CIS 602-01: Computational Reproducibility

Provenance

Dr. David Koop
Provenance Capture Mechanisms

• Workflow-based
  - Since workflow execution is controlled, keep track of all the workflow modules, parameters, etc. as they are executed

• Process-based
  - Each process is required to write out its own provenance information (not centralized like workflow-based)

• OS-based
  - The OS or filesystem is modified so that any activity it does is monitored and the provenance subsystem organizes it

• Tradeoffs:
  - Workflow- and process-based have better abstraction, OS-based requires minimal user effort once installed and can capture "hidden dependencies"
Prospective and Retrospective Provenance

• Recipe for baking a cake versus the actual process & outcome
• Prospective provenance is what was specified/intended
  - a workflow, script, list of steps
• Retrospective provenance is what actually happened
  - actual data, actual parameters, errors that occurred, timestamps, machine information
• Do not need prospective provenance to have retrospective provenance!
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
Workflow Evolution Provenance of MTA Fare Data

- initial data
  - corrected data
    - November ff
    - November 2 data
- station locations
  - August 16 Tab
  - Full fares map
- added fares
  - Station map
  - Broadway line
  - Broadway diff map
- difference
  - August 16 with labels
  - Concourse line
  - Filtered
  - Heatmap

D. Koop, CIS 602-01, Fall 2016
Workflow Evolution Provenance of MTA Fare Data
Workflow Evolution Provenance

with labels

filtered

HTTPFile
CSVFile
JSONFile
JoinTables
SheetReference
SelectFromTable
CellLocation
ProjectTable
GMapCell

HTTPFile
CSVFile
JSONFile
JoinTables
SelectFromTable
SelectFromTable
ProjectTable
GMapCircleCell
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 “latitude > 40.6”
  - delete parameter “float_expr” from “SelectFromTable”
  - add parameter “float_expr” to “SelectFromTable” with value “latitude > 40.7”
  - delete parameter “float_expr” from “SelectFromTable”
  - add parameter “float_expr” to “SelectFromTable” with value “latitude > 40.8”

...
Execution Provenance
Execution Provenance

```xml
<module id="12" name="vtkDataSetReader"
  start_time="2010-02-19 11:01:05"
  end_time="2010-02-19 11:01:07">
  <annotation key="hash"
    value="c54bea63cb7d912a43ce"/>
</module>

<module id="13" name="vtkContourFilter"
  start_time="2010-02-19 11:01:07"
  end_time="2010-02-19 11:01:08"/>

<module id="15" name="vtkDataSetMapper"
  start_time="2010-02-19 11:01:09"
  end_time="2010-02-19 11:01:12"/>

<module id="16" name="vtkActor"
  start_time="2010-02-19 11:01:12"
  end_time="2010-02-19 11:01:13"/>

<module id="17" name="vtkCamera"
  start_time="2010-02-19 11:01:13"
  end_time="2010-02-19 11:01:14"/>

<module id="18" name="vtkRenderer"
  start_time="2010-02-19 11:01:14"
  end_time="2010-02-19 11:01:14"/>
...
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 MplFigure → 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
...
Full Data Provenance

- newfilename.dat
  - HASH CONTENTS
    - 0ab678cd...
      - QUERY FILE STORE
        - OBTAIN FILE REFERENCE
          - 12ab3-45ef2...
            - QUERY PROVENANCE
              - INPUT FILES
                - input files
        - OBTAIN INPUT FILES
          - 12ab3-45ef2...
            - INPUT REFERENCES
              - 12ab3-45ef2...
                - HASH CONTENTS
                  - 0ab678cd...
                    - QUERY FILE STORE
                      - OBTAIN FILE REFERENCE
                        - 12ab3-45ef2...
                          - QUERY PROVENANCE
                            - INPUT FILES
                              - input files

[VisTrails Persistence Package]
Visualization Completions

• 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
Generating Visualizations by Analogy

A is to B as C is to D

PDB Report

<table>
<thead>
<tr>
<th>Protein Title</th>
<th>NEURAL CELL ADHESION MOLECULE, MODULE 2, NMR, 20 STRUCTURES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>P.H. JENSEN, V.SOROKA, N.K.TROMSEN, V.BEREZIN, E.BOCK, F.M.POUlsen</td>
</tr>
<tr>
<td>Atom Count</td>
<td>C: 9560, H: 15440, N: 2580, O: 2680, S: 60</td>
</tr>
<tr>
<td>Links</td>
<td>PDB Entry</td>
</tr>
</tbody>
</table>

D. Koop, CIS 602-01, Fall 2016
Generating Visualizations by Analogy

A is to B as C is to D
Generating Visualizations by Analogy

- Compute difference $\Delta(A,B)$ from provenance
  
  - $D = \Delta(A,B) \circ C$ is often not a valid workflow
Generating Visualizations by Analogy

- Compute difference $\Delta(A,B)$ from provenance
  - $D = \Delta(A,B) \circ C$ is often not a valid workflow
- Find map between $A$ & $C$: $\text{map}(A,C)$
Generating Visualizations by Analogy

• Compute difference $\Delta(A,B)$ from provenance
  - $D = \Delta(A,B) \circ C$ is often not a valid workflow

• Find map between A & C: $\text{map}(A,C)$

• Compute mapped difference
  $\Delta AC(A,B) = \text{map}(A,C) \Delta(A,B)$
  - $D = \Delta AC(A,B) \circ C$
Provenance for Teaching

Comparing Paths to Solutions for Two Students
VisTrails Demo

- [www.vistrails.org](http://www.vistrails.org)
- Creating workflows
- Tagging
- Annotations
- Running workflows from the command-line
Projects

• Survey:
  - Have not heard from many of you about the paper you will be reproducing…

• Research:
  - How can you test if your work is reproducible?
  - Why are the specific approaches (code versioning, containers) important to your research?
  - What does your work add to existing reproducibility tools?
Assignment 2

- [http://www.cis.umassd.edu/~dkoop/cis602/assignment2.html](http://www.cis.umassd.edu/~dkoop/cis602/assignment2.html)
- Reproducible Analysis of Tuberculosis Data using workflows and containers
- Create workflow using VisTrails
- Create Docker image using Dockerfile to run VisTrails
- Create Docker image that includes the workflow and runs it
- Create Docker image that allows reuse of that image
Provenance and data differencing for workflow reproducibility analysis

P. Missier, S. Woodman, H Hiden, P. Watson
Repeatability vs. Reproducibility

• Differences in data and workflows
• Important to understand when something in this spectrum fails and why
PDdiff

• Find differences in provenance traces
• Use differences between data (can use hashing)
• Can also have service (module/actor) mismatches