"New" Challenges in Data Integration: Unstructured, Large-Scale, Entity-Centric

Lucian Popa
IBM Research – Almaden
January 2018



Outline for the Rest of the Presentation

- **New types of data that need to be accessed and integrated**
  - Non-enterprise / public, unstructured / semi-structured, large-scale, wildly heterogeneous …
  - Very high-value, with appropriate level of integration!

    - Need for domain-specific entity-centric views (e.g., companies, people, products) on top of raw data/documents ...

- **Challenges in building entity-centric views**

    - Make it easier!
    - Need for higher-level abstractions & languages, integrated with learning algorithms

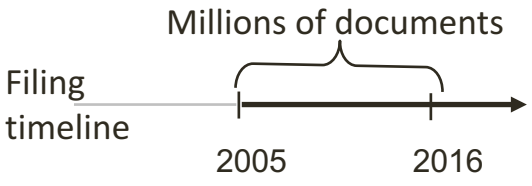
- **Survey recent research on declarative entity linking and integration**

    - Theoretical foundations + A practical language developed at IBM Research
    - Other interesting research directions



An Example Scenario: Building Entities from Public Data

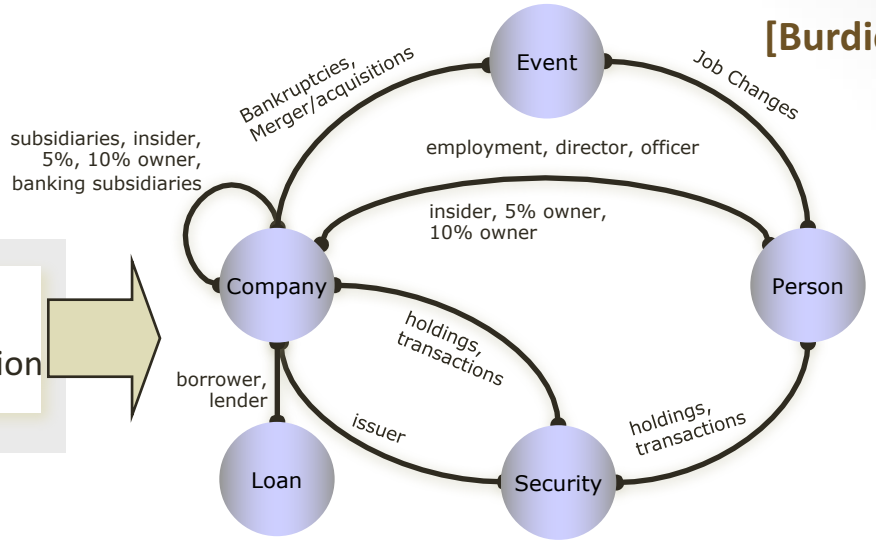
[Burdick et al, 2011]

