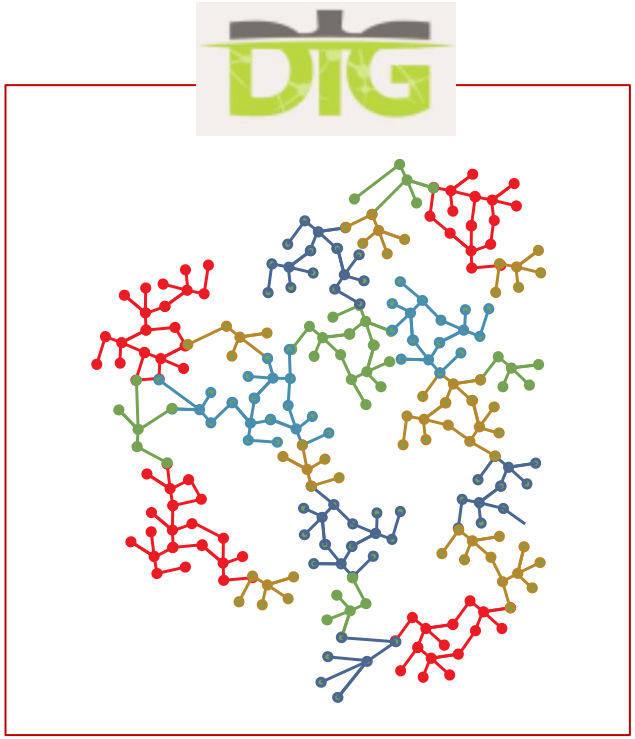
USC University of Southern California

Building knowledge graphs in DIG

Pedro Szekely and Craig Knoblock
University of Southern California
Information Sciences Institute
dig.isi.edu

Goal



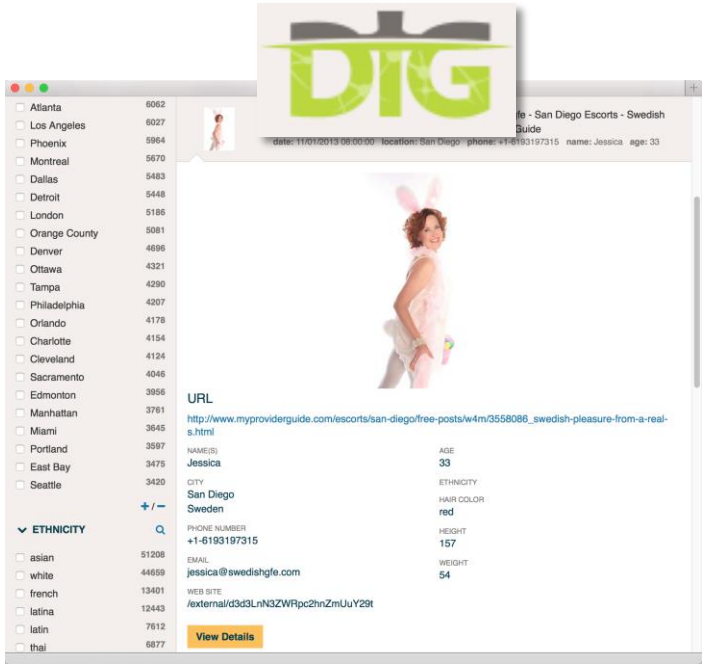
raw ◆ messy ◆ disconnected

hard to query, analyze & visualize

clean ◆ organized ◆ linked

easy to query, analyze & visualize

Use Case: Human Trafficking



raw ◆ messy ◆ disconnected

hard to query, analyze & visualize

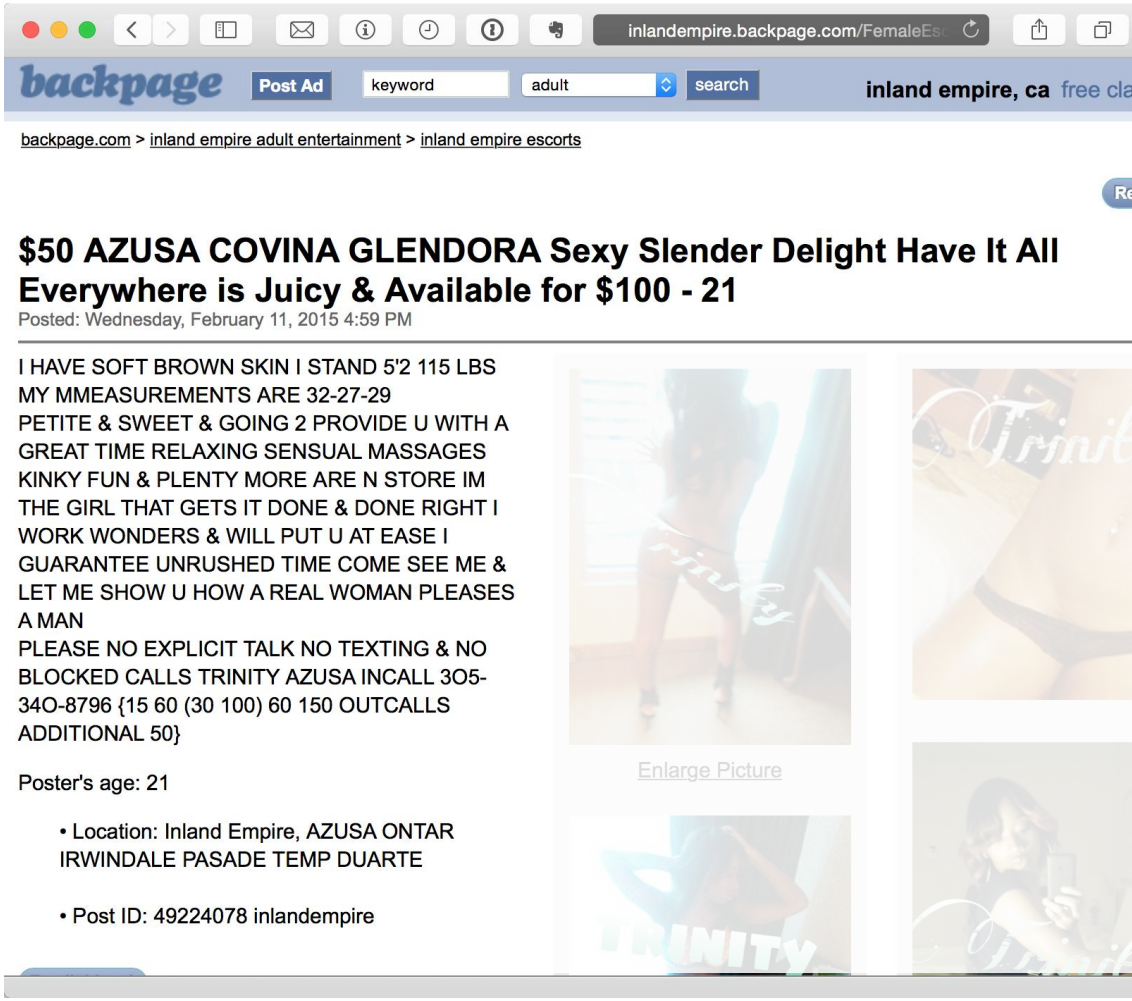
clean ◆ organized ◆ linked

easy to query, analyze & visualize

Use Case: Human Trafficking

100 million pages
~ 100 Web sites

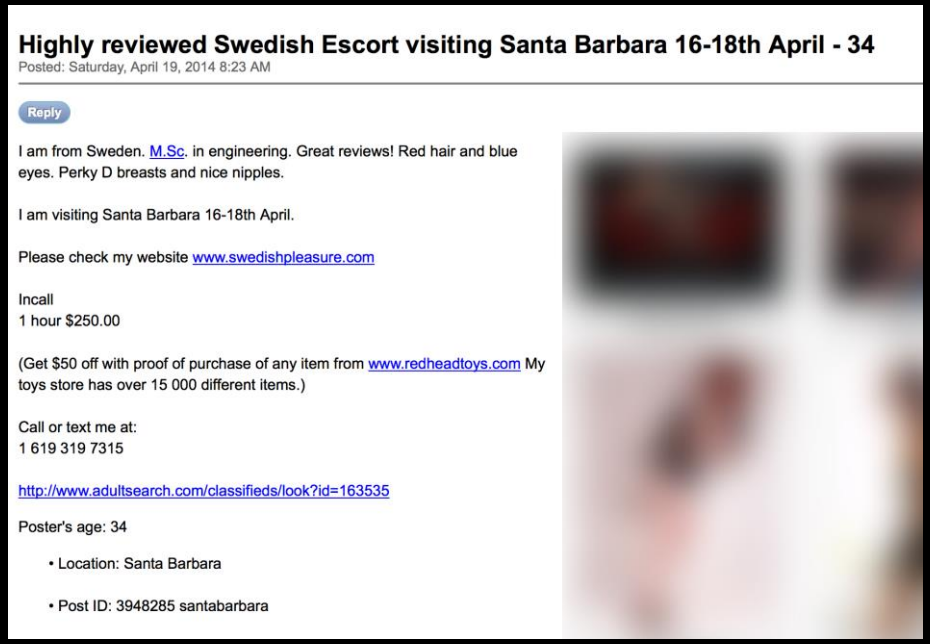
help victims
prosecute traffickers



Salient Statistics on Human Trafficking

- Profits per Year: $32 Billion
- Average Age of Entry To Prostitution in the US: 14
- PIMP's Profit Per Victim Per Year: $150,000
- Advertising Budget On the Web: $45 Million

Task: Tracking the Victim's



Highly reviewed Swedish Escort visiting Santa Barbara 16-18th April - 34

Posted: Saturday, April 19, 2014 8:23 AM

Reply

I am from Sweden. M.Sc. in engineering. Great reviews! Red hair and blue eyes. Perky D breasts and nice nipples.

