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.
- …

Duke UNIVERSITY

Ironically, Entity Resolution has many duplicate names

- Household matching
- Doubles
- Duplicate detection
- Mixed and split citation problem
- Record linkage
- Object identification
- Match
- Deduplication
- Fuzzy match
- Object consolidation
- Entity resolution
- Entity clustering
- Approximate match
- Identity uncertainty
- Reference reconciliation
- Merge/purge
- Hardening soft databases
- Householding
- Reference matching

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

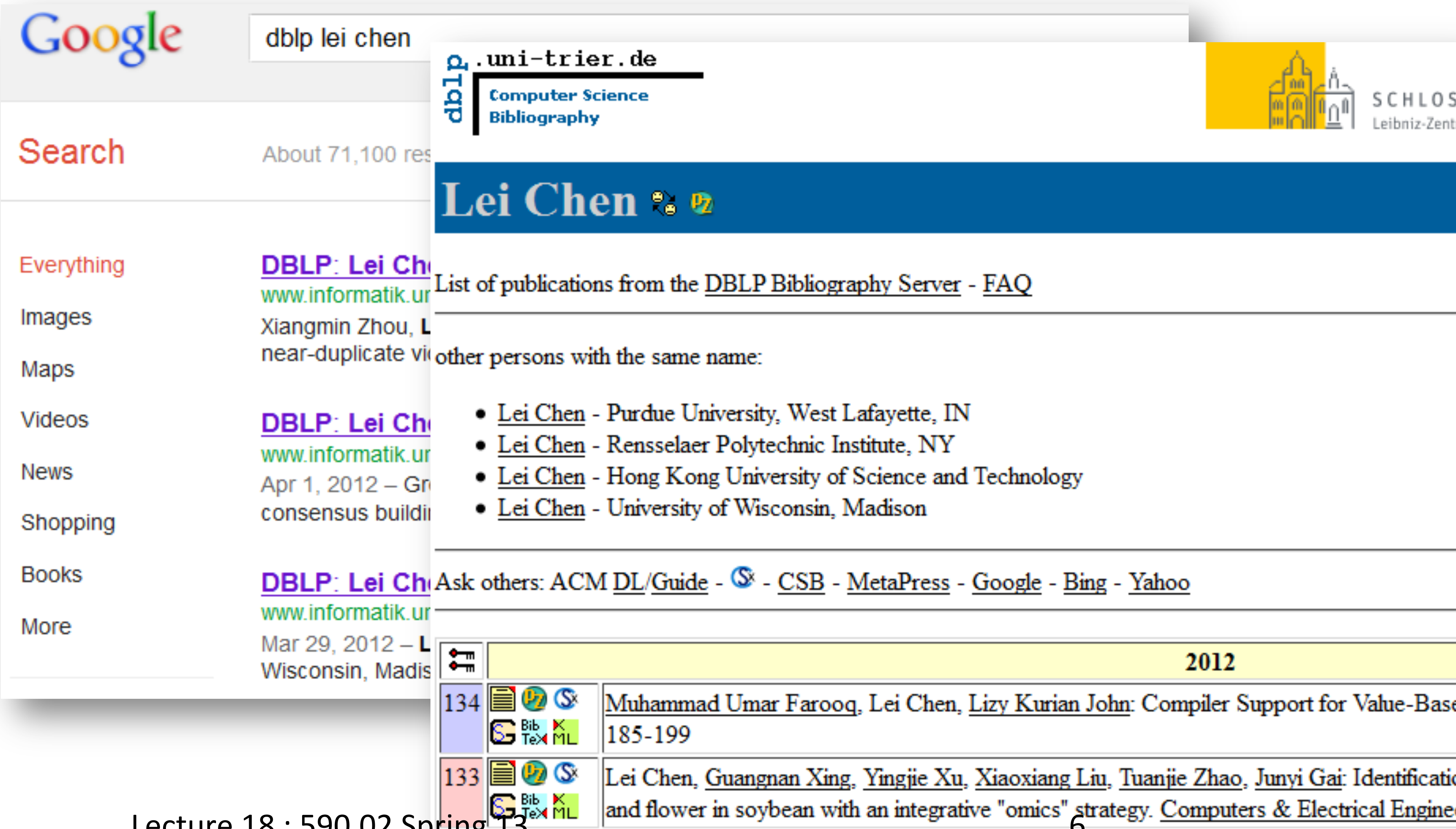
- "Overview of Record Linkage and Current Research Directions", William E Winkler, 2006

- 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).**

Duke UNIVERSITY

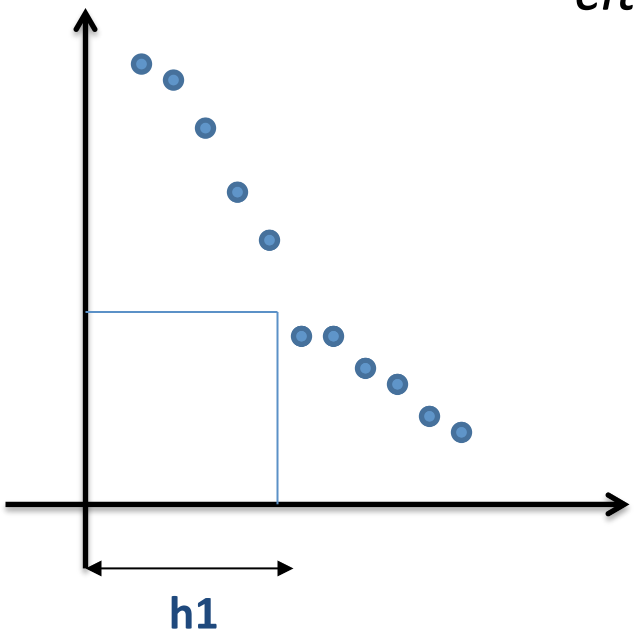
Motivation : Citation

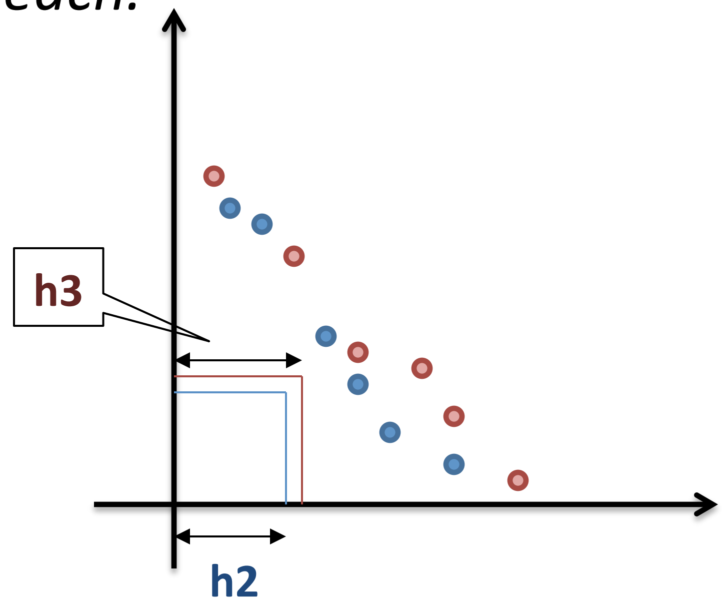
- What is the most recent publication of Lei Chen?



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.*

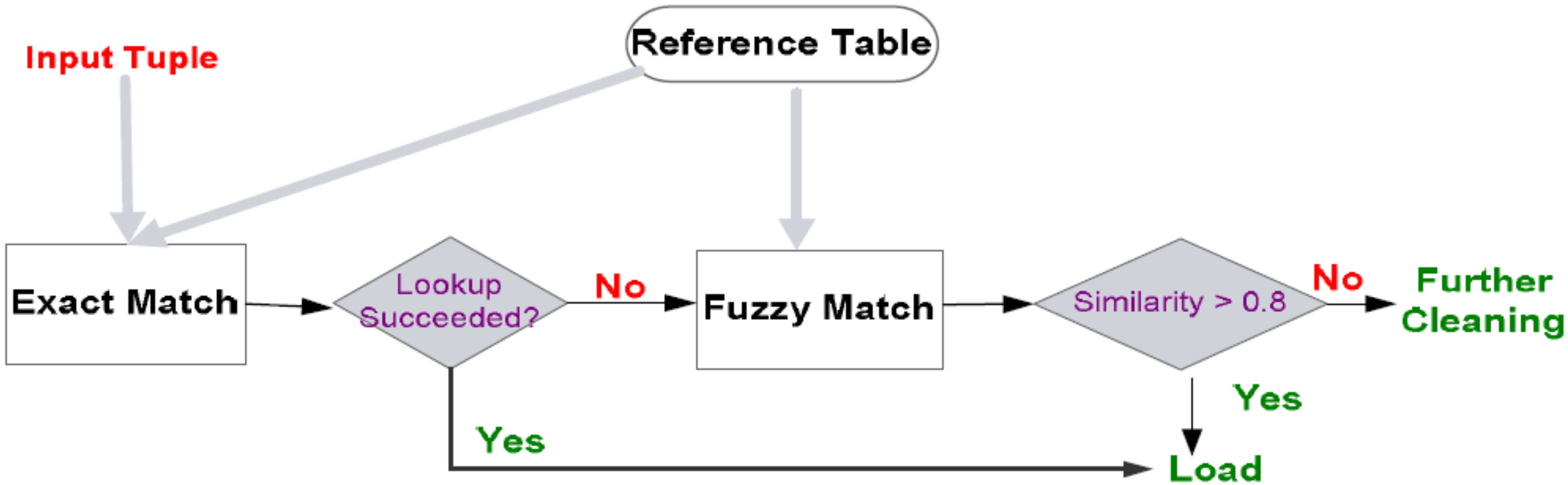




h1 > h2 and h1 > h3

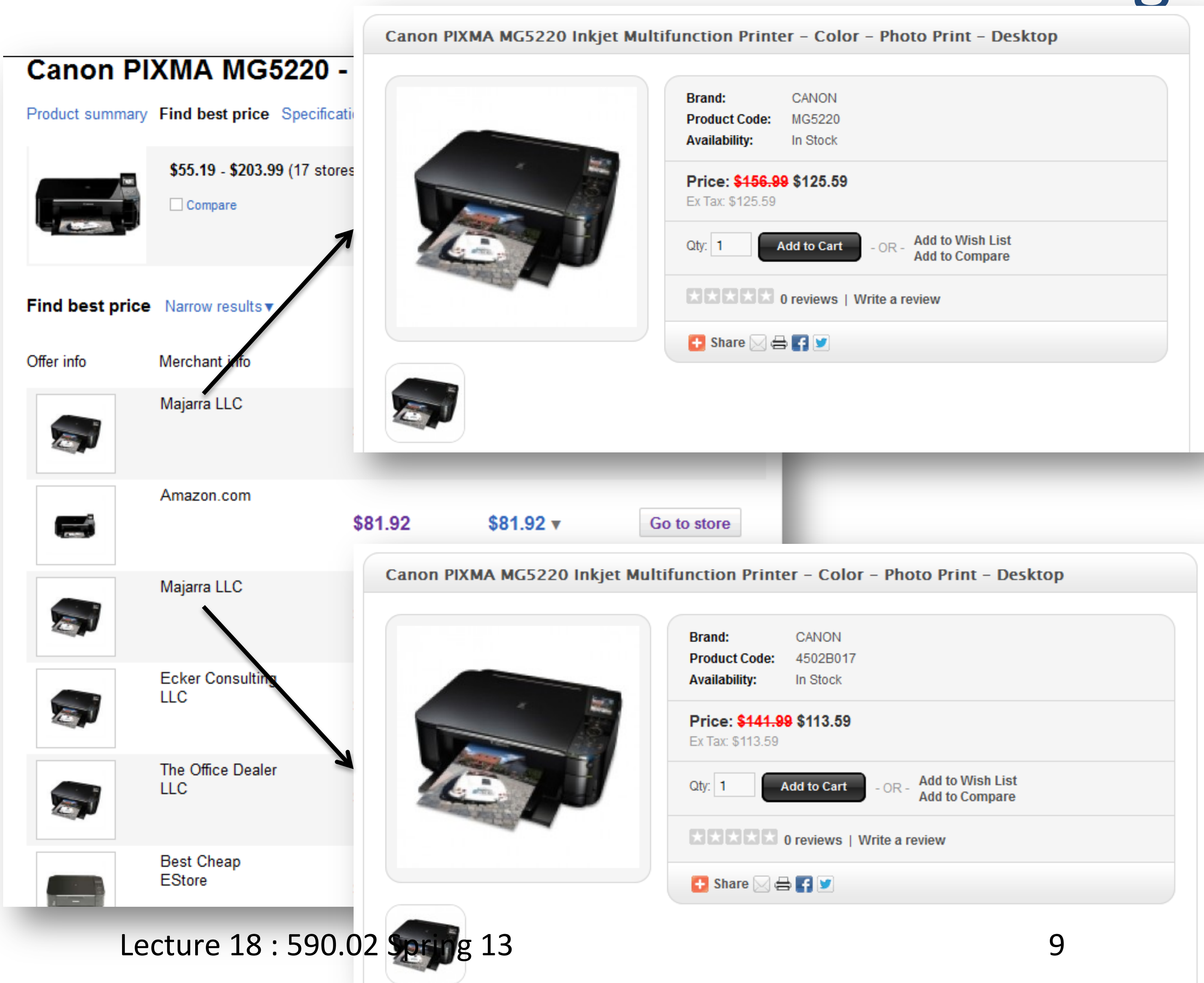
Motivation: Data Cleaning

- [Chaudhuri et al. SIGMOD 2003]

