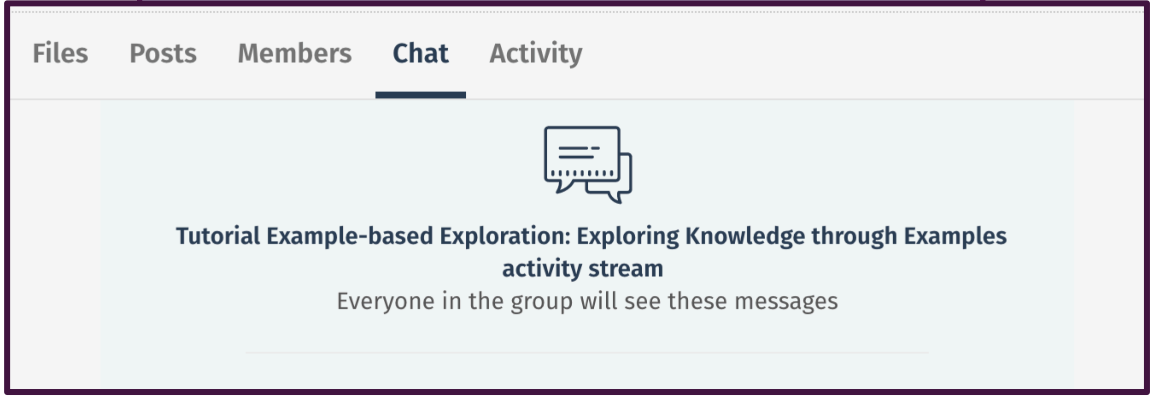
ESWC 2020 tutorial

Example-based exploration:
exploring knowledge through examples

Matteo Lissandrini (Aalborg University),
Davide Mottin (Aarhus University),
Yannis Velegrakis (Utrecht University)
Themis Palpanas (University of Paris)

AARHUS UNIVERSITY
AALBORG UNIVERSITET
Utrecht University
UNIVERSITÉ PARIS DESCARTES

You can use the MiTeam chat for questions

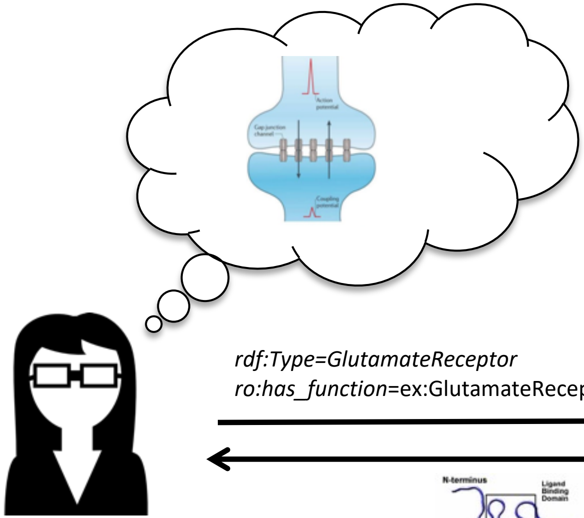


Tutorial Slides and Other Material

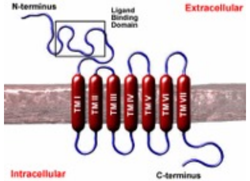
https://data-exploration.ml/eswc2020.html

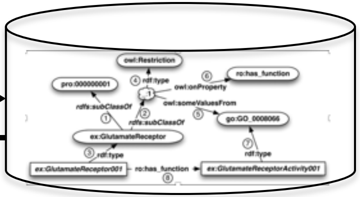

LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE
© marketoonist.com

Traditional Data Management Systems



rdf:Type=GlutamateReceptor
ro:has_function=ex:GlutamateReceptorActivity001

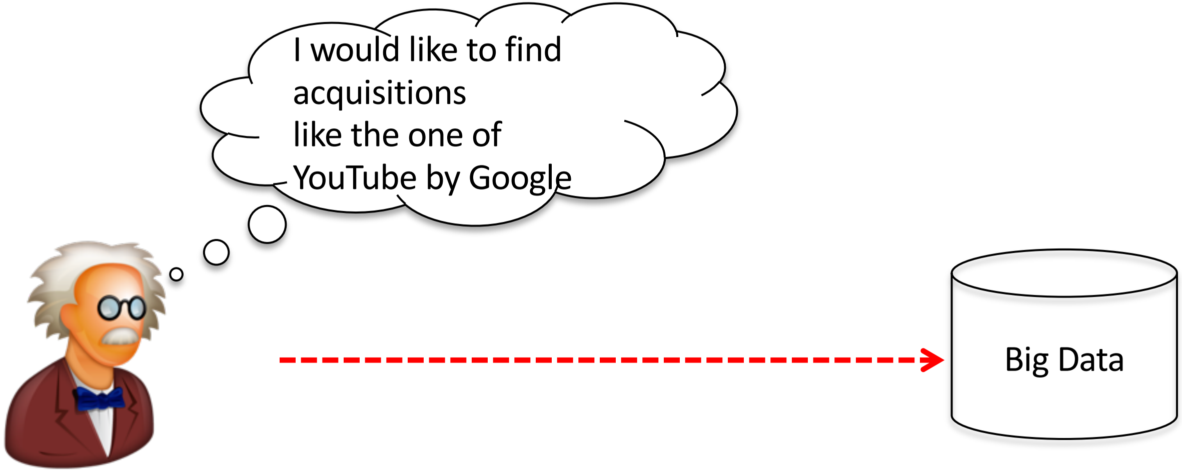




Data

Modern Data Management Systems

Not clear what we are looking for



Exploration

We know where we start
we don't know what we'll find

Exploration

Traditional

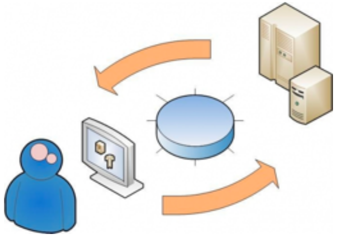
On data

Data exploration

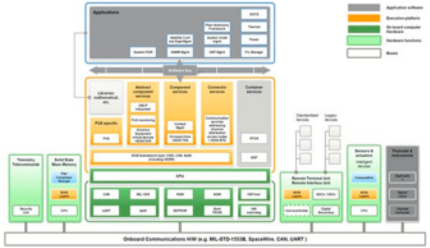
Cleaning and profiling

Visualization

Analysis

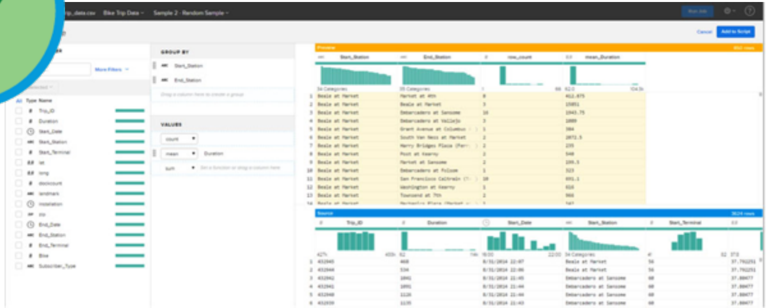


Interactions

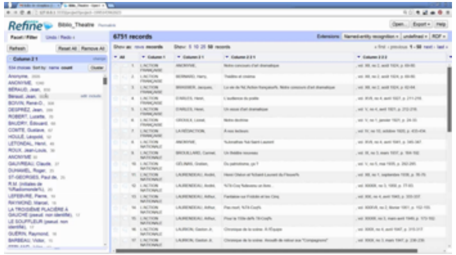


Architectures

Data exploration software



Trifacta: data preparation



OpenRefine: data preparation and cleanup

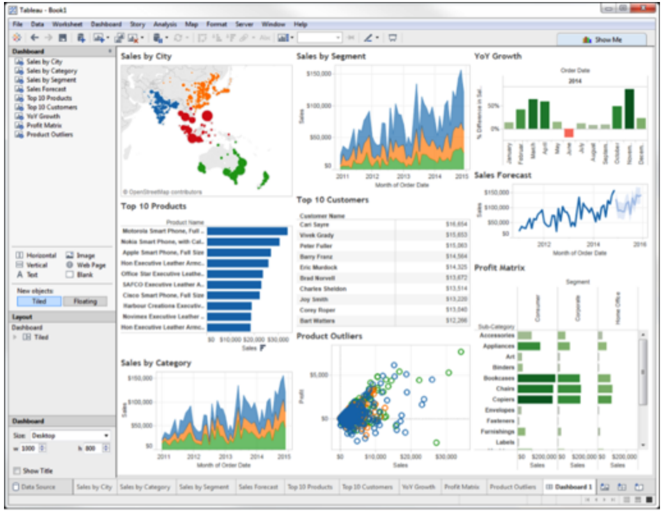


Tableau: analysis and statistics

Traditional data exploration methods

[Idreos et al., 2015]

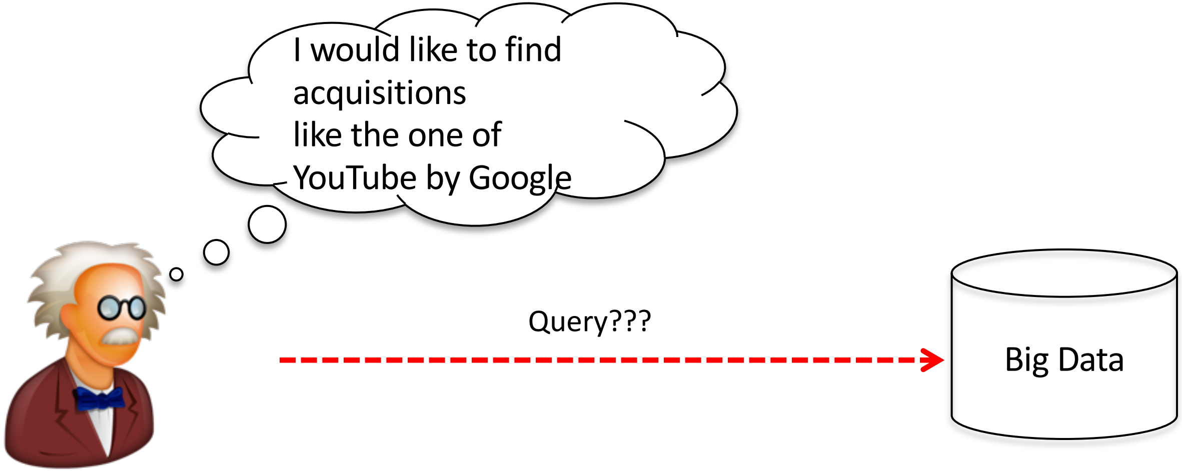
Efficiently extracting knowledge from data
even if we do not know exactly what we are looking for

SELECT avg(system-stars)
FROM Universe
WHERE system-stars > 10
GROUP BY galaxy

Not easy for novices

Modern Data Management Systems

How do we describe what we are looking for?



Declarative Exploratory methods

SELECT galaxy_name
FROM Universe.Galaxy

SELECT g.galaxy_name, SUM(s.stars) as st_s
FROM Universe.Galaxy **AS** g
JOIN Universe.Systems **AS** s
ON g.galaxy_name = s.galaxy_name
WHERE
g.st_s > 100B
AND diameter > 100k AND diameter > 180k
AND has_black_hole = TRUE
GROUP BY g.galaxy_name

Simple query (exploratory)

Over generic
100 billions results

Complex query
(for data experts)

Specific
Few results

Examples as Exploratory Methods

Example is always more efficacious than precept
Samuel Johnson, Rasselas (1759), Chapter 29.

Answers




Tutorial's goals

Techniques, Algorithms, Applications for using Examples to support Exploratory Search

- Exploratory methods using examples
- Algorithms for retrieving data without using query languages
- Interactive methods and user-in-the-loop feedback
- Machine learning for adaptive, online methods

But NOT

- Declarative query methods
- User interfaces and visualization
- Optimizations for fast data access
- Dynamic data

Our book on Example-based methods

Matteo Lissandrini
Aalborg University

Knowledge Graphs , Novel Query Paradigms, Graph Mining
http://people.cs.aau.dk/~matteo

Yannis Velegrakis
Utrecht University

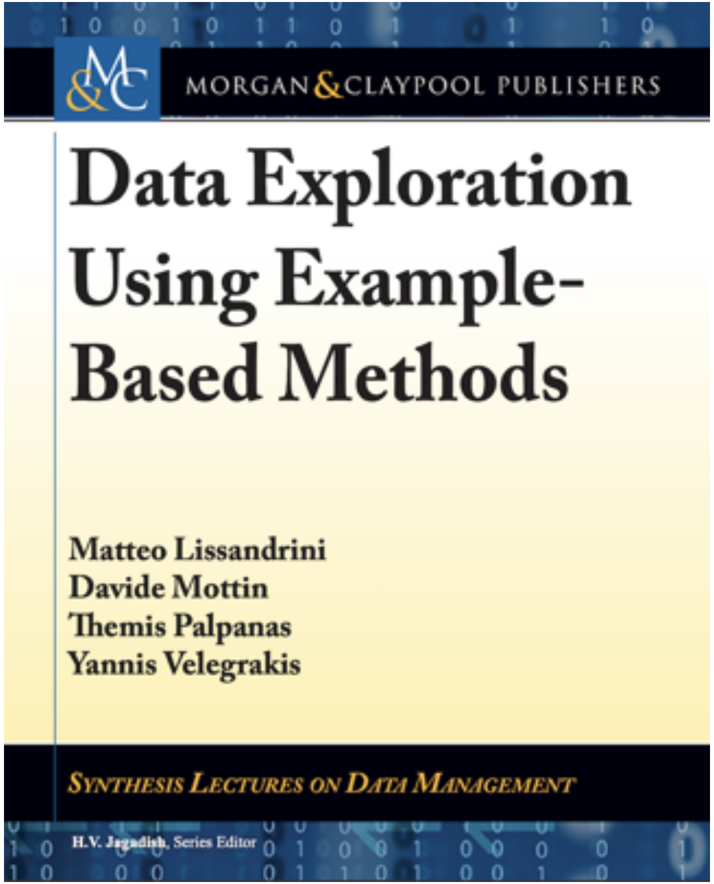
Big Data Management & Analytics, Information Integration, Data Curation
https://velgias.github.io

Davide Mottin
Aarhus University

Graph Mining, Novel Query Paradigms, Interactive Methods
https://mott.in

Themis Palpanas
Paris Descartes University

Data Series Indexing & Mining, Data Analytics & Management
http://www.mi.parisdescartes.fr/~themisp



Historical perspective: Query-by-example

[Zloof et al. 1975]

Specify a query by example tables, or skeletons.

Name	Stars	Diameter	Black_hole	Color	Life
P._	> 10B	>100k	TRUE	*	*
*	*	<180k	*	*	*

Incomplete values

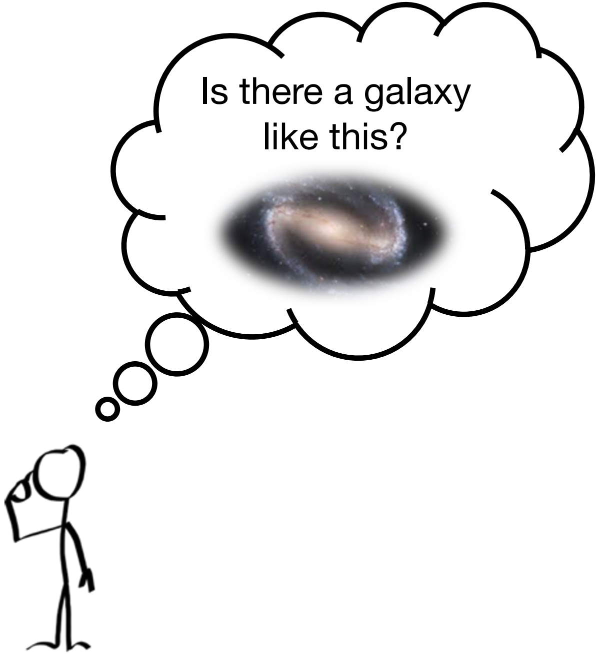
Unspecified values

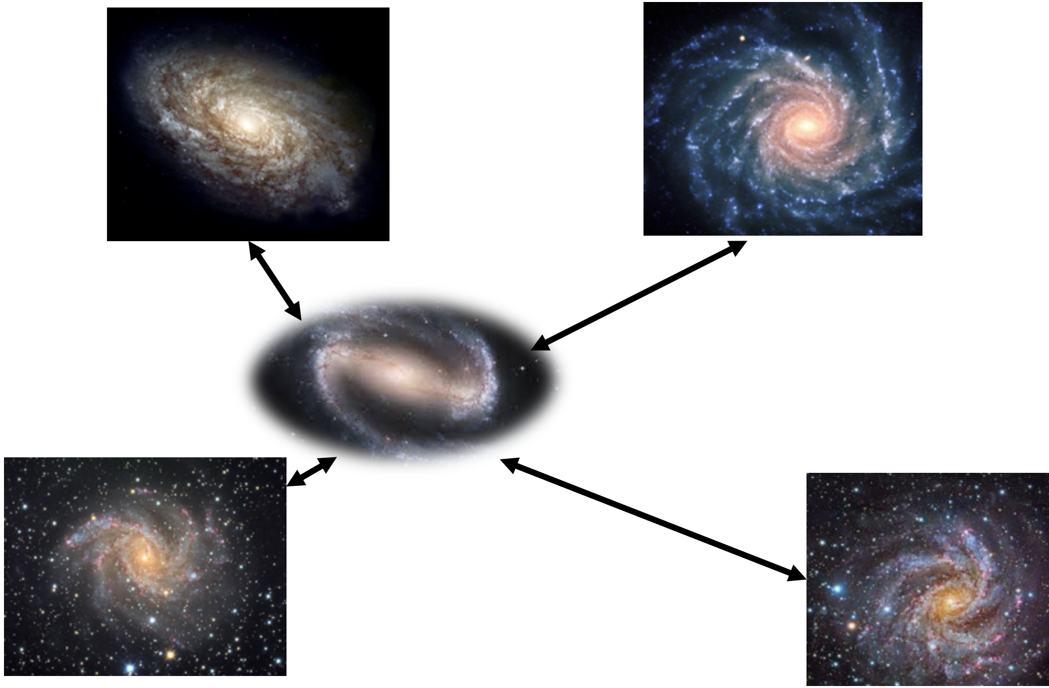
Value conditions

Intuitive interface for simple queries	Restricted to SQL syntax but not explicitly
SQL not required	Not example-based

Similarities are the key …

If we knew how similar each item is with respect to any other for **each** user, we would know the answer to





Similarities are the key …

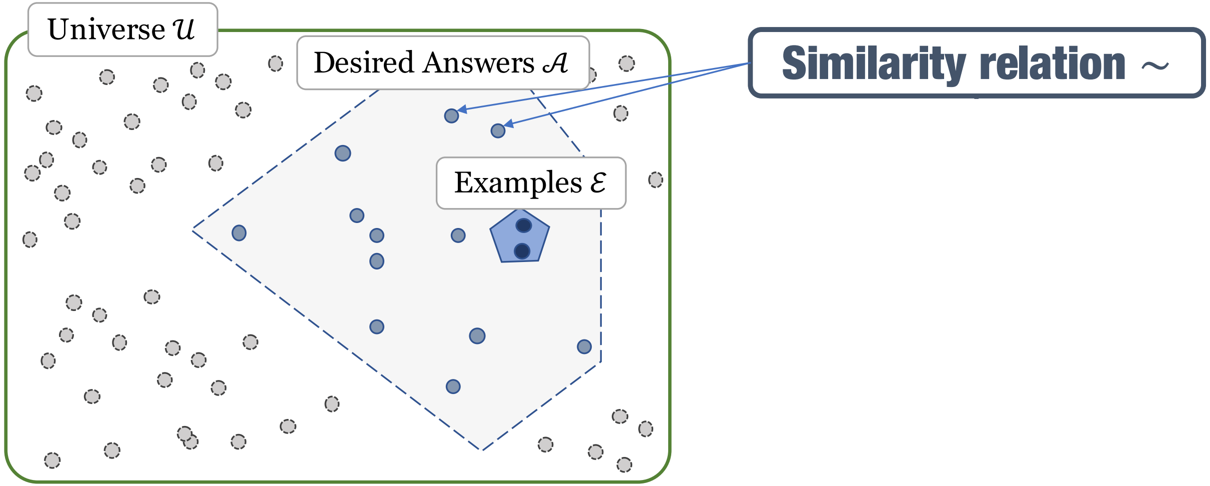
We define:

A universe $\mathcal{U}$ of items

A similarity among items $\sim$

A set of input examples $\mathcal{E}$

A set of output user desired answers $\mathcal{A}$



The example-based problem

Given

a set of examples $\mathcal{E}$ from a universe $\mathcal{U}$

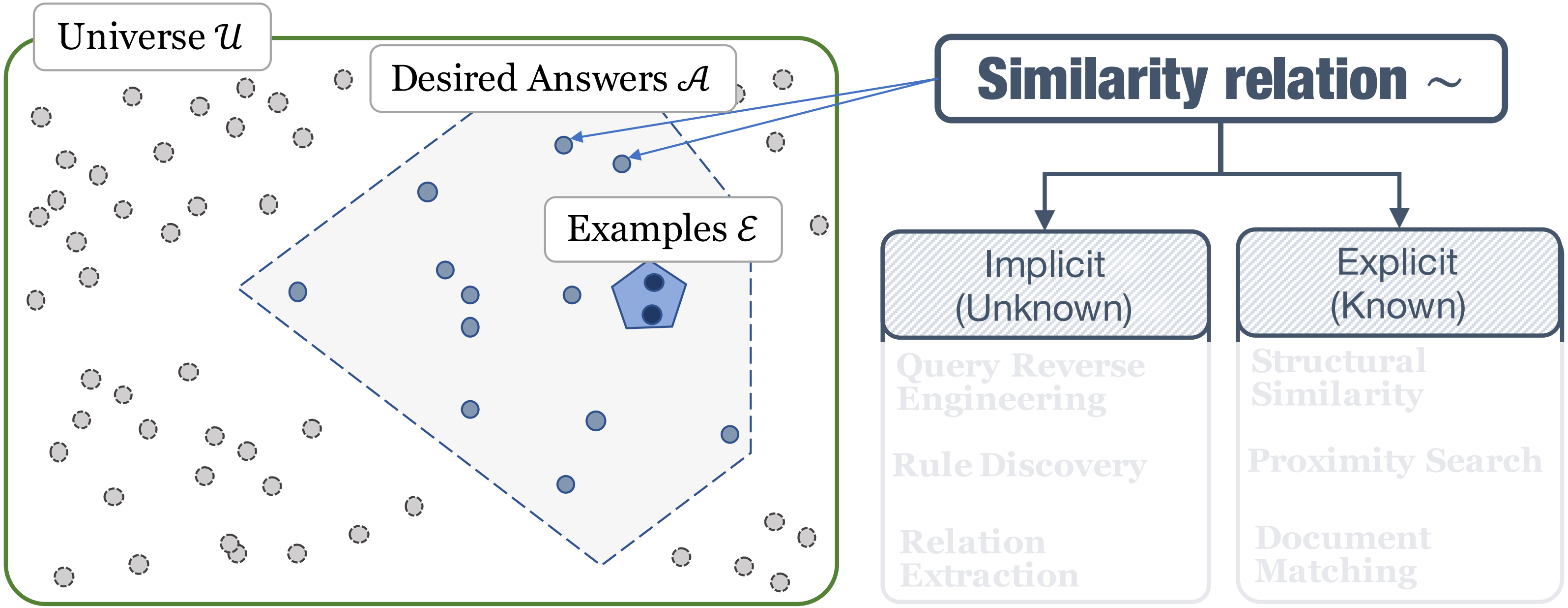
Find

a similarity $\sim$ such that

1. $\mathcal{E}$ is part of the answers $\mathcal{A}$ partially or totally
2. The answers in $\mathcal{A}$ are the most similar to the examples in $\mathcal{E}$ according to $\sim$

How do we find $\sim$ for each user?
Do we need to know exactly $\sim$?

Example-based methods



Example-based methods

Relational

Reverse engineering queries

Example-driven schema mapping

Interactive data repairing

Textual

Entity extraction by example text

Web table completion using examples

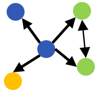
Search by example

Graph

Community-based Node-retrieval

Entity Search

Path and SPARQL queries
Graph structures as Examples



Tutorial structure

- Relational databases
- Textual data
- Graph and networks

Machine learning

Challenges and Remarks

Where we are

Relational databases

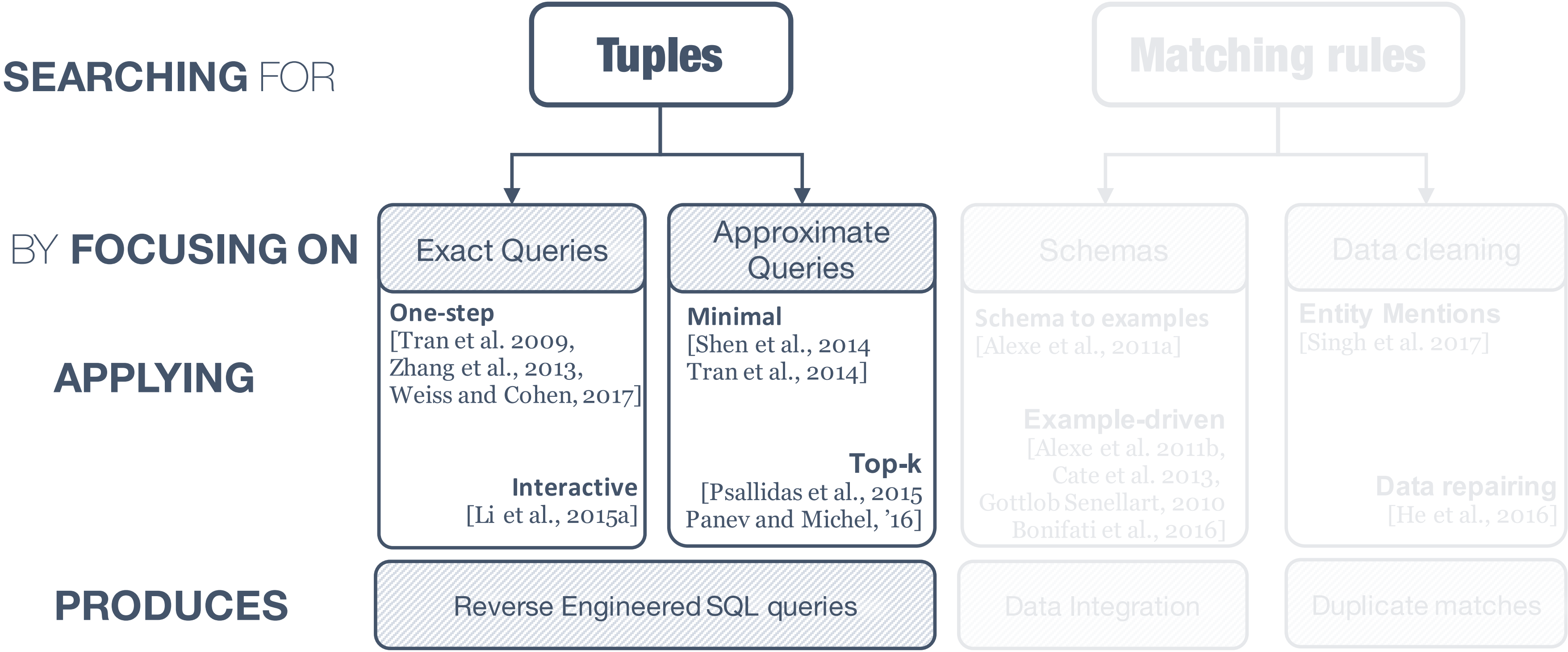
Textual data

Graphs and networks

Machine learning

Challenges and Remarks

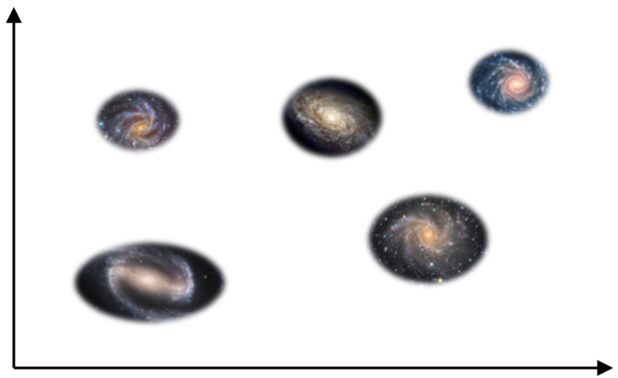
Searching for …



Reverse engineering queries (REQ)

Given a set of examples, find the query that generated that set of tuples

Example tuples



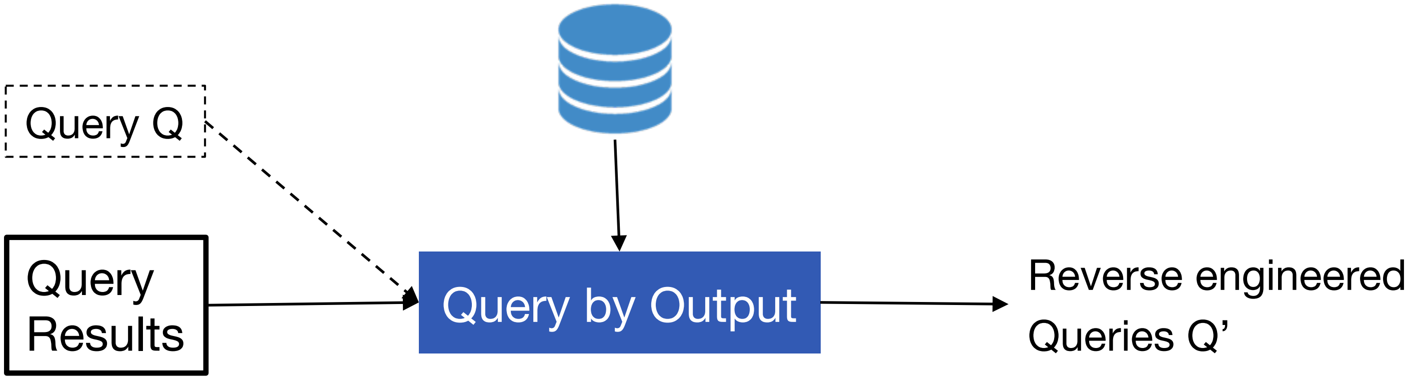
How do you find such queries?

```
SELECT g.galaxy_name, SUM(s.stars) AS st_s
FROM Universe.Galaxy  AS g
JOIN Universe.System AS s
ON g.galaxy_name = s.galaxy_name
WHERE
        g.st_s > 100B
        AND diameter > 100k AND diameter > 180k
        AND has_black_hole = TRUE
GROUP BY g.galaxy_name
```

```
SELECT galaxy_name
FROM Universe.Galaxy
```

Query by Output – TALOS (classification-based) [Tran et al. 2013]

Main idea: Find the set of queries that exactly return a set of examples



Two queries Q and Q’ are instance equivalent on a database D, if the results of Q are the same of the results of Q’

How many reverse engineered queries?

Master

	name	bat	throw	stint	weight	team
t_1	A	L	R	2	40	PIT
t_2	A	L	R	2	50	MT1
t_3	C	R	L	2	35	CHA
t_4	D	L	R	3	30	PIT
t_5	B	R	R	1	73	PIT
t_6	B	R	R	1	40	PIT
t_7	E	R	R	3	60	CHA

r_1	B	PIT
r_2	E	CHA

$Q(D)$

What queries generated Q(D)?

Q1 = SELECT name, team FROM Master WHERE bat = 'R' AND throw = 'R'

Q2 = SELECT name, team FROM Master WHERE bat = 'R' AND weight > 35

Q3 = SELECT name, team FROM Master WHERE bat = 'R' AND stint <> 2

...

Instance Equivalent Queries

TALOS

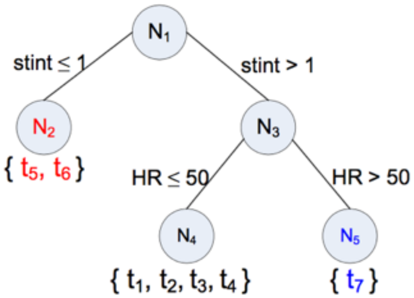
[Tran et al. 2013]

B	PIT
E	CHA

	name	bat	throw	stint	HR	team	
t_1	A	L	R	2	40	PIT	X
t_2	A	L	R	2	50	MT1	X
t_3	C	R	L	2	35	CHA	X
t_4	D	L	R	3	30	PIT	X
t_5	B	R	R	1	73	PIT	✓
t_6	B	R	R	1	40	PIT	✓
t_7	E	R	R	3	60	CHA	✓

Idea: treat the problem as a binary classification

1. **Strict:** all tuples must be captured
2. **At-Least-one:** one tuple for example must be captured



Decision tree

$$Gini(S_1, S_2) = \frac{(|S_1|Gini(S_1) + |S_2|Gini(S_2))}{|S_1| + |S_2|}$$

How complex is exact REQ?

[Weiss et al., 2017]

Database D

Relational Operators:

σ selection $\{=, \neq, \geq, \leq\}$

π projection

$\bowtie$ natural join

E^+ Positive examples
E^- Negative examples

REQ

Q such that results contain
- All positive examples
- No negative example

How difficult is to find:
A bounded size Q? an unbounded Q?

Complexity - No parameters

[Weiss et al., 2017]

Operator	Unbounded Queries	Bounded Queries
π	P	P
$\bowtie$	P	NPC
σ	P	NPC
$\sigma, \bowtie$	P	NPC
π, σ	NPC	NPC
$\sigma, \bowtie$	DP	DP
$\pi, \sigma, \bowtie$	DP	DP

Only projections: Easy

Unbounded selections: Easy

Bounded selections: HARD

Combination of operators: HARD!!!

Reduction from SAT

Unbounded Select

[Weiss et al., 2017]

	A	B	C	D	E
☑	1	2	3	4	5
☒	1	3	2	3	4
	2	4	4	1	3
	5	3	2	4	2
☒	4	2	3	1	2
	2	2	4	3	2
☑	1	1	2	1	5
☑	1	5	4	2	3

Possible queries?

$A = 1$ AND

$B \geq 1$ AND $B \leq 5$ AND

$C \geq 2$ AND $C \leq 4$ AND

$D \geq 1$ AND $D \leq 4$ AND $D \neq 3$

$E \geq 3$ AND $E \leq 5$ AND $E \neq 4$

Bounded select

Reduction from Set Cover

NP-C

INPUT: Database D, Examples E, Query size k

OUTPUT: Does there exist a query satisfying D and E, of size at most k?

