Lecture 13: Record Linkage

Core 109S In Data We Trust?, Spring 2017
Michael Hay

Logistics

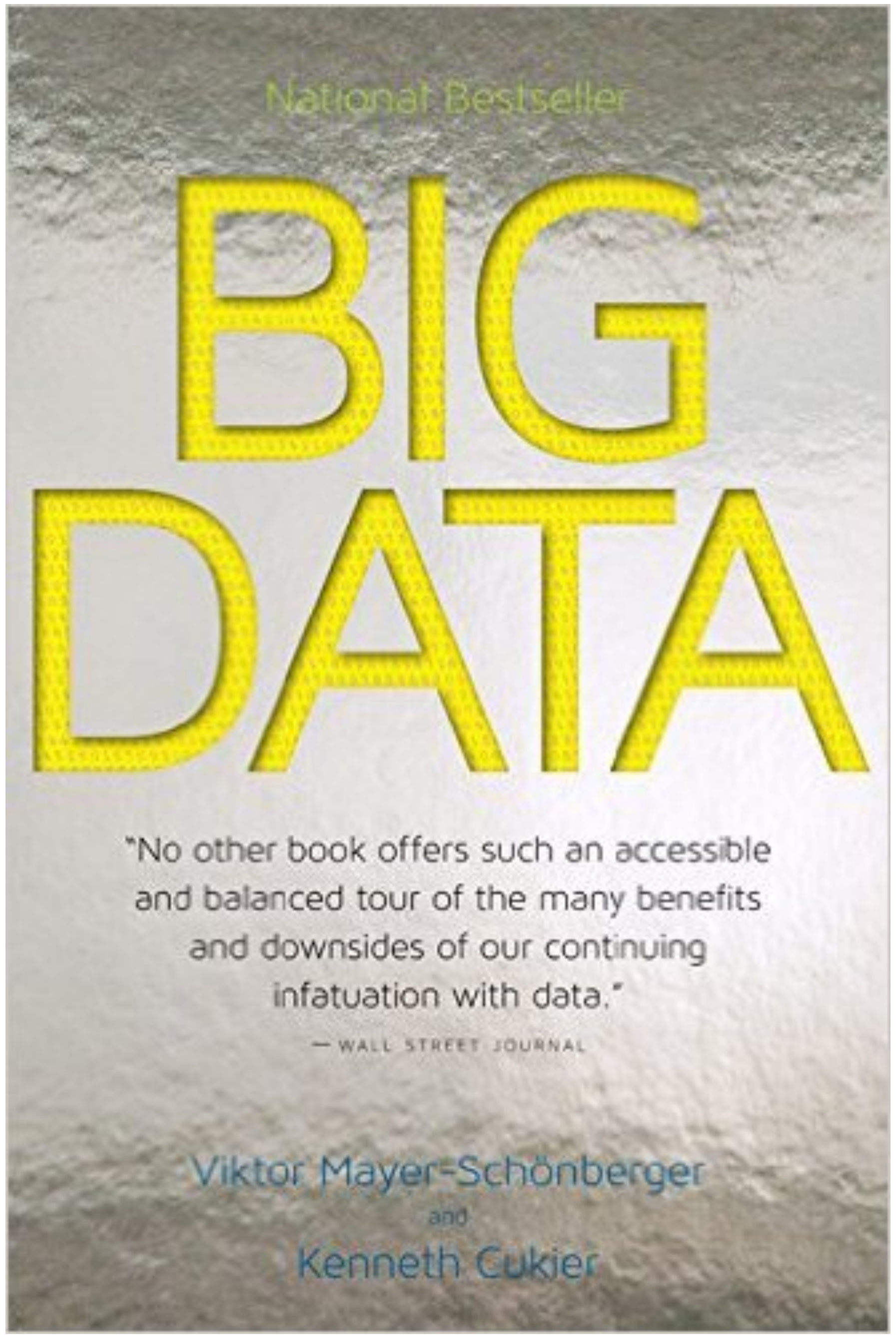
- HW 2 Coming out soon… algorithm analysis and record linkage

- Midterm exam will probably get pushed back (until after HW2 is due)

- Course schedule will be updated tonight with readings, etc.

Today

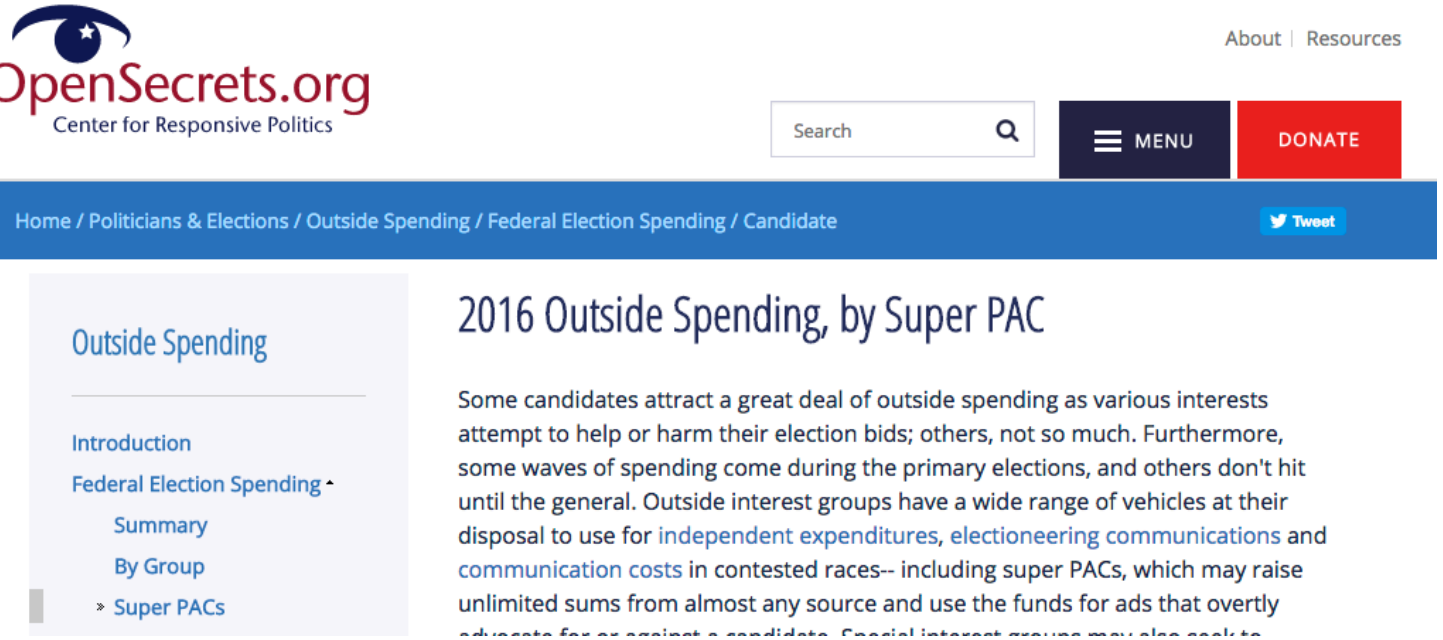
- Today we explore technical details related to two ideas from book
  - Messy: big data often means messy data
  - Recombinant data: such as Danish cancer patient data linked with mobile phone records


National Bestseller
BIG DATA
"No other book offers such an accessible and balanced tour of the many benefits and downsides of our continuing infatuation with data."
— WALL STREET JOURNAL
Viktor Mayer-Schönberger
and
Kenneth Cukier

Goals for today

- Record linkage (aka fuzzy matching)
- From this lecture, you should…
    - … have a general understanding of what record linkage is, and its potential "big data" applications
    - … understand what edit distance is
    - … be able to fill in a matrix of edit distance calculations and find the least cost sequence of edits

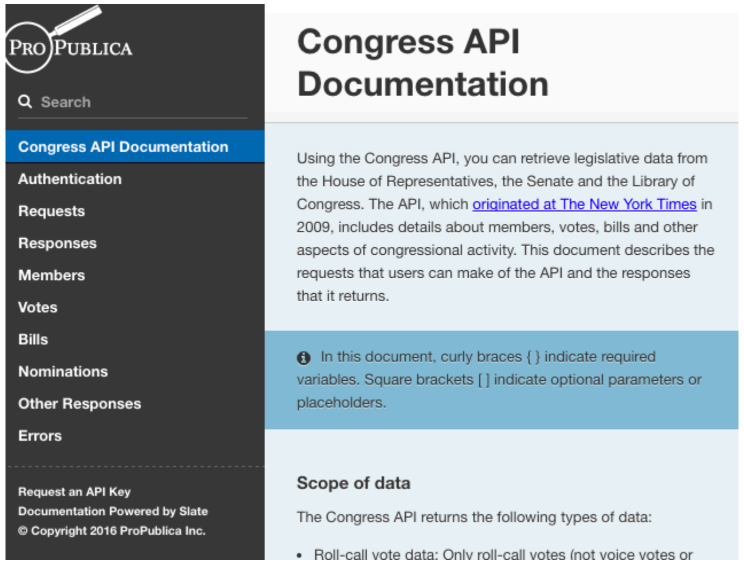
Motivating example


OpenSecrets.org
Center for Responsive Politics
About | Resources
Search
MENU
DONATE
Home / Politicians & Elections / Outside Spending / Federal Election Spending / Candidate
Tweet
Outside Spending
Introduction
Federal Election Spending
Summary
By Group
Super PACs

2016 Outside Spending, by Super PAC

Some candidates attract a great deal of outside spending as various interests attempt to help or harm their election bids; others, not so much. Furthermore, some waves of spending come during the primary elections, and others don't hit until the general. Outside interest groups have a wide range of vehicles at their disposal to use for independent expenditures, electioneering communications and communication costs in contested races-- including super PACs, which may raise unlimited sums from almost any source and use the funds for ads that overtly advocate for or against a candidate. Special interest groups may also seek to

https://www.opensecrets.org/outsidespending/summ.php



Congress API Documentation

Using the Congress API, you can retrieve legislative data from the House of Representatives, the Senate and the Library of Congress. The API, which originated at The New York Times in 2009, includes details about members, votes, bills and other aspects of congressional activity. This document describes the requests that users can make of the API and the responses that it returns.

