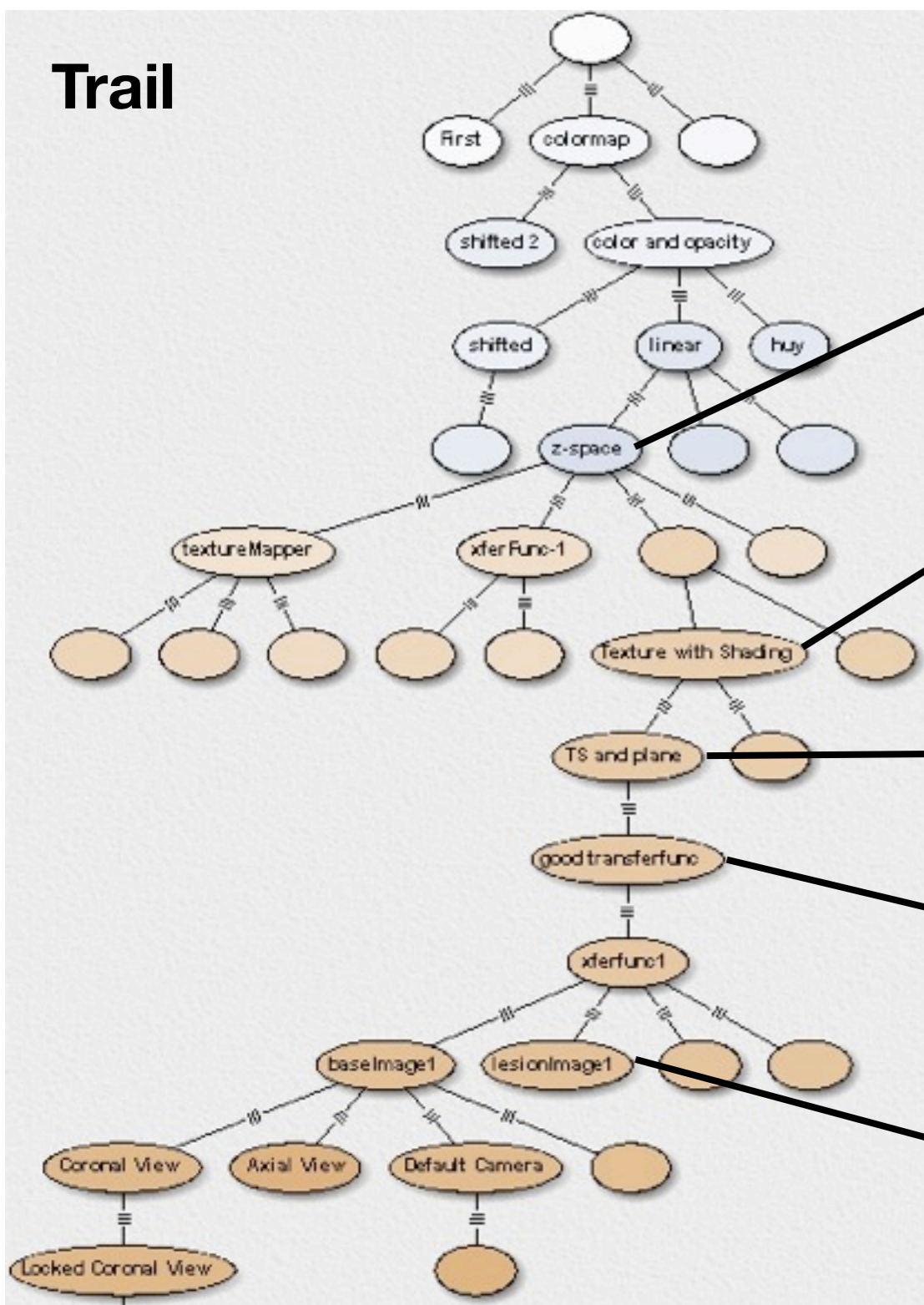
Data Products

Trail

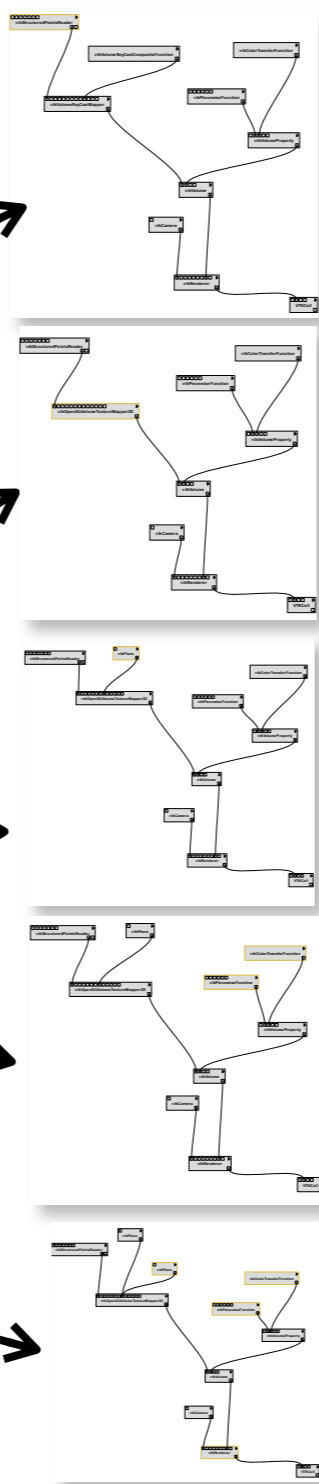


Keeping Exploration Trails

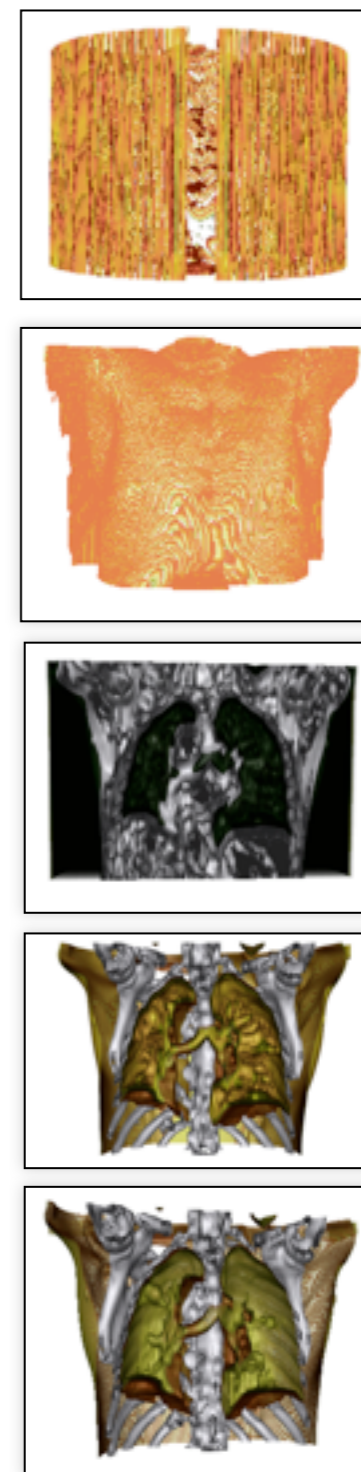
Trail



Workflows



Data Products

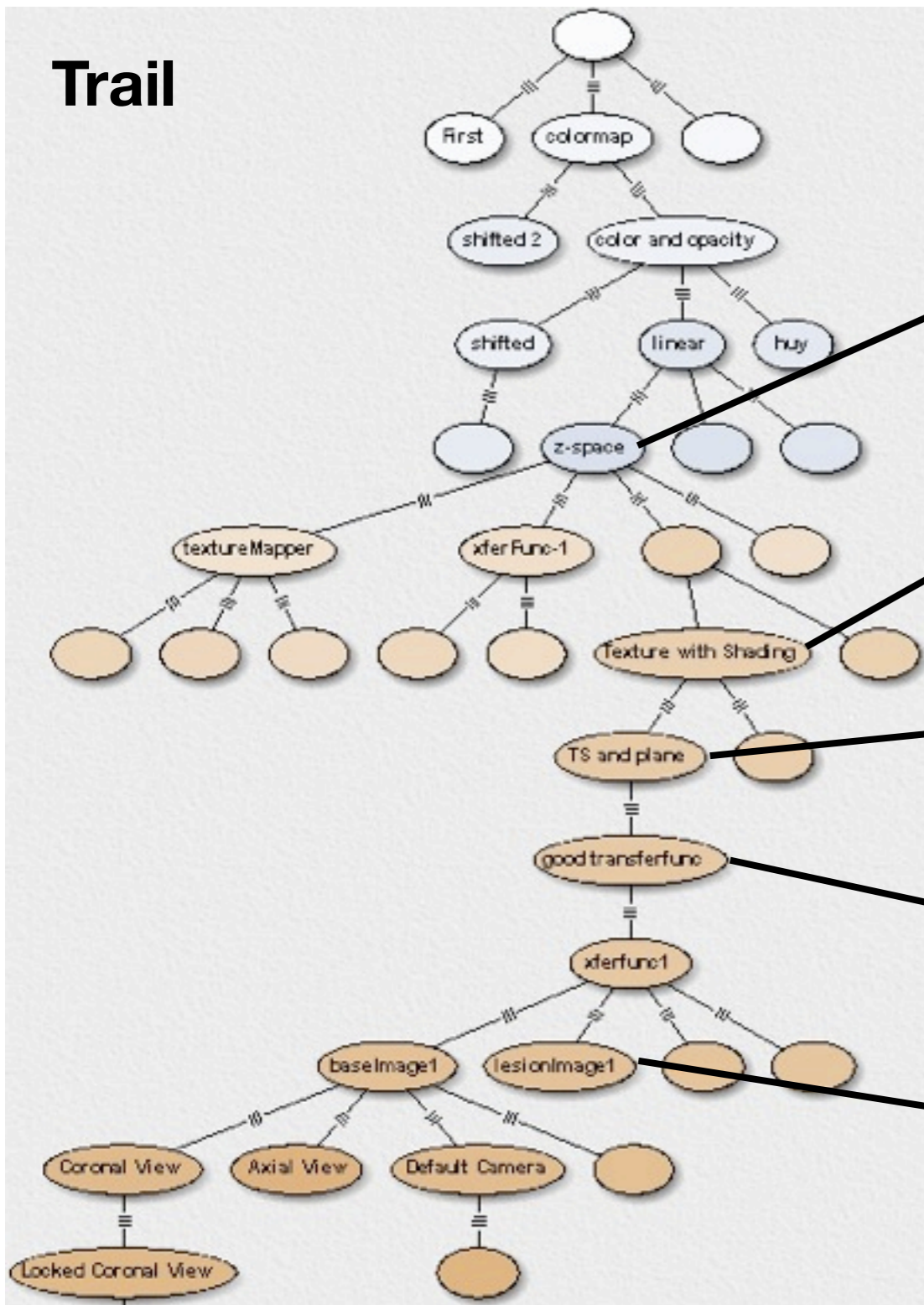


Keeping Exploration Trails

Notes

User

Trail



juliana

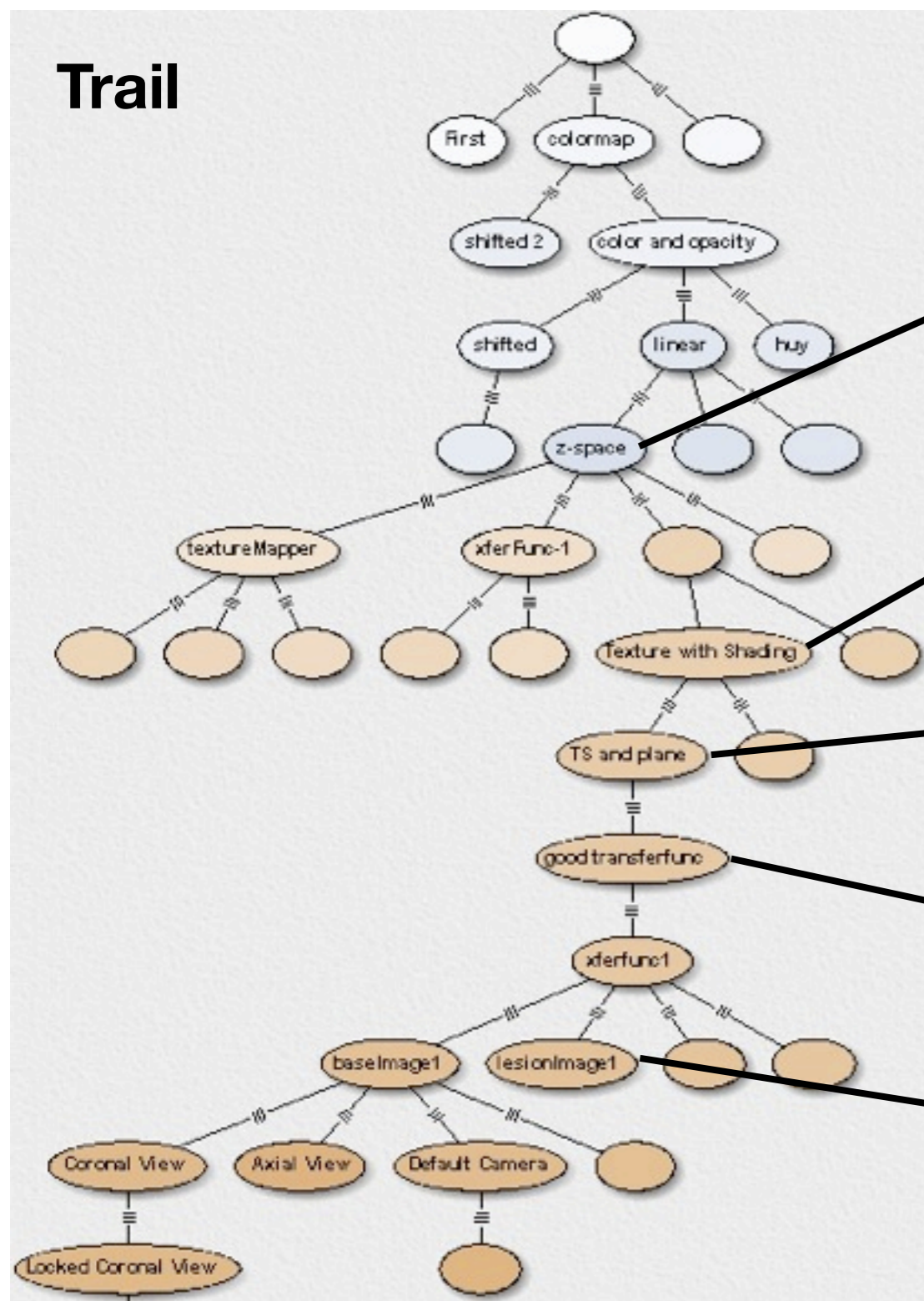
eranders

eranders

eranders

stevec

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

Demo 1

- Change-based provenance model
- Scalable generation of data products
- Understanding exploratory process: visual workflow difference

Change-Based Provenance

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

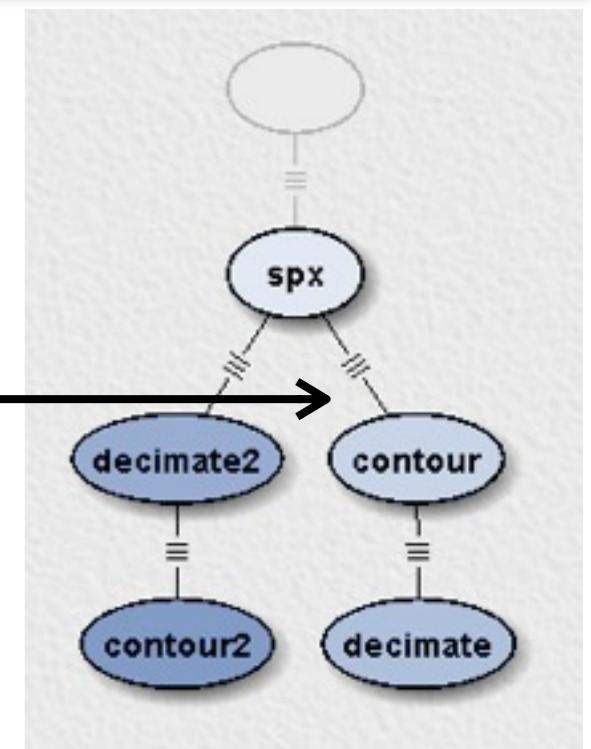
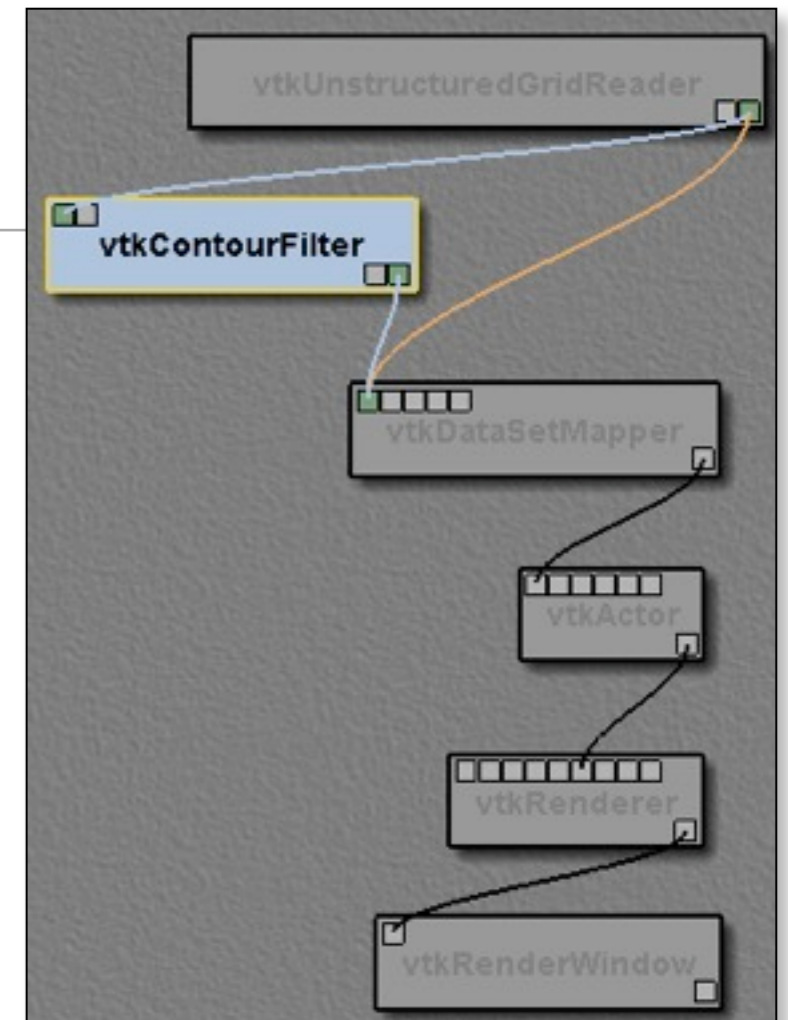
addModule

deleteConnection

addConnection

addConnection

setParameter



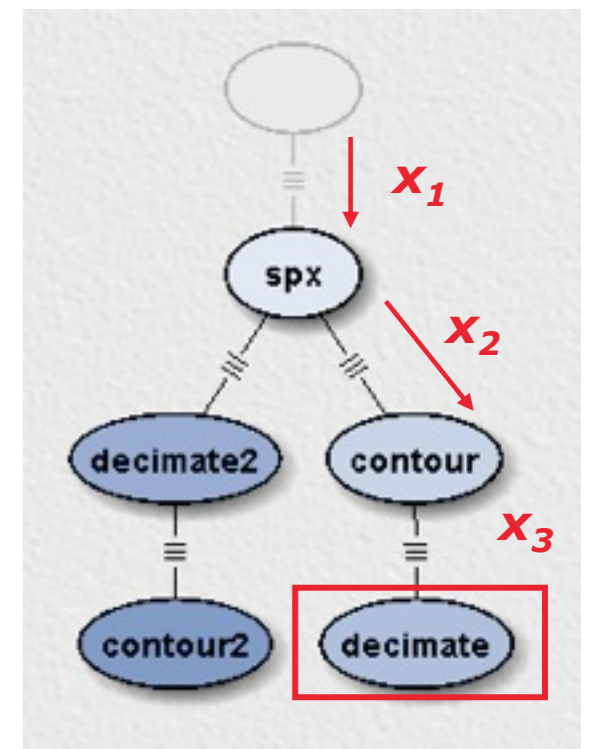
Change-Based Provenance

- Records user 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

$$V_t = X_n \circ X_{n-1} \circ \dots \circ X_1 \circ \emptyset$$

- Concise representation---no need to save all versions!

