



# *A Five-Star Rating Scheme to Assess Application Seamlessness*

↔ <http://bit.ly/lebo-issues-2014>



**Timothy Lebo**  
Tetherless World Constellation  
Rensselaer Polytechnic Institute



**Rensselaer**



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

DISTRIBUTION STATEMENT A. Approved  
for public release; distribution is unlimited.  
Case Number: 88ABW-2014-5577



# 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.



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**



# 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**.



Tim Berners-Lee  
[w3.org/DesignIssues/LinkedData](http://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)



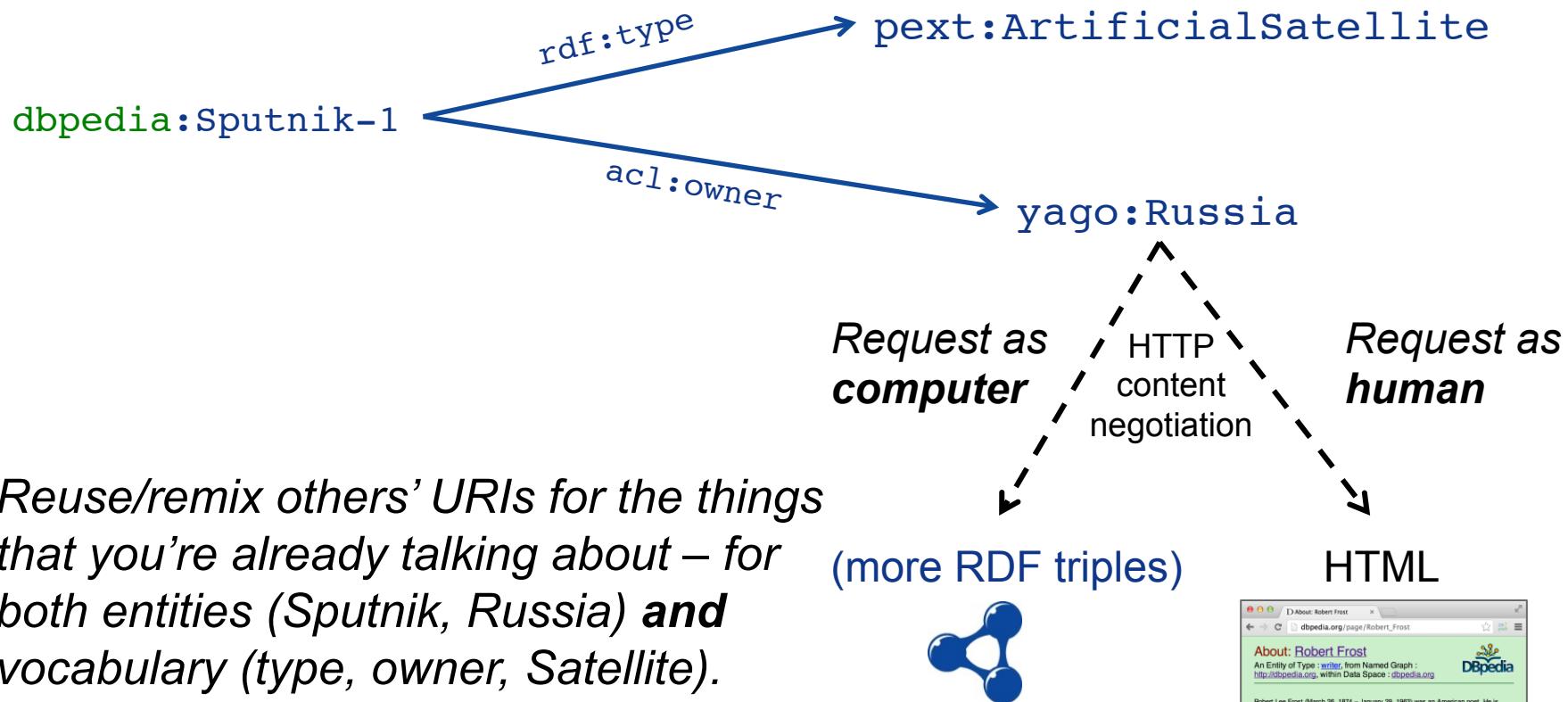
**Available** on the web (whatever format) but with an open license.

<http://5stardata.info>



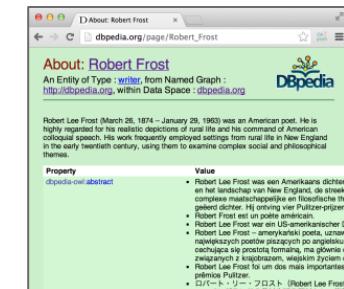
# Linked Data: Simple, Powerful, Emergent Design

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#>>





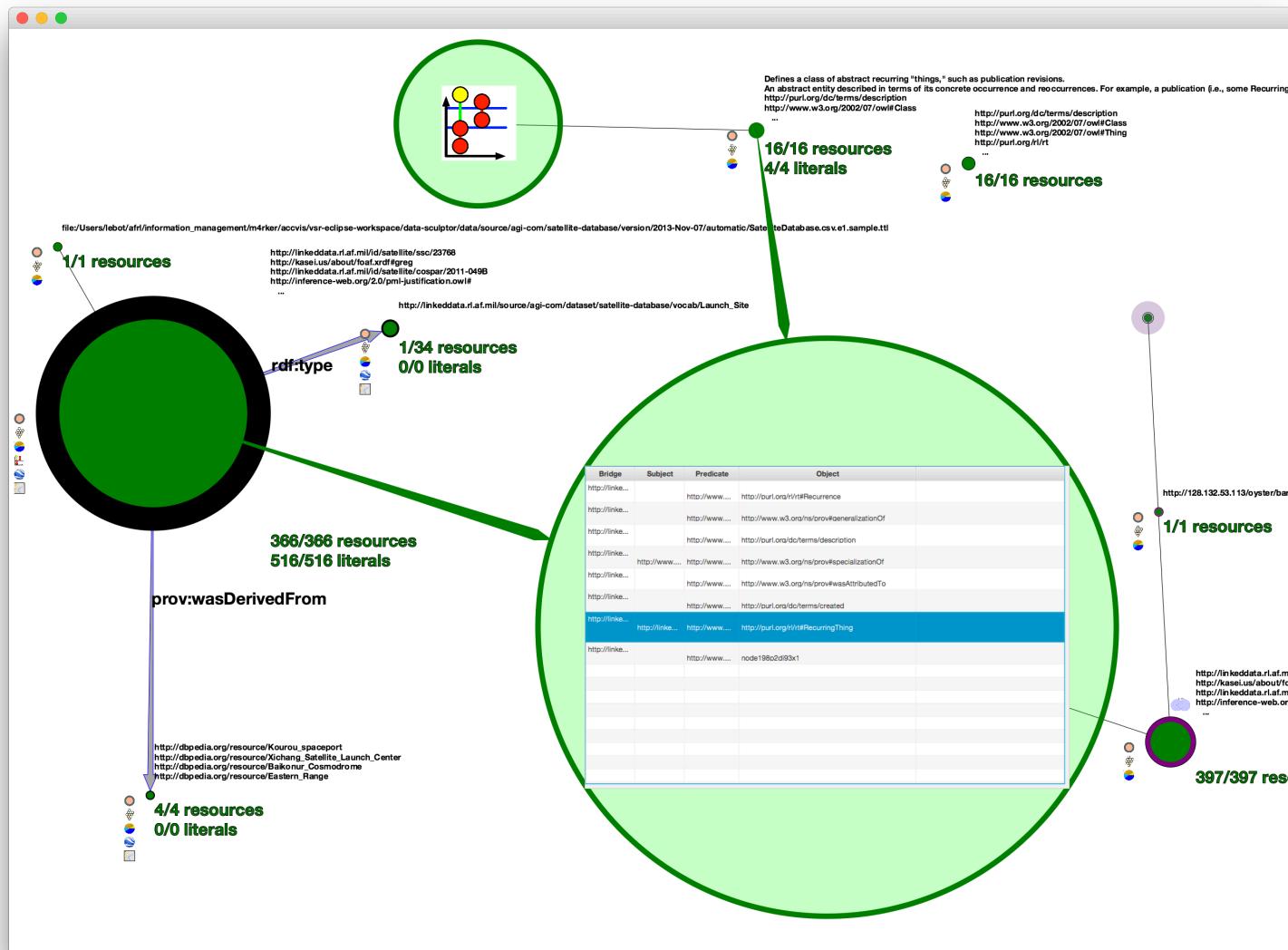
# Motivation

- 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

Activity

- CCS-descriptors-four-factors-circa-2010-FY-2014
- 2013.10.07
  - Fixed BTE bugs, applied to data carver selected graphs 8:00
  - gave Patrick literals in query let, gave him prefix mappings for curies in GUI 12:00
- 2013.10.08
  - D
  - D
  - (c
  - T
  - R
  - d
- 2013

Start

WillTurman's block #4631136 January 26, 2013

## D3 Interactive Streamgraph

public-prov-wg@w3.org from API

13 messages: Starting Thursday, 14 April 2011 10:33:00 GMT, i  
22:59:18 GMT  
sort by: [ thread ] [ author ] [ date ] [ subject ]  
Mail actions: [ mail a new topic ]  
Help: [ How to use the archives ] [ Search in the archives ]

