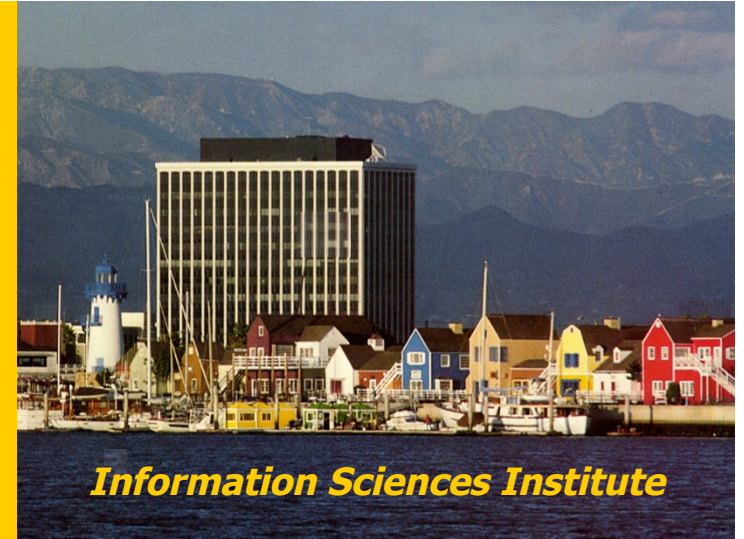


USC Viterbi
School of Engineering



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

- **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?**

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

probe_id	probe_name	gene_id	gene_symbol	gene_name	entrez_id	chromosome
1058685	A_23_P20713	729	C8G	complement com	733	9
1058684	CUST_15185_PI41	731	C9	complement com	735	5

ENTRY	NAME	DESCRIPTION	DISEASE	DRUG	GENE
map00010	Glycolysis /	Glycolysis is the	H00071		
map00020	Citrate cycle (TCA cycle)	The citrate cycle	H00073		

Entity1_id	Entity1_name	Entity2_id	Entity2_name	Relationship
PA446850	Blindness, Cortical	PA446850	Blindness, Cortical	PMID:18945600
PA446858	Neurodegenerative	PA446858	Neurodegenerative	PMID:18945600,PM

PharmGKB Accession Id	Name	Alternate Names	Type	Cross References	SMILES	External Vocabulary
PA10390	sulfonamides, urea derivatives		Drug Class			ATC:A10BB(Sulfonamides, ATC:G03C(Estrogens),ATC
PA449509	estrogens		Drug Class			ATC:L01BB(Purine analog

Entity1_id	Entity1_name	Entity2_id	Entity2_name	Relationship
PA55	APOE	PA446850	Blindness, Cortical	PMID:9804125
PA55	APOE	PA443970	Dystonia	PMID:9804125

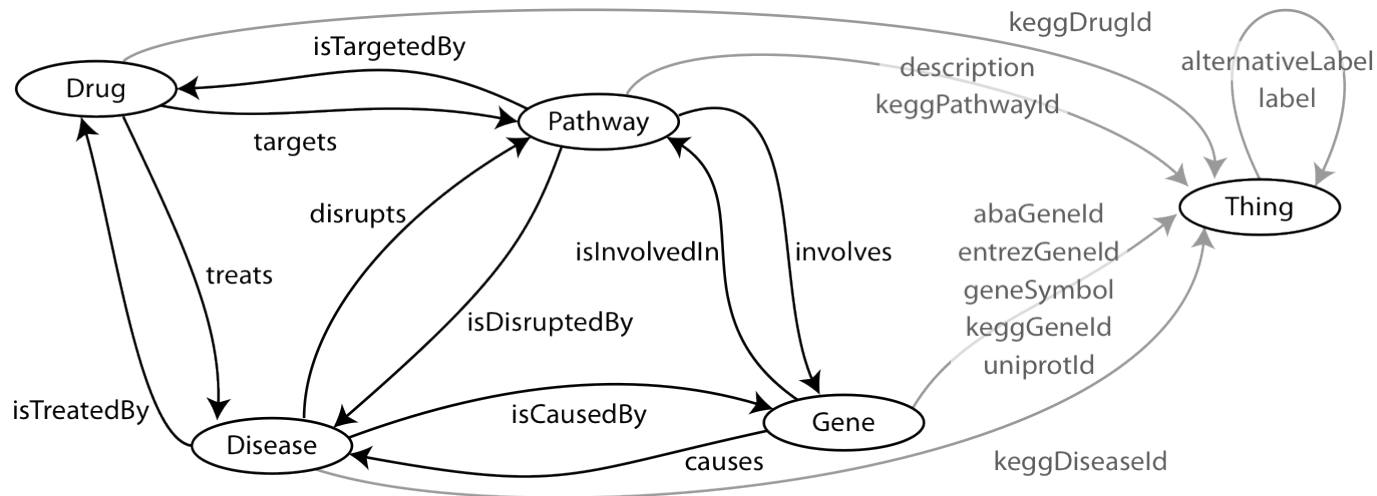
PharmGKB Accession Id	Entrez Id	Ensembl Id	UniProt Id	Name	Symbol	Alternate Names
PA117	1312		P21964	catechol-O- COMT	OTTHUMP00000197750,"OTTHUMP0000019	
PA121	1548		P11509	cytochrome CYP2A6	CYP1A6,"coumarin 7-LCPA6,"CYP2A" "CYP2A1	

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	APA2A1	PA164743471	adenosine triphosphate		

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		
PA154432660	nitric oxide synth	PA29366	CD3A				

Motivating Example (cont.)

Accession_Id	Name	Gene_Accession_Id	Gene_Name	Drug_Accession_Id	Drug_Name	Disease_Accession_Id	Disease_Name
PA2039	Methotrexate Pathway	PA267	ABCB1	PA452621	antineoplastic agents	PA443434	Arthritis, Rheumatoid
PA2040	Thiopurine Pathway	PA397	ABCC4	PA452621	antineoplastic agents	PA446116	Inflammatory Bowel Disease
PA145011108	Statin Pathway (P)	PA267	ABCB1	PA448500	atorvastatin	PA443635	Cardiovascular Disease
PA145011115	Phenytoin Pathway	PA27093	CYP1A2	PA450947	phenytoin	PA444065	Epilepsy
PA164713560	il22 soluble receptor	PA29779	IL10RA				
PA164713561	alpha-synuclein aggregation	PA32942	PARK2				
PA164713575	endocytotic role of	PA24852	AP2A1	PA164743471	adenosine triphosphate		
PA154422660	Bisphosphonate	PA26266	CDC42				

