Data Mining Concepts & Tasks

Duen Horng (Polo) Chau

Georgia Tech

CSE6242 / CX4242

Jan 16, 2014

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos

Last Time

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

Data Cleaning

- Google Refine, Data Wrangler

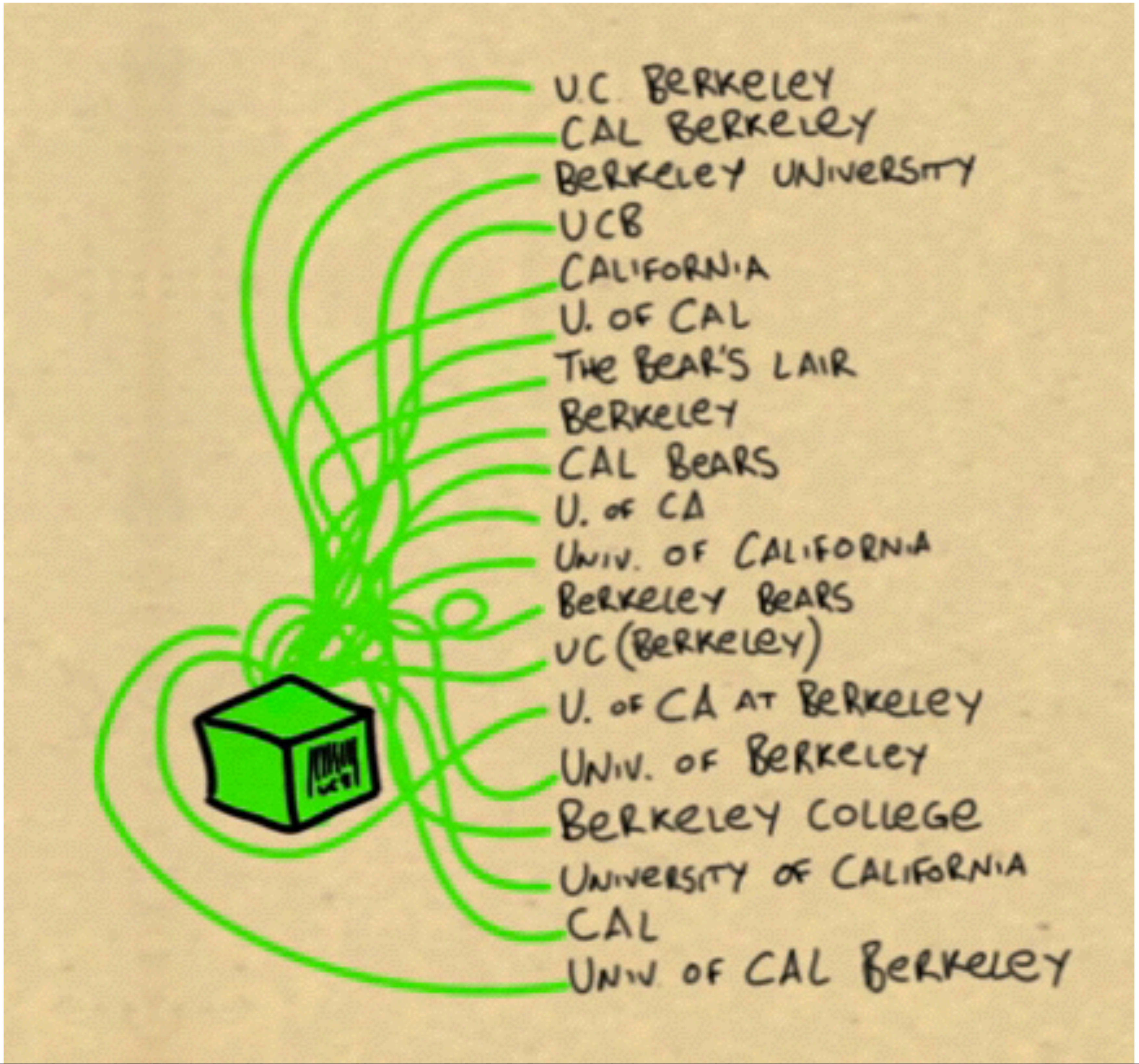
Data Integration

- Many examples: Google knowledge graph, Facebook Graph Search, Freebase, Feldspar, Kayak, Apple Siri, etc.
- We previewed the "**D-Dupe**" tool for "entity resolution"

