SFB-TRR 161
Quantitative Methods for Visual Computing
Transregional Collaborative Research Center

Provenance: On and Behind the Screens

Melanie Herschel, Marcel Hlawatsch

ACM SIGMOD/PODS 2016 | San Francisco, CA, USA | 2016-07-01

Universität Stuttgart
Universität Konstanz
Max-Planck-Institut für biologische Kybernetik

DFG

Agenda

- Part 1: Provenance
  - Overview
  - Workflow provenance
  - Data provenance
- Part 2: Visualization
  - Visualization Basics
  - Provenance Visualization 1 – Modules and events
  - Provenance Visualization 2 – Graph structure
- Open Research Questions

What is Provenance?

- Who is responsible for this production step?
- At what time did this step start?
- What was the input / output of a step?

Provenance

=

meta-data describing the **production process** of some **end product**

- Production process and supply chain of food.
- Experimental setup and chemical reactions leading to chemical compound.
- Data transformation and analysis underlying a business report.

- Food
- A chemical compound
- A business report

Provenance Applications
Supply Chains

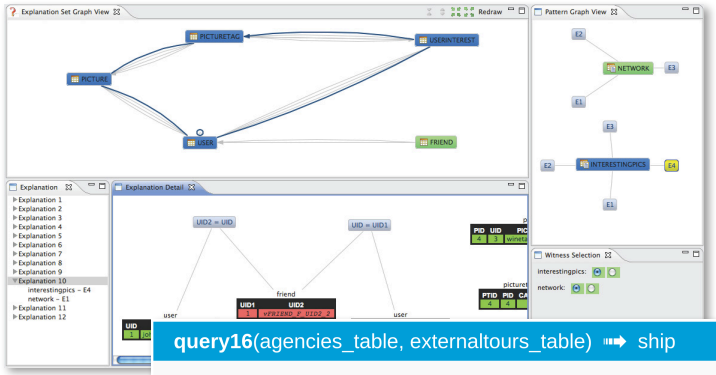
- **Trust** and **traceability**
  - *"Provenance and trust in food buying is past the point of no return"* [Art1]
  - *"End consumers may be concerned with the authenticity and the ethics of the products they buy, but companies also seek reassurance about the goods they procure."* [Art2]

Provenance Applications
Scientific Experiments

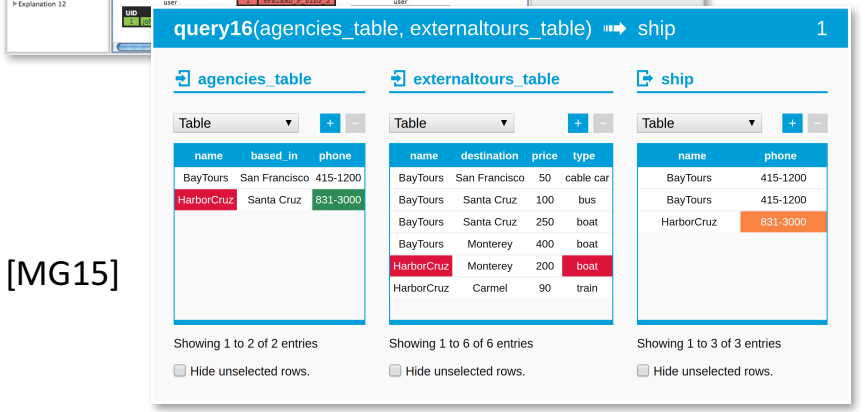
- **Preservation** and **repeatability**
  - *"The ultimate goal of the analysis preservation platform is to be able to reproduce an analysis even many years after its initial publication, permitting to extend impact of preserved analyses through validation and recasting services." [Art3]*

Provenance Applications
Complex Data Processing

- **Analysis** and **debugging**
  - Of Big Data pipelines
  - Of workflows
  - Of declarative queries

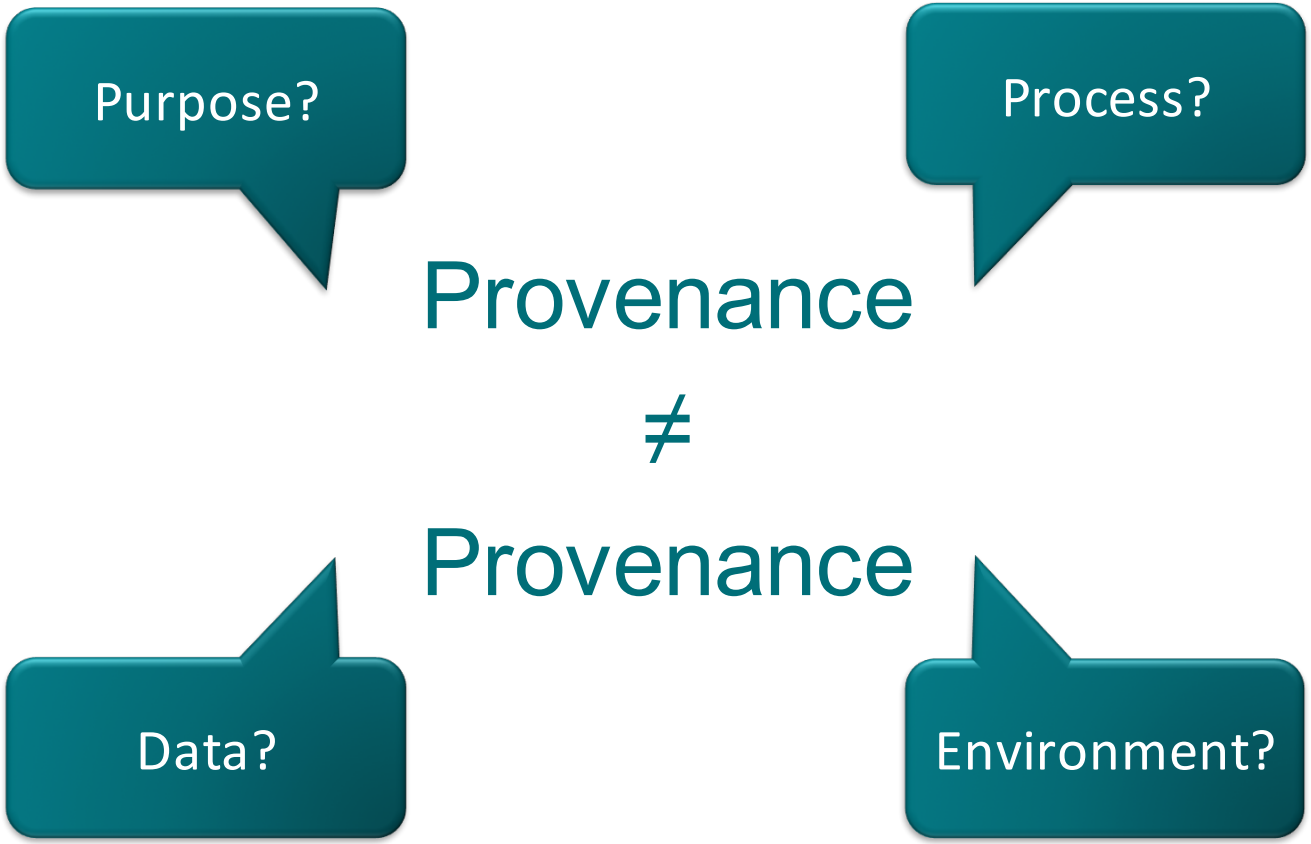


[Nautilus]

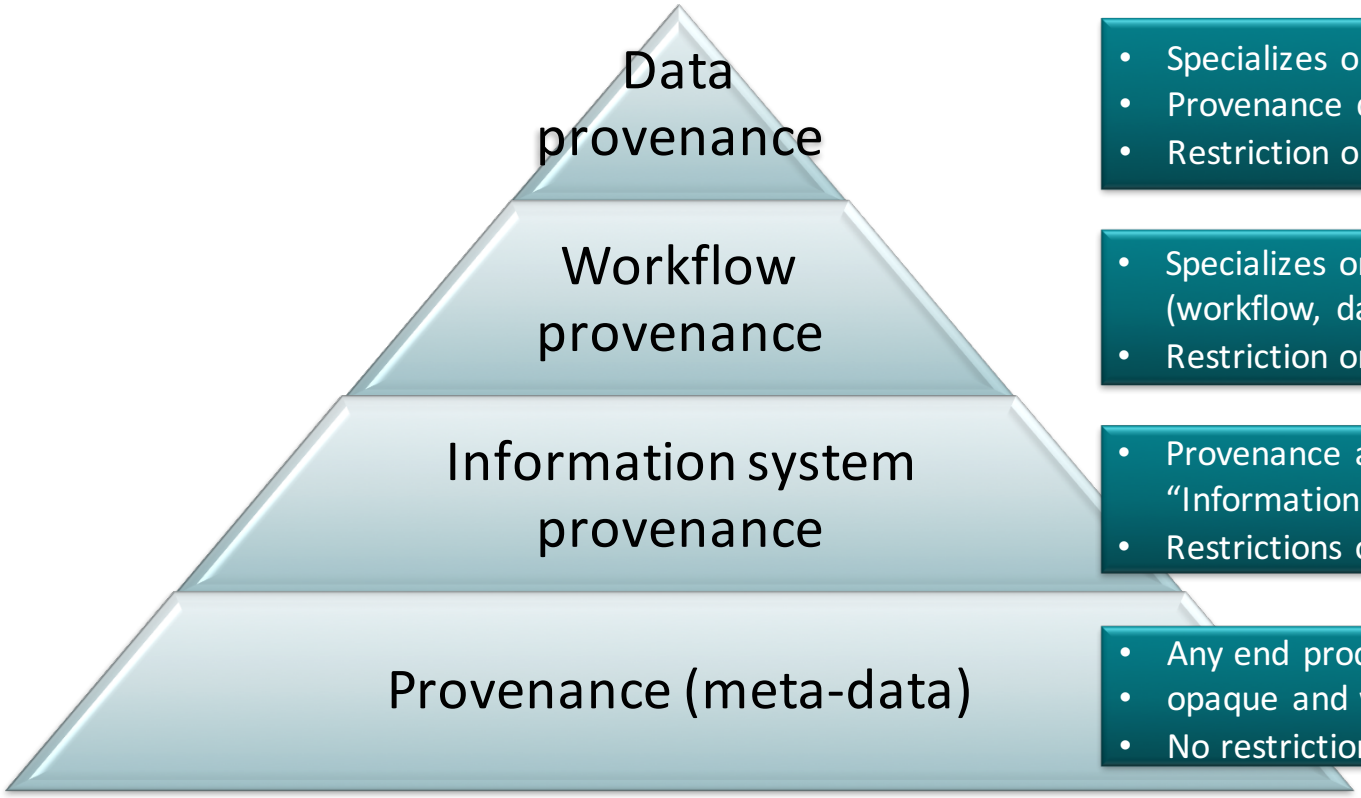


[MG15]

Provenance Types



Provenance Types



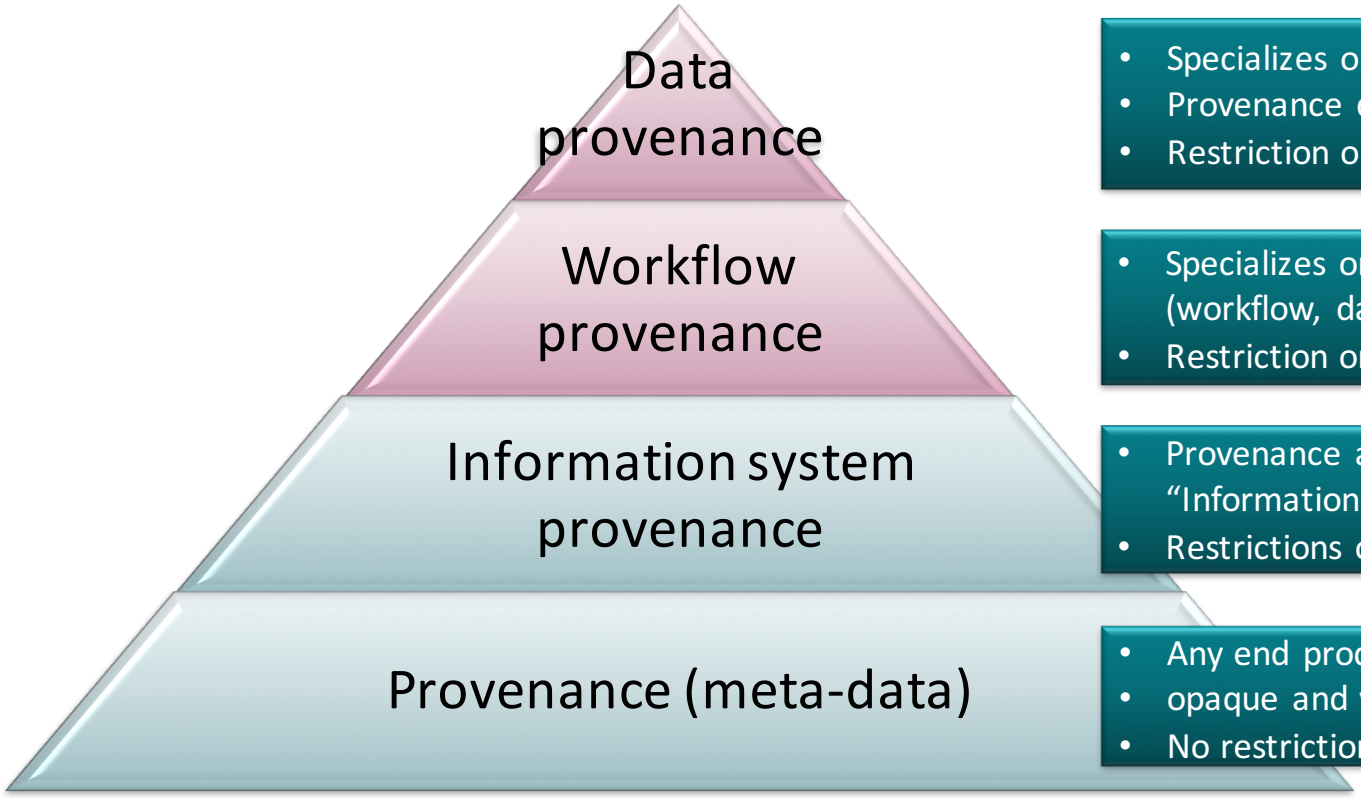
- Specializes on one structured query languages
- Provenance data model specific to query language
- Restriction on process and data

- Specializes on one particular type of process (workflow, dataflow)
- Restriction on process

- Provenance about “anything” pertaining to “Information Systems” in a very large sense.
- Restrictions on environment

- Any end product, any (proprietary) solution
- opaque and wide variety of environments
- No restrictions

Provenance Types



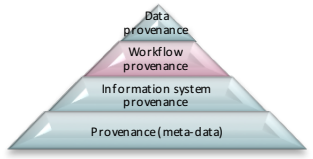
- Specializes on one structured query languages
- Provenance data model specific to query language
- Restriction on process and data

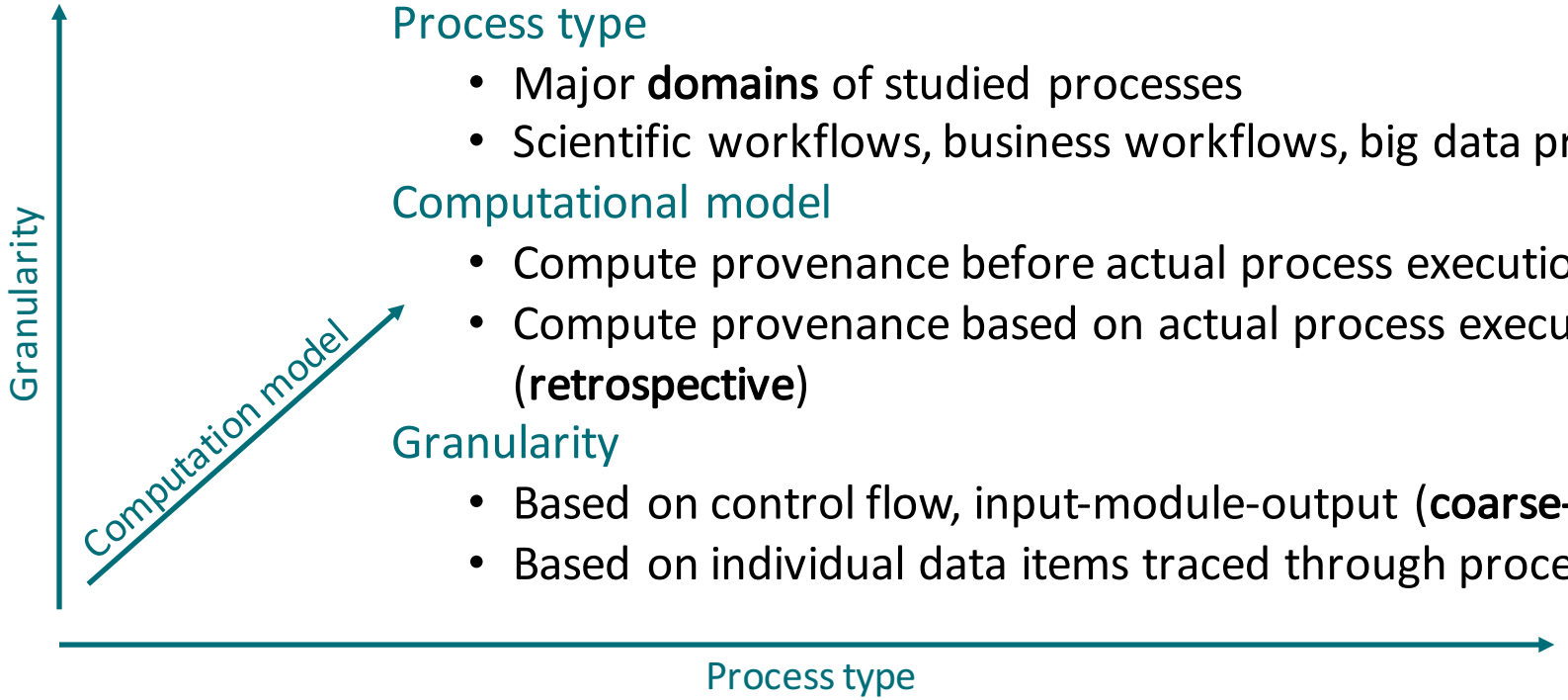
- Specializes on one particular type of process (workflow, dataflow)
- Restriction on process

- Provenance about “anything” pertaining to “Information Systems” in a very large sense.
- Restrictions on environment

- Any end product, any (proprietary) solution
- opaque and wide variety of environments
- No restrictions

Workflow Provenance





Process type

- Major **domains** of studied processes
- Scientific workflows, business workflows, big data processing

