

Entity Resolution: Tutorial

Lise Getoor

*University of Maryland
College Park, MD*

Ashwin Machanavajjhala

*Duke University
Durham, NC*

What is Entity Resolution?



Article [Talk](#)

Read

[Edit](#)

[View history](#)

Search



Record linkage

From Wikipedia, the free encyclopedia

(Redirected from [Entity resolution](#))

Record linkage (RL) refers to the task of finding **records** in a data set that refer to the same **entity** across different data sources (e.g., data files, books, websites, databases). Record linkage is necessary when **joining** data sets based on entities that may or may not share a common identifier (e.g., **database key**, **URI**, **National identification number**), as may

Name resolution

From Wikipedia, the free encyclopedia

Coreference

From Wikipedia, the free encyclopedia

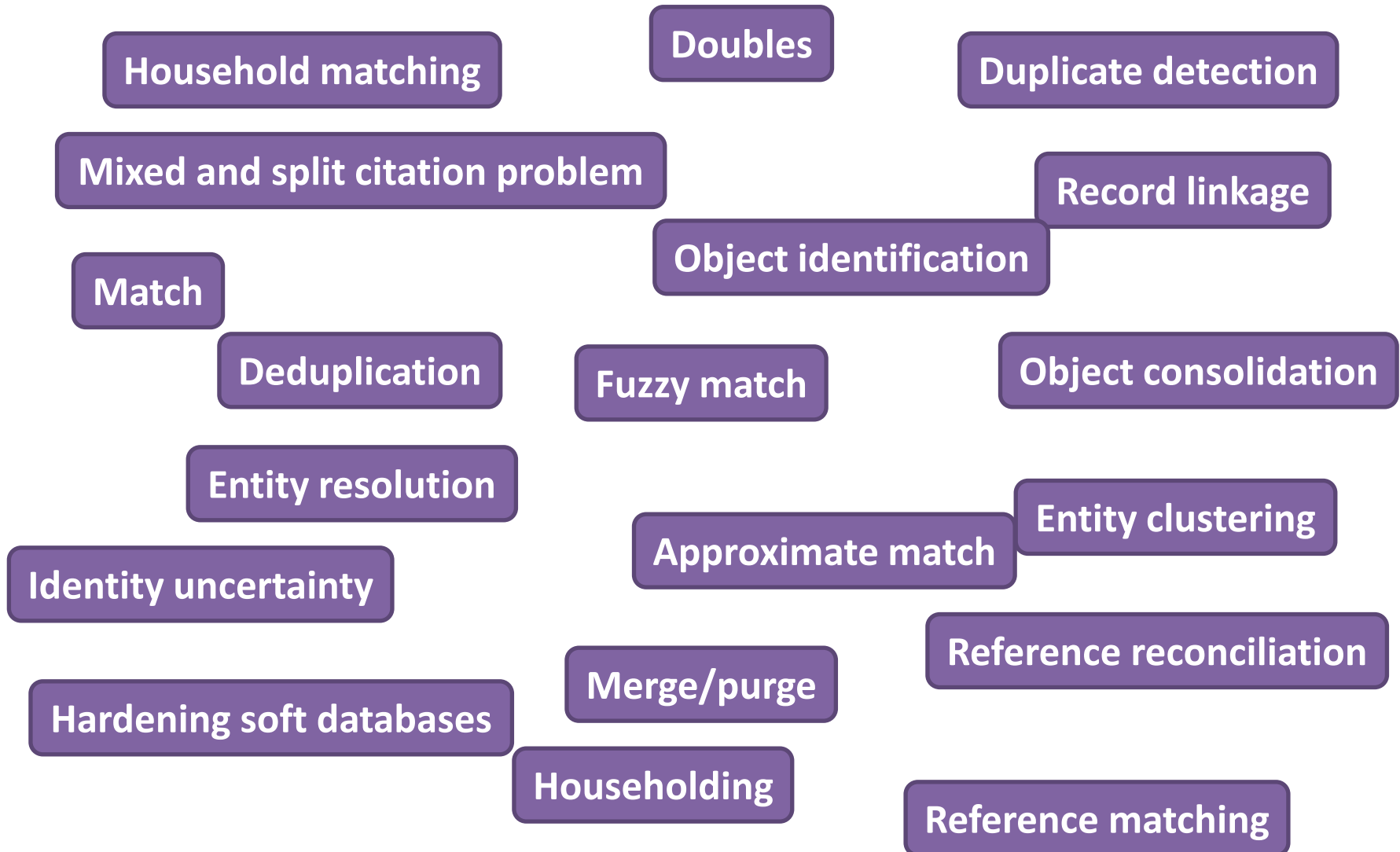
Deduplication

From Wikipedia, the free encyclopedia

Identity resolution

From Wikipedia, the free encyclopedia

Ironically, Entity Resolution has many duplicate names



Outline

1. Introduction & Motivation
 - a) Driving Applications
 - b) ER History
 - c) Growing Important of ER
2. Classical Single Entity ER
3. Efficiency: Blocking/Canopies
4. Relational & MultiEntity ER
5. Demo
6. Challenges & Future Directions

Outline

1. Introduction & Motivation
2. Classical Single Entity ER
 - a) Problem Statement
 - b) Data Preparation
 - c) Matching Features
 - d) Pairwise Approaches
 - e) Clustering-Based Approaches
 - f) Canonicalization
3. Efficiency: Blocking/Canopies
4. Relational & MultiEntity ER
5. Demo
6. Challenges & Future Directions

