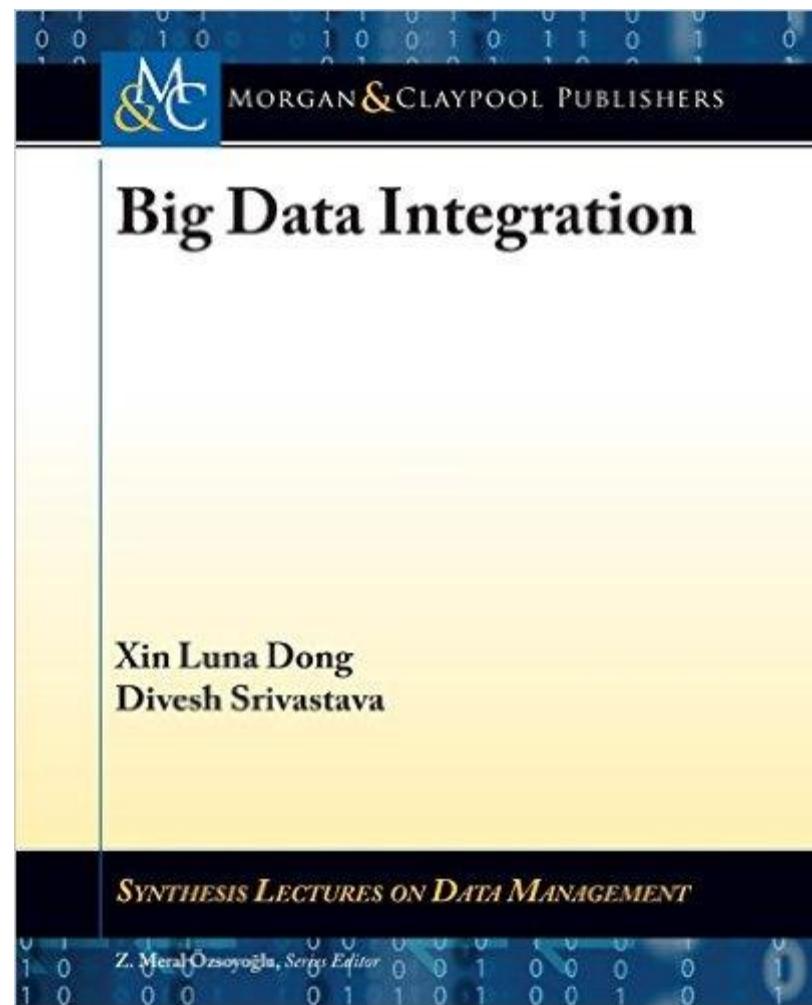


Big Data Integration

Xin Luna Dong (Amazon)

Divesh Srivastava (AT&T Labs-Research)

A Shameless Plug



What is “Big Data Integration?”

- ◆ Big data integration = Big data + data integration
- ◆ Data integration: easy access to multiple data sources [DHI12]
 - Virtual: mediated schema, query reformulation, link + fuse answers
 - Warehouse: materialized data, easy querying, consistency issues
- ◆ Big data: all about the V's ☺
 - Size: large **volume** of data, collected and analyzed at high **velocity**
 - Complexity: huge **variety** of data, of questionable **veracity**
 - Utility: data of considerable **value**

What is “Big Data Integration?”

- ◆ Big data integration = Big data + data integration
- ◆ Data integration: easy access to multiple data sources [DHI12]
 - Virtual: mediated schema, query reformulation, link + fuse answers
 - Warehouse: materialized data, easy querying, consistency issues
- ◆ Big data in the context of data integration: still about the V's ☺
 - Size: large **volume** of sources, changing at high **velocity**
 - Complexity: huge **variety** of sources, of questionable **veracity**
 - Utility: sources of considerable **value**

Outline

◆ Motivation

- Why do we need big data integration?
- How has “small” data integration been done?
- Challenges in big data integration

Why Do We Need “Big Data Integration?”

- ◆ Building web-scale knowledge bases (KB)



Google knowledge graph



NELL: Never-Ending Language Learning

ProBase



DBpedia



Walmart

Using KB in Search

koko restaurant crown melbourne  +Divesh

Web Maps Shopping Images News More ▾ Search tools

About 86,800 results (0.51 seconds)

Koko Melbourne - Japanese Restaurant | Crown Melbourne
<https://www.crownmelbourne.com.au> ... ▾ Crown Casino and Entertainment Complex ▾
Experience Koko, traditional Japanese restaurant and five star service at Crown Melbourne. Explore Melbourne's best restaurants and book online.
4.5 ★★★★★ 21 Google reviews · Write a review · Google+ page

📍 Level 3, Crown Towers, Crown Entertainment Complex, 8 Whiteman Street, Southbank VIC 3006, Australia
+61 3 9292 5777

Koko - Southbank | Urbanspoon
www.urbanspoon.com › Melbourne › City › Southbank ▾ Urbanspoon ▾
★★★★★ Rating: 83% - 911 votes
I was lucky enough to tag along to an event hosted by Crown Casino at Koko. ... at Koko, one of Melbourne's premier Japanese restaurants located at Crown ...

Koko Japanese Restaurant, Melbourne - TripAdvisor
www.tripadvisor.com.au › ... › Melbourne Restaurants ▾ TripAdvisor LLC ▾
★★★★★ Rating: 4 - 119 reviews
My partner and I had a lovely teppanyaki meal at Koko restaurant in the Crown Melbourne. We were very impressed with the quality and quantity of food.



See photos

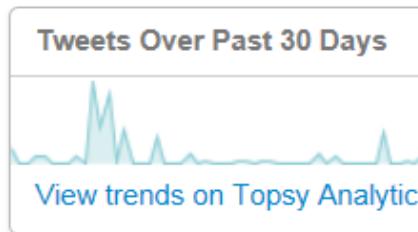


Immigration Museum
SEA LIFE Melbourne Aquarium
Yarra River
Southbank
Crown Melbourne

Koko 

4.5 ★★★★★ 21 Google reviews
Japanese Restaurant
Wood-lined restaurant with indoor water garden, serving traditional Japanese cuisine and aged sake.
Address: Level 3, Crown Towers, Crown Entertainment Complex, 8 Whiteman Street, Southbank VIC 3006, Australia
Phone: +61 3 9292 5777
Hours: Closed now · Hours

Using KB in Social Media



Turing Award



The ACM A.M. Turing Award is an annual prize given by the Association for Computing Machinery to "an individual selected for contributions of a technical nature made to the computing community". [Wikipedia](#)

Ceremony date (2015): June 20, 2015

People also search for: [Fields Medal](#), [Abel Prize](#), [National Medal of Technology and Innovation](#)

H1GU1a



Alan Turing

Computer Scientist

Alan Mathison Turing, OBE, FRS was a British pioneering computer scientist, mathematician, logician, cryptanalyst, philosopher, mathematical biologist, and marathon and ultra distance runner. [Wikipedia](#)

Born: June 23, 1912, Maida Vale, London, United Kingdom

Died: June 7, 1954, Wilmslow, United Kingdom

Education: Princeton University (1936–1938), [More](#)

Parents: Julius Mathison Turing, Ethel Sara Stoney

Siblings: John Turing

A.M. Turing Award

☆ Favorite  24 more

ng contributor to da
greatest-contributor-dat

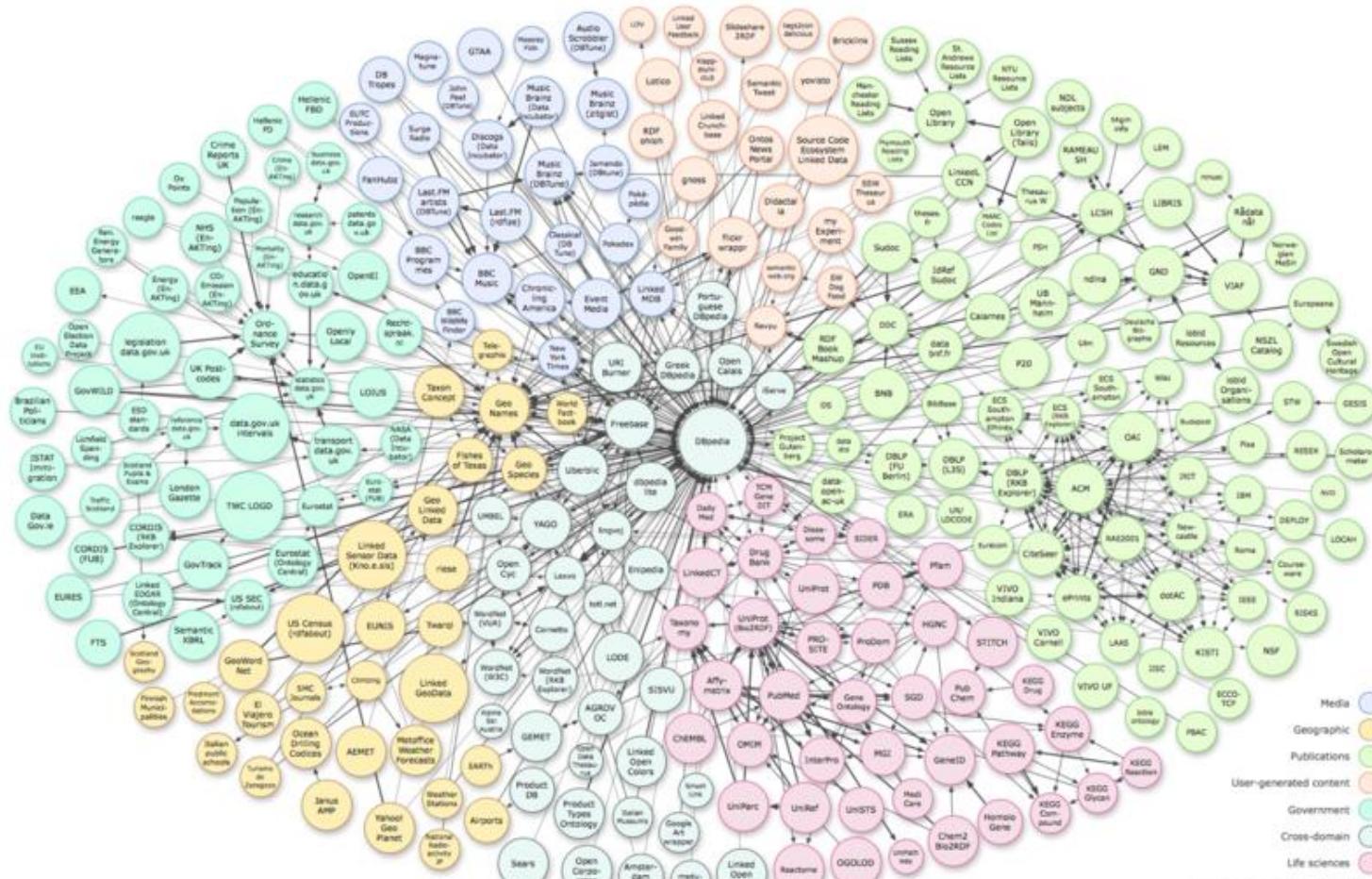
onebraker, greatest living contributor to database technology bit.ly/1IT4dN1 #Analytics
☆ Favorite  44 more

atest living contributor to database technology bit.ly/1IT4dN1 #Analytics

☆ Favorite  4 more

Why Do We Need “Big Data Integration?”

- ◆ Reasoning over linked data

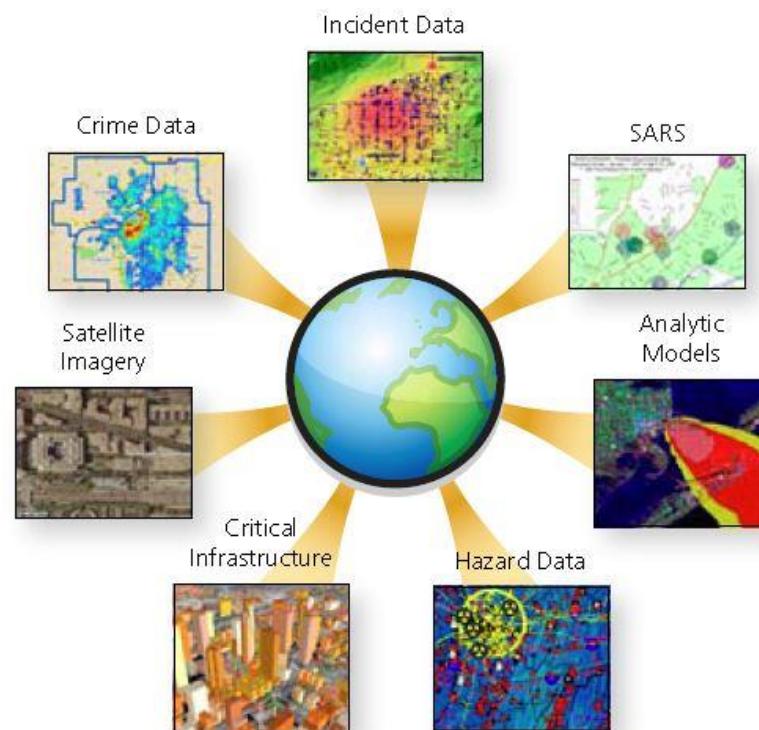


As of September 2011



Why Do We Need “Big Data Integration?”

- ◆ Geo-spatial data fusion

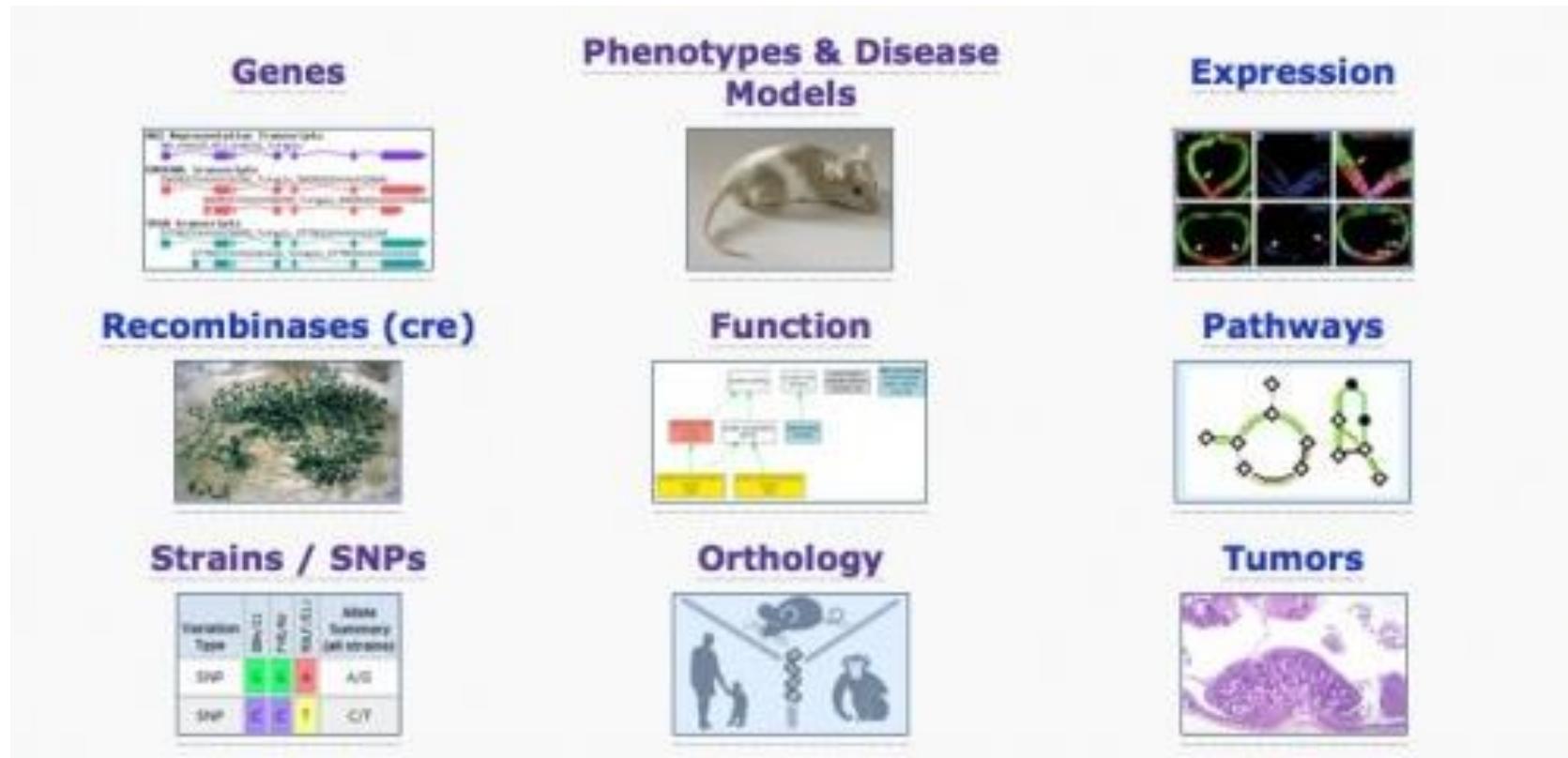


Geospatial Data Fusion

<http://axiomamuse.wordpress.com/2011/04/18/>

Why Do We Need “Big Data Integration?”

- ◆ Scientific data analysis



<http://sciencline.org/2012/01/from-index-cards-to-information-overload/>

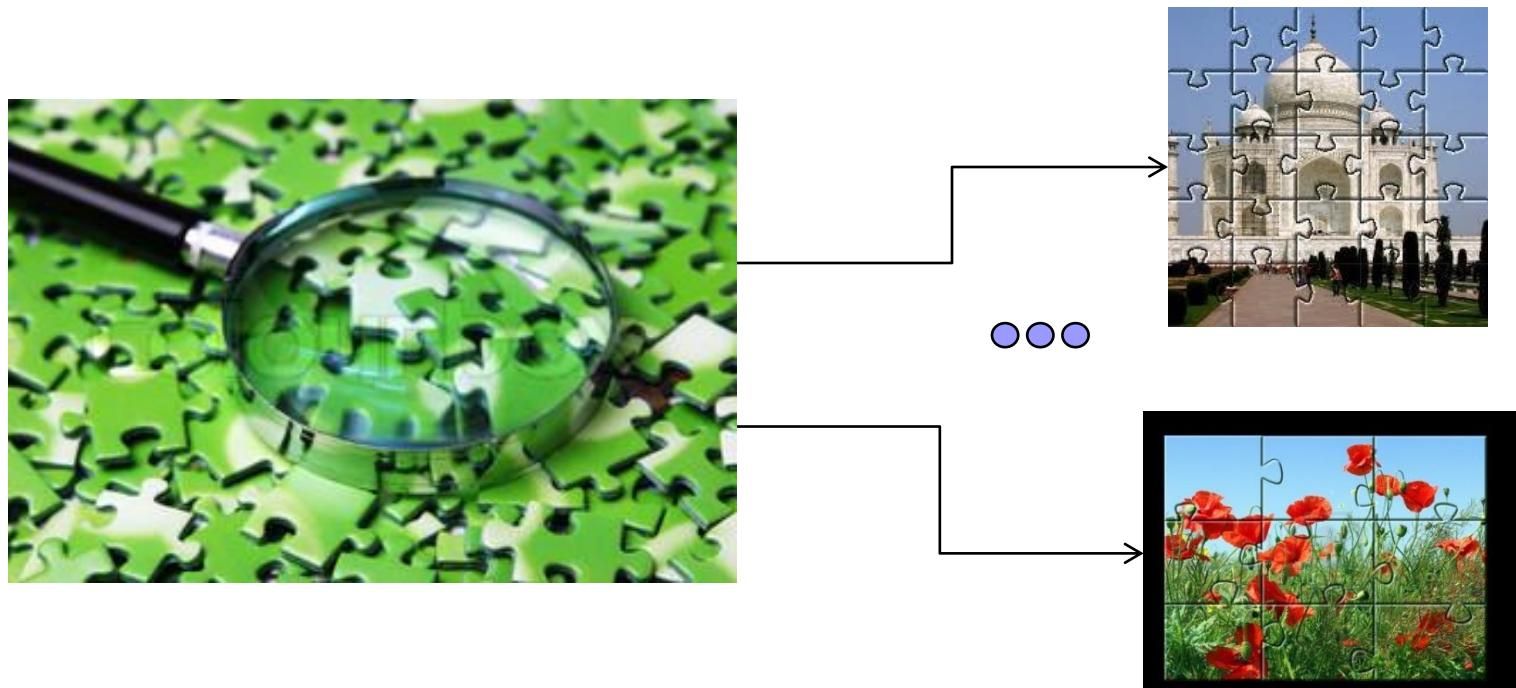
Outline

◆ Motivation

- Why do we need big data integration?
- How has “small” data integration been done?
- Challenges in big data integration

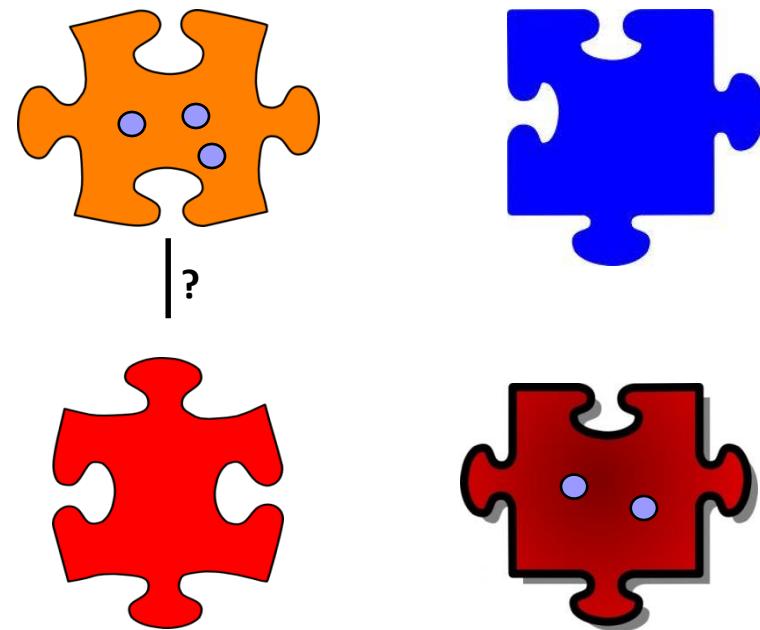
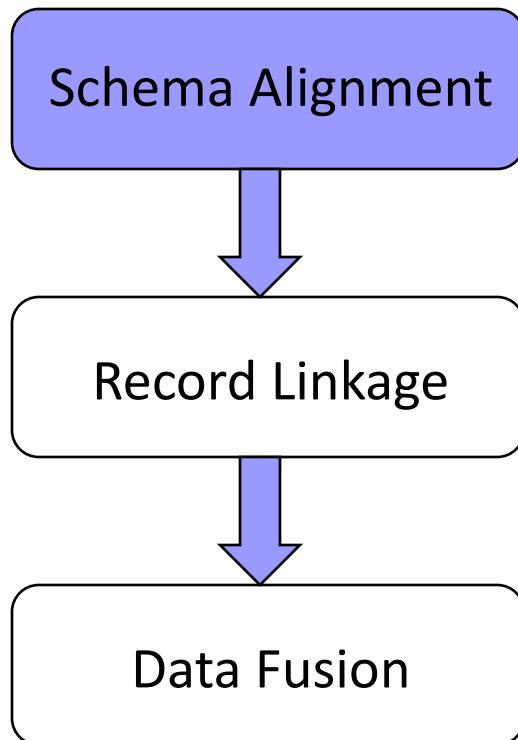
“Small” Data Integration: What Is It?

- ◆ Data integration = solving lots of jigsaw puzzles
 - Each jigsaw puzzle (e.g., Taj Mahal) is an **integrated entity**
 - Each piece of a puzzle comes from some **source**
 - Small data integration → solving small puzzles



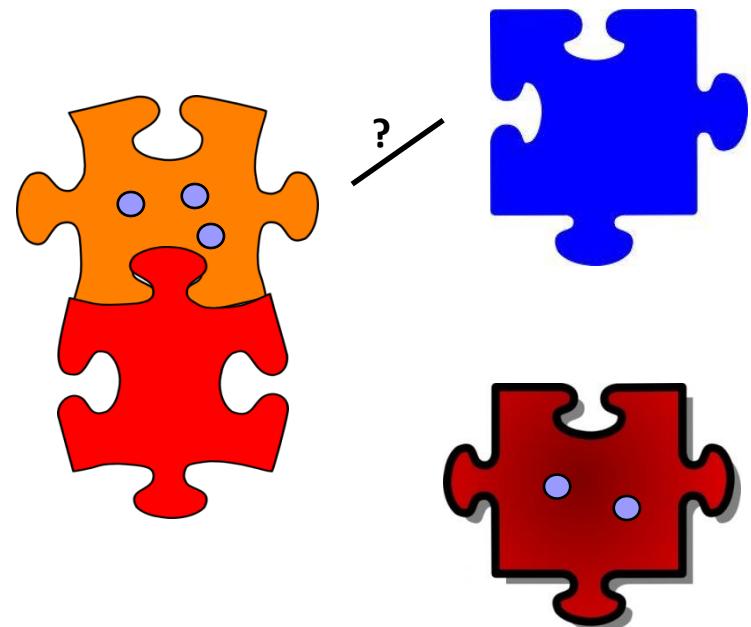
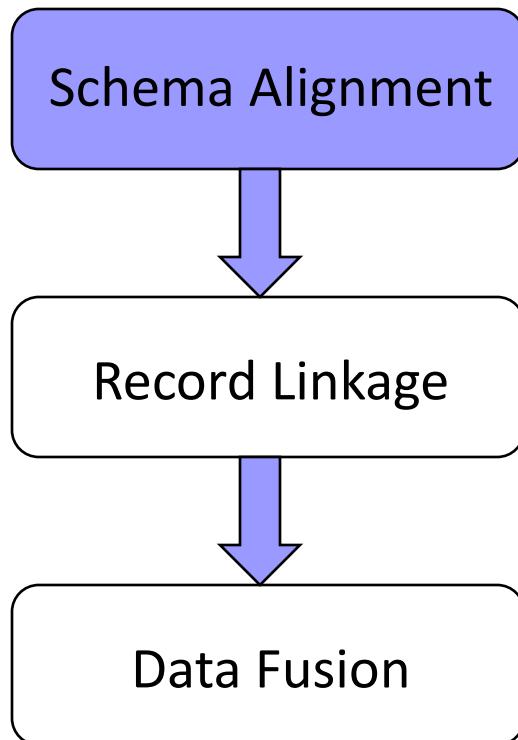
“Small” Data Integration: How is it Done?

- ◆ “Small” data integration: alignment + linkage + fusion
 - Schema alignment: mapping of **structure** (e.g., shape)



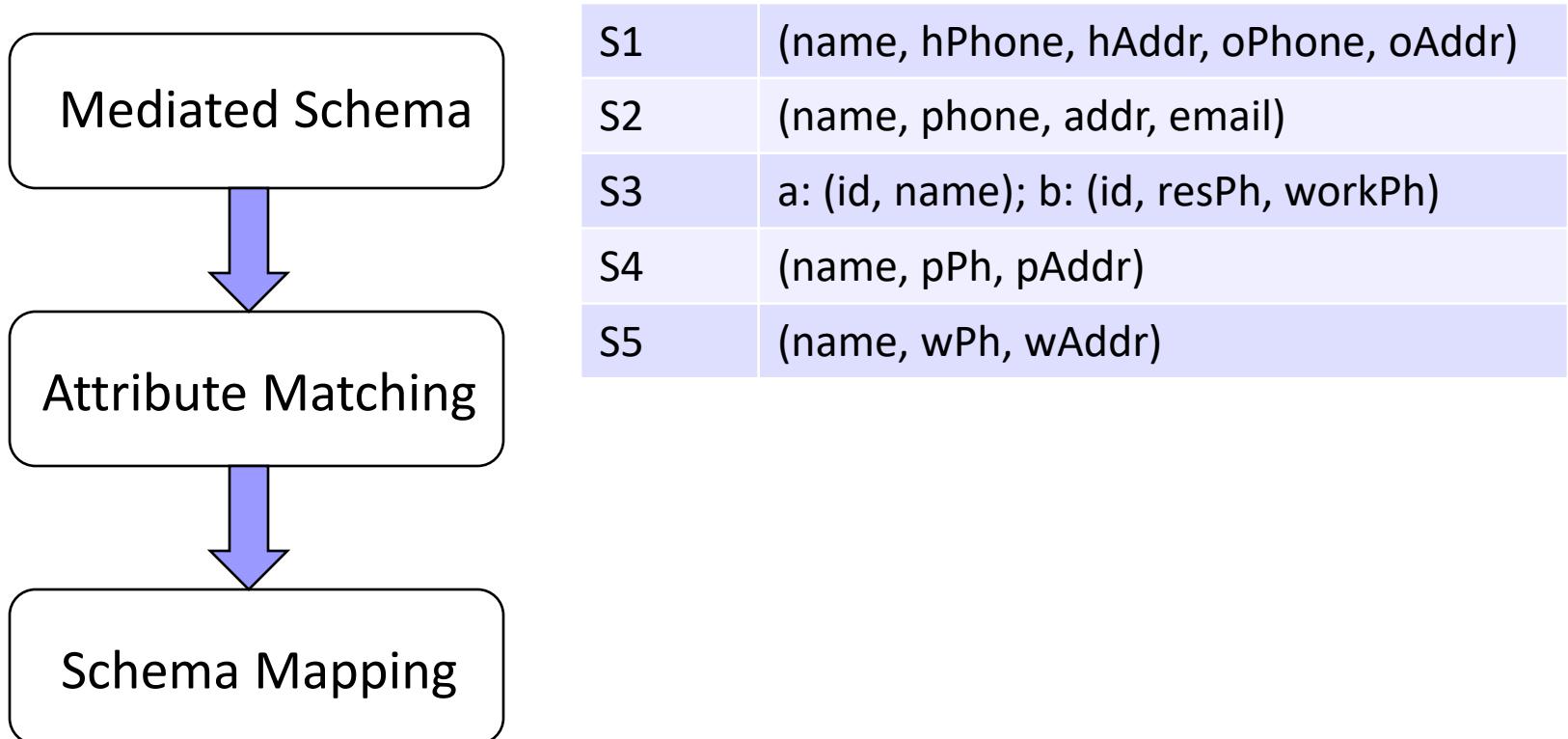
“Small” Data Integration: How is it Done?

- ◆ “Small” data integration: alignment + linkage + fusion
 - Schema alignment: mapping of **structure** (e.g., shape)



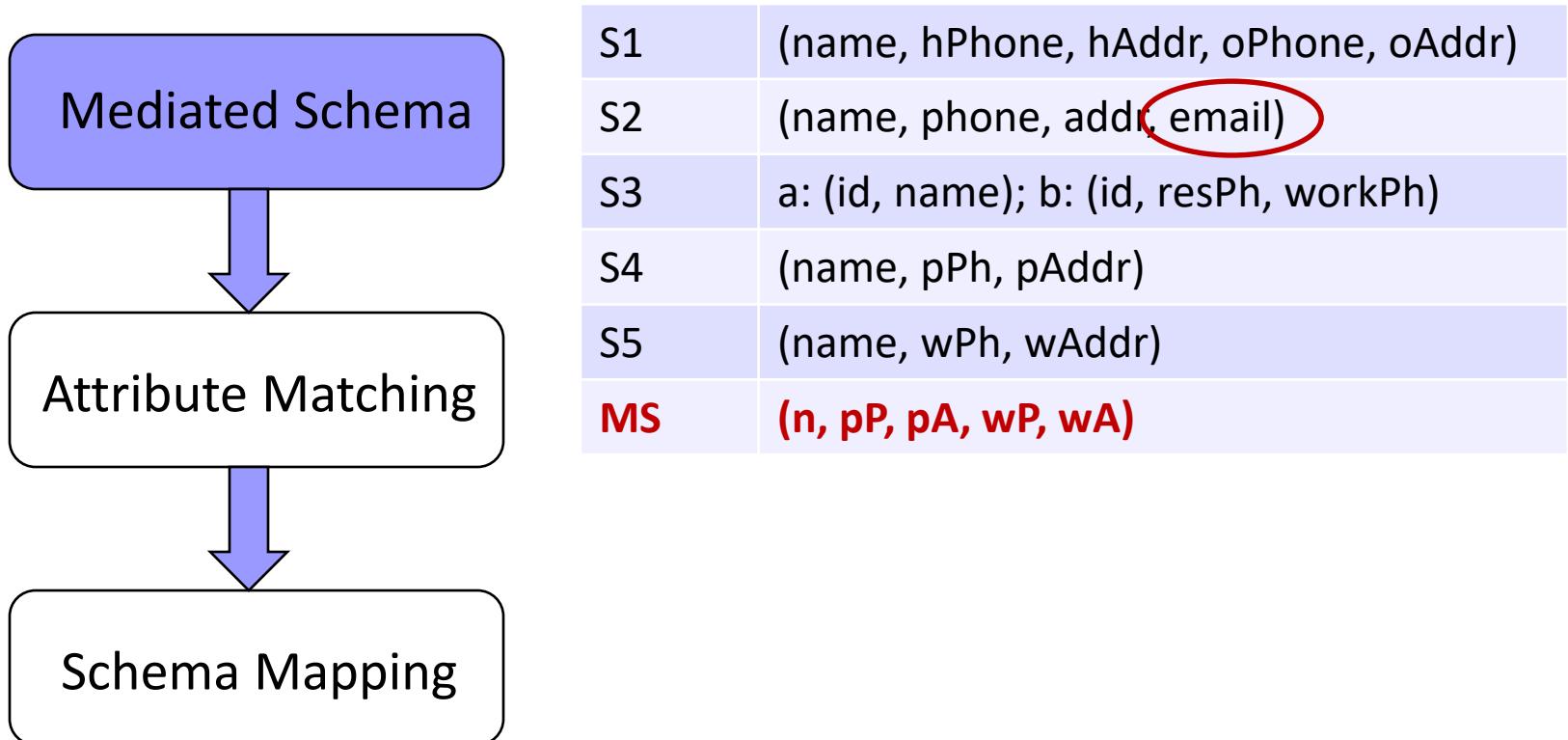
Schema Alignment: Three Steps [BBR II]

- ◆ Schema alignment: mediated schema + matching + mapping
 - Enables linkage, fusion to be semantically meaningful



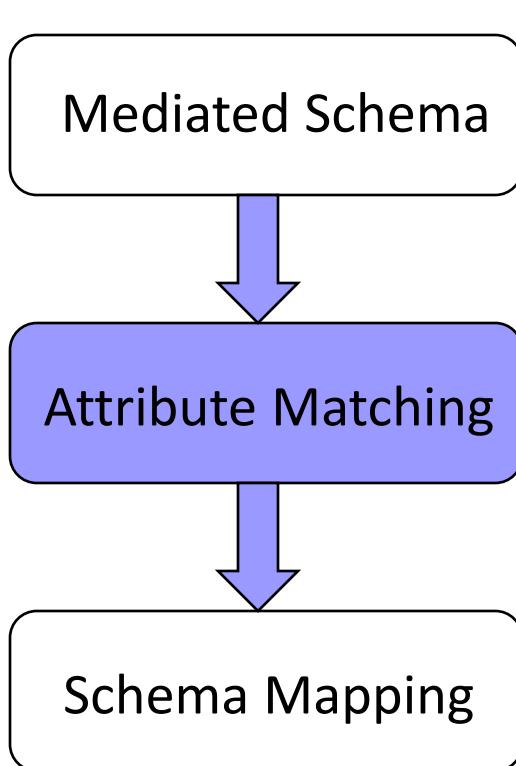
Schema Alignment: Three Steps

- ◆ Schema alignment: mediated schema + matching + mapping
 - Enables domain specific modeling



Schema Alignment: Three Steps

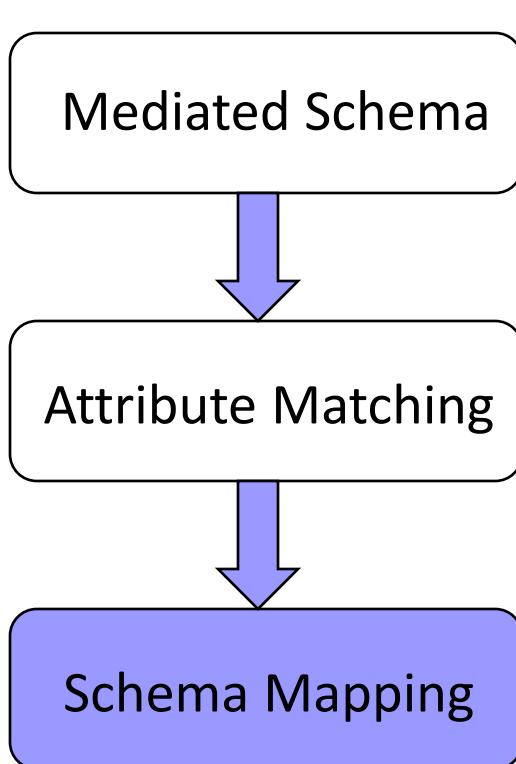
- ◆ Schema alignment: mediated schema + matching + mapping
 - Identifies correspondences between schema attributes



| | |
|-------------|--|
| S1 | (name, hPhone, hAddr, oPhone, oAddr) |
| S2 | (name, phone, addr, email) |
| S3 | a: (id, name); b: (id, resPh, workPh) |
| S4 | (name, pPh, pAddr) |
| S5 | (name, wPh, wAddr) |
| MS | (n, pP, pA, wP, wA) |
| MSAM | MS.n : S1.name, S2.name, S3a.name, ... MS.pP : S1.hPhone, S3b.resPh, S4.pPh MS.pA : S1.hAddr, S4.pAddr MS.wP : S1.oPhone, S2.phone, ... MS.wA : S1.oAddr, S2.addr, S5.wAddr |

Schema Alignment: Three Steps

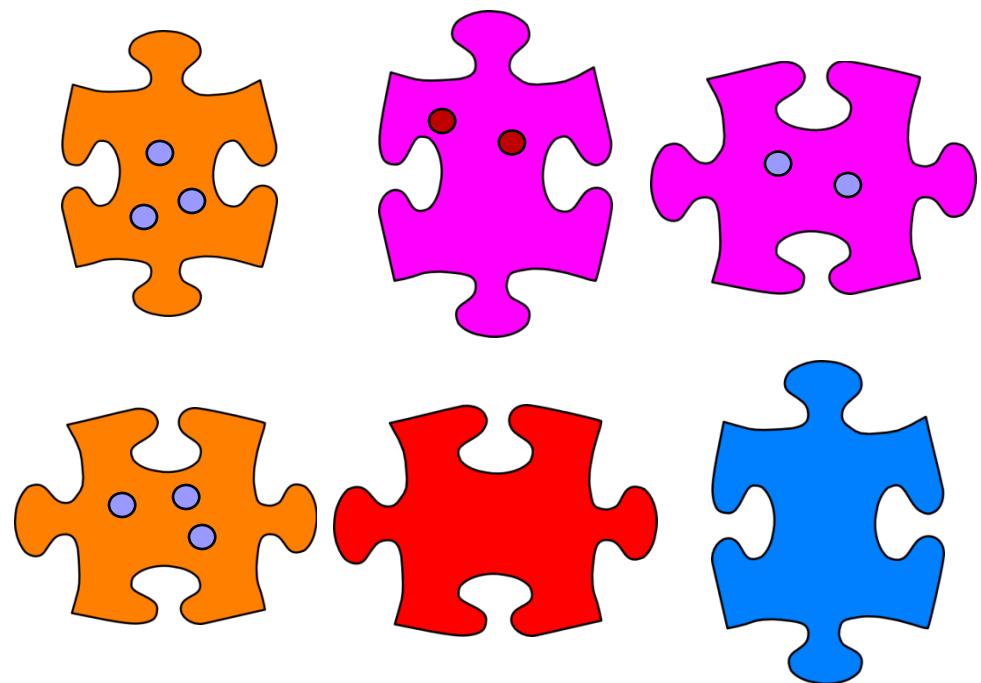
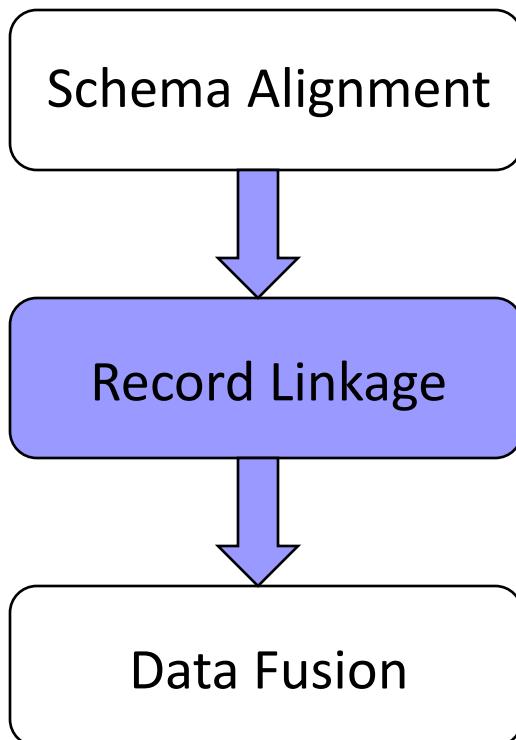
- ◆ Schema alignment: mediated schema + matching + mapping
 - Specifies transformation between records in different schemas



| | |
|-----------------------|---|
| S1 | (name, hPhone, hAddr, oPhone, oAddr) |
| S2 | (name, phone, addr, email) |
| S3 | a: (id, name); b: (id, resPh, workPh) |
| S4 | (name, pPh, pAddr) |
| S5 | (name, wPh, wAddr) |
| MS | (n, pP, pA, wP, wA) |
| MSSM (GAV) | MS(n, pP, pA, wP, wA) :- S1(n, pP, pA, wP, wA) MS(n, _, _, wP, wA) :- S2(n, wP, wA, e) MS(n, pP, _, wP, _) :- S3a(i, n), S3b(i, pP, wP) MS(n, pP, pA, _, _) :- S4(n, pP, pA) MS(n, _, _, wP, wA) :- S5(n, wP, wA) |

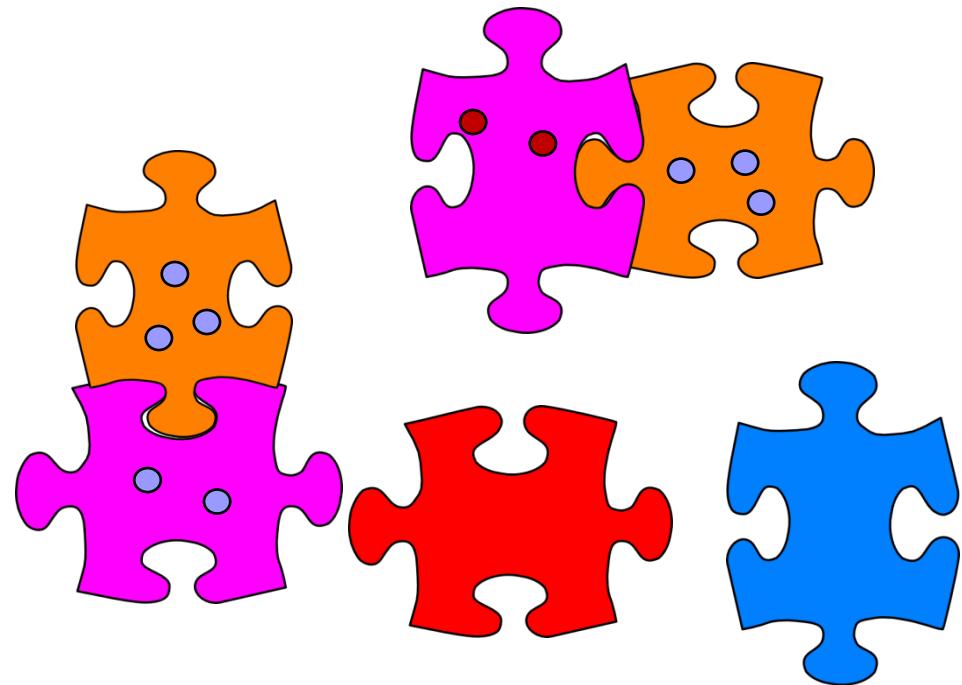
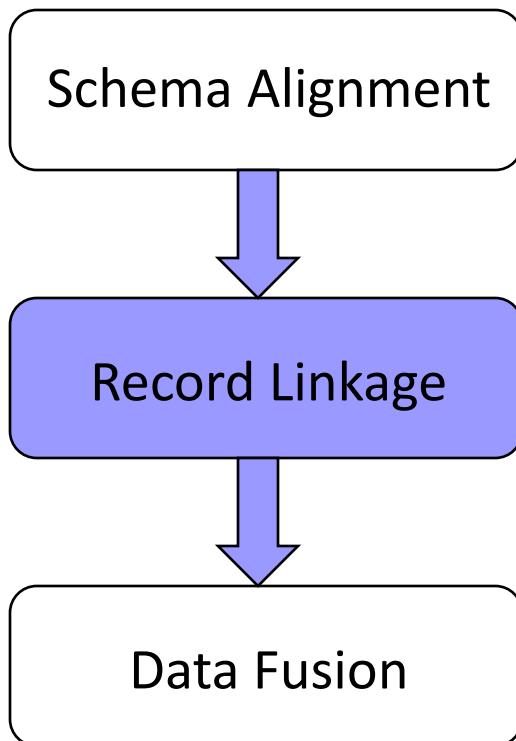
“Small” Data Integration: How is it Done?

- ◆ “Small” data integration: alignment + linkage + fusion
 - Record linkage: matching based on **content** (e.g., color, pattern)



“Small” Data Integration: How is it Done?

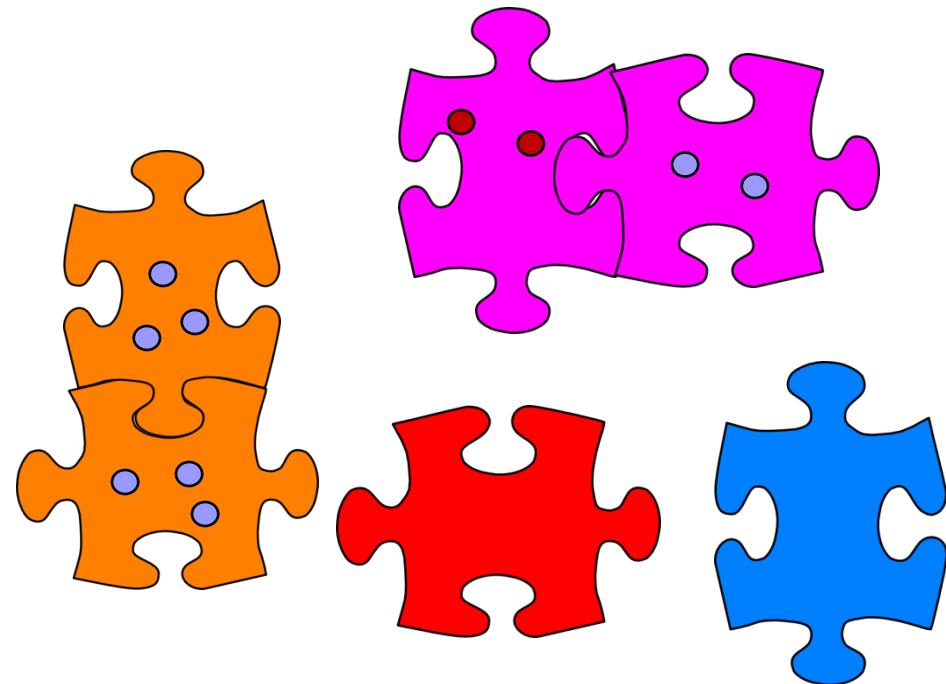
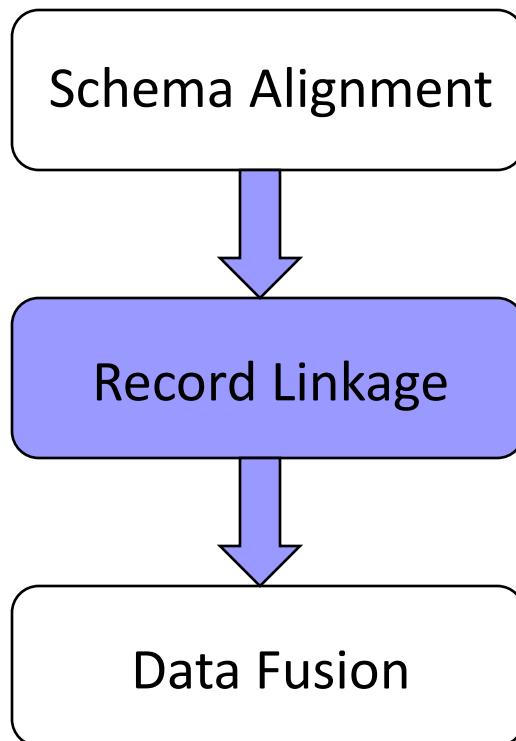
- ◆ “Small” data integration: alignment + linkage + fusion
 - Record linkage: matching based on **content** (e.g., color, pattern)





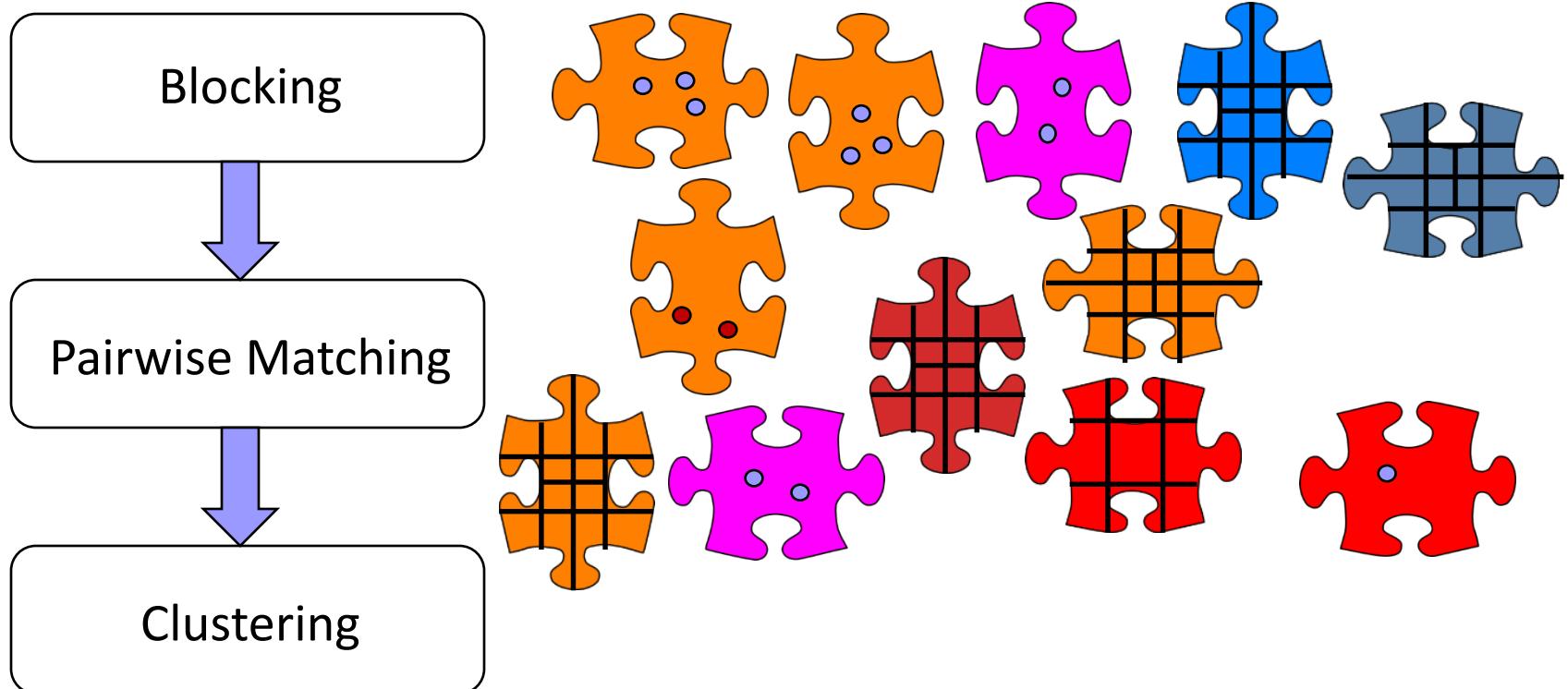
“Small” Data Integration: How is it Done?

- ◆ “Small” data integration: alignment + linkage + fusion
 - Record linkage: matching based on **content** (e.g., color, pattern)



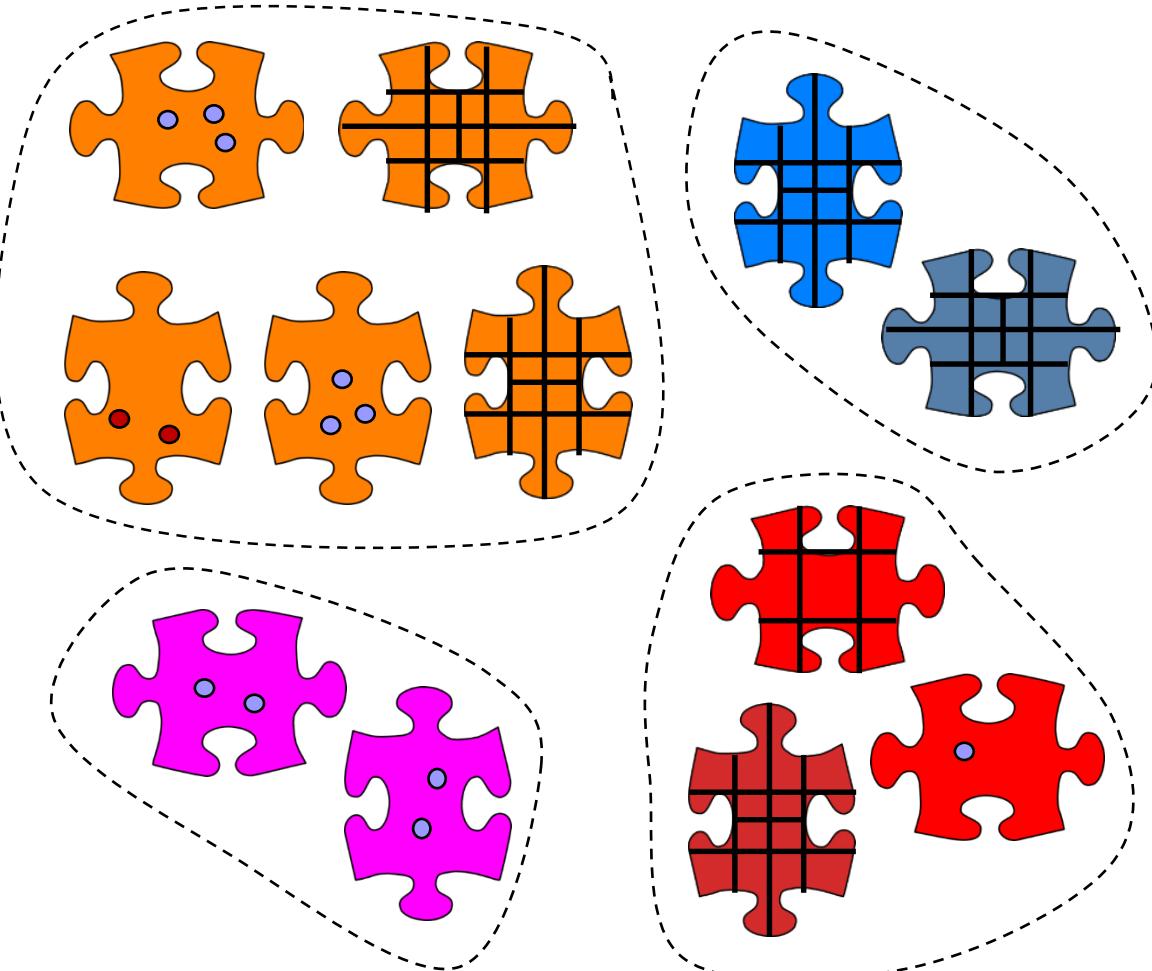
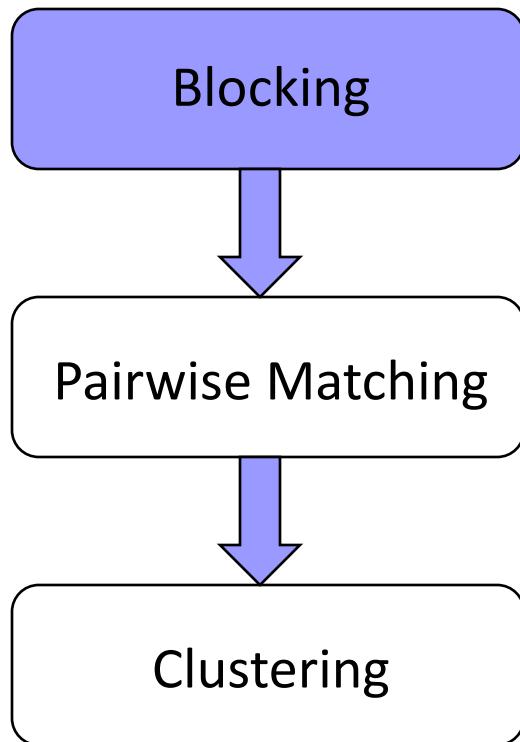
Record Linkage: Three Steps [EIV07, GM12]

- ◆ Record linkage: blocking + pairwise matching + clustering
 - Scalability, similarity, semantics



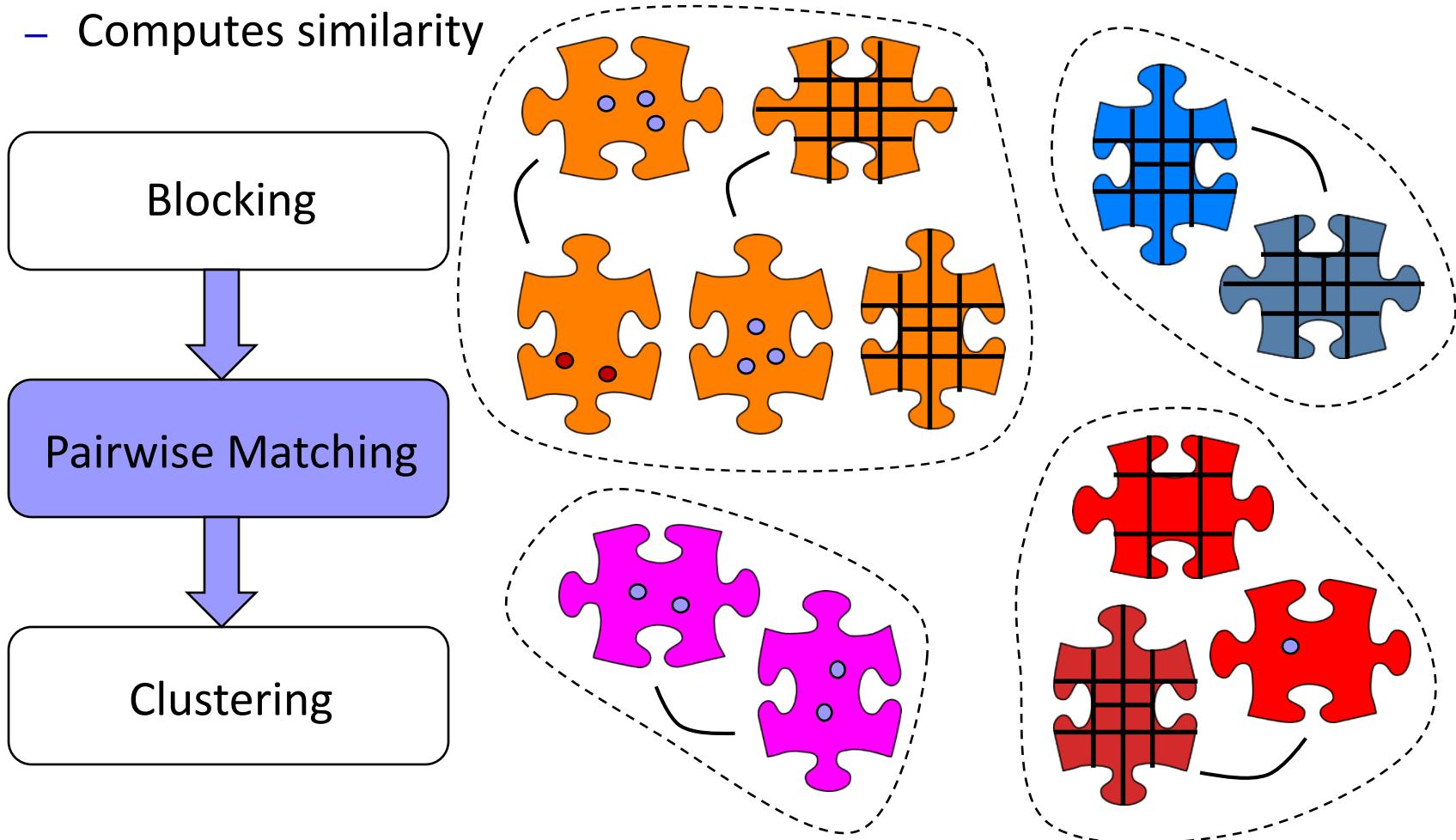
Record Linkage: Three Steps

- ◆ Blocking: **efficiently** create **small** blocks of **similar** records
 - Ensures scalability



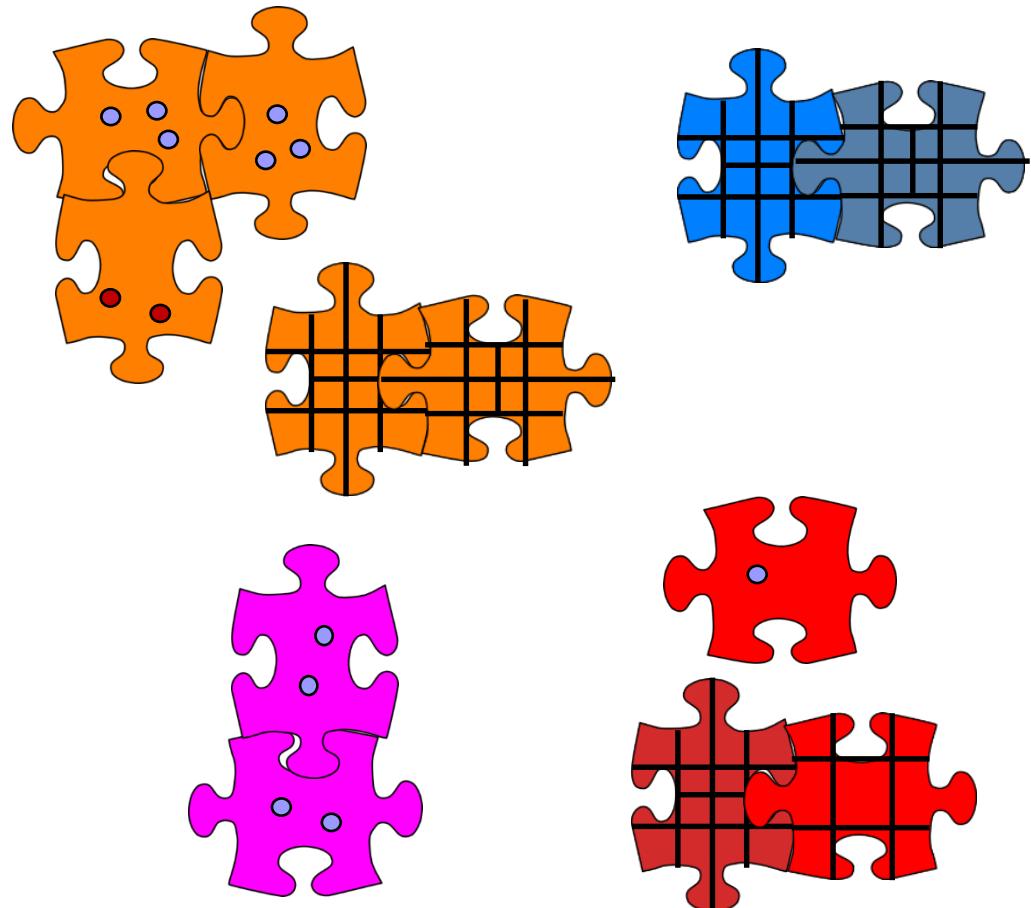
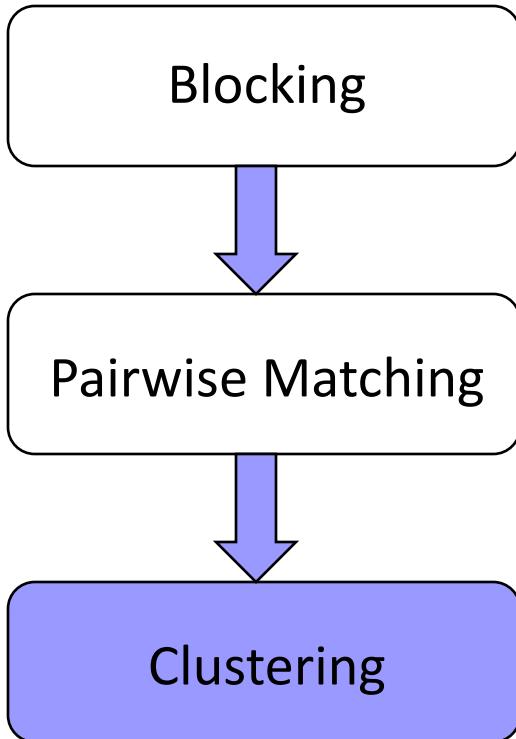
Record Linkage: Three Steps

- ◆ Pairwise matching: compares all record pairs in a block
 - Computes similarity



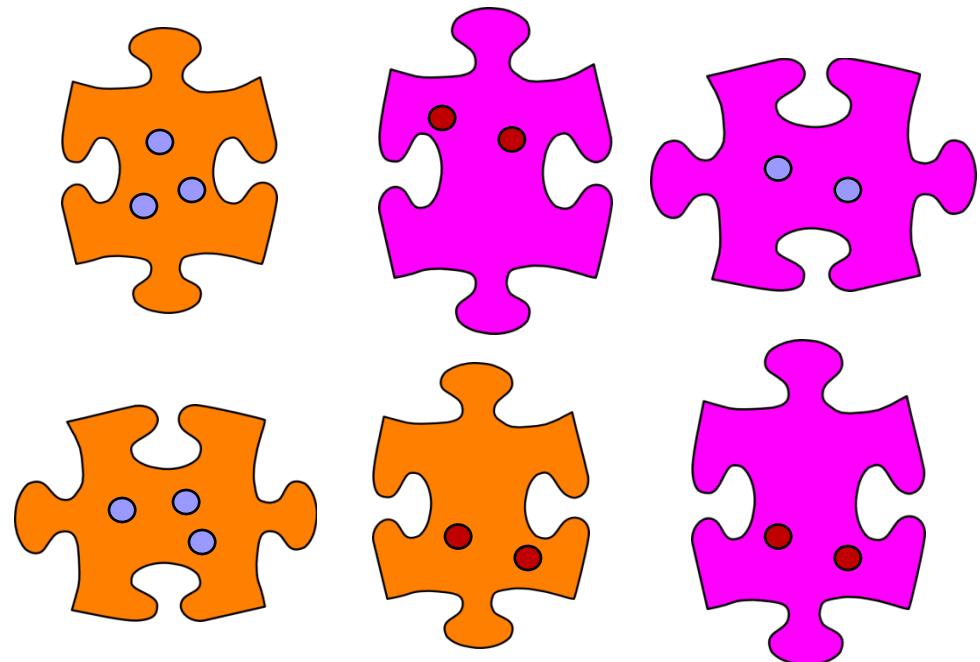
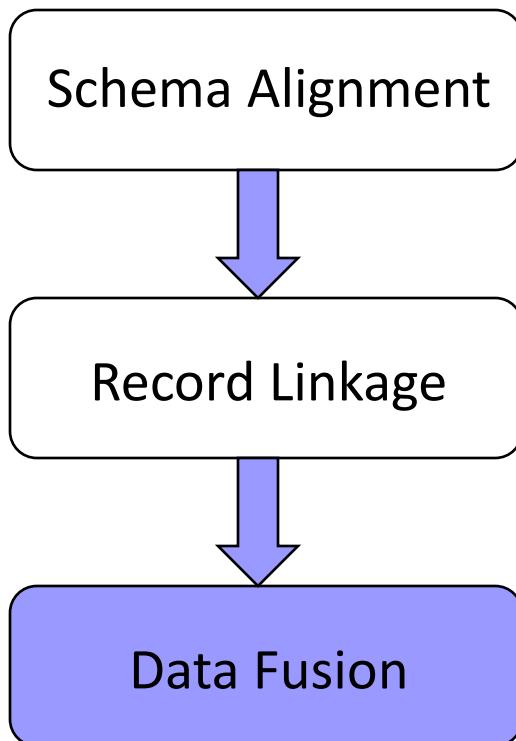
Record Linkage: Three Steps

- ◆ Clustering: groups sets of records into entities
 - Ensures semantics



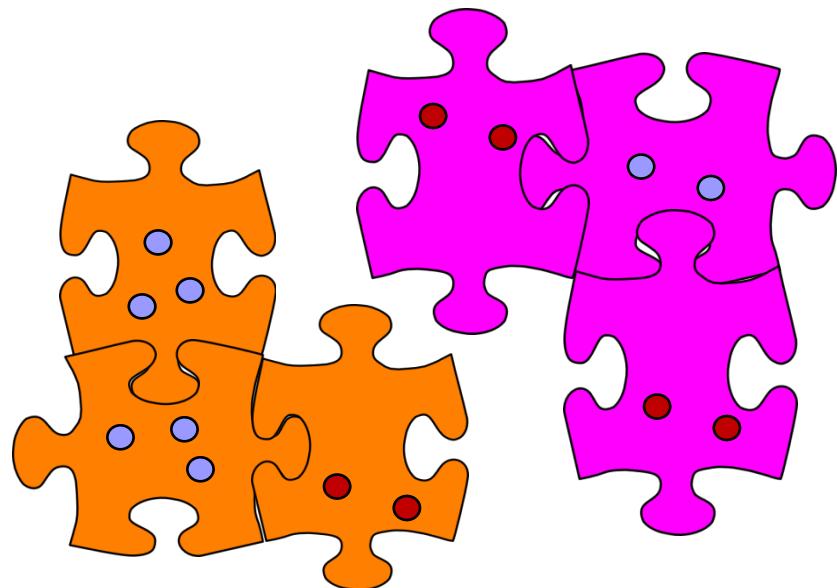
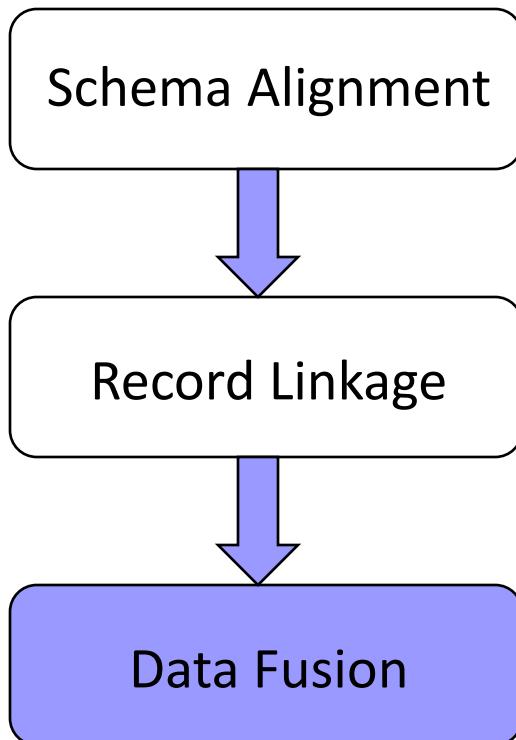
“Small” Data Integration: How is it Done?