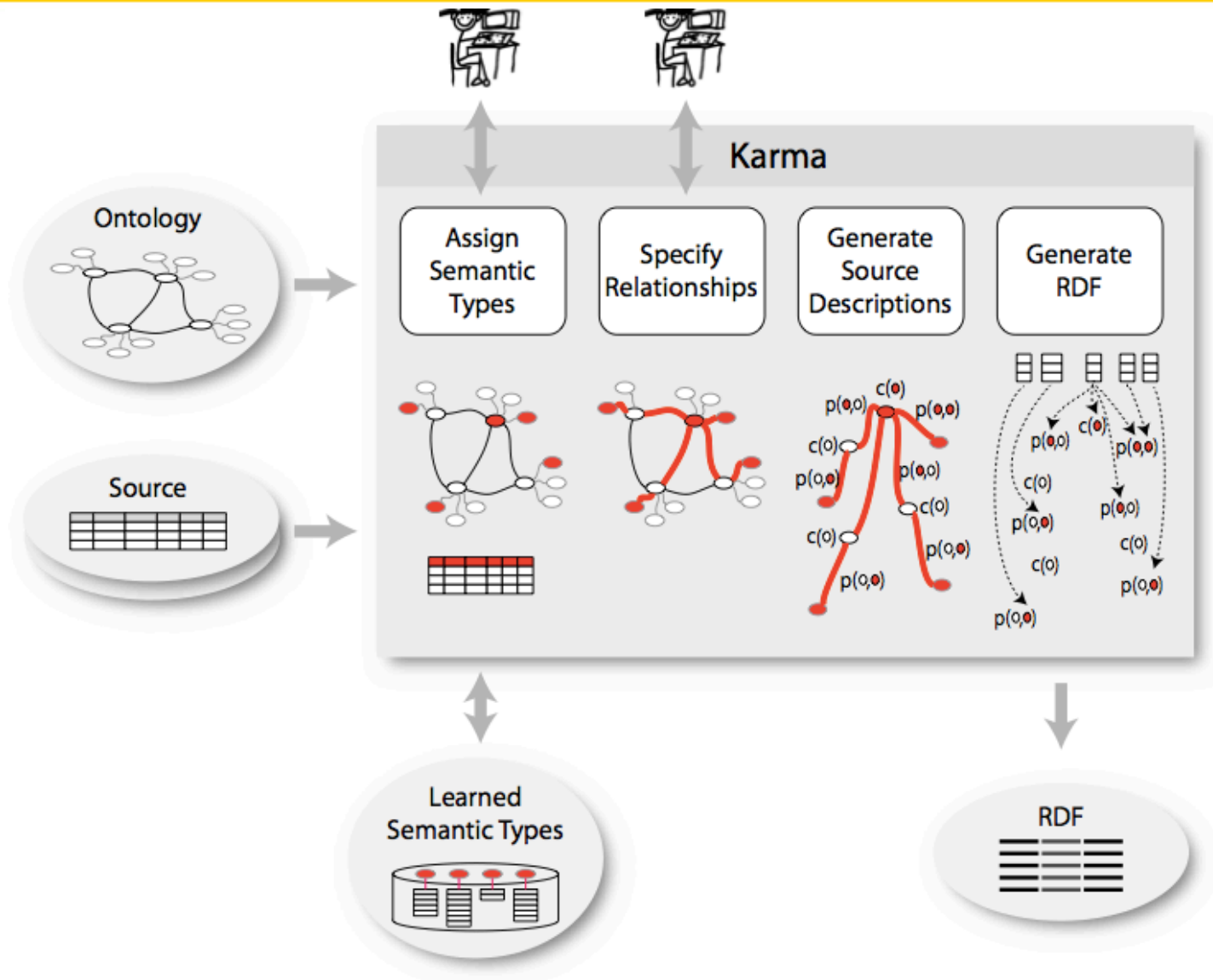
1, 2 3 4, 7, 8 9 5, 10, 11 12 6, 13, 14 15

```

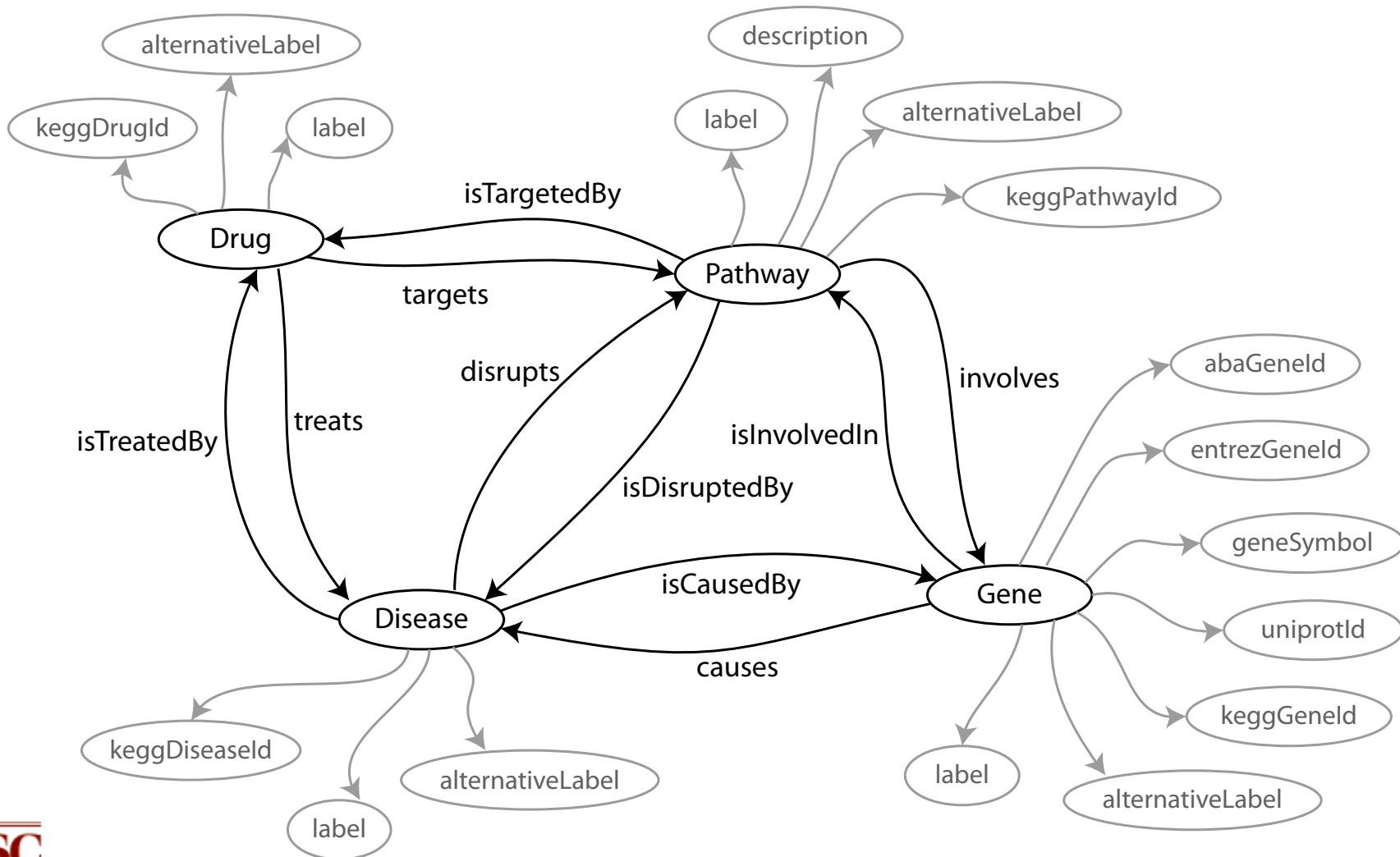
1.      :Pathway/Accession_Id/PA2039 a :Pathway;
2.      :Accession_Id "PA2039";
3.      :Label "Methotrexate Pathway";
4.      :Involves :Gene/Accession_Id/PA267;
5.      :IsTargetedBy :Drug/Accession_Id/PA452621 ;
6.      :IsDisruptedBy :Disease/Accession_Id/PA443434.
7.      :Gene/Accession_Id/PA267 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" .