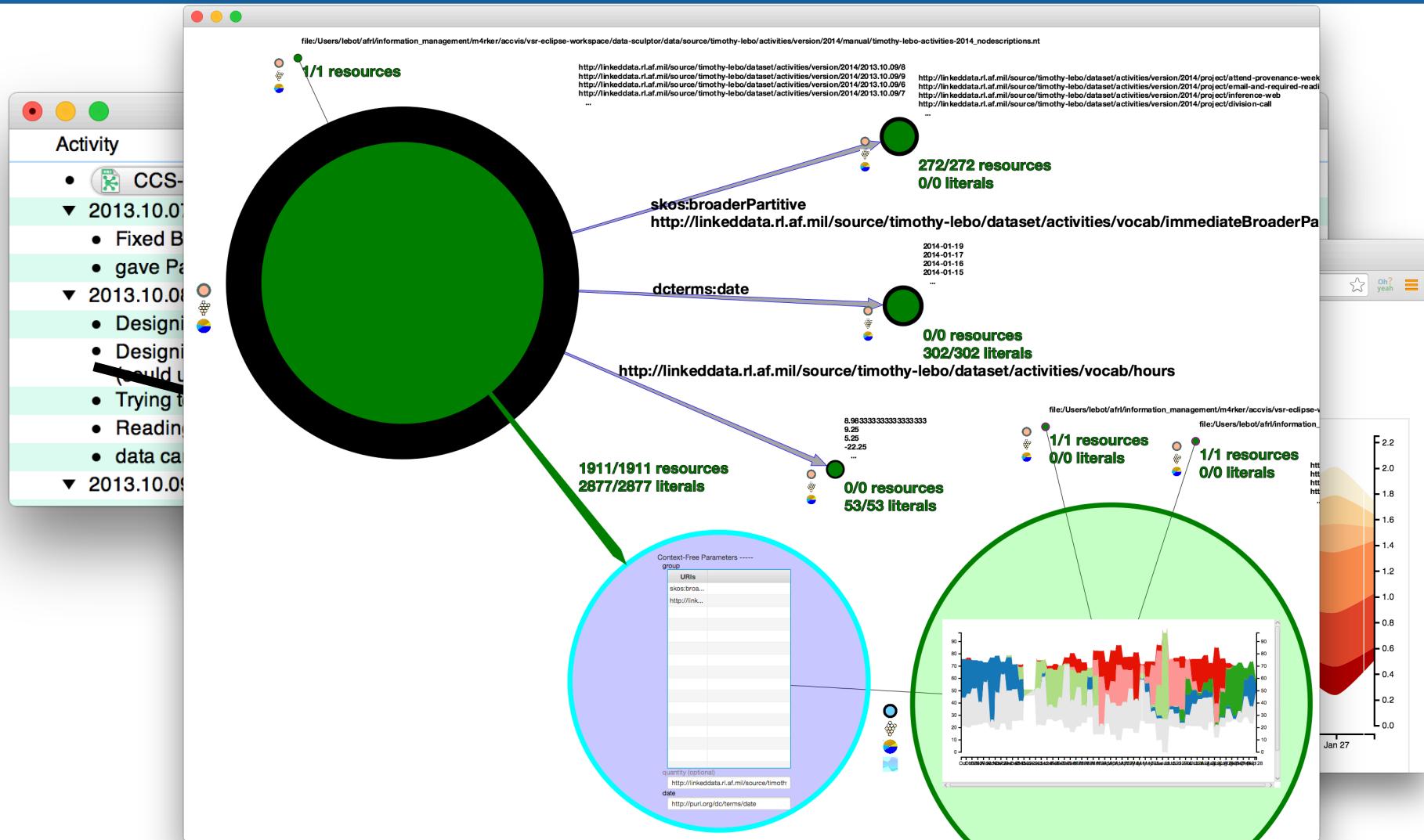
- Friday, 29 April 2011
  - Bootstrapping the "Model Task Force" Luc Moreau
  - Provenance Task Forces Luc Moreau
  - Re: Telecon Followup Daniel Garijo
- Thursday, 28 April 2011
  - Telecon Followup Paul Groth
- Tuesday, 26 April 2011
  - First Telecon Agenda Paul Groth
- Wednesday, 20 April 2011
  - Re: Working Group Telecon time Daniel Garijo
- Monday, 18 April 2011
  - Working Group Telecon time Paul Groth
- Friday, 15 April 2011
  - Re: Introduction Luc Moreau
- Thursday, 14 April 2011
  - RE: Introduction Deus, Helena
- Friday, 15 April 2011
  - Re: Introduction Iker Huerga
- Thursday, 14 April 2011
  - Re: Introduction Daniel Garijo
  - Introduction Paul Groth
  - Welcome to the Provenance WG Paul Groth
- Last message date: Friday, 29 April 2011 22:59:18 GMT
- Archived on: Thursday, 26 April 2012 13:06:28 GMT

13 messages sort by: [ thread ] [ author ] [ date ] [ subject ]  
Mail actions: [ mail a new topic ]

This archive was generated by [hypermail 2.2.0+W3C-0.50](#) : Thursday, 26 April 2012



# Application 1: Individual Time-Tracking





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

The image displays several overlapping windows illustrating different applications and data visualizations:

- Top Left Window:** A web browser showing the W3C public-prov-wg mailing list archive. It lists 13 messages from April 2011, sorted by date. The interface includes standard navigation buttons and a sidebar with links for Help and Mail actions.
- Top Center Window:** A circular visualization showing the distribution of resources and literals. A large green circle represents 8848/8848 resources (1332/1332 literals). A smaller black circle represents 0/0 resources (64/64 literals). A purple circle represents 0/0 resources (657/657 literals).
- Top Right Window:** A detailed view of the 0/0 resources (64/64 literals) circle. It shows a list of URIs and their dates, along with a list of authors: runnegar@w3.org, Creswell, Stephen, Eric, Jim McCusker, and others. Below this is a timeline visualization showing the frequency of messages over time.
- Bottom Left Window:** A screenshot of a software interface titled "Context-Free Parameters ----- group". It shows a table with columns for URIs, dcterms..., quantity (optional), and date. A specific row is highlighted with the URI <http://purl.org/dc/terms/date>.
- Bottom Center Window:** A circular visualization showing the distribution of resources and literals. A large green circle represents 0/0 resources (657/657 literals). A smaller black circle represents 1/1 resources (0/0 literals).
- Bottom Right Window:** A timeline visualization showing the frequency of messages over time, with a color scale from 0.0 to 2.2. The x-axis is labeled "Jan 27".

This archive was generated by [hypermail 2.2.0+W3C-0.50](#) : Thursday, 26 April 2012 13:06:28 GMT

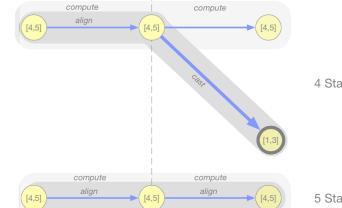
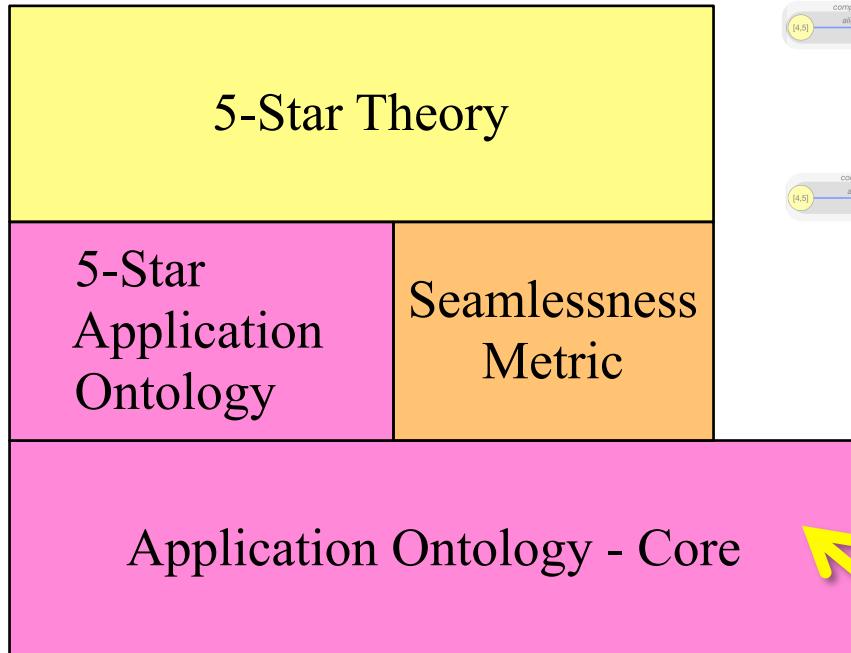


# Outline

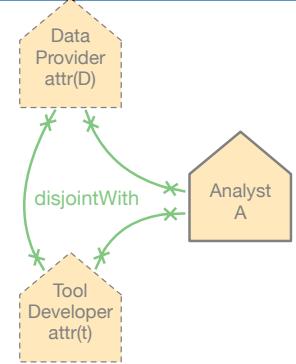
- Motivation
  - Visual Analytics [some problems]
  - Linked Data [some potential]
- Demo: Two-and-a-Half Example Applications
- A Nascent Theory
  - Application Ontology (a PROV extension)
  - Munging Ontology (7 types)
  - Cost Model (*moving* data across 5 stars)
  - 5-Star Applications (AO extension)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



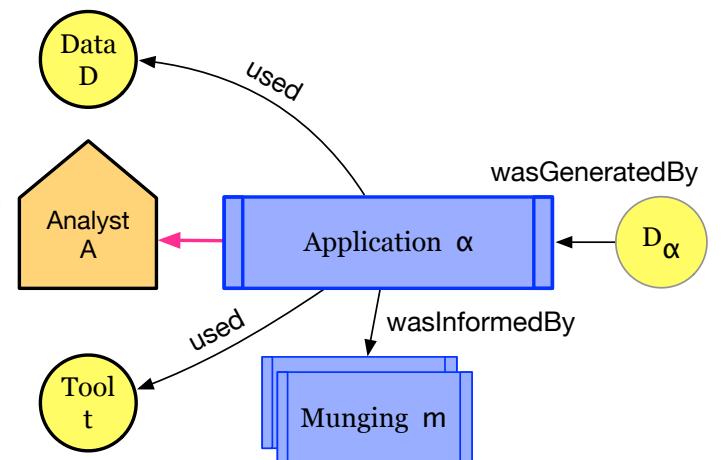
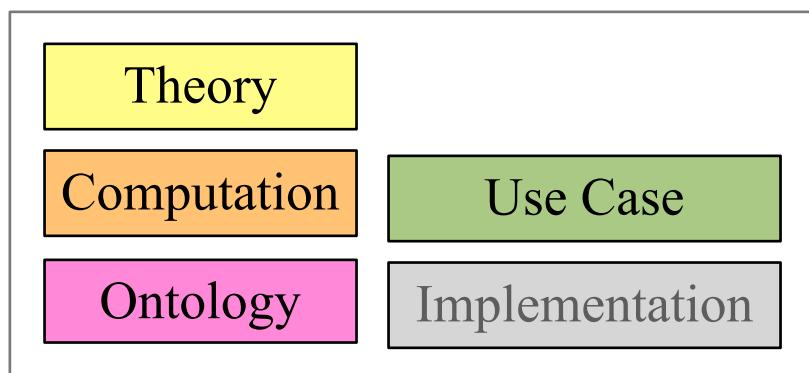
# Building to a Theory



[6, 38]  
[6, 24]