In this document, curly braces { } indicate required variables. Square brackets [ ] indicate optional parameters or placeholders.

Scope of data

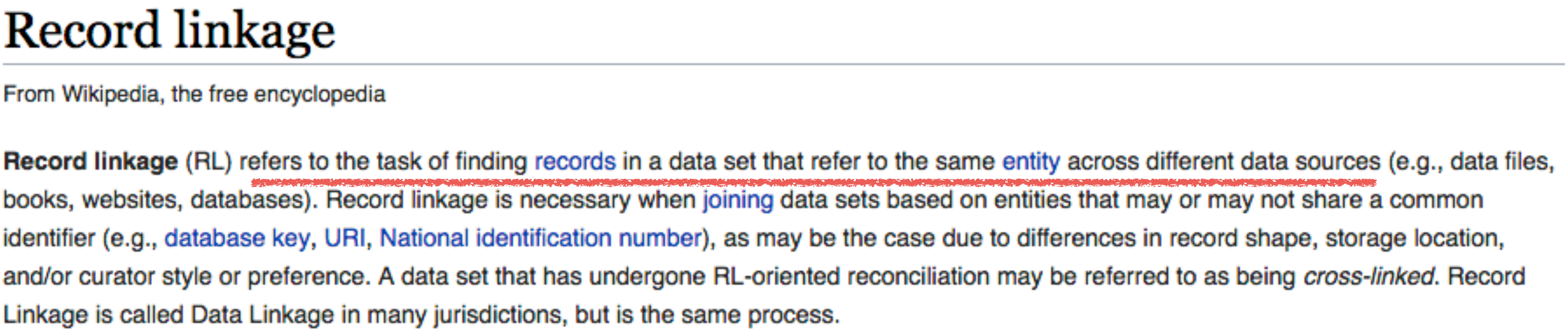
The Congress API returns the following types of data:

- Roll-call vote data: Only roll-call votes (not voice votes or

https://propublica.github.io/congress-api-docs/

- You have two different data sources, both describing the same set of entities (Congress members)
- You can put them both into the same relational database, but how do you join them? There is no key…

Record linkage



Record linkage

From Wikipedia, the free encyclopedia

Record linkage (RL) refers to the task of finding records in a data set that refer to the same entity across different data sources (e.g., data files, books, websites, databases). Record linkage is necessary when joining data sets based on entities that may or may not share a common identifier (e.g., database key, URI, National identification number), as may be the case due to differences in record shape, storage location, and/or curator style or preference. A data set that has undergone RL-oriented reconciliation may be referred to as being *cross-linked*. Record Linkage is called Data Linkage in many jurisdictions, but is the same process.

https://en.wikipedia.org/wiki/Record_linkage

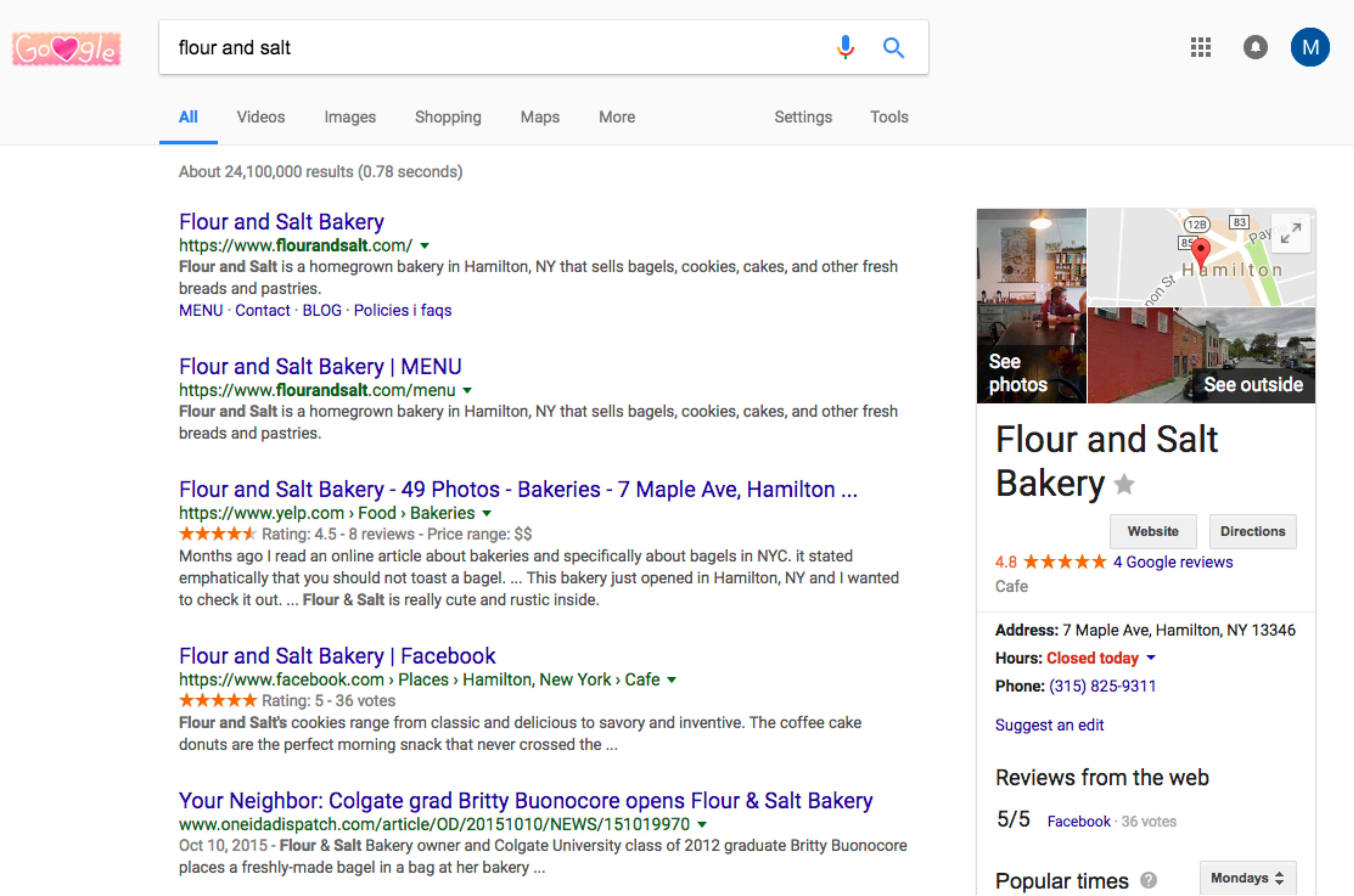
Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Ironically, record linkage has many names

- Coreference resolution
- Fuzzy match
- Deduplication
- Identity uncertainty
- Hardening soft databases
- Doubles
- Entity Resolution
- Reference reconciliation
- Object identification
- Approximate match
- Merge/purge
- Householding
- Duplicate detection
- Object consolidation
- Entity clustering
- Household matching
- Reference matching

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Motivating example: web



About 24,100,000 results (0.78 seconds)

Flour and Salt Bakery
https://www.flourandsalt.com/ ▾
Flour and Salt is a homegrown bakery in Hamilton, NY that sells bagels, cookies, cakes, and other fresh breads and pastries.
MENU · Contact · BLOG · Policies i faqs

Flour and Salt Bakery | MENU
https://www.flourandsalt.com/menu ▾
Flour and Salt is a homegrown bakery in Hamilton, NY that sells bagels, cookies, cakes, and other fresh breads and pastries.

Flour and Salt Bakery - 49 Photos - Bakeries - 7 Maple Ave, Hamilton ...
https://www.yelp.com › Food › Bakeries ▾
★★★★★ Rating: 4.5 - 8 reviews - Price range: $$
Months ago I read an online article about bakeries and specifically about bagels in NYC. it stated emphatically that you should not toast a bagel. ... This bakery just opened in Hamilton, NY and I wanted to check it out. ... **Flour** & **Salt** is really cute and rustic inside.

Flour and Salt Bakery | Facebook
https://www.facebook.com › Places › Hamilton, New York › Cafe ▾
★★★★★ Rating: 5 - 36 votes
Flour and Salt's cookies range from classic and delicious to savory and inventive. The coffee cake donuts are the perfect morning snack that never crossed the ...

Your Neighbor: Colgate grad Britty Buonocore opens Flour & Salt Bakery
www.oneidadispatch.com/article/OD/20151010/NEWS/151019970 ▾
Oct 10, 2015 - **Flour** & **Salt** Bakery owner and Colgate University class of 2012 graduate Britty Buonocore places a freshly-made bagel in a bag at her bakery ...