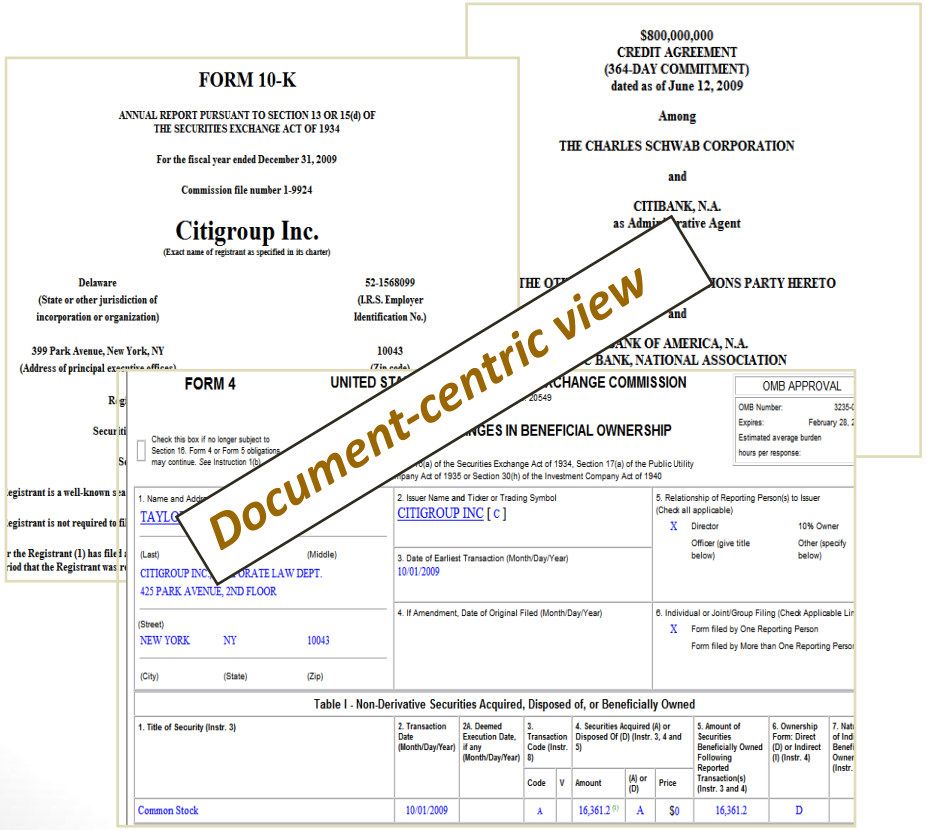


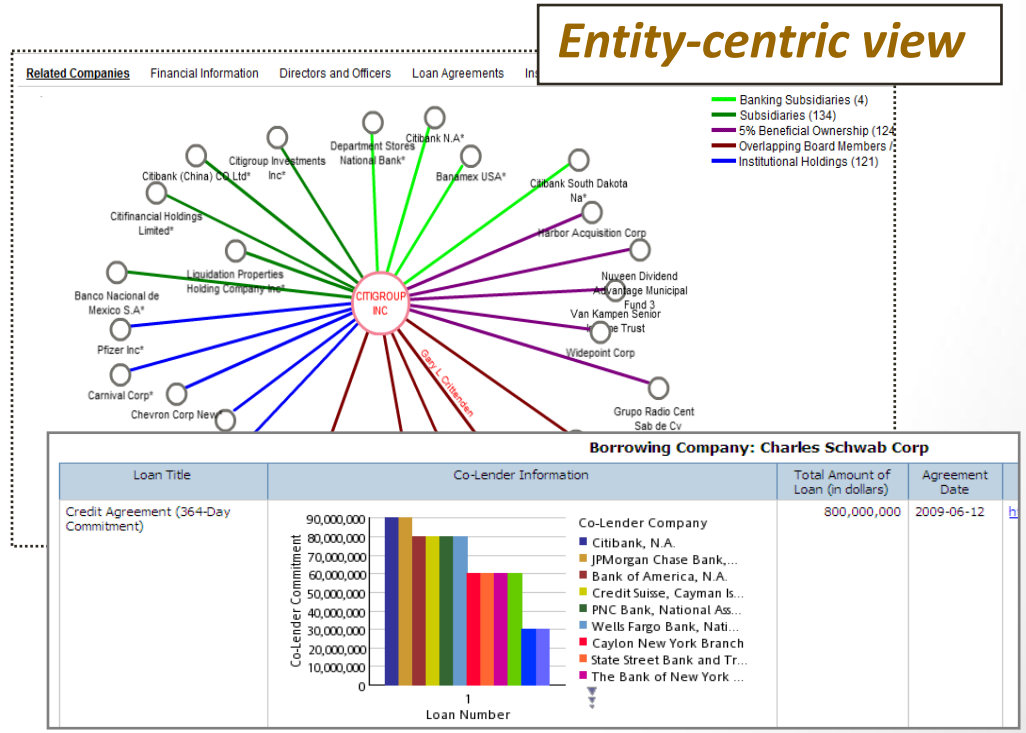
SEC/FDIC Filings of Financial Companies

Extract

Entity Integration







Challenge: Entity Resolution & Integration across Documents

Do these mentions refer to the same person ?

- variability in the person's name, lack of a key identifier
- supporting attributes vary depending on the context

All these facts need to be *linked* and *integrated*

Who Is James Dimon?

Signatures

Insider Transactions

Biographies

Committee memberships

Sincerely,



James Dimon
Chairman and Chief Executive O

Director	Audit	Compensation & Management Development	Corporate Governance & Nominating	Public Responsibility	Risk Policy
Crandall C. Bowles	Member				
Stephen B. Burke		Member	Member		
David M. Cote				Member	Member
James S. Crown				Member	Chair
James Dimon					

Name	Grant date	Approval date	Stock awards: Number of shares of stock or units (#) [2]	Option awards: Number of securities underlying options (#)	Option awards: Exercise price ($/Sh)	Grant date fair value ($)
James Dimon	1/22/2008	1/15/2008	364,048			$14,500,000
	1/22/2008 [3]	1/15/2008		2,000,000 [3]	$39.830	19,868,000 [3]
Michael J. Cavanagh [4]	1/22/2008	1/15/2008	94,151			3,750,000
	1/22/2008	1/15/2008		300,000	39.830	2,980,200
	1/30/2008	N/A		54,271	47.835	360,577
	1/30/2008	N/A		10,625	47.835	91,821
Frank J. Bisignano	1/22/2008	1/15/2008	94,151			3,750,000
	1/22/2008	1/15/2008		300,000	39.830	2,980,200

