Big Data Integration

Prof. Sonia Bergamaschi

Department of Engineering "Enzo Ferrari"

University of Modena and Reggio Emilia

Outline

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

What is "Big Data Integration?"

- Big Data Integration (BDI)= Big data + Data Integration

- Data Integration: easy access to multiple data sources
  - 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

What are Big Data? Often described using Five Vs

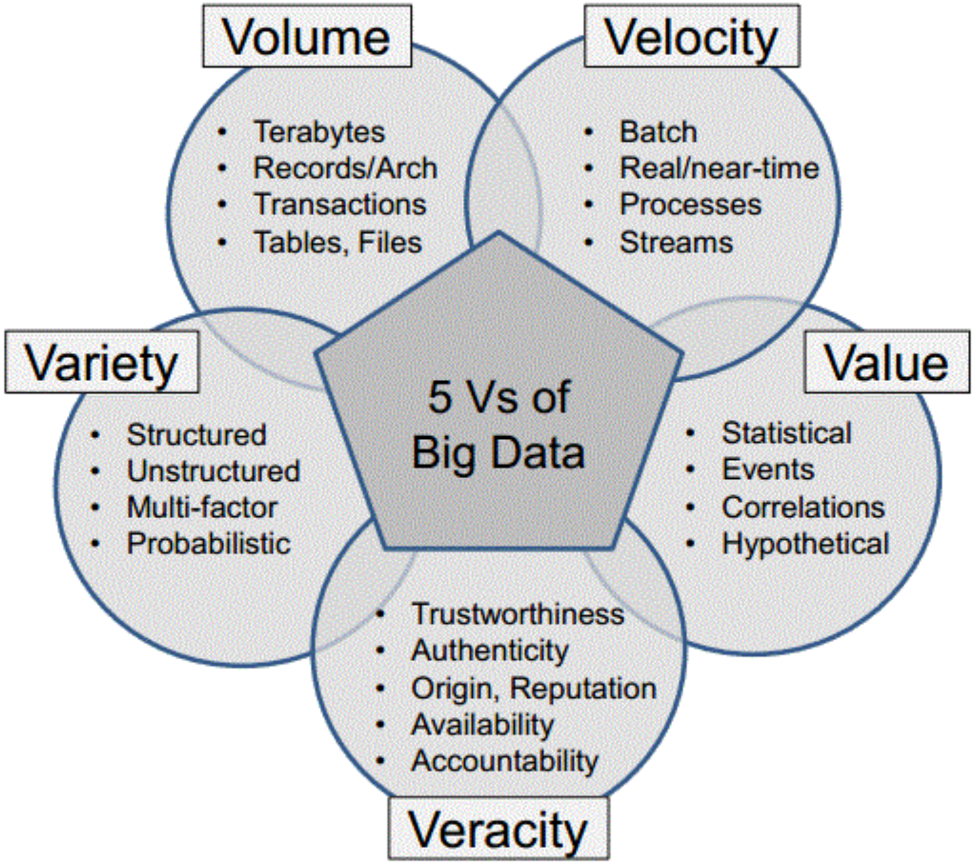
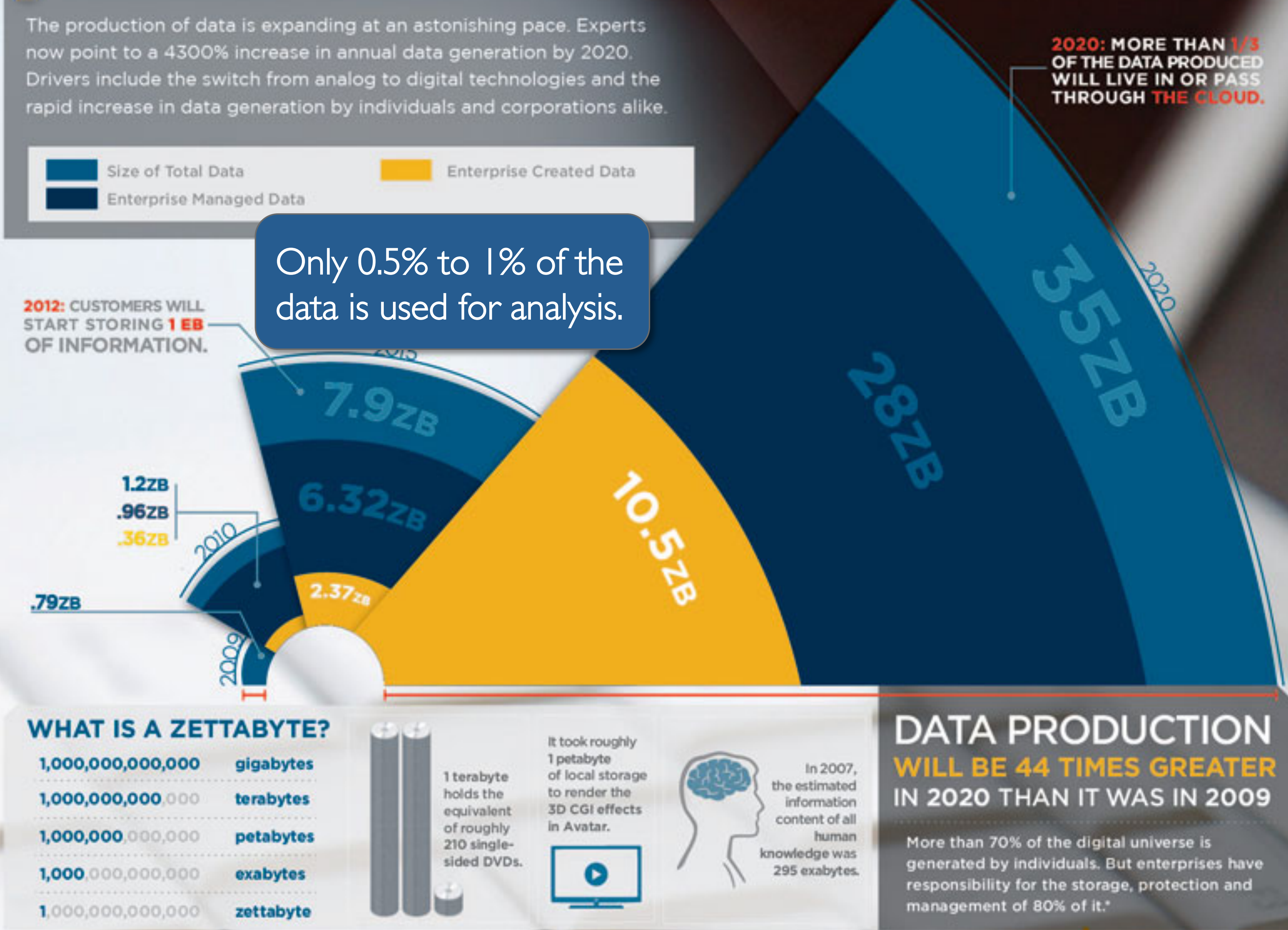


image from http://goo.gl/iz2Zig

Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari" – Università di Modena e Reggio Emilia



Only 0.5% to 1% of the data is used for analysis.

http://www.csc.com/insights/flxwd/78931-big_data_growth_just_beginning_to_explode

http://www.guardian.co.uk/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume

Questions typically asked on Big Data

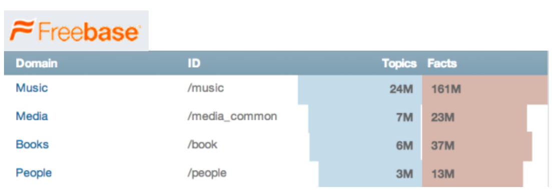
- What if your data volume gets so large and varied you don't know how to deal with it?
- Do you store all your data?
- Do you analyze it all?
- What is coverage, skew, quality?
- How can you find out which data points are really important?
- How can you use it to your best advantage?

[Seth 2014]

Do we really need Big Data Integration?

- Building web-scale knowledge bases

Google knowledge graph
enhance Google search engine's search results with semantic-search information gathered from a wide variety of sources.



Freebase

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

Freebase (Google) is an open, Creative Commons licensed repository of structured data of almost 23 million entities.
An **entity** is a single person, place, thing, or fact.
Freebase connects entities together as a graph.

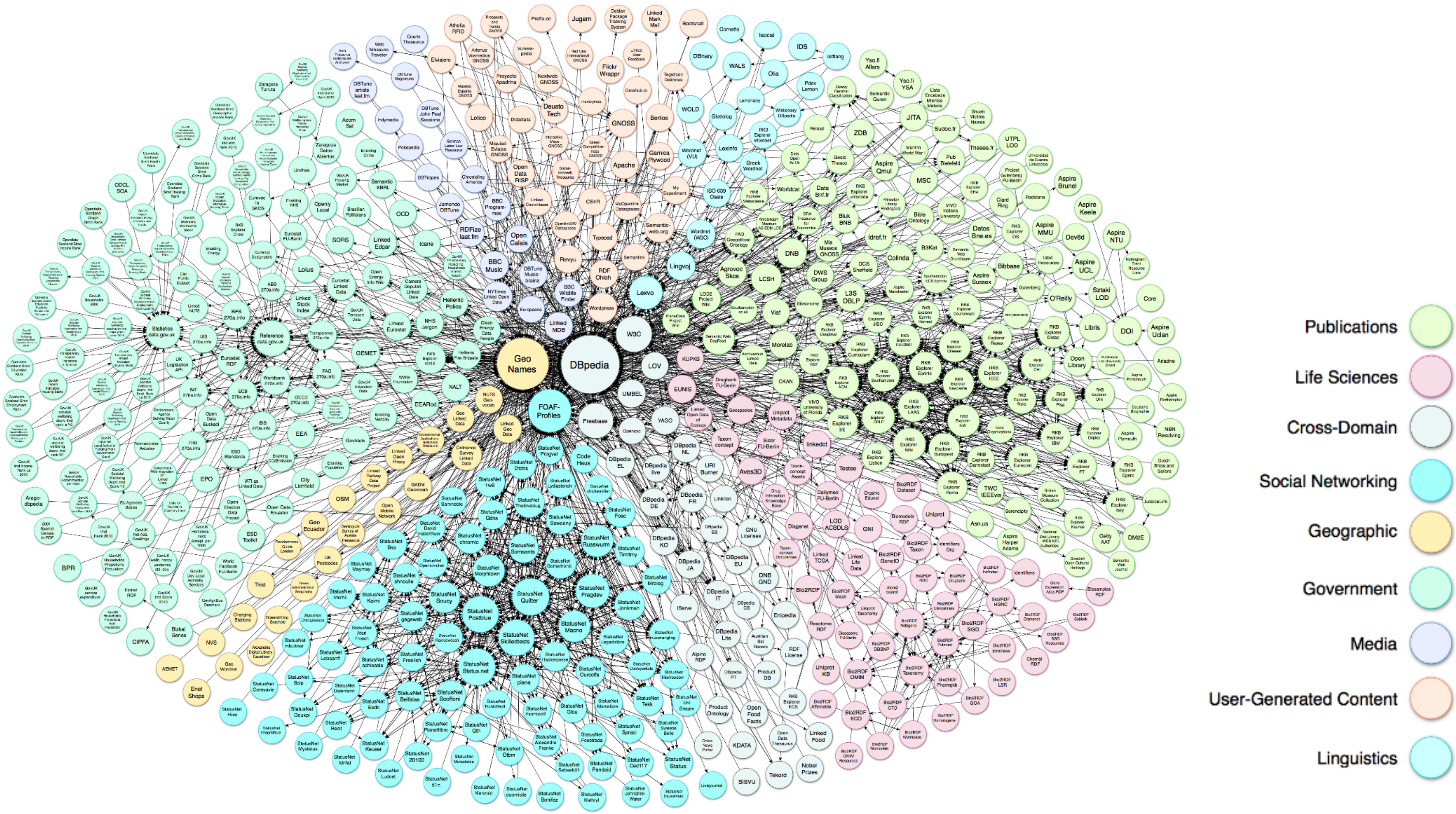
ProBase
Microsoft

- A universal, general-purpose, probabilistic taxonomy automatically constructed from a corpus of 1.6 billion web pages.
- Its goal is to open the mental world of human beings to machines. By injecting certain "general knowledge" into computing machines a better understanding of human communication can be achieved.

yago
select knowledge

- is a huge semantic knowledge base, derived from Wikipedia WordNet and GeoNames;
- has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities

Why Do We Need "Big Data Integration?"

