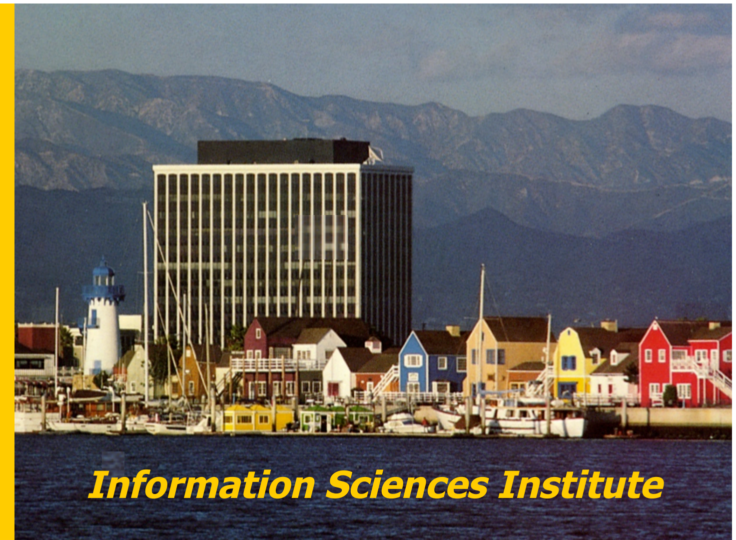
USC Viterbi
School of Engineering



Interactively Mapping Data Sources into the Semantic Web

Craig A. Knoblock, Pedro Szekely, Jose Luis Ambite, Shubham Gupta, Aman Goel, Maria Muslea, Kristina Lerman
University of Southern California

Parag Mallick
Stanford University

USC

Introduction

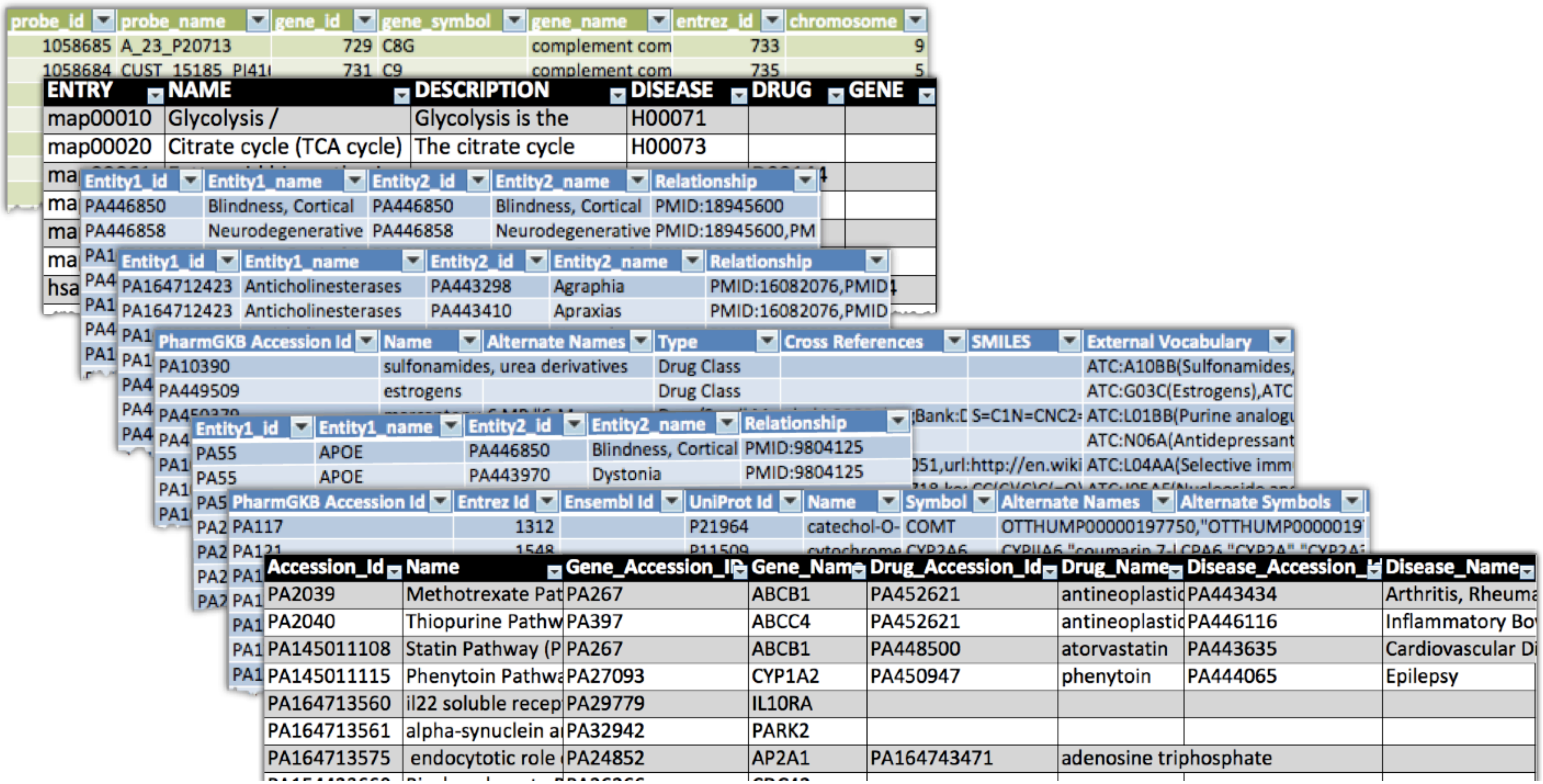
- Huge amount of data has been published to the Linked Open Data (> 28.5M triples)
- Remarkably little of this data has a detailed semantic description
- Challenge is how to allow users to easily publish data with respect to an ontology
- Can we automate the mapping to such an ontology?

USC Viterbi School of Engineering

Motivating Example

Information Sciences Institute

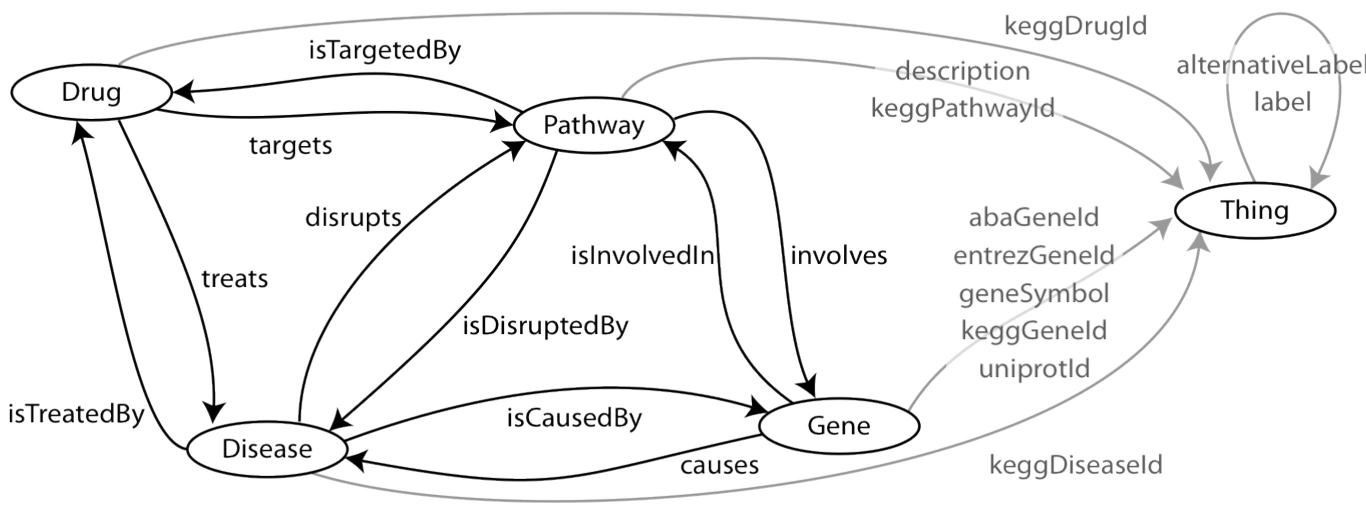
- **Integrate data from the Allen Brain Atlas (ABA) with standard neuroscience data sources** **[Bizer & Cyganiak, 2006]**
  - — UniProt, KEGG Pathway, PharmGKB, Linking Open Drug Data



USC

Motivating Example (cont.)

- **Challenge:**
  - — Create formal mappings from each of the sources into a shared ontology
  - — Use the mappings to create RDF



Accession_Id	Name	Gene_Accession_Id	Gene_Name	Drug_Accession_Id	Drug_Name	Disease_Accession_Id	Disease_Name
PA2039	Methotrexate Pat	PA267	ABCB1	PA452621	antineoplastic	PA443434	Arthritis, Rheuma
PA2040	Thiopurine Pathw	PA397	ABCC4	PA452621	antineoplastic	PA446116	Inflammatory Bo
PA145011108	Statin Pathway (P	PA267	ABCB1	PA448500	atorvastatin	PA443635	Cardiovascular Di
PA145011115	Phenytoin Pathwa	PA27093	CYP1A2	PA450947	phenytoin	PA444065	Epilepsy
PA164713560	il22 soluble recep	PA29779	IL10RA				
PA164713561	alpha-synuclein a	PA32942	PARK2				
PA164713575	endocytotic role	PA24852	AP2A1	PA164743471	adenosine triphosphate		

USC Viterbi School of Engineering

Motivating Example (cont.)

