UofT:DB Group

BIG DATA CURATION

DIMACS
Big Data Integration
Workshop

Renée J. Miller

Curation

- An art **curator** is responsible for the ... *acquisition and care* of *works of art.*

  A curator may:

  - make decisions regarding what objects to collect,
  - oversee their care and documentation,
  - conduct research based on the collection,
  - provide proper packaging of art for transport,
  - and share that research with the public and scholarly community through exhibitions and publications…

Data Curation

- *Acquisition* and *care* of *data*
  - make decisions regarding what data to collect,
  - oversee data care and documentation (metadata)
  - conduct research based on the collection
    - data-driven decision making
  - ensure proper packaging of data for reuse
  - and share that data with the public

- *Ensure data maintains its value over time*

Database Curation, Buneman, 2003

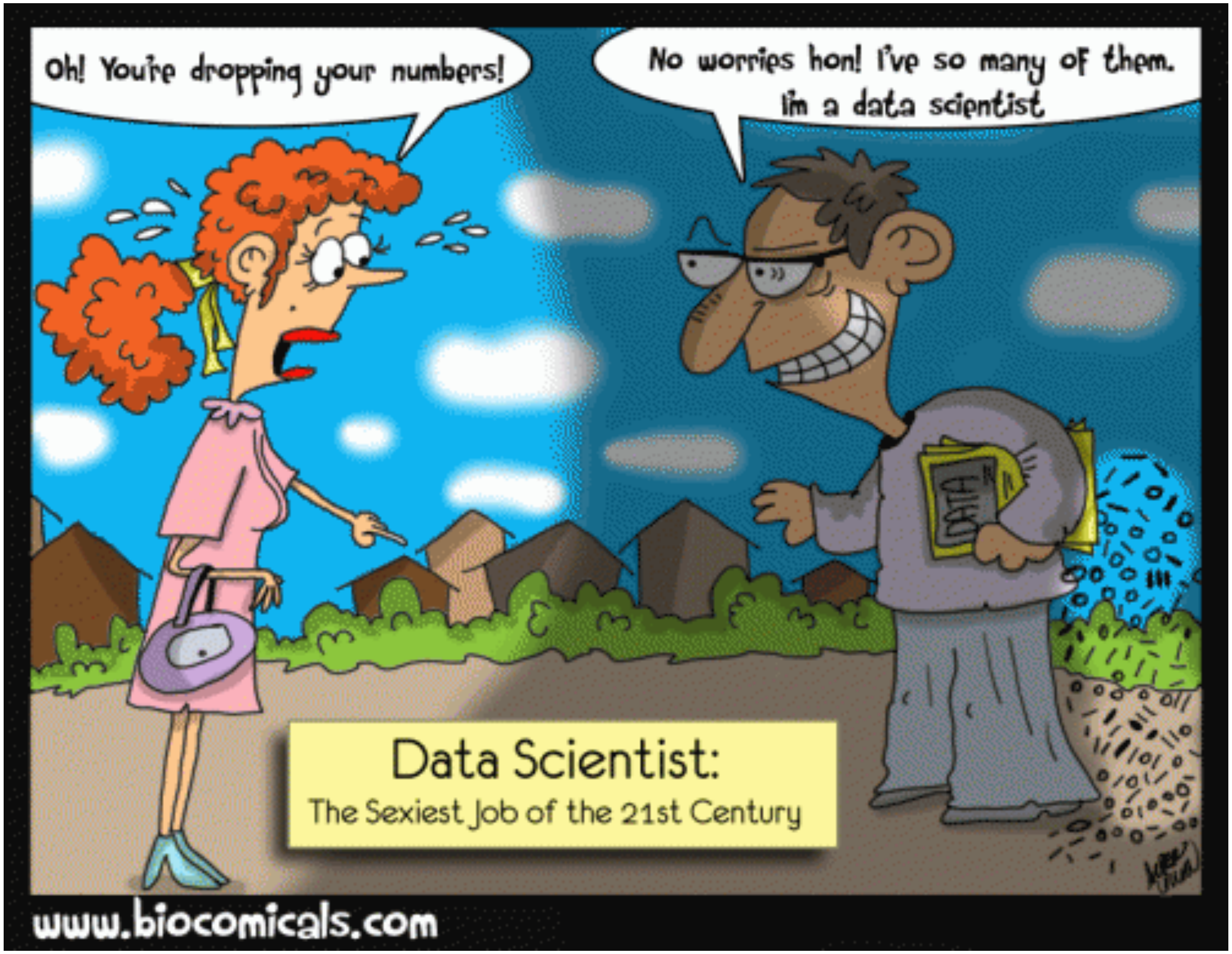
Big Data: 2012 Word of the Year

- Big data is no more exact a notion than big hair
- Data isn't a plural noun like pebbles, it's a mass noun like dust
- When you've got algorithms weighing hundreds of factors over a huge data set, you can't really know ***why they come to a particular decision*** or whether it ***really makes sense***
- A cult of infallibility - a vision of prediction ***obviating explanation***

Why 'Big Data' Should be the Word of the Year
Geoff Nunberg, Fresh Air, Copyright © 2012 NPR
http://www.npr.org/2012/12/20/167702665/geoff-nunbergs-word-of-the-year-big-data

- Society has become ***data-driven***
  - Data-driven education, evidence-based medicine, BI...
- Are we making wrong decisions with our data?

Big Data to Data Science


Oh! You're dropping your numbers!
No worries hon! I've so many of them. Im a data scientist
DATA
Data Scientist:
The Sexiest Job of the 21st Century
www.biocomicals.com

Data Scientist: The Sexiest Job of the 21st Century

Harvard Business Review Oct. 2012

© 2012 Biocomicals
Biocomicals by Dr. Alper Uzun

Talk Themes

- Curation is ultimately about ***semantics***
  - Exploit modeled semantics & be principled in how missing semantics is created
- Curation is for ***humans***
  - Facilitate human understanding and decision making
  - People must be able to correct and understand curation decisions
- Curation focus on ***small(ish) valuable datasets***
  - Leverage ***Big Data*** to add value to curated data
  - Automation required not just for scale, but to manage deep complexity of curation tasks

Credits

- LinkedCT.org
  - ***Oktie Hassanzadeh***, now IBM Watson
    - 2012 Toronto PhD “Record Linkage over Web Data”
- xCurator
  - ***Hassanzadeh, Soheil Hassas Yeganeh***
- LinQuer
  - ***Hassanzadeh, Kementsietsidis, Lim, Wang (IBM)***
- Provenance
  - ***Boris Glavic (IIT Chicago), Alonso (ETH Zurich), Haas (IBM), Saddique (Toronto)***

Talk Focus

- □ Publishing and using Open Data: xCurator
  - ◘ data.gov, data.gov.ca, data.gov.bc.ca
  - ◘ www.toronto.ca/open, ...
- □ Available in variety of semi-structured formats

Curation Examples

□ LinkedCT.org: clinical drug trials

LinkedCT

ClinicalTrials.gov

DBpedia

EMR

NCI Term Browser

□ MIDAS: company data

U.S. Securities and Exchange Commission

Freebase

DBpedia

Selected Problems

- Customization of Linkage Rules
  - Linking living data
- Provenance in Data Curation
  - Vision for data mining provenance
- Linkage point discovery
  - Big data challenge to a traditional data integration problem

Running Example

LinkedCT

- Source: **ClinicalTrials.gov**
  - Online registry of international clinical drug trials
  - 139,000+ XML files, updated daily
  - Provides web search interface
- Permits downloading relational (static) dump of DB
  - Permits structure querying
  - Relatively ***high cost of ownership***
  - Still stand-alone DB ***not integrated*** with other sources or even linked to common Web knowledge
    - "Find trials on Alzheimer's disease near Toronto"
  - Although (mostly manually) curated, data still contains ***errors and inconsistencies***
    - Thalassemia vs. Thalassaemia

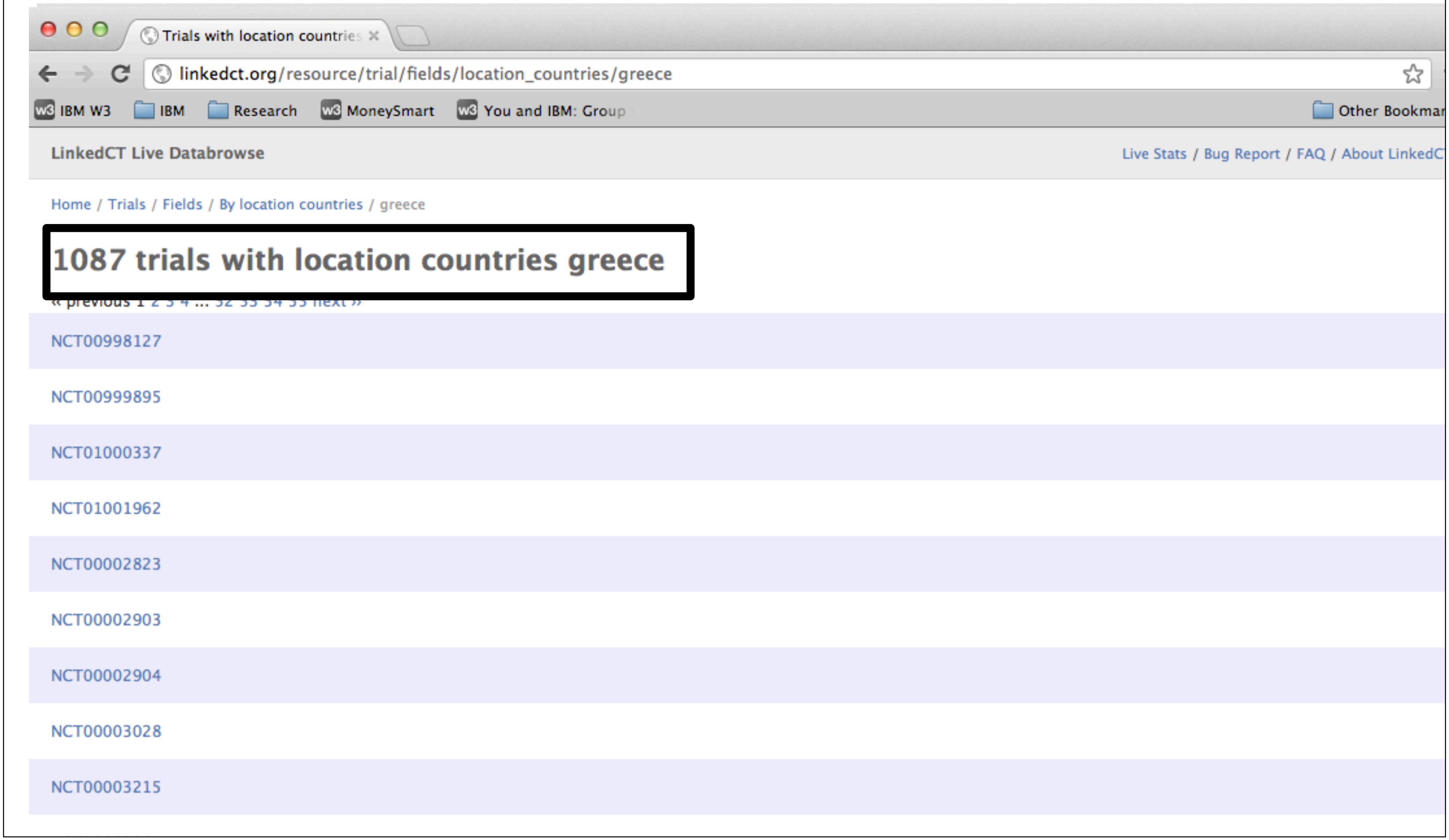
Goals LinkedCT.org

- Apply and study *large-scale data curation*
  - Original data is not massive, but adding value to data requires linking to big data
- Create an engaged, *incentivized user community*
  - Human knowledge (correction) critical
  - For most data, this is not going to come from MTurk...
  - *Raise level of abstraction* in how curation decisions communicated so domain experts can contribute
- Study curation *over time*
  - ClinicalTrials.gov publishes updates daily
  - Tracability (*provenance*) of curation decisions

Publish as Linked Open Data

- Choice to use RDF as target model largely orthogonal from other curation tasks
- Permitted the creation of a user community of domain experts
  - Critical part of any curation activity
- Life Sciences big driver Linked Open Data
  - Life Sciences large portion of open linked data sources
- Special opportunities afforded by Linked Open Data
  - Our focus is on more general curation issues

Why Curate?