Outline

1. Introduction & Motivation
2. Classical Single Entity ER
- 3. Efficiency: Blocking / Canopies**
4. Relational & MultiEntity ER
5. Demo
6. Challenges & Future Directions

Outline

1. Introduction & Motivation
2. Classical Single Entity ER
3. Efficiency: Blocking/Canopies
4. Relational & MultiEntity ER
 - a) Problem Statement
 - b) Relational Features
 - c) Graph-based Approaches
 - d) Agglomerative Approaches
 - e) Generative Model Based Approaches
 - f) Declarative Approaches
 - g) Constraint Based
 - h) Efficiency
5. Demo
6. Challenges & Future Directions

Outline

1. Introduction & Motivation
2. Classical Single Entity ER
3. Efficiency: Blocking/Canopies
4. Relational & MultiEntity ER
5. Demo
 - a) DeDupe: Interactive Deduplication Tool
6. Challenges & Future Directions

PART 1


INTRODUCTION & MOTIVATION

Motivation: Census

- “Overview of Record Linkage and Current Research Directions”, William E Winkler, 2006
- The Post Enumeration Survey (PES) provided an independent re-enumeration of a large number of blocks (small Census regions) that corresponded to approximately 70 individuals. The PES was matched to the Census so that a capture-recapture methodology could be used to estimate both undercoverage and overcoverage to improve Census estimates. **In a very large 1990 Decennial Census application, the computerized procedures were able to reduce the need for clerks and field follow-up from an estimated 3000 individuals over 3 months to 200 individuals over 6 weeks (Winkler 1995).**

Motivation : Citation

- What is the most recent publication of Lei Chen?



Search

About 71,100 results

Everything

Images

Maps

Videos

News

Shopping

Books


More

DBLP: Lei Chen
[www.informatik.uni-trier.de](http://www.informatik.uni-trier.de/~dblp/people/LeiChen/)
Xiangmin Zhou, Lei Chen: Near-duplicate video retrieval
Apr 1, 2012 – Grasp@VT
consensus building

DBLP: Lei Chen
[www.informatik.uni-trier.de](http://www.informatik.uni-trier.de/~dblp/people/LeiChen/)
Mar 29, 2012 – L
Wisconsin, Madison

DBLP: Lei Chen
[www.informatik.uni-trier.de](http://www.informatik.uni-trier.de/~dblp/people/LeiChen/)
Mar 29, 2012 – L
Wisconsin, Madison

dblp.uni-trier.de
Computer Science
Bibliography

 SCHLOSS
Leibniz-Zentrum für Informationstechnik



Lei Chen

List of publications from the [DBLP Bibliography Server](#) - [FAQ](#)

other persons with the same name:

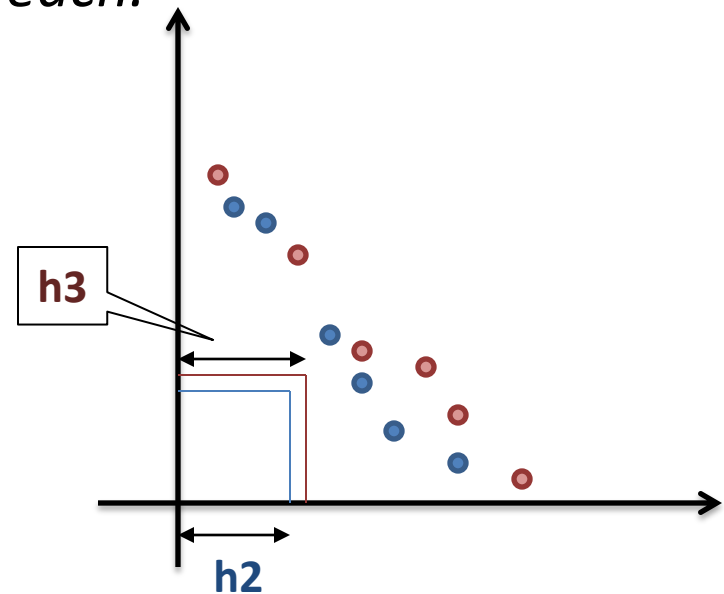
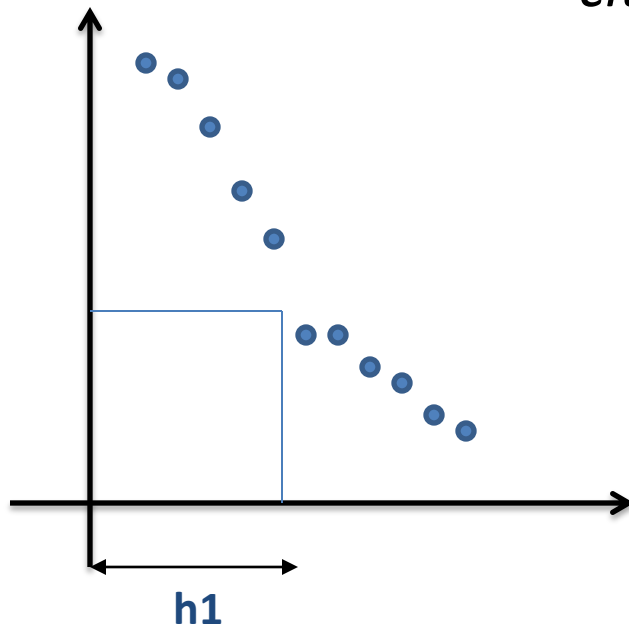
- [Lei Chen](#) - Purdue University, West Lafayette, IN
- [Lei Chen](#) - Rensselaer Polytechnic Institute, NY
- [Lei Chen](#) - Hong Kong University of Science and Technology
- [Lei Chen](#) - University of Wisconsin, Madison

