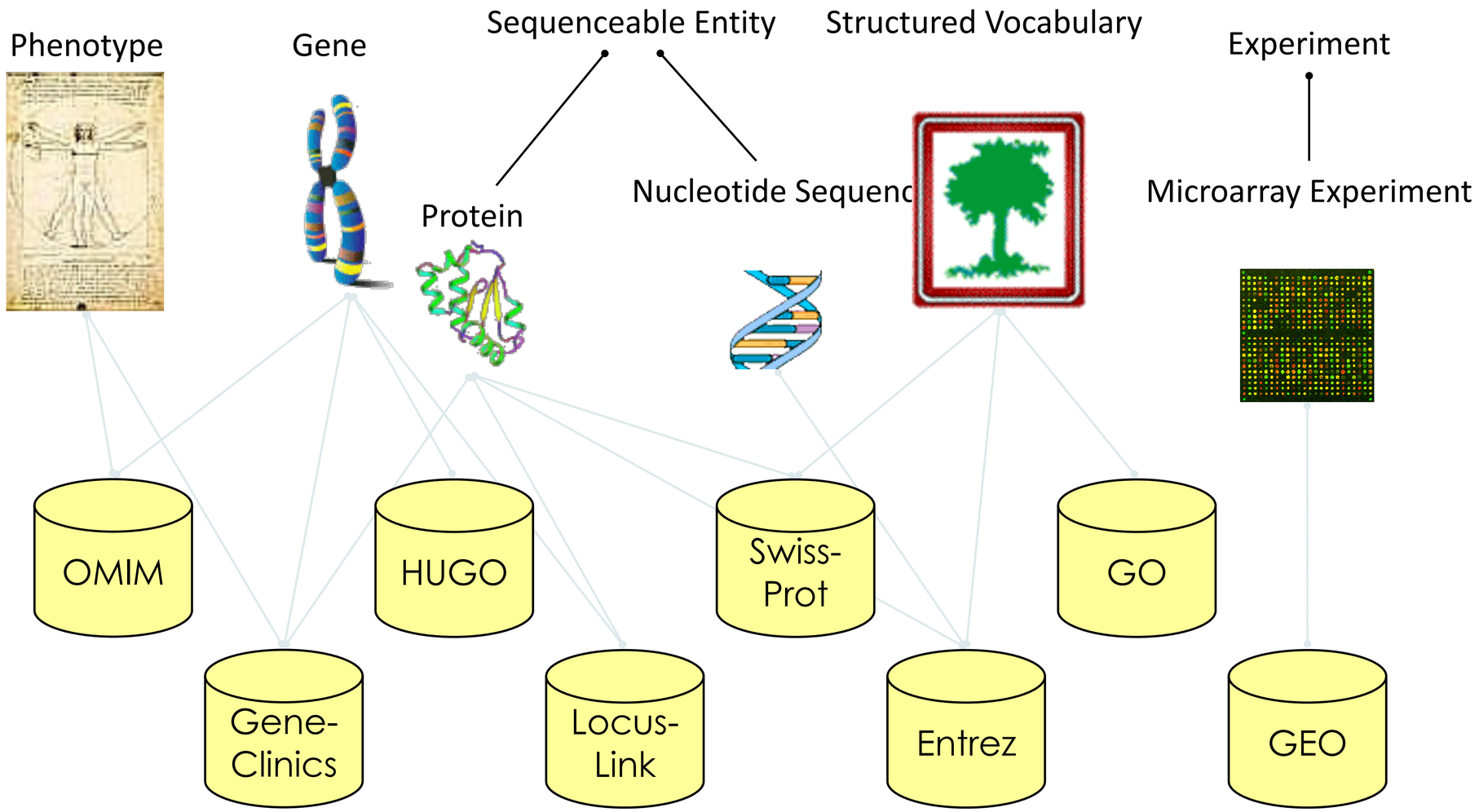
Advanced Data Management (CSCI 490/680)