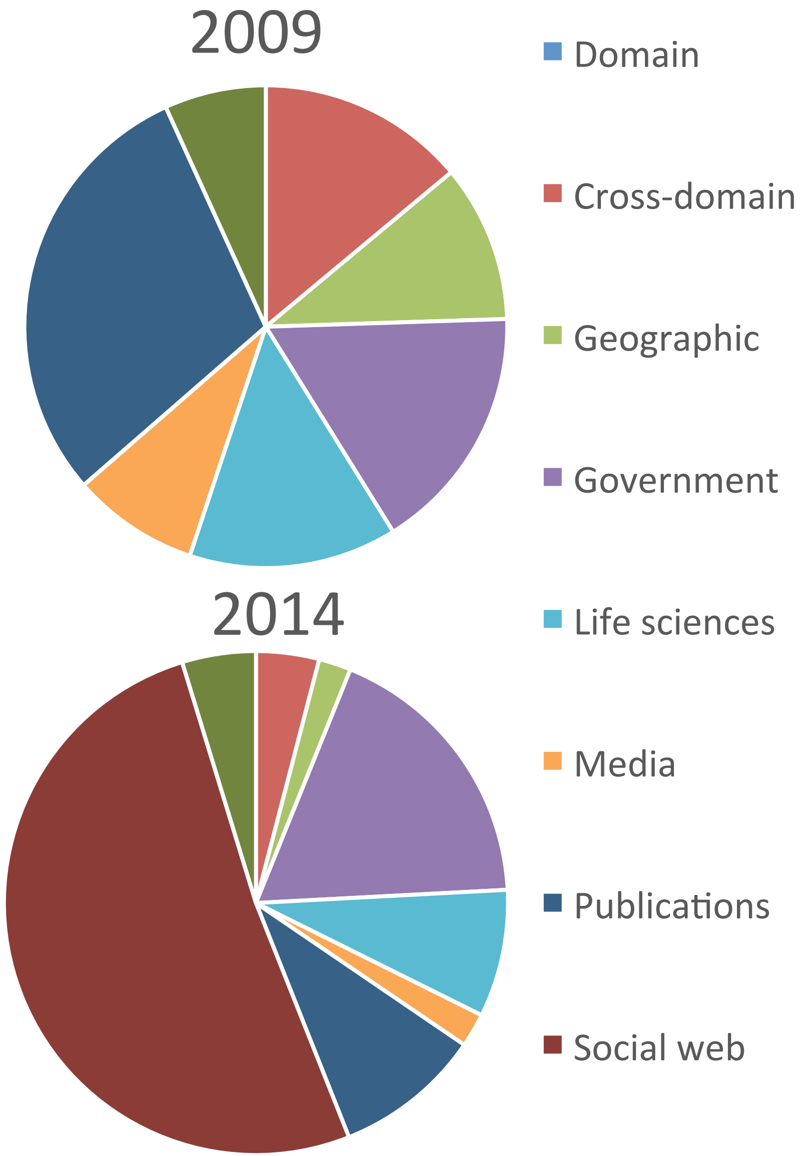


LOD Cloud Status

	2009		2014*	
Domain	Number	%	Number	%
Cross-domain	41	13.95%	41	4.04%
Geographic	31	10.54%	21	2.07%
Government	49	16.67%	183	18.05%
Life sciences	41	13.95%	83	8.19%
Media	25	8.50%	22	2.17%
Publications	87	29.59%	96	9.47%
Social web	0	0.00%	520	51.28%
User-generated content	20	6.80%	48	4.73%
Total	**294**		**1014**	

http://lod-cloud.net/

*Only 570 datasets belong to the LOD cloud,
the remaining datasets do not contain ingoing/outgoing links to the LOD Cloud.



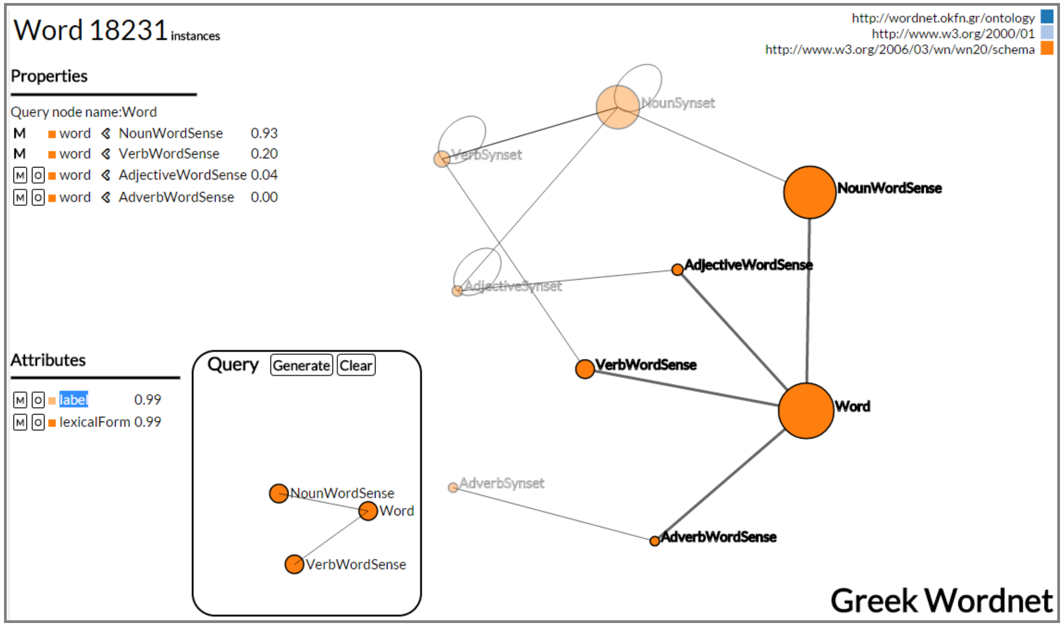
Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari" – Università di Modena e Reggio Emilia

Automatic LOD Schema Extraction & Summarization

- Using LOD sources in a Big Data Integration task is difficult as we do not have a high level view of their contents

- LODeX tries to address this issue by performing LOD Schema Extraction and Summarization from LOD sources

The output of LODeX is the Schema Summary of a source and it can be used:



- To provide a high level view of LOD sources contents

- To handle a visual interface for SPARQL query generation

[Bergamaschi et al. 2014]
[Benedetti et al. 2014]

Current Focus on Big Data

- Focus on verticals
  advertising, social media, retail, financial services, telecom and healthcare

  – Aggregate data, focused on transactions, **limited integration** (limited complexity), analytics to find (simple) patterns

  – Emphasis on technologies to handle volume/scale, and to lesser extent velocity: Hadoop, NoSQL, MPP (Massive Parallel Processing) for data warehouse: DWA (Data Warehousing Appliance), ….

  – Full faith in the power of data (no hypothesis), bottom up analysis

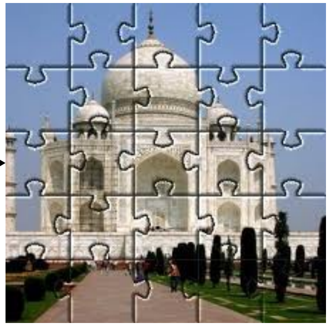
Full faith in the power of data

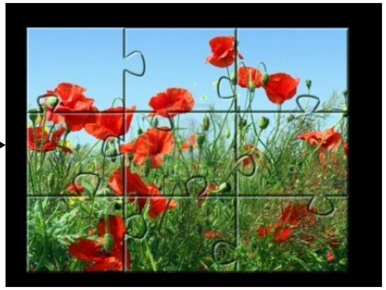
The quest for knowledge used to begin with grand theories.
Now it begins with massive amounts of data. Welcome to the Petabyte Age!

Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari" – Università di Modena e Reggio Emilia

"Small" Data Integration: What Is It?

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





[Dong and Srivastava 2013]

Small Data Integration, an Overview

Schema Alignment

provides a global mediated schema of local sources schemata on the basis of local attributes matching and global to local mappings

↓

Record Linkage

identifies different instantiations of the same entity coming from local sources

↓

Data Fusion

fuses in a single entity its different instantiations coming from local sources

Virtual Data Integration

MOMIS

www.datariver.it

[Berg.SIGMOD99]

SCHEMA ALIGNMENT
semi automatic

Global as View Mapping

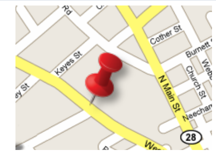
Companies Mediated Schema

1. Attribute Matching
2. Companies Mediated Schema
3. Global as View mapping
4. Query

MoMIS

www.datariver.it

MoMIS

Name	Address	Sector	Revenue	Map
Software Inc.	Nimitz Fwy, Newark, US	Information Technology	€ 6.000 mln	
Fashion Inc.	Via Savona, Cuneo, IT	Textile	€ 930 mln	

VIRTUAL INTEGRATION DATA CONFLICTS RESOLUTION

Data stored in Local sources

XML

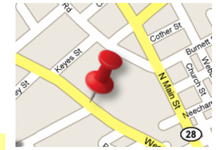
Name	Address	Sector	N° Emp.
Fashion Inc.	Via Savona, Cuneo, IT	Textile	8000
Software Inc.	Nimitz Fwy, Newark, US	Information Technology	600

Company	Location	Revenue
Software Inc.	Nimitz Fwy, Newark, US	€ 6.000 mln
Fashion Inc.	Via Libertà, Cuneo, IT	€ 930 mln

Name	Address	Latitude	Longitude
Software Inc.	Nimitz Fwy, Newark, US	37'44 N	122'13 W

ALWAYS UP TO DATE

MoMIS

Name	Address	Sector	Revenue	Map
Software Inc.	Nimitz Fwy, Newark, US	Information Technology	€ 6.000 mln	
Fashion Inc.	Via Savona, Cuneo, IT	Textile	€ 1.200 mln	

