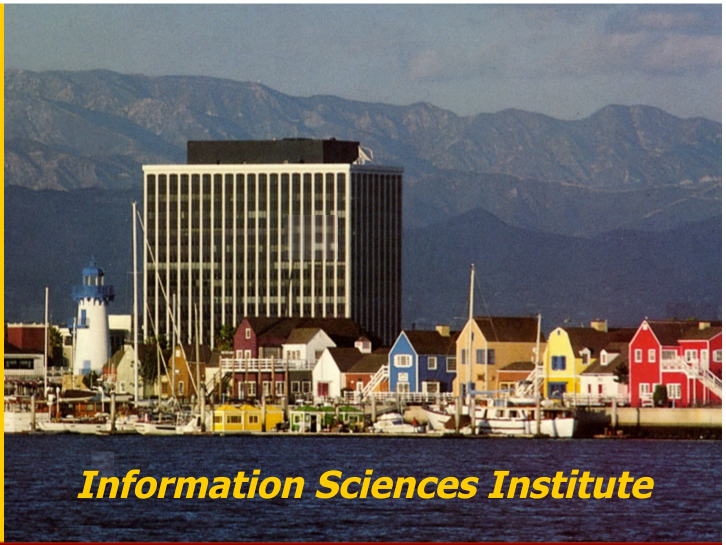
USC Viterbi
School of Engineering



Semi-Automatically Mapping Structured Sources into the Semantic Web

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

Parag Mallick
Stanford University

USC

Introduction

- **Ultimate goal of the Semantic Web:**
  - — provide access to data
  - — under clear semantics
  - — to facilitate seamless data integration
- **Linked Open Data publishing large amounts of data**
  - — but few detailed semantic descriptions
  - — little ontology reuse
- **Challenge: empower users to *easily* map and publish data with respect to desired ontologies**
  - — automate process as much as possible
  - — user corrects system interactively by demonstration (no programming)

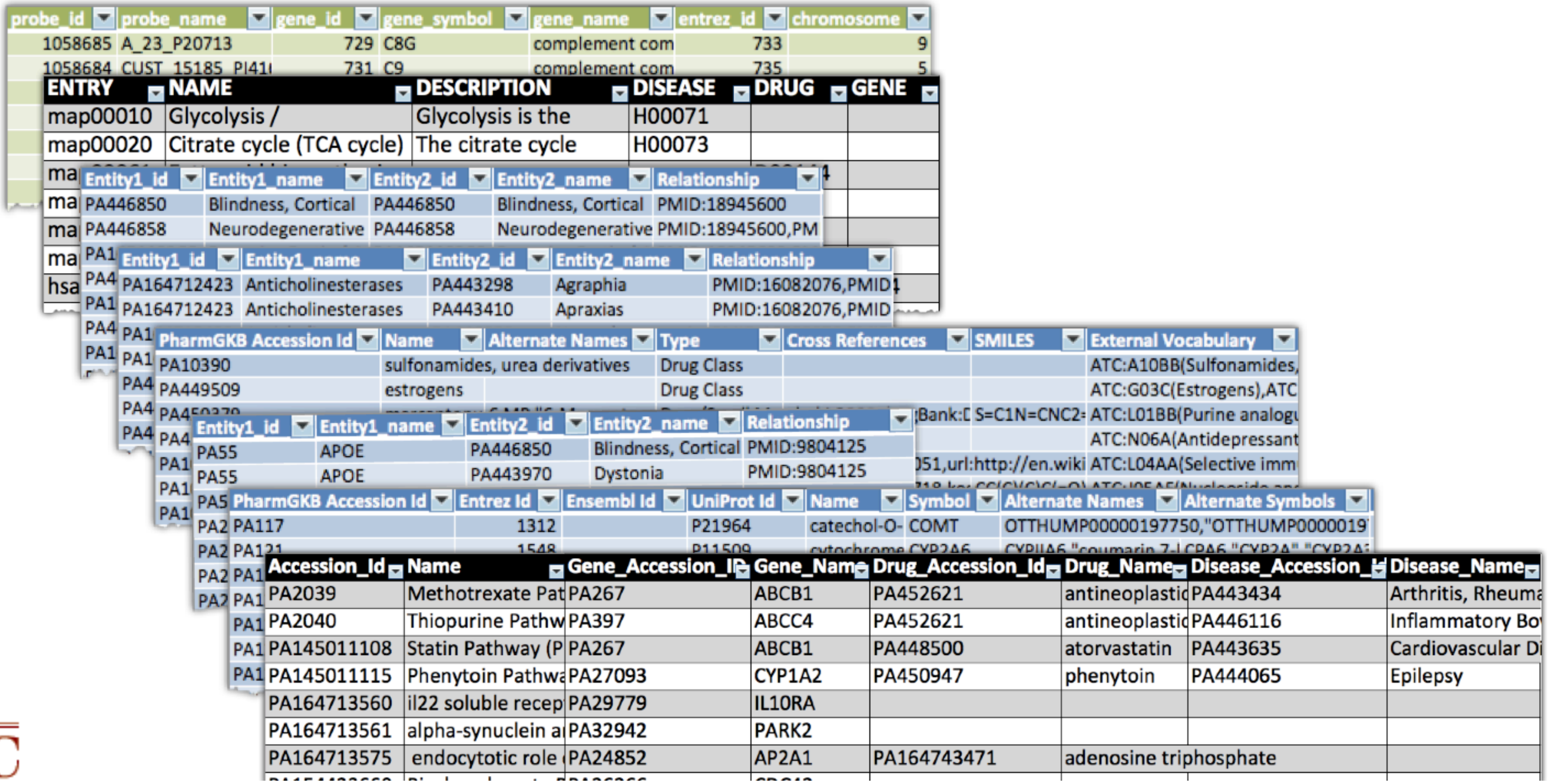
USC Viterbi
School of Engineering

Motivating Example: Sources

ISI
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, …



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

Entity1_id	Entity1_name	Entity2_id	Entity2_name	Relationship
PA164712423	Anticholinesterases	PA443298	Agraphia	PMID:16082076,PMID
PA164712423	Anticholinesterases	PA443410	Apraxias	PMID:16082076,PMID

PharmGKB Accession Id	Name	Alternate Names	Type	Cross References	SMILES	External Vocabulary
PA10390	sulfonamides, urea derivatives		Drug Class			ATC:A10BB(Sulfonamides,
PA449509	estrogens		Drug Class			ATC:G03C(Estrogens),ATC
				Bank:[	S=C1N=CNC2=	ATC:L01BB(Purine analogu
						ATC:N06A(Antidepressant
				051,url:http://en.wiki		ATC:L04AA(Selective imm

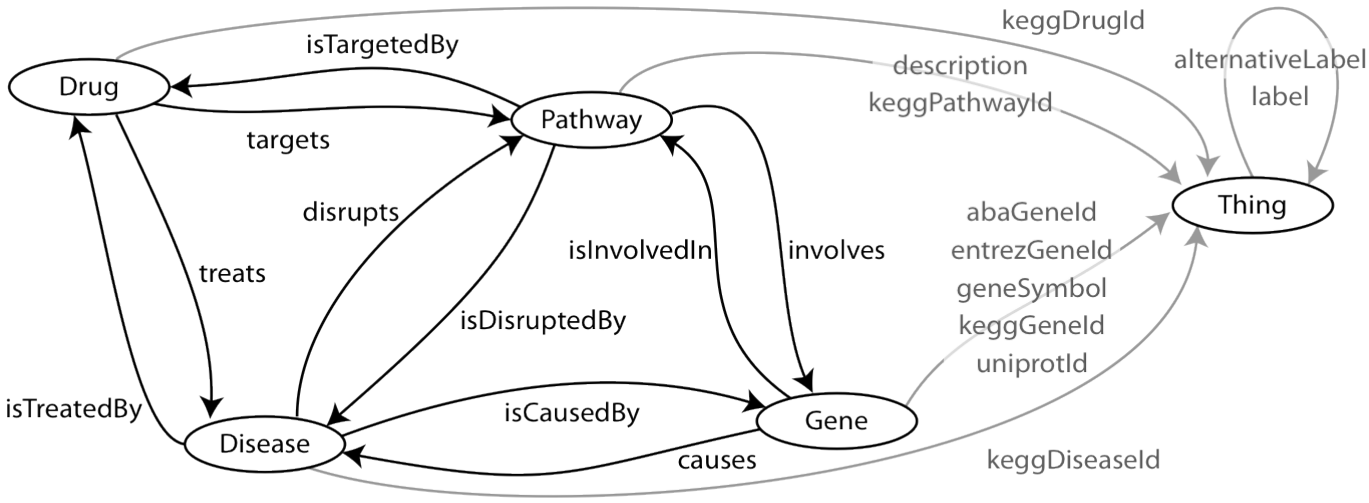
Entity1_id	Entity1_name	Entity2_id	Entity2_name	Relationship
PA55	APOE	PA446850	Blindness, Cortical	PMID:9804125
PA55	APOE	PA443970	Dystonia	PMID:9804125

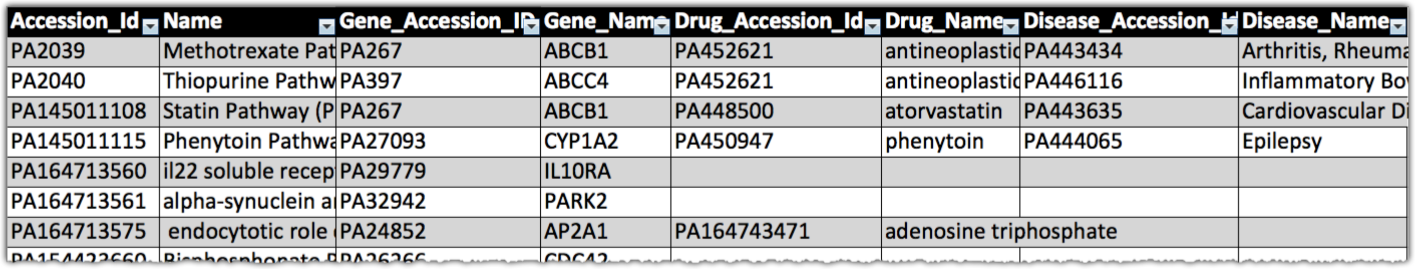
PharmGKB Accession Id	Entrez Id	Ensembl Id	UniProt Id	Name	Symbol	Alternate Names	Alternate Symbols
PA117	1312		P21964	catechol-O-	COMT	OTTHUMP00000197750,"OTTHUMP0000019	

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

Motivating Example: Ontology

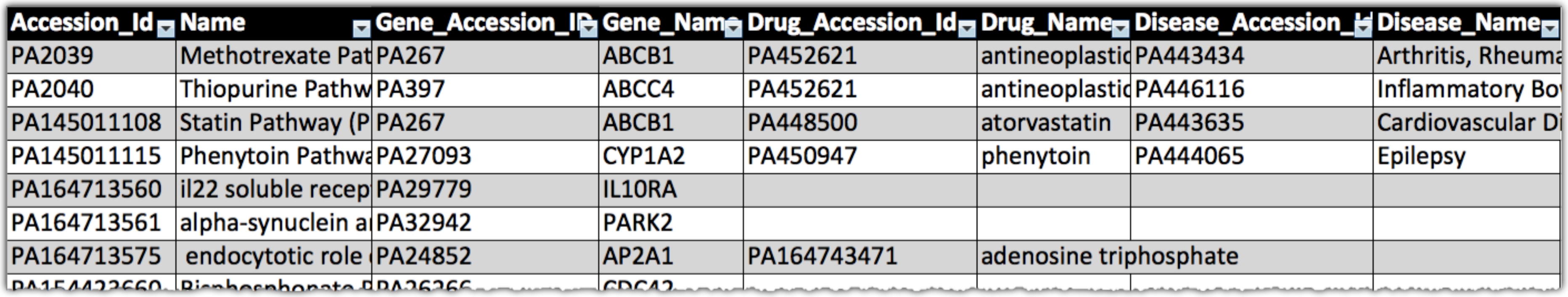




Accession_Id	Name	Gene_Accession_I	Gene_Name	Drug_Accession_Id	Drug_Name	Disease_Accession_I	Disease_Name
PA2039	Methotrexate Pat	PA267	ABCB1	PA452621	antineoplastic	PA443434	Arthritis, Rheuma
PA2040	Thiopurine Pathw	PA397	ABCC4	PA452621	antineoplastic	PA446116	Inflammatory Bo
PA145011108	Statin Pathway (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		
PA154423660	Bisphosphonate P	PA26266	CDC42				

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

Motivating Example: Formal Mapping

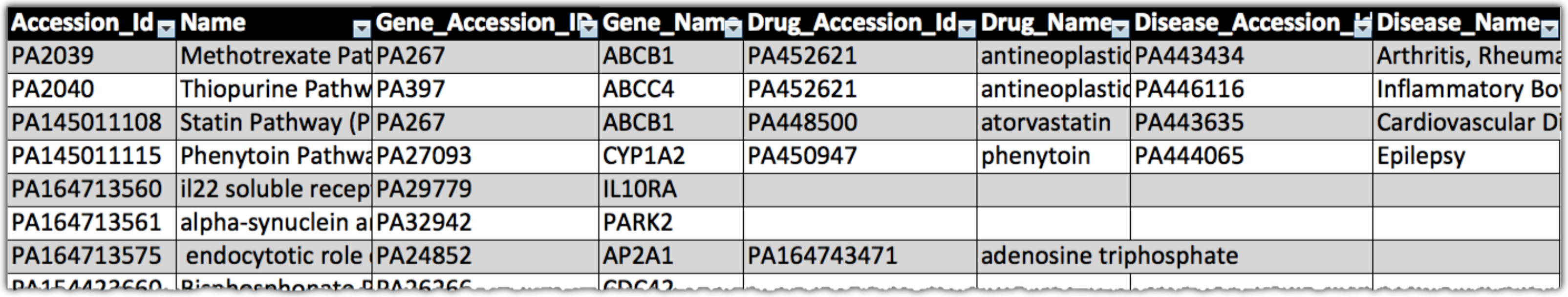


Accession_Id	Name	Gene_Accession_I	Gene_Nam	Drug_Accession_Id	Drug_Name	Disease_Accession_I	Disease_Name
PA2039	Methotrexate Pat	PA267	ABCB1	PA452621	antineoplastic	PA443434	Arthritis, Rheuma
PA2040	Thiopurine Pathw	PA397	ABCC4	PA452621	antineoplastic	PA446116	Inflammatory Bo
PA145011108	Statin Pathway (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		
PA154423660	Bisphosphonate P	PA26266	CDC42				

PharmGKBPathways(PathwayId, Name, GeneId, GeneName, DrugId, DrugName, DiseaseId, DiseaseName) →

Pathway(uri(PathwayId)) ^ name(uri(PathwayId), Name) ^
involves(uri(PathwayId), uri(GeneId)) ^
Gene(uri(GeneId)) ^ geneSymbol(uri(GeneId), GeneName) ^
isDisruptedBy(uri(PathwayId), uri(DiseaseId)) ^
Disease(uri(DiseaseId)) ^ name(uri(DiseaseId), DiseaseName)
isTargetedBy(uri(PathwayId), uri(DrugId)) ^
Drug(uri(DrugId)) ^ name(uri(DrugId), DrugName)

