Data Mining Concepts & Tasks

Duen Horng (Polo) Chau
Georgia Tech

CSE6242 / CX4242
Sept 9, 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.

Continuing with

Data Integration

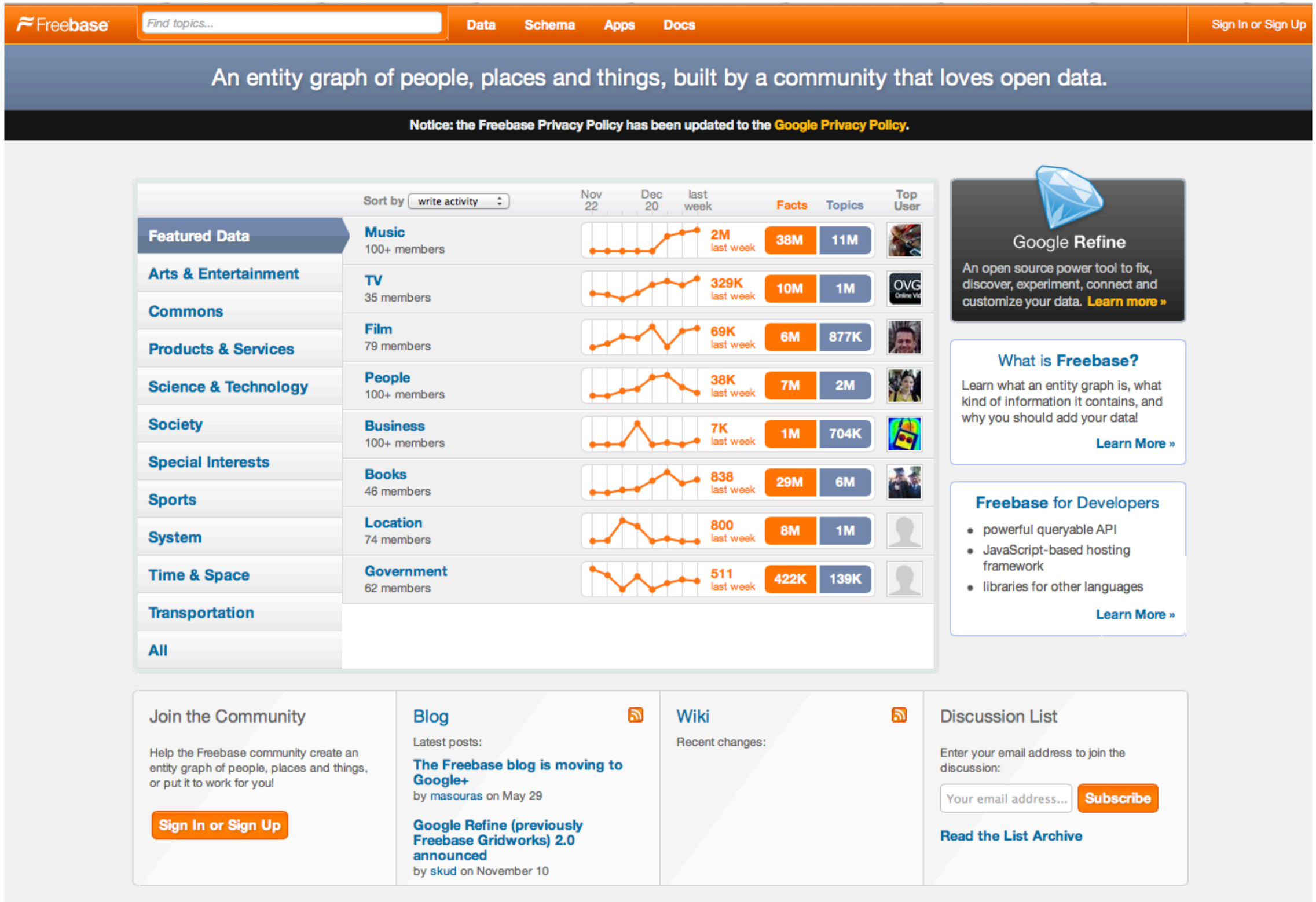
Freebase

(a graph of entities)