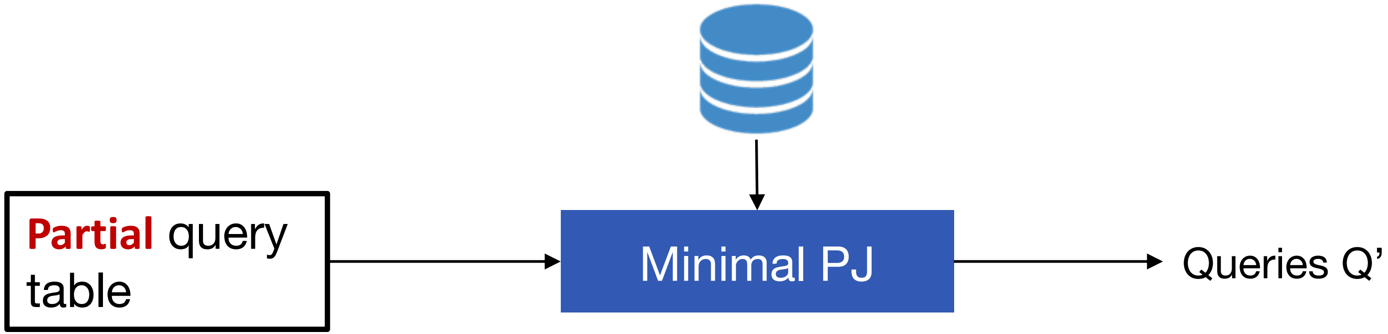
U = {1,2,3,4,5} S = { {1,2,3}, {2,4}, {3,4}, {4,5} }

	S_1	S_2	S_3	S_4
☒	1	0	0	0
☒	1	1	0	0
☒	1	0	1	0
☒	0	1	1	1
☒	0	0	0	1
☑	1	1	1	1

Minimal Project Join REQ

[Shen et al., 2014]

Main idea: Find the set of queries that **approximately** return a set of examples



	A	B	C
1	Mike	ThinkPad	Office
2	Mary	iPad	
3	Bob		Dropbox

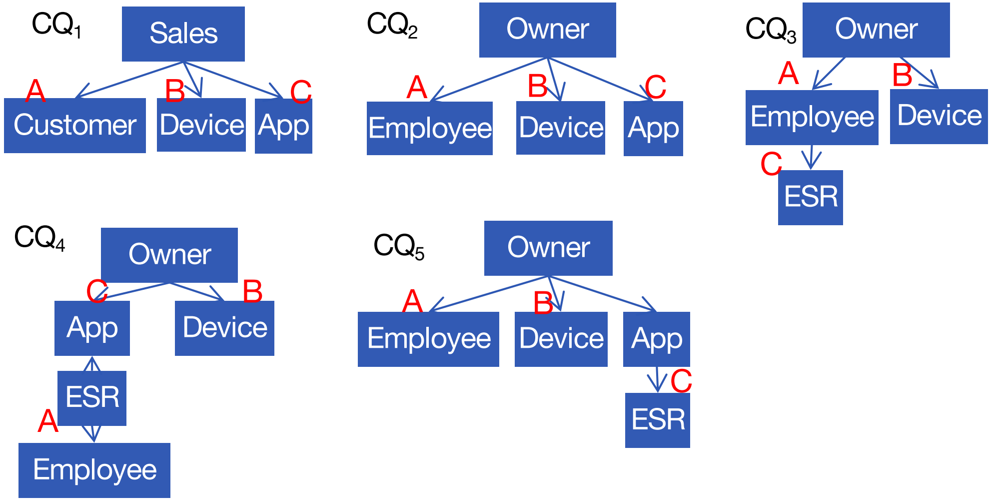
- valid: every tuple is present in query results
- minimal: any removal in query tree gets to an invalid query

Candidate Query Generation

[Shen et al., 2014]

- Use candidate network generation algorithm (Hristidis 2002)

	A	B	C
1	Mike	ThinkPad	Office
2	Mary	iPad	
3	Bob		Dropbox



1. Generate join tree J
2. Generate mapping ϕ
3. Check minimal:
   - Every leaf node contains a column that is mapped by an input column

Validity verification

[Shen et al., 2014]

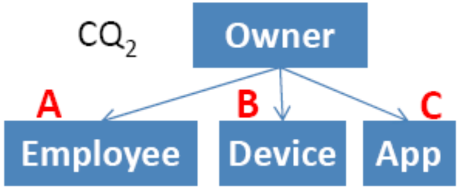
Naïve: check all candidate queries singularly if they return ALL examples

Better: exploit substructures in candidate queries for pruning

Best: adaptively select the substructures to have the min number of evaluations

NP-hard

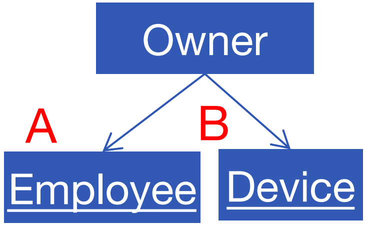
CQ_2



Candidate query

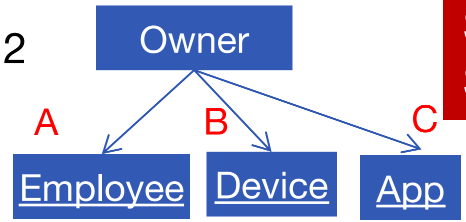
Substructures

Sub 1



Sub 1 fails => CQ_2 invalid

Sub 2



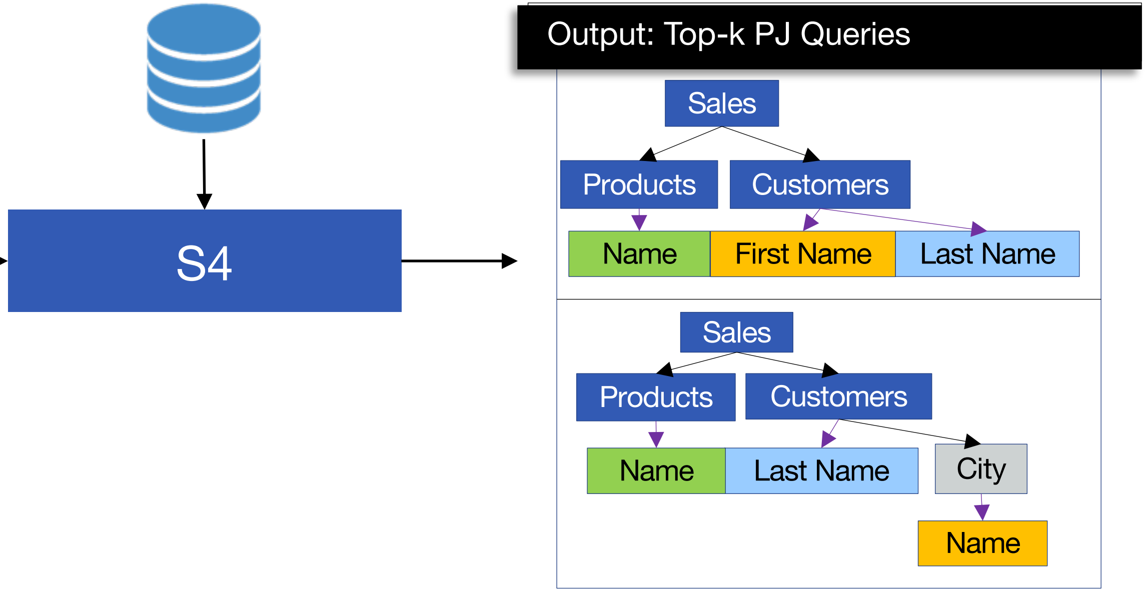
Sub 1 fails => Sub 2 fails

Minimal Project Join REQ

[Psallidas et al., 2015]

Main idea: Allow missing rows/columns and rank the k best queries

Partial query table



	A	B	C
1	John	Smith	Xbox
2	Jill	Hans	Surface

Ranking score

[Psallidas et al., 2015]

Linear combination of row score and column score

(Overlapping with the example table)

$$\frac{\alpha * score_{row}(Q) + (1 - \alpha) * score_{col}(Q)}{|Q|}$$

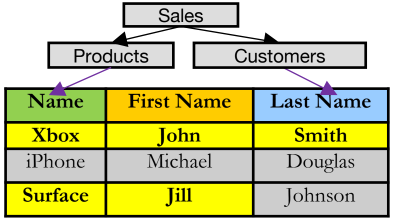
- $\alpha = 1$ penalizes missing rows
- $\alpha = 0$ penalizes missing columns

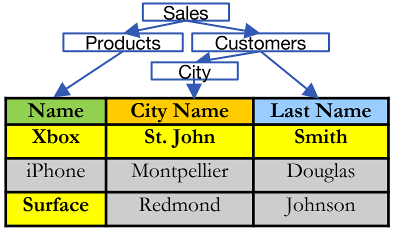
Row score

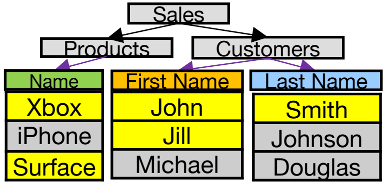
			Row Score	
John	Smith	Xbox	3	3
Jill	Hans	Surface	2	1
			5	4

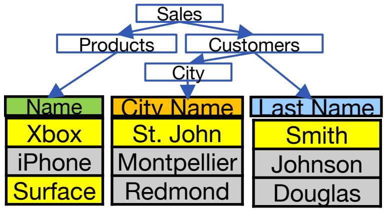
Column score

	John	Smith	Xbox	
	Jill	Hans	Surface	
Column Score	2	1	2	5
	2	1	1	4





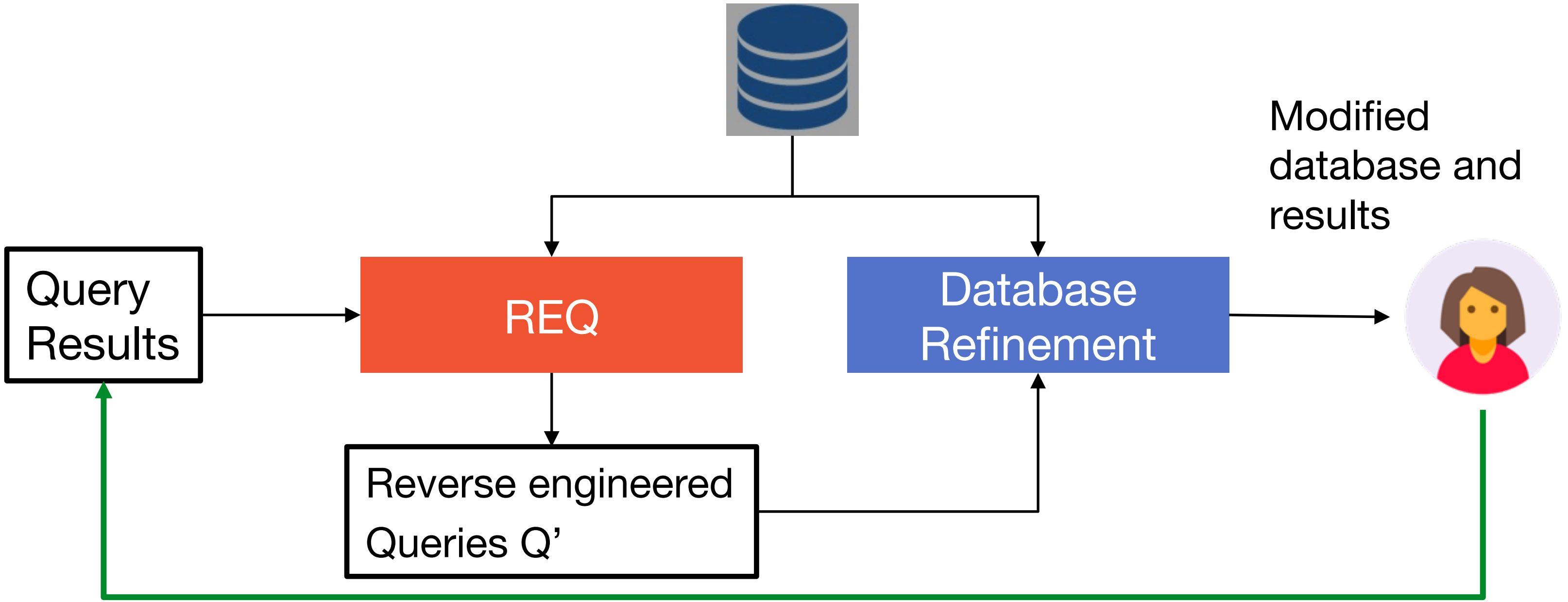




Interactive REQ – Query from Examples (cost model)

[Li et al., 2015]

Main idea: Interactively remove candidate queries proposing a new set of query results from a modified database



Database Refinement

[Li et al., 2015]

eid	name	gender	dept	salary
1	Alice	F	Sales	3700
2	Bob	M	IT	4200
3	Carol	F	Service	3000
4	Dave	M	IT	5000

Results

name
Bob
Dave

REQs =

- $Q_1 = \sigma_{gender=M}(D)$
- $Q_2 = \sigma_{salary>3700}(D)$
- $Q_3 = \sigma_{dept=IT}(D)$

Database Refinement

eid	name	gender	dept	salary
1	Alice	F	Sales	3700
2	Bob	M	IT	3000 ~~4200~~
3	Carol	F	Service	3000
4	Dave	M	IT	5000

name
Bob
Dave

Result $R_1' = Q_1(D') = Q_3(D')$

name
~~Bob~~
Dave

Result $R_2' = Q_2(D')$

Cost model

[Li et al., 2015]

Number of modified tables

Choice to go with the max tuples modified

Number of new result sets

$$cost(D') = \boxed{edit(D, D') + \beta \cdot n} + \boxed{\sum_{i=1}^{k} edit(R, R_i)} + \boxed{N \cdot \frac{edit(D, D')}{\mu} + \beta} + \boxed{\frac{2}{k}\sum_{i=1}^{k} edit(R, R_i)}$$

DB cost | Results cost | Effort to examine D’ | Effort to examine new results

Current cost | Residual cost

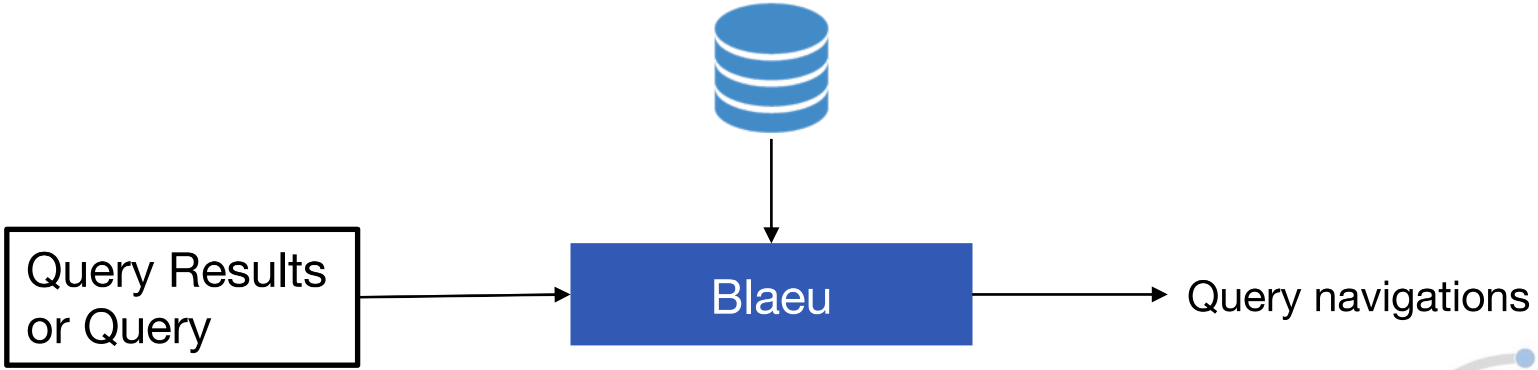
Main idea: Find a refined db D’ and results $R_1, \ldots R_k$ with:

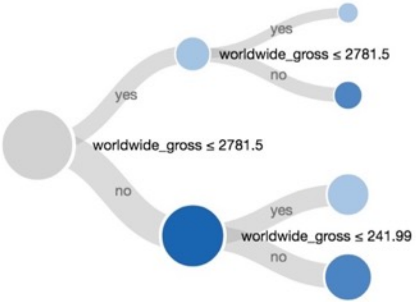
1. Minimum number of results k
2. Minimum differences i the database
3. The query are balanced (less interactions)

Examples for query suggestion: Blaeu (Clustering)

[Sellam et al., 2016]

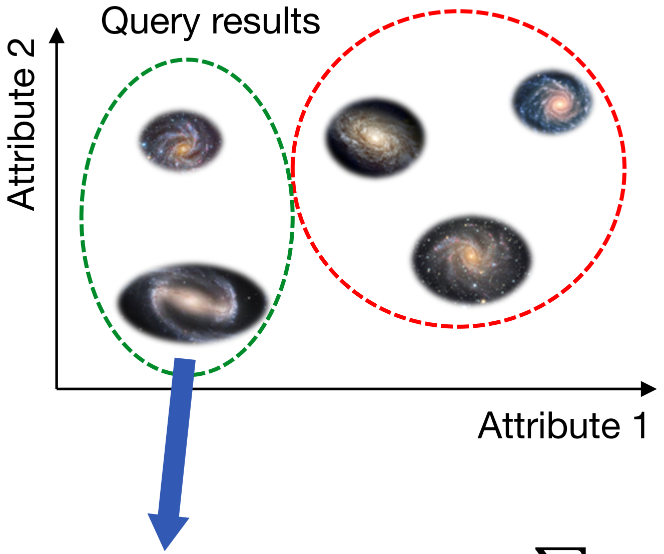
Main idea: Allow interactive navigation of the query space in a hierarchy





Examples for query suggestion: Blaeu

[Sellam et al., 2016]



$$u: DB \rightarrow \{-1,1\}, U(Q) = \sum_{t \in Q} u(t)$$

User utility

Given a result of an example query Q, explore the data through data maps = partitions

Output: Set of query refinements

Problem: User utility is unknown

- Cluster analysis for result exploration
- Zoom and projection operations
- User model

Examples for query suggestion: Blaeu

[Sellam et al., 2016]

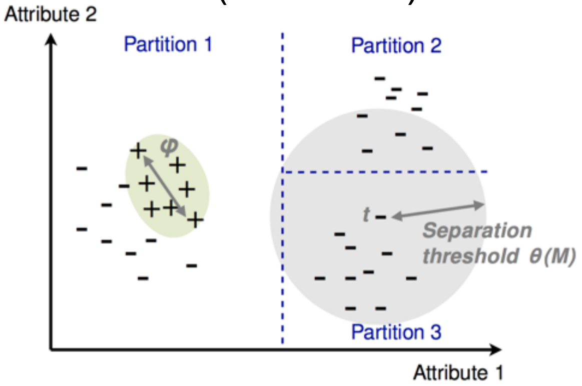
$$u: DB \rightarrow \{-1,1\}, U(C) = \sum_{t \in C} u(t)$$

Unknown User utility

Find the partition $\mathcal{C} = \{C_1, \ldots, C_n\}$ of the results of Q such that exists $C_j \in \mathcal{C}: U(C_j) > U(Q)$

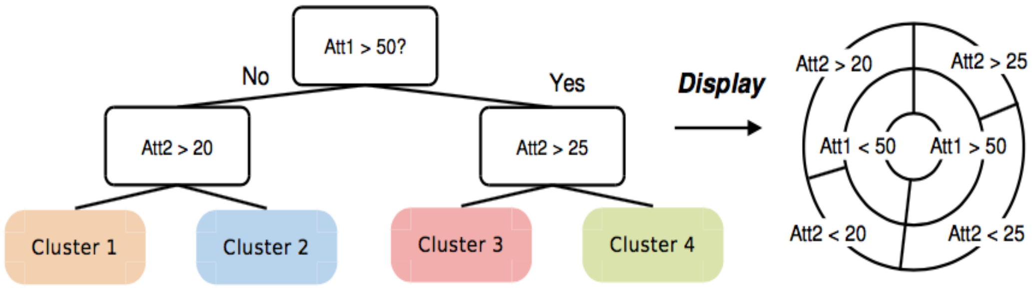
Solution: interesting tuples are close to each other within a maximum separation threshold $\theta(\mathcal{C})$

Detect clusters (k-medoid)

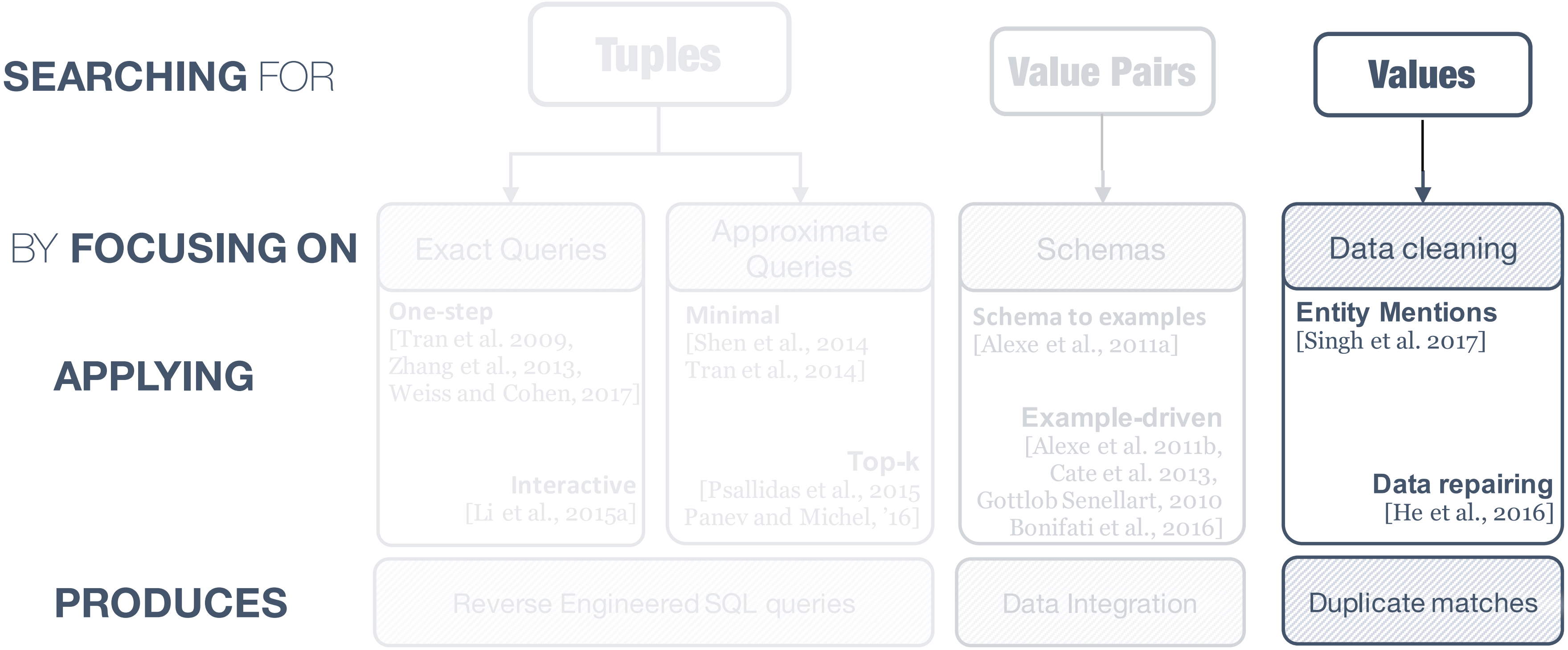


Inference

Organize clusters



Searching for …



Data Cleaning

- Often data have redundancy, wrong values, and missing values
- Different values can represent the same object (e.g., N.Y. and New York)
- Values can be simply wrong

Data cleaning refers to ways of making the data consistent and correct

tid	Date	Molecule	Laboratory	Quantity
t1	11 Nov	$C_{16}H_{16}Cl$	Austin	200
t2	12 Nov	statin	Austin	100
t3	12 Nov	$C_{24}H_{75}S_6$	N.Y.	100
t4	12 Nov	statin	Boston	200
t5	13 Nov	statin	Austin	200
t6	15 Nov	$C_{17}H_{20}N$	Dubai	1000

⇩

tid	Date	Molecule	Laboratory	Quantity
t1	11 Nov	$C_{16}H_{16}Cl$	Austin	200
t2	12 Nov	$C_{22}H_{28}F$	Austin	100
t3	12 Nov	$C_{24}H_{75}S_6$	New York	100
t4	12 Nov	statin	Boston	200
t5	13 Nov	$C_{22}H_{28}F$	Austin	200
t6	15 Nov	$C_{17}H_{20}N$	Dubai	100

Data repairing: rules

[He, J. et al. 2016]

A rule is a logical formula which determines how to change the value in a cell or a group of cells.

IF $[X_1 = C_1 \ldots X_n = C_n]$ UPDATE X_i to some value

- The update t_3[Laboratory] ← ”New York” can be obtained by the rule
- IF [Laboratory = ”N.Y.”] UPDATE Laboratory to ”New York”
- ```
  UPDATE Table
  SET Laboratory='New York'
  WHERE tid=t3
  ```

BUT it needs to be done for each cell!!

tid	Date	Molecule	Laboratory	Quantity
t1	11 Nov	$C_{16}H_{16}Cl$	Austin	200
t2	12 Nov	statin	Austin	100
t3	12 Nov	$C_{24}H_{75}S_6$	N.Y.	100
t4	12 Nov	statin	Boston	200
t5	13 Nov	statin	Austin	200
t6	15 Nov	$C_{17}H_{20}N$	Dubai	1000