Data Discovery

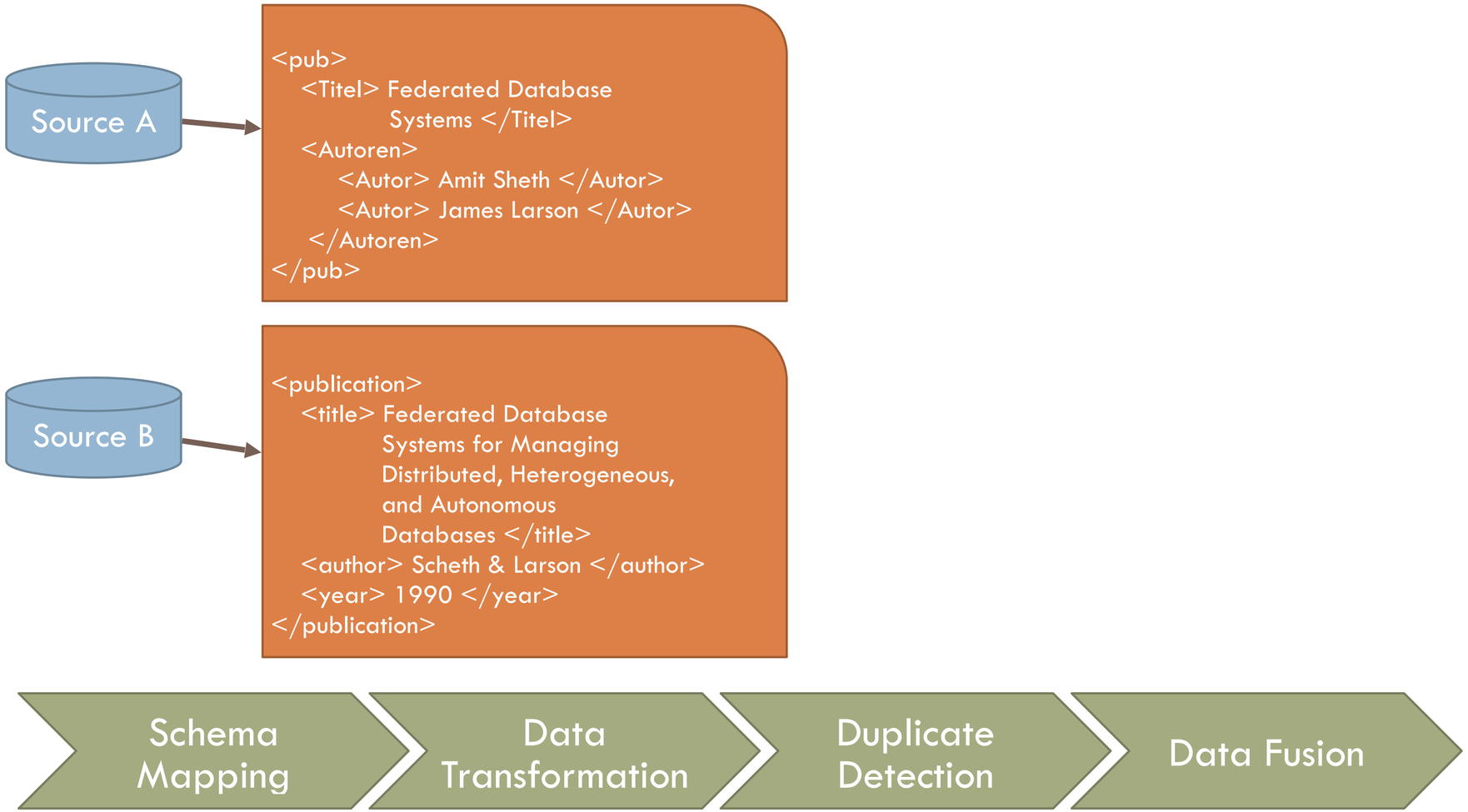
Dr. David Koop

Data Integration: Combine Datasets with Different Data



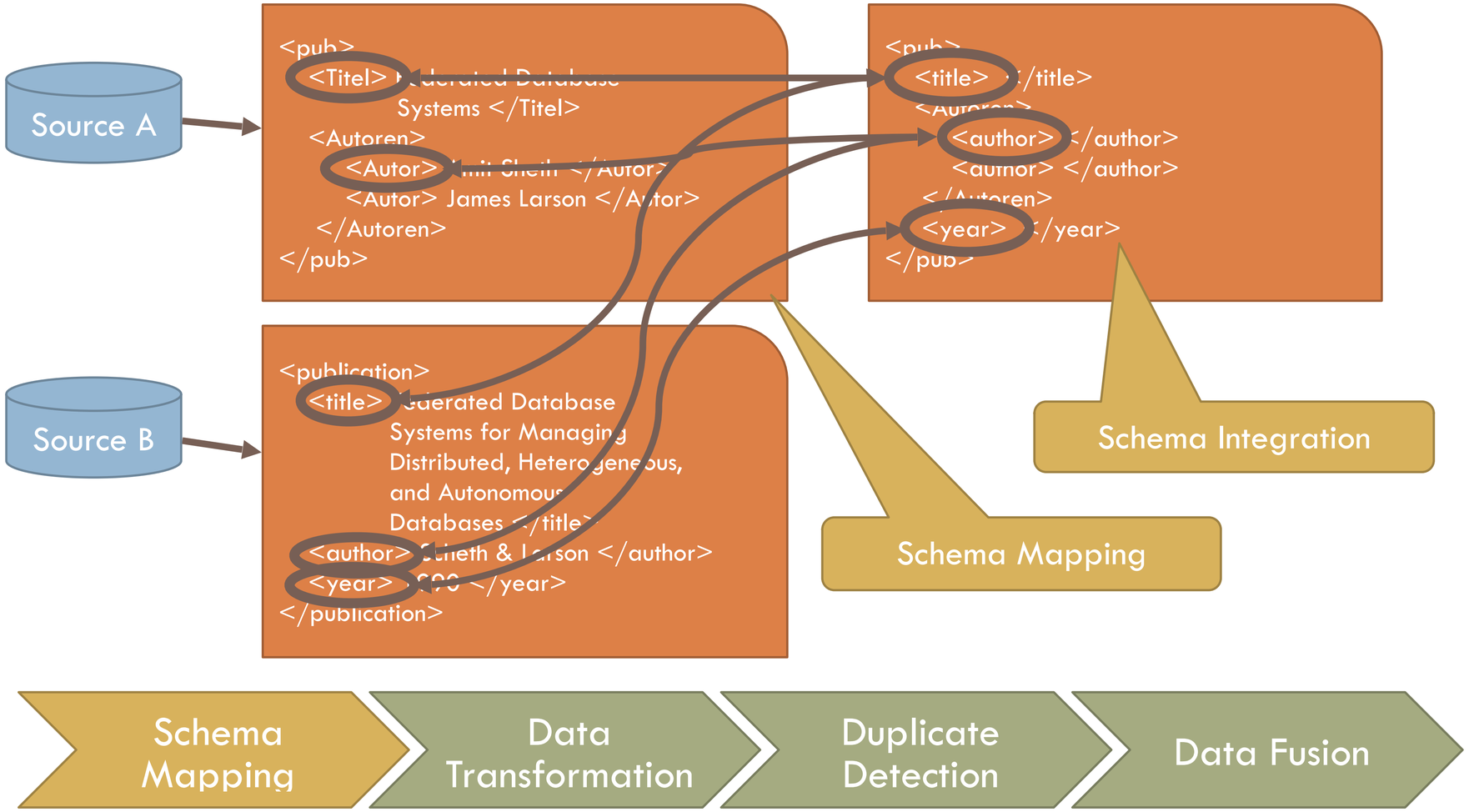
[A. Doan et al., 2012]

Information Integration



[L. Dong and F. Naumann, 2009]