I am visiting Santa Barbara 16-18th April.

Please check my website www.swedishpleasure.com

Incall
1 hour $250.00

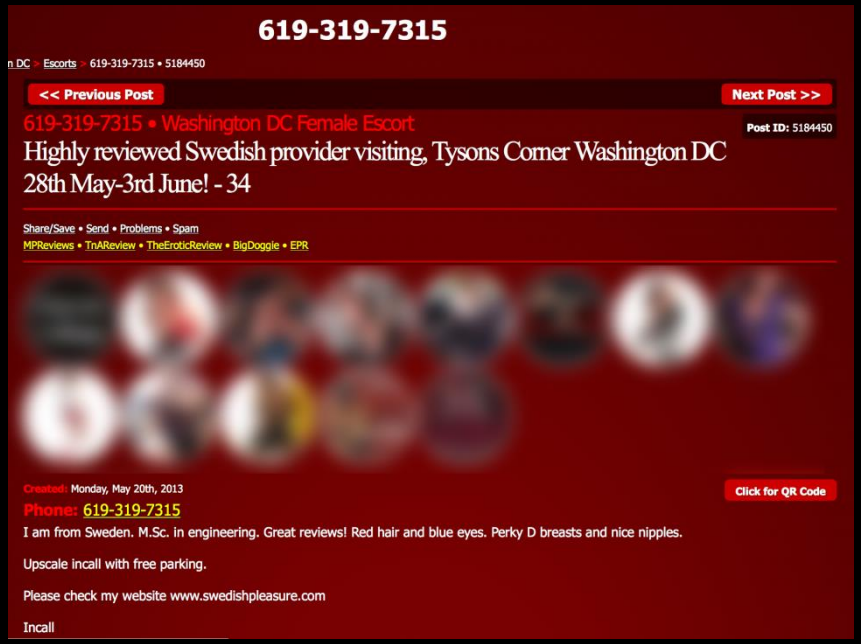
(Get $50 off with proof of purchase of any item from www.redheadtoys.com My toys store has over 15 000 different items.)

Call or text me at:
1 619 319 7315

http://www.adultsearch.com/classifieds/look?id=163535

Poster's age: 34

- Location: Santa Barbara
- Post ID: 3948285 santabarbara



619-319-7315

n DC > Escorts > 619-319-7315 • 5184450

<< Previous Post | Next Post >>

619-319-7315 • Washington DC Female Escort

Post ID: 5184450

Highly reviewed Swedish provider visiting, Tysons Corner Washington DC 28th May-3rd June! - 34

Share/Save • Send • Problems • Spam

MPReviews • TnAReview • TheEroticReview • BigDoggie • EPR

Created: Monday, May 20th, 2013

Click for QR Code

Phone: 619-319-7315

I am from Sweden. M.Sc. in engineering. Great reviews! Red hair and blue eyes. Perky D breasts and nice nipples.

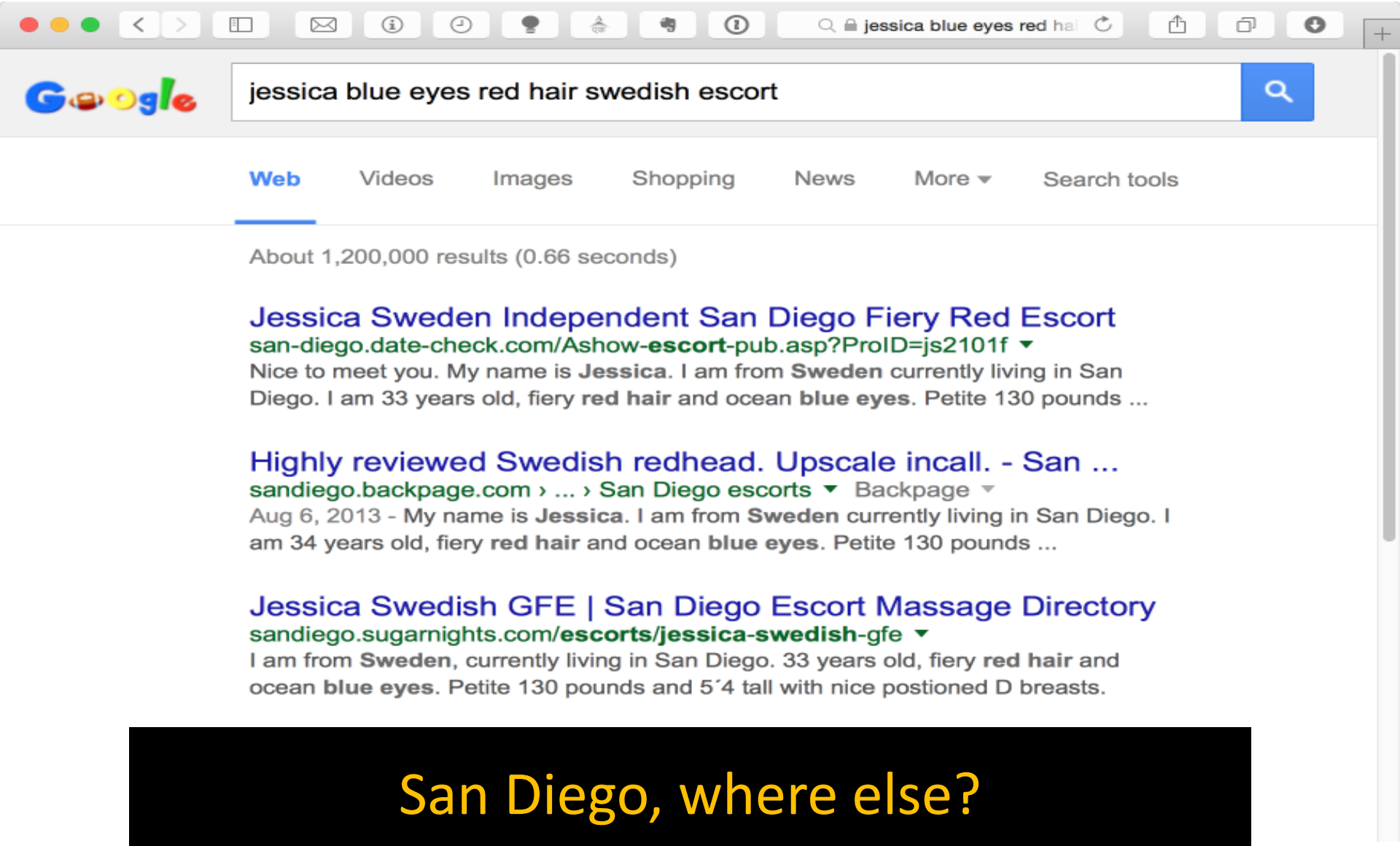
Upscale incall with free parking.

Please check my website www.swedishpleasure.com

Incall

> 100 million pages advertising adult services

Example: Investigating a Reported Victim

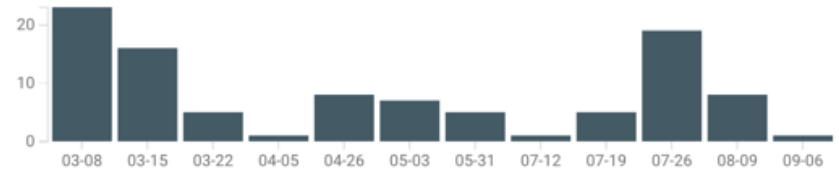


San Diego, where else?

DIG Interface: Find the locations where a potential victim was advertised

202-596-5980

Offers between 2015-03-08 and 2015-09-06

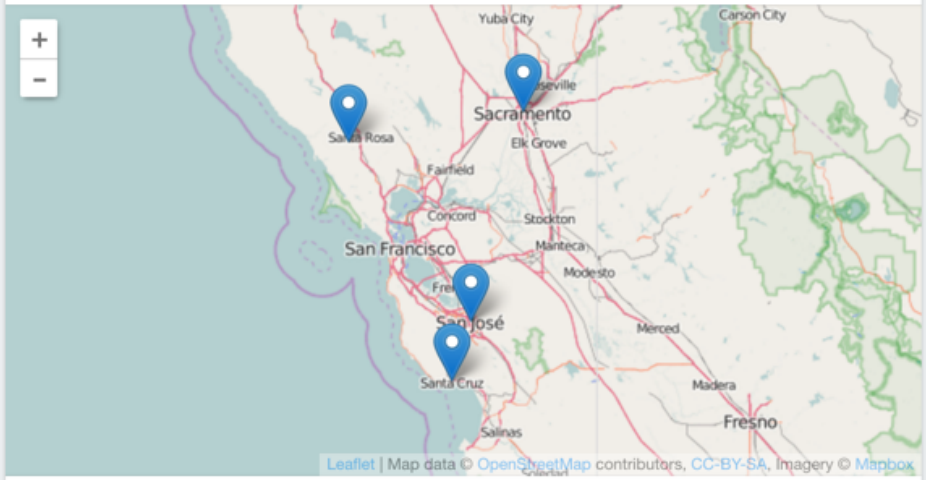


Services offered

NAME		AGE	
lexi	3	21	10
		24	7
		22	2
		23	2

(photos section)

