Big Data Integration

Xin Luna Dong (Google Inc.)

Divesh Srivastava (AT&T Labs-Research)

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

Why Do We Need "Big Data Integration?"

- Building web-scale knowledge bases

Google knowledge graph

Freebase

Domain	ID	Topics	Facts
Music	/music	24M	161M
Media	/media_common	7M	23M
Books	/book	6M	37M
People	/people	3M	13M

ProBase

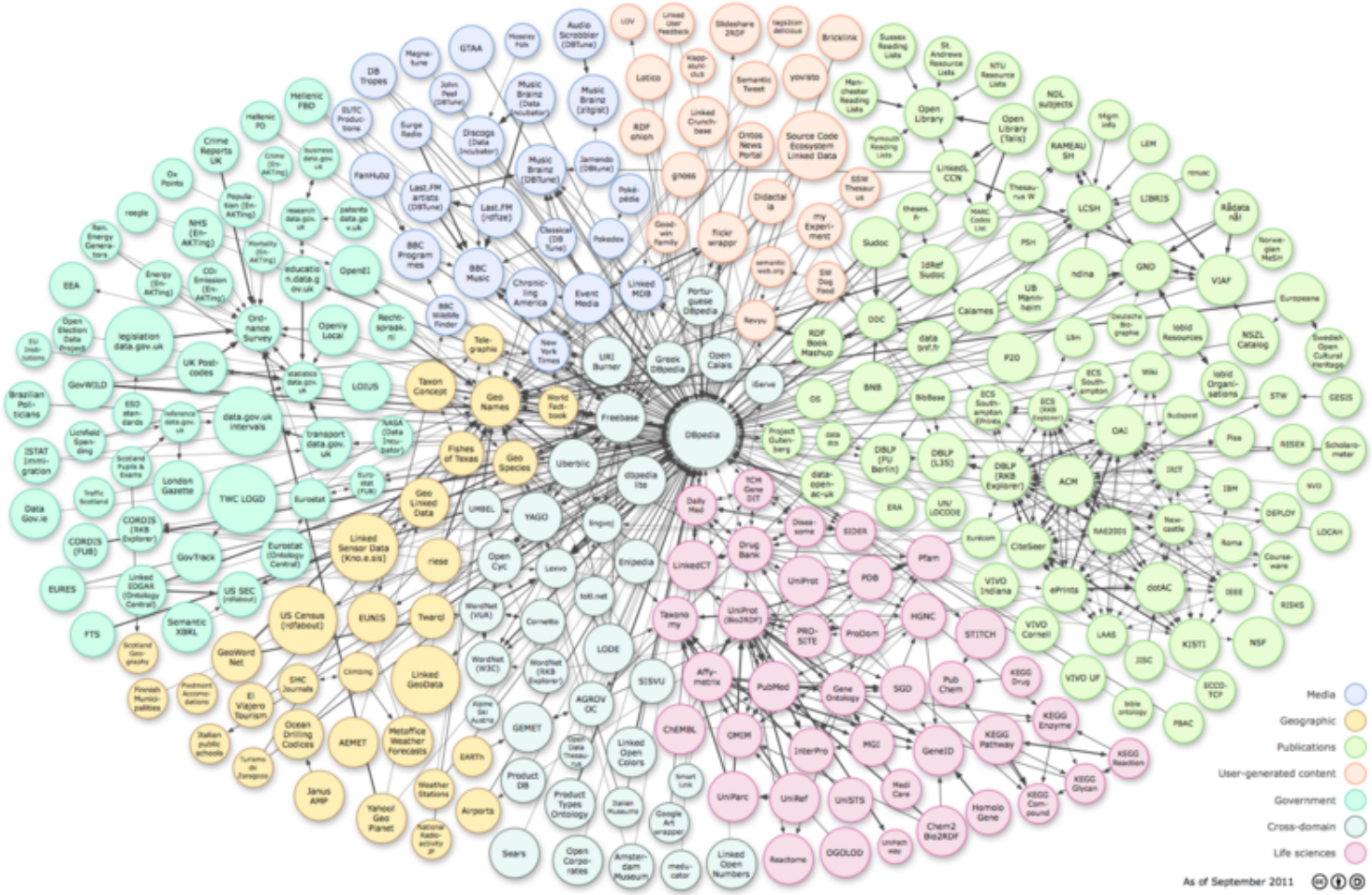
MSR knowledge base

A Little Knowledge Goes a Long Way.

NELL

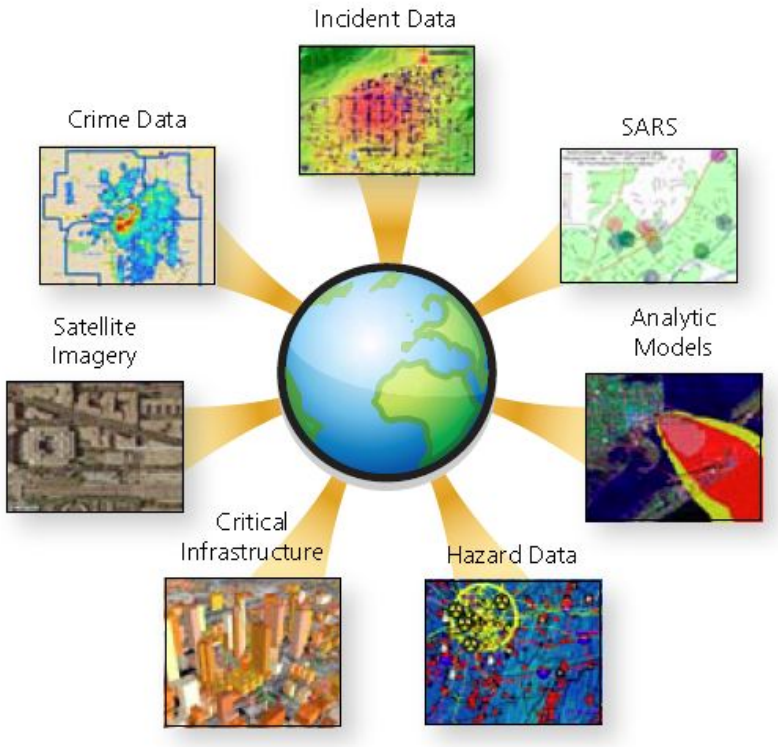
Why Do We Need "Big Data Integration?"

- Reasoning over linked data



Why Do We Need "Big Data Integration?"

- Geo-spatial data fusion

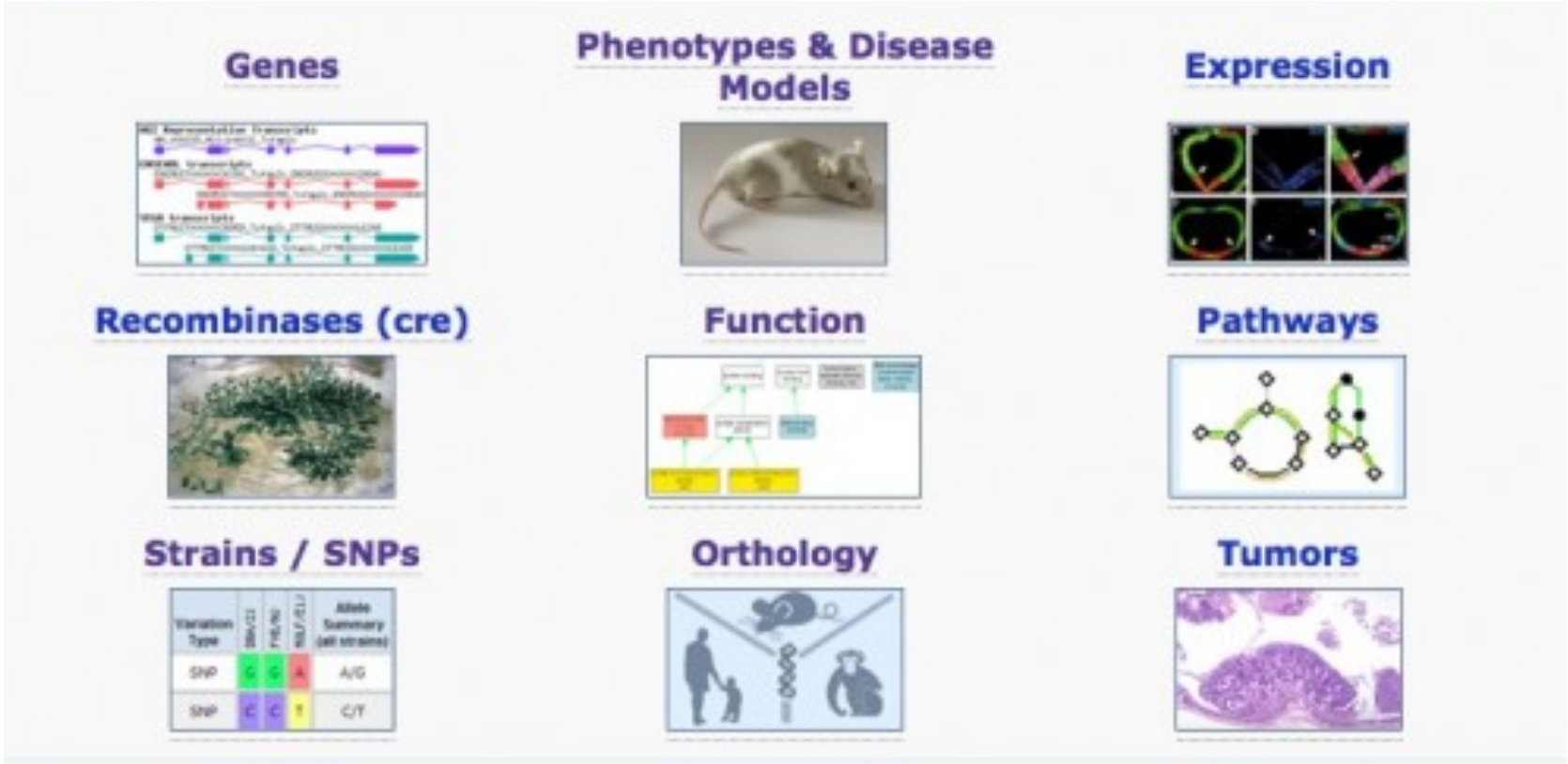


Geospatial Data Fusion

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

Why Do We Need "Big Data Integration?"

- Scientific data analysis



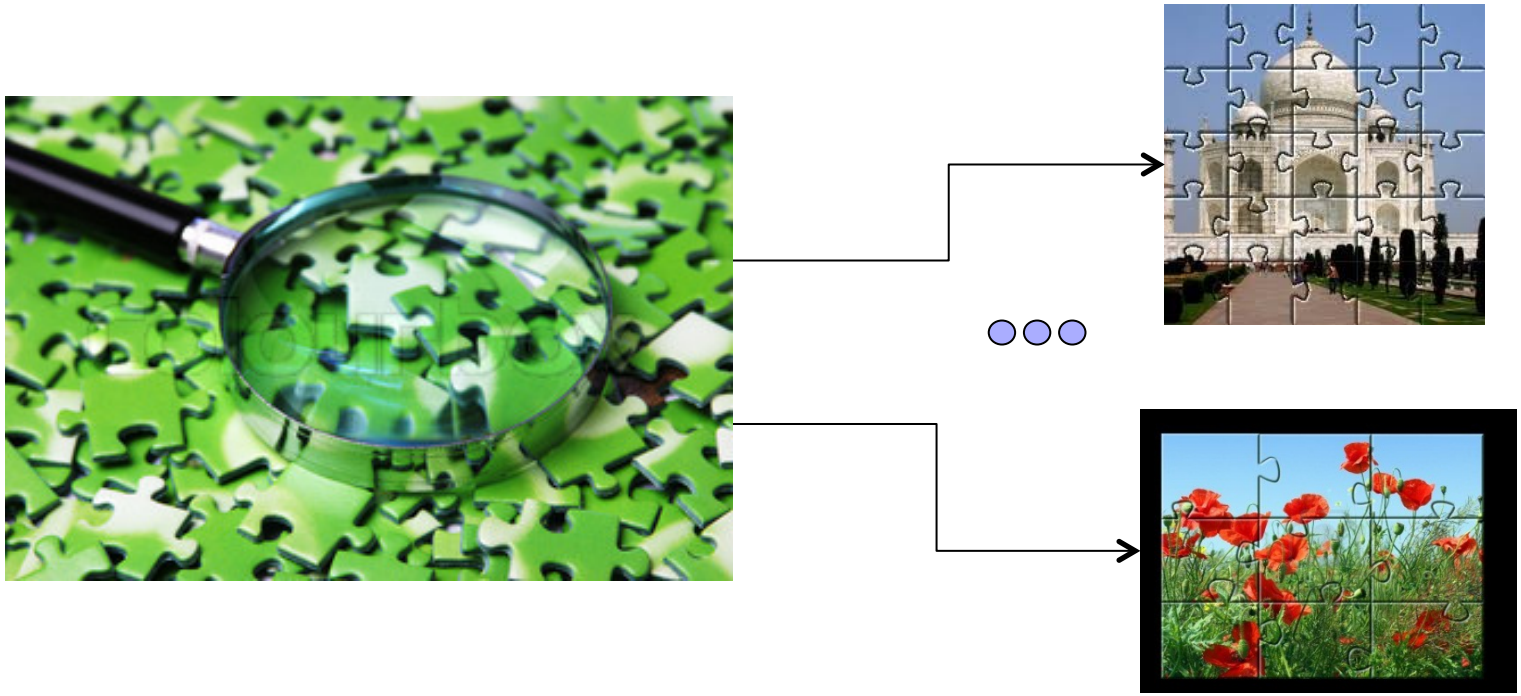
http://scienceline.org/2012/01/from-index-cards-to-information-overload/

Outline

- Motivation
  - Why do we need big data integration?
  - How has “small” data integration been done?
  - Challenges in big data integration
- Schema alignment
- Record linkage
- Data fusion
- Emerging topics

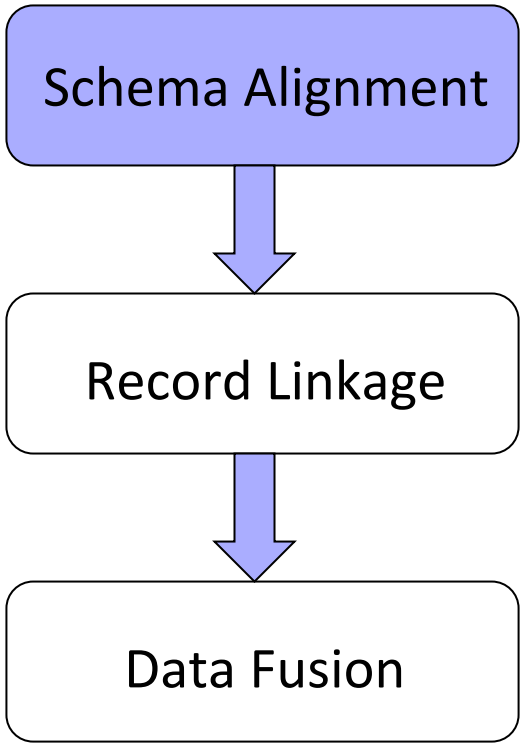
"Small" Data Integration: What Is It?

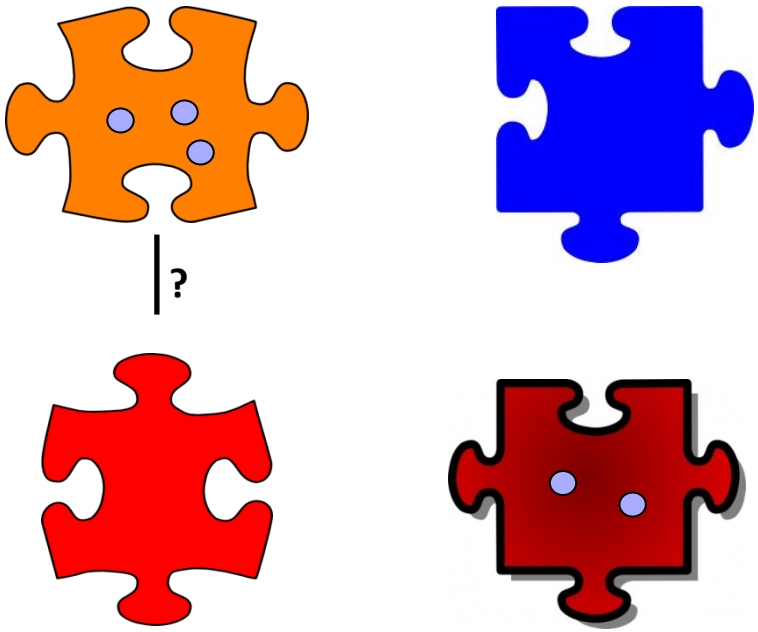
- Data integration = solving lots of jigsaw puzzles
  - Each jigsaw puzzle (e.g., Taj Mahal) is an **integrated entity**
  - Each piece of a puzzle comes from some **source**
  - Small data integration → solving small puzzles



"Small" Data Integration: How is it Done?

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

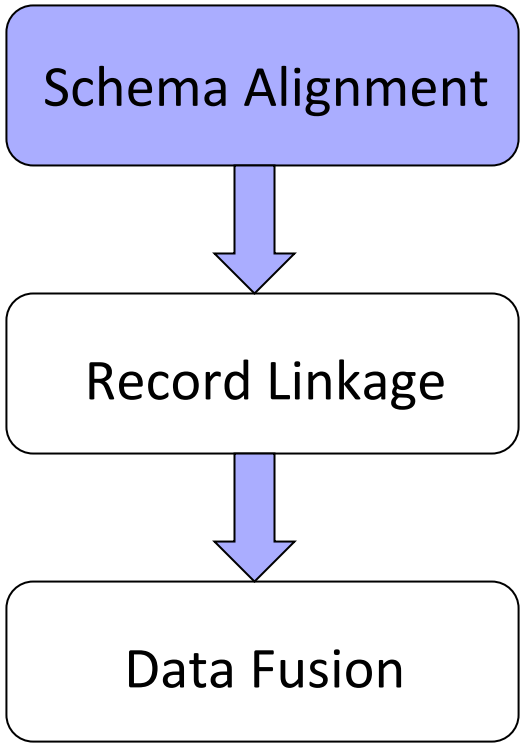


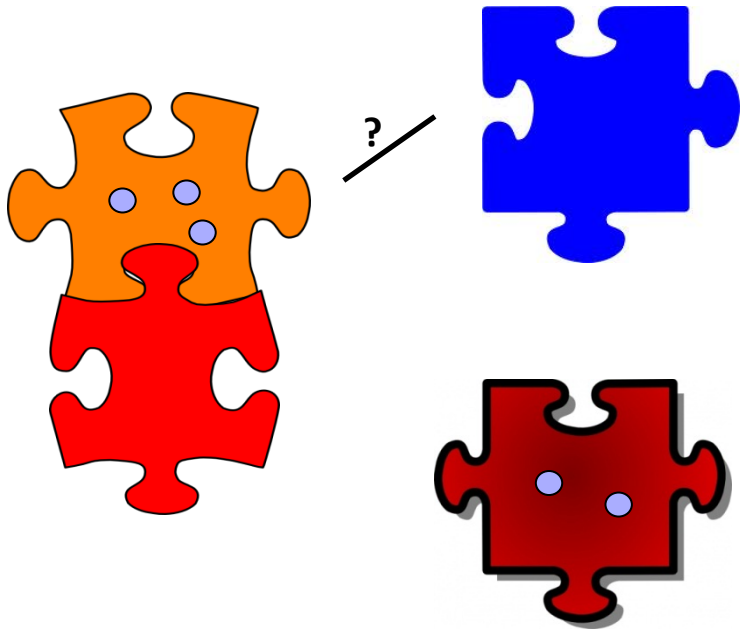


"Small" Data Integration: How is it Done?

X

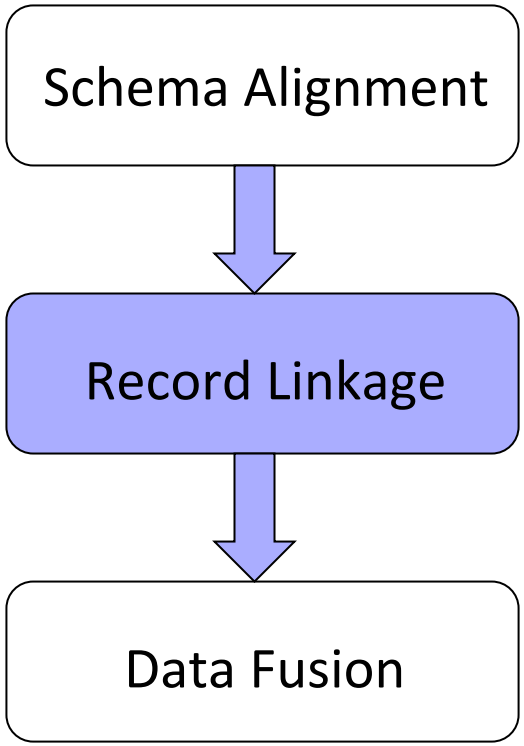
- "Small" data integration: alignment + linkage + fusion
  - Schema alignment: mapping of **structure** (e.g., shape)

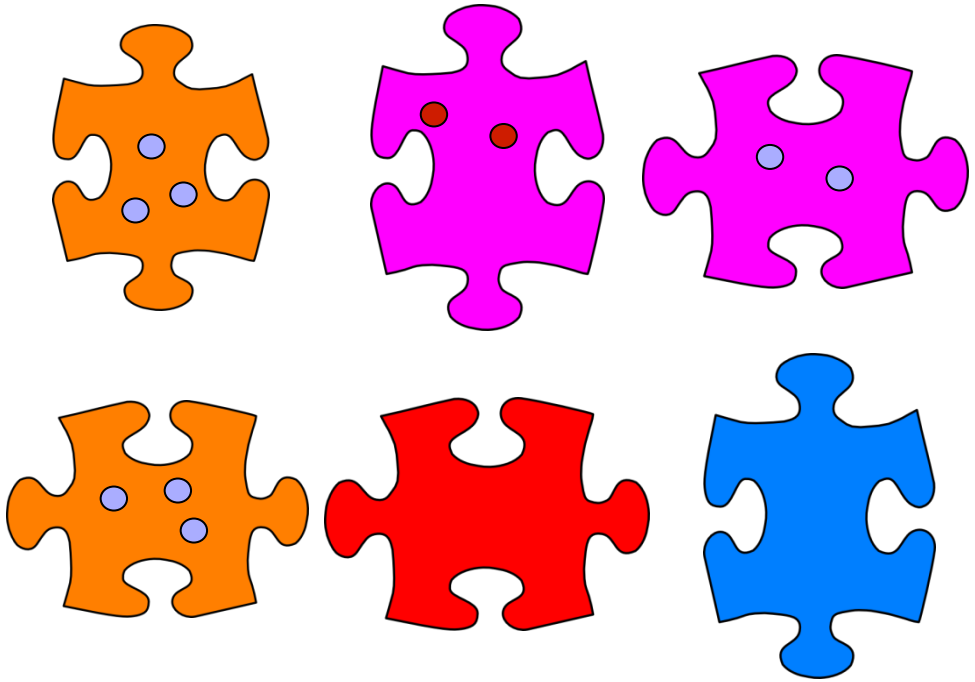




"Small" Data Integration: How is it Done?

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

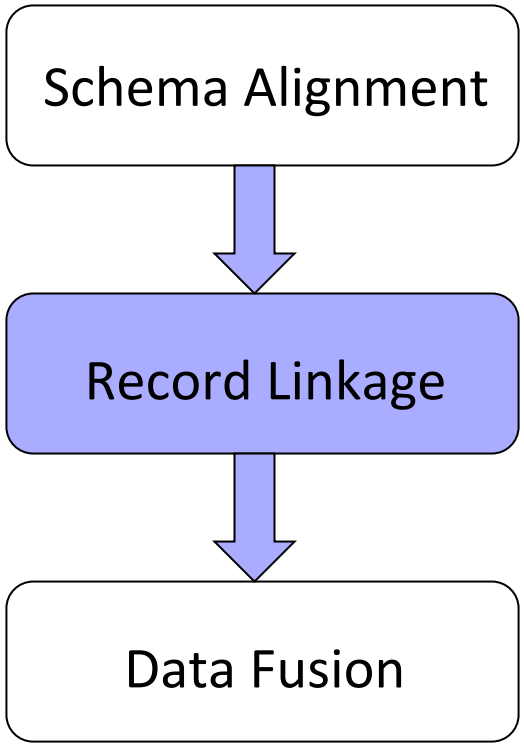


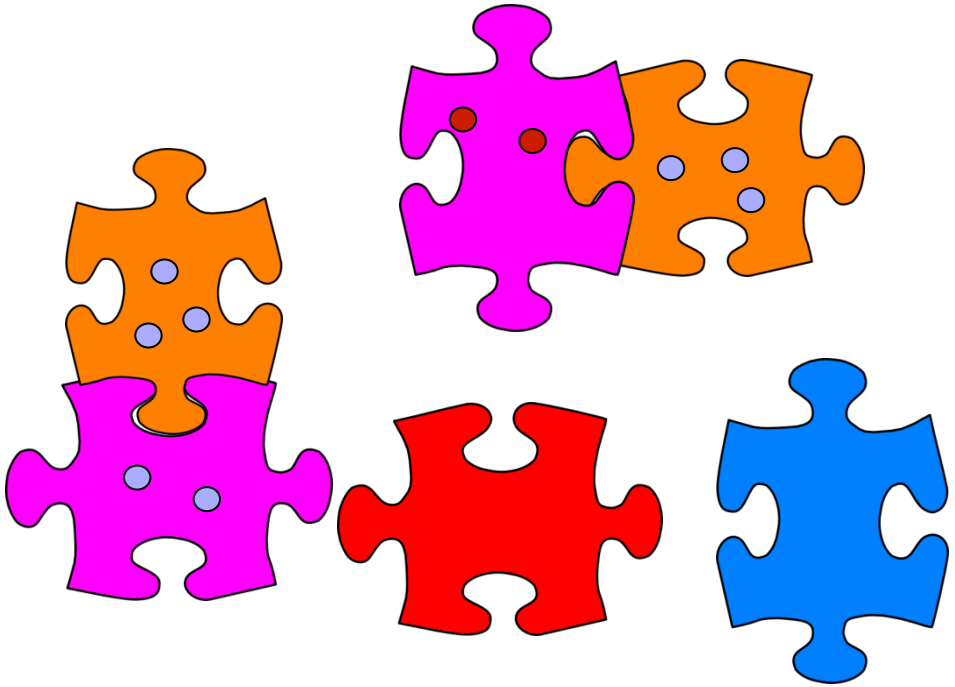


"Small" Data Integration: How is it Done?

X

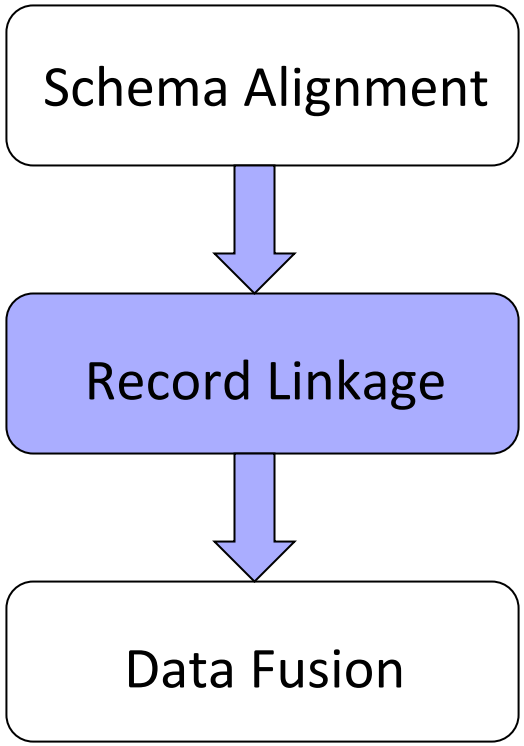
- "Small" data integration: alignment + linkage + fusion
  - Record linkage: matching based on **content** (e.g., color, pattern)

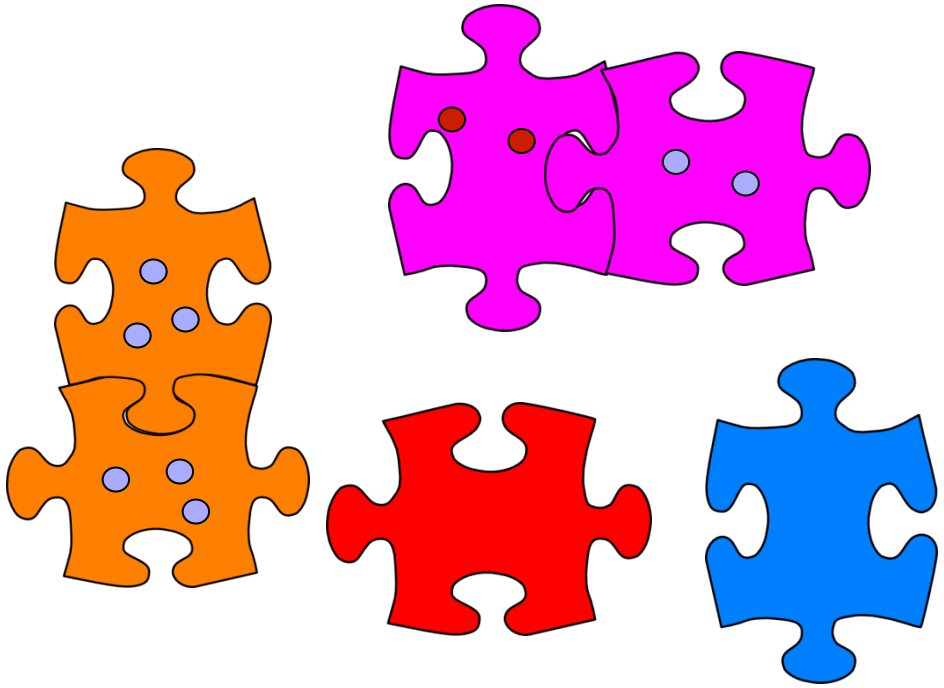




"Small" Data Integration: How is it Done?

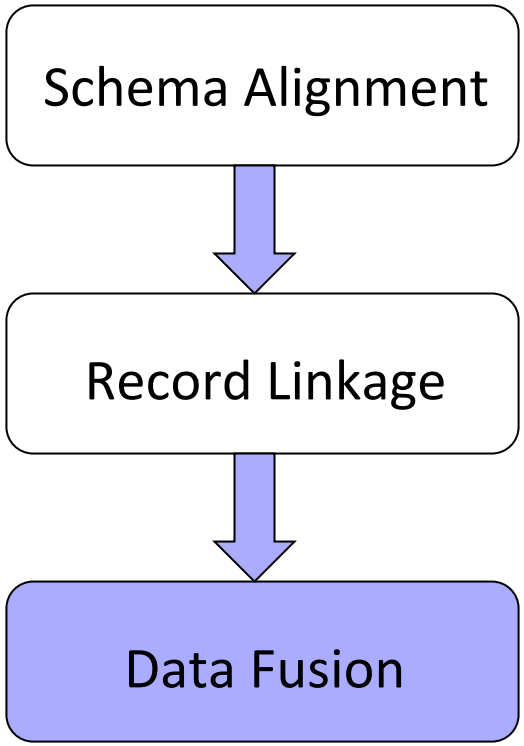
- "Small" data integration: alignment + linkage + fusion
  - Record linkage: matching based on **content** (e.g., color, pattern)

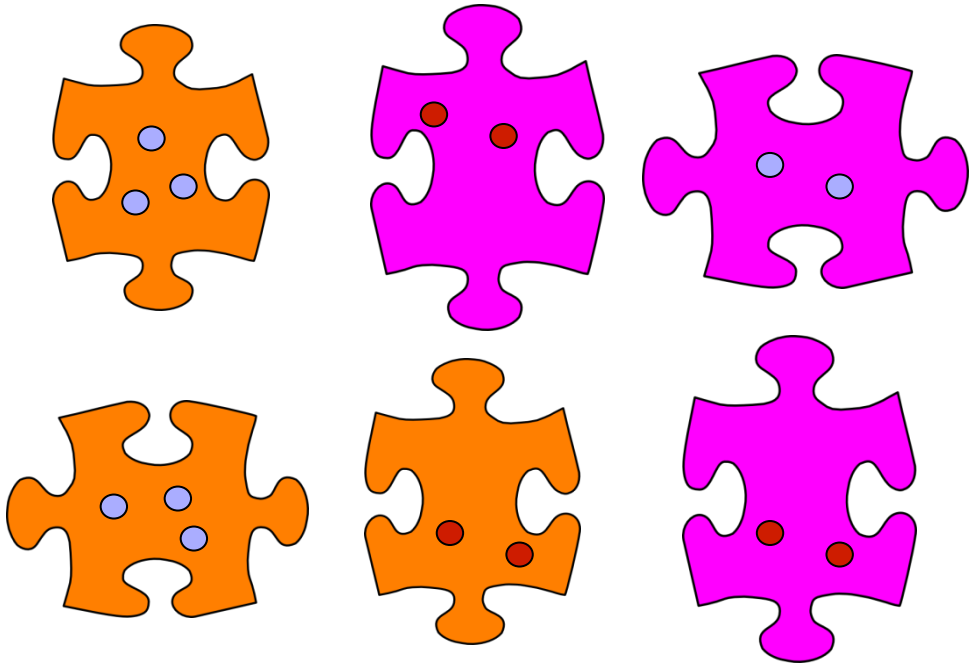




"Small" Data Integration: How is it Done?

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

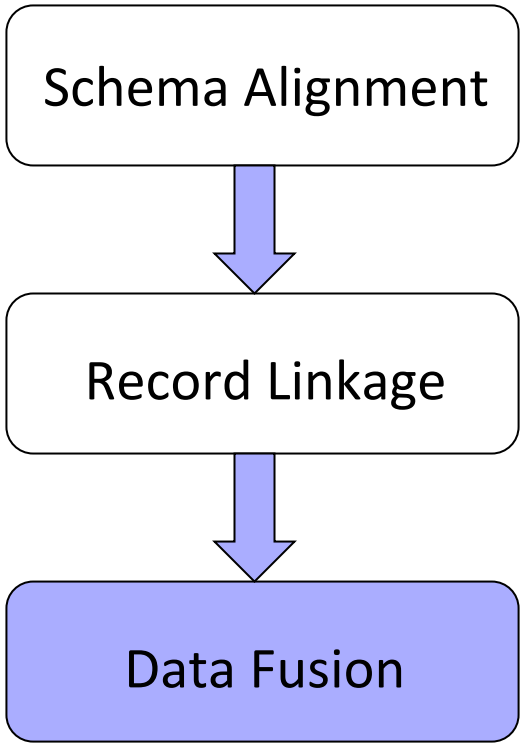


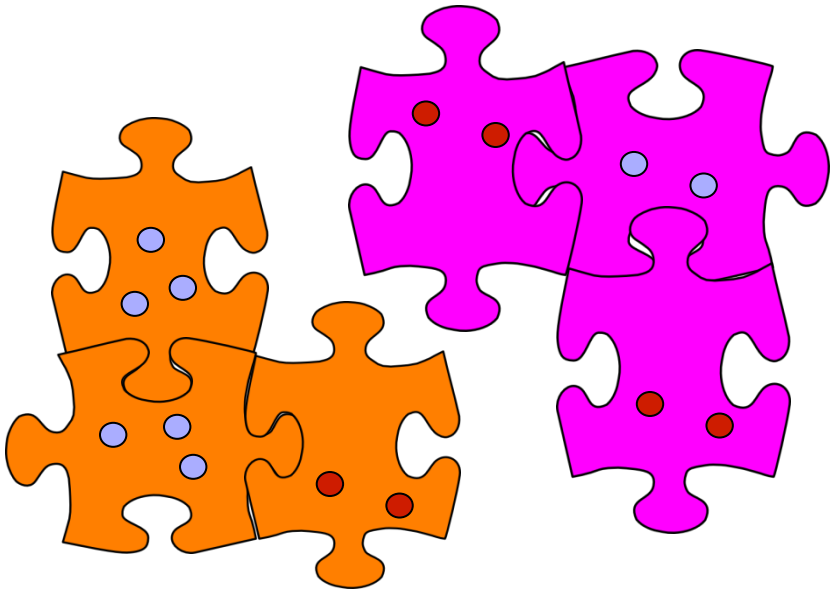


"Small" Data Integration: How is it Done?

X

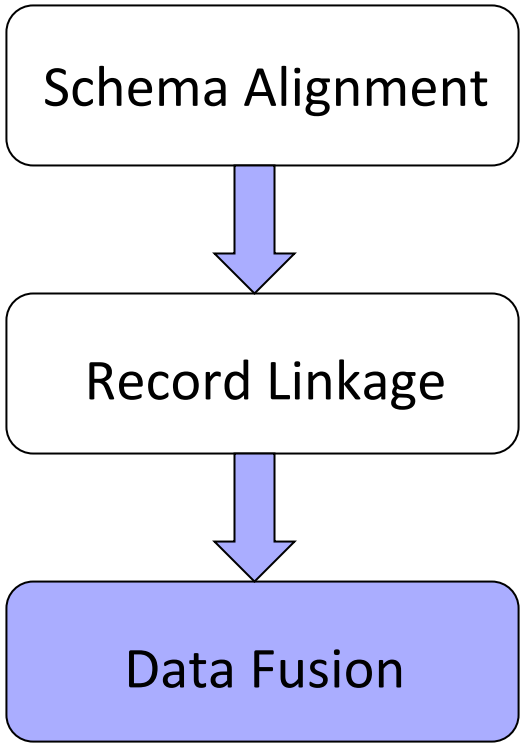
- "Small" data integration: alignment + linkage + fusion
  - Data fusion: reconciliation of **mismatching content** (e.g., pattern)

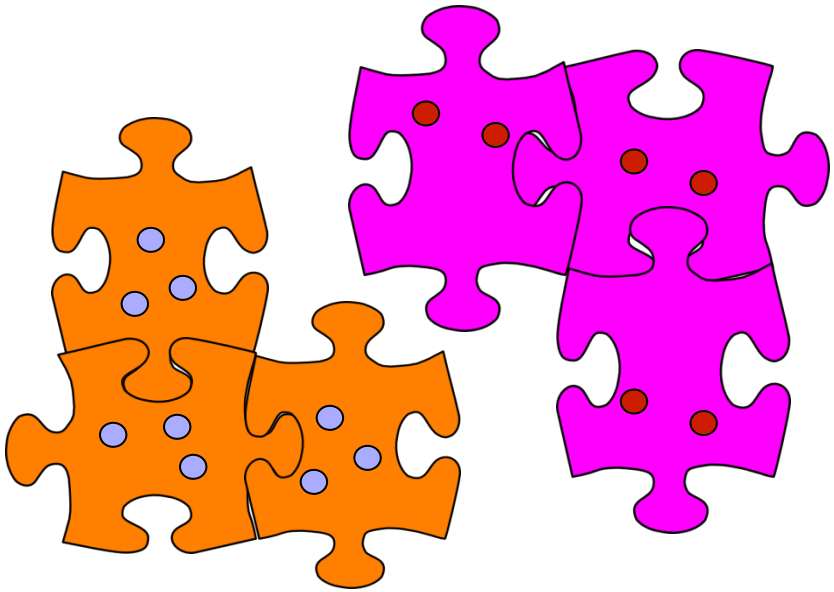




"Small" Data Integration: How is it Done?

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





Outline

- Motivation
  - Why do we need big data integration?
  - How has “small” data integration been done?
  - Challenges in big data integration
- Schema alignment
- Record linkage
- Data fusion
- Emerging topics

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?

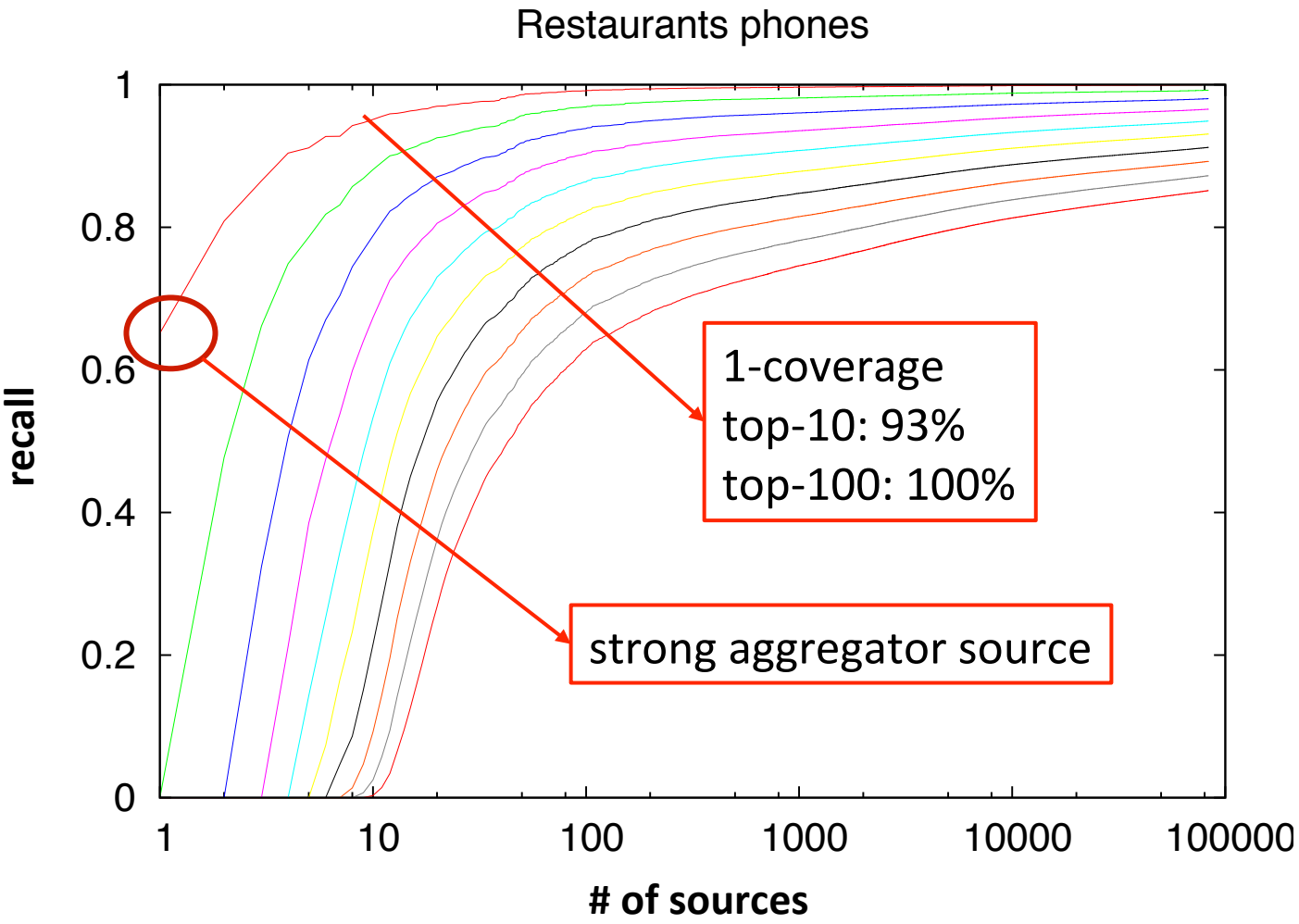
Domain Specific Data: Spread

- How many sources needed to build a complete DB for a domain?