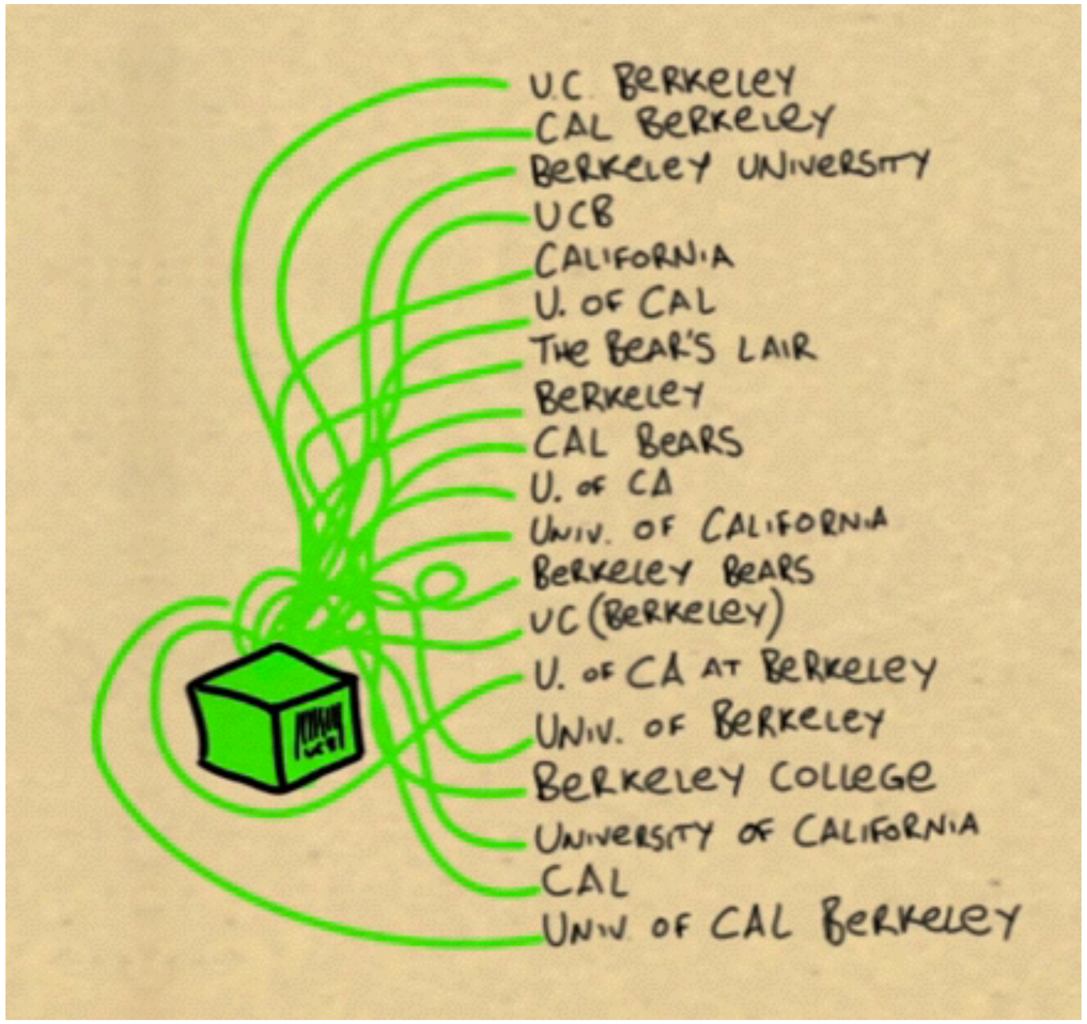
Continuing with

Data Integration

What do we **need** before we can even integrate datasets/tables/schemas?



What do we **need** before we can even integrate datasets/tables/schemas?



You need an ID for every unique entity/item/object/thing… Easy?