4 location(s)



OFFERS PER CITY

North Bay	41
Sacramento	33
San Jose	18
San Francisco	6
Santa Cruz	1

40 offer(s)

202-596-5980
??Busty Bombshell In Town??OutCalls ONLY - Sacramento escorts - backpage.com

Other related items

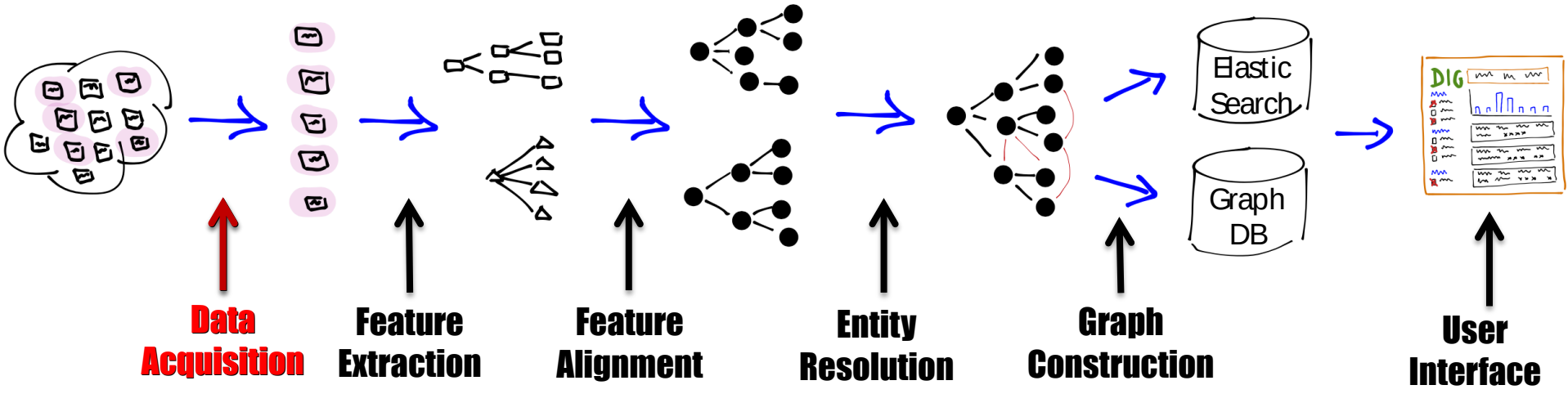
2 0 2

CO-OCCURRING EMAILS

CO-OCCURRING PHONES

2025965980	99
9162721072	1

Steps To Build a DIG



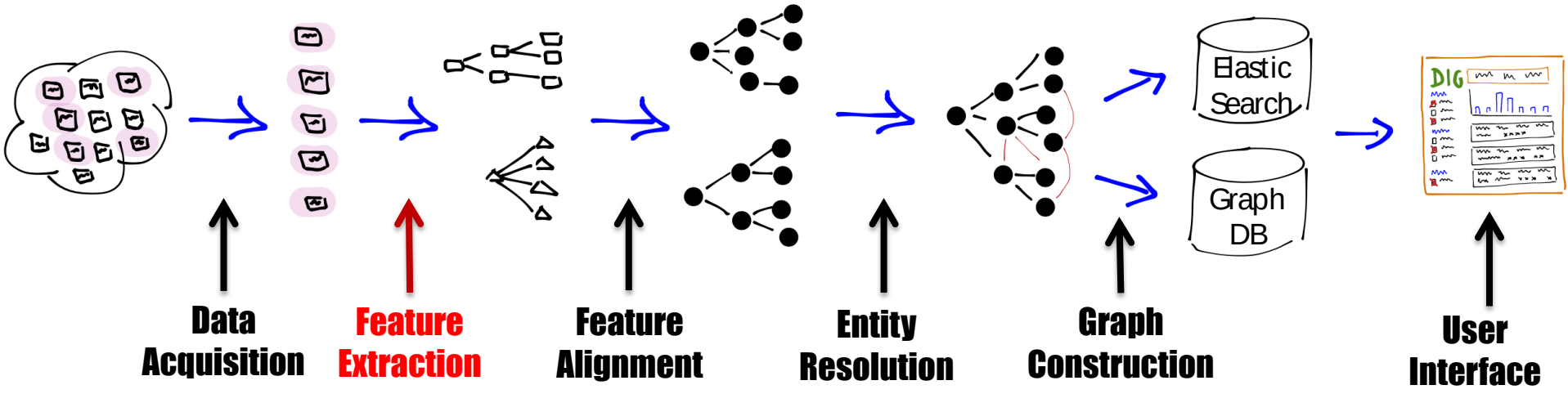
Data Acquisition

downloading relevant data

batch ◆ real-time

Web pages ◆ Web service ◆ database ◆
CSV ◆ Excel ◆ XML ◆ JSON

Steps To Build a DIG



Feature Extraction

from raw sources to structured data

- trainable text extractors
- extraction from structured Web pages
- image features
- PDF extractor

Feature Extraction from Text

"YOU don't wanna miss out on ME :) Perfect lil booty Green eyes Long curly black hair Im a Irish, Armenian and Filipino mixed princess :) ❤ Kim ❤ 7○7~7two7~7four77 ❤ HH 80 roses ❤ Hour 120 roses ❤ 15 mins 60 roses"

name:	Kim
eye-color:	green
hair-color:	black
phone:	707-727-7477
rate:	$60/15min $80/30min $120/60min

20 Examples

Mary has brunette hair and an athletic build. She can be at your door in minutes.

Ashley has great assets and a friendly demeanor that you'll love. Eye color blue. Call or text @ 213-555-9876

Name Loren Age 23 Breast Size 34 C Eye Color Blue Hair Color Brown No Games No Drama. Great Personality And Easy To Get Along With. Love To Meet New People. 100% Real Pic.

Noemi 22 Ans Hair: Noir Eyes: Green Kelly : 22 Ans 5p7 135Lbs 36C Hair : Blonde Eyes : Green Sabrinna : 21 Ans Quebecoise / European 5p7 130 Lbs

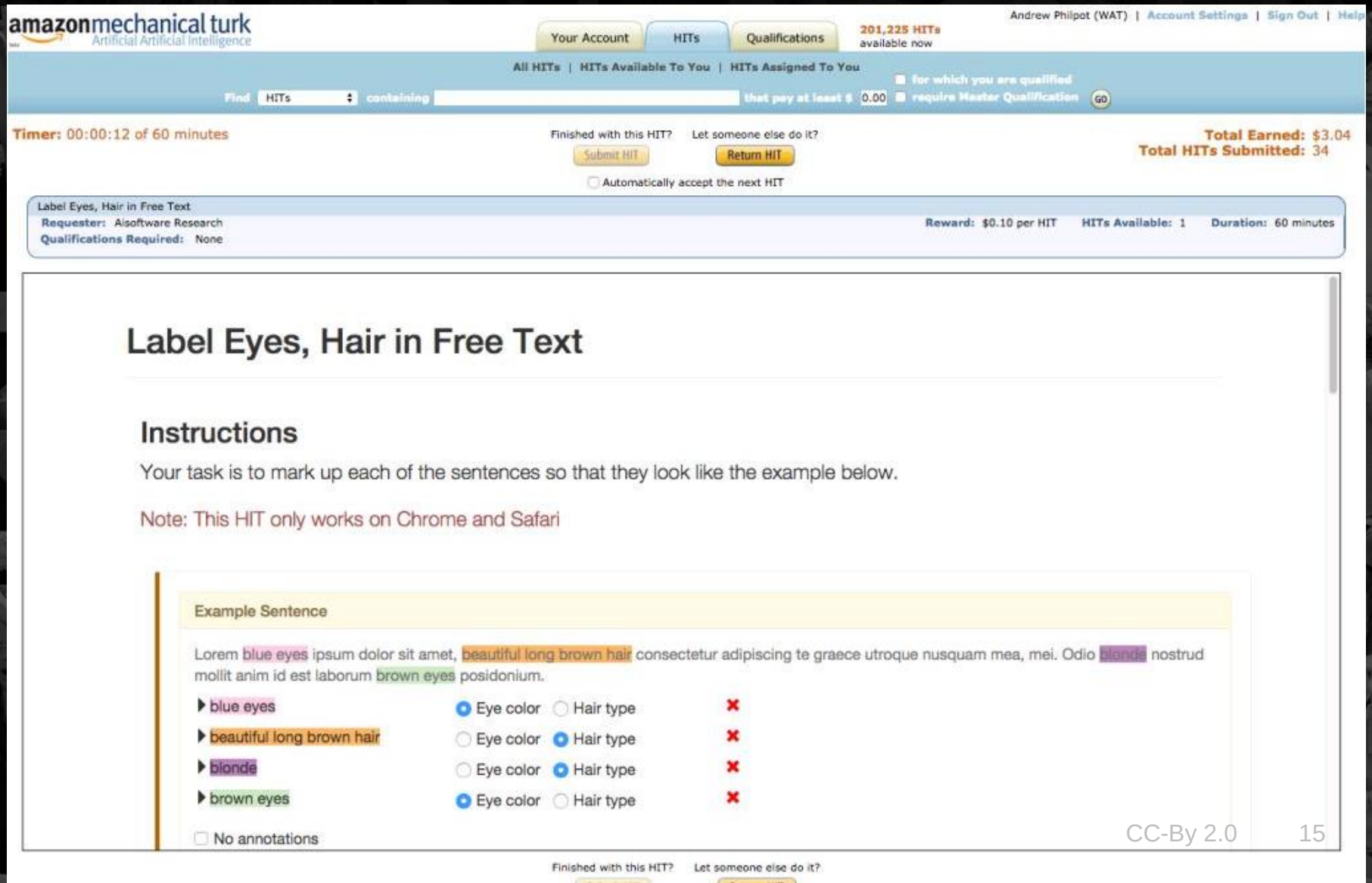
Rubie is an all natural British redhead. Justyce 5'7, 135 lbs B36 - 26 - 36. Curves in all right places. Blue eyed. 100% PROFESSIONAL.