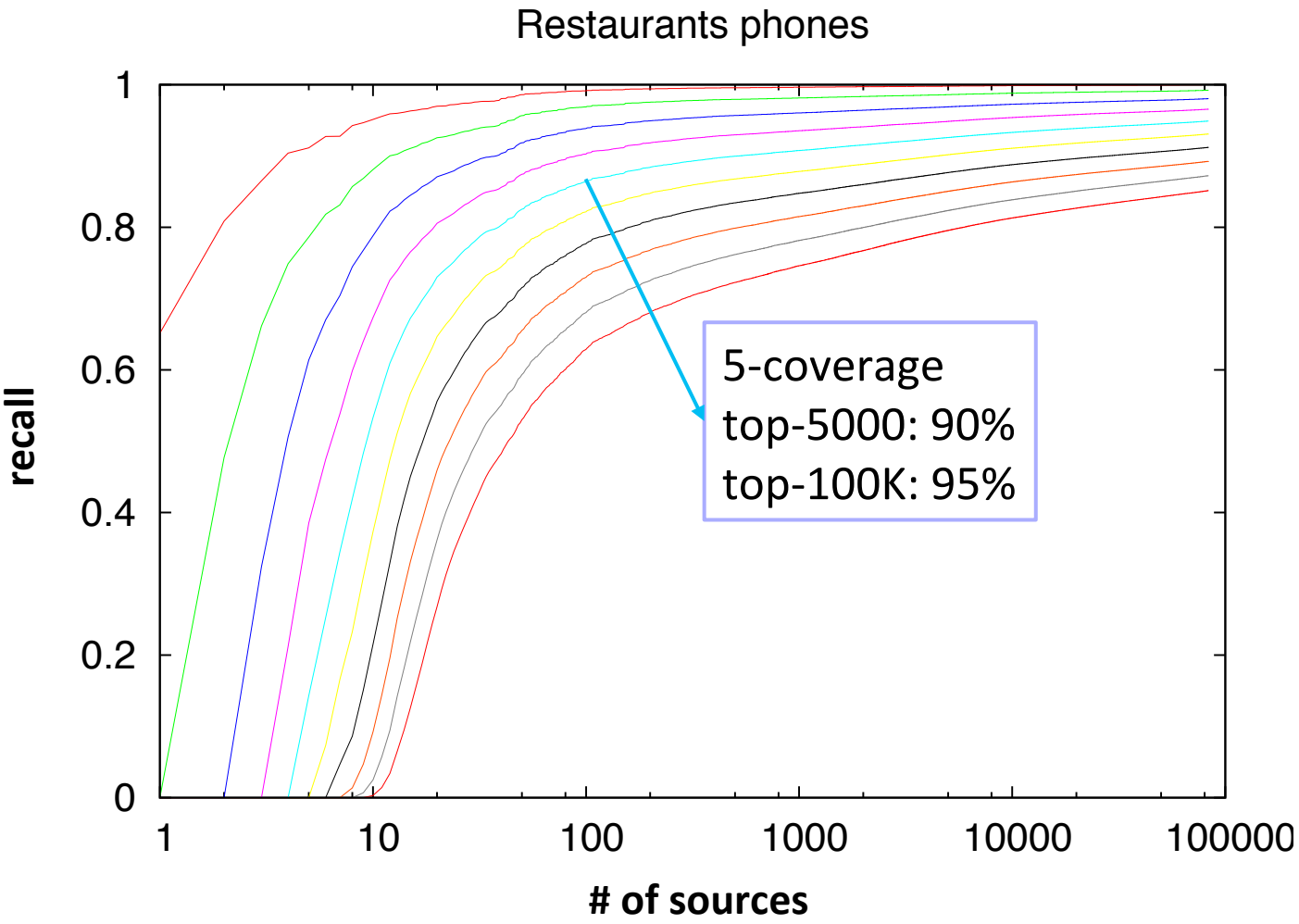
- [DMP12] looked at 9 domains with the following properties
  - Access to large comprehensive databases of entities in the domain
  - Entities have attributes that are (nearly) unique identifiers, e.g., ISBN for Books, phone number or homepage for Restaurants

- 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)

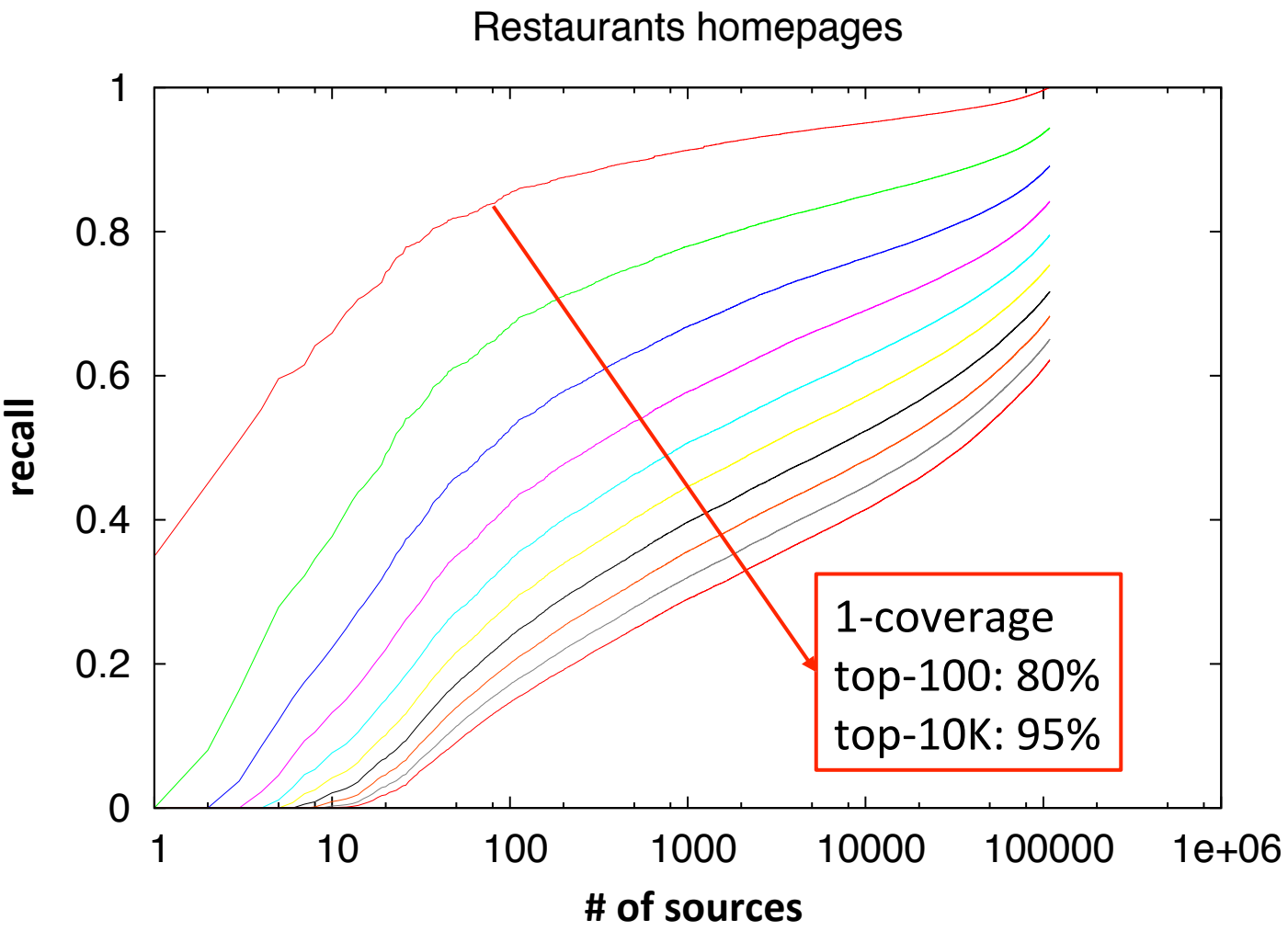
Domain Specific Data: Spread



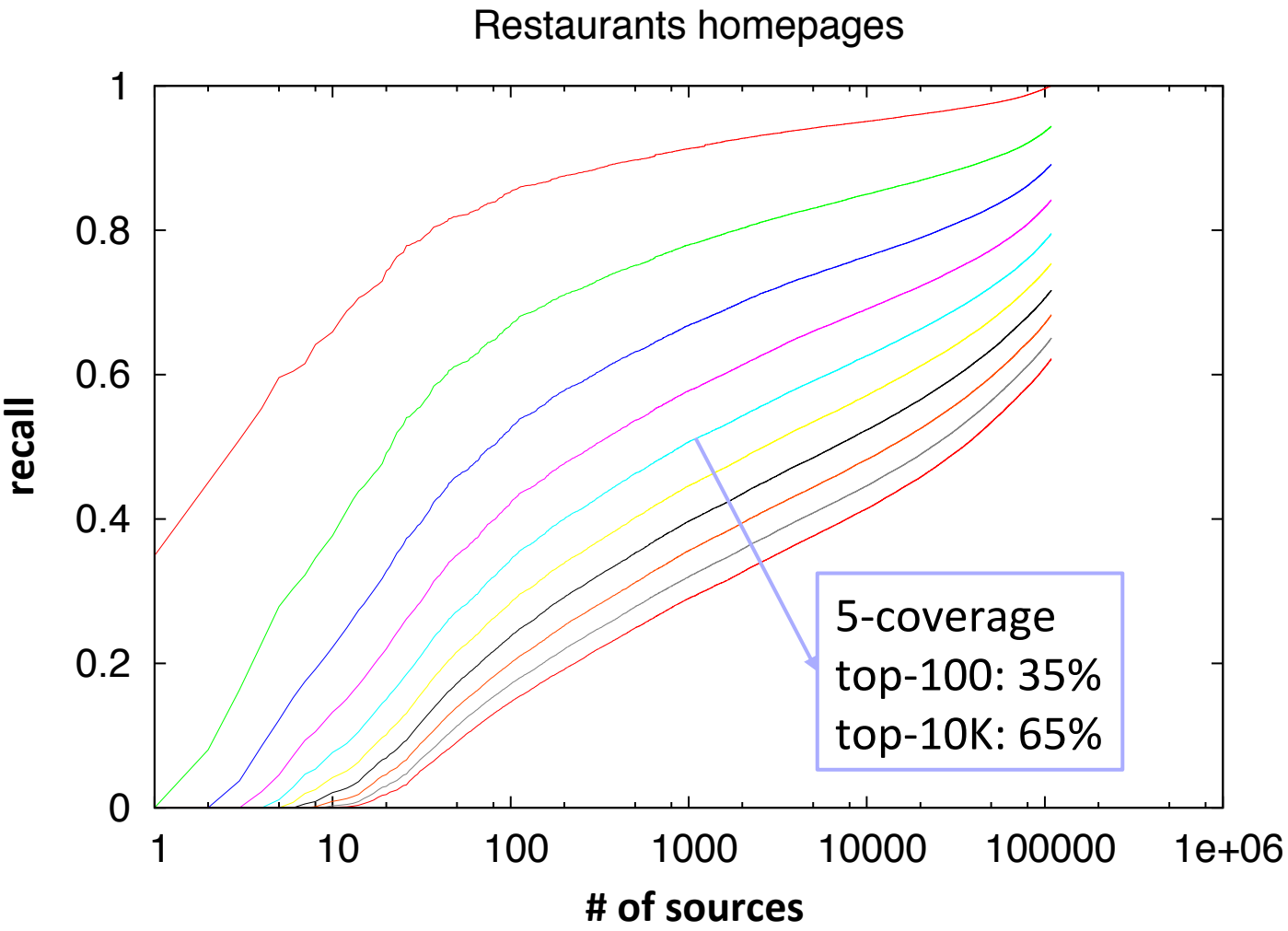
Domain Specific Data: Spread



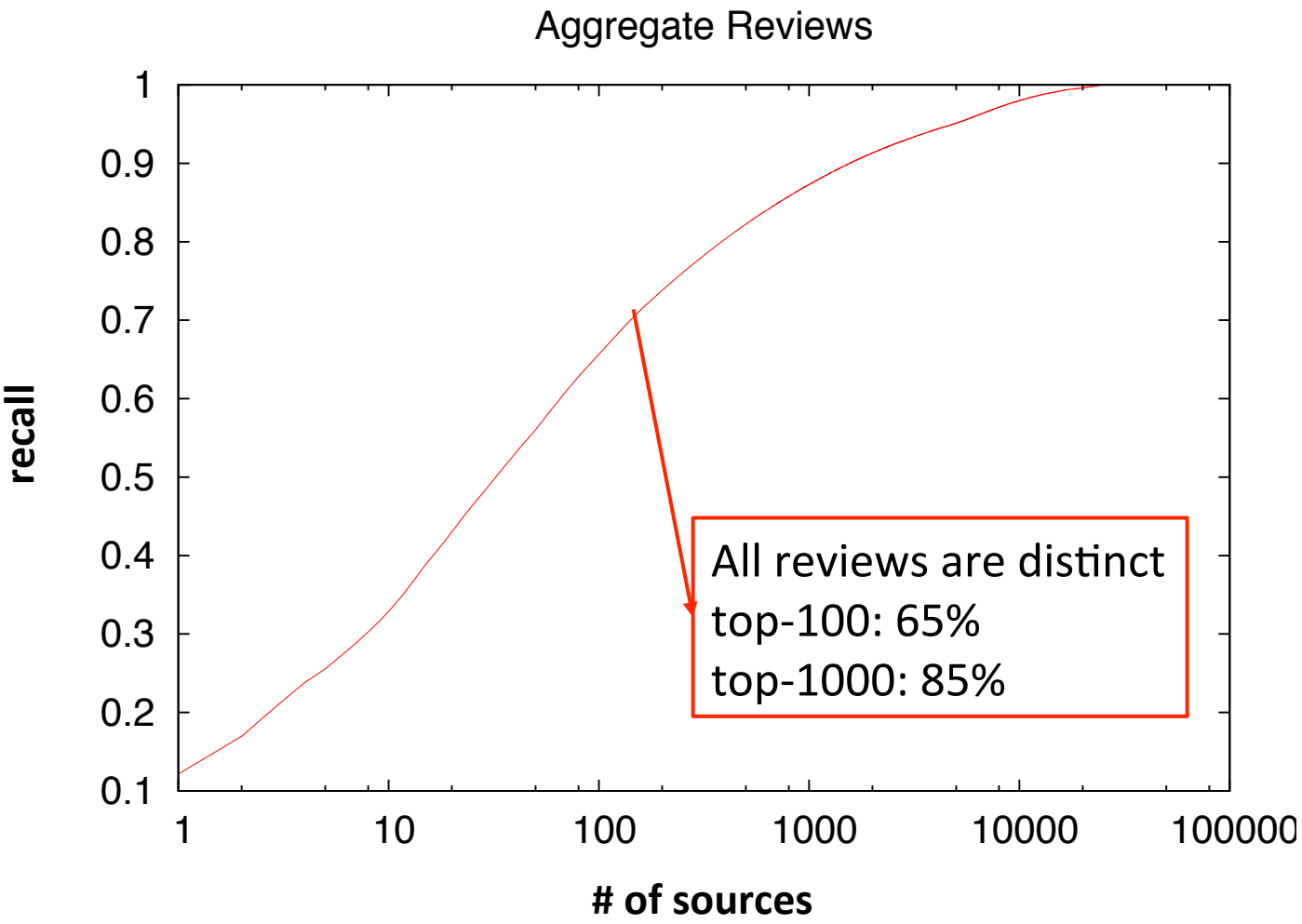
Domain Specific Data: Spread



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

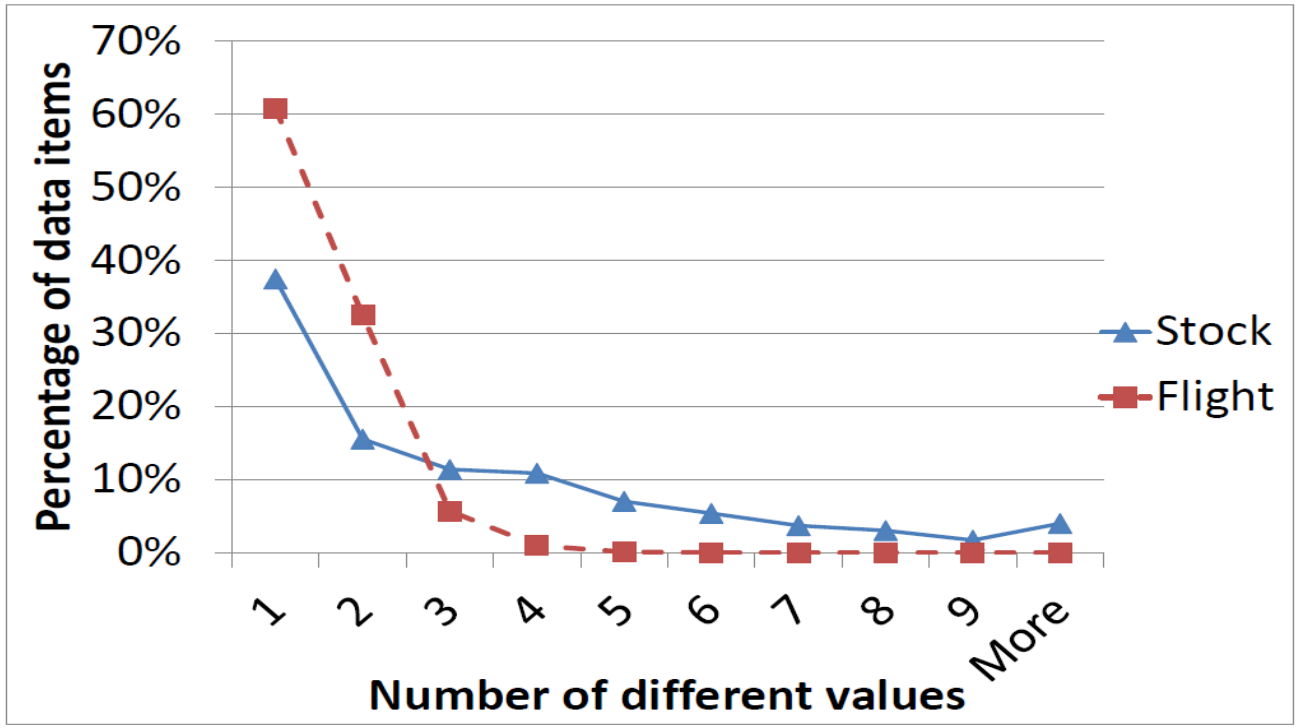
Case Study II: Deep Web Quality [LDL+13]

- 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	Day's Range:	93.80 - 95.71
Trade Time:	4:00PM EDT	52wk Range:	25.38 - 95.71
Change:	↑ 1.69 (1.81%)	Volume:	2,384,075
Prev Close:	93.45	Avg Vol (3m):	2,512,070
Open:	94.01	Market Cap:	13.51B
Bid:	95.03 x 100	P/E (ttm):	119.82
Ask:	95.94 x 100	EPS (ttm):	0.79
1y Target Est:	92.50	Div & Yi	N/A (N/A)

52wk Range: 25.38-95.71

52 Wk: 25.38-93.72

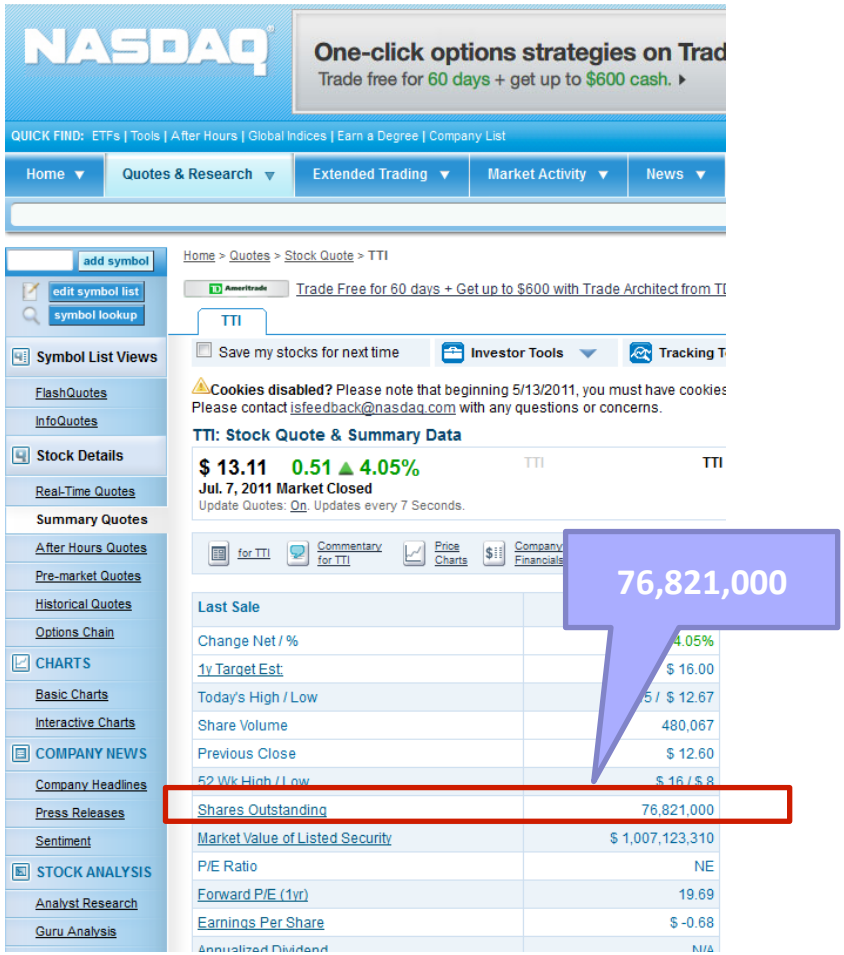
Nasdaq

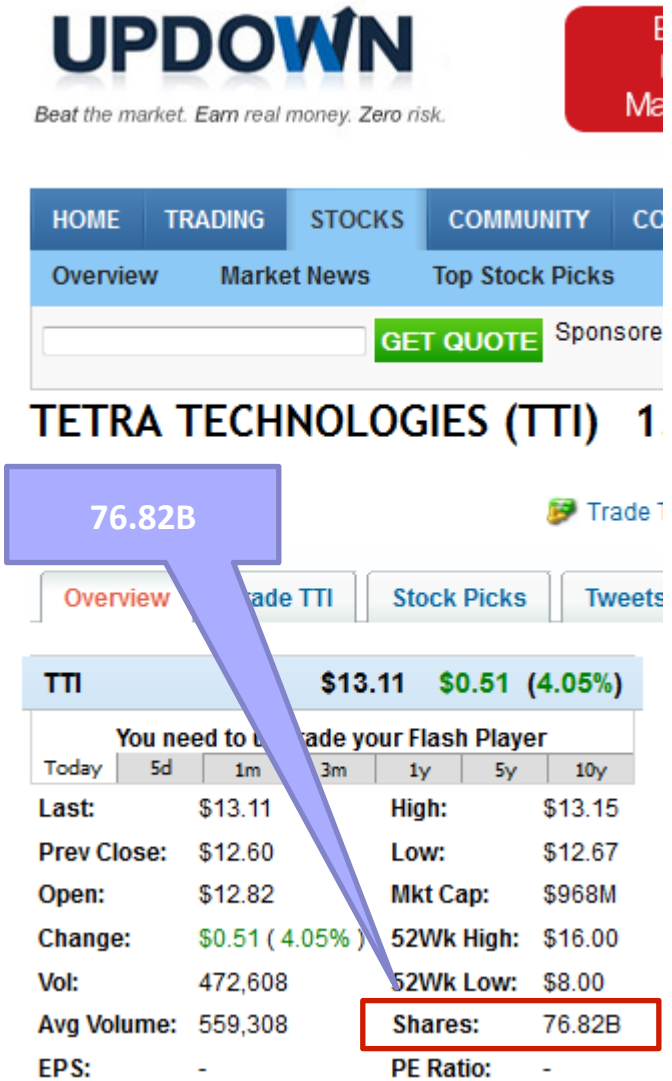
Last Sale	$ 95.14
Change Net / %	1.69 ▲ 1.81%
Best Bid / Ask	$ 95.03 / $ 95.94
1y Target Est:	$ 95.00
Today's High / Low	$ 95.71 / $ 93.80
Share Volume	2,384,175
50 Day Avg. Daily Volume	2,751,062
Previous Close	$ 93.45
52 Wk High / Low	3.72 / $ 25.38
Shares Outstanding	152,785,000
Market Value of Listed 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 Da	N/A
Dividend	N/A
Cu d	N/A
	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

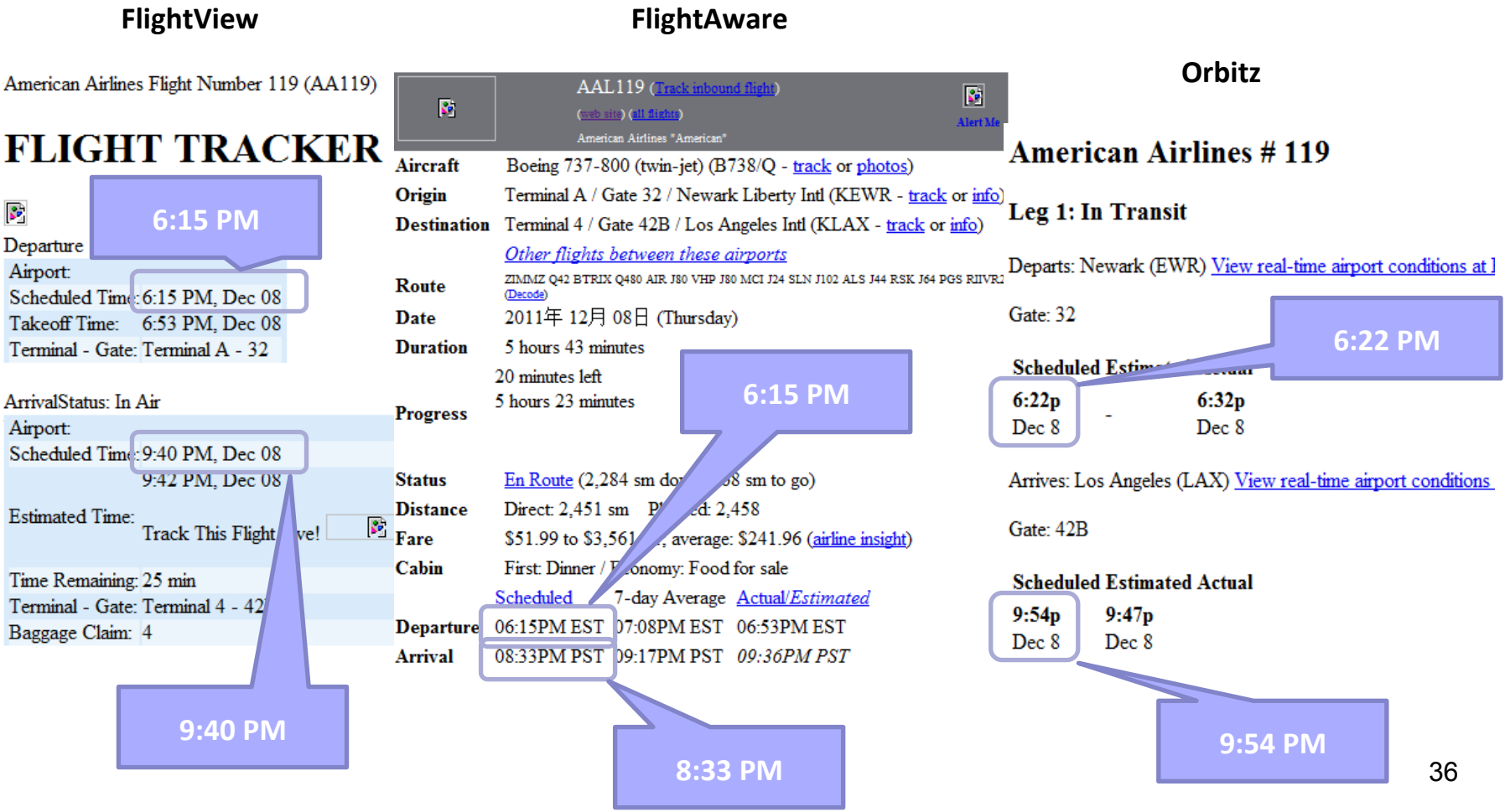
- Why such inconsistency?
  - Unit errors





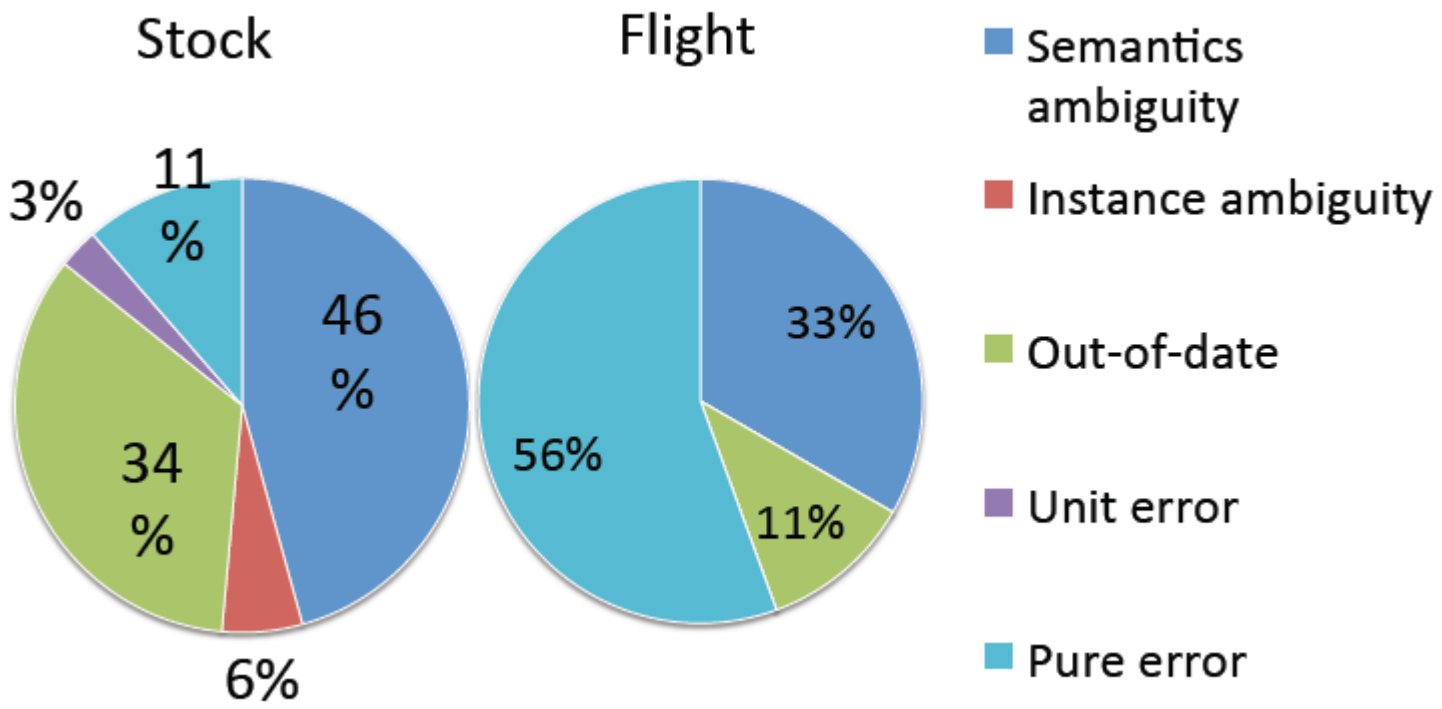
Deep Web Quality

- Why such inconsistency?
  - Pure errors



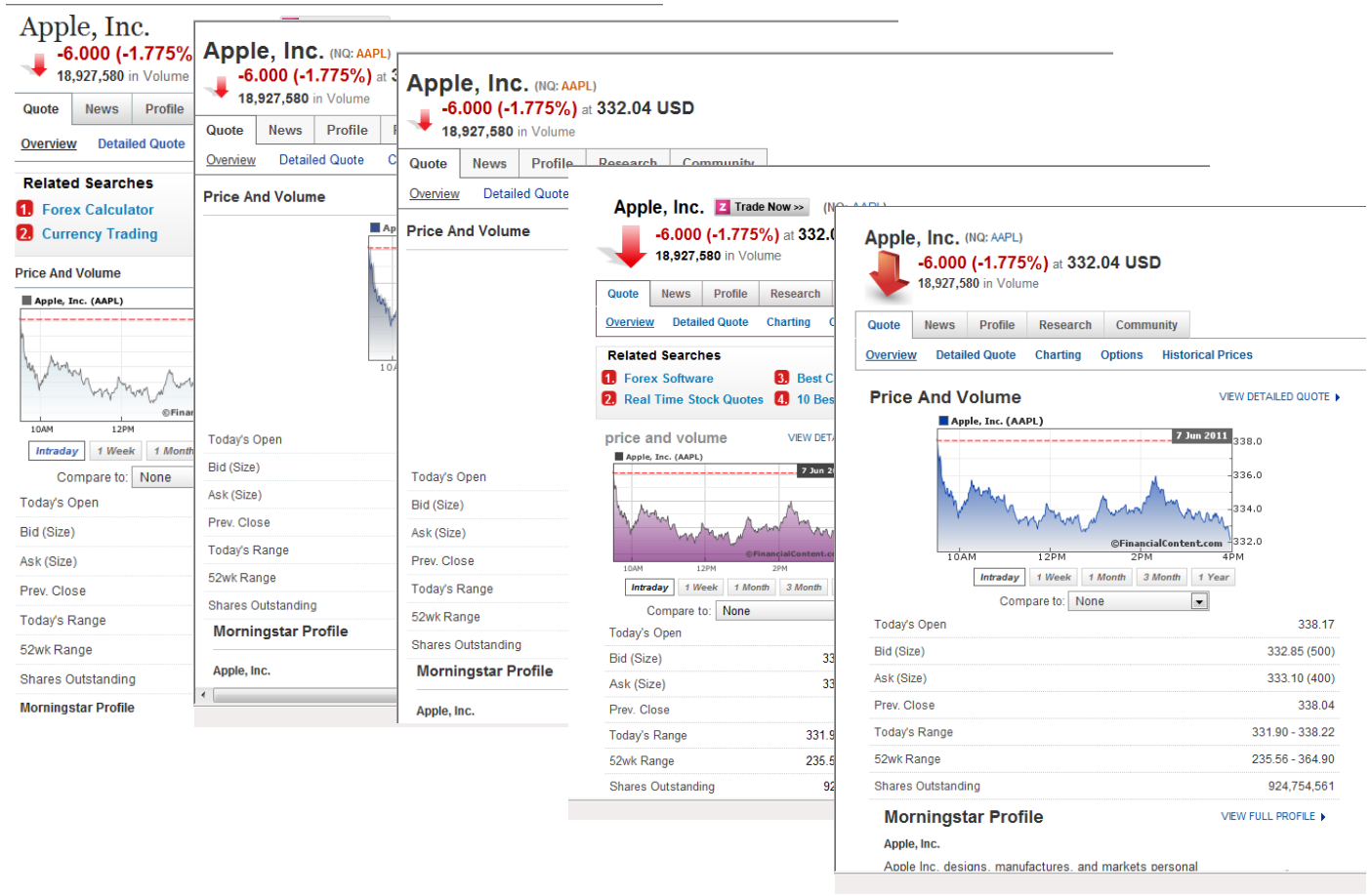
Deep Web Quality

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



Deep Web Quality

- Copying between sources?



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

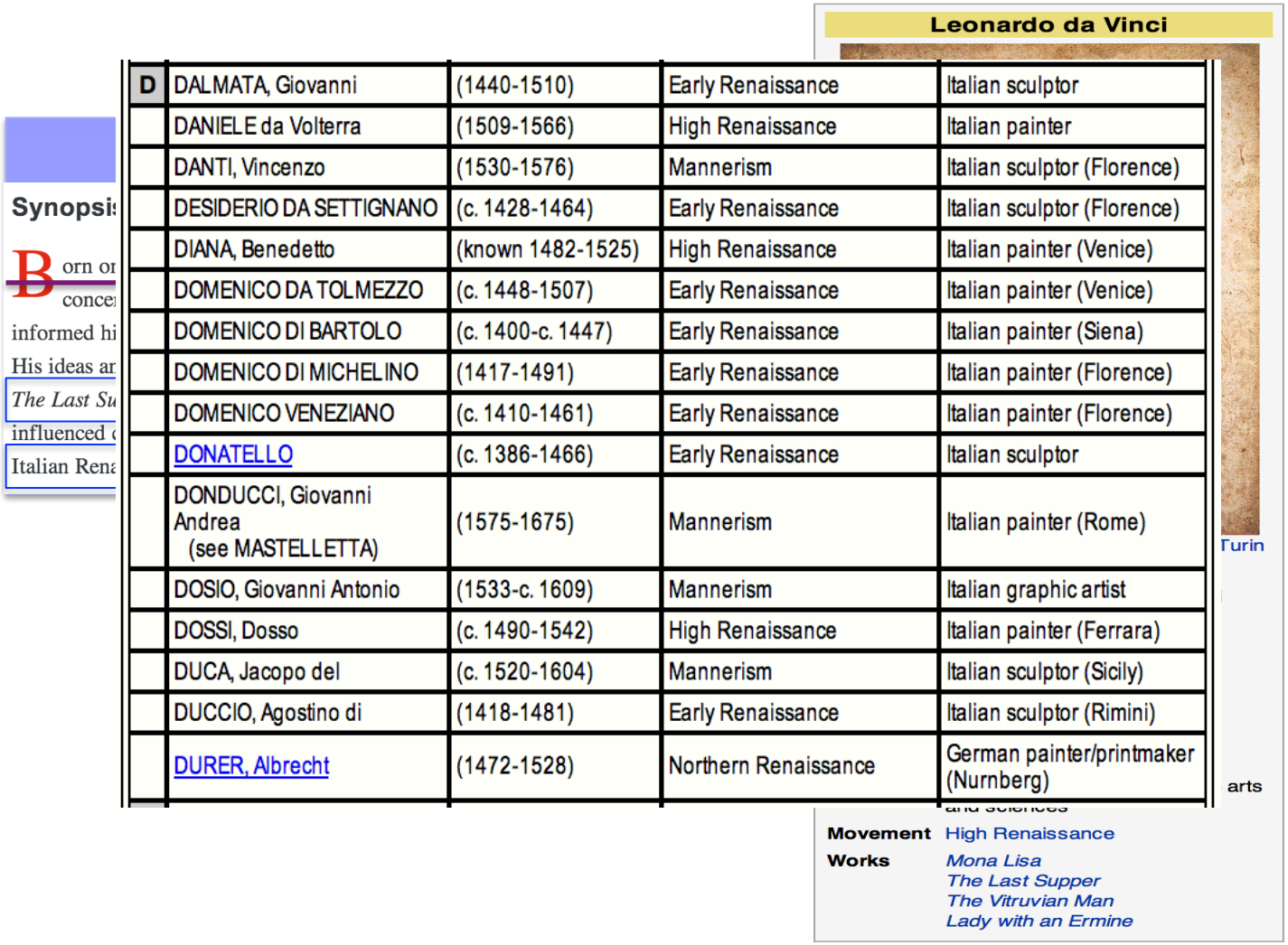
BDI: Why is it Challenging?

- 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**



Synopsi

B orn or
concei
informed hi
His ideas ar
The Last Su
influenced
Italian Rena

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	(1472-1528)	Northern Renaissance	German painter/printmaker (Nurnberg)

Leonardo da Vinci

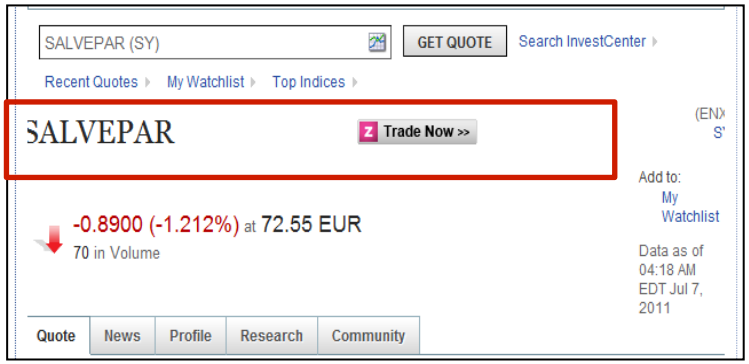
Turin

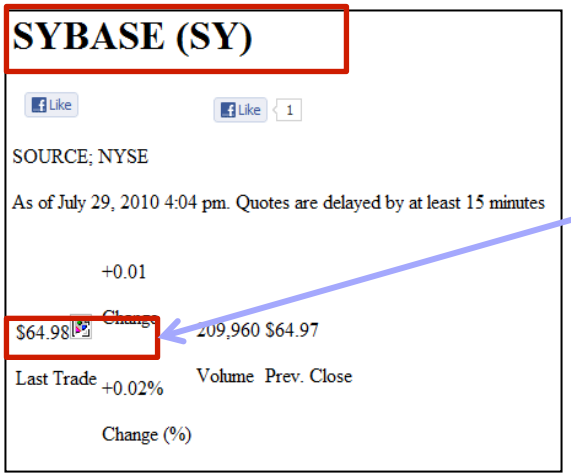
arts

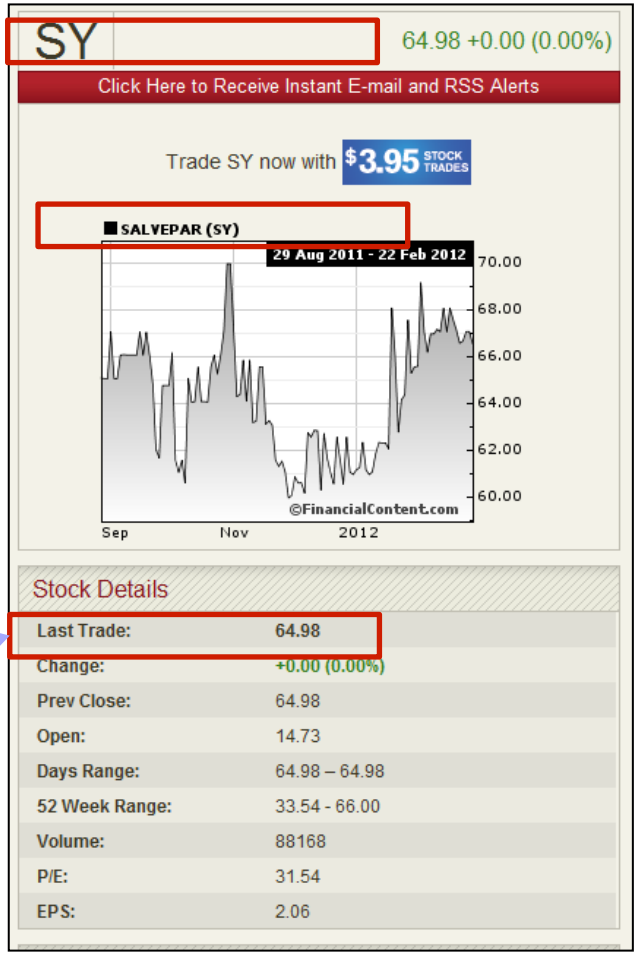
Movement High Renaissance
Works *Mona Lisa*
The Last Supper
The Vitruvian Man
Lady with an Ermine

BDI: Why is it Challenging?