- ◆ “Small” data integration: alignment + linkage + fusion
 - Data fusion: reconciliation of **mismatching content** (e.g., pattern)



“Small” Data Integration: How is it Done?

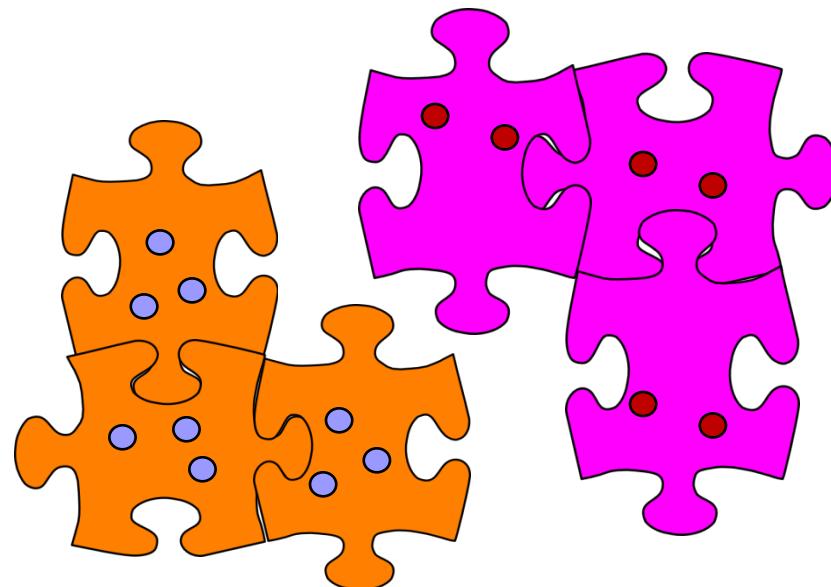
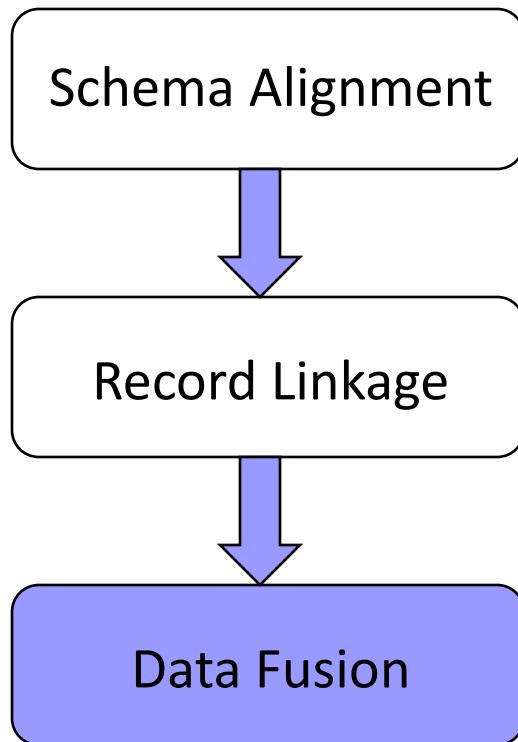
- ◆ “Small” data integration: alignment + linkage + fusion
 - Data fusion: reconciliation of **mismatching content** (e.g., pattern)





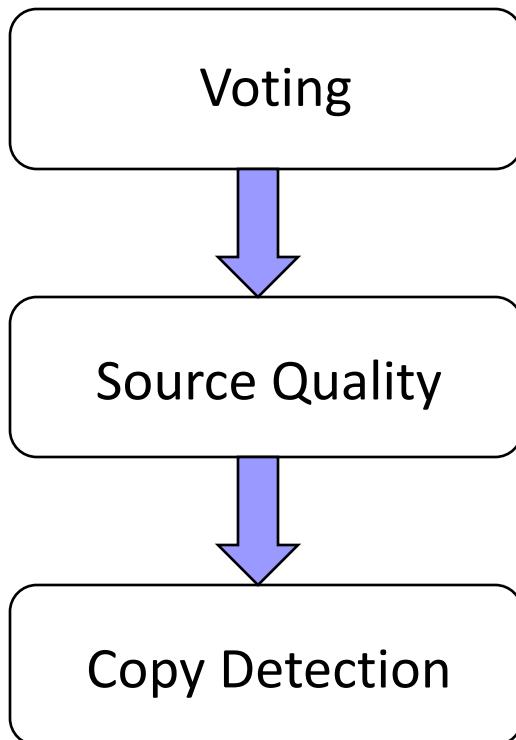
“Small” Data Integration: How is it Done?

- ◆ “Small” data integration: alignment + linkage + fusion
 - Data fusion: reconciliation of **mismatching content** (e.g., pattern)



Data Fusion: Three Components [DBS09a]

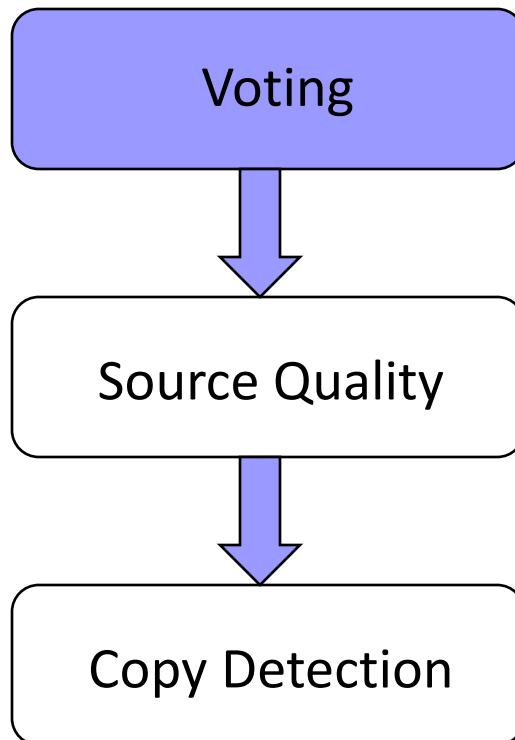
- ◆ Data fusion: voting + source quality + copy detection
 - Resolves inconsistency across diversity of sources



| | S1 | S2 | S3 | S4 | S5 |
|-----------|-----|------------|------------|------------|------------|
| Jagadish | UM | <u>ATT</u> | UM | UM | <u>UI</u> |
| Dewitt | MSR | MSR | <u>UW</u> | <u>UW</u> | <u>UW</u> |
| Bernstein | MSR | MSR | MSR | MSR | MSR |
| Carey | UCI | <u>ATT</u> | <u>BEA</u> | <u>BEA</u> | <u>BEA</u> |
| Franklin | UCB | UCB | <u>UMD</u> | <u>UMD</u> | <u>UMD</u> |

Data Fusion: Three Components [DBS09a]

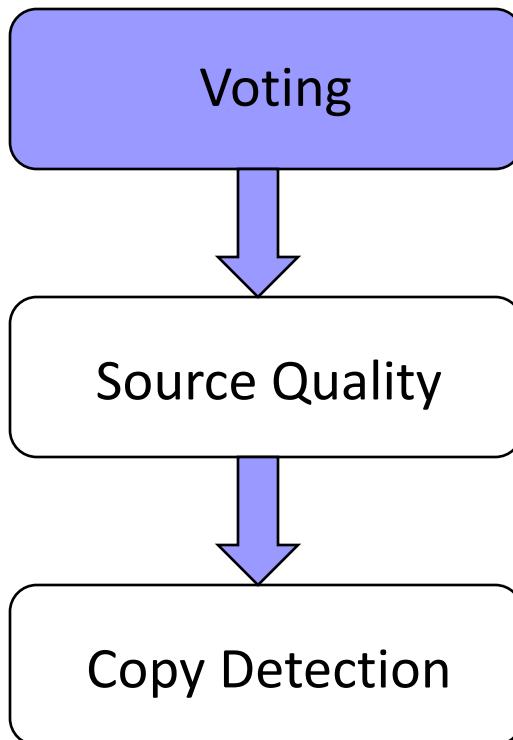
- ◆ Data fusion: voting + source quality + copy detection



| | S1 | S2 | S3 |
|-----------|-----|-----|-----|
| Jagadish | UM | ATT | UM |
| Dewitt | MSR | MSR | UW |
| Bernstein | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA |
| Franklin | UCB | UCB | UMD |

Data Fusion: Three Components [DBS09a]

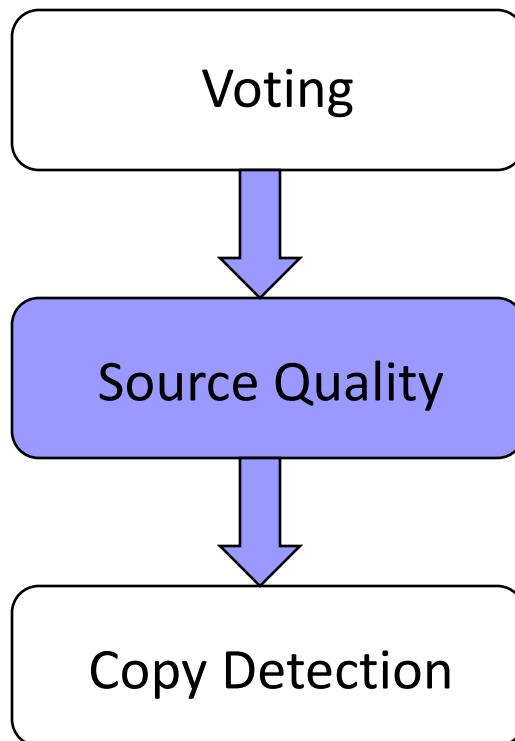
- ◆ Data fusion: voting + source quality + copy detection
 - Supports difference of opinion



| | S1 | S2 | S3 |
|-----------|-----|-----|-----|
| Jagadish | UM | ATT | UM |
| Dewitt | MSR | MSR | UW |
| Bernstein | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA |
| Franklin | UCB | UCB | UMD |

Data Fusion: Three Components [DBS09a]

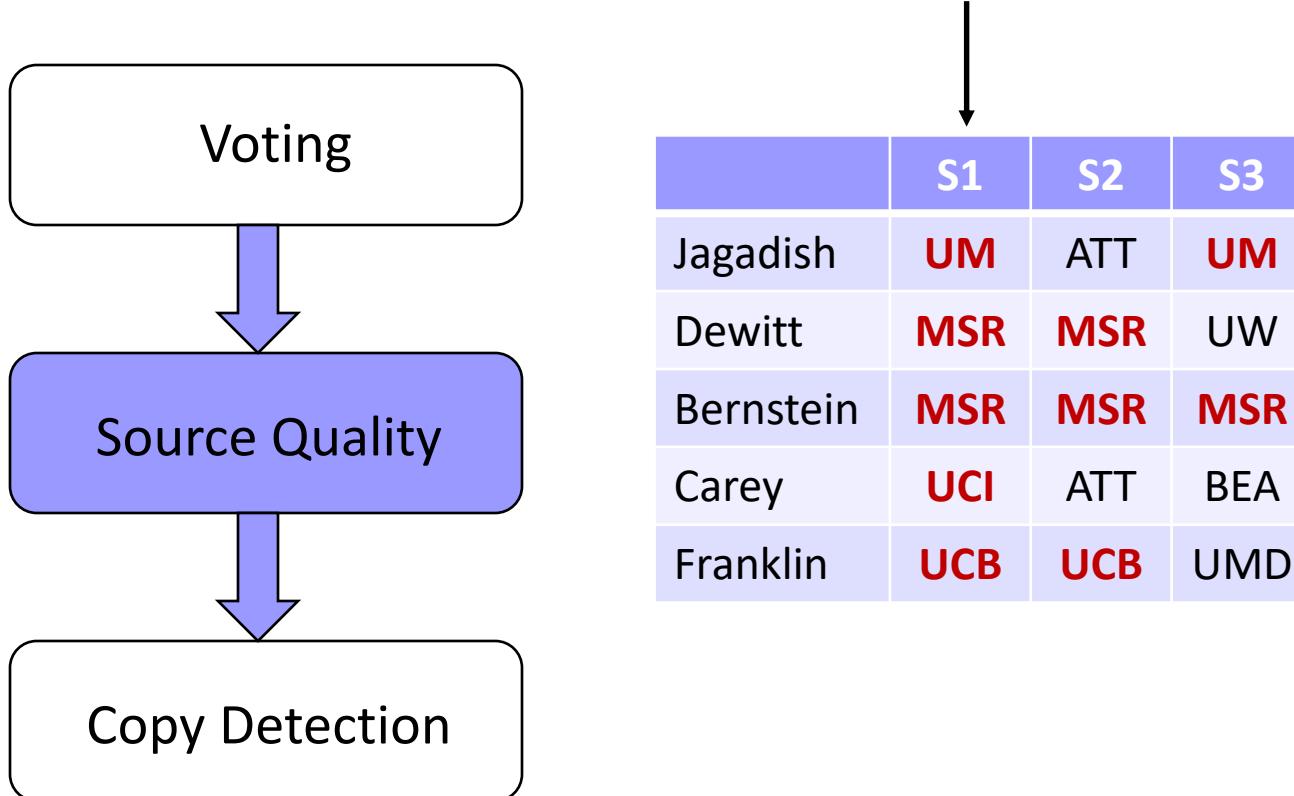
- ◆ Data fusion: voting + source quality + copy detection



| | S1 | S2 | S3 |
|-----------|-----|-----|-----|
| Jagadish | UM | ATT | UM |
| Dewitt | MSR | MSR | UW |
| Bernstein | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA |
| Franklin | UCB | UCB | UMD |

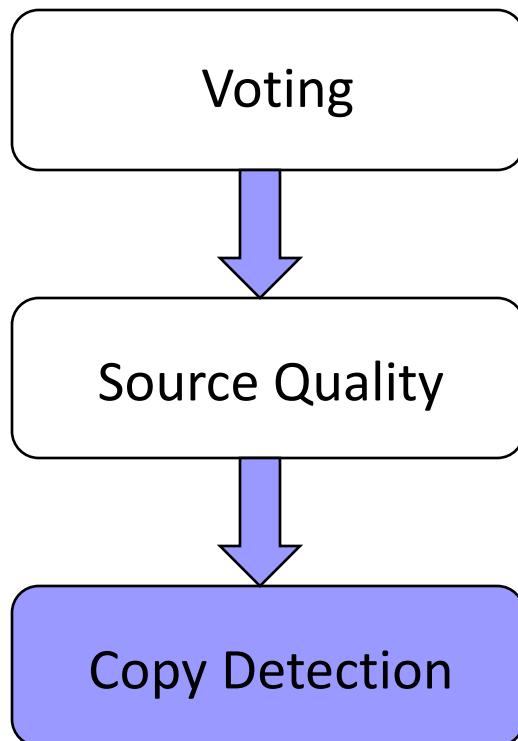
Data Fusion: Three Components [DBS09a]

- ◆ Data fusion: voting + source quality + copy detection
 - Gives more weight to knowledgeable sources



Data Fusion: Three Components [DBS09a]

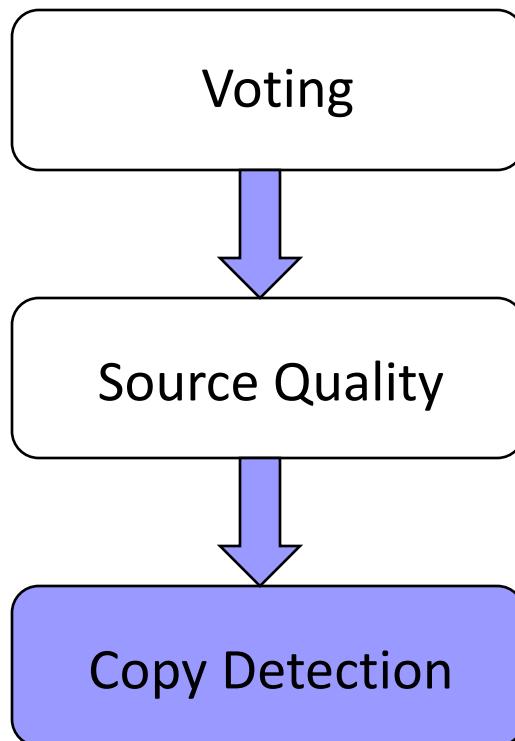
- ◆ Data fusion: voting + source quality + copy detection



| | S1 | S2 | S3 | S4 | S5 |
|-----------|-----|-----|-----|-----|-----|
| Jagadish | UM | ATT | UM | UM | UI |
| Dewitt | MSR | MSR | UW | UW | UW |
| Bernstein | MSR | MSR | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA | BEA | BEA |
| Franklin | UCB | UCB | UMD | UMD | UMD |

Data Fusion: Three Components [DBS09a]

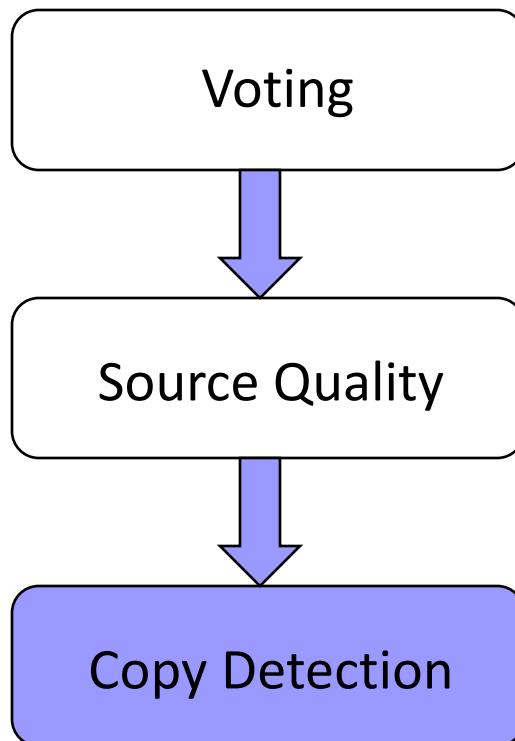
- ◆ Data fusion: voting + source quality + copy detection



| | S1 | S2 | S3 | S4 | S5 |
|-----------|-----|-----|-----|-----|-----|
| Jagadish | UM | ATT | UM | UM | UI |
| Dewitt | MSR | MSR | UW | UW | UW |
| Bernstein | MSR | MSR | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA | BEA | BEA |
| Franklin | UCB | UCB | UMD | UMD | UMD |

Data Fusion: Three Components [DBS09a]

- ◆ Data fusion: voting + source quality + copy detection
 - Reduces weight of copier sources



| | S1 | S2 | S3 | S4 | S5 |
|-----------|-----|-----|-----|-----|-----|
| Jagadish | UM | ATT | UM | UM | UI |
| Dewitt | MSR | MSR | UW | UW | UW |
| Bernstein | MSR | MSR | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA | BEA | BEA |
| Franklin | UCB | UCB | UMD | UMD | UMD |

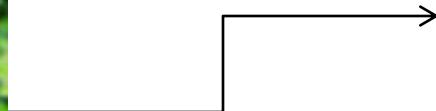
Outline

◆ Motivation

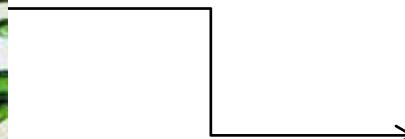
- Why do we need big data integration?
- How has “small” data integration been done?
- Challenges in big data integration

BDI: Why is it Challenging?

- ◆ Data integration = solving lots of jigsaw puzzles
 - Big data integration → **big, messy** puzzles
 - E.g., missing, duplicate, damaged pieces



...



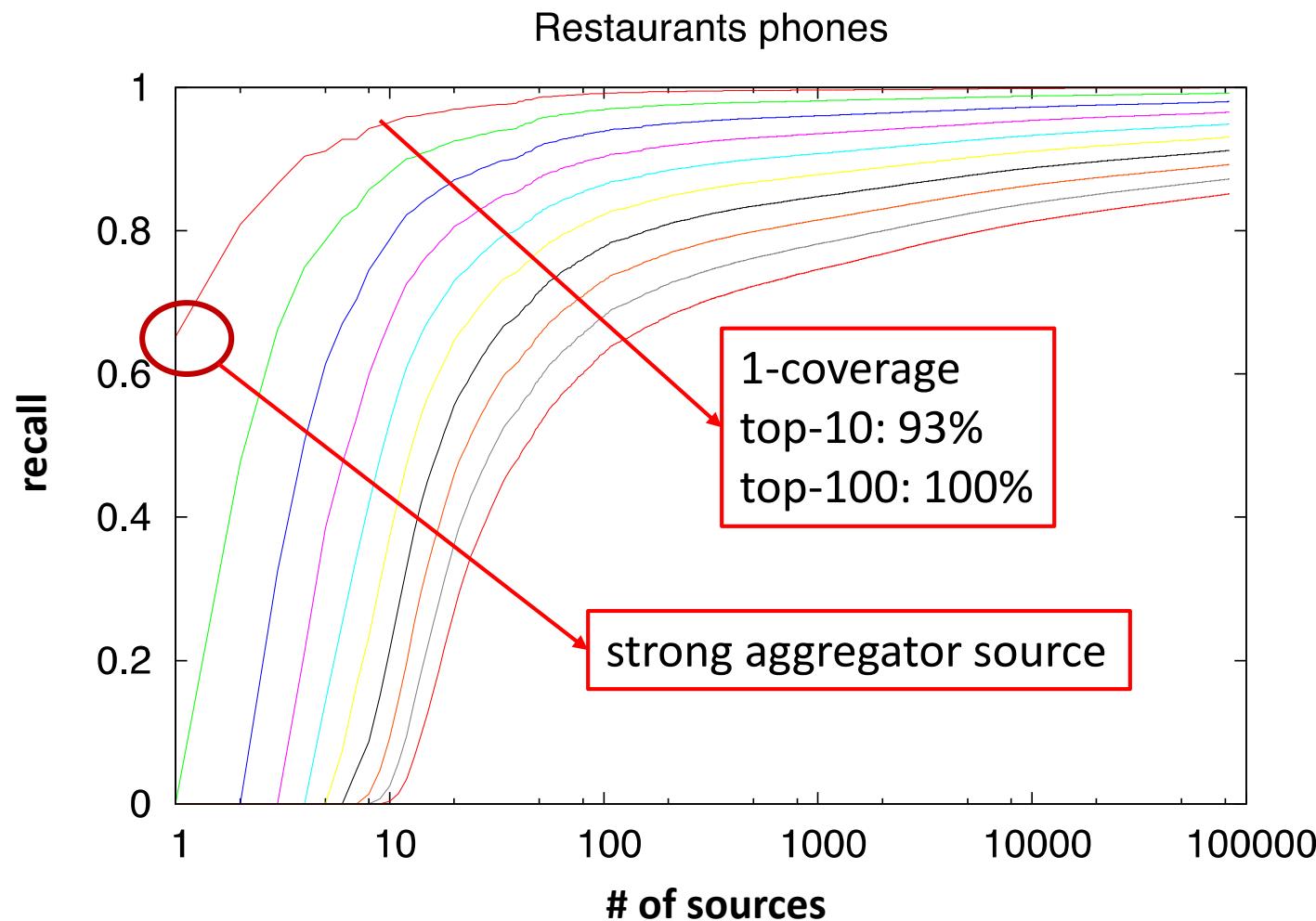
Case Study I: Domain Specific Data [DMP12]

- ◆ Goal: analysis of domain-specific structured data across the Web
- ◆ Questions addressed:
 - How is the data about a given domain spread across the Web?
 - How easy is it to discover entities, sources in a given domain?
 - How much value do the tail entities in a given domain have?

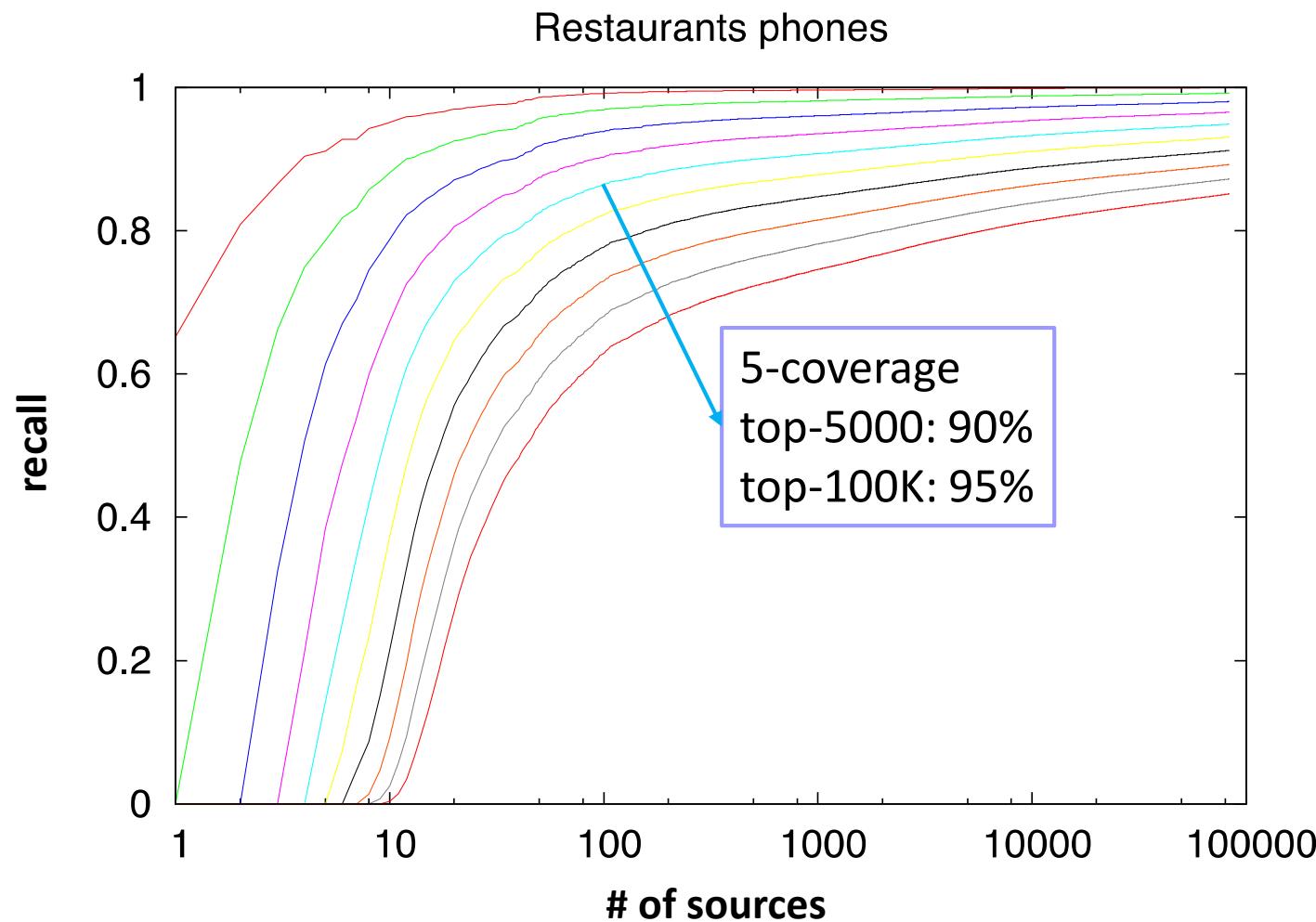
Domain Specific Data: Spread

- ◆ How many sources needed to build a complete DB for a domain?
- ◆ [DMP12] looked at 9 domains with the following properties
 - Access to large comprehensive databases of entities in the domain
 - Entities have attributes that are (nearly) unique identifiers, e.g., ISBN for Books, phone number or homepage for Restaurants
- ◆ Methodology of case study:
 - Used the entire web cache of Yahoo! search engine
 - Webpage has an entity if it contains an identifying attribute
 - Aggregate the set of all entities found on each website (source)

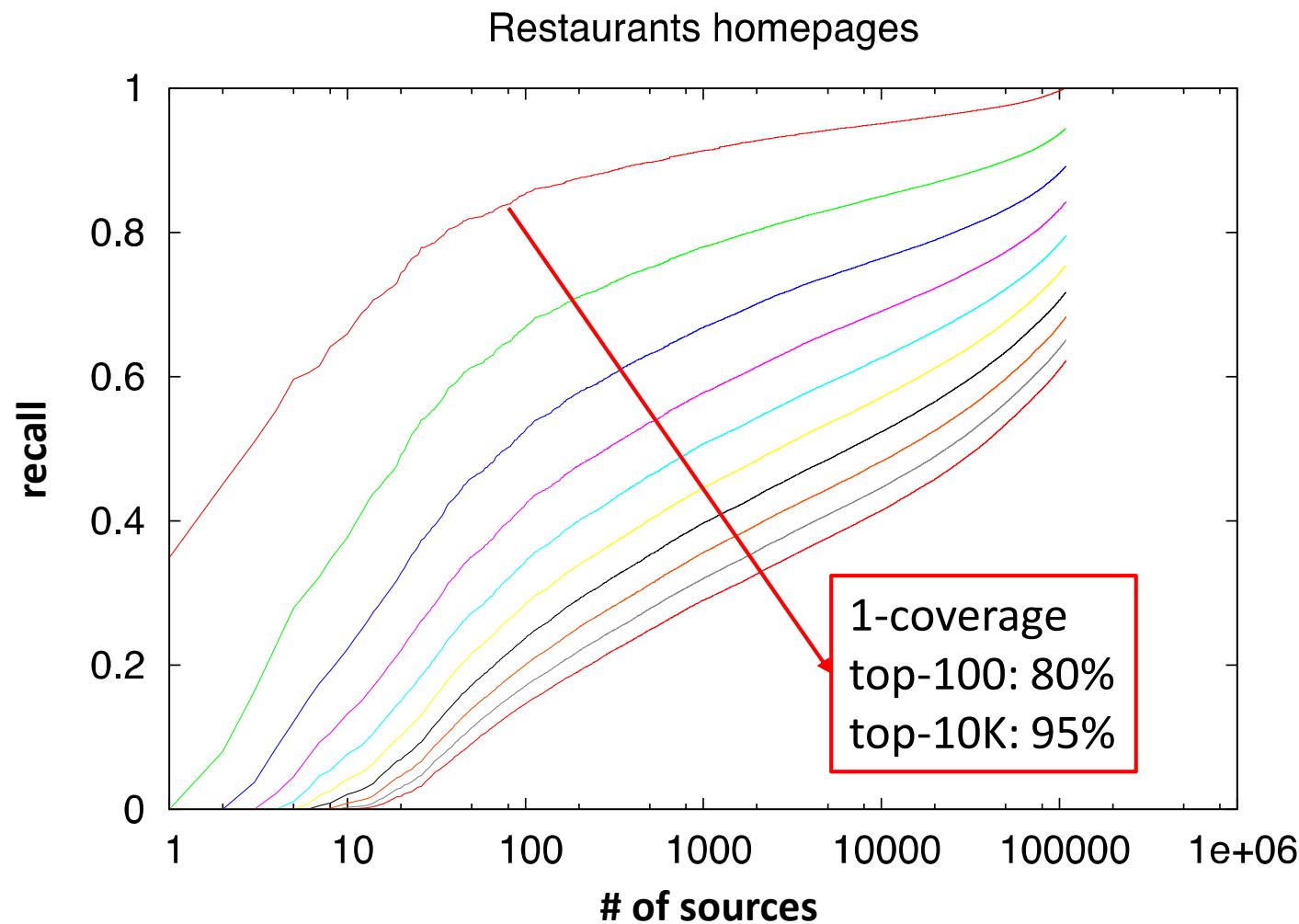
Domain Specific Data: Spread



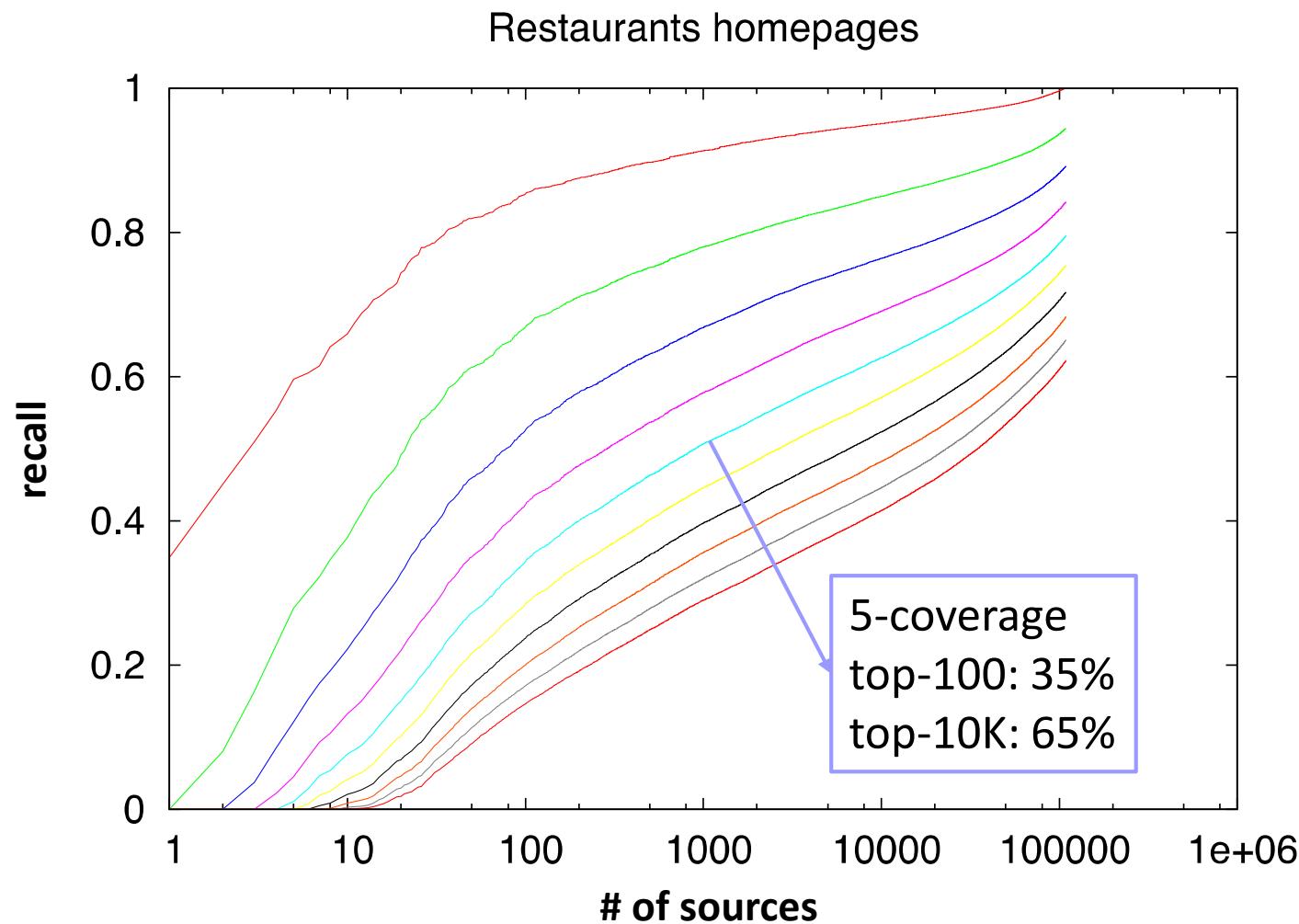
Domain Specific Data: Spread



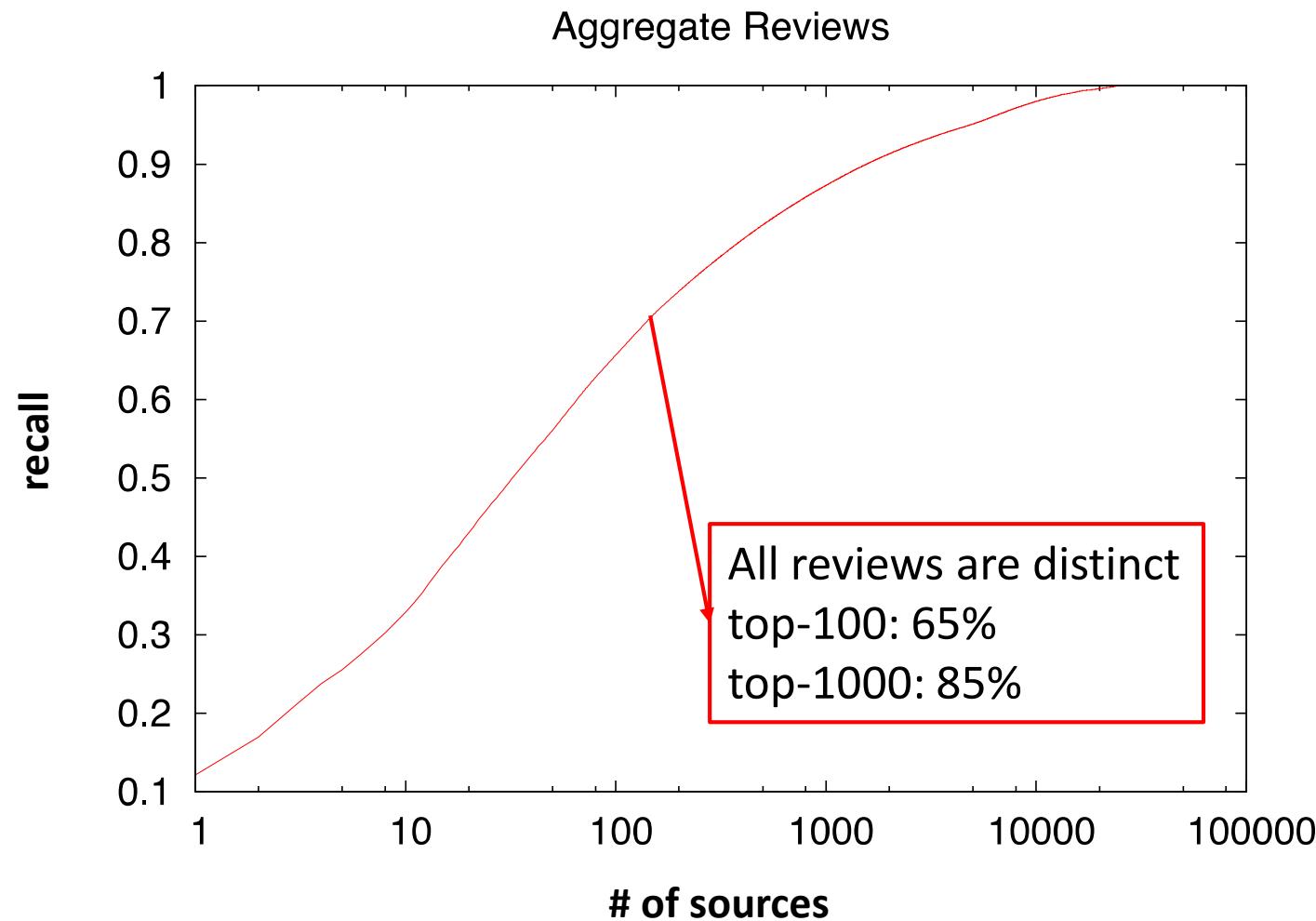
Domain Specific Data: Spread



Domain Specific Data: Spread



Domain Specific Data: Spread



Domain Specific Data: Connectivity

- ◆ How well are the sources “connected” in a given domain?
 - Do you have to be a search engine to find domain-specific sources?
- ◆ [DMP12] considered the entity-source graph for various domains
 - Bipartite graph with entities and sources (websites) as nodes
 - Edge between entity e and source s if some webpage in s contains e
- ◆ Methodology of case study:
 - Study graph properties, e.g., diameter and connected components

Domain Specific Data: Connectivity

- ◆ Almost all entities are connected to each other
 - Largest connected component has more than 99% of entities

| Graph Domain | Attr | Avg. #sites per entity | diameter | # conn. comp. | % entities in largest comp. |
|--------------|----------|------------------------|----------|---------------|-----------------------------|
| Books | ISBN | 8 | 8 | 439 | 99.96 |
| Automotive | phone | 13 | 6 | 9 | 99.99 |
| Banks | phone | 22 | 6 | 15 | 99.99 |
| Home | phone | 13 | 8 | 4507 | 99.76 |
| Hotels | phone | 56 | 6 | 11 | 99.99 |
| Libraries | phone | 47 | 6 | 3 | 99.99 |
| Restaurants | phone | 32 | 6 | 52 | 99.99 |
| Retail | phone | 19 | 7 | 628 | 99.93 |
| Schools | phone | 37 | 6 | 48 | 99.97 |
| Automotive | homepage | 115 | 6 | 10 | 98.52 |
| Banks | homepage | 68 | 8 | 30 | 99.57 |
| Home | homepage | 20 | 8 | 5496 | 97.87 |
| Hotels | homepage | 56 | 8 | 24 | 99.90 |
| Libraries | homepage | 251 | 6 | 4 | 99.86 |
| Restaurants | homepage | 46 | 6 | 146 | 99.82 |
| Retail | homepage | 45 | 7 | 1260 | 99.20 |
| Schools | homepage | 74 | 6 | 122 | 99.57 |

Domain Specific Data: Connectivity

- ◆ High redundancy and overlap enable use of bootstrapping
 - Low diameter ensures that most sources can be found quickly

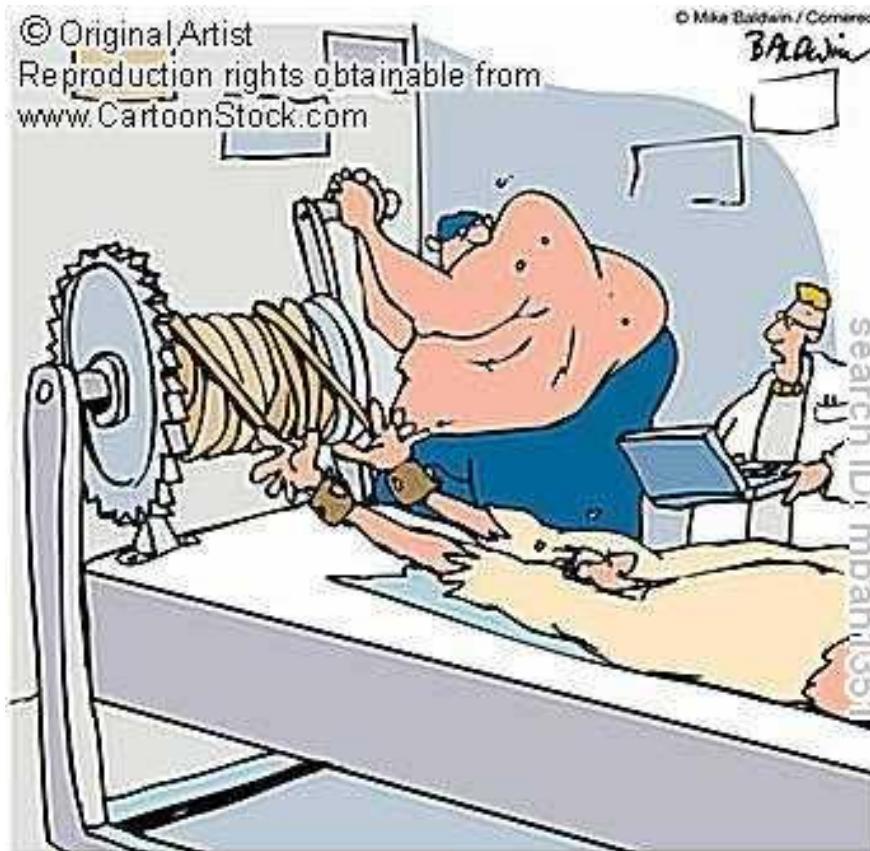
| Graph Domain | Attr | Avg. #sites per entity | diameter | # conn. comp. | % entities in largest comp. |
|--------------|----------|------------------------|----------|---------------|-----------------------------|
| Books | ISBN | 8 | 8 | 439 | 99.96 |
| Automotive | phone | 13 | 6 | 9 | 99.99 |
| Banks | phone | 22 | 6 | 15 | 99.99 |
| Home | phone | 13 | 8 | 4507 | 99.76 |
| Hotels | phone | 56 | 6 | 11 | 99.99 |
| Libraries | phone | 47 | 6 | 3 | 99.99 |
| Restaurants | phone | 32 | 6 | 52 | 99.99 |
| Retail | phone | 19 | 7 | 628 | 99.93 |
| Schools | phone | 37 | 6 | 48 | 99.97 |
| Automotive | homepage | 115 | 6 | 10 | 98.52 |
| Banks | homepage | 68 | 8 | 30 | 99.57 |
| Home | homepage | 20 | 8 | 5496 | 97.87 |
| Hotels | homepage | 56 | 8 | 24 | 99.90 |
| Libraries | homepage | 251 | 6 | 4 | 99.86 |
| Restaurants | homepage | 46 | 6 | 146 | 99.82 |
| Retail | homepage | 45 | 7 | 1260 | 99.20 |
| Schools | homepage | 74 | 6 | 122 | 99.57 |

Domain Specific Data: Lessons Learned

- ◆ Spread:
 - Even for domains with strong aggregators, we need to go to the long tail of sources to build a reasonably complete database
 - Especially true if we want k-coverage for boosting confidence

- ◆ Connectivity:
 - Sources in a domain are well-connected, with a high degree of content redundancy and overlap
 - Remains true even when head aggregator sources are removed

A Lot of Information on the Web



"Come to think of it, he doesn't need to give us the information. I can just look it up on the Internet."

Information Can Be Erroneous

Telegraph.co.uk

Home News Sport Finance Comment Travel Lifestyle Culture Fas
UK World Politics Celebrities Obituaries Weird Earth Science Health News Education

HOME > NEWS > NEWS TOPICS > HOW ABOUT THAT?

Steve Jobs obituary published by Bloomberg

An obituary of very-much-alive Apple founder Steve Jobs has been accidentally published by the respected Bloomberg business news wire.

By Matthew Moore

Last Updated: 7:05PM BST 28 Aug 2008



Steve Jobs was described as the man who 'refashioned the mobile phone' in the erroneous obituary. Photo: REUTERS

The story, marked "Hold for release – Do not use", was sent in error to the news service's thousands of corporate clients.

T Text Size + -

E Email this article

P Print this article

D Share this article

91 diggs [digg it](#)

How about that? [RSS](#)

USA [RSS](#)

News [RSS](#)

The week in pictures



[IN PICS](#)

Pictures of the day

The story, marked "Hold for release – Do not use", was sent in error to the news service's thousands of corporate clients.

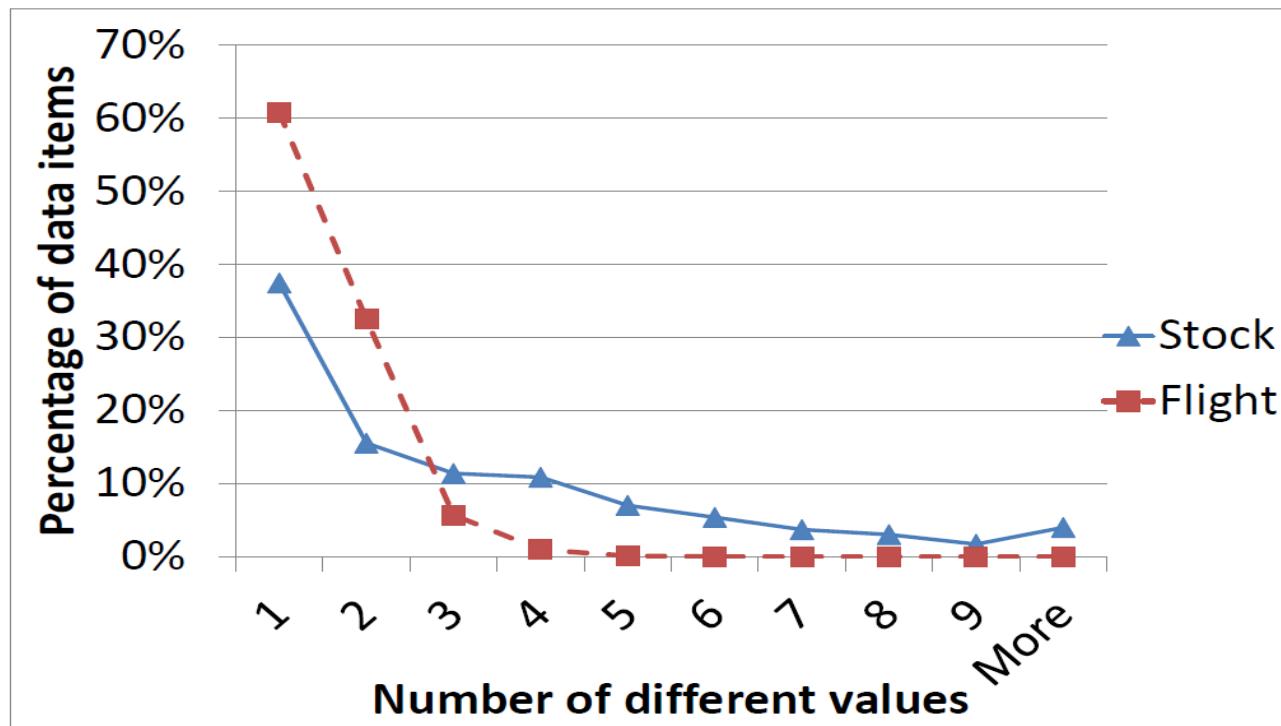
Case Study II: Deep Web Quality [LDL+13]

- ◆ Study on two domains
 - Belief of clean data
 - Poor quality data can have big impact

| | #Sources | Period | #Objects | #Local- attrs | #Global- attrs | Considered items |
|--------|----------|---------|----------|---------------|----------------|------------------|
| Stock | 55 | 7/2011 | 1000*20 | 333 | 153 | 16000*20 |
| Flight | 38 | 12/2011 | 1200*31 | 43 | 15 | 7200*31 |

Deep Web Quality

- ◆ Is the data consistent?
 - Tolerance to 1% value difference



Deep Web Quality

- ◆ Why such inconsistency?
 - Semantic ambiguity

Yahoo! Finance

Green Mountain Coffee Roasters, (NasdaqGS: GMCR)

After Hours: 95.13 ↓ -0.01 (-0.02%) 4:07PM EDT

| | |
|----------------|---|
| Last Trade: | 95.14 |
| Trade Time: | 4:00PM EDT |
| Change: | ↑ 1.69 (1.81%) |
| Prev Close: | 93.45 |
| Open: | 94.01 |
| Bid: | 95.03 x 100 |
| Ask: | 95.94 x 100 |
| 1y Target Est: | 92.50 |

| | |
|---------------|----------------------|
| Day's Range: | 93.80 - 95.71 |
| 52wk Range: | 25.38 - 95.71 |
| Volume: | 2,384,075 |
| Avg Vol (3m): | 2,512,070 |
| Market Cap: | 13.51B |
| P/E (ttm): | 119.82 |
| EPS (ttm): | 0.79 |
| Div & Yield: | N/A (N/A) |

52wk Range: 25.38-95.71

52 Wk: 25.38-93.72

Nasdaq

| | |
|--------------------------------------|---|
| Last Sale | \$ 95.14 |
| Change Net / % | 1.69 ▲ 1.81% |
| Best Bid / Ask | \$ 95.03 / \$ 95.94 |
| 1y Target Est: | \$ 95.00 |
| Today's High / Low | \$ 95.71 / \$ 93.80 |
| Share Volume | 2,384,175 |
| 50 Day Avg. Daily Volume | 2,751,062 |
| Previous Close | \$ 93.45 |
| 52 Wk High / Low | 3.72 / \$ 25.38 |
| Shares Outstanding | 152,785,000 |
| Market Value of Listed Security | ,535,964,900 |
| P/E Ratio | 120.43 |
| Forward P/E (1yr) | 63.57 |
| Earnings Per Share | \$ 0.79 |
| Annualized Dividend | N/A |
| Ex Dividend Date | N/A |
| Dividend Yield | N/A |
| Cash Dividend | N/A |
| EPS (ttm) | 0.82 |
| NASDAQ Official Open Price: | \$ 94.01 |
| Date of NASDAQ Official Open Price: | Jul. 7, 2011 |
| NASDAQ Official Close Price: | \$ 95.14 |
| Date of NASDAQ Official Close Price: | Jul. 7, 2011 |

Day's Range: 93.80-95.71

Deep Web Quality

- ◆ Why such inconsistency?
 - Unit errors

The screenshot shows the NASDAQ website's stock quote page for TTI. The top navigation bar includes links for Home, Quotes & Research, Extended Trading, Market Activity, and News. A sidebar on the left provides links for add symbol, edit symbol list, symbol lookup, Symbol List Views, FlashQuotes, InfoQuotes, Stock Details, Real-Time Quotes, Summary Quotes, After Hours Quotes, Pre-market Quotes, Historical Quotes, Options Chain, CHARTS, Basic Charts, Interactive Charts, COMPANY NEWS, Company Headlines, Press Releases, Sentiment, STOCK ANALYSIS, Analyst Research, and Guru Analysis. The main content area displays the stock quote for TTI: \$13.11, +\$0.51 (4.05%), last updated on Jul 7, 2011, market closed. It also notes that quotes are updated every 7 seconds. Below this, there is a table for 'Last Sale' with columns for Change Net / %, 1y Target Est., Today's High / Low, Share Volume, Previous Close, 52 Wk High / Low, and Shares Outstanding. The 'Shares Outstanding' row is highlighted with a red box and a blue callout bubble containing the value '76,821,000'. Other data in the table includes Market Value of Listed Security (\$1,007,123,310), P/E Ratio (NE), Forward P/E (1yr) (19.69), Earnings Per Share (\$-0.68), and Annualized Dividend (\$NA).

The screenshot shows the UPDOWN website's stock quote page for TTI. The top navigation bar includes links for HOME, TRADING, STOCKS, COMMUNITY, and CO. Below this are links for Overview, Market News, and Top Stock Picks. A green 'GET QUOTE' button is visible. The main content area displays the stock quote for TTI: \$13.11, +\$0.51 (4.05%). It includes a purple callout bubble pointing to the 'Shares Outstanding' value '76,821,000'. Other data shown includes a large '76.82B' (likely a typo for 76.82B), a 'Trade' button, and tabs for Overview, Trade TTI, Stock Picks, and Tweets. Below the quote, there is a message about upgrading Flash Player, followed by a table of historical data for TTI.

| Today | 5d | 1m | 3m | 1y | 5y | 10y |
|-------------|----------------|----|----|------------|---------|-----|
| Last: | \$13.11 | | | High: | \$13.15 | |
| Prev Close: | \$12.60 | | | Low: | \$12.67 | |
| Open: | \$12.82 | | | Mkt Cap: | \$968M | |
| Change: | \$0.51 (4.05%) | | | 52Wk High: | \$16.00 | |
| Vol: | 472,608 | | | 52Wk Low: | \$8.00 | |
| Avg Volume: | 559,308 | | | Shares: | 76.82B | 56 |
| EPS: | - | | | PE Ratio: | - | |

Deep Web Quality

- ◆ Why such inconsistency?

- Pure errors

FlightView

American Airlines Flight Number 119 (AA119)

FLIGHT TRACKER

6:15 PM
Departure
Airport: Newark Liberty Intl (KEWR)
Scheduled Time: 6:15 PM, Dec 08
Takeoff Time: 6:53 PM, Dec 08
Terminal - Gate: Terminal A - 32

ArrivalStatus: In Air

Airport: Newark Liberty Intl (KEWR)

Scheduled Time: 9:40 PM, Dec 08

9:42 PM, Dec 08

Estimated Time:

Track This Flight

Time Remaining: 25 min

Terminal - Gate: Terminal 4 - 42

Baggage Claim: 4

9:40 PM

FlightAware

AAL119 ([Track inbound flight](#))

([web site](#)) ([all flights](#))

American Airlines "American"

Aircraft: Boeing 737-800 (twin-jet) (B738/Q - [track](#) or [photos](#))

Origin: Terminal A / Gate 32 / Newark Liberty Intl (KEWR - [track](#) or [info](#))

Destination: Terminal 4 / Gate 42B / Los Angeles Intl (KLAX - [track](#) or [info](#))

Other flights between these airports

ZIMMZ Q42 BTRIX Q480 AIR J80 VHP J80 MCI J24 SLN J102 ALS J44 RSK J64 PGS RIIVR2
[Decode](#)

Route: ZIMMZ Q42 BTRIX Q480 AIR J80 VHP J80 MCI J24 SLN J102 ALS J44 RSK J64 PGS RIIVR2

Date: 2011年 12月 08日 (Thursday)

Duration: 5 hours 43 minutes

20 minutes left

5 hours 23 minutes

Progress: 6:15 PM

Status: [En Route](#) (2,284 sm down / 38 sm to go)

Distance: Direct: 2,451 sm Planned: 2,458

Fare: \$51.99 to \$3,561.00, average: \$241.96 ([airline insight](#))

Cabin: First Class / Premium Economy: Food for sale

Scheduled: 7-day Average [Actual/Estimated](#)

Departure: 06:15PM EST 07:08PM EST 06:53PM EST

Arrival: 08:33PM PST 09:17PM PST 09:36PM PST

9:40 PM

8:33 PM

Orbitz

American Airlines # 119

Leg 1: In Transit

Departs: Newark (EWR) [View real-time airport conditions at](#)

Gate: 32

Scheduled Estimated Actual

6:22p - 6:32p
Dec 8 Dec 8

6:22 PM

Arrives: Los Angeles (LAX) [View real-time airport conditions](#)

Gate: 42B

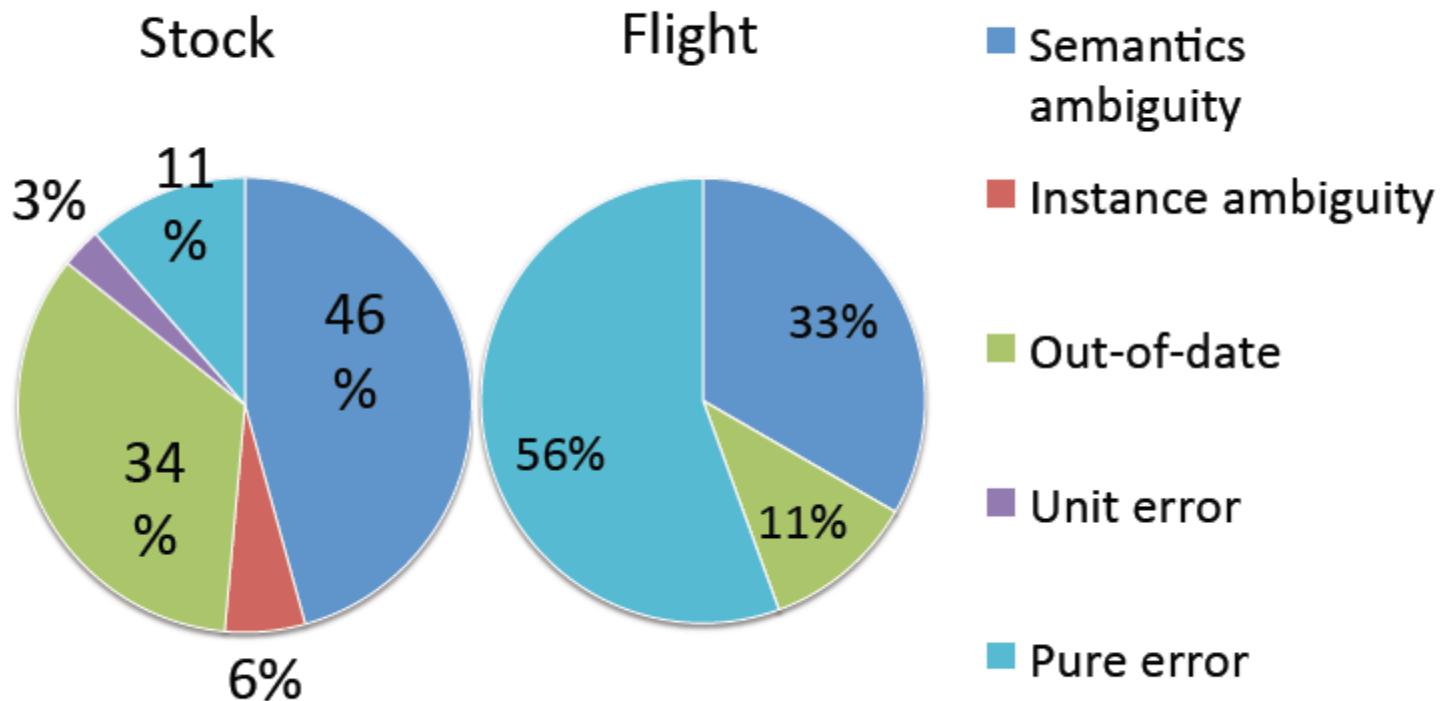
Scheduled Estimated Actual

9:54p 9:47p
Dec 8 Dec 8

9:54 PM

Deep Web Quality

- ◆ Why such inconsistency?
 - Random sample of 20 data items + 5 items with largest # of values



Deep Web Quality

◆ Copying between sources?