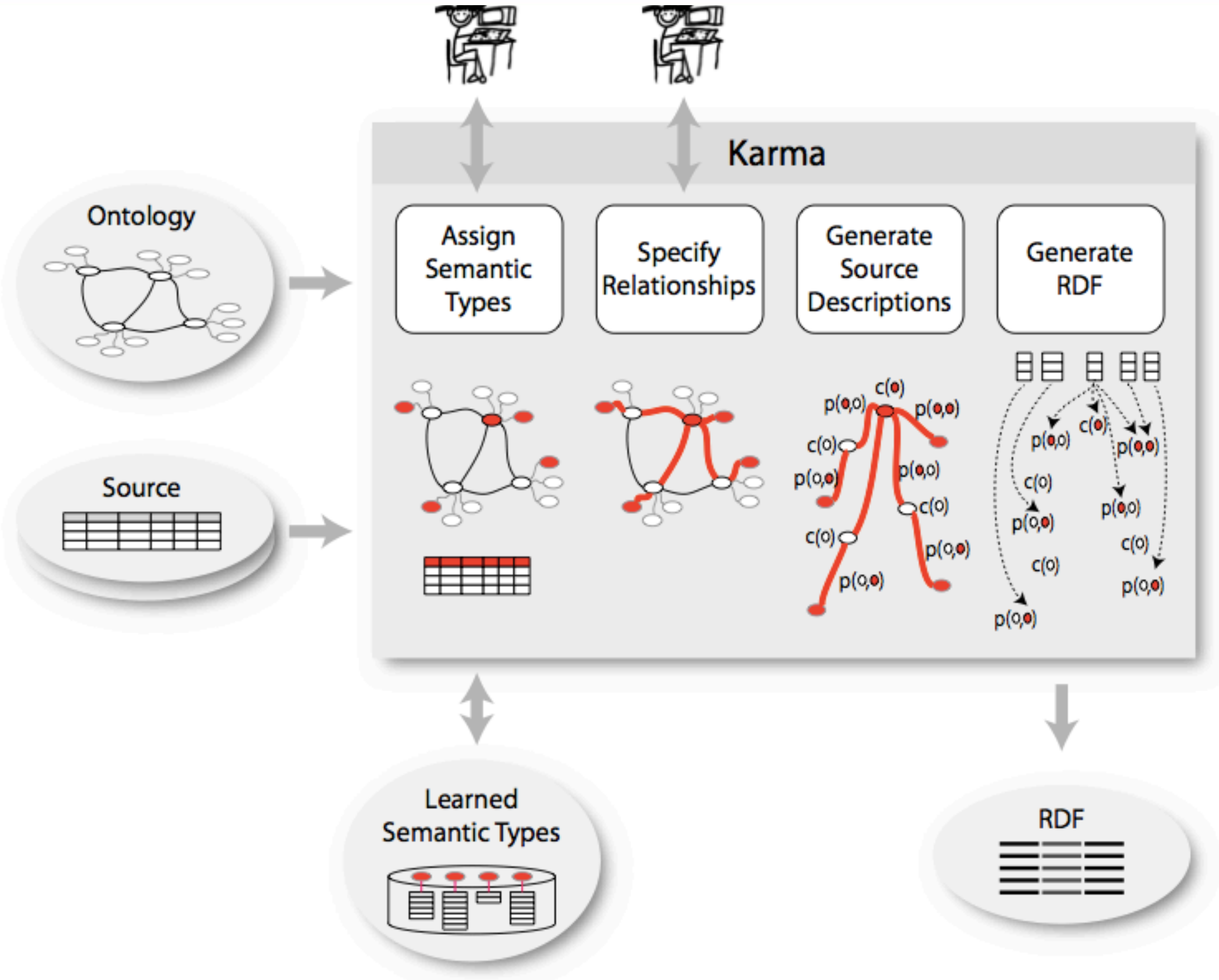
Information Sciences Institute

Accession_Id	Name	Gene_Accession_I	Gene_Nam	Drug_Accession_Id	Drug_Name	Disease_Accession_I	Disease_Name
PA2039	Methotrexate Pat	PA267	ABCB1	PA452621	antineoplastic	PA443434	Arthritis, Rheuma
PA2040	Thiopurine Pathw	PA397	ABCC4	PA452621	antineoplastic	PA446116	Inflammatory Bo
PA145011108	Statin Pathway (P	PA267	ABCB1	PA448500	atorvastatin	PA443635	Cardiovascular Di
PA145011115	Phenytoin Pathwa	PA27093	CYP1A2	PA450947	phenytoin	PA444065	Epilepsy
PA164713560	il22 soluble recep	PA29779	IL10RA				
PA164713561	alpha-synuclein a	PA32942	PARK2				
PA164713575	endocytotic role	PA24852	AP2A1	PA164743471	adenosine triphosphate		
1, 2	3	4, 7, 8	9	5, 10, 11	12	6, 13, 14	15

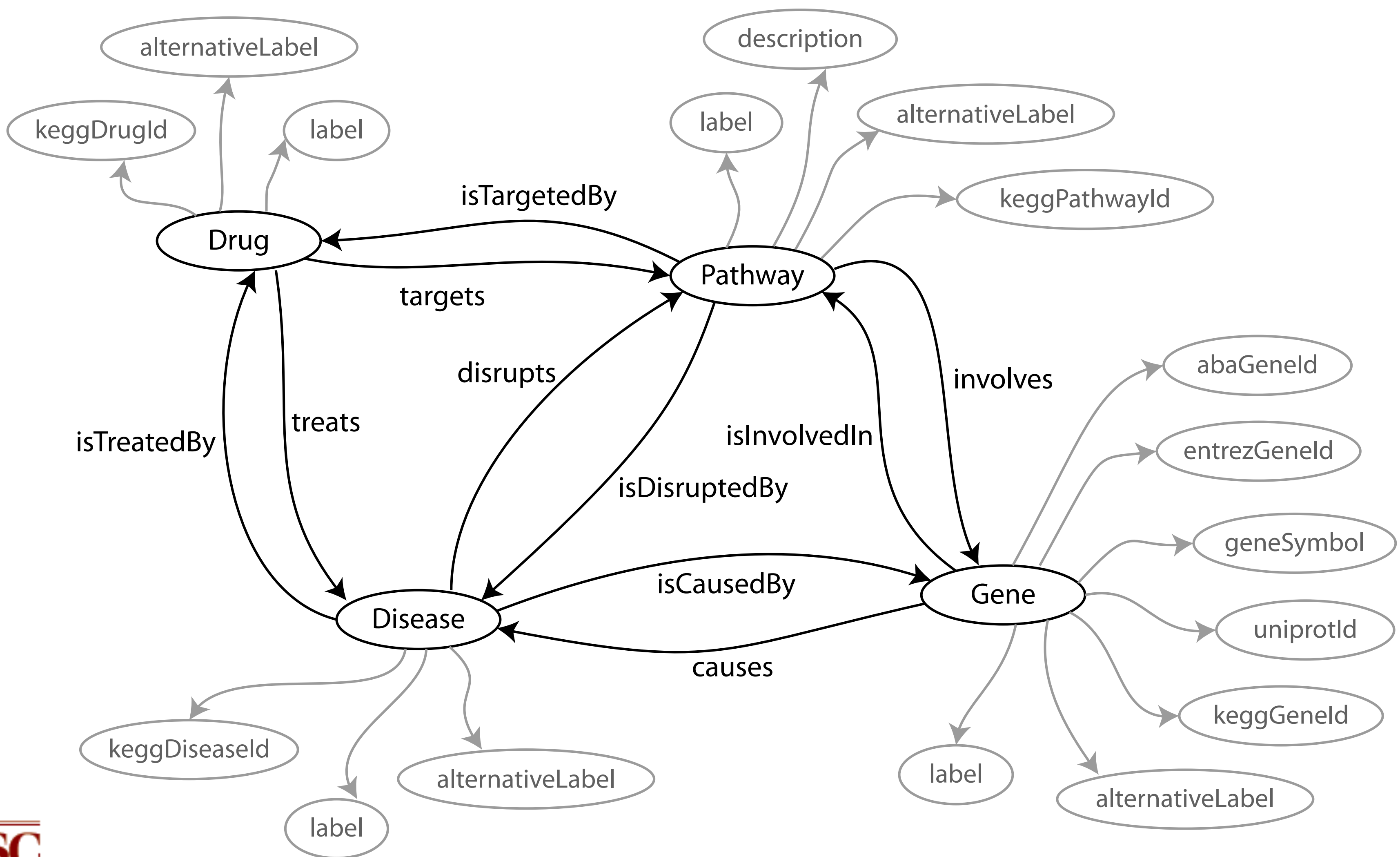
```
1.   :Pathway/Accession_Id/PA2039 a :Pathway;
2.      :Accession_Id "PA2039";
3.      :Label "Methotrexate Pathway";
4.      :Involves :Gene/Accession_Id/PA267;
5.      :IsTargetedBy :Drug/Accession_Id/PA452621 ;
6.      :IsDisruptedBy :Disease/Accession_Id/PA443434.
7.    :Gene/Accession_Id/PA267 a :Gene;
8.      :Accession_Id "PA267";
9.      :Label "ABCB1".
10.  :Drug/Accession_Id/PA452621 a :Drug;
11.     :Accession_Id "PA452621";
12.     :Label "antineoplastic agents".
13.  :Disease/Accession_Id/PA443434 a :Disease ;
14.     :Accession_Id "PA443434";
15.     :Label "Arthritis, Rheumatoid" .
```