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





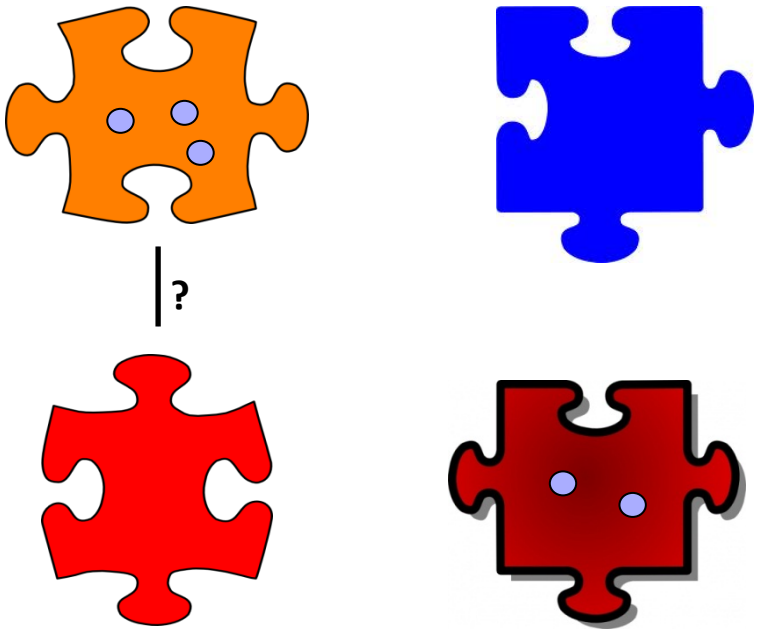


Outline

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

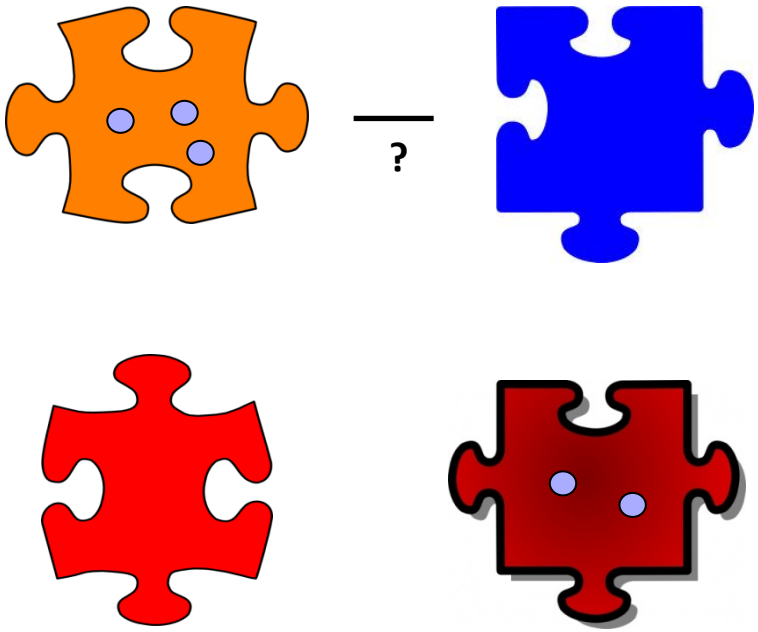
Schema Alignment

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



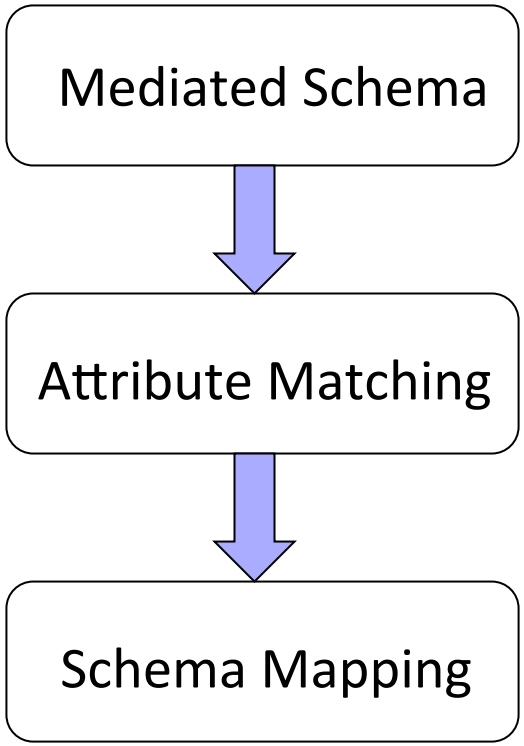
Schema Alignment

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



Schema Alignment: Three Steps [BBR11]

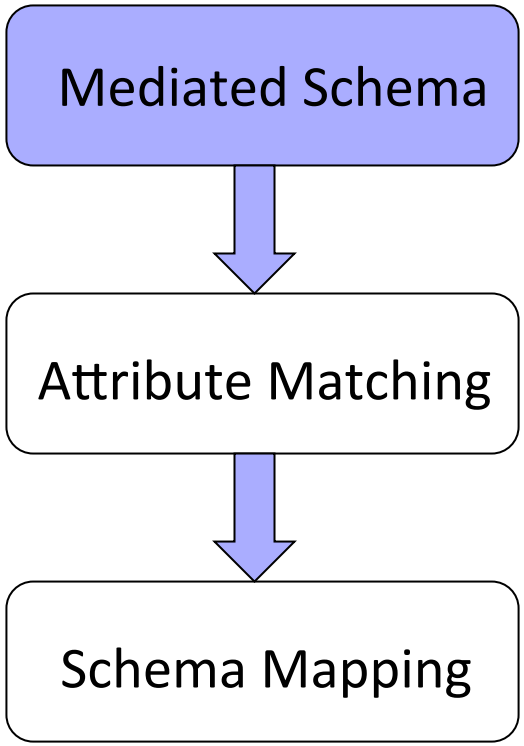
- Schema alignment: mediated schema + matching + mapping
  - Enables linkage, fusion to be semantically meaningful



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)

Schema Alignment: Three Steps

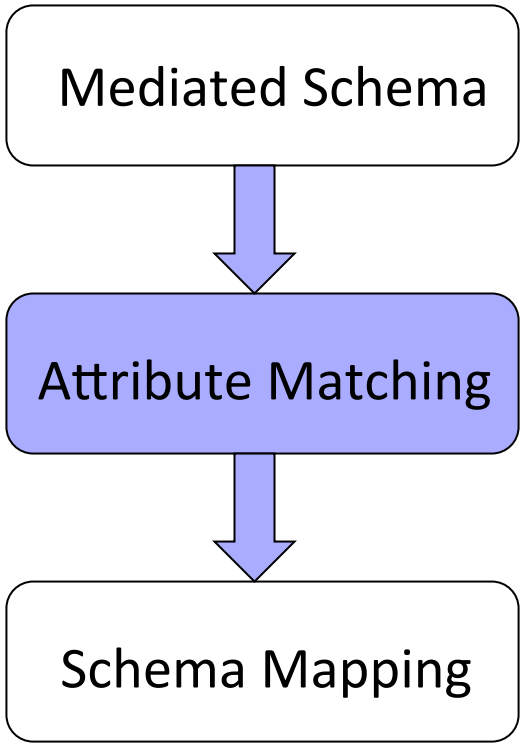
- Schema alignment: mediated schema + matching + mapping
  - Enables domain specific modeling



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)**

Schema Alignment: Three Steps

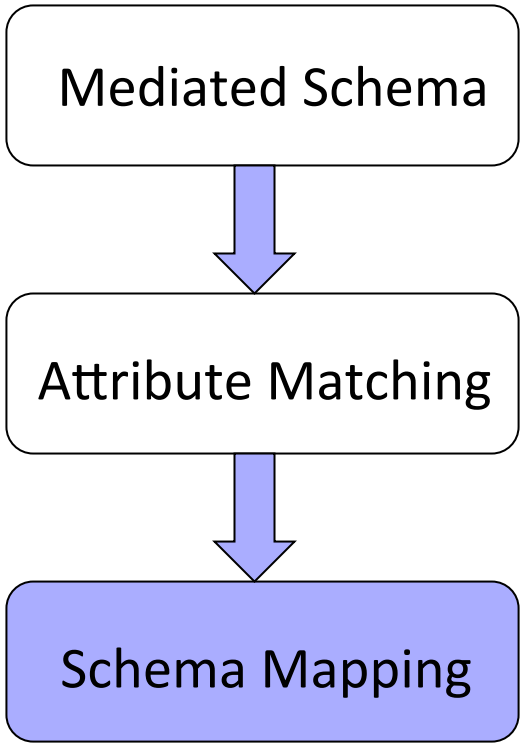
- Schema alignment: mediated schema + matching + mapping
  - Identifies correspondences between schema attributes



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

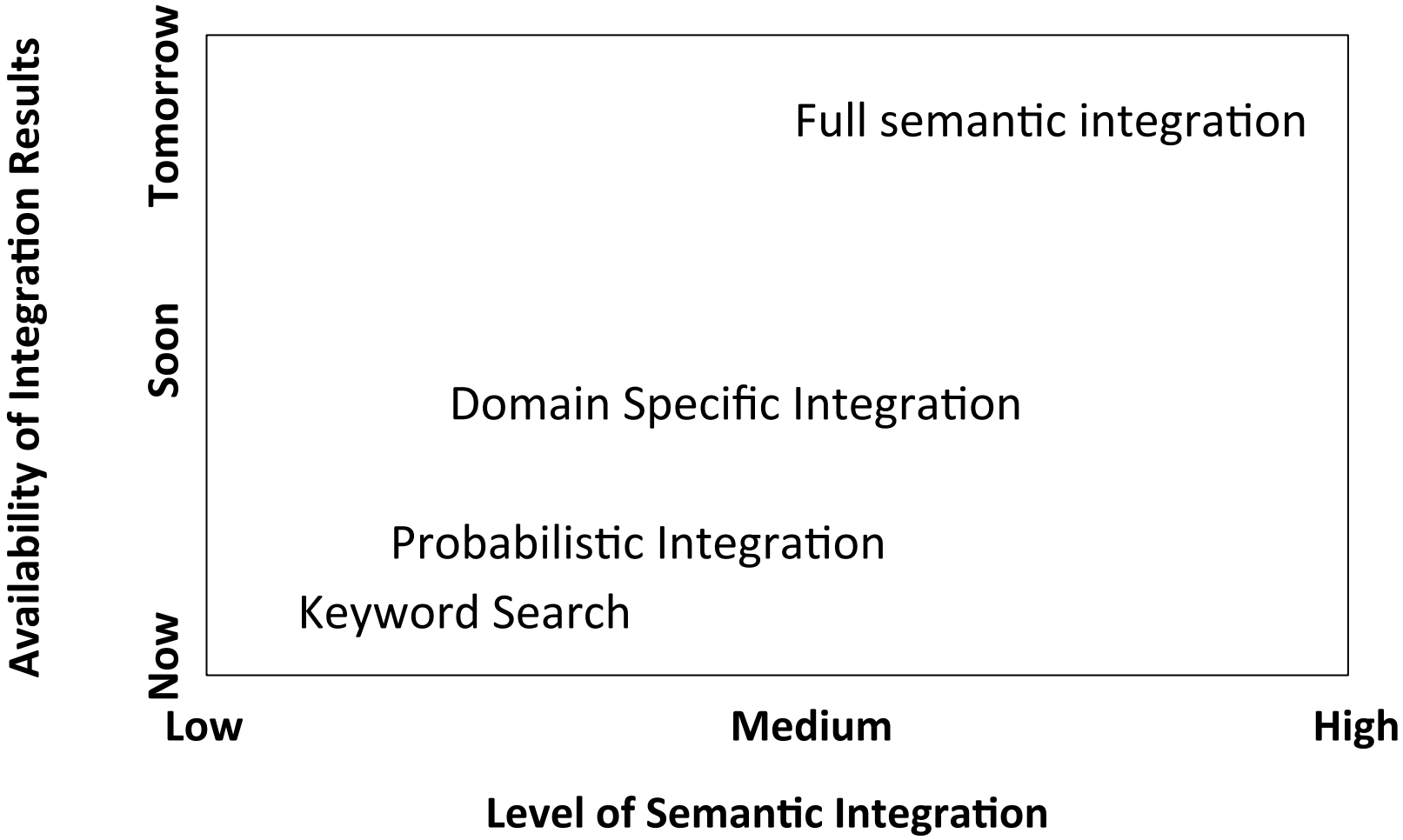
- **Volume, Variety**
  - Integrating deep web query interfaces [WYD+04, CHZ05]
  - Crawl, index deep web data [MKK+08]
  - Extract structured data from web tables [CHW+08, LSC10, PS12, DFG+12] and web lists [GS09, EMH09]
  - Dataspace systems [FHM05, HFM06, DHY07]
  - Keyword search based data integration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



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 …

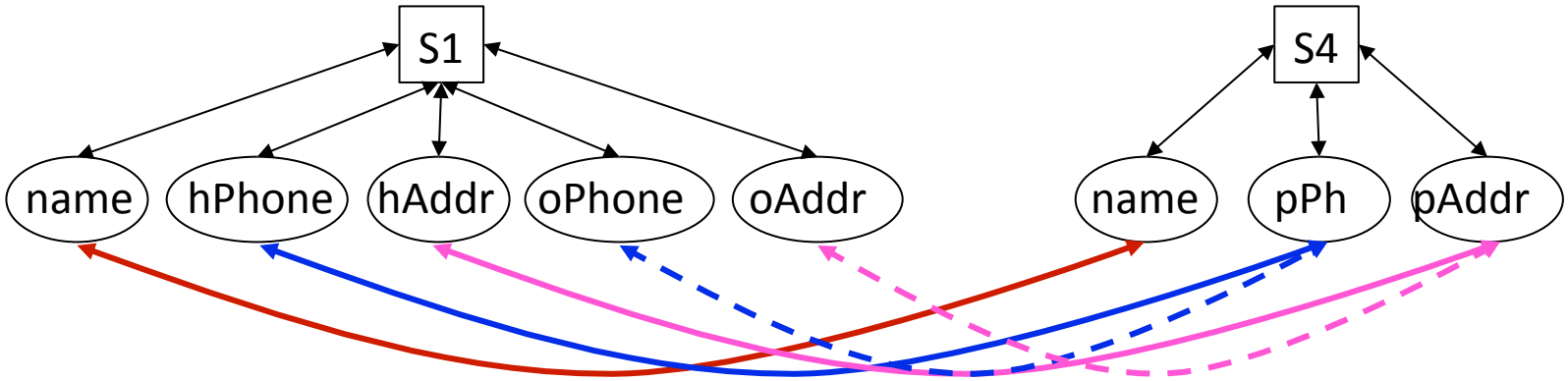
Bootstrapping DI Systems [DDH08]

- 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

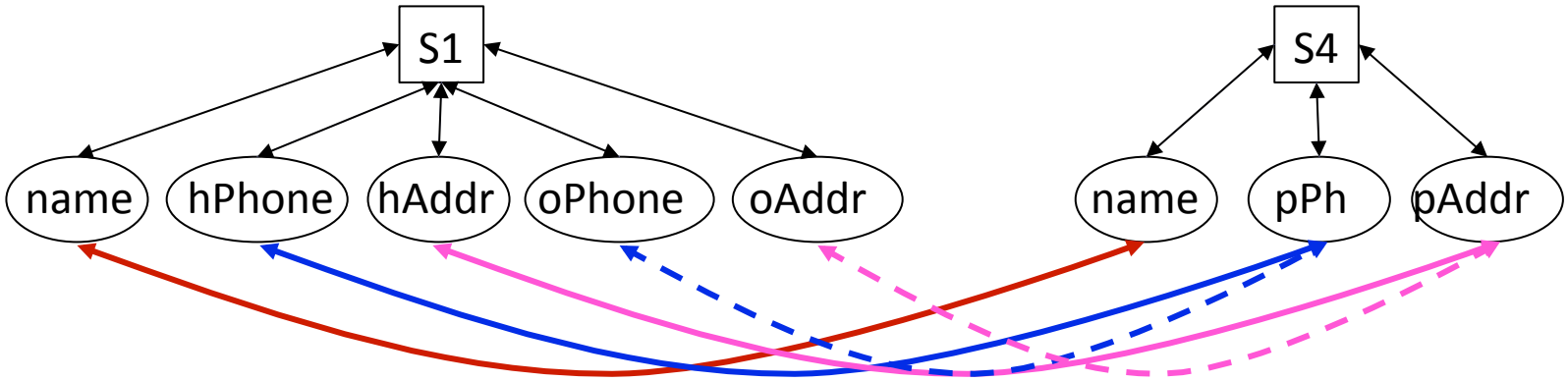
- Automatically create mappings from sources to mediated schema
  - Probabilistic mappings use weighted attribute correspondences

Probabilistic Mediated Schemas [DDH08]



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

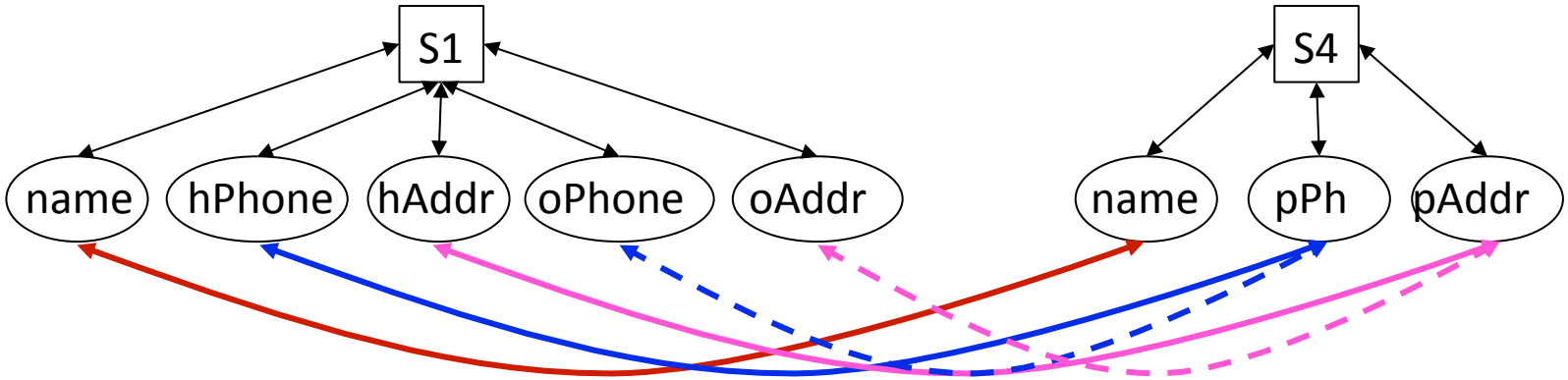
Probabilistic Mediated Schemas [DDH08]



- 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]

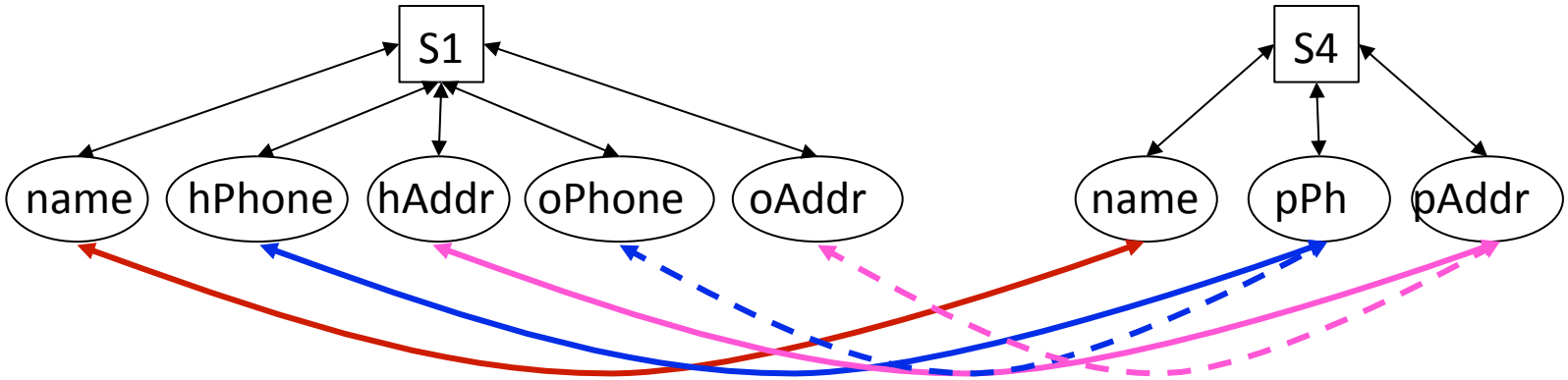
- Mapping between P-mediated schema and a source schema



- Example mappings between M1 and S1
  - G1({**M1.n**, name}, {**M1.phP**, hPhone}, {**M1.phA**, hAddr}, …)
  - G2({**M1.n**, name}, {**M1.phP**, oPhone}, {**M1.phA**, oAddr}, …)
  - G = {(G1, 0.6), (G2, 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

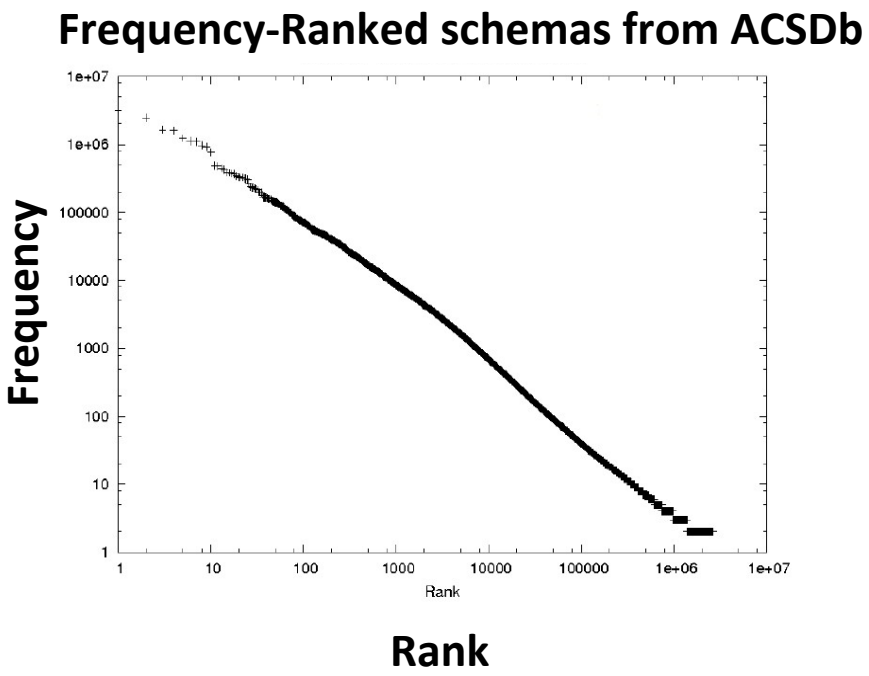
- Result of Q2, under by tuple semantics, across all possible worlds
  - Note the difference with the result of Q2, under by table semantics

Q2R	pAddr	Prob
	Summit	0.76
	New York	0.76

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
  - (schema, count)

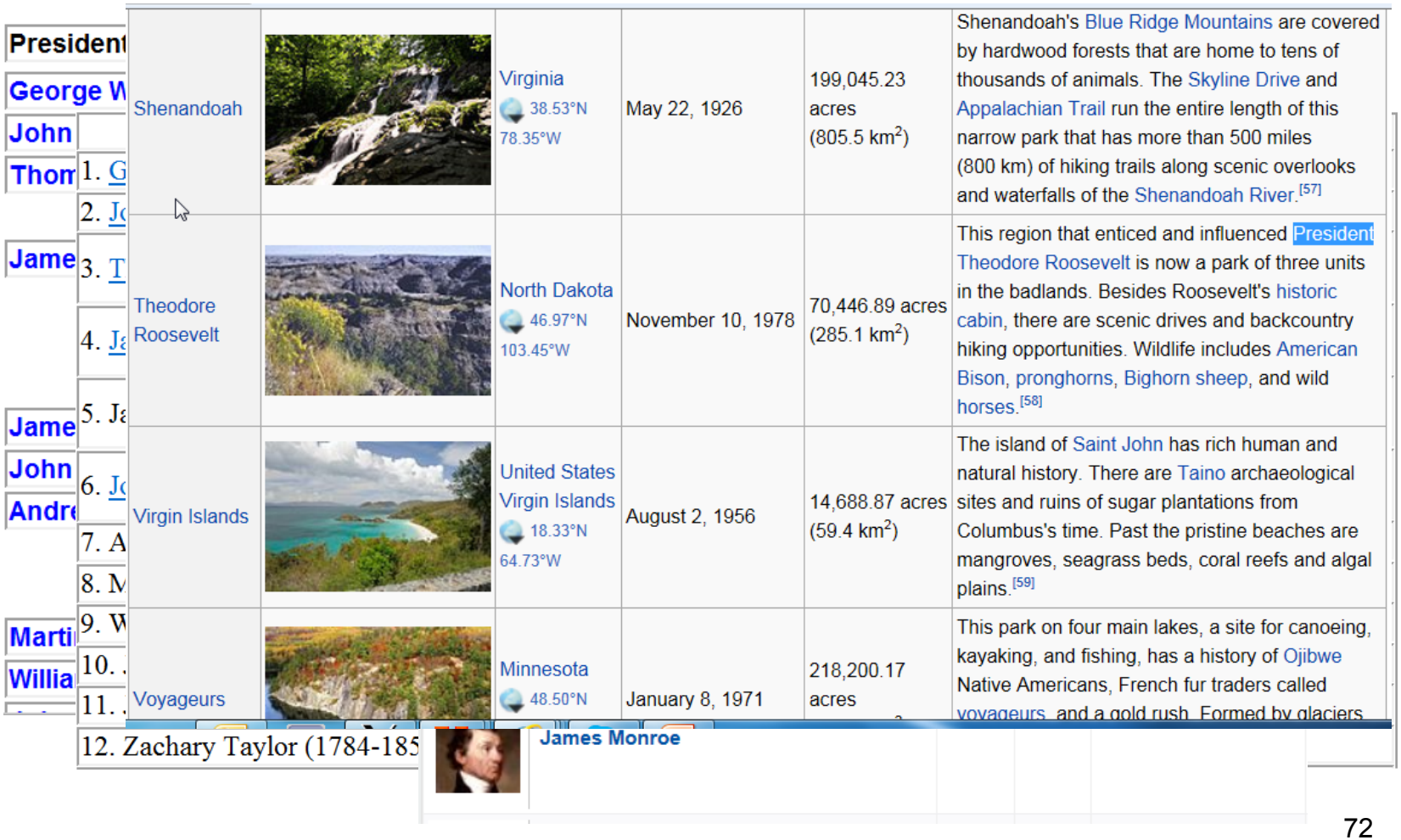


WebTables: Keyword Search [CHW+08]

- 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

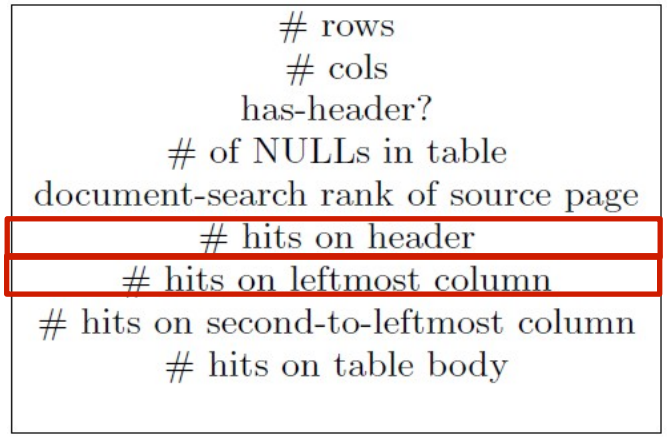
- Example keyword query: “presidents of the US”



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]
Theodore Roosevelt		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]
Virgin Islands		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]
Voyageurs		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

WebTables: Keyword Search

- FeatureRank: use table specific features
  - Query independent features
  - Query dependent features
  - Linear regression estimator
  - Heavily weighted features

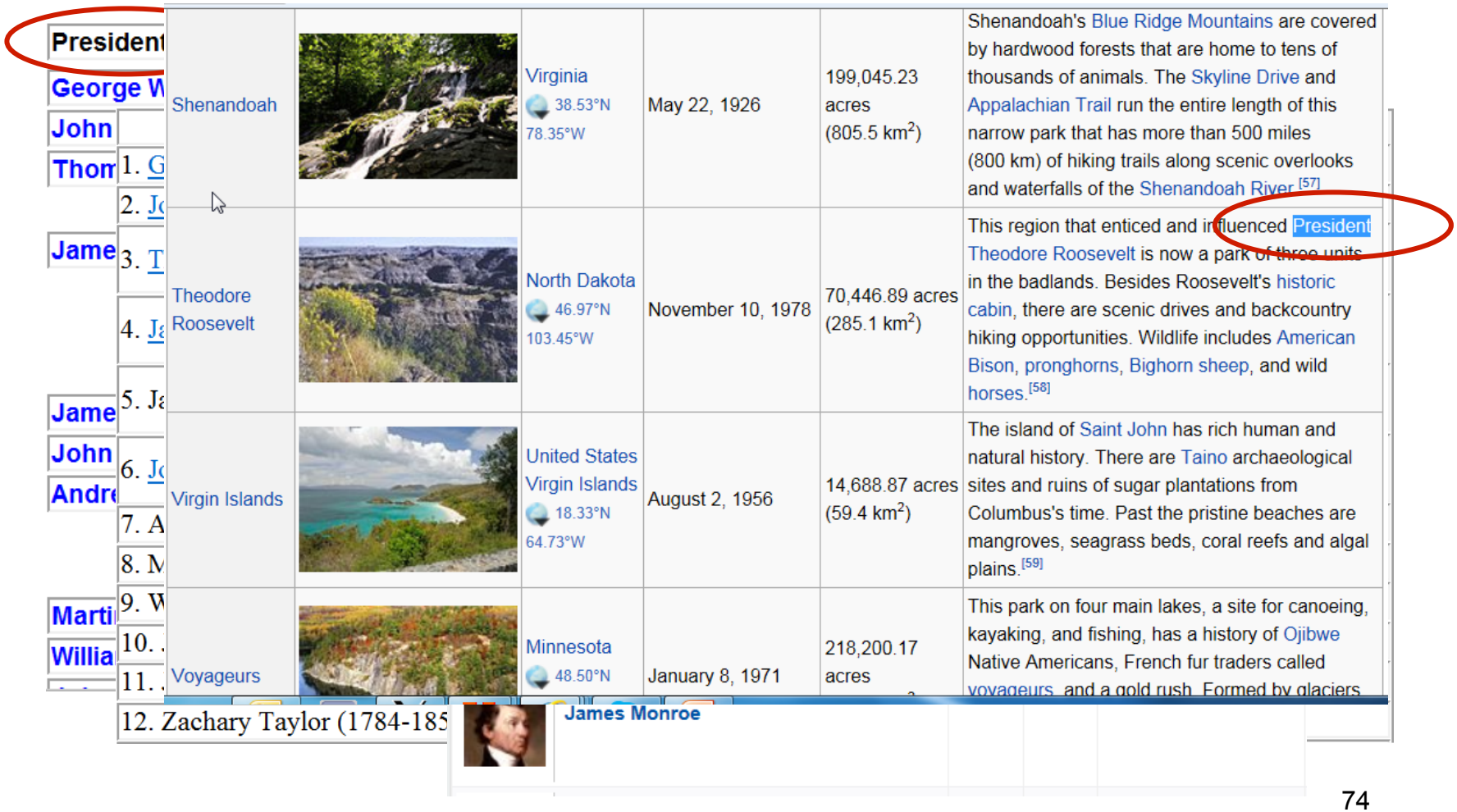


- Result quality: fraction of high scoring relevant tables

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						
Thom	1. G					
	2. Jo					
Jame	3. T					
	4. Ja	Theodore Roosevelt		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]
	5. Ja					
Jame						
John	6. Jo	Virgin Islands		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]
Andre	7. A					
	8. M					
Marti	9. W					
Willia	10. .					
	11. .	Voyageurs		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
	12. Zachary Taylor (1784-185			James Monroe			

WebTables: Keyword Search

- SchemaRank: also include **schema coherency** as a table feature
  - Use point-wise mutual information (pmi) derived from ACSDb
  - p(a) = fraction of unique schemas containing attributes a
  - 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)

- pmi(a,b) = $\log_2(p(a,b)/(p(a)*p(b)))$
  - pmi(**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**)

- pmi(a,b) = $\log_2(p(a,b)/(p(a)*p(b)))$
  - pmi(President, Vice President) = $\log_2(0.4/(0.4 * 0.4)) = 1.32$
  - pmi(**President**, **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 pmi(a,b) over all a, b in attrs(R)
  - coherency(T1) = avg({1.32}) = 1.32
  - coherency(T2) = avg({1.32, -0.68, -0.26, 0.32, -0.26, -0.68}} = -0.15

Annotating Web Tables [LSC10]

- 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

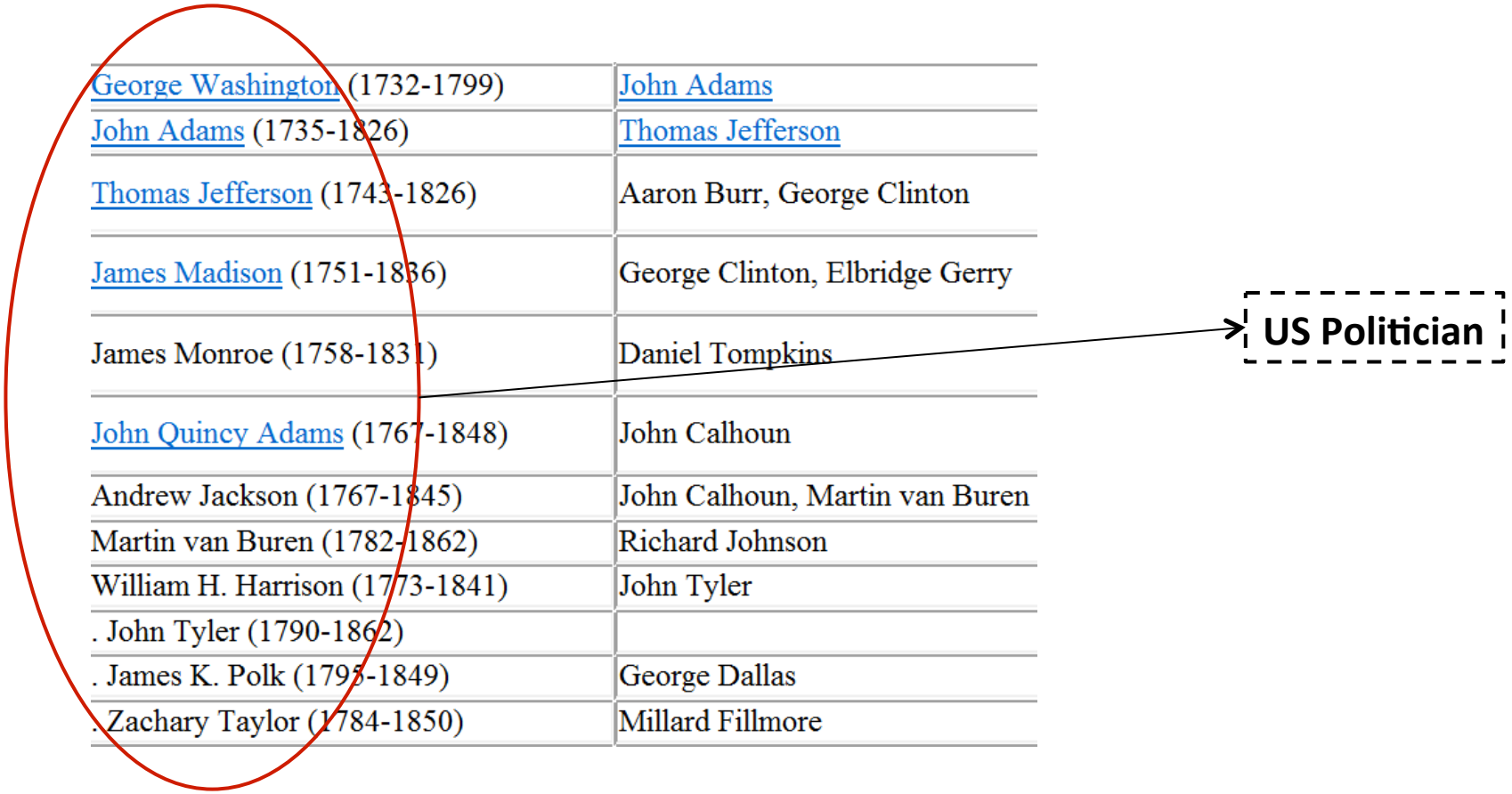
- Goal: given a Web table, which entities occur in which cells, what are the column types, and the relationships between columns?

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?

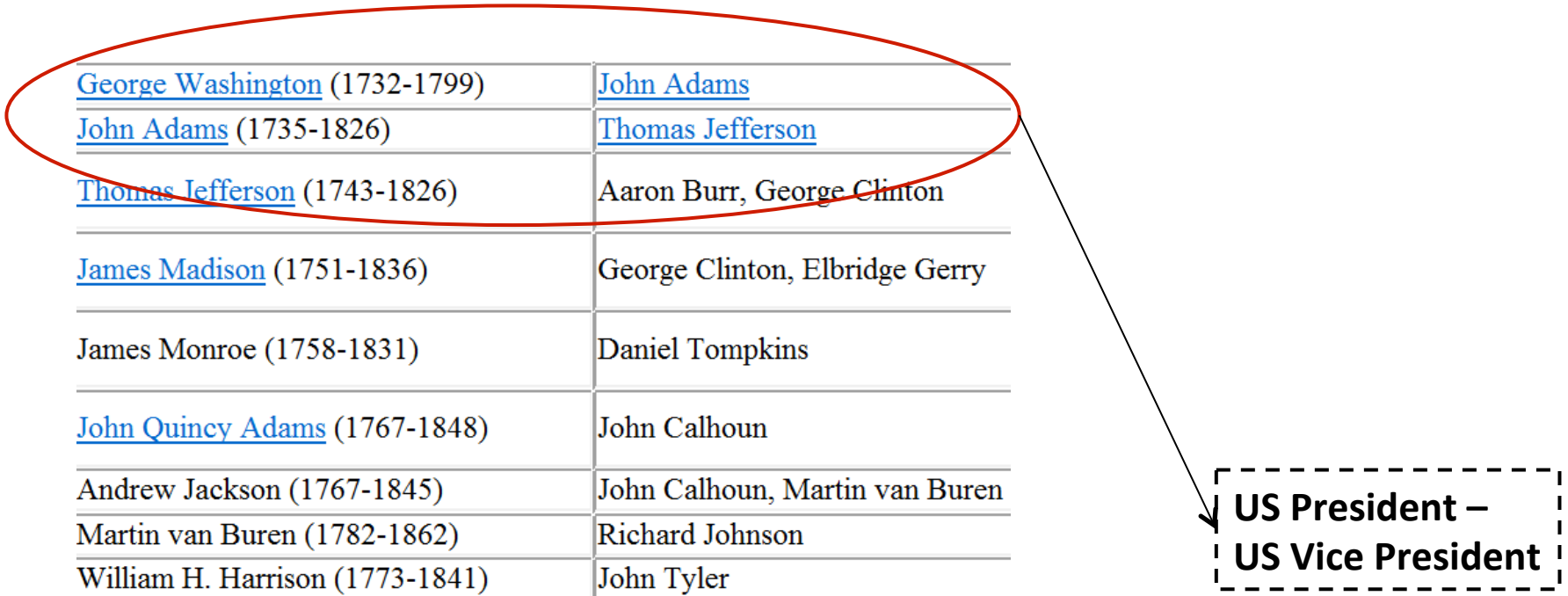
George Washington (1732-1799)	John Adams
John Adams (1735-1826)	Thomas Jefferson
Thomas Jefferson (1743-1826)	Aaron Burr, George Clinton
James Madison (1751-1836)	George Clinton, Elbridge Gerry
James Monroe (1758-1831)	Daniel Tompkins
John Quincy Adams (1767-1848)	John Calhoun
Andrew Jackson (1767-1845)	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



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?

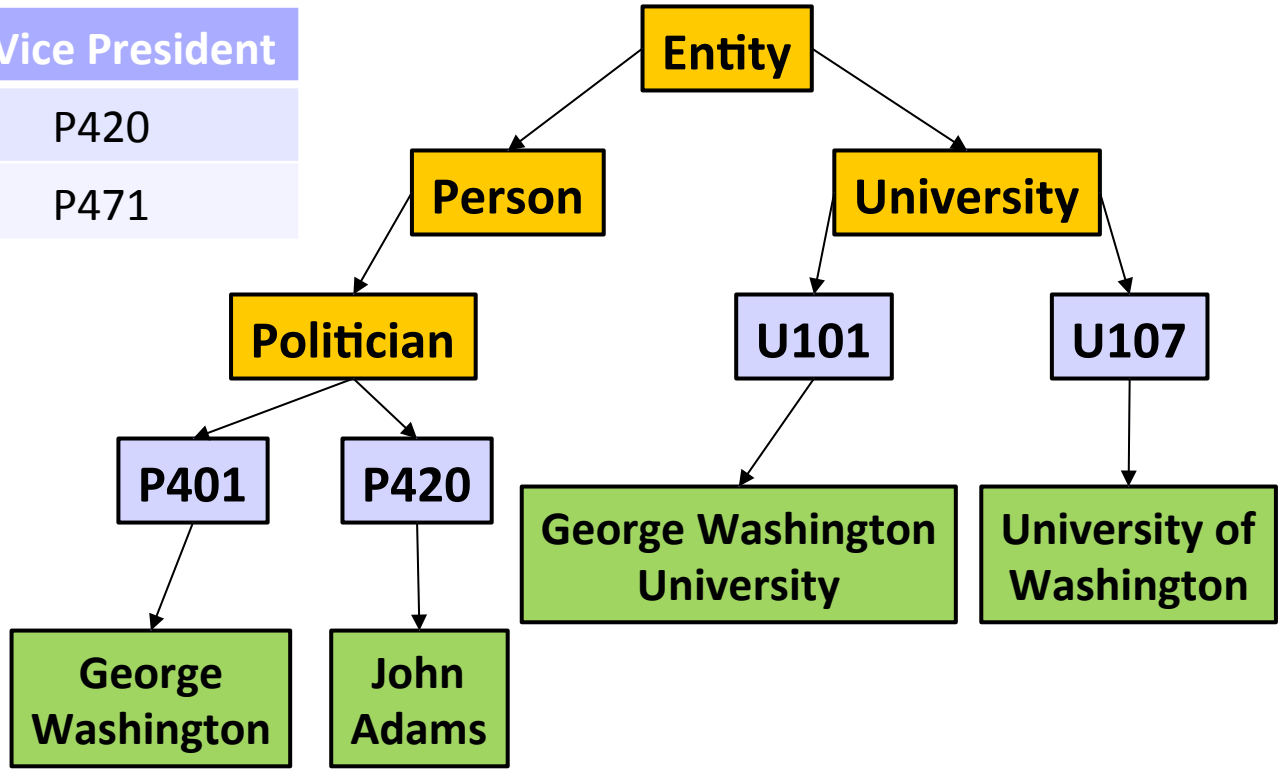
George Washington (1732-1799)	John Adams
John Adams (1735-1826)	Thomas Jefferson
Thomas Jefferson (1743-1826)	Aaron Burr, George Clinton
James Madison (1751-1836)	George Clinton, Elbridge Gerry
James Monroe (1758-1831)	Daniel Tompkins
John Quincy Adams (1767-1848)	John Calhoun
Andrew Jackson (1767-1845)	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



Annotating Web Tables: Using a Catalog

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

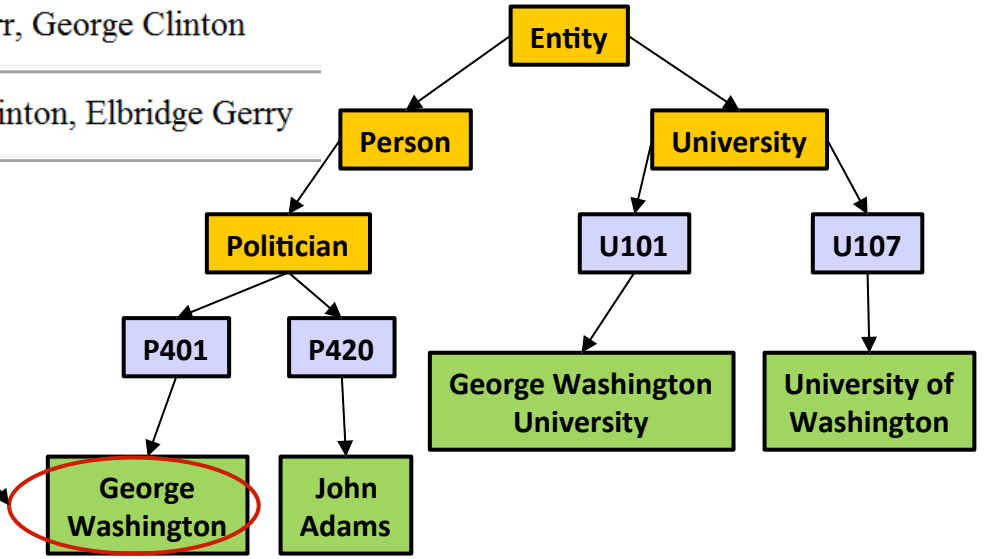
US President	US Vice President
P401	P420
P420	P471



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)

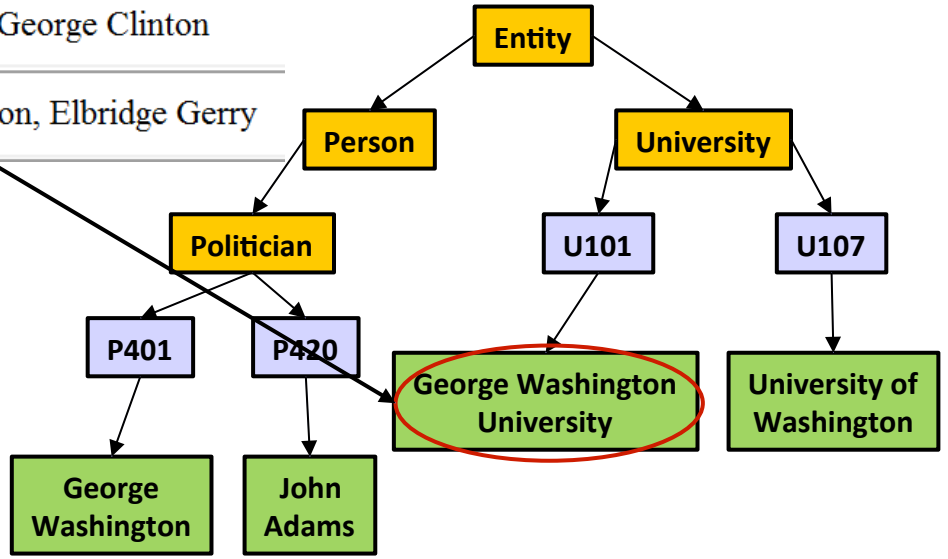
George Washington (1732-1799)	John Adams
John Adams (1735-1826)	Thomas Jefferson
Thomas Jefferson (1743-1826)	Aaron Burr, George Clinton
James Madison (1751-1836)	George Clinton, Elbridge Gerry



Annotating Web Tables: Using a Catalog

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

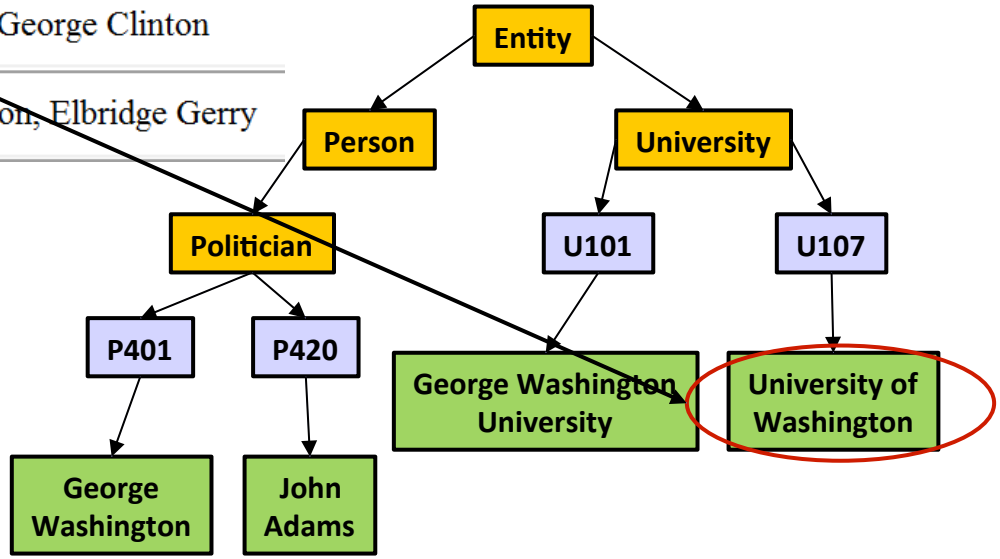
George Washington (1732-1799)	John Adams
John Adams (1735-1826)	Thomas Jefferson
Thomas Jefferson (1743-1826)	Aaron Burr, George Clinton
James Madison (1751-1836)	George Clinton, Elbridge Gerry