Trials with location countries
IBM W3
IBM
Research
MoneySmart
You and IBM: Group
Other Bookmar
Live Stats / Bug Report / FAQ / About LinkedC

linkedct.org/resource/trial/fields/location_countries/greece

LinkedCT Live Databrowse

Home / Trials / Fields / By location countries / greece

1087 trials with location countries greece

« previous 1 2 3 4 ... 32 33 34 35 next »

NCT00998127

NCT00999895

NCT01000337

NCT01001962

NCT00002823

NCT00002903

NCT00002904

NCT00003028

NCT00003215

Example LinkedCT Application

- □ Evaluating Research-Disease Disparity [Zaveri11]

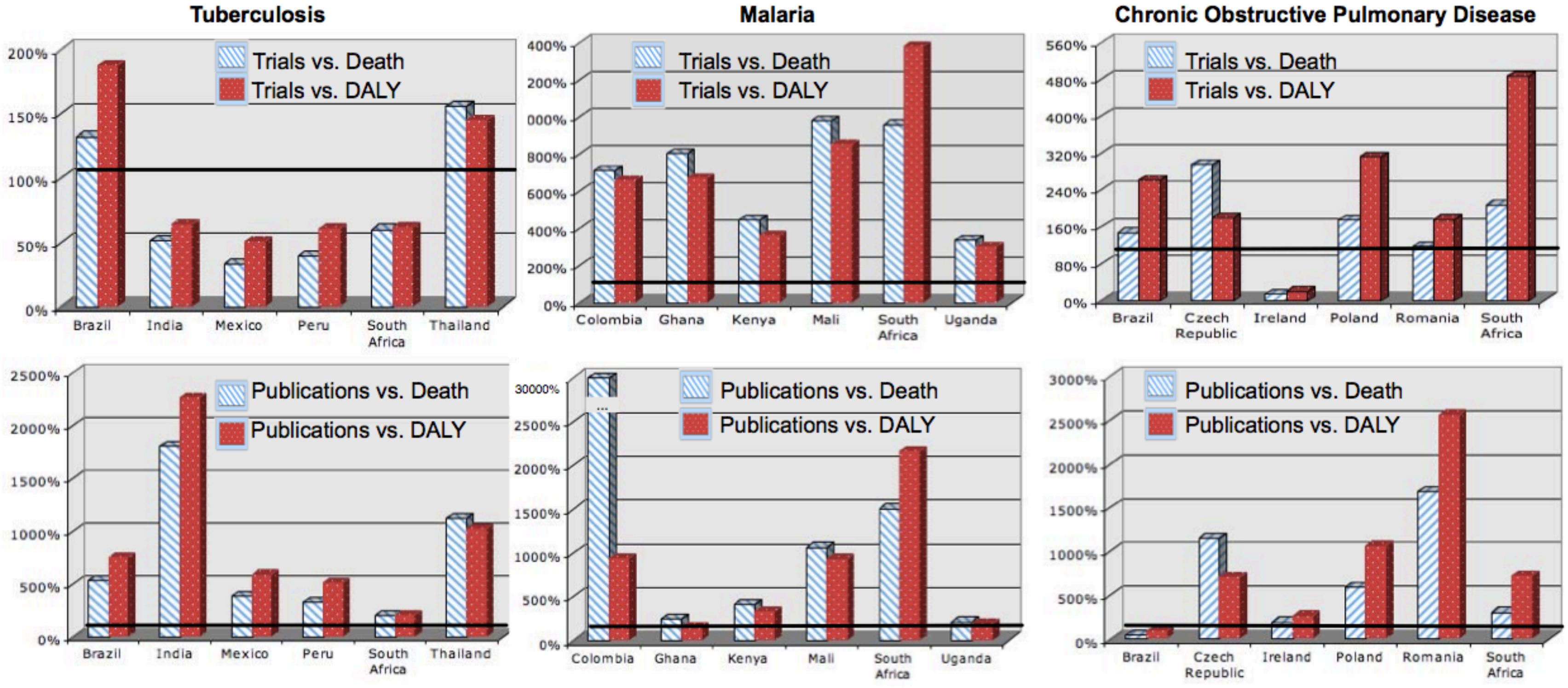
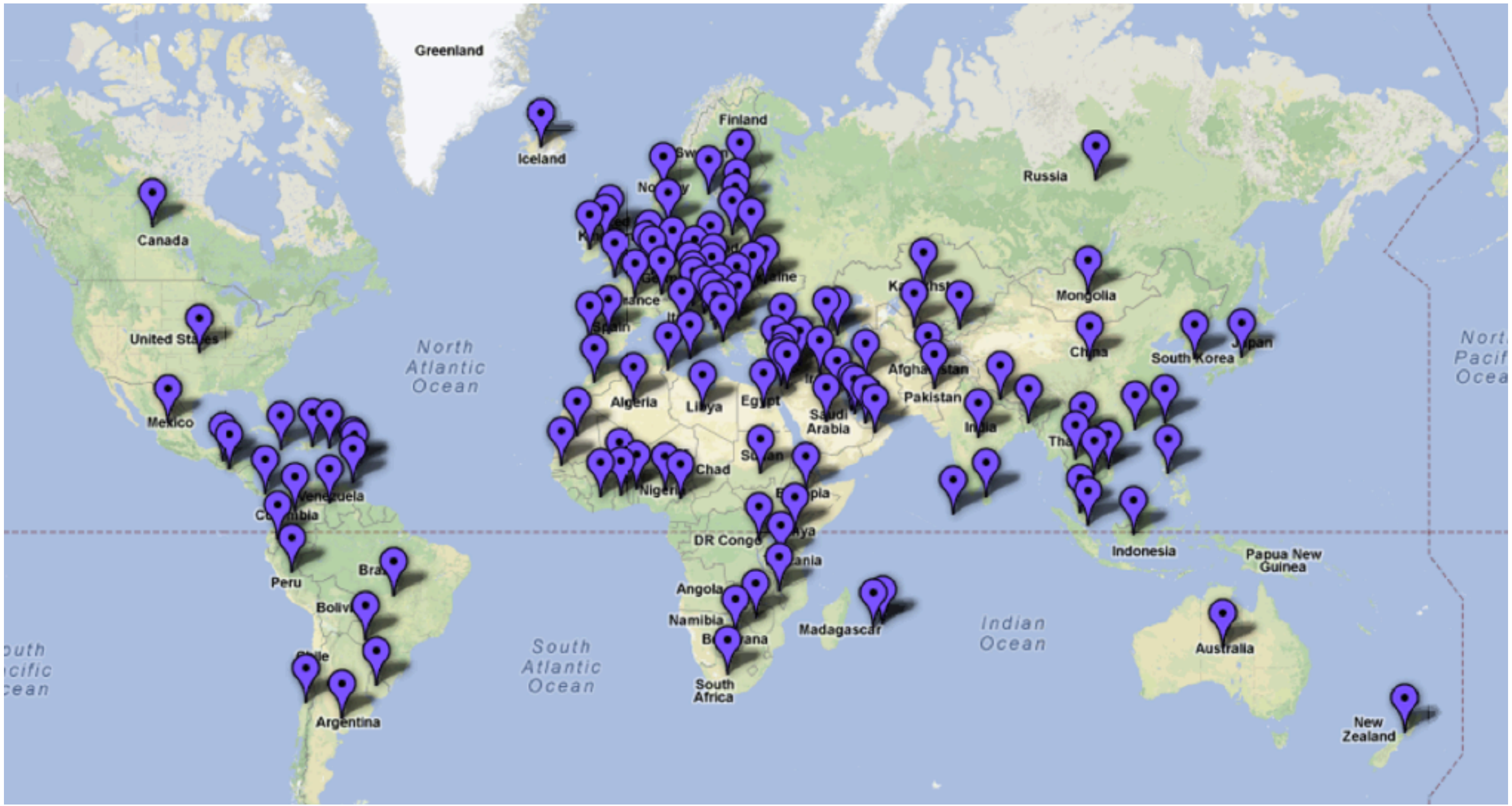


Figure 5. Depiction of the four ReDD indices for the diseases Tuberculosis, Malaria and COPD and selected countries. The black line indicates the level of a balanced distribution of research resources and is low in some figures due to the zero research investments not visualized. Source: [Zaveri11]

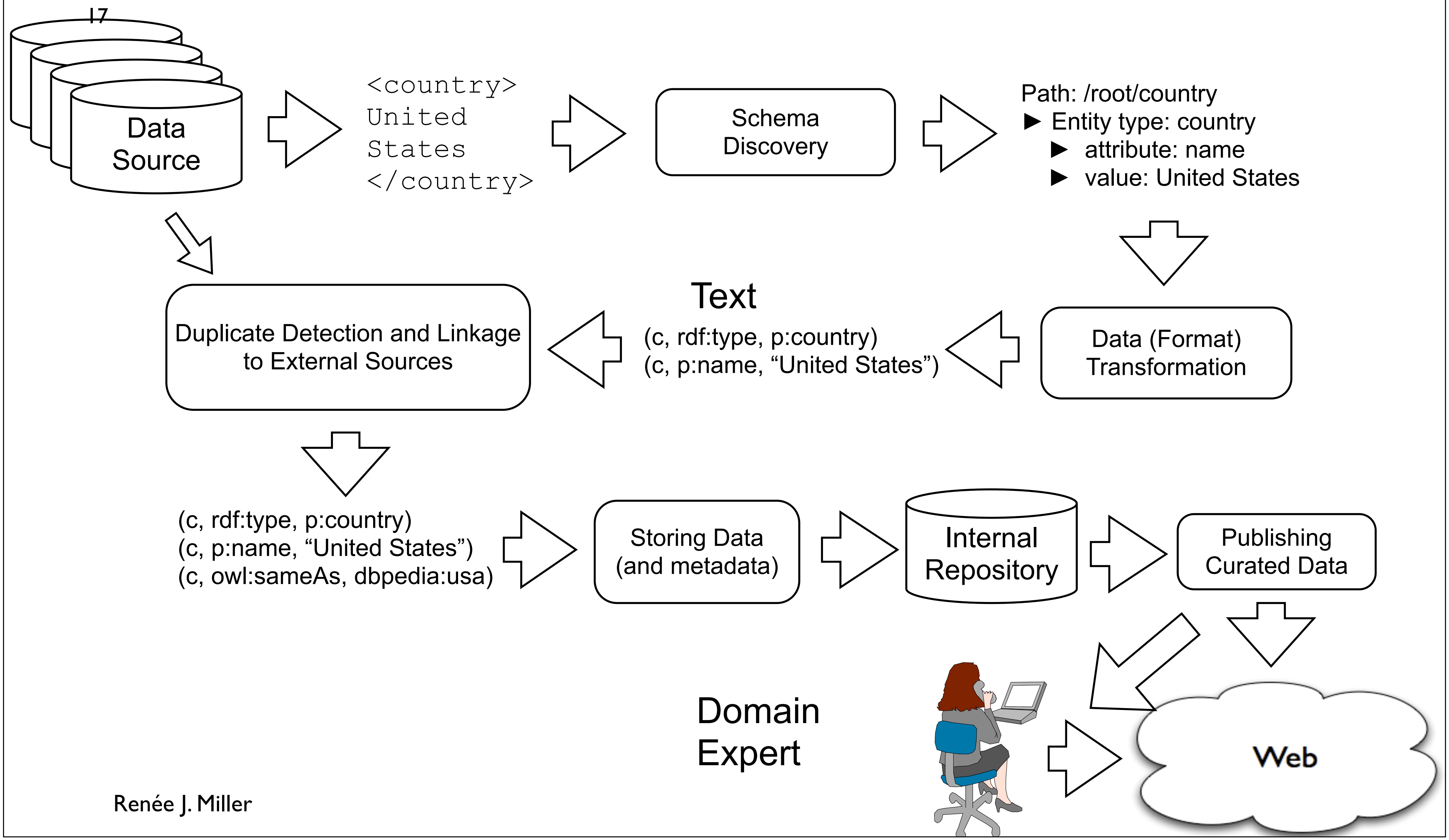
LinkedCT.org

The following map shows the users of the ***Linked Clinical Trials (LinkedCT) project*** *based on data provided by Google Analytics from the project's website www.linkedct.org. The project website currently has users (visitors) from* ***131 countries****, with over 10,500 visitors per month (over 18,600 page views)*



