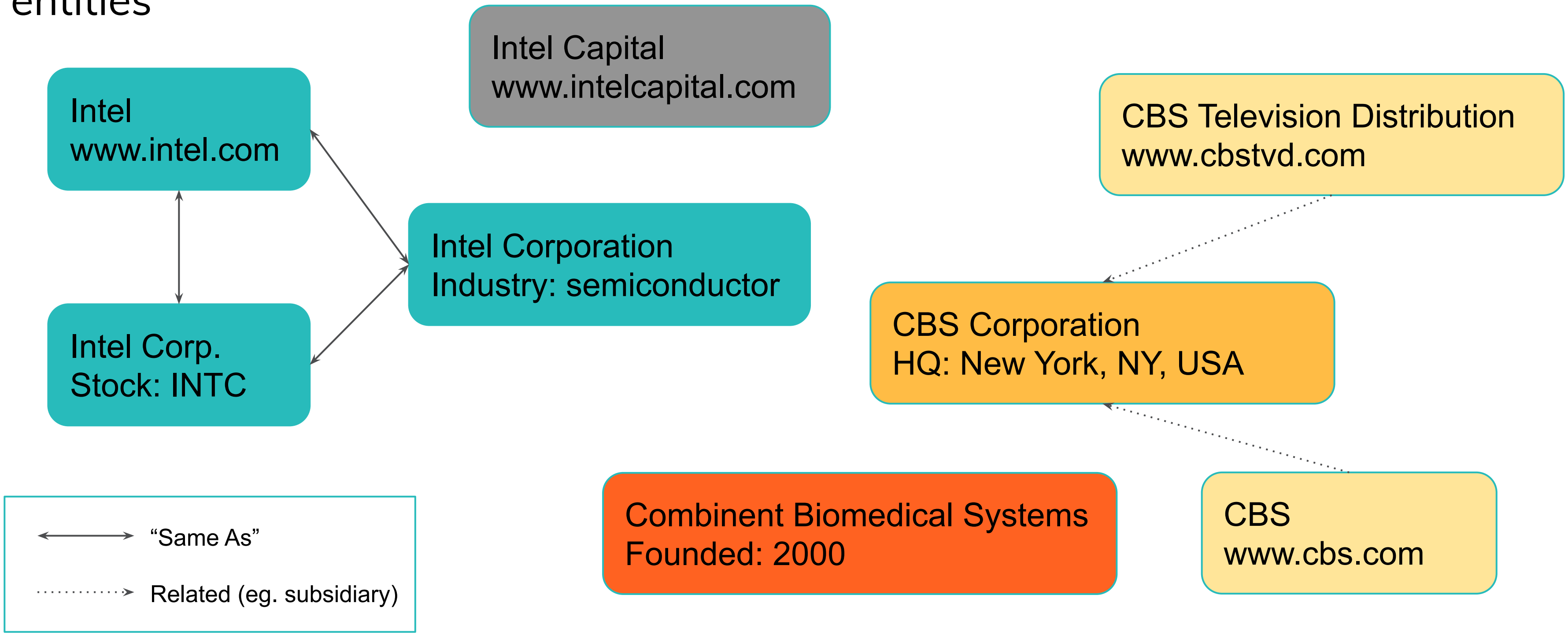
Record Linking for Meltwater's Knowledge Graph

Márton Miháltz
marton.mihaltz@meltwater.com

Meltwater

Introduction

Challenge: link objects from multiple sources that refer to the same real world entities



Overview

- About Meltwater
- MW’s Knowledge Graph
- Record Linking for the KG
- Blocking
- First models for organizations and persons
- Improved Models