I am a hot brunette with chestnut brown eyes living in New York. I have long curly hair and great assets, curves in right places. Call or text me.

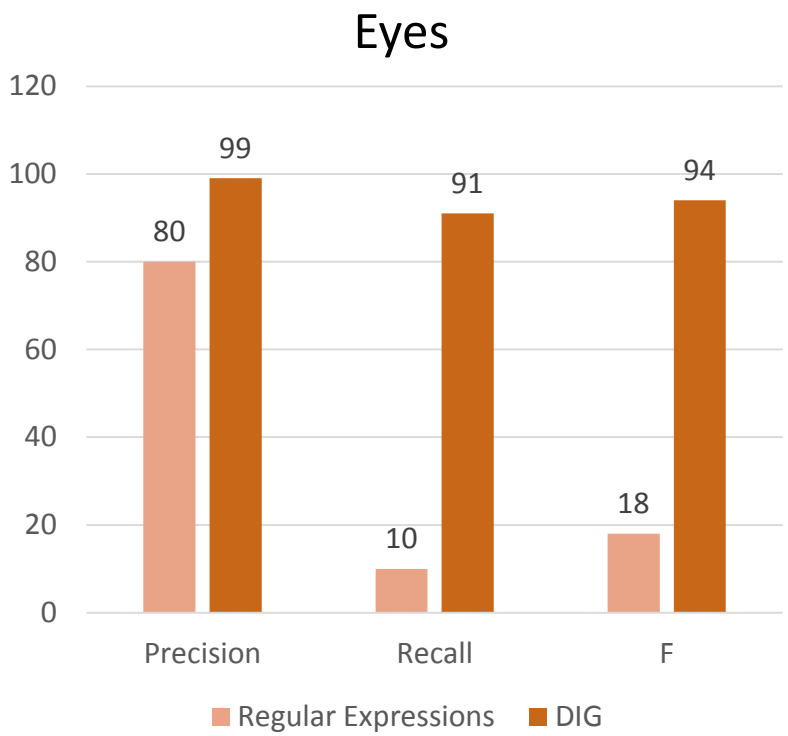
Lovely Ladies (315) 555-1234. Ashley 24 # 34c blond, green, Kayla 27 # 34b brunette, Candy 40 # 32b blonde blue, Kelly 29 # 32b blonde.

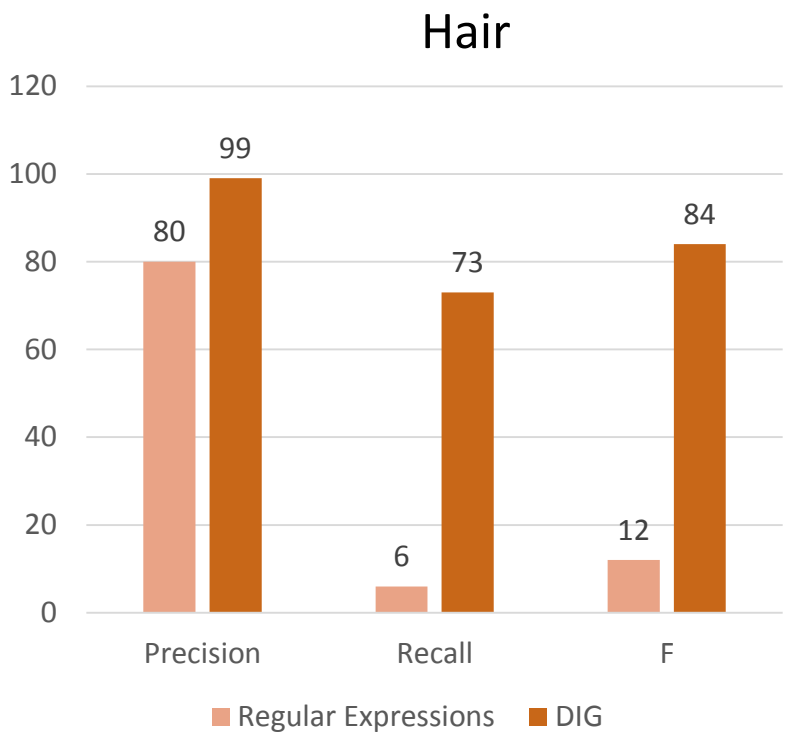
You'll love spending some quality time with me. I have long brown hair with blonde high lights and my eyes are light green. Im a great companion if u need one.

1,000's of Tasks (2 Cents/Sentence)



Performance of CRF Extractors





Structured Extraction

FOR SALE: STOEGER M3500

post id: 4700468
share:

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
Caliber:	12 Gauge
Action:	Semi-automatic
Firearm Type:	Shotgun

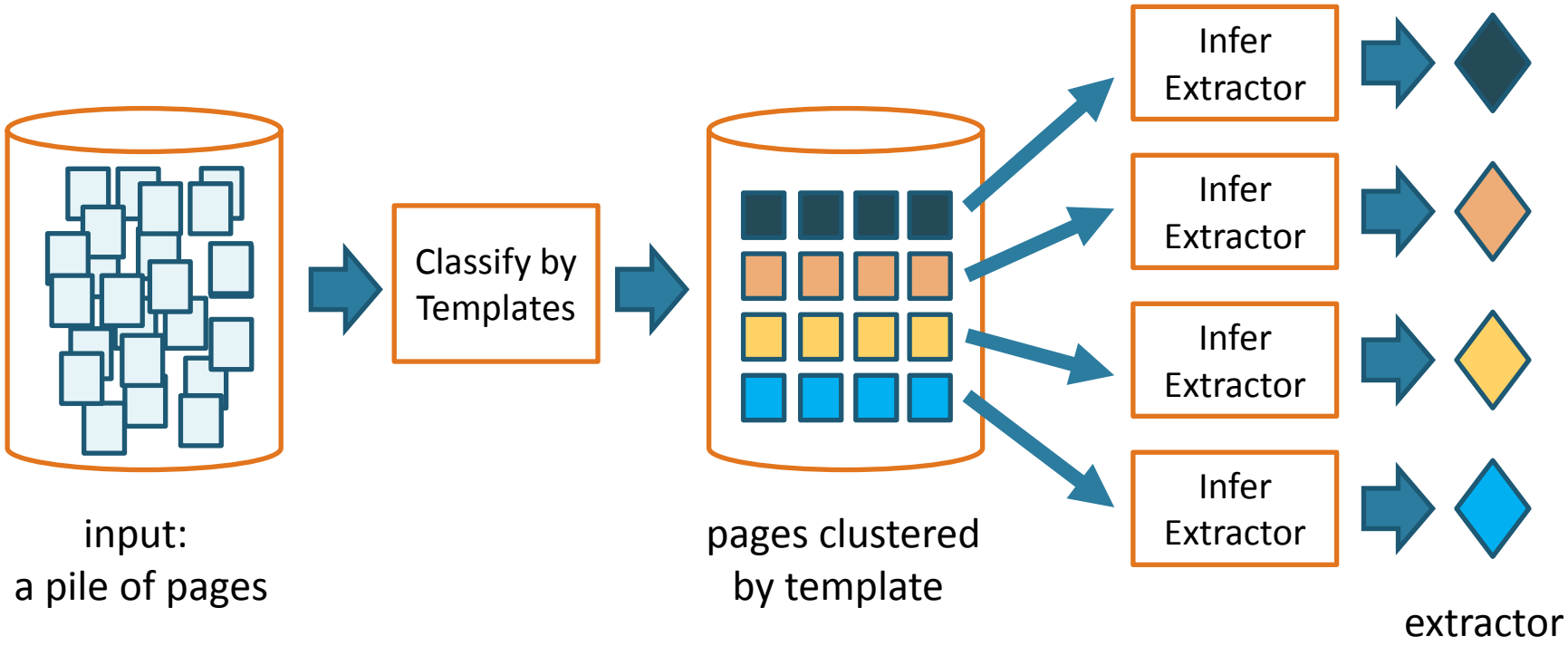
Flag | Edit | Favorite

Contact Seller

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

Automated Extraction



Unsupervised Extraction Tool

InferLink {Landmark} | tennesseegunexchange | Markup | Extraction | Downloads

+ Add Page	page_2.html	page_3.html	page_4.html	page_5.html
0036	Winchester model 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), PRO-TECH OUTDOORS cordura shoulder/belt holster, original plastic box/lock & 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 slight .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. Must 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

