A Five-Star Rating Scheme to Assess Application Seamlessness

⇔ http://bit.ly/lebo-issues-2014

Timothy Lebo
Tetherless World Constellation
Rensselaer Polytechnic Institute

Timothy Lebo, Nicholas Del Rio, Patrick Fisher, Chad Salisbury
Information Directorate
Air Force Research Laboratory

Outline

- Motivation
  - Visual Analytics [some problems]
  - Linked Data [some potential]
- Demo: A-Half-and-Two Example Applications
- A Nascent Theory
  - Application Ontology (a PROV extension)
  - Munging Ontology (7 types)
  - Cost Model (*moving* data across 5 stars)
  - 5-Star Applications (AO extension)
  - Predictions!
- Conclusions and Future Work

Visual Analytics

- *Visual analytics* is the science of analytical reasoning facilitated by interactive visual interfaces.
  - ***Analytical reasoning techniques*** that enable users to obtain deep insights that directly support assessment, planning, and decision making
  - ***Visual representations and interaction techniques*** that take advantage of the human eye's broad bandwidth pathway into the mind to allow users to see, explore, and understand large amounts of information at once
  - ***Data representations and transformations*** that convert all types of conflicting and dynamic data in ways that support visualization and analysis
  - Techniques to support ***production, presentation, and dissemination*** of the results of an analysis to communicate information in the appropriate context to a variety of audiences.

Human cognition ⟷ Machine representation

Thomas and Cook, 2005.

Practical Challenges in Visual Analytics

Among **35** data analysts from **25** commercial organizations:

- Most tedious and time-consuming task is **discovering** and **wrangling** data
- Analytical results are **static**
- Analytical results are **shared** via email, a shared file system, or during group meetings
- Difficulties discovering when **relevant** data becomes available
- Visualizations avoided because considered *a **barrier** to underlying data*

S. Kandel *et al.*, 2012.

Linked Data Avoids Archaeological Endeavors

Linked Data offers a huge **potential** for establishing explicit, understandable connections within and across data sources.

★★★★★ 

Include **links** to other URIs, so that people can **discover** more things.

★★★★ 

Use HTTP URIs to name things, so that people can **look up those names**.
When someone looks up a URI, provide useful information in **RDF**.

Tim Berners-Lee
w3.org/DesignIssues/LinkedData

★★★ 

Available as **non-proprietary** format.
(e.g. CSV instead of Excel)

★★ 

Available as machine-readable **structured data**.
(e.g. Excel instead of image scan of a table)

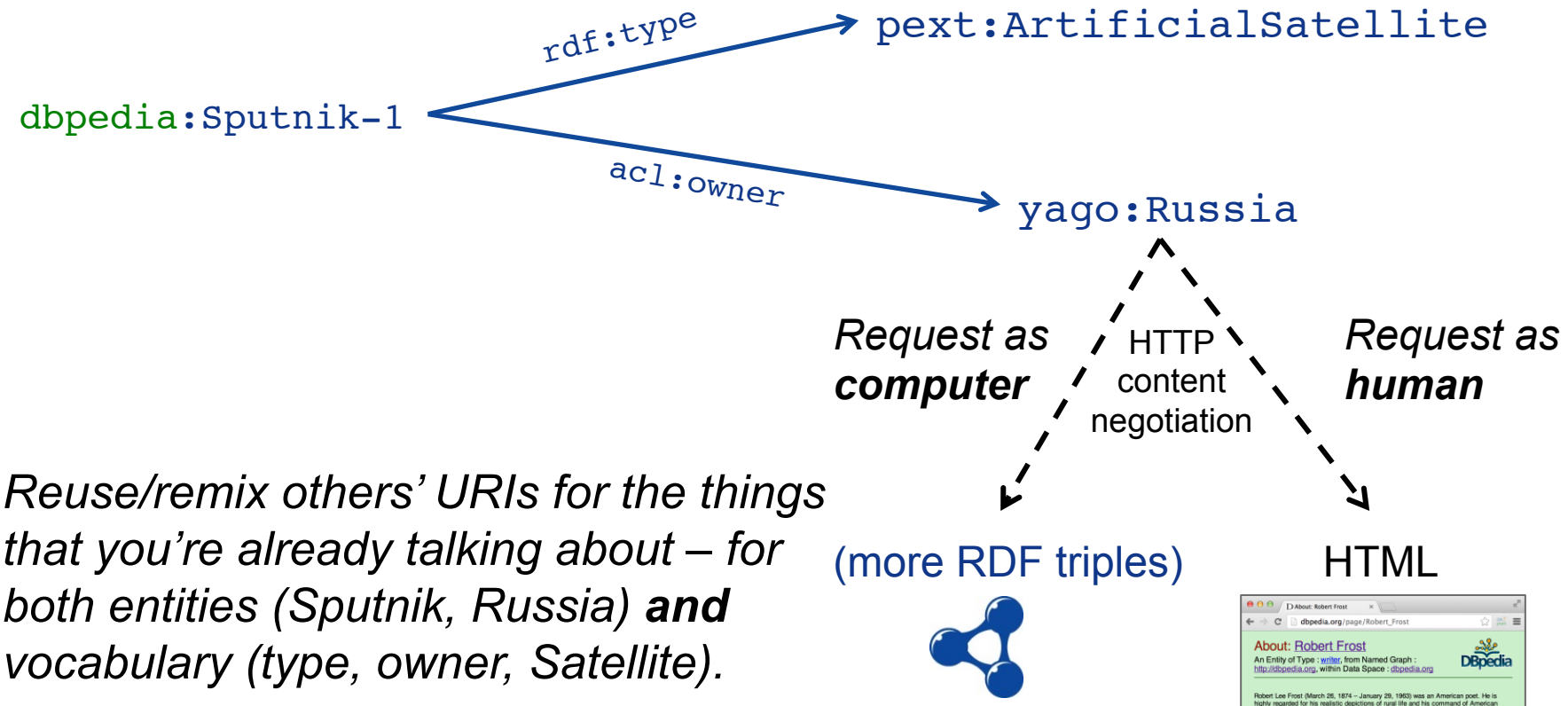
★ 

Available on the web (whatever format) but with an open license.

http://5stardata.info

Linked Data: Simple, Powerful, Emergent Design

Linked Data offers a huge **potential** for establishing explicit, understandable connections within and across data sources.



*Reuse/remix others' URIs for the things that you're already talking about – for both entities (Sputnik, Russia) **and** vocabulary (type, owner, Satellite).*

```
dbpedia : http://dbpedia.org/resource/
acl     : http://www.w3.org/ns/auth/acl#
yago    : http://yago-knowledge.org/resource/
pext    : http://www.ontotext.com/proton/protonext#>
```

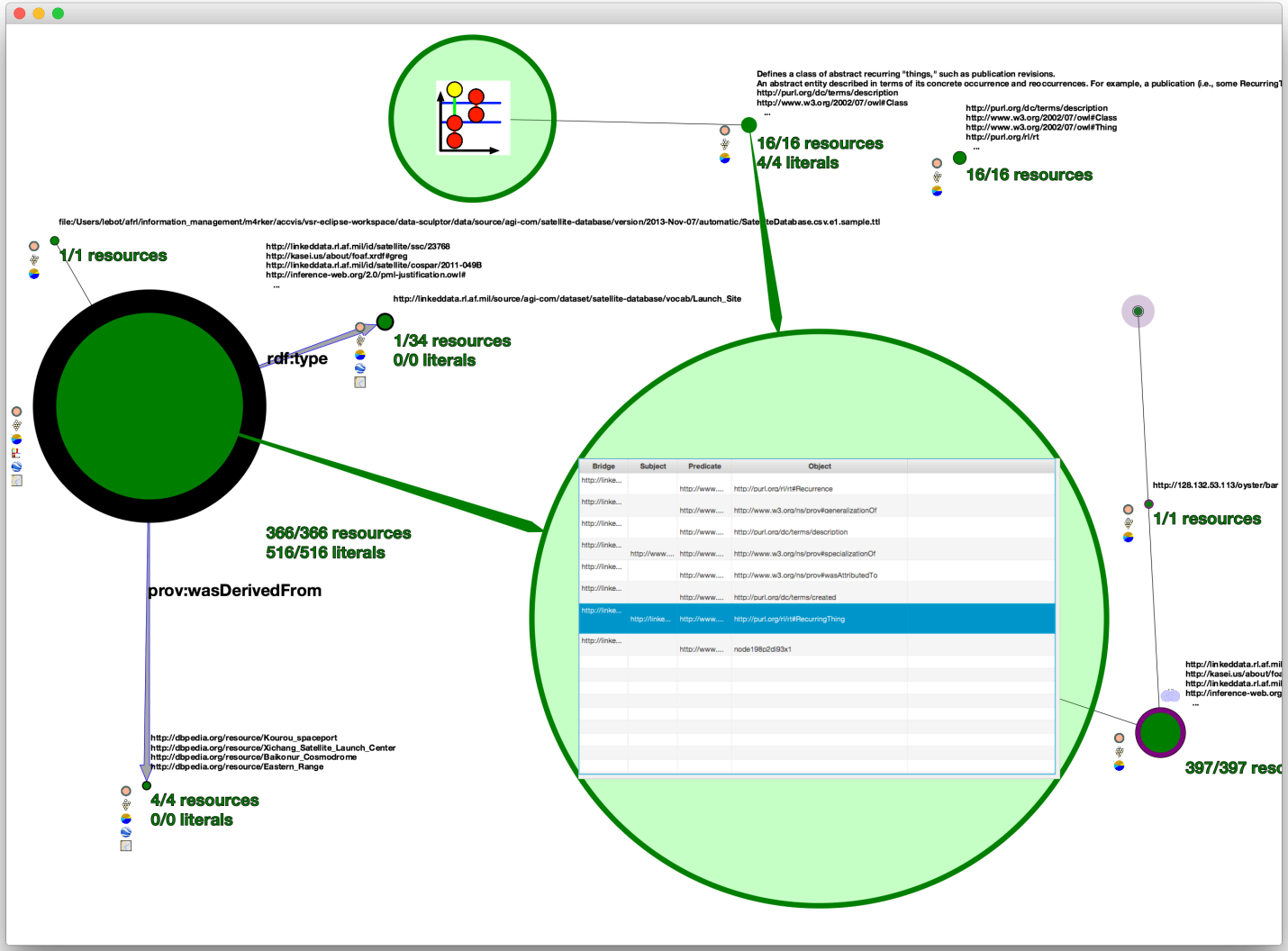
Motivation

- Technical barriers slow **analysts** down in *creating, understanding, trusting, and re-purposing* results.