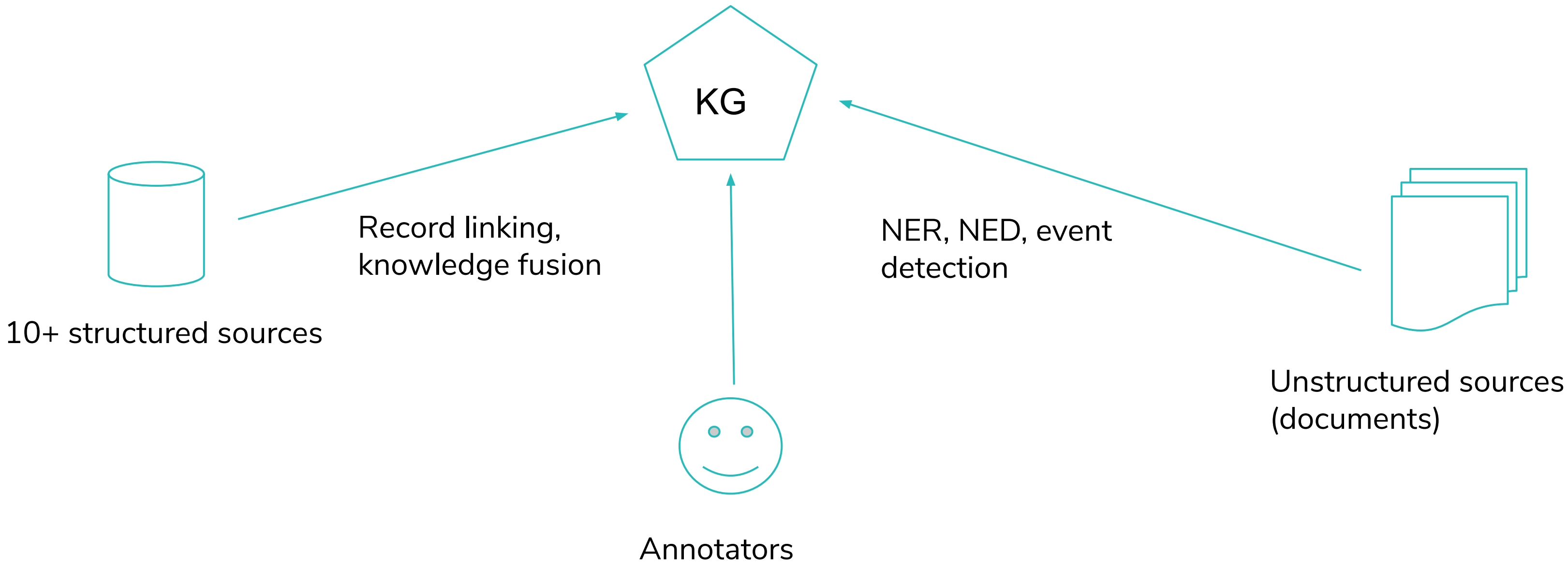
Meltwater's Knowledge Graph

About Meltwater

- **Media Intelligence** solution
  - Media monitoring, social media engagement, competitive intelligence, smart alerts, reports etc.
- 30K clients
- 2K employees in 55 offices, 25 countries
- 10M sources globally: news, social media, print media, broadcasts, podcasts etc.
- 17 NLP languages, 500K docs/s
- 1.4*10^12 documents (2 years rolling)

Meltwater's *Fairhair.ai* Knowledge Graph

- **Nodes**: *organizations, key persons, industries, stock indices, addresses* etc.
- **Edges**: relationships (*affiliations, subsidiaries, industry associations* etc.) or events (*mergers and acquisitions* etc.)



Events Detected via the KG

Executive Departure

Apple

Apple loses an executive

Ars Technica Samuel Axon
Jan 26 • 5:49 PM

Apple's long-time hardware lead steps down to work on mysterious "new project"

It's unusual for the leader of department to shift focus to just one product.

10.8M Reach | Neutral

Launch

Apple

Apple has had a launch related event

MyBroadband Bloomberg
Oct 7 • 5:59 AM

Apples reveals launch date for 5G iPhones

Apple Inc. announced that its biggest product launch event of the year will be held Oct. 13. The Cupertino, California-based technology...

266k Reach | Neutral

Acquisition

Cisco | IMImobile

Cisco & IMImobile are involved in an acquisition event

CISCO

PR Newswire
Feb 19 • 1:30 PM

Cisco Completes Acquisition of IMImobile PLC

SAN JOSE, Calif., and LONDON, Feb. 19, 2021 /PRNewswire/ -- News Summary: Cisco has completed the acquisition of IMImobile PLC. ...

8.49M Reach | Neutral

Recognition

HCL | Cisco

HCL receives an award from Cisco

HCL

India Infoline Ltd
Feb 11 • 1:28 AM

HCL Technologies wins Prestigious Quality Award from Cisco; stock trades higher

HCL is recognized for its Engineering and R&D services provided to Cisco, including its execution, agility and highest quality delivered ...

1.13M Reach | Positive

Meltwater

About Record Linking

What is Record Linking?

- **Cluster** database records / knowledge base entries such that each cluster corresponds to a **single distinct real-world entity** (e.g., a business, a person).

ID	Name	Street Address	City	Phone
r1	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
r2	Starbucks	123 MISSION ST	SAN FRANCISCO	4155431510
r3	Starbucks	123 Mission St	San Francisco	4155431510
r4	Starbucks Coffee	340 MISSION ST	SAN FRANCISCO	4155431510
r5	Starbucks Coffee	333 MARKET ST	SAN FRANCISCO	4155434786
r6	Starbucks	MARKET ST	San Francisco	-

(source)