Ask others: [ACM DL/Guide](#) - [CS](#) - [CSB](#) - [MetaPress](#) - [Google](#) - [Bing](#) - [Yahoo](#)

		2012
134		Muhammad Umar Farooq, Lei Chen, Lizy Kurian John: Compiler Support for Value-Based 185-199
133		Lei Chen, Guangnan Xing, Yingjie Xu, Xiaoxiang Liu, Tuanjie Zhao, Junyi Gai: Identification and flower in soybean with an integrative "omics" strategy. Computers & Electrical Engineering

ER and H-Index

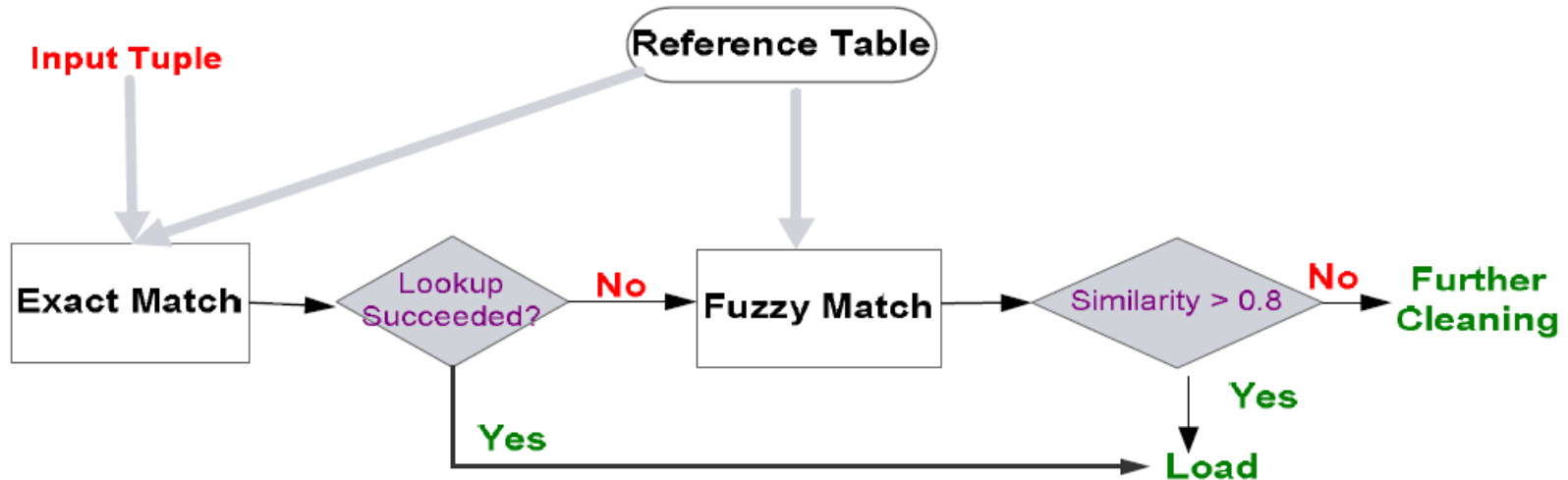
A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each.



$$h1 > h2 \text{ and } h1 > h3$$

Motivation: Data Cleaning

- [Chaudhuri et al, SIGMOD 2003]



- Reference table contains “clean” records
- Input table has “noisy” records
- Applications
 - Geocoding incoming queries
 - Match new customers to old ones
 - Products

Motivation: Data Cleaning

Canon PIXMA MG5220 -

[Product summary](#) [Find best price](#) [Specifications](#)



\$55.19 - \$203.99 (17 stores)

☐ [Compare](#)

[Find best price](#) [Narrow results](#)

Offer info

Merchant info



Majarra LLC



Amazon.com



Majarra LLC



Ecker Consulting LLC



The Office Dealer LLC



Best Cheap EStore

Canon PIXMA MG5220 Inkjet Multifunction Printer – Color – Photo Print – Desktop



Brand: CANON

Product Code: MG5220

Availability: In Stock

Price: ~~\$156.99~~ \$125.59

Ex Tax: \$125.59

Qty:

[Add to Cart](#)

- OR -

[Add to Wish List](#)
[Add to Compare](#)

★★★★★ 0 reviews | [Write a review](#)

[Share](#) [Email](#) [Print](#) [Facebook](#) [Twitter](#)

\$81.92

\$81.92 ▼

[Go to store](#)

Canon PIXMA MG5220 Inkjet Multifunction Printer – Color – Photo Print – Desktop



Brand: CANON

Product Code: 4502B017

Availability: In Stock

Price: ~~\$141.99~~ \$113.59

Ex Tax: \$113.59

Qty:

[Add to Cart](#)

- OR -

[Add to Wish List](#)
[Add to Compare](#)

★★★★★ 0 reviews | [Write a review](#)

[Share](#) [Email](#) [Print](#) [Facebook](#) [Twitter](#)