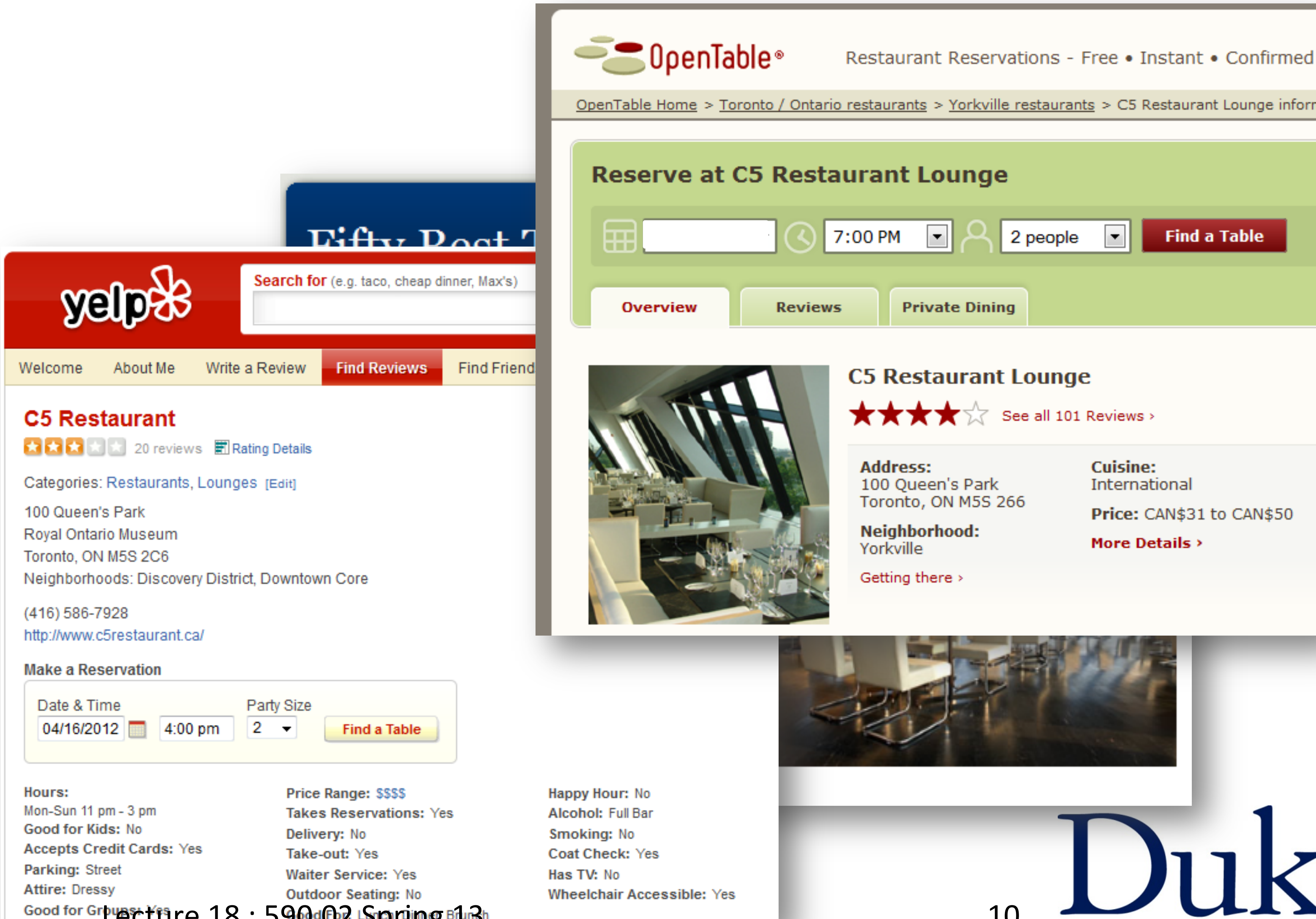


- Reference table contains "clean" records
- Input table has "noisy" records
- Applications
  - Geocoding incoming queries
  - Match new customers to old ones
  - Products