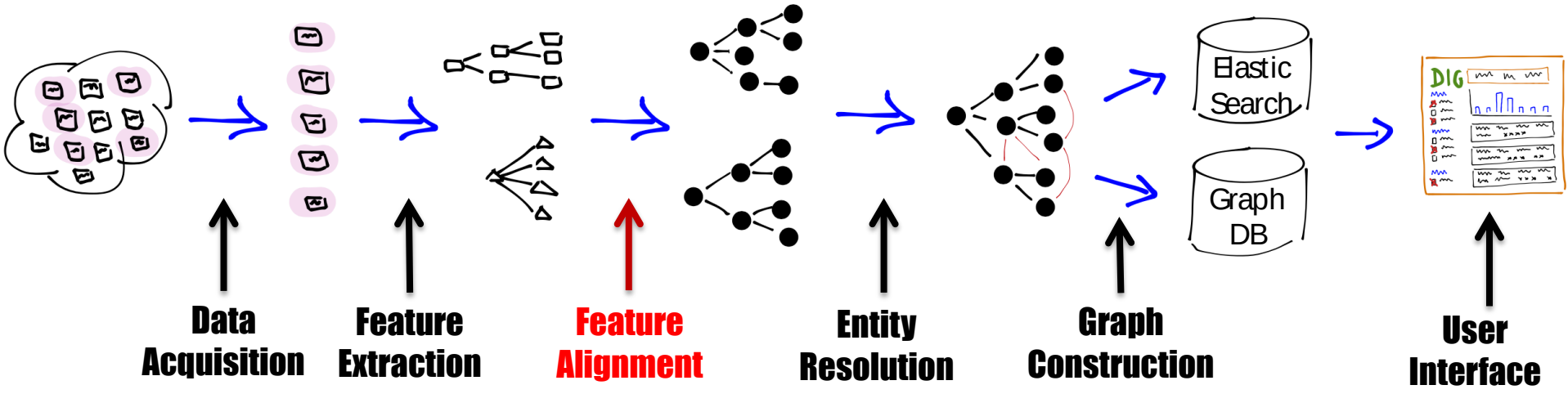
Extraction Evaluation

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)
Pretty Good	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)

Steps To Build a DIG



Feature Alignment

from multiple schemas to a common domain schema

Multiple Schemas

- CSV, Excel
- Database tables
- Web services
- Extractors
- Nomenclature
- Spelling

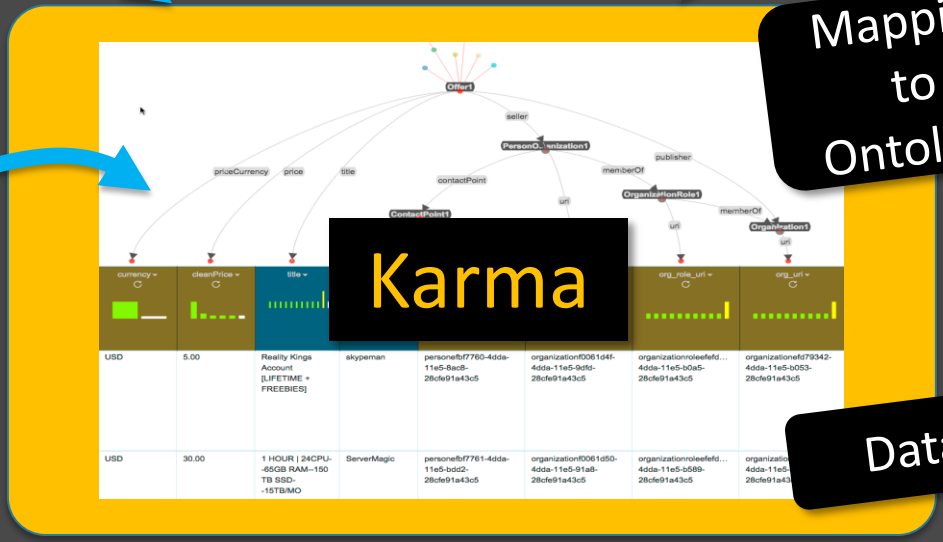
Karma: Mapping Data to Ontologies

Relational Sources

Hierarchical Sources

Services

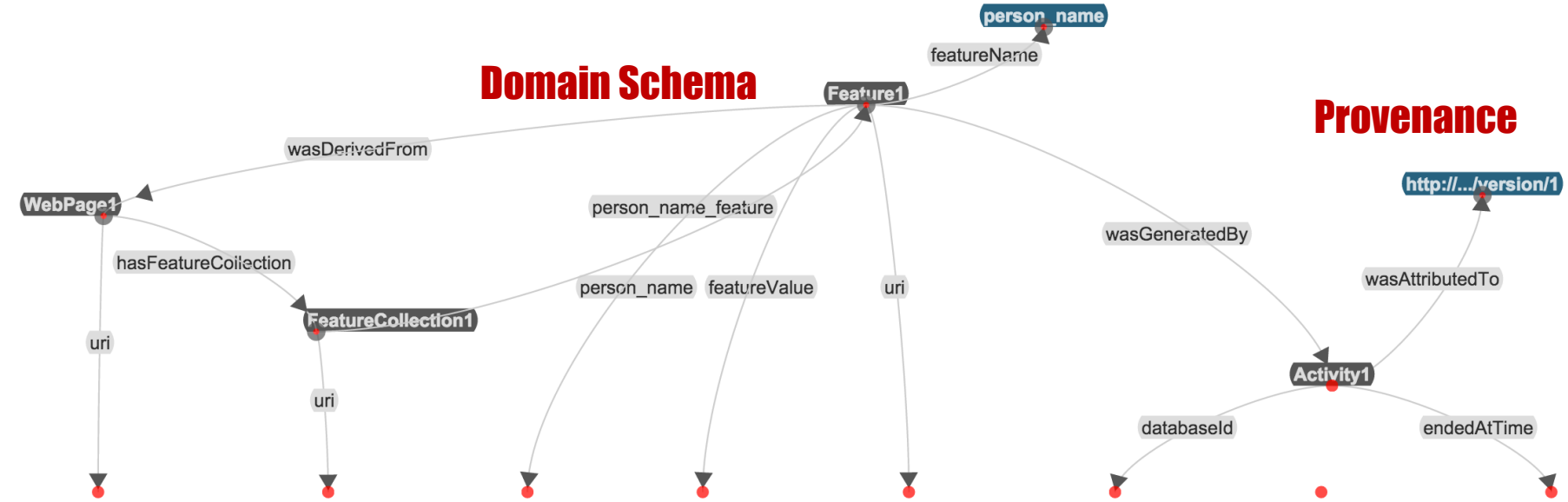
Schema.org



{ JSON-LD }

RDF

Karma Solves Feature Alignment



ad_uri	featureCollection_uri	name_clean	name_clean2	feature_uri	database_id	stanford_gentime	stanford_gentime_iso
http://memex.zapto.or…	http://memex.zapto.or…	Steph	Steph	http://memex.zapto.or…	10464463	2014-11-18	2014-11-18T00:00:00
http://memex.zapto.or…							2014-11-18T00:00:00

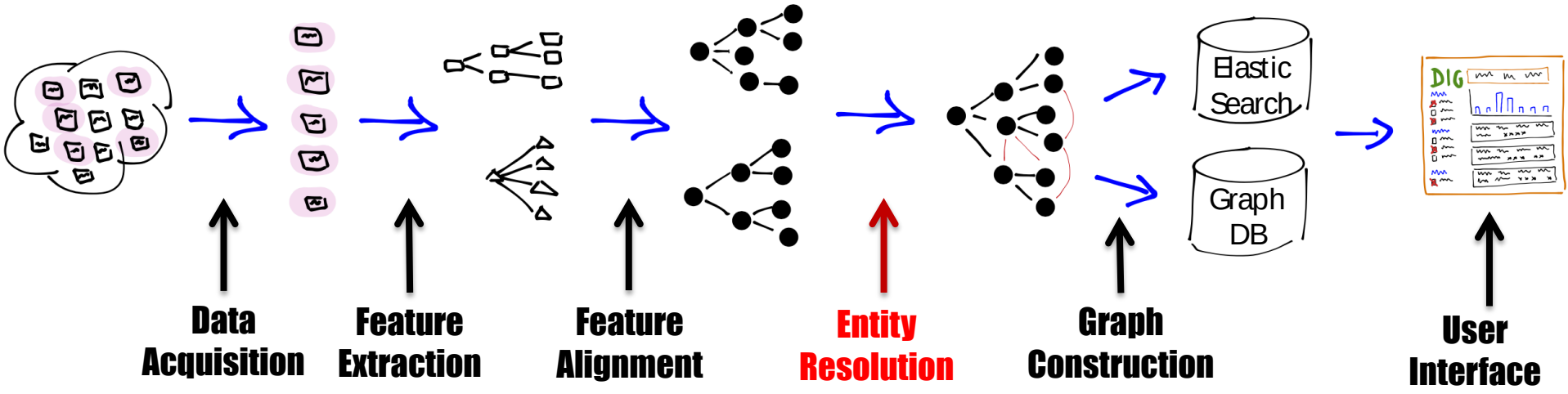
took ~30 minutes to align the output of the Stanford name extractor

Feature Alignment Statistics

- 5 contractors provided data
- ~ 15 datasets
- > 30 Karma models
- > 200 million records

