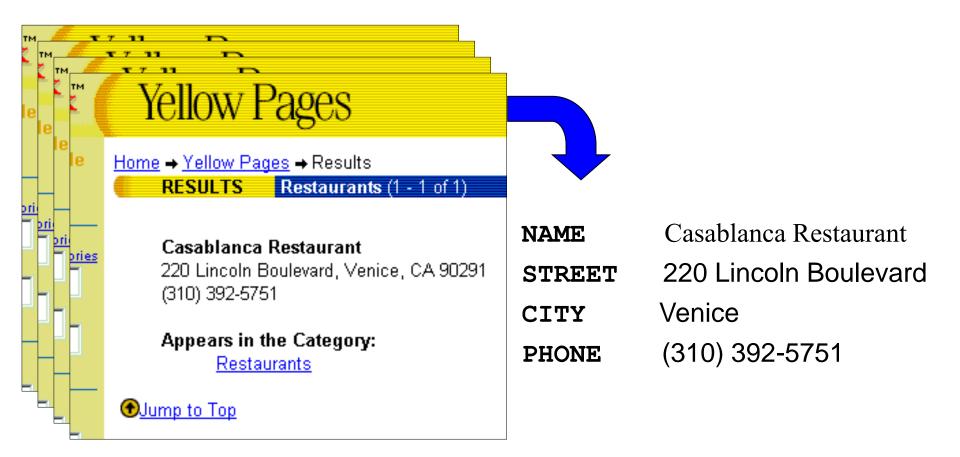


AAAI 2018 Tutorial Building Knowledge Graphs

Craig Knoblock University of Southern California

Wrappers for Web Data Extraction

Extracting Data from Semistructured Sources



Approaches to Wrapper Construction

- Manual Wrapper Construction
- Learning Wrappers from Labelled Examples
- Grammar Induction for Automatic
 Wrapper Construction

Grammar Induction Approach

- Pages automatically generated by scripts that encode results of db query into HTML
 - Script = grammar
- Given a set of pages generated by the same script
 - Learn the grammar of the pages
 - Wrapper induction step
 - Use the grammar to parse the pages
 - Data extraction step

Belive Bity 30 2908 thern California

RoadRunner

[Crescenzi, Mecca, & Merialdo]

- Automatically generates a wrapper from large web pages
 - Pages of the same *class*
 - No dynamic content from javascript, ajax, etc
- Infers source schema
 - Supports nested structures and lists
 - Extracts data from pages
- Efficient approach to large, complex pages with regular structure



- Compares two pages at a time to find similarities and differences
- Infers nested structure (schema) of page
- Extracts fields

http://www.csbooks.com/author?John+Smith

In light point (whith			
	CSbooks.com	(main many <u>manual</u> ma	
Lands, Allerando (K)	040 000 000 0000 00000	and the second of the second s	10
Browse Bubjects: Addicial Ptampence Addicial USa	You searched for beeks by	John Smith	
Additional Line Computational	Database Primer		- 1
Computer Graphics Database Theory	First Edition, Paperback 1998	Our Price: \$20	1
Disk Baked Camputing Parathel Camputing Software Engineering Visite	Second Ed Bon, Hard Cover 2000	Our Price: \$30)
	Book Description This look introduces the reader to it topics are the relational models and	te Decey and technology of database systems. The mat the SOL query language	•
	Computer Systems First Edition, Paperback 1985	Our Price: \$40	
	Book Description An undergraduate level introduction machines and moves on to parallelin	le computer and declares. I starts from the Van Resonan nachines	•
		naciatista i Batilatina i Menilitanalitinana natilatia i Diata i Dianak.Den	
	Market 1	ANT COMPLEX TON CONTRACTOR	

http://www.csbooks.com/author?Paul+Jones

Climiteren Research Manual	d Singhamp		_ight:
In these particularies	Ensenti I.		11
	CSbooks.com	Manager (1999) (1999)	· ·
	, saws , doing , second , balance	the state of the second of the second	· · · · · · · · · · · · · · · · · · ·
and Alexandre St.	- O		Read (2004-1-hot 2)
Brewse Subjects: Addicid intelligence Addiced Life	You searched for books by	Paul Jones	
Compatitional	XML at Work		
Georgeder Grephics Gerigeder Grephics Gelehese Theory	First Edition, Paperback 1999	Our Price: \$30	0.000 North
Disk Balest Camputing Parattel Camputing Settype Engineering	Book Description A comprehensive description of XM	, and all relative standards (20),	inkels)
Shana	HTML and Scripts		
	1993	Our Price: \$30	and more the
	Second Edition, Hard Cover 1999	Our Price: \$45	C HALSON D
	Book Description Anies HTM, handbook, with a full	rial an itse use of scripts for the para	enter of page on the _
	JavaScripts		
	2990	Our Price: \$50	C and a set
	Book Description A read in every Webmadier's book	n+ť	

Extracted Result

		Total number of	I SCHEWA	s found: 1	
Scheme Nun	where $1: A + B \in \{0\}$	> D E >* F >*		Total Ti	me: 0* 180 ms
sample? Abel	_				
John Smith	8				F
	Database Primer	c	D	6	This book introduces the
		First Edition, Paperback	1998	\$20	reader to the theory and technology(TRUNCATED)
		Second Edition, Hard Cover	2000	\$30	
	Computer Systems	1996 \$40		An undergraduate level introduction to computer [TRUNCATED]	
sample3.html		3 			
A					
Paul Jones	B				F
	XML at Work	G	D	0	A comprehensive description
		First Edition, Paperback	1999	\$30	of XML and all related standards [TRUNCATED]
	HTML and Soriets	nul	1993	\$30	A useful HTML handbook, with a good tutorial on the
	and the second s	Second Edition, Hard Cover	1999	\$45	use of sc. [TRUNCATED]
	JavaScripts	nu#	2000	\$50	A must in every Webmaster's bookshelf

Belive Bity 30 2908 thern California

Union-Free Regular Expression (UFRE)

- Web page structure can be represented as Union-Free Regular Expression (UFRE)
 - UFRE is Regular Expressions without *disjunctions*
 - If *a* and *b* are UFRE, then the following are also UFREs
 - a.b
 - (a)+
 - (a)?

Boliversity3of2908thern California

Union-Free Regular Expression (UFRE)

- Web page structure can be represented as Union-Free Regular Expression (UFRE)
 - UFRE is Regular Expressions without *disjunctions*
 - If *a* and *b* are UFRE, then the following are also UFREs
 - a.b \rightarrow string fields
 - (a)+ \rightarrow lists (possibly nested)
 - (a)? \rightarrow optional fields
 - Strong assumption that usually holds

Approach

- Given a set of example pages
- Generate the *Union-Free Regular Expression* which contains example pages
- Find the least upper bounds on the RE lattice to generate a wrapper in *linear time*
- Reduces to finding the least upper bound on two UFREs

Matching/Mismatches

Given a set of pages of the same type

- Take the first page to be the *wrapper* (UFRE)
- Match each successive sample page against the wrapper
- Mismatches result in generalizations of wrapper
 - String mismatches
 - Tag mismatches

Matching/Mismatches

Given a set of pages of the same type

- Take the first page to be the *wrapper* (UFRE)
- Match each successive sample page against the wrapper
- Mismatches result in generalizations of wrapper
 - String mismatches
 - Discover fields
 - Tag mismatches
 - Discover optional fields
 - Discover iterators

Beliversity39f2908thern California

Example Matching

- Wrapper (initially Page)	1):	- Sample (Page 2):
01: <html></html>	parsing	01: <html></html>
02: Books of:		→ 02: Books of:
03: 		→ 03:
04: John Smith	string mil atch (#PGDATA)	→ 04: Paul Jones
05:	• • •	05:
06: 	tag mismatch (?)	06:
		→ 07:
07: 		08:
08-10: <i>Title:</i>	ł	09-11: <i>Title:</i>
11: DB Primer	string mismatch (#PCDATA)	12: XML at Work
12:		13:
13: 		14:
14-16: <i>Title:</i>	ł	15-17: <i>Title:</i>
17: Comp. Sys.	string mismatch (#PCDATA)	18: HTML Scripts
18:	+	19:
19:	tag mismatch (+)	20:
20:	terminal tag search and	21-23: <i>Title:</i>
	-	24: Javascript
	square matching	+25:
- Wrapper after solving mi	smatches:	26:
• •		27:
<html>Books of: #PCDATA<</html>	/B>	
()?		
	1. (IT T)) +	
(<i>Title:</i>#PCDAT	A)'	