```

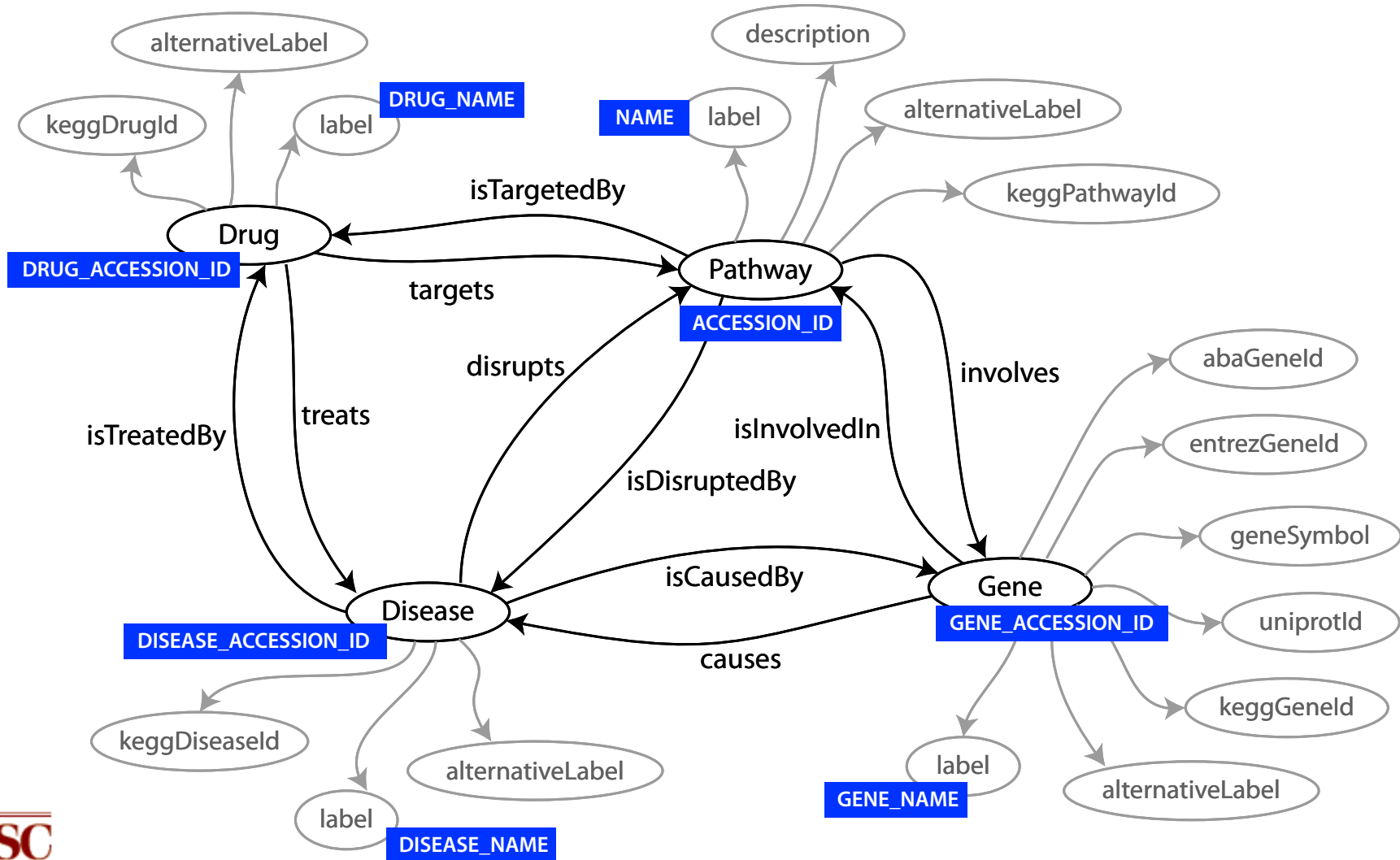
Overall Approach



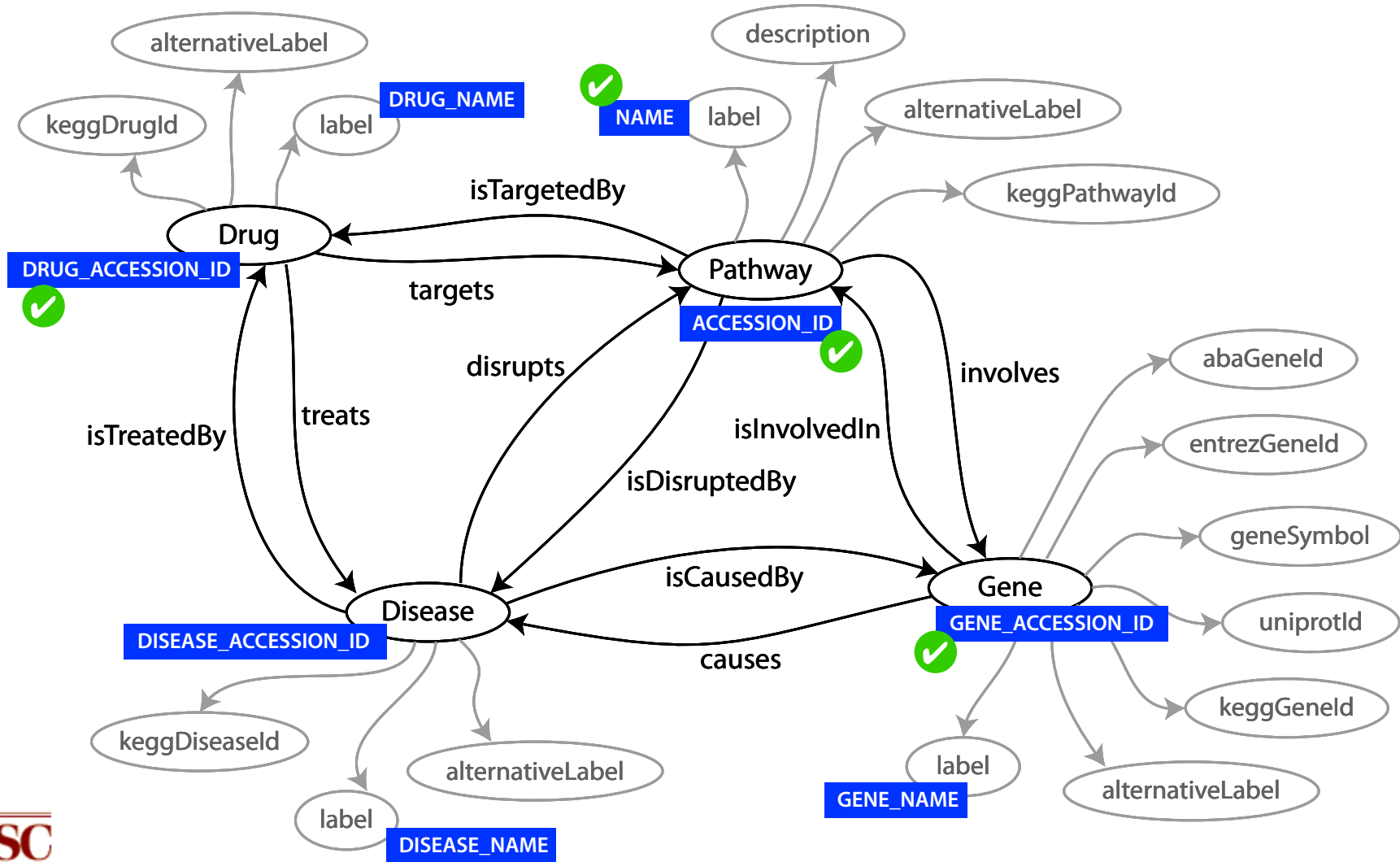
Building the Ontology Graph



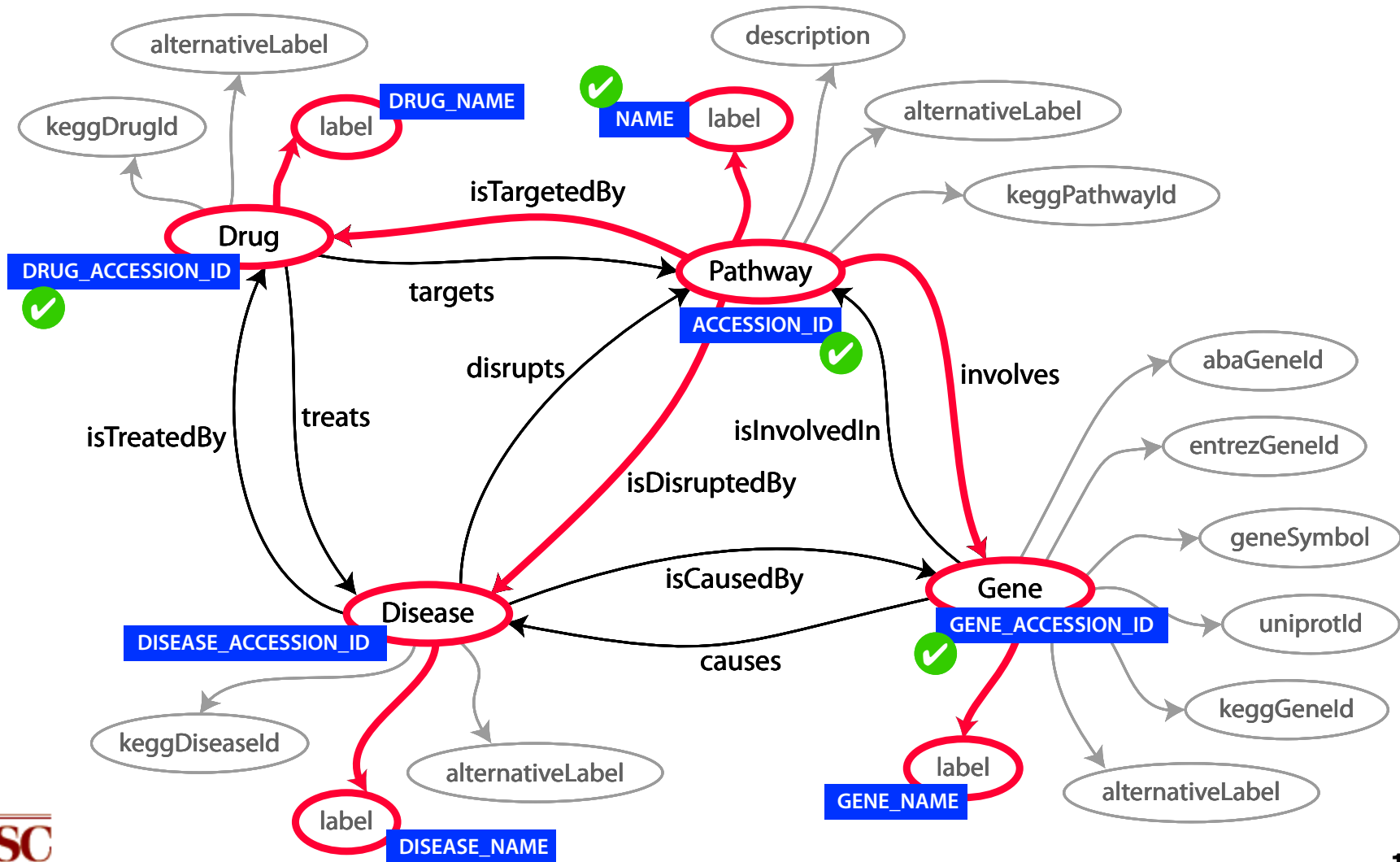
Building the Ontology Graph



Building the Ontology Graph



Building the Ontology Graph



Inferring the Semantic Types

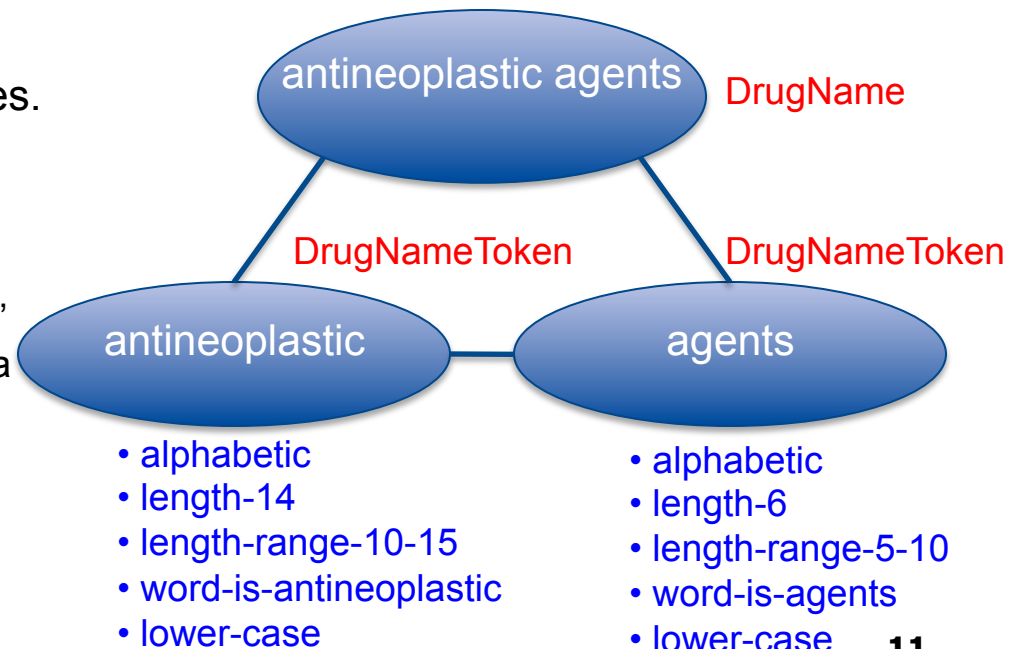
Problem: Given some columns of data, identify their semantic class.

- 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

Table Script Alignment Column

GeneName Accession_Id	PathwayName Name	GeneName Gene_Accession_ID	GeneName Gene_Name
PA2039	Methotrexate Pathway	PA267	ABCB1
PA2040	Thiopurine Pathway	PA397	ABCC4
PA145011108			
PA145011115			
PA164713560			
PA164713561			
PA164713575			
PA154423660			
PA2025			
PA2027			
PA2038			



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

Choose Semantic Type:

Class:

- Acquital
- Address
- Airport
- hasCharacteristic
- hasCharacteristic

Range of Data Properties

Create new semantic type: **PathwayID**

OK



Assigning correct label to a column of type **PathwayID**.

The CRF model discriminates between **PathwayID** and **GeneName**.

