Attend Part 1 (2-3pm) to get 1 point extra credit.
Polo will announce on Piazza options for DL students.

Data Science/Data Analytics and Scaling to Big Data with MathWorks

Using Data Analytics to turn large volumes of complex data into actionable information can help you improve design and decision-making processes. However, developing effective analytics and integrating them into business systems can be challenging. In this seminar you will learn approaches and techniques available in MATLAB® to tackle these challenges

Data Science/Data Analytics (Part 1) — Fri, Feb 6, 2-3pm l Klaus 1116

* **Access and Explore data**: Accessing, exploring, and analyzing data stored in files, the web, and databases
* **Data Munging**: Techniques for cleaning, exploring, visualizing, and combining complex multivariate data sets
* **Developing predictive models**: Prototyping, testing, and refining predictive models using machine learning methods
* I**ntegrating analytics with systems**: Integrating and running analytics within enterprise business systems and interactive web applications

Scale to Big Data (Part 2) — Fri, Feb 6, 3-4pm l Klaus 2443

* Work with out-of-memory datasets with MATLAB
* MapReduce algorithms and Hadoop integration in MATLAB

CSE 6242 / CX 4242

Data Mining Concepts & Tasks

Duen Horng (Polo) Chau
Georgia Tech

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

Final words about

Data Integration

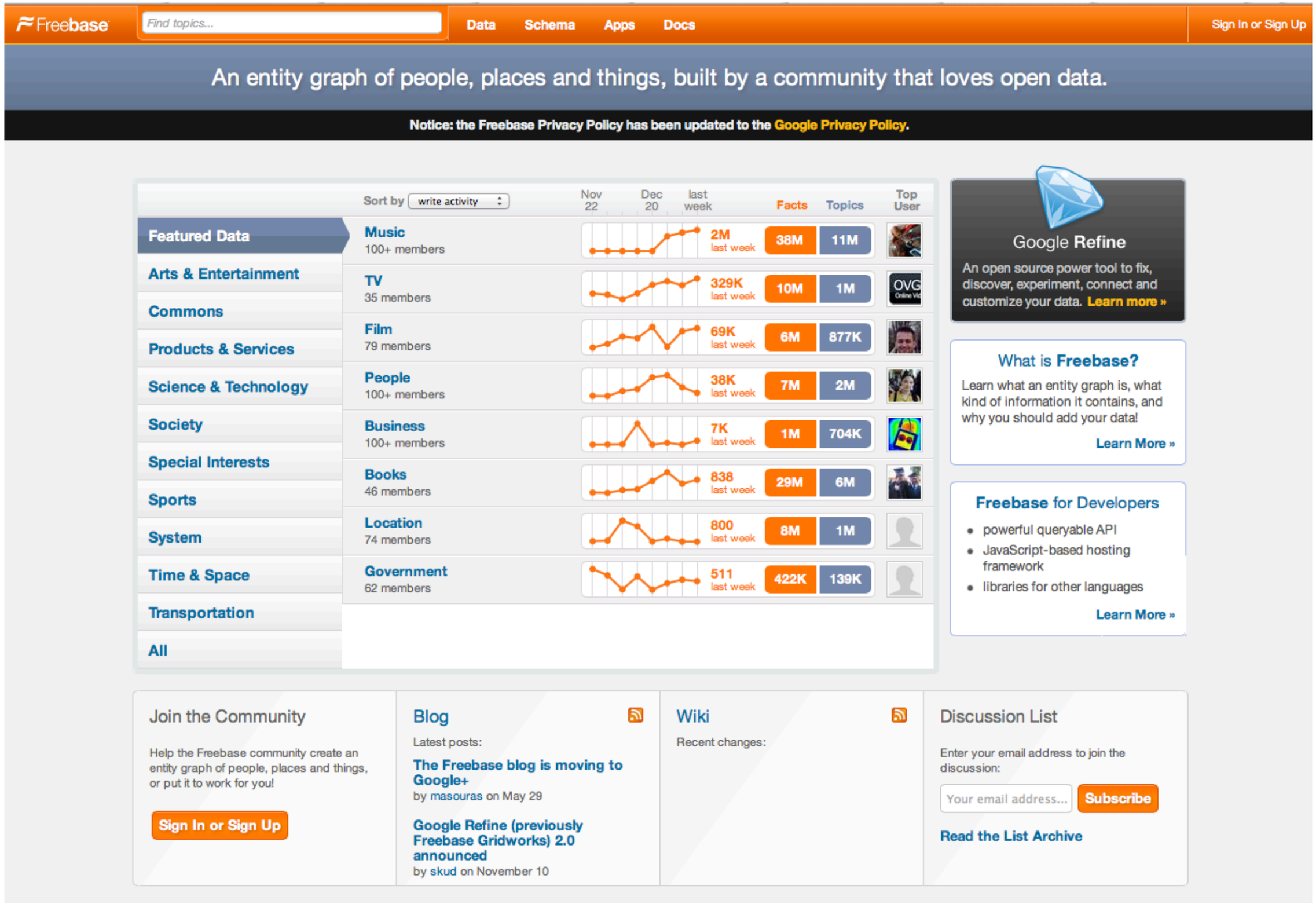
Freebase

(a graph of entities)

"…a large collaborative knowledge base consisting of metadata composed mainly by its **community members**…"

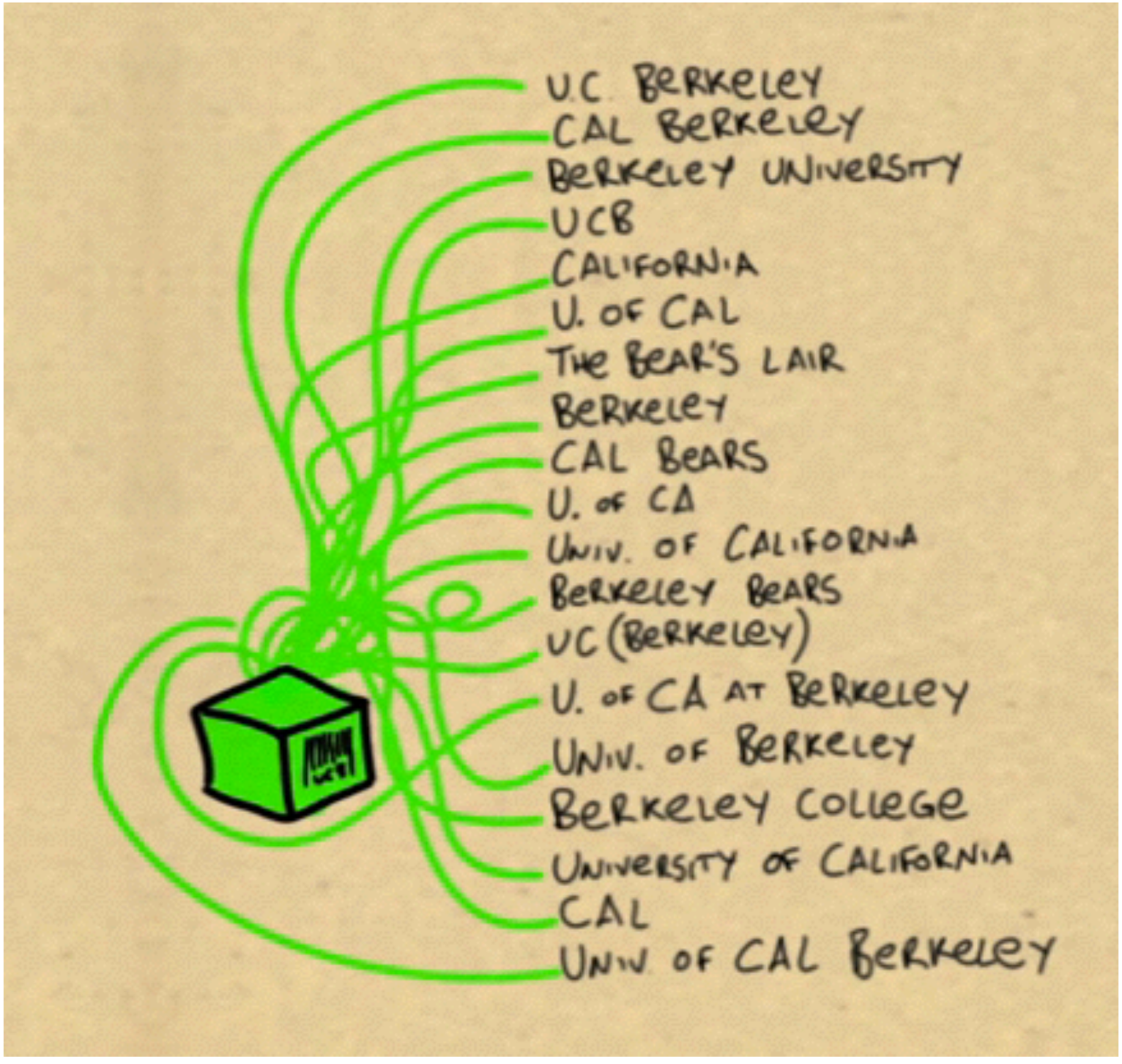
Wikipedia.