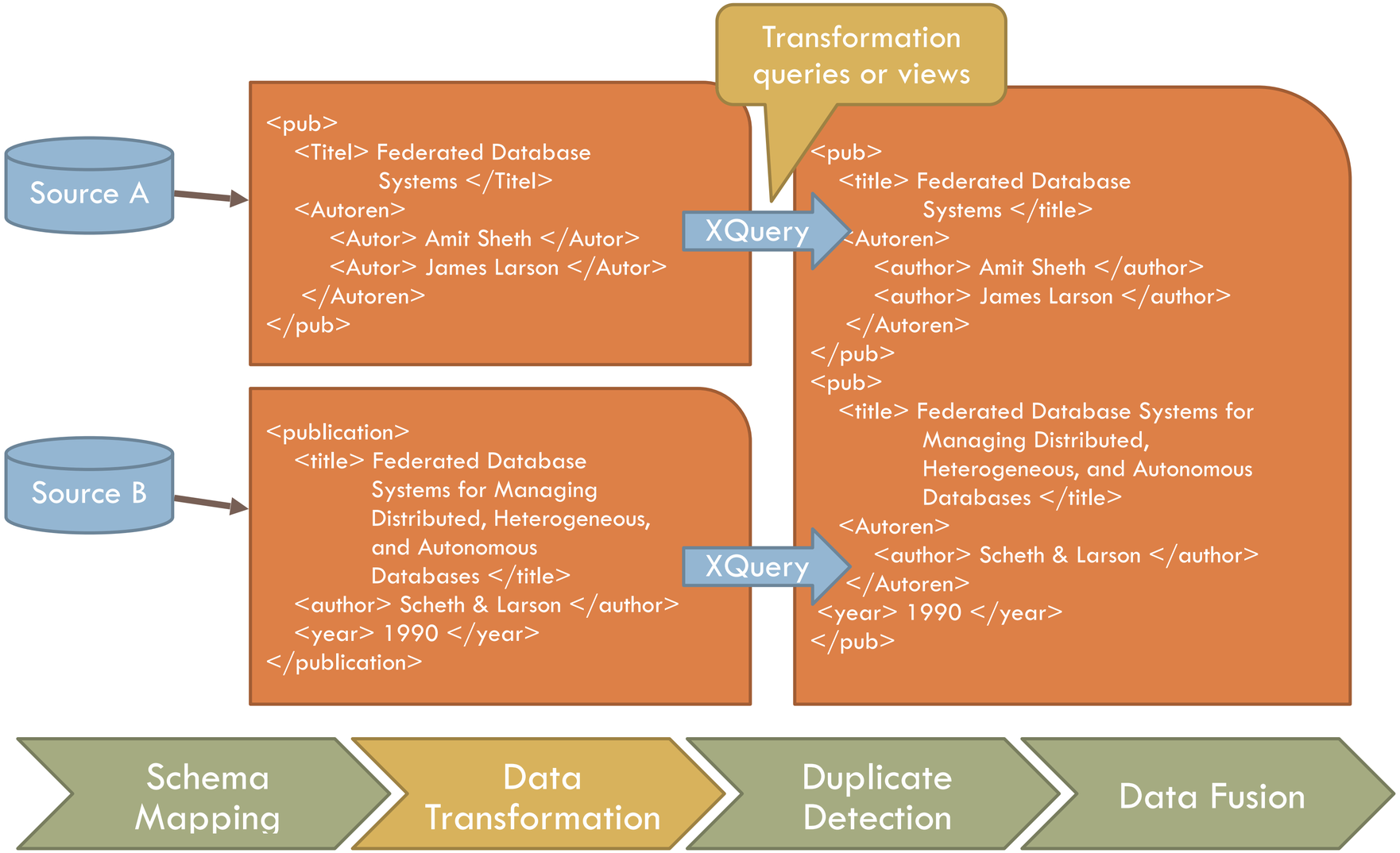
Information Integration



Schema Mapping → Data Transformation → Duplicate Detection → Data Fusion

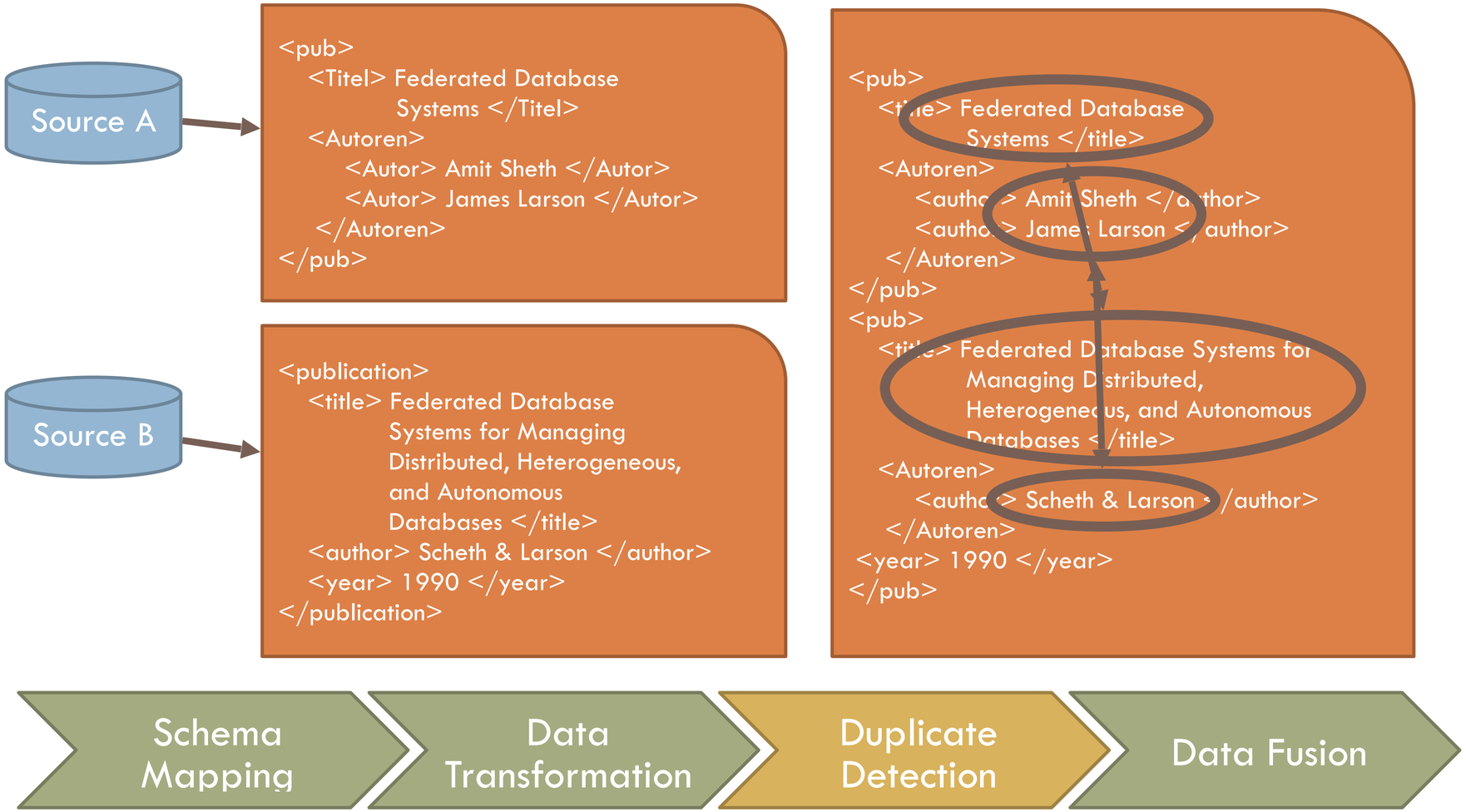
[L. Dong and F. Naumann, 2009]

Information Integration



[L. Dong and F. Naumann, 2009]

Information Integration



[L. Dong and F. Naumann, 2009]

"Duplicate Detection" has many Duplicates

[L. Dong and F. Naumann, 2009]

"Duplicate Detection" has many Duplicates

Household matching

Doubles

Duplicate detection

Mixed and split citation problem

Record linkage

Object identification

Match

Deduplication

Fuzzy match

Object consolidation

Entity resolution

Entity clustering

Approximate match

Identity uncertainty

Reference reconciliation

Merge/purge

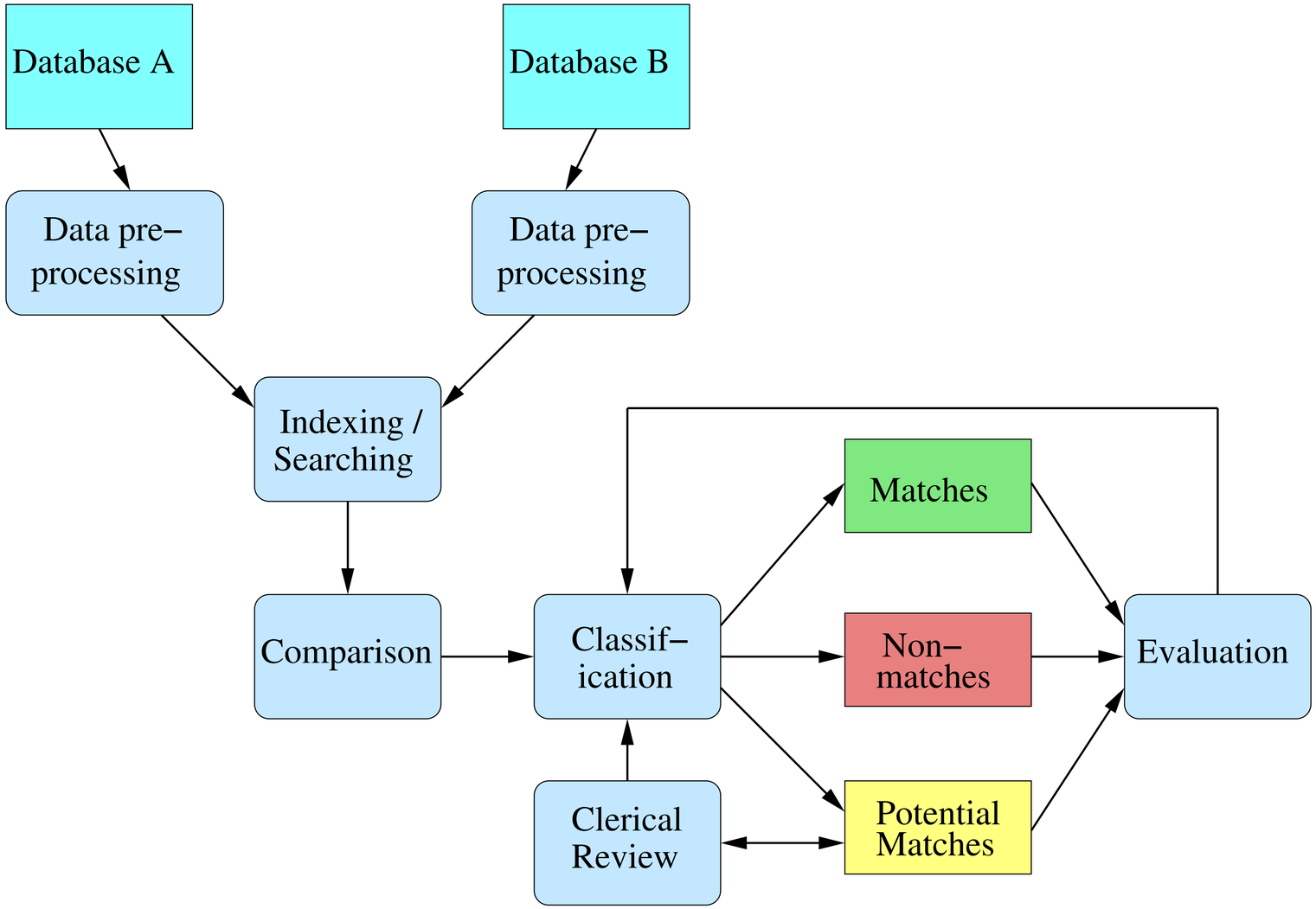
Hardening soft databases

Reference matching

Householding

[L. Dong and F. Naumann, 2009]

