Informationsintegration: Stand der Technik

This presentation is a synopsis of the excellent ICDE 2013-tutorial on big data integration by Xin Luna Dong and Divesh Srivastava:

A Small Tutorial on Big Data Integration

Xin Luna Dong (Google Inc.)
Divesh Srivastava (AT&T Labs-Research)

http://www.research.att.com/~divesh/papers/bdi-icde2013.pptx

Future Challenge:
Big Data Integration!

What is "Big Data Integration?"

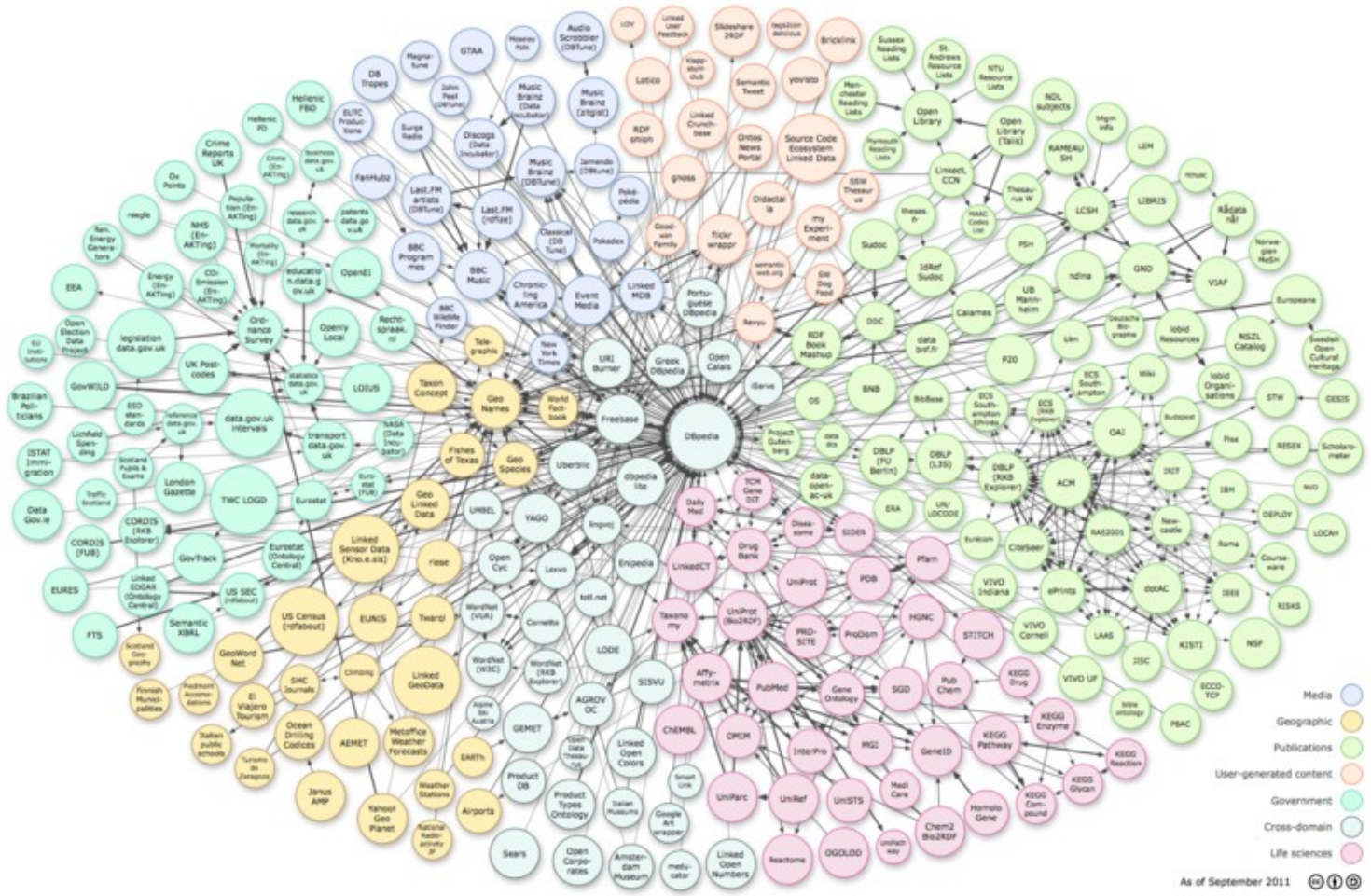
- Big data integration = Big data + data integration

- Data integration: easy access to multiple data sources [DHI12]
  - Virtual: mediated schema, query redirection, link + fuse answers
  - Warehouse: materialized data, easy querying, consistency issues

- Big data in the context of data integration: all about the V's ☺
  - Size: large **volume** of sources, changing at high **velocity**
  - Complexity: huge **variety** of sources, of questionable **veracity**

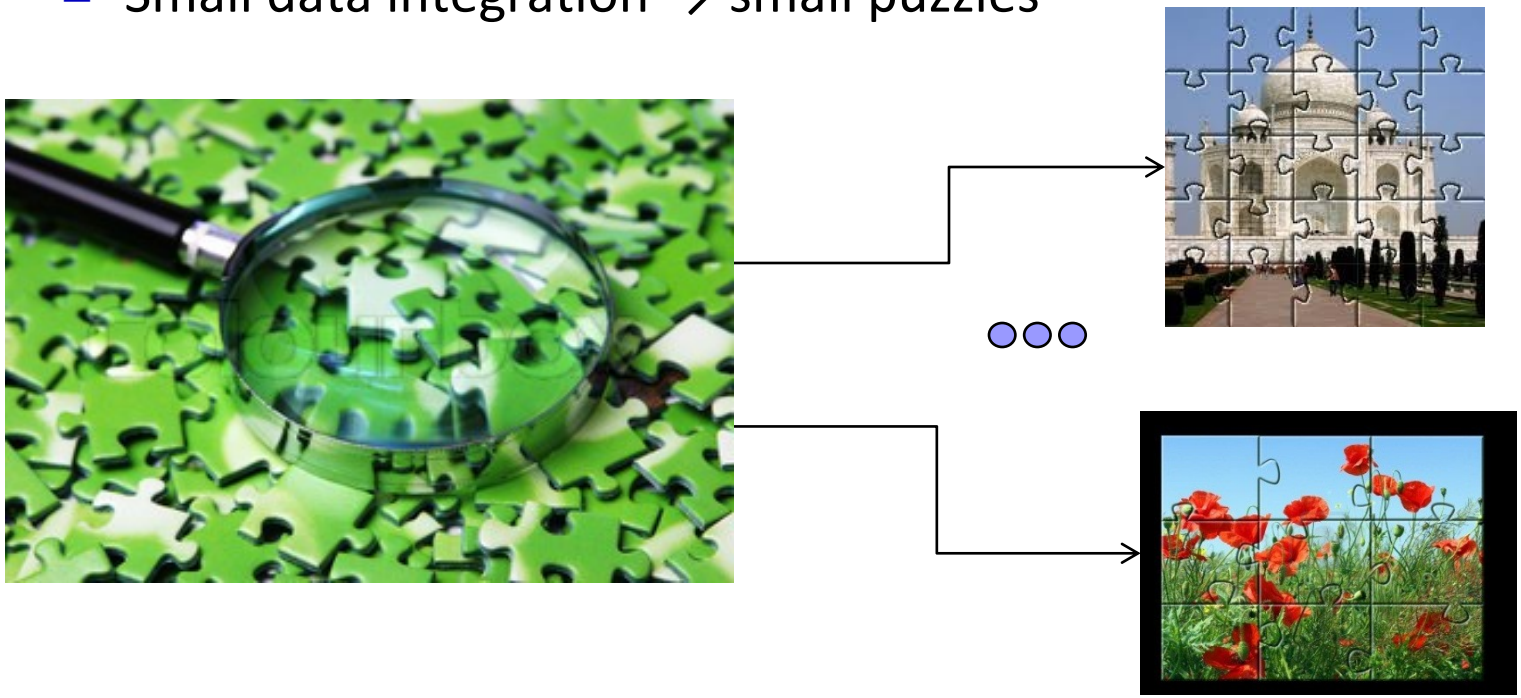
Why Do We Need "Big Data Integration?"

- Reasoning over linked data



"Small" Data Integration: Why is it Hard?

- Data integration = solving lots of jigsaw puzzles
  - Each jigsaw puzzle (e.g., Taj Mahal) is an **integrated entity**
  - Each type of puzzle (e.g., flowers) is an **entity domain**
  - Small data integration → small puzzles



BDI: Why is it Challenging?

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

BDI: Why is it Challenging?

- Number of structured sources: **Volume**
  - 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**

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

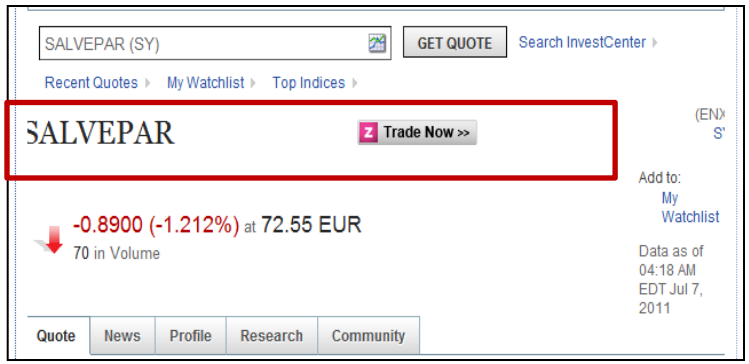
arts

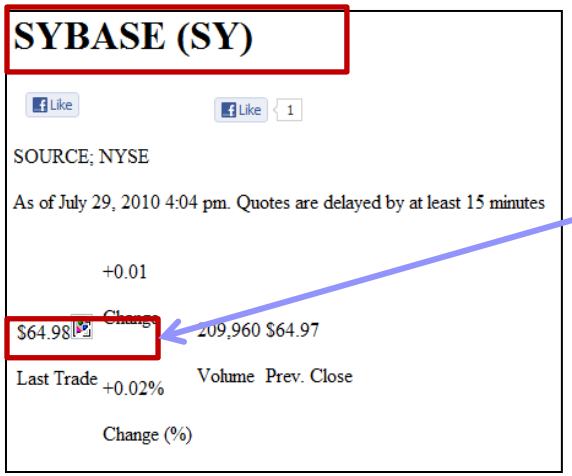
Movement High Renaissance

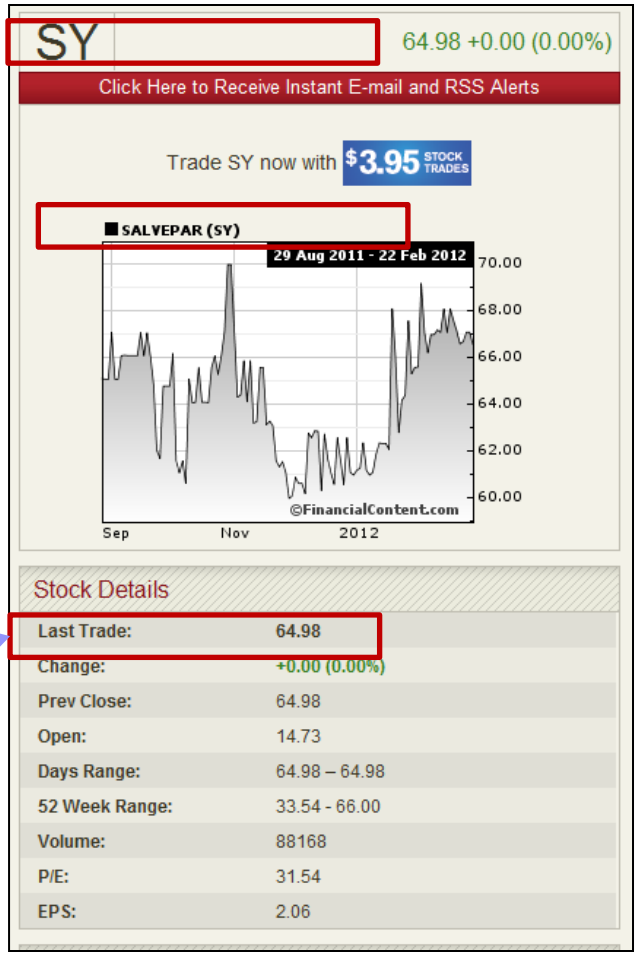
Works *Mona Lisa*
The Last Supper
The Vitruvian Man
Lady with an Ermine

BDI: Why is it Challenging?