USC

Overall Approach



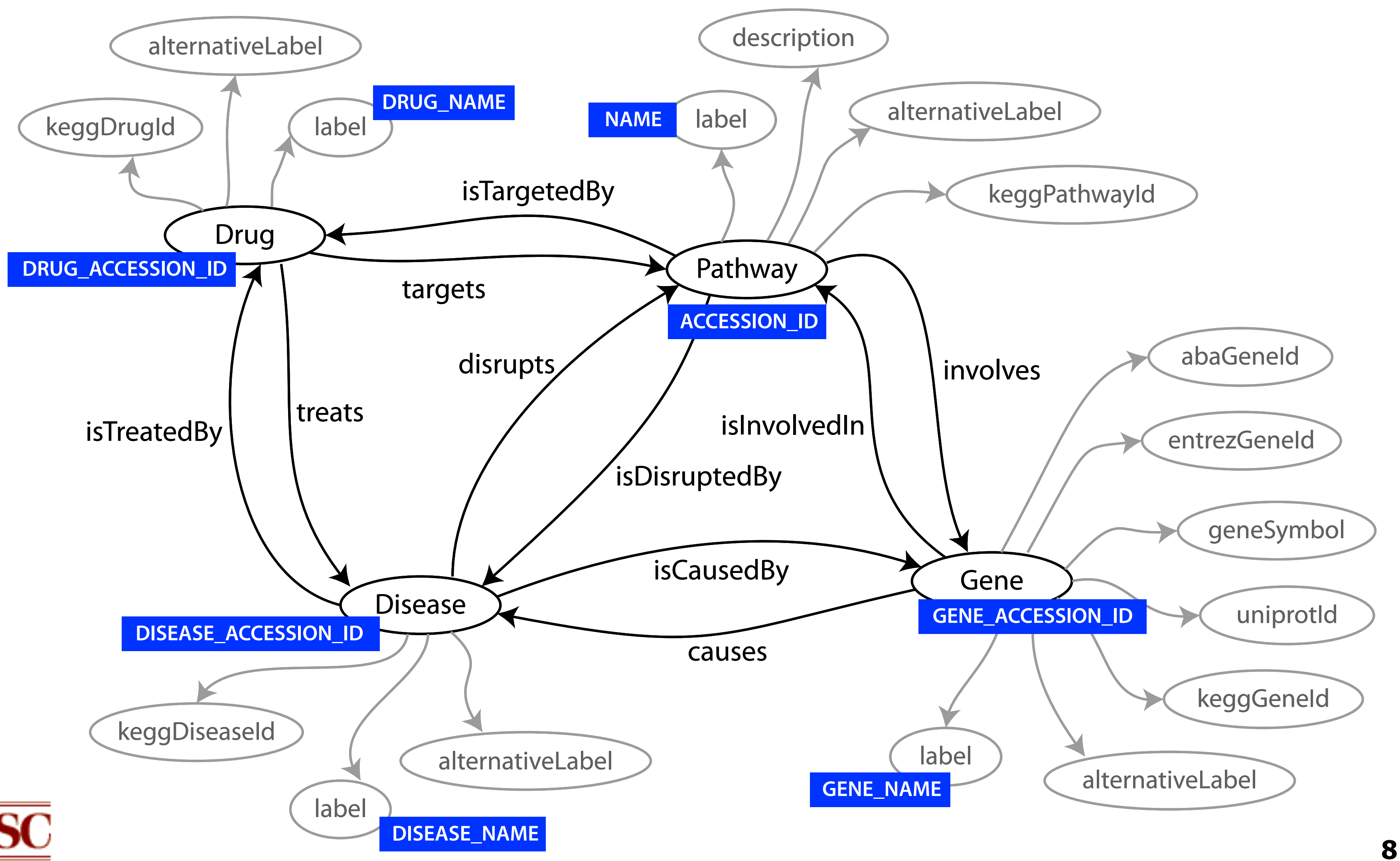
Building the Ontology Graph



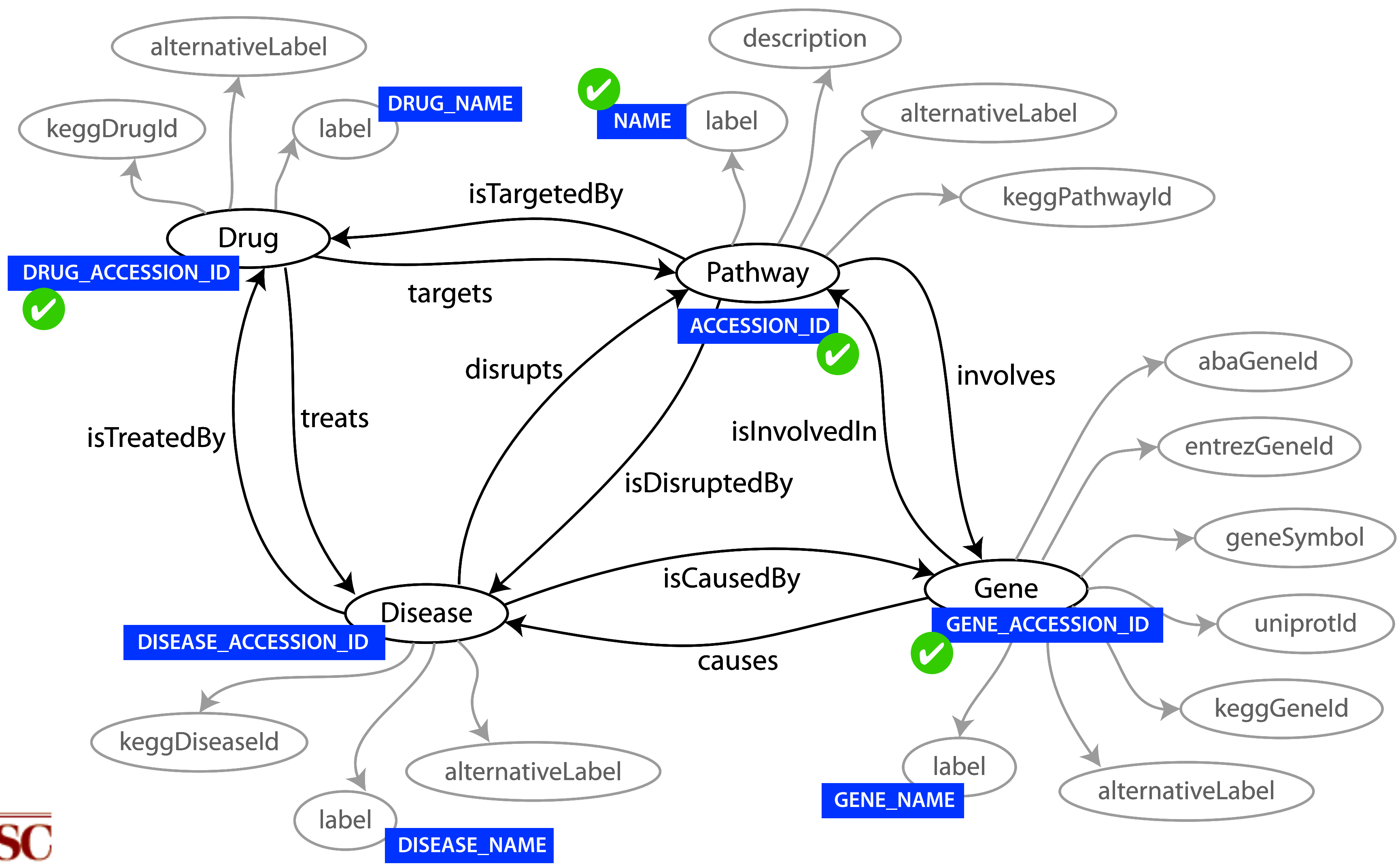
USC Viterbi School of Engineering

Building the Ontology Graph

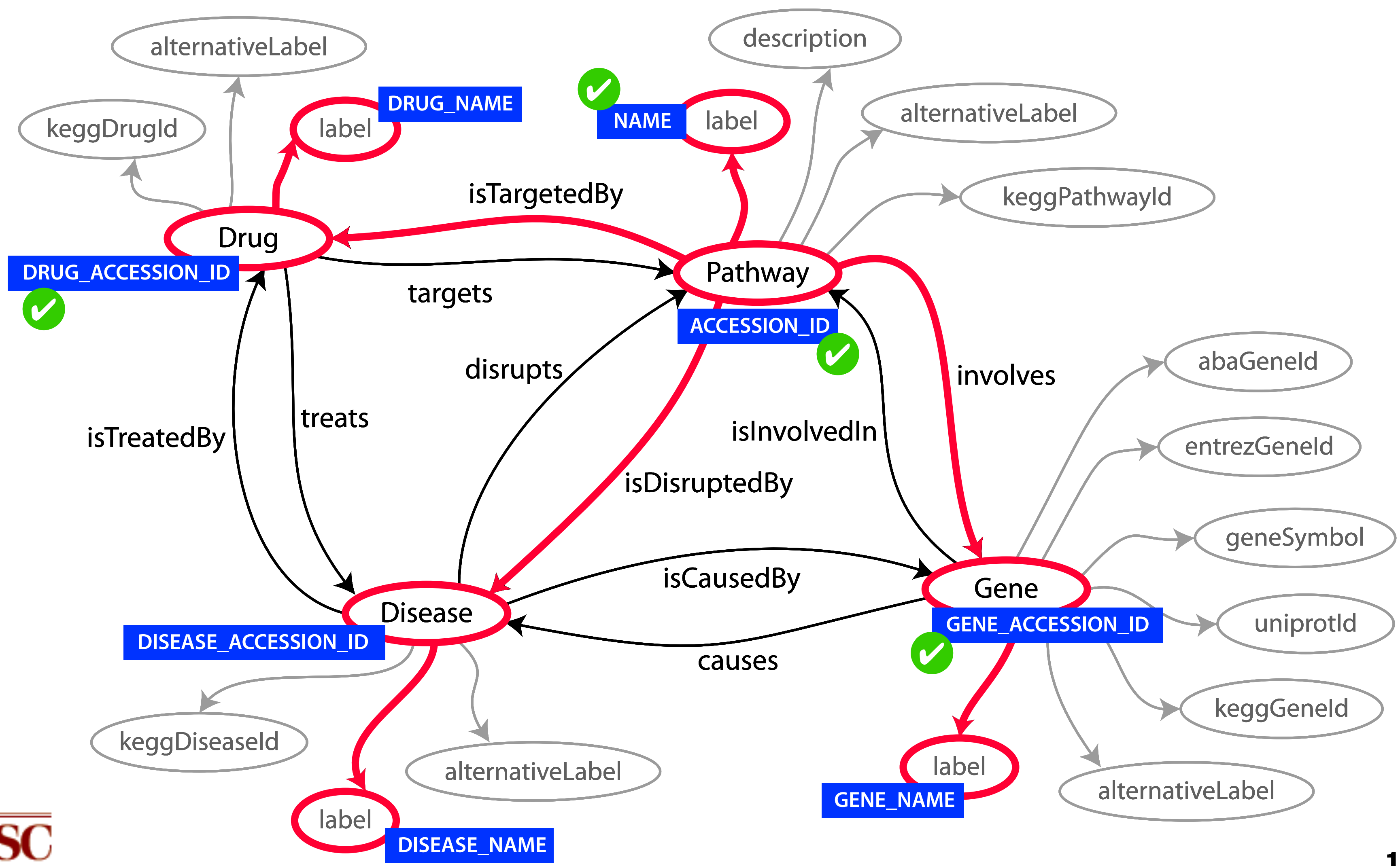
Information Sciences Institute



Building the Ontology Graph



Building the Ontology Graph



Inferring the Semantic Types

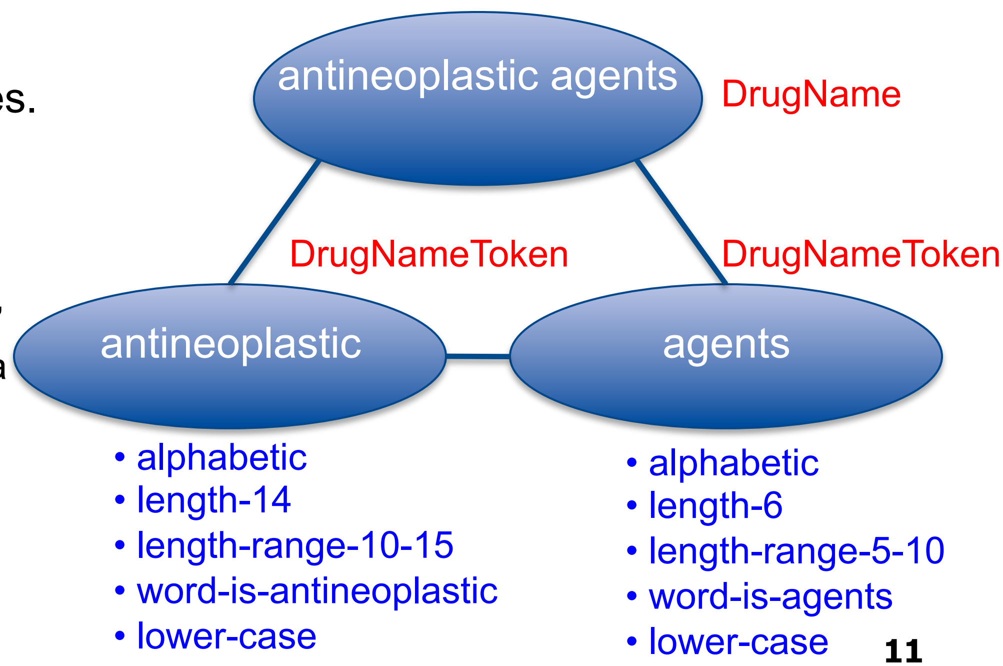
Problem: Given some columns of data, identify their semantic class.