Motivation: Data Cleaning

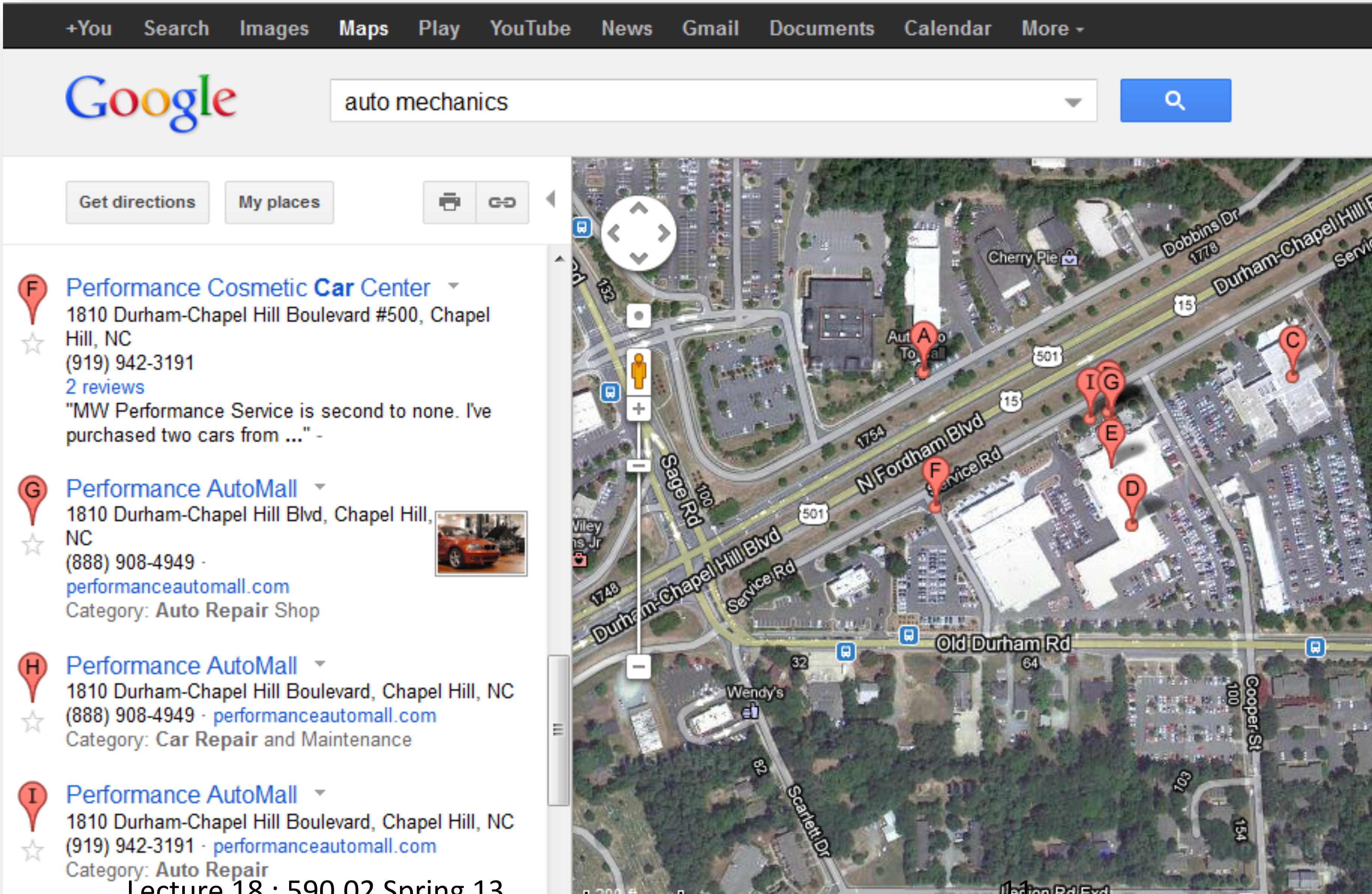


Motivation : Web Search

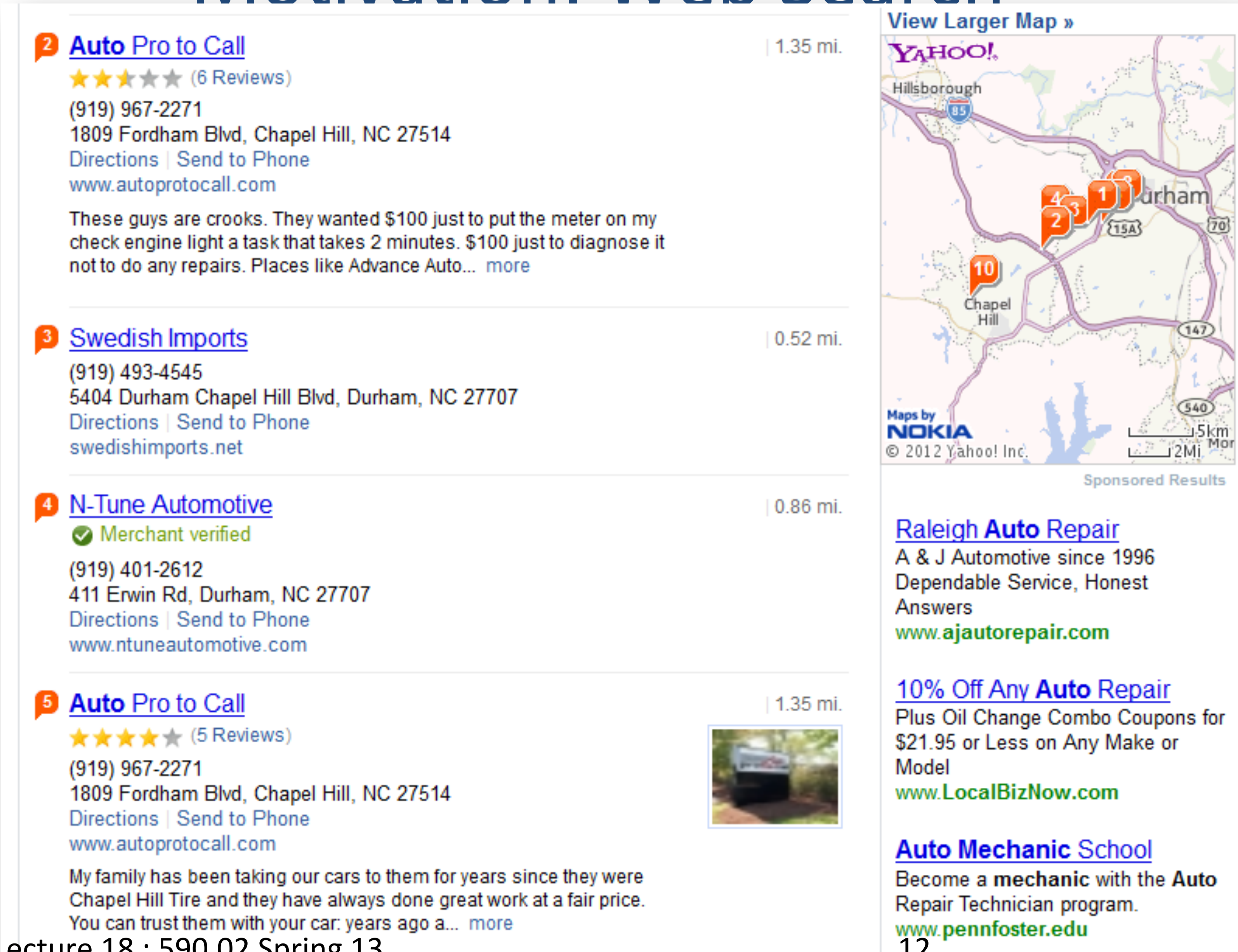


Duke UNIVERSITY

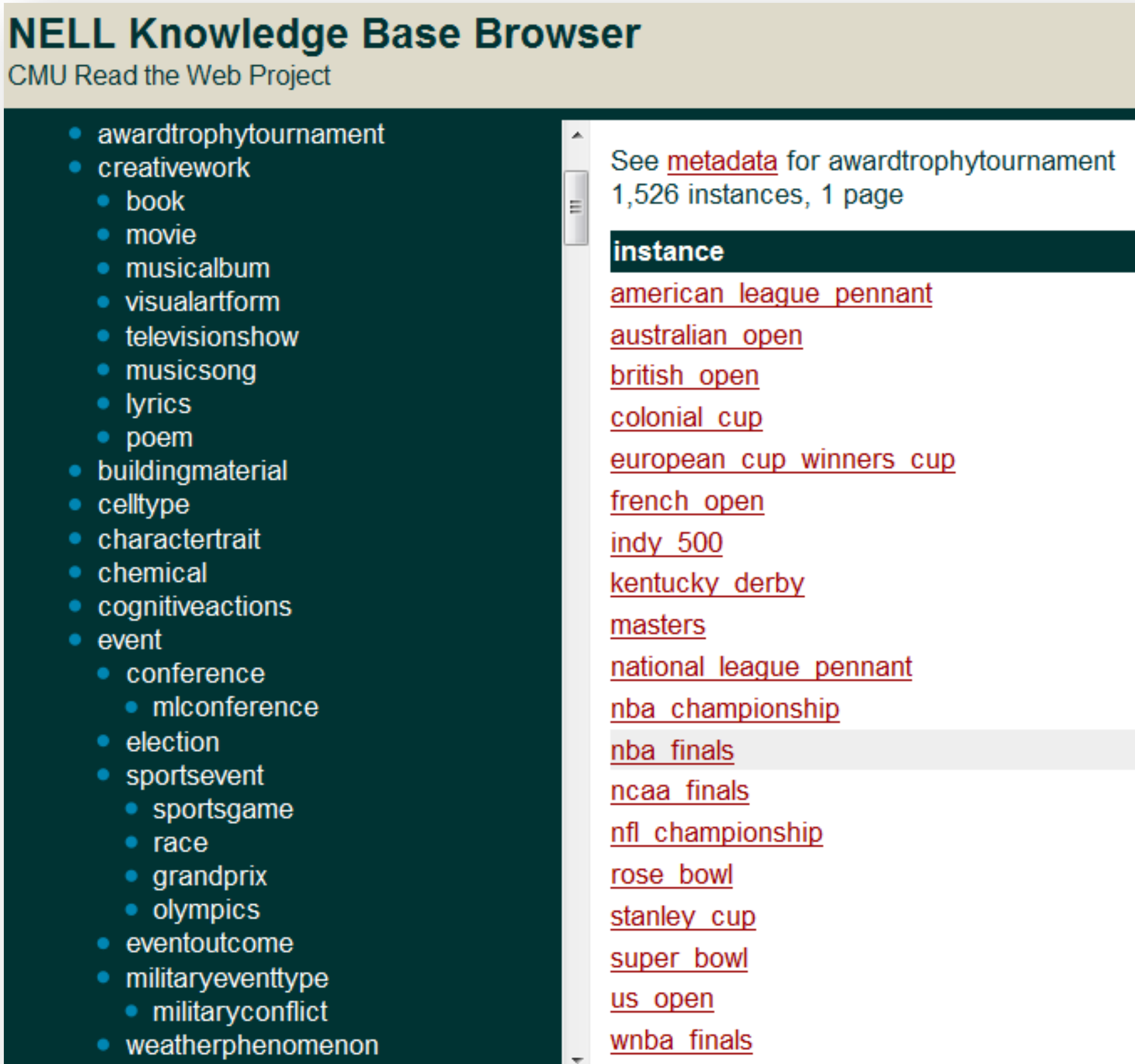
Motivation: Web Search



Motivation: Web Search



Motivation: Machine Reading



ER helps improve information extraction

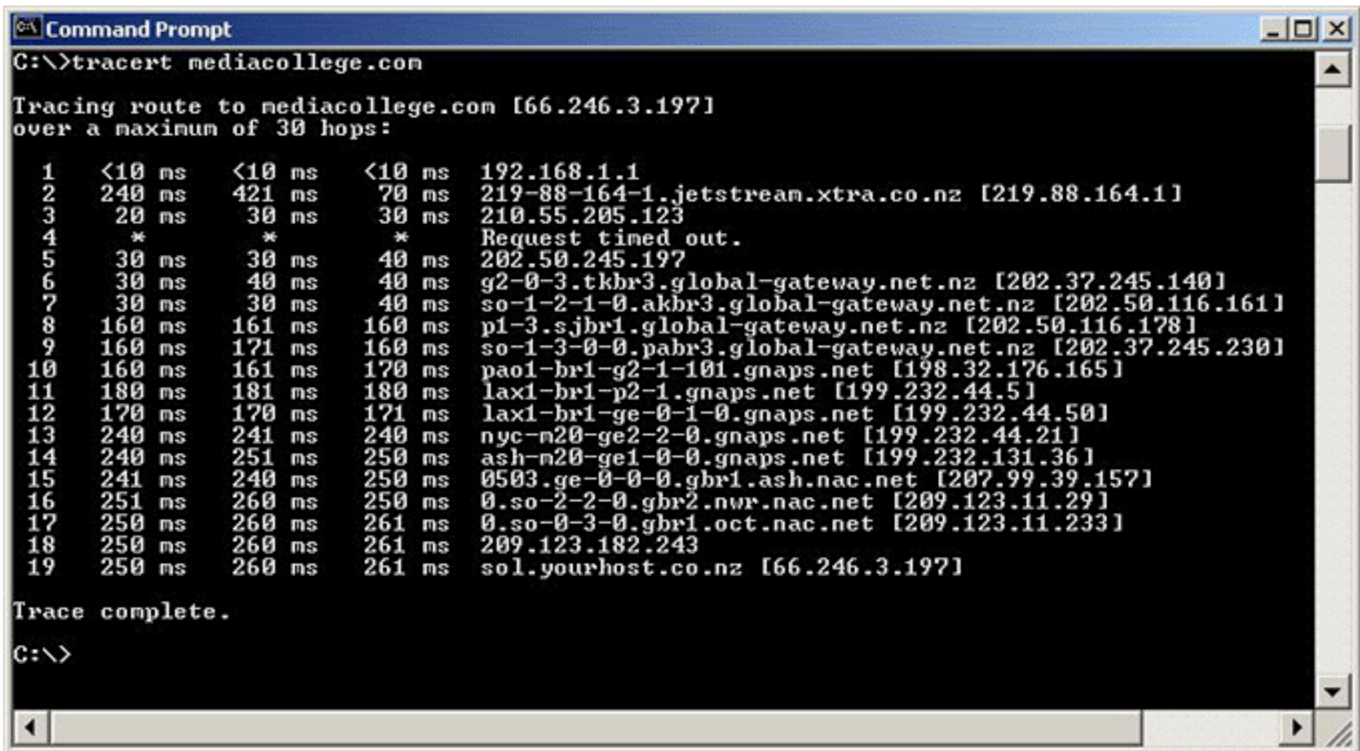
- If we know how to extract from one list, and the same entity appear on another differently formatted list, we can use the overlap for training an extractor on the second list. [Gupta et al VLDB11, Machanavajjhala et al WSDM11]

- *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`



IP Aliasing Problem [Willinger et al. 2009]

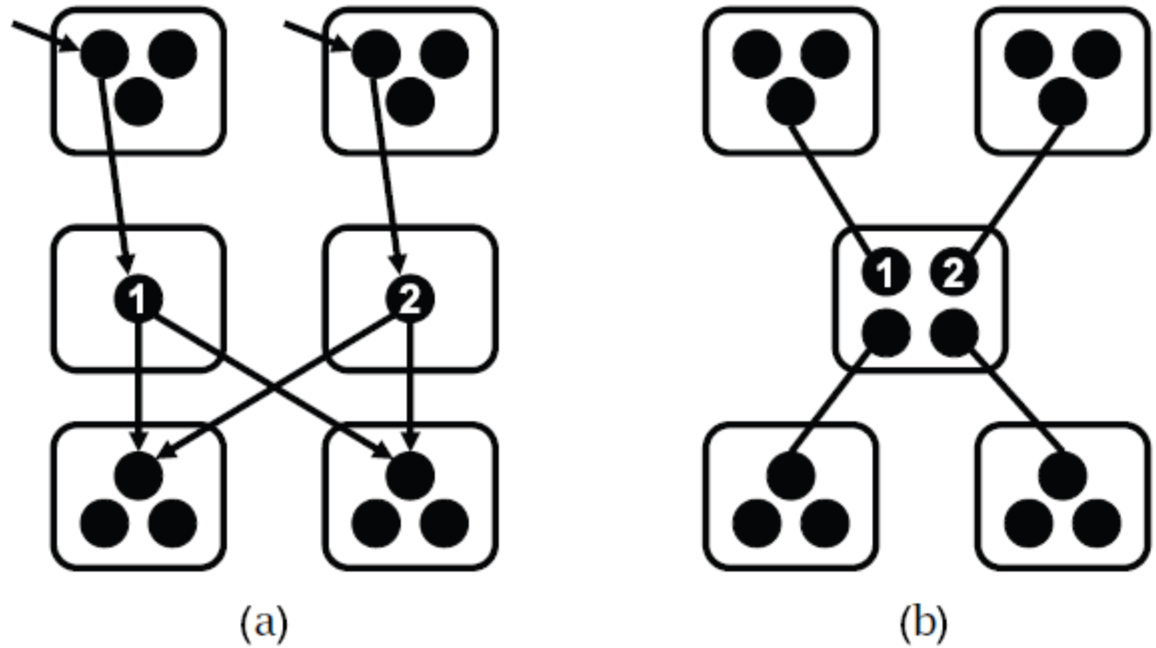


Figure 2. The IP alias resolution problem. Paraphrasing Fig. 4 of [50], `traceroute` does not list routers (boxes) along paths but IP addresses of input interfaces (circles), and alias resolution refers to the correct mapping of interfaces to routers to reveal the actual topology. In the case where interfaces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

