VisTrails: Using Provenance and Workflows for Scientific Exploration

David Koop
School of Engineering, New York University

Orchestrate Different Tools & Libraries

Orchestrate Different Tools & Libraries

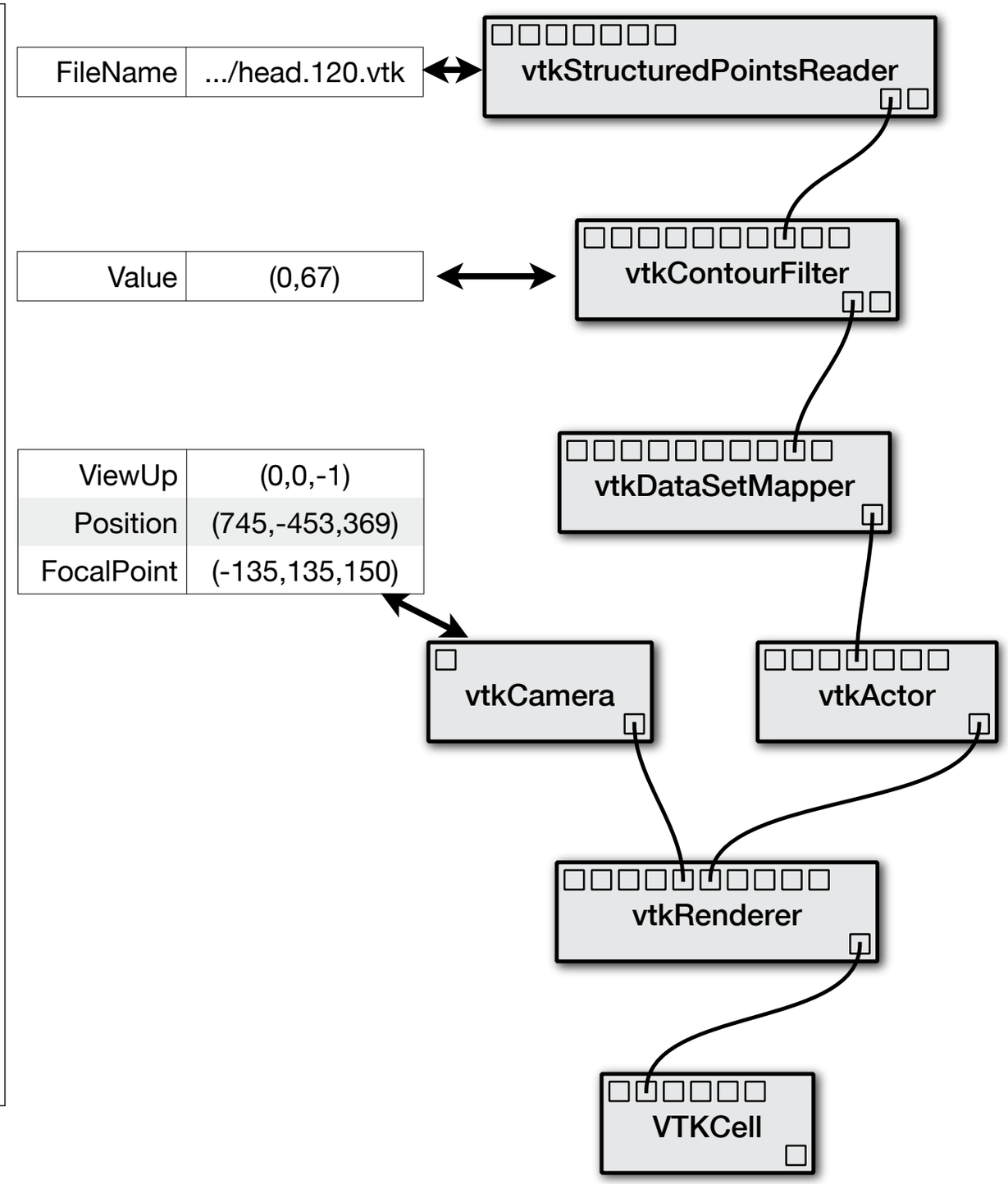
OpenMP OpenCL

Abstract & Structure Computations

[Paul Klee]

Structured & Abstracted Code

```
data = vtk.vtkStructuredPointsReader()
data.SetFileName(../examples/data/head.120.vtk)

contour = vtk.vtkContourFilter()
contour.SetInput(data.GetOutput())
contour.SetValue(0, 67)

mapper = vtk.vtkPolyDataMapper()
mapper.SetInput(contour.GetOutput())
mapper.ScalarVisibilityOff()

actor = vtk.vtkActor()
actor.SetMapper(mapper)

cam = vtk.vtkCamera()
cam.SetViewUp(0,0,-1)
cam.SetPosition(745,-453,369)
cam.SetFocalPoint(135,135,150)
cam.ComputeViewPlaneNormal()

ren = vtk.vtkRenderer()
ren.AddActor(actor)
ren.SetActiveCamera(cam)
ren.ResetCamera()
renwin = vtk.vtkRenderWindow()
renwin.AddRenderer(ren)

style = vtk.vtkInteractorStyleTrackballCamera()
iren = vtk.vtkRenderWindowInteractor()
iren.SetRenderWindow(renwin)
iren.SetInteractorStyle(style)
iren.Initialize()
iren.Start()
```



Ensure Quality, Reliability, & Trustworthiness

[Pentium Chip with FDIV Bug, Photo by Konstantin Lanzet, CC BY-SA 3.0]

Provenance in Art

Rembrandt van Rijn
Dutch, 1606 - 1669
***Self-Portrait*, 1659**
oil on canvas
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.
- Mellon Educational and Charitable Trust, The A.W.
- Montagu, and 4th Earl of Cardigan, George, 3rd Duke of

[National Gallery of Art]

Provenance in Art

Rembrandt van Rijn
Dutch, 1606 - 1669
***Self-Portrait*, 1659**
oil on canvas
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.

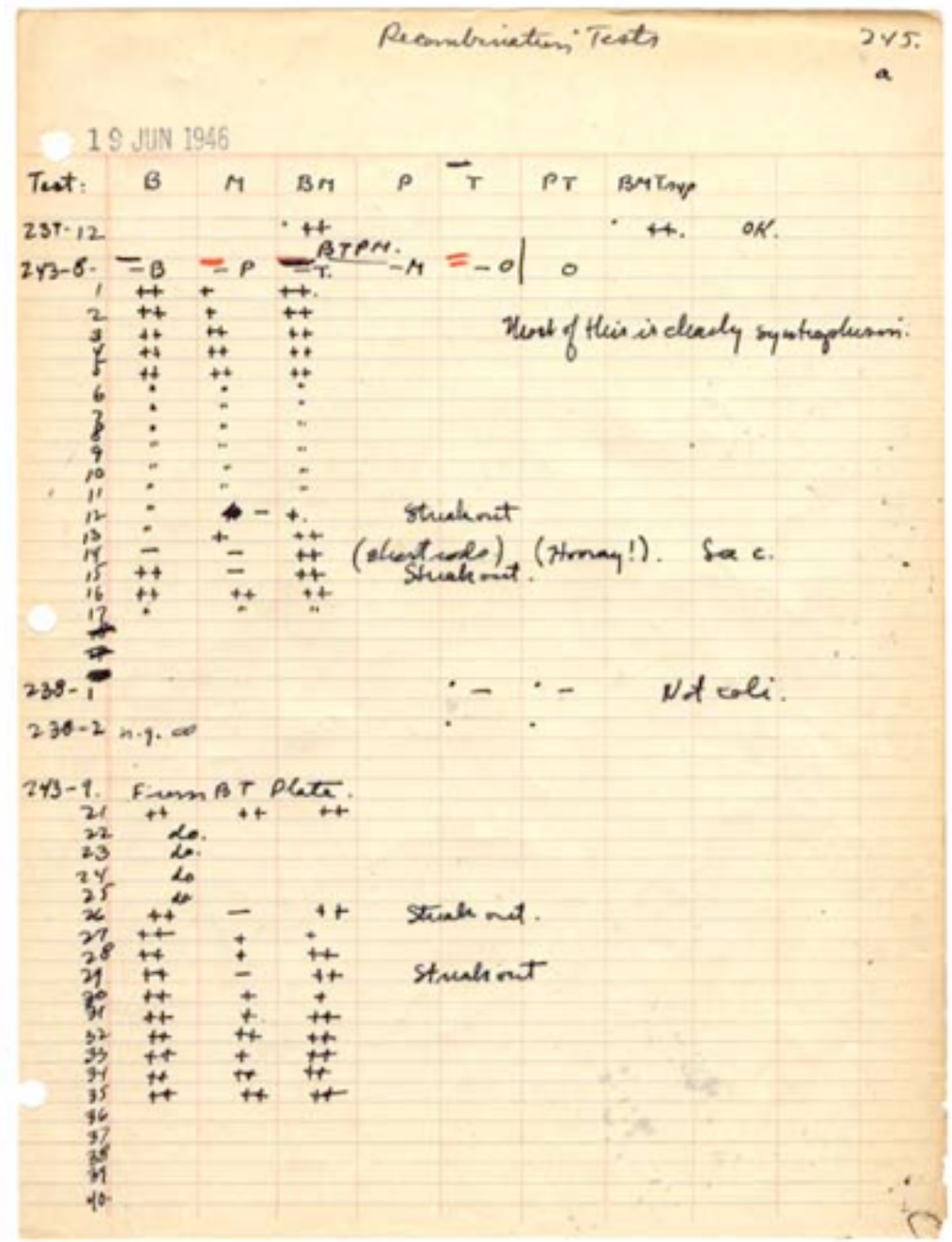
Associated Names

- Buccleuch, Henry, 3rd Duke of
- Buccleuch, John Charles, 7th Duke of
- Colnaghi & Co., Ltd., P. & D.
- Knoedler & Company, M.
- Mellon, Andrew W.
- Mellon Educational and Charitable Trust, The A.W.
- Montagu, and 4th Earl of Cardigan, George, 3rd Duke of

[National Gallery of Art]

Provenance in Science

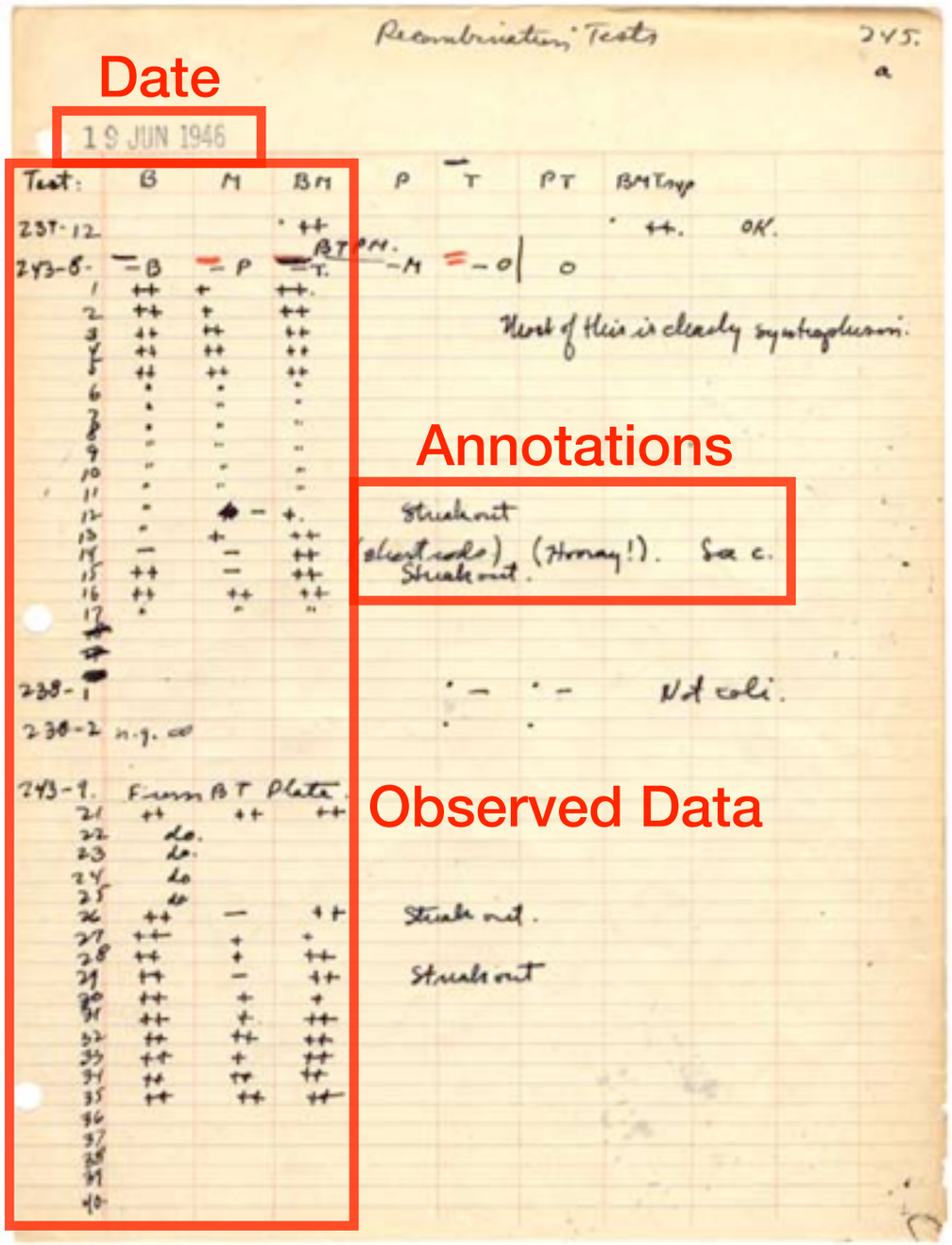
- Provenance: the **lineage** of data, a computation, or a visualization
- **Provenance is as (or more) important as the result!**
- Old solution:
  - Lab notebooks
- New problems:
  - Large volumes of data
  - Complex analyses
  - Writing notes doesn't scale



[DNA Recombination, Lederberg]

Provenance in Science

- Provenance: the **lineage** of data, a computation, or a visualization
- **Provenance is as (or more) important as the result!**
- Old solution:
  - Lab notebooks
- New problems:
  - Large volumes of data
  - Complex analyses
  - Writing notes doesn't scale

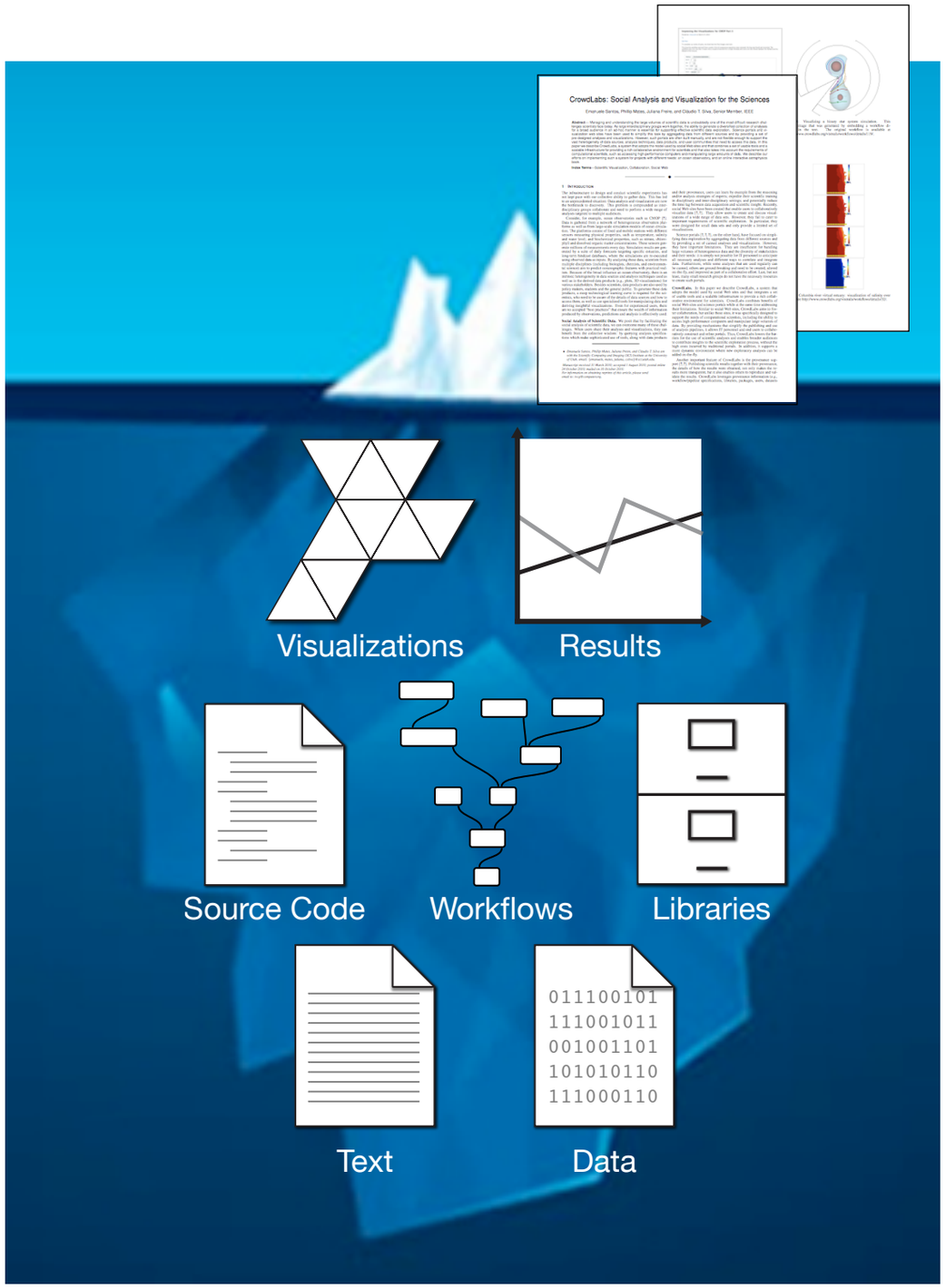


[DNA Recombination, Lederberg]

Reproducibility

[Rube Goldberg Contest, Photo by Argonne National Laboratory, CC BY-NC-SA 2.0]

Reproducibility

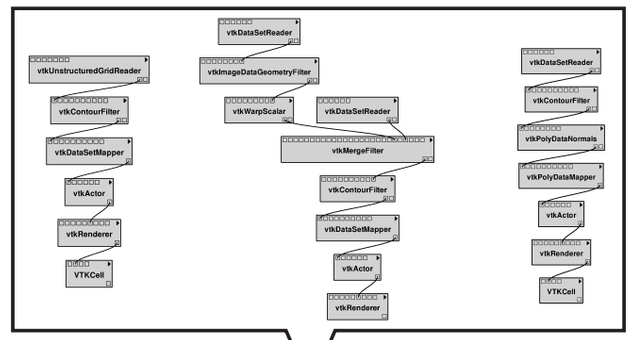


- Capture **how** results were achieved
- Includes **many** different items
- Improve **collaboration** and sharing

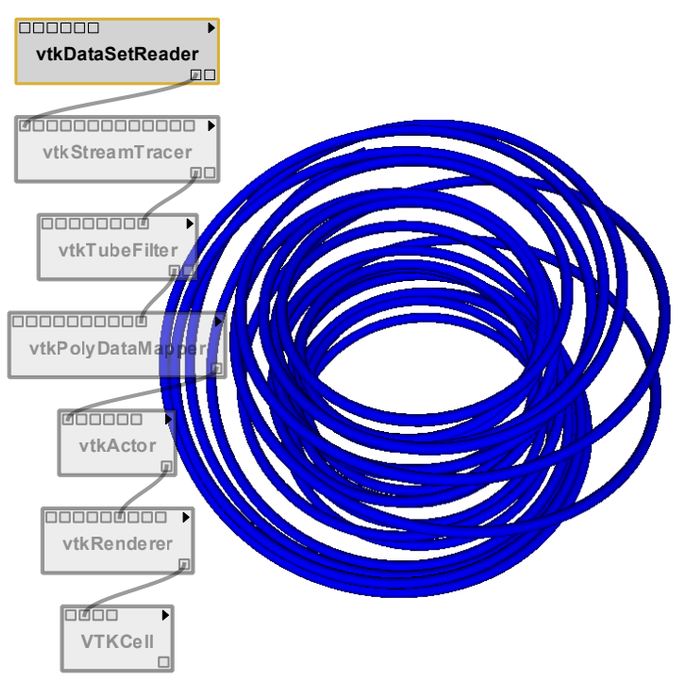
Reuse Past Work

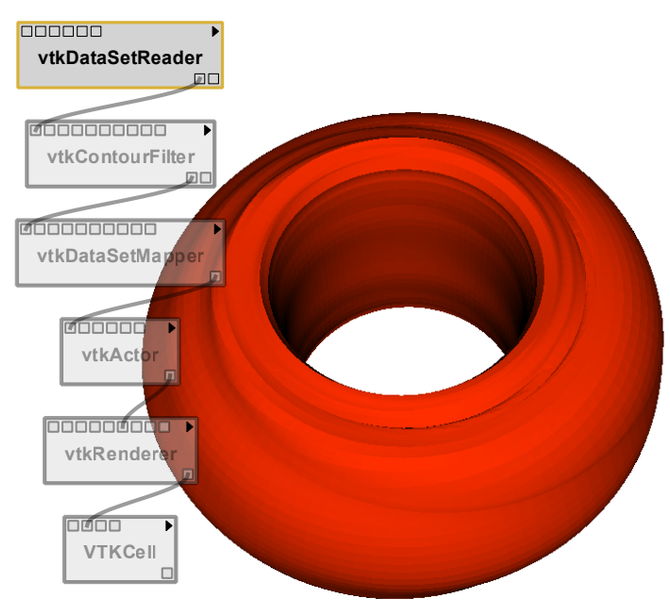
[PETCO Park, Photo by Edward O'Connor, CC BY-SA 2.0]

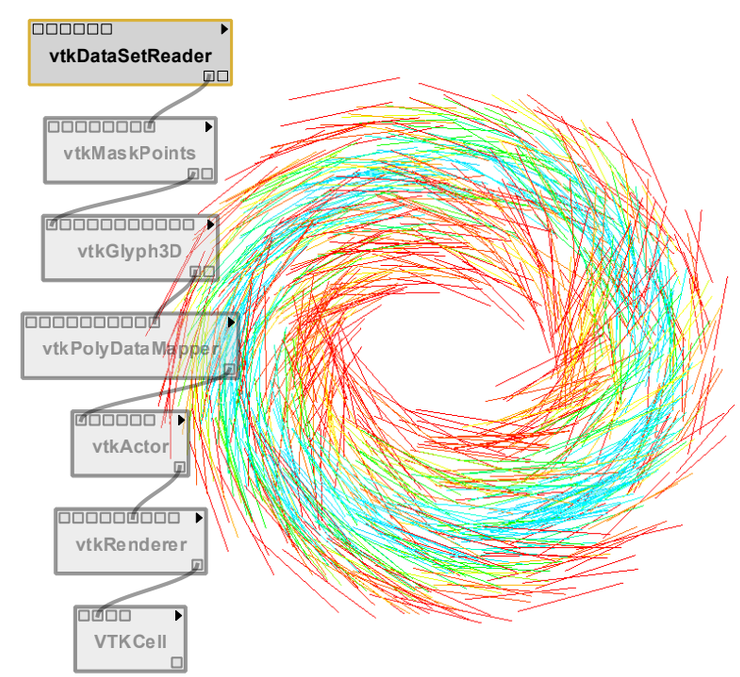
Reuse Past Work



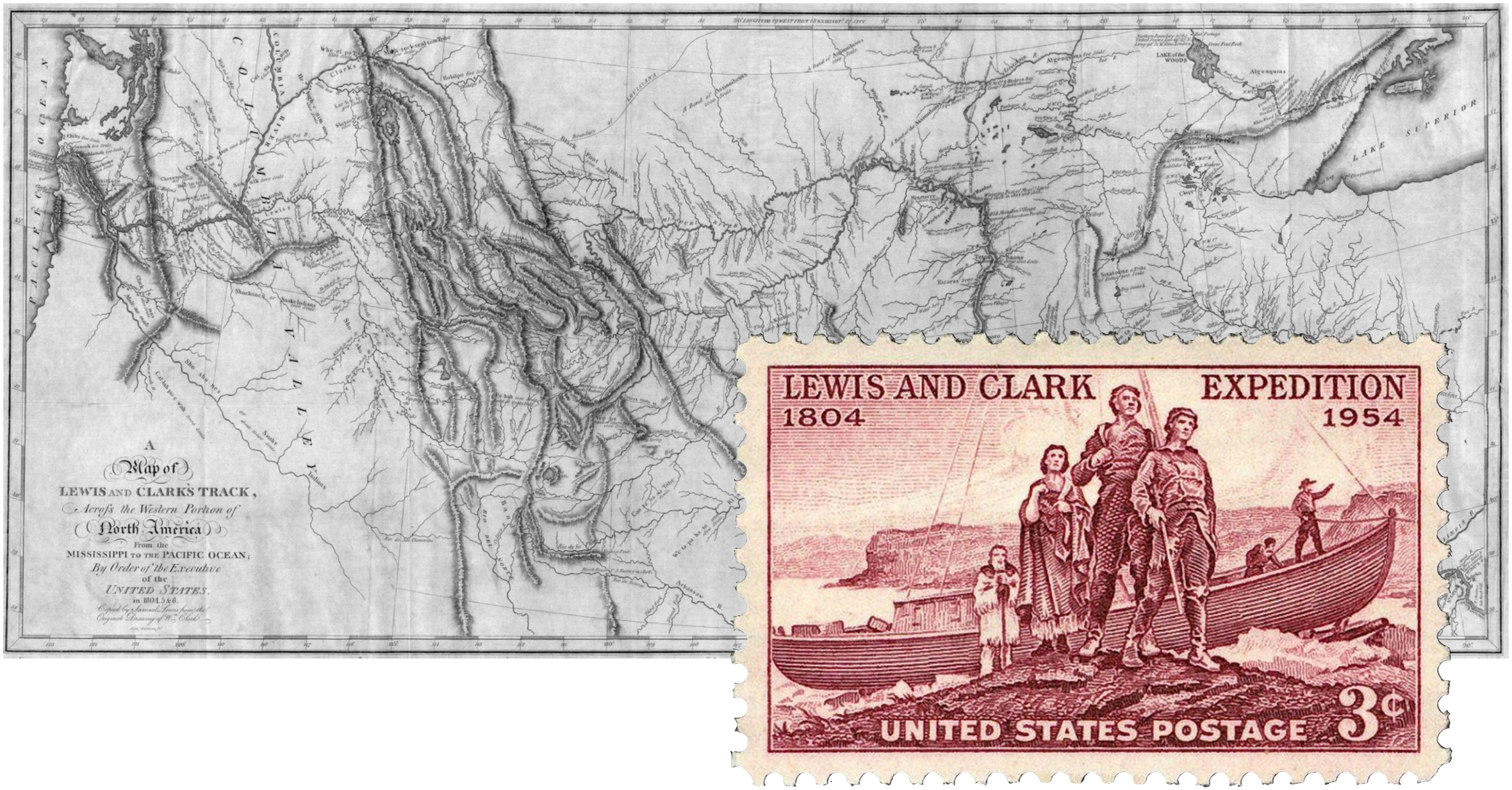
Database of Workflows



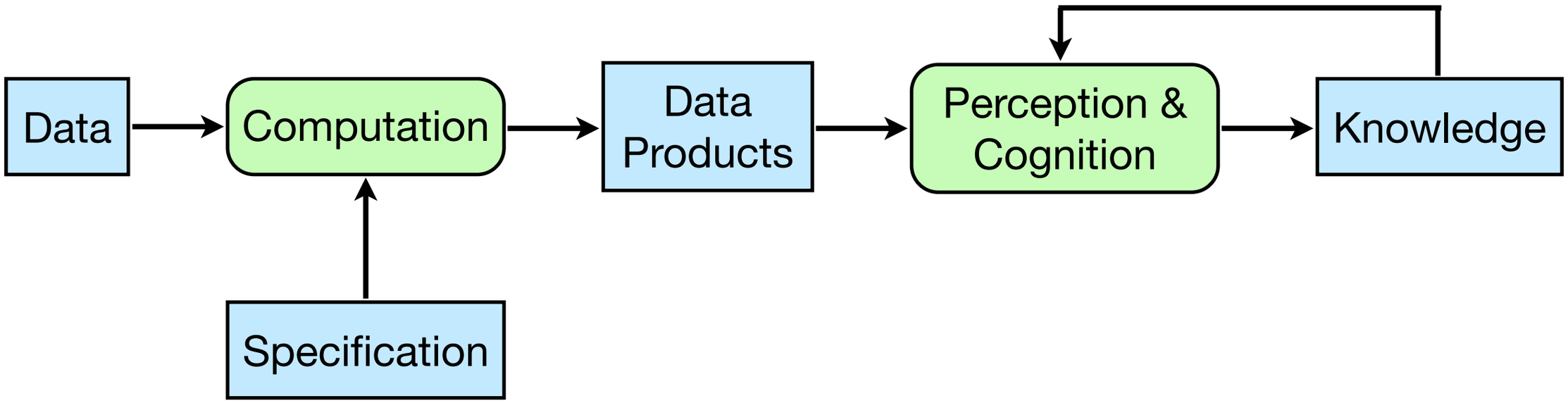




Unencumbered Exploration & Learning

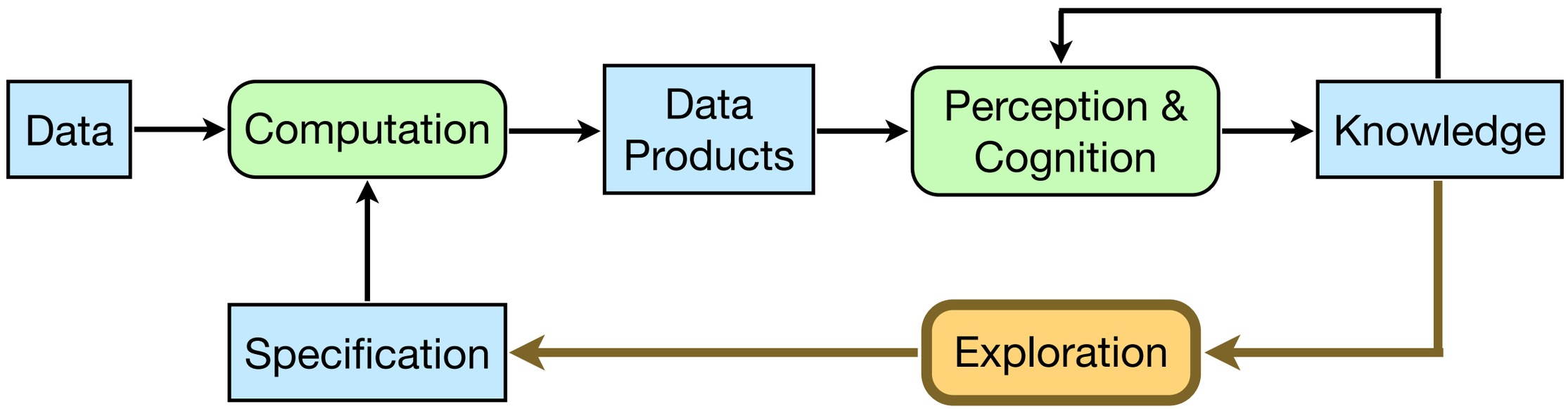


Data Exploration



[Modified from Van Wijk, Vis 2005]

Data Exploration



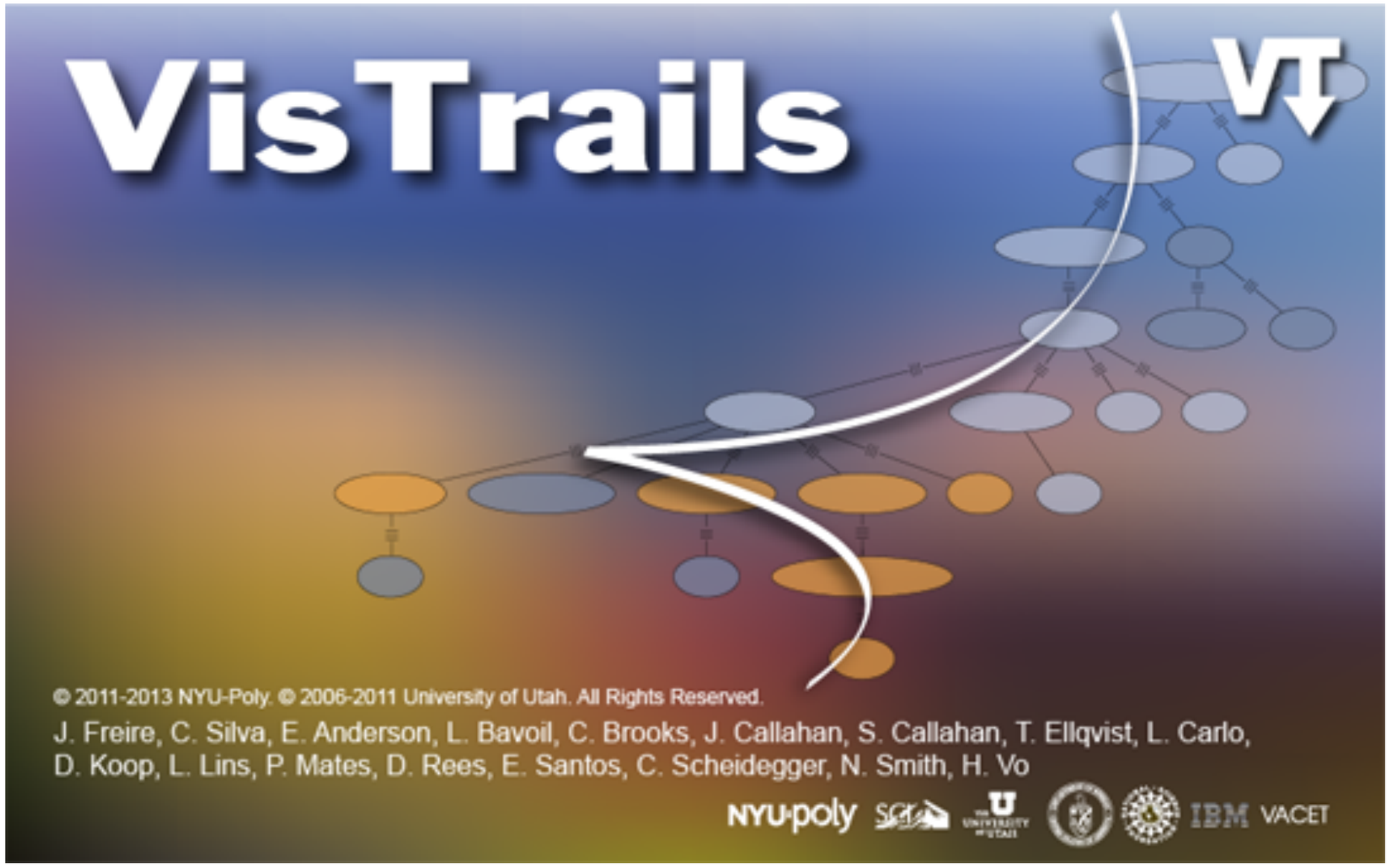
[Modified from Van Wijk, Vis 2005]

- Data analysis and visualization are **iterative** processes
- In **exploratory** tasks, **change is the norm**!