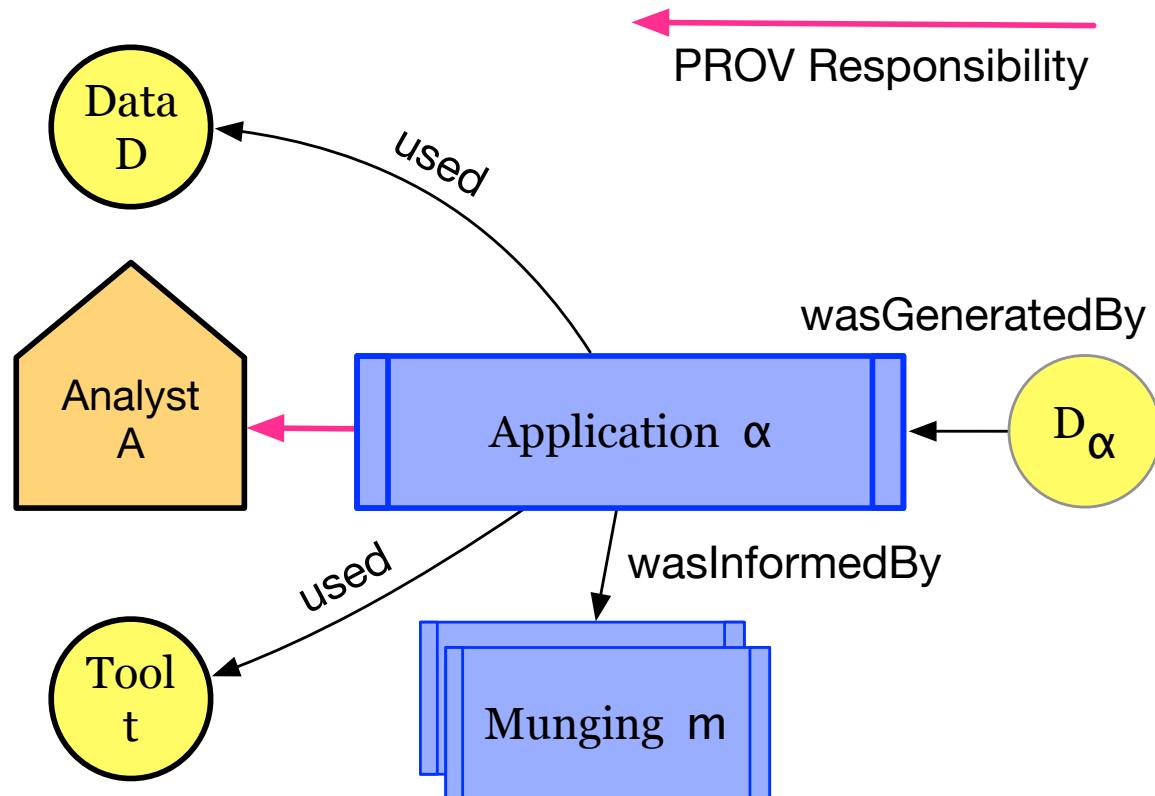
$$S_*(E) = \frac{\sum_{\alpha \in E} \sum_{m \in M_\alpha} cost(shim)}{\sum_{\alpha \in E} \sum_{m \in M_\alpha} pot(D_\alpha)cost(m)}$$





# 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.*

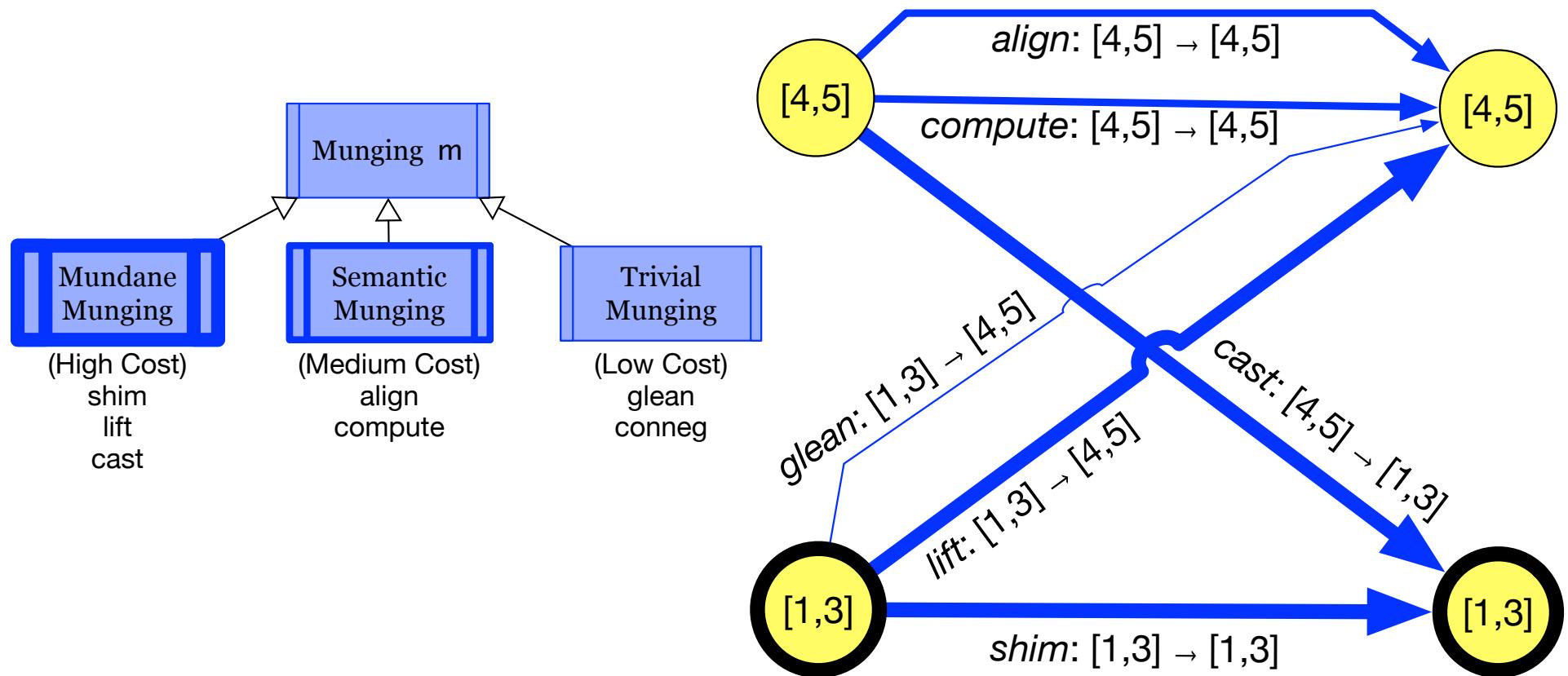


*Munging (or wrangling) is the imperfect manipulation of data into a usable form.*



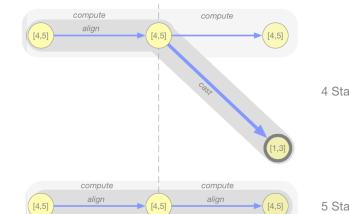
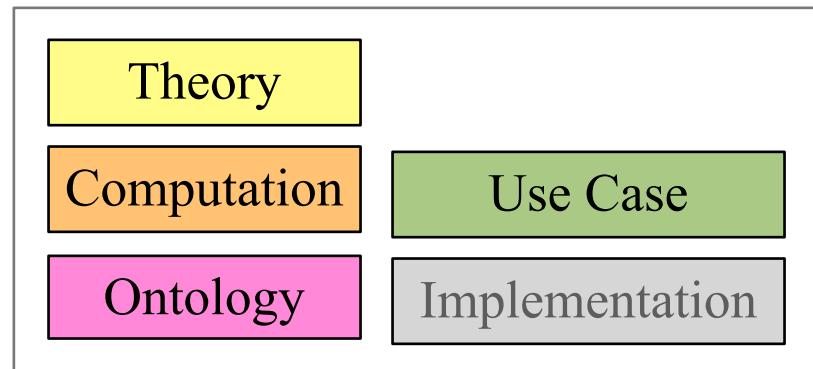
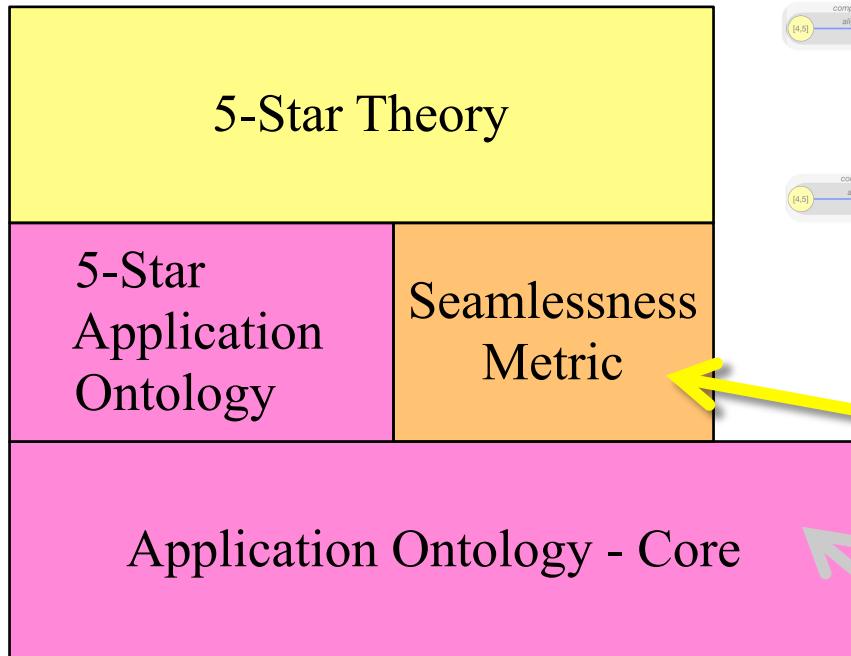
# Munging Ontology