UofT:DB Group

The Process



17

Data Source

```
<country>
United
States
</country>
```

Schema Discovery

Path: /root/country
- ▶ Entity type: country
  - ▶ attribute: name
  - ▶ value: United States

Data (Format) Transformation

Text

(c, rdf:type, p:country)
(c, p:name, “United States”)

Duplicate Detection and Linkage to External Sources

(c, rdf:type, p:country)
(c, p:name, “United States”)
(c, owl:sameAs, dbpedia:usa)

Storing Data (and metadata)

Internal Repository

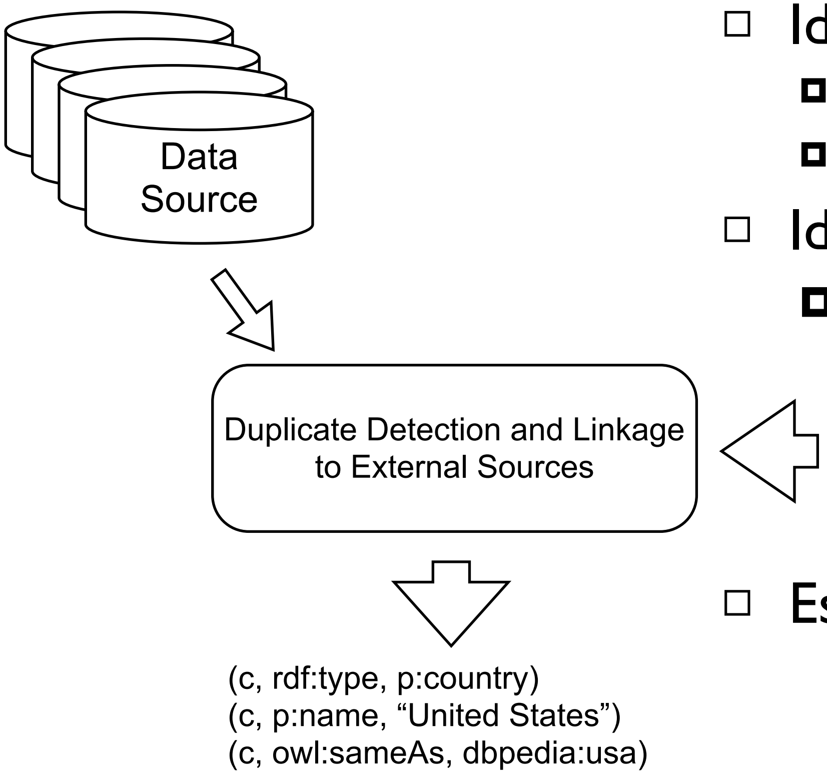
Publishing Curated Data

Web

Domain Expert

Duplicate Detection & Instance Linkage

- □ Identify and properly link duplicate entities
  - ◘ Entities that refer to the same real-world entity
  - ◘ "same-as" links
- □ Identify and properly link duplicate entity types
  - ◘ "equivalent-class" links



(c, rdf:type, p:country)
(c, p:name, "United States")

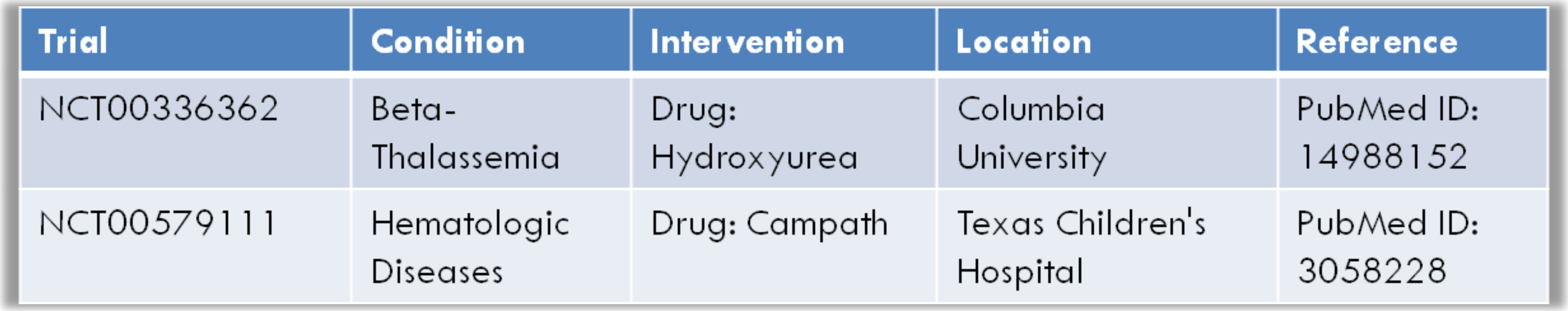
(c, rdf:type, p:country)
(c, p:name, "United States")
(c, owl:sameAs, dbpedia:usa)

- □ Establish links to external sources

- □ Many tools and techniques exist for duplicate detection and linkage [Christen12]
  - □ Scalability is a key issue [Elmagarmid07]
  - □ ***Getting linkage rules right even more of a challenge***

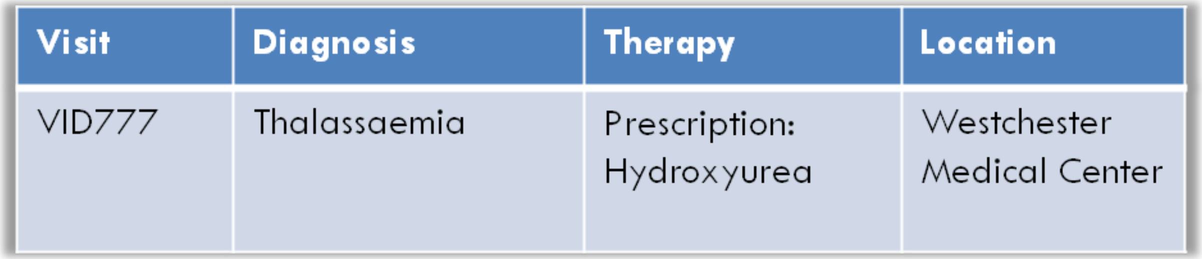
Data Linking

Clinical Trials (CT) from ClinicalTrials.gov/LinkedCT.org



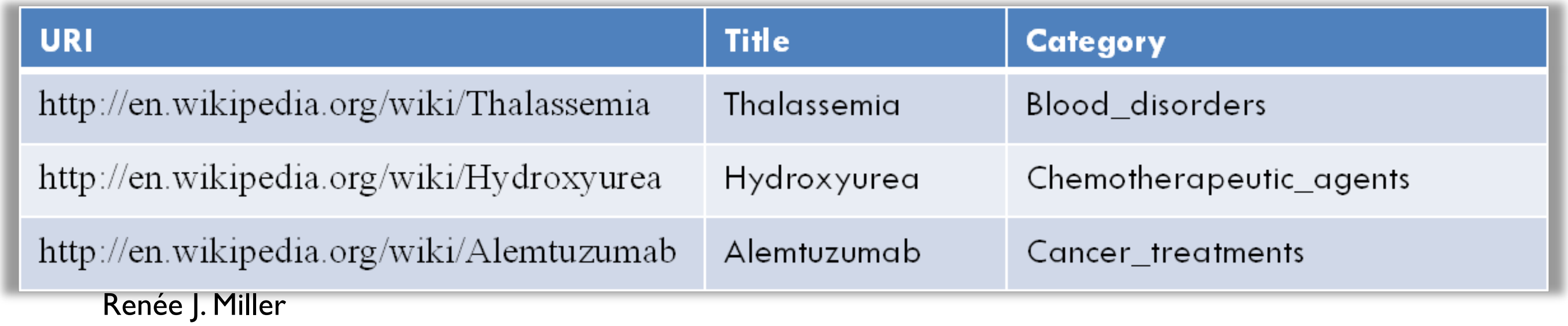
Trial	Condition	Intervention	Location	Reference
NCT00336362	Beta-Thalassemia	Drug: Hydroxyurea	Columbia University	PubMed ID: 14988152
NCT00579111	Hematologic Diseases	Drug: Campath	Texas Children's Hospital	PubMed ID: 3058228

Patient Visits (PV)



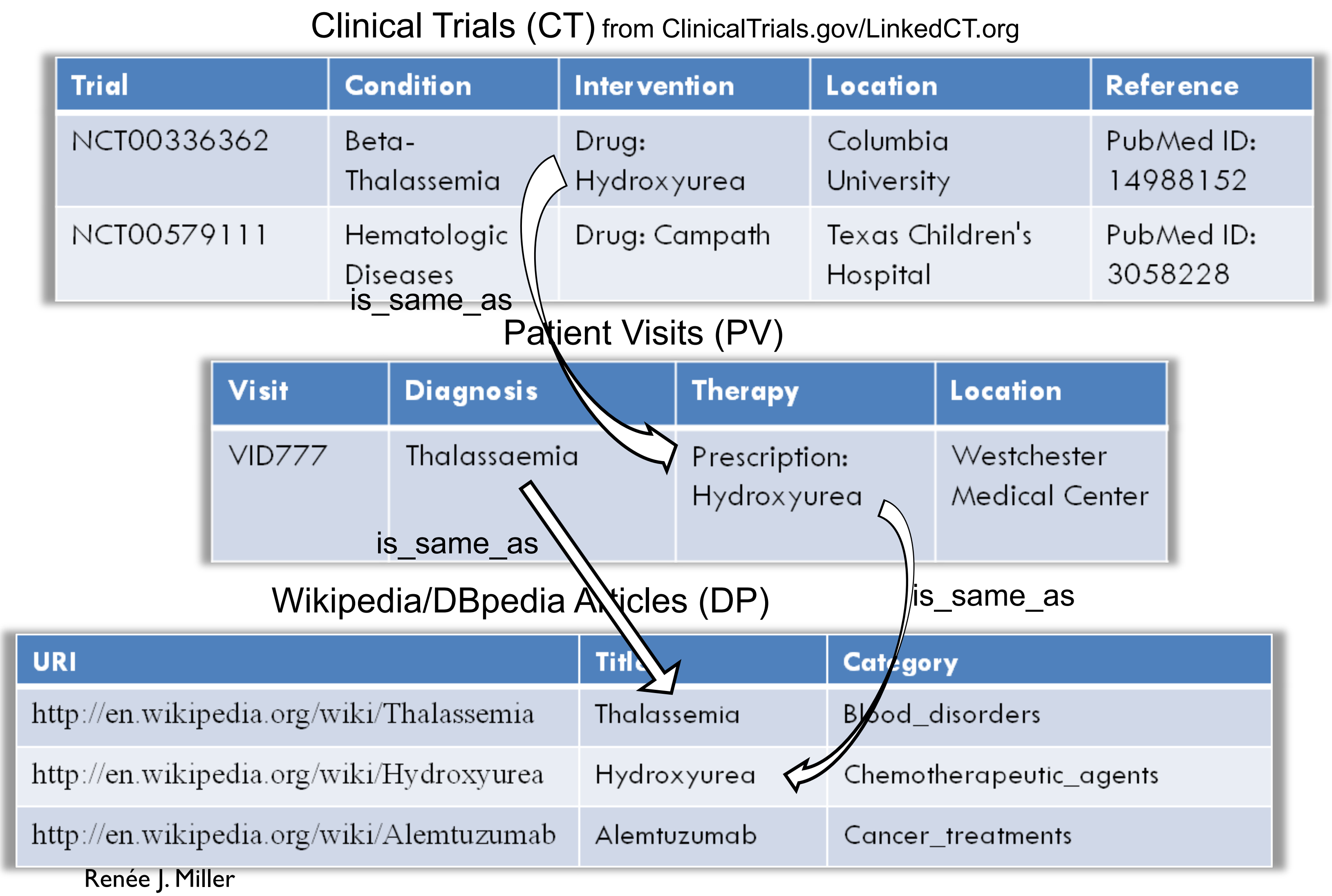
Visit	Diagnosis	Therapy	Location
VID777	Thalassaemia	Prescription: Hydroxyurea	Westchester Medical Center