⇩

tid	Date	Molecule	Laboratory	Quantity
t1	11 Nov	$C_{16}H_{16}Cl$	Austin	200
t2	12 Nov	$C_{22}H_{28}F$	Austin	100
t3	12 Nov	$C_{24}H_{75}S_6$	New York	100
t4	12 Nov	statin	Boston	200
t5	13 Nov	$C_{22}H_{28}F$	Austin	200
t6	15 Nov	$C_{17}H_{20}N$	Dubai	100
  ```

Interactive data cleaning: problem

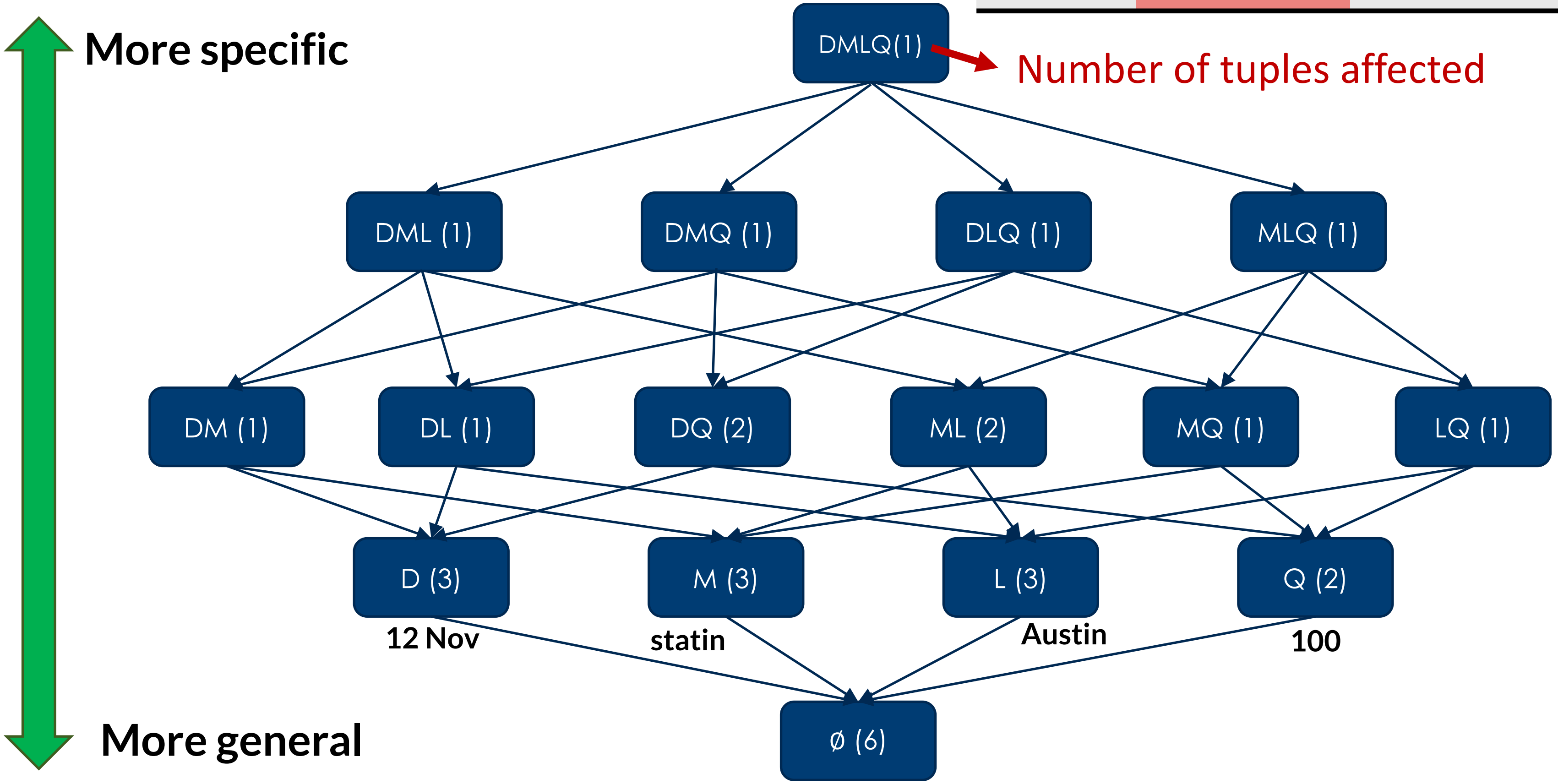
User validates rules, but has no capacity to validate all rules for each update.

- **Budget Repair Problem:** Given a set $\mathcal{Q}$ of rules, a table T and a budget B, **find B rules from $\mathcal{Q}$ to maximize the number of repairs over T**
- Budget repair problem is an *online problem*

Corresponding *offline problem* is: given as input $\mathcal{Q}$ rules where validity of each rule is known, select B rules from $\mathcal{Q}$ to maximize the number of repairs over T. (NP-Hard)

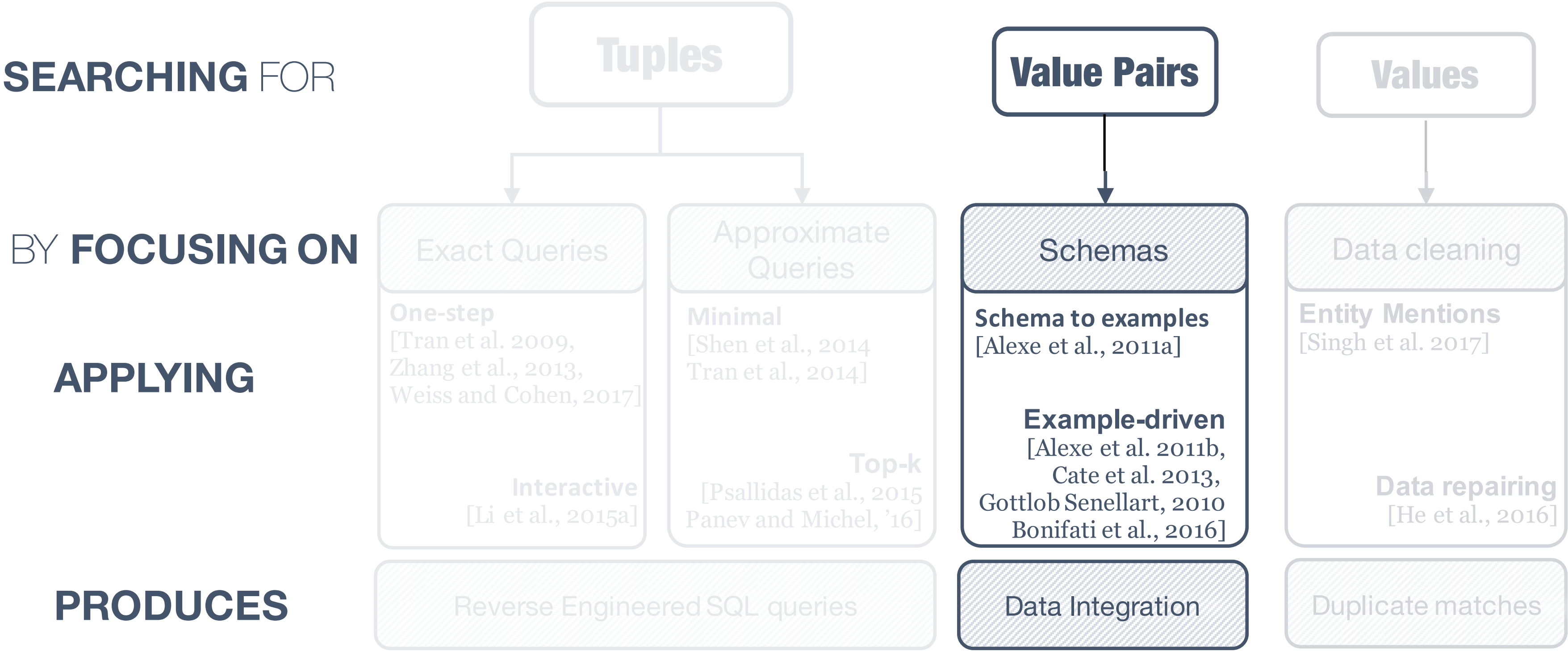
Rule lattice

tid	Date	Molecule	Laboratory	Quantity
t2	12 Nov	statin→$C_{22}H_{28}F$	Austin	100



M. Lissandrini, D. Mottin, T. Palpanas, Y. Velegrakis

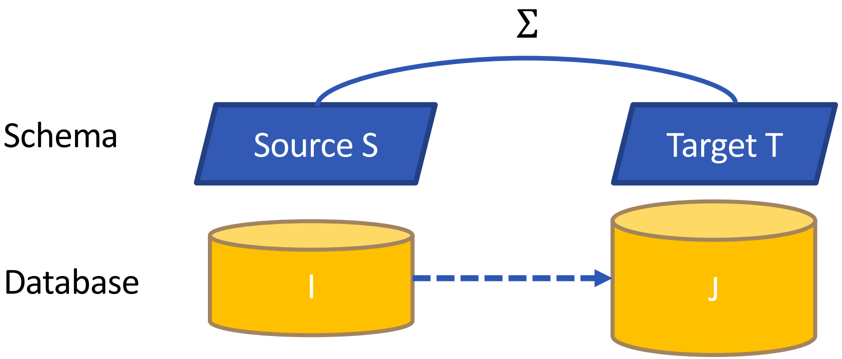
Searching for …



Schema mapping

- Schema mapping finds a way to represent items on one database to items on another database
- Finds a mapping Σ between two schemas such that a query on one database can be converted to a query on the other database
- Schema mappings in Σ are rules in first-order logic that specifies the relationships between schema S and T

$$\forall x \forall y\; S(x,y) \wedge U(x,z) \rightarrow \exists v\; T(v,y) \wedge T'(v,z)$$



Mapping generation

[Bonifati et al. 2017]

E_S:

Company

IdCompany	*Name*	*Town*
'C1'	'AA'	'Paris'
'C2'	'Ev'	'Lyon'

Flight

Departure	*Arrival*	*IdCompany*
'Lyon'	'Paris'	'C1'
'Paris'	'Lyon'	'C2'

Travel Agency

IdAgency	*Name*	*Town*
'A1'	'TC'	'L.A.'

E_T:

Firm

Id	*Name*	*Town*
'Id1'	'AA'	'Paris'
'Id2'	'Ev'	'Lyon'
'Id3'	'TC'	'L.A.'

Departure

Town	*IdFirm*
'Lyon'	'Id1'
'Paris'	'Id2'

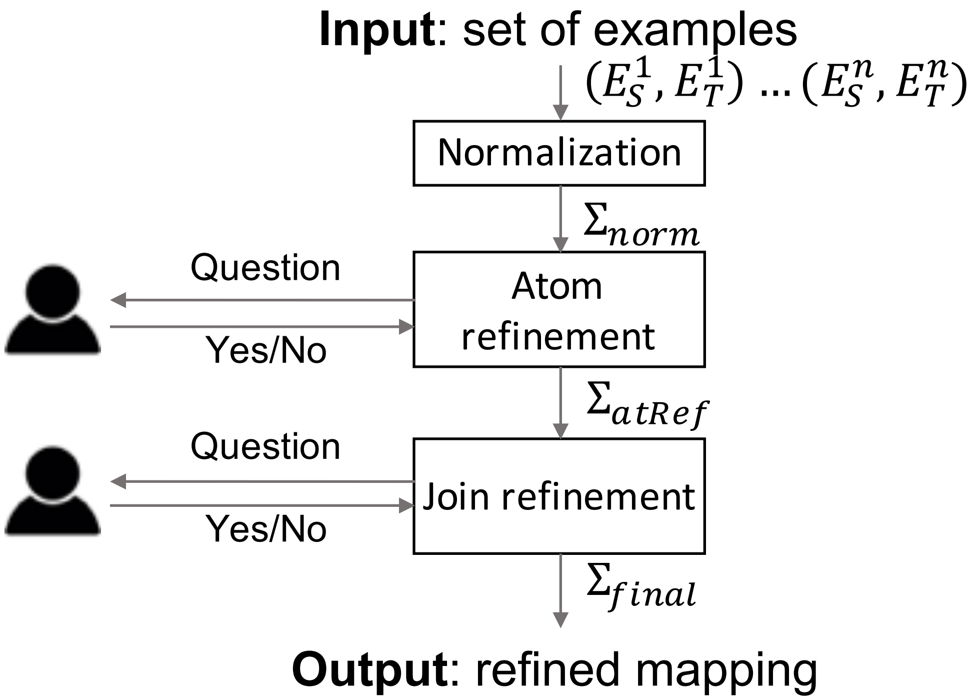
Arrival

Town	*IdFirm*
'Paris'	'Id1'
'Lyon'	'Id2'

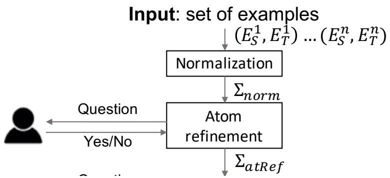
$$\begin{aligned}\mathbf{m}: &\ \mathit{Company}(\mathrm{c1}, \mathrm{aa}, \mathrm{paris}) \wedge \mathit{Company}(\mathrm{c2}, \mathrm{ev}, \mathrm{lyon}) \wedge \mathit{TravelAgency}(\mathrm{a1}, \mathrm{tc}, \mathrm{la}) \\ &\wedge \mathit{Flight}(\mathrm{lyon}, \mathrm{paris}, \mathrm{c1}) \wedge \mathit{Flight}(\mathrm{paris}, \mathrm{lyon}, \mathrm{c2}) \\ &\rightarrow \exists \mathrm{id1}, \mathrm{id2}, \mathrm{id3}, \mathit{Firm}(\mathrm{id1}, \mathrm{aa}, \mathrm{paris}) \wedge \mathit{Departure}(\mathrm{lyon}, \mathrm{id1}) \wedge \mathit{Arrival}(\mathrm{paris}, \mathrm{id1}) \\ &\quad \wedge \mathit{Firm}(\mathrm{id2}, \mathrm{ev}, \mathrm{lyon}) \wedge \mathit{Departure}(\mathrm{paris}, \mathrm{id2}) \wedge \mathit{Arrival}(\mathrm{lyon}, \mathrm{id2}) \wedge \mathit{Firm}(\mathrm{id3}, \mathrm{tc}, \mathrm{la})\end{aligned}$$

Interactive Mapping

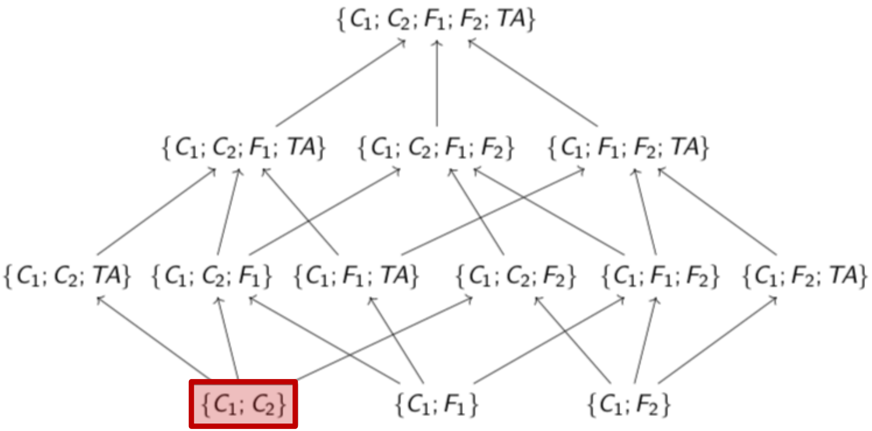
[Bonifati et al. 2017]



Atom Refinement

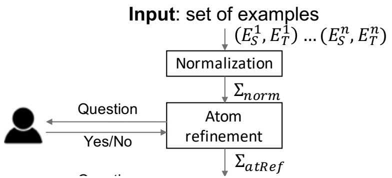


Ask the user and refine the left part of the rule

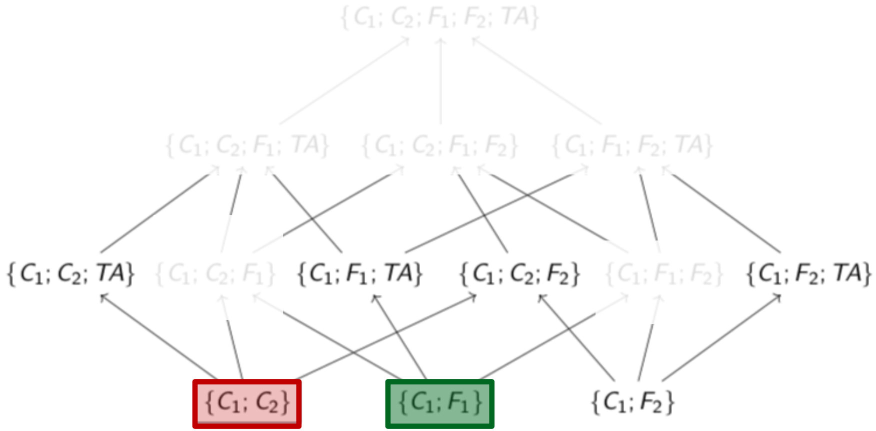


Are the tuples Company(c1,aa,paris); Company (c2, ev, lyon) enough to produce Firm(id, aa, Paris); Departure (Lyon, id); Arrival(Paris, id)?

Atom Refinement



Ask the user and refine the left part of the rule



Are the tuples Company(c1,aa,paris); Flight (lyon, paris, c1) enough to produce Firm(id, aa, Paris); Departure (Lyon, id); Arrival(Paris, id)?

Where we are

Relational databases

Textual data

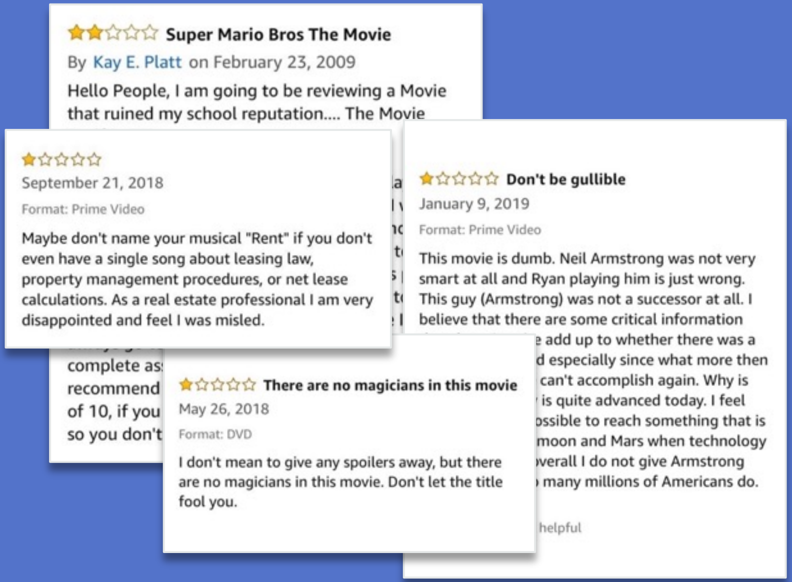
Graphs and networks

Machine learning

Challenges and Remarks

SIMILARITY for TEXTUAL DATA

Unstructured

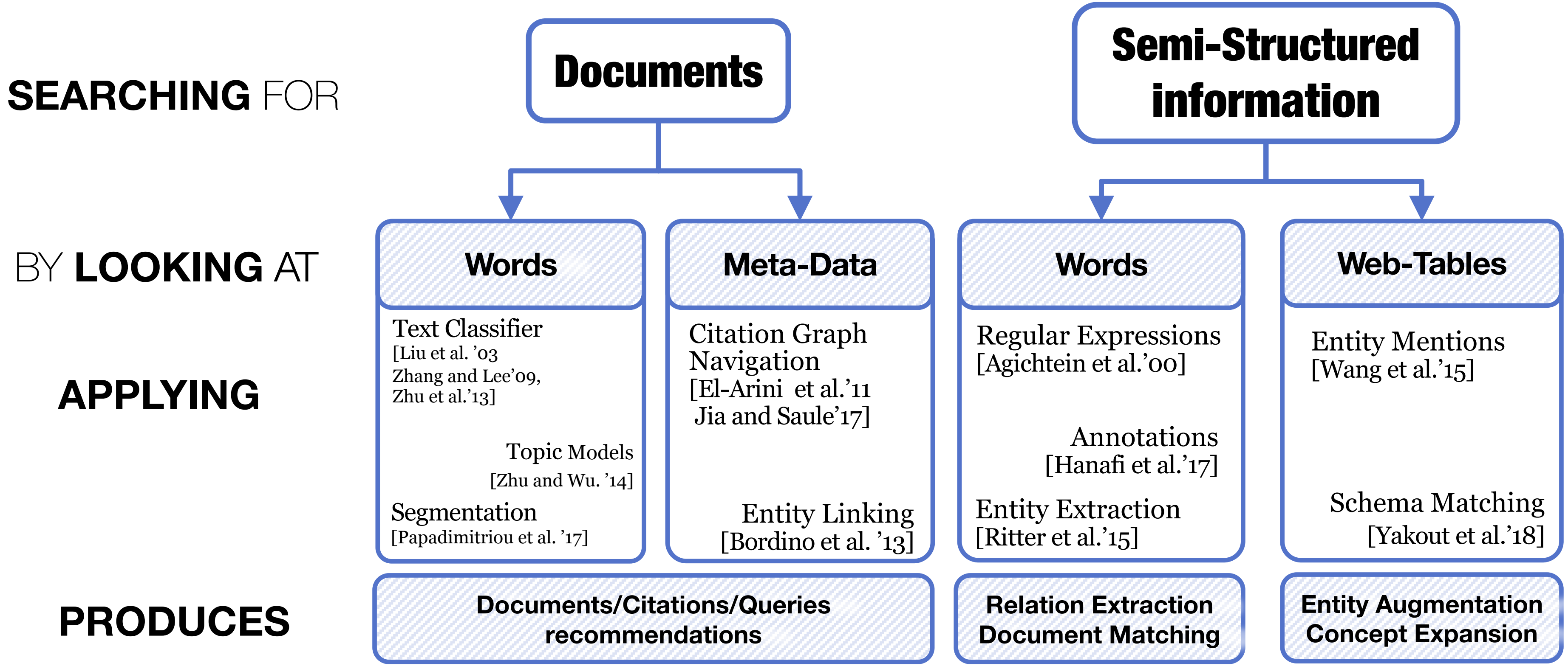


Semi-Structured

HR Information		Contact	
Position	Salary	Office	Extn.
Accountant	$162,700	Tokyo	5407
Chief Executive Officer (CEO)	$1,200,000	London	5797
Junior Technical Author	$86,000	San Francisco	1562
Software Engineer	$1		
Software Engineer	$2		
Integration Specialist	$3		
Software Engineer	$1		
Pre-Sales Support	$1		
Sales Assistant	$1		
Senior Javascript Developer	$4		

Category	Structure	Country	City	Height (metres)	Height (feet)	Year built	Coordinates
Mixed use	Burj Khalifa	United Arab Emirates	Dubai	828.1	2,717	2010	25°11′50.0″N 55°16′26.6″E
Self-supporting tower	Tokyo Skytree	Japan	Tokyo	634	2,080	2011	35°42′36.5″N 139°48′39″E
Guyed steel lattice mast	KVLY-TV mast	United States	Blanchard, North Dakota	628.8	2,063	1963	47°20′32″N 97°17′25″W
Clock building	Abraj Al Bait Towers	Saudi Arabia	Mecca	601	1,972	2011	21°25′08″N 39°49′35″E
Mixed use	Lotte World Tower	South Korea	Seoul	555.7	1823	2017	37°30′45″N 127°6′10″E
Office	One World Trade Center	United States	New York, NY	541	1,776	2013	40°42′46.8″N 74°0′48.6″W
Military structure	Large masts of INS Kattabomman	India	Tirunelveli	471	1,545	2014	8°22′42.52″N 77°44′38.45″E ; 8°22′30.13″N 77°45′21.07″E
		United ...es	Lualualei, Hawaii	458	1,503	1972	21°25′11.87″N 158°08′53.67″W ; 21°25′13.38″N 158°09′14.35″W
		Malaysia	Kuala Lumpur	452	1,482	1998	3°09′27.45″N 101°42′40.7″E; 3°09′29.45″N 101°42′43.4″E
		United ...es	New York	425.5	1,396	2015	

Company	Contact	Country
Launch Pad	Maria Anders	Germany
Pay Talk	Francisco Chang	Mexico
Earn More	Roland Mendel	Austria
Island Trading	Helen Bennett	UK



SEARCHING FOR
BY LOOKING AT
APPLYING
PRODUCES
Documents
Semi-Structured information
Words
Meta-Data
Words
Web-Tables
Text Classifier
[Liu et al. '03
Zhang and Lee'09,
Zhu et al.'13]
Topic Models
[Zhu and Wu. '14]
Segmentation
[Papadimitriou et al. '17]
Citation Graph Navigation
[El-Arini et al.'11
Jia and Saule'17]
Entity Linking
[Bordino et al. '13]
Regular Expressions
[Agichtein et al.'00]
Annotations
[Hanafi et al.'17]
Entity Extraction
[Ritter et al.'15]
Entity Mentions
[Wang et al.'15]
Schema Matching
[Yakout et al.'18]
Documents/Citations/Queries recommendations
Relation Extraction
Document Matching
Entity Augmentation
Concept Expansion

Document Search

Keyword Queries
& Relevance

Keyword query: search text with text

“Action movie with magic”

Search documents containing *those exact words*

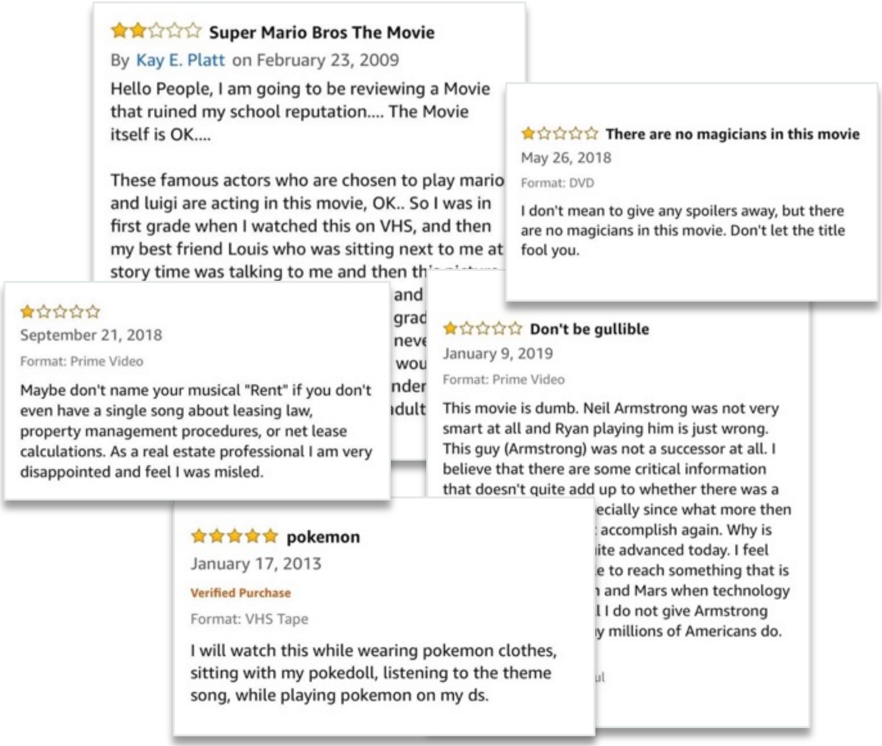
… a live action movie…

…. there is plenty of action…

… packed with action…

… Magic Mike is comedy movie …

… in Harry Potter magic is everywhere..



Is this enough?
Identify “relevant words” and “relevant documents”

Document Search

Relevant Keywords

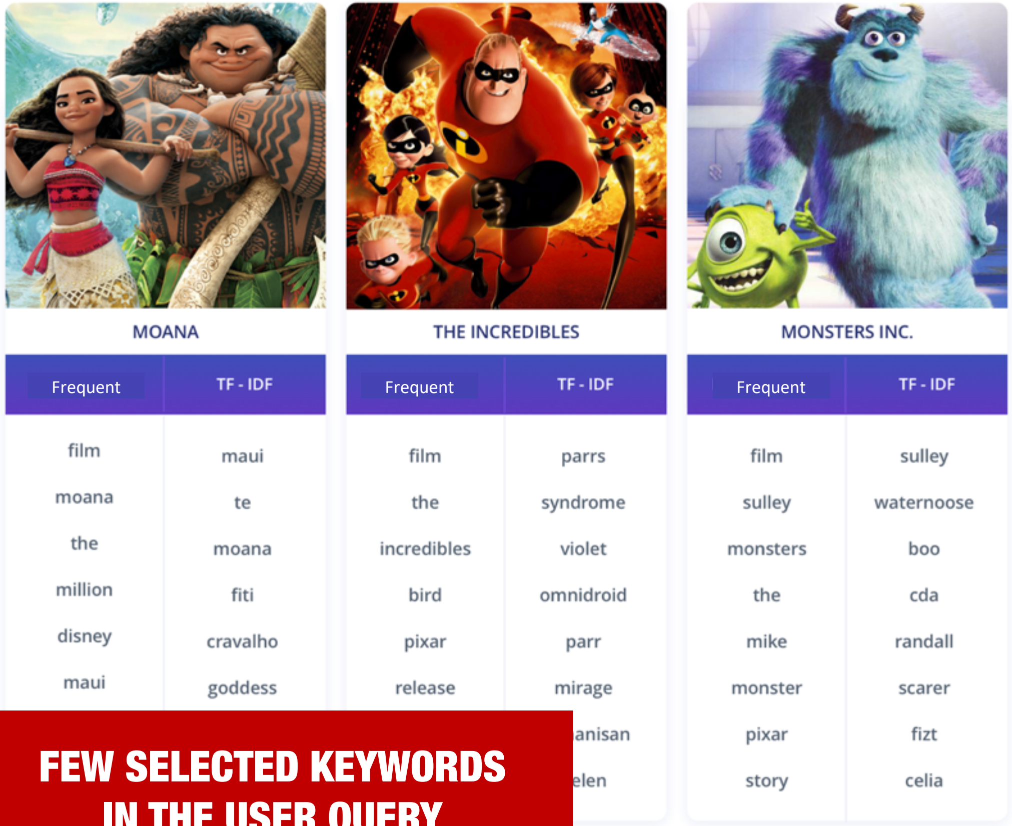
Relevance: which keywords are more helpful in describing the content of the document?

Relevance ≠ Frequency

What keywords are more likely to be used to describe the document we want and not other documents

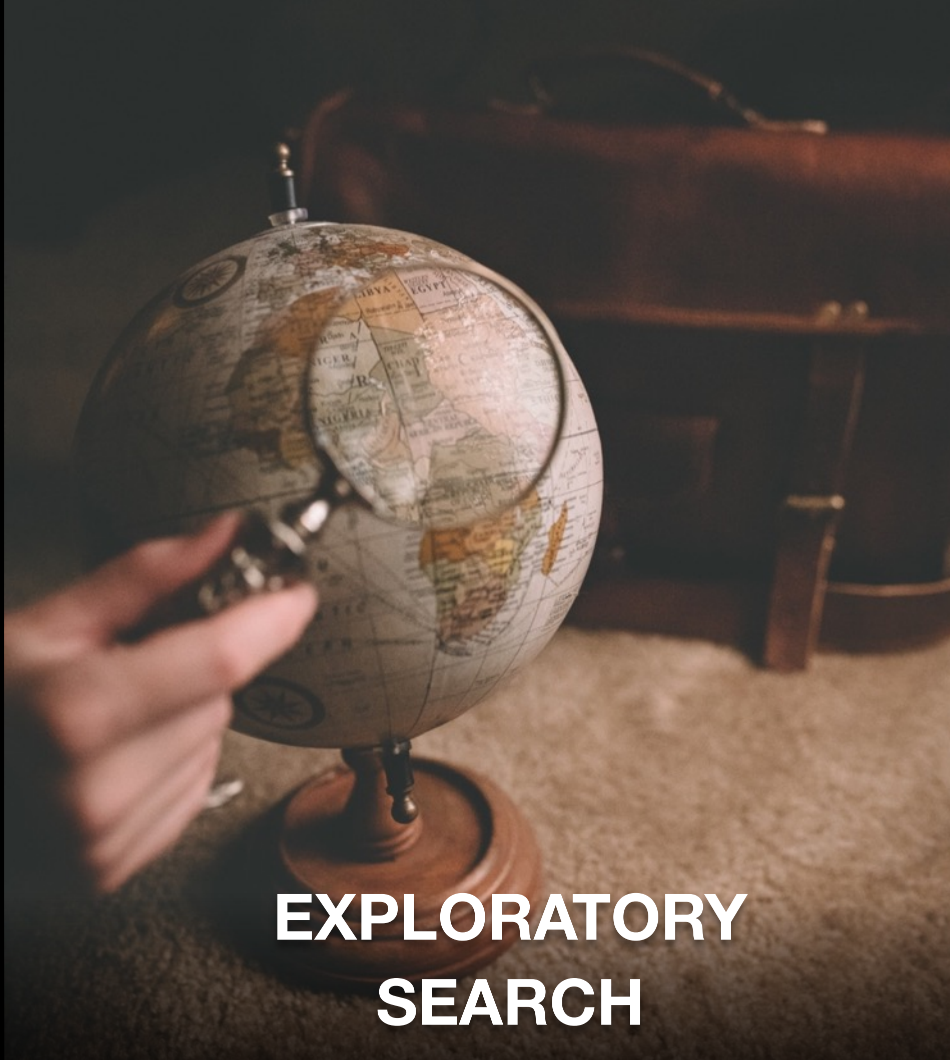
1. Term-frequency: how many times the term appears in the document
2. Document-frequency: In how many documents the term appears

TF-IDF: Term Frequency Inverse Document Frequency



FEW SELECTED KEYWORDS IN THE USER QUERY

What keywords to choose?


TRADITIONAL
SEARCH

EXPLORATORY
SEARCH

SEARCHING FOR

Documents

Semi-Structured information

BY **LOOKING** AT

APPLYING

PRODUCES

	Documents		Semi-Structured information	
BY **LOOKING** AT	**Words**	**Meta-Data**	Words	Web-Tables
APPLYING	Text Classifier [Liu et al. '03 Zhang and Lee'09, Zhu et al.'13] Topic Models [Zhu and Wu. '14] Segmentation [Papadimitriou et al. '17]	Citation Graph Navigation [El-Arini et al.'11 Jia and Saule'17] Entity Linking [Bordino et al. '13]	Regular Expressions [Agichtein et al.'00] Annotations [Hanafi et al.'17] Entity Extraction [Ritter et al.'15]	Entity Mentions [Wang et al.'15] Schema Matching [Yakout et al.'18]
PRODUCES	**Documents/Citations/Queries recommendations**		Relation Extraction Document Matching	Entity Augmentation Concept Expansion

Documents as Examples

Liu et al. [2003]

Exemplar documents

Set of exemplar documents

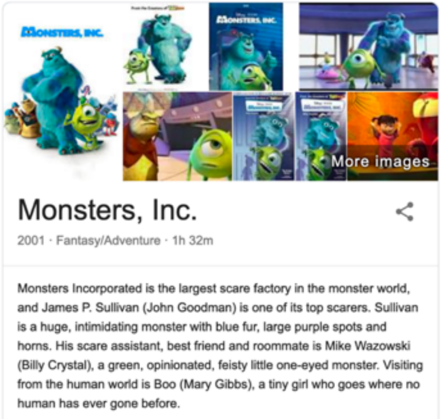
rather than a set of keywords.

An entire document may contain more information!

It also contains more noise

Identify what makes them special, i.e., relevant

Find me movies like these:



Example-based Document Search
Given a corpus of documents D,
and a small set of relevant documents (D_{rel}),
identify a set of answer documents D_A
such that $D_{rel} \subseteq D_A \subseteq D$.

Model as a classification problem!

PROBLEM:MISSING NEGATIVE CLASS

Few positive examples and
a ***large set of unknown.***

What *features* can *discriminate* relevant and irrelevant?

Would be better to have *some negative examples*

Text Classifiers

Liu et al. [2003]

Using Positive and Unlabeled Examples

Positive Unlabeled learning

- a corpus of documents D,
- 2 Classes: relevant $\top$ & irrelevant $\perp$
- relevant documents (D_{rel})

  $\forall d \in D_{rel}$. class(d) = $\top$
- Unlabeled documents $U = D - D_{rel}$

Goal:

train a classifier $C : D \rightarrow \{\top, \perp\}$,
to predict class(u) $\forall$ u $\in$ U.

Missing:
To train C we need examples
for the negative class $\perp$

Algorithm 4.9 Document Classification with Positive and Unalabled Data

Input: Relevant Documents $\mathbf{D}_{rel} \subseteq \mathcal{D}$, Unlabeled Documents $\mathbf{U} \subseteq \mathcal{D}$
Output: Classifier $\mathbb{C}$

1: $\mathbf{D}_{neg} \leftarrow$ getNegativeSample($\mathbf{U}$) $\triangleright$ See Li and Liu [2003], Liu et al. [2002], Yu et al. [2002]
2: $\mathbb{C} \leftarrow$ trainClassifier($\mathbf{D}_{rel}, \mathbf{D}_{neg}, \mathbf{U} \setminus \mathbf{D}_{neg}$) $\triangleright$ E.g., Expectation Maximization, SVM, or Rocchio
3: **return** $\mathbb{C}$

Inferring Negative Examples (I)

Liu et al. [2003]

Assign a label to Unlabeled data:

how to determine a negative sample set without asking the user

Goal:
Determine set of elements to be regarded as reliable negatives (RN)

Train a "simplistic" classifier

4 Alternative approaches

- **Naïve Bayes** (McCallum et al. [1998])
  - All unlabeled data are assumed negatives
  - NB-Classifier estimates **P(c|d)** based on on **P(w|c)** with $\mathbf{c} \in \{\top, \bot\}$, $\mathbf{d} \in \mathbf{D}$, and words $\mathbf{w} \in \mathbf{W}$

- The **Rocchio** technique (Raskutti et al. [2002])
  - $\forall d \in D$ $\vec{d}$ is the TF-IDF vector representation
  - Build prototype vectors $\vec{c}_\top$ for documents in D_{rel}
  - and $\vec{c}_\bot$ for documents in U
  - Compare each $\forall d \in U$ with $\vec{c}_\top$ and $\vec{c}_\bot$
  - assign the class of the most similar vector

$$\vec{c}_\top = \alpha \frac{1}{|\mathbf{D}_{rel}|} \sum_{d \in \mathbf{D}_{rel}} \frac{\vec{d}}{||\vec{d}||} - \beta \frac{1}{|\mathbf{U}|} \sum_{d \in \mathbf{U}} \frac{\vec{d}}{||\vec{d}||}$$

Inferring Negative Examples (II)

Liu et al. [2003]

Assign a label to Unlabeled data:

how to determine a negative sample set without asking the user

Goal:
Determine set of elements to be regarded as reliable negatives (RN)

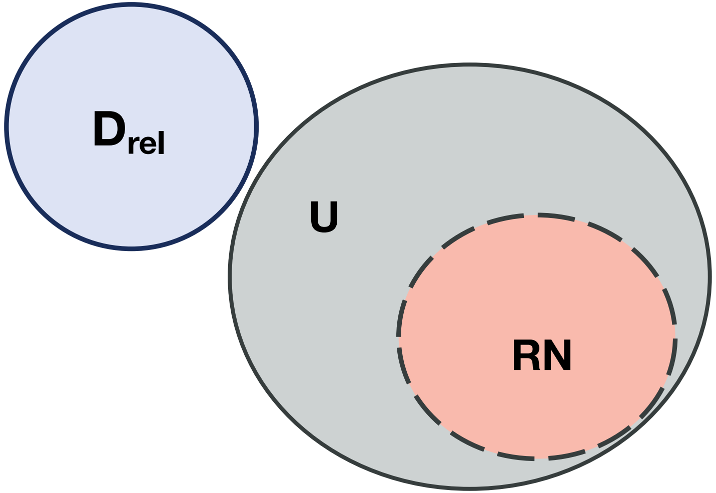
Train a "simplistic" classifier

4 Alternative approaches

- The **Spy** technique (Liu et al. [2002])
  - Extract a sample S from the positive example
  - Merge S in U (deploy the spies!)
  - Build NB classifier with EM
  - Determine threshold t such that all spies are correctly classified
  - Document above the threshold are considered negative

- **1-DNF*** technique (Yu et al. [2002]).

  **Disjunctive Normal Form*
  - *Positive Example Based Learning*
  - Get words $W_f \subset W$. $freq(w, D_{rel})/|D_{rel}| > freq(w, U)/|U|$
  - Remove from U all documents containing any word in W_f



Training the Expert Classifier

Exploit the partial-supervision

Liu et al. [2003]

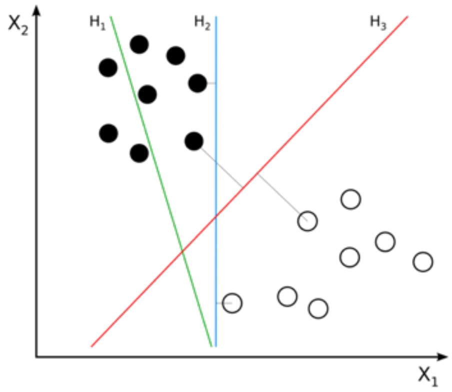
Expert Classifier

Builds on the result of the first step to train a much more sophisticated and precise classifier.

- **1-shot approach**
  - Use D_{rel} and RN and train a classifier (SVM or EM)
- **Iterative approach**
  - Use D_{rel} and RN and train a classifier C_i
  - Use C_i and extract new negative documents Q
  - Add Q to RN, train a new classifier C_{i+1}
  - Continue until no more negative documents are retrieved

[*Optionally*] evaluate the last trained classifier over D_{rel} and discard it if it performs poorly

(**Danger:***Concept Drift*)



Methods perform **poorly** when **the initial set of documents is very small**

The Rocchio approach + EM is best for this case

Advanced models with TF-IDF or Topic models
Zhu et al. [2013] - Zhu and Wu [2014]

Beware of Class Imbalance!
SMOTE: Synthetic Minority Over-sampling Technique

Document Segmentation

Intention-based relatedness

Papadimitriou et al. [2017]

Model documents as Composite Objects
Do not perform matching across the posts as a whole but across ***fragments*** of them that are ***written for the same intention***

Intuition:
Different parts of the document
Have different Purposes:

- Provide background information
- Describe Problem
- Ask question…

I have an HP system with a RAID 0 controller and 4 disks in form of a JBOD. I would like to install Hadoop with a replication 4 HDFS and only 320GB of disk space used from every disc. **Do you know whether it would perform ok or whether the partial use of the disk would degrade performance.** Friends have downloaded the Cloudera distribution but it didn't work. It stopped since the web site was suggesting to have 1TB disks. I am asking because I do not want to install Linux and then realize that my **hardware configuration is not the right one.**

Doc A

Extra RAID disk drives seem to be the solution to my problem but does adding RAID drives requires a **reformat and rebuild of the system to improve performance?**

Doc C

My boss gave me yesterday an HP Pavilion computer with Intel Matrix Storage System, a 320GB drive and Linux pre-installed. **I am thinking to add an extra disk drive using a RAID 0 or 1. Can I do it without having to rebuild the entire system?** I have already looked at the HP official web site for how to use a JBOD. But I have not found anything related to it.

Doc B

My HP Pavilion stops working after 15 min of activity. I called our technical department but no luck. Despite the many calls, I did not manage to find **a person with adequate knowledge to find out what is wrong.** All they said is bring it to up and we will see, which frustrated me. At the end I had the brilliant idea to move it to a cooler place and voila. No more p

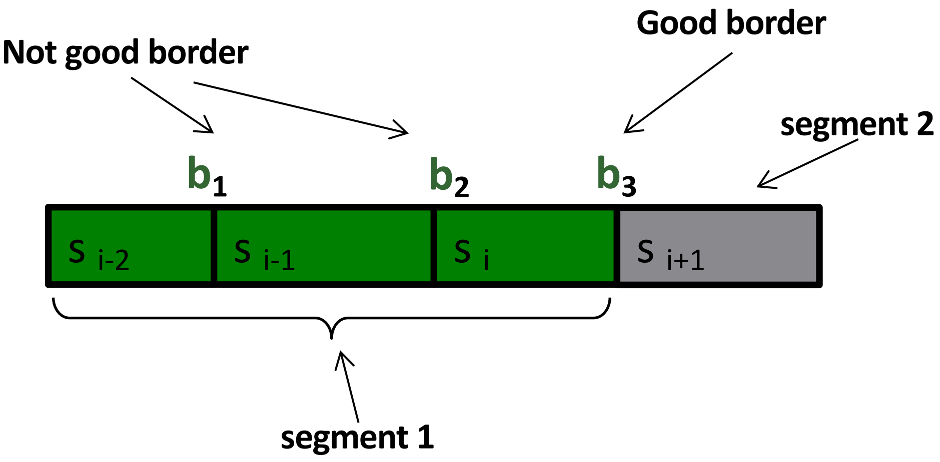
Doc D

Segmentation

Boundaries

Papadimitriou et al. [2017]

Use text characteristics and identify points in which a significant variation of these characteristics occurs, and place a segmentation border there.



Communication means & Text Features

Tense(CM_{tense})	present	past	future
Subject (CM_{subj})	I/we	you	it/they/(s)he
Style (CM_{qneg})	interrog.	negative	affirmative
Status (CM_{pasact})	passive	active	
Part of Speech(CM_{pos})	verb	noun	adj./adverb

0 I have an HP system with a RAID 0 controller and 4 disks in form of a JBOD. 75 I would like to install Hadoop with a replication 4 HDFS and only 320GB of disk space used from every disc. 182 Do you know whether 201 it would perform ok or whether the partial use of the disk 259 would degrade performance. 285 Friends have downloaded the Cloudera distribution but 338 it didn't work. 355 It stopped since 371 the web site was suggesting to have 1TB disks. 418 I am asking because 436 I do not want to install Linux and then realize that 488 my hardware configuration is not the right one. 535

Bottom-up approach

1. Start with single words as segments
2. Compute a **Diversity Index** in each segment
3. Merge segments with low diversity

Intention Clustering & Matching

Papadimitriou et al. [2017]

Matching among segments with the same intention

Clusters are based on intentions

Given a document d_q,

1. the system will **segment** d_q ,
2. identify for each segment the **segments in the same cluster**
3. **aggregate the similarity** of those segments into a score for each document.

C3

Friends have downloaded the Cloudera distribution but it didn't work. It stopped since the web site was suggesting to have 1TB disks.

I am thinking to add an extra disk drive using a RAID 0 or 1. Can I do it without having to rebuild the entire system?

Do you know whetherit would perform ok or whether the

Despite the many calls, I did not manage to find a person with adequate knowledge to find out what is wrong.

