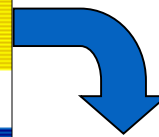


Wrap-up

# Part 1

Web IE, Wrappers and Information Integration using Karma

# Extracting Data from Semi-structured Sources

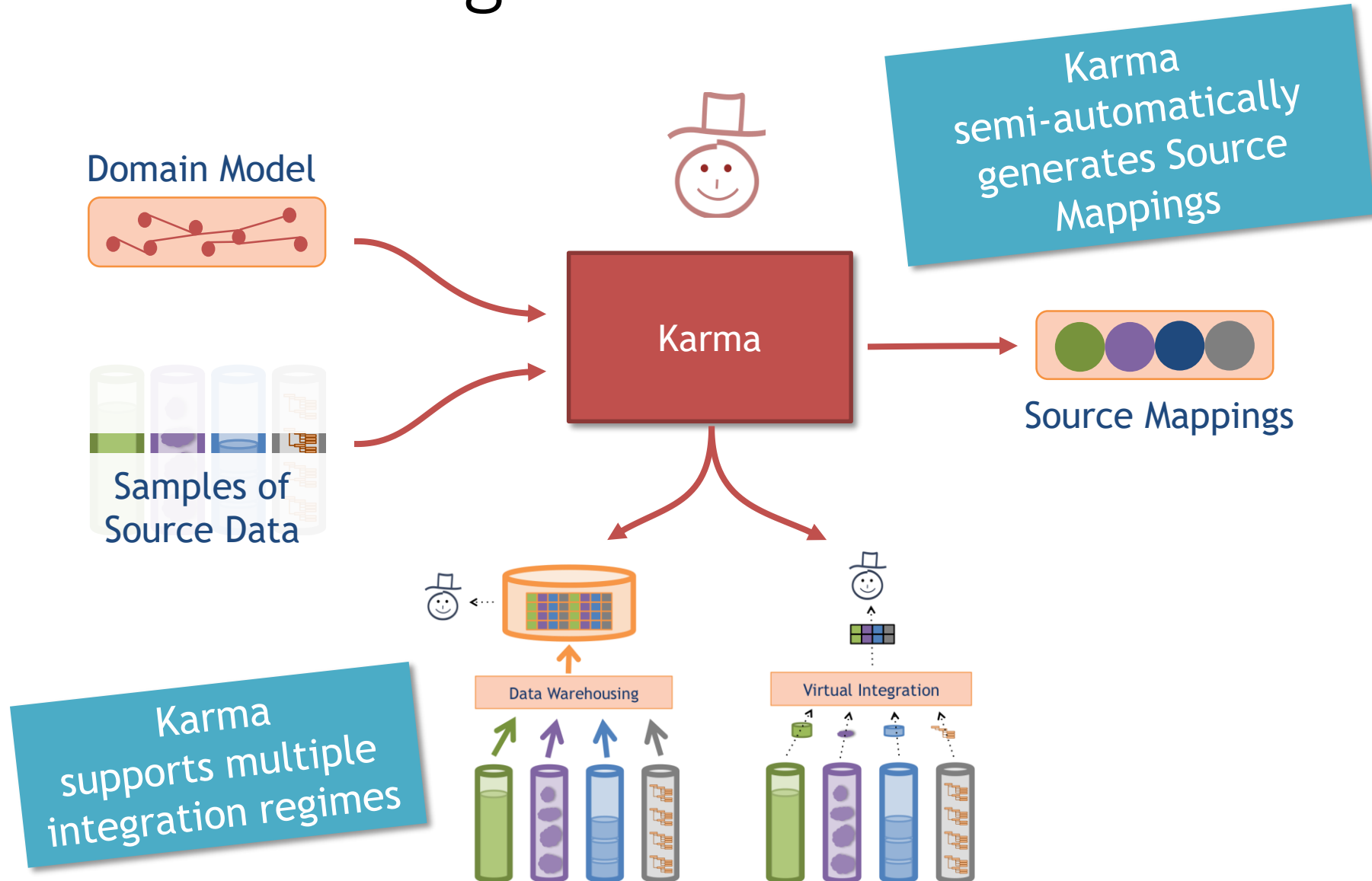


<b>NAME</b>	Casablanca
Restaurant	
<b>STREET</b>	220 Lincoln Boulevard
<b>CITY</b>	Venice
<b>PHONE</b>	(310) 392-5751