Entity Resolution

(A hard problem in data integration)

Polo Chau
P. Chau
Duen Horng Chau
Duen Chau
D. Chau

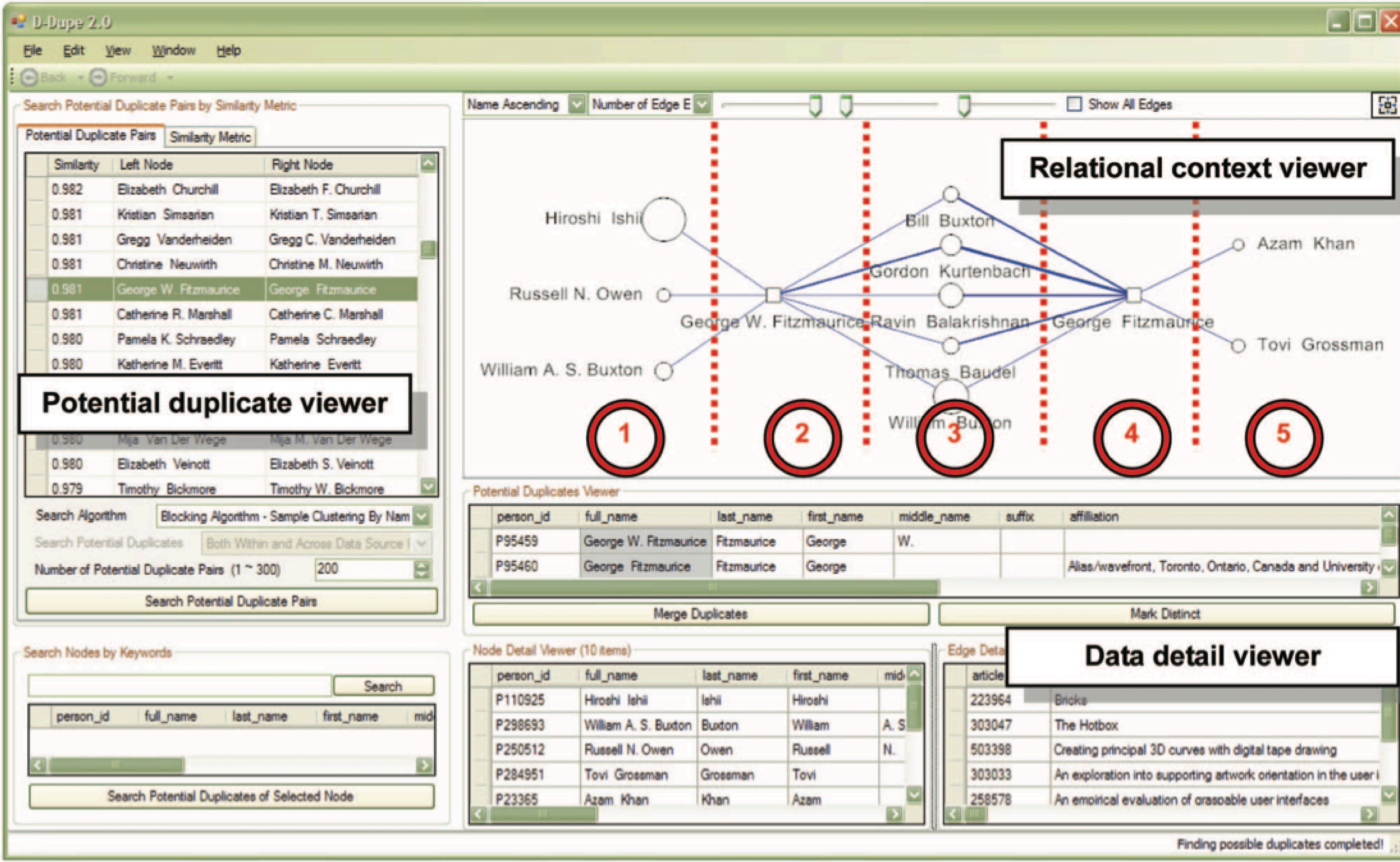
D-Dupe

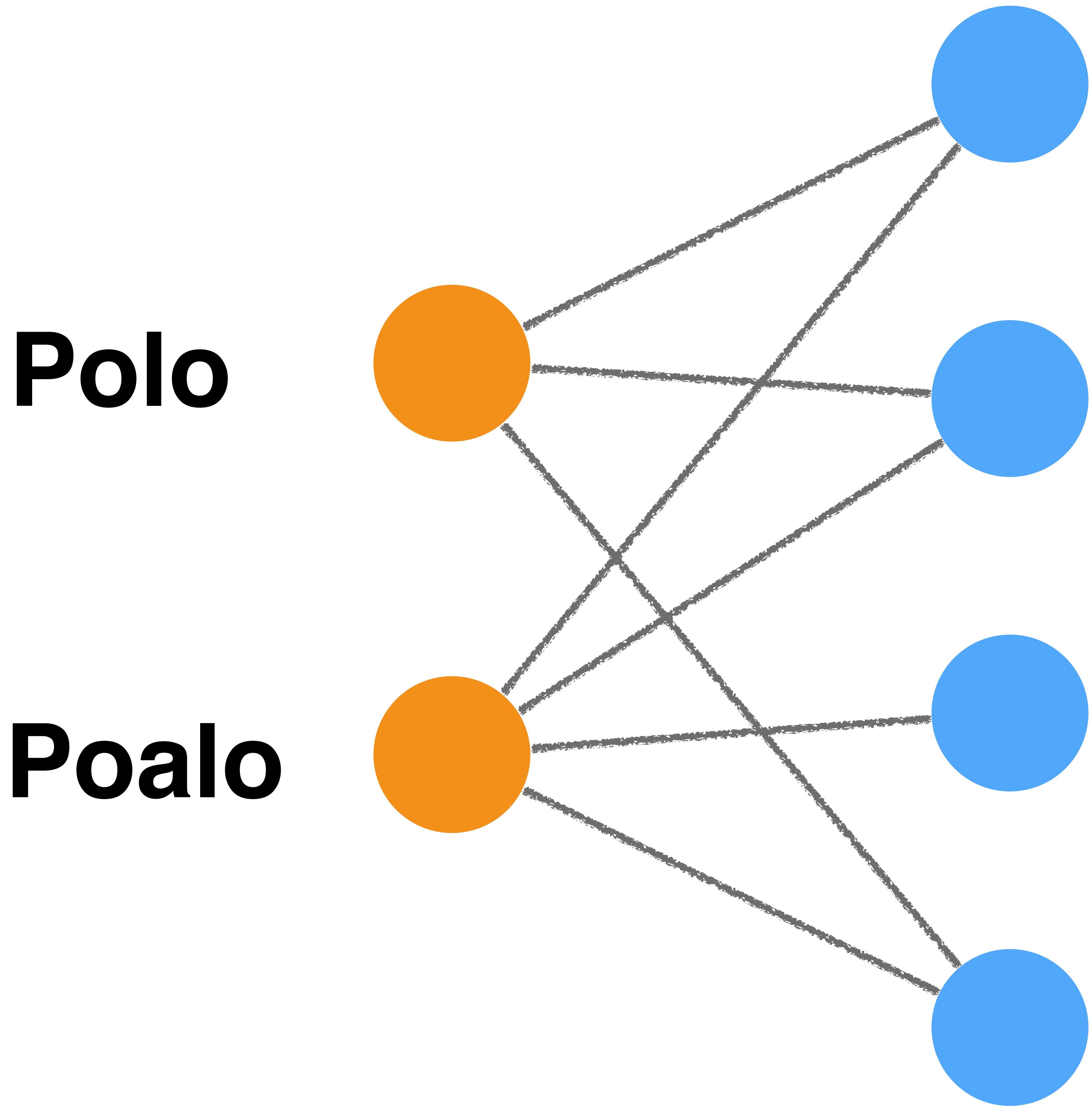
Interactive Data Deduplication and Integration
TVCG 2008

University of Maryland
Bilgic, Licamele, Getoor, Kang, Shneiderman

http://linqs.cs.umd.edu/basilic/web/Publications/2008/kang:tvcg08/kang-tvcg08.pdf

http://www.cs.umd.edu/projects/linqs/ddupe/ (skip to 0:55)