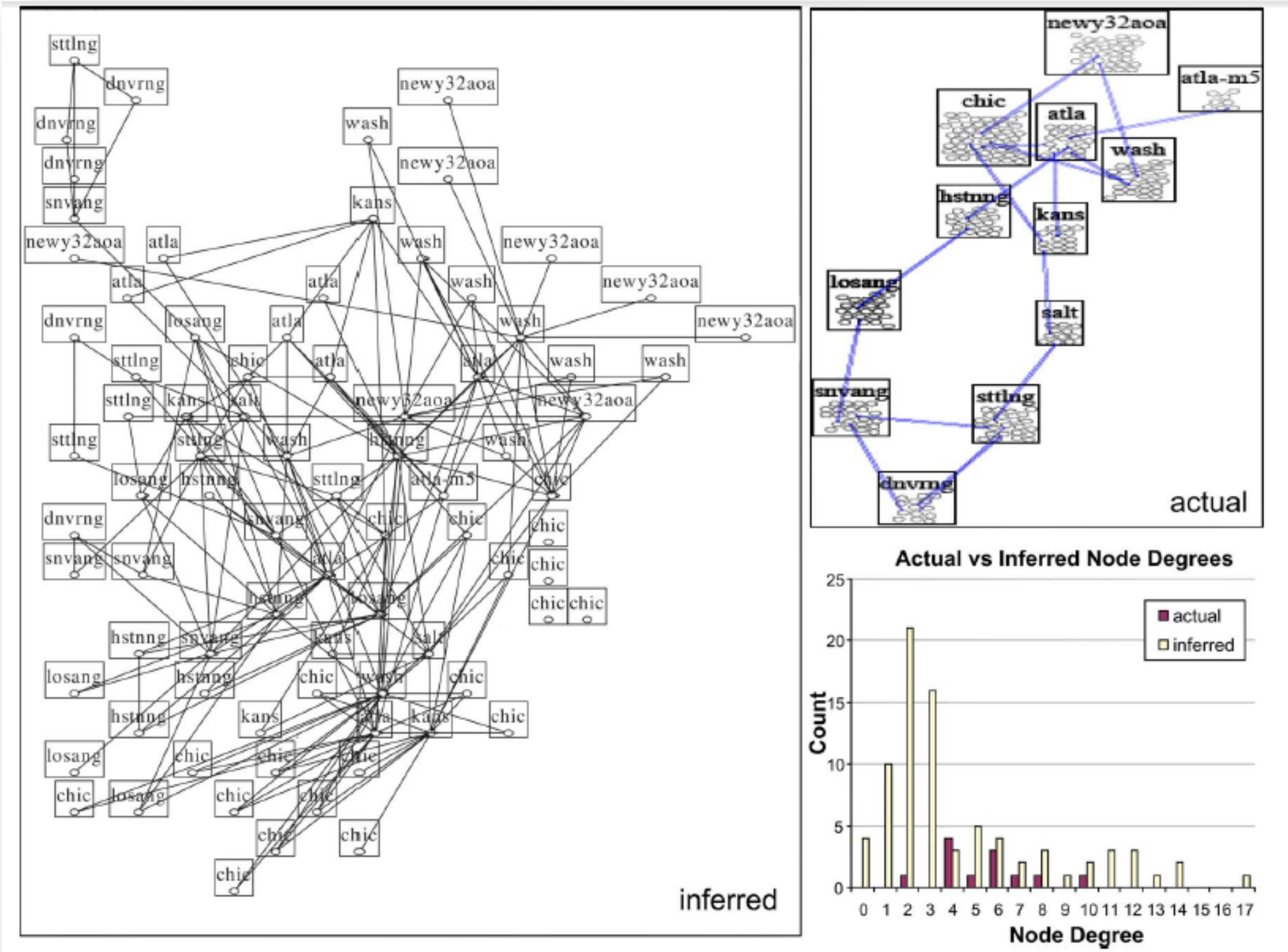


Figure 3. The IP alias resolution problem in practice. This is re-produced from [48] and shows a comparison between the Abilene/Internet2 topology inferred by Rocketfuel (left) and the actual topology (top right). Rectangles represent routers with interior ovals denoting interfaces. The histograms of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

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.
  - …

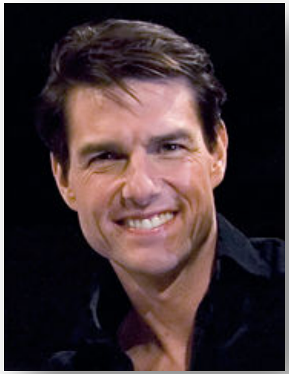
Outline

- 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

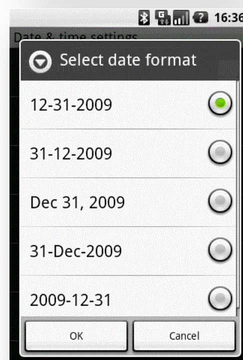
Administrative dates:	set to 0101YY, 010199, 999999
Date of Birth	0101YY, 1506YY, 3006YY, 0107YY, 1507YY, 0101YEAR
Names:	set to spaces, NK, UNKNOWN, or ZZZZ BABY, MALE, FEMALE, TWIN, TRIPLET, INFANT
Other variables:	set to 9, 99, 9999, -1 NK (Not Known) NA (Not applicable) NC (Not coded) U (Unknown)

[Gill et al; Univ of Oxford 2003]

Traditional Challenges in ER

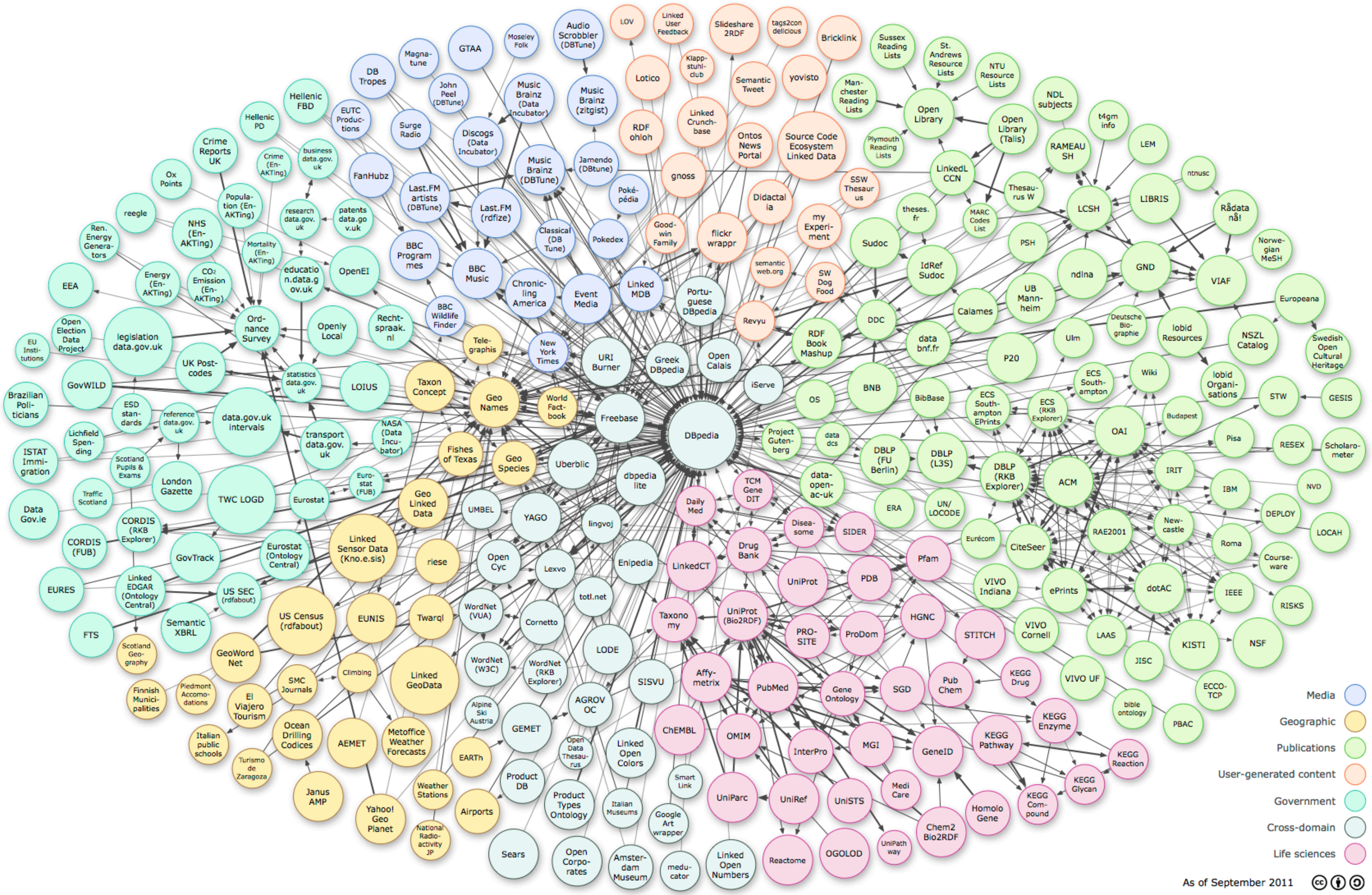
- Name/Attribute ambiguity
- Errors due to data entry
- Missing Values
- Changing Attributes

- Data formatting



- Abbreviations / Data Truncation

Big-Data ER Challenges



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

Duke UNIVERSITY

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]

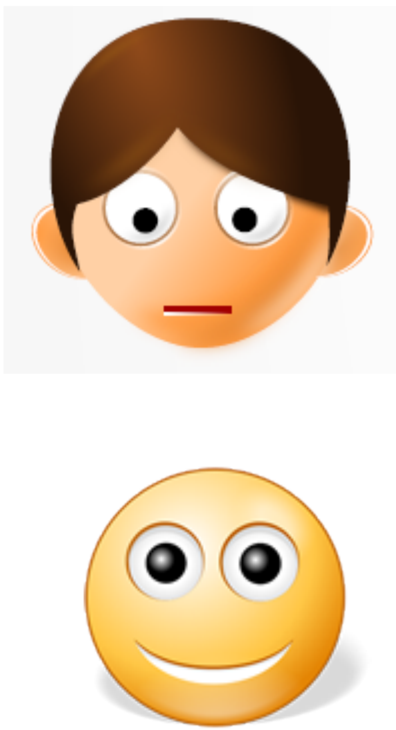
- Tutorials
  - Record Linkage: Similarity measures and Algorithms
    [N. Koudas, S. Sarawagi, D. Srivatsava SIGMOD '06]
  - Data fusion--Resolving data conflicts for integration
    [X. Dong, F. Naumann VLDB '09]
  - Entity Resolution: Theory, Practice and Open Challenges
    http://goo.gl/Ui38o [L. 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