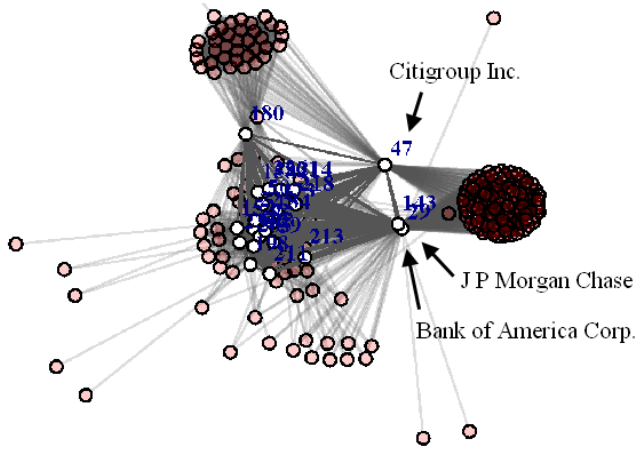


James Dimon, 53, Chairman and Chief Executive Officer of JPMorgan Chase. Director since 2000.

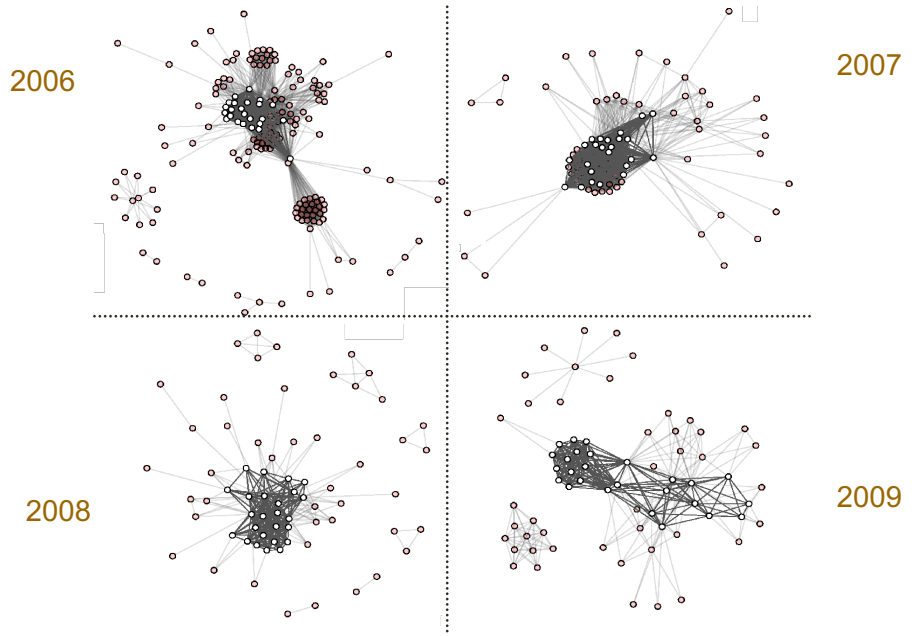
Mr. Dimon became Chairman of the Board on December 31, 2006, and has been Chief Executive Officer and President since December 31, 2005. He had been President and Chief Operating Officer since JPMorgan Chase's merger with Bank One Corporation in July 2004. At Bank One he had been Chairman and Chief Executive Officer since March 2000. Mr. Dimon is a graduate of Tufts University and received an MBA from Harvard Business School. He is a director of The College Fund/UNCF and serves on the Board of Directors of The Federal Reserve Bank of New York, The National Center on Addiction and Substance Abuse, Harvard Business School and Catalyst. He is on the Board of Trustees of New York University School of Medicine.

High-Value Entity-View Leads to Further Insights

Systemic Analysis [Burdick et al, 2011]



Co-lending network for financial institutions (2005)