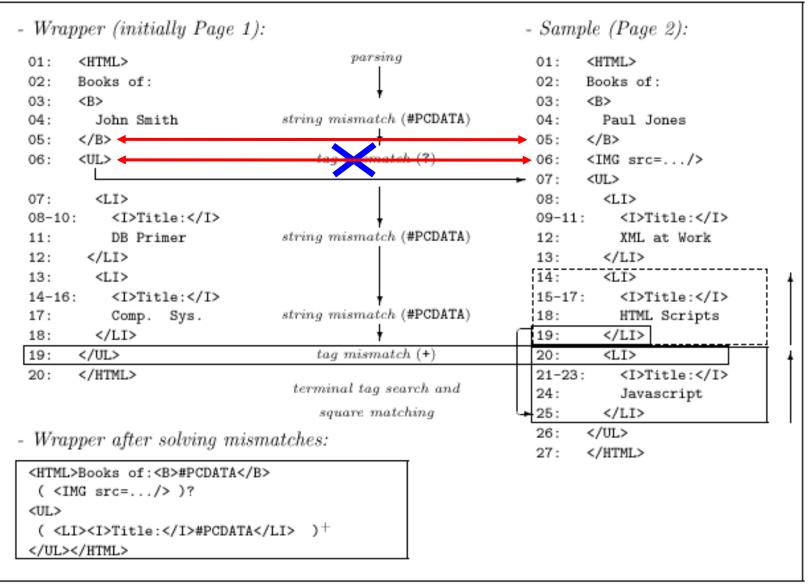
Boliversity30f2908thern California

String Mismatches: Discovering Fields

- String mismatches are used to discover fields of the document
- Wrapper is generalized by replacing "John Smith" with #PCDATA
- <HTML>Books of: John Smith
- \rightarrow <HTML> Books of: #PCDATA

Beiversity3of2908thern California

Example Matching



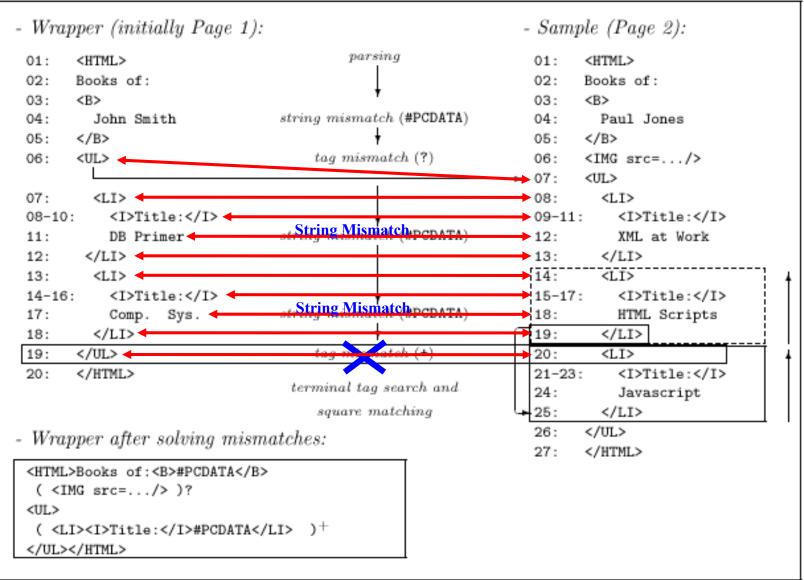
Beiverity3of2908thern California

Tag Mismatches: Discovering Optionals

- First check to see if mismatch is caused by an iterator (described next)
- If not, could be an optional field in wrapper or sample
- Cross search used to determine possible optionals
- Image field determined to be optional:
 - ()?

Bebreesity39f2908thern California

Example Matching



Boliversity3of2908thern California

Tag Mismatches: Discovering Iterators

- Assume mismatch is caused by repeated elements in a list
 - End of the list corresponds to last matching token:
 - Beginning of list corresponds to one of the mismatched tokens: or
 - These create possible "squares"
- Match possible squares against earlier squares
- Generalize the wrapper by finding all contiguous repeated occurrences:
 - (<I>Title:</I>#PCDATA)+

Belive Bity 30 29 18 thern California

Example Matching

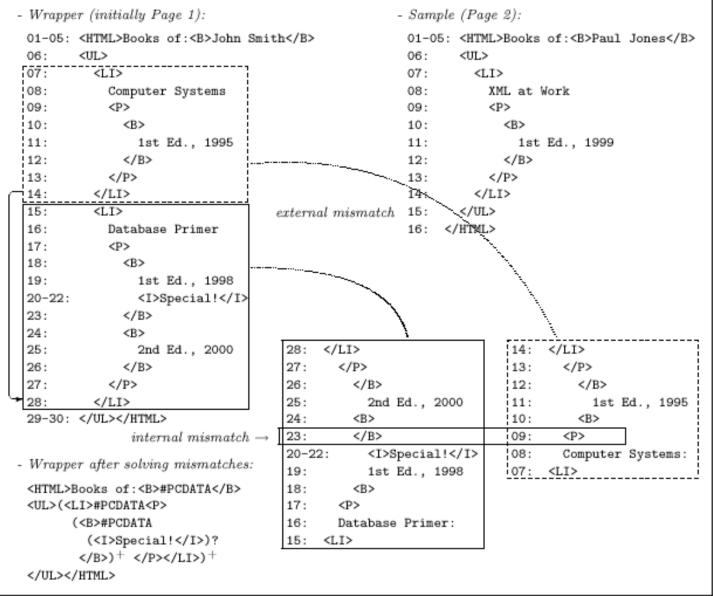
- Wraz	pper (initially Page 1)):	- Sample (Page 2):
01:	<html></html>	parsing	01: <html></html>
02:	Books of:		02: Books of:
03:		+	03:
04:	John Smith	string mismatch (#PCDATA)	04: Paul Jones
05:		ŧ	05:
06:		tag mismatch (?)	06:
			→ 07:
07:			08:
08-10	: <i>Title:</i>	ł	09-11: <i>Title:</i>
11:	DB Primer	string mismatch (#PCDATA)	12: XML at Work
12:			13:
13:			14:
14-16	: <i>Title:</i>	ŧ	15-17: <i>Title:</i>
17:	Comp. Sys.	string mismatch (#PCDATA)	18: HTML Scripts
18:		+	19:
19:		tag mismatch (+)	20:
20:		to an in the second second	21-23: <i>Title:</i>
		terminal tag search and	24: Javascript
		square matching	→ 25:
- Wra	pper after solving mis	matches:	26:
			27:
	>Books of: #PCDATA<td>3></td><td></td>	3>	
	MG src=/>)?		
			
	I> <i>Title:</i> #PCDATA)+	

Internal Mismatches

- Generate *internal mismatch* while trying to match square against earlier squares on *the same page*
 - Solving internal mismatches yield further refinements in the wrapper
 - List of book editions
 - I>Special!</I>

Beliversity 30f2908thern California

Recursive Example



Discussion

Assumptions:

- Pages are well-structured
- Structure can be modeled by UFRE (no disjunctions)
- Search space for explaining mismatches is huge
 - Uses a number of heuristics to prune space
 - Limited backtracking
 - Limit on number of choices to explore
 - Patterns cannot be delimited by optionals
 - Will result in pruning possible wrappers

Limitations

- Learnable grammars
 - Union-Free Regular Expressions (RoadRunner)
 - Variety of schema structure: tuples (with optional attributes) and lists of (nested) tuples
 - Does not efficiently handle disjunctions pages with alternate presentations of the same attribute
 - Context-free Grammars
 - Limited learning ability
- User needs to provide a set of pages of the same type

Inferlink Web Extraction Software

USCBIMBORMATIO2Sciences Institute Extraction

FOR SALE: STOEGER M3500

post id: 4700468 share: **f v p**

Price:	\$ 500	Listed On:	Thursday, September 17, 2015
Seller:	Private Party	Listed In:	Shotguns
Account:	Registered on 5/9/2013	Location:	Keenesburg, Denver, Colorado - Map
	Listings by this user	Shipping:	No

Manufacturer:	Stoeger	Flag Edit Favorite
Caliber:	12 Gauge	
Action:	Semi-automatic	Contact Seller
Firearm Type:	Shotgun	

I have a Stoeger m3500. It is a year old. It has 200 rounds through it from clay shooting. Its in perfect condition. If you have any questions email or text me. 9703427061. I'm asking 500



Contact Seller

CC-By 2.0 27

Structured Extraction

FOR SALE	E: STOEGER M35	500	
post id: <mark>4700468</mark> share: f 🖂 💆			
Price:	\$ 500	Listed On:	Thursday, September 17, 2015
Seller:	Private Party	Listed In:	Shotguns
Account:	Registered on 5/9/2013 Listings by this user	Location: Shipping:	Keenesburg, Denver, Colorado Map No
Manufacturer: Caliber: Action: Firearm Type:	Stoeger 12 Gauge Semi-automatic Shotgun		Flag Edit Favorite Contact Seller
	m3500. It is a year old. It has or text me. 9703427061. I'm a		shooting. Its in perfect condition. If you have any



Contact Seller

USEBInformation Sciences Institute Automated Extraction [Minton et al., Inferlink]

