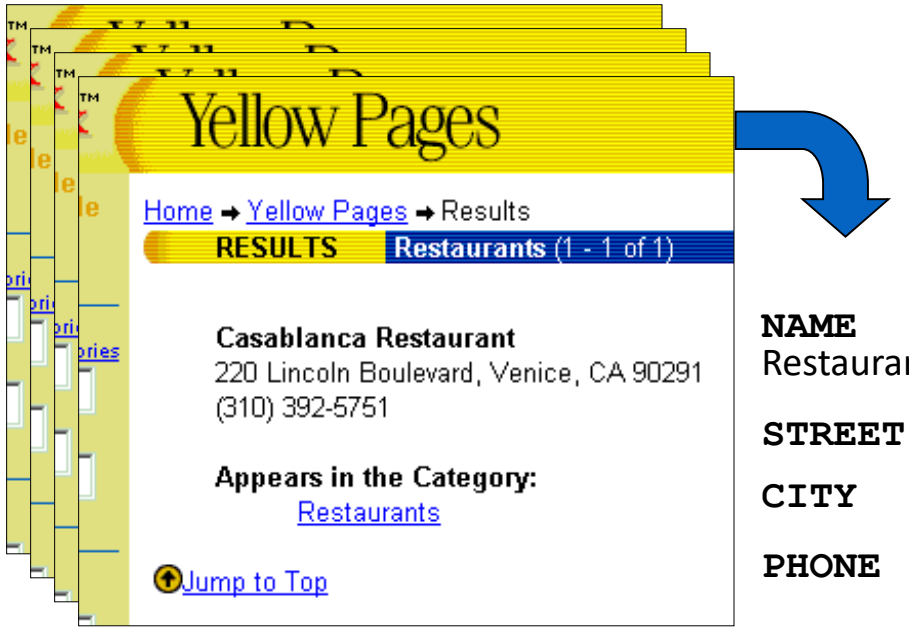
Wrap-up

Part 1

Web IE, Wrappers and Information Integration using Karma

Extracting Data from Semi-structured Sources



NAME Casablanca Restaurant

STREET 220 Lincoln Boulevard

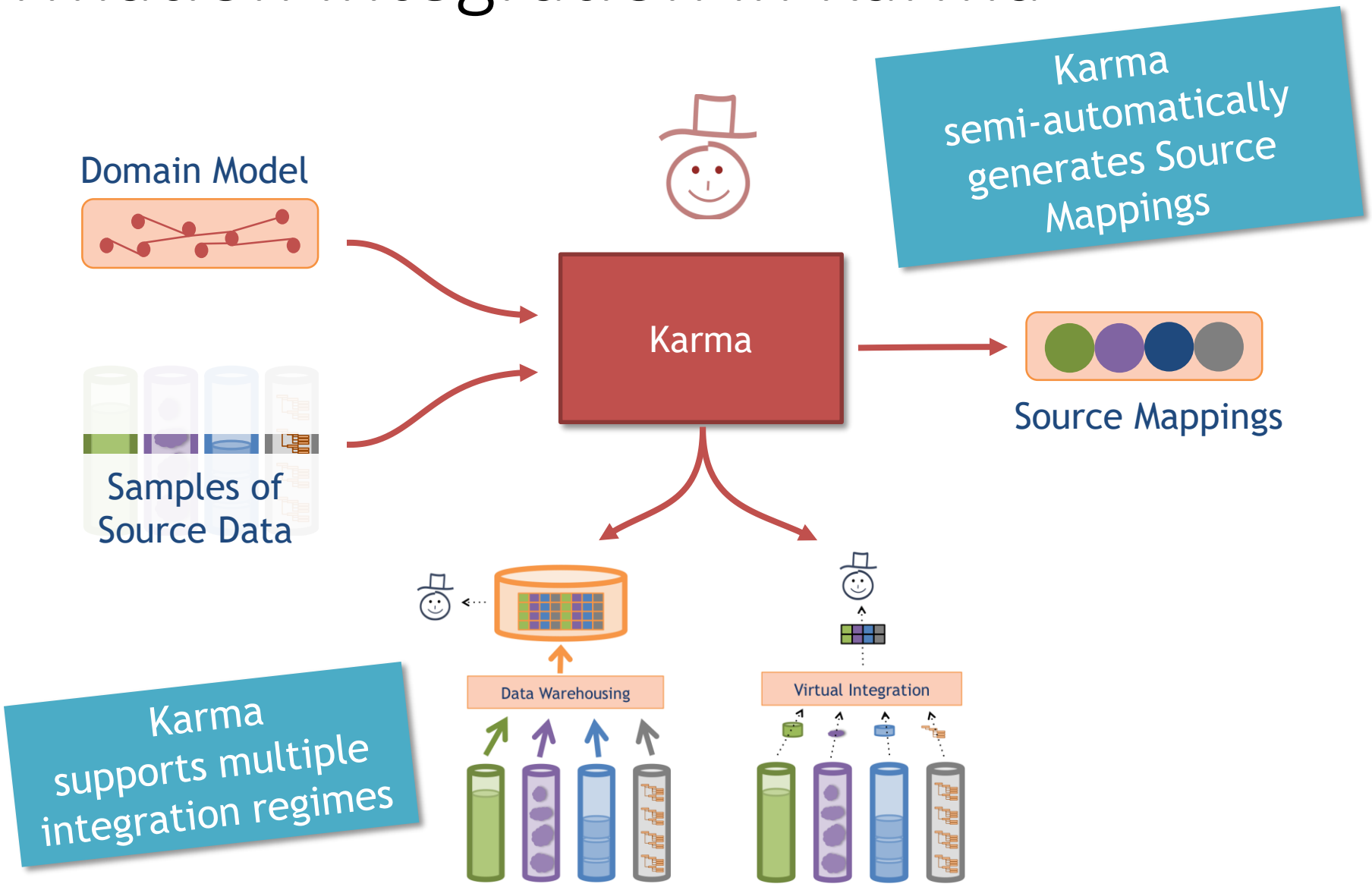
CITY Venice

PHONE (310) 392-5751

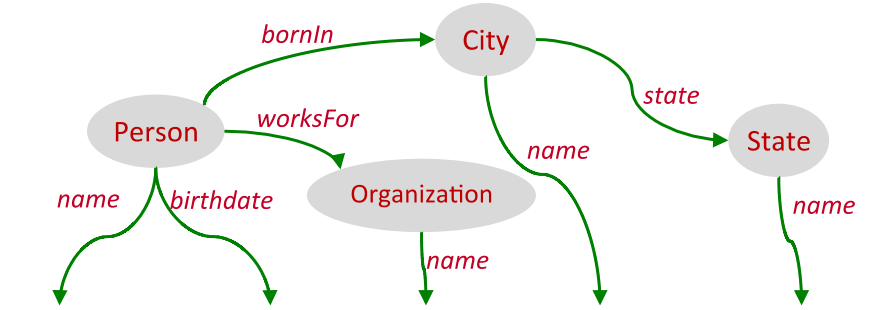
Approaches to Wrapper Construction

- 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



Column 1	Column 2	Column 3	Column 4	Column 5
Bill Gates	Oct 1955	Microsoft	Seattle	WA
Mark Zuckerberg	May 1984	Facebook	White Plains	NY
Larry Page	Mar 1973	Google	East Lansing	MI

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.