VIRTUAL INTEGRATION

Data stored in Local sources

XML

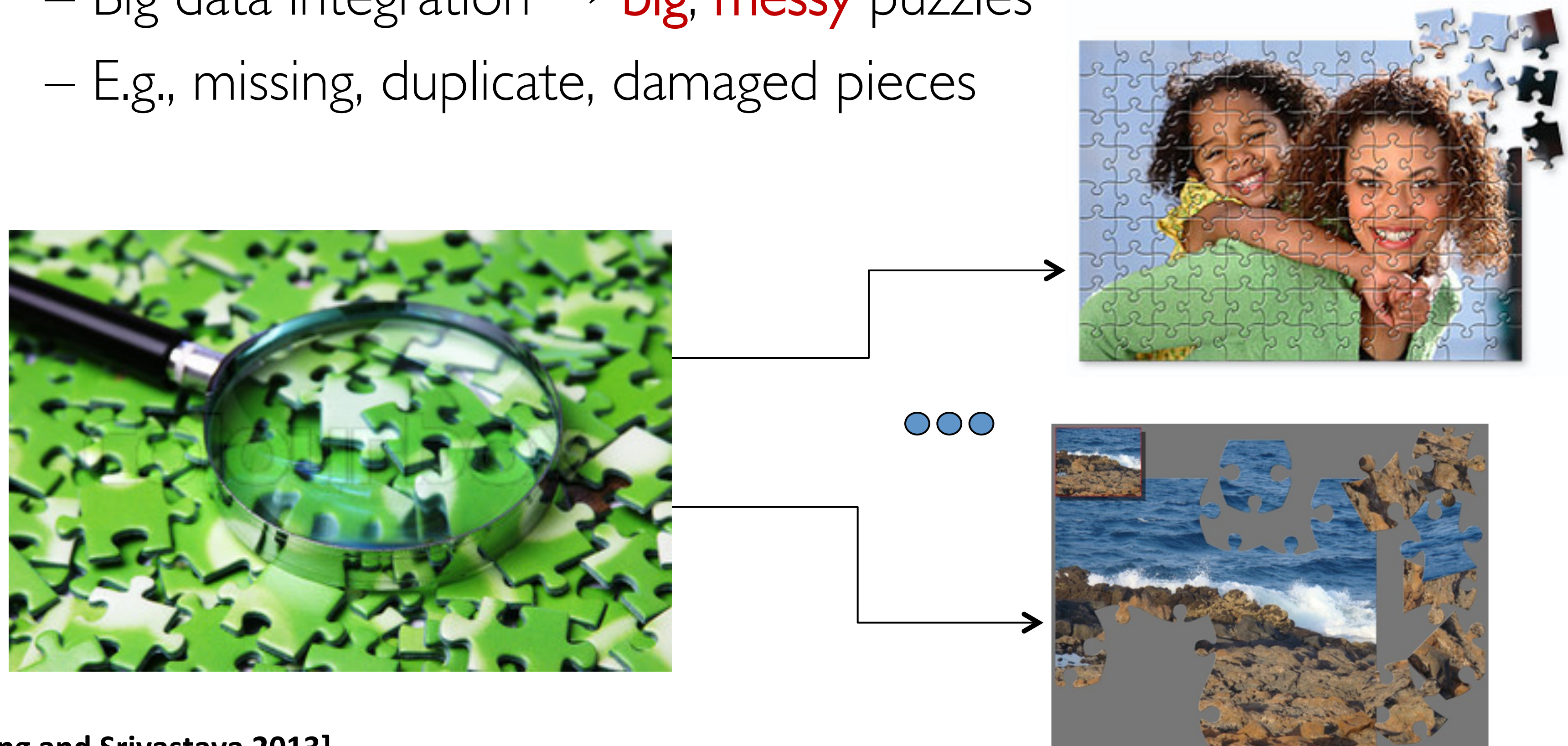
Name	Address	Sector	N° Emp.
Fashion Inc.	Via Savona, Cuneo, IT	Textile	8000
Software Inc.	Nimitz Fwy, Newark, US	Information Technology	600

Company	Location	Revenue
Software Inc.	Nimitz Fwy, Newark, US	€ 6.000 mln
Fashion Inc.	Via Libertà, Cuneo, IT	€ 1.200 mln

Name	Address	Latitude	Longitude
Software Inc.	Nimitz Fwy, Newark, US	37’44 N	122’13 W

BDI: Why is it Challenging?

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



[Dong and Srivastava 2013]

BDI: Why is it Challenging?

Number of structured sources: Volume

- 150 million high quality relational tables on the web
- 10s of millions of high quality deep web sources
- Challenges:
  - Difficult to do schema alignment even if we restrict the integration within a single specific domain
  - Expensive to warehouse all the integrated data
  - Infeasible to support virtual integration

Rate of change in structured sources: Velocity

- Many sources provide rapidly changing data, e.g., stock prices
- 450,000 databases, 1.25M query interfaces on the web
- Challenges:
  - Difficult to understand evolution of semantics
  - Extremely expensive to warehouse data history
  - Infeasible to capture rapid data changes in a timely fashion

[Dong and Srivastava 2013]

BDI: Why is it Challenging?

- Representation differences among sources: Variety

Synopsis

Born or
conce
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

Deep Web Quality: a Case Study

- Deep Web Quality: Volume, Velocity, Veracity

Study on two high quality domains: Stock Market, Flights

- sources: Top 100 results from Google and Yahoo!, then filtered the ones that can be crawled or queried by means of API
- Sources have different attributes ("local attrs"), but many of them have the same semantic. Global attributes = attributes after manual matching

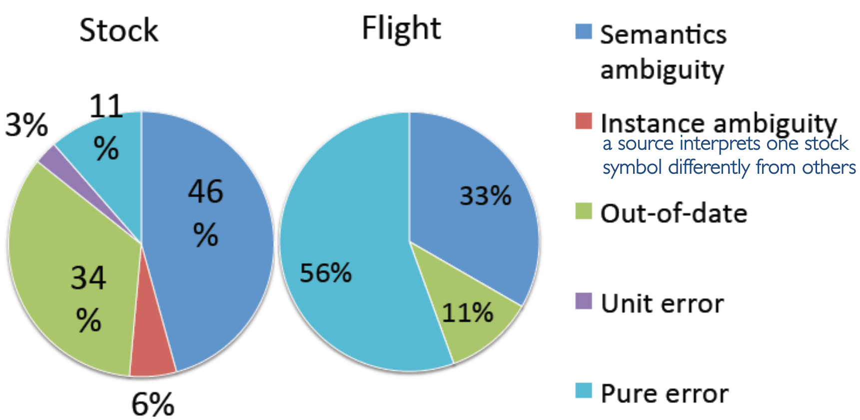
	#Sources	Period	#Objects *days	#Local-attrs	#Global-attrs	Considered items
Stock	55	7/2011	1000*20	333	153	16000*20
Flight	38	12/2011	1200*31	43	15	7200*31

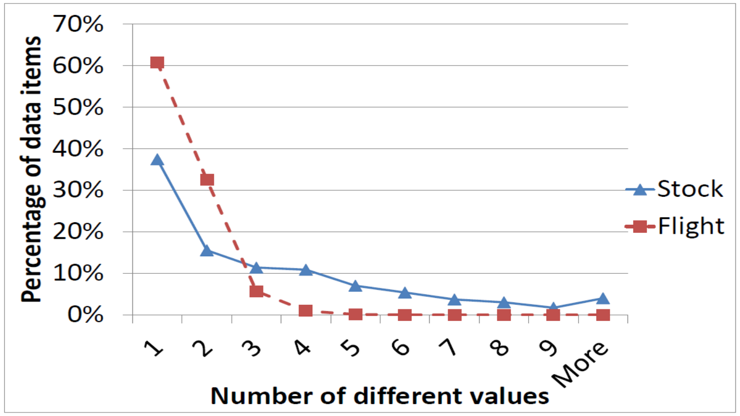
- Spread: we need to go to the long tail of sources to build a reasonably complete database
- Connectivity: Sources are well-connected, with a high degree of content redundancy and overlap

[Dong et al. 2013]

Deep Web Quality: a Case Study (2)

Deep Web data has considerable inconsistency





- For the 60% of items in stock sources, we find 2 different values

Outline

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

KEYWORD SEARCH ON STRUCTURED DATABASES

Hotels in London next
to a Train Station

Current structured query
languages are not suitable for end users

Keyword queries can be
the solution

Structured data on the web are
believed to be at least 500 times more
than those that exist as web pages

[Bergamaschi et al. 2011]
[Bergamaschi et al. 2013]

Keyword Search in BDI: Beyond Schema Alignment

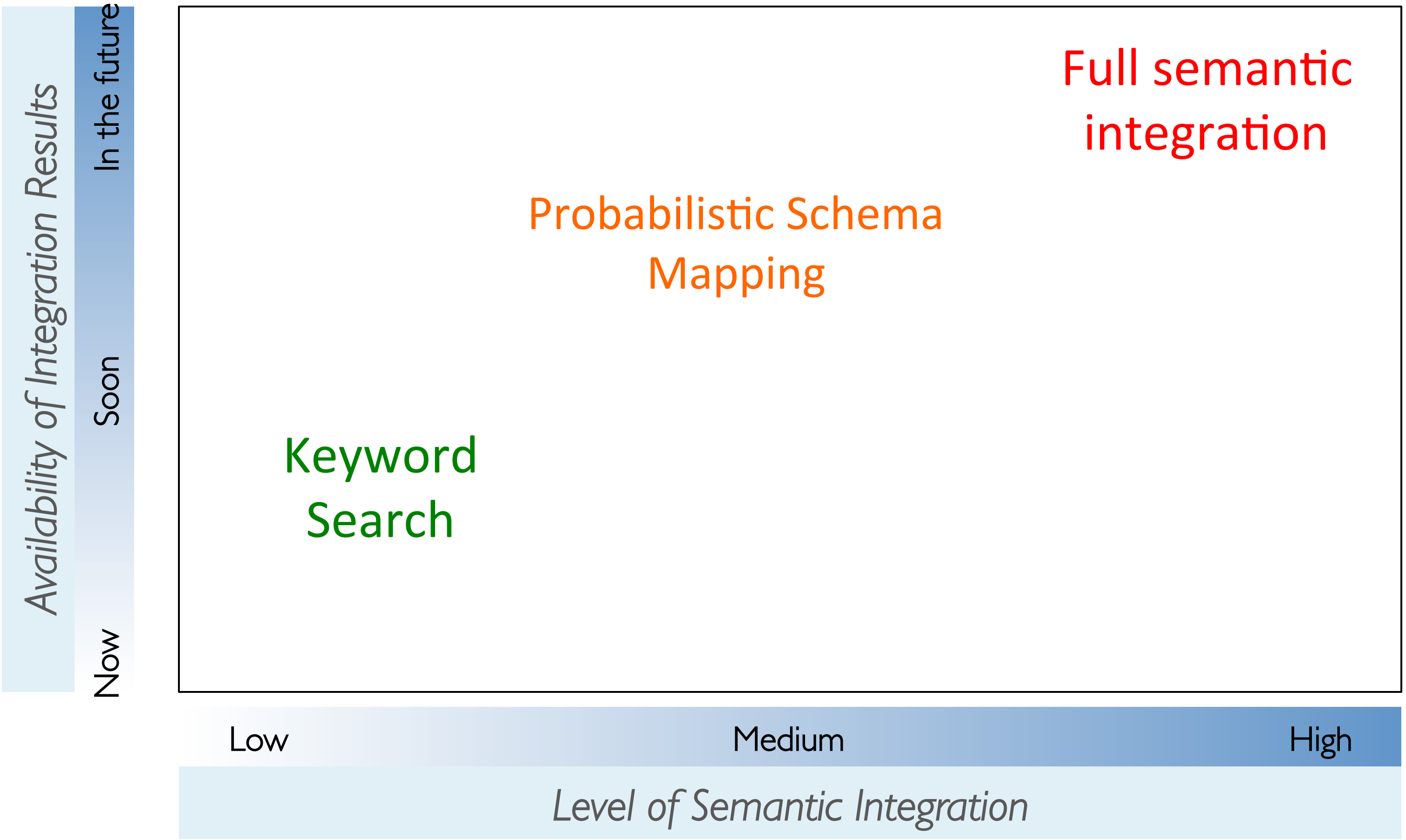
- The user poses keyword queries that are matched against source relations and their attributes;

- the system uses metadata (e.g., foreign keys, links, schema mappings, synonyms, and taxonomies) to create multiple ranked queries linking the matches to keywords; the set of queries is attached to a Web query form and the user may pose specific queries by filling in parameters in the form

- the answers are ranked and annotated with data provenance, and the user provides feedback on the utility of the answers, from which the system ultimately learns to assign costs to sources and associations, as a result changing the ranking of the queries used to generate results.