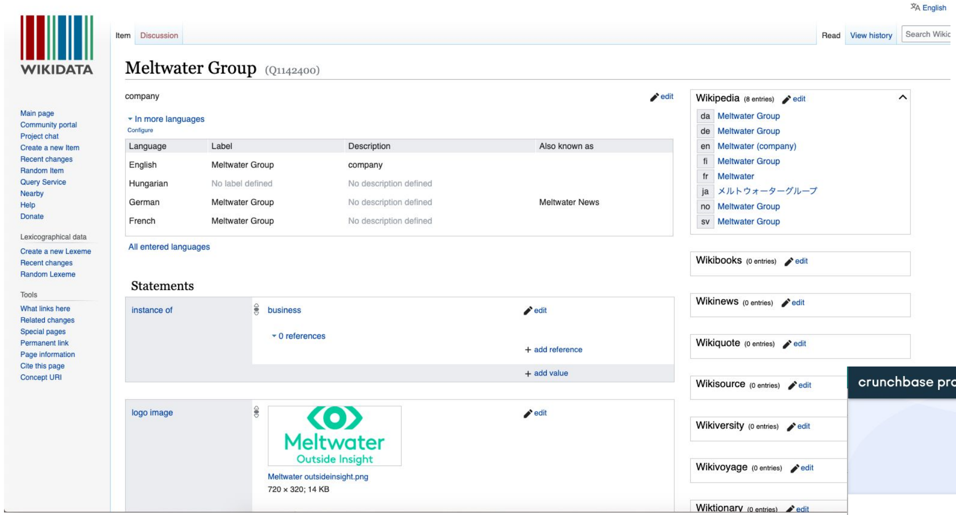
Why is Record Linking Challenging?

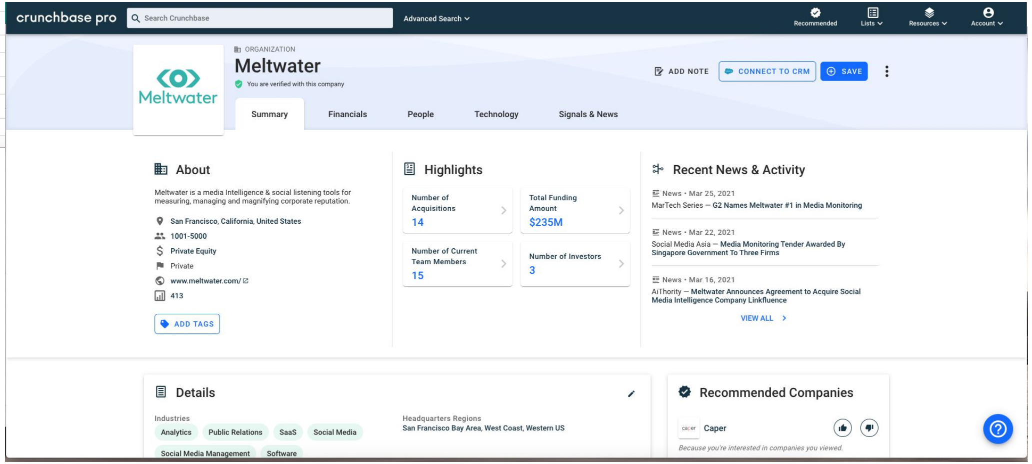
- No literal match (r1, r2)
- Literal match, but not same cluster (r1, r3)
- Very different value, but same cluster (r3, r4)
- Missing attributes (Starbucks / r6)

ID	Person name	Affiliation
r1	Tim Cook	Apple Inc.
r2	Timothy Donald Cook	Apple
r3	Tim Cook	Canadian War Museum
r4	Tim Cook	CWM

ID	Name	Street Address	City	Phone
r1	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
r2	Starbucks	123 MISSION ST	SAN FRANCISCO	4155431510
r3	Starbucks	123 Mission St	San Francisco	4155431510
r4	Starbucks Coffee	340 MISSION ST	SAN FRANCISCO	4155431510
r5	Starbucks Coffee	333 MARKET ST	SAN FRANCISCO	4155434786
r6	Starbucks	MARKET ST	San Francisco	-

Record Linking Example

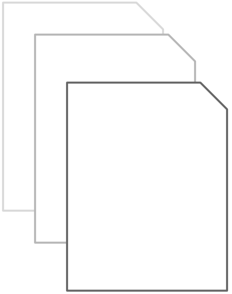




Why is Record Linking Challenging 2: Scalability

- Comparing every record to every other one would be $n * (n-1)/2 = O(n^2)$
- We can do better
- We want support for **parallelization**

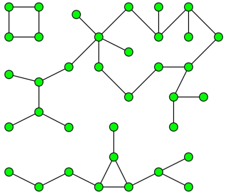
Record Linking Workflow



Source records

Schema mapping, normalization

Blocking



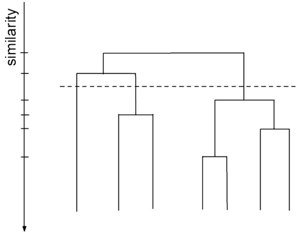
Pairwise similarity score (Machine Learning)

{name1, url1, …}

{name2, url2, …}

$$\begin{bmatrix} .65 \\ .82 \\ \vdots \end{bmatrix} \rightarrow .55$$

Agglomerative clustering



Knowledge Fusion (*TruthFinder*)

Name: ~~name1~~, name2
URL: url1, ~~url2~~

Meltwater

Record Linking Workflow (Details)

1. Mapping to **common schema** (*KG Ontology*)
2. **Blocking**
   - Reduce number of comparisons << O(n*n)
   - Blocking keys: easy to compute &
     minimize P. that objects in same cluster can be in different blocks
3. Pairwise similarity **classifier**
   - Similarity score ([0, 1]), for each record pair in block
   - Features: custom normalization & similarity functions
4. Hierarchical agglomerative **clustering**
   - Via pairwise similarity scores
   - Cut-off threshold
5. Knowledge **fusion** (*TruthFinder*)

