

# Entity Resolution

*CompSci 590.03*

*Instructor: Ashwin Machanavajjhala*

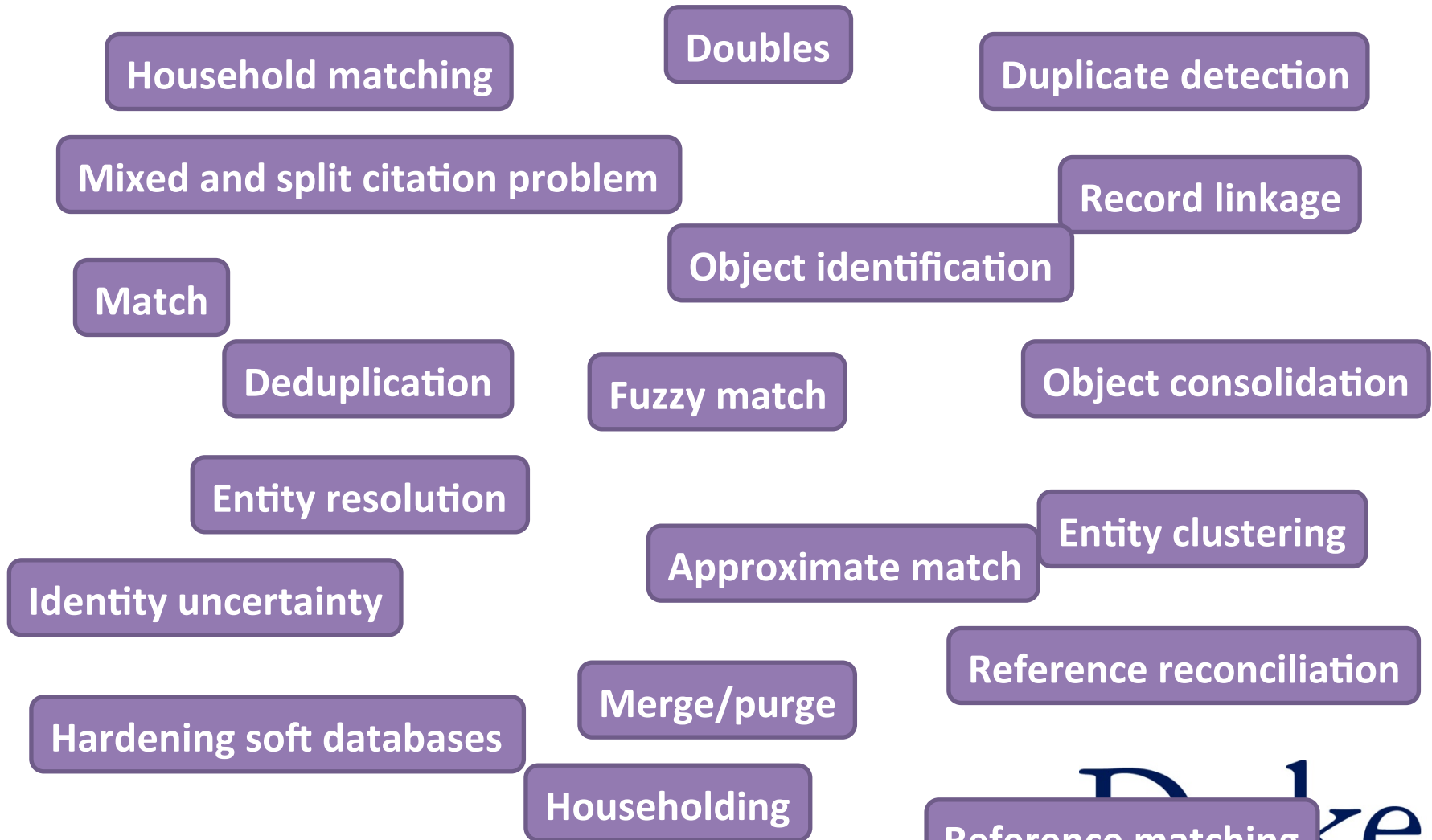
# What is Entity Resolution?

*Problem of identifying and linking/grouping different manifestations of the same real world object.*

Examples of manifestations and objects:

- Different ways of addressing (names, email addresses, FaceBook accounts) the same person in text.
- Web pages with differing descriptions of the same business.
- Different photos of the same object.
- ...

# Ironically, Entity Resolution has many duplicate names



# Outline

- Introduction
  - Driving Applications
  - Challenges
- Problem Formulation
  - Single Entity ER
  - Relational & Multi-Entity ER
- Algorithms for Single Entity ER
  - Computing Pairwise Match scores
  - Blocking: Efficiently Identifying of Near-Duplicates
  - Correlation Clustering: Enforcing Transitivity Constraints
- Algorithms for Relational & Multi-Entity ER




# 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 .uni-trier.de

Computer Science Bibliography













**Lei Chen** 🤖 📄

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

other persons with the same name:

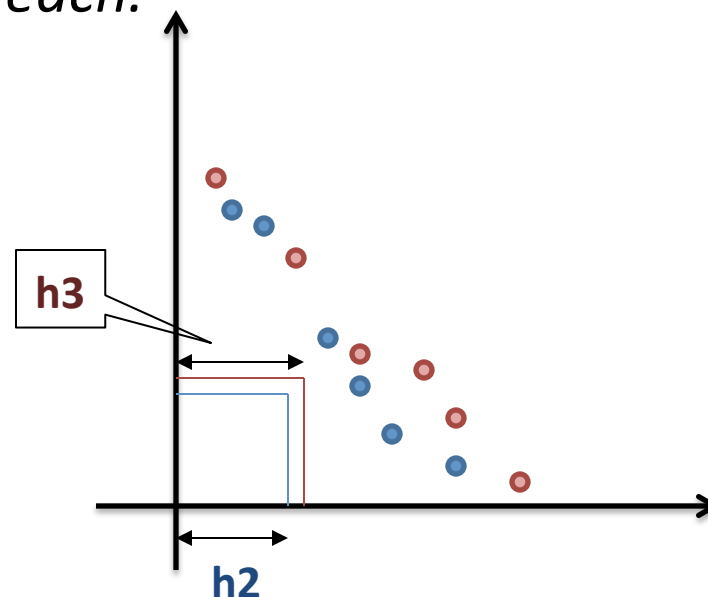
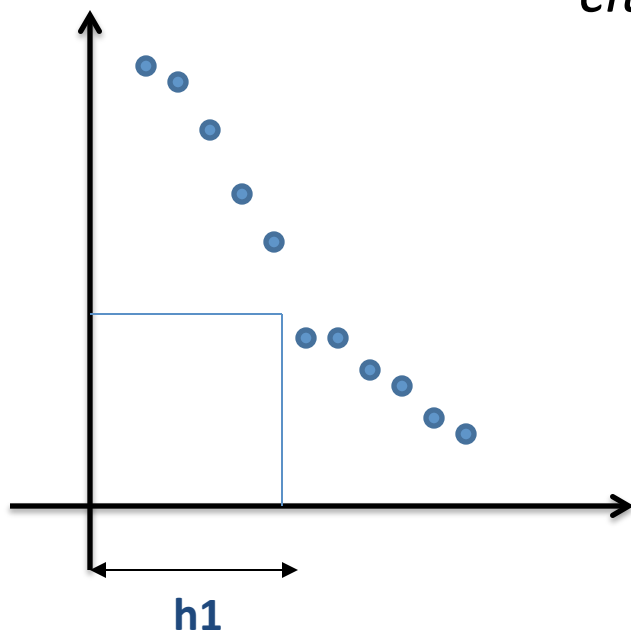
- [Lei Chen](#) - Purdue University, West Lafayette, IN
- [Lei Chen](#) - Rensselaer Polytechnic Institute, NY
- [Lei Chen](#) - Hong Kong University of Science and Technology
- [Lei Chen](#) - University of Wisconsin, Madison

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

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

# ER and H-Index

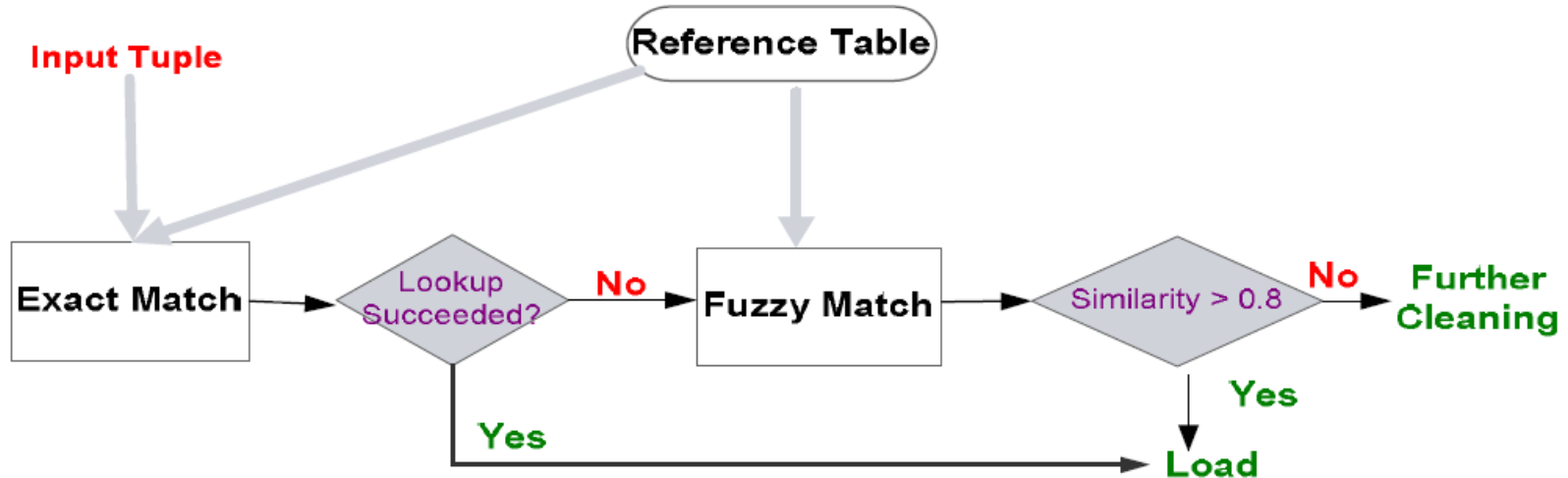
*A scientist has index  $h$  if  $h$  of his/her  $N_p$  papers have at least  $h$  citations each, and the other  $(N_p - h)$  papers have no more than  $h$  citations each.*



**$h1 > h2$  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: 1

[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: ~~\$144.99~~ \$113.59

Ex Tax: \$113.59

Qty: 1

[Add to Cart](#)

- OR -

[Add to Wish List](#)

[Add to Compare](#)

★★★★★ 0 reviews | [Write a review](#)

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

# Motivation : Web Search



**yelp** Search for (e.g. taco, cheap dinner, Max's)

Welcome About Me Write a Review Find Reviews Find Friends

### C5 Restaurant

★★★★☆ 20 reviews Rating Details

Categories: Restaurants, Lounges [Edit]

100 Queen's Park  
Royal Ontario Museum  
Toronto, ON M5S 2C6  
Neighborhoods: Discovery District, Downtown Core

(416) 586-7928  
<http://www.c5restaurant.ca/>

**Make a Reservation**

Date & Time: 04/16/2012 Party Size: 2 Find a Table

**Hours:**  
Mon-Sun 11 pm - 3 pm  
**Good for Kids:** No  
**Accepts Credit Cards:** Yes  
**Parking:** Street  
**Attire:** Dressy  
**Good for Groups:** Yes

**Price Range:** \$\$\$\$  
**Takes Reservations:** Yes  
**Delivery:** No  
**Take-out:** Yes  
**Waiter Service:** Yes  
**Outdoor Seating:** No

**Happy Hour:** No  
**Alcohol:** Full Bar  
**Smoking:** No  
**Coat Check:** Yes  
**Has TV:** No  
**Wheelchair Accessible:** Yes