Exploration and Creativity Support

- Reasoning is key to the exploratory processes
- *"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
- Need external aids—tools to facilitate this process
  - Creativity support tools [Ben Shneiderman]
- Need aid from people—collaboration

VisTrails


VisTrails
VT
© 2011-2013 NYU-Poly. © 2006-2011 University of Utah. All Rights Reserved.
J. Freire, C. Silva, E. Anderson, L. Bavoil, C. Brooks, J. Callahan, S. Callahan, T. Ellqvist, L. Carlo,
D. Koop, L. Lins, P. Mates, D. Rees, E. Santos, C. Scheidegger, N. Smith, H. Vo
NYU·poly
IBM
VACET

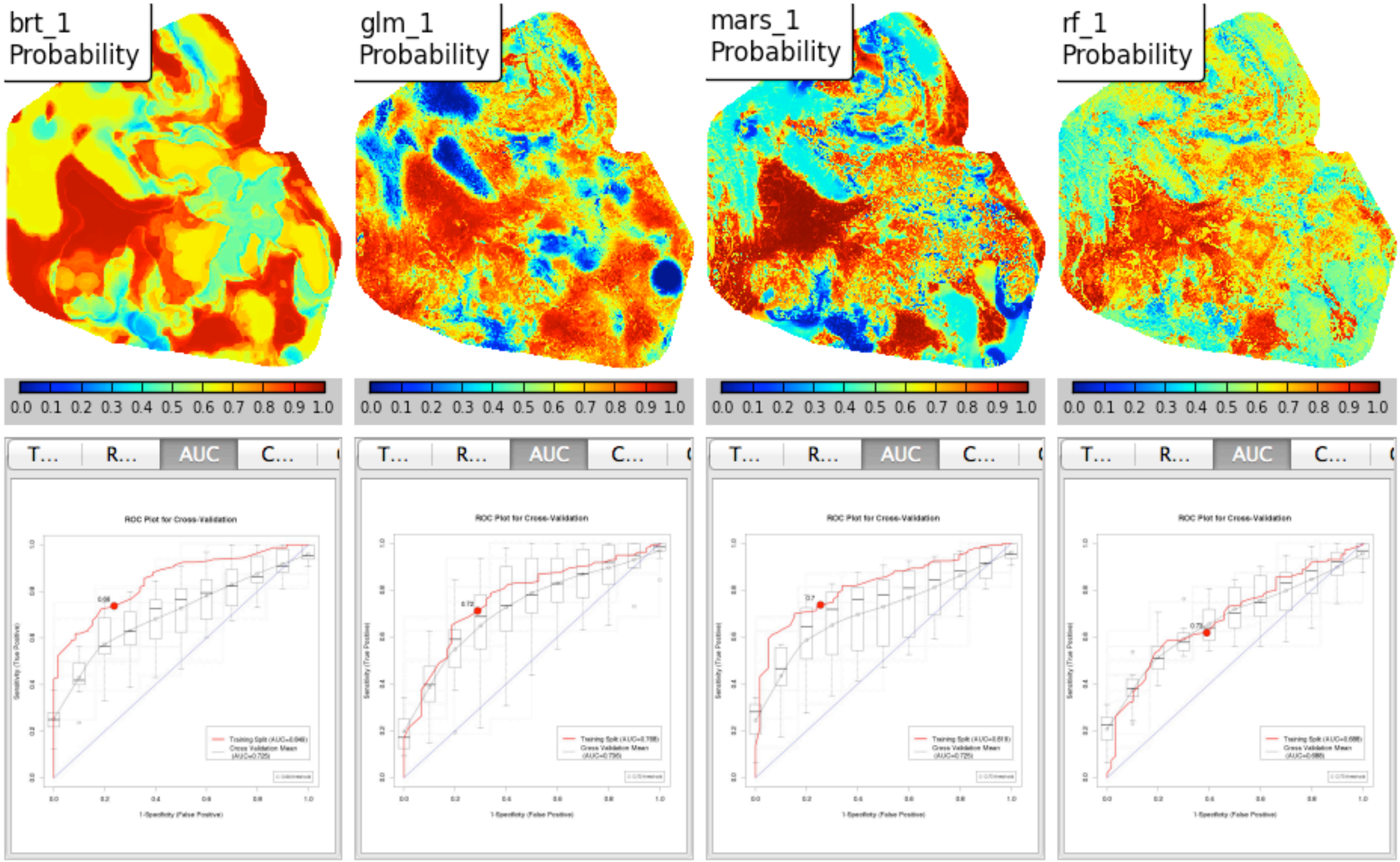
VisTrails

- Comprehensive **provenance infrastructure** for computational tasks
- Focus on **exploratory** tasks such as simulation, visualization, and data analysis
- **Transparently** tracks provenance of the discovery process—from data acquisition to visualization
  - The **trail** followed as users generate and test hypotheses
  - Users can refer back to any point along this trail at any time
- Leverage provenance to **streamline exploration**
- Focus on **usability**—build tools for scientists

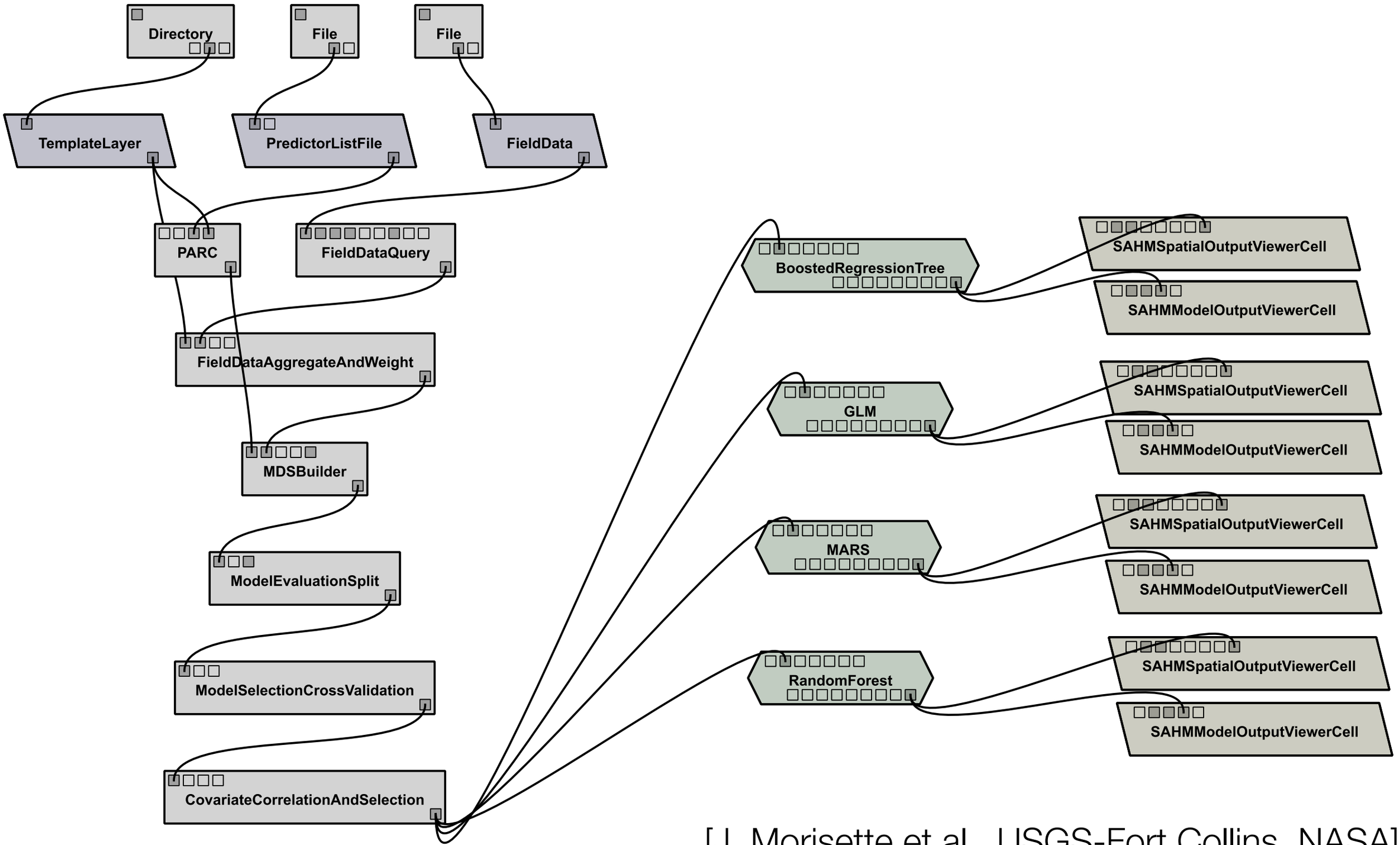
VisTrails

- Open-source, freely downloadable system (www.vistrails.org)
  - Also on GitHub (github.com/vistrails) and PyPI
- Multi-platform: users on Mac, Linux, and Windows
- Python code and uses PyQt and Qt for the interface
- Over 35,000 downloads
- User's guide, wiki, and mailing list
- Many users in different disciplines and countries:
  - Visualizing environmental simulations (CMOP STC)
  - Simulation for solid, fluid and structural mechanics (Galileo Network, UFRJ Brazil)
  - Quantum physics simulations (ALPS, ETH Zurich)
  - Climate analysis (UV-CDAT, LLNL)
  - Habitat modeling (USGS)
  - Open Wildland Fire Modeling (U. Colorado, NCAR)
  - High-energy physics (LEPP, Cornell)
  - Cosmology simulations (LANL)
  - Using tms for improving memory (Pyschiatry, U. Utah)
  - eBird (Cornell, NSF DataONE)
  - Astrophysical Systems (LSU)
  - NIH NBCR (UCSD)
  - Pervasive Technology Labs (Indiana University)
  - Linköping University
  - University of North Carolina, Chapel Hill
  - UTEP

SAHM: Modeling the Spread of Invasive Species

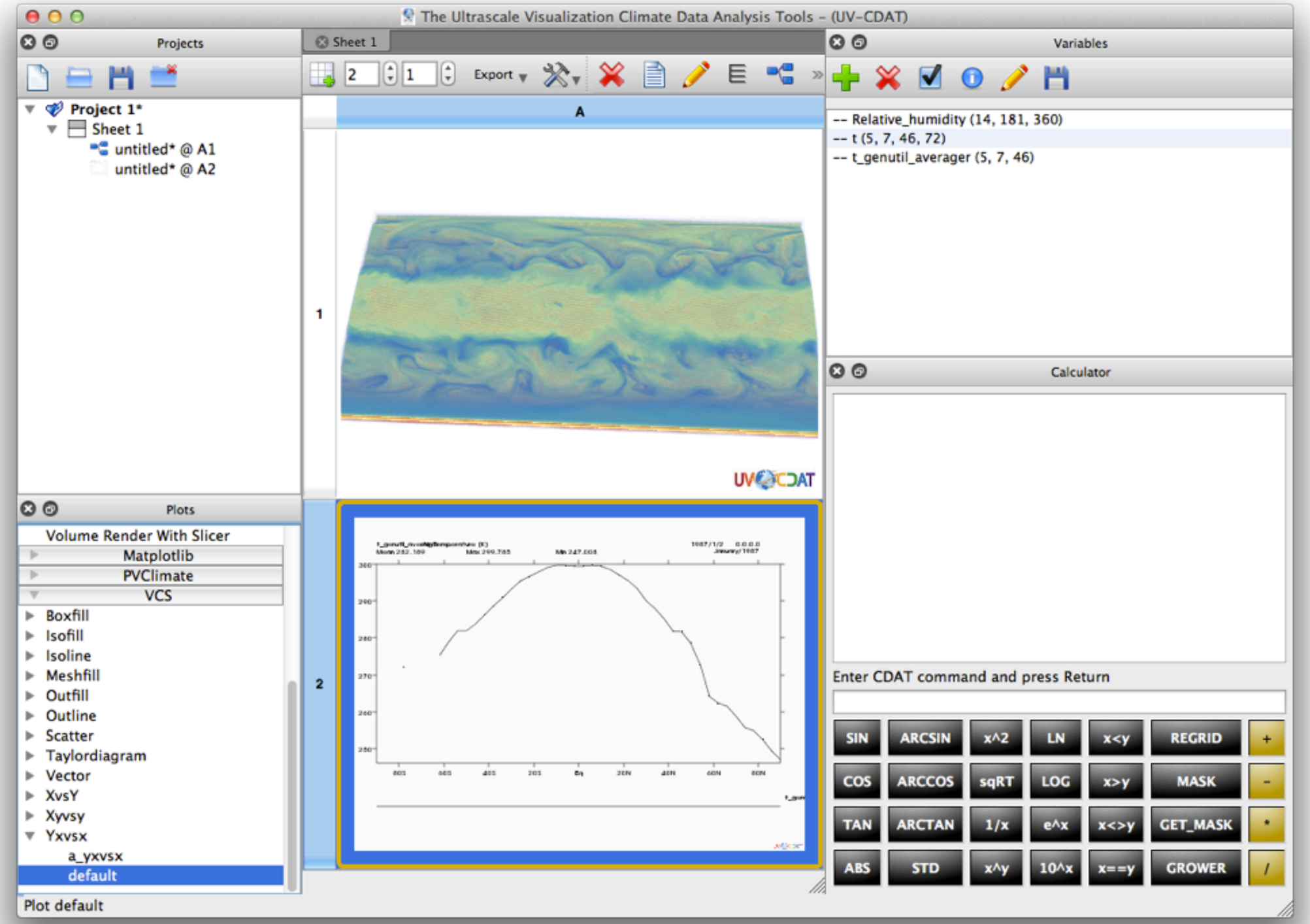


SAHM: Modeling the Spread of Invasive Species

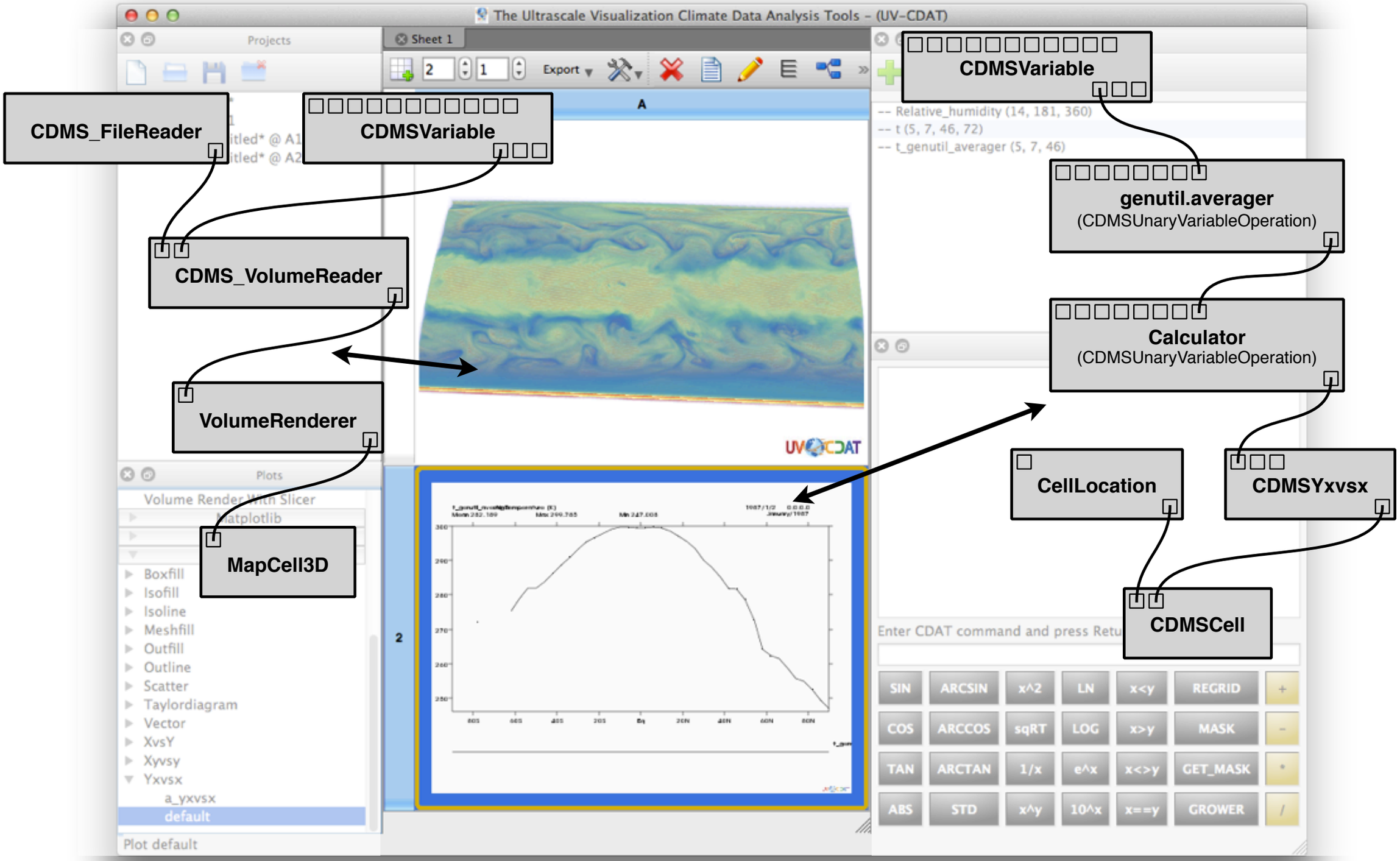


[J. Morisette et al., USGS-Fort Collins, NASA]

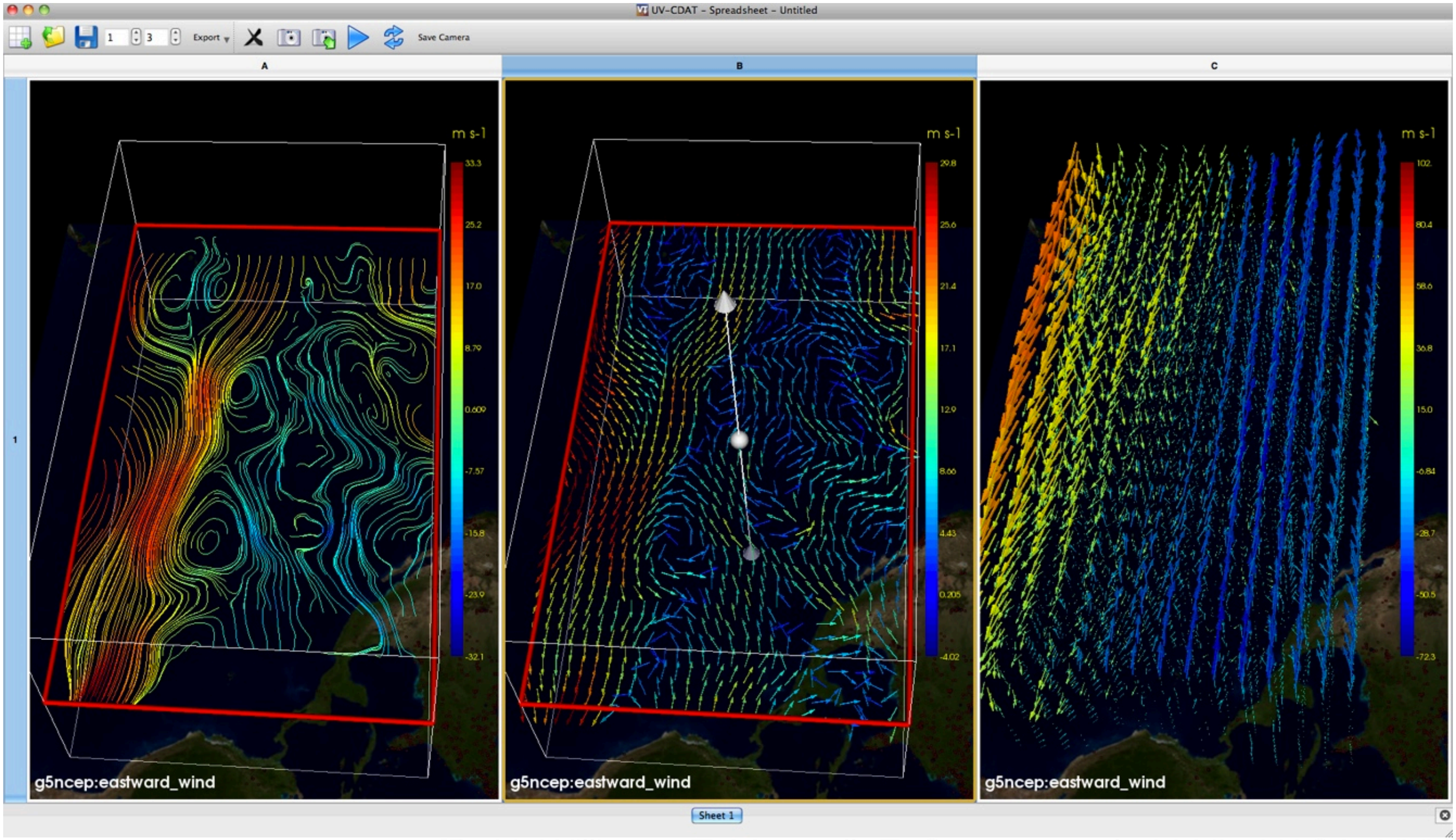
UV-CDAT: Climate Science



UV-CDAT: Climate Science

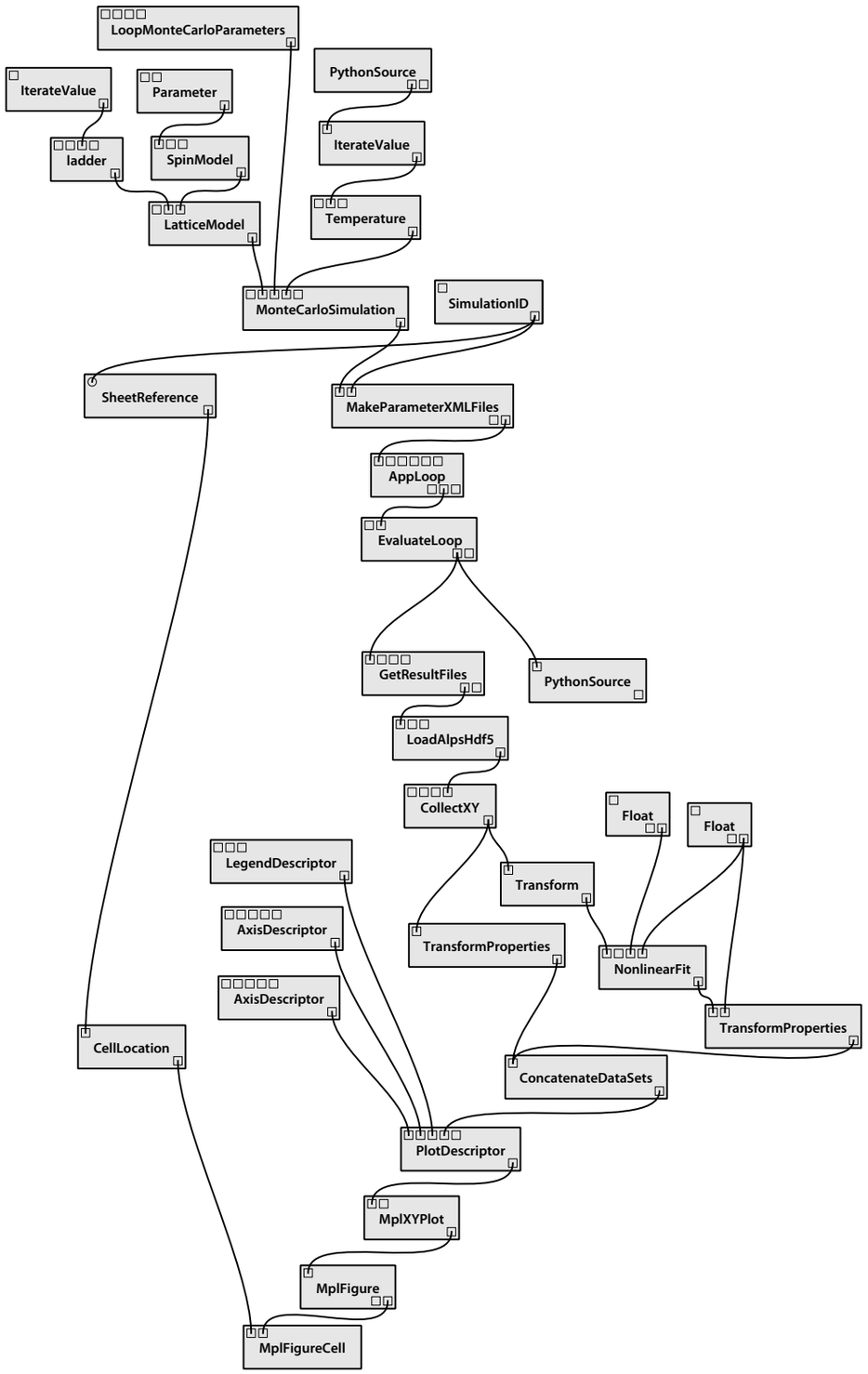


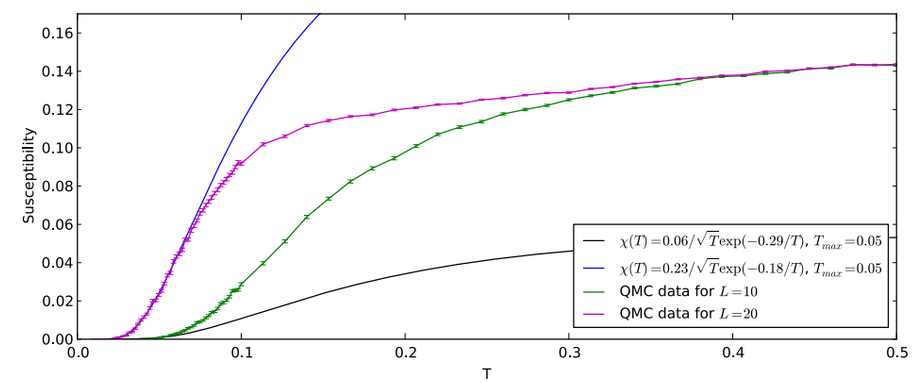
UV-CDAT: Climate Science

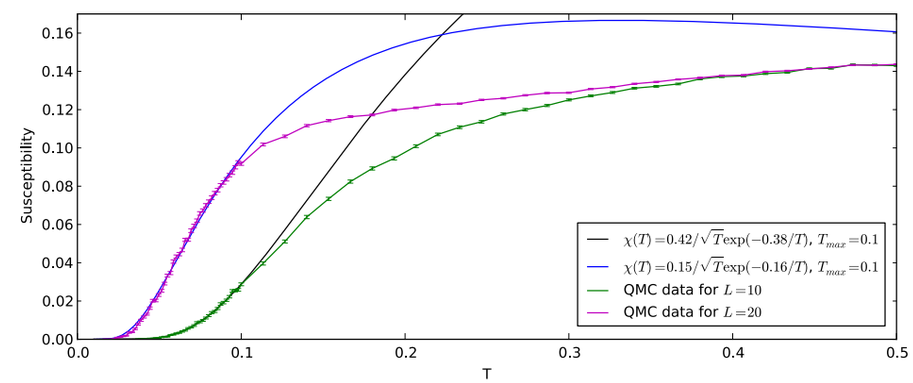


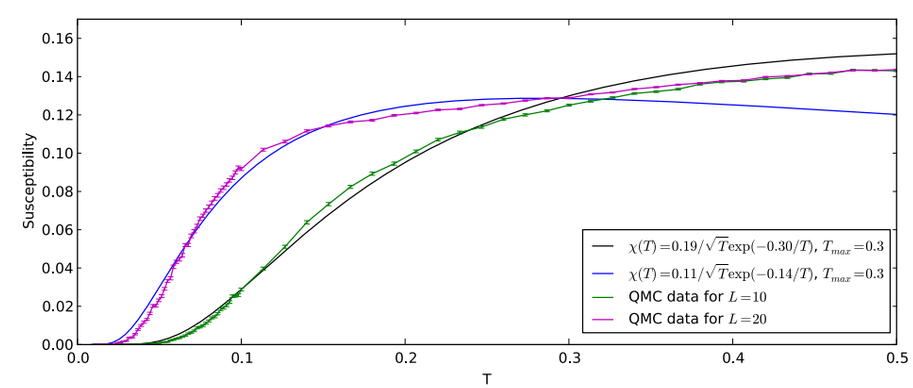
[D. N. Williams, T. Maxwell, E. Santos, et al., LLNL, NASA, NYU]

ALPS: Large Quantum Simulations





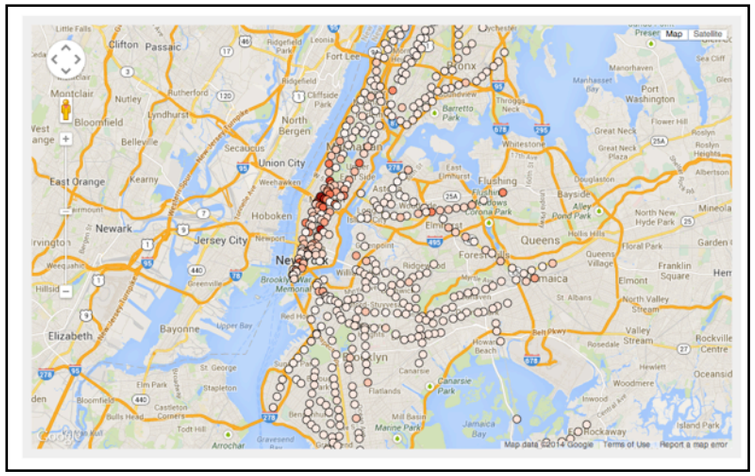


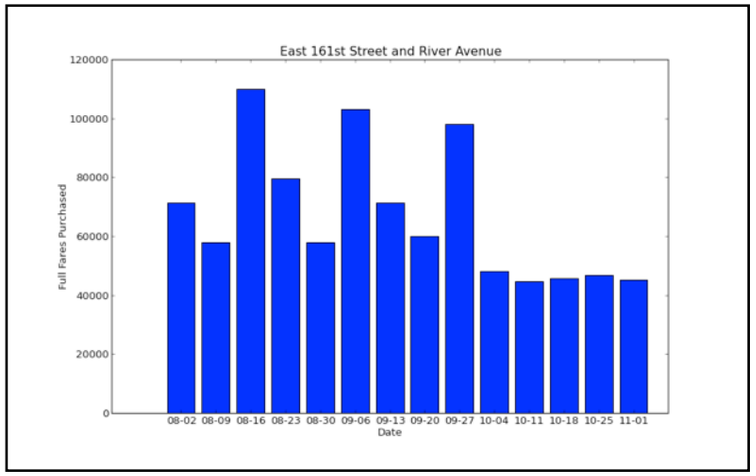


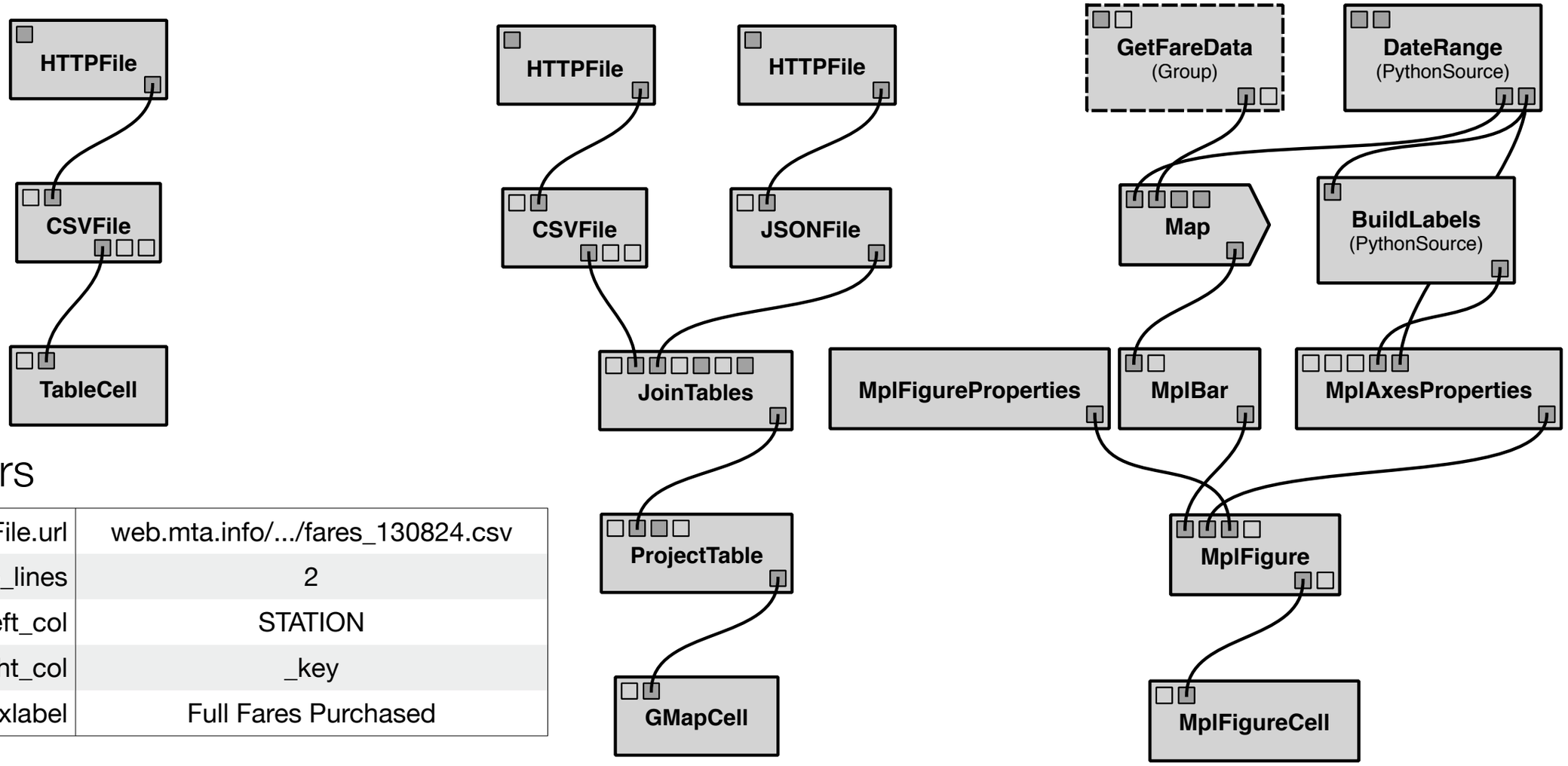
[M. Troyer et al., ETH-Zurich]

Example: MTA Fare Data

	REMOTE	STATION	FF	SEN/DIS	7-D AFAS UNL	D AFAS/RMF I	JOINT RR TKT	7-D UNL	30-D UNL
1	R011	42ND STREET & 8TH AVENUE	00228985	00008471	00000441	00001455	00000134	00033341	00071255
2	R170	14TH STREET-UNION SQUARE	00224603	00011051	00000827	00003026	00000660	00089367	00199841
3	R046	42ND STREET & GRAND CENTRAL	00207758	00007908	00000323	00001183	00003001	00040759	00096613
4	R012	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00001279	00003622	00035527	00067483
5	R293	34TH STREET - PENN STATION	00168768	00006155	00000523	00001065	00005031	00030645	00054376
6	R033	42ND STREET/TIMES SQUARE	00159382	00005945	00000378	00001205	00000690	00058931	00078644
7	R022	34TH STREET & 6TH AVENUE	00156008	00006276	00000487	00001543	00000712	00058910	00110466
8	R084	59TH STREET/COLUMBUS CIRCLE	00155262	00009484	00000589	00002071	00000542	00053397	00113966
9	R020	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000384	00001159	00000723	00037978	00090745
10	R179	86TH STREET-LEXINGTON AVE	00142169	00010367	00000470	00001839	00000271	00050328	00125250
11	R023	34TH STREET & 6TH AVENUE	00134052	00005005	00000348	00001112	00000649	00031531	00075040
12	R029	PARK PLACE	00121614	00004311	00000287	00000931	00000792	00025404	00065362
13	R047	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000704	00001241	00022808	00068216
14	R031	34TH STREET & 7TH AVENUE	00095076	00003990	00000232	00000727	00001459	00024284	00038671
15	R017	LEXINGTON AVENUE	00094655	00004688	00000190	00000833	00000754	00020018	00055066
16	R175	8TH AVENUE-14TH STREET	00094313	00003907	00000286	00001144	00000256	00038272	00074661
17	R057	BARCLAYS CENTER	00093804	00004204	00000454	00001386	00001491	00039113	00068119
18	R138	WEST 4TH ST-WASHINGTON SQ	00093562	00004677	00000251	00000965	00000127	00031628	00074458



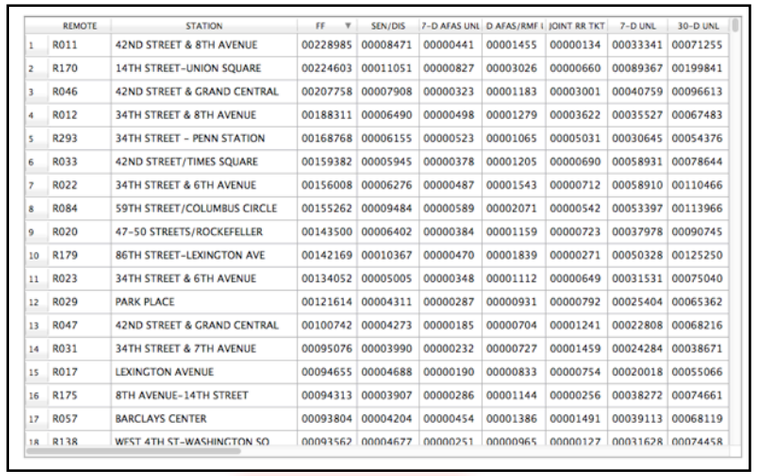




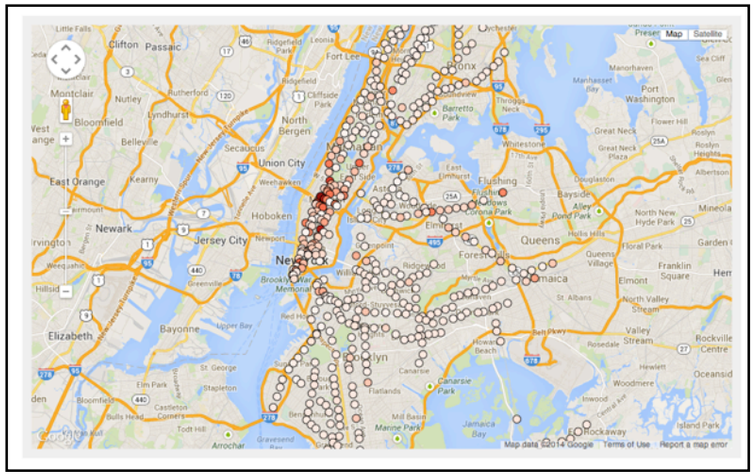
Parameters

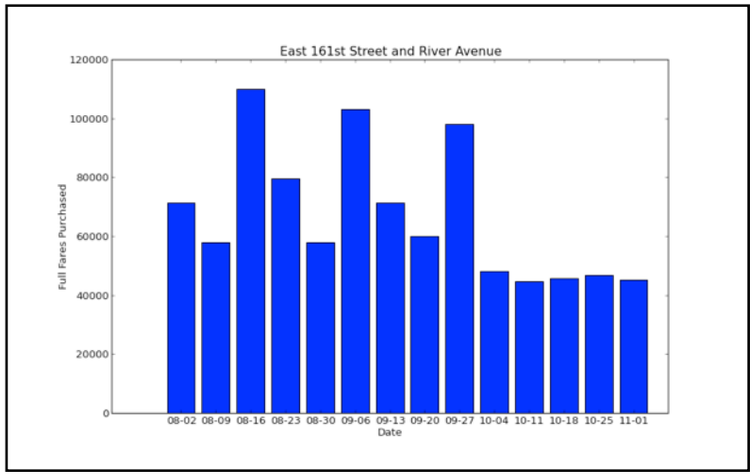
HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased

