



<http://bit.ly/lebo-halloween-issues>

Towards Data-Centric Digital Communities

*An Example for the Department of
Health and Human Services*



Timothy Lebo
Tetherless World Constellation
Rensselaer Polytechnic Institute



Rensselaer



Motivation: Avoiding Archaeological Endeavors

*What are the **things** that you're talking about?
How do those things **relate**?*

RUBES by Leigh Rubin



At last, the mystery of the
Mayan calendar revealed.



A few years later, or the **same moment**
somewhere else on the web.



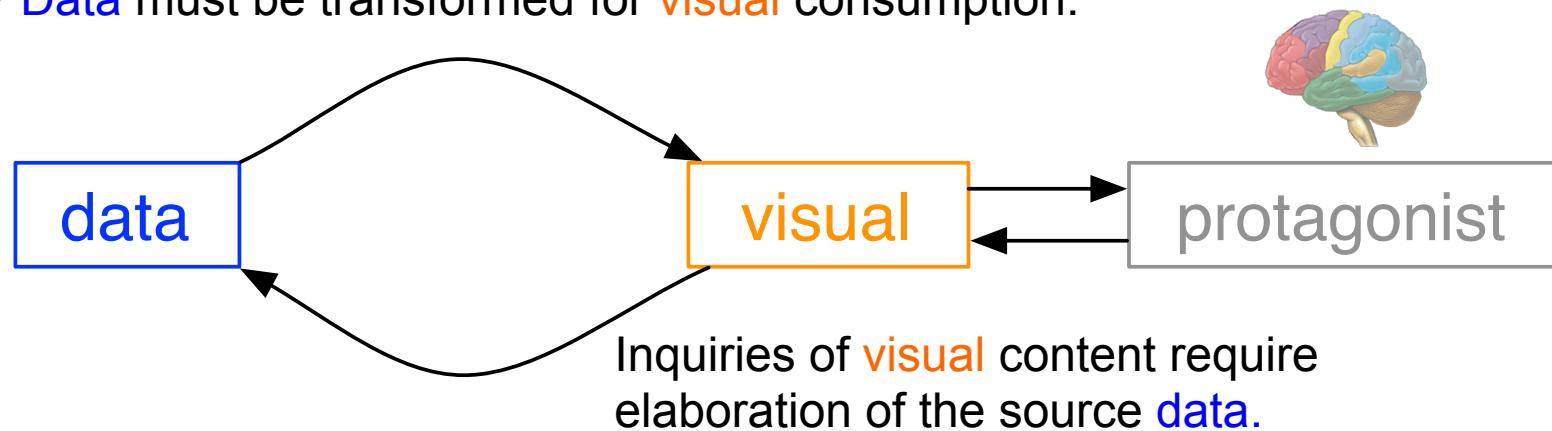
Abstract Objective:

Protagonist is more informed about the world.

Protagonist cannot consume **data** directly.

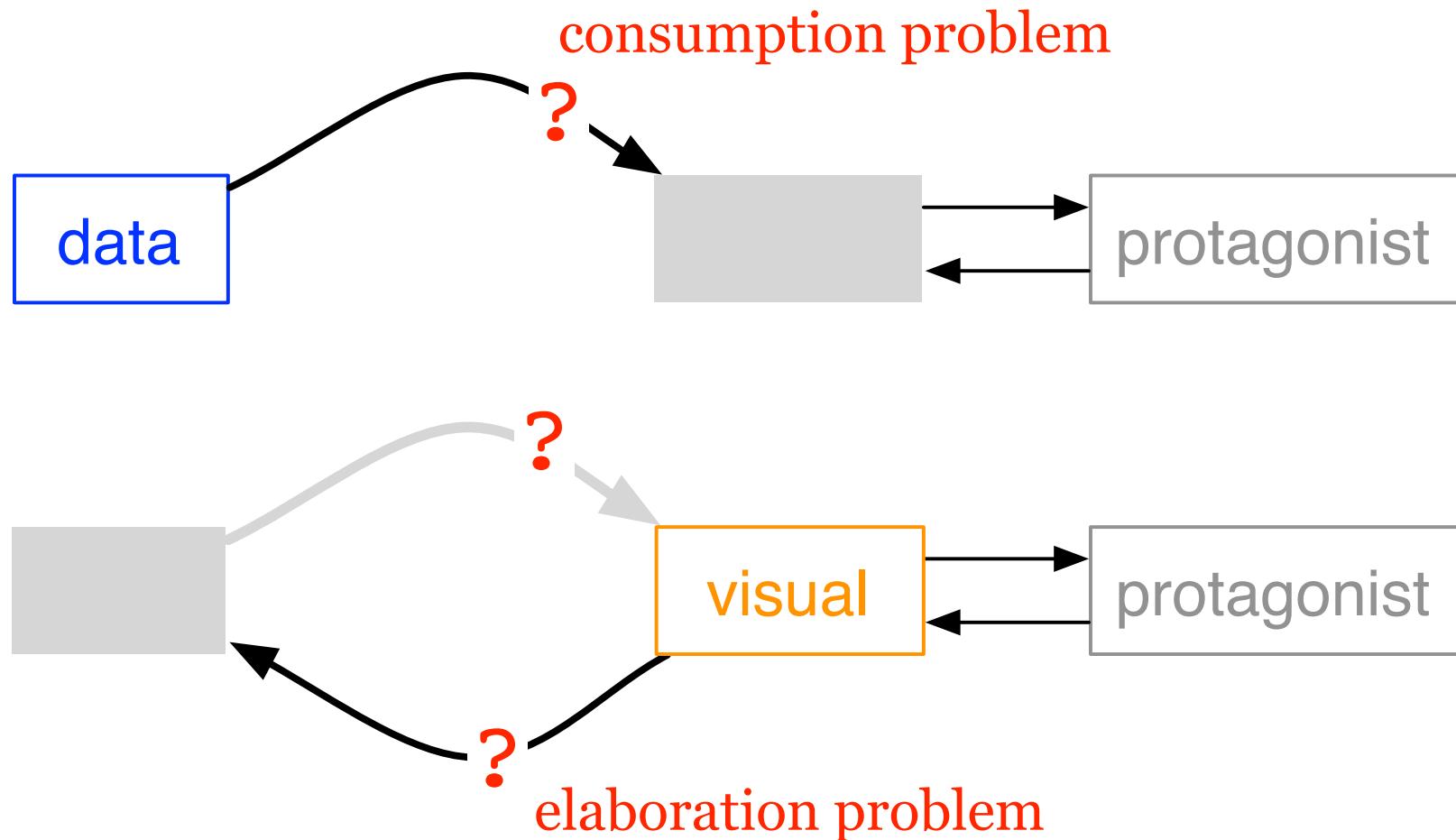
Vision is the predominant method of consumption.

→ **Data** must be transformed for **visual** consumption.



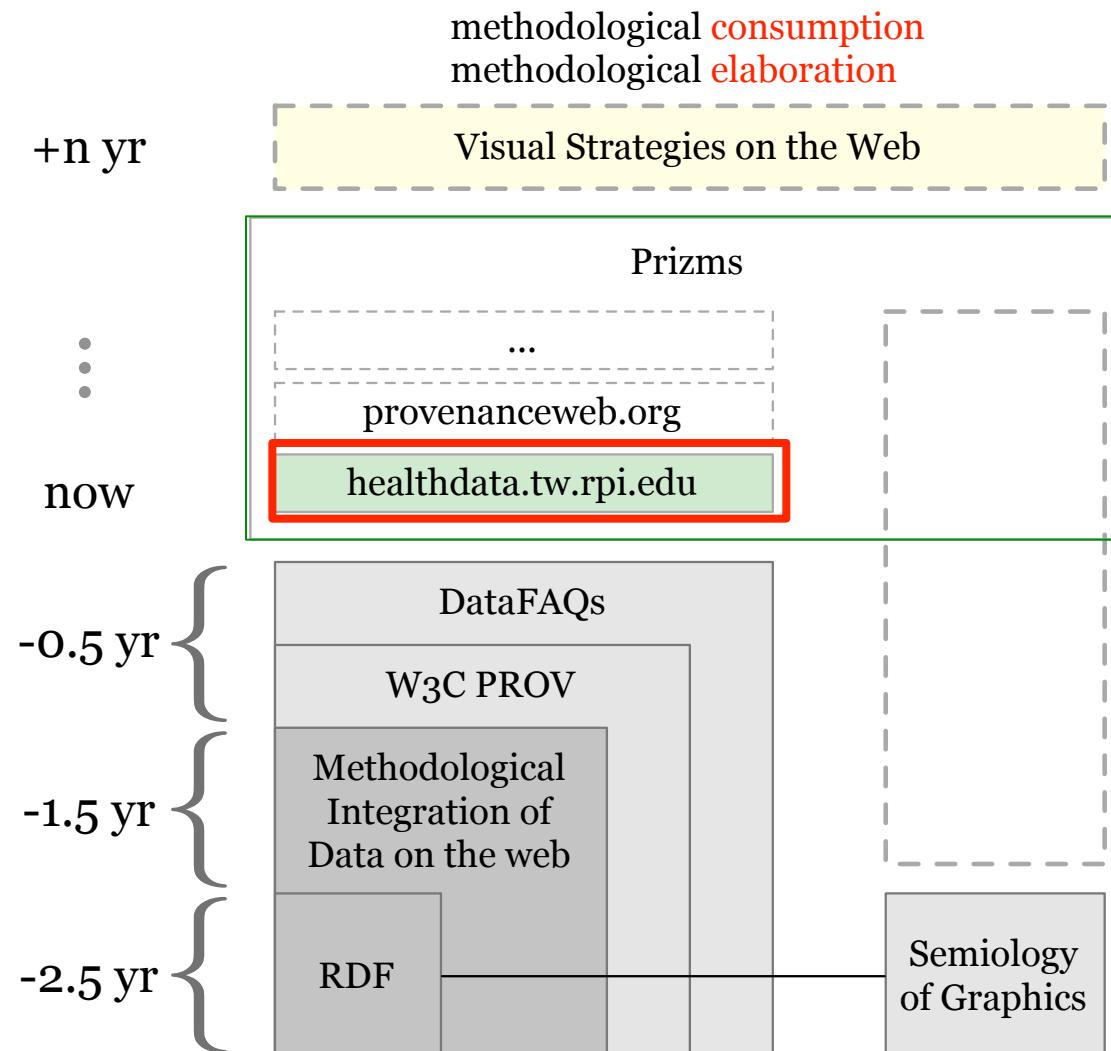


Two Symmetric Challenges





Talk Overview





Dept. Health and Human Services Developer Challenges

“Establish learning communities that collaboratively evolve and mature the utility and usability of a broad range of health and human service data.”