Record Linkage Process



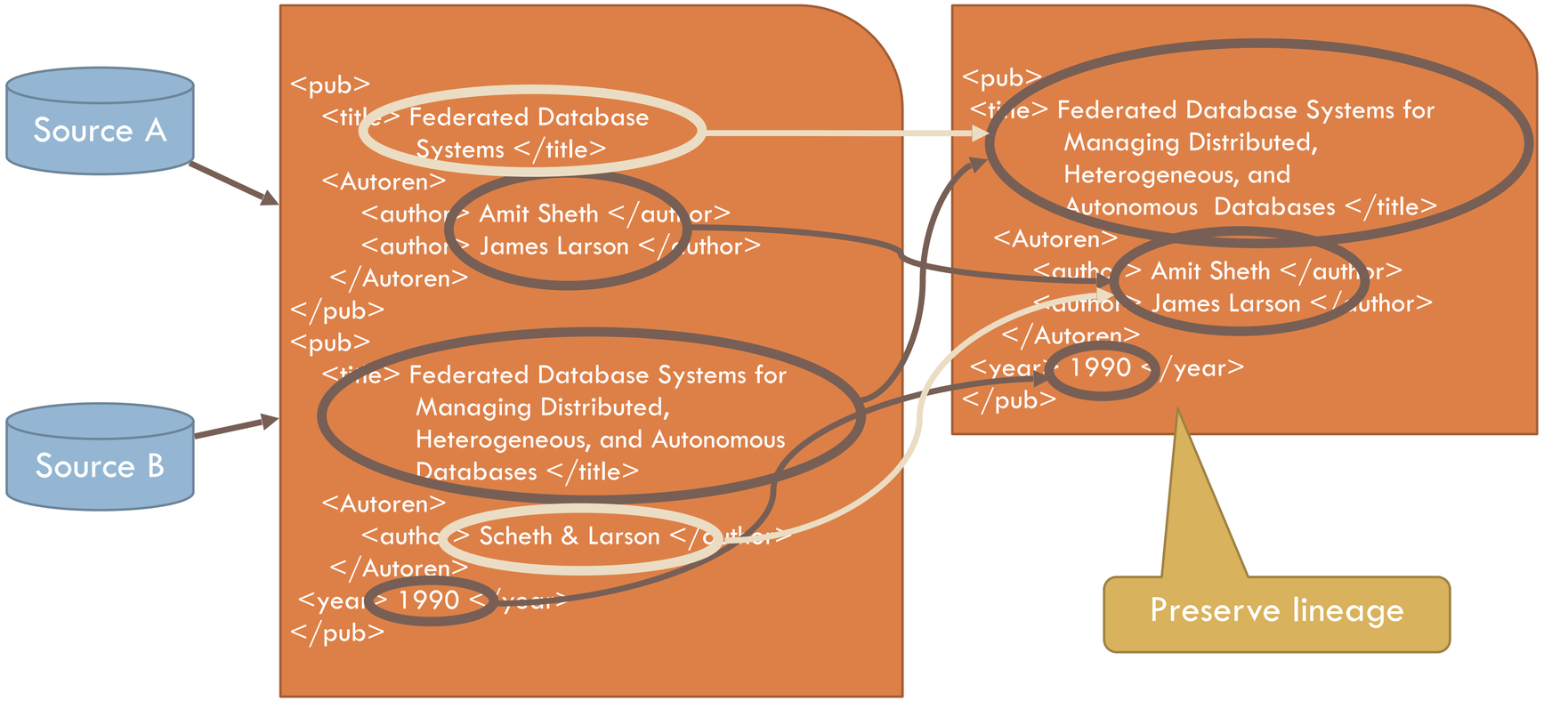
[P. Christen , 2019]

Record Linkage Techniques

- Deterministic matching
  - Rule-based matching (complex to build and maintain)
- Probabilistic record linkage [Fellegi and Sunter, 1969]
  - Use available attributes for linking (often personal information, like names, addresses, dates of birth, etc.)
  - Calculate match weights for attributes
- “Computer science” approaches
  - Based on machine learning, data mining, database, or information retrieval techniques
  - Supervised classification: Requires training data (true matches)
  - Unsupervised: Clustering, collective, and graph based

[P. Christen , 2019]

Information Integration



[L. Dong and F. Naumann, 2009]

Assignment 3

- Same Info Wanted data
- Data wrangling with
  - Trifacta Wrangler
  - pandas
- For place, date extraction: 2 regexs, don't try to standardize anything, CS680 need to extract place details, date is EC
- Trifacta # of Rows Issue
- Due Wednesday, March 3

recid	order	date	place	state
1 - 41.23k	1 - 5	1 - 1.87k	5,431 Categories	44 Categories
38575	1	null	MA, BROOKLINE	MA
34452	1	1857	NY, NYC	NY
34453	1	1857	NY, NYC	NY
34454	1	1857	NY, NYC	NY
35259	1	1855	OH, CINCINATTI	OH
37781	1	1864	MA, ABINGTON	MA
37781	2	05/67	MA, BOSTON	MA
37781	3	null	CA	CA
39120	1	null	TX, MILLICAN	TX
34455	1	null	AUSTRALIA	null
34776	1	null	IL, CHICAGO	IL
34881	1	64	NY, BINGHAMPTON, BROOME CO.	NY
35309	1	1860	IL	IL
35537	1	1861	MA, BOSTON	MA
34757	1	null	TN, NASHVILLE	TN
38439	1	null	MA, BOSTON	MA
38439	2	null	CA, SAN FRANCISCO	CA
41070	2	null	CINCINNATI	null
33438	1	1862	MA, BOSTON	MA
33478	1	10/64	AL, MOBILE	AL
33478	2	null	IL, ST. TRELIA	IL
33940	1	1857	NC	NC
34331	1	02/65	MA, BOSTON	MA
33693	1	null	NY	NY
33693	2	null	CANADAS	null
34306	1	02/65	MA, BOSTON	MA
36900	1	null	PA, PHILADELPHIA	PA
37541	1	null	AUSTRALIA, SIDNEY	null
33485	1	1858	MA, NEW BEDFORD	MA

Test 1

- Comments at the end of class

Integrating Conflicting Data: The Role of Source Dependence

X. L. Dong, L. Berti-Equille, and D. Srivastava

Example Problem

[X L Dong et al., 2009]

Example Problem

	S1	S2	S3
Stonebraker	MIT	Berkeley	MIT
Dewitt	MSR	MSR	UWisc
Bernstein	MSR	MSR	MSR
Carey	UCI	AT&T	BEA
Halevy	Google	Google	UW

[X L Dong et al., 2009]

Naive Voting Works

	S1	S2	S3
Stonebraker	MIT	Berkeley	MIT
Dewitt	MSR	MSR	UWisc
Bernstein	MSR	MSR	MSR
Carey	UCI	AT&T	BEA
Halevy	Google	Google	UW

[X L Dong et al., 2009]

Naive Voting Only Works if Data Sources are Independent

[X L Dong et al., 2009]

Naive Voting Only Works if Data Sources are Independent

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

S4 and S5 copy from S3

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

S4 and S5 copy from S3

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

2. With only a snapshot it is hard to decide which source is a copier.

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

1. Sharing common data does not in itself imply copying.

2. With only a snapshot it is hard to decide which source is a copier.

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

3. A copier can also provide or verify some data by itself, so it is inappropriate to ignore all of its data.

[X L Dong et al., 2009]

Source Dependence

- Source dependence: two sources S and T deriving the same part of data directly or transitively from a common source (can be one of S or T).
  - Independent source
  - Copier
    - copying part (or all) of data from other sources
    - may verify or revise some of the copied values
    - may add additional values
- Assumptions
  - Independent values
  - Independent copying
  - No loop copying

[X L Dong et al., 2009]

Core Case

- Conditions
  - Same source accuracy
  - Uniform false-value distribution
  - Categorical value
- Proposition: W. independent “good” sources, Naïve voting selects values with highest probability to be true.