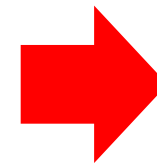
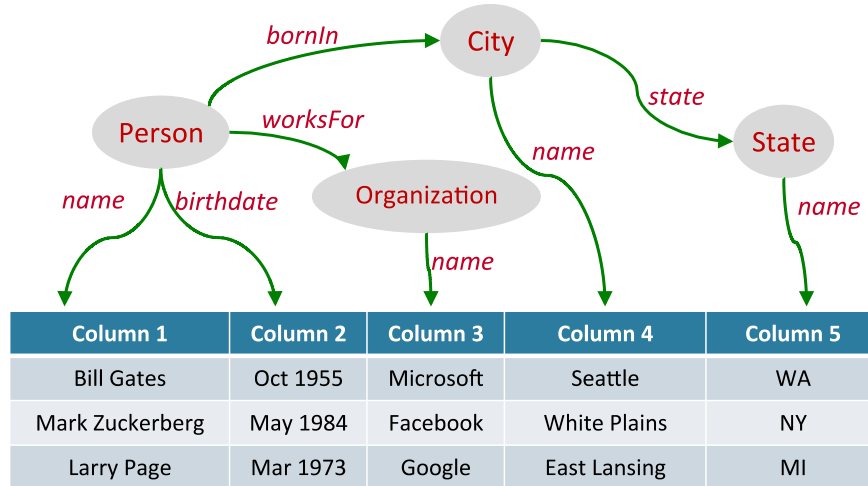
# Approaches to Wrapper Construction

- Manual Wrapper Construction
- Learning-based Wrapper Construction
- Automatic Wrapper Construction
  - Grammar learning using Roadrunner
  - Clustering and learning the structure of the clustered pages using the Inferlink tool

# Information Integration in Karma



Karma **semi-automatically builds** semantic models



Knowledge  
Graphs

Karma uses **semantic models** to **create** knowledge graphs

# Part 2

Information Extraction from 'unstructured' data

# Document Features

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.











Grammatical sentences plus some formatting & links

**Dr. Steven Minton** - Founder/CTO  
 Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

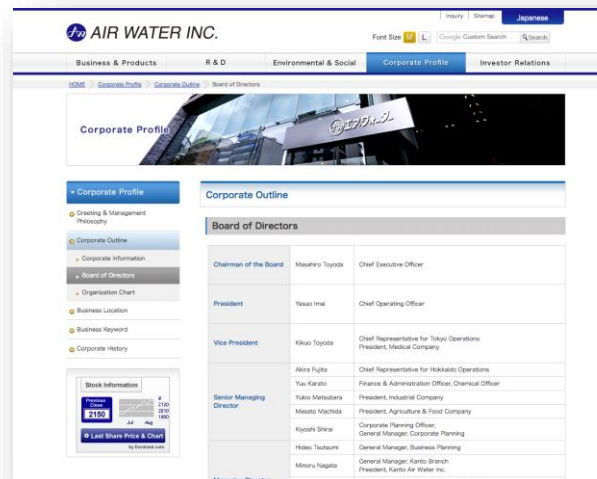
- Press
- **Contact**
- General information
- Directions maps

**Frank Huybrechts** - COO  
 Mr. Huybrechts has over 20 years of

Non-grammatical snippets, rich formatting & links

<b>Barto, Andrew G.</b>	(413) 545-2109	<a href="mailto:barto@cs.umass.edu">barto@cs.umass.edu</a>	CS276
Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			 
<b>Berger, Emery D.</b>	(413) 577-4211	<a href="mailto:emery@cs.umass.edu">emery@cs.umass.edu</a>	CS344
Assistant Professor.			 
<b>Brock, Oliver</b>	(413) 577-0334	<a href="mailto:oli@cs.umass.edu">oli@cs.umass.edu</a>	CS246
Assistant Professor.			 
<b>Clarke, Lori A.</b>	(413) 545-1328	<a href="mailto:clarke@cs.umass.edu">clarke@cs.umass.edu</a>	CS304
Professor. Software verification, testing, and analysis; software architecture and design.			 
<b>Cohen, Paul R.</b>	(413) 545-3638	<a href="mailto:cohen@cs.umass.edu">cohen@cs.umass.edu</a>	CS278
Professor. Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.			 