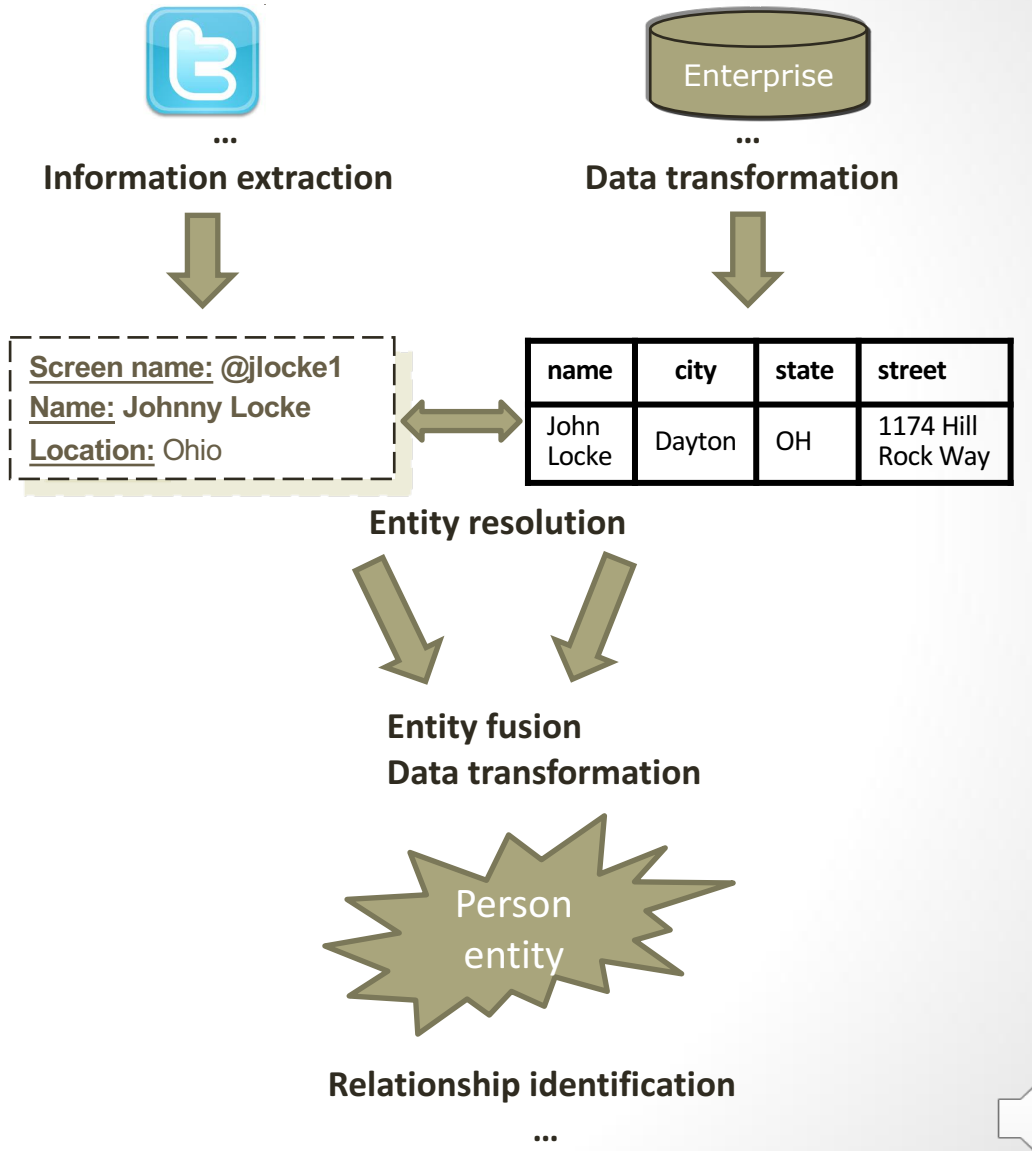


- Entities: **major financial institutions**
- Relationship: **co-lending**, extracted from regulatory SEC filings of companies that borrow money.
  - Gives an indication of the **strength of the connections (stress)** in the global financial system

This global analysis is entirely driven by explicit linkages extracted from public data

Core Operations to Build Entity-Centric Views

- **Information Extraction** (prerequisite)
  - Extraction of the relevant concept mentions (person, locations, job change, etc.) from unstructured documents
- **Entity Resolution (ER)**
  - *Matching of entities in one or more datasets*
    - Semi-structured vs. structured
    - Example: **social media** (sparse, heterogeneous) to **enterprise** (regular, high-quality)
- **Entity fusion**
  - *Attribute-level normalization and fusion*
- **Data transformation/ETL**
  - Before entity resolution, or as part of entity fusion
- **Relationship identification**
  - *Entity linking* (e.g., Bank to a Parent holding company)



Key Challenges

- **How do we make it easier, and more accessible, to build entity-centric algorithms ?**
  - Existing approaches require combinations of technologies & systems:
    - ETL-like systems and database technologies (SQL, XQuery, etc.) for ***data transformations***
    - Dedicated (black-box) engines for ***linking entities***
    - Scripting (PigLatin, Scala, Python) in data-parallel systems (Hadoop, Spark, etc.) for ***scalability***

  Drawbacks:

  - **Data-parallel programming & maintenance requires serious expertise and time**
  - **Hard to reason: integration logic is lost among physical operations and multiple systems**
  - **Not focused on the higher-level entities, more worried about performance and scale.**