[Talukdar et al. 2008]

BDI Space of Strategies

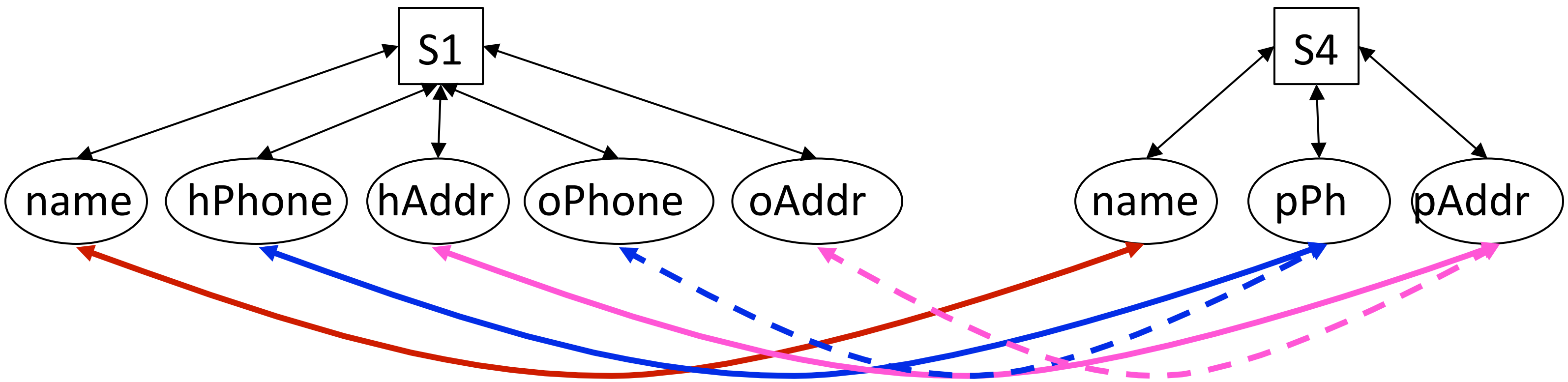


Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari" – Università di Modena e Reggio Emilia

Bootstrapping DI Systems

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

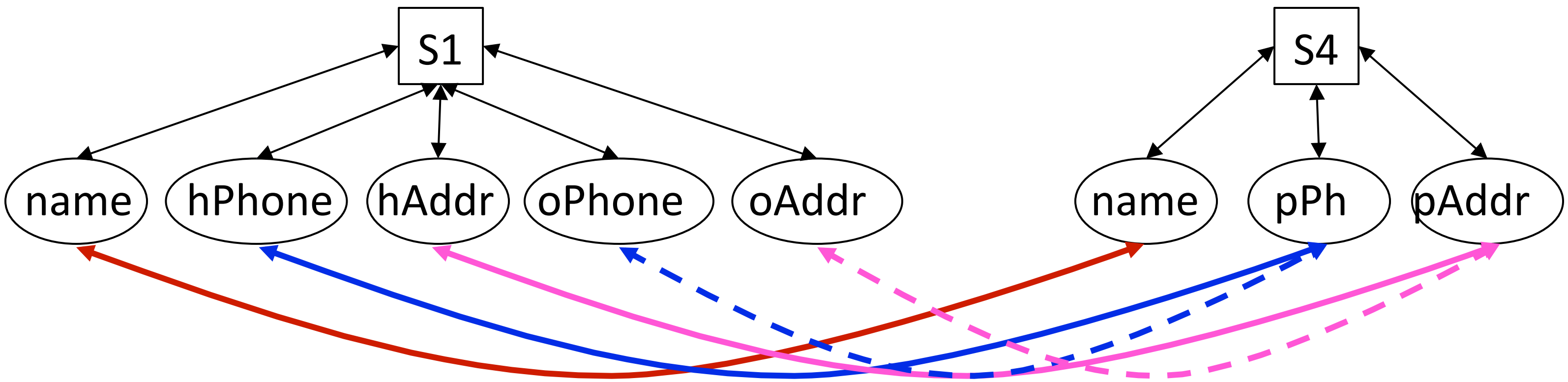
Probabilistic Mediated Schemas



- Mediated schemas: automatically created by inspecting sources
  - Clustering of source attributes (ex: S1.hPhone, S1.oPhone, S4.pPh)
  - Variety of sources → uncertainty in accuracy of clustering

[Sarma et al. 2008]

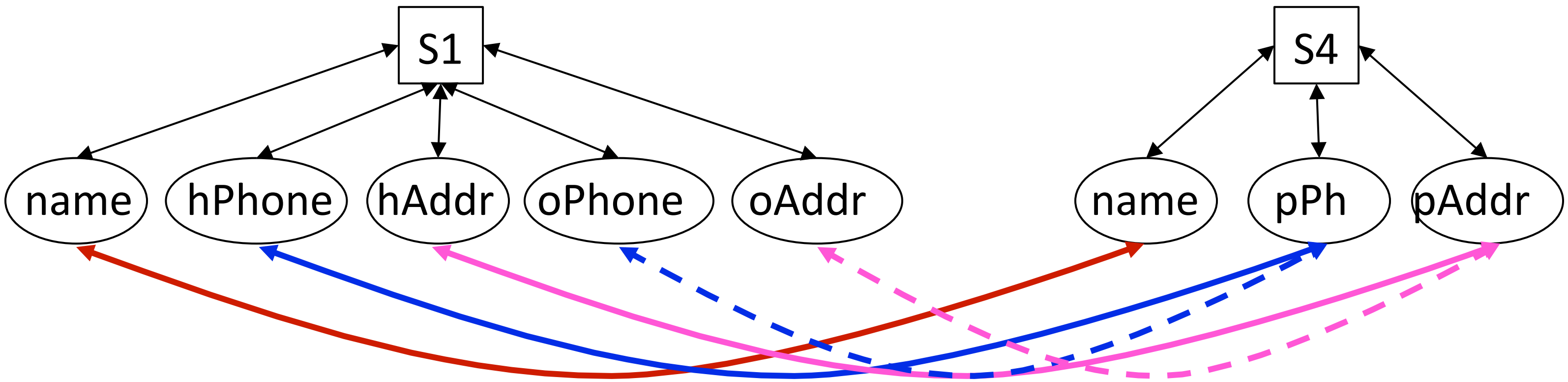
Probabilistic Mediated Schemas



- 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

- 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 with 2 possible semantics for such mappings:
  - *By table semantics*: one mapping for all tuples in a table
    - assumes that there exists a correct mapping but we don't know what it is
  - *By tuple semantics*: different mappings are okay in a table
    - assumes that the correct mapping may depend on the particular tuple in the source data

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

Outline

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

BDI Record Linkage: Three Steps

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

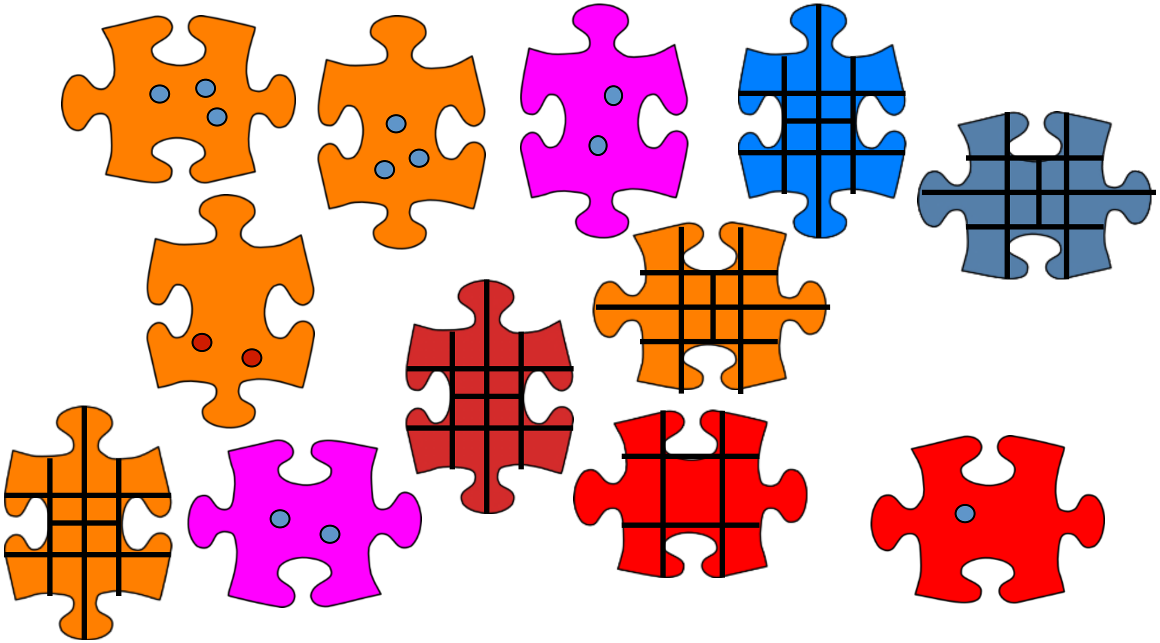
Blocking

↓

Pairwise Matching

↓

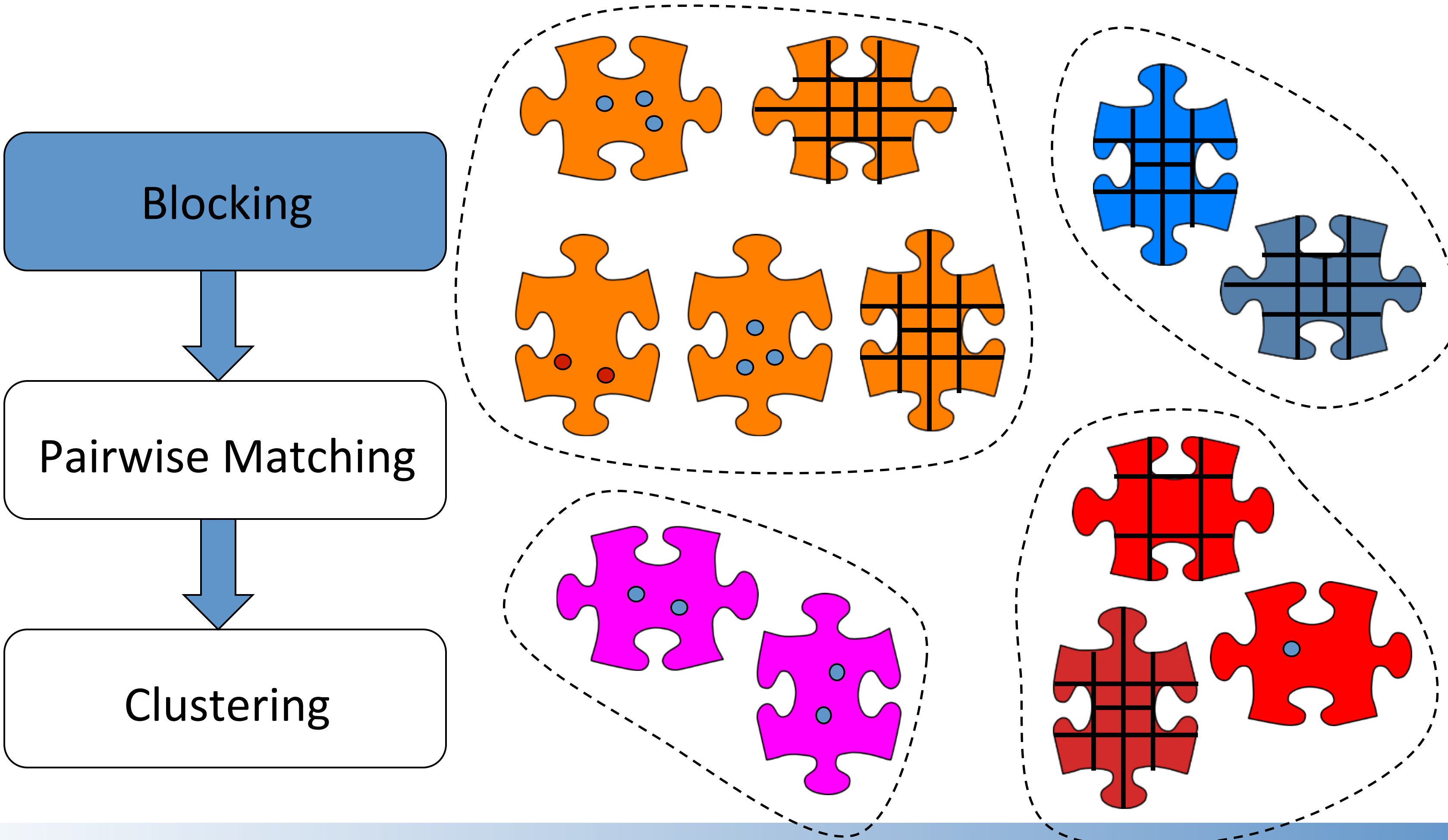
Clustering



Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari" – Università di Modena e Reggio Emilia