vistrail



VisTrails Provenance

Provenance Layers:

- Workflow evolution provenance
- Prospective provenance:
 - Recipes for how to produce data
- Retrospective provenance:
 - Invocation records of run time environments and resources used: site, host, executable, execution time, file stats ...
- Annotations: user-defined provenance
 - Metadata annotations about procedures and data
- Using this information, we can reconstruct the causality graph or derivation lineage, i.e., relationships among data, programs and computations

VisTrails Provenance

Provenance Layers:

- Workflow evolution provenance
- Prospective provenance: workflow definition
 - Recipes for how to produce data
- Retrospective provenance:
 - Invocation records of run time environments and resources used: site, host, executable, execution time, file stats ...
- Annotations: user-defined provenance
 - Metadata annotations about procedures and data
- Using this information, we can reconstruct the causality graph or derivation lineage, i.e., relationships among data, programs and computations

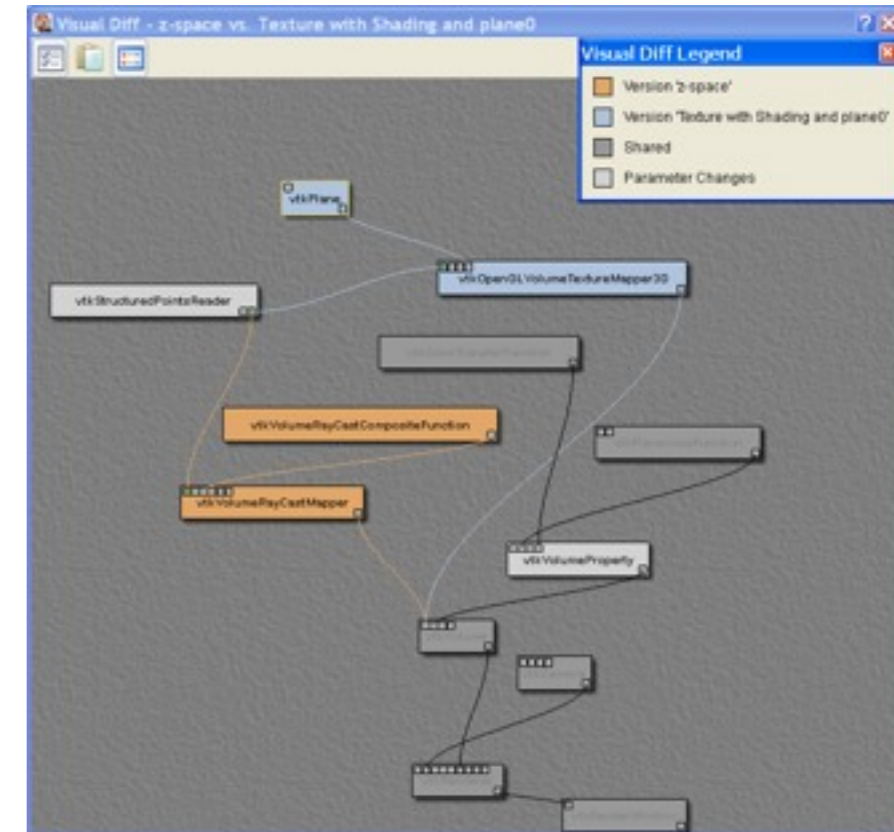
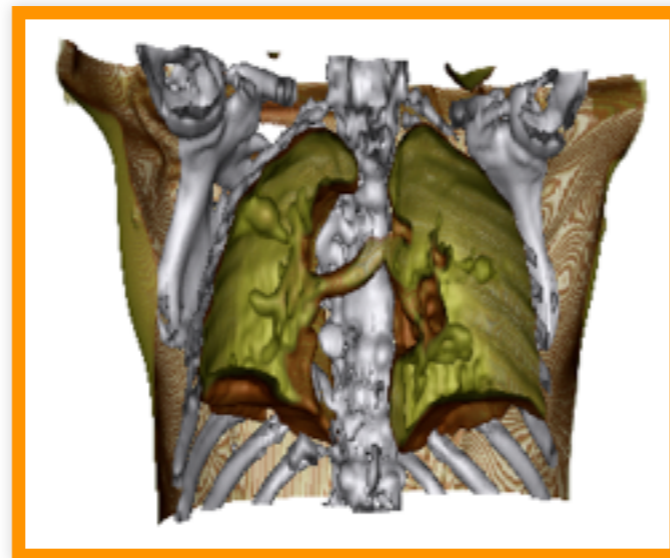
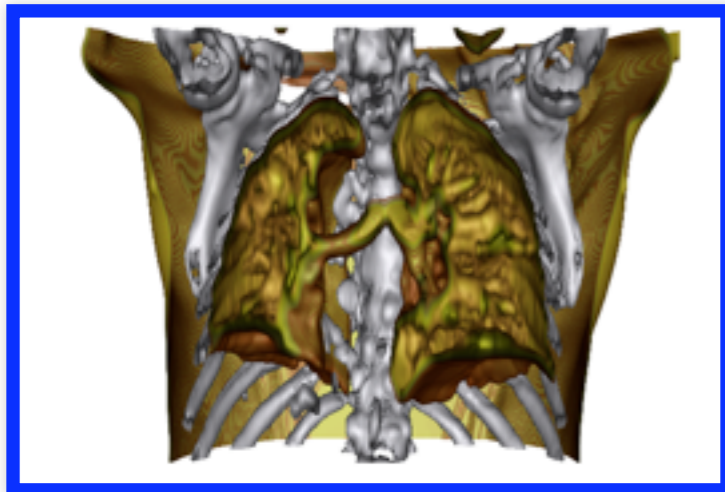
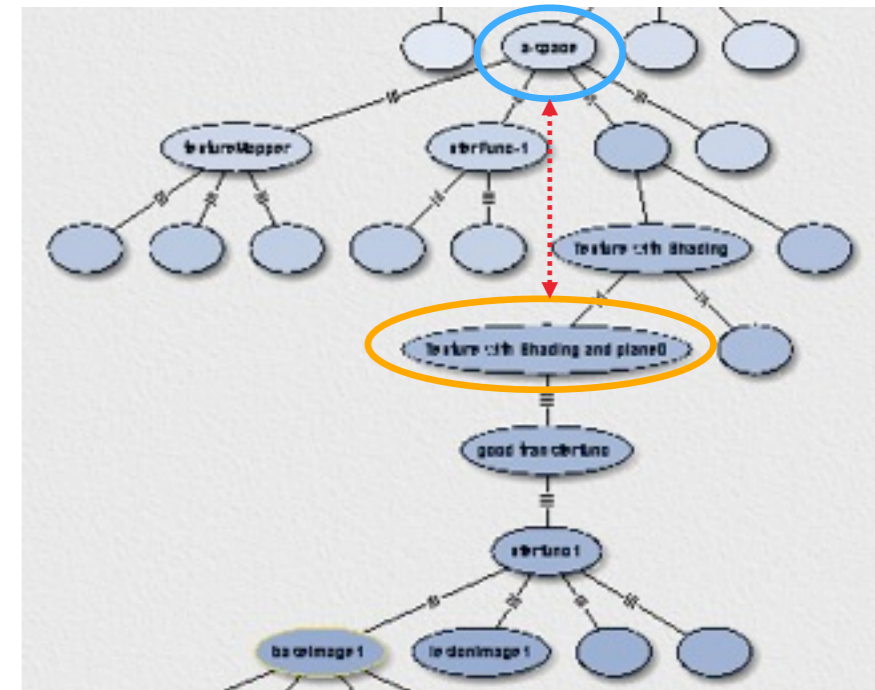
VisTrails Provenance

Provenance Layers:

- Workflow evolution provenance
- Prospective provenance: workflow definition
 - Recipes for how to produce data
- Retrospective provenance: execution log
 - Invocation records of run time environments and resources used: site, host, executable, execution time, file stats ...
- Annotations: user-defined provenance
 - Metadata annotations about procedures and data
- Using this information, we can reconstruct the causality graph or derivation lineage, i.e., relationships among data, programs and computations

Provenance Beyond Reproducibility

- Support for reflective reasoning
- Ability to compare data products

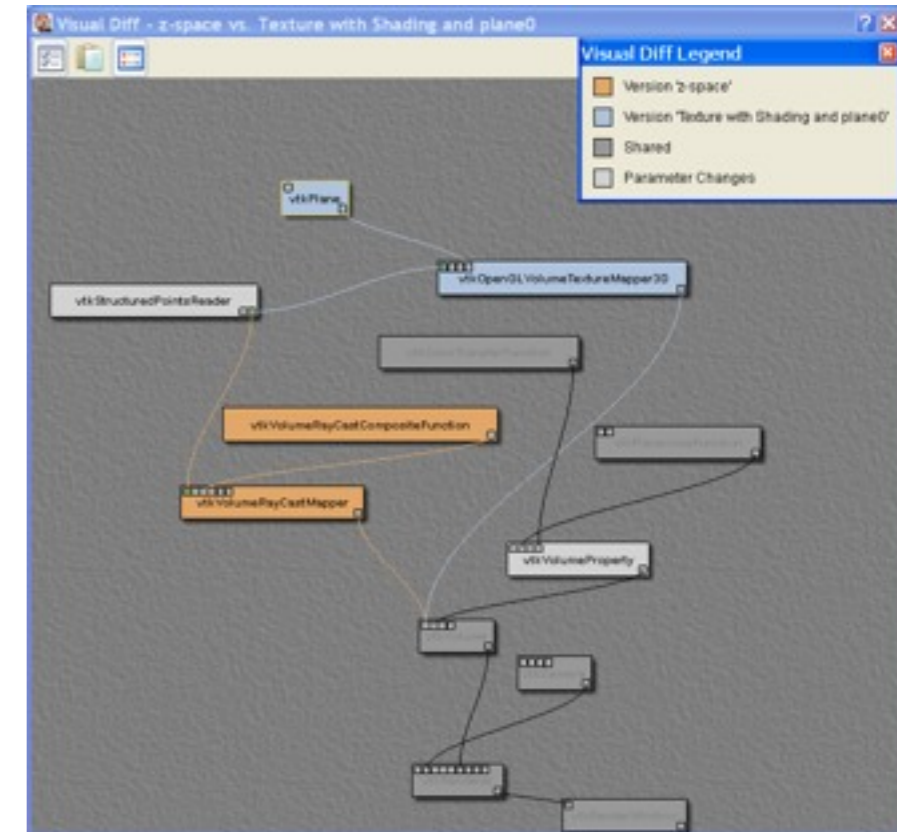
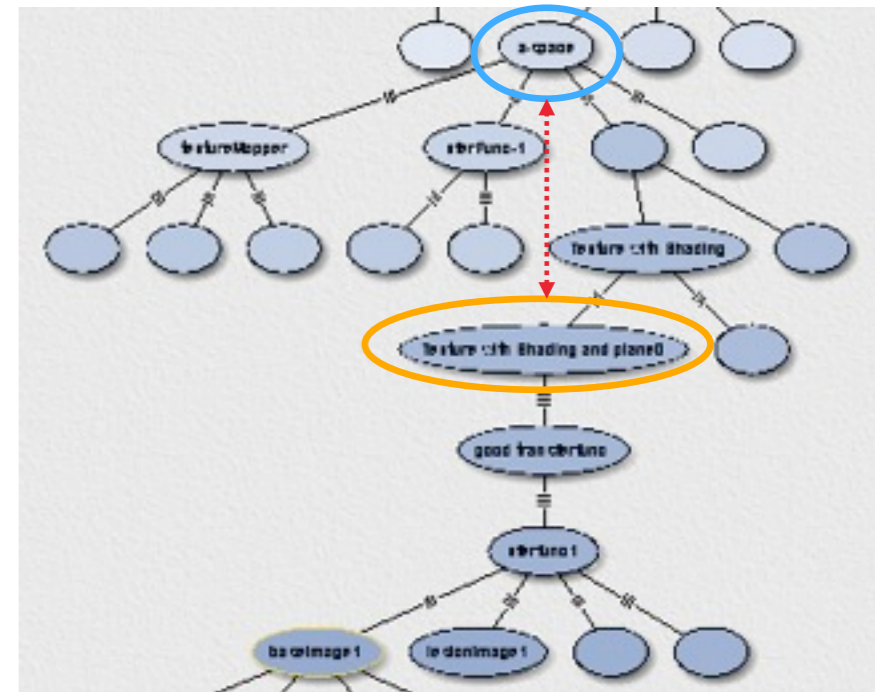


Provenance Beyond Reproducibility

- Support for reflective reasoning
- Ability to compare data products

“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



Computing workflow Differences

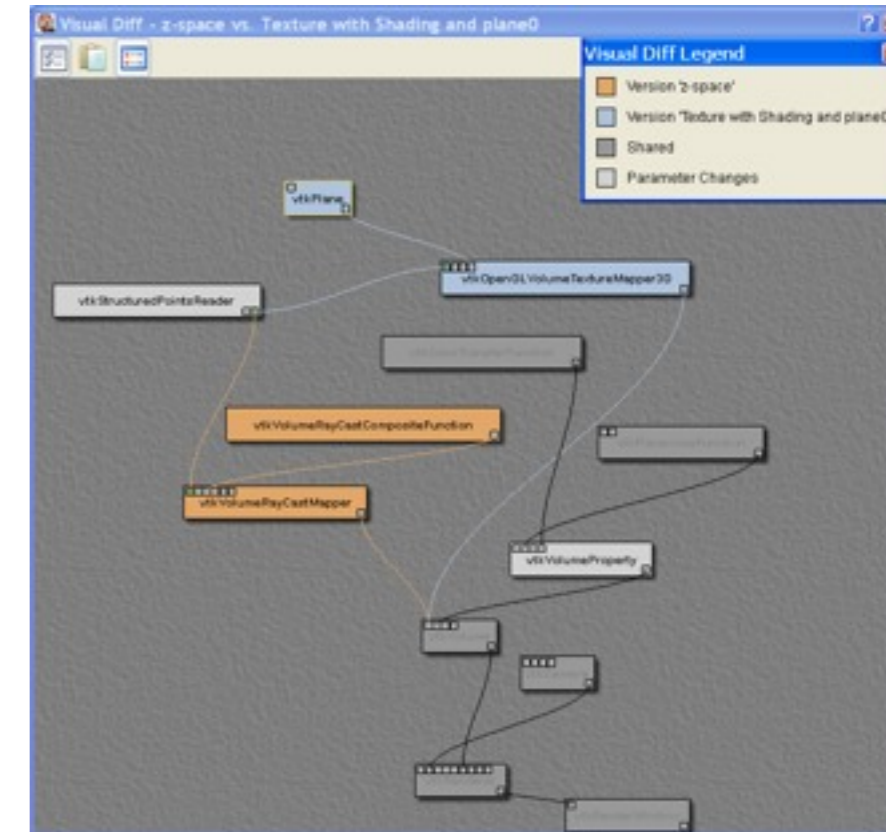
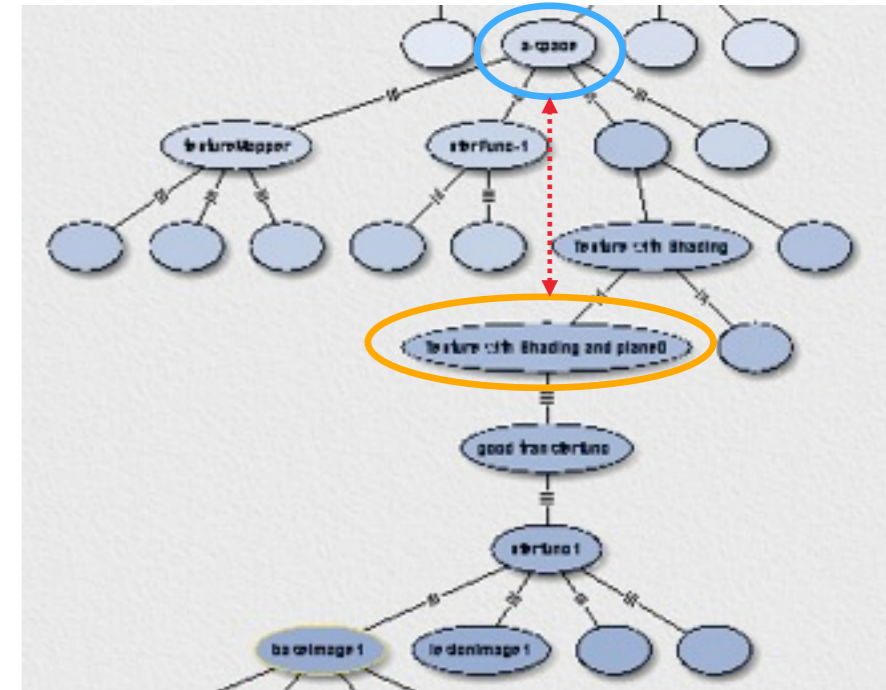
- No need to compute subgraph isomorphism!
- A vistrail is a rooted tree: all nodes have a common ancestor — diffs are well-defined and simple to compute