Flour and Salt Bakery ★

4.8 ★★★★★ 4 Google reviews
Cafe

Address: 7 Maple Ave, Hamilton, NY 13346
Hours: Closed today ▾
Phone: (315) 825-9311

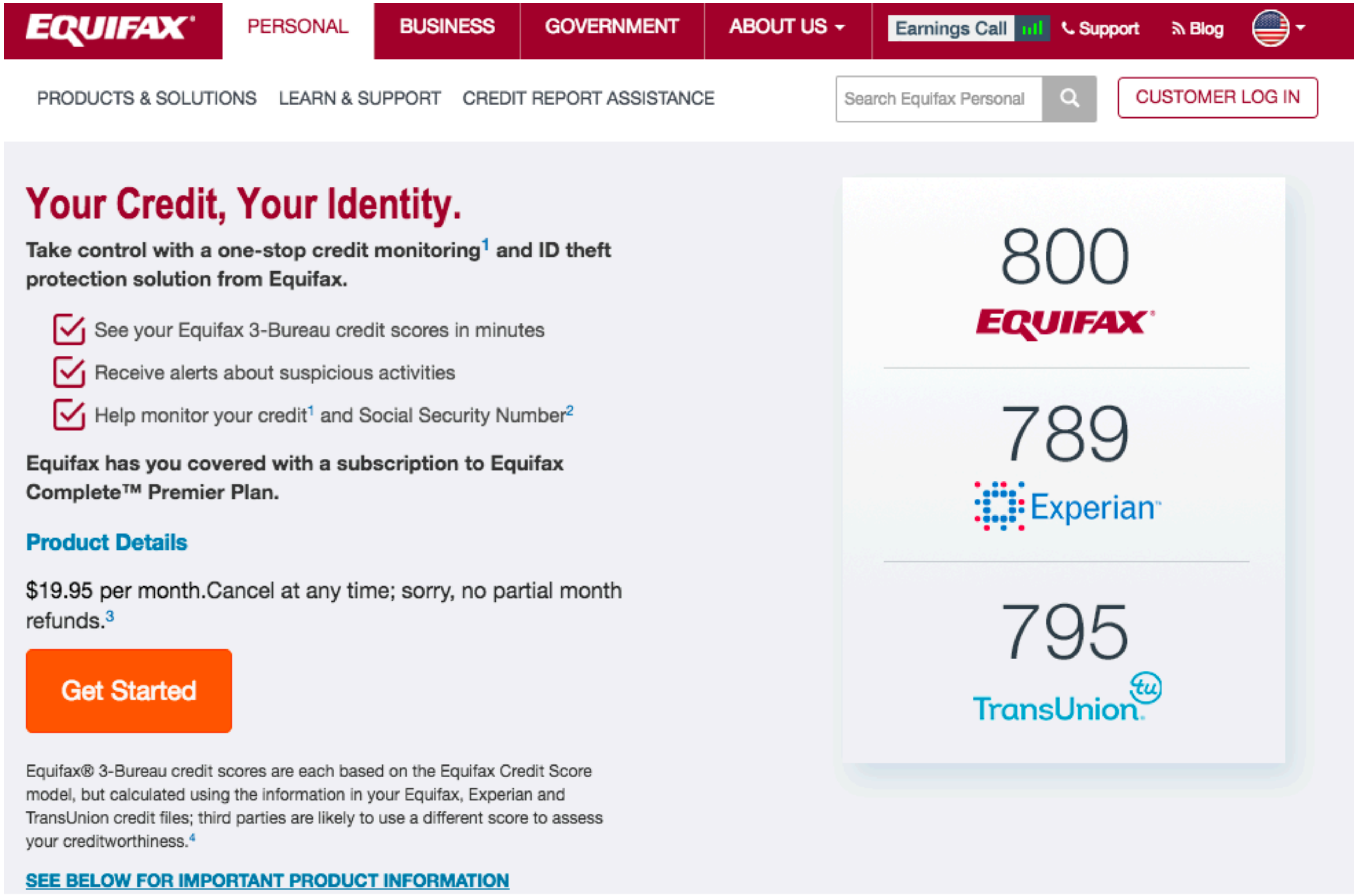
Suggest an edit

Reviews from the web

5/5 Facebook · 36 votes

Popular times ?

Motivating example: credit reports



Your Credit, Your Identity.

Take control with a one-stop credit monitoring[1] and ID theft protection solution from Equifax.

- ☑ See your Equifax 3-Bureau credit scores in minutes
- ☑ Receive alerts about suspicious activities
- ☑ Help monitor your credit[1] and Social Security Number[2]

Equifax has you covered with a subscription to Equifax Complete™ Premier Plan.

Product Details

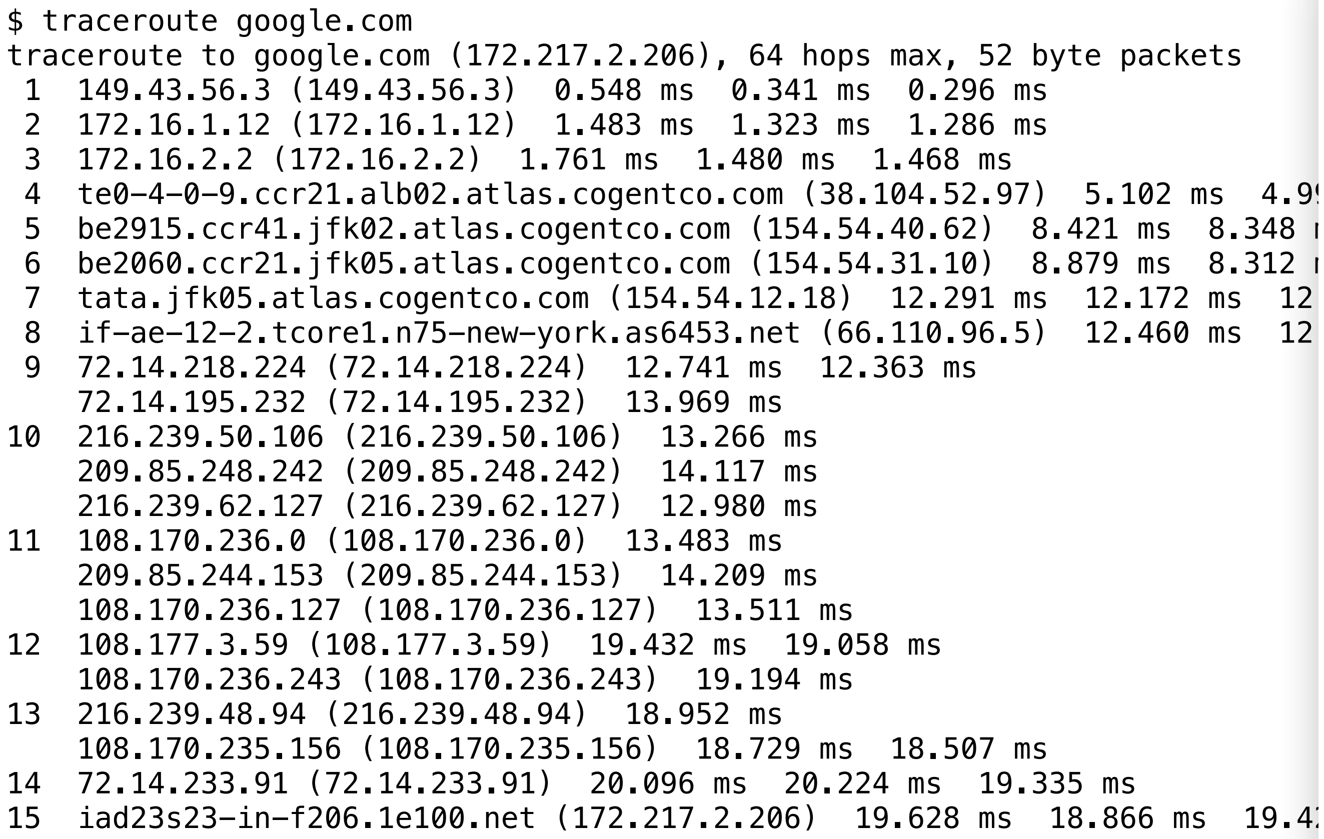
$19.95 per month.Cancel at any time; sorry, no partial month refunds.[3]

Equifax® 3-Bureau credit scores are each based on the Equifax Credit Score model, but calculated using the information in your Equifax, Experian and TransUnion credit files; third parties are likely to use a different score to assess your creditworthiness.[4]

SEE BELOW FOR IMPORTANT PRODUCT INFORMATION

Motivating example: networks

Measuring topology of internet using traceroute. IP aliasing problem:



```
$ traceroute google.com
traceroute to google.com (172.217.2.206), 64 hops max, 52 byte packets
 1  149.43.56.3 (149.43.56.3)  0.548 ms  0.341 ms  0.296 ms
 2  172.16.1.12 (172.16.1.12)  1.483 ms  1.323 ms  1.286 ms
 3  172.16.2.2 (172.16.2.2)  1.761 ms  1.480 ms  1.468 ms
 4  te0-4-0-9.ccr21.alb02.atlas.cogentco.com (38.104.52.97)  5.102 ms  4.9
 5  be2915.ccr41.jfk02.atlas.cogentco.com (154.54.40.62)  8.421 ms  8.348
 6  be2060.ccr21.jfk05.atlas.cogentco.com (154.54.31.10)  8.879 ms  8.312
 7  tata.jfk05.atlas.cogentco.com (154.54.12.18)  12.291 ms  12.172 ms  12
 8  if-ae-12-2.tcore1.n75-new-york.as6453.net (66.110.96.5)  12.460 ms  12
 9  72.14.218.224 (72.14.218.224)  12.741 ms  12.363 ms
    72.14.195.232 (72.14.195.232)  13.969 ms
10  216.239.50.106 (216.239.50.106)  13.266 ms
    209.85.248.242 (209.85.248.242)  14.117 ms
    216.239.62.127 (216.239.62.127)  12.980 ms
11  108.170.236.0 (108.170.236.0)  13.483 ms
    209.85.244.153 (209.85.244.153)  14.209 ms
    108.170.236.127 (108.170.236.127)  13.511 ms
12  108.177.3.59 (108.177.3.59)  19.432 ms  19.058 ms
    108.170.236.243 (108.170.236.243)  19.194 ms
13  216.239.48.94 (216.239.48.94)  18.952 ms
    108.170.235.156 (108.170.235.156)  18.729 ms  18.507 ms
14  72.14.233.91 (72.14.233.91)  20.096 ms  20.224 ms  19.335 ms
15  iad23s23-in-f206.1e100.net (172.217.2.206)  19.628 ms  18.866 ms  19.4
```