“Facilitate better integration, analysis, and interpretation of our data, helping meet the growing demand for value added information services.”

HealthData.gov

<http://www.healthdata.gov/developer-challenges-overview>



http://hub.healthdata.gov

349
dataset
pages



http://healthdata.tw.rpi.edu

A screenshot of a web browser displaying the homepage of healthdata.tw.rpi.edu. The browser window shows the title 'Welcome to healthdata.tw.rpi.edu' and the URL 'healthdata.tw.rpi.edu/index.html'. The website has a dark header bar with links for Home, Linked Datasets, CKAN Datasets, Named Graphs, and Vocabularies. The main content area features three sections: 'Explore our datasets', 'CKAN repository', and 'Behind the curtain'.

Welcome to healthdata.tw.rpi.edu

Explore our datasets
Check the [datasets we have available](#) as Linked RDF.

CKAN repository
Check out our [CKAN repository](#), a [writable mirror](#) of <http://hub.healthdata.gov>.

Behind the curtain
This site is maintained through our [github repository](#); see the [wiki](#). The RDF data is available in [this SPARQL endpoint](#). All data that we provide on this site can be automatically reproduced from the original government sources using [csv2rdf4lod](#).



Dataset listings

The datasets listed here were derived from the datasets listed at <http://hub.healthdata.gov>. The CKAN instance at <http://healthdata.tw.rpi.edu/hub> mirrors the listings at <http://hub.healthdata.gov> and contains additional annotations. Many of the datasets listed in the CKAN instances were converted to RDF.

Content

- [Stage 9: RDF Datasets with Shared Vocabulary](#)
- [Stage 8: RDF Datasets with Shared Resources](#)
- [Stage 7: Enhanced RDF Datasets](#)
- [Stage 6: RDF Datasets with similar predicates](#)
- [Stage 5: RDF Datasets with similar objects](#)
- [Stage 4: RDF Datasets](#)
- [Stage 3: Retrieved, but Defunct tabular CKAN Datasets](#)
- [Stage 2: Unretrieved Tabular CKAN Datasets with Distribution Metadata](#)
- [Stage 1: CKAN Datasets](#)

Linked Datasets that Share Vocabulary
These RDF datasets share vocabulary.

Linked Datasets that Share Resources
These RDF datasets mention common resources.

- [hub-healthdata-gov hospital-compare](#)

Datasets listings are partitioned according to their Linked Data maturity (by querying their provenance).



Dataset metadata

Screenshot of a web browser showing dataset metadata for "hub-healthdata-gov/hospital-compare".

The page includes:

- Datasets with common predicates:** No datasets found.
- Datasets with common vocabularies:** No datasets found.
- Datasets with shared raw columns:**
 - hub-healthdata-gov food-recalls
- Dataset Modifications:** A timeline visualization showing modifications over time. Blue circles represent the number of modifications made to this dataset, while gray circles represent the number of modifications made to other datasets on a given day.
- Attribute Value Table:**

Attribute	Value
Identifier	hub-healthdata-gov hospital-compare

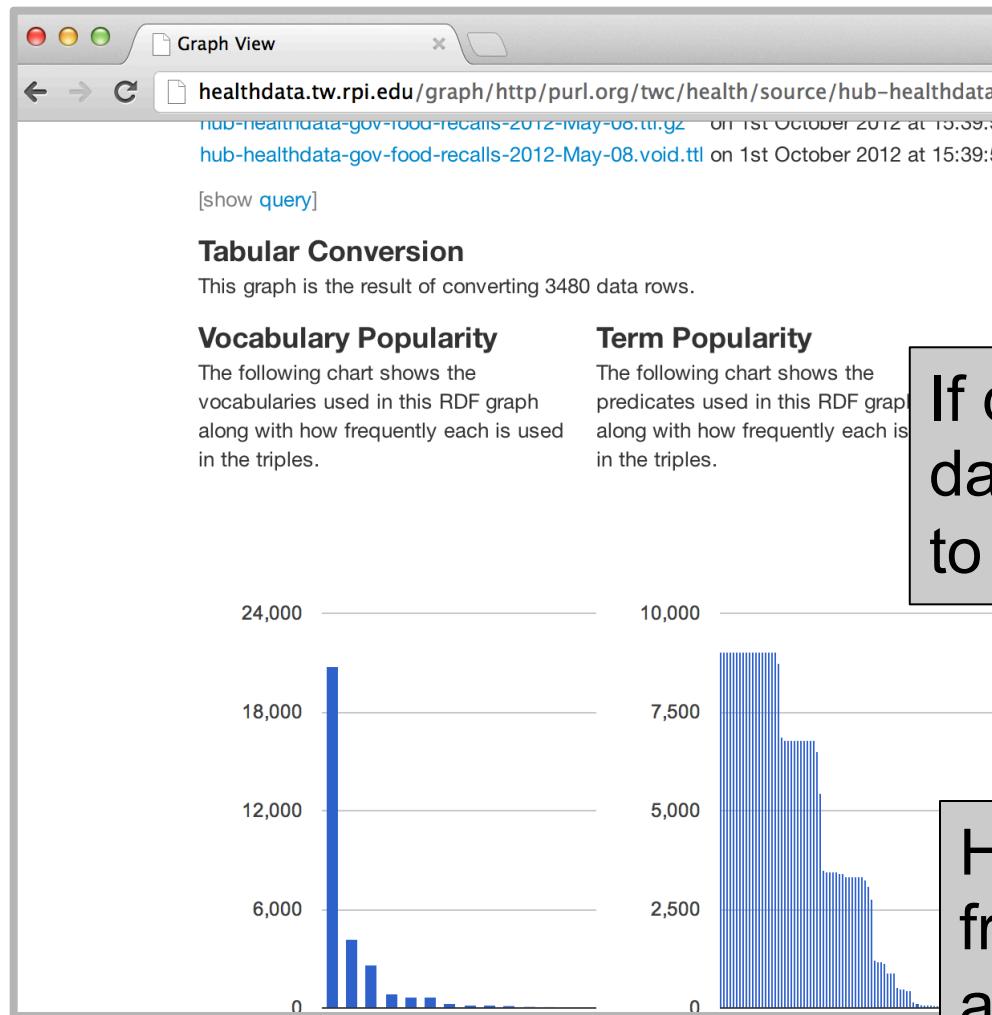
Links back to the government's original dataset listing.

Dataset pages show connections to other datasets.

Dataset modification times are shown relative to all other dataset modification times.



Overview of loaded graphs



Links to the loaded files, and the time they were loaded (a “provenance freebie”).

If derived from tabular data, a row count is given to indicate dataset size.

Histograms show how frequently class and properties are used in the data.