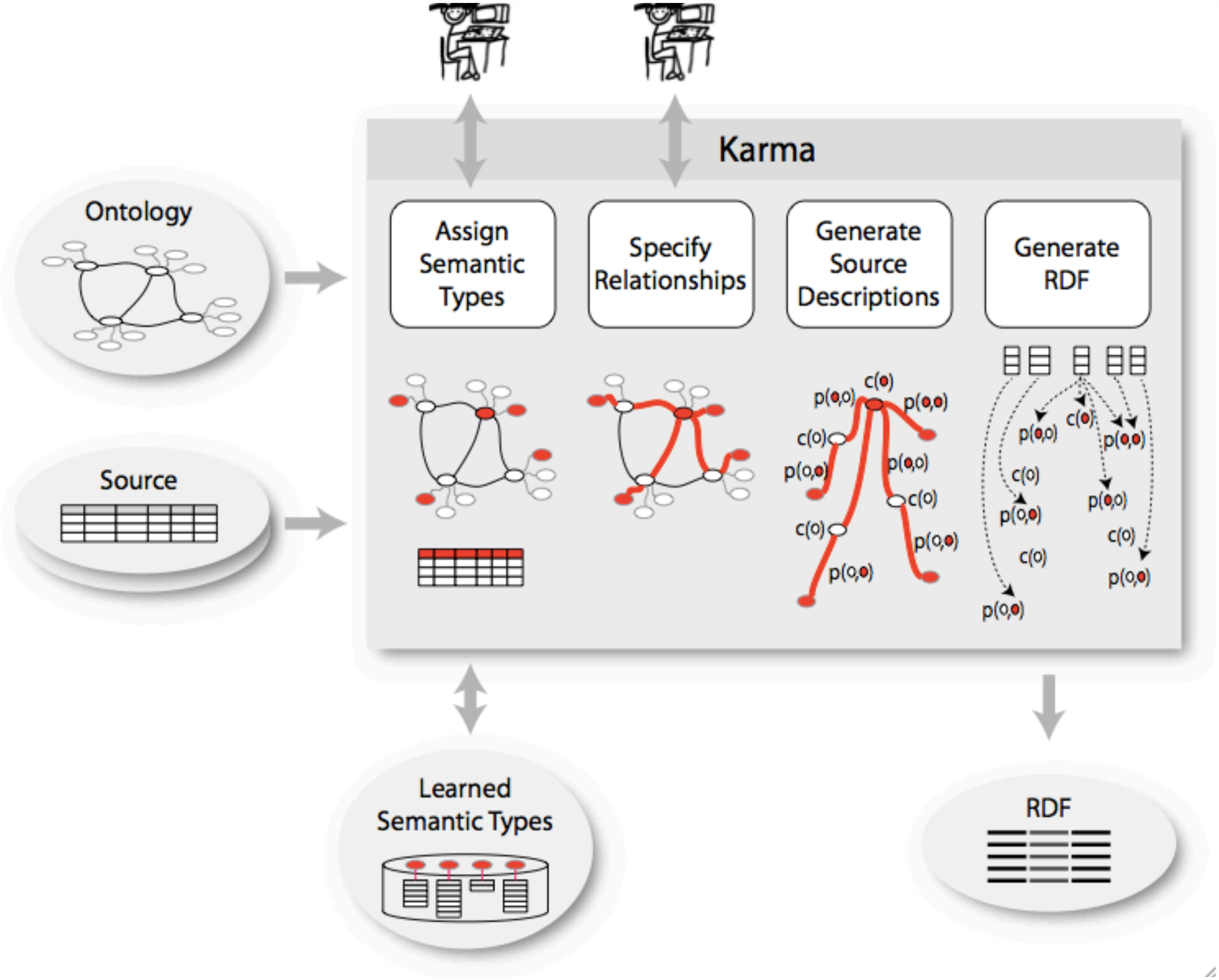
Motivating Example: Publish Data in RDF wrt Ontology