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

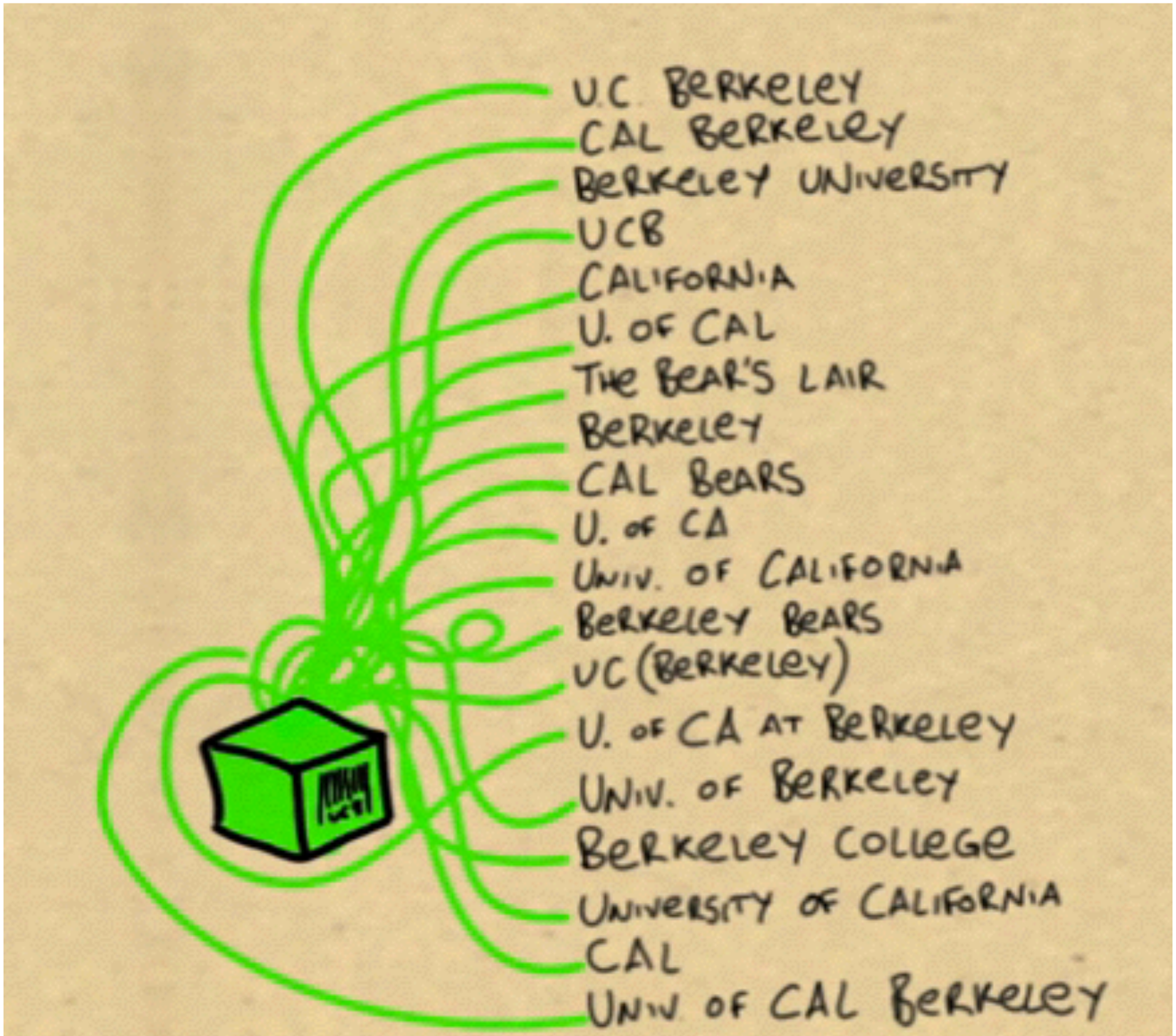
Wikipedia.