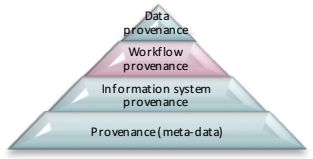
Computational model

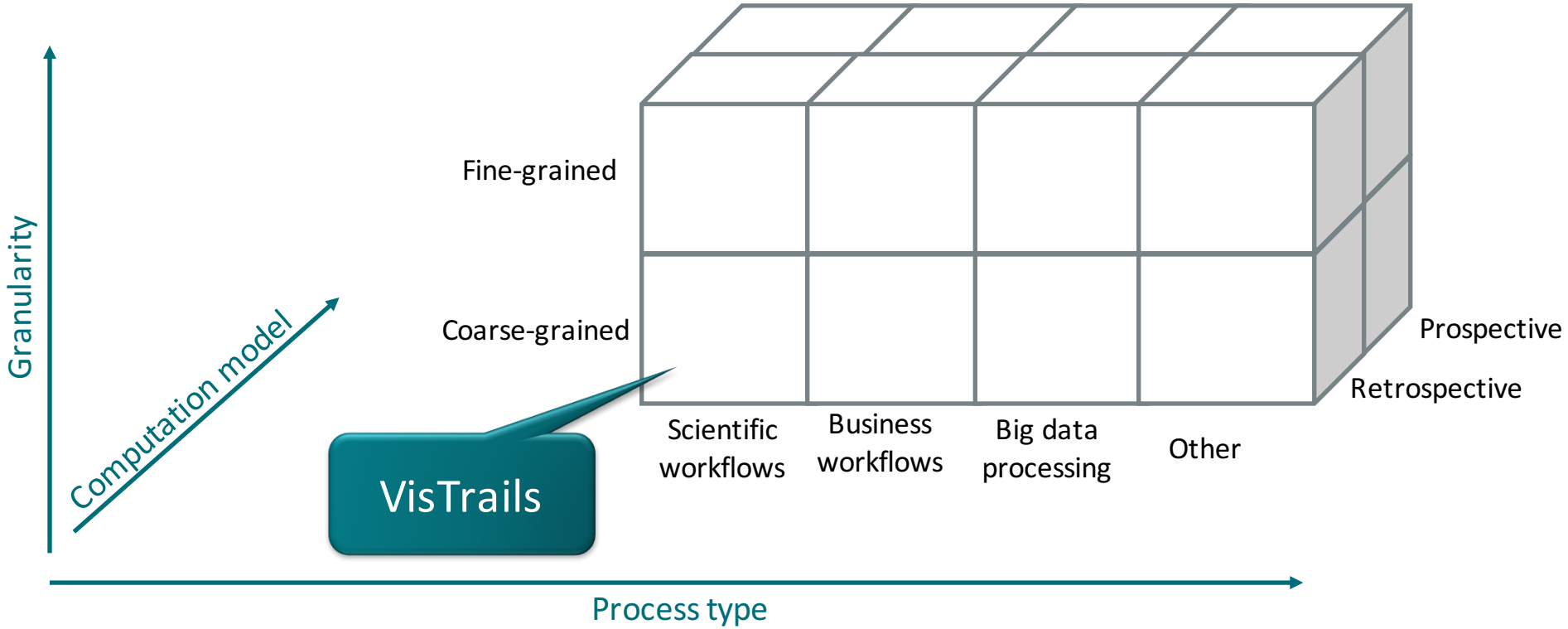
- Compute provenance before actual process execution (**prospective**)
- Compute provenance based on actual process execution (**retrospective**)

Granularity

- Based on control flow, input-module-output (**coarse-grained**)
- Based on individual data items traced through process (**fine-grained**)

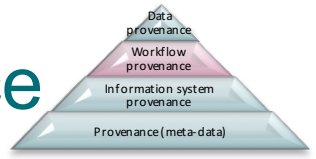
Workflow Provenance





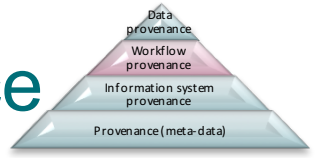
For a (partial) survey, see [DCL+07]

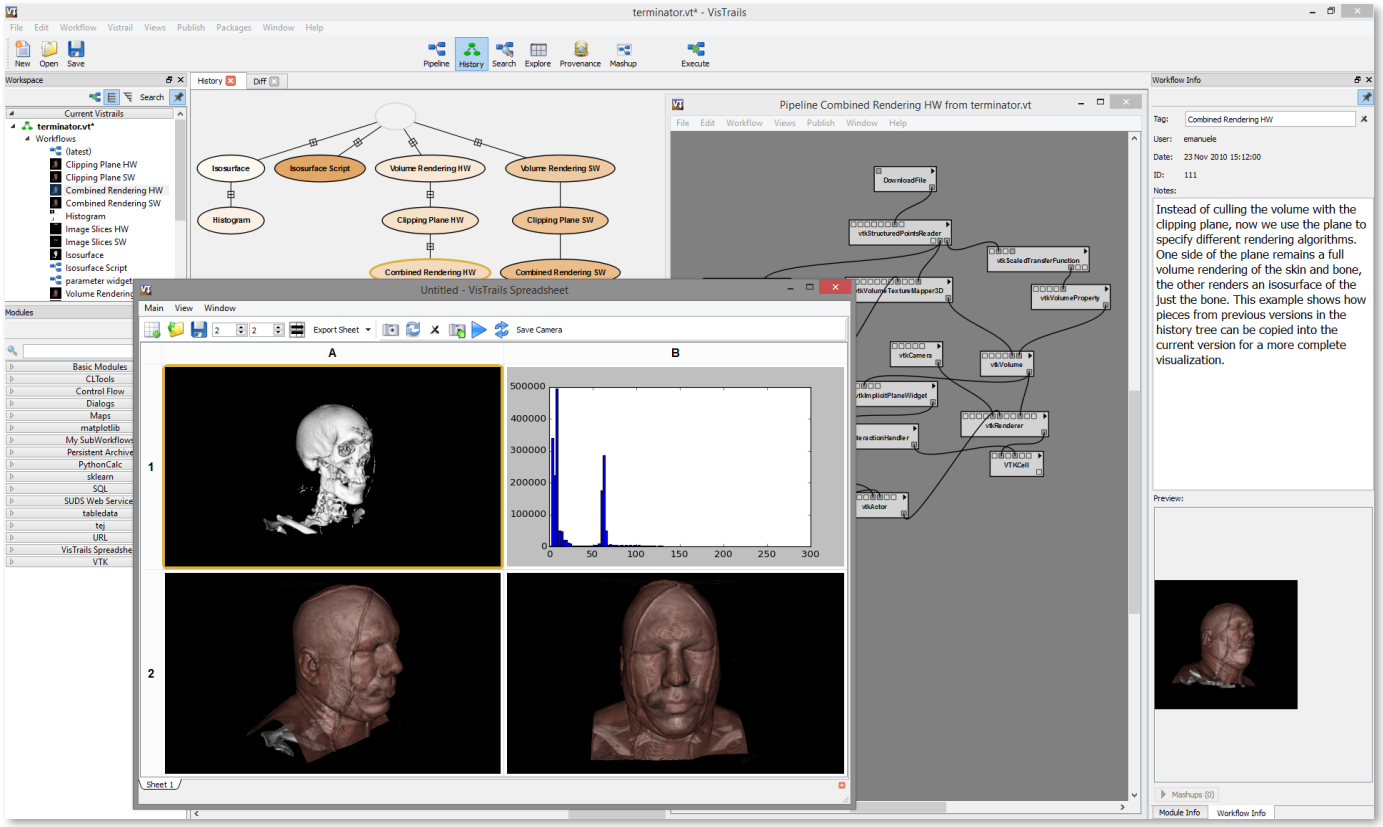
Coarse-grained Scientific Workflow Provenance VisTrails [FTC+06]



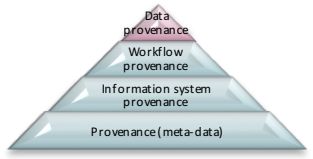
- Initially developed to manage the design of workflows producing visualizations.
- Provenance tracks evolution
  - Records all steps performed by visualization expert while designing an image production process
  - Provenance model includes actions applied to the workflow
- Typical actions
  - Add/ replace / delete a module
  - Add a connection between modules
  - Set parameter values
- Tree data model
  - Each node corresponds to a version of a workflow
  - An edge between parent node P and child node C corresponds to one or more actions applied to P to obtain C.

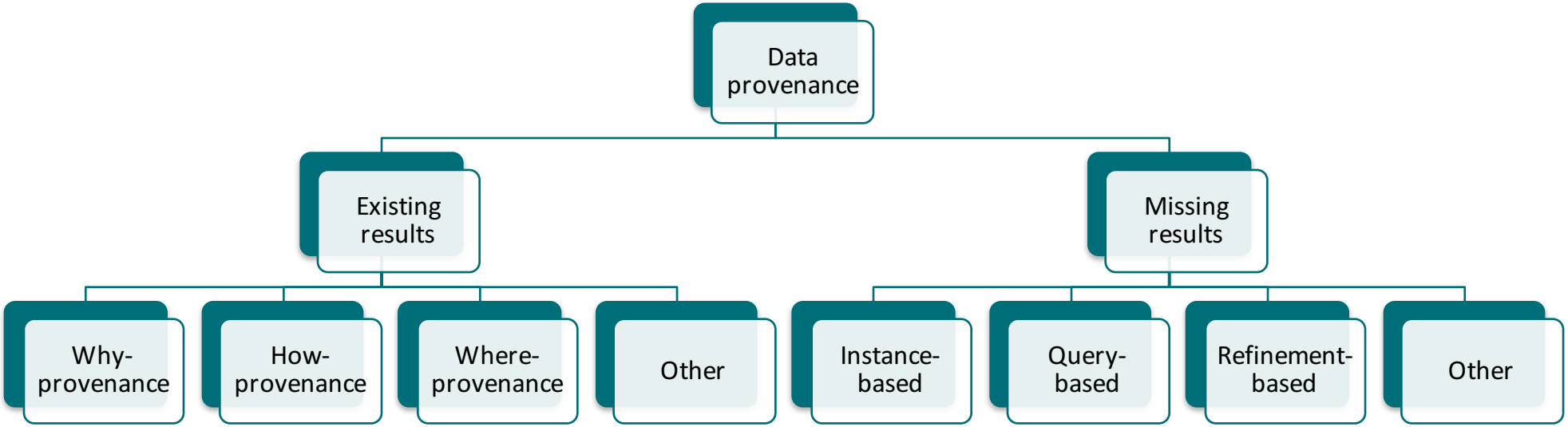
Coarse-grained Scientific Workflow Provenance



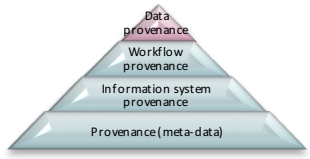


Data Provenance



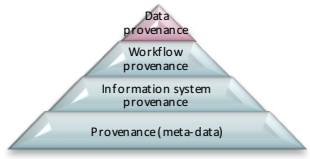


Existing Results



- Given:
  - Database D
  - Query Q
  - Q(D): result of executing Q over D
- Data provenance of existing data = origin of a tuple t in Q(D)?
  - Why-provenance
    What data in D contributes to t?
  - How-provenance
    How are tuples from D combined to produce t?
  - Where-provenance
    Which source values in D were copied to produce t?

Existing Results



Flights

	Number	Origin	Destination
t1	AF1234	Nice	Frankfurt
t2	BA5678	London	Paris

Trains

	Number	Origin	Destination
t3	TGV1	Paris	Stuttgart
t4	TGV2	London	Paris
t5	DB1	Frankfurt	Berlin
t6	DB2	Munich	Stuttgart

Hotels

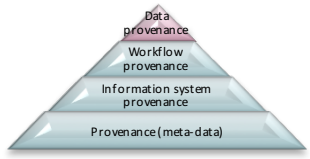
	Hotel	city
t7	Hilton	Paris
t8	Concorde	Paris
t9	Holiday Inn	Frankfurt
t10	Hyatt	London

Query Q:

```
SELECT hotel, destination
FROM  (
        SELECT destination FROM Flights
        UNION SELECT destination FROM Trains
) AS Transports T, Hotels H
WHERE T.destination= hotels.city
```

hotel	destination
Hilton	Paris
Concorde	Paris
Holiday Inn	Frankfurt

Existing Results



Flights

	Number	Origin	Destination
t1	AF1234	Nice	Frankfurt
t2	BA5678	London	Paris

Trains

	Number	Origin	Destination
t3	TGV1	Paris	Stuttgart
t4	TGV2	London	Paris
t5	DB1	Frankfurt	Berlin
t6	DB2	Munich	Stuttgart

Hotels

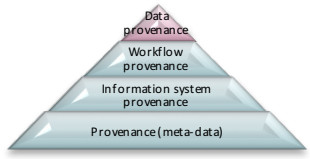
	Hotel	city
t7	Hilton	Paris
t8	Concorde	Paris
t9	Holiday Inn	Frankfurt
t10	Hyatt	London

Query Q:

```
SELECT hotel, destination
FROM  (
        SELECT destination FROM Flights
        UNION SELECT destination FROM Trains
) AS Transports T, Hotels H
WHERE T.destination= hotels.city
```

hotel	destination
Hilton	Paris
Concorde	Paris
Holiday Inn	Frankfurt

Existing Results – Why-Provenance



Flights

	Number	Origin	Destination
t1	AF1234	Nice	Frankfurt
t2	BA5678	London	Paris

Trains

	Number	Origin	Destination
t3	TGV1	Paris	Stuttgart
t4	TGV2	London	Paris
t5	DB1	Frankfurt	Berlin
t6	DB2	Munich	Stuttgart

Hotels

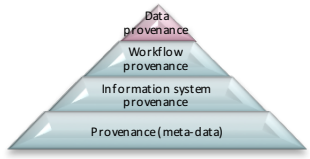
	Hotel	city
t7	Hilton	Paris
t8	Concorde	Paris
t9	Holiday Inn	Frankfurt
t10	Hyatt	London

Query Q:

```
SELECT hotel, destination
FROM  (
        SELECT destination FROM Flights
        UNION SELECT destination FROM Trains
) AS Transports T, Hotels H
WHERE T.destination= hotels.city
```

hotel	destination
Hilton	Paris
Concorde	Paris
Holiday Inn	Frankfurt

Existing Results – How-Provenance



Flights

	Number	Origin	Destination
t1	AF1234	Nice	Frankfurt
t2	BA5678	London	Paris

Trains

	Number	Origin	Destination
t3	TGV1	Paris	Stuttgart
t4	TGV2	London	Paris
t5	DB1	Frankfurt	Berlin
t6	DB2	Munich	Stuttgart

Hotels

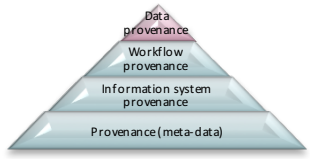
	Hotel	city
t7	Hilton	Paris
t8	Concorde	Paris
t9	Holiday Inn	Frankfurt
t10	Hyatt	London

Query Q:

```
SELECT hotel, destination
FROM  (
        SELECT destination FROM Flights
        UNION SELECT destination FROM Trains
) AS Transports T, Hotels H
WHERE T.destination= hotels.city
```

	hotel	destination
t2*t7 + t4*t7	Hilton	Paris
t2*t8 + t4*t7	Concorde	Paris
t1 * t9	Holiday Inn	Frankfurt

Existing Results – Where-Provenance



Flights

	Number	Origin	Destination
t1	AF1234	Nice	Frankfurt
t2	BA5678	London	Paris

Trains

	Number	Origin	Destination
t3	TGV1	Paris	Stuttgart
t4	TGV2	London	Paris
t5	DB1	Frankfurt	Berlin
t6	DB2	Munich	Stuttgart

Hotels

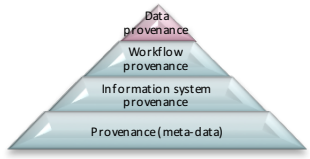
	Hotel	city
t7	Hilton	Paris
t8	Concorde	Paris
t9	Holiday Inn	Frankfurt
t10	Hyatt	London

Query Q:

```
SELECT hotel, destination
FROM  (
        SELECT destination FROM Flights
        UNION SELECT destination FROM Trains
) AS Transports T, Hotels H
WHERE T.destination= hotels.city
```

hotel	destination
Hilton	Paris
Concorde	Paris
Holiday Inn	Frankfurt

Missing Results



Flights

	Number	Origin	Destination
t1	AF1234	Nice	Frankfurt
t2	BA5678	London	Paris

Trains

	Number	Origin	Destination
t3	TGV1	Paris	Stuttgart
t4	TGV2	London	Paris
t5	DB1	Frankfurt	Berlin
t6	DB2	Munich	Stuttgart

Hotels

	Hotel	city
t7	Hilton	Paris
t8	Concorde	Paris
t9	Holiday Inn	Frankfurt
t10	Hyatt	London

Query Q:

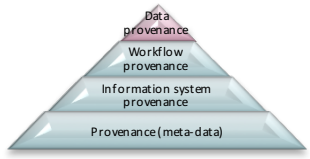
```
SELECT hotel, destination
FROM  (
        SELECT destination FROM Flights
        UNION SELECT destination FROM Trains
) AS Transports T, Hotels H
WHERE T.destination= hotels.city
```

hotel	destination
Hilton	Paris
Concorde	Paris
Holiday Inn	Frankfurt

Why not in Q(D)?

?	Stuttgart

Missing Results – Instance-Based



Flights

	Number	Origin	Destination
t1	AF1234	Nice	Frankfurt
t2	BA5678	London	Paris

Trains

	Number	Origin	Destination
t3	TGV1	Paris	Stuttgart
t4	TGV2	London	Paris
t5	DB1	Frankfurt	Berlin
t6	DB2	Munich	Stuttgart

Hotels

	Hotel	city
t7	Hilton	Paris
t8	Concorde	Paris
t9	Holiday Inn	Frankfurt
t10	Hyatt	London
	?	Stuttgart

Query Q:

```
SELECT hotel, destination
FROM  (
        SELECT destination FROM Flights
        UNION SELECT destination FROM Trains
) AS Transports T, Hotels H
WHERE T.destination= hotels.city
```

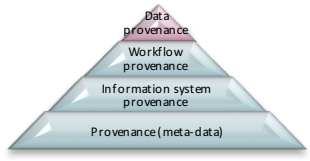
hotel	destination
Hilton	Paris
Concorde	Paris
Holiday Inn	Frankfurt

Why not in Q(D)?

?	Stuttgart

No hotels in Stuttgart (instance-based)

Missing Results – Query-Based



Join is too selective!

Flights

	Number	Origin	Destination
t1	AF1234	Nice	Frankfurt
t2	BA5678	London	Paris

Trains

	Number	Origin	Destination
t3	TGV1	Paris	Stuttgart
t4	TGV2	London	Paris
t5	DB1	Frankfurt	Berlin
t6	DB2	Munich	Stuttgart

Hotels

	Hotel	city
t7	Hilton	Paris
t8	Concorde	Paris
t9	Holiday Inn	Frankfurt
t10	Hyatt	London

Query Q:

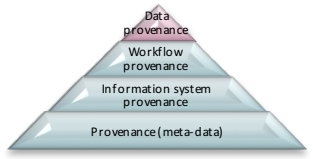
```
SELECT hotel, destination
FROM  (
        SELECT destination FROM Flights
        UNION SELECT destination FROM Trains
) AS Transports T, Hotels H
WHERE T.destination= hotels.city
```

hotel	destination
Hilton	Paris
Concorde	Paris
Holiday Inn	Frankfurt

Why not in Q(D)?

?	Stuttgart

Missing Results – Modification-Based



Flights

	Number	Origin	Destination
t1	AF1234	Nice	Frankfurt
t2	BA5678	London	Paris

Trains

	Number	Origin	Destination
t3	TGV1	Paris	Stuttgart
t4	TGV2	London	Paris
t5	DB1	Frankfurt	Berlin
t6	DB2	Munich	Stuttgart

Hotels

	Hotel	city
t7	Hilton	Paris
t8	Concorde	Paris
t9	Holiday Inn	Frankfurt
t10	Hyatt	London

Query is buggy and we fix it!