- Title
- Description
- Seller
- Post Date
- Expiry Date
- Price
- Location
- Category
- Member Since
- Num Views
- Post ID

			Post a Listing
What are you looking for?		Panolver	÷ 👂
tint + Edmatrix + Childle + Mandolat + Rugar Nove Man	Citrage dis 11.12*		
Ruger New Model Single-Si	x 9 1/2" 🖓	1539.00 Map C	orized Postor
	Neurol: 815-181-3201 Sandi Gen Tiwe Oger Neurville Neuron: United States Spartwatel Codes: 52211 Spatier 21, 2018 11:21 am Diplema: 21 days, 2 houre	1/2*	ew Model Eingle-Sia 9 × Serverates 2711
Description Sup, 3:0" barrol, original vestinut grtup, Notizer 4 barrows. 2: VID is hopper (professional pol), PRID- tolater, original plastic boxfook & legs 2228 op bowyhing appears in one except sight. 2228 bowyhing appears in one except sight. 2228 bowyhing appears in one except sight. 2228 bowyhing appears in one except. Notice Participation and a sight of the sight sight. 2020 parentl, or delivery to FFL resulted with IN QOW permit, or delivery to FFL resulted with	TECH OUTDOORS centure shouldent inderNactury test annelope willned gind hyper springprotective gun size IN cylinder spin mark with only 1 to 10 to set for 50 pts. using hismady 45 pt f 7N, \$508. Must here TN chiver's locat	nel 130-round 1X 1X 1X	E.
Other Poderal see, you, may not sell a frame provide party transaction transmission attracts of tradeout macromonal diary you and call for transaction with the main, model, annu number and called of the hyperge sensing framework in the maximum matcher Ad Independent 7: 5205cbase177428.	Transmite Laterant). While not respired, it is refl to if of also (units here for a protocol one) the entry you self or buy. Do sum to verify that the p	Hy Formation - Ser Wester	Foday Rective 201 Magnum (Virviewa) Bank Regellius (Virviewa) Rada Regellius (Virviewa)
		> mince Na	80 (3 viver) 18 645 335 Auto, possible trade for

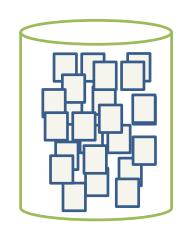






US CBlyn for mation Sciences Institute

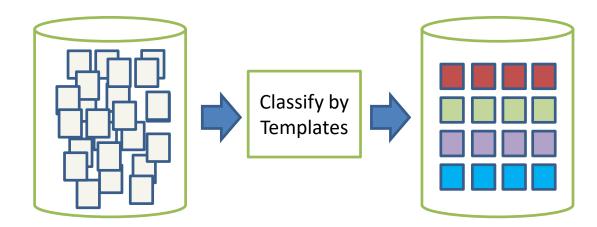
Automated Extraction



Input: A Pile of Pages

US CBlyn for mation Sciences Institute

Automated Extraction



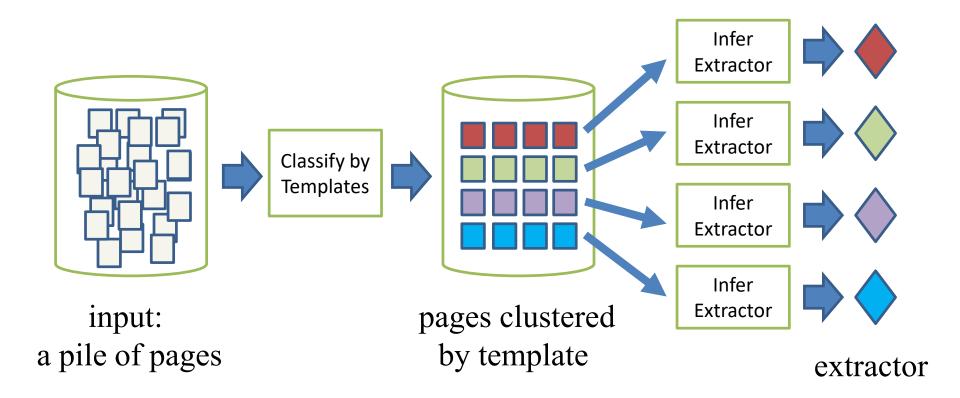
input: a pile of pages pages clustered by template

Clustering

- Cluster
 - Based on the visible text
 - Page is broken into chunks
 - These are continuous blocks of text
 - Search for common visible chunks
 - Remove chunks that occur in all pages
 - Remove those that occur in fewer than 10 pages
 - Greedy algorithm to cluster the pages based on the remaining chunks
 - Sort by the size of the clusters created by each chunk

USEBInformation Sciences Institute

Automated Extraction



Extractor Learning

- Input: cluster
- Select 5 random pages to build a template
 - Tokenize on space & punctuation
 - Start with n-grams of tuples of size n
 - Find those n-grams that occur on all pages
 - Keep only those n-grams that occur exactly once per page
 - Decompose pages based on these n-grams
 - Run algorithm recursively on decomposed page
 - Repeat above for size n-1 down to n=2
 - Construct template based on the decomposition

USCBlynfo@mation34Sciences Institute

Unsupervised Extraction Tool

	InferLink	{Landmark}
--	-----------	------------

tennesseegunexchange

Markup

Extraction Downloads

8 11

+ Add Page	page_2.html	page_3.html	page_4.html	page_5.html
0036 Winchester mo 140		Beretta hand gun for sale	1956 Mossin nagant	Ruger New Model Single-Six 9 1/2″
Code1354	37122	37201	37122	37211
Description1392	Winchester model 140 12 guage semiauto. 28inch ribbed and vented barrel with modified choke. Gun fires and cycles well. Text for pics.	9mm Beretta hand gun for sale at a very good price with a delivery to any interested buyer in the state.contact me at silven2016@yandex.com or text me at	1956 Mossin nagant. 7.62×54 bolt action Russian war rifle. Missing bayonet but rifle fires and cycles well.	Blue, 9½" barrel, original walnut grips, NcStar 4x32E scope/mount + lighted reticle and lens covers, 2 1/2 lb trigger (professional job), PRC TECH OUTDOORS cordura shoulder/belt holster, original plastic box/lock & amp; key/.22LR cylinder/factory test envelope w/fired casing/instruction manual/original rear sight/ original trigger spring/protective gun sleeve. Everything appears as new except sligh .22WRM cylinder spin mark with only 1 to 1½ 50-round boxes fired. .22LR cylinder never used. Scope is set for 50 yds. using Hornady 45 gr FTX .22WMR ammo. Total package f.o.b. Nashville, TN, \$539. Mus have TN driver's license and/ or TN CCW permit, or delivery to FFL required with photo of Driver's License and cleared payment.
Expires1366	45 days, 23 hours	This ad has expired	46 days, 4 hours	89 days, 7 hours
Facebook2081	2015	2015	2015	2016
Firearms1276	raquo; Shotguns » Winchester model 140	raquo; Pistols » Semi- Auto » Beretta hand gun for sale	raquo; Rifles » Bolt Action » 1956 Mossin nagant	raquo; Pistols » Revolver » Ruger New Model Single-Six 9 1/2″
Group2122	Non-Felon / Legal	Non-Felon / Legal	Non-Felon / Legal	Mentally Capable / Non-Felon / Legal
ID1421	47955db13ce85f8e	930552dc84196efa	8655da9d2455236	6256c6aa4070496

US (Blyn for mation Sciences Institute

Extraction Evaluation

10 websites, 5 pages each

fields

	Title	Desc	Seller	Date	Price	Loc	Cat	Member Since	Expires	Views	ID
Perfect	1.0 (50/50)	.76 (37/49)	.95 (40/42)	.83 (40/48)	.87 (39/45)	.51 (23/45)	.68 (34/50)	1.0 (35/35)	.52 (15/29)	.76 (19/25)	.97 (35/36)
Including partial and extra data	1.0 (50/50)	.98 (48/49)	.95 (40/42)	.83 (40/48)	.98 (44/45)	.84 (38/45)	.88 (44/50)	1.0 (35/35)	.55 (16/29)	1.0 (25/25)	1.0 (36/36)

Discussion

- Inferlink approach solves some of the key limitations of Roadrunner
 - Pages do not all have to be of the same type
 - Multiple optionals would be treated as different page types
 - Scales well with complex pages

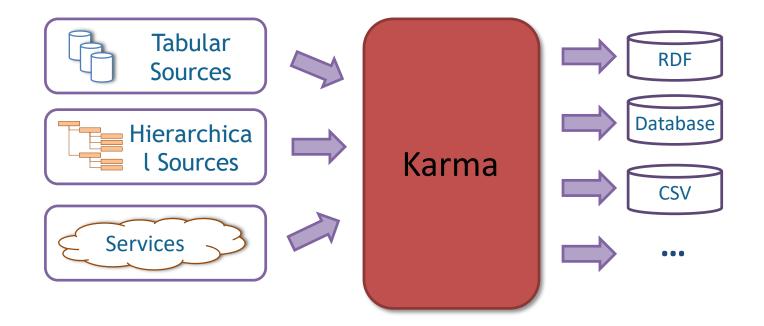
Web Data Extraction Software

- Beautiful Soup
 - <u>http://www.crummy.com/software/BeautifulSoup/</u>
 - Python library to manually write wrappers
- Jsoup
 - http://jsoup.org/
 - Java library to manually write wrappers
- ScrapingHub
 - http://scrapinghub.com/
 - Portia provides a wrapper learner
- Others
 - <u>https://www.quora.com/Which-are-some-of-the-best-web-data-scraping-tools</u>
 - Tell us if you find a good one!