Semantic classes:
- DrugName
- DiseaseID
- DiseaseName
- GeneName

?	?	?
antineoplastic agents	PA443434	Arthritis, Rheumatoid
antineoplastic agents	PA446116	Inflammatory Bowel Diseases
atorvastatin	PA443635	Cardiovascular Diseases
phenytoin	PA444065	Epilepsy
adenosine triphosphate		
	PA443560	Breast Neoplasms
budesonide		
ifosfamide		

Solution: Train a CRF model that learns the association between the features of the tokens and their labels.

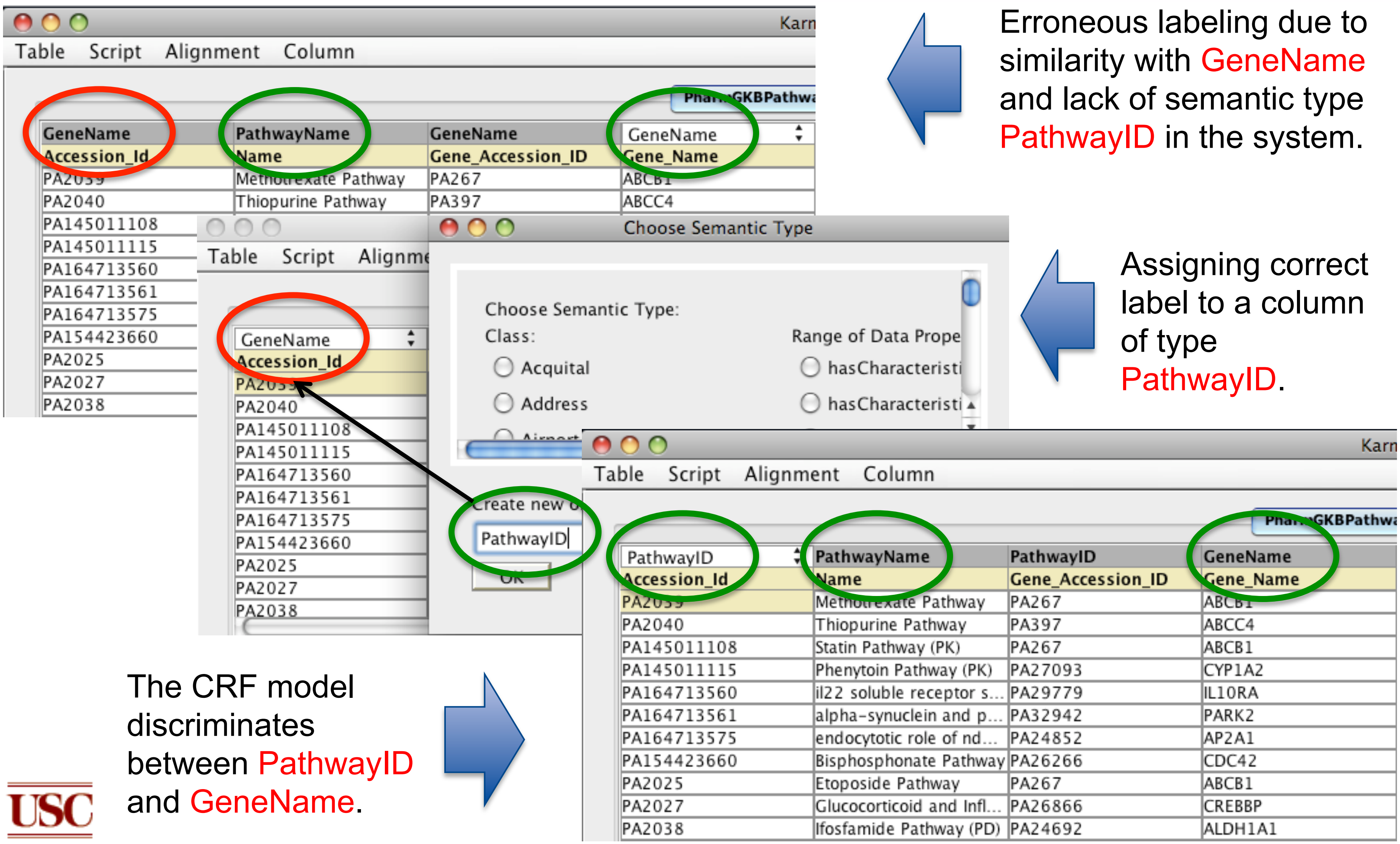
- Tokenize each field and extract their features.
- Create feature functions and learn their weights.
  - DrugNameToken is alphabetic
  - DrugNameToken is lowercase
  - DrugNameToken is the word "agents"
  - Field with label DrugName will have a token of label DrugNameToken
- Predict label for new column based on how many high-weight feature functions apply.



Interactively Refining the Semantic Types

Erroneous labeling due to similarity with GeneName and lack of semantic type PathwayID in the system.

Assigning correct label to a column of type PathwayID.

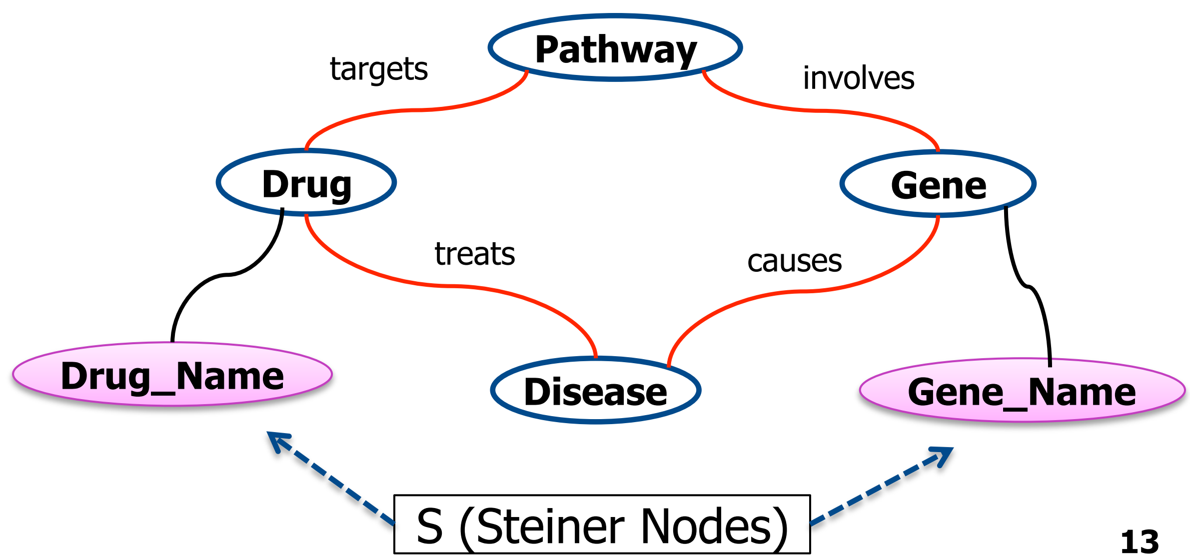


The CRF model discriminates between PathwayID and GeneName.

Inferring the Relationships

- **Apply a fast Steiner tree algorithm**
  - $G=(V,E)$, $S \subset V$, $c: E \rightarrow \Re$
  - Find a tree of G that spans S with minimal total cost
- **Approximation Alg. [Kou & Markowsky, 1981]**
  - Worst case time complexity: $O(|V|^2|S|)$
  - Approximation Ratio: less than 2
- **Example**

Drug_Name	Gene_Name
Antineoplastic	ABCB1
Antineoplastic	ABCC4
Atorvastatin	ABCB1

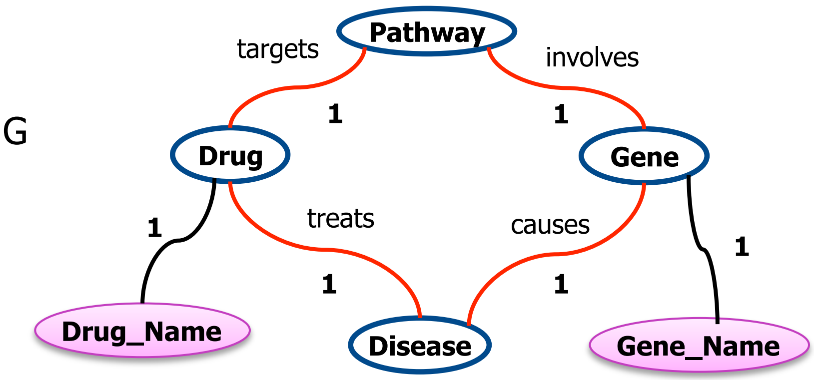


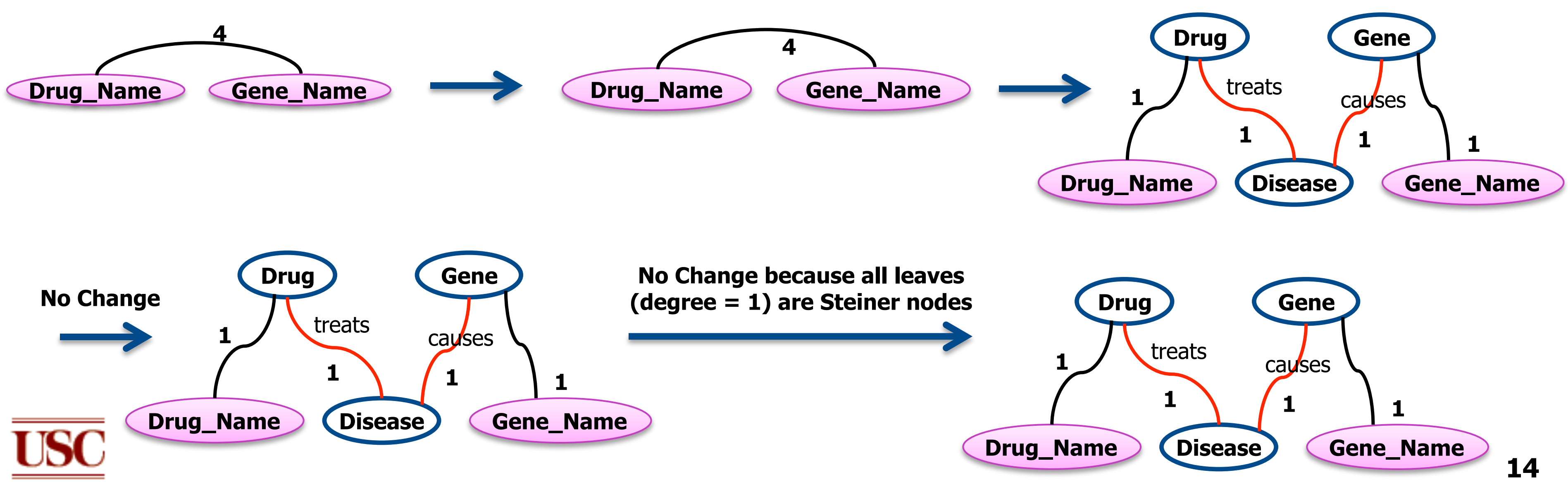
USC Viterbi School of Engineering

Steiner Tree algorithm (cont.)