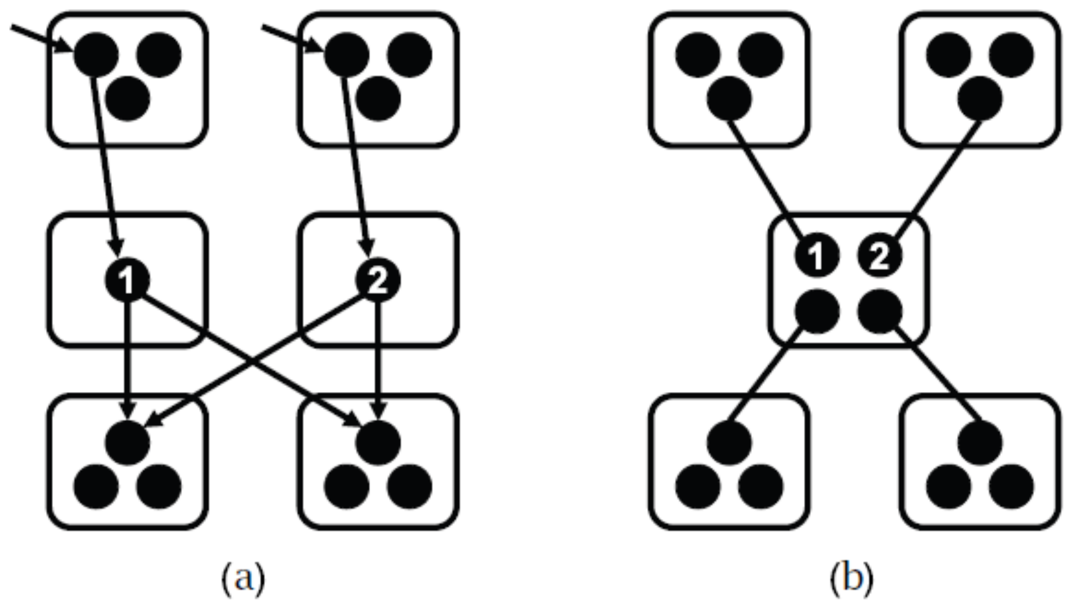
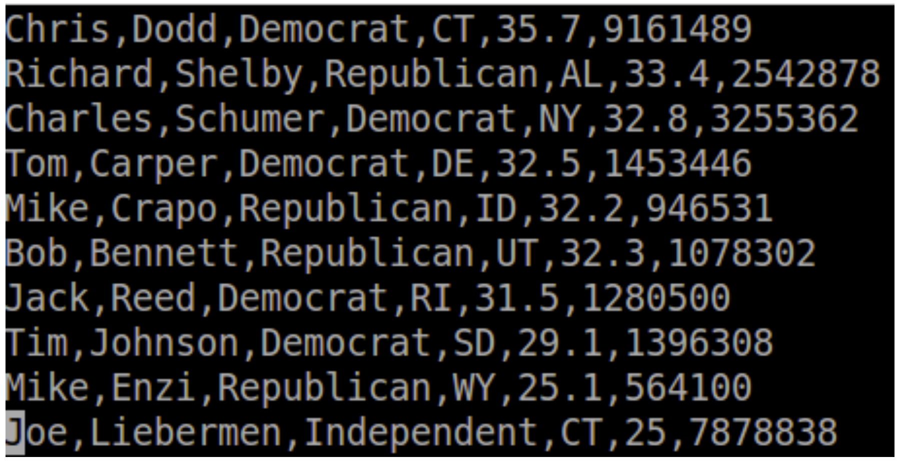


Figure 2. The IP alias resolution problem. Paraphrasing Fig. 4 of [50], traceroute does not list routers (boxes) along paths but IP addresses of input interfaces (circles), and alias resolution refers to the correct mapping of interfaces to routers to reveal the actual topology. In the case where interfaces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

[Willinger et al. 2009]

Back to example

- How to link senator's records in two different data sources.
  - Join on (firstname, lastname)?
    - *Too specific* ("Joe" vs. "Joseph")
  - Join on just last name?
    - *Too inclusive* ("Smith")
  - Where is "Joe Liebermen"?
    - *Spelling mistakes, etc. Want approximate matching!*



```
Chris,Dodd,Democrat,CT,35.7,9161489
Richard,Shelby,Republican,AL,33.4,2542878
Charles,Schumer,Democrat,NY,32.8,3255362
Tom,Carper,Democrat,DE,32.5,1453446
Mike,Crapo,Republican,ID,32.2,946531
Bob,Bennett,Republican,UT,32.3,1078302
Jack,Reed,Democrat,RI,31.5,1280500
Tim,Johnson,Democrat,SD,29.1,1396308
Mike,Enzi,Republican,WY,25.1,564100
Joe,Liebermen,Independent,CT,25,7878838
```

Levenshtein (or edit) distance

- The minimum number of character edit operations needed to turn one string ("string" = an array of characters) into the other.

  LIEBERMAN

  LIEBERMEN

- Substitute A to E. Edit distance = 1

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Levenshtein (or edit) distance

- Distance between two string *s* and *t* is the lowest cost sequence of edit commands that transform s to t.
- Edit commands
  - Copy character from *s* to *t*
  - Delete a character from *s*
    - Ex: *s* = "Joey" and *t* = "Joe"
  - Insert a character into *s*
    - Ex: *s* = "Hilary" and *t* = "Hillary"
  - Substitute one character for another
    - Ex: *s* = "Smyth" and *t* = "Smith"

	Cost
Copy	0
Delete	1
Insert	1
Sub	1

In general, costs could be different

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

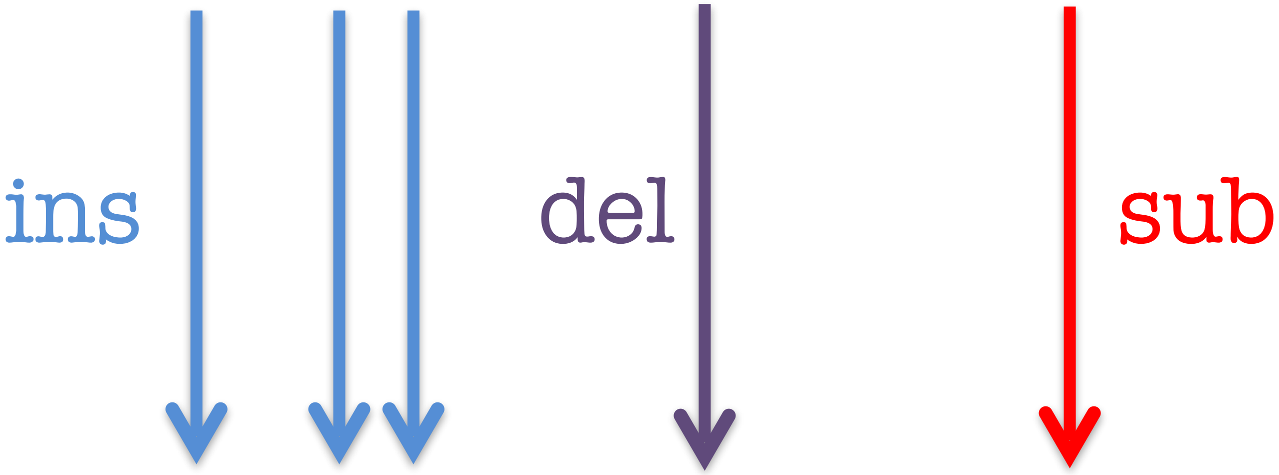
Example

s = `Joe Liebermen`

t = `Joseph Liberman`

Example

s = Jo_e__ Liebermen



t = Joseph Li_berman

Total cost: 3 + 1 + 1 = 5

5 min. break

Computing edit distance

- Two key observations

  1. We can contemplate edit distance between any substrings of *s* and *t*

     `cost(i,j)` = edit distance between `s[1..i]` and `t[1..j]`

Computing edit distance

t = "_Joey"

s = "_Joseph"