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

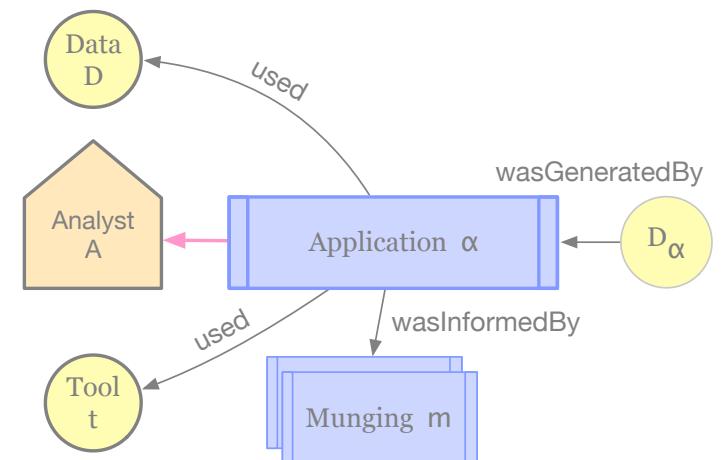
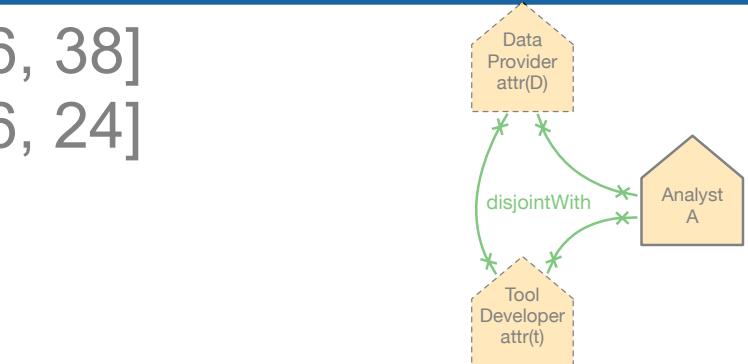




# Building to a Theory

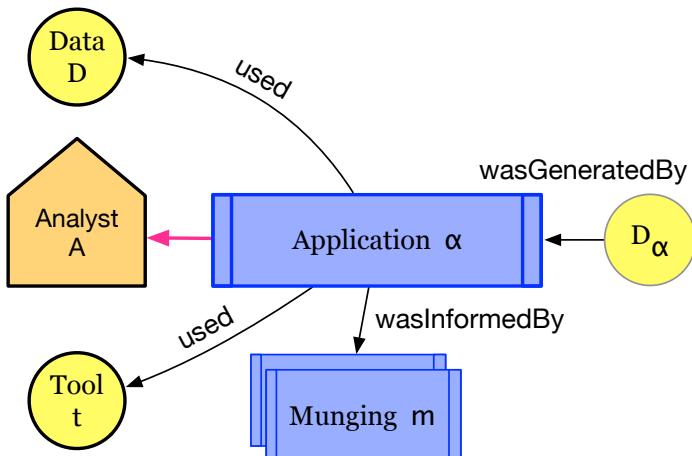


$$S_*(E) = \frac{\sum_{\alpha \in E} \sum_{m \in M_\alpha} cost(shim)}{\sum_{\alpha \in E} \sum_{m \in M_\alpha} pot(D_\alpha)cost(m)}$$





# Ecosystem Seamlessness



Let an analytical ecosystem  $E$  be the set of applications that influenced<sup>16</sup> a particular analysis:

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

Let an application  $\alpha$  be a tuple comprising its set of munging activities  $M$  and the resulting dataset  $D_\alpha$ :

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

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].

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



# Ecosystem Seamlessness: Full Ordering of Munge Costs

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

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

---

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

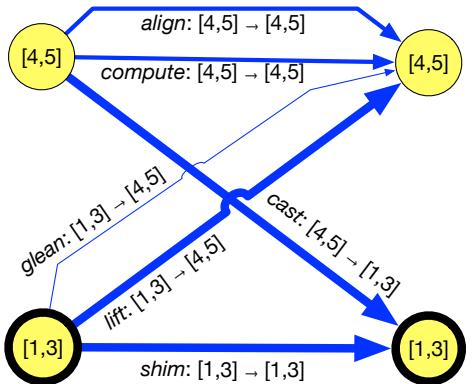
$$cost(align) > cost(comp)$$

---

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

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

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



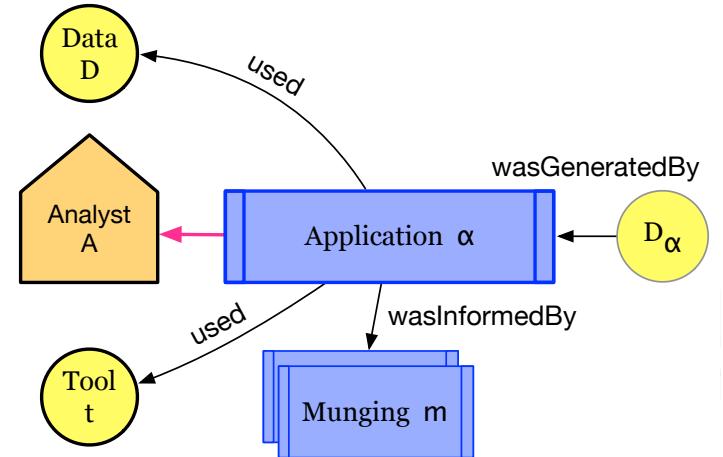


# Ecosystem Seamlessness

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

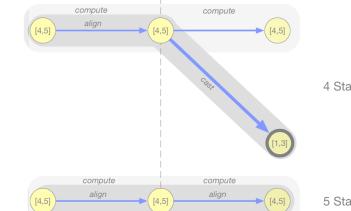
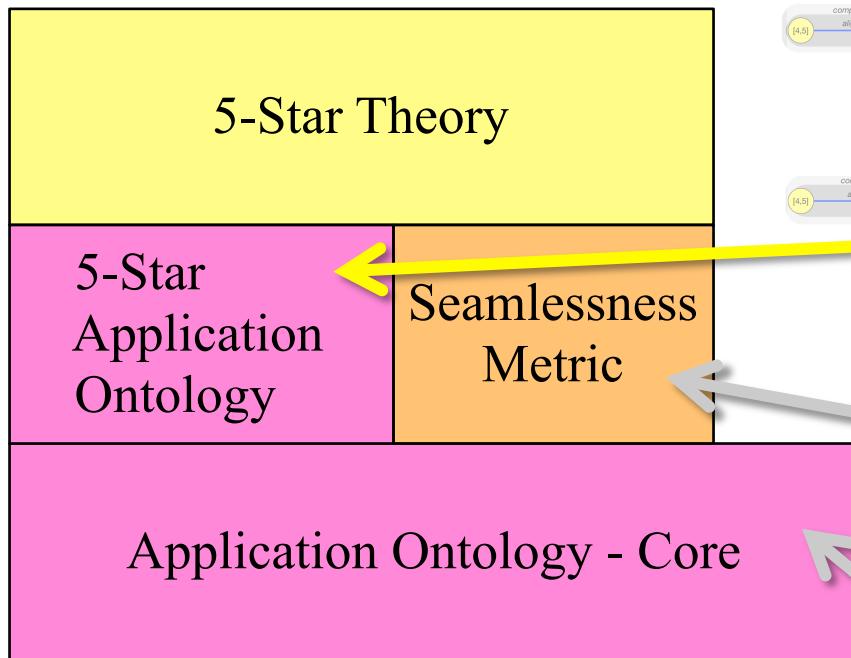
$$1 \leq S_*(E)$$

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

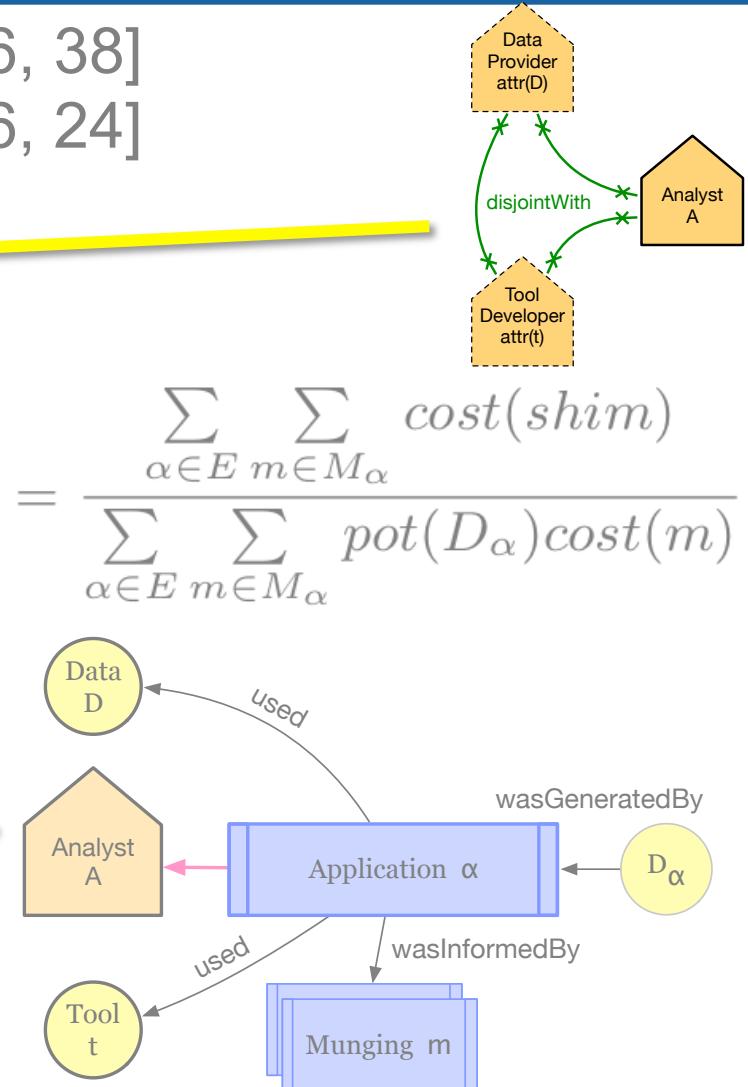




# Building to a Theory



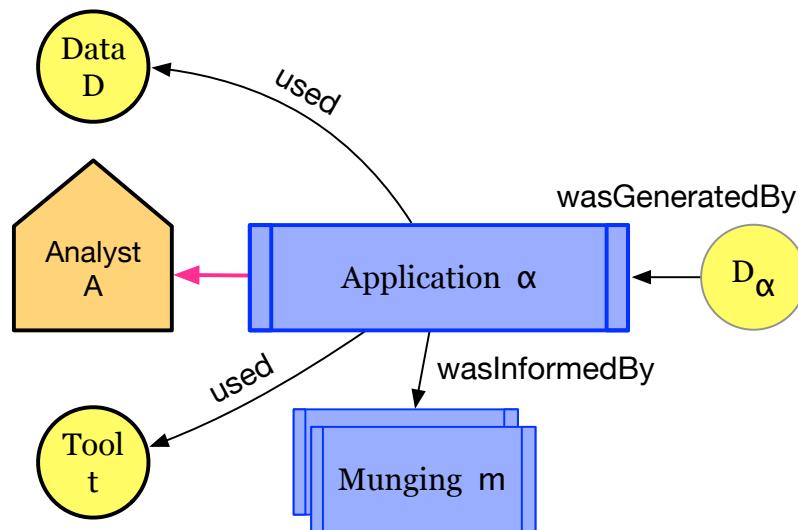
$$S_*(E) = \frac{\sum_{\alpha \in E} \sum_{m \in M_\alpha} cost(shim)}{\sum_{\alpha \in E} \sum_{m \in M_\alpha} pot(D_\alpha)cost(m)}$$