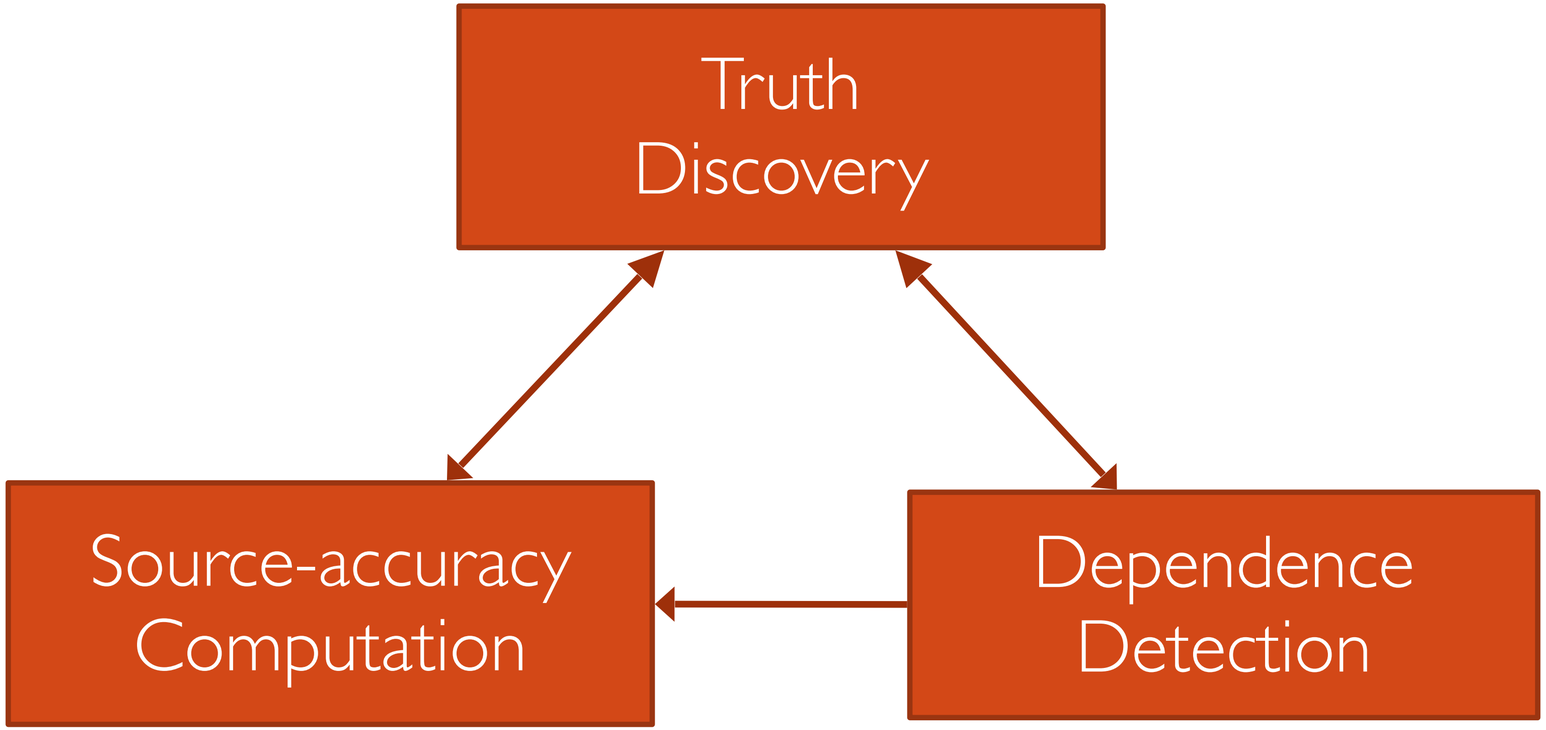
[X L Dong et al., 2009]

Ideas

- If two sources share a lot of false values, they are more likely to be dependent.
- S1 is more likely to copy from S2, if the accuracy of the common data is highly different from the accuracy of S1.

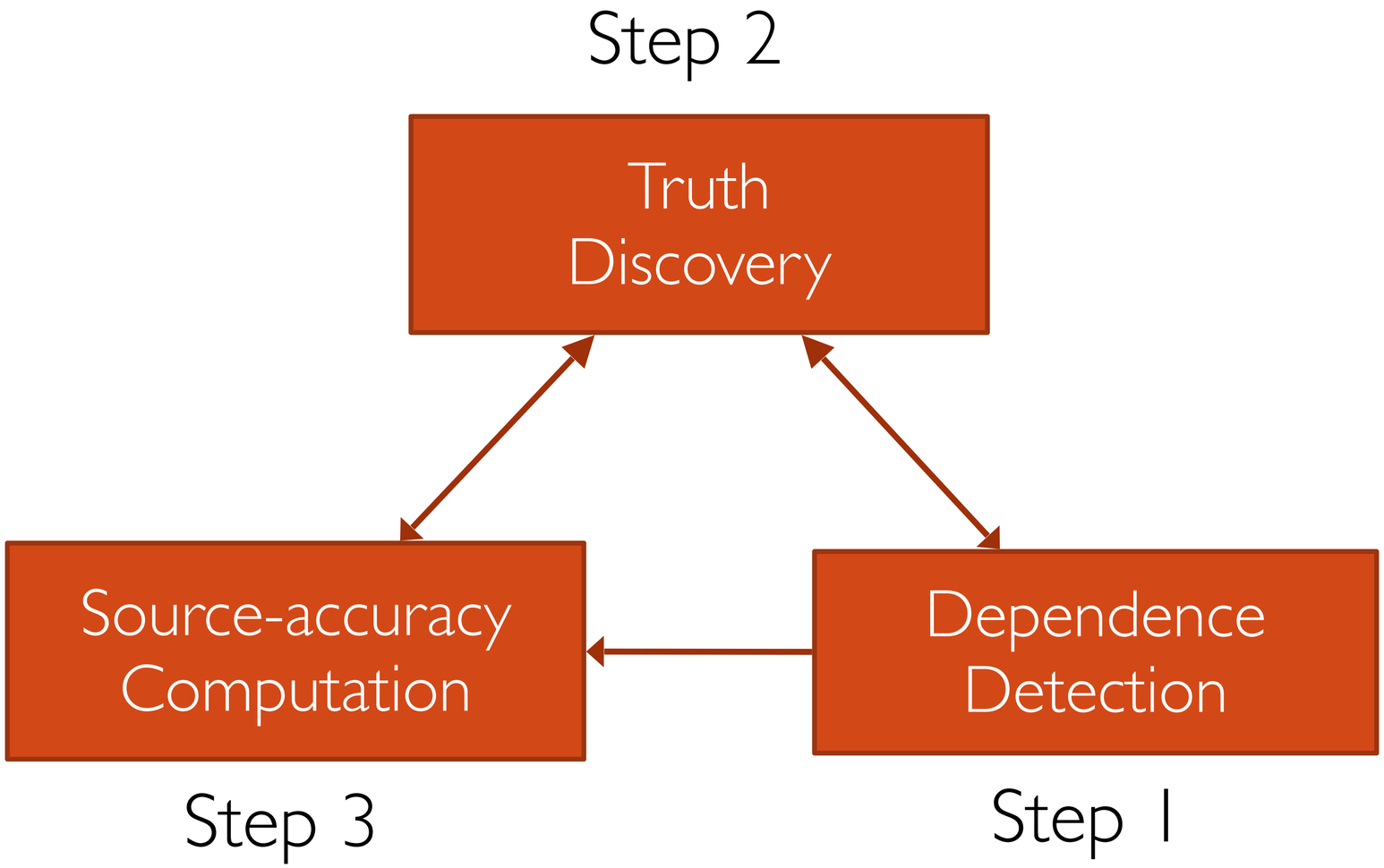
[X L Dong et al., 2009]

Combining Accuracy and Dependence



[X L Dong et al., 2009]

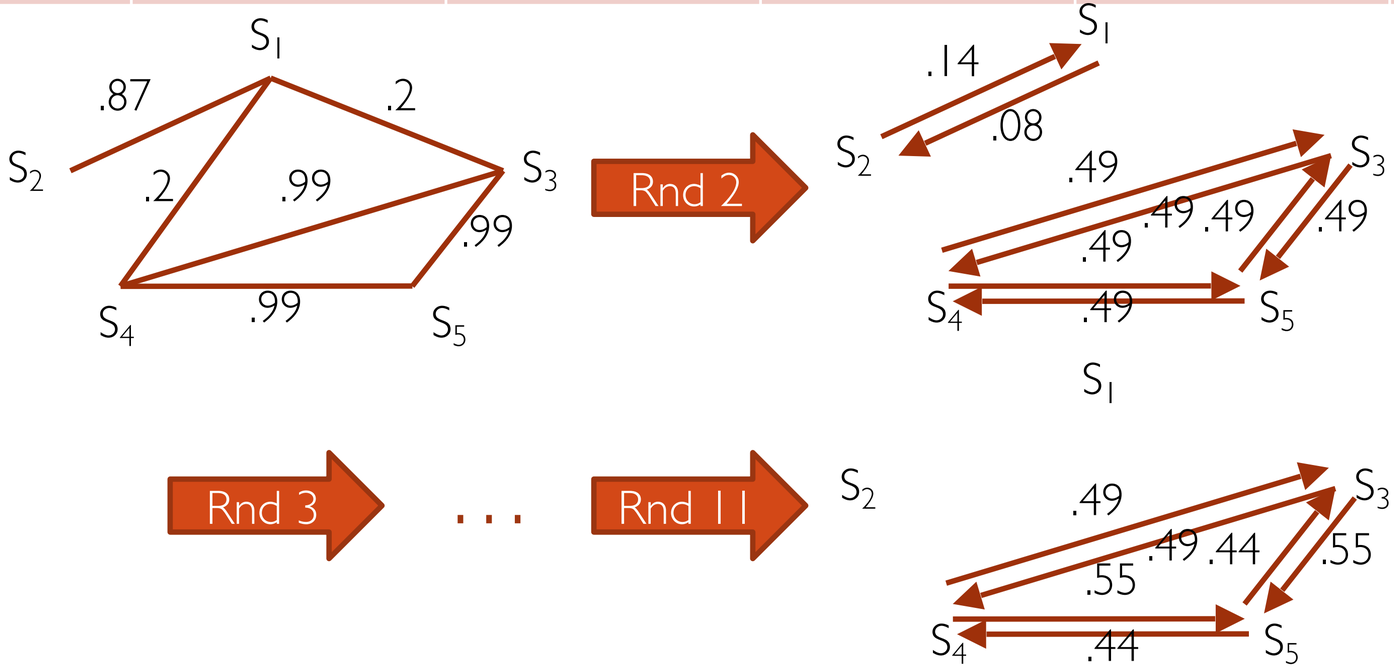
Combining Accuracy and Dependence



[X L Dong et al., 2009]