	_	J	o	e	y
_					
J					
o					
s					
e					
p					
h					

Cost of changing
_ ➡ _J

Cost of changing
_Jose ➡ _Jo

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Computing edit distance

- Two key observations

    1. We can contemplate edit distance between any substrings of *s* and *t*

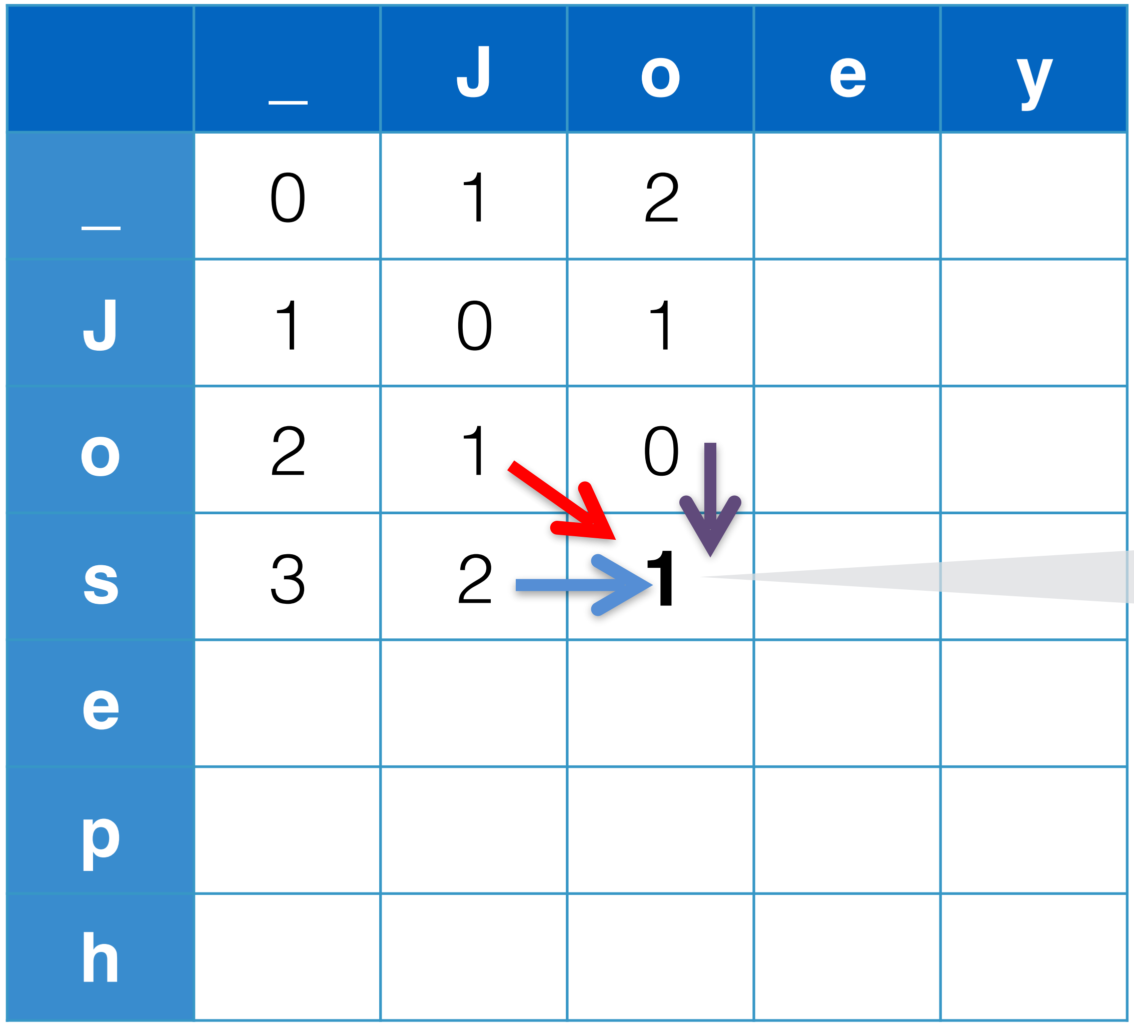
       `cost(i,j)` = edit distance between `s[1..i]` and `t[1..j]`

    2. To compute cost(i,j), focus on effect of last edit command

Computing edit distance

t = "_Joey"

s = "_Joseph"



	_	J	o	e	y
_	0	1	2		
J	1	0	1		
o	2	1	0		
s	3	2	**1**		
e					
p					
h					

Cost of changing `_Jos` ➡ `_Jo.` Last edit command could be:

- Delete s: 1+ Cost(_Jo ➡ _Jo)
- Insert o: 1 + Cost(_Jos ➡ _J)
- Sub s with o: 1 + Cost(_Jo ➡ _J)

Set Cost(_Jos ➡ _Jo) to be the *minimum* of these options.

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Computing edit distance

t = "_Joey"

s = "_Joseph"

	_	J	o	e	y
_	0	1	2	3	4
J	1	0	1	2	3
o	2	1	0	1	2
s	3	2	1	1	2
e	4	3	2	1	2
p	5	4	3	2	2
h	6	5	4	3	3

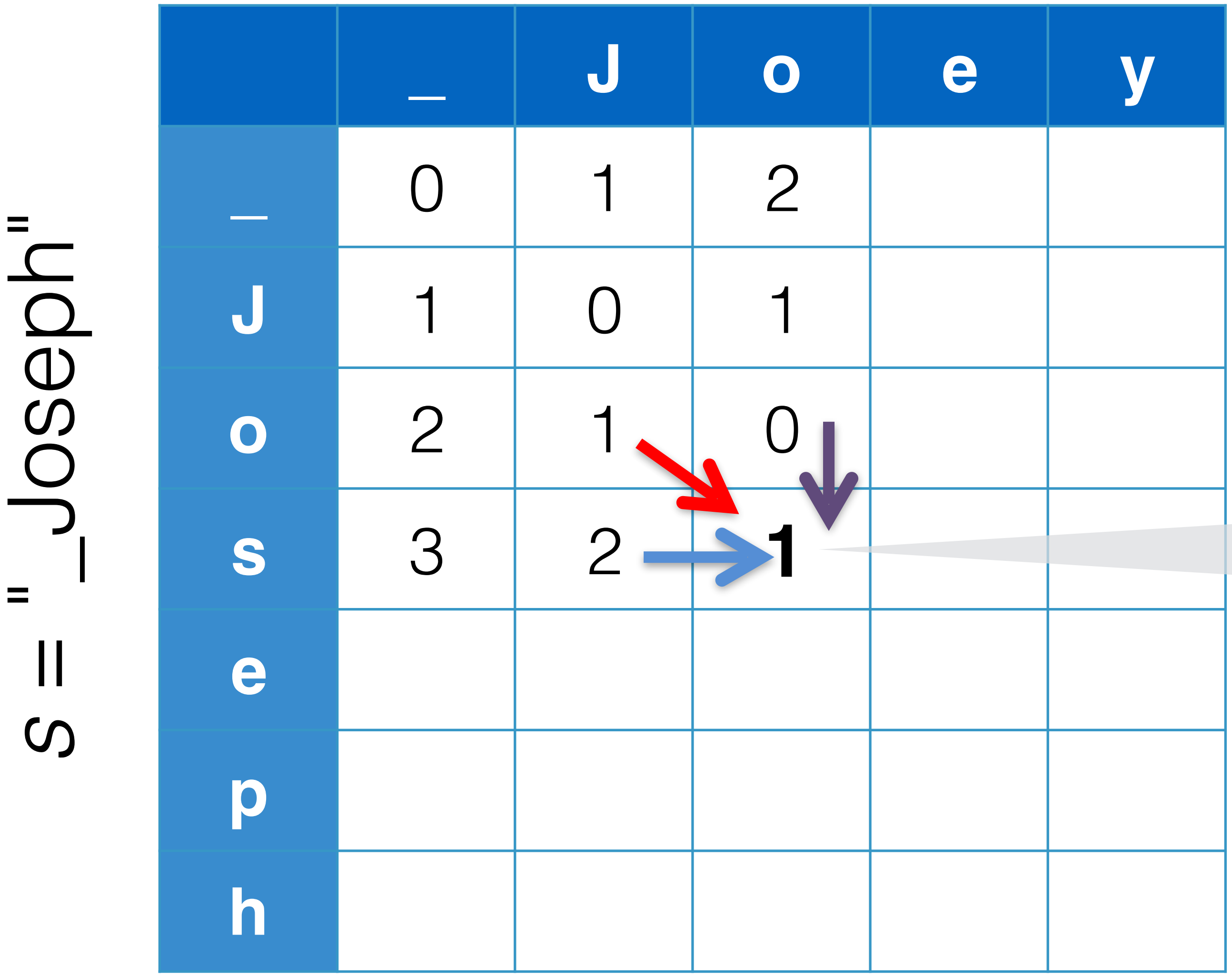
Exercise

Instructions: ~work in groups of 2-3 to fill this in; then I will call on one of you

Compute edit distance between s = "_BCD" and t="_ABC". (The Joey example below is just for reference.)

t = "_Joey"

s = "_Joseph"



	_	J	o	e	y
_	0	1	2		
J	1	0	1		
o	2	1	0		
s	3	2	**1**		
e					
p					
h					

Cost of changing _Jos ➡ _Jo. Last edit command could be:

- Delete s: 1+ Cost(_Jo ➡ _Jo)
- Insert o: 1 + Cost(_Jos ➡ _J)
- Sub s with o: 1 + Cost(_Jo ➡ _J)

Set Cost(_Jos ➡ _Jo) to be the *minimum* of these options.