Information Sciences Institute

- **Step1: construct the complete graph**
  - Nodes: Steiner Nodes
  - Links Weights: shortest path from each pair in original G
- **Step2: compute MST (minimal spanning tree)**
- **Step3: replace each <u>link</u> with the corresponding shortest <u>path</u> in original G**
- **Step4: compute MST again**
- **Step5: remove extra links until all <u>leaves</u> are Steiner nodes**



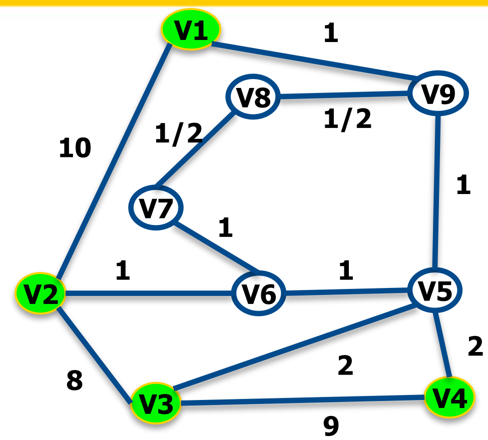


USC

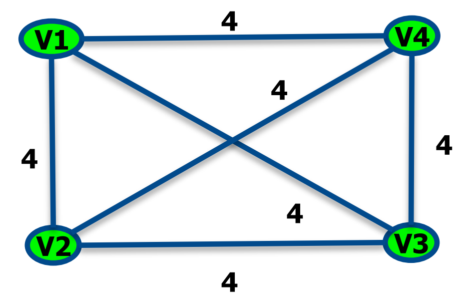
USC Viterbi School of Engineering

Steiner Tree Algorithm

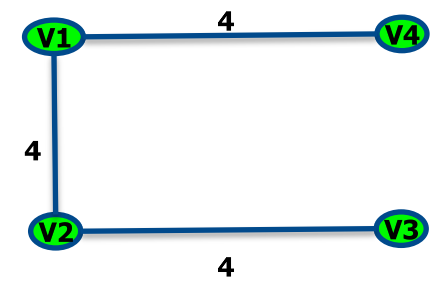
Information Sciences Institute



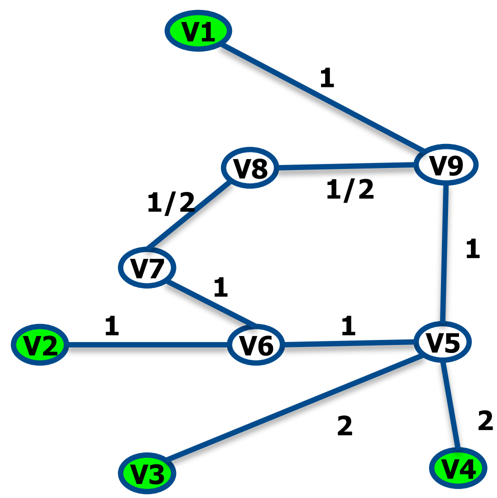
Steiner nodes: {V1, V2, V3, V4}



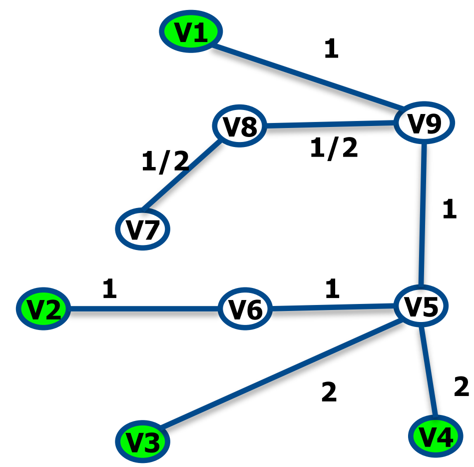
1. construct the complete graph (Nodes: Steiner Nodes, Links Weights: shortest path from each pair in original G)



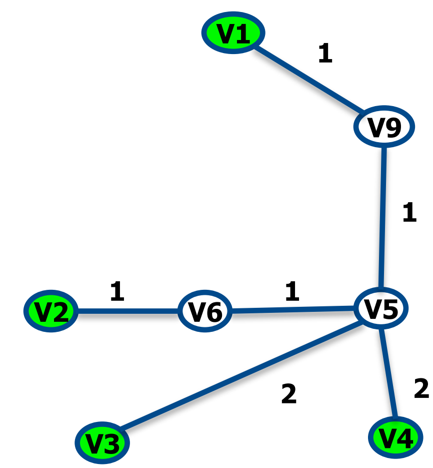
2. Compute MST



3. replace each link with the corresponding shortest path in original G



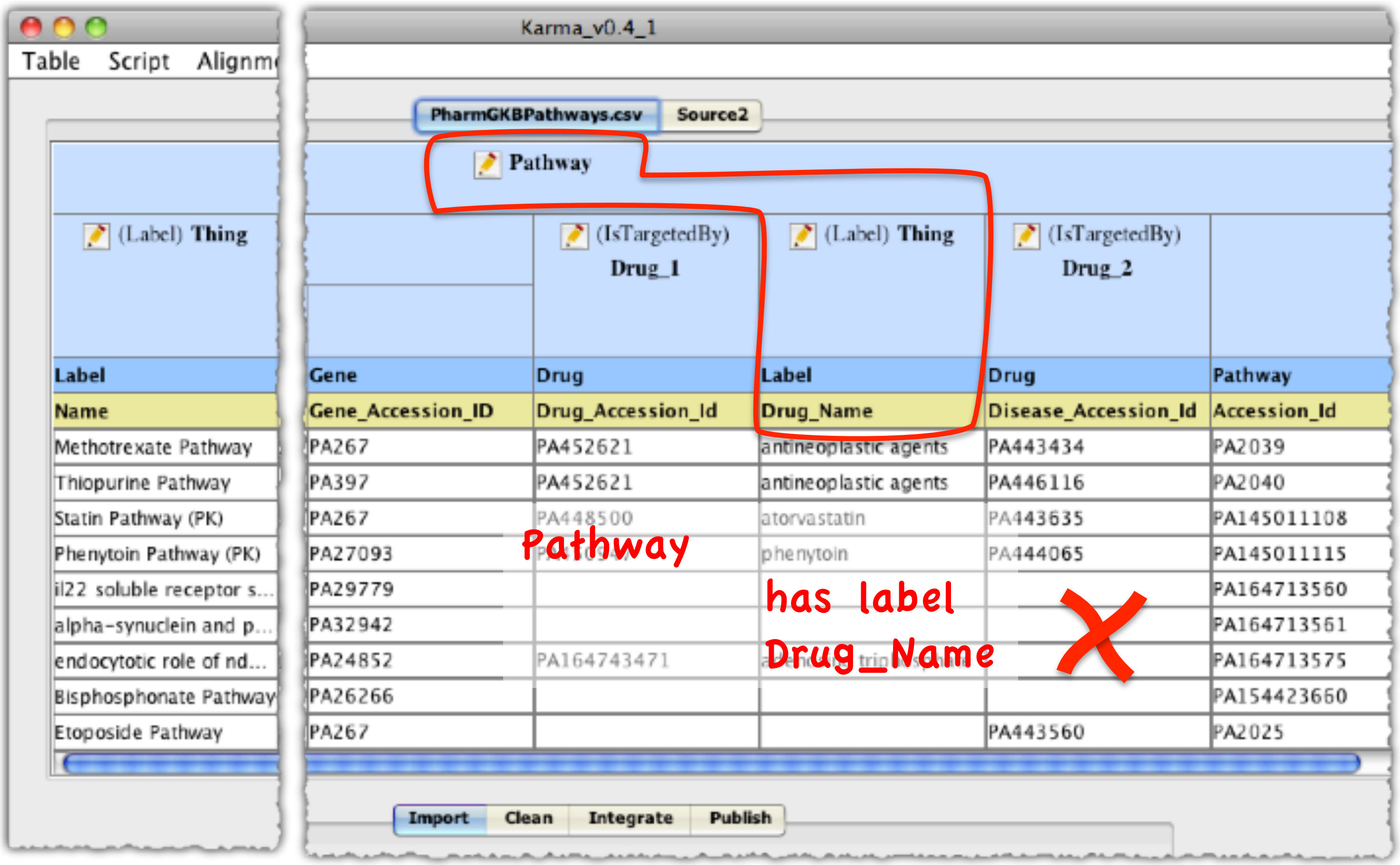
4. Compute MST



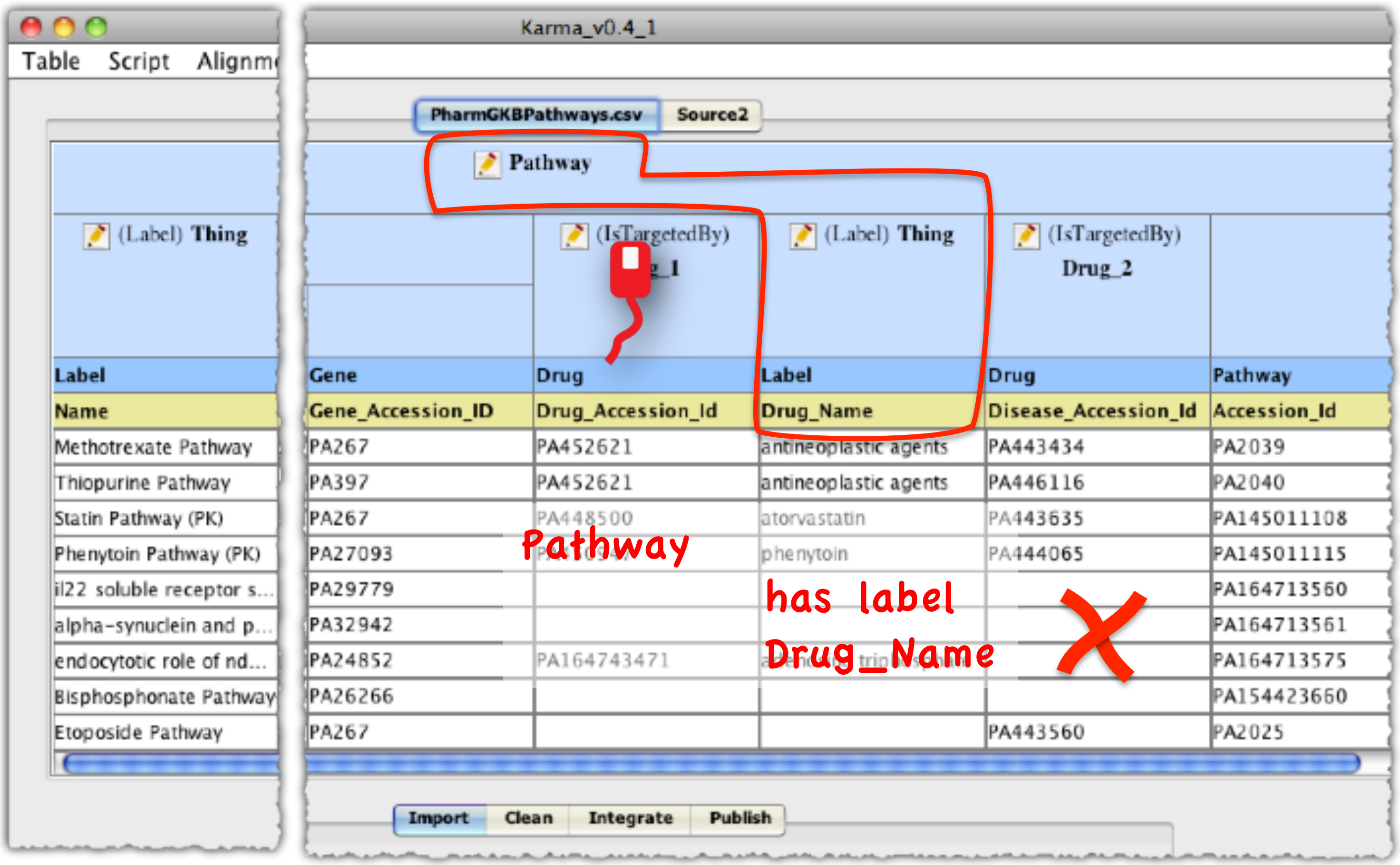
5. remove extra links until all leaves are Steiner nodes