Annotating Web Tables: Using a Catalog

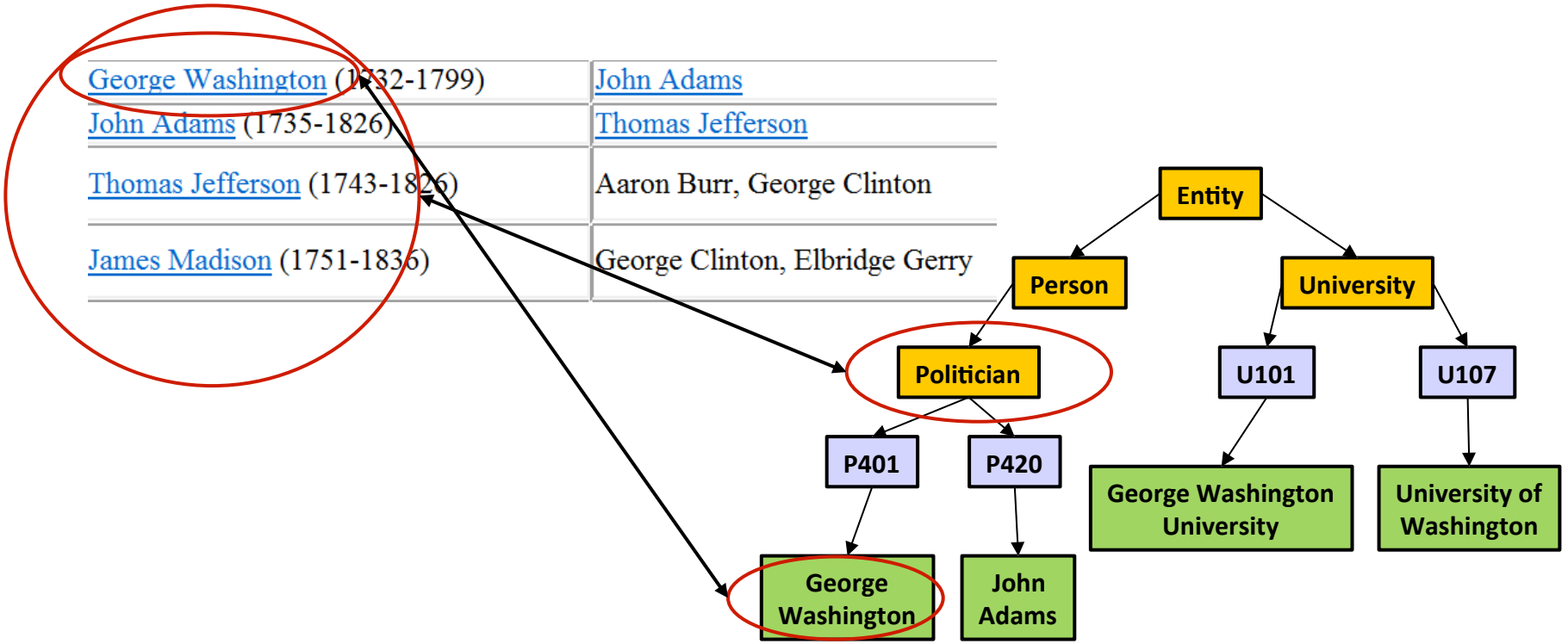
- How good is it to label cell (r, c), containing text D_{rc}, with entity E?
  - Similarity between D_{rc} and L(E)

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



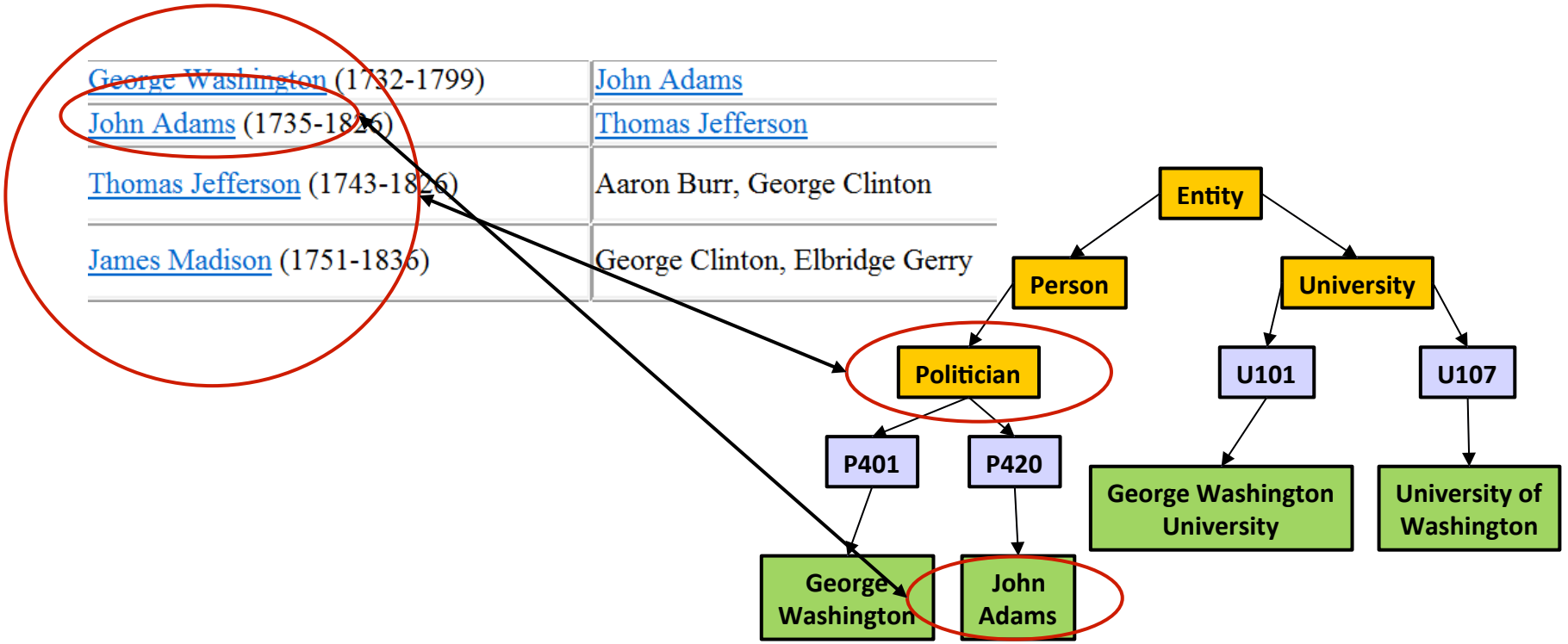
Annotating Web Tables: Using a Catalog

- 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

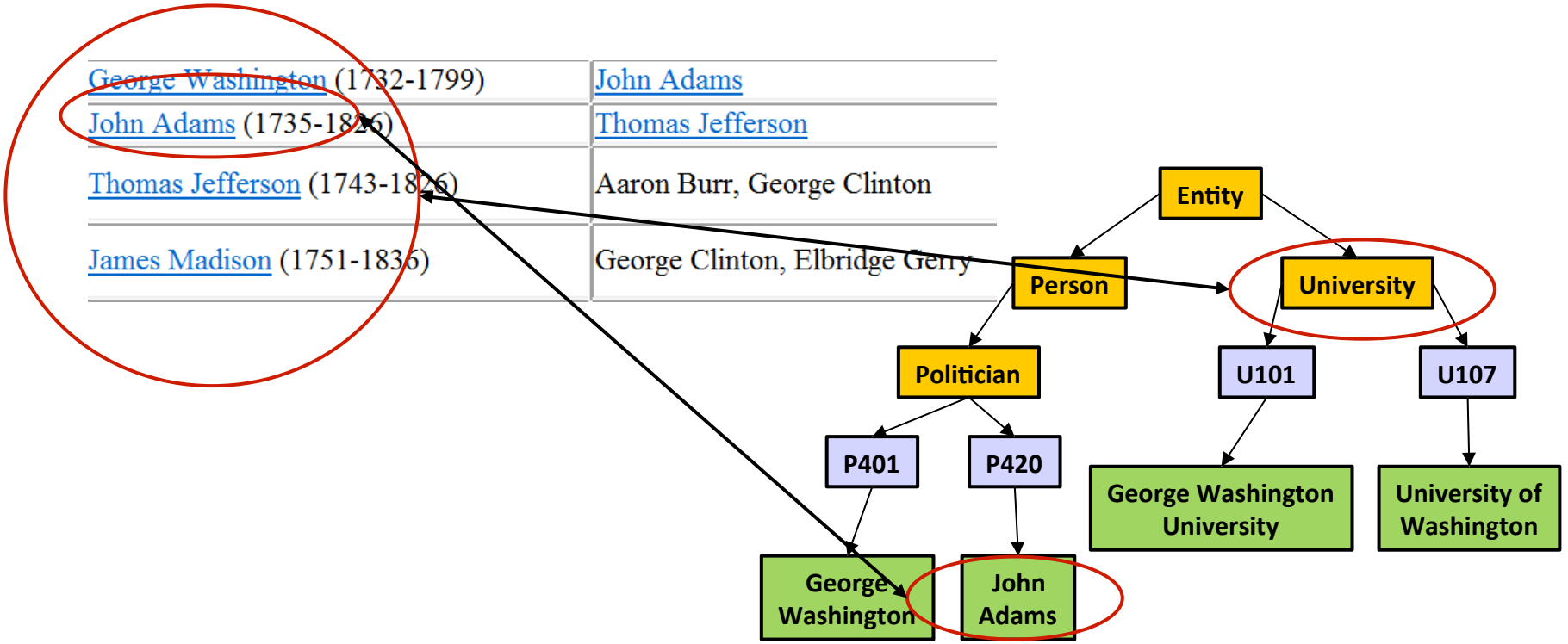
- How good is it to label column c with type T and cell (r, c) with E?
  - Entity E should belong to type T (but catalog may be incomplete)



Annotating Web Tables: Using a Catalog

X

- 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)

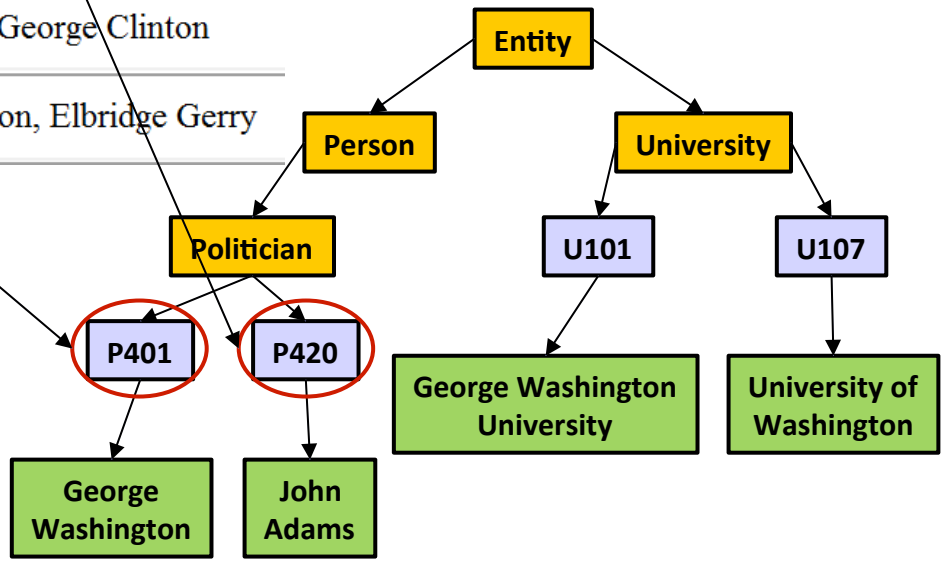


Annotating Web Tables: Using a Catalog

- 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?

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

US President	US Vice President
P401	P420
P420	P471

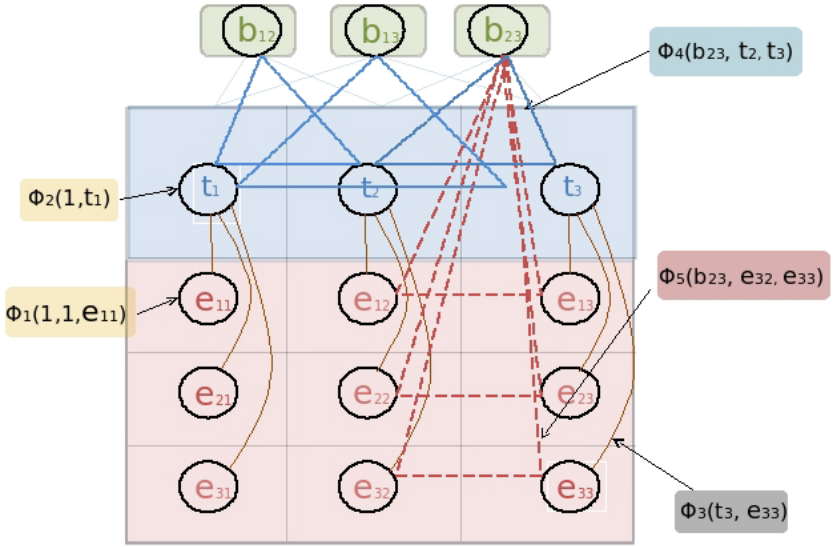


Annotating Web Tables: Using a Catalog

- 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)

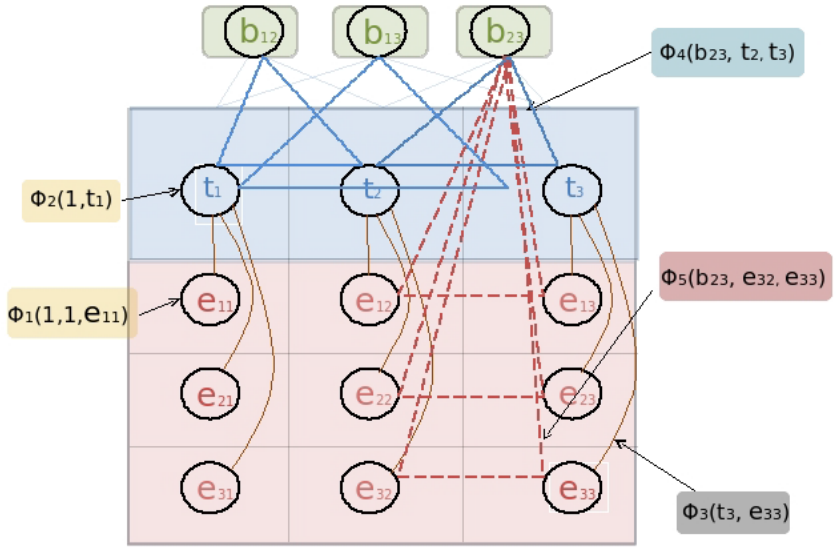
Annotating Web Tables: Using a Catalog

- 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)	John C. Breckinridge (1857-1861)
	Abraham Lincoln (1861-1865)	Hannibal Hamlin (1861-1865)
		Andrew Johnson (1865)
James Monroe (1817-1825)	Andrew Johnson (1865-1869)	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

X

- 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

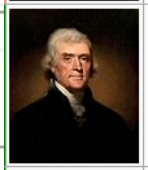
President	James Madison (1809-1817)	George Clinton (1809-1812)
George Washington (1789-1797)		none (1812-1813)
John Adams (1797-1801)		Elbridge Gerry (1813-1814)
Thomas Jefferson (1801-1809)		none (1814-1817)
	James Monroe (1817-1825)	Daniel D. Tompkins (1817-1825)
	John Quincy Adams (1825-1829)	John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

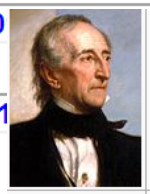
	President	Took office	Left office	Party	Term [n 1]	Previous office
	George Washington (1732–1799) [11][12][13]	April 30, 1789	March 4, 1797	Independent [14]	1 (1789) 2 (1792)	Commander-in-Chief of the Continental Army (1775–1783)
	John Adams (1735–1826) [15][16][17]	March 4, 1797	March 4, 1801 [n 2]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

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 (180	
James Madison (1809-1	
James Monroe (1817-18	
John Quincy Adams (18	
Andrew Jackson (1829-	
Martin Van Buren (1837	
William Henry Harrison	

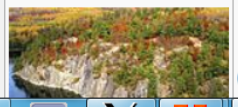
	John Tyler (1790–1862) [39][40][41]	April 4, 1841	March 4, 1845	Whig April 4, 1841 – September 13, 1841; no party[n 7] September 13, 1841 – March 4, 1845	(1840)	Vice President
	James K. Polk (1795–1849) [42][43][44]	March 4, 1845	March 4, 1849	Democratic	15 (1844)	Governor of Tennessee (1839–1841)
	Zachary Taylor (1784–1850) [45][46][47]	March 4, 1849	July 9, 1850 [n 3]	Whig	16 (1848)	U.S. Army Major general (1846–1849)
	Millard Fillmore (1800–1874) [48][49][50]	July 9, 1850	March 4, 1853 [n 8]	Whig		Vice President

Finding Related Tables

X

- 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 (179	
Thomas Jefferso	
James Madison (	
James Monroe (1	
John Quincy Ada	
Andrew Jackson	
Martin Van Buren	
William Henry Har	

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]
Theodore Roosevelt		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]
Virgin Islands		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 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]
Voyageurs		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

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)	
James Madison (1809-1817)	
James Monroe (1817-1825)	
John Quincy Adams (1825-1829)	
Andrew Jackson (1829-1837)	
Martin Van Buren (1837-1841)	
William Henry Harrison (1841)	

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

John Tyler (1841-1845)	none (1841-1845)
James K. Polk (1845-1849)	George M. Dallas (1845-1849)
Zachary Taylor (1849-1850)	Millard Fillmore (1849-1850)
Millard Fillmore (1850-1853)	none (1850-1853)
Franklin Pierce (1853-1857)	William King (1853)
	none (1853-1857)
James Buchanan (1857-1861)	John C. Breckinridge (1857-1861)
Abraham Lincoln (1861-1865)	Hannibal Hamlin (1861-1865)
	Andrew Johnson (1865)
Andrew Johnson (1865-1869)	none (1865-1869)
Ulysses S. Grant (1869-1877)	Schuyler Colfax (1869-1873)
	Henry Wilson (1873-1875)
	none (1875-1877)
Rutherford B. Hayes (1877-1881)	William Wheeler (1877-1881)
James A. Garfield (1881)	Chester Arthur (1881)
Chester Arthur (1881-1885)	none (1881-1885)
Grover Cleveland (1885-1889)	Thomas Hendricks (1885)
	none (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

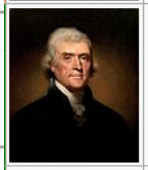
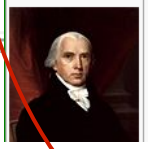
- Methodology
  - Coverage of entity set: T' should contain most of T's entities
  - Coherent schema expansion: use the ACSDb to measure the maximum benefit that a subset of attributes of T' can provide to T

- 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

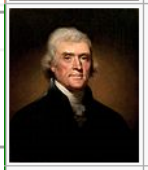
President	Vice President
George Washington (1789-1797)	John Adams (1789-1797)
John Adams (1797-1801)	
Thomas Jefferson (1801-1809)	
James Madison (1809-1817)	
James Monroe (1817-1825)	
John Quincy Adams (1825-1829)	
Andrew Jackson (1829-1837)	
Martin Van Buren (1837-1841)	
William Henry Harrison (1841)	

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

Finding Related Tables: Schema Complement

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

President	Vice President
George Washington (1789-1797)	John Adams (1789-1797)
John Adams (1797-1801)	
Thomas Jefferson (1801-1809)	
James Madison (1809-1817)	
James Monroe (1817-1825)	
John Quincy Adams (1825-1829	
Andrew Jackson (1829-1837)	
Martin Van Buren (1837-1841)	
William Henry Harrison (1841)	

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

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

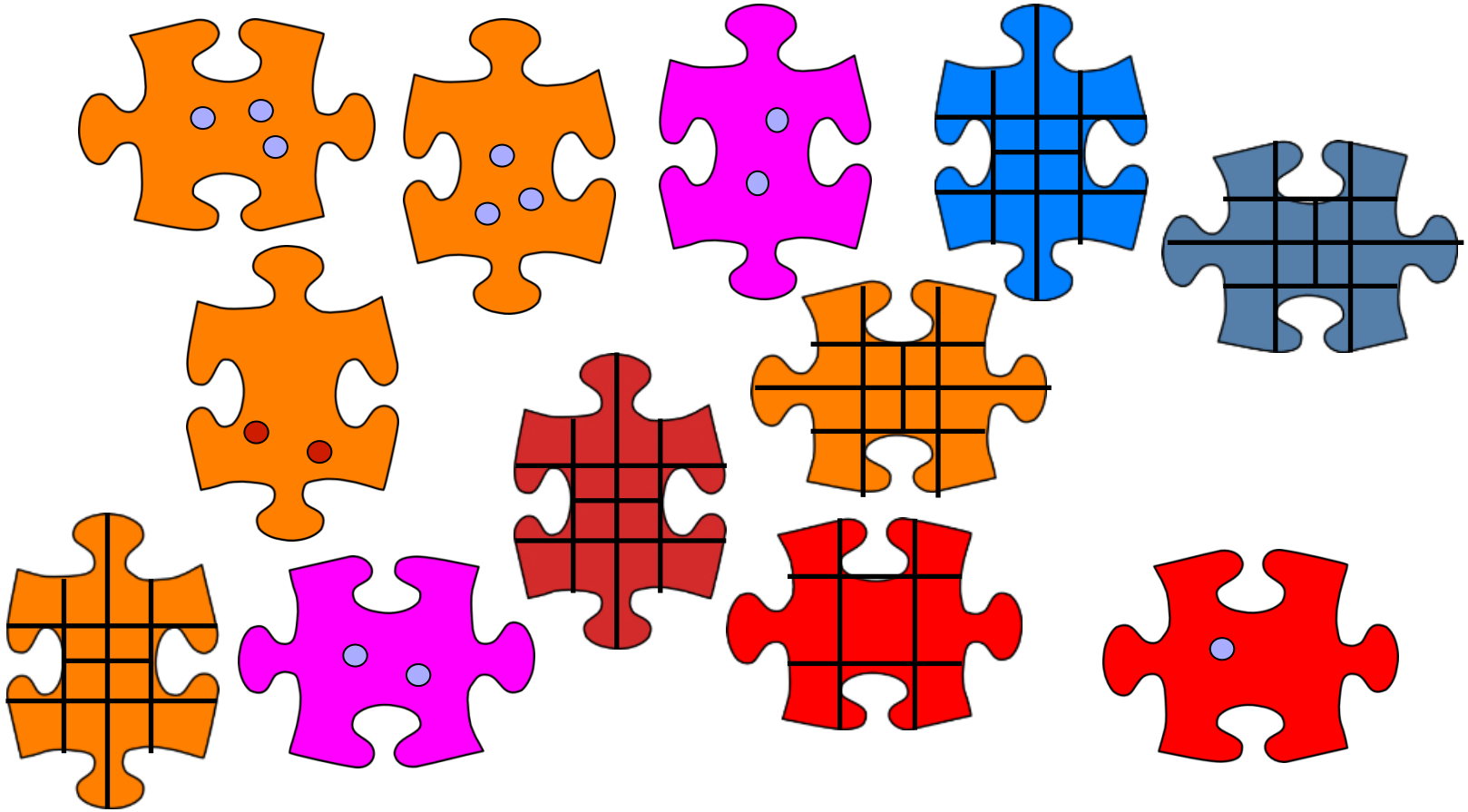
- Useful filters:
  - Two tables must share entity column name or inferred label
  - Two tables must share at least n entities, n = 1, 2, 3

Outline

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

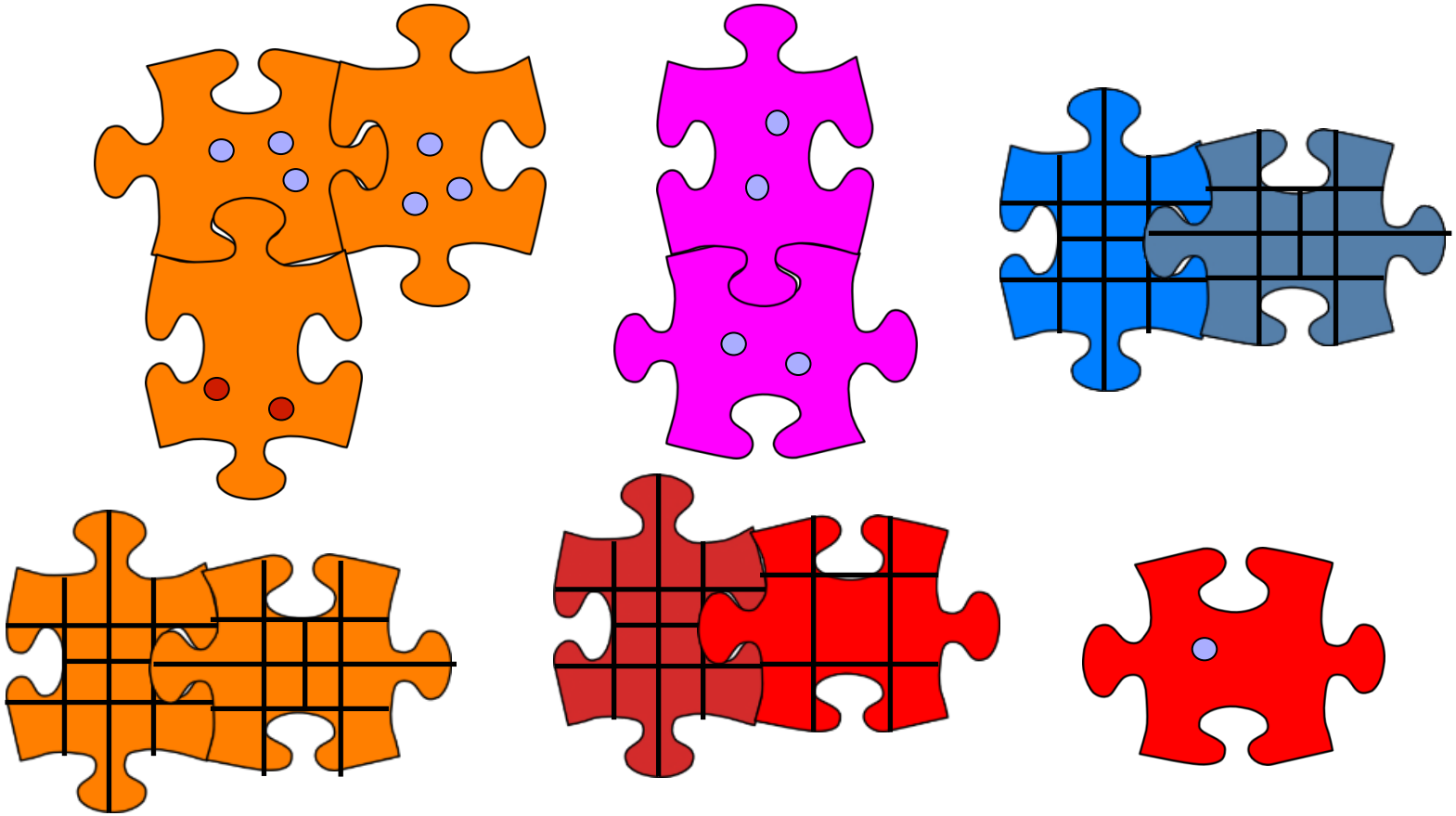
Record Linkage

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



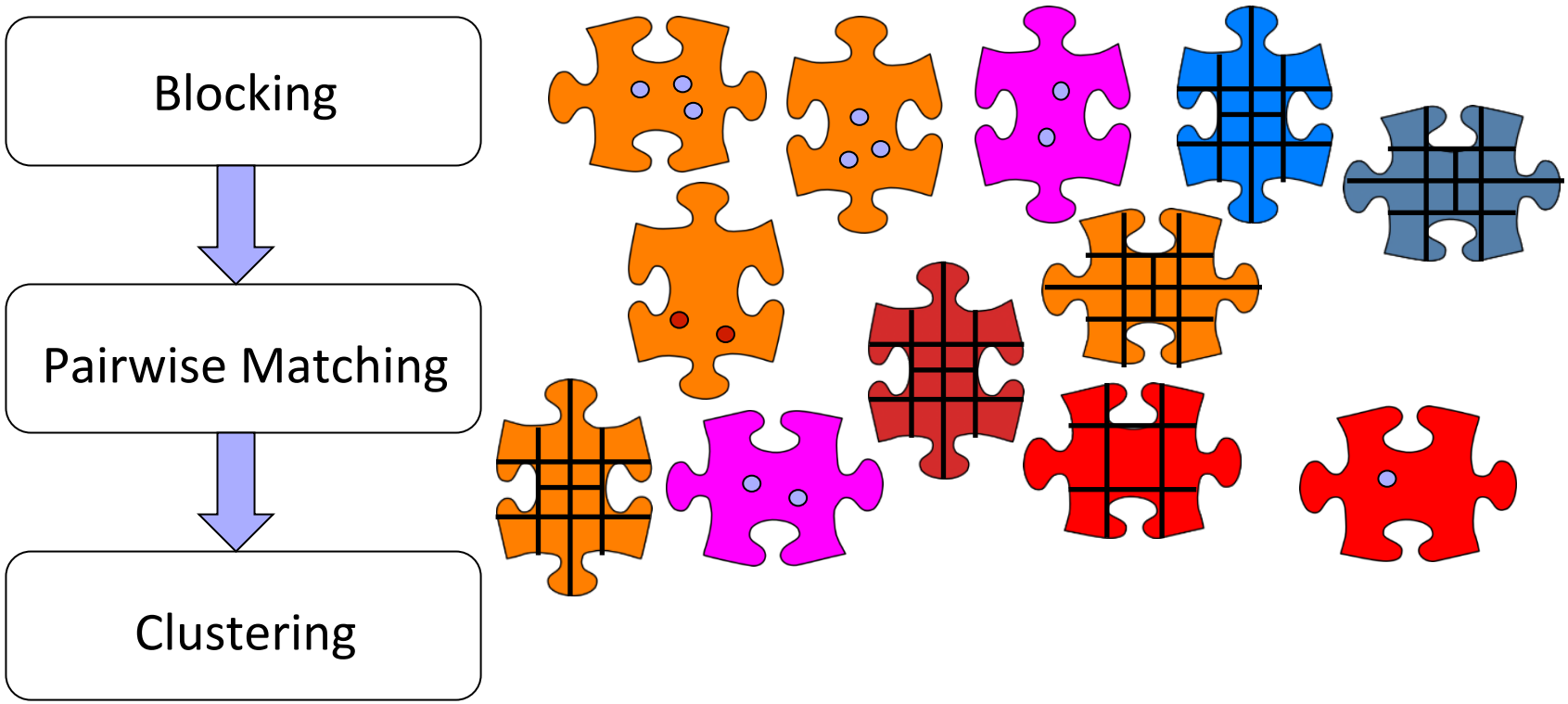
Record Linkage

- Matching based on identifying content: color, pattern



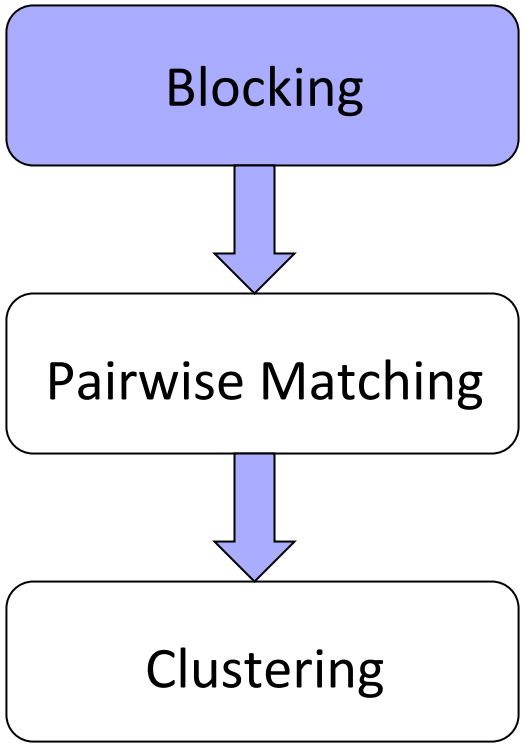
Record Linkage: Three Steps [EIV07, GM12]

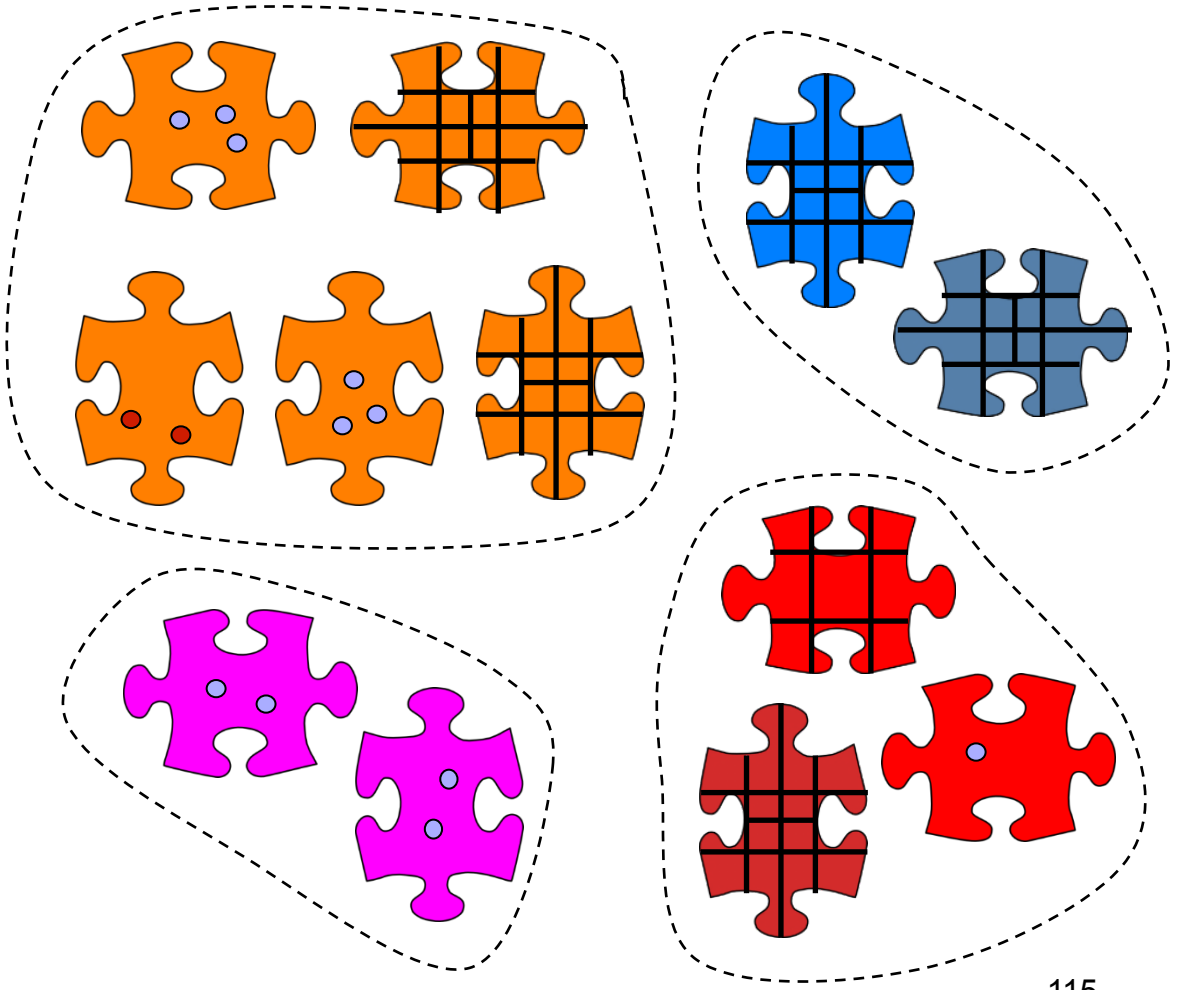
- Record linkage: blocking + pairwise matching + clustering
  - Scalability, similarity, semantics



Record Linkage: Three Steps

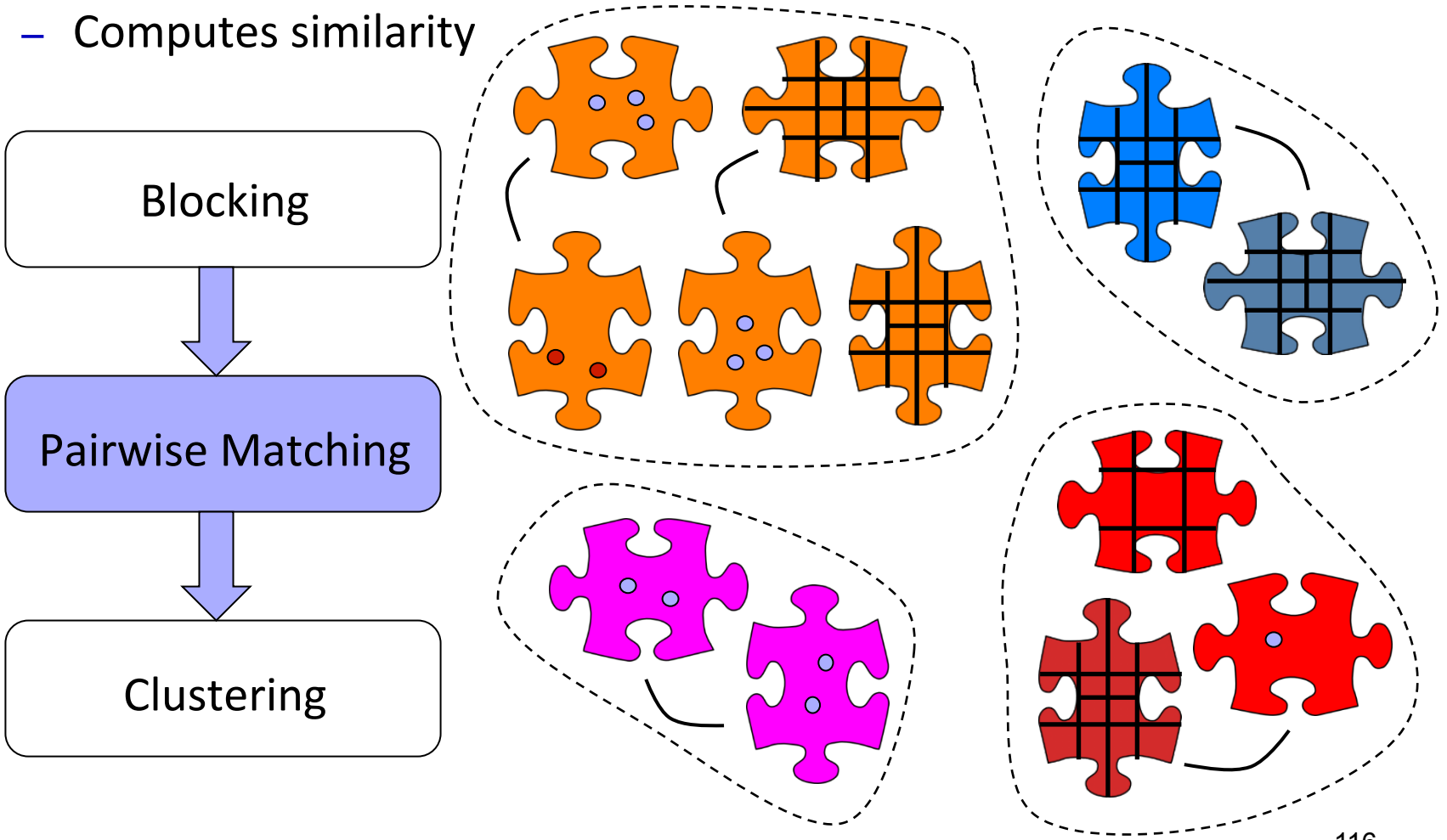
- Blocking: **efficiently** create **small** blocks of **similar** records
  - Ensures scalability





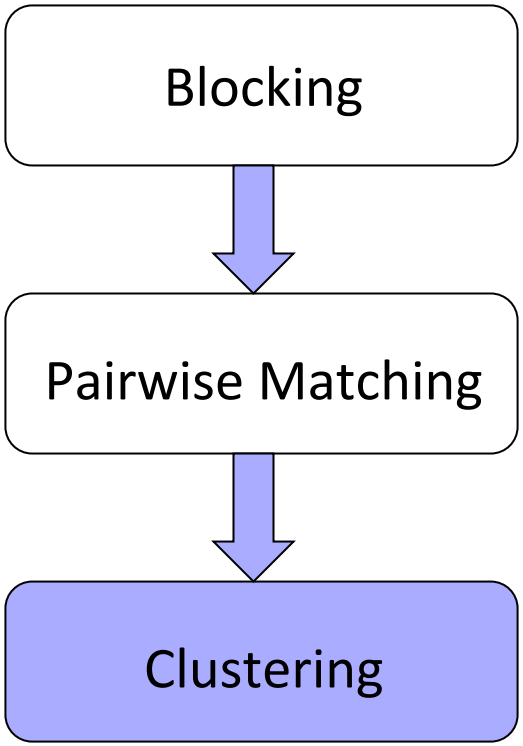
Record Linkage: Three Steps

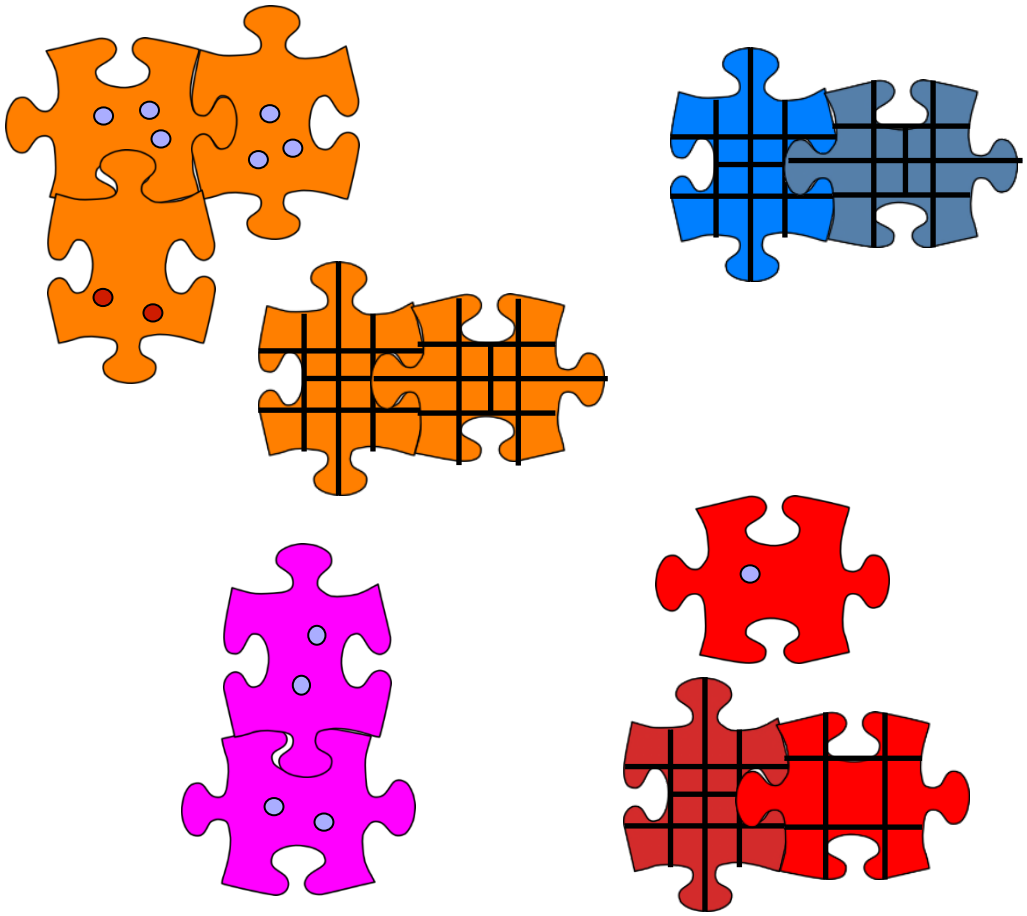
- Pairwise matching: compares all record pairs in a block
  - Computes similarity



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]

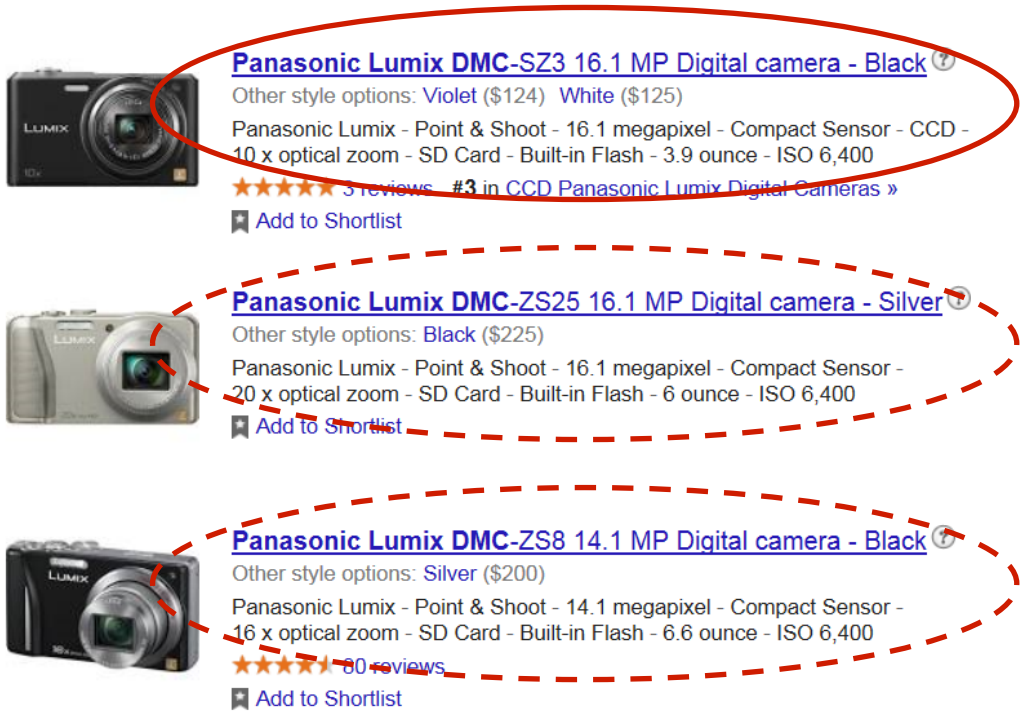
BDI: Record Linkage

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

- **Veracity**
  - Linking temporal records [LDM+11]