Crowd-sourcing Approaches: Freebase



http://wiki.freebase.com/wiki/What_is_Freebase%3F

We need ways to identify the **many ways** that **one thing** may be called. How?



Entity Resolution

(A hard problem in data integration)

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

Why is **Entity Resolution** so Important?

ebay Shop by category

iphone

All Categories

Search Advanced

Related: iphone 5 iphone 4 iphone unlocked iphone 3gs iphone verizon iphone 5c iphone 4 unlocked iphone 3 samsung galaxy s3 ...

Include description

Categories

Cell Phones & Accessories (2,653,244)

- Cell Phone Accessories (2,414,030)
- **Cell Phones & Smartphones**
- Other (144,703)
- Replacement Parts & Tools (56,276)
- Wholesale Lots (4,886)
- More

See all categories

Features see all

Contract see all

Condition see all

- New (3,232)
- New other (see details) (1,831)
- Manufacturer refurbished (559)
- Seller refurbished (1,563)
- Used (19,256)
- For parts or not working (6,281)

Price

- Under $25
- $25 - $50
- $50 - $100
- Over $100

$ to $

Format see all

- All Listings (32,722)
- Auction (15,207)
- Buy It Now (23,263)

Item Location see all

- Default
- Within

100 mi of 30332

- US Only
- North America
- Worldwide

Delivery Options see all

- Free shipping

All Listings | Auction | Buy It Now

Sort: Best Match View:

All > Cell Phones & Accessories > Cell Phones & Smartphones

iphone 32,722 listings + Follow this search

Find Your iPhone

iPhone 5s

- 4" Retina Display
- True Tone Flash
- Slo-mo Video
- Touch ID

Shop iPhone 5s

iPhone 5c

- 4" Retina Display
- 8 MP Camera
- 1080p HD Video
- Multi Color Opt.

Shop iPhone 5c

iPhone 5

- 4" Retina Display
- 8 MP Camera
- 1080p HD Video
- Face Time, Siri

Shop iPhone 5

Apple iPhone 4S A1387, Sprint, 16GB, White, Clean ESN

$54.00 2m left (Today 7:53AM)

2 bids

Apple iPhone 4 - 8GB - (Verizon) Smartphone - Black or White - Good

USA SELLER *** WARRANTY *** ACCESSORIES INCLUDED

$79.88

Buy It Now

Free shipping

Save $10 for every 3 items you buy

FAST 'N FREE

Get it on or before **Sat, Sep. 13**

2053+ Watchers

Apple iPhone 4 - 16GB - Black (Verizon) Smartphone 7.1.2 MC676LL/A Clean ESN

$79.00 2m left (Today 7:54AM)

0 bids

$125.00

Buy It Now

Popular on eBay

Apple iPhone 4 - 8GB - Verizon Straight Talk...

$109.95

Buy It Now

Free shipping

U Apple iPhone 4 - 8GB - Black (Verizon)...

$78.95

Buy It Now

Free shipping

U Apple iPhone 4 - 8GB - White (Verizo...