**OpenTable®** Restaurant Reservations - Free • Instant • Confirmed

[OpenTable Home](#) > [Toronto / Ontario restaurants](#) > [Yorkville restaurants](#) > [C5 Restaurant Lounge inform](#)

### Reserve at C5 Restaurant Lounge

7:00 PM 2 people Find a Table

Overview Reviews Private Dining

### C5 Restaurant Lounge

★★★★☆ See all 101 Reviews >

**Address:**  
100 Queen's Park  
Toronto, ON M5S 266

**Cuisine:**  
International

**Price:** CAN\$31 to CAN\$50

**Neighborhood:**  
Yorkville

**More Details >**

[Getting there >](#)





# Motivation: Web Search

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



auto mechanics



Get directions

My places



**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 ..." -

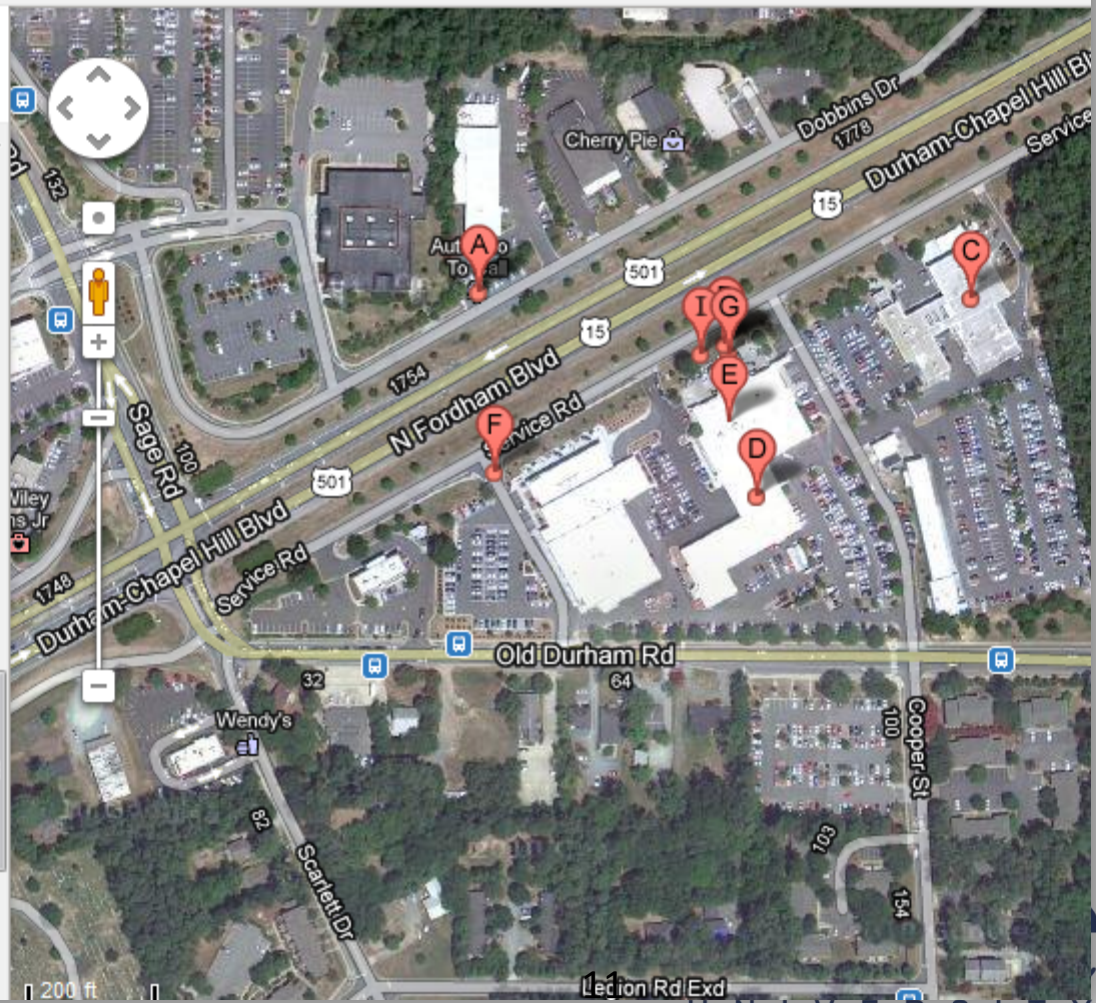
**Performance AutoMall** ▾  
1810 Durham-Chapel Hill Blvd, Chapel Hill, NC  
(888) 908-4949 - [performanceautomall.com](http://performanceautomall.com)  
Category: Auto Repair Shop



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

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

Lecture 18 : 590.02 Spring 13



UNIVERSITY

# 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](http://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](http://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](http://www.ntuneautomotive.com)

## 5 [Auto Pro to Call](#)

| 1.35 mi.

★★★★★ (5 Reviews)

(919) 967-2271

1809 Fordham Blvd, Chapel Hill, NC 27514

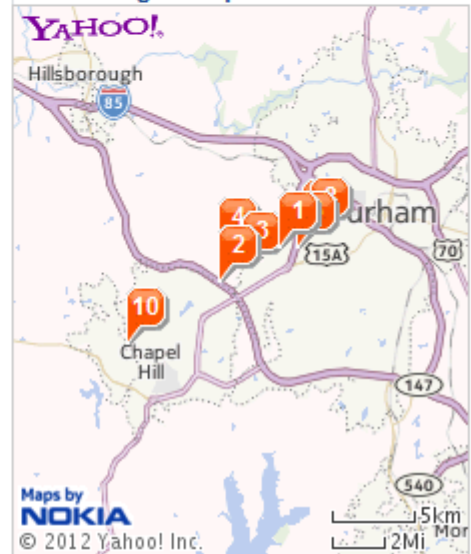
[Directions](#) | [Send to Phone](#)

[www.autoprotocall.com](http://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](http://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](http://www.LocalBizNow.com)

### [Auto Mechanic School](#)

Become a mechanic with the Auto  
Repair Technician program.

[www.pennfoster.edu](http://www.pennfoster.edu)



# Motivation: Machine Reading

## NELL Knowledge Base Browser

CMU Read the Web Project

- awardtrophytournament
- creativework
  - book
  - movie
  - musicalalbum
  - visualartform
- televisionshow
- musicson
- 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

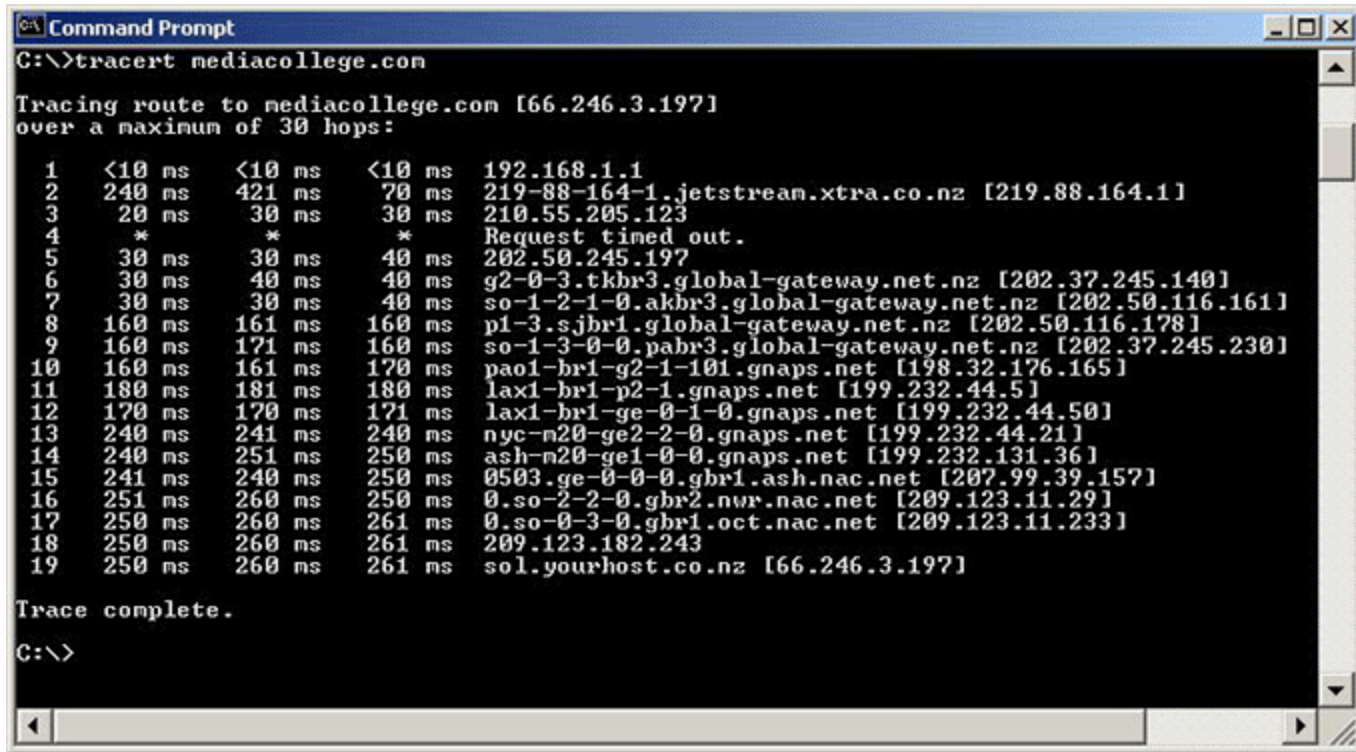
- If we know how to extract from one list, and the same entity appear on another differently formatted list, we can use the overlap for training an extractor on the second list. [Gupta et al VLDB11, Machanavajjhala et al WSDM11]