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

- Press

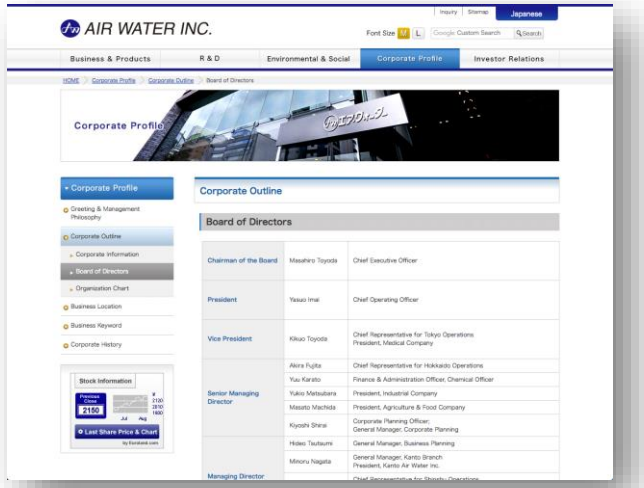
Contact

- General information
- Directions maps

Non-grammatical snippets, rich formatting & links

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

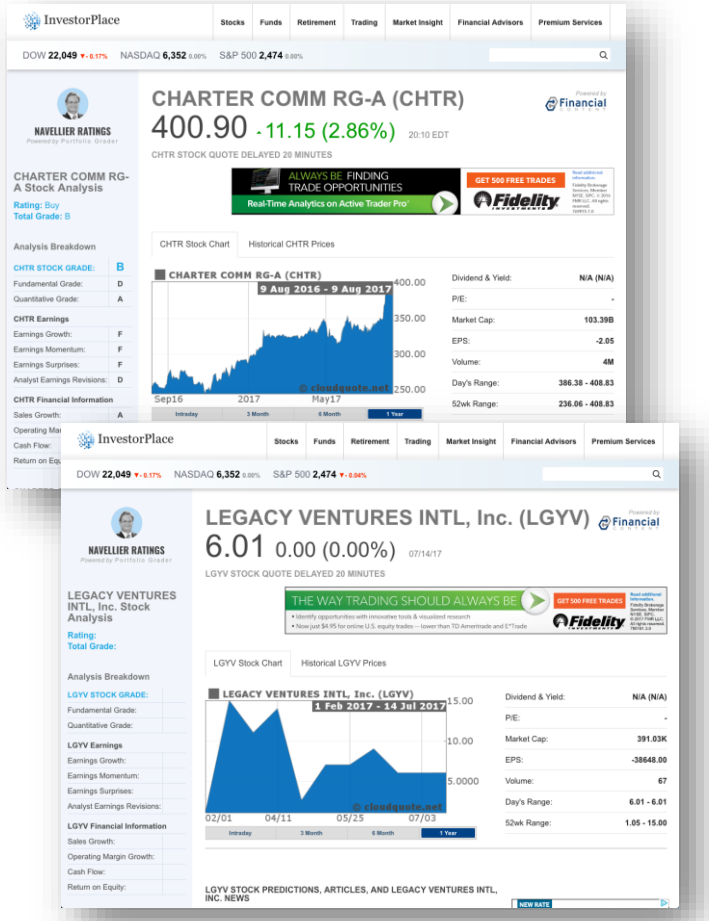
Tables



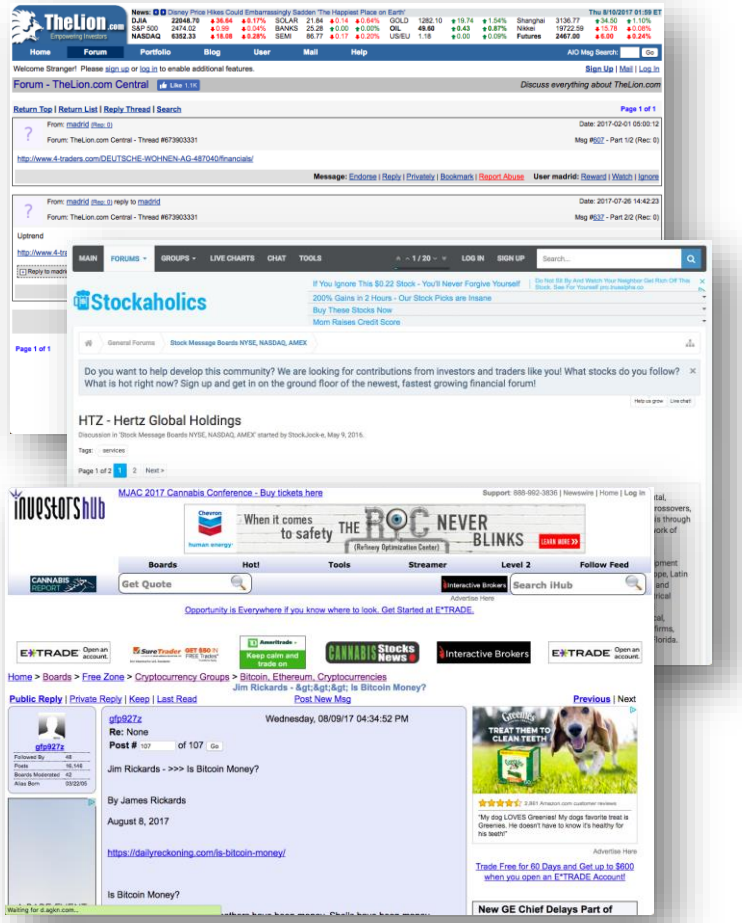