$79.95

Buy It Now

Free shipping

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 I

Number of Potential Duplicate Pairs (1 ~ 300): 200

Search Potential Duplicate Pairs

Search Nodes by Keywords

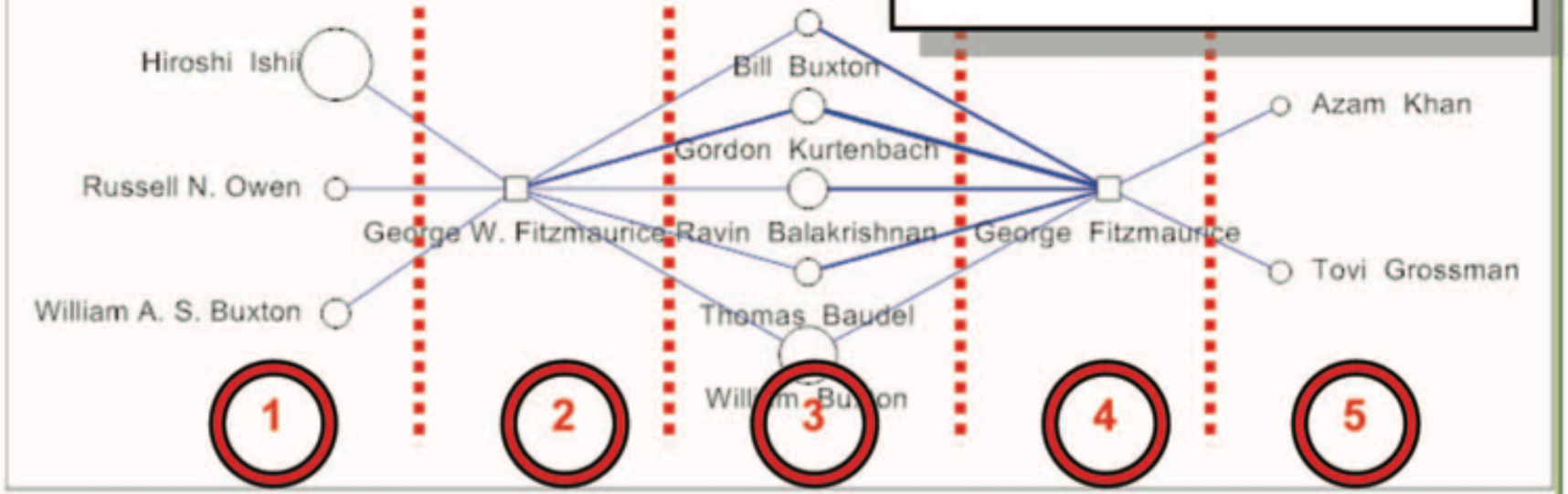
Search

person_id	full_name	last_name	first_name	mid

Search Potential Duplicates of Selected Node

Name Ascending | Number of Edge E | Show All Edges

Relational context viewer



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	mid
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