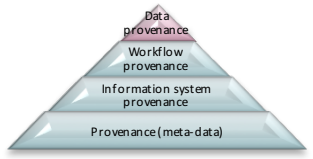
Query Q:

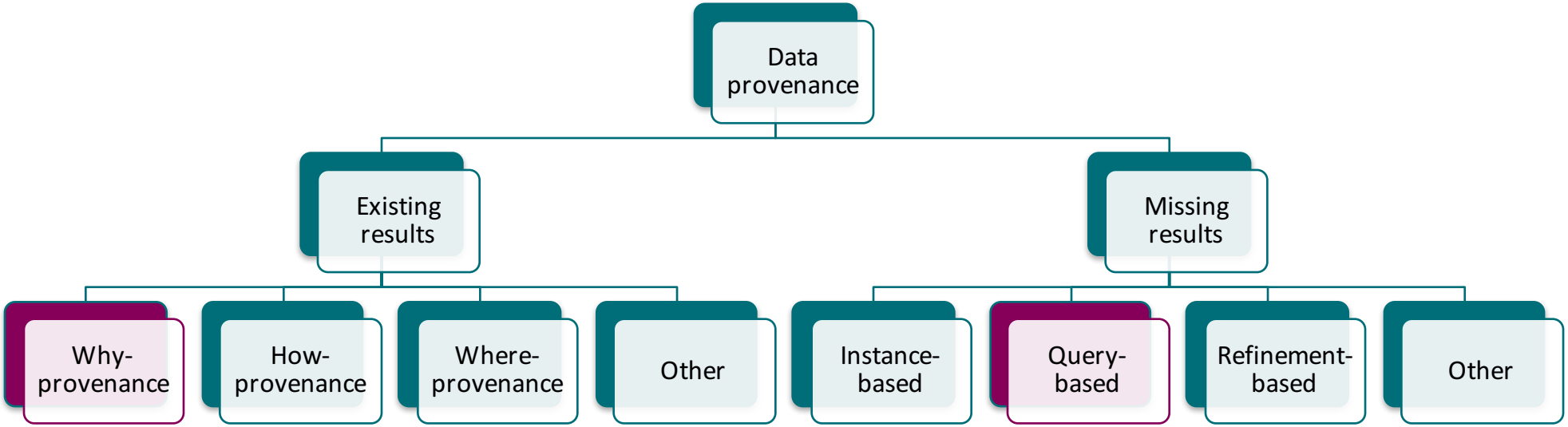
```
SELECT hotel, destination
FROM  (
        SELECT destination FROM Flights
        UNION SELECT destination FROM Trains
) AS Transports T
        RIGHT OUTER JOIN  Hotels H
        ON T.destination= hotels.city
```

hotel	destination
Hilton	Paris
Concorde	Paris
Holiday Inn	Frankfurt
?	Stuttgart

Why not in Q(D)?

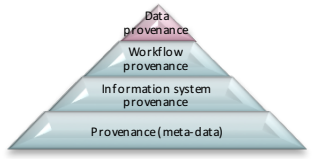
Data Provenance

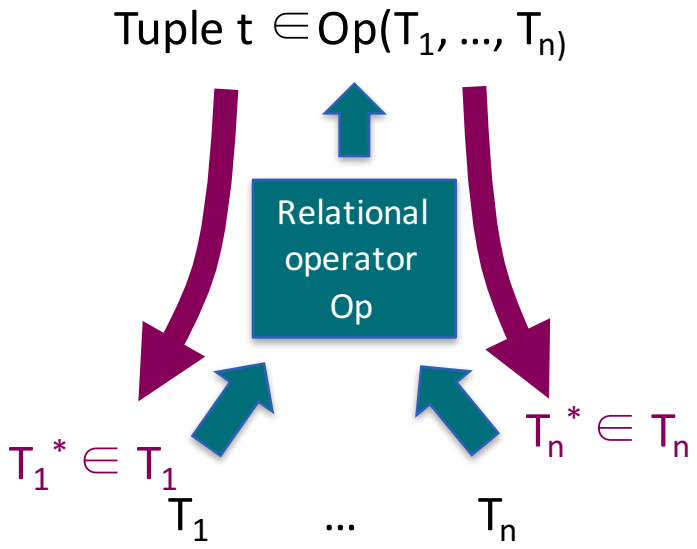




Extends classifications of [CCT09]

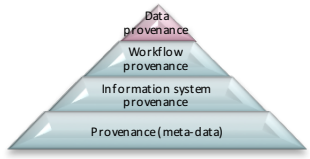
Why-Provenance for Existing Results Lineage Tracing [CW00]





Find necessary and sufficient **predecessors** of t

Why-Provenance for Existing Results



bid	title	price
b1	Odyssey	15
b2	Iliad	45
b3	Antigone	49
b4	Medea	45

Book B

aid	bid
a1	b2
a1	b1
a2	b3
a3	b4

AuthorBook AB

$\bowtie_{bid}$

$\bowtie_{aid}$

$\pi_{A.name,\ p.price}$

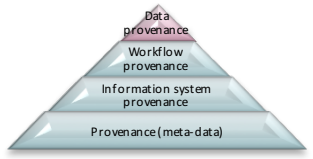
name	price
Sophocles	49
Euripides	45

Author A

$\sigma_{A.date > 800BC}$

aid	name	date
a1	Homer	800BC
a2	Sophocles	400BC
a3	Euripides	400BC

Why-Provenance for Existing Results



Book B

bid	title	price
b1	Odyssey	15
b2	Iliad	45
b3	Antigone	49
b4	Medea	45

AuthorBook AB

aid	bid
a1	b2
a1	b1
a2	b3
a3	b4

Author A

aid	name	date
a1	Homer	800BC
a2	Sophocles	400BC
a3	Euripides	400BC

$\sigma_{A.date > 800BC}$

aid	name	date
a2	Sophocles	400BC
a3	Euripides	400BC

$\bowtie_{bid}$

bid	title	price	aid
b1	Odyssey	15	a1
b2	Iliad	45	a1
b3	Antigone	49	a2
b4	Medea	45	a3

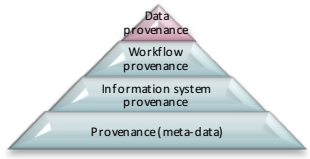
$\bowtie_{aid}$

bid	title	price	aid	name	date
b3	Antigone	49	a2	Sophocles	400BC
b4	Medea	45	a3	Euripides	400BC

$\pi_{A.name,\ p.price}$

name	price
Sophocles	49
Euripides	45

Why-Provenance for Existing Results



bid	title	price
b1	Odyssey	15
b2	Iliad	45
b3	Antigone	49
b4	Medea	45

Book B

aid	bid
a1	b2
a1	b1
a2	b3
a3	b4

AuthorBook AB

$\bowtie_{bid}$

bid	title	price	aid
b1	Odyssey	15	a1
b2	Iliad	45	a1
b3	Antigone	49	a2
b4	Medea	45	a3

$\bowtie_{aid}$

$\pi_{A.name,\ p.price}$

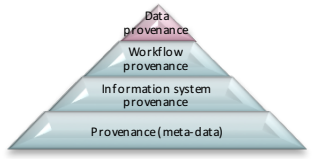
Author A

$\sigma_{A.date > 800BC}$

aid	name	date
a1	Homer	800BC
a2	Sophocles	400BC
a3	Euripides	400BC

aid	name	date
a2	Sophocles	400BC
a3	Euripides	400BC

Why-Provenance for Existing Results



bid	title	price
b1	Odyssey	15
b2	Iliad	45
b3	Antigone	49
b4	Medea	45

Book B

aid	bid
a1	b2
a1	b1
a2	b3
a3	b4

AuthorBook AB

$\bowtie_{bid}$

$\bowtie_{aid}$

$\pi_{A.name,\ p.price}$

Author A

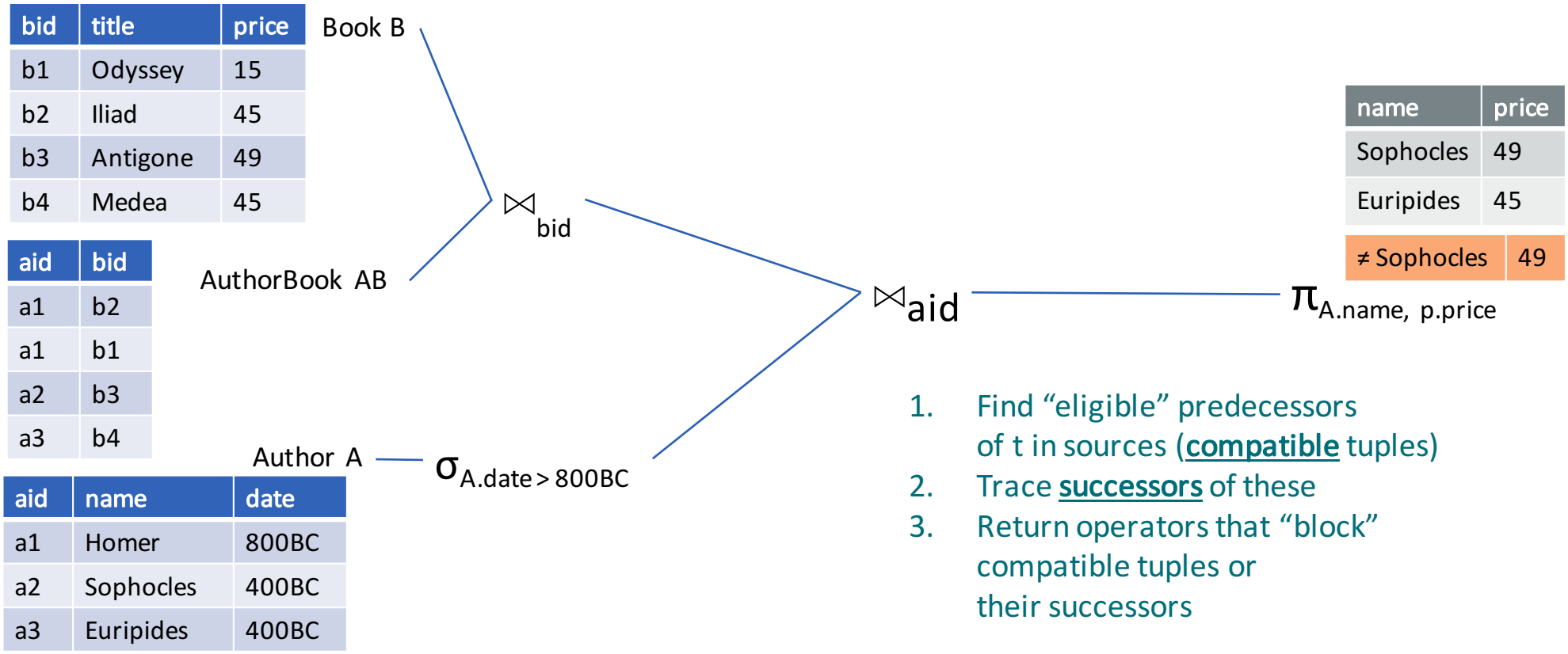
$\sigma_{A.date > 800BC}$

aid	name	date
a1	Homer	800BC
a2	Sophocles	400BC
a3	Euripides	400BC

aid	name	date
a2	Sophocles	400BC
a3	Euripides	400BC

Why-provenance of (Sophocles, 49)
= B(b3, Antigone, 49), AB(a2, b3), A(a2, Sophocles, 400BC)

Query-based Provenance for Missing Results [BHT14]



bid	title	price
b1	Odyssey	15
b2	Iliad	45
b3	Antigone	49
b4	Medea	45

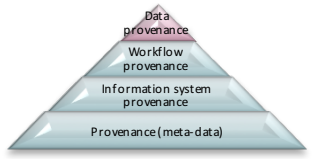
aid	bid
a1	b2
a1	b1
a2	b3
a3	b4

aid	name	date
a1	Homer	800BC
a2	Sophocles	400BC
a3	Euripides	400BC

name	price
Sophocles	49
Euripides	45
≠ Sophocles	49

1. Find "eligible" predecessors of t in sources (**compatible** tuples)
2. Trace **successors** of these
3. Return operators that "block" compatible tuples or their successors

Query-based Provenance for Missing Results



bid	title	price
b1	Odyssey	15
b2	Iliad	45
b3	Antigone	49
b4	Medea	45

Book B

aid	bid
a1	b2
a1	b1
a2	b3
a3	b4

AuthorBook AB

$\bowtie_{bid}$

Author A

aid	name	date
a1	Homer	800BC
a2	Sophocles	400BC
a3	Euripides	400BC

$\sigma_{A.date > 800BC}$

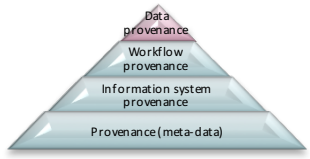
$\bowtie_{aid}$

$\pi_{A.name,\ p.price}$

name	price
Sophocles	49
Euripides	45
≠ Sophocles	49

1. Find "eligible" predecessors of t in sources (**compatible** tuples)
2. Trace **successors** of these
3. Return operators that "block" compatible tuples or their successors

Query-based Provenance for Missing Results



bid	title	price
b1	Odyssey	15
b2	Iliad	45
b3	Antigone	49
b4	Medea	45

Book B

aid	bid
a1	b2
a1	b1
a2	b3
a3	b4

AuthorBook AB

$\bowtie_{bid}$

$\bowtie_{aid}$

$\pi_{A.name,\ p.price}$

Author A

aid	name	date
a1	Homer	800BC
a2	Sophocles	400BC
a3	Euripides	400BC

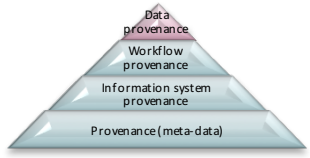
$\sigma_{A.date > 800BC}$

aid	name	date
a2	Sophocles	400BC
a3	Euripides	400BC

Picky wrt a1

1. Find "eligible" predecessors of t in sources (**compatible** tuples)
2. Trace successors of these
3. Return operators that "block" compatible tuples or their successors

Query-based Provenance for Missing Results



bid	title	price
b1	Odyssey	15
b2	Iliad	45
b3	Antigone	49
b4	Medea	45

Book B

aid	bid
a1	b2
a1	b1
a2	b3
a3	b4

AuthorBook AB

aid	name	date
a1	Homer	800BC
a2	Sophocles	400BC
a3	Euripides	400BC

Author A

$\sigma_{A.date > 800BC}$

Picky wrt a1

bid	title	price	aid
b1	Odyssey	15	a1
b2	Iliad	45	a1
b3	Antigone	49	a2
b4	Medea	45	a3

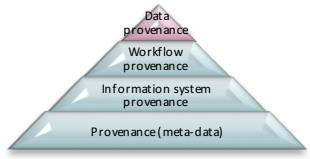
$\bowtie_{bid}$

$\bowtie_{aid}$

$\pi_{A.name,\ p.price}$

1. Find "eligible" predecessors of t in sources (**compatible** tuples)
2. Trace successors of these
3. Return operators that "block" compatible tuples or their successors

Query-based Provenance for Missing Results



bid	title	price
b1	Odyssey	15
b2	Iliad	45
b3	Antigone	49
b4	Medea	45

Book B

aid	bid
a1	b2
a1	b1
a2	b3
a3	b4

AuthorBook AB

aid	name	date
a1	Homer	800BC
a2	Sophocles	400BC
a3	Euripides	400BC

Author A

$\bowtie_{bid}$

$\sigma_{A.date > 800BC}$

Picky wrt a1

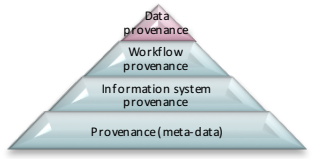
bid	title	price	aid	name	date
b3	Antigone	49	a2	Sophocles	400BC
b4	Medea	45	a3	Euripides	400BC

Picky wrt all remaining

$\pi_{A.name,\ p.price}$

1. Find "eligible" predecessors of t in sources (**compatible** tuples)
2. Trace successors of these
3. Return operators that "block" compatible tuples or their successors

Query-based Provenance for Missing Results



bid	title	price
b1	Odyssey	15
b2	Iliad	45
b3	Antigone	49
b4	Medea	45

Book B

aid	bid
a1	b2
a1	b1
a2	b3
a3	b4

AuthorBook AB

$\bowtie_{bid}$

Author A

aid	name	date
a1	Homer	800BC
a2	Sophocles	400BC
a3	Euripides	400BC

$\sigma_{A.date > 800BC}$

Picky wrt a1

Picky wrt all remaining

$\bowtie_{aid}$

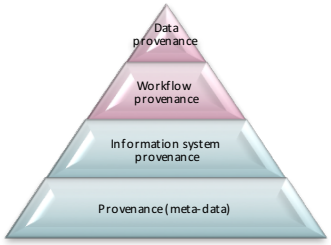
$\pi_{A.name,\ p.price}$

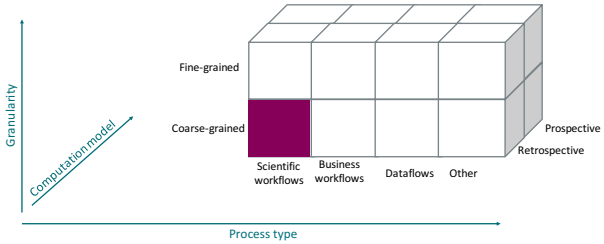
Query-based why-not provenance for (≠Sophocles, 49) :

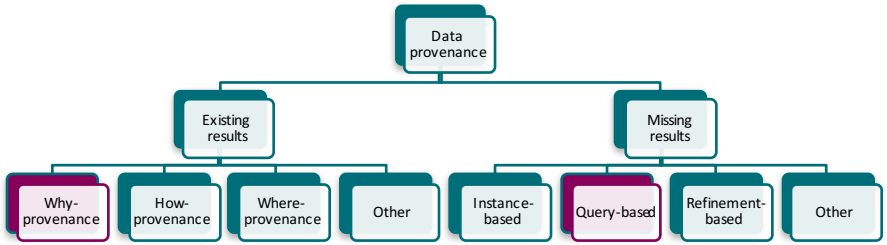
$\{\bowtie_{aid}, \sigma_{A.date > 800BC}\}$

1. Find "eligible" predecessors of t in sources (**compatible** tuples)
2. Trace **successors** of these
3. Return operators that "block" compatible tuples or their successors

Provenance and Visualization







- Many different types of provenance data
- Provenance data can easily become large
- Provenance typically used to convey some information about the process either for experts or non-experts.
- Convey information over (large) data sets using proper visualization

Agenda

- Part 1: Provenance
  - Overview
  - Workflow provenance
  - Data provenance
- Part 2: Visualization
  - Visualization Basics
  - Provenance Visualization 1 – Modules and events
  - Provenance Visualization 2 – Graph structure
- Open Research Questions

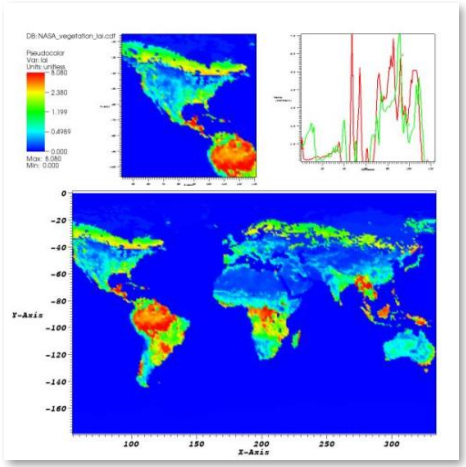
Visualization?

What is Visualization?

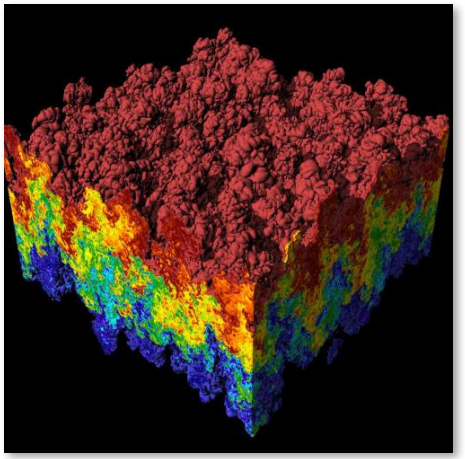
Visual representation of data

- Charts (line, bar, pie etc.), plots, diagrams
- 3D images, renderings
- Even numbers/digits!