Running on Apache **Spark** on AWS **EMR clusters**

Blocking

Single-Attribute Blocking

ID	Expected BlockID	Name	HomepageURL	Blocking key (=domain of URL)
c1	b1	Exxon Mobil Corporation	http://exxonmobil.com	exxonmobil
c2	b1	Exxon	http://exxonmobil.com	exxonmobil
c3	b1	Exxon	http://exxon.com	exxon
c4	b2	Lincoln National Corporation	http://www.lfg.com	lfg
c5	b2	Lincoln Financial Group	http://www.lincolnfinancial.com	lincolnfinancial
c6	b3	John Deere	http://www.deere.com	deere
c7	b3	Deere & Company	http://www.johndeere.com	johndeere

Single-Attribute Blocking

ID	Expected BlockID	Name	HomepageURL	Blocking key (=domain of URL)
c1	b1	Exxon Mobil Corporation	http://exxonmobil.com	exxonmobil
c2	b1	Exxon	http://exxonmobil.com	exxonmobil
c3	b1	Exxon	http://exxon.com	Exxon
c4	b2	Lincoln National Corporation	http://www.lfg.com	lfg
c5	b2	Lincoln Financial Group	http://www.lincolnfinancial.com	lincolnfinancial
c6	b3	John Deere	http://www.deere.com	deere
c7	b3	Deere & Company	http://www.johndeere.com	johndeere

Multi-Attribute, Multi-Value Blocking

ID	Expected BlockID	Name	HomepageURL	Blocking key1 (=domain of URL)	Blocking key2 (=tokens of Name)
c1	b1	Exxon Mobil Corporation	http://exxonmobil.com	exxonmobil	exxon, mobil
c2	b1	Exxon	http://exxonmobil.com	exxonmobil	exxon
c3	b1	Exxon	http://exxon.com	exxon	exxon
c4	b2	Lincoln National Corporation	http://www.lfg.com	lfg	lincoln, national
c5	b2	Lincoln Financial Group	http://www. lincolnfinancial.com	lincolnfinancial	lincoln, financial
c6	b3	John Deere	http://www.deere.com	deere	john, deere
c7	b3	Deere & Company	http://www.johndeere .com	johndeere	deere

Multi-Attribute, Multi-Value Blocking

ID	Expected BlockID	Name	HomepageURL	Blocking key1 (=domain of URL)	Blocking key2 (=tokens of Name)
c1	b1	Exxon Mobil Corporation	http://exxonmobil.com	exxonmobil	exxon, mobil
c2	b1	Exxon	http://exxonmobil.com	exxonmobil	exxon
c3	b1	Exxon	http://exxon.com	exxon	exxon
c4	b2	Lincoln National Corporation	http://www.lfg.com	lfg	lincoln, national
c5	b2	Lincoln Financial Group	http://www.lincolnfinancial.com	lincolnfinancial	lincoln, financial
c6	b3	John Deere	http://www.deere.com	deere	john, deere
c7	b3	Deere & Company	http://www.johndeere.com	johndeere	deere

Meltwater

Multi-Attribute Blocking With Connected Components Analysis

1. **Build graph**
   - Vertices: *records* (id, blocking key-value pairs, fields)
   - Edges: *connect any 2 vertices if they share at least 1 blocking key-value pair*

2. **Find connected components**
(connected component: *subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the supergraph*)

3. For **each component: do clustering** inside

Blocking Key Graph 1: Vertices

```
c1
{k1=exxonmobil,
 k2=exxon,
 k2=mobil}
```

```
c4
{k1=lfg,
 k2=lincoln,
 k2=national}
```

```
c3
{k1=exxon,
 k2=exxon}
```