Explore based on related topics linked to common goals

C1

I am asking because I do not want to install Linux and then

ight one.

I have already looked at the HP official web site for how to use a JBOD. But I have not found anything related to it.

I have an HP system with a RAID 0 controller and 4 disks in

Extra RAID disk drives seem to be the solution to my problem but does adding RAID drives requires a reformat and rebuild of the system to improve performance?

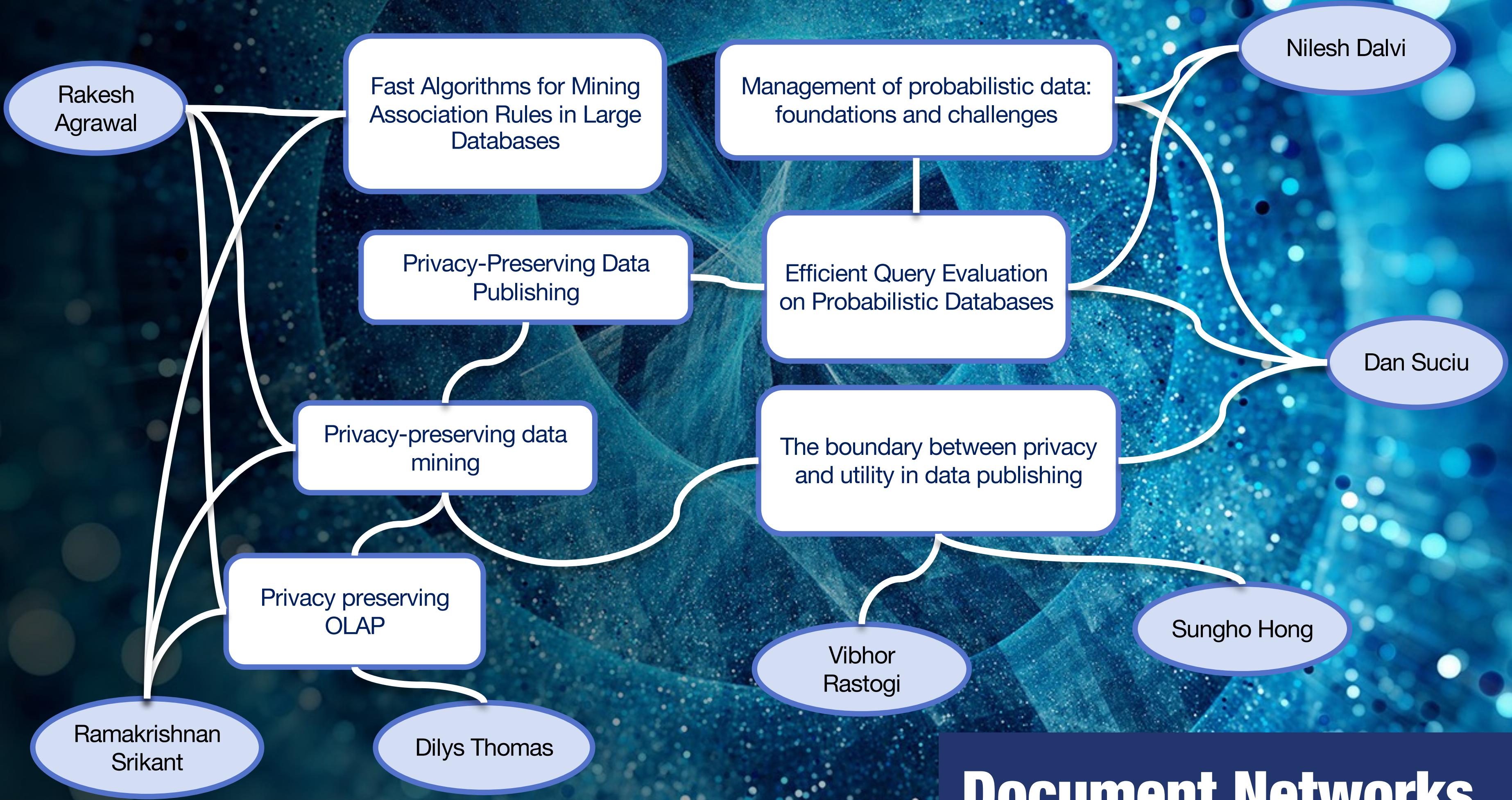
C2

All they said is bring it to up and we will see, which frustrated me. At the end I had the brilliant idea to move it to a cooler place and voila. No more problems.

My HP Pavilion stops working after 15 min of activity. I called our technical department but no luck.

Linux pre-installed.

Document Networks



https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

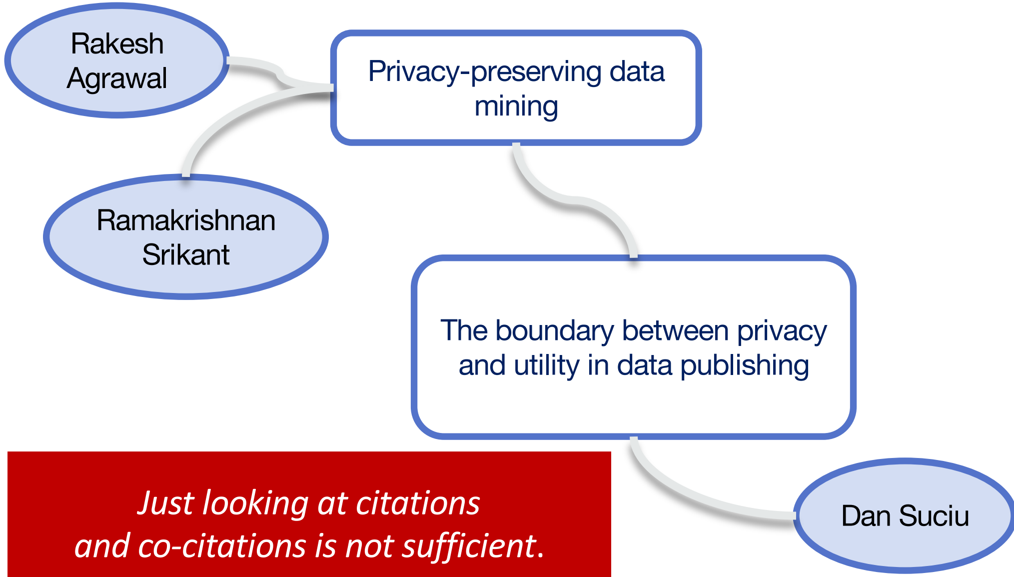
Influence in Citation Networks

Document relevance based on influence

El-Arini and Guestrin [2011]
Jia and Saule [2017]

Citation Network

- Nodes are Authors and Papers
- Edges are Authorship and Citations
- Influence is based on connecting Paths



Just looking at citations and co-citations is not sufficient.

Advance Models

- El-Arini and Guestrin [2011] :
  - Condition influence on topics

    Iterate for each topic T: Select topic T, keep only papers relevant for T, compute connecting Paths.
  - Weight edges with Influence-Probability
- Jia and Saule [2017]
  - Enrich graph with Keywords & Venues

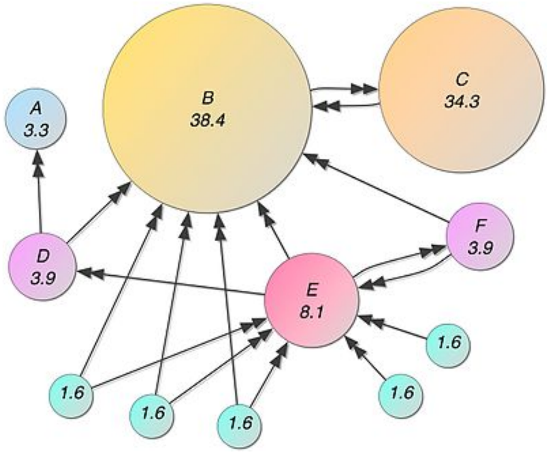
Start from a known document
Explore new related topics, authors, venues…

https://en.wikipedia.org/wiki/PageRank
SublinearAlgorithms for Personalized PageRank, with Applications – Ashish Goel et al

Traverse (Document) Networks

How to navigate links and connections

El-Arini and Guestrin [2011]
Jia and Saule [2017]



Global Page Rank

Starting from a random node, traversing randomly, **random restart point** anywhere in the graph

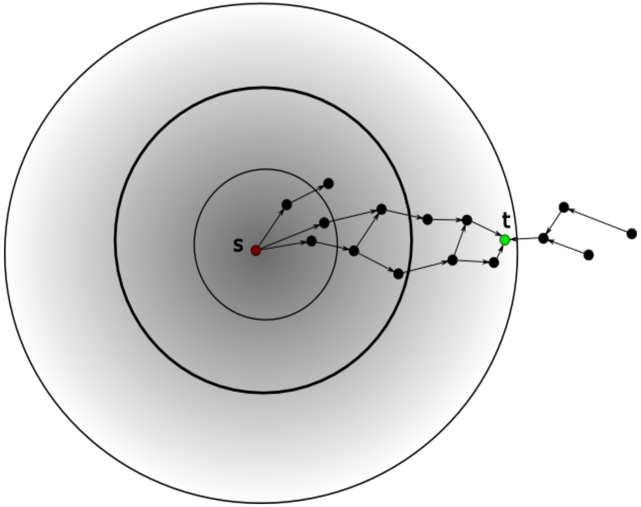
Personalized Page Rank

- Start from seed nodes, i.e. the documents D_{rel}
- Navigate towards locally connected nodes

Example based Exploration implies locality

CHALLENGE:
Identify meaningful transition probabilities

E.g., El-Arini and Guestrin [2011]



Personalized Page Rank

Starting from a **limited set of nodes**, traversing randomly, restart point is one in **the initial set**.
Bound not to travel too far

Serendipitous Search

Bordino et al. [2013]

Enhance document links with Entities and Query-logs

America

Peru

Machu Picchu

Connected entities

Query Logs

Francisco Pizarro
Rafting excursion
Amazon river
...

Serenditpity
Related topics potentially come to mind *after* consulting the page.

Input: Query/Document
Output: Queries


Machu Picchu

From Wikipedia, the free encyclopedia

Machu Picchu (Spanish pronunciation: [ˈmatʃu ˈpiktʃu], Quechua: *Machu Picchu* [ˈmɑtʃu ˈpixtʃu], "Old Peak") is a pre-Columbian 15th-century Inca site located 2,430 metres (7,970 ft) above sea level.[1][2] Machu Picchu is located in the Cusco Region of Peru, South America. It is situated on a mountain ridge above the Urubamba Valley in Peru, which is 80 kilometres (50 mi) northwest of Cusco and through which the Urubamba River flows. Most

UNESCO World Heritage Site

Historic Sanctuary of Machu Picchu

Name as inscribed on the World Heritage List

Document

Serendipitous Search

rafting excursion down the urubamba river
el dorado temple of sun
indios quechuas
map of peru
sapa inca

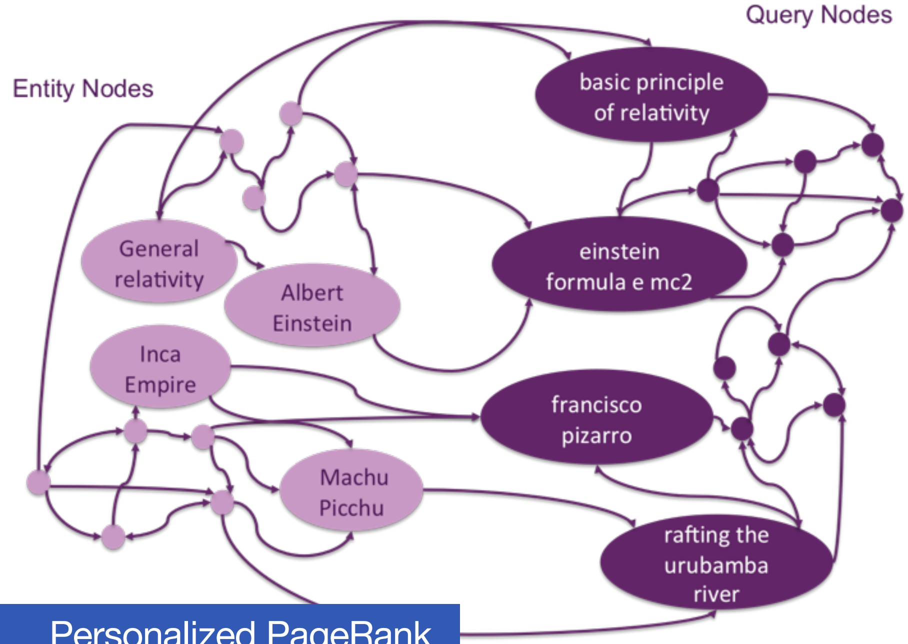
Searches related to Document content

Exploit "lateral connections" in User Search Behaviors

Entity Query Graph

Bordino et al. [2013]

Entity-Query graph from queries to entities and back



Personalized PageRank to score suggested queries

EQGraph Weighted Edges

Queries in the same session

1. query to query:

$$w_Q(q_i \to q_j) = w_{QFG}(q_i \to q_j)$$

Frequency-based approach

2. entity to query

$$w_{EQ}(e \to q) = \frac{f(q)}{\sum_{q_i | e \in X_E(q_i)} f(q_i)}$$

The more queries entities share the higher the probability

3. entity to entity

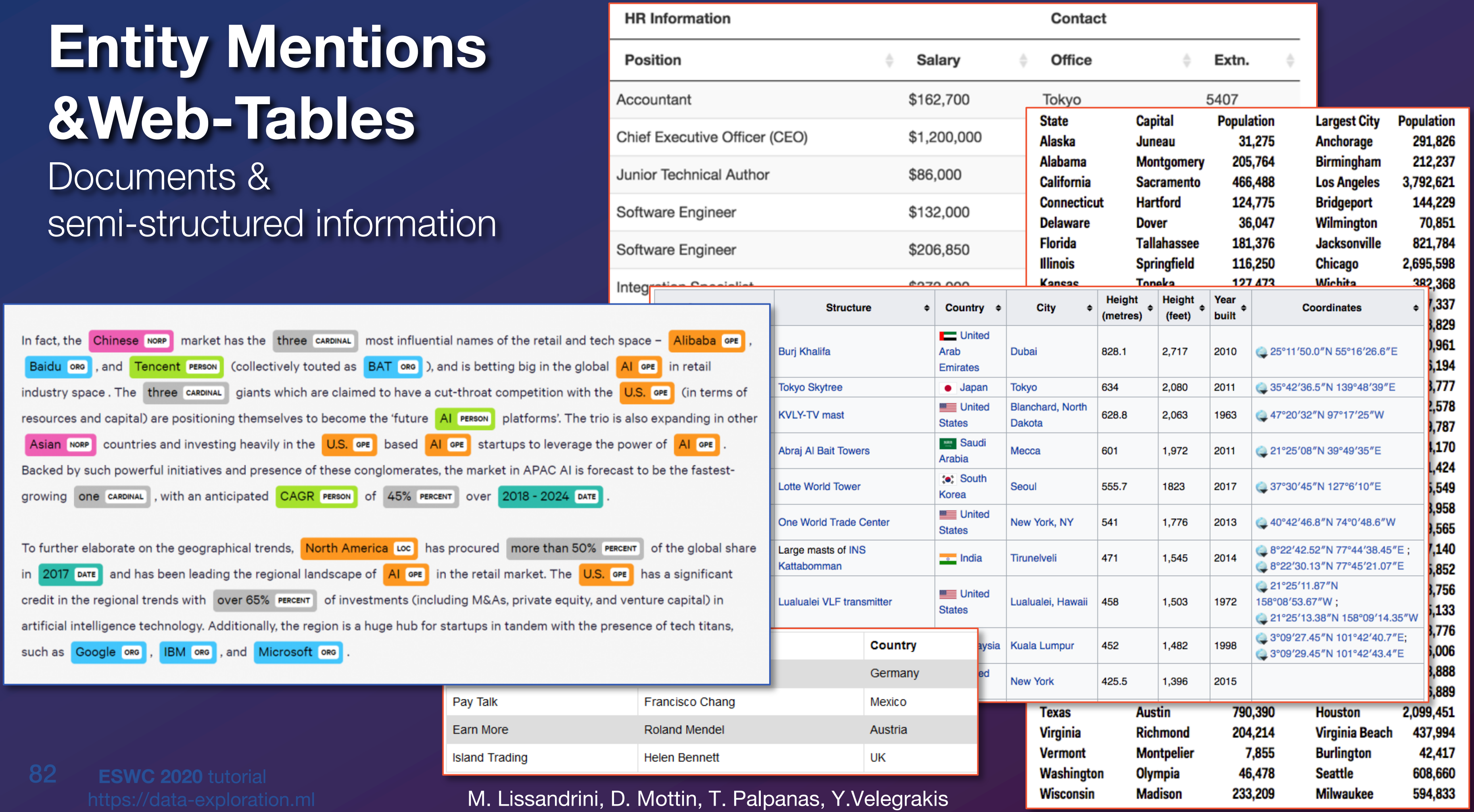
$$w_E(e_u \to e_v) = 1 - \prod_{i=1,\dots,r} (1 - p_{q_{i_s} \to q_{i_t}}(e_u \to e_v))$$

Based on query to query edges

SEARCHING FOR	Documents		Semi-Structured information	
BY **LOOKING** AT	Words	Meta-Data	Words	Web-Tables
APPLYING	Text Classifier [Liu et al. '03 Zhang and Lee'09, Zhu et al.'13] Topic Models [Zhu and Wu. '14] Segmentation [Papadimitriou et al. '17]	Citation Graph Navigation [El-Arini et al.'11 Jia and Saule'17] Entity Linking [Bordino et al. '13]	Regular Expressions [Agichtein et al.'00] Annotations [Hanafi et al.'17] Entity Extraction [Ritter et al.'15]	Entity Mentions [Wang et al.'15] Schema Matching [Yakout et al.'18]
PRODUCES	Documents/Citations/Queries recommendations		Relation Extraction Document Matching	Entity Augmentation Concept Expansion

Entity Mentions &Web-Tables

Documents &
semi-structured information



M. Lissandrini, D. Mottin, T. Palpanas, Y.Velegrakis

Entity-relation tuples

Brin [1998]
Agichtein and Gravano [2000]

Example-based extraction of Entity mentions and Relations

Search for Information WITHIN Documents
Explore new Entities
and new ways to express relations

Works bests with Binary relation
Can work with multiple mentions:
`Bob born in U.S.A. in 1978`

1. Example

`⟨ Google ; Menlo Park ⟩`

2. Match

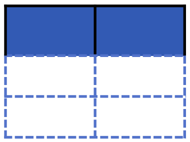
`Google founded in Menlo Park…`

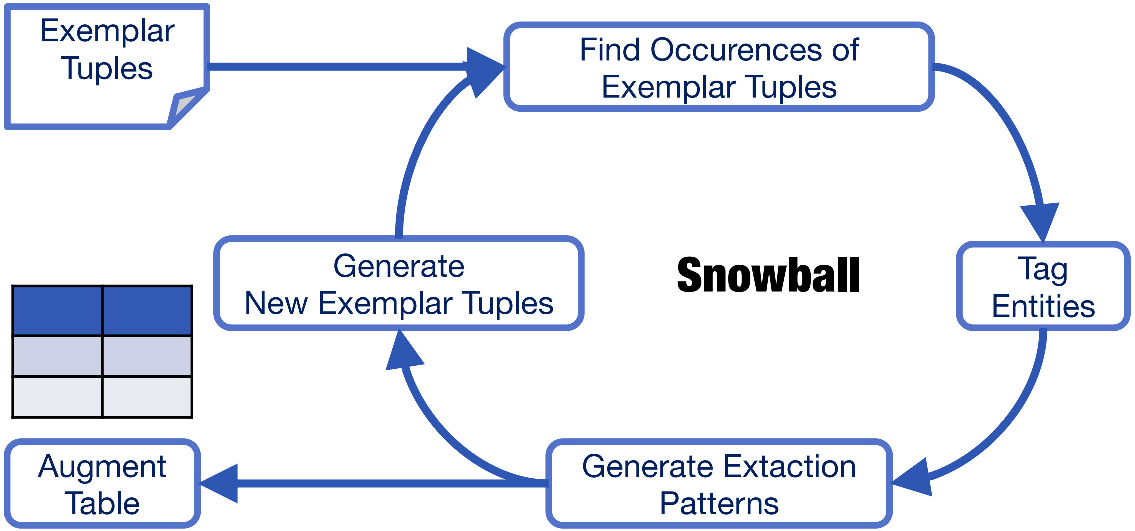
3. Extract Pattern

`… [X] founded in [Y] …`

4. Extract New Mentions & Patterns

`Apple founded in Coupertino …`
`Apple headquarters in Coupertino`





Goal: Enrich a list of Entity-relationships data

Entity-relation tuples

Example-based extraction of Entity mentions and Relations

Brin [1998]
Agichtein and Gravano [2000]

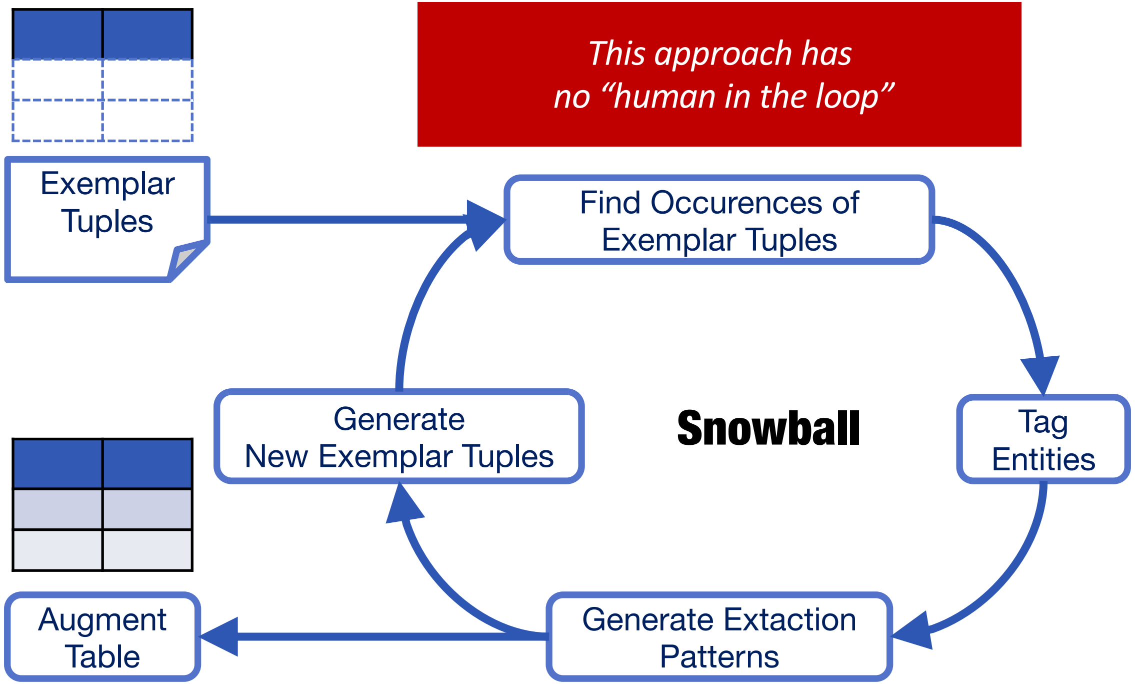
How to validate the new rules extracted automatically?

1. **Compare extracted rules with known tuples:** confidence of R is based on how many known tuples extracts

2. **Compare extracted tuples with known rules:** confidence of T is based on how many known rules also extract T

New extracted Rules and Tuples should be supported by known information

This approach has no "human in the loop"



IN MY DEFENSE
I WAS LEFT UNSUPERVISED

Entity-extraction by Example

Learn extraction rules from example

Hanafi et al., [2017]

Allow to match from text
both **Positive** and **Negative** examples

Goal: Supervised Extraction

definition) increased 9.6 percent, the number of murders increased 6.2 percent, aggravated assaults increased 2.3 percent, the number of rapes (revised definition) rose 1.1 percent, and robbery violations were up 0.3 percent.
Violent crime increased in all but two city groupings. In cities with populations from 50,000 to 99,999 inhabitants, violent crime was down 0.3 percent, and in cities with 500,000 to 999,999 in population, violent crime decreased 0.1 percent. The largest increase in violent crime, 5.3 percent, was noted in cities with 250,000

SEER

Output: Extraction rules

P: Percentage = 1.0 = 1.0

D: {5, 6} = 0.4 D: {percent, %} = 0.4 = 0.4

R: [0-9]+ = 0.2 D: {percent, %} = 0.4 = 0.3

Matching Rules

From string tokens to “semantics”

Hanafi et al., [2017]

Intuition: Exploit a vocabulary of simple specialized patters with known semantics

Example: 5 percent up in Dubai

5

P: Number

P: Integer

L: ‘5’

R: [0-9]+

percent

L: ‘percent’

R: [A-Za-z]+

T: 0-1

…

Each rule has a “class” and a preference score

Token gap $\prec$ Regex $\prec$ Literal Dictionary $\prec$ Pre-builts

Dubai : T: 0-1 $\prec$ L: ‘Dubai’ $\prec$ P: City

0 1

Each token may have different candidate “matching rules”

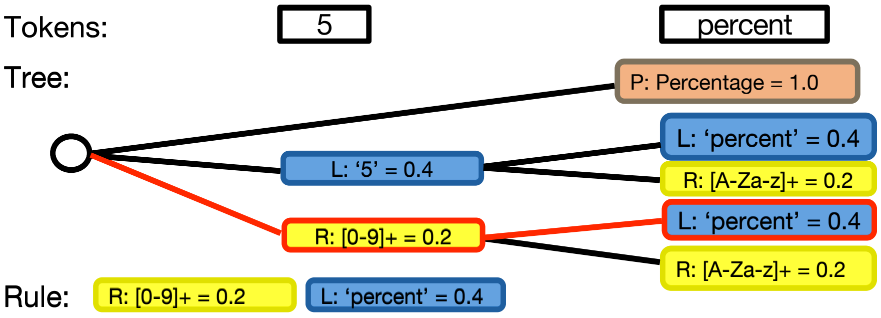
M. Lissandrini, D. Mottin, T. Palpanas, Y. Velegrakis

Merging Rules

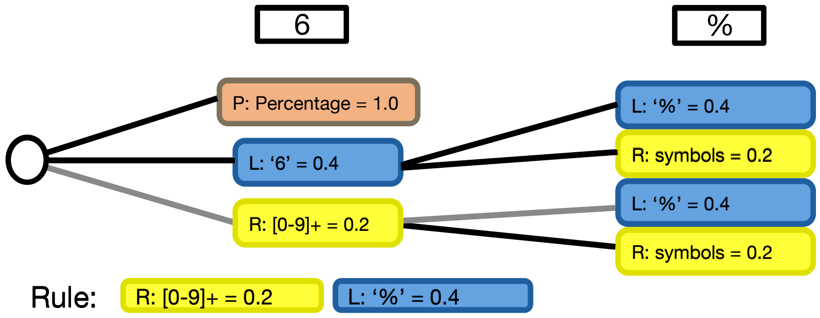
Hanafi et al., [2017]

Reconcile multiple interpretations

Example: 5 percent

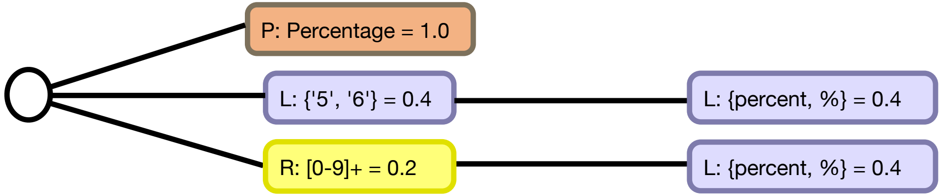


Example: 6 %



Intersection: [5 percent, 6%]

Consider also Negative Examples to prune candidates



https://vimeo.com/208729128

Web Tables

Semi-structured data on the web

https://en.wikipedia.org/wiki/Denmark#Regions

Regions

The governing bodies of the regions are the regional councils, each with forty-one councillors elected for four-year terms. The councils headed by regional district chairmen (*regionsrådsformanden*), who are elected by the council.[79] The areas of responsibility for the re councils are the national health service, social services and regional development.[79][80] Unlike the counties they replaced, the region allowed to levy taxes and the health service is partly financed by a national health care contribution until 2018 (*sundhedsbidrag*), partl from both government and municipalities.[18] From 1 January 2019 this contribution will be abolished, as it is bein tax instead.

The area and populations of the regions vary widely; for example, the Capital Region, which encompasses the C with the exception of the subtracted province East Zealand but includes the Baltic Sea island of Bornholm, has a than that of North Denmark Region, which covers the more sparsely populated area of northern Jutland. Under th densely populated municipalities, such as Copenhagen Municipality and Frederiksberg, had been given a status making them first-level administrative divisions. These *sui generis* municipalities were incorporated into the new r reforms.

Danish name	English name	Admin. centre	Largest city (populous)	Population (January 2017)	Total area (km²)
Hovedstaden	Capital Region of Denmark	Hillerød	Copenhagen	1,807,404	2,568.29
Midtjylland	Central Denmark Region	Viborg	Aarhus	1,304,253	13,095.80
Nordjylland	North Denmark Region	Aalborg	Aalborg	587,335	7,907.09
Sjælland	Region Zealand	Sorø	Roskilde	832,553	7,268.75
Syddanmark	Region of Southern Denmark	Vejle	Odense	1,217,224	12,132.21

Source: Regional and municipal key figures

Google

country capitals

All Images Maps Videos News More Settings Tools

About 2.460.000.000 results (0,77 seconds)

According to countries-ofthe-world.com View 40+ more

Albania Tirana Algeria Algiers Andorra Andorra la Vella Angola

List of world capitals

Country	Capital city
Albania	Tirana
Algeria	Algiers

Google

food calories

All Images Videos Maps News More Settings Tools

About 317.000.000 results (0,57 seconds)

Food Group	Carbohydrates (Grams)	Calories
Milk (higher % of simple carbohydrates; less nutrient dense)		
Chocolate milk (1 cup)	26	208
Low fat (2%) milk	12	121
Pudding (any flavor) (1/2 cup)	30	161

67 more rows

Carbohydrate and Calorie Content of Foods By Item | MomsTeam

Web Table Extraction, Retrieval, and Augmentation: A Survey.
Shuo Zhang and Krisztian Balog. 2020.
https://doi.org/10.1145/3372117

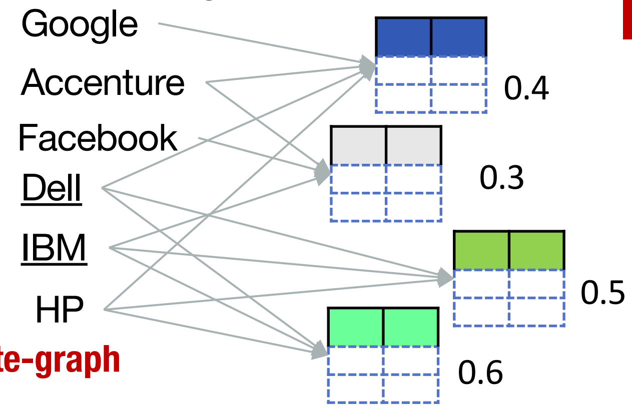
Entity List Expansion

Wang et al. [2015]

Augmentation: identify entities to complete the list

Goal: Given some seed entity mentions, retrieve more entities of the same type

1. Input: Incomplete list + Keyword query
2. Retrieve tables from pages based on the keyword query
3. Assign Score to tables based on relevance
4. Extract entity mentions from tables
5. Analyze Entity mention co-occurrence
6. Pick ”co-occurring” Entities



Bipartite-graph

Problem: *entities may appear together for different reasons*

Score Propagation

Problem: *Here PPR Causes concept drift*

Heuristic Propagation

Incomplete table

IT Company
Dell
IBM
Lenovo
….?

Augmented table

IT Company
Dell
IBM
Lenovo
Apple
Samsung
HP
Acer

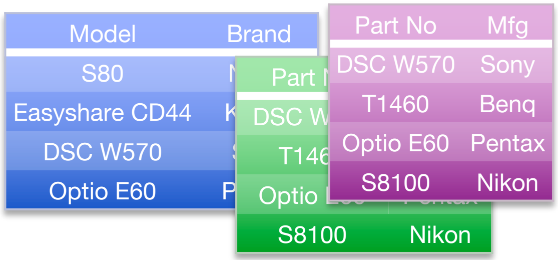
Web-Table Completion

Yakout et al. [2012]

Identify relevant content, retrieve missing information

Goal: Retrieve missing attribute values

Intuition: If there is a structure, we can match it!



Web tables

Model	Brand
S80	Benq
A10	
GX-1S	
T1460	

Incomplete table

→ **InfoGather** →

Model	Brand
S80	Benq
A10	**Innostream**
GX-1S	**Samsung**
T1460	**Benq**

Complete table

Problem: *entities may appear together for different reasons*

Extra Input: table header
target attribute name or example of completing attribute

Table Correlation Graph

Yakout et al. [2012]

Schema matching for web-page and web-tables
Binary-relations only

Goal: Retrieve missing attribute values

Determine Table Match

Direct Match between Q(K,A) and T(K,B)
K=entity names in a column
A,B = attribute column name (header)

Can use approximate matching and thesaurus

$$S_{DMA}(T) = \begin{cases} \frac{|T \cap_K Q|}{\min(|Q|, |T|)} & if\ Q.A \approx T.B \\ 0 & otherwise \end{cases}$$

***Problem:** considers only direct links between Q and T*

Query

Name	Developer
SQL Server	Microsoft
MySQL	
Teradata	
Firebird	

T1

Product	Vendor
MySQL	Oracle corp.
PostgreSQL	PostgreSQL Grp
MongoDB	MongoDB inc.
Berkley DB	Oracle corp.

List of Open Source **database software**

T2

Name	Max Row Size
MySQL	64Kb
Oracle	8Kb
Firebird	64Kb
Berkley DB	8kb

Information about **database** size limits

T3

Name	Developer
MySQL	Oracle
SQL Server	Microsoft
Office	Micrsoft
Photoshop	Adobe

Best selling **software** in 2010

T6

Vendor	Software
Oracle corp.	Oracle DB
IBM	DB2
Teradata	Teradata Corp.

Companies developing **database software**

T5

Vendor	Revenue
Oracle	11787M
IBM	4870M
Microsoft	4098M
Teradata	882M

Database software, 2011 revenue by vendor

T4

Name	Windows	Linux
Oracle	Yes	Yes
MySQL	Yes	Yes
SQL Server	Yes	No
PostgreSQL	Yes	Yes

OS support for top **database software**

Table Correlation Graph

Yakout et al. [2012]

Schema matching for web-page and web-tables
Binary-relations only

Goal: Retrieve missing attribute values

Determine Table Match

Holistic Match

1. Assign Direct Match Score from Query to Tables
2. Scores >0 are starting nodes

3. Use classifier to add weight to other table pairs

   Build Classifier using

   - Context similarity
   - Table-to-content similarity
   - URL similarity
   - Tuples Similarity

   the model predicts the match between two tables with a probability

4. Use starting node and execute PPR

5. Use PPR scores to rank matching tables

Overcomes problems due to poor matching with the query

Query

Name	Developer
SQL Server	Microsoft
MySQL	
Teradata	
Firebird	

T1

Product	Vendor
MySQL	Oracle corp.
PostgreSQL	PostgreSQL Grp
MongoDB	MongoDB inc.
Berkley DB	Oracle corp.

List of Open Source **database software**

T6

Vendor	Software
Oracle corp.	Oracle DB
IBM	DB2
Teradata	Teradata Corp.

Companies developing **database software**

T2

Name	Max Row Size
MySQL	64Kb
Oracle	8Kb
Firebird	64Kb
Berkley DB	8kb

Information about **database** size limits

T5

Vendor	Revenue
Oracle	11787M
IBM	4870M
Microsoft	4098M
Teradata	882M

Database software, 2011 revenue by vendor

T3

Name	Developer
MySQL	Oracle
SQL Server	Microsoft
Office	Micrsoft
Photoshop	Adobe

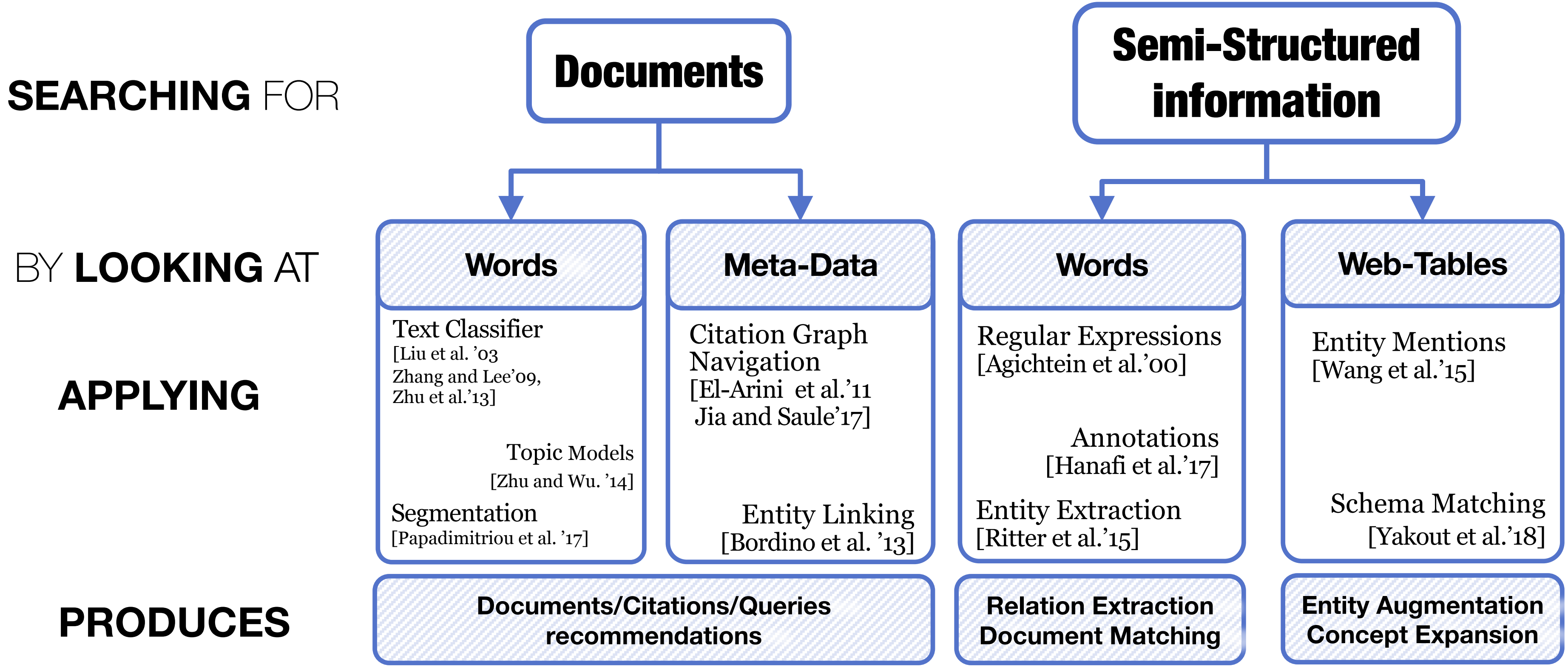
Best selling **software** in 2010

T4

Name	Windows	Linux
Oracle	Yes	Yes
MySQL	Yes	Yes
SQL Server	Yes	No
PostgreSQL	Yes	Yes

OS support for top **database software**

0.1 0.2 0.1 0.1 0.3 0.7 0.2 0.1 0.2 0.5


SEARCHING FOR
BY LOOKING AT
APPLYING
PRODUCES
Documents
Semi-Structured information
Words
Meta-Data
Words
Web-Tables
Text Classifier
[Liu et al. '03
Zhang and Lee'09,
Zhu et al.'13]
Topic Models
[Zhu and Wu. '14]
Segmentation
[Papadimitriou et al. '17]
Citation Graph Navigation
[El-Arini et al.'11
Jia and Saule'17]
Entity Linking
[Bordino et al. '13]
Regular Expressions
[Agichtein et al.'00]
Annotations
[Hanafi et al.'17]
Entity Extraction
[Ritter et al.'15]
Entity Mentions
[Wang et al.'15]
Schema Matching
[Yakout et al.'18]
Documents/Citations/Queries recommendations
Relation Extraction
Document Matching
Entity Augmentation
Concept Expansion

Where we are

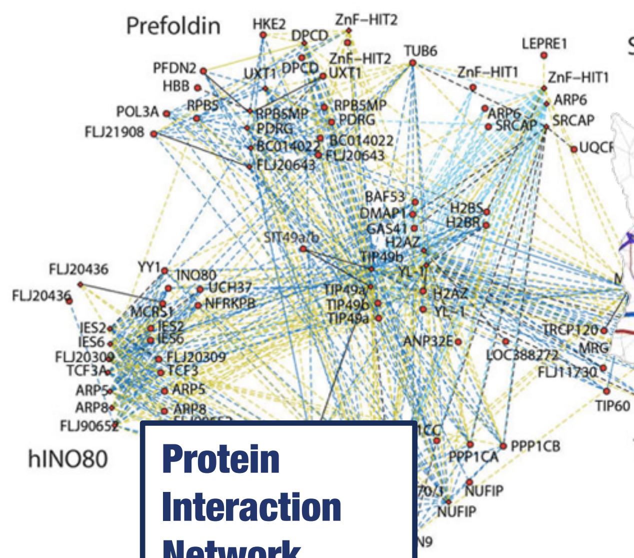
Relational databases

Textual data

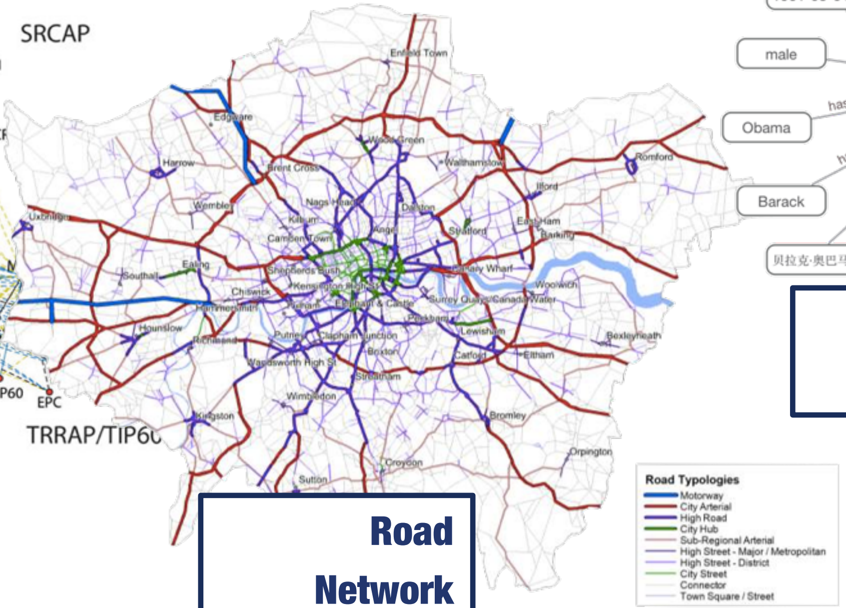
Graphs and networks

Machine learning

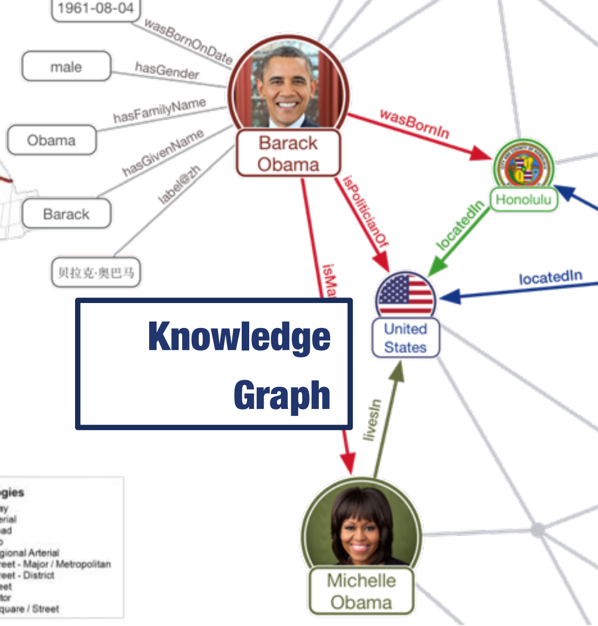
Challenges and Remarks



Protein Interaction Network



Road Network



Knowledge Graph

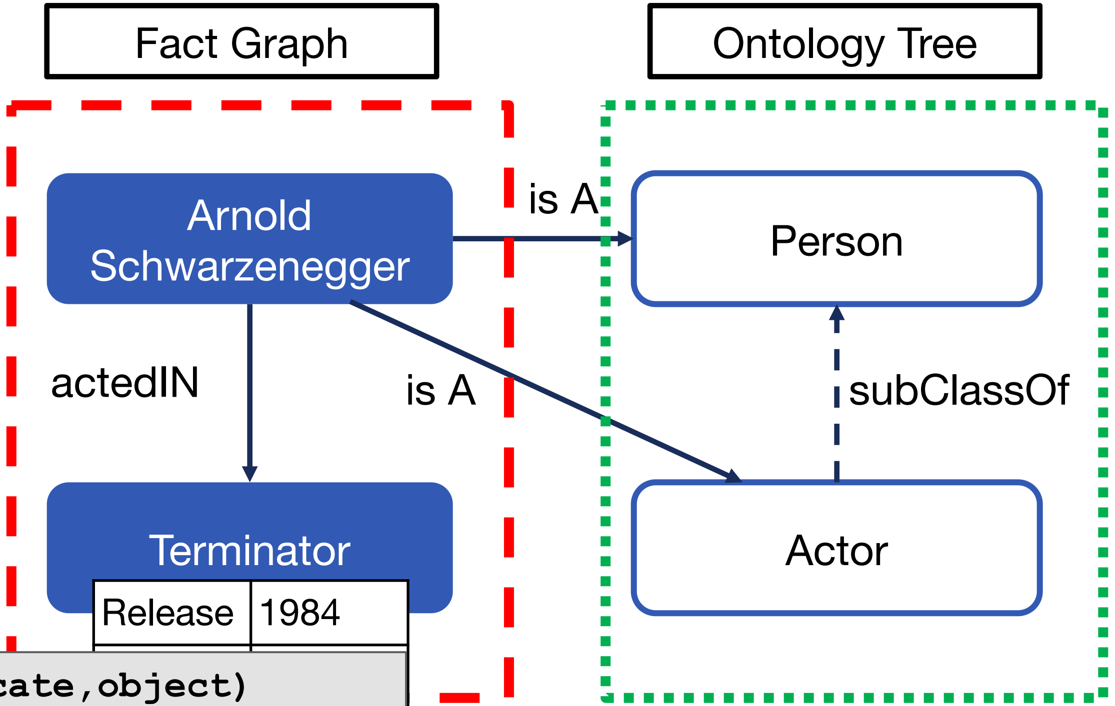
Graphs *are* Everywhere

Social Network

Graphs

Connected Data

A graph is a graph is a graph



Edge-labelled Multigraphs

G: $\langle V, E, L, \ell \rangle$

Attributes:

V/E:<key,value>

RDF **(subject,predicate,object)**