Wikipedia/DBpedia Articles (DP)



URI	Title	Category
http://en.wikipedia.org/wiki/Thalassemia	Thalassemia	Blood_disorders
http://en.wikipedia.org/wiki/Hydroxyurea	Hydroxyurea	Chemotherapeutic_agents
http://en.wikipedia.org/wiki/Alemtuzumab	Alemtuzumab	Cancer_treatments

Data Linking



Clinical Trials (CT) from ClinicalTrials.gov/LinkedCT.org

Trial	Condition	Intervention	Location	Reference
NCT00336362	Beta-Thalassemia	Drug: Hydroxyurea	Columbia University	PubMed ID: 14988152
NCT00579111	Hematologic Diseases	Drug: Campath	Texas Children's Hospital	PubMed ID: 3058228

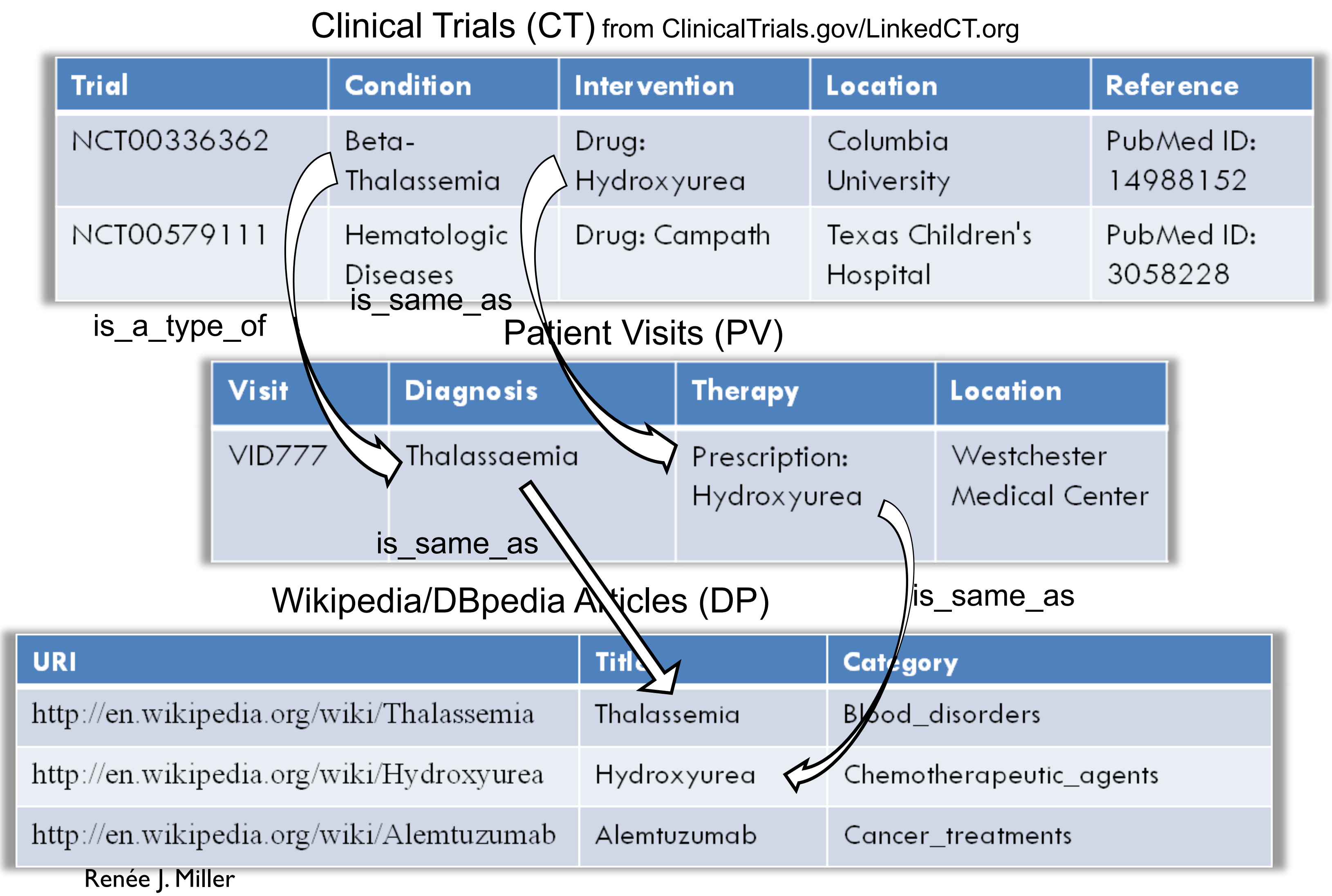
Patient Visits (PV)

Visit	Diagnosis	Therapy	Location
VID777	Thalassaemia	Prescription: Hydroxyurea	Westchester Medical Center

Wikipedia/DBpedia Articles (DP)

URI	Title	Category
http://en.wikipedia.org/wiki/Thalassemia	Thalassemia	Blood_disorders
http://en.wikipedia.org/wiki/Hydroxyurea	Hydroxyurea	Chemotherapeutic_agents
http://en.wikipedia.org/wiki/Alemtuzumab	Alemtuzumab	Cancer_treatments

Data Linking



Clinical Trials (CT) from ClinicalTrials.gov/LinkedCT.org

Trial	Condition	Intervention	Location	Reference
NCT00336362	Beta-Thalassemia	Drug: Hydroxyurea	Columbia University	PubMed ID: 14988152
NCT00579111	Hematologic Diseases	Drug: Campath	Texas Children's Hospital	PubMed ID: 3058228

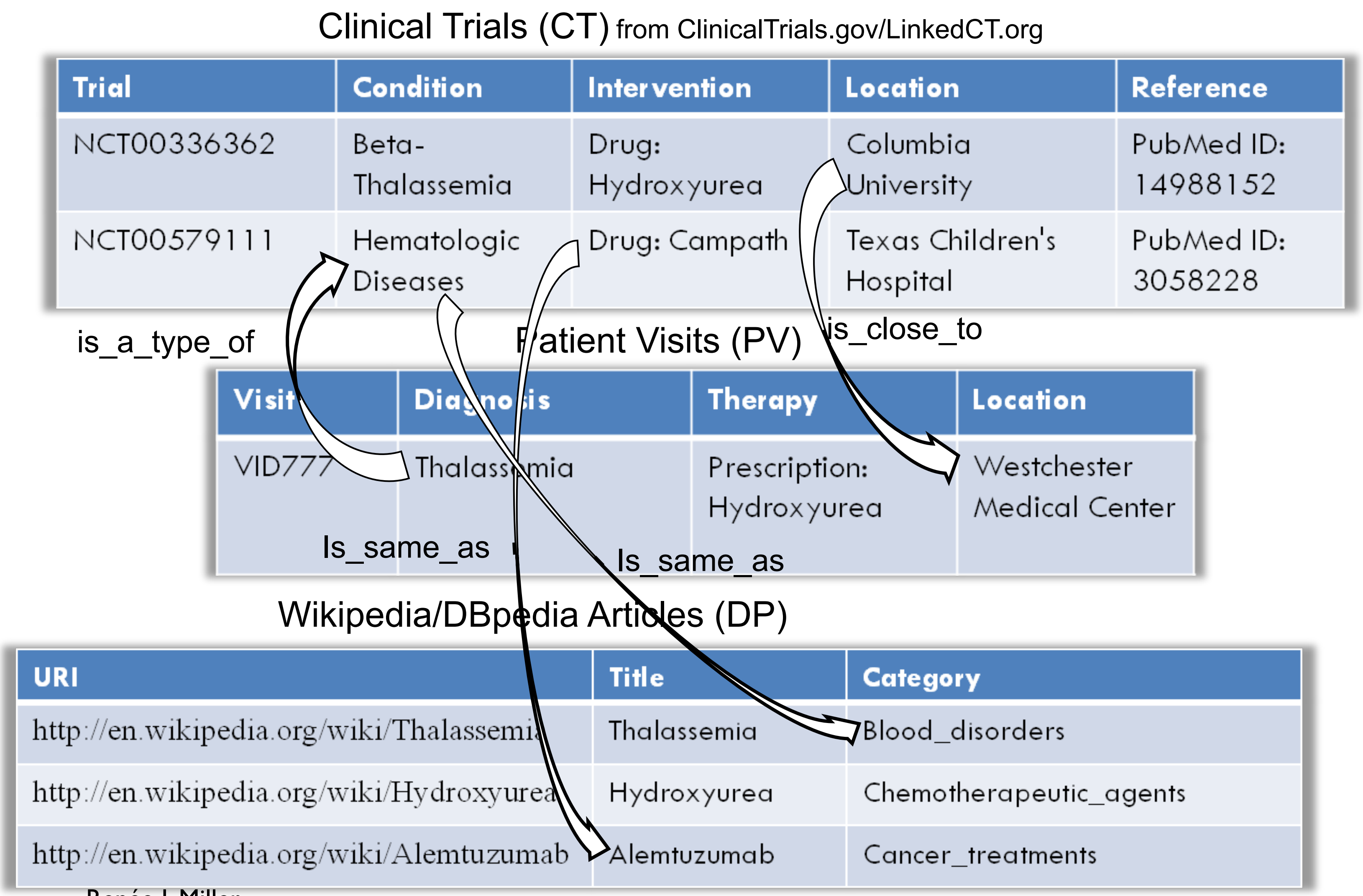
Patient Visits (PV)

Visit	Diagnosis	Therapy	Location
VID777	Thalassaemia	Prescription: Hydroxyurea	Westchester Medical Center

Wikipedia/DBpedia Articles (DP)

URI	Title	Category
http://en.wikipedia.org/wiki/Thalassemia	Thalassemia	Blood_disorders
http://en.wikipedia.org/wiki/Hydroxyurea	Hydroxyurea	Chemotherapeutic_agents
http://en.wikipedia.org/wiki/Alemtuzumab	Alemtuzumab	Cancer_treatments

Data Linking



Clinical Trials (CT) from ClinicalTrials.gov/LinkedCT.org

Trial	Condition	Intervention	Location	Reference
NCT00336362	Beta-Thalassemia	Drug: Hydroxyurea	Columbia University	PubMed ID: 14988152
NCT00579111	Hematologic Diseases	Drug: Campath	Texas Children's Hospital	PubMed ID: 3058228

is_a_type_of

is_close_to

Patient Visits (PV)

Visit	Diagnosis	Therapy	Location
VID777	Thalassemia	Prescription: Hydroxyurea	Westchester Medical Center

Is_same_as

Is_same_as

Wikipedia/DBpedia Articles (DP)

URI	Title	Category
http://en.wikipedia.org/wiki/Thalassemia	Thalassemia	Blood_disorders
http://en.wikipedia.org/wiki/Hydroxyurea	Hydroxyurea	Chemotherapeutic_agents
http://en.wikipedia.org/wiki/Alemtuzumab	Alemtuzumab	Cancer_treatments

Data Linking: Problem Definition

- Find an ***effective linkage method*** accurately models pairs of values match (in a given semantic relationship)
- Great research on finding duplicates
  - SERF (Stanford), Dedupalog (UW), PSL (UMD), ...
- We wanted something easy to
  - Customize (e.g., with domain knowledge)
  - Understand
    - Automated learning does not obviate understanding!