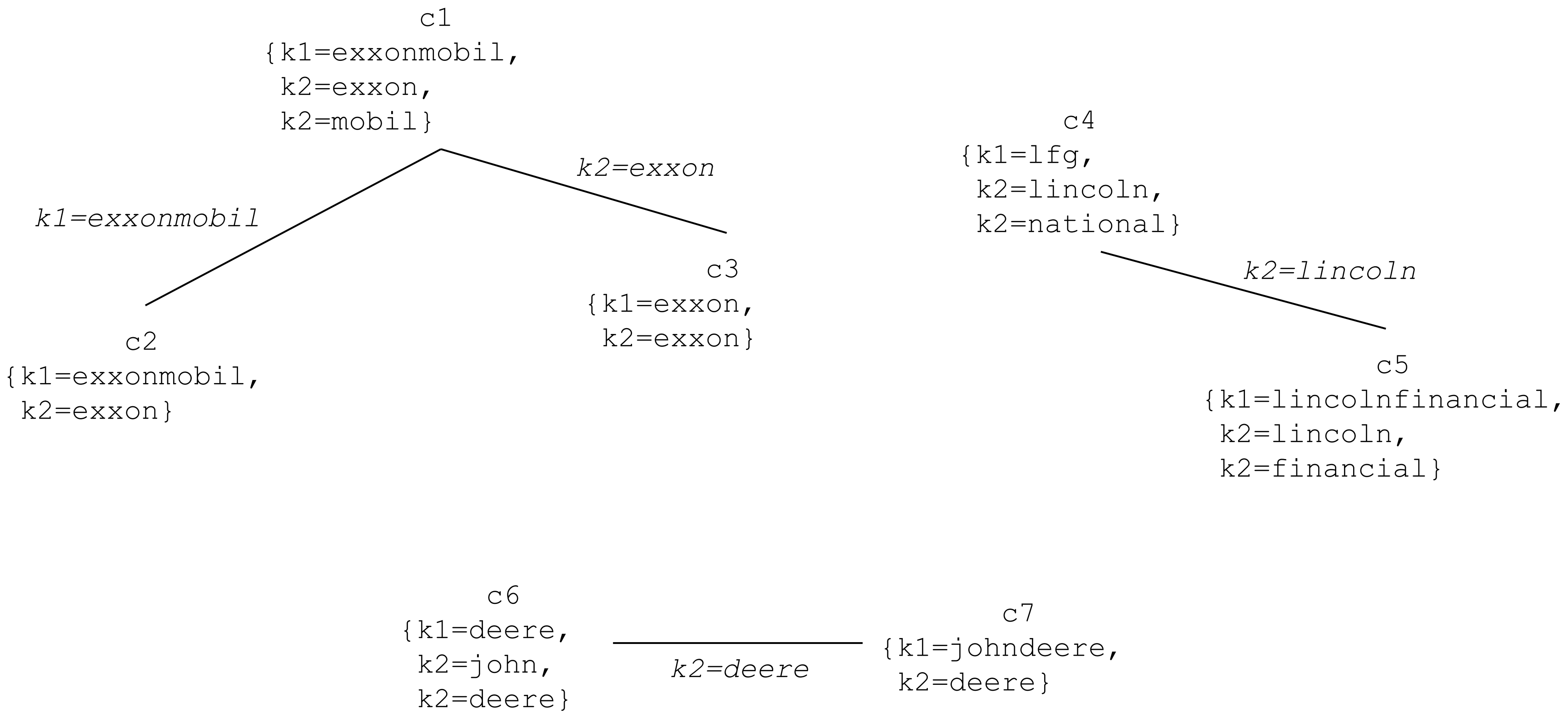
```
c2
{k1=exxonmobil,
 k2=exxon}
```

```
c5
{k1=lincolnfinancial,
 k2=lincoln,
 k2=financial}
```

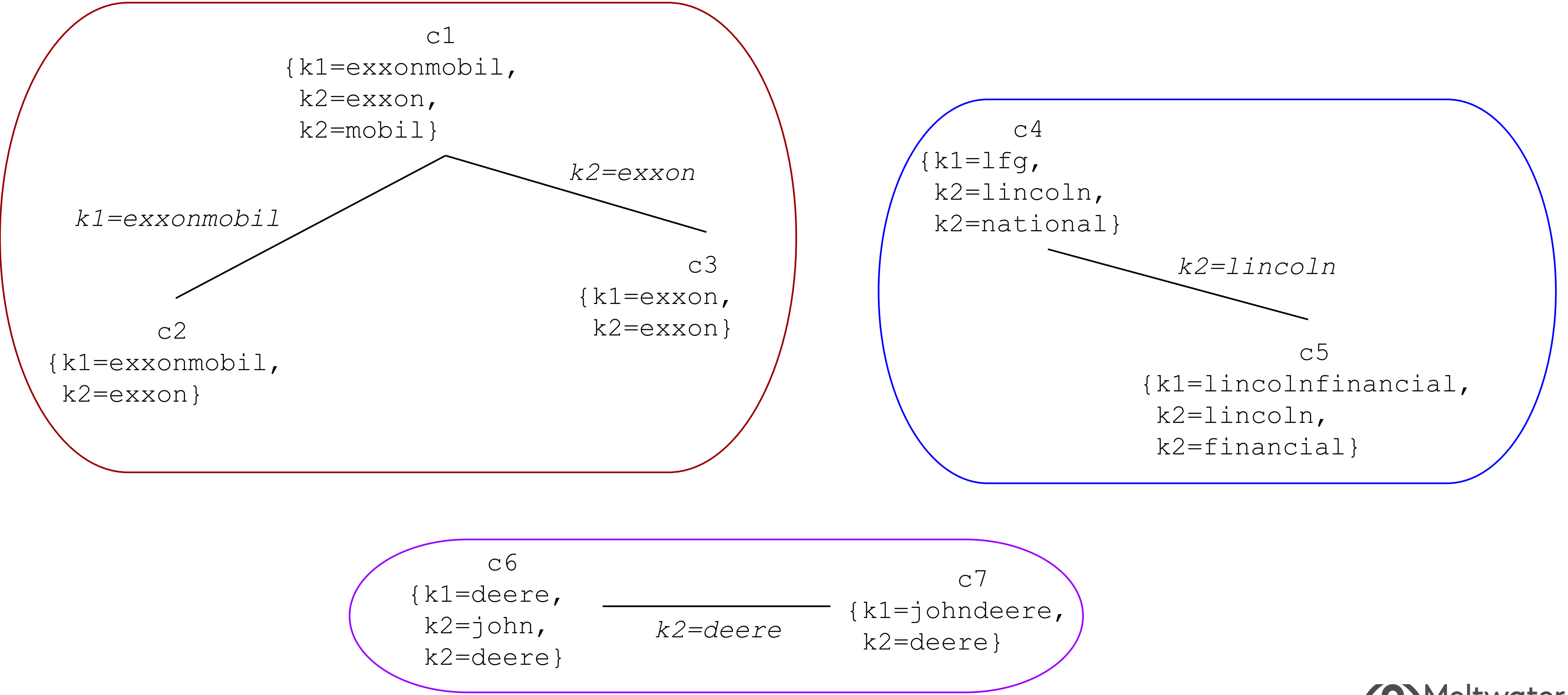
```
c6
{k1=deere,
 k2=john,
 k2=deere}
```

```
c7
{k1=johndeere,
 k2=deere}
```