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

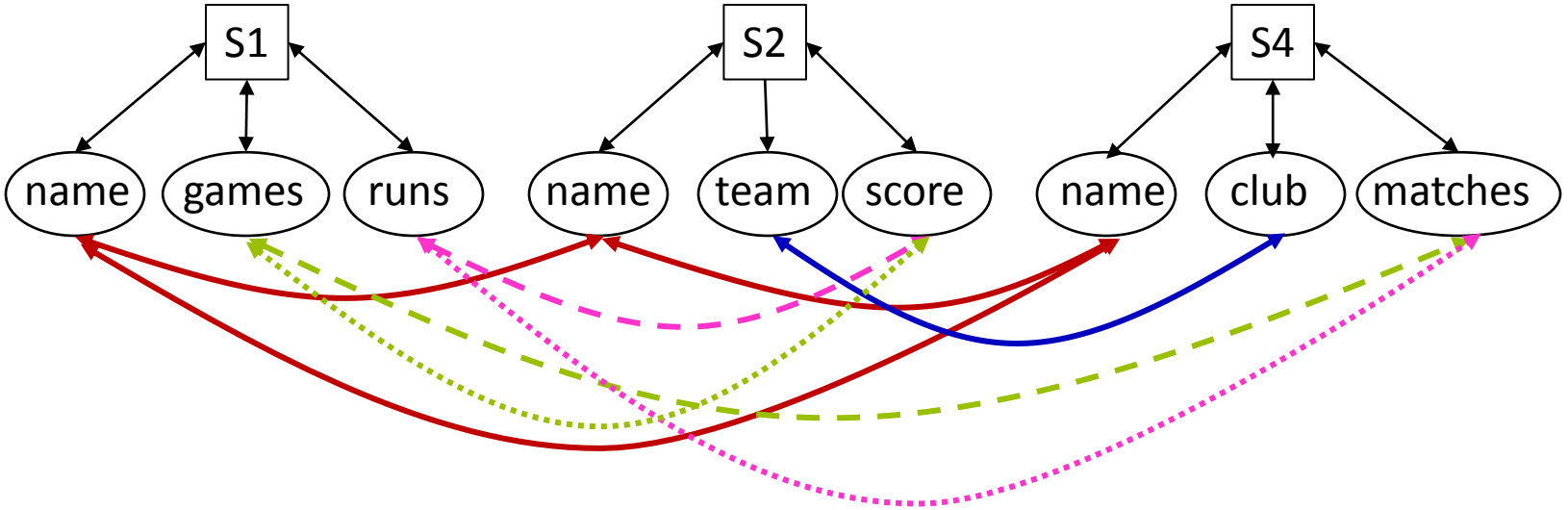






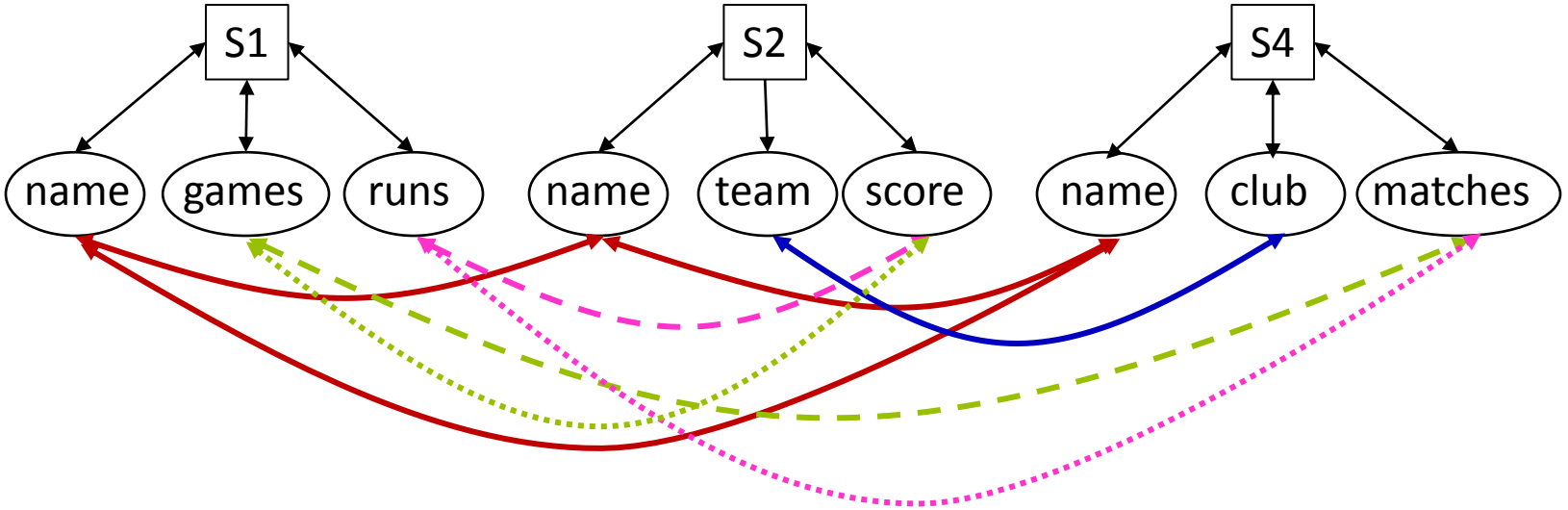
Schema Mapping/Integration

Probabilistic Mediated Schemas [DDH08]



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

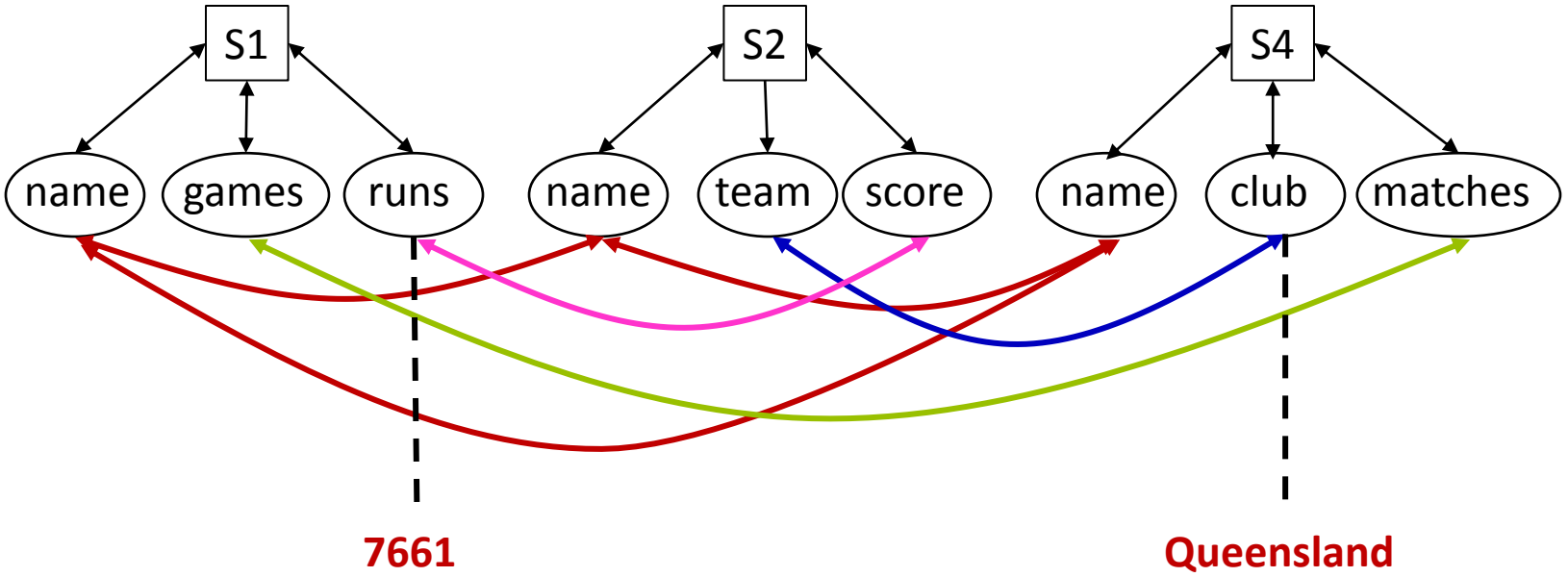
Probabilistic Mediated Schemas [DDH08]



- Example P-mediated schema
  - M1({S1.games, S4.matches}, {S1.runs, S2.score})
  - M2({S1.games, S2.score}, {S1.runs, S4.matches})
  - M = {(M1, 0.6), (M2, 0.2), (M3, 0.1), (M4, 0.1)}

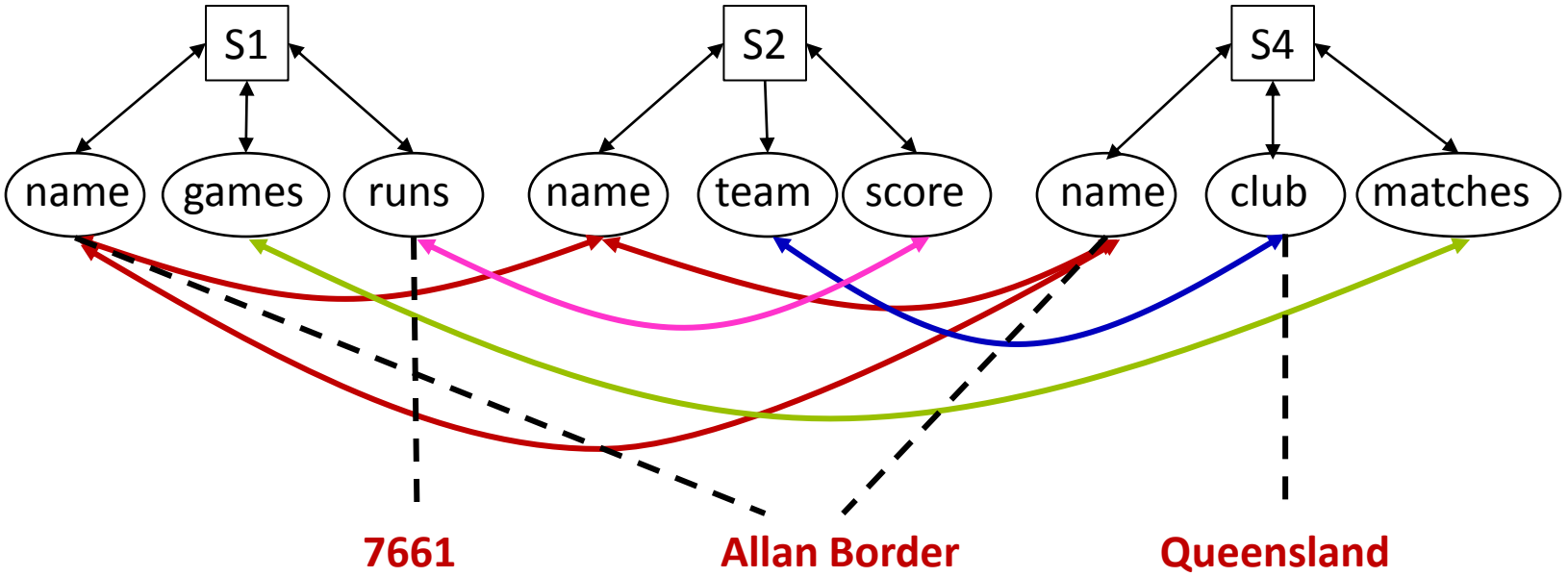
Keyword Search Based Integration [TJM+08]