The figure displays four separate web pages from different financial sources, all showing the same basic information for Apple, Inc. (AAPL). Each page includes a red arrow pointing downwards, indicating a price decrease.

Page 1 (Left): Shows a chart for Apple, Inc. (AAPL) with a price drop of -6.00 (-1.775%) at 3:30 PM. It includes tabs for Quote, News, Profile, Overview, Detailed Quote, and Related Searches (Forex Calculator, Currency Trading).

Page 2 (Second from Left): Shows a similar chart for Apple, Inc. (AAPL) with a price drop of -6.00 (-1.775%) at 3:30 PM. It includes tabs for Quote, News, Profile, Research, and Community.

Page 3 (Third from Left): Shows a chart for Apple, Inc. (AAPL) with a price drop of -6.00 (-1.775%) at 3:30 PM. It includes tabs for Quote, News, Profile, Research, and Community, along with a 'Trade Now' button.

Page 4 (Right): Shows a chart for Apple, Inc. (AAPL) with a price drop of -6.00 (-1.775%) at 3:30 PM. It includes tabs for Quote, News, Profile, Research, and Community, along with a 'VIEW DETAILED QUOTE' link and a detailed price table.

Deep Web Quality

- ◆ Copying on erroneous data?

| | Remarks | Size | Schema sim | Object sim | Value sim | Avg accu |
|--------|--------------------|------|------------|------------|-----------|----------|
| Stock | Depen claimed | 11 | 1 | .99 | .99 | .92 |
| | Depen claimed | 2 | 1 | 1 | .99 | .75 |
| Flight | Depen claimed | 5 | 0.80 | 1 | 1 | .71 |
| | Query redirection | 4 | 0.83 | 1 | 1 | .53 |
| | Dependence claimed | 3 | 1 | 1 | 1 | .92 |
| | Embedded interface | 2 | 1 | 1 | 1 | .93 |
| | Embedded interface | 2 | 1 | 1 | 1 | .61 |

Deep Web Quality: Lessons Learned

- ◆ Deep Web data has considerable inconsistency
 - Even in domains where poor quality data can have big impact
 - Semantics ambiguity, out of date data, unexplainable errors
- ◆ Deep Web sources often copy from each other
 - Copying can happen on erroneous data, spreading poor quality data

BDI: Why is it Challenging?

- ◆ Number of structured sources: **Volume**
 - Millions of websites with domain specific structured data [DMP12]
 - 154 million high quality relational tables on the web [CHW+08]
 - 10s of millions of high quality deep web sources [MKK+08]
 - 10s of millions of useful relational tables from web lists [EMH09]
- ◆ Challenges:
 - Difficult to do schema alignment
 - Expensive to warehouse all the integrated data
 - Infeasible to support virtual integration

BDI: Why is it Challenging?

- ◆ Rate of change in structured sources: **Velocity**
 - 43,000 – 96,000 deep web sources (with HTML forms) [B01]
 - 450,000 databases, 1.25M query interfaces on the web [CHZ05]
 - 10s of millions of high quality deep web sources [MKK+08]
 - Many sources provide rapidly changing data, e.g., stock prices
- ◆ Challenges:
 - Difficult to understand evolution of semantics
 - Extremely expensive to warehouse data history
 - Infeasible to capture rapid data changes in a timely fashion

BDI: Why is it Challenging?

- ◆ Representation differences among sources: **Variety**

Synopsis: Born or conceived informed him. His ideas are influenced by Italian Renaissance.

| Leonardo da Vinci | | | | |
|-------------------|--|-------------------|----------------------|---------------------------------------|
| D | DALMATA, Giovanni | (1440-1510) | Early Renaissance | Italian sculptor |
| | DANIELE da Volterra | (1509-1566) | High Renaissance | Italian painter |
| | DANTI, Vincenzo | (1530-1576) | Mannerism | Italian sculptor (Florence) |
| | DESIDERIO DA SETTIGNANO | (c. 1428-1464) | Early Renaissance | Italian sculptor (Florence) |
| | DIANA, Benedetto | (known 1482-1525) | High Renaissance | Italian painter (Venice) |
| | DOMENICO DA TOLMEZZO | (c. 1448-1507) | Early Renaissance | Italian painter (Venice) |
| | DOMENICO DI BARTOLO | (c. 1400-c. 1447) | Early Renaissance | Italian painter (Siena) |
| | DOMENICO DI MICHELINO | (1417-1491) | Early Renaissance | Italian painter (Florence) |
| | DOMENICO VENEZIANO | (c. 1410-1461) | Early Renaissance | Italian painter (Florence) |
| | DONATELLO | (c. 1386-1466) | Early Renaissance | Italian sculptor |
| | DONDUCCI, Giovanni Andrea (see MASTELLETTA) | (1575-1675) | Mannerism | Italian painter (Rome) |
| | DOSIO, Giovanni Antonio | (1533-c. 1609) | Mannerism | Italian graphic artist |
| | DOSSI, Dosso | (c. 1490-1542) | High Renaissance | Italian painter (Ferrara) |
| | DUCA, Jacopo del | (c. 1520-1604) | Mannerism | Italian sculptor (Sicily) |
| | DUCCIO, Agostino di | (1418-1481) | Early Renaissance | Italian sculptor (Rimini) |
| | DURER, Albrecht | (1471-1528) | Northern Renaissance | German painter/printmaker (Nuremberg) |

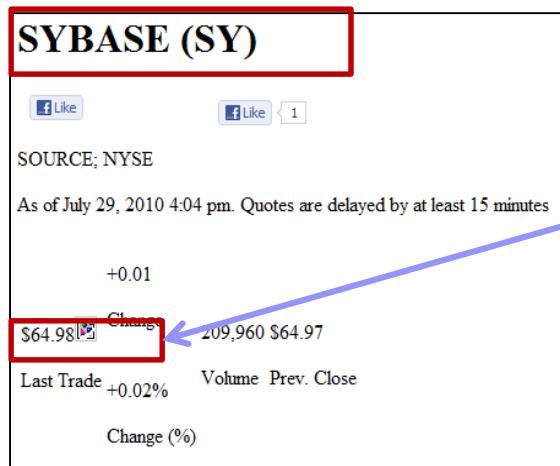
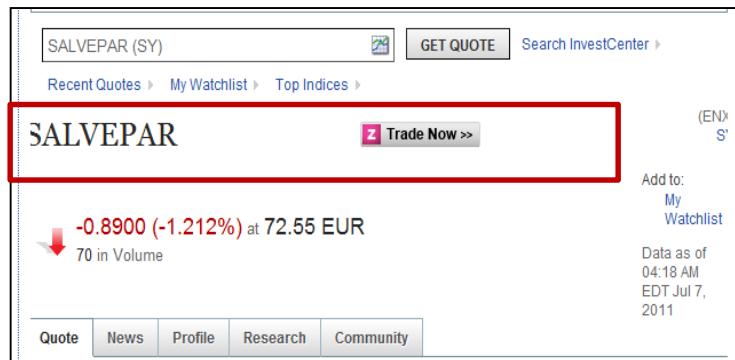
Movement High Renaissance
Works *Mona Lisa*
The Last Supper
The Vitruvian Man
Lady with an Ermine



Turin
arts
and sciences

BDI: Why is it Challenging?

- ◆ Poor data quality of deep web sources [LDL+13]: **Veracity**



| Stock Details | |
|----------------|---------------|
| Last Trade: | 64.98 |
| Change: | +0.00 (0.00%) |
| Prev Close: | 64.98 |
| Open: | 14.73 |
| Days Range: | 64.98 – 64.98 |
| 52 Week Range: | 33.54 - 66.00 |
| Volume: | 88168 |
| P/E: | 31.54 |
| EPS: | 2.06 |

Outline

- ◆ Motivation
- ◆ Schema alignment
- ◆ Record linkage
- ◆ Data fusion
- ◆ Emerging topics

BDI: Schema Alignment

◆ **Volume, Variety**

- Integrating deep web query interfaces [WYD+04, CHZ05]
- Crawl, index deep web data [MKK+08]
- Extract structured data from web tables [CHW+08, LSC10, PS12, DFG+12] and web lists [GS09, EMH09]
- Dataspace systems [FHM05, HFM06, DHY07]
- Keyword search based data integration [TJM+08]

◆ **Velocity**

- Keyword search-based dynamic data integration [TIP10]

BDI: Record Linkage

- ◆ **Volume**: dealing with billions of records
 - Map-reduce based record linkage [VCL10, KTR12]
 - Adaptive record blocking [DNS+12, MKB12, VN12]
 - Blocking in heterogeneous data spaces [PIP+12, PKP+13]

- ◆ **Velocity**
 - Incremental record linkage [WGM10, WGM13, GDS14]

BDI: Record Linkage

◆ Variety

- Matching structured and unstructured data [KGA+11, KTT+12]
- Matching Web tables and catalogs [LSC10]

◆ Veracity

- Linking temporal records [LDM+11]
- Using crowdsourcing oracle [WLK+13, VBD14, FSS16]

BDI: Data Fusion

◆ **Veracity**

- Using source trustworthiness [YHY08, GAM+10, PR11, YT11, GSH11, PR13]
- Combining source accuracy and copy detection [DBS09a, QAH+13]
- Multiple truth values [ZRG+12]
- Erroneous numeric data [ZH12]
- Experimental comparison on deep web data [LDL+13]

BDI: Data Fusion

- ◆ **Volume:**

- Online data fusion [LDO+11]

- ◆ **Velocity**

- Truth discovery for dynamic data [DBS09b, PRM+12]

- ◆ **Variety**

- Combining record linkage with data fusion [GDS+10]

Questions? Suggestions? Criticisms?



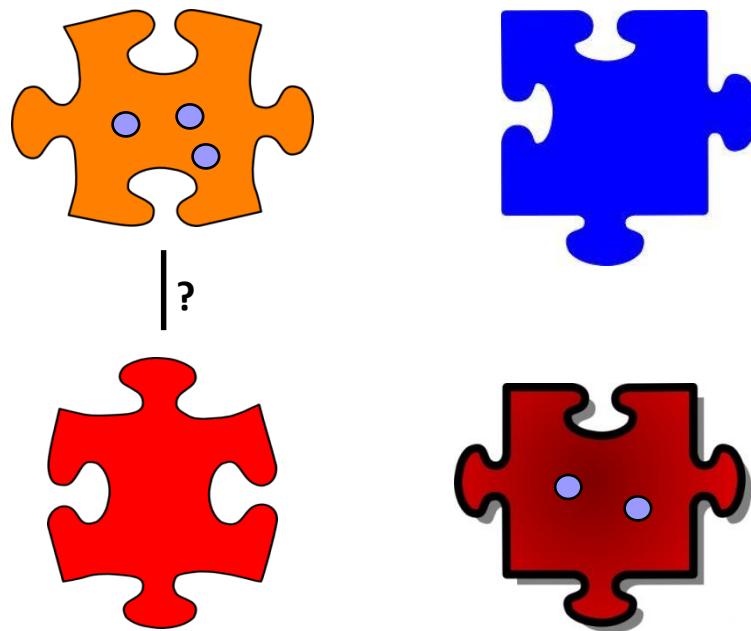
Outline

- ◆ Motivation
- ◆ Schema alignment
 - Overview
 - Techniques for big data
- ◆ Record linkage
- ◆ Data fusion
- ◆ Emerging topics



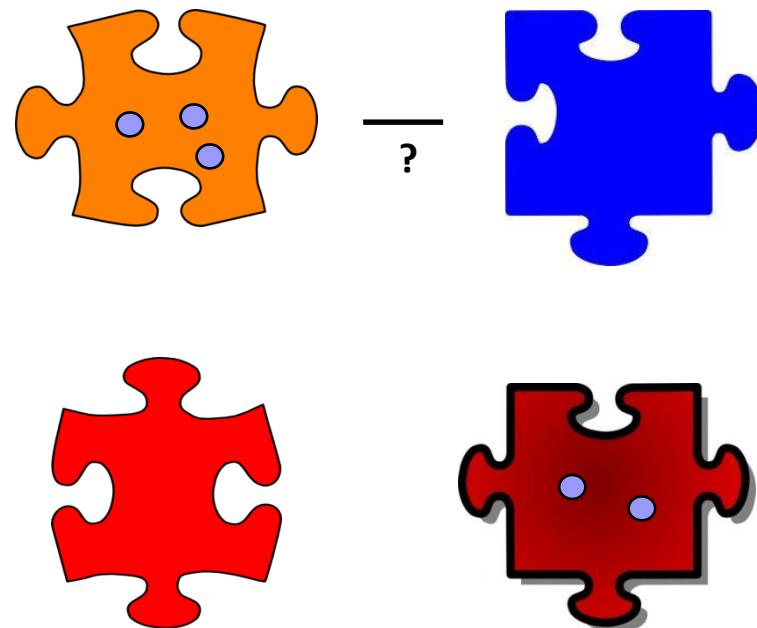
Schema Alignment

- ◆ Matching based on structure (e.g., shape)



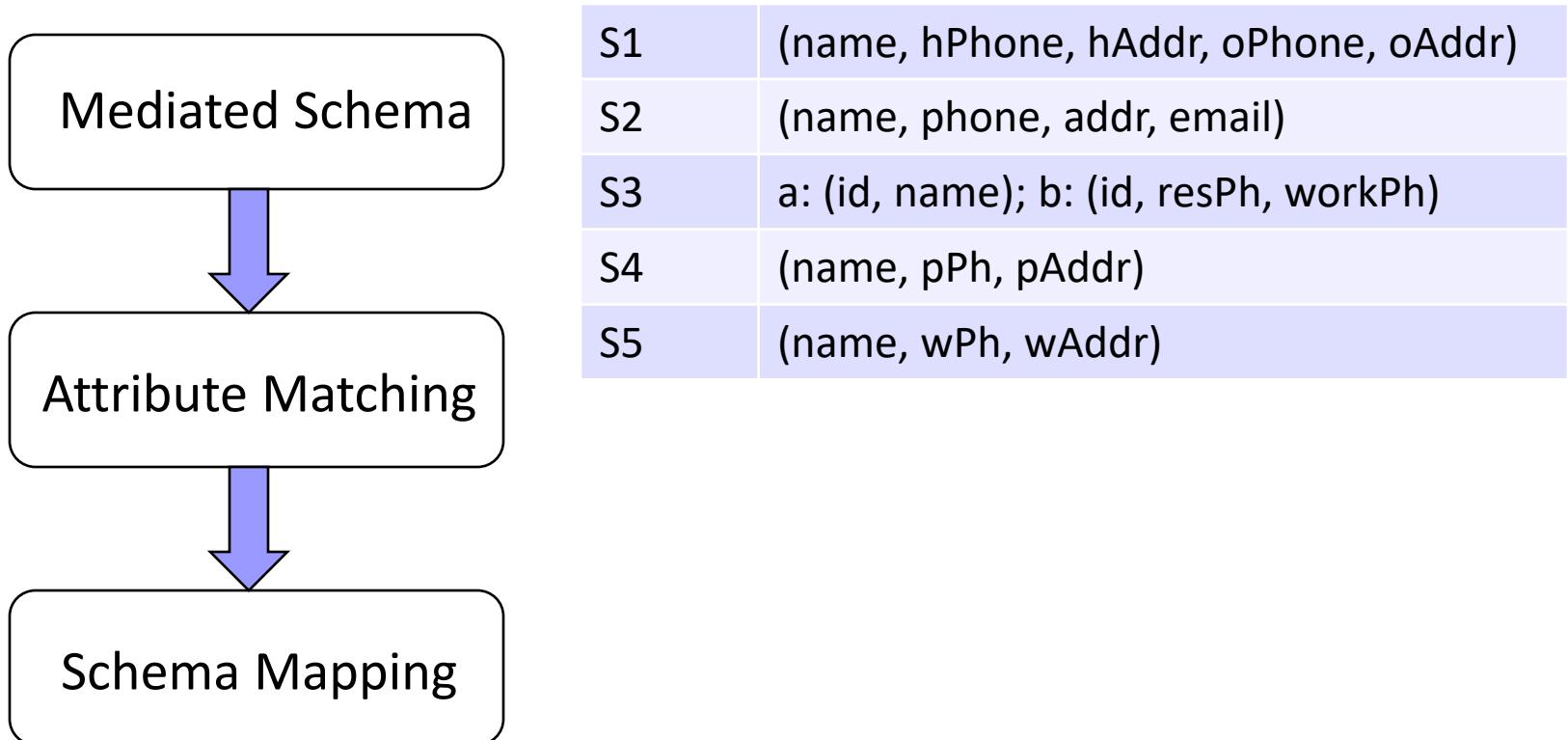
Schema Alignment

- ◆ Matching based on structure (e.g., shape)



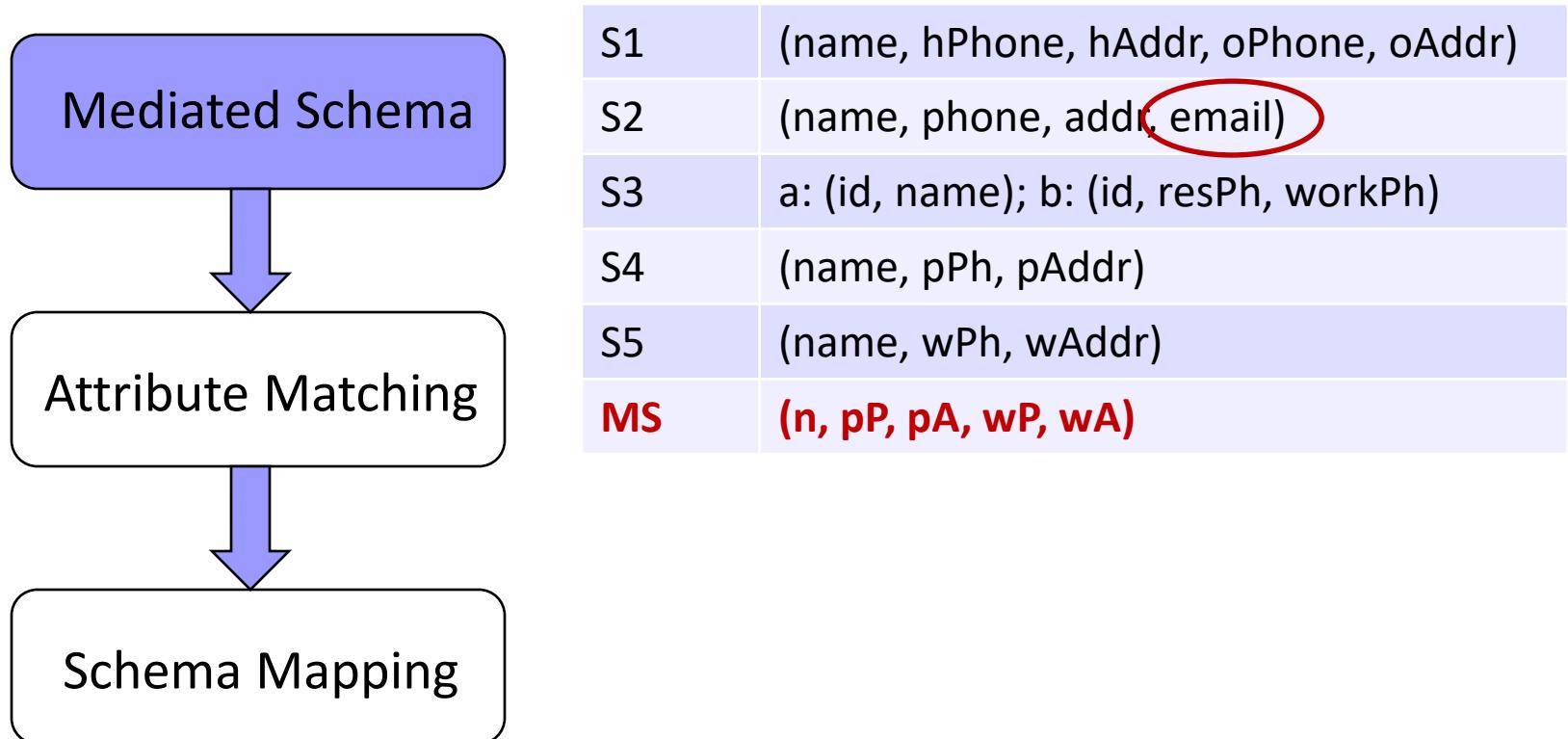
Schema Alignment: Three Steps [BBR II]

- ◆ Schema alignment: mediated schema + matching + mapping
 - Enables linkage, fusion to be semantically meaningful



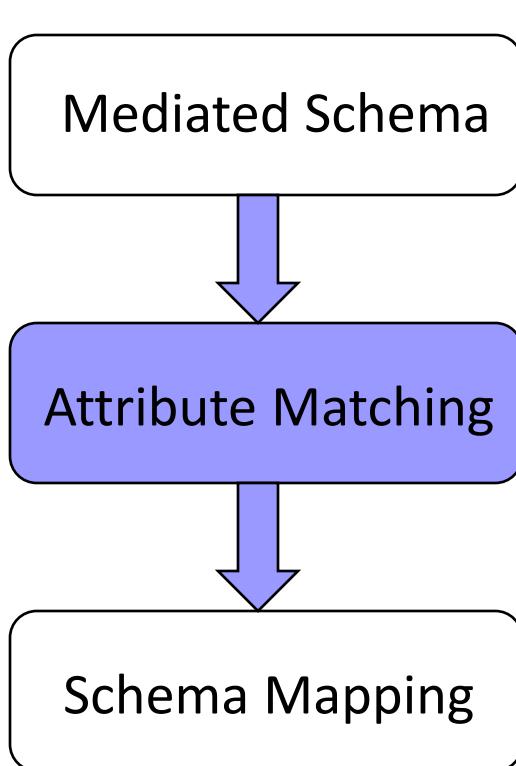
Schema Alignment: Three Steps

- ◆ Schema alignment: mediated schema + matching + mapping
 - Enables domain specific modeling



Schema Alignment: Three Steps

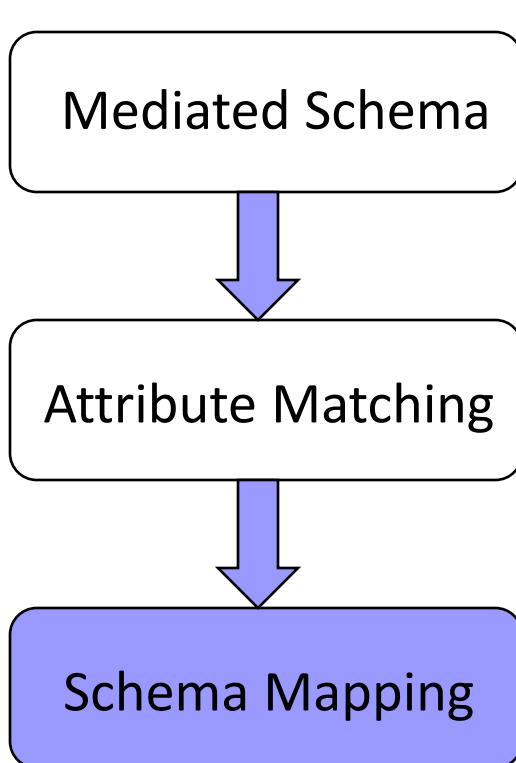
- ◆ Schema alignment: mediated schema + matching + mapping
 - Identifies correspondences between schema attributes



| | |
|-------------|--|
| S1 | (name, hPhone, hAddr, oPhone, oAddr) |
| S2 | (name, phone, addr, email) |
| S3 | a: (id, name); b: (id, resPh, workPh) |
| S4 | (name, pPh, pAddr) |
| S5 | (name, wPh, wAddr) |
| MS | (n, pP, pA, wP, wA) |
| MSAM | MS.n : S1.name, S2.name, S3a.name, ... MS.pP : S1.hPhone, S3b.resPh, S4.pPh MS.pA : S1.hAddr, S4.pAddr MS.wP : S1.oPhone, S2.phone, ... MS.wA : S1.oAddr, S2.addr, S5.wAddr |

Schema Alignment: Three Steps

- ◆ Schema alignment: mediated schema + matching + mapping
 - Specifies transformation between records in different schemas



| | |
|-----------------------|---|
| S1 | (name, hPhone, hAddr, oPhone, oAddr) |
| S2 | (name, phone, addr, email) |
| S3 | a: (id, name); b: (id, resPh, workPh) |
| S4 | (name, pPh, pAddr) |
| S5 | (name, wPh, wAddr) |
| MS | (n, pP, pA, wP, wA) |
| MSSM (GAV) | MS(n, pP, pA, wP, wA) :- S1(n, pP, pA, wP, wA) MS(n, _, _, wP, wA) :- S2(n, wP, wA, e) MS(n, pP, _, wP, _) :- S3a(i, n), S3b(i, pP, wP) MS(n, pP, pA, _, _) :- S4(n, pP, pA) MS(n, _, _, wP, wA) :- S5(n, wP, wA) |

Outline

- ◆ Motivation
- ◆ Schema alignment
 - Overview
 - Techniques for big data
- ◆ Record linkage
- ◆ Data fusion
- ◆ Emerging topics

BDI: Schema Alignment

◆ **Volume, Variety**

- Integrating deep web query interfaces [WYD+04, CHZ05]
- Crawl, index deep web data [MKK+08]
- Extract structured data from web tables [CHW+08, LSC10, PS12, DFG+12] and web lists [GS09, EMH09]
- Dataspace systems [FHM05, HFM06, DHY07]
- Keyword search based data integration [TJM+08]

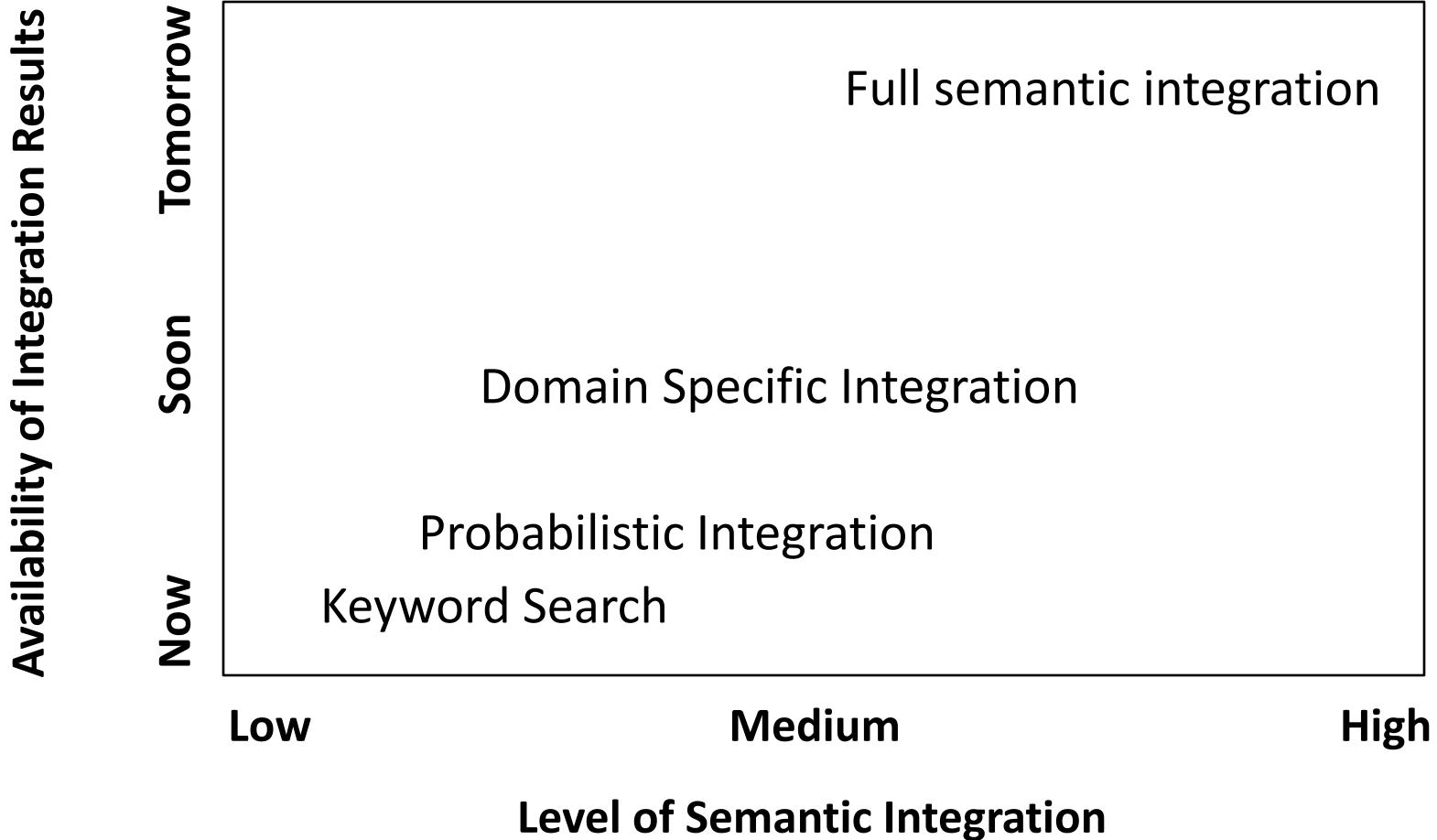
◆ **Velocity**

- Keyword search-based dynamic data integration [TIP10]

Space of Strategies

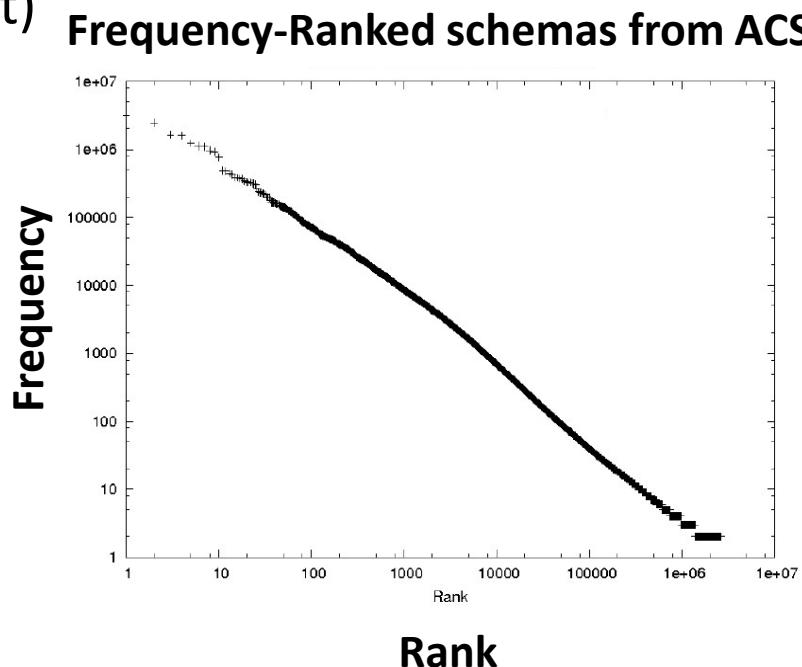
- ◆ Now: keyword search over **all** data sources
 - Keywords used to query, integrate sources [CHW+08, TJM+08]
- ◆ Now and soon: automatic **lightweight** integration
 - Model uncertainty: probabilistic schema, mappings [DDH09, DHY07]
 - Cluster sources, enable domain specific integration [CHZ05]
- ◆ Tomorrow: full-fledged semantic data integration across domains

Space of Strategies



WebTables [CHW+08]

- ◆ Background: Google crawl of the surface web, reported in 2008
 - 154M good relational tables, 5.4M attribute names, 2.6M schemas
- ◆ ACSDb: <https://web.eecs.umich.edu/~michjc/data/acsdb.html>
 - (schema, count)



WebTables: Keyword Search [CHW+08]

- ◆ Query model: keyword search
- ◆ Goal: Rank tables on web in response to query keywords
 - Not web pages (can have multiple tables), not individual records
- ◆ Challenges:
 - Web page features apply ambiguously to embedded tables
 - Web tables on a page may not all be relevant to a query
 - Web tables have specific features (e.g., schema elements)

WebTables: Keyword Search

- ◆ Example keyword query: “presidents of the US”

| President | | | | | |
|-----------|------------------------------------|---|---|-------------------|--|
| George W. | Shenandoah |  | Virginia 38.53°N 78.35°W | May 22, 1926 | 199,045.23 acres (805.5 km ²) Shenandoah's Blue Ridge Mountains are covered by hardwood forests that are home to tens of thousands of animals. The Skyline Drive and Appalachian Trail run the entire length of this narrow park that has more than 500 miles (800 km) of hiking trails along scenic overlooks and waterfalls of the Shenandoah River. ^[57] |
| John | 1. G. 2. J. |  | North Dakota 46.97°N 103.45°W | November 10, 1978 | 70,446.89 acres (285.1 km ²) This region that enticed and influenced President Theodore Roosevelt is now a park of three units in the badlands. Besides Roosevelt's historic cabin, there are scenic drives and backcountry hiking opportunities. Wildlife includes American Bison, pronghorns, Bighorn sheep, and wild horses. ^[58] |
| James | 3. T. 4. Ja. |  | United States Virgin Islands 18.33°N 64.73°W | August 2, 1956 | 14,688.87 acres (59.4 km ²) The island of Saint John has rich human and natural history. There are Taíno archaeological sites and ruins of sugar plantations from Columbus's time. Past the pristine beaches are mangroves, seagrass beds, coral reefs and algal plains. ^[59] |
| John | 5. Ja. 6. Je. 7. A. 8. M. |  | Minnesota 48.50°N | January 8, 1971 | 218,200.17 acres This park on four main lakes, a site for canoeing, kayaking, and fishing, has a history of Ojibwe Native Americans, French fur traders called voyageurs, and a gold rush. Formed by glaciers |
| Martin | 9. V. 10. . 11. . |  | James Monroe | | |
| William | 12. Zachary Taylor (1784-1850) | | | | |

WebTables: Keyword Search

- ◆ FeatureRank: use table specific features

- Query independent features
- Query dependent features
- Linear regression estimator
- Heavily weighted features

| |
|-------------------------------------|
| # rows |
| # cols |
| has-header? |
| # of NULLs in table |
| document-search rank of source page |
| # hits on header |
| # hits on leftmost column |
| # hits on second-to-leftmost column |
| # hits on table body |

- ◆ Result quality: fraction of high scoring relevant tables

| k | Naïve | FeatureRank |
|----|-------|-------------|
| 10 | 0.26 | 0.43 |
| 20 | 0.33 | 0.56 |
| 30 | 0.34 | 0.66 |

WebTables: Keyword Search

- ◆ Example keyword query: “presidents of the US”

| President | | | | | | |
|--------------------------------|--------------------------------|---|--|-------------------|---------------------------------|--|
| George W. | Shenandoah |  | Virginia 38.53°N 78.35°W | May 22, 1926 | 199,045.23 acres (805.5 km²) | Shenandoah's Blue Ridge Mountains are covered by hardwood forests that are home to tens of thousands of animals. The Skyline Drive and Appalachian Trail run the entire length of this narrow park that has more than 500 miles (800 km) of hiking trails along scenic overlooks and waterfalls of the Shenandoah River. ^[57] |
| John | 1. George Washington | | | | | |
| Thomas | 2. John Adams | | | | | |
| James | 3. Thomas Jefferson |  | North Dakota 46.97°N 103.45°W | November 10, 1978 | 70,446.89 acres (285.1 km²) | This region that enticed and influenced President Theodore Roosevelt is now a park of three units in the badlands. Besides Roosevelt's historic cabin, there are scenic drives and backcountry hiking opportunities. Wildlife includes American Bison, pronghorns, Bighorn sheep, and wild horses. ^[58] |
| James | 5. James Monroe | | | | | |
| John | 6. John Quincy Adams | | | | | |
| Andrew | 7. Andrew Jackson |  | United States Virgin Islands 18.33°N 64.73°W | August 2, 1956 | 14,688.87 acres (59.4 km²) | The island of Saint John has rich human and natural history. There are Taino archaeological sites and ruins of sugar plantations from Columbus's time. Past the pristine beaches are mangroves, seagrass beds, coral reefs and algal plains. ^[59] |
| Martin | 9. Martin Van Buren |  | Minnesota 48.50°N | January 8, 1971 | 218,200.17 acres | This park on four main lakes, a site for canoeing, kayaking, and fishing, has a history of Ojibwe Native Americans, French fur traders called voyageurs, and a gold rush. Formed by glaciers |
| William | 10. William H. Harrison | | | | | |
| | 11. Zachary Taylor (1784-1850) |  | James Monroe | | | |
| 12. Zachary Taylor (1784-1850) | | | | | | |

WebTables: Keyword Search

- ◆ SchemaRank: also include **schema coherency** as a table feature
 - Use point-wise mutual information (pmi) derived from ACSDb
 - $p(a) = \text{fraction of unique schemas containing attributes } a$
 - $\text{pmi}(a,b) = \log_2(p(a,b)/(p(a)*p(b)))$
 - Coherency = average pmi(a,b) over all a, b in attrs(R)
- ◆ Result quality: fraction of high scoring relevant tables

| k | Naïve | FeatureRank | SchemaRank |
|----|-------|-------------|------------|
| 10 | 0.26 | 0.43 | 0.47 |
| 20 | 0.33 | 0.56 | 0.59 |
| 30 | 0.34 | 0.66 | 0.68 |

WebTables: Keyword Search

- ◆ Example keyword query: “presidents of the US”
 - T1(President, Vice President)
 - T2(President, Term, Party, Vice President)
 - T3(State, Governor, Party, Term)
 - T4(State, Senator, Party, Term, Born On)
 - T5(Chief Justice, Nominated By, Term)

WebTables: Keyword Search

- ◆ Example keyword query: “presidents of the US”
 - T1(**President, Vice President**)
 - T2(**President**, Term, Party, **Vice President**)
 - T3(State, Governor, Party, Term)
 - T4(State, Senator, Party, Term, Born On)
 - T5(Chief Justice, Nominated By, Term)

- ◆ $\text{pmi}(a,b) = \log_2(p(a,b)/(p(a)*p(b)))$
 - $\text{pmi}(\text{President, Vice President}) = \log_2(0.4/(0.4 * 0.4)) = 1.32$

WebTables: Keyword Search

- ◆ Example keyword query: “presidents of the US”
 - T1(**President**, Vice President)
 - T2(**President**, **Term**, Party, Vice President)
 - T3(State, Governor, Party, **Term**)
 - T4(State, Senator, Party, **Term**, Born On)
 - T5(Chief Justice, Nominated By, **Term**)

- ◆ $\text{pmi}(a,b) = \log_2(p(a,b)/(p(a)*p(b)))$
 - $\text{pmi}(\text{President}, \text{Vice President}) = \log_2(0.4/(0.4 * 0.4)) = 1.32$
 - $\text{pmi}(\text{President}, \text{Term}) = \log_2(0.2/(0.4*0.8)) = -0.68$

WebTables: Keyword Search

- ◆ Example keyword query: “presidents of the US”
 - T1(President, Vice President) **1**
 - T2(President, Term, Party, Vice President) **2**
 - T3(State, Governor, Party, Term)
 - T4(State, Senator, Party, Term, Born On)
 - T5(Chief Justice, Nominated By, Term)

- ◆ Schema coherency = average $\text{pmi}(a,b)$ over all a, b in $\text{attrs}(R)$
 - coherency(T1) = $\text{avg}(\{1.32\}) = 1.32$
 - coherency(T2) = $\text{avg}(\{1.32, -0.68, -0.26, 0.32, -0.26, -0.68\}) = -0.15$

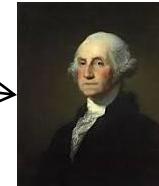
Annotating Web Tables [LSCI0]

- ◆ Goal: given a Web table, which entities occur in which cells, what are the column types, and the relationships between columns?
- ◆ Why is this challenging?
 - Text in table cells often mention entities, but can be ambiguous
 - Column headers, if present, do not use controlled vocabulary
- ◆ Benefits of solving this problem
 - Permits use of relational, metadata-aware queries on Web tables
 - Extracts knowledge from Web tables

Annotating Web Tables: Entities

- ◆ Goal: given a Web table, which entities occur in which cells, what are the column types, and the relationships between columns?

| | |
|---|----------------------------------|
| George Washington (1732-1799) | John Adams |
| John Adams (1735-1826) | Thomas Jefferson |
| Thomas Jefferson (1743-1826) | Aaron Burr, George Clinton |
| James Madison (1751-1836) | George Clinton, Elbridge Gerry |
| James Monroe (1758-1831) | Daniel Tompkins |
| John Quincy Adams (1767-1848) | John Calhoun |
| Andrew Jackson (1767-1845) | John Calhoun, Martin van Buren |
| Martin van Buren (1782-1862) | Richard Johnson |
| William H. Harrison (1773-1841) | John Tyler |
| . John Tyler (1790-1862) | |
| . James K. Polk (1795-1849) | George Dallas |
| . Zachary Taylor (1784-1850) | Millard Fillmore |



Annotating Web Tables: Column Types

- ◆ Goal: given a Web table, which entities occur in which cells, what are the column types, and the relationships between columns?

| | |
|---|----------------------------------|
| George Washington (1732-1799) | John Adams |
| John Adams (1735-1826) | Thomas Jefferson |
| Thomas Jefferson (1743-1826) | Aaron Burr, George Clinton |
| James Madison (1751-1836) | George Clinton, Elbridge Gerry |
| James Monroe (1758-1831) | Daniel Tompkins |
| John Quincy Adams (1767-1848) | John Calhoun |
| Andrew Jackson (1767-1845) | John Calhoun, Martin van Buren |
| Martin van Buren (1782-1862) | Richard Johnson |
| William H. Harrison (1773-1841) | John Tyler |
| . John Tyler (1790-1862) | |
| . James K. Polk (1795-1849) | George Dallas |
| Zachary Taylor (1784-1850) | Millard Fillmore |

US Politician

Annotating Web Tables: Relationships

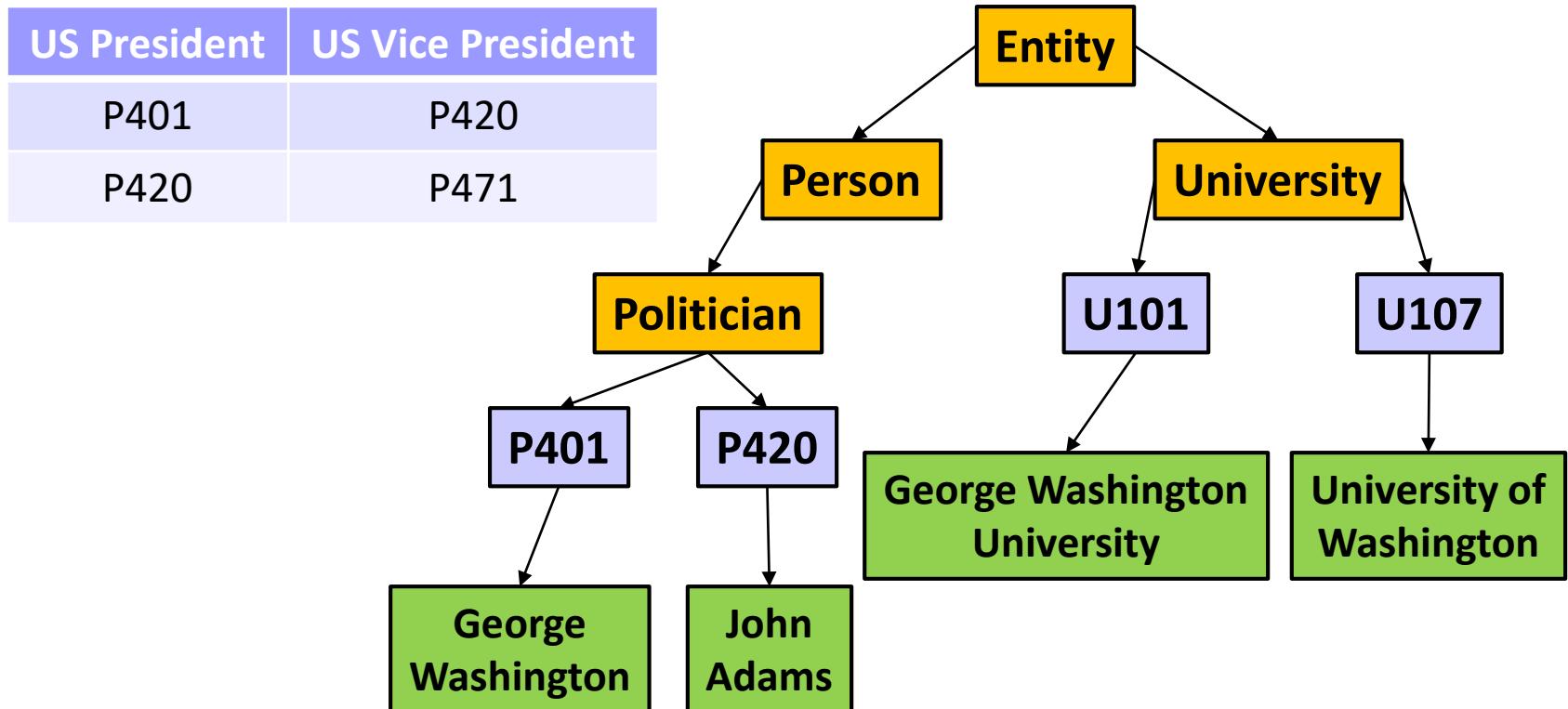
- ◆ Goal: given a Web table, which entities occur in which cells, what are the column types, and the relationships between columns?

| | |
|---|----------------------------------|
| George Washington (1732-1799) | John Adams |
| John Adams (1735-1826) | Thomas Jefferson |
| Thomas Jefferson (1743-1826) | Aaron Burr, George Clinton |
| James Madison (1751-1836) | George Clinton, Elbridge Gerry |
| James Monroe (1758-1831) | Daniel Tompkins |
| John Quincy Adams (1767-1848) | John Calhoun |
| Andrew Jackson (1767-1845) | John Calhoun, Martin van Buren |
| Martin van Buren (1782-1862) | Richard Johnson |
| William H. Harrison (1773-1841) | John Tyler |
| . John Tyler (1790-1862) | |
| . James K. Polk (1795-1849) | George Dallas |
| . Zachary Taylor (1784-1850) | Millard Fillmore |

US President – US Vice President

Annotating Web Tables: Using a Catalog

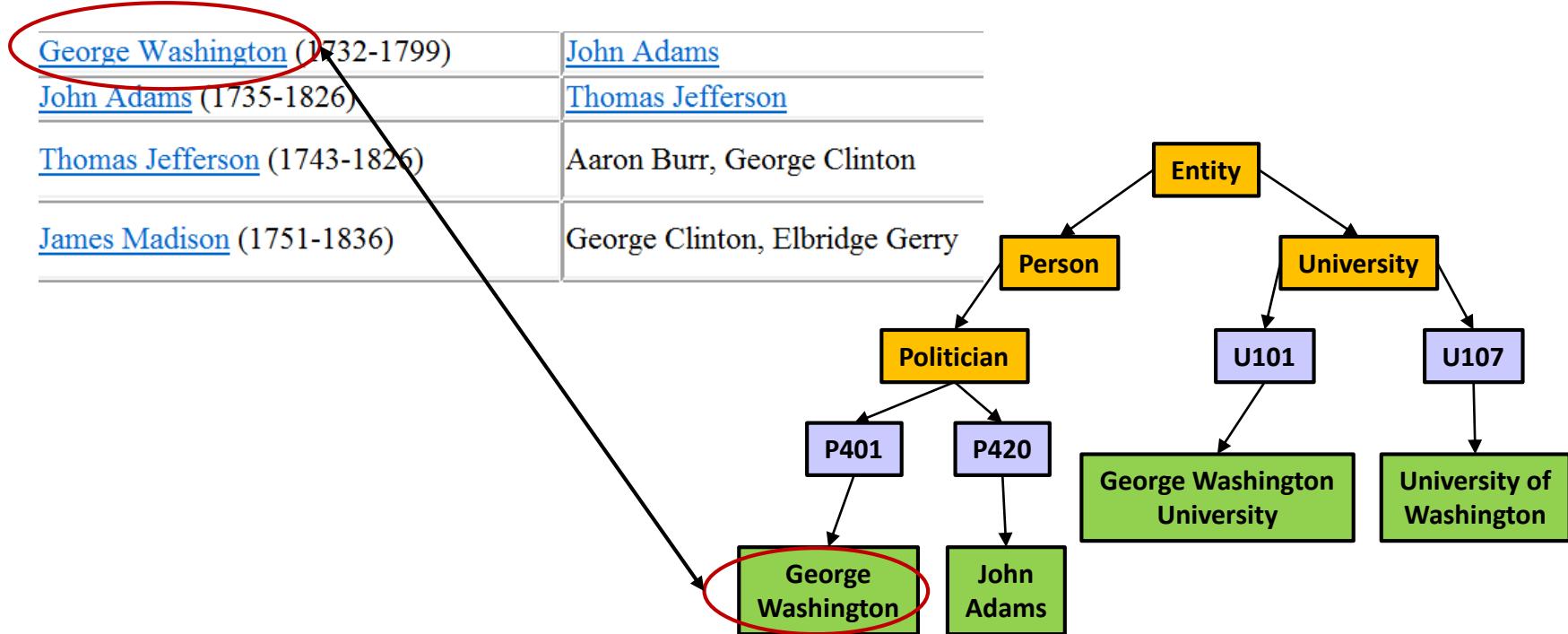
- ◆ A catalog consists of a type hierarchy, entities that are instances of (possibly multiple) types, and binary relationships





Annotating Web Tables: Using a Catalog

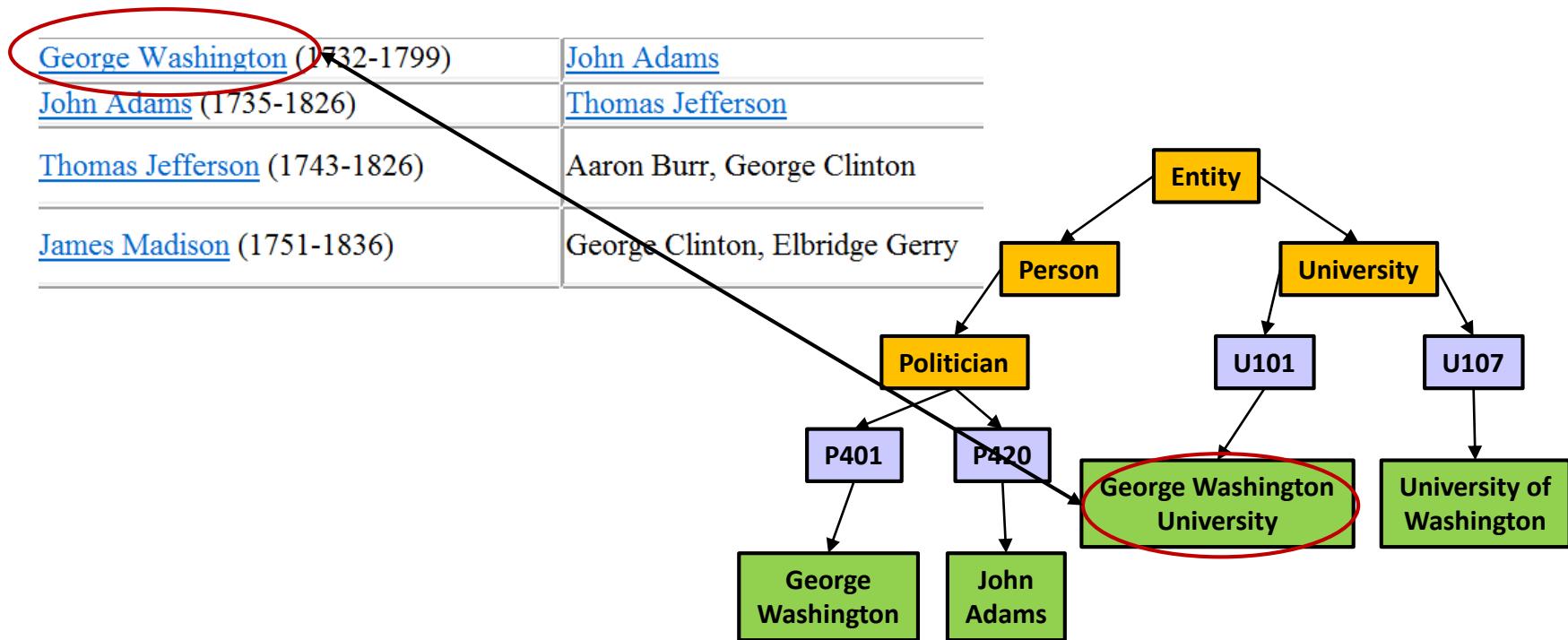
- ◆ How good is it to label cell (r, c), containing text D_{rc} , with entity E?
 - Similarity between D_{rc} and $L(E)$





Annotating Web Tables: Using a Catalog

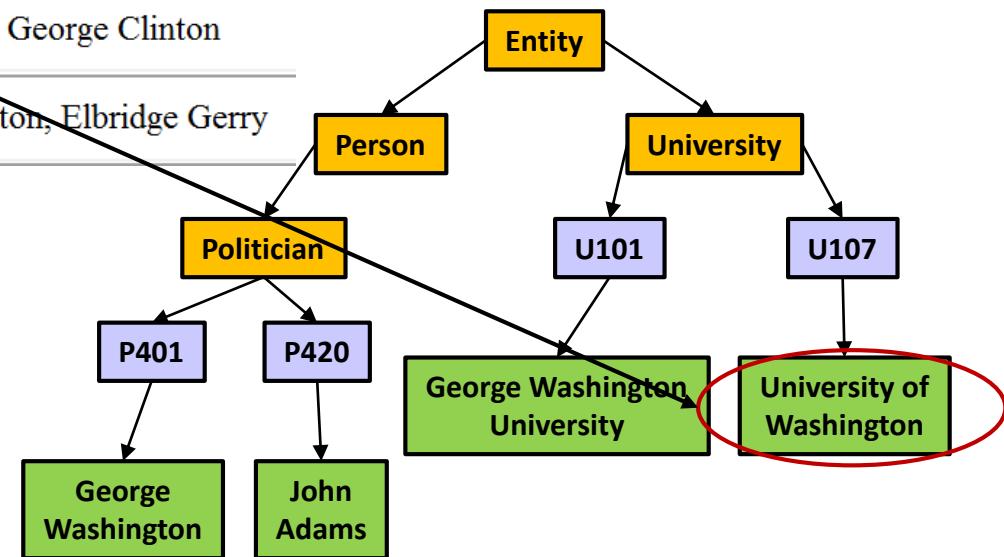
- ◆ How good is it to label cell (r, c), containing text D_{rc} , with entity E?
 - Similarity between D_{rc} and $L(E)$



Annotating Web Tables: Using a Catalog

- ◆ How good is it to label cell (r, c), containing text D_{rc} , with entity E?
 - Similarity between D_{rc} and $L(E)$

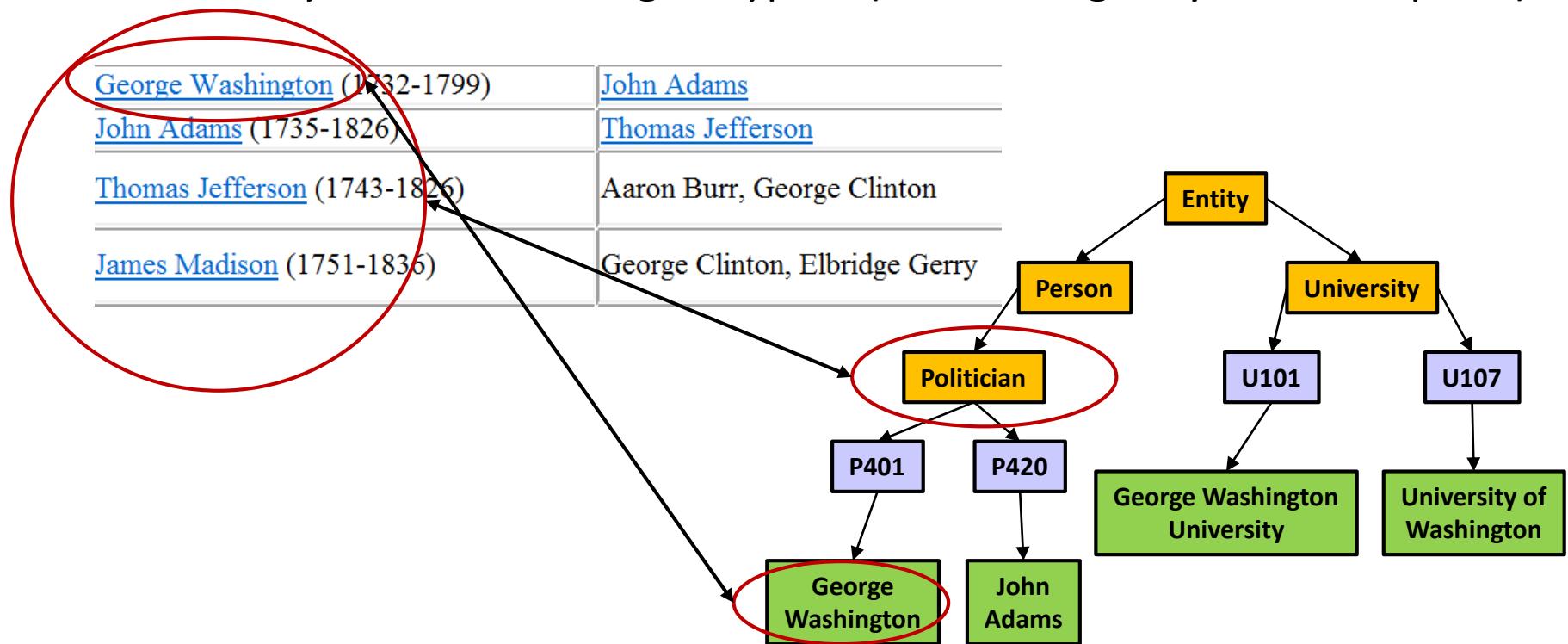
| | |
|---|----------------------------------|
| George Washington (1732-1799) | John Adams |
| John Adams (1735-1826) | Thomas Jefferson |
| Thomas Jefferson (1743-1826) | Aaron Burr, George Clinton |
| James Madison (1751-1836) | George Clinton, Elbridge Gerry |





Annotating Web Tables: Using a Catalog

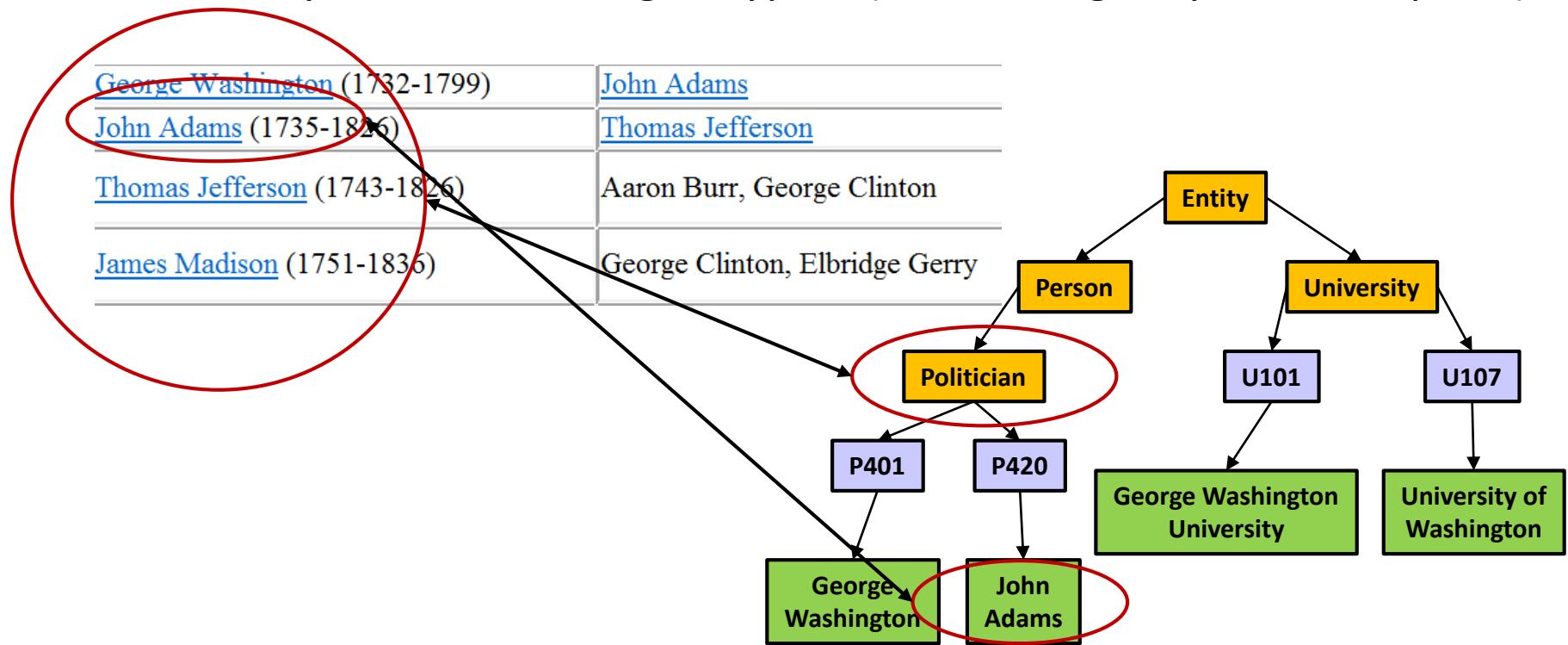
- ◆ How good is it to label column c with type T and cell (r, c) with E?
 - Entity E should belong to type T (but catalog may be incomplete)





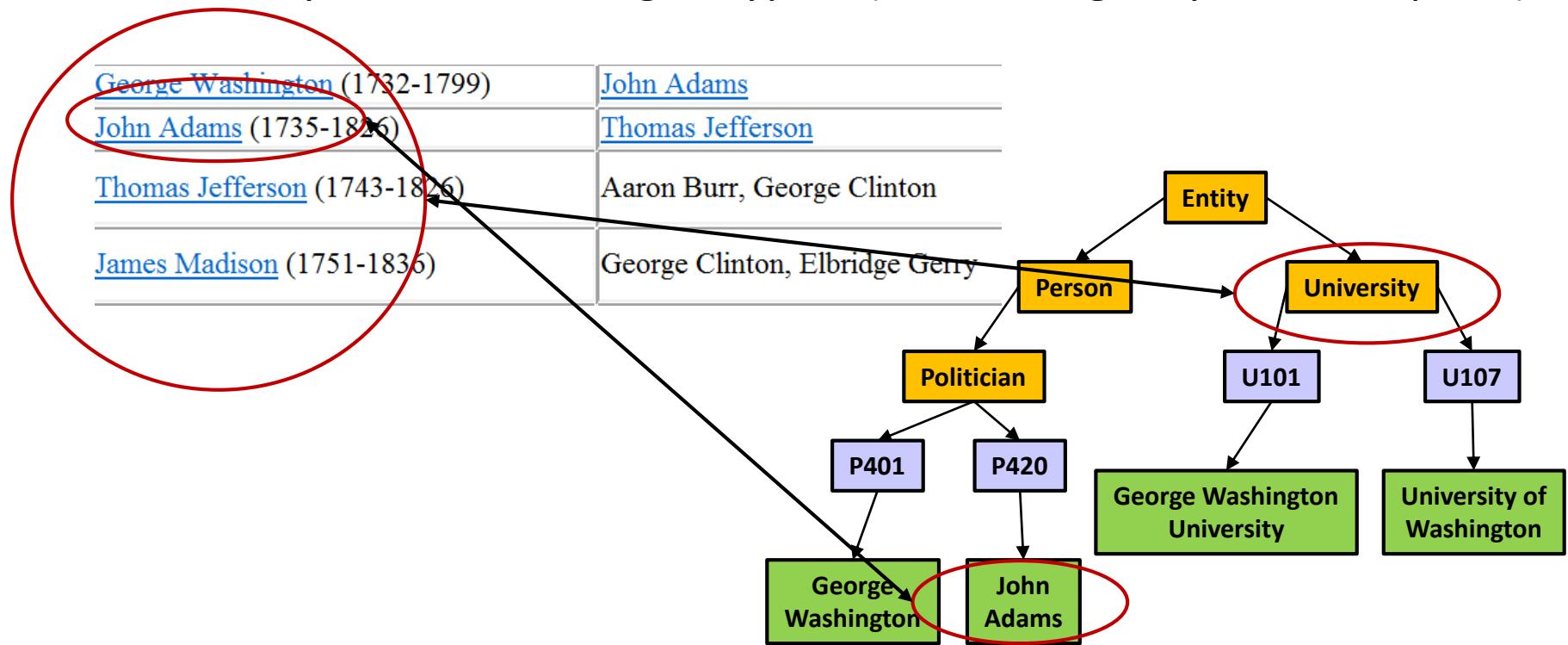
Annotating Web Tables: Using a Catalog

- ♦ How good is it to label column c with type T and cell (r, c) with E?
 - Entity E should belong to type T (but catalog may be incomplete)



Annotating Web Tables: Using a Catalog

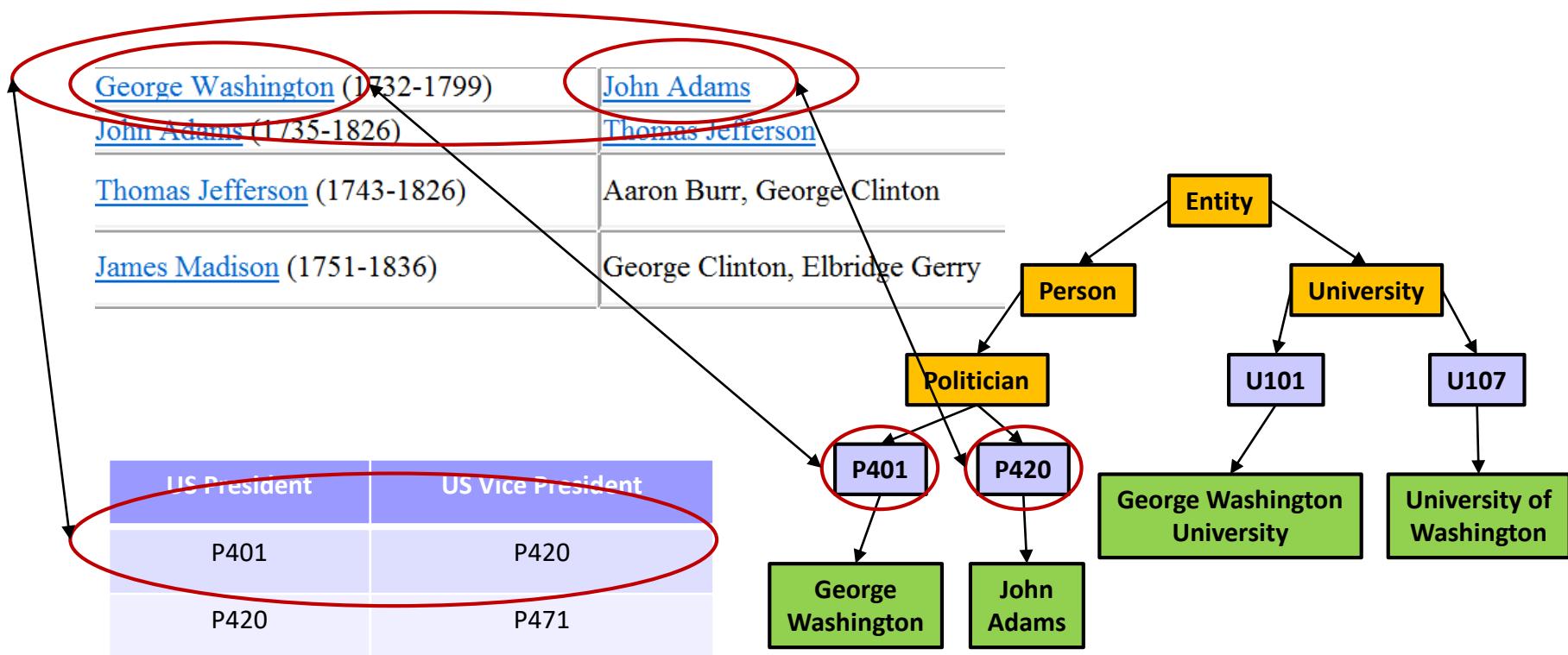
- ♦ How good is it to label column c with type T and cell (r, c) with E?
 - Entity E should belong to type T (but catalog may be incomplete)





Annotating Web Tables: Using a Catalog

- ◆ Do entity annotations e_{rc} for cell (r, c) and $e_{rc'}$ for cell (r, c') vote for or against annotating column pair (c, c') with relation R?