Blocking Key Graph 2: Edges



Blocking Key Graph 3: Connected Components



RL for Organizations

RL for Organizations: Blocking Keys

1. Domain part of `HomepageURLField`
   Eg. `http://www.intel.com/welcome -> intel`
   - Whitelists and heuristics
     - Eg. *sites.google.com/site/lirepublicairporths* -> *google/lirepublicairporths*
     - Using blog site's subdomain eg: <*site*>*.wordpress*
     - Using path for social media profiles eg: *twitter/user*
2. Tokens of normalized `OrganizationNameField`
   `Tapestry, Inc. -> tapestry`
   `Exxon Mobil Corporation -> exxon, mobil`
   - Blacklisted tokens: general words in names (`Technology, Energy, Data,` …)

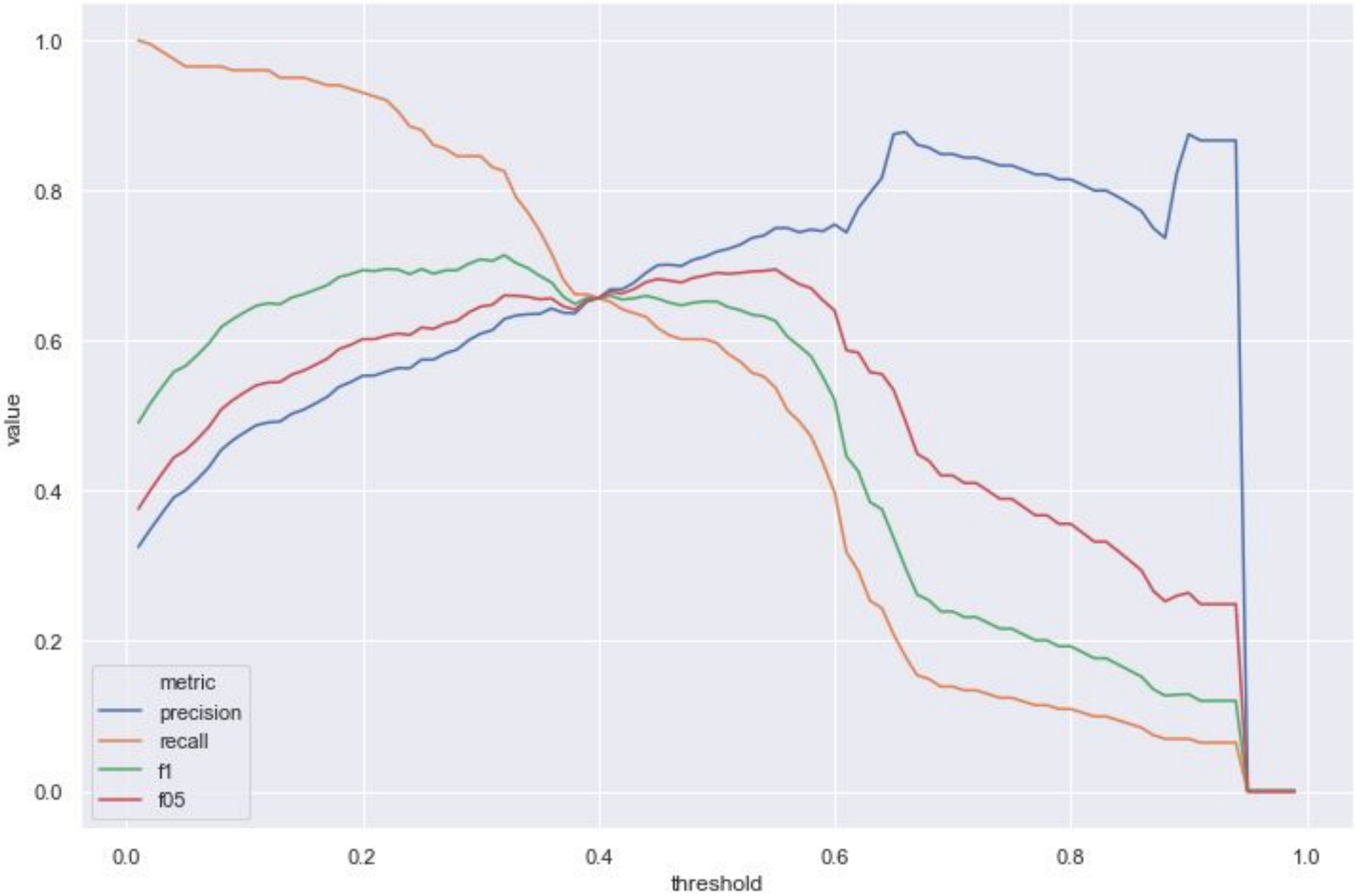
RL for Organizations: Similarity Classifier Features