- Visual Analytics
  - Sub-discipline
  - Large, complex data
  - Visualization + data analytics (data mining, machine learning etc.)
  - Interactive tools

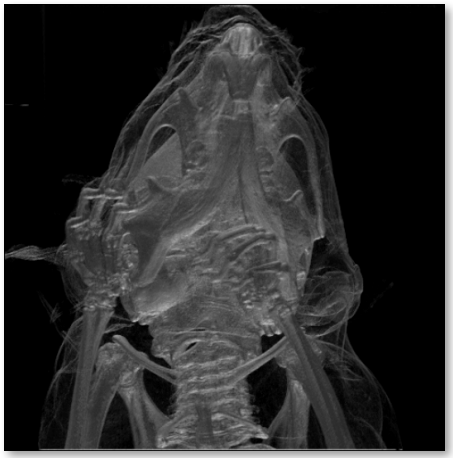


[By UCRL [Public domain], via Wikimedia Commons]



[By Lawrence Livermore National Laboratory [Public domain], via Wikimedia Commons]

434232
43432
4341
43455322
43489



[Lackas at English Wikipedia [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons]

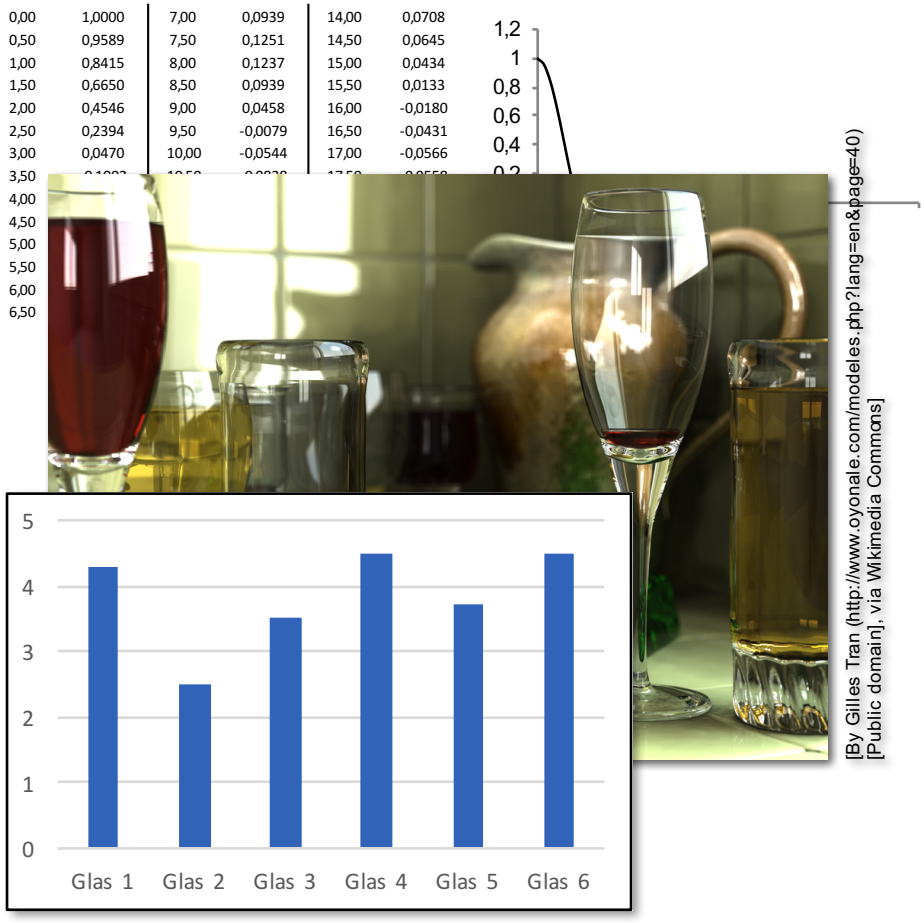
SFB-TRR 161

Why Visualization?

Transport information, provide insight

- Human visual system largest information channel
- Trained to recognize patterns

- It‘s not about realism (computer graphics)
- It‘s about showing only important/relevant things
- Filtering, highlighting, aggregation
- Provide insight, support reasoning

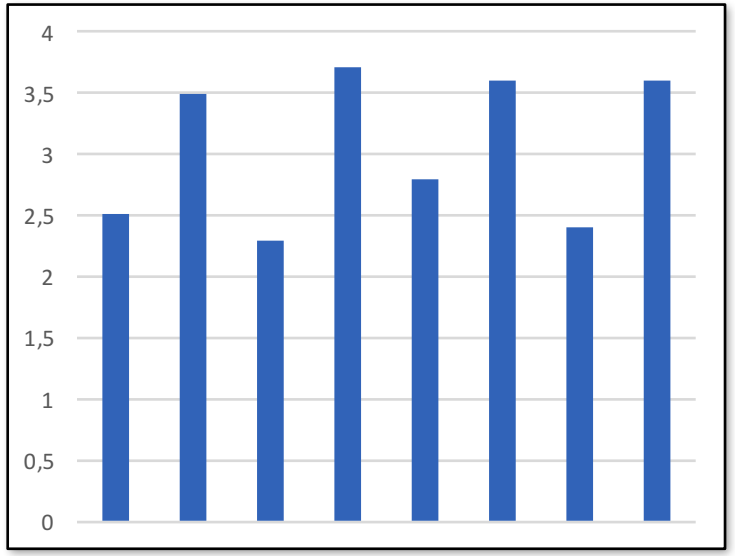


How should we use visualization?

Exploration, unclear questions, unknown data

- Exploration of data
- Searching for patterns
- Answer complex, fuzzy questions
- Communicate results

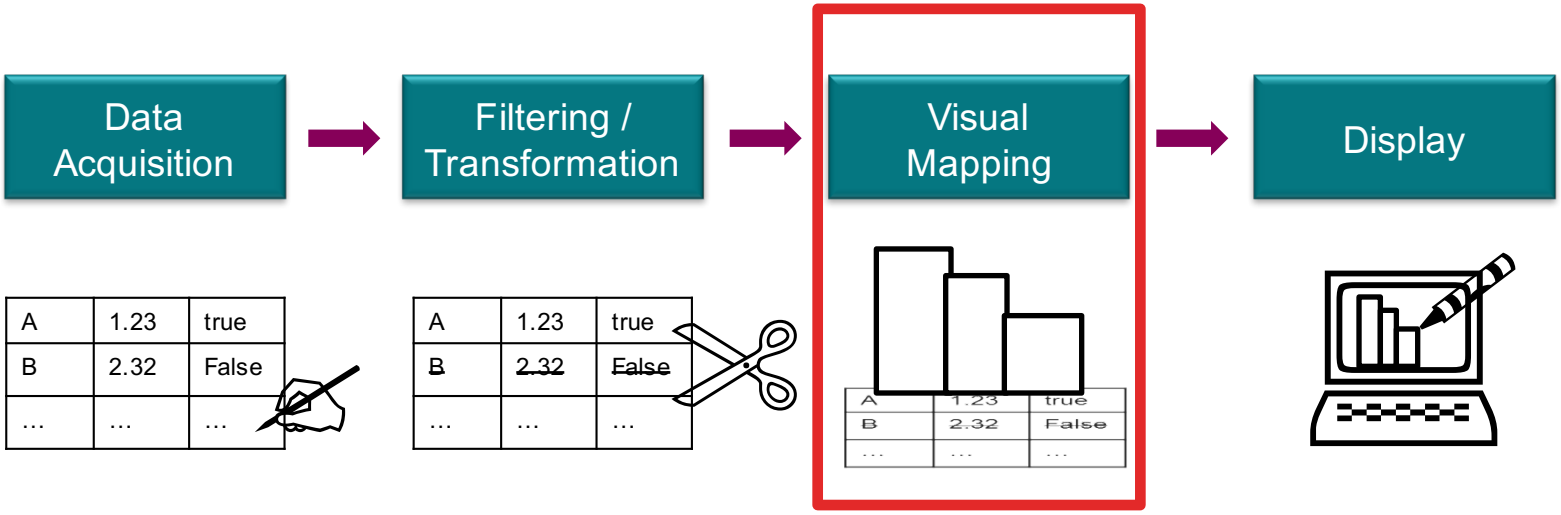
- If you know what are you looking for ➡ use queries
- Example chart
  - Minimum/maximum ➡ visualization not required
  - Oscillation ➡ visualization



Visualization Basics

Visualization Pipeline

Basic steps

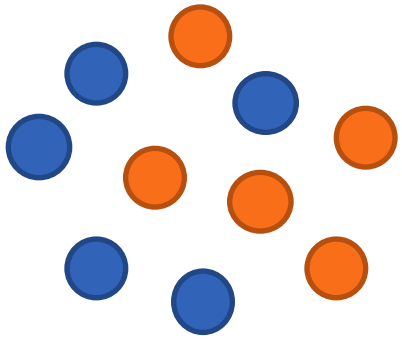


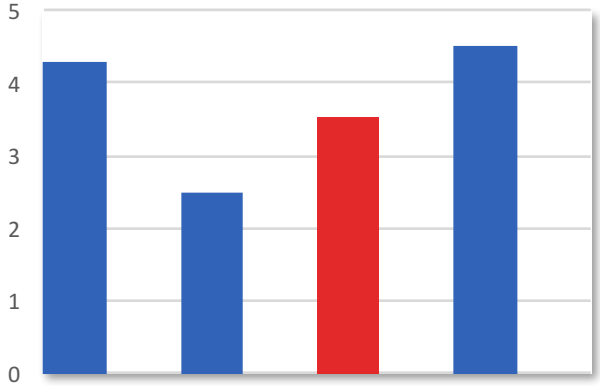
- Focus on mapping step in this tutorial

Color – colorful is beautiful?

Usage of color

- Represent values
- Highlight items
- Group / separate

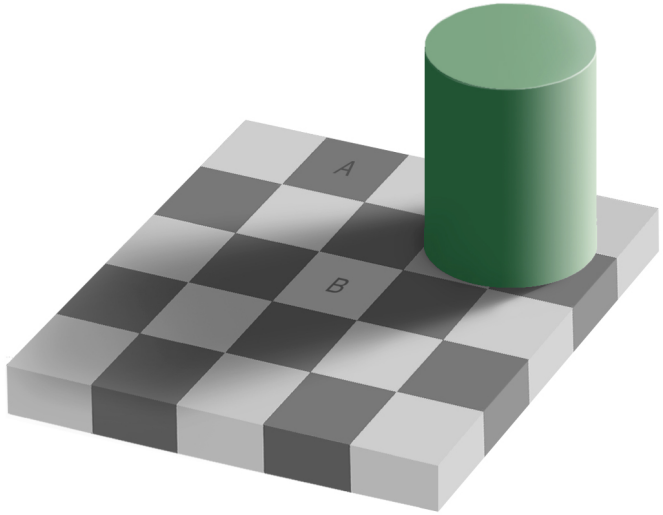




Color – colorful is beautiful?

Issues

- Perception problems:
  limited resolution, influence of surrounding
- Bad combinations (color blindness, etc.)



[By Original by Edward H. Adelson, this file by Gustavb
[Copyrighted free use], via Wikimedia Commons]

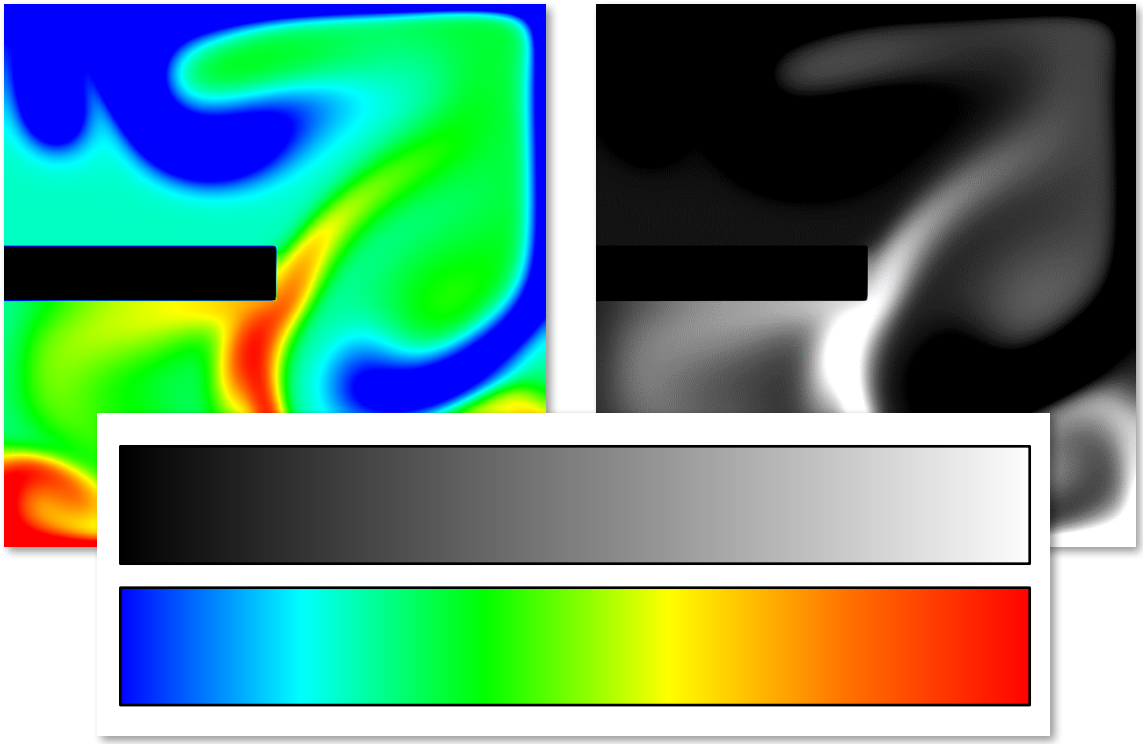
Bad Combination

Bad Combination

Color – colorful is beautiful?

Color maps

- Which color map?
- Interpretation (red = hot etc.)
- Linearity

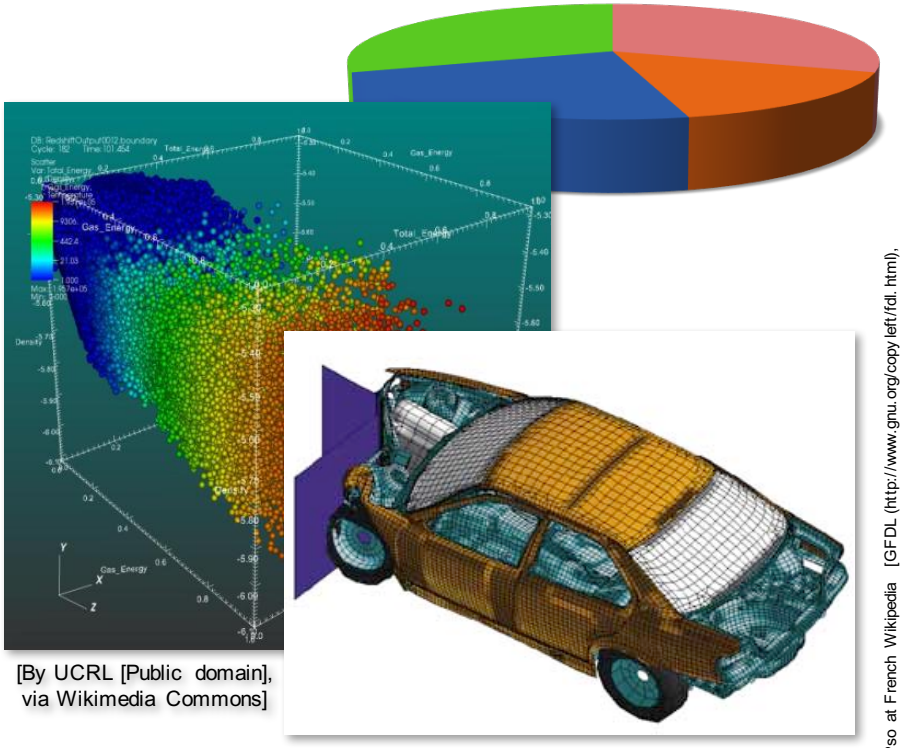


[1] D. Borland and R. M. Taylor, "Rainbow Color Map (Still) Considered Harmful," in IEEE Computer Graphics and Applications, vol. 27, no. 2, pp. 14-17, 2007.
[2] Brewer, C. A. 1999. Color Use Guidelines for Data Representation, Proceedings of the Section on Statistical Graphics, American Statistical Association, pp. 55-60.
[3] http://colorbrewer2.org/

3D – more dimensions are better?

Issues

- Occlusion, perspective, depth perception
- When can 3D be suitable?
  - Representation of 3D spatial information (e.g., CT Scans, crash simulations)
  - Navigation in 3D space
  - …



[By UCRL [Public domain], via Wikimedia Commons]

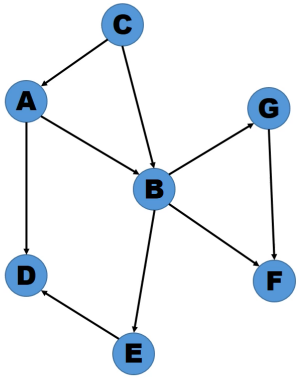
[Pso at French Wikipedia [GFDL (http://www.gnu.org/copyleft/fdl.html), CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/) or CC BY-SA 2.5-2.0-1.0 (http://creativecommons.org/licenses/by-sa/2.5-2.0-1.0)], via Wikimedia Commons]

[4] R. Brath. 3D infovis is here to stay: Deal with it. In Proceedings of IEEE VIS International Workshop on 3DVis, 2014.
[5] M. Tory, A. E. Kirkpatrick, M. S. Atkins, and T. Moller. Visualization task performance with 2D, 3D, and combination displays. IEEE Trans. On Visualization and Computer Graphics, 12(1), 2006.

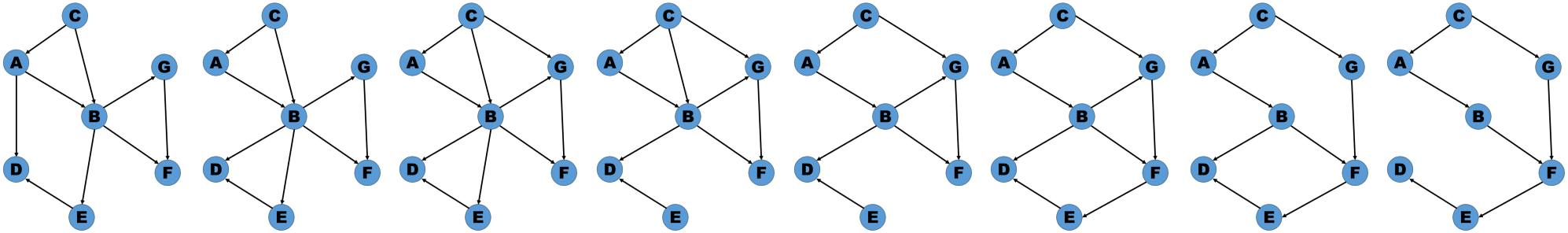
Animation

Sequential display of information

- Relies on short-term memory
- Good for direct comparison (change blindness)
- Bad for exploration (repeated playback)



[By Cary Bass (Own wok) [GFDL (http://www.gnu.org/copyleft/fdl.html), CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/) or CC BY-SA 2.5-2.0-1.0 (http://creativecommons.org/licenses/by-sa/2.5-2.0-1.0)], via Wikimedia Commons]



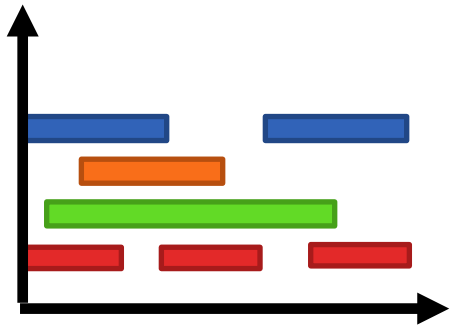
[6] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? International Journal of Human-Computer Studies, 57(4), 2002.