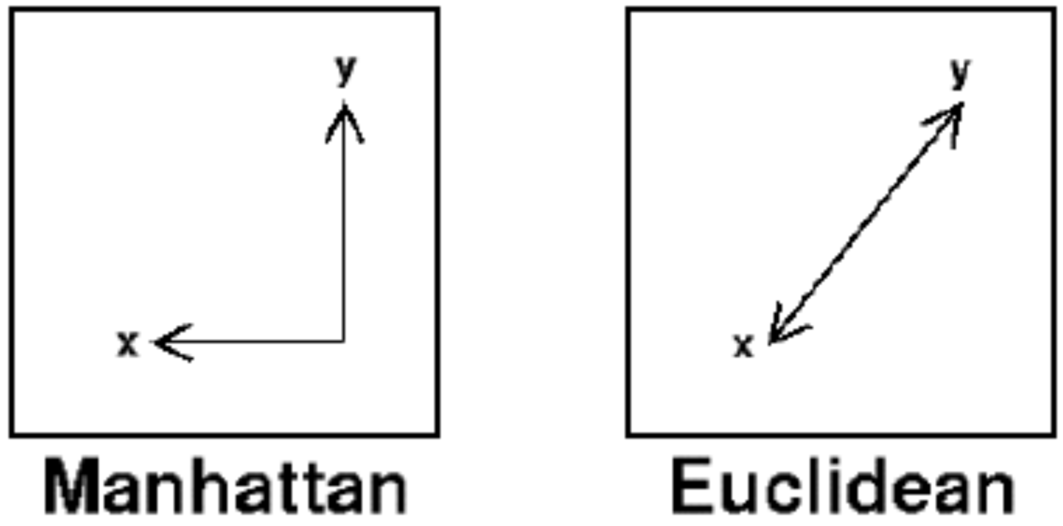

D-Dupe 2.0
File Edit View Window Help
Back Forward
Search Potential Duplicate Pairs by Similarity Metric
Potential Duplicate Pairs
Similarity Metric
Similarity Left Node Right Node
0.982 Elizabeth Churchill Elizabeth F. Churchill
0.981 Kristian Simsarian Kristian T. Simsarian
0.981 Gregg Vanderheiden Gregg C. Vanderheiden
0.981 Christine Neuwirth Christine M. Neuwirth
0.981 George W. Fitzmaurice George Fitzmaurice
0.981 Catherine R. Marshall Catherine C. Marshall
0.980 Pamela K. Schraedley Pamela Schraedley
0.980 Katherine M. Everitt Katherine Everitt
0.980 Mija Van Der Wege Mija M. Van Der Wege
0.980 Elizabeth Veinott Elizabeth S. Veinott
0.979 Timothy Bickmore Timothy W. Bickmore
Potential duplicate viewer
Search Algorithm
Blocking Algorithm - Sample Clustering By Nam
Search Potential Duplicates
Both Within and Across Data Source
Number of Potential Duplicate Pairs (1 ~ 300)
200
Search Potential Duplicate Pairs
Search Nodes by Keywords
Search
person_id full_name last_name first_name
Search Potential Duplicates of Selected Node
Name Ascending
Number of Edge E
Show All Edges
Relational context viewer
Hiroshi Ishii
Russell N. Owen
William A. S. Buxton
George W. Fitzmaurice
Bill Buxton
Gordon Kurtenbach
Ravin Balakrishnan
Thomas Baudel
William Buxton
George Fitzmaurice
Azam Khan
Tovi Grossman
1
2
3
4
5
Potential Duplicates Viewer
person_id full_name last_name first_name middle_name suffix affiliation
P95459 George W. Fitzmaurice Fitzmaurice George W.
P95460 George Fitzmaurice Fitzmaurice George Alias/wavefront, Toronto, Ontario, Canada and University
Merge Duplicates
Mark Distinct
Node Detail Viewer (10 items)
person_id full_name last_name first_name
P110925 Hiroshi Ishii Ishii Hiroshi
P298693 William A. S. Buxton Buxton William A. S
P250512 Russell N. Owen Owen Russell N.
P284951 Tovi Grossman Grossman Tovi
P23365 Azam Khan Khan Azam
Edge Deta
Data detail viewer
article
223964 Bricks
303047 The Hotbox
503398 Creating principal 3D curves with digital tape drawing
303033 An exploration into supporting artwork orientation in the user i
258578 An empirical evaluation of graspable user interfaces
Finding possible duplicates completed!