# 5-Star Applications

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

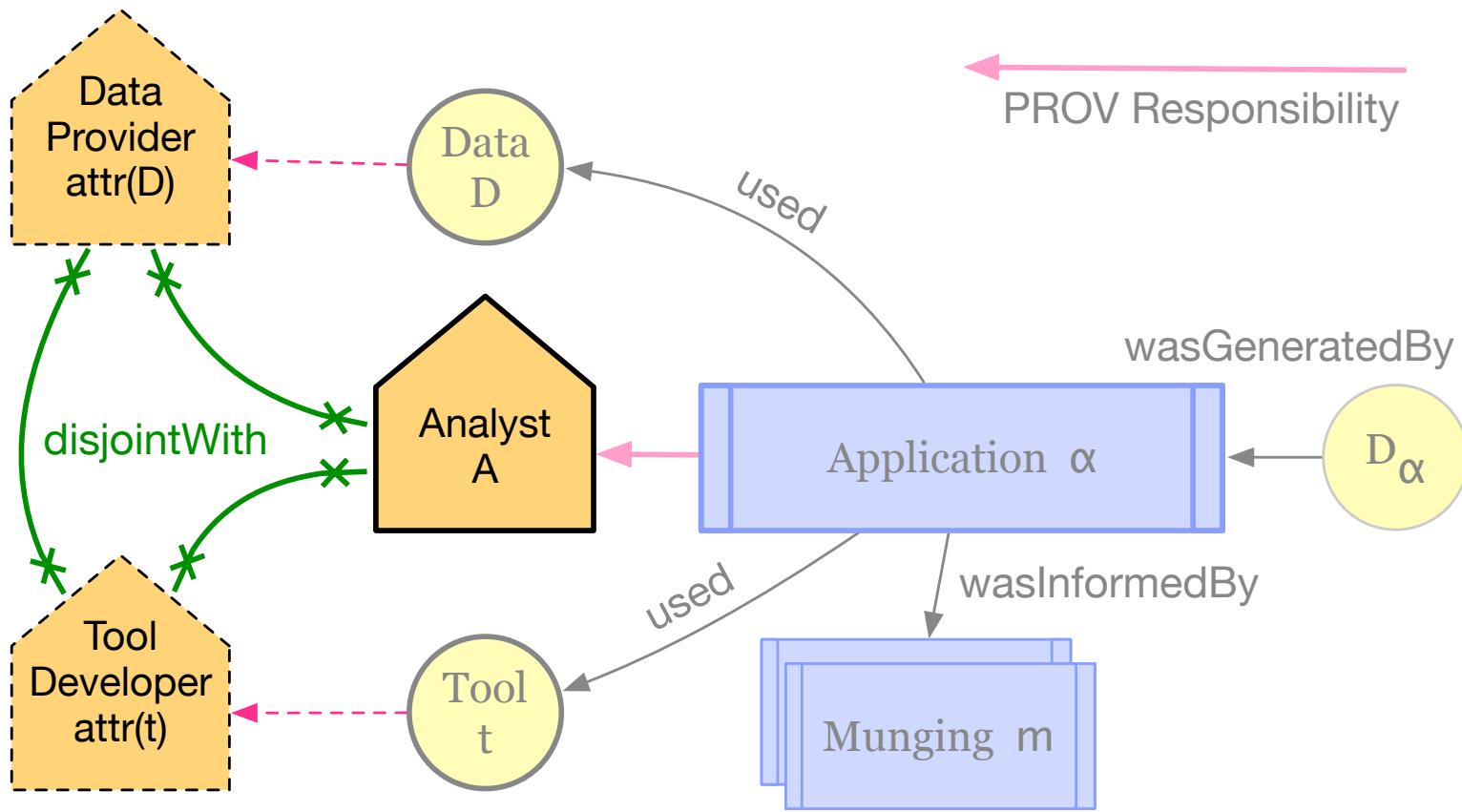


☆	Informal Restriction	Formally
1	data providers, analysts, and tool developers are disjoint	$attr(D) \cap A \cap attr(t) = \emptyset$
2	accept data (any format) via URL; cite that URL in the future	$tbl(D) \geq 1 \wedge URL \in D_\alpha$
3	accept data (RDF format) via URL; cite that URL in the future	$tbl(D) \geq 4$
4	use a tool's input semantics (OWL, SPARQL) when performing 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

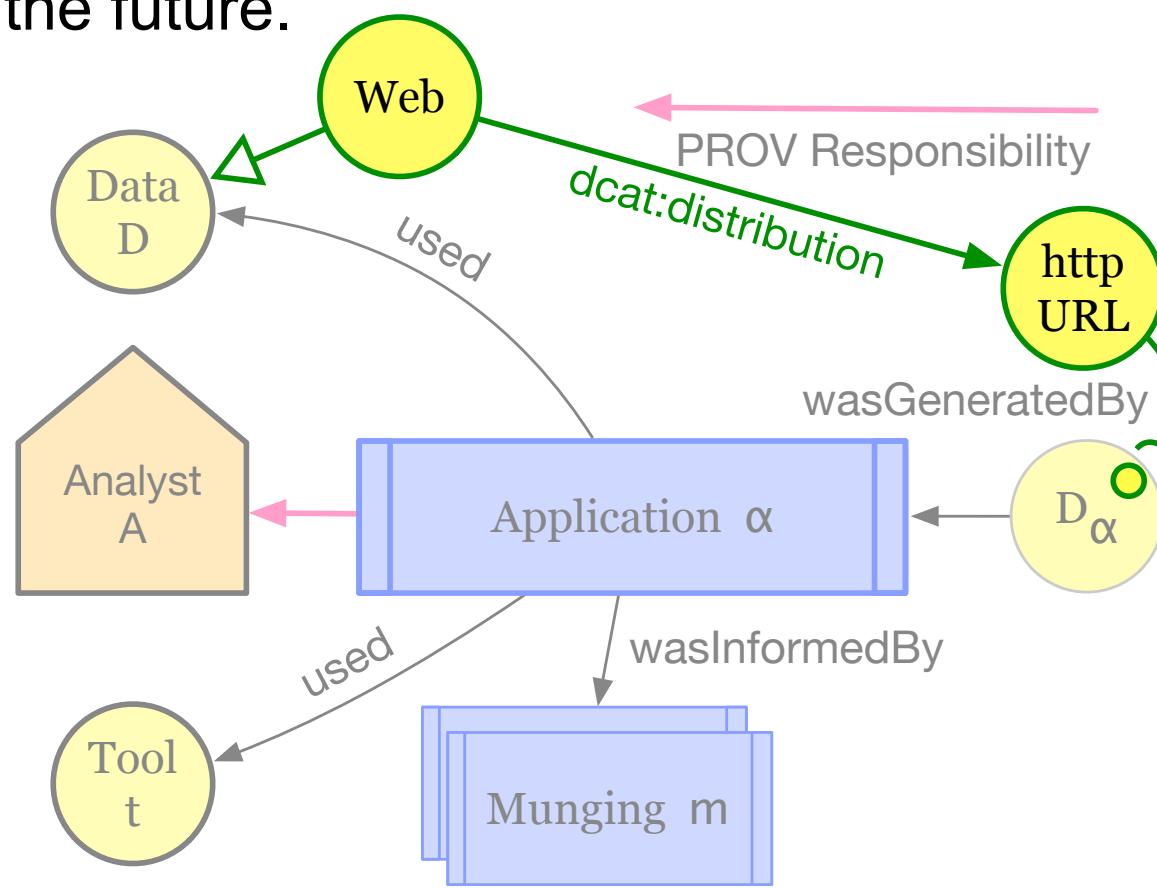
Data providers, analysts, and tool developers are disjoint.





# 2-Star Applications

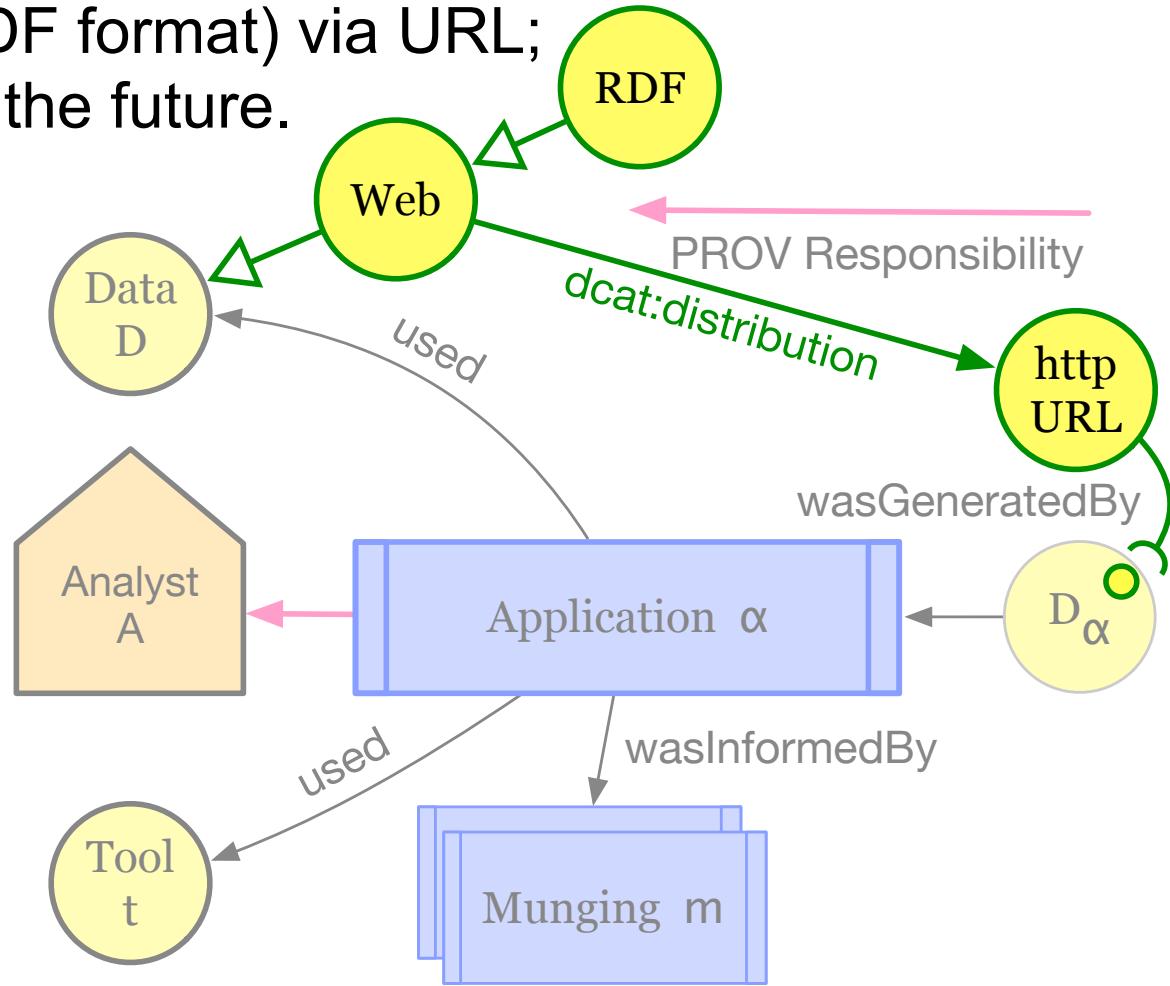
Accept data (any format) via URL;  
cite that URL in the future.