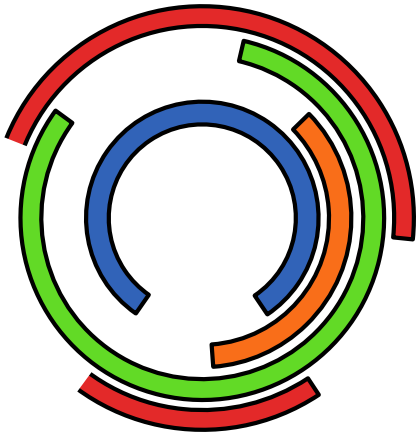
Visual Representations

Time-Dependent Data

How to represent time?

- Static vs. dynamic representations (animation, video)
- Temporal dimension as spatial dimension (maybe 3D)
- Timelines (Gantt charts etc.)
  - Easy comparison
  - Continuous
- Radial layout
  - Periodicity



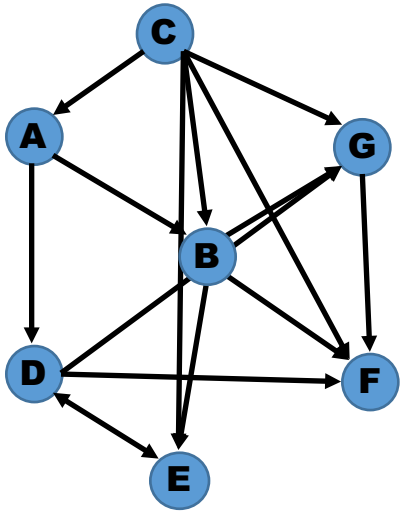


[7] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. Visualization of Time-Oriented Data. Human–Computer Interaction Series. Springer London, 2011.
[8] M. Burch and D. Weiskopf. On the benefits and drawbacks of radial diagrams. In W. Huang, editor, Handbook of Human Centric Visualization. Springer New York, 2014.

Graph Visualization

Different visual representations

Node-Link Diagram



Adjacency Matrix

	A	B	C	D	E	F	G
A		■		■			
B					■	■	■
C	■	■			■	■	■
D					■	■	■
E				■			
F							
G						■	

Adjacency List

A	B	D			
B	E	F	G		
C	A	B	E	F	G
D	E	F	G		
E	D				
F					
G	F				

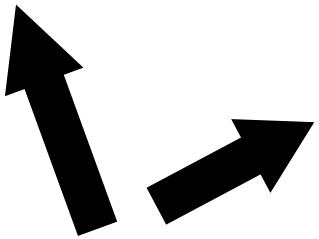
Glyphs

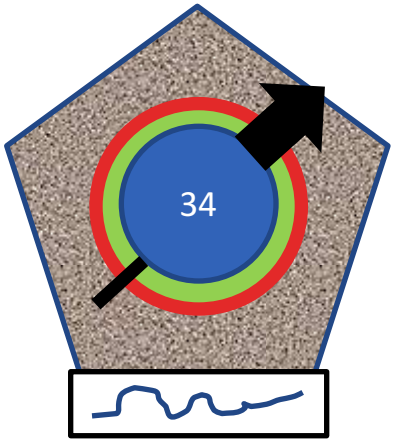
Symbolic encoding of data

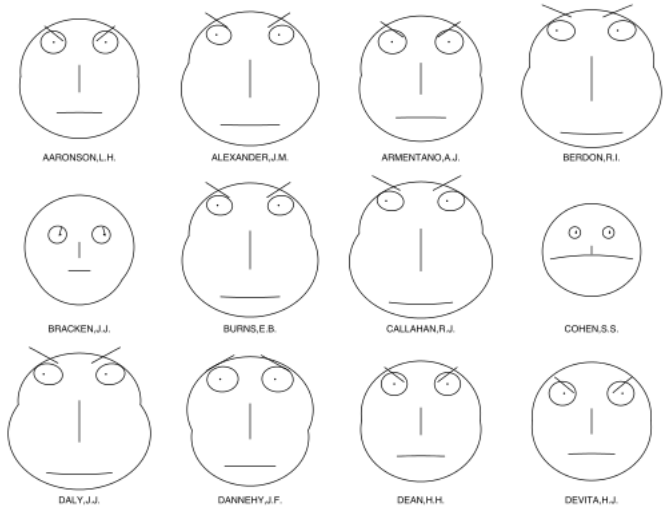
- Represent multiple data dimensions
- Many design parameters

- Try to keep it intuitive
- Keep the complexity low
- Metaphors from everyday life

name	angle	magnitude
A	110°	4
B	28°	3







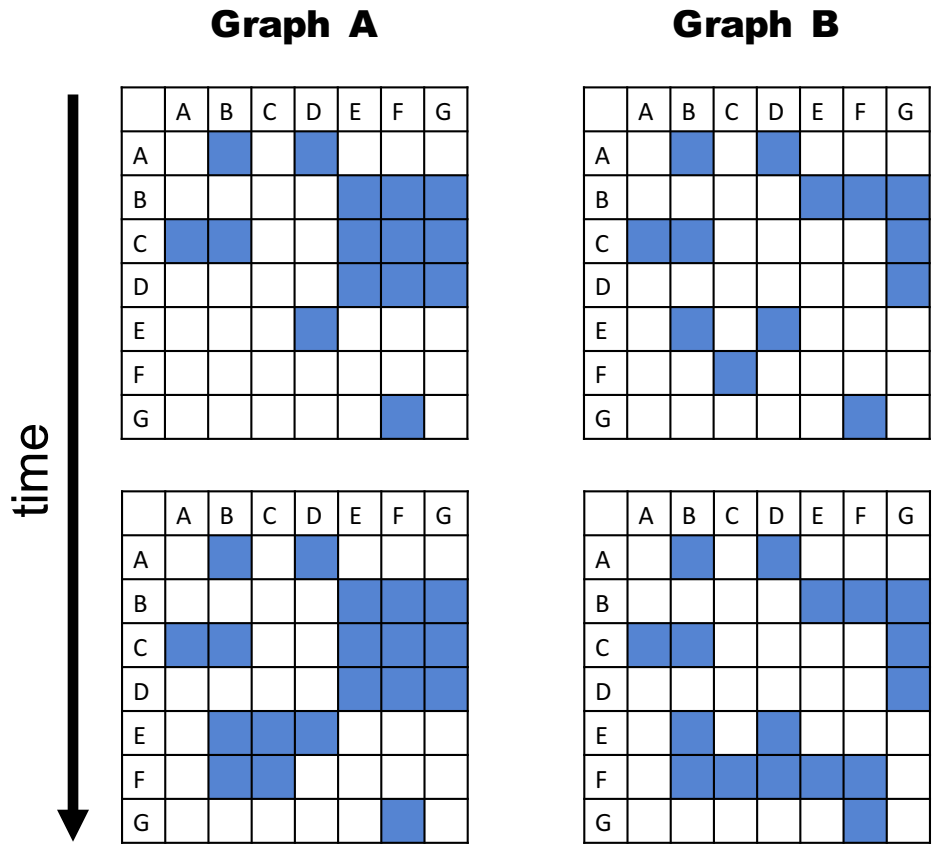
[By Avenue (Own work) [Public domain], via Wikimedia Commons]

[9] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. Eurographics State of the Art Reports, 2013.

Small Multiples

Multiple aligned plots / charts

- Comparison of different datasets or partitions of the data
- Alignment for easier comparison (grid)
- Semantics of grouping



[10] Edward Tufte, Envisioning information, Graphics Press, Cheshire, CT, 1990.

Sparklines

Word-sized plots / graphics

- Combined with text
- Smooth integration
- Overview, indicator for detailed exploration
- Plots, charts, symbols
- Readability on small sizes
- Intuitive understanding

	today	month
Company A	**7.85**	
Company B	**5.12**	
Company C	**10.32**	

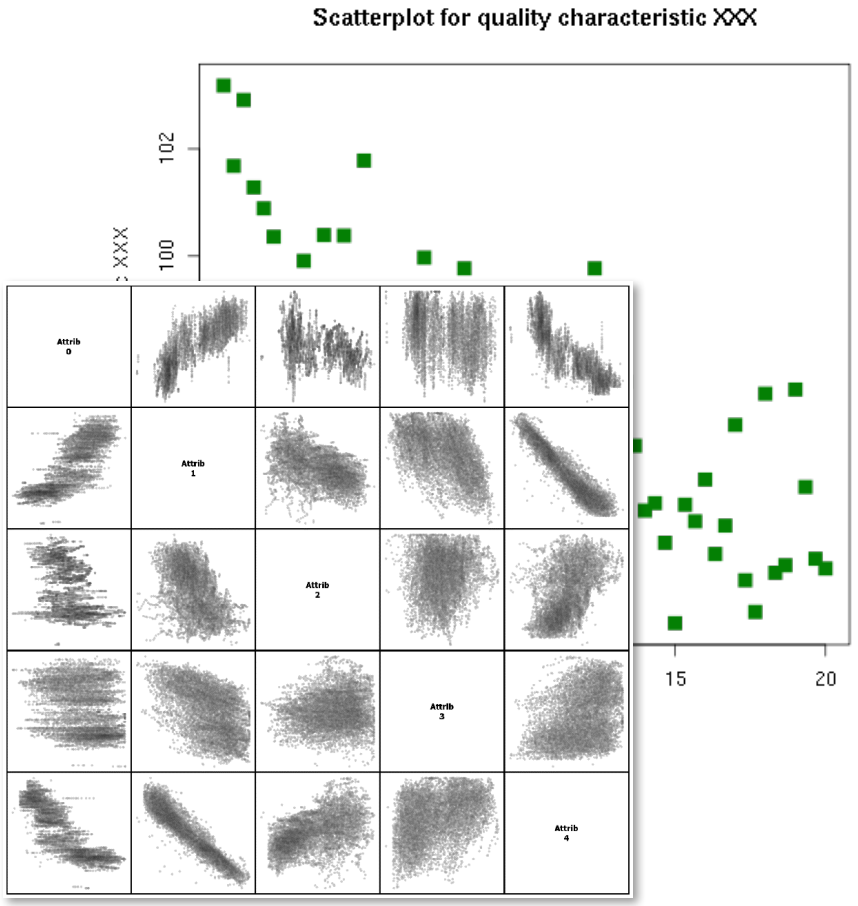
	output	state
Engine A	**89.34 %**	
Engine B	**0.0 %**	
Engine C	**73.45 %**	

[11] Edward Tufte, Beautiful Evidence, Graphis Press, Cheshire, CT, 2006.

Multi-Dimensional Data

Scatterplots

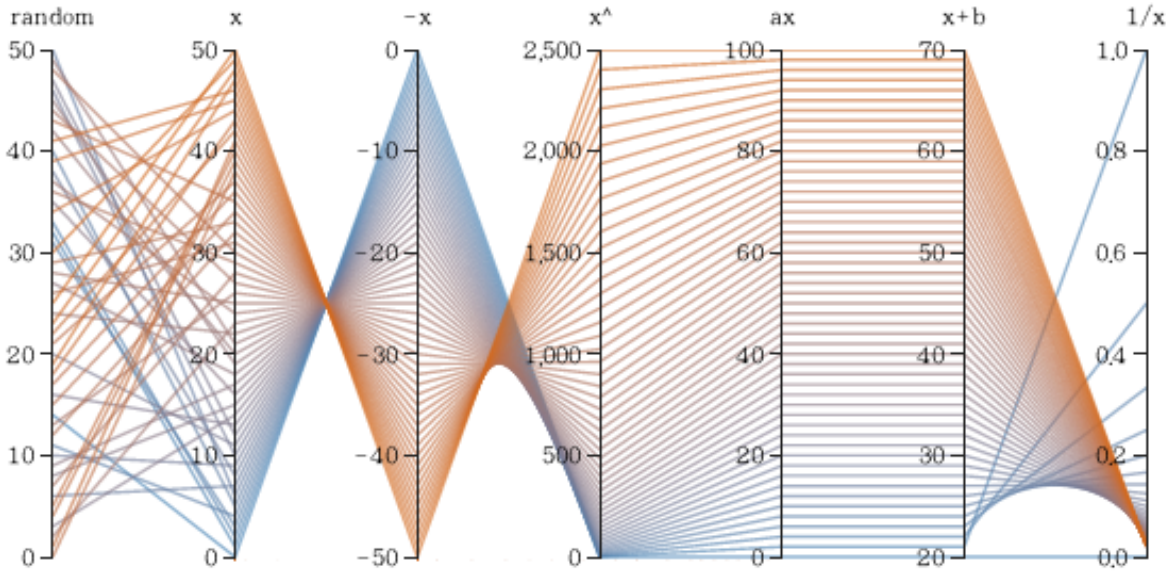
- Shows relation between 2 dimensions
- Matrix for multiple dimensions
- All combinations of dimensions



Multi-Dimensional Data

Parallel coordinates

- Parallel axes for data dimensions
- Data point represented by line strip
- Relation between two dimensions
- Ordering issues



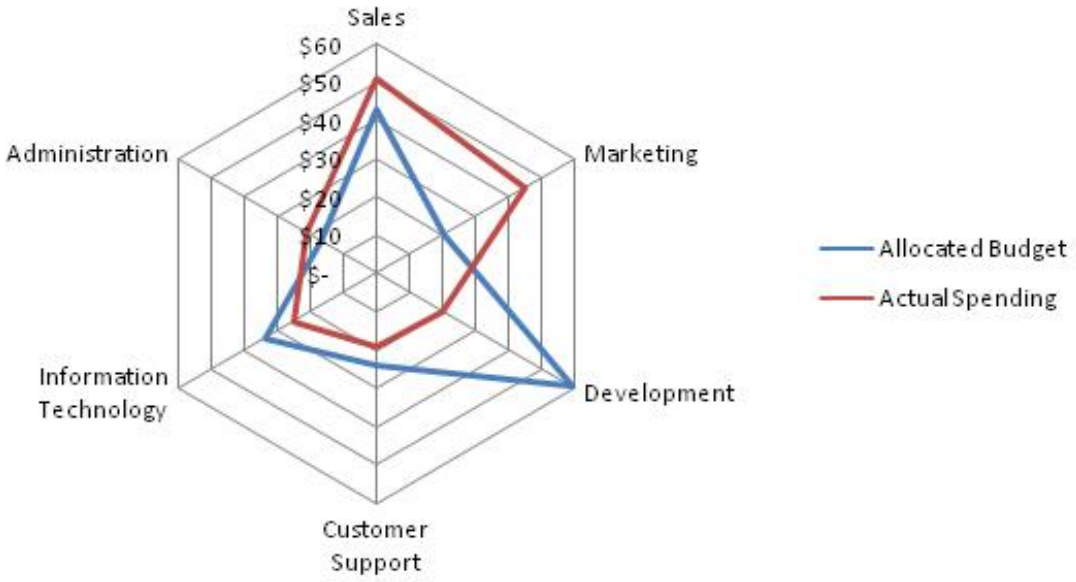
[By Yug (Own work) [CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons]

[12] Julian Heinrich, Daniel Weiskopf: State of the Art of Parallel Coordinates. Eurographics (STARs) 2013: 95-116.

Multi-Dimensional Data

Radar chart

- Radial layout of dimensions
- Polygon for each data point
- Interpretation difficult (meaning of area?)



[By David Clement (Own work by the original uploader) [Public domain], via Wikimedia Commons]

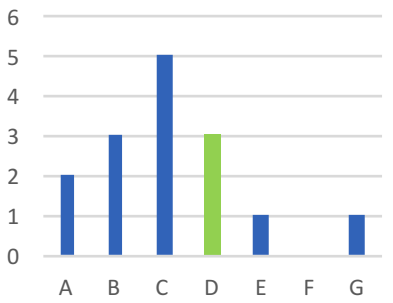
Multiple Coordinated Views

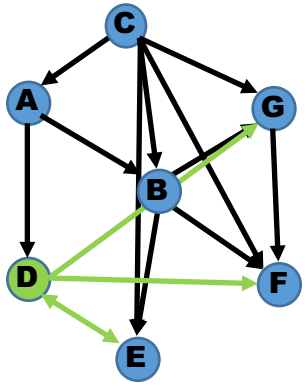
Multiples views to show different aspects

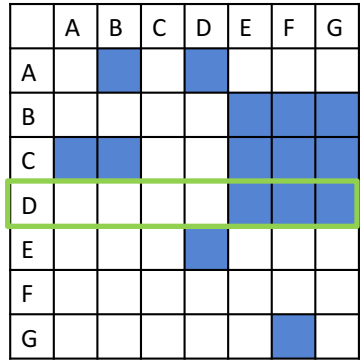
- Combine the advantages of different visualizations
- Consistent view is important
- Update all views

- Brushing and linking
  - Select in one view
  - Highlight in other views

Vertex	Edges
A	2
B	3
C	5
D	3
E	1
F	0
G	1







[13] J. C. Roberts, "State of the Art: Coordinated & Multiple Views in Exploratory Visualization," Internat. Conf. on Coordinated and Multiple Views in Exploratory Visualization, 2007, pp. 61-71.
[14] C. North and B. Shneiderman. Snap-together visualization: can users construct and operate coordinated visualizations? International Journal of Human-Computer Studies, 53(5), 2000.
[15] A. Buja, J. A. McDonald, J. Michalak, and W. Stuetzle. Interactive data visualization using focusing and linking. In International Conference on Visualization (VIS), 1991.