Aligning and Integrating Data in Karma



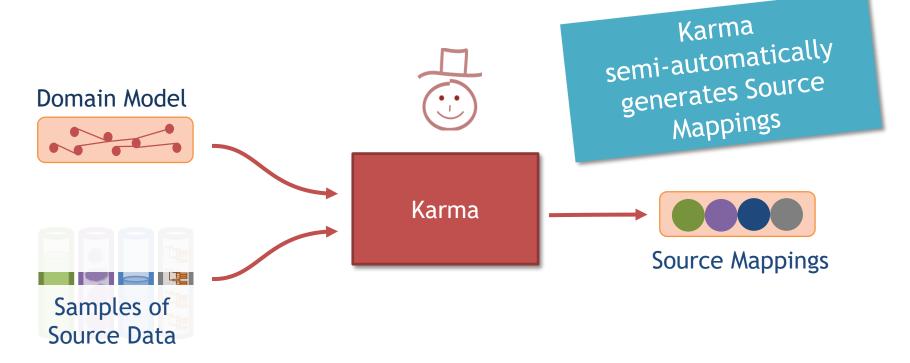
Interactive tool for rapidly extracting, cleaning, transforming, integrating and publishing data



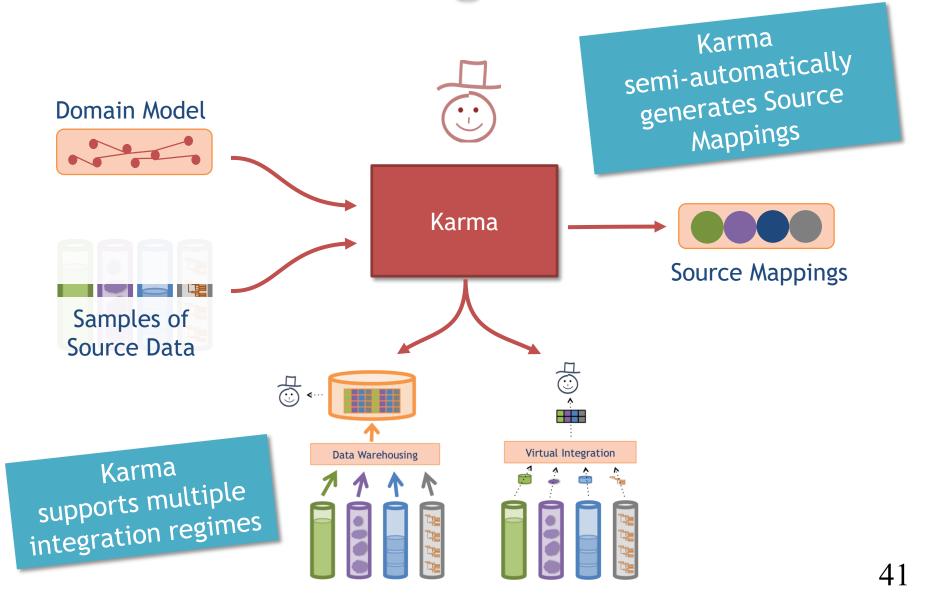




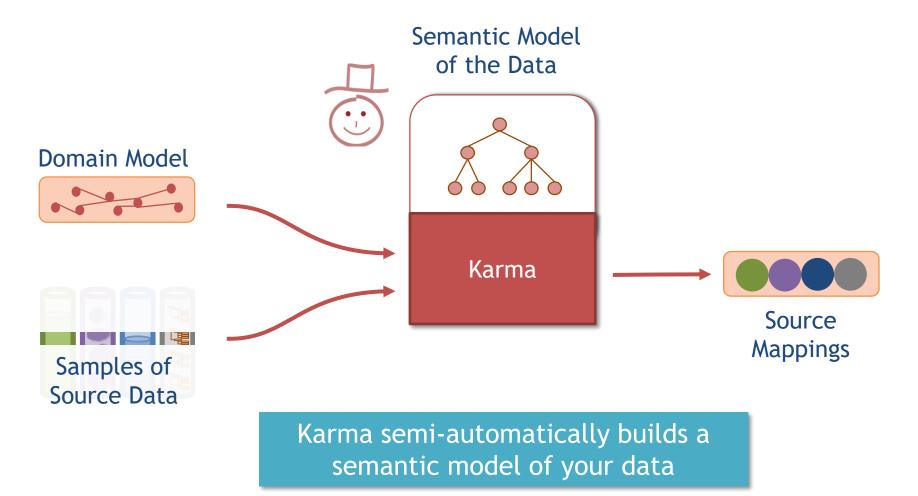
Information Integration in Karma



Information Integration in Karma



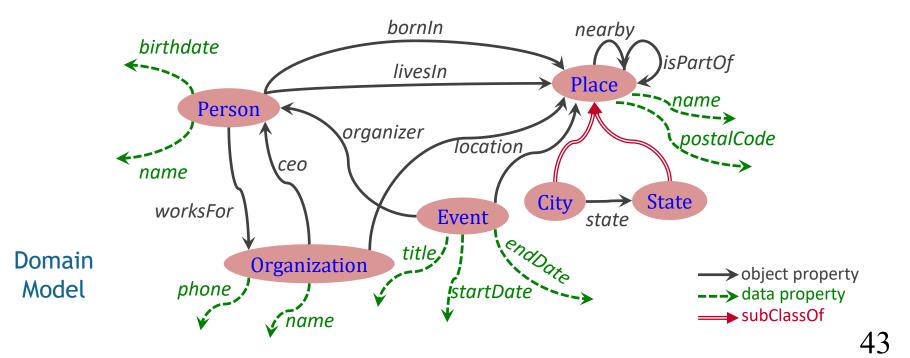
Secret Sauce: Karma Understands Your Data



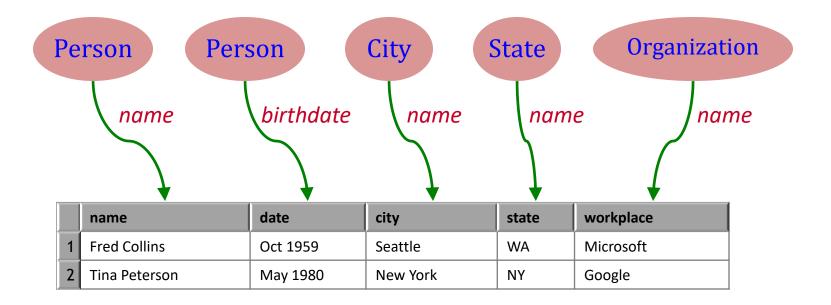
What is a Semantic Model?

Describe sources using classes & relationships in an ontology

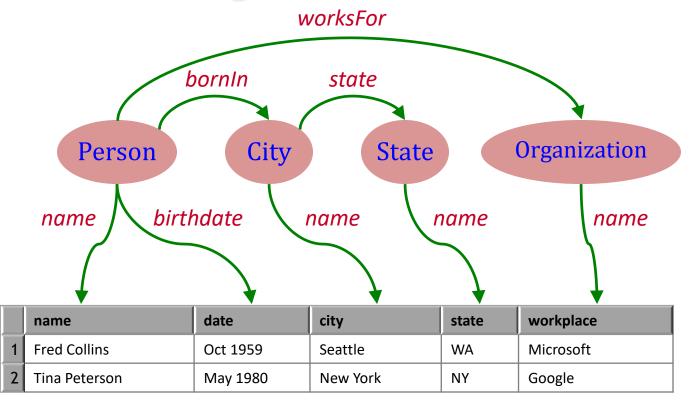
Source date city workplace state name Fred Collins Oct 1959 Microsoft Seattle WA 2 **Tina Peterson** May 1980 New York NY Google