Motivation: Web Search

+You Search Images Maps Play YouTube News Gmail Documents Calendar More ▾



auto mechanics



Get directions

My places

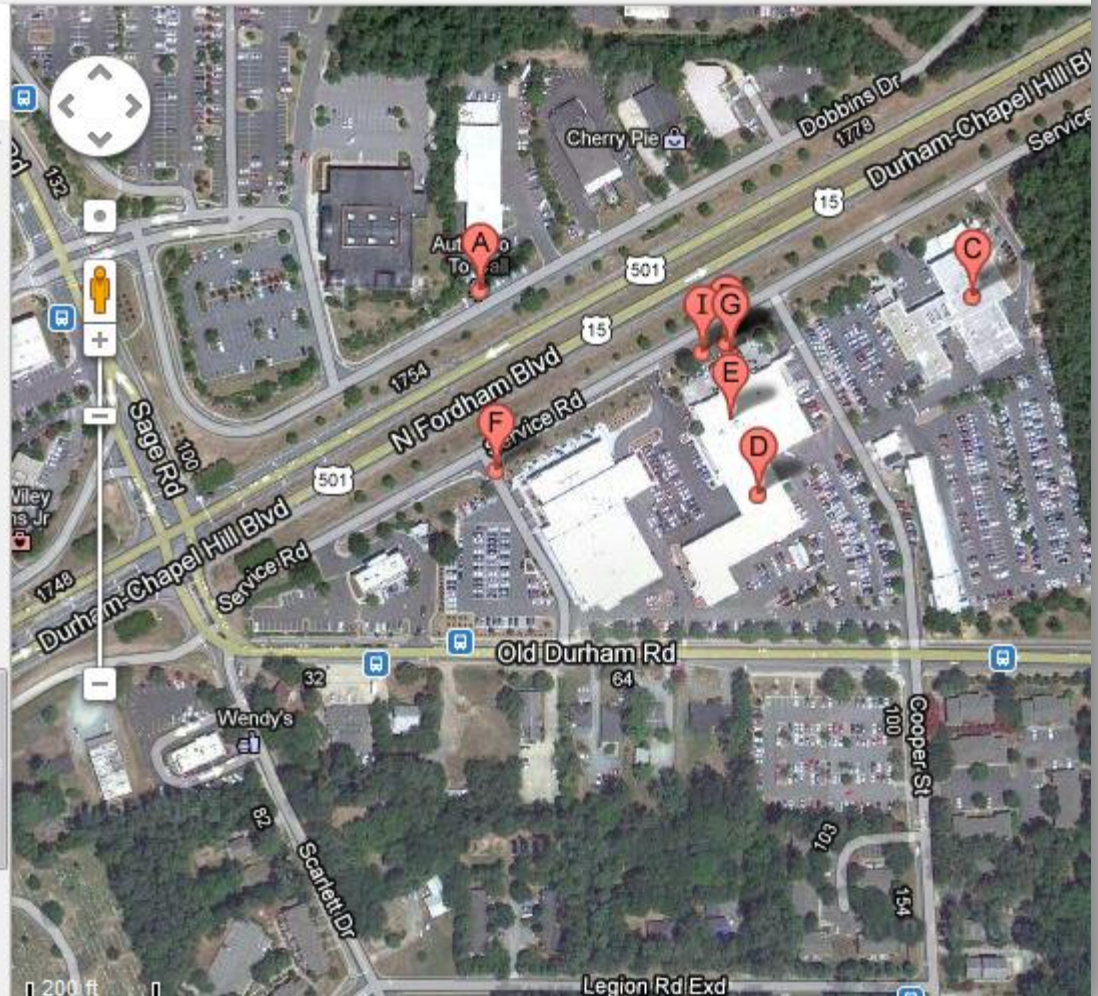


F **Performance Cosmetic Car Center** ▾
1810 Durham-Chapel Hill Boulevard #500, Chapel Hill, NC
(919) 942-3191
2 reviews
"MW Performance Service is second to none. I've purchased two cars from ..." -

G **Performance AutoMall** ▾
1810 Durham-Chapel Hill Blvd, Chapel Hill, NC
(888) 908-4949 - performanceautomall.com
Category: Auto Repair Shop

H **Performance AutoMall** ▾
1810 Durham-Chapel Hill Boulevard, Chapel Hill, NC
(888) 908-4949 - performanceautomall.com
Category: Car Repair and Maintenance

I **Performance AutoMall** ▾
1810 Durham-Chapel Hill Boulevard, Chapel Hill, NC
(919) 942-3191 - performanceautomall.com
Category: Auto Repair



Motivation: Web Search

2 [Auto Pro to Call](#)

| 1.35 mi.

★★★★★ (6 Reviews)

(919) 967-2271

1809 Fordham Blvd, Chapel Hill, NC 27514

[Directions](#) | [Send to Phone](#)

www.autoprotocall.com

These guys are crooks. They wanted \$100 just to put the meter on my check engine light a task that takes 2 minutes. \$100 just to diagnose it not to do any repairs. Places like Advance Auto... [more](#)

3 [Swedish Imports](#)

| 0.52 mi.

(919) 493-4545

5404 Durham Chapel Hill Blvd, Durham, NC 27707

[Directions](#) | [Send to Phone](#)

swedishimports.net

4 [N-Tune Automotive](#)

| 0.86 mi.

✓ Merchant verified

(919) 401-2612

411 Erwin Rd, Durham, NC 27707

[Directions](#) | [Send to Phone](#)

www.ntuneautomotive.com

5 [Auto Pro to Call](#)

| 1.35 mi.