- *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



```
Command Prompt
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]

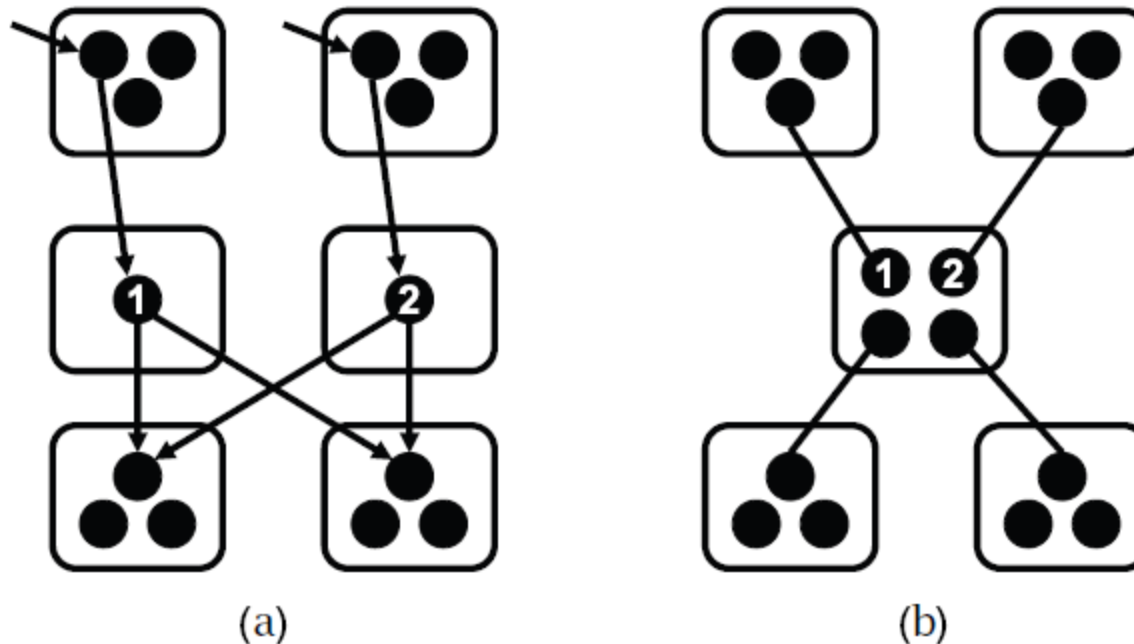


Figure 2. The IP alias resolution problem. Paraphrasing Fig. 4 of [50], traceroute does not list routers (boxes) along paths but IP addresses of input interfaces (circles), and alias resolution refers to the correct mapping of interfaces to routers to reveal the actual topology. In the case where interfaces 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]

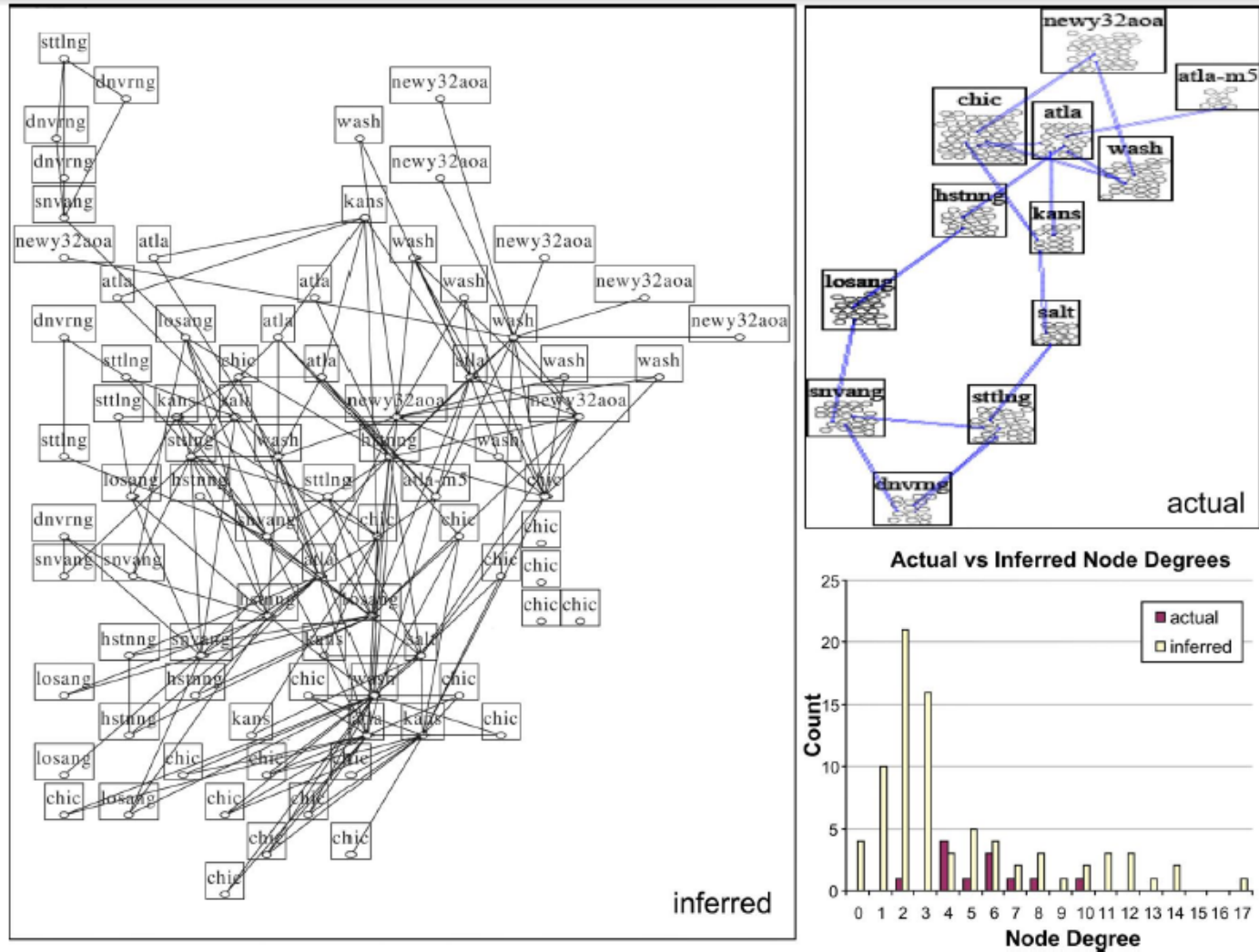


Figure 3. The IP alias resolution problem in practice. This is re-produced from [48] and shows a comparison between the Abilene/Internet2 topology inferred by Rocketfuel (left) and the actual topology (top right). Rectangles represent routers with interior ovals denoting interfaces. The

# Motivation: Privacy in Big-Data Analysis

- Datasets collected by different organizations can't be shared as is due to privacy concerns
- Individuals are de-identified before publishing the data
- May want to identify correlations between de-identified datasets
  - Join medical records from a hospital with locations tracked by a cell phone provider to identify correlations between activity and health.
  - Google Flu: correlation search logs with flu incidence.
  - ...

# Outline

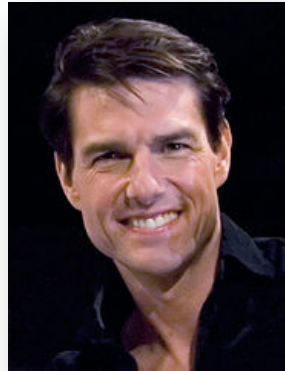
- Introduction
  - Driving Applications
  - Challenges
- Problem Formulation
- Algorithms for Single Entity ER
- Algorithms for Relational & Multi-Entity ER



# Traditional Challenges in ER

- Name/Attribute ambiguity

**Thomas Cruise**



**Michael Jordan**





# Traditional Challenges in ER

- Name/Attribute ambiguity
- 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

# Traditional Challenges in ER

- Name/Attribute ambiguity
- Errors due to data entry
- 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

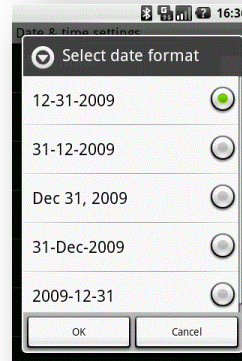
NK (Not Known)  
NA (Not applicable)  
NC (Not coded)  
U (Unknown)

# Traditional Challenges in ER

- Name/Attribute ambiguity
- Errors due to data entry
- Missing Values
- Changing Attributes

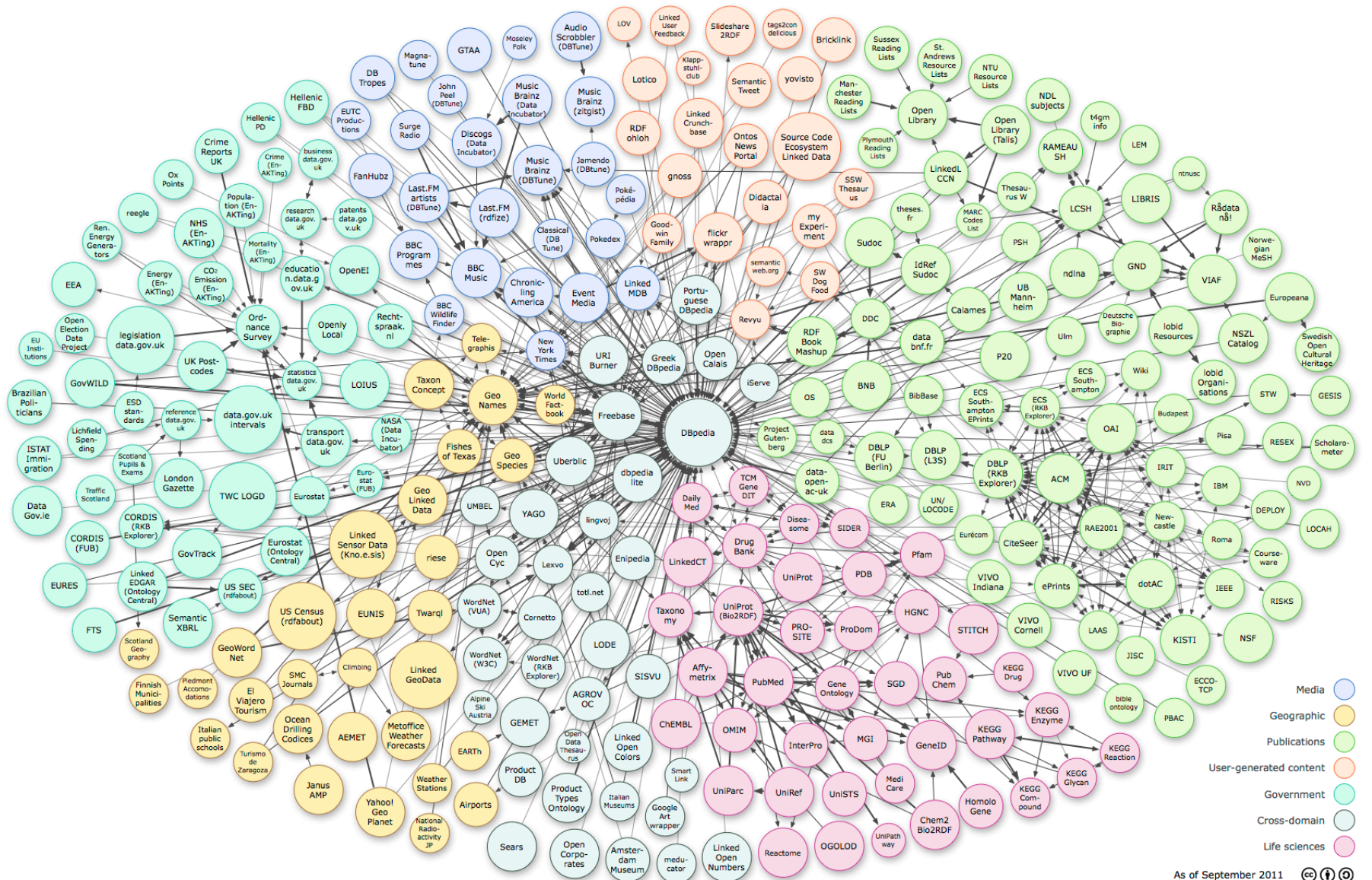


- Data formatting



- Abbreviations / Data Truncation

# Big-Data ER Challenges



As of September 2011

# Big-Data ER Challenges

- Larger and more Datasets
  - Need efficient parallel techniques
- More Heterogeneity
  - Unstructured, Unclean and Incomplete data. Diverse data types.