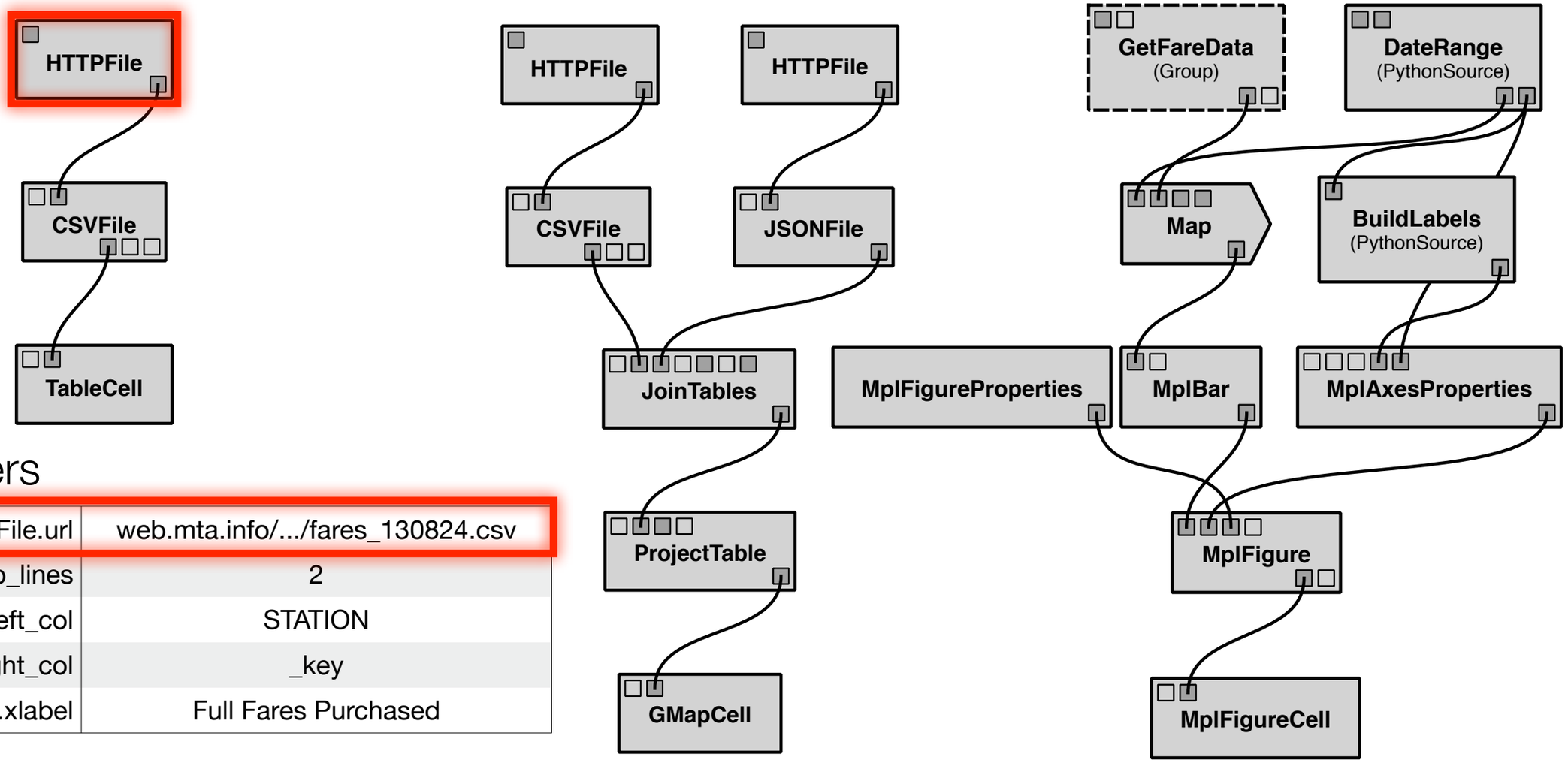
Example: MTA Fare Data



	REMOTE	STATION	FF	SEN/DIS	7-D AFAS UNL	D AFAS/RMF I	JOINT RR TKT	7-D UNL	30-D UNL
1	R011	42ND STREET & 8TH AVENUE	00228985	00008471	00000441	00001455	00000134	00033341	00071255
2	R170	14TH STREET-UNION SQUARE	00224603	00011051	00000827	00003026	00000660	00089367	00199841
3	R046	42ND STREET & GRAND CENTRAL	00207758	00007908	00000323	00001183	00003001	00040759	00096613
4	R012	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00001279	00003622	00035527	00067483
5	R293	34TH STREET - PENN STATION	00168768	00006155	00000523	00001065	00005031	00030645	00054376
6	R033	42ND STREET/TIMES SQUARE	00159382	00005945	00000378	00001205	00000690	00058931	00078644
7	R022	34TH STREET & 6TH AVENUE	00156008	00006276	00000487	00001543	00000712	00058910	00110466
8	R084	59TH STREET/COLUMBUS CIRCLE	00155262	00009484	00000589	00002071	00000542	00053397	00113966
9	R020	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000384	00001159	00000723	00037978	00090745
10	R179	86TH STREET-LEXINGTON AVE	00142169	00010367	00000470	00001839	00000271	00050328	00125250
11	R023	34TH STREET & 6TH AVENUE	00134052	00005005	00000348	00001112	00000649	00031531	00075040
12	R029	PARK PLACE	00121614	00004311	00000287	00000931	00000792	00025404	00065362
13	R047	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000704	00001241	00022808	00068216
14	R031	34TH STREET & 7TH AVENUE	00095076	00003990	00000232	00000727	00001459	00024284	00038671
15	R017	LEXINGTON AVENUE	00094655	00004688	00000190	00000833	00000754	00020018	00055066
16	R175	8TH AVENUE-14TH STREET	00094313	00003907	00000286	00001144	00000256	00038272	00074661
17	R057	BARCLAYS CENTER	00093804	00004204	00000454	00001386	00001491	00039113	00068119
18	R138	WEST 4TH ST-WASHINGTON SQ	00093562	00004677	00000251	00000965	00000127	00031628	00074458





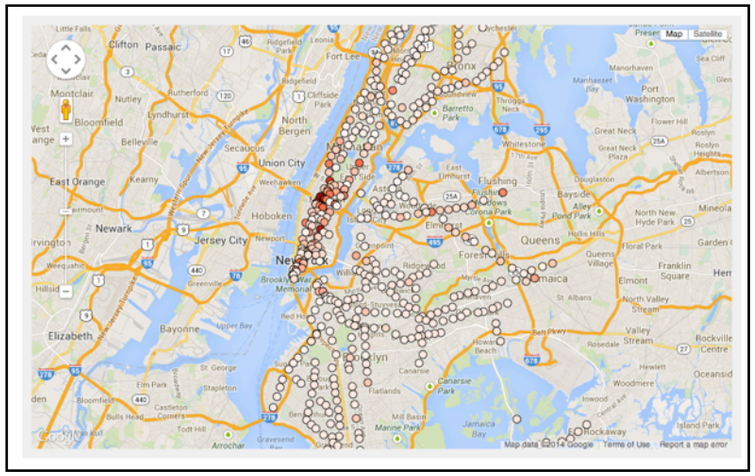


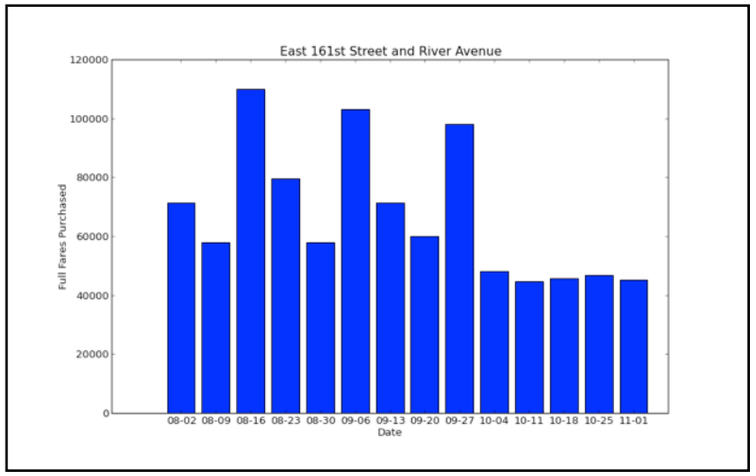
Parameters

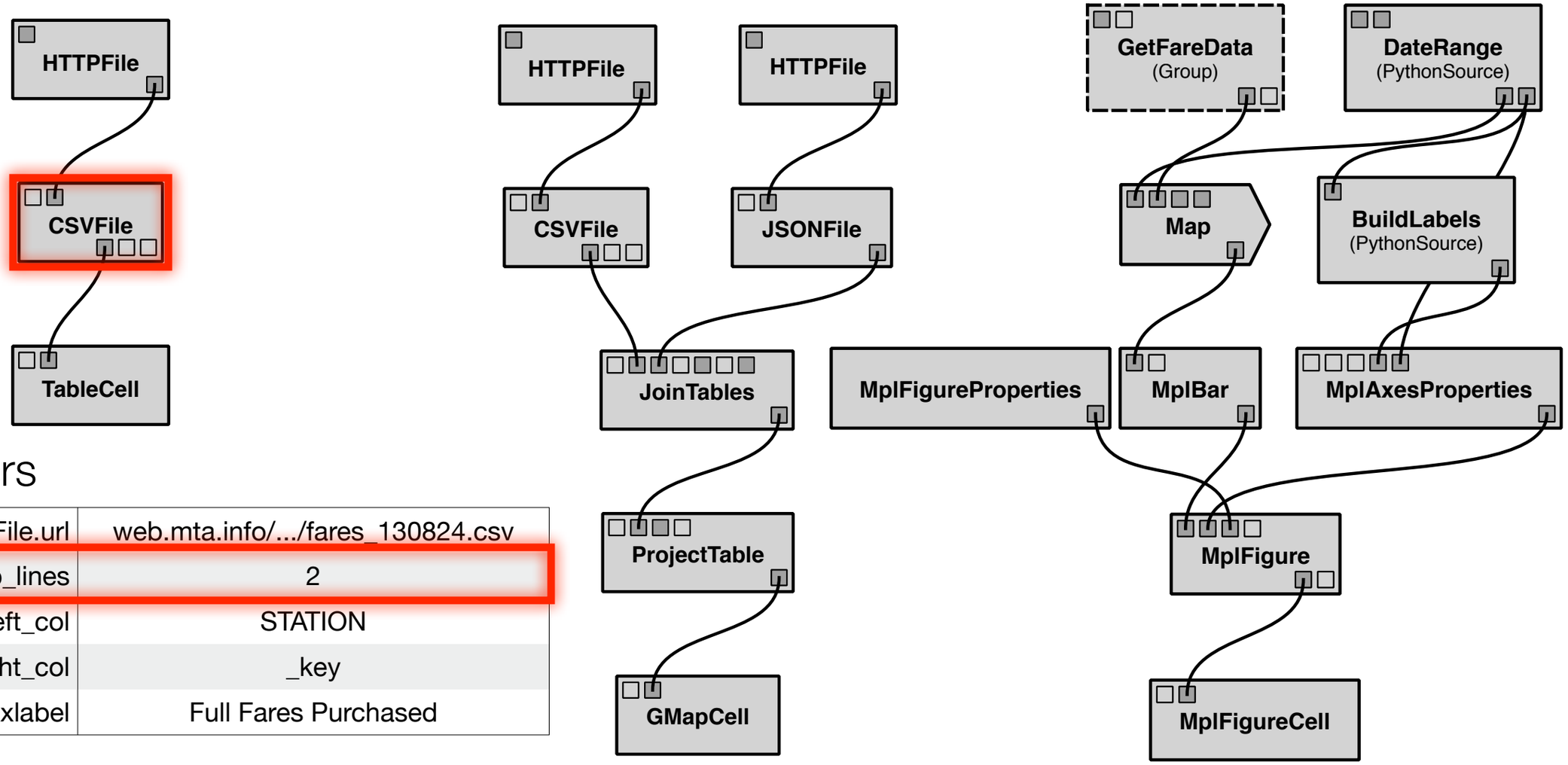
HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased

Example: MTA Fare Data

	REMOTE	STATION	FF	SEN/DIS	7-D AFAS UNL	D AFAS/RMF I	JOINT RR TKT	7-D UNL	30-D UNL
1	R011	42ND STREET & 8TH AVENUE	00228985	00008471	00000441	00001455	00000134	00033341	00071255
2	R170	14TH STREET-UNION SQUARE	00224603	00011051	00000827	00003026	00000660	00089367	00199841
3	R046	42ND STREET & GRAND CENTRAL	00207758	00007908	00000323	00001183	00003001	00040759	00096613
4	R012	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00001279	00003622	00035527	00067483
5	R293	34TH STREET - PENN STATION	00168768	00006155	00000523	00001065	00005031	00030645	00054376
6	R033	42ND STREET/TIMES SQUARE	00159382	00005945	00000378	00001205	00000690	00058931	00078644
7	R022	34TH STREET & 6TH AVENUE	00156008	00006276	00000487	00001543	00000712	00058910	00110466
8	R084	59TH STREET/COLUMBUS CIRCLE	00155262	00009484	00000589	00002071	00000542	00053397	00113966
9	R020	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000384	00001159	00000723	00037978	00090745
10	R179	86TH STREET-LEXINGTON AVE	00142169	00010367	00000470	00001839	00000271	00050328	00125250
11	R023	34TH STREET & 6TH AVENUE	00134052	00005005	00000348	00001112	00000649	00031531	00075040
12	R029	PARK PLACE	00121614	00004311	00000287	00000931	00000792	00025404	00065362
13	R047	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000704	00001241	00022808	00068216
14	R031	34TH STREET & 7TH AVENUE	00095076	00003990	00000232	00000727	00001459	00024284	00038671
15	R017	LEXINGTON AVENUE	00094655	00004688	00000190	00000833	00000754	00020018	00055066
16	R175	8TH AVENUE-14TH STREET	00094313	00003907	00000286	00001144	00000256	00038272	00074661
17	R057	BARCLAYS CENTER	00093804	00004204	00000454	00001386	00001491	00039113	00068119
18	R138	WEST 4TH ST-WASHINGTON SQ	00093562	00004677	00000251	00000965	00000127	00031628	00074458



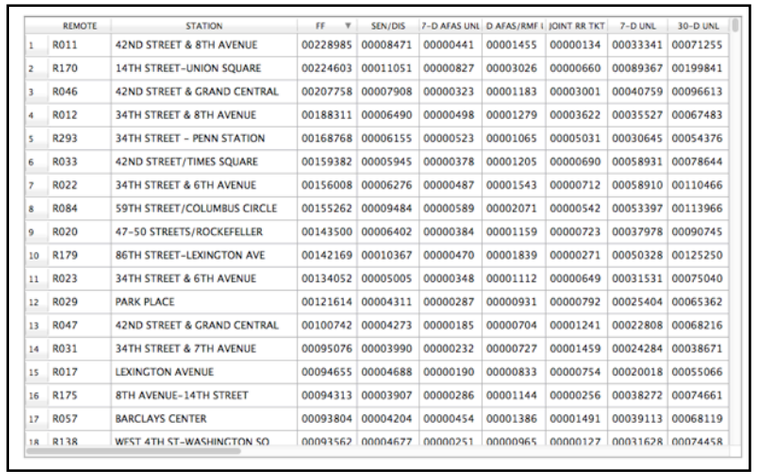




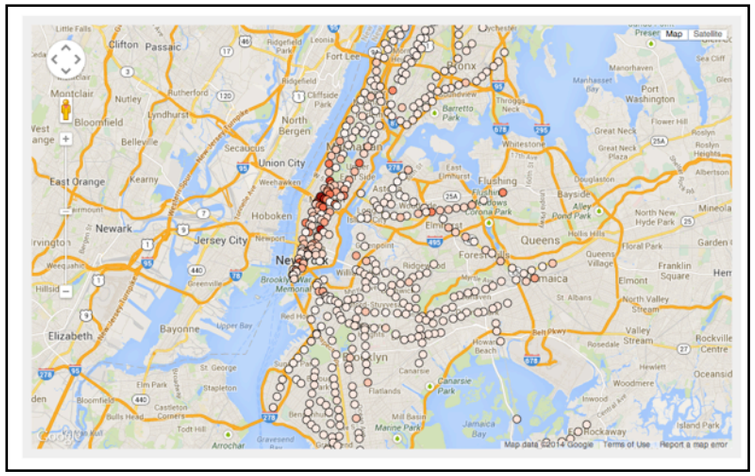
Parameters

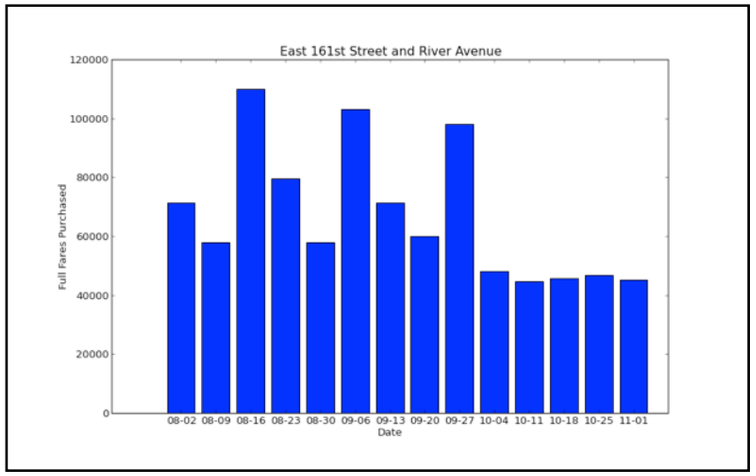
HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased

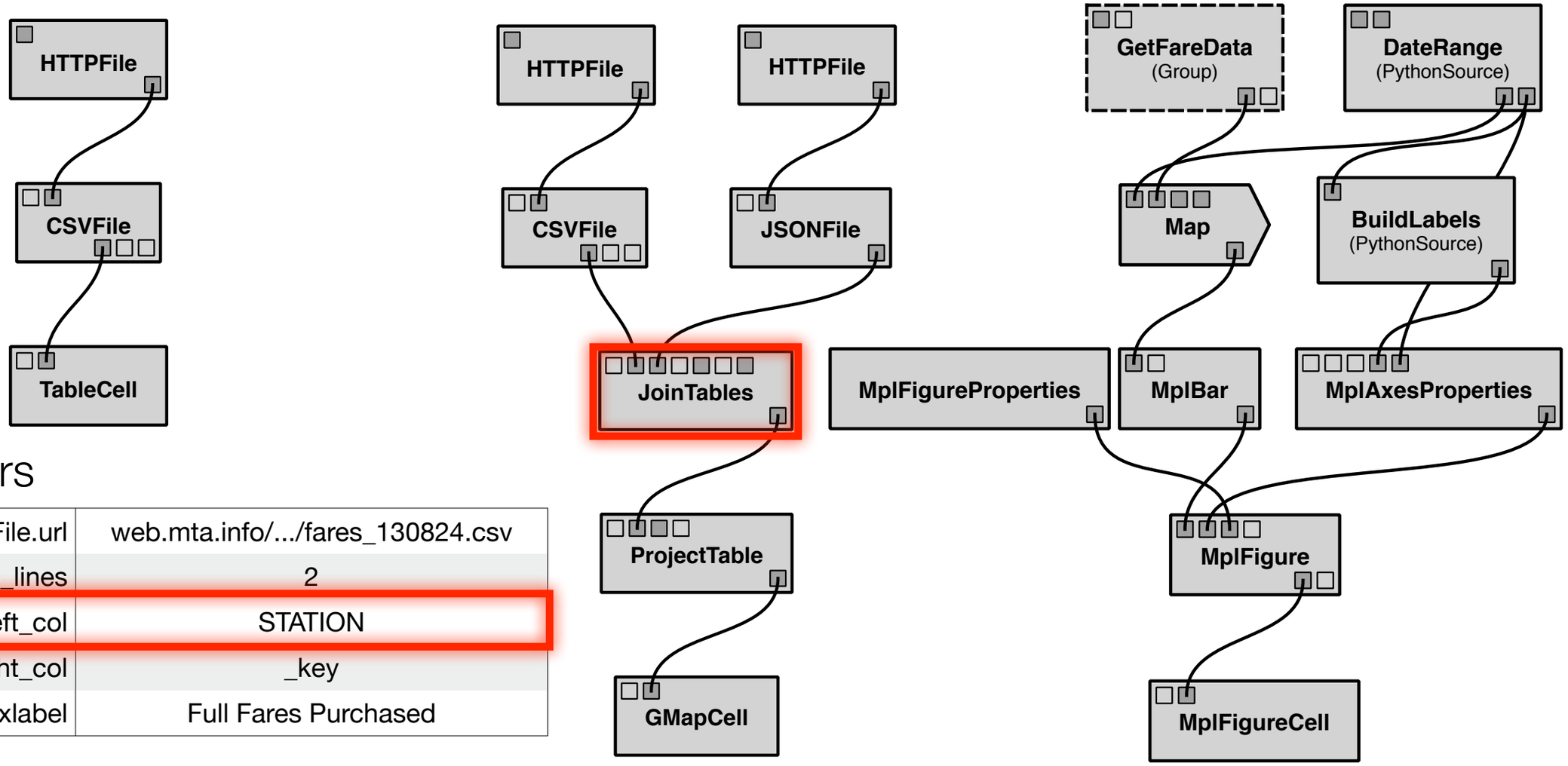
Example: MTA Fare Data



	REMOTE	STATION	FF	SEN/DIS	7-D AFAS UNL	D AFAS/RMF I	JOINT RR TKT	7-D UNL	30-D UNL
1	R011	42ND STREET & 8TH AVENUE	00228985	00008471	00000441	00001455	00000134	00033341	00071255
2	R170	14TH STREET-UNION SQUARE	00224603	00011051	00000827	00003026	00000660	00089367	00199841
3	R046	42ND STREET & GRAND CENTRAL	00207758	00007908	00000323	00001183	00003001	00040759	00096613
4	R012	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00001279	00003622	00035527	00067483
5	R293	34TH STREET - PENN STATION	00168768	00006155	00000523	00001065	00005031	00030645	00054376
6	R033	42ND STREET/TIMES SQUARE	00159382	00005945	00000378	00001205	00000690	00058931	00078644
7	R022	34TH STREET & 6TH AVENUE	00156008	00006276	00000487	00001543	00000712	00058910	00110466
8	R084	59TH STREET/COLUMBUS CIRCLE	00155262	00009484	00000589	00002071	00000542	00053397	00113966
9	R020	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000384	00001159	00000723	00037978	00090745
10	R179	86TH STREET-LEXINGTON AVE	00142169	00010367	00000470	00001839	00000271	00050328	00125250
11	R023	34TH STREET & 6TH AVENUE	00134052	00005005	00000348	00001112	00000649	00031531	00075040
12	R029	PARK PLACE	00121614	00004311	00000287	00000931	00000792	00025404	00065362
13	R047	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000704	00001241	00022808	00068216
14	R031	34TH STREET & 7TH AVENUE	00095076	00003990	00000232	00000727	00001459	00024284	00038671
15	R017	LEXINGTON AVENUE	00094655	00004688	00000190	00000833	00000754	00020018	00055066
16	R175	8TH AVENUE-14TH STREET	00094313	00003907	00000286	00001144	00000256	00038272	00074661
17	R057	BARCLAYS CENTER	00093804	00004204	00000454	00001386	00001491	00039113	00068119
18	R138	WEST 4TH ST-WASHINGTON SQ	00093562	00004677	00000251	00000965	00000127	00031628	00074458





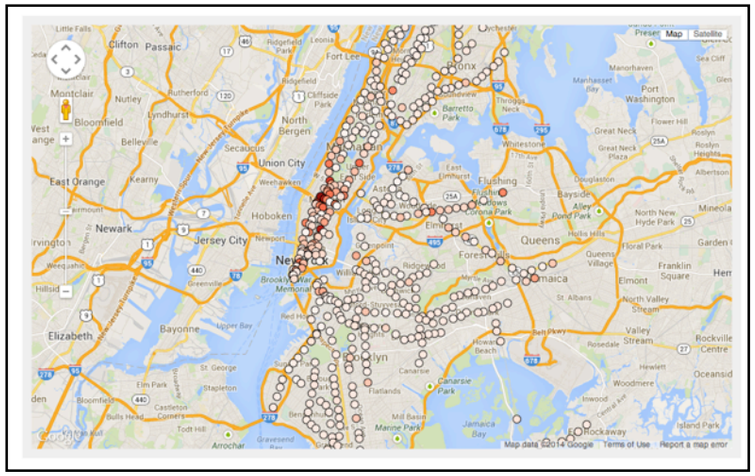


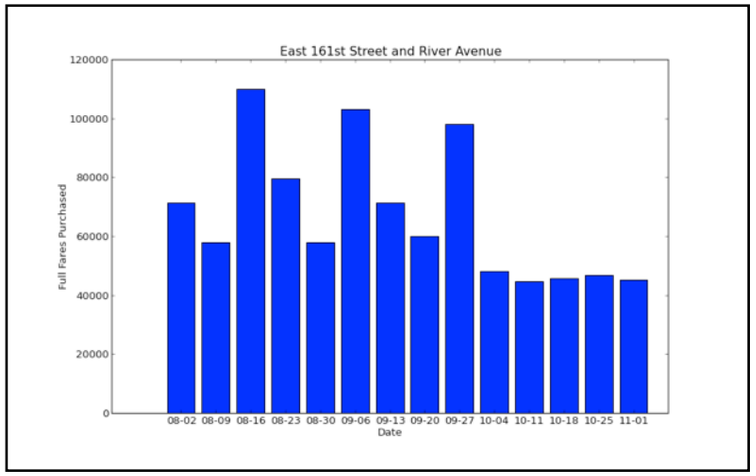
Parameters

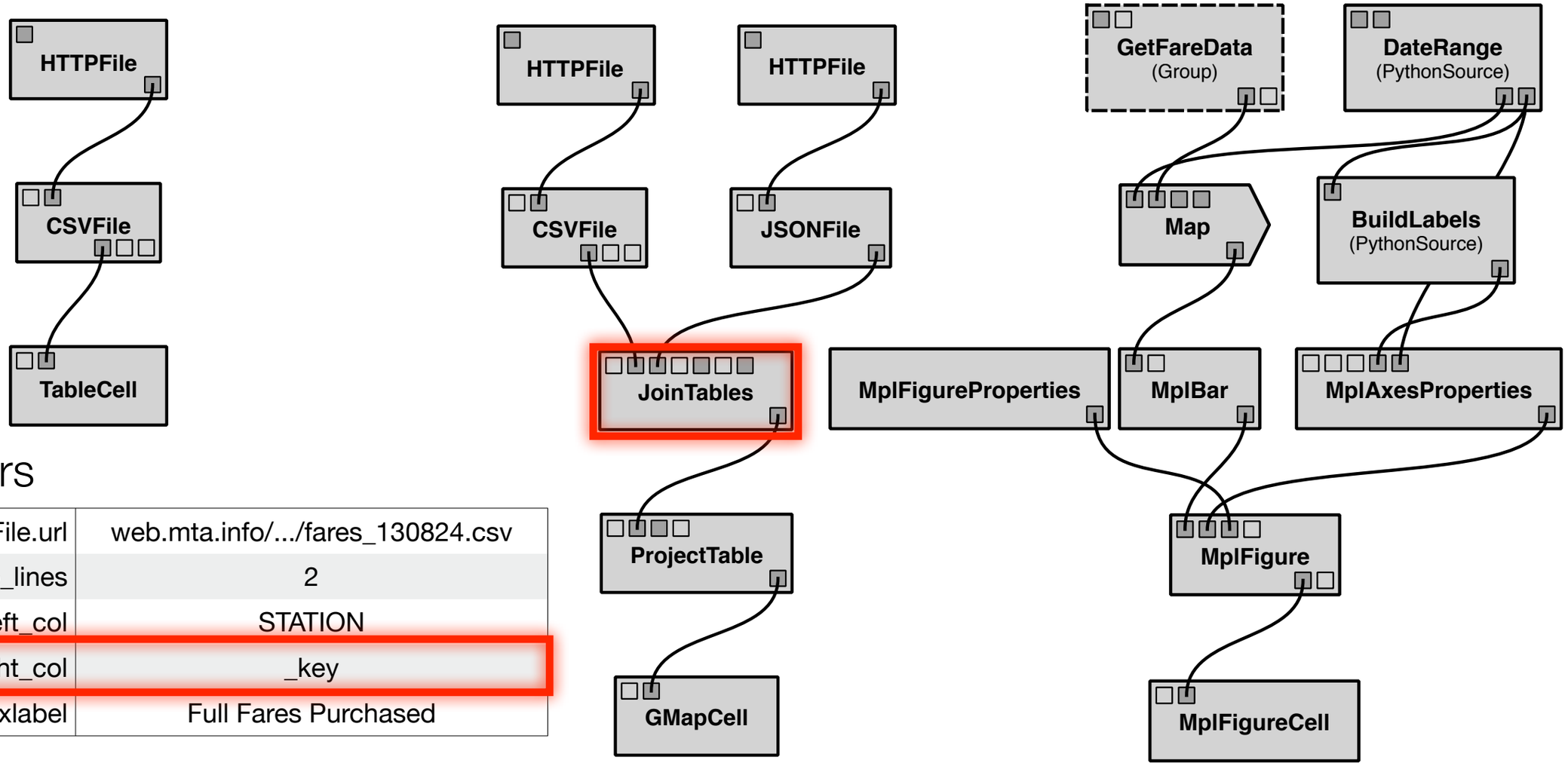
HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased

Example: MTA Fare Data

	REMOTE	STATION	FF	SEN/DIS	7-D AFAS UNL	D AFAS/RMF I	JOINT RR TKT	7-D UNL	30-D UNL
1	R011	42ND STREET & 8TH AVENUE	00228985	00008471	00000441	00001455	00000134	00033341	00071255
2	R170	14TH STREET-UNION SQUARE	00224603	00011051	00000827	00003026	00000660	00089367	00199841
3	R046	42ND STREET & GRAND CENTRAL	00207758	00007908	00000323	00001183	00003001	00040759	00096613
4	R012	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00001279	00003622	00035527	00067483
5	R293	34TH STREET - PENN STATION	00168768	00006155	00000523	00001065	00005031	00030645	00054376
6	R033	42ND STREET/TIMES SQUARE	00159382	00005945	00000378	00001205	00000690	00058931	00078644
7	R022	34TH STREET & 6TH AVENUE	00156008	00006276	00000487	00001543	00000712	00058910	00110466
8	R084	59TH STREET/COLUMBUS CIRCLE	00155262	00009484	00000589	00002071	00000542	00053397	00113966
9	R020	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000384	00001159	00000723	00037978	00090745
10	R179	86TH STREET-LEXINGTON AVE	00142169	00010367	00000470	00001839	00000271	00050328	00125250
11	R023	34TH STREET & 6TH AVENUE	00134052	00005005	00000348	00001112	00000649	00031531	00075040
12	R029	PARK PLACE	00121614	00004311	00000287	00000931	00000792	00025404	00065362
13	R047	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000704	00001241	00022808	00068216
14	R031	34TH STREET & 7TH AVENUE	00095076	00003990	00000232	00000727	00001459	00024284	00038671
15	R017	LEXINGTON AVENUE	00094655	00004688	00000190	00000833	00000754	00020018	00055066
16	R175	8TH AVENUE-14TH STREET	00094313	00003907	00000286	00001144	00000256	00038272	00074661
17	R057	BARCLAYS CENTER	00093804	00004204	00000454	00001386	00001491	00039113	00068119
18	R138	WEST 4TH ST-WASHINGTON SQ	00093562	00004677	00000251	00000965	00000127	00031628	00074458





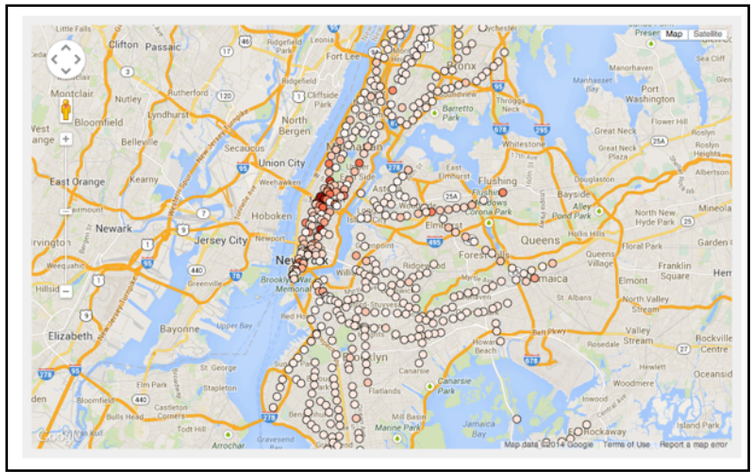


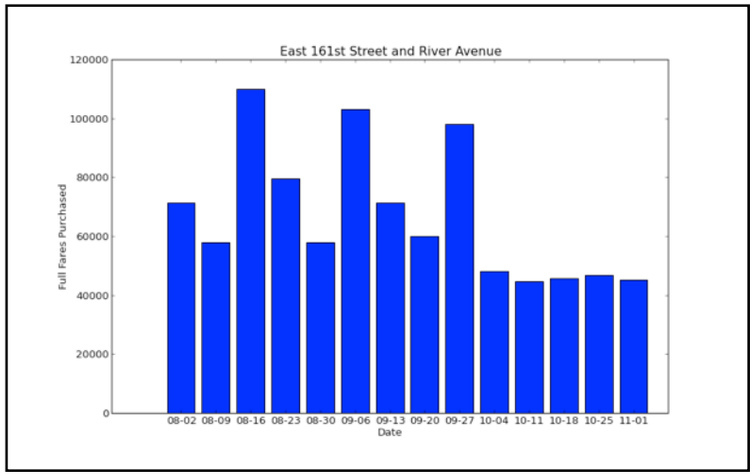
Parameters

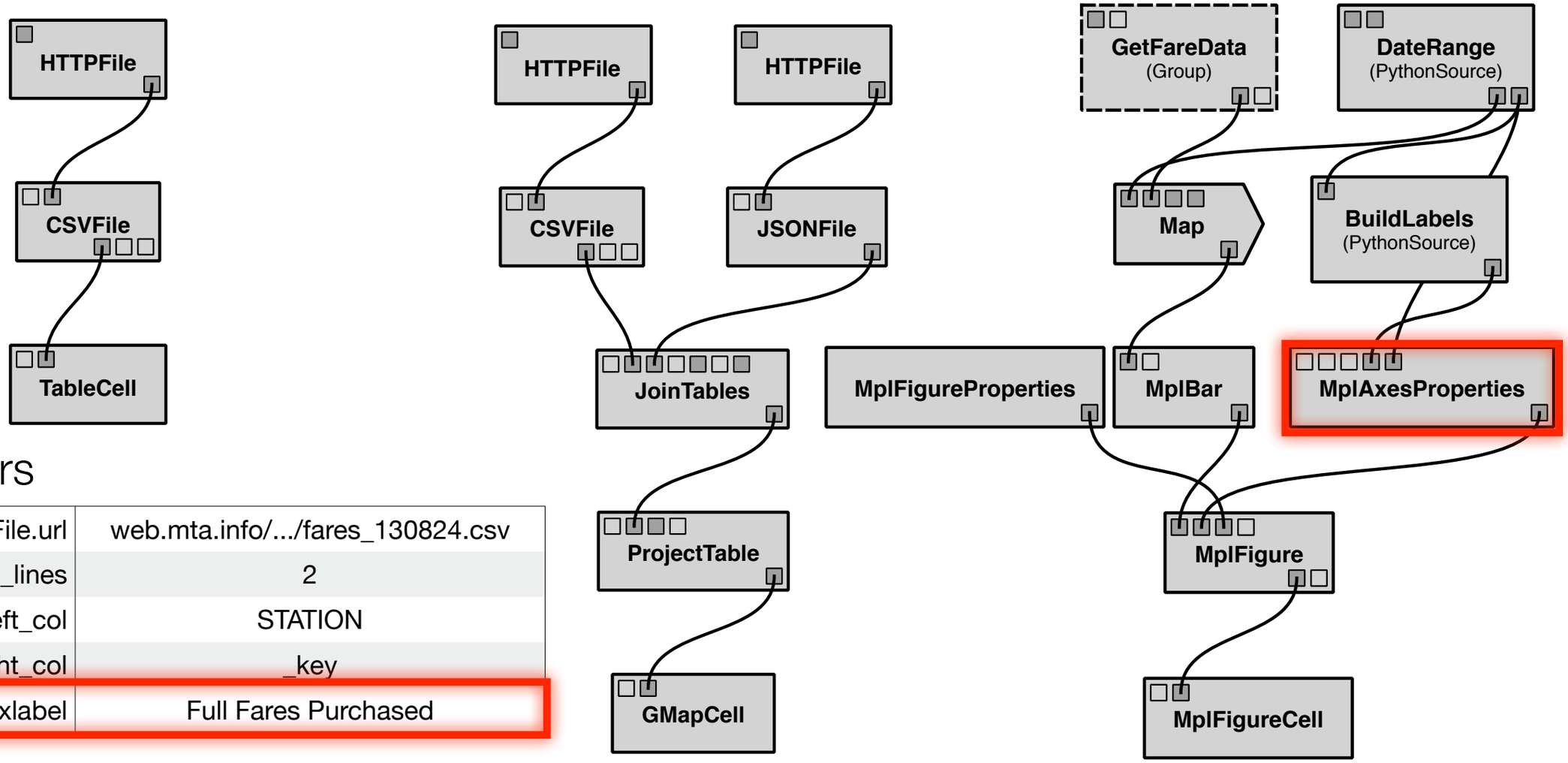
HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased

Example: MTA Fare Data

	REMOTE	STATION	FF	SEN/DIS	7-D AFAS UNL	D AFAS/RMF I	JOINT RR TKT	7-D UNL	30-D UNL
1	R011	42ND STREET & 8TH AVENUE	00228985	00008471	00000441	00001455	00000134	00033341	00071255
2	R170	14TH STREET-UNION SQUARE	00224603	00011051	00000827	00003026	00000660	00089367	00199841
3	R046	42ND STREET & GRAND CENTRAL	00207758	00007908	00000323	00001183	00003001	00040759	00096613
4	R012	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00001279	00003622	00035527	00067483
5	R293	34TH STREET - PENN STATION	00168768	00006155	00000523	00001065	00005031	00030645	00054376
6	R033	42ND STREET/TIMES SQUARE	00159382	00005945	00000378	00001205	00000690	00058931	00078644
7	R022	34TH STREET & 6TH AVENUE	00156008	00006276	00000487	00001543	00000712	00058910	00110466
8	R084	59TH STREET/COLUMBUS CIRCLE	00155262	00009484	00000589	00002071	00000542	00053397	00113966
9	R020	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000384	00001159	00000723	00037978	00090745
10	R179	86TH STREET-LEXINGTON AVE	00142169	00010367	00000470	00001839	00000271	00050328	00125250
11	R023	34TH STREET & 6TH AVENUE	00134052	00005005	00000348	00001112	00000649	00031531	00075040
12	R029	PARK PLACE	00121614	00004311	00000287	00000931	00000792	00025404	00065362
13	R047	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000704	00001241	00022808	00068216
14	R031	34TH STREET & 7TH AVENUE	00095076	00003990	00000232	00000727	00001459	00024284	00038671
15	R017	LEXINGTON AVENUE	00094655	00004688	00000190	00000833	00000754	00020018	00055066
16	R175	8TH AVENUE-14TH STREET	00094313	00003907	00000286	00001144	00000256	00038272	00074661
17	R057	BARCLAYS CENTER	00093804	00004204	00000454	00001386	00001491	00039113	00068119
18	R138	WEST 4TH ST-WASHINGTON SQ	00093562	00004677	00000251	00000965	00000127	00031628	00074458





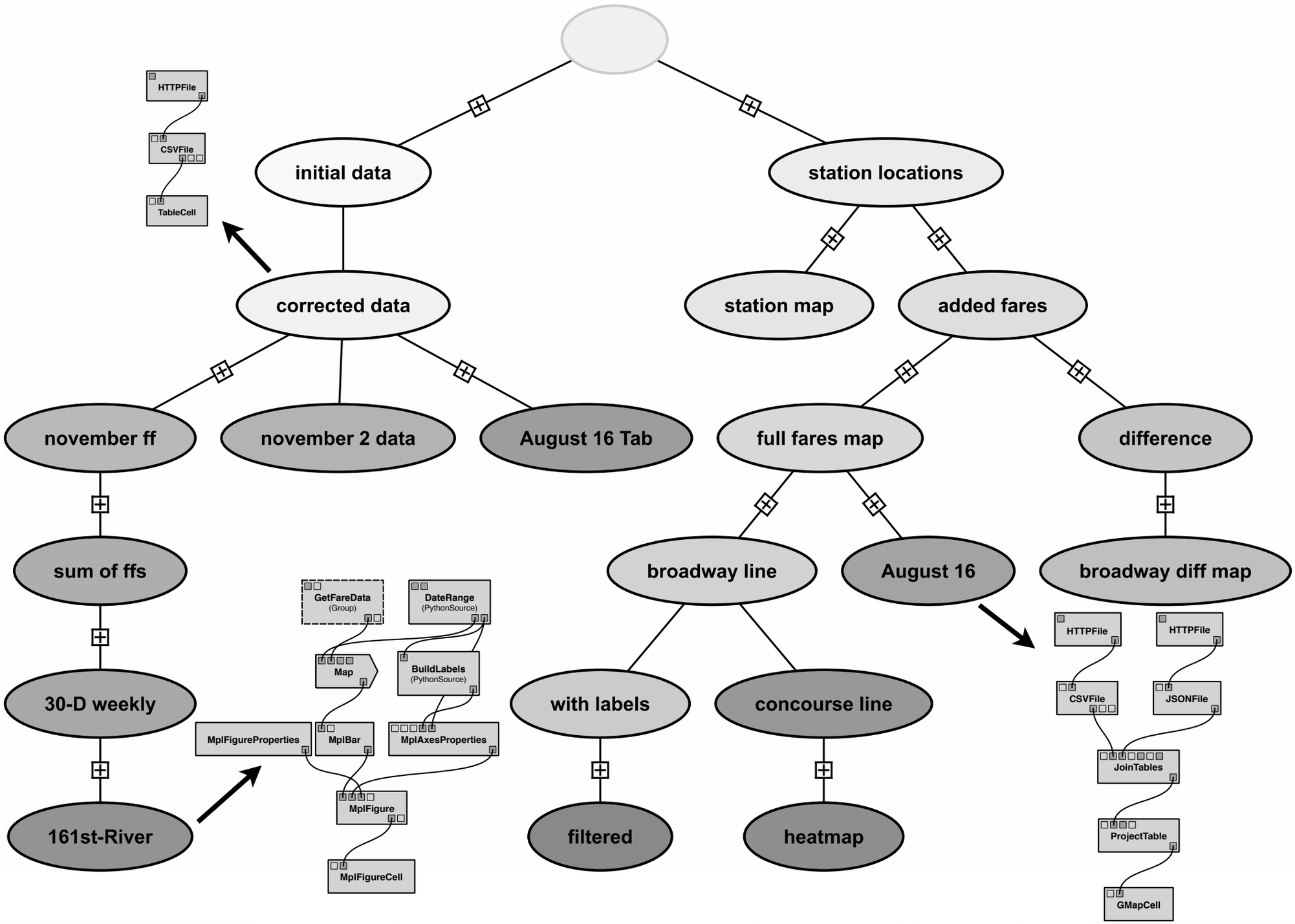


Parameters

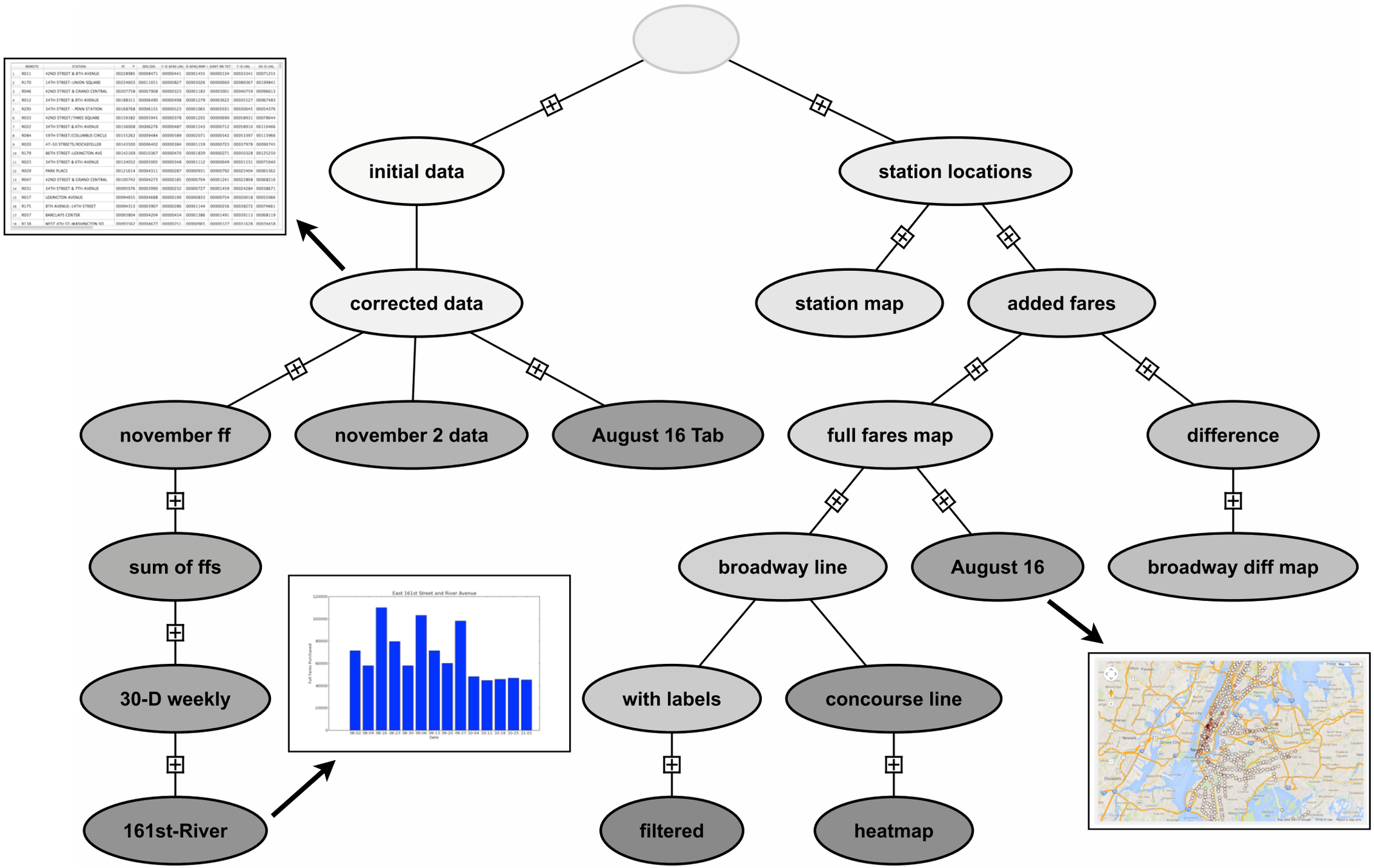
HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased

Capturing Provenance

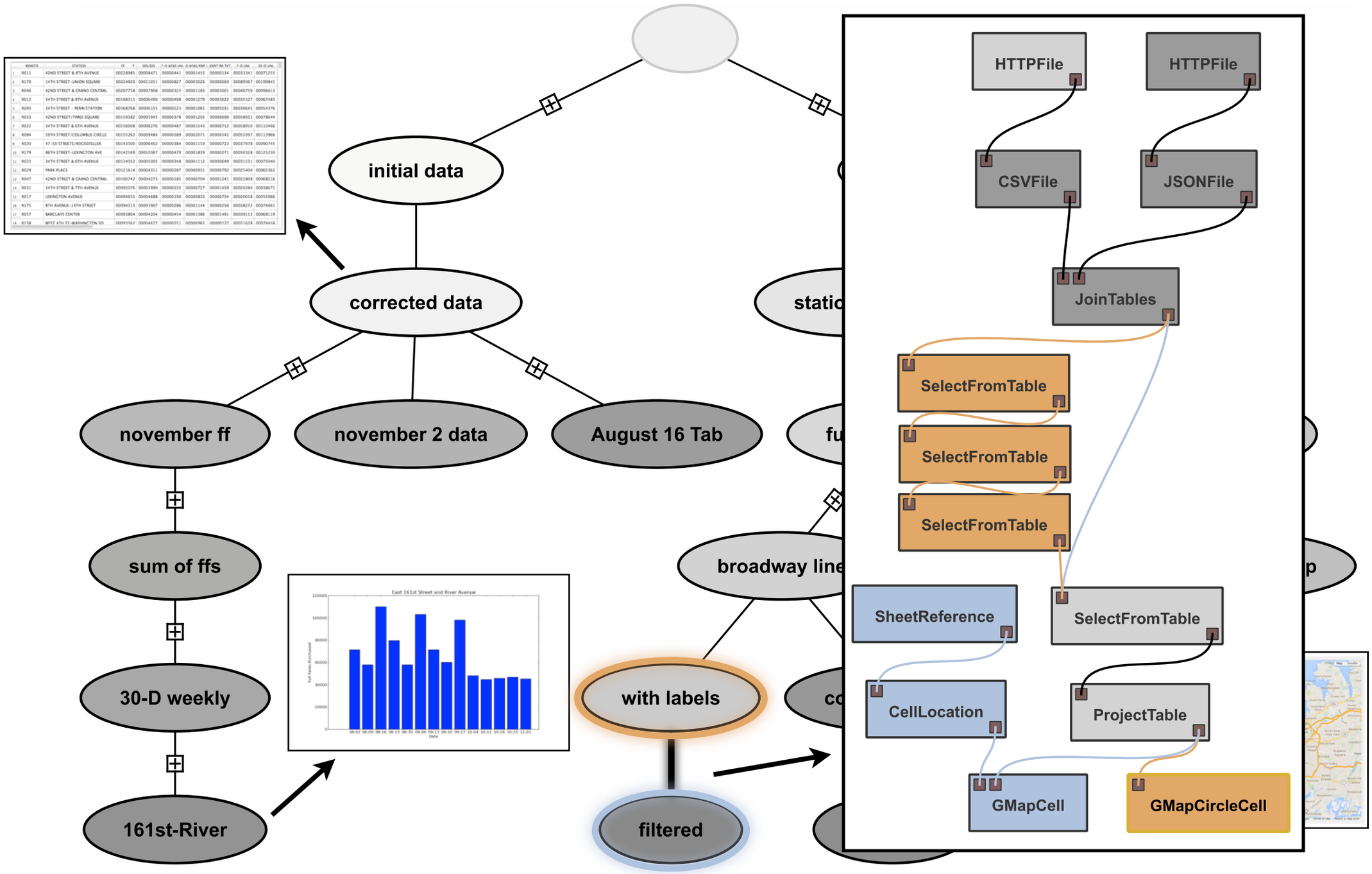
Workflow Evolution Provenance of MTA Fare Data



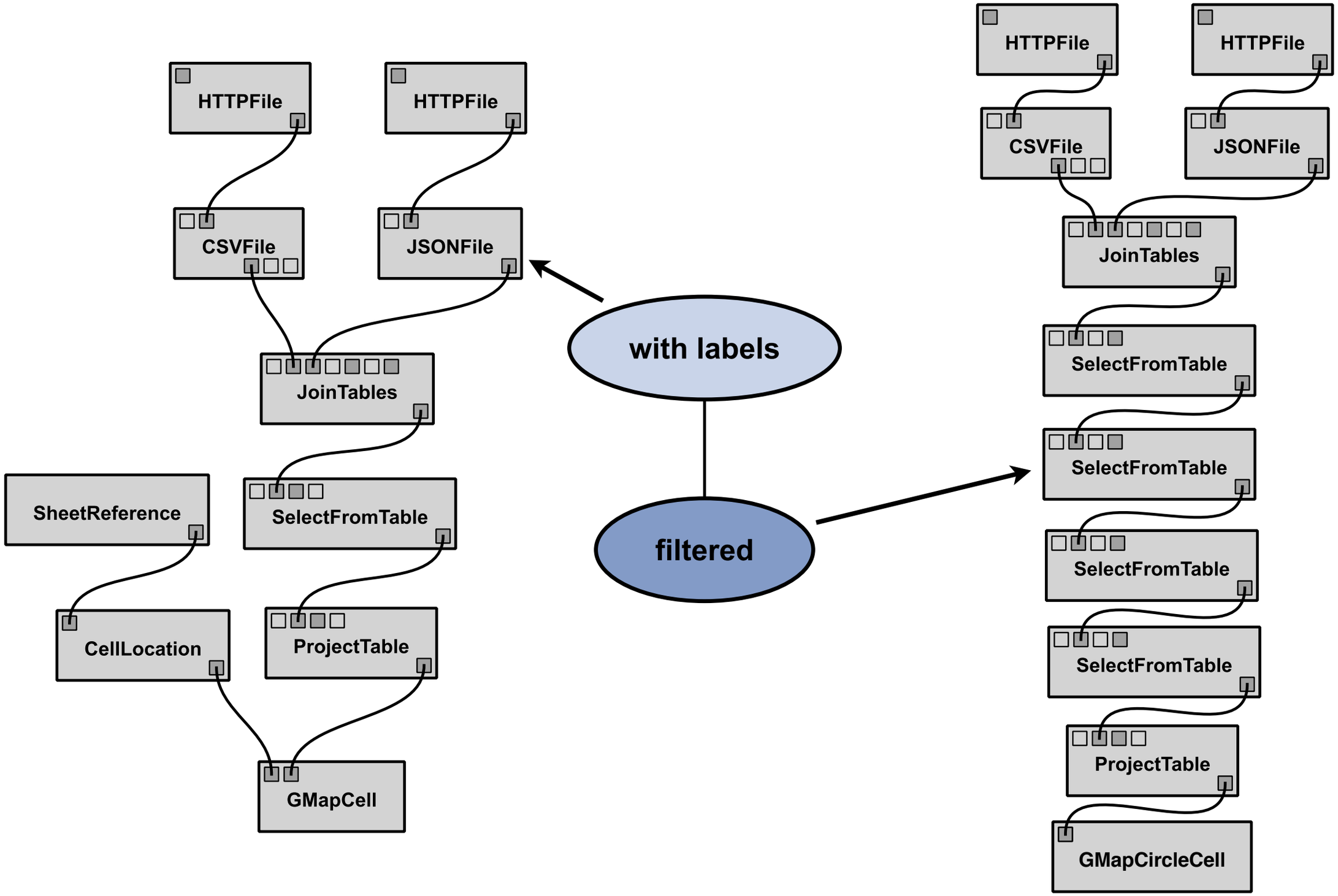
Workflow Evolution Provenance of MTA Fare Data



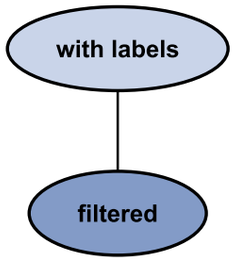
Workflow Evolution Provenance of MTA Fare Data

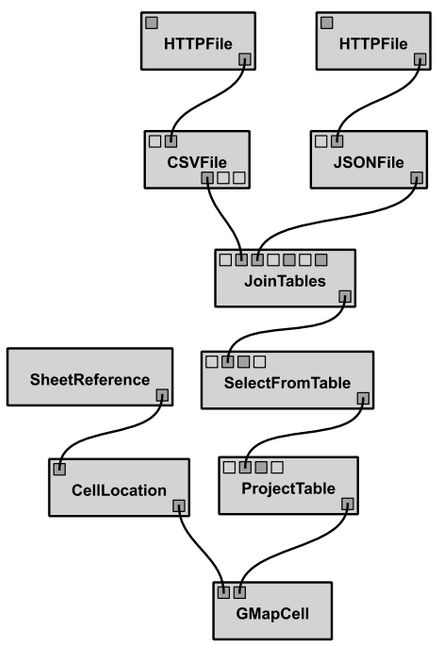


Workflow Evolution Provenance

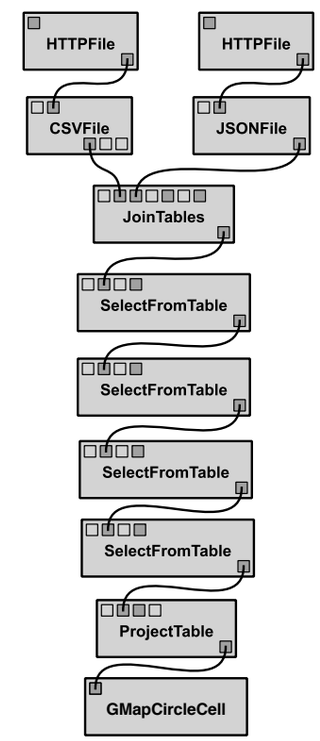


Workflow Evolution Provenance

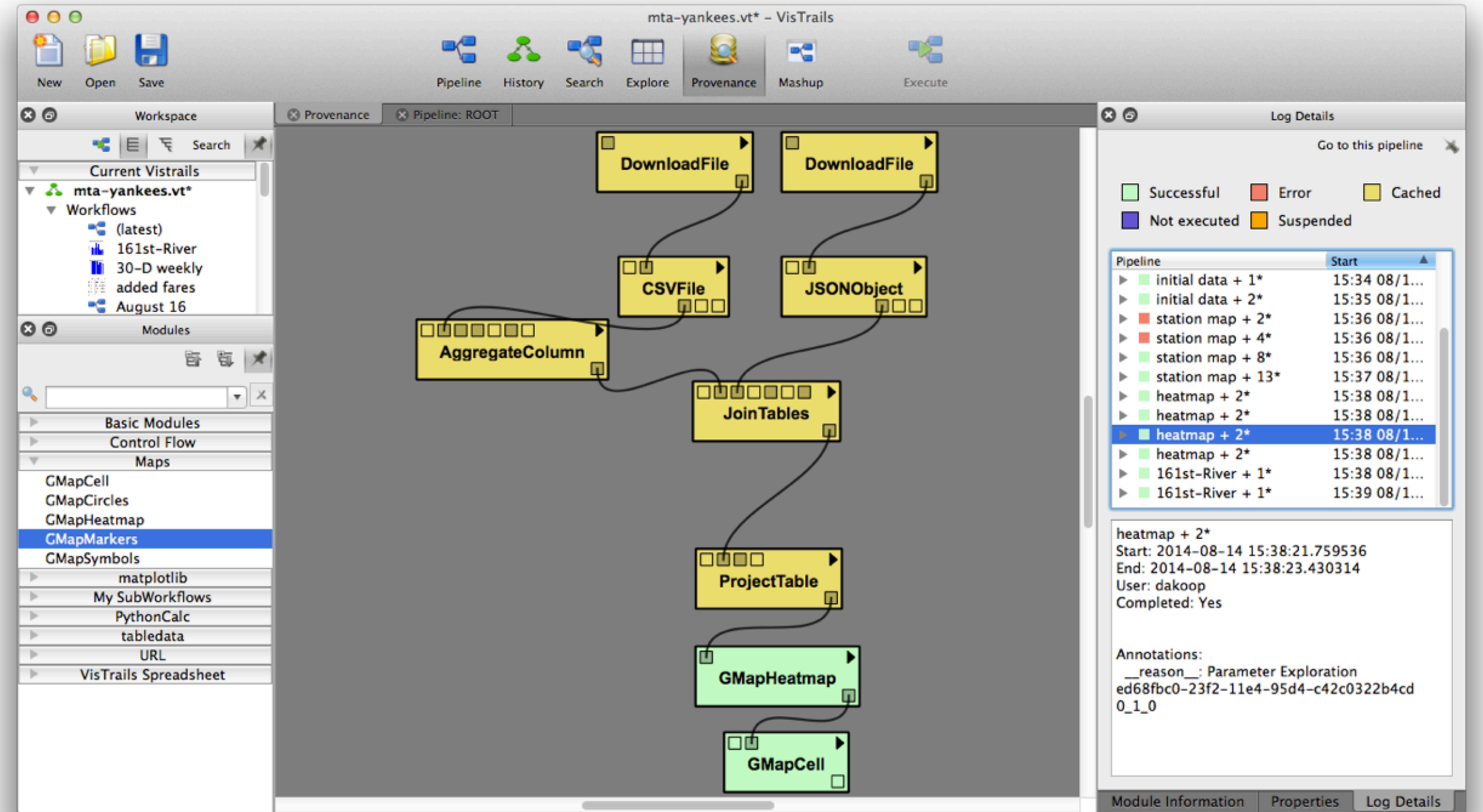




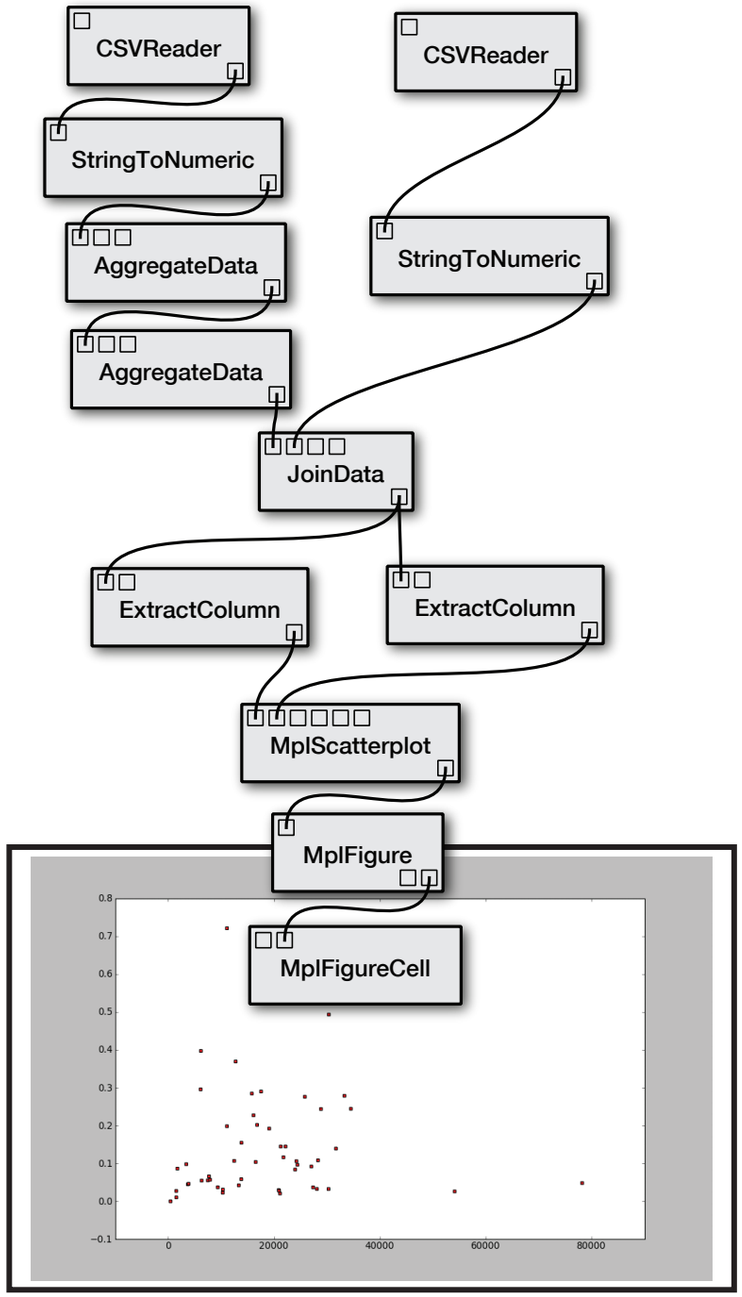
delete module “GMapCell”
delete module “CellLocation”
delete module “ProjectTable”
delete module “SelectFromTable”
…
add module “SelectFromTable”
add parameter “float_expr” to “SelectFromTable” with value “latitutde > 40.6”
delete parameter “float_expr” from “SelectFromTable”
add parameter “float_expr” to “SelectFromTable” with value “latitutde > 40.7”
delete parameter “float_expr” from “SelectFromTable”
add parameter “float_expr” to “SelectFromTable” with value “latitutde > 40.8”
…