```
(Arnold_Schwarzenegger,isA,Person)
(Actor, subClassOf, Person)
(Arnold_Schwarzenegger, actedIn, Terminator)
```

The Structure of the Graph Is as important as the Data-values

Exemplar Queries

Example-driven graph search

Mottin et al. [2014,2016]

Input: Q_e, an example element of interest

Output: set of elements in the desired result set

Nodes/Entities
Edges/Facts
Structures

Exemplar Query Evaluation

- **evaluate** Q_e in a database D, finding a sample S
- find the set of elements A **similar** to S given a **similarity relation**
- [*OPTIONAL*] return only the subset A^R that are relevant

Usually requires an intermediate step:
User input (keywords) → Element in the graph

SIMILARITY for GRAPHS

- **Nodes**
  - Connectivity
  - Properties
- **Structures**
  - Queries
  - (Edge-)Labels

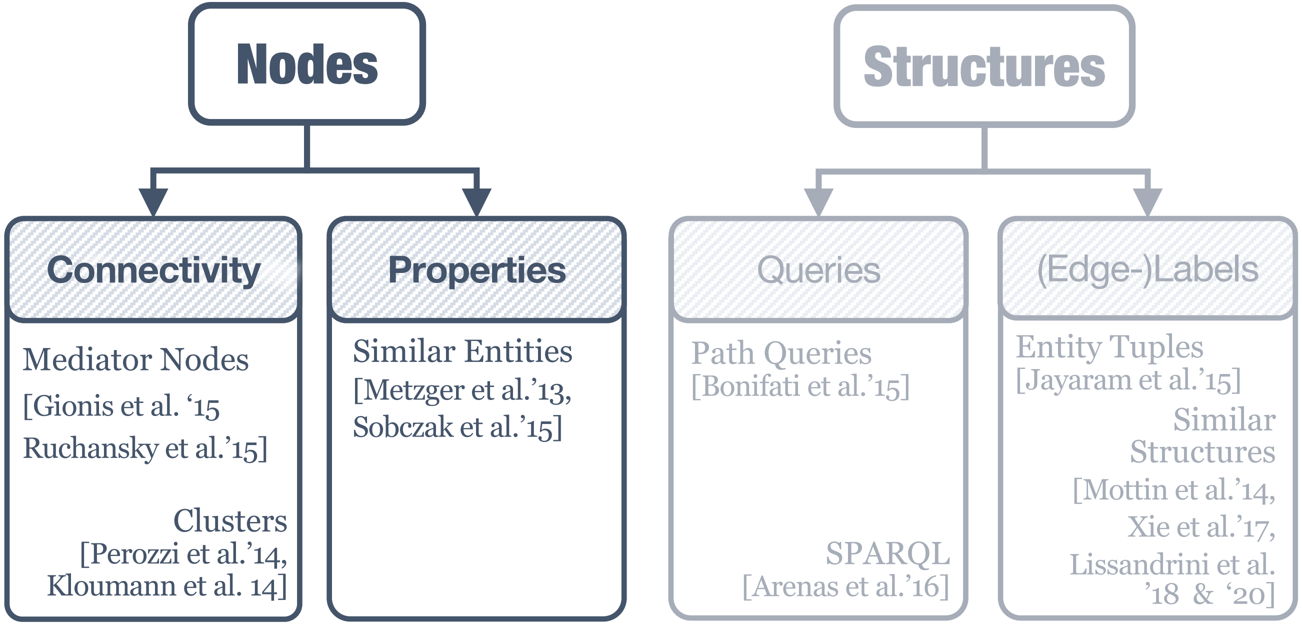
CHALLENGE: DISCOVER USER PREFERENCE

CHALLENGE: EFFICIENT SEARCH

SEARCHING FOR

BY **LOOKING** AT

PRODUCES

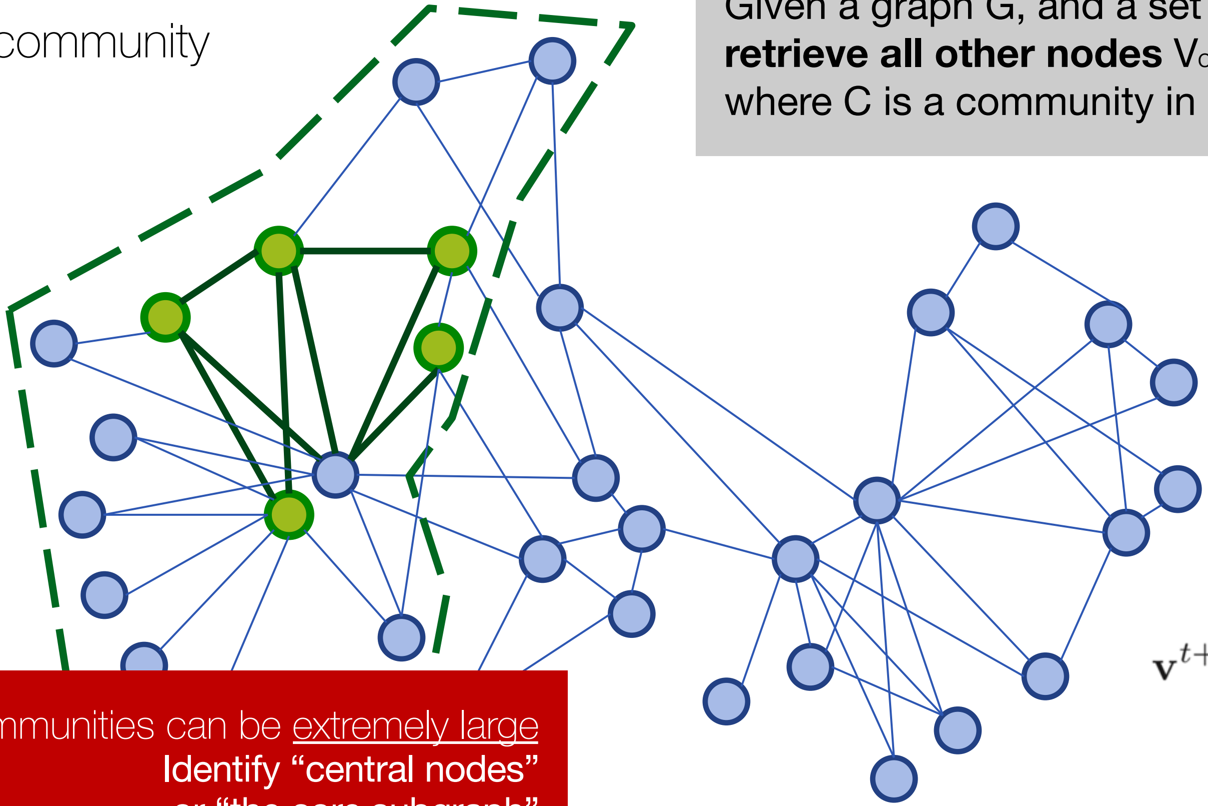


Seed Set Expansion

Nodes connected
by a community

Kloumann and Kleinberg [2014]

Given a graph G, and a set of **query nodes** $V_Q \subseteq V_G$,
retrieve all other nodes $V_C \subseteq V_G$,
where C is a community in G, and $V_Q \subseteq V_C$.



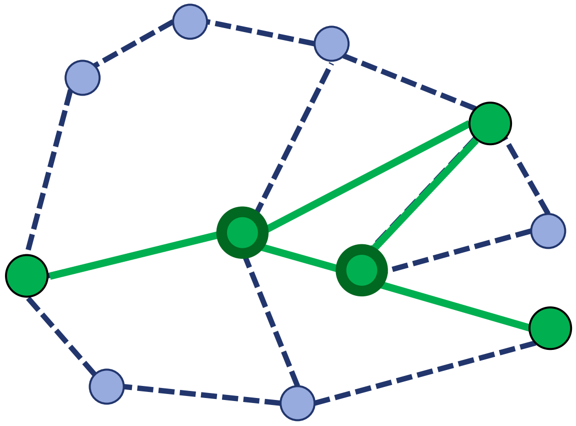
Solution: PPR

$$\mathbf{v}^{t+1} = (1-\alpha)\mathbf{M} \cdot \mathbf{v}^{t} + \alpha \mathbf{v}^{0}$$

Communities can be <u>extremely large</u>
Identify "central nodes"
or "the core subgraph"

The Minimum Wiener Connector Problem

Ruchansky et al. [2015]



Model: Unlabeled Undirected Graph

Query: A set of Nodes Q

Similarity: Shortest-Path distance

Output: A Set of Connector Nodes H

"explains" connections in Q

Connectors:
Nodes with HIGH closeness
to ALL the inputs

Similar to a Steiner-Tree but
overall pairwise distances are optimized

Case: Infected Patients
→ Culprit/Other Infected

Case: Target Audience
→ Influencers

The Minimum Wiener Connector Problem

Ruchansky et al. [2015]

Model: Unlabeled Undirected Graph

Query: A set of Nodes Q

Similarity: Shortest-Path distance

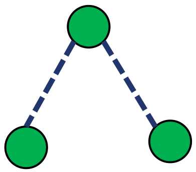
Output: A Set of Connector Nodes H

"explains" connections in Q

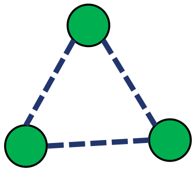
minimize the sum of pairwise shortest-path-distances between nodes in the connector H

Called: Wiener Index.

tradeoff between size and average distance



W=1+2+1 =4



W=1+1+1 = 3

Sometimes The Best Solution is NOT A Tree

NP-Hard

$$min \sum_{(u,v)\in H} d(u,v)$$

$\mathrm{d(u, v)}$ is the shortest-path distance

Approximate minimum Wiener Index Connector

Ruchansky et al. [2015]

CHOOSE $r \in Q$ & $\lambda \in \left[1,\ \log_{(1+\beta)} |V|\right]$

All Pairwise Distances

→ Distances from a root r

Measure distance in H (i.e., subgraph-induced)

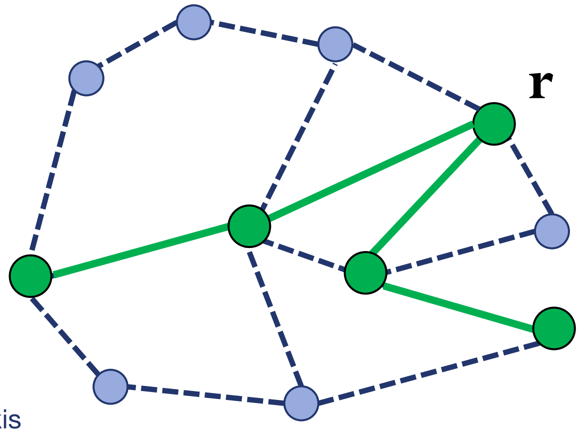
→ Precomputed distance in G

Edge Weights

$$w(u, v) = \lambda + \frac{max\{d_G(r,u), d_G(r,v)\}}{\lambda}$$

Approximated with Edge-Weighted SteinerTree

Enumerate Candidate Solutions for r ∈ Q & λ and keep best tree



Focused Clustering and Outlier Detection

Similarity based on attributes

Model: Unlabeled Undirected Graph <u>with Node Attributes</u>

Query: A set of Nodes Q

Similarity: <u>To Be Inferred</u>

based on Attribute Values & Connectivity

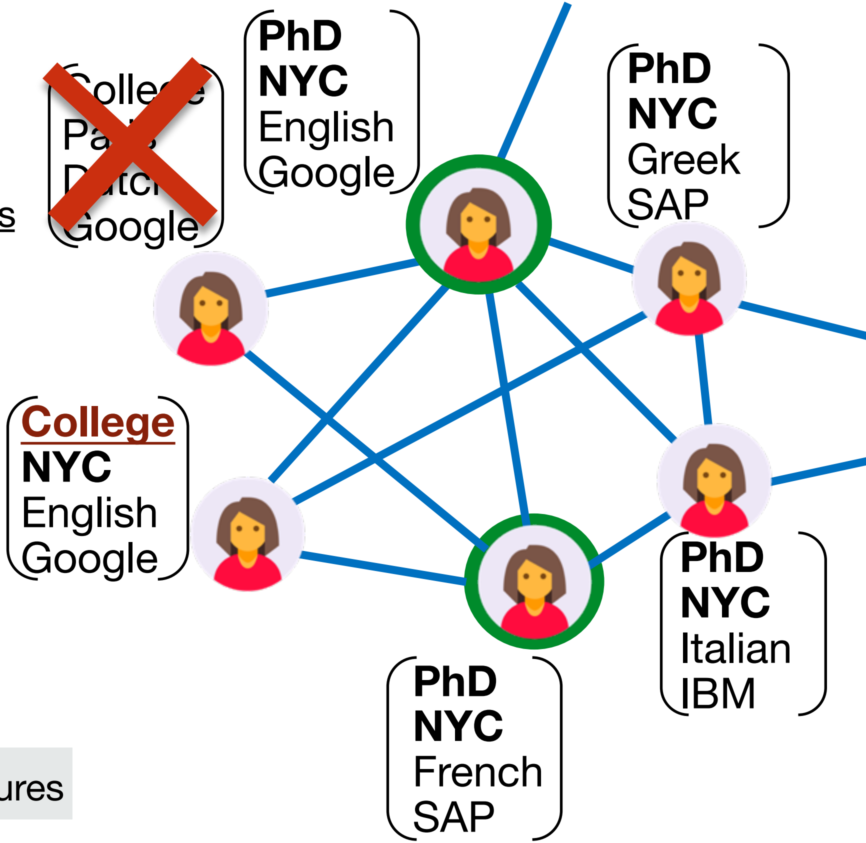
Output: <u>Clusters</u> of Nodes: Dense & Coherent

\+ <u>Outliers</u>

Case: Target Users ⟶ Community with same interests

Case: Products⟶ Co-purchased products with similar features

Perozzi et al. [2014]



M. Lissandrini, D. Mottin, T. Palpanas, Y. Velegrakis

Focused Clustering

Infer User Focus

TASK: Infer "FOCUS" , important attributes

attribute weights β

$$\begin{pmatrix}\textbf{PhD}\\ \textbf{NYC}\\ \text{English}\\ \text{Google}\end{pmatrix} \begin{pmatrix}\textbf{PhD}\\ \textbf{NYC}\\ \text{French}\\ \text{SAP}\end{pmatrix} \Rightarrow \begin{pmatrix}\mathbf{0.5}\\ \mathbf{0.5}\\ 0\\ 0\end{pmatrix}$$

1. Set of similar pairs, PS (from Q)

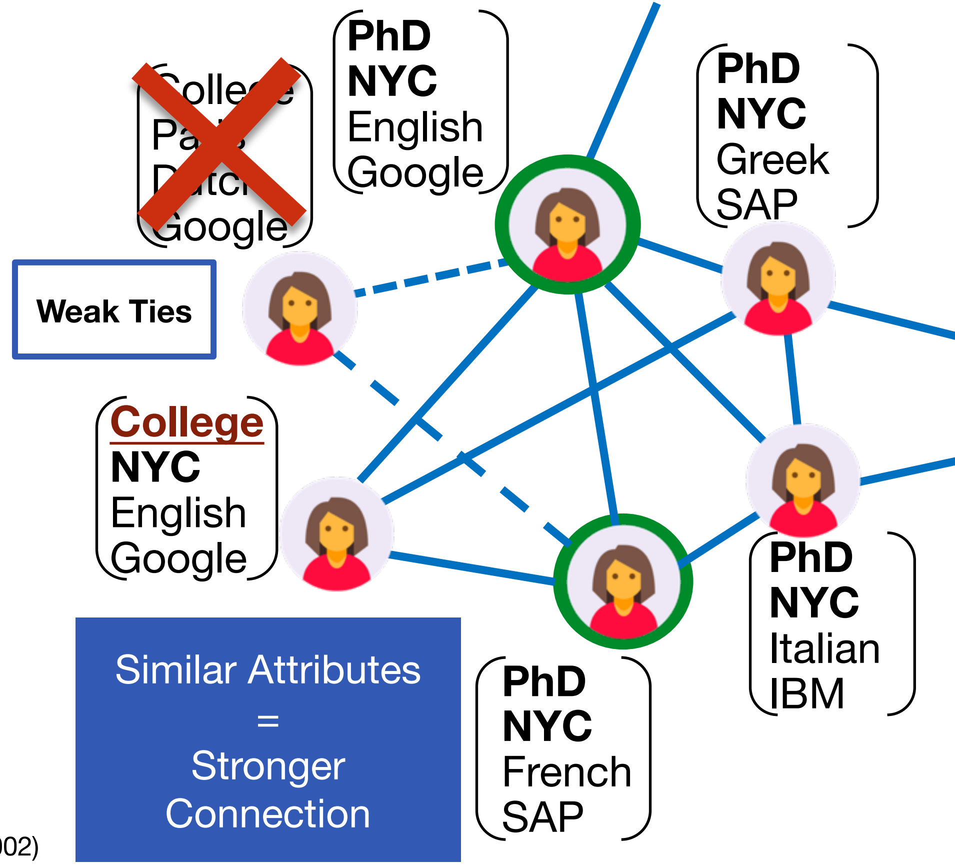
2. Set of dissimilar pairs, PD (random sample)

3. Learn a distance metric between PS and PD

$$\min_A \sum_{(u,v)\in P_S} (f_i - f_j)^T \mathbf{A} (f_i - f_j) - \gamma \log\left(\sum_{(u,v)\in P_D} \sqrt{(f_i - f_j)^T \mathbf{A} (f_i - f_j)}\right)$$

(Distance Metric Learning, inverse Mahalanobis distance: Xing, et al 2002)

Perozzi et al. [2014]



Focused Clustering

Perozzi et al. [2014]

Prune the Graph and keep dense communities

TASK: Extract Clusters on Focused Graph

attribute weights β → Edge Weight

1. Find Starting Set of Small Candidate Clusters

1.a Drop low-weight edges

1.b Extract Strongly Connected Component C_1, C_2, ...

2. Grow Clusters around Candidates

2.a Compute conductance of C: $\phi^{(w)}$ (C, G)

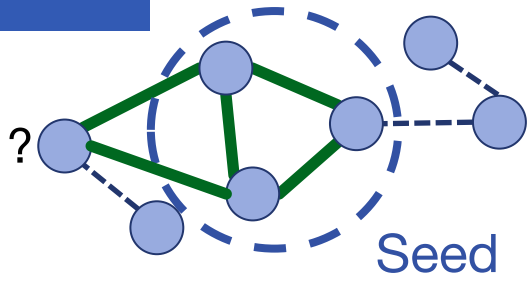
2.b Select node to add to C': best improvement to $\Delta\phi^{(w)}$ (C,C') (greedy)

2.c Prune Underperforming nodes

3. Detect Outliers: High unweighted conductance

w.r.t. low weighted conductance

LOCAL clusters



Weighted Conductance:
ratio between the weighted sum of edges crossing the boundaries of the cluster and the weighted sum of those residing within it.

Performant Strategy:
Start with local solution and expand around them to avoid complete scans of the graph

iQBEES: Entity Search by Example

Knowledge Graph Search

Metzger et al. [2013]
Sobczak et al. [2015]

Model: Knowledge Graph (Edge-labels)

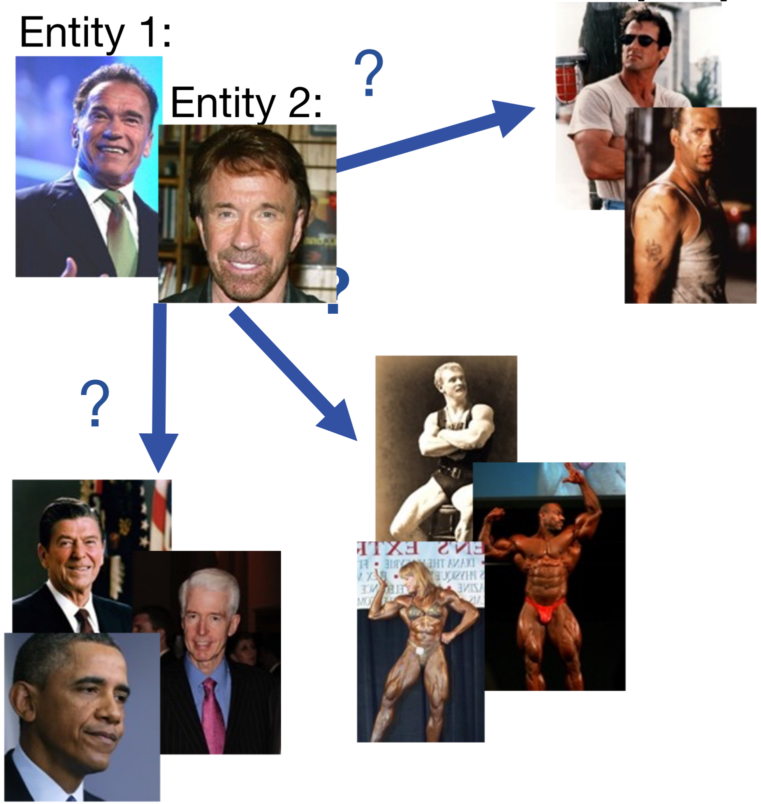
Query: A set of Entities Q

Similarity: Shared semantic properties

Output: A Set of <u>Similar Entities</u> (ranked)

Case: Products⟶ Products with similar aspects

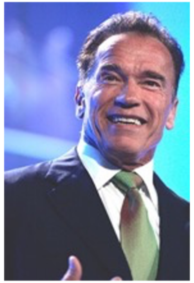
Case: KG Exploration⟶ Related Entities



Maximal Aspects

Selecting Features of Entity Similarity

Metzger et al. [2013]
Sobczak et al. [2015]



?x sport BodyBuilding

?x type AmericanActor

Is not maximal if
Adding any aspect
→ E(A)={Arnold}

?x type AmericanActor

?x governorOf California

Include
Typical Types

1. Prune generic aspects

?x hasHeight 1.88m

?x type Entity

Use most
Specific Type

2. Rank Set of aspects

?x type AmericanActor

?x actedIn TheExpendables

?x type ActionActor

REPEATABLE
Update Q

SIMILARITY for GRAPHS

Nodes
- ✓ Connectivity
- ✓ Properties

Structures
- Queries
- (Edge-)Labels

Queries can retrieve both Nodes and Structures

SEARCHING FOR

- **Nodes**
  - Connectivity
  - Properties
- **Structures**
  - **Queries**
  - **(Edge-)Labels**

BY **LOOKING** AT

- Connectivity
- Properties
- Queries
- (Edge-)Labels

PRODUCES

Connectivity	Properties	Queries	(Edge-)Labels
Mediator Nodes [Gionis et al. '15 Ruchansky et al.'15]	Similar Entities [Metzger et al.'13, Sobczak et al.'15]	Path Queries [Bonifati et al.'15]	Entity Tuples [Jayaram et al.'15]
Clusters [Perozzi et al.'14, Kloumann et al. 14]		SPARQL [Arenas et al.'16]	Similar Structures [Mottin et al.'14, Xie et al.'17, Lissandrini et al. '18 & '20]

Learning Path Queries on Graphs

Bonifati et al. [2015]

Queries from Examples

Model: Edge Labeled Graph

Query: 2 sets of Entities Q^{+} , Q^{-}

Positive, Negative

Negative Examples to disambiguate intention

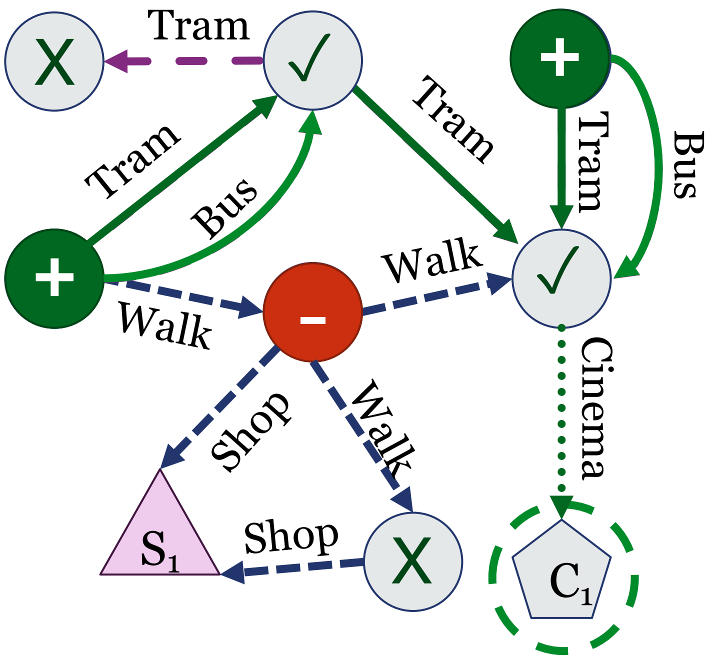
Similarity: Common Path Query (RegExp)

$$q := \epsilon \mid a (a \in \Sigma) \mid q_1 + q_2 \mid q_1 \cdot q_2 \mid q^*$$

(bus|tram)*+ Cinema

Output: Set of Nodes satisfying paths for Q^{+}

but not paths for Q^{-}



Case: Proteins⟶ Similar interactions/co-expression

Case: Tasks Initiator⟶ Similar Processes/Behaviours

MONADIC: only starting nodes

extensible to

BINARY/ N-ARY : path from X to Y

Learnability of Path Queries

When is possible and How

Bonifati et al. [2015]

Sometimes Positive & Negative Examples Cannot be reconciled!

Query: 2 sets of Entities Q^{+} , Q^{-}

Consistency:

1. Select Smallest Consistent Path

$$\forall v \in Q^{+}.\, paths_G(v) \not\subseteq paths_G(Q^{-})$$

2. Loops cause infinite paths? Fix Maximal Length K

   When to use Kleene star * ?

$$\mathrm{C} \mid (\mathrm{A.B.C}) \longrightarrow (\mathrm{A.B})^{*}.\mathrm{C}$$

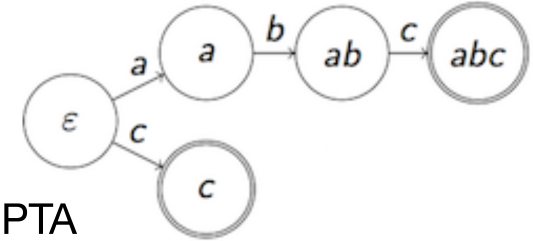
3. Generalize SCP
   a) Construct Prefix Tree Acceptor
   b) Generalize into DFA with Merge

Consistency Check: PSPACE-complete

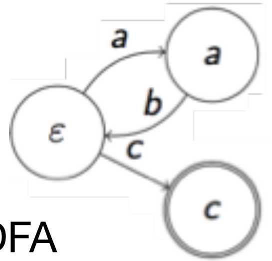
Enumerate Paths Up to Fixed distance

For paths of Length N

K = 2 x N +1



PTA



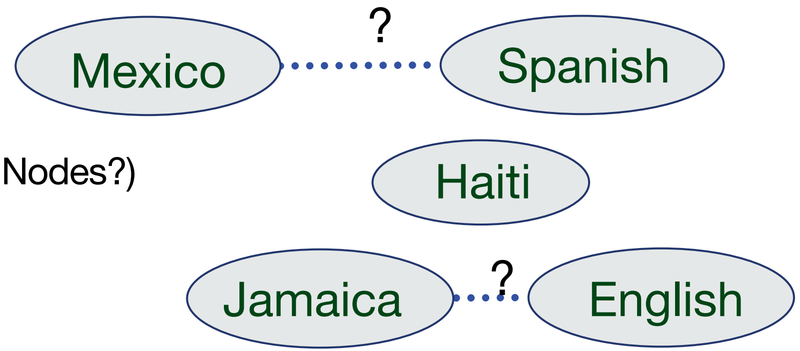
DFA

Can be INTERACTIVE! The system presents to the user nodes to label as Positive/Negative

Reverse engineering SPARQL queries

Arenas et al. [2016]

Knowledge Graph Search



Model: Knowledge Graph (Edge-labels)

Query: Set of Answers → Not Graphs but Tuples (of Nodes?)

Similarity: common AND/OPT/FILTER query

Output: a SPARQL query / query results

	?e1	?e2
M1	**Mexico**	**Spanish**
M2	**Haiti**	
M3	**Jamaica**	**English**