Approach 1 – Modules and Events

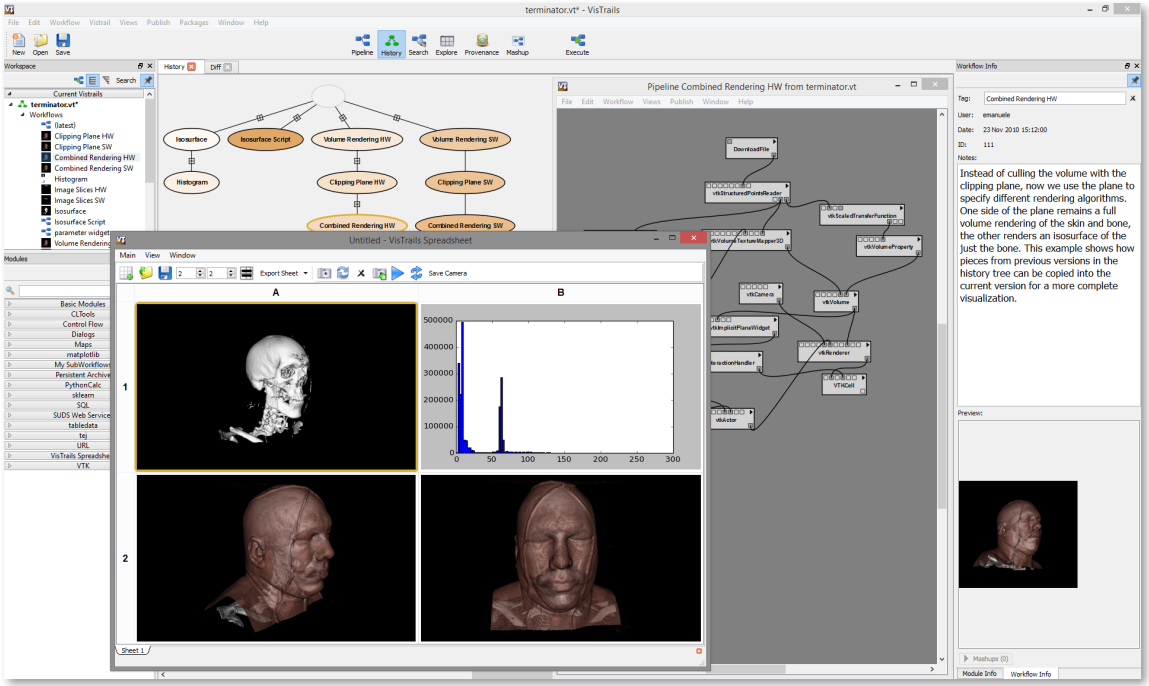
The Data

VisTrails

- Modular visualization framework
- Combine visualization modules

- Integrated provenance support
- All actions/changes are logged

- Workflow provenance

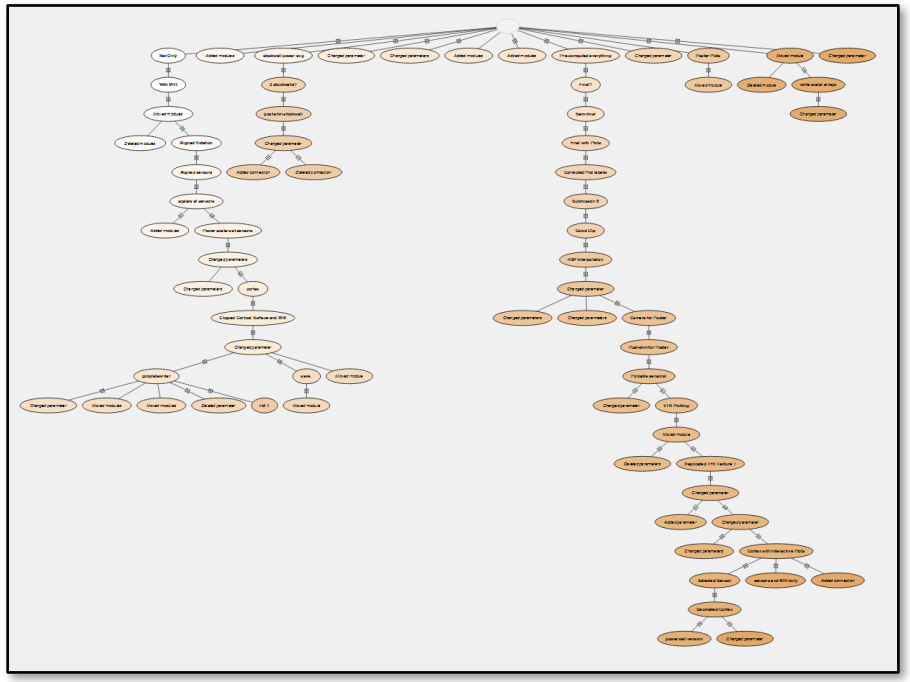


[16] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Vistrails: Visualization meets data management. In Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, 2006.

VisTrails – Provenance Visualization

History View

- History tree of changes
- Node link visualization
- Color for temporal dimension
- Labels
- Focus on branching



New Visualization Approach

Analysis of user behavior

- How was the result obtained?
- Which steps were performed?

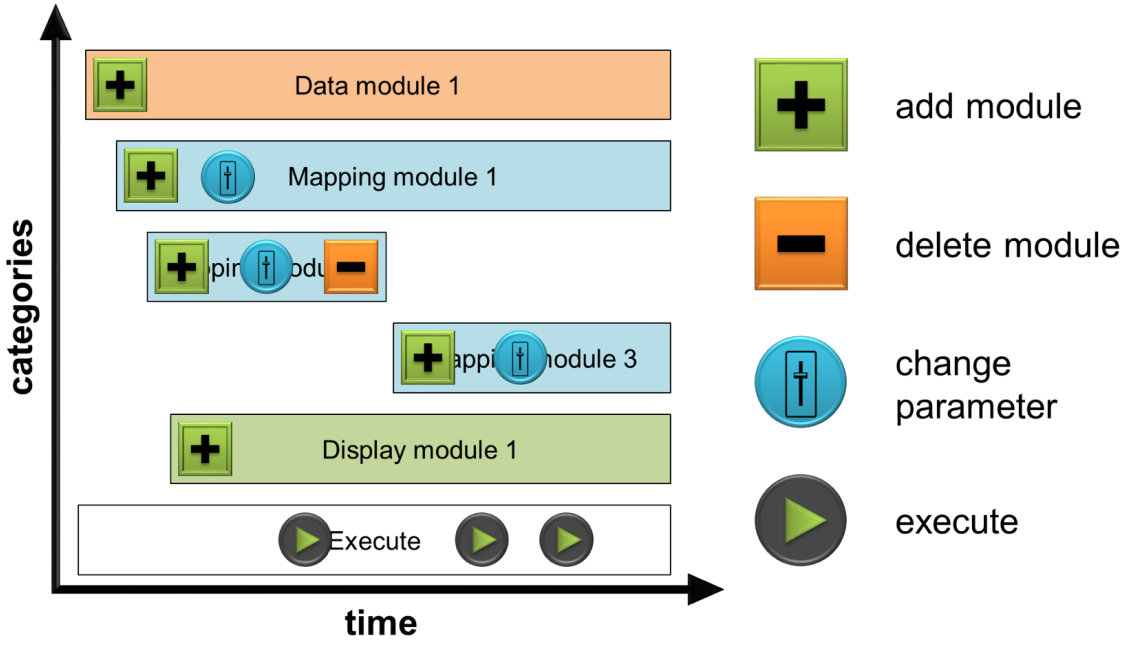
- Relevant aspects
  - Temporal order
  - Action/event type
  - Affected modules

New Visualization Approach

Visualization concept

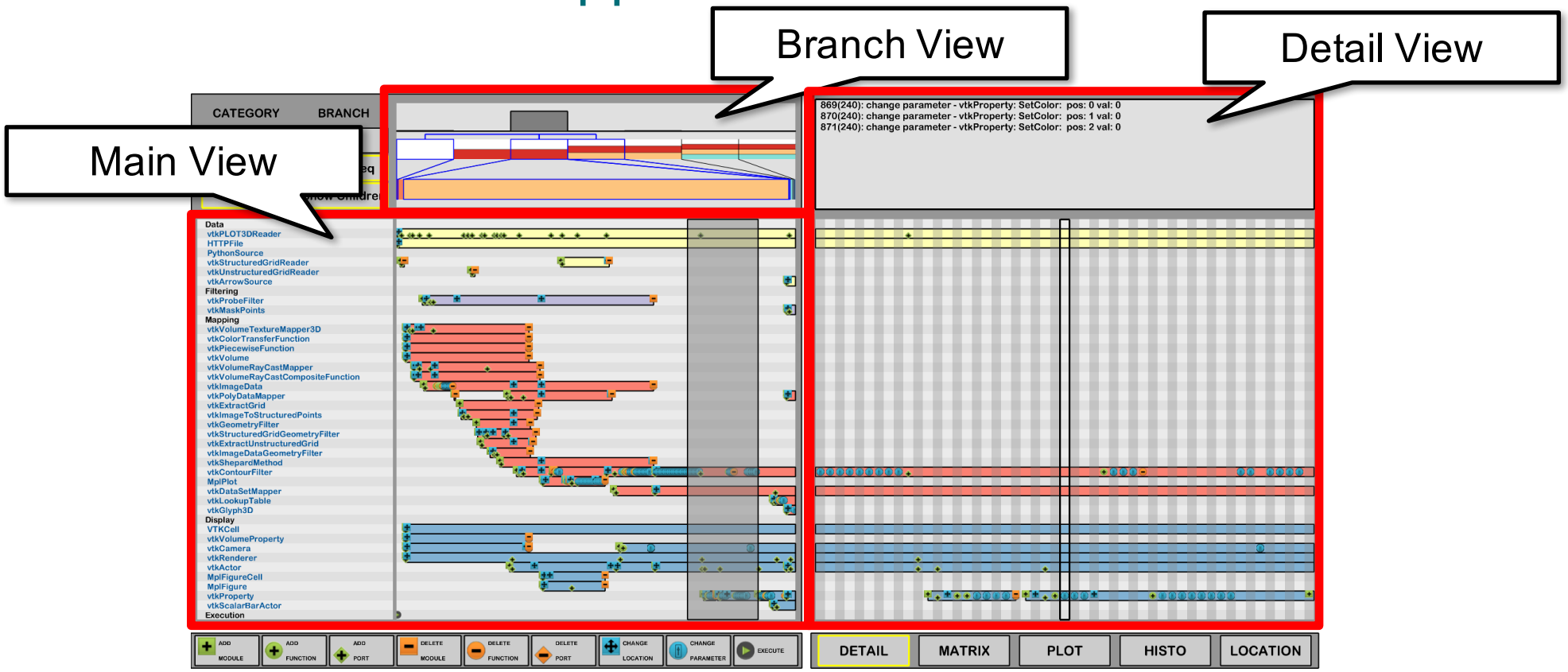
- Lifetime of modules
- Bars for timespans
- Coloring/grouping of module type

- Events as symbols
- Placed on related module

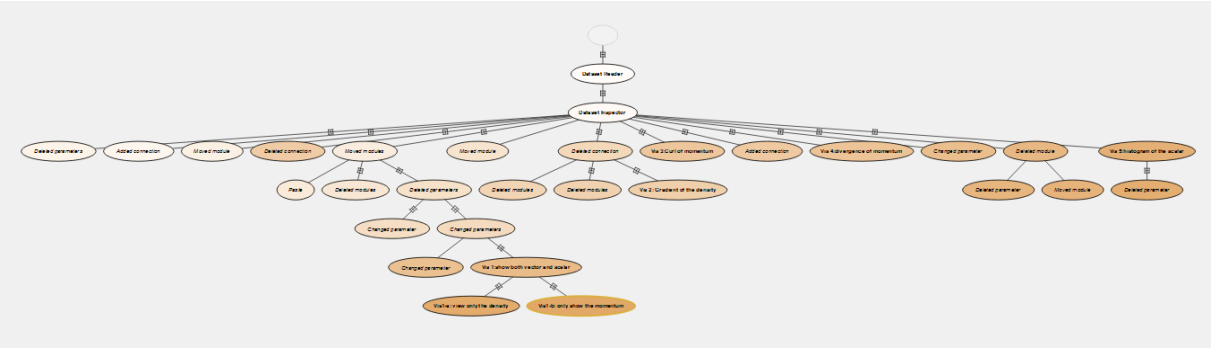


[17] M. Hlawatsch, M. Burch, F. Beck, J. Freire, C. Silva, and D. Weiskopf. Visualizing the evolution of module workflows. In International Conference on Information Visualisation (IV), 2015

New Visualization Approach



Branch View

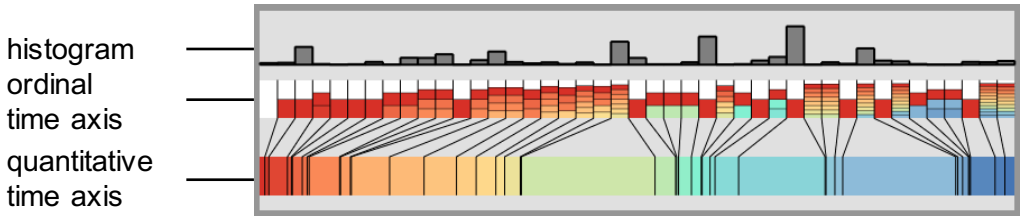


Tree

- Good for branching structure
- Requires space
- Time difficult to see

Our View

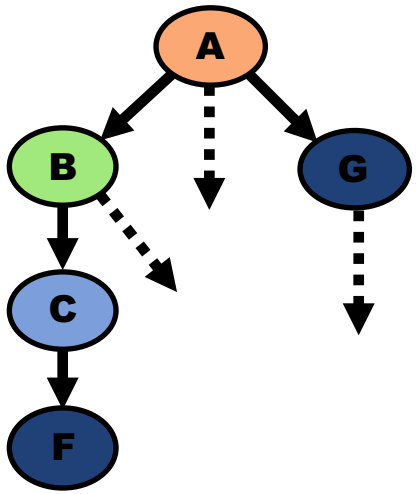
- More compact representation
- Temporal order
- Branching still visible

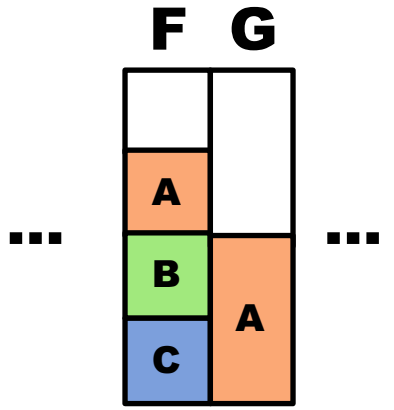


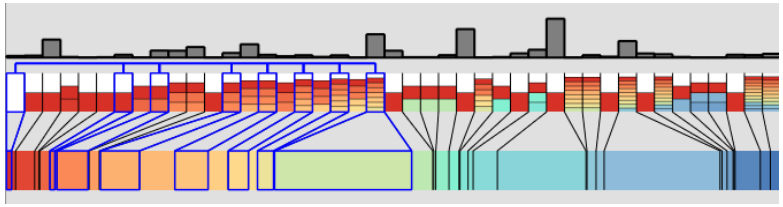
Branch View

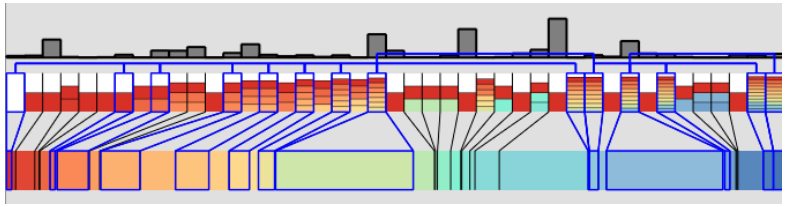
Block Diagram

- All previous branches stacked blocks
- Temporally ordered
- Color shows time of predecessor

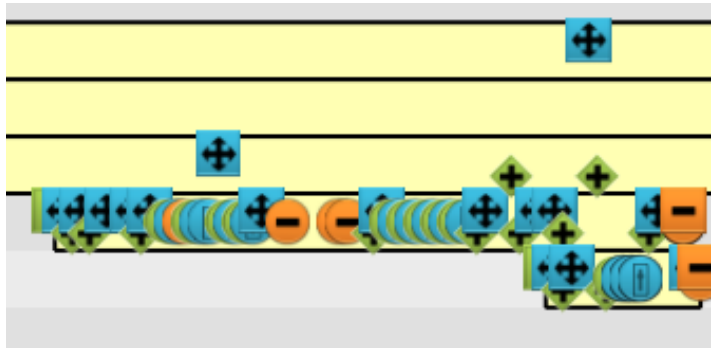








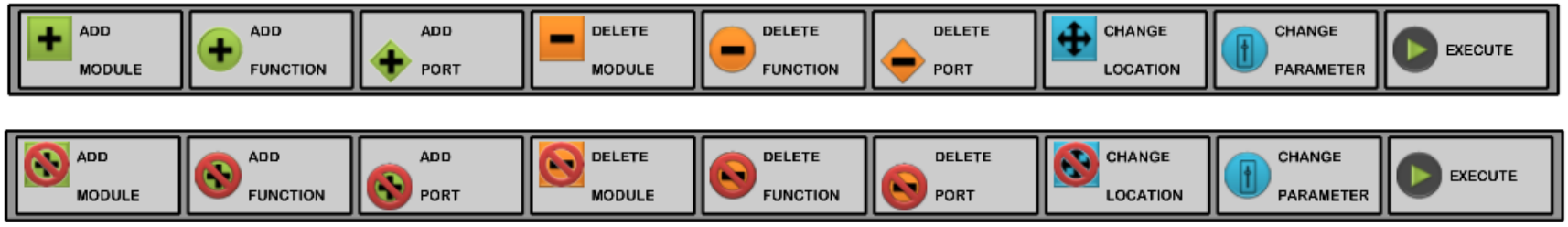
Symbol design



Scalability

- Large number of symbols
- Densely packed, overlap

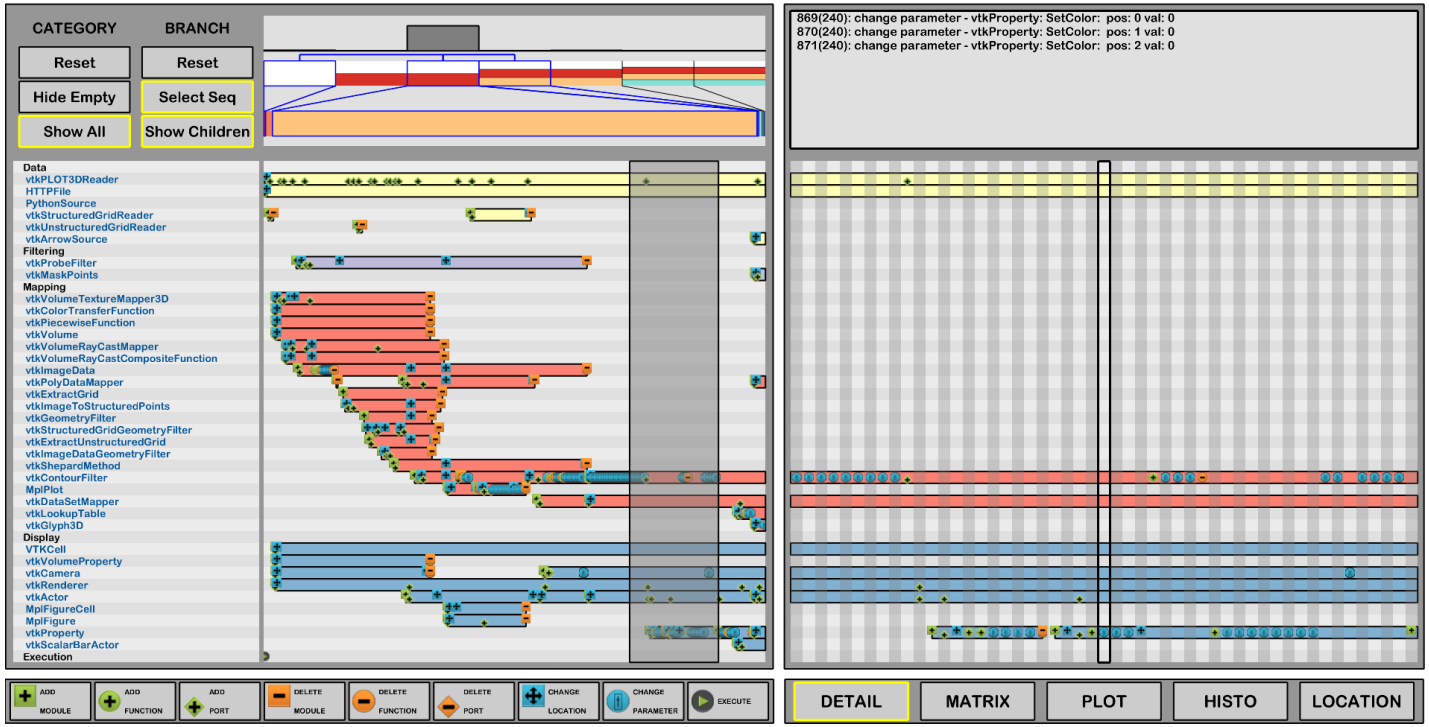
- Color for action type (green – add, orange – delete, blue – change)
- Shape for target type (square – module, circle – function, diamond – port)
- Different vertical placement