$$vt_1 = X_i \circ X_{i-1} \circ \dots \circ X_1 \circ \emptyset$$

$$vt_2 = X_j \circ X_{j-1} \circ \dots \circ X_1 \circ \emptyset$$

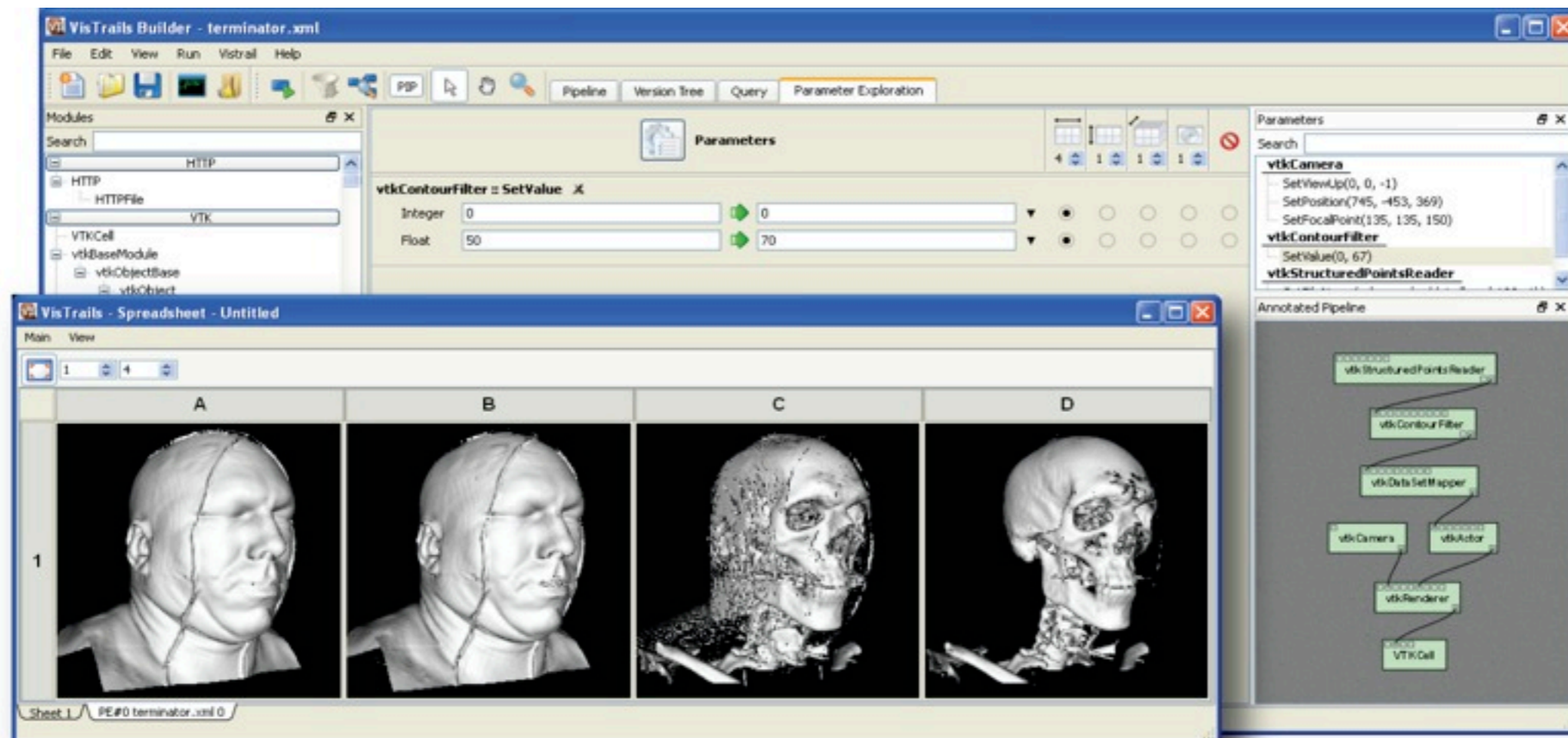
$$vt_1 - vt_2 = \{X_i, X_{i-1}, \dots, X_1, \emptyset\} - \{X_j, X_{j-1}, \dots, X_1, \emptyset\}$$

- 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



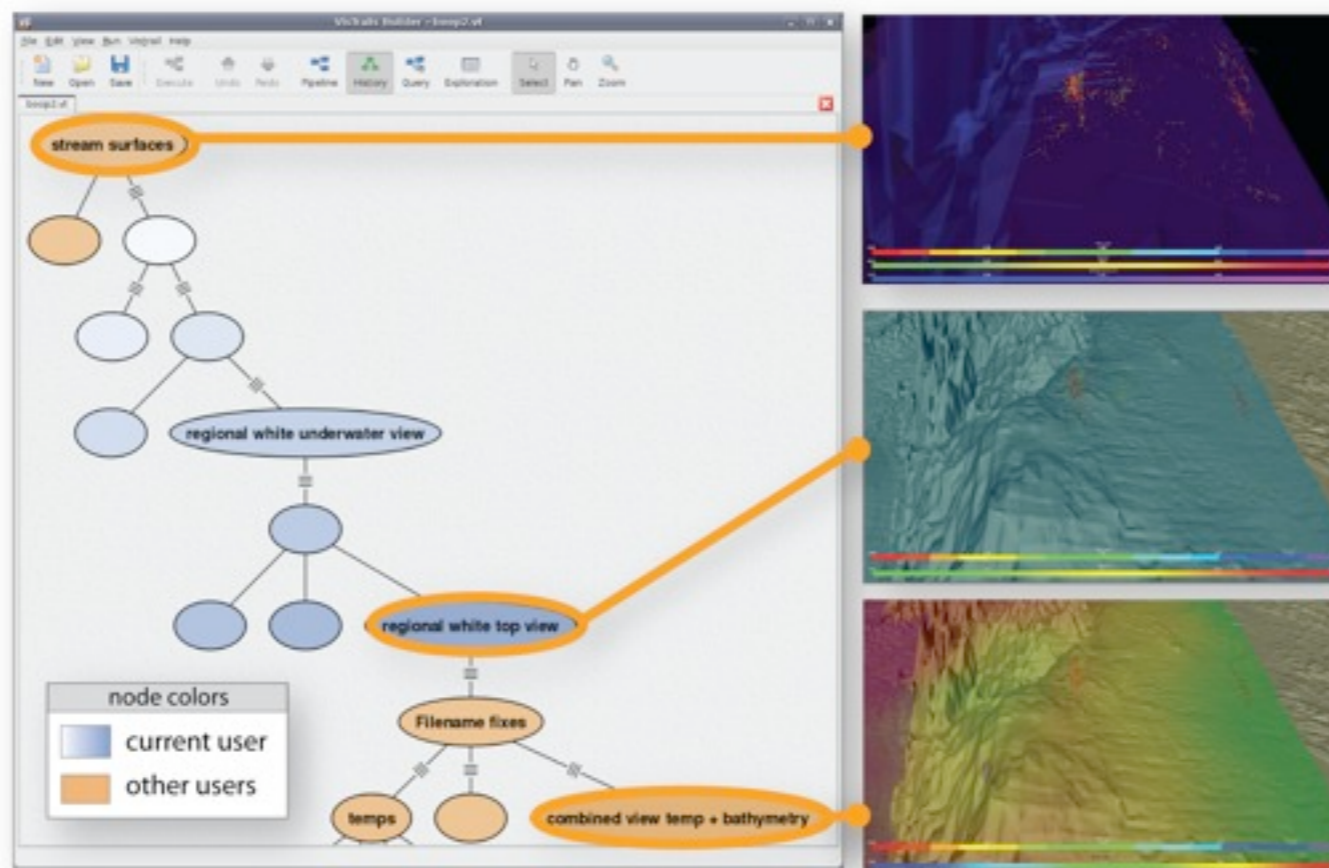
[Freire et al., IPAW 2006]

Exploring the Change Space

- Scripting workflows: Parameter explorations are simple to specify and apply
- Exploration of parameter space for a workflow \mathbf{v}_t
- $(\text{setParameter}(\text{id}_n, \text{value}_n) \circ \dots \circ (\text{setParameter}(\text{id}_1, \text{value}_1) \circ \mathbf{v}_t))$
- Exploration of multiple workflow specifications
 $(\text{addModule}(\text{id}_i, \dots) \circ (\text{deleteModule}(\text{id}_i) \circ \mathbf{v}_1))$
...
 $(\text{addModule}(\text{id}_i, \dots) \circ (\text{deleteModule}(\text{id}_i) \circ \mathbf{v}_n))$
- 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



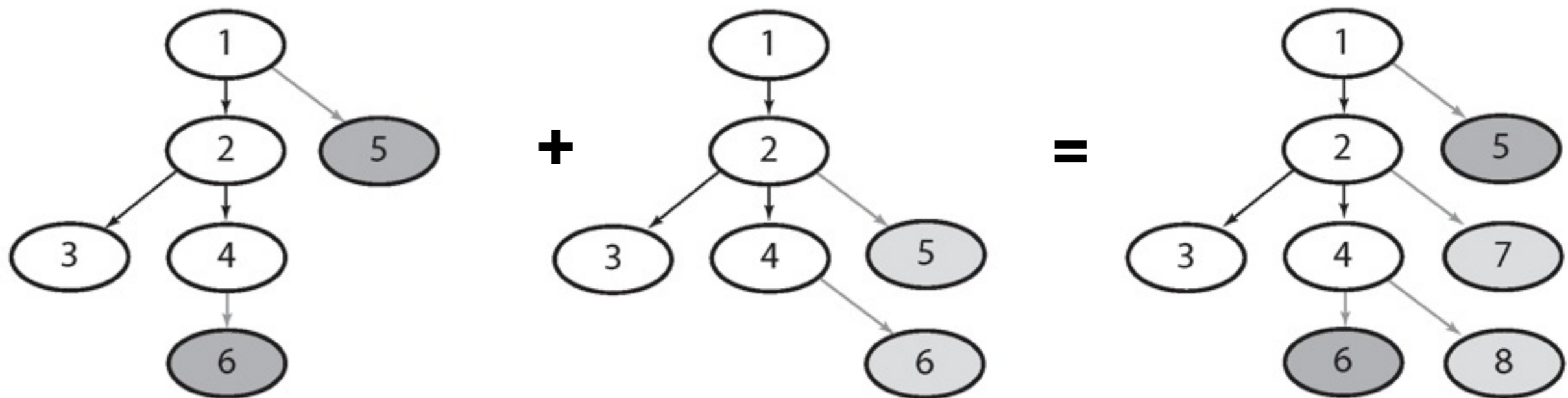
[Ellkvist et al., IPAW 2008]

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
- Synchronization is simple—provenance is monotonic
- 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

VisTrails Synchronization

- Version tree is monotonic
 - Actions are always added, never deleted
- Merging two vistrails is simple



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

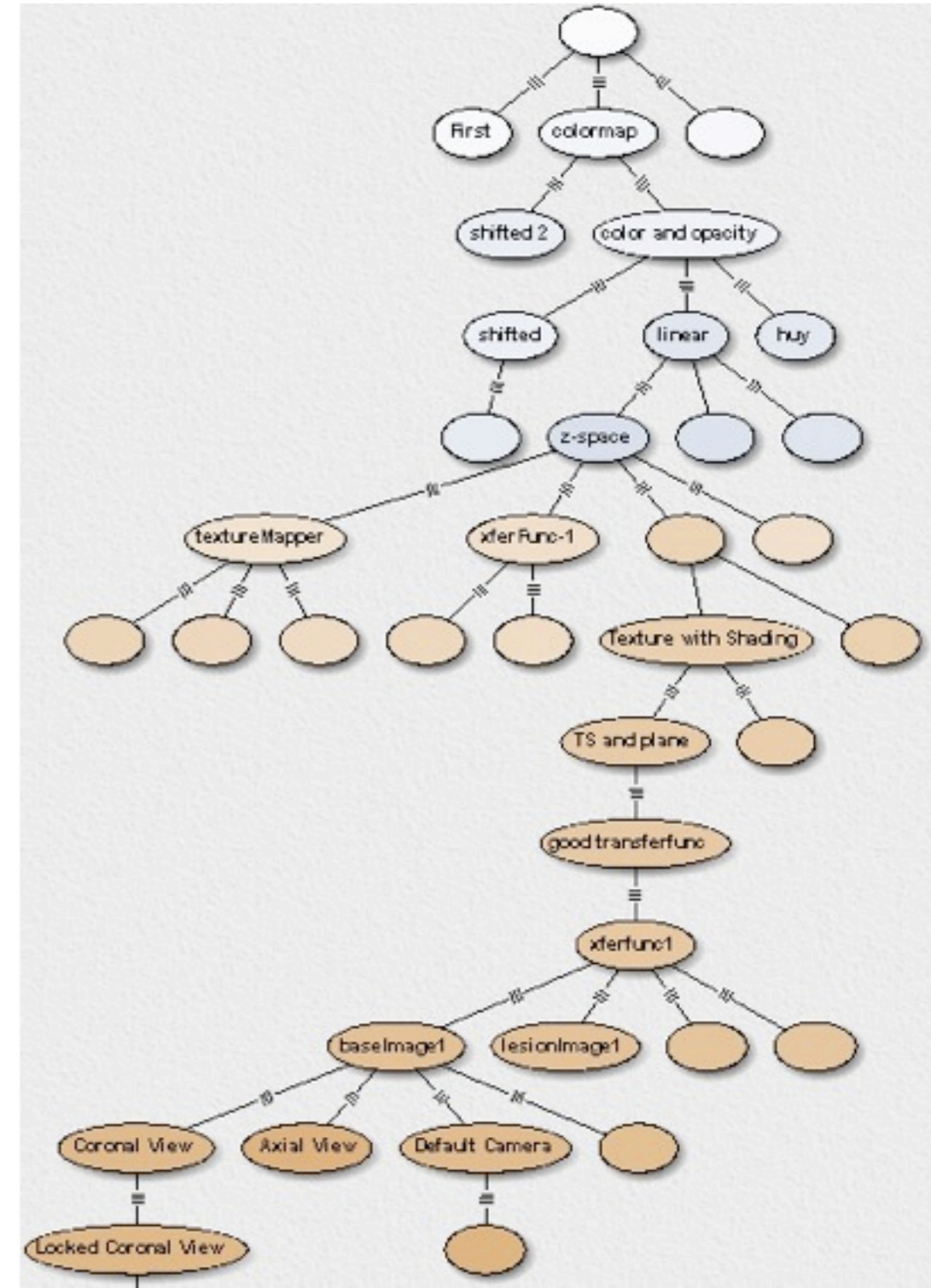
Exploring and Re-Using Provenance



Interacting with Workflow Evolution

Provenance

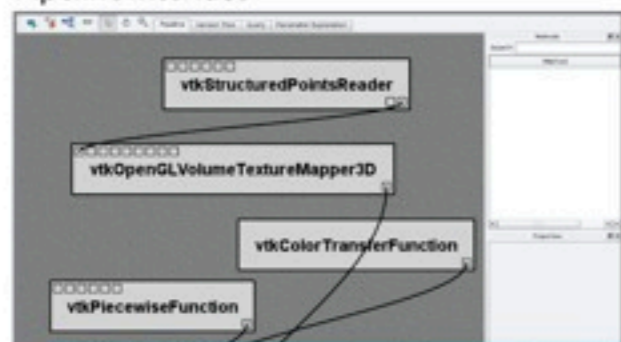
- Storing detailed information is important
- Need appropriate user interface to
 - leverage information, and
 - deal with the information overload
- Understanding the history
 - Different colors for different users
 - Node age represented by saturation level
- Create views over the version tree



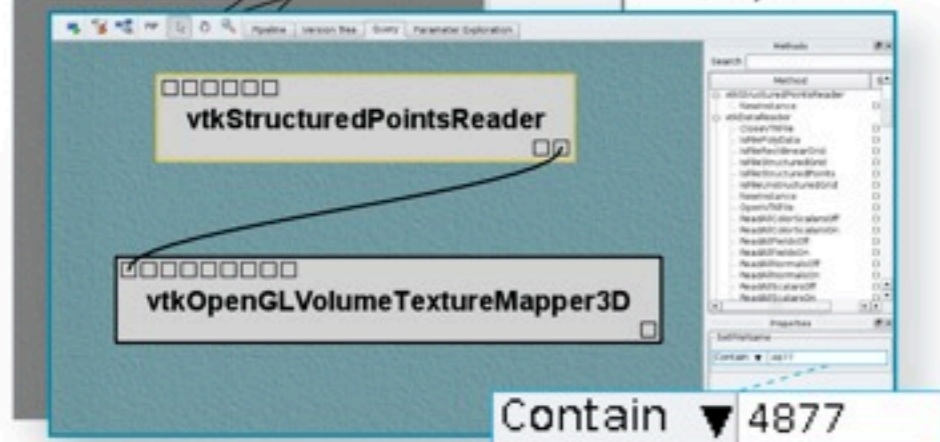
Querying Workflows and Provenance

- Workflow and provenance are graphs: hard to specify queries using text!
- Querying workflows by example [Scheidegger et al., TVCG 2007]
- WYSIWYQ -- What You See Is What You Query
- Interface to create workflow is same as to query