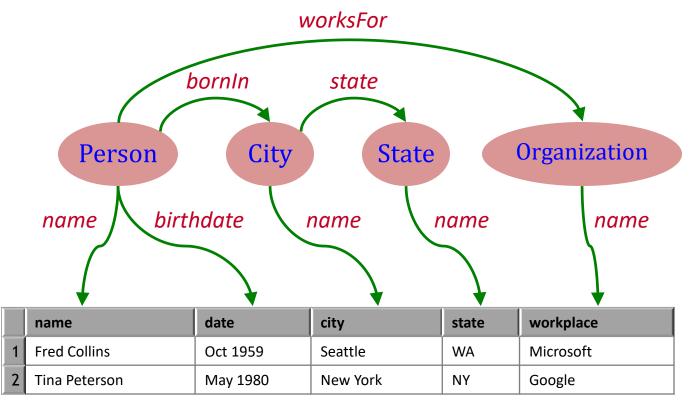
Semantic Types



Relationships

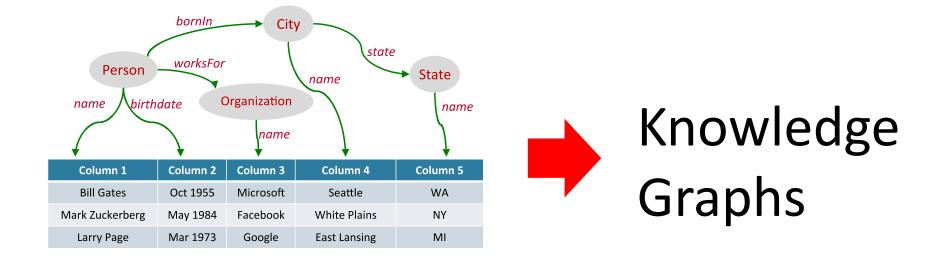


Semantic Model



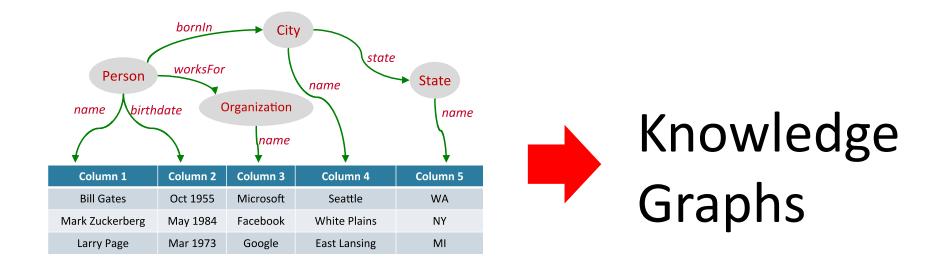
Semantic models will be formalized as Source Mappings

Key ingredient to automate <u>source discovery</u>, <u>data integration</u>, and <u>publishing semantic data (RDF triples)</u> 46



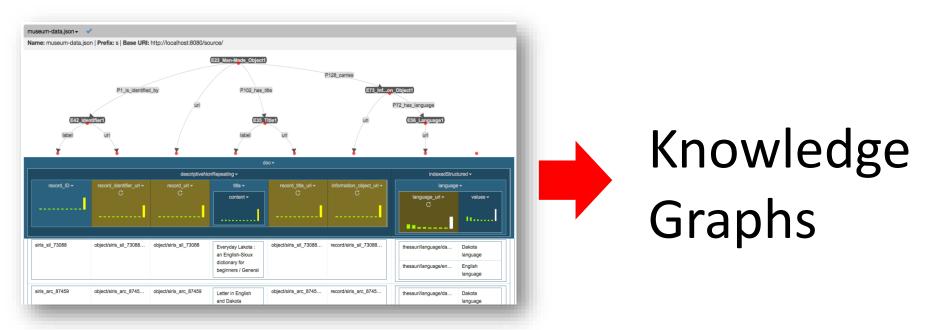
Karma uses semantic models to create knowledge grap

Karma semi-automatically builds semantic models



Karma uses semantic models to create knowledge grap

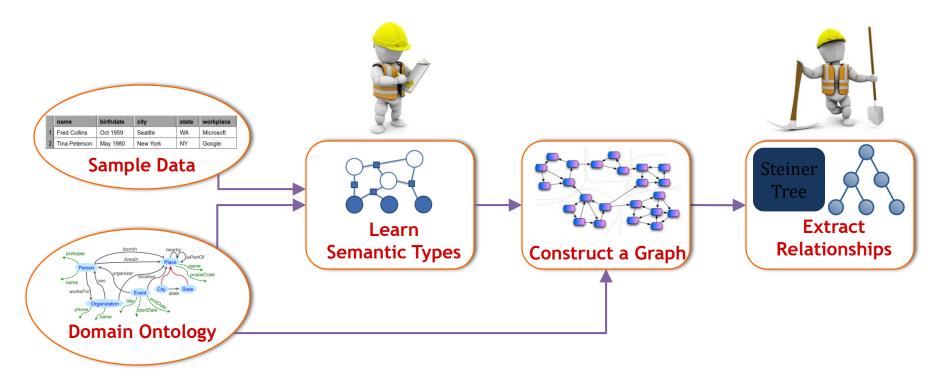
Karma semi-automatically builds semantic models ... and provides a nice GUI to edit them



Karma uses semantic models to create knowledge grap

Semi-automatically Building Semantic Models in Karma

Approach [Knoblock et al, ESWC 2012]

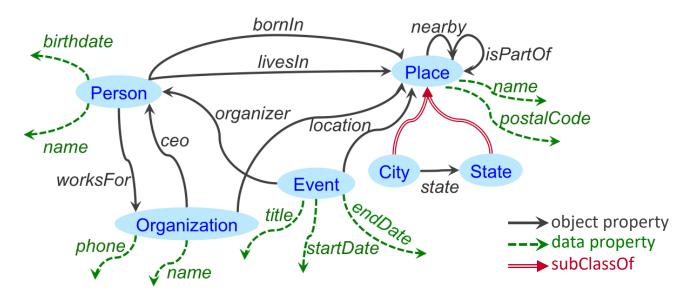




Source

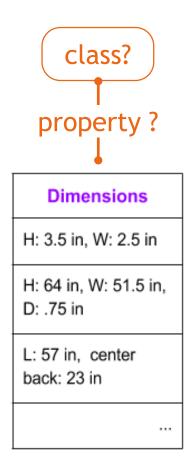
name	date	city	state	workplace
1 Fred Collins	Oct 1959	Seattle	WA	Microsoft
2 Tina Peterson	May 1980	New York	NY	Google

Domain Ontology

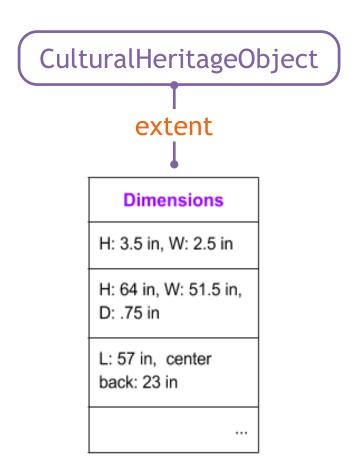


Find a semantic model for the source (map the source to the ontology)

Learning Semantic Types [Krishnamurthy et al., ESWC 2015]



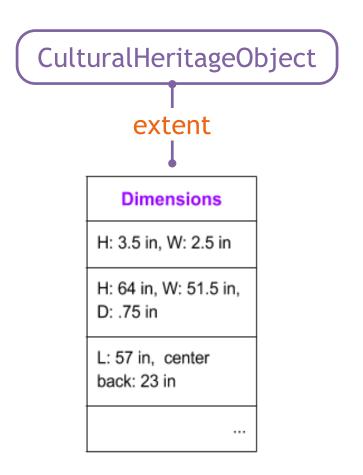
Learning Semantic Types



1- User specifies

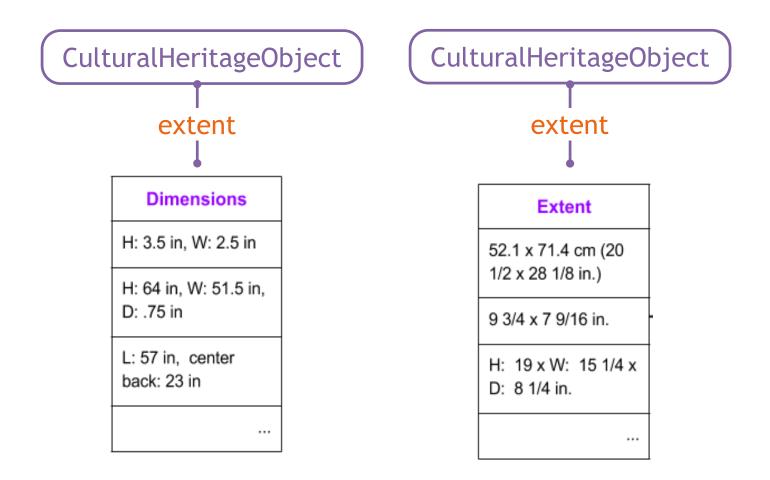
2- System learns

Learning Semantic Types



Extent				
52.1 x 71.4 cm (20 1/2 x 28 1/8 in.)				
9 3/4 x 7 9/16 in.				
H: 19 x W: 15 1/4 x D: 8 1/4 in.				