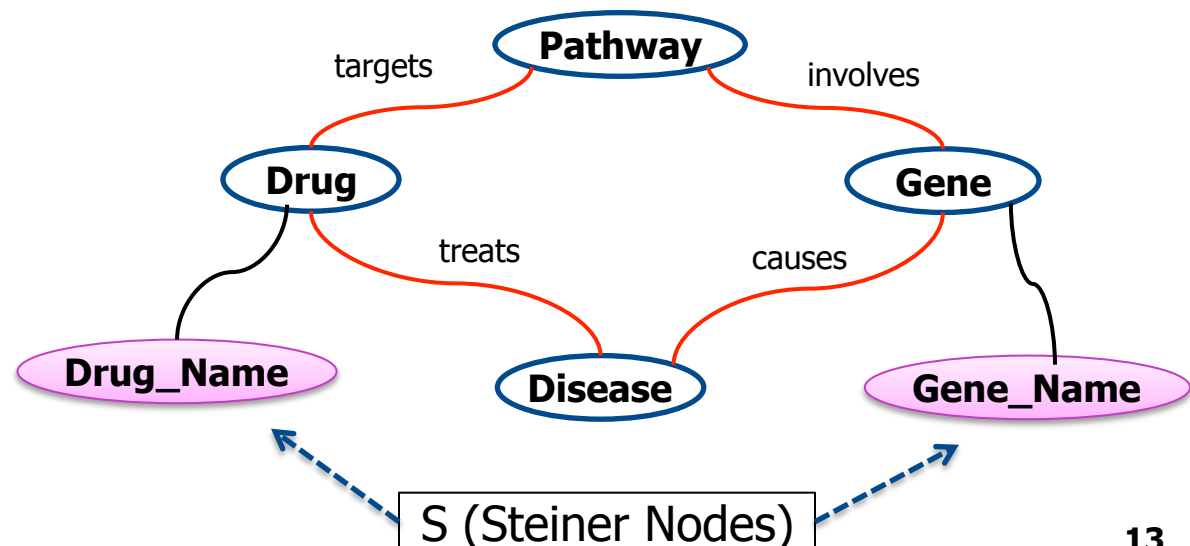


Table Script Alignment Column

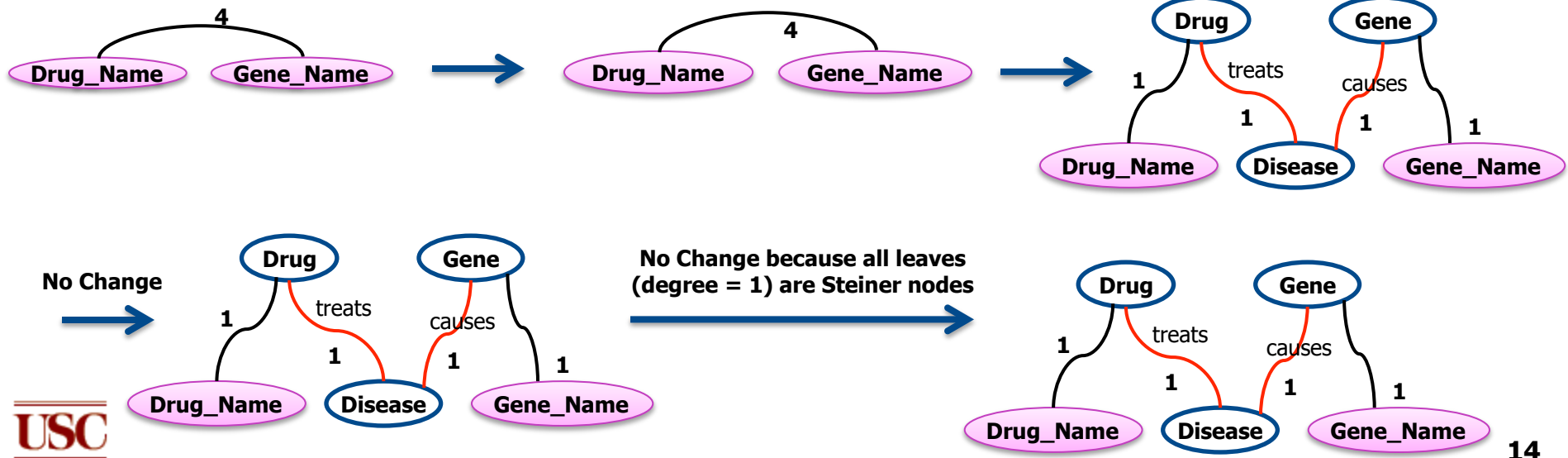
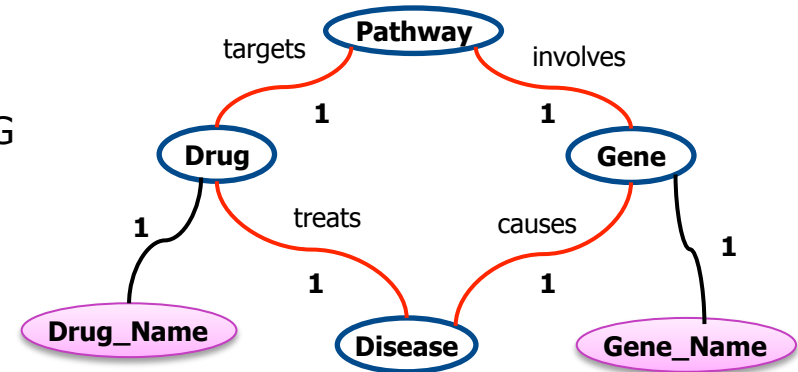
PathwayID Accession_Id	PathwayName Name	PathwayID Gene_Accession_ID	GeneName Gene_Name
PA2039	Methotrexate Pathway	PA267	ABCB1
PA2040	Thiopurine Pathway	PA397	ABCC4
PA145011108	Statin Pathway (PK)	PA267	ABCB1
PA145011115	Phenytoin Pathway (PK)	PA27093	CYP1A2
PA164713560	IL22 soluble receptor s...	PA29779	IL10RA
PA164713561	alpha-synuclein and p...	PA32942	PARK2
PA164713575	endocytotic role of nd...	PA24852	AP2A1
PA154423660	Bisphosphonate Pathway	PA26266	CDC42
PA2025	Etoposide Pathway	PA267	ABCB1
PA2027	Glucocorticoid and Infl...	PA26866	CREBBP
PA2038	Ifosfamide Pathway (PD)	PA24692	ALDH1A1

- **Apply a fast Steiner tree algorithm**
 - $G=(V,E)$, $S \subset V$, $c: E \rightarrow \mathfrak{R}$
 - 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**

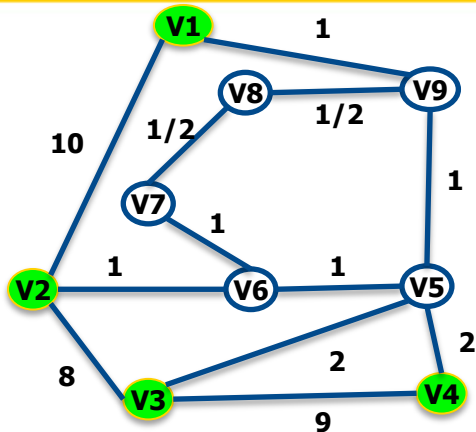
Drug_Name	Gene_Name
Antineoplastic	ABCB1
Antineoplastic	ABCC4
Atorvastatin	ABCB1



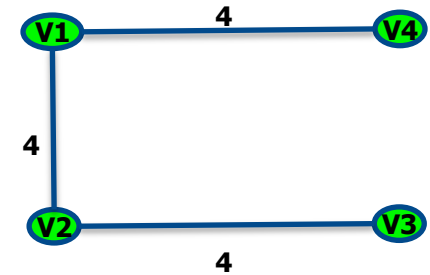
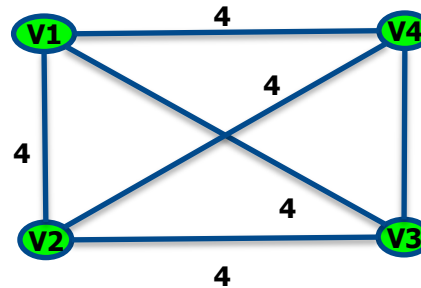
- **Step1: construct the complete graph**
 - Nodes: Steiner Nodes
 - Links Weights: shortest path from each pair in original G
- **Step2: compute MST (minimal spanning tree)**
- **Step3: replace each link with the corresponding shortest path in original G**
- **Step4: compute MST again**
- **Step5: remove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



Steiner Tree Algorithm

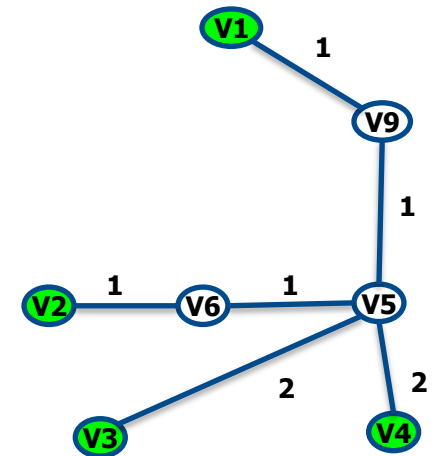
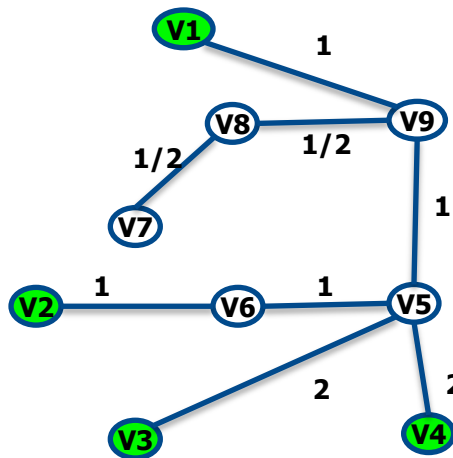
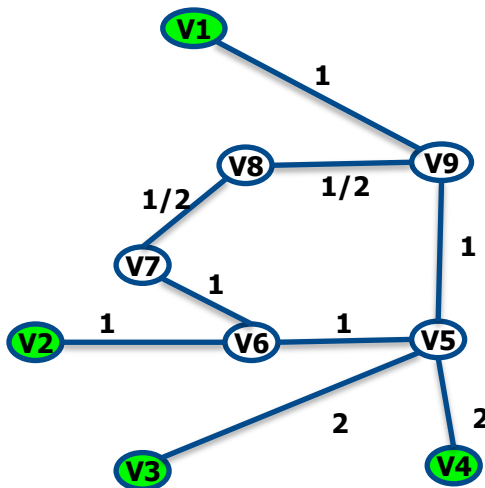