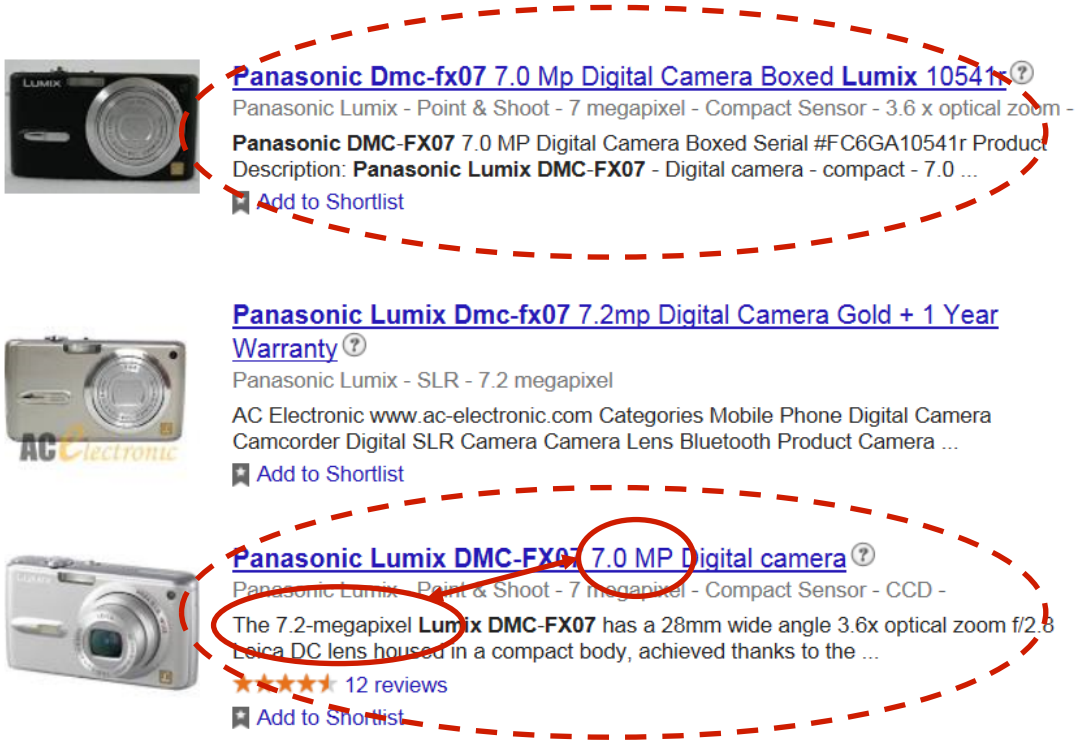
Matching with Unstructured Data

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



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



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 ...
★★★★½ 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

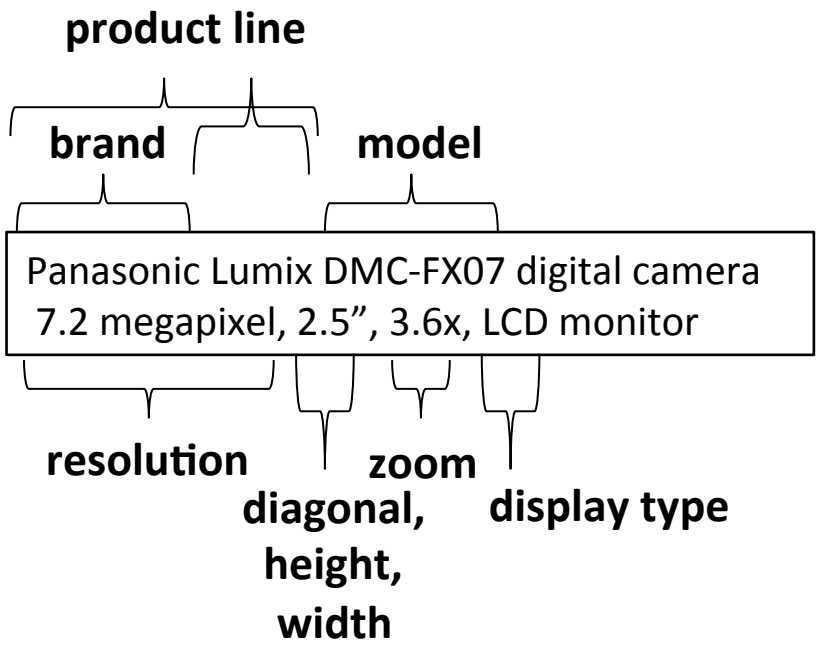
Panasonic DMC-FX07EB digital camera silver

Lumix FX07EB-S, 7.2MP

Structured + Unstructured Data

- Key idea: optimal parse of (unstructured) offer wrt specification

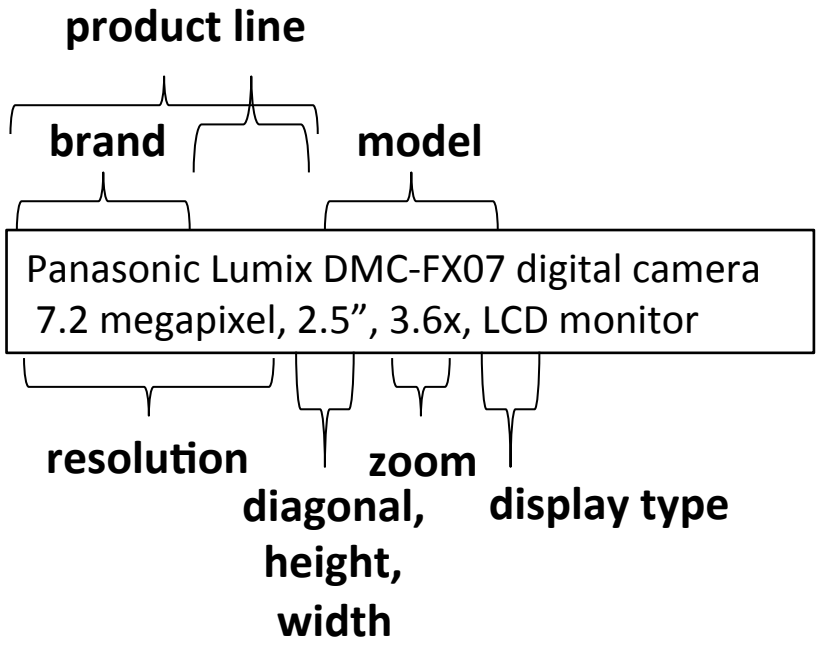
- Semantic parse of offers: tagging
  - Use inverted index built on specification values
  - Tag all n-grams



Structured + Unstructured Data

- Key idea: optimal parse of (unstructured) offer wrt specification

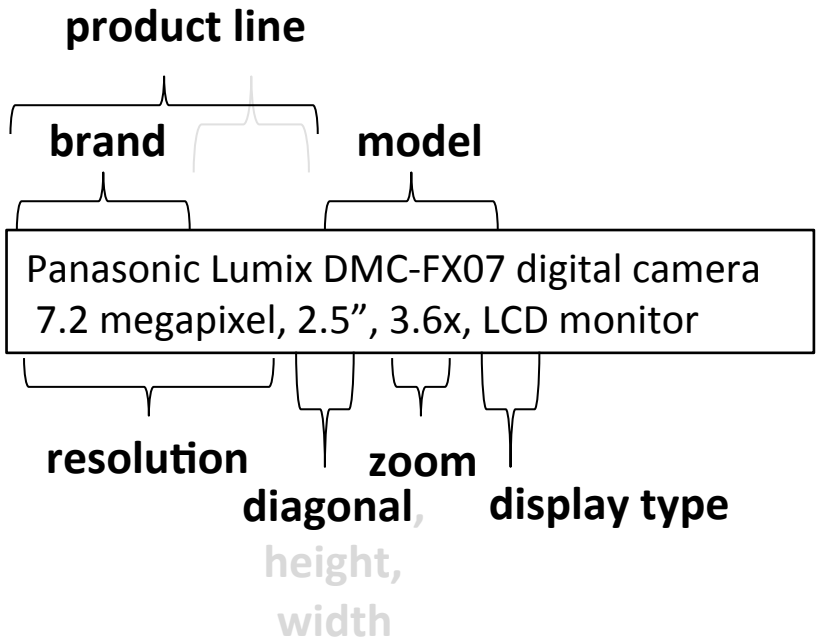
- Semantic parse of offers: tagging, plausible parse
  - Combination of tags such that each attribute has distinct value



Structured + Unstructured Data

- Key idea: optimal parse of (unstructured) offer wrt specification

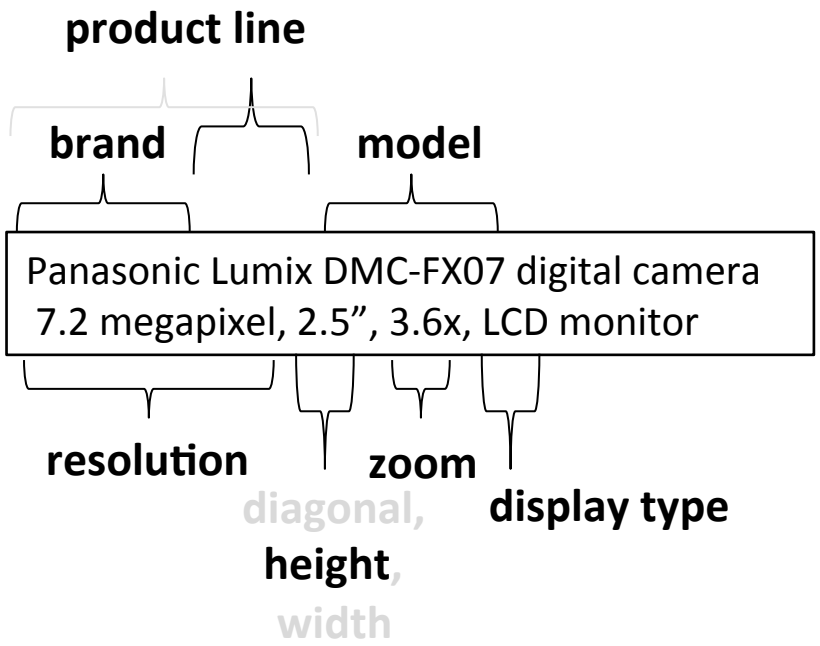
- Semantic parse of offers: tagging, plausible parse
  - Combination of tags such that each attribute has distinct value



Structured + Unstructured Data

- Key idea: optimal parse of (unstructured) offer wrt specification

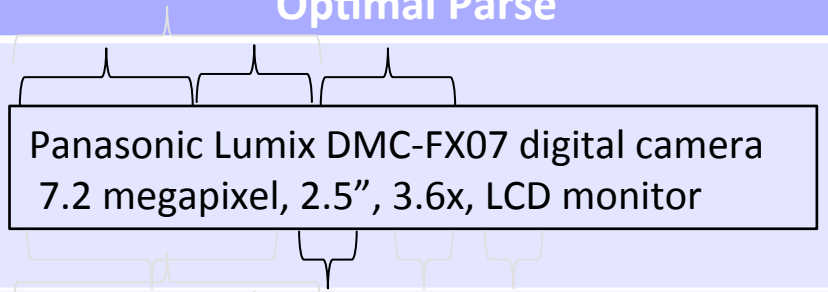
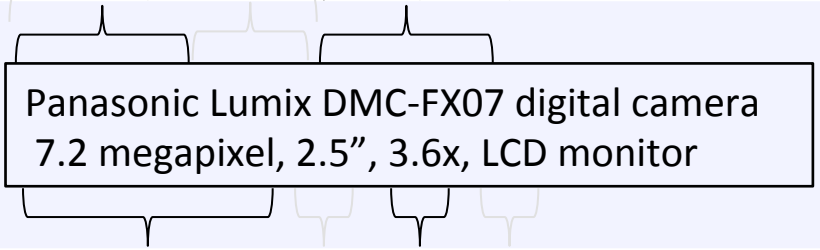
- Semantic parse of offers: tagging, plausible parse
  - Combination of tags such that each attribute has distinct value
  - # depends on ambiguities



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

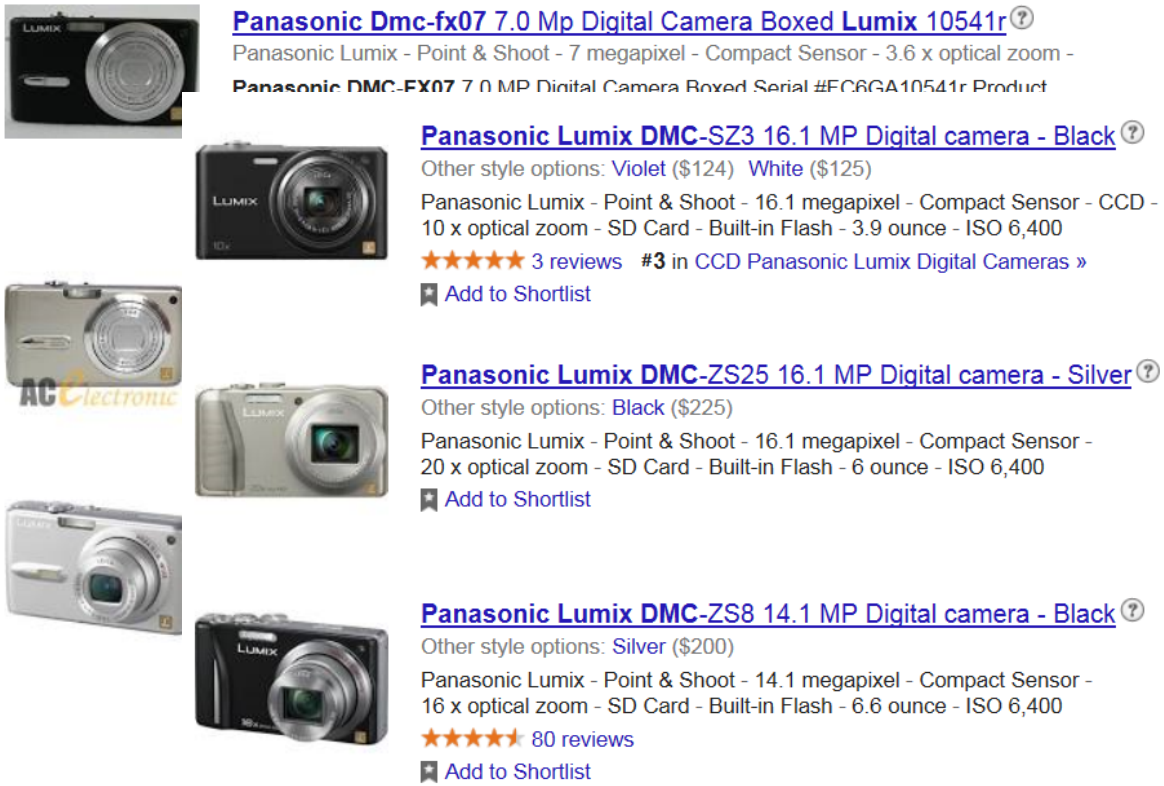
Product specification		Optimal Parse
brand product line model diagonal	Panasonic Lumix DMC-FX05 2.5 in	
brand model resolution zoom	Panasonic DMC-FX07 7.2 megapixel 3.6x	

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

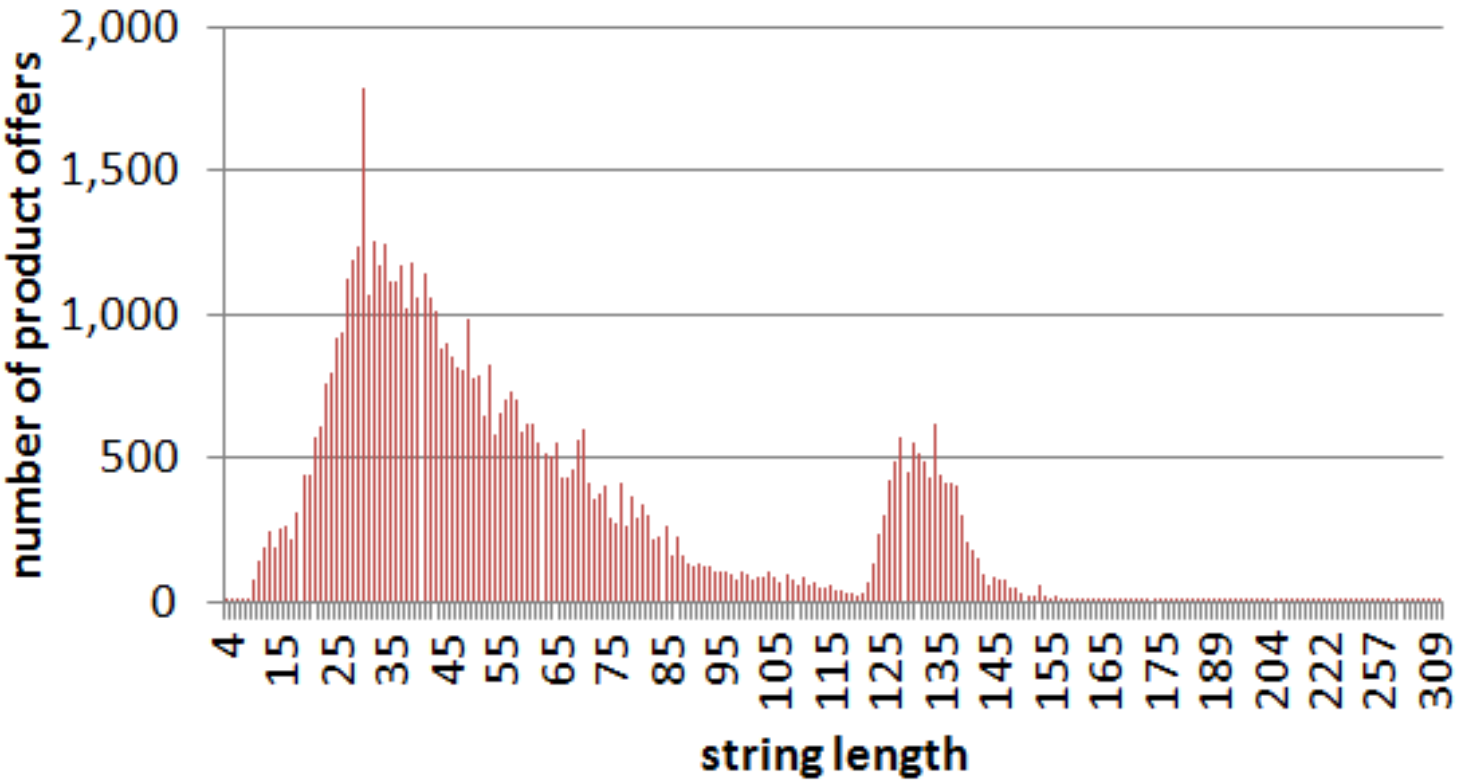
Matching Unstructured Data [KTT+12]

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



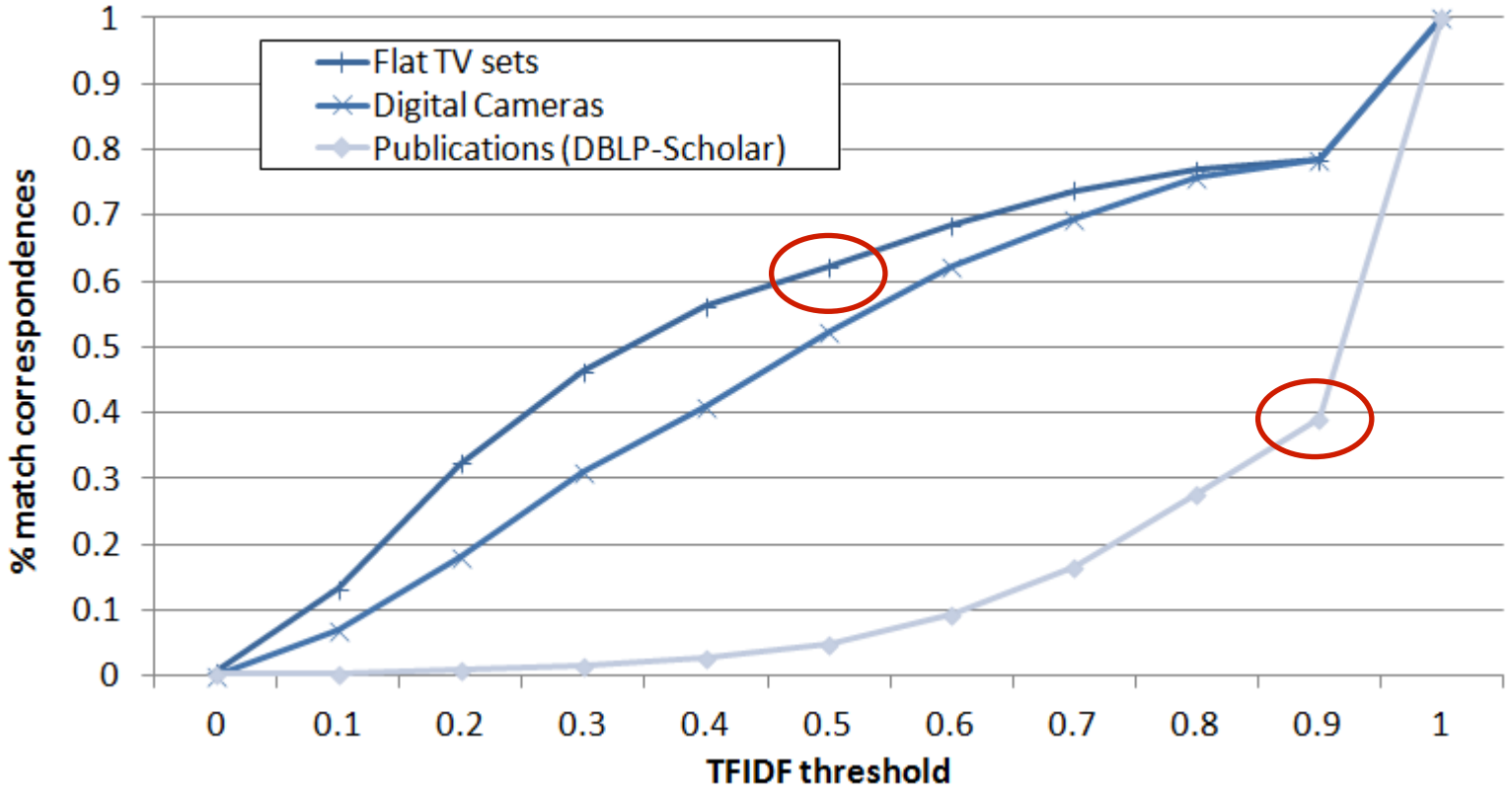
Matching Unstructured Data

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



Matching Unstructured Data

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



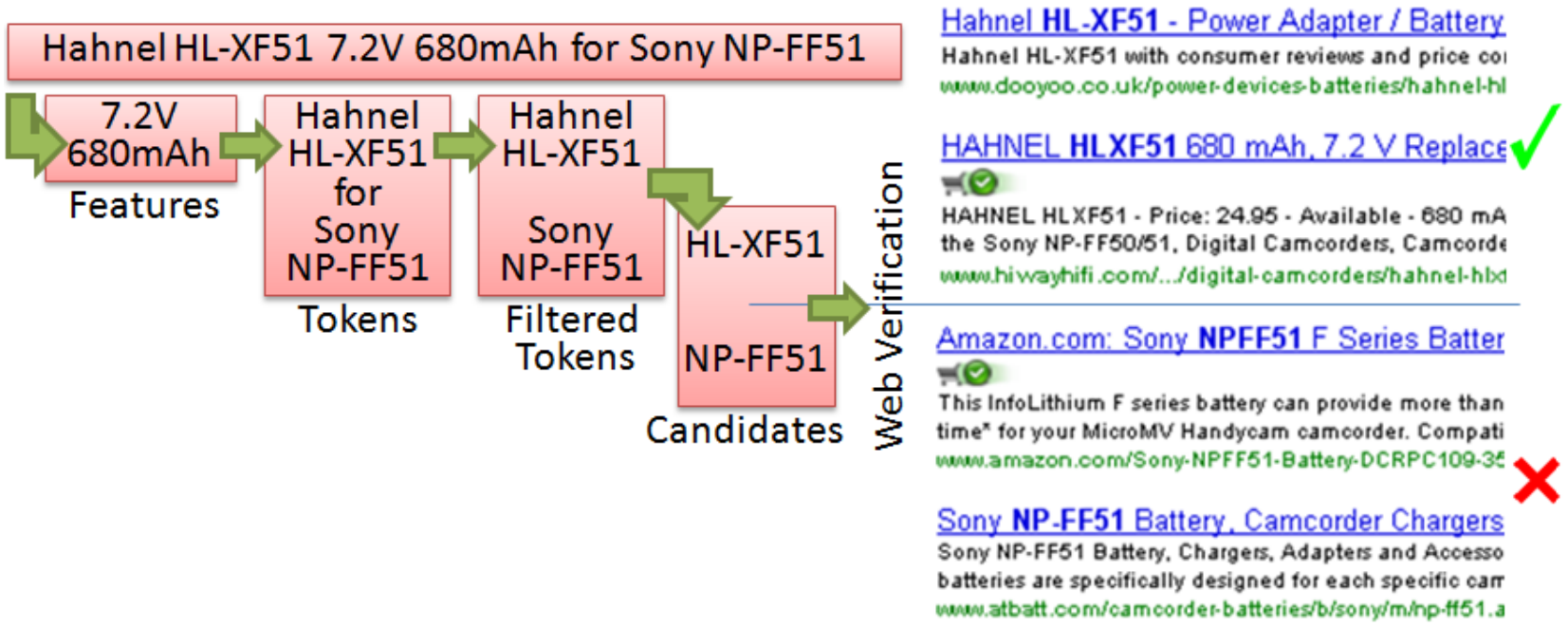
Matching Unstructured Data

- 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



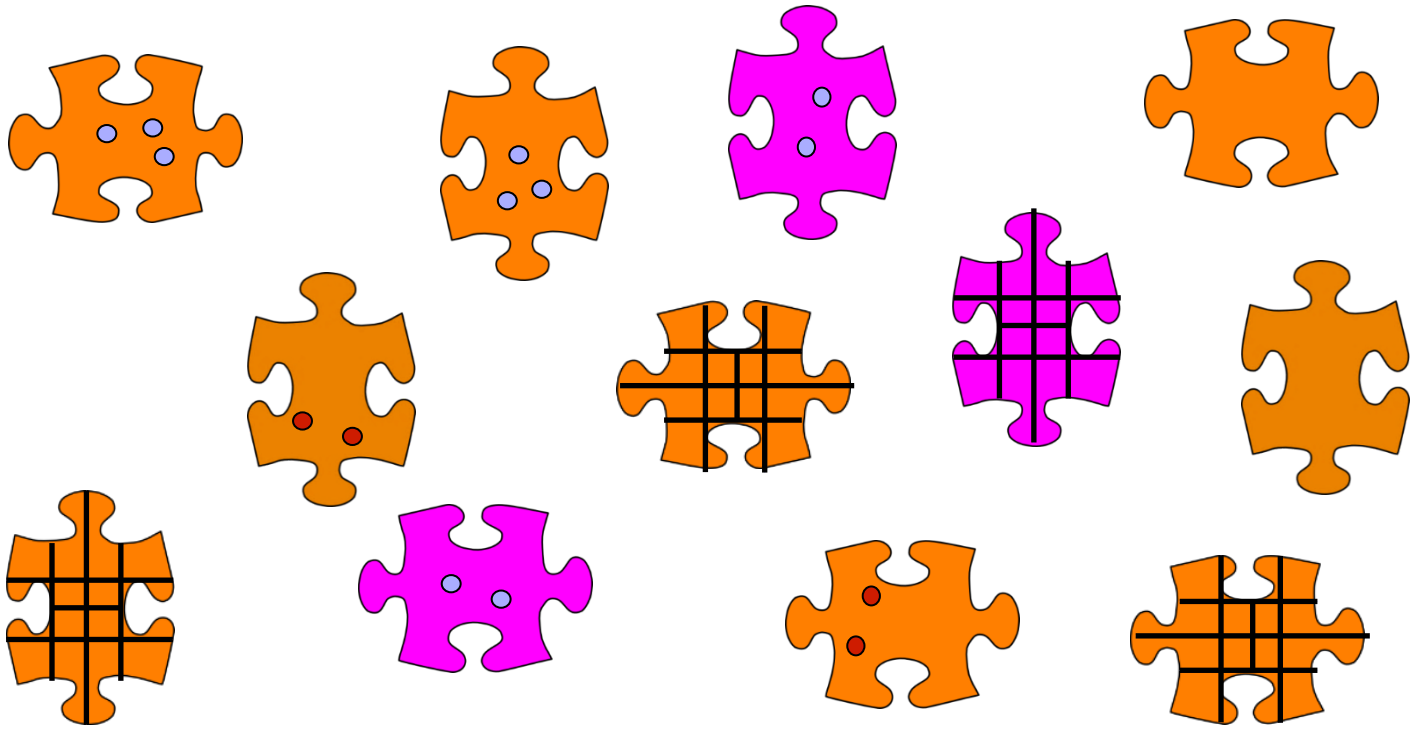
Record Linkage Using MapReduce [KTR12]

- 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

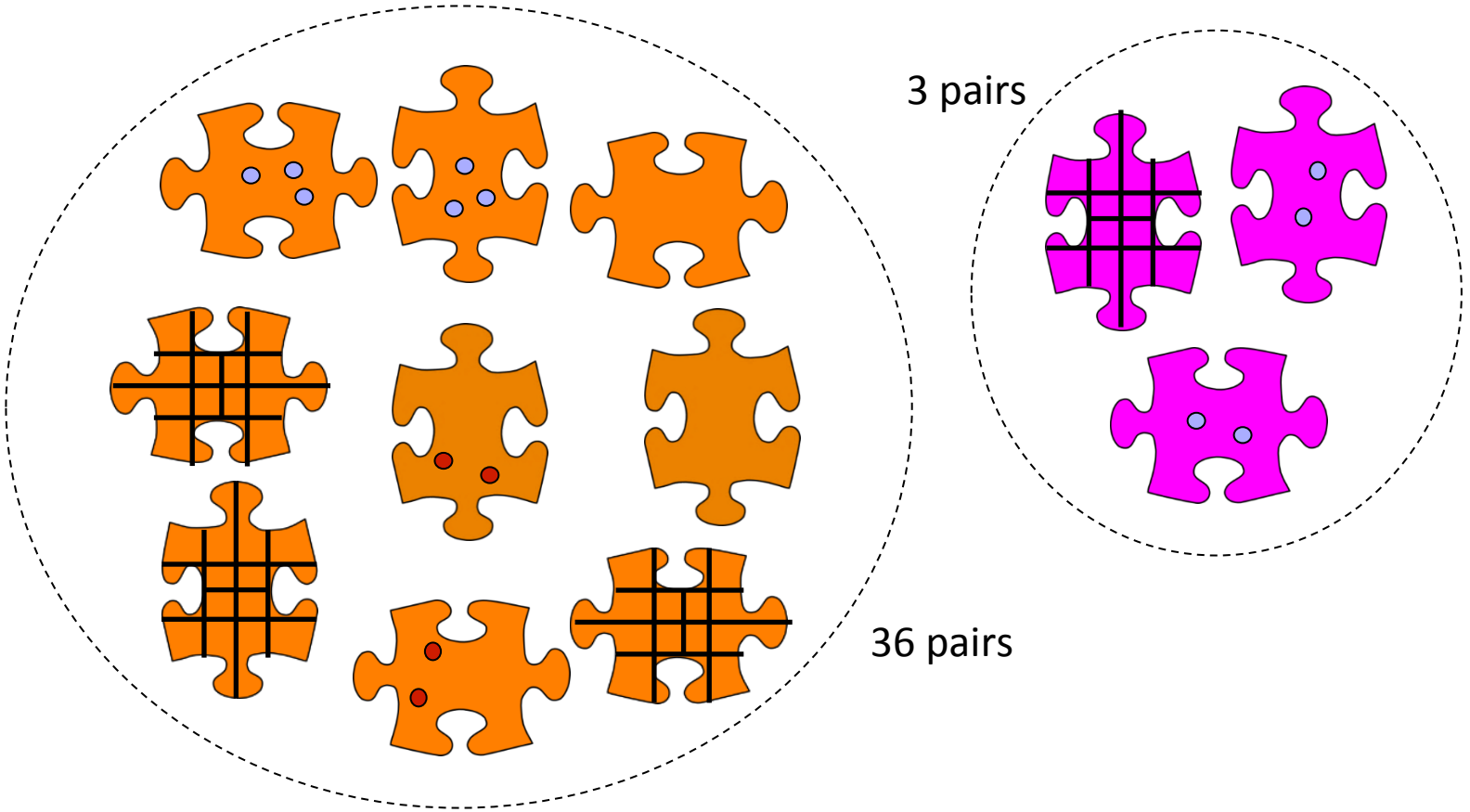
Record Linkage Using MapReduce

- Challenge: data skew → unbalanced workload



Record Linkage Using MapReduce

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



Load Balancing

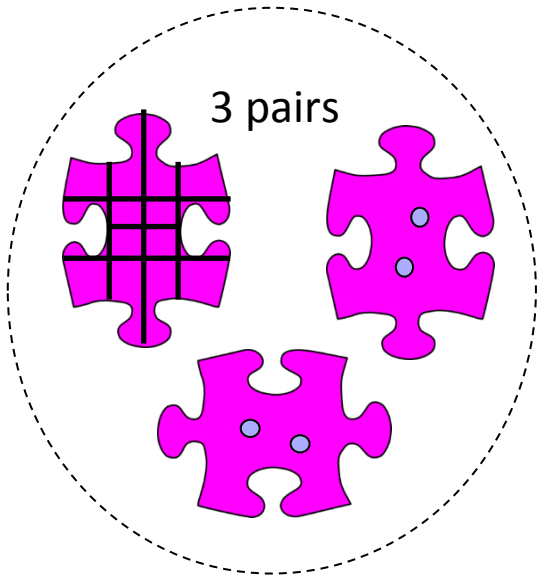
- Challenge: data skew → unbalanced workload
  - Difficult to tune blocking function to get balanced workload