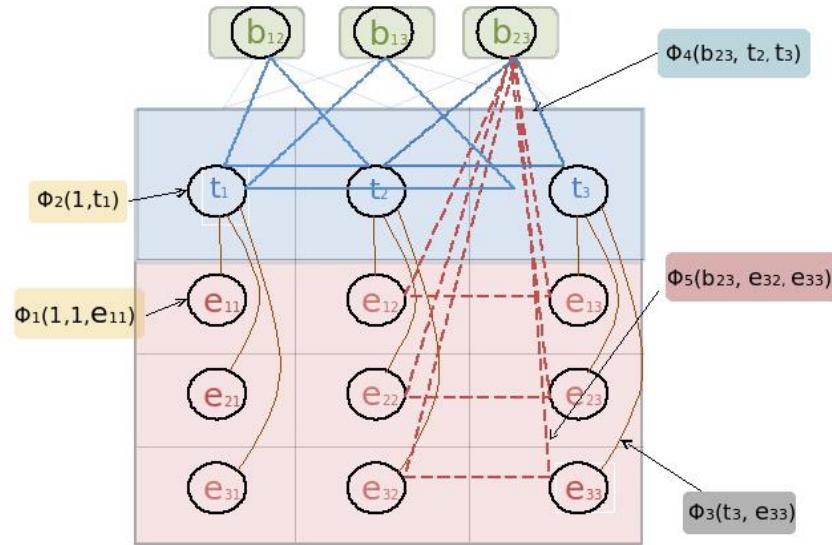


Annotating Web Tables: Using a Catalog

- ◆ Model table annotation using interrelated random variables, represented by a probabilistic graphical model
 - Cell text (in Web table) and entity label (in catalog)
 - Column header (in Web table) and type label (in catalog)
 - Column type and cell entity (in Web table)
 - Pair of column types (in Web table) and relation (in catalog)
 - Entity pairs (in Web table) and relation (in catalog)

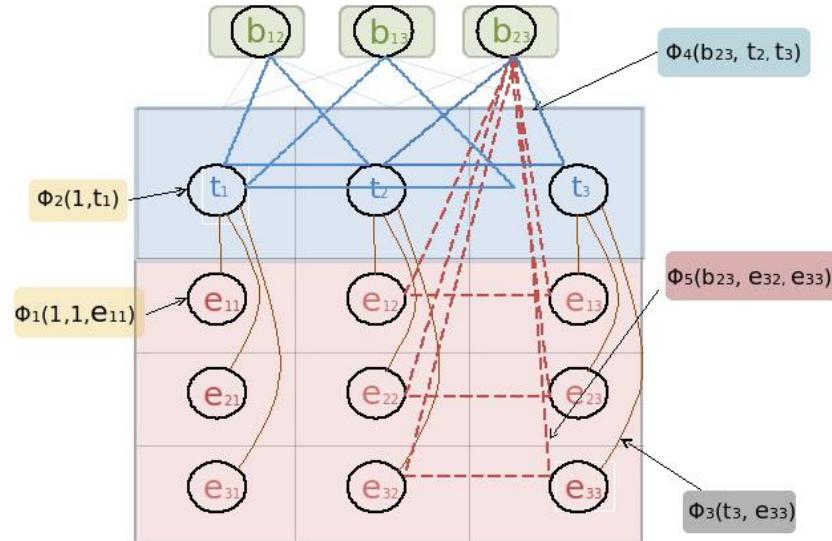
Annotating Web Tables: Using a Catalog

- ◆ Model table annotation using interrelated random variables, represented by a probabilistic graphical model
 - Cell text (in Web table) and entity label (in catalog)
 - Column header (in Web table) and type label (in catalog)
 - Column type and cell entity (in Web table)



Annotating Web Tables: Using a Catalog

- ◆ Model table annotation using interrelated random variables, represented by a probabilistic graphical model
 - Pair of column types (in Web table) and relation (in catalog)
 - Entity pairs (in Web table) and relation (in catalog)



Annotating Web Tables: Using a Catalog

- ◆ Model table annotation using interrelated random variables, represented by a probabilistic graphical model
 - Cell text (in Web table) and entity label (in catalog)
 - Column header (in Web table) and type label (in catalog)
 - Column type and cell entity (in Web table)
 - Pair of column types (in Web table) and relation (in catalog)
 - Entity pairs (in Web table) and relation (in catalog)
- ◆ Task of annotation amounts to searching for an assignment of values to the variables that maximizes the joint probability
 - Problem is NP-hard in the general case
 - Use iterative belief propagation in “factor graphs” until convergence



Finding Related Tables [DFG+12]

- ♦ Motivation: given a table T and a corpus C of tables, find tables T' in C that can be integrated with T to augment T's information

| President | Vice President |
|-------------------------------|--|
| George Washington (1789-1797) | John Tyler (1841-1845) none (1841-1845) |
| John Adams (1797-1801) | James K. Polk (1845-1849) George M. Dallas (1845-1849) |
| Thomas Jefferson (1801-1809) | Zachary Taylor (1849-1850) Millard Fillmore (1849-1850) |
| | Millard Fillmore (1850-1853) none (1850-1853) |
| James Madison (1809-1817) | Franklin Pierce (1853-1857) William King (1853) none (1853-1857) |
| | James Buchanan (1857-1861) Abraham Lincoln (1861-1865) John C. Breckinridge (1857-1861) Hannibal Hamlin (1861-1865) |
| James Monroe (1817-1825) | Andrew Johnson (1865) none (1865-1869) |
| John Quincy Adams (1825-1829) | Ulysses S. Grant (1869-1877) Schuyler Colfax (1869-1873) |
| Andrew Jackson (1829-1837) | Henry Wilson (1873-1875) none (1875-1877) |
| | Rutherford B. Hayes (1877-1881) William Wheeler (1877-1881) |
| Martin Van Buren (1837-1841) | James A. Garfield (1881) Chester Arthur (1881) |
| William Henry Harrison (1841) | Chester Arthur (1881-1885) none (1881-1885) |
| | Grover Cleveland (1885-1889) Thomas Hendricks (1885) none (1885-1889) |

Finding Related Tables

- ♦ Motivation: given a table T and a corpus C of tables, find tables T' in C that can be integrated with T to augment T's information

| | | |
|-------------------------------|--------------------------------|--------------------------------|
| President | James Madison (1809-1817) | George Clinton (1809-1812) |
| George Washington (1789-1797) | | none (1812-1813) |
| John Adams (1797-1801) | | Elbridge Gerry (1813-1814) |
| Thomas Jefferson (1801-1809) | James Monroe (1817-1825) | none (1814-1817) |
| | John Quincy Adams (1825-1829) | Daniel D. Tompkins (1817-1825) |
| James Madison (1809-1817) | Andrew Jackson (1829-1837) | John C. Calhoun (1825-1829) |
| | | John C. Calhoun (1829-1832) |
| | | none (1832-1833) |
| | | Martin Van Buren (1833-1837) |
| | none (1814-1817) | |
| James Monroe (1817-1825) | Daniel D. Tompkins (1817-1825) | |
| John Quincy Adams (1825-1829) | John C. Calhoun (1825-1829) | |
| Andrew Jackson (1829-1837) | John C. Calhoun (1829-1832) | |
| | none (1832-1833) | |
| | Martin Van Buren (1833-1837) | |
| Martin Van Buren (1837-1841) | Richard M. Johnson (1837-1841) | |
| William Henry Harrison (1841) | John Tyler (1841) | |



Finding Related Tables

- ♦ Motivation: given a table T and a corpus C of tables, find tables T' in C that can be integrated with T to augment T's information

| President | Vice President | | | | | |
|-------------------------------|---|---|---------------|------------------------|-----------------------|--|
| | John Adams (1789-1797) | | | | | |
| | President | Took office | Left office | Party | Term [n 1] | Previous office |
| George Washington (1789-1797) | George Washington (1732-1799) [11][12][13] | April 30, 1789 | March 4, 1797 | Independent [14] | 1 (1789) | Commander-in-Chief of the Continental Army (1775-1783) |
| John Adams (1797-1801) | | | | | 2 (1792) | |
| Thomas Jefferson (1801-1809) | | | | | | |
| James Madison (1809-1817) |  | John Adams (1735-1826) [15][16][17] | March 4, 1797 | March 4, 1801 [n 2] | Federalist | 3 (1796) |
| James Monroe (1817-1825) |  | | | | | Vice President |
| John Quincy Adams (1825-1829) | | | | | | |
| Andrew Jackson (1829-1837) |  | Thomas Jefferson (1743-1826) [18][19][20] | March 4, 1801 | March 4, 1809 | Democratic-Republican | 4 (1800) |
| Martin Van Buren (1837-1841) | | | | | | Vice President |
| William Henry Harrison (1841) |  | James Madison (1751-1836) [21][22][23] | March 4, 1809 | March 4, 1817 | Democratic-Republican | 5 (1804) |
| | | | | | | |
| | | | | | | 6 (1808) |
| | | | | | | Secretary of State (1801-1809) |
| | | | | | | 7 |

Finding Related Tables

- ◆ Motivation: given a table T and a corpus C of tables, find tables T' in C that can be integrated with T to augment T's information

| President | Vice President | | | | | | |
|------------------------------------|---|---------------|------------------------|--|--------------|--|--|
| George Washington (1789-1797) | John Adams (1789-1797) | | | | | | |
| John Adams (1797-1801) | Thomas Jefferson (1797-1801) | | | | | | |
| Thomas Jefferson (1801-1809) | John Tyler (1790-1862) [39][40][41] | April 4, 1841 | March 4, 1845 | Whig April 4, 1841 – September 13, 1841 | (1840) | Vice President | |
| James Madison (1809-1817) | | | | no party ^[n 7] September 13, 1841 – March 4, 1845 | | | |
| James Monroe (1817-1829) | James K. Polk (1795-1849) [42][43][44] | March 4, 1845 | March 4, 1849 | Democratic | 15 (1844) | Governor of Tennessee (1839-1841) | |
| John Quincy Adams (1829-1837) | Zachary Taylor (1784-1850) [45][46][47] | March 4, 1849 | July 9, 1850 [n 3] | Whig | | U.S. Army Major general (1846-1849) | |
| Andrew Jackson (1829-1837) | | | | | 16 (1848) | | |
| Martin Van Buren (1837-1845) | Millard Fillmore (1800-1874) [48][49][50] | July 9, 1850 | March 4, 1853 [n 8] | Whig | | Vice President | |
| William Henry Harrison (1841-1845) | | | | | | | |

Finding Related Tables

- ♦ Motivation: given a table T and a corpus C of tables, find tables T' in C that can be integrated with T to augment T's information

| President | Vice President |
|------------------------------------|---|
| George Washington (1789-1797) | John Adams (1789-1797) |
| John Adams (1797-1801) | Thomas Jefferson (1801-1809) |
| Thomas Jefferson (1801-1809) | James Madison (1809-1817) |
| James Madison (1809-1817) | Shenandoah  Shenandoah's Blue Ridge Mountains are covered by hardwood forests that are home to tens of thousands of animals. The Skyline Drive and Appalachian Trail run the entire length of this narrow park that has more than 500 miles (800 km) of hiking trails along scenic overlooks and waterfalls of the Shenandoah River. ^[57] |
| James Monroe (1817-1825) | Theodore Roosevelt  This region that enticed and influenced President Theodore Roosevelt is now a park of three units in the badlands. Besides Roosevelt's historic cabin, there are scenic drives and backcountry hiking opportunities. Wildlife includes American Bison, pronghorns, Bighorn sheep, and wild horses. ^[58] |
| John Quincy Adams (1825-1829) | North Dakota  North Dakota is a state in the Great Plains of the United States. It is the 19th largest state by area, and the 5th least populous. |
| Andrew Jackson (1829-1837) | United States Virgin Islands  The United States Virgin Islands are a territory of the United States located in the Caribbean Sea. They consist of several islands and cays, including St. Thomas, St. John, and St. Croix. |
| Martin Van Buren (1837-1841) | August 2, 1956  Voyageurs National Park is a national park located in the state of Minnesota, United States. It is known for its numerous lakes and rivers, as well as its boreal forest ecosystem. |
| William Henry Harrison (1841-1845) | January 8, 1971  Minnesota is a state in the Upper Midwest region of the United States. It is known for its large lakes, forests, and cold climate. |

Finding Related Tables

- ◆ Motivation: given a table T and a corpus C of tables, find tables T' in C that can be integrated with T to augment T 's information
- ◆ Examples of related tables
 - Tables that are candidates for union, and add new entities
 - Tables that are candidates for join, and add new attributes

Finding Related Tables

- ◆ Motivation: given a table T and a corpus C of tables, find tables T' in C that can be integrated with T to augment T 's information
- ◆ More generally:
 - Are tables T and T' the results of applying queries Q and Q' on U ?
 - Are Q and Q' different, but have a similar select-project structure?
 - Is virtual table U coherent?
- ◆ Problem: Find top- k tables with highest relatedness scores to T

Finding Related Tables: Entity Complement

- ◆ Goal: Find top-k tables T' that are candidates for union with T
- ◆ Methodology
 - Entity consistency: T' should have the same type of entities as T
 - Entity expansion: T' should substantially add new entities to T
 - Schema consistency: T and T' should have similar schemas

Finding Related Tables: Entity Complement

- ◆ Goal: Find top-k tables T' that are candidates for union with T
 - Entity consistency, entity expansion

| President | Vice President |
|-------------------------------|------------------------------|
| George Washington (1789-1797) | John Adams (1789-1797) |
| John Adams (1797-1801) | |
| Thomas Jefferson (1801-1809) | John Tyler (1841-1845) |
| James Madison (1809-1817) | James K. Polk (1845-1849) |
| James Monroe (1817-1825) | Zachary Taylor (1849-1850) |
| John Quincy Adams (1825-1829) | Millard Fillmore (1850-1853) |
| Andrew Jackson (1829-1837) | Franklin Pierce (1853-1857) |
| Martin Van Buren (1837-1841) | James Buchanan (1857-1861) |
| William Henry Harrison (1841) | Abraham Lincoln (1861-1865) |
| | Andrew Johnson (1865) |
| | none (1865-1869) |
| | Schuyler Colfax (1869-1873) |
| | Henry Wilson (1873-1875) |
| | none (1875-1877) |
| | William Wheeler (1877-1881) |
| | Chester Arthur (1881) |
| | none (1881-1885) |
| | Thomas Hendricks (1885) |
| | none (1885-1889) |



Finding Related Tables: Entity Complement

- ◆ Goal: Find top-k tables T' that are candidates for union with T
 - Schema consistency

| President | Vice President |
|-------------------------------|---------------------------------|
| George Washington (1789-1797) | John Adams (1789-1797) |
| John Adams (1797-1801) | John Tyler (1841-1845) |
| Thomas Jefferson (1801-1809) | James K. Polk (1845-1849) |
| James Madison (1809-1817) | Zachary Taylor (1849-1850) |
| | Millard Fillmore (1850-1853) |
| | Franklin Pierce (1853-1857) |
| James Monroe (1817-1825) | James Buchanan (1857-1861) |
| John Quincy Adams (1825-1829) | Abraham Lincoln (1861-1865) |
| Andrew Jackson (1829-1837) | Andrew Johnson (1865-1869) |
| | Ulysses S. Grant (1869-1877) |
| Martin Van Buren (1837-1841) | Rutherford B. Hayes (1877-1881) |
| William Henry Harrison (1841) | James A. Garfield (1881) |
| | Chester Arthur (1881-1885) |
| | Grover Cleveland (1885-1889) |

Finding Related Tables: Entity Complement

- ◆ Goal: Find top- k tables T' that are candidates for union with T
- ◆ [DFG+12] use three signals to ensure entity complement tables
 - WebIsA: noisy database of entities (155M) and types (1.5M)
 - Freebase: curated database of entities (16M) and types (600K)
 - WebTable labels: count co-occurrence of entities in Web tables
- ◆ Relatedness score:
 - Use weighted Jaccard similarity on label sets for entity consistency
 - Use bipartite max-weight matching for schema consistency

Finding Related Tables: Schema Complement

- ◆ Goal: Find top-k tables T' that are candidates for join with T
- ◆ Methodology
 - Coverage of entity set: T' should contain most of T 's entities
 - Coherent schema expansion: use the ACSDb to measure the maximum benefit that a subset of attributes of T' can provide to T
- ◆ Recall, ACSDb(Schema, Count) can be used for schema coherency

Finding Related Tables: Schema Complement

- ◆ Goal: Find top-k tables T' that are candidates for join with T
 - Entity coverage

| President | Vice President | | | | | | |
|-------------------------------|---|----------------|------------------------|-----------------------|-------------|--|--|
| | President | Took office | Left office | Party | Term [n 1] | Previous office | |
| George Washington (1789-1797) | John Adams (1789-1797) | | | | 1 (1789) | | |
| John Adams (1797-1801) | George Washington (1732-1799) [11][12][13] | April 30, 1789 | March 4, 1797 | Independent [14] | 2 (1792) | Commander-in-Chief of the Continental Army (1775-1783) | |
| Thomas Jefferson (1801-1809) | John Adams (1735-1826) [15][16][17] | March 4, 1797 | March 4, 1801 [n 2] | Federalist | 3 (1796) | Vice President | |
| James Madison (1809-1817) | Thomas Jefferson (1743-1826) [18][19][20] | March 4, 1801 | March 4, 1809 | Democratic-Republican | 4 (1800) | Vice President | |
| James Monroe (1817-1825) | James Madison (1751-1836) [21][22][23] | March 4, 1809 | March 4, 1817 | Democratic-Republican | 5 (1804) | | |
| John Quincy Adams (1825-1829) | | | | | 6 (1808) | | |
| Andrew Jackson (1829-1837) | | | | | 7 | Secretary of State (1801-1809) | |
| Martin Van Buren (1837-1841) | | | | | | | |
| William Henry Harrison (1841) | | | | | | | |



Finding Related Tables: Schema Complement

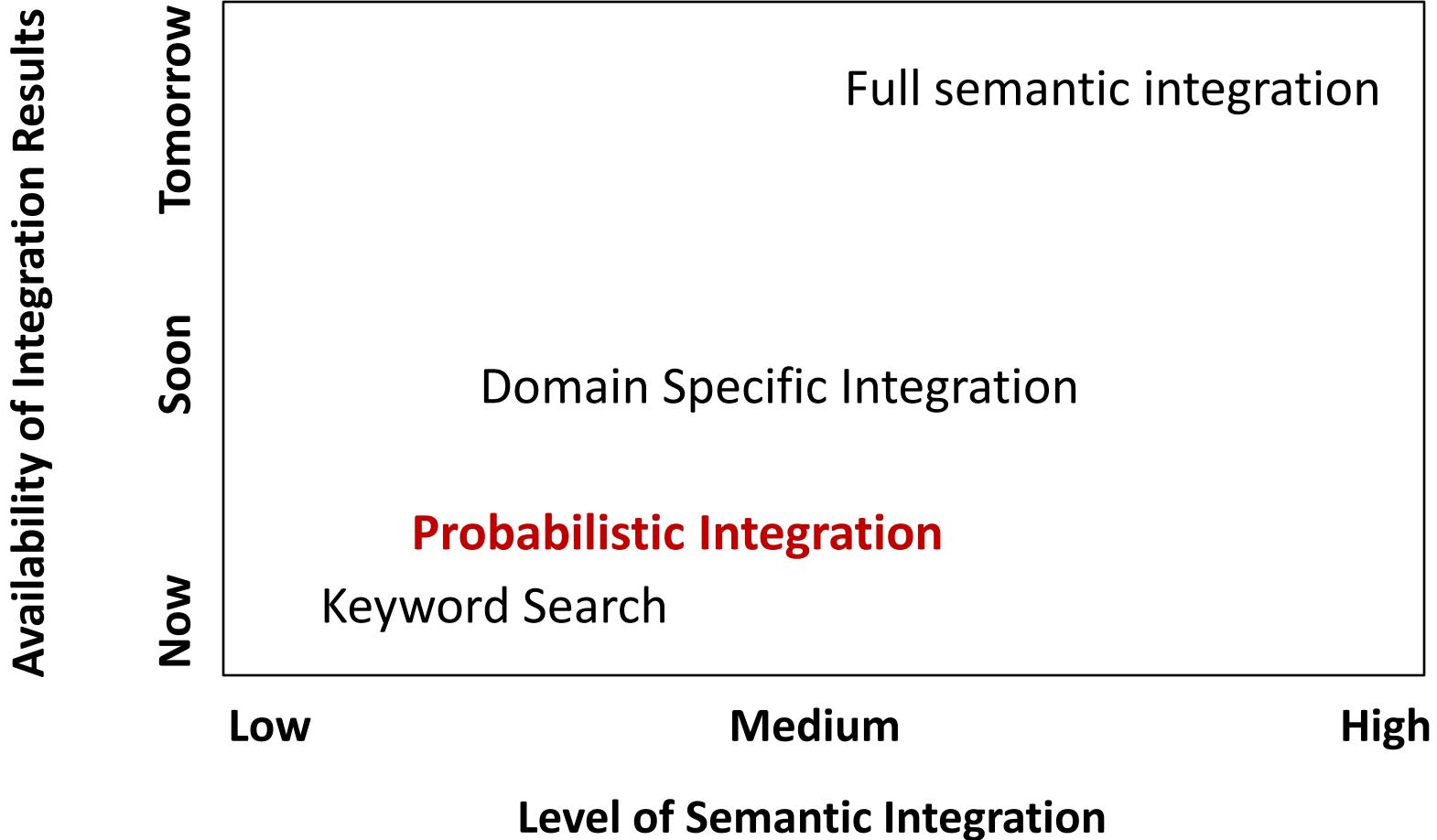
- ◆ Goal: Find top-k tables T' that are candidates for join with T
 - Coherent schema expansion

| President | Vice President | President | Took office | Left office | Party | Term [n 1] | Previous office |
|-------------------------------|------------------------|--|----------------|------------------------|-----------------------|---------------|--|
| George Washington (1789-1797) | John Adams (1789-1797) | George Washington (1732-1799) [11][12][13] | April 30, 1789 | March 4, 1797 | Independent [14] | 1 (1789) | Commander-in-Chief of the Continental Army (1775-1783) |
| John Adams (1797-1801) | | John Adams (1735-1826) [15][16][17] | March 4, 1797 | March 4, 1801 [n 2] | Federalist | 2 (1797) | Vice President |
| Thomas Jefferson (1801-1809) | | Thomas Jefferson (1743-1826) [18][19][20] | March 4, 1801 | March 4, 1809 | Democratic-Republican | 3 (1796) | Vice President |
| James Madison (1809-1817) | | James Madison (1751-1836) [21][22][23] | March 4, 1809 | March 4, 1817 | Democratic-Republican | 4 (1800) | |
| | | | | | | 5 (1804) | |
| James Monroe (1817-1825) | | | | | | 6 (1808) | |
| John Quincy Adams (1825-1829) | | | | | | 7 | Secretary of State (1801-1809) |
| Andrew Jackson (1829-1837) | | | | | | | |
| Martin Van Buren (1837-1841) | | | | | | | |
| William Henry Harrison (1841) | | | | | | | |

Finding Related Tables: Efficiency Issues

- ◆ Naïve approach: compute relatedness score for every table pair
 - Very expensive on large table corpora
- ◆ Key idea: use filters to scale up computation of table relatedness
 - **Fewer comparisons**: use filters as blocking criteria to bucketize tables, and only perform relatedness comparisons within buckets
 - **Faster comparisons**: apply sequence of filters, based on selectivity and computational efficiency of filters
- ◆ Useful filters:
 - Two tables must share entity column name or inferred label
 - Two tables must share at least n entities, $n = 1, 2, 3$

Space of Strategies



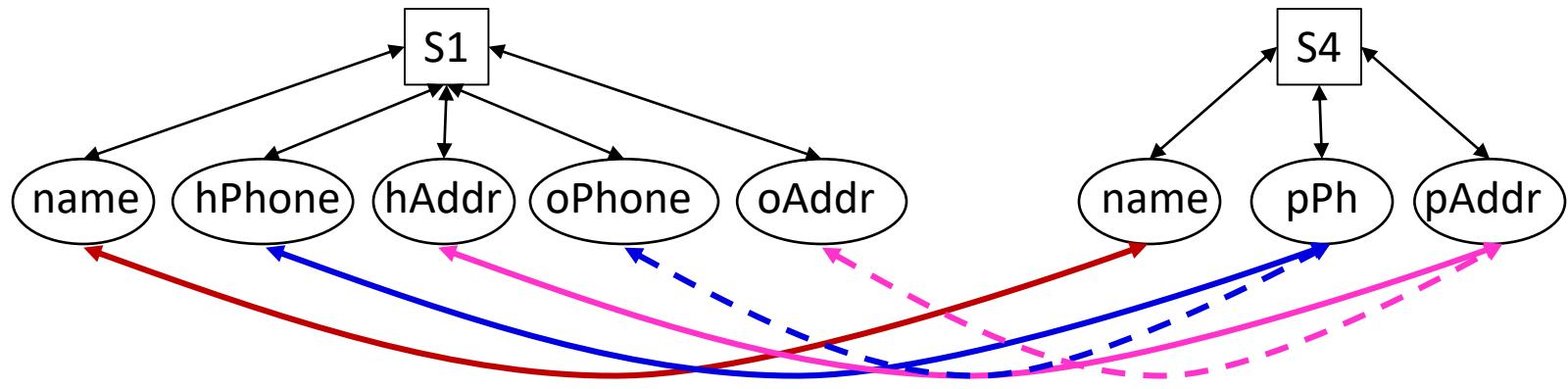
Dataspace Approach [FHM05, HFM06]

- ◆ Motivation: SDI approach (as-is) is infeasible for BDI
 - **Volume, variety** of sources → unacceptable up-front modeling cost
 - **Velocity** of sources → expensive to maintain integration results
- ◆ Key insight: **pay-as-you-go** approach may be feasible
 - Start with simple, universally useful service
 - Iteratively add complexity when and where needed [JFH08]
- ◆ Approach has worked for RDBMS, Web, Hadoop ...

Bootstrapping DI Systems [DDH08]

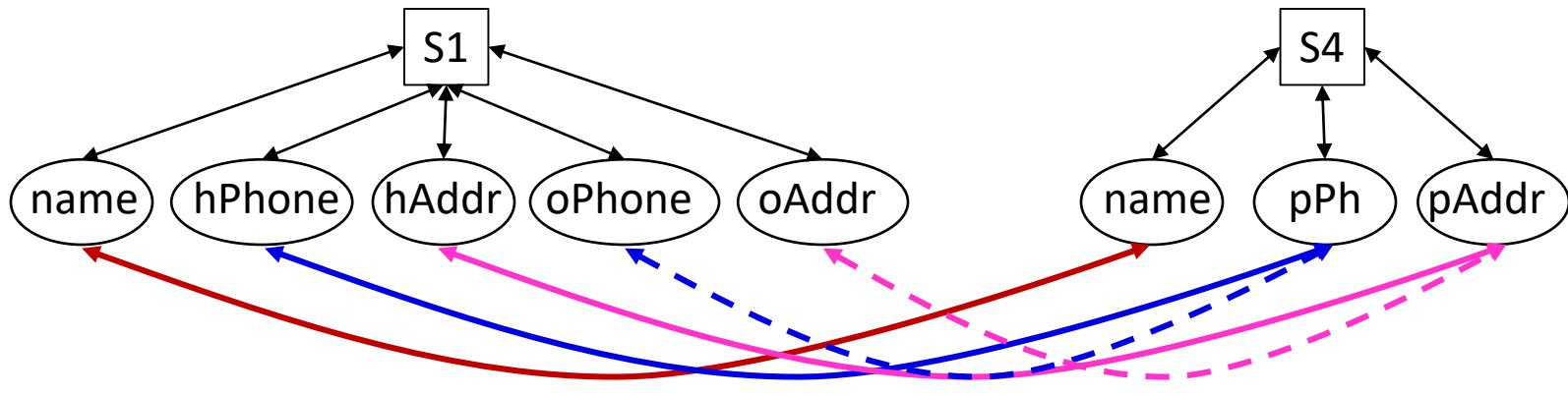
- ◆ Thesis: completely automated data integration is feasible, but ...
 - Need to model uncertainty about semantics of attributes in sources
- ◆ Automatically create a mediated schema from a set of sources
 - Uncertainty → probabilistic mediated schemas
 - P-mediated schemas offer benefits in modeling uncertainty
- ◆ Automatically create mappings from sources to mediated schema
 - Probabilistic mappings use weighted attribute correspondences

Probabilistic Mediated Schemas [DDH08]



- ◆ Mediated schemas: automatically created by inspecting sources
 - Clustering of source attributes
 - **Volume, variety** of sources → uncertainty in accuracy of clustering

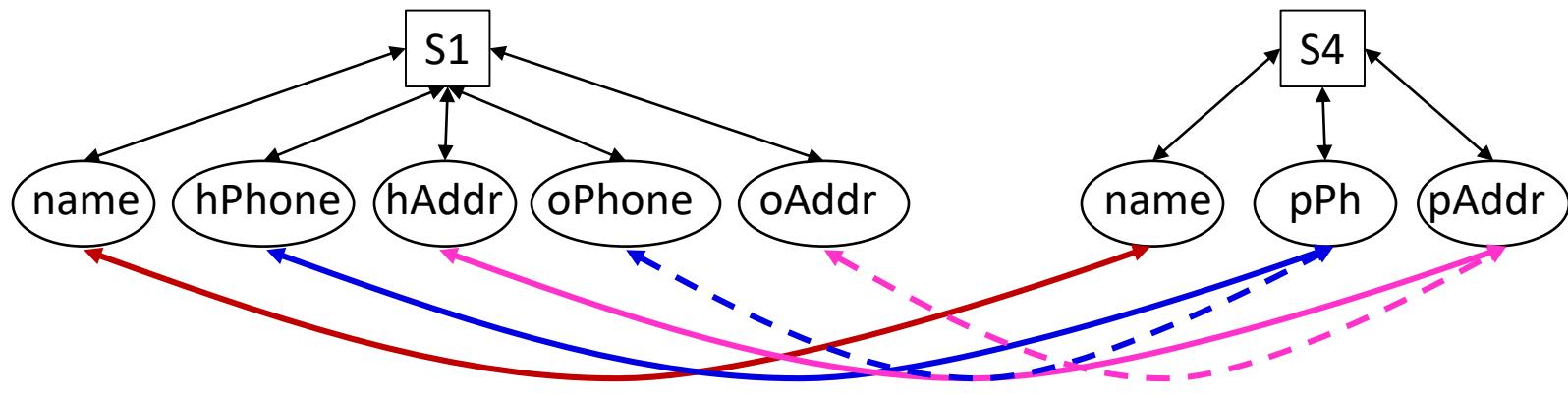
Probabilistic Mediated Schemas [DDH08]



- ◆ Example P-mediated schema MS
 - M1({name}, {hPhone, pPh}, {oPhone}, {hAddr, pAddr}, {oAddr})
 - M2({name}, {hPhone}, {pPh, oPhone}, {hAddr}, {pAddr, oAddr})
 - M3({name}, {hPhone, pPh}, {oPhone}, {hAddr}, {pAddr}, {oAddr})
 - M4({name}, {hPhone}, {pPh, oPhone}, {hAddr}, {pAddr}, {oAddr})
 - MS = {(M1, 0.6), (M2, 0.4)}

Probabilistic Mappings [DHY07, DDH08]

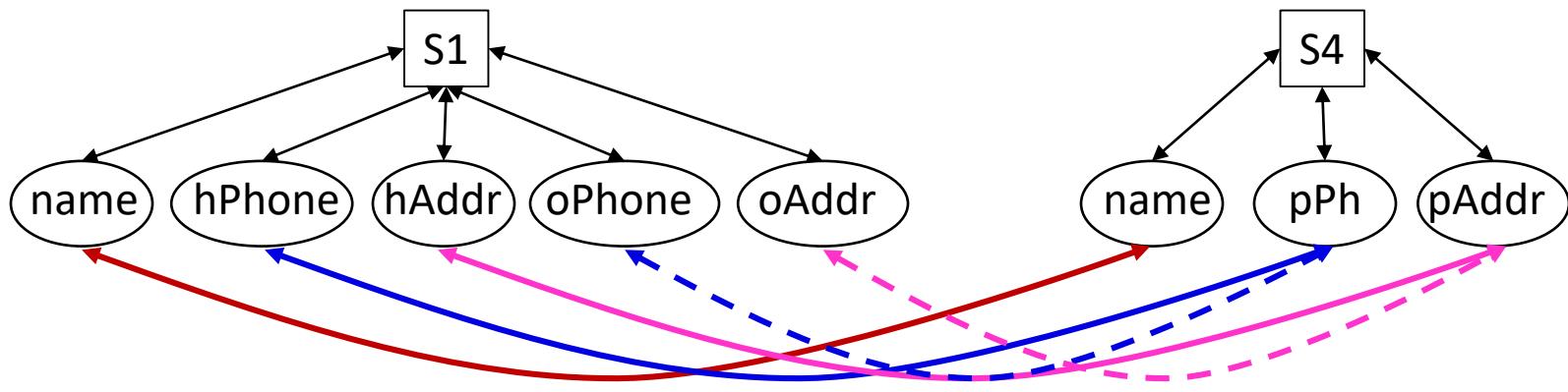
- ◆ Mapping between P-mediated schema and a source schema



- ◆ Example mappings between M_1 and S_1
 - $G_1(\{M_1.n\}, name), \{M_1.phP\}, hPhone), \{M_1.phA\}, hAddr\}, \dots)$
 - $G_2(\{M_1.n\}, name), \{M_1.phP\}, oPhone), \{M_1.phA\}, oAddr\}, \dots)$
 - $G = \{(G_1, 0.6), (G_2, 0.4)\}$

Probabilistic Mappings

- ◆ Mapping between P-mediated schema and a source schema



- ◆ Answering queries on P-mediated schema based on P-mappings
 - By table semantics: one mapping for all tuples in a table
 - By tuple semantics: different mappings are okay in a table

Probabilistic Mappings: By Table Semantics

- ◆ Consider query Q1: SELECT name, pPh, pAddr FROM MS

| S1 | name | hPhone | hAddr | oPhone | oAddr |
|--------|----------|----------|----------|----------|-------|
| Ken | 111-1111 | New York | 222-2222 | Summit | |
| Barbie | 333-3333 | Summit | 444-4444 | New York | |

- ◆ Result of Q1, under by table semantics, in a possible world
 - G1({**M1.n**, name}, {**M1.phP**, hPhone}, {**M1.phA**, hAddr}, ...)

| Q1R (Prob = 0.60) | name | pPh | pAddr | Map |
|----------------------|----------|----------|-------|-----|
| Ken | 111-1111 | New York | G1 | |
| Barbie | 333-3333 | Summit | G1 | |

Probabilistic Mappings: By Table Semantics

- ◆ Consider query Q1: SELECT name, pPh, pAddr FROM MS

| S1 | name | hPhone | hAddr | oPhone | oAddr |
|--------|----------|----------|----------|----------|-------|
| Ken | 111-1111 | New York | 222-2222 | Summit | |
| Barbie | 333-3333 | Summit | 444-4444 | New York | |

- ◆ Result of Q1, under by table semantics, in a possible world
 - G2({**M1.n**, name}, {**M1.phP**, oPhone}, {**M1.phA**, oAddr}, ...)

| Q1R (Prob = 0.40) | name | pPh | pAddr | Map |
|----------------------|----------|----------|-------|-----|
| Ken | 222-2222 | Summit | G2 | |
| Barbie | 444-4444 | New York | G2 | |

Probabilistic Mappings: By Table Semantics

- ◆ Now consider query Q2: SELECT pAddr FROM MS

| S1 | name | hPhone | hAddr | oPhone | oAddr |
|--------|----------|----------|----------|----------|-------|
| Ken | 111-1111 | New York | 222-2222 | Summit | |
| Barbie | 333-3333 | Summit | 444-4444 | New York | |

- ◆ Result of Q2, under by table semantics, across all possible worlds

| Q2R | pAddr | Prob |
|----------|-------|------|
| Summit | 1.0 | |
| New York | 1.0 | |

Probabilistic Mappings: By Tuple Semantics

- ◆ Consider query Q1: SELECT name, pPh, pAddr FROM MS

| S1 | name | hPhone | hAddr | oPhone | oAddr |
|--------|----------|----------|----------|----------|-------|
| Ken | 111-1111 | New York | 222-2222 | Summit | |
| Barbie | 333-3333 | Summit | 444-4444 | New York | |

- ◆ Result of Q1, under by tuple semantics, in a possible world
 - G1({**M1.n**, name}, {**M1.phP**, hPhone}, {**M1.phA**, hAddr}, ...)
 - G2({**M1.n**, name}, {**M1.phP**, oPhone}, {**M1.phA**, oAddr}, ...)

| Q1R (Prob = 0.36) | name | pPh | pAddr | Map |
|----------------------|--------|----------|----------|-----|
| | Ken | 111-1111 | New York | G1 |
| | Barbie | 333-3333 | Summit | G1 |

Probabilistic Mappings: By Tuple Semantics

- ◆ Consider query Q1: SELECT name, pPh, pAddr FROM MS

| S1 | name | hPhone | hAddr | oPhone | oAddr |
|--------|----------|----------|----------|----------|-------|
| Ken | 111-1111 | New York | 222-2222 | Summit | |
| Barbie | 333-3333 | Summit | 444-4444 | New York | |

- ◆ Result of Q1, under by tuple semantics, in a possible world
 - G1({**M1.n**, name}, {**M1.phP**, hPhone}, {**M1.phA**, hAddr}, ...)
 - G2({**M1.n**, name}, {**M1.phP**, oPhone}, {**M1.phA**, oAddr}, ...)

| Q1R (Prob = 0.16) | name | pPh | pAddr | Map |
|----------------------|--------|----------|----------|-----|
| | Ken | 222-2222 | Summit | G2 |
| | Barbie | 444-4444 | New York | G2 |

Probabilistic Mappings: By Tuple Semantics

- ◆ Consider query Q1: SELECT name, pPh, pAddr FROM MS

| S1 | name | hPhone | hAddr | oPhone | oAddr |
|--------|----------|----------|----------|----------|-------|
| Ken | 111-1111 | New York | 222-2222 | Summit | |
| Barbie | 333-3333 | Summit | 444-4444 | New York | |

- ◆ Result of Q1, under by tuple semantics, in a possible world
 - G1({**M1.n**, name}, {**M1.phP**, hPhone}, {**M1.phA**, hAddr}, ...)
 - G2({**M1.n**, name}, {**M1.phP**, oPhone}, {**M1.phA**, oAddr}, ...)

| Q1R (Prob = 0.24) | name | pPh | pAddr | Map |
|----------------------|--------|----------|----------|-----|
| | Ken | 111-1111 | New York | G1 |
| | Barbie | 444-4444 | New York | G2 |

Probabilistic Mappings: By Tuple Semantics

- ◆ Consider query Q1: SELECT name, pPh, pAddr FROM MS

| S1 | name | hPhone | hAddr | oPhone | oAddr |
|--------|----------|----------|----------|----------|-------|
| Ken | 111-1111 | New York | 222-2222 | Summit | |
| Barbie | 333-3333 | Summit | 444-4444 | New York | |

- ◆ Result of Q1, under by tuple semantics, in a possible world
 - G1({**M1.n**, name}, {**M1.phP**, hPhone}, {**M1.phA**, hAddr}, ...)
 - G2({**M1.n**, name}, {**M1.phP**, oPhone}, {**M1.phA**, oAddr}, ...)

| Q1R (Prob = 0.24) | name | pPh | pAddr | Map |
|----------------------|----------|--------|-------|-----|
| Ken | 222-2222 | Summit | G2 | |
| Barbie | 333-3333 | Summit | G1 | |

Probabilistic Mappings: By Tuple Semantics

- ◆ Now consider query Q2: SELECT pAddr FROM MS

| S1 | name | hPhone | hAddr | oPhone | oAddr |
|--------|----------|----------|----------|----------|-------|
| Ken | 111-1111 | New York | 222-2222 | Summit | |
| Barbie | 333-3333 | Summit | 444-4444 | New York | |

- ◆ Result of Q2, under by tuple semantics, across all possible worlds
 - Note the difference with the result of Q2, under by table semantics

| Q2R | pAddr | Prob |
|----------|-------|------|
| Summit | 0.76 | |
| New York | 0.76 | |

Questions? Suggestions? Criticisms?

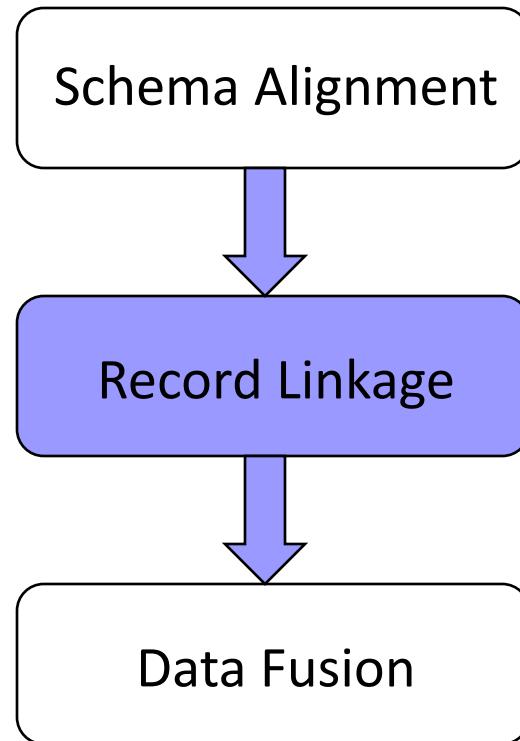


Outline

- ◆ Motivation
- ◆ Schema alignment
- ◆ Record linkage
 - Overview
 - Techniques for big data
- ◆ Data fusion
- ◆ Emerging topics

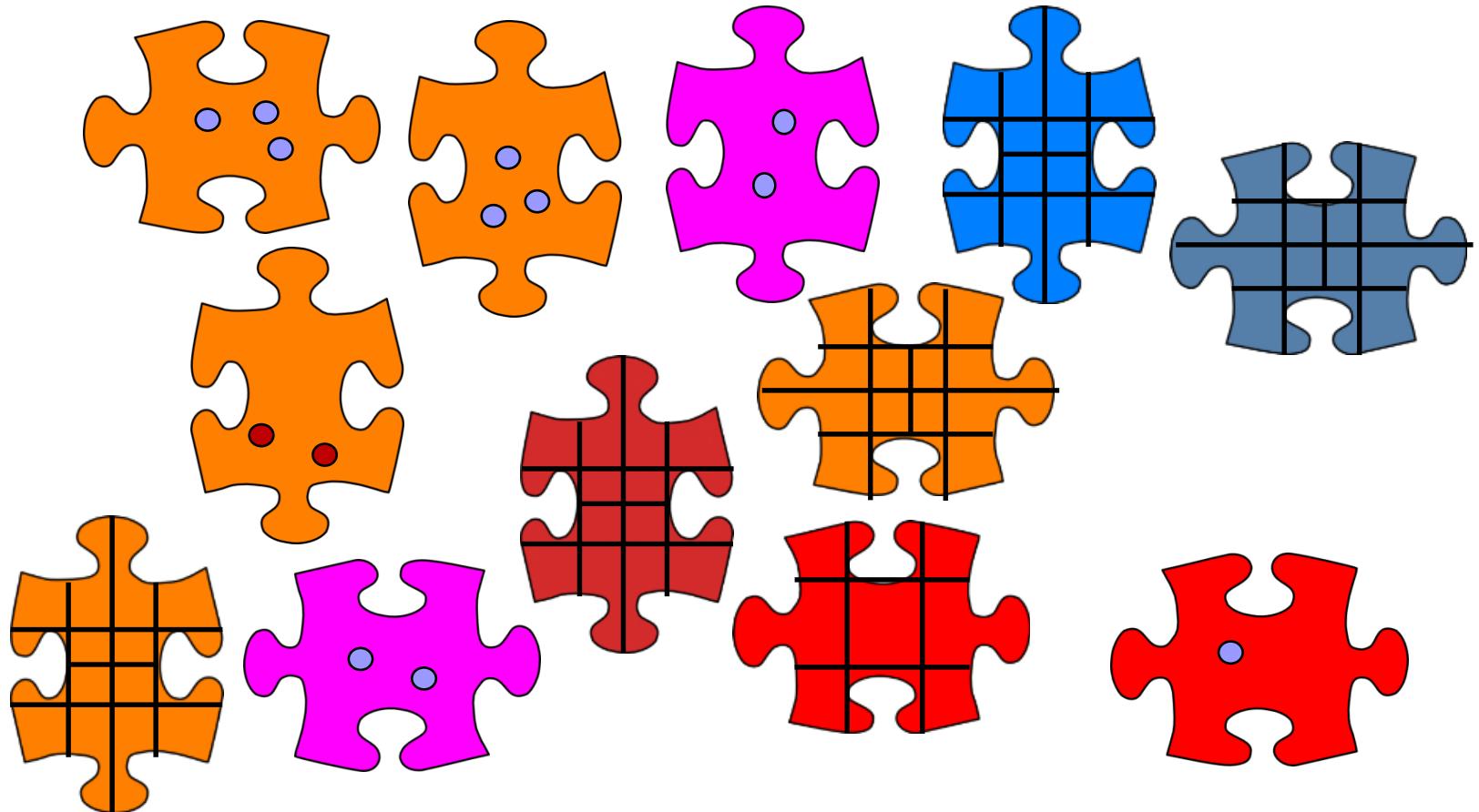
Record Linkage: Where Does It Fit In?

- ◆ “Small” data integration: alignment + linkage + fusion



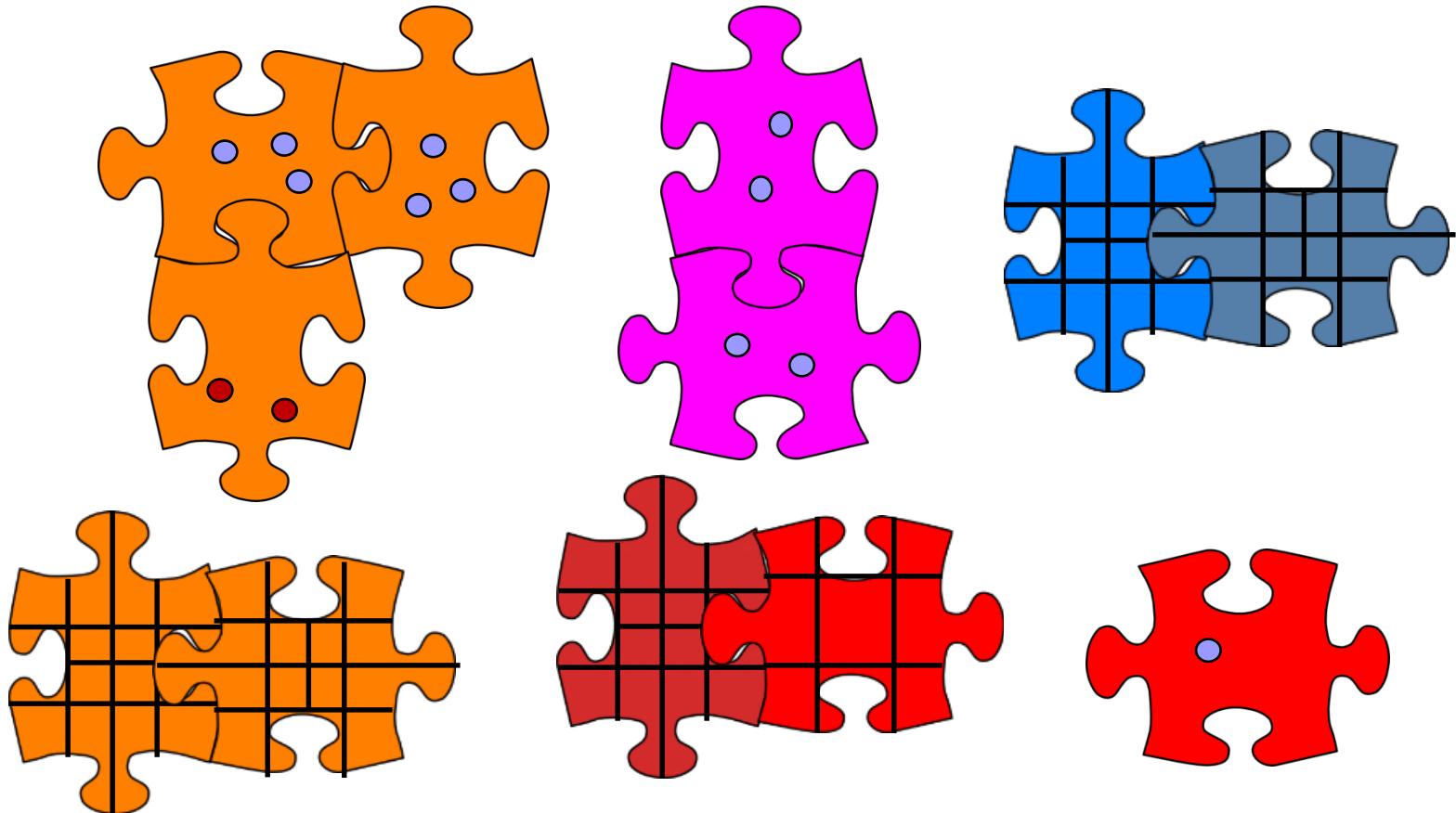
Record Linkage

- ◆ Matching based on **identifying** content: color, pattern



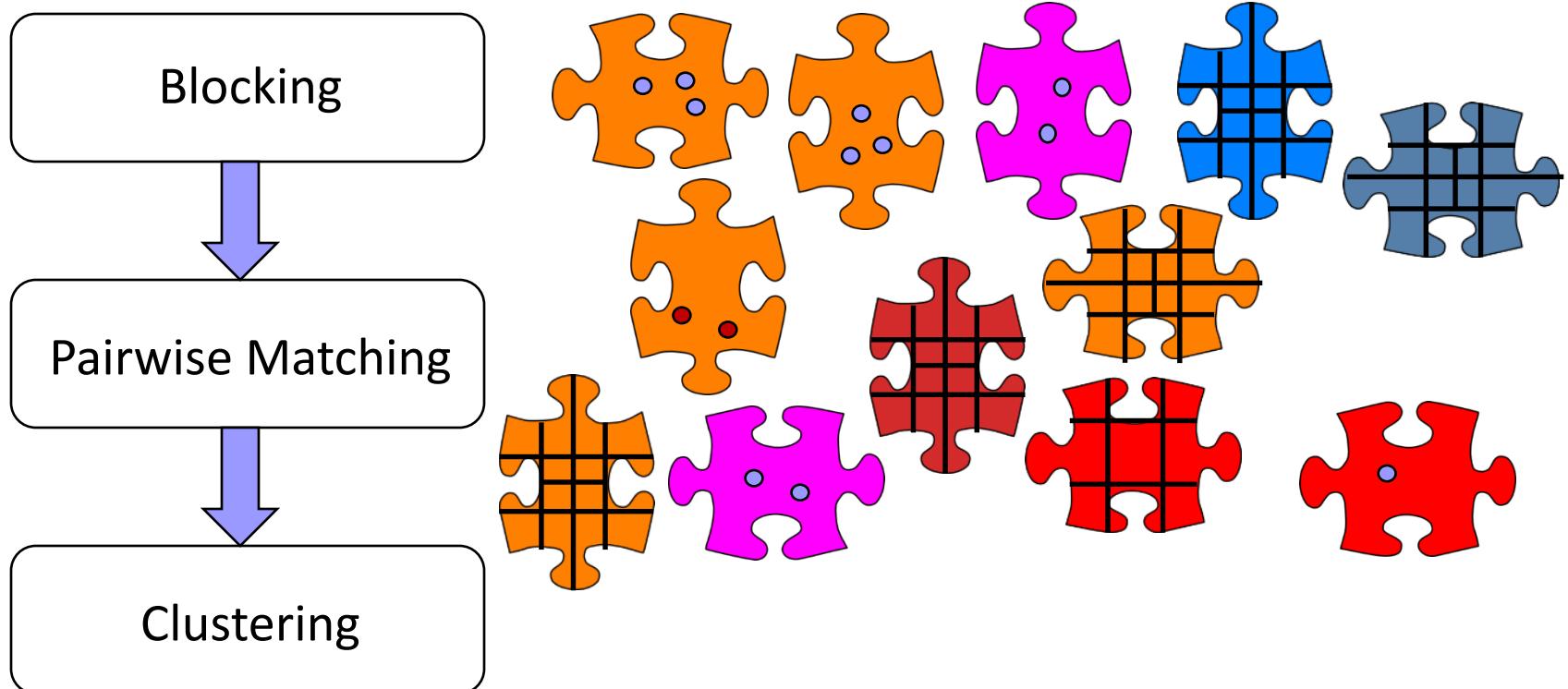
Record Linkage

- ◆ Matching based on identifying content: color, pattern



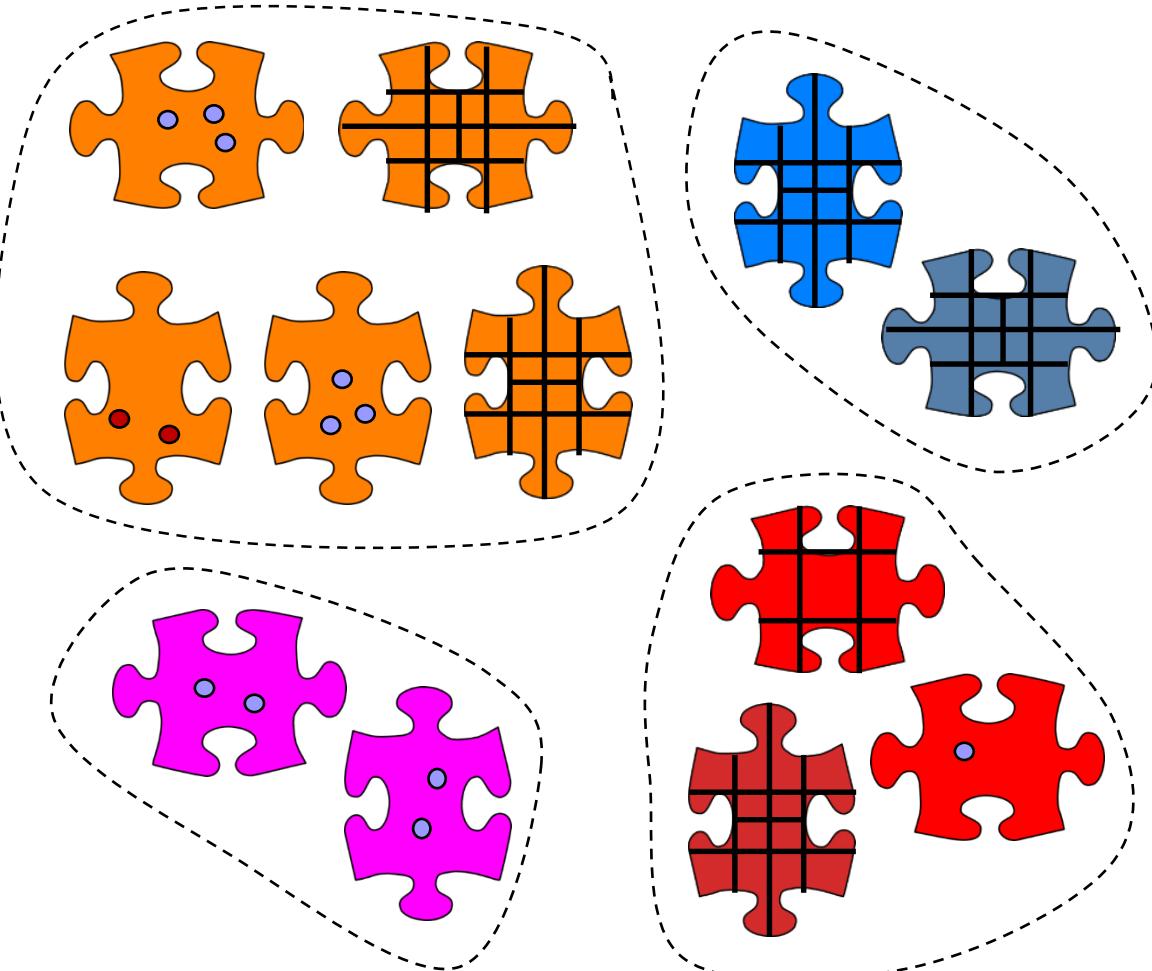
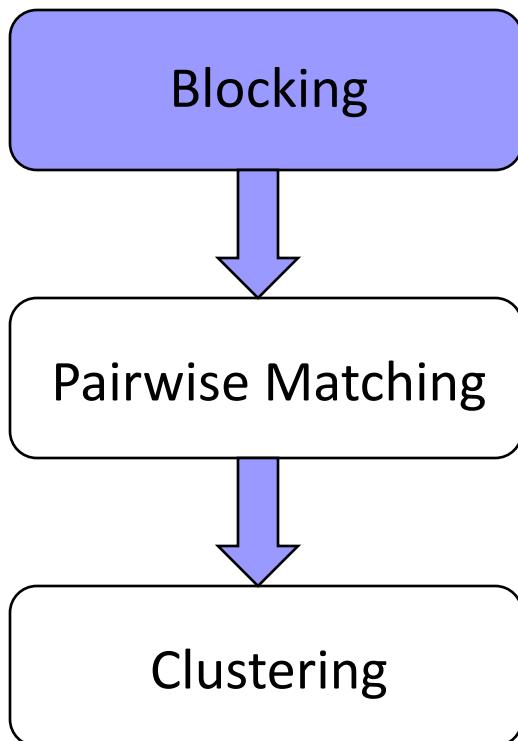
Record Linkage: Three Steps [EIV07, GM12]

- ◆ Record linkage: blocking + pairwise matching + clustering
 - Scalability, similarity, semantics



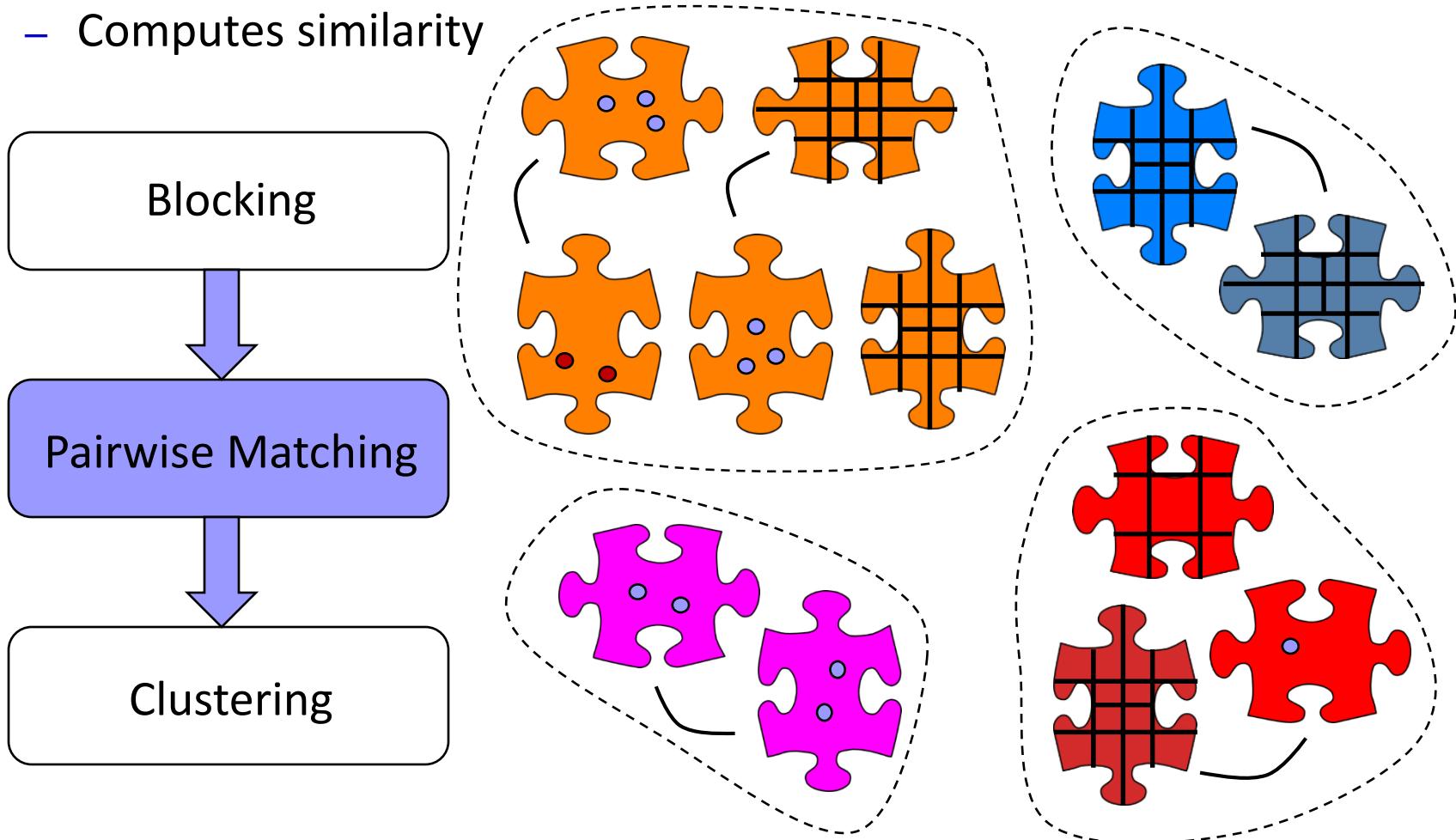
Record Linkage: Three Steps

- ◆ Blocking: **efficiently** create **small** blocks of **similar** records
 - Ensures scalability