Reusing Others' Tools

CKAN mirror

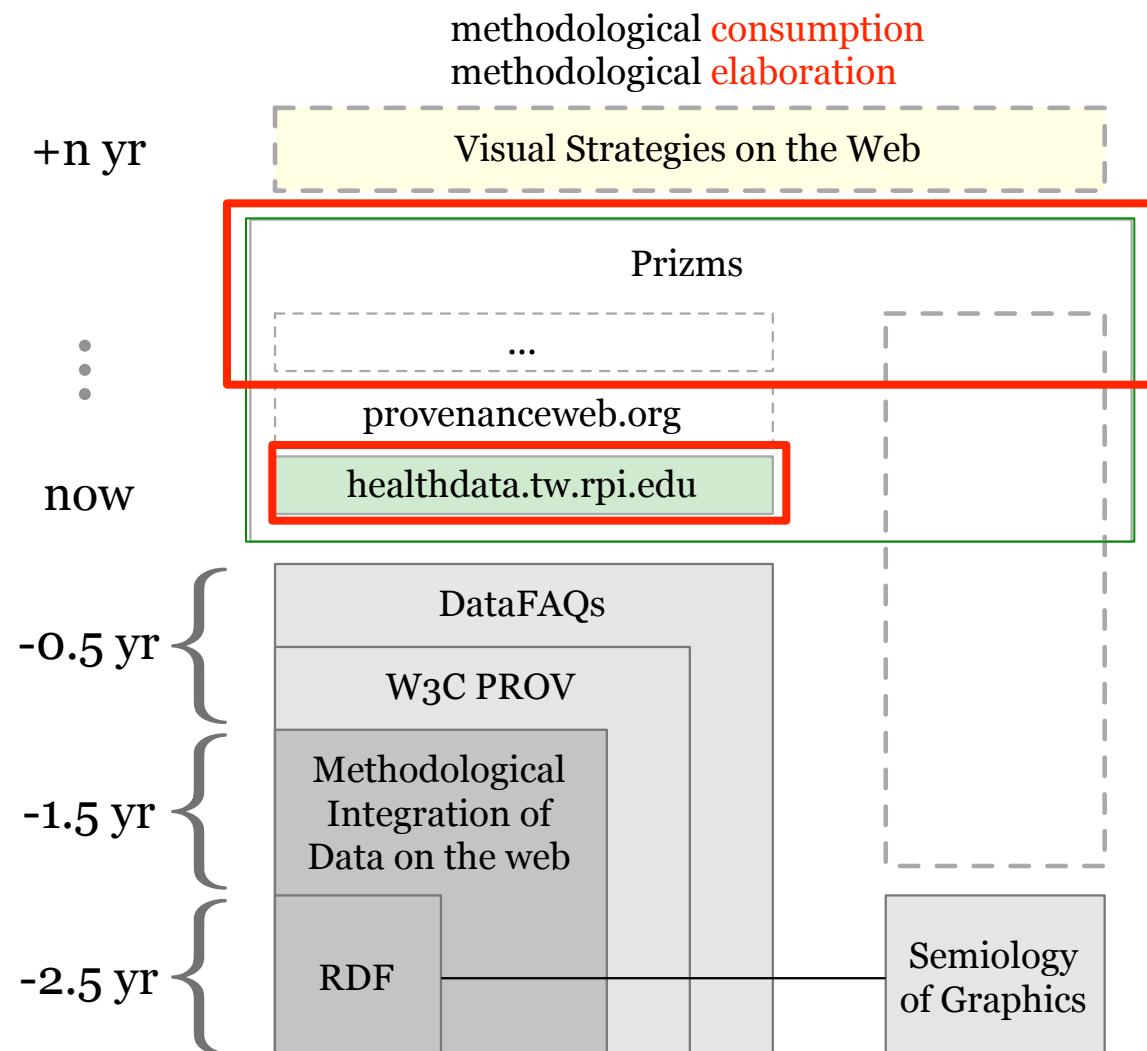
SPARQL endpoint

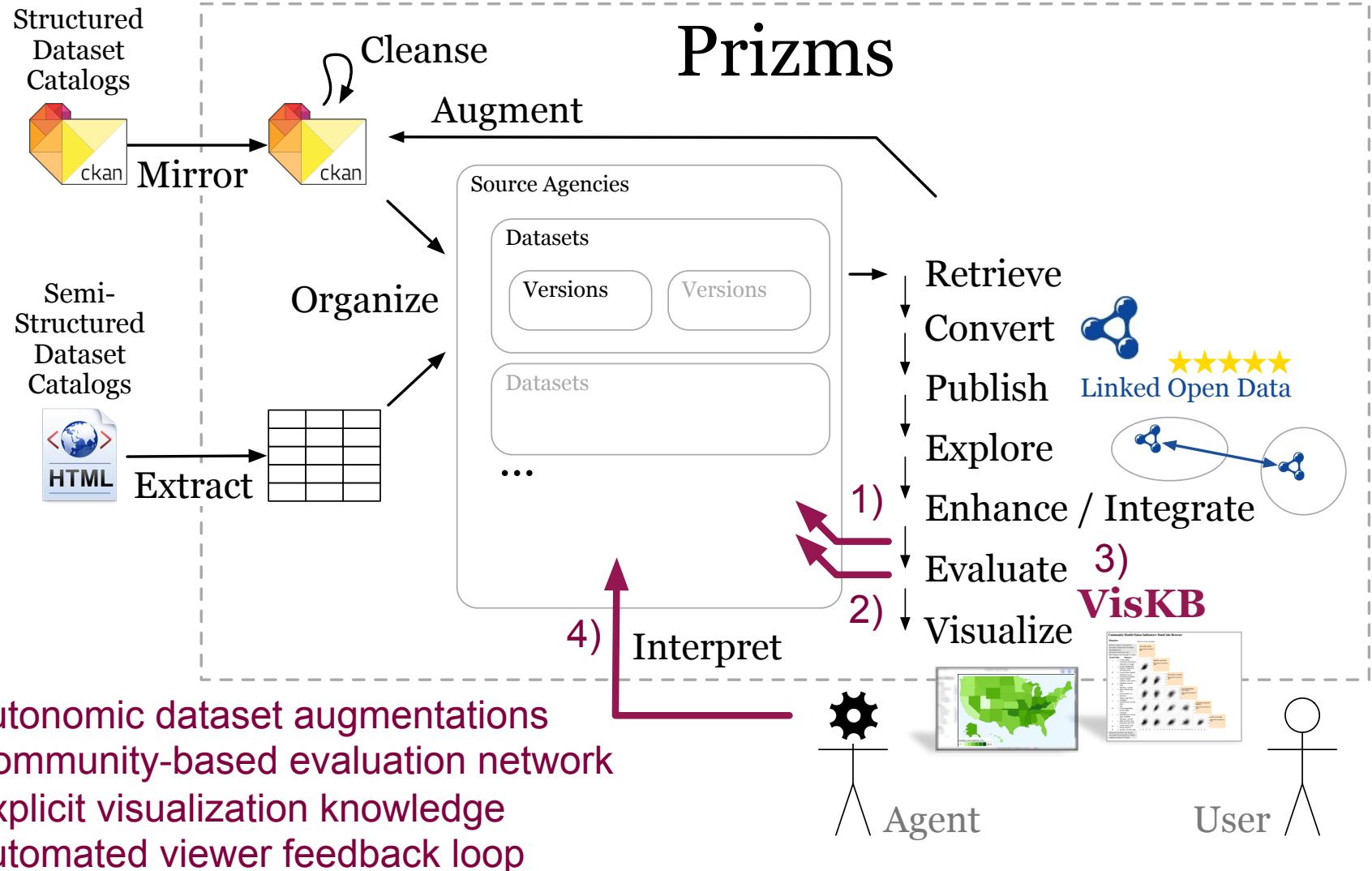
Linked Data:

```
curl -H "Accept: application/rdf+xml" -L http://purl.org/twc/health/source/hub-healthdata-gov/dataset/hospital-compare
```



Overview



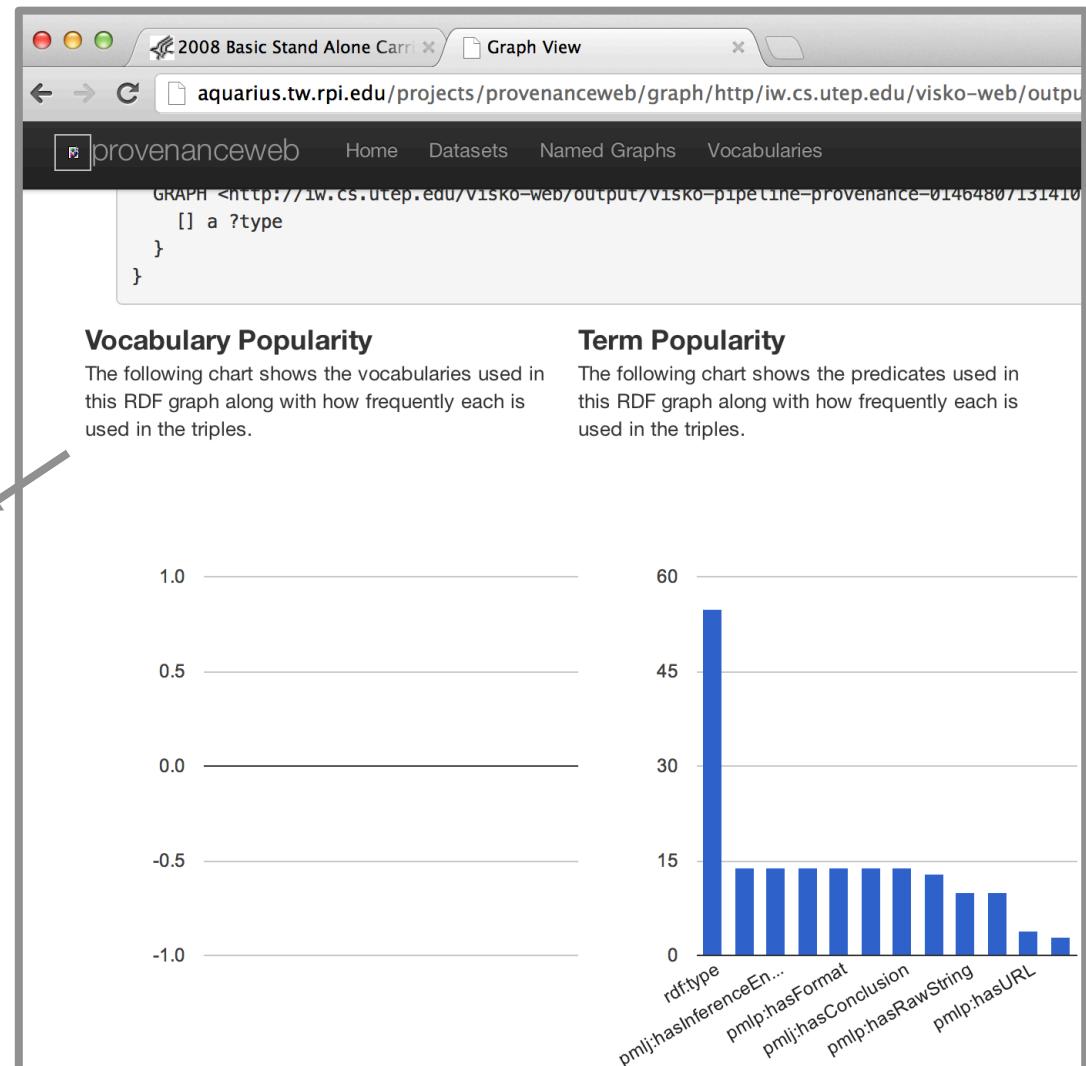


- 1) Autonomic dataset augmentations
- 2) Community-based evaluation network
- 3) Explicit visualization knowledge
- 4) Automated viewer feedback loop