★★★★★ (5 Reviews)

(919) 967-2271

1809 Fordham Blvd, Chapel Hill, NC 27514

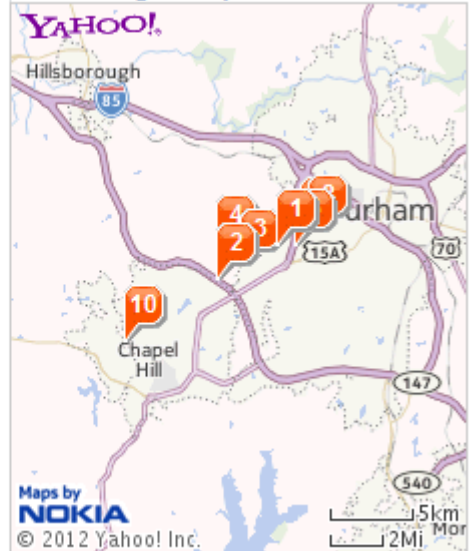
[Directions](#) | [Send to Phone](#)

www.autoprotocall.com

My family has been taking our cars to them for years since they were Chapel Hill Tire and they have always done great work at a fair price. You can trust them with your car: years ago a... [more](#)



[View Larger Map »](#)



Sponsored Results

[Raleigh Auto Repair](#)

A & J Automotive since 1996
Dependable Service, Honest
Answers

www.ajautorepair.com

[10% Off Any Auto Repair](#)

Plus Oil Change Combo Coupons for
\$21.95 or Less on Any Make or
Model

www.LocalBizNow.com

[Auto Mechanic School](#)

Become a **mechanic** with the Auto
Repair Technician program.

www.pennfoster.edu

Motivation: Machine Reading

NELL Knowledge Base Browser

CMU Read the Web Project

- awardtrophytournament
- creativework
 - book
 - movie
 - musicalalbum
 - visualartform
 - televisionshow
 - musicsong
 - lyrics
 - poem
- buildingmaterial
- celltype
- charactertrait
- chemical
- cognitiveactions
- event
 - conference
 - mlconference
 - election
 - sportsevent
 - sportsgame
 - race
 - grandprix
 - olympics
 - eventoutcome
- militaryeventtype
 - militaryconflict
- weatherphenomenon

See [metadata](#) for awardtrophytournament
1,526 instances, 1 page

instance

[american league pennant](#)
[australian open](#)
[british open](#)
[colonial cup](#)
[european cup winners cup](#)
[french open](#)
[indy 500](#)
[kentucky derby](#)
[masters](#)
[national league pennant](#)
[nba championship](#)
[nba finals](#)
[ncaa finals](#)
[nfl championship](#)
[rose bowl](#)
[stanley cup](#)
[super bowl](#)
[us open](#)
[wnba finals](#)

ER helps improve information extraction

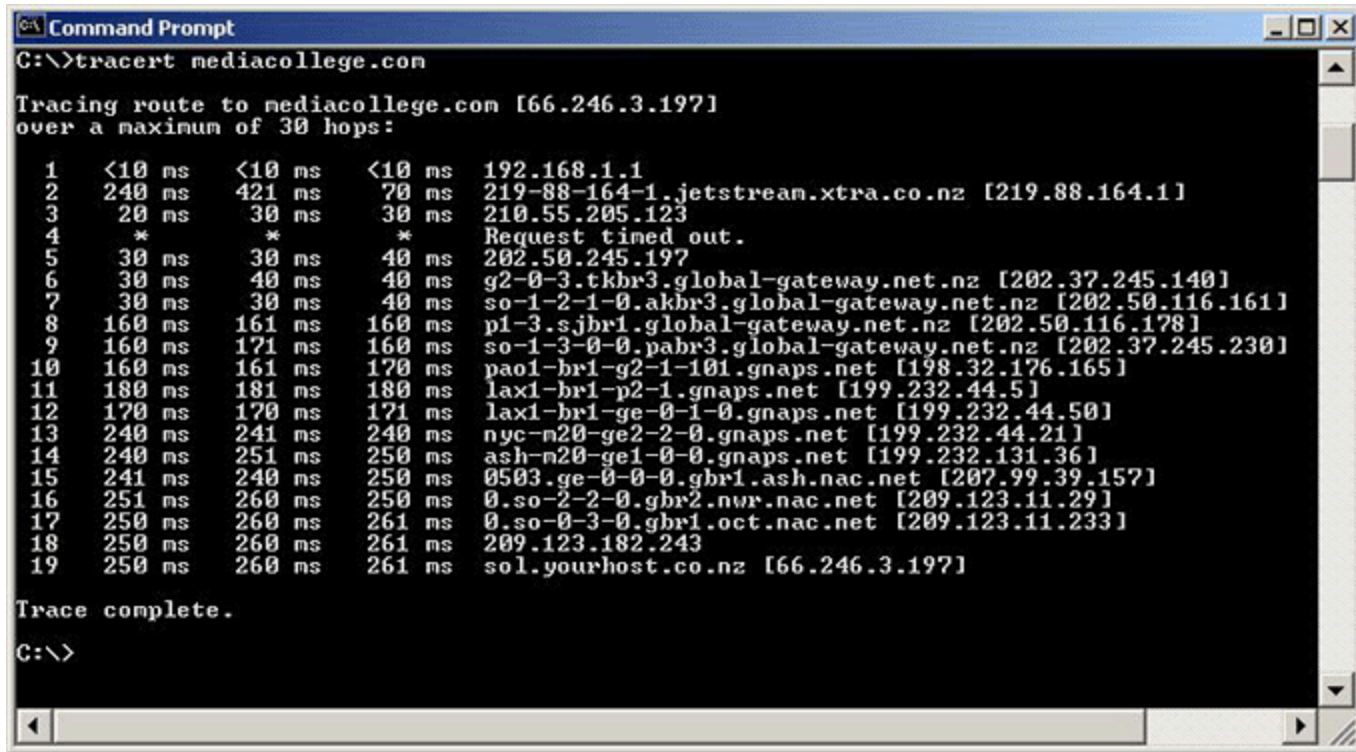
- If we know how to extract from one list, and the same entity appear on another differently formatted list, we can use the overlap for training an extractor on the second list. [Gupta et al VLDB11, Machanavajjhala et al WSDM11]