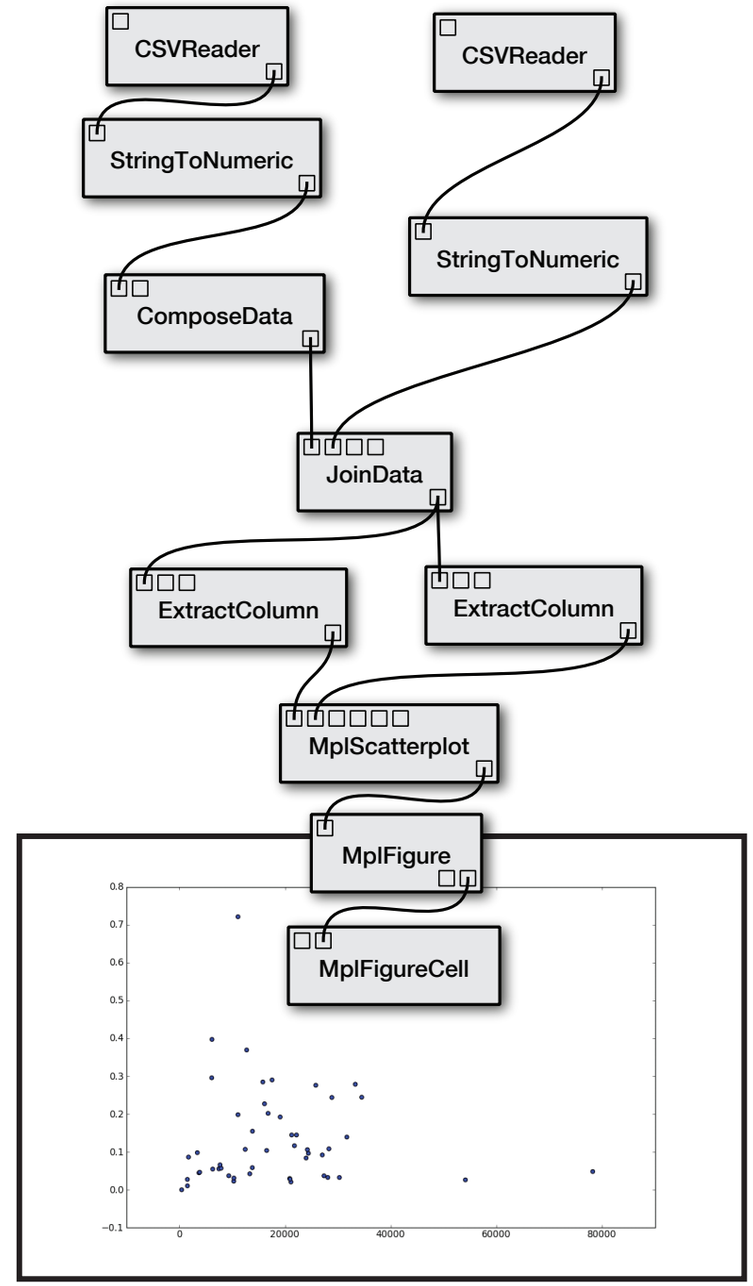


Execution Provenance

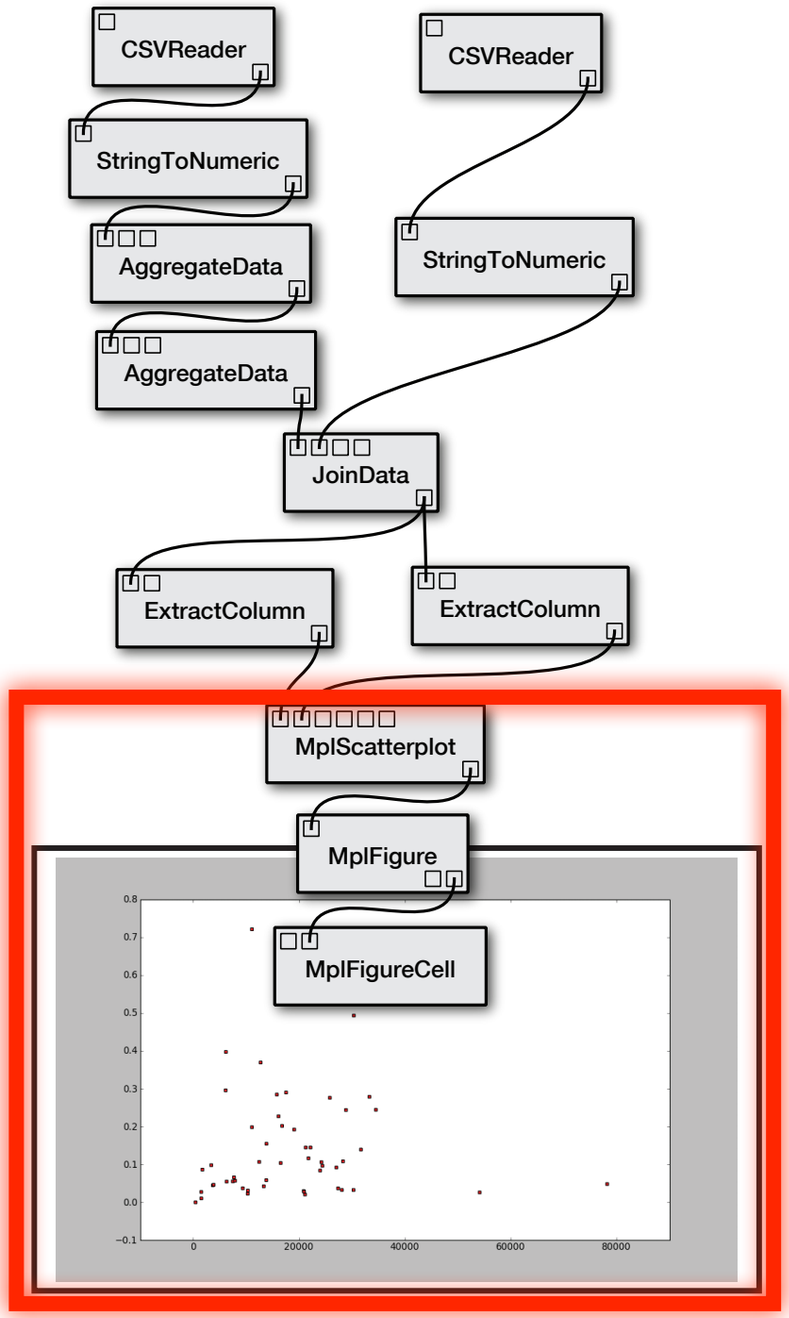


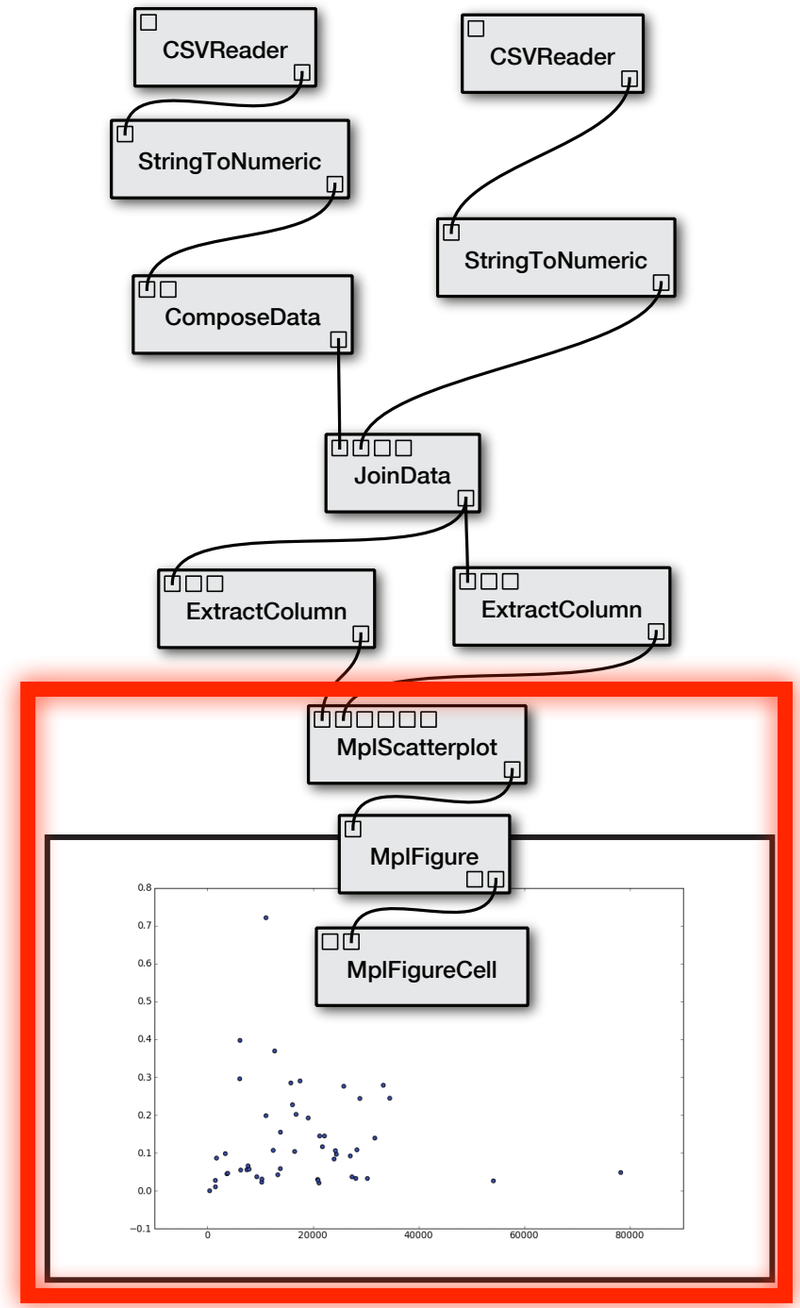
Workflow Upgrades



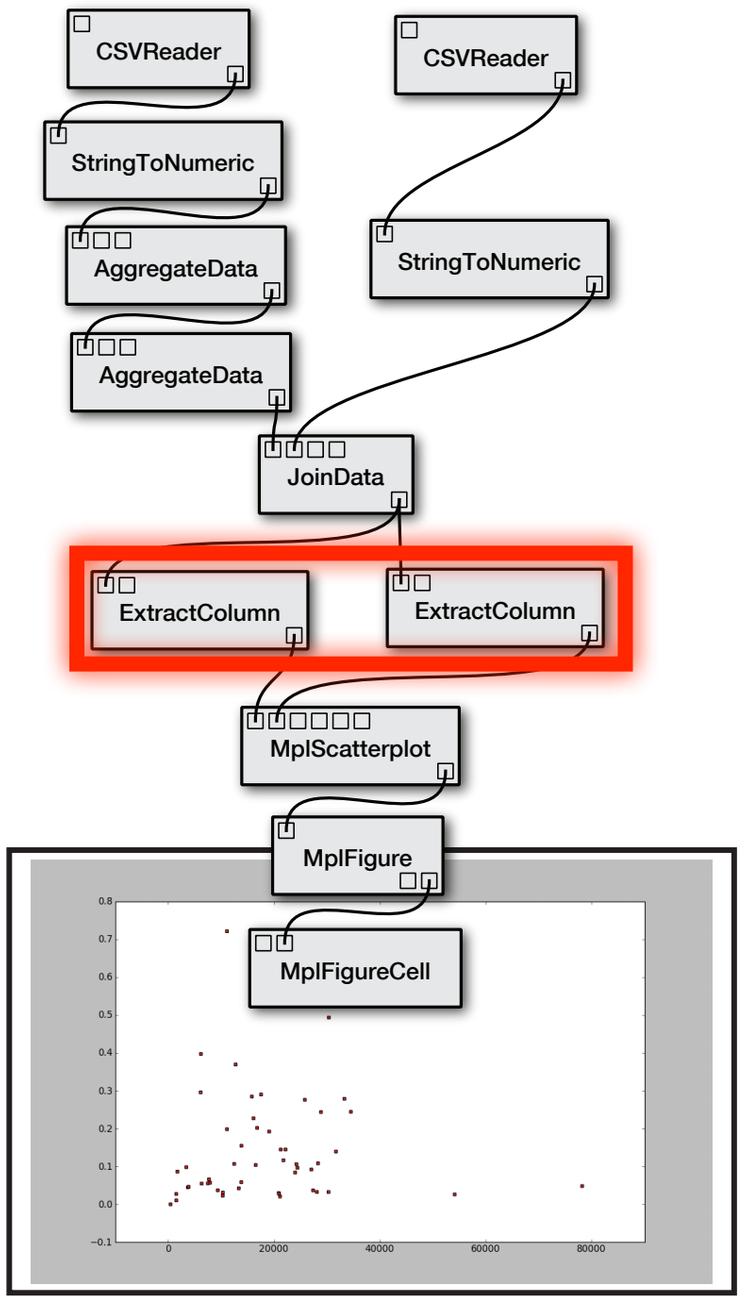


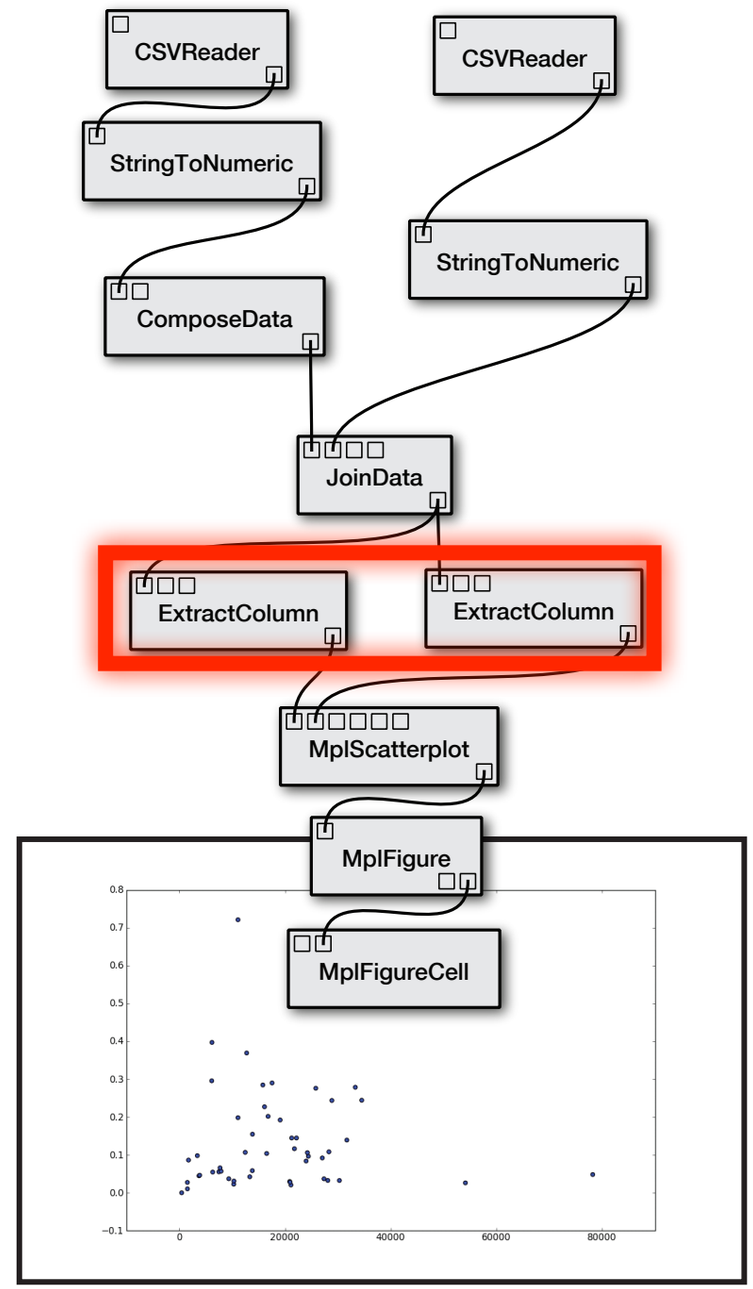
Workflow Upgrades



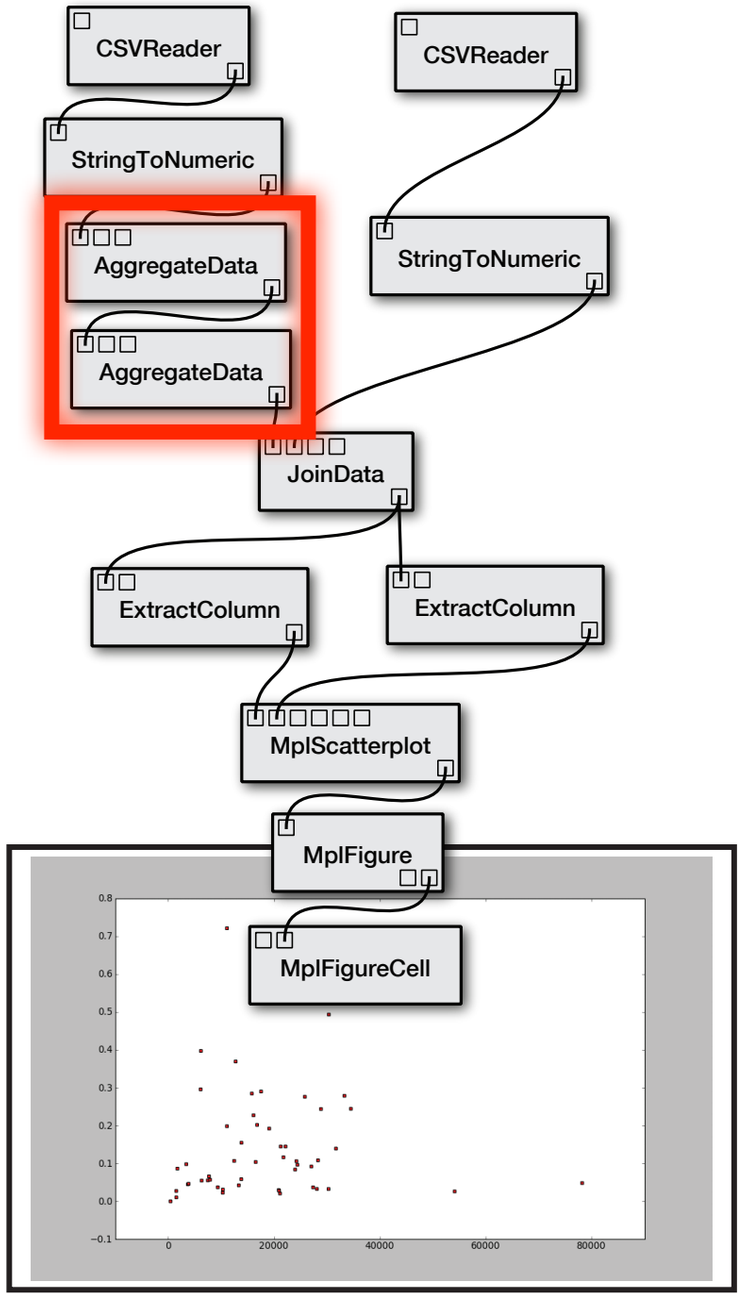


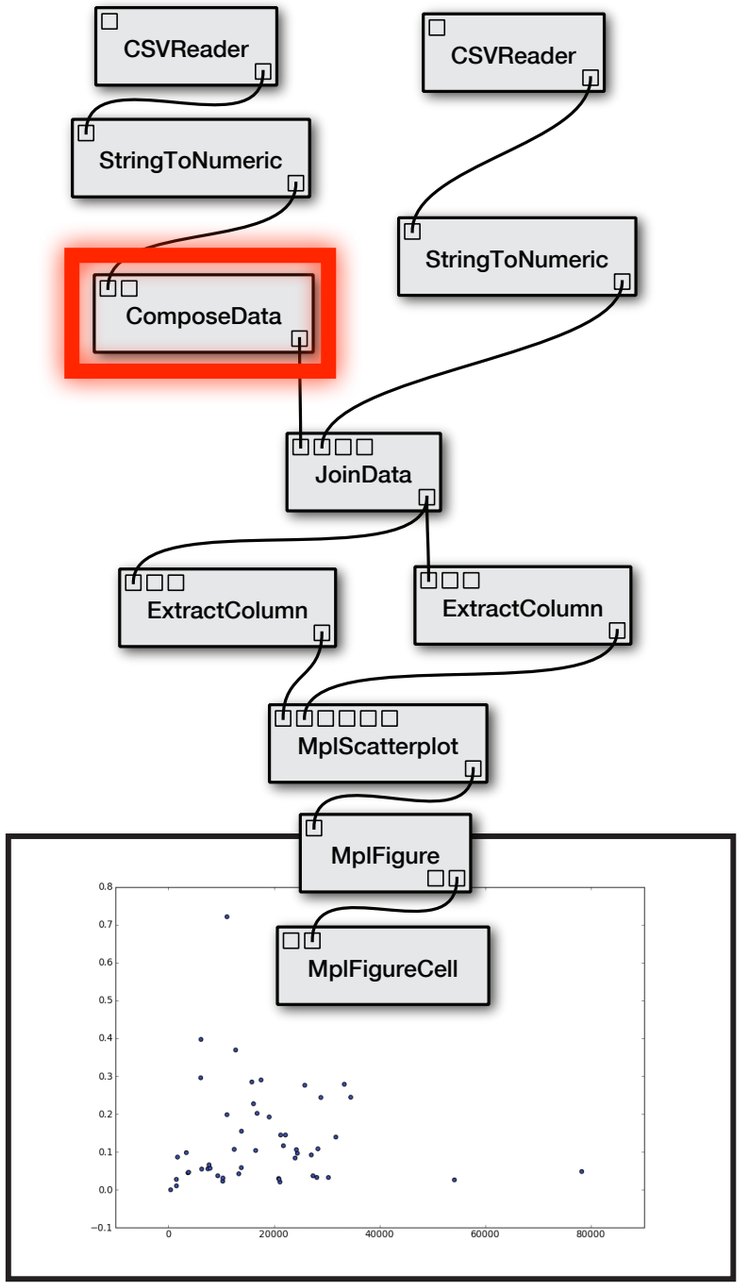
Workflow Upgrades



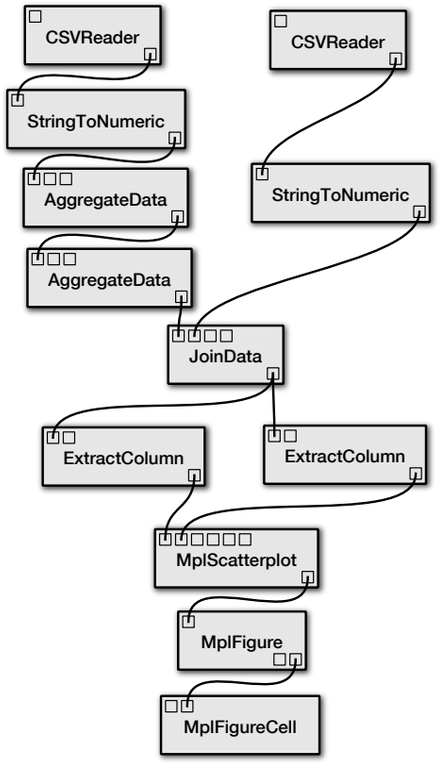


Workflow Upgrades



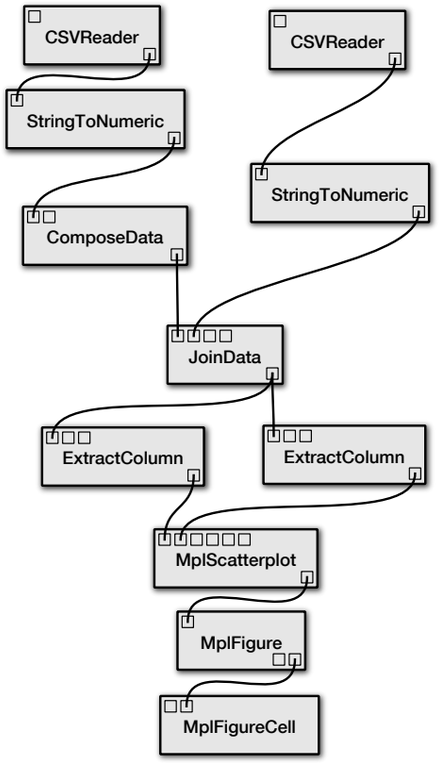


Provenance of Workflow Upgrades

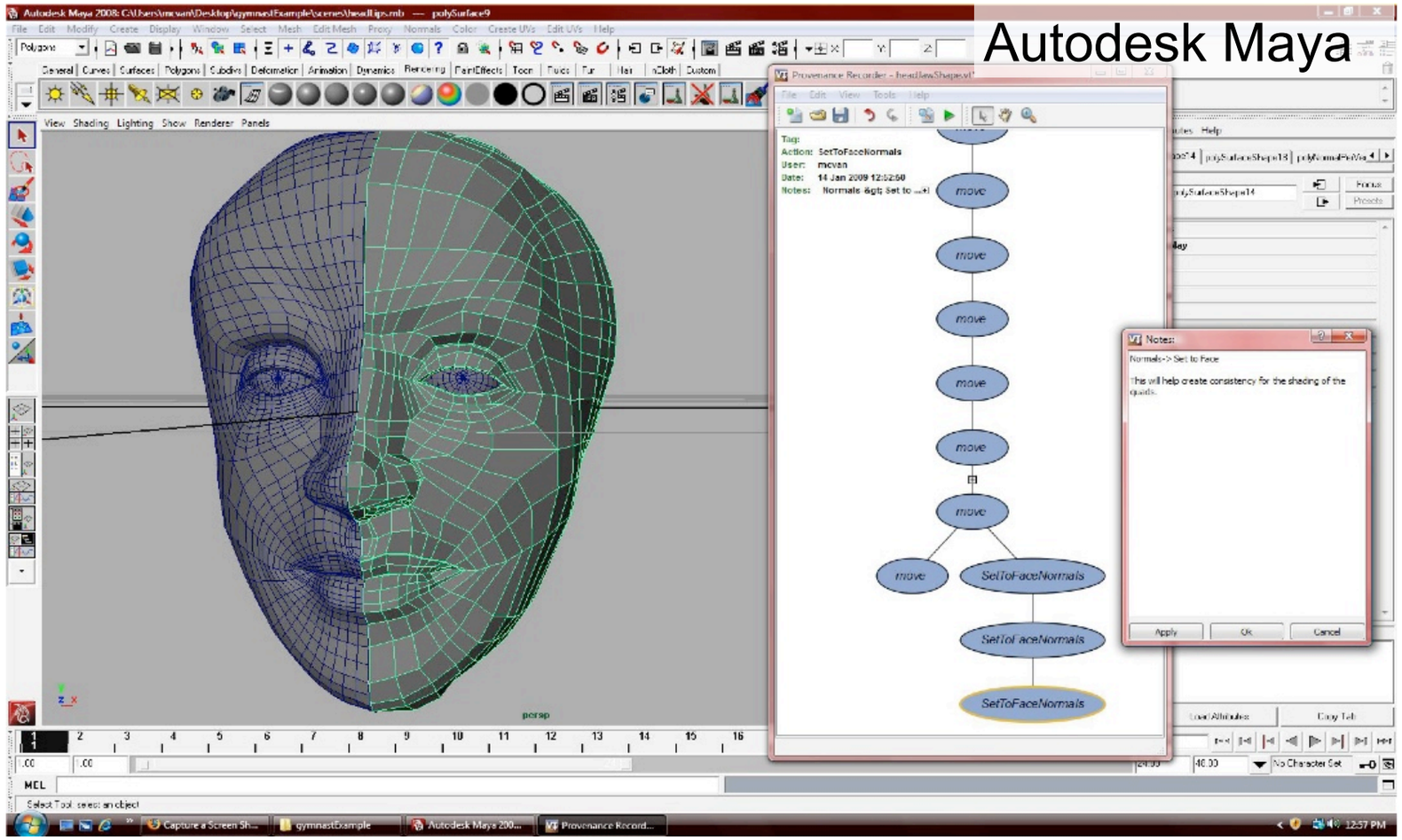


Change-based Provenance:

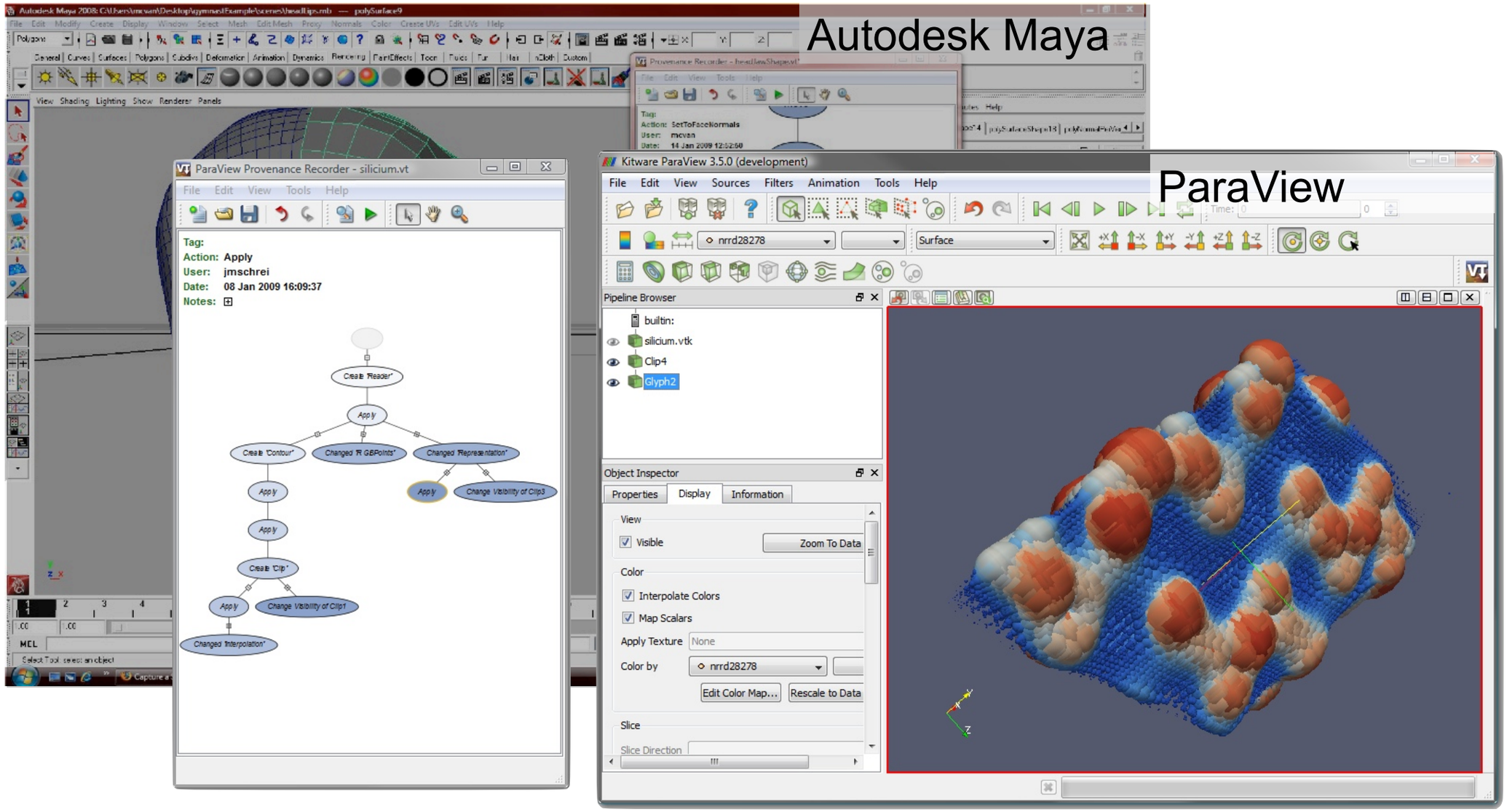
```
delete connection StringToNumeric → AggregateData
delete connection AggregateData → AggregateData
delete connection AggregateData → JoinData
delete connection JoinData → ExtractColumn
delete connection JoinData → ExtractColumn
delete connection ExtractColumn → MplScatterplot
delete connection ExtractColumn → MplScatterplot
delete connection MplScatterplot → MplFigure
delete connection MplFfigure → MplFigureCell
delete module AggregateData version 1.0.4
delete module AggregateData version 1.0.4
delete module ExtractColumn version 0.9.7
delete module ExtractColumn version 0.9.7
delete module MplScatterplot version 2.0.0
delete module MplFigure version 2.0.0
delete module MplFigureCell version 2.0.0
add module ComposeData version 1.1.0
add module ExtractColumn version 1.0.2
add module ExtractColumn version 1.0.2
add module MplScatterplot version 2.0.1
add module MplFigure version 2.0.1
add module MplFigureCell version 2.0.1
add connection StringToNumeric → ComposeData
add connection ComposeData → JoinData
add connection JoinData → ExtractColumn
add connection JoinData → ExtractColumn
add connection ExtractColumn → MplScatterplot
add connection ExtractColumn → MplScatterplot
...
```



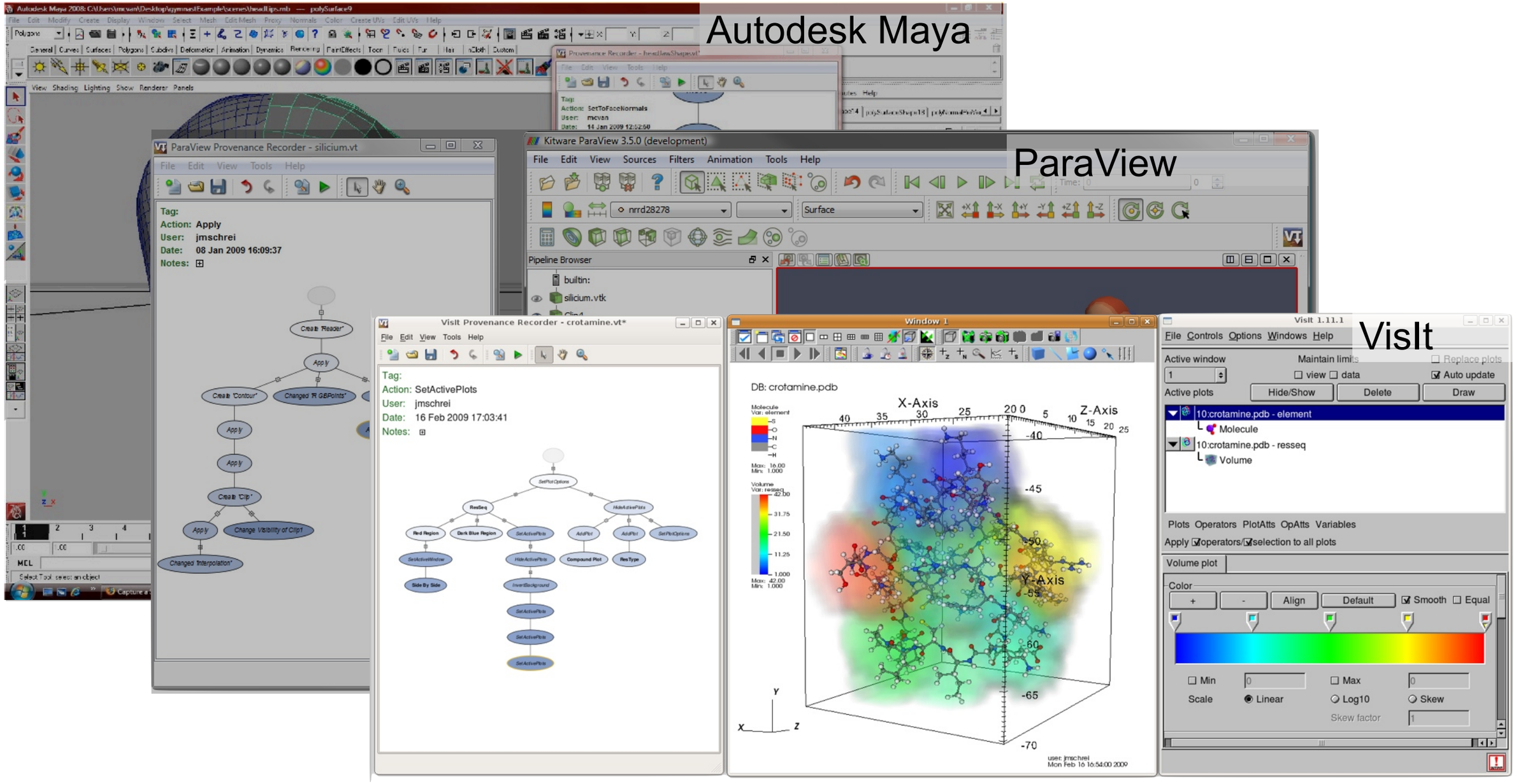
Adding Provenance to 3rd-Party Tools



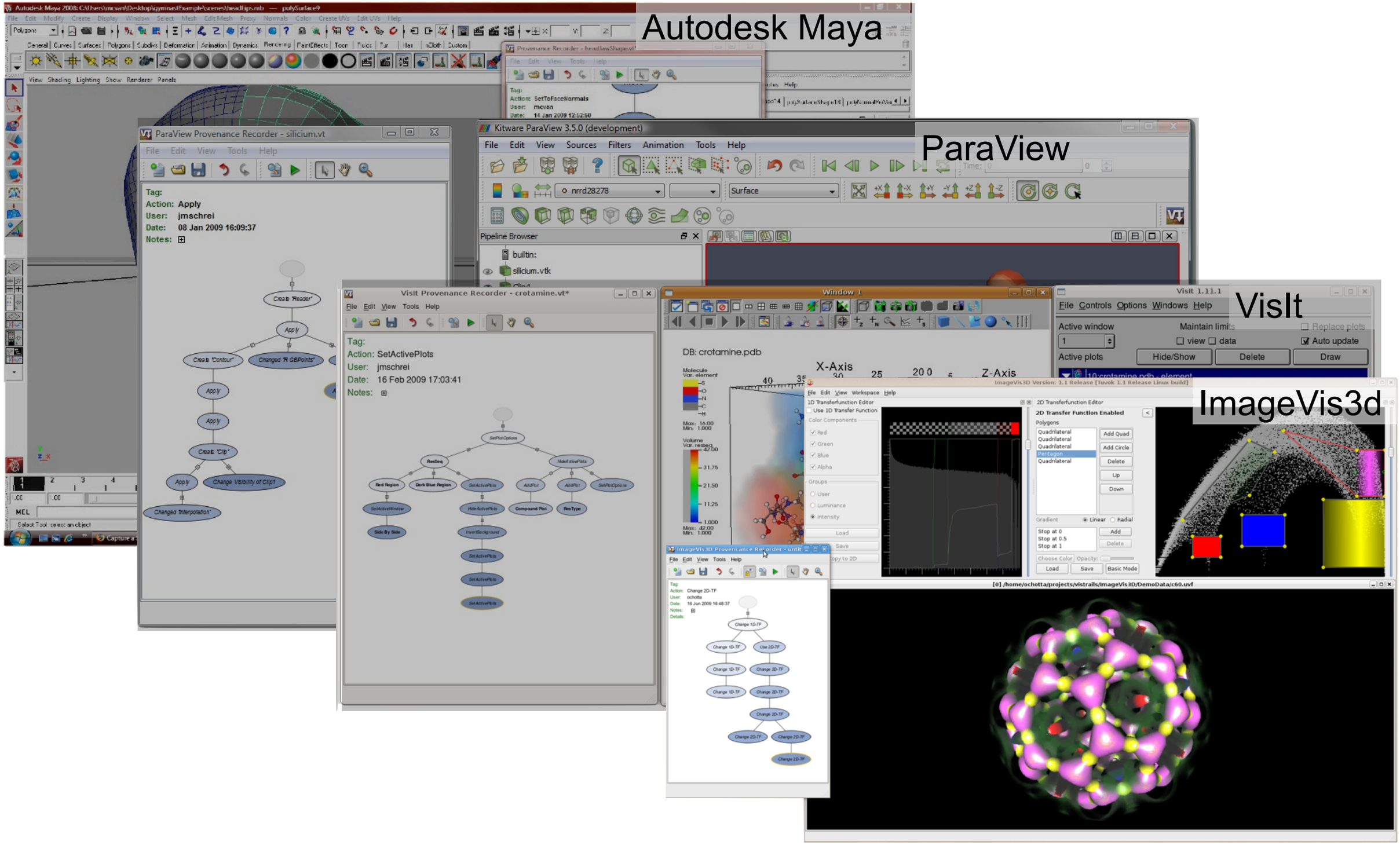
Adding Provenance to 3rd-Party Tools



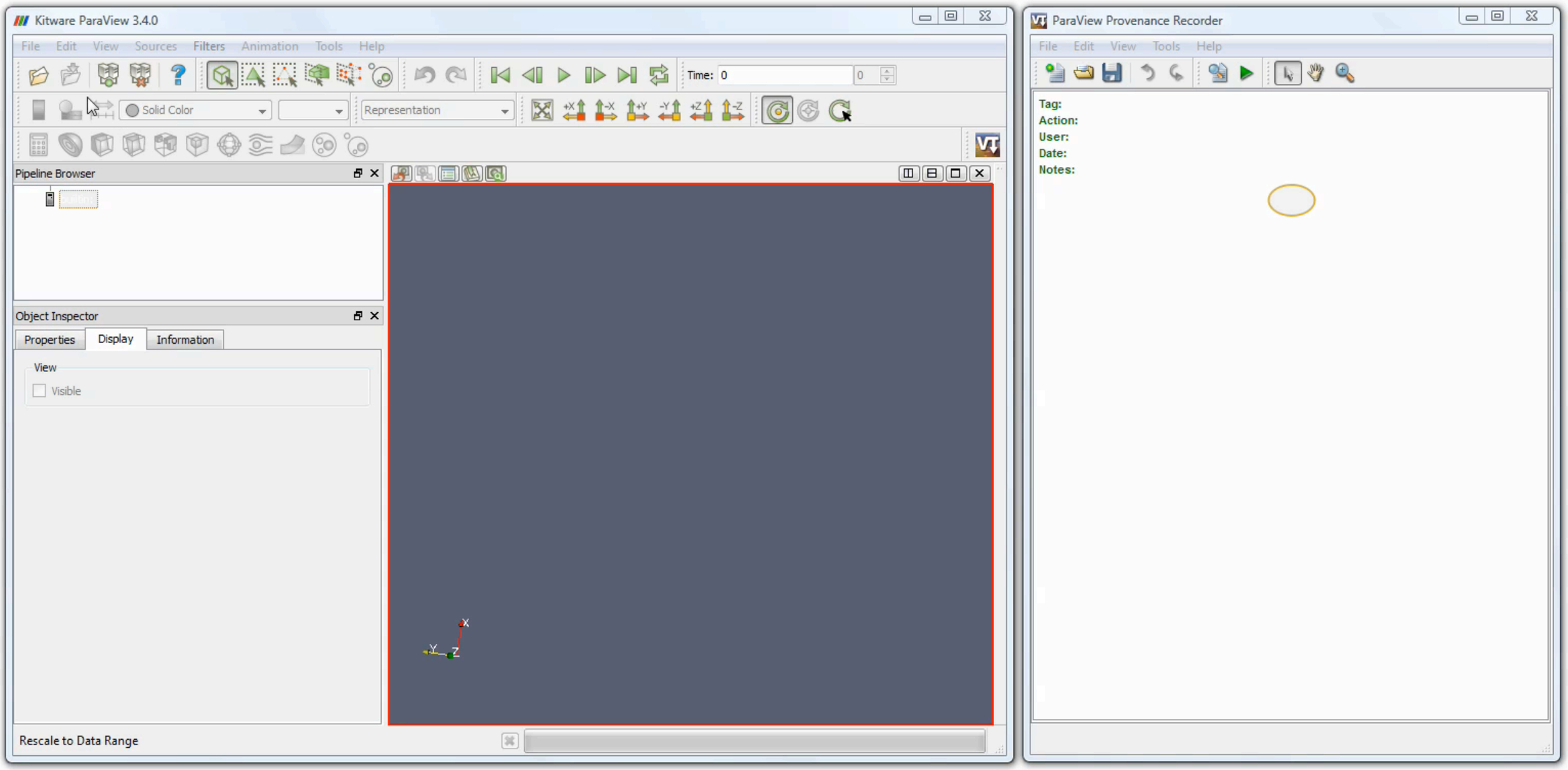
Adding Provenance to 3rd-Party Tools



Adding Provenance to 3rd-Party Tools

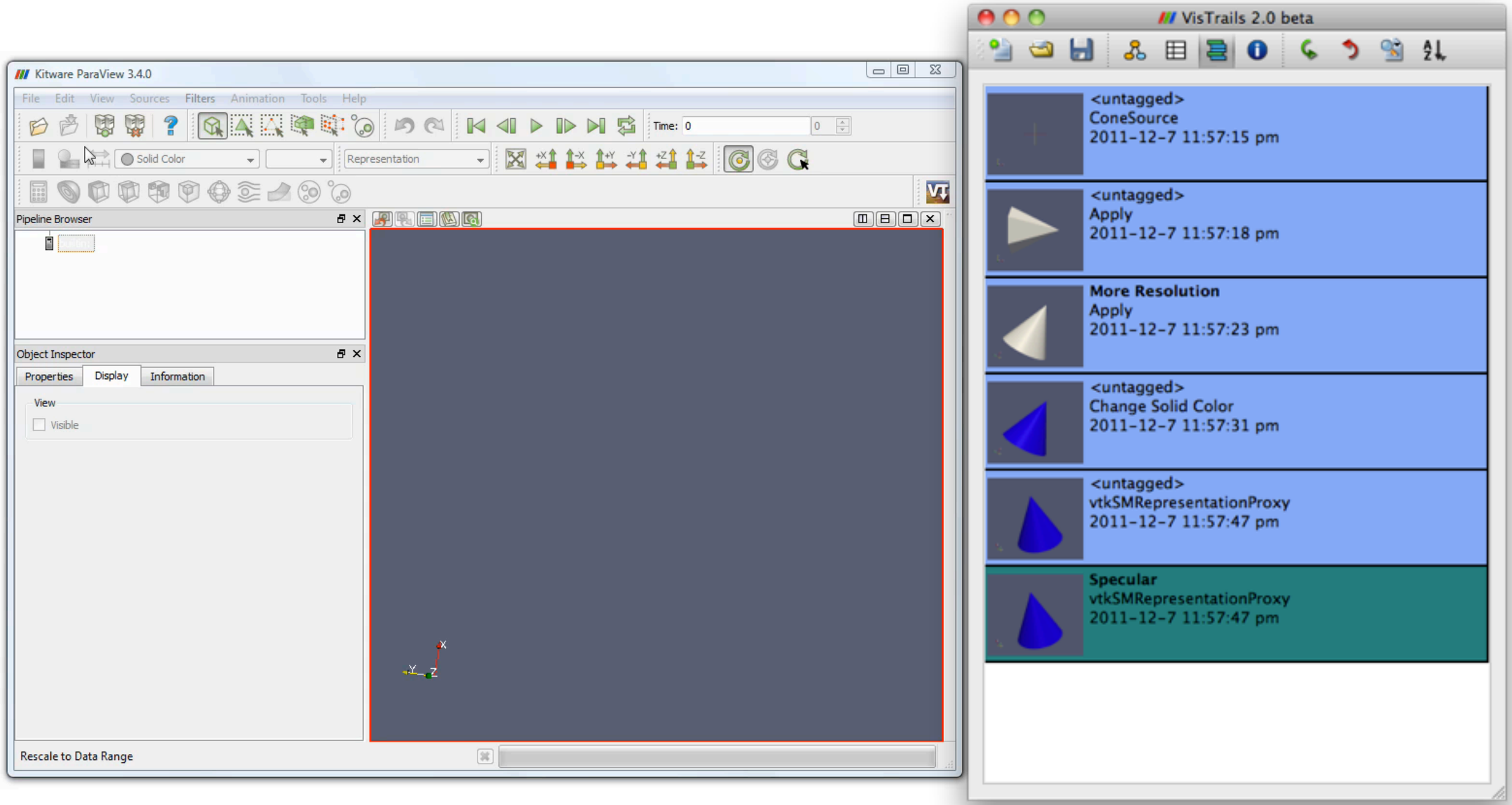


VisTrails Provenance Plugin for ParaView



[VisTrails, Inc.]