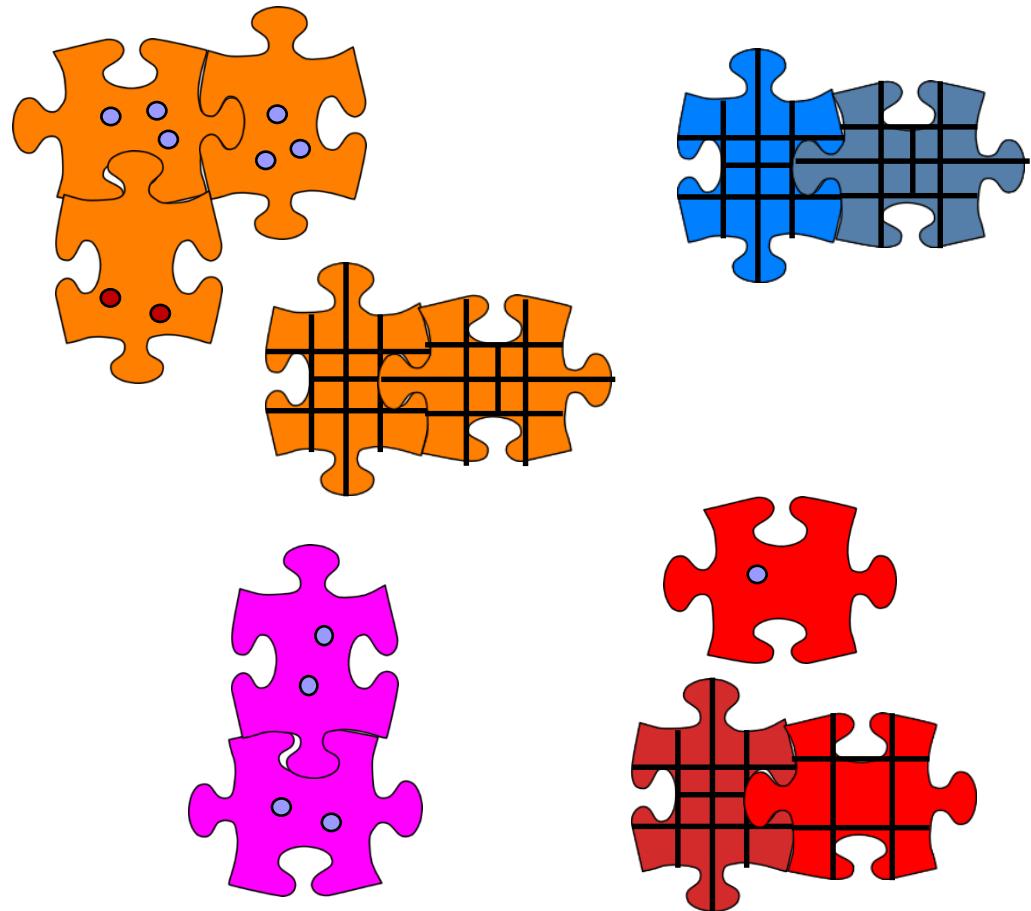
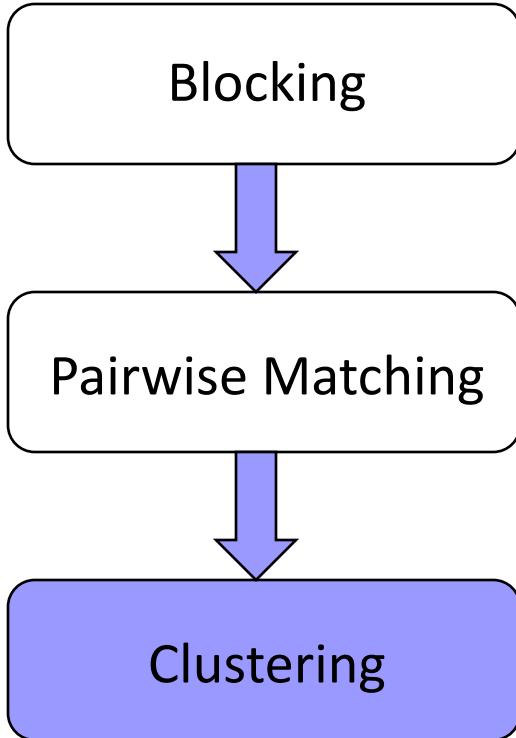
Record Linkage: Three Steps

- ◆ Pairwise matching: compares all record pairs in a block
 - Computes similarity



Record Linkage: Three Steps

- ◆ Clustering: groups sets of records into entities
 - Ensures semantics



Outline

- ◆ Motivation
- ◆ Schema alignment
- ◆ Record linkage
 - Overview
 - Techniques for big data
- ◆ Data fusion
- ◆ Emerging topics

BDI: Record Linkage

- ◆ **Volume**: dealing with billions of records
 - Map-reduce based record linkage [VCL10, KTR12]
 - Adaptive record blocking [DNS+12, MKB12, VN12]
 - Blocking in heterogeneous data spaces [PIP+12, PKP+13]

- ◆ **Velocity**
 - Incremental record linkage [WGM10, WGM13, GDS14]

BDI: Record Linkage

◆ Variety

- Matching structured and unstructured data [KGA+11, KTT+12]
- Matching Web tables and catalogs [LSC10]
- Matching product pages [QBC+18]

◆ Veracity

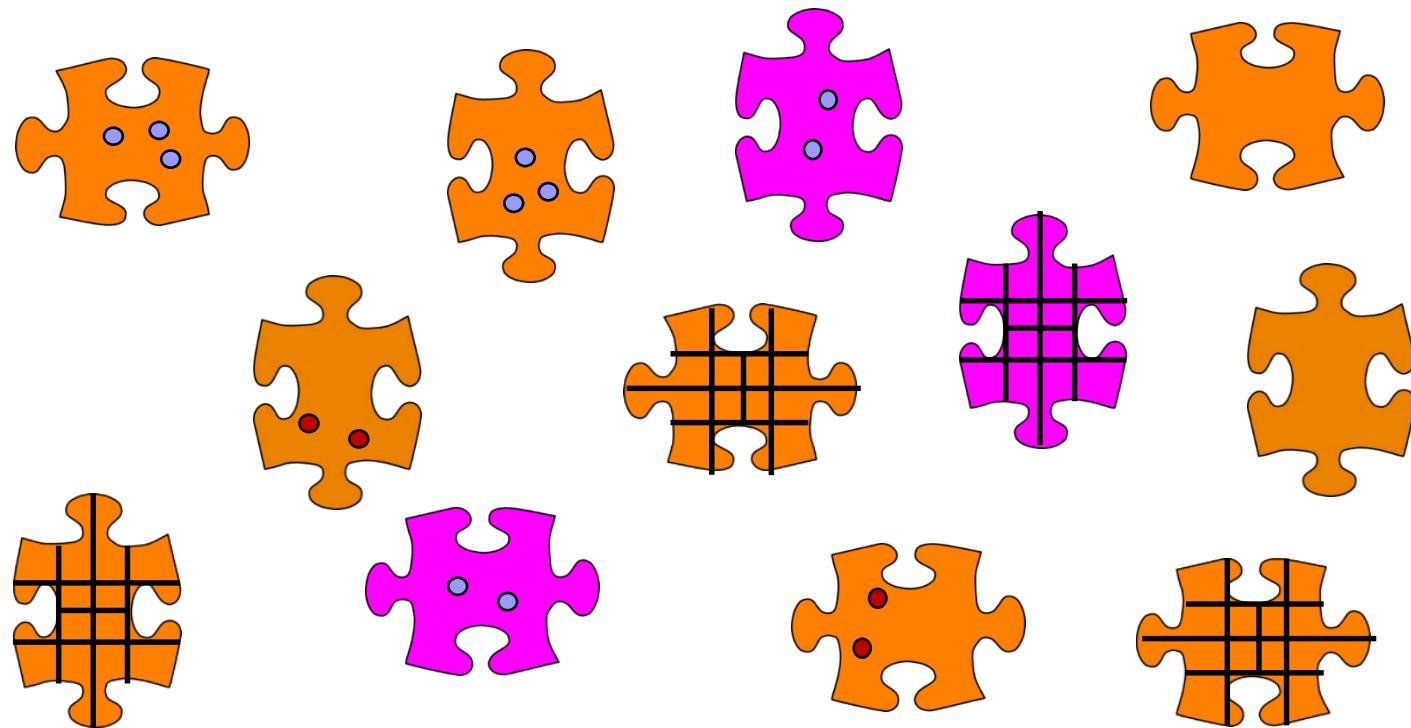
- Linking temporal records [LDM+11]
- Using crowdsourcing oracle [WLK+13, VBD14, FSS16, GFS+18]

Record Linkage Using MapReduce [KTR12]

- ◆ Motivation: despite use of blocking, record linkage is expensive
 - Can record linkage be effectively parallelized?
- ◆ Basic: use MapReduce to execute blocking-based RL in parallel
 - **Map** tasks can read records, redistribute based on blocking key
 - All entities of the same block are assigned to same **Reduce** task
 - Different blocks matched in **parallel** by multiple Reduce tasks

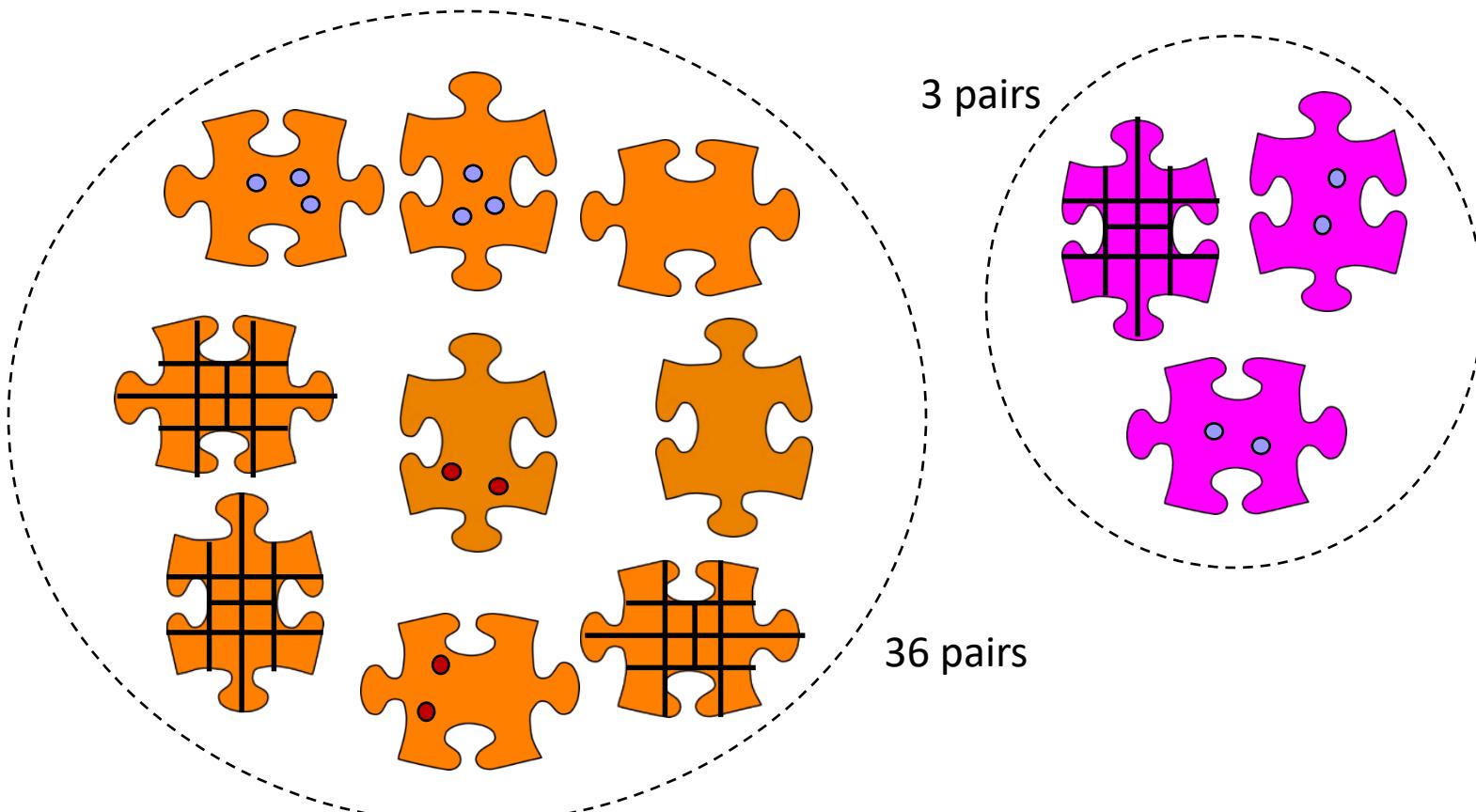
Record Linkage Using MapReduce

- ◆ Challenge: data skew → unbalanced workload



Record Linkage Using MapReduce

- ◆ Challenge: data skew → unbalanced workload
 - Speedup: $39/36 = 1.083$

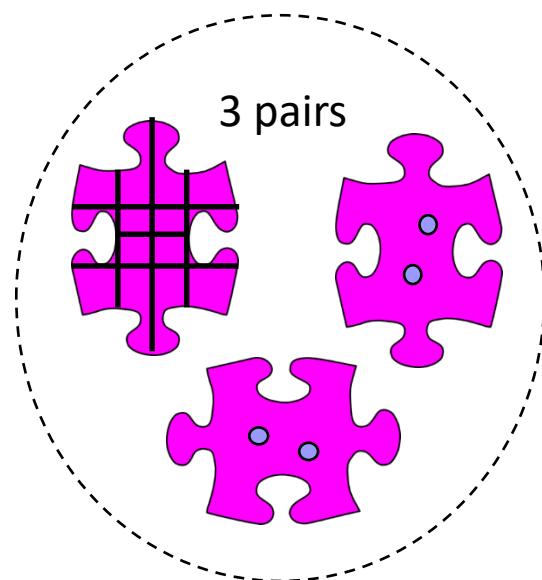


Load Balancing

- ◆ Challenge: data skew → unbalanced workload
 - Difficult to tune blocking function to get balanced workload
- ◆ Key ideas for load balancing
 - **Preprocessing** MR job to determine blocking key distribution
 - Redistribution of **Match** tasks to **Reduce** tasks to balance workload
- ◆ Two load balancing strategies:
 - BlockSplit: split large blocks into sub-blocks
 - PairRange: global enumeration and redistribution of all pairs

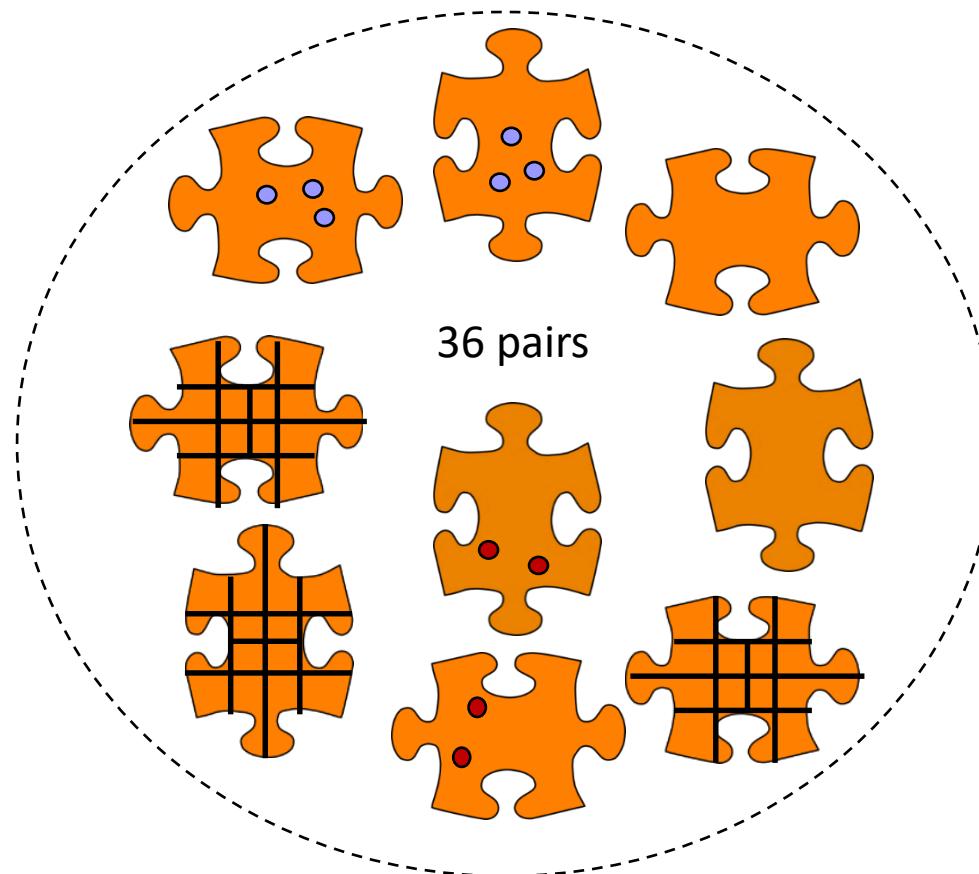
Load Balancing: BlockSplit

- ◆ Small blocks: processed by a single match task (as in Basic)



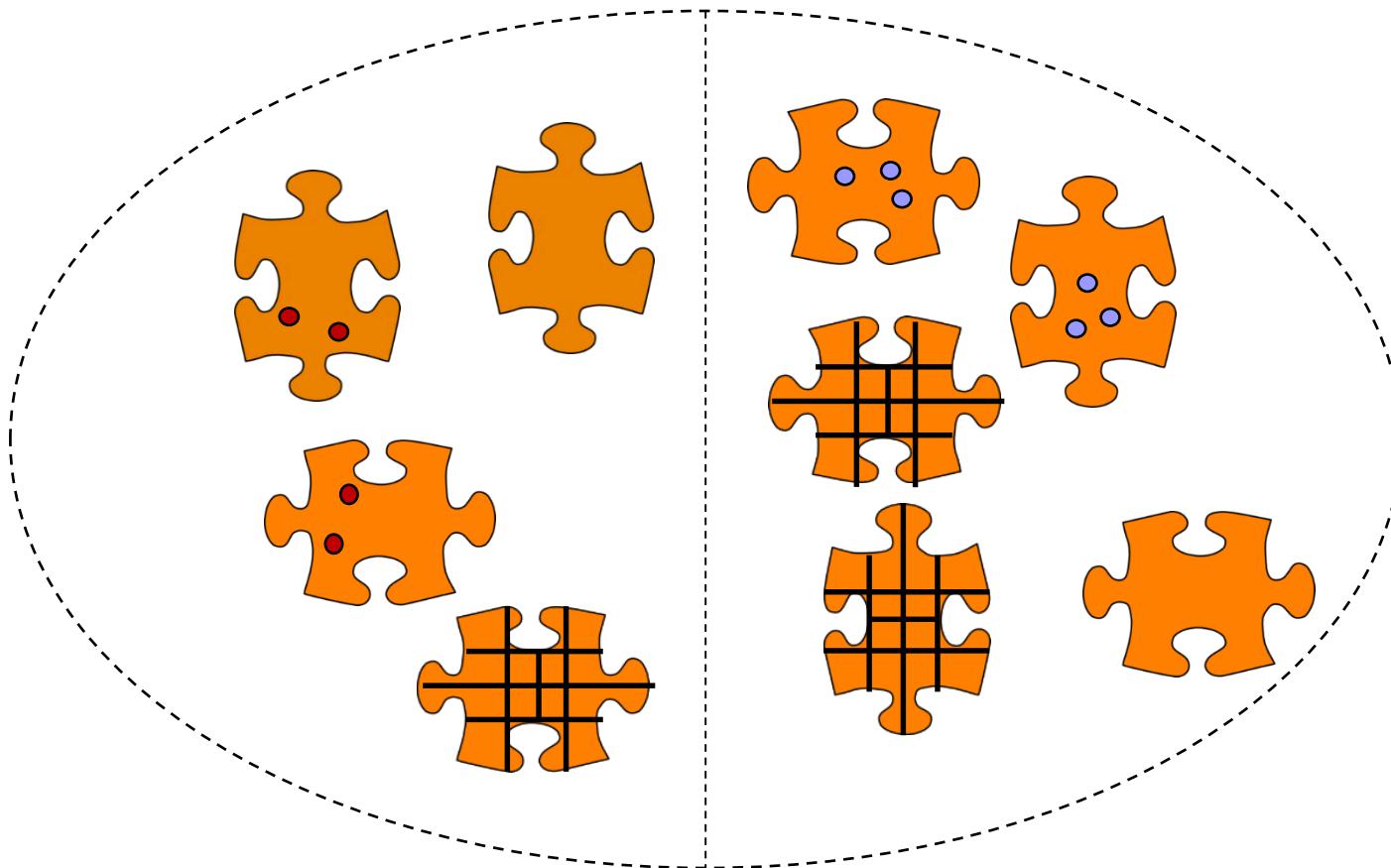
Load Balancing: BlockSplit

- ◆ Large blocks: split randomly into multiple sub-blocks



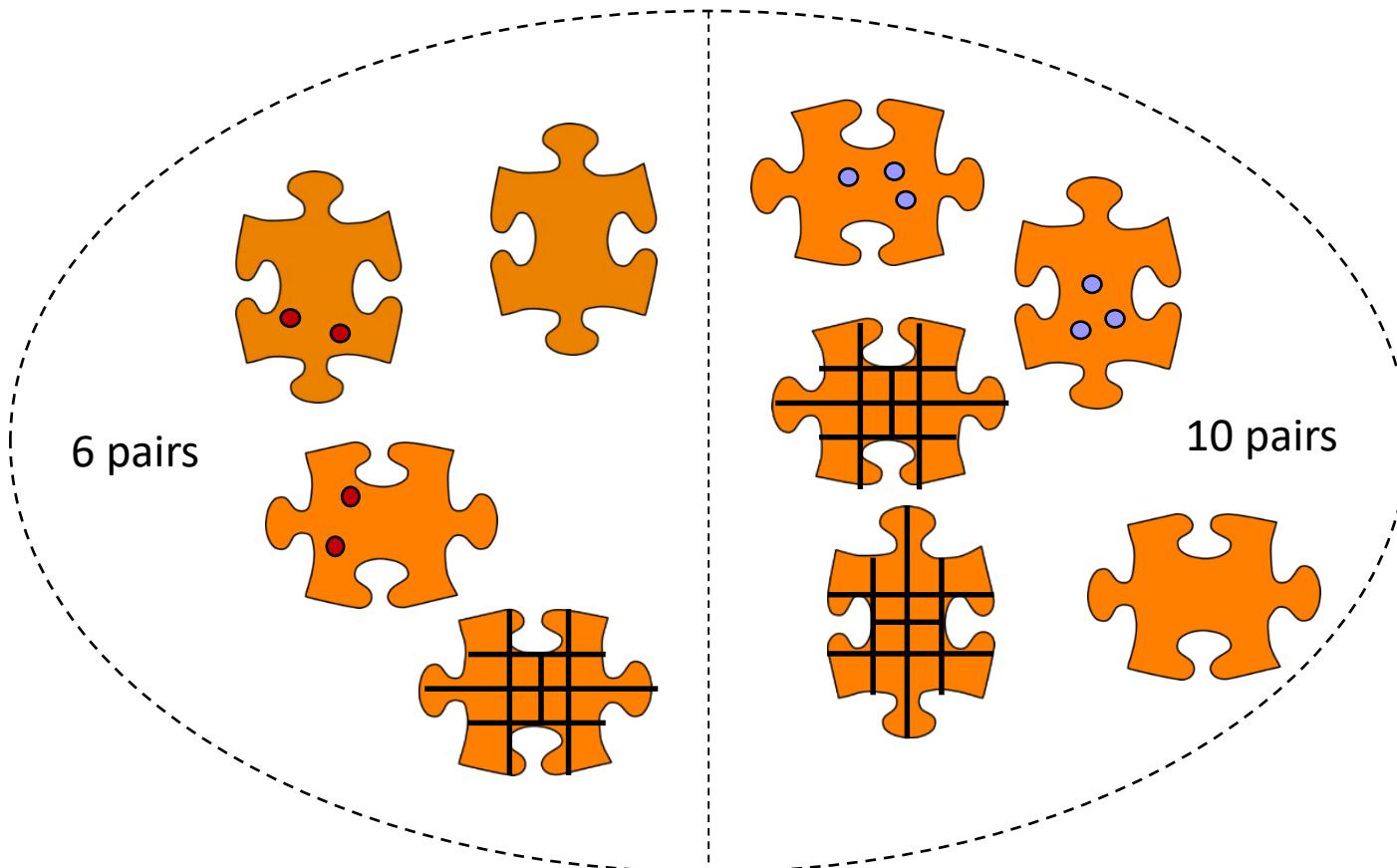
Load Balancing: BlockSplit

- ◆ Large blocks: split randomly into multiple sub-blocks



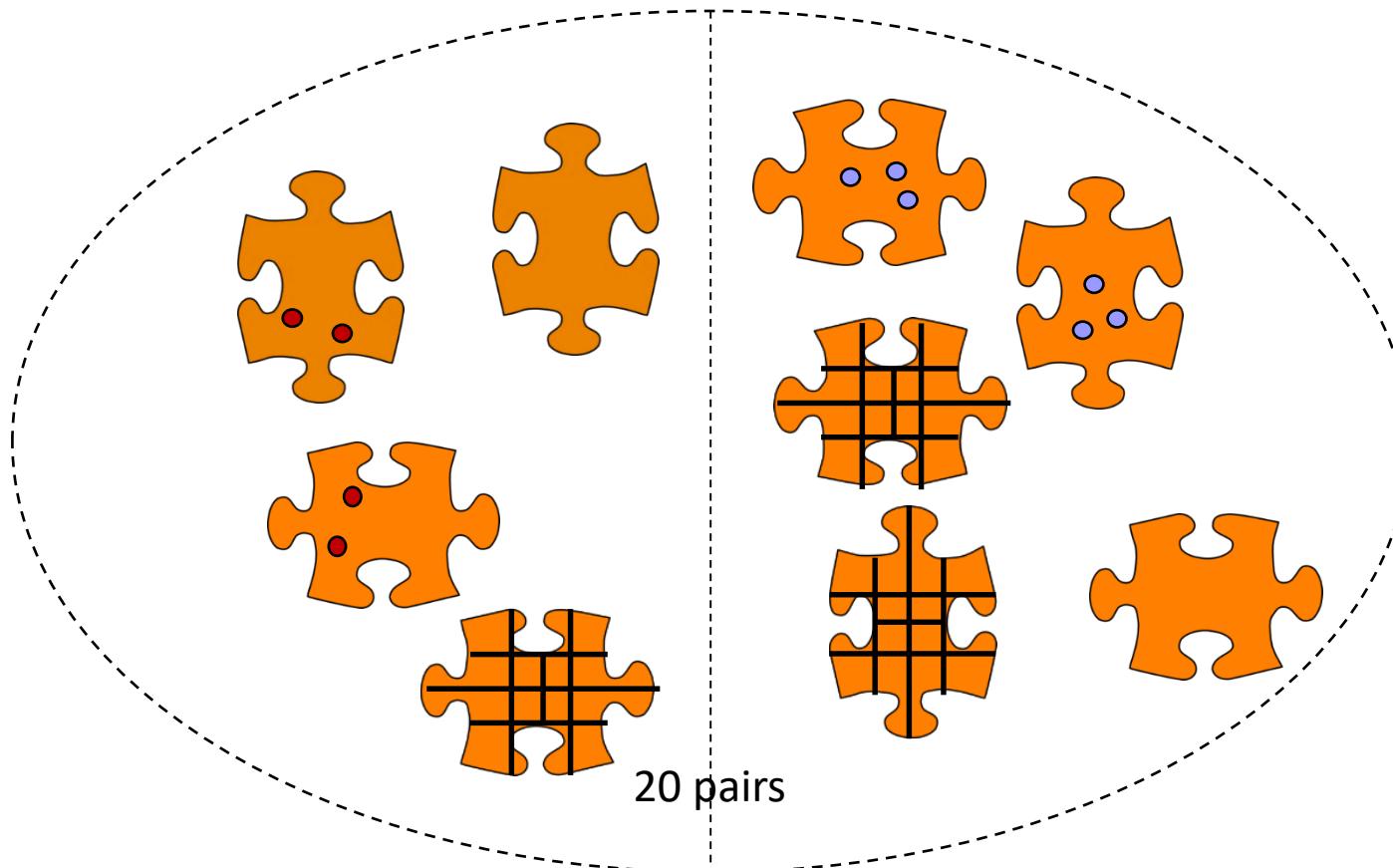
Load Balancing: BlockSplit

- ◆ Large blocks: split randomly into multiple sub-blocks
 - Each sub-block processed (like unsplit block) by single match task



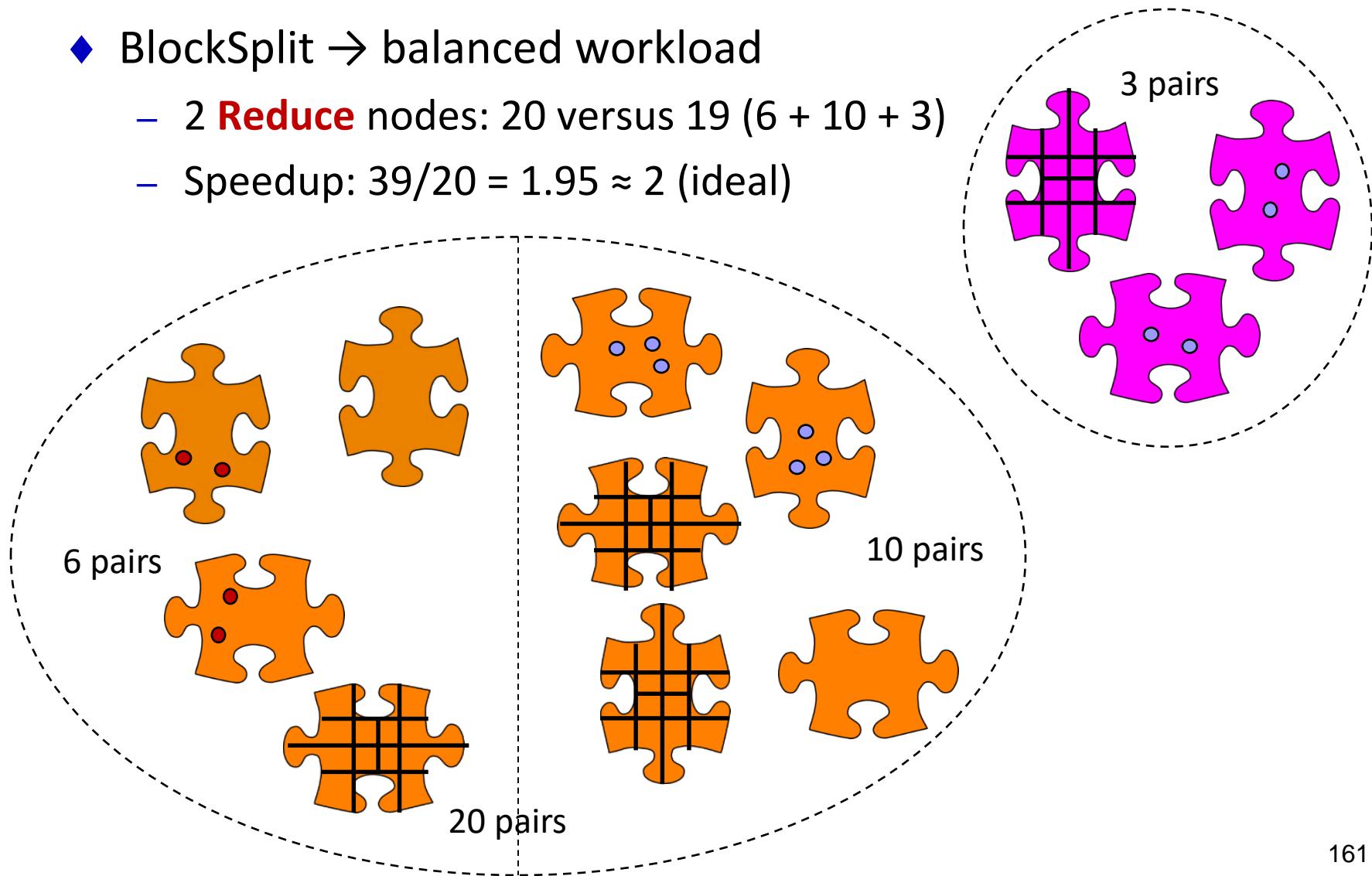
Load Balancing: BlockSplit

- ◆ Large blocks: split into multiple sub-blocks
 - Pair of sub-blocks is processed by “cartesian product” match task



Load Balancing: BlockSplit

- ◆ BlockSplit → balanced workload
 - 2 **Reduce** nodes: 20 versus 19 ($6 + 10 + 3$)
 - Speedup: $39/20 = 1.95 \approx 2$ (ideal)



Improving Blocking Recall

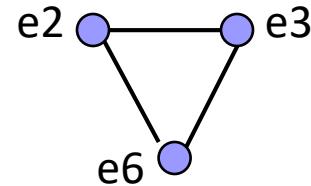
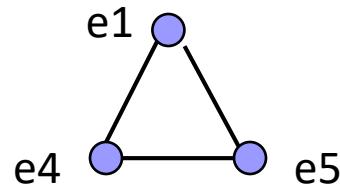
- ◆ Observation: a single, pairwise-disjoint blocking has poor recall

| Id | First Name | Last Name | YOB | State | Elected |
|----|------------|-----------|------|--------------|---------|
| e1 | Bob | Menendez | 1954 | New Jersey | 2006 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Robert | Menendez | 1954 | NJ | 2006 |
| e5 | Robert | Menendez | 1954 | New Jersey | 2007 |
| e6 | Bob | Casey, Jr | 1961 | Pennsylvania | 2007 |

Improving Blocking Recall

- ◆ Observation: a single, pairwise-disjoint blocking has poor recall

| Id | First Name | Last Name | YOB | State | Elected |
|----|------------|-----------|------|--------------|---------|
| e1 | Bob | Menendez | 1954 | New Jersey | 2006 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Robert | Menendez | 1954 | NJ | 2006 |
| e5 | Robert | Menendez | 1954 | New Jersey | 2007 |
| e6 | Bob | Casey, Jr | 1961 | Pennsylvania | 2007 |



Improving Blocking Recall

- ◆ Observation: a single, pairwise-disjoint blocking has poor recall

| Id | First Name | Last Name | YOB | State | Elected |
|----|------------|-----------|------|--------------|---------|
| e1 | Bob | Menendez | 1954 | New Jersey | 2006 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Robert | Menendez | 1954 | NJ | 2006 |
| e5 | Robert | Menendez | 1954 | New Jersey | 2007 |
| e6 | Bob | Casey, Jr | 1961 | Pennsylvania | 2007 |

- ◆ Block by Last Name: $\{\{e1, e4, e5\}, \{e2\}, \{e3, e6\}\}$



Improving Blocking Recall

- ◆ Observation: a single, pairwise-disjoint blocking has poor recall

| Id | First Name | Last Name | YOB | State | Elected |
|----|------------|-----------|------|--------------|---------|
| e1 | Bob | Menendez | 1954 | New Jersey | 2006 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Robert | Menendez | 1954 | NJ | 2006 |
| e5 | Robert | Menendez | 1954 | New Jersey | 2007 |
| e6 | Bob | Casey, Jr | 1961 | Pennsylvania | 2007 |

- ◆ Block by First Name: $\{\{e1, e2, e6\}, \{e3, e4, e5\}\}$



Improving Blocking Recall

- ◆ Observation: a single, pairwise-disjoint blocking has poor recall

| Id | First Name | Last Name | YOB | State | Elected |
|----|------------|-----------|------|--------------|---------|
| e1 | Bob | Menendez | 1954 | New Jersey | 2006 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Robert | Menendez | 1954 | NJ | 2006 |
| e5 | Robert | Menendez | 1954 | New Jersey | 2007 |
| e6 | Bob | Casey, Jr | 1961 | Pennsylvania | 2007 |

- ◆ Block by Elected: $\{\{e1, e4\}, \{e2, e3, e5, e6\}\}$



Improving Blocking Recall

- ◆ Solution: use multiple, overlapping blocking strategies

| Id | First Name | Last Name | YOB | State | Elected |
|----|------------|-----------|------|--------------|---------|
| e1 | Bob | Menendez | 1954 | New Jersey | 2006 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Robert | Menendez | 1954 | NJ | 2006 |
| e5 | Robert | Menendez | 1954 | New Jersey | 2007 |
| e6 | Bob | Casey, Jr | 1961 | Pennsylvania | 2007 |

- ◆ Block by Elected: $\{\{e1, e4\}, \{e2, e3, e5, e6\}\}$



Improving Blocking Recall

- ◆ Solution: use multiple, overlapping blocking strategies

| Id | First Name | Last Name | YOB | State | Elected |
|----|------------|-----------|------|--------------|---------|
| e1 | Bob | Menendez | 1954 | New Jersey | 2006 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Robert | Menendez | 1954 | NJ | 2006 |
| e5 | Robert | Menendez | 1954 | New Jersey | 2007 |
| e6 | Bob | Casey, Jr | 1961 | Pennsylvania | 2007 |

- ◆ Also block by Last Name: $\{\{e1, e4, e5\}, \{e2\}, \{e3, e6\}\}$

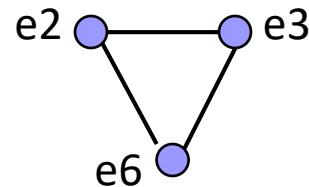
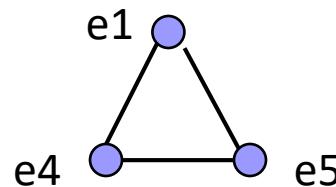


Improving Blocking Recall

- ◆ Solution: use multiple, overlapping blocking strategies

| Id | First Name | Last Name | YOB | State | Elected |
|----|------------|-----------|------|--------------|---------|
| e1 | Bob | Menendez | 1954 | New Jersey | 2006 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Robert | Menendez | 1954 | NJ | 2006 |
| e5 | Robert | Menendez | 1954 | New Jersey | 2007 |
| e6 | Bob | Casey, Jr | 1961 | Pennsylvania | 2007 |

- ◆ Also block by Last Name: $\{\{e1, e4, e5\}, \{e2\}, \{e3, e6\}\}$



Meta-Blocking [PKP+13]

- ◆ Observation: using multiple blocking strategies can be **inefficient**

| Id | First Name | Last Name | YOB | State | Elected |
|----|------------|-----------|------|--------------|---------|
| e1 | Bob | Menendez | 1954 | New Jersey | 2006 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Robert | Menendez | 1954 | NJ | 2006 |
| e5 | Robert | Menendez | 1954 | New Jersey | 2007 |
| e6 | Bob | Casey, Jr | 1961 | Pennsylvania | 2007 |

- ◆ If we block by each of: First Name, Last Name, YOB, State, Elected
 - (e1, e4) is compared 3 times, (e3, e5) is compared 2 times
 - Total number of pair comparisons is worse than not using blocking

Meta-Blocking: Dealing with BDI Variety

- ◆ Why consider blocking using all possible values?
 - Variety in BDI schemas may effectively result in schema-less data

| | | | | | |
|----|----------|------------|------|--------------|------|
| e1 | Menendez | New Jersey | 2006 | Bob | 1954 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Menendez | NJ | 1954 | Robert | 2006 |
| e5 | Menendez | New Jersey | 1954 | Robert | 2007 |
| e6 | Bob | Casey, Jr | 2007 | Pennsylvania | 1961 |

- ◆ Efficiency can be improved using ad hoc approaches
 - Block purging drops the largest blocks (akin to “stop words”)
 - May (or may not) result in big drop in recall

Meta-Blocking: Improve Blocking Efficiency

- ◆ Goal: substantially fewer pair comparisons, equally high recall
- ◆ Approach: represent multiple blockings using a weighted graph
 - Each record e_i in the data set is a node n_i in the graph
 - Create edge (n_i, n_j) if e_i and e_j are in same block in some blocking
 - Weight of edge (n_i, n_j) depends on the likelihood of a match
 - Choose high weighted subset of edges for pair comparisons
- ◆ Benefits: substantially reduce the number of pair comparisons

Meta-Blocking: Improve Blocking Efficiency

- ◆ Example: represent multiple blockings using a (weighted) graph

| | | | | | |
|----|----------|------------|------|--------------|------|
| e1 | Menendez | New Jersey | 2006 | Bob | 1954 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Menendez | NJ | 1954 | Robert | 2006 |
| e5 | Menendez | New Jersey | 1954 | Robert | 2007 |
| e6 | Bob | Casey, Jr | 2007 | Pennsylvania | 1961 |

e1 ○ e2

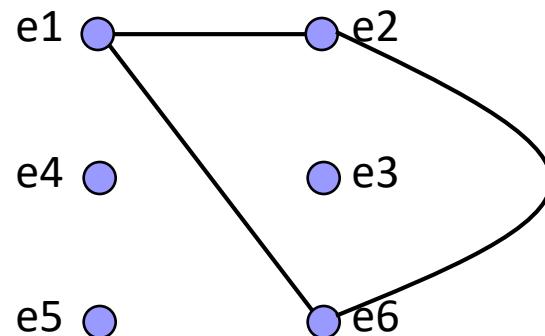
e4 ○ e3

e5 ○ e6

Meta-Blocking: Improve Blocking Efficiency

- ◆ Example: represent multiple blockings using a (weighted) graph

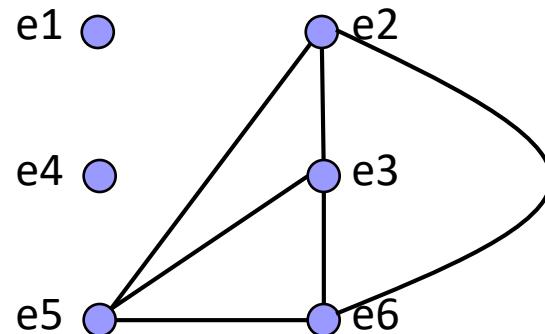
| | | | | | |
|----|----------|------------|------|--------------|------|
| e1 | Menendez | New Jersey | 2006 | Bob | 1954 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Menendez | NJ | 1954 | Robert | 2006 |
| e5 | Menendez | New Jersey | 1954 | Robert | 2007 |
| e6 | Bob | Casey, Jr | 2007 | Pennsylvania | 1961 |



Meta-Blocking: Improve Blocking Efficiency

- ◆ Example: represent multiple blockings using a (weighted) graph

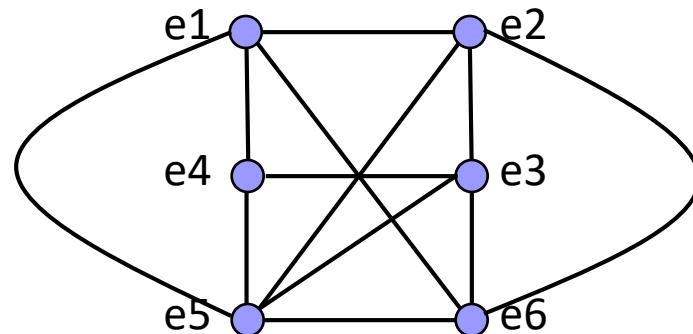
| | | | | | |
|----|----------|------------|------|--------------|------|
| e1 | Menendez | New Jersey | 2006 | Bob | 1954 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Menendez | NJ | 1954 | Robert | 2006 |
| e5 | Menendez | New Jersey | 1954 | Robert | 2007 |
| e6 | Bob | Casey, Jr | 2007 | Pennsylvania | 1961 |



Meta-Blocking: Improve Blocking Efficiency

- ◆ Example: represent multiple blockings using a (weighted) graph

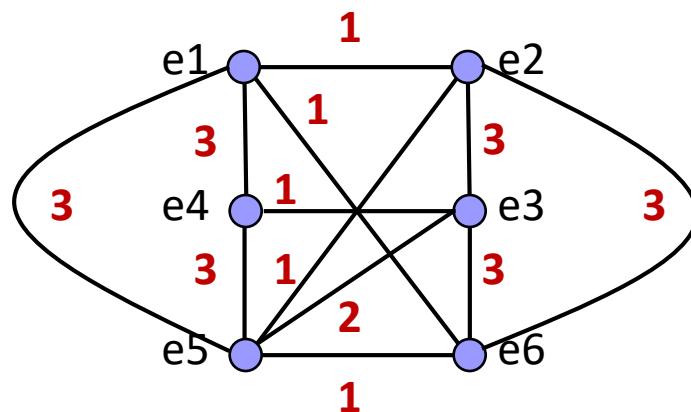
| | | | | | |
|----|----------|------------|------|--------------|------|
| e1 | Menendez | New Jersey | 2006 | Bob | 1954 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Menendez | NJ | 1954 | Robert | 2006 |
| e5 | Menendez | New Jersey | 1954 | Robert | 2007 |
| e6 | Bob | Casey, Jr | 2007 | Pennsylvania | 1961 |



Meta-Blocking: Improve Blocking Efficiency

- ◆ Example: weight of edge is number of co-occurring blocks

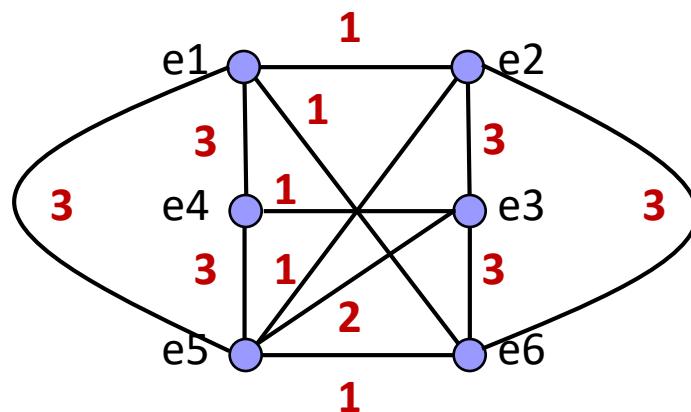
| | | | | | |
|----|----------|------------|------|--------------|------|
| e1 | Menendez | New Jersey | 2006 | Bob | 1954 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Menendez | NJ | 1954 | Robert | 2006 |
| e5 | Menendez | New Jersey | 1954 | Robert | 2007 |
| e6 | Bob | Casey, Jr | 2007 | Pennsylvania | 1961 |



Meta-Blocking: Improve Blocking Efficiency

- ◆ Example: drop all edges with weight < average edge weight (2.08)

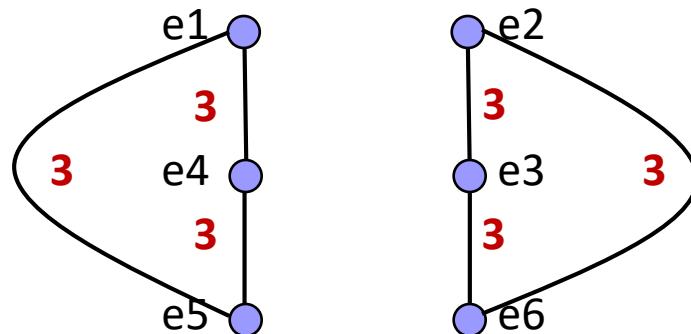
| | | | | | |
|----|----------|------------|------|--------------|------|
| e1 | Menendez | New Jersey | 2006 | Bob | 1954 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Menendez | NJ | 1954 | Robert | 2006 |
| e5 | Menendez | New Jersey | 1954 | Robert | 2007 |
| e6 | Bob | Casey, Jr | 2007 | Pennsylvania | 1961 |



Meta-Blocking: Improve Blocking Efficiency

- ◆ Example: drop all edges with weight < average edge weight (2.08)

| | | | | | |
|----|----------|------------|------|--------------|------|
| e1 | Menendez | New Jersey | 2006 | Bob | 1954 |
| e2 | Bob | Casey | 1960 | Pennsylvania | 2007 |
| e3 | Robert | Casey, Jr | 1960 | Pennsylvania | 2007 |
| e4 | Menendez | NJ | 1954 | Robert | 2006 |
| e5 | Menendez | New Jersey | 1954 | Robert | 2007 |
| e6 | Bob | Casey, Jr | 2007 | Pennsylvania | 1961 |



Meta-Blocking: Improve Blocking Efficiency

- ◆ Alternate edge weighting strategies
 - Edge weight (n_i, n_j) = number of co-occurring blocks of e_i and e_j
 - Edge weight = $\sum_{B_k} (1/|B_k|)$, e_i and e_j co-occur in B_k
- ◆ Alternate edge pruning strategies
 - Prune edges with weight below threshold (average edge weight)
 - Prune all but edges with top-k edge weights
 - For each node, prune edges based on local thresholds, counts

Meta-Blocking: Improve Blocking Efficiency

- ◆ Why does meta-blocking improve blocking efficiency?
 - Cost of graph construction = $O(\# \text{ of pair comparisons})$
 - But, pair comparison is more expensive than weight computation
 - Use inverted indexes for implementation efficiency
 - Each high weight record pair is compared only once
 - A large number of low weight record pairs are not compared

- ◆ When does meta-blocking preserve a high recall?
 - High weight record pairs correspond to high likelihood of match

Matching with Unstructured Data

- ◆ Matching product offers: 1000s of stores, millions of products
 - Product offers are terse, unstructured text
 - Many similar but different product offers

The image shows three Panasonic Lumix cameras listed vertically, each with a red circle highlighting its product description.

Top Camera: A black compact camera with a red circle around its description.

Panasonic Lumix DMC-SZ3 16.1 MP Digital camera - Black ⓘ
Other style options: Violet (\$124) White (\$125)
Panasonic Lumix - Point & Shoot - 16.1 megapixel - Compact Sensor - CCD - 10 x optical zoom - SD Card - Built-in Flash - 3.9 ounce - ISO 6,400
★★★★★ 3 reviews #3 in CCD Panasonic Lumix Digital Cameras »
Add to Shortlist

Middle Camera: A silver compact camera with a dashed red circle around its description.

Panasonic Lumix DMC-ZS25 16.1 MP Digital camera - Silver ⓘ
Other style options: Black (\$225)
Panasonic Lumix - Point & Shoot - 16.1 megapixel - Compact Sensor - 20 x optical zoom - SD Card - Built-in Flash - 6 ounce - ISO 6,400
Add to Shortlist

Bottom Camera: A black compact camera with a dashed red circle around its description.

Panasonic Lumix DMC-ZS8 14.1 MP Digital camera - Black ⓘ
Other style options: Silver (\$200)
Panasonic Lumix - Point & Shoot - 14.1 megapixel - Compact Sensor - 16 x optical zoom - SD Card - Built-in Flash - 6.6 ounce - ISO 6,400
★★★★★ 60 reviews
Add to Shortlist

Matching with Unstructured Data

- ◆ Matching product offers: 1000s of stores, millions of products
 - Product offers are terse, unstructured text
 - Many similar but different product offers
 - Same product has different descriptions, missing + wrong values



[Panasonic Dmc-fx07 7.0 Mp Digital Camera Boxed Lumix 10541r](#)

Panasonic Lumix - Point & Shoot - 7 megapixel - Compact Sensor - 3.6 x optical zoom -
Panasonic DMC-FX07 7.0 MP Digital Camera Boxed Serial #FC6GA10541r Product
Description: **Panasonic Lumix DMC-FX07** - Digital camera - compact - 7.0 ...

[Add to Shortlist](#)



[Panasonic Lumix Dmc-fx07 7.2mp Digital Camera Gold + 1 Year
Warranty](#)

Panasonic Lumix - SLR - 7.2 megapixel
AC Electronic www.ac-electronic.com Categories Mobile Phone Digital Camera
Camcorder Digital SLR Camera Camera Lens Bluetooth Product Camera ...

[Add to Shortlist](#)



[Panasonic Lumix DMC-FX07 7.0 MP Digital camera](#)

Panasonic Lumix - Point & Shoot - 7 megapixel - Compact Sensor - CCD -
The 7.2-megapixel **Lumix DMC-FX07** has a 28mm wide angle 3.6x optical zoom f/2.8
Leica DC lens housed in a compact body, achieved thanks to the ...

12 reviews

[Add to Shortlist](#)

Matching with Unstructured Data

- ◆ Matching product offers: 1000s of stores, millions of products
 - Product offers are terse, unstructured text
 - Many similar but different product offers
 - Same product has different descriptions, missing + wrong values
- ◆ Challenging scenarios for record linkage
 - Matching structured specifications with unstructured offers
 - Matching unstructured offers with each other

Structured + Unstructured Data [KGA+11]

- ◆ Motivation: matching offers to specifications with high precision
 - Product specifications are structured: set of (name, value) pairs
 - Product offers are terse, unstructured text

| Attribute Name | Attribute Value |
|----------------|-----------------|
| category | digital camera |
| brand | Panasonic |
| product line | Panasonic Lumix |
| model | DMC-FX07 |
| resolution | 7 megapixel |
| color | silver |



[Panasonic Dmc-fx07 7.0 Mp Digital Camera Boxed Lumix 10541r](#) ⓘ
Panasonic Lumix - Point & Shoot - 7 megapixel - Compact Sensor - 3.6 x optical zoom -
[Panasonic DMC-FX07 7.0 MP Digital Camera Boxed Serial #FC6GA10541r](#) Product Description: [Panasonic Lumix DMC-FX07](#) - Digital camera - compact - 7.0 ...
[Add to Shortlist](#)



[Panasonic Lumix Dmc-fx07 7.2mp Digital Camera Gold + 1 Year Warranty](#) ⓘ
Panasonic Lumix - SLR - 7.2 megapixel
AC Electronic www.ac-electronic.com Categories Mobile Phone Digital Camera Camcorder Digital SLR Camera Camera Lens Bluetooth Product Camera ...
[Add to Shortlist](#)



[Panasonic Lumix DMC-FX07 7.0 MP Digital camera](#) ⓘ
Panasonic Lumix - Point & Shoot - 7 megapixel - Compact Sensor - CCD -
The 7.2-megapixel **Lumix DMC-FX07** has a 28mm wide angle 3.6x optical zoom f/2.8 Leica DC lens housed in a compact body, achieved thanks to the ...
 12 reviews
[Add to Shortlist](#)

Structured + Unstructured Data

- ◆ Motivation: matching offers to specifications with high precision
 - Product specifications are structured: set of (name, value) pairs
 - Product offers are terse, unstructured text

| Attribute Name | Attribute Value |
|----------------|-----------------|
| category | digital camera |
| brand | Panasonic |
| product line | Panasonic Lumix |
| model | DMC-FX07 |
| resolution | 7 megapixel |
| color | silver |
| | |

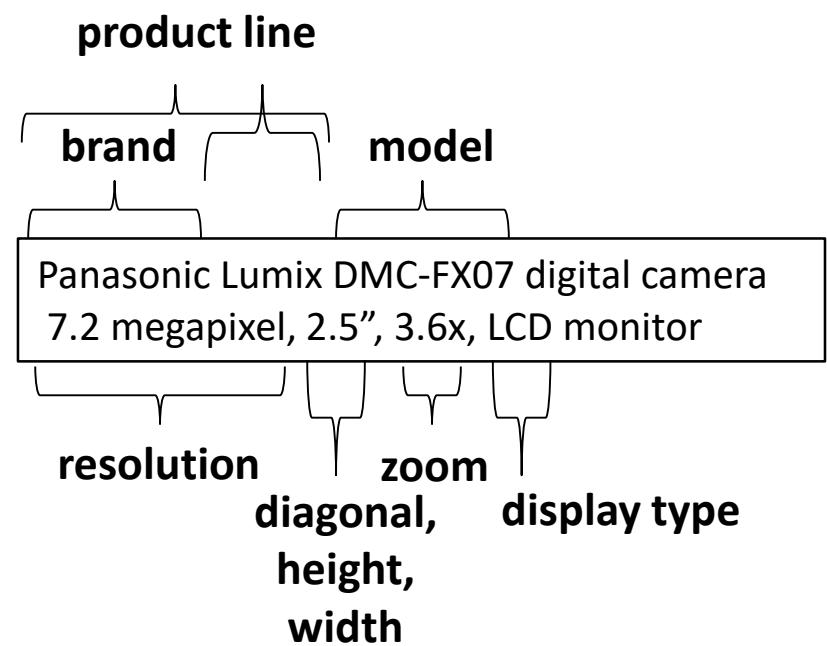
Panasonic Lumix DMC-FX07 digital camera
7.2 megapixel, 2.5", 3.6x , LCD monitor

Panasonic DMC-FX07EB digital
camera silver

Lumix FX07EB-S, 7.2MP

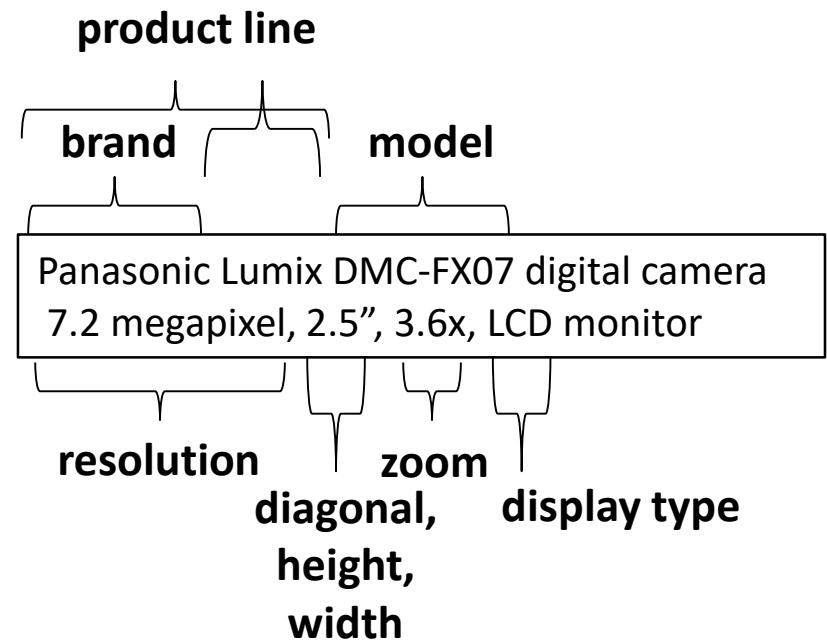
Structured + Unstructured Data

- ◆ Key idea: optimal parse of (unstructured) offer wrt specification
 - ◆ Semantic parse of offers: tagging
 - Use inverted index built on specification values
 - Tag all n-grams



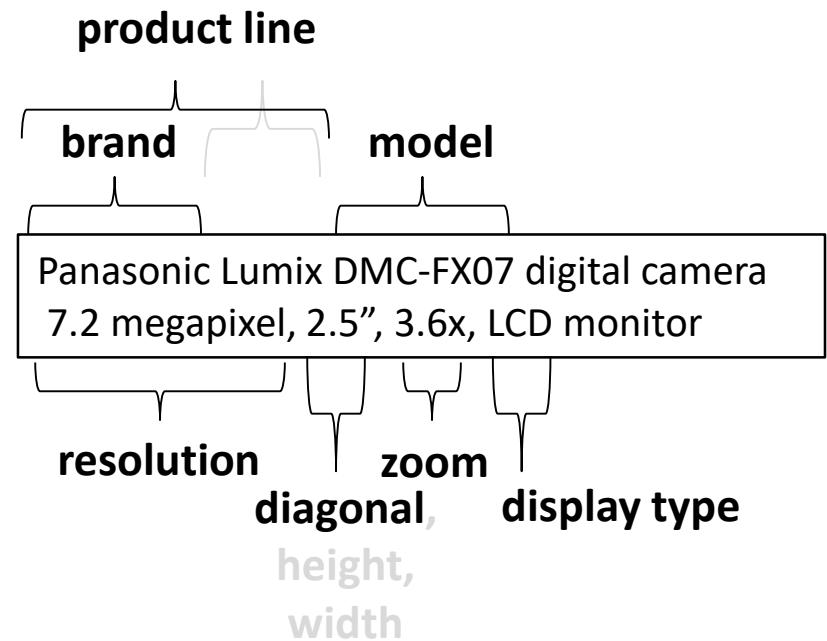
Structured + Unstructured Data

- ◆ Key idea: optimal parse of (unstructured) offer wrt specification
- ◆ Semantic parse of offers: tagging, plausible parse
 - Combination of tags such that each attribute has distinct value



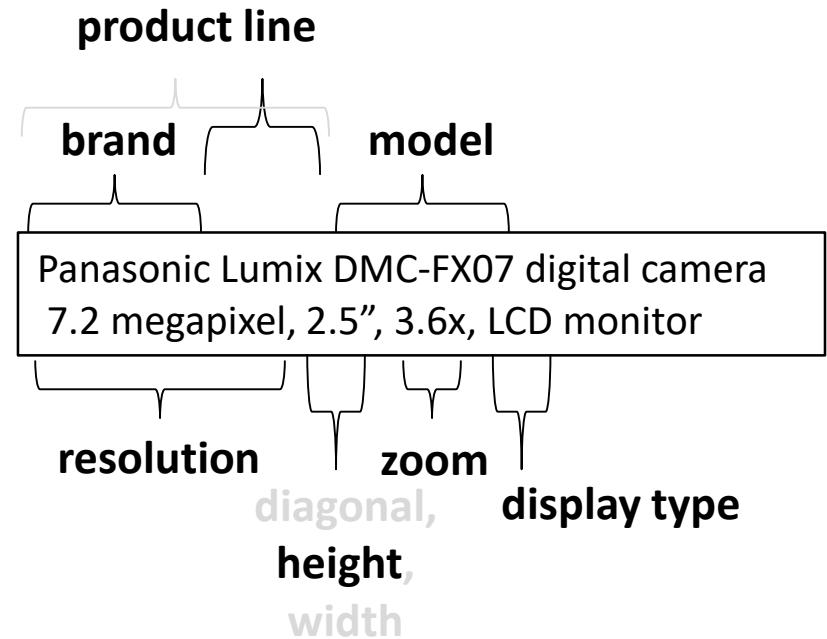
Structured + Unstructured Data

- ◆ Key idea: optimal parse of (unstructured) offer wrt specification
- ◆ Semantic parse of offers: tagging, plausible parse
 - Combination of tags such that each attribute has distinct value



Structured + Unstructured Data

- ◆ Key idea: optimal parse of (unstructured) offer wrt specification
- ◆ Semantic parse of offers: tagging, plausible parse
 - Combination of tags such that each attribute has distinct value
 - # depends on ambiguities



Structured + Unstructured Data

- ◆ Key idea: optimal parse of (unstructured) offer wrt specification
- ◆ Semantic parse of offers: tagging, plausible parse, optimal parse
 - Optimal parse depends on the product specification

| Product specification | | Optimal Parse |
|-----------------------|---------------|---|
| brand | Panasonic | Panasonic Lumix DMC-FX07 digital camera |
| product line | Lumix | 7.2 megapixel, 2.5", 3.6x, LCD monitor |
| model | DMC-FX05 | |
| diagonal | 2.5 in | |
| brand | Panasonic | Panasonic Lumix DMC-FX07 digital camera |
| model | DMC-FX07 | 7.2 megapixel, 2.5", 3.6x, LCD monitor |
| resolution | 7.2 megapixel | |
| zoom | 3.6x | |

Structured + Unstructured Data

- ◆ Key idea: optimal parse of (unstructured) offer wrt specification
- ◆ Semantic parse of offers: tagging, plausible parse, optimal parse
- ◆ Finding specification with largest match probability is now easy
 - Similarity feature vector between offer and specification: $\{-1, 0, 1\}^*$
 - Use binary logistic regression to learn weights of each feature
 - Blocking 1: use classifier to categorize offers into product categories
 - Blocking 2: identify candidates with ≥ 1 high-weighted features

Matching Unstructured Data [KTT+12]

- ◆ Motivation: matching product offers with each other
 - No structured specification available



[Panasonic Dmc-fx07 7.0 Mp Digital Camera Boxed Lumix 10541r](#) ?

Panasonic Lumix - Point & Shoot - 7 megapixel - Compact Sensor - 3.6 x optical zoom -

[Panasonic DMC-FX07 7.0 MP Digital Camera Boxed Serial #FC6GA10541r](#) Product



[Panasonic Lumix DMC-SZ3 16.1 MP Digital camera - Black](#) ?

Other style options: Violet (\$124) White (\$125)

Panasonic Lumix - Point & Shoot - 16.1 megapixel - Compact Sensor - CCD -
10 x optical zoom - SD Card - Built-in Flash - 3.9 ounce - ISO 6,400

★★★★★ 3 reviews #3 in CCD Panasonic Lumix Digital Cameras »



[Add to Shortlist](#)



AC Electronic



[Panasonic Lumix DMC-ZS25 16.1 MP Digital camera - Silver](#) ?

Other style options: Black (\$225)

Panasonic Lumix - Point & Shoot - 16.1 megapixel - Compact Sensor -
20 x optical zoom - SD Card - Built-in Flash - 6 ounce - ISO 6,400

[Add to Shortlist](#)



[Panasonic Lumix DMC-ZS8 14.1 MP Digital camera - Black](#) ?