Need better abstraction and tools

Towards Better Abstraction: Declarative Approaches

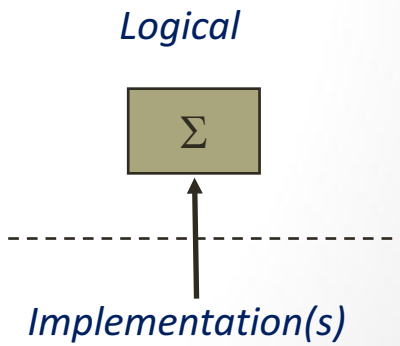


Entity Resolution: Need for Declarative Foundations

- Long-standing research problem, starting with **[Fellegi and Sunter, 1969]**
  - Much work devoted to "physical" algorithms (matching, blocking, clustering, etc.)
  - See surveys **[Elmagarmid et al, 2007], [Ganti and Sarma, 2013].**

- Less work on the logical/declarative foundations of ER.
  - *Other related sub-areas greatly benefit from declarative specifications:*
    - ***Conjunctive queries*** for DB query processing **[Chandra and Merlin, 1977]**
    - ***Constraints*** for consistent query answering (data cleaning) **[Arenas et al, 1999]**
    - ***Constraints*** for schema mapping and data exchange **[Miller, Haas, Hernandez, 1999]**, **[Fagin et al, 2003]**

- Logical specification: expresses, up-front, the **intended results** (the "what").
  - Implementation (the "how") is cleanly separated from the specification.
  - **Ability to reason about systems, including optimization.**



Can we come up with a declarative framework for ER systems ?