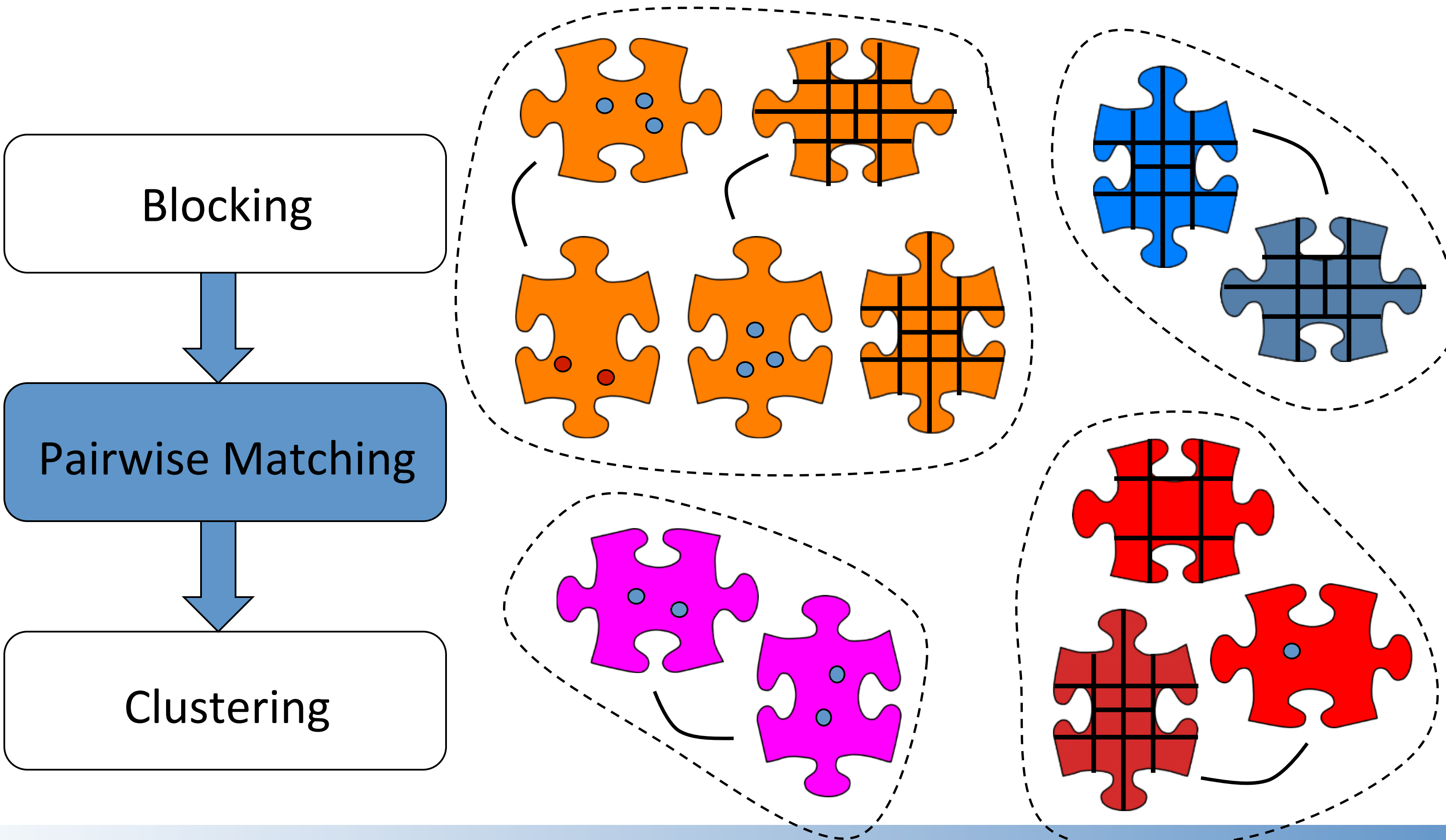
BDI Record Linkage: Three Steps

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



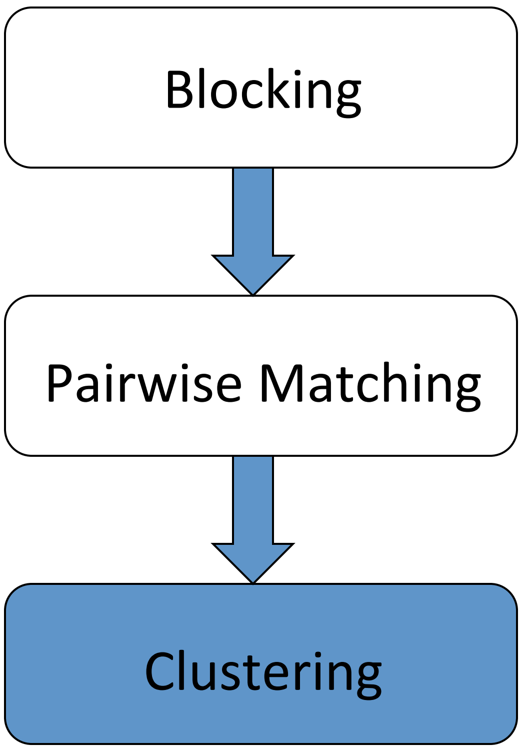
BDI Record Linkage: Three Steps

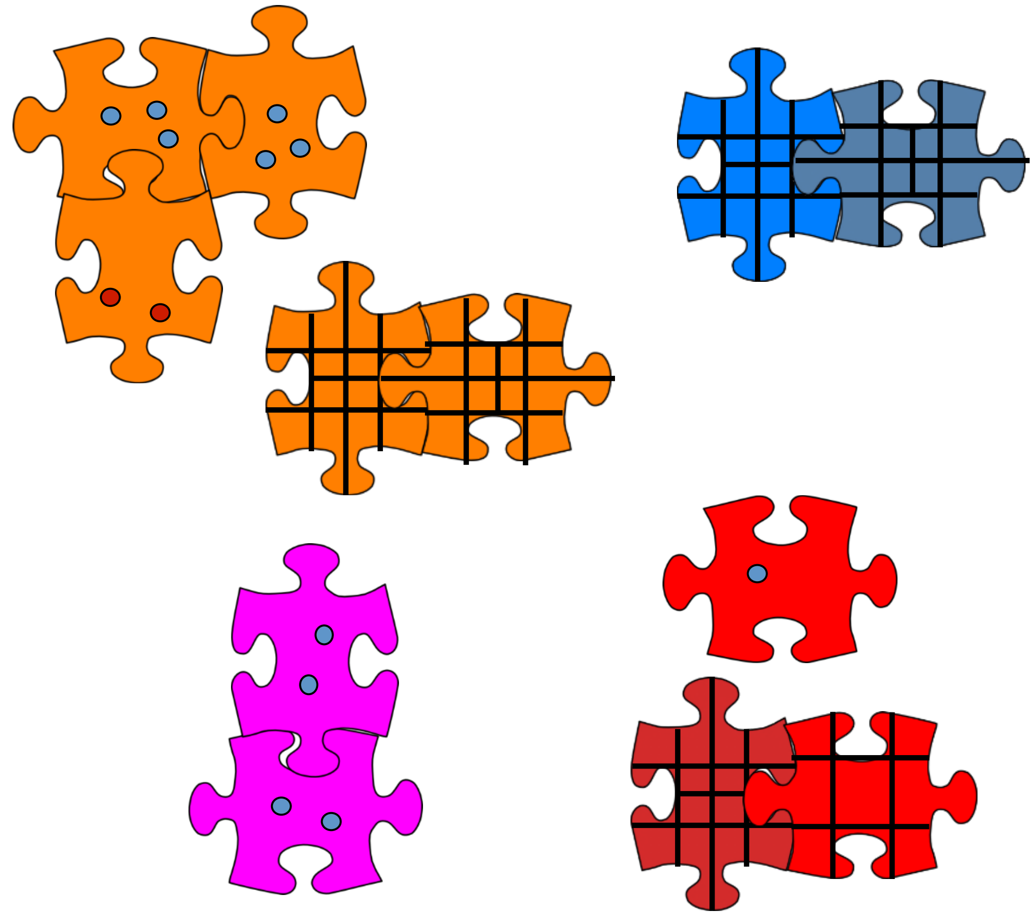
- Pairwise matching: compares all record pairs in a block
  - Computes similarity



BDI Record Linkage: Three Steps

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





Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari" – Università di Modena e Reggio Emilia

Challenge for Record Linkage: 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

[Kannan et al. 2008]

Challenge for Record Linkage: 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 Dmc-fx07 7.0 Mp Digital Camera Boxed **Lumix** 10541r
Panasonic Lumix - Point & Shoot - 7 megapixel - Compact Sensor - 3.6 x optical zoom -
Panasonic DMC-FX07 7.0 MP Digital Camera Boxed Serial #FC6GA10541r Product Description: **Panasonic Lumix DMC-FX07** - Digital camera - compact - 7.0 ...
Add to Shortlist

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

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

Technologies for BDI: Record Linkage Using MapReduce

- Motivation: despite use of blocking, record linkage is expensive
  – Can record linkage be effectively parallelized?