- 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

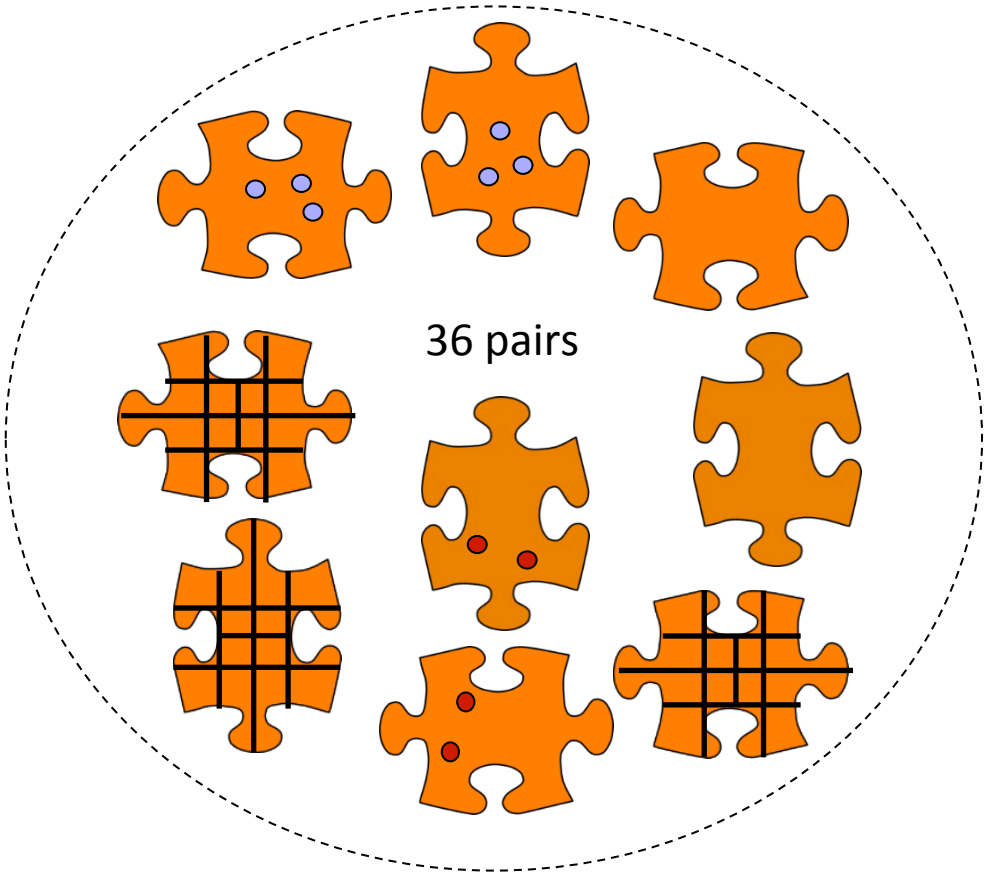
Load Balancing: BlockSplit

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



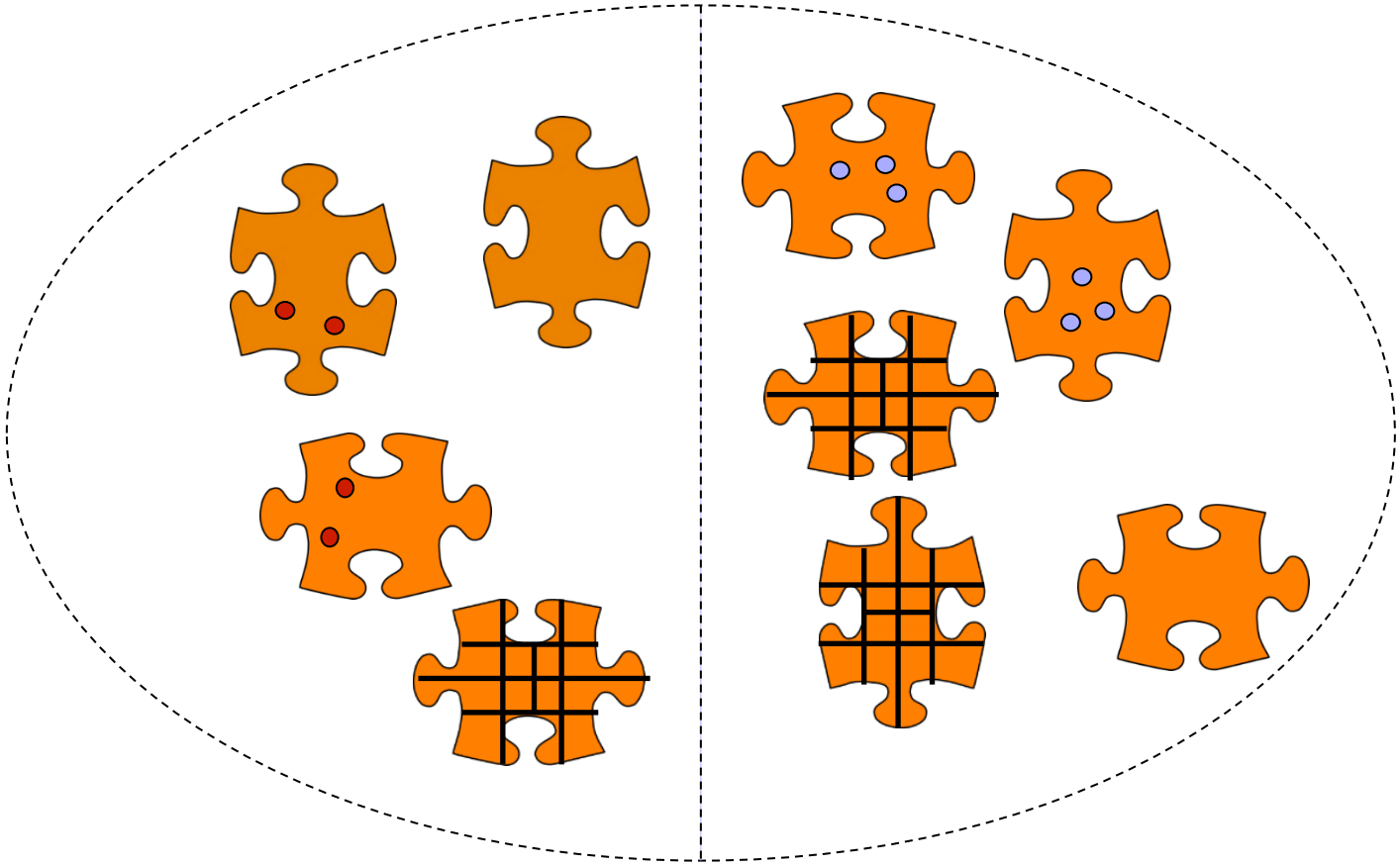
Load Balancing: BlockSplit

- Large blocks: split randomly into multiple sub-blocks



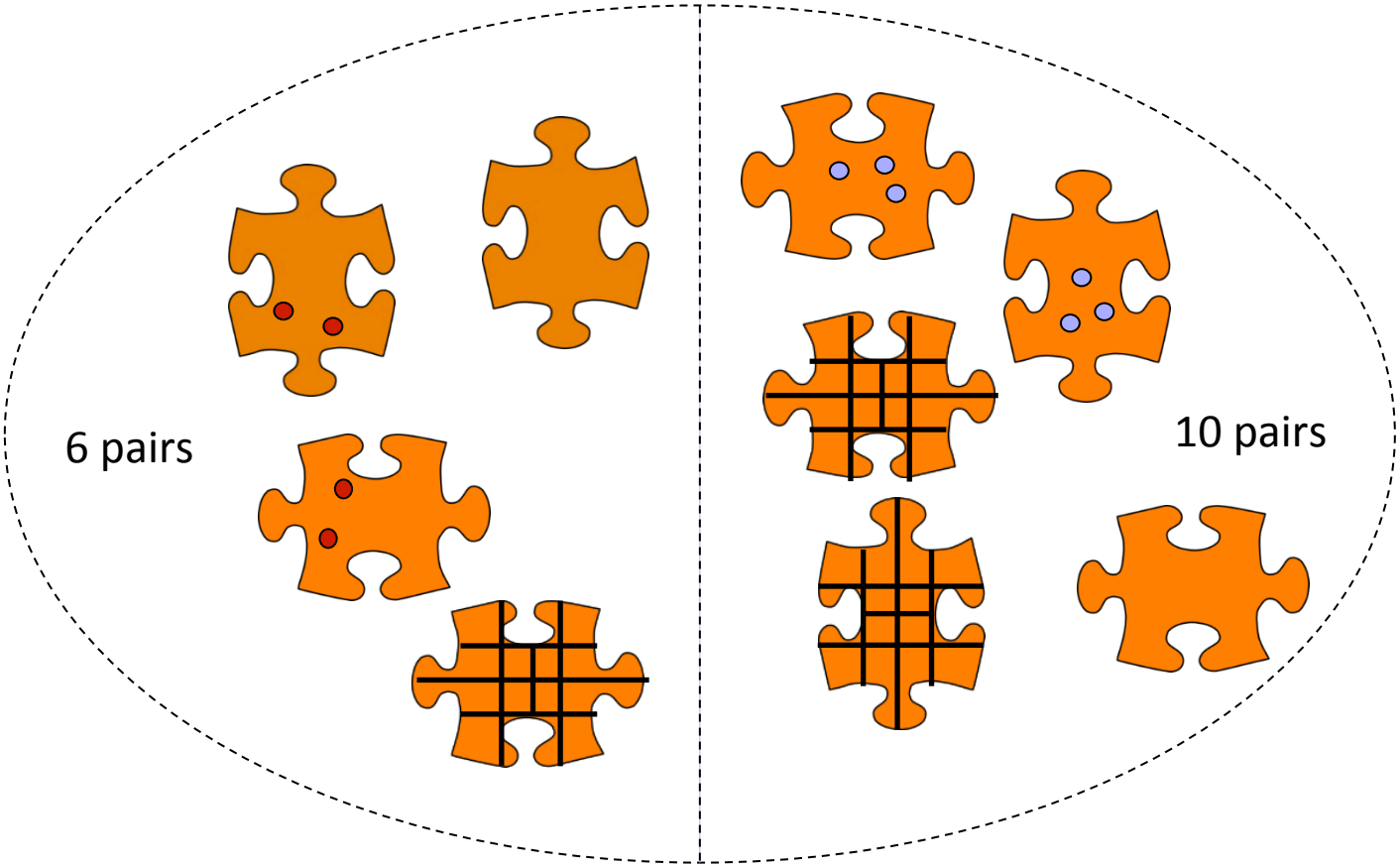
Load Balancing: BlockSplit

- Large blocks: split randomly into multiple sub-blocks



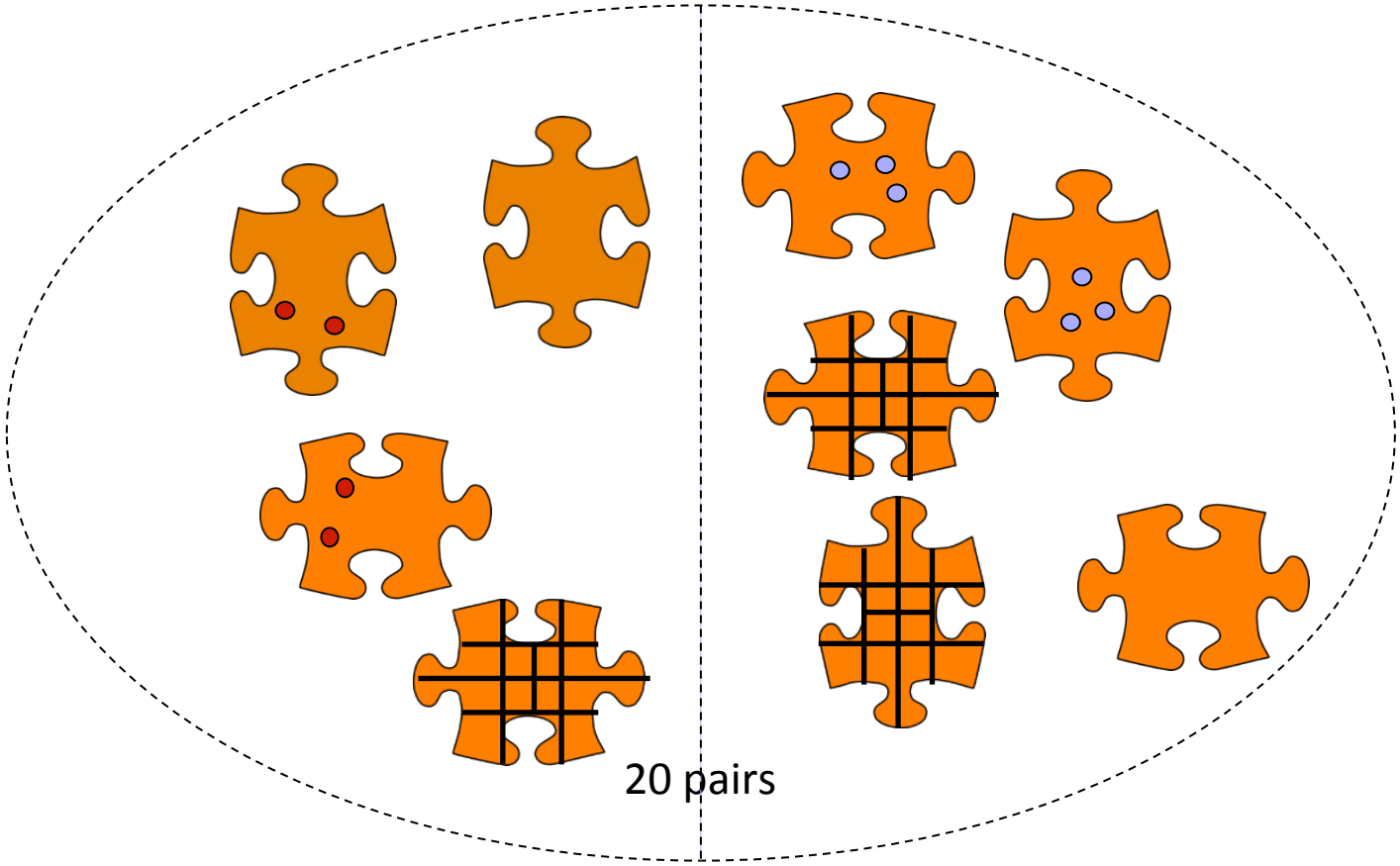
Load Balancing: BlockSplit

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



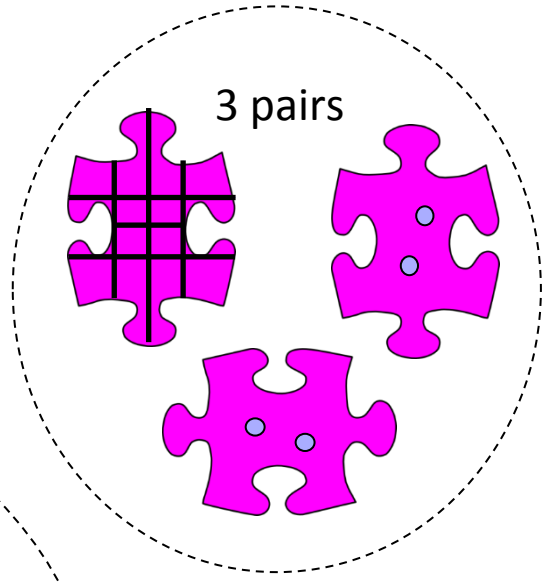
Load Balancing: BlockSplit

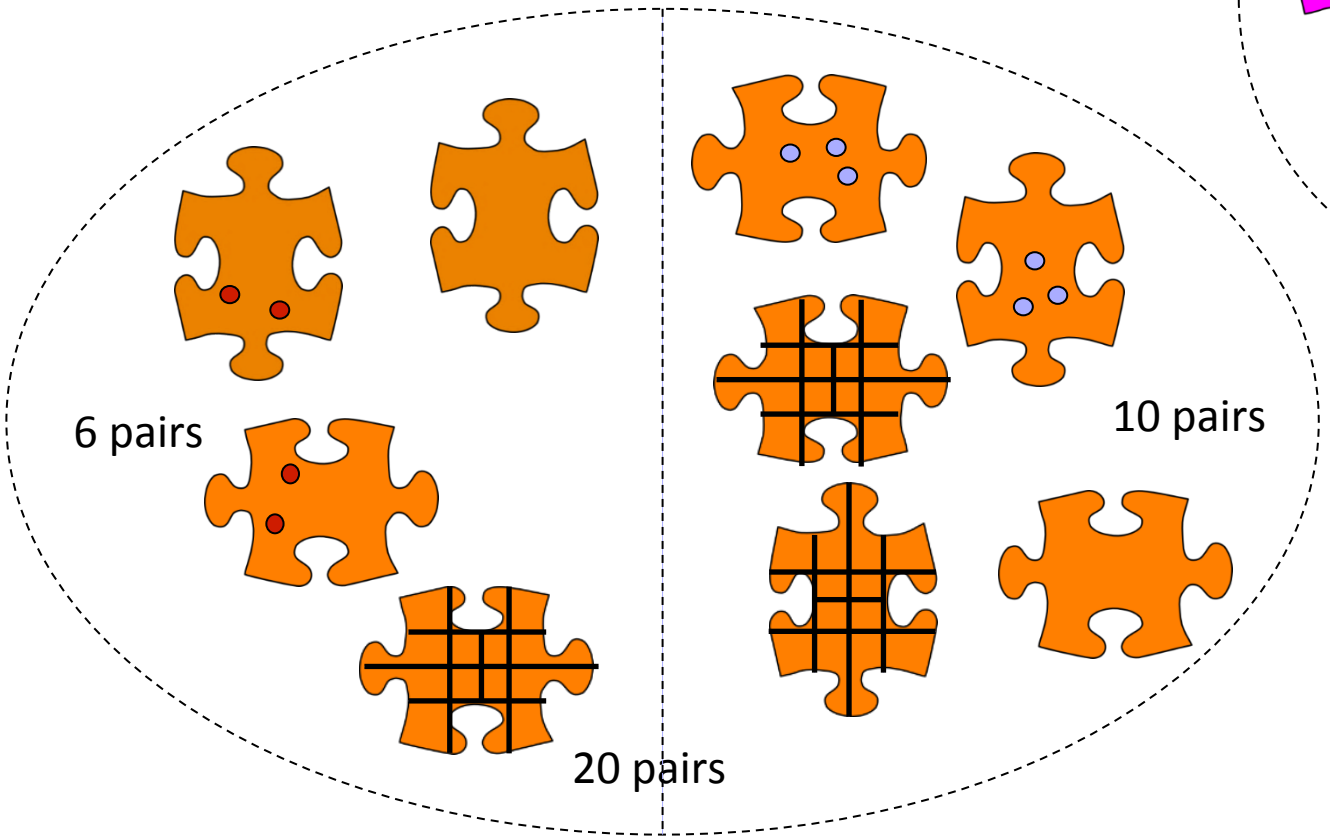
- Large blocks: split into multiple sub-blocks
  - Pair of sub-blocks is processed by "cartesian product" match task



Load Balancing: BlockSplit

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





Improving Blocking Recall

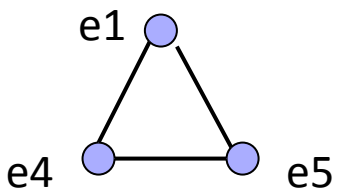
- Observation: a single, pairwise-disjoint blocking has poor recall

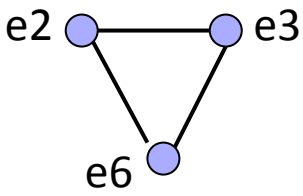
Id	First Name	Last Name	YOB	State	Elected
e1	Bob	Menendez	1954	New Jersey	2006
e2	Bob	Casey	1960	Pennsylvania	2007
e3	Robert	Casey, Jr	1960	Pennsylvania	2007
e4	Robert	Menendez	1954	NJ	2006
e5	Robert	Menendez	1954	New Jersey	2007
e6	Bob	Casey, Jr	1961	Pennsylvania	2007

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

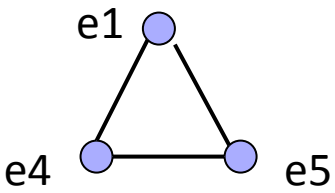
- Also block by Last Name: {{e1, e4, e5}, {e2}, {e3, e6}}

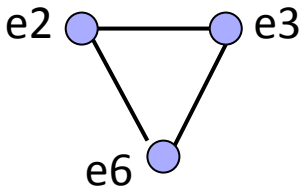
Improving Blocking Recall

- 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

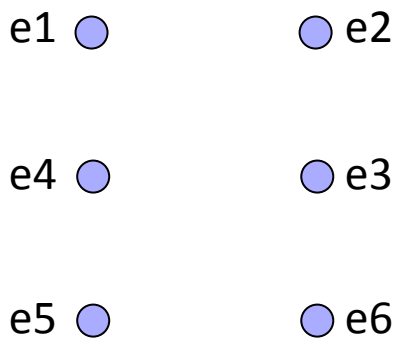
Meta-Blocking: Improve Blocking Efficiency

- 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

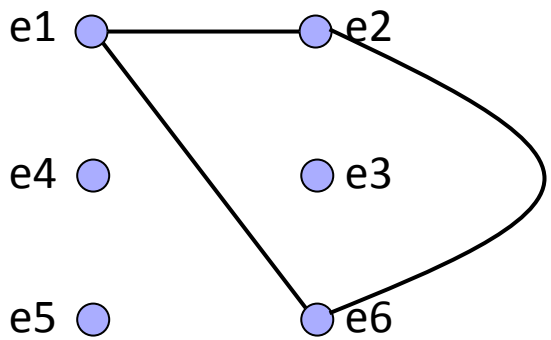
e1	Menendez	New Jersey	2006	Bob	1954
e2	Bob	Casey	1960	Pennsylvania	2007
e3	Robert	Casey, Jr	1960	Pennsylvania	2007
e4	Menendez	NJ	1954	Robert	2006
e5	Menendez	New Jersey	1954	Robert	2007
e6	Bob	Casey, Jr	2007	Pennsylvania	1961



Meta-Blocking: Improve Blocking Efficiency

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

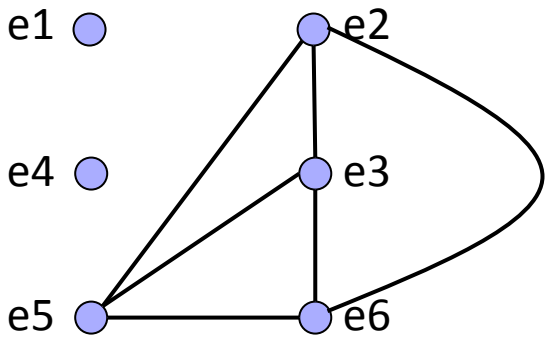
e1	Menendez	New Jersey	2006	Bob	1954
e2	Bob	Casey	1960	Pennsylvania	2007
e3	Robert	Casey, Jr	1960	Pennsylvania	2007
e4	Menendez	NJ	1954	Robert	2006
e5	Menendez	New Jersey	1954	Robert	2007
e6	Bob	Casey, Jr	2007	Pennsylvania	1961



Meta-Blocking: Improve Blocking Efficiency

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

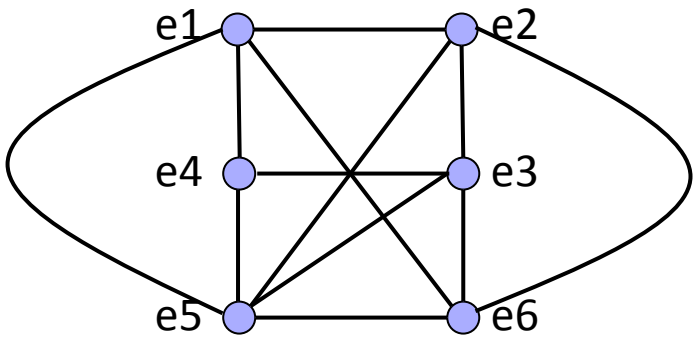
e1	Menendez	New Jersey	2006	Bob	1954
e2	Bob	Casey	1960	Pennsylvania	2007
e3	Robert	Casey, Jr	1960	Pennsylvania	2007
e4	Menendez	NJ	1954	Robert	2006
e5	Menendez	New Jersey	1954	Robert	2007
e6	Bob	Casey, Jr	2007	Pennsylvania	1961



Meta-Blocking: Improve Blocking Efficiency

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

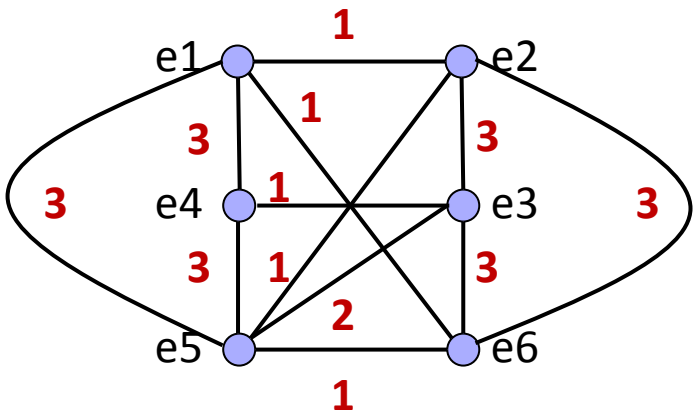
e1	Menendez	New Jersey	2006	Bob	1954
e2	Bob	Casey	1960	Pennsylvania	2007
e3	Robert	Casey, Jr	1960	Pennsylvania	2007
e4	Menendez	NJ	1954	Robert	2006
e5	Menendez	New Jersey	1954	Robert	2007
e6	Bob	Casey, Jr	2007	Pennsylvania	1961



Meta-Blocking: Improve Blocking Efficiency

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

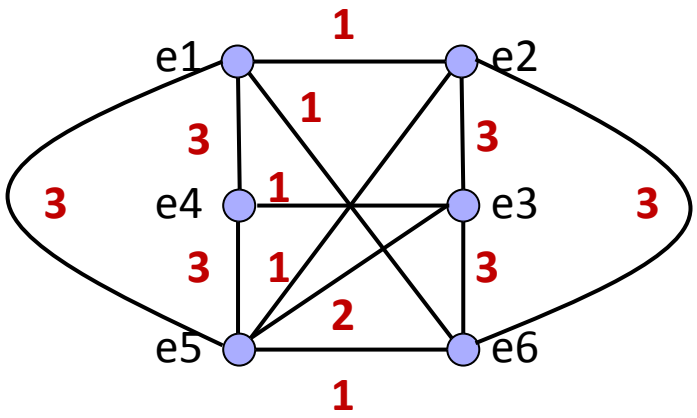
e1	Menendez	New Jersey	2006	Bob	1954
e2	Bob	Casey	1960	Pennsylvania	2007
e3	Robert	Casey, Jr	1960	Pennsylvania	2007
e4	Menendez	NJ	1954	Robert	2006
e5	Menendez	New Jersey	1954	Robert	2007
e6	Bob	Casey, Jr	2007	Pennsylvania	1961



Meta-Blocking: Improve Blocking Efficiency

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

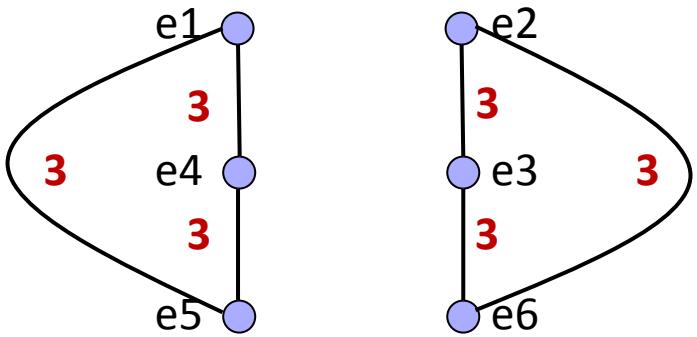
e1	Menendez	New Jersey	2006	Bob	1954
e2	Bob	Casey	1960	Pennsylvania	2007
e3	Robert	Casey, Jr	1960	Pennsylvania	2007
e4	Menendez	NJ	1954	Robert	2006
e5	Menendez	New Jersey	1954	Robert	2007
e6	Bob	Casey, Jr	2007	Pennsylvania	1961



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

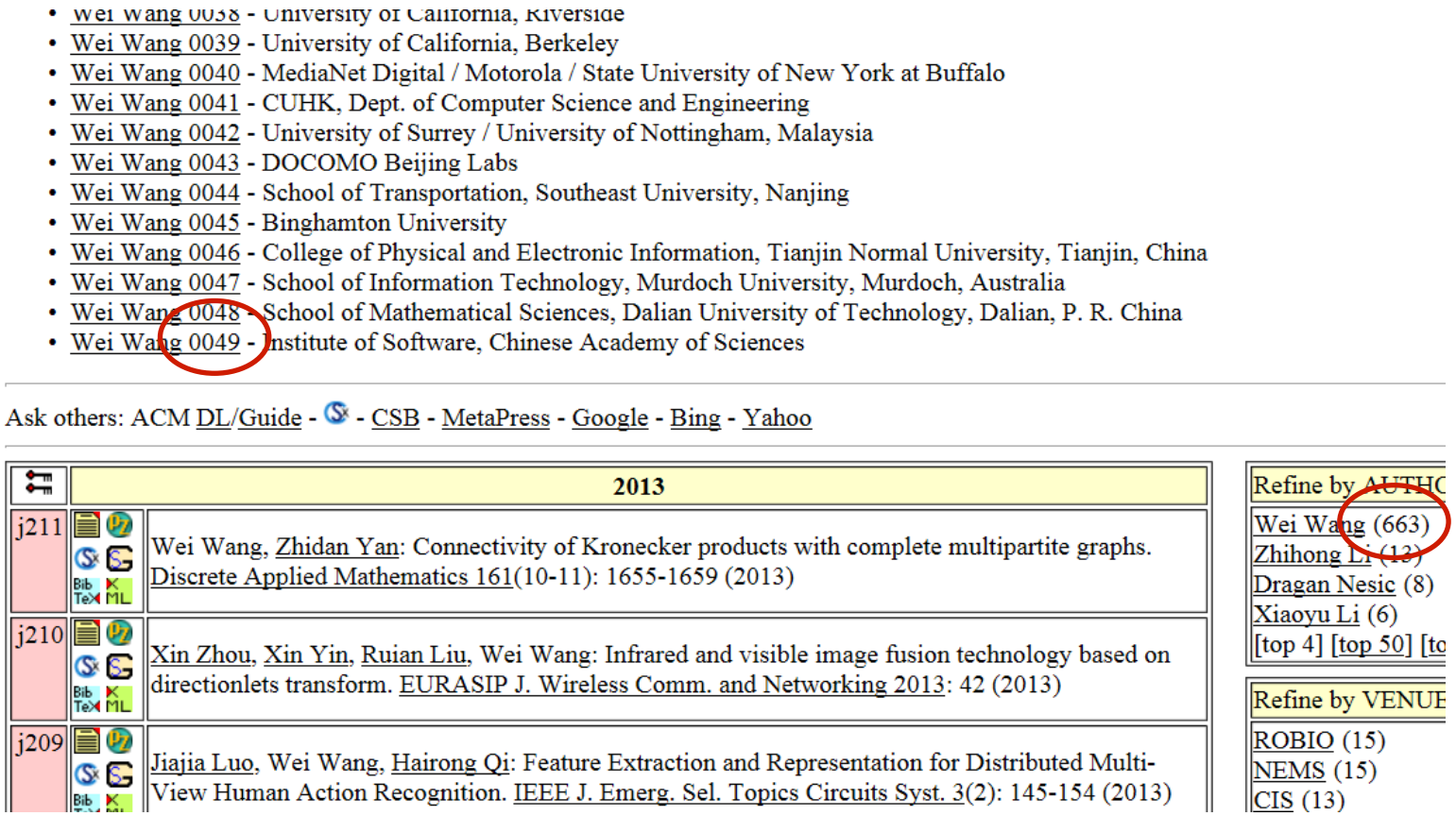
Meta-Blocking: Improve Blocking Efficiency

- Why does meta-blocking improve blocking efficiency?
  - Cost of graph construction = O (# 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

Linking Temporal Records [LDM+11]

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



- Wei Wang 0038 - University of California, Riverside
- Wei Wang 0039 - University of California, Berkeley
- Wei Wang 0040 - MediaNet Digital / Motorola / State University of New York at Buffalo
- Wei Wang 0041 - CUHK, Dept. of Computer Science and Engineering
- Wei Wang 0042 - University of Surrey / University of Nottingham, Malaysia
- Wei Wang 0043 - DOCOMO Beijing Labs
- Wei Wang 0044 - School of Transportation, Southeast University, Nanjing
- Wei Wang 0045 - Binghamton University
- Wei Wang 0046 - College of Physical and Electronic Information, Tianjin Normal University, Tianjin, China
- Wei Wang 0047 - School of Information Technology, Murdoch University, Murdoch, Australia
- Wei Wang 0048 - School of Mathematical Sciences, Dalian University of Technology, Dalian, P. R. China
- Wei Wang 0049 - Institute of Software, Chinese Academy of Sciences

Ask others: ACM DL/Guide - CSB - MetaPress - Google - Bing - Yahoo

	2013
j211	Wei Wang, Zhidan Yan: Connectivity of Kronecker products with complete multipartite graphs. Discrete Applied Mathematics 161(10-11): 1655-1659 (2013)
j210	Xin Zhou, Xin Yin, Ruian Liu, Wei Wang: Infrared and visible image fusion technology based on directionlets transform. EURASIP J. Wireless Comm. and Networking 2013: 42 (2013)
j209	Jiajia Luo, Wei Wang, Hairong Qi: Feature Extraction and Representation for Distributed Multi-View Human Action Recognition. IEEE J. Emerg. Sel. Topics Circuits Syst. 3(2): 145-154 (2013)

Refine by AUTHO
Wei Wang (663)
Zhihong Li (13)
Dragan Nesic (8)
Xiaoyu Li (6)
[top 4] [top 50] [to

Refine by VENUE
ROBIO (15)
NEMS (15)
CIS (13)

Linking Temporal Records

- How many Xin Dong's are in DBLP, with which publications?

- Xin Dong
- Xin Dong 0002
- Xin Dong 0003
- Xin Dong 0004
- Xin Dong 0005
- Xin Dong 0006

Ask others: ACM DL/Guide - S - CSB - MetaPress - Google - Bing - Yahoo

	2013
c26	Xin Luna Dong, Divesh Srivastava: Big data integration. ICDE 2013: 1245-1248
c25	Anish Das Sarma, Xin Luna Dong: SIGMOD 2013 new researcher symposium. SIGMOD Conference 2013: 1293-1294
c24	Xin Luna Dong, Laure Berti-Equille, Divesh Srivastava: Data Fusion: Resolving Conflicts from Multiple Sources. WAIM 2013: 64-76
c23	Xin Luna Dong, Divesh Srivastava: Compact explanation of data fusion decisions. WWW 2013: 379-390
e3	Lei Chen, Xin Luna Dong (Eds.): Proceedings of the 2013 SIGMOD/PODS Ph.D. Symposium, New York, NY, USA, June 23, 2013. ACM 2013, ISBN 978-1-4503-2155-6

Refine by AUTHOR

Xin Dong (50)
Alon Y. Halevy (20)
Divesh Srivastava (19)
Anish Das Sarma (8)
[top 4] [top 50] [all 57]

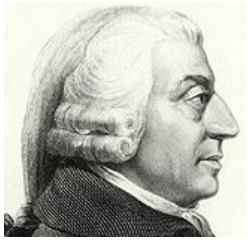
Refine by VENUE

PVLDB (9)
SIGMOD (8)
WebDB (4)
VLDB (3)
[top 4] [all 24]

Linking Temporal Records: Motivation

- Traditional record linkage
  - Links records of an entity from multiple sources **at a point in time**

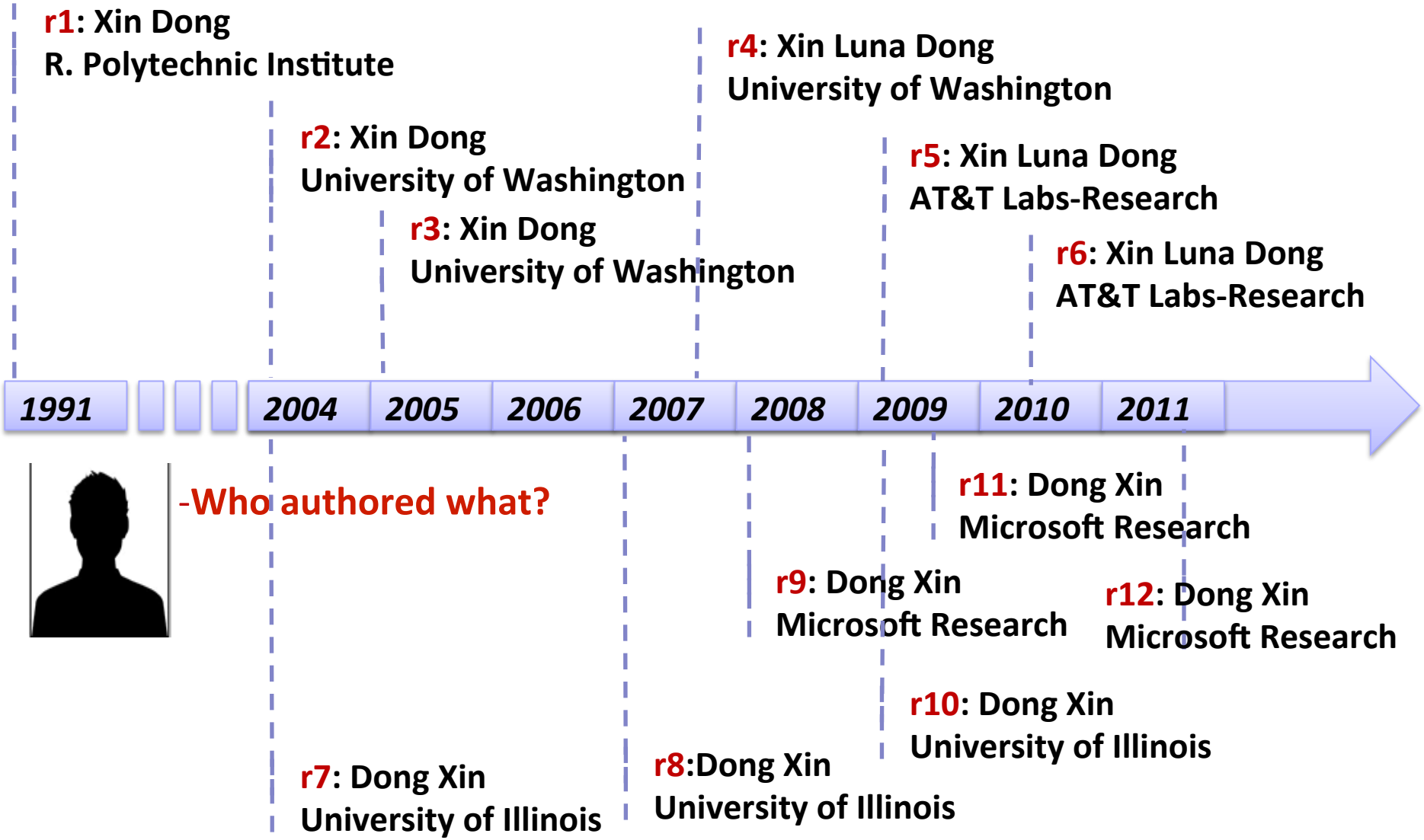
- Record linkage in Long Data
  - Links records of an entity **over a long time period**
  - Attribute values of an entity evolve over time
  - Different entities across time may have the same attribute value



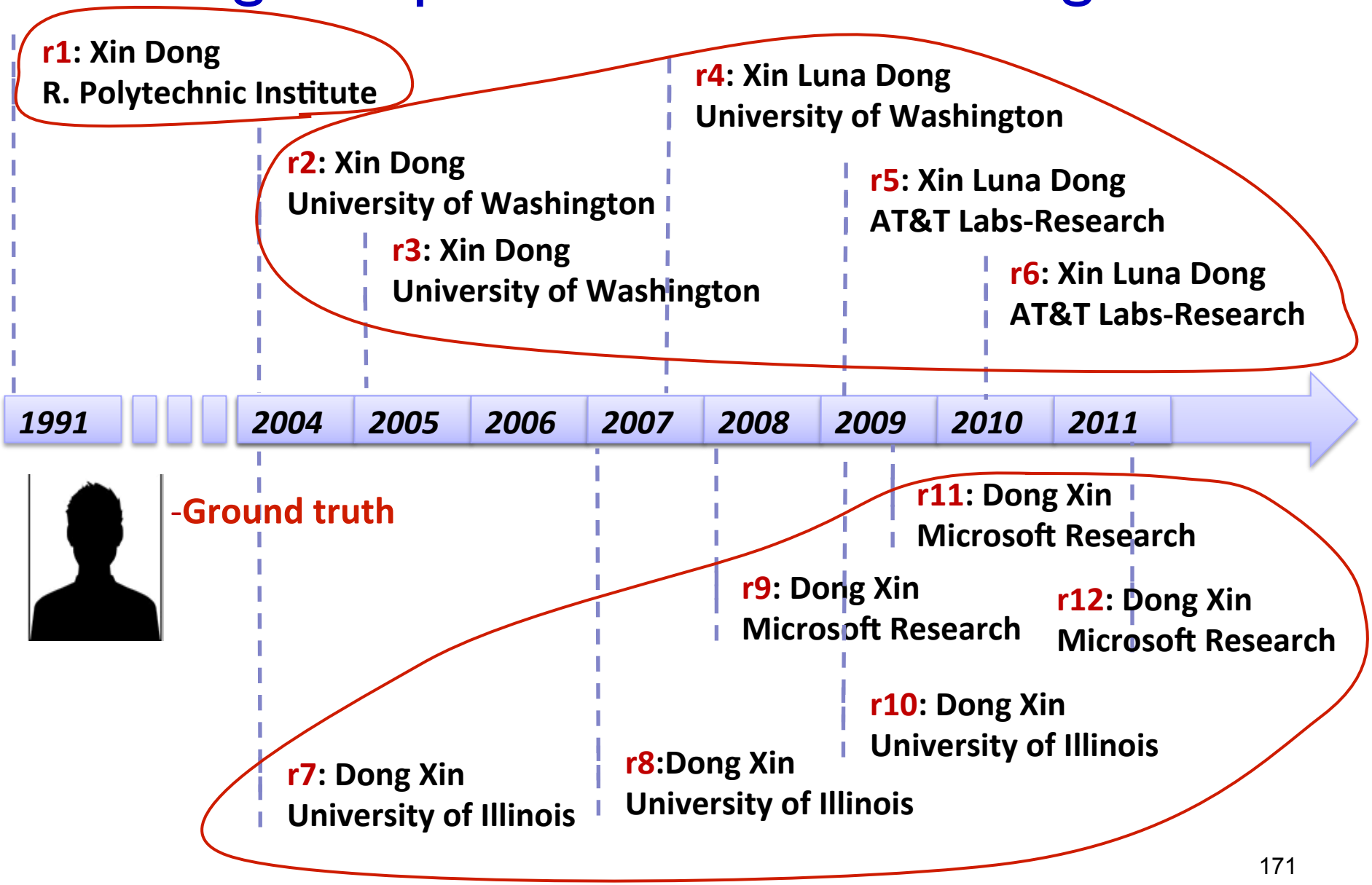
Adam Smith (1723-1790)

Adam Smith (1965-)

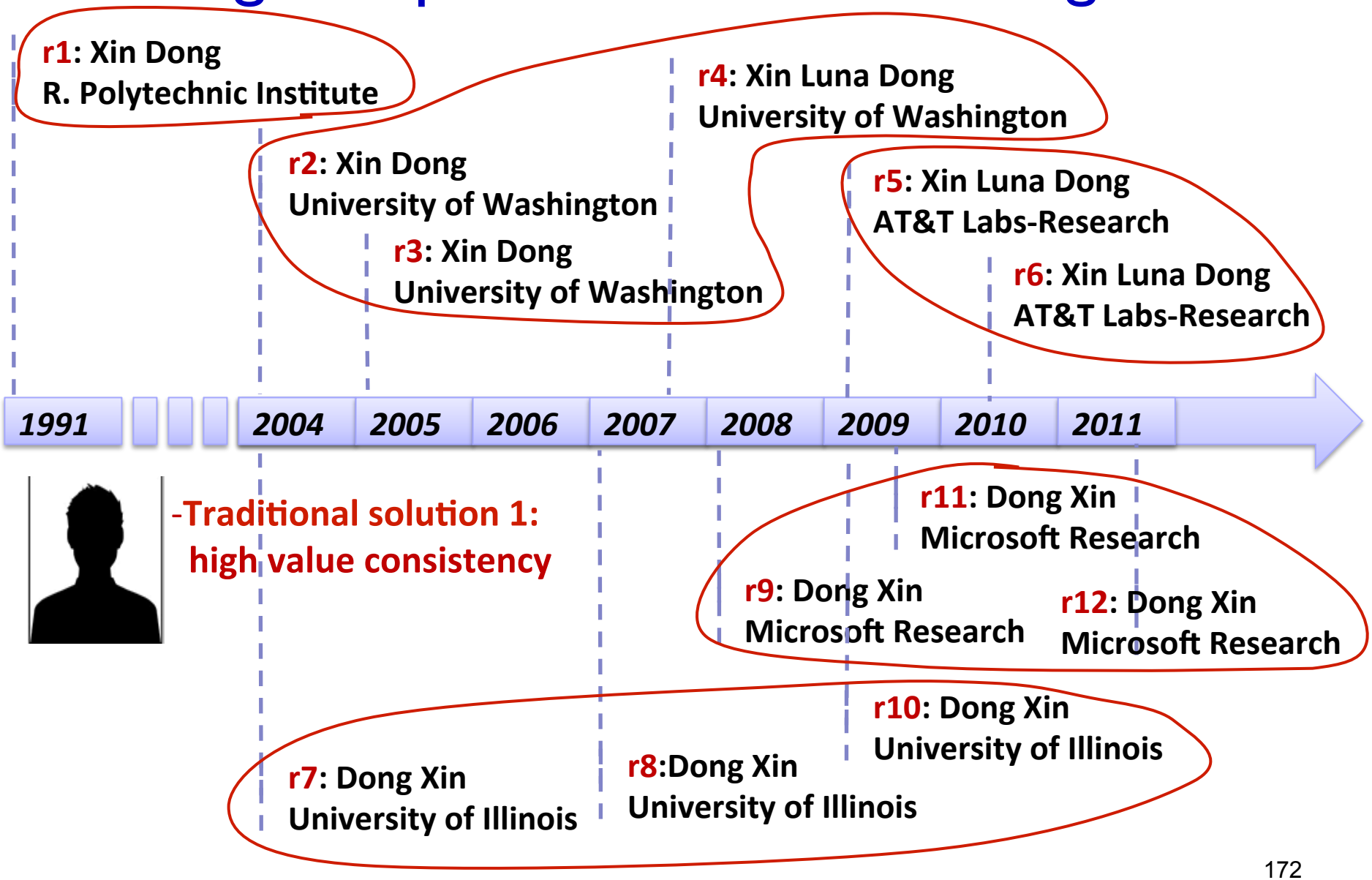
Linking Temporal Records: Challenges



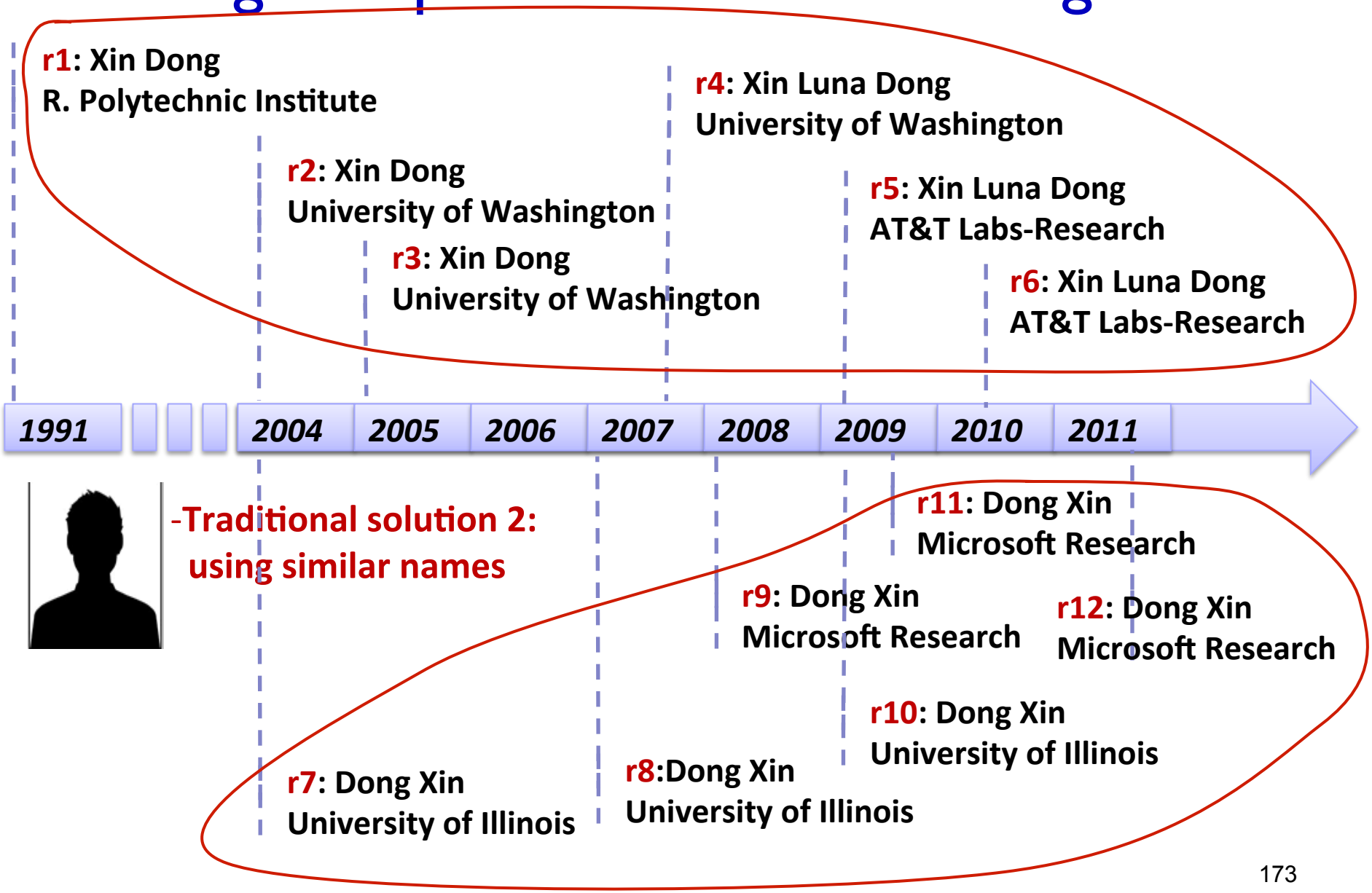
Linking Temporal Records: Challenges



Linking Temporal Records: Challenges



Linking Temporal Records: Challenges



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

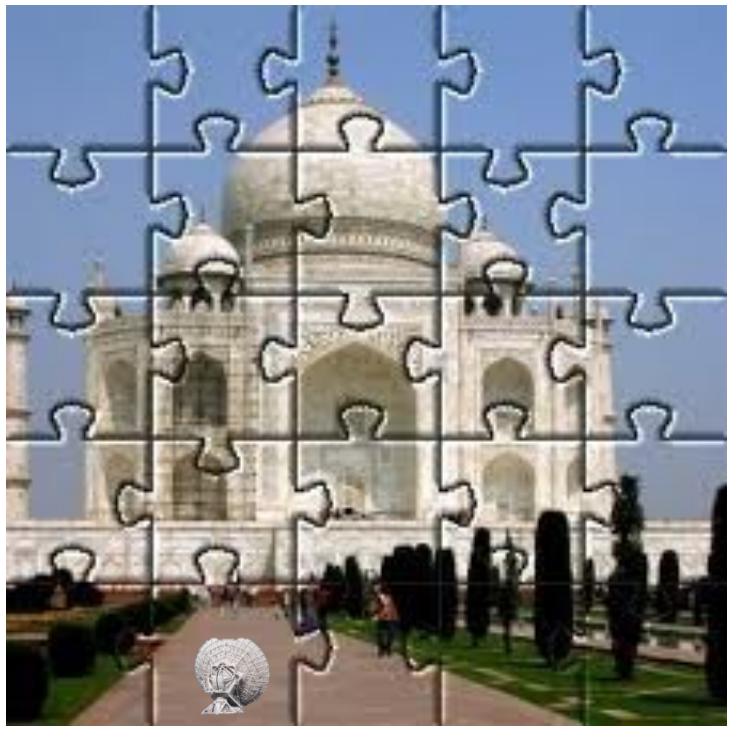
ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	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

Outline

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

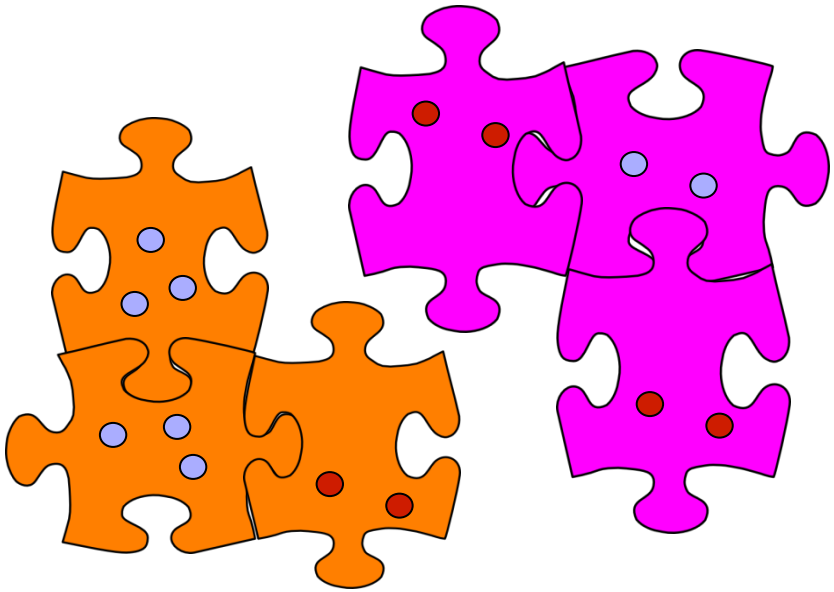
Data Fusion

- Reconciliation of conflicting content: pattern



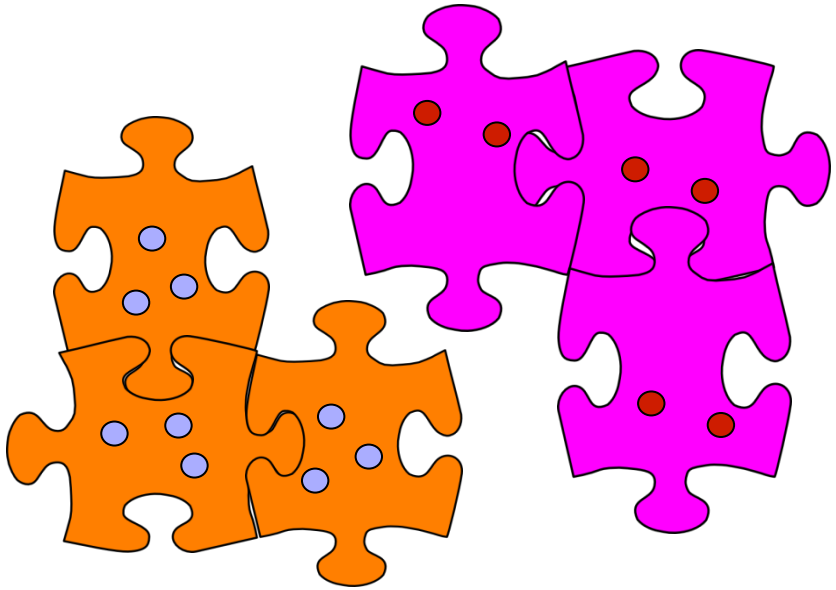
Data Fusion

- Reconciliation of conflicting content: pattern



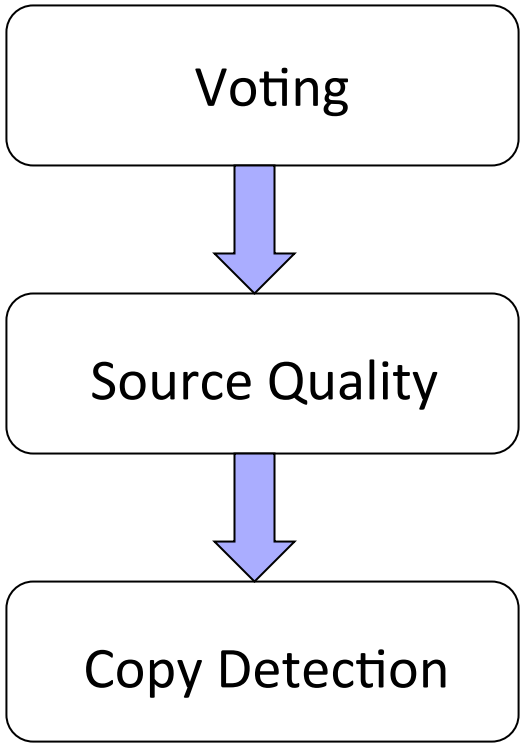
Data Fusion

- Reconciliation of conflicting content: pattern



Data Fusion: Three Components [DBS09a]