# 3-Star Applications

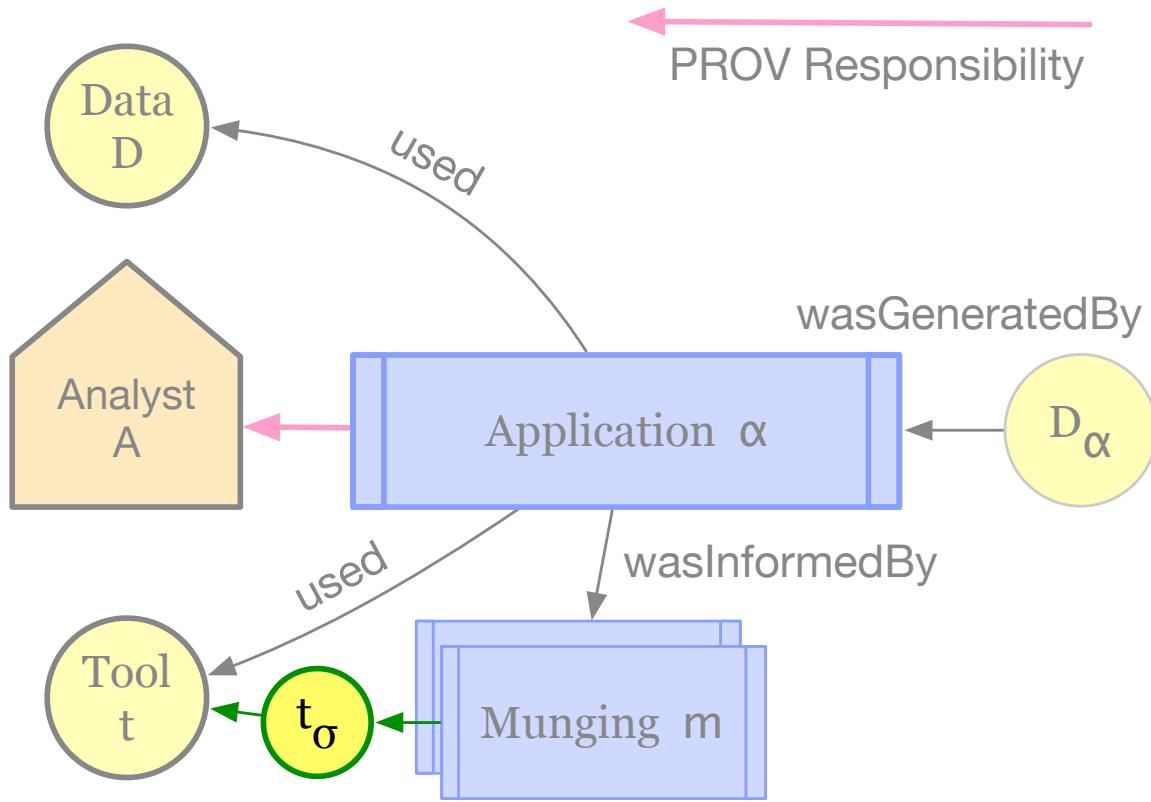
Accept data (RDF format) via URL;  
cite that URL in the future.





# 4-Star Applications

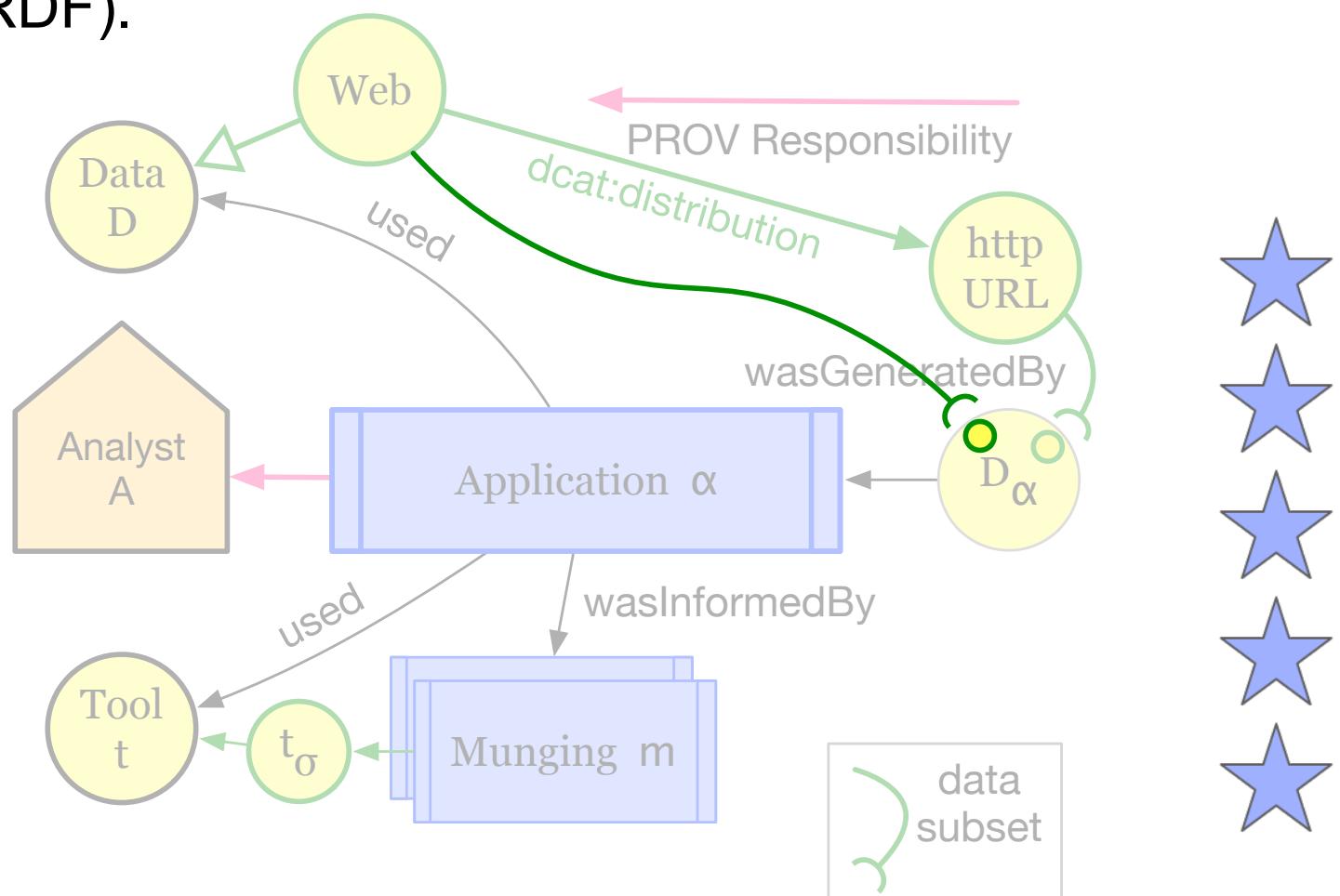
Use a tool's input semantics (OWL, SPARQL) to munge.





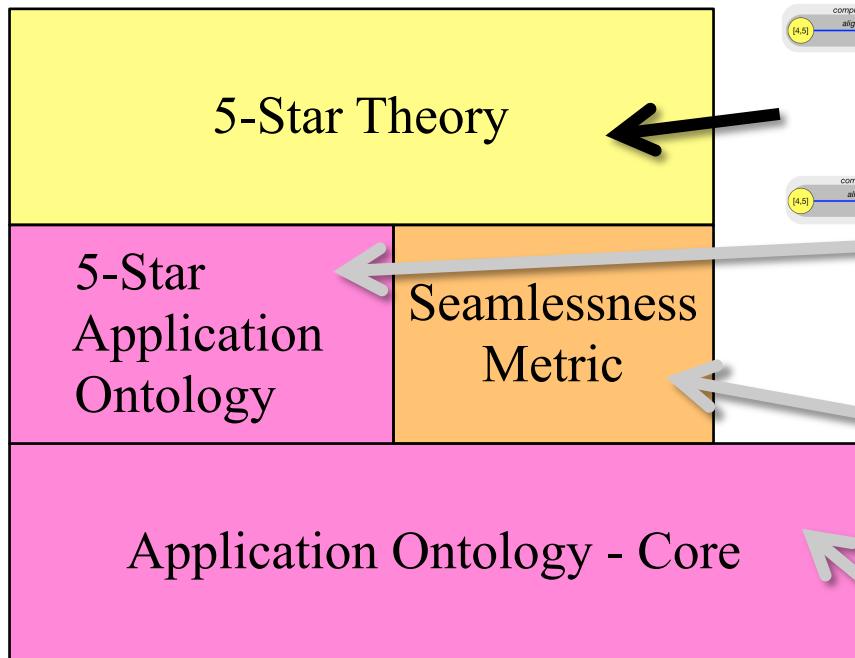
# 5-Star Applications

Provide any information derived during use (as RDF).