Tables



Board of Directors		
Chairman of the Board	Masahiro Toyoda	Chief Executive Officer
President	Yasuo Imai	Chief Operating Officer
Vice President	Kiisaku Toyoda	Chief Representative for Tokyo Operations President, Medical Company
	Akira Fujita	Chief Representative for Hokkaido Operations
	Yuu Kuroki	Finance & Administration Officer, Chemical Officer
	Takao Masahara	President, Industrial Company
	Masato Machida	President, Agriculture & Food Company
	Kiyoshi Shira	Corporate Planning Officer
	Hideo Tachibana	General Manager, Business Planning
	General Manager, Korea Branch	
	Minoru Nagata	President, Korea Air Water Inc.
		Chief Representative for South America

Charts





# Scope

Web site specific

Genre specific  
(e.g., forums)

Wide, non-specific

InvestorPlace  
DOW 22,049 +0.67% NASDAQ 6,352 +0.05% S&P 500 2,474 +0.05%

### CHARTER COMM RG-A (CHTR)

400.90 -11.15 (2.86%) 20:10 EDT

CHTR STOCK QUOTE DELAYED 20 MINUTES

NAVELLIER RATINGS  
Rating: Buy  
Total Grade: B

Analysis Breakdown  
CHTR STOCK GRADE: B  
Fundamental Grade: D  
Quantitative Grade: A

CHTR Earnings  
Earnings Growth: F  
Earnings Momentum: F  
Earnings Surprises: F  
Analyst Earnings Revisions: D

CHTR Financial Information  
Sales Growth: A

CHTR Stock Chart  
Historical CHTR Prices

Dividend & Yield: N/A (N/A)  
PIE: -  
Market Cap: 103.30B  
EPS: -2.05  
Volume: 4M  
Day's Range: 386.38 - 408.83  
52wk Range: 236.06 - 408.83

InvestorPlace  
DOW 22,049 +0.67% NASDAQ 6,352 +0.05% S&P 500 2,474 +0.05%

### LEGACY VENTURES INTL, Inc. (LGVV)

6.01 0.00 (0.00%) 07/14/17

LGVV STOCK QUOTE DELAYED 20 MINUTES

NAVELLIER RATINGS  
Rating: Buy  
Total Grade: B

Analysis Breakdown  
LGVV STOCK GRADE: B  
Fundamental Grade: D  
Quantitative Grade: A

LGVV Earnings  
Earnings Growth: F  
Earnings Momentum: F  
Earnings Surprises: F  
Analyst Earnings Revisions: D

LGVV Financial Information  
Sales Growth: A

LGVV Stock Chart  
Historical LGVV Prices

Dividend & Yield: N/A (N/A)  
PIE: -  
Market Cap: 391.03K  
EPS: -38648.00  
Volume: 67  
Day's Range: 6.01 - 6.01  
52wk Range: 1.05 - 15.00

TheLion.com  
Home Forum Portfolio Blog User Mail Help

Welcome Stranger! Please sign up or log in to enable additional features.

### Forum - TheLion.com Central

Return Top | Return List | Reply Thread | Search

From: madrid (Sec. 0) Date: 2017-02-01 05:00:12  
Forum: TheLion.com Central - Thread #87303331  
Mag #552 - Part 1/2 (Rec: 0)

Message: Endorse | Reply | Favorite | Bookmark | Report Abuse | User madrid: Reward | Watch | Ignore

Upvoted  
http://www.4traders.com/forums/stock-message-board/nyse-nasdaq-amex/htz-hertz-global-holdings

HTZ - Hertz Global Holdings  
Discussion in 'Stock Message Boards NYSE, NASDAQ, AMEX' started by StockJock, May 9, 2016.