Extracted from KG Field	Feature Name	Normalization	Semantics
`HomepageURLField`	`Homepage_dissimilarity`	Remove protocol, remove common paths eg. `/index.htm, /en-us`	Normalized Levenshtein distance
	`Homepage_exact_match`		1 if match, 0 if no match, .5 if either missing
	`Homepage_suffix_no_match`	Extract url suffix(es), eg. `.com, .co.uk`	1 if no match, 0 if match, .5 if either missing
`OrganizationNameField`	`Name_dissimilarity`	Remove prefixes/suffixes & slugify, eg. `The Coca-Cola Company -> coca-cola`	Jaro-Winkler distance
	`Name_suffix`		1 if either name is real suffix of the other, 0 otherwise
`FacebookURLField`	`facebook_handle_match`	Extract handle from URL	1 if match, 0 if no match, .5 if either missing
`TwitterURLField`	`twitter_handle_match`	Extract handle from URL	1 if match, 0 if no match, .5 if either missing
`LinkedInURLField`	`linkedin_handle_match`	Extract handle from URL	1 if match, 0 if no match, .5 if either missing

Training Data

- Initial model
  - 18 company records from 4 sources
  - 7 clusters
  - 36 positive pairs (same cluster)
  - 135 negative pairs (different clusters)
- Improved model
  - 6K Manually identified clusters
  - 13K positive pairs (same cluster)
  - 13K negative pairs (same block, different cluster)

Org. Similarity Classifier Evaluation



Meltwater

Org. Evaluation 1.

- Similarity classifier on test set: 640 public company pairs manually annotated (same/not same)

	Initial	Improved			
Threshold	.3	.66 (max. prec.)	.01 (max. rec.)	.32 (max. F1)	.47 (max. F0.5)
Precision	83.3%	**87.8%**	32.5%	62.9%	75.0%
Recall	2.5%	17.9%	**100.0%**	82.6%	53.7%
F1	4.8%	29.8%	49.1%	**71.4%**	62.6%
F0.5	11.2%	49.3%	37.6%	66.0%	**69.5%**

Org. Clustering Evaluation

Evaluation of Clustering against Gold Standard

	Initial	Improved			
Threshold	.3	.66	.01	.32	.47
Precision	93%	95.4%	99.3%	98.7%	98.4%
Recall (with missing*)	77% (47%)	81% (49.8%)	96.5% (59.4%)	95.5% (58.7%)	93.3% (57.4%)
F1-score (with missing*)	84% (63%)	87.6% (65.4%)	97.9% (74.3%)	97% (73.6%)	95.8% (72.5%)
F0.5-score (with missing*)	89% (77%)	92.12% (80.6%)	98.7% (87.5%)	98.0% (86.8%)	97.3% (86,1%)

Gold Standard:

- 50 from Fortune-1k
- 51 from Wikidata
- 50 from DBpedia
- 51 from Crunchbase
- 50 from Nasdaq
- 50 from Linkedin

Actual clustering input in Proto-Graph:

- 50 from Fortune-1k
- 51 from Wikidata
- 48 from DBpedia
- 51 from Crunchbase
- 35 from Nasdaq
- 7 from Linkedin

* accounting for objects present in the gold standard but lost during ETL before record linking (invalid/missing data etc.)