  - No longer just matching names with names, but Amazon profiles with browsing history on Google and friends network in Facebook.



# Big-Data ER Challenges

- Larger and more Datasets
  - Need efficient 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)
  - Serve diverse application with different accuracy requirements

# ER References

- Book / Survey Articles
  - Data Quality and Record Linkage Techniques [T. Herzog, F. Scheuren, W. Winkler, Springer, '07]
  - Duplicate Record Detection [A. Elmagrid, P. Ipeirotis, V. Verykios, TKDE '07]
  - An Introduction to Duplicate Detection [F. Naumann, M. Herschel, M&P synthesis lectures 2010]
  - Evaluation of Entity Resolution Approached on Real-world Match Problems [H. Kopke, A. Thor, E. Rahm, PVLDB 2010]
  - Data Matching [P. Christen, Springer 2012]
- Tutorials
  - Record Linkage: Similarity measures and Algorithms [N. Koudas, S. Sarawagi, D. Srivatsava SIGMOD '06]
  - Data fusion--Resolving data conflicts for integration [X. Dong, F. Naumann VLDB '09]
  - Entity Resolution: Theory, Practice and Open Challenges <http://goo.gl/Ui38o> [L. Getoor, A. Machanavajjhala AAAI '12]

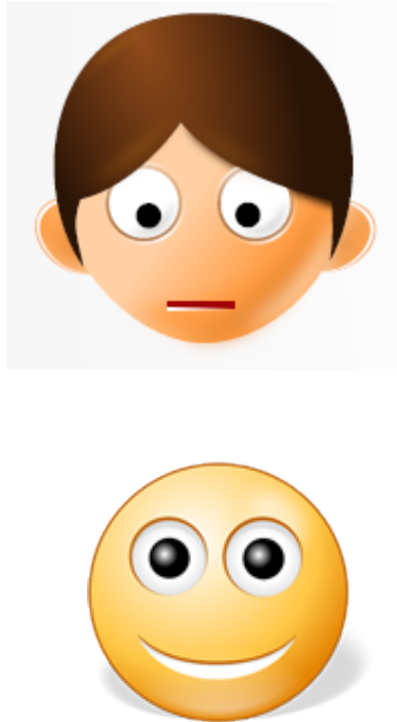
# Outline

- Introduction
- Problem Formulation
  - Single Entity ER
  - Relational & Multi-Entity ER
- Algorithms for Single Entity ER
- Algorithms for Relational & Multi-Entity ER



# Single Entity Problem Statement

## Real World

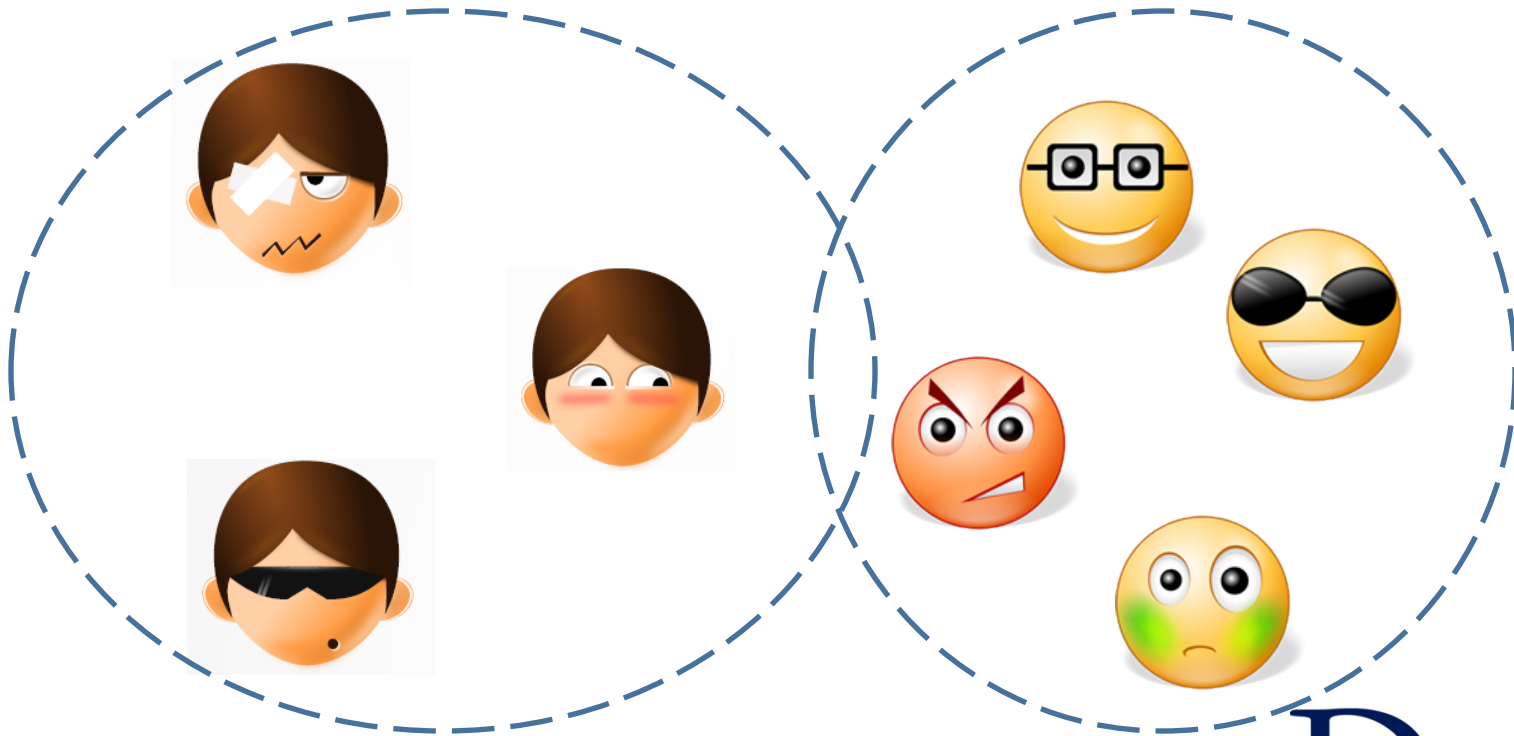


## Digital World



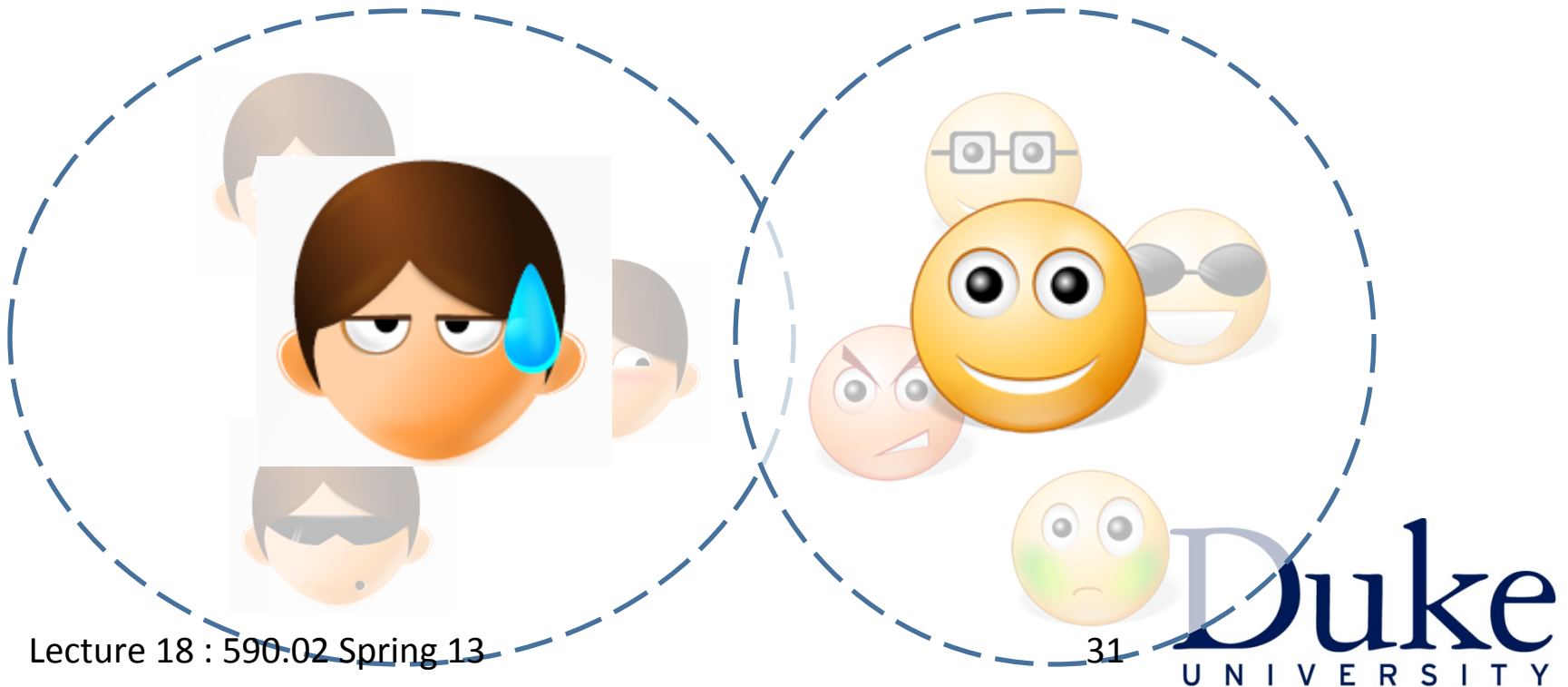
# Deduplication Problem Statement

- Cluster the records/mentions that correspond to same entity



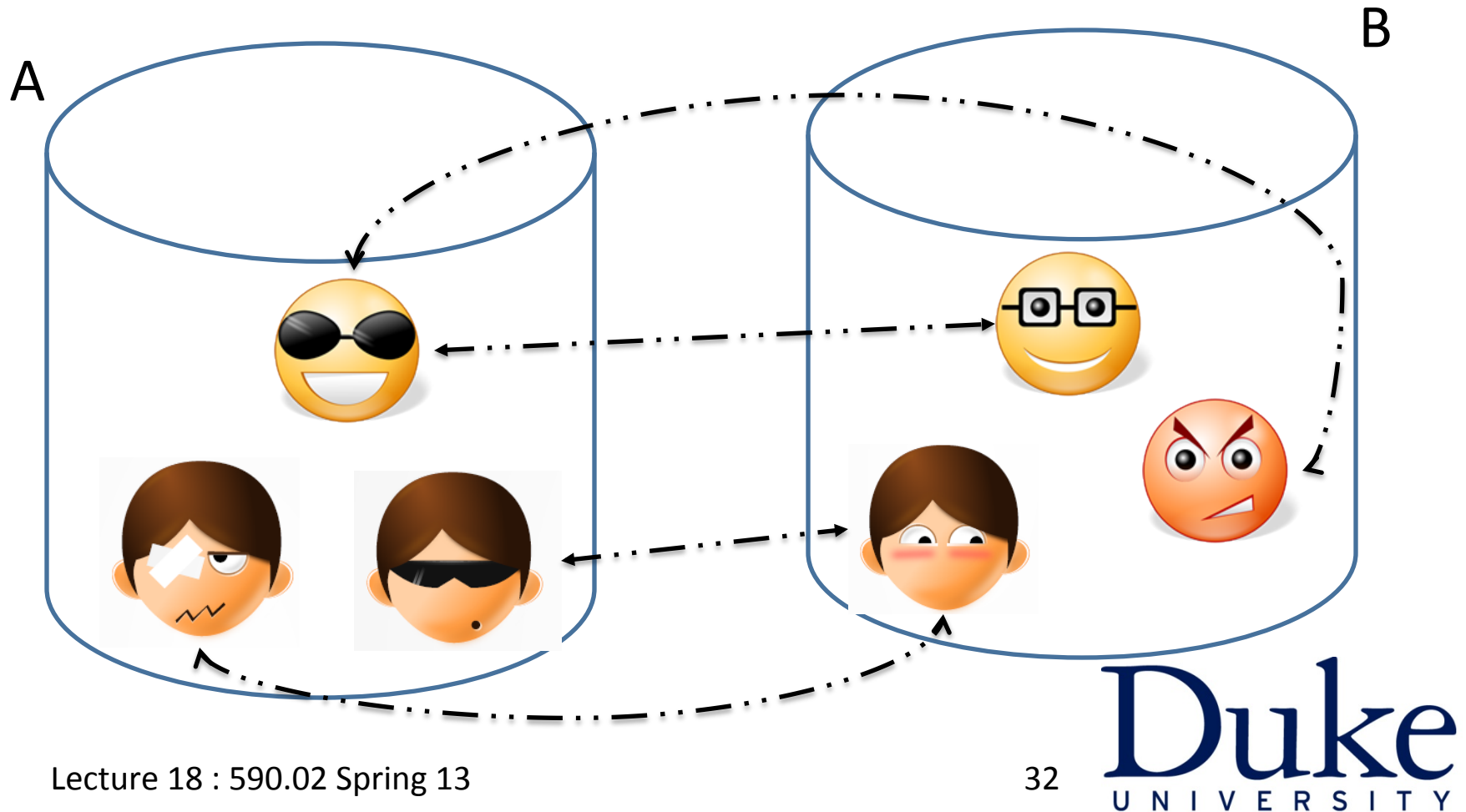
# Deduplication Problem Statement

- Cluster the records/mentions that correspond to same entity
  - Intensional Variant:** Compute cluster representative



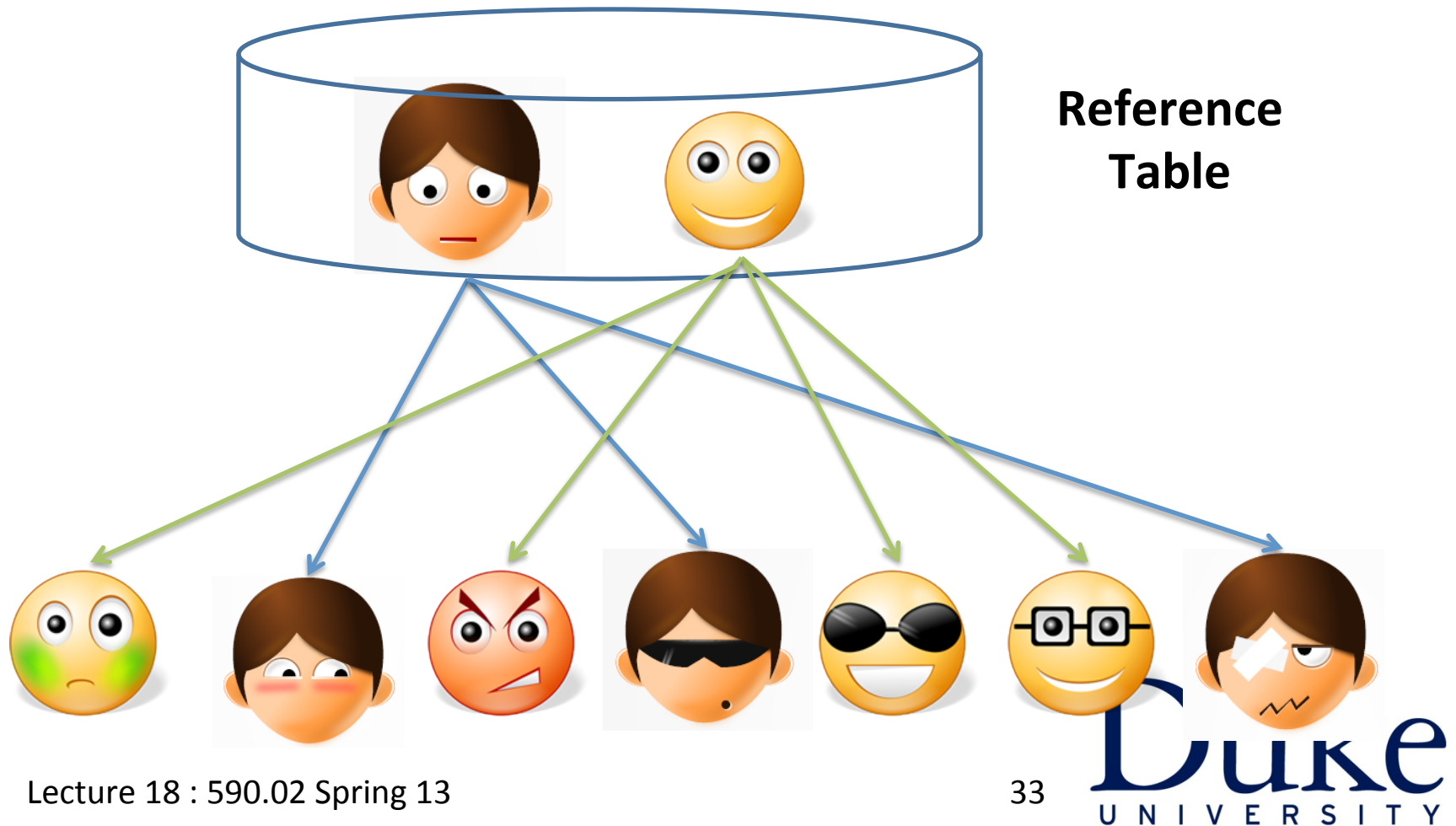
# Record Linkage Problem Statement

- Link records that match across databases



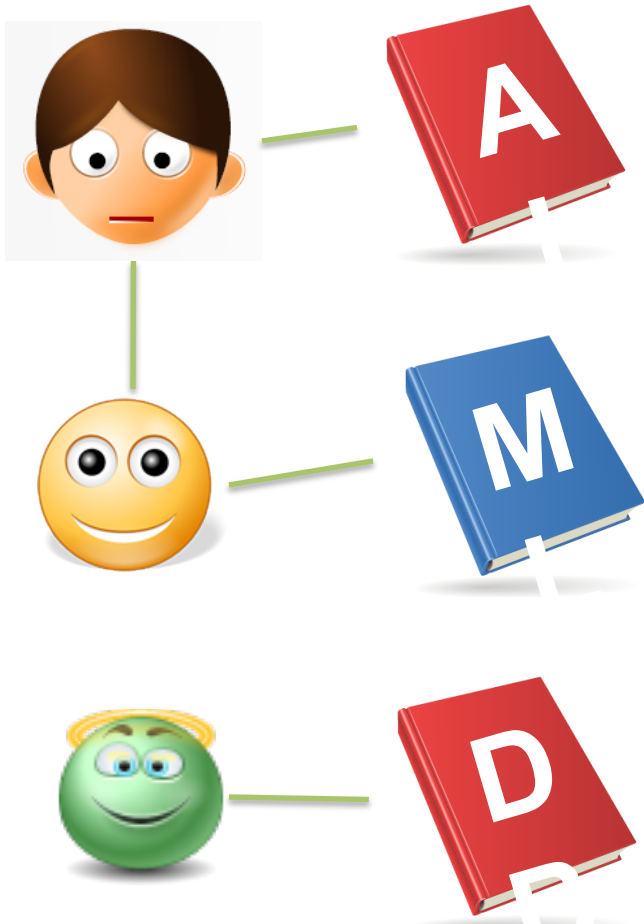
# Reference Matching Problem

- Match noisy records to clean records in a reference table

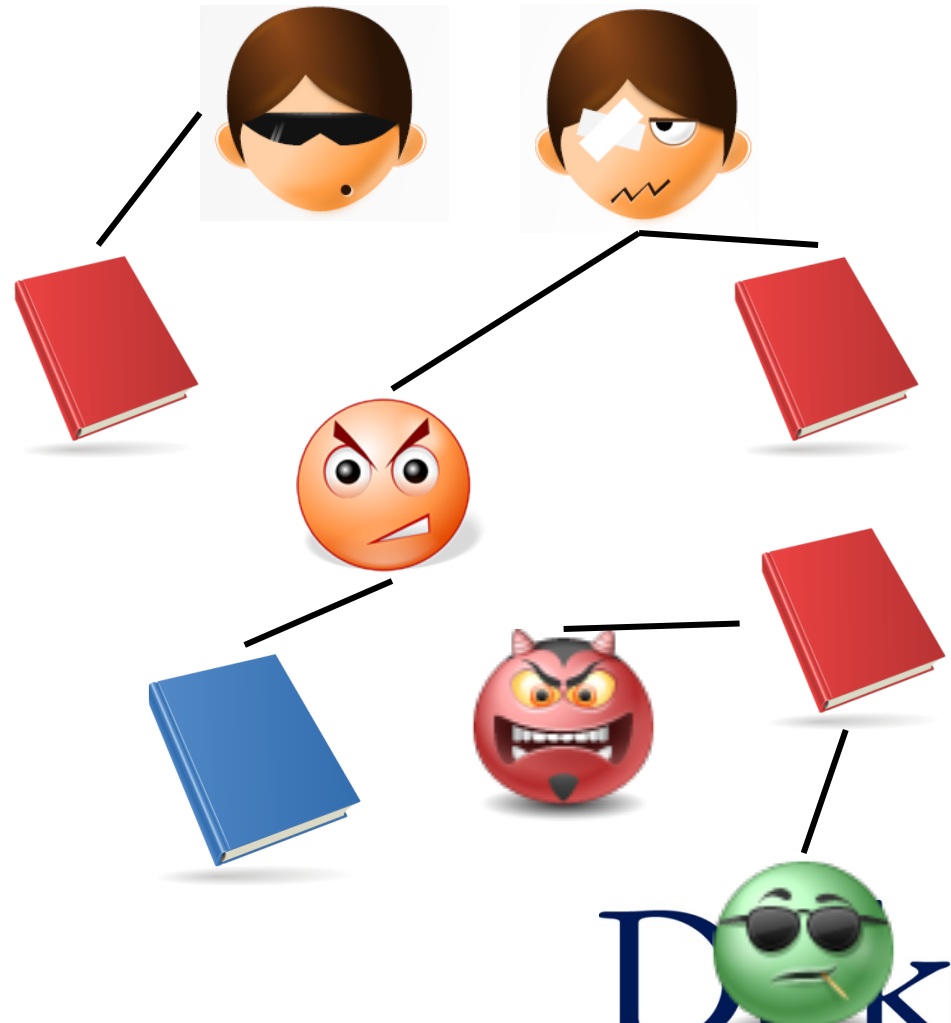


# Relational/Multi-Entity Problem Statement

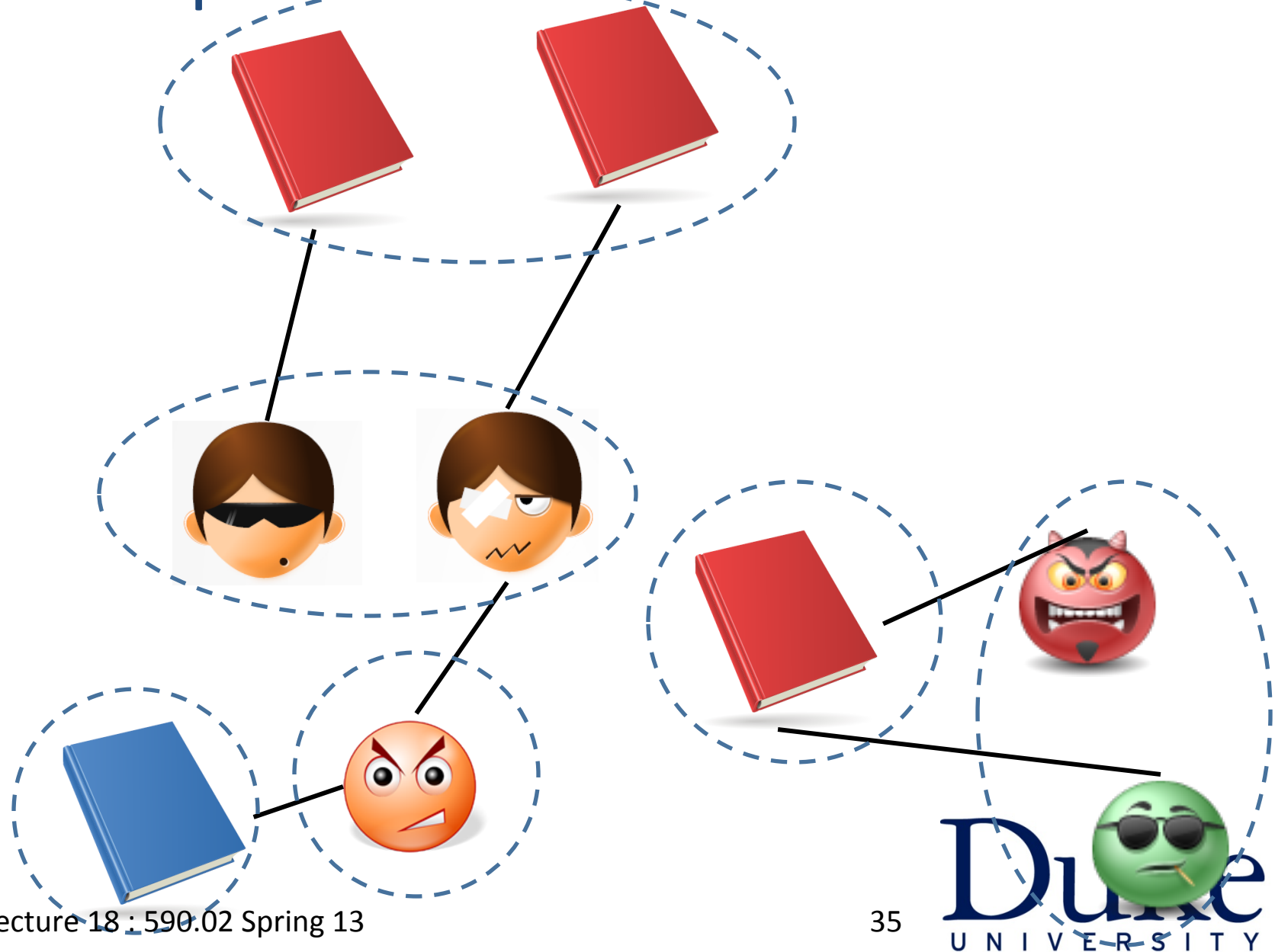
## Real World



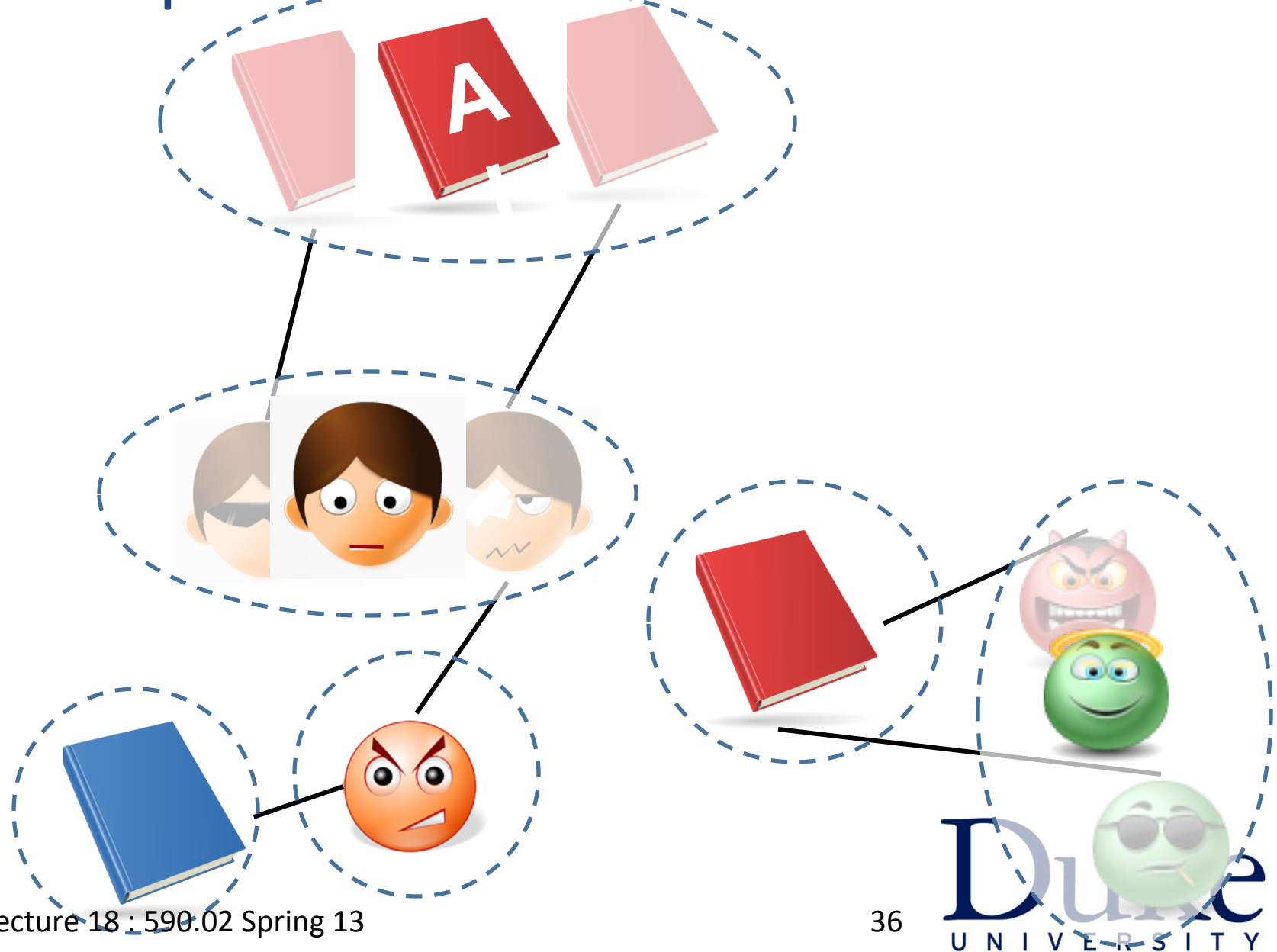
## Digital World



# Deduplication-Problem Statement

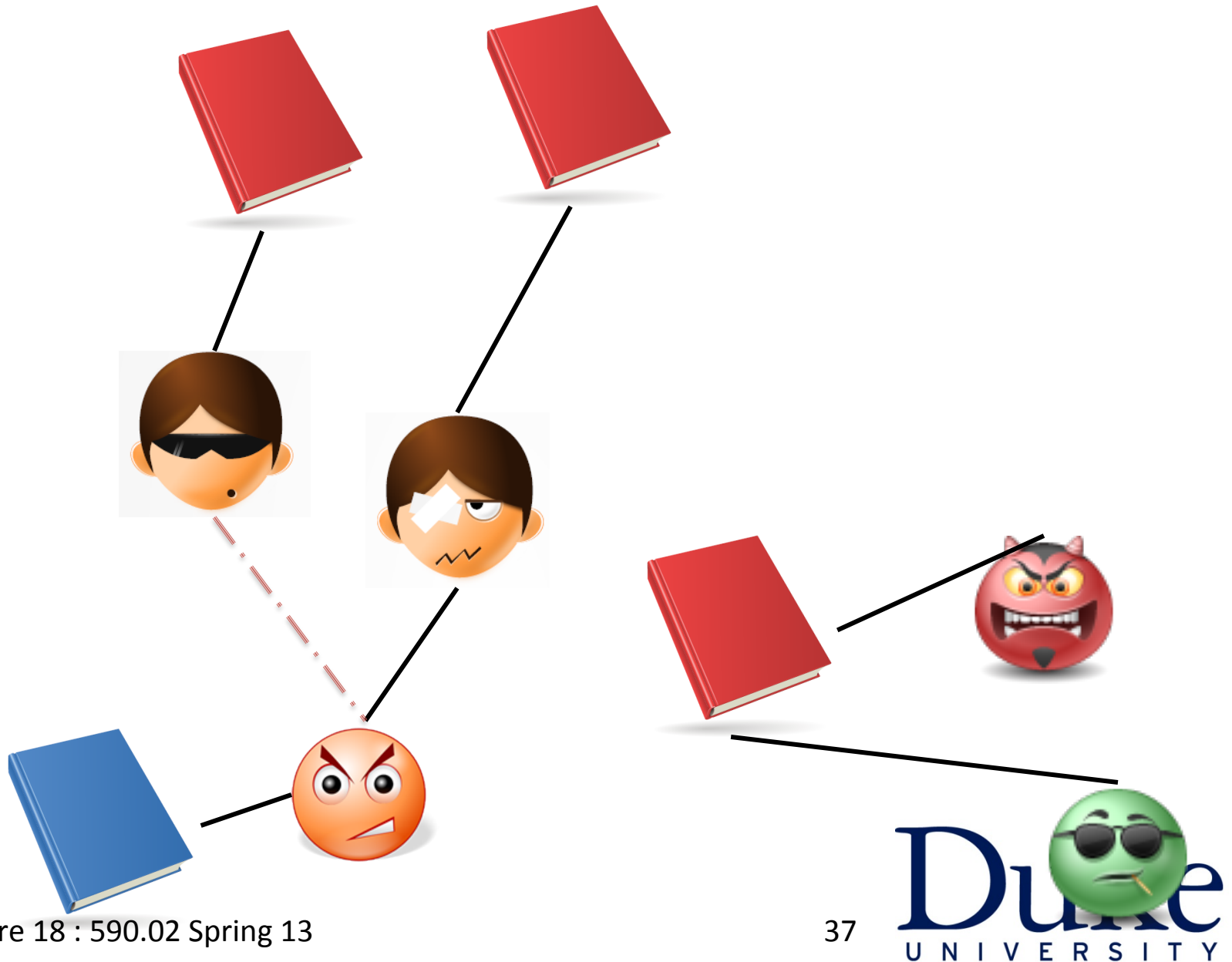


# Deduplication with Canonicalization

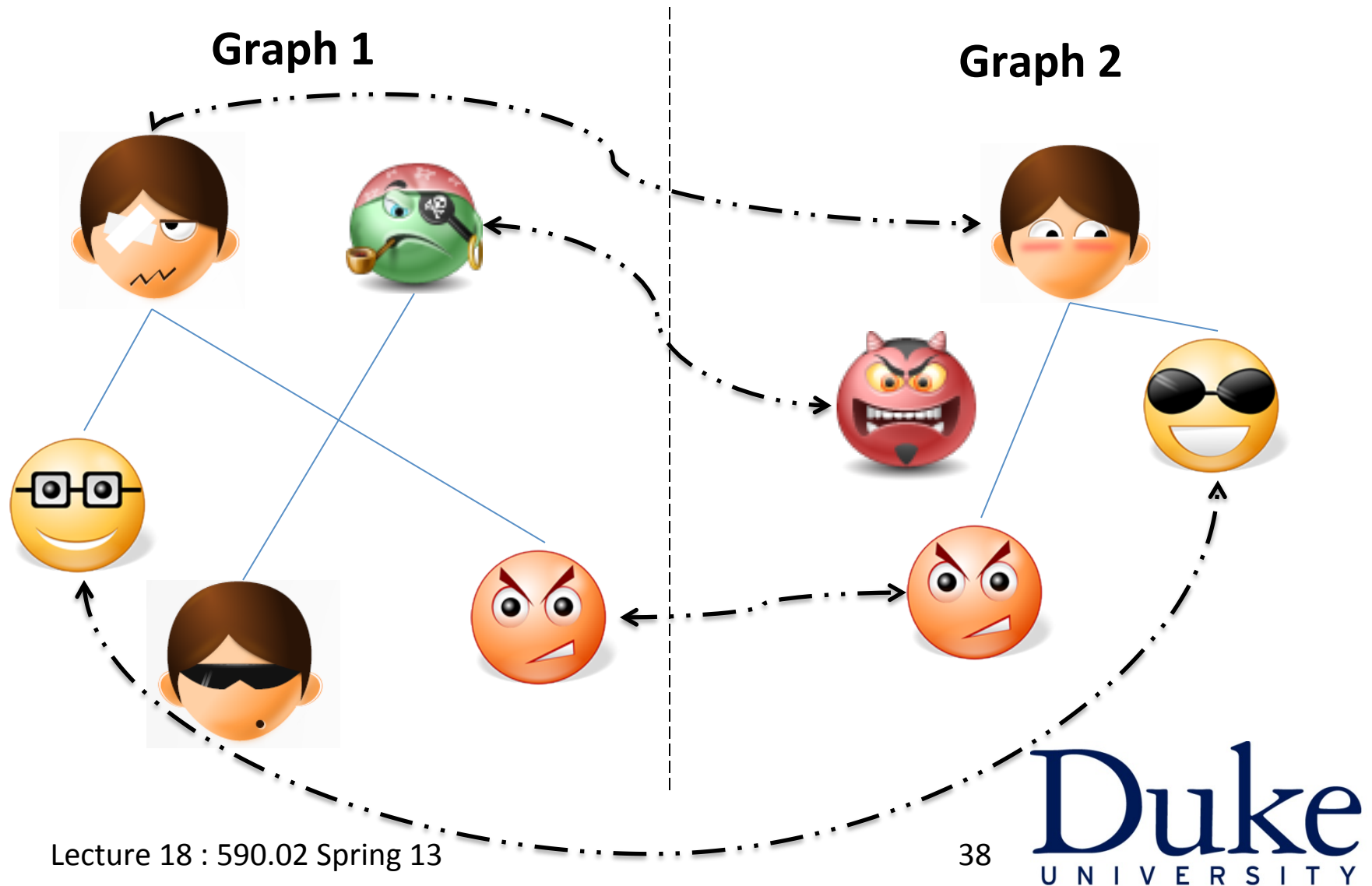




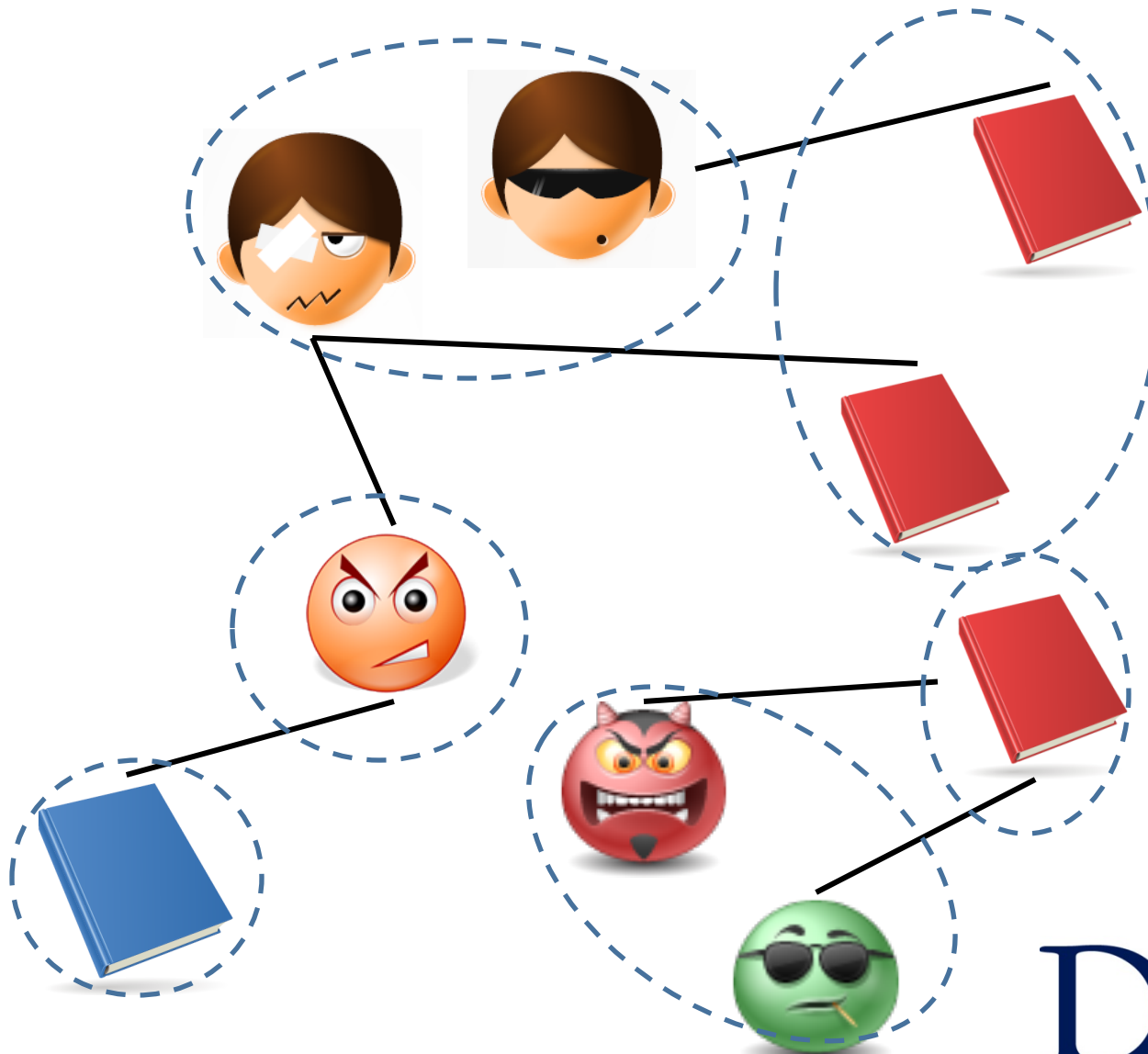