```
MATCH (?X, is_a, Country)
OPT (?X, has_language, ?Y)
```

Case: Open Data⟶ Query Unknown Schema

Case: Novice User ⟶ Avoid SPARQL

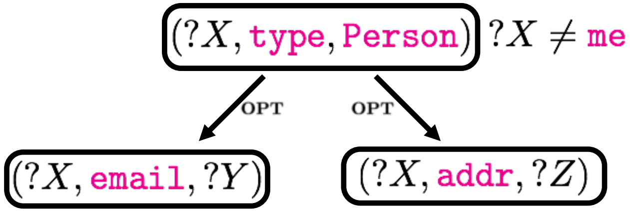
Reverse engineering SPARQL queries

Challenges and Complexity

Arenas et al. [2016]

Query: Set of Variable Mappings

	?X	?Y	?Z
M1	John		
M2	Mary	mary@email.eu	
M3	Lucy		Roses Street



Incomplete Mappings are treated as OPTIONAL
Typical of RDF queries

Enumerate all possible SPARQL queries satisfied by the mappings

INTRACTABLE

Σ_2^p−complete

Build tree-shaped SPARQL queries IMPLIED by the mappings

SEARCHING FOR

- Nodes
- Structures

BY **LOOKING** AT

- Connectivity (Nodes)
- Properties (Nodes)
- Queries (Structures)
- (Edge-)Labels (Structures)

PRODUCES

Connectivity

Mediator Nodes
[Gionis et al. '15
Ruchansky et al.'15]

Clusters
[Perozzi et al.'14,
Kloumann et al. 14]

Properties

Similar Entities
[Metzger et al.'13,
Sobczak et al.'15]

Queries

Path Queries
[Bonifati et al.'15]

SPARQL
[Arenas et al.'16]

(Edge-)Labels

Entity Tuples
[Jayaram et al.'15]

Similar Structures
[Mottin et al.'14,
Xie et al.'17,
Lissandrini et al. '18 & '20]

Graph Exemplar Queries

Search for Structures

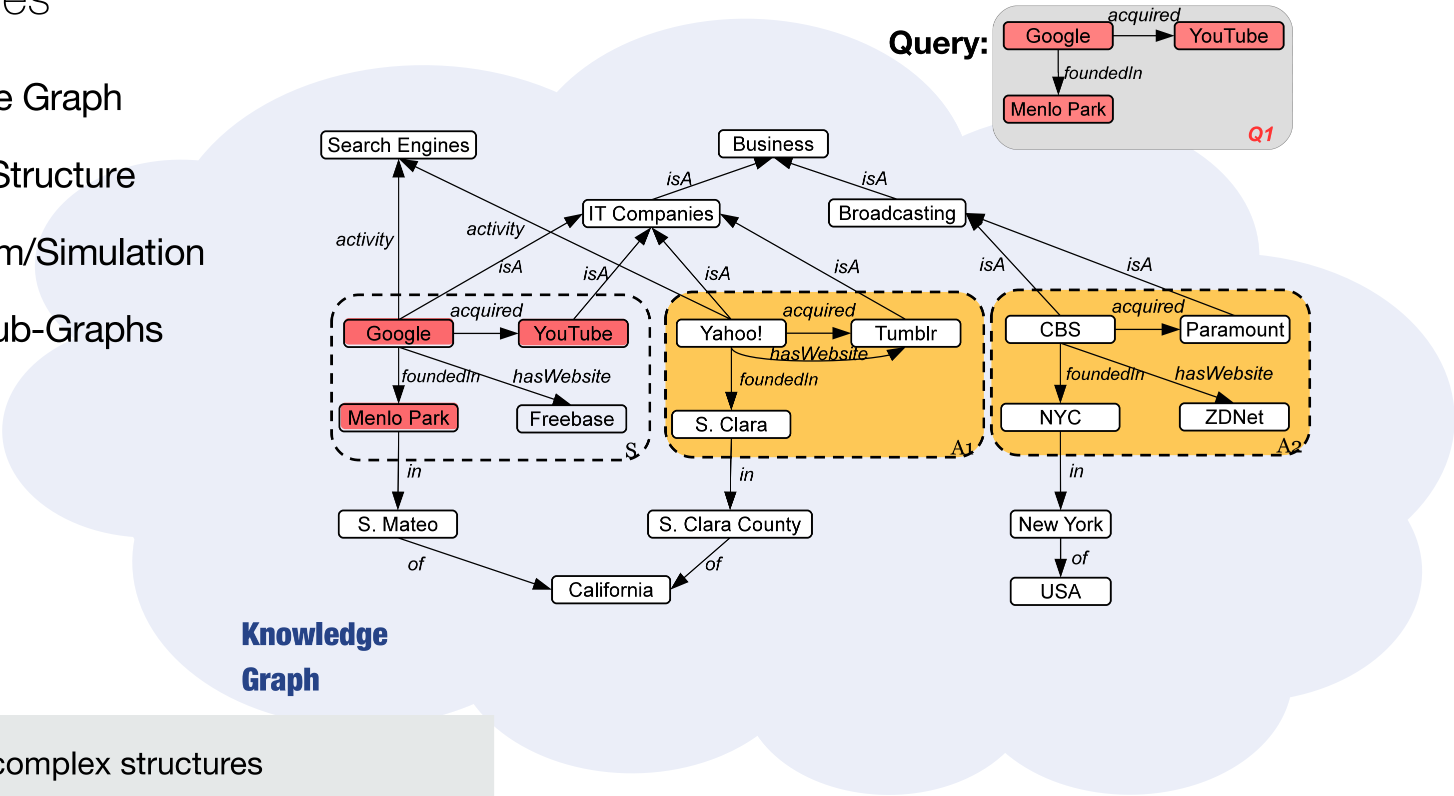
Mottin et al. [2016]

Model: Knowledge Graph

Query: Example Structure

Similarity: Isomorphism/Simulation

Output: A set of Sub-Graphs



Case: Rich Schema ⟶ Find complex structures

Graph Query by Example(GQBE)

Search for example Tuples

Jayaram et al. [2015]

Model: Knowledge Graph

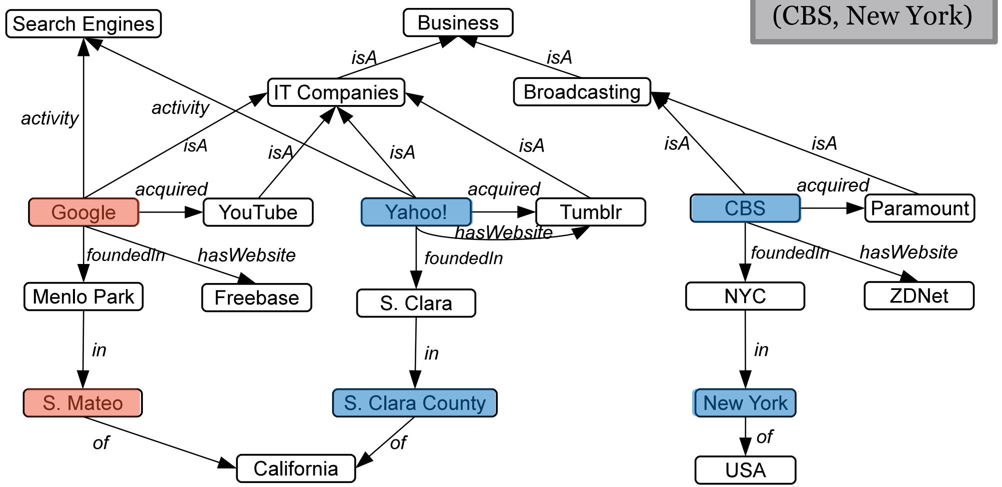
Query: Entity Tuples

Similarity: ~Isomorphism

Output: A set of Tuples

Q = (Google, S. Mateo)

Results =
(Yahoo, S. Clara)
(CBS, New York)



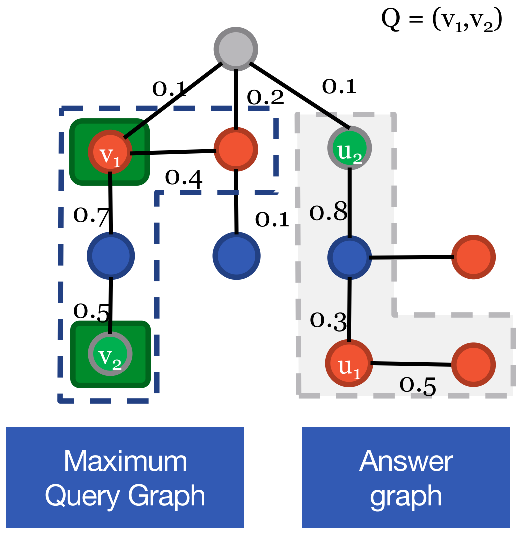
In GQBE Input is a set of (disconnected) entity mention tuples

Case: Known Entities+Uknown Connections ⟶ Find Complex Connections

GQBE: Maximum Query Graph

Jayaram et al. [2015]

Understand the connections implied by the tuples



1. Find the maximum query graph
   - Graph with M edges having the maximum weight
2. Answers subgraph-isomorphic to the query graph NP-hard
3. Return top-k

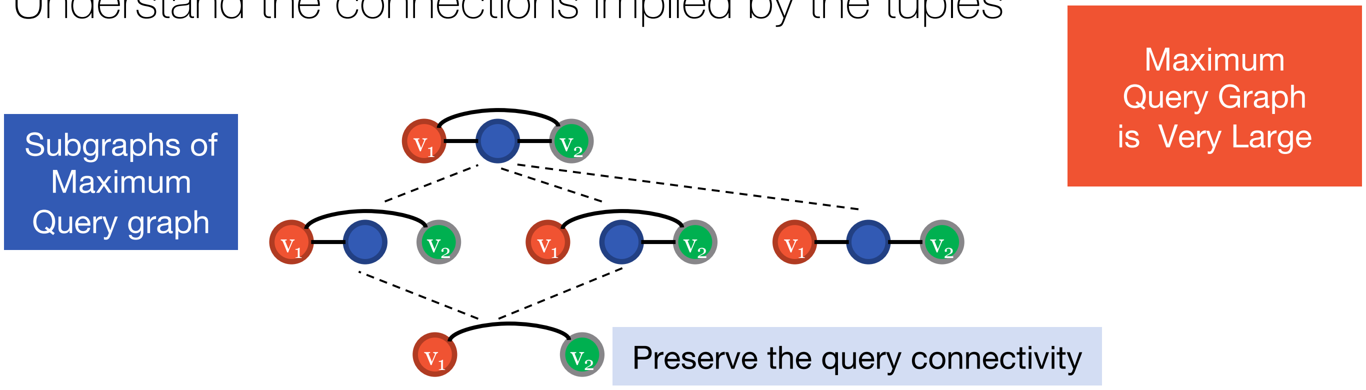
Answer score:
- Sum of query graph weights
- Similarity match between edges in the answer and the query (shared nodes take extra credit)

$$\text{match}(e, e') = \begin{cases} \frac{w(e)}{|E(u)|} & \text{if } u{=}f(u) \\ \frac{w(e)}{|E(v)|} & \text{if } v{=}f(v) \\ \frac{w(e)}{min(|E(u)|,|E(v)|)} & \text{if } u{=}f(u), v{=}f(v) \\ 0 & \text{otherwise} \end{cases}$$

GQBE: Multiple Query Tuples

Understand the connections implied by the tuples

Jayaram et al. [2015]

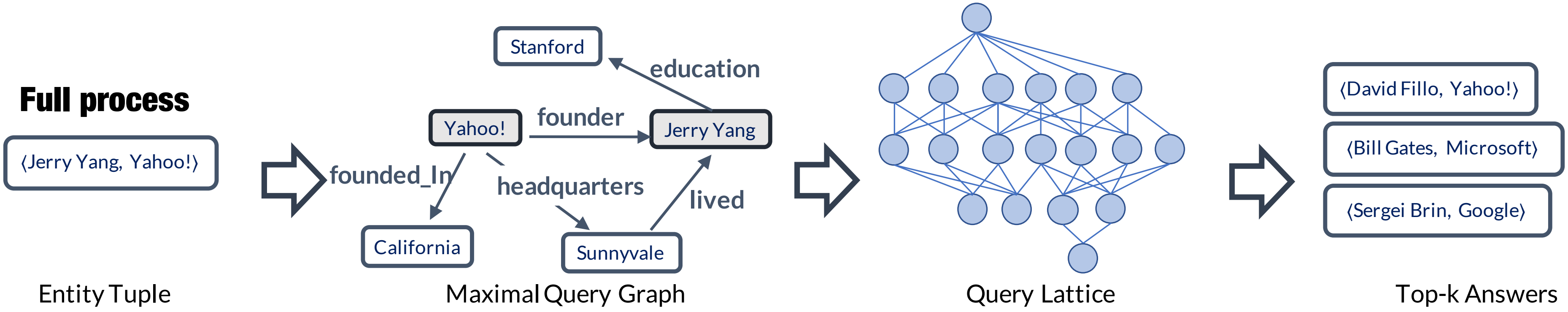


Find answers using a lattice obtained removing edges from the union graph

GQBE finds answers for multiple query tuples

Compute a re-weighted union graph of the individual query graphs

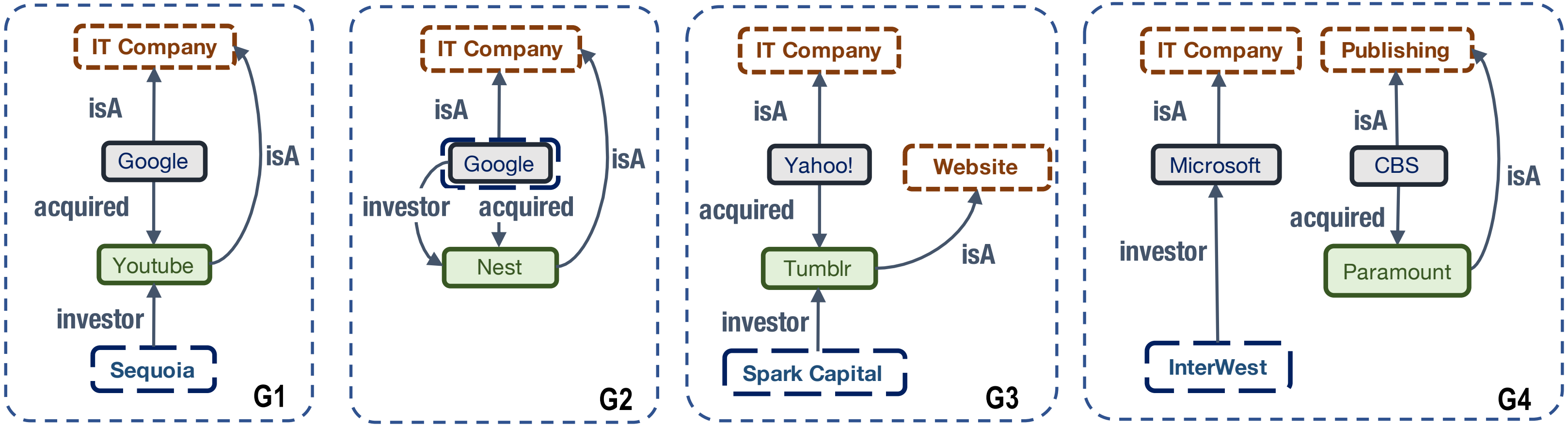
Full process



Graph Isomorphism vs. Simulation Variants

Structural Congruence/Similarity

Isomorphism requires an bijective function
Simulation requires only a surjective relation
Preserves only Parent → Child relationships



Example of **Simulating** (G1∼ {G2,G3,G4}) and **Strong-simulating** Graphs (G1≈G2)

Strong Simulation preserves close connectivity

Strong simulation: Capturing topology in graph pattern matching
– Shuai Ma et al., 2014

Computing Exemplar Queries (i)

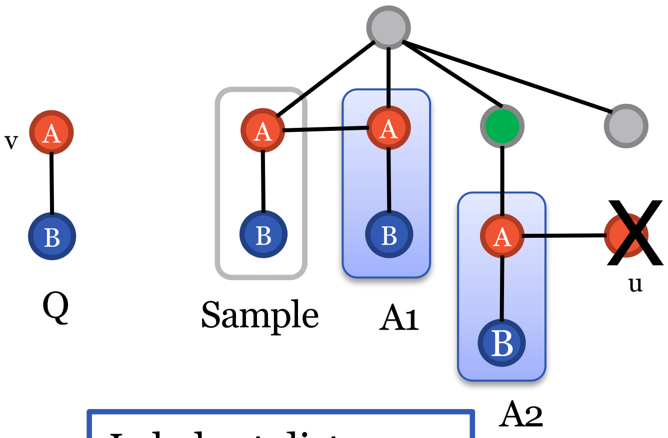
Fast Structure Matching

Mottin et al. [2016]

NP-complete
(subgraph isomorphism)

$O(|V|^4)$ (simulation)

Reduce Search Space:
Removes nodes that **cannot be part of a solution**



Labels at distance 1

Exact Pruning technique:

- Compute the neighbor labels of each node

$$W_{n,a,i} = \{n_1 | l(n_1, n_2) = a \vee \in N_{i-1}(n)\}$$

- **Prune nodes not matching query** nodes **neighborhood** labels
- Apply iteratively on the query nodes

neighborhood (v) = {(B,1)}
$\not\subseteq$
neighborhood (u) = {(A,1)}

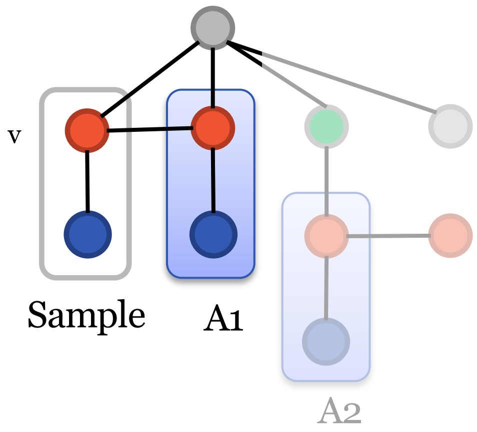
No Match

Computing Exemplar Queries (ii)

Mottin et al. [2016]

Prune Irrelevant Answers

Reduce Search Space:
Removes nodes that **are likely to be less relevant**



NP-complete (subgraph isomorphism)

$O(|V|^4)$ (simulation)

Approximation:

- Nodes closed to the sample are more important
- Use **Personalized PageRank** with a weighted matrix

$$v = (1 - c)Av + cp$$

- Weight edges: frequency of the edge-label

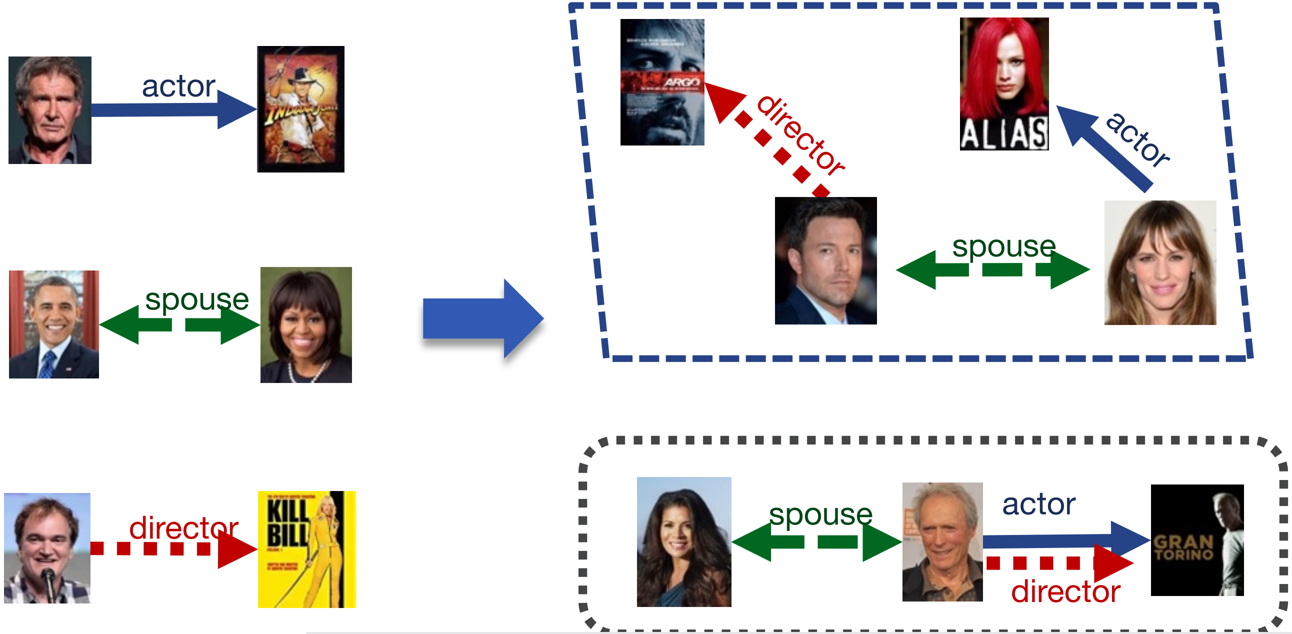
$$I(e_{ij}^{\ell}) = I(\ell) = \log \frac{1}{P(\ell)} = -\log P(\ell)$$

$$P(\ell) = \frac{|E^{\ell}|}{|E|}$$

Search with Multiple Examples

Combining partial answers

Lissandrini et al. [2018]



- Multiple Simple Examples
- Each Example describes an Aspect
- Results are Combinations of aspects
- Results have possibly Multiple Structures

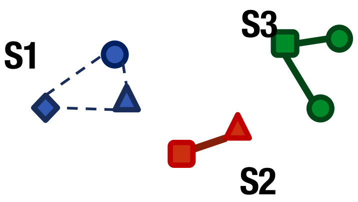
Case: Unknown Structures ⟶ Find Complex Connections with Simpler Components

M. Lissandrini, D. Mottin, T. Palpanas, Y. Velegrakis

Search Framework

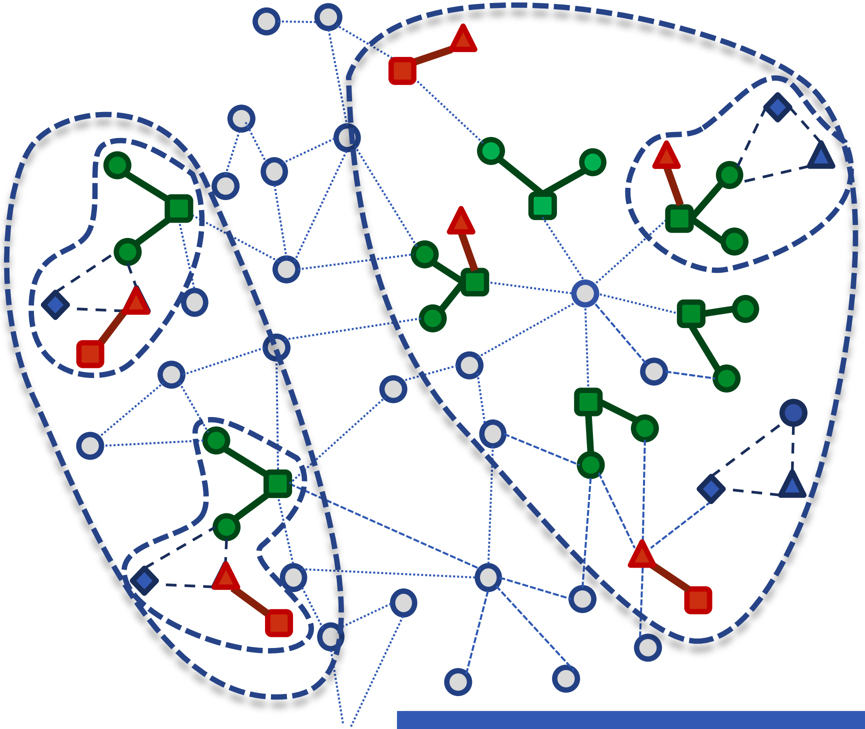
Pruning and Partial matching

Lissandrini et al. [2018]



Multi-exemplar Answering

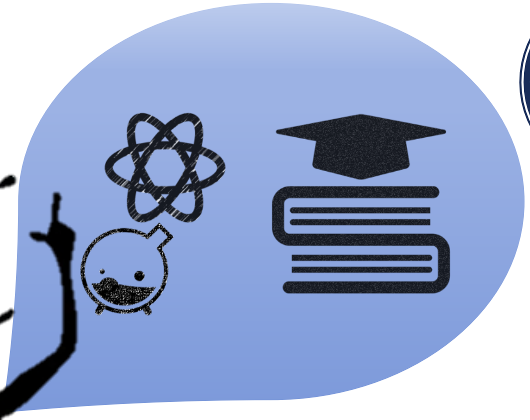
Input: Database $G : \langle V, E, \ell \rangle$
Input: Samples $\mathcal{S} : \langle s_1, ..., s_m \rangle$
Output: Answers $\mathcal{A}$
1: $\mathcal{G} \leftarrow \textsc{Partial}(G, \mathcal{S})$
2: $\mathcal{A} \leftarrow \textsc{Search}(\mathcal{G}, \mathcal{S})$
3: **return** $\mathcal{A}$



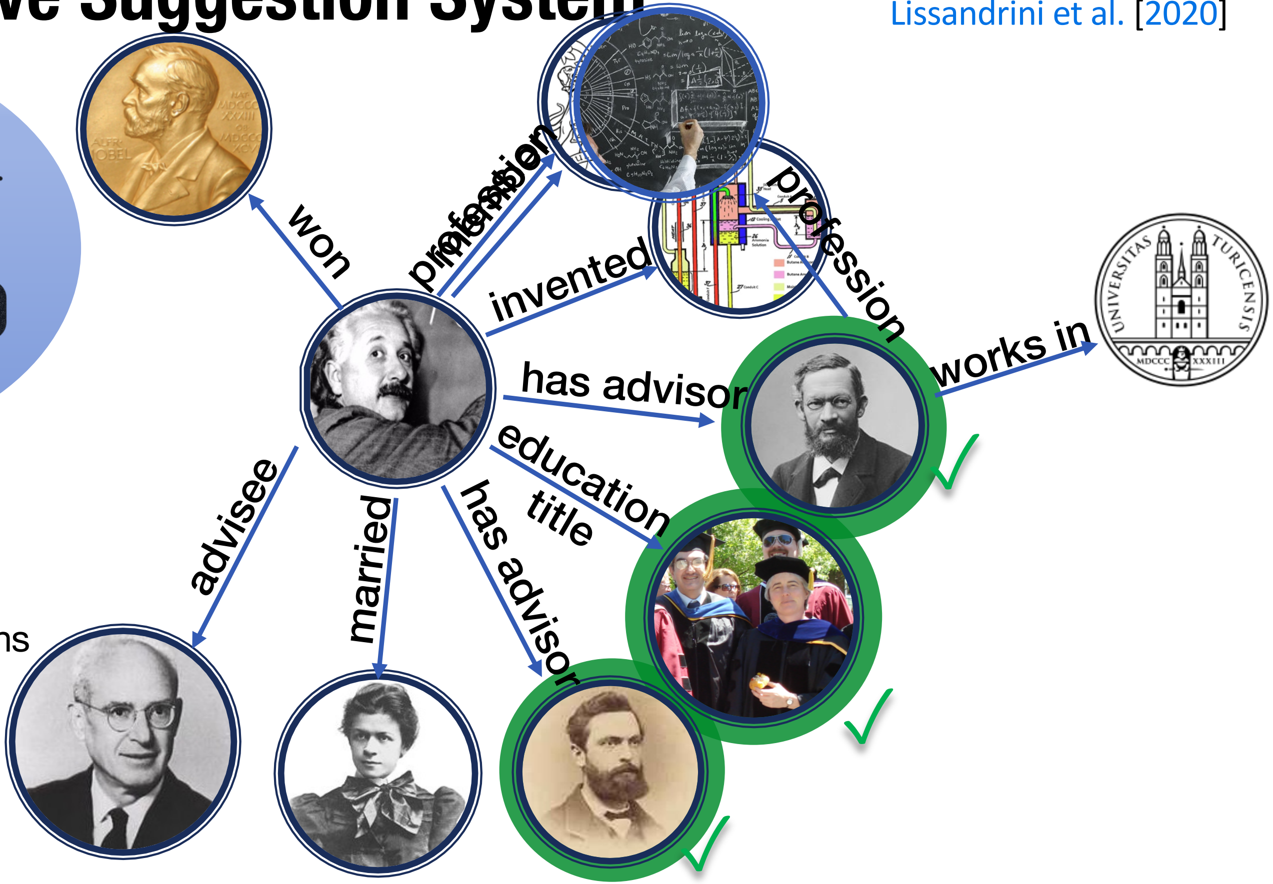
Exploit Localized Search

Solution: **Interactive Suggestion System**

Lissandrini et al. [2020]



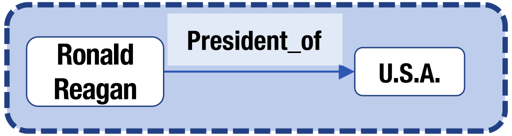
User Interaction
- Narrow the options
- Select **suggestion**
- **Focused** Expansions



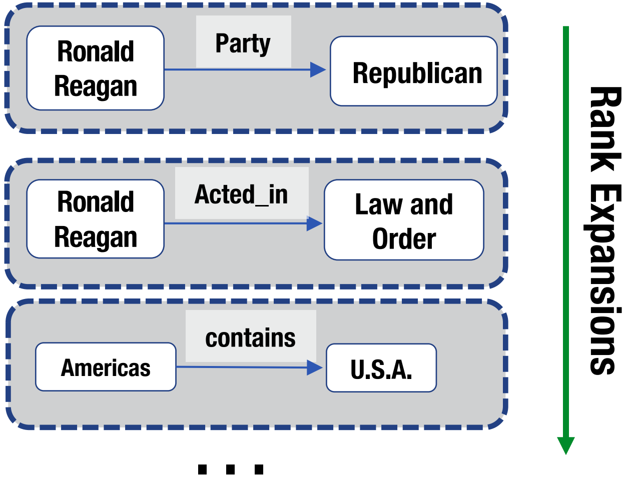
Suggesting Expansions

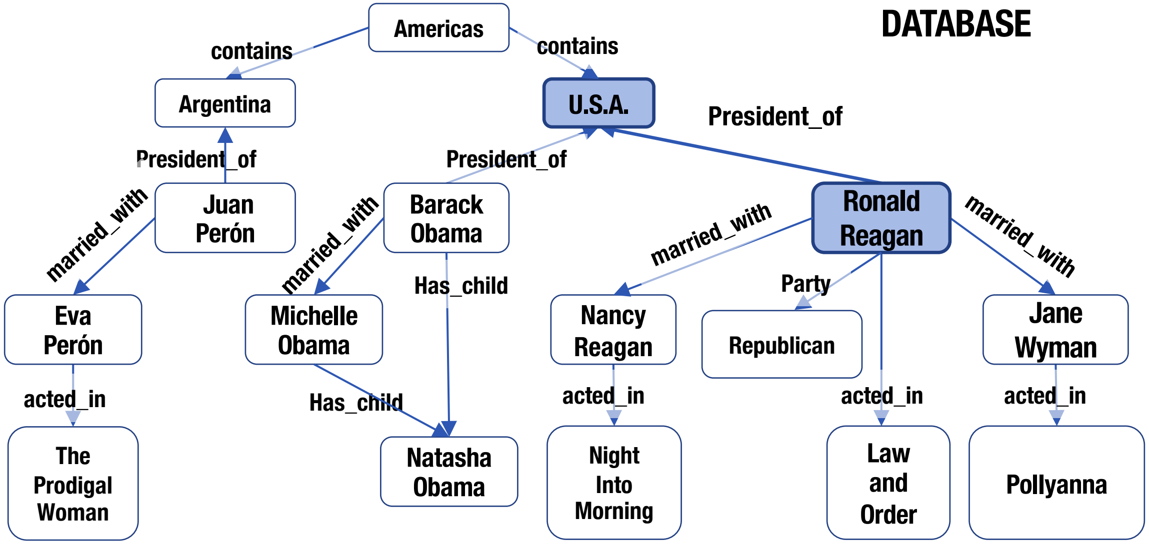
Lissandrini et al. [2020]

The User Search



Which Expansion to Suggest?





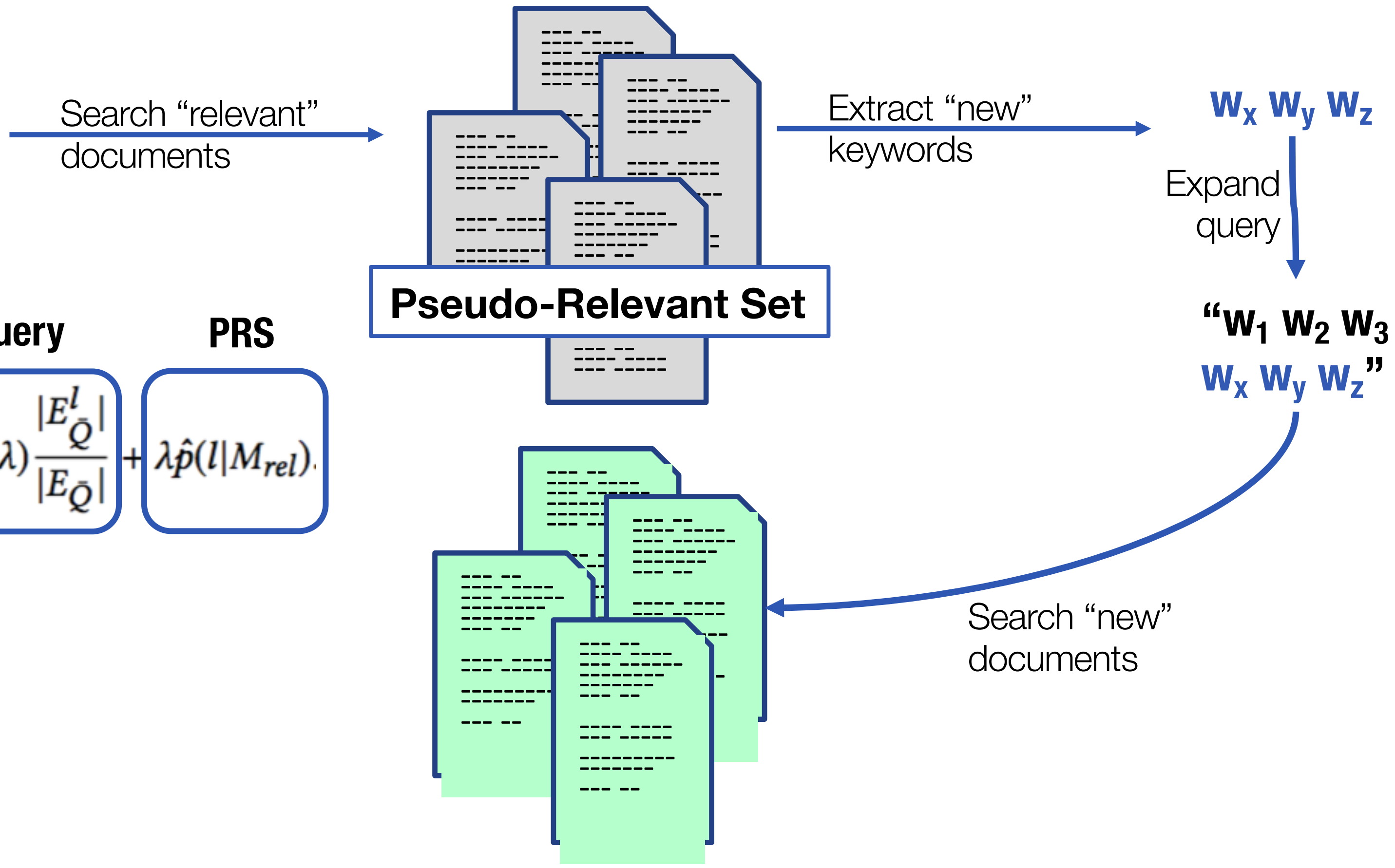
Approach:
Rank Edges based on Label characteristics

Pseudo Relevance Feedback for Document Search

Lissandrini et al. [2020]

Keyword Query

"w_1 w_2 w_3"



Query **PRS**

$$\hat{p}(l|M_{\bar{Q}})_{REL} = \boxed{(1-\lambda)\frac{|E^l_{\bar{Q}}|}{|E_{\bar{Q}}|}} + \boxed{\lambda\hat{p}(l|M_{rel})}.$$

Pseudo Relevance Feedback Models

Lissandrini et al. [2020]

Maximum Likelihood Estimation

$$\hat{p}(l|M_{rel})_{MLE} \approx \sum_{\bar{G}\in\bar{\mathcal{G}}_{rel}} \hat{p}(l|M_{\bar{G}})\hat{p}(\bar{Q}|M_{\bar{G}})$$

$$\hat{p}(\bar{Q}|M_{\bar{G}}) \propto \prod_{l\in\bar{Q}} \hat{p}(l|M_{\bar{G}})$$

PRS

$$\hat{p}(l|M_{\bar{Q}})_{REL} = (1-\lambda)\frac{|E^{l}_{\bar{Q}}|}{|E_{\bar{Q}}|} + \boxed{\lambda\hat{p}(l|M_{rel})}.$$

KL-Divergence

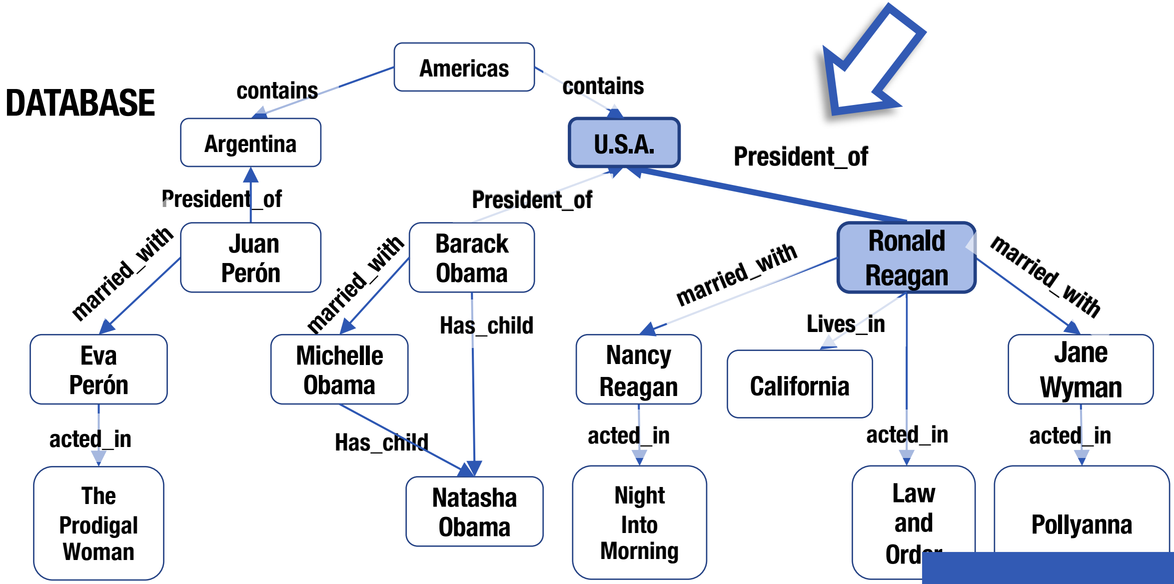
$$\hat{p}(l|M_{rel})_{KL} \propto$$

$$exp\left(\frac{1}{(1-\lambda)}\frac{1}{|\bar{\mathcal{G}}_{rel}|}\sum_{\bar{G}}^{\bar{\mathcal{G}}_{rel}} \log\left(\hat{p}(l|M_{\bar{G}})\right) - \frac{\lambda}{(1-\lambda)}\log\left(\hat{p}(l|\mathcal{K})\right)\right)$$

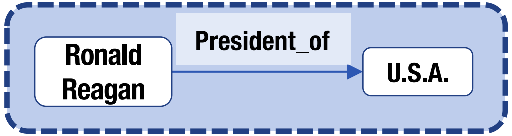
2 Models of Estimation
MLE & KL-Divergence

Bag Model for Graphs

Lissandrini et al. [2020]



The User Search



How can we convert to the document model?

The Bag-of-Labels Model

{ President_of, contains, married_with, married_with, acted_in, lives_in }

- Graphs can be modeled as **Bag of Words**
- Describes **MORE than what is in the query**

https://www.youtube.com/watch?v=A1_dKvX5ZRk

SEARCHING FOR

Nodes

Structures

BY **LOOKING** AT

PRODUCES

Connectivity

Mediator Nodes
[Gionis et al. ‘15
Ruchansky et al.’15]

Clusters
[Perozzi et al.’14,
Kloumann et al. 14]

Properties

Similar Entities
[Metzger et al.’13,
Sobczak et al.’15]

Queries

Path Queries
[Bonifati et al.’15]

SPARQL
[Arenas et al.’16]

(Edge-)Labels

Entity Tuples
[Jayaram et al.’15]

Similar
Structures
[Mottin et al.’14,
Xie et al.’17,
Lissandrini et al’18]

Few Approaches accept User Feedback

M. Lissandrini, D. Mottin, T. Palpanas, Y. Velegrakis

Where we are

Relational databases

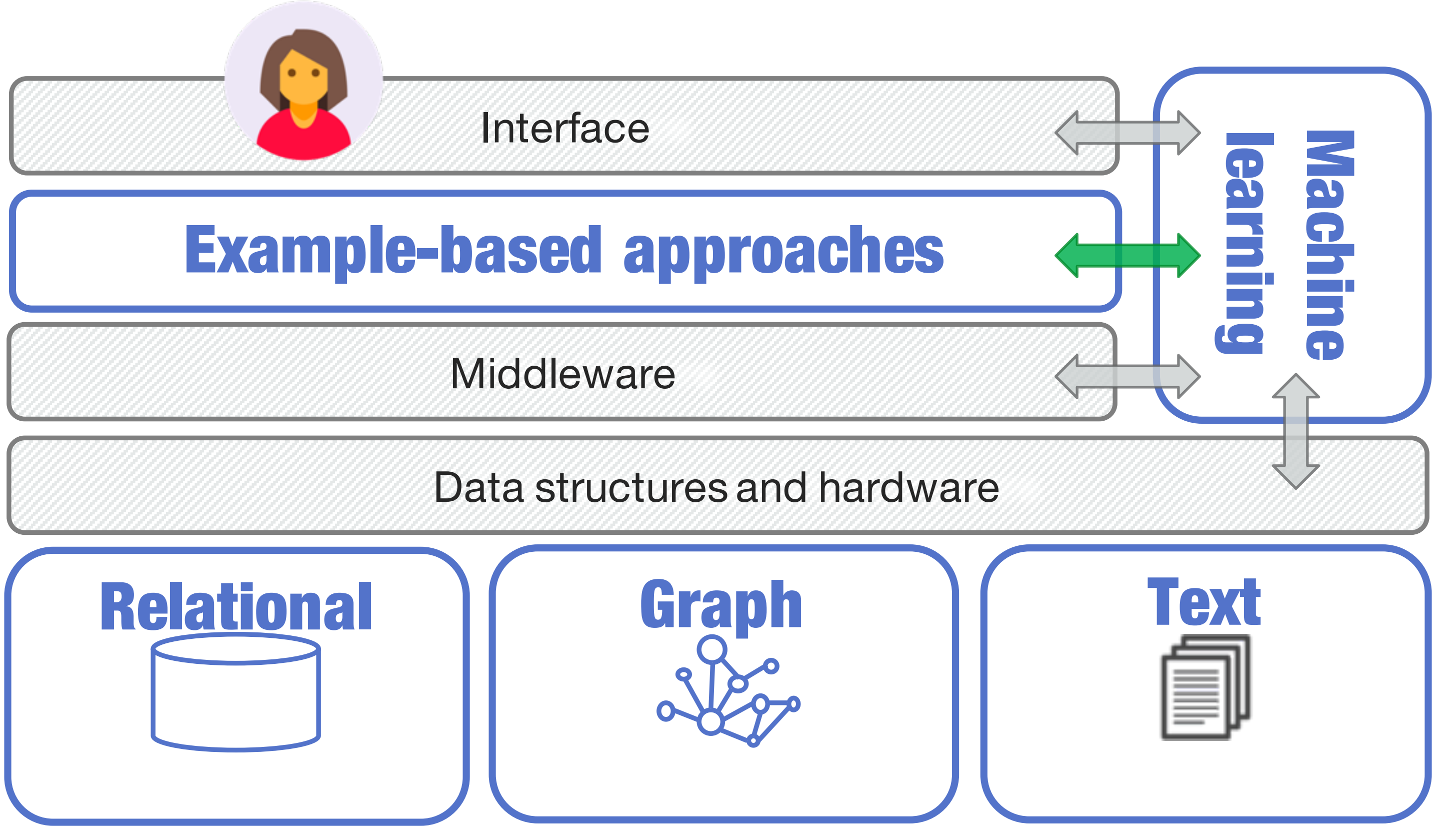
Textual data

Graphs and networks

Machine learning

Challenges and Remarks

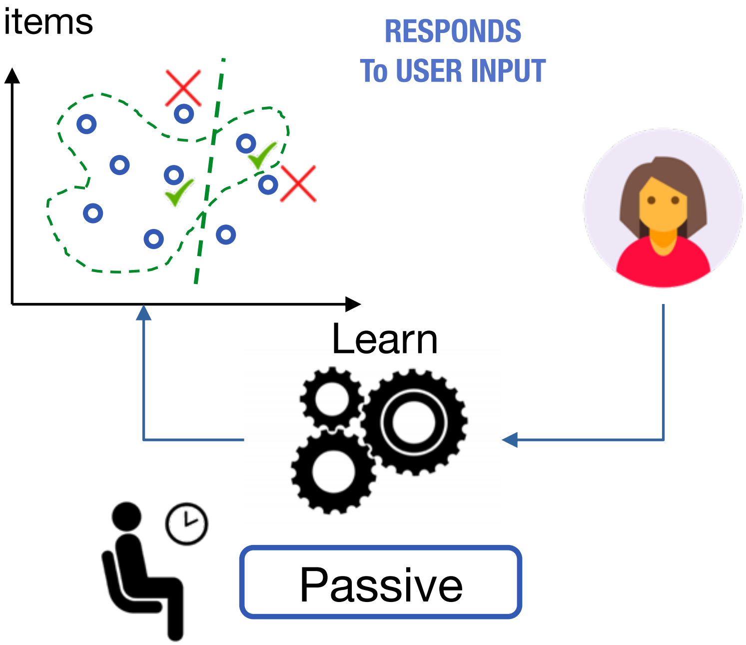
How ML fits the Big Picture

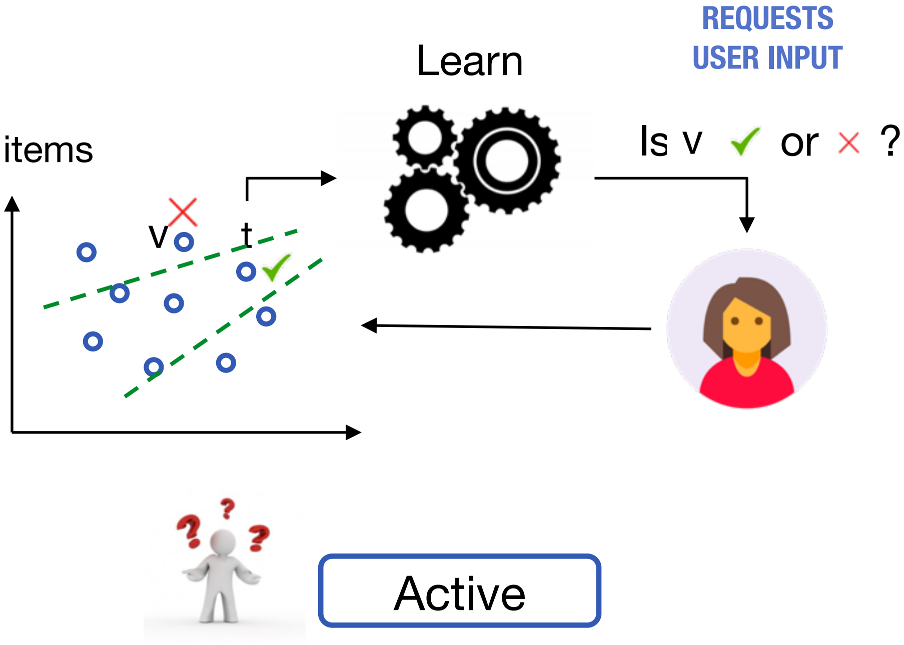


Interactive exploration of datasets

Main idea: Learn the items to show online as more points are acquired

Two ways of learning: passive and active





M. Lissandrini, D. Mottin, T. Palpanas, Y. Velegrakis

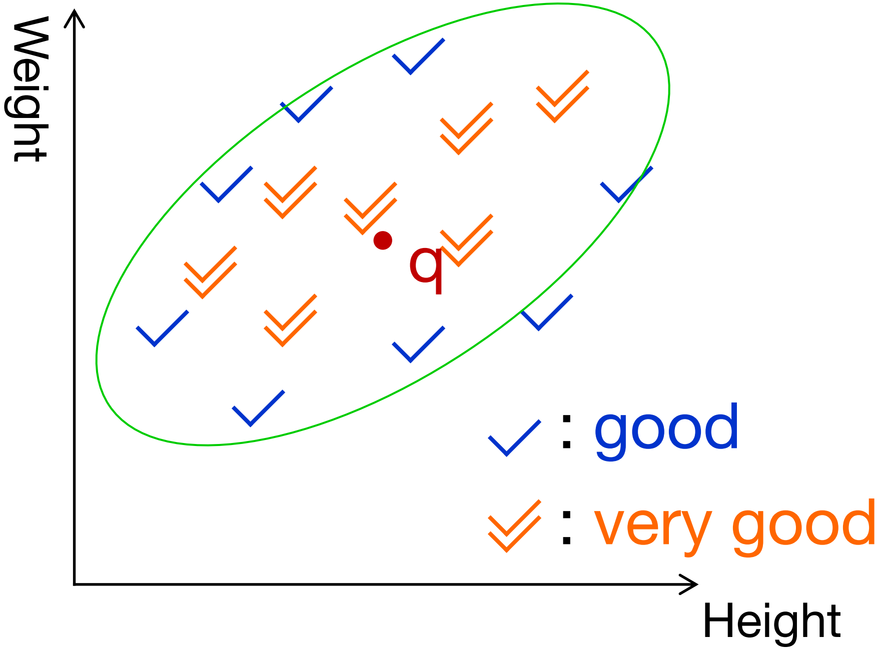
MindReader

[Ishikawa et al., 1999]

Main idea: learn an implicit query from user examples and optional scores

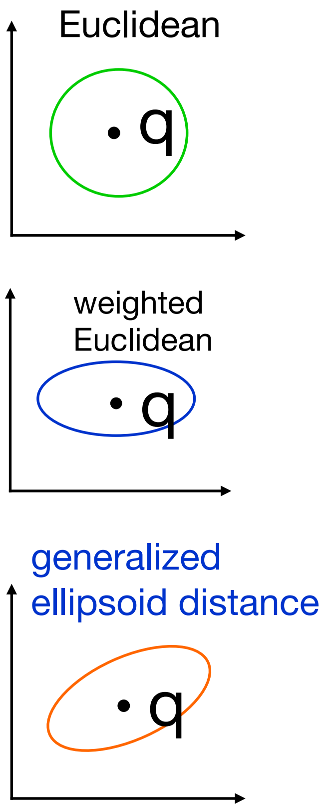