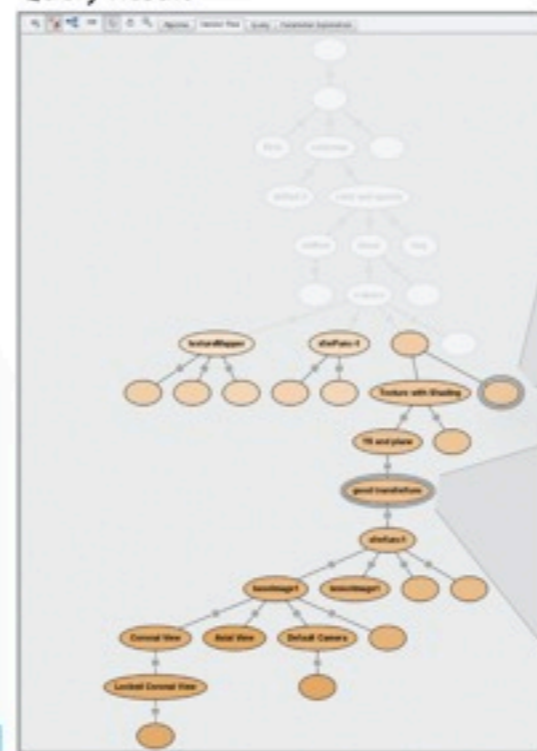
Pipeline Interface



Query Interface



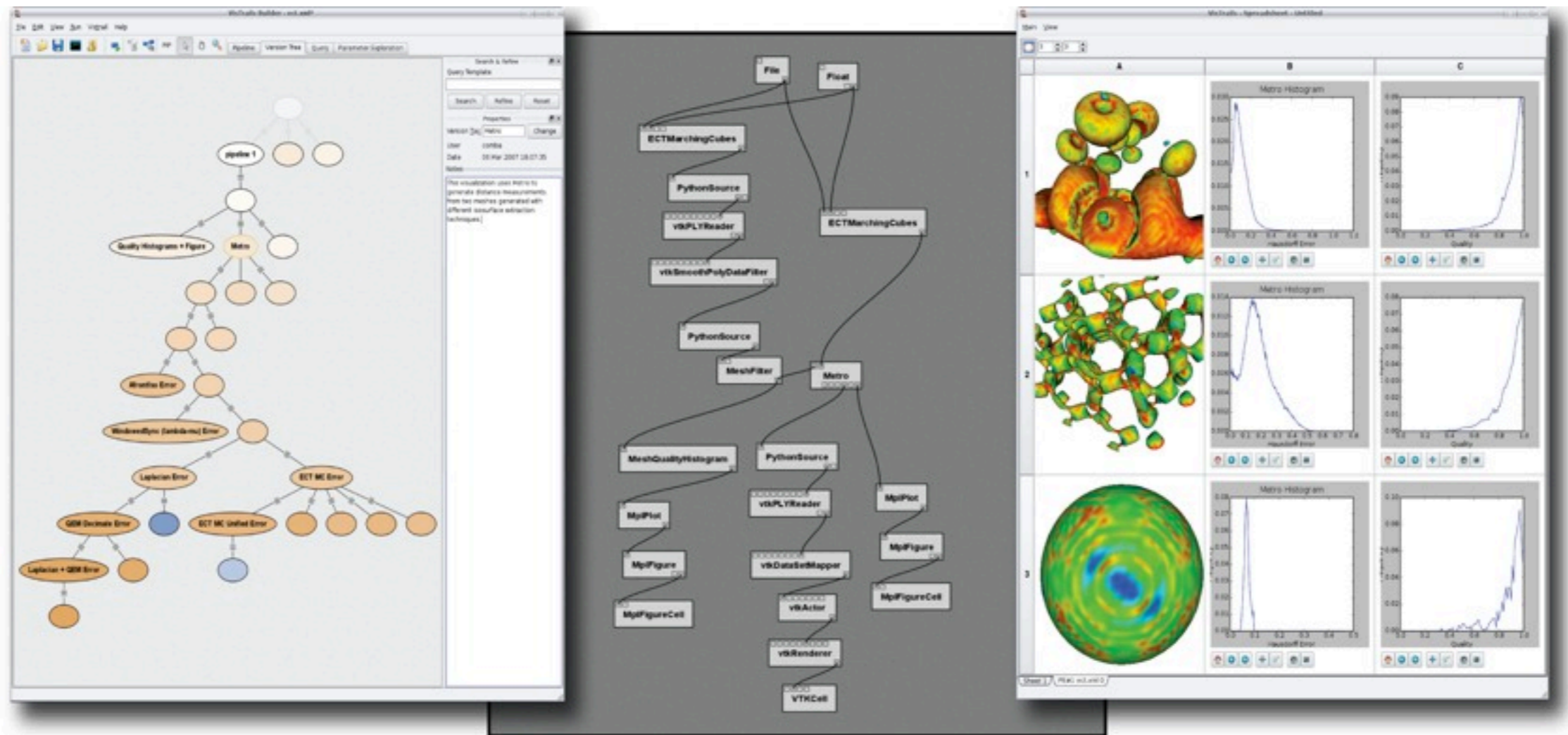
Query Result



Refining Workflows

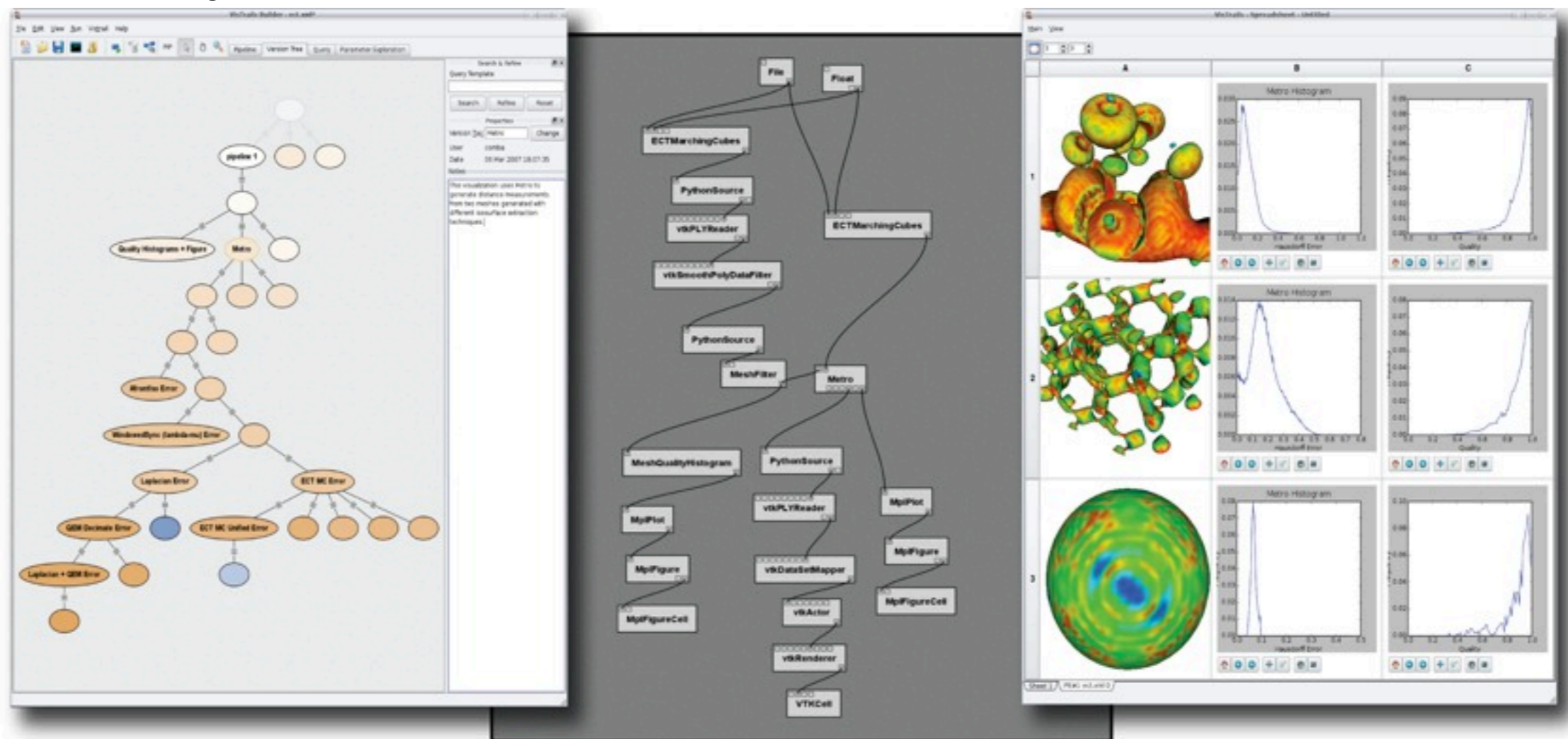
- Complex workflows are hard to create
 - Programming expertise
 - Domain knowledge
 - Familiarity with different tools

Steep learning curve



Refining Workflows

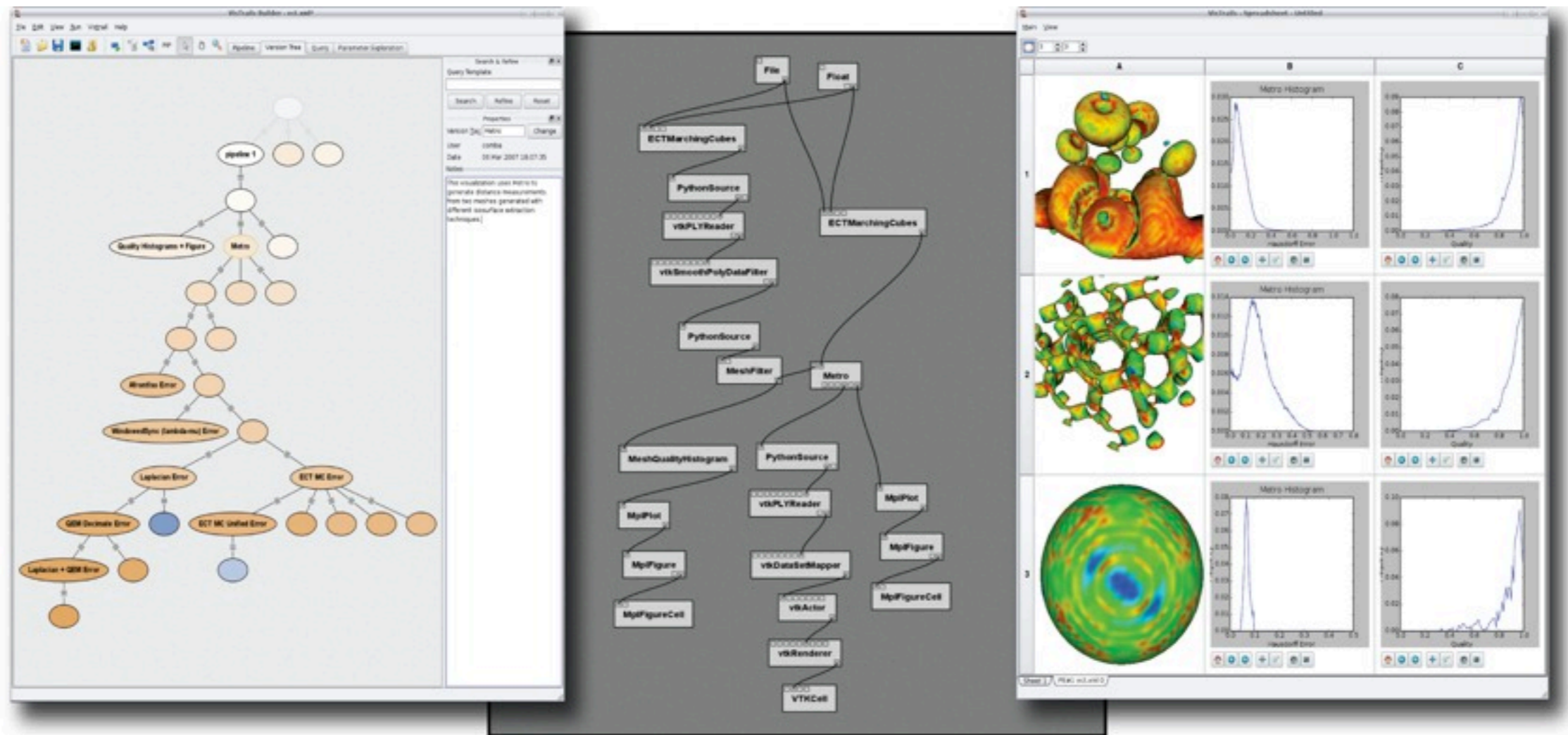
- Complex workflows are hard to create
 - Programming expertise
 - Domain knowledge
 - Familiarity with different tools



Refining Workflows

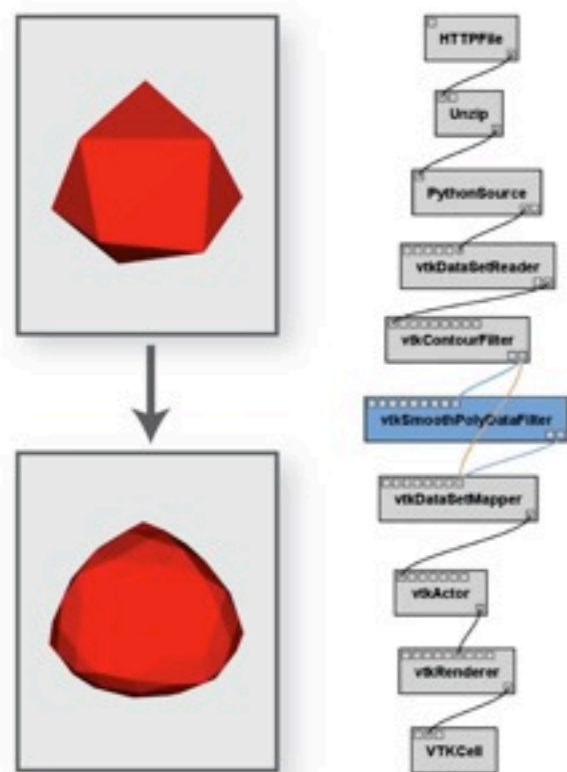
- Complex workflows are hard to create
 - Programming expertise
 - Domain knowledge
 - Familiarity with different tools

Steep learning curve

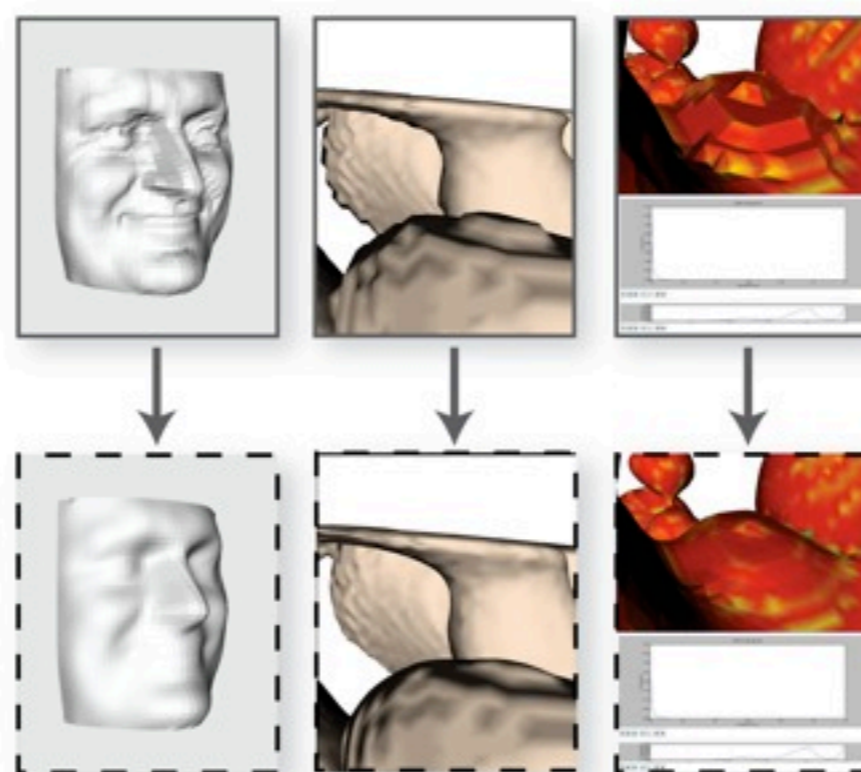


Refining Analyses by Analogy

- 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 [Scheidegger et al, IEEE TVCG 2007]



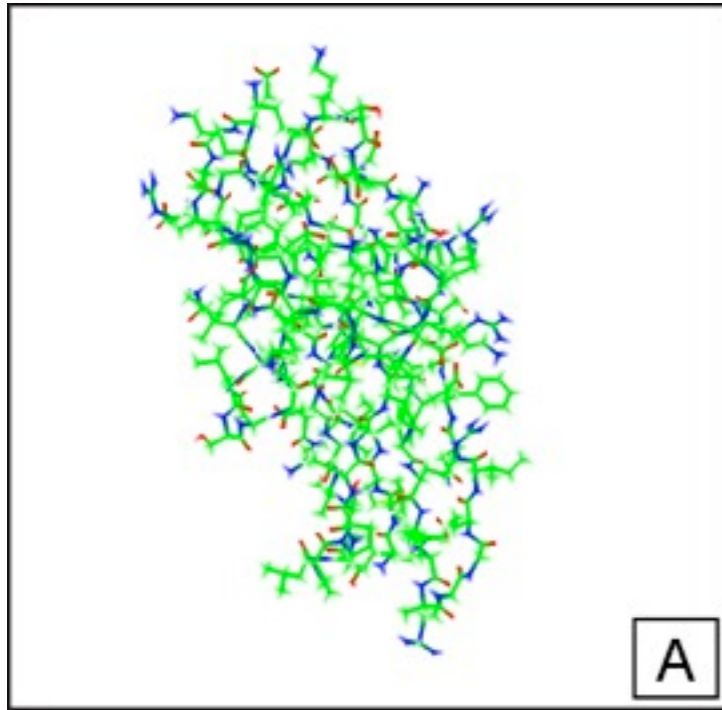
Analogy Template



Automatically constructed visualizations

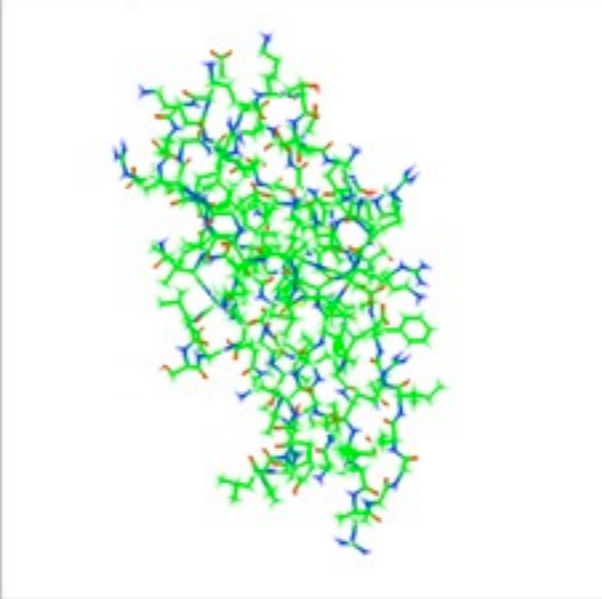


Refining Workflows by Analogy



is to

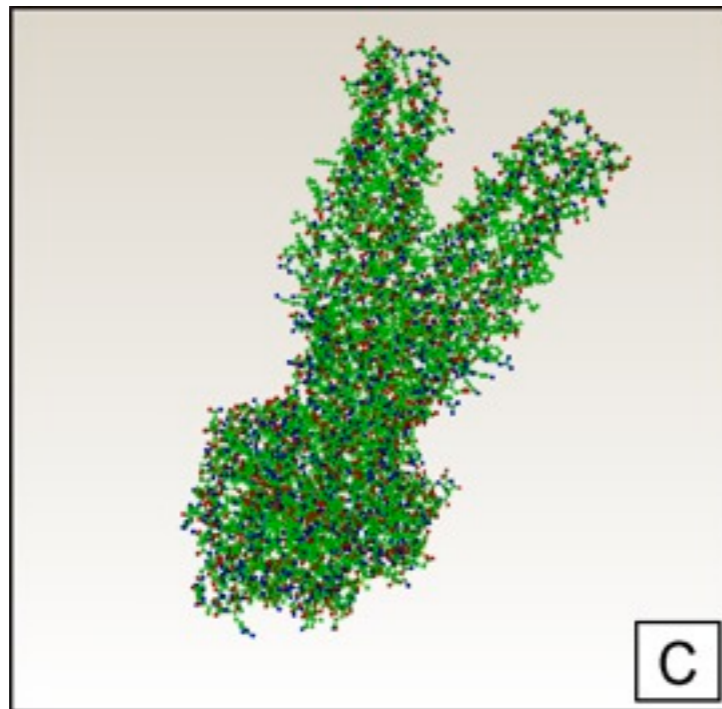
PDB Report



Protein Title	NEURAL CELL ADHESION MOLECULE, MODULE 2, NMR, 20 STRUCTURES
Authors	P.H.JENSEN, V.SOROKA, N.K.THOMSEN, V.BEREZIN, E.BOCK, F.M.POULSEN
Atom Count	C: 9560 H: 15440 N: 2580 O: 2680 S: 60
Links	PDB Entry