Charts

Scope

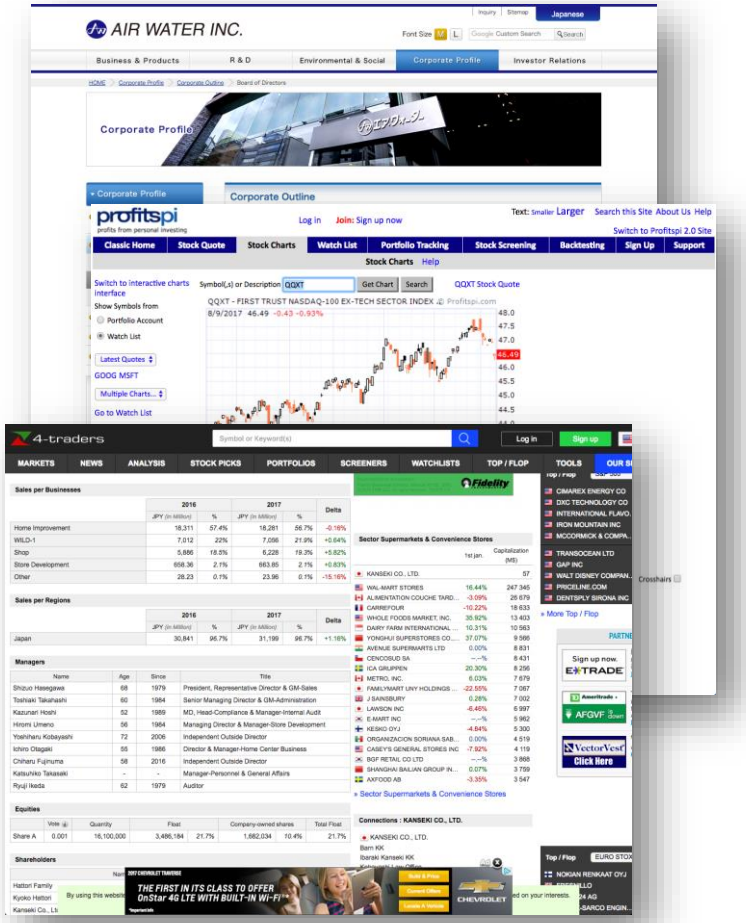
Web site specific



Genre specific (e.g., forums)



Wide, non-specific



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 ❤ 7○7~7two7~7four77 ❤ HH 80 roses ❤ Hour 120 roses ❤ 15 mins 60 roses"

Courtesy of Andrew McCallum

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?