Steiner nodes: {V1, V2, V3, V4}



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

2. Compute MST

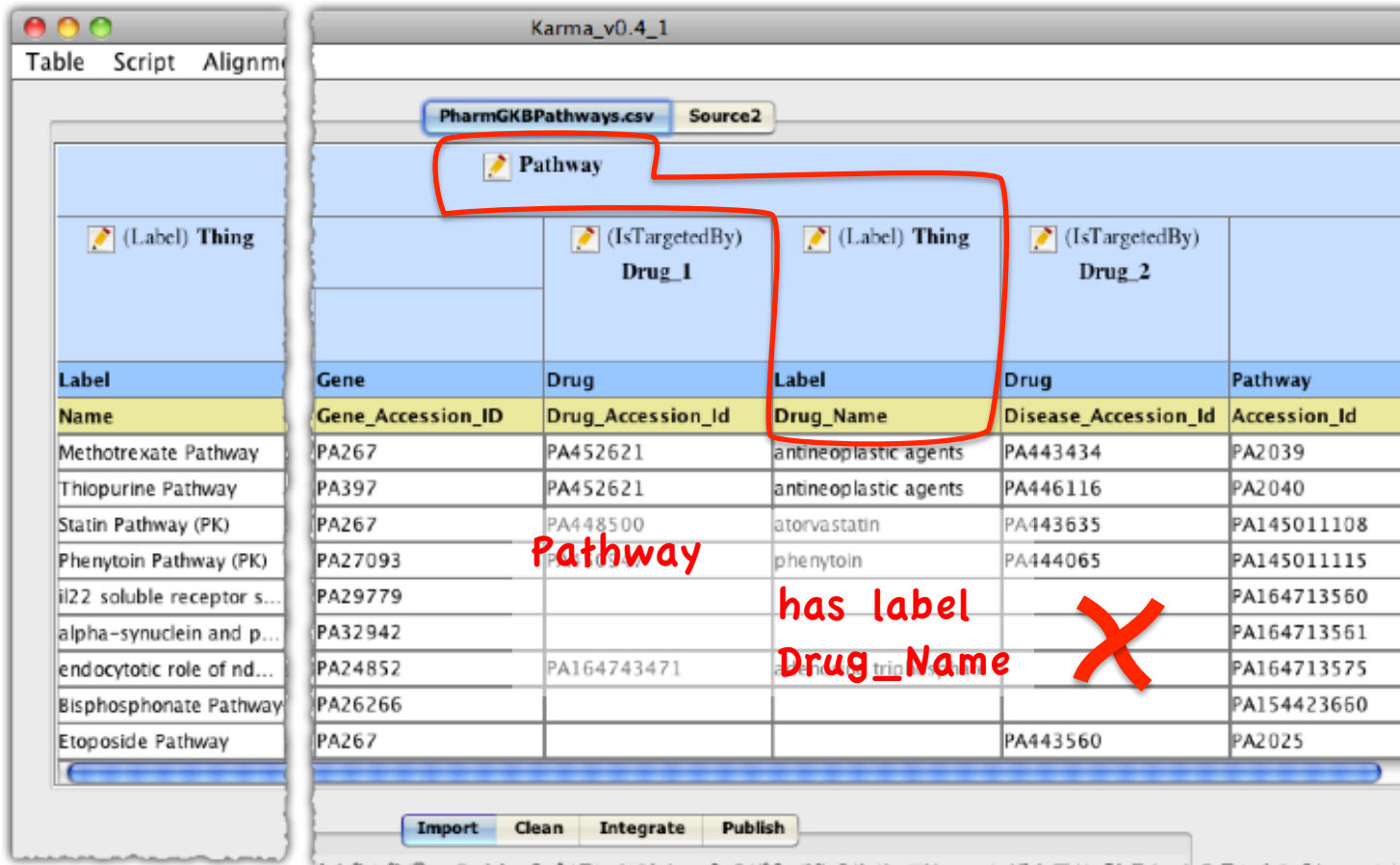


3. replace each link with the corresponding shortest path in original G

4. Compute MST

5. remove extra links until all leaves are Steiner nodes

Interactive Refinement of the Relationships



The screenshot shows the Karma_v0.4_1 interface with a table titled 'PharmGKBPathways.csv' under 'Source2'. A red box highlights the 'Pathway' column header. Below the table, red text annotations are present: 'Pathway' is written over the 'Drug' column header, and 'has label Drug_Name' is written over the 'Drug' column header with a large red 'X' next to it.

Gene	Drug	Label	Drug	Pathway
Gene_Accession_ID	Drug_Accession_Id	Drug_Name	Disease_Accession_Id	Accession_Id
PA267	PA452621	antineoplastic agents	PA443434	PA2039
PA397	PA452621	antineoplastic agents	PA446116	PA2040
PA267	PA448500	atorvastatin	PA443635	PA145011108
PA27093		phenytoin	PA444065	PA145011115
PA29779				PA164713560
PA32942				PA164713561
PA24852	PA164743471			PA164713575
PA26266				PA154423660
PA267			PA443560	PA2025

Interactive Refinement of the Relationships

PharmGKBPathways.csv Source2

Pathway

(Label) Thing

(IsTargetedBy) Drug_1

(Label) Thing

(IsTargetedBy) Drug_2

Gene	Drug	Label	Drug	Pathway
Gene_Accession_ID	Drug_Accession_Id	Drug_Name	Disease_Accession_Id	Accession_Id
PA267	PA452621	antineoplastic agents	PA443434	PA2039
PA397	PA452621	antineoplastic agents	PA446116	PA2040
PA267	PA448500	atorvastatin	PA443635	PA145011108
PA27093		phenytoin	PA444065	PA145011115
PA29779				PA164713560
PA32942				PA164713561
PA24852	PA164743471			PA164713575
PA26266				PA154423660
PA267			PA443560	PA2025

Import Clean Integrate Publish

Pathway

has label Drug_Name

Interactive Refinement of the Relationships

The screenshot shows the Karma_v0.4_1 software interface. The main window displays a table with columns: Gene, Drug, Label, and another Label. The table contains data for various pathways and drugs. A dialog box titled "Choose Relationship:" is open, showing four radio button options: "Pathway -- Label --> Thing" (selected), "Gene -- Label --> Thing", "Drug -- Label --> Thing" (highlighted with a red box), and "Disease -- Label --> Thing". The "Drug -- Label --> Thing" option is highlighted with a red box, and a red arrow points to it from the table.

Gene	Drug	Label	Label
Gene_Accession_ID	Drug_Accession_Id	Drug_Name	
PA267	PA452621	antineoplastic agent	
PA397	PA452621	antineoplastic agent	
PA267	PA448500	atorvastatin	
PA27093	PA450947	phenytoin	
PA29779			PA164713560
PA32942			PA164713561
PA24852	PA164743471	adenosine triphosphate	PA164713575
PA26266			PA154423660
PA267			PA443560 PA2025

Interactive Refinement of the Relationships

The screenshot shows the Karma_v0.4_1 interface with a data table. A red box highlights a row in the table, and red text with a checkmark indicates that the pathway is targeted by a drug with a specific label.