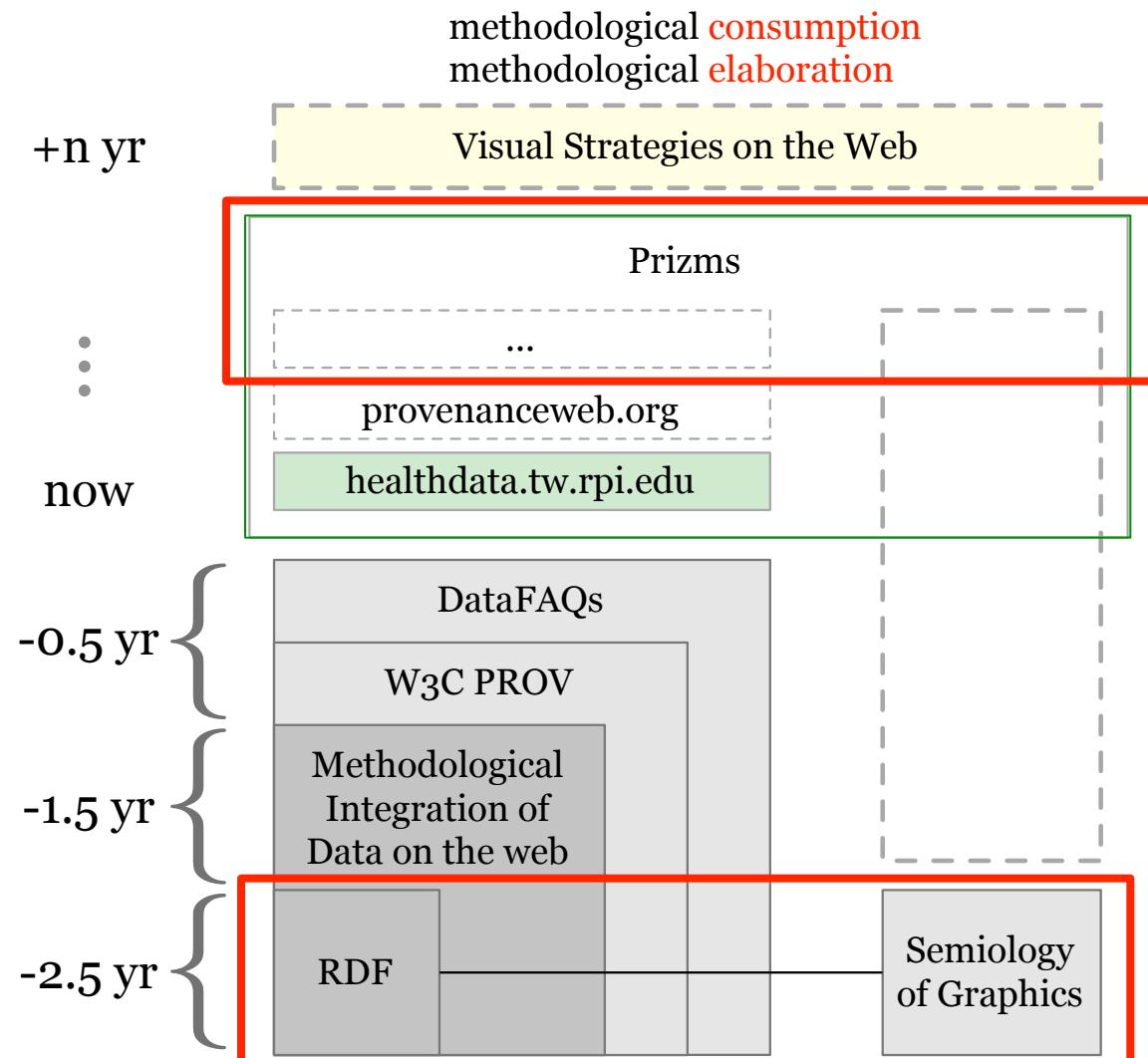
Applying Prizms

...
<http://ieeveis.tw.rpi.edu>
<http://provenanceweb.org> ↗
<http://healthdata.tw.rpi.edu>
<http://logd.tw.rpi.edu>





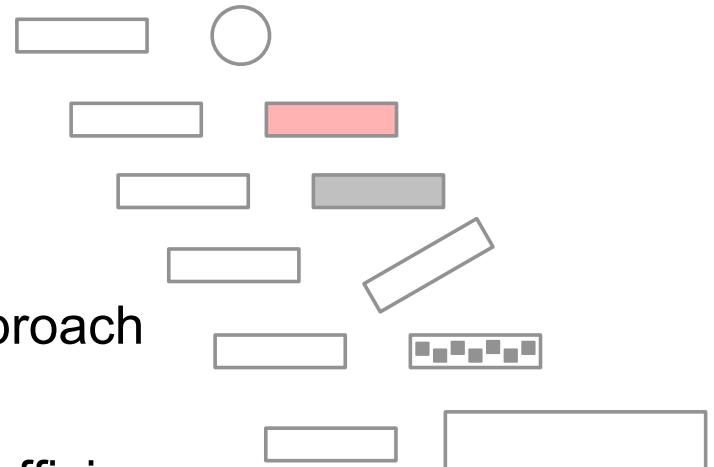
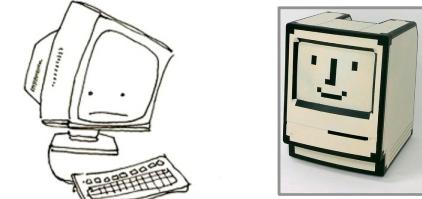
Review





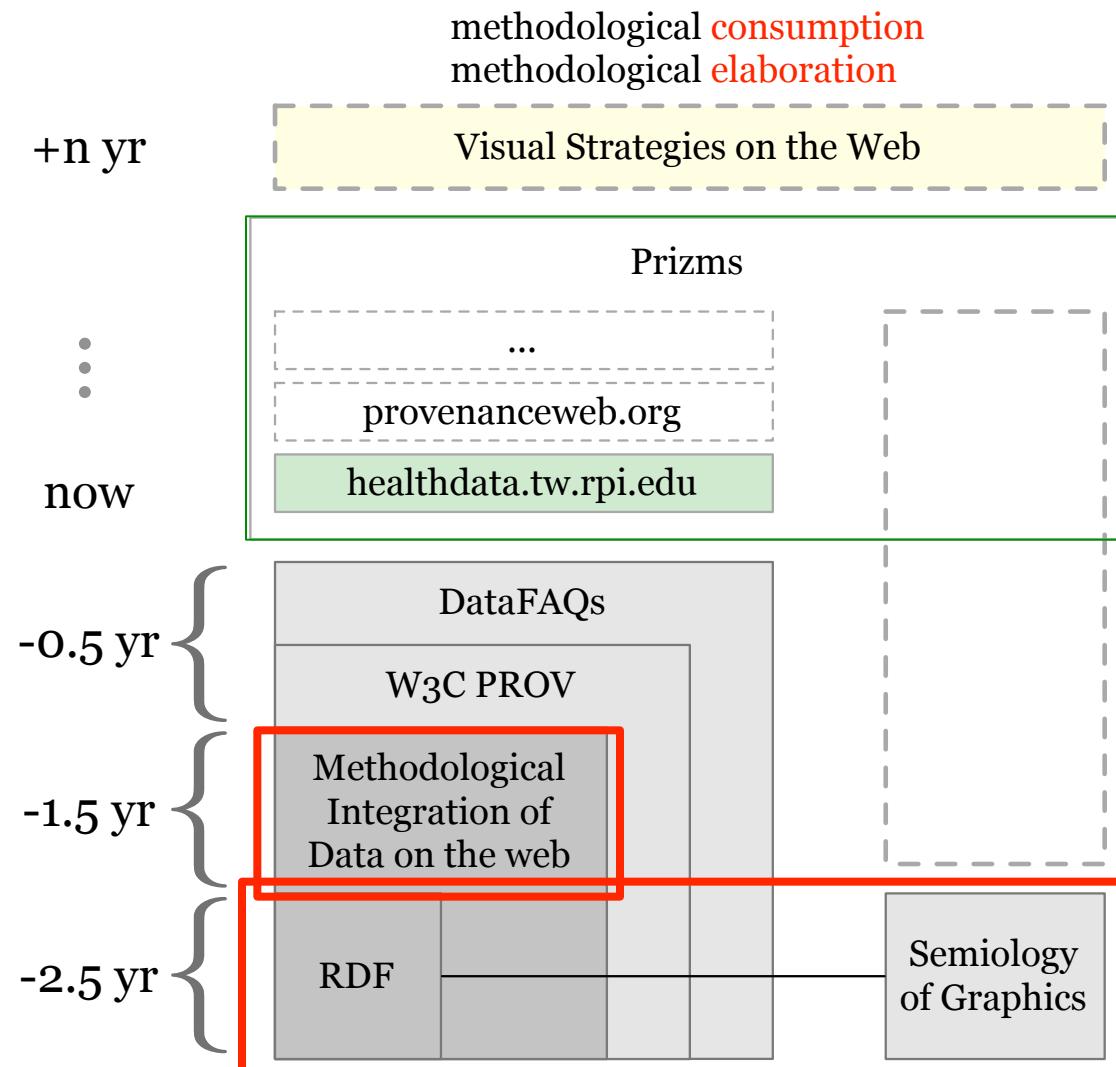
Inspiration for Visual Strategies on the Web