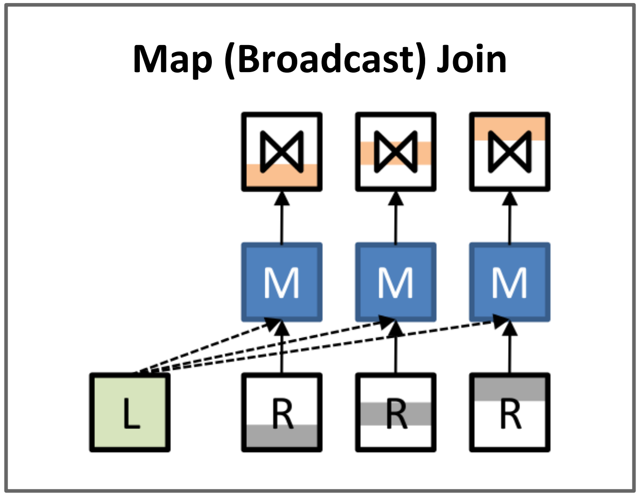
- Basic: use MapReduce to execute blocking-based record linkage in parallel
  – Map tasks can read records and perform redistribution based on blocking key
  – All entities of the same block are assigned to same Reduce task
  – Different blocks matched in parallel by multiple Reduce tasks
  → Difficult to tune the blocking function to get balanced workload

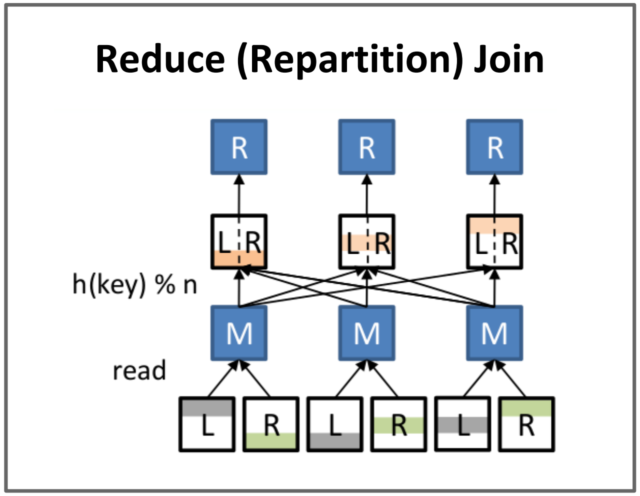
Challenge: data skew → unbalanced workload

[Kolb et al. 2012]

Technologies for BDI: Record Linkage

Similar skew issue with join operation in Hadoop:
which strategy to choose? How to configure it?





- Joins do not naturally fit MapReduce
- Very time consuming to implement
- Hand optimization necessary

Image from Robert Metzger's speech – "Stratosphere: System Overview" – Big Data Beers Meetup, Nov. 19th 2013

Technologies for BDI: Record Linkage (2)

Challenge: data skew → unbalanced workload

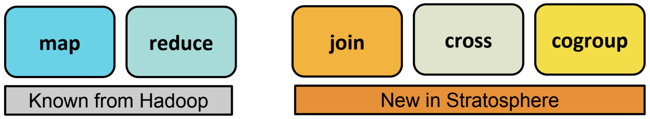
- Key ideas for load balancing:
  - Preprocessing MR job to determine blocking key distribution
    - It worth the overhead since the reduce phase consumes the vast majority of the overall runtime (95%)
  - Redistribution of Map tasks to Reduce tasks to balance workload
- Two load balancing strategies:
  - BlockSplit: split large blocks into sub-blocks
  - PairRange: global enumeration and redistribution of all pairs

[Kolb et al. 2012]

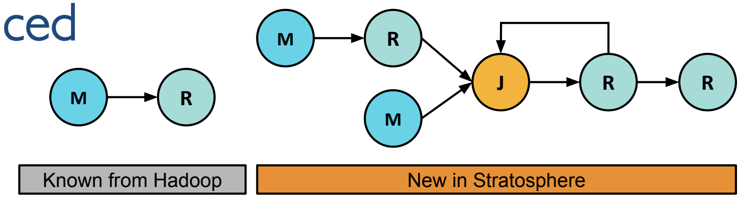
Solving MapReduce Problems

Stratosphere [Alexandrov et al. 2014]

- Extends MapReduce with more operators



- Support for advanced data flow graphs



- Only write to disk if necessary (otherwise in-memory)
- Natively implemented JOINS into the system
  - Optimizer decides join strategy (e.g. Hybrid Hash Join starts in-memory and gracefully degrade)

Image from Robert Metzger's speech – "Stratosphere: System Overview" – Big Data Beers Meetup, Nov. 19th 2013

Similar operator are implemented also in:

- Spark [Matel et al. 2012]
  - Focus on *in-memory computation*.
- Hyracks [Borkar et al. 2011]
  - Focus on expressing computation as a DAG (directed acyclic graph) of data operator

Outline

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

Technologies for BDI: Data Fusion

Basic Solution: Naïve Voting

- Supports difference of opinion, allows conflict resolution
- Works well for independent sources that have similar accuracy

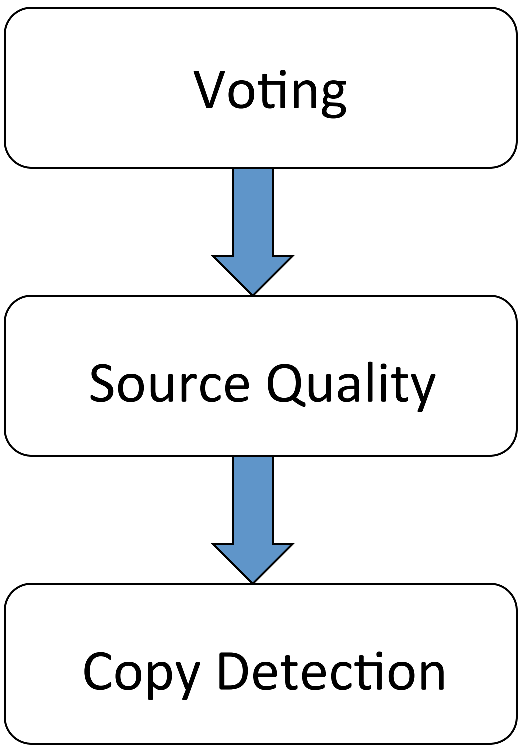
But…

- 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

Data Fusion when Conflicts Arises: Three Components

Reconciliation of conflicting content

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



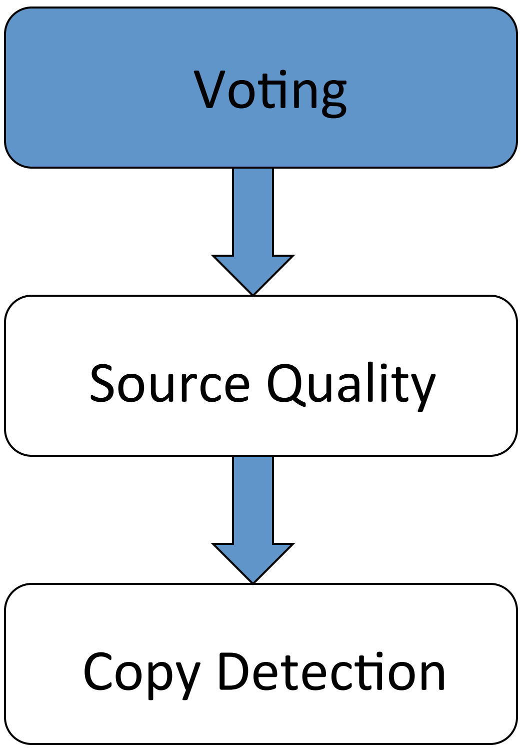
	S1	S2	S3	S4	S5
Jagadish	UM	ATT	UM	UM	UI
Dewitt	MSR	MSR	UW	UW	UW
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	ATT	BEA	BEA	BEA
Franklin	UCB	UCB	UMD	UMD	UMD

[Dong et al. 2009]

Data Fusion: Three Components

Reconciliation of conflicting content

- Data fusion: voting + source quality + copy detection
  - Initially we know only 3 sources

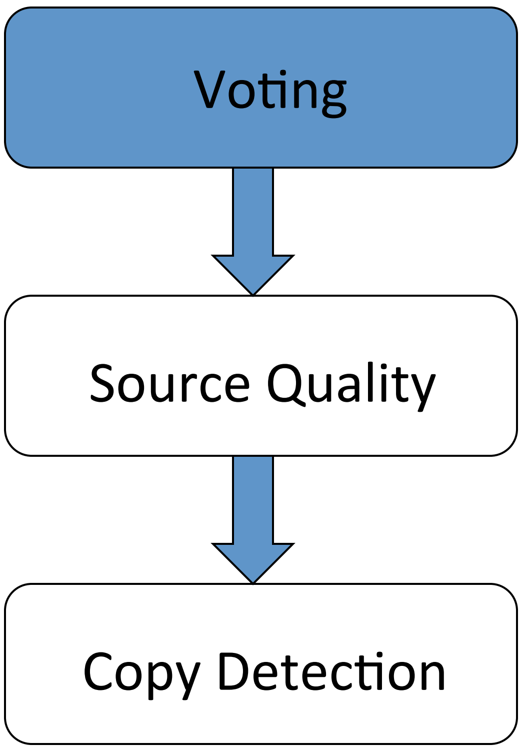


	S1	S2	S3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

Data Fusion: Three Components

Reconciliation of conflicting content

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

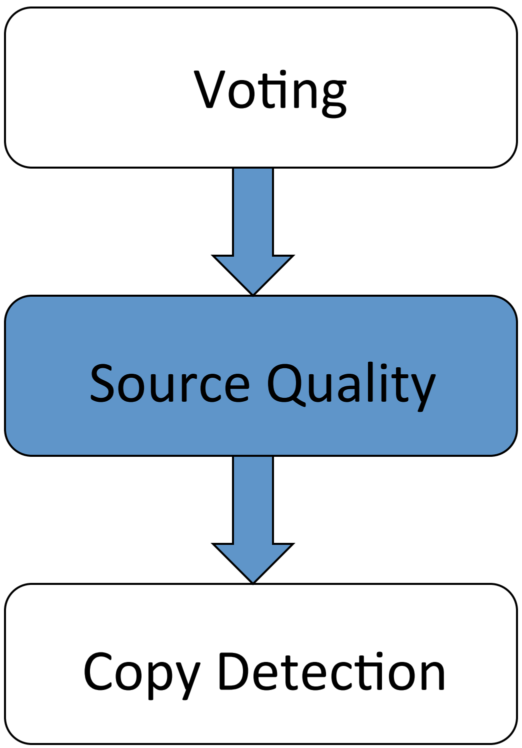


	S1	S2	S3
Jagadish	**UM**	ATT	**UM**
Dewitt	**MSR**	**MSR**	UW
Bernstein	**MSR**	**MSR**	**MSR**
Carey	UCI	ATT	BEA
Franklin	**UCB**	**UCB**	UMD

Data Fusion: Three Components

Reconciliation of conflicting content

- Data fusion: voting + source quality + copy detection
  - S1 wins providing the highest number of agreed content

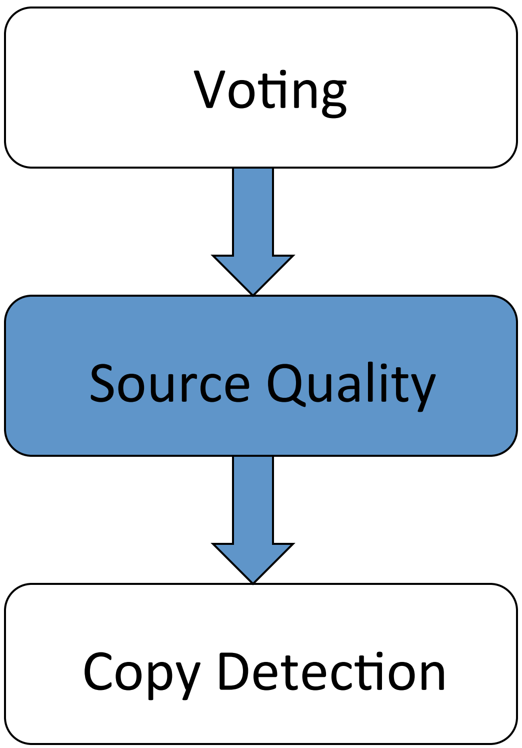


	S1	S2	S3
Jagadish	**UM**	ATT	**UM**
Dewitt	**MSR**	**MSR**	UW
Bernstein	**MSR**	**MSR**	**MSR**
Carey	UCI	ATT	BEA
Franklin	**UCB**	**UCB**	UMD