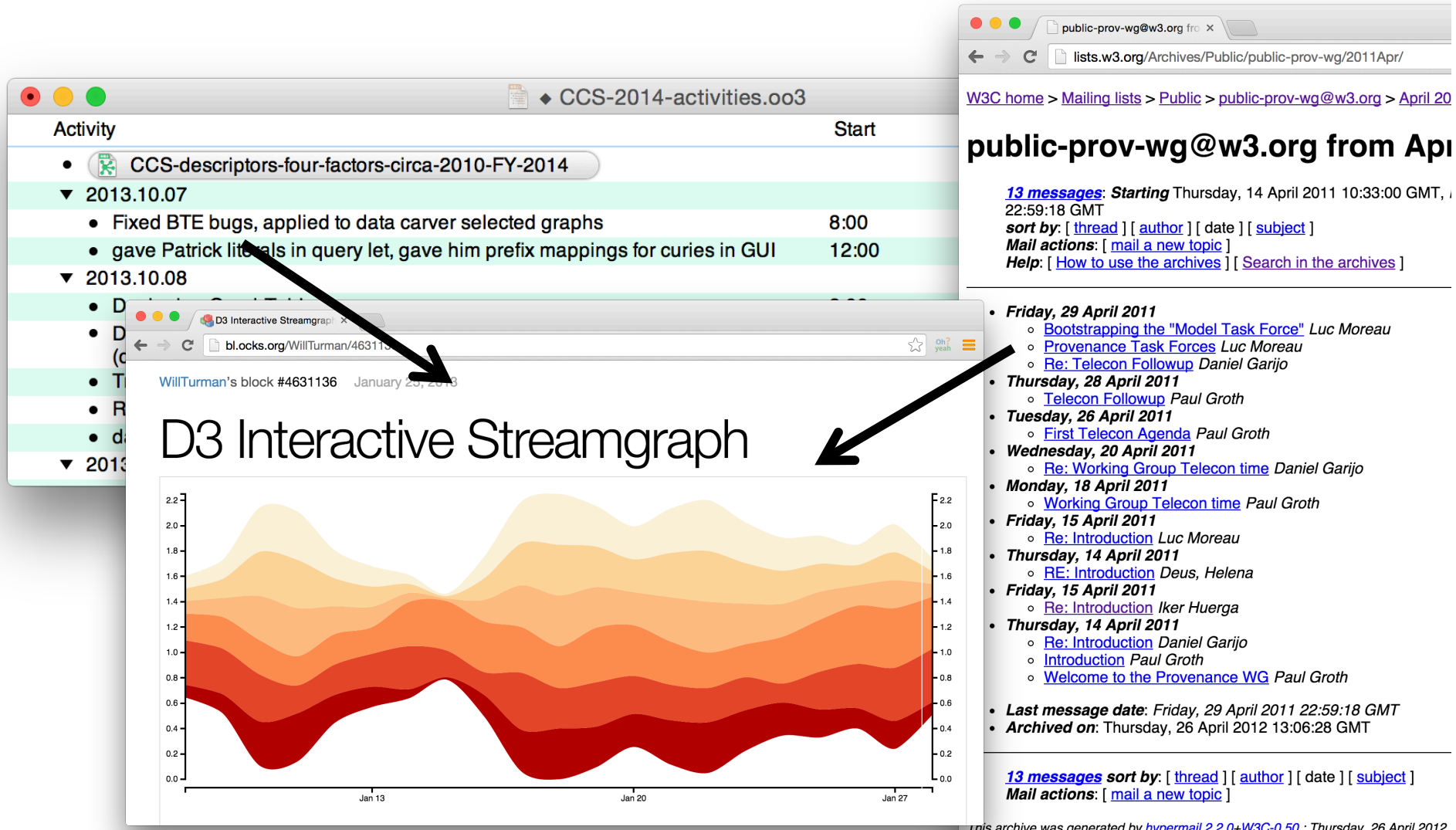
- Linked Data offers a huge **potential** for establishing explicit, understandable connections within and across data sources.
- Linked Data is **ready to use**, but not yet useful.

Hypothesis: Applying Linked Data principles to the analytical process **reduces the time required** for **analysts** to *create, understand, trust, and re-purpose* any existing result.

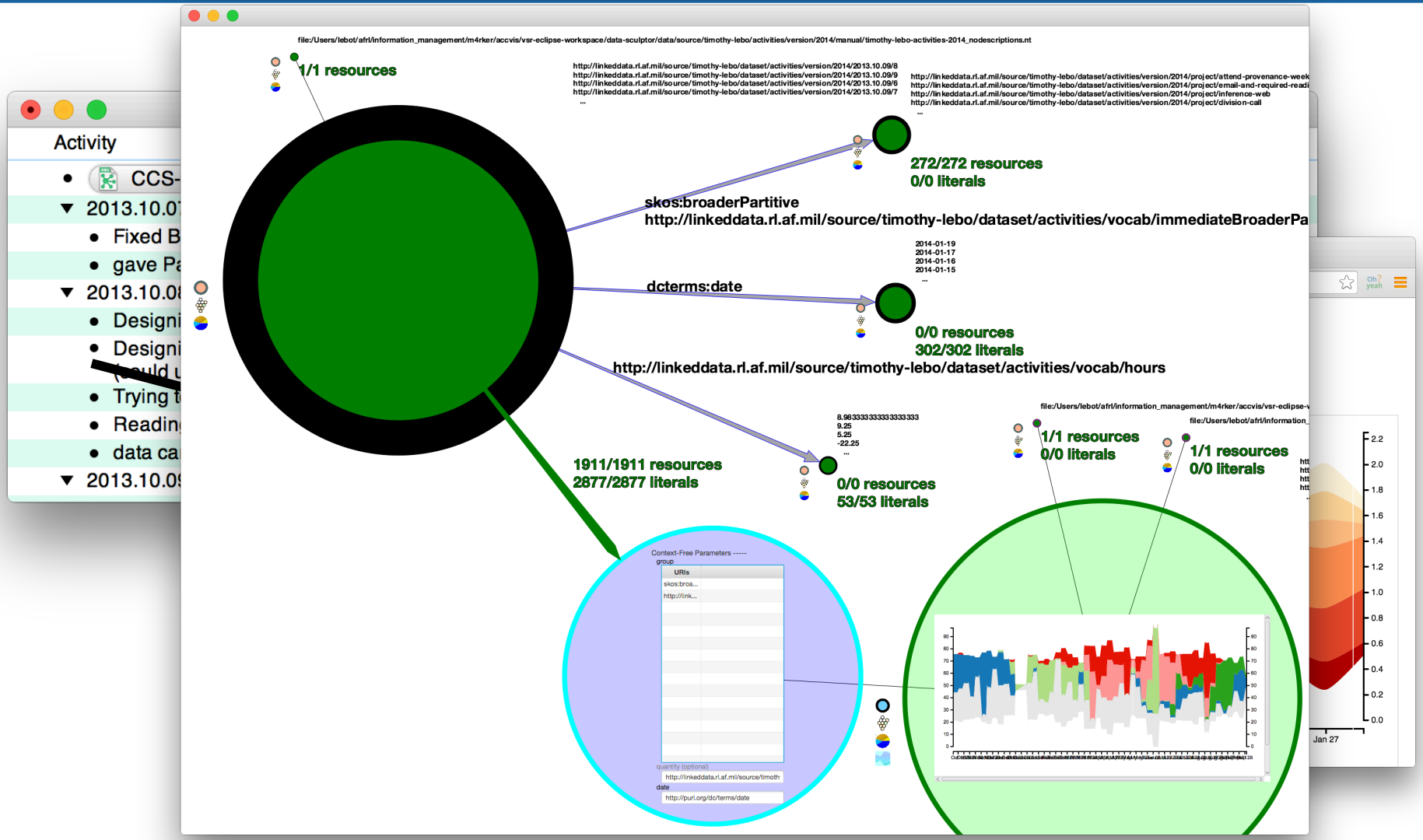
A Half Example: Munging by Aligning with a Visualization's Input Semantics



