Record Linkage

Everything Data
CompSci 290.01 Spring 2014

DUKE
COMPUTER SCIENCE

Announcements (Wed. Jan. 28)

- **Homework #3** will be posted by tomorrow morning
  - Due midnight Sunday

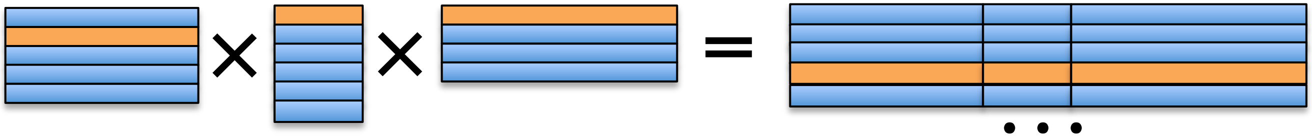
Recap: *Querying Relational Databases in SQL*

SELECT *columns or expressions*

5. Compute one output row for each "wide row" (or for each group of them if query has grouping/aggregation)

FROM *tables*

1. Generate all combinations of rows, one from each table; each combination forms a "wide row"

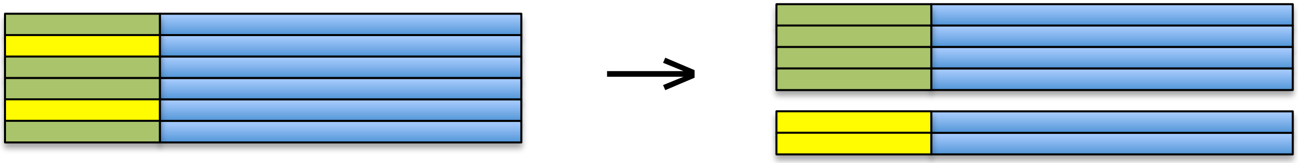


WHERE *conditions*

2. Filter—keep only "wide rows" satisfying *conditions*

GROUP BY *columns*

3. Group—"wide rows" with matching values for *columns* go into the same group

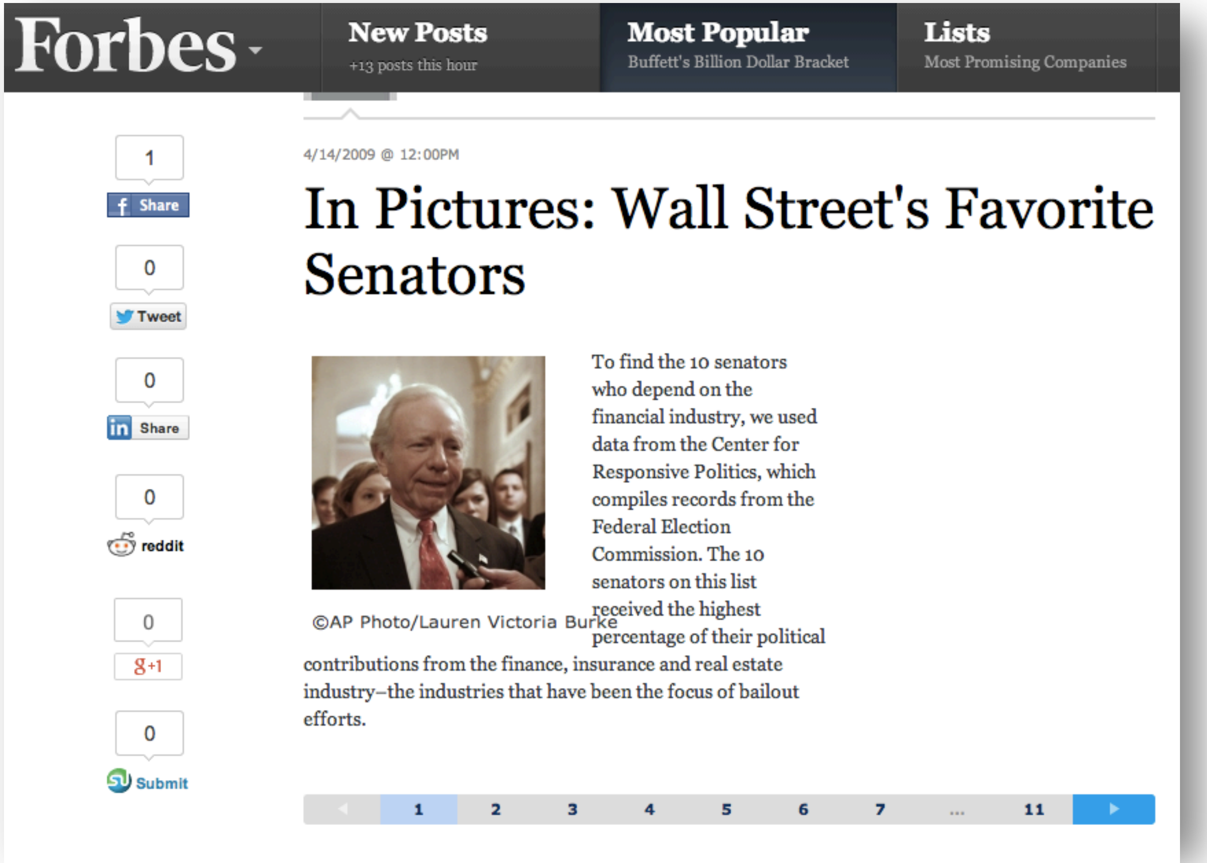


ORDER BY *output columns*;

4. Sort the output rows

Problem

- Forbes magazine article: “Wall Street’s favorite senators”



4/14/2009 @ 12:00PM

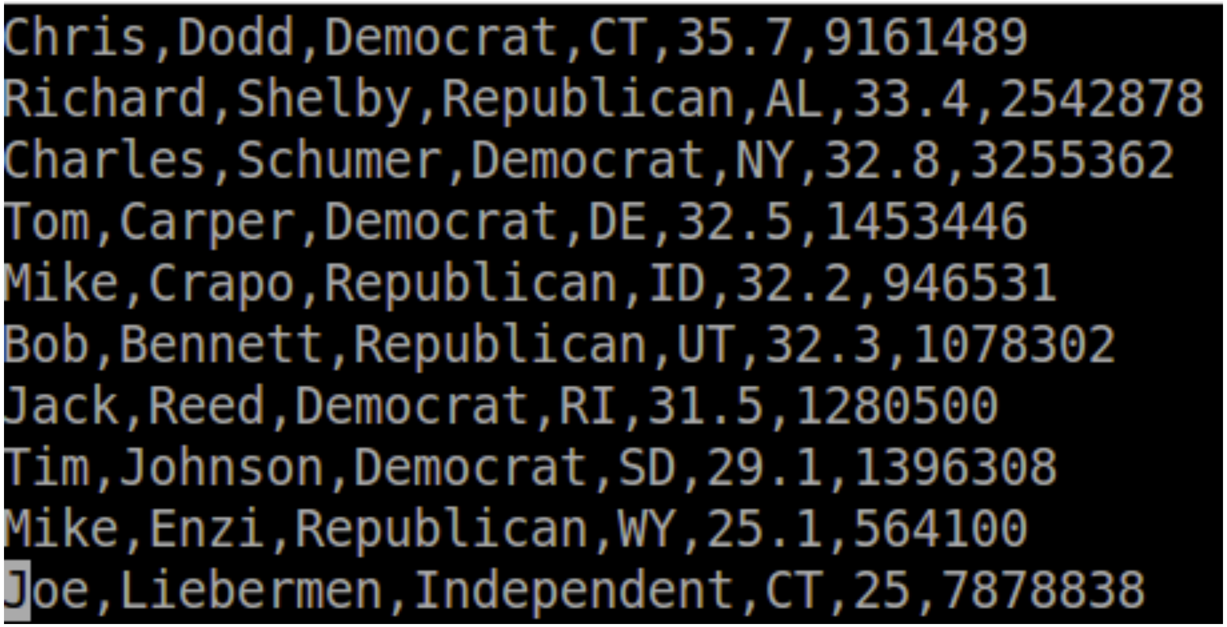
In Pictures: Wall Street's Favorite Senators