Other style options: Silver (\$200)

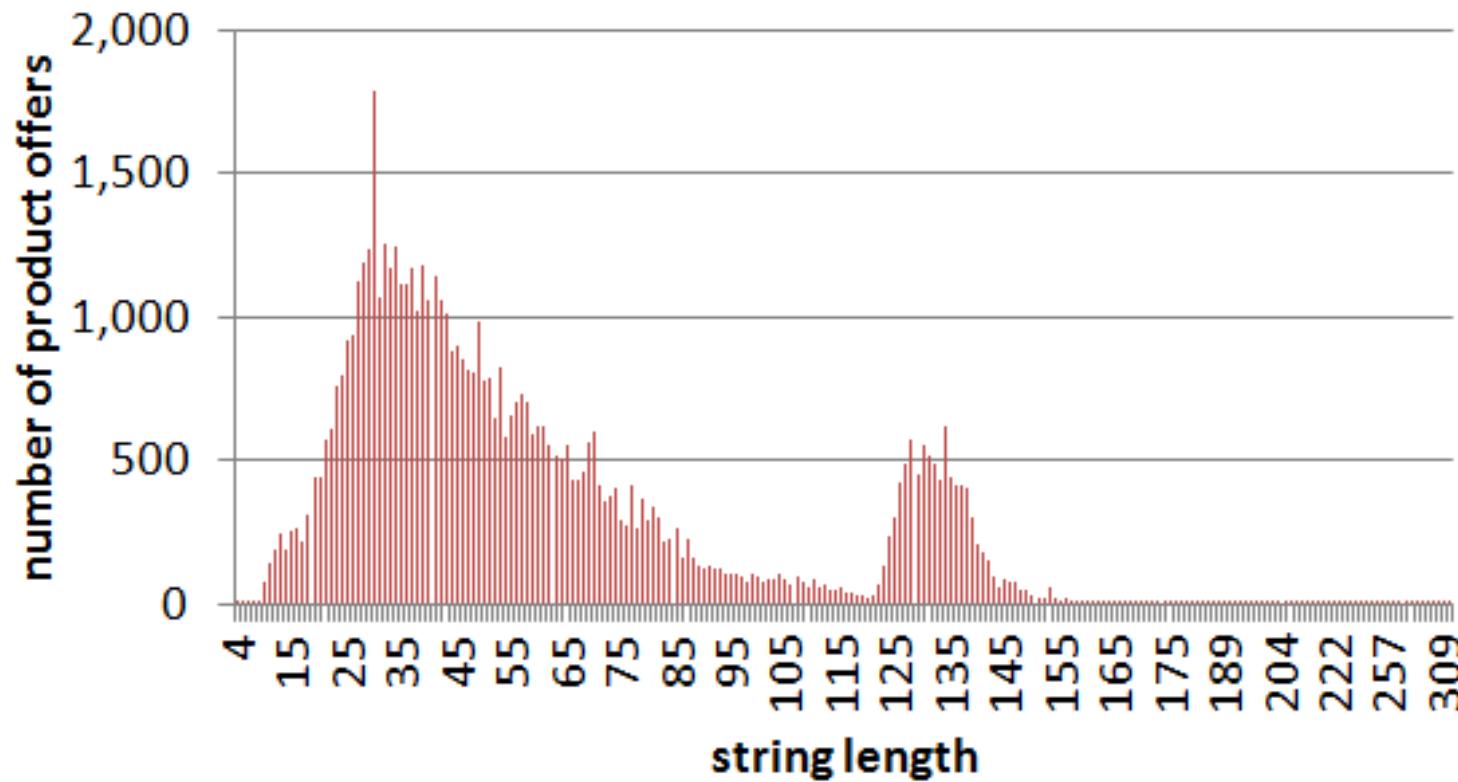
Panasonic Lumix - Point & Shoot - 14.1 megapixel - Compact Sensor -
16 x optical zoom - SD Card - Built-in Flash - 6.6 ounce - ISO 6,400

★★★★★ 80 reviews

[Add to Shortlist](#)

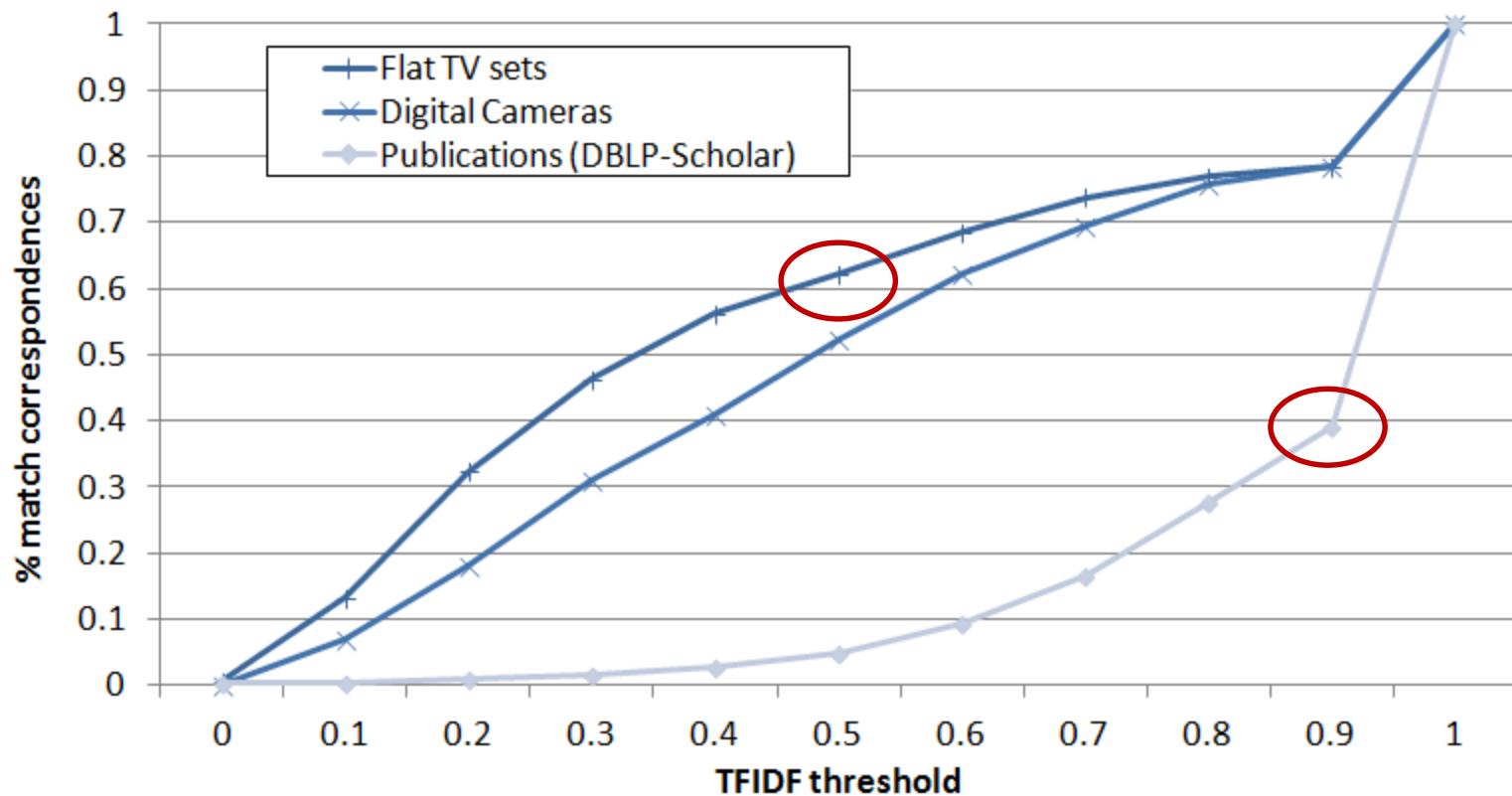
Matching Unstructured Data

- ◆ Challenge: product titles are verbose, heterogeneous [KTT+12]



Matching Unstructured Data

- ◆ Challenge: matching based on product titles is difficult [KTT+12]

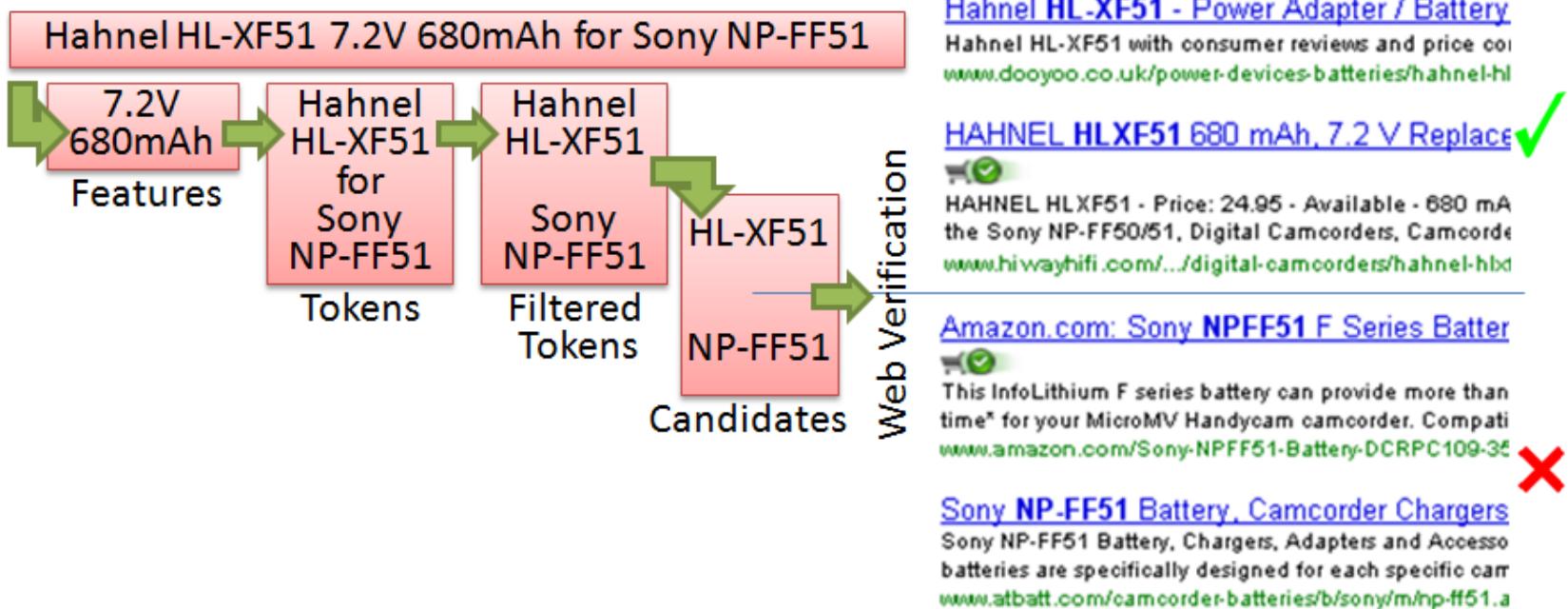


Matching Unstructured Data

- ◆ Potential solution: extract and use identifiers
 - UPC, GTIN (global trade item number) often unavailable
- ◆ Product code
 - Manufacturer-specific identifier, e.g., DMC-FX07, DMC-SZ3
 - Utilize to differentiate similar but different products

Matching Unstructured Data

- ◆ Product code extraction
 - Key step: web verification via consistency of manufacturer



Linking Temporal Records [LDM+II]

◆ How many Wei Wang's are in DBLP, with which publications?

- Wei Wang 0124 — China University of PetroleumBeijing, National Engineering Laboratory for Pipeline Safety/Beijing Key Laboratory of Urban Oil and Gas Distribution Technology, China
- Wei Wang 0125 — The Hong Kong University of Science and Technology, Department of Chemistry, Kowloon, Hong Kong
- Wei Wang 0126 — East China University of Science and Technology, Department of Mathematics, Shanghai, China
- Wei Wang 0127 — Chongqing University, School of Software Engineering, China
- Wei Wang 0128 — Beijing Institute of Technology, Institute of Application Specific Instruction-Set Processors, China

[+] Other persons with a similar name 

[–] 2010 – today 

2018

■ [j375]     Wei Wang, Wanbiao Ma:
A diffusive HIV infection model with nonlocal delayed transmission. Appl. Math. Lett. 75: 96-101 (2018)

2017

■ [j374]     Shanshan Lu, Wei Wang, Guo-yu Wang:
Stationary Points of a Kurtosis Maximization Criterion for Noisy Blind Source Extraction. IEEE Access 5: 8736-8740 (2017)

■ [j373]     Peiyan Yuan, Wei Wang, MingYang Song:
Ties in Overlapping Community Structures: Strong or Weak? IEEE Access 5: 10012-10016 (2017)

[–] Refine list

showing all 964 records

refine by search term

refine by type

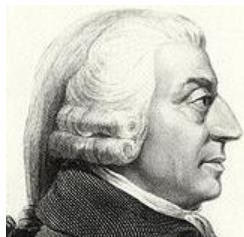
- Books and Theses (only)
 - Journal Articles (only)
 - Conference and Workshop Papers (c)
 - Editorship (only)
 - Informal Publications (only)
- [select all](#) | [deselect all](#)

refine by coauthor

Yi Zhang (6)
Weiyuan Li (5)

Linking Temporal Records: Motivation

- ◆ Traditional record linkage
 - Links records of an entity from multiple sources **at a point in time**
- ◆ Record linkage in Long Data
 - Links records of an entity **over a long time period**
 - Attribute values of an entity evolve over time
 - Different entities across time may have the same attribute value



Adam Smith (1723-1790)

Adam Smith (1965-)



Linking Temporal Records: Challenges

r1: Xin Dong

R. Polytechnic Institute

r4: Xin Luna Dong

University of Washington

r2: Xin Dong

University of Washington

r3: Xin Dong

University of Washington

r5: Xin Luna Dong

AT&T Labs-Research

r6: Xin Luna Dong

AT&T Labs-Research

1991

2004

2005

2006

2007

2008

2009

2010

2011



-Who authored what?

r7: Dong Xin

University of Illinois

r8:Dong Xin

University of Illinois

r11: Dong Xin

Microsoft Research

r9: Dong Xin

Microsoft Research

r12: Dong Xin

Microsoft Research

r10: Dong Xin

University of Illinois

Linking Temporal Records: Challenges

r1: Xin Dong

R. Polytechnic Institute

r2: Xin Dong

University of Washington

r3: Xin Dong

University of Washington

r4: Xin Luna Dong

University of Washington

r5: Xin Luna Dong

AT&T Labs-Research

r6: Xin Luna Dong

AT&T Labs-Research

1991

2004

2005

2006

2007

2008

2009

2010

2011



-Ground truth

r11: Dong Xin

Microsoft Research

r12: Dong Xin

Microsoft Research

r9: Dong Xin

Microsoft Research

r10: Dong Xin

University of Illinois

r7: Dong Xin

University of Illinois

r8:Dong Xin

University of Illinois

Linking Temporal Records: Challenges

r1: Xin Dong
R. Polytechnic Institute

r2: Xin Dong
University of Washington

r3: Xin Dong
University of Washington

r4: Xin Luna Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r6: Xin Luna Dong
AT&T Labs-Research

1991

2004

2005

2006

2007

2008

2009

2010

2011



-Traditional solution 1:
high value consistency

r7: Dong Xin
University of Illinois

r8:Dong Xin
University of Illinois

r11: Dong Xin
Microsoft Research

r9: Dong Xin
Microsoft Research

r12: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

Linking Temporal Records: Challenges

r1: Xin Dong

R. Polytechnic Institute

r2: Xin Dong

University of Washington

r3: Xin Dong

University of Washington

r4: Xin Luna Dong

University of Washington

r5: Xin Luna Dong

AT&T Labs-Research

r6: Xin Luna Dong

AT&T Labs-Research

1991

2004

2005

2006

2007

2008

2009

2010

2011



-Traditional solution 2:
using similar names

r7: Dong Xin

University of Illinois

r8:Dong Xin

University of Illinois

r9: Dong Xin

Microsoft Research

r12: Dong Xin

Microsoft Research

r10: Dong Xin

University of Illinois

r11: Dong Xin

Microsoft Research

Linking Temporal Records: Opportunities

- ◆ Smooth transition in one attribute, despite evolution of another

| ID | Name | Affiliation | Co-authors | Year |
|-----|---------------|--------------------------|-------------------|------|
| r1 | Xin Dong | R. Polytechnic Institute | Wozny | 1991 |
| r2 | Xin Dong | University of Washington | Halevy, Tatarinov | 2004 |
| r7 | Dong Xin | University of Illinois | Han, Wah | 2004 |
| r3 | Xin Dong | University of Washington | Halevy | 2005 |
| r4 | Xin Luna Dong | University of Washington | Halevy, Yu | 2007 |
| r8 | Dong Xin | University of Illinois | Wah | 2007 |
| r9 | Dong Xin | Microsoft Research | Wu, Han | 2008 |
| r10 | Dong Xin | University of Illinois | Ling, He | 2009 |
| r11 | Dong Xin | Microsoft Research | Chaudhuri, Ganti | 2009 |
| r5 | Xin Luna Dong | AT&T Labs-Research | Das Sarma, Halevy | 2009 |
| r6 | Xin Luna Dong | AT&T Labs-Research | Naumann | 2010 |
| r12 | Dong Xin | Microsoft Research | He | 2011 |

Linking Temporal Records: Opportunities

- ◆ Erratic changes in an attribute value are quite unlikely

| ID | Name | Affiliation | Co-authors | Year |
|-----|---------------|--------------------------|-------------------|------|
| r1 | Xin Dong | R. Polytechnic Institute | Wozny | 1991 |
| r2 | Xin Dong | University of Washington | Halevy, Tatarinov | 2004 |
| r7 | Dong Xin | University of Illinois | Han, Wah | 2004 |
| r3 | Xin Dong | University of Washington | Halevy | 2005 |
| r4 | Xin Luna Dong | University of Washington | Halevy, Yu | 2007 |
| r8 | Dong Xin | University of Illinois | Wah | 2007 |
| r9 | Dong Xin | Microsoft Research | Wu, Han | 2008 |
| r10 | Dong Xin | University of Illinois | Ling, He | 2009 |
| r11 | Dong Xin | Microsoft Research | Chaudhuri, Ganti | 2009 |
| r5 | Xin Luna Dong | AT&T Labs-Research | Das Sarma, Halevy | 2009 |
| r6 | Xin Luna Dong | AT&T Labs-Research | Naumann | 2010 |
| r12 | Dong Xin | Microsoft Research | He | 2011 |

Linking Temporal Records: Opportunities

- ◆ Typically, there is continuity of history, i.e., no big gaps in time

| ID | Name | Affiliation | Co-authors | Year |
|-----|---------------|--------------------------|-------------------|------|
| r1 | Xin Dong | R. Polytechnic Institute | Wozny | 1991 |
| r2 | Xin Dong | University of Washington | Halevy, Tatarinov | 2004 |
| r7 | Dong Xin | University of Illinois | Han, Wah | 2004 |
| r3 | Xin Dong | University of Washington | Halevy | 2005 |
| r4 | Xin Luna Dong | University of Washington | Halevy, Yu | 2007 |
| r8 | Dong Xin | University of Illinois | Wah | 2007 |
| r9 | Dong Xin | Microsoft Research | Wu, Han | 2008 |
| r10 | Dong Xin | University of Illinois | Ling, He | 2009 |
| r11 | Dong Xin | Microsoft Research | Chaudhuri, Ganti | 2009 |
| r5 | Xin Luna Dong | AT&T Labs-Research | Das Sarma, Halevy | 2009 |
| r6 | Xin Luna Dong | AT&T Labs-Research | Naumann | 2010 |
| r12 | Dong Xin | Microsoft Research | He | 2011 |

Linking Temporal Records: Solution

- ◆ High penalty for value disagreement over a short time period

| ID | Name | Affiliation | Co-authors | Year |
|-----|---------------|--------------------------|-------------------|------|
| r1 | Xin Dong | R. Polytechnic Institute | Wozny | 1991 |
| r2 | Xin Dong | University of Washington | Halevy, Tatarinov | 2004 |
| r7 | Dong Xin | University of Illinois | Han, Wah | 2004 |
| r3 | Xin Dong | University of Washington | Halevy | 2005 |
| r4 | Xin Luna Dong | University of Washington | Halevy, Yu | 2007 |
| r8 | Dong Xin | University of Illinois | Wah | 2007 |
| r9 | Dong Xin | Microsoft Research | Wu, Han | 2008 |
| r10 | Dong Xin | University of Illinois | Ling, He | 2009 |
| r11 | Dong Xin | Microsoft Research | Chaudhuri, Ganti | 2009 |
| r5 | Xin Luna Dong | AT&T Labs-Research | Das Sarma, Halevy | 2009 |
| r6 | Xin Luna Dong | AT&T Labs-Research | Naumann | 2010 |
| r12 | Dong Xin | Microsoft Research | He | 2011 |

Linking Temporal Records: Solution

- ◆ Lower penalty for value disagreement over a long time period

| ID | Name | Affiliation | Co-authors | Year |
|-----|---------------|--------------------------|-------------------|------|
| r1 | Xin Dong | R. Polytechnic Institute | Wozny | 1991 |
| r2 | Xin Dong | University of Washington | Halevy, Tatarinov | 2004 |
| r7 | Dong Xin | University of Illinois | Han, Wah | 2004 |
| r3 | Xin Dong | University of Washington | Halevy | 2005 |
| r4 | Xin Luna Dong | University of Washington | Halevy, Yu | 2007 |
| r8 | Dong Xin | University of Illinois | Wah | 2007 |
| r9 | Dong Xin | Microsoft Research | Wu, Han | 2008 |
| r10 | Dong Xin | University of Illinois | Ling, He | 2009 |
| r11 | Dong Xin | Microsoft Research | Chaudhuri, Ganti | 2009 |
| r5 | Xin Luna Dong | AT&T Labs-Research | Das Sarma, Halevy | 2009 |
| r6 | Xin Luna Dong | AT&T Labs-Research | Naumann | 2010 |
| r12 | Dong Xin | Microsoft Research | He | 2011 |

Linking Temporal Records: Solution

- ◆ High reward for value agreement across a small time gap

| ID | Name | Affiliation | Co-authors | Year |
|-----|---------------|--------------------------|-------------------|------|
| r1 | Xin Dong | R. Polytechnic Institute | Wozny | 1991 |
| r2 | Xin Dong | University of Washington | Halevy, Tatarinov | 2004 |
| r7 | Dong Xin | University of Illinois | Han, Wah | 2004 |
| r3 | Xin Dong | University of Washington | Halevy | 2005 |
| r4 | Xin Luna Dong | University of Washington | Halevy, Yu | 2007 |
| r8 | Dong Xin | University of Illinois | Wah | 2007 |
| r9 | Dong Xin | Microsoft Research | Wu, Han | 2008 |
| r10 | Dong Xin | University of Illinois | Ling, He | 2009 |
| r11 | Dong Xin | Microsoft Research | Chaudhuri, Ganti | 2009 |
| r5 | Xin Luna Dong | AT&T Labs-Research | Das Sarma, Halevy | 2009 |
| r6 | Xin Luna Dong | AT&T Labs-Research | Naumann | 2010 |
| r12 | Dong Xin | Microsoft Research | He | 2011 |

Linking Temporal Records: Solution

- ◆ Lower reward for value agreement across a big time gap

| ID | Name | Affiliation | Co-authors | Year |
|-----|---------------|--------------------------|-------------------|------|
| r1 | Xin Dong | R. Polytechnic Institute | Wozny | 1991 |
| r2 | Xin Dong | University of Washington | Halevy, Tatarinov | 2004 |
| r7 | Dong Xin | University of Illinois | Han, Wah | 2004 |
| r3 | Xin Dong | University of Washington | Halevy | 2005 |
| r4 | Xin Luna Dong | University of Washington | Halevy, Yu | 2007 |
| r8 | Dong Xin | University of Illinois | Wah | 2007 |
| r9 | Dong Xin | Microsoft Research | Wu, Han | 2008 |
| r10 | Dong Xin | University of Illinois | Ling, He | 2009 |
| r11 | Dong Xin | Microsoft Research | Chaudhuri, Ganti | 2009 |
| r5 | Xin Luna Dong | AT&T Labs-Research | Das Sarma, Halevy | 2009 |
| r6 | Xin Luna Dong | AT&T Labs-Research | Naumann | 2010 |
| r12 | Dong Xin | Microsoft Research | He | 2011 |

Linking Temporal Records: Intuitions

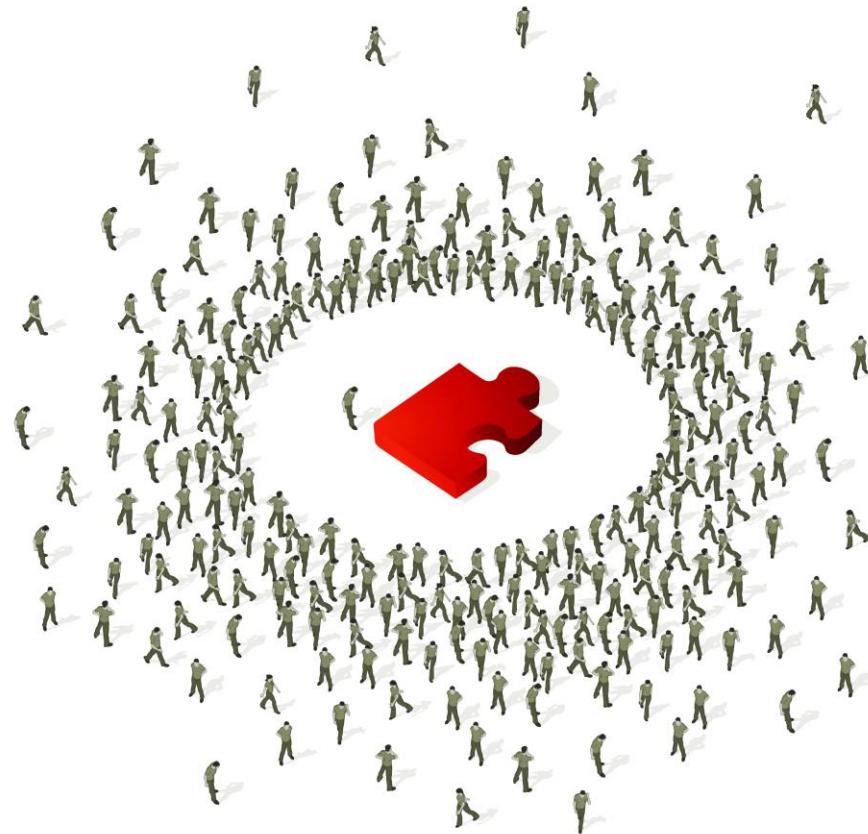
- ◆ Consider records in time order for clustering

| ID | Name | Affiliation | Co-authors | Year |
|-----|---------------|--------------------------|-------------------|------|
| r1 | Xin Dong | R. Polytechnic Institute | Wozny | 1991 |
| r2 | Xin Dong | University of Washington | Halevy, Tatarinov | 2004 |
| r7 | Dong Xin | University of Illinois | Han, Wah | 2004 |
| r3 | Xin Dong | University of Washington | Halevy | 2005 |
| r4 | Xin Luna Dong | University of Washington | Halevy, Yu | 2007 |
| r8 | Dong Xin | University of Illinois | Wah | 2007 |
| r9 | Dong Xin | Microsoft Research | Wu, Han | 2008 |
| r10 | Dong Xin | University of Illinois | Ling, He | 2009 |
| r11 | Dong Xin | Microsoft Research | Chaudhuri, Ganti | 2009 |
| r5 | Xin Luna Dong | AT&T Labs-Research | Das Sarma, Halevy | 2009 |
| r6 | Xin Luna Dong | AT&T Labs-Research | Naumann | 2010 |
| r12 | Dong Xin | Microsoft Research | He | 2011 |

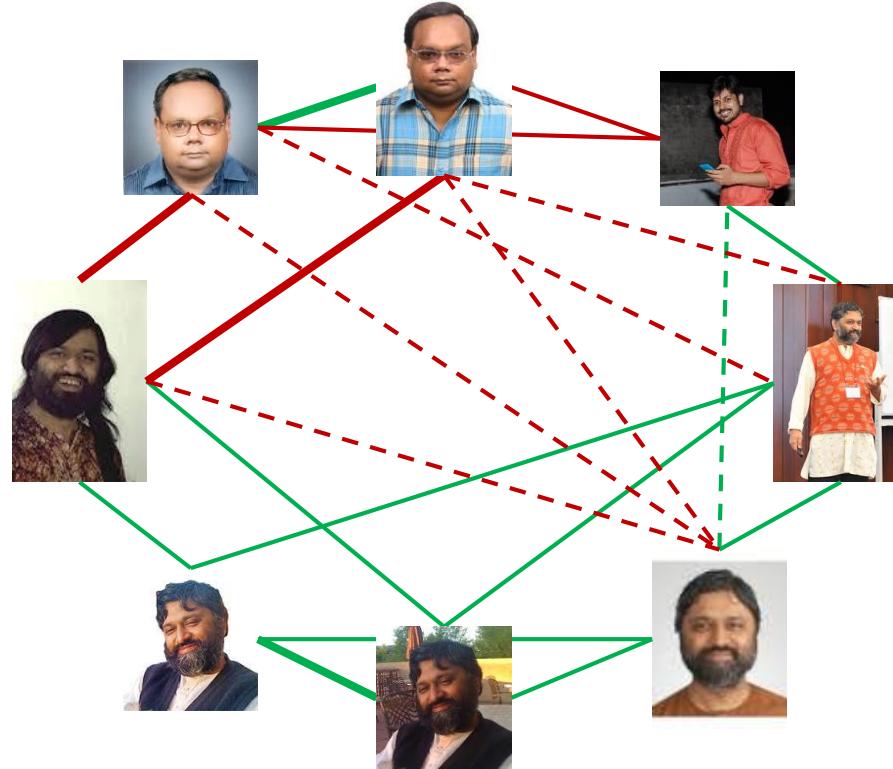


Using An Oracle [WLK+13, VBDI14, FSSI16]

- ◆ Veracity: crowdsourcing can help with difficult pairwise matching

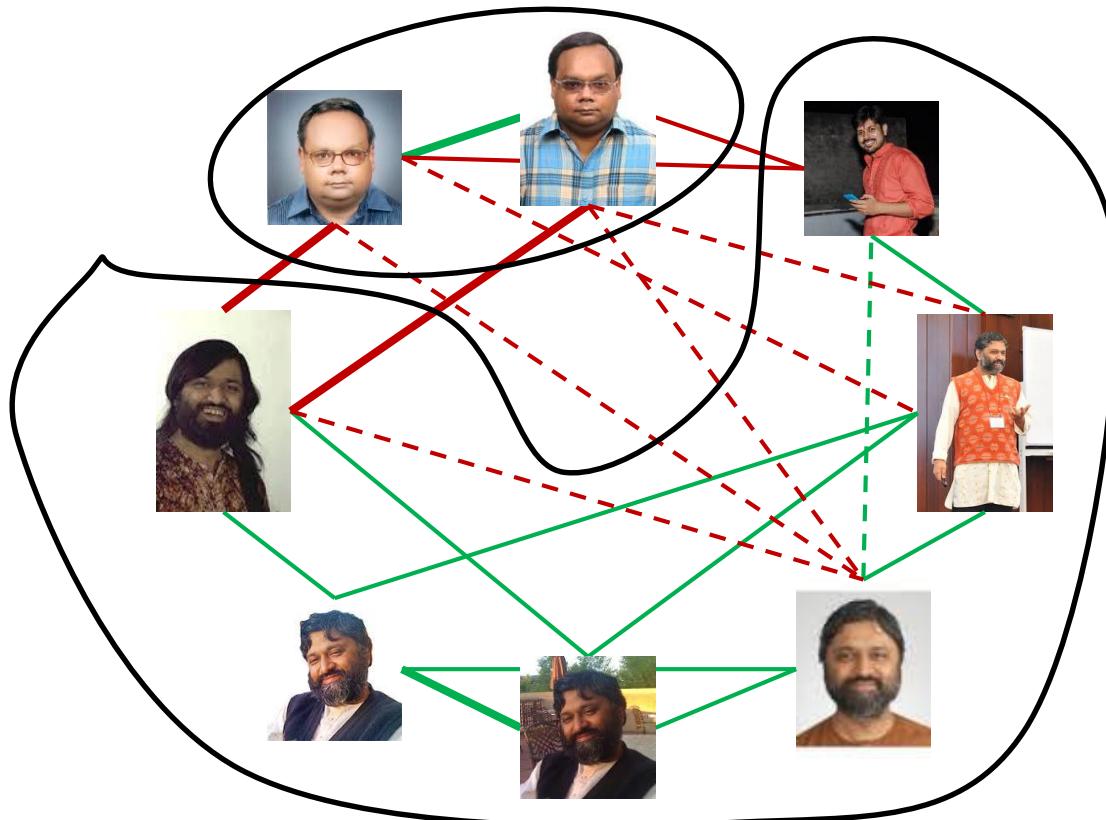


Using An Oracle: Motivation



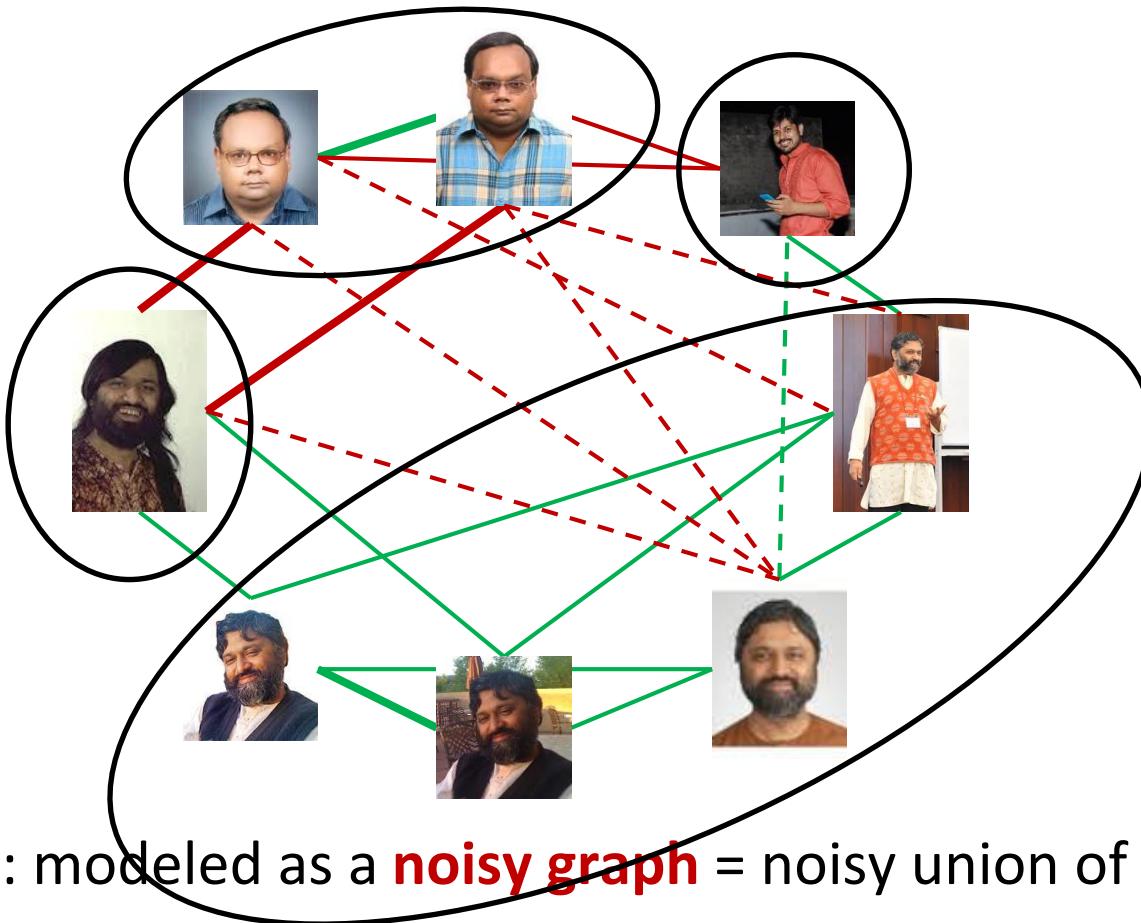
- ◆ Input data: modeled as a **noisy graph** = noisy union of cliques.
 - Node = description, image, structured record, etc. of an entity.
 - Edge weight = **probability** that node pair represents same entity.

Using An Oracle: Motivation



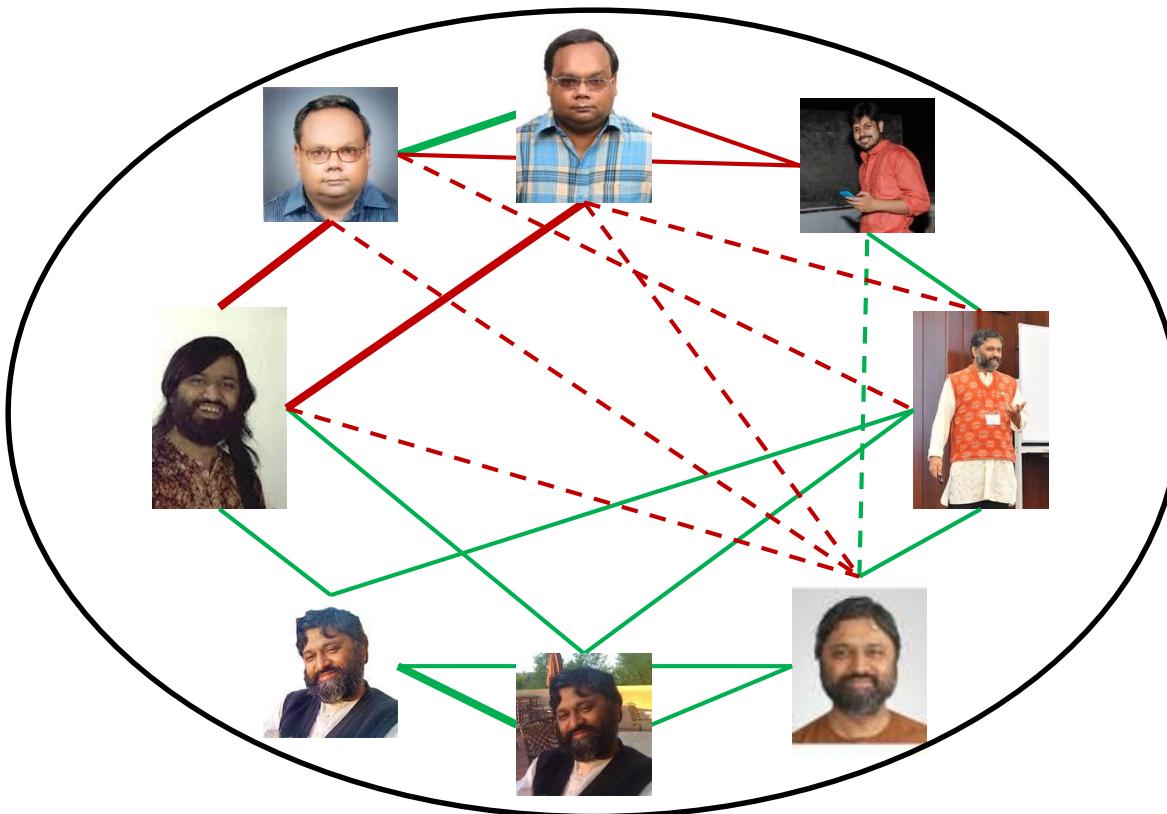
- ◆ Input data: modeled as a **noisy graph** = noisy union of cliques.
- ◆ Output: a **set of clusters**, each of which corresponds to an entity.
 - Based on **connected components** of green edges.

Using An Oracle: Motivation



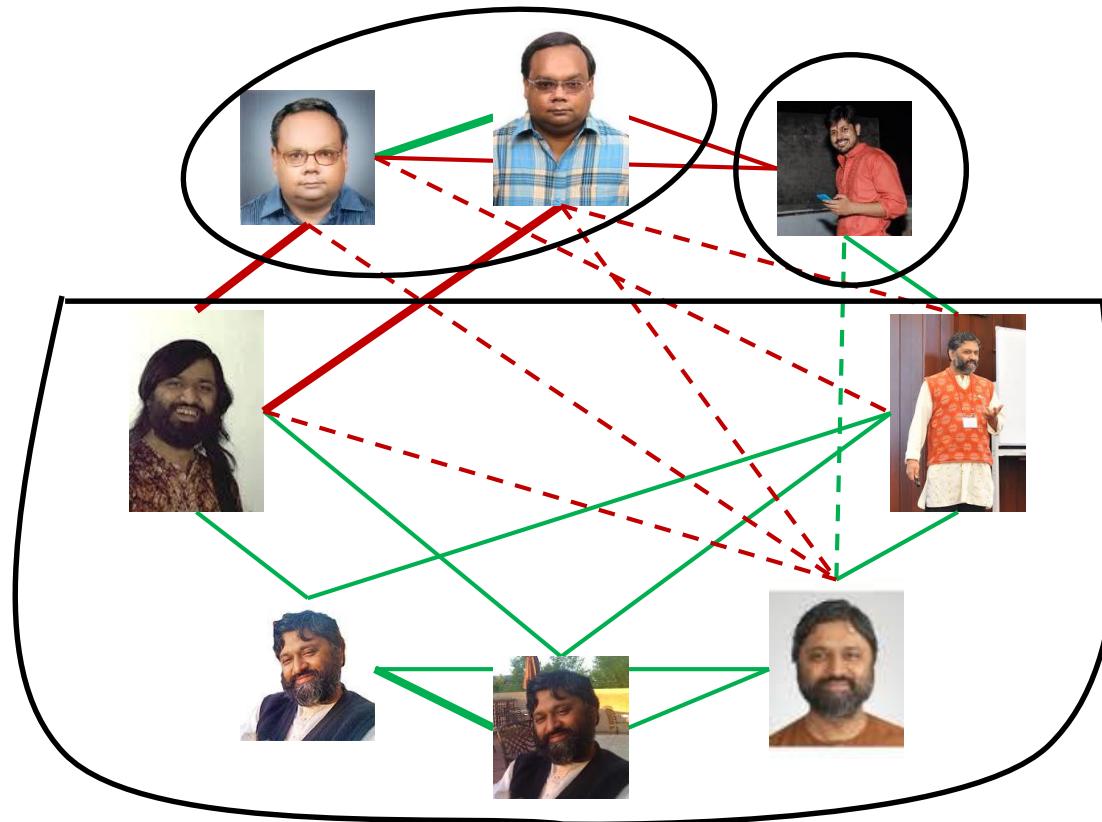
- ◆ Input data: modeled as a **noisy graph** = noisy union of cliques.
- ◆ Output: a **set of clusters**, each of which corresponds to an entity.
 - Based on **correlation clustering** using edge weights.

Using An Oracle: Motivation



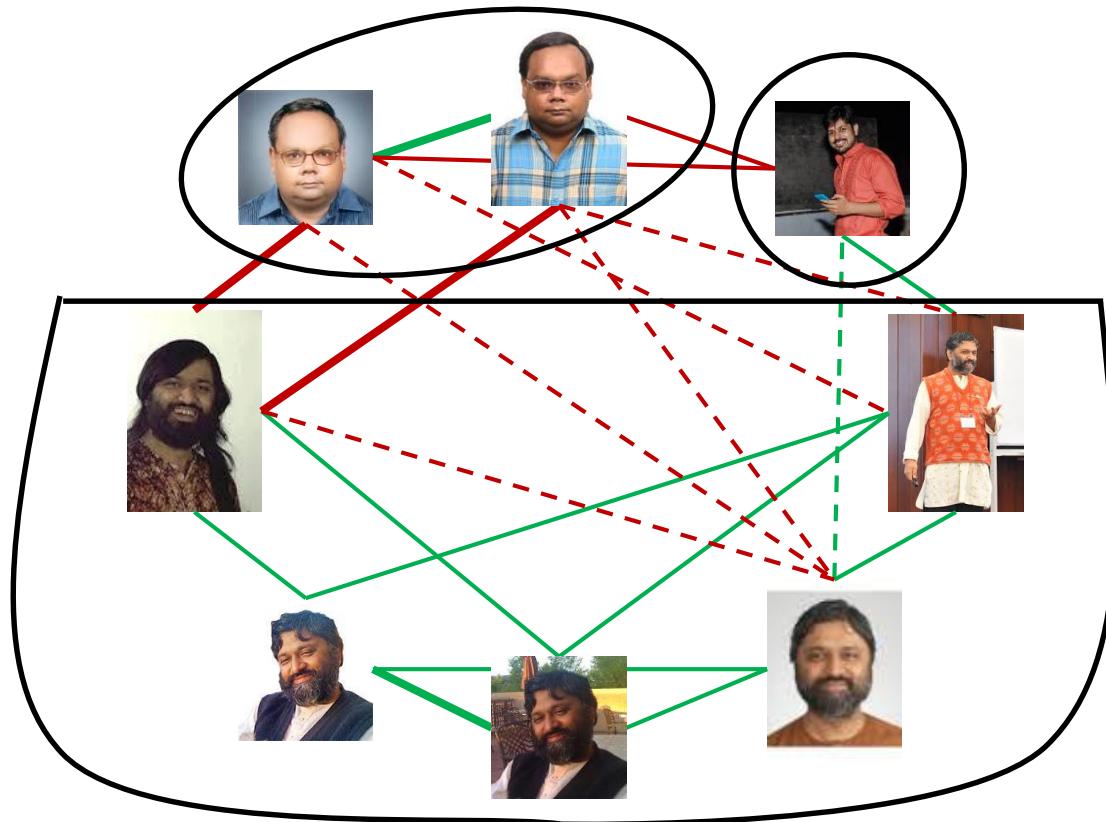
- ◆ Input data: modeled as a **noisy graph** = noisy union of cliques.
- ◆ Output: a **set of clusters**, each of which corresponds to an entity.
 - Based **only on name** = “Divesh Srivastava”.

Using An Oracle: Motivation



- ◆ Input data: modeled as a **noisy graph** = noisy union of cliques.
- ◆ Output: a **set of clusters**, each of which corresponds to an entity.
 - **Ground truth** consists of 3 clusters.

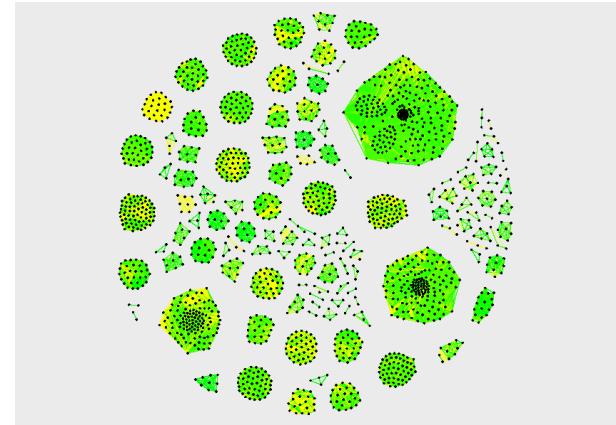
Using An Oracle: Motivation



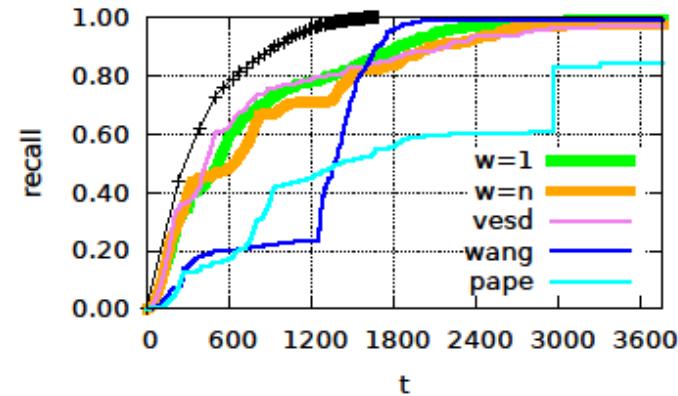
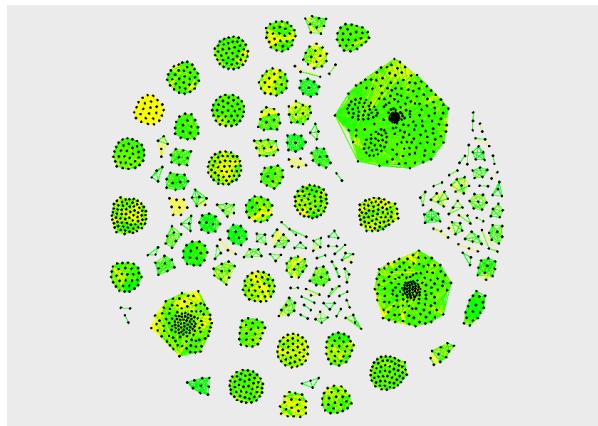
- ◆ Naïve approach: ask oracle node pair queries randomly.
 - May need to ask $O(n^2)$ queries for complete entity resolution.
- ◆ Smarter approach: ask oracle queries **guided by probabilities**.

Entity Resolution Using an Oracle

- ◆ Input data: modeled as a noisy graph.
- ◆ Output: a set of clusters = entities.
- ◆ **Formal problem** [WL+13, VBD14]:
 - Given an **oracle** that can **correctly answer** if a node pair is a match, what is an optimal strategy to ask oracle queries so as to minimize the number of queries for **correctly** resolving the entire graph?
- ◆ Motivation: reduce crowdsourcing ER cost for data set.

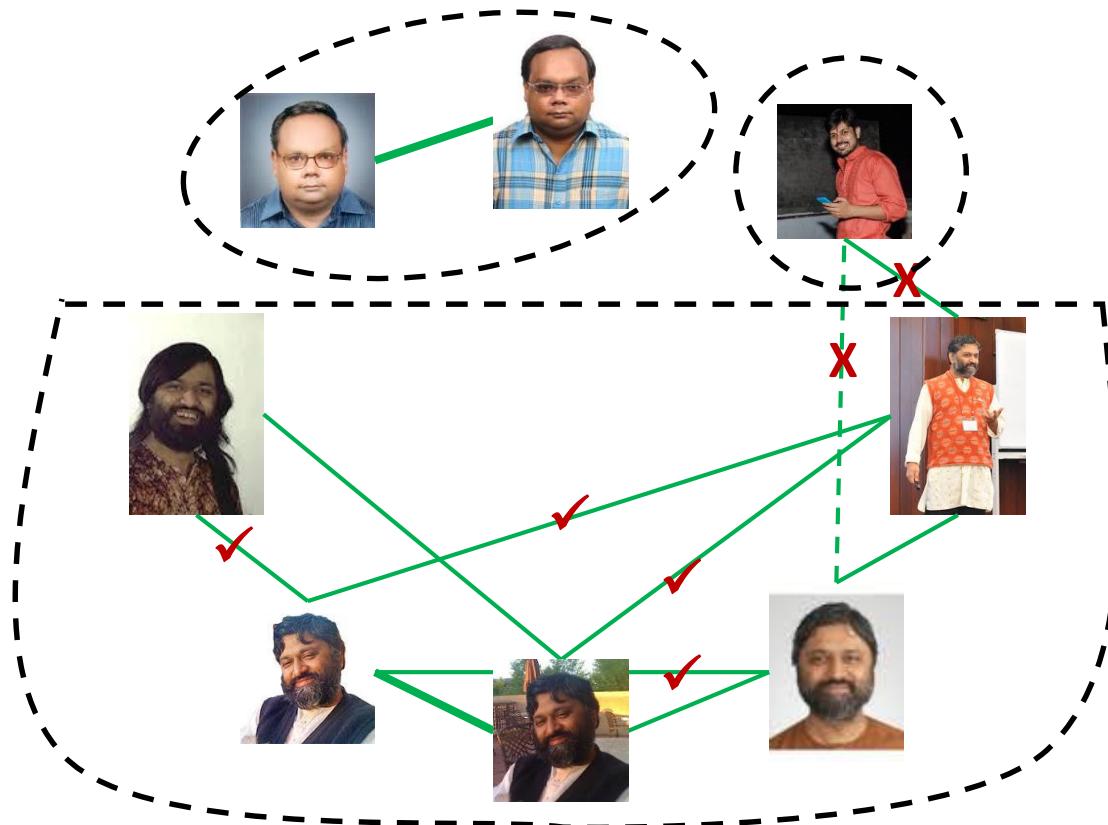


Online Entity Resolution Using an Oracle



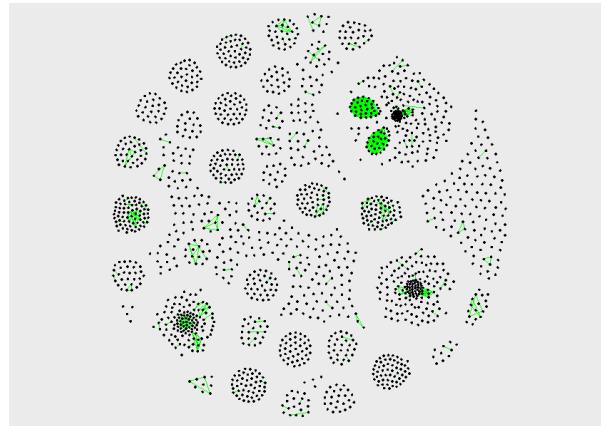
- ◆ Motivation: limited resolution time, early user termination.
- ◆ New formal problem:
 - Given an **oracle** that can **correctly** or **noisily answer** if a node pair is a match, what is an optimal strategy to ask oracle queries so as to **maximize progressive recall** wrt the sequence of oracle queries?
 - Progressive recall = area under “recall vs query sequence” curve.

Edge Ordering [WL+13]



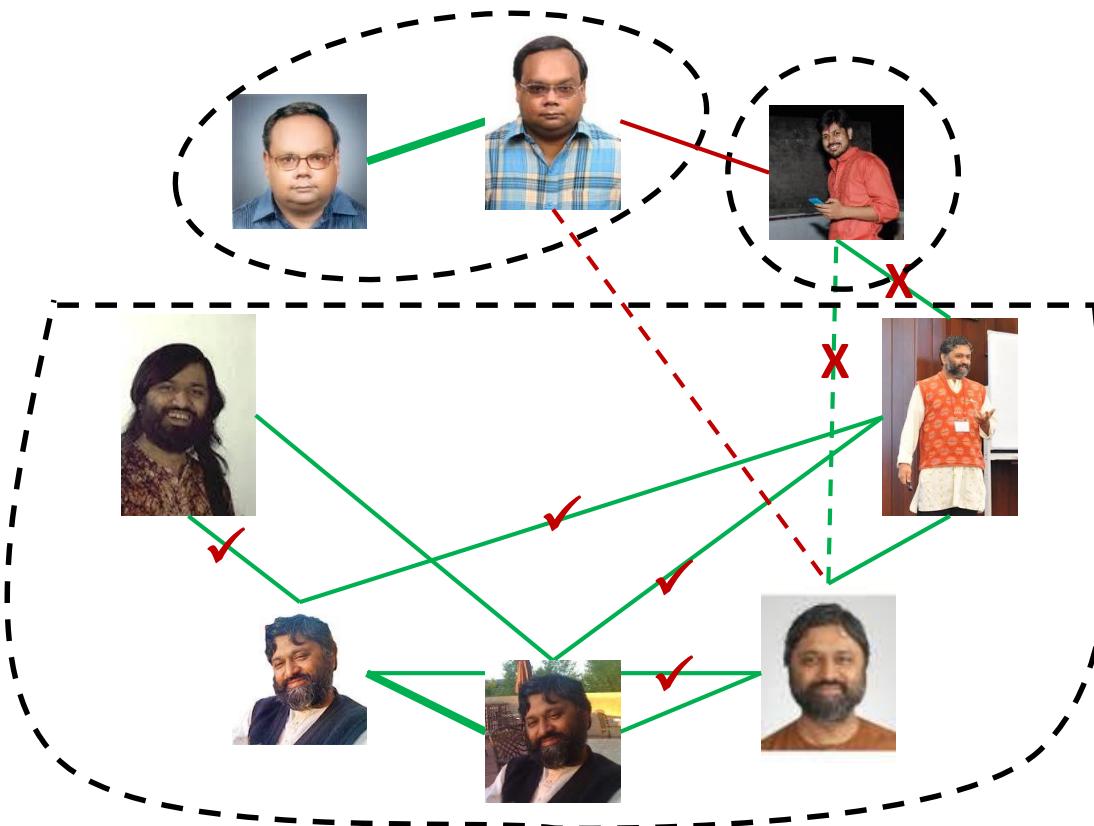
- ♦ EO: ask oracle queries in \downarrow **edge probability order**.
 - Can grow multiple clusters and sub-clusters in parallel.

Edge Ordering



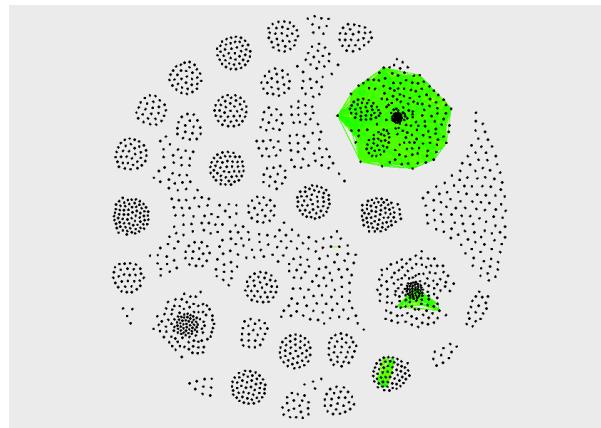
- ◆ Optimal strategy needs to ask $N - K + (K \text{ choose } 2)$ oracle queries.
 - Takes advantage of (matching and non-matching) transitivity.
- ◆ EO: ask oracle queries in **↓ edge probability order**.
 - Can grow multiple clusters and sub-clusters in parallel.
 - Worst-case approximation ratio of $O(N)$ [VBD14].

Node Ordering [VBD14]



- ◆ NO: process nodes in **↓ order of their expected cluster sizes**.
 - Ask oracle queries in **↓ edge probability order to processed nodes**.
 - Does not grow multiple sub-clusters of a cluster in parallel.

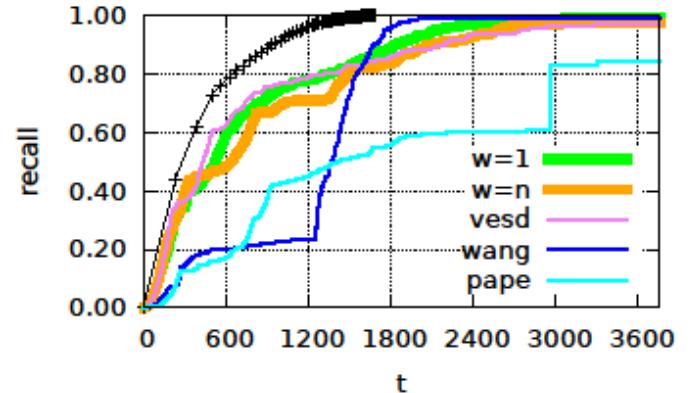
Node Ordering



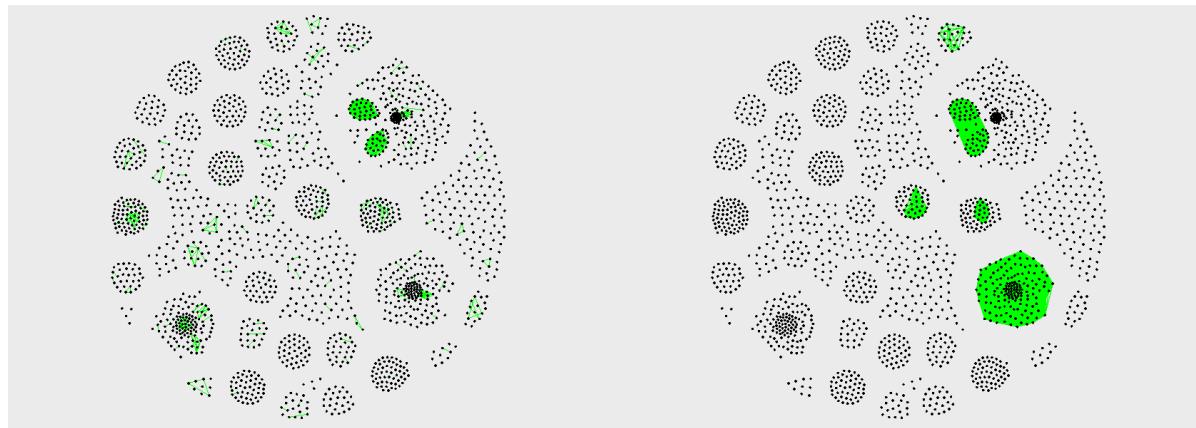
- ◆ Optimal strategy needs to ask $N - K + (K \text{ choose } 2)$ oracle queries.
 - Takes advantage of (matching and non-matching) transitivity.
- ◆ NO: process nodes in **↓ order of their expected cluster sizes**.
 - Ask oracle queries in **↓ edge probability order to processed nodes**.
 - Can grow similar-sized clusters (but not sub-clusters) in parallel.
 - Worst-case approximation ratio of $O(K)$ [VBD14].

Progressive Recall, Benefit Metric

- ◆ Progressive recall = area under “recall vs query sequence” curve
- ◆ Maximizing progressive recall:
 - Grow one cluster at a time in **decreasing order of true cluster sizes**
 - Need to ask $N - K + (K \text{ choose } 2)$ oracle queries
- ◆ Benefit metric: robust estimate of **marginal gain in recall**
 - Input: $p(u, v) = \text{probability that nodes } u \text{ and } v \text{ are matching}$
 - Benefit of record pair (u, v) : $b_e(u, v) = |C_T(u)| * |C_T(v)| * p(u, v)$
 - Benefit of adding node u to cluster C : $b_n(u, C) = \sum p(u, v), v \in C$
 - Benefit of processing node u : $b_n(u) = \max b_n(u, C)$, for all clusters C

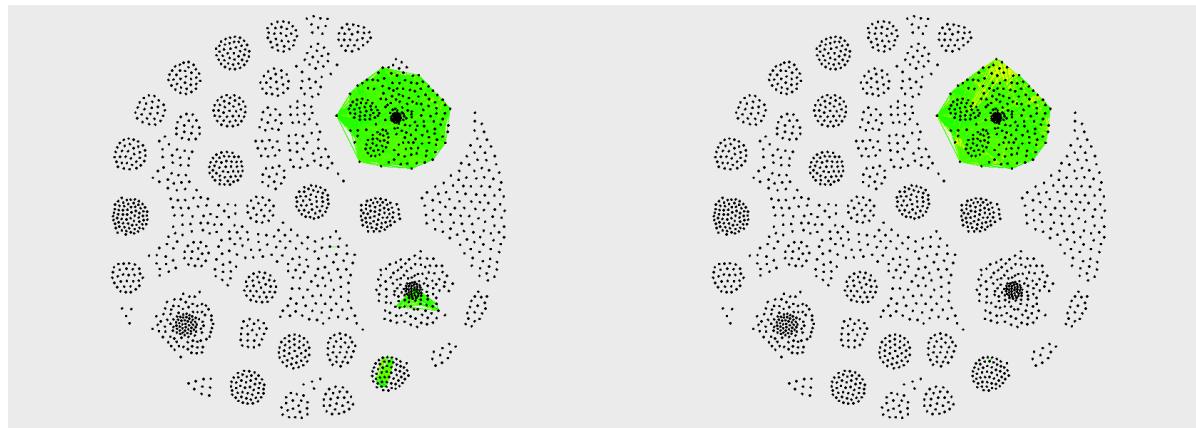


Progressive Recall Oracle Strategy [FSS16]



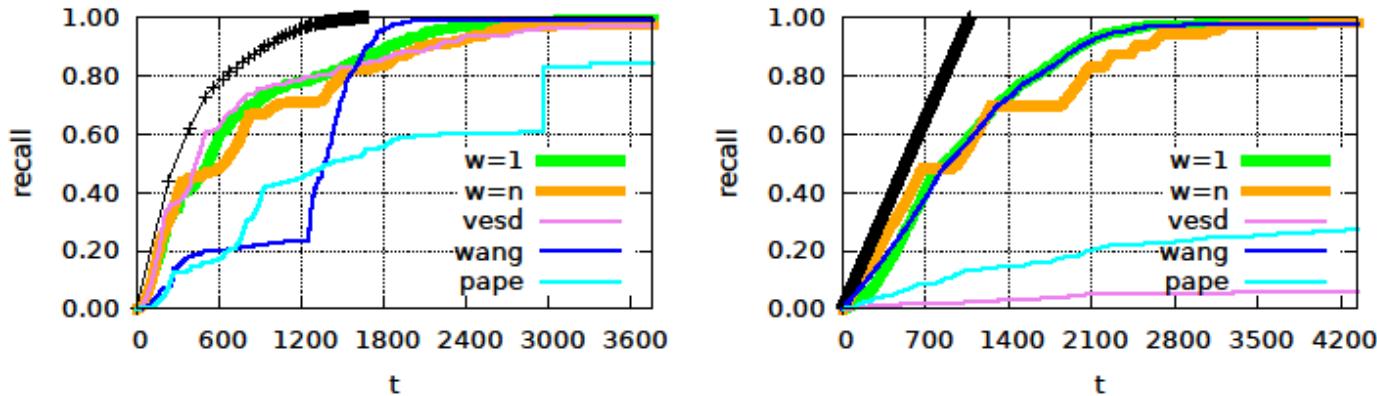
- ◆ **Edge ordering:** use benefit metric instead of edge probability
 - Iteratively query oracle with (u, v) having **highest value of $b_e(u, v)$**
 - Initially, edge with highest value of $p(u, v)$ is queried
 - Subsequently, can query lower probability, higher benefit edge

Progressive Recall Oracle Strategy [FSS16]



- ◆ **Hybrid ordering:** use node ordering, then edge ordering
 - Iteratively: select node u with **highest value of $b_n(u)$** , then query oracle with (u, v) , $v \in C$, in **decreasing order of $b_n(u, C)$**
 - Heuristic: use a threshold on benefit $b_n(u, C)$
 - Finally, process non-inferable edges (u, v) in **\downarrow order of $b_e(u, v)$**

Experimental Results



- ◆ Progressive recall of **sequential strategies**
 - Cora data set (skewed cluster sizes): Hybrid \approx [VBD14] $>>$ [WLK+13]
 - Prod data set (small cluster sizes): Hybrid \approx [WLK+13] $>>$ [VBD14]
- ◆ Key result: hybrid strategy is **robust**, dominates [WLK+13, VBD14]

Questions? Suggestions? Criticisms?