Crowd-sourcing Approaches: Freebase

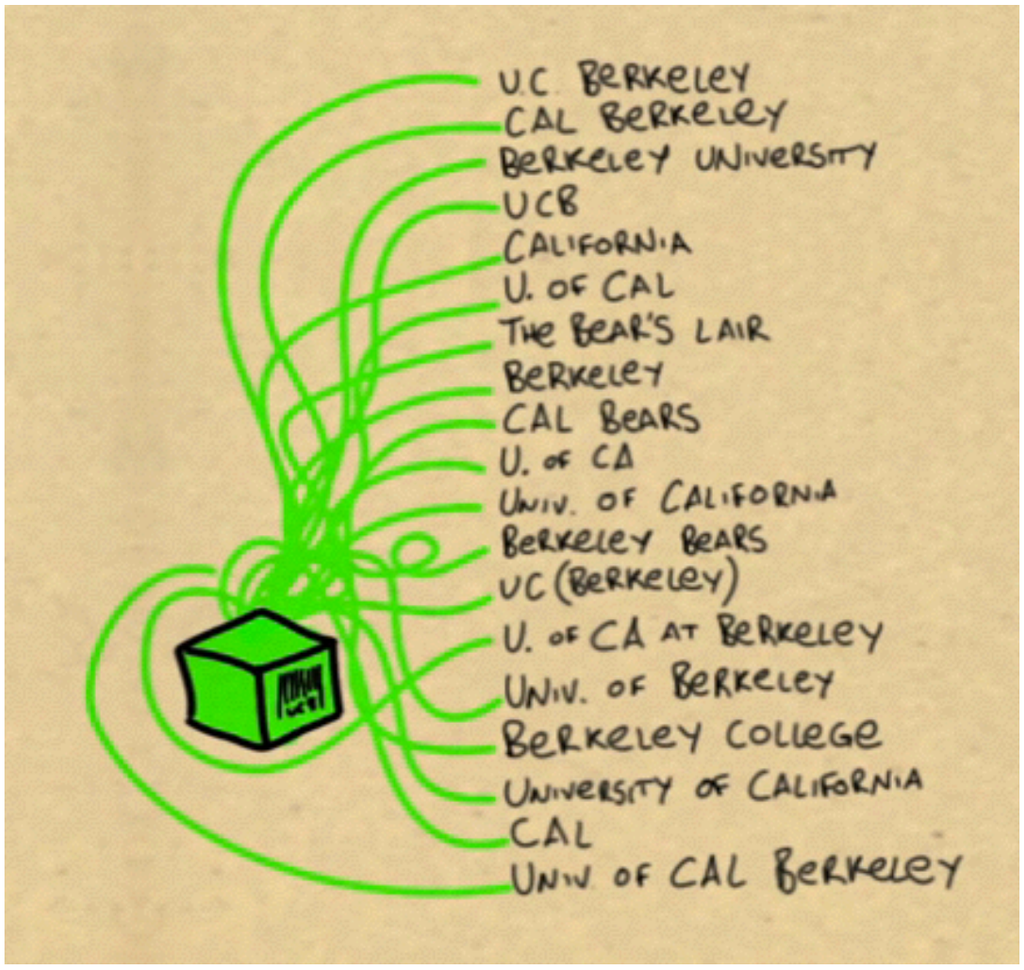


http://wiki.freebase.com/wiki/What_is_Freebase%3F

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?