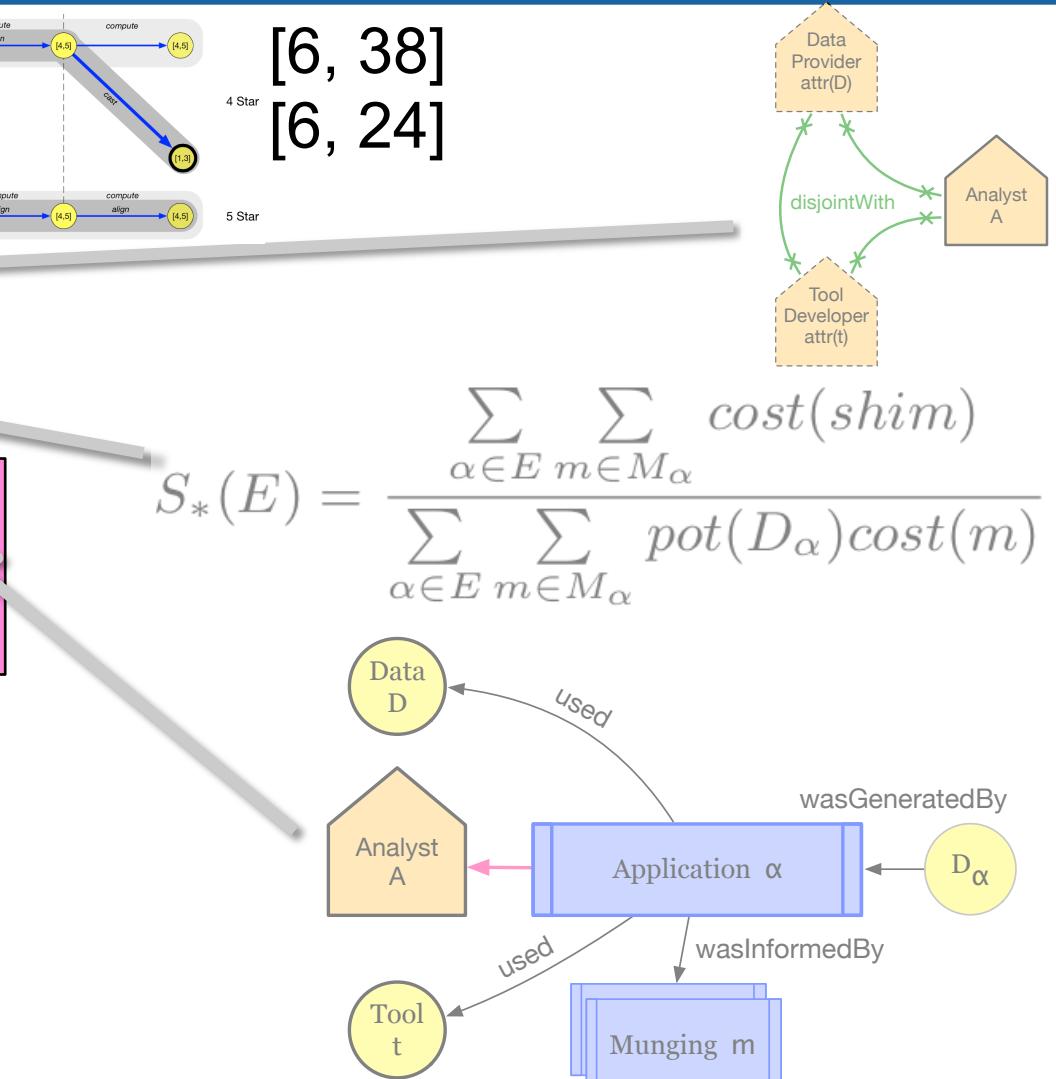
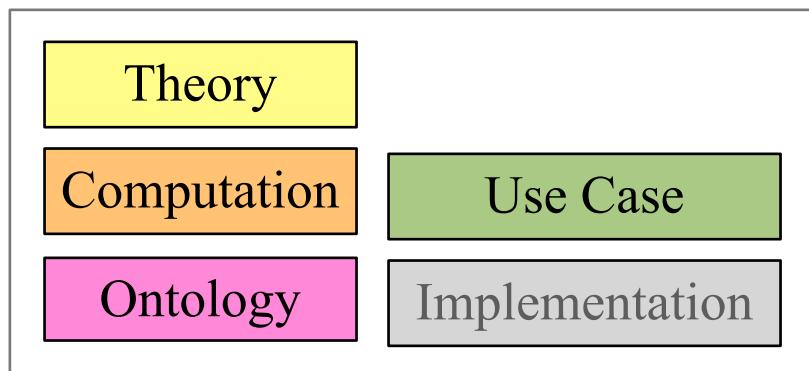




# Building to a Theory

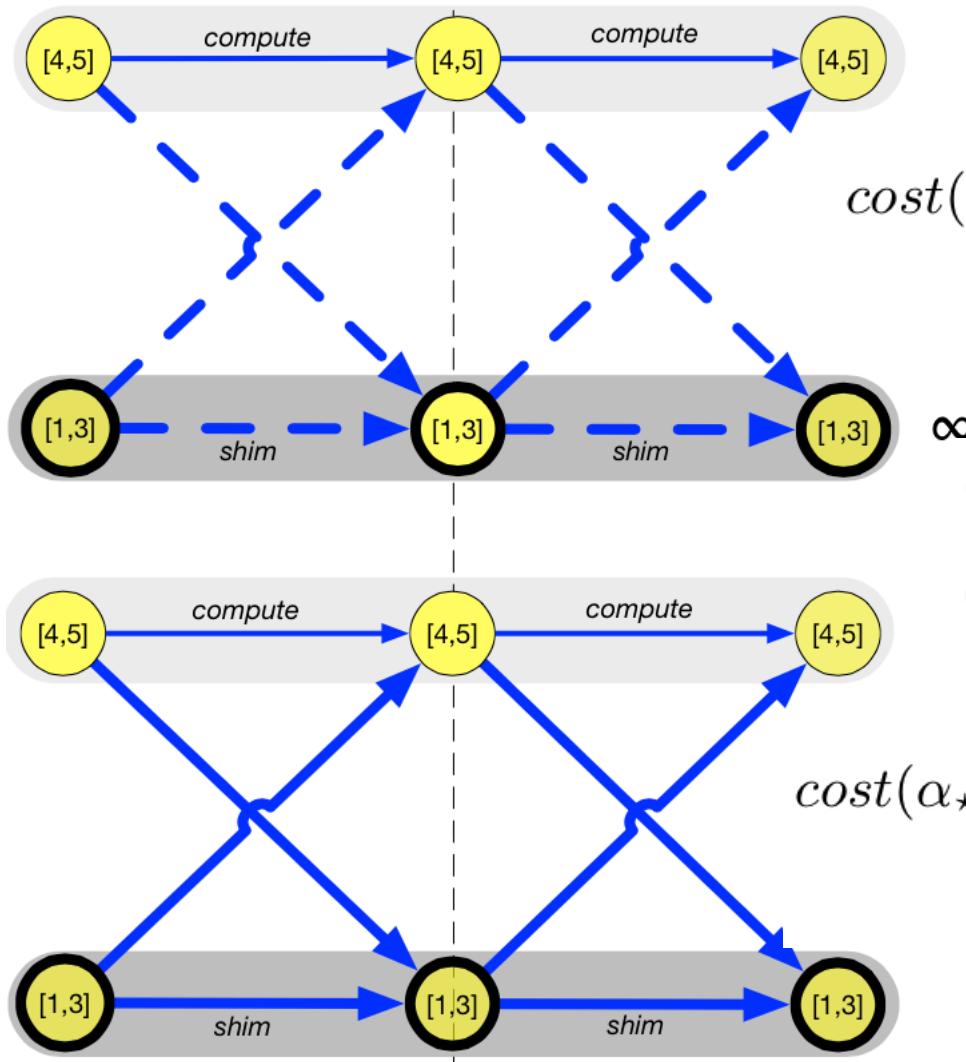


$$S_*(E) = \frac{\sum_{\alpha \in E} \sum_{m \in M_\alpha} cost(shim)}{\sum_{\alpha \in E} \sum_{m \in M_\alpha} pot(D_\alpha)cost(m)}$$





# Predicted Cost Bounds for any Hypothetical Application Chain



$$\begin{aligned} \text{cost}(\alpha_*) &= [2 \times \text{cost}(comp), \infty] \\ &= \underline{[6, \infty]} \end{aligned}$$

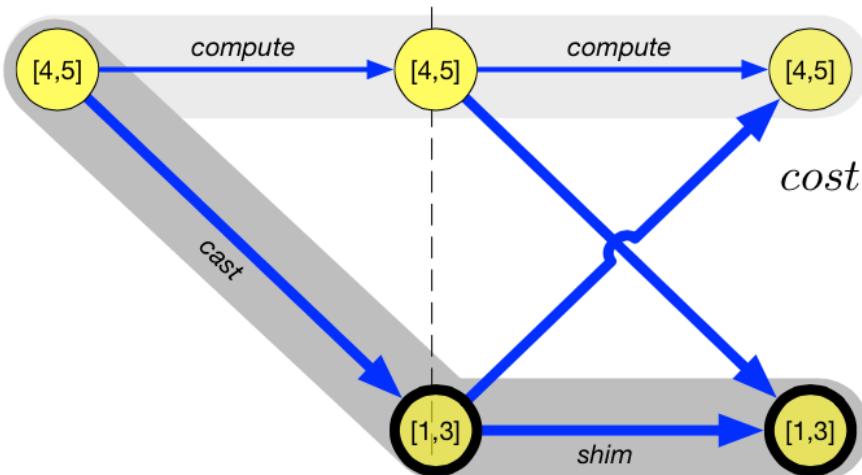
Lower Cost Bound Path

Upper Cost Bound Path

$$\begin{aligned} \text{cost}(\alpha_{**}) &= [2 \times \text{cost}(comp), 2 \times \text{cost}(shim)] \\ &= [6, 38] < \text{cost}(\alpha_*) = [6, \infty] \end{aligned}$$