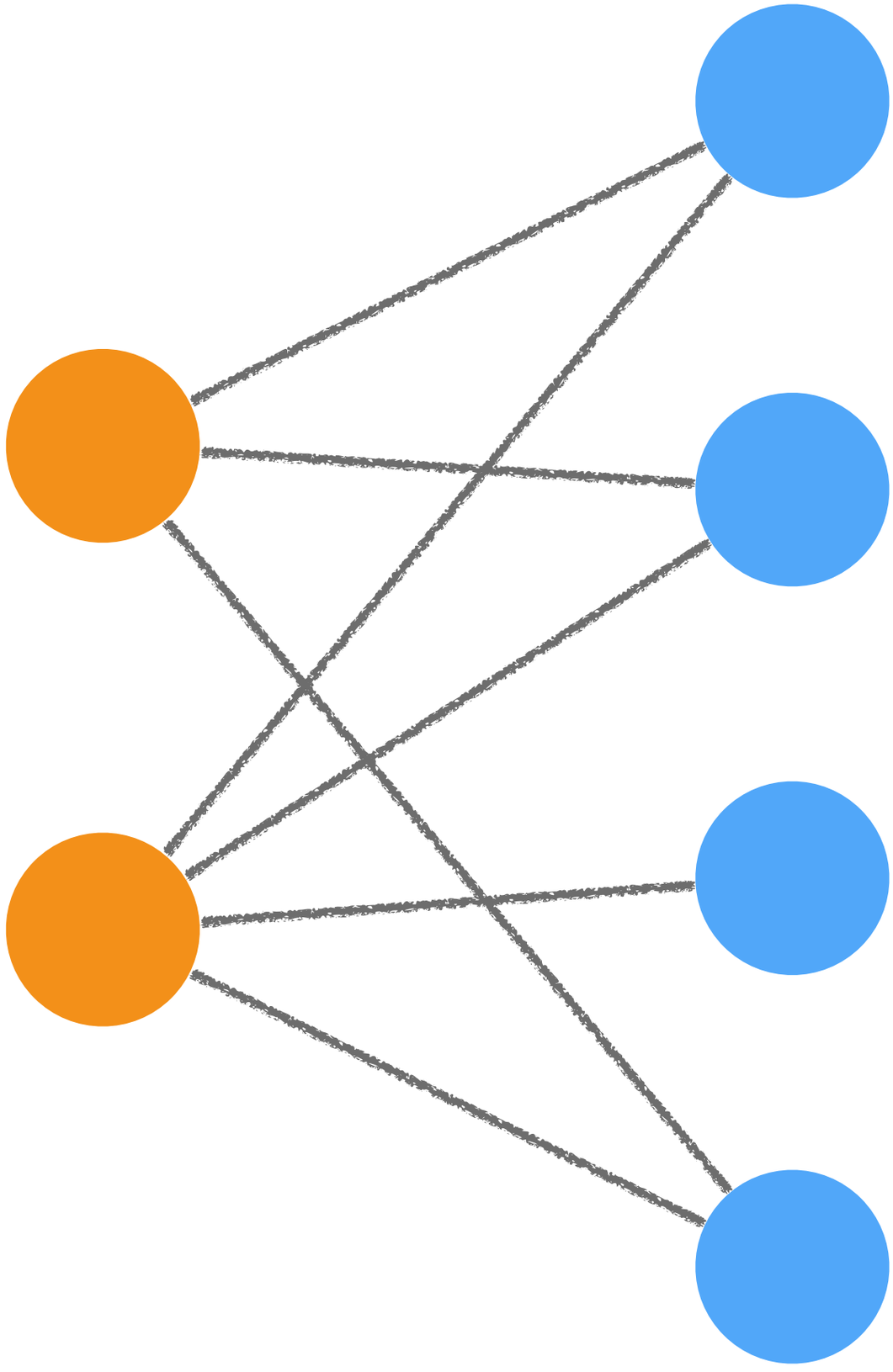
Data detail viewer

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 emoirical evaluation of orasoable user interfaces

Finding possible duplicates completed!