The Motivating Example

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW



[X L Dong et al., 2009]

The Motivating Example

Accuracy	S1	S2	S3	S4	S5
Round 1	.52	.42	.53	.53	.53
Round 2	.63	.46	.55	.55	.55
Round 3	.71	.52	.53	.53	.37
Round 4	.79	.57	.48	.48	.31
…	…	…	…	…	…
Round 11	.97	.61	.40	.40	.21

Value Confidence	Carey			Halevy	
	UCI	AT&T	BEA	**Google**	UW
Round 1	1.61	1.61	2.0	2.1	2.0
Round 2	1.68	1.3	2.12	2.74	2.12
Round 3	2.12	1.47	2.24	3.59	2.24
Round 4	2.51	1.68	2.14	4.01	2.14
…	…	…	…	…	…
Round 11	4.73	2.08	1.47	6.67	1.47

[X L Dong et al., 2009]

How do you find data?

What is a dataset?

- SDMX: a collection of related observations, organized according to a predefined structure
- DataCube (W3C): a collection of observations, possibly organized into various slices, conforming to some common dimensional structure
- Data Catalog Vocab: a collection of data, published or curated by a single agent, and available for access or download in one or more formats

- [Chapman et al., 2020]: a collection of related observations organized and formatted for a particular purpose
  - Can be table or images, graphs, documents, etc.

[Chapman et al., 2020]

Goal of Dataset Search: Accurate (A) vs. Timely (B)

A

New York City

ALLIANCE ENERGY	239 10TH AVE	New York	NY	10001
EASTSIDE SERVICE STATION	253 E 2ND ST	New York	NY	10009
BP	21 E HOUSTON ST	New York	NY	10012
FREDERICK BP	2040 FREDERICK DOUGLASS BLVD	New York	NY	10026
ORLANDO TEJEDA	3225 BROADWAY	New York	NY	10027
RIVER DRIVE CAR WASH AND GAS	673 W 125TH ST	New York	NY	10027
SHELL	1599 LEXINGTON AVE	New York	NY	10029
GETTY	348 E 106TH ST	New York	NY	10029
MOBIL ON THE RUN	2165 AMSTERDAM AVE	New York	NY	10032
BROADWAY MOBIL	3740 BRO			
GETTY	89 SAINT			
COCO 4633	3936 10T			
HESS 32517	401 W 20			
BP	2326 1ST			
SHELL	2276 1ST			
BP	255 E 125			
EASTSIDE GAS	1890 PAR			
HESS 32215	502 W 45			
145TH STREET MOBIL	150 W 14			
SHELL	232 W 14			
NEW YORK GETTY	119 W 14			
HESS 32520	120 W 14			
SHELL	1855 1ST			
ADAMS GAS STATION	248 BAY S			
STATEN ISLAND GETTY	1201 VICT			
7-ELEVEN	1252 FOR			
LIBERTY GAS	745 PORT			
FOREST AND RICHMOND CI	1810 FOR			
HESS 32581	2121 FOR			
FOREST GULF	2151 FOR			
BP	1098 RICH			

B

Tweets

NYC GAS @NYC_GAS 30m
RT [redacted]: 30 minute gas line at Shell on Long beach road and Merrick road near South Nassau #ligas
Expand

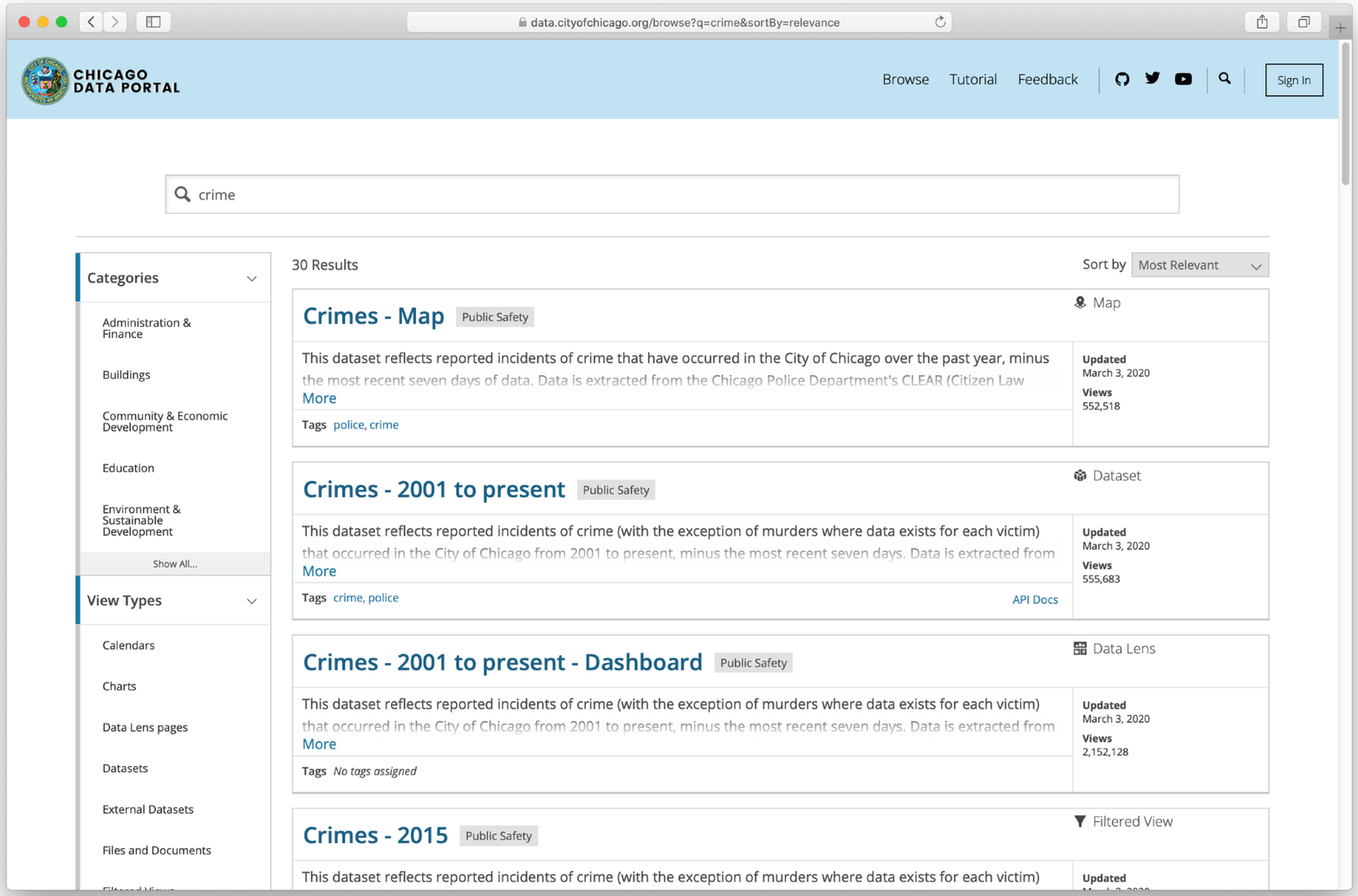
NYC GAS @NYC_GAS 32m
#nycgas #brooklyngas RT [redacted]: 7-Eleven 301 65th & 3rd Brooklyn, NY 11220 gas now
Expand

NYC GAS @NYC_GAS 42m
#siopen RT [redacted]: Mobil station on Richmond Ave & Arthur Kill in Staten Island has gas. Minimal line. Regular only. #sigas
Expand

[redacted] 2h
Gas line not bad at all @45th street Hess on 10th av in NYC. Maybe 10 min.
Retweeted by NYC GAS

[Chapman et al., 2020]

Dataset Search Example



[data.cityofchicago.org]