- Key idea: information need driven integration
  - Search graph: source tables with weighted associations
  - Query keywords: matched to elements in different sources
  - Derive top-k SQL view, using Steiner tree on search graph



Keyword Search Based Integration [TJM+08]

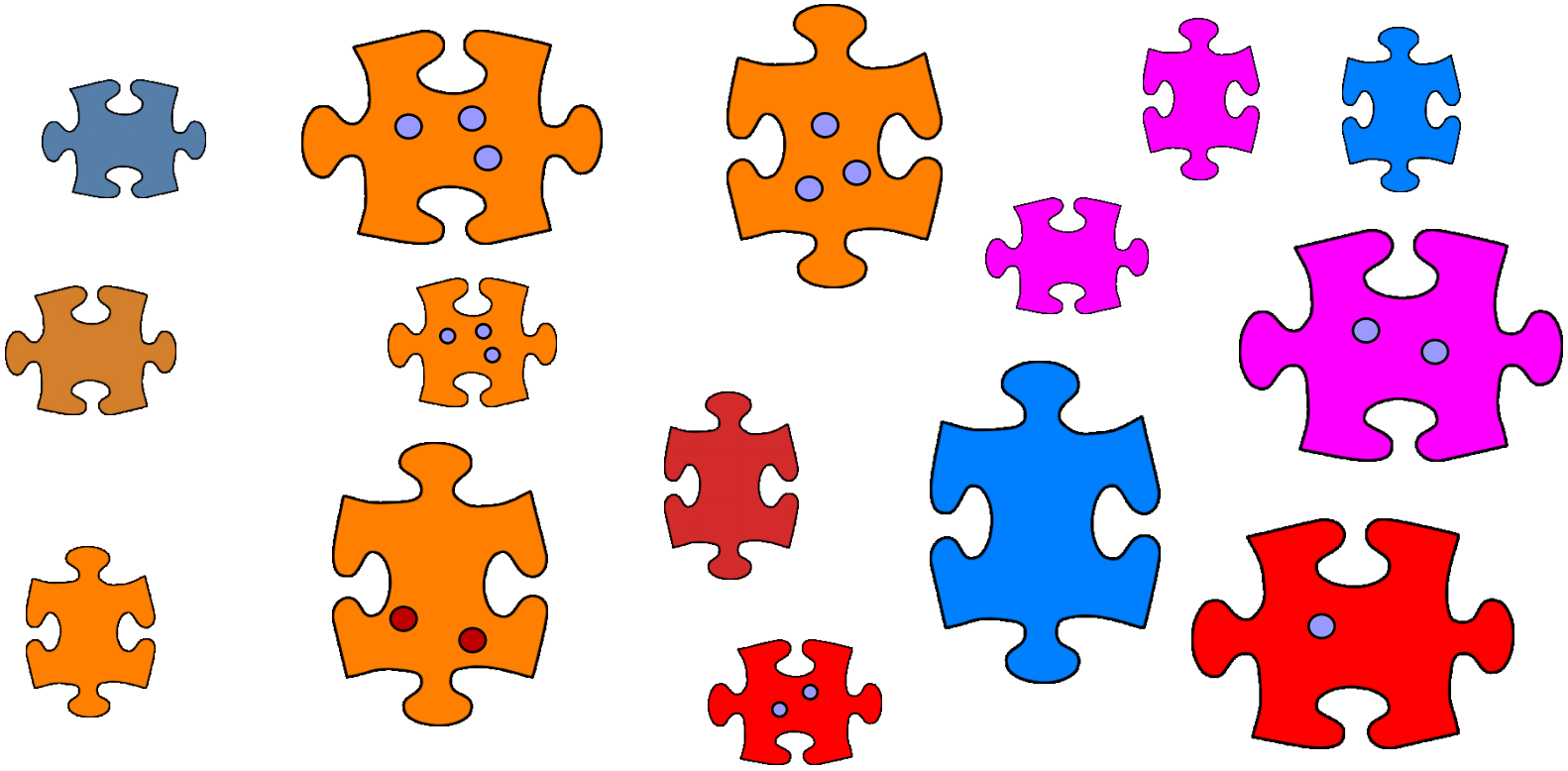
- Key idea: information need driven integration
  - Search graph: source tables with weighted associations
  - Query keywords: matched to elements in different sources
  - Derive top-k SQL view, using Steiner tree on search graph



Record Linkage

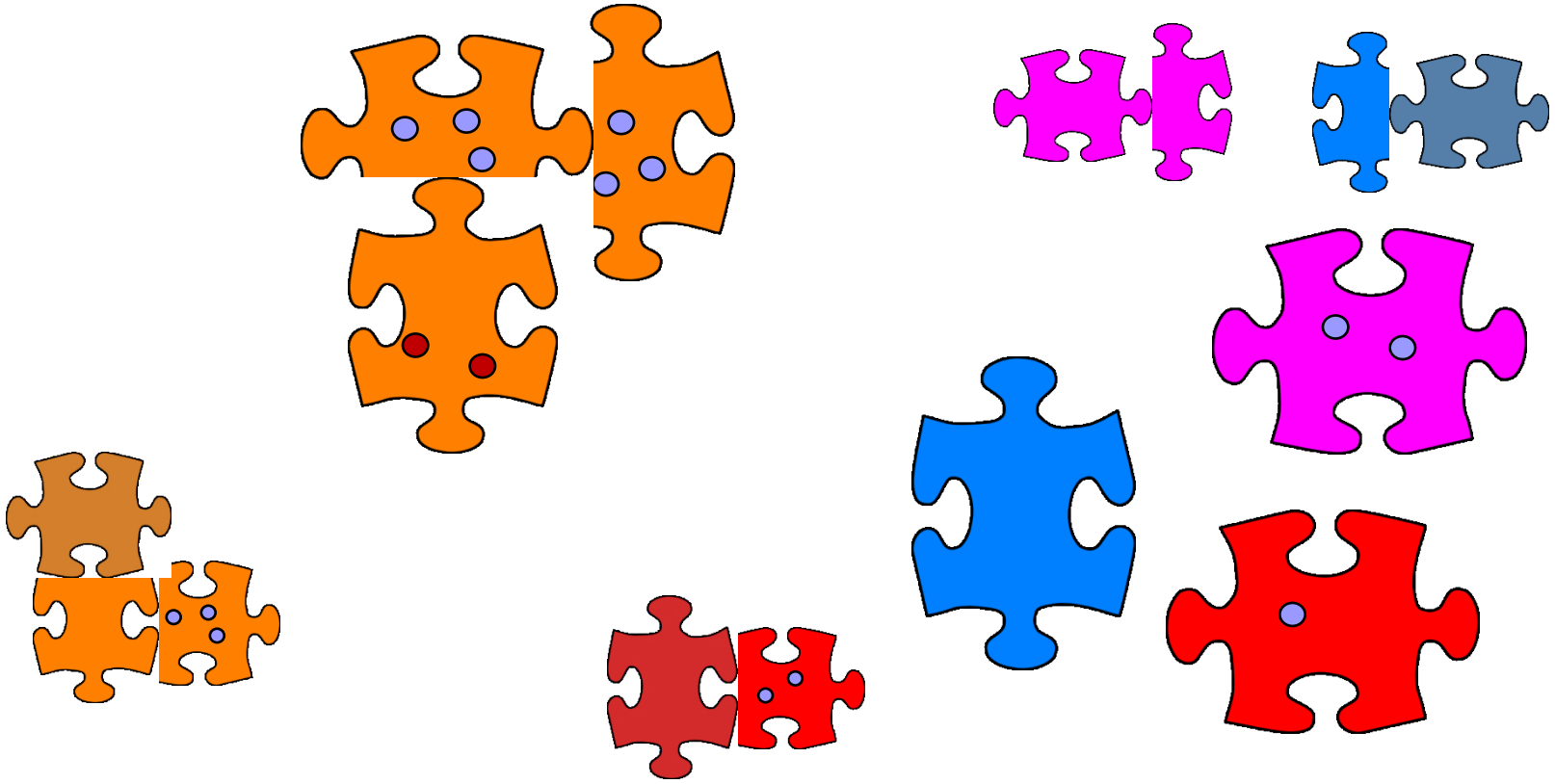
Record Linkage

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



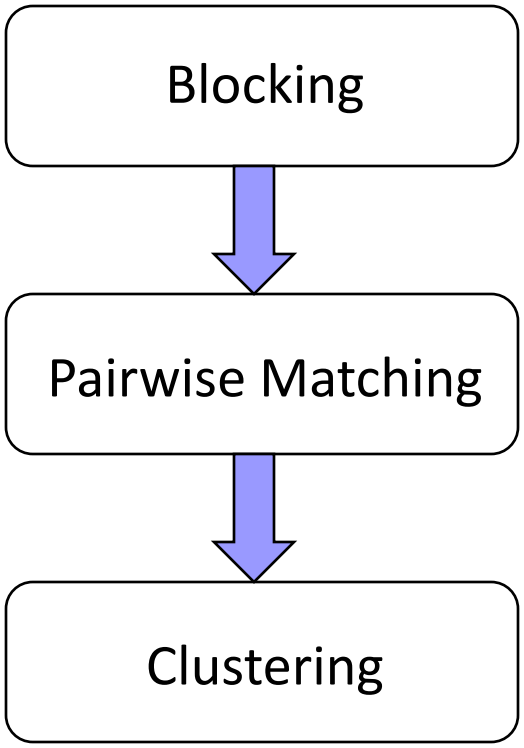
Record Linkage

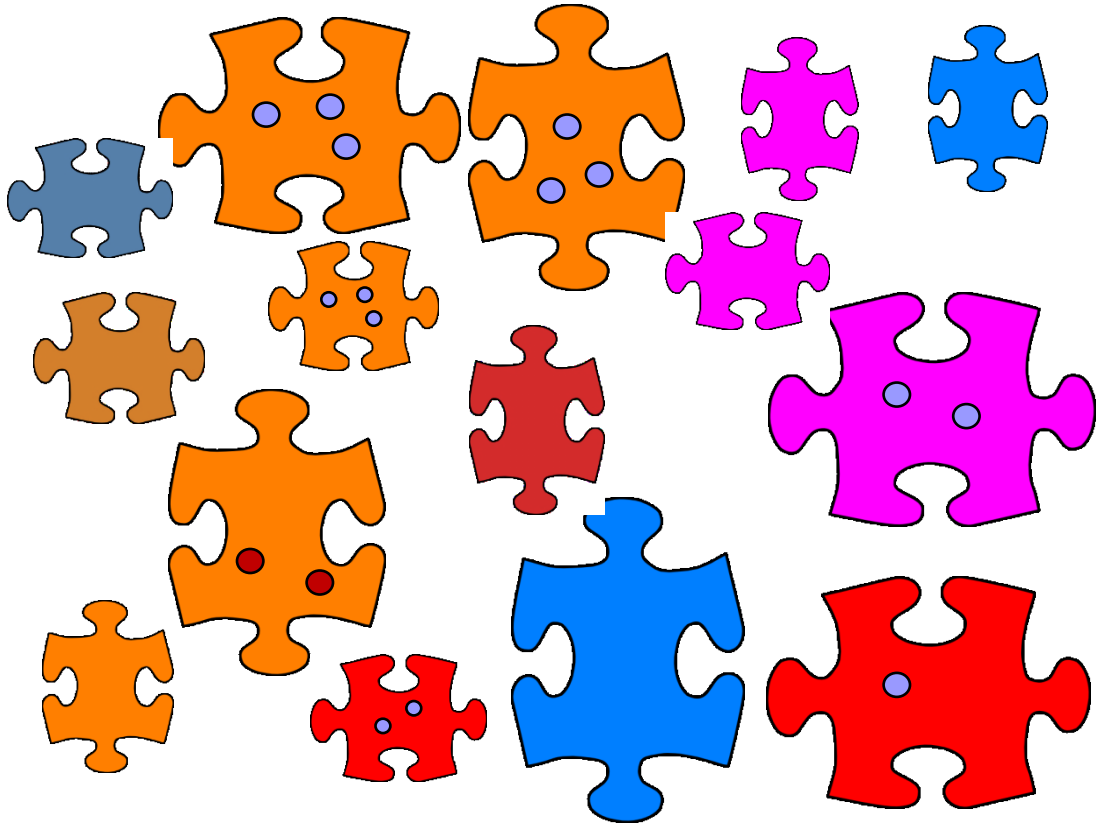
- Matching based on identifying content: color, size