B

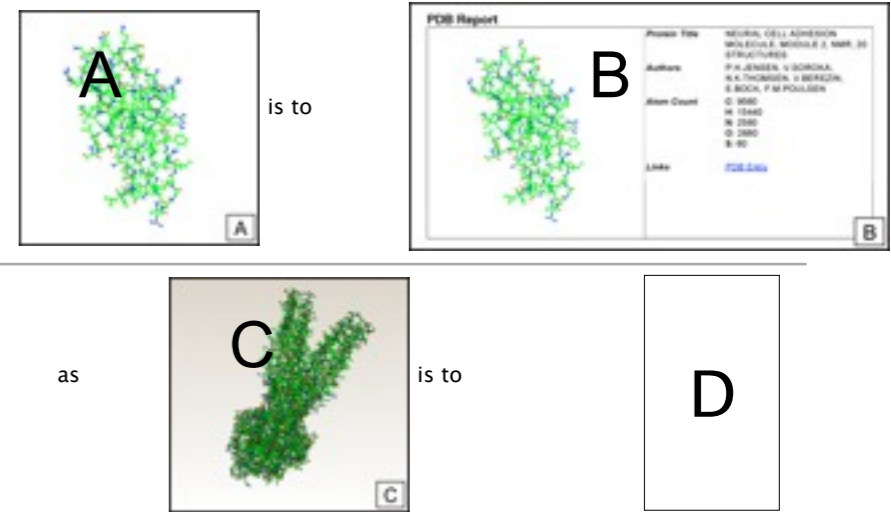
as



is to

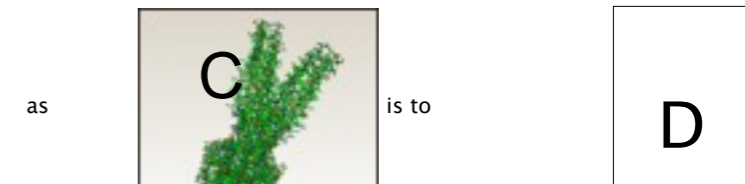
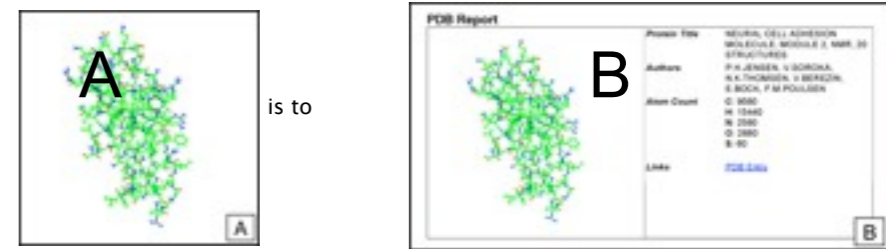
?

Refining Workflows by Analogy

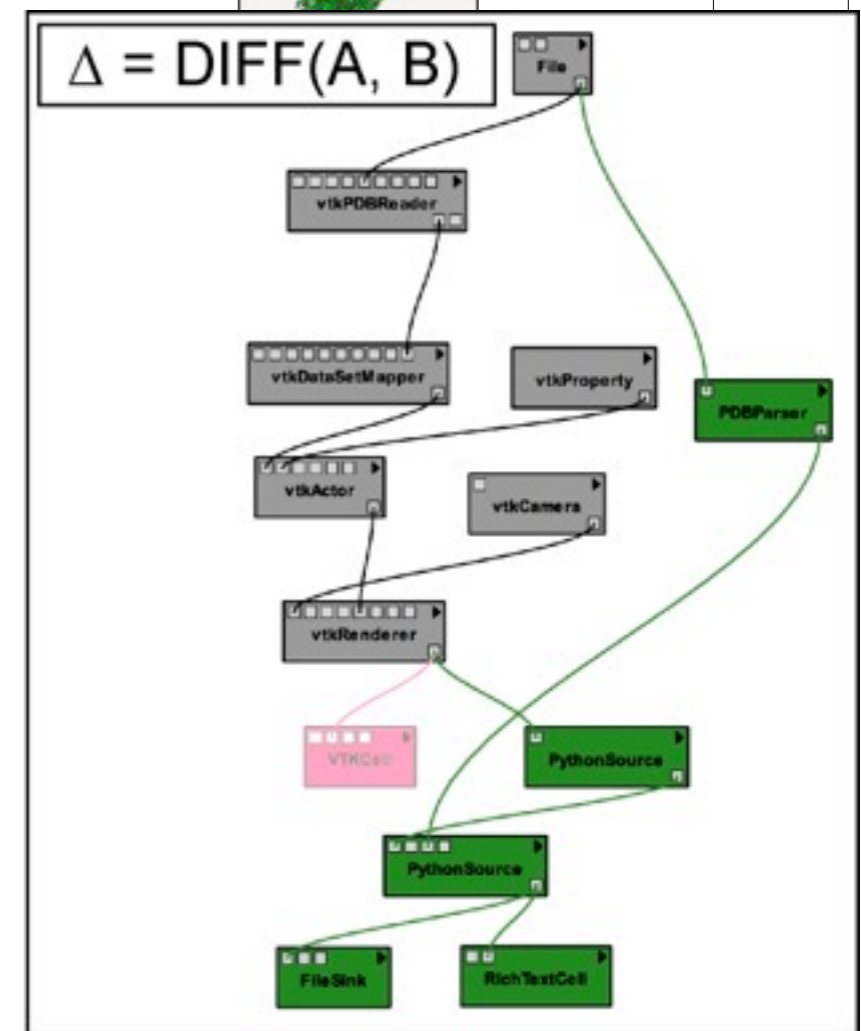


- Compute difference: $\Delta(A,B)$
 - Just like a patch!
 - But...
 - $D = \Delta(A,B) \circ C$ may not be a valid workflow
- Find correspondences between A and C: $\text{map}(A,C)$
 - Diffuse similarity scores across the product graph $A \times C$ using Eigenvalue decompositions
- Compute mapped difference $\Delta_{AC}(A,B) = \text{map}(A,C) \Delta(A,B)$
- $D = \Delta_{AC}(A,B) \circ C$

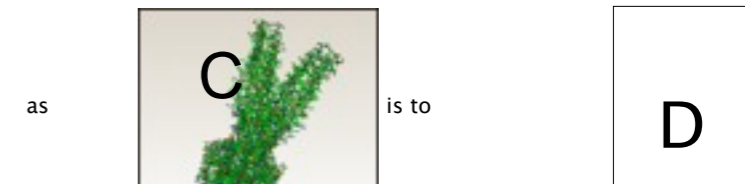
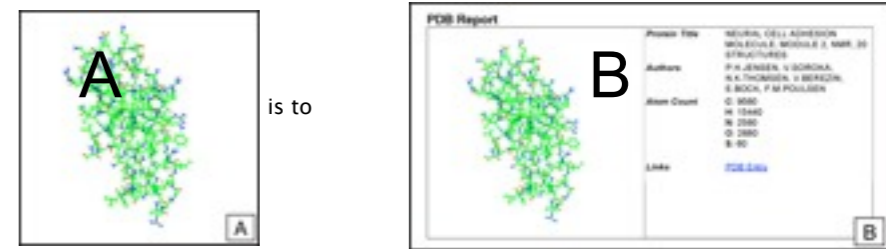
Refining Workflows by Analogy



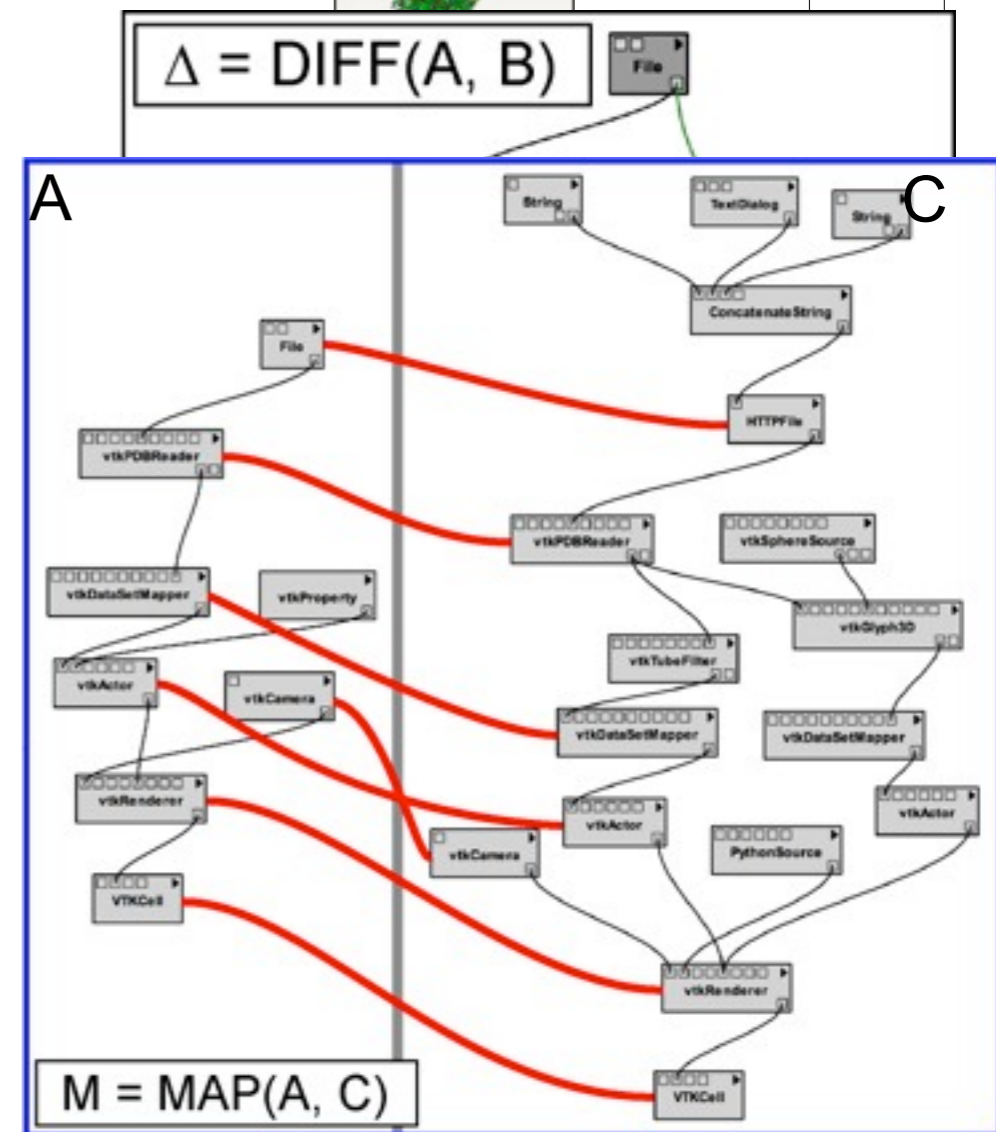
- Compute difference: $\Delta(A, B)$
 - Just like a patch!
 - But...
 - $D = \Delta(A, B) \circ C$ may not be a valid workflow
- Find correspondences between A and C: $\text{map}(A, C)$
 - Diffuse similarity scores across the product graph $A \times C$ using Eigenvalue decompositions
- Compute mapped difference $\Delta_{AC}(A, B) = \text{map}(A, C) \Delta(A, B)$
- $D = \Delta_{AC}(A, B) \circ C$



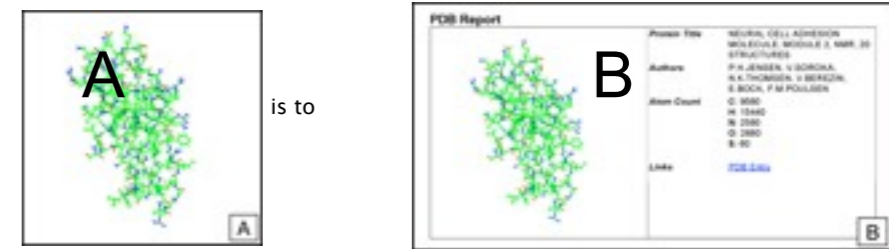
Refining Workflows by Analogy



- Compute difference: $\Delta(A, B)$
 - Just like a patch!
 - But...
 - $D = \Delta(A, B) \circ C$ may not be a valid workflow
- Find correspondences between A and C: $\text{map}(A, C)$
 - Diffuse similarity scores across the product graph $A \times C$ using Eigenvalue decompositions
- Compute mapped difference $\Delta_{AC}(A, B) = \text{map}(A, C) \Delta(A, B)$
- $D = \Delta_{AC}(A, B) \circ C$

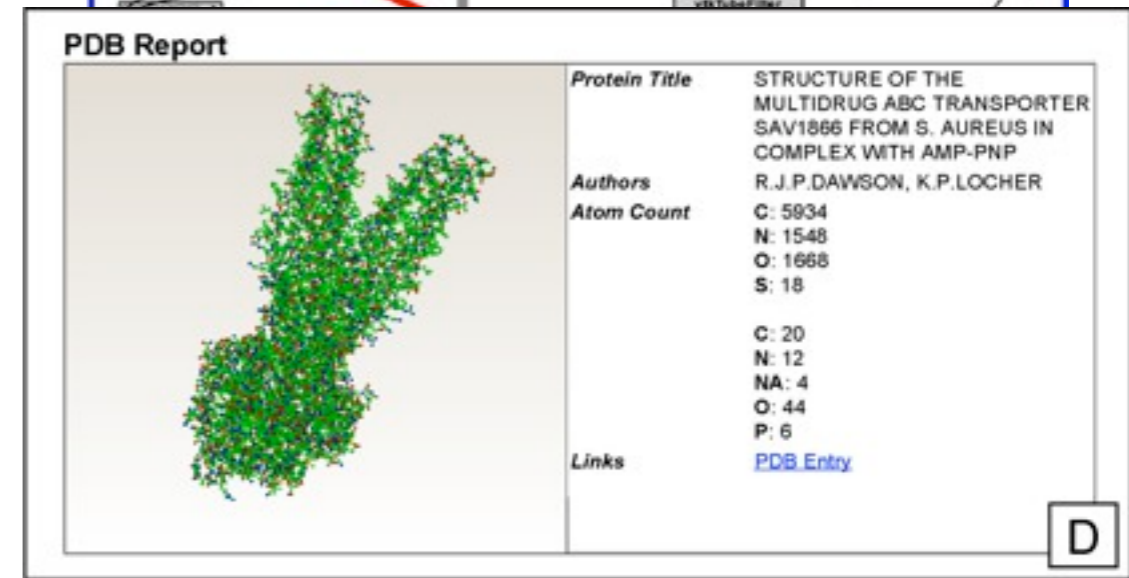
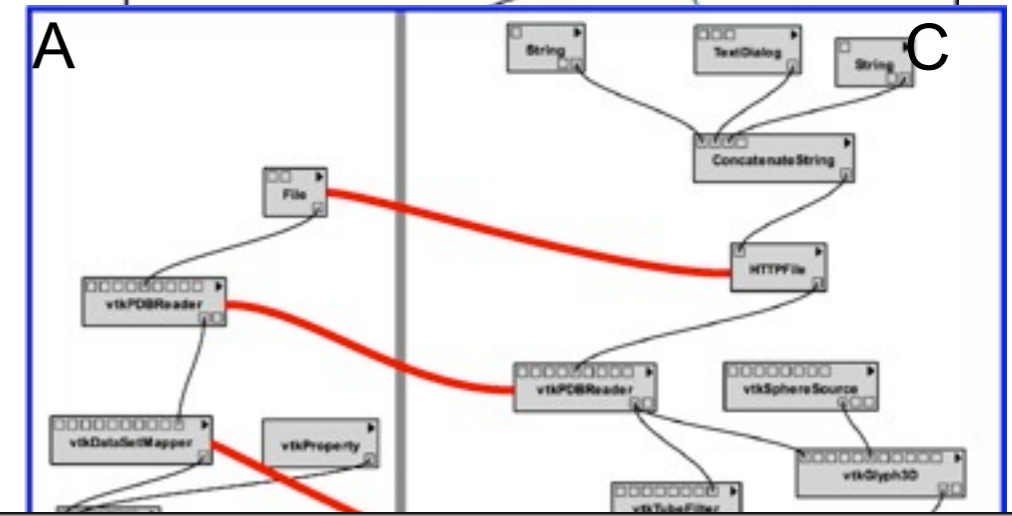