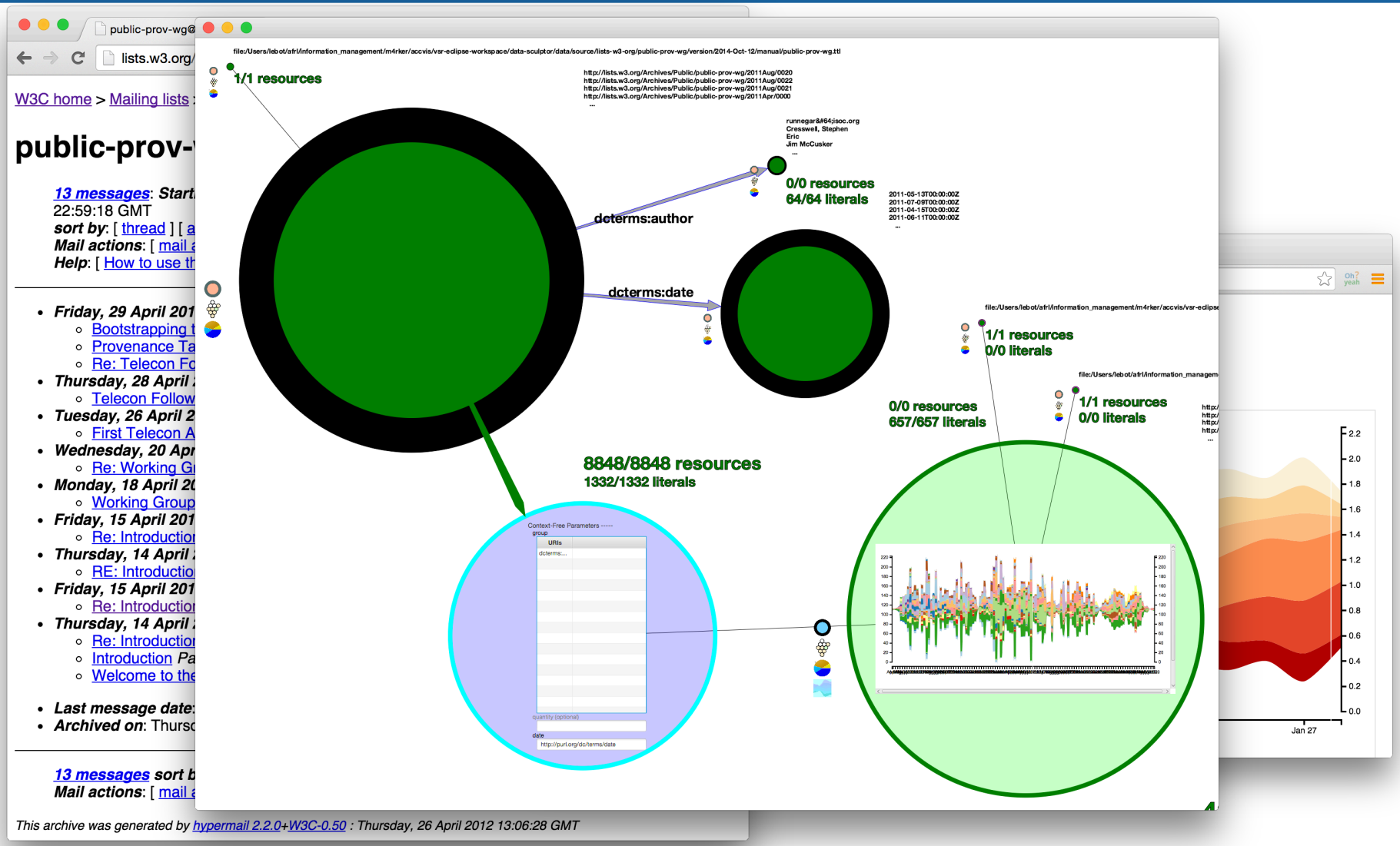
Two Example Applications



Application 1: Individual Time-Tracking



Application 2: public-prov-wg@w3.org Mail Archives



Outline

- Motivation
  - Visual Analytics [some problems]
  - Linked Data [some potential]
- Demo: Two-and-a-Half Example Applications
- A Nascent Theory
  - Application Ontology (a PROV extension)
  - Munging Ontology (7 types)
  - Cost Model (*moving* data across 5 stars)
  - 5-Star Applications (AO extension)
  - Predictions!
- Conclusions and Future Work

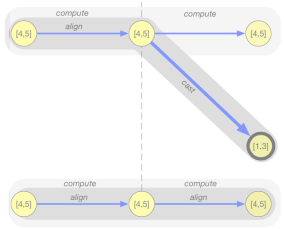
Building to a Theory

5-Star Theory

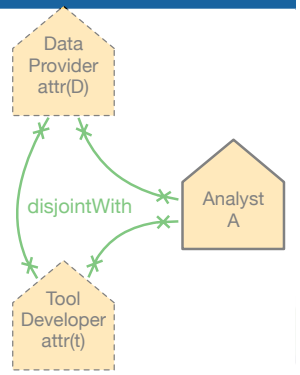
5-Star Application Ontology

Seamlessness Metric

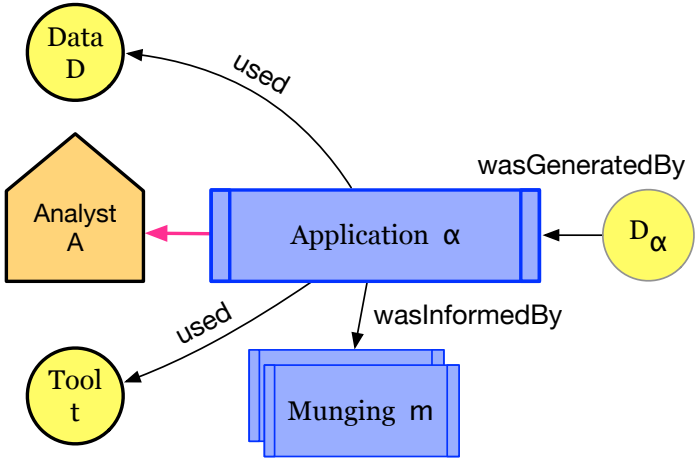
Application Ontology - Core



[6, 38]
[6, 24]



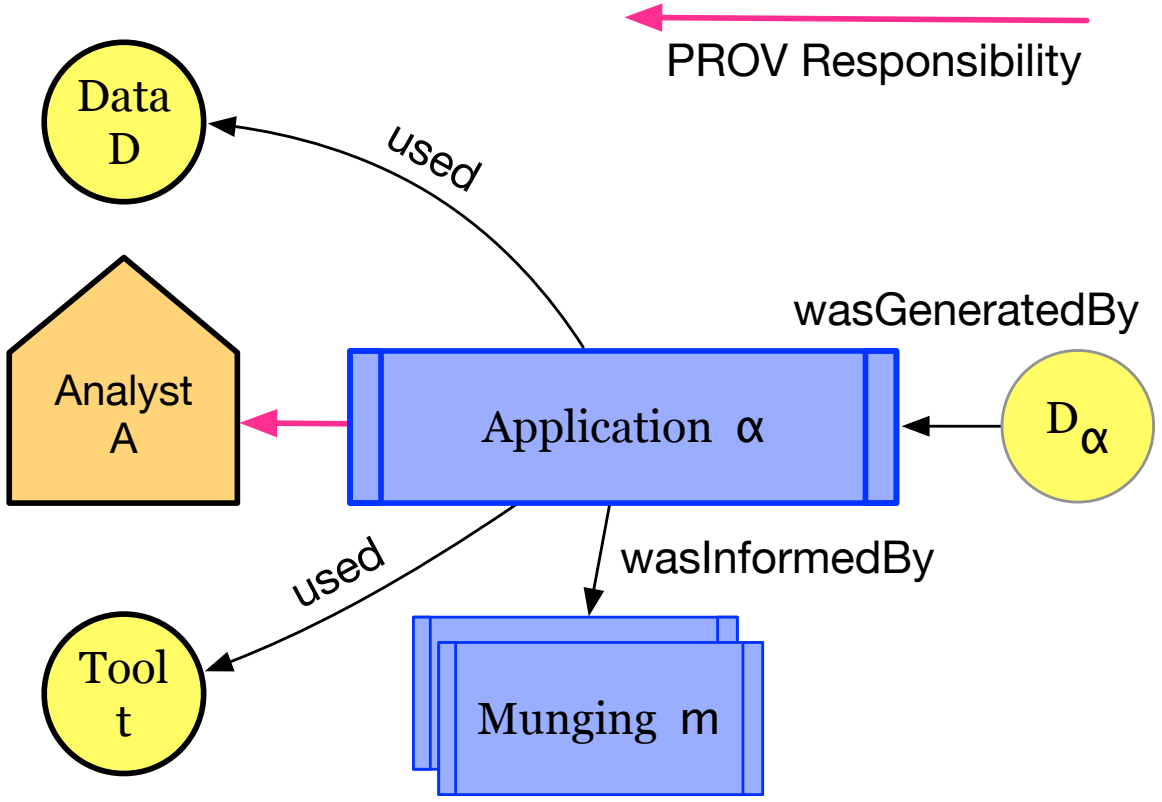
$$S_*(E) = \frac{\sum_{\alpha \in E} \sum_{m \in M_\alpha} cost(shim)}{\sum_{\alpha \in E} \sum_{m \in M_\alpha} pot(D_\alpha) cost(m)}$$



Theory
Computation
Ontology
Use Case
Implementation

Application Ontology - Core

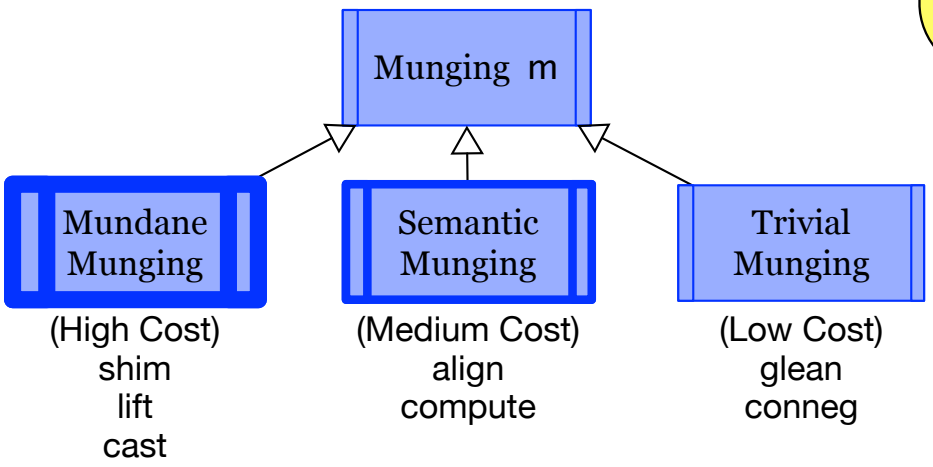
An activity is something that occurs over a period of time and acts upon or with entities; it may include consuming, processing, transforming, modifying, relocating, using, or generating entities.