Our Solution: LinQuer

- Generic, extensible and easy-to-use framework
  - LinQL: SQL extension specification of linkage methods
  - [Hassanzadeh, Kementsietsidis, Lim, M-, Wang 09]
- Library of large variety similarity functions
  - Syntactic (string) errors or differences
  - Semantic relationship or equivalence
  - A mix of syntactic similarity and semantic relationship
- Efficient implementation and integration with SQL
  - Declarative approximate string joins [Hassanzadeh 07]
- ***Open incremental incorporation human knowledge***

Customization of linkage methods

- Linkage sources, semantic tables, weight tables all stored as native SQL
- Easy incorporation of ***prior knowledge***
  - E.g., “Disease”, “Disorder”, “Cancer” and “Syndrome” are relatively unimportant for matching
    - “Hematologic Disorders” = “Hematologic Diseases”
- Easy to incorporate exceptions or human provided links

```
SELECT P.*, CT.*
FROM   emr P, clinicalTrial CT
WHERE  CT.interventiontype=`drug’ AND
       P.diagnosis LINK CT.condition
       USING synonym(NCI,NCIconcept,NCIsynonym)
       AND CT.condition = NCIconcept
       AND weightedJaccard(P.diagnosis,NCIsynonym,.7)
```

Data Provenance

- SQL specification of linkage methods
- Use ***provenance*** to explain results
  - “Why was this link produced?”
  - “Explain the link (Beta-Thalassemia is-type-of Thalassaemia)?”
- Data provenance models (why-provenance) give set of facts (data) used to derive a link, e.g.,
  - NCIThessaurus.TypeOf(Beta-Thalassemia, Thalassemia)
  - Jaccard(Thalassemia, Thalassaemia, .7)

vs.

- expertLink(Beta-Thalassemia, is-type-of, Thalassaemia, JaneDoe)

- [Buneman, Khanna, Tan 01], [Green, Karvounarakis, Tannen 07]

Beyond Data Provenance

- Given link (or any data derived from curation):
  - What data is it derived from (data provenance)?
  - Which linkage methods were used to create it?
  - Who created it?
  - Which data sources contributed to it?
- Given an erroneous result, is the error in the base data, the curation (linking) process, or human curation decisions?

Contributions

- Relational representation of provenance
  - Support querying of provenance
  - Which links were derived from the assertion that (Thalessemia is-type-of, BloodDisorder)?
  - Requires not only ability to generate provenance but to represent and query it relationally *[Glavic, Alonso, M-- 13]*
- What part of a linkage method or transformation program (mapping) contributed to a link?
  - If data has been mapped (transformed) into another schema, could the transformation code be wrong?
  - Using provenance to debug data exchange
    - *[Glavic, Alonso, Haas M--10]*

Provenance for Curation

- Determine curation effect as input data changes
    - We ran a frequent itemset mining algorithm to discovery entity types last week. Users have modified the repository since then. What is the effect on the mining result?
- Determine trust based on responsibility
    - I ran a clustering algorithm over manually supplied links between LinkedCT and my EMR DB. Information about the provence of each link (the trustworthiness of each user is available). How trustworthy is each of the resulting clusters.

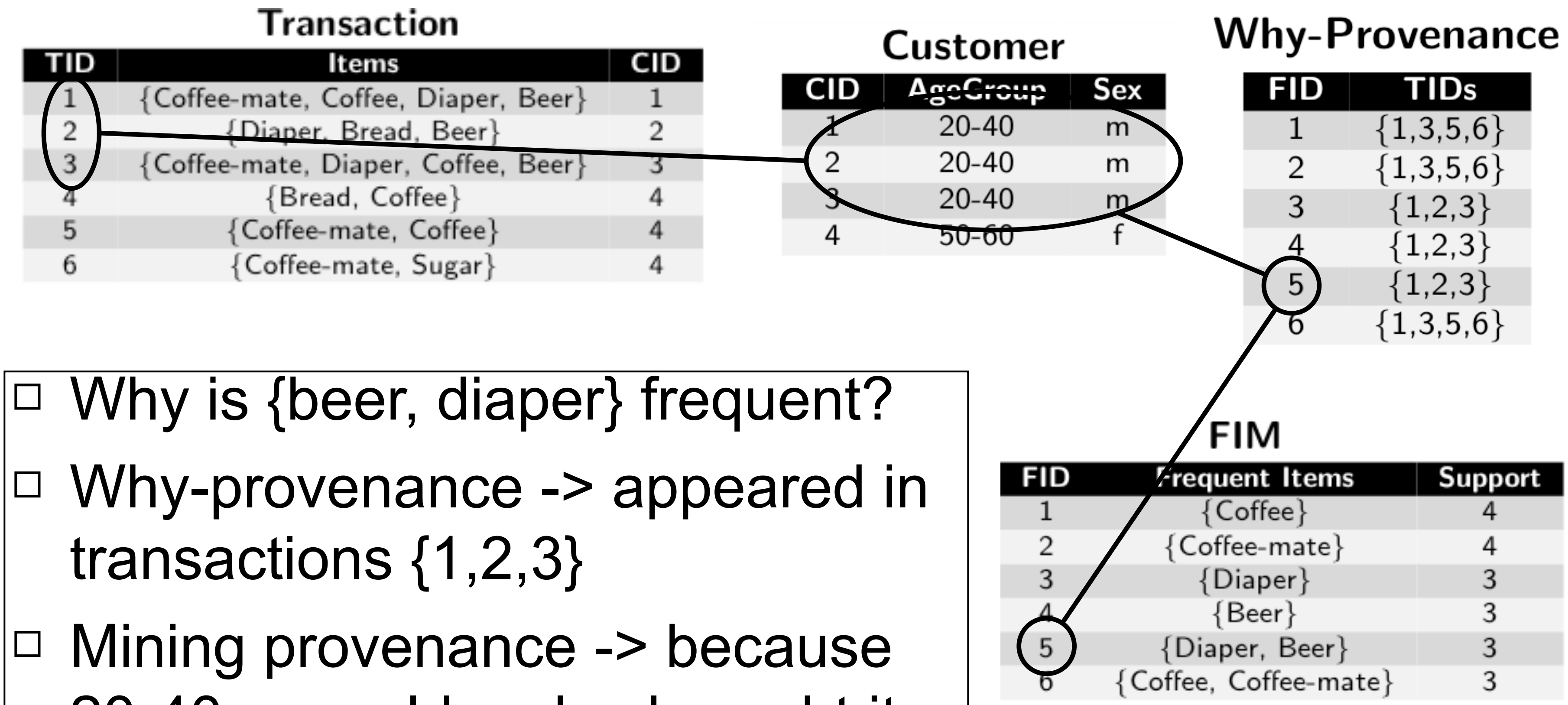
Provenance for Data Mining

- Data mining
  - Extract useful information from data
    - Summarization, simplification, filtering
  - Underlies many curation tasks like schema discovery and link discovery
- Raw data mining outputs are often hard to interpret
  - Drill-down to relevant inputs (**Provenance+**)
    - Find related inputs and summarize this information (**Mine Provenance**)
  - Quantify amount of responsibility (**Responsibility**)

Frequent Itemset Mining

- □ One of the most notorious mining task
- □ **Input:** Set of transactions (each is a subset of items from a domain D)
- □ **Output:** Set of frequent itemsets (subsets of D) which appear in fraction larger than minimum support threshold σ
- □ **Provenance for FIM** *[Glavic, Siddique, M-- 13]*
  - ◘ Why-Provenance
  - ◘ I-Provenance
  - ◘ Mining provenance and related data

Provenance



Transaction

TID	Items	CID
1	{Coffee-mate, Coffee, Diaper, Beer}	1
2	{Diaper, Bread, Beer}	2
3	{Coffee-mate, Diaper, Coffee, Beer}	3
4	{Bread, Coffee}	4
5	{Coffee-mate, Coffee}	4
6	{Coffee-mate, Sugar}	4

Customer

CID	AgeGroup	Sex
1	20-40	m
2	20-40	m
3	20-40	m
4	50-60	f

Why-Provenance

FID	TIDs
1	{1,3,5,6}
2	{1,3,5,6}
3	{1,2,3}
4	{1,2,3}
5	{1,2,3}
6	{1,3,5,6}

FIM

FID	Frequent Items	Support
1	{Coffee}	4
2	{Coffee-mate}	4
3	{Diaper}	3
4	{Beer}	3
5	{Diaper, Beer}	3
6	{Coffee, Coffee-mate}	3

- Why is {beer, diaper} frequent?
- Why-provenance -> appeared in transactions {1,2,3}
- Mining provenance -> because 20-40 year old males brought it

Some Ideas for Clustering

- Provenance for clusters is huge
  - Provenance of a single result contains up to all inputs
  - But, influence is non-uniform!
- Quantify amount of responsibility
  - Follow ideas from [Meliou et al.] and [Halpern et al.]?
  - Responsibility = How much does clustering change when a certain input is removed?
- Parameter settings vs. data responsibility
  - Mining algorithms are often sensitive to param. settings
  - To what extend does a result depend on data vs. parameters?