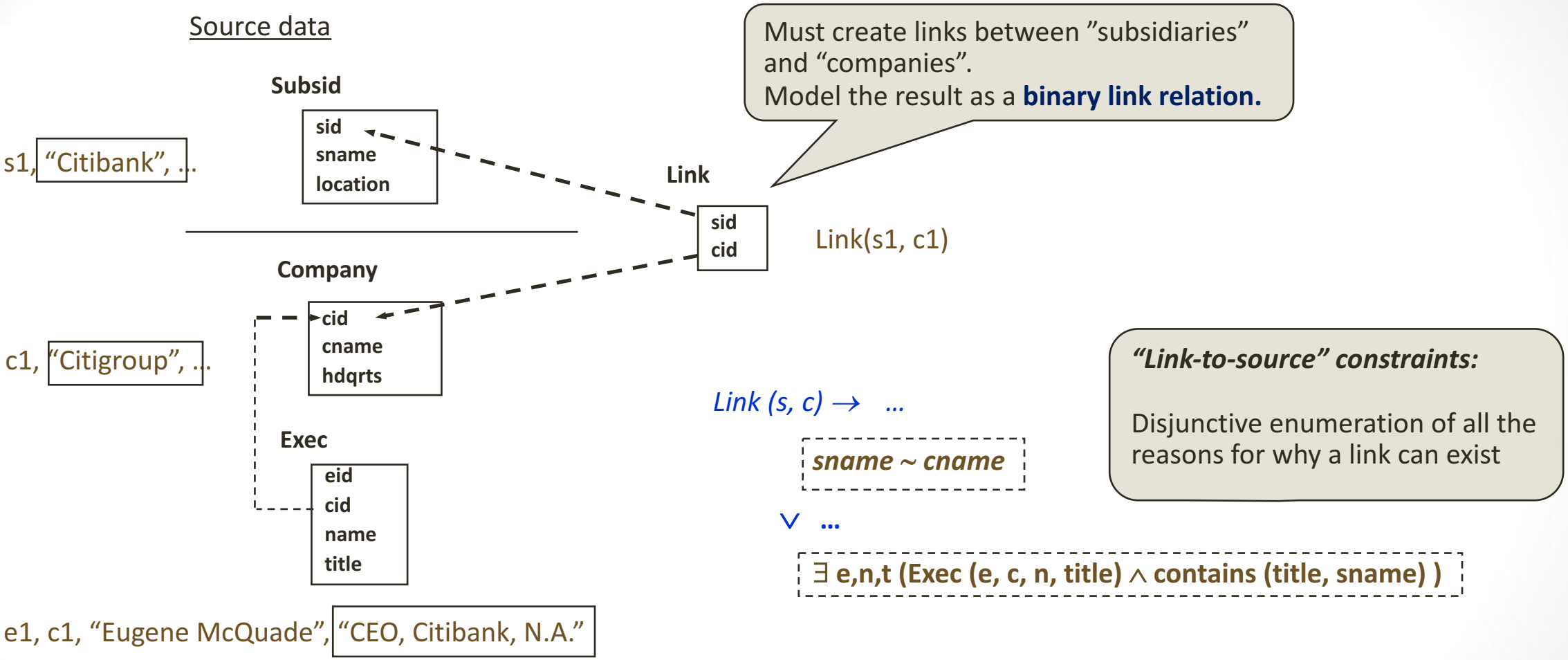
Recent Work

Some recent approaches have started to focus on **specification:**

- **High-level languages, based on rules.**
  - **Dedupalog [Arasu et al, 2009]:** combinations of hard and "soft" constraints, together with heuristics to minimize the number of violations
  - **Not entirely declarative:** The user must be aware of how the system interprets the rules
- **Probabilistic approaches** (e.g., Markov Logic Networks - MLNs) **[Singla and Domingos, 2006]**
  - **FO syntax is a guideline,** to be **interpreted probabilistically**
  - In general, require complex inference; not easy to explain answers
- **Purely declarative, logic-based**: "link-to-source" constraints **[Burdick et al, 2015]**
  - Gives rise to a formal, deterministic, semantics based on ***certain links***



Example of Declarative Entity Linking [Burdick et al, 2015]



Logical specification states the desired ***properties*** of the links (***"what" not "how"***)

Why is Such Framework Useful ?

- Can relate multiple entity-linking specifications via static analysis of the constraints
  - Can define containment / equivalence of entity-linking specifications

- Can analyze the expressive power of entity-linking frameworks under various languages of constraints **[Burdick et al, 2017]**

Probabilistic matching **[Fellegi & Sunter, 1969]** (used in commercial products)

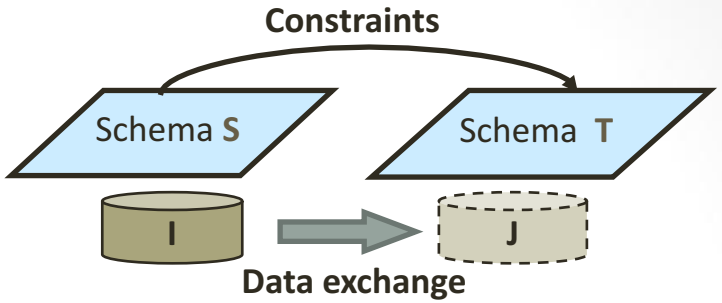
Linear MLNs

- **Surprisingly, probabilistic approaches can be captured by the declarative approach.**
- ***Results/algorithms can be transferred by exploiting such connection*** **[Burdick et al, 2015]**
  - PTIME results from declarative entity linking $\Rightarrow$ PTIME results (not known before) for linear MLNs

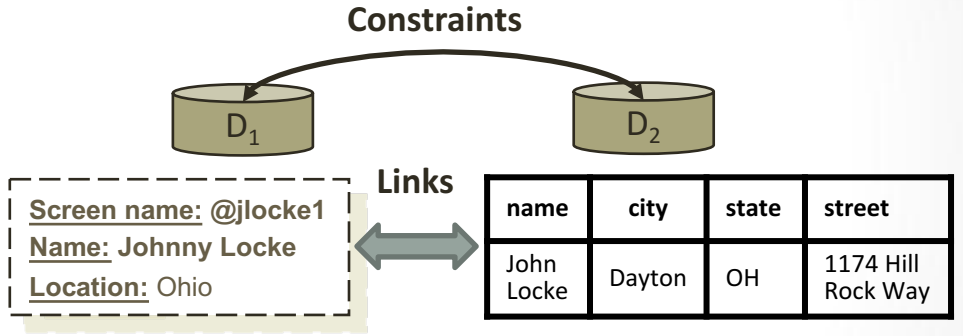
Provides rigorous foundations behind practical languages for data integration …

HIL: A High-Level Language for Entity-Centric Operations

- **Concepts from schema mapping** and **data transformation**.
  - Clio schema mapping work **[Miller, Haas, Hernandez, 1999]**
  - Data exchange based on ***constraints***
    - Foundations for "declarative" data transformation **[Fagin et al, 2003]**

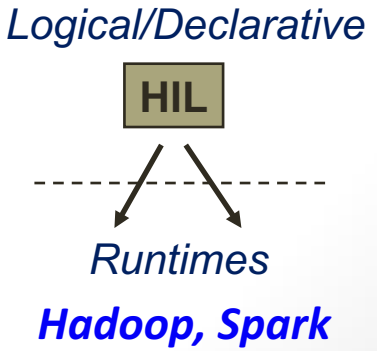


- **Concepts from declarative entity linking.**
  - ***Constraints**-based* specification for matching of entities **[Burdick et al, 2015]**



- **Programming language & compilation techniques.**
  - *Flexible type system* (loosely structured & sparse data, many schemas)
  - *Advanced compilation for large-scale data* **[Hernandez et al, 2013]**

*Ships with **IBM BigMatch**, used in many entity-centric applications.*



HIL Example: Company Matching across Regulatory Agencies (FFIEC vs. SEC)

```
create link FFIEC_SEC as

select  [ffiec_id: f.IDRSSD,  sec_id: s.CIK ]

from FFIEC f,  SEC s

match using
      rule1: toUpper (f.city) = toUpper (s.city) and
            normalize (f.stateprov) = normalize (s.stateprov) and
            companyNameMatch (f.name, s.name),
      rule2:
            …

cardinality ffiec_id 1:1 sec_id;
```

ER Algorithm:
Various combinations of attribute-level matching, filtering and normalizing predicates.