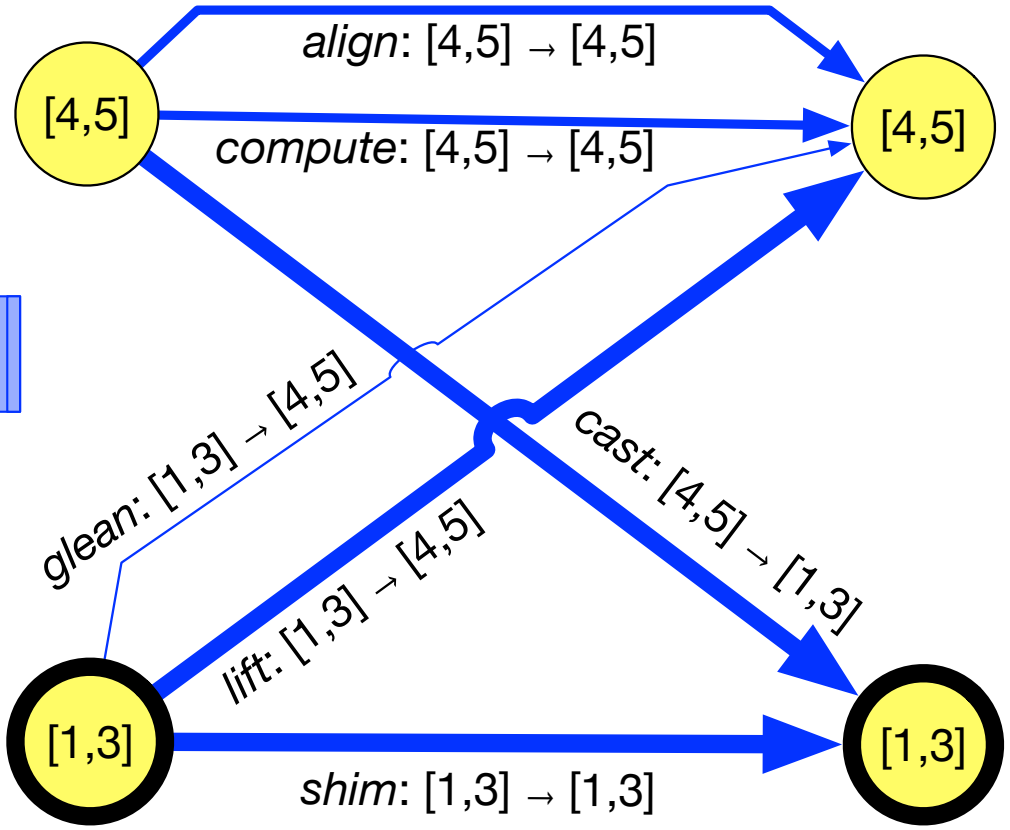


Munging (or wrangling) is the imperfect manipulation of data into a usable form.

Munging Ontology

$$\text{munge} : \{D_{[1,3]}, D_{[4,5]}\} \mapsto \{D_{[1,3]}, D_{[4,5]}\}$$





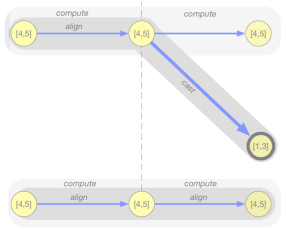
Building to a Theory

5-Star Theory

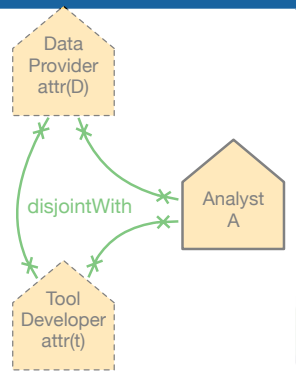
5-Star Application Ontology

Seamlessness Metric

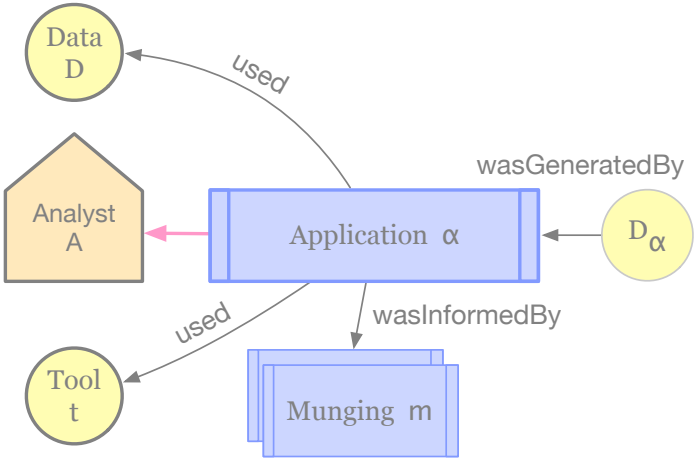
Application Ontology - Core



[6, 38]
[6, 24]

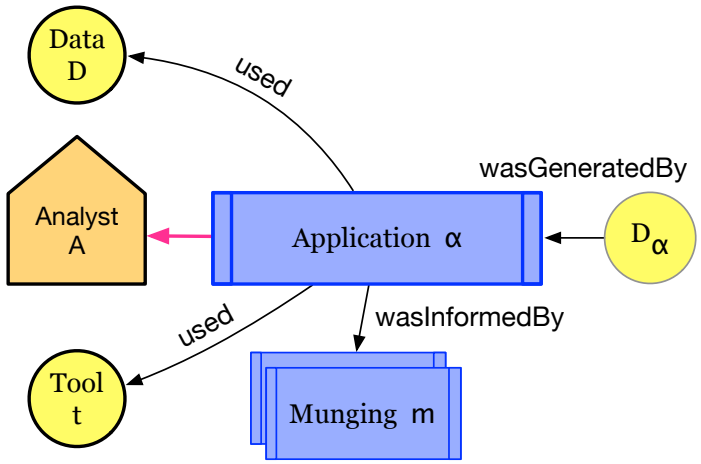


$$S_*(E) = \frac{\sum_{\alpha \in E} \sum_{m \in M_\alpha} cost(shim)}{\sum_{\alpha \in E} \sum_{m \in M_\alpha} pot(D_\alpha) cost(m)}$$



Theory
Computation
Ontology
Use Case
Implementation

Ecosystem Seamlessness



Let an analytical ecosystem E be the set of applications that influenced[16] a particular analysis:

$$E = \{\alpha_1, \alpha_2, ..., \alpha_n\}$$

Let an application α be a tuple comprising its set of munging activities M and the resulting dataset D_α:

$$\alpha = (M_\alpha = \{m_1, m_2, ..., m_m\}, D_\alpha)$$

The cost to perform an application is *at least* equal to the cost incurred by its munges. This inequality is based on existing work that shows that munging costs are a non-trivial portion of the overall analytical costs [6].

$$0 \ll cost(M_\alpha) = \sum_{m \in M_\alpha} cost(m) < cost(\alpha)$$

Ecosystem Seamlessness: Full Ordering of Munge Costs

$$cost(shim) > cost(lift) + 2\,cost(align) + cost(cast)$$

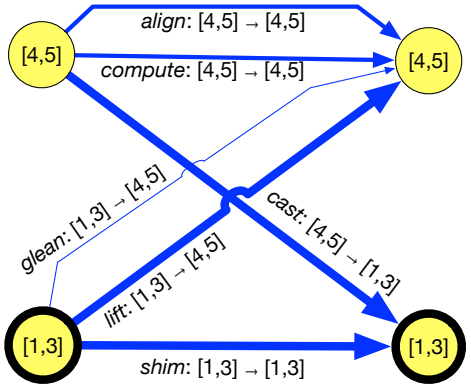
$$cost(lift) > cost(cast)$$

$$cost(cast) > cost(align)$$

$$cost(align) > cost(comp)$$

$$cost(comp) > cost(glean)$$

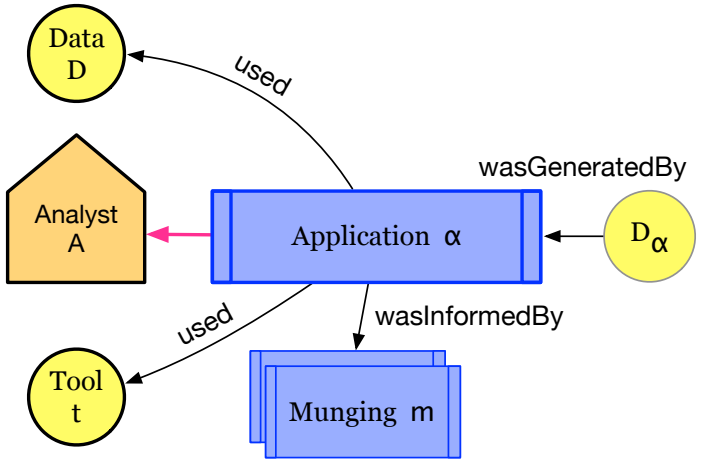
$$cost(glean) > cost(conneg)$$



$$cost(m) = \begin{cases} 19 = 6 + 2(4) + 5 : \text{ if } shim \\ 6 \ : \text{ if } lift \\ 5 \ : \text{ if } cast \\ 4 \ : \text{ if } align \\ 3 \ : \text{ if } comp \\ 2 \ : \text{ if } glean \\ 1 \ : \text{ if } conneg \end{cases}$$

Ecosystem Seamlessness

$$S_*(E) = \frac{\sum_{\alpha \in E} \sum_{m \in M_\alpha} cost(shim)}{\sum_{\alpha \in E} \sum_{m \in M_\alpha} pot(D_\alpha) cost(m)}$$



$$1 \leq S_*(E)$$

$$pot(D_\alpha) = \begin{cases} \frac{1}{cost(shim)} & : tbl(D_\alpha) > 3 \\ \frac{1}{cost(conneg)} & : conneg(D_\alpha) \supset \emptyset \\ \frac{1}{cost(glean)} & : glean(D_\alpha) \supset \emptyset \\ 1 & : \text{otherwise} \end{cases}$$

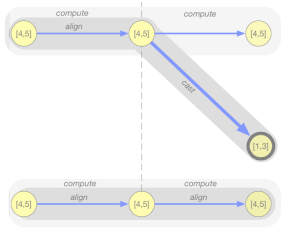
Building to a Theory

5-Star Theory

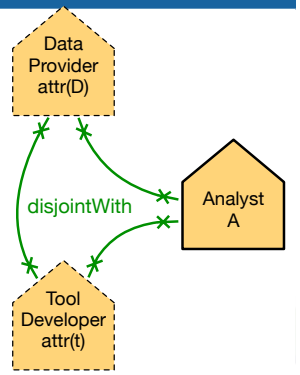
5-Star Application Ontology

Seamlessness Metric

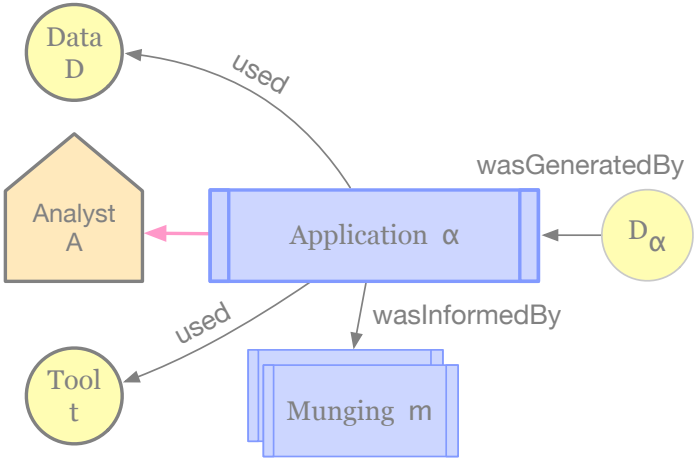
Application Ontology - Core



[6, 38]
[6, 24]