Computing edit distance

t = "_Joey"

s = "_Joseph"

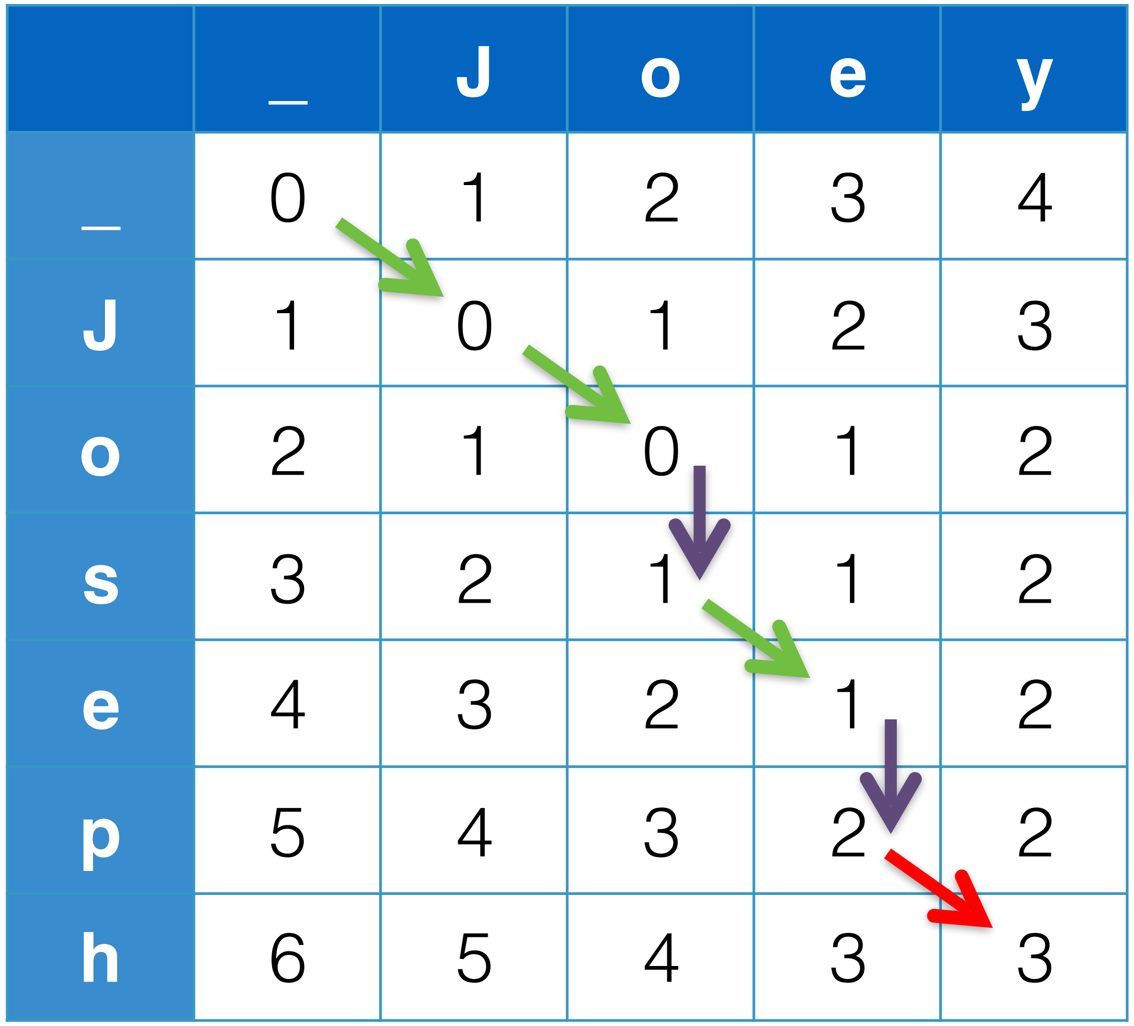
	_	J	o	e	y
_	0	1	2	3	4
J	1	0	1	2	3
o	2	1	0	1	2
s	3	2	1	1	2
e	4	3	2	1	2
p	5	4	3	2	2
h	6	5	4	3	3

The edit distance between `Joseph` and `Joey` is 3, *but which edit commands achieve this?*

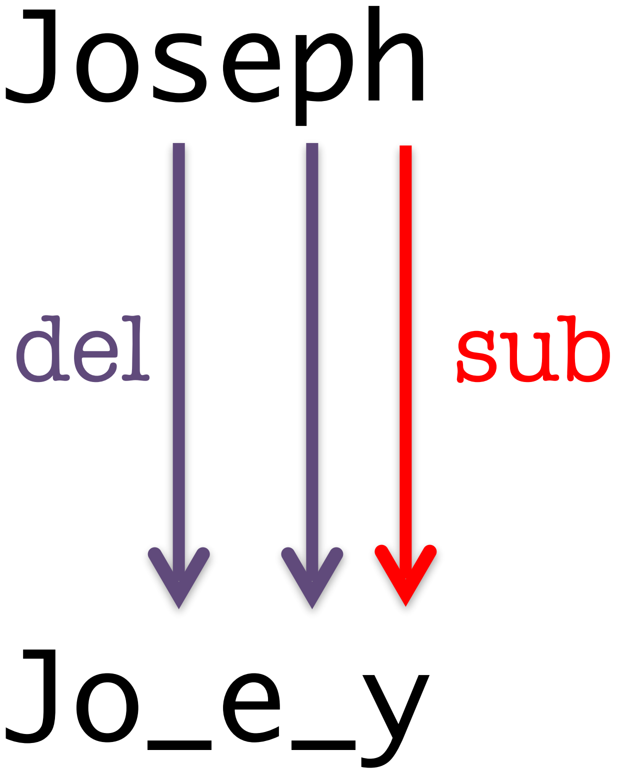
Computing edit distance

t = "_Joey"

s = "_Joseph"



	_	J	o	e	y
_	0	1	2	3	4
J	1	0	1	2	3
o	2	1	0	1	2
s	3	2	1	1	2
e	4	3	2	1	2
p	5	4	3	2	2
h	6	5	4	3	3



Joseph

del

sub

Jo_e_y

Remember the minimum in each step and retrace your path.

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

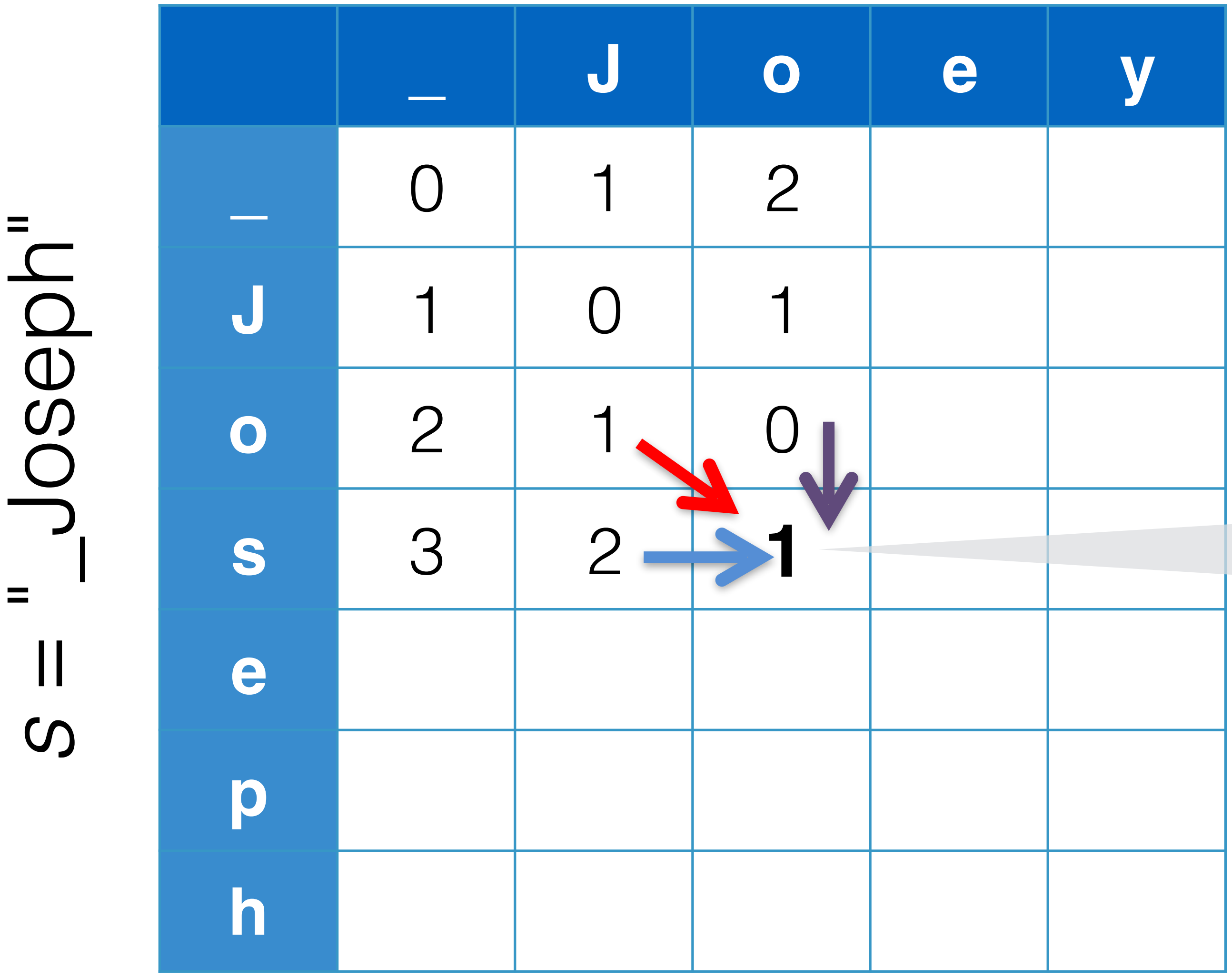
Exercise

Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you

Return to your previous example and note the minimum cost edit command at each step. (The Joey example below is just for reference.)

t = "_Joey"

s = "_Joseph"



	_	J	o	e	y
_	0	1	2		
J	1	0	1		
o	2	1	0		
s	3	2	**1**		
e					
p					
h					

Cost of changing `_Jos` ➡ `_Jo.` Last edit command could be:

- Delete s: 1+ Cost(`_Jo` ➡ `_Jo`)
- Insert o: 1 + Cost(`_Jos` ➡ `_J`)
- Sub s with o: 1 + Cost(`_Jo` ➡ `_J`)

Set Cost(`_Jos` ➡ `_Jo`) to be the *minimum* of these options.