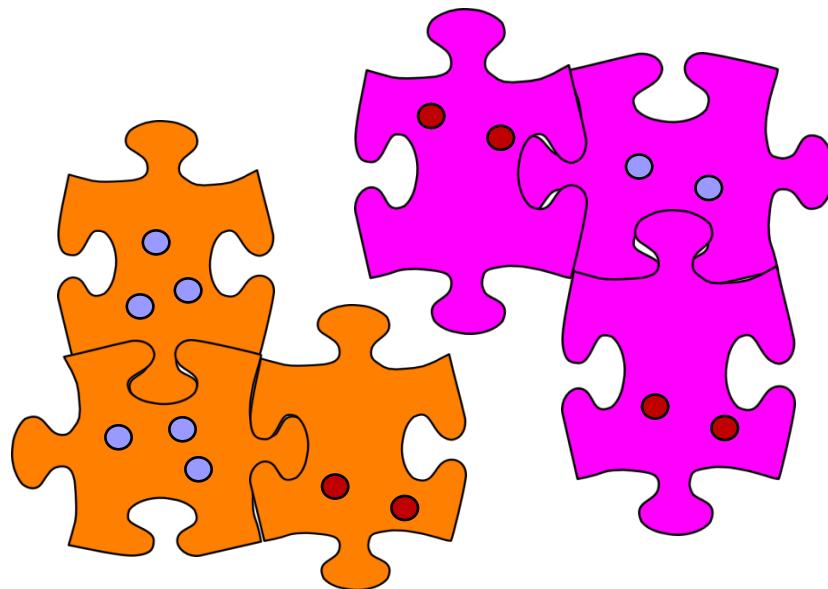


Outline

- ◆ Motivation
- ◆ Schema alignment
- ◆ Record linkage
- ◆ Data fusion
 - Overview
 - Techniques for big data
- ◆ Emerging topics

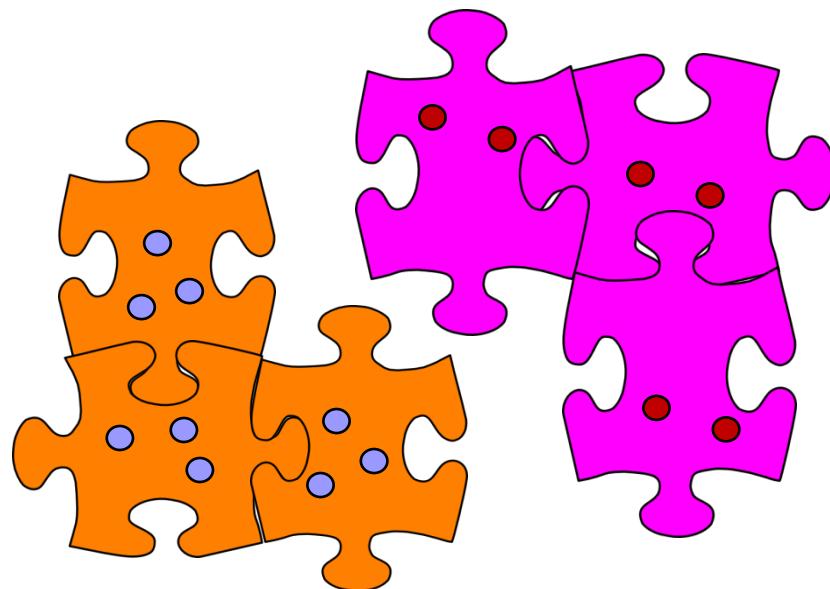
Data Fusion

- ◆ Reconciliation of conflicting content: pattern



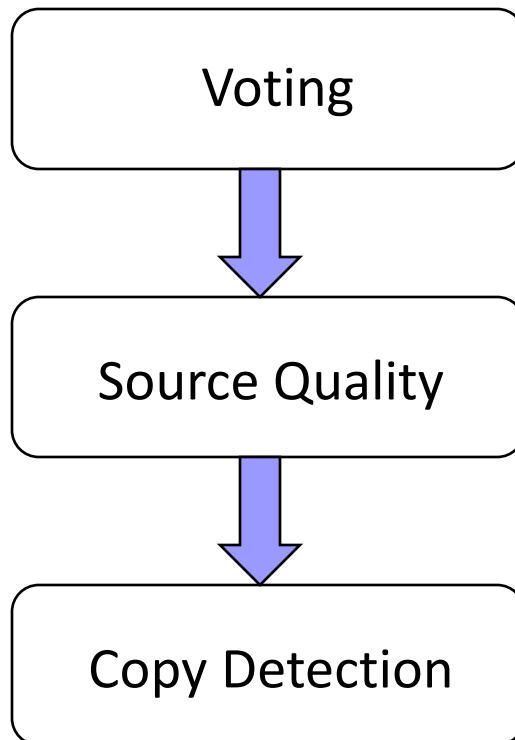
Data Fusion

- ◆ Reconciliation of conflicting content: pattern



Data Fusion: Three Components [DBS09a]

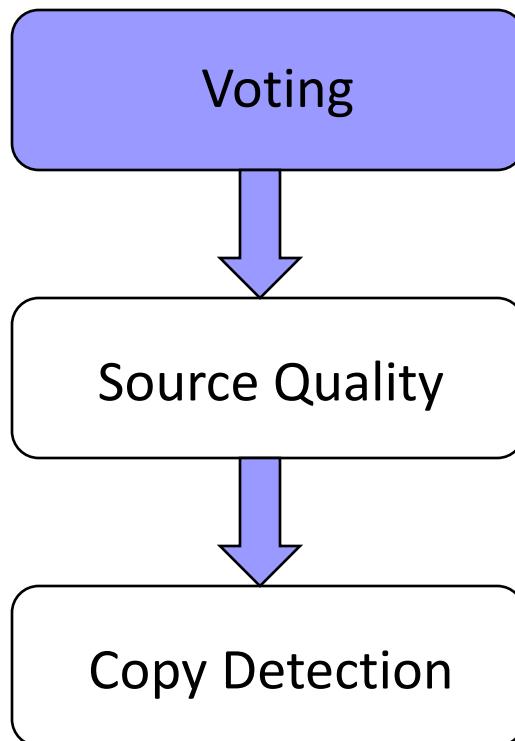
- ◆ Data fusion: voting + source quality + copy detection
 - Resolves inconsistency across diversity of sources



| | S1 | S2 | S3 | S4 | S5 |
|-----------|-----|------------|------------|------------|------------|
| Jagadish | UM | <u>ATT</u> | UM | UM | <u>UI</u> |
| Dewitt | MSR | MSR | <u>UW</u> | <u>UW</u> | <u>UW</u> |
| Bernstein | MSR | MSR | MSR | MSR | MSR |
| Carey | UCI | <u>ATT</u> | <u>BEA</u> | <u>BEA</u> | <u>BEA</u> |
| Franklin | UCB | UCB | <u>UMD</u> | <u>UMD</u> | <u>UMD</u> |

Data Fusion: Three Components [DBS09a]

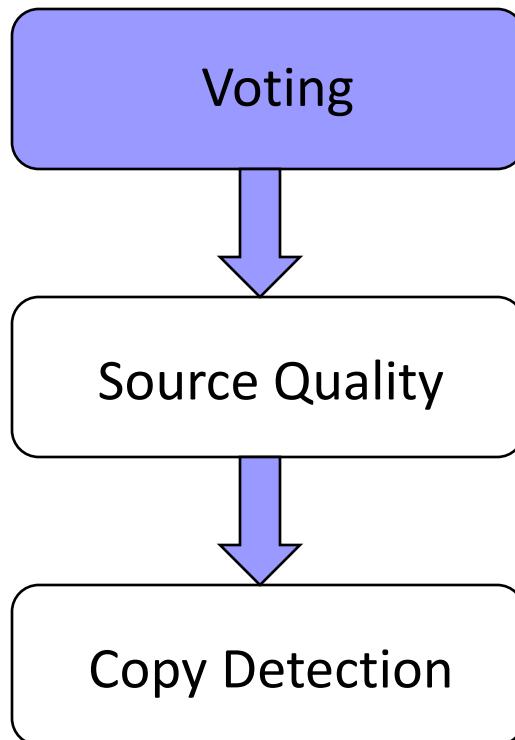
- ◆ Data fusion: voting + source quality + copy detection



| | S1 | S2 | S3 |
|-----------|-----|-----|-----|
| Jagadish | UM | ATT | UM |
| Dewitt | MSR | MSR | UW |
| Bernstein | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA |
| Franklin | UCB | UCB | UMD |

Data Fusion: Three Components [DBS09a]

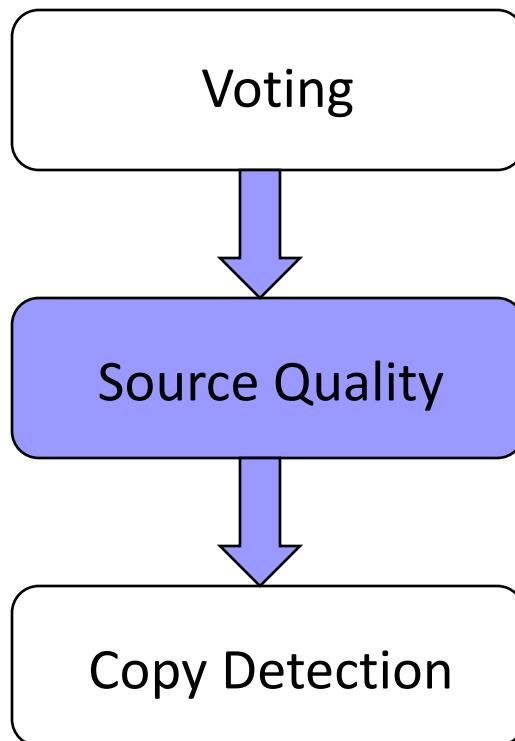
- ◆ Data fusion: voting + source quality + copy detection
 - Supports difference of opinion



| | S1 | S2 | S3 |
|-----------|-----|-----|-----|
| Jagadish | UM | ATT | UM |
| Dewitt | MSR | MSR | UW |
| Bernstein | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA |
| Franklin | UCB | UCB | UMD |

Data Fusion: Three Components [DBS09a]

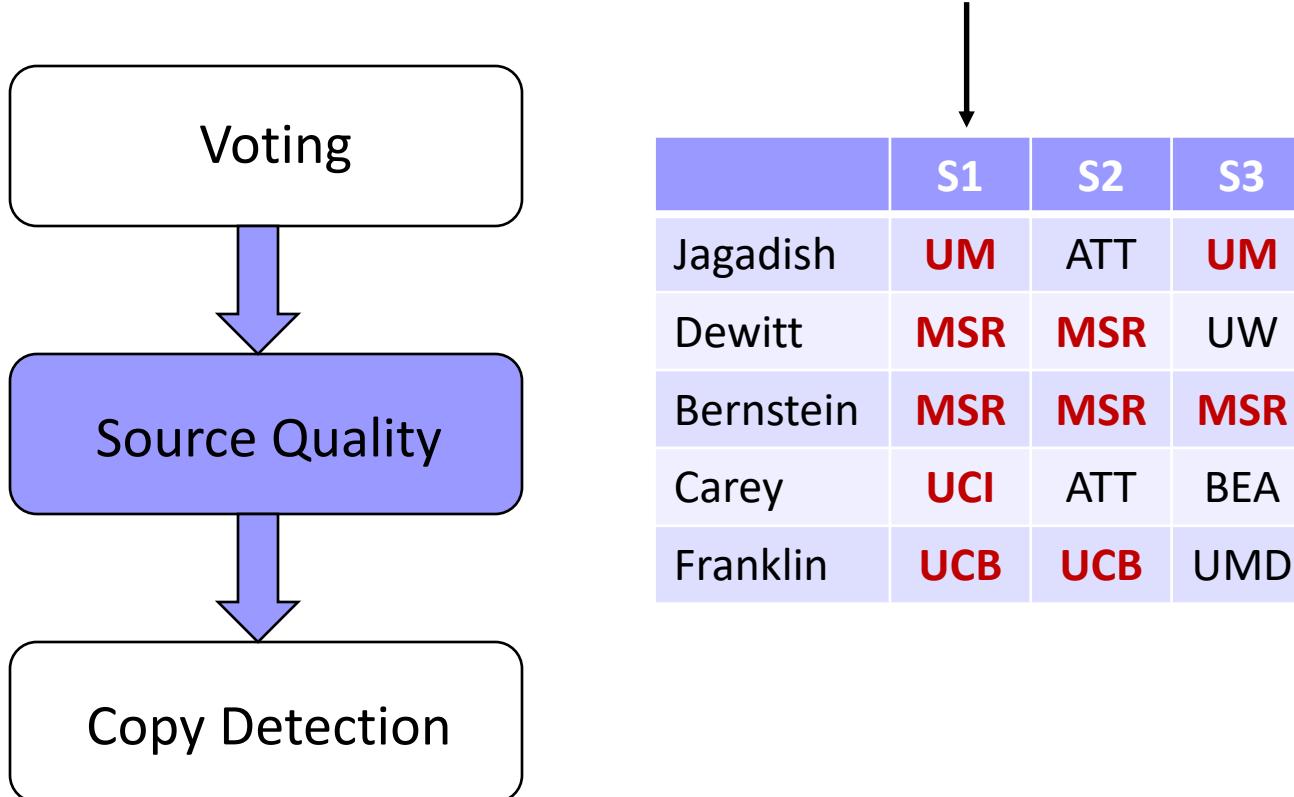
- ◆ Data fusion: voting + source quality + copy detection



| | S1 | S2 | S3 |
|-----------|-----|-----|-----|
| Jagadish | UM | ATT | UM |
| Dewitt | MSR | MSR | UW |
| Bernstein | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA |
| Franklin | UCB | UCB | UMD |

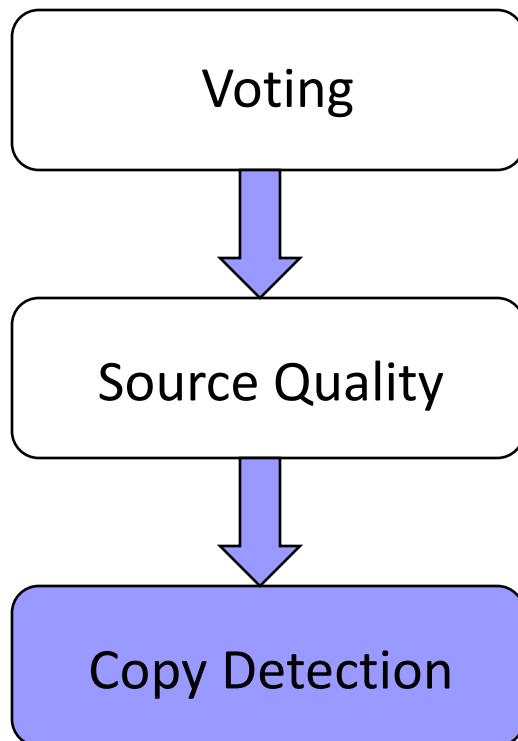
Data Fusion: Three Components [DBS09a]

- ◆ Data fusion: voting + source quality + copy detection
 - Gives more weight to knowledgeable sources



Data Fusion: Three Components [DBS09a]

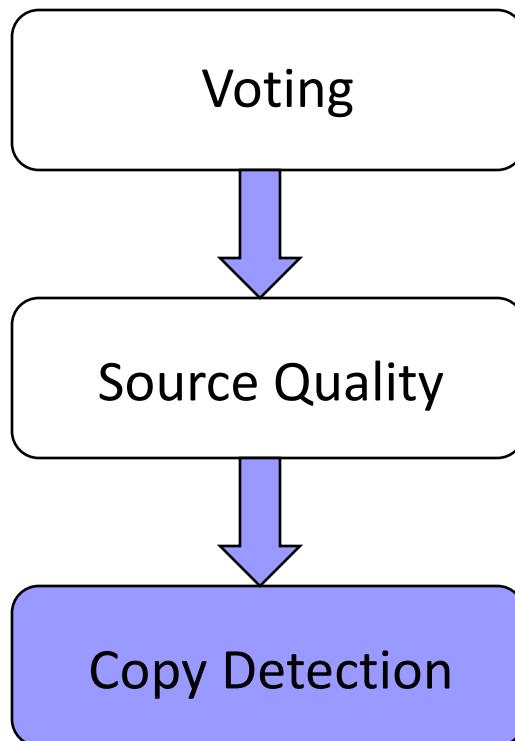
- ◆ Data fusion: voting + source quality + copy detection



| | S1 | S2 | S3 | S4 | S5 |
|-----------|-----|-----|-----|-----|-----|
| Jagadish | UM | ATT | UM | UM | UI |
| Dewitt | MSR | MSR | UW | UW | UW |
| Bernstein | MSR | MSR | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA | BEA | BEA |
| Franklin | UCB | UCB | UMD | UMD | UMD |

Data Fusion: Three Components [DBS09a]

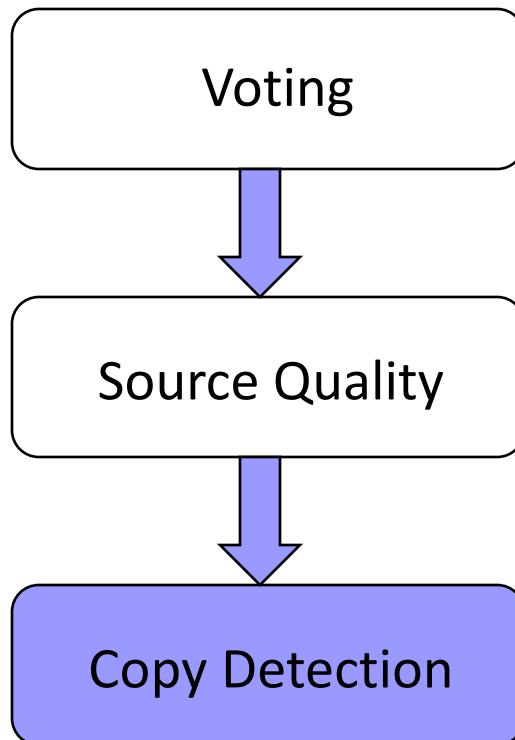
- ◆ Data fusion: voting + source quality + copy detection



| | S1 | S2 | S3 | S4 | S5 |
|-----------|-----|-----|-----|-----|-----|
| Jagadish | UM | ATT | UM | UM | UI |
| Dewitt | MSR | MSR | UW | UW | UW |
| Bernstein | MSR | MSR | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA | BEA | BEA |
| Franklin | UCB | UCB | UMD | UMD | UMD |

Data Fusion: Three Components [DBS09a]

- ◆ Data fusion: voting + source quality + copy detection
 - Reduces weight of copier sources



| | S1 | S2 | S3 | S4 | S5 |
|-----------|-----|-----|-----|-----|-----|
| Jagadish | UM | ATT | UM | UM | UI |
| Dewitt | MSR | MSR | UW | UW | UW |
| Bernstein | MSR | MSR | MSR | MSR | MSR |
| Carey | UCI | ATT | BEA | BEA | BEA |
| Franklin | UCB | UCB | UMD | UMD | UMD |

Outline

- ◆ Motivation
- ◆ Schema alignment
- ◆ Record linkage
- ◆ Data fusion
 - Overview
 - Techniques for big data
- ◆ Emerging topics

BDI: Data Fusion

◆ **Veracity**

- Using source trustworthiness [YHY08, GAM+10, PR11, YT11, GSH11, PR13]
- Combining source accuracy and copy detection [DBS09a, QAH+13]
- Multiple truth values [ZRG+12]
- Erroneous numeric data [ZH12]
- Experimental comparison on deep web data [LDL+13]

BDI: Data Fusion

- ◆ **Volume:**

- Online data fusion [LDO+11]

- ◆ **Velocity**

- Truth discovery for dynamic data [DBS09b, PRM+12]

- ◆ **Variety**

- Combining record linkage with data fusion [GDS+10]

Basic Solution: Naïve Voting

- ◆ Supports difference of opinion, allows conflict resolution
- ◆ Works well for independent sources that have similar accuracy
- ◆ When sources have different accuracies
 - Need to give more weight to votes by knowledgeable sources
- ◆ When sources copy from other sources
 - Need to reduce the weight of votes by copiers

Source Accuracy [YHY08, DBS09a]

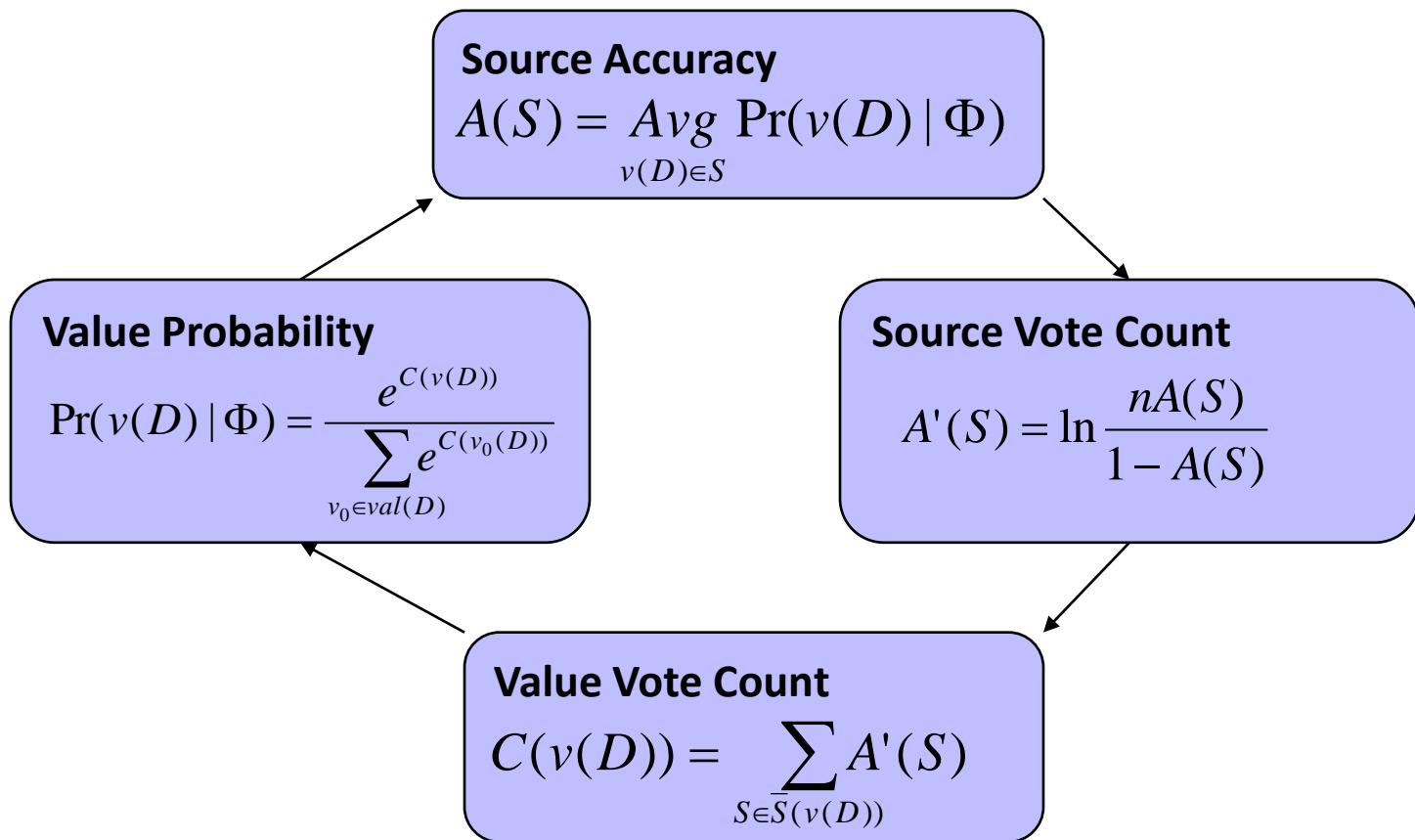
- ◆ Need to give more weight to knowledgeable sources
- ◆ Computing source accuracy: $A(S) = \text{Avg}_{v_i(D) \in S} \Pr(v_i(D) \text{ true} \mid \Phi)$
 - $v_i(D) \in S$: S provides value v_i on data item D
 - Φ : observations on all data items by sources S
 - $\Pr(v_i(D) \text{ true} \mid \Phi)$: probability of $v_i(D)$ being true
- ◆ How to compute $\Pr(v_i(D) \text{ true} \mid \Phi)$?

Source Accuracy

- ◆ Input: data item D, $\text{val}(D) = \{v_0, v_1, \dots, v_n\}$, Φ
- ◆ Output: $\Pr(v_i(D) \text{ true} \mid \Phi)$, for $i=0, \dots, n$ (sum=1)
- ◆ Based on Bayes Rule, need $\Pr(\Phi \mid v_i(D) \text{ true})$
 - Under independence, need $\Pr(\Phi_D(S) \mid v_i(D) \text{ true})$
 - If S provides v_i : $\Pr(\Phi_D(S) \mid v_i(D) \text{ true}) = A(S)$
 - If S does not : $\Pr(\Phi_D(S) \mid v_i(D) \text{ true}) = (1-A(S))/n$
- ◆ Challenge:
 - Inter-dependence between source accuracy and value probability?

Source Accuracy

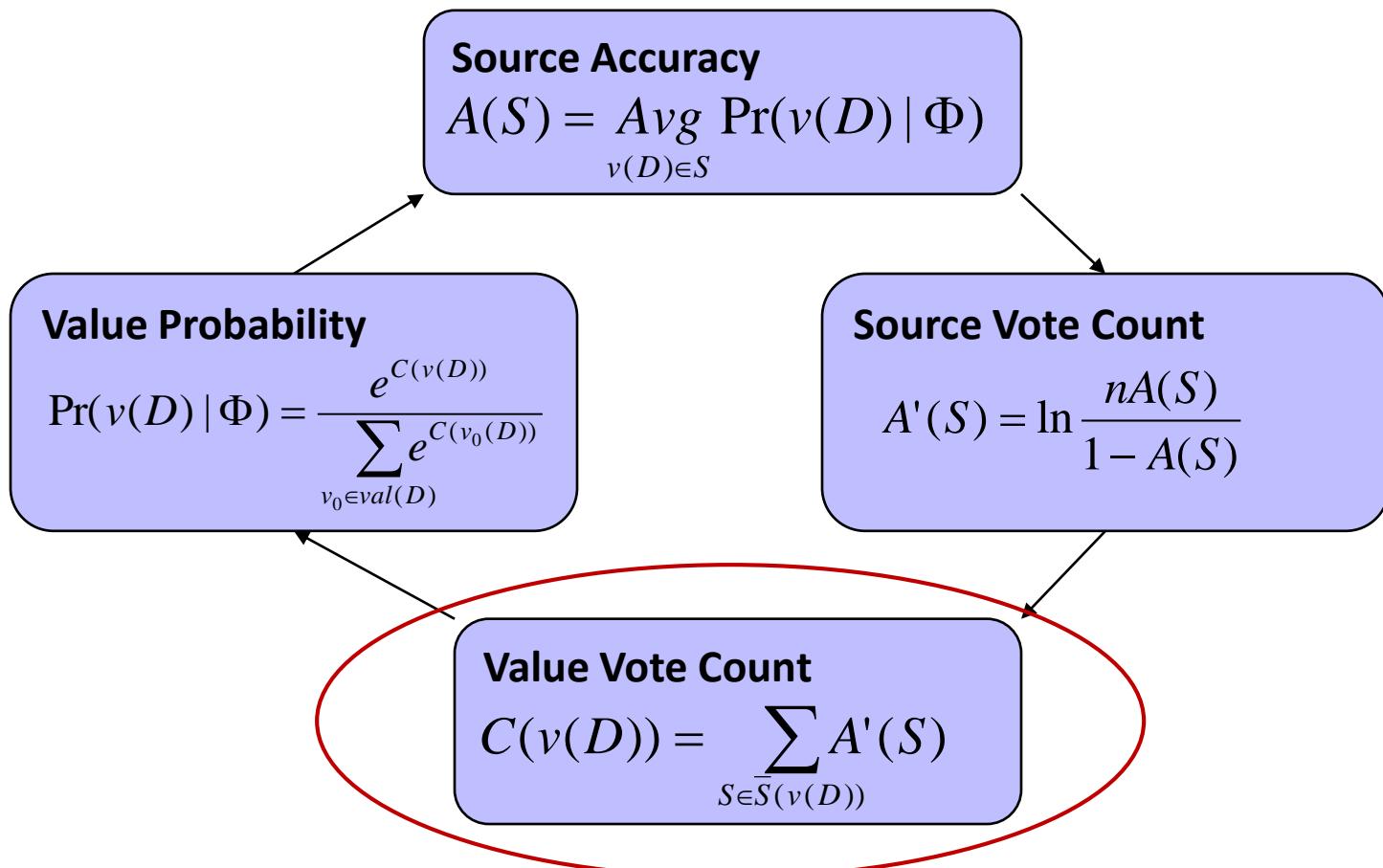
- ◆ Continue until source accuracy converges



Source Accuracy

- ◆ Continue until source accuracy converges

- Use value similarity in vote count $C^*(v) = C(v) + \rho \sum_{v' \neq v} C(v') \bullet sim(v, v')$



Copy Detection

Are Source 1 and Source 2 dependent? Not necessarily

Source 1 on USA Presidents:

1st : George Washington

2nd : John Adams

3rd : Thomas Jefferson

4th : James Madison

...

41st : George H.W. Bush

42nd : William J. Clinton

43rd : George W. Bush

44th : Barack Obama

Source 2 on USA Presidents:

1st : George Washington

2nd : John Adams

3rd : Thomas Jefferson

4th : James Madison

...

41st : George H.W. Bush

42nd : William J. Clinton

43rd : George W. Bush

44th : Barack Obama



Copy Detection

Are Source 1 and Source 2 dependent? Very likely

Source 1 on USA Presidents:

1st : George Washington

2nd : Benjamin Franklin

3rd : John F. Kennedy

4th : Abraham Lincoln

Source 2 on USA Presidents:

1st : George Washington

2nd : Benjamin Franklin

3rd : John F. Kennedy

4th : Abraham Lincoln

...

...

41st : George W. Bush

41st : George W. Bush

42nd : Hillary Clinton

42nd : Hillary Clinton

43rd : Dick Cheney

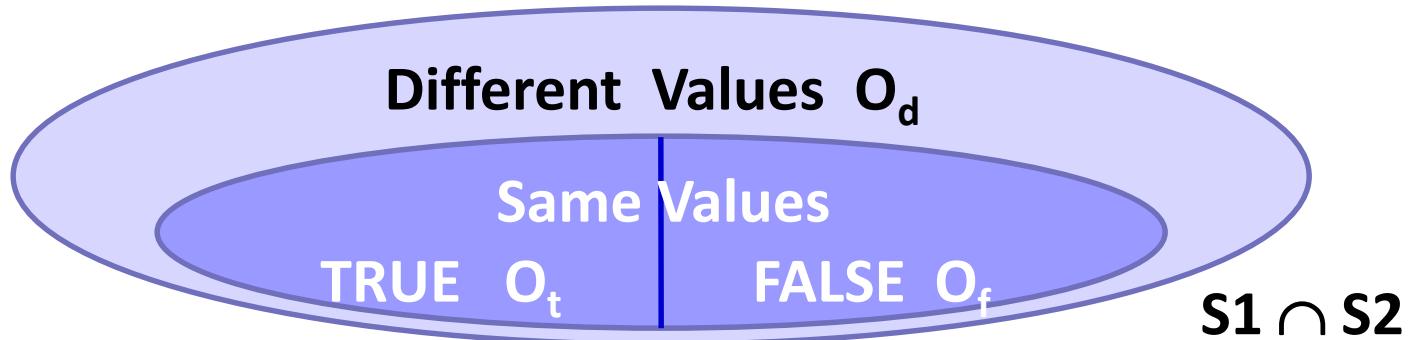
43rd : Dick Cheney

44th : Barack Obama

44th : John McCain

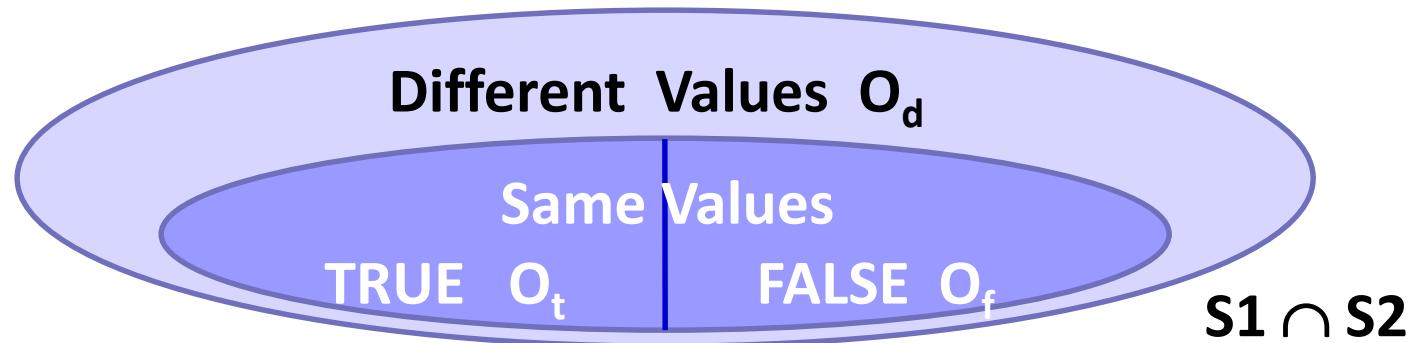


Copy Detection: Bayesian Analysis



- ◆ Goal: $\Pr(S_1 \perp S_2 | \Phi)$, $\Pr(S_1 \sim S_2 | \Phi)$ (sum = 1)
- ◆ According to Bayes Rule, we need $\Pr(\Phi | S_1 \perp S_2)$, $\Pr(\Phi | S_1 \sim S_2)$
- ◆ Key: compute $\Pr(\Phi_D | S_1 \perp S_2)$, $\Pr(\Phi_D | S_1 \sim S_2)$, for each $D \in S_1 \cap S_2$

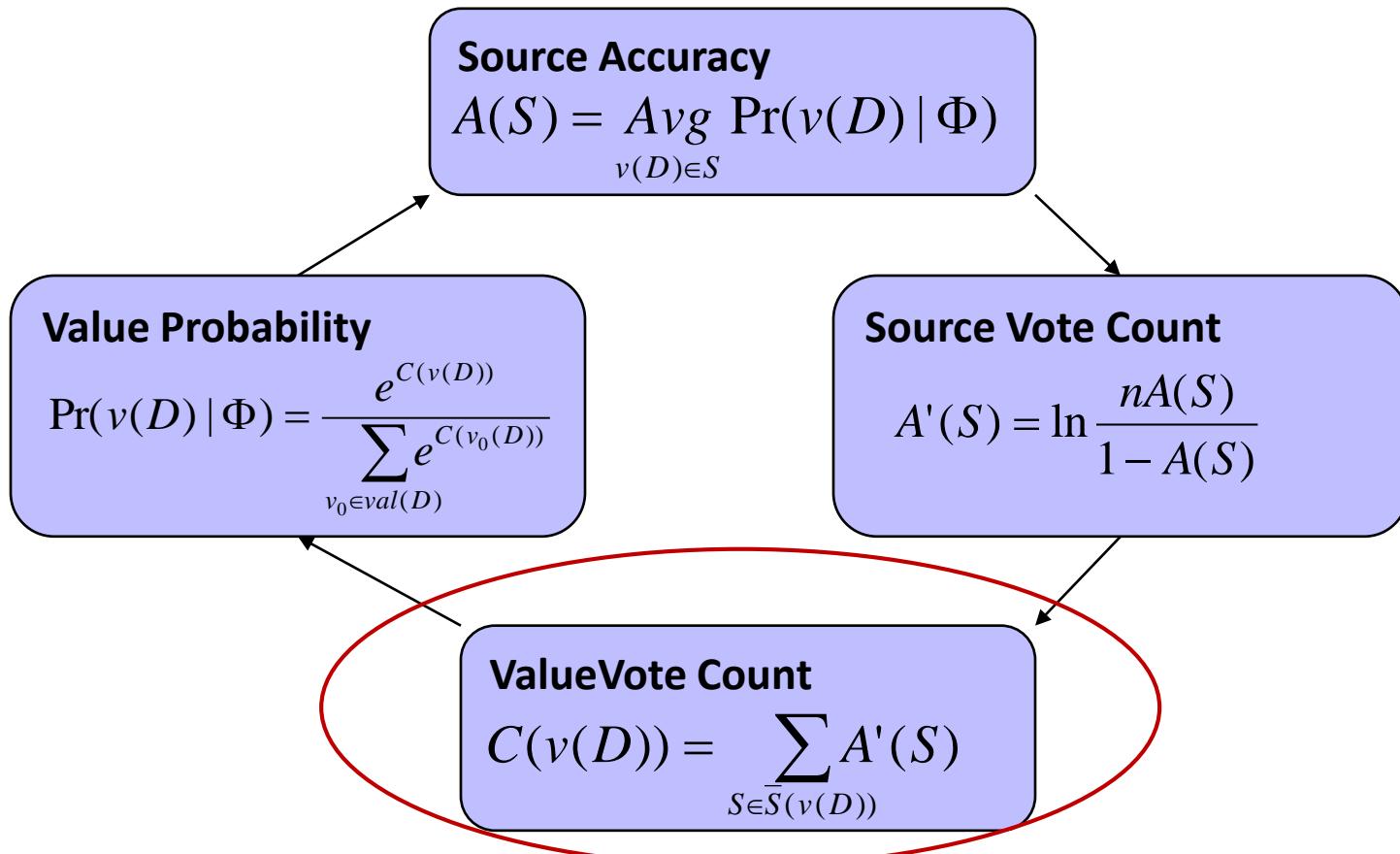
Copy Detection: Bayesian Analysis



| Pr | Independence | Copying |
|-------|---------------------------------------|--|
| O_t | A^2 | $A \cdot c + A^2(1 - c)$ |
| O_f | $\frac{(1 - A)^2}{n}$ | $(1 - A) \bullet c + \frac{(1 - A)^2}{n}(1 - c)$ |
| O_d | $P_d = 1 - A^2 - \frac{(1 - A)^2}{n}$ | $P_d(1 - c)$ |

Discount Copied Values

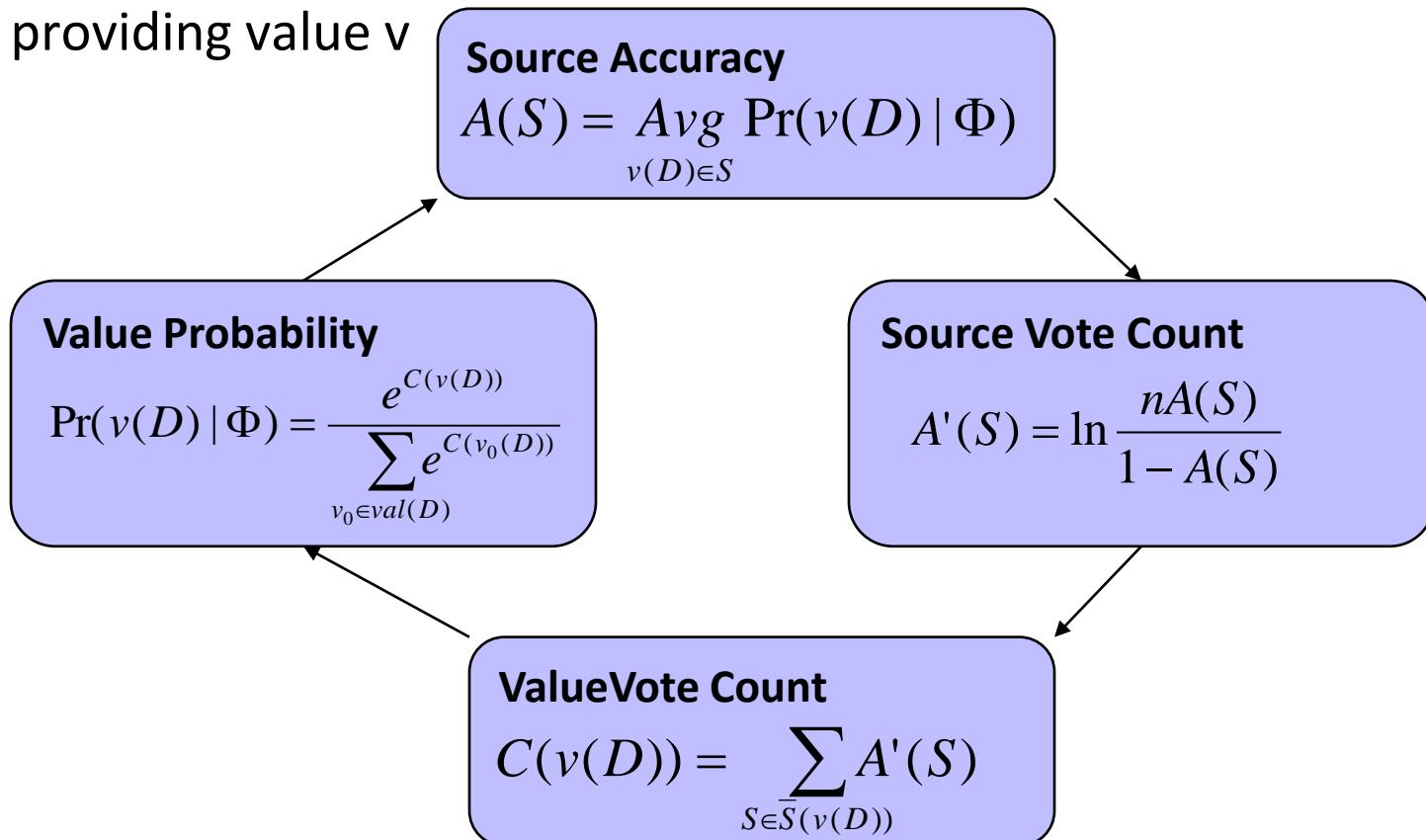
- ◆ Continue until convergence



Discount Copied Values

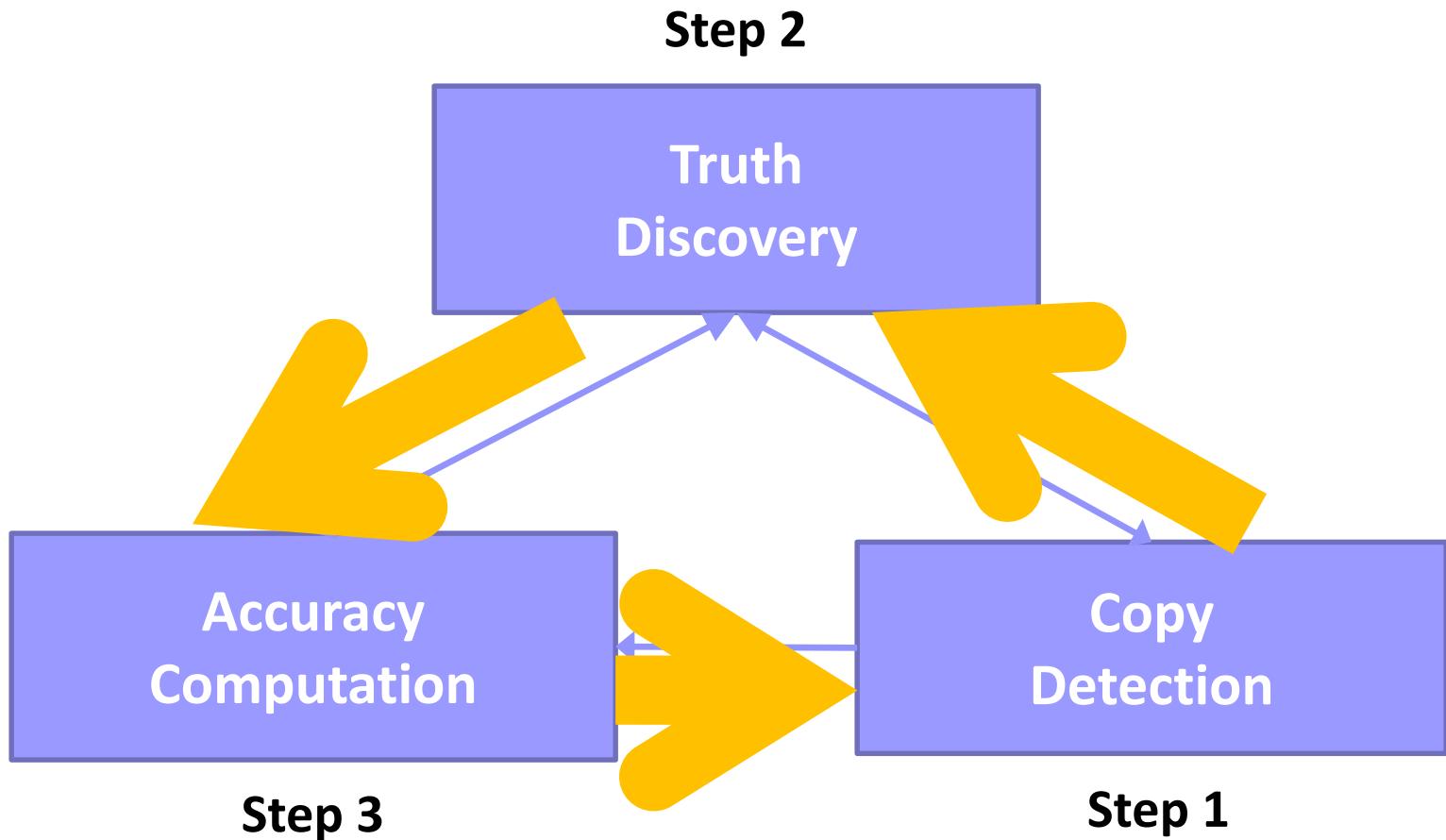
- ◆ Continue until convergence

- Consider dependence $C(v) = \sum_{S \in S(v)} A'(S) \bullet I(S)$
- $I(S)$: pr of independently providing value v



Iterative Process

- ◆ Typically converges when #objs >> #srcs



Challenges in a Dynamic World [DBS09b]

| | S1 | S2 | S3 | S4 | S5 |
|-------------|--------|--------|------|------|------|
| Stonebraker | MIT | UCB | MIT | MIT | MS |
| Dewitt | MSR | MSR | Wisc | Wisc | Wisc |
| Bernstein | MSR | MSR | MSR | MSR | MSR |
| Carey | UCI | AT&T | BEA | BEA | BEA |
| Halevy | Google | Google | UW | UW | UW |

Challenges in a Dynamic World [DBS09b]

| | S1 | S2 | S3 | S4 | S5 |
|--|--------------------------|--|---|-----------------------------------|--|
| Stonebraker (Θ , UCB), (02, <i>MIT</i>) | (03, MIT) | (00, UCB) <i>Out-of-date!</i> | (01, UCB) (06, MIT) | (05, MIT) | (03, UCB) (05, MS) <i>ERR!</i> |
| Dewitt (Θ , Wisc), (08, <i>MSR</i>) | (00, Wisc) (09, MSR) | (00, UW) (01, Wisc) (08, MSR) | (01, UW) (02, Wisc) <i>Out-of-date!</i> | (05, Wisc) <i>Out-of-date!</i> | (03, UW) (05, \perp) (07, Wisc) |
| Bernstein (Θ , <i>MSR</i>) | (00, MSR) | (00, MSR) | (01, MSR) | (07, MSR) | (03, MSR) |
| Carey (Θ , Propell), (02, BEA), (08, <i>UCI</i>) | (04, BEA) (09, UCI) | (05, AT&T) <i>ERR!</i> | (06, BEA) <i>Out-of-date!</i> | (07, BEA) <i>Out-of-date!</i> | (07, BEA) <i>Out-of-date!</i> |
| Halevy (Θ , UW), (05, <i>Google</i>) | (00, UW) (07, Google) | (00, Wisc) (02, UW) (05, Google) | (01, Wisc) (06, UW) <i>SLOW!</i> | (05, UW) <i>SLOW!</i> | (03, Wisc) (05, Google) (07, UW) <i>SLOW!</i> |

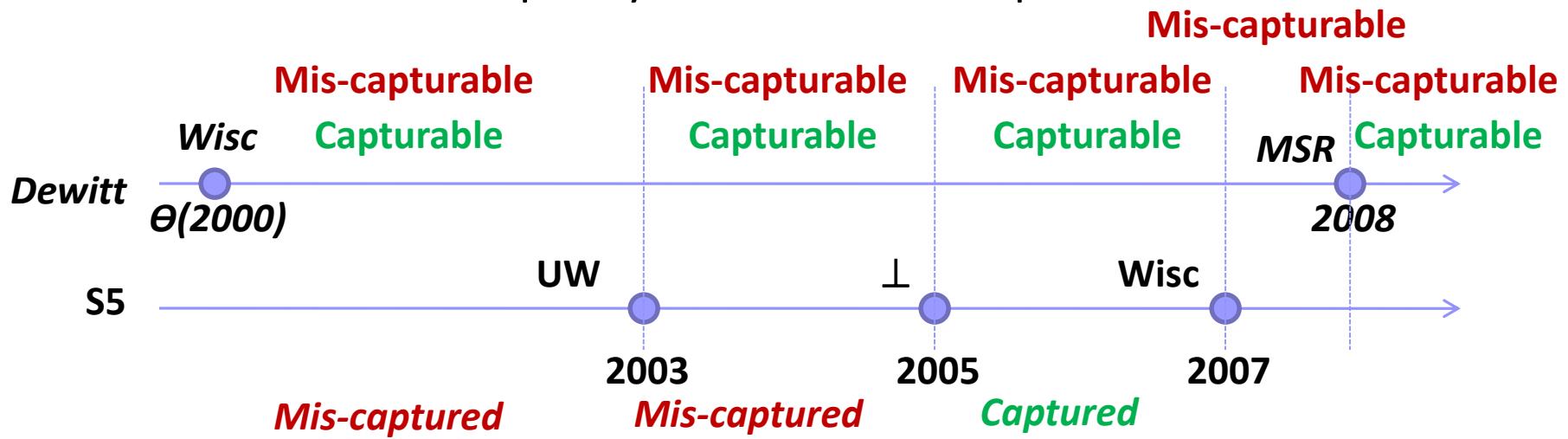
- ◆ True values can evolve over time
- ◆ Low-quality data can be caused by different reasons

Problem Definition

| Problem Definition | Static World | Dynamic World |
|--------------------|---|--|
| Objects | Each associated with a value; e.g., Google for Halevy | Each associated with a <i>lifespan</i> ; e.g., (00, UW), (05, Google) for Halevy |
| Sources | Each can provide a value for an object; e.g., S1 providing Google | Each can have a list of updates for an object; e.g., S1's updates for Halevy (00, UW), (07, Google) |
| Output | true value for each object | <ol style="list-style-type: none">1. Life span: true value for each object at each time point2. Copying: pr of S1 being a copier of S2 and pr of S1 being actively copying at each time point |

Quality of Data Sources

- ◆ CEF: three orthogonal quality measures
 - Coverage: how many transitions are **captured**
 - Exactness: how many transitions are not **mis-captured**
 - Freshness: how quickly transitions are captured



Coverage = #*Captured*/#*Capturable* (e.g., $\frac{1}{4}=.25$)

Exactness= $1 - \# \text{Mis-Captured} / \# \text{Mis-Capturable}$ (e.g., $1 - 2/5 = .6$)

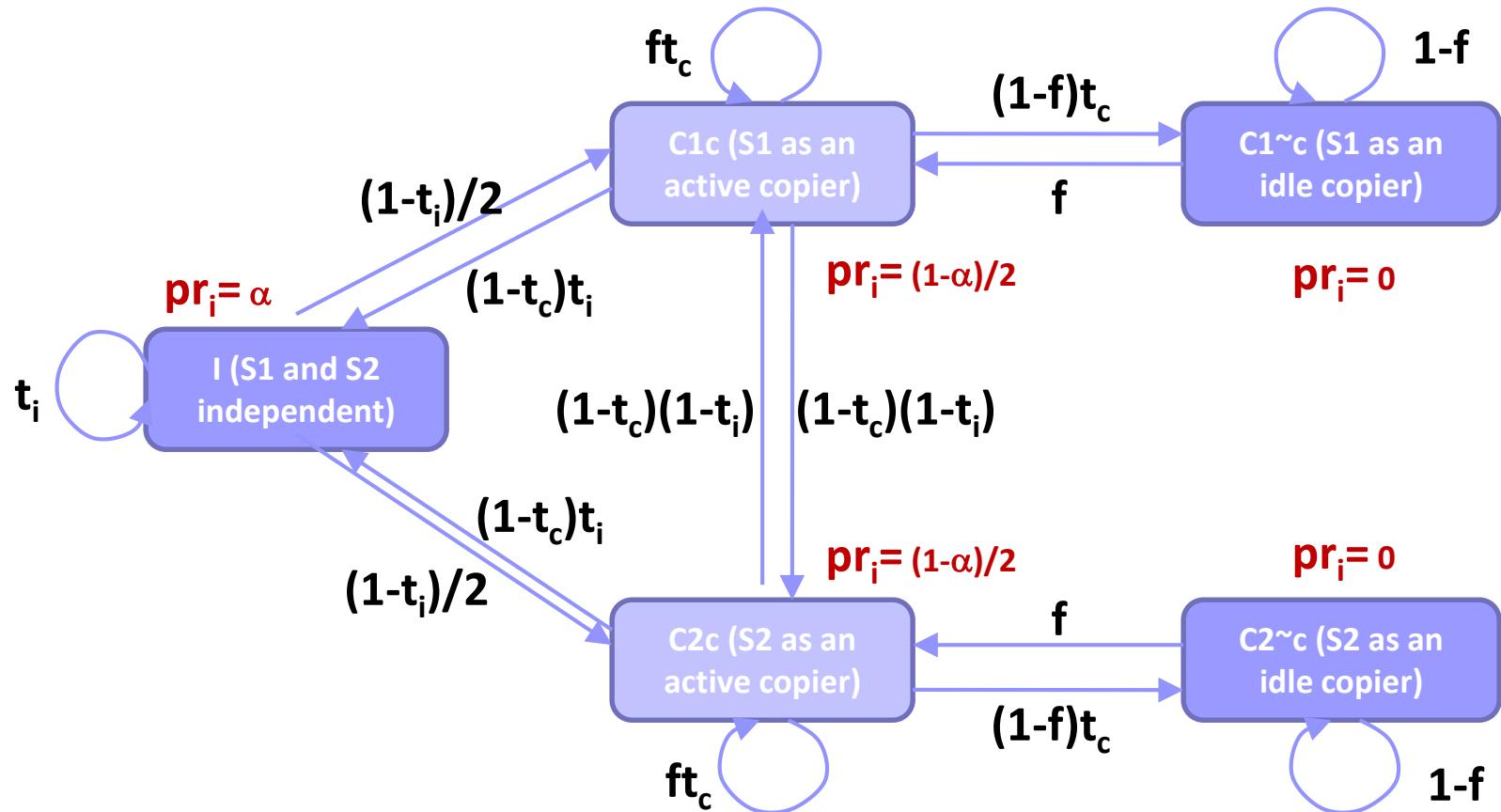
Freshness(Δ)= #(Captured w. length $\leq \Delta$)/#Captured (e.g., F(0)=0, F(1)=0, F(2)=1/1 = 1...)

Copy Detection

- ◆ Intuition: Copying is likely between S1 and S2 if
 - They make common mistakes
 - Overlapping updates are performed after real values have changed

| | S1 | S2 | S3 | S4 | S5 |
|--|--------------------------|--|------------------------|------------|--|
| Stonebraker (00, UCB), (02, MIT) | (03, MIT) | (00, UCB) | (01, UCB) (06, MIT) | (05, MIT) | (03, UCB) (05, MS) |
| Dewitt (00, Wisc), (08, MSR) | (00, Wisc) (09, MSR) | (00, UW) (01, Wisc) (08, MSR) | (01, UW) (02, Wisc) | (05, Wisc) | (03, UW) (05, ⊥) (07, Wisc) |
| Bernstein (00, MSR) | (00, MSR) | (00, MSR) | (01, MSR) | (07, MSR) | (03, MSR) |
| Carey (00, Propell), (02, BEA), (08, UCI) | (04, BEA) (09, UCI) | (05, AT&T) | (06, BEA) | (07, BEA) | (07, BEA) |
| Halevy (00, UW), (05, Google) | (00, UW) (07, Google) | (00, Wisc) (02, UW) (05, Google) | (01, Wisc) (06, UW) | (05, UW) | (03, Wisc) (05, Google) (07, UW) |

The Copying Detection HMM Model



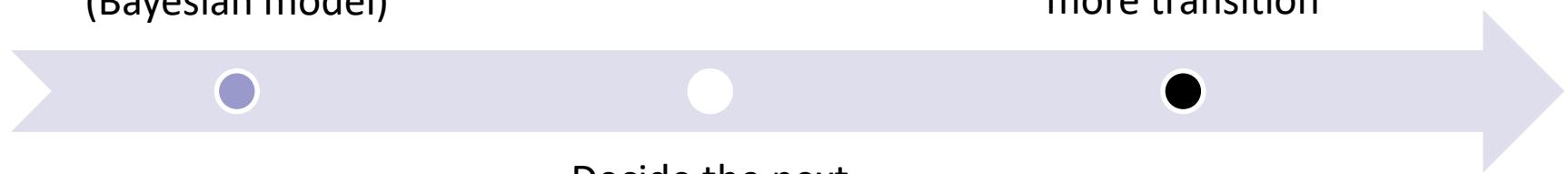
- ◆ A period of copying starts from and ends with a real copying
- ◆ α – Pr(init independence); f – Pr(a copier actively copying);
 t_i – Pr(remaining independent); t_c – Pr(remaining as a copier);

Lifespan Discovery

- ◆ Algorithm: for each object O

Decide the initial
value v_0
(Bayesian model)

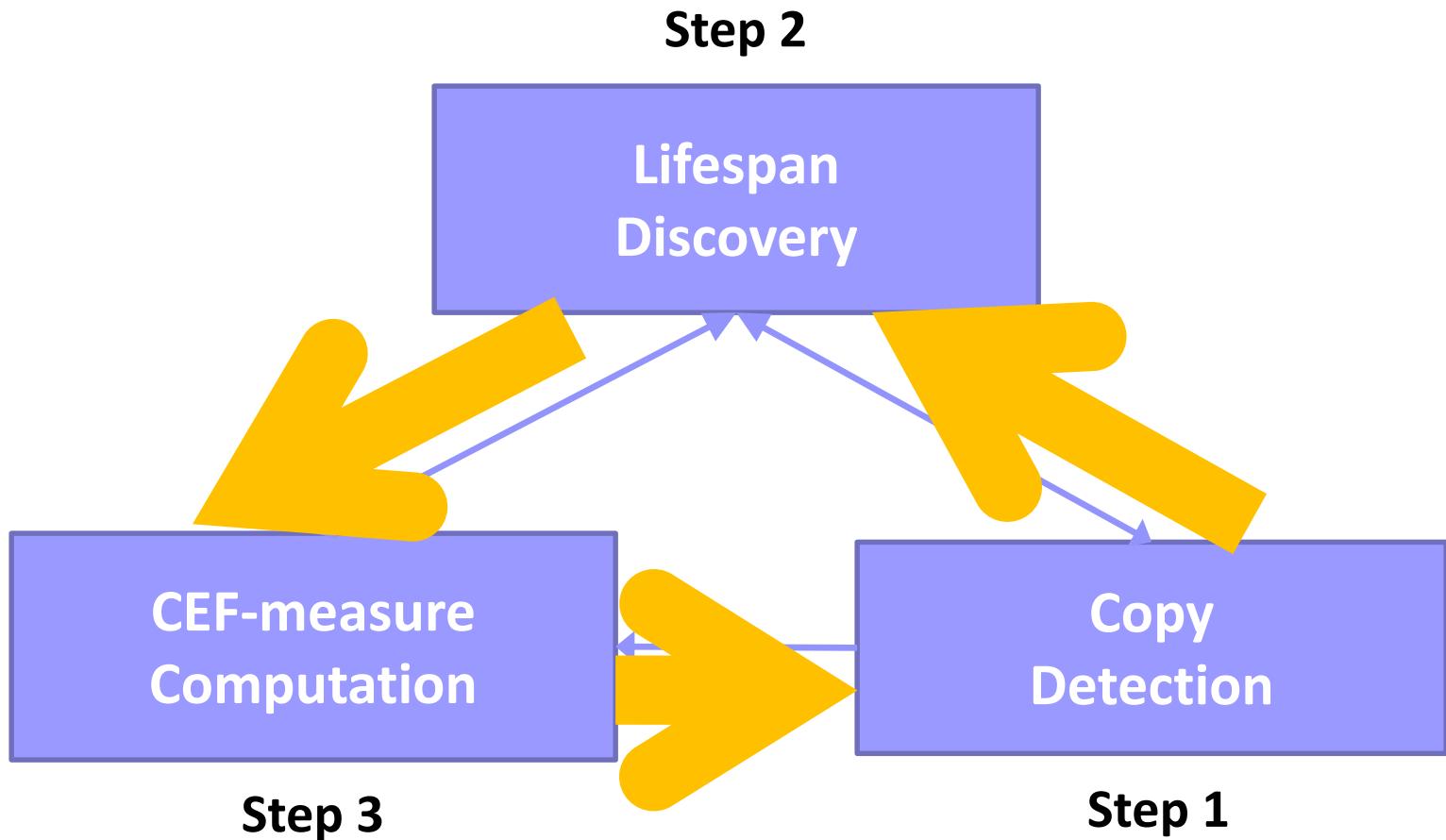
Terminate when no
more transition



Decide the next
transition (t, v)
(Bayesian model)

Iterative Process

- ◆ Typically converges when $\# \text{objs} \gg \# \text{srcs}$



Example Revisited

| | S1 | S2 | S3 | S4 | S5 |
|---|--------------------------|--|------------------------|----------|--|
| Halevy (θ, UW), (05, <i>Google</i>) | (00, UW) (07, Google) | (00, Wisc) (02, UW) (05, Google) | (01, Wisc) (06, UW) | (05, UW) | (03, Wisc) (05, Google) (07, UW) |

- ◆ Lifespan for Halevy and CEF-measure for S1 and S2

| Rnd | Halevy | C(S1) | E(S1) | F(S1,0) | F(S1,1) | C(S2) | E(S2) | F(S2,0) | F(S2,1) |
|-----|---|-------|-------|---------|---------|-------|-------|---------|---------|
| 0 | | .99 | .95 | .1 | .2 | .99 | .95 | .1 | .2 |
| 1 | (θ, Wisc) (2002, UW) (2003, Google) | .97 | .94 | .27 | .4 | .57 | .83 | .17 | .3 |
| 2 | (θ, UW) (2002, Google) | .92 | .99 | .27 | .4 | .64 | .8 | .18 | .27 |
| 3 | (θ, UW) (2005, Google) | .92 | .99 | .27 | .4 | .64 | .8 | .25 | .42 |

Summary

| | Schema alignment | Record linkage | Data fusion |
|----------|---|---|---|
| Volume | <ul style="list-style-type: none">Integrating deep webWeb tables | <ul style="list-style-type: none">Adaptive blockingMeta blocking | <ul style="list-style-type: none">Online fusion |
| Velocity | <ul style="list-style-type: none">Keyword-based integration for dynamic data | <ul style="list-style-type: none">Incremental linkage | <ul style="list-style-type: none">Fusion for dynamic data |
| Variety | <ul style="list-style-type: none">Data spacesKeyword-based integration | <ul style="list-style-type: none">Linking text to structured data | <ul style="list-style-type: none">Combining fusion with linkage |
| Veracity | | <ul style="list-style-type: none">Value-variety tolerant linkage | <ul style="list-style-type: none">Truth discovery |

Outline

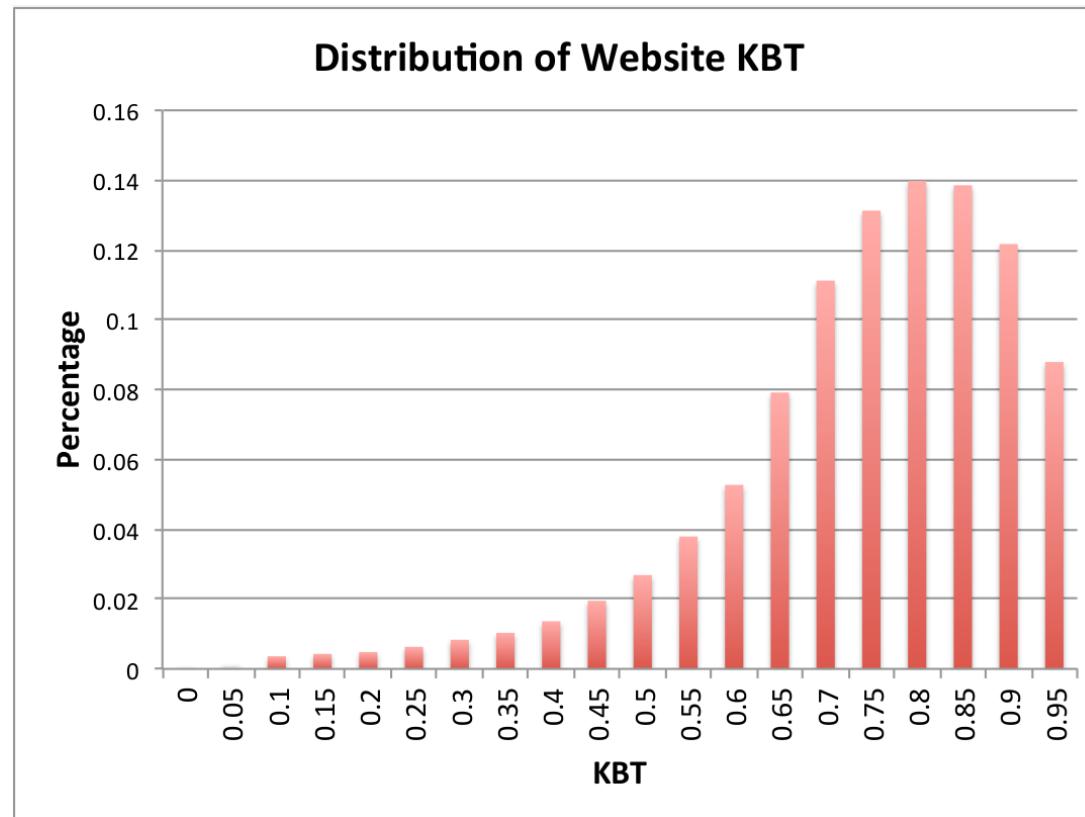
- ◆ Motivation
- ◆ Schema alignment
- ◆ Record linkage
- ◆ Data fusion
- ◆ Emerging topics
 - Knowledge-based trust, source selection, ...