Data Fusion: Three Components

Reconciliation of conflicting content

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

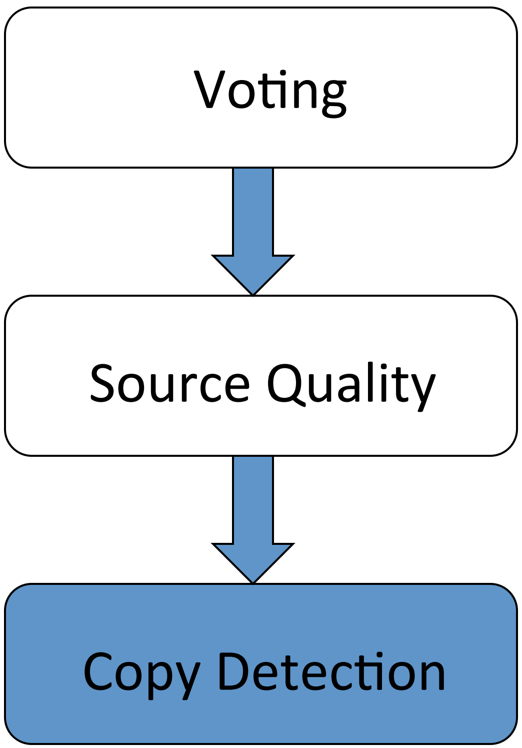


	S1	S2	S3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

Data Fusion: Three Components

Reconciliation of conflicting content

- Data fusion: voting + source quality + copy detection
  - Two more sources considered

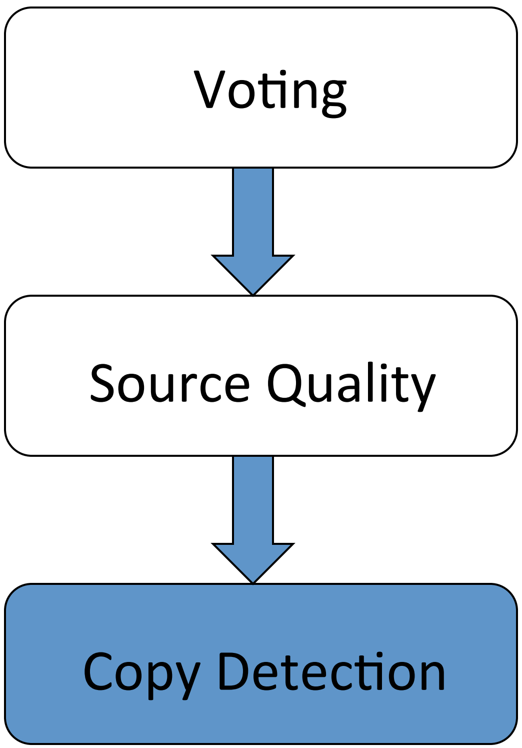


	S1	S2	S3	S4	S5
Jagadish	UM	ATT	UM	UM	UI
Dewitt	MSR	MSR	UW	UW	UW
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	ATT	BEA	BEA	BEA
Franklin	UCB	UCB	UMD	UMD	UMD

Data Fusion: Three Components

Reconciliation of conflicting content

- Data fusion: voting + source quality + copy detection



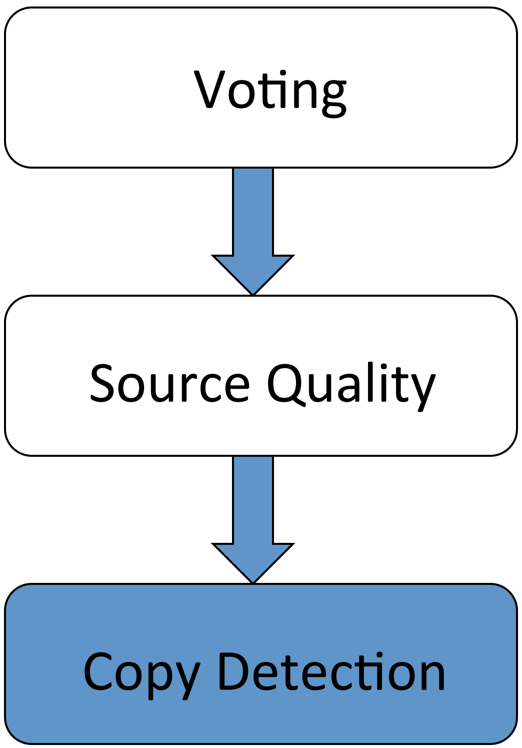
	S1	S2	S3	S4	S5
Jagadish	UM	ATT	UM	UM	UI
Dewitt	MSR	MSR	UW	UW	UW
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	ATT	BEA	BEA	BEA
Franklin	UCB	UCB	UMD	UMD	UMD

Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari" – Università di Modena e Reggio Emilia

Data Fusion: Three Components

Reconciliation of conflicting content

- 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?

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 M	on	✓
…		
41st : George	. Bush	✓
42nd : William J. Clinton	42nd : William J. Clinton	✓
43rd : George W. Bush	43rd : George W. Bush	✓
44th: Barack Obama	44th: Barack Obama	✓

Not necessarily, because it is the correct order of presidents.

- 2 independent correct sources will always have the same data

Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari" – Università di Modena e Reggio Emilia

Copy Detections

Are Source 1 and Source 2 dependent?

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. Ke	3rd : John F. Kennedy	✗
4th : Abrahan	coln	✗
…	…	
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	

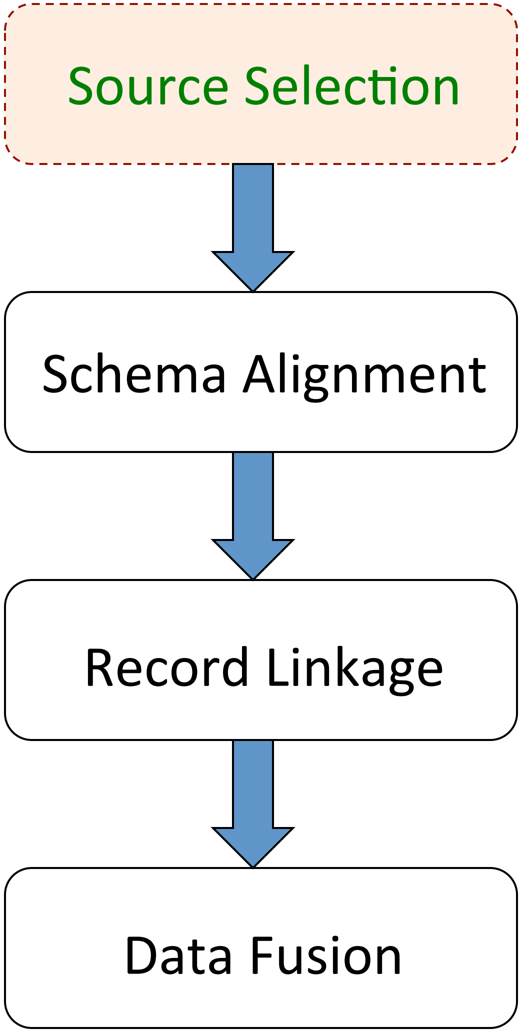
Very Likely one source have copied the incorrectness of the other source.

58

Outline

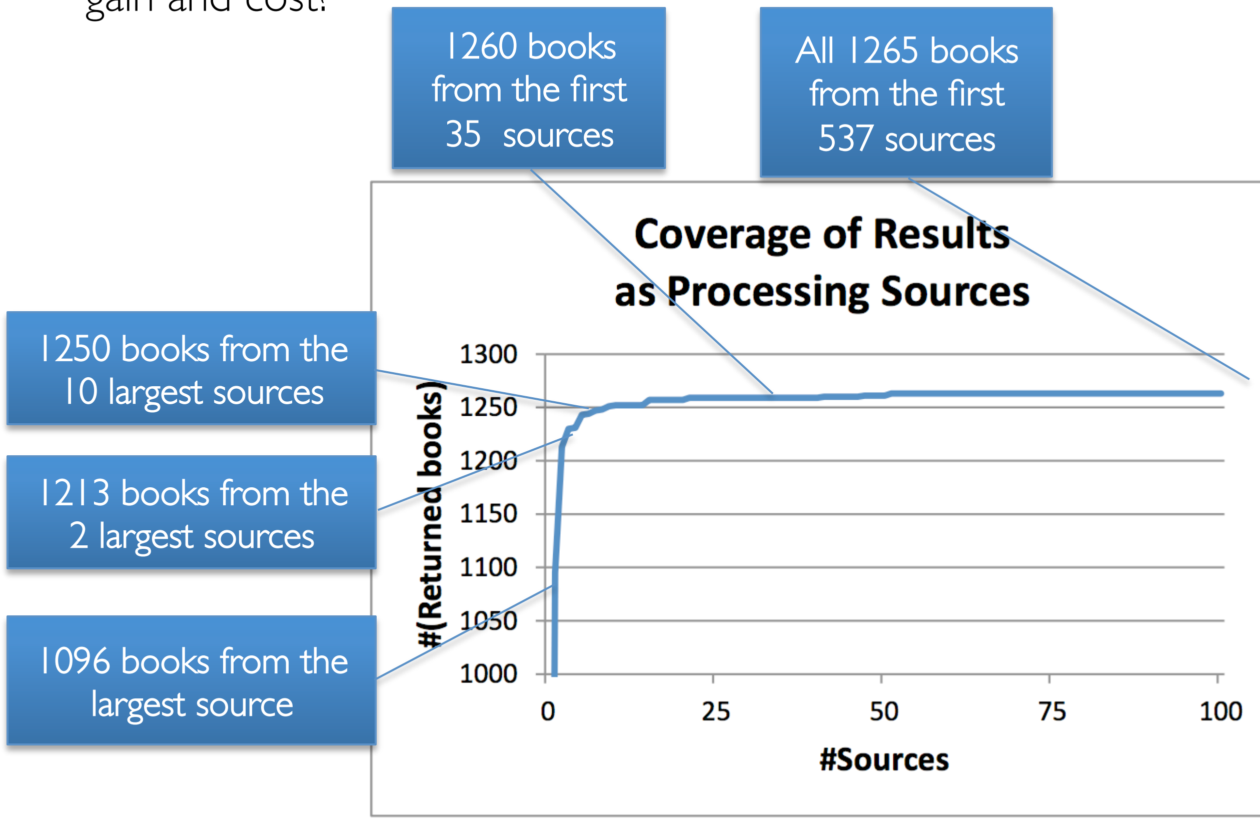
- Motivation
- Schema alignment
- Record linkage
- Data fusion
- Emerging topics

BDI: Source Selection



Is it best to integrate all data?