InvestorsHub  
When it comes to safety, THE BOC NEVER BLINKS (Robbery Optimization Droids)

Boards | Hot | Tools | Streamer | Level 2 | Follow Feed

Opportunity is Everywhere if you know where to look. Get Started at E\*TRADE

### Home > Boards > Free Zone > Cryptocurrency Groups > Bitcoin, Ethereum, Cryptocurrencies

Public Reply | Private Reply | Keep | Last Read

gfb272 Wednesday, 08/09/17 04:34:52 PM  
Re: None  
Post # 107 of 107 Go

Jim Rickards ->>> Is Bitcoin Money?

By James Rickards  
August 8, 2017  
https://dailyreckoning.com/is-bitcoin-money/

Is Bitcoin Money?

AIR WATER INC.  
Business & Products R & D Environmental & Social Corporate Profile Investor Relations

Corporate Profile

Corporate Outline

profitspi  
Classic Home Stock Quote Stock Charts Watch List Portfolio Tracking Stock Screening Backtesting Sign Up Support

Switch to Interactive charts  
Symbol(s) or Description: QQQT  
Get Chart Search QQQT Stock Quote

Show Symbols from: 8/9/2017 46.49 -0.43 -0.93%

GOOG MSFT  
Multiple Charts: Line

4-traders  
MARKETS NEWS ANALYSIS STOCK PICKS PORTFOLIOS SCREENERS WATCHLISTS TOP / FLOP OUR SERVICES

Sales by Business

2016	2017	Delta	
Home Improvement	18,311 57.4%	18,281 55.7%	-0.16%
WLD-1	7,512 22%	7,056 21.9%	-0.64%
Shop	5,806 19.3%	6,228 19.2%	+5.82%
Store Development	698.36 2.1%	683.85 2.1%	-0.53%
Other	26.23 0.1%	23.96 0.1%	-18.16%

Sales by Region

2016	2017	Delta	
Japan	30,841 96.7%	31,199 96.7%	+1.16%

Managers

Name	Age	Since	Title
Shozo Hasegawa	68	1979	President, Representative Director & GM-Sales
Toshitaki Tanihara	60	1984	Senior Managing Director & GM-Administration
Kazumasa Inaba	52	1989	MD, Head Compliance & Manager Internal Audit
Hiroyuki Uemura	66	1984	Managing Director & Manager Store Development
Yoshihiko Kobayashi	72	2006	Independent Outside Director
Ichiro Otagiri	55	1986	Director & Manager Home Center Business
Chihiro Fujimura	58	2016	Independent Outside Director
Katsuhiko Tatemaki	-	-	Manager Personnel & General Affairs
Ryoji Ikeda	62	1979	Auditor

Equities

Share A	Quantity	Price	Company-owned shares	Total Price			
Share A	0.001	16,100,000	3,486,184	21.7%	1,882,024	19.4%	21.7%

Shareholders

Company Name	Share A	Share B
Kansei Co., Ltd.	100%	0%

# Pattern Complexity

E.g., word patterns

## Closed set

U.S. states

He was born in Alabama...

The big Wyoming sky...

## Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

## Complex pattern

U.S. postal addresses

University of Arkansas  
P.O. Box 140  
Hope, AR 71802

Headquarters:  
1128 Main Street, 4th Floor  
Cincinnati, Ohio 45210

## Ambiguous patterns, needing context and many sources of evidence

Person names

...was among the six houses sold by Hope Feldman that

Pawel Opalinski, Software Engineer at WhizBang Labs.

“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 ♥ 7o7~7two7~7four77 ♥ HH 80 roses ♥ Hour 120 roses ♥ 15 mins 60 roses”

# Practical Considerations

- How good (precision/recall) is necessary?
  - High precision when showing extractions to users
  - High recall when used for ranking results
- How long does it take to construct?
  - Minutes, hours, days, months
- What expertise do I need?
  - None (domain expertise), patience (annotation), simple scripting, machine learning guru
- What tools can I use?
  - Many ...