- *Arthur Charles Clarke*, born in Somerset, 1917.
- *Dave Barry*, born in Armonk, 1947.
- *Frank Herbert*, born in 1920.
- *Dame Agatha Christie*, born in Devon (UK), 1890.
- *Noam Chomsky*, born in Philadelphia.

- Noam Chomsky -- 7 December 1928.
- Agatha Christie -- 15 September 1890.
- John R. R. Tolkien -- 3 January 1892.
- Salman Rushdie -- 19 June 1947.

Motivation : Network Science

- Measuring the topology of the internet ... using traceroute



```
C:\>tracert mediacollege.com

Tracing route to mediacollege.com [66.246.3.197]
over a maximum of 30 hops:

  1  <10 ms  <10 ms  <10 ms  192.168.1.1
  2  240 ms  421 ms  70 ms  219-88-164-1.jetstream.xtra.co.nz [219.88.164.1]
  3  20 ms  30 ms  30 ms  210.55.205.123
  4  *      *      *      Request timed out.
  5  30 ms  30 ms  40 ms  202.50.245.197
  6  30 ms  40 ms  40 ms  g2-0-3.ttkbr3.global-gateway.net.nz [202.37.245.140]
  7  30 ms  30 ms  40 ms  so-1-2-1-0.akbr3.global-gateway.net.nz [202.50.116.161]
  8  160 ms  161 ms  160 ms  p1-3.sjbr1.global-gateway.net.nz [202.50.116.178]
  9  160 ms  171 ms  160 ms  so-1-3-0-0.pabr3.global-gateway.net.nz [202.37.245.230]
 10  160 ms  161 ms  170 ms  pao1-br1-g2-1-101.gnaps.net [198.32.176.165]
 11  180 ms  181 ms  180 ms  lax1-br1-p2-1.gnaps.net [199.232.44.5]
 12  170 ms  170 ms  171 ms  lax1-br1-ge-0-1-0.gnaps.net [199.232.44.50]
 13  240 ms  241 ms  240 ms  nyc-n20-ge2-2-0.gnaps.net [199.232.44.21]
 14  240 ms  251 ms  250 ms  ash-n20-ge1-0-0.gnaps.net [199.232.131.36]
 15  241 ms  240 ms  250 ms  0503.ge-0-0-0.gbr1.ash.nac.net [207.99.39.157]
 16  251 ms  260 ms  250 ms  0.so-2-2-0.gbr2.nwr.nac.net [209.123.11.29]
 17  250 ms  260 ms  261 ms  0.so-0-3-0.gbr1.oct.nac.net [209.123.11.233]
 18  250 ms  260 ms  261 ms  209.123.182.243
 19  250 ms  260 ms  261 ms  sol.yourhost.co.nz [66.246.3.197]

Trace complete.

C:\>
```

IP Aliasing Problem [Willinger et al. 2009]