Learning Semantic Types



Requirements

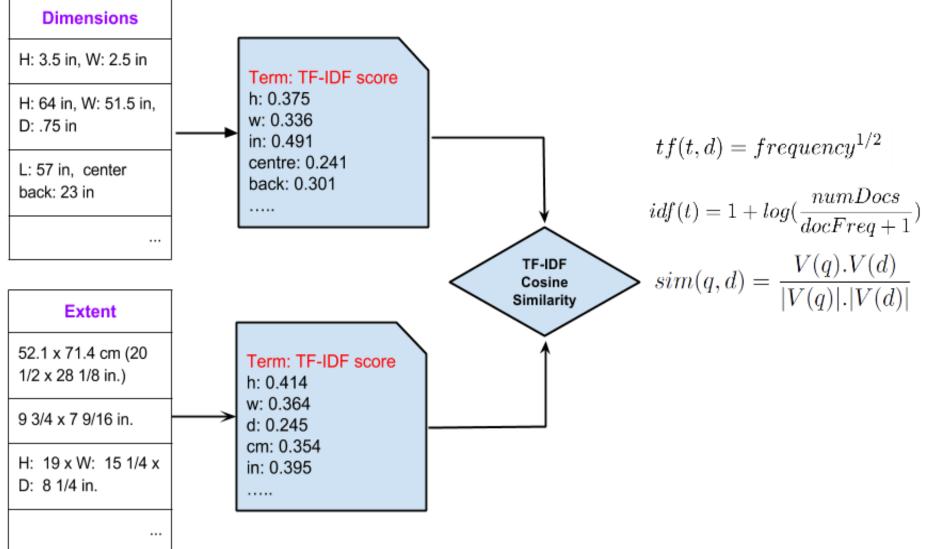
- Learn from a small number of examples
- Work on both textual and numeric values
- Learn quickly and highly scalable to large number of semantic types

Approach for Textual Data

- Document: each column of data
- Label: each semantic type
- Use Apache Lucene to index the labeled documents
- Compute TF/IDF vectors for documents
- Compare documents using Cosine Similarity between TF/IDF vectors

Dimensions
H: 3.5 in, W: 2.5 in
H: 64 in, W: 51.5 in, D: .75 in
L: 57 in, center back: 23 in

Approach for Textual Data

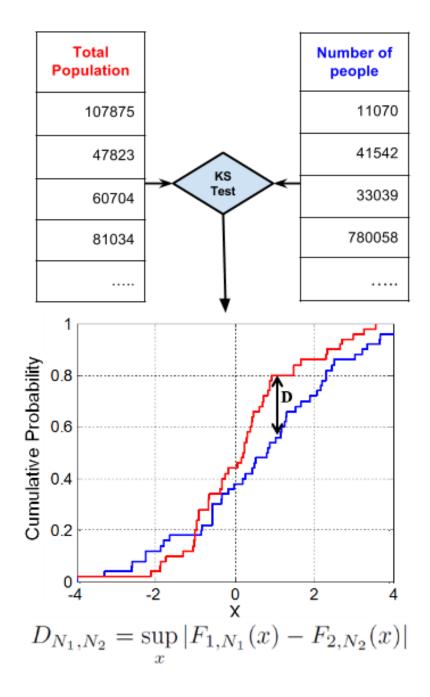


Approach for Numeric Data

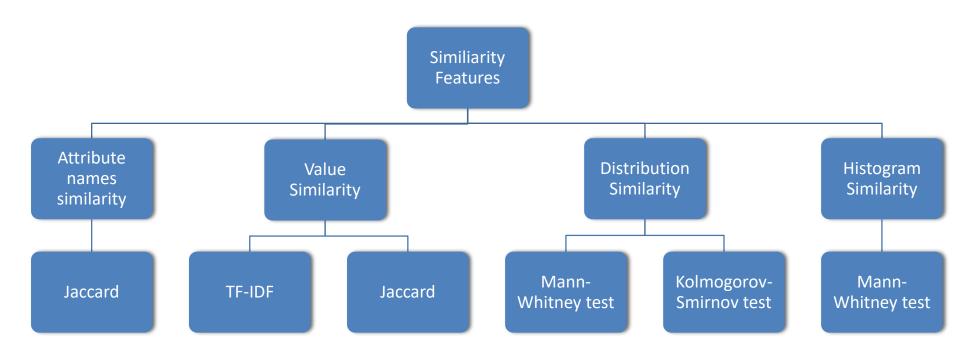
- Distribution of values in different semantic types is different, e.g., temperature vs. population
- Use Statistical Hypothesis Testing to see which distribution fits best
- Welch's T-test, Mann-Whitney U-test and Kolmogorov-Smirnov Test

Total Population	Number of people
107875	11070
47823	41542
60704	33039
81034	780058

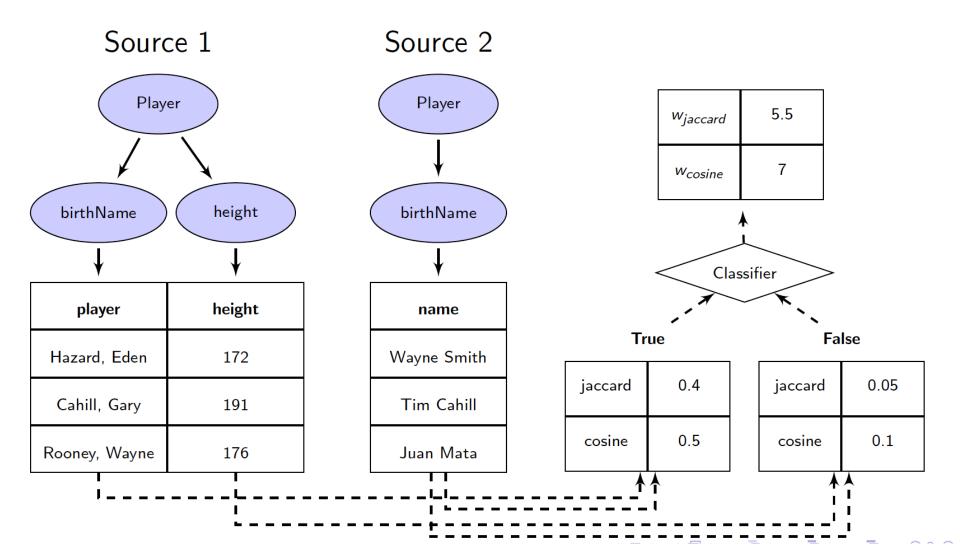
Approach for Numeric Data



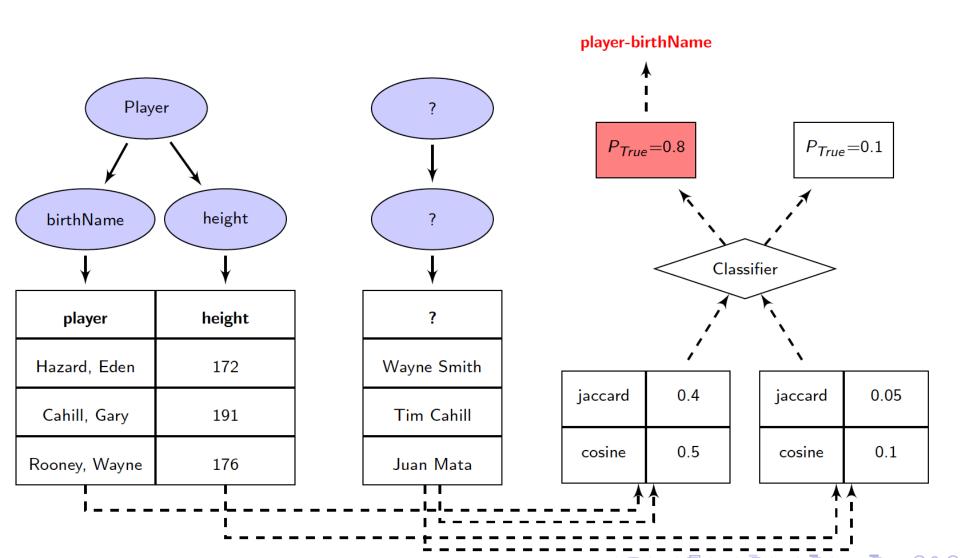
Similarity features



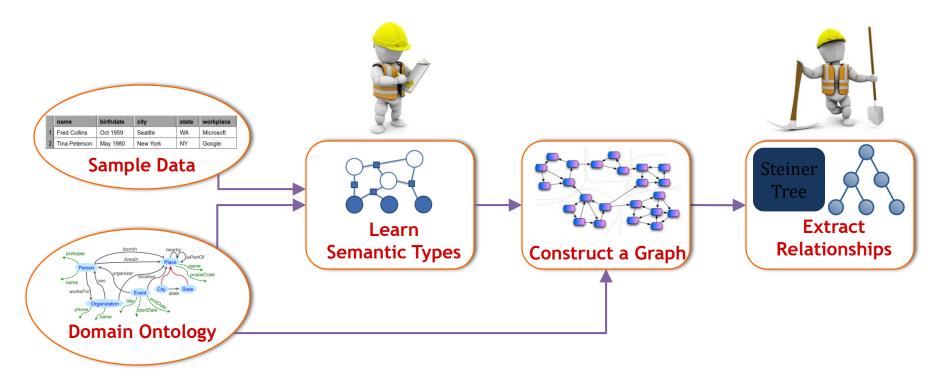
Training machine learning model [Pham et al., ISWC 2016]