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

person_id	name	state_id
1	Smith	111
2	Johnson	222
3	Obama	222

+

state_id	state_name
111	GA
222	NY
333	CA

person_id	name	state
1	Smith	GA
2	Johnson	NY
3	Obama	NY

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
- Free In-store pickup

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

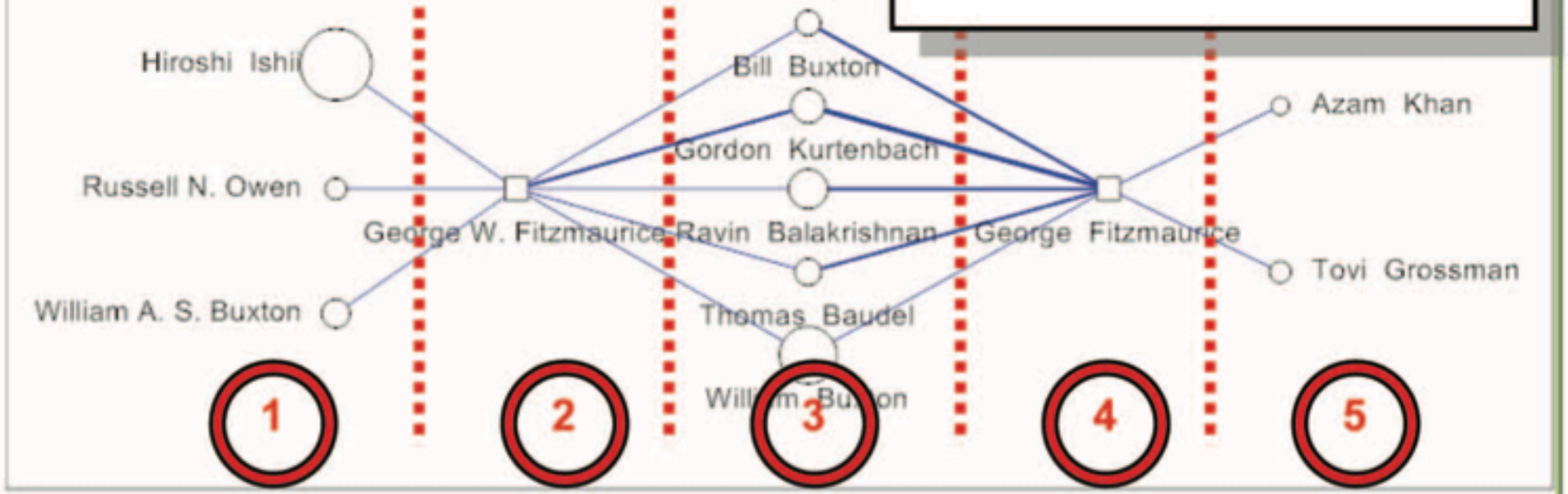
Search

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