Applying edit distance

- Back to motivating example: joining data about senators
- Some databases (e.g. postgresql) have built-in support for edit distance.
- Compute edit distance between firstname fields, and between last name fields
- Consider match if sum of distances below some *threshold*.
- Obviously, errors are possible: https://youtu.be/aRrDsbUdY_k?t=371

Edit distance variants

- Needleman-Wunsch
  - Different costs for each operation
- Affine gap penalty
  - "Joe Lieberman" vs. "Jo**seph I.** Lieberman"
  - Penalty for consecutive inserts: penalty for first insert + smaller penalty for each subsequent insert
- Edit distance has numerous applications (especially bioinformatics)

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Other Similarity Methods

Easiest and most efficient

- Equality on a boolean predicate
- Edit distance
  - Levenshtein, Affine
- Set similarity
  - Jaccard
- Vector Based
  - Cosine similarity, TFIDF

- Translation-based
- Numeric distance between values
- Phonetic Similarity
  - Soundex, Metaphone
- Other
  - Jaro-Winkler, Soft-TFIDF, Monge-Elkan

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Summary of Similarity Methods

Handle Typographical errors

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

Good for Text (reviews/ tweets), sets, class membership, …

Useful for abbreviations, alternate names.

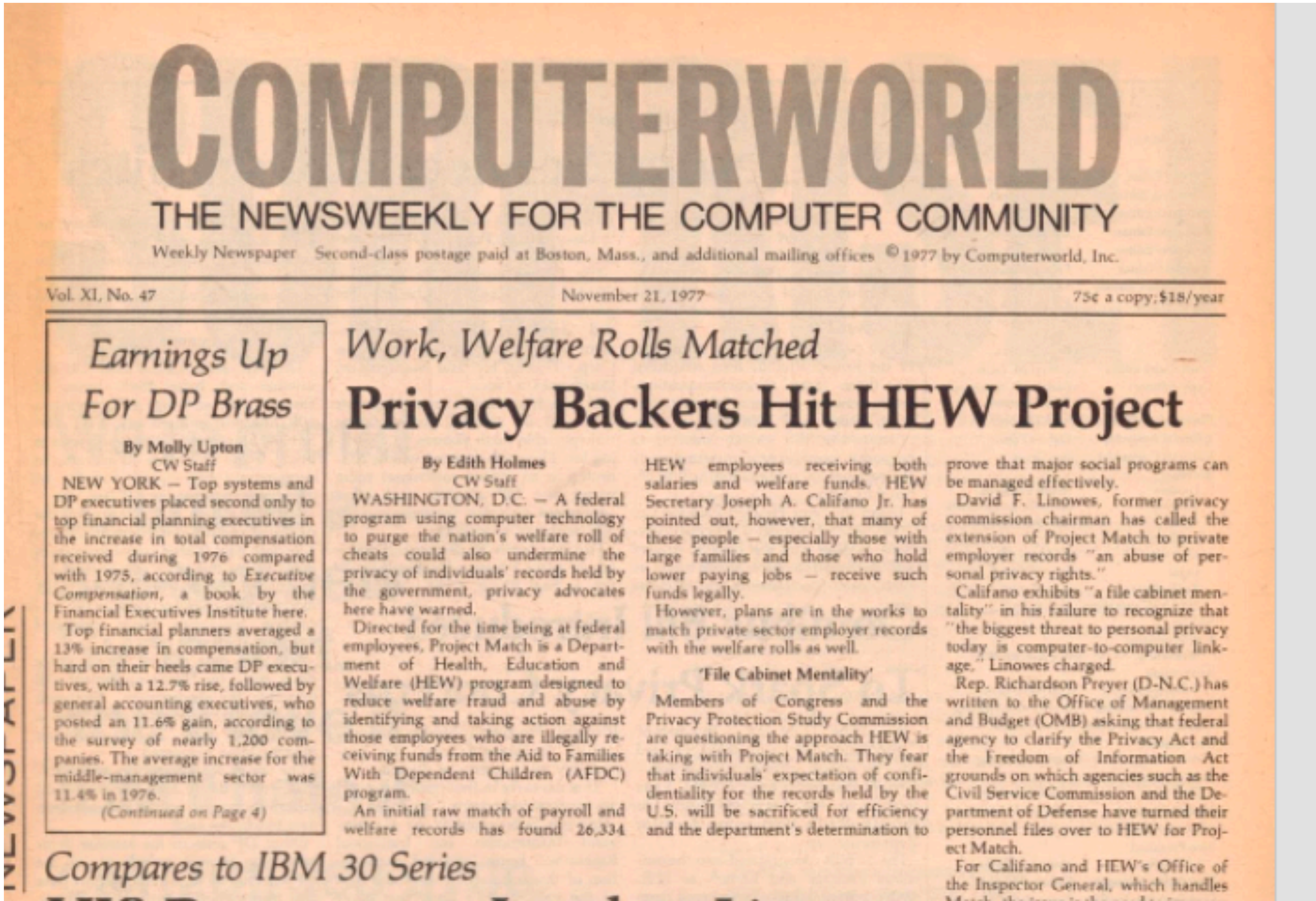
- Translation-based
- Numeric distance between values
- Phonetic Similarity
  - Soundex, Metaphone
- Other
  - Jaro-Winkler, Soft-TFIDF, Monge-Elkan

Good for Names

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Ugly side of record linkage

- What was this story about? Discussion.



COMPUTERWORLD

THE NEWSWEEKLY FOR THE COMPUTER COMMUNITY

Weekly Newspaper   Second-class postage paid at Boston, Mass., and additional mailing offices ©1977 by Computerworld, Inc.

Vol. XI, No. 47   November 21, 1977   75¢ a copy; $18/year

Earnings Up For DP Brass

By Molly Upton
CW Staff

NEW YORK – Top systems and DP executives placed second only to top financial planning executives in the increase in total compensation received during 1976 compared with 1975, according to *Executive Compensation*, a book by the Financial Executives Institute here.

Top financial planners averaged a 13% increase in compensation, but hard on their heels came DP executives, with a 12.7% rise, followed by general accounting executives, who posted an 11.6% gain, according to the survey of nearly 1,200 companies. The average increase for the middle-management sector was 11.4% in 1976.

(Continued on Page 4)

Work, Welfare Rolls Matched

Privacy Backers Hit HEW Project

By Edith Holmes
CW Staff

WASHINGTON, D.C. – A federal program using computer technology to purge the nation's welfare roll of cheats could also undermine the privacy of individuals' records held by the government, privacy advocates here have warned.

Directed for the time being at federal employees, Project Match is a Department of Health, Education and Welfare (HEW) program designed to reduce welfare fraud and abuse by identifying and taking action against those employees who are illegally receiving funds from the Aid to Families With Dependent Children (AFDC) program.

An initial raw match of payroll and welfare records has found 26,334 HEW employees receiving both salaries and welfare funds. HEW Secretary Joseph A. Califano Jr. has pointed out, however, that many of these people – especially those with large families and those who hold lower paying jobs – receive such funds legally.

However, plans are in the works to match private sector employer records with the welfare rolls as well.

'File Cabinet Mentality'

Members of Congress and the Privacy Protection Study Commission are questioning the approach HEW is taking with Project Match. They fear that individuals' expectation of confidentiality for the records held by the U.S. will be sacrificed for efficiency and the department's determination to prove that major social programs can be managed effectively.

David F. Linowes, former privacy commission chairman has called the extension of Project Match to private employer records "an abuse of personal privacy rights."

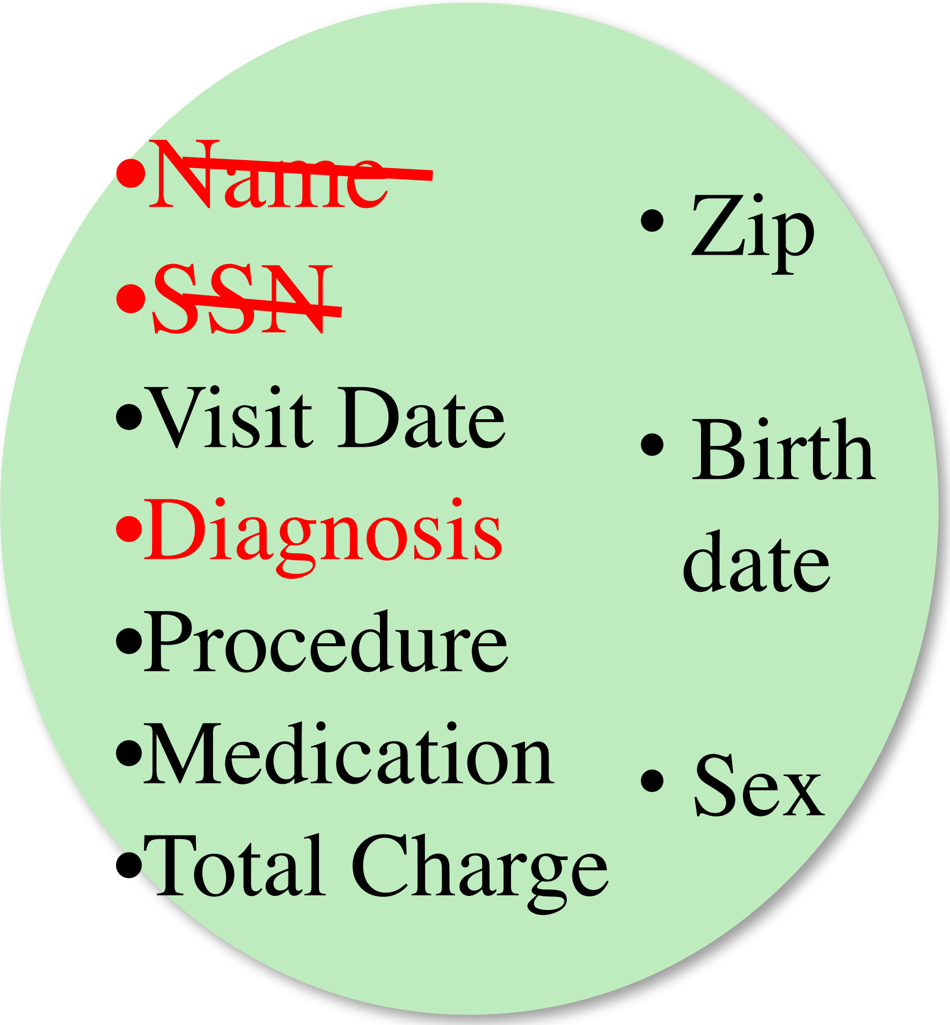
Califano exhibits "a file cabinet mentality" in his failure to recognize that "the biggest threat to personal privacy today is computer-to-computer linkage," Linowes charged.