©AP Photo/Lauren Victoria Burke

To find the 10 senators who depend on the financial industry, we used data from the Center for Responsive Politics, which compiles records from the Federal Election Commission. The 10 senators on this list received the highest percentage of their political contributions from the finance, insurance and real estate industry–the industries that have been the focus of bailout efforts.

Problem

- Forbes magazine article: “Wall Street’s favorite senators”



```
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
```

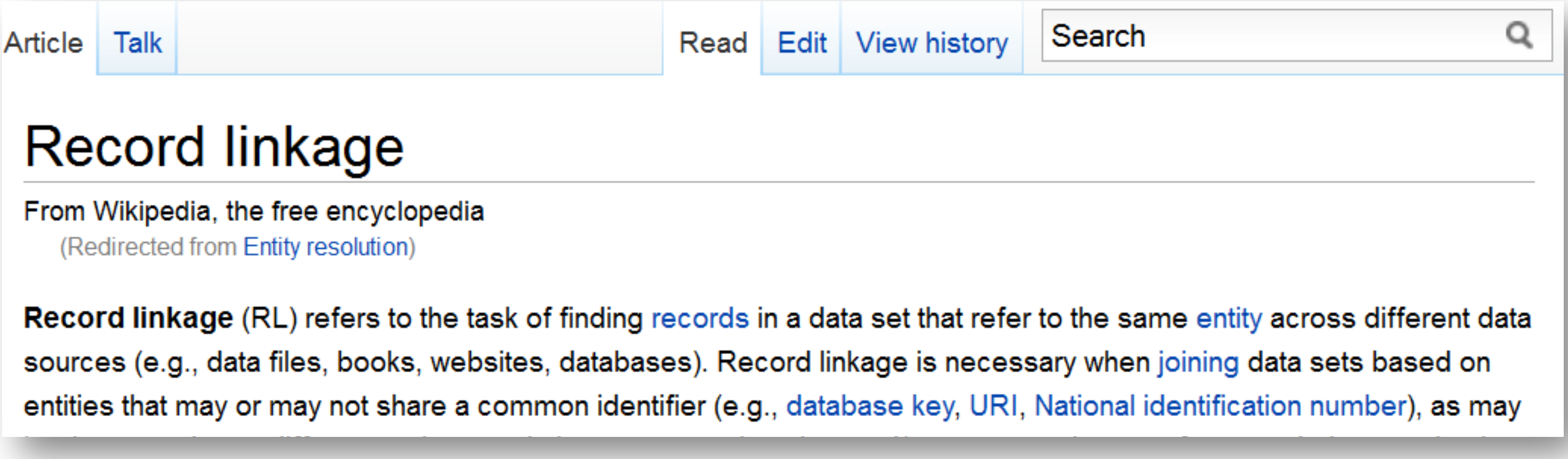
- What are their ages?

Solution

- Join with the `persons` table (from `govtrack`)
- But there is no key to join on …

Record Linkage

- Problem of finding duplicate entities across different sources (*or even within a single dataset*).



Record linkage

From Wikipedia, the free encyclopedia
(Redirected from Entity resolution)