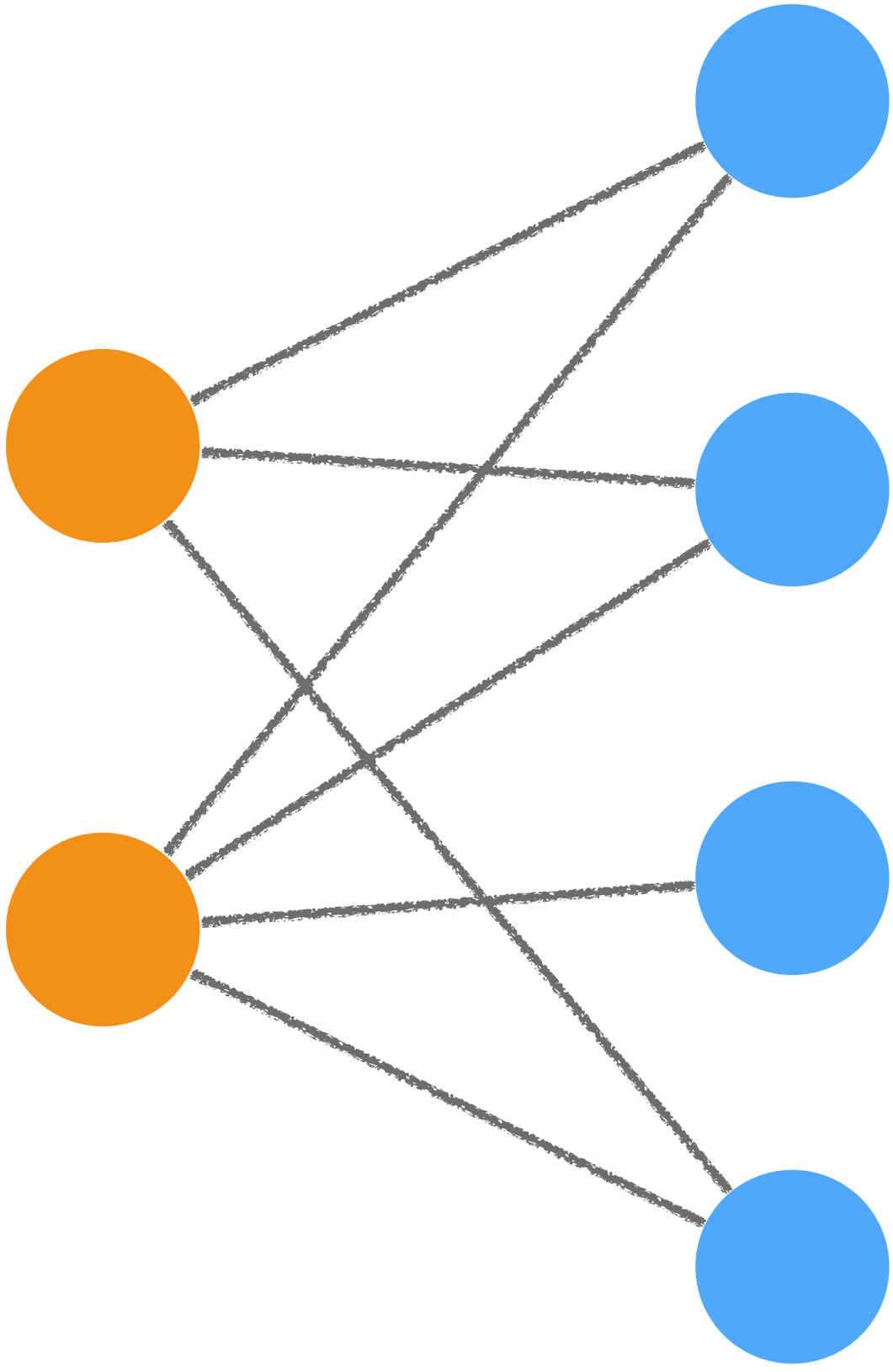
Data detail viewer

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

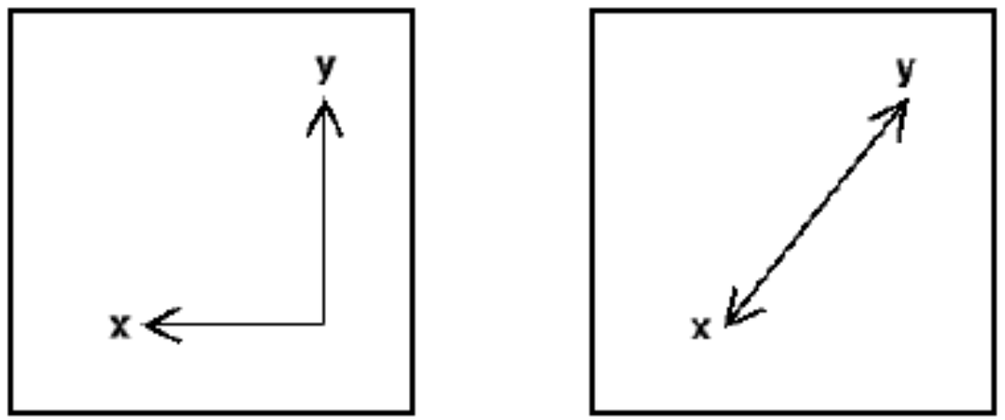
Poalo



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"
- Many more…



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

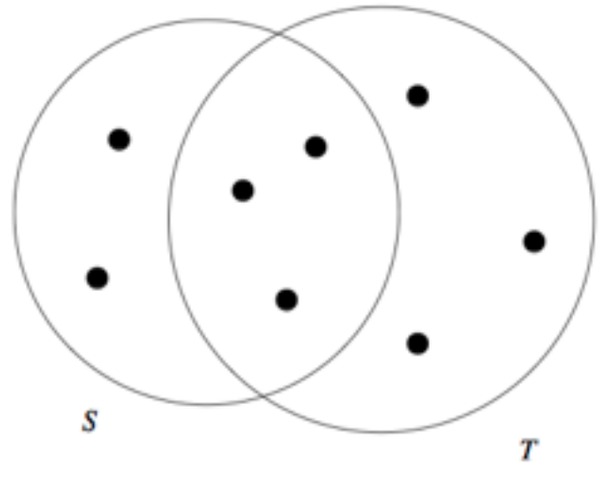


Figure 3.1: Two sets with Jaccard similarity 3/8

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)