Rep. Richardson Preyer (D-N.C.) has written to the Office of Management and Budget (OMB) asking that federal agency to clarify the Privacy Act and the Freedom of Information Act grounds on which agencies such as the Civil Service Commission and the Department of Defense have turned their personnel files over to HEW for Project Match.

For Califano and HEW's Office of the Inspector General, which handles

Compares to IBM 30 Series

The Ugly side of Record Linkage
[Sweeney IJUFKS 2002]



- ~~Name~~
- ~~SSN~~
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

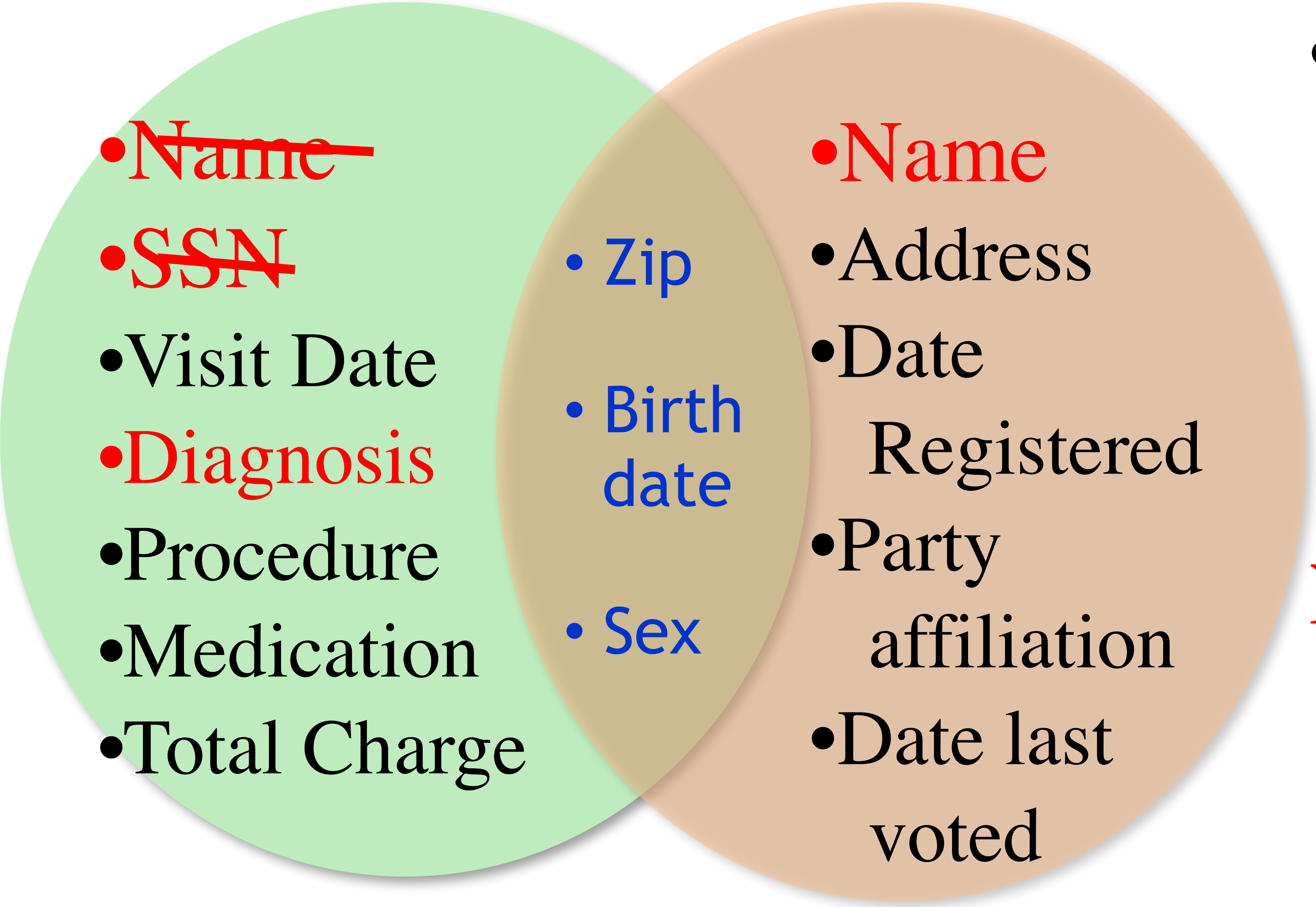
- Zip
- Birth date
- Sex

Medical Data

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

The Ugly side of Record Linkage

[Sweeney IJUFKS 2002]



- ~~Name~~
- ~~SSN~~
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

- Zip
- Birth date
- Sex

- Name
- Address
- Date Registered
- Party affiliation
- Date last voted

Medical Data **Voter List**

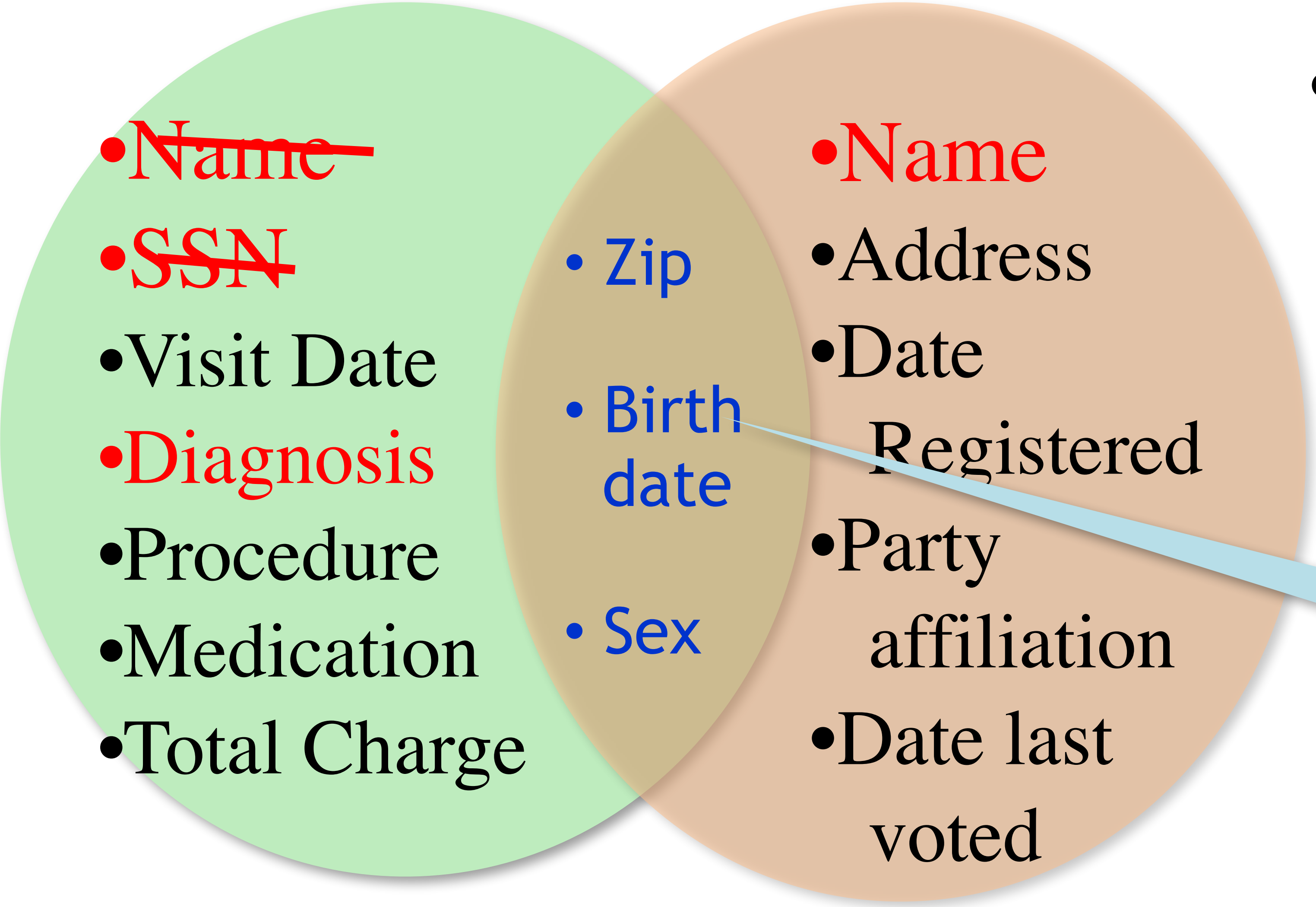
- Governor of MA **uniquely identified** using ZipCode, Birth Date, and Sex.

Name linked to Diagnosis

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

The Ugly side of Record Linkage

[Sweeney IJUFKS 2002]



- ~~Name~~
- ~~SSN~~
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

- Zip
- Birth date
- Sex

- Name
- Address
- Date Registered
- Party affiliation
- Date last voted

Medical Data **Voter List**

- (87 % of US population **uniquely identified** using ZipCode, Birth Date, and Sex.

Quasi Identifier

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University