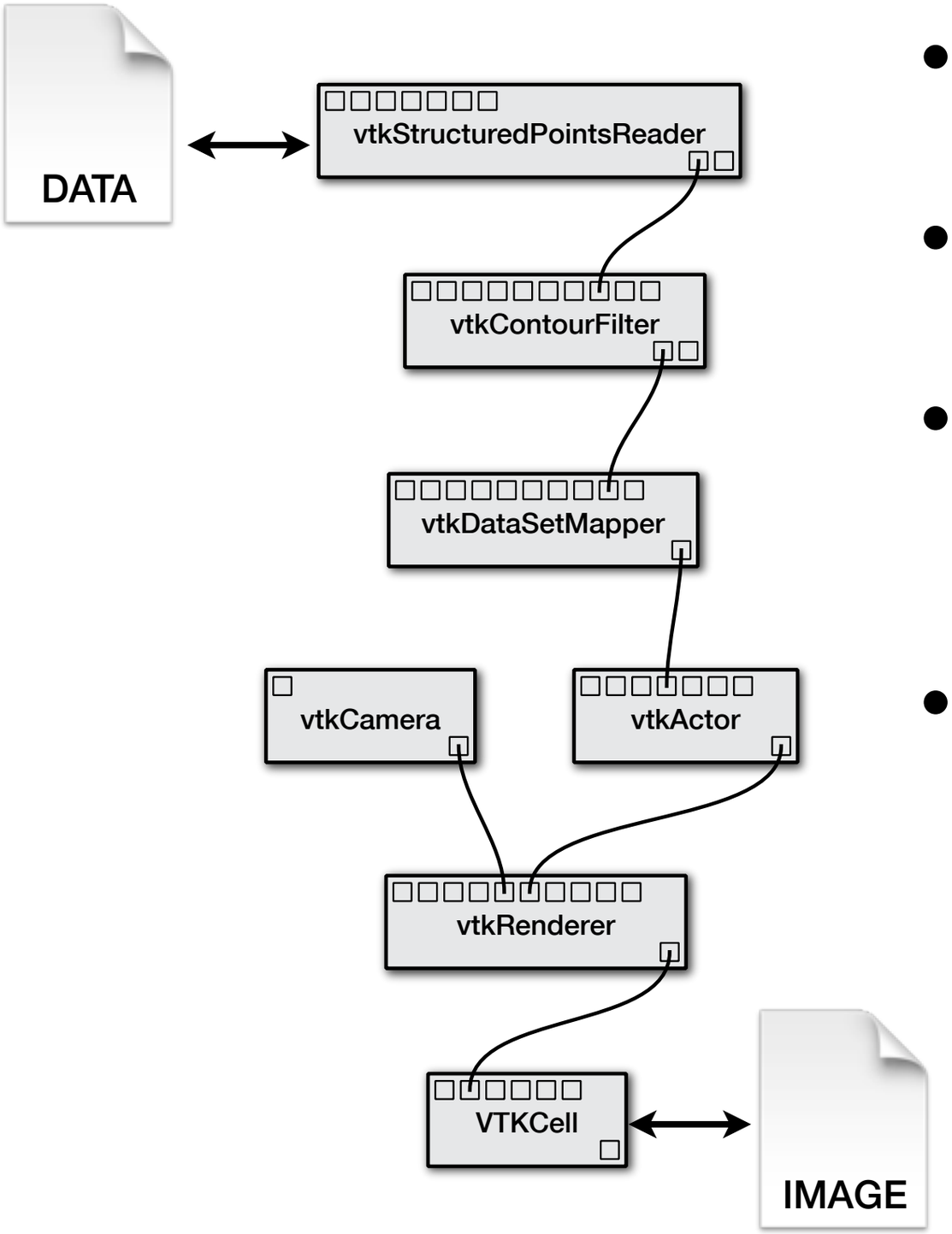
VisTrails Provenance Plugin for ParaView



[VisTrails, Inc.]

Querying and Re-using Provenance

Querying Provenance

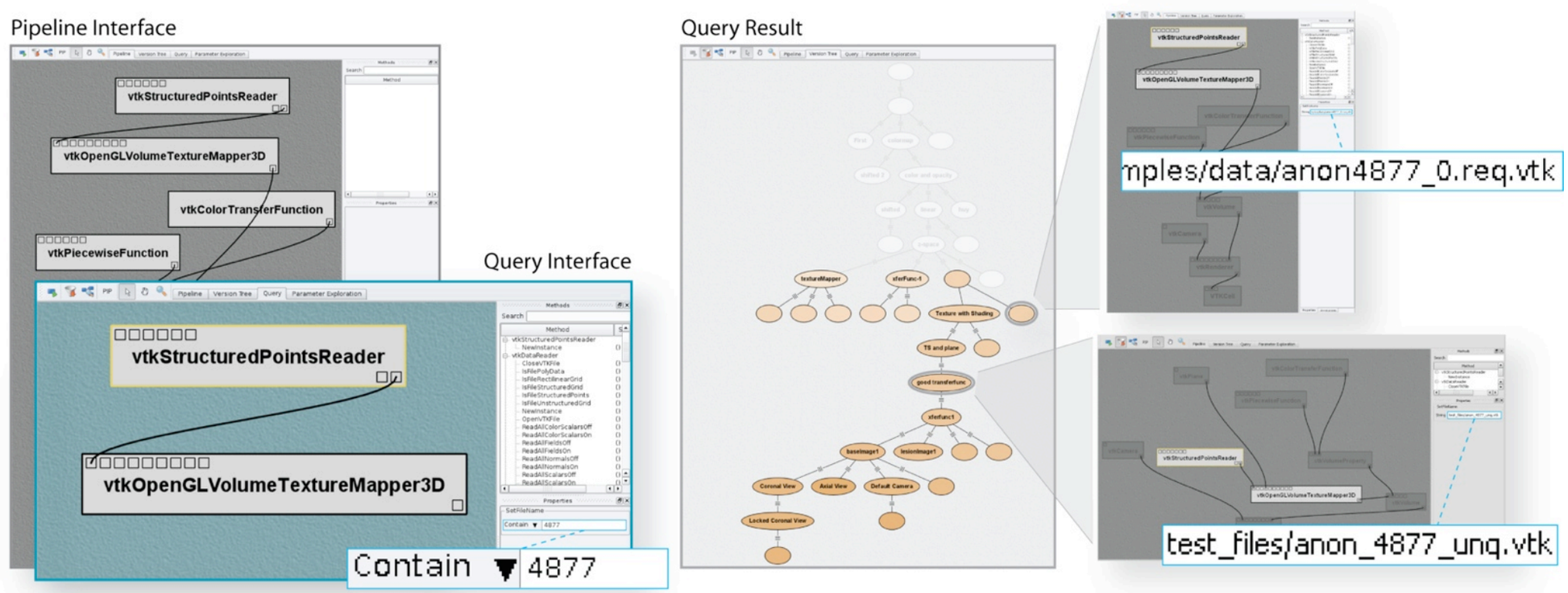


- *What process led to the output image?*
- *What input datasets contributed to the output image?*
- *What workflows include resampling and isosurfacing with isovalue 57?*

- Graph traversal or graph patterns
  - How do we write such queries?

Querying Provenance by Example

- Provenance is represented as graphs: hard to specify queries using text!
- Querying workflows by example [Scheidegger et al., TVCG 2007; Beeri et al., VLDB 2006; Beeri et al. VLDB 2007]
  - WYSIWYQ -- What You See Is What You Query
  - Interface to create workflow is same as to query



Linking Provenance and Data

```
<workflow_exec id="1">
  <m_exec id="5"
          name="vtkStructuredDataReader"
          package="edu.utah.sci.vistrails.vtk"
          version="5.6.0">
    <param id="2" name="SetFile"
           value="/MyData/05-12-sc2.dat"/>
  </m_exec>
  <m_exec id="6"
          name="vtkContourFilter"
          package="edu.utah.sci.vistrails.vtk"
          version="5.6.0">
    <param id="3" name="SetValue"
           value="[1, 57]"/>
    <param id="4" name="ComputeScalarsOn"
           value="True"/>
  </m_exec>

...

<m_exec id="11"
        name="FileSink"
        package="edu.utah.sci.vistrails.basic"
        version="1.5">
  <param id="15" name="path"
         value="/home/a/results/23.out"/>
</m_exec>
```

- Filenames are often the mode of identification in data exploration
- We might also use URIs or access curated data stores
  - Can this always be expected for exploratory tasks?
  - What happens if offline?
- Solution:
  - Managed store for data associated with computations
  - Improved data identification
  - Automatic versioning

Linking Provenance and Data

```
<workflow_exec id="1">
  <m_exec id="5"
          name="vtkStructuredDataReader"
          package="edu.utah.sci.vistrails.vtk"
          version="5.6.0">
    <param id="2" name="Se
           value="/MyData      2.dat"/>
  </m_exec>
  <m_exec id="6"
          name="vtkContourFilter"
          package="edu.utah.sci.vistrails.vtk"
          version="5.6.0">
    <param id="3" name="SetValue"
           value="[1, 57]"/>
    <param id="4" name="ComputeScalarsOn"
           value="True"/>
  </m_exec>

...

<m_exec id="11"
        name="FileSink"
        package="edu.utah.sci.vistrails.basic"
        version="1.5">
  <param id="15" name="path"
         value="/home/a/results/23.out"/>
</m_exec>
```

FILE NOT FOUND

- Filenames are often the mode of identification in data exploration
- We might also use URIs or access curated data stores
  - Can this always be expected for exploratory tasks?
  - What happens if offline?
- Solution:
  - Managed store for data associated with computations
  - Improved data identification
  - Automatic versioning

Linking Provenance and Data

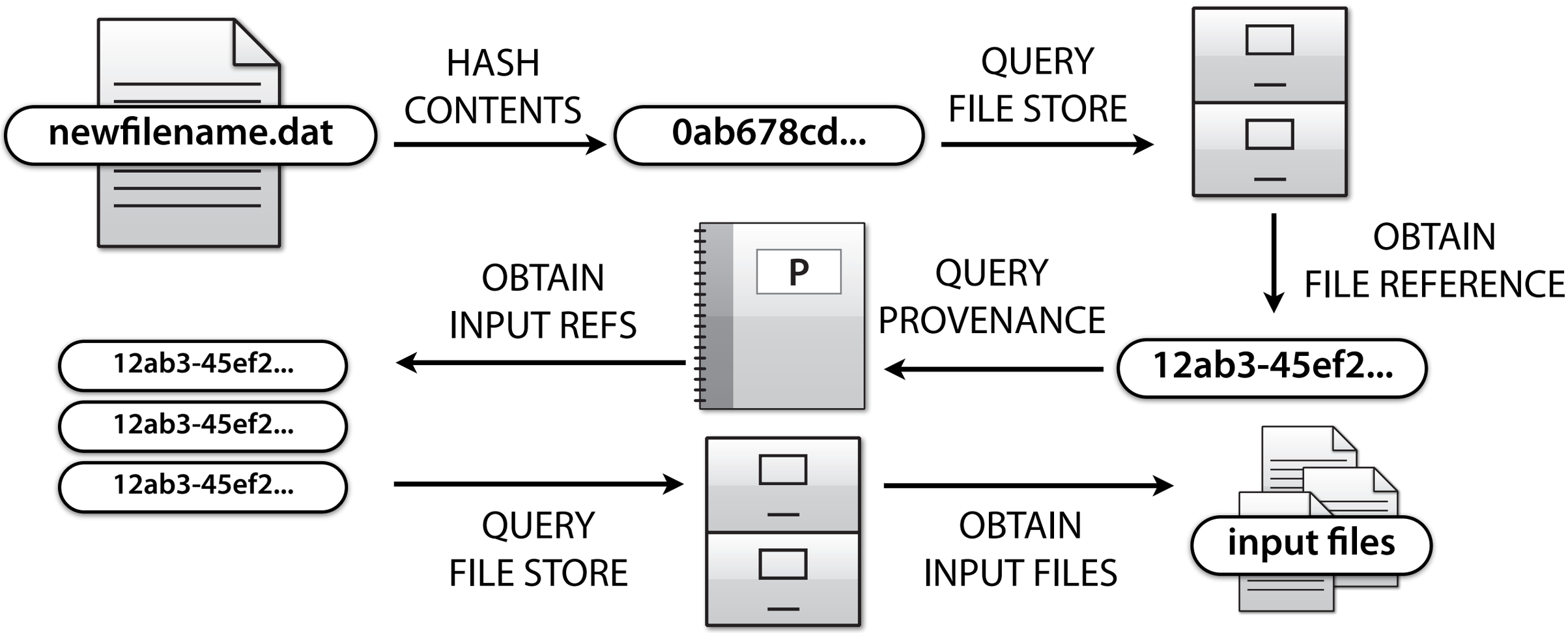
```
<workflow_exec id="1">
  <m_exec id="5"
          name="vtkStructuredDataReader"
          package="edu.utah.sci.vistrails.vtk"
          version="5.6.0">
    <param id="2" name="Se
           value="/MyData      2.dat"/>
  </m_exec>
  <m_exec id="6"
          name="vtkContourFilter"
          package="edu.utah.sci.vistrails.vtk"
          version="5.6.0">
    <param id="3" name="SetValue"
           value="[1, 57]"/>
    <param id="4" name="ComputeScalarsOn"
           value="True"/>
  </m_exec>

...

<m_exec id="11"
        name="FileSink"
        package="edu.utah.sc.vistrails.basic"
        version="1.5">
  <param id="15" name="pa
         value="/home/a      /23.out"/>
</m_exec>
```

FILE NOT FOUND

FILE NOT FOUND

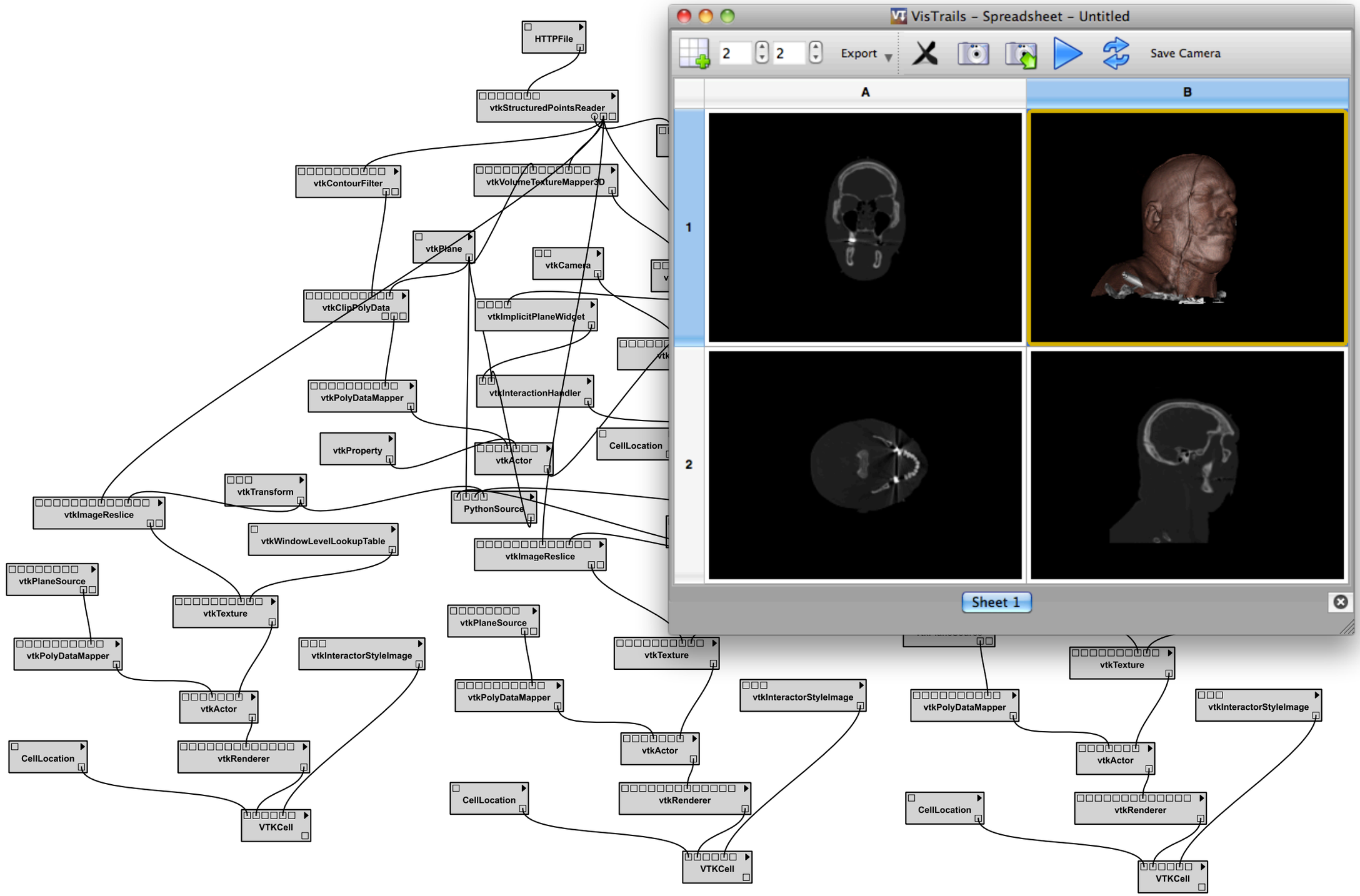
- Filenames are often the mode of identification in data exploration
- We might also use URIs or access curated data stores
  - Can this always be expected for exploratory tasks?
  - What happens if offline?
- Solution:
  - Managed store for data associated with computations
  - Improved data identification
  - Automatic versioning

Full Data Provenance

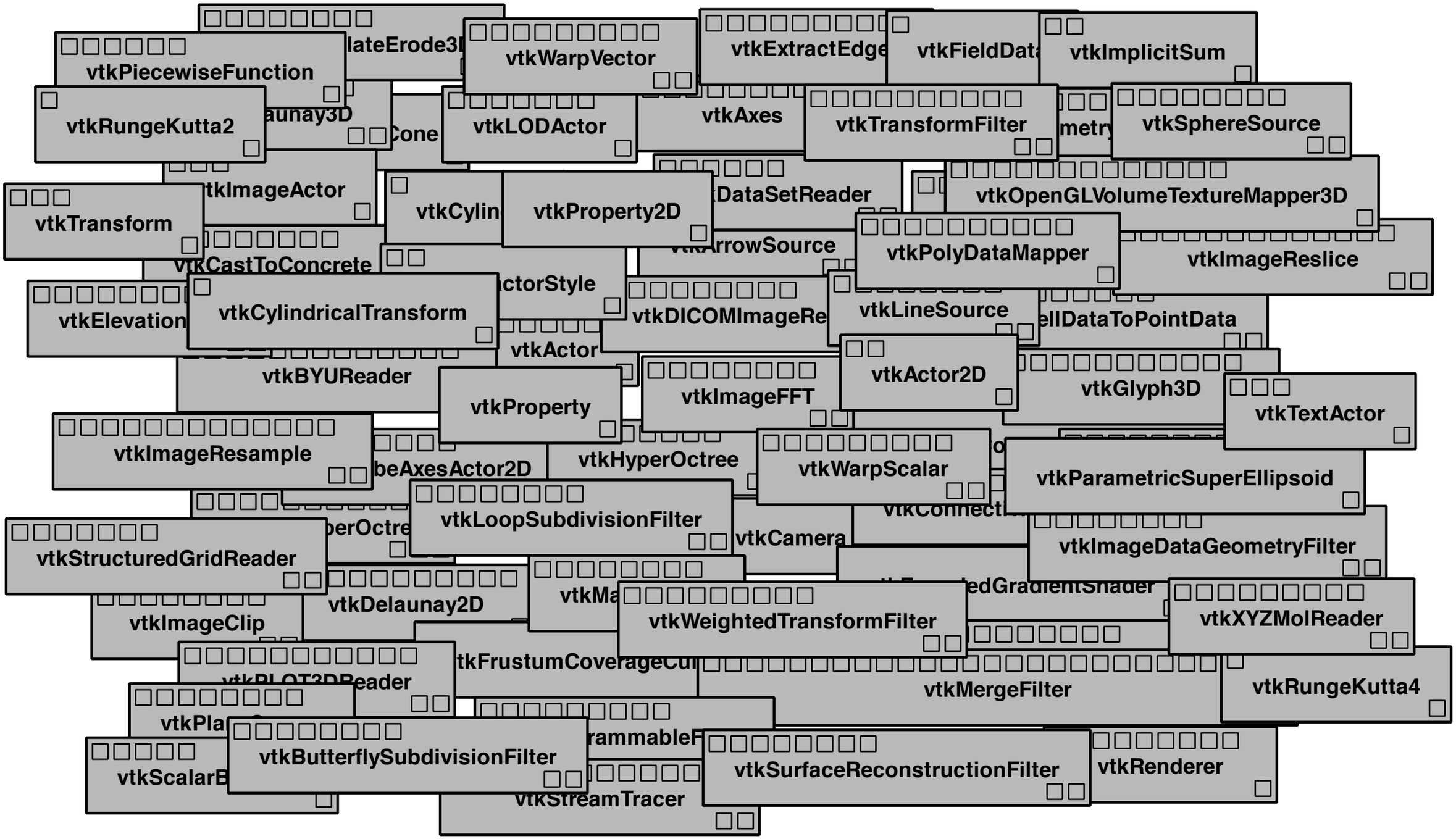


[VisTrails Persistence Package]

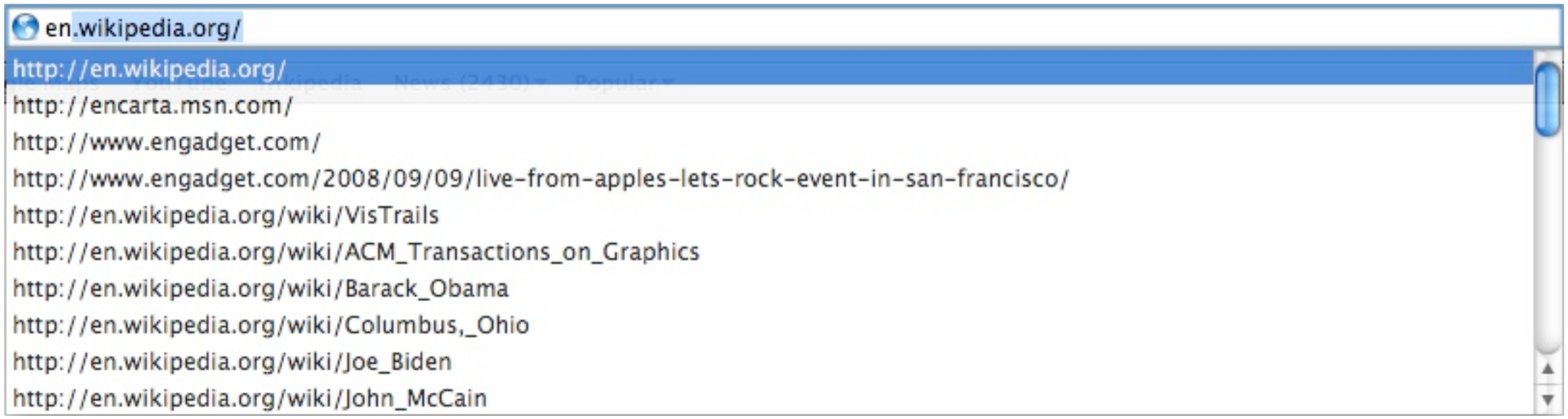
Building Visualization Pipelines



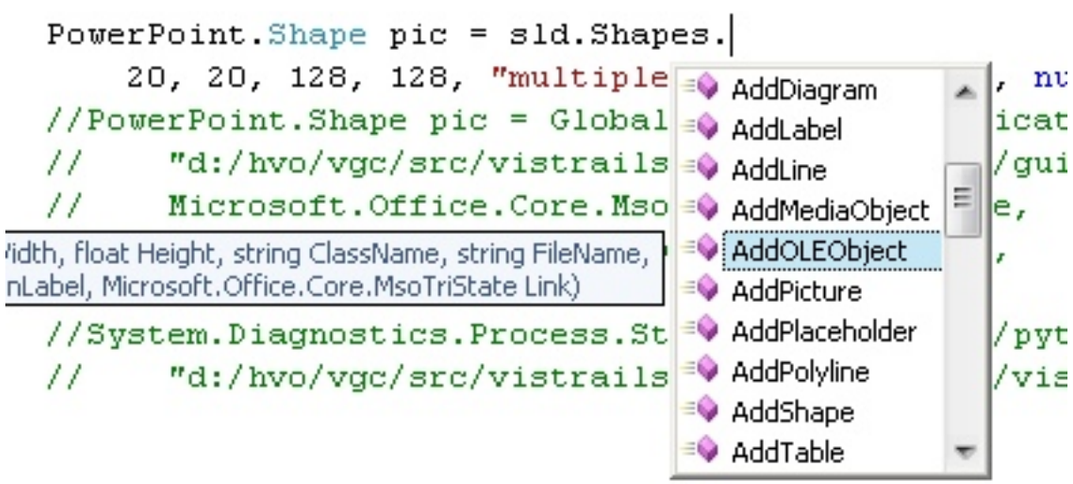
Building Visualization Pipelines



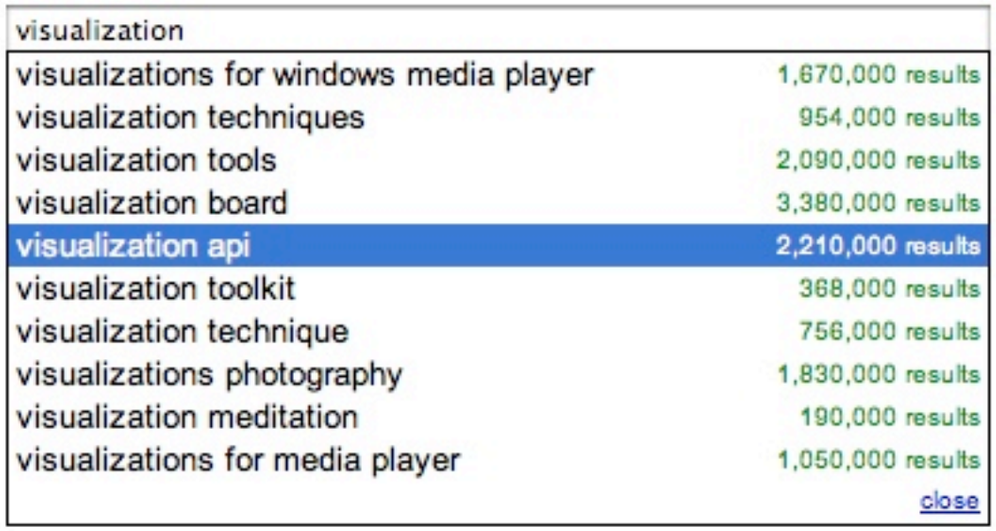
Completions



[URL Completion, Safari]

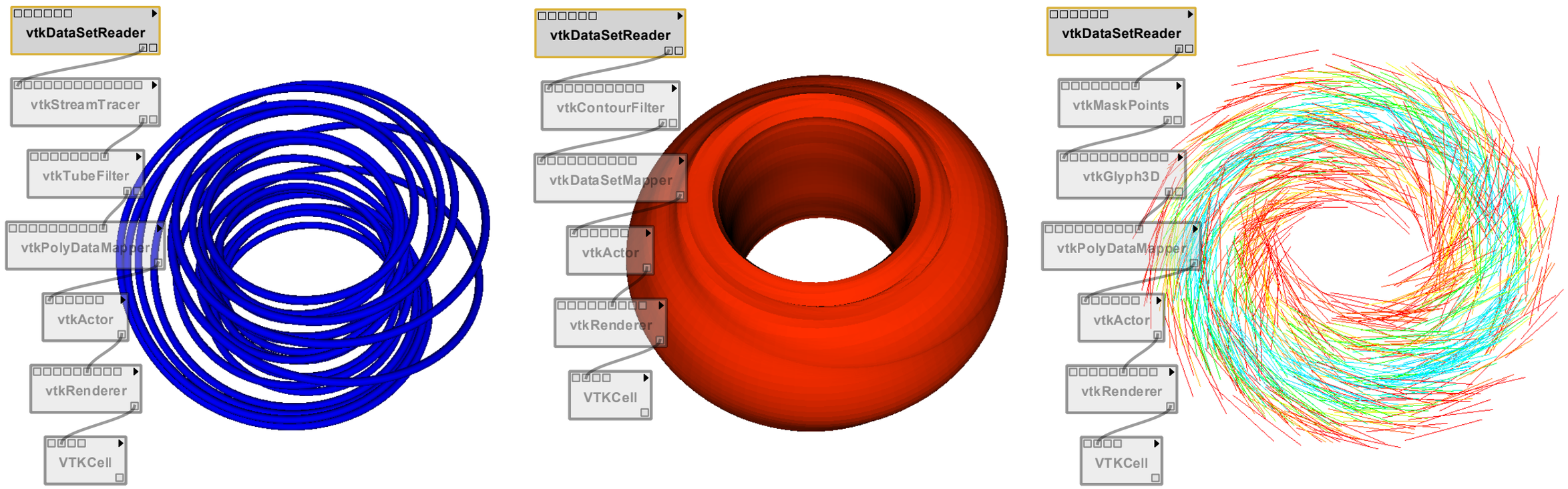


[Code Completion, Intellisense]



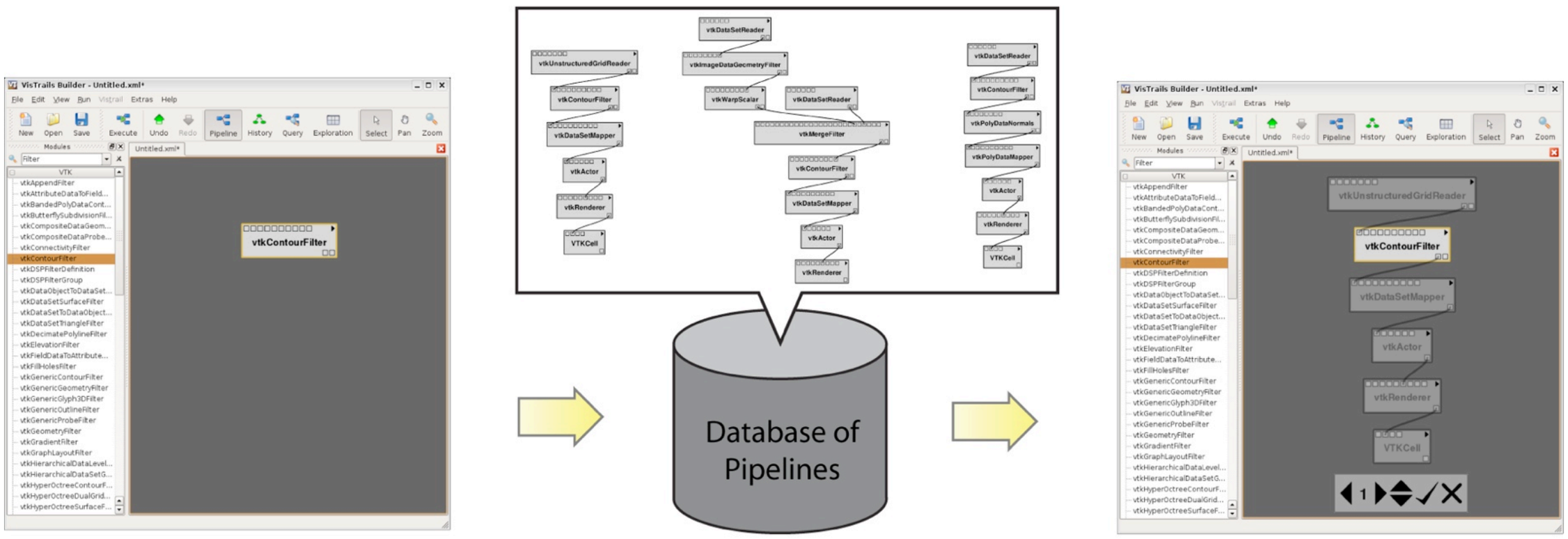
[Web Search Completion, Google]