USC

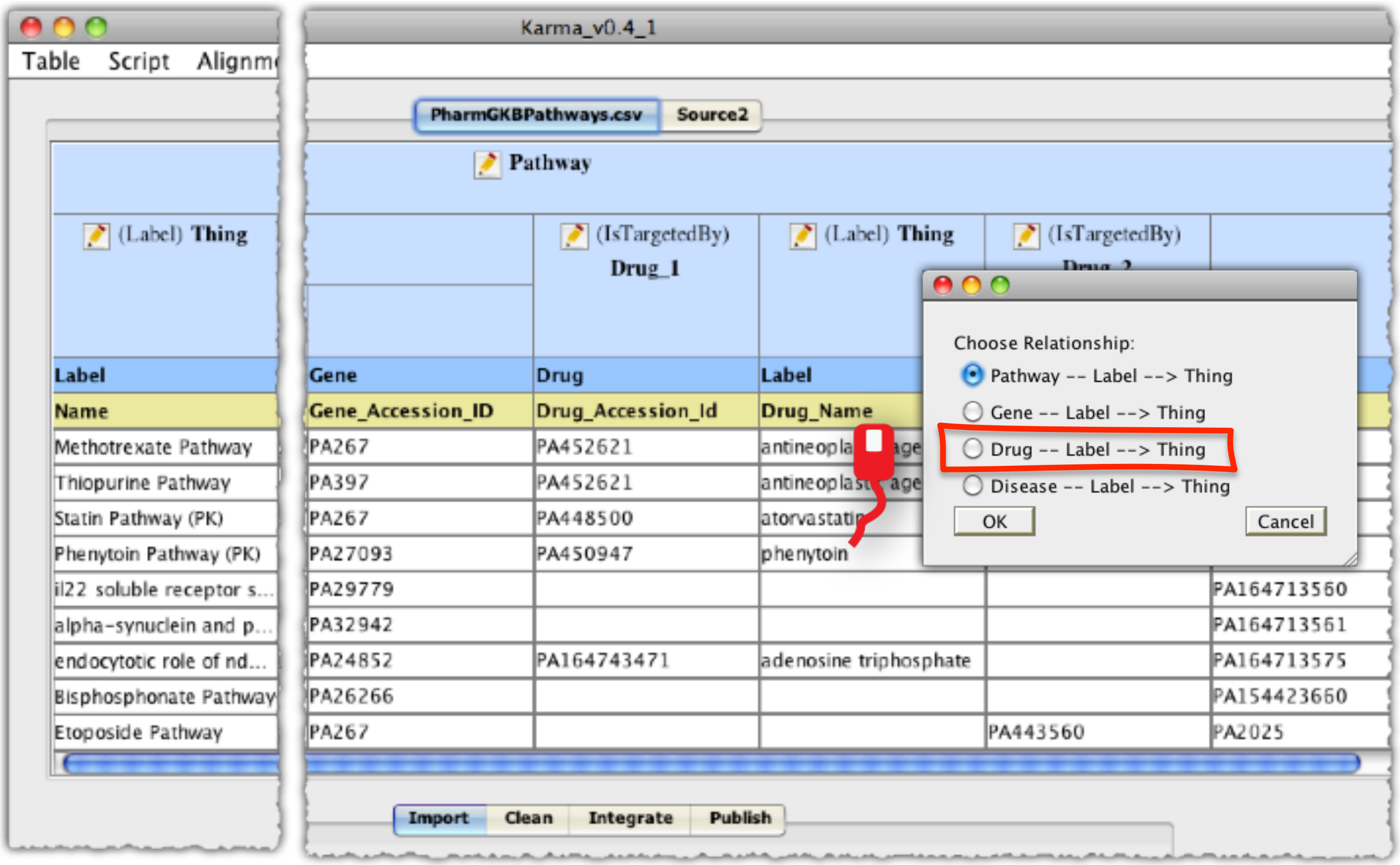
Interactive Refinement of the Relationships



Interactive Refinement of the Relationships



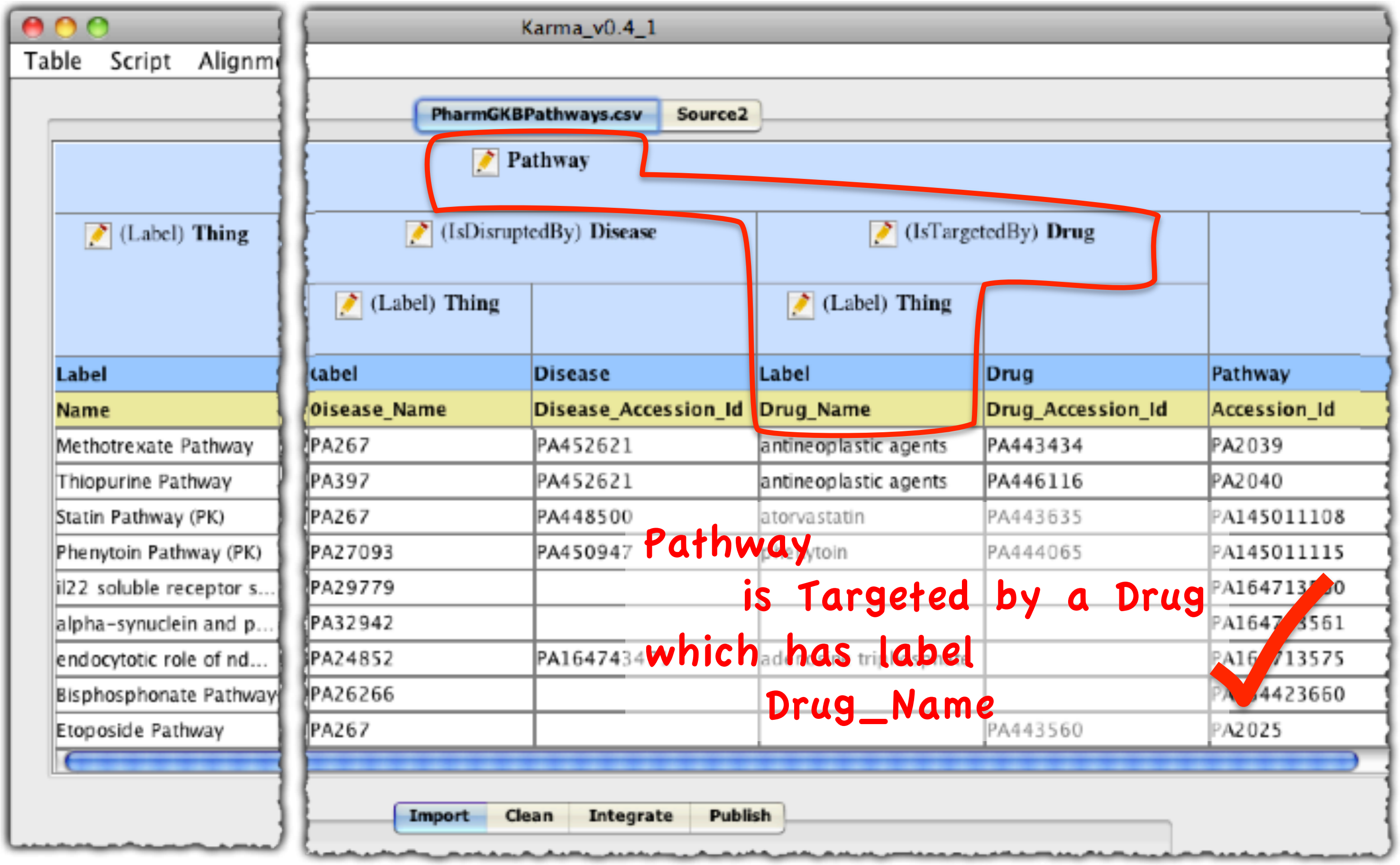
Interactive Refinement of the Relationships



USC Viterbi
School of Engineering

Interactive Refinement of the Relationships

ISI
Information Sciences Institute



USC

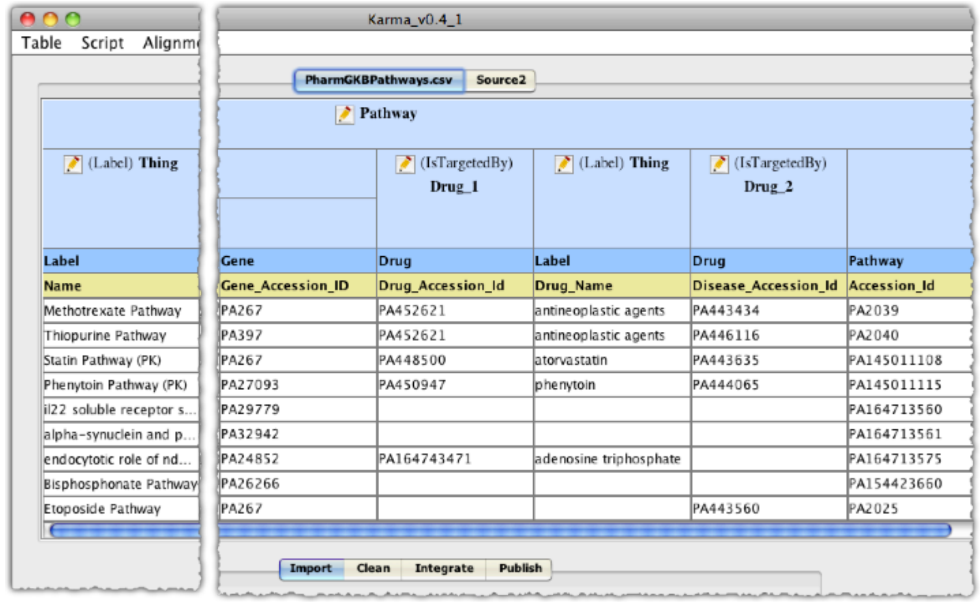
Generation of the Source Descriptions: Idea