- RDF: My Data Hero
 - What are the **things** you're talking about?
 - How do those things **relate**?
 - Can I **get more information** about a thing with JUST its name?
 - Can I **merge** your data with **any other data** in the same format?
- Bertin's Semiology of Graphics
 - Set of Information
 - Analysis of the Information
 - Domain vs. Graphical sign-systems
 - Imposition, implantation, perceptual approach
 - Stages in reading a graphic
 - Level of reading, instant of perception, efficiency





Review

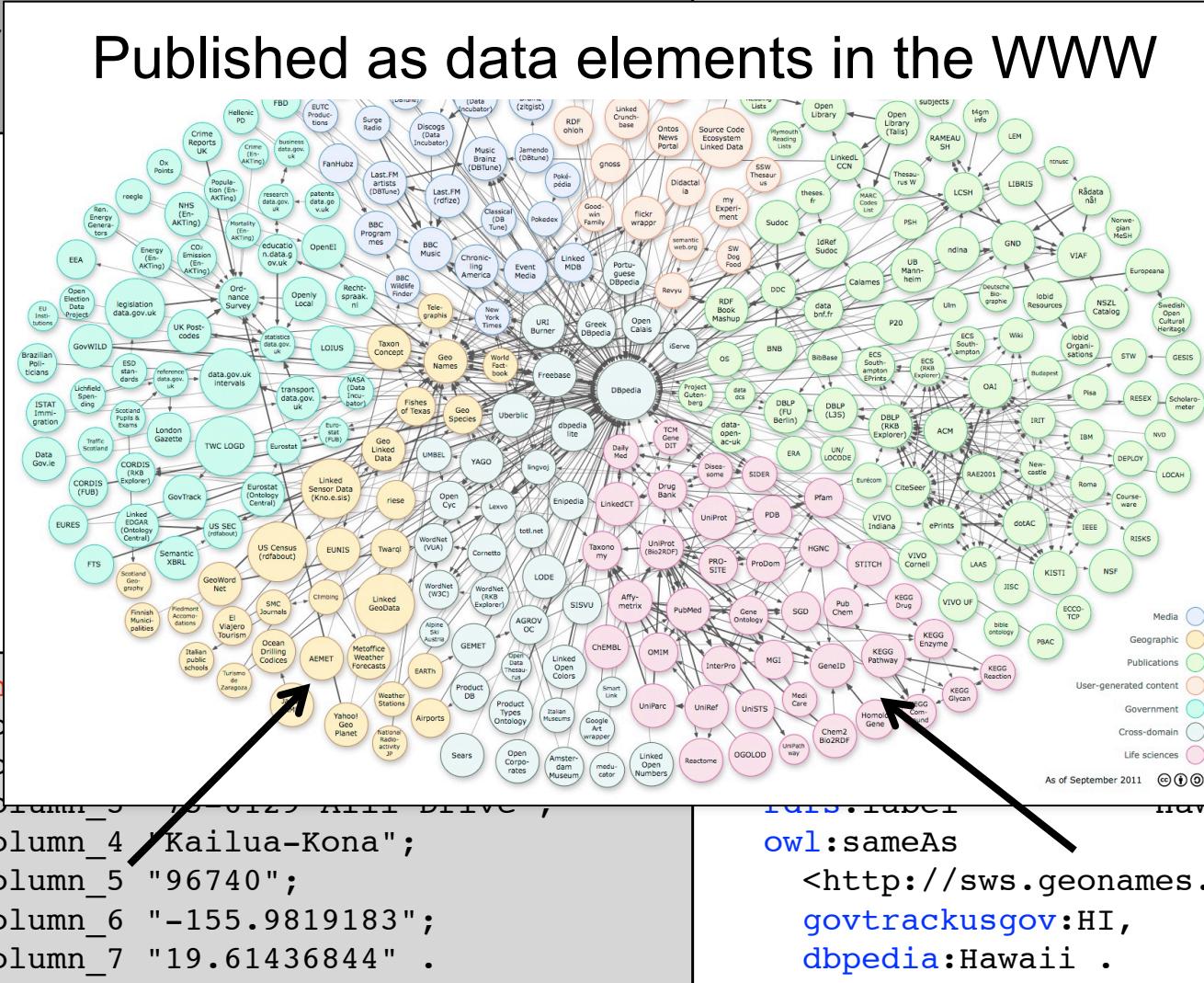




Good RDF and Bad RDF

"Hawaii",
"75-6129"
"96740",

Published as data elements in the WWW



```
ds4383:th
raw:co
raw:co
raw:co
raw:column_3 "75-6129 Alii Drive",
raw:column_4 "Kailua-Kona";
raw:column_5 "96740";
raw:column_6 "-155.9819183";
raw:column_7 "19.61436844" .
```

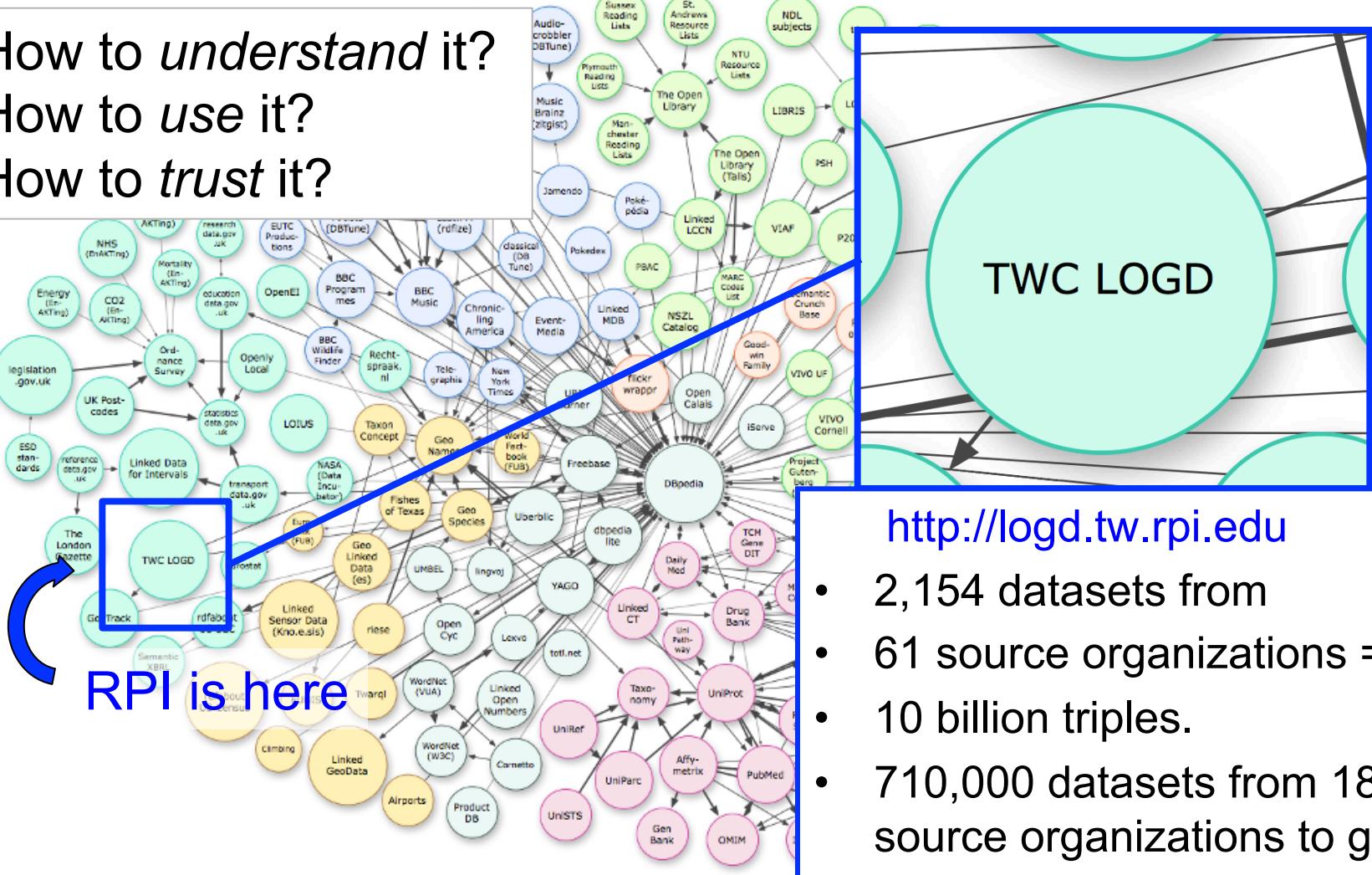
owl:sameAs

<<http://sws.geonames.org/5855797/>>,
govtrackusgov:HI,
dbpedia:Hawaii.