- 1 hour processing in 20 node Hadoop cluster

Steps To Build a DIG



Entity Resolution

merging records that refer to the same entity

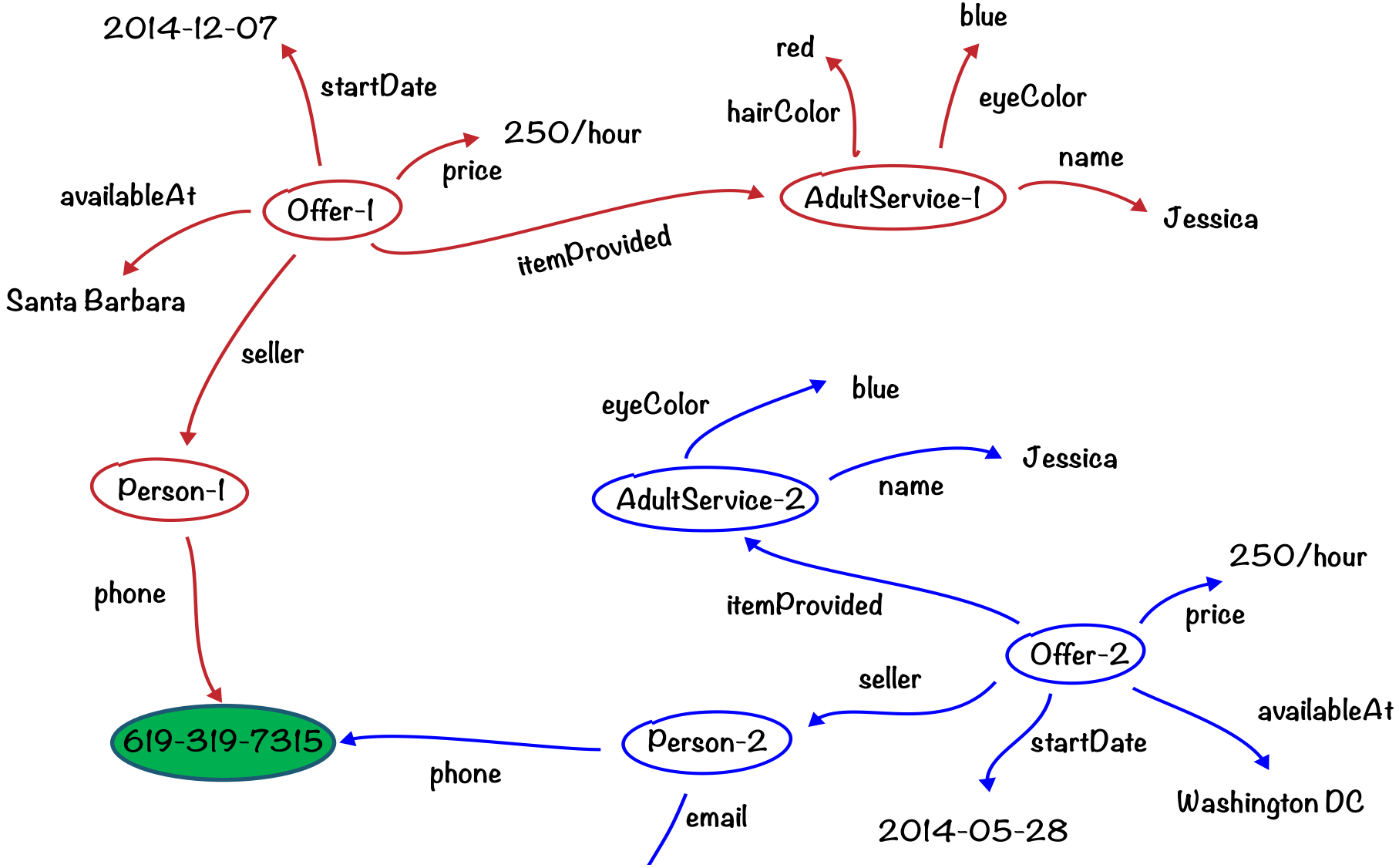
currently working on techniques to address

missing data

incorrect data

scale (~50 million records)

Entity Resolutuion on Strong Attributes



Linking Using Text Similarity

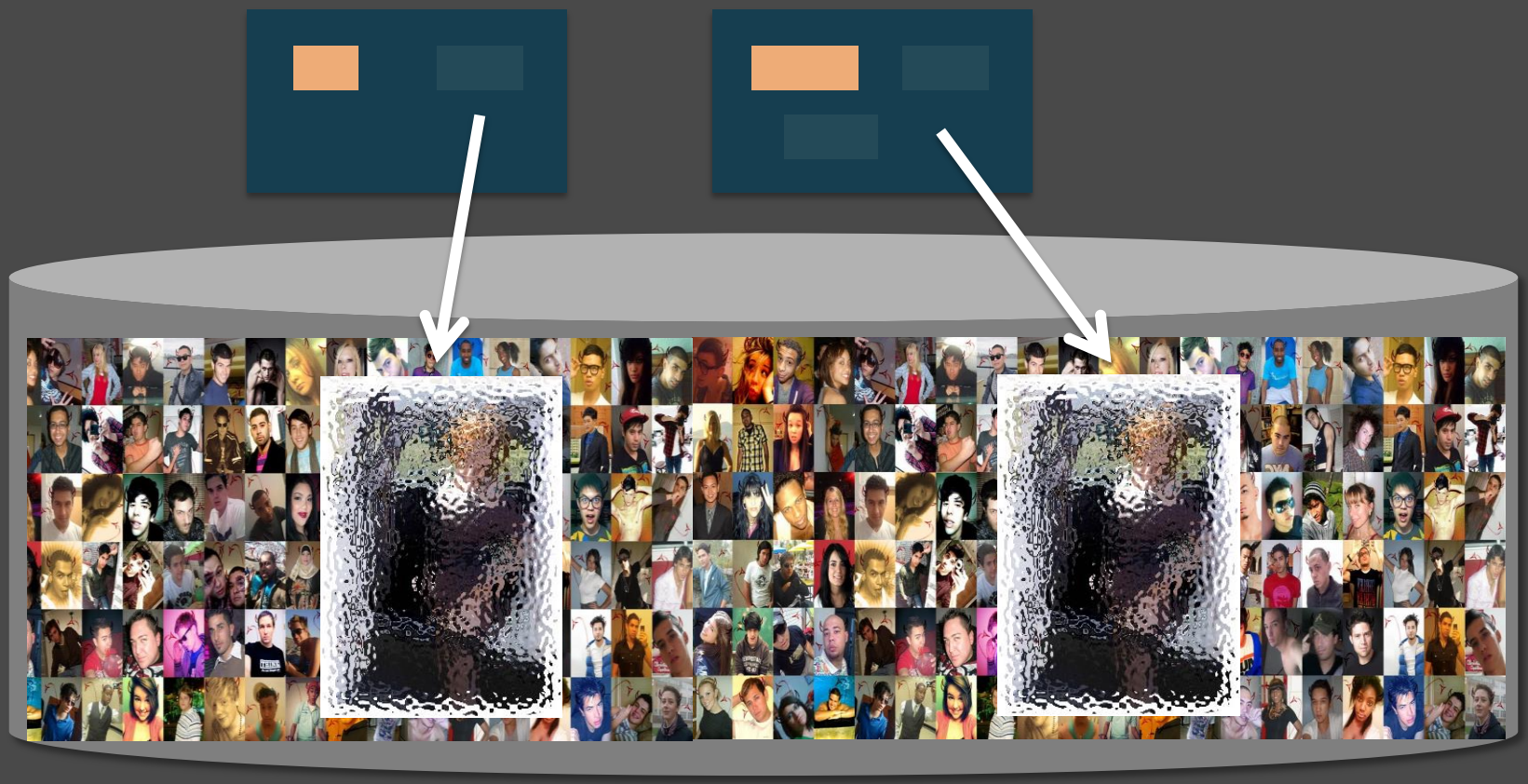
E M I L Y SEXY. ** wHiTe/lATin girl ** bUsTy SWEET. LoTs Of fUn. Call Me. O_U_T_C___A___L_L_S

L A Y L A SEXY. ** wHiTe girl ** bUsTy SWEET. LoTs Of fUn. Call Me.
O____U____T____C___A___L____L____S