# Link Prediction Problem Statement



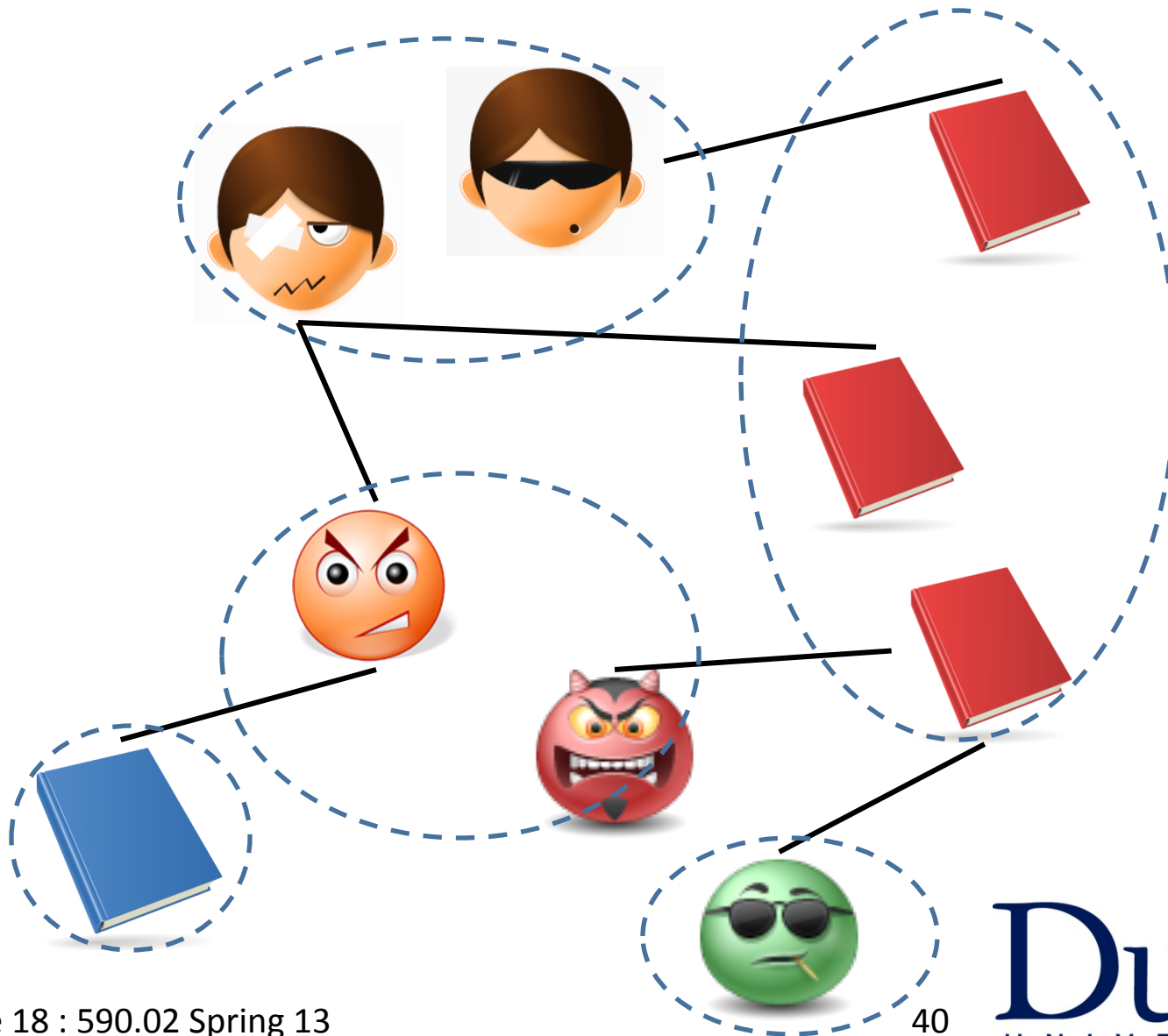
# Graph Alignment (& motif search)



# Relationships are crucial



# Relationships are crucial

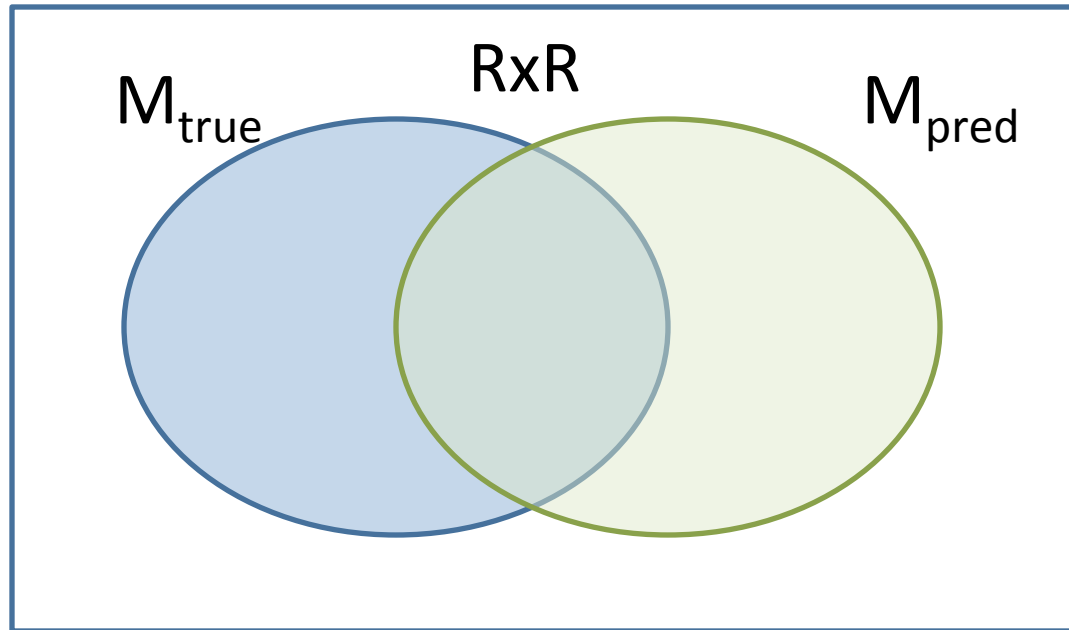


# Notation

- $R$ : set of records / mentions (typed)
- $H$ : set of relations / hyperedges (typed)
- $M$ : set of *matches* (record pairs that correspond to same entity )
- $N$ : set of *non-matches* (record pairs corresponding to different entities)
- $E$ : set of entities
- $L$ : set of links
  
- True ( $M_{true}, N_{true}, E_{true}, L_{true}$ ): according to real world  
vs Predicted ( $M_{pred}, N_{pred}, E_{pred}, L_{pred}$ ): by algorithm

# Relationship between $M_{\text{true}}$ and $M_{\text{pred}}$

- $M_{\text{true}}$  (SameAs , Equivalence)
- $M_{\text{pred}}$  (Similar representations and similar attributes)

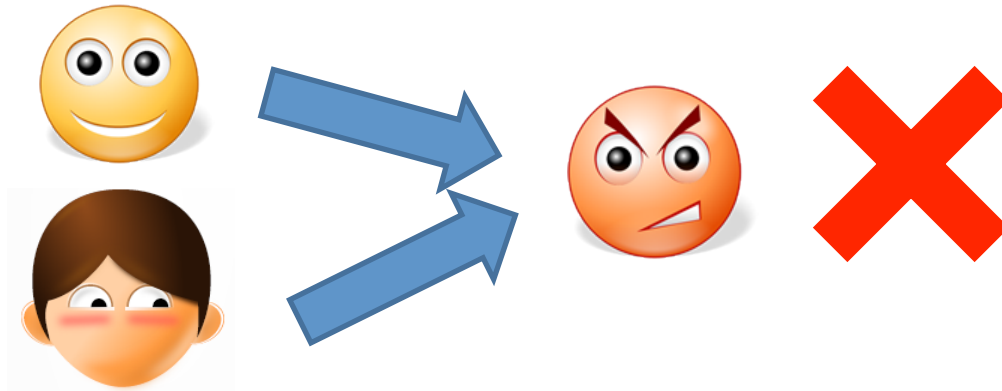


# Metrics

- Pairwise metrics
  - Precision/Recall, F1
  - # of predicted matching pairs
- Cluster level metrics
  - purity, completeness, complexity
  - Precision/Recall/F1: Cluster-level, closest cluster, MUC,  $B^3$ , Rand Index
  - Generalized merge distance [Menestrina et al, PVLDB10]
- Little work that evaluates correct prediction of links

# Typical Assumptions Made

- *Each record/mention is associated with a single real world entity.*



- *In record linkage, no duplicates in the same source*
- *If two records/mentions are identical, then they are true matches*

$$(\text{🕶️}, \text{🕶️}) \in M_{\text{true}}$$



# ER versus Classification

Finding matches vs non-matches is a classification problem

- Imbalanced: typically  $O(R)$  matches,  $O(R^2)$  non-matches
- Instances are pairs of records. Pairs are not IID

  $\in M_{\text{true}}$

AND

  $\in M_{\text{true}}$



  $\in M_{\text{true}}$

# ER vs (Multi-relational) Clustering

Computing entities from records is a clustering problem

- In typical clustering algorithms (k-means, LDA, etc.) *number of clusters is a constant or sub linear in  $R$ .*
- In ER: *number of clusters is linear in  $R$ , and average cluster size is a constant. Significant fraction of clusters are singletons.*

# Outline

- Introduction
- Problem Formulation
- Algorithms for Single Entity ER
  - Computing Pairwise Match scores
  - Blocking: Efficiently Identifying of Near-Duplicates
  - Correlation Clustering: Enforcing Transitivity Constraints
- Algorithms for Relational & Multi-Entity ER

# Matching Features

- For two references x and y, compute a “comparison” vector of *similarity scores* of component attribute.