Polo
Poalo

Numerous **similarity functions**

Excellent read: http://infolab.stanford.edu/~ullman/mmds/ch3a.pdf

- Euclidean distance
  Euclidean norm / L2 norm
- Manhattan distance
- Jaccard Similarity
  e.g., overlap of nodes' #neighbors



Jaccard similarity of sets S and T is $|S \cap T|/|S \cup T|$

- String edit distance
  e.g., "Polo Chau" vs "Polo Chan"
- Many more…

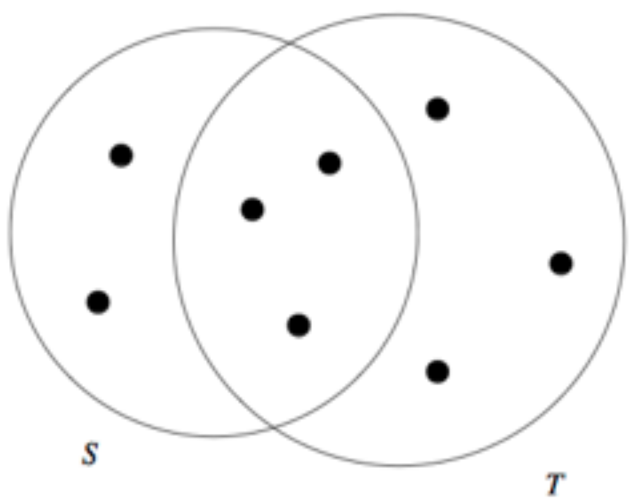


Figure 3.1: Two sets with Jaccard similarity 3/8

Core components: **Similarity functions**

Determine how two entities are similar.

D-Dupe's approach:
Attribute similarity + **relational similarity**

$$sim(e_i, e_j) = (1 - \alpha) \times sim_A(e_i, e_j) + \alpha \times sim_R(e_i, e_j),$$
$$0 \leq \alpha \leq 1,$$

Similarity score for a pair of entities

Attribute similarity (a weighted sum)

$$sim_A(e_i, e_j) = \sum_{k=1}^{n} w_k \times sim\_fun_k(e_i \cdot a_k, e_j \cdot a_k),$$

$$-1 \leq w_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{n} |w_k| = 1,$$

Summary for data integration

Opportunities

- enable new services (Siri, padmapper)
- enable new ways to discover info
- improve existing services
- reduce redundancy
- new way to interactive with data
- promote knowledge transfer (e.g., between companies)

Data Mining Concepts & Tasks

- Collection
- Cleaning
- Integration
- Analysis
- Visualization
- Presentation
- Dissemination

Each data-driven (business, decision-making) problem is **unique,** e.g., different goals, constraints.

Good news: many (sub)tasks that underlie these problems are **common**

Here is an **overview** of the common tasks.

1. (soft) Classification, Probability Estimation (supervised learning)

Predict which of a (small) set of classes an entity belong to.

Examples: Is this app malicious or benign? Will this customer click on this ad?

More Examples?
payment transaction -> fraudulent?
news/emails -> spam?
tumor -> benign?
sentiment analysis -> +, -, neutral
weather -> rain, storm, sunny
movies genres -> action, etc.
friends -> close, acquaintance, etc.
online dating -> will work out or not?
surveillance system -> suspicious or not

2. Regression ("value estimation") (supervised learning)

Predict the **numerical value** of some variable for an entity.

Example: how much minutes will this cellphone customer use?