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		
PA154423660	Bisphosphonate P	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" .
```

Approach



Ontology

Source

Karma

Assign Semantic Types

Specify Relationships

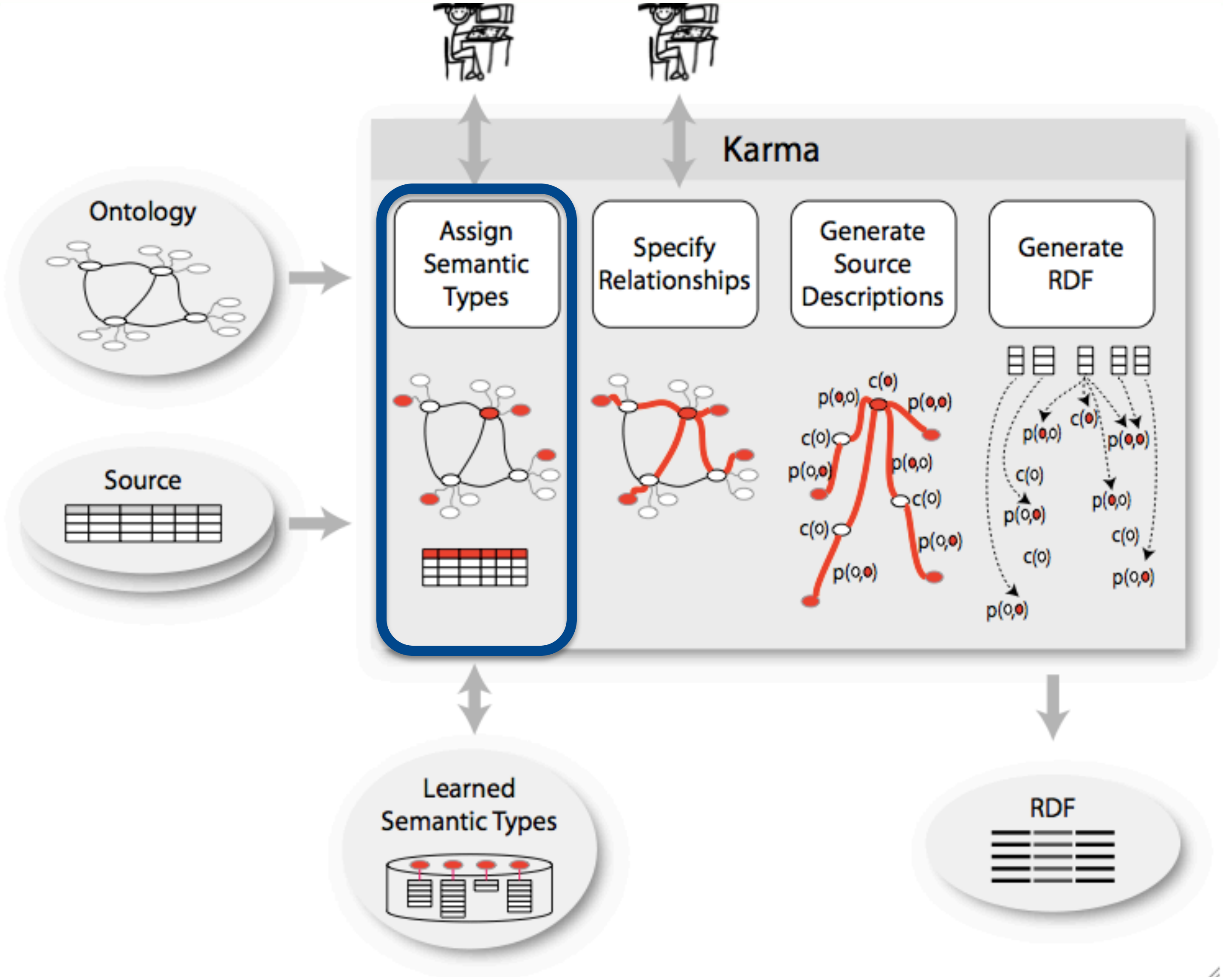
Generate Source Descriptions

Generate RDF

Learned Semantic Types

RDF

Approach: Semantic Typing



Karma

Ontology

Source

Assign Semantic Types

Specify Relationships

Generate Source Descriptions

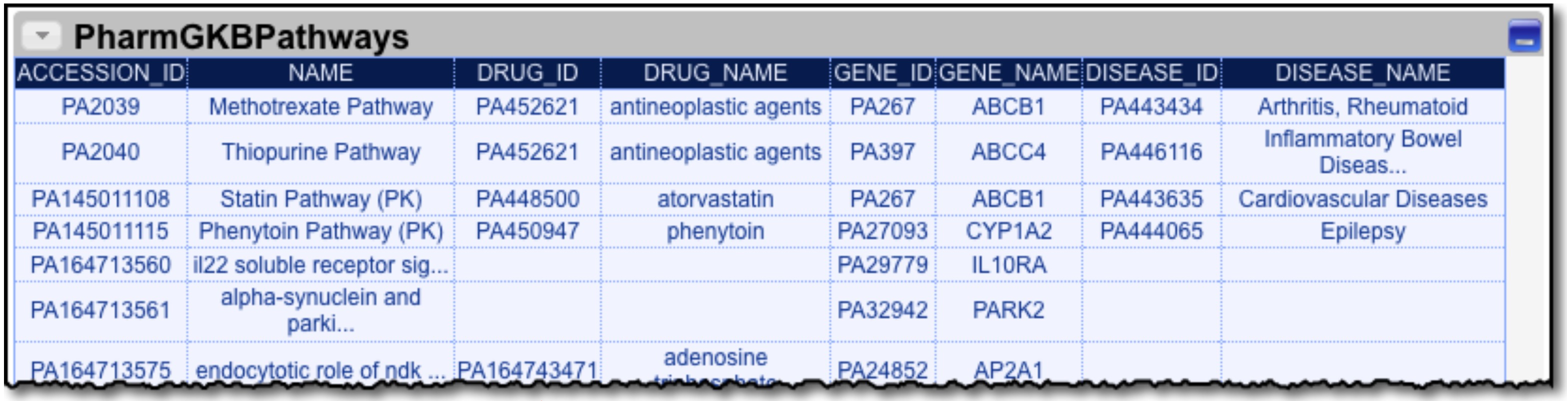
Generate RDF

Learned Semantic Types

RDF

Identify the Semantic Types: Problem Description

Given some columns



PharmGKBPathways

ACCESSION_ID	NAME	DRUG_ID	DRUG_NAME	GENE_ID	GENE_NAME	DISEASE_ID	DISEASE_NAME
PA2039	Methotrexate Pathway	PA452621	antineoplastic agents	PA267	ABCB1	PA443434	Arthritis, Rheumatoid
PA2040	Thiopurine Pathway	PA452621	antineoplastic agents	PA397	ABCC4	PA446116	Inflammatory Bowel Diseas...
PA145011108	Statin Pathway (PK)	PA448500	atorvastatin	PA267	ABCB1	PA443635	Cardiovascular Diseases
PA145011115	Phenytoin Pathway (PK)	PA450947	phenytoin	PA27093	CYP1A2	PA444065	Epilepsy
PA164713560	il22 soluble receptor sig...			PA29779	IL10RA		
PA164713561	alpha-synuclein and parki...			PA32942	PARK2		
PA164713575	endocytotic role of ndk ...	PA164743471	adenosine triphosphate	PA24852	AP2A1		

Predict their Semantic Types

Pathway.PharmGKBPathwayId

Pathway.Label

Drug.PharmGKBGeneId

Drug.Label

Gene.PharmGKBGeneId

Gene.Symbol

Disease.PharmGKBDiseaseId

Disease.Label

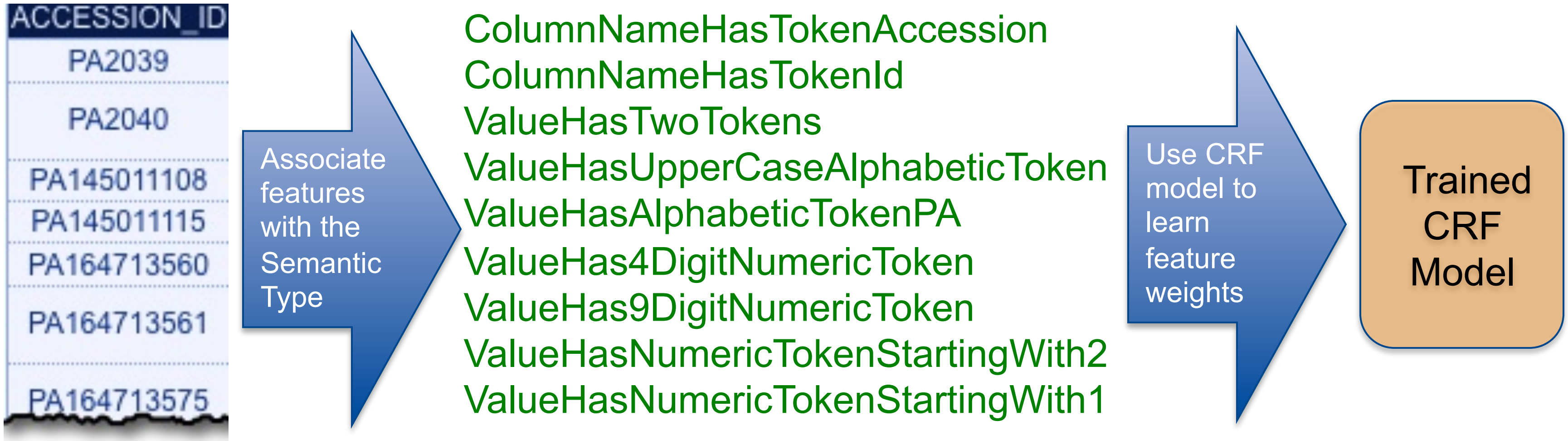
Identify the Semantic Types: Problem Definition

- **Definitions:**
  - — Semantic types: $T = \{t_1, t_2, \ldots\}$
  - — Column of values: $(n, \{v_1, v_2, \ldots\})$
    - *n: column name, v_i : value*
- **Training phase:**
  - — Given labeled columns of data: $\{(n, \{v_1, v_2, \ldots\})^1 \rightarrow t^1, \ldots\}$
  - — Learn a labeling function, $\phi(n, v) \rightarrow \{p_1, p_2, \ldots\}$
    - *Input: column name n, values v*
    - *Output: probability distribution $\{p_1, p_2, \ldots\}$ over semantic types T*
- **Labeling phase:**
  - — Given an unlabeled column: $(n, \{v_1, v_2, \ldots\})$
  - — Generate probability distribution for all values:
    - $\phi(n, v_i)$ -> $\{p_{i,1}, p_{i,2}, p_{i,3}, \ldots\}$
  - — Find probability distribution for column $\{p_{col,1}, p_{col,2}, p_{col,3}, \ldots\}$
    - $\Sigma_i \{p_{i,1}, p_{i,2}, p_{i,3}, \ldots\}$ / N, *where N is number of values in the column*
    - *Output semantic type with highest average likelihood*

Identify the Semantic Types: Training phase

- **Tokenize values in a given labeled column into pure alphabetic, numeric and symbol tokens**
- **Extract features from the tokens and the column name and associate them with column’s semantic type**

Training Column:



ACCESSION_ID
PA2039
PA2040
PA145011108
PA145011115
PA164713560
PA164713561
PA164713575

Associate features with the Semantic Type

ColumnNameHasTokenAccession
ColumnNameHasTokenId
ValueHasTwoTokens
ValueHasUpperCaseAlphabeticToken
ValueHasAlphabeticTokenPA
ValueHas4DigitNumericToken
ValueHas9DigitNumericToken
ValueHasNumericTokenStartingWith2
ValueHasNumericTokenStartingWith1

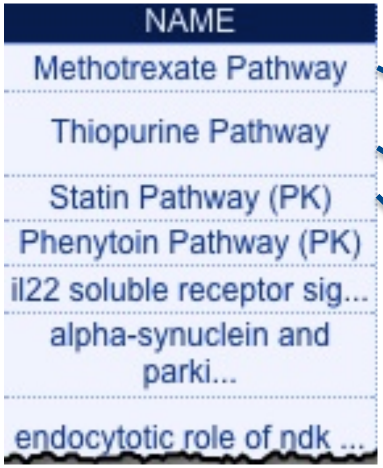
Use CRF model to learn feature weights

Trained CRF Model

Semantic Type:
Pathway.PharmGKBPathwayId

Identify the Semantic Types: Semantic Type Prediction

Given a new column of data, predict the semantic type for each value and them combine them to predict the overall semantic type



NAME

Methotrexate Pathway

Thiopurine Pathway

Statin Pathway (PK)

Phenytoin Pathway (PK)

il22 soluble receptor sig...

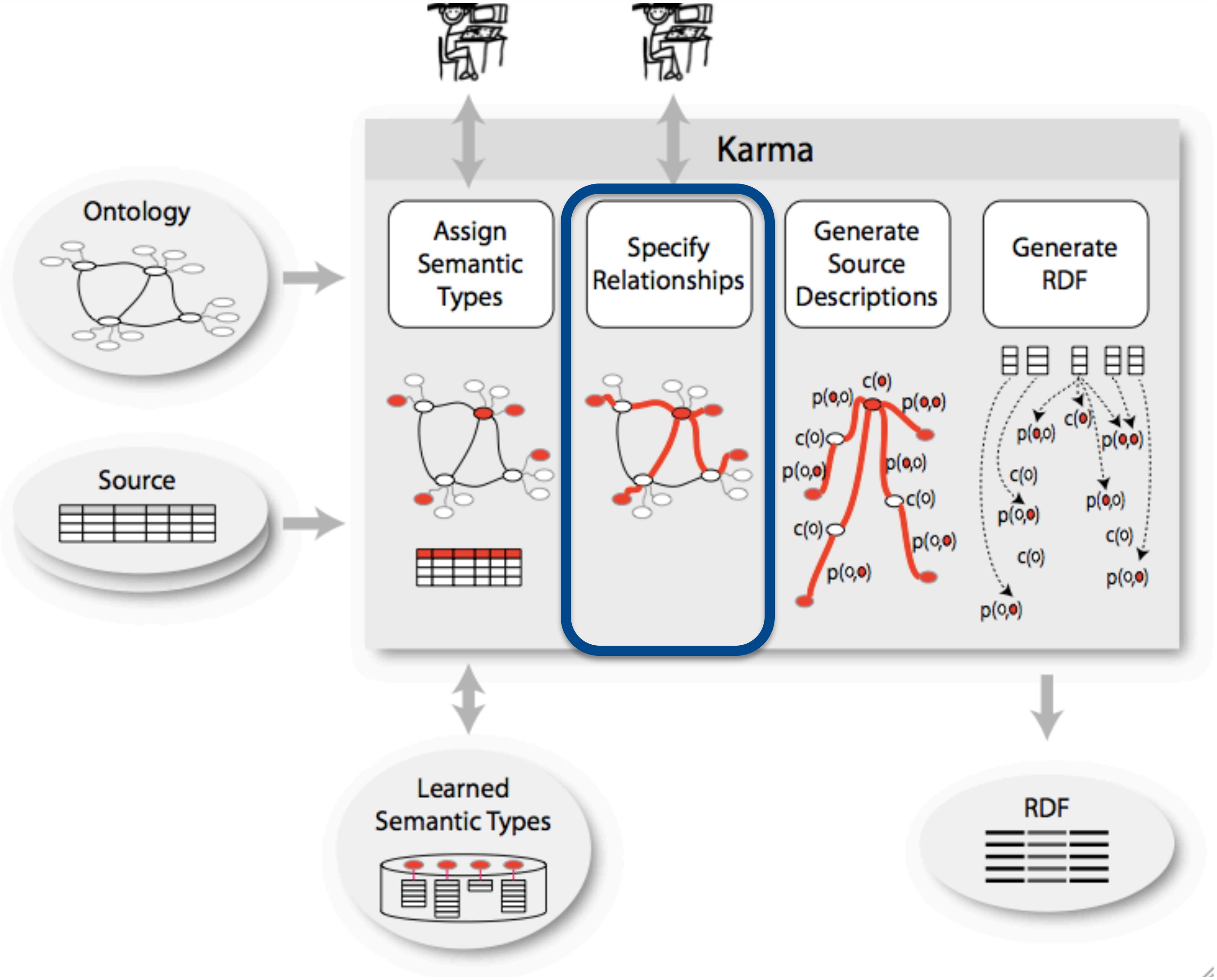
alpha-synuclein and parki...

endocytotic role of ndk ...

	Pathway. Label	Pathway. PharmGKBPathwayId	Gene. Name
	1.0	0.0	0.0
	1.0	0.0	0.0
	0.8	0.0	0.2
		…	
Overall	**0.93**	0.0	0.7

Predicted semantic type for the column

Overall Approach



Karma

Ontology

Source

Assign Semantic Types

Specify Relationships

Generate Source Descriptions

Generate RDF

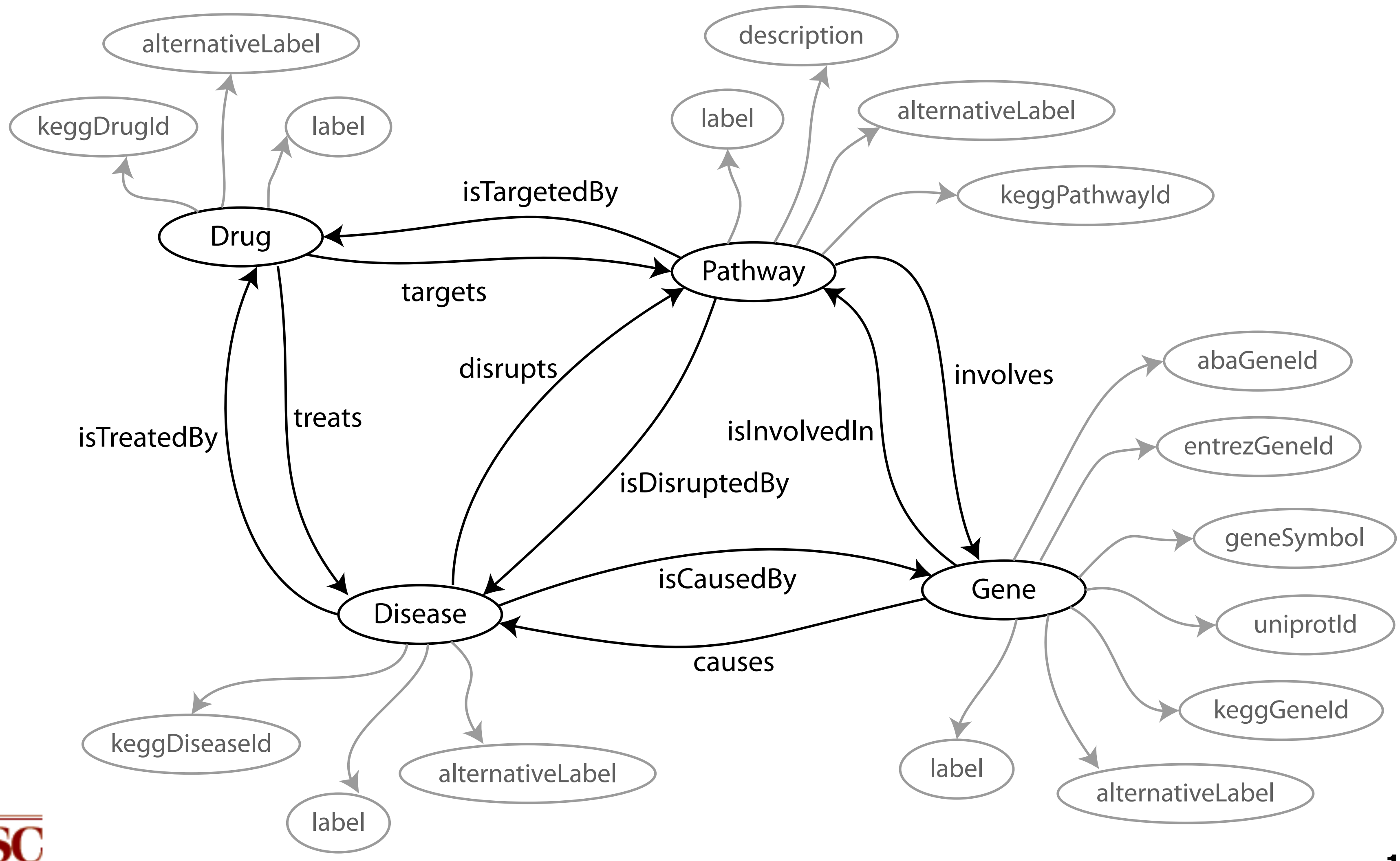
Learned Semantic Types

RDF

USC Viterbi School of Engineering

Build an Ontology Graph

ISI Information Sciences Institute

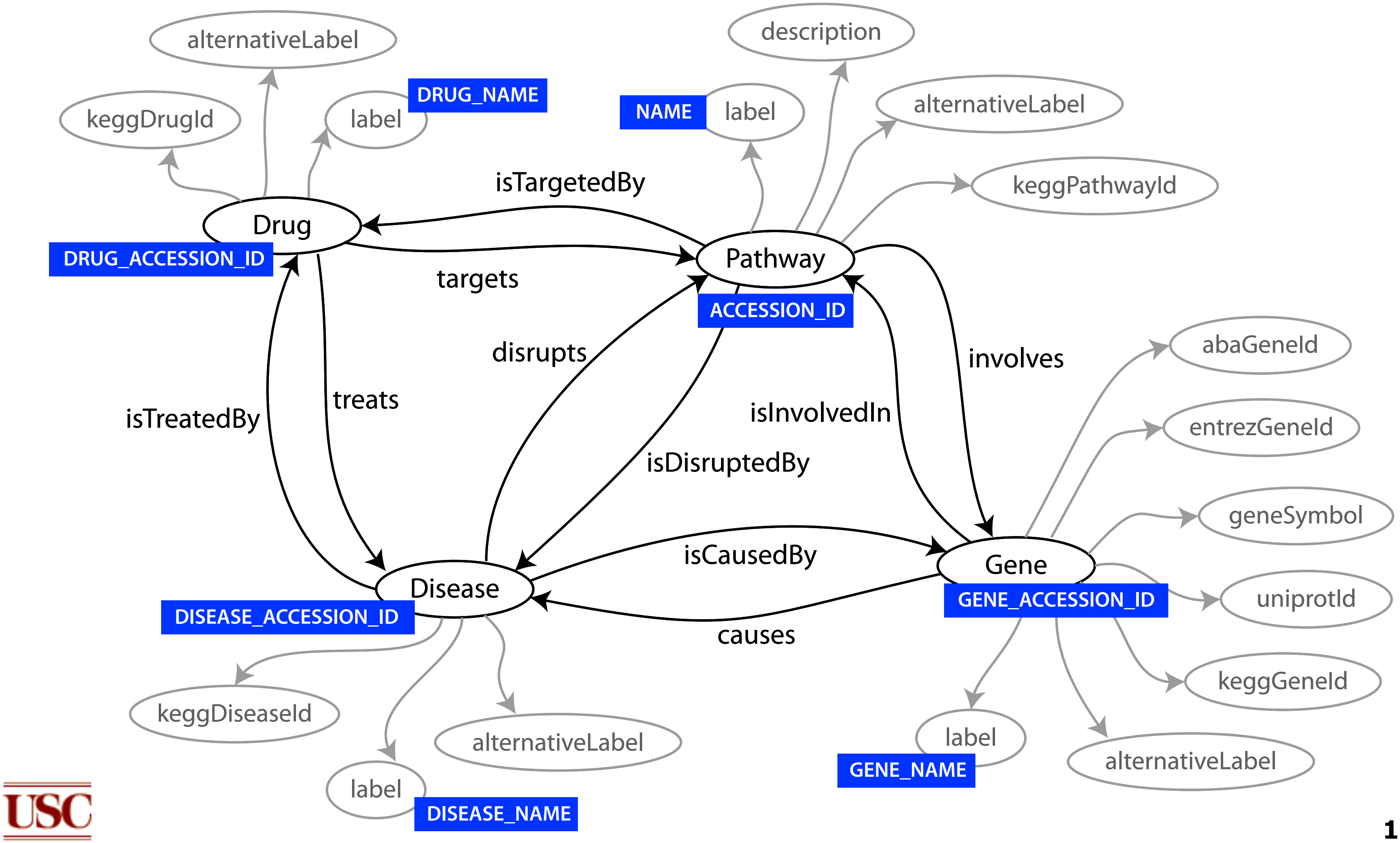


USC

USC Viterbi
School of Engineering

Semantic Types Used to Construct Ontology Graph

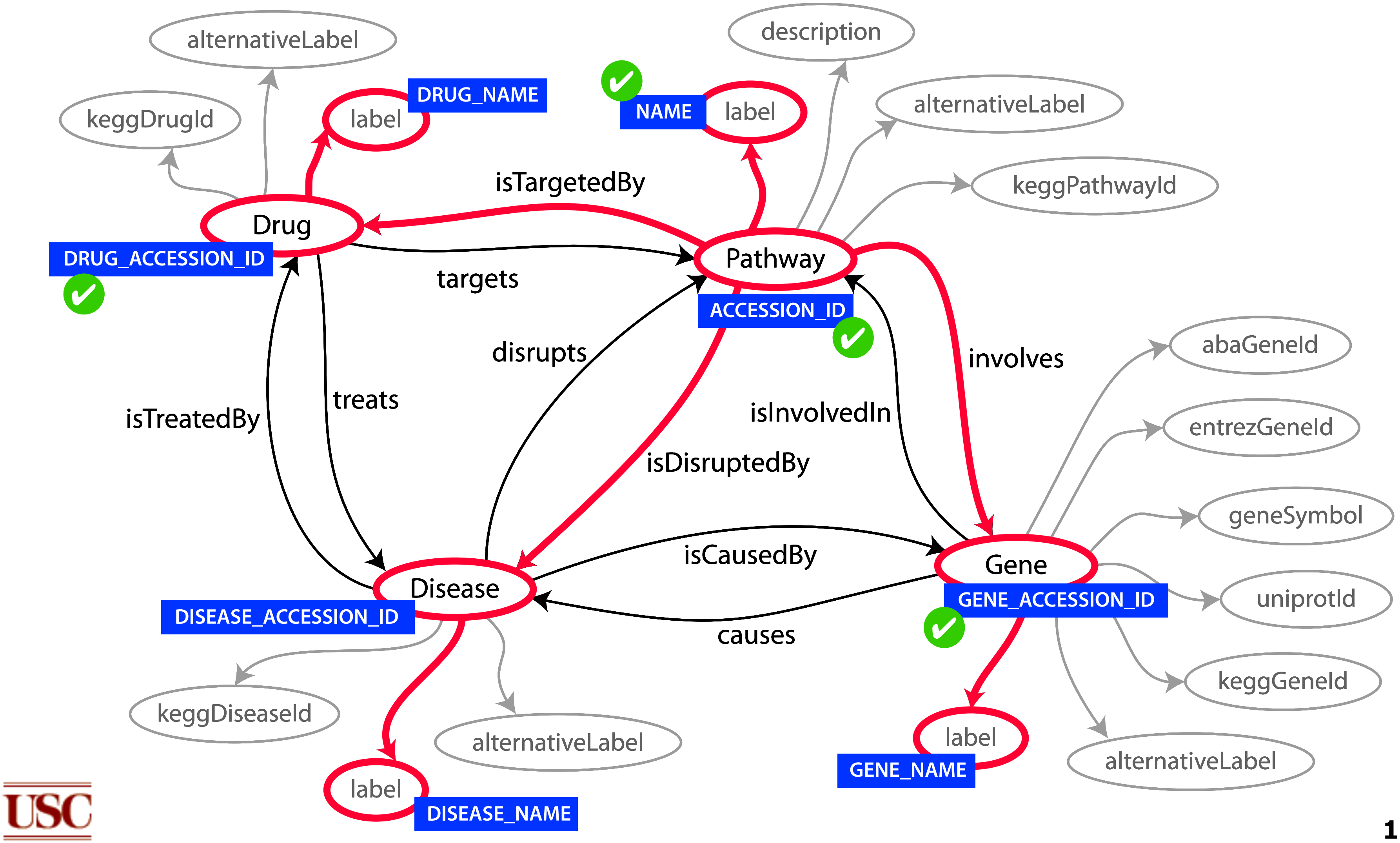
Information Sciences Institute



USC Viterbi
School of Engineering

Infer Relationships Connecting Source Types within Ontology Graph

Information Sciences Institute

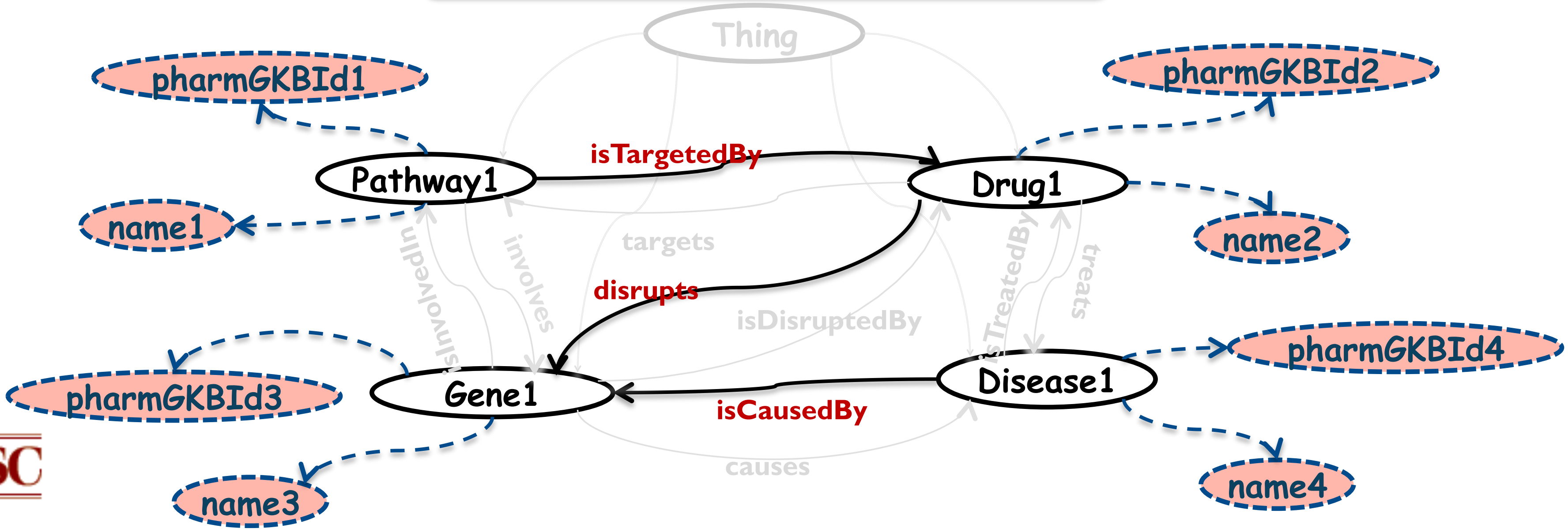


USC

Computing the Minimal Tree

- **Steiner minimal tree (SMT) - General Steiner Tree**
  - $G=(V,E)$, $S \subset V$, $c: E \rightarrow \Re$
  - Shortest network connecting vertices of T
- **Approximation Alg. [Kou & Markowsky, 1981]**