# myDIG: A KG Construction Toolkit

Python, MIT license, <https://github.com/usc-isi-i2/dig-etl-engine>

- **Enable end-users to construct domain-specific KGs**
  - end users from 5 government orgs constructed KGs in less than one day
- **Suite of extraction techniques**
  - semi-structured HTML pages, glossaries, NLP rules, NER, tables (coming soon)
- **KG includes provenance and confidences**
  - enable research to improve extractions and KG quality
- **Scalable**
  - runs on laptop (~100K docs), cluster (> 100M docs)
- **Robust**
  - Deployed to many law enforcement agencies
- **Easy to install**
  - Docker deployment with single “docker compose up” installation

# Part 3

Knowledge Graph Completion

# What is knowledge graph completion?

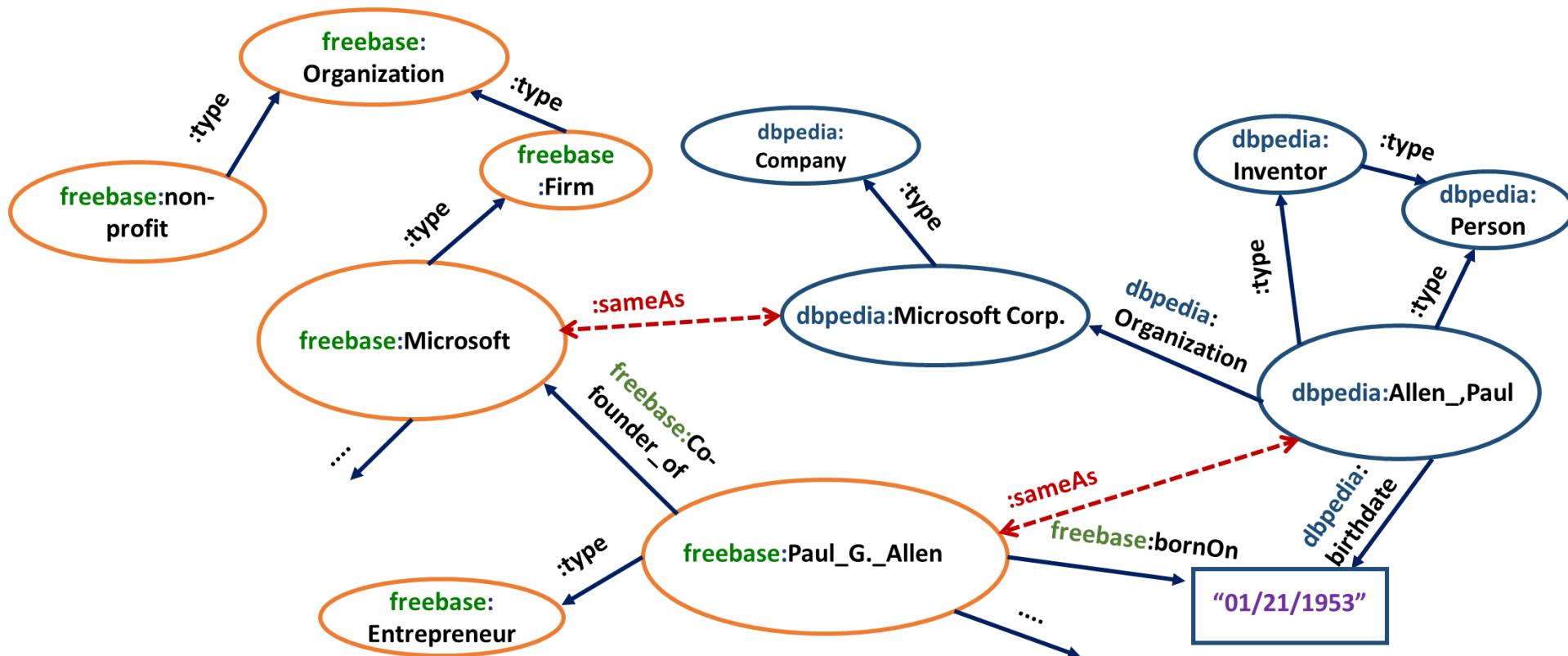
- An 'intelligent' way of doing data cleaning
  - Deduplicating entity nodes (**entity resolution**)
  - Collective reasoning (**probabilistic soft logic**)
  - **Link prediction**
  - Dealing with **missing values**
  - Anything that improves an existing knowledge graph!
- Also known as **knowledge base identification**