Visualization Pipeline Completions

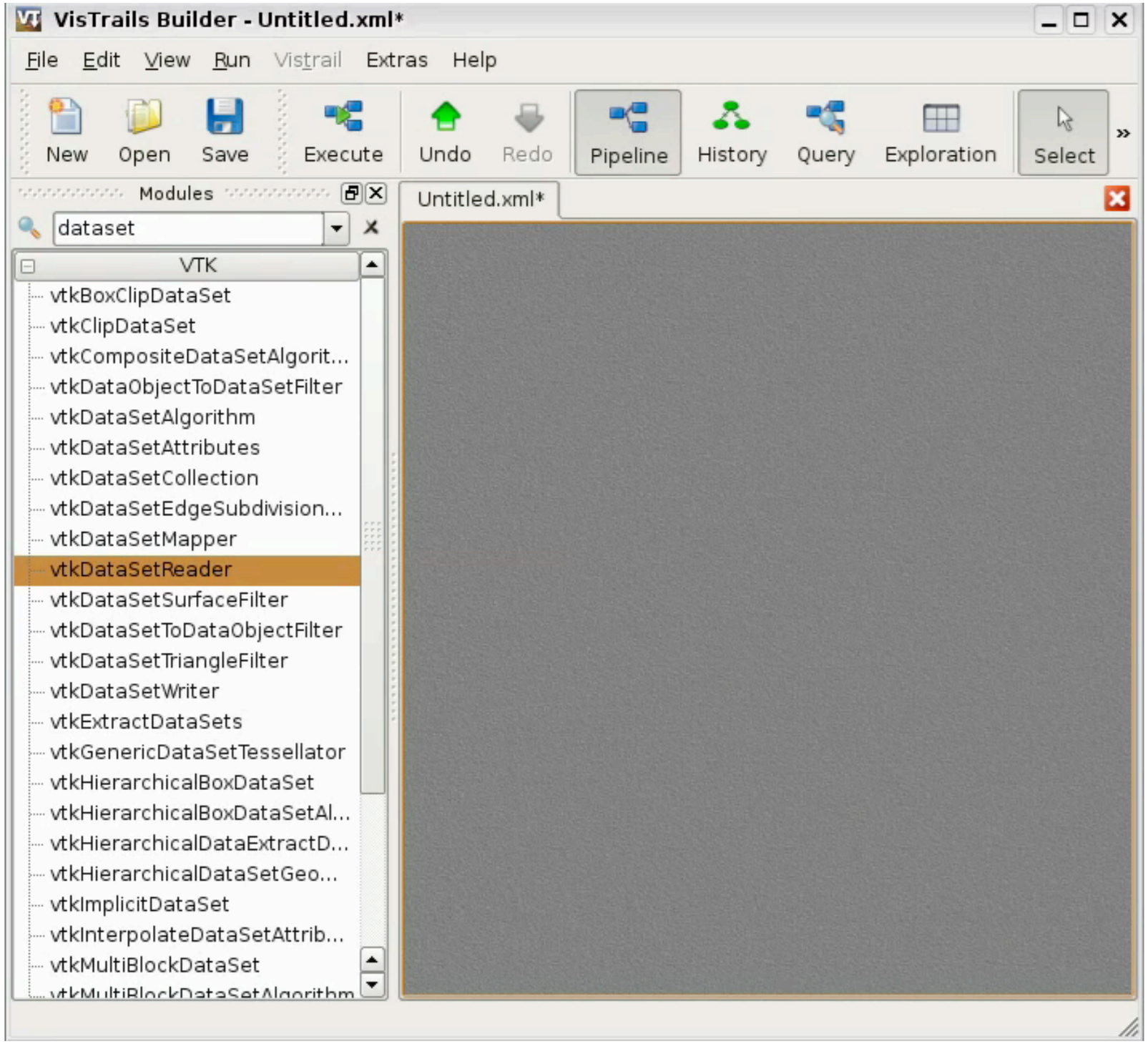


VisComplete Overview

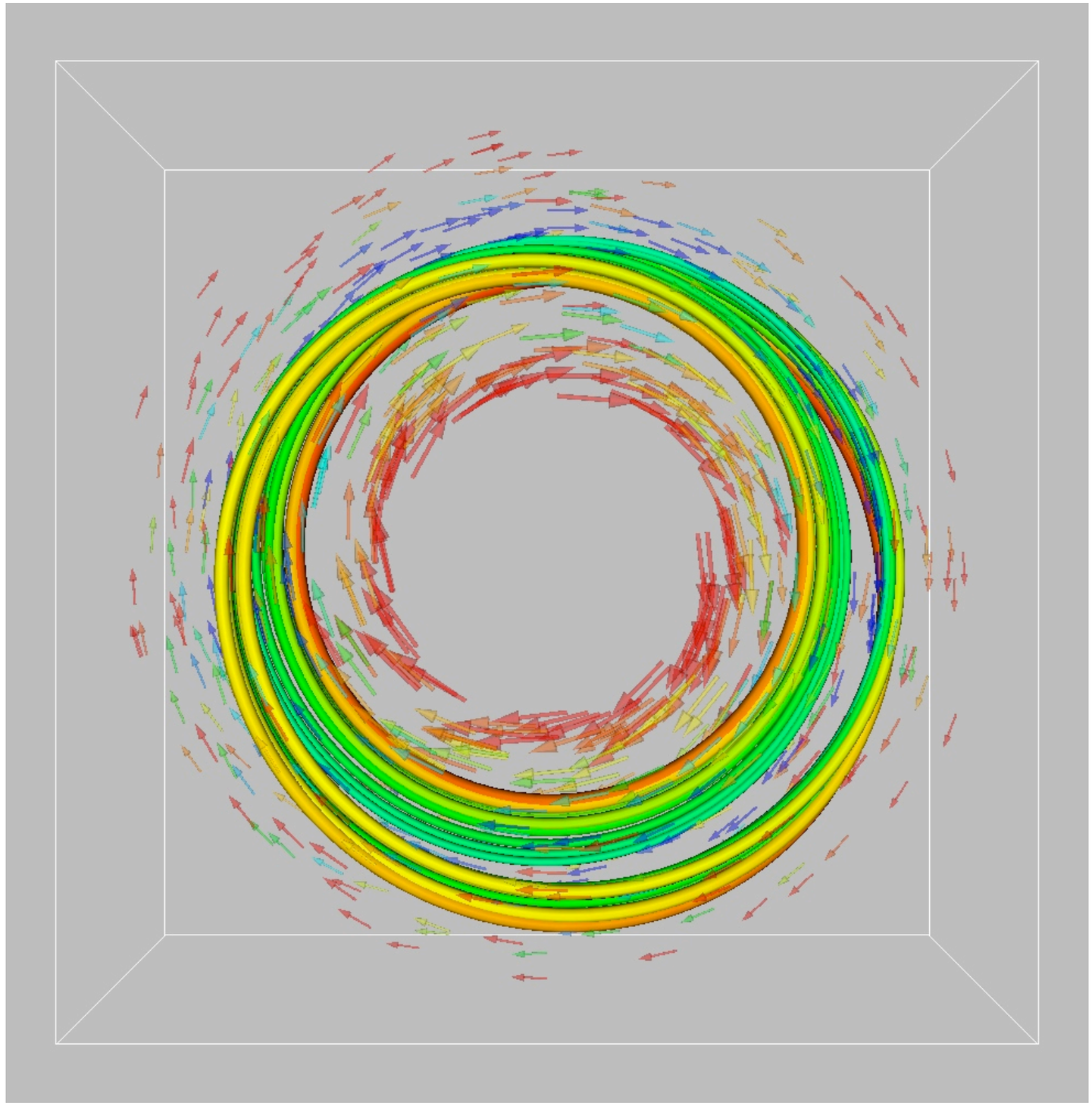
- **Mine** provenance collection: Identify graph fragments that co-occur in a collection of workflows (**Data-Driven**)
- Predict sets of likely workflow additions to a given partial workflow



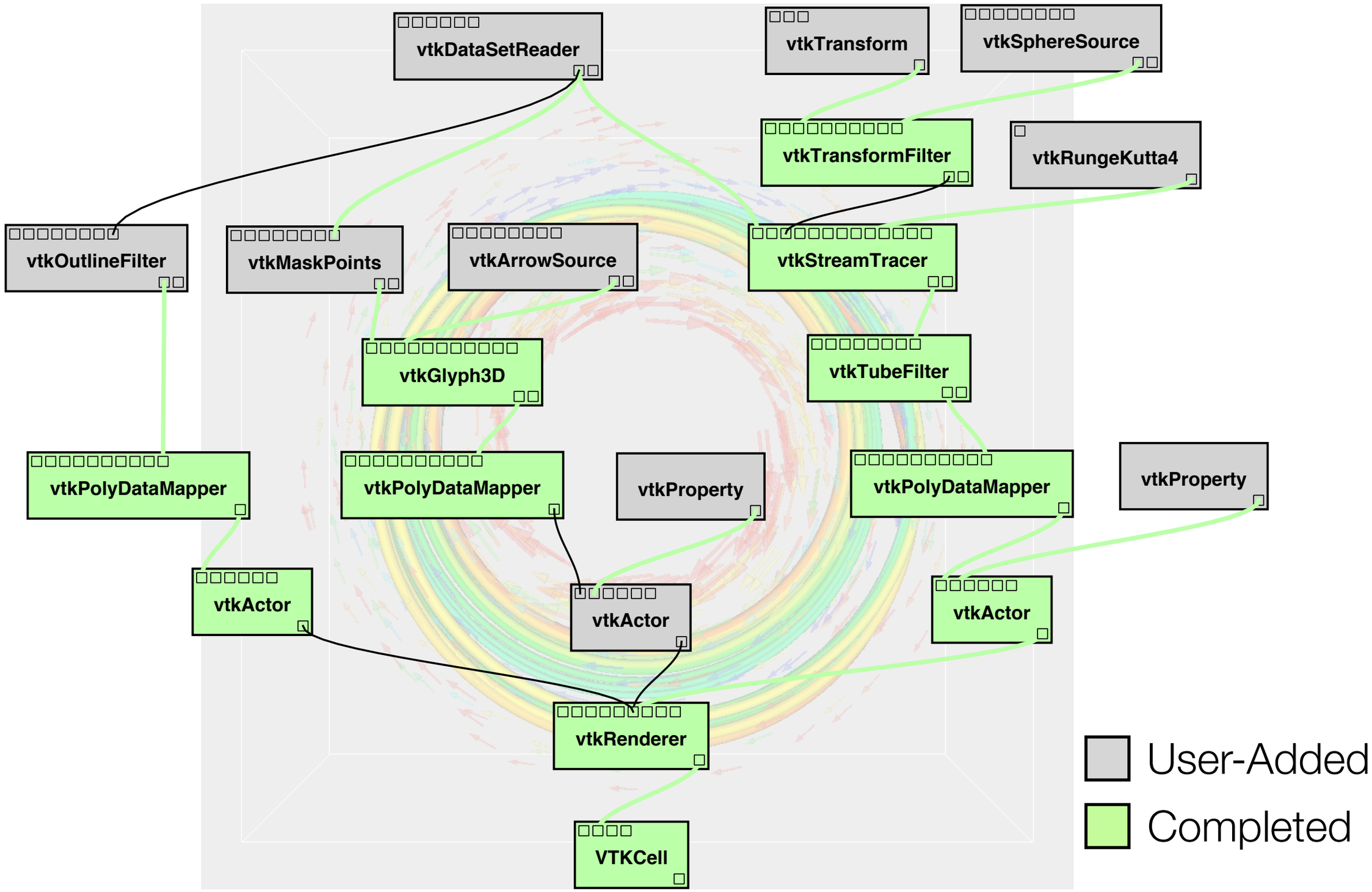
Suggestion Interface



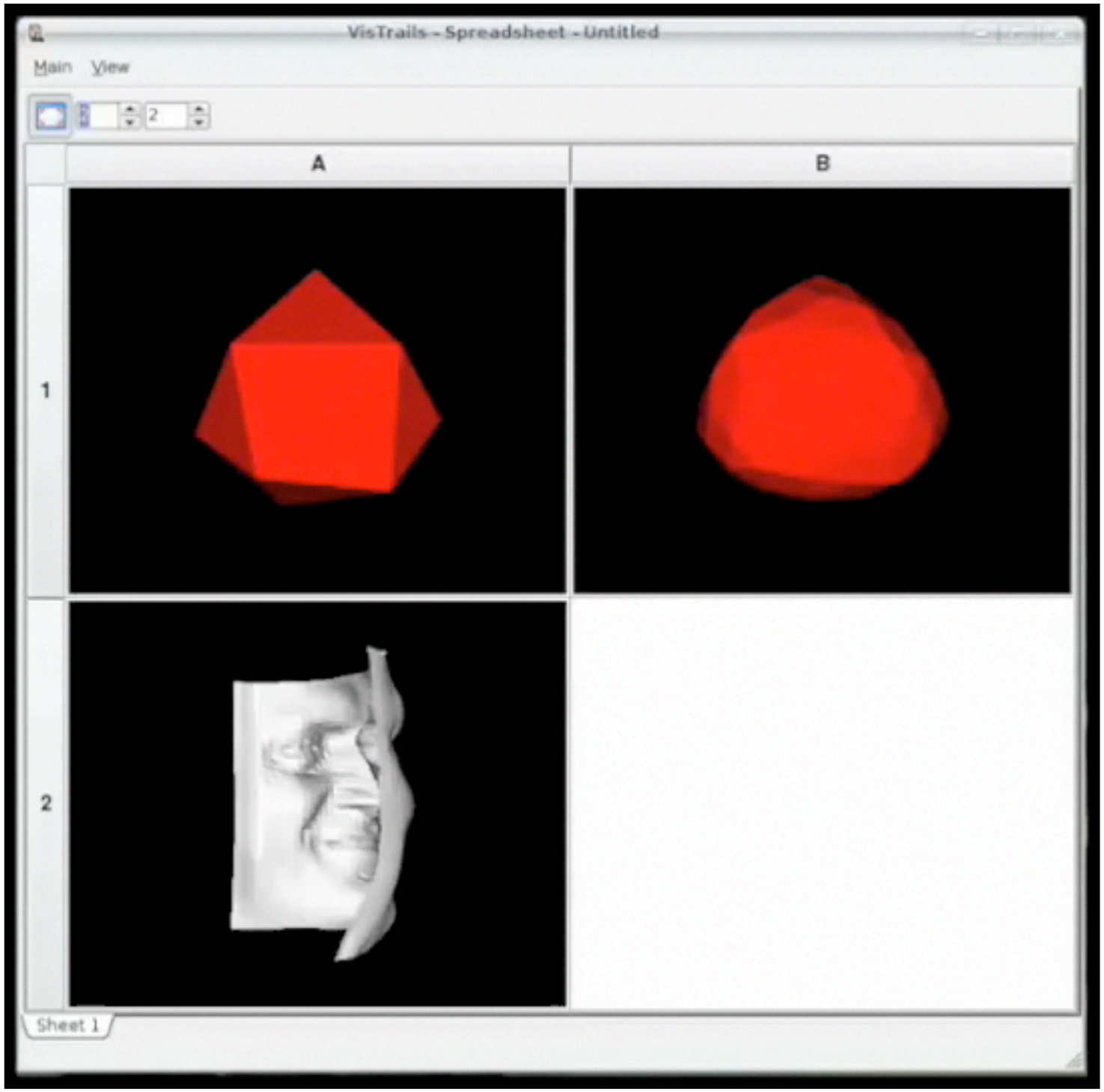
VisComplete Results



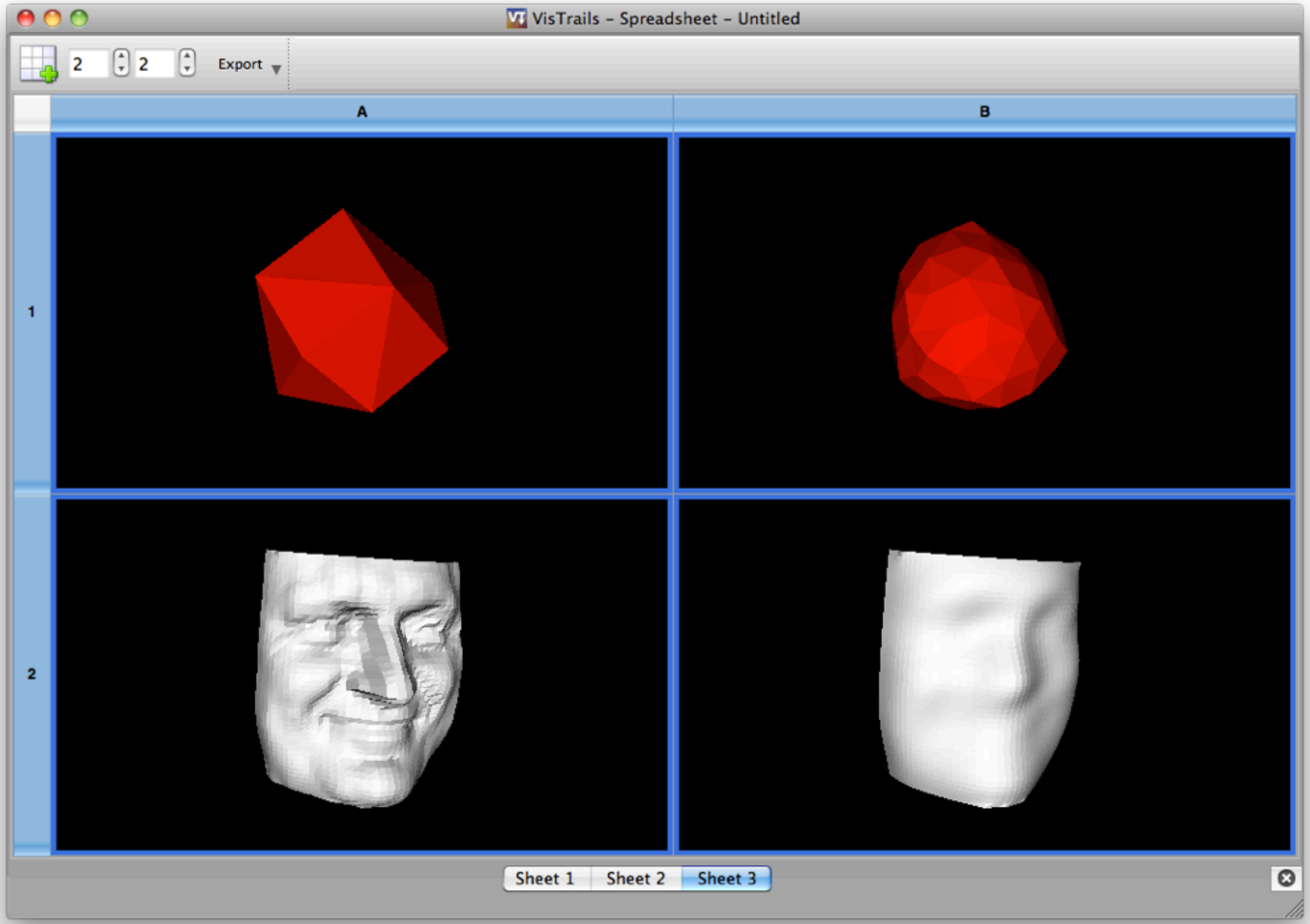
VisComplete Results



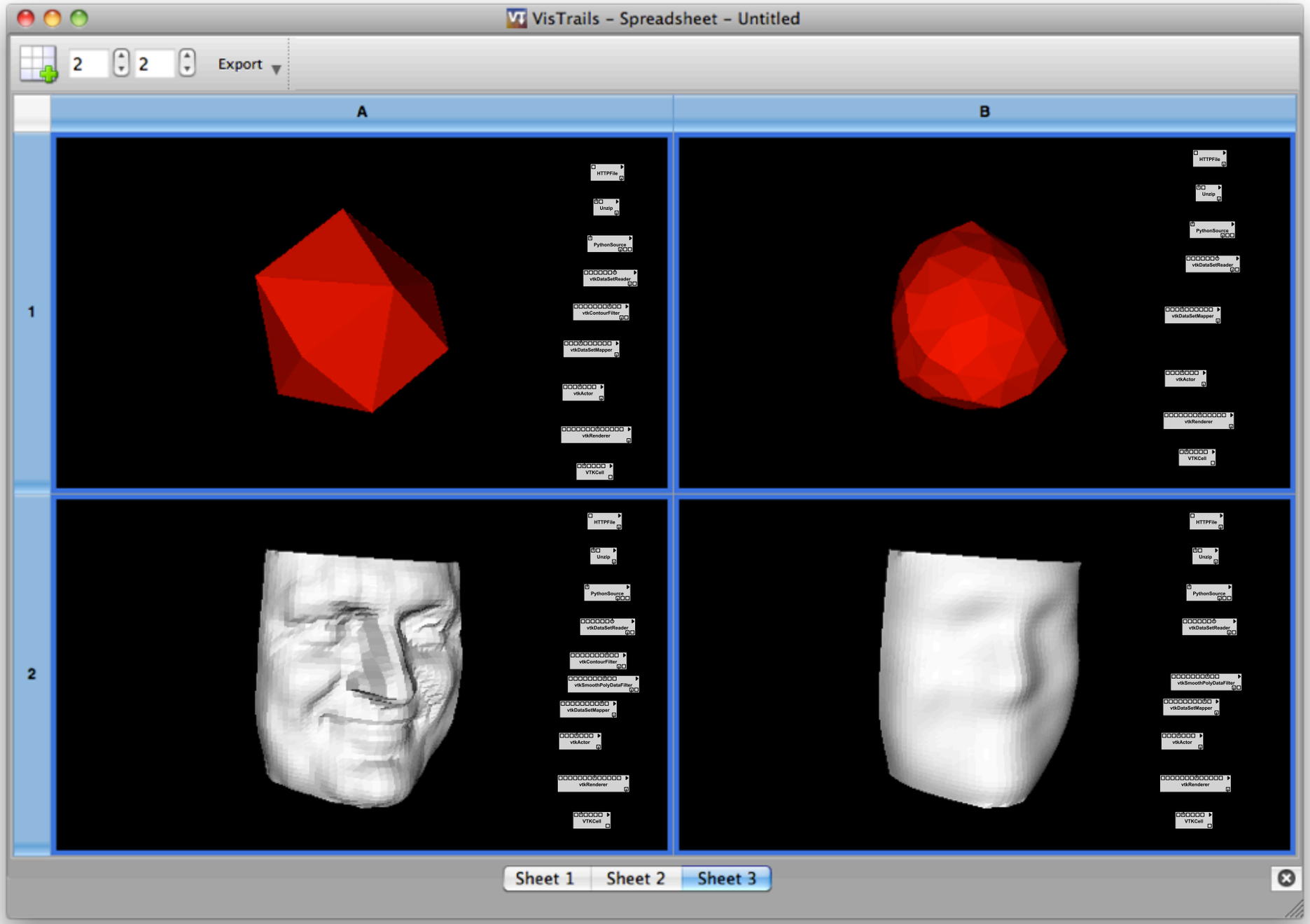
Visualization by Analogy



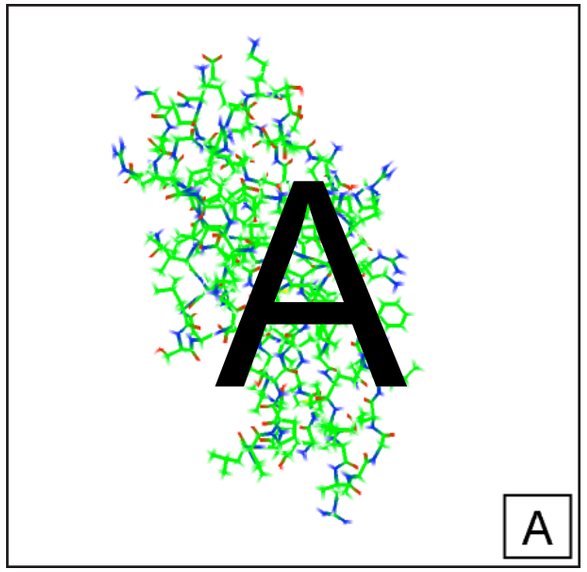
Visualization by Analogy



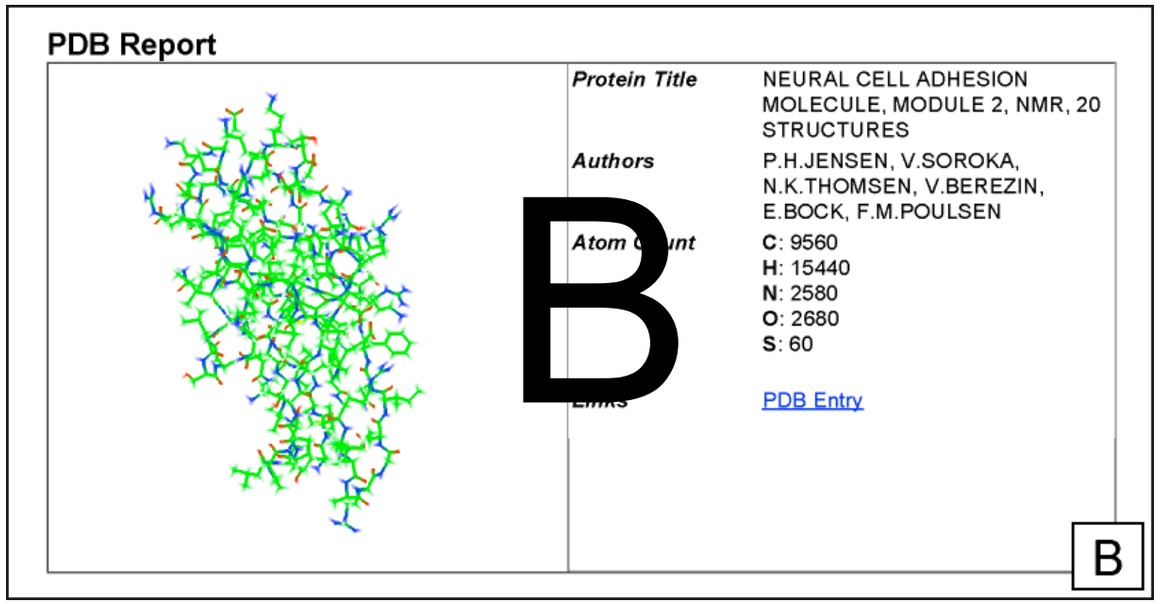
Visualization by Analogy



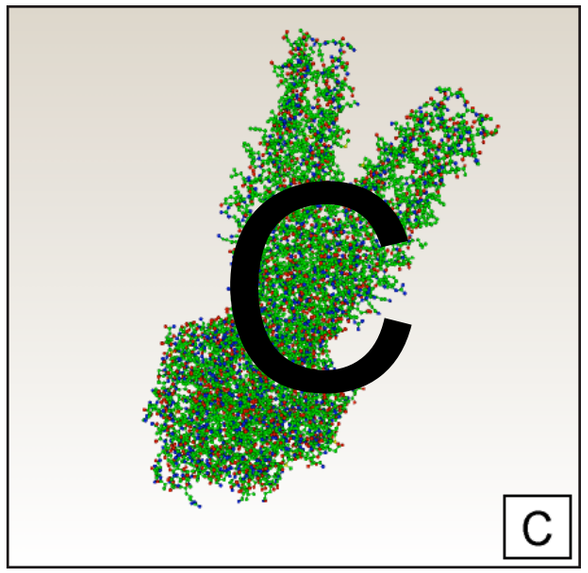
Generating Visualizations by Analogy



is to



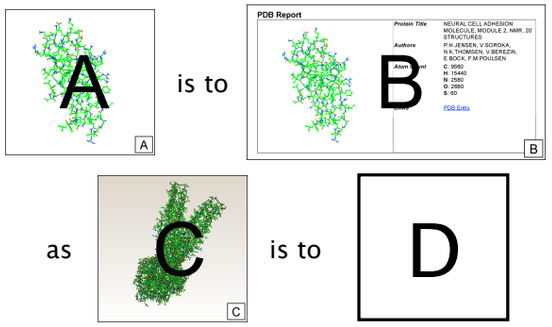
as



is to

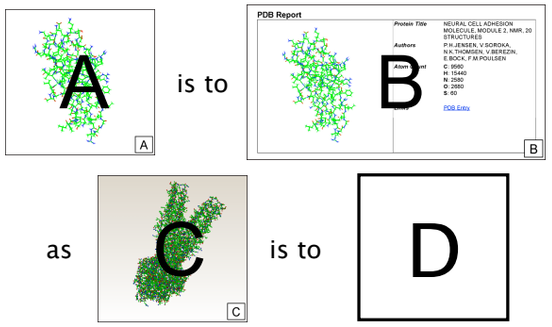
D

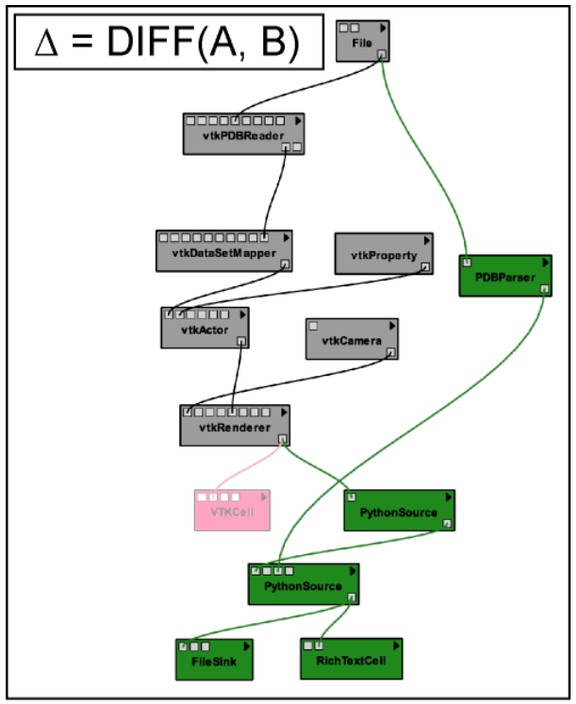
Generating Visualizations by Analogy



Generating Visualizations by Analogy

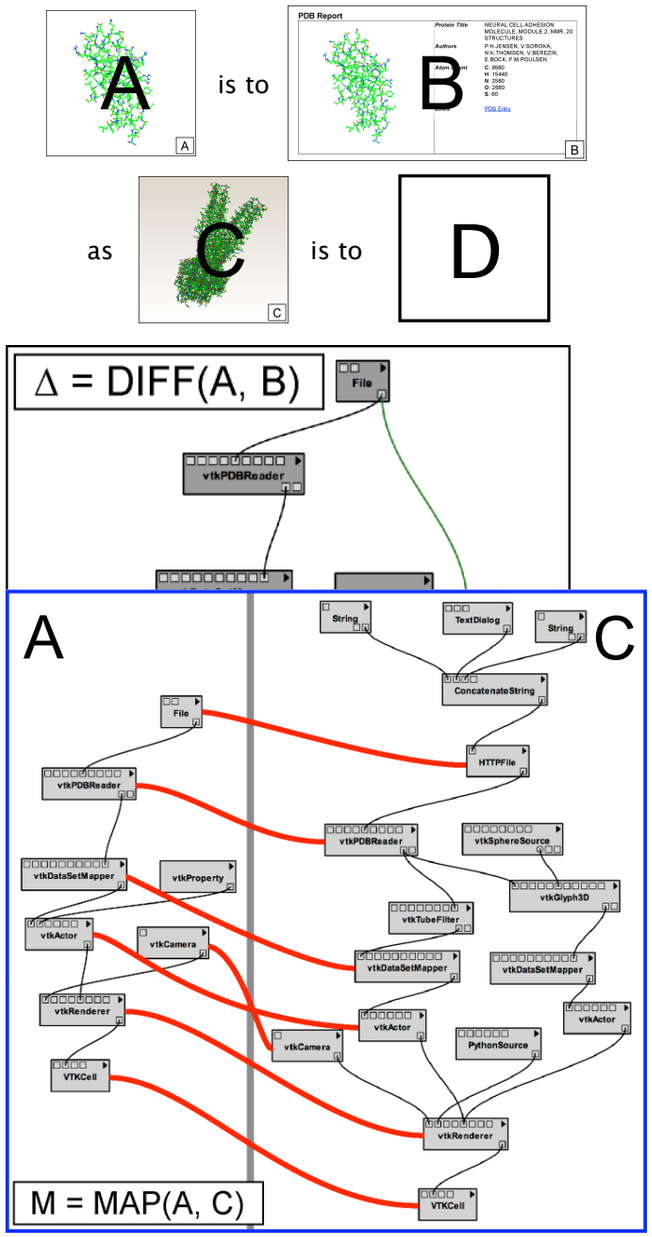
- Compute **difference** Δ(A,B) from provenance
  - D = Δ(A,B) ∘ C is often not a valid workflow





Generating Visualizations by Analogy

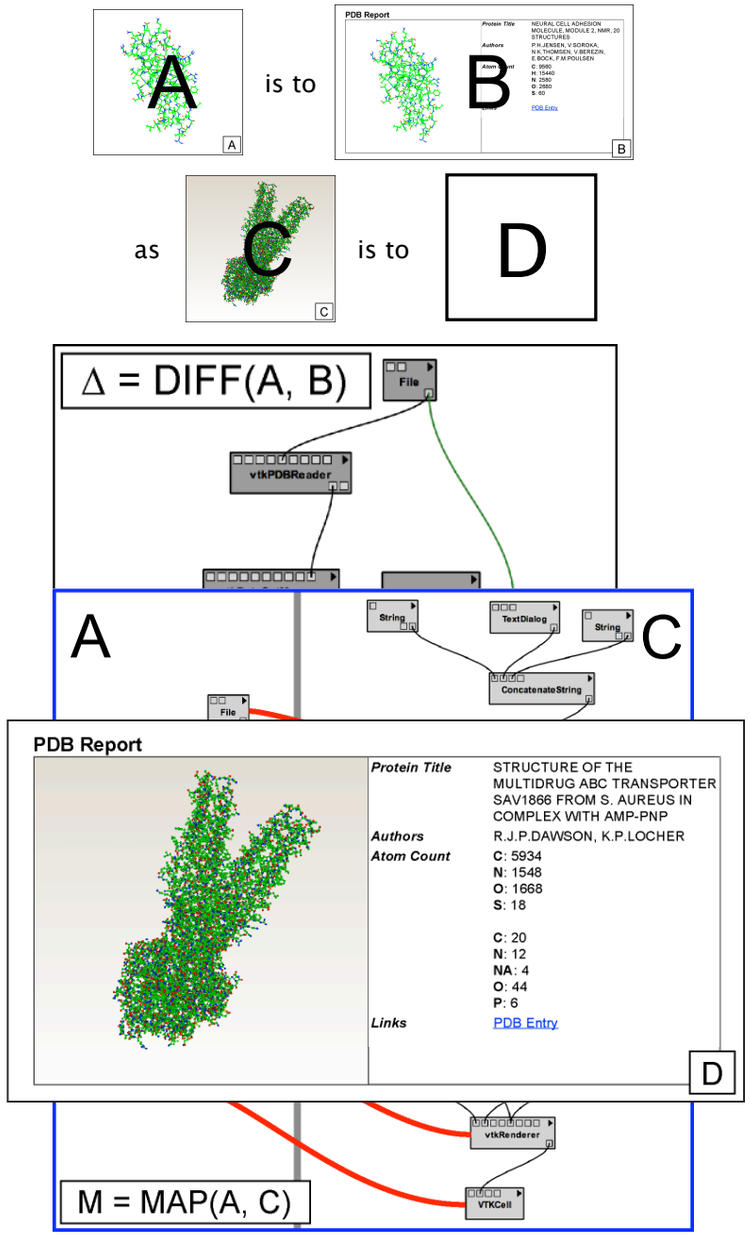
- Compute **difference** Δ(A,B) from provenance
  - D = Δ(A,B) ∘ C is often not a valid workflow
- Find **map** between A & C: map(A,C)