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

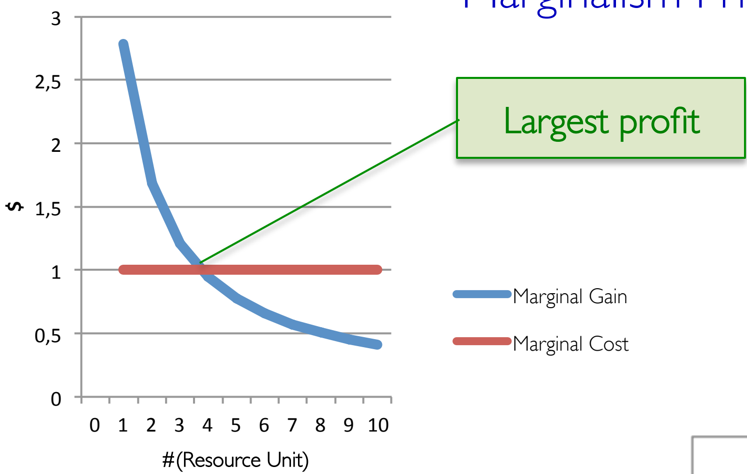


Redundant Data Do Not Bring Much Gain

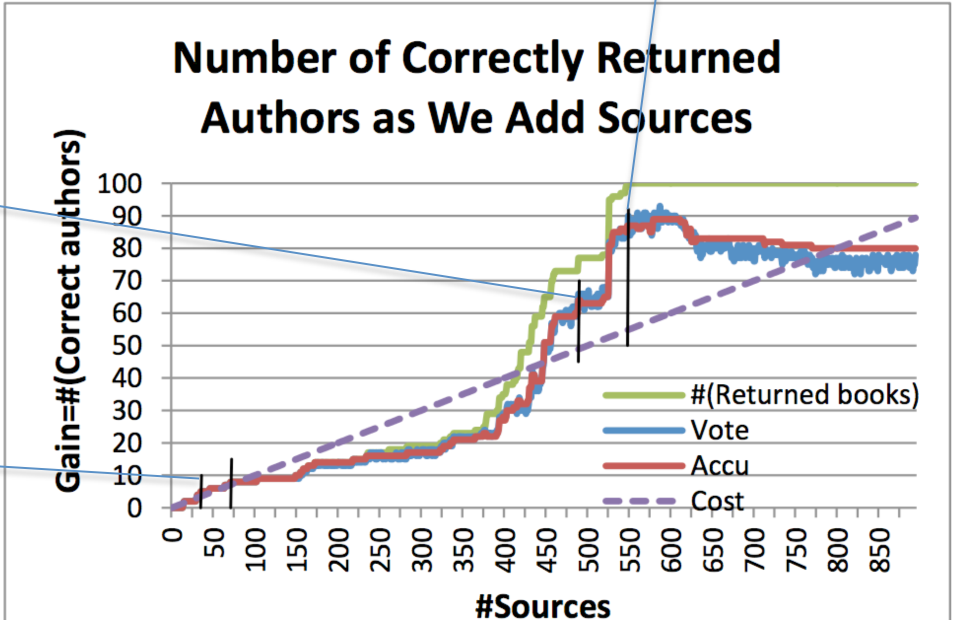
[Dong X.L. et al. 2013]

Marginalism for Source Selection

Marginalism Principle in Economic Theory



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



CS books from AbeBooks.com

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

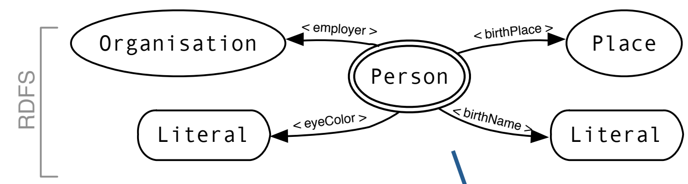
Challenge 2. Each source is different in quality, so different ordering leads to different marginal points: best solution integrates 26 sources

Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari" – Università di Modena e Reggio Emilia

A Statistical Approach to Discover the Topics of a Data Source

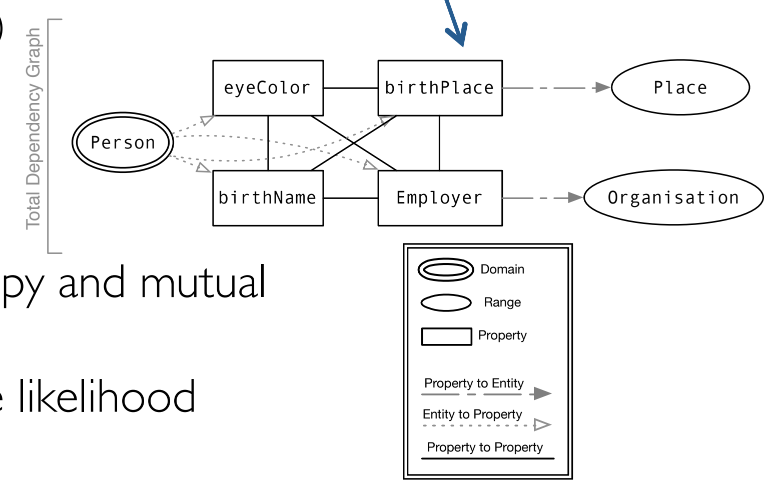
How to find the data sources satisfying specific information needs ?

- In portals (e.g. DataHub) data sources are indexed on the basis of metadata (e.g., title, author, content description, …) provided by the content owner
- Searching in portals is a tedious and error prone work generating biased results if the metadata are not accurate



Challenges:

1. To provide an automatic data-driven approach for extracting metadata from a target source with respect to a reference ontology (e.g. DBPedia)
2. To handle the huge size of involved data



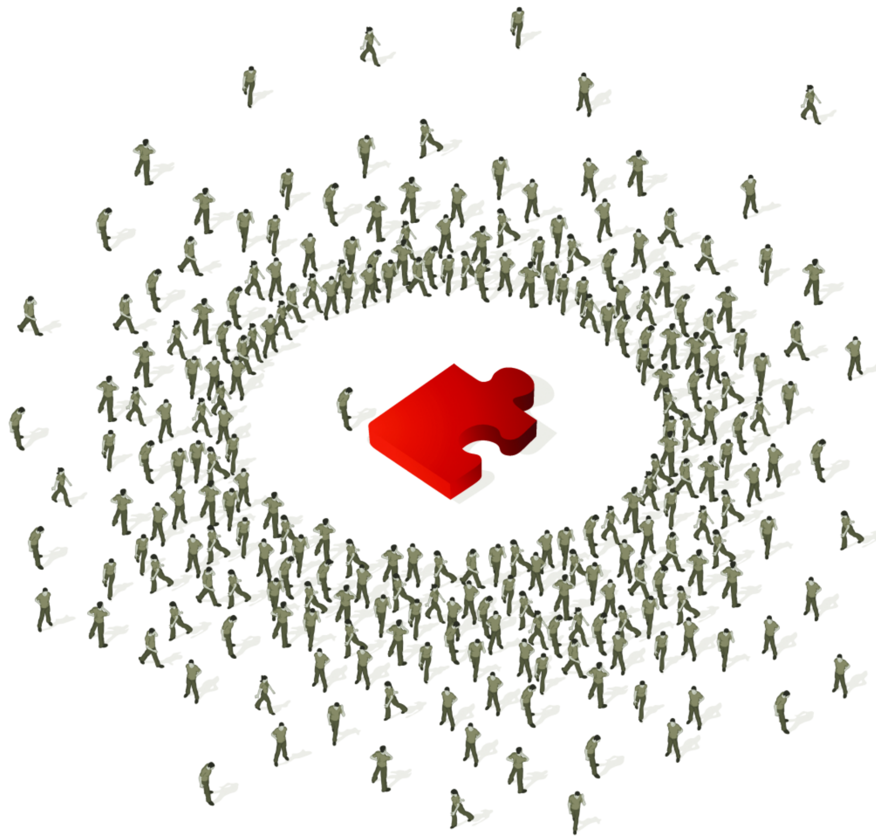
The method:

- Weights assigned to properties by using entropy and mutual information (correct but not scalable)
- Estimation of the weights based on composite likelihood
  - less sensitive to overfitting
  - allow to handle big data sources

[Bergamaschi S. et al. 2014]

Future Work

- Active integration by crowdsourcing



Future Work

- Source exploration tool

DATA AND TOOLS

Data.gov

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

Browse Raw Datasets

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

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!

Slides partially taken from the ICDE 2013 tutorial of my colleague Divesh Srivastava

Thank You!

Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari" – Università di Modena e Reggio Emilia

References

Motivation:

[Seth 2014] Amit Seth: "Transforming Big Data into Smart Data " , 2014 www.slideshare.net/apsheth/transforming-big-data-into-smart-data-deriving-value-via-harnessing-volume-variety-and-velocity-using-semantic-techniques-and-technologies

[LOD14] http://lod-cloud.net/

[Benedetti et al. 2014] F. Benedetti, S. Bergamaschi, and L. Po, "A visual summary for linked open data sources," 2014, to appear in International Semantic Web Conference (Posters & Demos).

[Bergamaschi et al. 2014] F. Benedetti, S. Bergamaschi, and L. Po, "Online index extraction from linked open data sources," 2014, to appear in Linked Data for Information Extraction (LD4IE) Workshop held at International Semantic Web Conference.

[Dong and Srivastava 2013] X.L. Dong, and D. Srivastava. "Big data integration." ICDE, 2013 IEEE 29th

[Bergamaschi et al. 1999] Sonia Bergamaschi, Silvana Castano, and Maurizio Vincini. "Semantic integration of semistructured and structured data sources." ACM Sigmod Record 28.1 (1999): 54-59.

[Dong et al. 2013] 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)

Schema alignment:

[Bergamaschi et al. 2011] Sonia Bergamaschi, et al. "Keyword search over relational databases: a metadata approach." Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011.

References

[Bergamaschi et al. 2013] Sonia Bergamaschi, Francesco Guerra, and Giovanni Simonini. "Keyword Search over Relational Databases: Issues, Approaches and Open Challenges." Bridging Between Information Retrieval and Databases. Springer Berlin Heidelberg, 2014. 54-73.

[Talukdar et al. 2008] 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)

[Sarma et al. 2008] Anish Das Sarma, Xin Dong, Alon Y. Halevy: Bootstrapping pay-as-you-go data integration systems. SIGMOD Conference 2008: 861-874

Record linkage

[Kannan et al. 2008] Anitha Kannan, Inmar E. Givoni, Rakesh Agrawal, Ariel Fuxman: Matching unstructured product offers to structured product specifications. KDD 2011: 404-412

[Kolb et al. 2012] Lars Kolb, Andreas Thor, Erhard Rahm: Load Balancing for MapReduce-based Entity Resolution. ICDE 2012: 618-629

[Alexandrov et al. 2014] Alexander Alexandrov , Rico Bergmann , Stephan Ewen , Johann-Christoph Freytag , Fabian Hueske , Arvid Heise , Odej Kao , Marcus Leich , Ulf Leser , Volker Markl , Felix Naumann , Mathias Peters , Astrid Rheinländer , Matthias J. Sax , Sebastian Schelter , Mareike Höger , Kostas Tzoumas , Daniel Warneke, “The Stratosphere platform for Big Data Analytics” . VLDB Journal 2014

[Matel et al. 2012] Zaharia, Matei, et al. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing." Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012.

References

[Borkar et al. 2011] Borkar, V., Carey, M., Grover, R., Onose, N., & Vernica, R. (2011, April). Hyracks: A flexible and extensible foundation for data-intensive computing. In Data Engineering (ICDE), 2011 IEEE 27th International Conference on (pp. 1151-1162). IEEE.

Data fusion:

[Dong et al. 2009] Xin Luna Dong, Laure Berti-Equille, Divesh Srivastava: Integrating Conflicting Data: The Role of Source Dependence. PVLDB 2(1): 550-561 (2009)

Emerging topics:

[Dong X.L. et al. 2013] Xin Luna Dong, Barna Saha, Divesh Srivastava: Less is More: Selecting Sources Wisely for Integration. PVLDB 6(2): 37-48 (2013)

[Bergamaschi S. et al. 2014] S. Bergamaschi, D. Ferrari, F. Guerra, G. Simonini "Discovering the topics of a data source: a statistical approach " 2014, Surfacing the Deep and the Social Web (SDSW) Workshop held at ISWC