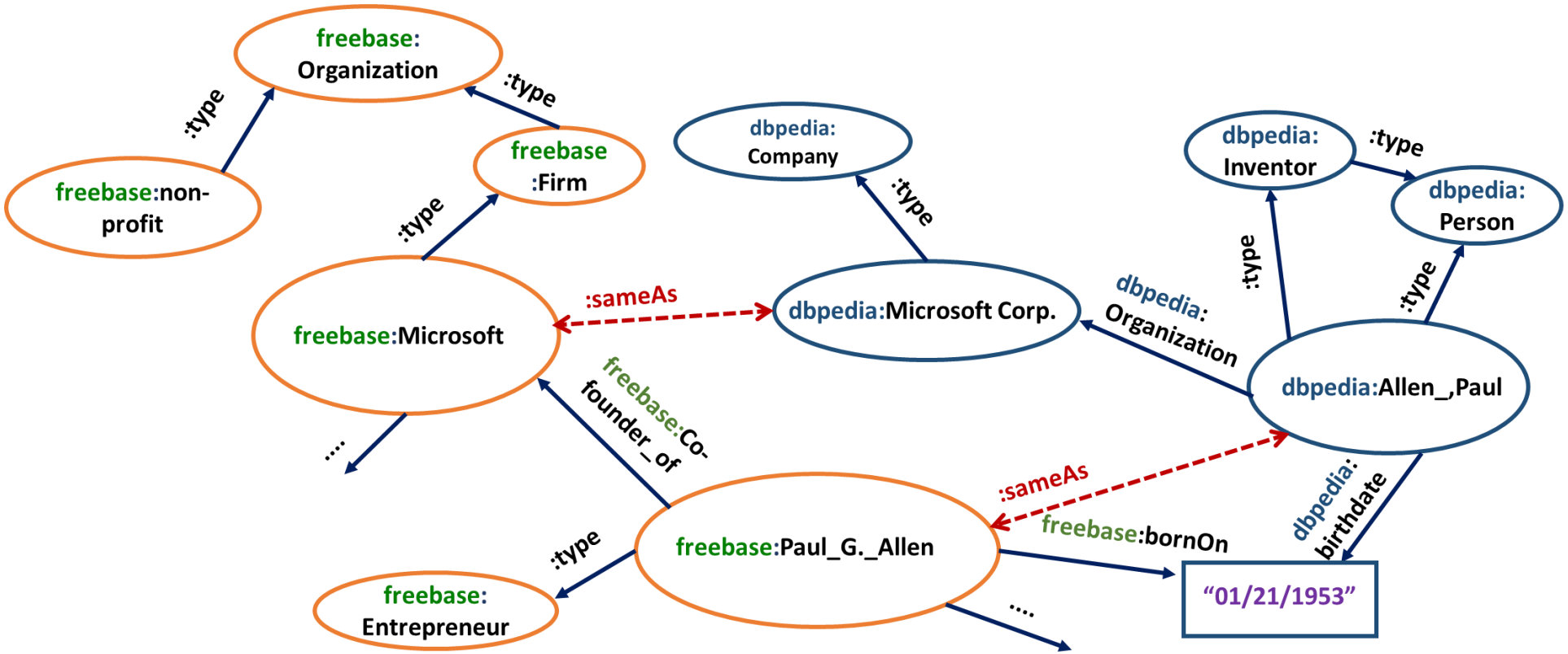
- An ‘intelligent’ way of doing data cleaning
  - Deduplicating entity nodes (entity resolution)
  - Collective reasoning (probabilistic soft logic)
  - Link prediction
  - Dealing with missing values
  - Anything that improves an existing knowledge graph!
- Also known as knowledge base identification