Selected Problems

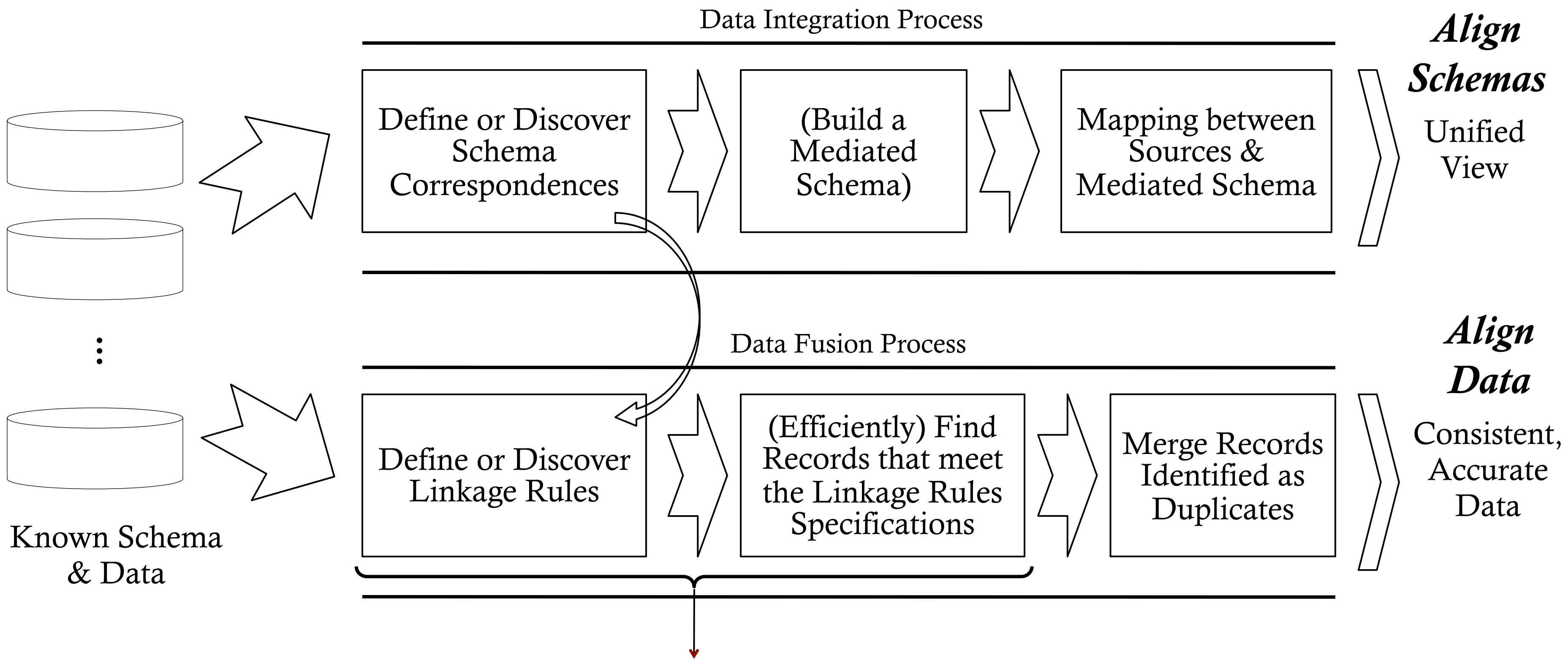
- Customization of Linkage Rules
  - Linking living data

- Provenance in Data Curation
  - Vision for data mining provenance

- Linkage point discovery
  - Big data challenge to a traditional data integration problem
  - [Hassanzadeh et al., VLDB 2013]

Traditional Integration & Fusion

Traditional Data Integration or Fusion Process



Data Integration Process

Define or Discover Schema Correspondences

(Build a Mediated Schema)

Mapping between Sources & Mediated Schema

Align Schemas

Unified View

Data Fusion Process

Define or Discover Linkage Rules

(Efficiently) Find Records that meet the Linkage Rules Specifications

Merge Records Identified as Duplicates

Align Data

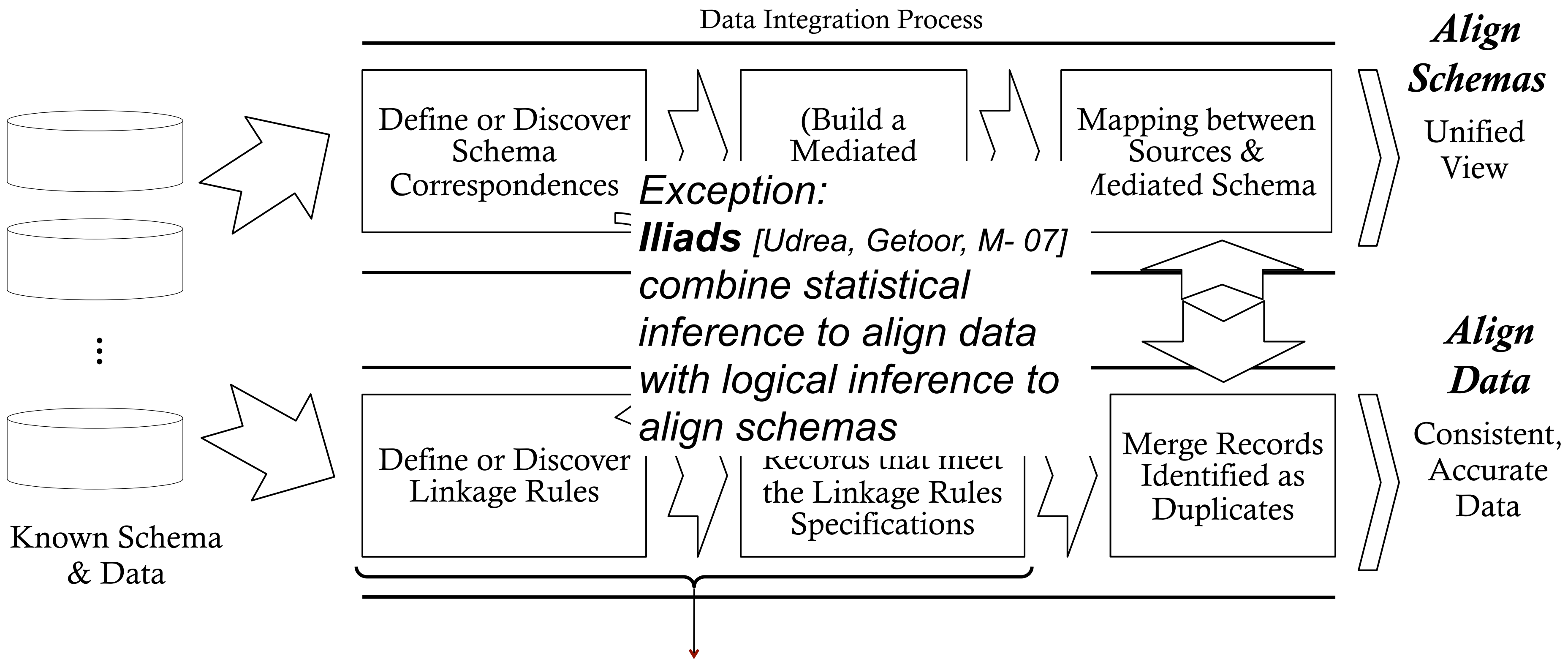
Consistent, Accurate Data

Known Schema & Data

Goal: linking records that refer to the same real-world entity (problem referred to as **Record Linkage**, or **Entity Resolution**)

Traditional Integration & Fusion

Traditional Data Integration or Fusion Process


Data Integration Process
Define or Discover Schema Correspondences
(Build a Mediated
Mapping between Sources & Mediated Schema
Align Schemas
Unified View
Define or Discover Linkage Rules
Records that meet the Linkage Rules Specifications
Merge Records Identified as Duplicates
Align Data
Consistent, Accurate Data
Known Schema & Data

Exception:

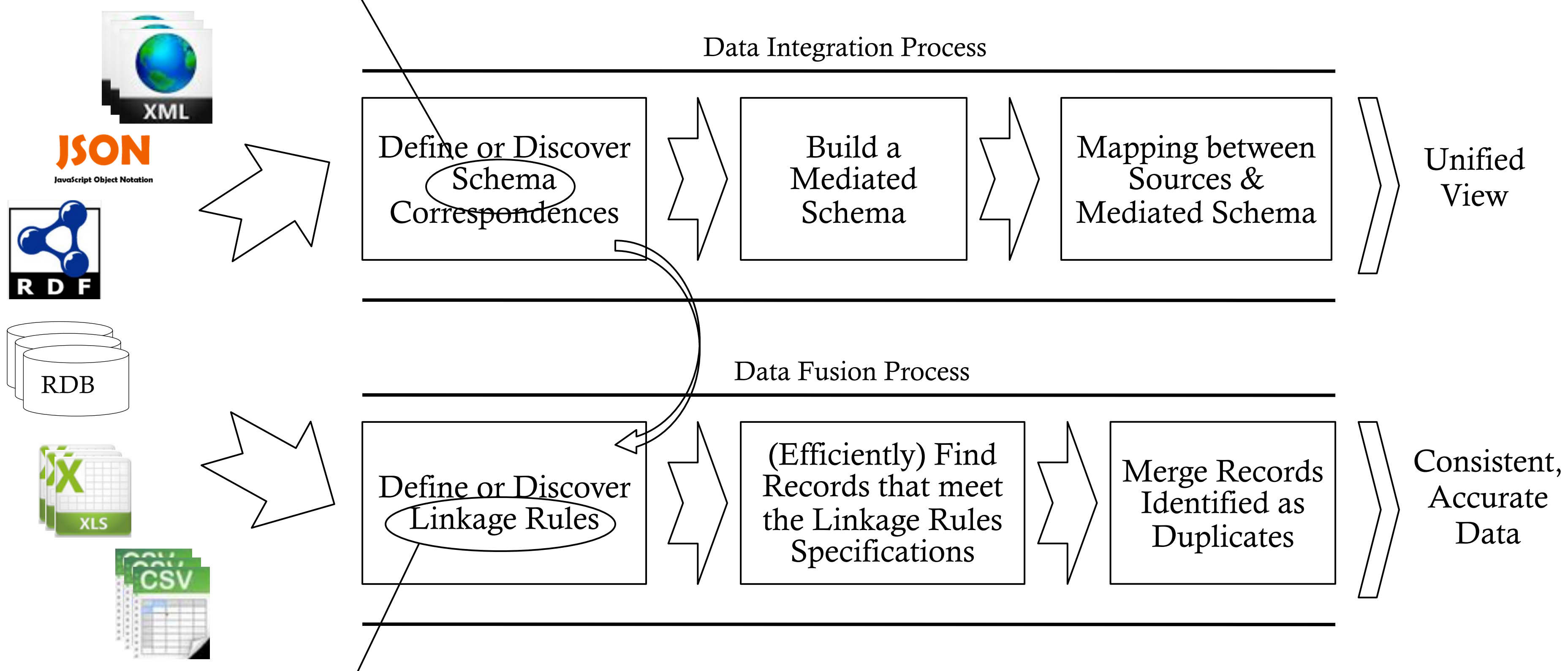
***Iliads** [Udrea, Getoor, M- 07]*

combine statistical inference to align data with logical inference to align schemas

Goal: linking records that refer to the same real-world entity (problem referred to as **Record Linkage**, or **Entity Resolution**)

Big Data Challenge

Schema can be: **unknown**, **very large**, and **noisy**



Data Integration Process

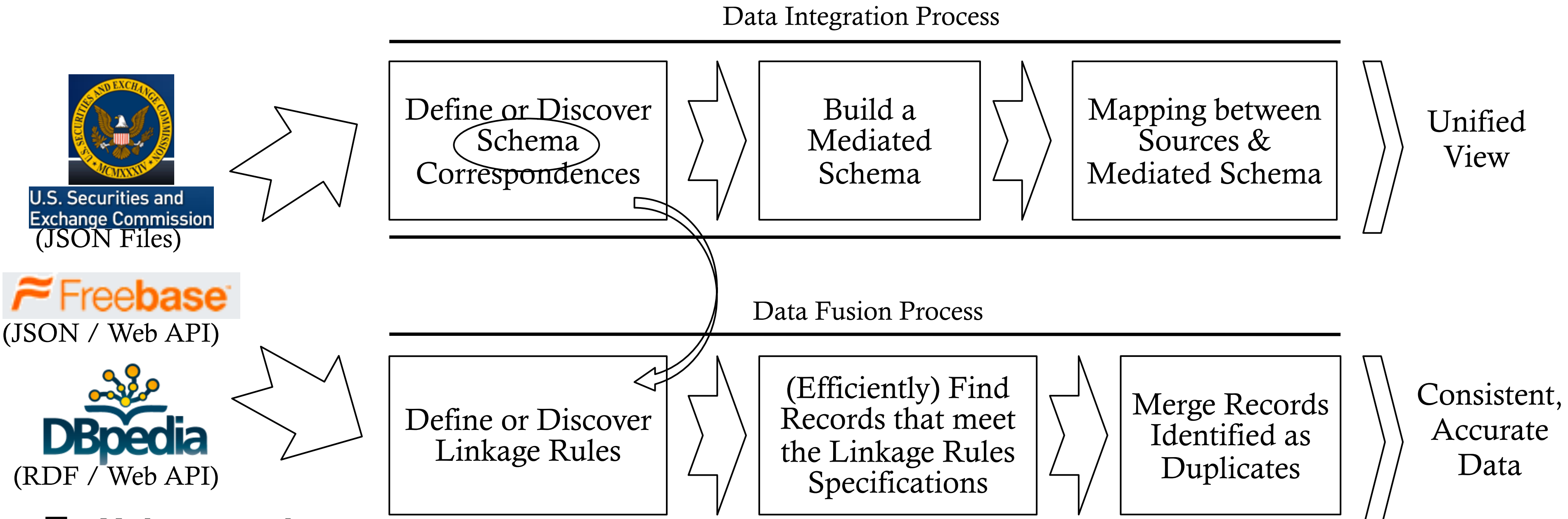
Define or Discover Schema Correspondences → Build a Mediated Schema → Mapping between Sources & Mediated Schema → Unified View