  - [ 1<sup>st</sup>-author-match-score,  
paper-match-score,  
venue-match-score,  
year-match-score, .... ]
- Similarity scores
  - Boolean (match or not-match)
  - Real values based on distance functions

# Summary of Matching Features

Handle  
Typographical  
errors

- Equality on a boolean predicate
- Edit distance
  - Levenstein, Smith-Waterman, Affine
- Set similarity
  - Jaccard, Dice
- Vector Based
  - Cosine similarity, TFIDF

Good for Text like  
reviews/ tweets

Good for Names

- Alignment-based or Two-tiered
  - Jaro-Winkler, Soft-TFIDF, Monge-Elkan
- Phonetic Similarity
  - Soundex
- Translation-based
- Numeric distance between values
- Domain-specific

Useful for  
abbreviations,  
alternate names.

- Useful packages:
  - SecondString, <http://secondstring.sourceforge.net/>
  - Simmetrics: <http://sourceforge.net/projects/simmetrics/>
  - LingPipe, <http://alias-i.com/lingpipe/index.html>

# Pairwise Match Score

Problem: Given a vector of component-wise similarities for a pair of records (x,y), compute  $P(x \text{ and } y \text{ match})$ .

Solutions:

1. Weighted sum or average of component-wise similarity scores.  
Threshold determines match or non-match.
  - $0.5 * 1^{\text{st}}\text{-author-match-score} + 0.2 * \text{venue-match-score} + 0.3 * \text{paper-match-score}$ .
  - Hard to pick weights.
    - Match on last name match *more predictive* than login name.
    - Match on “Smith” *less predictive* than match on “Machanavajjhala”.
  - Hard to tune a threshold.

# Pairwise Match Score

Problem: Given a vector of component-wise similarities for a pair of records (x,y), compute  $P(x \text{ and } y \text{ match})$ .

Solutions:

1. Weighted sum or average of component-wise similarity scores.  
Threshold determines match or non-match.
2. Formulate rules about what constitutes a match.
  - $(1^{\text{st}}\text{-author-match-score} > 0.7 \text{ AND venue-match-score} > 0.8)$   
OR  $(\text{paper-match-score} > 0.9 \text{ AND venue-match-score} > 0.9)$
  - Manually formulating the right set of rules is hard.



# Basic ML Approach

- $r = (x, y)$  is record pair,  $\gamma$  is comparison vector,  $M$  matches,  $U$  non-matches

- Decision rule 
$$R = \frac{P(\gamma \mid r \in M)}{P(\gamma \mid r \in U)}$$

$$R > t \Rightarrow r \rightarrow \text{Match}$$

$$R \leq t \Rightarrow r \rightarrow \text{Non - Match}$$

# Fellegi & Sunter Model [FS, Science '69]

- Record pair:  $r = (x, y)$  in  $A \times B$