Searching "mildly overweighted" patients

- The doctor selects examples by browsing patient database
- The examples have **"oblique"** correlation
- We can "guess" the implied query



Learning an ellipsoid distance

[Ishikawa et al., 1999]



Weighted distance matrix

$$D(x,q) = (x-q)^{\top} M (x-q)$$

Implicit query

$$D(x,q) = \sum_{j}^{n} \sum_{k}^{n} m_{jk}(x_j - q_j)(x_k - q_k)$$

Learn the query minimizing the penalty = weighted sum of distances between query point and sample vectors

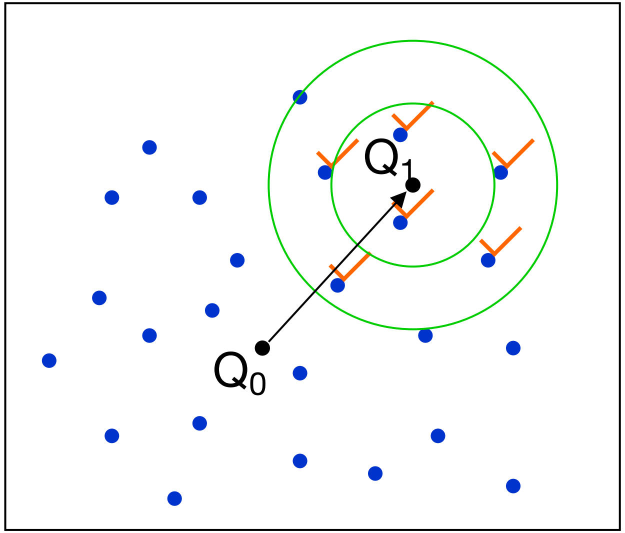
$$\text{minimize} \sum_{i} (x_i - q)^{\top} M (x_i - q)$$

$$\text{subject to} \quad \det(M) = 1$$

Learning the distance

[Ishikawa et al., 1999]

Query point is moved towards “good” examples — Rocchio formula in IR



Q_0: query point

: retrieved data

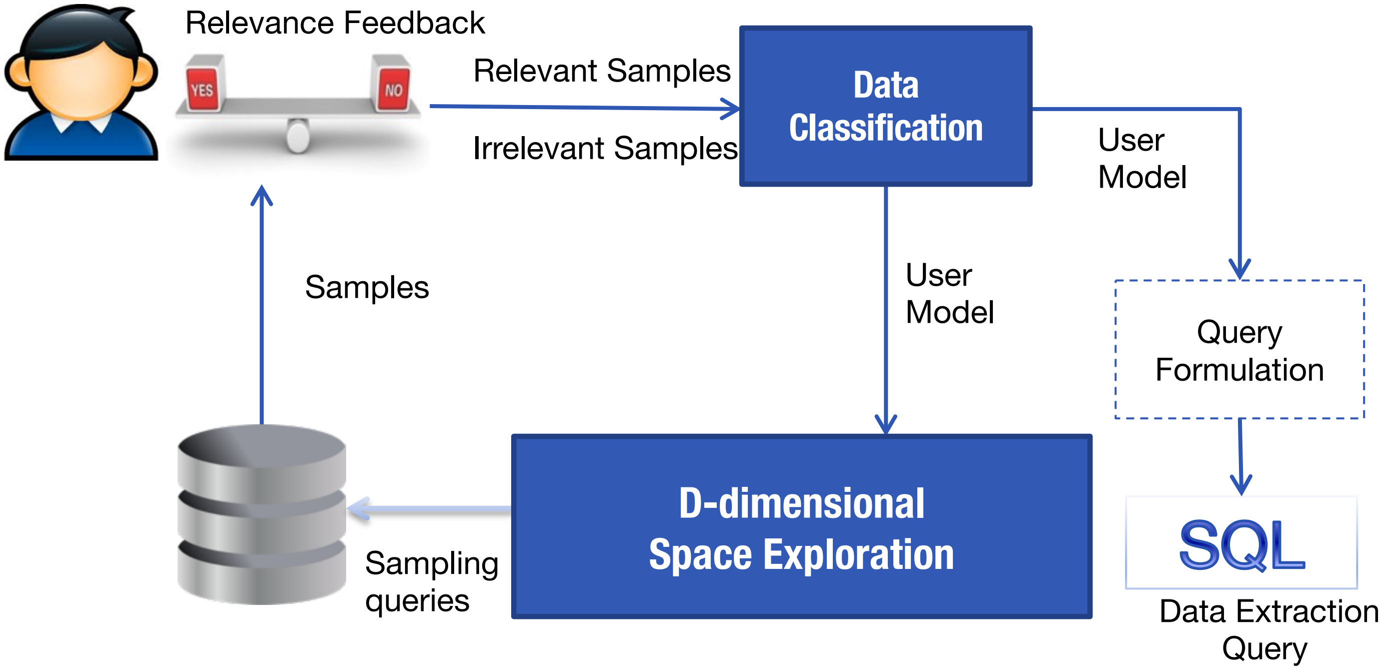
: relevance judgments

Q_1: new query point

Learning can be done online!!!

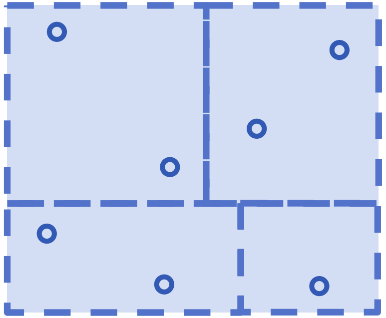
Explore-by-Example: AIDE

[Dimitriadou et al., 2014,2016]

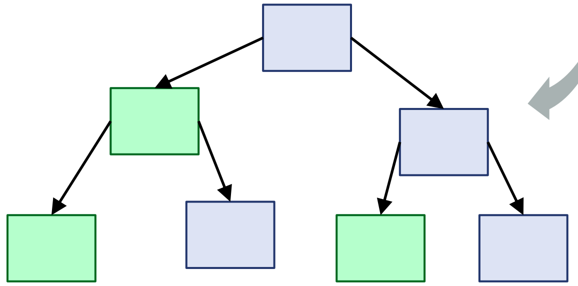


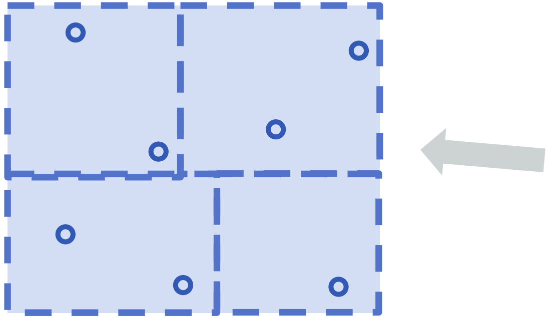
The AIDE algorithm

[Dimitriadou et al., 2014, 2016]



1. Divide the space into d-dimensional cubes
2. Find the sample points in the cubes (medoids)
3. Train the classifier
4. Refine the training sampling from neighbors of misclassified points
5. Boundary refinement

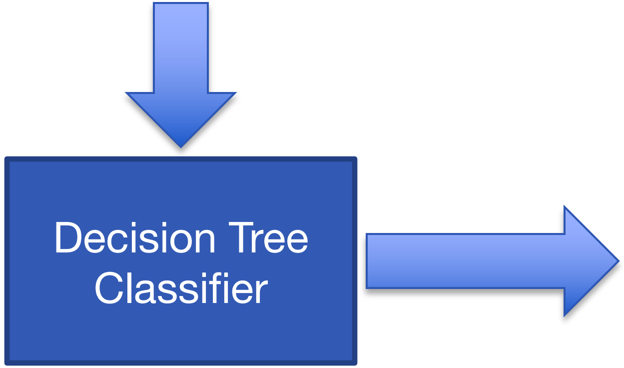


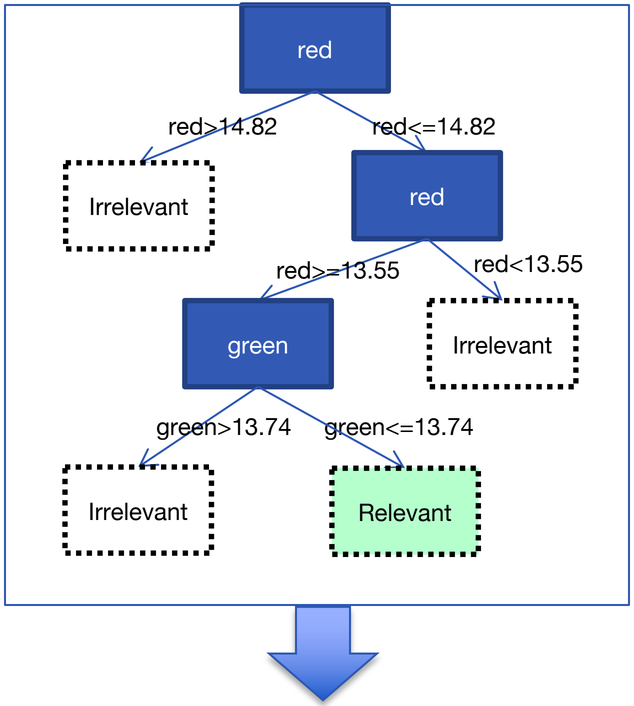


Classification & Query Formulation

[Dimitriadou et al., 2014, 2016]

Sample	Red	Green	Relevant
Object A	13.67	12.34	Yes
Object B	15.32	14.50	No
..	..	..	...
Object X	14.21	13.57	Yes

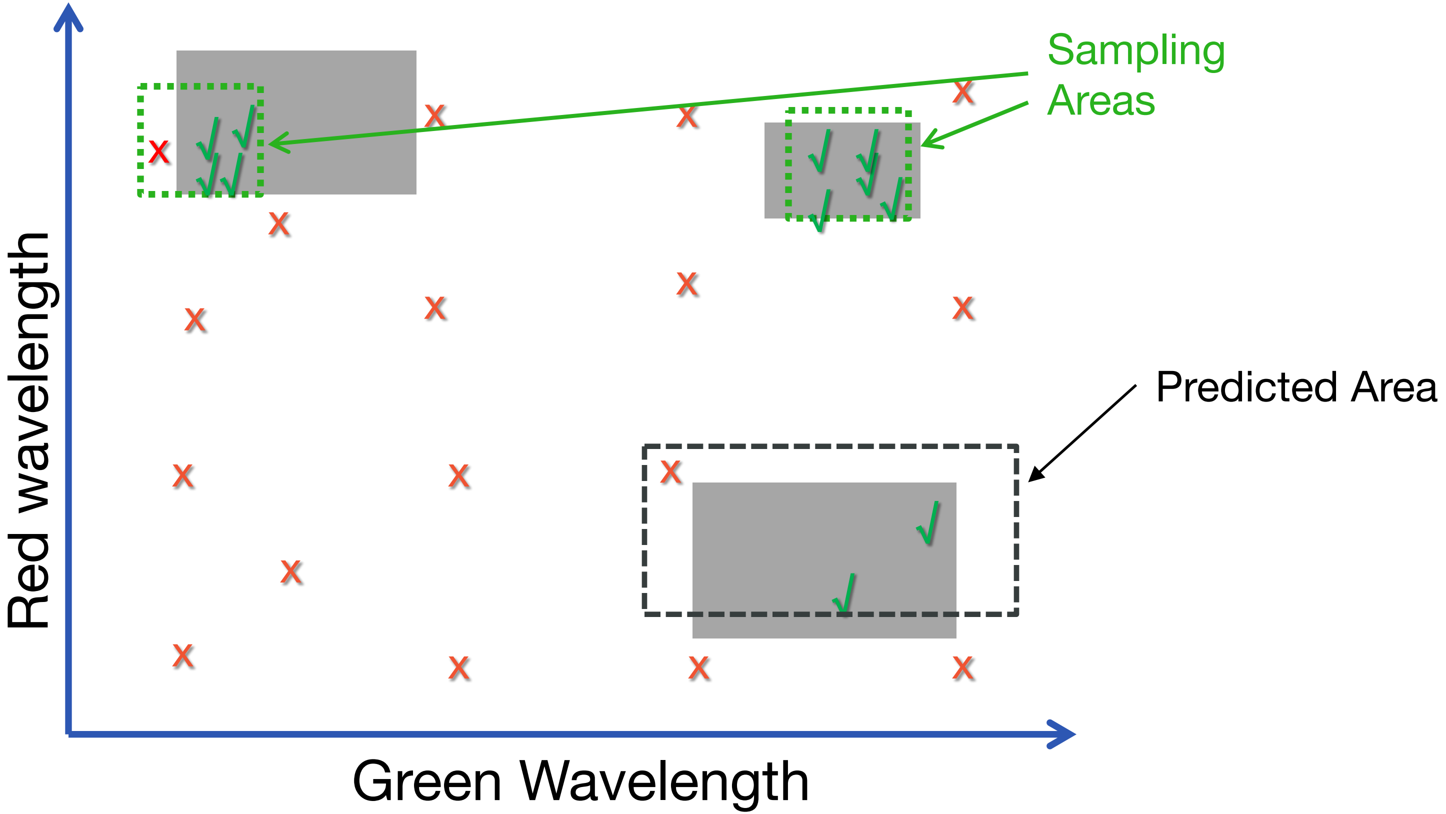




SELECT * FROM galaxy WHERE red<= 14.82 AND red>= 13.5 AND green<=13.74

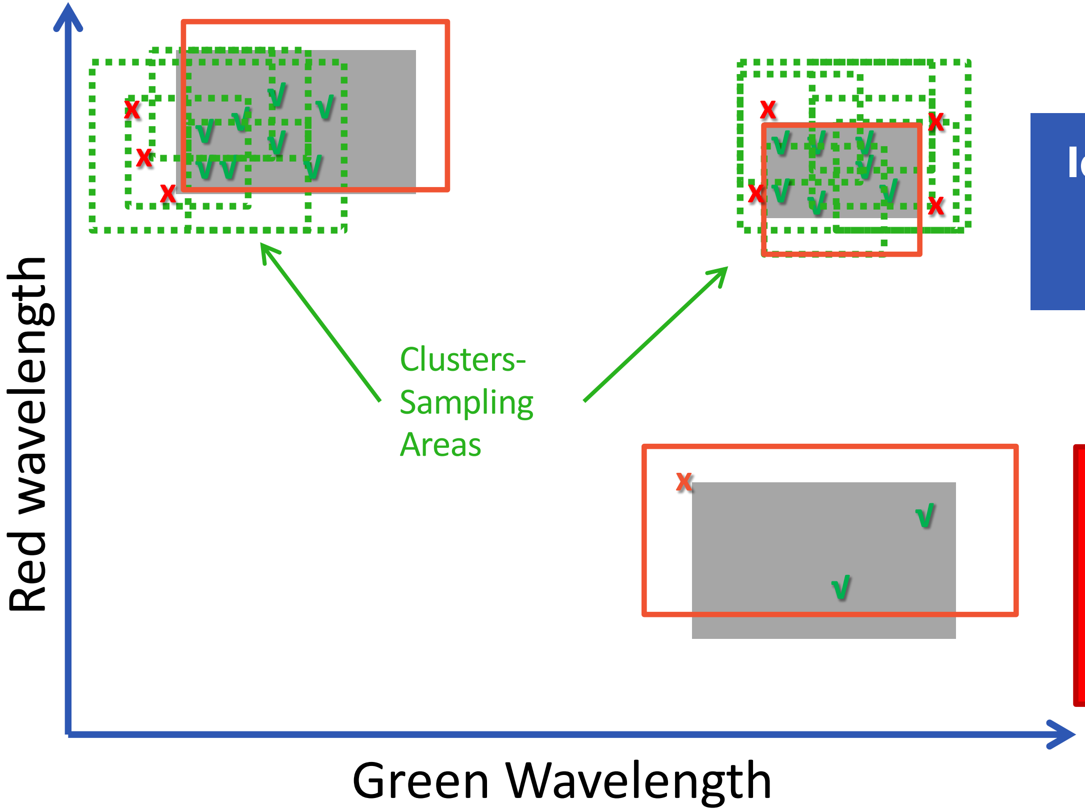
Misclassified Sample Exploitation

[Dimitriadou et al., 2014,2016]



Clustering-based Sampling

[Dimitriadou et al., 2014,2016]



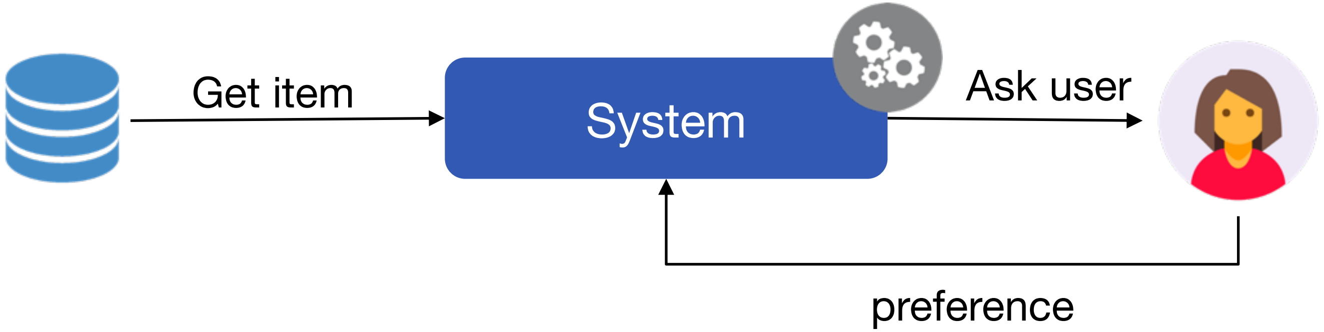
Idea: Use a k-medoid approach to find sampling areas

Iterative approach: How many samples does it take to reach the desired result?

Active learning for online query systems

[Vanchinathan et al., 2015]

Main idea: the system "queries" the user to understand their preferences

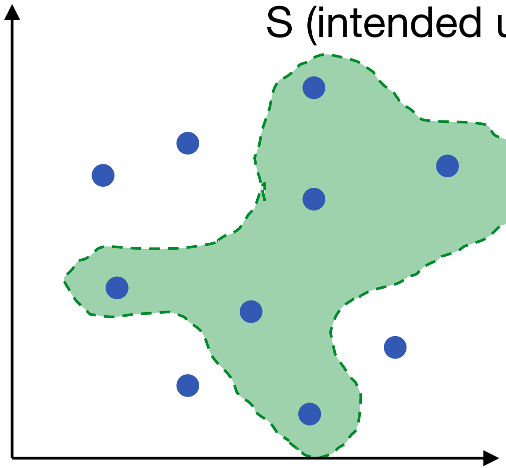


Learn unknown preferences and minimize the number of questions to the user

Learning unknown preferences

[Vanchinathan et al., 2015]

Problem: Find a set S that maximize the unknown user preference within a budget (e.g., number of interactions)



User preferences

$$\arg\max \sum_{v \in S} pref(v)$$

$$\text{subject to } \; Cost(S) \leq budget$$

Cost for the set S

A step back …

Learning from an unknown environment …

Multi-armed bandits

- Maximize the reward by successively playing gamble machines (the ‘arms’ of the bandits)
- Invented in **early 1950s** by Robbins for decision making under uncertainty when the environment is unknown
- The reward is unknown ahead of time

…

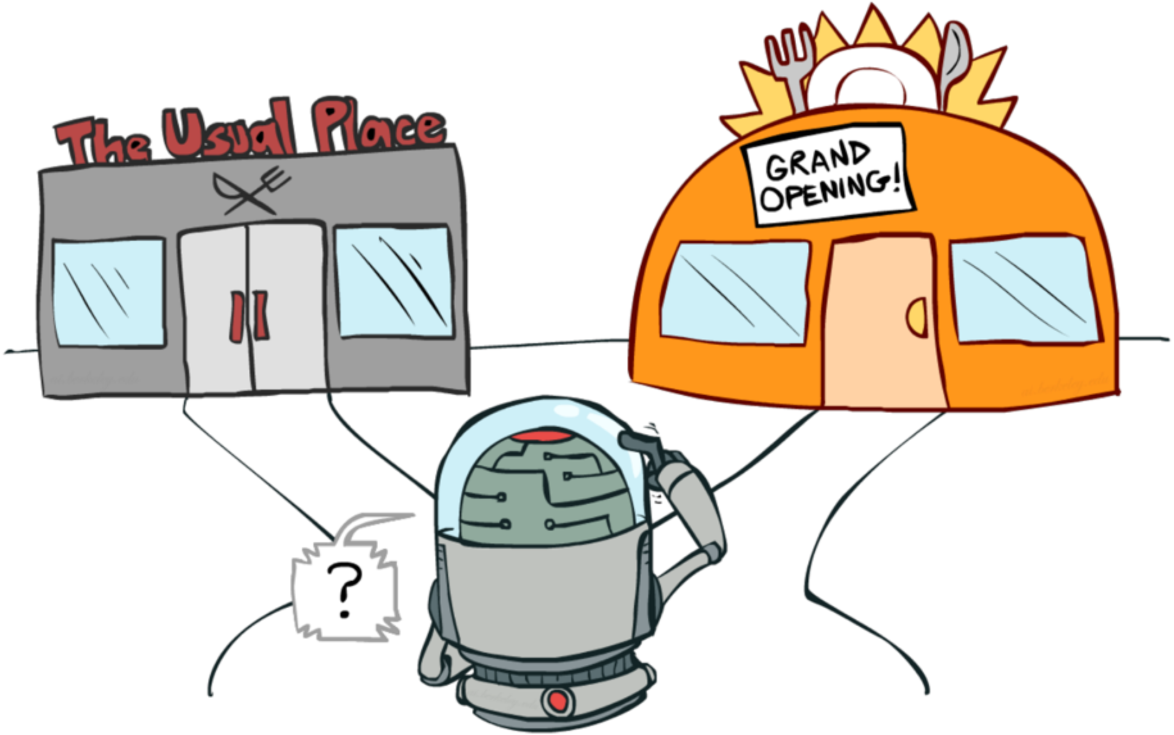
Reward X_1 Reward X_2 Reward X_3

Multi-armed bandits

- Reward = random variable $X_{i,n}$; $1 \leq i \leq K, n \geq 1$
- i = index of the gambling machine
- n = number of plays
- μ_i = expected reward of machine i.

A policy, or allocation strategy A is an algorithm that chooses the next machine to play based on the sequence of past plays and obtained rewards.

Exploration vs Exploitation



https://lilianweng.github.io/lil-log/2018/01/23/the-multi-armed-bandit-problem-and-its-solutions.html

Greedy: A pure exploitation algorithm

Choose the machine with current best expected reward

- **Exploitation vs exploration dilemma:** Should you exploit the information you've learned or explore new options in the hope of greater payoff?

- In the **greedy case**, the balance is completely towards **exploitation**
- Yet, **only exploitation will not lead to a good solution**

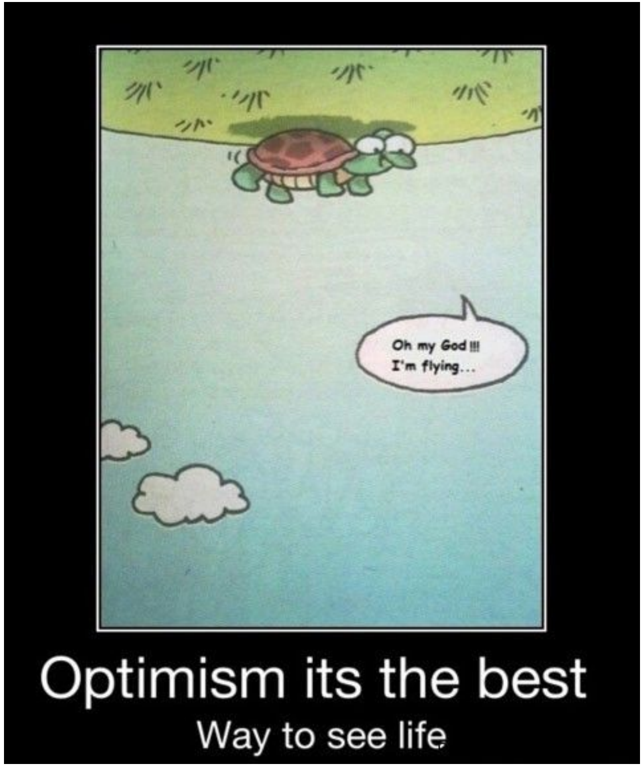
Quality measure - Regret

Total expected regret (after T plays):

$$R_T \ = \mu^* \cdot T - \sum_{i=1}^{K} \mu_j \cdot \mathbb{E}\left[N_{i,T}\right]$$

μ^*: highest expected reward
$\mathbb{E}[N_{i,T}]$: expected number of times machine *i* is played

An optimistic view



Upper confidence bound (UCB) algorithm

Optimistic estimate of the mean of arm = ‘largest value it could plausibly be’

1. Pull at each time t the arm with the maximum probability of being the best

$$\frac{1}{n_j}\sum_{s=1}^{n_j} X_{j,s} + \sqrt{\frac{2\log(1/t)}{n_j}}$$

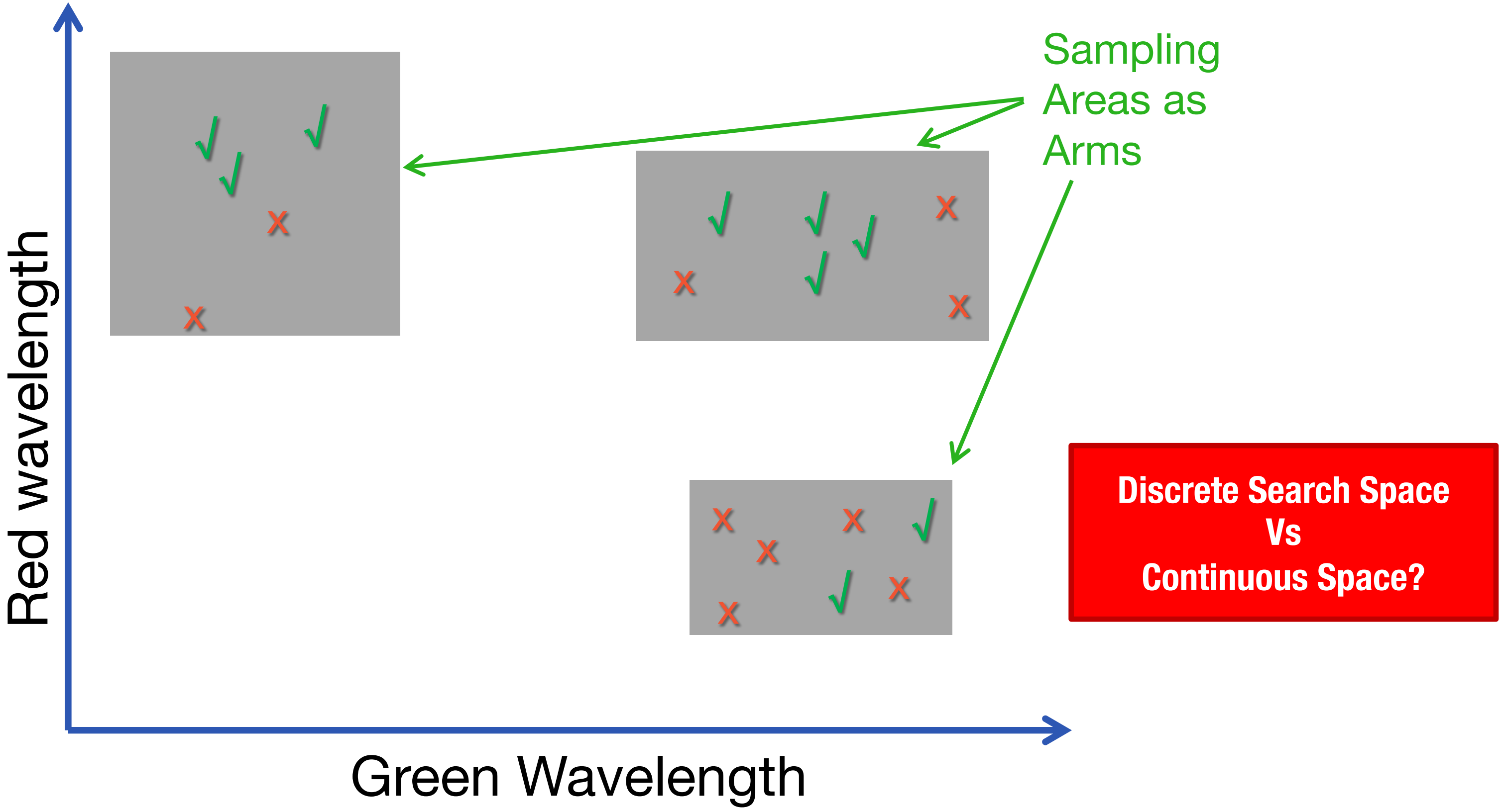
2. Repeat until the budget (number of steps T) is depleted

n_j: number of times the arm j has been pulled

Balance exploration and exploitation: The uncertainty diminishes as the time passes

Back to our problem

Modeling the same problem as a Multi-Armed Bandit



Background: Gaussian processes

[Bishop et al., 2006]

Idea: Model the user preferences as a Gaussian Process

A Gaussian Process (GP) is an infinite set of variables, any subset of this is Gaussian

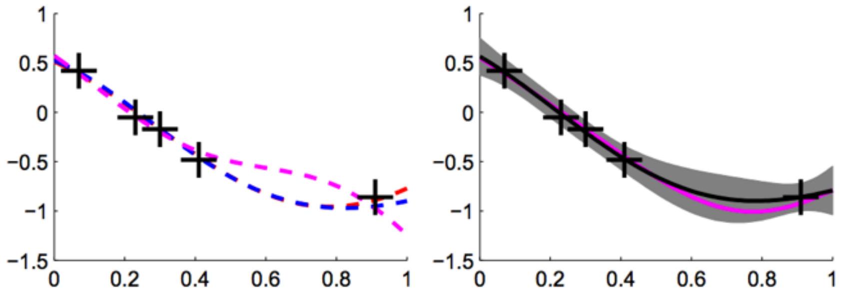
$$P(\mathbf{f}|\Sigma,\mu) = |2\pi\Sigma|^{\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{f}-\mu)^{\top}\Sigma^{-1}(\mathbf{f}-\mu))$$

Gaussian prior

Specified only by mean and covariance

Given observations $\{x,y\}_{i=1}^{n}$ over an unknown function f drawn from a Gaussian prior, the posterior is Gaussian

$$P(\mathbf{f}|\mathbf{y}) \propto \int dx\, P(\mathbf{f},\mathbf{x},\mathbf{y})$$



M. Lissandrini, D. Mottin, T. Palpanas, Y. Velegrakis

GP-Select

[Vanchinathan et al., 2015]

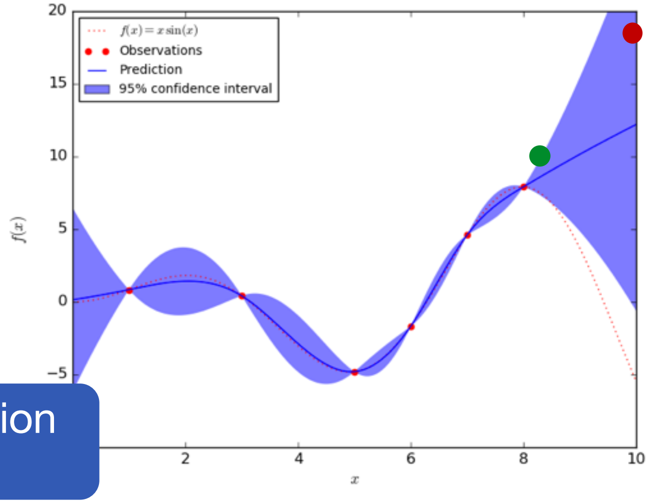
Algorithm 1 GP-SELECT

Input: Ground Set $\mathbf{V}$, kernel κ and budget B
Initialize selection set S
for $t = 1, 2, \ldots, B$ **do**
  Model Update:
    $[\mu_{t-1}(\cdot), \sigma^2_{t-1}(\cdot)] \leftarrow \text{GP-Inference}(\kappa, (S, y_{\{1:t-1\}}))$
  Item Selection:
    Set $v_t \leftarrow \underset{v \in \mathbf{V}/\{v_{1:t-1}\}}{\operatorname{argmax}} \; \mu_{t-1}(v) + \beta_t^{1/2} \sigma_{t-1}(v)$
  $S \leftarrow S \cup \{v_t\}$
  Receive feedback $y_t = f(v_t) + \epsilon_t$
end for

Learn posterior

Trades off exploration exploitation

Ask user feedback

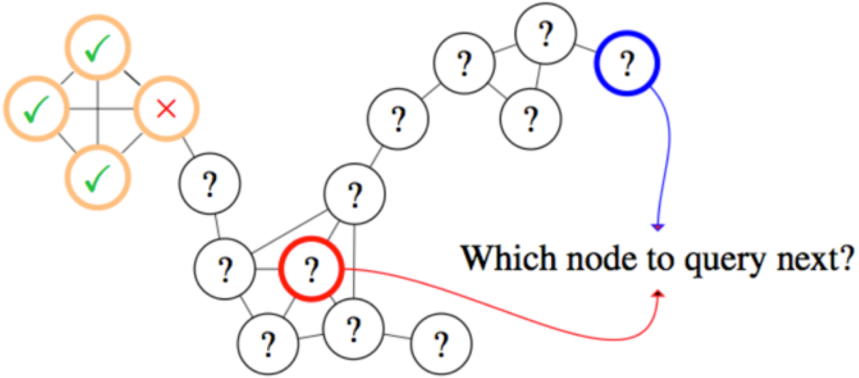


- Exploration: select items with high-variance
- Exploitation: select items with high-value

Active learning on graphs – which prior?

[Ma et al., 2015]

Idea: Use the graph structure to infer the node classes



Use graph Laplacian as prior
$L = D - A$, A is the adjacency matrix

$$p(\mathbf{f}) \sim \mathcal{N}(0, L^{-1})$$

Laplacian: higher probability of having the same class if two nodes are connected

Where could Active learning help?

Reverse engineering queries and rules

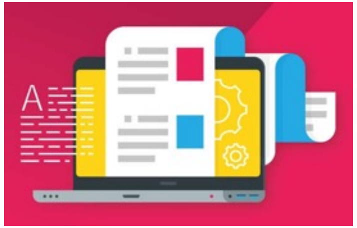
- Interactive Refinement of example tuples
- Learning the most probable queries from their results

Graph exploration

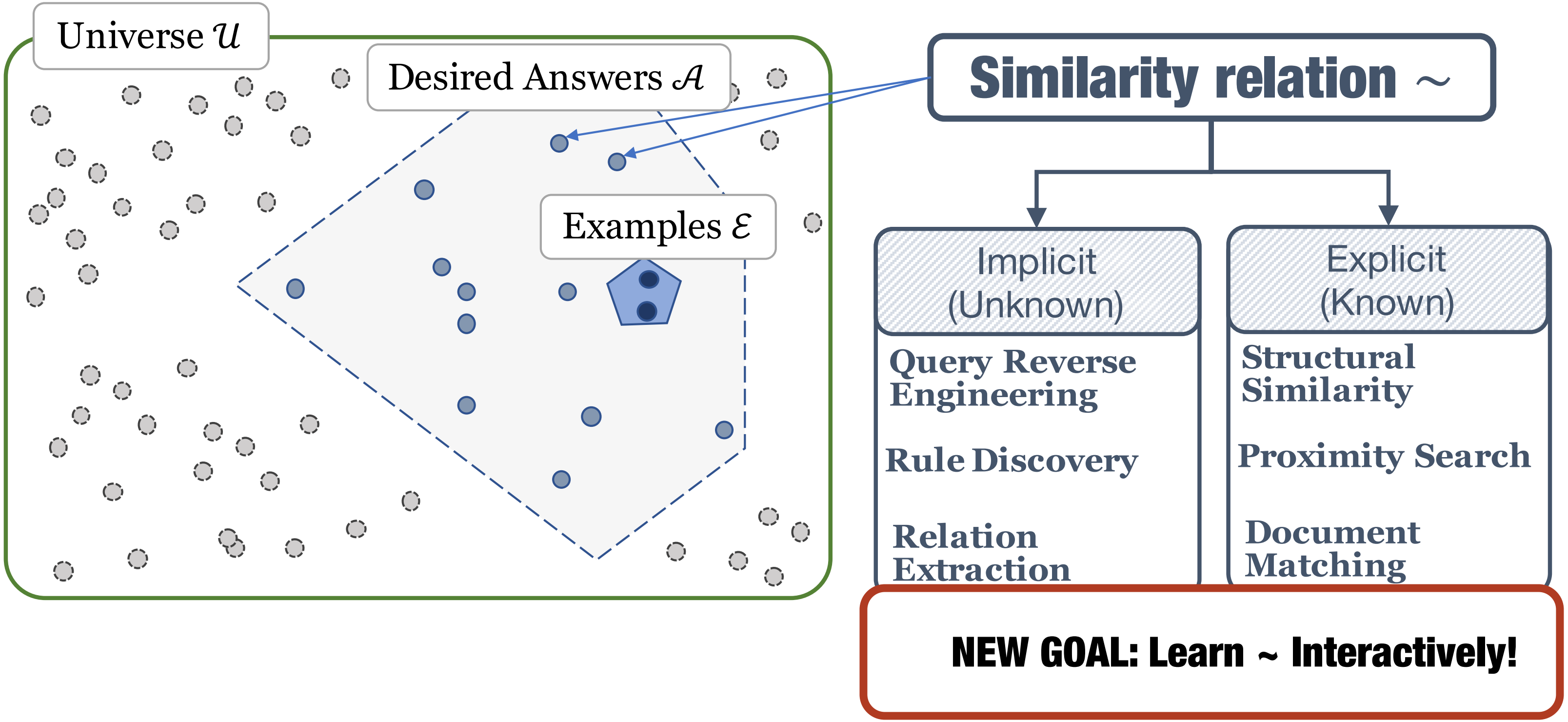
- Summarization of knowledge graphs with preferences
- Seed set expansion
- Recommendation of relevant nodes

Text processing

- Fast entity matching
- Advertising based on documents search



Example-based methods



MAB: good resources

Books and surveys

- http://slivkins.com/work/MAB-book.pdf
- http://downloads.tor-lattimore.com/book.pdf
- http://sbubeck.com/SurveyBCB12.pdf

Tutorials

- Lattimore - AAAI 2018: part 1 - part 2
- Tutorial on bayesian optimization of expensive cost functions
- Blog on bandits: http://banditalgs.com/

Where we are

Relational databases

Textual data

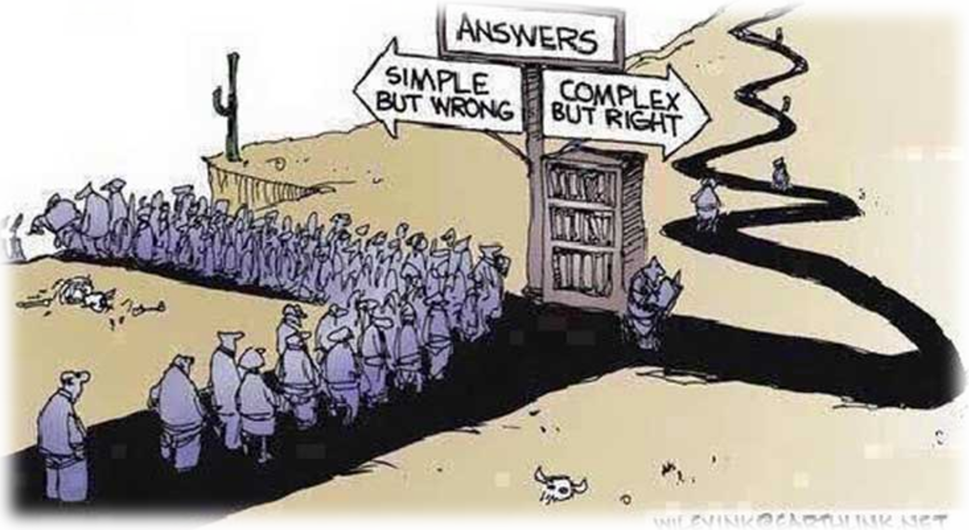
Graphs and networks

Machine learning

Challenges and Remarks

Big data – Easy value?

© marketoonist.com



Exploration

We know where we start
we don't know what we'll find

Traditional Search Methods are not Enough
We need Specialized Methods for Data Exploration

From broad views

to Detailed view

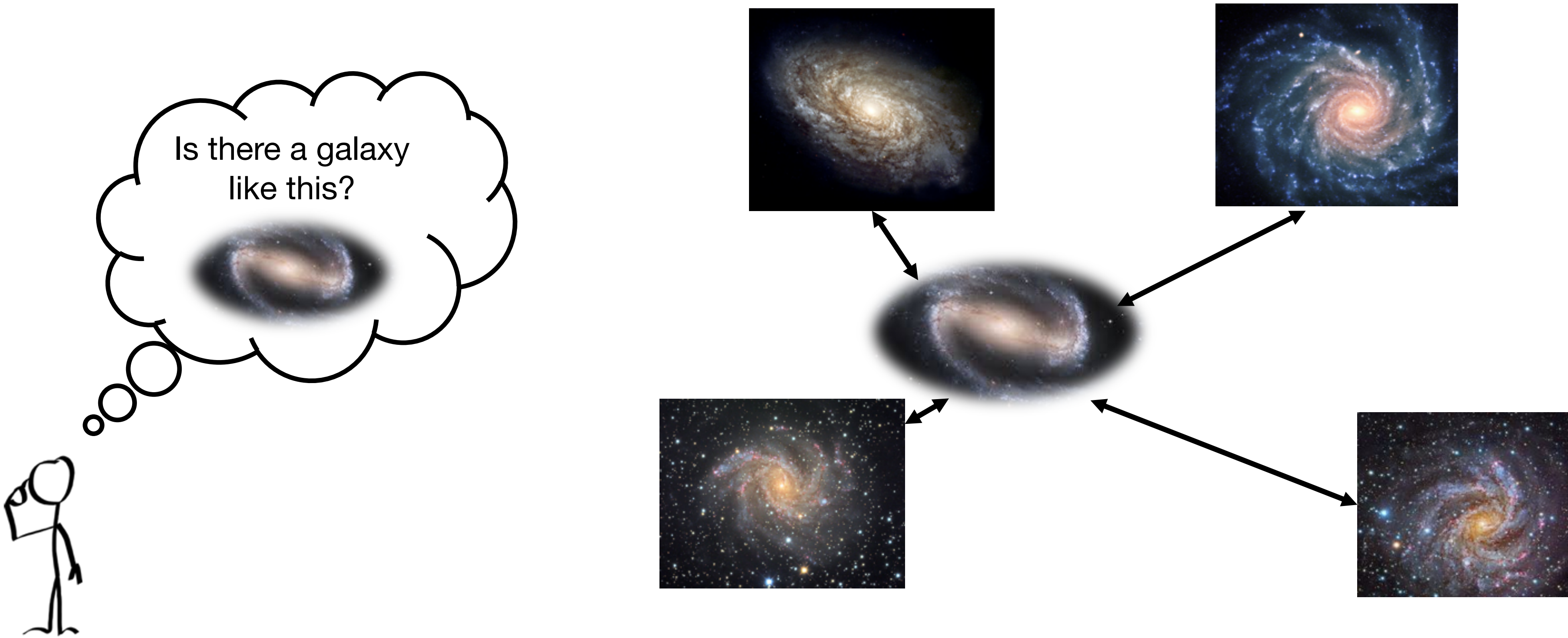
From exploration as
`select count(*)`

to find what is interesting

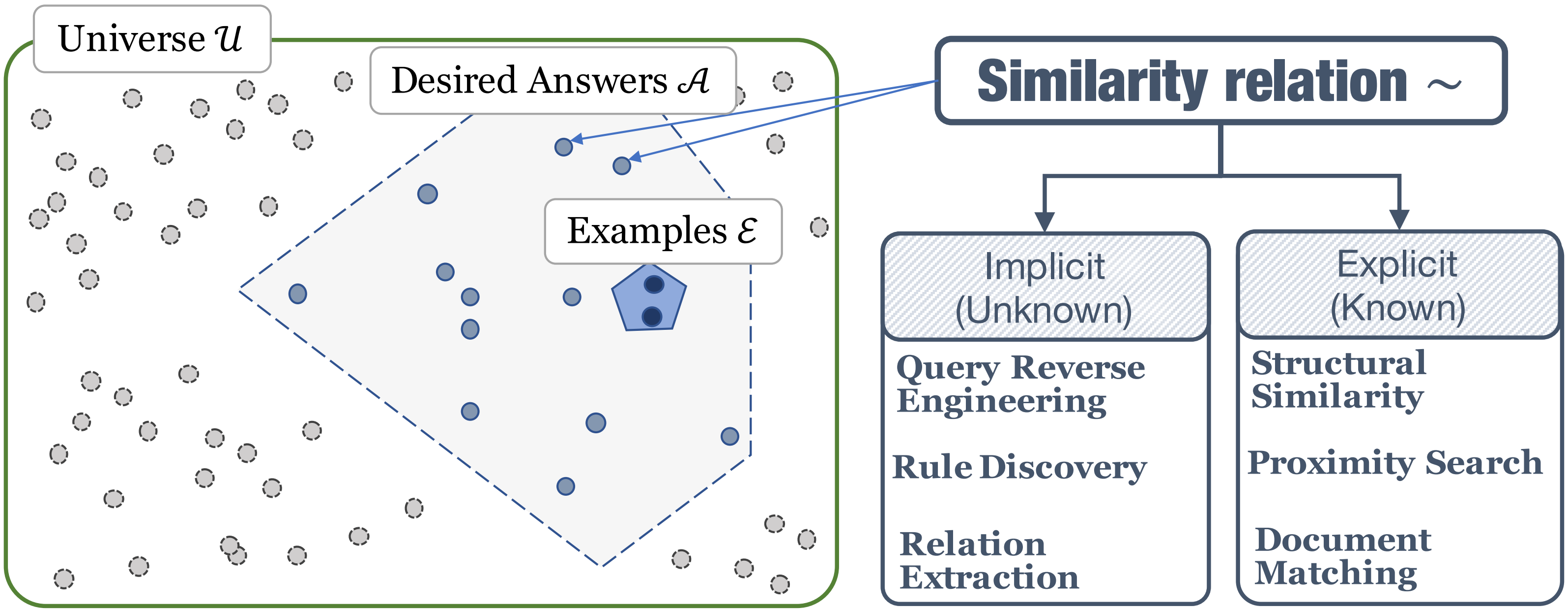
From Exact Search
based on explicit conditions

to Exploratory Search
based on Implicit needs

Similarities are the key …



Example-based methods: All You Need is …



Example-based methods

Relational

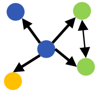
- Reverse engineering queries
- Example-driven schema mapping
- Interactive data repairing

Textual

- Search documents by example
- Entity extraction by example text
- Web table completion using examples

Graph

- Community-based Node-retrieval
- Entity Search
- Path and SPARQL queries
- Graph structures as Examples



Example-based methods: takeaways

Relational

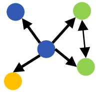