# Some solutions we covered

- Entity Resolution (ER)
- Probabilistic Soft Logic (PSL)
- Knowledge Graph Embeddings (KGEs), with applications

# Entity Resolution (ER)

- The **algorithmic** problem of **grouping** entities referring to the **same** underlying entity





# Extraction Graph+Ontology + ER+PSL

## Uncertain Extractions:

- .5: Lbl(Kyrgyzstan, bird)
- .7: Lbl(Kyrgyzstan, country)
- .9: Lbl(Kyrgyz Republic, country)
- .8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

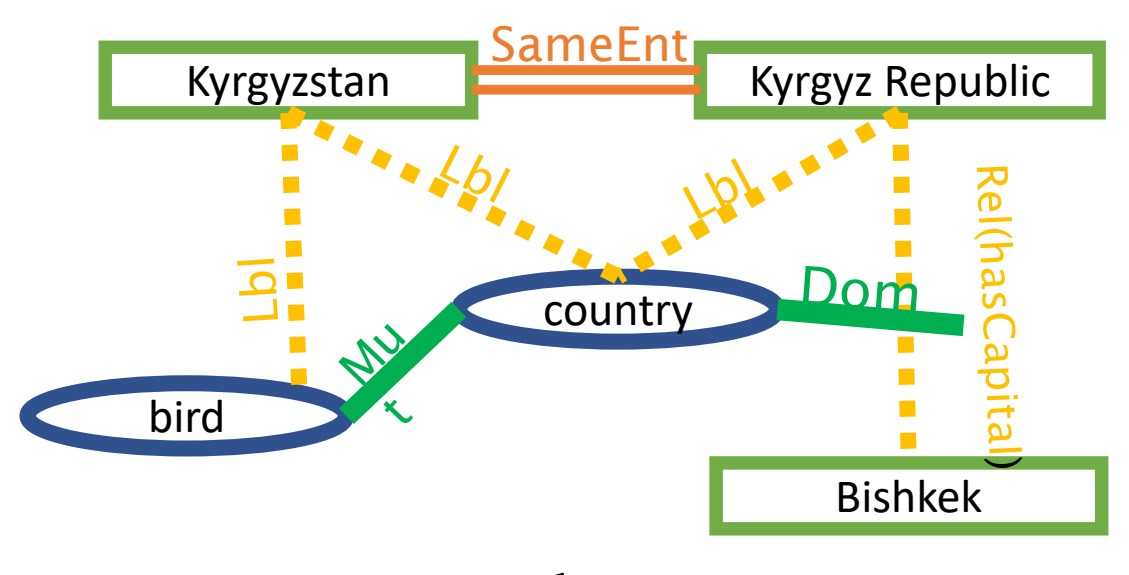
## Ontology:

- Dom(hasCapital, country)
- Mut(country, bird)

## Entity Resolution:

- SameEnt(Kyrgyz Republic, Kyrgyzstan)