- **From**
  - sources combined by the user in the interface, and
  - selected steiner tree over the ontology
- **Construct**
  - GLAV rule (st-tgd): logical implication with conjunctive formulas in antecedent and consequent
  - Use function symbols to generate URIs (object IDs)
  - Typical of data integration (e.g., [Halevy 2001]) and data exchange (e.g., [Arenas et al, 2010])

- **To generate RDF use the GLAV rule in data exchange mode**

Generation of the Source Descriptions: rule antecedent

- **From**
  - sources combined by the user in the interface
    - → **antecedent** of GLAV rule
  - selected steiner tree over the ontology
- **Construct**
  - logical GLAV rule (st-tgd)

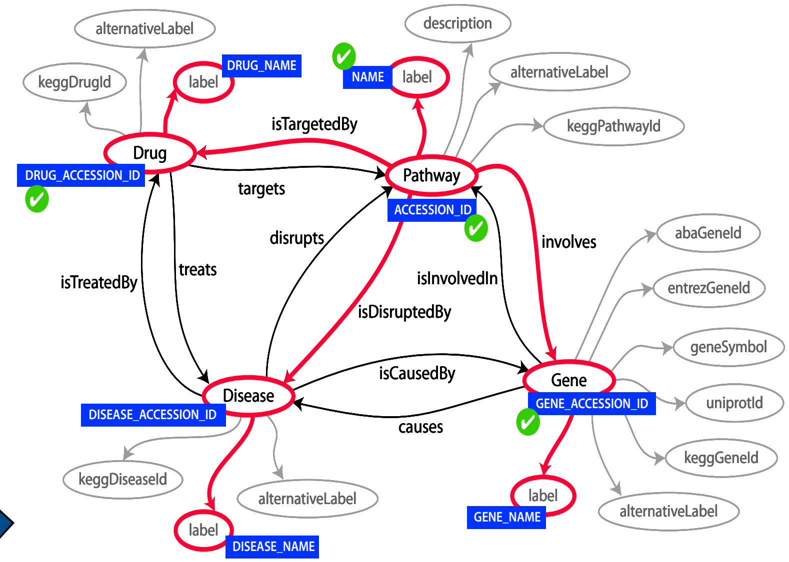


$$\text{PharmGKBPathways}(\textsc{Name}, \textsc{Accession\_Id},\ \textsc{Gene\_Accession\_ID},\ \textsc{Disease\_Name},\ \textsc{Gene\_Name}, \textsc{Disease\_Accession\_Id}, \textsc{Drug\_Name}, \textsc{Drug\_Accession\_Id})$$

(One source predicate in this example, but in general it could be a conjunction (join) of several source predicates)

Generation of the Source Descriptions: rule consequent

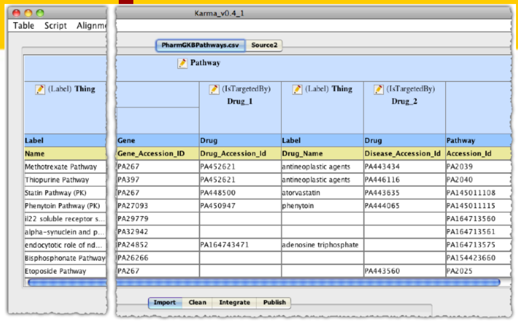
- From
  - sources combined by the user in the interface
    → antecedent of GLAV rule
  - selected steiner tree over the ontology
    → **consequent** of GLAV rule
- Construct
  - logical GLAV rule (st-tgd)



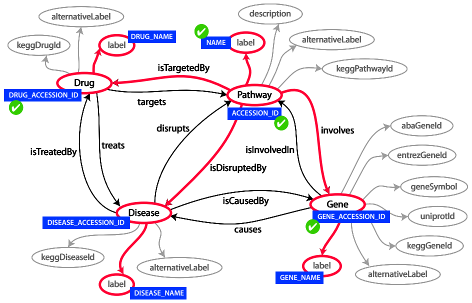
Pathway(**uri**(ACCESSION_ID)) ^ label(**uri**(ACCESSION_ID), NAME) ^
involves(**uri**(ACCESSION_ID), **uri**(GENE_ACCESSION_ID)) ^
isTargetedBy(**uri**(ACCESSION_ID), **uri**(DRUG_ACCESSION_ID)) ^
isDisruptedBy(**uri**(ACCESSION_ID), **uri**(DISEASE_ACCESSION_ID)) ^
Gene(**uri**(GENE_ACCESSION_ID)) ^ label(**uri**(GENE_ACCESSION_ID), GENE_NAME) ^
Drug(**uri**(DRUG_ACCESSION_ID)) ^ label(**uri**(DRUG_ACCESSION_ID), DRUG_NAME) ^
Disease(**uri**(DISEASE_ACCESSION_ID)) ^
label(**uri**(DISEASE_ACCESSION_ID), DISEASE_NAME)

Generation of the Source Descriptions

- **From**
  - — sources combined by the user in the interface, and
  - — selected steiner tree over the ontology
- **Construct**
  - — logical GLAV rule (st-tgd)



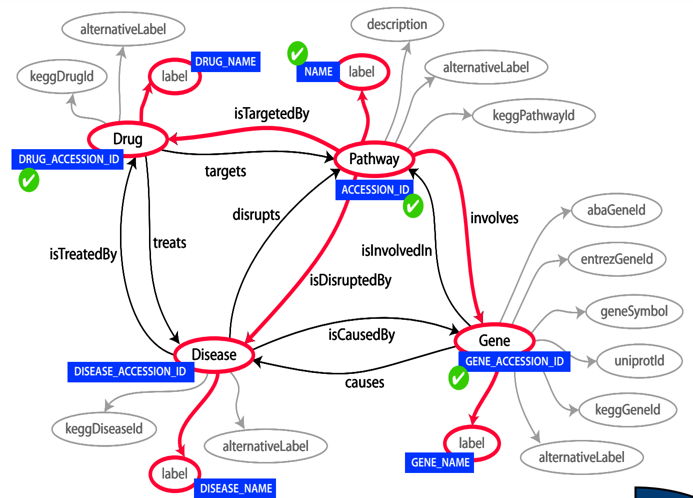
+



=

$$\begin{aligned}
&\text{PharmGKBPathways}(\textsc{Name}, \textsc{Accession_Id}, \textsc{Gene_Accession_ID}, \textsc{Disease_Name},\\
&\quad \textsc{Gene_Name}, \textsc{Disease_Accession_Id}, \textsc{Drug_Name}, \textsc{Drug_Accession_Id}) \rightarrow\\
&\quad \text{Pathway}(\textbf{uri}(\textsc{Accession_Id})) \wedge \text{label}(\textbf{uri}(\textsc{Accession_Id}), \textsc{Name}) \wedge\\
&\quad \text{involves}(\textbf{uri}(\textsc{Accession_Id}), \textbf{uri}(\textsc{Gene_Accession_ID})) \wedge\\
&\quad \text{isTargetedBy}(\textbf{uri}(\textsc{Accession_Id}), \textbf{uri}(\textsc{Drug_Accession_Id})) \wedge\\
&\quad \text{isDisruptedBy}(\textbf{uri}(\textsc{Accession_Id}), \textbf{uri}(\textsc{Disease_Accession_Id})) \wedge\\
&\quad \text{Gene}(\textbf{uri}(\textsc{Gene_Accession_ID})) \wedge \text{label}(\textbf{uri}(\textsc{Gene_Accession_ID}), \textsc{Gene_Name}) \wedge\\
&\quad \text{Drug}(\textbf{uri}(\textsc{Drug_Accession_Id})) \wedge \text{label}(\textbf{uri}(\textsc{Drug_Accession_Id}), \textsc{Drug_Name}) \wedge\\
&\quad \text{Disease}(\textbf{uri}(\textsc{Disease_Accession_Id})) \wedge\\
&\quad \text{label}(\textbf{uri}(\textsc{Disease_Accession_Id}), \textsc{Disease_Name})
\end{aligned}$$

Generation of the Source Descriptions: rule consequent



Node →Class (unary predicate)

Edge → binary predicate

- **Object property (class to class)**
- **Data property (class to literal)**

Use function symbols to create URIs:

- Pathway Accession ID = PA164713560
- **uri**(PA164713560) = http://www.semanticweb.org/ontologies/bio#Pathway_PA164713560

Pathway(**uri**(ACCESSION_ID)) ^ label(**uri**(ACCESSION_ID), NAME) ^
involves(**uri**(ACCESSION_ID), **uri**(GENE_ACCESSION_ID)) ^
isTargetedBy(**uri**(ACCESSION_ID), **uri**(DRUG_ACCESSION_ID)) ^
isDisruptedBy(**uri**(ACCESSION_ID), **uri**(DISEASE_ACCESSION_ID)) ^
Gene(**uri**(GENE_ACCESSION_ID)) ^ label(**uri**(GENE_ACCESSION_ID), GENE_NAME) ^
Drug(**uri**(DRUG_ACCESSION_ID)) ^ label(**uri**(DRUG_ACCESSION_ID), DRUG_NAME) ^
Disease(**uri**(DISEASE_ACCESSION_ID)) ^
label(**uri**(DISEASE_ACCESSION_ID), DISEASE_NAME)

Generating the RDF

Evaluating the GLAV rule generates the desired RDF

- **Data exchange from relational to RDF data (triples)**
- **Unary predicate → rdf:type triple**
- **Binary predicates → object or data property triples**
  - If uri() function in both arguments of predicate, then object property, otherwise data property

USC Viterbi School of Engineering

Generating the RDF

Information Sciences Institute

Input Tuple

[Name:PhenytoinPathway(PK); Gene_Accession_ID:PA27093; Accession_Id:PA145011115; Disease_Name:Epilepsy; Gene_Name:CYP1A2; Disease_Accession_Id:PA444065; Drug_Name:phenytoin; Drug_Accession_Id:PA450947;]

GLAV Rule

PharmGKBPathways(NAME, ACCESSION_ID, GENE_ACCESSION_ID, DISEASE_NAME, GENE_NAME, DISEASE_ACCESSION_ID, DRUG_NAME, DRUG_ACCESSION_ID) →
Pathway(**uri**(ACCESSION_ID)) ^ label(**uri**(ACCESSION_ID), NAME) ^
involves(**uri**(ACCESSION_ID), **uri**(GENE_ACCESSION_ID)) ^
isTargetedBy(**uri**(ACCESSION_ID), **uri**(DRUG_ACCESSION_ID)) ^
isDisruptedBy(**uri**(ACCESSION_ID), **uri**(DISEASE_ACCESSION_ID)) ^
Gene(**uri**(GENE_ACCESSION_ID)) ^ label(**uri**(GENE_ACCESSION_ID), GENE_NAME) ^
Drug(**uri**(DRUG_ACCESSION_ID)) ^ label(**uri**(DRUG_ACCESSION_ID), DRUG_NAME) ^
Disease(**uri**(DISEASE_ACCESSION_ID)) ^
label(**uri**(DISEASE_ACCESSION_ID), DISEASE_NAME)