Refining Workflows by Analogy

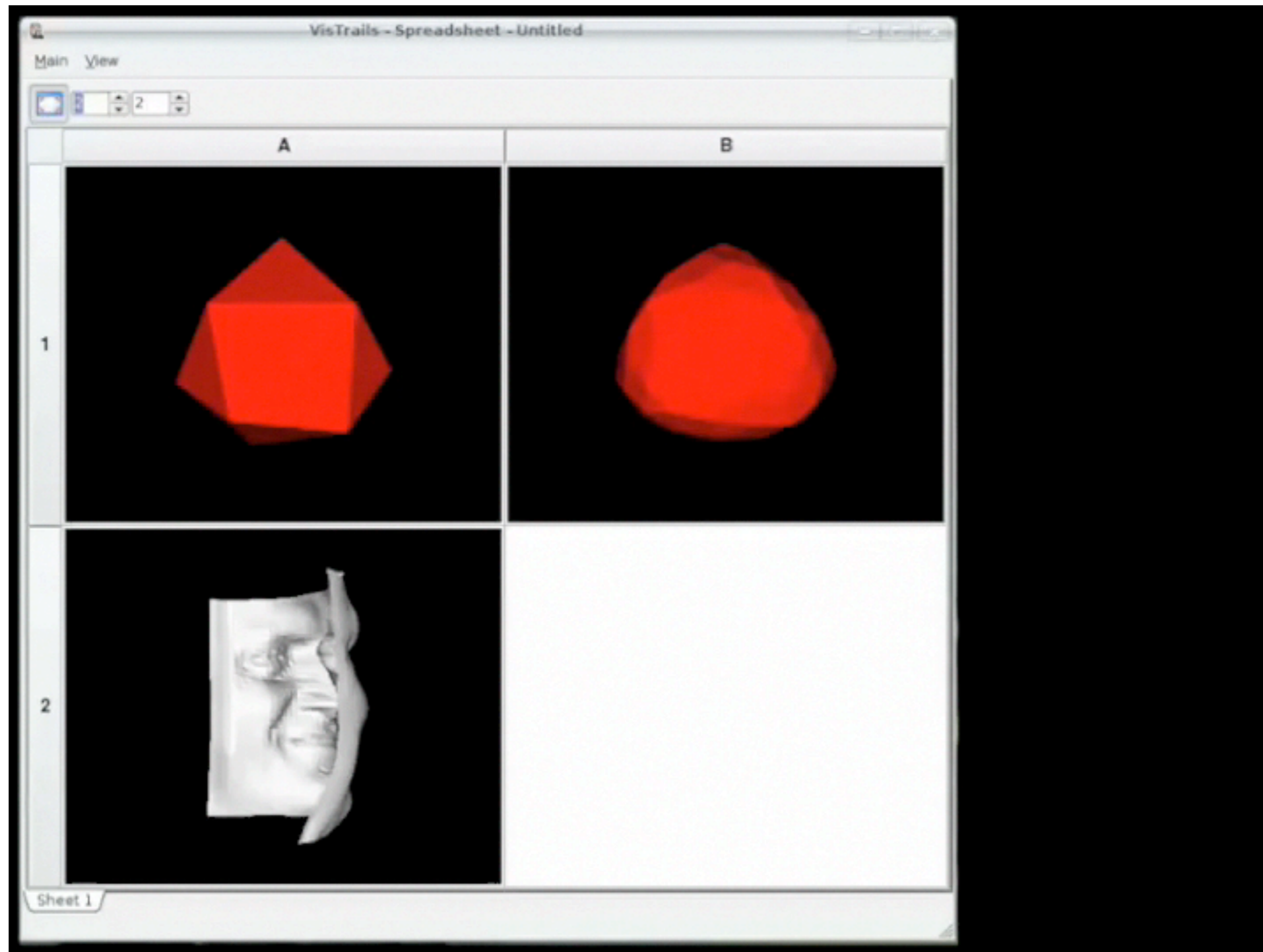


$\Delta = \text{DIFF}(A, B)$

- Compute difference: $\Delta(A, B)$
 - Just like a patch!
 - But...
 - $D = \Delta(A, B) \circ C$ may not be a valid workflow
- Find correspondences between A and C: $\text{map}(A, C)$
 - Diffuse similarity scores across the product graph $A \times C$ using Eigenvalue decompositions
- Compute mapped difference $\Delta_{AC}(A, B) = \text{map}(A, C) \Delta(A, B)$
- $D = \Delta_{AC}(A, B) \circ C$



Generating Visualizations by Analogy

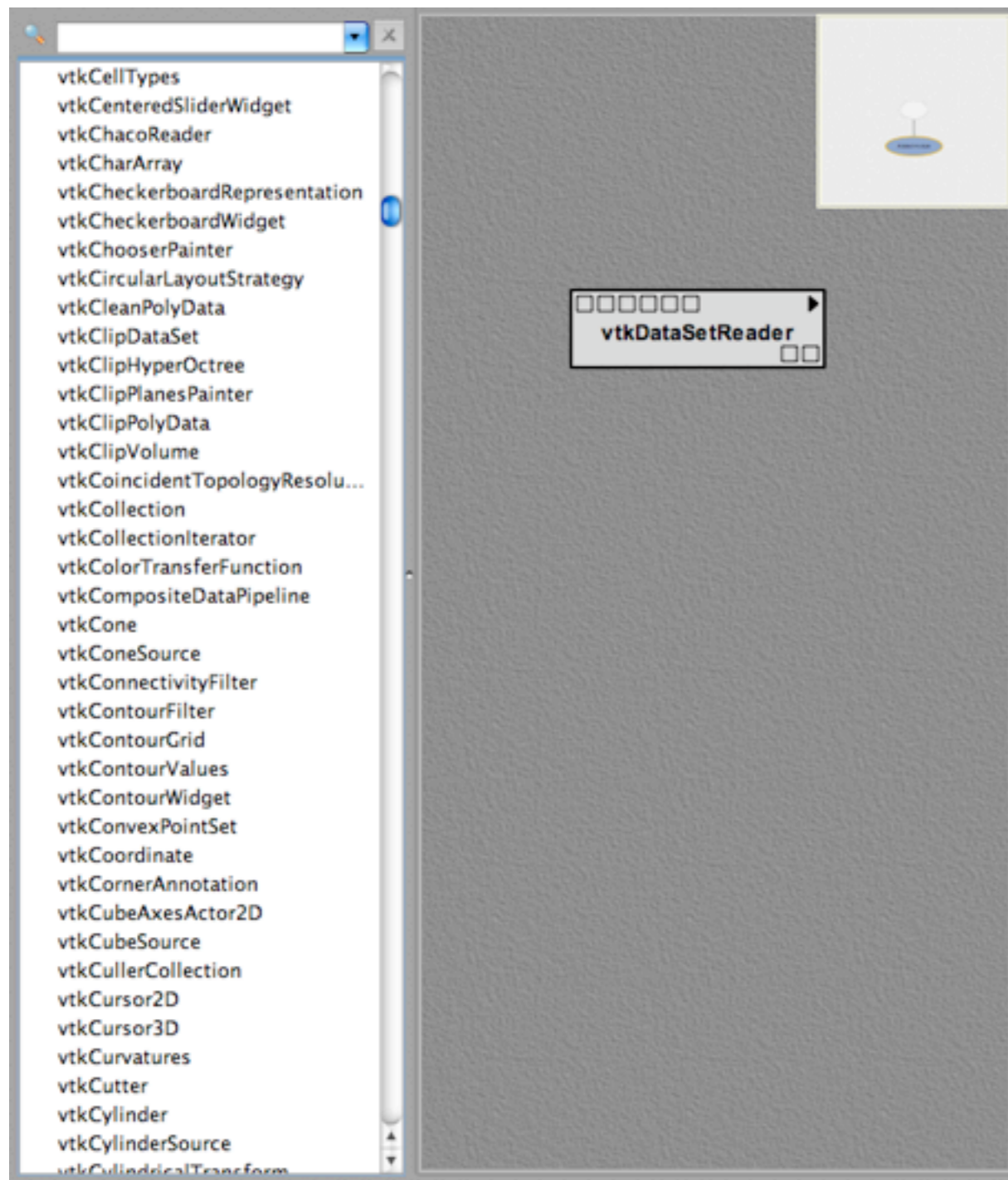


[Scheidegger et al, IEEE TVCG 2007]

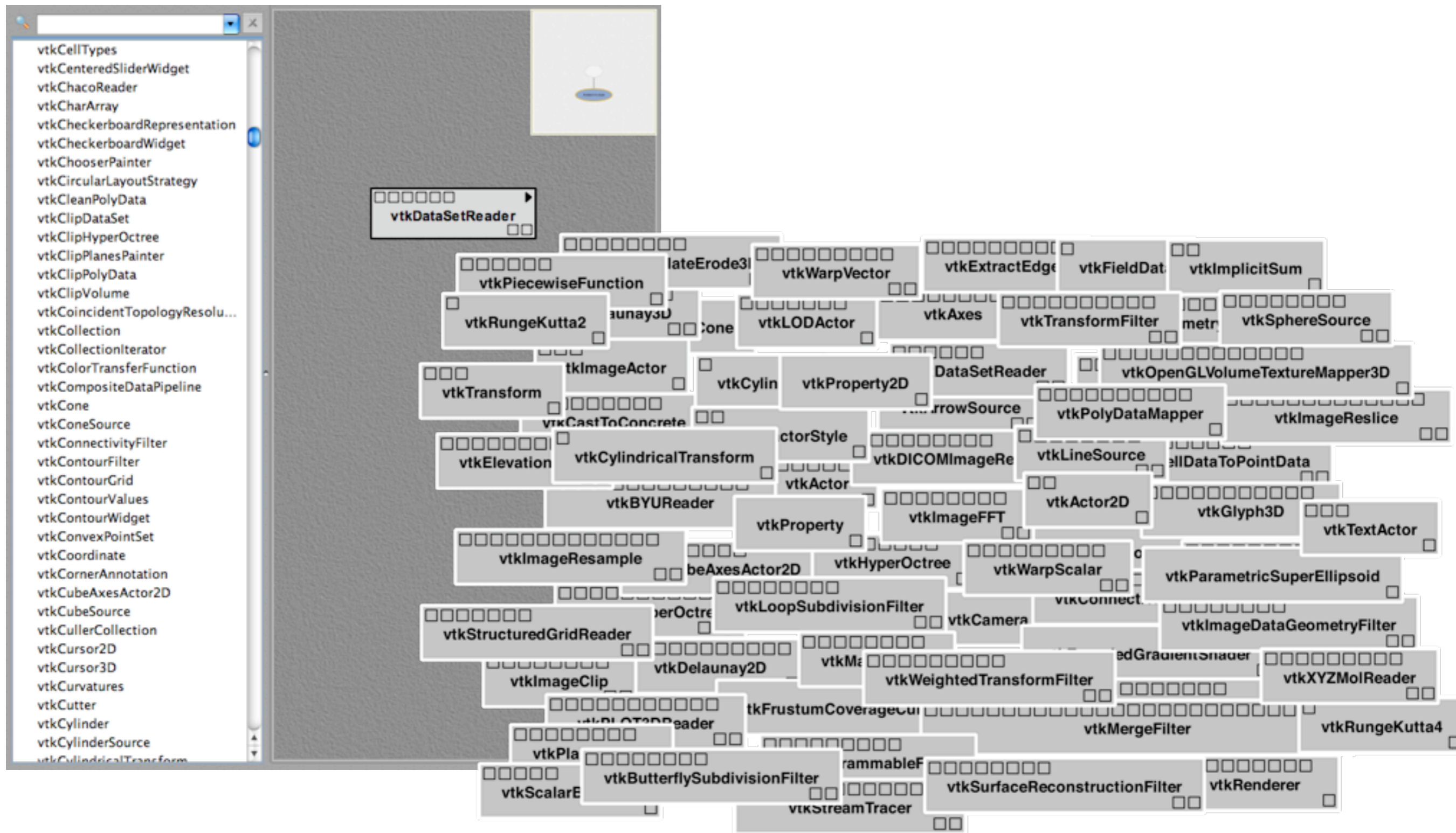
Demo 2

Querying and Searching
Workflow refinement by analogy

The need for Guidance in Workflow Design

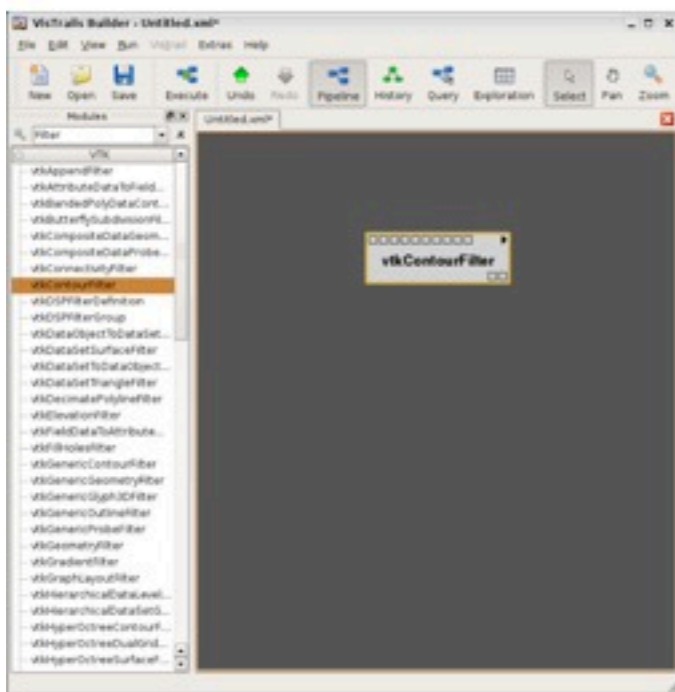


The need for Guidance in Workflow Design

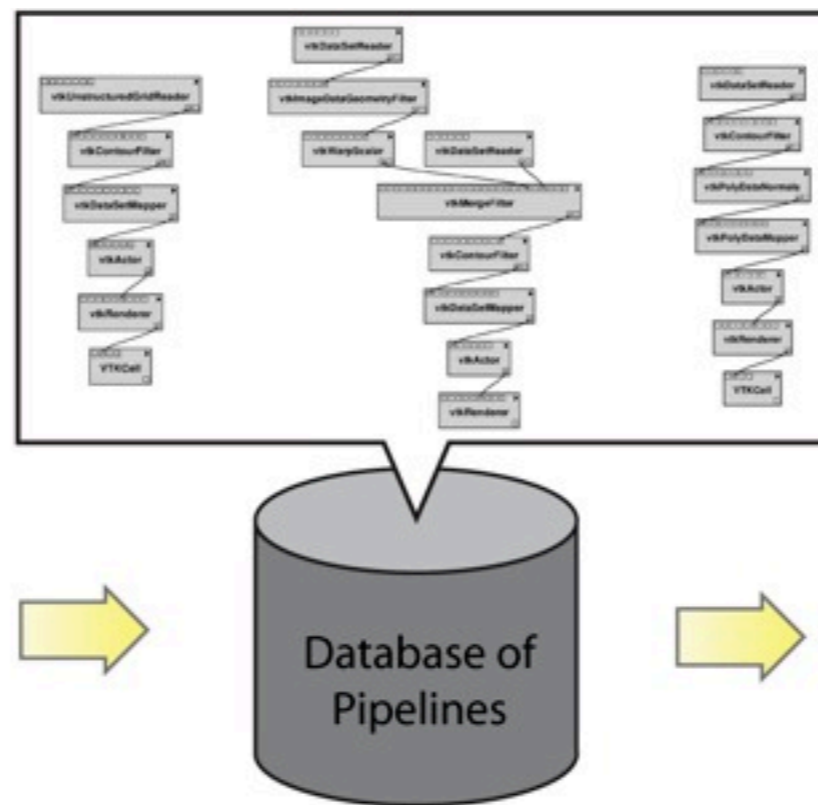


VisComplete: A Workflow Recommendation System

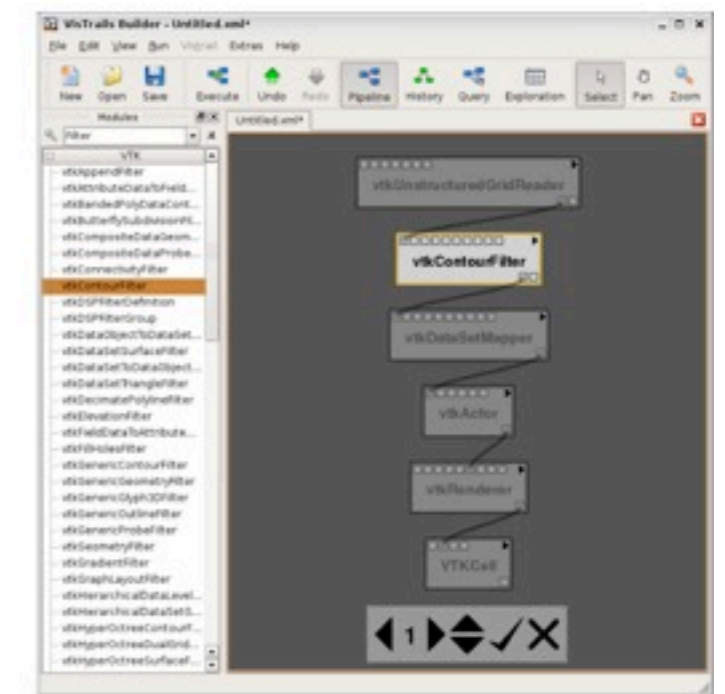
- Mine provenance collection: Identify fragments that co-occur in a collection of workflows
- Predict sets of likely workflow additions to a given partial workflow
- Similar to a Web browser suggesting URL completions



(a)



(b)

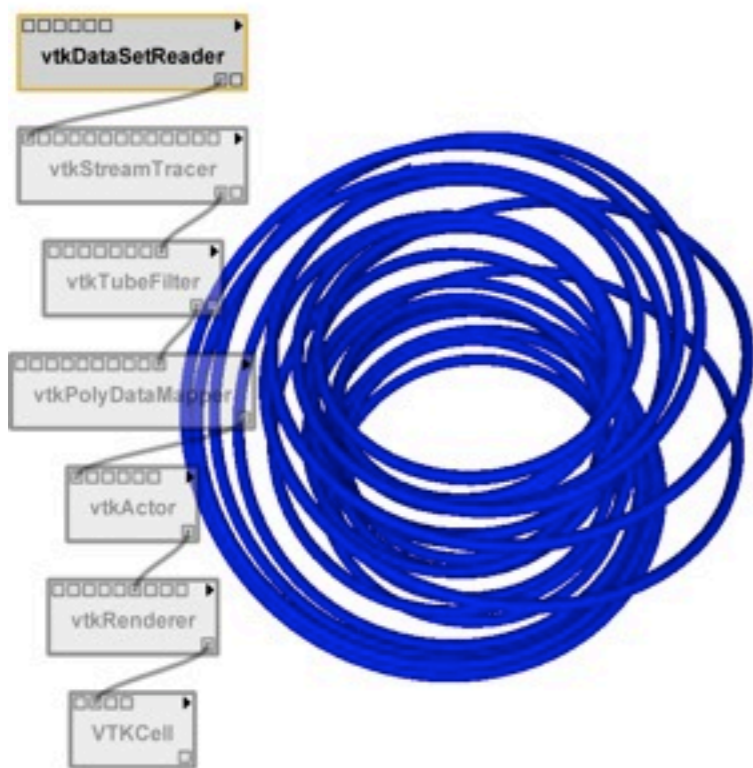


(c)