Polo

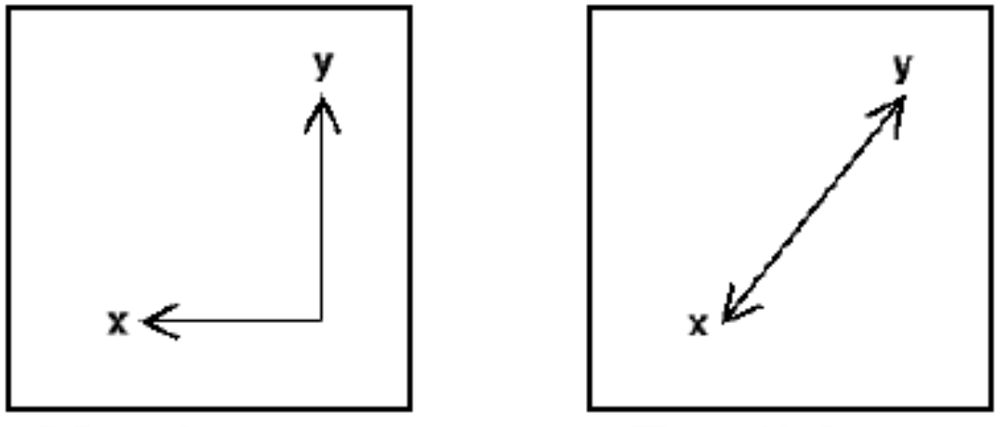
Poalo



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



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

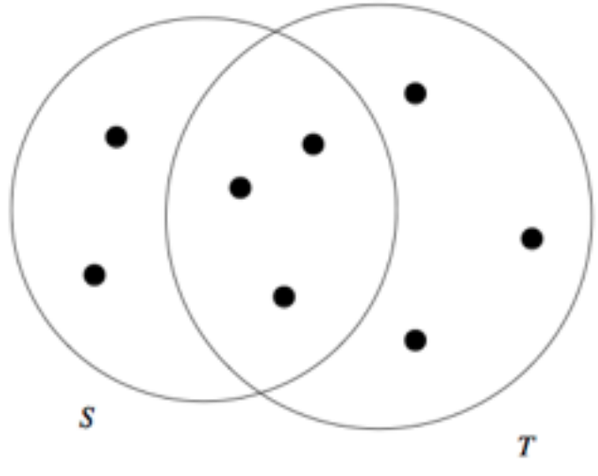


Figure 3.1: Two sets with Jaccard similarity 3/8

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

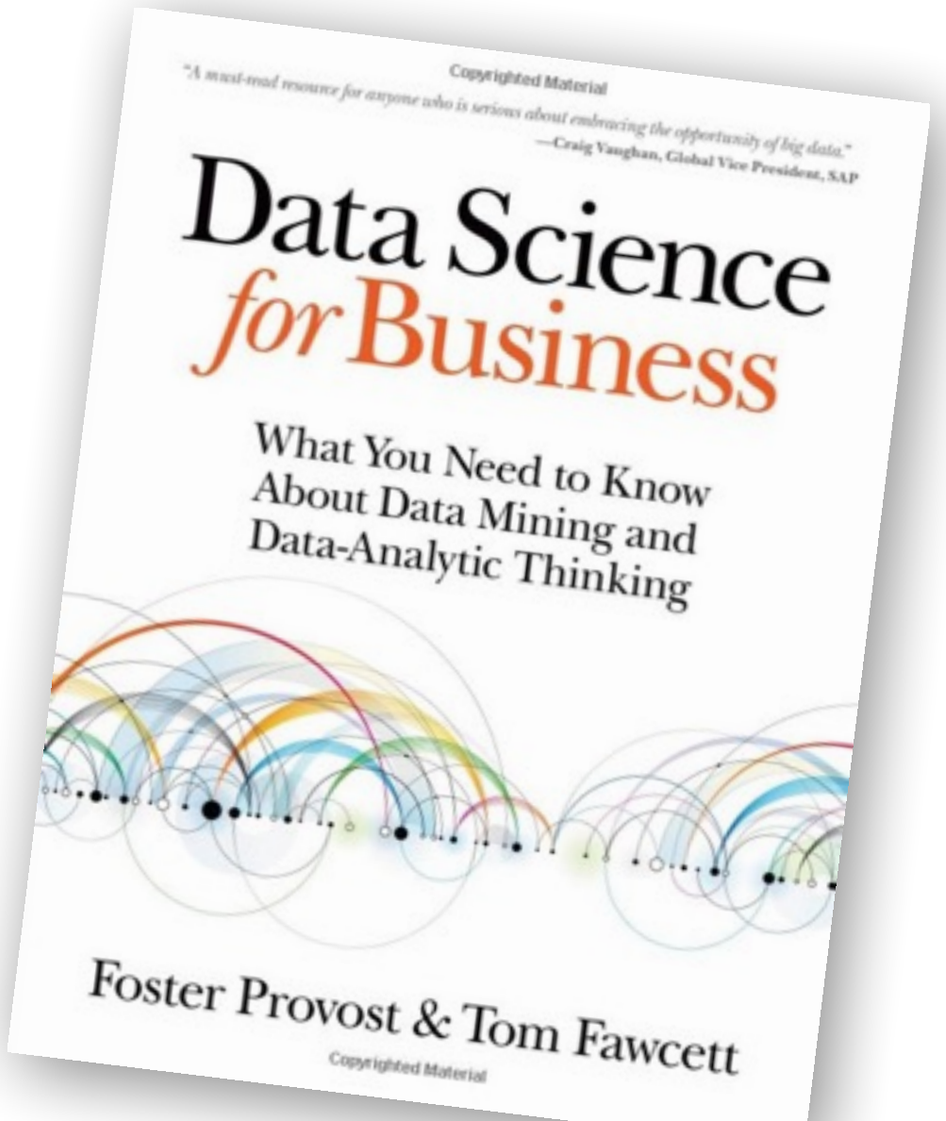
Attribute similarity (a weighted sum)