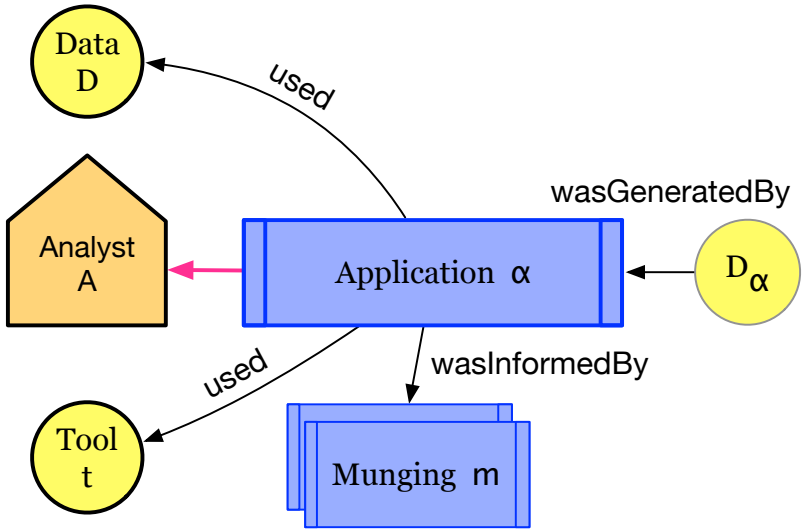
$$S_*(E) = \frac{\sum_{\alpha \in E} \sum_{m \in M_\alpha} cost(shim)}{\sum_{\alpha \in E} \sum_{m \in M_\alpha} pot(D_\alpha) cost(m)}$$



Theory
Computation
Ontology
Use Case
Implementation

5-Star Applications

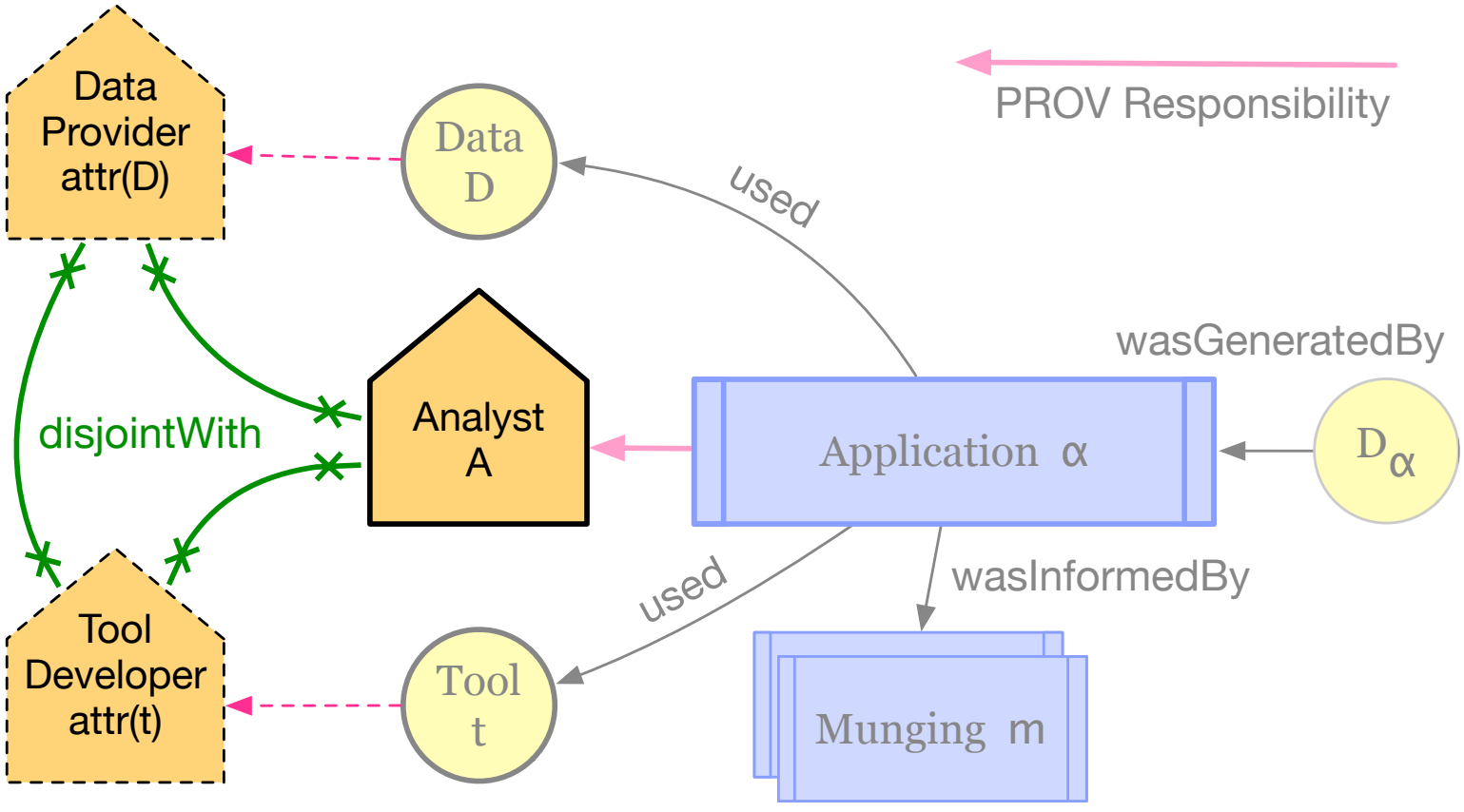
Five progressive restrictions on the form of applications,
based on the **structure** of data involved and the **behavior** of tools used.



☆	Informal Restriction	Formally
1	data providers, analysts, and tool developers are disjoint	$attr(D) \cap A \cap attr(t) = \emptyset$
2	accept data (any format) via URL; cite that URL in the future	$tbl(D) >= 1 \wedge URL \in D_\alpha$
3	accept data (RDF format) via URL; cite that URL in the future	$tbl(D) >= 4$
4	use a tool's input semantics (OWL, SPARQL) when preforming munges	$used(m, t_\sigma) \wedge m \in M$
5	provide any information (RDF format) derived during use	$D \subset D_\alpha$

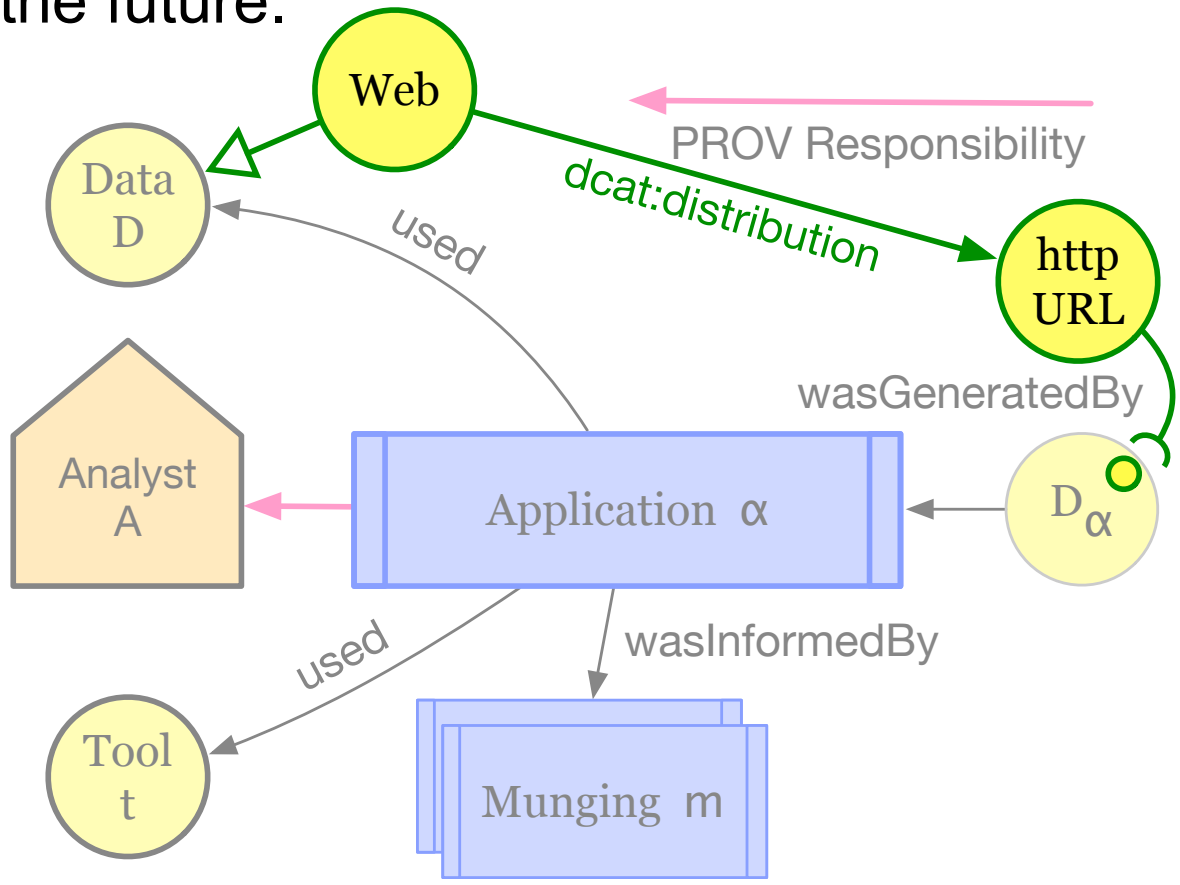
1-Star Applications

Data providers, analysts, and tool developers are disjoint.