  - $O(|V|^2|S|)$, Approximation Ratio: less than 2

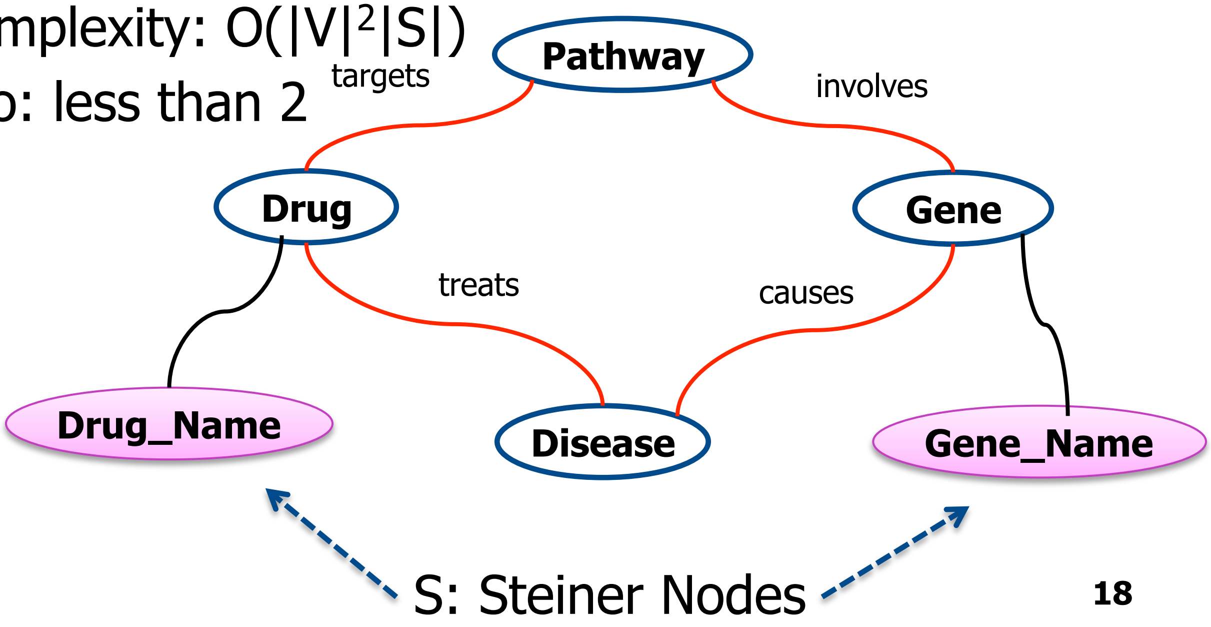
Steiner Nodes: Semantic Types



Inferring the Relationships

- **Search for minimal explanation (source description)**
- **Steiner tree connecting semantic types over ontology graph**
  - Given graph $G=(V,E)$, nodes $S \subset V$, cost $c: E \rightarrow \Re$
  - Find a tree of G that spans S with minimal total cost
  - Unfortunately, NP-complete
- **Approximation Algorithm [KMB, 1981]**
  - Worst-case time complexity: $O(|V|^2|S|)$
  - Approximation Ratio: less than 2

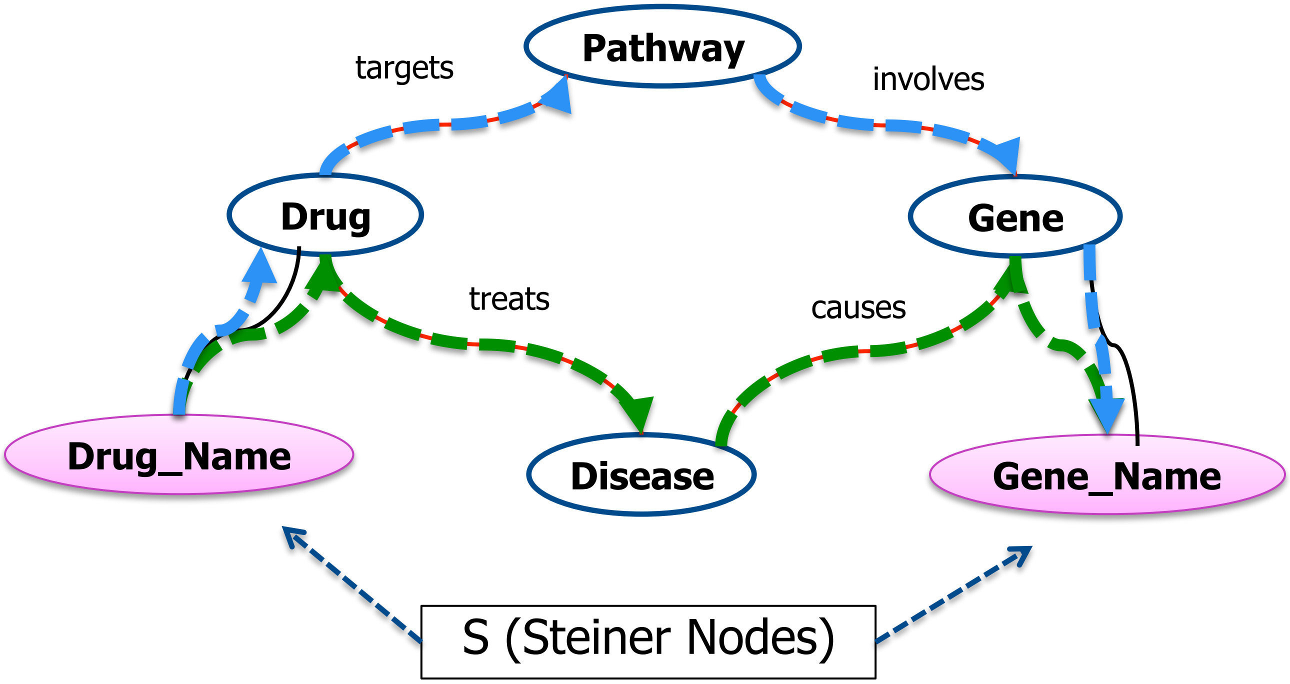
Drug_Name	Gene_Name
Antineoplastic	ABCB1
Antineoplastic	ABCC4
Atorvastatin	ABCB1



Inferring the Relationships

- **Search for minimal explanation (source description)**
- **Multiple explanations:**
  - Drug that targets pathway that involves gene (- - ▶)
  - Drug that treats disease caused by gene (- - ▶)

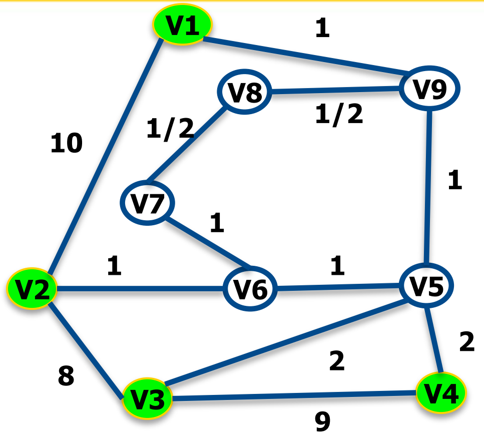
Drug_Name	Gene_Name
Antineoplastic	ABCB1
Antineoplastic	ABCC4
Atorvastatin	ABCB1



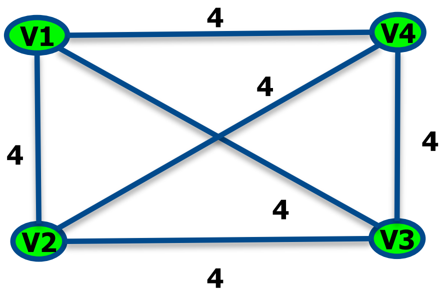
USC Viterbi
School of Engineering

Steiner Tree Algorithm

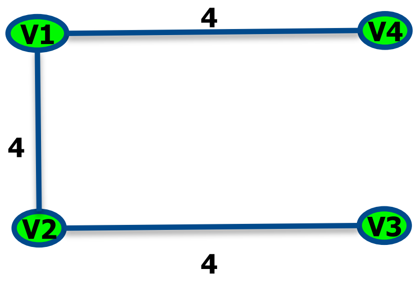
Information Sciences Institute



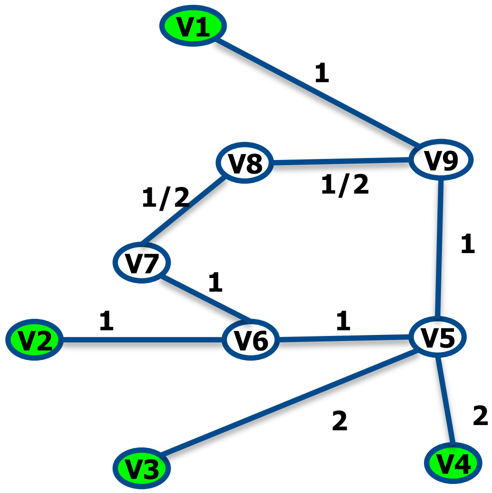
Steiner nodes: {V1, V2, V3, V4}



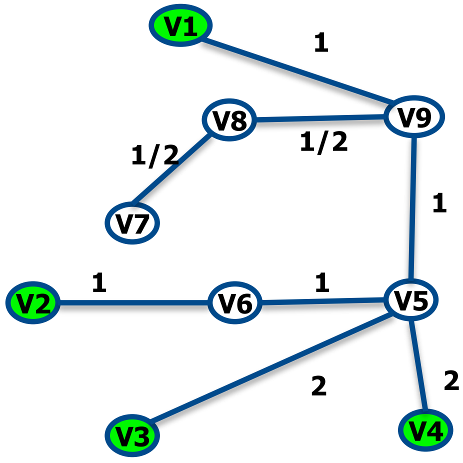
1. construct the complete graph (Nodes: Steiner Nodes, Links Weights: shortest path from each pair in original G)



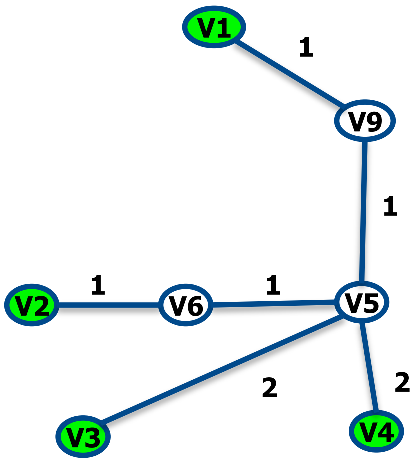
2. Compute MST



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



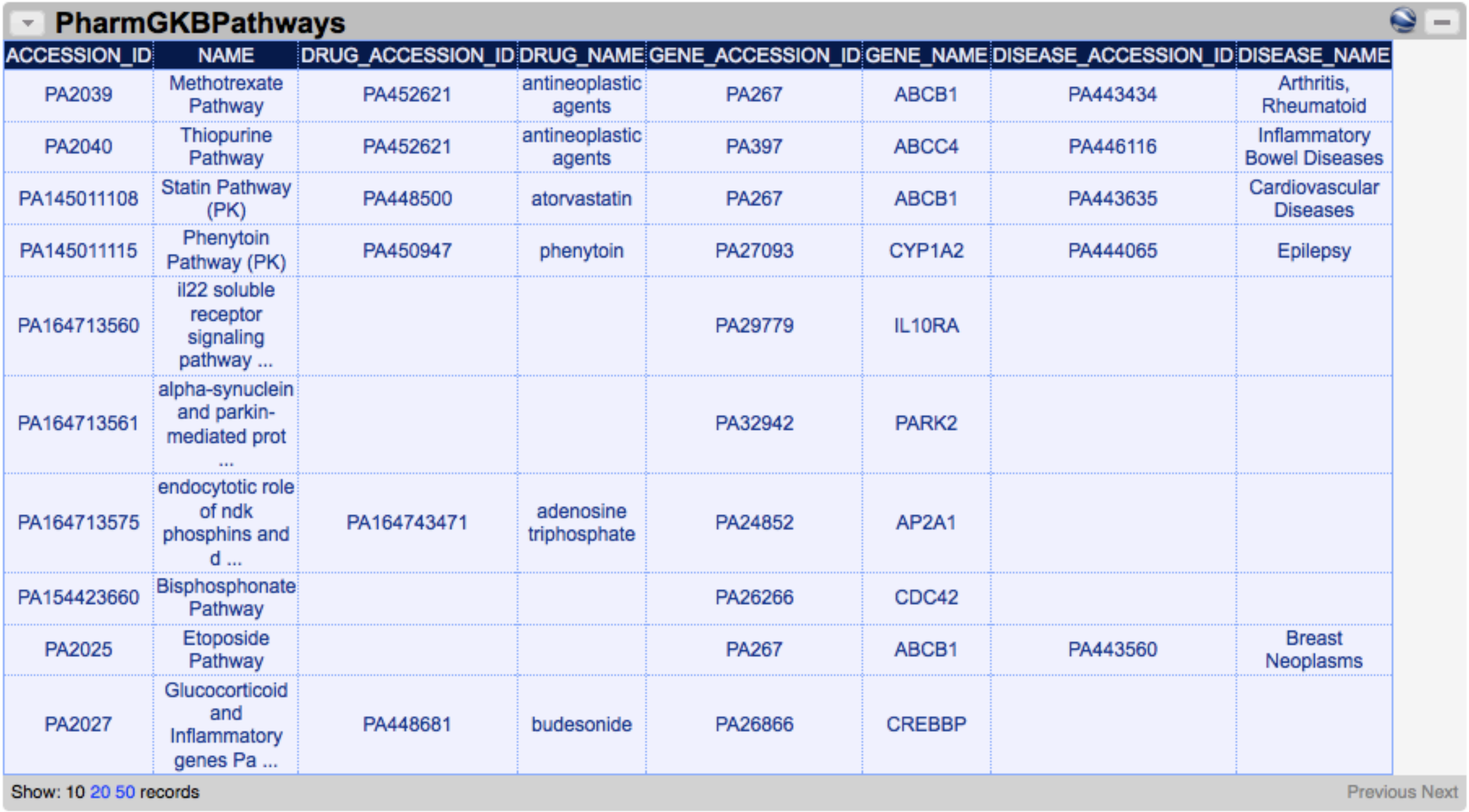
4. Compute MST



5. remove extra links until all leaves are Steiner nodes

USC

Source Data



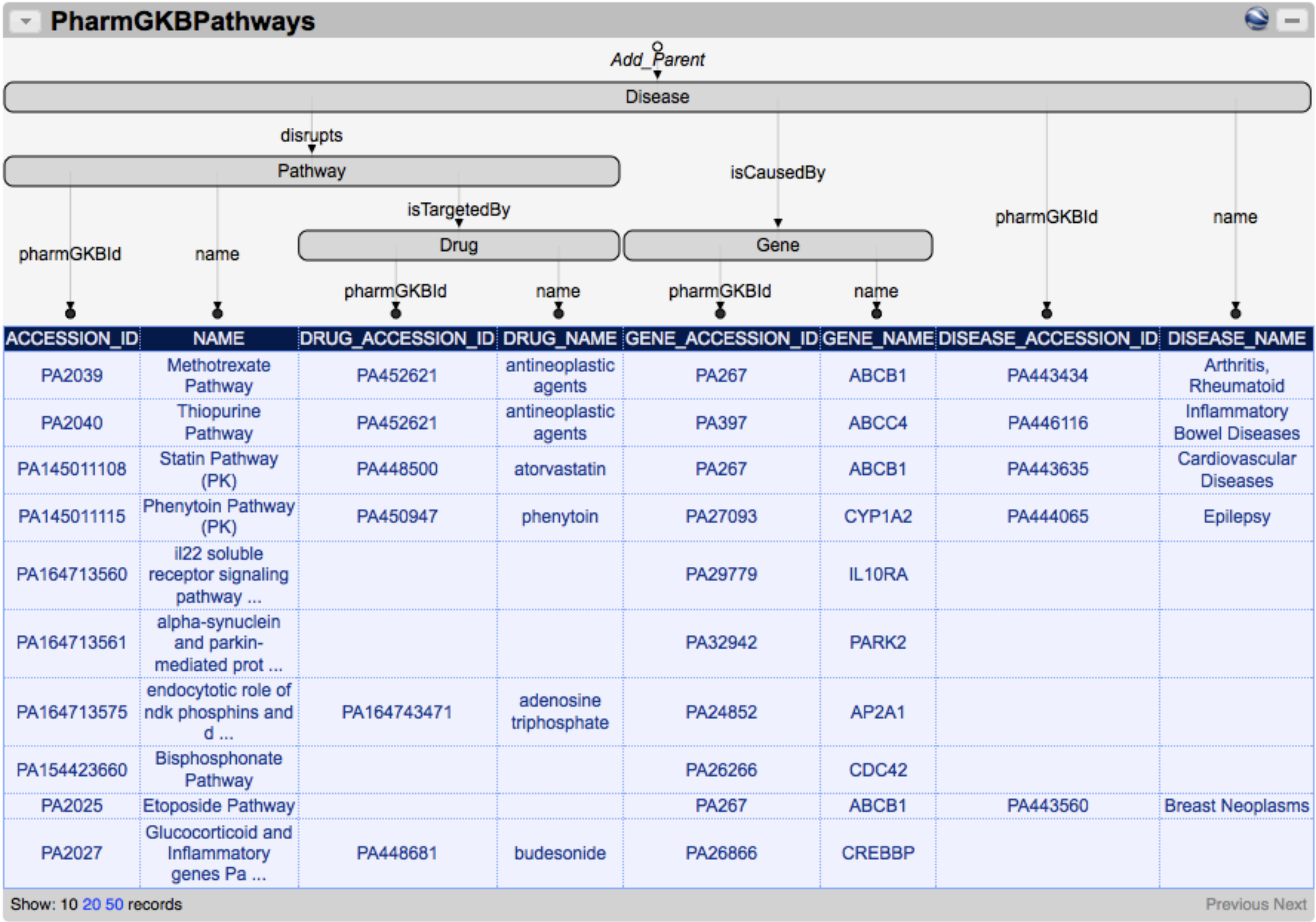
PharmGKBPathways

ACCESSION_ID	NAME	DRUG_ACCESSION_ID	DRUG_NAME	GENE_ACCESSION_ID	GENE_NAME	DISEASE_ACCESSION_ID	DISEASE_NAME
PA2039	Methotrexate Pathway	PA452621	antineoplastic agents	PA267	ABCB1	PA443434	Arthritis, Rheumatoid
PA2040	Thiopurine Pathway	PA452621	antineoplastic agents	PA397	ABCC4	PA446116	Inflammatory Bowel Diseases
PA145011108	Statin Pathway (PK)	PA448500	atorvastatin	PA267	ABCB1	PA443635	Cardiovascular Diseases
PA145011115	Phenytoin Pathway (PK)	PA450947	phenytoin	PA27093	CYP1A2	PA444065	Epilepsy
PA164713560	il22 soluble receptor signaling pathway ...			PA29779	IL10RA		
PA164713561	alpha-synuclein and parkin-mediated prot ...			PA32942	PARK2		
PA164713575	endocytotic role of ndk phosphins and d ...	PA164743471	adenosine triphosphate	PA24852	AP2A1		
PA154423660	Bisphosphonate Pathway			PA26266	CDC42		
PA2025	Etoposide Pathway			PA267	ABCB1	PA443560	Breast Neoplasms
PA2027	Glucocorticoid and Inflammatory genes Pa ...	PA448681	budesonide	PA26866	CREBBP		

Show: 10 20 50 records

Previous Next

Proposed Semantic model



ACCESSION_ID	NAME	DRUG_ACCESSION_ID	DRUG_NAME	GENE_ACCESSION_ID	GENE_NAME	DISEASE_ACCESSION_ID	DISEASE_NAME
PA2039	Methotrexate Pathway	PA452621	antineoplastic agents	PA267	ABCB1	PA443434	Arthritis, Rheumatoid
PA2040	Thiopurine Pathway	PA452621	antineoplastic agents	PA397	ABCC4	PA446116	Inflammatory Bowel Diseases
PA145011108	Statin Pathway (PK)	PA448500	atorvastatin	PA267	ABCB1	PA443635	Cardiovascular Diseases
PA145011115	Phenytoin Pathway (PK)	PA450947	phenytoin	PA27093	CYP1A2	PA444065	Epilepsy
PA164713560	il22 soluble receptor signaling pathway ...			PA29779	IL10RA		
PA164713561	alpha-synuclein and parkin-mediated prot ...			PA32942	PARK2		
PA164713575	endocytotic role of ndk phosphins and d ...	PA164743471	adenosine triphosphate	PA24852	AP2A1		
PA154423660	Bisphosphonate Pathway			PA26266	CDC42		
PA2025	Etoposide Pathway			PA267	ABCB1	PA443560	Breast Neoplasms
PA2027	Glucocorticoid and Inflammatory genes Pa ...	PA448681	budesonide	PA26866	CREBBP		

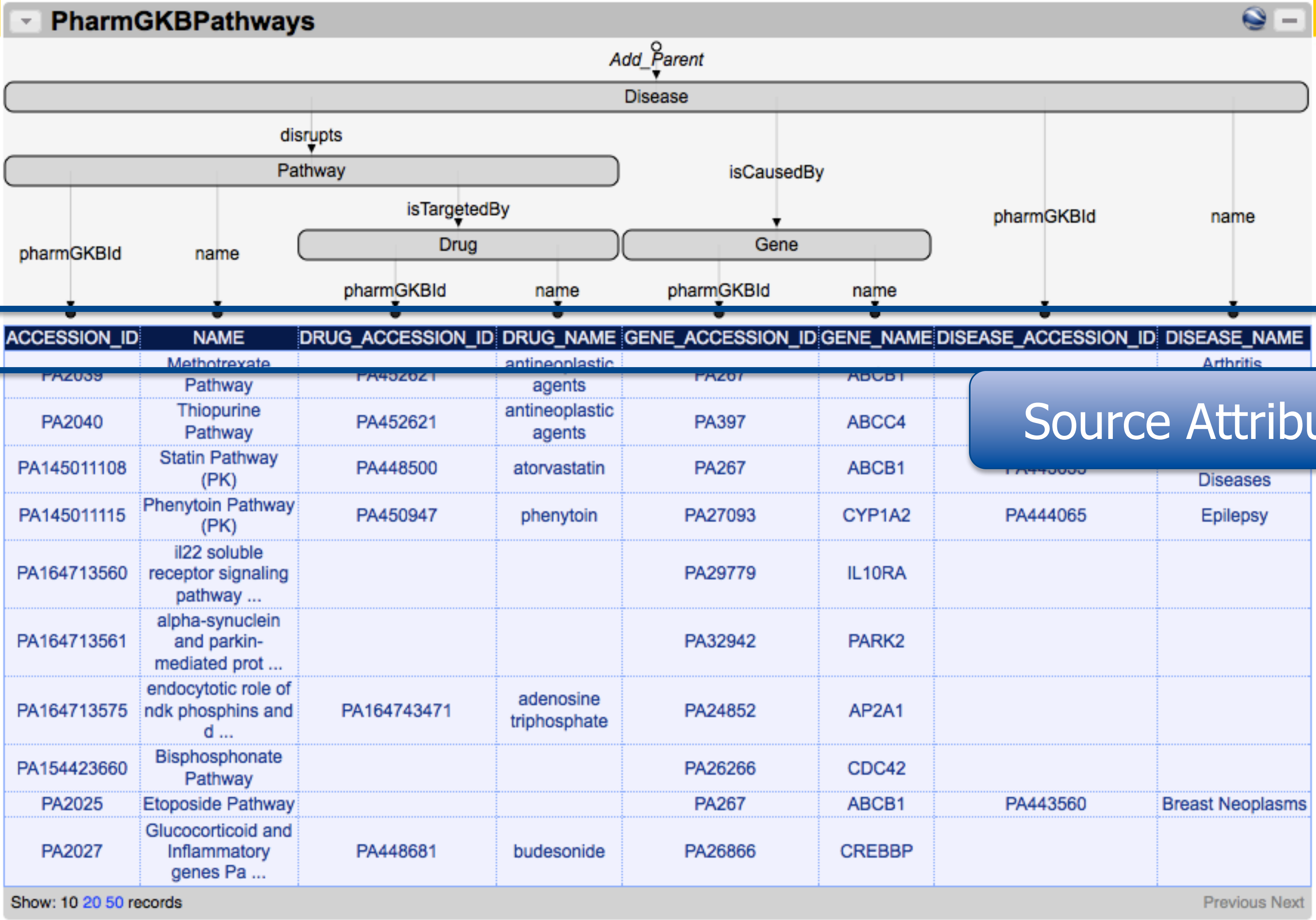
Show: 10 20 50 records

Previous Next

USC Viterbi
School of Engineering

Proposed Semantic model

Information Sciences Institute



ACCESSION_ID	NAME	DRUG_ACCESSION_ID	DRUG_NAME	GENE_ACCESSION_ID	GENE_NAME	DISEASE_ACCESSION_ID	DISEASE_NAME
PA2039	Methotrexate Pathway	PA452621	antineoplastic agents	PA267	ABCB1	[illegible]	Arthritis
PA2040	Thiopurine Pathway	PA452621	antineoplastic agents	PA397	ABCC4	[illegible]	[illegible]
PA145011108	Statin Pathway (PK)	PA448500	atorvastatin	PA267	ABCB1	[illegible]	[illegible] Diseases
PA145011115	Phenytoin Pathway (PK)	PA450947	phenytoin	PA27093	CYP1A2	PA444065	Epilepsy
PA164713560	il22 soluble receptor signaling pathway ...			PA29779	IL10RA		
PA164713561	alpha-synuclein and parkin-mediated prot ...			PA32942	PARK2		
PA164713575	endocytotic role of ndk phosphins and d ...	PA164743471	adenosine triphosphate	PA24852	AP2A1		
PA154423660	Bisphosphonate Pathway			PA26266	CDC42		
PA2025	Etoposide Pathway			PA267	ABCB1	PA443560	Breast Neoplasms