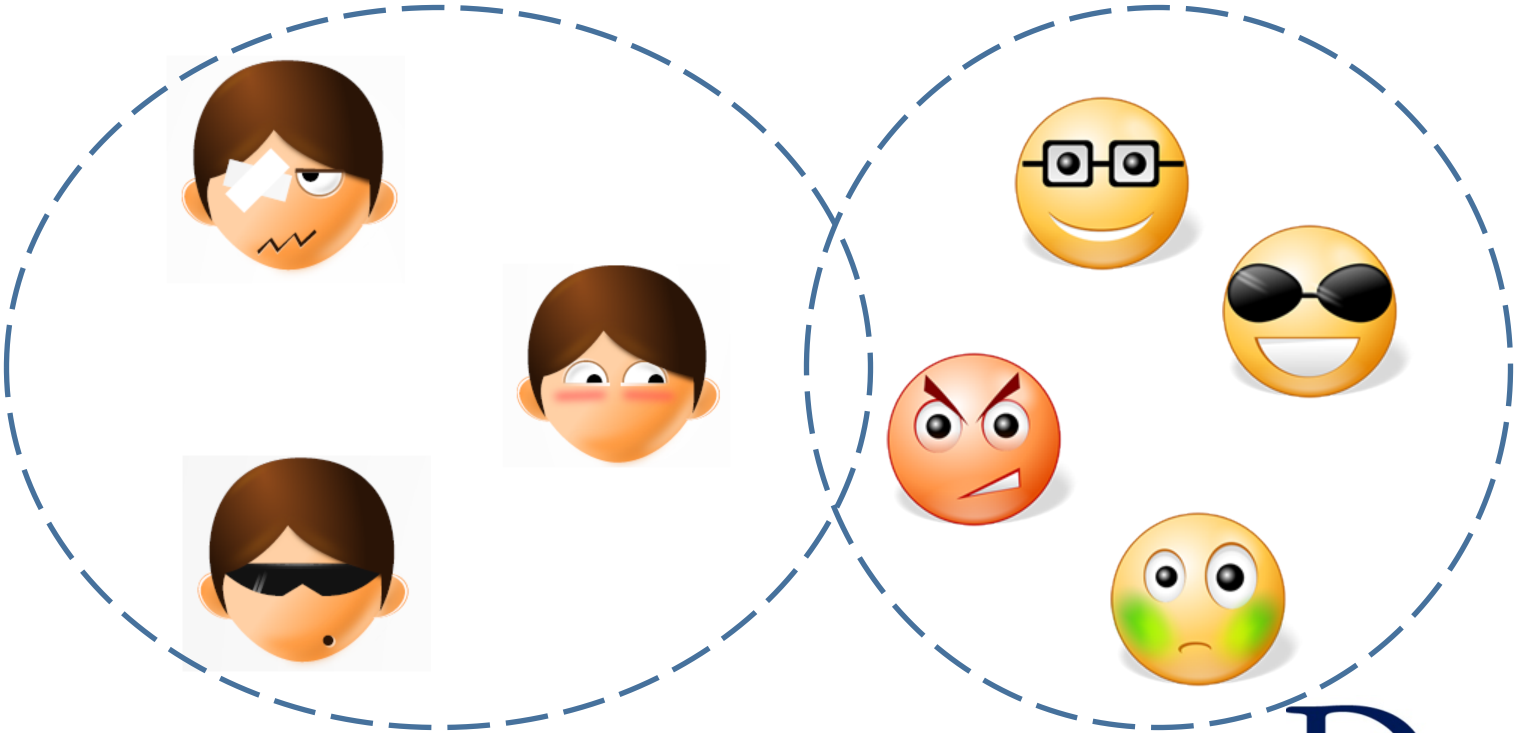
Real World



Digital World

Deduplication Problem Statement

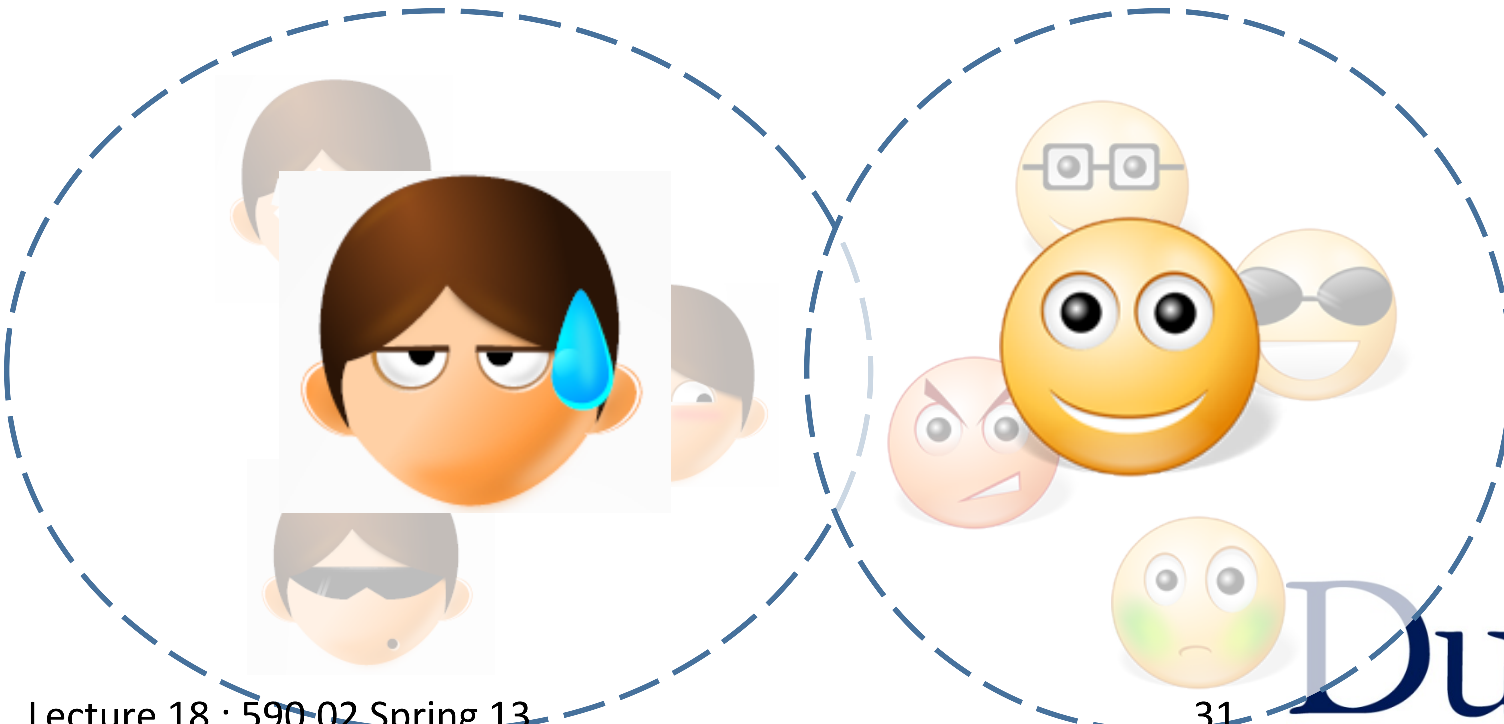
- Cluster the records/mentions that correspond to same entity