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

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

Data Mining Concepts & Tasks

A critical skill in data science is the ability to decompose a data-analytics problem into pieces such that each piece matches a known task for which tools are available. Recognizing familiar problems and their solutions avoids wasting time and resources reinventing the wheel. It also allows people to focus attention on more interesting parts of the process that require human involvement—parts that have not been automated, so human creativity and intelligence must come into play.



http://www.amazon.com/Data-Science-Business-data-analytic-thinking/dp/1449361323

1. Classification or class Probability Estimation

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

- Cancer testing (yes, no)
- Movie genre (action, drama, etc.)
- sports (win, loss)
- email spam filter (spam, or not)
- gesture detection (pinch, swipe…)
- planet zone habitable or not
- gene prediction
- news types (sports, entertainment)
- virus scanning (malware or not)

2. Regression ("value estimation")

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

- point value of wine (50-100)
- credit score (start with classification; default or not)
- stock prices — wall street
- relationship between price and sales
- weather
- sports and game scores

3. Similarity Matching

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

- recommending items you may want to buy
- find similar gene sequences (that may be repeating, or does similar things)
- online dating
- building auditing (energy consumption)
- patent search
- carpool matching (find people to carpool)
- detecting fake identies

4. Clustering (unsupervised learning)

Group entities together by their similarity. (User provides # of clusters)

- cluster people into demographics groups (young, old, etc)
- cluster people by accents (y'all, you all)
- hierarchical clustering for metabolomics
- clustering images on the web (cat?)
- ~=dimensionality reduction

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

http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

6. Profiling / Pattern Mining / Anomaly Detection (unsupervised)

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

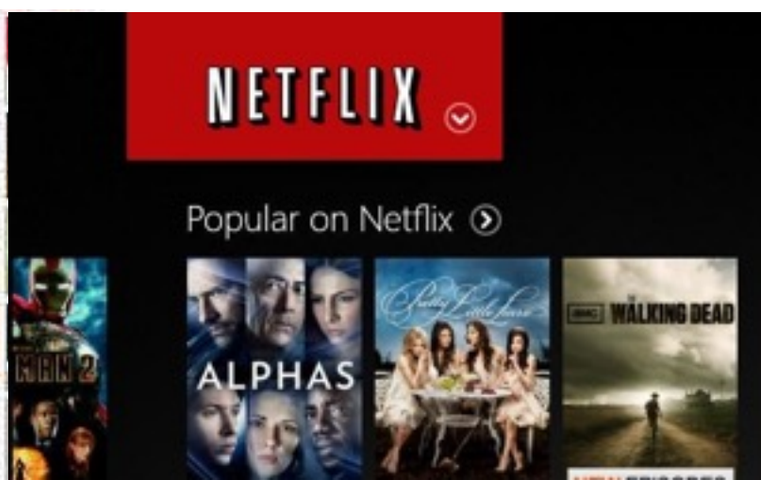
7. Link Prediction / Recommendation

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

linkedin/facebook: people you may know

amazon/netflix: because you like terminator…

suggest other movies you may also like



8. Data reduction ("dimensionality reduction")

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

1. if you want to visualize the data (in 2D/3D)

2. faster computation/less storage

3. reduce noise

Start Thinking About Project!

- What problems do you want to solve?
- Using what (large) datasets?
- What techniques do you need?