PA2027	Glucocorticoid and Inflammatory genes Pa ...	PA448681	budesonide	PA26866	CREBBP		

Show: 10 20 50 records

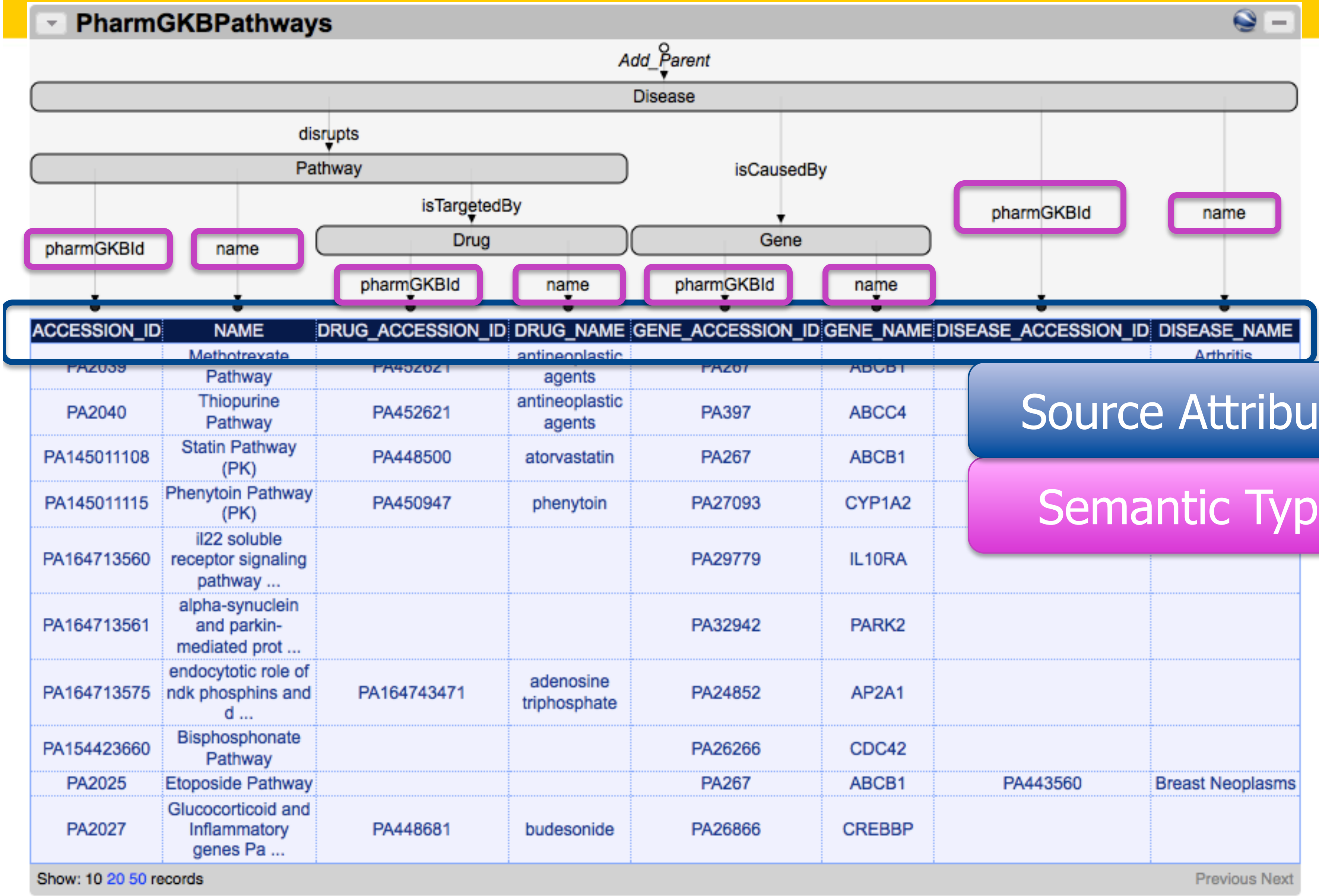
Previous Next

Source Attributes

USC Viterbi School of Engineering

Proposed Semantic model

ISI Information Sciences Institute



ACCESSION_ID	NAME	DRUG_ACCESSION_ID	DRUG_NAME	GENE_ACCESSION_ID	GENE_NAME	DISEASE_ACCESSION_ID	DISEASE_NAME
PA2039	Methotrexate Pathway	PA452621	antineoplastic agents	PA267	ABCB1		Arthritis
PA2040	Thiopurine Pathway	PA452621	antineoplastic agents	PA397	ABCC4		
PA145011108	Statin Pathway (PK)	PA448500	atorvastatin	PA267	ABCB1		
PA145011115	Phenytoin Pathway (PK)	PA450947	phenytoin	PA27093	CYP1A2		
PA164713560	il22 soluble receptor signaling pathway ...			PA29779	IL10RA		
PA164713561	alpha-synuclein and parkin-mediated prot ...			PA32942	PARK2		
PA164713575	endocytotic role of ndk phosphins and d ...	PA164743471	adenosine triphosphate	PA24852	AP2A1		
PA154423660	Bisphosphonate Pathway			PA26266	CDC42		
PA2025	Etoposide Pathway			PA267	ABCB1	PA443560	Breast Neoplasms
PA2027	Glucocorticoid and Inflammatory genes Pa ...	PA448681	budesonide	PA26866	CREBBP		

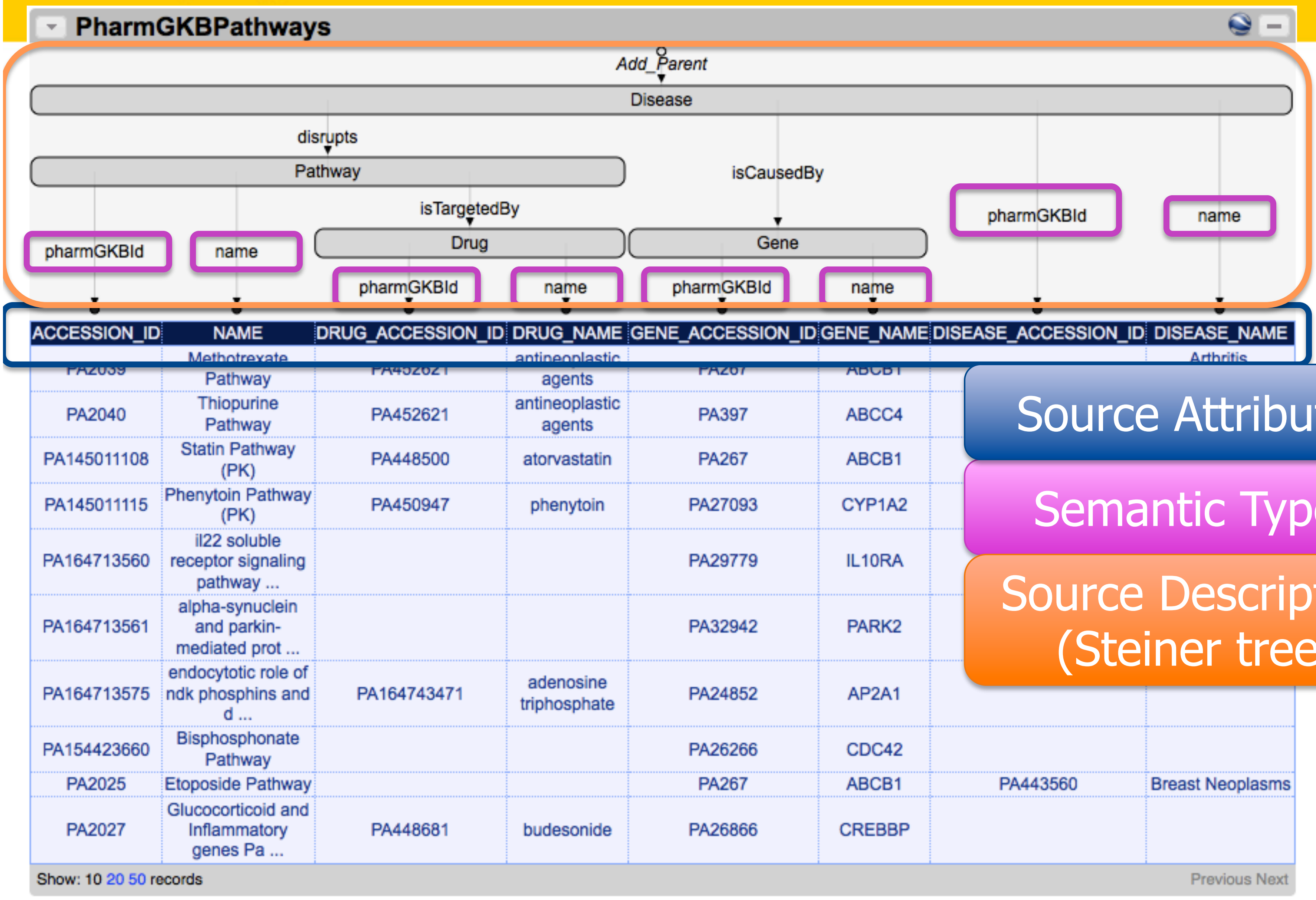
Show: 10 20 50 records

Previous Next

Source Attributes

Semantic Types

Proposed Semantic model



ACCESSION_ID	NAME	DRUG_ACCESSION_ID	DRUG_NAME	GENE_ACCESSION_ID	GENE_NAME	DISEASE_ACCESSION_ID	DISEASE_NAME
PA2039	Methotrexate Pathway	PA452621	antineoplastic agents	PA267	ABCB1		Arthritis
PA2040	Thiopurine Pathway	PA452621	antineoplastic agents	PA397	ABCC4		
PA145011108	Statin Pathway (PK)	PA448500	atorvastatin	PA267	ABCB1		
PA145011115	Phenytoin Pathway (PK)	PA450947	phenytoin	PA27093	CYP1A2		
PA164713560	il22 soluble receptor signaling pathway ...			PA29779	IL10RA		
PA164713561	alpha-synuclein and parkin-mediated prot ...			PA32942	PARK2		
PA164713575	endocytotic role of ndk phosphins and d ...	PA164743471	adenosine triphosphate	PA24852	AP2A1		
PA154423660	Bisphosphonate Pathway			PA26266	CDC42		
PA2025	Etoposide Pathway			PA267	ABCB1	PA443560	Breast Neoplasms
PA2027	Glucocorticoid and Inflammatory genes Pa ...	PA448681	budesonide	PA26866	CREBBP		

Show: 10 20 50 records

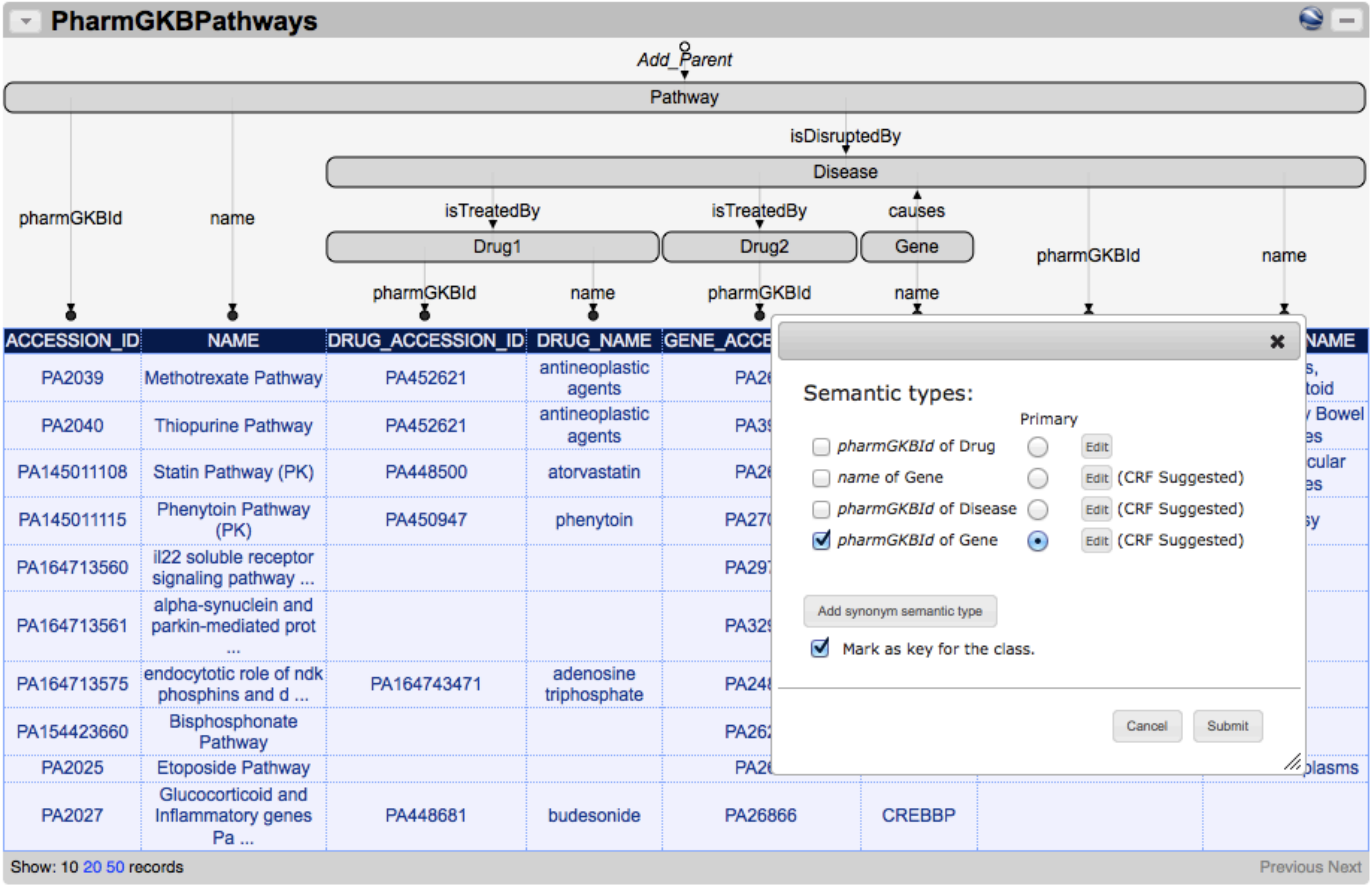
Previous Next

Source Attributes

Semantic Types

Source Description (Steiner tree)

User Correction of Semantic Type



ACCESSION_ID	NAME	DRUG_ACCESSION_ID	DRUG_NAME	GENE_ACCE			NAME
PA2039	Methotrexate Pathway	PA452621	antineoplastic agents	PA26			s, toid
PA2040	Thiopurine Pathway	PA452621	antineoplastic agents	PA39			/ Bowel es
PA145011108	Statin Pathway (PK)	PA448500	atorvastatin	PA26			cular es
PA145011115	Phenytoin Pathway (PK)	PA450947	phenytoin	PA270			sy
PA164713560	il22 soluble receptor signaling pathway ...			PA297			
PA164713561	alpha-synuclein and parkin-mediated prot ...			PA329			
PA164713575	endocytotic role of ndk phosphins and d ...	PA164743471	adenosine triphosphate	PA248			
PA154423660	Bisphosphonate Pathway			PA262			
PA2025	Etoposide Pathway			PA26			plasms
PA2027	Glucocorticoid and Inflammatory genes Pa ...	PA448681	budesonide	PA26866	CREBBP		

Show: 10 20 50 records

Previous Next

✖

Semantic types:

	Primary		
□ *pharmGKBId* of Drug	○	Edit	
□ *name* of Gene	○	Edit	(CRF Suggested)
□ *pharmGKBId* of Disease	○	Edit	(CRF Suggested)
☑ *pharmGKBId* of Gene	●	Edit	(CRF Suggested)

Add synonym semantic type

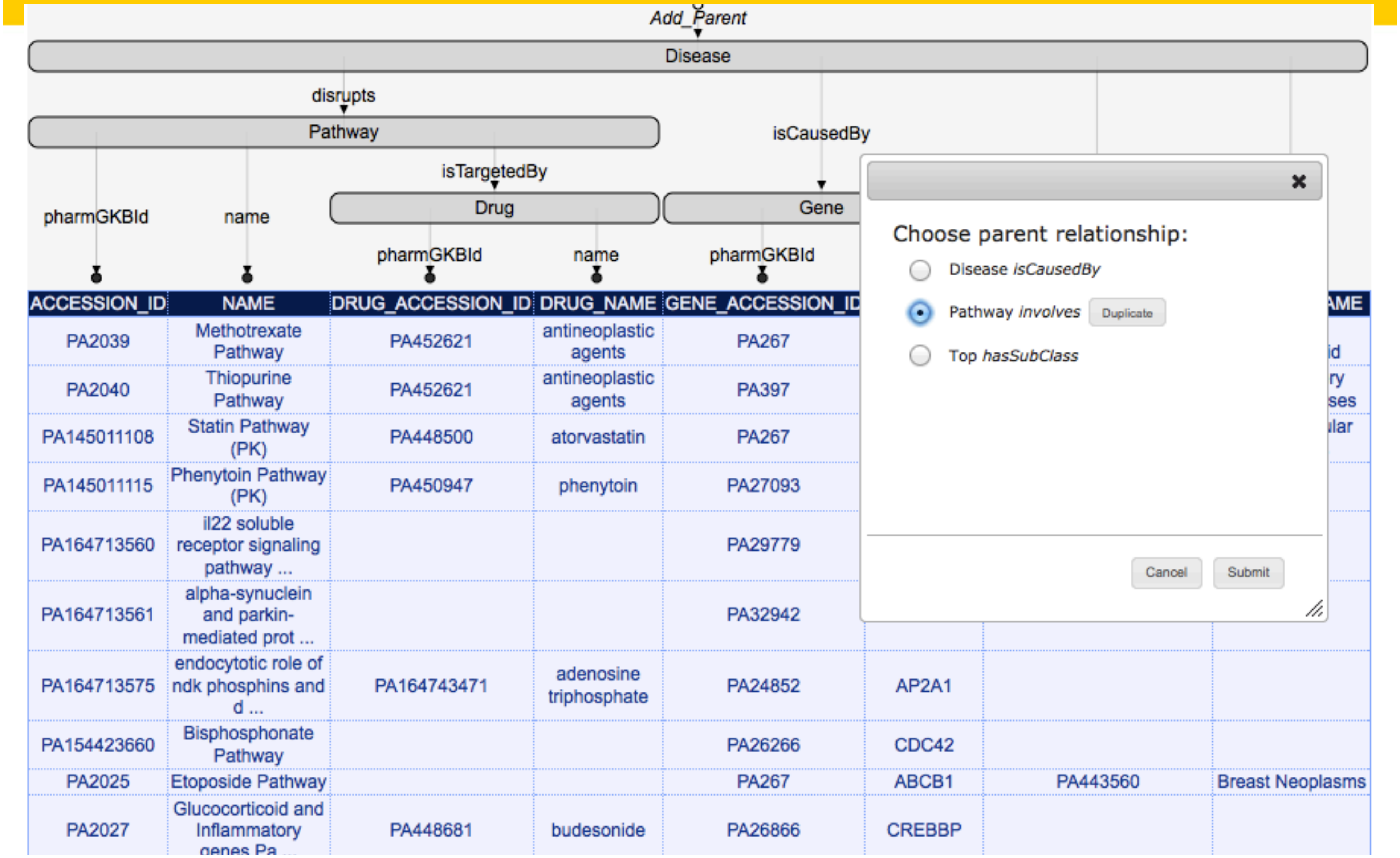
☑ Mark as key for the class.

Cancel Submit

USC Viterbi
School of Engineering

User Corrects Model

ISI
Information Sciences Institute



Choose parent relationship:

○ Disease *isCausedBy*

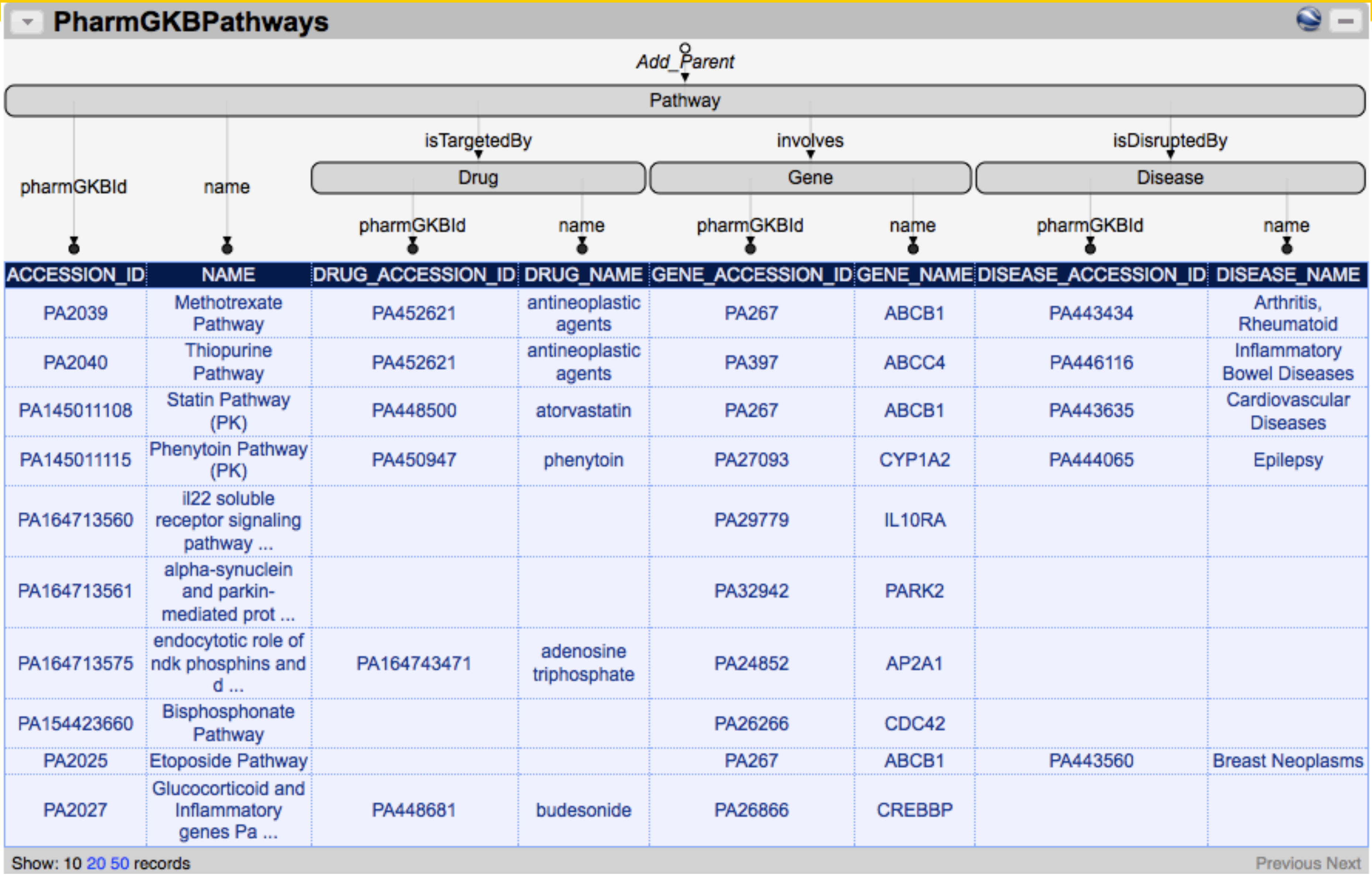
● Pathway *involves* Duplicate

○ Top *hasSubClass*

Cancel Submit

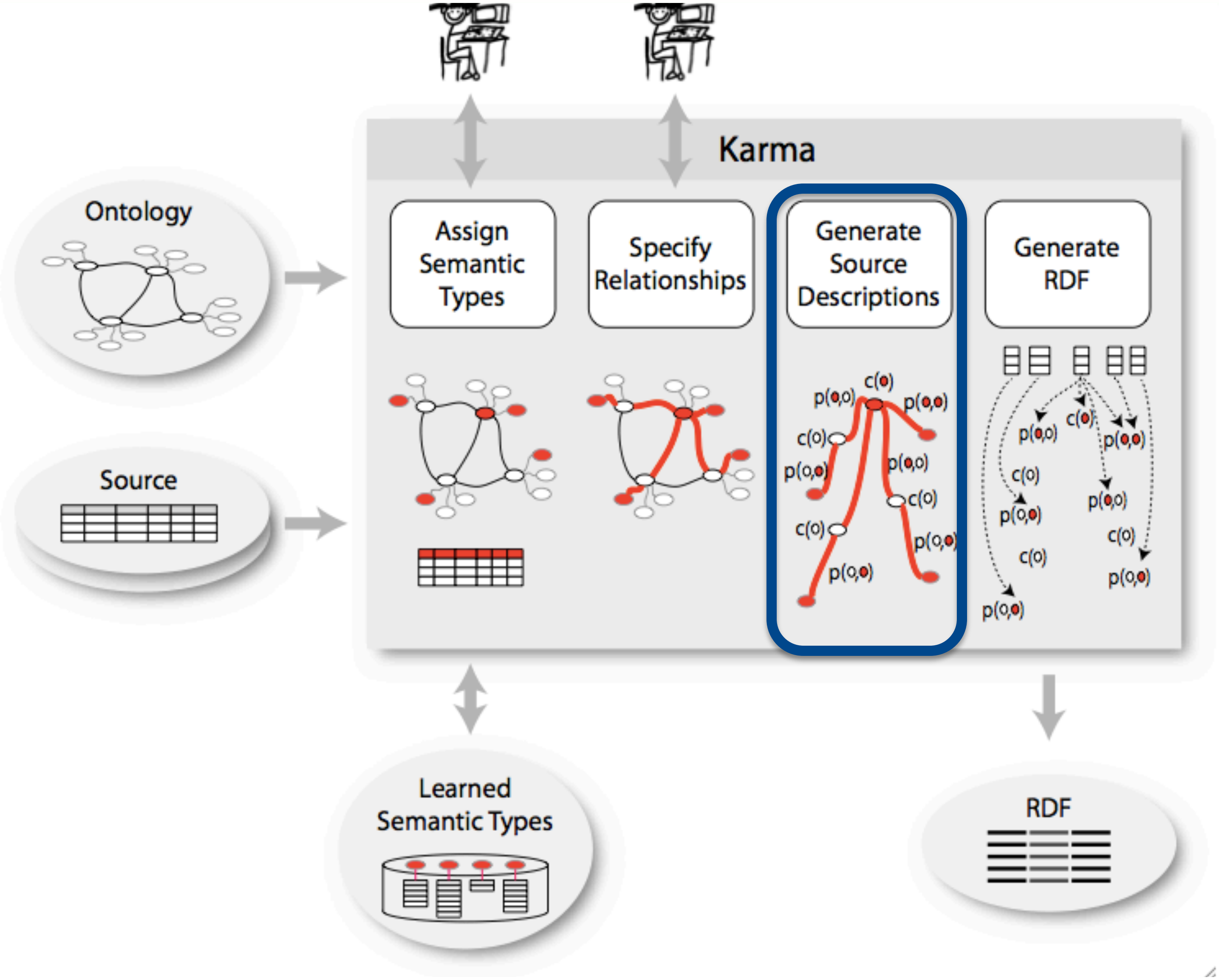
ACCESSION_ID	NAME	DRUG_ACCESSION_ID	DRUG_NAME	GENE_ACCESSION_ID			AME
PA2039	Methotrexate Pathway	PA452621	antineoplastic agents	PA267			id
PA2040	Thiopurine Pathway	PA452621	antineoplastic agents	PA397			ry ses
PA145011108	Statin Pathway (PK)	PA448500	atorvastatin	PA267			lar
PA145011115	Phenytoin Pathway (PK)	PA450947	phenytoin	PA27093			
PA164713560	il22 soluble receptor signaling pathway ...			PA29779			
PA164713561	alpha-synuclein and parkin-mediated prot ...			PA32942			
PA164713575	endocytotic role of ndk phosphins and d ...	PA164743471	adenosine triphosphate	PA24852	AP2A1		
PA154423660	Bisphosphonate Pathway			PA26266	CDC42		
PA2025	Etoposide Pathway			PA267	ABCB1	PA443560	Breast Neoplasms
PA2027	Glucocorticoid and Inflammatory genes Pa ...	PA448681	budesonide	PA26866	CREBBP		

Final source model



ACCESSION_ID	NAME	DRUG_ACCESSION_ID	DRUG_NAME	GENE_ACCESSION_ID	GENE_NAME	DISEASE_ACCESSION_ID	DISEASE_NAME
PA2039	Methotrexate Pathway	PA452621	antineoplastic agents	PA267	ABCB1	PA443434	Arthritis, Rheumatoid