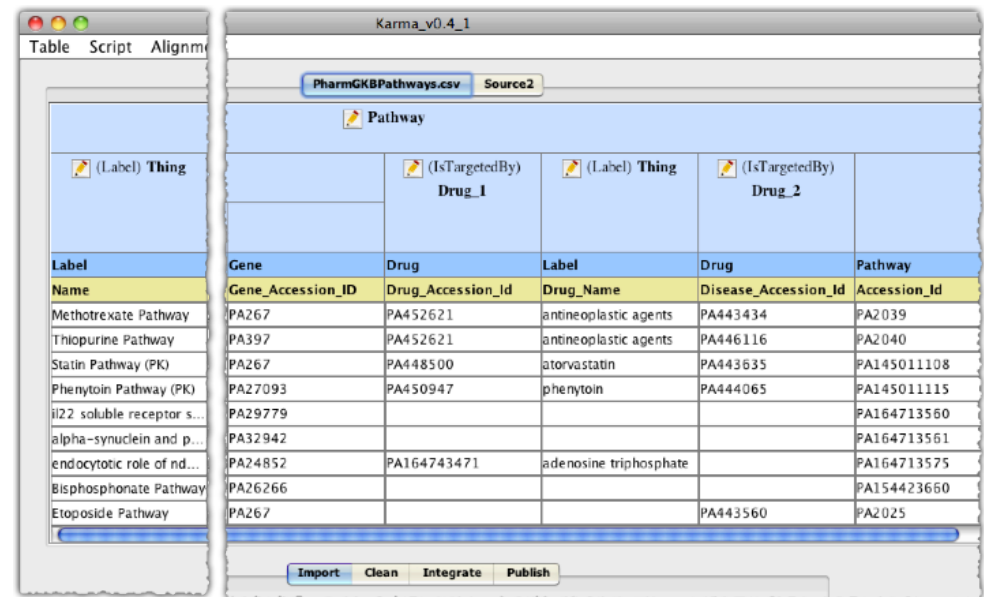
Label	Disease	Drug	Pathway
Label	Disease	Drug	Pathway
Name	Disease_Accession_Id	Drug_Name	Drug_Accession_Id
Methotrexate Pathway	PA267	antineoplastic agents	PA443434
Thiopurine Pathway	PA397	antineoplastic agents	PA446116
Statin Pathway (PK)	PA267	atorvastatin	PA443635
Phenytoin Pathway (PK)	PA27093	phenytoin	PA444065
il22 soluble receptor s...	PA29779		
alpha-synuclein and p...	PA32942		
endocytotic role of nd...	PA24852	adrenomedullary	PA164713575
Bisphosphonate Pathway	PA26266		
Etoposide Pathway	PA267		PA443560

Pathway is Targeted by a Drug which has label Drug_Name ✓

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**

- 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)

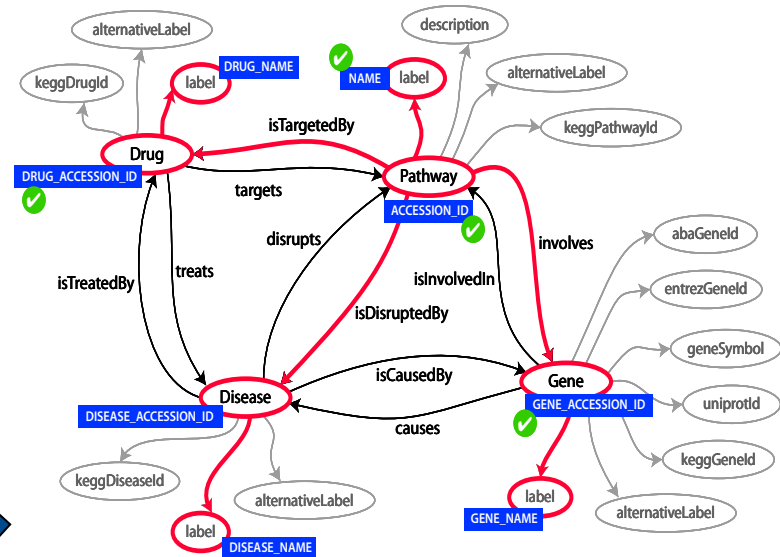


Gene	Drug	Label	Drug	Pathway
Gene_Accession_ID	Drug_Accession_Id	Drug_Name	Disease_Accession_Id	Accession_Id
PA267	PA452621	antineoplastic agents	PA443434	PA2039
PA397	PA452621	antineoplastic agents	PA446116	PA2040
PA267	PA448500	atorvastatin	PA443635	PA145011108
PA27093	PA450947	phenytoin	PA444065	PA145011115
PA29779				PA164713560
PA32942				PA164713561
PA24852	PA164743471	adenosine triphosphate		PA164713575
PA26266				PA154423660
PA267			PA443560	PA2025



PharmGKBPathways(NAME,ACCESSION_ID, GENE_ACCESSION_ID, DISEASE_NAME, GENE_NAME,DISEASE_ACCESSION_ID,DRUG_NAME,DRUG_ACCESSION_ID)

- **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)



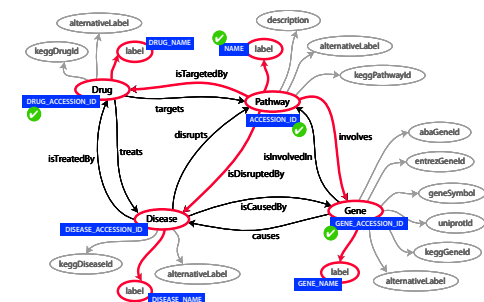
$$\begin{aligned}
 & \text{Pathway}(\text{uri}(\text{ACCESSION_ID})) \wedge \text{label}(\text{uri}(\text{ACCESSION_ID}), \text{NAME}) \wedge \\
 & \text{involves}(\text{uri}(\text{ACCESSION_ID}), \text{uri}(\text{GENE_ACCESSION_ID})) \wedge \\
 & \text{isTargetedBy}(\text{uri}(\text{ACCESSION_ID}), \text{uri}(\text{DRUG_ACCESSION_ID})) \wedge \\
 & \text{isDisruptedBy}(\text{uri}(\text{ACCESSION_ID}), \text{uri}(\text{DISEASE_ACCESSION_ID})) \wedge \\
 & \text{Gene}(\text{uri}(\text{GENE_ACCESSION_ID})) \wedge \text{label}(\text{uri}(\text{GENE_ACCESSION_ID}), \text{GENE_NAME}) \wedge \\
 & \text{Drug}(\text{uri}(\text{DRUG_ACCESSION_ID})) \wedge \text{label}(\text{uri}(\text{DRUG_ACCESSION_ID}), \text{DRUG_NAME}) \wedge \\
 & \text{Disease}(\text{uri}(\text{DISEASE_ACCESSION_ID})) \wedge \\
 & \text{label}(\text{uri}(\text{DISEASE_ACCESSION_ID}), \text{DISEASE_NAME})
 \end{aligned}$$

Generation of the Source Descriptions

Gene	Drug	Label	Drug	Pathway
Gene_Accession_Id	Drug_Accession_Id	Drug_Name	Disease_Accession_Id	Accession_Id
Methotrexate Pathway	PA43267	antimetabolite agents	PA441434	PA2039
Tiazopine Pathway	PA43262	antimetabolite agents	PA446116	PA2040
Statins Pathway (PK)	PA267	antivascular	PA443835	PA145011108
Phenytoin Pathway (PK)	PA27093	phenytoin	PA444265	PA145011115
G22 soluble receptor s...	PA29779			PA164733500
alpha-synuclein and p...	PA33842			PA164733501
endocrine role of ric...	PA24852	adenosine triphosphate		PA164733575
triphosphonate Pathway	PA26766			PA154423650
Etoposide Pathway	PA267		PA443560	PA2025

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

+



=

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)