Detail Views

Zoom view

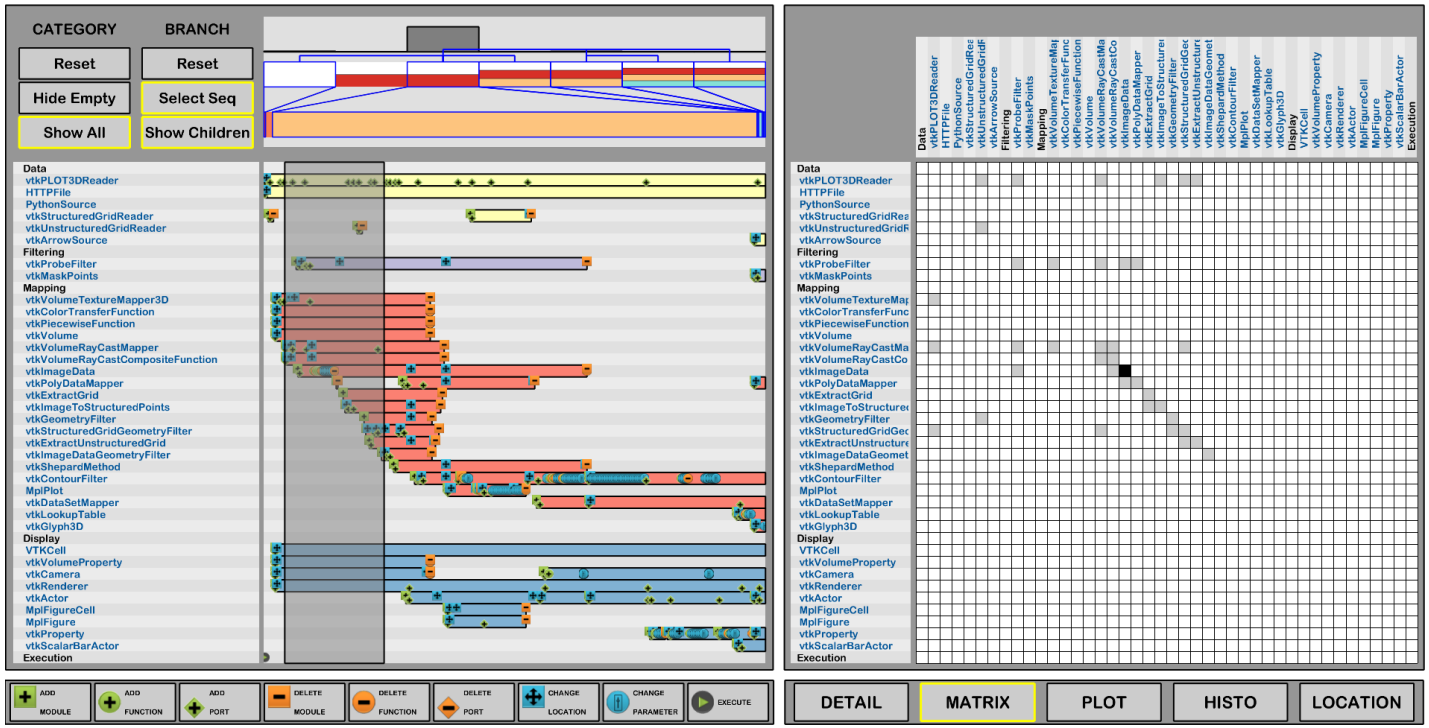
- Close-up view of event sequence
- Avoid overlapping symbols
- Raw data log on top



Detail Views

Transition matrix

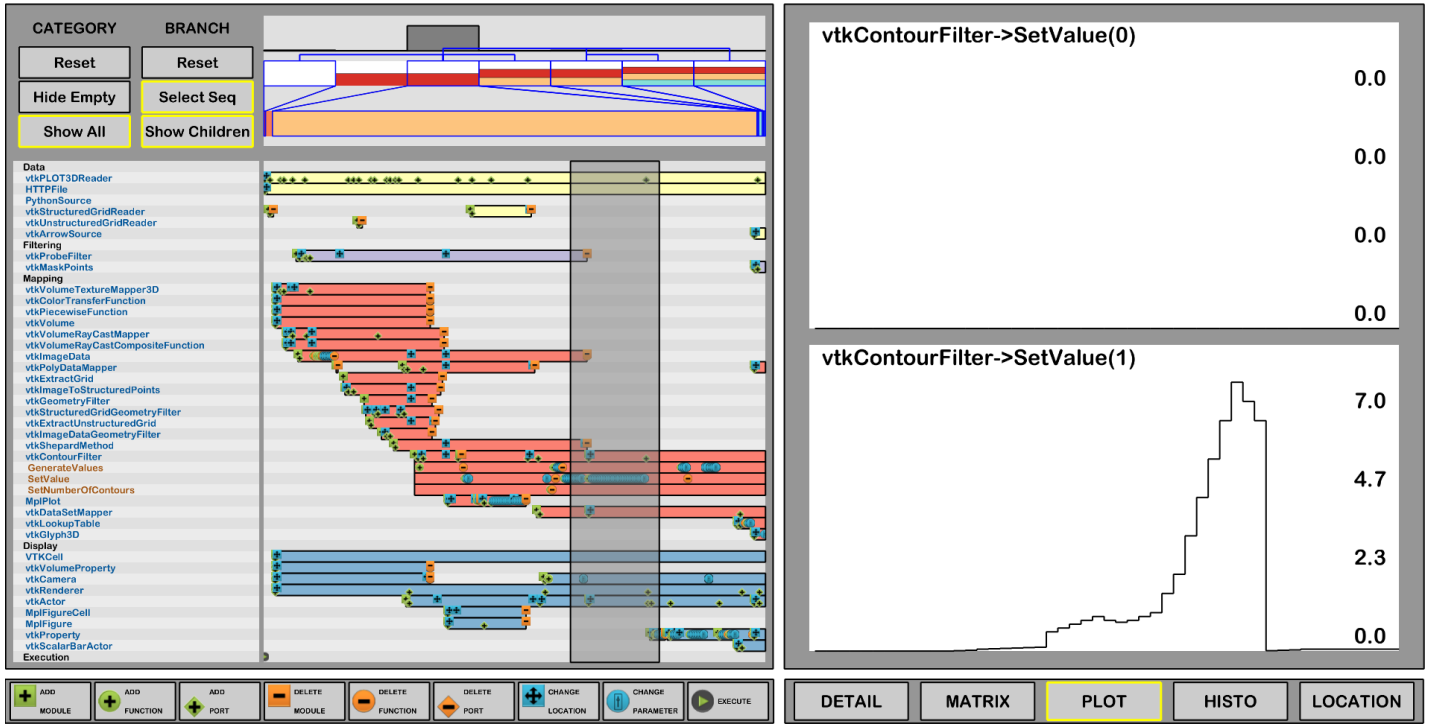
- Matrix of subsequent events
- Sequence patterns



Detail Views

Parameter plot

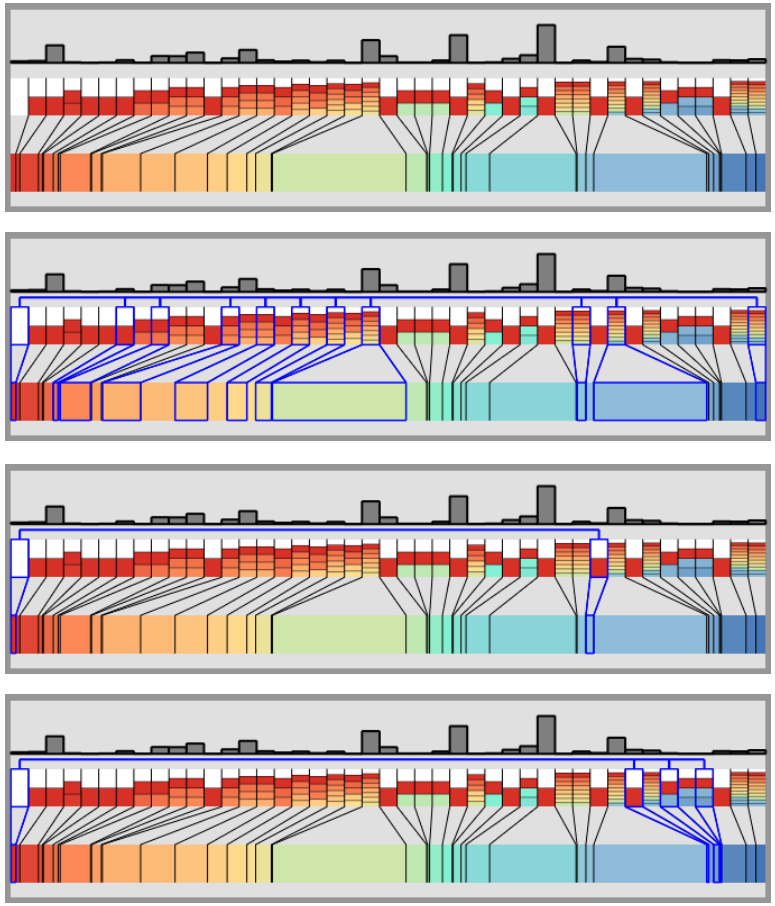
- Plot of parameter values over time
- Change behavior



Examples

Branch View

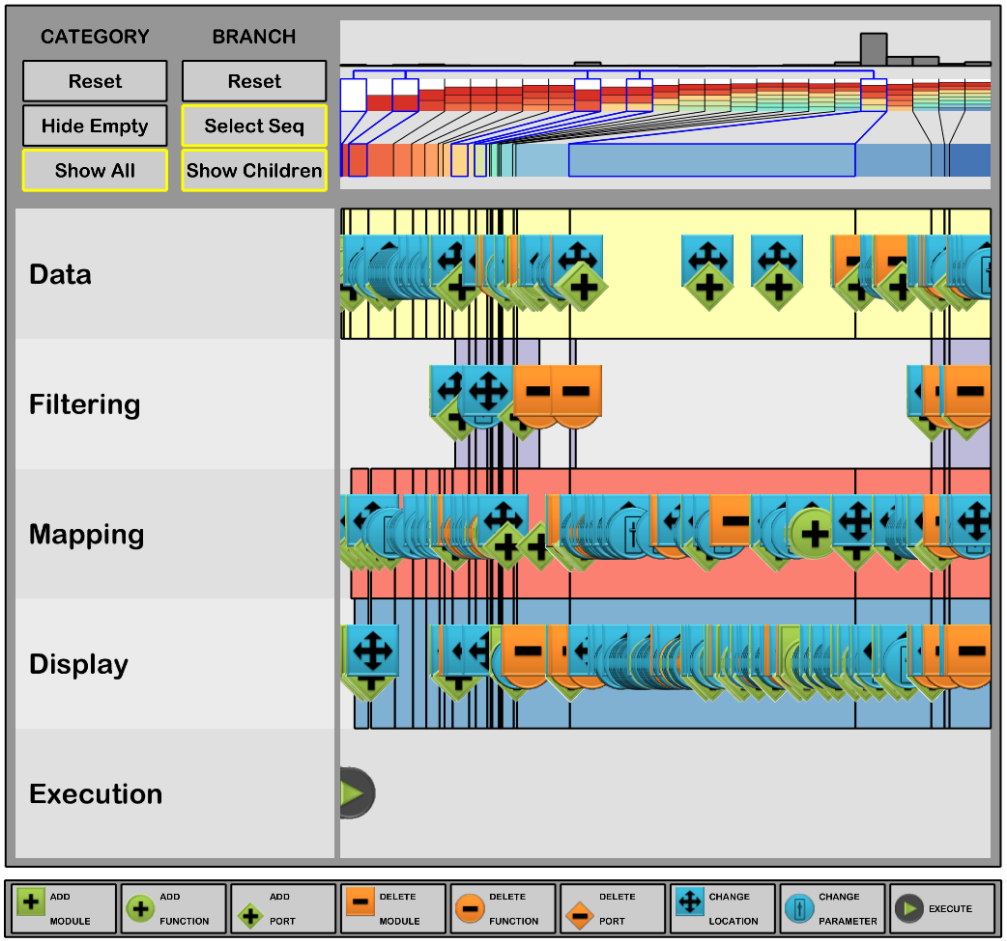
- Different time spans of sequences
- Restarted several times from beginning
- Last sequence has many parents



Examples

Main View

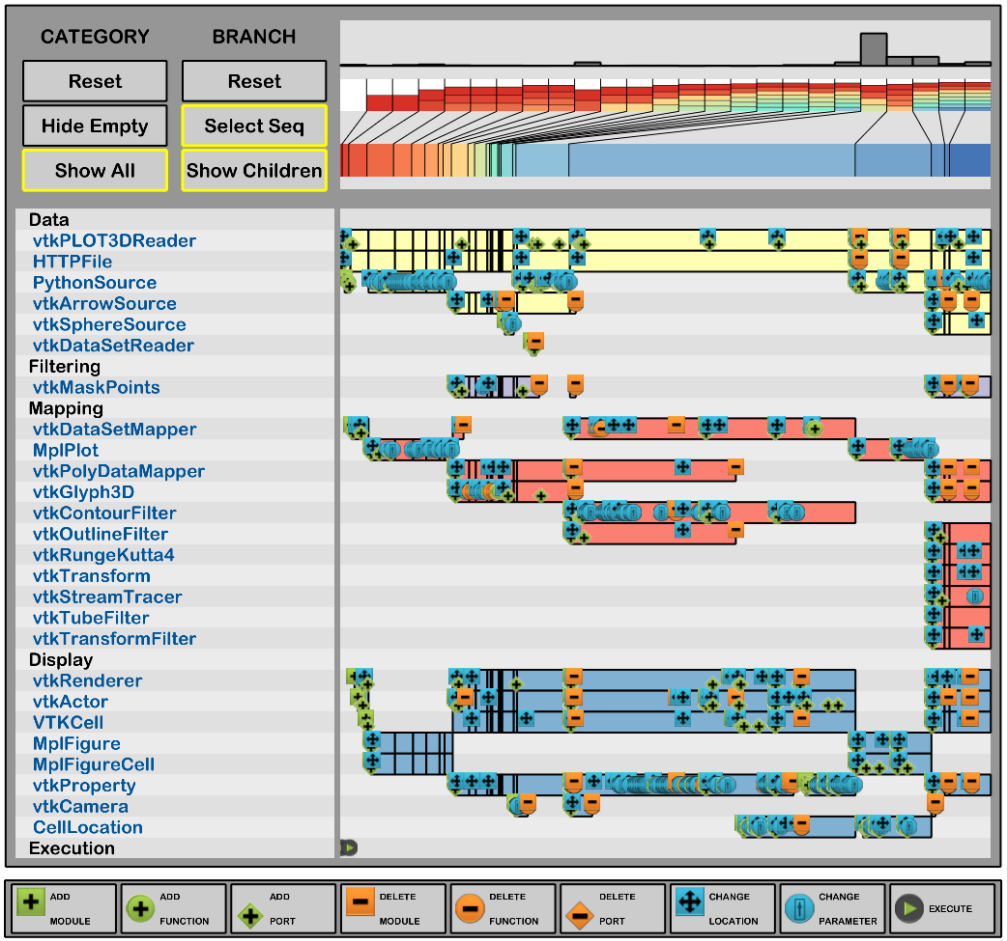
- Quite sequential progress
- One large sequence
- Belongs to early sequences



Examples

Main View

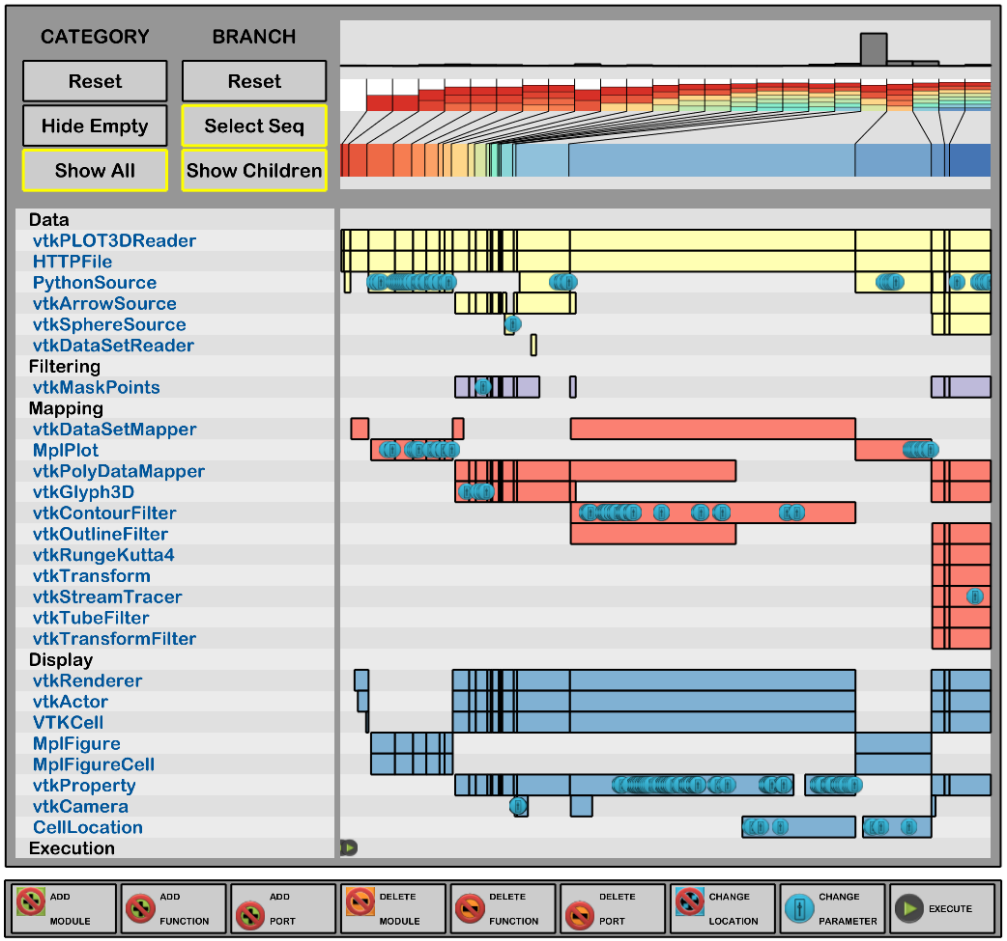
- Long sequence has different mapping modules
- No filtering in the long sequence