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

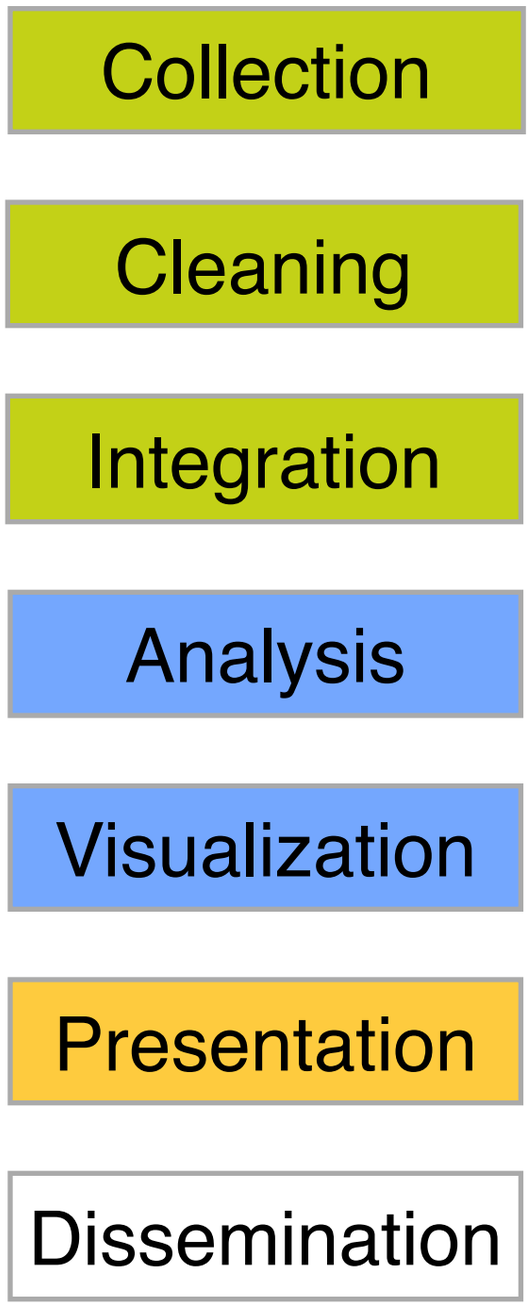
$$-1 \leq w_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{n} |w_k| = 1,$$