Record Linkage: Three Steps [EIV07, GM12]

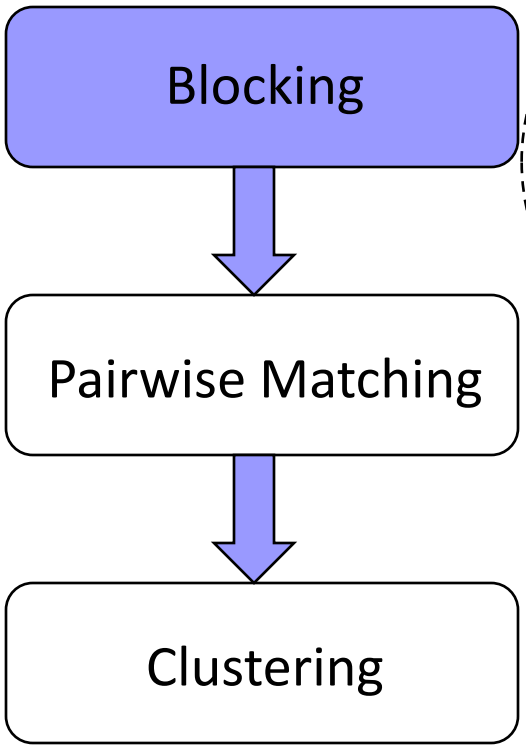
- Record linkage: blocking + pairwise matching + clustering
  - Scalability, similarity, semantics

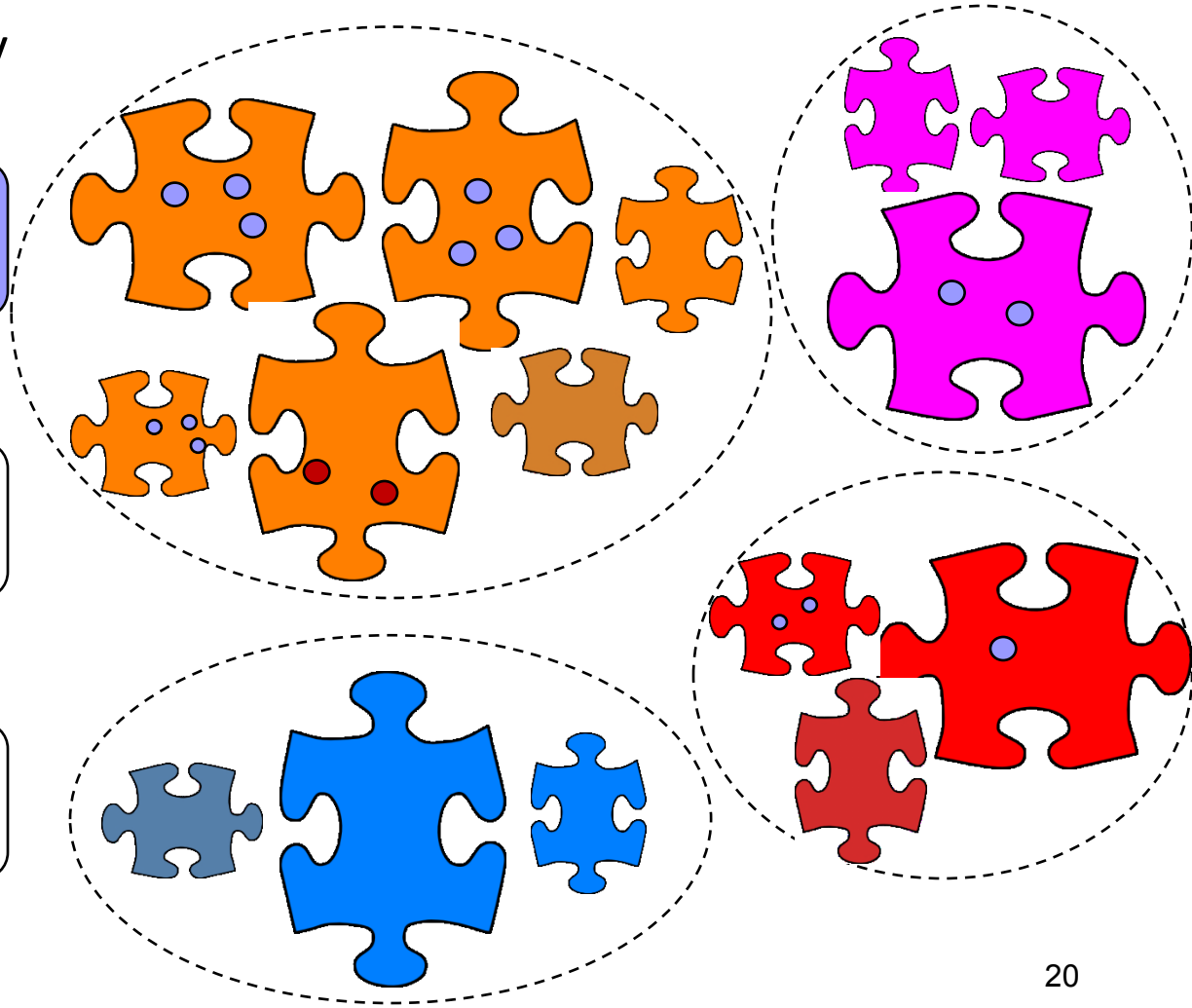




Record Linkage: Three Steps

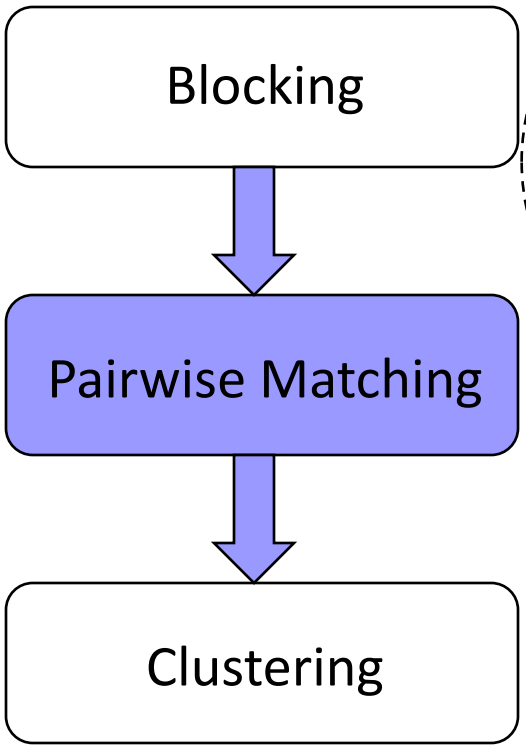
- Blocking: **efficiently** create **small** blocks of **similar** records
  - Ensures scalability

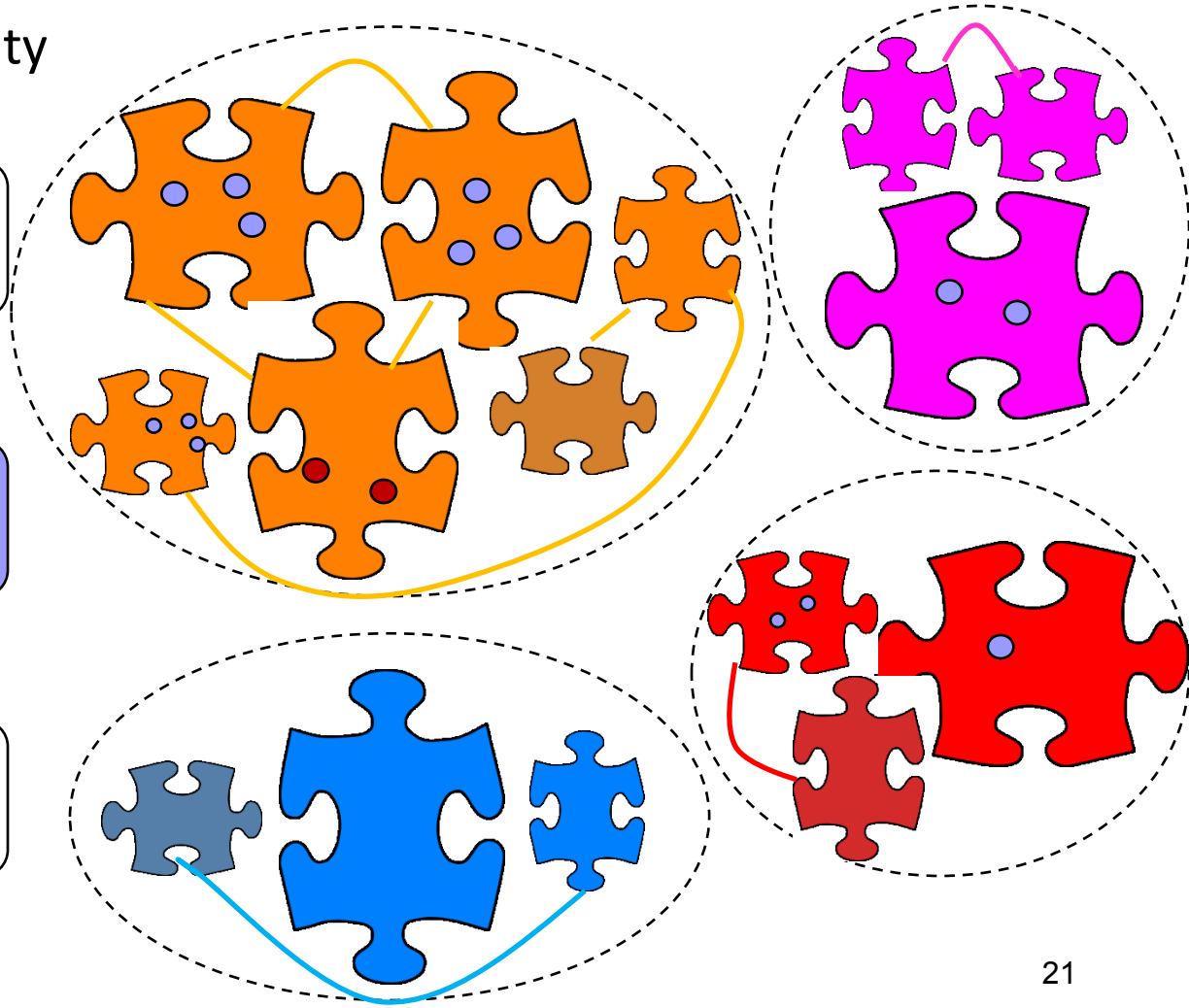




Record Linkage: Three Steps

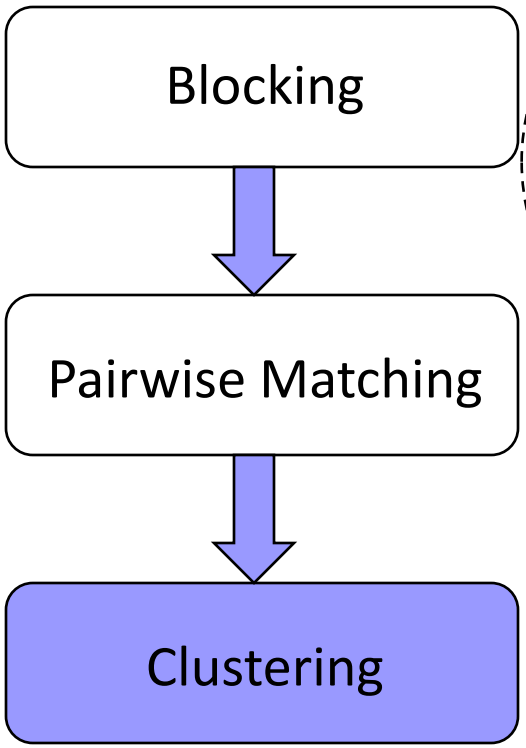
- Pairwise matching: compares all record pairs in a block
  - Computes similarity