Examples

Main View

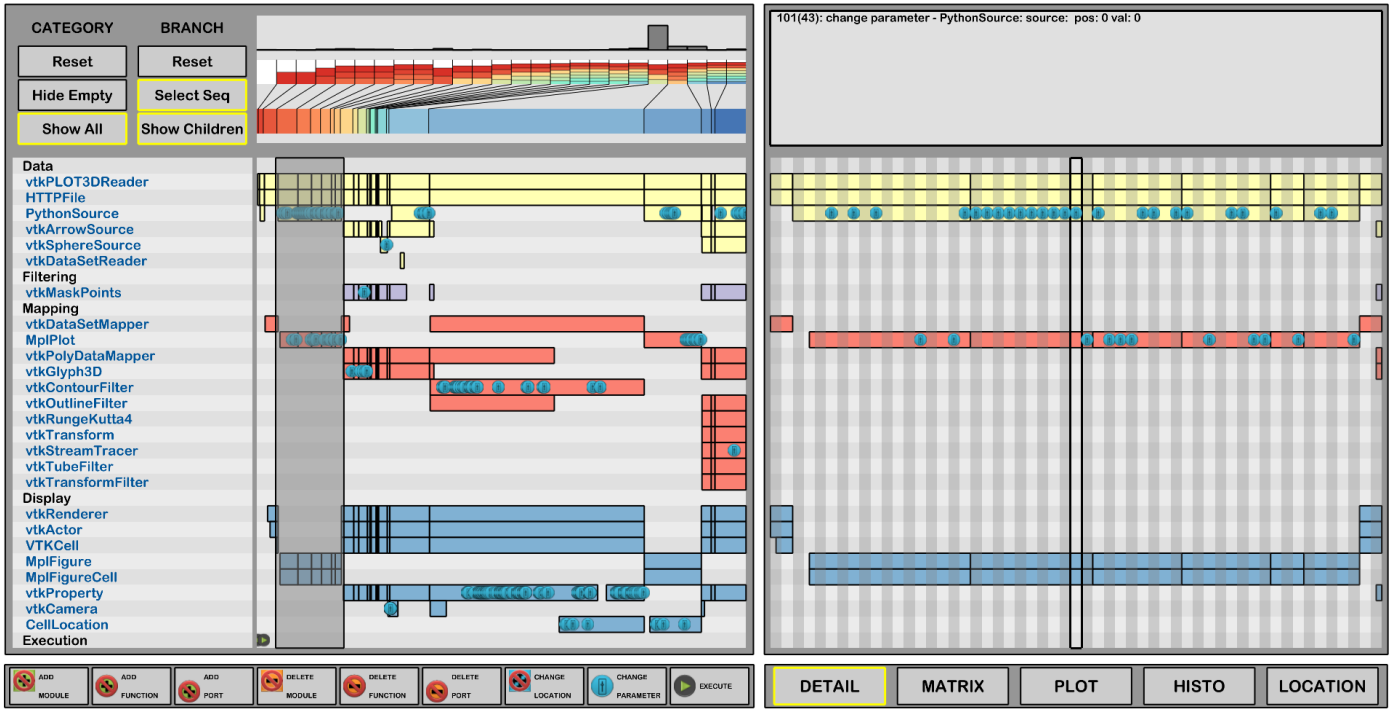
- Too many events / symbols
- Focus on parameter changes
- “vtkContour Filter”
  (isosurface visualization)
- “vtkProperty”
  (color and opacity)



Examples

Detail Views

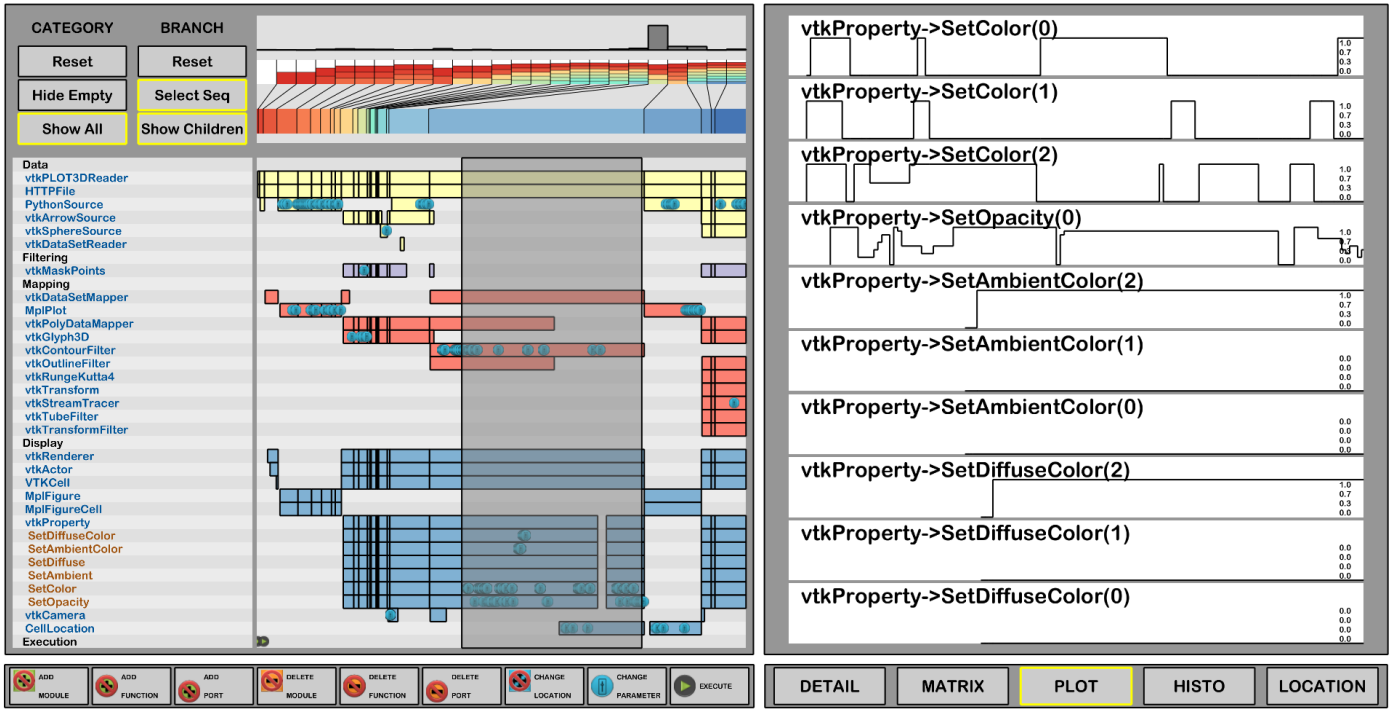
- Zoomed view
- Events do not overlap



Examples

Detail Views

- Parameter plot
- Changes of color and opacity

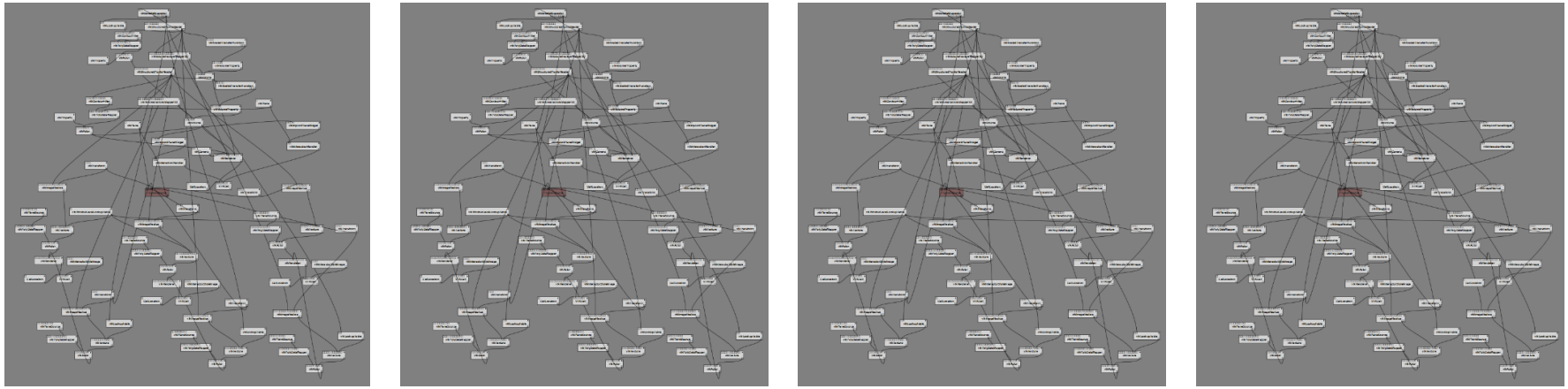


Approach 2 – Graph Structure

Graph Visualization

Changes in graph structure?

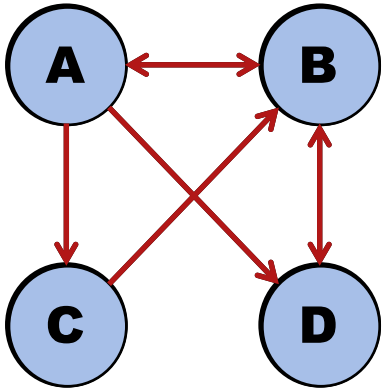
- Side-by-side visualization not always helpful
- Difficult to track individual vertices or edges over time

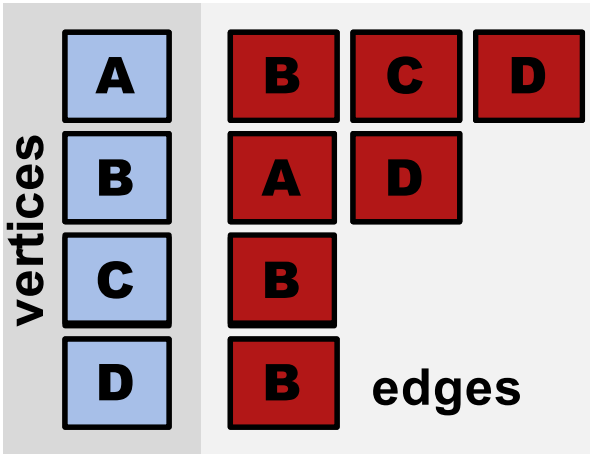


Adjacency Lists

Compact representation

- List as graph representation
  - Vertical axis of nodes
  - Each row contains vertices
- Space saving
- Flexible layout
- Easy to see
  - Direct connections
  - Number of vertices



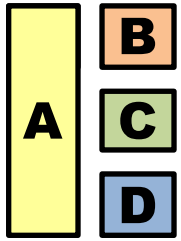


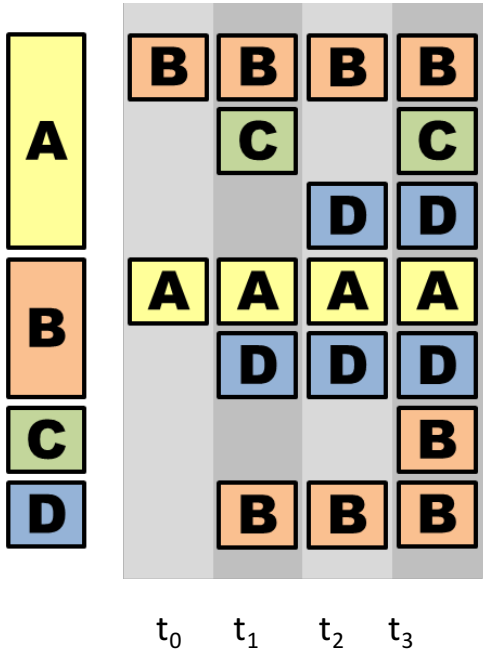
[18] M. Hlawatsch, M. Burch, and D. Weiskopf. Visual adjacency lists for dynamic graphs. IEEE Trans. on Visualization and Computer Graphics, 20(11), 2014.

Dynamic Graphs

Gantt layout

- Separate row for every vertex
- Time steps next to each other
- Visual fusion ➡ bars for time spans



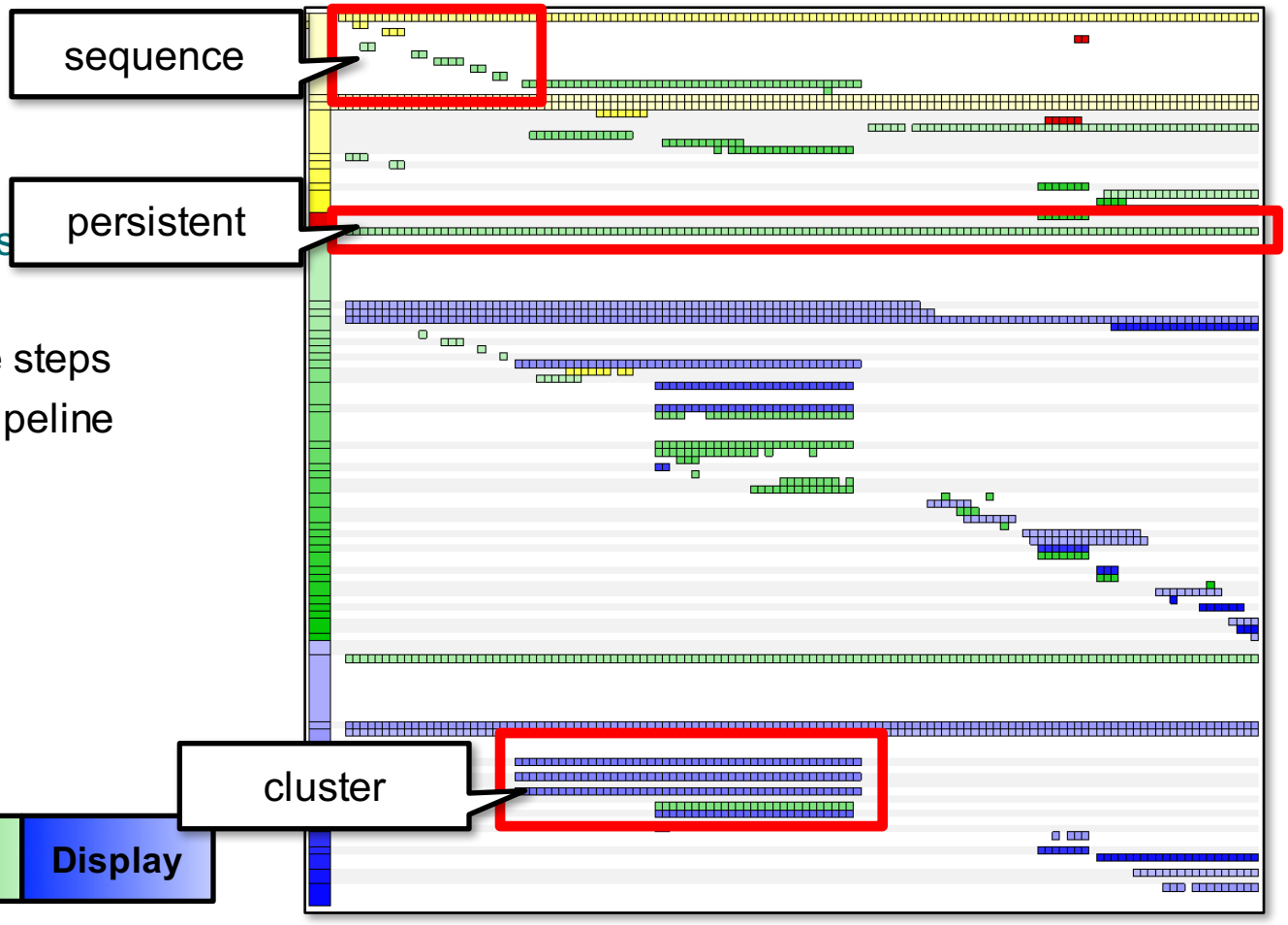


Example

Visualization workflow (VisTrails)

- 126 time steps
- Average 33 vertices per time steps
- Colored after visualization pipeline

- Sequences
- Persistent edges
- Cluster



General Hints

- Keep it simple, only show relevant data/information
- Avoid visual clutter (less visual elements)
- Carefully use colors, 3D, animation
- Scalability? Test your approach with large data
- Is it intuitive? Test your approach with persons not involved/non-experts

Agenda

- Part 1: Provenance
  - Overview
  - Workflow provenance
  - Data provenance
- Part 2: Visualization
  - Visualization Basics
  - Provenance Visualization 1 – Modules and events
  - Provenance Visualization 2 – Graph structure
- Open Research Questions

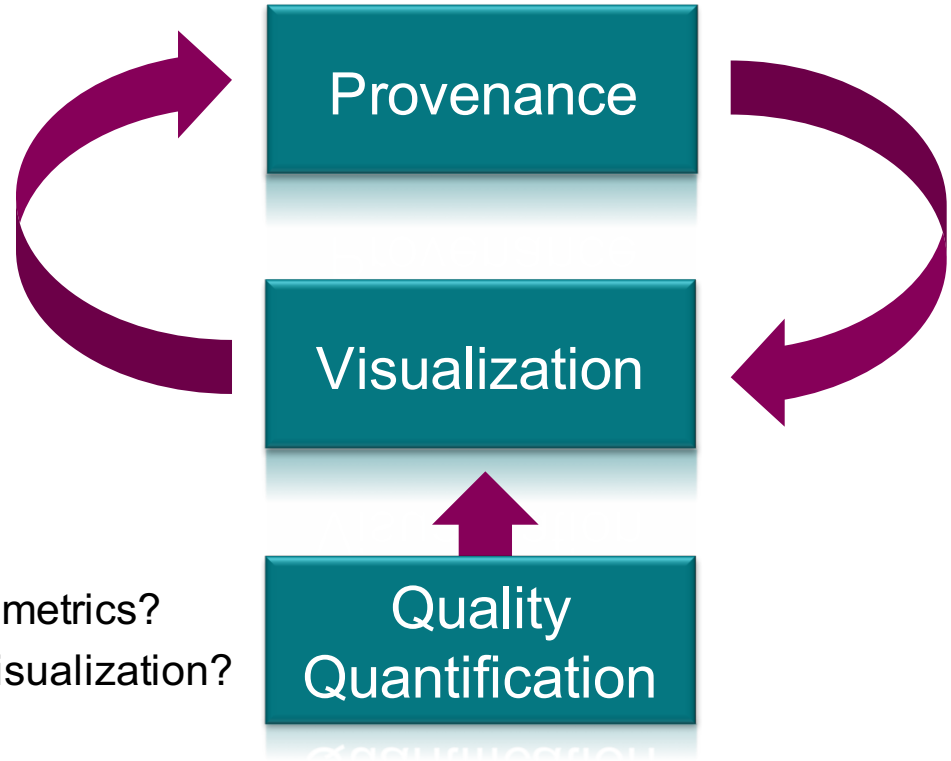
Summary

- Effective visualization requires careful design
- Human perception and knowledge must be considered
- Application dependent, visualization must be adapted
- No single approach suitable for all applications, use cases, data types

- **Cooperate with experts!**

Open Research Questions

- Support usage of provenance?
- Automatic selection of visualization method?



- Improve visualization methods?
- Prediction of user behavior?
- Support user?

- Objective quality metrics?
- What is a good visualization?

[19] E. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. IEEE Trans. On Visualization and Computer Graphics, 22(1), 2016.

Provenance References

- **Provenance Overview**
  - [Art1] https://www.marketingweek.com/2014/01/15/horse-meat-scandal-has-had-a-lasting-effect-on-how-food-is-marketed/
  - [Art2] https://hbr.org/2010/10/the-transparent-supply-chain
  - [Art3] http://information-technology.web.cern.ch/CHEP/data-analysis-preservation-services-lhc-experiments
  - [Nautilus] Nautilus Website: http://nautilus-system.org
  - [MG15] Screenshot produced using tool presented in Tobias Müller, Torsten Grust:Provenance for SQL through Abstract Interpretation: Value-less, but Worthwhile. PVLDB 8(12): 1872-1875 (2015)
- **Workflow provenance**
  - [FTC+06] Juliana Freire, Cláudio T. Silva, Steven P. Callahan, Emanuele Santos, Carlos Eduardo Scheidegger, Huy T. Vo:Managing Rapidly-Evolving Scientific Workflows. IPAW 2006.
  - [DCL+-7] S. B. Davidson, S. Cohen-Boulakia, A. Eyal, B. Ludäscher, T. M. McPhillips, S. Bowers, M. K. Anand, and J. Freire. Provenance in scientific workflow systems. *IEEE Data Eng. Bull.*, 30(4):44– 50, 2007.
- **Data provenance**
  - [CW00] Yingwei Cui, Jennifer Widom, and Janet L. Wiener. 2000. Tracing the lineage of view data in a warehousing environment. *ACM Trans. Database Syst.* 25, 2 (June 2000), 179-227.
  - [BHT14] Nicole Bidoit, Melanie Herschel, Katerina Tzompanaki: Query-Based Why-Not Provenance with NedExplain. EDBT 2014.
  - [CCT09] J. Cheney, L. Chiticariu, and W. C. Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4), 2009.