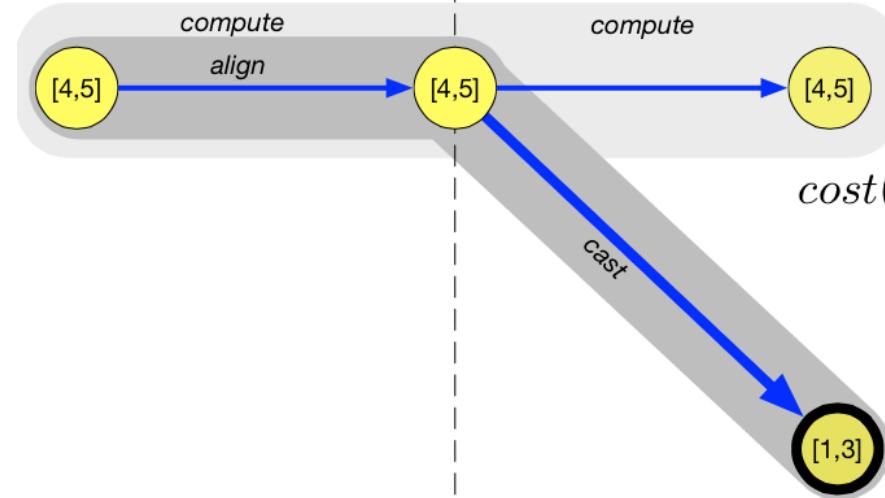
# Predicted Cost Bounds for any Hypothetical Application Chain



$$\begin{aligned} \text{cost}(\alpha_{***}) &= [2 \times \text{cost}(comp), \text{cost}(cast) + \text{cost}(shim)] \\ &= [6, 24] < \text{cost}(\alpha_{**}) = [6, 38] \end{aligned}$$

Lower Cost Bound Path

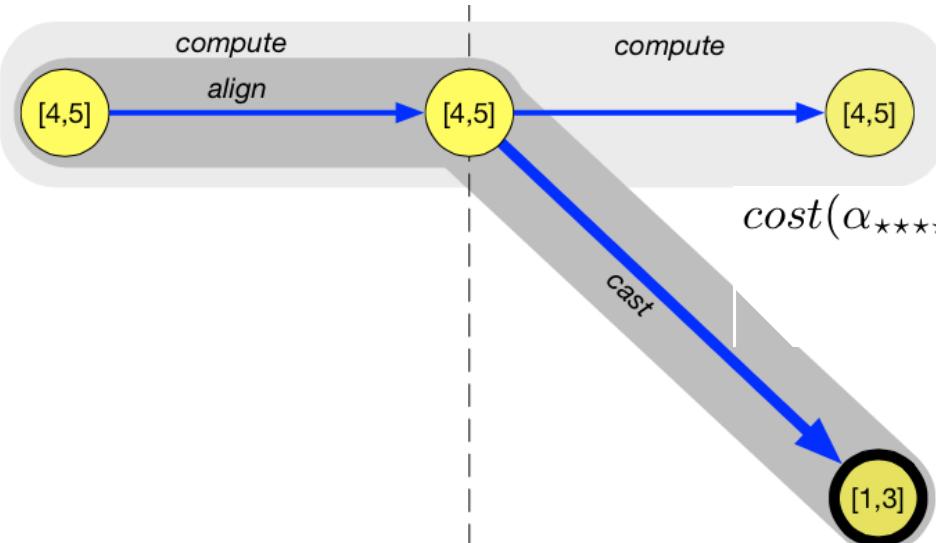
Upper Cost Bound Path



$$\begin{aligned} \text{cost}(\alpha_{****}) &= [2 \times \text{cost}(comp), \text{cost}(align) + \text{cost}(cast)] \\ &= [6, 9] < \text{cost}(\alpha_{***}) = [6, 24] \end{aligned}$$

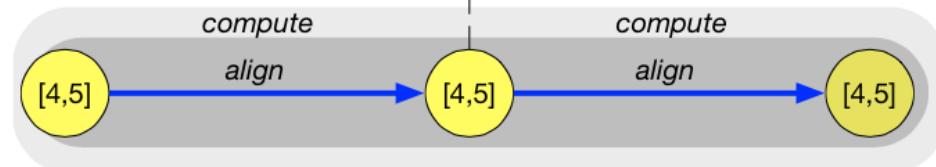


# Predicted Cost Bounds for any Hypothetical Application Chain



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

---



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

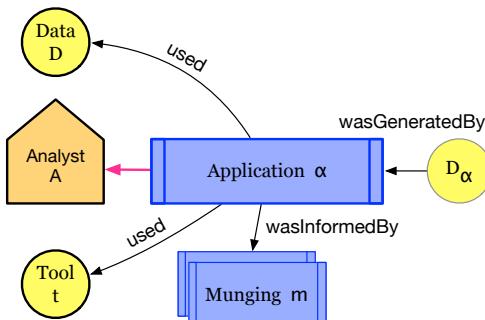
---

- Lower Cost Bound Path
- Upper Cost Bound Path

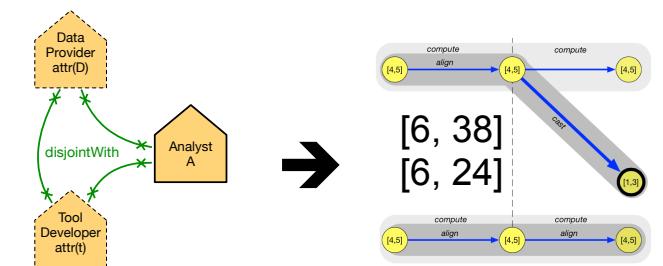


# Conclusion

- 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

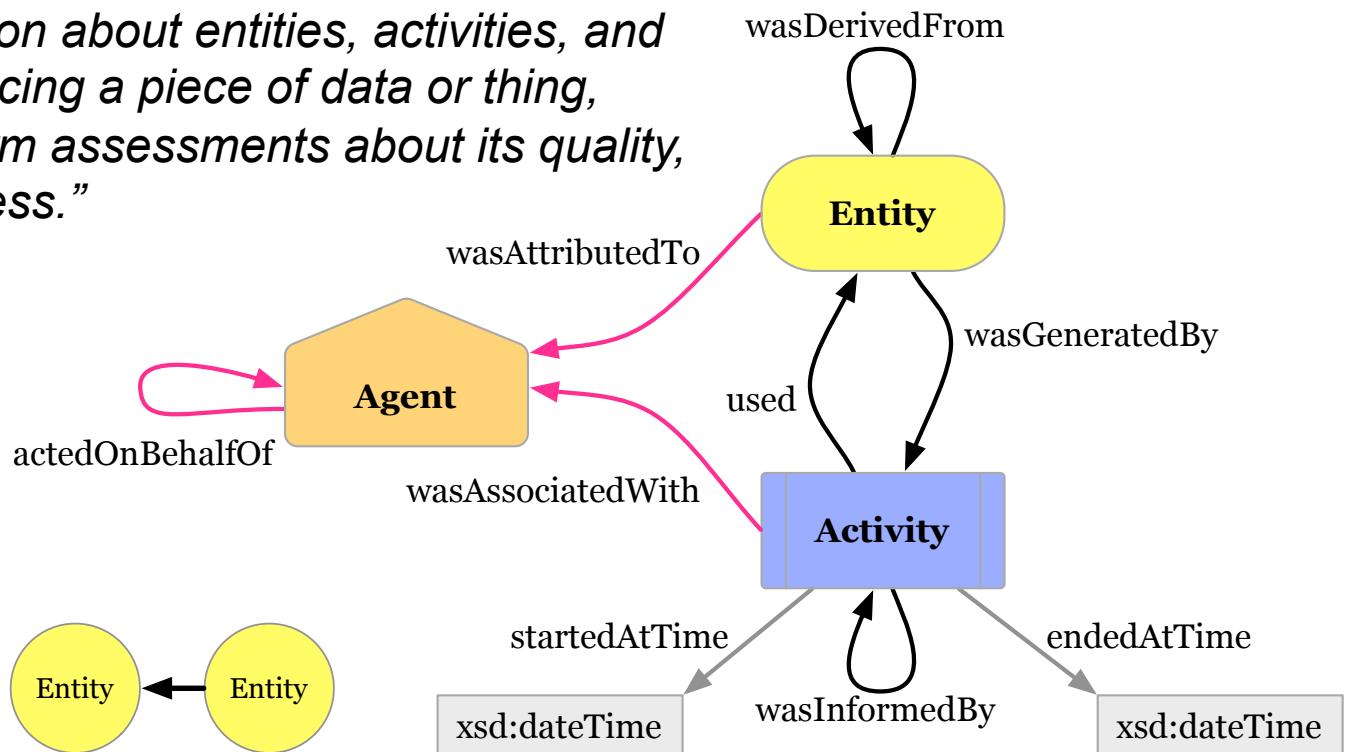
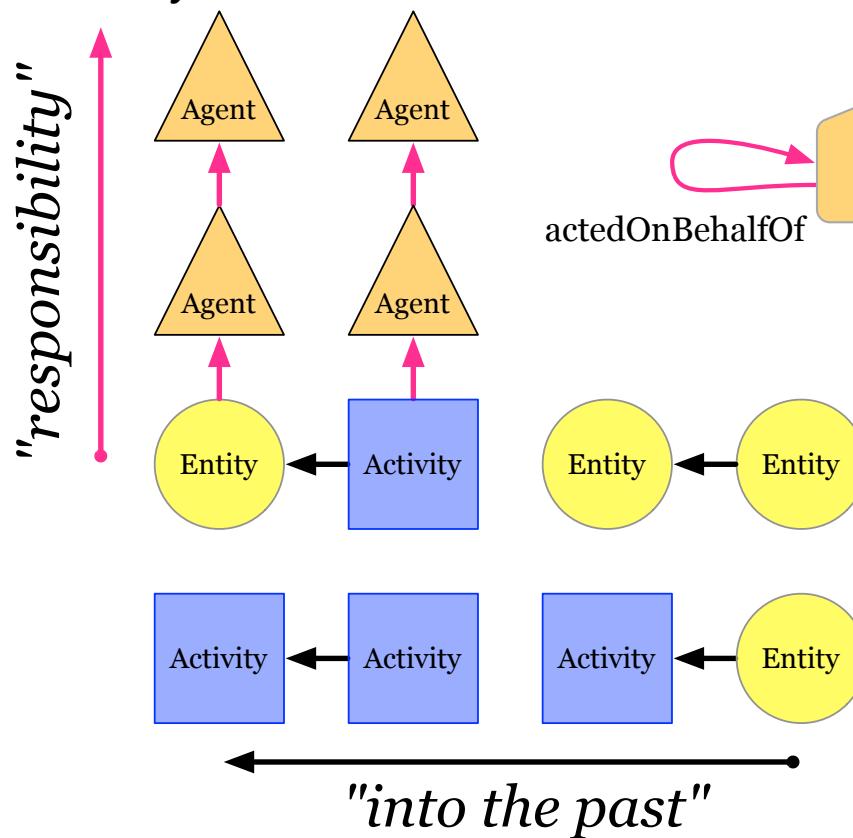
- S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2917–2926, Dec. 2012.
- J. Thomas and K. Cook. *Illuminating the path: The research and development agenda for visual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



# W3C PROVance

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>