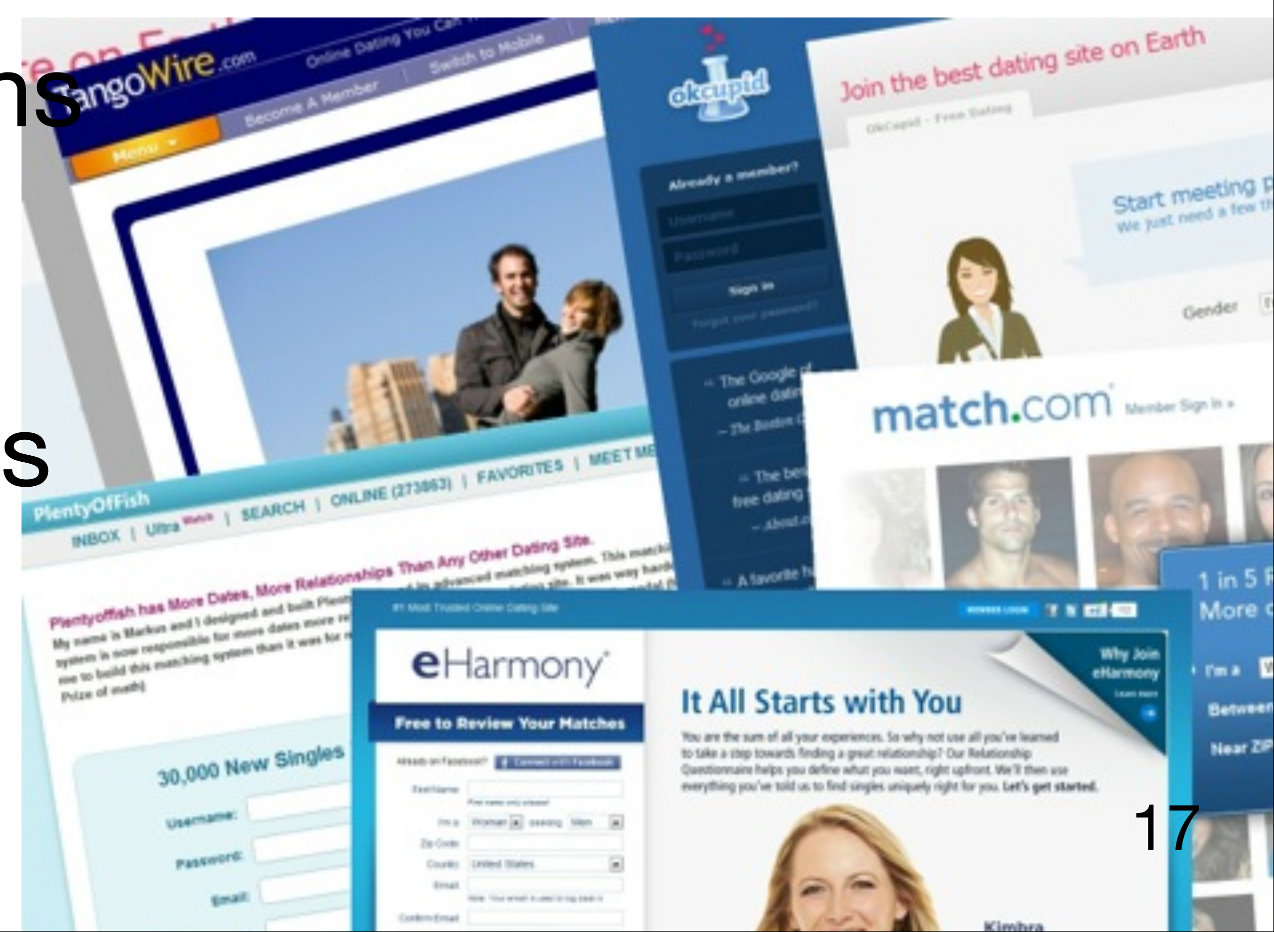
Related to classification, but predict **how much**, instead of **discrete decisions** (e.g., yes, no)

More Examples?
#cancer cells
length of stay of patients in hospital
loan limits to approve for a customer
rent of a house
stock price
online traffic
population
rating of movies
election results (#votes)
rainfall
scores (soccer/football)

3. Similarity Matching

Find similar entities (from a large dataset) based on what we know about them.

Examples?
online dating
similar songs/artists
netflix video recommendations
amazon products
flight deals, hotels
restaurants, tourist attractions
google: similar keywords
auto-correction



4. Clustering (unsupervised learning)

Group entities together by their similarity.

Examples?
organisms by environment
material recognition
biking groups (group bikers by interests, hobbies)
group stack overflow posts by tags
meetup atlanta
group locality according crime rate
grouping pixels in images -> differentiate between foreground and background -> object recognition
group plants (bare fruits or not?)
article grouping (academic or otherwise)
google news (world, sports, etc.)