Dimensions of Data

	Dimension	Categories	Question to be answered
objective	Type	Web Crawler, Customizable Crawler, Search Engine, Pure Data Vendor, Complex Data Vendor, Matching Vendor, Enrichment Tagging, Enrichment Sentiment, Enrichment Analysis, Data Market Place	What is the type of the core offering?
	Time Frame	Static/Factual, Up To Date	Is the data static or real-time?
	Domain	All, Finance/Economy, Bio Medicine, Social Media, Geo Data, Address Data	What is the data about?
	Data Origin	Internet, Self-Generated, User, Community, Government, Authority	Where does the data come from? Who is the author?
	Pricing Model	Free, Freemium, Pay-Per-Use, Flat Rate	Is the offer free, pay-per-use or usable with a flat rate?
	Data Access	API, Download, Specialized Software, Web Interface	What technical means are offered to access the data?
	Data Output	XML, CSV/XLS, JSON, RDF, Report	In what way is the data formatted for the user?
	Language	English, German, More	What is the language of the website? Does it differ from the language of the data?
	Target Audience	Business, Customer	Towards whom is the product geared?
subjective	Trustworthiness	Low, Medium, High	How trustworthy is the vendor? Can the original data source be tracked or verified?
	Size of Vendor	Startup, Medium, Big, Global Player	How big is the vendor?
	Maturity	Research Project, Beta, Medium, High	Is the product still in beta or already established?

[Schomm et al., 2013]

Enriched Data

- Tagging: add searchable keywords
- Sentiment: add information about how people feel about item
- Analysis: start processing the data

Search Process

Query Language	Query Handling	Data Handling	Results Presentation
Structured (DB)	Parser, Optimizer, Evaluator, Executor	Access Managers, Distribution Managers, indexes, system catalogue	Result Set
Keyword (Classic IR)	Parser, Vector representation, ranking	Indexes and files	Ranked List
Entity Search (IR + Semantic Web)	Parser, reasoners, ontologies, evaluator	Indexes; extraction of Hidden Web	Result Set
Tabular (Web Tables)	Parser, optimizers, evaluators reasoners, ontologies, evaluator	Dataset building; Indexes; extraction of Hidden Web	Table

Goods: Organizing Google's Datasets

- Tool for Google to help its employees find internal data
- Keep data where it is, how it is, but extract metadata to aid search
- Challenges:
  - Dataset size and scale: >26 billion datasets
  - Variety: formats (text, csv, Bigtable), storage (GoogleFS, db server)
  - Churn: ~5% of datasets deleted each day
  - Metadata uncertainty: protocol buffers, primary key identification
  - Computing importance: need to understand users
  - Recovering semantics: understanding the data aids metadata extraction

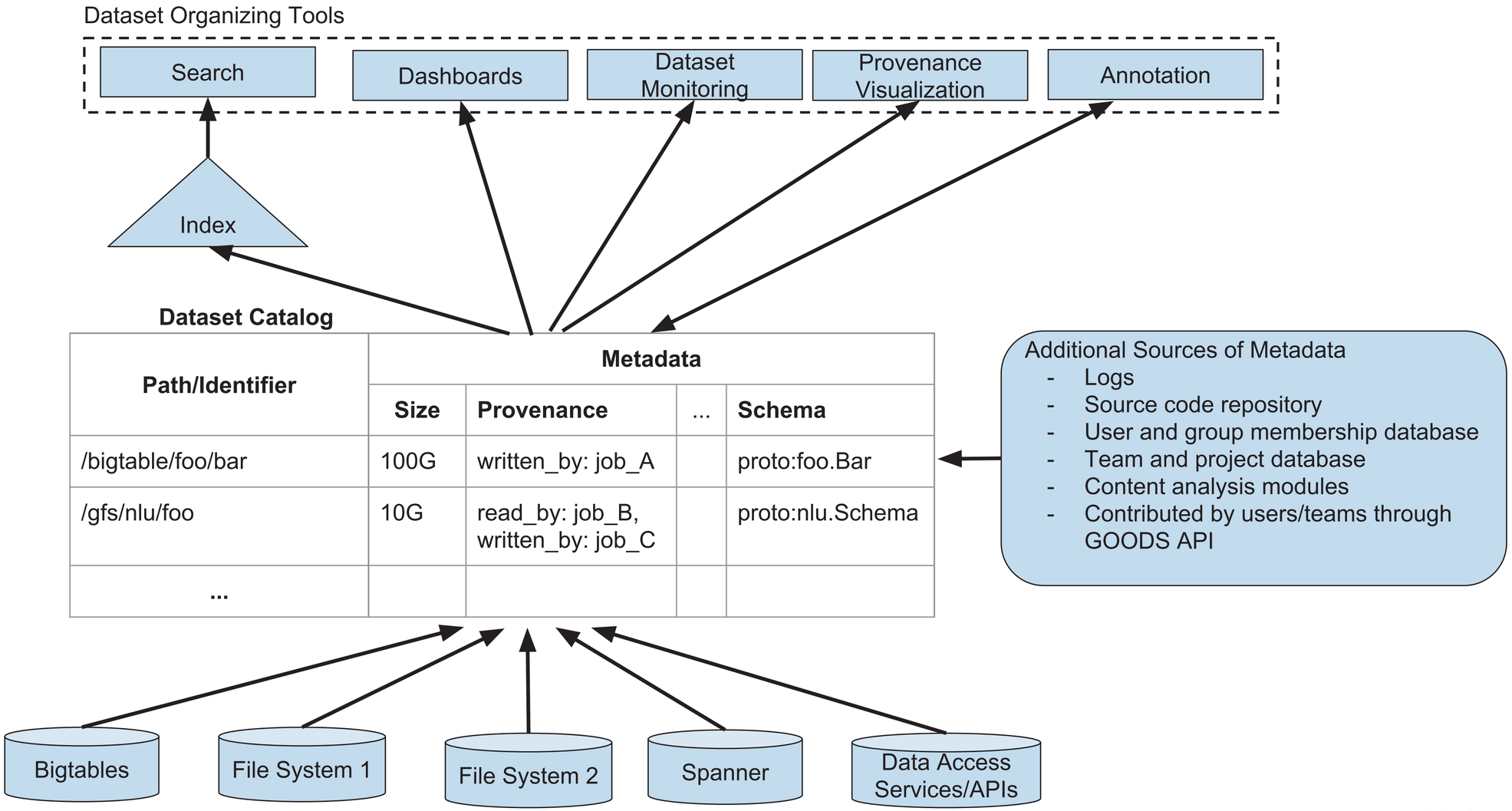
[Halevy et al., 2016]

Goods Metadata Organization

Metadata Groups	Metadata
Basic	size, format, aliases, last modified time, access control lists
Content-based	schema, number of records, data fingerprint, key field, frequent tokens, similar datasets
Provenance	reading jobs, writing jobs, downstream datasets, upstream datasets
User-supplied	description, annotations
Team and Project	project description, owner team name
Temporal	change history

[Halevy et al., 2016]

Goods: Organizing Google's Datasets



[Halevy et al., 2016]

Goods Lessons

- Need for evolution: users bookmark and annotate dataset pages,
- Ranking is domain-specific: a dataset used by another team should be higher
- Expect unusual datasets: metadata extraction can cause crashes
- Data export required: e.g. for visualization
- Ensure recoverability: expensive work so retain snapshots of data

[Halevy et al., 2016]

Google Dataset Search

Google

ski revenue

Updated Date | Download Format | Usage Rights | Free

100+ datasets found

statista Revenue ski & snowboard resorts in the U.S., 2008-2013
www.statista.com
Updated Nov 27, 2018

F Total Revenue for Skiing Facilities, Establishments...
fred.stlouisfed.org
Updated Jan 30, 2020

statista U.S. ski and snowboard rental industry revenue from 2013 to...
www.statista.com
Updated Jul 22, 2019

statista U.S. ski and snowboard rental

Total Revenue for Skiing Facilities, Establishments Subject to Federal Income Tax, Employer Firms

REVEF71392TAXABL

Explore at FRED

Dataset updated Jan 30, 2020

License

https://research.stlouisfed.org/fred_terms.html#copyright-public-domain

Description

Graph and download economic data for Total Revenue for Skiing Facilities, Establishments Subject to Federal Income Tax, Employer Firms (REVEF71392TAXABL) from 19 recreation, employer firms, accounting, revenue, establishments, tax, services, and USA.