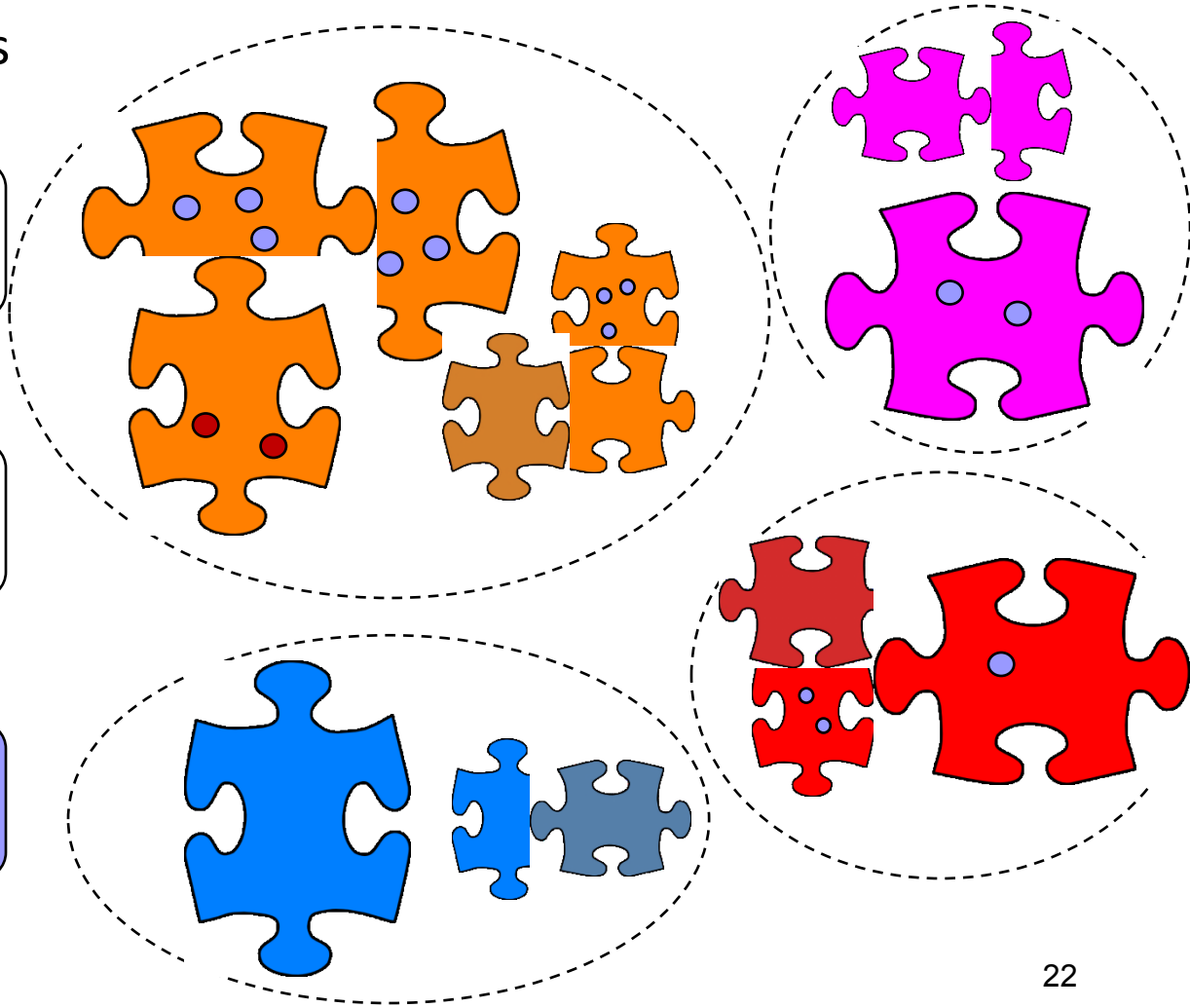




Record Linkage: Three Steps

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





Record Linkage Using MapReduce

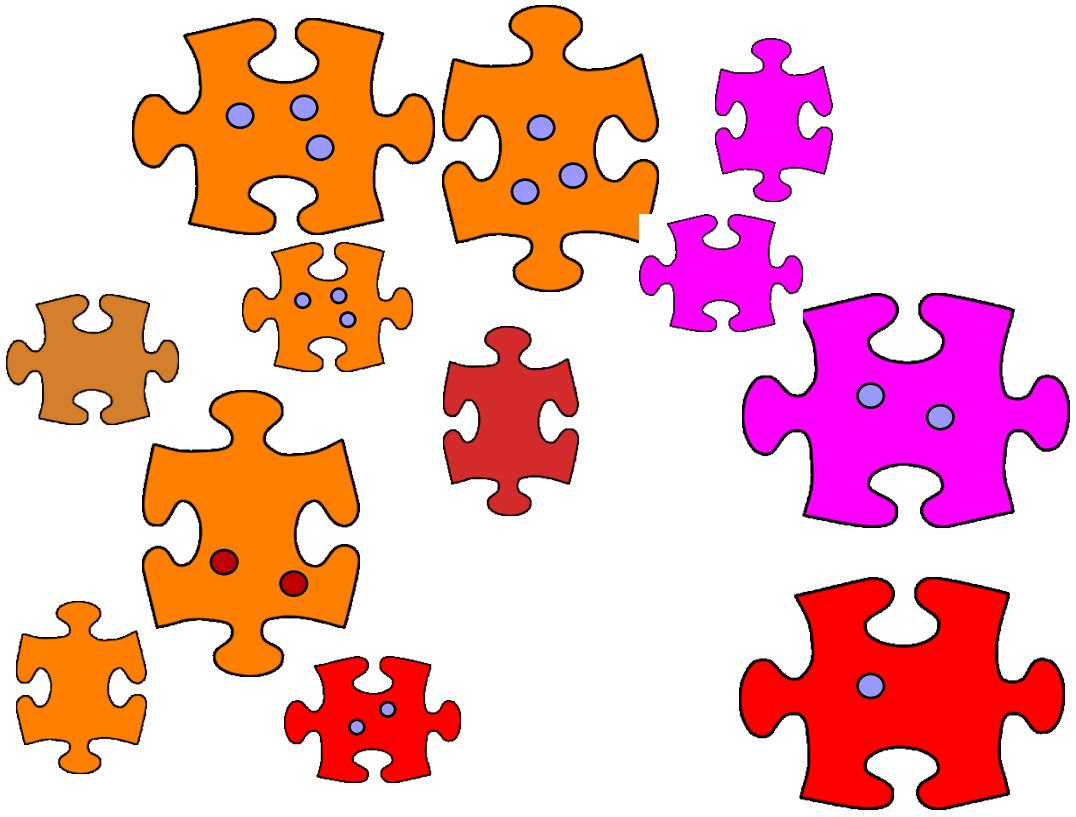
Record Linkage Using MapReduce [KTR12]