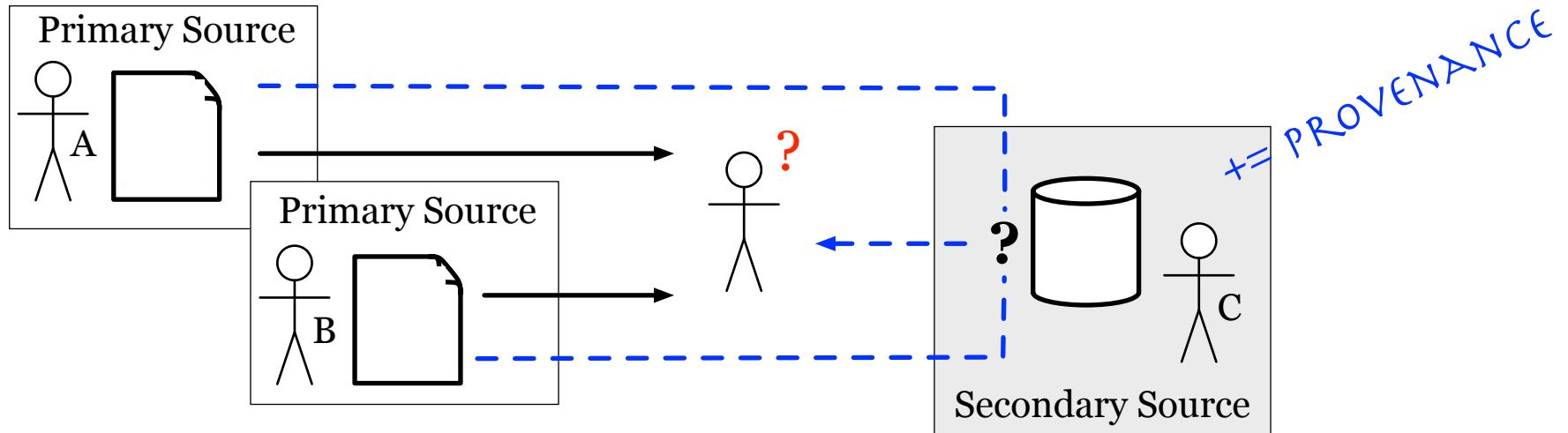
LOGD: Linking Open Government Data

- How to *understand* it?
- How to *use* it?
- How to *trust* it?





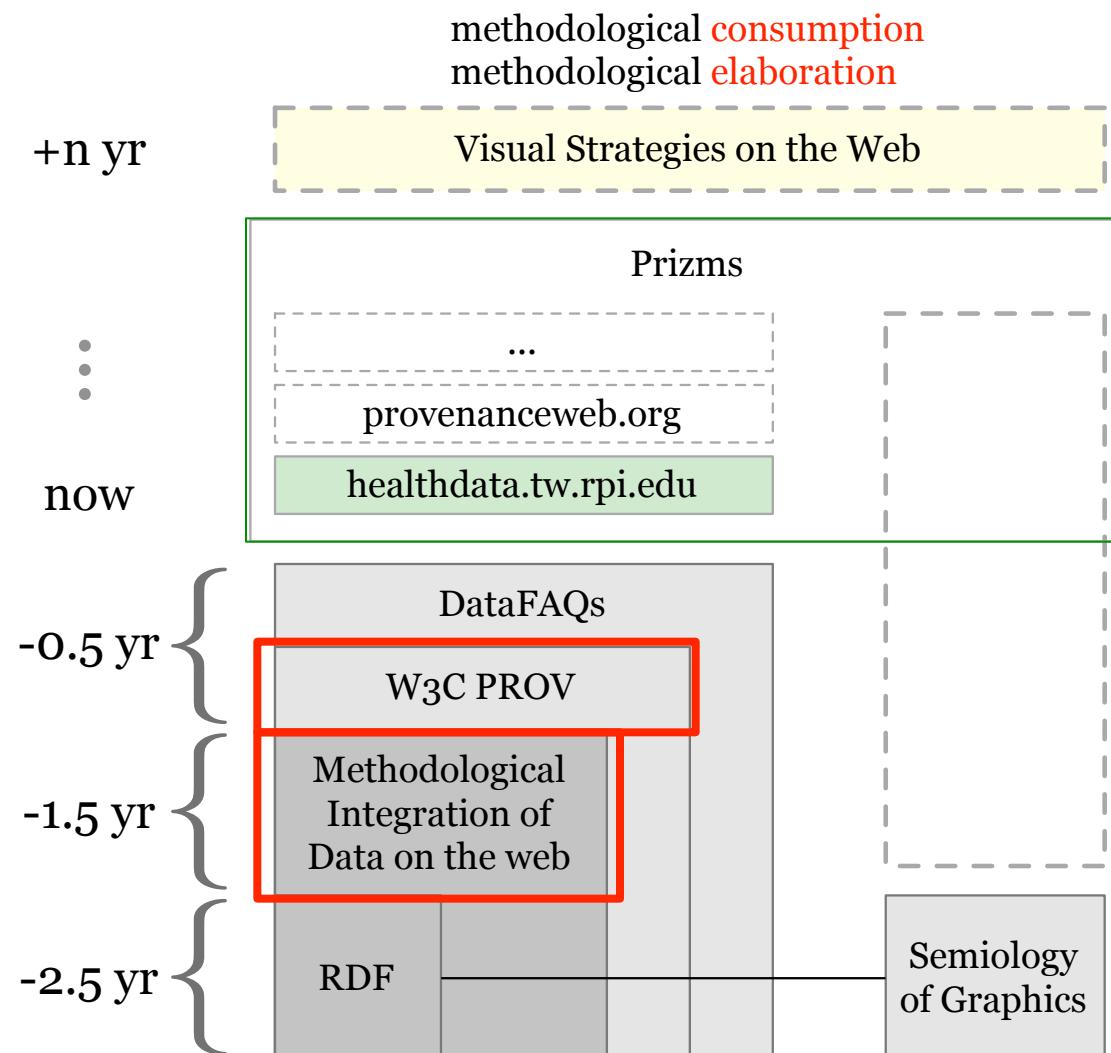
The Integrator's Dilemma



- | | |
|---|---|
| <ul style="list-style-type: none">+ “Original”+ Authoritative+ Domain Expertise | <ul style="list-style-type: none">- “Transformed”- Non-Authoritative- No Domain Expertise |
| <hr style="width: 100px; margin: 10px auto;"/> | |
| <ul style="list-style-type: none">- Not uniform structure- Not linked | <ul style="list-style-type: none">+ Uniform structure+ Linked (RDF) |

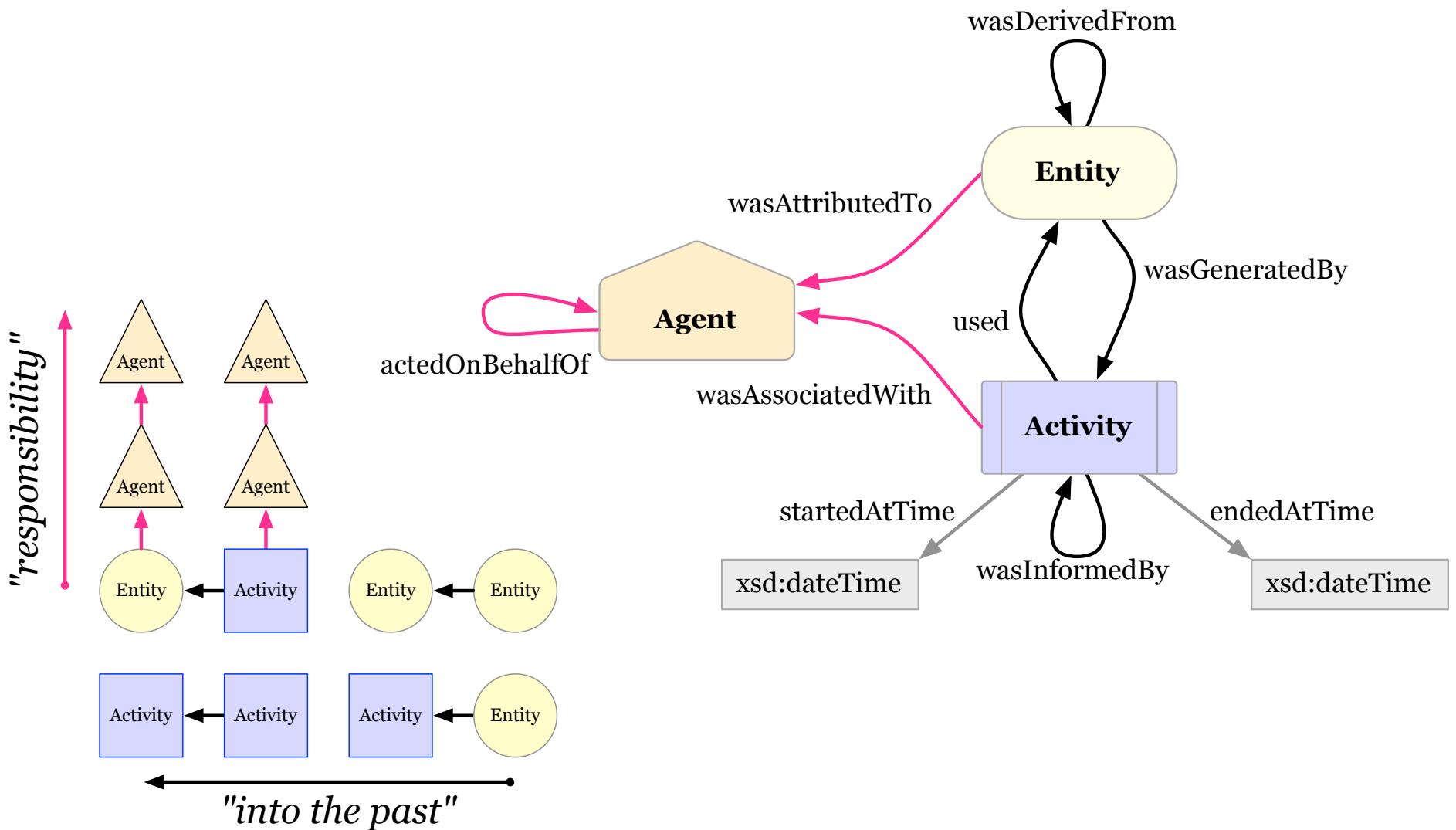


Update





W3C Provenance Interchange Model





W3C Provenance Interchange Model

W3C Working Draft



PROV-O: The PROV Ontology

W3C Working Draft 24 July 2012

This version:

<http://www.w3.org/TR/2012/WD-prov-o-20120724/>

Latest published version:

<http://www.w3.org/TR/prov-o/>

Latest editor's draft:

<https://dvcs.w3.org/hg/prov/raw-file/default/ontology/Overview.html>

Previous version:

<http://www.w3.org/TR/2012/WD-prov-o-20120503/>

Editors:

[Timothy Lebo](#), Rensselaer Polytechnic Institute, USA

[Satya Sahoo](#), Case Western Reserve University, USA

[Deborah McGuinness](#), Rensselaer Polytechnic Institute, USA

Authors:

(In alphabetical order)

[Khalid Bolhajjama](#), University of Manchester, UK



Last Call Draft

Candidate Recommendation

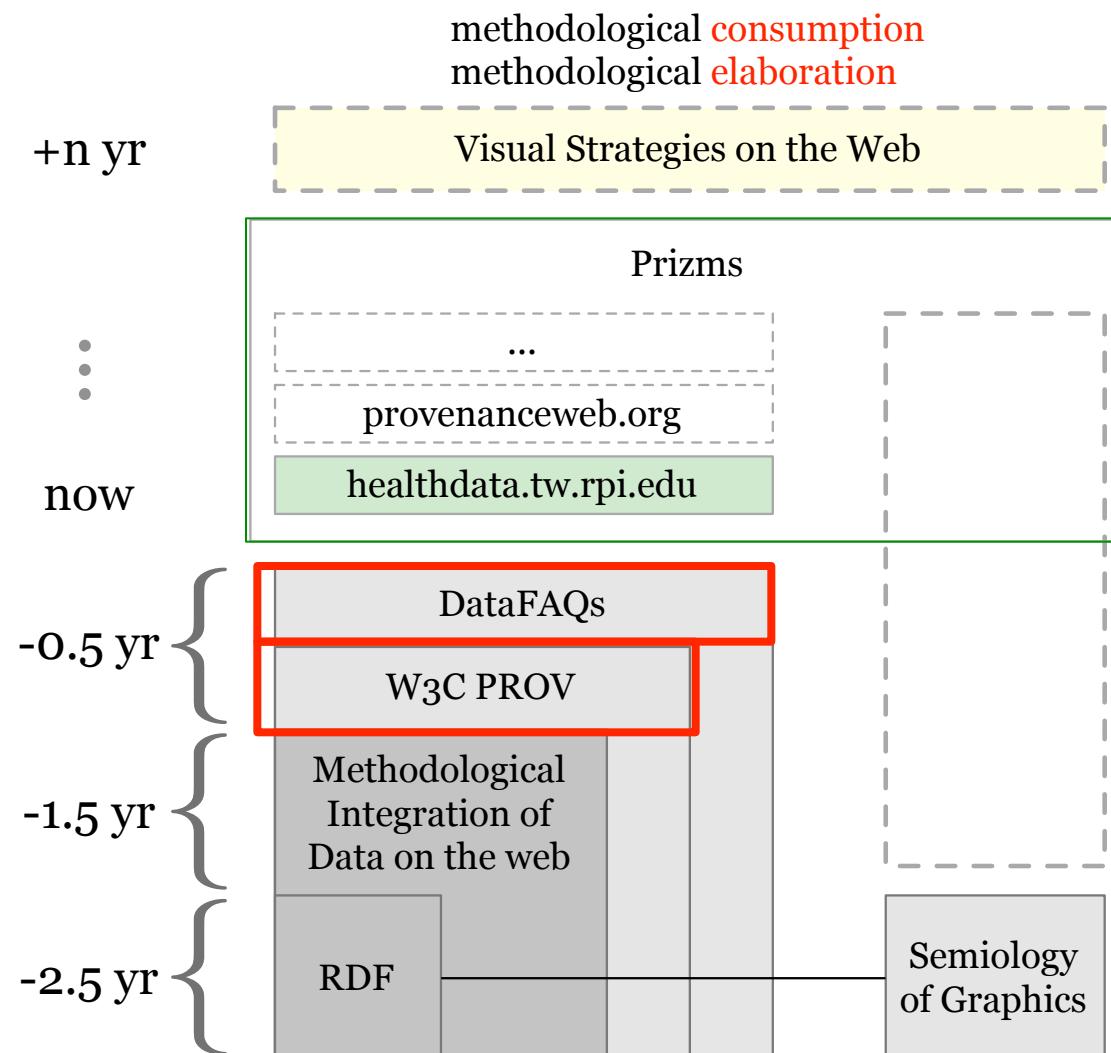
Proposed Recommendation

Final Recommendation

March
2013



Update

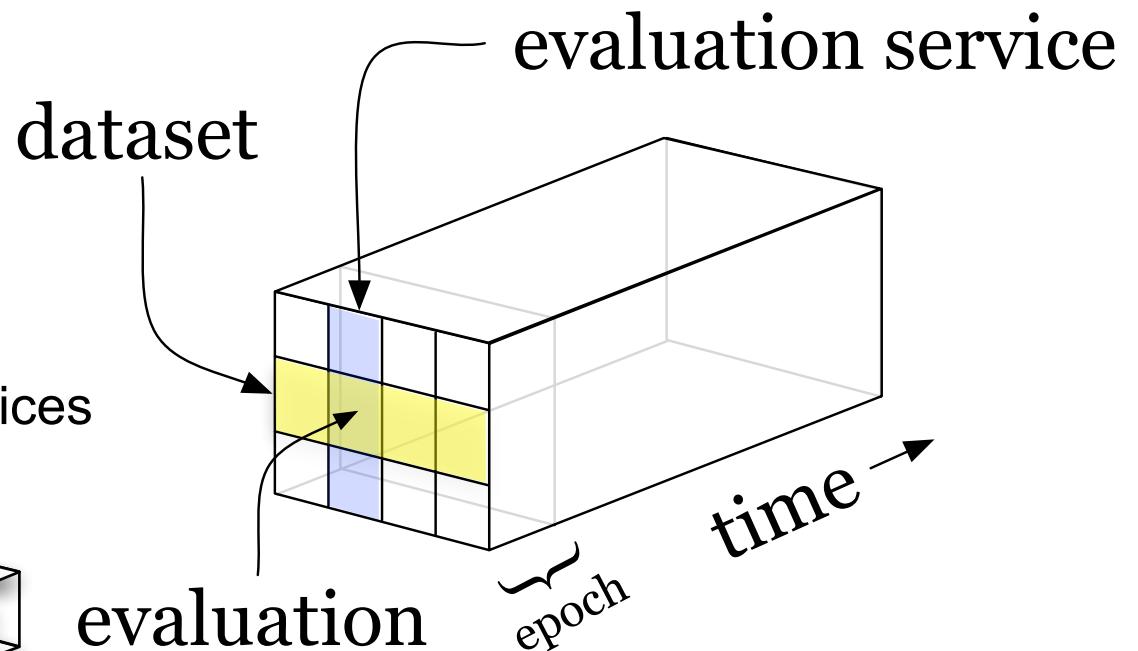
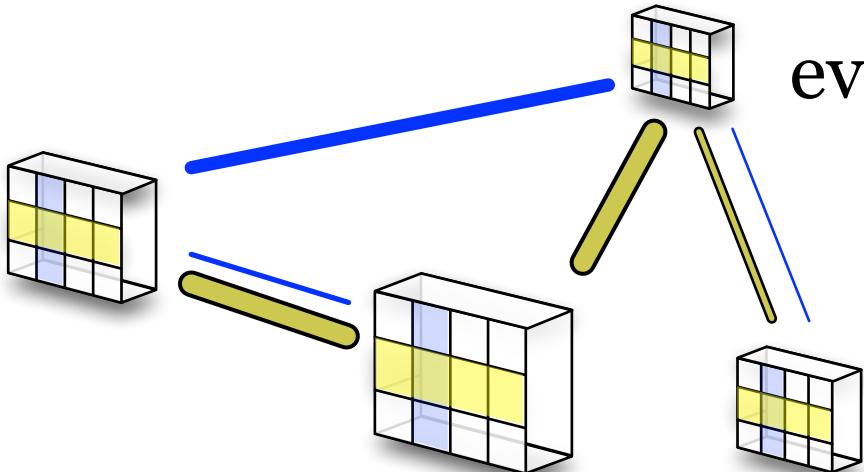