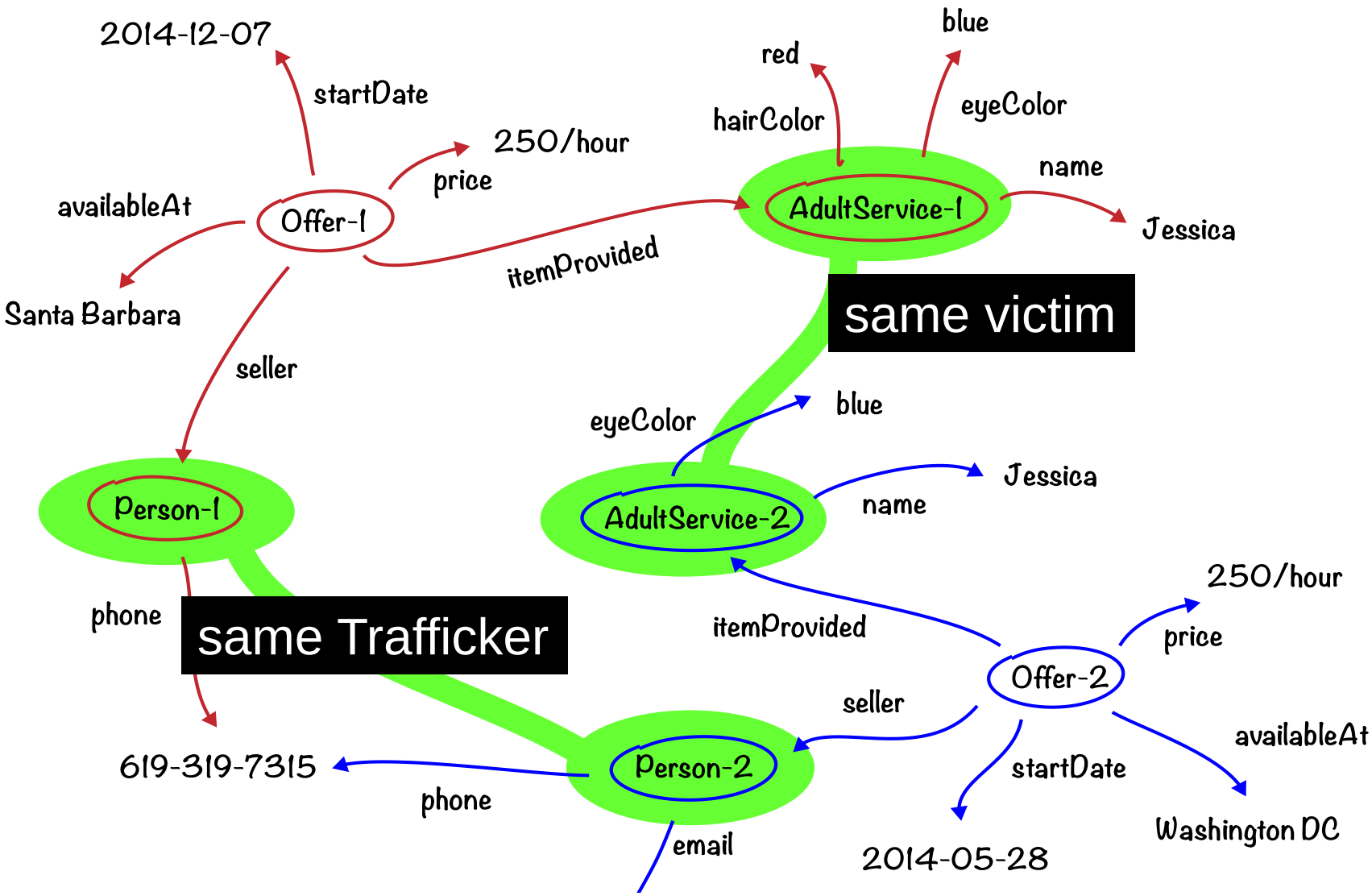
L I L A SEXY. ** WhiTe girl ** bUsTy SWEET. LoTs Of fUn. Call Me. O_U_T_C___A___L_L_S

Linking Using Image Similarity

100 Million Images Technology: Deep Learning



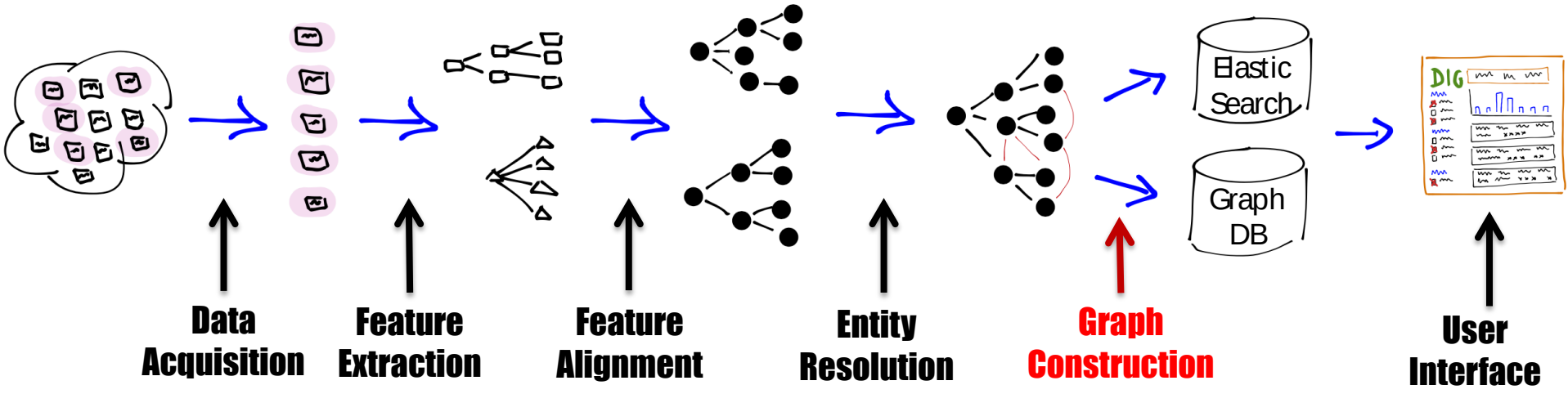
Unsupervised Collective Entity Resolution



Unsupervised Collective Entity Resolution

	Precision			Recall			Fmeasure		
	Author	Paper	Product	Author	Paper	Product	Author	Paper	Product
Limes-F	0.958	0.827	0.446	0.864	0.761	0.160	0.909	0.792	0.236
Silk-F	0.846	0.877	0.459	0.986	0.756	0.348	0.910	0.812	0.395
Gsum	0.727	0.668	0.01	0.569	0.624	0.587	0.638	0.645	0.02
CoSum-B	0.993	0.871	0.58	0.940	0.611	0.477	*0.966*	*0.718*	*0.524*
Limes-MO	0.912	0.827	0.446	0.944	0.761	0.160	0.928	0.792	0.236
Silk-MO	0.932	**0.877**	0.459	0.958	0.756	0.348	0.945	0.812	0.395
Serf	0.985	0.837	0.436	0.687	0.808	0.186	0.809	0.822	0.261
CoSum-P	**0.999**	0.771	**0.639**	**0.997**	**0.997**	**0.695**	***0.998***	***0.87***	***0.666***
Best in Literature	NA	NA	0.615 [19]	NA	NA	0.63 [19]	0.995 [5]	NA	0.622 [19]

Steps To Build a DIG



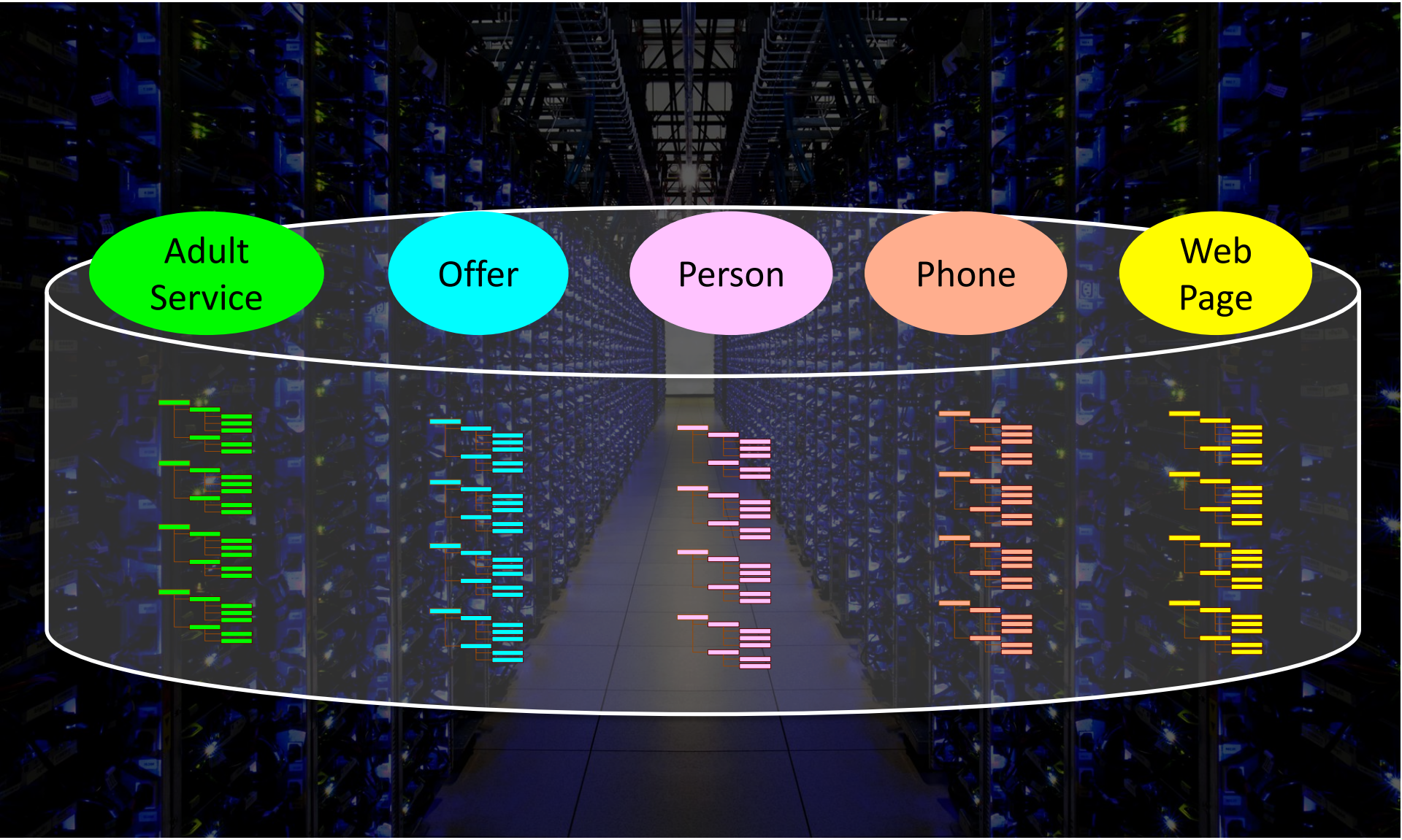
Graph Construction

assembling the data for efficient query & analysis

- ElasticSearch: scalable, efficient query
- graph databases: network analytics
- NoSQL: scalable analytics