Duke UNIVERSITY

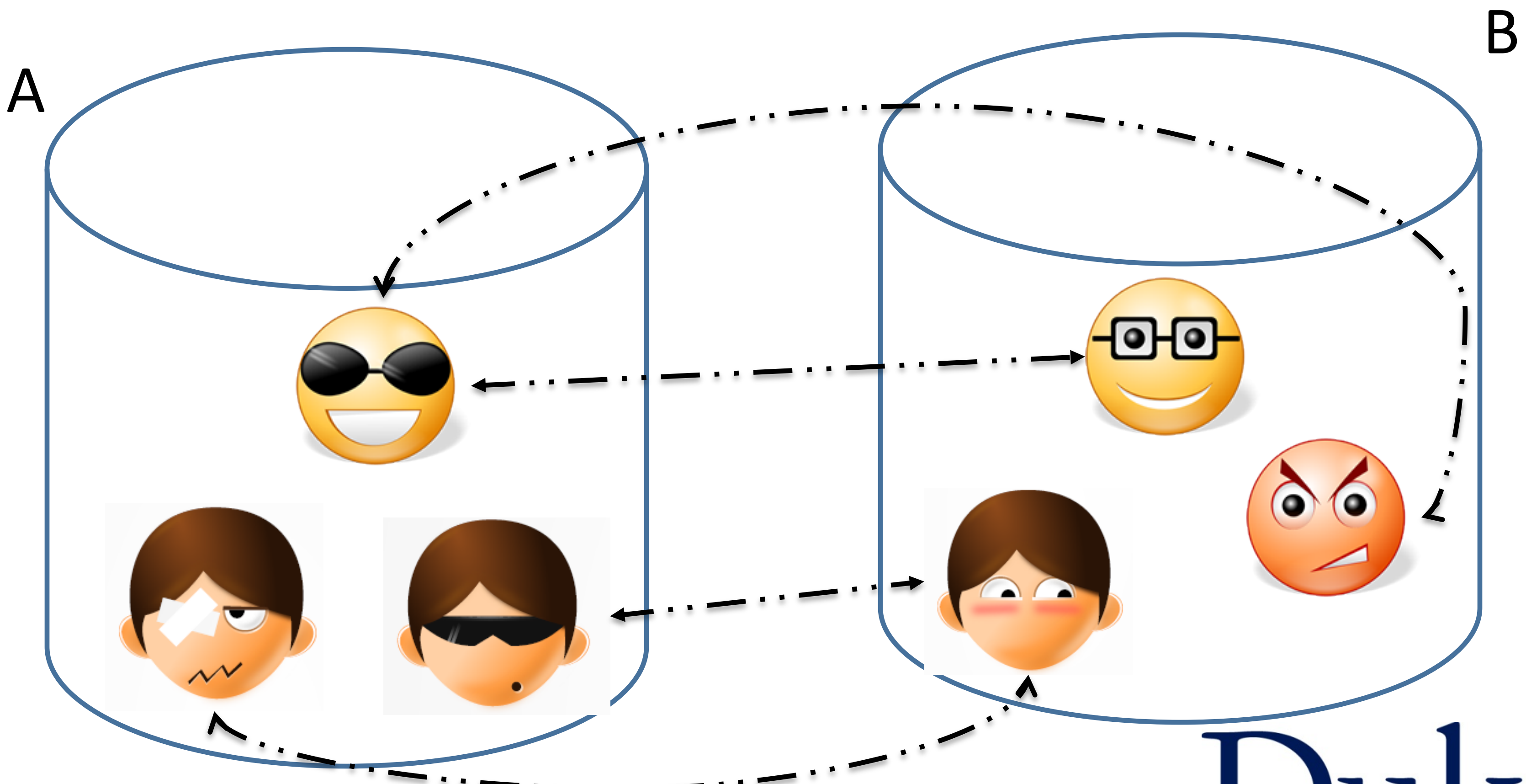
Deduplication Problem Statement

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



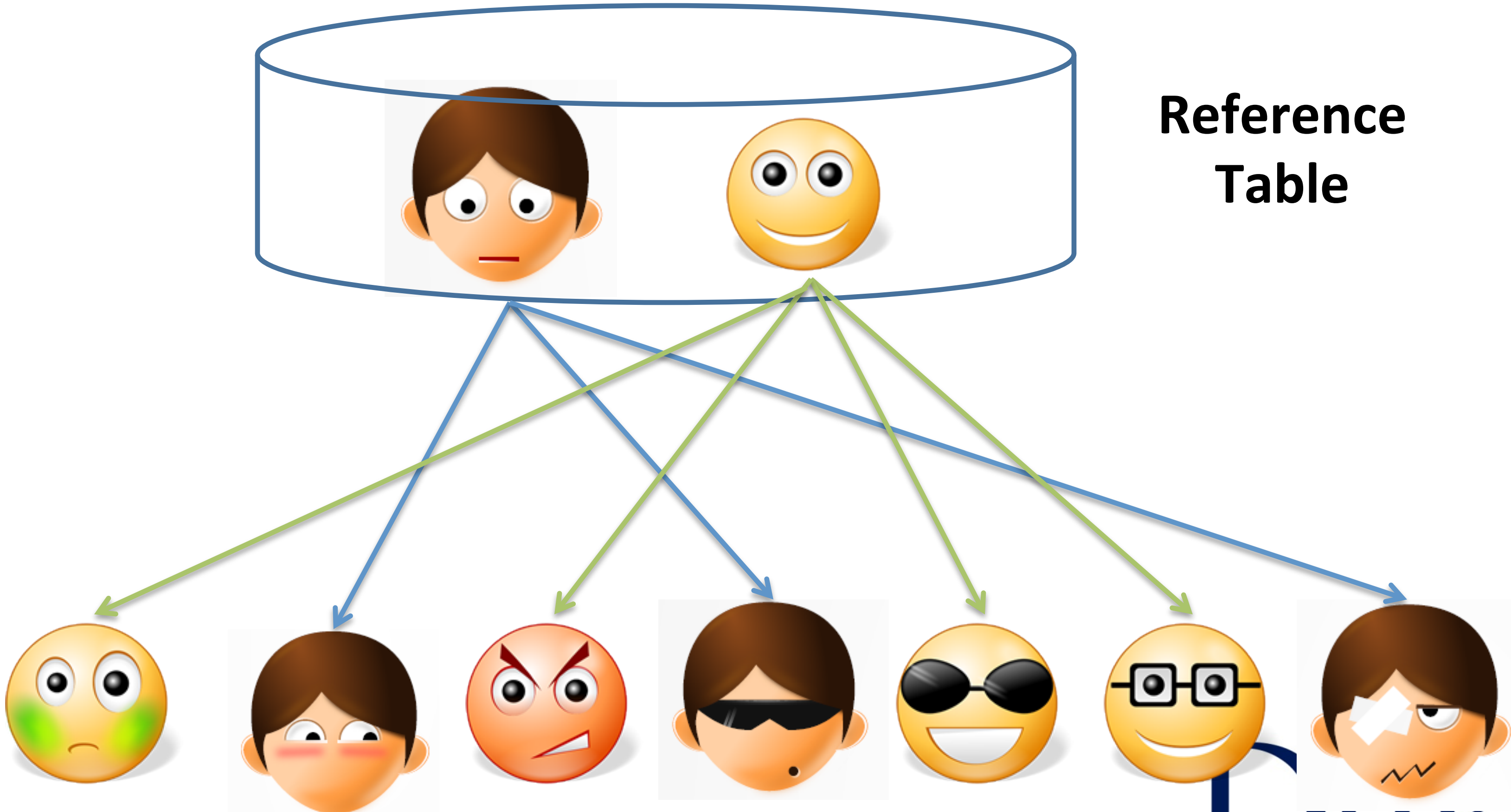
Record Linkage Problem Statement

- Link records that match across databases



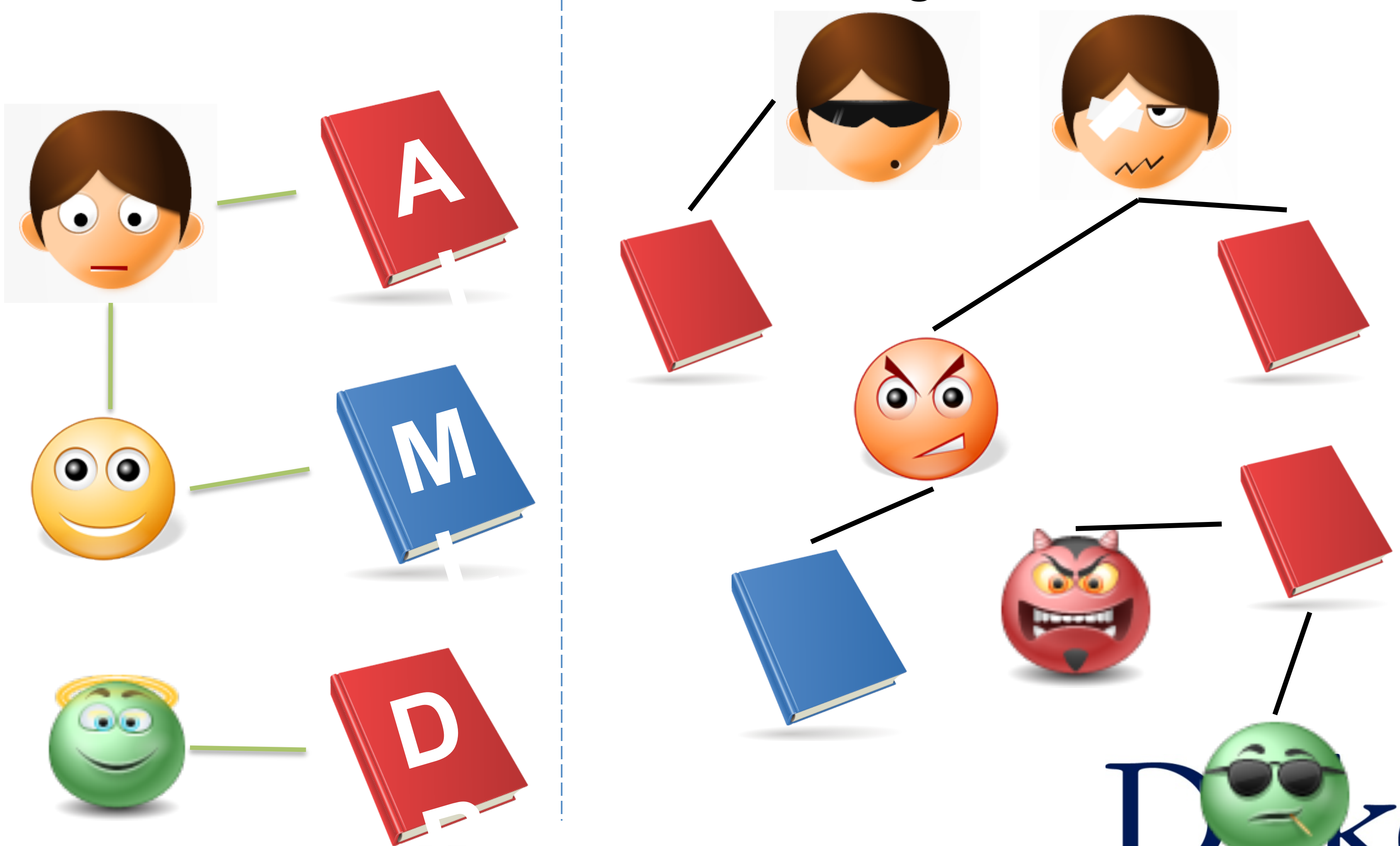
Reference Matching Problem

- Match noisy records to clean records in a reference table

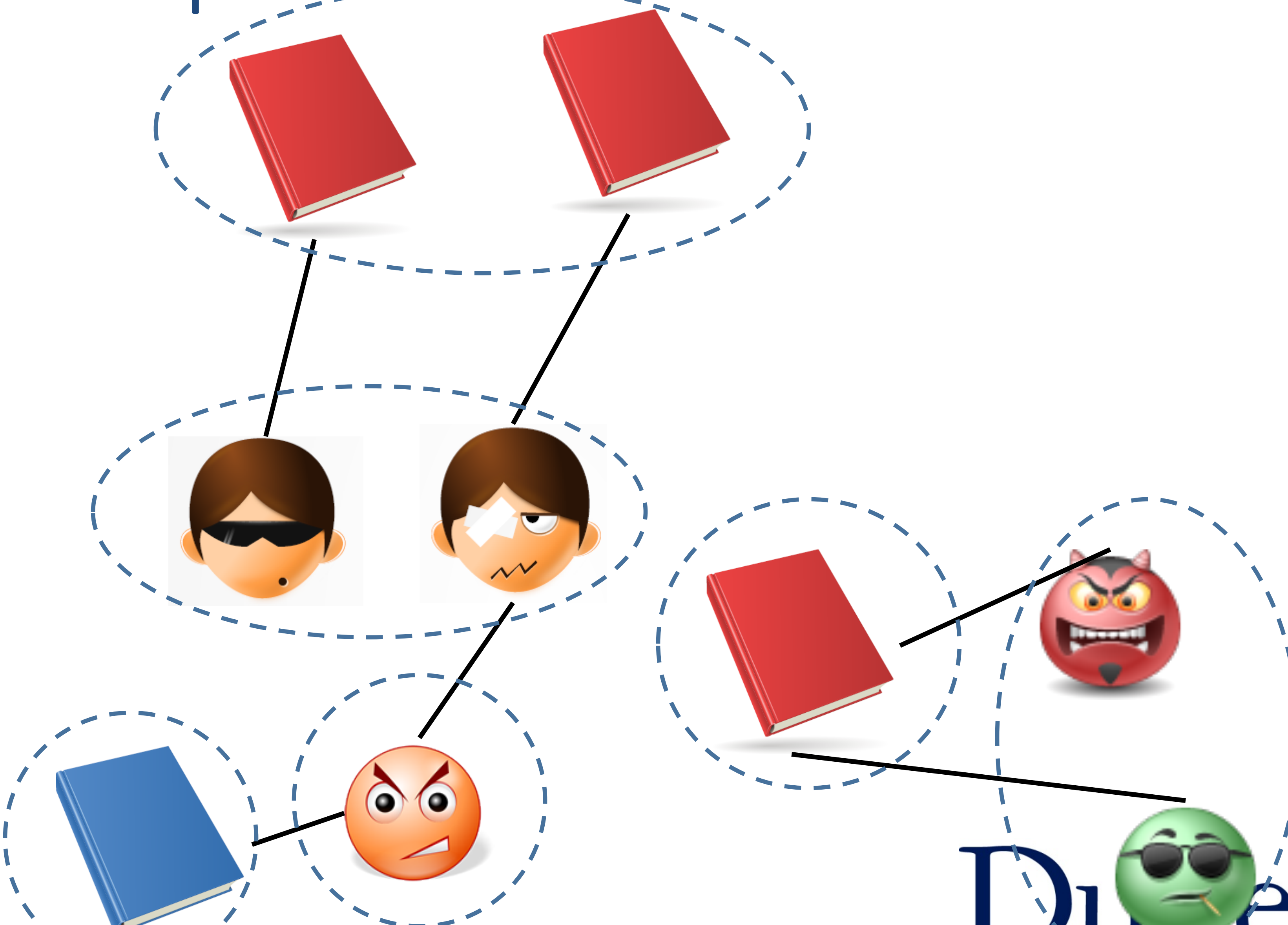