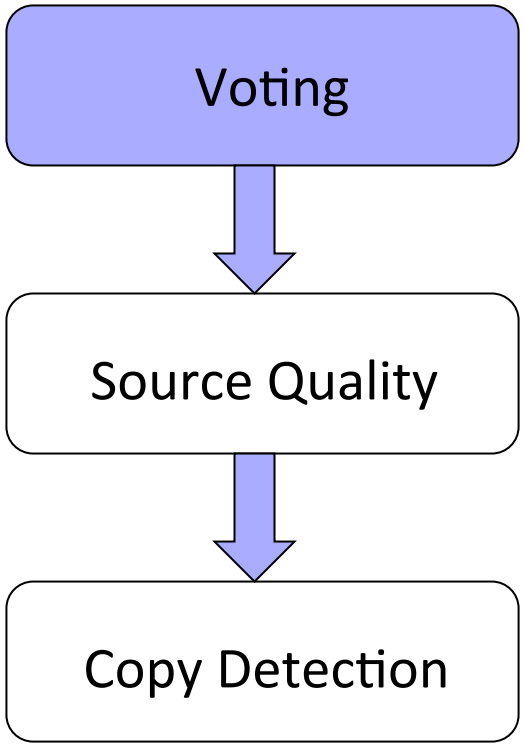
- Data fusion: voting + source quality + copy detection
  - Resolves inconsistency across diversity of sources



	S1	S2	S3	S4	S5
Jagadish	UM	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]

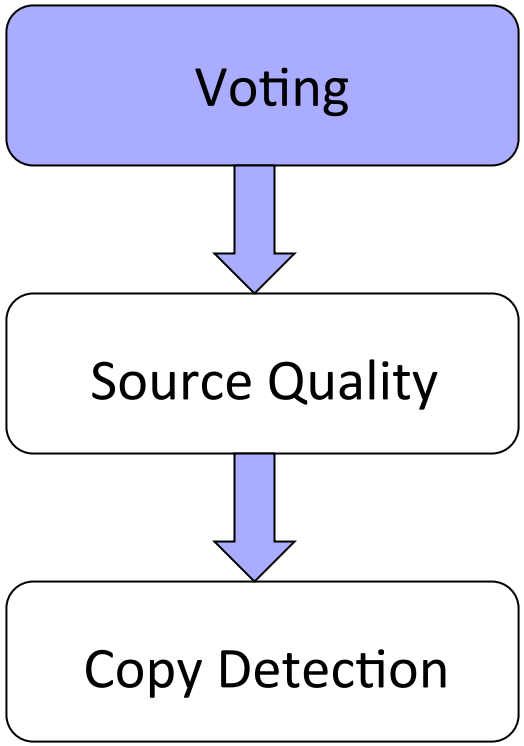
- Data fusion: voting + source quality + copy detection



	S1	S2	S3
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]

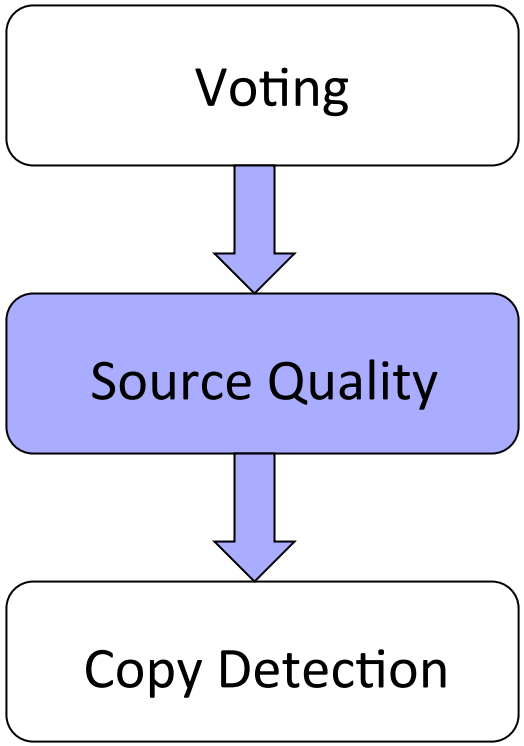
- Data fusion: voting + source quality + copy detection
  - Supports difference of opinion



	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]

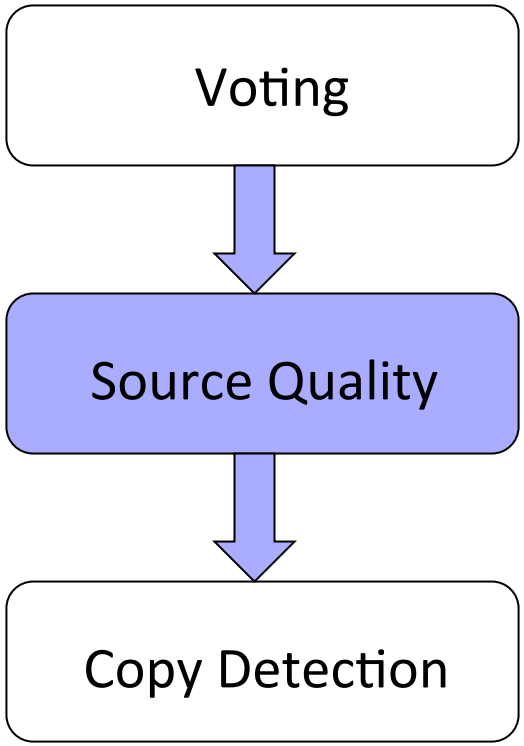
- Data fusion: voting + source quality + copy detection



	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]

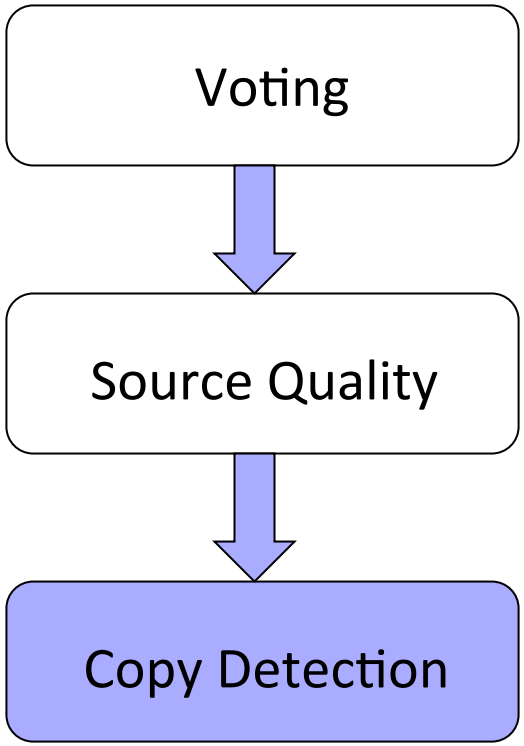
- Data fusion: voting + source quality + copy detection
  - Gives more weight to knowledgeable sources



	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]

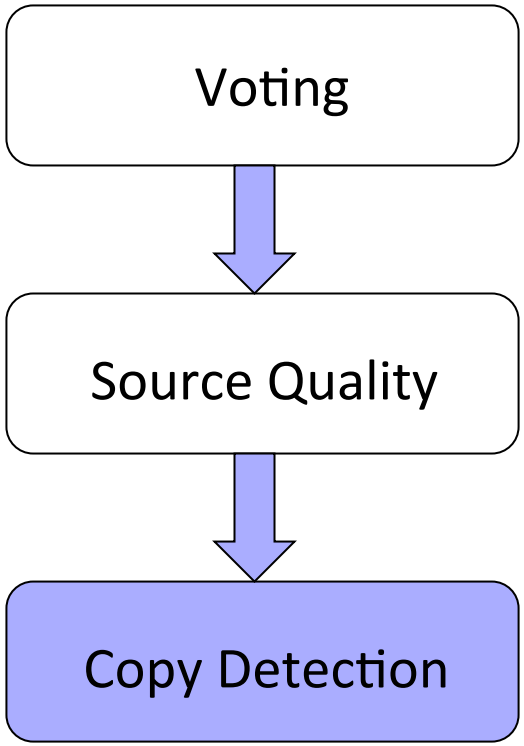
- Data fusion: voting + source quality + copy detection



	S1	S2	S3	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]

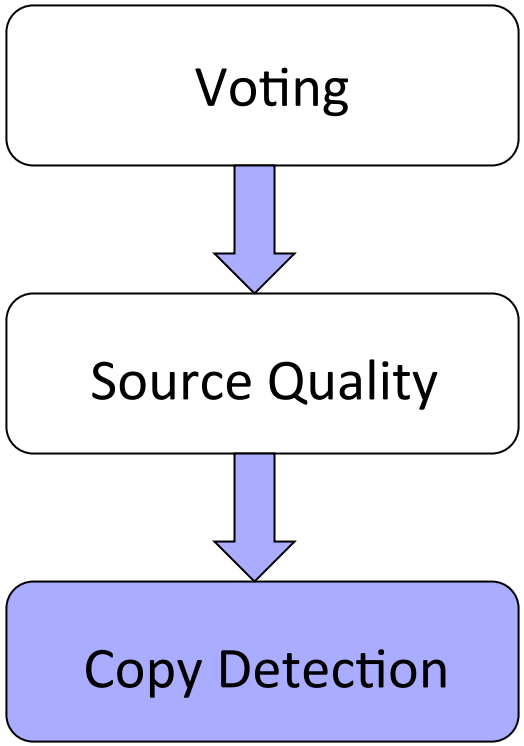
- Data fusion: voting + source quality + copy detection



	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

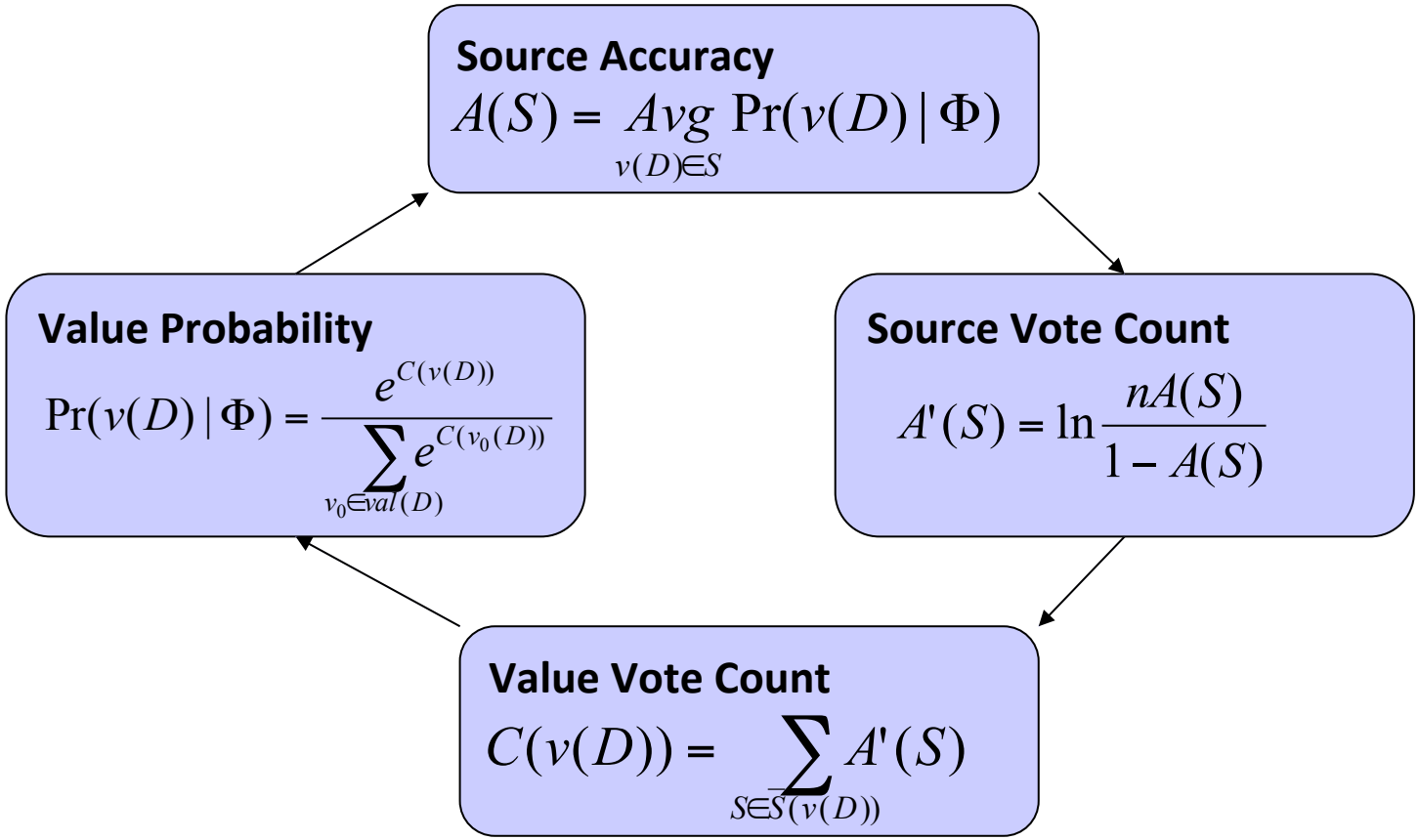
- How to compute $\Pr(v_i(D) \text{ true} \mid \Phi)$?

Source Accuracy

- Input: data item D, val(D) = $\{v_0, v_1, \ldots, v_n\}$, Φ
- Output: $Pr(v_i(D)\ \text{true} \mid \Phi)$, for i=0,…, 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?

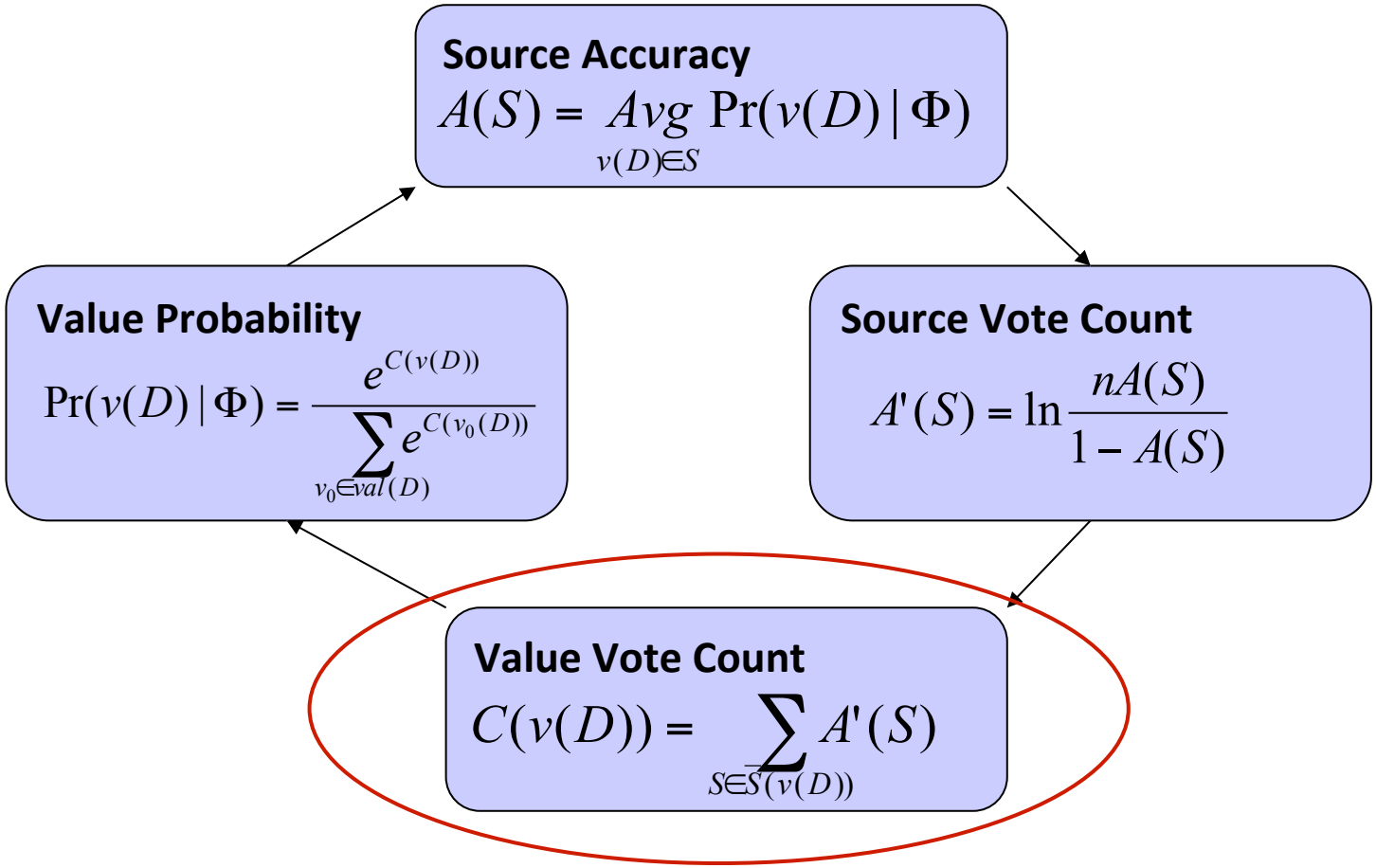
Source Accuracy

- 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

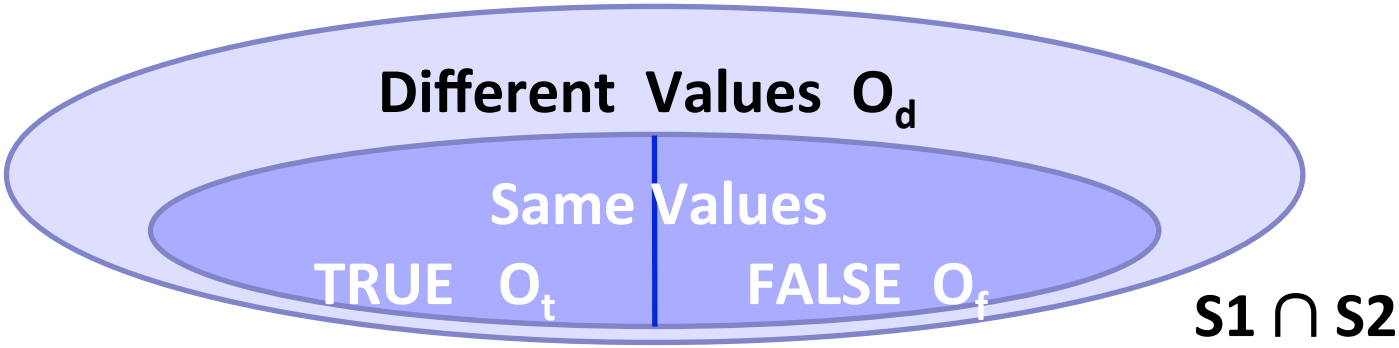
Source 1 on USA Presidents:	Source 2 on USA Presidents:	
1st : George Washington	1st : George Washington	✓
2nd : John Adams	2nd : John Adams	✓
3rd : Thomas Jefferson	3rd : Thomas Jefferson	✓
4th : James Madison	4th : James Madison	✓
…	…	
41st : George H.W. Bush	41st : George H.W. Bush	✓
42nd : William J. Clinton	42nd : William J. Clinton	✓
43rd : George W. Bush	43rd : George W. Bush	✓
44th: Barack Obama	44th: Barack Obama	✓

Copy Detection

Are Source 1 and Source 2 dependent? Very likely

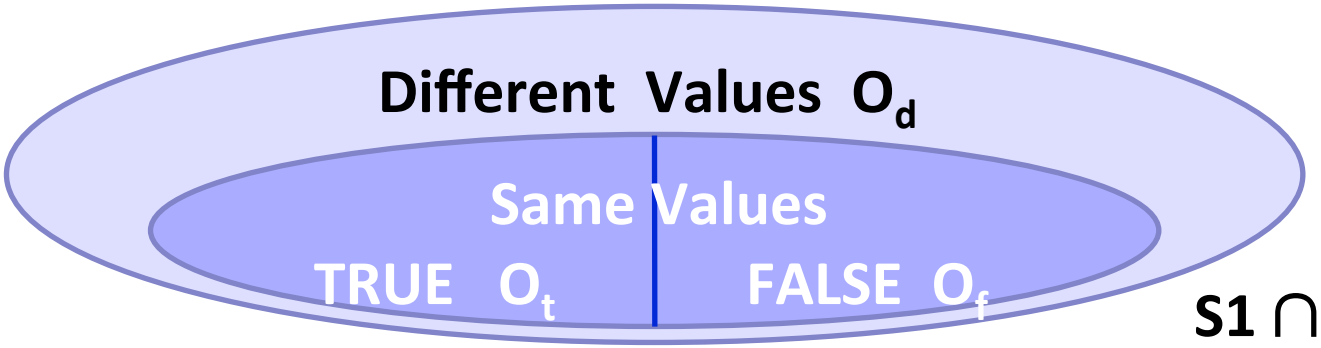
Source 1 on USA Presidents:	Source 2 on USA Presidents:	
1st : George Washington	1st : George Washington	✓
2nd : Benjamin Franklin	2nd : Benjamin Franklin	✗
3rd : John F. Kennedy	3rd : John F. Kennedy	✗
4th : Abraham Lincoln	4th : Abraham Lincoln	✗
…	…	
41st : George W. Bush	41st : George W. Bush	✗
42nd : Hillary Clinton	42nd : Hillary Clinton	✗
43rd : Dick Cheney	43rd : Dick Cheney	✗
44th: Barack Obama	44th: John McCain	

Copy Detection: Bayesian Analysis



- Goal: Pr(S1⊥S2| Φ), Pr(S1~S2| Φ) (sum = 1)
- According to Bayes Rule, we need Pr(Φ|S1⊥S2), Pr(Φ|S1~S2)
- Key: compute $Pr(\Phi_D|S1\perp S2)$, $Pr(\Phi_D|S1\sim S2)$, for each $D \in S1 \cap S2$

Copy Detection: Bayesian Analysis

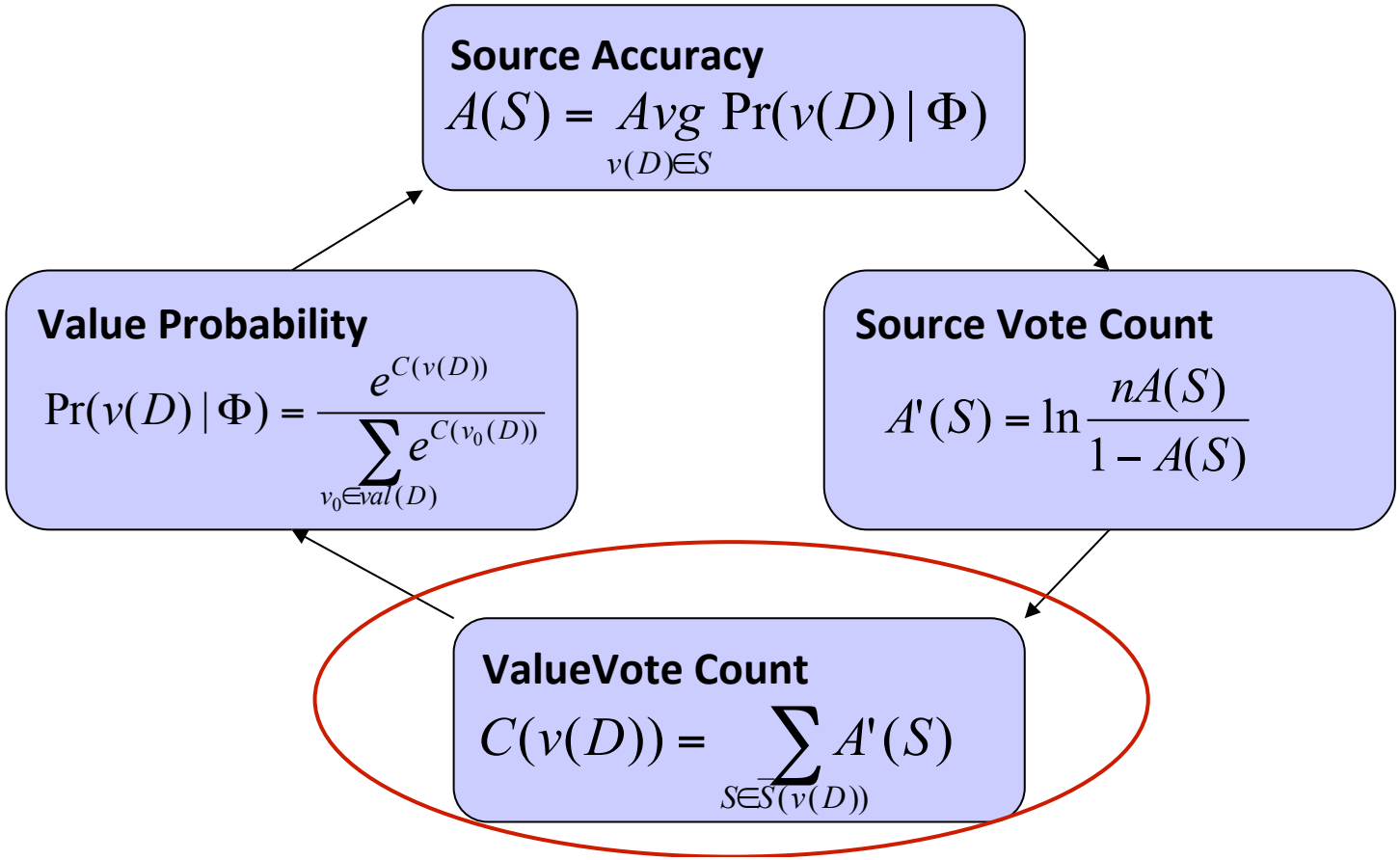


S1 ∩ S2

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

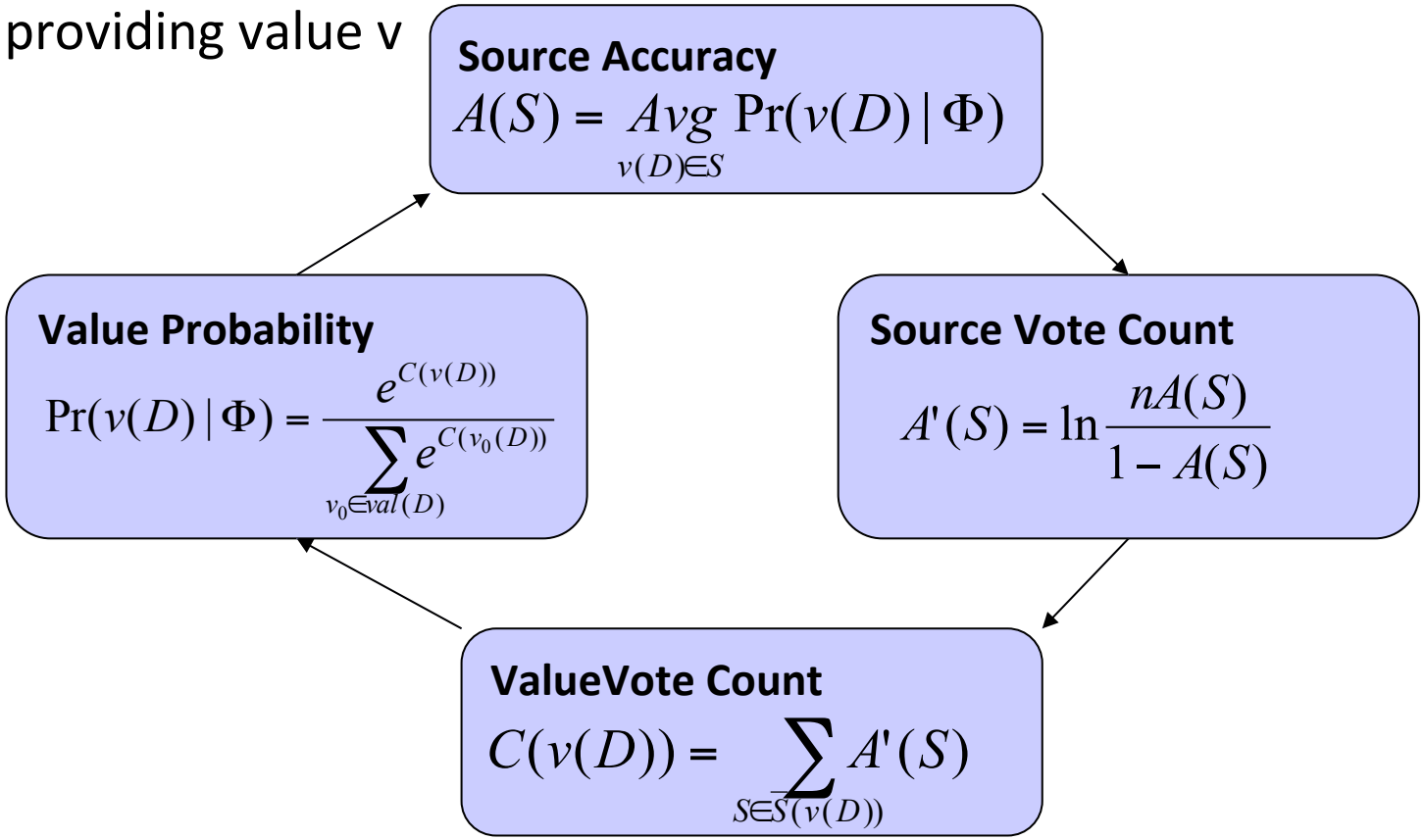
Discount Copied Values

- Continue until convergence



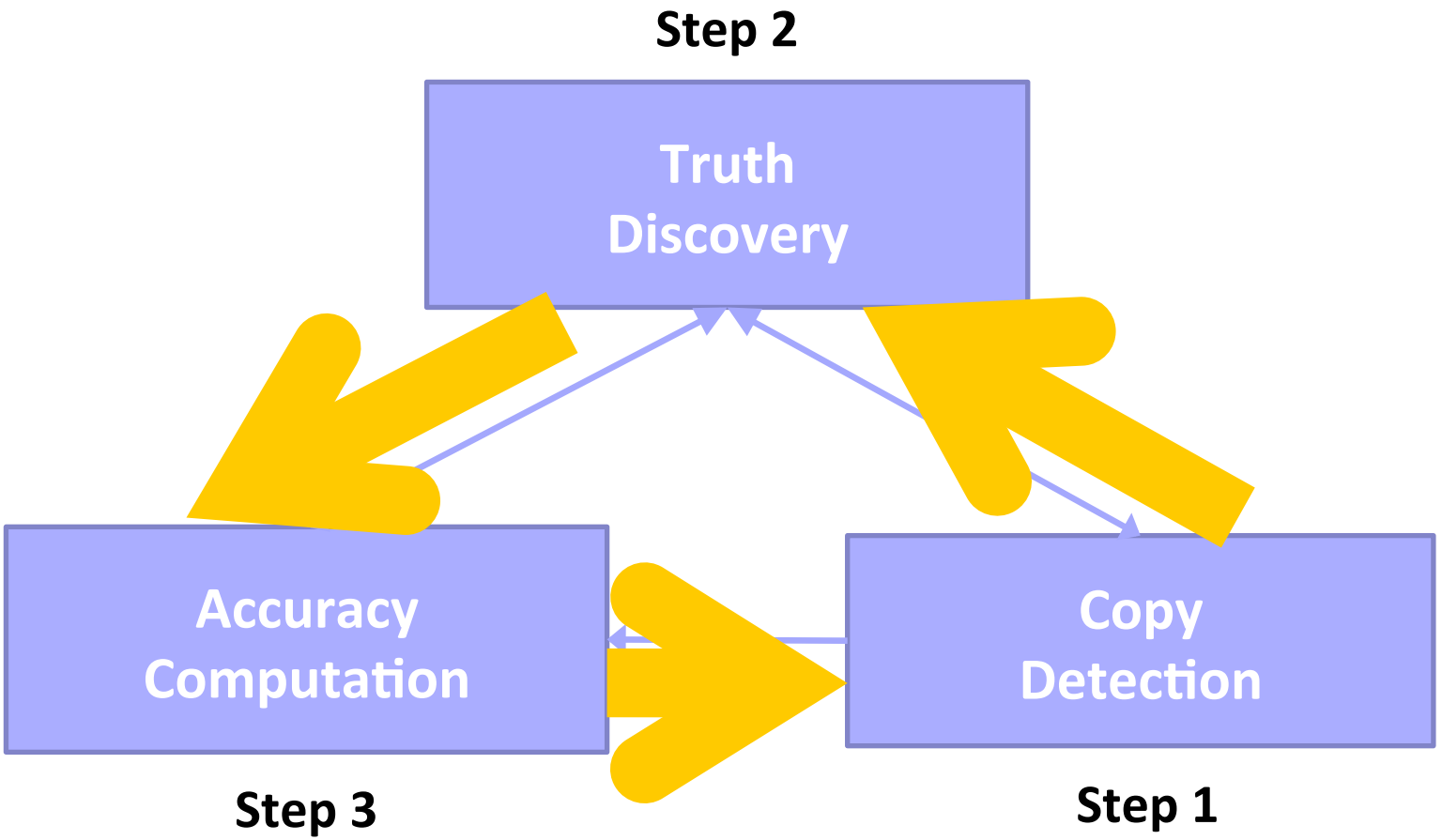
Discount Copied Values

- Continue until convergence
  - Consider dependence $C(v) = \sum_{S \in \mathcal{S}(v)} A'(S) \bullet I(S)$
  - 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, ***MIT***)	**(03, MIT)**	**(00, UCB)** Out-of-date!	(01, UCB) **(06, MIT)**	**(05, MIT)**	(03, UCB) **(05, MS)** ERR!
Dewitt (Θ, Wisc), (08, ***MSR***)	(00, Wisc) **(09, MSR)**	(00, UW) (01, Wisc) **(08, MSR)**	(01, UW) **(02, Wisc)** Out-of-date!	**(05, Wisc)** Out-of-date!	(03, UW) (05, ⊥) **(07, Wisc)**
Bernstein (Θ, ***MSR***)	**(00, MSR)**	**(00, MSR)**	**(01, MSR)**	**(07, MSR)**	**(03, MSR)**
Carey (Θ, Propell), (02, BEA), (08, ***UCI***)	(04, BEA) **(09, UCI)**	**(05, AT&T)** ERR!	**(06, BEA)** Out-of-date	**(07, BEA)** Out-of-date!	**(07, BEA)** Out-of-date!
Halevy (Θ, UW), (05, ***Google***)	(00, UW) **(07, Google)**	(00, Wisc) (02, UW) **(05, Google)**	(01, Wisc) **(06, UW)** SLOW!	**(05, UW)** SLOW!	(03, Wisc) (05, Google) **(07, UW)** SLOW!

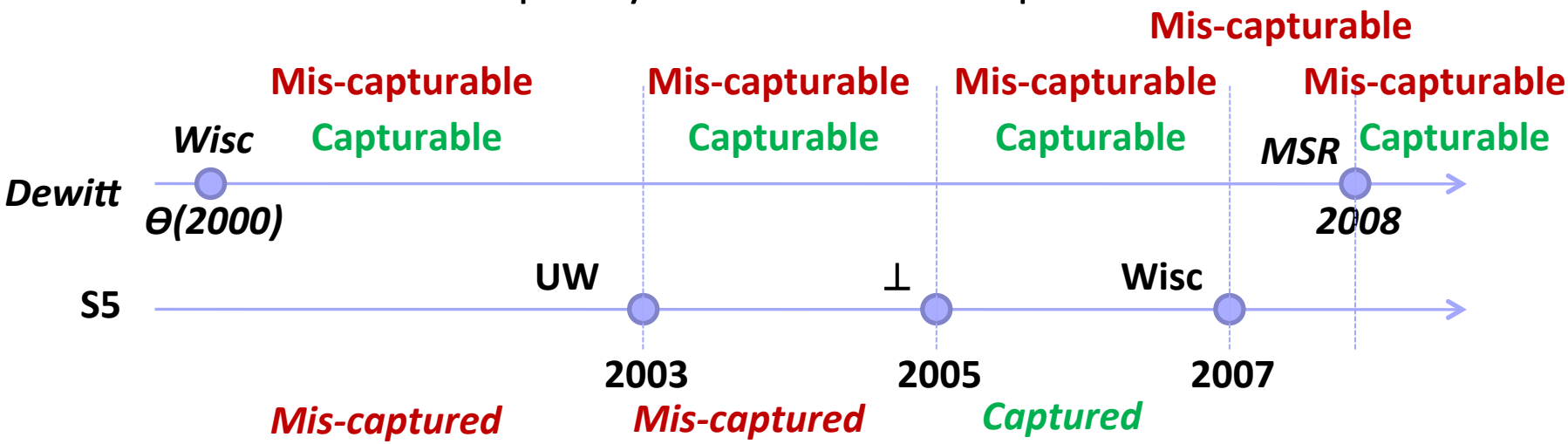
- True values can evolve over time
- Low-quality data can be caused by different reasons

Problem Definition

Problem Definition	Static World	Dynamic World
Objects	Each associated with a value; e.g., Google for Halevy	Each associated with a *lifespan*; 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	1. Life span: true value for each object at each time point 2. 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., ¼=.25)

Exactness= 1-#*Mis-Captured*/#Mis-Capturable (e.g., 1-2/5=.6)

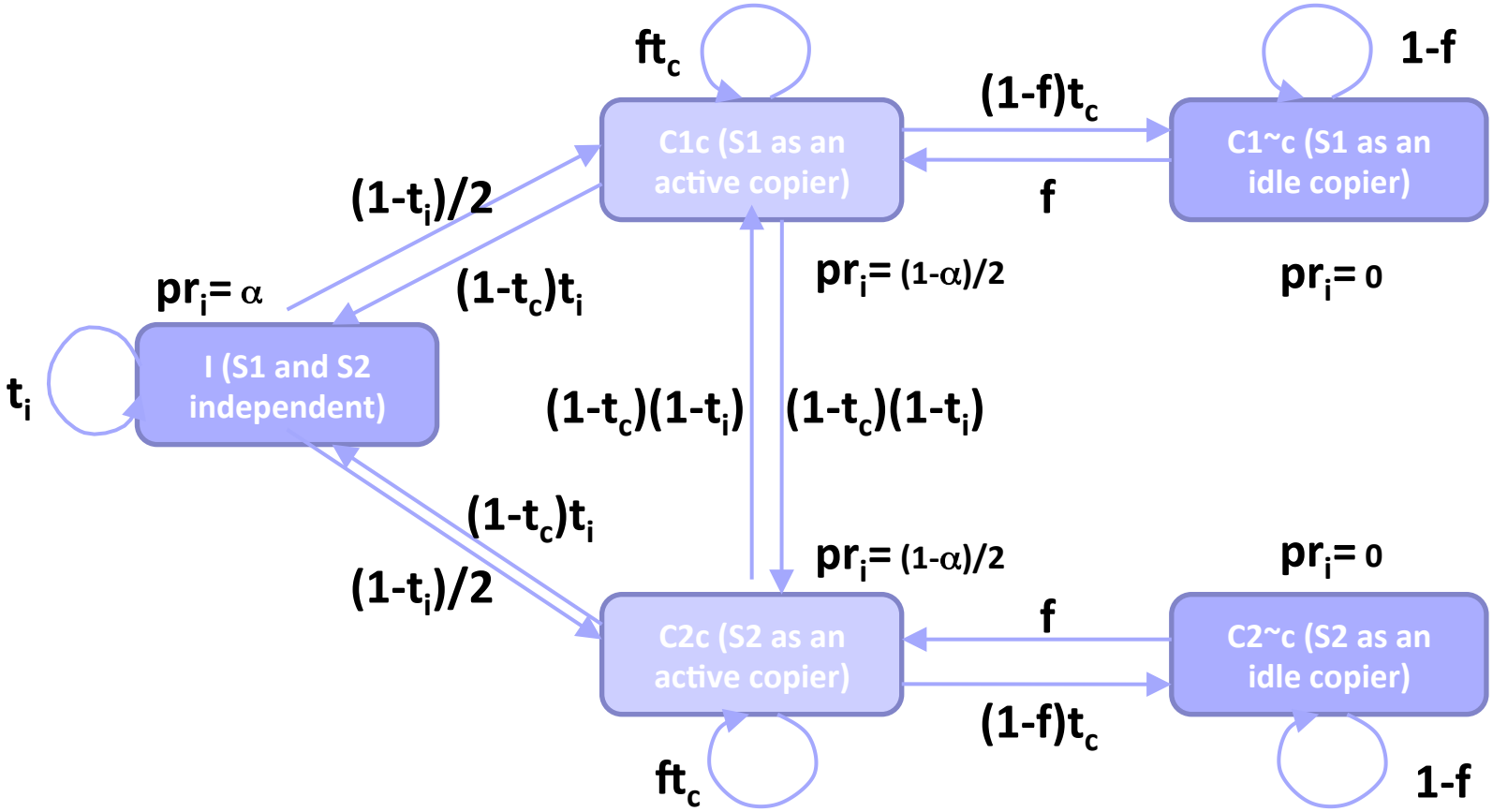
Freshness(Δ)= #(*Captured w. length*<=Δ)/#*Captured* (e.g., F(0)=0, F(1)=0, F(2)=1/1 = 1…)

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)

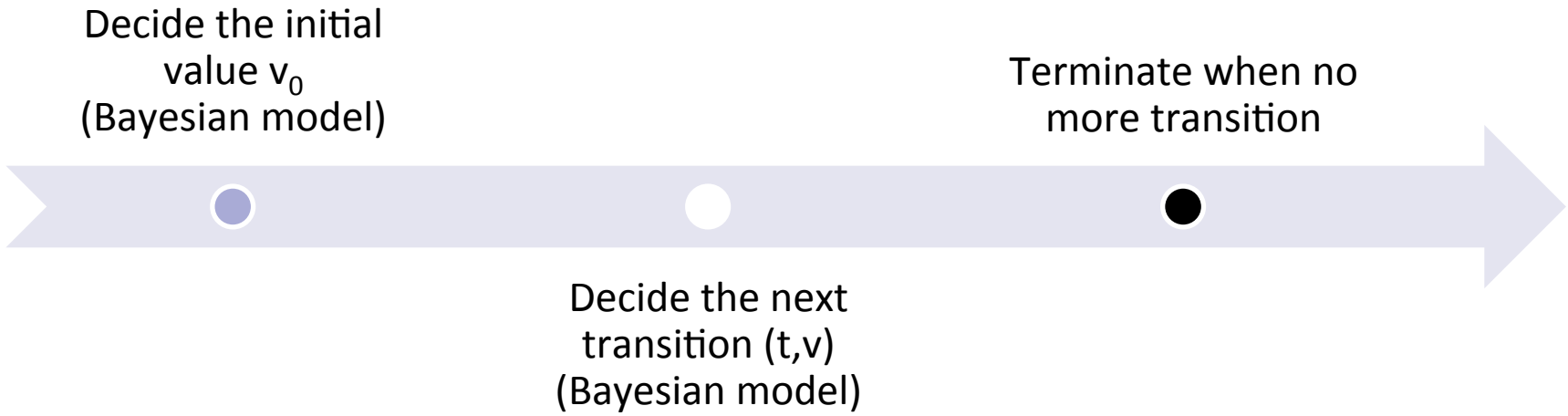
The Copying Detection HMM Model



- 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);

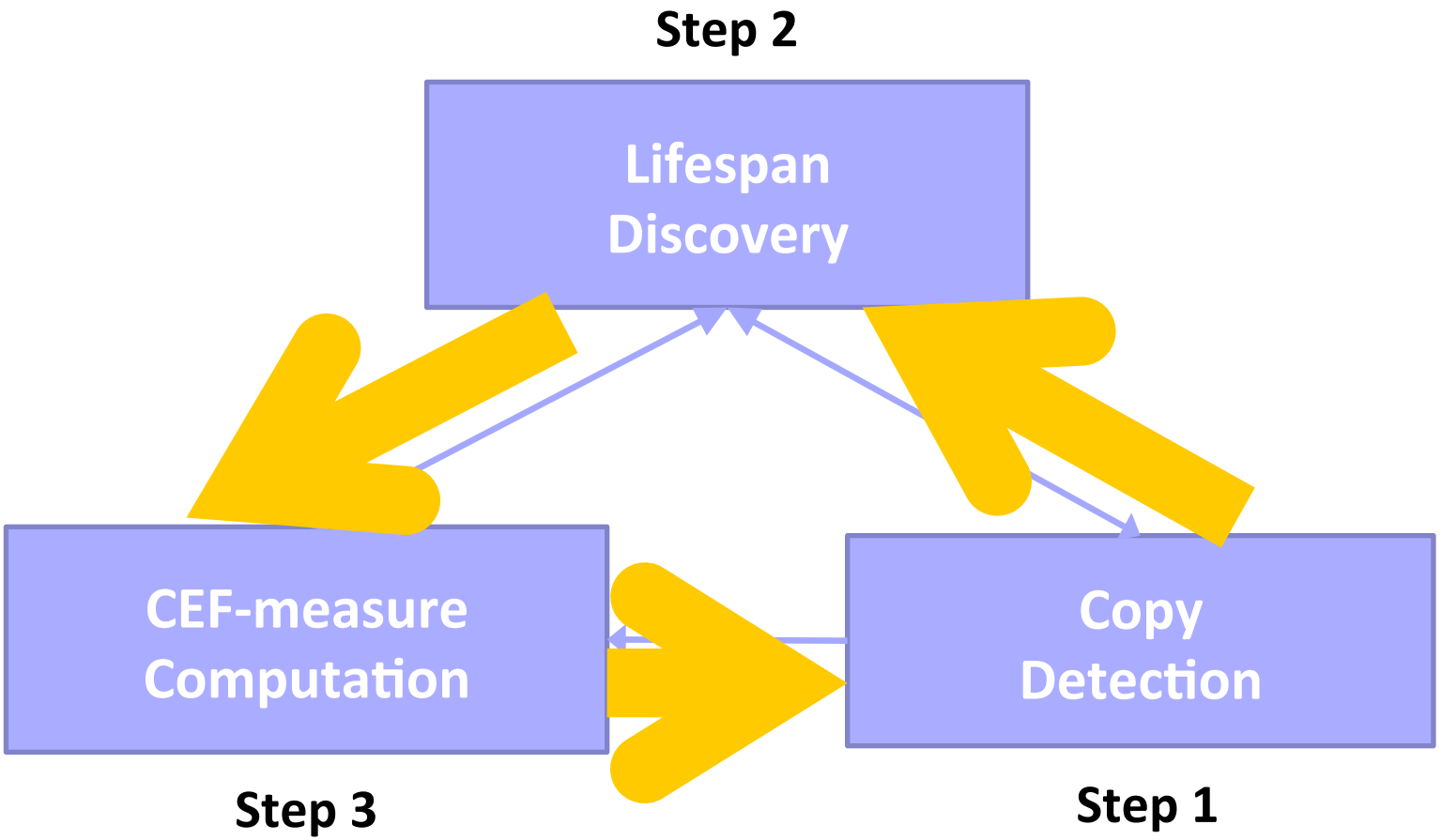
Lifespan Discovery

- Algorithm: for each object O



Iterative Process

- Typically converges when #objs >> #srcs



Example Revisited

	S1	S2	S3	S4	S5
Halevy (Θ, 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)