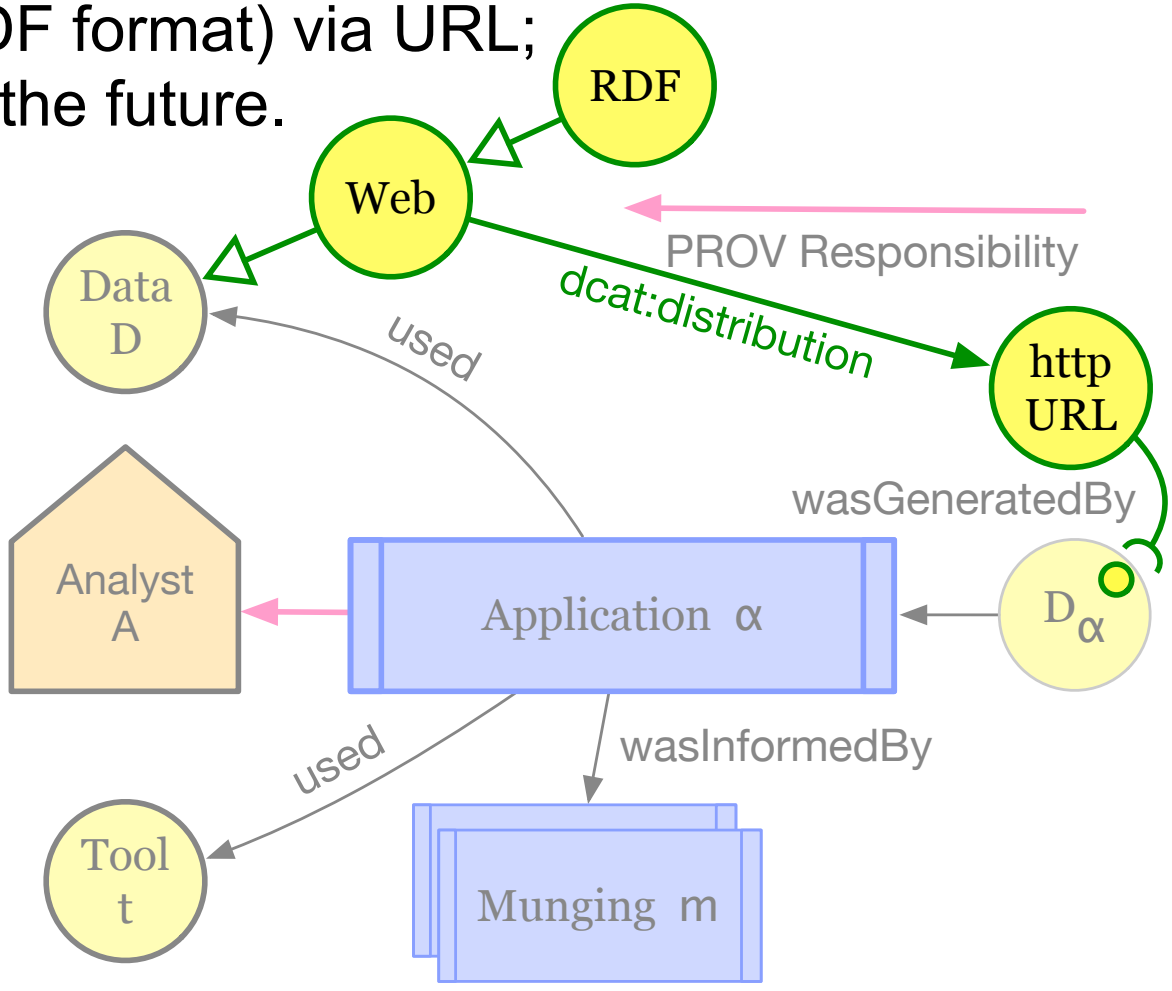
2-Star Applications

Accept data (any format) via URL;
cite that URL in the future.



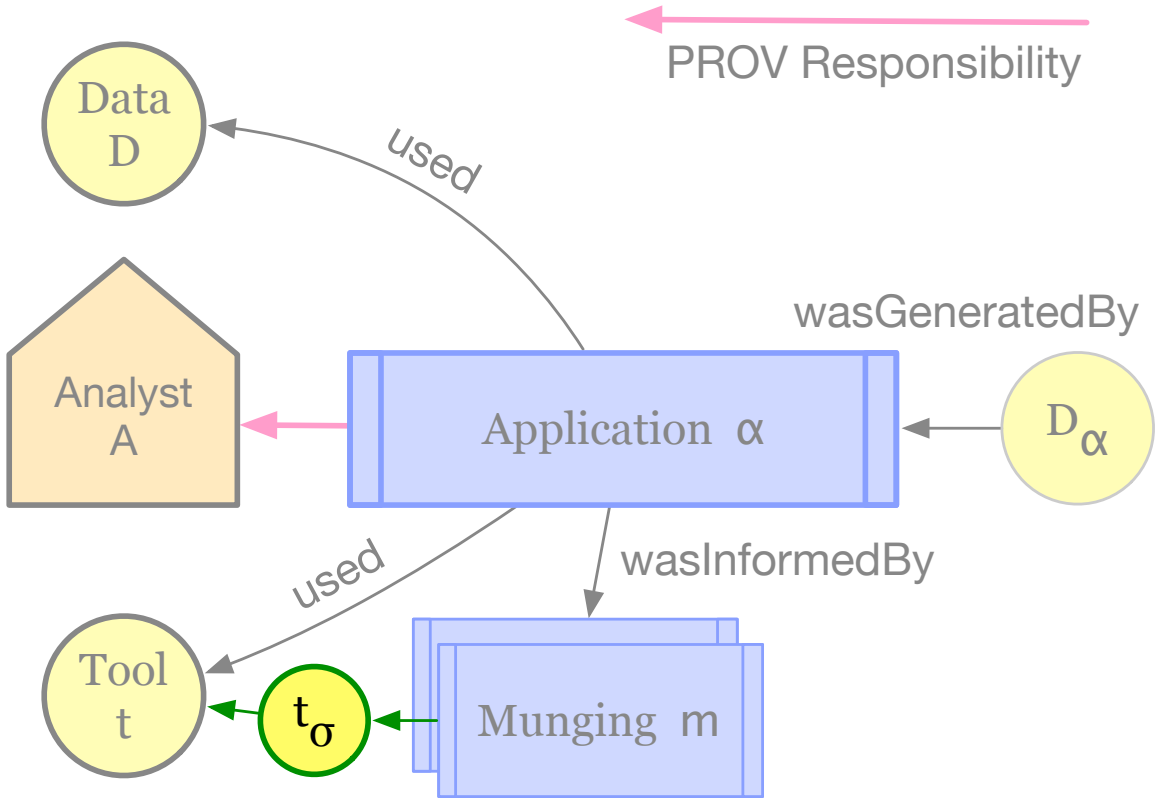
3-Star Applications

Accept data (RDF format) via URL;
cite that URL in the future.



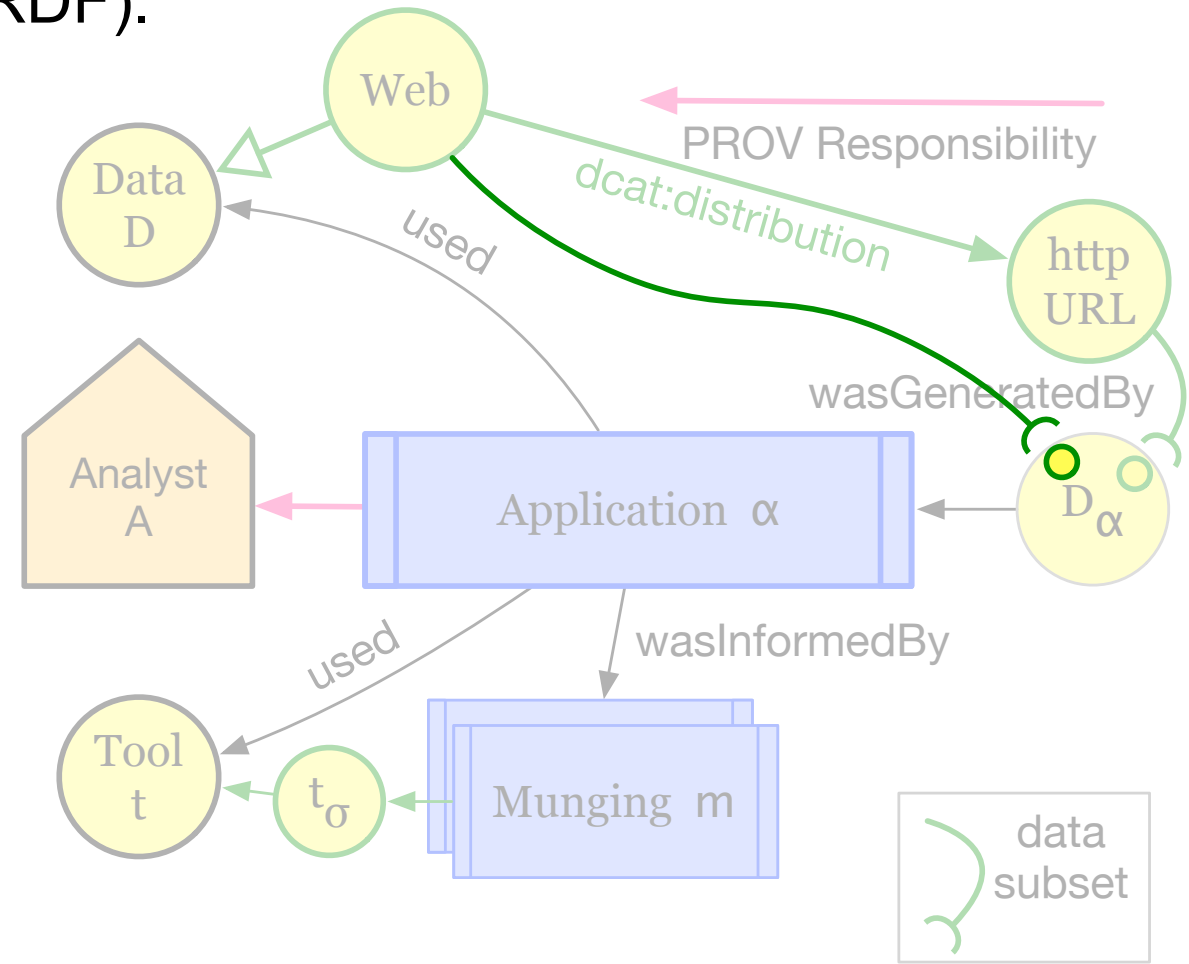
4-Star Applications

Use a tool's input semantics (OWL, SPARQL) to munge.