[Koop et al., IEEE Vis 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

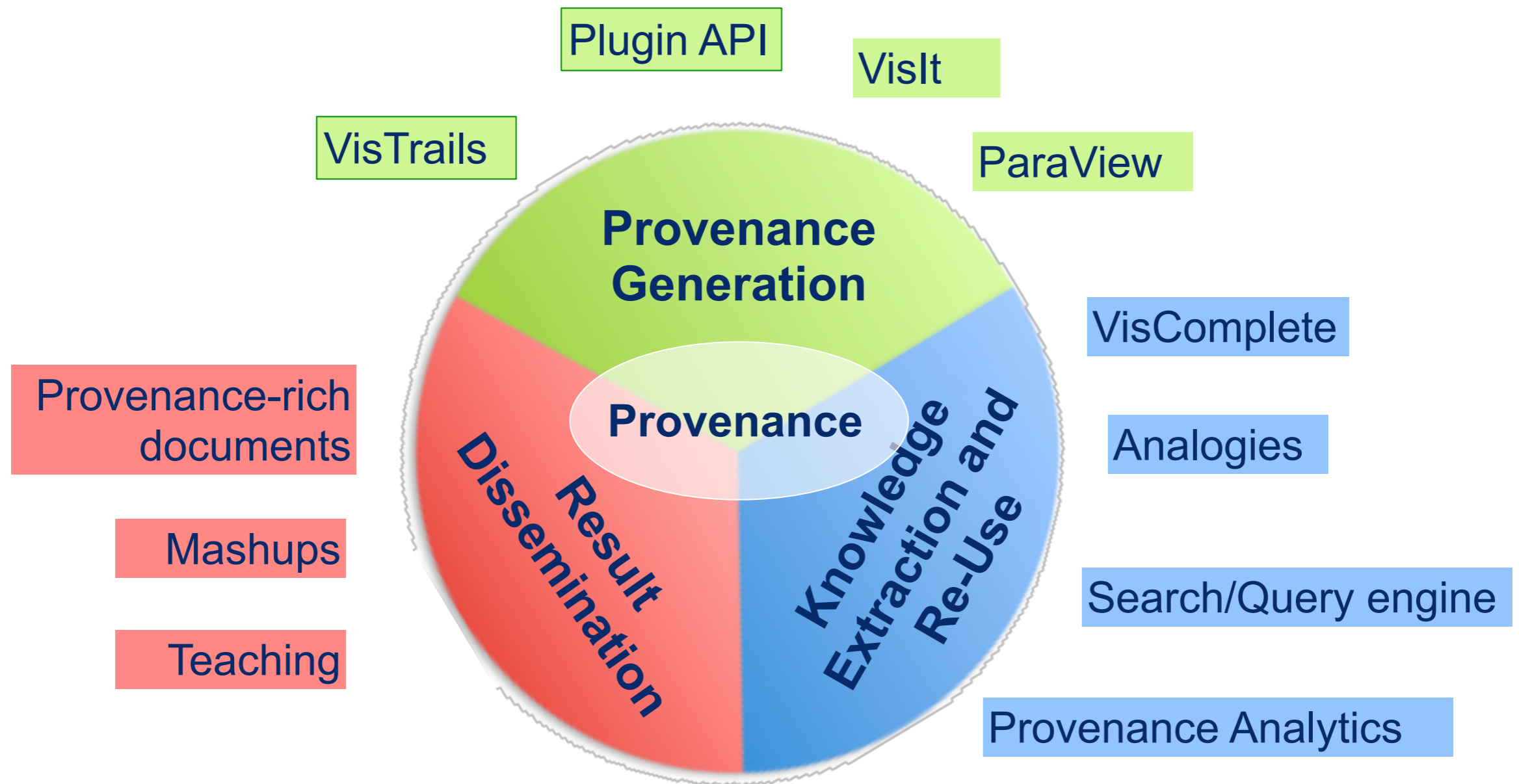


VisComplete: Data-driven Suggestions for Visualization Systems

Emerging Applications



Provenance Infrastructure



Scientific Publications and Provenance

J Appl Physiol 98: 2191–2196, 2005.

First published March 17, 2005; doi:10.1152/jappphysiol.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/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 ~ 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., $\sim 7\%$). 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 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 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 (± 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 gases 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

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.

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: coyle@mail.utexas.edu).

http://www.jap.org

8750-7587/05 \$8.00 Copyright © 2005 the American Physiological Society

2191

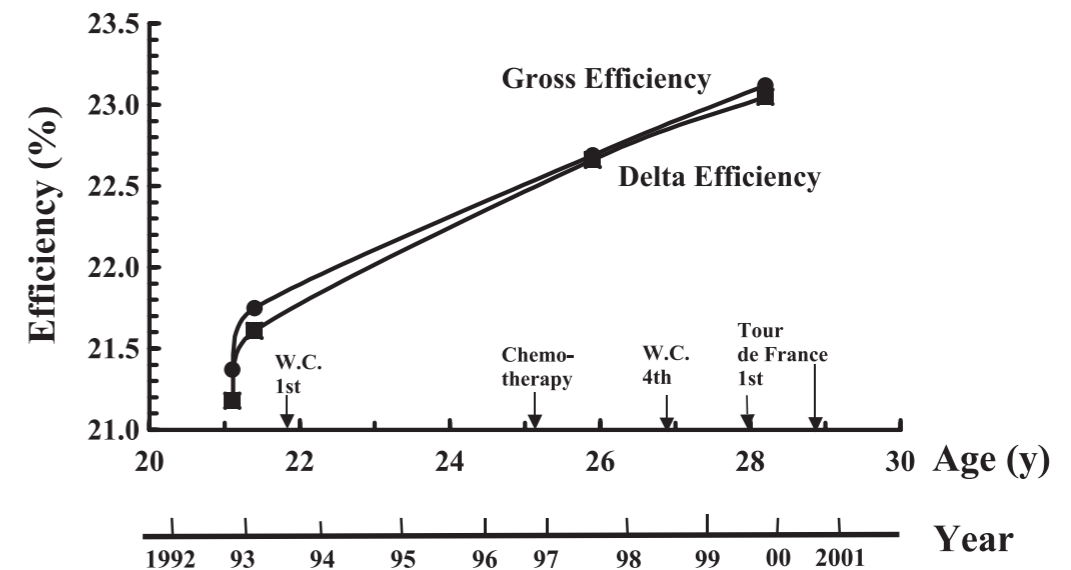


Fig. 1. Mechanical efficiency when bicycling expressed as "gross efficiency" and "delta efficiency" over the 7-yr period in this individual. WC, World Bicycle Road Racing Championships, 1st and 4th place, respectively. Tour de France 1st, Grand Champion of the Tour de France in 1999–2004.

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

J Appl Physiol 98: 2191-2196, 2005.
First published March 17, 2005; doi:10.1152/jap.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, TX 78712

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/jap.00216.2005. This case describes the cyclist Champion of the world (WC), who has now become the oldest cyclist to win the Tour de France. It is hypothesized that the improved muscular efficiency changes in muscle myosin type stimulated by training intensify for 3-6 h on most days.

MUCH HAS BEEN contributed to the understanding of elite endurance athletes in sports such as distance running, cycling, and cross-country skiing. The numerous physiological variables that have been measured to determine an athlete's profile with that of a population of peers (4, 6, 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 the specific physiological determinants of endurance with years of training. Only a few longitudinal reports document the changes in physiological factors that accompany years of continued endurance training at the level performed by elite endurance athletes.

This case study reports on the changes in an individual bicycle racer during a 7-yr period spanning from 1992 to 2004.

"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>

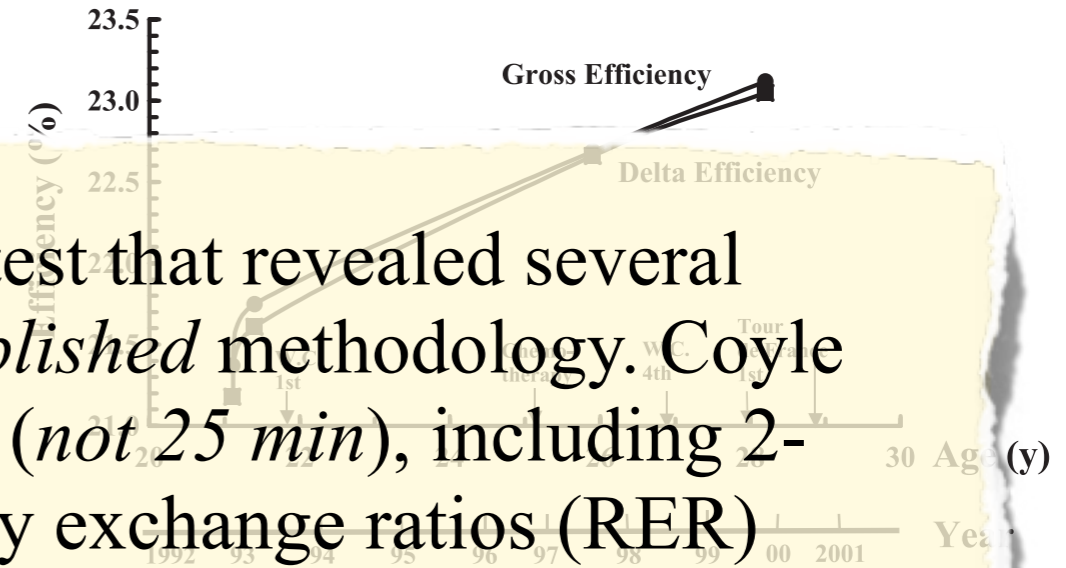


Fig. 1. Muscular efficiency, the bicycling expressed as "gross efficiency" and "delta efficiency" over the 7-yr period in this individual. WC, World Bicycle Road Racing Championships, 1st and 4th place, respectively. Tour de France, 1st place, respectively. Tour de France champion of the Tour de France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

METHODS

On the day of the test, the subject was taken 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).

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: coyle@ruil.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.

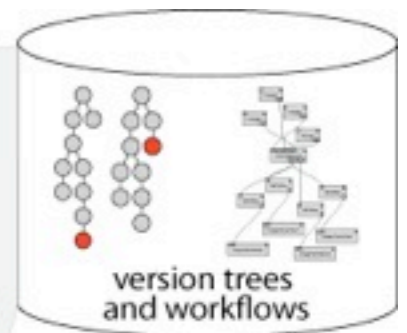
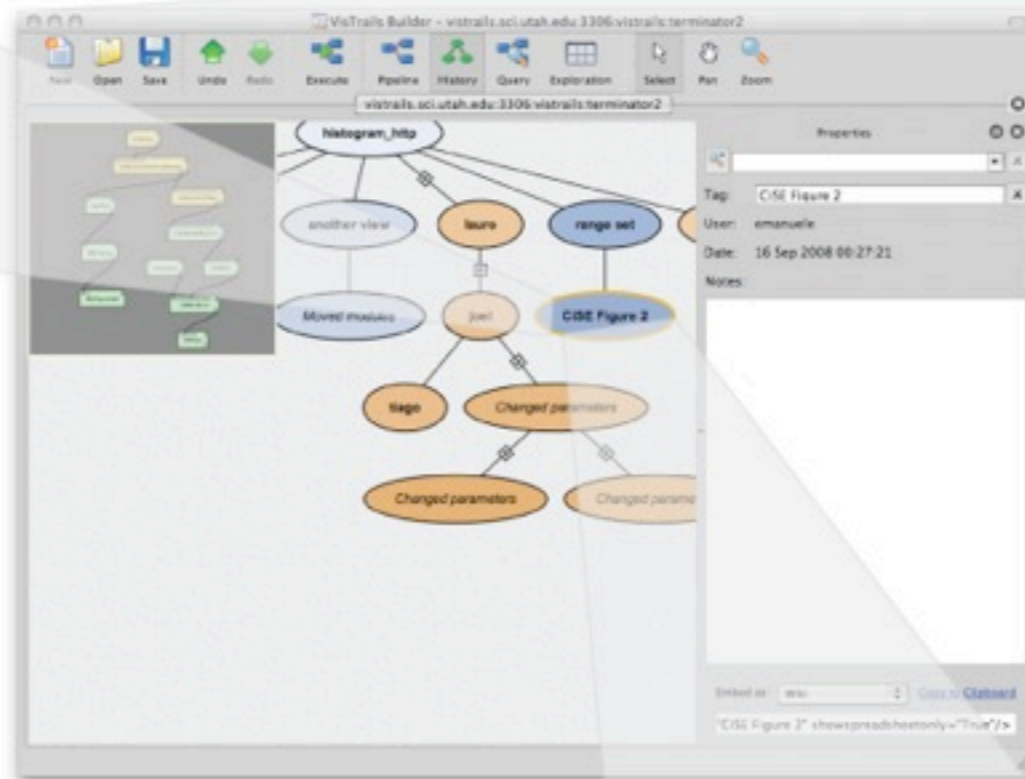
Provenance-Rich Publications

- 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
- Describe more of the discovery process: people only describe successes, can we learn from mistakes?
- Dynamic (interactive) publications
 - Evolve over time
 - Blog/wiki like=> Science 2.0
- Need tools to support this!

Provenance-Rich Documents

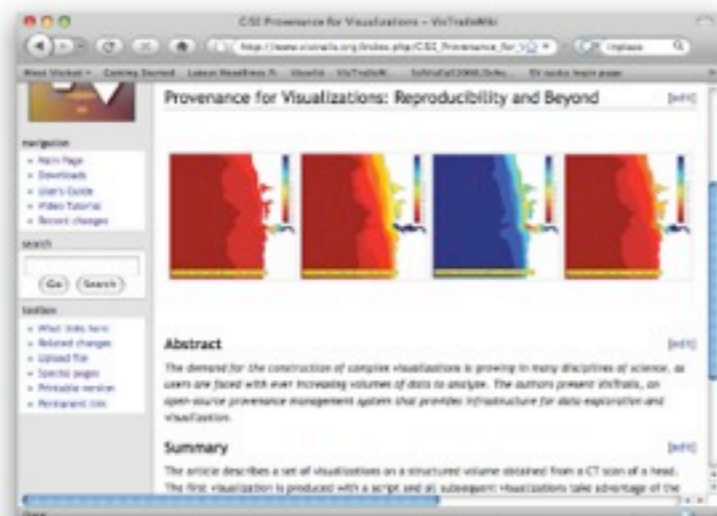


Interactive visualization



version trees and workflows

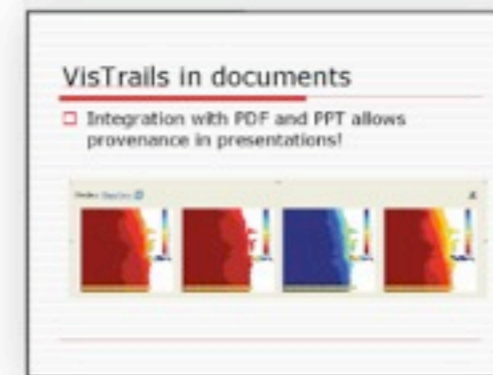
provenance repository



wiki page



pdf document



powerpoint presentation

The Provenance-Rich Paper

Provenance and Teaching (1)

- Leverage provenance to improve the way we teach CS and Science
- <http://www.vistrails.org/index.php/SciVisFall2008>
- Also used at UNC, Linkoping (Sweden), UTEP
- Lecture provenance: student can reproduce results

Provenance and Teaching (1)

- Leverage provenance to improve the way we teach CS and Science
- <http://www.vistrails.org/index.php/SciVisFall2008>
- Also used at UNC, Linkoping (Sweden), UTEP
- Lecture provenance: student can reproduce results

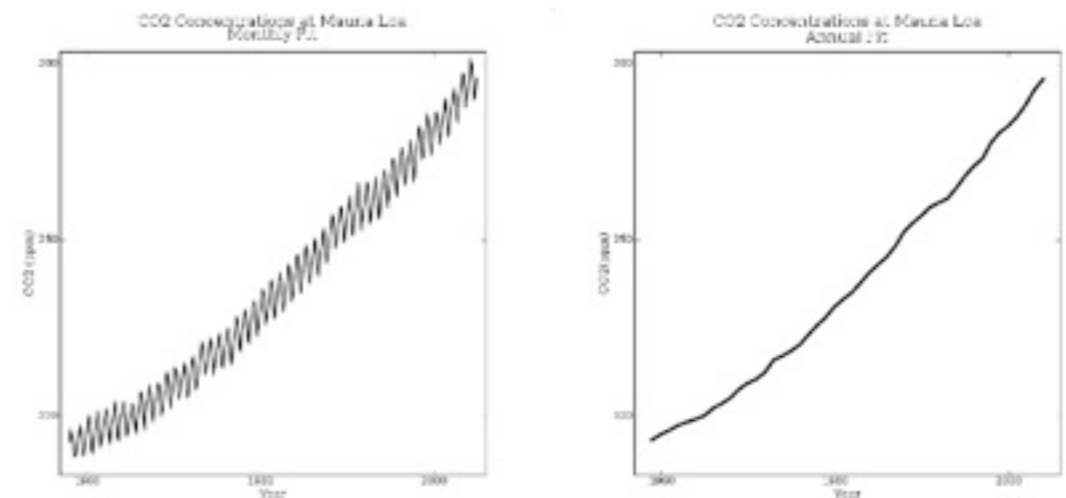
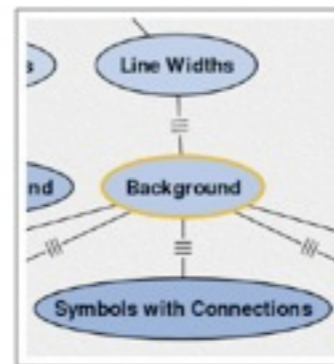


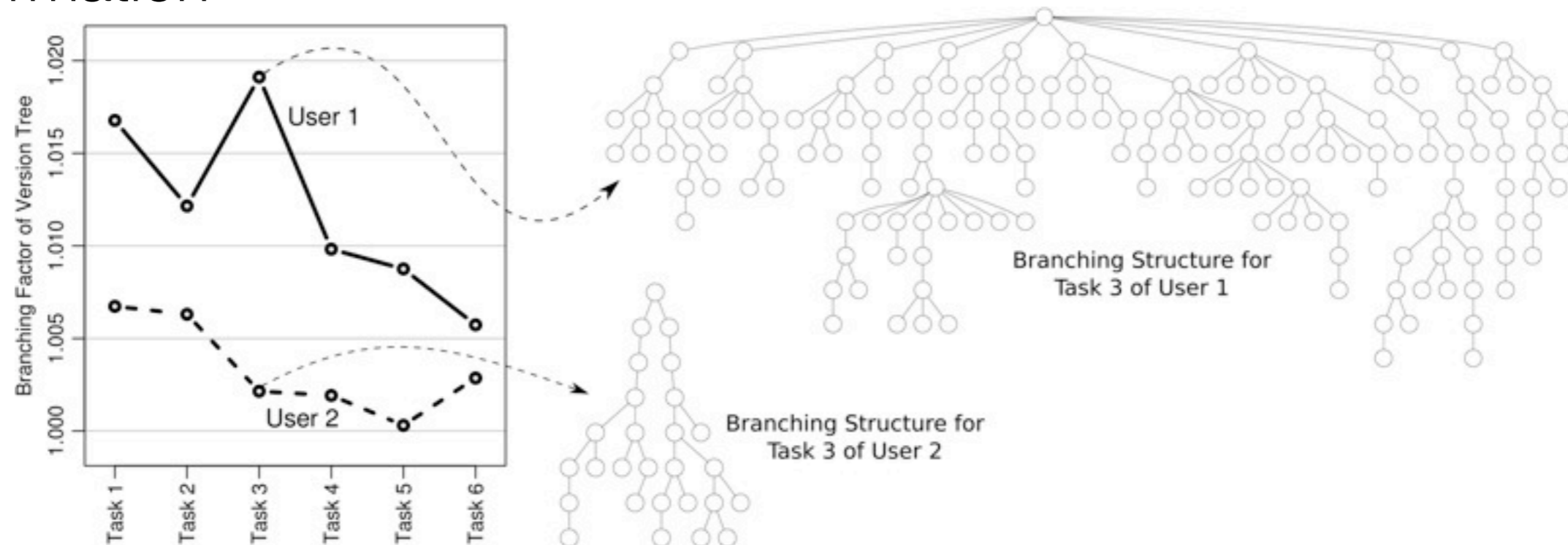
Figure 5.2: Plots of the Mauna Loa data set showing monthly measurements (left) with the yearly trend (right) using the principles for improving vision. The plot on the right is the same that was shown previously in Figure 5.1.