- 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

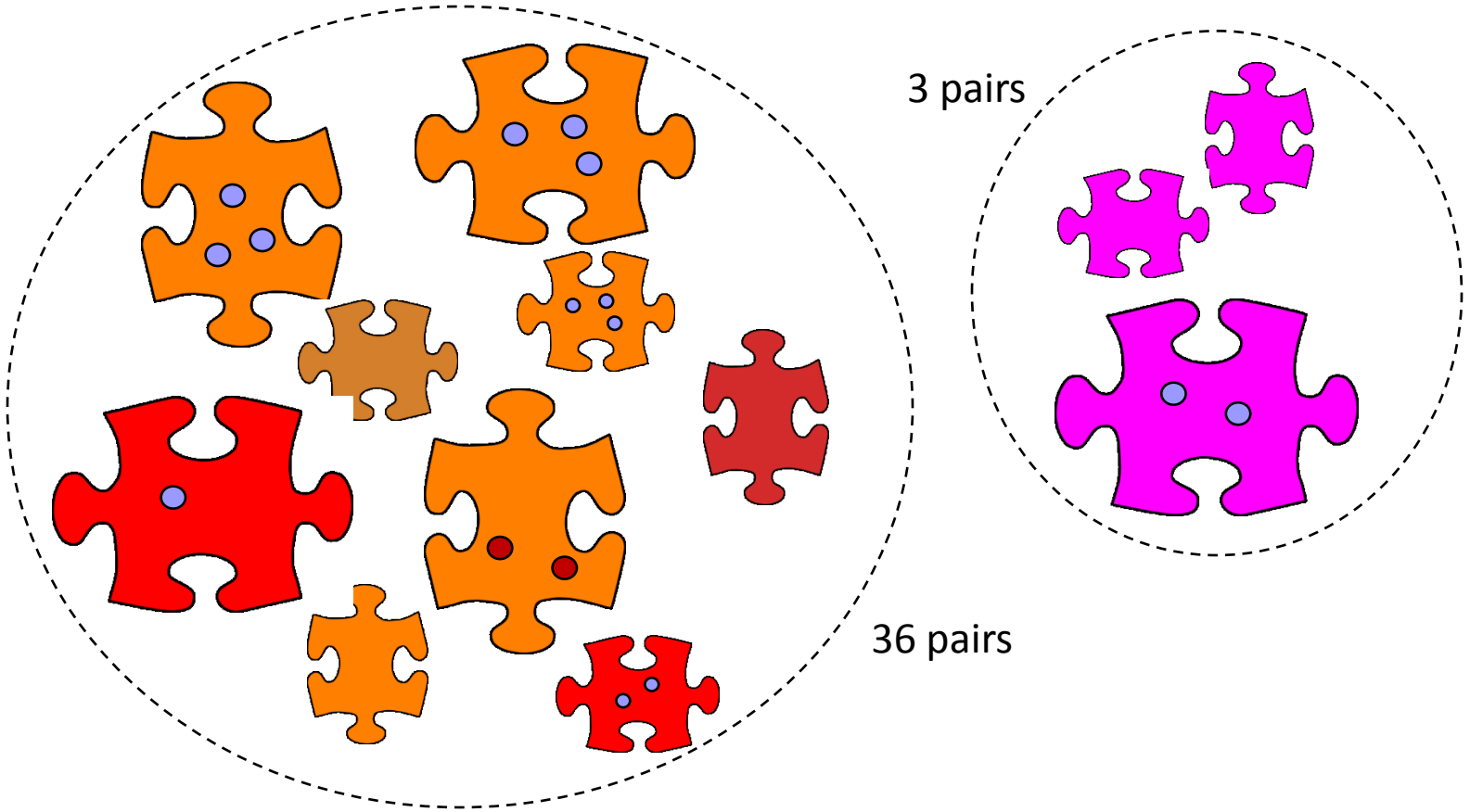
Record Linkage Using MapReduce

- Challenge: data skew → unbalanced workload



Record Linkage Using MapReduce

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



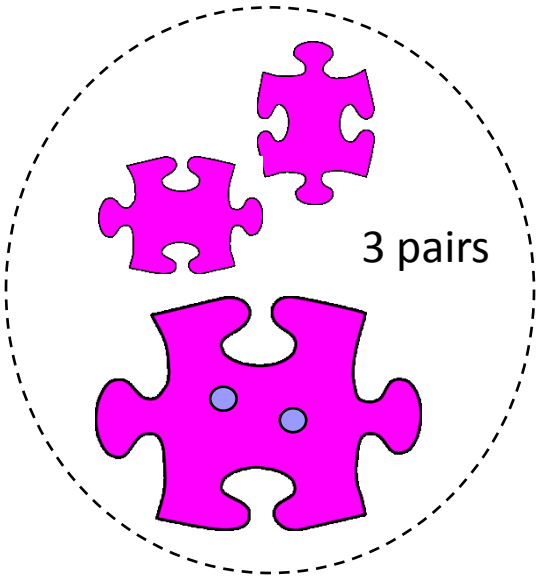
Load Balancing

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

- Load balancing strategy:
  - BlockSplit: split large blocks into sub-blocks

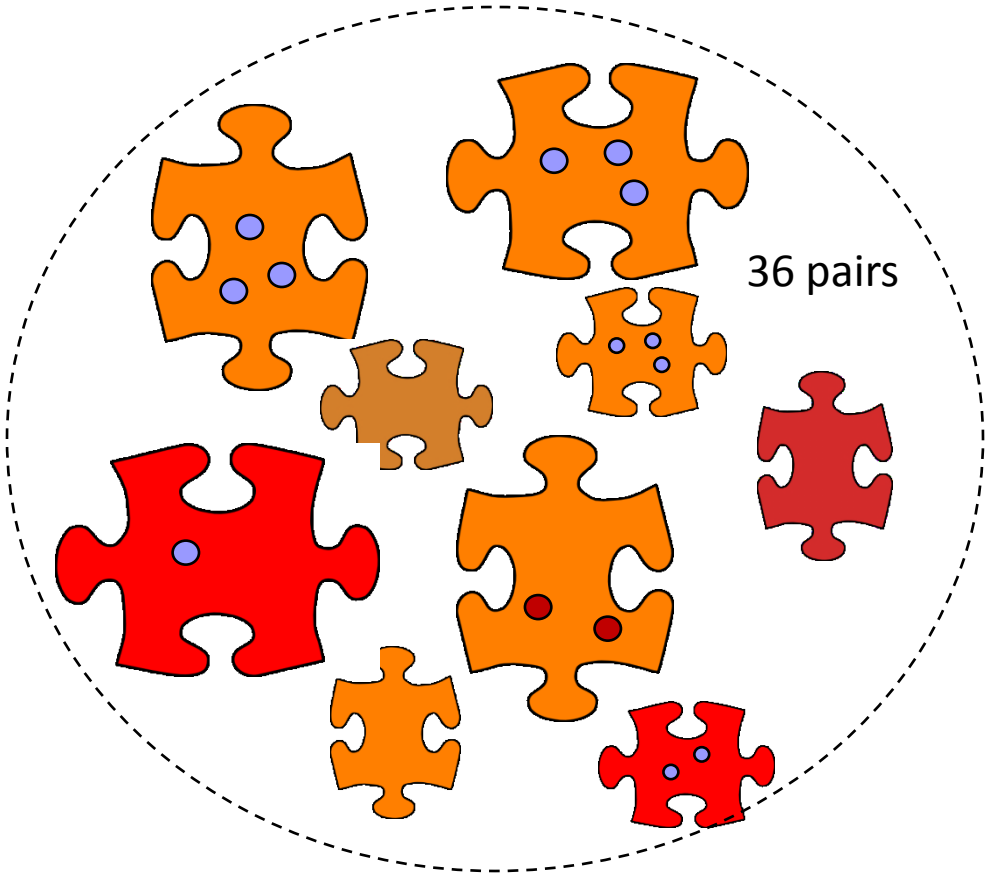
Load Balancing: BlockSplit

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



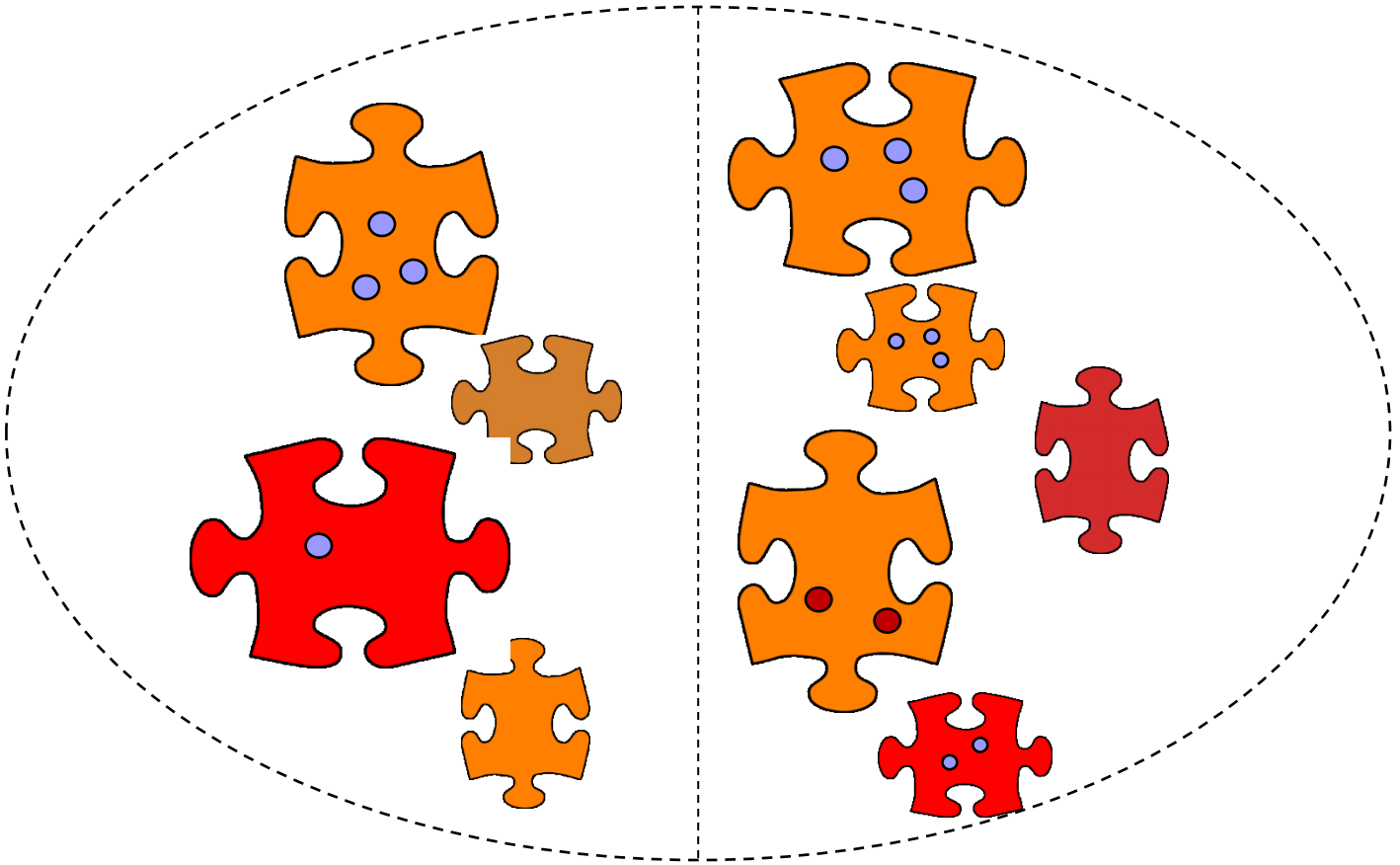
Load Balancing: BlockSplit

- Large blocks: split into multiple sub-blocks



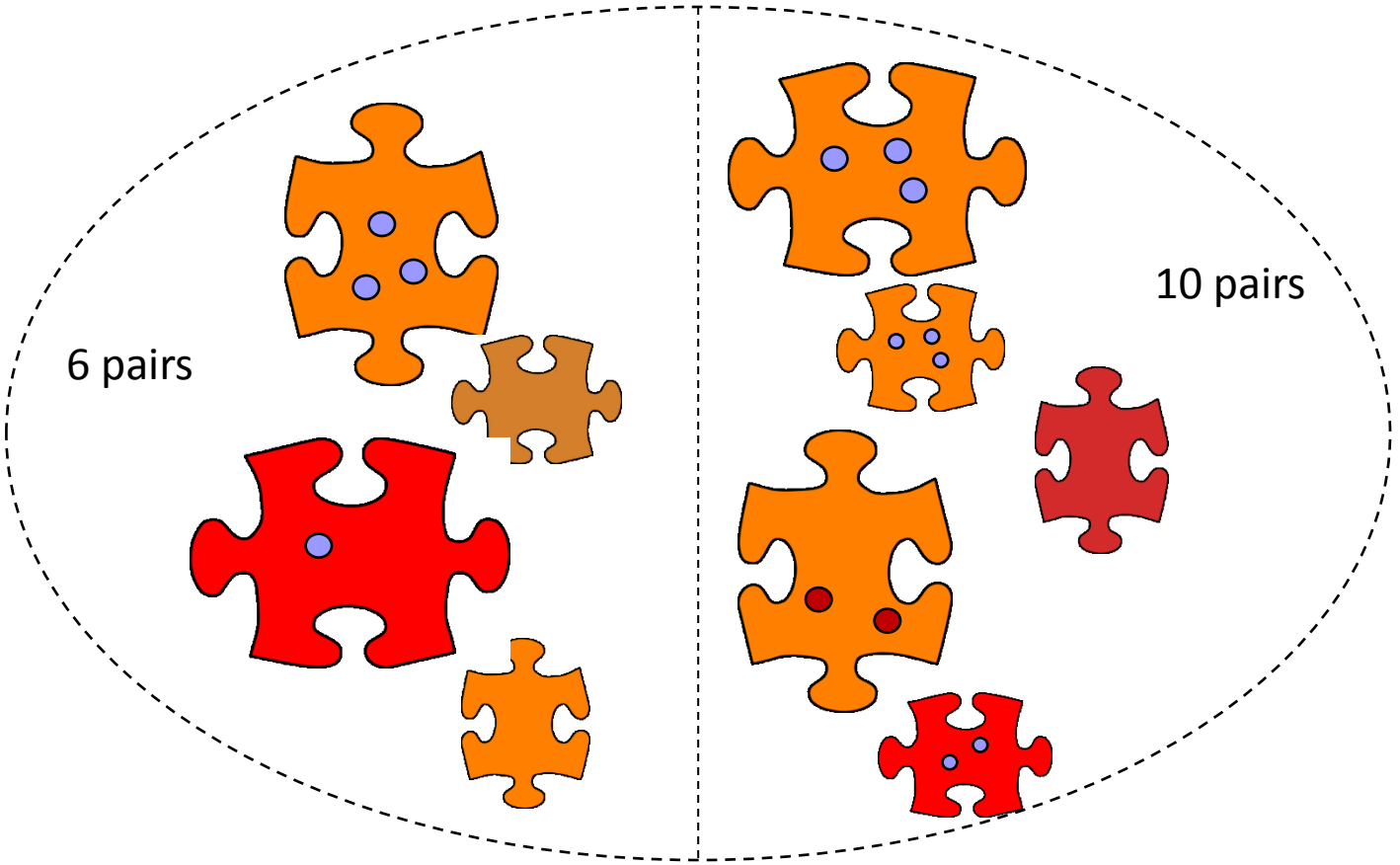
Load Balancing: BlockSplit

- Large blocks: split into multiple sub-blocks



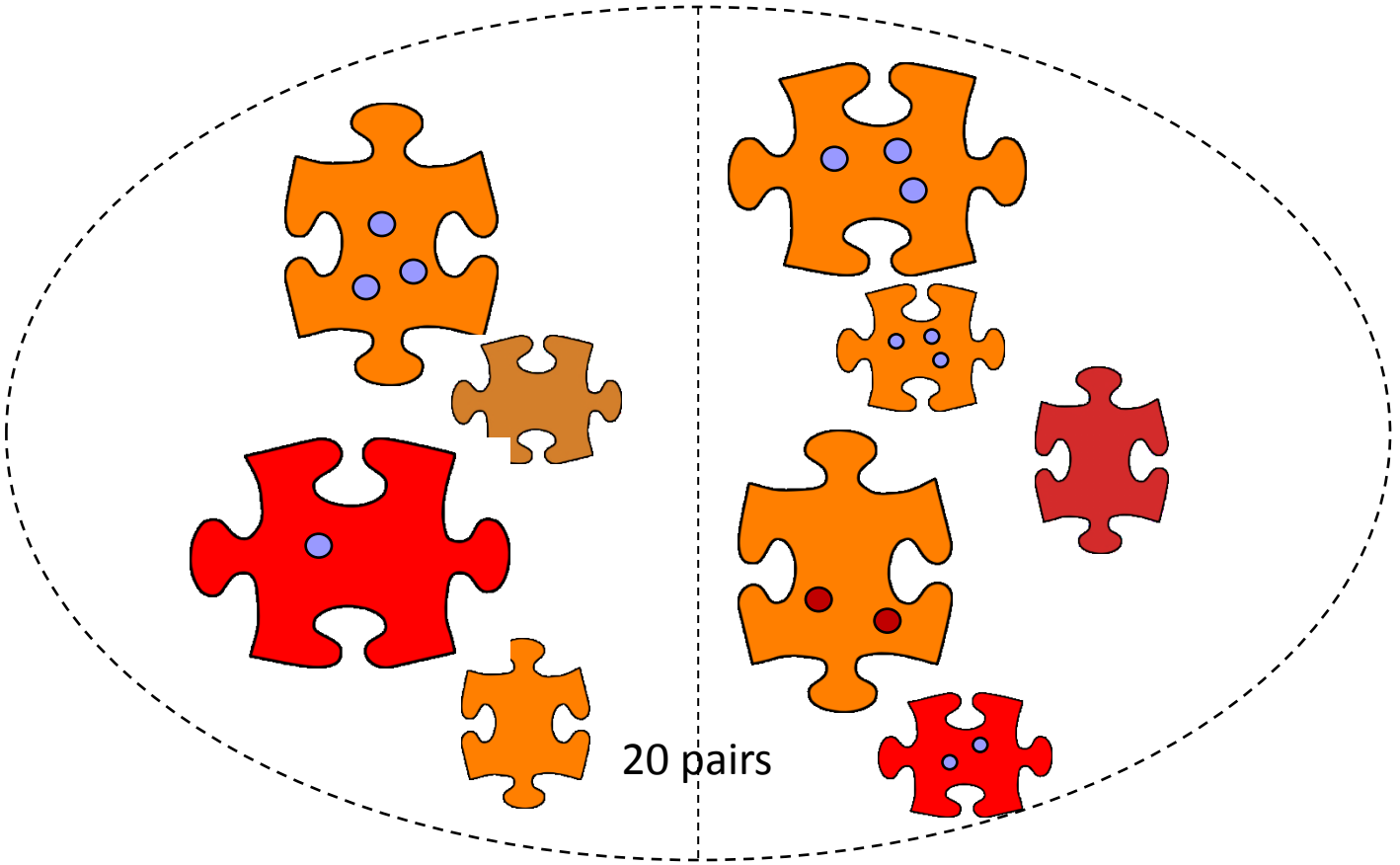
Load Balancing: BlockSplit

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



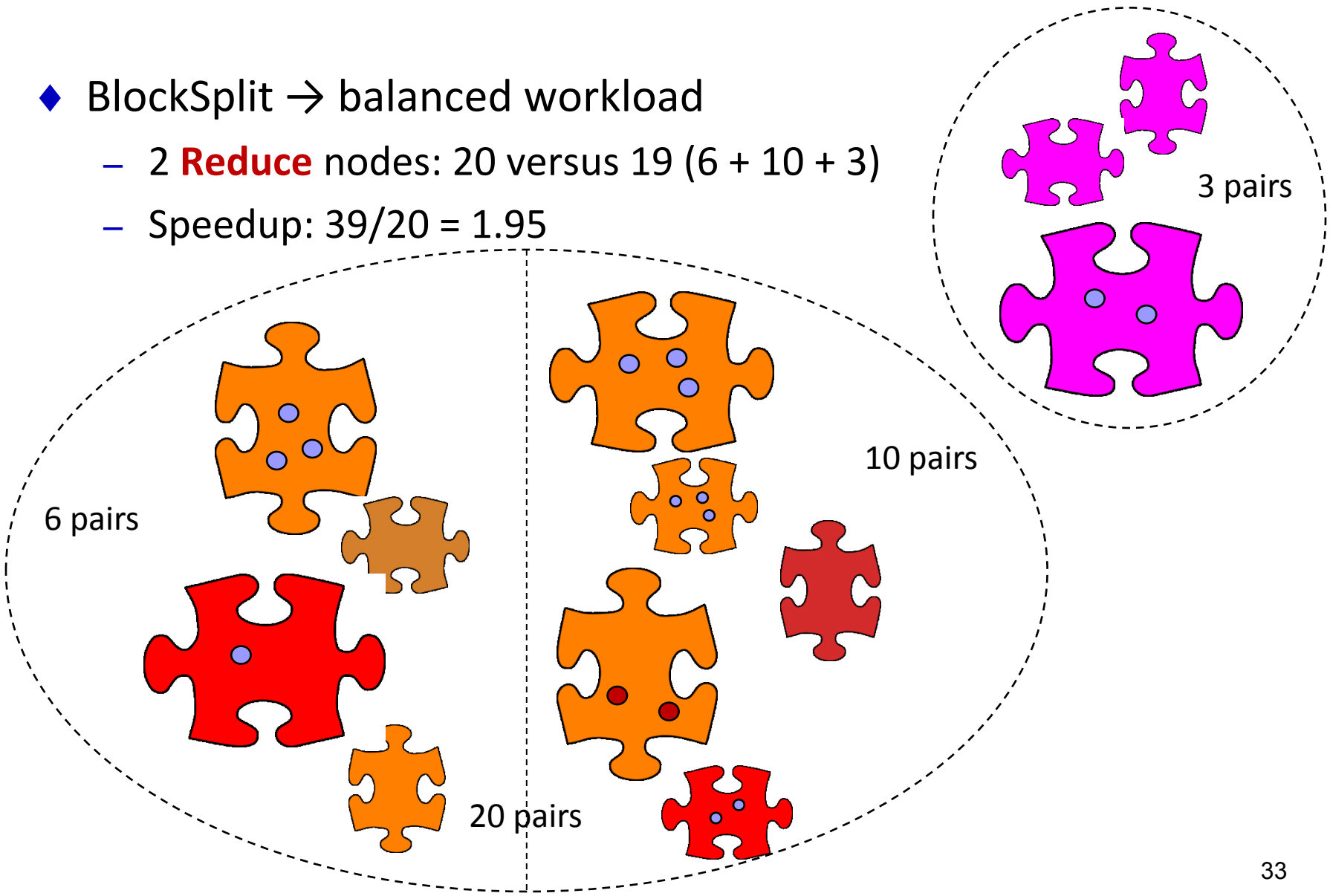
Load Balancing: BlockSplit

- 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