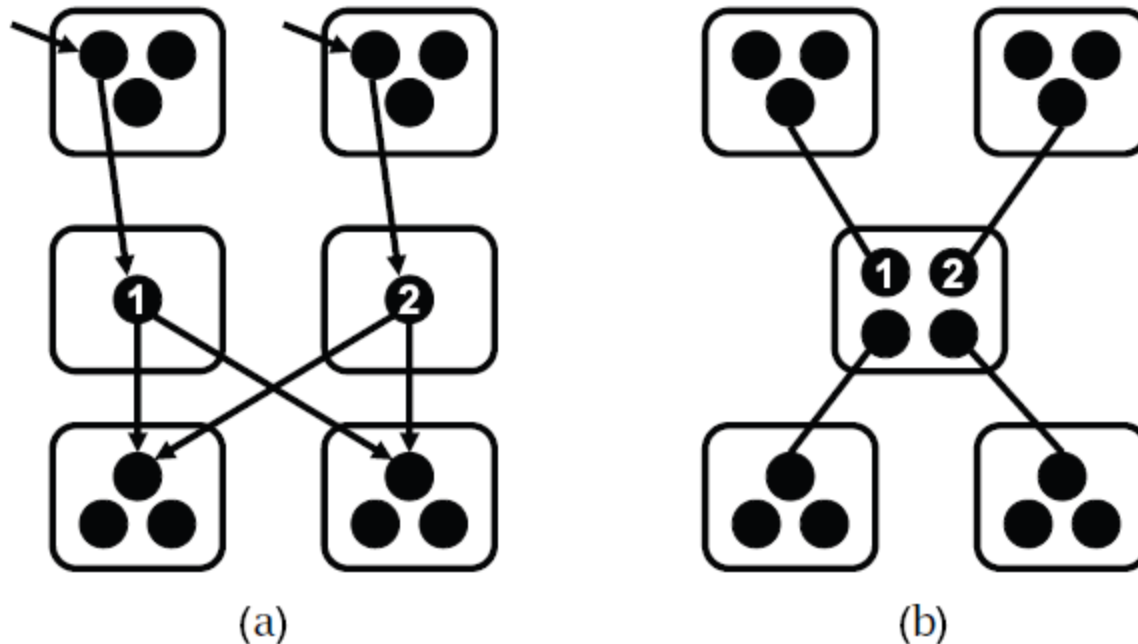


Figure 2. The IP alias resolution problem. Paraphrasing Fig. 4 of [50], traceroute does not list routers (boxes) along paths but IP addresses of input interfaces (circles), and alias resolution refers to the correct mapping of interfaces to routers to reveal the actual topology. In the case where interfaces 1 and 2 are aliases, (b) depicts the actual topology while (a) yields an “inflated” topology with more routers and links than the real one.

IP Aliasing Problem [Willinger et al. 2009]

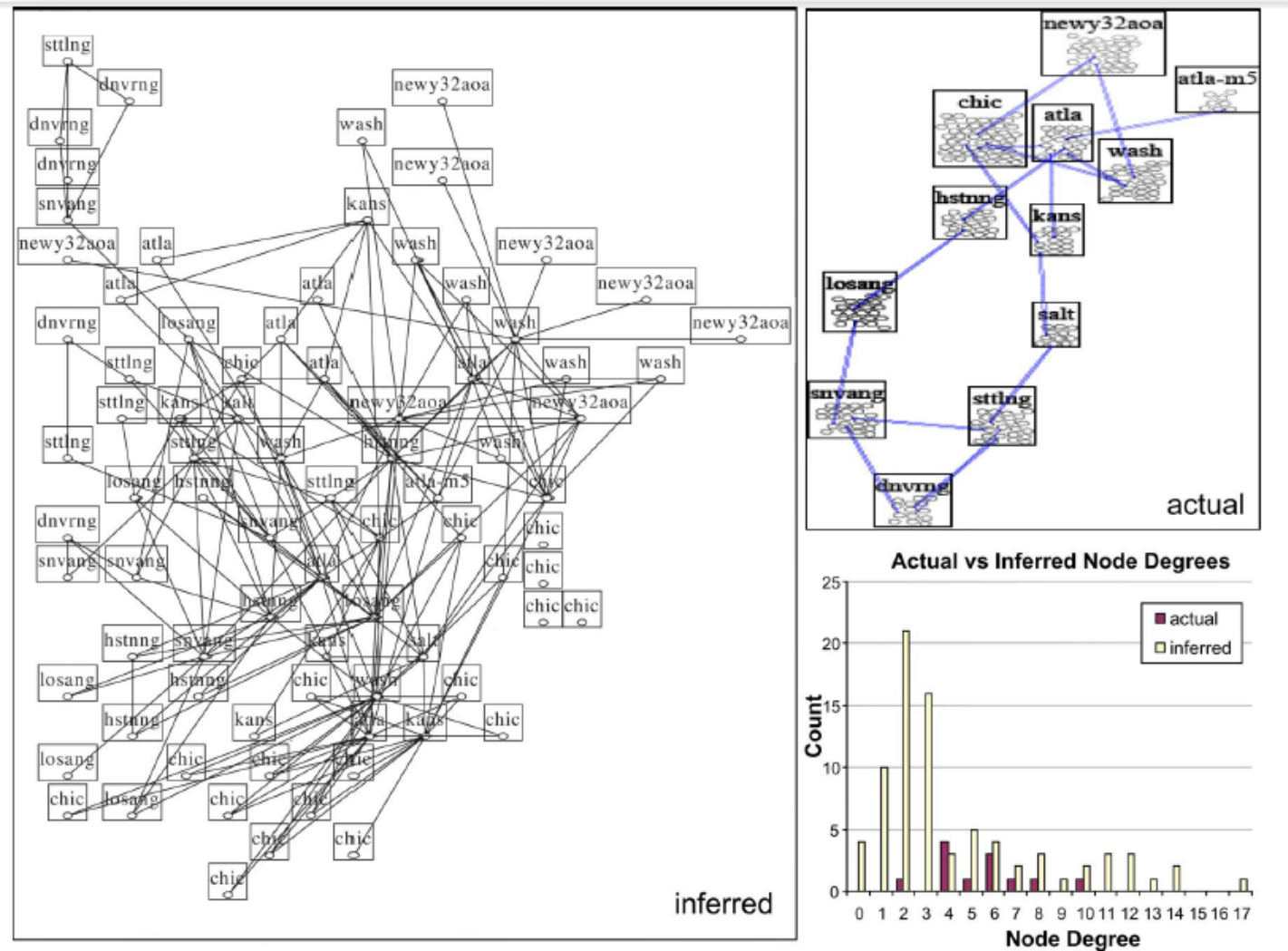


Figure 3. The IP alias resolution problem in practice. This is re-produced from [48] and shows a comparison between the Abilene/Internet2 topology inferred by Rocketfuel (left) and the actual topology (top right). Rectangles represent routers with interior ovals denoting interfaces. The histograms of the corresponding node degrees are shown in the bottom right plot. © 2008 ACM.