5-Star Applications

Provide any information derived during use (as RDF).



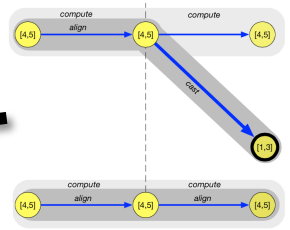
Building to a Theory

5-Star Theory

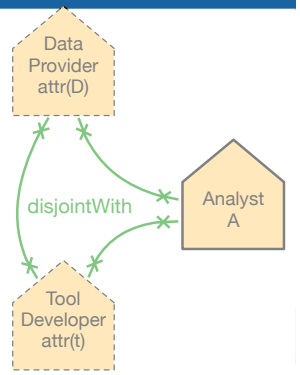
5-Star Application Ontology

Seamlessness Metric

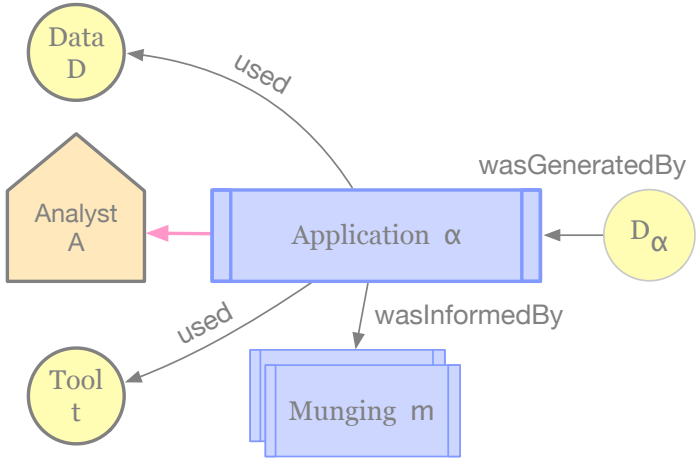
Application Ontology - Core



[6, 38]
[6, 24]

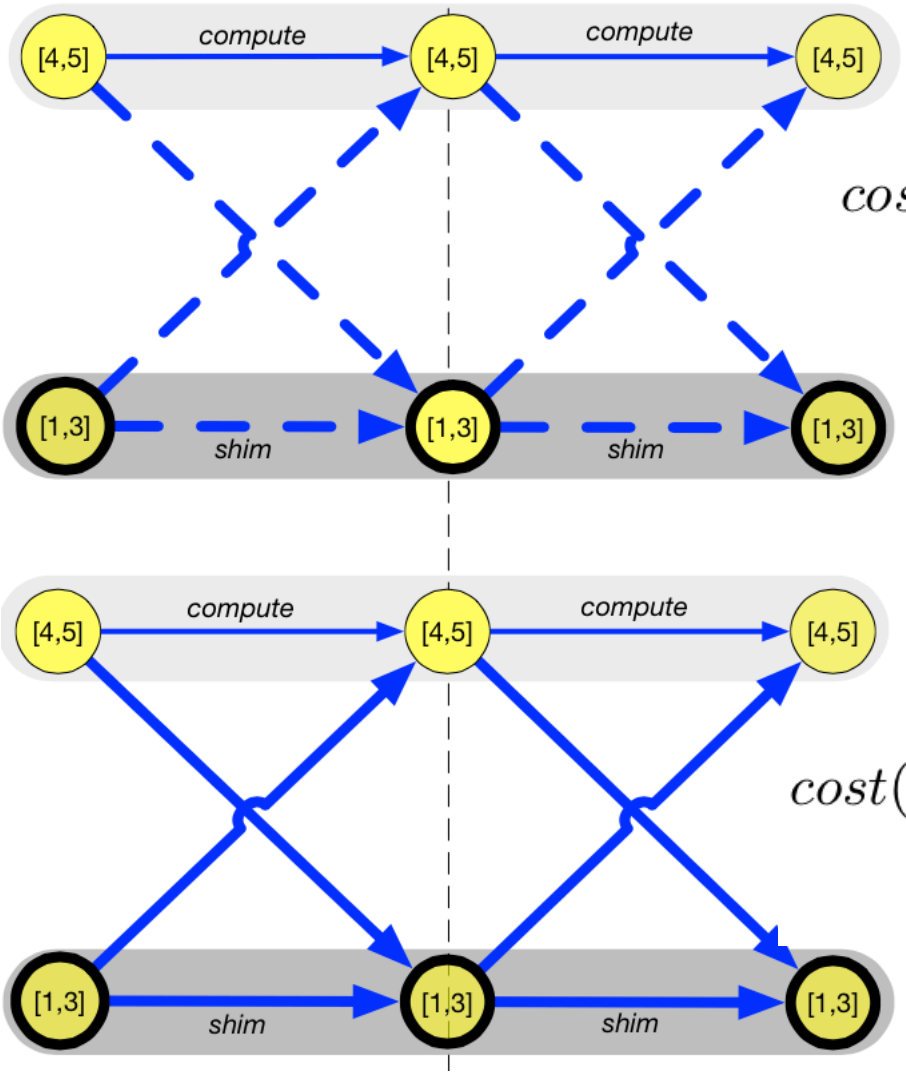


$$S_*(E) = \frac{\sum_{\alpha \in E} \sum_{m \in M_\alpha} cost(shim)}{\sum_{\alpha \in E} \sum_{m \in M_\alpha} pot(D_\alpha) cost(m)}$$



Theory
Computation
Ontology
Use Case
Implementation

Predicted Cost Bounds for any Hypothetical Application Chain



$$cost(\alpha_\star) = [2 \times cost(comp), \infty]$$

$$= \underline{[6, \infty]}$$

∞

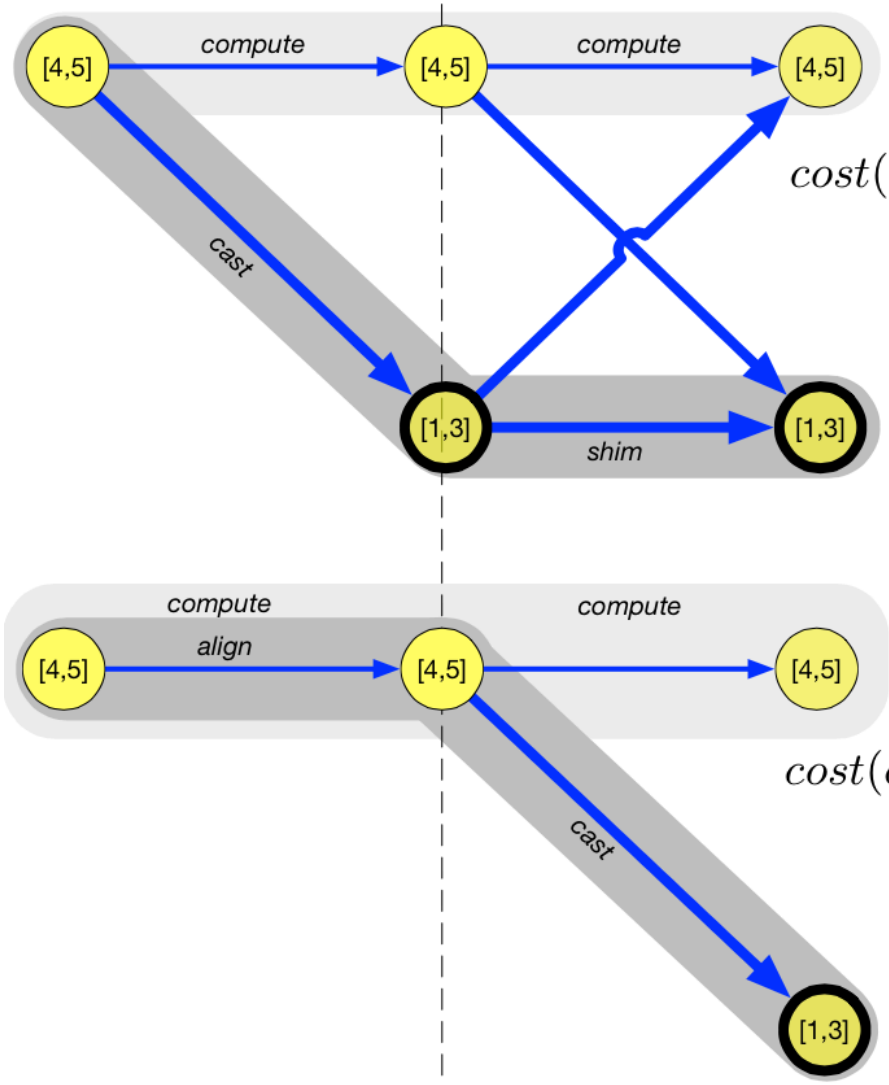
Lower Cost Bound Path

Upper Cost Bound Path

$$cost(\alpha_{\star\star}) = [2 \times cost(comp), 2 \times cost(shim)]$$

$$= \underline{[6, 38]} < cost(\alpha_\star) = [6, \infty]$$

Predicted Cost Bounds for any Hypothetical Application Chain



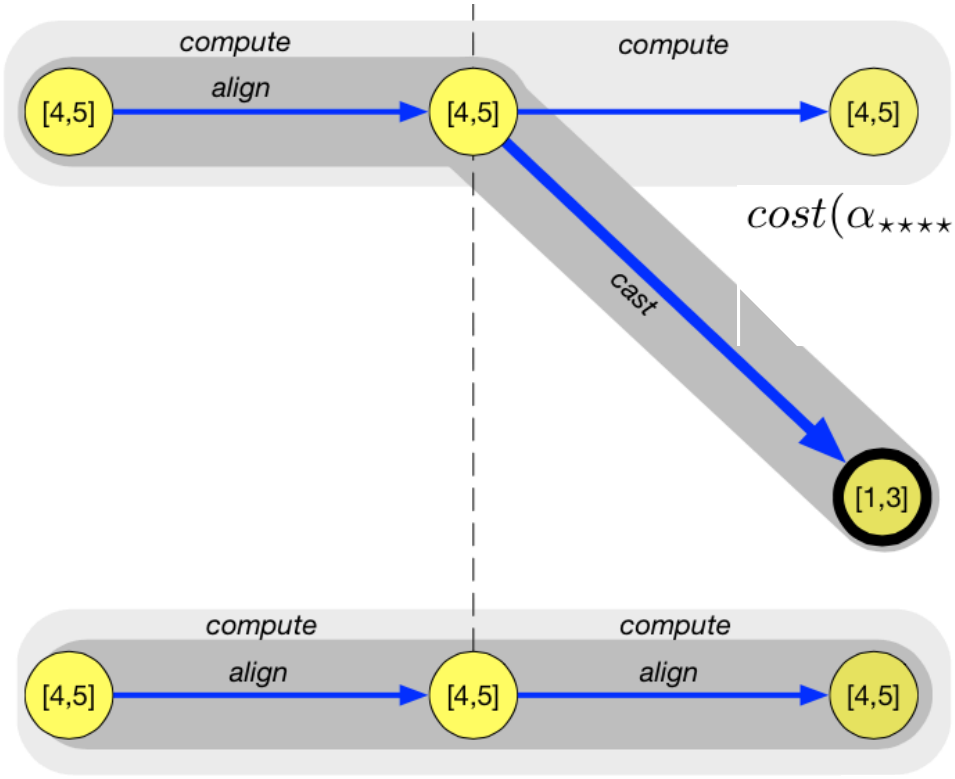
$$cost(\alpha_{\star\star\star}) = [2 \times cost(comp), cost(cast) + cost(shim)]$$
$$= [6, 24] < cost(\alpha_{\star\star}) = [6, 38]$$