Predicting new attribute



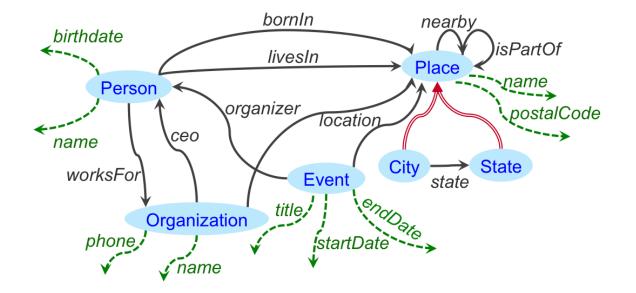
Approach [Knoblock et al, ESWC 2012]



Construct a Graph

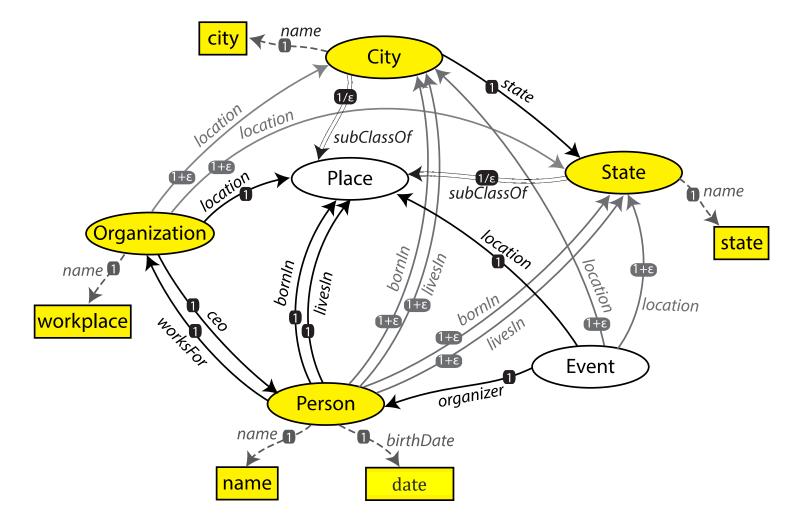
Construct a graph from semantic types and ontology

Person City State Organization name birthdate name name name						
	name	date	city	state	workplace	
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft	
2	Tina Peterson	May 1980	New York	NY	Google	



Construct a Graph

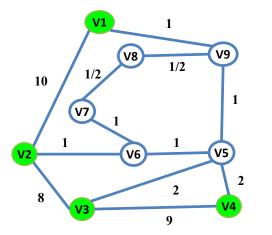
Construct a graph from semantic types and ontology



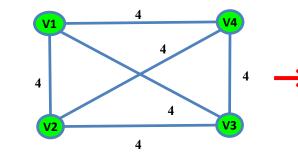
Inferring the Relationships

- Search for minimal explanation
- 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 [Kou et al., 1981]
 - Worst-case time complexity: O(|V|²|S|)
 - Approximation Ratio: less than 2

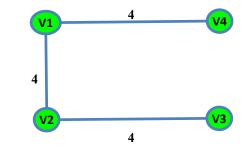
Steiner Tree



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

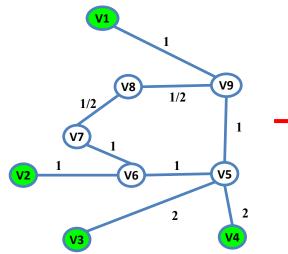


1. construct the complete graph (Nodes: Steiner Nodes, Links Weights: shortest path from each pair in original G)

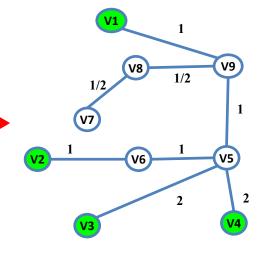


2. Compute MST

V1



3. replace each <u>link</u> with the corresponding shortest <u>path</u> in original G