BDI: Source Quality

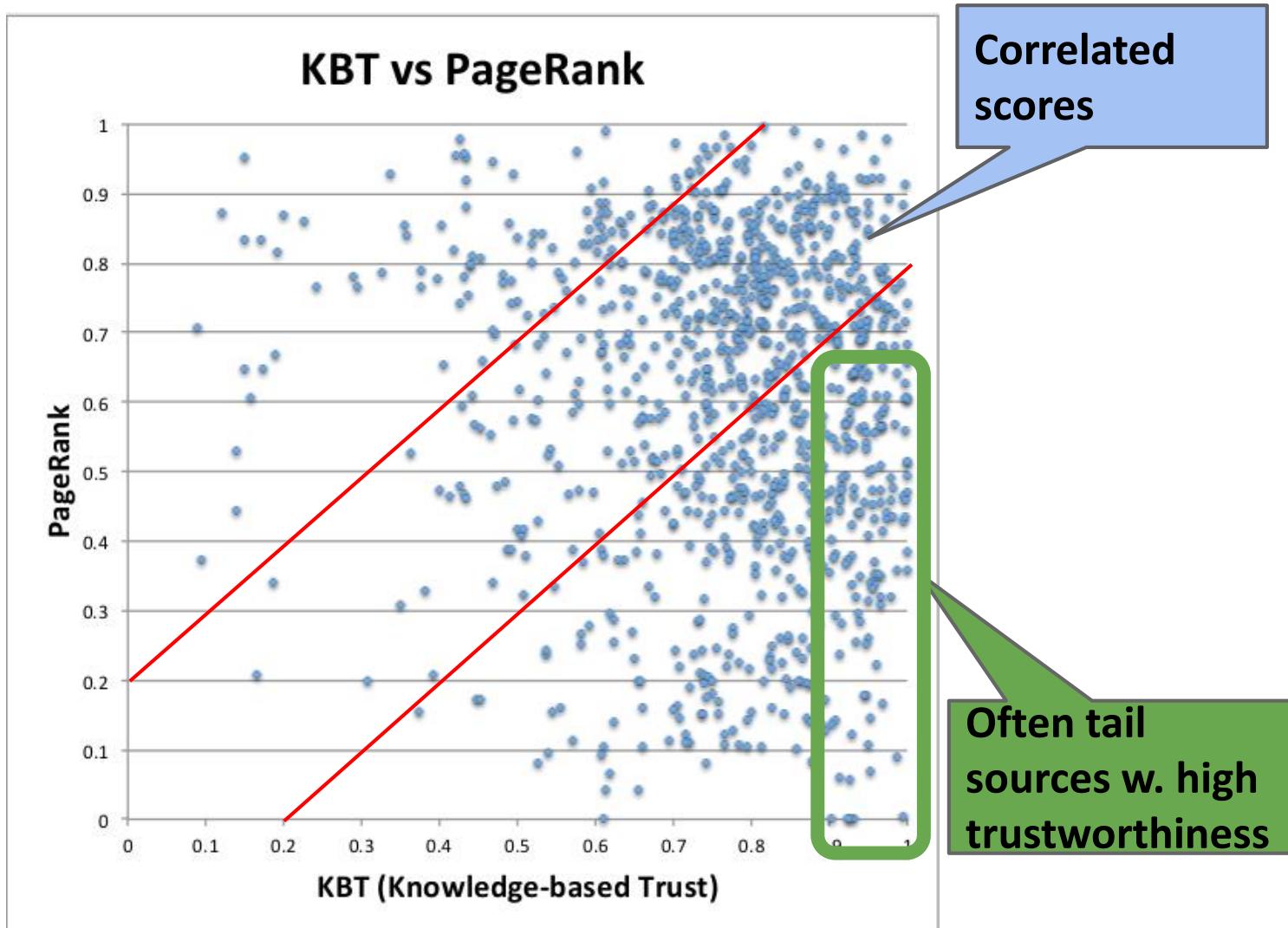
- ◆ What we have now for Web sources:
 - Page Rank: links between Websites/Webpages
 - Log based: search log and click-through rate
 - etc.

Knowledge-Based Trust (KBT) [DGM+15]

- ◆ Trustworthiness in $[0,1]$ for 5.6M websites and 119M webpages



Knowledge-Based Trust versus PageRank



Tail Sources with Low PageRank

The image displays four different website interfaces, each representing a tail source with low PageRank:

- salary.com**: A job search and salary comparison platform. It features sections for employees and employers, various job search filters, and a sidebar asking about workplace bullying.
- WOYLAAC**: A website dedicated to providing English subtitles for Korean dramas. It lists several drama titles with their episode counts and English subtitles available.
- Backingtrackguitar.com**: A resource for guitarists, offering backing tracks and promotional offers for Amazon diapers.
- UK Regional Portal**: A general information portal for the United Kingdom, covering topics like local branches, countries, regions, and specific news items like Kate Middleton's photos.

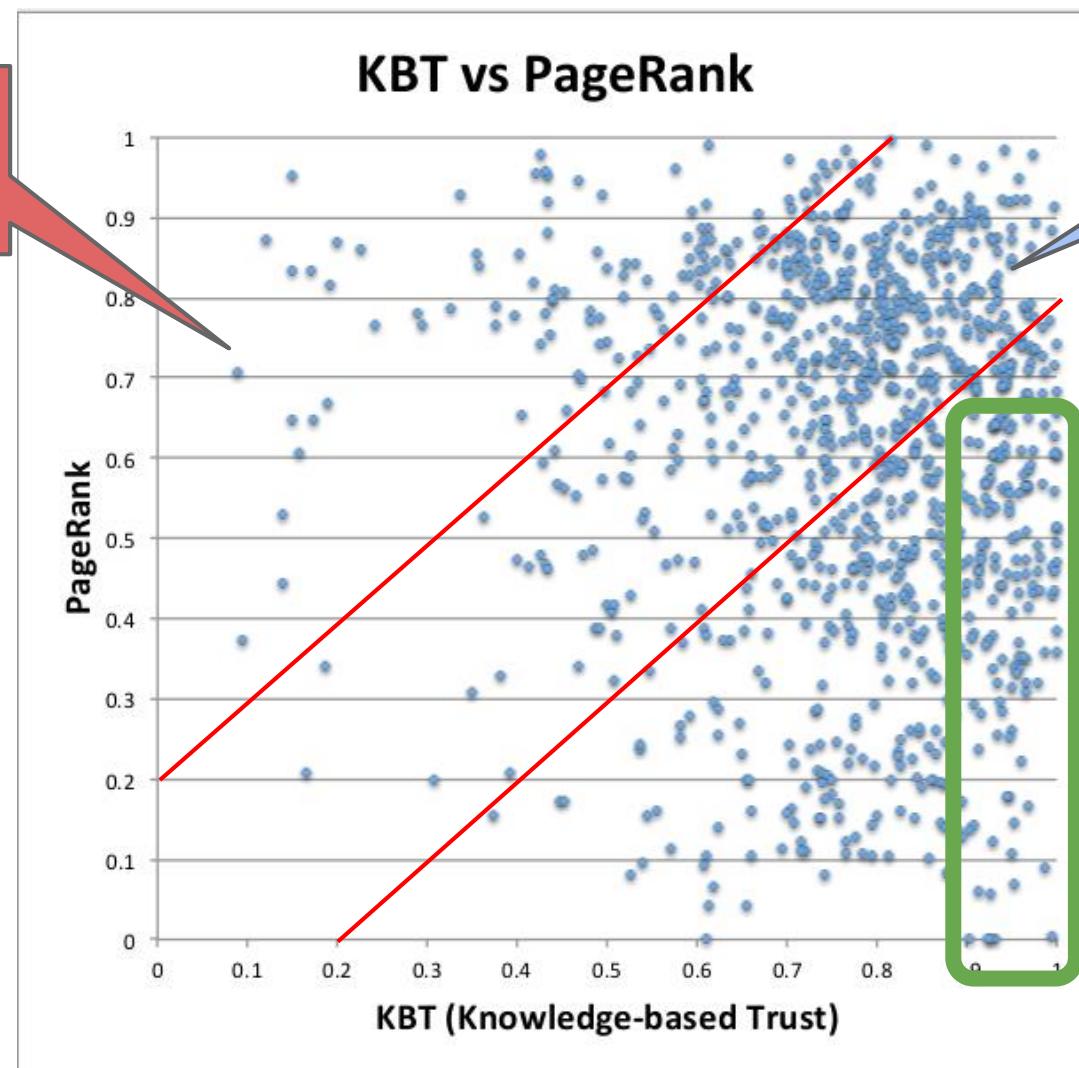
- ◆ Among 100 sampled websites, 85 are indeed trustworthy

Knowledge-Based Trust versus PageRank

Often sources
w. low accuracy

Correlated
scores

Often tail
sources w. high
trustworthiness



Popular Websites May Not Be Trustworthy

Gossip Websites

<http://www.ebizmba.com/articles/gossip-websites>

| Domain |
|---------------------|
| www.eonline.com |
| perezhilton.com |
| radaronline.com |
| www.zimbio.com |
| mediatakeout.com |
| gawker.com |
| www.popsugar.com |
| www.people.com |
| www.tmz.com |
| www.fishwrapper.com |
| celebrity.yahoo.com |
| wonderwall.msn.com |
| hollywoodlife.com |
| www.wetpaint.com |

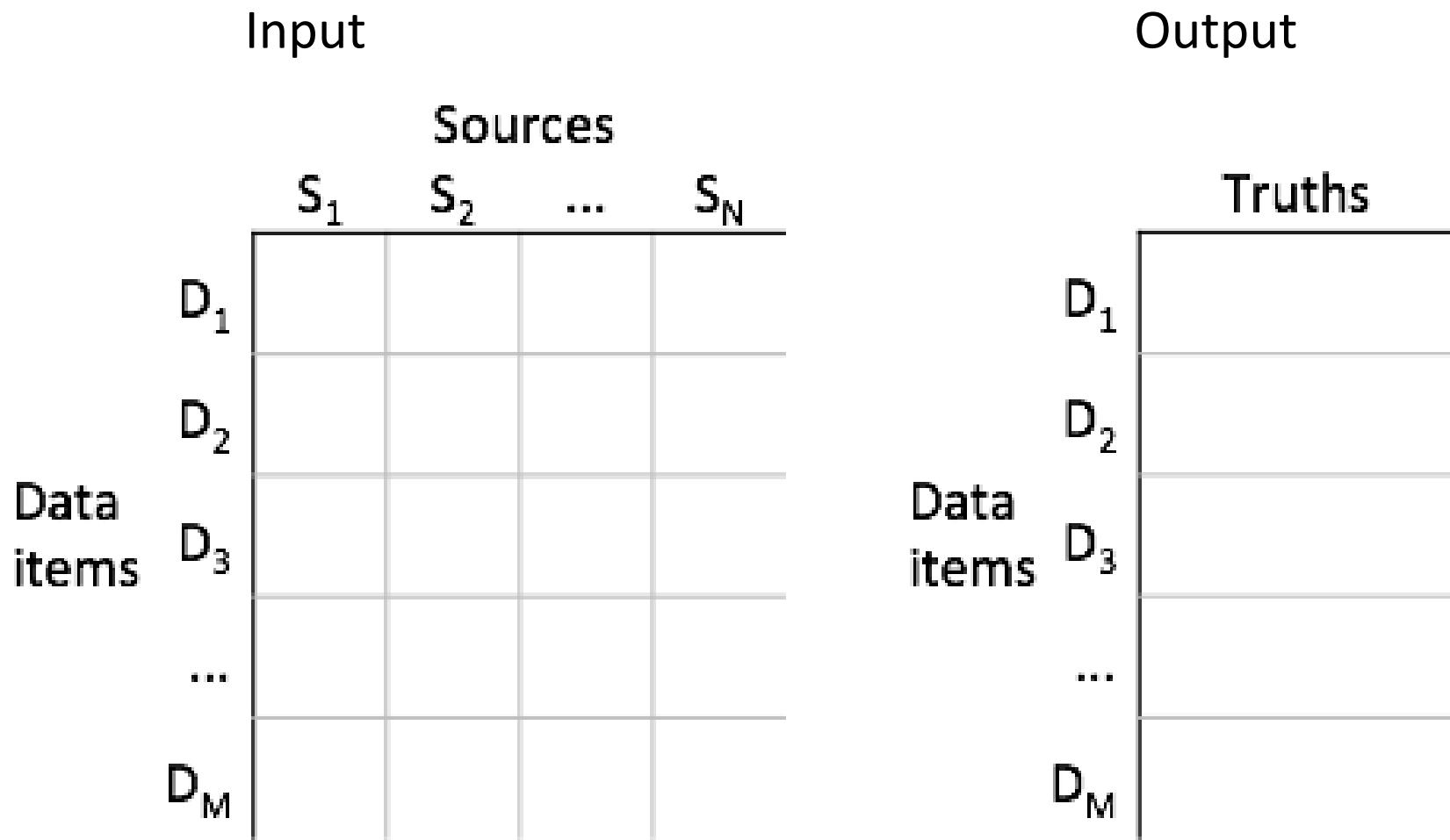
**14 out of 15 have a
PageRank among top
15% of the websites**

All have knowledge-based trust in bottom 50%

Goal: Judge Knowledge Triple Correctness

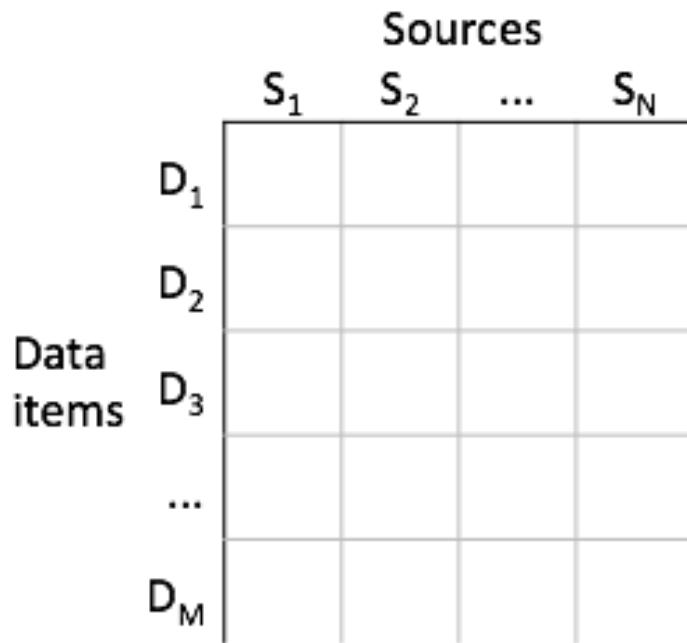
- ◆ Input: Knowledge triples and their provenance
 - Which extractor extracts from which source
- ◆ Output: a probability in [0,1] for each triple
 - Probabilistic decisions versus deterministic decisions

Data Fusion: Definition

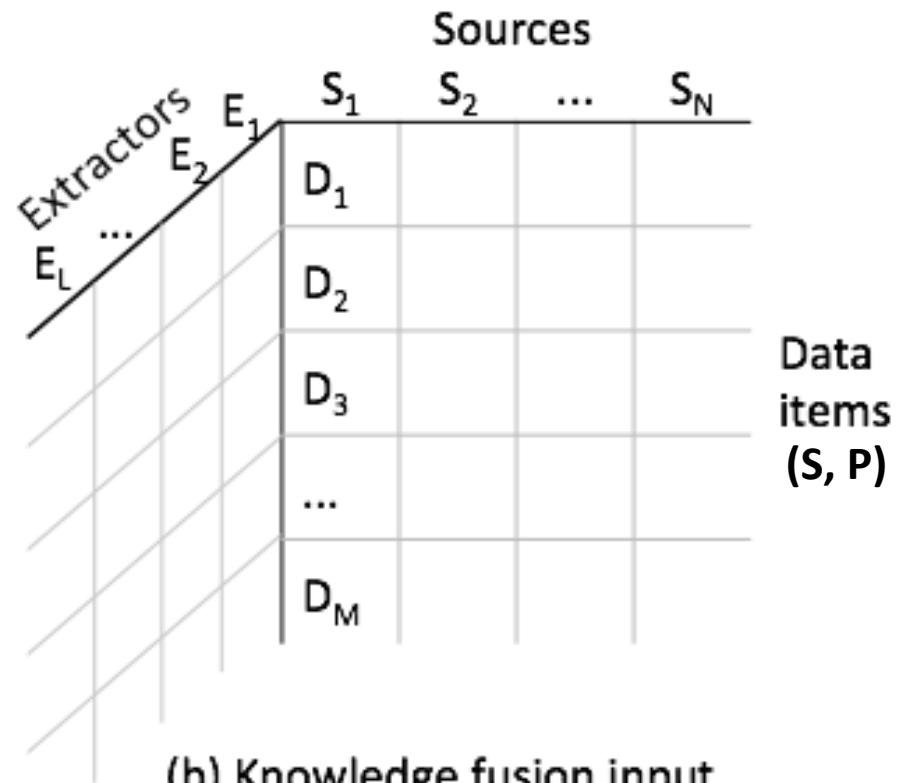


Knowledge Fusion Challenges

- ◆ Input is *three-dimensional*



(a) Data fusion input



(b) Knowledge fusion input

What About Extractions [PSD+14]

- ◆ Extracted Harry Potter actors/actresses

| Harry Potter | Ext1 | Ext2 | Ext3 |
|--------------|------|------|------|
| Daniel | ✓ | ✓ | ✓ |
| Emma | ✓ | | ✓ |
| Rupert | ✓ | ✓ | |
| Jonny | ✓ | | |
| Eric | | | ✓ |

What About Extractions

- ◆ Extracted Harry Potter actors/actresses

| Harry Potter | Ext1 | Ext2 | Ext3 |
|--------------|------|------|------|
| Daniel | ✓ | ✓ | ✓ |
| Emma | ✓ | | ✓ |
| Rupert | ✓ | ✓ | |
| Jonny | ✓ | | |
| Eric | | | ✓ |

- ◆ Voting: trust the majority

What About Extractions

- ◆ Extracted Harry Potter actors/actresses

| Harry Potter | Ext1 (high rec) | Ext2 (high prec) | Ext3 (med prec/rec) |
|--------------|--------------------|---------------------|------------------------|
| Daniel | ✓ | ✓ | ✓ |
| Emma | ✓ | | ✓ |
| Rupert | ✓ | ✓ | |
| Jonny | ✓ | | |
| Eric | | | ✓ |

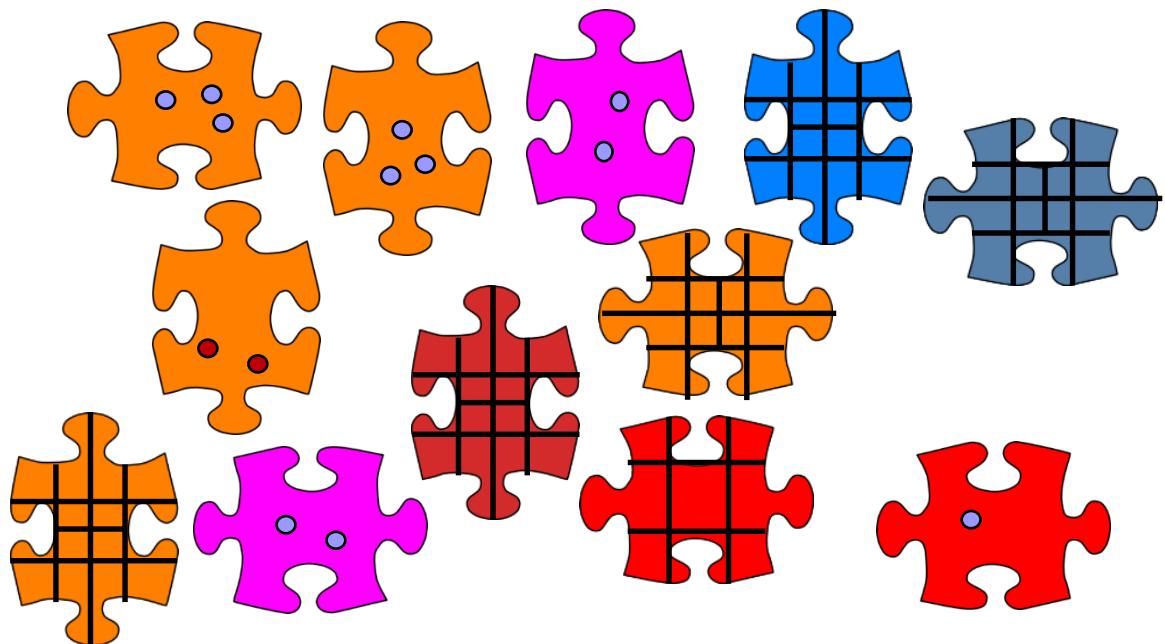
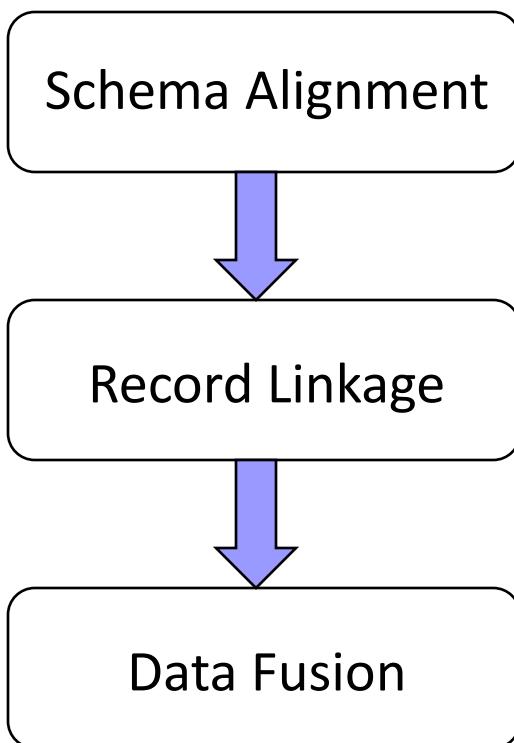
- ◆ Quality-based:
 - More likely to be correct if extracted by high-precision sources
 - More likely to be wrong if not extracted by high-recall sources

Outline

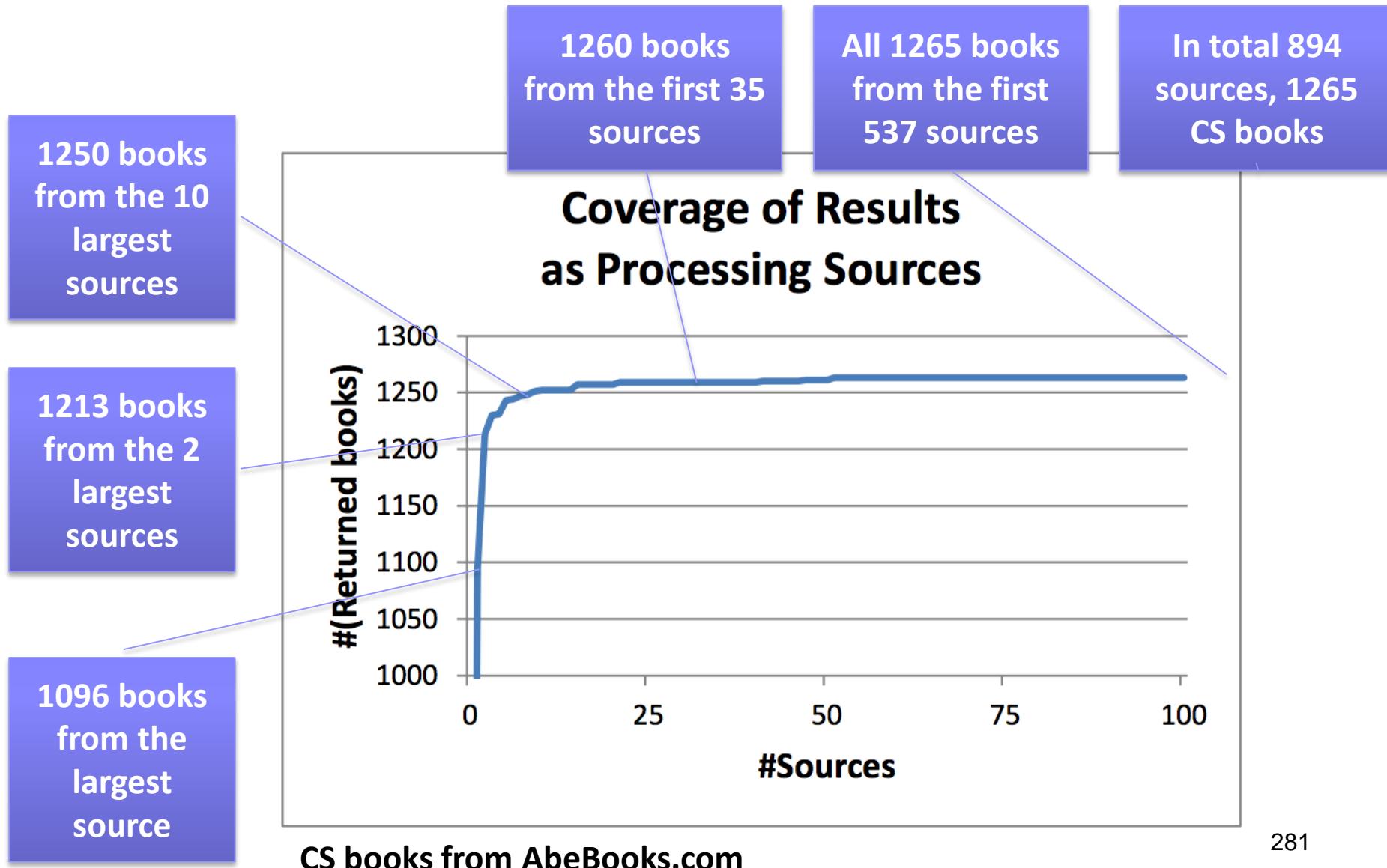
- ◆ Motivation
- ◆ Schema alignment
- ◆ Record linkage
- ◆ Data fusion
- ◆ Emerging topics
 - Knowledge-based trust, source selection, ...

BDI: Source Selection [DSS13]

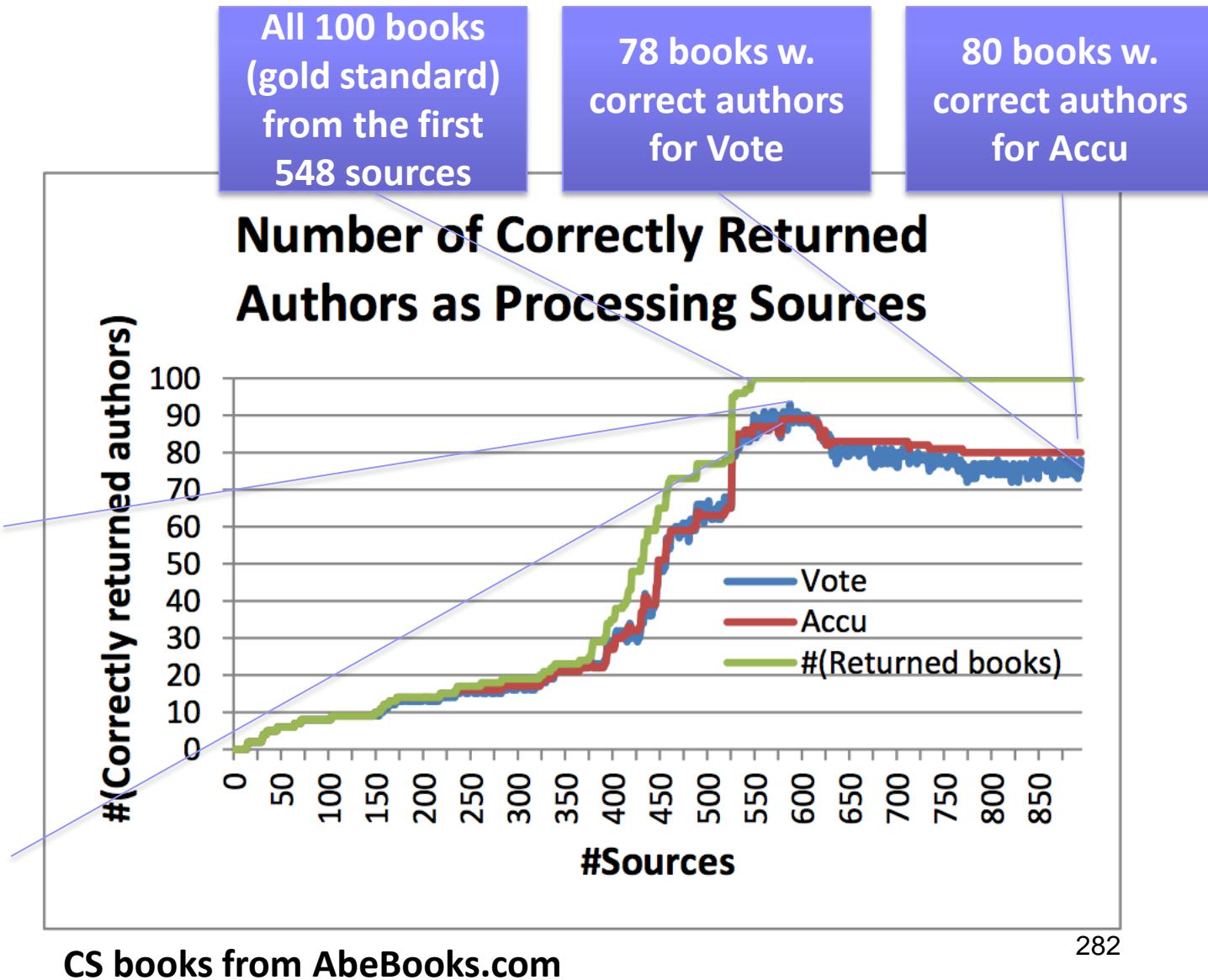
- ◆ Is it best to integrate **all** data?
 - Some data may be redundant or low-quality



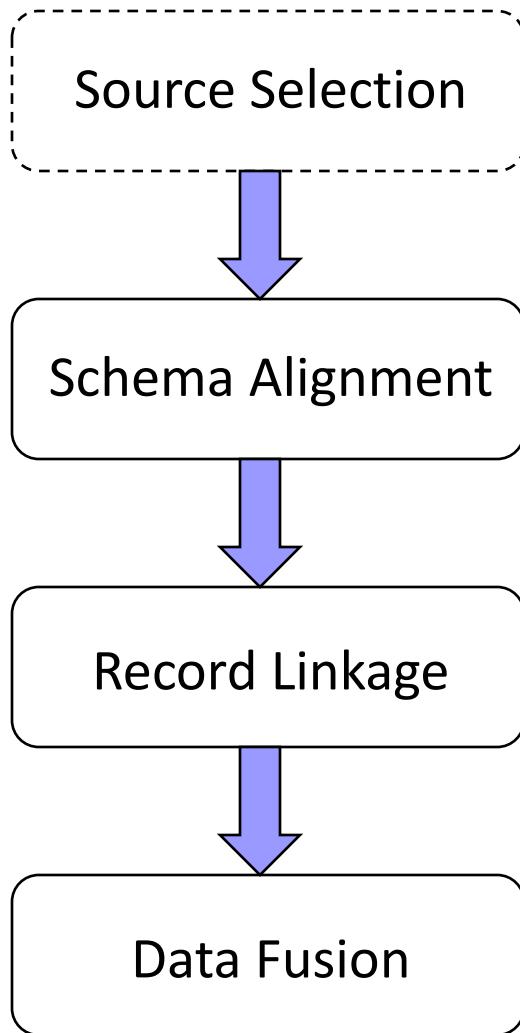
Redundant Data Do Not Bring Much Gain



Erroneous Data May Hurt Quality



BDI: Source Selection [DSS13]



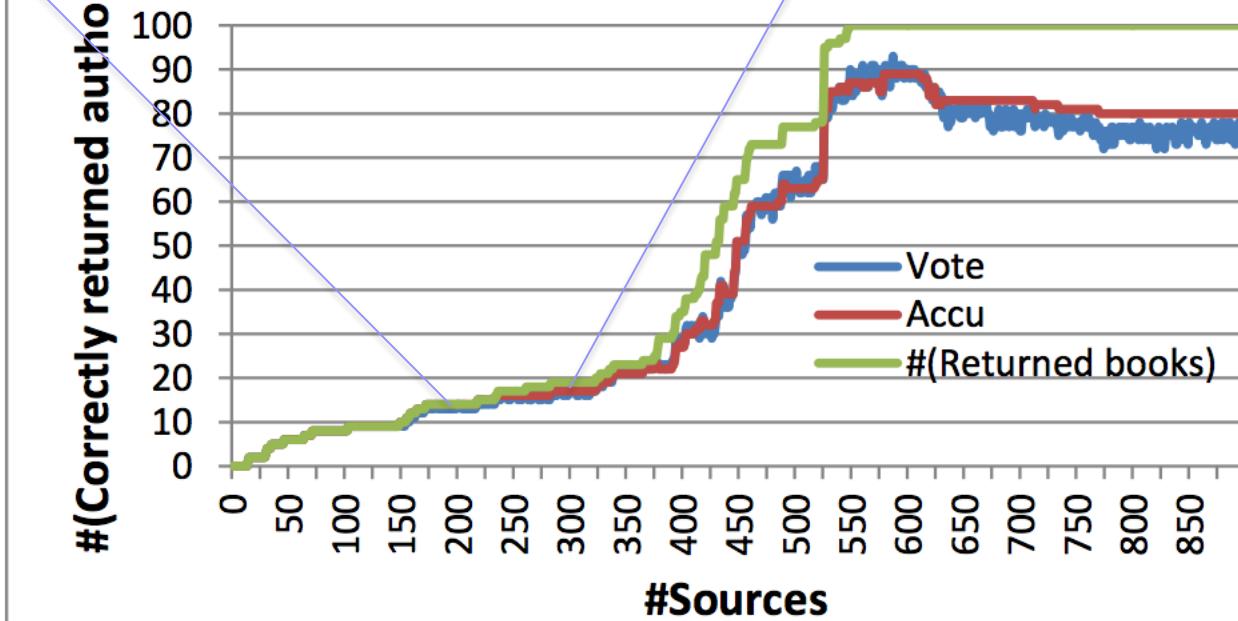
- ◆ How to wisely select sources before integration to balance gain and cost?

Maximize Quality Under Budget?

14 books (17.6% fewer) w. correct authors from the first 200 (33% less resources) sources

17 books w. correct authors from 300 sources (*budget*)

Number of Correctly Returned Authors as Processing Sources

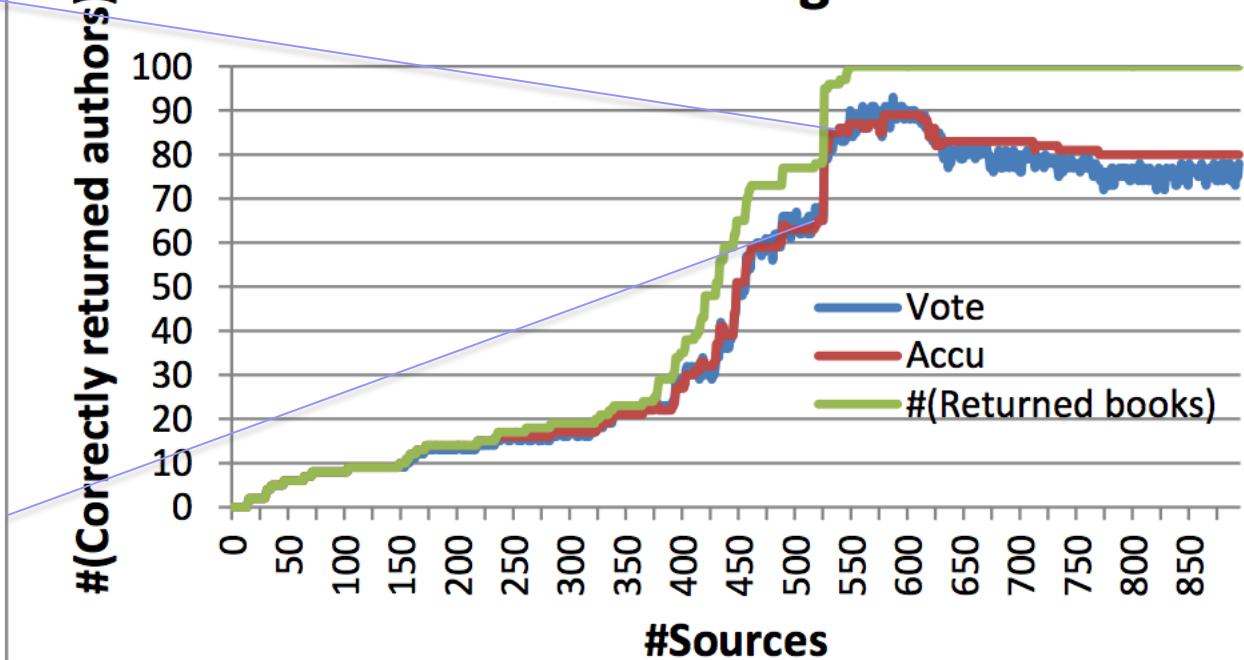


Minimize Cost with Certain Quality?

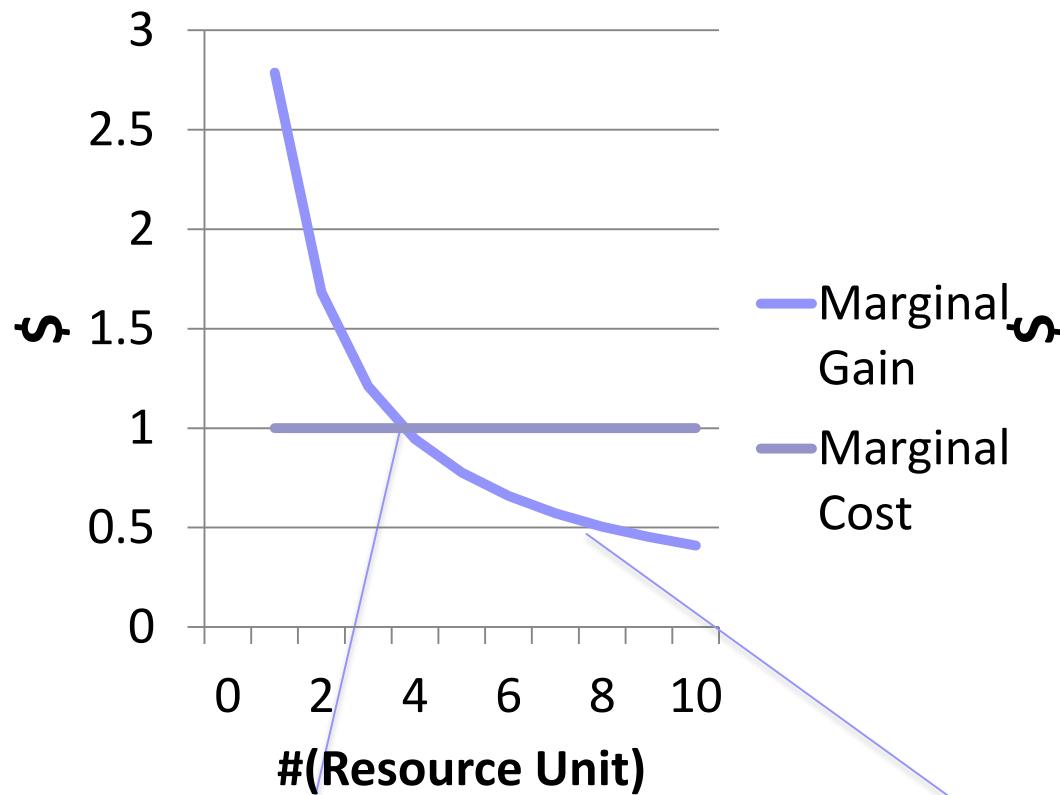
81 books (25% more) w. correct authors from 526 sources (1% more)

65 books w. correct authors (quality requirement) from the first 520 sources

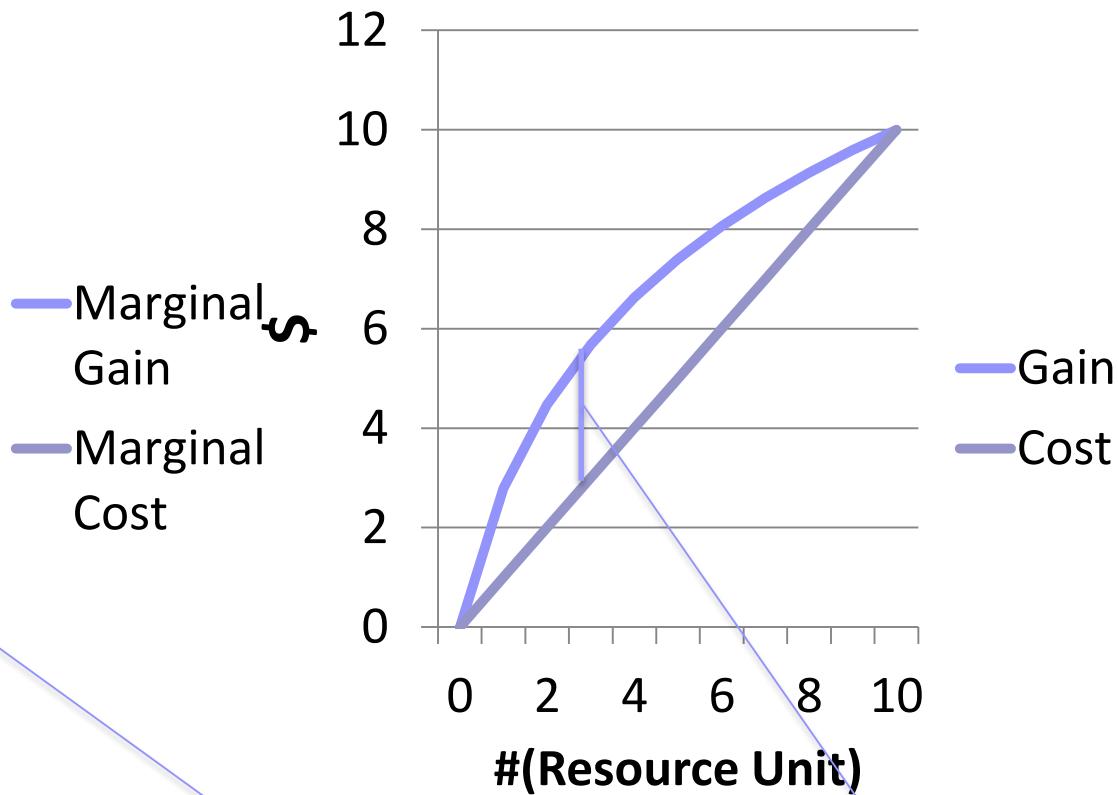
Number of Correctly Returned Authors as Processing Sources



Marginalism Principle in Economic Theory



Marginal gain
II
Marginal cost



The law of
Diminishing Returns

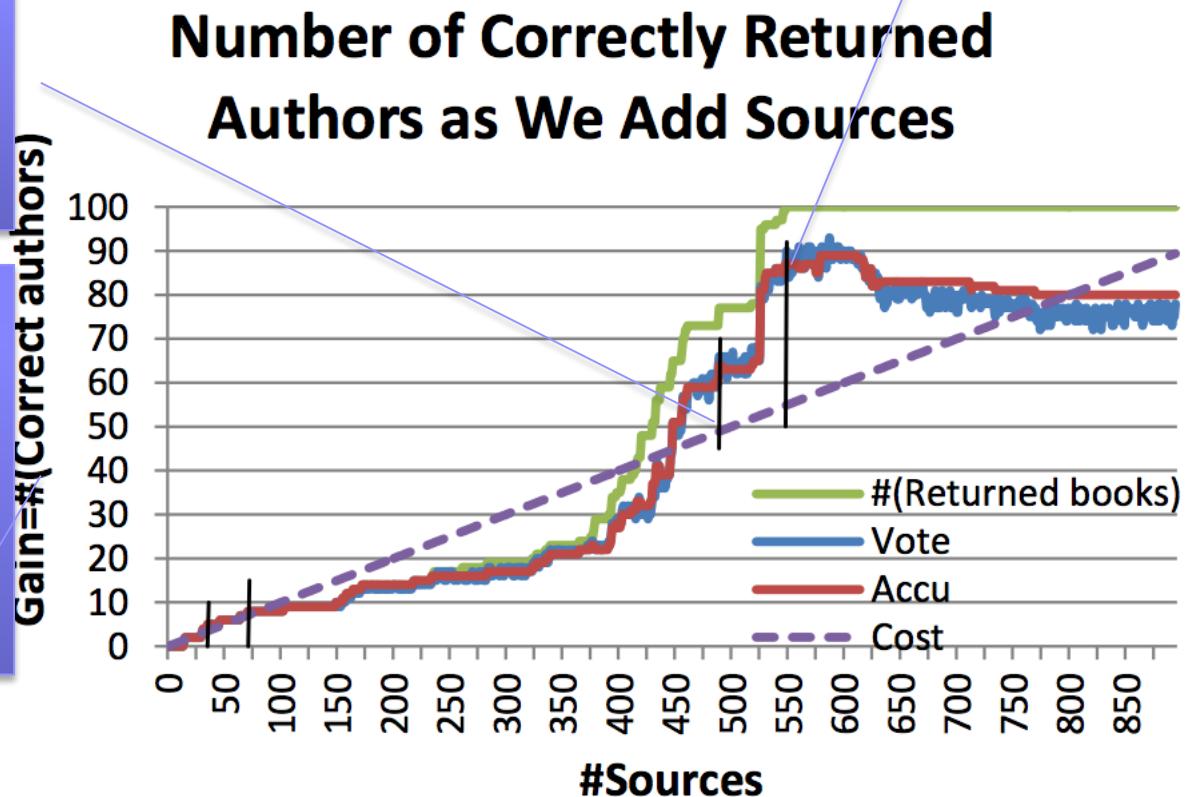
Largest profit

Marginalism for Source Selection

Challenge 1. The Law of Diminishing Returns does not necessarily hold, so multiple marginal points

Challenge 2. Each source is different in quality, so different ordering leads to different marginal points: best solution integrates 26 sources

Marginal point with the largest profit in this ordering: 548 sources



Outline

- ◆ Motivation
- ◆ Schema alignment
- ◆ Record linkage
- ◆ Data fusion
- ◆ Emerging topics
 - Knowledge-based trust, source selection, ...

Emerging Work

- ◆ Reconsider the architecture



Data warehousing

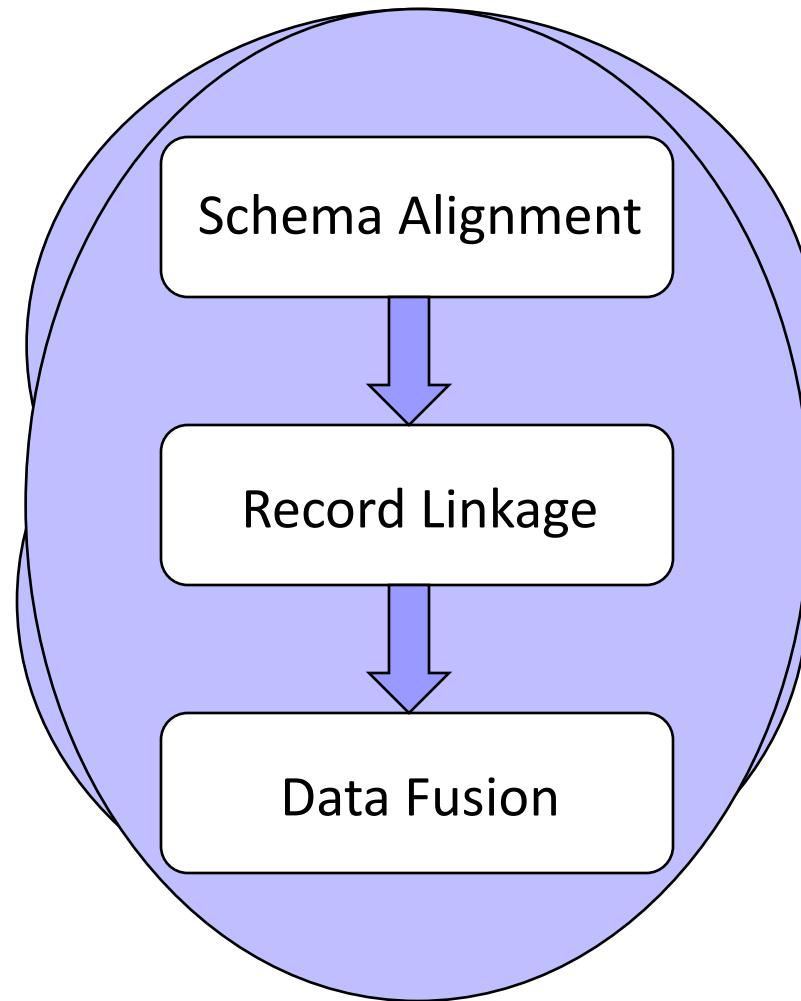


Virtual integration



Emerging Work

- ◆ Combining different components



Emerging Work

- ◆ Quality diagnosis



Emerging Work

◆ Source exploration tool

DATA AND TOOLS

Data.gov



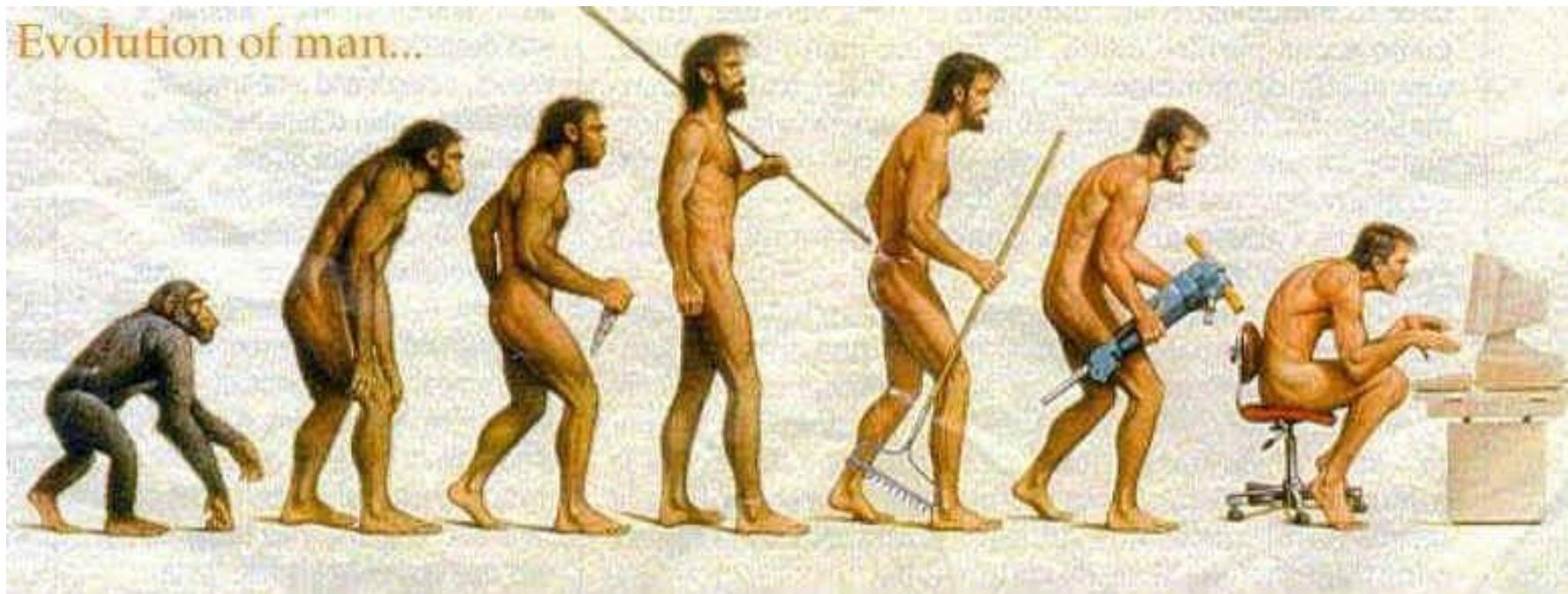
- 373,029 raw and geospatial datasets
- 1,209 data tools
- 308 apps
- 137 mobile apps
- 171 agencies and subagencies
- Suggest a dataset

Browse Raw Datasets RSS

| Name |
|--|
| 1. Worldwide M1+ Earthquakes, Past 7 Days Geography and Environment ANSS, geologist, plate, real time, environment Real-time, worldwide earthquake list for the past 7 days |
| 2. U.S. Overseas Loans and Grants (Greenbook) Foreign Commerce and Aid foreign assistance, economic assistance, These data are U.S economic and military assistance by country from 1946 to 2011. This is the authoritative data set |
| 3. Federal Data Center Consolidation Initiative (FDCCI) Data Center Closings 2010-2013 Federal Government Finan fddci, ... Updated February 8, 2013. Federal Data Center Consolidation Initiative (FDCCI) Data Center Closings 2010-2013. |
| 4. TSCA Inventory Geography and Environment new chemicals, manufactured chemicals, ... This dataset consists of the non confidential identities of chemical substances submitted under the Toxic Substances Control Act. |
| 5. Data.gov Catalog Other dataset, metadata, catalog, data extraction tool, ... An interactive dataset containing the metadata for the Data.gov raw datasets and tools catalogs. |
| 6. National Stock Number Extract Information and Communications Vendor, Product, NSN, National Stock Number, ... National Stock Number extract includes the current listing of National Stock Numbers (NSNs), NSN item name and description, vendor, product, and unit of measure. |
| 7. MyPyramid Food Raw Data Health and Nutrition Calories, Food, Nutrition, Fat, Nutrients, ... MyPyramid Food Data provides information on the total calories; calories from solid fats, added sugars, and alcohol |
| 8. Central Contractor Registration (CCR) FOIA Extract Information and Communications vendor, registration, contract This dataset lists all government contractors previously available under FOIA. |
| 9. FDIC Failed Bank List Banking, Finance, and Insurance closing, financial institutions, failed, failure, ... The FDIC is often appointed as receiver for failed banks. This list includes banks which have failed since October 1, 1980. |
| 10. Personnel Trends by Gender/Race Population American Indian, Black, Military, Hawaiian, ... Number of Service members by Gender, Race, Branch |
| 11. Local Area Unemployment Statistics Labor Force, Employment, and Earnings State and area labor force statistics, ... The Local Area Unemployment Statistics (LAUS) program produces monthly and annual employment, unemployment, and wage data for areas across the United States. |
| 12. FDCCI Map for CIO.gov Federal Government Finances and Employment The Federal CIO Council launched a government-wide Data Center Consolidation Task Force to consolidate and improve the way the federal government uses its data centers. |
| 13. Farmers Markets Geographic Data Agriculture Organic, Plants, Prepared Food, Nuts, ... longitude and latitude, state, address, name, and zip code of Farmers Markets in the United States |

Emerging Work

- ◆ Integrate data over time



Conclusions

- ◆ Big data integration is an important area of research
 - Knowledge bases, linked data, geo-spatial fusion, scientific data
- ◆ Much interesting work has been done in this area
 - Schema alignment, record linkage, data fusion
 - Challenges due to **volume, velocity, variety, veracity**
- ◆ A lot more research needs to be done!

Thank You!

References

- ◆ [B01] Michael K. Bergman: The Deep Web: Surfacing Hidden Value (2001)
- ◆ [BBR11] Zohra Bellahsene, Angela Bonifati, Erhard Rahm (Eds.): Schema Matching and Mapping. Springer 2011
- ◆ [CHW+08] Michael J. Cafarella, Alon Y. Halevy, Daisy Zhe Wang, Eugene Wu, Yang Zhang: WebTables: exploring the power of tables on the web. PVLDB 1(1): 538-549 (2008)
- ◆ [CHZ05] Kevin Chen-Chuan Chang, Bin He, Zhen Zhang: Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web. CIDR 2005: 44-55

References

- ◆ [DBS09a] Xin Luna Dong, Laure Berti-Equille, Divesh Srivastava: Integrating Conflicting Data: The Role of Source Dependence. PVLDB 2(1): 550-561 (2009)
- ◆ [DBS09b] Xin Luna Dong, Laure Berti-Equille, Divesh Srivastava: Truth Discovery and Copying Detection in a Dynamic World. PVLDB 2(1): 562-573 (2009)
- ◆ [DDH08] Anish Das Sarma, Xin Dong, Alon Y. Halevy: Bootstrapping pay-as-you-go data integration systems. SIGMOD Conference 2008: 861-874
- ◆ [DDH09] Anish Das Sarma, Xin Luna Dong, Alon Y. Halevy: Data Modeling in Dataspace Support Platforms. Conceptual Modeling: Foundations and Applications 2009: 122-138

References

- ◆ [DFG+12] Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Y. Halevy, Hongrae Lee, Fei Wu, Reynold Xin, Cong Yu: Finding related tables. SIGMOD Conference 2012: 817-828
- ◆ [DGM+15] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, Wei Zhang: Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. PVLDB 8(9): 938-949 (2015)
- ◆ [DHI12] AnHai Doan, Alon Y. Halevy, Zachary G. Ives: Principles of Data Integration. Morgan Kaufmann 2012
- ◆ [DHY07] Xin Luna Dong, Alon Y. Halevy, Cong Yu: Data Integration with Uncertainty. VLDB 2007: 687-698

References

- ◆ [DMP12] Nilesh N. Dalvi, Ashwin Machanavajjhala, Bo Pang: An Analysis of Structured Data on the Web. PVLDB 5(7): 680-691 (2012)
- ◆ [DNS+12] Uwe Draisbach, Felix Naumann, Sascha Szott, Oliver Wonneberg: Adaptive Windows for Duplicate Detection. ICDE 2012: 1073-1083
- ◆ [DSS13] Xin Luna Dong, Barna Saha, Divesh Srivastava: Less is More: Selecting Sources Wisely for Integration. PVLDB 6(2): 37-48 (2013)

References

- ◆ [EIV07] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios: Duplicate Record Detection: A Survey. IEEE Trans. Knowl. Data Eng. 19(1): 1-16 (2007)
- ◆ [EMH09] Hazem Elmeleegy, Jayant Madhavan, Alon Y. Halevy: Harvesting Relational Tables from Lists on the Web. PVLDB 2(1): 1078-1089 (2009)
- ◆ [FHM05] Michael J. Franklin, Alon Y. Halevy, David Maier: From databases to dataspaces: a new abstraction for information management. SIGMOD Record 34(4): 27-33 (2005)
- ◆ [FSS16] Donatella Firmani, Barna Saha, Divesh Srivastava: Online Entity Resolution Using an Oracle. PVLDB 9(5): 384-395 (2016)

References

- ◆ [GAM+10] Alban Galland, Serge Abiteboul, Amélie Marian, Pierre Senellart: Corroborating information from disagreeing views. WSDM 2010: 131-140
- ◆ [GDS+10] Songtao Guo, Xin Dong, Divesh Srivastava, Remi Zajac: Record Linkage with Uniqueness Constraints and Erroneous Values. PVLDB 3(1): 417-428 (2010)
- ◆ [GDS14] Anja Gruenheid, Xin Luna Dong, Divesh Srivastava: Incremental Record Linkage. PVLDB 7(9): 697-708 (2014)
- ◆ [GFS+18] Sainyam Galhotra, Donatella Firmani, Barna Saha, Divesh Srivastava: Robust entity resolution using random graphs. SIGMOD 2018
- ◆ [GM12] Lise Getoor, Ashwin Machanavajjhala: Entity Resolution: Theory, Practice & Open Challenges. PVLDB 5(12): 2018-2019 (2012)

References

- ◆ [GS09] Rahul Gupta, Sunita Sarawagi: Answering Table Augmentation Queries from Unstructured Lists on the Web. PVLDB 2(1): 289-300 (2009)
- ◆ [GSH11] Manish Gupta, Yizhou Sun, Jiawei Han: Trust analysis with clustering. WWW (Companion Volume) 2011: 53-54
- ◆ [HFM06] Alon Y. Halevy, Michael J. Franklin, David Maier: Principles of dataspace systems. PODS 2006: 1-9

References

- ◆ [JFH08] Shawn R. Jeffery, Michael J. Franklin, Alon Y. Halevy: Pay-as-you-go user feedback for dataspace systems. SIGMOD Conference 2008: 847-860
- ◆ [KGA+11] Anitha Kannan, Inmar E. Givoni, Rakesh Agrawal, Ariel Fuxman: Matching unstructured product offers to structured product specifications. KDD 2011: 404-412
- ◆ [KTR12] Lars Kolb, Andreas Thor, Erhard Rahm: Load Balancing for MapReduce-based Entity Resolution. ICDE 2012: 618-629
- ◆ [KTT+12] Hanna Köpcke, Andreas Thor, Stefan Thomas, Erhard Rahm: Tailoring entity resolution for matching product offers. EDBT 2012: 545-550

References

- ◆ [LDL+13] Xian Li, Xin Luna Dong, Kenneth B. Lyons, Weiyi Meng, Divesh Srivastava: Truth Finding on the deep web: Is the problem solved? PVLDB, 6(2) (2013)
- ◆ [LDM+11] Pei Li, Xin Luna Dong, Andrea Maurino, Divesh Srivastava: Linking Temporal Records. PVLDB 4(11): 956-967 (2011)
- ◆ [LDO+11] Xuan Liu, Xin Luna Dong, Beng Chin Ooi, Divesh Srivastava: Online Data Fusion. PVLDB 4(11): 932-943 (2011)
- ◆ [LSC10] Girija Limaye, Sunita Sarawagi, Soumen Chakrabarti: Annotating and Searching Web Tables Using Entities, Types and Relationships. PVLDB 3(1): 1338-1347 (2010)

References

- ◆ [MKB12] Bill McNeill, Hakan Kardes, Andrew Borthwick : Dynamic Record Blocking: Efficient Linking of Massive Databases in MapReduce. QDB 2012
- ◆ [MKK+08] Jayant Madhavan, David Ko, Lucja Kot, Vignesh Ganapathy, Alex Rasmussen, Alon Y. Halevy: Google's Deep Web crawl. PVLDB 1(2): 1241-1252 (2008)
- ◆ [MSS10] Claire Mathieu, Ocan Sankur, Warren Schudy: Online Correlation Clustering. STACS 2010: 573-584

References

- ◆ [PKP+13] George Papadakis, Georgia Koutrika, Themis Palpanas, Wolfgang Nejdl: Meta-blocking: taking entity resolution to the next level. TKDE (2013).
- ◆ [PIP+12] George Papadakis, Ekaterini Ioannou, Themis Palpanas, Claudia Niederee, Wolfgang Nejdl: A blocking framework for entity resolution in highly heterogeneous information spaces. TKDE (2012)
- ◆ [PR11] Jeff Pasternack, Dan Roth: Making Better Informed Trust Decisions with Generalized Fact-Finding. IJCAI 2011: 2324-2329
- ◆ [PR13] Jeff Pasternack, Dan Roth: Latent credibility analysis. WWW 2013: 1009-1020

References

- ◆ [PRM+12] Aditya Pal, Vibhor Rastogi, Ashwin Machanavajjhala, Philip Bohannon: Information integration over time in unreliable and uncertain environments. WWW 2012: 789-798
- ◆ [PS12] Rakesh Pimplikar, Sunita Sarawagi: Answering Table Queries on the Web using Column Keywords. PVLDB 5(10): 908-919 (2012)
- ◆ [PSD+14] Raveli Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, Divesh Srivastava: Fusing data with correlations. SIGMOD Conference 2014: 433-444

References

- ◆ [QAH+13] Guo-Jun Qi, Charu C. Aggarwal, Jiawei Han, Thomas S. Huang: Mining collective intelligence in diverse groups. WWW 2013: 1041-1052
- ◆ [QBC+18] Disheng Qiu, Luciano Barbosa, Valter Crescenzi, Paolo Merialdo, Divesh Srivastava: Big data linkage for product specification pages. SIGMOD 2018
- ◆ [TIP10] Partha Pratim Talukdar, Zachary G. Ives, Fernando Pereira: Automatically incorporating new sources in keyword search-based data integration. SIGMOD Conference 2010: 387-398
- ◆ [TJM+08] Partha Pratim Talukdar, Marie Jacob, Muhammad Salman Mehmood, Koby Crammer, Zachary G. Ives, Fernando Pereira, Sudipto Guha: Learning to create data-integrating queries. PVLDB 1(1): 785-796 (2008)

References

- ◆ [VBD14] Norases Vesdapunt, Kedar Bellare, Nilesh N. Dalvi: Crowdsourcing Algorithms for Entity Resolution. PVLDB 7(12): 1071-1082 (2014)
- ◆ [VCL10] Rares Vernica, Michael J. Carey, Chen Li: Efficient parallel set-similarity joins using MapReduce. SIGMOD Conference 2010: 495-506
- ◆ [VN12] Tobias Vogel, Felix Naumann: Automatic Blocking Key Selection for Duplicate Detection based on Unigram Combinations. QDB 2012

References

- ◆ [WGM10] Steven Whang, Hector Garcia-Molina: Entity Resolution with Evolving Rules. PVLDB 3(1): 1326-1337 (2010)
- ◆ [WGM13] Steven Whang, Hector Garcia-Molina: Incremental Entity Resolution on Rules and Data. VLDB J. (2013)
- ◆ [WLK+13] Jiannan Wang, Guoliang Li, Tim Kraska, Michael J. Franklin, Jianhua Feng: Leveraging transitive relations for crowdsourced joins. SIGMOD Conference 2013: 229-240
- ◆ [WYD+04] Wensheng Wu, Clement T. Yu, AnHai Doan, Weiyi Meng: An Interactive Clustering-based Approach to Integrating Source Query interfaces on the Deep Web. SIGMOD Conference 2004: 95-106

References

- ◆ [HY08] Xiaoxin Yin, Jiawei Han, Philip S. Yu: Truth Discovery with Multiple Conflicting Information Providers on the Web. IEEE Trans. Knowl. Data Eng. 20(6): 796-808 (2008)
- ◆ [YT11] Xiaoxin Yin, Wenzhao Tan: Semi-supervised truth discovery. WWW 2011: 217-226
- ◆ [ZH12] Bo Zhao, Jiawei Han: A probabilistic model for estimating real-valued truth from conflicting sources. QDB 2012
- ◆ [ZRG+12] Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, Jiawei Han: A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. PVLDB 5(6): 550-561 (2012)