Opportunities for learning!

- **Learn the overall ER algorithm**
- **Learn each attribute structure and variations** (i.e., the logic behind companyNameMatch, normalize, etc.)

California, CA, Calif. → CA

"Isabella Bank" (a bank) — **not a match** — "Isabella Bank Corp" (parent holding company)

"Bank of America N.A." — **match** — "Bank of America National Association"

Key Challenges (Revisited)

- **How do we make it easier, and more accessible, to build entity-centric algorithms ?**
  - Existing approaches require combinations of technologies & systems:
    - ETL-like systems and database technologies (SQL, XQuery, etc.) for ***data transformations***
    - Dedicated (black-box) engines for ***linking entities***
    - Scripting (PigLatin, Scala, Python) in data-parallel systems (Hadoop, Spark, etc.) for ***scalability***

  Drawbacks:

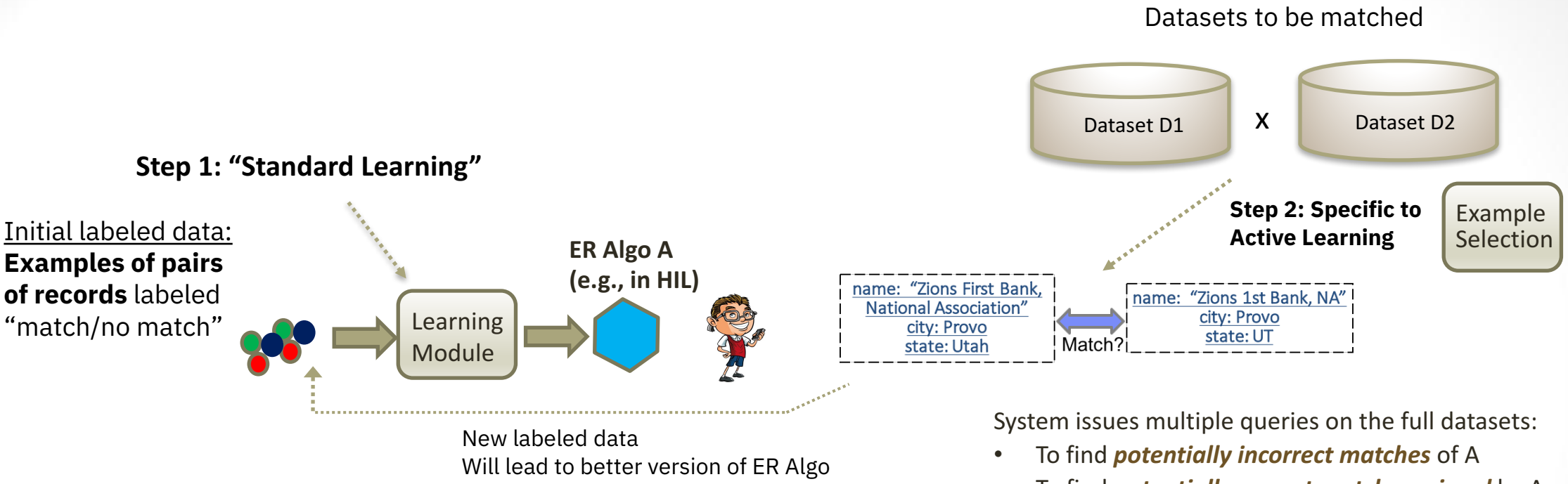
  - **Data-parallel programming & maintenance requires serious expertise and time**
  - **Hard to reason: integration logic is lost among physical operations and multiple systems**
  - **Not focused on the higher-level entities, more worried about performance and scale.**

Need better abstraction and tools

- ***More ambitious: Scale this technology to a larger class of users (non-programmers)***
  - **System & Domain Expert "co-learn" the right algorithms**
    - use data (not code) to interact with the expert,
    - use active learning techniques to identify the hard cases where the human input is required



Example: Active Learning of ER Algorithms



- Expert ready to give feedback on current algorithm A.
- The active learning system must find first a small set of "interesting" examples to be labeled.