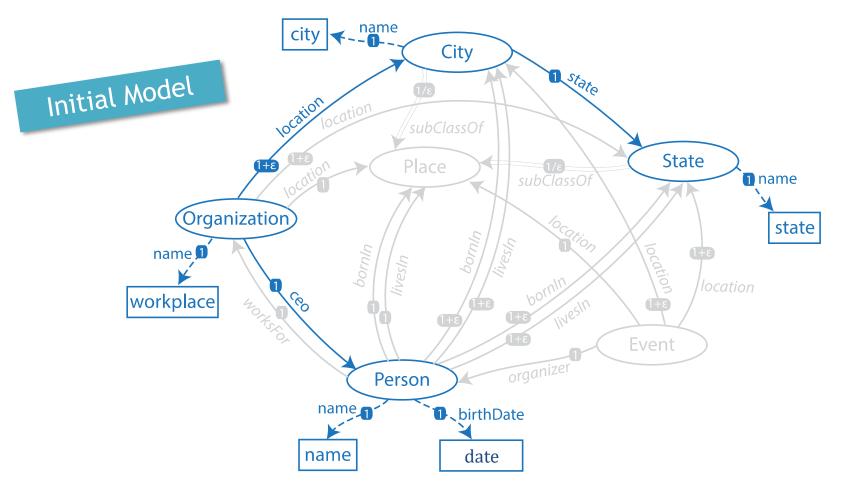
V9 1 V2 1 V6 1 V5 2 2 V3 V4

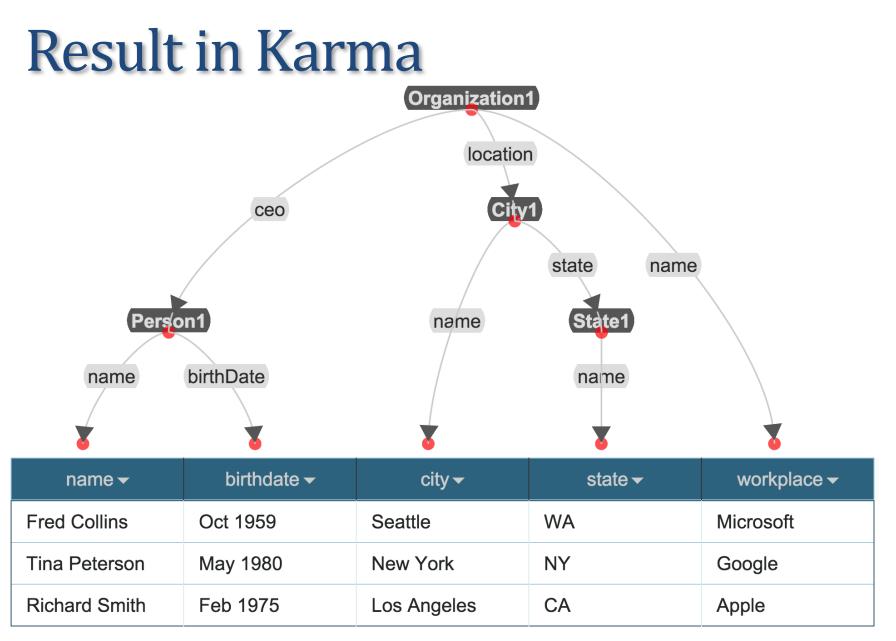
4. Compute MST

5. remove extra links until all <u>leaves</u> are Steiner nodes 69

Inferring the Relationships Select minimal tree that connects all semantic types

• A customized Steiner tree algorithm

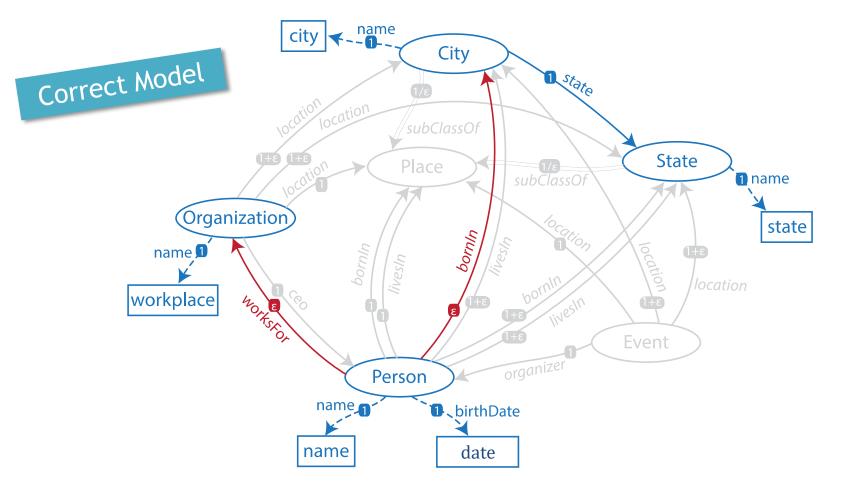


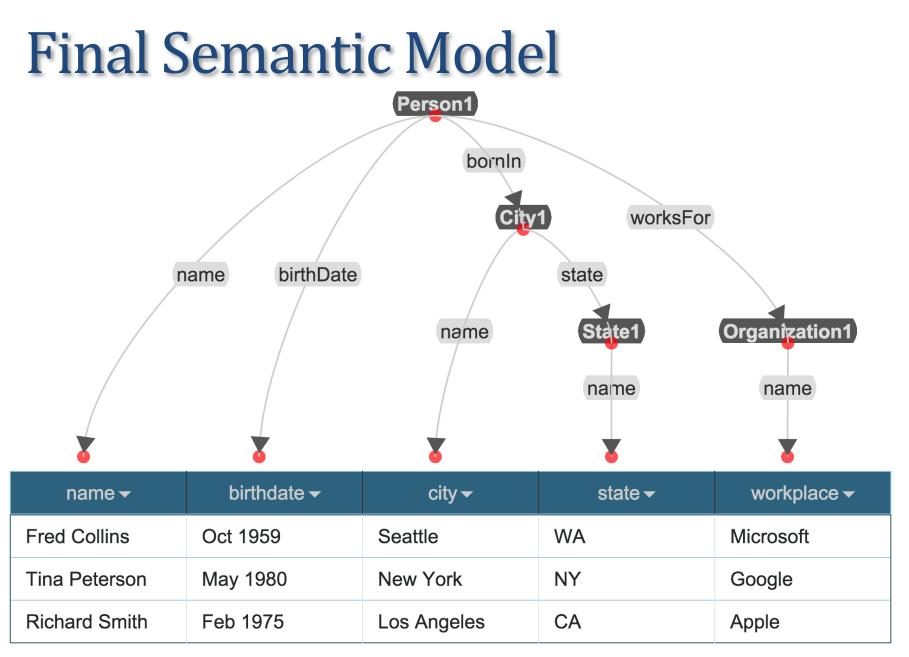


Refining the Model

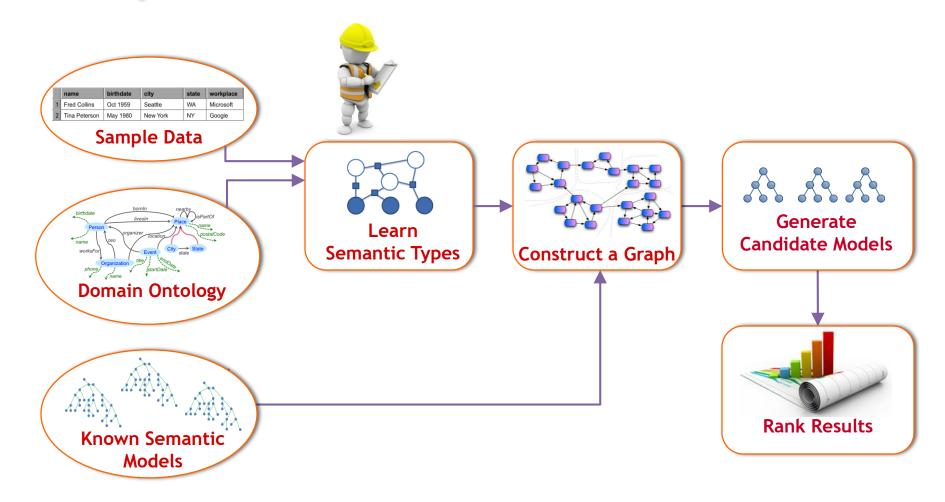
Impose constraints on Steiner Tree Algorithm

- Change weight of selected links to ϵ
- Add source and target of selected link to Steiner nodes





Karma Learns the Source Models Taheriyan et al., ISWC 2013, ICSC 2014



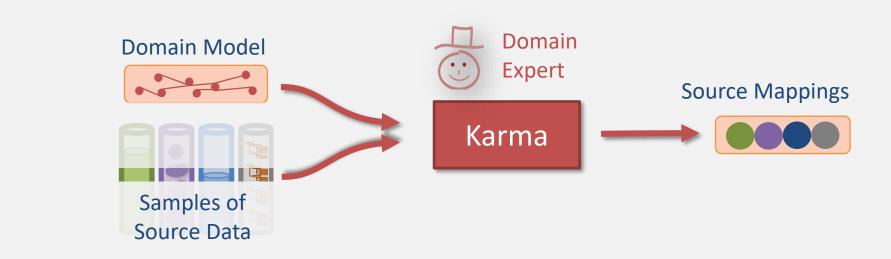
Karma Use Cases



University of Southern California

Pedro Szekely and Craig Knoblock

Source Mapping Phase



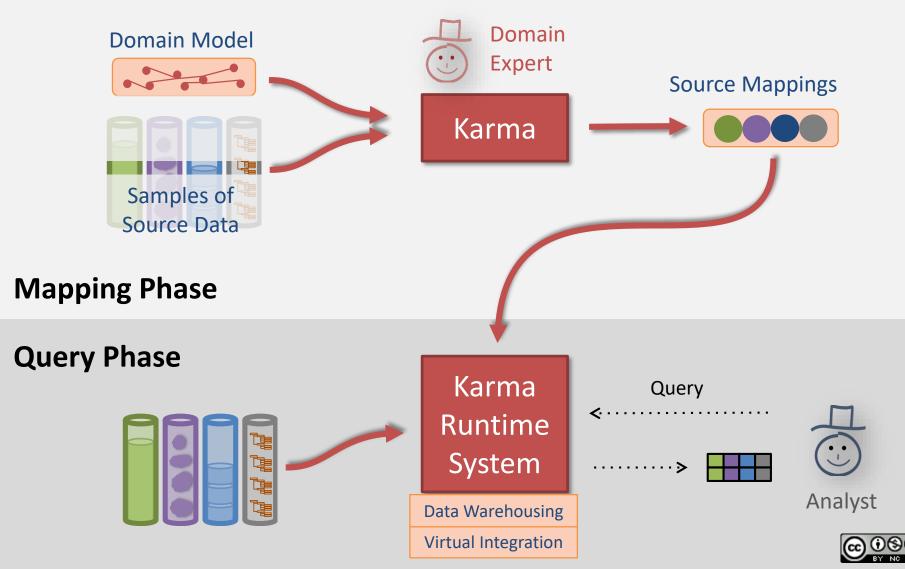
Mapping Phase



University of Southern California

Pedro Szekely and Craig Knoblock

Source Mapping and Query Time



University of Southern California

Pedro Szekely and Craig Knoblock

VIVO

- <u>VIVO</u> is a system to build researcher networks across institutions
- Used Karma to map the data about USC faculty to VIVO ontology and publish it as RDF
- VIVO ingest the RDF data



An interdisciplinary network

Enabling collaboration and discovery among scientists across all disciplines.

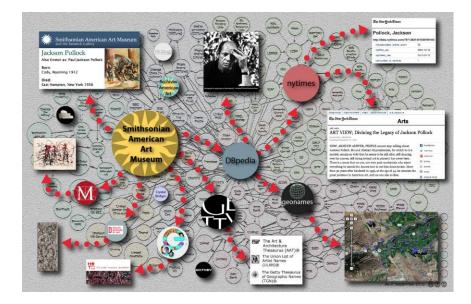
The network of scientists will facilitate scholarly discovery. Institutions will participate in the network by installing VIVO, or by providing semantic web-compliant data to the network.



• <u>Video</u>

American Art Collaborative [Knoblock et al., ISWC 2017]

- Used Karma to convert data of 13 American Art Museums to Linked Open Data
- Modeled according to CIDOC-CRM Ontology
- Linked the generated RDF to DBPedia and ULAN





Using Karma to map museum data to the CIDOC CRM ontology

https://www.youtube.com/watch?v=h3_yiBhAJlc

Discussion

- Automatically build rich semantic descriptions of data sources
- Exploit the background knowledge from (i) the domain ontology, and (ii) the known source models
- Semantic descriptions are the key ingredients to automate many tasks, e.g.,
 - Source Discovery
 - Data Integration
 - Service Composition



More Info

karma.isi.edu