Integrating structured and unstructured data

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
  - Many similar but different product offers, specifications

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]

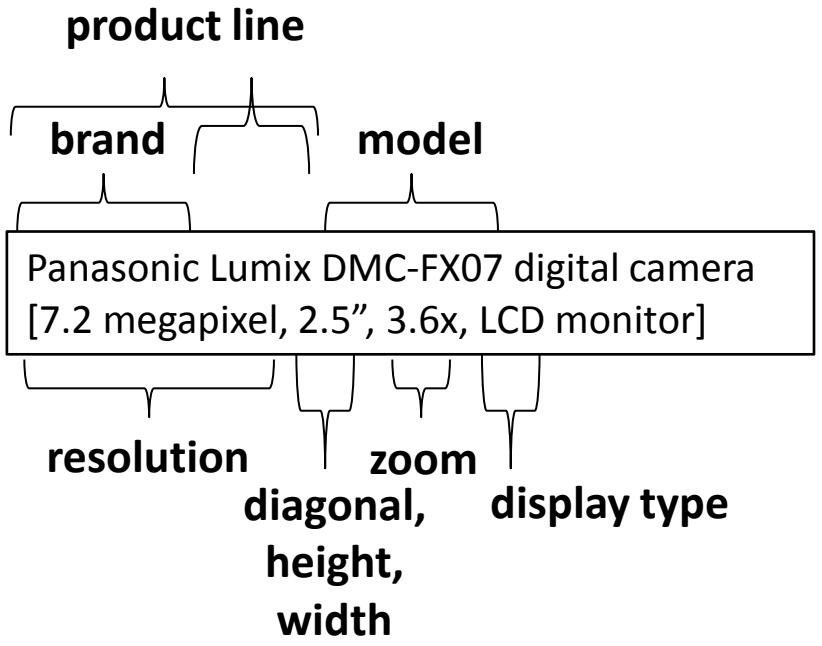
Panasonic DMC-FX07EB digital camera silver

Lumix FX07EB-S, 7.2MP

Structured + Unstructured Data

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

- Semantic parse of offers: tagging, 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, 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	Panasonic Lumix DMC-FX07 digital camera [7.2 megapixel, 2.5", 3.6x, LCD monitor]
brand model resolution zoom	Panasonic DMC-FX07 7.2 megapixel 3.6x	Panasonic Lumix DMC-FX07 digital camera [7.2 megapixel, 2.5", 3.6x, LCD monitor]

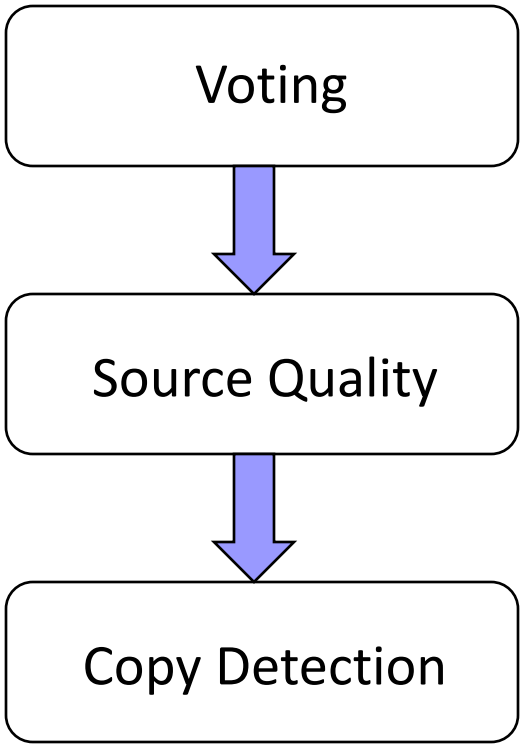
Structured + Unstructured Data

- 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 offer into product category
  - Blocking 2: identify candidates with ≥ 1 high weighted feature

Data Fusion

Data Fusion: Three Components [DBS09a]