Data Fusion Process

Define or Discover Linkage Rules → (Efficiently) Find Records that meet the Linkage Rules Specifications → Merge Records Identified as Duplicates → Consistent, Accurate Data

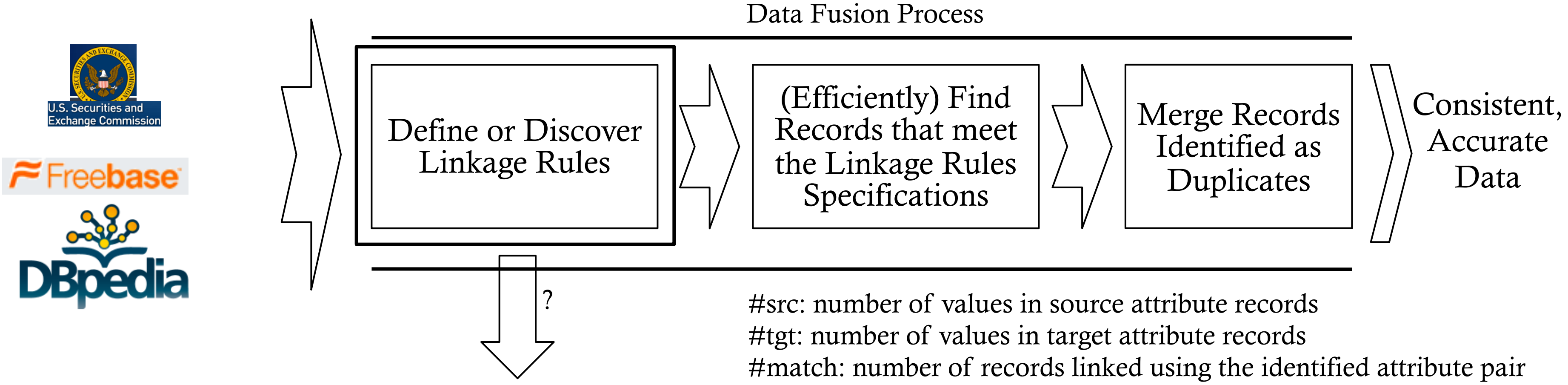
- Linkage rules are no longer simple relationships between known schema elements
- Manual definition of rules is no longer feasible
- Semi-automatic discovery of rules can be **very challenging**

Dynamic, Noisy Schemas


Data Integration Process
U.S. Securities and Exchange Commission
(JSON Files)
Freebase
(JSON / Web API)
DBpedia
(RDF / Web API)
Define or Discover Schema Correspondences
Build a Mediated Schema
Mapping between Sources & Mediated Schema
Unified View
Data Fusion Process
Define or Discover Linkage Rules
(Efficiently) Find Records that meet the Linkage Rules Specifications
Merge Records Identified as Duplicates
Consistent, Accurate Data

- **Unknown schema**
  - SEC data has no given schema, Freebase & DBpedia have dynamic and highly heterogeneous schemas
- **Very large schema**
  - Considering only "company" entities in Freebase & DBpedia
    - DBpedia has over 3,000 attributes, Freebase has 167 attributes, SEC has 72 attributes
  - Freebase and DBpedia have thousands of types
- **Noisy schema**
  - All sources are automatically extracted from text and contain noise
    e.g., dateOfDeath & stockTicket [sic] attributes for a Company in DBpedia

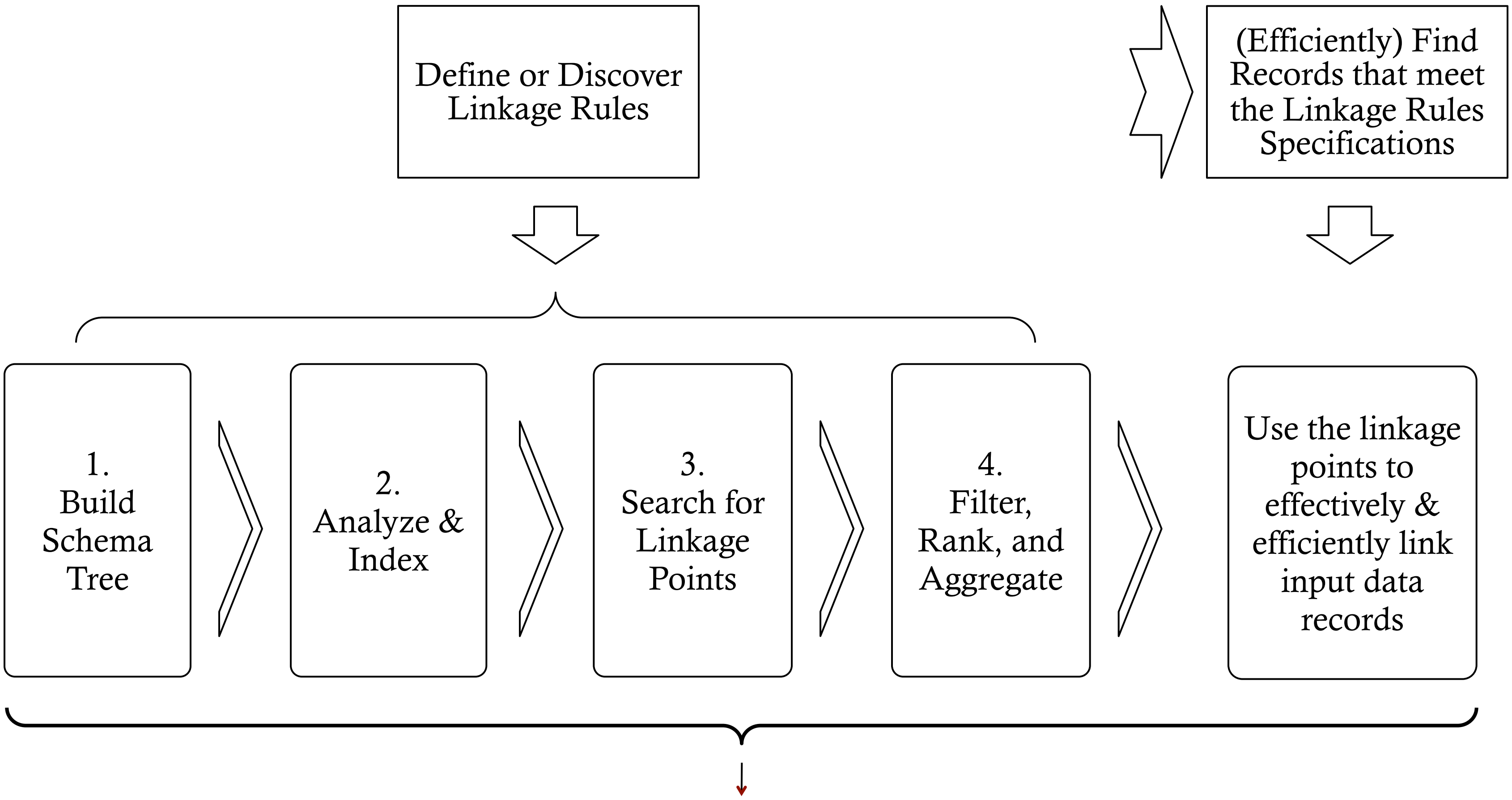
Examples


U.S. Securities and Exchange Commission
Freebase
DBpedia
Data Fusion Process
Define or Discover Linkage Rules
(Efficiently) Find Records that meet the Linkage Rules Specifications
Merge Records Identified as Duplicates
Consistent, Accurate Data
?

#src: number of values in source attribute records
#tgt: number of values in target attribute records
#match: number of records linked using the identified attribute pair

Source Attributes	Target Attributes	#src	#tgt	#match
DBpedia→company→SECcik	SEC→company→cik	20	1,981	1
DBpedia→company→owl#SameAs http://rdf.freebase.com/ns/guid.9202a8c04000641f80000000e6bb8ff	Freebase→company→id '/guid/9202a8c04000641f80000000e6bb8ff'	19,397	14,509	606
DBpedia→company→owl#SameAs http://rdf.freebase.com/ns/guid.9202a8c04000641f80000000e6bb8ff	Freebase→company→guid '#9202a8c04000641f80000000e6bb8ff'	19,397	73,273	17,423
Freebase→company→webpage	DBpedia→company→foaf#page	78,012	24,367	7
Freebase→company→webpage	DBpedia→company→webpagesURLs	78,012	74,589	16,216
Freebase→company→ticker_symbol	SEC→company→stockSymbol	75,374	1,981	1,439

Linkage Point Discovery



Define or Discover Linkage Rules

(Efficiently) Find Records that meet the Linkage Rules Specifications

1. Build Schema Tree

2. Analyze & Index

3. Search for Linkage Points

4. Filter, Rank, and Aggregate

Use the linkage points to effectively & efficiently link input data records

Novel record linkage pipeline, linking records that refer to the same or related real-world entities