Output RDF

```
@prefix s: <http://www.semanticweb.org/ontologies/bio/> .
s:Pathway_PA145011115 a category:Pathway .
s:Gene_PA27093 a category:Gene .
s:Drug_PA450947 a category:Drug .
s:Disease_PA444065 a category:Disease .
s:Pathway_PA145011115 property:Label "Phenytoin Pathway (PK)" .
s:Pathway_PA145011115 property:Involves s:Gene_PA27093 .
s:Pathway_PA145011115 property:IsTargetedBy s:Drug_PA450947 .
s:Pathway_PA145011115 property:IsDisruptedBy s:Disease_PA444065 .
s:Gene_PA27093 property:Label "CYP1A2" .
s:Drug_PA450947 property:Label "phenytoin" .
s:Disease_PA444065 property:Label "Epilepsy" .
```

USC

Evaluation Methodology

- **We evaluated our approach by integrating the same bioinformatics sources integrated by Becker et al.**
  — PharmGKB
  — ABA
  — KEGG Pathway
  — UniProt

- **We measured the following metrics:**
  — Equivalence of the mappings generated by Karma to the manually generated Becker et al. R2R mappings
  — The effort required to produce the mappings in terms of the user actions required per source

USC Viterbi School of Engineering

Evaluation Results

Information Sciences Institute

Source	Table Name	# Columns	# User Actions		
			Assigning Type	Choosing Path	Total
PharmGKB	Genes	8	8	0	8
	Drugs	3	1	2	3
	Diseases	4	2	3	5
	Pathways	5	3	0	3
ABA	Genes	4	1	1	2
KEGG Pathway	Pathways	6	5	0	5
	Diseases	2	0	1	1
	Genes	1	1	0	1
	Drugs	2	2	1	3
UniProt	Genes	4	1	1	2
		Total: 39	Total: 24	Total: 9	Total: 33
			Avg. User Actions/Property = 33/39 = 0.85		

Thee were 41 mappings, but there was no data for 2 of the mappings

Of the remaining 39 mappings, 38 were semantically equivalent to the R2R mappings

The remaining case required a data normalization rule in the mapping

USC

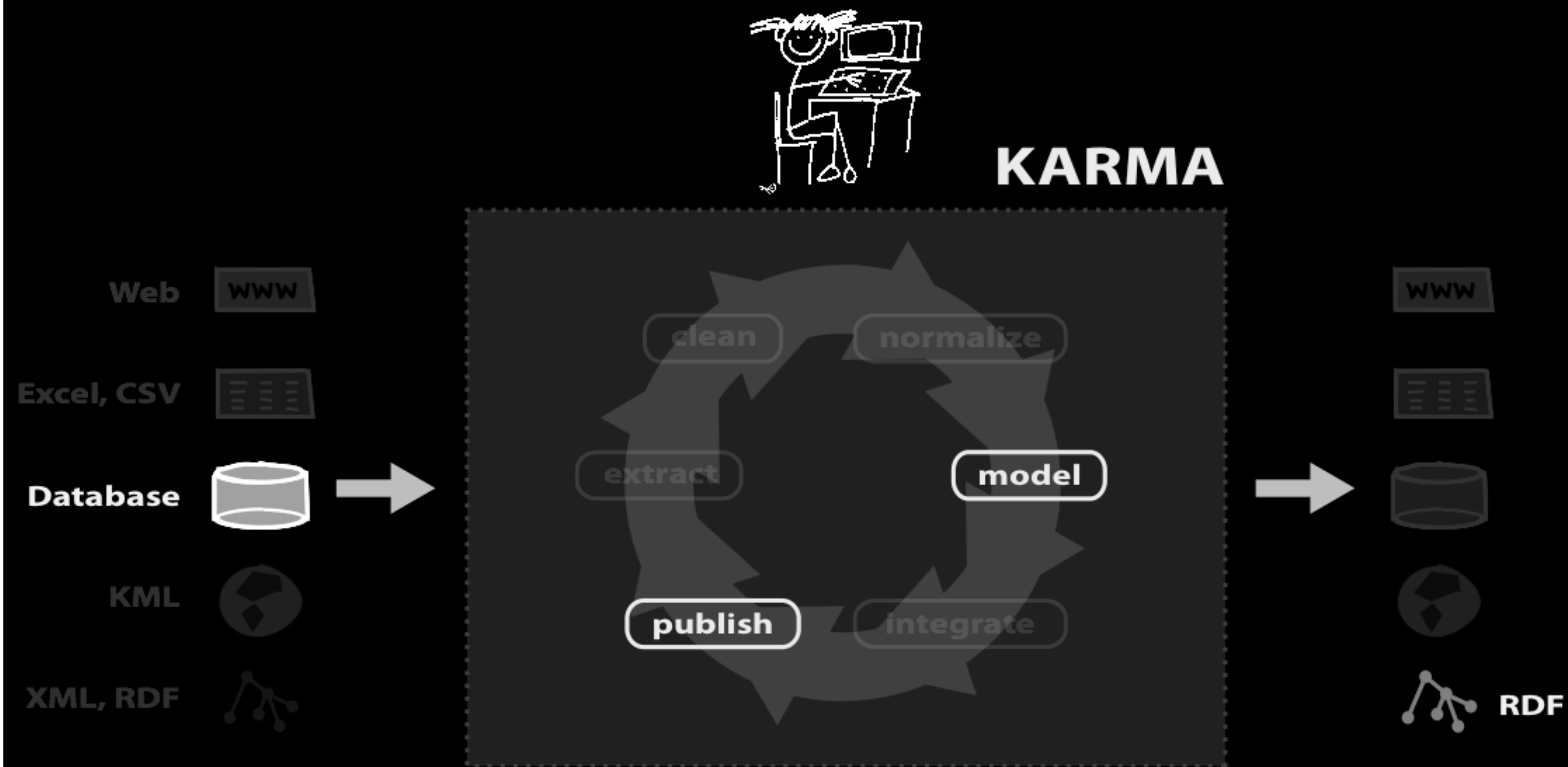
Related Work

- **Mapping Databases into RDF**
  - D2R [Bizer & Cyganiak, 2006]
    - *Maps a database into RDF using the DB schema*
  - R2R [Bizer & Shultz, 2010]
    - *Mannually defines the mappings of D2R triples to another ontology*
- **Ontology Matching**
  - [Doan et al., 2000]
    - *Learn mappings to the ontology using data, but would be analogous to just doing the semantic typing*
- **Schema Matching**
  - [Rahm et al., 2001]
    - *Generates alignments between schemas, not a fine-grained model of the data*
- **Semantic Integration of Bioinformatics Data**
  - Bio2RDF [Belleau et al., 2008]
    - *Manual conversion of sources into RDF*

Discussion

- **Presented an approach to map existing data sources directly into an ontology and generate the RDF**
  —Automates as much of the mapping as possible
  —Allows the user to easily refine the mapping
- **Makes it possible to rapidly integate data sources over an integrated domain model**
- **Using the generated mapping rule, we are now working on supporting a SPARQL endpoint**
  —The RDF data will be generated on the fly

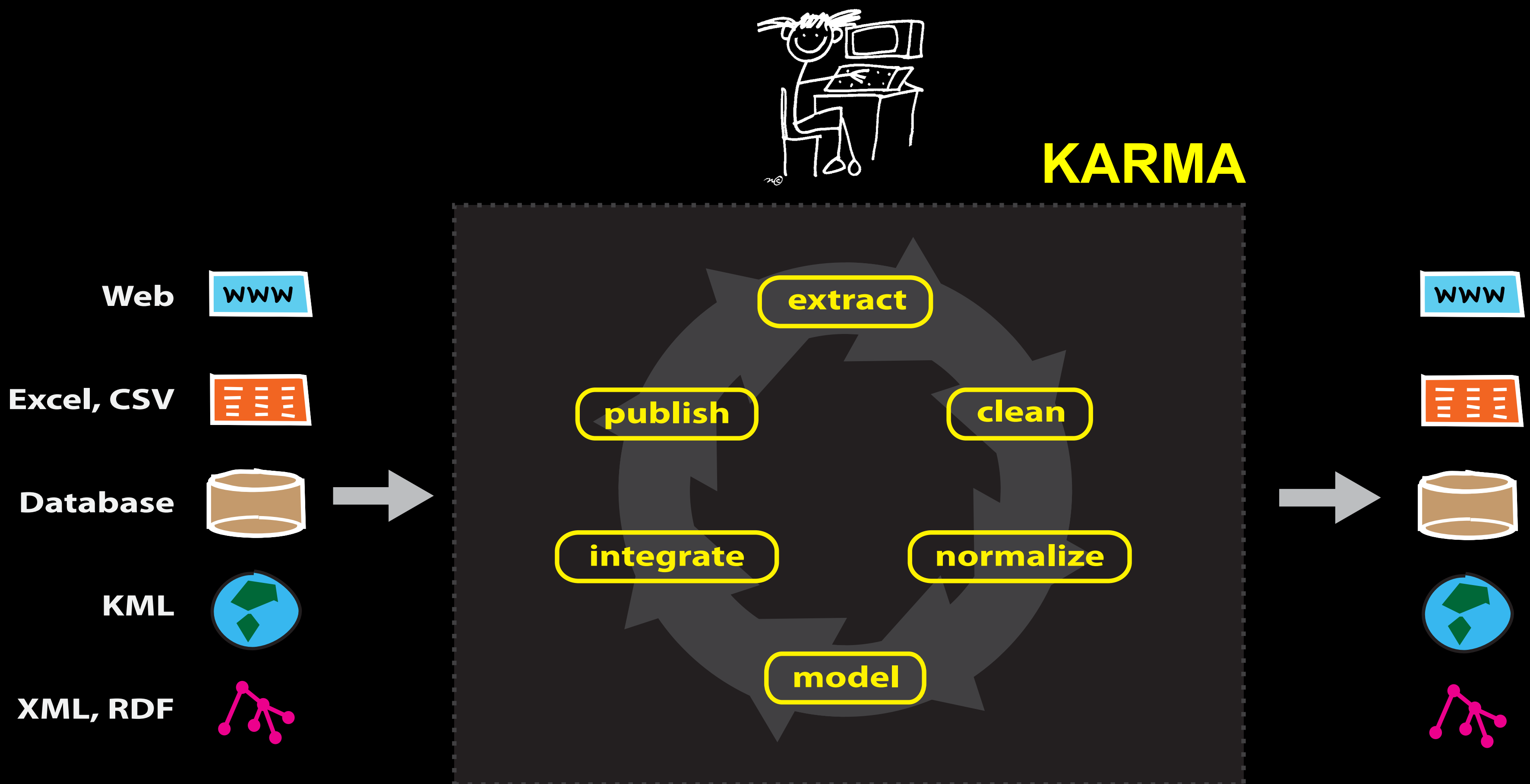
Focus of This Paper



USC Viterbi School of Engineering

Overall Karma Effort

Information Sciences Institute



More Information

- **More information available on Karma:**
  - http://www.isi.edu/~knoblock
- **Contact:**
  - knoblock@isi.edu or pszekely@isi.edu
- **Software:**
  - Software will be available as open source under the Apache license as soon as we complete the next version