PA2040	Thiopurine Pathway	PA452621	antineoplastic agents	PA397	ABCC4	PA446116	Inflammatory Bowel Diseases
PA145011108	Statin Pathway (PK)	PA448500	atorvastatin	PA267	ABCB1	PA443635	Cardiovascular Diseases
PA145011115	Phenytoin Pathway (PK)	PA450947	phenytoin	PA27093	CYP1A2	PA444065	Epilepsy
PA164713560	il22 soluble receptor signaling pathway ...			PA29779	IL10RA		
PA164713561	alpha-synuclein and parkin-mediated prot ...			PA32942	PARK2		
PA164713575	endocytotic role of ndk phosphins and d ...	PA164743471	adenosine triphosphate	PA24852	AP2A1		
PA154423660	Bisphosphonate Pathway			PA26266	CDC42		
PA2025	Etoposide Pathway			PA267	ABCB1	PA443560	Breast Neoplasms
PA2027	Glucocorticoid and Inflammatory genes Pa ...	PA448681	budesonide	PA26866	CREBBP		

Approach: Generate Logical Schema Mappings



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)

Uses of Source Descriptions

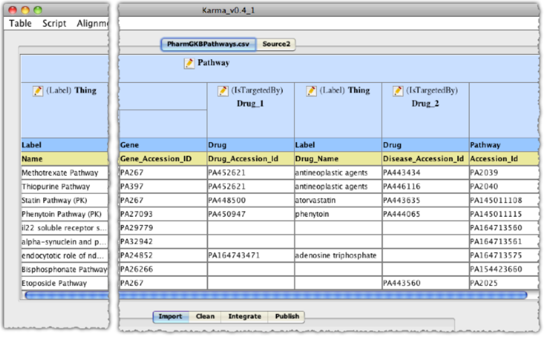
- **Data integration (e.g., [Halevy 2001])**
  - — Answer queries under GLAV rules in mediator
- **Data exchange (e.g., [Arenas et al, 2010])**
  - — Materialize RDF using the GLAV rules

USC Viterbi School of Engineering

Generation of the Source Descriptions

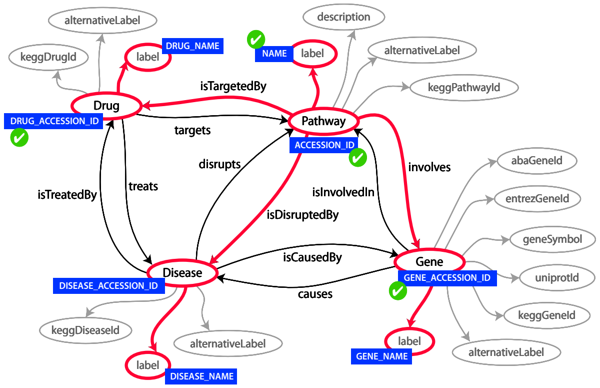
ISI Information Sciences Institute

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



	Pathway				
(Label) Thing		(IsTargetedBy) Drug_1	(Label) Thing	(IsTargetedBy) Drug_2	
Label	Gene	Drug	Label	Drug	Pathway
Name	Gene_Accession_ID	Drug_Accession_Id	Drug_Name	Disease_Accession_Id	Accession_Id
Methotrexate Pathway	PA267	PA452621	antineoplastic agents	PA443434	PA2039
Thiopurine Pathway	PA397	PA452621	antineoplastic agents	PA446116	PA2040
Statin Pathway (PK)	PA267	PA448500	atorvastatin	PA443635	PA145011108
Phenytoin Pathway (PK)	PA27093	PA450947	phenytoin	PA444065	PA145011115
il22 soluble receptor s...	PA29779				PA164713560
alpha-synuclein and p...	PA32942				PA164713561
endocytotic role of nd...	PA24852	PA164743471	adenosine triphosphate		PA164713575
Bisphosphonate Pathway	PA26266				PA154423660
Etoposide Pathway	PA267			PA443560	PA2025

+



=

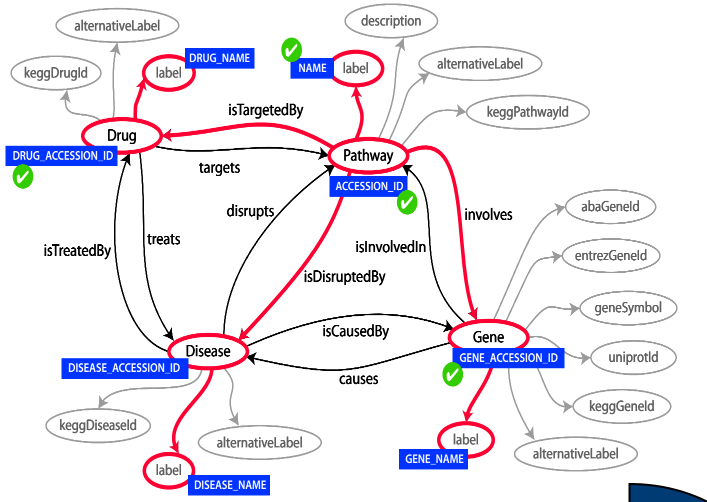
$$
\begin{aligned}
&\text{PharmGKBPathways}(\textsc{Name},\textsc{Accession\_Id},\ \textsc{Gene\_Accession\_ID},\ \textsc{Disease\_Name},\\
&\quad \textsc{Gene\_Name},\textsc{Disease\_Accession\_Id},\textsc{Drug\_Name},\textsc{Drug\_Accession\_Id}) \rightarrow\\
&\text{Pathway}(\textbf{uri}(\textsc{Accession\_Id}))\ \hat{}\ \text{label}(\textbf{uri}(\textsc{Accession\_Id}),\ \textsc{Name})\ \hat{}\\
&\text{involves}(\textbf{uri}(\textsc{Accession\_Id}),\ \textbf{uri}(\textsc{Gene\_Accession\_ID}))\ \hat{}\\
&\text{isTargetedBy}(\textbf{uri}(\textsc{Accession\_Id}),\ \textbf{uri}(\textsc{Drug\_Accession\_Id}))\ \hat{}\\
&\text{isDisruptedBy}(\textbf{uri}(\textsc{Accession\_Id}),\ \textbf{uri}(\textsc{Disease\_Accession\_Id}))\ \hat{}\\
&\text{Gene}(\textbf{uri}(\textsc{Gene\_Accession\_ID}))\ \hat{}\ \text{label}(\textbf{uri}(\textsc{Gene\_Accession\_ID}),\ \textsc{Gene\_Name})\ \hat{}\\
&\text{Drug}(\textbf{uri}(\textsc{Drug\_Accession\_Id}))\ \hat{}\ \text{label}(\textbf{uri}(\textsc{Drug\_Accession\_Id}),\ \textsc{Drug\_Name})\ \hat{}\\
&\text{Disease}(\textbf{uri}(\textsc{Disease\_Accession\_Id}))\ \hat{}\\
&\text{label}(\textbf{uri}(\textsc{Disease\_Accession\_Id}),\ \textsc{Disease\_Name})
\end{aligned}
$$

Generation of the Source Descriptions: rule consequent

Node → Class (unary predicate)

Edge → binary predicate

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

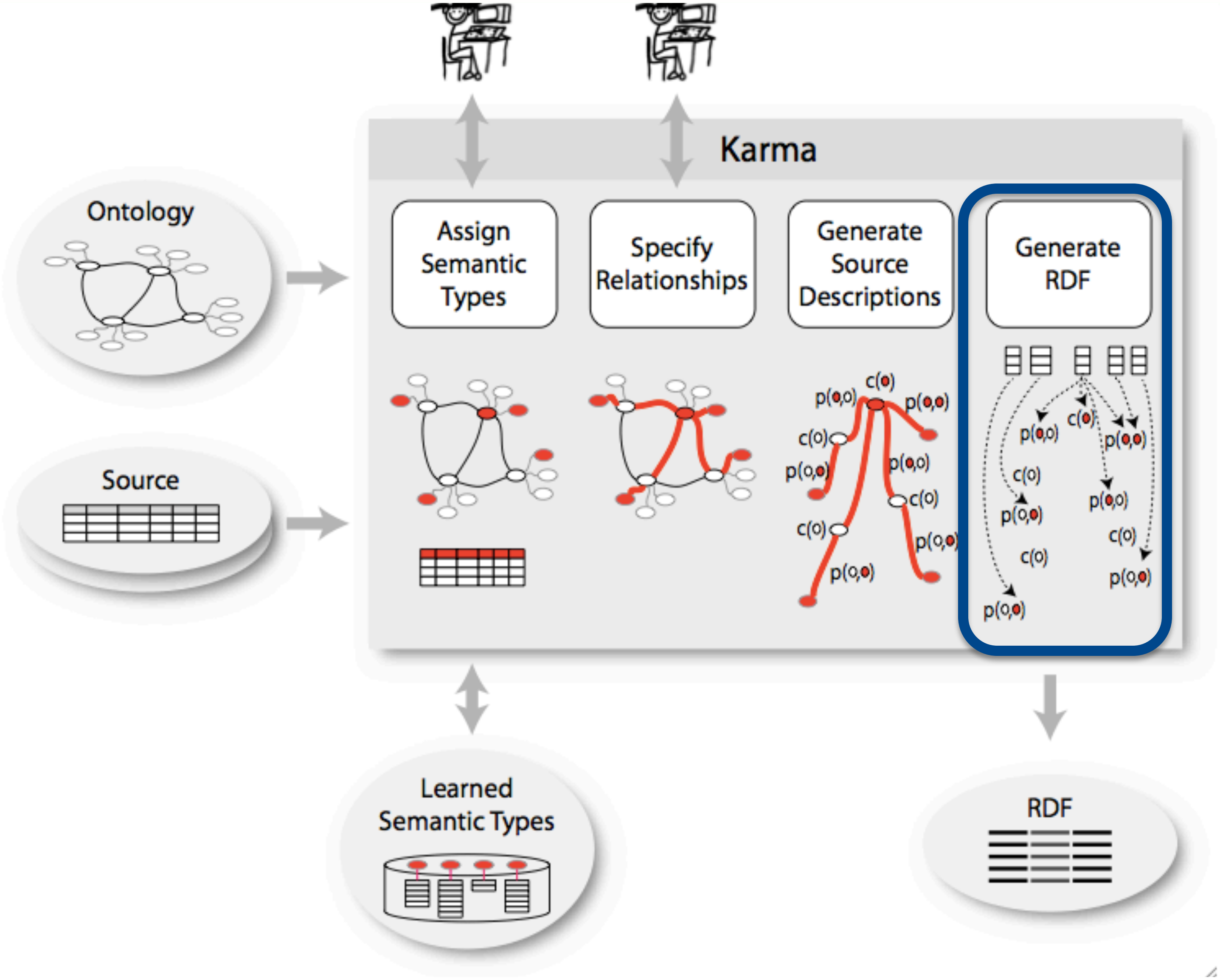


Use function symbols to create URIs:

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

$$\begin{gathered}
\mathrm{Pathway}(\mathbf{uri}(\textsc{Accession\_Id}))\ \hat{}\ \mathrm{label}(\mathbf{uri}(\textsc{Accession\_Id}), \textsc{Name})\ \hat{} \\
\mathrm{involves}(\mathbf{uri}(\textsc{Accession\_Id}), \mathbf{uri}(\textsc{Gene\_Accession\_ID}))\ \hat{} \\
\mathrm{isTargetedBy}(\mathbf{uri}(\textsc{Accession\_Id}), \mathbf{uri}(\textsc{Drug\_Accession\_Id}))\ \hat{} \\
\mathrm{isDisruptedBy}(\mathbf{uri}(\textsc{Accession\_Id}), \mathbf{uri}(\textsc{Disease\_Accession\_Id}))\ \hat{} \\
\mathrm{Gene}(\mathbf{uri}(\textsc{Gene\_Accession\_ID}))\ \hat{}\ \mathrm{label}(\mathbf{uri}(\textsc{Gene\_Accession\_ID}), \textsc{Gene\_Name})\ \hat{} \\
\mathrm{Drug}(\mathbf{uri}(\textsc{Drug\_Accession\_Id}))\ \hat{}\ \mathrm{label}(\mathbf{uri}(\textsc{Drug\_Accession\_Id}), \textsc{Drug\_Name})\ \hat{} \\
\mathrm{Disease}(\mathbf{uri}(\textsc{Disease\_Accession\_Id}))\ \hat{} \\
\mathrm{label}(\mathbf{uri}(\textsc{Disease\_Accession\_Id}), \textsc{Disease\_Name})
\end{gathered}$$

Approach: Generate RDF



Generating the RDF

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;]

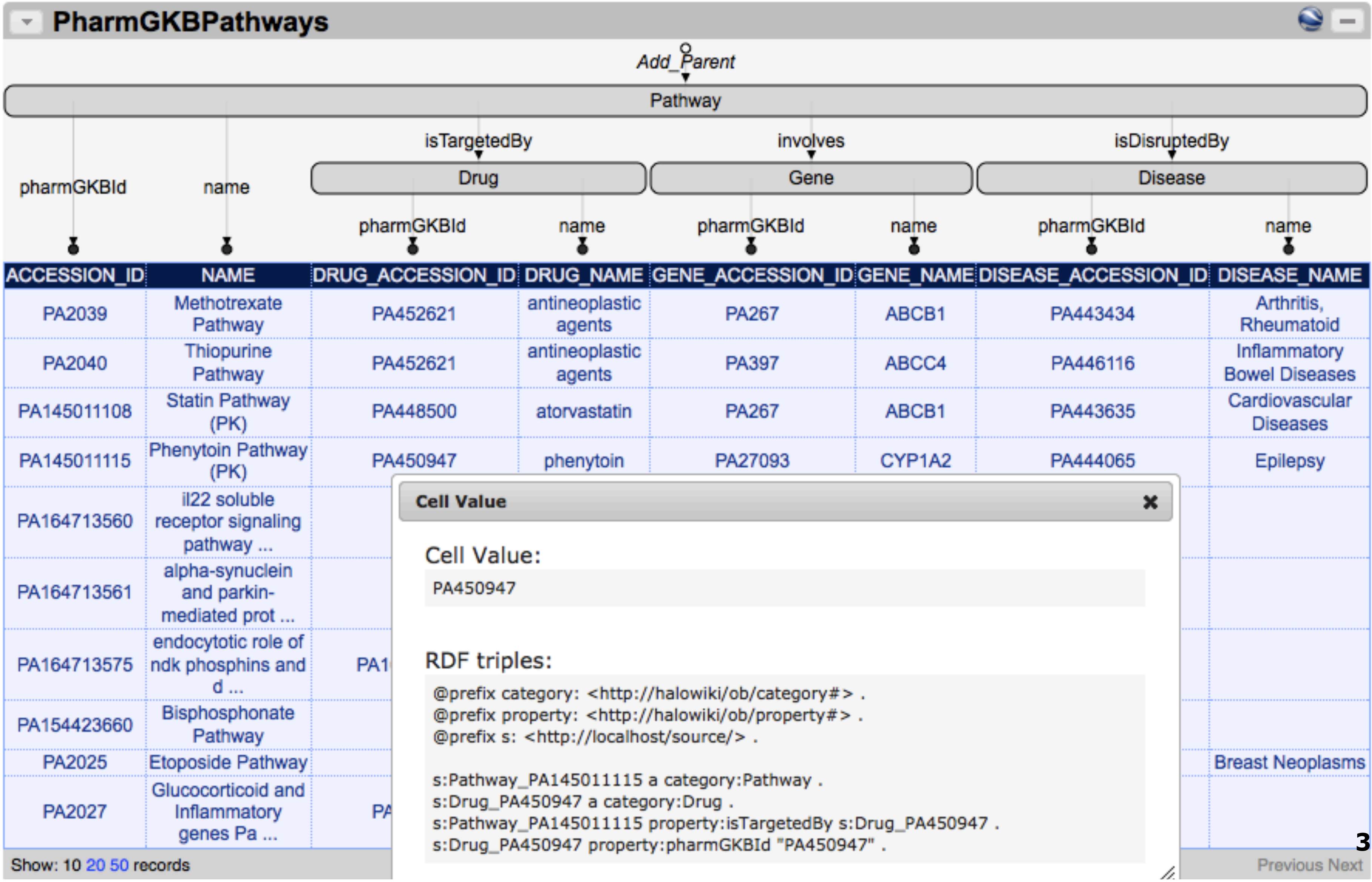
GLAV Rule