Provenance and Teaching (2)

- Homework 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
- Very detailed (and honest!) feedback: instructors can leverage this information

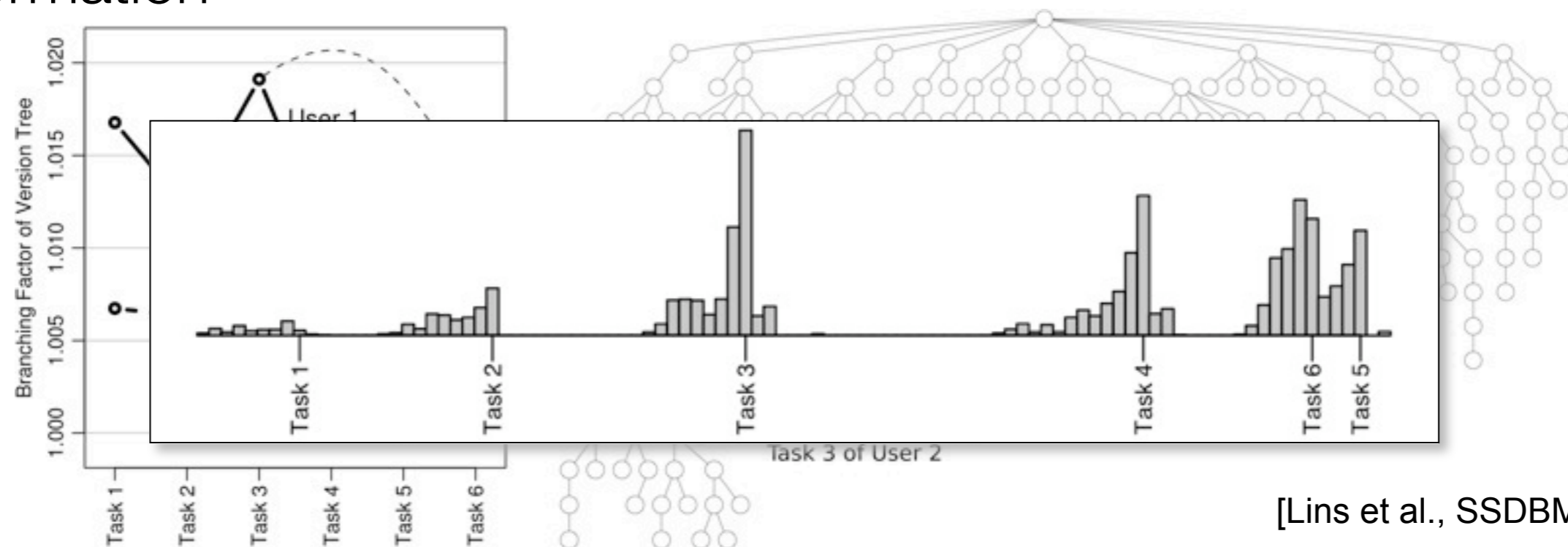
Provenance and Teaching (2)

- Homework 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
- Very detailed (and honest!) feedback: instructors can leverage this information



Provenance and Teaching (2)

- Homework 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
- Very detailed (and honest!) feedback: instructors can leverage this information

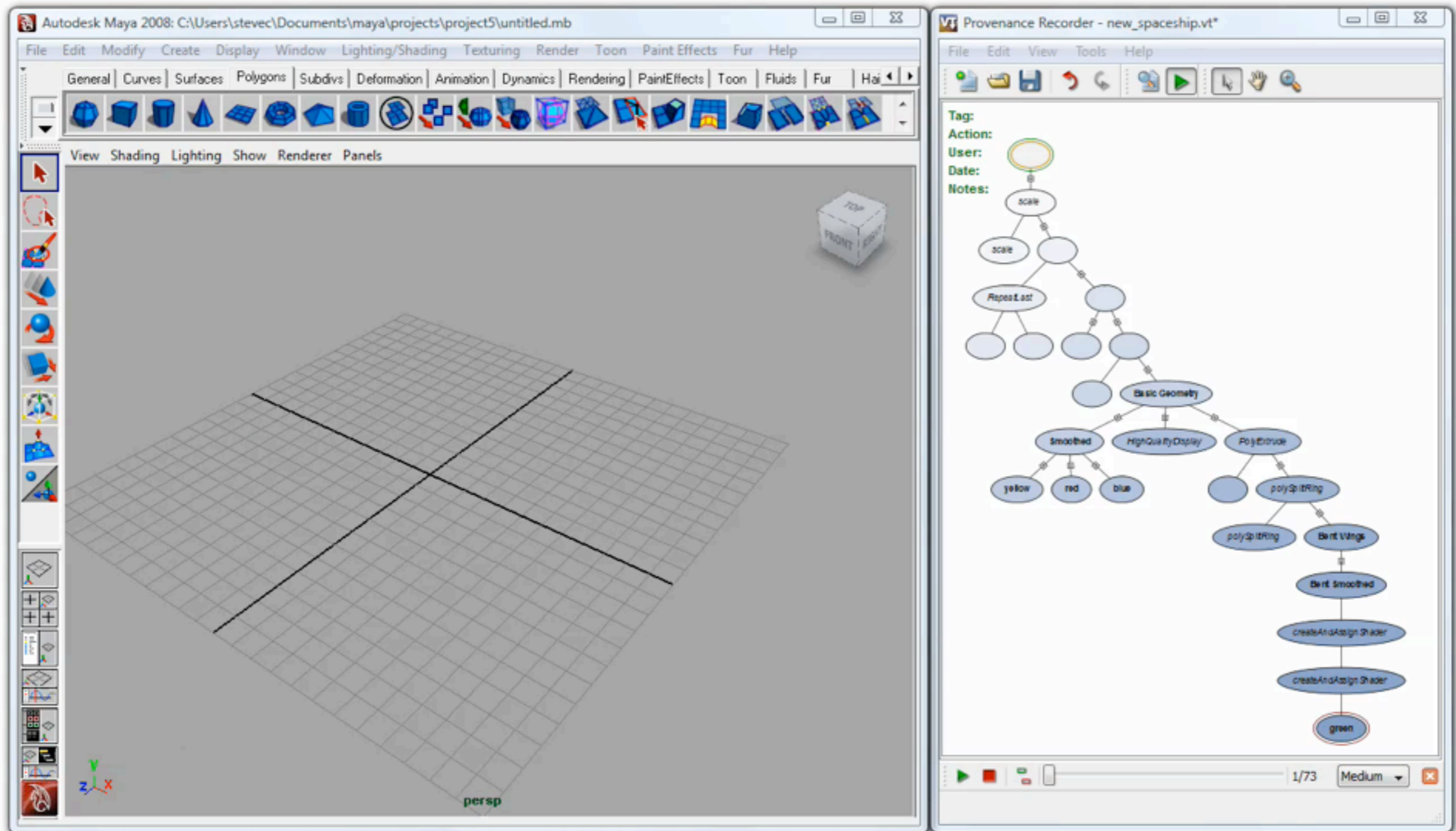


[Lins et al., SSDBM 2008]

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



Using Provenance To Teach Electronic Media



“[...] 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 Analytics: Opportunities

- Volume of collected provenance is growing
- Workflow and provenance repositories
 - myExperiments (EU), Provenance Repository (Indiana), ManyEyes (IBM), Yahoo! Pipes
- Opportunity for knowledge discovery, sharing and re-use
 - Discover workflow patterns -> a recommendation system that suggests alternatives to users as they construct a workflow
 - Discover workflow refinement patterns -> automatically extract analogies from shared repositories
 - Cluster (organize) workflow collections -> simplify query and search over repositories
 - Infer workflow specification from execution log [Aalst et al., TKDE 2004]

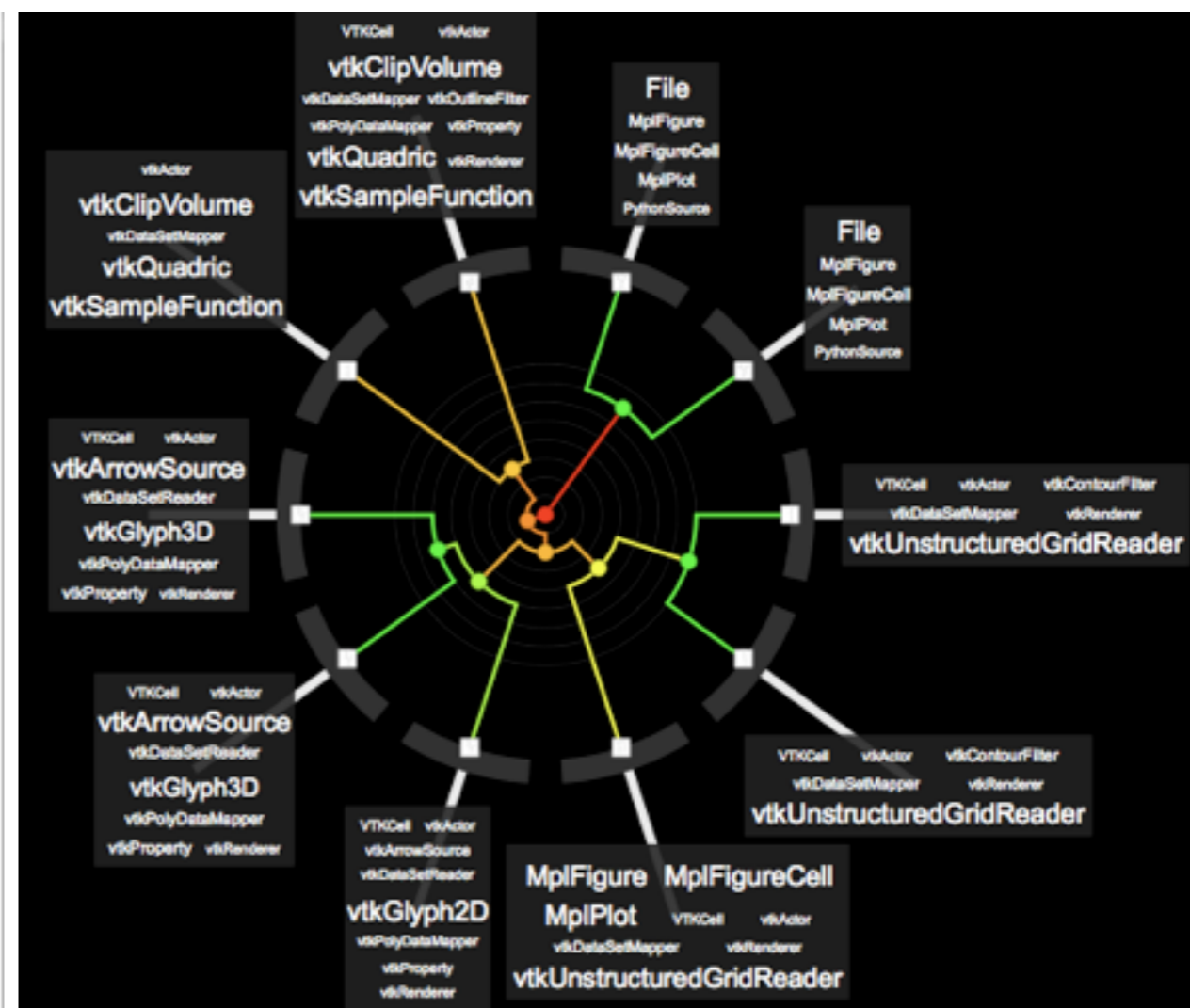
Search Engine: Clustering Workflows

Workflow query:

Snippet type: Snippet size: Result size: force min k

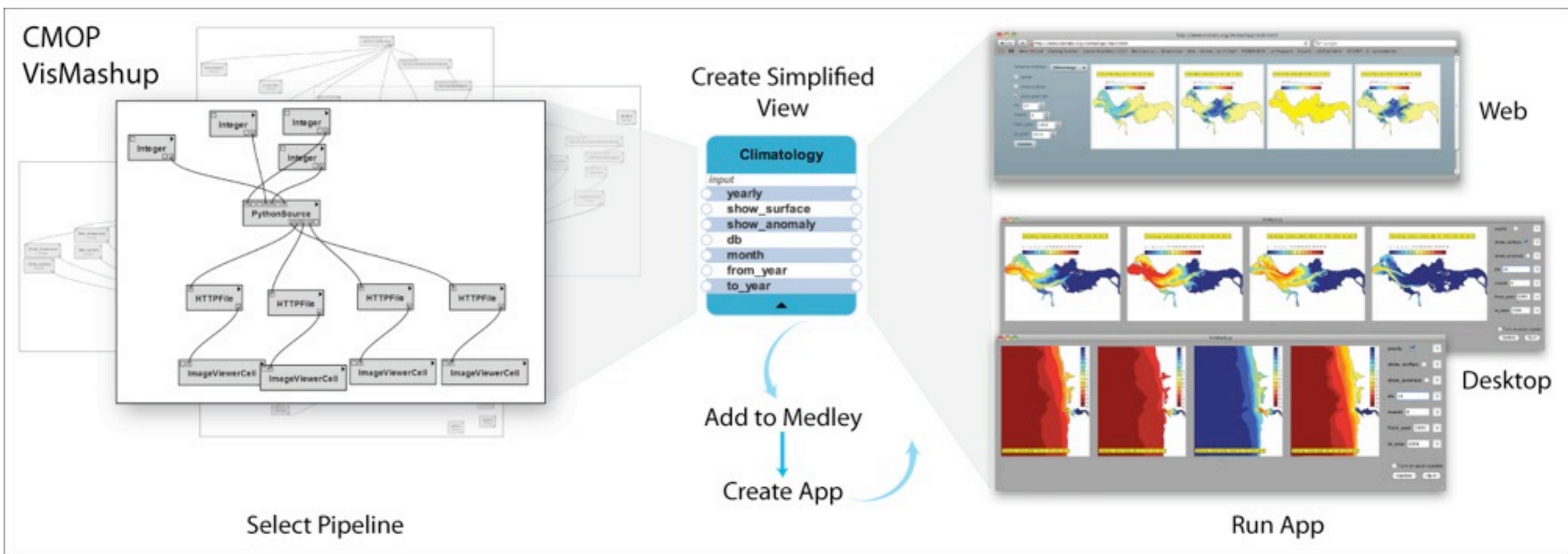
68 result(s) found for "mindi":

1. [Problem1](#) **mindi** 10 Sep 2007 15:40:45
2. [Problem2](#) **mindi** 10 Sep 2007 15:42:50
3. [Problem3](#) **mindi** 10 Sep 2007 17:12:41
4. [Problem4](#) **mindi** 13 Sep 2007 18:58:33
5. [Problem1b](#) **mindi** 30 Sep 2007 12:26:54
6. [Problem 3a](#) **mindi** 24 Oct 2007 19:36:56
7. [Problem1a](#) **mindi** 05 Dec 2007 16:22:56
8. [Problem1 finalvector](#) **mindi** 06 Dec 2007 21:17:41
9. [Problem3](#) **mindi** 13 Sep 2007 12:45:27
10. [Problem1b legend](#) **mindi** 02 Oct 2007 10:19:31
"Add legend to make the data easier to understand."



Building Customized Applications with VisMashup

- Portals increasingly using workflows in the backend,
- but with canned, one-of-a-kind interfaces--very little flexibility
- Use workflows to build mashups and customized applications



[Santos et al., IEEE Vis 2009 (to appear)]

Conclusions and Future Work

- Provenance management is essential for computational science
- Leveraging provenance, VisTrails
 - Supports reflective reasoning
 - Provides intuitive interfaces for simplifying the construction and refinement of workflows
- Sharing provenance creates new opportunities [Freire and Silva, CHI SDA, 2008]
 - Workflow/provenance repositories; provenance-enabled publications
 - Expose users to different techniques and tools
 - Users can learn by example; expedite their training; and potentially reduce their time to insight
- Provenance + Workflows + Sharing have the potential to revolutionize science!

Acknowledgments

Thanks to 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.



VACET



Thanks to Juliana for providing the slides