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



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

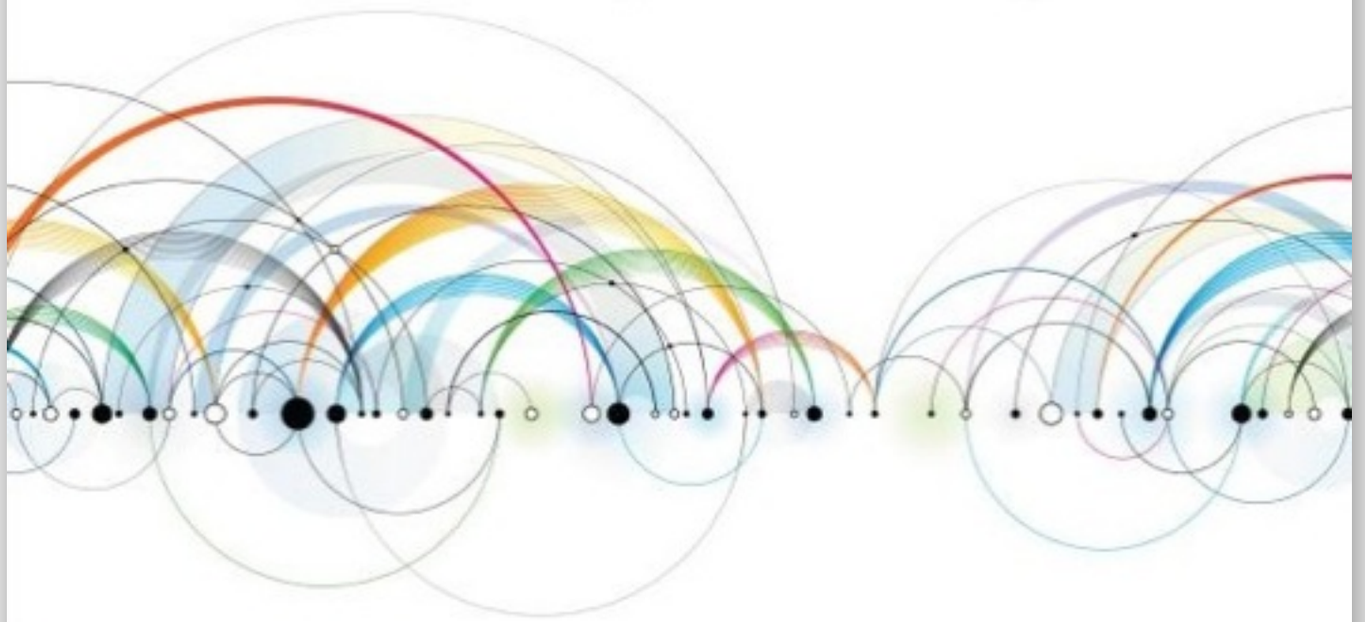
Good news: many (sub)tasks that underlie these problems are **common**

Here is an **overview** of the common tasks, based on *Data Science for Business: What you need to know about data mining and data-analytic thinking*

"A must-read resource for anyone who is serious about embracing the opportunity of big data."
—Craig Vaughan, Global Vice President, SAP

Data Science *for* Business

What You Need to Know About Data Mining and Data-Analytic Thinking



Foster Provost & Tom Fawcett

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

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?

stock prices

price of plane tickets

weather prediction

credit scores

time until machine fails (data center)

inventory management (supply chain)

population change (city, population planning)

sports stat (gambling)

3. Similarity Matching

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

Examples?
Online dating

recommendation systems (similar songs, movies)

image “classifier” (find all sunset images)

suggestions for online shopping

market segmentation

suggestion of friends on facebook

online advertisement

-> restaurant “classification” (italian, Chinese)

search results (google “similar” results)

search query matching

4. Clustering (unsupervised learning)

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

Examples?
factors for diseases

movie categories (genres; soft clustering)

market segmentation for targeted advertisement

social network analysis (whether people like the same thing)

geographical data (identify “neighborhood”, popular landmarks)

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 (unsupervised)

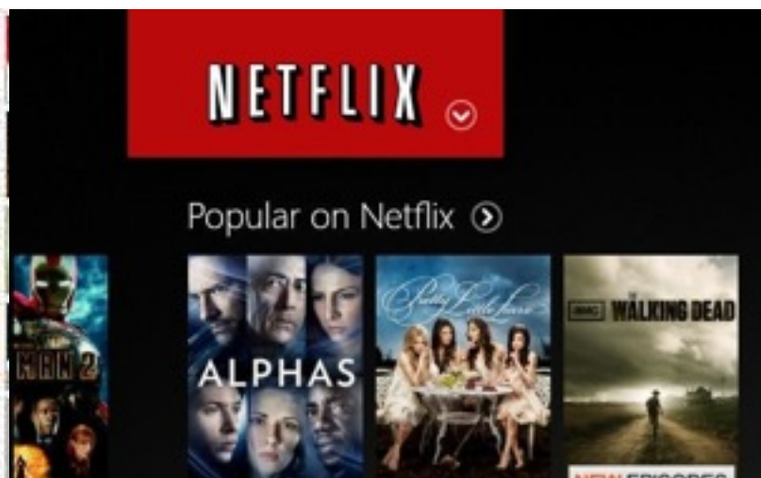
Characterize **typical** behaviors of an entity (person, computer router, etc.) so you can find **trends** and **outliers**.

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.

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?
- Will describe project requirements in next lecture