```
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" .
```

Interface can show RDF of single data cell



ACCESSION_ID	NAME	DRUG_ACCESSION_ID	DRUG_NAME	GENE_ACCESSION_ID	GENE_NAME	DISEASE_ACCESSION_ID	DISEASE_NAME
PA2039	Methotrexate Pathway	PA452621	antineoplastic agents	PA267	ABCB1	PA443434	Arthritis, Rheumatoid
PA2040	Thiopurine Pathway	PA452621	antineoplastic agents	PA397	ABCC4	PA446116	Inflammatory Bowel Diseases
PA145011108	Statin Pathway (PK)	PA448500	atorvastatin	PA267	ABCB1	PA443635	Cardiovascular Diseases
PA145011115	Phenytoin Pathway (PK)	PA450947	phenytoin	PA27093	CYP1A2	PA444065	Epilepsy
PA164713560	il22 soluble receptor signaling pathway ...						
PA164713561	alpha-synuclein and parkin-mediated prot ...						
PA164713575	endocytotic role of ndk phosphins and d ...	PA1					
PA154423660	Bisphosphonate Pathway						
PA2025	Etoposide Pathway						Breast Neoplasms
PA2027	Glucocorticoid and Inflammatory genes Pa ...	PA					

Show: 10 20 50 records

Previous Next

Cell Value ✖

Cell Value:

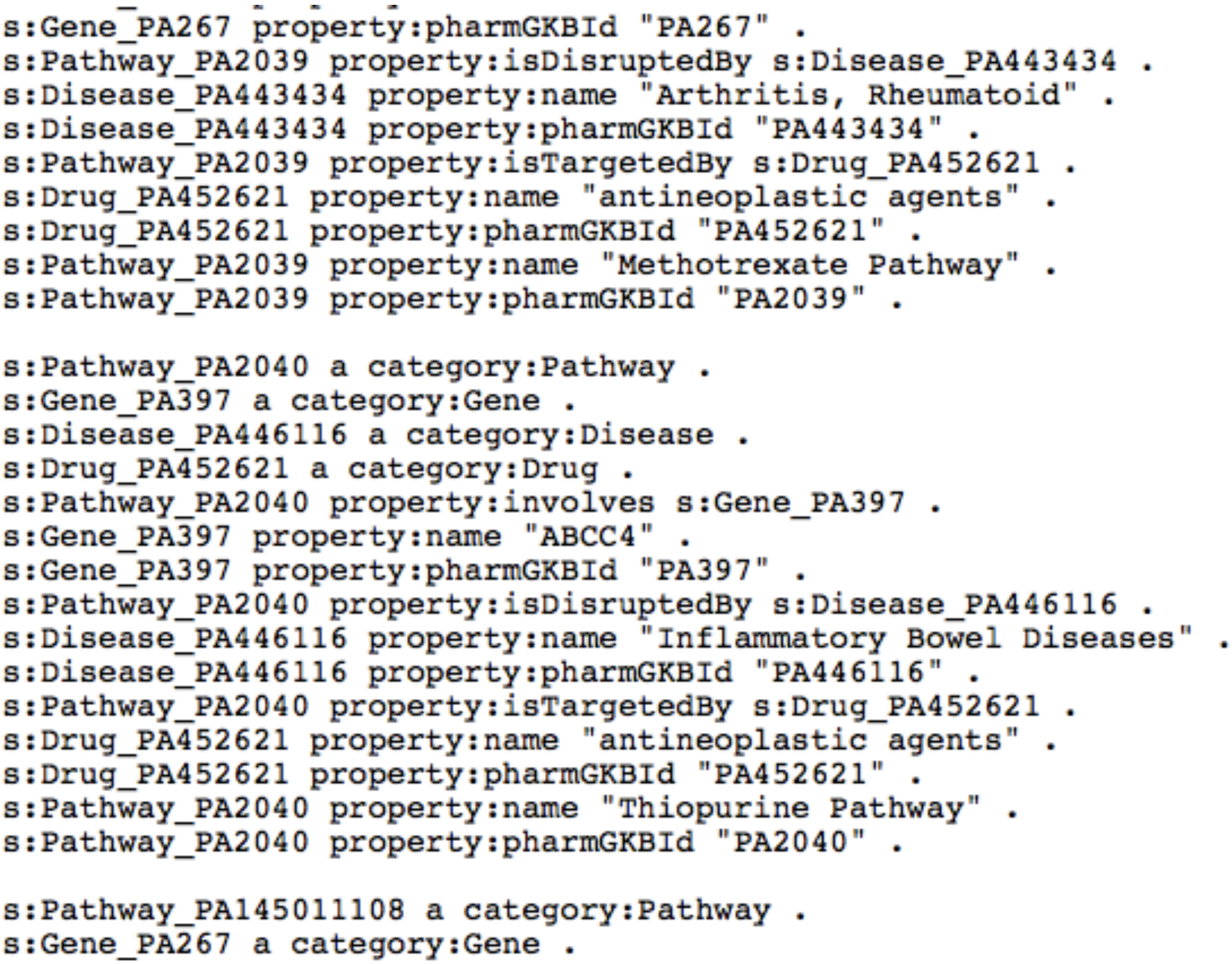
PA450947

RDF triples:

```
@prefix category: <http://halowiki/ob/category#> .
@prefix property: <http://halowiki/ob/property#> .
@prefix s: <http://localhost/source/> .

s:Pathway_PA145011115 a category:Pathway .
s:Drug_PA450947 a category:Drug .
s:Pathway_PA145011115 property:isTargetedBy s:Drug_PA450947 .
s:Drug_PA450947 property:pharmGKBId "PA450947" .
```

Export RDF



```
s:Gene_PA267 property:pharmGKBId "PA267" .
s:Pathway_PA2039 property:isDisruptedBy s:Disease_PA443434 .
s:Disease_PA443434 property:name "Arthritis, Rheumatoid" .
s:Disease_PA443434 property:pharmGKBId "PA443434" .
s:Pathway_PA2039 property:isTargetedBy s:Drug_PA452621 .
s:Drug_PA452621 property:name "antineoplastic agents" .
s:Drug_PA452621 property:pharmGKBId "PA452621" .
s:Pathway_PA2039 property:name "Methotrexate Pathway" .
s:Pathway_PA2039 property:pharmGKBId "PA2039" .

s:Pathway_PA2040 a category:Pathway .
s:Gene_PA397 a category:Gene .
s:Disease_PA446116 a category:Disease .
s:Drug_PA452621 a category:Drug .
s:Pathway_PA2040 property:involves s:Gene_PA397 .
s:Gene_PA397 property:name "ABCC4" .
s:Gene_PA397 property:pharmGKBId "PA397" .
s:Pathway_PA2040 property:isDisruptedBy s:Disease_PA446116 .
s:Disease_PA446116 property:name "Inflammatory Bowel Diseases" .
s:Disease_PA446116 property:pharmGKBId "PA446116" .
s:Pathway_PA2040 property:isTargetedBy s:Drug_PA452621 .
s:Drug_PA452621 property:name "antineoplastic agents" .
s:Drug_PA452621 property:pharmGKBId "PA452621" .
s:Pathway_PA2040 property:name "Thiopurine Pathway" .
s:Pathway_PA2040 property:pharmGKBId "PA2040" .

s:Pathway_PA145011108 a category:Pathway .
s:Gene_PA267 a category:Gene .
```

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

Evaluation Results

Source	Table Name	# Columns	# User Actions		
			Assign Type	Specify Relationship	Total
PharmGKB	Genes	8	8	0	8
	Drugs	3	3	0	3
	Diseases	4	4	0	4
	Pathways	5	2	1	3
ABA	Genes	6	3	0	3
KEGG Pathway	Pathways	6	3	1	4
	Diseases	2	2	0	2
	Genes	1	1	0	1
	Drugs	2	2	0	2
UniProt	Genes	4	1	0	1
		Total: 41	Total: 29	Total: 2	Total: 31
			Avg. User Actions/Property = 31/41 = 0.76		

There are 24 unique semantic types and the system started with no training data

If we had already learned those, a total of 7 user actions would have been required

→ 0.17 avg user actions!

Related Work

- **Mapping Databases into RDF**
  - — D2R & R2R [Bizer & Cyganiak, 2006, Bizer & Shultz, 2010]
  - — Semion [Nuzzolese, Gangemi, Presutti, & Ciancarini, 2010]
    - *Maps a database into RDF using the DB schema*
    - *Mannually defines the mappings of 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

- ***Rapidly* map existing data sources into ontology**
  - Automates as much of the mapping as possible
  - Allows the user to easily refine the mapping
- **Data exchange: generate RDF**
- **Data integration: query sources using ontology**

- **Current/Future Directions:**
  - Model semantic (linked data) services
  - XML, RDF, nested relational sources
  - Add data cleaning and record linkage capabilities
  - Mediator support

More Information

- **More information/papers available on Karma:**
  - http://www.isi.edu/~knoblock

- **Contact:**
  - Craig Knoblock: knoblock@isi.edu
  - Pedro Szekely: pszekely@isi.edu
  - Jose Luis Ambite: ambite@isi.edu

- **Software:**
  - Software is available as open source under the Apache license
  - https://github.com/InformationIntegrationGroup/Web-Karma-Public