## (Annotated) Extraction Graph

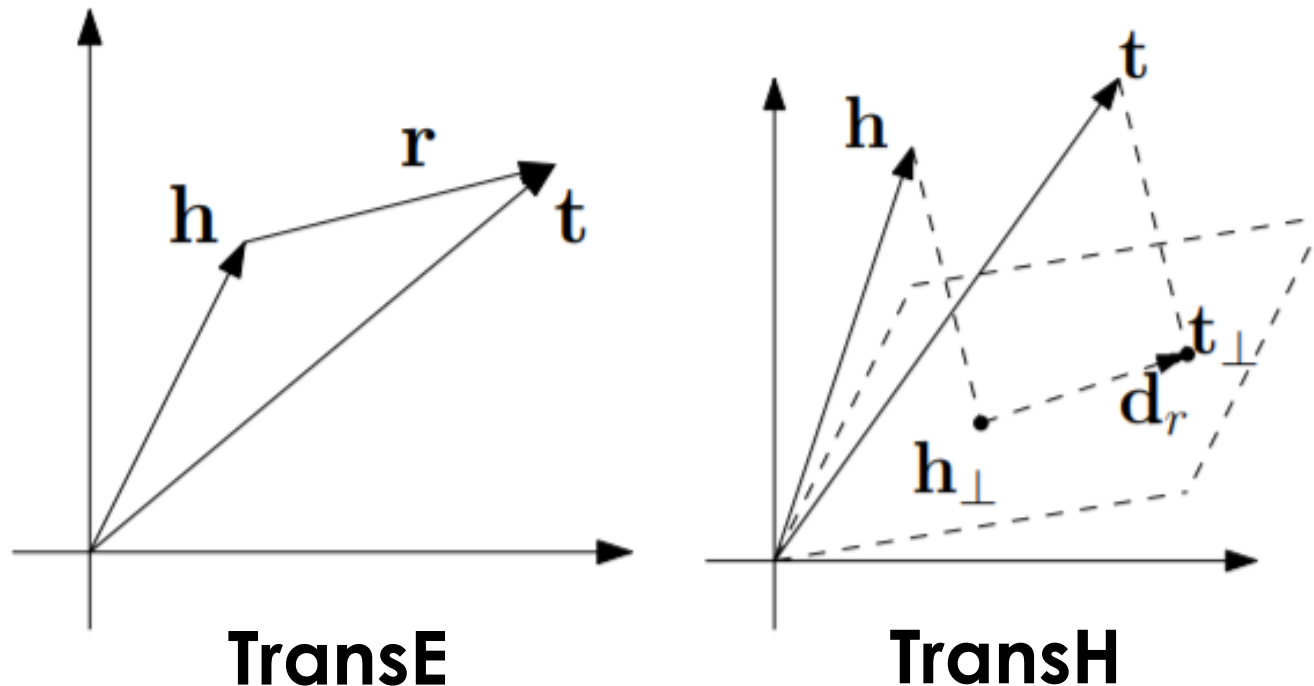


## After Knowledge Graph Identification



# Knowledge graph embeddings

- Many ways to **model** the problem: entities are usually **vectors**, relations could be **vectors** or **matrices**



# Objective/loss/energy functions

- What is an ‘optimal’ vector/matrix for an entity or relation?

Model	Score function $f_r(\mathbf{h}, \mathbf{t})$	# Parameters
TransE (Bordes et al. 2013b)	$\ \mathbf{h} + \mathbf{r} - \mathbf{t}\ _{\ell_1/2}, \mathbf{r} \in \mathbb{R}^k$	$O(n_e k + n_r k)$
Unstructured (Bordes et al. 2012)	$\ \mathbf{h} - \mathbf{t}\ _2^2$	$O(n_e k)$
Distant (Bordes et al. 2011)	$\ W_{rh}\mathbf{h} - W_{rt}\mathbf{t}\ _1, W_{rh}, W_{rt} \in \mathbb{R}^{k \times k}$	$O(n_e k + 2n_r k^2)$
Bilinear (Jenatton et al. 2012)	$\mathbf{h}^\top W_r \mathbf{t}, W_r \in \mathbb{R}^{k \times k}$	$O(n_e k + n_r k^2)$
Single Layer	$\mathbf{u}_r^\top f(W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$ $\mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^s, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$	$O(n_e k + n_r (sk + s))$
NTN (Socher et al. 2013)	$\mathbf{u}_r^\top f(\mathbf{h}^\top \mathbf{W}_r \mathbf{t} + W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$ $\mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^s, \mathbf{W}_r \in \mathbb{R}^{k \times k \times s}, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$	$O(n_e k + n_r (sk^2 + 2sk + 2s))$
TransH (	$\ (\mathbf{h} - \mathbf{w}_r^\top \mathbf{h} \mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^\top \mathbf{t} \mathbf{w}_r)\ _2^2$ $\mathbf{w}_r, \mathbf{d}_r \in \mathbb{R}^k$	$O(n_e k + 2n_r k)$

# Applications

- Triples classification
- Link prediction
- Toponym Featurization
- Many more!

Hands-on activities