Lower Cost Bound Path

Upper Cost Bound Path

$$cost(\alpha_{\star\star\star\star}) = [2 \times cost(comp), cost(align) + cost(cast)]$$
$$= [6, 9] < cost(\alpha_{\star\star\star}) = [6, 24]$$

Predicted Cost Bounds for any Hypothetical Application Chain



$$cost(\alpha_{\star\star\star\star}) = [2 \times cost(comp), cost(align) + cost(cast)]$$
$$= [6, 9] < cost(\alpha_{\star\star\star}) = [6, 24]$$

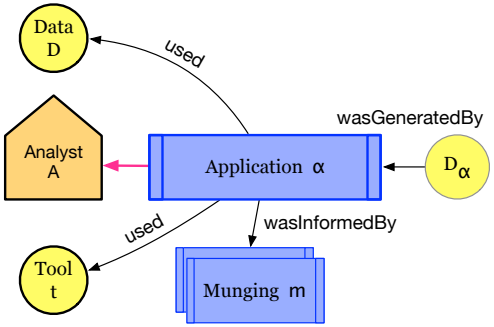
$$cost(\alpha_{\star\star\star\star\star}) = [2 \times cost(comp), 2 \times cost(align)]$$
$$= [6, 8] < cost(\alpha_{\star\star\star\star}) = [6, 9]$$

Lower Cost Bound Path

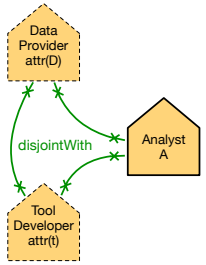
Upper Cost Bound Path

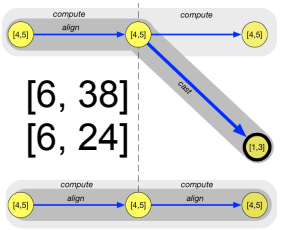
Conclusion

- A simple theory of seamless analytics
  - embodies tenets of the semantic web
  - based on the structure of data involved and the behavior of tools used
  - provides explanations, falsifiable predictions



$$S_*(E) = \frac{\sum_{\alpha \in E} \sum_{m \in M_\alpha} cost(shim)}{\sum_{\alpha \in E} \sum_{m \in M_\alpha} pot(D_\alpha) cost(m)}$$





Future Work

- Theoretically
  - Demonstrate coverage of phenomena
  - Apply to a variety of costs (existing measures)
  - Test, break, revise!
- Practically
  - Tool Developers: methodologies and techniques
    - Knowledge-based Software Engineering
    - Content-Preserving URLs
  - Tool Users: usability and effectiveness

Future Work

- Towards Formal Visualization Knowledge
  - Multi-Expression Semantics (i.e., Java, OWL, SPARQL)
  - Ontology for views and mappings
  - Show coverage across existing toolsets
  - “Zombie Apocalypse”, views dying to show you something
  - Modeling user needs and capabilities
  - Intelligent, contextualized pruning of the flood of views

Thanks!

- Questions?

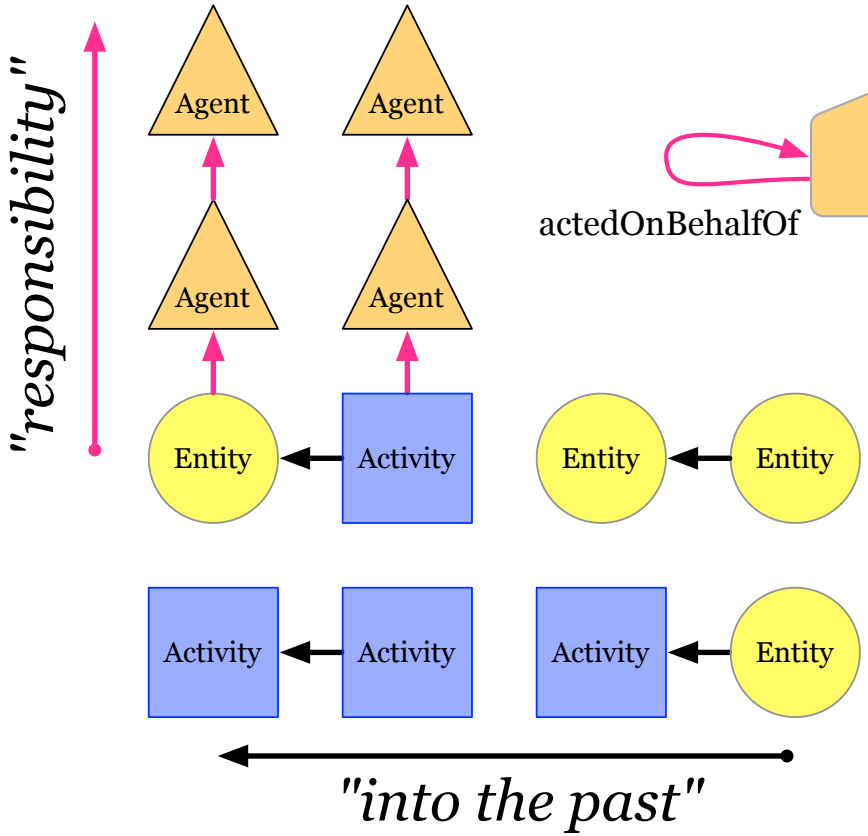
References

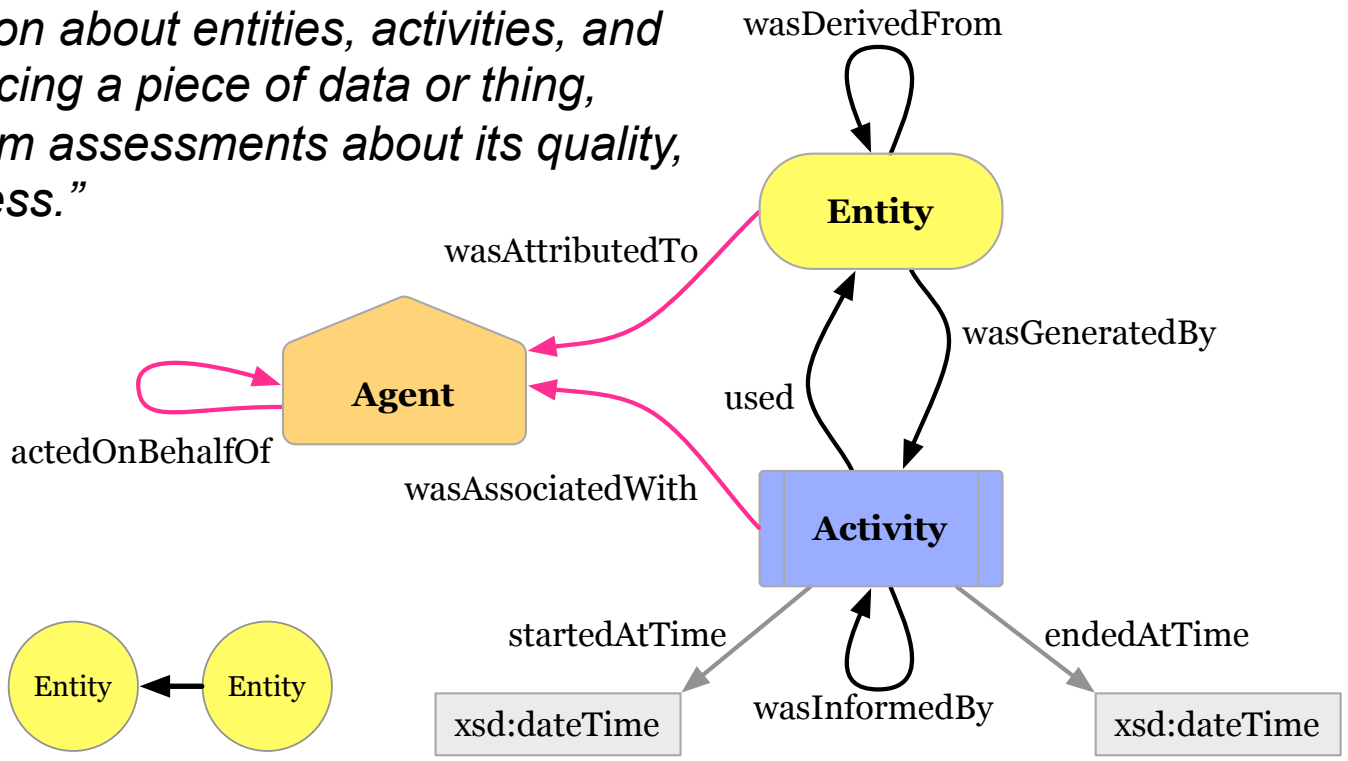
- S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. Visualization and Computer Graphics, IEEE Transactions on, 18(12):2917–2926, Dec. 2012.
- J. Thomas and K. Cook. Illuminating the path: The research and development agenda for visual analytics. IEEE Computer Society Press, 2005.

W3C PROVenance

Recommendation 20 Apr 2013

"Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness."





Core model

http://www.w3.org/TR/prov-overview