Relational/Multi-Entity Problem Statement

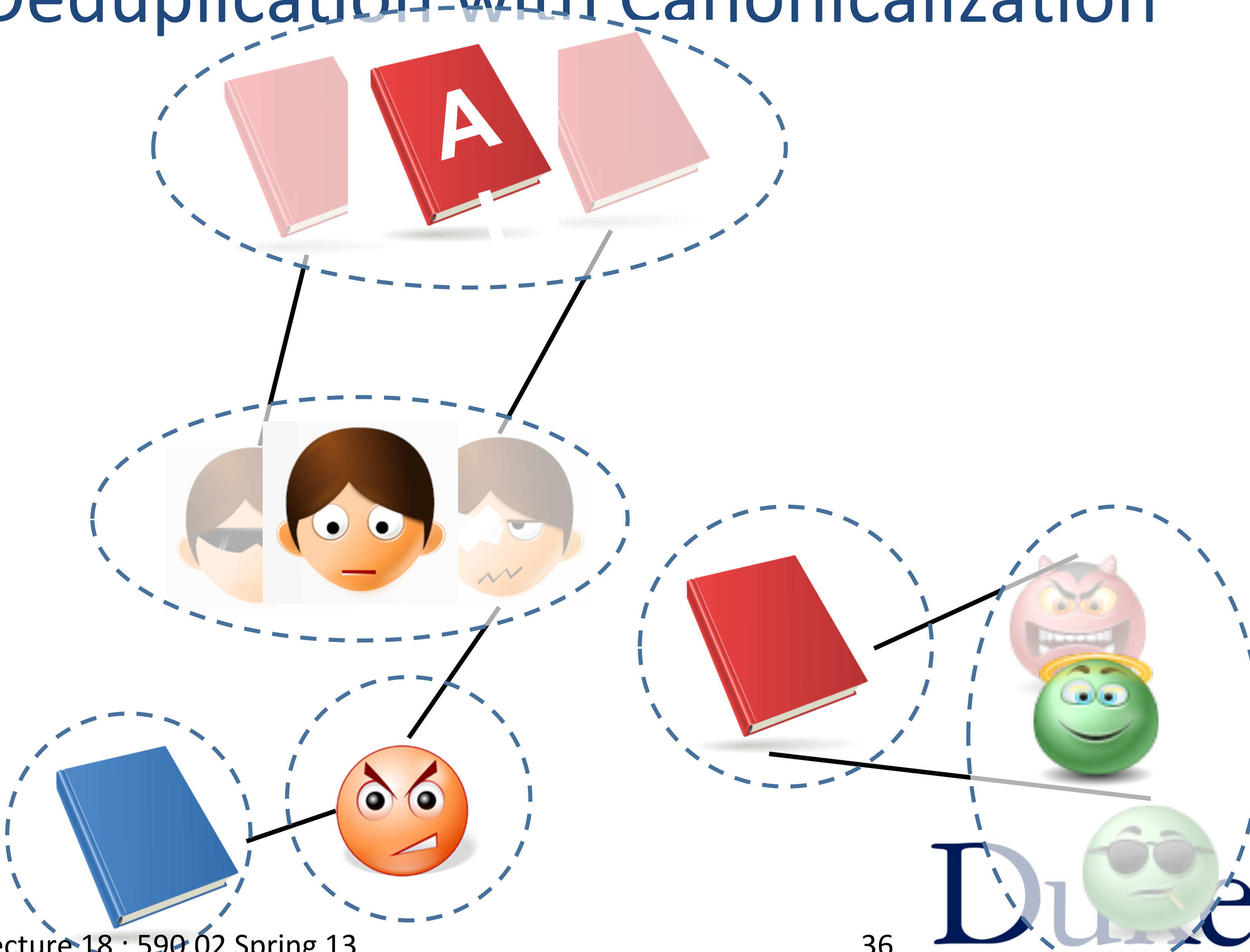
Real World | **Digital World**



Deduplication Problem Statement

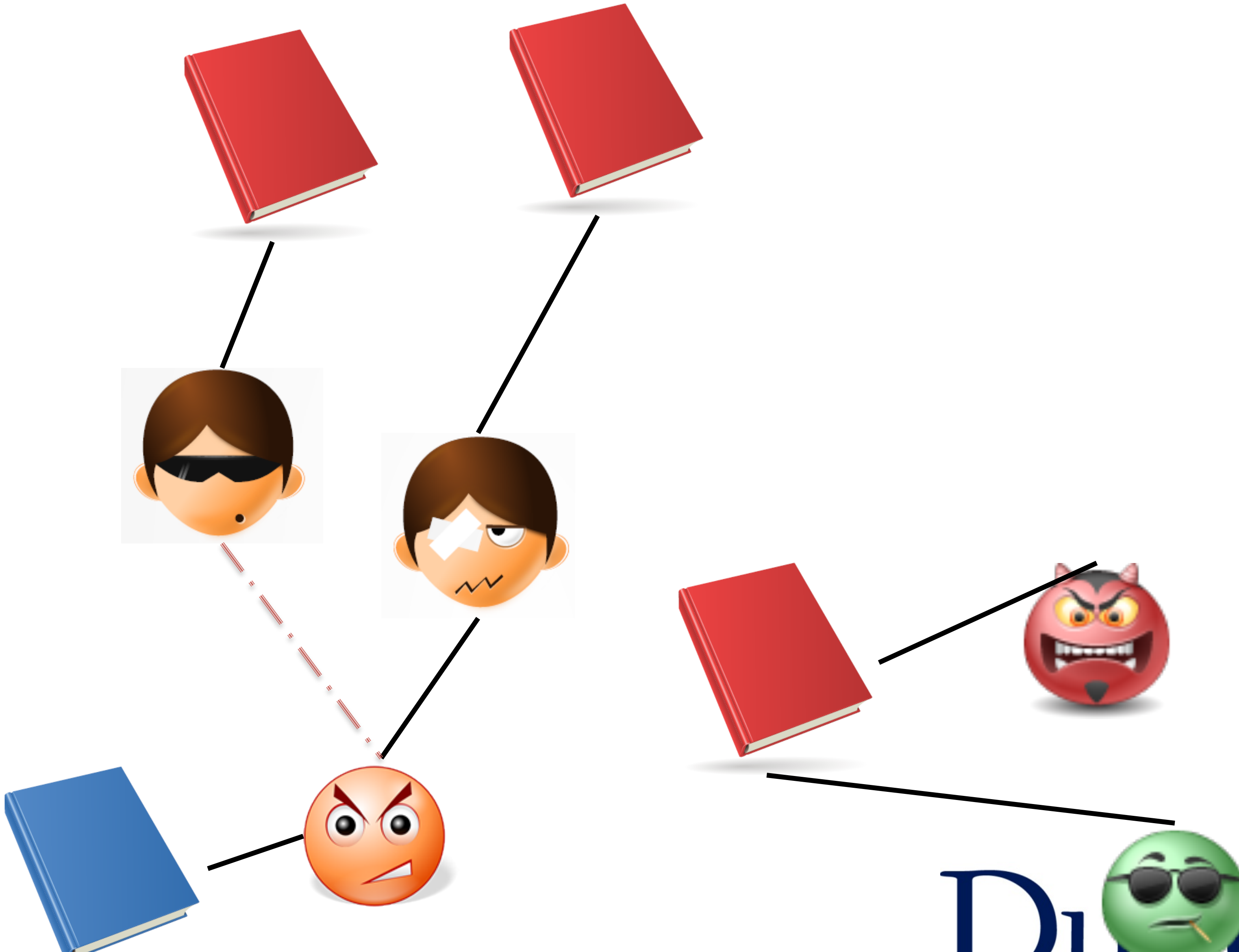


Deduplication with Canonicalization

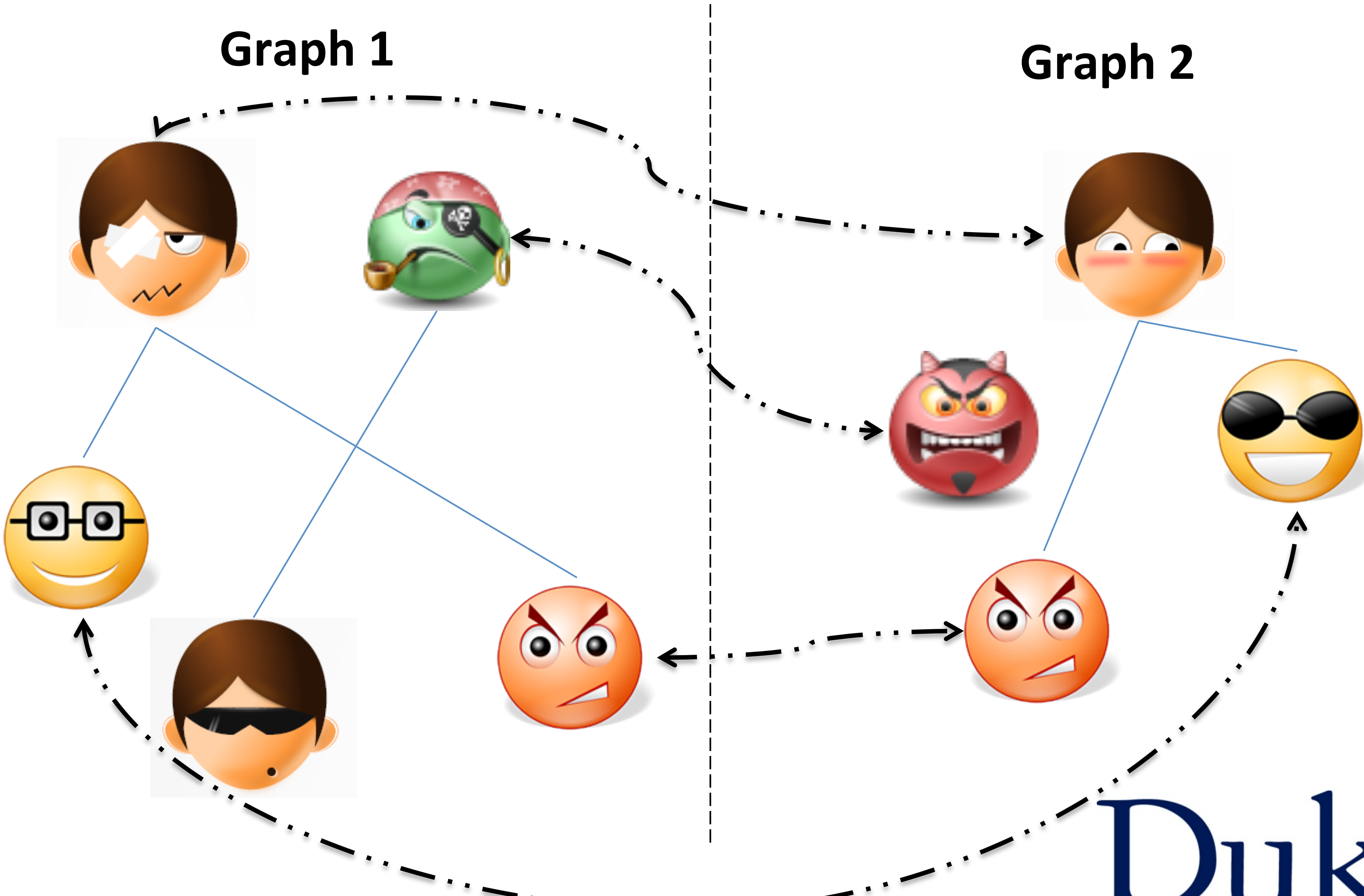


DUKE UNIVERSITY

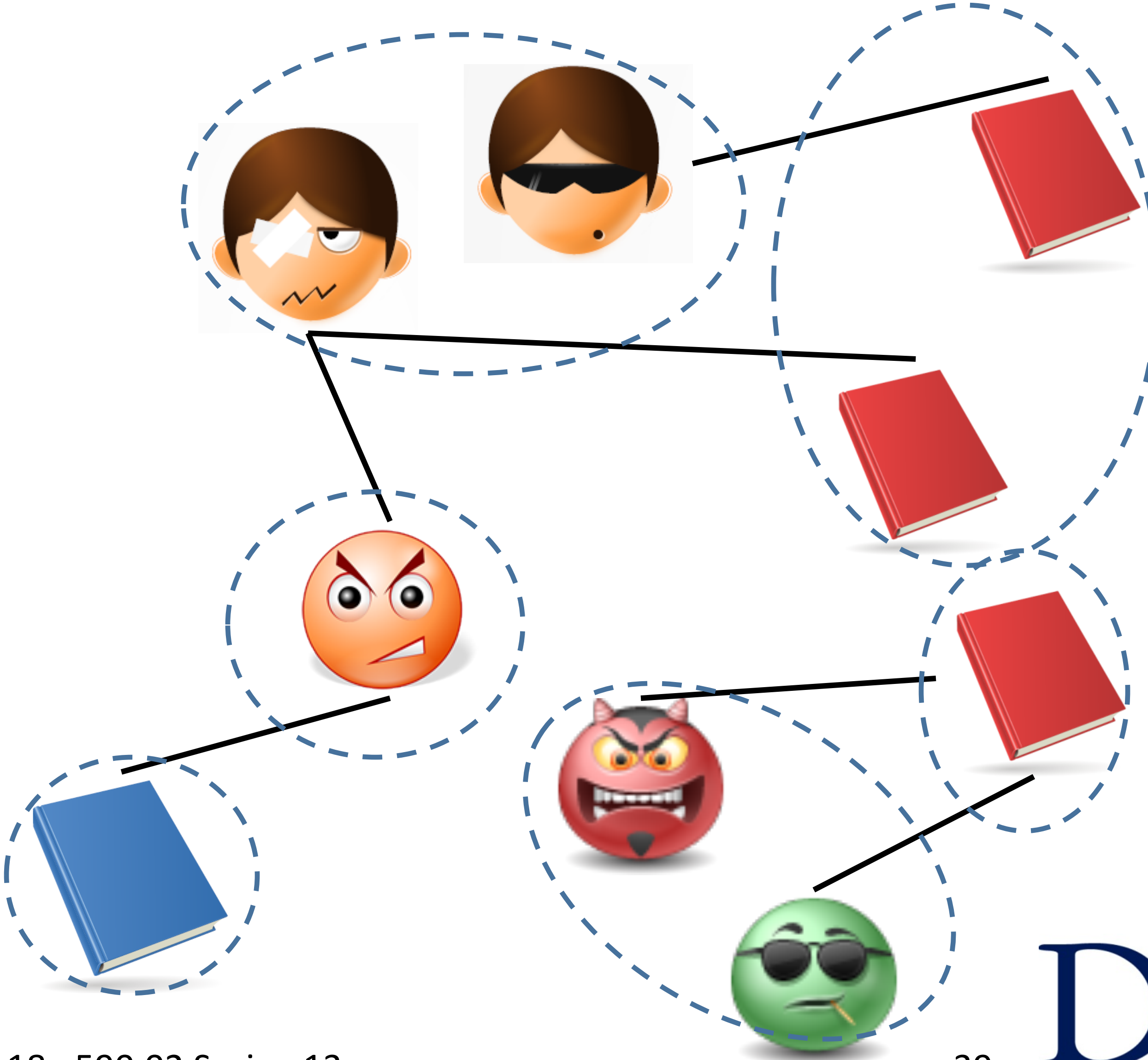
Link Prediction Problem Statement



Graph Alignment (& motif search)