DataFAQs

LINKED DATA QUALITY REPORTS

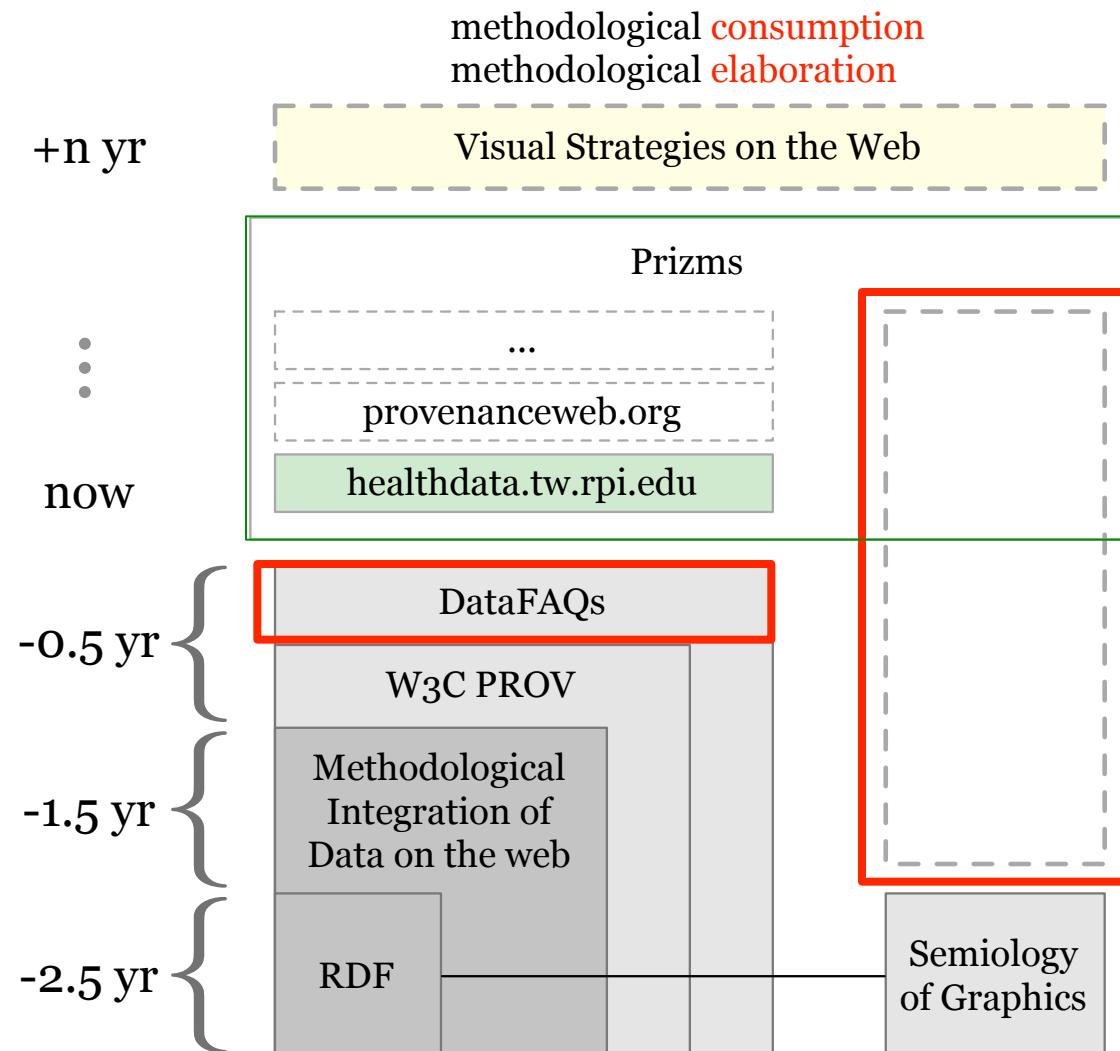
1. Choose Datasets
2. Choose Evaluation Services
3. Explore Evaluations



- Individual efforts benefit community
- Automated, quantitative feedback
- Recommendations from the aggregate
- Network effect

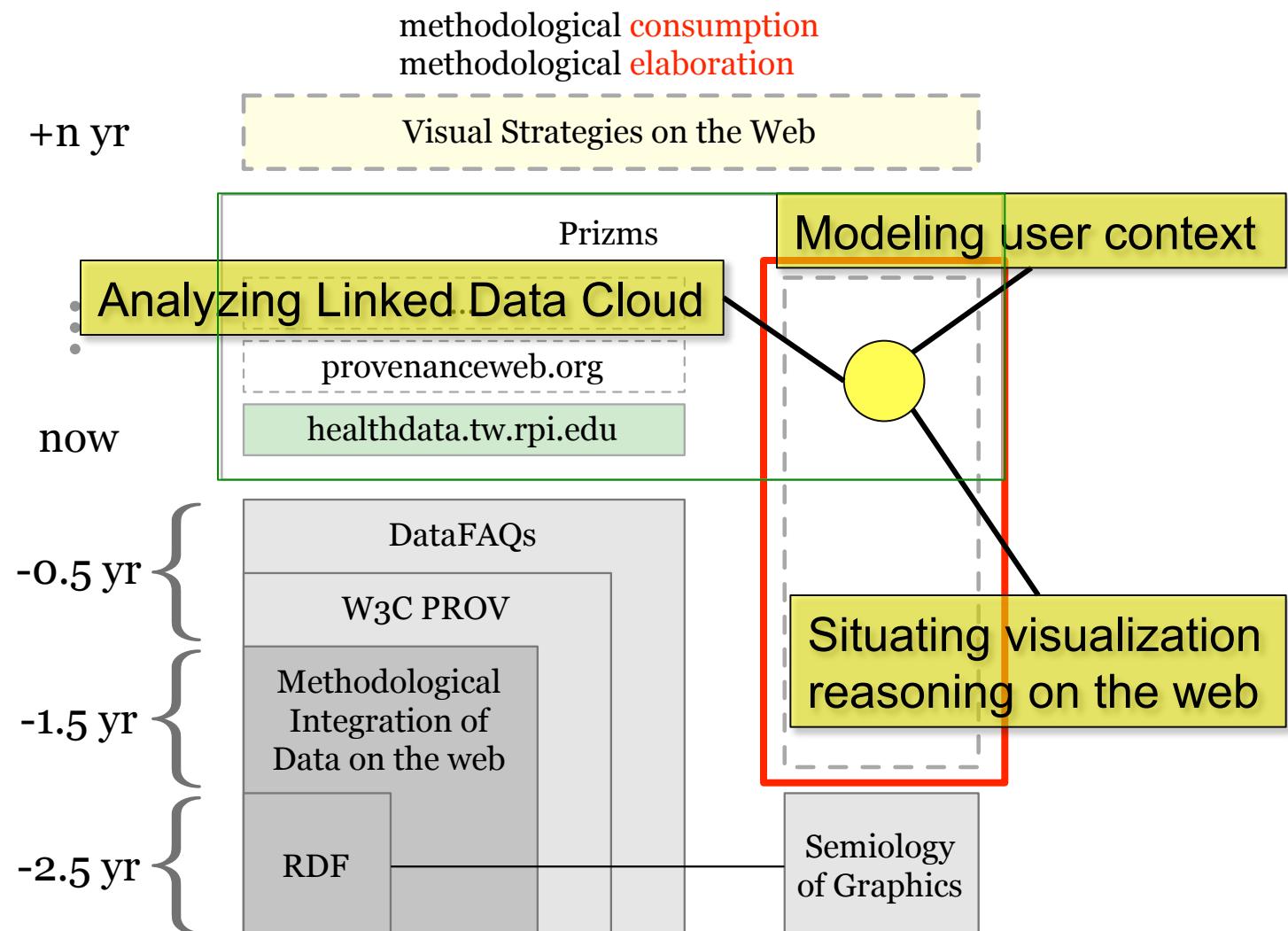


Future Work





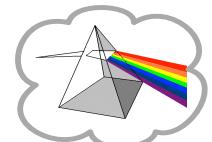
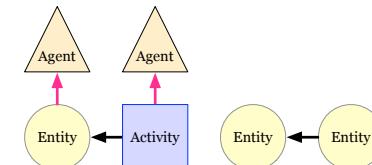
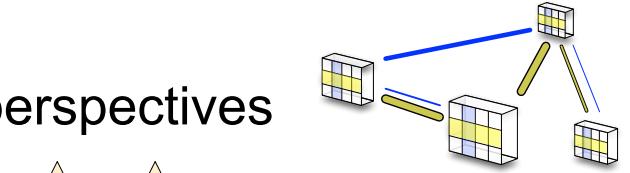
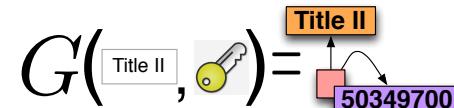
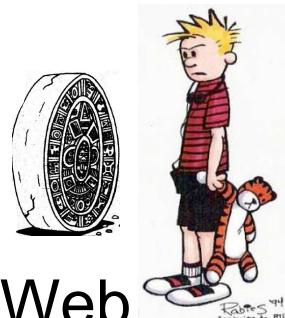
Future Work





Conclusion

- Motivation with Calvin
- Integrating data on the Web
 - Contextualizing entity names, relations
- DataFAQs quality evaluation framework
 - Encapsulating and gathering contextualized perspectives
- W3C Provenance Data Model
 - Relating to an historical context
- Prizms (and future work)
 - Visualization knowledge: pairing user context with data context





Thanks!



- Ping Wang
- Alvaro Graves
- Jim McCusker
- Dominic DiFranzo
- Yu Chen
- Josh Shinavier



Deborah McGuinness