[Google Dataset Search]

Google Dataset Search

- Index datasets all over the web (~25 million datasets)
- Use an open standard (schema.org) to describe properties of dataset
- Largest topics: geosciences, biology, and agriculture
- Filter:
  - Updated date
  - Dataset format: tables, images, text
  - Usage Rights
  - Cost

[N. Noy, 2020]

Requirements

- System must be **open** so new providers can add their own datasets
- Search is over **metadata** (a provider may require users to pay/create account)
- Metadata must be published by the data publishers themselves, adhering to a **standard**

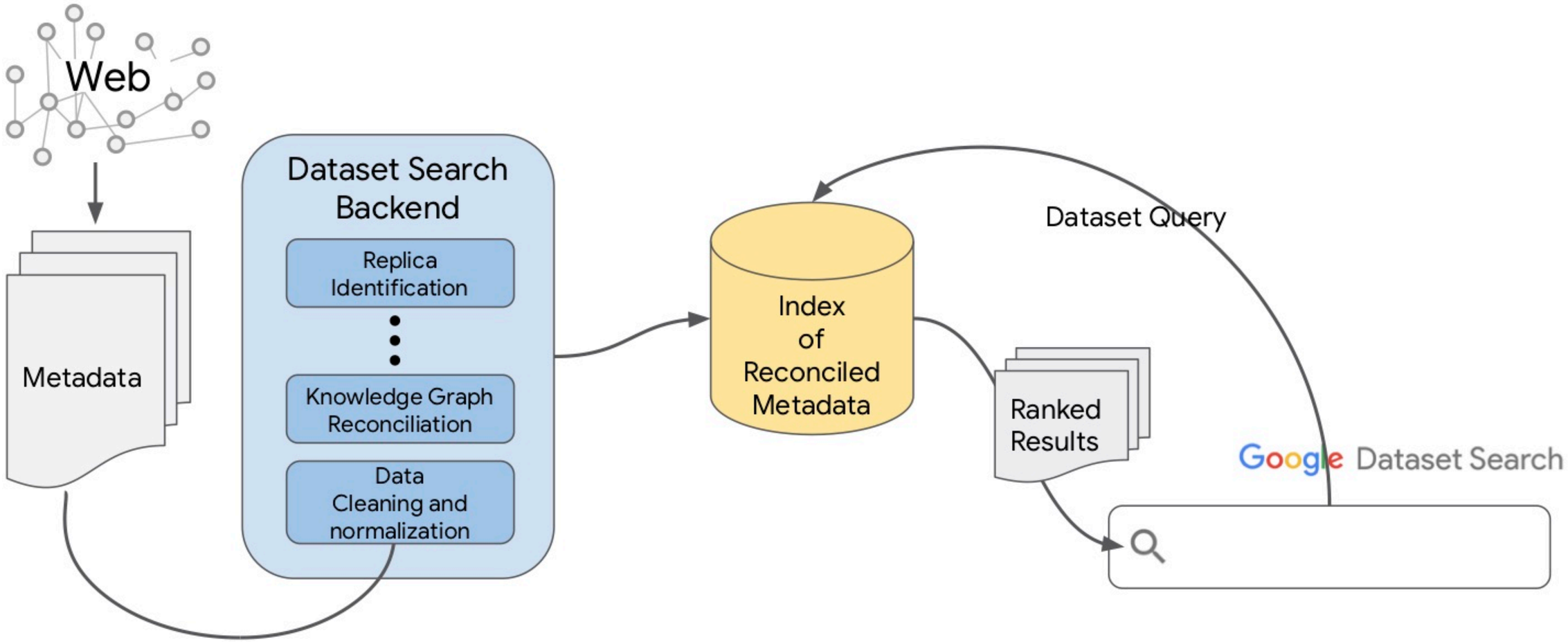
[N. Noy et al., 2019]

Challenges

- Metadata Quality: providers don't adhere to the specs
- Metadata Duplication in Search Results: search results vs. profile pages
- Dataset Replication and Provenance: identify replicas across providers
- Churn and Stale Sites:
  - 3% deleted, 7-10% added per day
  - standard web crawlers check high-traffic sites more often
- Ranking/Relevance: data citation might help
- Multiple Dataset-Metadata Standards: schema.org vs DCAT

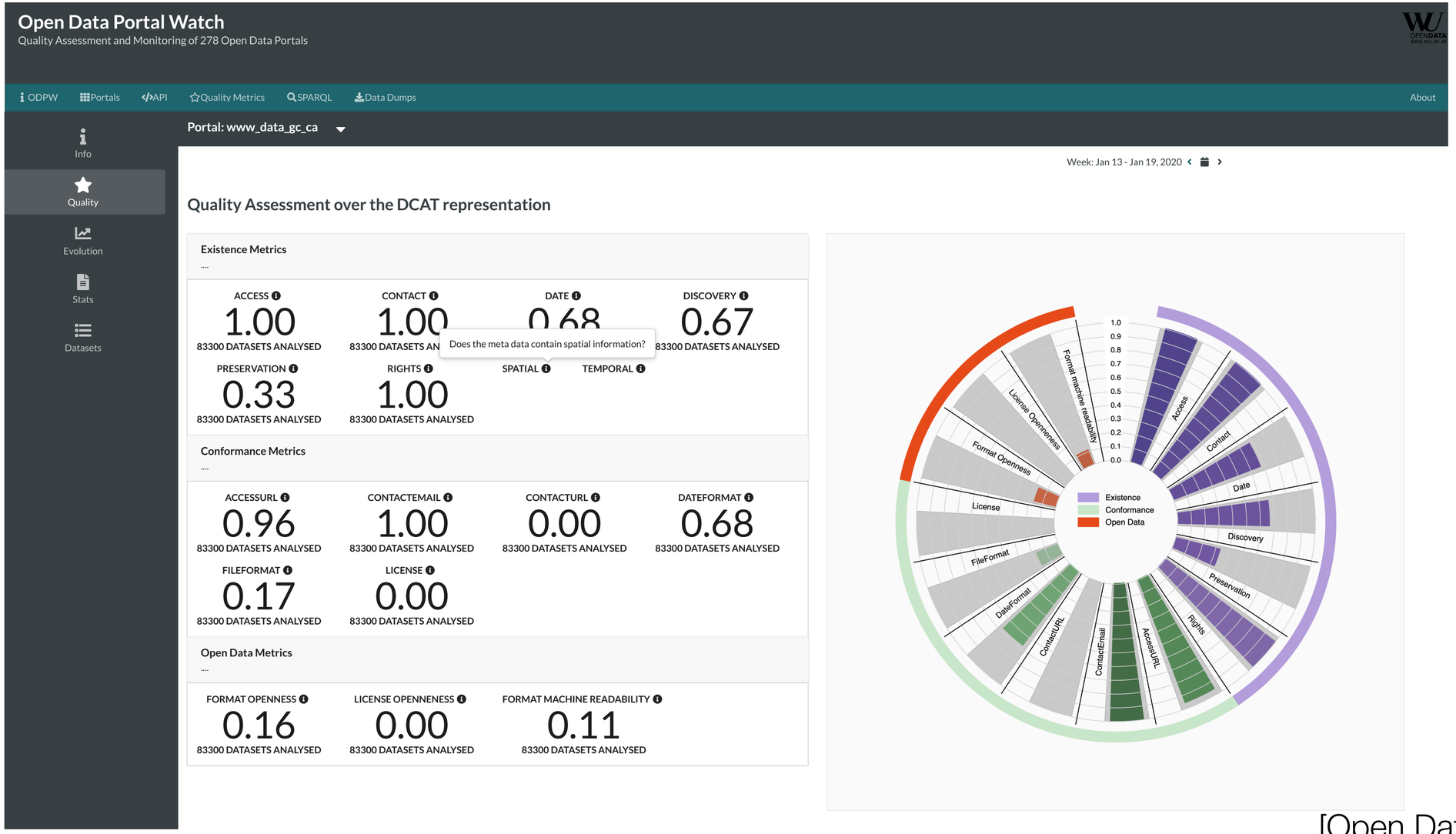
[N. Noy et al., 2019]

Google Dataset Search Overview



[N. Noy et al., 2019]

Dataset Quality Metrics



[Open Data Portal Watch]

Remaining Challenges for Dataset Search

- Query languages: moving beyond keywords
- Query handling: differentiated access
- Data handling: extra knowledge (external and dataset-intrinsic)
- Results presentation: interactivity