- **Complex search space**
- Exact and approximate
- Interactivity can improve the quality
- Limited to query inference

Textual

- **Allows serendipitous search**
- Easier document finding
- Speed up entity matching
- Extract semi-structure data

Graph

- **Heterogenous Structures**
- Exploit locality
- Entity attributes are expressive
- Large result-sets require ranking



The use of examples

Examples can ease data exploration

- … reduce need for complex queries / simplify user input
- … require no schema knowledge
- … allow uncertainity in search conditions
- … require little data analytics expertise

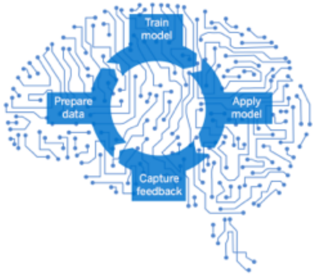
Acknowledgments

We would like to thank the authors of the papers who kindly provided us the slides and materials

Angela Bonifati, Radu Ciucianu, Marcelo Arenas, Gonzalo Diaz, Egor Kostylev, Yaacov Weiss, Sarah Cohen, Fotis Psallidas, Li Hao, Chan Chee Yong, Ilaria Bordino, Mohamed Yakout, Kris Ganjam, Kaushik Chakrabati, Thibault Sellam, Rohit Singh, Maeda Hanafi, Dmitri Kalashnikov, Marcin Sydow, Mingzhu Zhu, Yoshiharu Ishikawa, Daniel Deutch, Nandish Jayaram, Paolo Papotti, Bryan Perozzi, Kiriaki Dimitriadou, Yifei Ma, Natali Ruchansky, Quoc Trung Tran, Hastagiri Prakash Vanchinathan

… and many others (see references)

Where should we invest time?



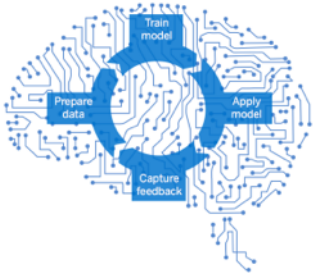
Machine learning

Approximate Methods

User models

Scalability

Where should we invest time?

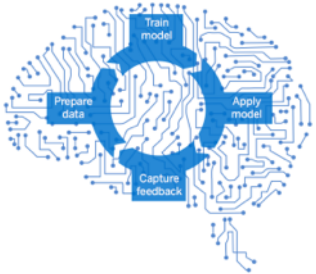


Machine learning

Learn from Examples

- … Similarity Measures: are often “fuzzy” and “implicit”
- … New representations of the search space
- Challenge: Scale! Exploration of large search spaces

Where should we invest time?



Machine learning → **User models**

Learn from Examples

- … Similarity Measures to represent User Interests
- … User-centric, dynamic, Exploration-strategies: learn as you go
- Challenge: Distinct User have Different Goals! Explore in different ways

We need more data!

Where should we invest time?

Scalability

Scale Example-based search

- … Huge search space, dynamic data, variety of data models
- … Exploration is Interactive, requires Interactive response time
- Adaptive Data-structures, localized access, flexible schema, incremental index

Where should we invest time?

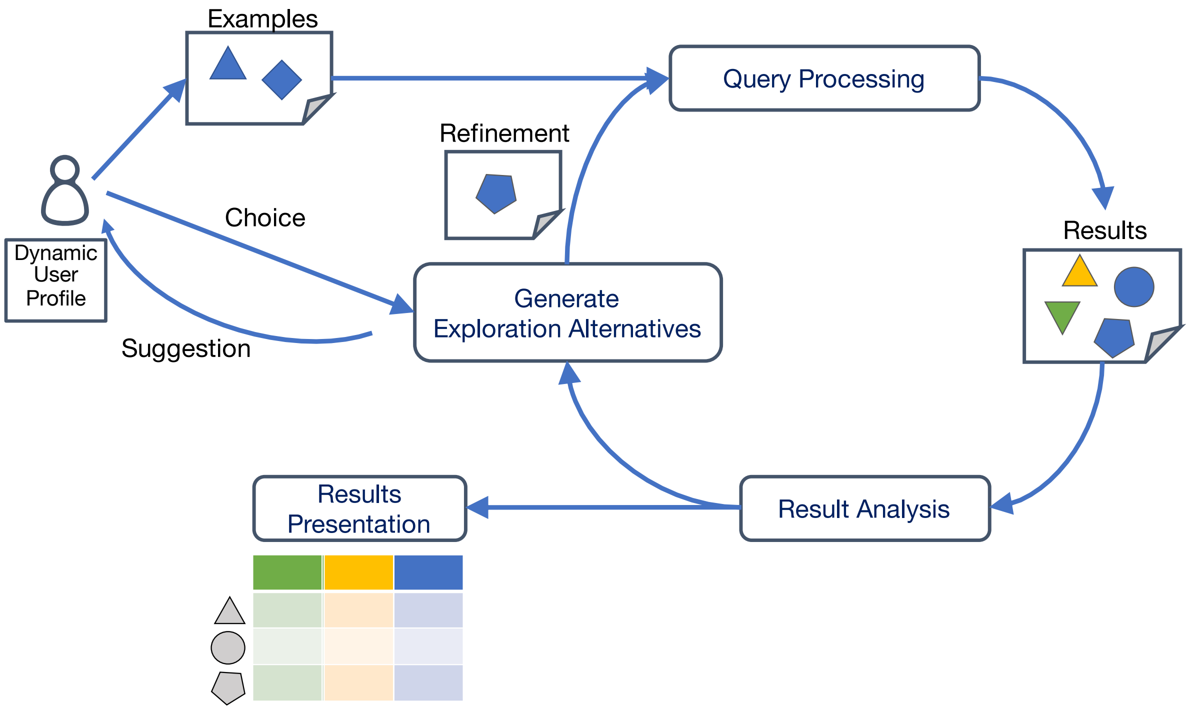
Scalability → Approximate Methods

Scale Example-based search

- … An approximate answer now is better than a precise answer in 1hour
- … Approximate answers can provide insights without being accurate

Exploratory queries retrieve large resultsets: the user needs only a glimpse to figure out if they are moving in the right direction!

Interactive Example Based Exploration System?



Requires:

Fast Query Processing

Avoid the full recomputation of a query

Limit the computation to only a sample

Adaptive query executions

Adaptive data-structures and indexes,

Automatic Result Analysis

Automatically identify peculiar characteristics,

Data-summarization techniques

Learn user interests automatically

ADOPT HETEROGENEITY

Need for solutions that
operate across different models

operate on heterogeneous datastores

dataset search

Data Lakes??
better
Semantic Data Lakes??

DEMOCRATIZATION
easy access to data

Tools that work on
commodity
hardware, mobile devices

Data-exploration for
everyday use-cases

Users want back
the control on their data

NATARUAL LANGUAGE INTERFACE

flexible, vague, imprecise input

Exploration through conversation

Understand the user Intention & context

Example is always more efficacious than precept

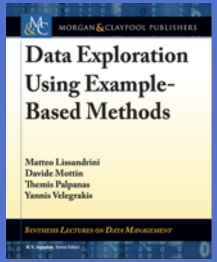
***Samuel Johnson**, Rasselas (1759), Chapter 29.*

"**New Trends on Exploratory Methods for Data Analytics.**" *PVLDB, 2017.*

"**Data Exploration Using Example-Based Methods.**" *M&C, 2018.*

"**Exploring the Data Wilderness through Examples.**" *SIGMOD, 2019.*

"**Example-driven Search: a New Frontier for Exploratory Search.**" *SIGIR 2019*



Slides: https://data-exploration.ml/eswc2020.html

References

M. Arenas, G. I. Diaz, and E. V. Kostylev. Reverse engineering sparql queries. WWW, 2016.

Agichtein, E. and Gravano, L. Snowball: Extracting relations from large plain-text collections. ICDL, 2000.

A.Bonifati, R.Ciucanu,and A.Lemay. Learning path queries on graph databases. EDBT, 2015.

A. Bonifati, R. Ciucanu, and S. Staworko. Learning join queries from user examples. TODS, 2016.

A. Bonifati, U. Comignani, E. Coquery, and R. Thion. Interactive mapping specification with exemplar tuples. SIGMOD, 2017.

I. Bordino, G. De Francisci Morales, I. Weber, and F. Bonchi. From machu picchu to rafting the urubamba river: anticipating information needs via the entity-query graph. WSDM, 2013.

D. Deutch and A. Gilad. Qplain: Query by explanation. ICDE, 2016.

References

G. Diaz, M. Arenas, and M. Benedikt. Sparqlbye: Querying rdf data by example. PVLDB, 2016.

K. Dimitriadou, O. Papaemmanouil, and Y. Diao. Explore-by-example: An automatic query steering framework for interactive data exploration. In SIGMOD, 2014.

B. Eravci and H. Ferhatosmanoglu. Diversity based relevance feedback for time series search. PVLDB, 2013.

A. Gionis, M. Mathioudakis, and A. Ukkonen. Bump hunting in the dark: Local discrepancy maximization on graphs. ICDE, 2015.

M. F. Hanafi, A. Abouzied, L. Chiticariu, and Y. Li. Synthesizing extraction rules from user examples with seer. SIGMOD, 2017.

He, J., Veltri, E., Santoro, D., Li, G., Mecca, G., Papotti, P. and Tang, N. Interactive and deterministic data cleaning. SIGMOD, 2016.

Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Querying databases through multiple examples. VLDB, 1998.

N. Jayaram, A. Khan, C. Li, X. Yan, and R. Elmasri. Querying knowledge graphs by example entity tuples.

TKDE, 2015.

References

H. Li, C.-Y. Chan, and D. Maier. Query from examples: An iterative, data-driven approach to query construction. PVLDB, 2015.

M. Lissandrini, D. Mottin, Y. Velegrakis, T. Palpanas. Multi-Example Search in Rich Information Graphs ICDE 2018

M. Lissandrini, D. Mottin, T. Palpanas, Y. Velegrakis. Graph-Query Suggestions for Knowledge Graph Exploration. WWW 2020

Y. Ma, T.-K. Huang, and J. G. Schneider. Active search and bandits on graphs using sigma-optimality. UAI, 2015.

S. Metzger, R. Schenkel, and M. Sydow. Qbees: query by entity examples. CIKM, 2013.

D. Mottin, M. Lissandrini, Y. Velegrakis, and T. Palpanas. Searching with xq: the exemplar query search engine. SIGMOD, 2014.

D. Mottin, M. Lissandrini, Y. Velegrakis, and T. Palpanas. Exemplar queries: a new way of searching. VLDB J., 2016.

B. Perozzi, L. Akoglu, P. Iglesias Sanchez, and E. Müller. Focused clustering and outlier detection in large attributed graphs. KDD, 2014.

References

F. Psallidas, B. Ding, K. Chakrabarti, and S. Chaudhuri. S4: Top-k spreadsheet-style search for query discovery. SIGMOD, 2015.

R. Rolim, G. Soares, L. D'Antoni, O. Polozov, S. Gulwani, R. Gheyi, R. Suzuki, and B. Hartmann. Learning syntactic program transformations from examples. ICSE, 2017.

N. Ruchansky, F. Bonchi, D. García-Soriano, F. Gullo, and N. Kourtellis. The minimum wiener connector problem. SIGMOD, 2015.

T. Sellam and M. Kersten. Cluster-driven navigation of the query space. TKDE, 2016.

Y. Shen, K. Chakrabarti, S. Chaudhuri, B. Ding, and L. Novik. Discovering queries based on example tuples. SIGMOD, 2014.

R. Singh. Blinkfill: Semi-supervised programming by example for syntactic string transformations. PVLDB, 2016.

G. Sobczak, M. Chochół, R. Schenkel, and M. Sydow. iqbees: Towards interactive semantic entity search based on maximal aspects. Foundations of Intelligent Systems, 2015.

References

Y. Su, S. Yang, H. Sun, M. Srivatsa, S. Kase, M. Vanni, and X. Yan. Exploiting relevance feedback in knowledge graph search. KDD, 2015.

Q. T. Tran, C.-Y. Chan, and S. Parthasarathy. Query reverse engineering. VLDB J., 2014.

H. P. Vanchinathan, A. Marfurt, C.-A. Robelin, D. Koss- mann, and A. Krause. Discovering valuable items from massive data. In KDD, 2015.

C.Wang, A.Cheung, and R.Bodik. Interactive query synthesis from input-output examples. In SIGMOD, 2017.

C. Wang, A. Cheung, and R. Bodik. Synthesizing highly expressive sql queries from input-output examples. In PLDI, 2017.

Y. Y. Weiss and S. Cohen. Reverse engineering spj-queries from examples. SIGMOD, 2017.

M. Yakout, K. Ganjam, K. Chakrabarti, and S. Chaudhuri. Infogather: Entity augmentation and attribute discovery by holistic matching with web tables. SIGMOD, 2012.

M. Zhu and Y.-F. B. Wu. Search by multiple examples. WSDM, 2014.

M. M. Zloof. Query by example. AFIPS NCC, 1975.