Org. RL Qualitative Evaluation

- Clustering of 11 companies from 7 sources similar to “Apple” (KG-1002)

Org. RL Qual. Eval.: Initial Model

Source	Name	Homepage	ClusterId
golden_set	Apple	apple.com	1 ✅
nasdaq	Apple Inc.	http://www.apple.com	1 ✅
dbpedia	Apple Store (online)	http://www.apple.com/	1 ❌
dbpedia	Apple Inc.	http://www.apple.com	1 ✅
fortune1k	Apple, Inc.	http://www.apple.com	1 ✅
wikidata	Apple (Germany)	http://www.apple.com/de/	1 ❌
wikidata	Apple (United Kingdom)	https://www.apple.com/uk/	1 ❌
wikidata	Apple Store Online	http://www.apple.com/	1 ❌
crunchbase	Apple	http://www.apple.com	1 ✅
linkedin	Apple Sign	http://www.apple.com/	1 ❌
barchart	Apple Inc	http://www.apple.com	1 ✅

Org. RL Qual. Eval.: Improved Model

Source	Name	Homepage	ClusterId
golden_set	Apple	apple.com	1 ✅
nasdaq	Apple Inc.	http://www.apple.com	1 ✅
dbpedia	Apple Store (online)	http://www.apple.com/	2 ✅
dbpedia	Apple Inc.	http://www.apple.com	1 ✅
fortune1k	Apple, Inc.	http://www.apple.com	1 ✅
wikidata	Apple (Germany)	http://www.apple.com/de/	3 ✅
wikidata	Apple (United Kingdom)	https://www.apple.com/uk/	4 ✅
wikidata	Apple Store Online	http://www.apple.com/	2 ✅
crunchbase	Apple	http://www.apple.com	1 ✅
linkedin	Apple Sign	http://www.apple.com/	5 ✅
barchart	Apple Inc	http://www.apple.com	1 ✅

Org. RL Qualitative Evaluation

- Clustering of 11 companies from 7 sources similar to “Apple”

	Before	After
Precision	29.09%	100.00%
Recall	84.21%	100.00%
F-measure	43.24%	100.00%

RL for Persons

RL for Persons: Similarity Classifier Features

Extracted from KG Field	Feature Name	Normalization	Semantics
PersonNameField, PersonNameAliasField	person_name_normalized_exact_match	slugification	1 if overlap in 2 sets, 0 otherwise
	max_person_name_similarity		Normalized Damerau-Levenshtein similarity (.5 if either missing)
PersonOrganization RelationField. OrganizationNameField	affiliated_organization_name_dissimilarity	Remove prefixes/suffixes & slugify, eg. `The Coca-Cola Company -> coca-cola`	Jaro-Winkler distance (.5 if either missing)
	affiliated_organization_name_suffix		1 if either name is real suffix of the other, 0 otherwise
PersonOrganization RelationField. HomepageURLField	affiliated_organization_homepage_dissimilarity	Remove protocol, remove common paths eg. `/index.htm, /en-us`	Normalized Levenshtein distance
	affiliated_organization_homepage_suffix_no_match	Extract url suffix(es), eg. `.com, .co.uk`	1 if no match, 0 if match, .5 if either missing
PersonOrganization RelationField. TwitterURLField	twitter_handle_match	Extract handle from URL	1 if match, 0 if no match, .5 if either missing
PersonOrganization RelationField. JobTitleField	average_job_titles_jaccard_for_matching_orgs	`(norm. org name, norm. job title)` pairs	Jaccard similarity bw. job title sets for matching org.

Meltwater

RL for Persons: Blocking Keys

- From all `PersonNameFields` and `PersonNameAliasFields`
- Identify (using `probablepeople`):
    - **given name** and **surname** parts (default: last token for family name)
- Convert given name to formal version (if available)
- Blocking key = slugify(formalized given name + “ “ + family name)
- Examples:
- `Jim Hackett, James Hackett            -> james-hackett`
- `Robert Iger, BOB IGER                 -> robert-iger`
- `Donald M. Casey Jr., Donald Casey     -> donald-casey`

Person RL: Job Title Normalization

- Keep only first 100 chars
- Use only first 10 tokens
- Segment (commas, ”&”)
- Slugify
- Remove prefixes (`co-, interim-, ...`)
- Resolve abbreviations (`ceo, cfo, cio, cto, coo, cmo, ...`)
- Replace suffixes (`-man|-woman -> -person`)

RL for Persons: Classifier Training & Test Sets

- Fortune-1000 company executives (2019)
- Manually annotated matching ids from
  - Crunchbase
  - Wikidata
  - EON people
- **1416 Positive** pairs from annotation
- **1869 Negative** pairs auto generated
  - same blocking key, but id != positive, from CB, WD & EON
  - Manually verified suspicious pairs (affiliation company ==)
- Train–test split: 75-25 %

RL for Persons: Classifier Evaluation

	Initial		Improved	
	Max. F.5	Max. F1	Max. F.5	Max. F1
Precision	88.6%	54.1%	98.6%	98.3%
Recall	43.5%	97.7%	89.1%	89.6%
F1-score		69.6%		93.7%
F.5-score	73.4%		96.5%	

RL for Persons: Clustering Evaluation

Evaluation of Clustering against Test split of Gold Standard

	Initial	Improved
Threshold	.55 (max. F1)	.74 (max. F.5)
Precision	78.1%	87.6%
Recall	74.5%	87.5%
F1-measure	76.3%	87.5%

Summary

- Meltwater's Knowledge Graph
  - Fused from structural sources via Record Linking
  - Improved by NER, NED and event detection from unstructured sources
  - Serving Signals for clients
- Record Linking
  - blocking, similarity classifier, hierarchical clustering
  - Models in production for Organizations, Persons
- More information