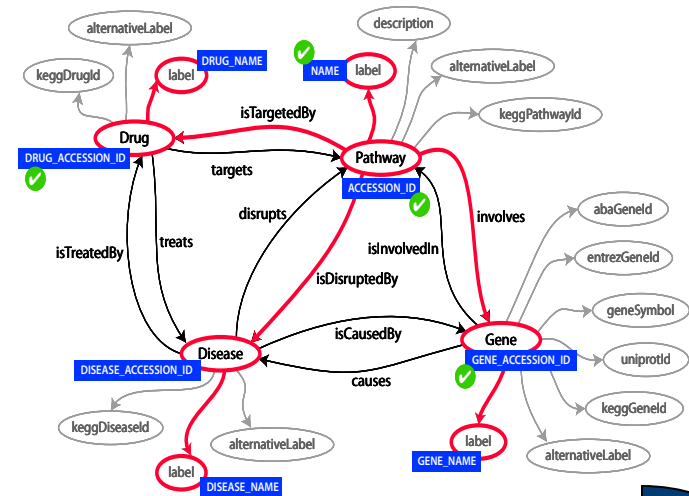
Node → Class (unary predicate)

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



$\text{Pathway}(\text{uri}(\text{ACCESSION_ID})) \wedge \text{label}(\text{uri}(\text{ACCESSION_ID}), \text{NAME}) \wedge$
 $\text{involves}(\text{uri}(\text{ACCESSION_ID}), \text{uri}(\text{GENE_ACCESSION_ID})) \wedge$
 $\text{isTargetedBy}(\text{uri}(\text{ACCESSION_ID}), \text{uri}(\text{DRUG_ACCESSION_ID})) \wedge$
 $\text{isDisruptedBy}(\text{uri}(\text{ACCESSION_ID}), \text{uri}(\text{DISEASE_ACCESSION_ID})) \wedge$
 $\text{Gene}(\text{uri}(\text{GENE_ACCESSION_ID})) \wedge \text{label}(\text{uri}(\text{GENE_ACCESSION_ID}), \text{GENE_NAME}) \wedge$
 $\text{Drug}(\text{uri}(\text{DRUG_ACCESSION_ID})) \wedge \text{label}(\text{uri}(\text{DRUG_ACCESSION_ID}), \text{DRUG_NAME}) \wedge$
 $\text{Disease}(\text{uri}(\text{DISEASE_ACCESSION_ID})) \wedge$
 $\text{label}(\text{uri}(\text{DISEASE_ACCESSION_ID}), \text{DISEASE_NAME})$

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

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" .
```

- **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

Evaluation Results

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		

There 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

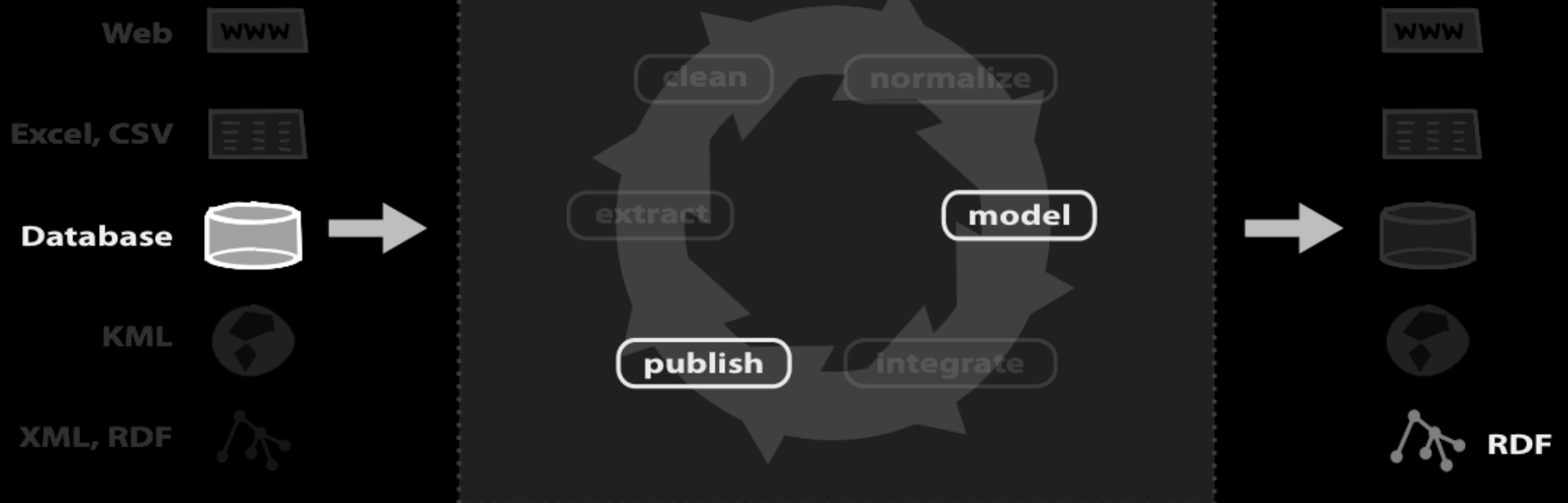
The remaining case required a data normalization rule in the mapping

- **Mapping Databases into RDF**
 - D2R [Bizer & Cyganiak, 2006]
 - *Maps a database into RDF using the DB schema*
 - R2R [Bizer & Shultz, 2010]
 - *Manually 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*

- **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 integrate 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

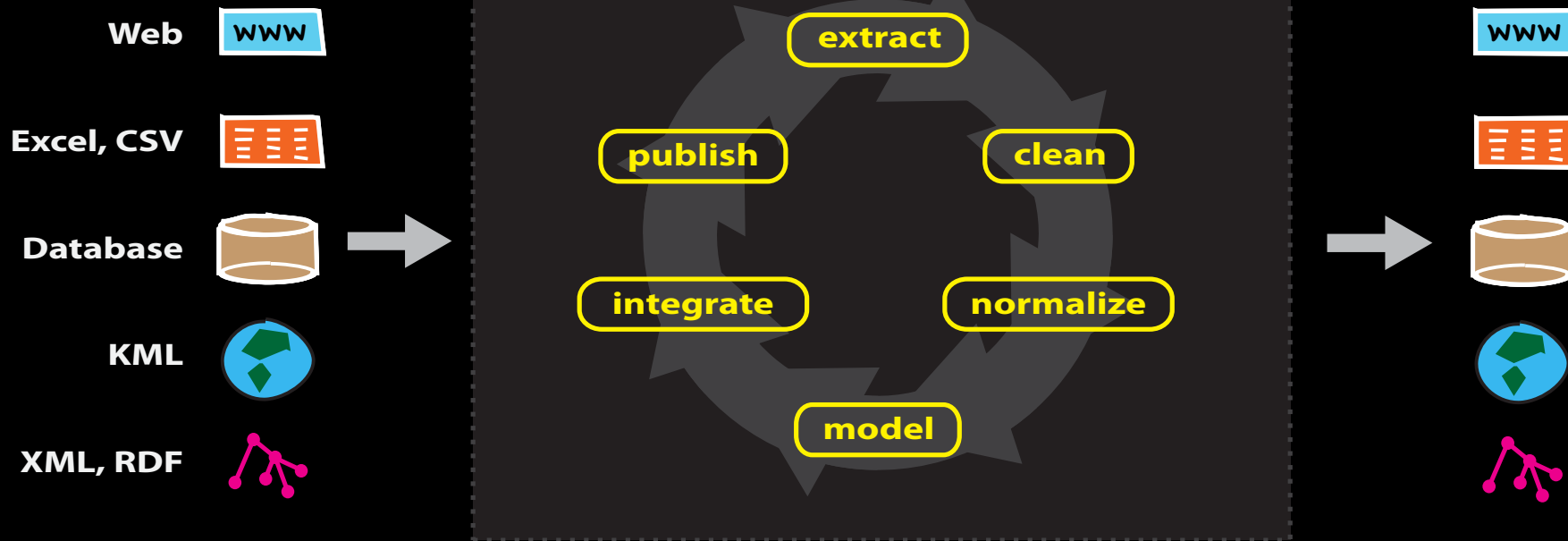


KARMA





KARMA



- **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