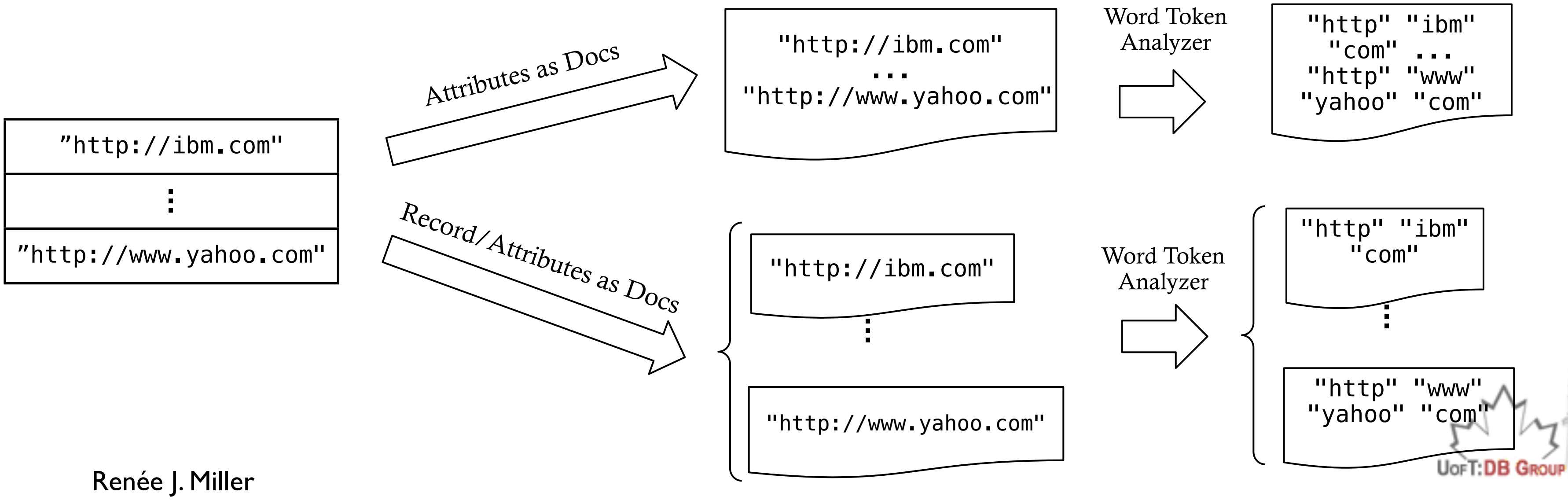
Analyze and Index

Instance Values ⇨ Virtual Documents

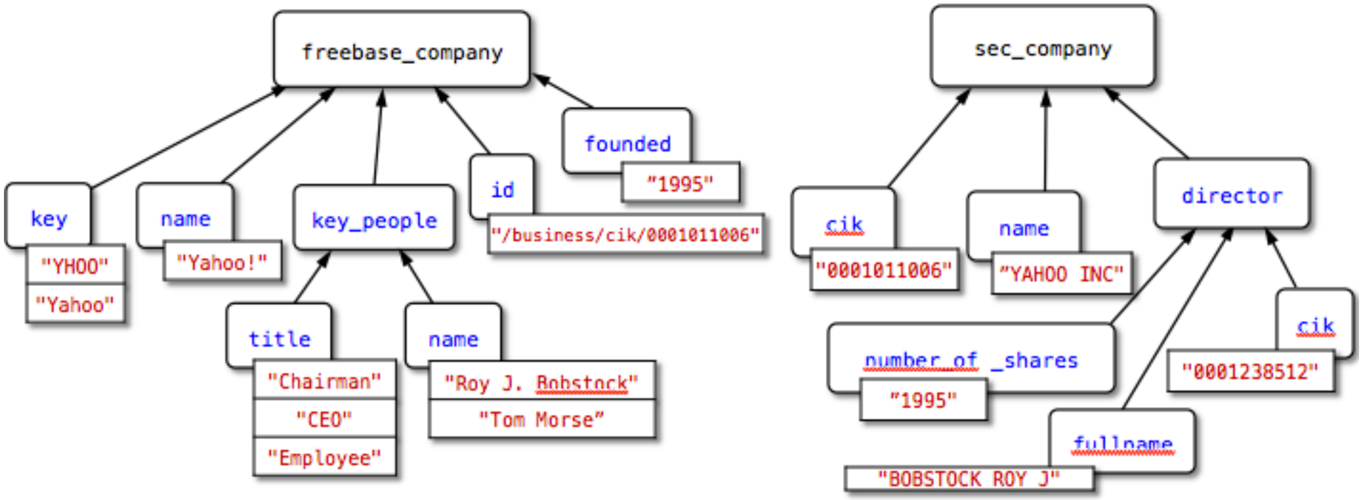
- Attribute value sets (multisets) as documents
- Record/Attribute values as documents

⇨ Analyze (tokenize) using a library of analyzers

- Exact: no transformation
- Lower: transform strings into lowercase
- Split: split strings by whitespace
- Word Token: replace special characters with whitespace and then split
- Q-gram: split strings into substrings of length *q*



Example linkage points



Example linkage points:

(`freebase_company` → `key`, `sec_company` → `name`)
using case insensitive substring matching as relevance function

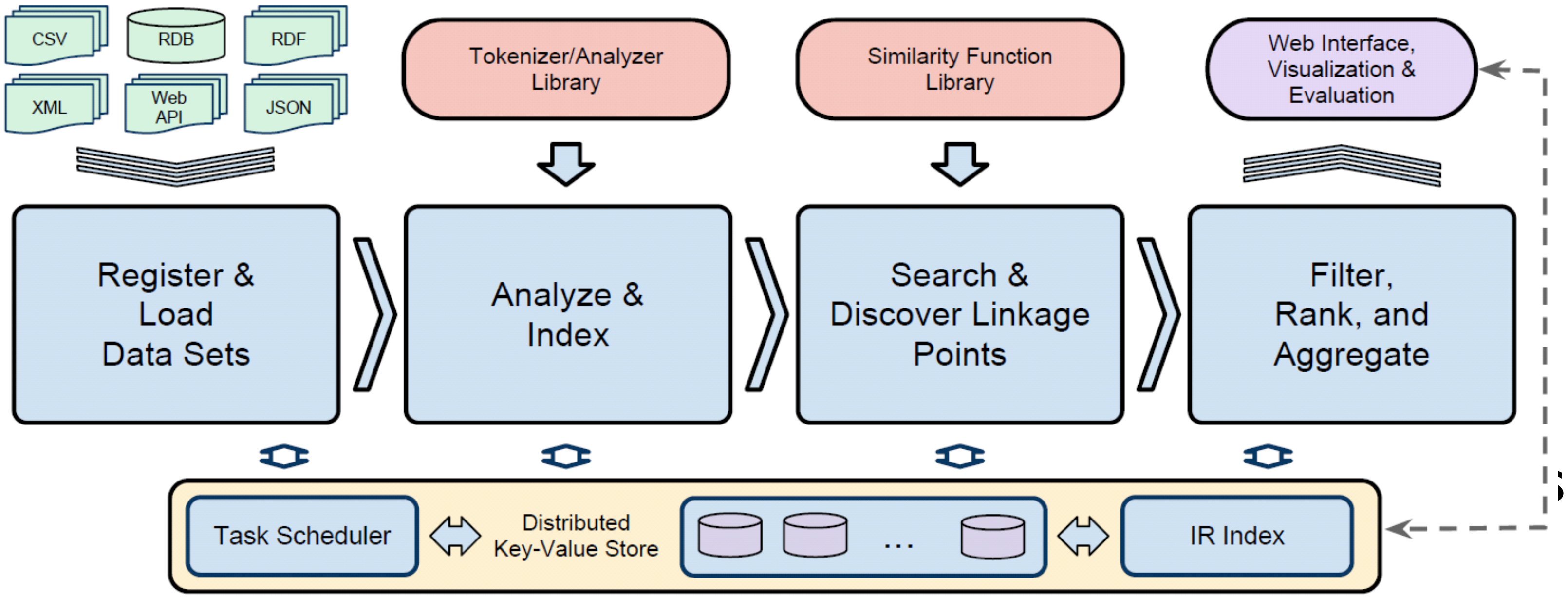
(`freebase_company` → `founded`, `sec_company` → `number_of_shares`)
using exact matching as relevance function

SMASH Algorithms

- Smash-S
  - Treat attributes (source and target) as documents and compare with set similarity measure
- Smash-R
  - Take sample of source attribute values and find best (k) matches to values in target attribute (uses indices to find efficiently)
- Smash-X: filter by
  - Cardinality: size of the linkage set
  - Coverage: % linked records in source or target data
  - Strength: % distinct records in the linkage set

Architecture

- Main features
  - Transformation module that supports any kind of semi-structured (Web) data


CSV
RDB
RDF
XML
Web API
JSON
Tokenizer/Analyzer Library
Similarity Function Library
Web Interface, Visualization & Evaluation
Register & Load Data Sets
Analyze & Index
Search & Discover Linkage Points
Filter, Rank, and Aggregate
Task Scheduler
Distributed Key-Value Store
…
IR Index

  - Web interface to visualize and evaluate results, and monitor tasks

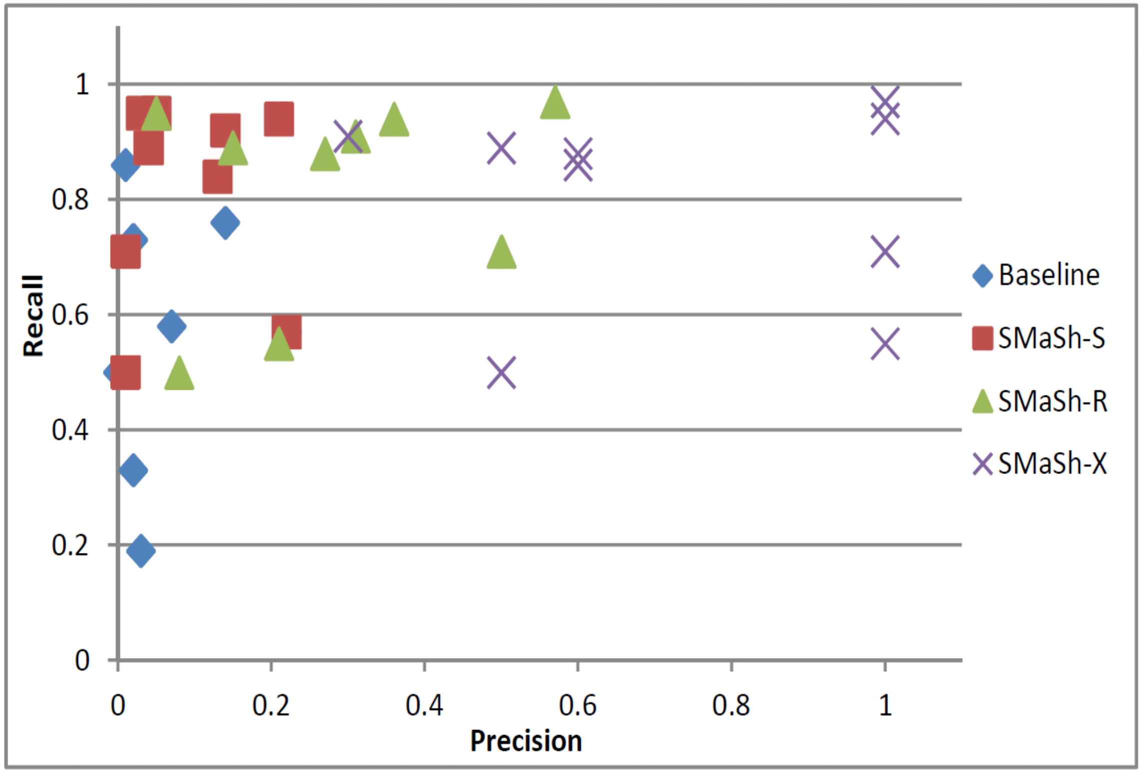
Quality Linkage Points

Entity	Source	Data Set	Rec#	Fact#	Attr#
Company	Freebase	`fbComp`	74,971	1.92M	167
	U.S. Securities and Exchange Commission	`secComp`	1,981	4.54M	72
	DBpedia	`dbpComp`	24,367	1.91M	1,738
Drug	Freebase	`fbDrug`	3,882	92K	56
	DrugBank Open Data Drug & Drug Target Database	`dbankDrug`	4,774	1.52M	145
	DBpedia	`dbpDrug`	3,662	216K	337
Movie	Freebase	`fbMovie`	42,265	899K	57
	IMDb	`imdbMovie`	14,405	483K	41
	DBpedia	`dbpMovie`	15,165	1.57M	1,021

Quality of Linkage Points

- □Each point represents one of the nine scenarios for each of the algorithms
- □Baseline: SMaSh-S with lower analyzer
  - ◘Resembles using a long-running script and the functionality of Web APIs to find linkage points
- □A linkage point is considered relevant if it consists of attributes that can be used to perform **record linkage** (finding records that refer to the same real-world entity)

Evaluation



- Precision: percentage of linkage points in the output that are relevant
- Recall: percentage of relevant linkage points in the output

Talk Themes

- Curation is ultimately about ***semantics***
  - Exploit modeled semantics & be principled in how missing semantics is created
- Curation is for ***humans***
  - Facilitate human understanding and decision making
  - People must be able to correct and understand curation decisions
- Curation focus on ***small(ish) valuable datasets***
  - Leverage ***Big Data*** to add value to curated data
  - Automation required not just for scale, but to manage deep complexity of curation tasks