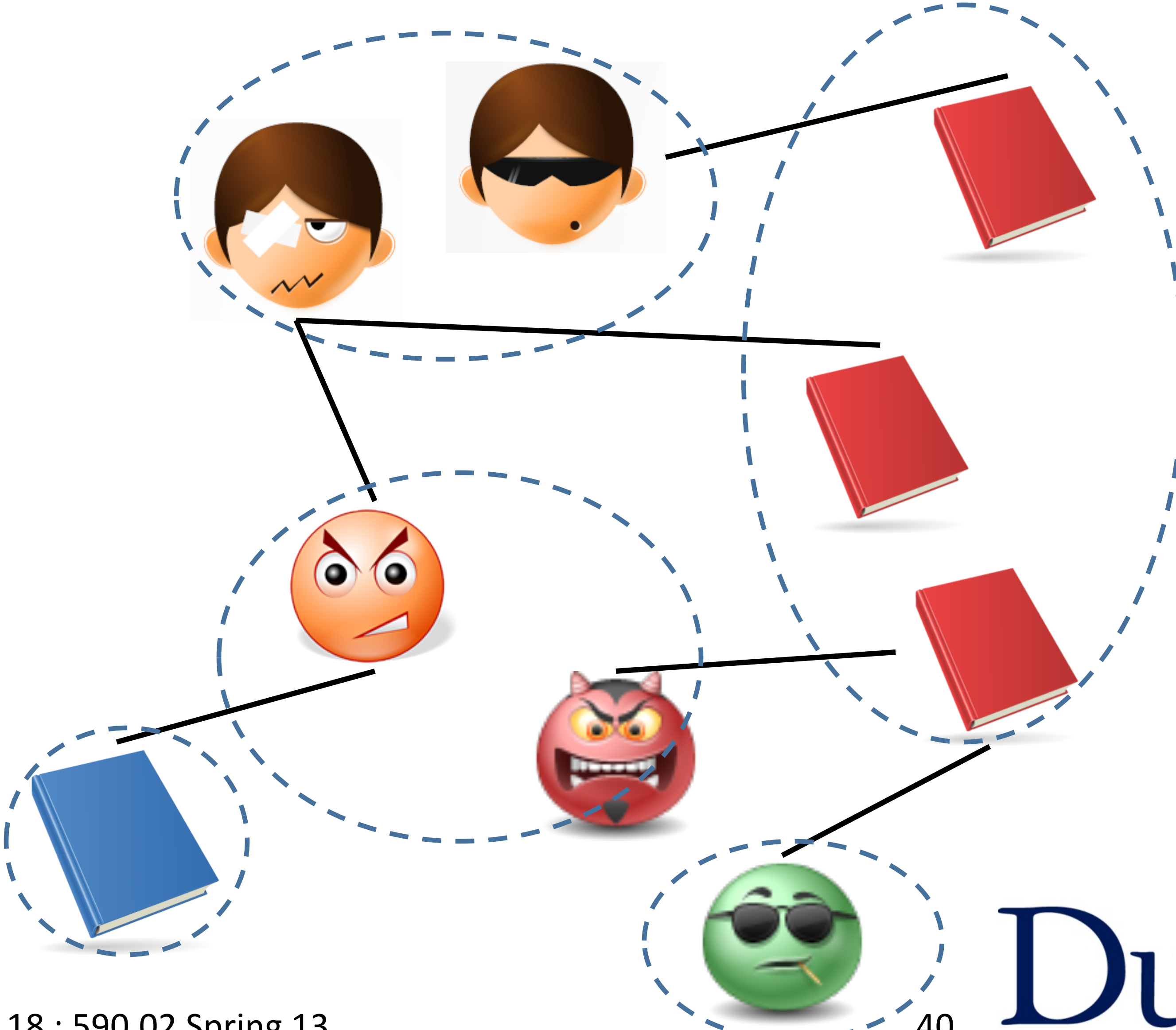


Relationships are crucial



Duke UNIVERSITY

Relationships are crucial



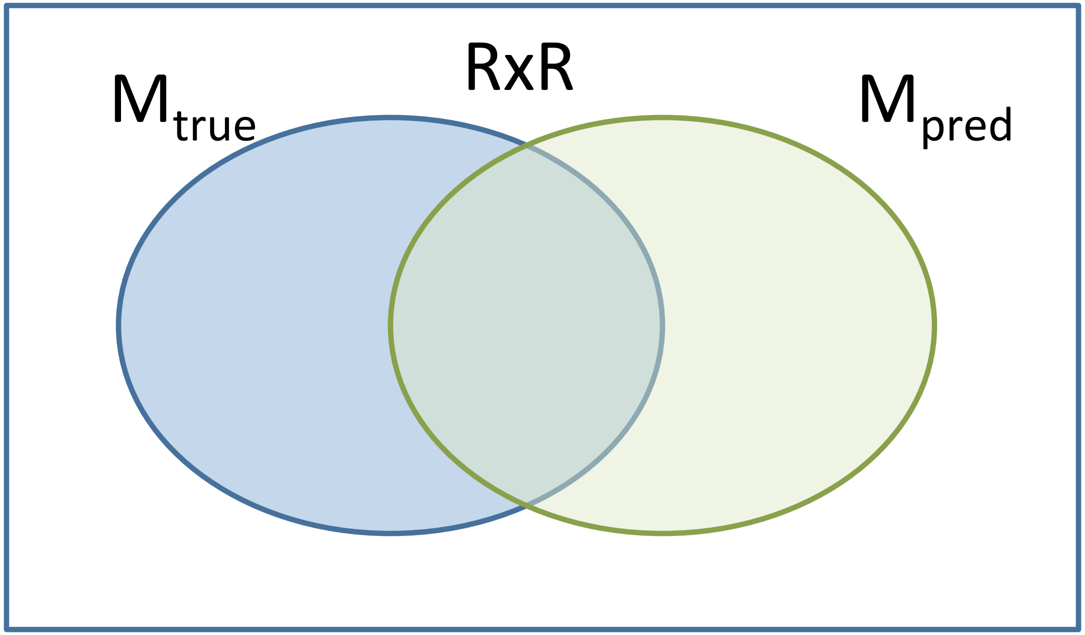
Notation

- *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_{true} and M_{pred}

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

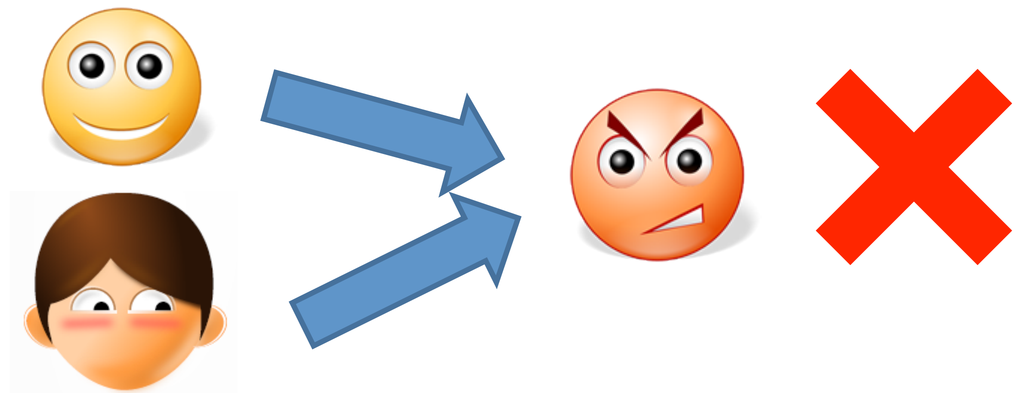


Metrics

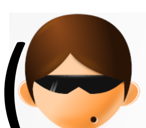
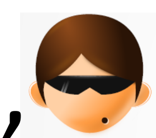
- Pairwise metrics
  - Precision/Recall, F1
  - # of predicted matching pairs
- Cluster level metrics
  - purity, completeness, complexity
  - Precision/Recall/F1: Cluster-level, closest cluster, MUC, B^3 , Rand Index
  - Generalized merge distance [Menestrina et al, PVLDB10]
- Little work that evaluates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

Typical Assumptions Made

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



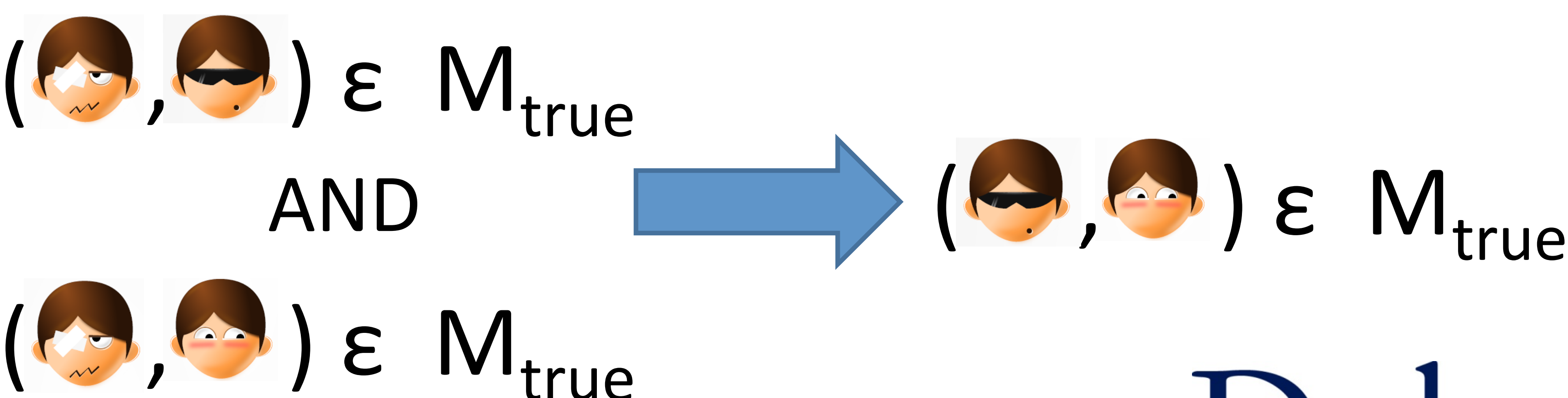
- *In record linkage, no duplicates in the same source*
- *If two records/mentions are identical, then they are true matches*

( , ) ε M_{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



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^{st}-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 and y match).

Solutions:

1. Weighted sum or average of component-wise similarity scores. Threshold determines match or non-match.
   - 0.5*1st-author-match-score + 0.2*venue-match-score + 0.3*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 and y 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^{st}-author-match-score > 0.7 AND venue-match-score > 0.8) OR (paper-match-score > 0.9 AND venue-match-score > 0.9)
   – Manually formulating the right set of rules is hard.

Duke
UNIVERSITY

Basic ML Approach

- *r = (x,y)* is record pair, γ is comparison vector, *M* matches, *U* non-matches

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

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

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

Duke UNIVERSITY

Fellegi & Sunter Model [FS, Science '69]

- Record pair: *r = (x,y)* in A x B

- γ is comparison vector
  - E.g., γ = ["Is x.name = y.name?", "Is x.address = y.address?" …]
  - 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, γ 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 x B,

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

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

Duke UNIVERSITY

Error due to a Linkage Rule

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

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

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

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

Optimal Linkage Rule

- Let μ and λ 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 Γ with error bounds μ and λ, if
  - L* meets the type I and type II requirements

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

  - L* has the least conditional probabilities of *not making a decision.* That is for all other decision rules L (with error bounds μ and λ),

$$P(A2(L^*)|U) \leq P(A2(L)|U),$$
$$P(A2(L^*)|M) \leq P(A2(L)|M)$$

Finding the Optimal Linkage Rule

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

Largest $\frac{m(\gamma)}{u(\gamma)}$

$$\underbrace{\gamma_1, \ldots, \gamma_n}_{A_1}, \underbrace{\gamma_{n+1}, \ldots, \gamma_{n'-1}}_{A_2}, \underbrace{\gamma_{n'} \ldots, \gamma_N}_{A_3}$$

Smallest $\frac{m(\gamma)}{u(\gamma)}$

- Set A1 to be the first n vectors, and A3 to be the last N – n' vectors such that:

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

Using Fellegi Sunter in Practice

- Γ is usually high dimensional (computing $m(\gamma)$ and $u(\gamma)$ is inefficient)
  - Use conditional independence of features in γ 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 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

Duke UNIVERSITY

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, …]

Duke UNIVERSITY

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