5. Co-occurrence grouping

(Many names: frequent itemset mining, association rule discovery, market-basket analysis)

Find associations between entities based on transactions that involve them

(e.g., bread and milk often bought together)

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

6. Profiling / Pattern Mining / Anomaly Detection

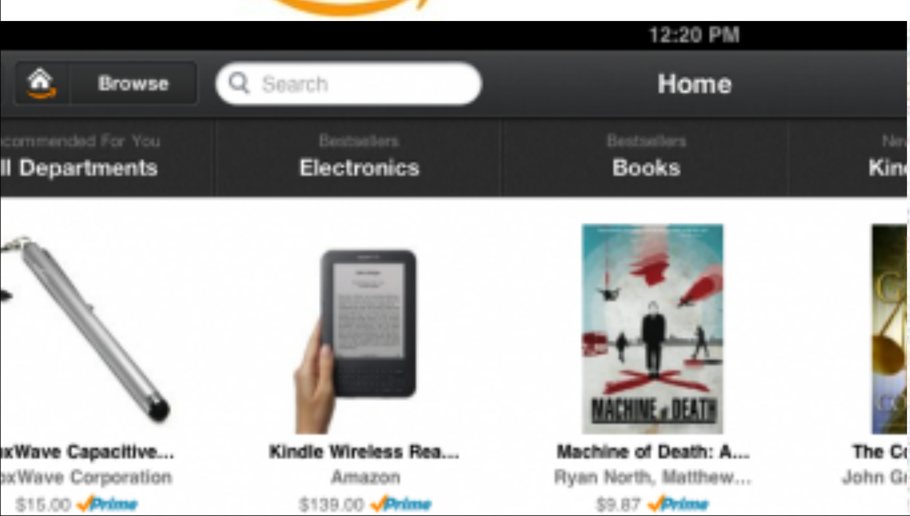
Characterize **typical** behaviors of an entity (person, computer router, etc.) so you can find **trends** and **outliers**.

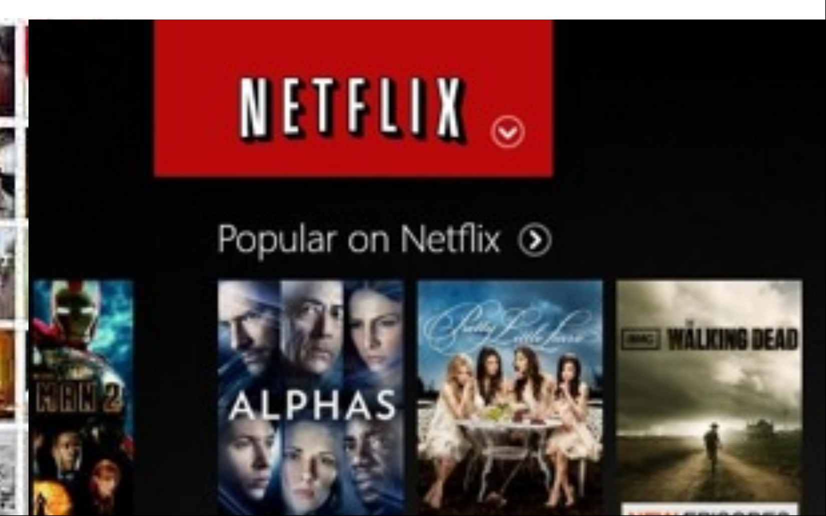
Examples?
computer instruction prediction
removing noise from experiment (data cleaning)
detect anomalies in network traffic
moneyball
weather anomalies (e.g., big storm)
google sign-in (alert)
smart security camera
embezzlement
trending articles

7. Link Prediction / Recommendation

Predict if two entities should be connected, and how strongly that link should be.

Examples?
two people on Facebook
amazon (things bought together); asssociation-rule mining
netflix: recommend jim carey movie
related questions on quora
top apps on apple store
crime group detection (bad guys on social network)
google search suggestions





8. Data reduction ("dimensionality reduction")

Shrink a large dataset into smaller one, with as little loss of information as possible

When to do it? Examples? Why do it?

Original data is too big -> too hard to process, or take too long

2D -> 1D (many Ds -> few Ds): for visualization, for more efficient algorithms

Graph partitioning - split a large graph into smaller subgraphs

Start thinking about project

- What kind of datasets and problems do you want to solve?
- What techniques do you need?

Survey

Why do you take this class?

* case studies/examples/end-to-end analysis
* applications
** methods to handle/visualize real-world big data
learn the right analytics steps, right way to display data