- $\gamma$  is comparison vector
  - E.g.,  $\gamma = [\text{"Is } x.\text{name} = y.\text{name?"}, \text{"Is } x.\text{address} = y.\text{address?"} \dots]$
  - Assume binary vector for simplicity
- $M$  : set of matching pairs of records
- $U$  : set of non-matching pairs of records

# Fellegi & Sunter Model [FS, Science '69]

- $r = (x, y)$  is record pair,  $\gamma$  is comparison vector,  $M$  matches,  $U$  non-matches

- Decision rule 
$$R = \frac{P(\gamma \mid r \in M)}{P(\gamma \mid r \in U)}$$

$$R \geq t_l \Rightarrow r \rightarrow \text{Match}$$

$$t_l < R < t_u \Rightarrow r \rightarrow \text{Potential Match}$$

$$R \leq t_u \Rightarrow r \rightarrow \text{Non - Match}$$

- Naïve Bayes Assumption: 
$$P(\gamma \mid r \in M) = \prod_i P(\gamma_i \mid r \in M)$$

# Quality of a Decision Rule ( $t_l, t_u$ )

- Consider three sets of record pairs (as classified by the decision rule)
  - A1:  $(x,y)$  is a match
  - A2:  $(x,y)$  is a possible match (uncertain)
  - A3:  $(x,y)$  is a non-match
- Given some distribution over  $A \times B$ ,

$$\begin{aligned} m(\gamma) &= P(\gamma | r \in M) \\ &= \sum_{(x,y) \in M} P(\gamma(x,y)) P(x,y|M) \end{aligned}$$

$$\begin{aligned} u(\gamma) &= P(\gamma | r \in U) \\ &= \sum_{(x,y) \in U} P(\gamma(x,y)) P(x,y|U) \end{aligned}$$

# Error due to a Linkage Rule

- Type I Error:  $(x,y)$  in  $U$ , but the linkage rule calls it a match

$$P(A1|U) = \sum_{\gamma \in \Gamma} u(\gamma)P(A1|\gamma)$$

- Type II Error:  $(x,y)$  in  $M$ , but the linkage rule calls it a non-match

$$P(A3|M) = \sum_{\gamma \in \Gamma} m(\gamma)P(A3|\gamma)$$

# Optimal Linkage Rule

- Let  $\mu$  and  $\lambda$  be bounds on the type I and type II error, resp.
- For any decision rule  $L = (t_l, t_u)$ , let  $A1(L)$ ,  $A2(L)$  and  $A3(L)$  denote the sets of matches, possible matches and non-matches.

# Optimal Linkage Rule

- $L^* = (t_l^*, t_u^*)$  is an optimal decision rule for comparison space  $\Gamma$  with error bounds  $\mu$  and  $\lambda$ , if
  - $L^*$  meets the type I and type II requirements

$$P(A1(L^*)|U) = \mu, \quad P(A3(L^*)|M) = \lambda$$

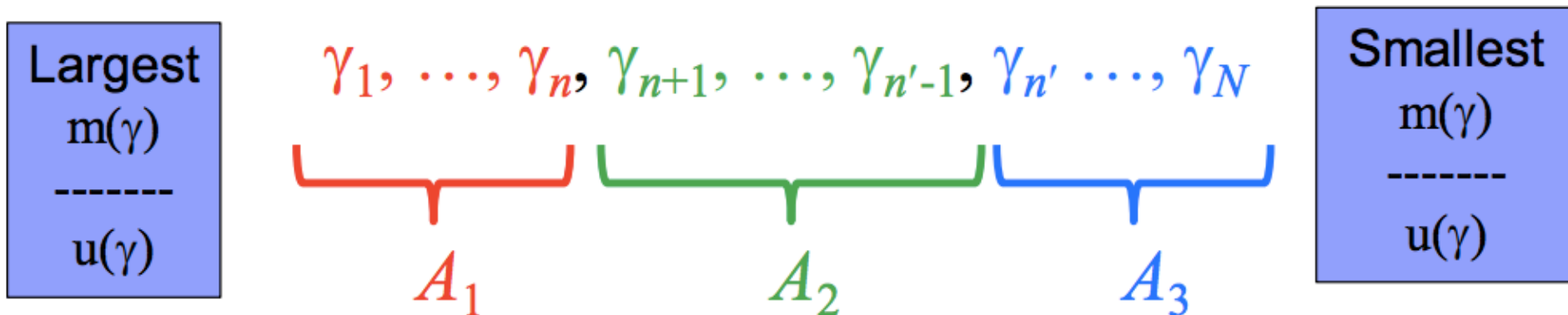
- $L^*$  has the least conditional probabilities of *not making a decision*.  
That is for all other decision rules  $L$  (with error bounds  $\mu$  and  $\lambda$ ),

$$\begin{aligned} P(A2(L^*)|U) &\leq P(A2(L)|U), \\ P(A2(L^*)|M) &\leq P(A2(L)|M) \end{aligned}$$



# Finding the Optimal Linkage Rule

- Suppose there are  $N$  comparison vectors