System issues multiple queries on the full datasets:

- To find ***potentially incorrect matches*** of A
- To find ***potentially correct matches missed*** by A

Some of the challenges:

- Find a small set of such examples at scale (e.g.,100M x 100M)
- Achieve as good quality as hand-written algorithms

[Arasu et al, 2010], [Qian et al, 2017]

Further Remarks and Pointers

- Closely related area: ***data cleaning, inconsistent databases***.
  - Entity resolution (deduplication) is an important form of data cleaning.
  - But there is more. See **[Arenas et al, 1999], [Galhardas et al, 2001], [Ilyas, Chu, 2015], [Ganti, Sarma, 2013]**

- Much work to be done in ***learning*** of the artifacts (rules, algorithms) for ER, data transformation, data normalization and attribute/entity fusion.
  - Learning of ***reusable*** and ***readable*** artifacts is important
  - Having the ***human in the loop*** is important. See also **[Doan et al, 2017]**
  - ***Data quality***, ***explainability*** and ***provenance*** **[Cheney et al, 2009]** are important.

- Also, more broadly: ***knowledge-base construction from unstructured data***.
  - Fertile ground that combines entity-centric algorithms with information extraction, deep learning techniques for link prediction or more general inference.
  - See also **[Nickel et al, 2016]**



Relevant Publications

- D. Burdick, R. Fagin, P. G. Kolaitis, L. Popa, W.-C. Tan. ***Expressive Power of Entity-Linking Frameworks.*** ICDT 2017, 25-43.
- K. Qian, L. Popa, P. Sen. ***Active Learning for Large-Scale Entity Resolution.*** CIKM 2017.
- A. Doan, A. Ardalan, J. R. Ballard, S. Das, Y. Govind, P. Konda, H. Li, S. Mudgal, E. Paulson, P. Suganthan, H. Zhang. ***Human-in-the-Loop Challenges for Entity Matching: A Midterm Report***. HILDA@SIGMOD 2017: 12:1-12:6
- M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich. ***A Review of Relational Machine Learning for Knowledge Graphs***. Proceedings of the IEEE 104(1): 11-33 (2016)
- D. Burdick, R. Fagin, P. G. Kolaitis, L. Popa, W.-C. Tan. ***A Declarative Framework for Linking Entities.*** ICDT 2015, 25-43. (Extended version appeared in ACM TODS, 2016).
- I. F. Ilyas, X. Chu. ***Trends in Cleaning Relational Data: Consistency and Deduplication***. Foundations and Trends in Databases 5(4): 281-393 (2015)
- M. A. Hernández, G. Koutrika, R. Krishnamurthy, L. Popa, R. Wisnesky. ***HIL: A High-Level Scripting Language for Entity Integration***. EDBT 2013: 549-560
- V. Ganti, A. Das Sarma. ***Data Cleaning: A Practical Perspective.*** Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2013, pp. 1-85.
- D. Burdick, S. Das, M. A. Hernández, C.T. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, S. Vaithyanathan. ***Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study.*** IEEE Data Eng. Bull, 2011.

- A. Arasu, M. Götz, R. Kaushik. ***On Active Learning of Record Matching Packages.*** SIGMOD 2010: 783-794.
- A. Arasu, C. Ré, D. Suciu. ***Large-Scale Deduplication with Constraints Using Dedupalog.*** ICDE 2009: 952-963.
- R. Fagin, L. M. Haas, M. A. Hernández, R. J. Miller, L. Popa, Y. Velegrakis. ***Clio: Schema Mapping Creation and Data Exchange.*** Conceptual Modeling: Foundations and Applications 2009: 198-236
- J. Cheney, L. Chiticariu, W.-C. Tan. ***Provenance in Databases: Why, How, and Where.*** Foundations and Trends in Databases 1(4): 379-474 (2009)
- A. K. Elmagarmid, P. G. Ipeirotis, V. S. Verykios. ***Duplicate Record Detection: A Survey***. IEEE Trans. Knowl. Data Eng. 19(1): 1-16 (2007)
- P. Singla, P. Domingos. ***Entity Resolution with Markov Logic***. ICDM 2006, 572-582.
- R. Fagin, P. G. Kolaitis, R. J. Miller, L. Popa. ***Data Exchange: Semantics and Query Answering***. ICDT 2003: 207-224. (Extended version appeared in Theor. Comp. Sci., 2005).
- H. Galhardas, D. Florescu, D. E. Shasha, E. Simon, C.-A. Saita. ***Declarative Data Cleaning: Language, Model, and Algorithms***. VLDB 2001: 371-380
- R. J. Miller, L. M. Haas, M. A. Hernández. ***Schema Mapping as Query Discovery***. VLDB 2000: 77-88
- M. Arenas, L. E. Bertossi, J. Chomicki. ***Consistent Query Answers in Inconsistent Databases***. PODS 1999: 68-79
- I. P. Fellegi, A. B. Sunter. ***A Theory of Record Linkage.*** J. Am. Statistical Assoc., 64, 328 (1969), 1183-1210