Historical ER Challenges

- Name/Attribute ambiguity

Thomas Cruise



Michael Jordan



Historical ER Challenges

- Errors due to data entry



↓	C1	C2
	Total Cholesterol_1	Total Cholesterol_2
682	214.4	214.4
683	184.4	184.4
684	183.5	183.5
685	240.7	240.7
686	215.1	215.1
687	198.6	198.6
688	2800.0	280.0
689	210.8	210.8
690	182.5	182.5
691	192.6	192.6

Historical ER Challenges

- Missing Values

Exhibit 2: Examples of variables that are set to unknown values

Administrative dates: set to 0101YY, 010199, 999999

Date of Birth 0101YY, 1506YY, 3006YY, 0107YY, 1507YY, 0101YEAR

Names: set to spaces, NK, UNKNOWN, or ZZZZ
BABY, MALE, FEMALE, TWIN, TRIPLET, INFANT

Other variables: set to 9, 99, 9999, -1

NK (Not Known)

NA (Not applicable)

NC (Not coded)

U (Unknown)

[Gill et al; Univ of Oxford 2003]

Historical ER Challenges

- Changing Attributes

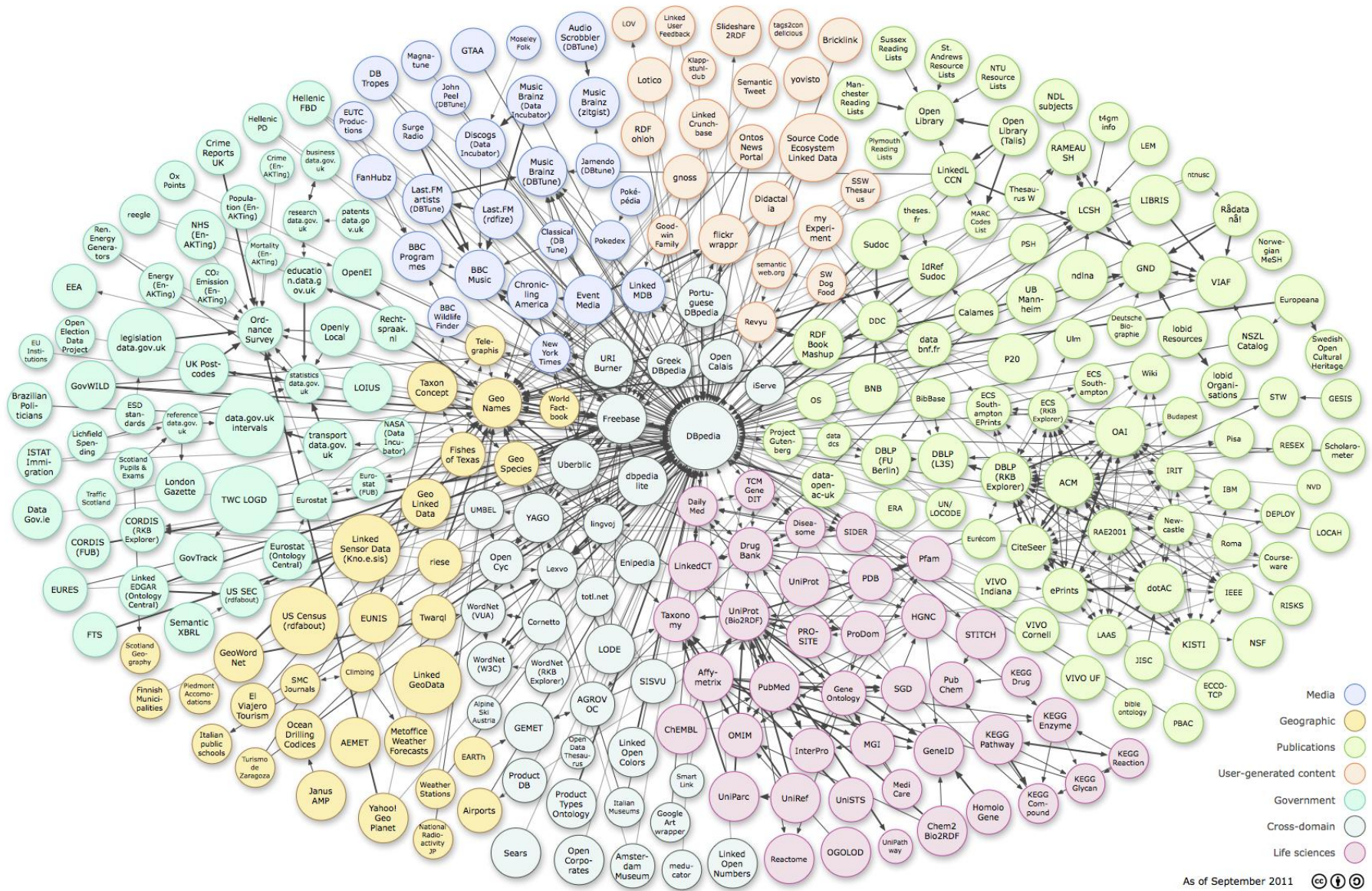


- Data formatting



- Abbreviations / Data Truncation

New ER Challenges



- **More Data**
 - Need parallel techniques
- **More Heterogeneity**
 - Unstructured, Unclean and Incomplete data. Diverse data types.
- **More linked**
 - Need to infer relationships in addition to “equality”
- **Multi-Relational**
 - Deal with structure of entities (Are Walmart and Walmart Pharmacy the same?)
- **Multi-domain**
 - Customizable methods that span across domains
- **Multiple applications** (web search versus comparison shopping)
 - Served diverse application with different accuracy requirements