- bulk loading: massive data imports
- real-time updates: live, changing data

Elastic Search Data Model



Indexing for High Performance Knowledge Graph Queries

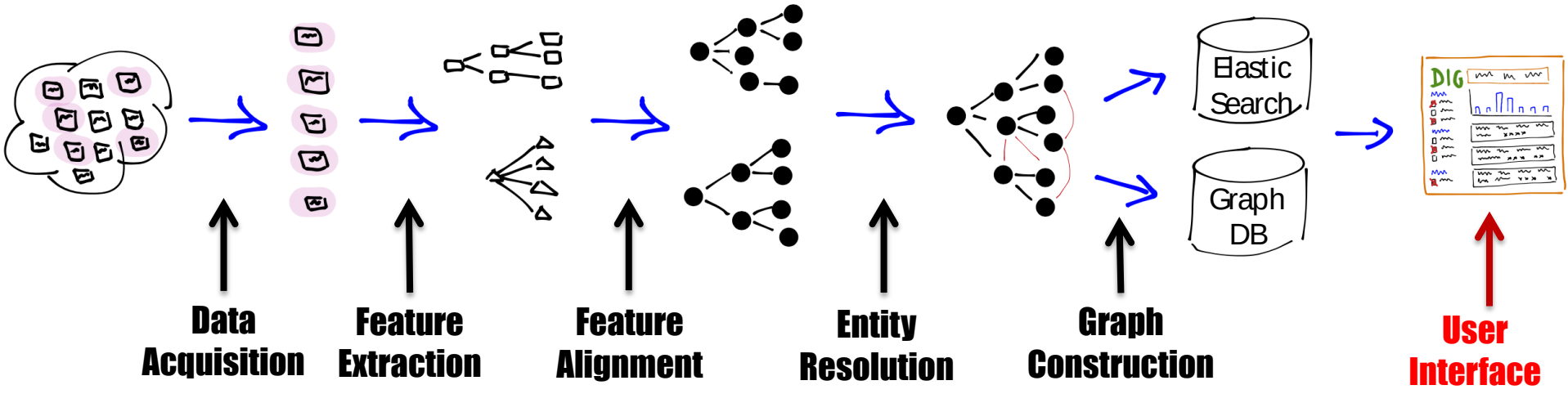
Avg. Query Times in Milliseconds
Single User Query Load
1.2 billion triples

State of the Art Graph Database (RDF)

Database	Keyword	Facet	Facet (Missing)	Click Search	Anchored	Click Viz
Virtuoso 7	4,649	13,368	12,402	25,891	6,778	27,452
ES 1.7.3 Standalone	70	80	79	124	3,565	3,768
ES 1.7.3 Cluster (5)	469	366	363	496	850	1,415
ES 1.7.3 Cluster (20)	108	105	106	148	1,413	1,510

DIG indexing deployed in ElasticSearch

Steps To Build a DIG



DIG

Enter Keywords

jessica red hair blue eyes swedish

Sort Results By

Best Match

Country

View All Sort By:

- United States (7020194)
- Canada (1275456)
- Australia (214762)
- England (113345)
- China (68904)
- South Africa (42371)
- India (39210)
- United Arab Emirates (38903)
- Turkey (27059)
- Hong Kong (18588)

City

View All Sort By:

- Toronto, Ontario (227627)
- New York, New York (189764)
- Ontario, California (186880)
- Washington, District of Columbia (136296)
- Chicago, Illinois (123528)
- Los Angeles, California (117045)
- San Diego, California (115255)
- Calgary, Alberta (102367)
- Denver, Colorado (102332)
- Houston, Texas (101766)

Telephone

View All Sort By:

- 4089077677 (12543)
- 4157382275 (11913)

25 of 10999713 Results

619-319-7315 - Escort ad in San Diego, California | Highly Reviewed Swedish...

escortphonelist.com July 29, 2013 6193197315

619-319-7315 - Escort ad in San Diego, California | Highly Reviewed Swedish...

escortphonelist.com July 29, 2013 6193197315

619-319-7315 - Escort ad in San Diego, California | Highly Reviewed Swedish...

escortphonelist.com July 29, 2013 6193197315

Perhaps a Swedish treat? Silky white skin... - san diego escorts - backpage.com

backpage.com August 27, 2014

Url: http://sandiego.backpage.com/FemaleEscorts/perhaps-a-swedish-treat-silky-white-skin-35/16983107

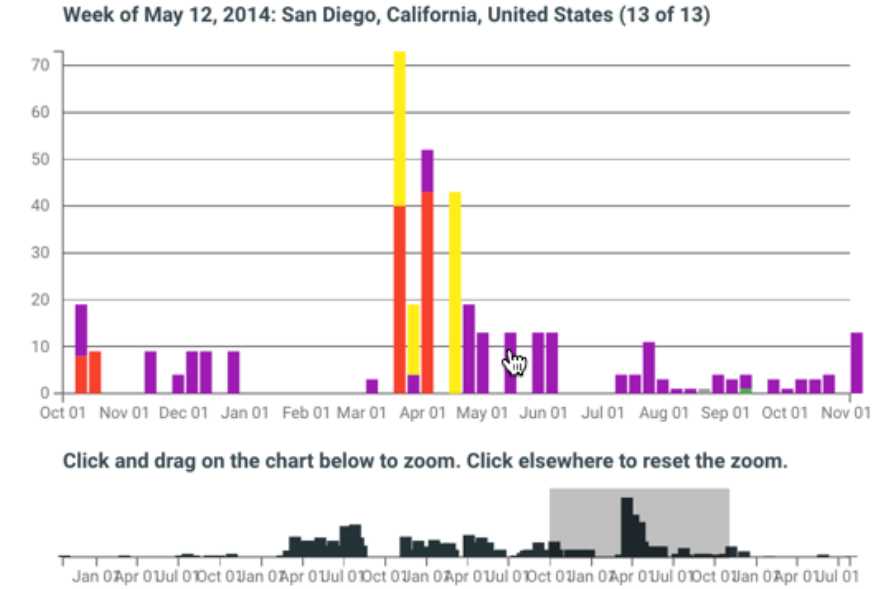
Body: I am #26 on top 100 list of the best providers in San Diego. I have a classy private upscale incall with fireplace, pool and jacuzzi. I am from Sweden. M.Sc . in engineering. Great reviews! Red hair and blue eyes. Perky D breasts and nice nipples. Incall 1 hour $250.00 Text me at: 1 619 319 7315 xoxo Jessica

Addresses: San Diego, California, United States

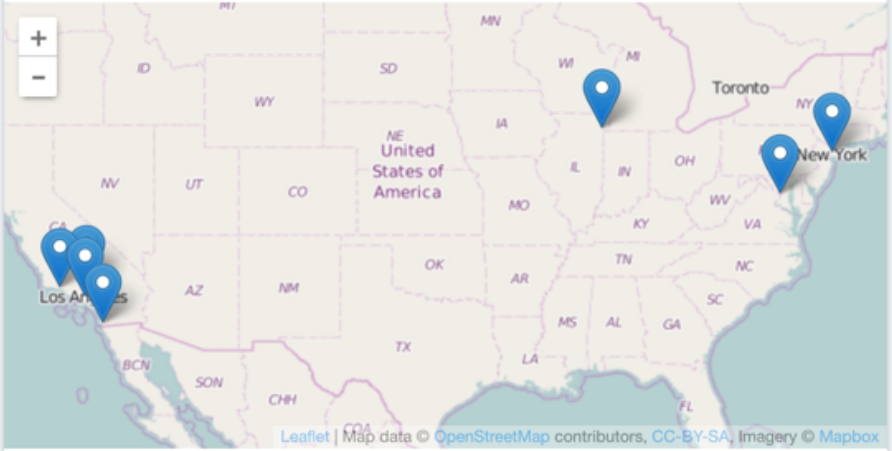
SHOW MORE RESULTS

Phone: 619-319-7315

Ads from Week of October 18, 2010 to Week of July 27, 2015



7 Locations



OFFERS PER CITY

San Diego, California	806
Palmdale, California	163
Santa Barbara, California	137
Los Angeles, California	34
Chicago, Illinois	19
Washington, District of Columbia	8
Manhattan, New York	1

Summary of Provider Attributes

NAME		AGE		ETHNICITY		HAIR COLOR	
jessica	187	33	478	caucasian	2	red	276
eve	3	34	239				
		32	193				
		35	74				

DIG Deployment for Human Trafficking

- 100 million Web pages
- Live updates (~5,000 pages/hour)
- ElasticSearch database (7 nodes)
- Hadoop workflows (20 nodes)

- District Attorney
- Law Enforcement
- NGOs

Deployed to 6 Law Enforcement Agencies and Successfully Used to Prosecute Traffickers

DIG Applications

Human Trafficking

large, real users

Material Science Research

70,000 paper abstracts (built in 1 week)

Arms Trafficking

Identify illegal sales

Patent Trolls

Identify patent trolls

Cyber Attacks

Predict cyber attacks from dark web data

Conclusions

- Complete tool-chain to build domain-specific knowledge graphs
- Integrates heterogeneous data: web pages, databases, CSV, web APIs, images, etc.
- Scales to ~100 million pages, ~3 billion facts
- Deployed to law enforcement

Questions?

dig.isi.edu

Open Source, Apache 2 License