- Sort them in decreasing order of  $m(\gamma) / u(\gamma)$



- Set  $A_1$  to be the first  $n$  vectors, and  $A_3$  to be the last  $N - n'$  vectors such that:

$$\mu = \sum_{i=1}^n u(\gamma_i), \quad \lambda = \sum_{i=1}^n m(\gamma_i)$$

# Using Fellegi Sunter in Practice

- $\Gamma$  is usually high dimensional (computing  $m(\gamma)$  and  $u(\gamma)$  is inefficient)
  - Use conditional independence of features in  $\gamma$  given match or non-match
  - Naïve Bayes assumption
- Computing  $P(\gamma \mid r \in M)$  requires some knowledge of matches.
  - Supervised learning (assume a training set is provided)
  - EM-based techniques can be used to learn the parameters jointly while identifying matches.

# Supervised Approaches

- Supervised machine learning algorithms
  - Decision trees
    - [Cochinwala et al, IS01]
  - Support vector machines
    - [Bilenko & Mooney, KDD03]; [Christen, KDD08]
  - Ensembles of classifiers
    - [Chen et al., SIGMOD09]
  - Conditional Random Fields (CRF)
    - [Gupta & Sarawagi, VLDB09]
- Issues:
  - **Training set generation**
  - Imbalanced classes – many more negatives than positives (even after eliminating obvious non-matches ... using *Blocking*)
  - Misclassification cost

# Creating a Training Set is a key issue

- Constructing a training set is hard – since most pairs of records are “easy non-matches”.
  - 100 records from 100 cities.
  - Only  $10^6$  pairs out of total  $10^8$  (1%) come from the same city
- Some pairs are hard to judge even by humans
  - Inherently ambiguous
    - E.g., Paris Hilton (person or business)
  - Missing attributes
    - Starbucks, Toronto vs Starbucks, Queen Street ,Toronto

# Avoiding Training Set Generation

- Unsupervised / Semi-supervised Techniques
  - EM based techniques to learn parameters
    - [Winkler '06, Herzog et al '07]
  - Generative Models
    - [Ravikumar & Cohen, UAI04]
- Active Learning
  - Committee of Classifiers
    - [Sarawagi et al KDD '00, Tajeda et al IS '01]
  - Provably optimizing precision/recall
    - [Arasu et al SIGMOD '10, Bellare et al KDD '12]
  - Crowdsourcing
    - [Wang et al VLDB '12, Marcus et al VLDB '12, ...]

# Next few classes

- Introduction
- Problem Formulation
- Algorithms for Single Entity ER
  - Computing Pairwise Match scores
  - Blocking: Efficiently Identifying of Near-Duplicates
  - Correlation Clustering: Enforcing Transitivity Constraints
- Algorithms for Relational & Multi-Entity ER