- 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

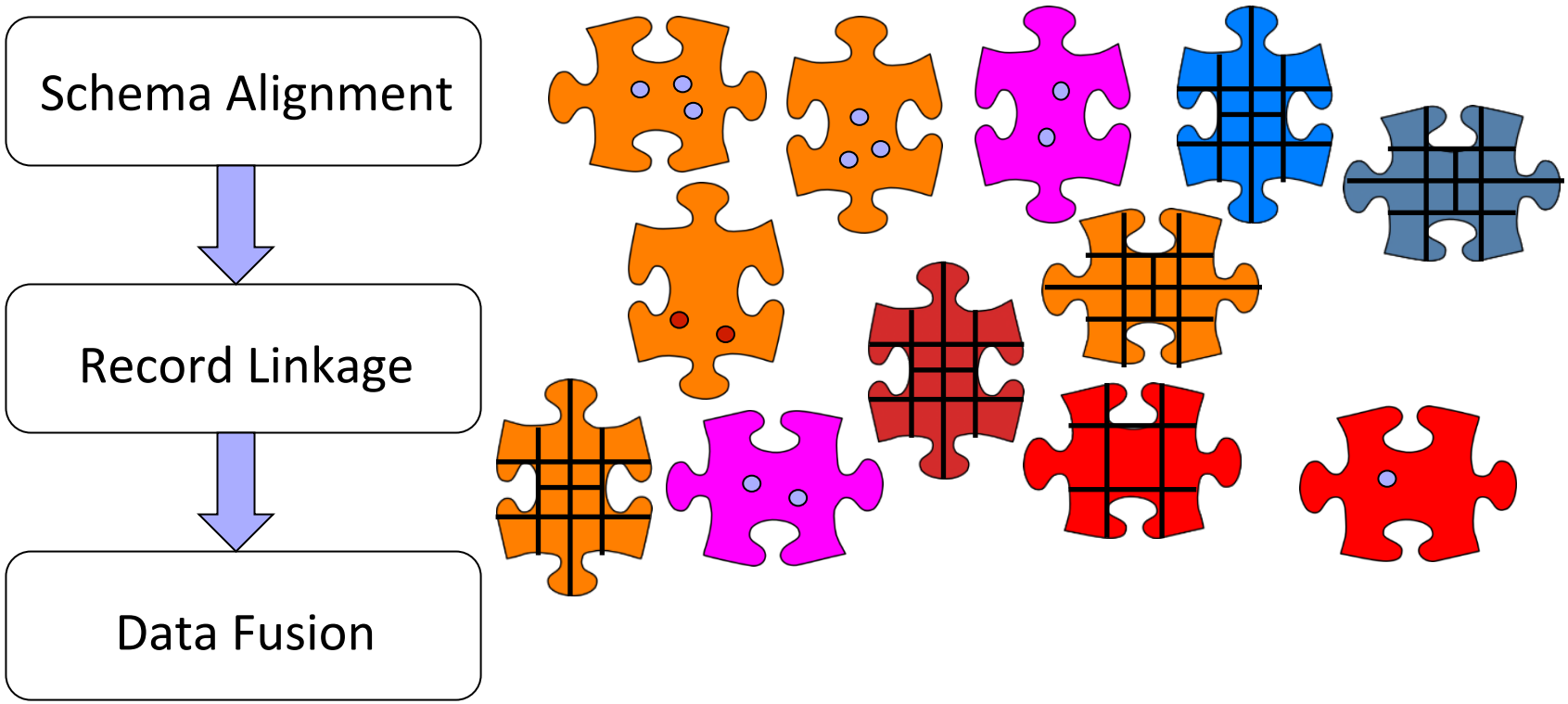
	Schema alignment	Record linkage	Data fusion
Volume	• Integrating deep web • Web tables	• Adaptive blocking • Meta blocking	• Online fusion
Velocity	• Keyword-based integration for dynamic data	• Incremental linkage	• Fusion for dynamic data
Variety	• Data spaces • Keyword-based integration	• Linking text to structured data	• Combining fusion with linkage
Veracity		• Value-variety tolerant linkage	• Truth discovery

Outline

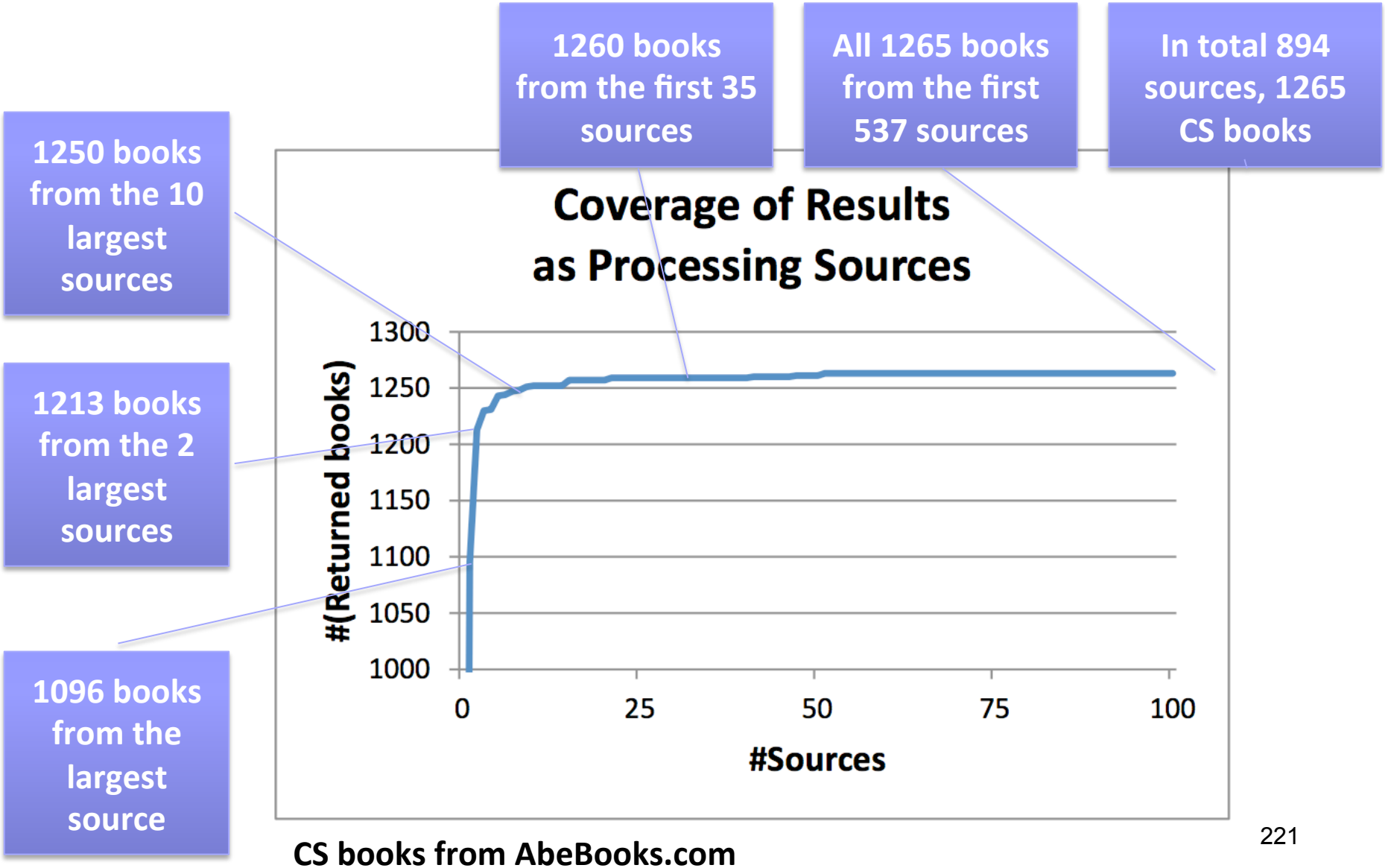
- Motivation
- Schema alignment
- Record linkage
- Data fusion
- Emerging topics
  - Source selection, future work

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



CS books from AbeBooks.com

Erroneous Data May Hurt Quality

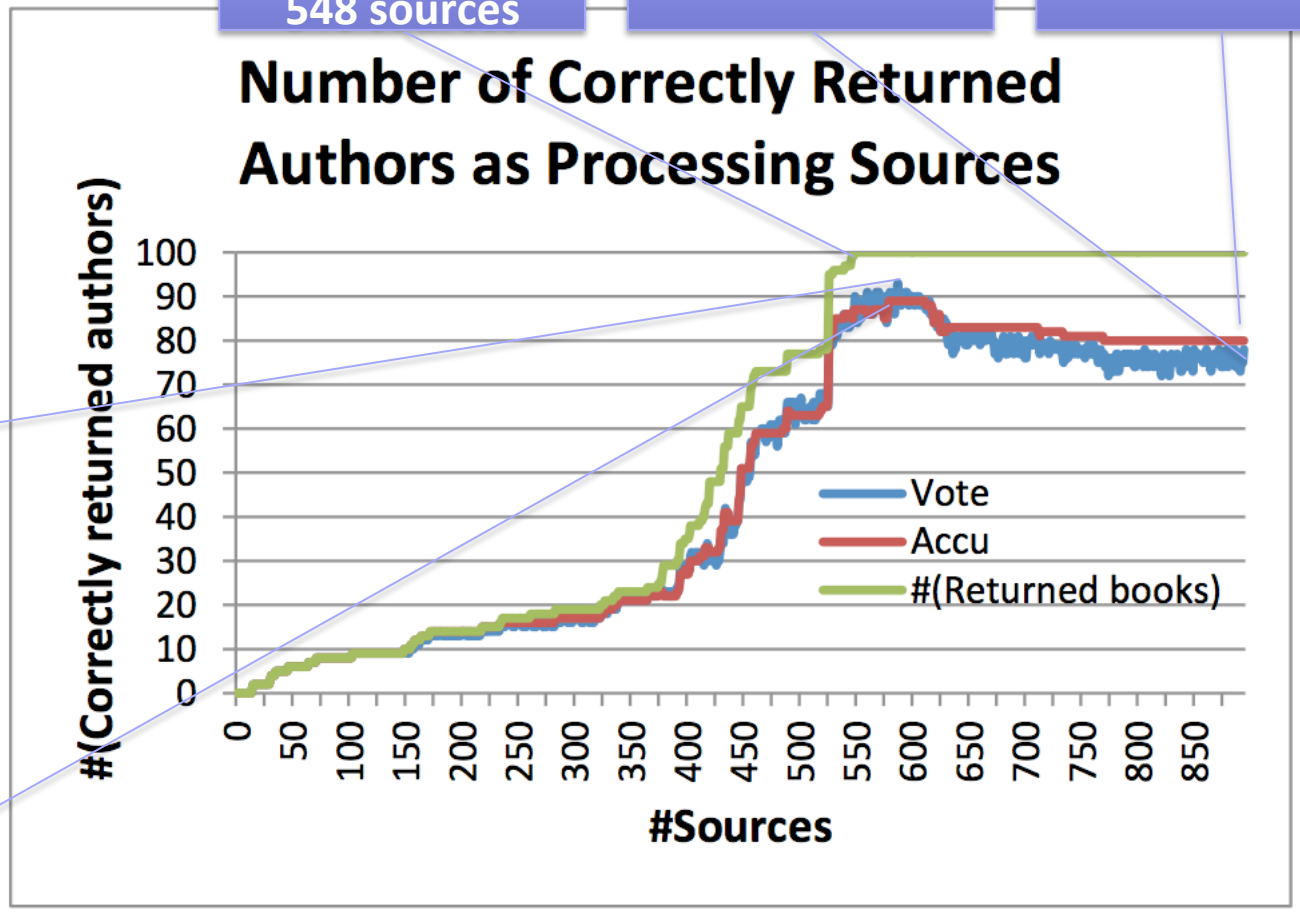
All 100 books (gold standard) from the first 548 sources

78 books w. correct authors for Vote

80 books w. correct authors for Accu

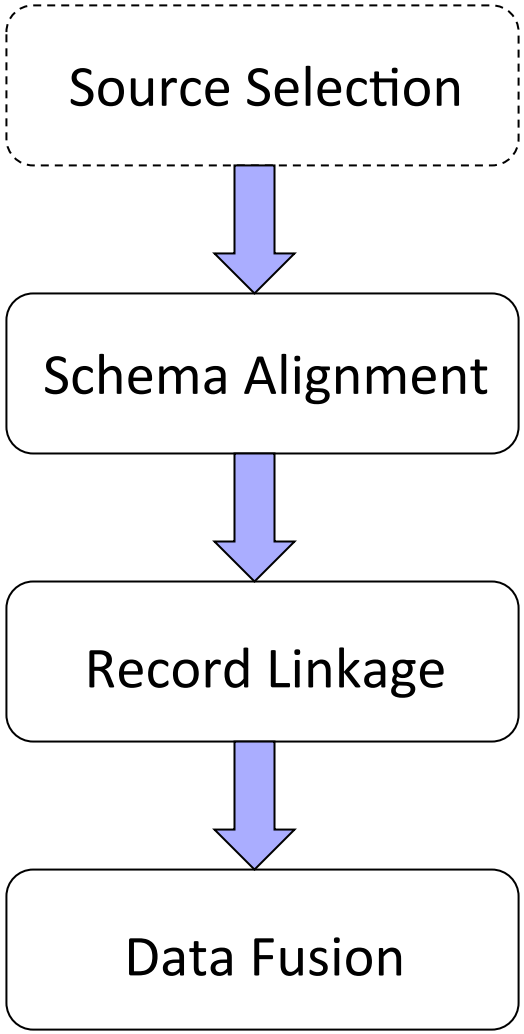
93 > 80 books w. correct authors after 583 sources (Vote)

90 > 80 books w. correct authors after 579 sources (Accu)



CS books from AbeBooks.com

BDI: Source Selection [DSS13]

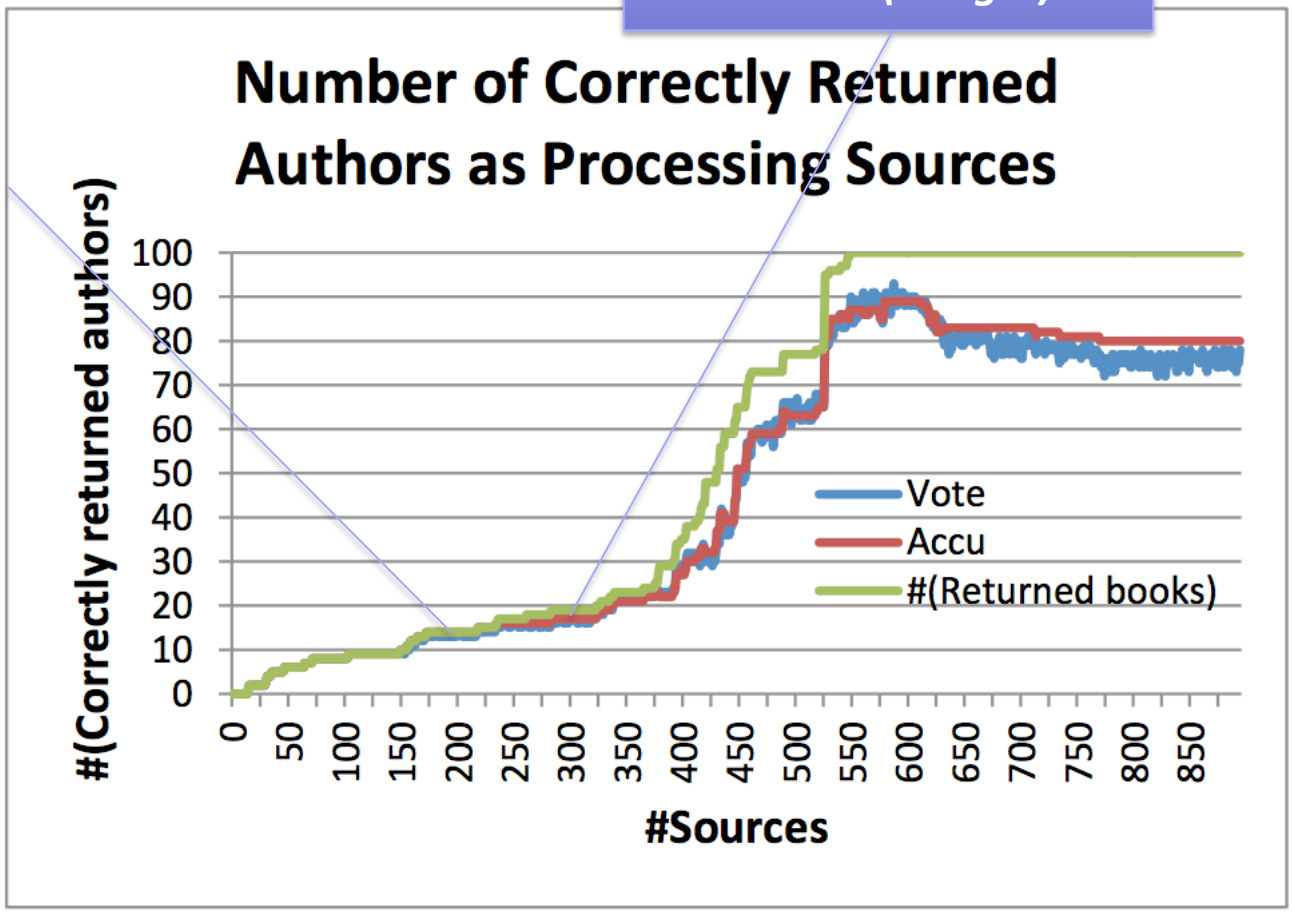


- How to wisely select sources before integration to balance gain and cost?

Maximize Quality Under Budget?

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

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

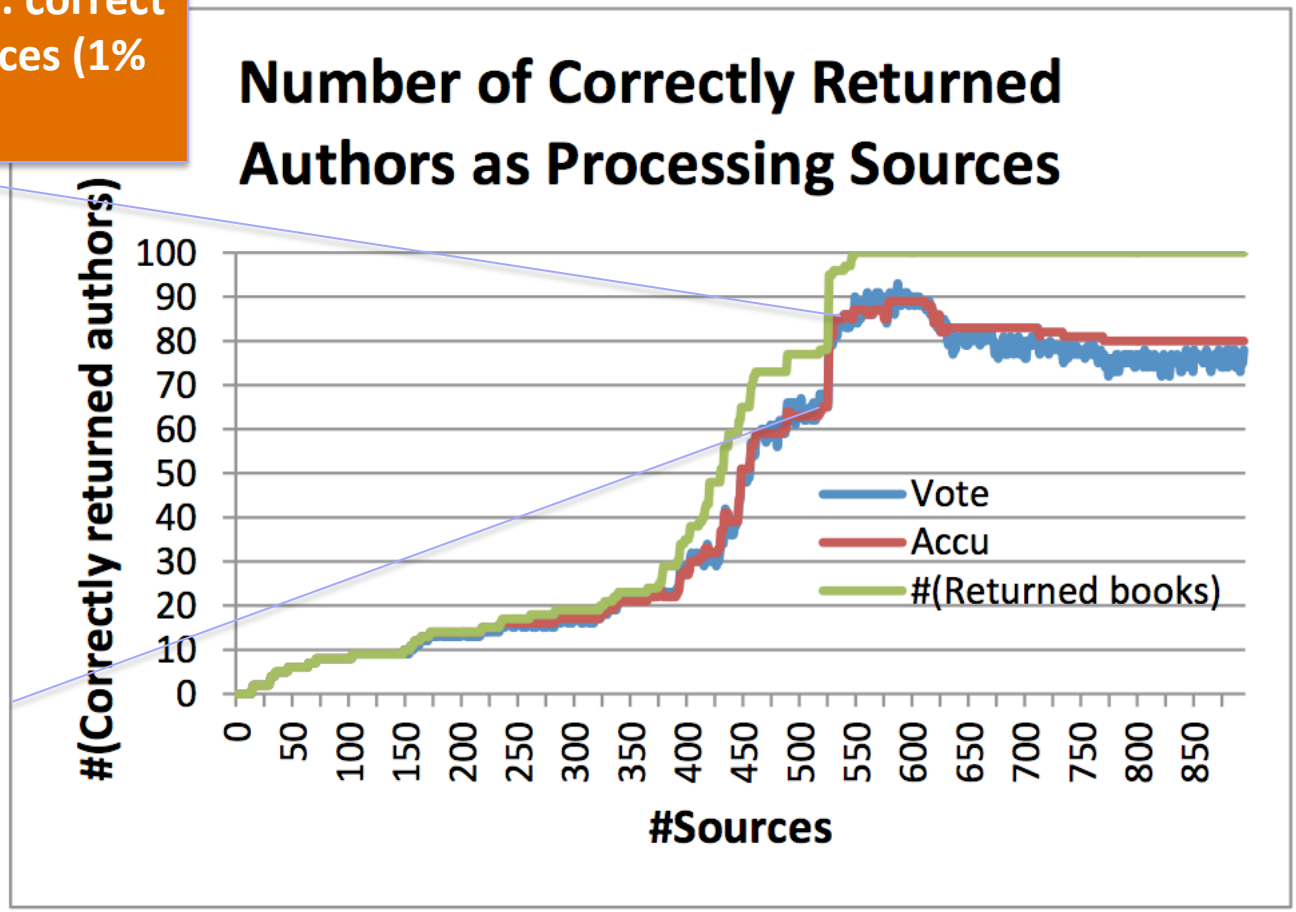


CS books from AbeBooks.com

Minimize Cost with Certain Quality?

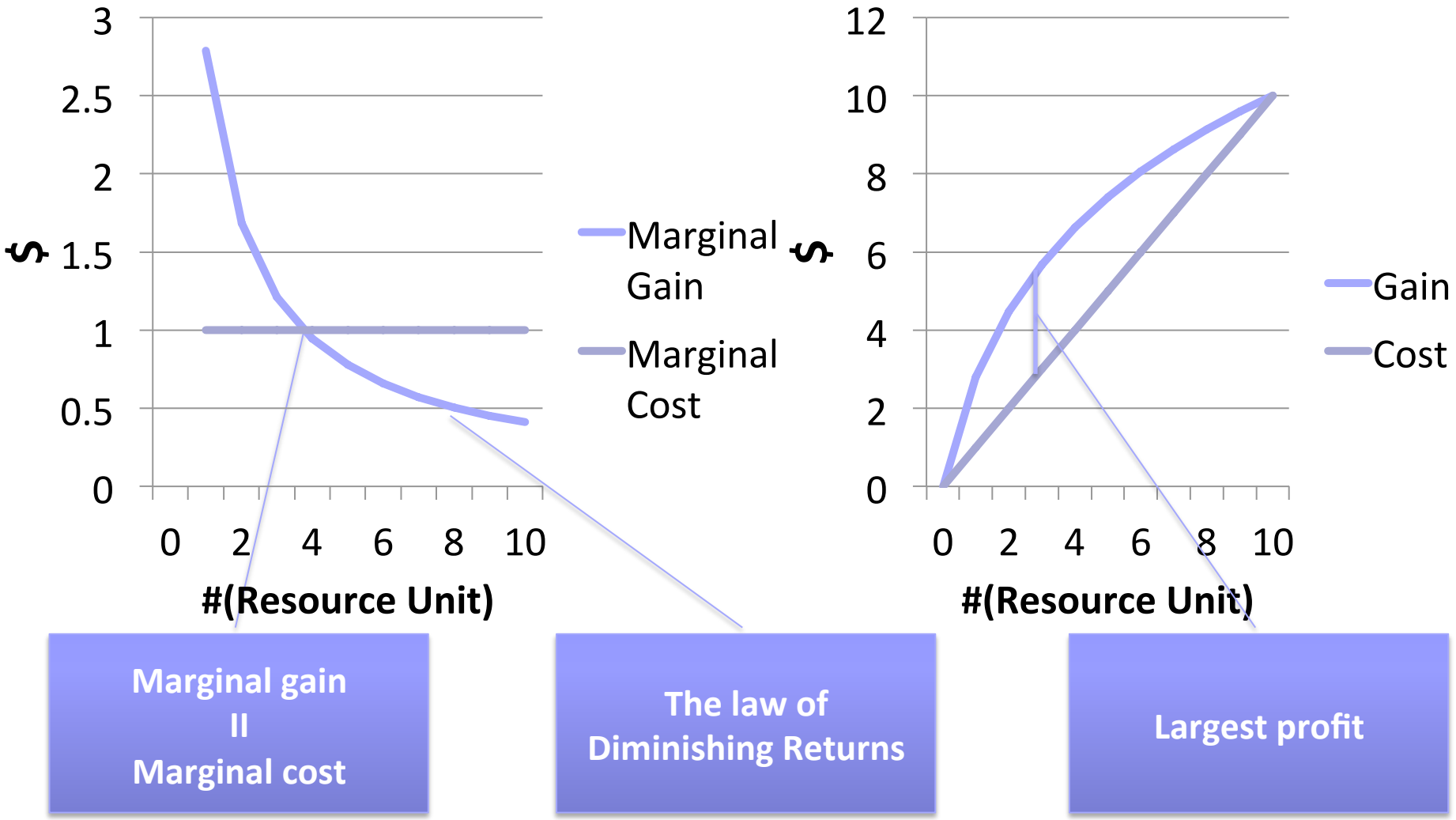
81 books (25% more) w. correct authors from 526 sources (1% more)

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



CS books from AbeBooks.com

Marginalism Principle in Economic Theory

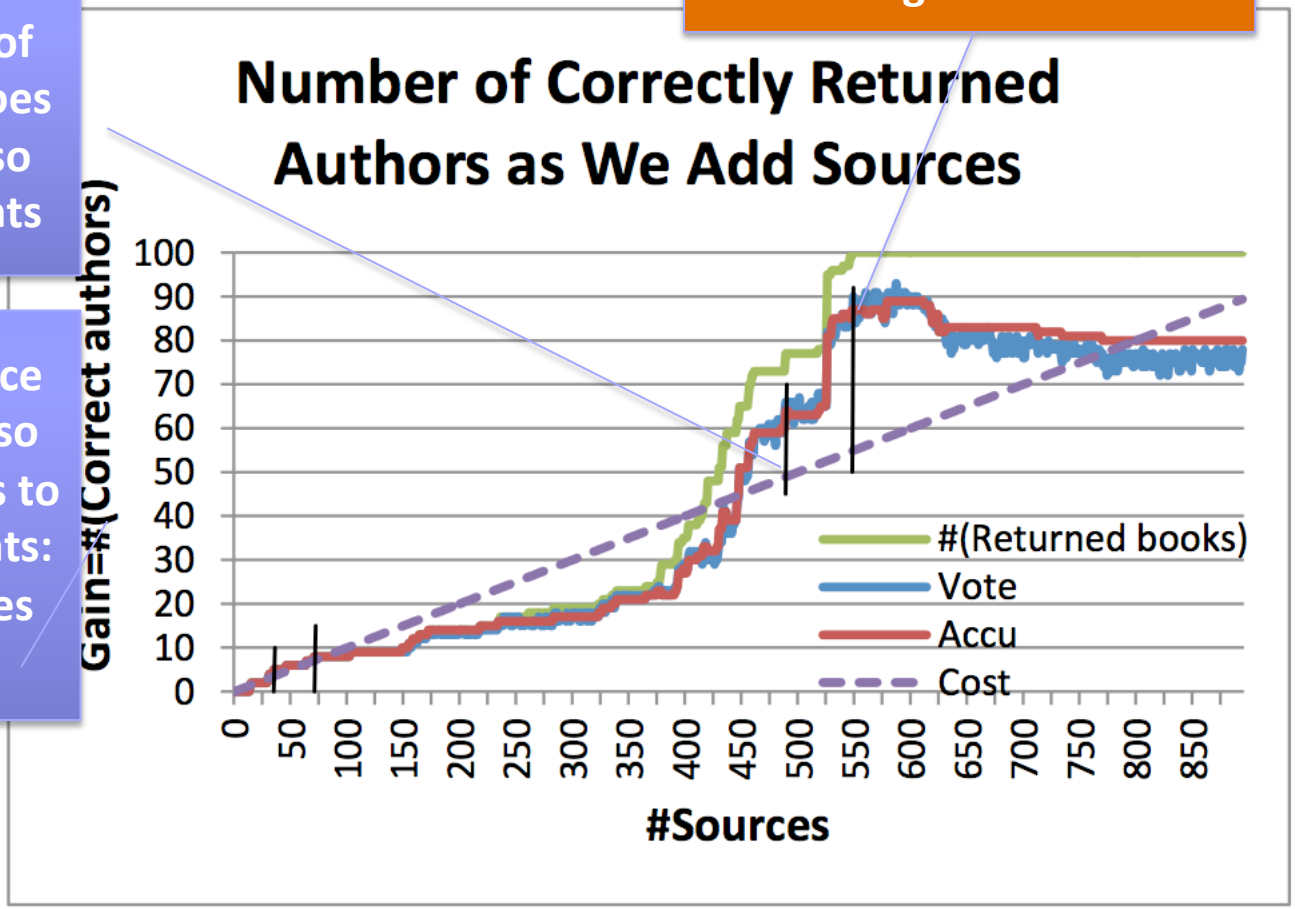


Marginalism for Source Selection

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



CS books from AbeBooks.com

Future Work

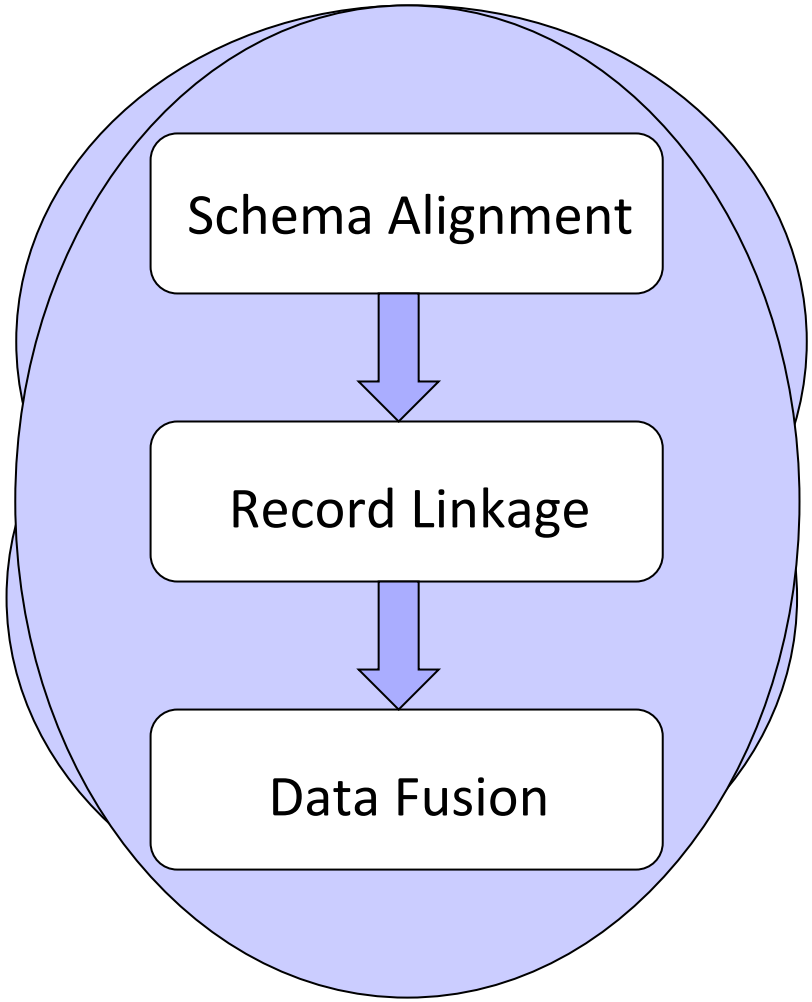
- Reconsider the architecture

Data warehousing

Virtual integration

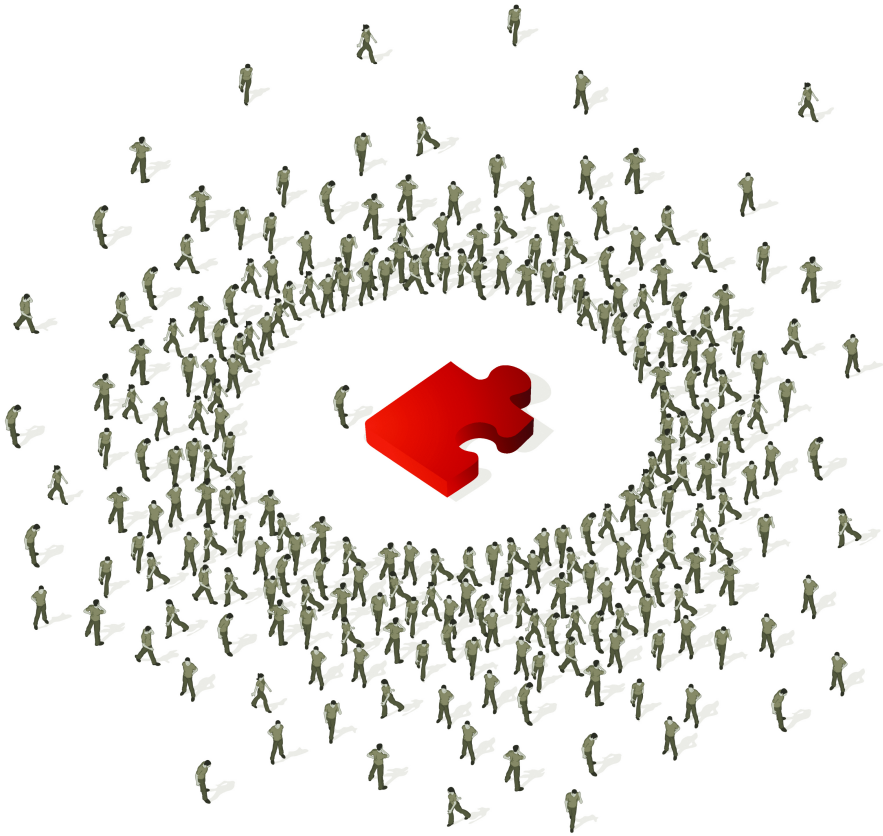
Future Work

- Combining different components



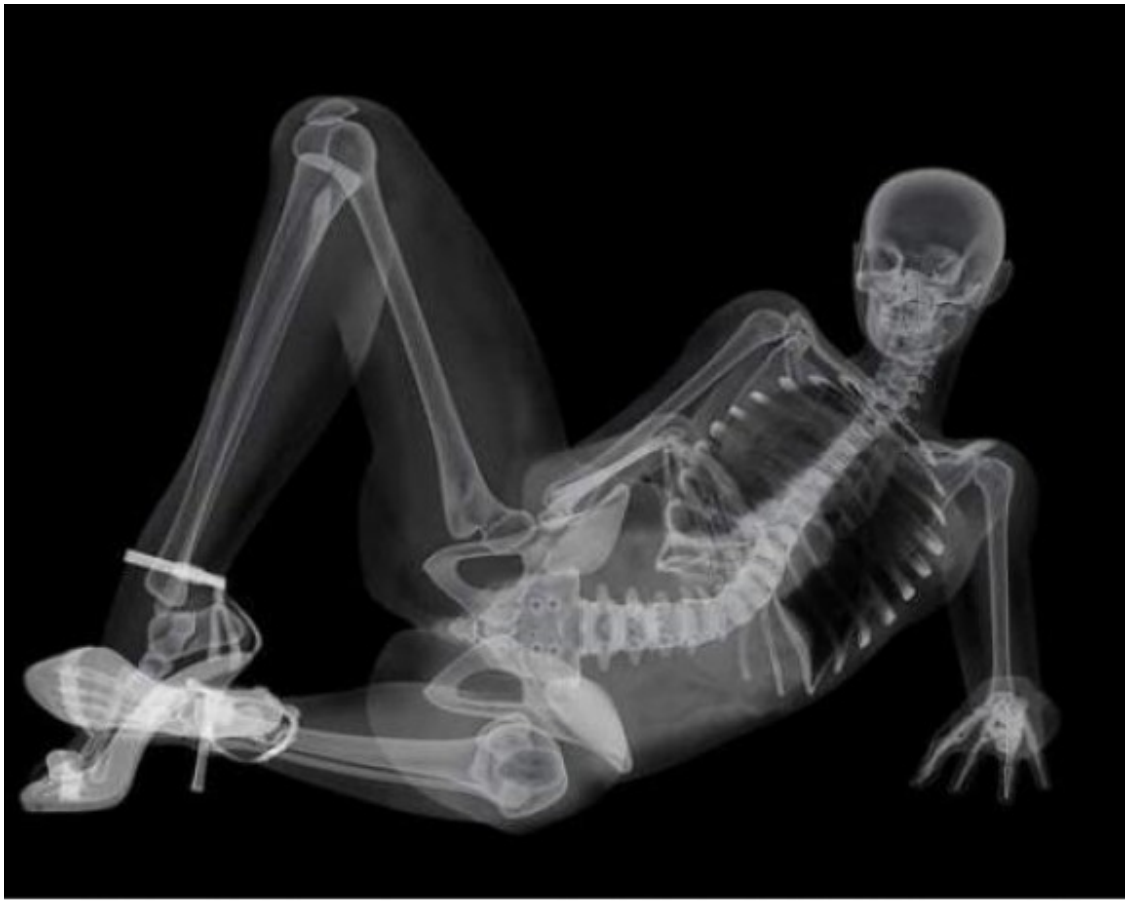
Future Work

- Active integration by crowdsourcing



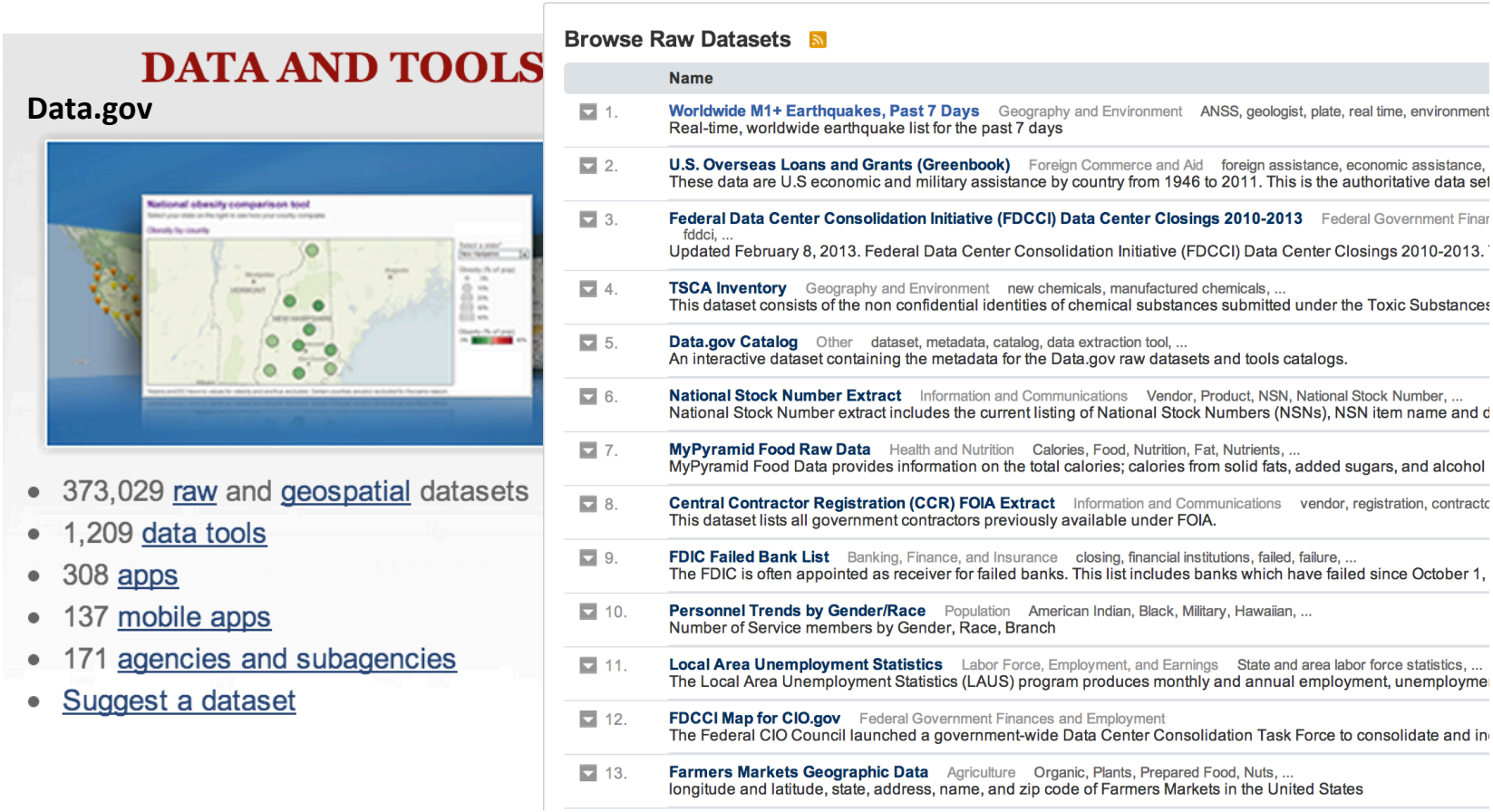
Future Work

- Quality diagnosis



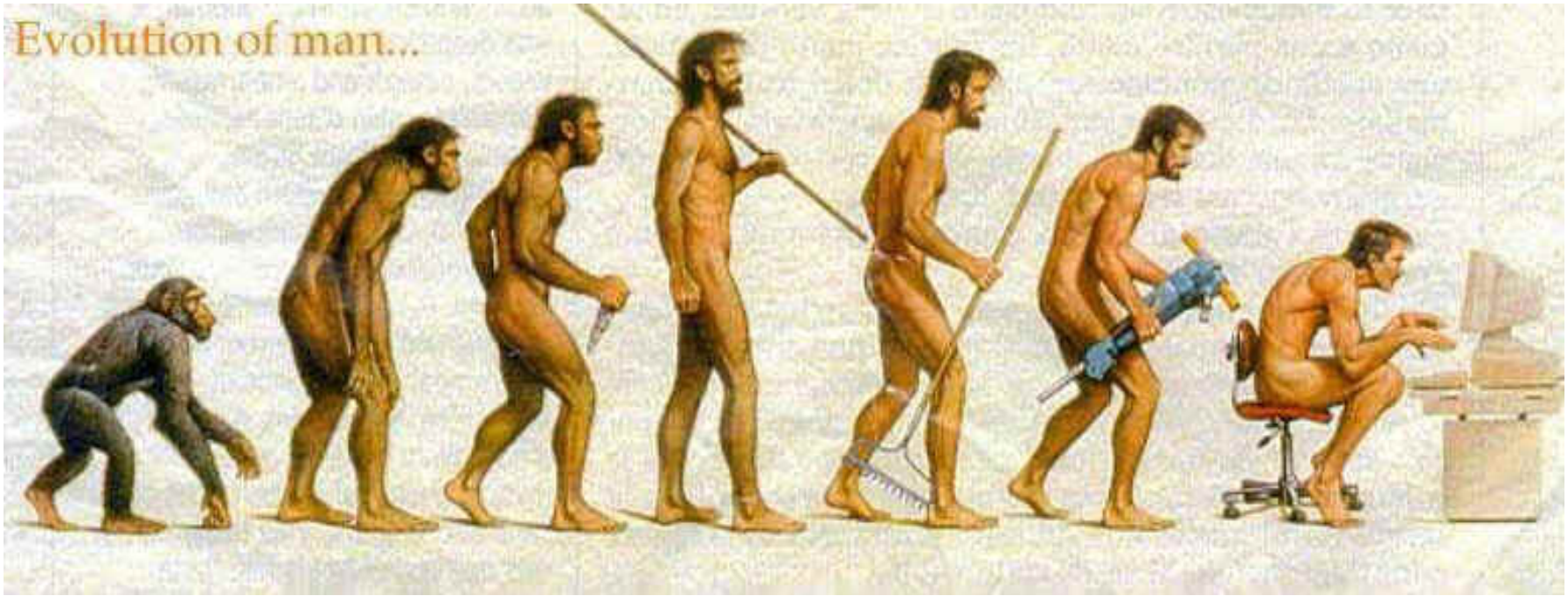
Future Work

- Source exploration tool



Future 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
- [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

References

- [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
- [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)

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)

- [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)

References

- [QAH+13] Guo-Jun Qi, Charu C. Aggarwal, Jiawei Han, Thomas S. Huang: Mining collective intelligence in diverse groups. WWW 2013: 1041-1052

- [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)

- [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)

- [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

- [YHY08] 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)