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



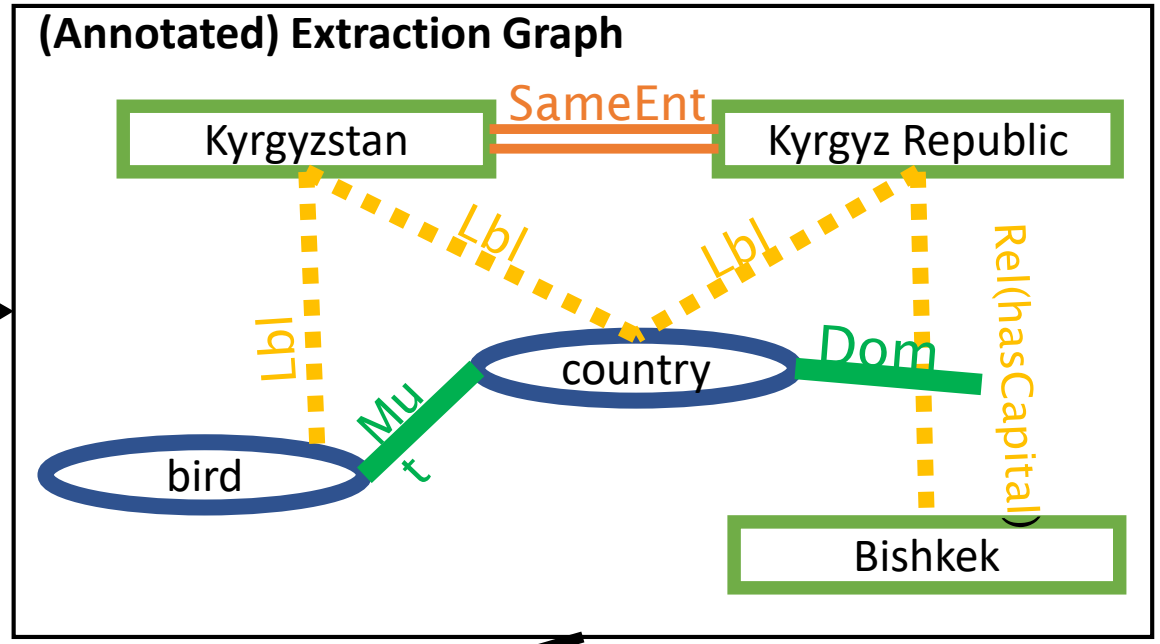
Extraction Graph+Ontology + ER+PSL

Uncertain Extractions:
.5: Lbl(Kyrgyzstan, bird)
.7: Lbl(Kyrgyzstan, country)
.9: Lbl(Kyrgyz Republic, country)

.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

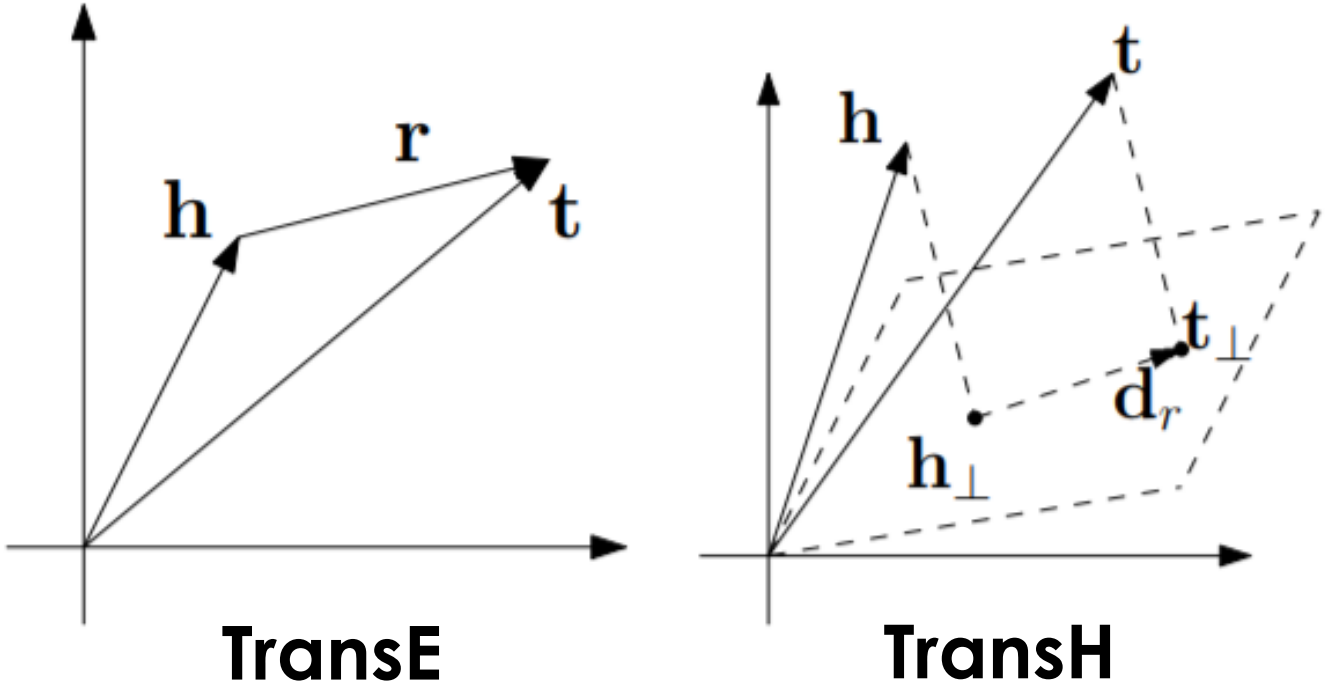
Ontology:
Dom(hasCapital, country)
Mut(country, bird)

Entity Resolution:
SameEnt(Kyrgyz Republic, Kyrgyzstan)



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