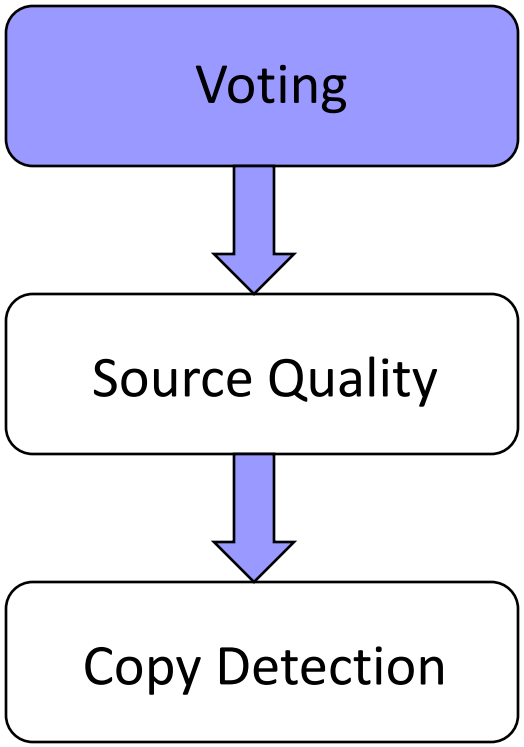
- Data fusion: voting + source quality + copy detection
  - Resolves inconsistency across diversity of sources



	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]

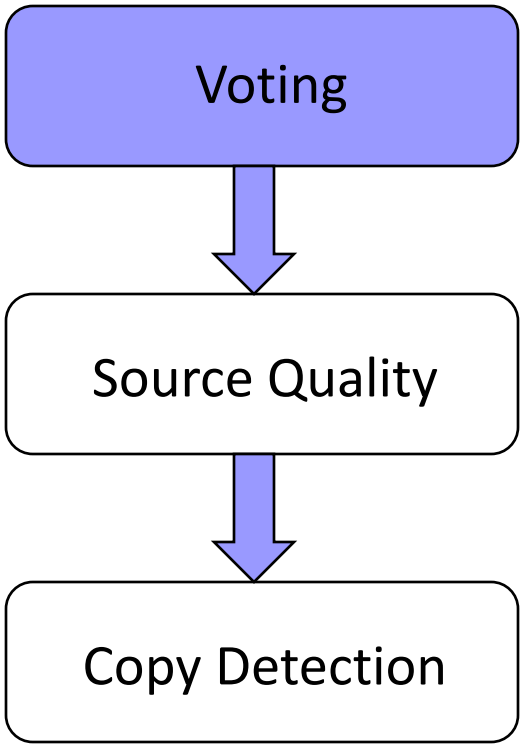
- Data fusion: voting + source quality + copy detection



	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]

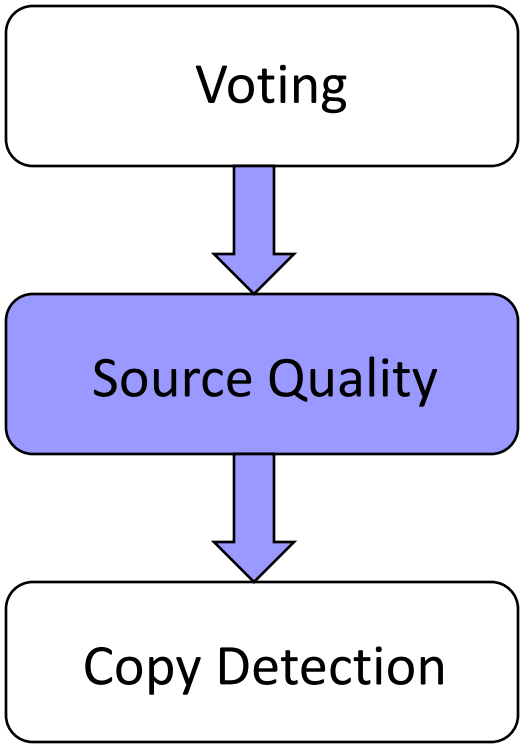
- Data fusion: voting + source quality + copy detection
  - Supports difference of opinion



	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]

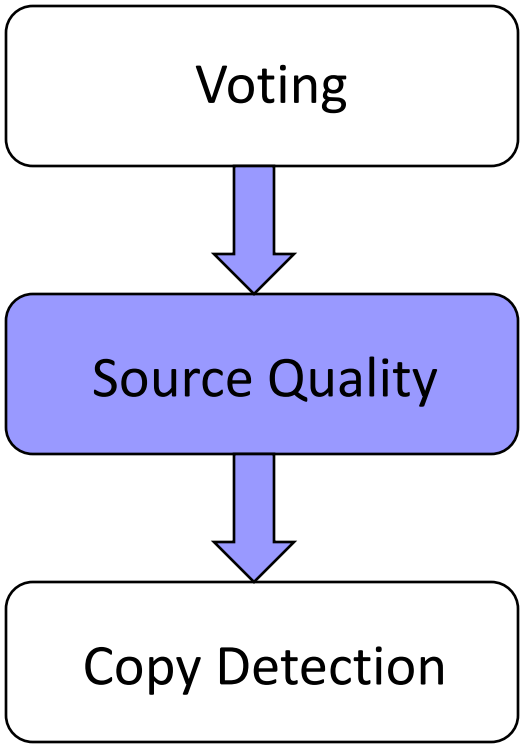
- Data fusion: voting + source quality + copy detection



	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]

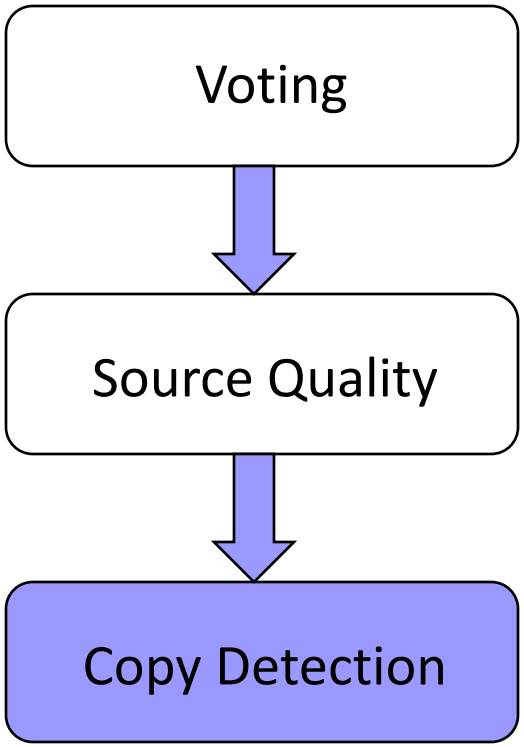
- Data fusion: voting + source quality + copy detection
  - Gives more weight to knowledgeable sources



	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]

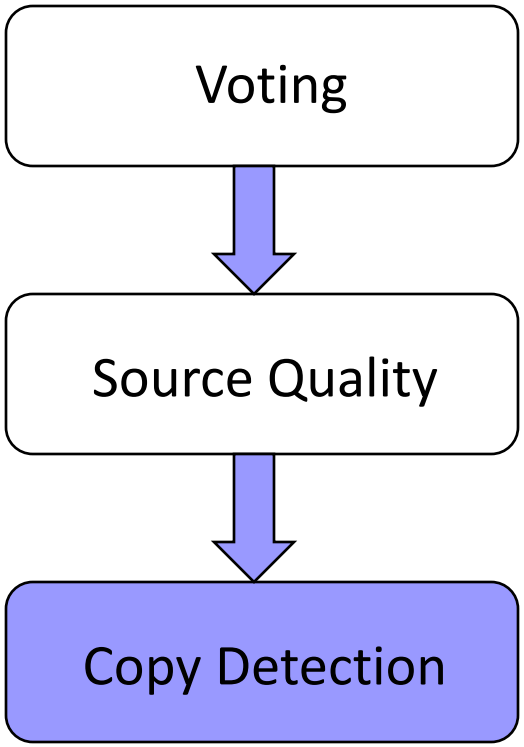
- Data fusion: voting + source quality + copy detection



	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]

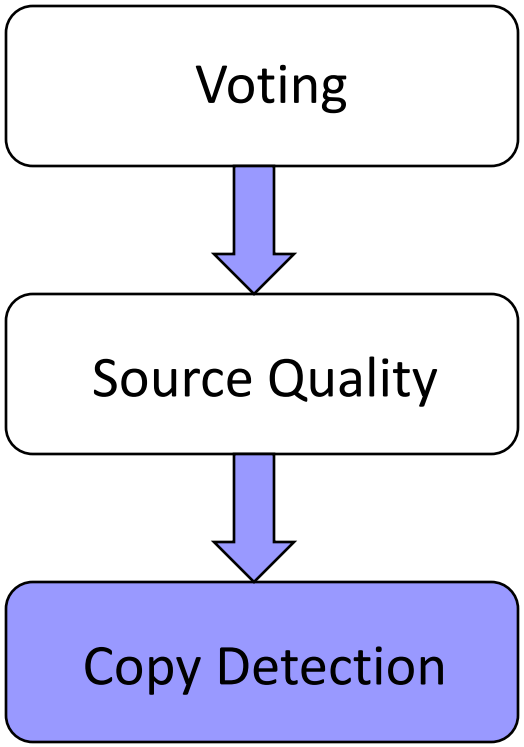
- Data fusion: voting + source quality + copy detection



	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

Copy Detection

Are Source 1 and Source 2 dependent? Not necessarily

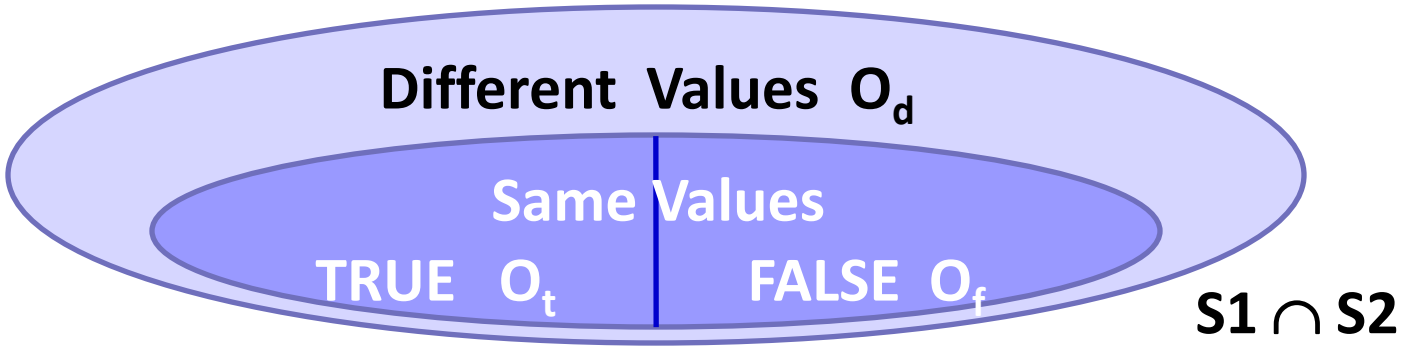
Source 1 on USA Presidents:	Source 2 on USA Presidents:	
1st : George Washington	1st : George Washington	✓
2nd : John Adams	2nd : John Adams	✓
3rd : Thomas Jefferson	3rd : Thomas Jefferson	✓
4th : James Madison	4th : James Madison	✓
…	…	
41st : George H.W. Bush	41st : George H.W. Bush	✓
42nd : William 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 $D \in S1 \cap S2$

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

- [DNS+12] Uwe Draisbach, Felix Naumann, Sascha Szott, Oliver Wonneberg: Adaptive Windows for Duplicate Detection. ICDE 2012: 1073-1083

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)
- [GS09] Rahul Gupta, Sunita Sarawagi: Answering Table Augmentation Queries from Unstructured Lists on the Web. PVLDB 2(1): 289-300 (2009)
- [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)

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

- [PIP+12] George Papadakis, Ekaterini Ioannou, Themis Palpanas, Claudia Niederee, Wolfgang Neidjl: 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

- [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

- [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

- [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
- [YJY08] 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)
- [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)