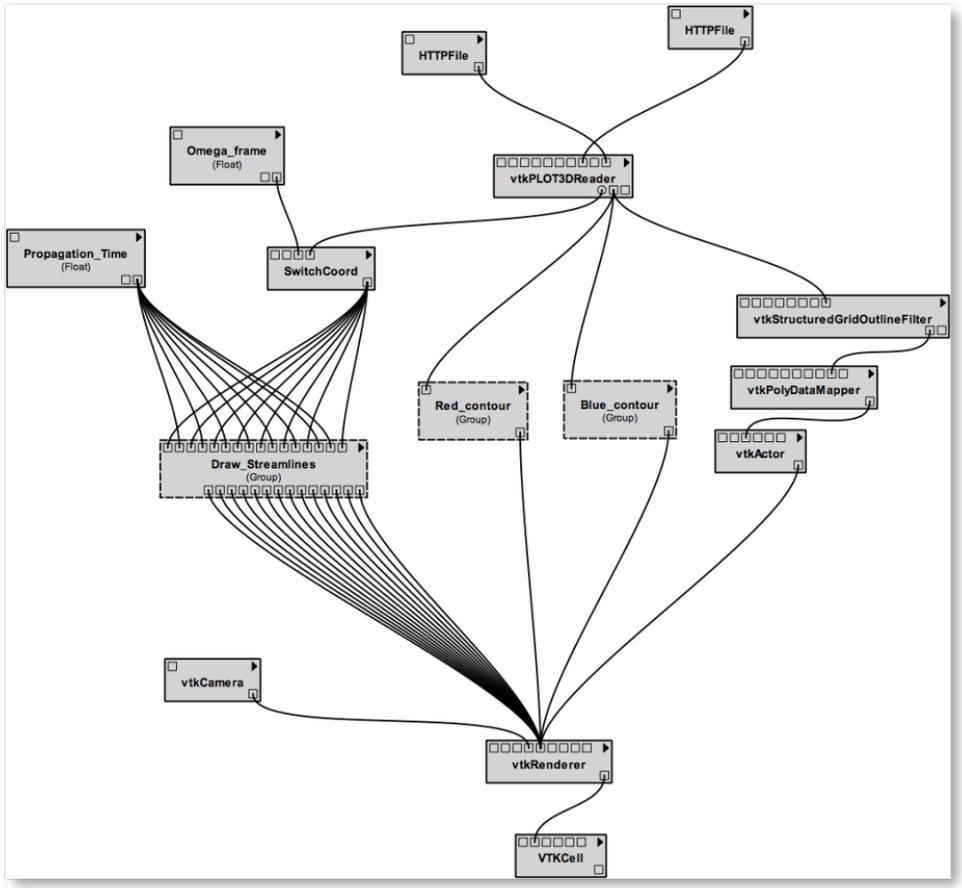
Generating Visualizations by Analogy

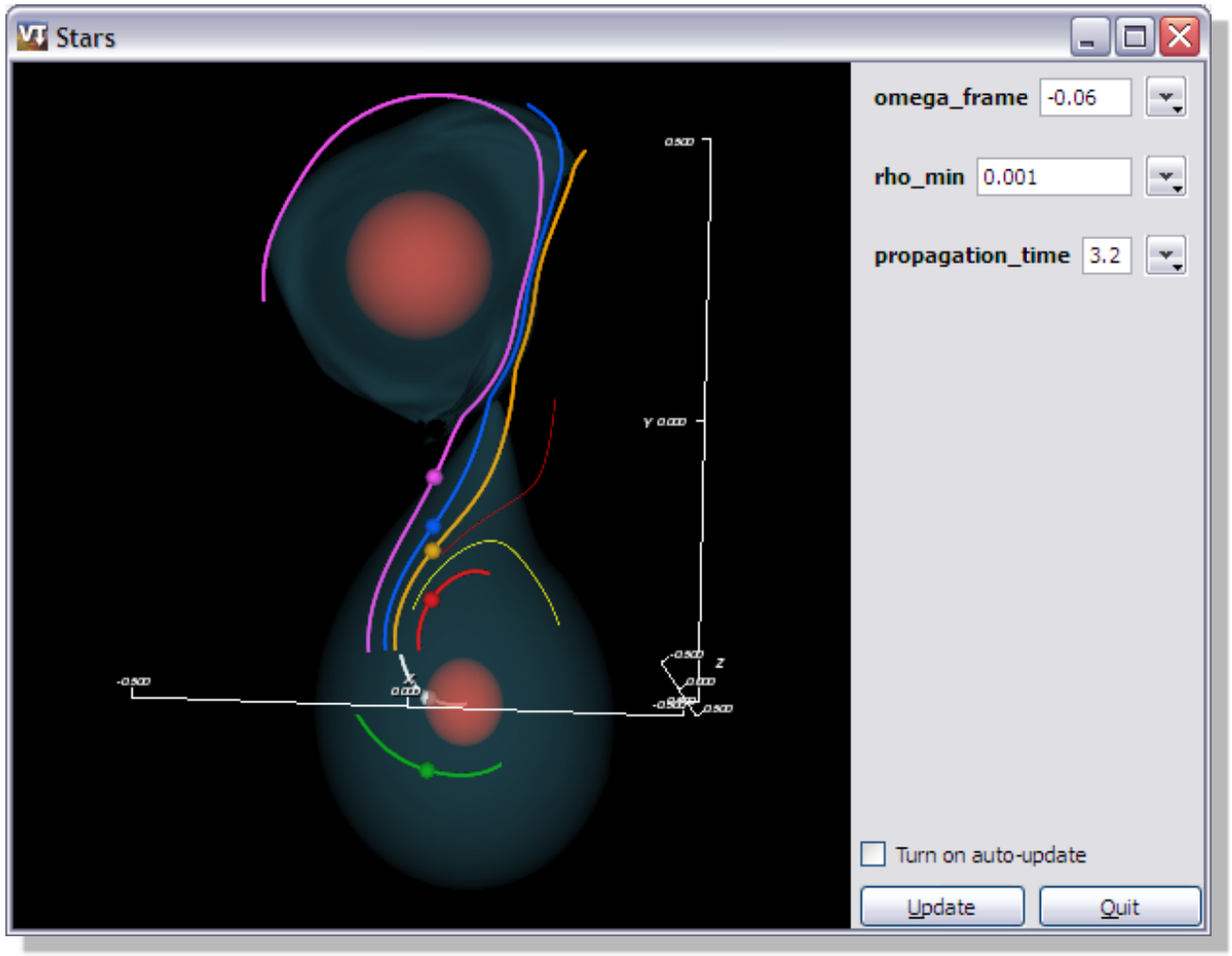
- Compute **difference** Δ(A,B) from provenance
  - D = Δ(A,B) ∘ C is often not a valid workflow
- Find **map** between A & C: map(A,C)
- Compute **mapped difference**

  $\Delta_{AC}(A,B) = map(A,C)\ \Delta(A,B)$
  - D = $\Delta_{AC}(A,B) \circ C$



VisMashup





Acquire and Analyze Pipelines → Create Views (Simplify Pipelines) → Combine Views → App generation and deployment

Provenance in the Wild

Provenance-Rich Publications

arXiv:1106.3267v3 [cond-mat.str-el] 5 Jul 2011

Galois Conjugates of Topological Phases

M. H. Freedman,[1] J. Gukelberger,[2] M. B. Hastings,[1] S. Trebst,[1] M. Troyer,[2] and Z. Wang[1]
[1]*Microsoft Research, Station Q, University of California, Santa Barbara, CA 93106, USA*
[2]*Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland*
(Dated: July 6, 2011)

Galois conjugation relates unitary conformal field theories (CFTs) and topological quantum field theories (TQFTs) to their non-unitary counterparts. Here we investigate Galois conjugates of quantum double models, such as the Levin-Wen model. While these Galois conjugated Hamiltonians are typically non-Hermitian, we find that their ground state wave functions still obey a generalized version of the usual code property (local operators do not act on the ground state manifold) and hence enjoy a generalized topological protection. The key question addressed in this paper is whether such non-unitary topological phases can also appear as the ground states of Hermitian Hamiltonians. Specific attempts at constructing Hermitian Hamiltonians with these ground states lead to a loss of the code property and topological protection of the degenerate ground states. Beyond this we rigorously prove that no local change of basis (IV.5) can transform the ground states of the Galois conjugated doubled Fibonacci theory into the ground states of a topological model whose Hermitian Hamiltonian satisfies Lieb-Robinson bounds. These include all gapped local or quasi-local Hamiltonians. A similar statement holds for many other non-unitary TQFTs. One consequence is that the "Gaffnian" wave function cannot be the ground state of a gapped fractional quantum Hall state.

PACS numbers: 05.30.Pr, 73.43.-f

I. INTRODUCTION

Galois conjugation, by definition, replaces a root of a polynomial by another one with identical algebraic properties. For example, i and $-i$ are Galois conjugate (consider $z^2+1=0$) as are $\phi = \frac{1+\sqrt{5}}{2}$ and $-\frac{1}{\phi} = \frac{1-\sqrt{5}}{2}$ (consider $z^2-z-1=0$), as well as $\sqrt[3]{2}$, $\sqrt[3]{2}e^{2\pi i/3}$, and $\sqrt[3]{2}e^{-2\pi i/3}$ (consider $z^3-2=0$). In physics Galois conjugation can be used to convert non-unitary conformal field theories (CFTs) to unitary ones, and vice versa. One famous example is the non-unitary Yang-Lee CFT, which is Galois conjugate to the Fibonacci CFT $(G_2)_1$, the even (or integer-spin) subset of $\mathrm{su}(2)_3$.

In statistical mechanics non-unitary conformal field theories have a venerable history.[1,2] However, it has remained less clear if there exist physical situations in which non-unitary models can provide a useful description of the low energy physics of a quantum mechanical system – after all, Galois conjugation typically destroys the Hermitian property of the Hamiltonian. Some non-Hermitian Hamiltonians, which surprisingly have totally real spectrum, have been found to arise in the study of PT-invariant one-particle systems[3] and in some Galois conjugate many-body systems[4] and might be seen to open the door a crack to the physical use of such models. Another situation, which has recently attracted some interest, is the question whether non-unitary models can describe 1D edge states of certain 2D bulk states (the edge holographic for the bulk). In particular, there is currently a discussion on whether or not the "Gaffnian" wave function could be the ground state for a *gapped* fractional quantum Hall (FQH) state albeit with a non-unitary "Yang-Lee" CFT describing its edge.[5–7] We conclude that this is not possible, further restricting the possible scope of non-unitary models in quantum mechanics.

We reach this conclusion quite indirectly. Our main thrust is the investigation of Galois conjugation in the simplest non-Abelian Levin-Wen model.[8] This model, which is also called "DFib", is a topological quantum field theory (TQFT) whose states are string-nets on a surface labeled by either a trivial or "Fibonacci" anyon. From this starting point, we give a rigorous argument that the "Gaffnian" ground state cannot be locally conjugated to the ground state of any topological phase, within a Hermitian model satisfying Lieb-Robinson (LR) bounds[9] (which includes but is not limited to gapped local and quasi-local Hamiltonians).

Lieb-Robinson bounds are a technical tool for local lattice models. In relativistically invariant field theories, the speed of light is a strict upper bound to the velocity of propagation. In lattice theories, the LR bounds provide a similar upper bound by a velocity called the LR velocity, but in contrast to the relativistic case there can be some exponentially small "leakage" outside the light-cone in the lattice case. The Lieb-Robinson bounds are a way of bounding the leakage outside the light-cone. The LR velocity is set by microscopic details of the Hamiltonian, such as the interaction strength and range. Combining the LR bounds with the spectral gap enables us to prove locality of various correlation and response functions. We will call a Hamiltonian a *Lieb-Robinson Hamiltonian* if it satisfies LR bounds.

We work primarily with a single example, but it should be clear that the concept of Galois conjugation can be widely applied to TQFTs. The essential idea is to retain the particle types and fusion rules of a unitary theory but when one comes to writing down the algebraic form of the F-matrices (also called $6j$ symbols), the entries are now Galois conjugated. A slight complication, which is actually an asset, is that writing an F-matrix requires a gauge choice and the most convenient choice may differ before and after Galois conjugation.

Our method is not restricted to Galois conjugated $\mathrm{DFib}^{\mathcal{G}}$ and its factors $\mathrm{Fib}^{\mathcal{G}}$ and $\overline{\mathrm{Fib}}^{\mathcal{G}}$, but can be generalized to infinitely many non-unitary TQFTs, showing that they will not arise as low energy models for a gapped 2D quantum mechan-

non-Hermitian DYL model

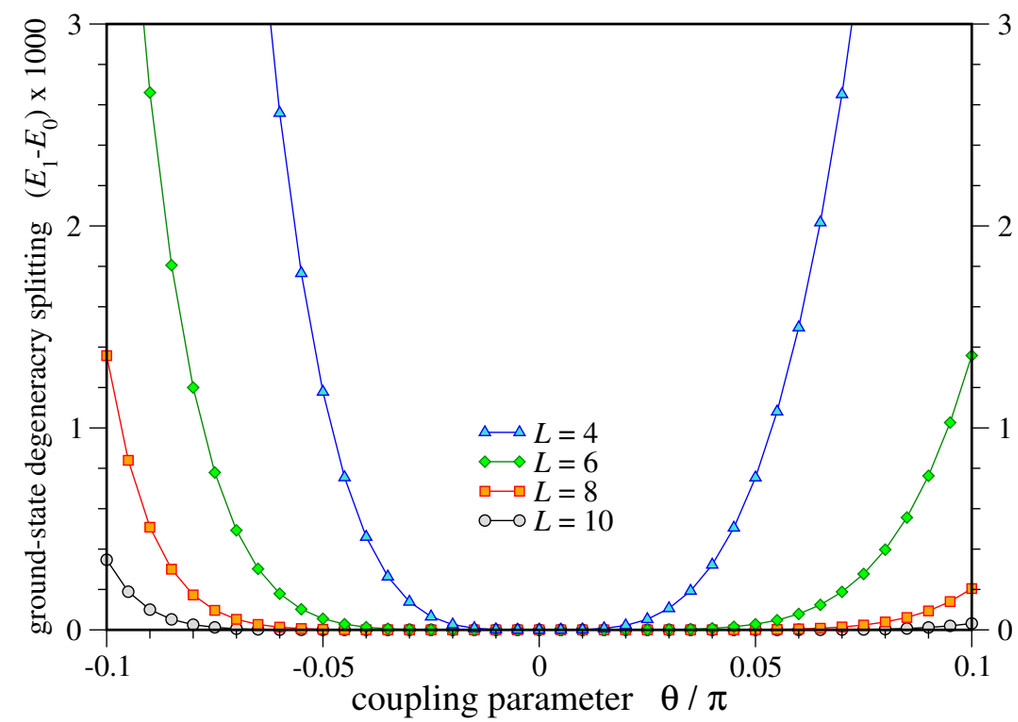


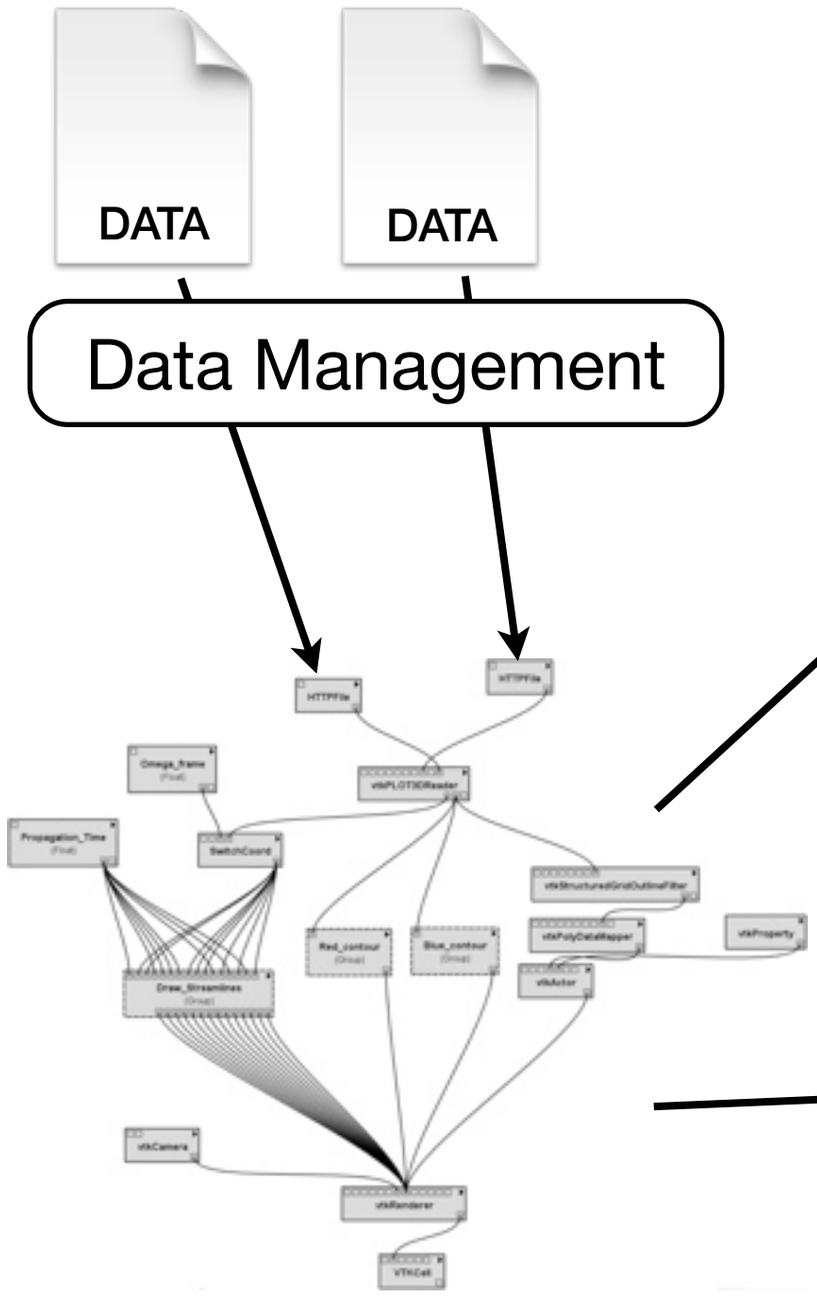
FIG. 6. (color online) Ground-state degeneracy splitting of the non-Hermitian doubled Yang-Lee model when perturbed by a string tension ($\theta \neq 0$).

Galois Conjugates of Topological Phases
[Freedman et al., 2012]

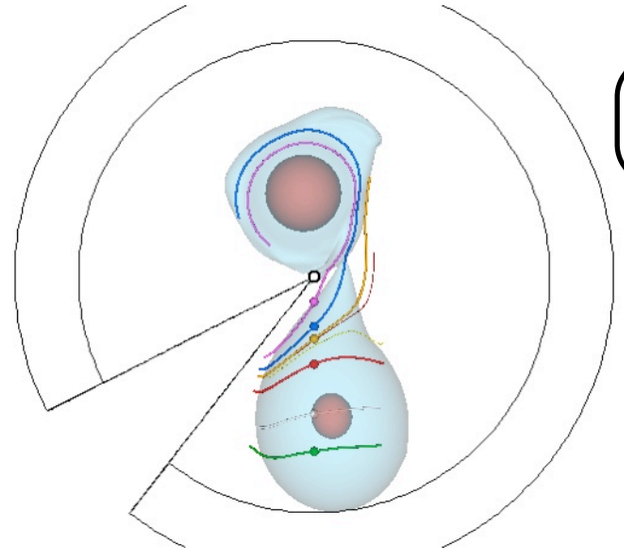
Benefits of Provenance-Rich Publications

- Produce more **knowledge**–not just text
- Allow scientists to stand on the shoulders of giants (and their own)
- Science can move **faster**!
- **Higher-quality** publications
- Authors will be more **careful**
- Many eyes to **check** results
- Describe more of the discovery process: people only describe successes, can we **learn from mistakes**?
- Expose users to different techniques and tools: expedite their training; and potentially **reduce their time to insight**

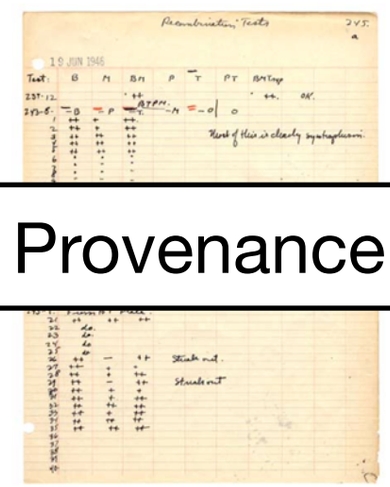
Vision: Provenance-Rich Science



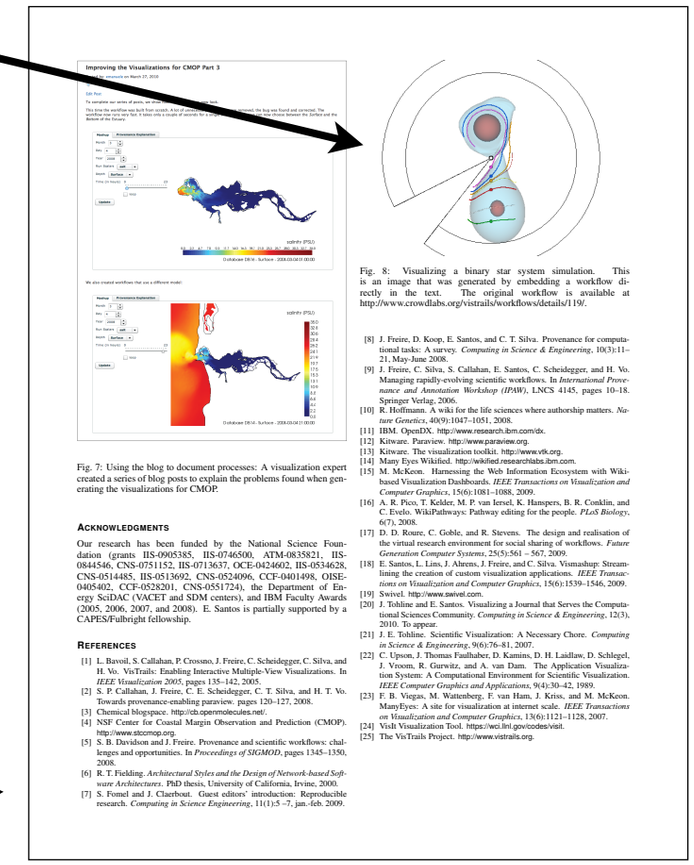
Computation



Visualization



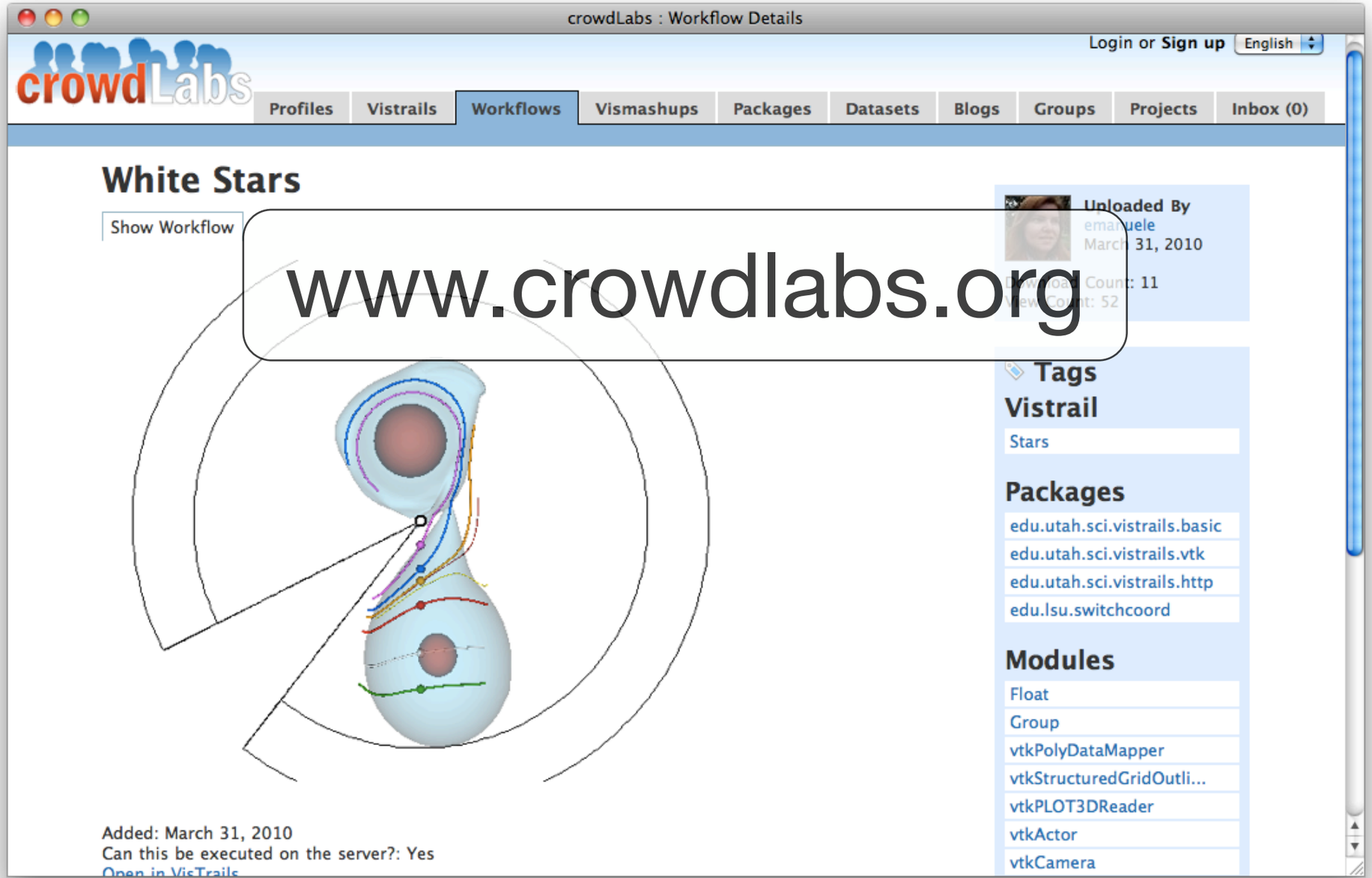
Provenance



Publishing

Provenance and the Web

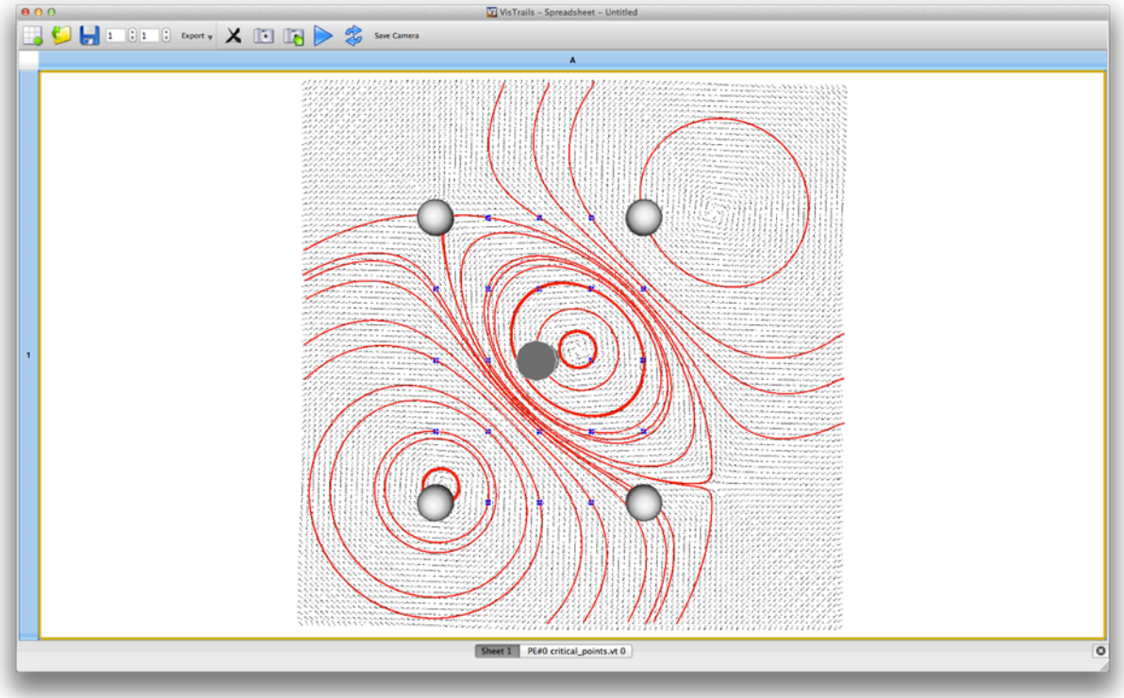
- CrowdLabs: share approaches, results, data, and explorations
- Support visualization, online exploration



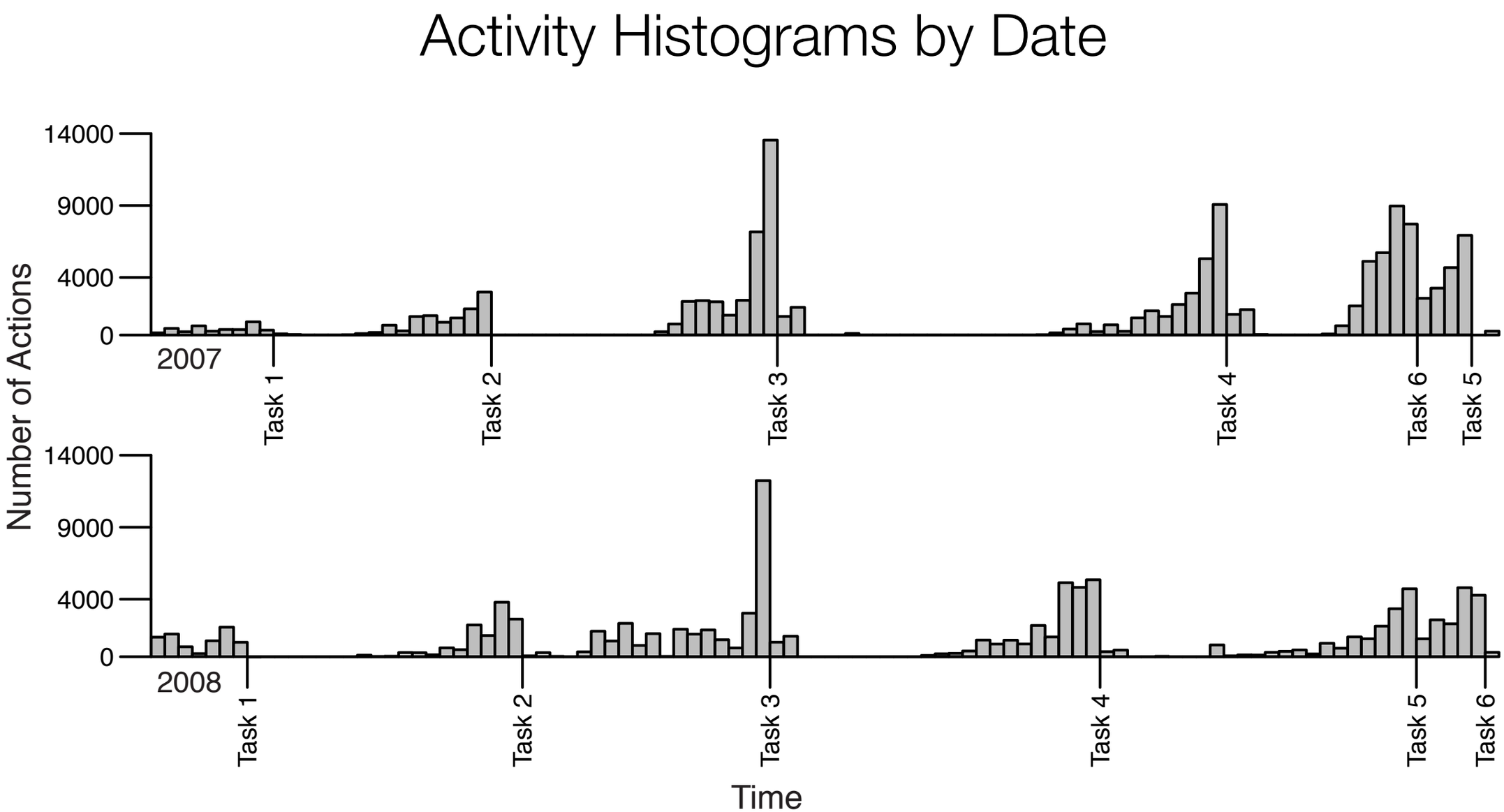
www.crowdlabs.org

VisTrails for Teaching Scientific Visualization

- "Using VisTrails and Provenance for Teaching Scientific Visualization" [Silva et al., Eurographics Educator Program, 2010]
- Same features that scientists use for exploratory tasks can also benefit students
  - Exploration: see all pipelines not just a "final" one
  - Comparison: see different pipelines and what changes exist
  - Assessment: see how a solution was developed

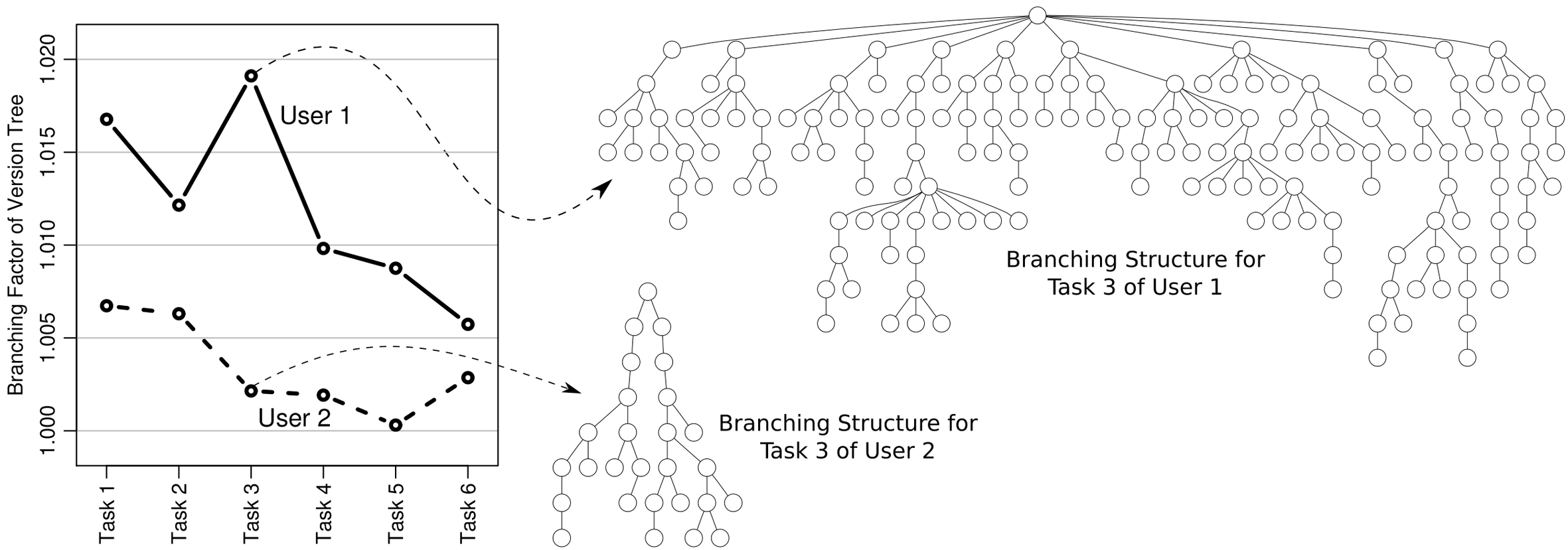


Provenance Analysis of Projects



Provenance Analysis of Projects

Comparing Paths to Solutions for Two Students



Conclusion

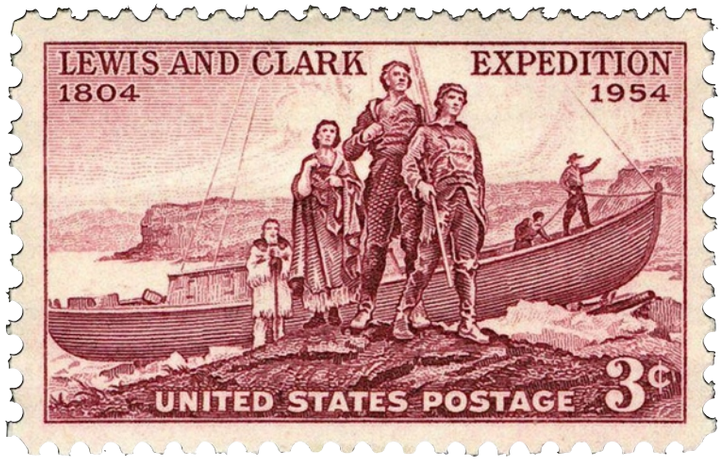
Orchestration

Abstraction

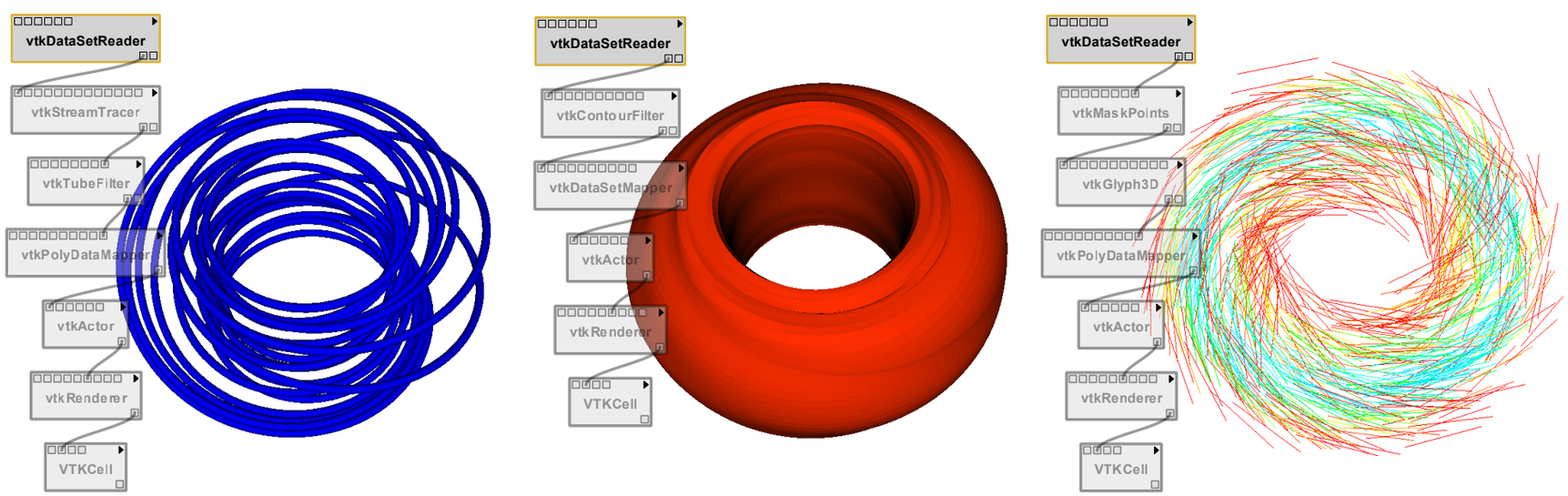
Quality

Reproducibility

Reuse



Exploration

Questions

www.vistrails.org