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

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

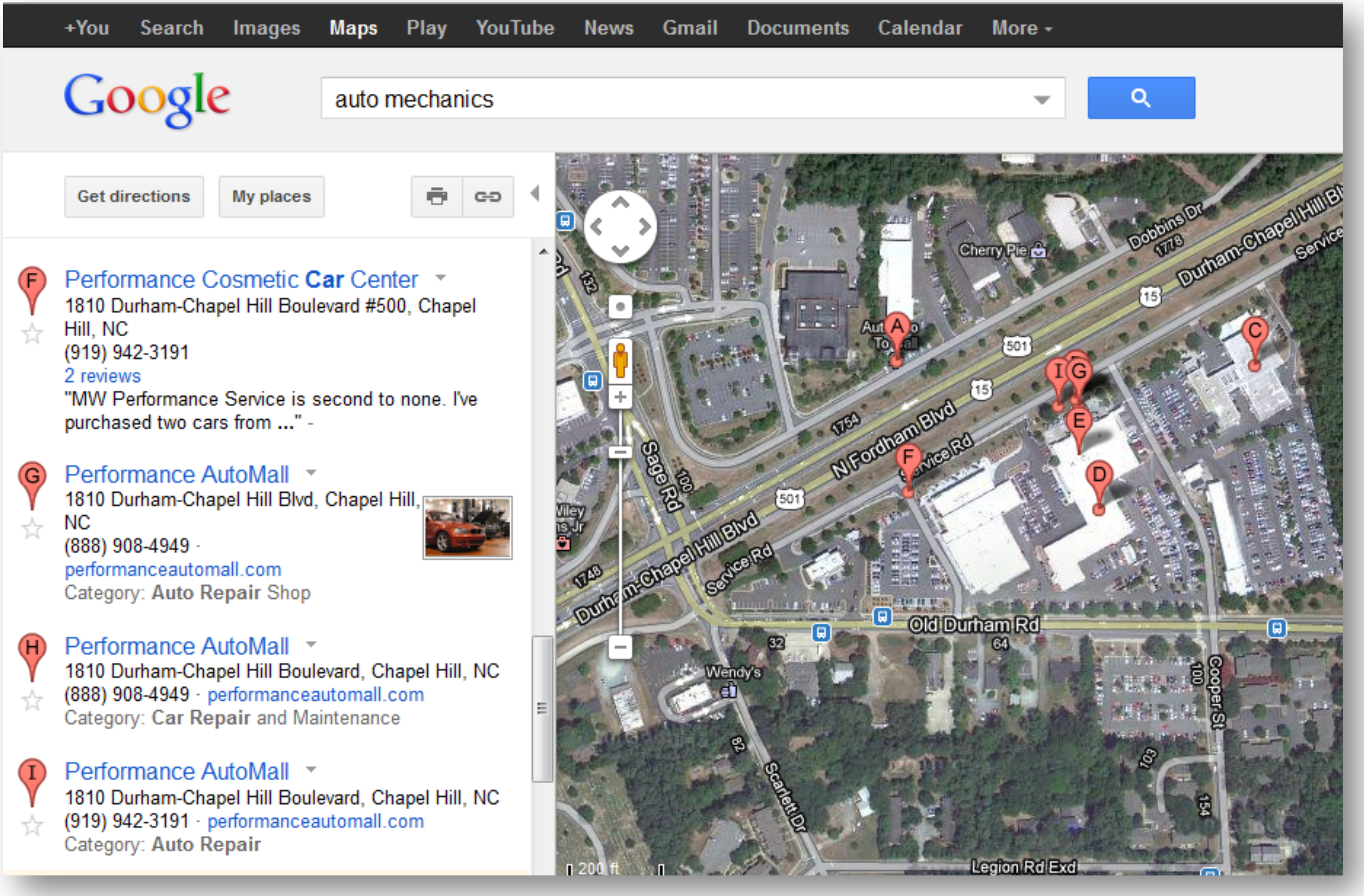
Object consolidation

Entity clustering

Household matching

Reference matching

Motivating Example 1: Web



F

Performance Cosmetic Car Center
1810 Durham-Chapel Hill Boulevard #500, Chapel Hill, NC
(919) 942-3191
2 reviews
"MW Performance Service is second to none. I've purchased two cars from ..." -

G

Performance AutoMall
1810 Durham-Chapel Hill Blvd, Chapel Hill, NC
(888) 908-4949 -
performanceautomall.com
Category: Auto Repair Shop

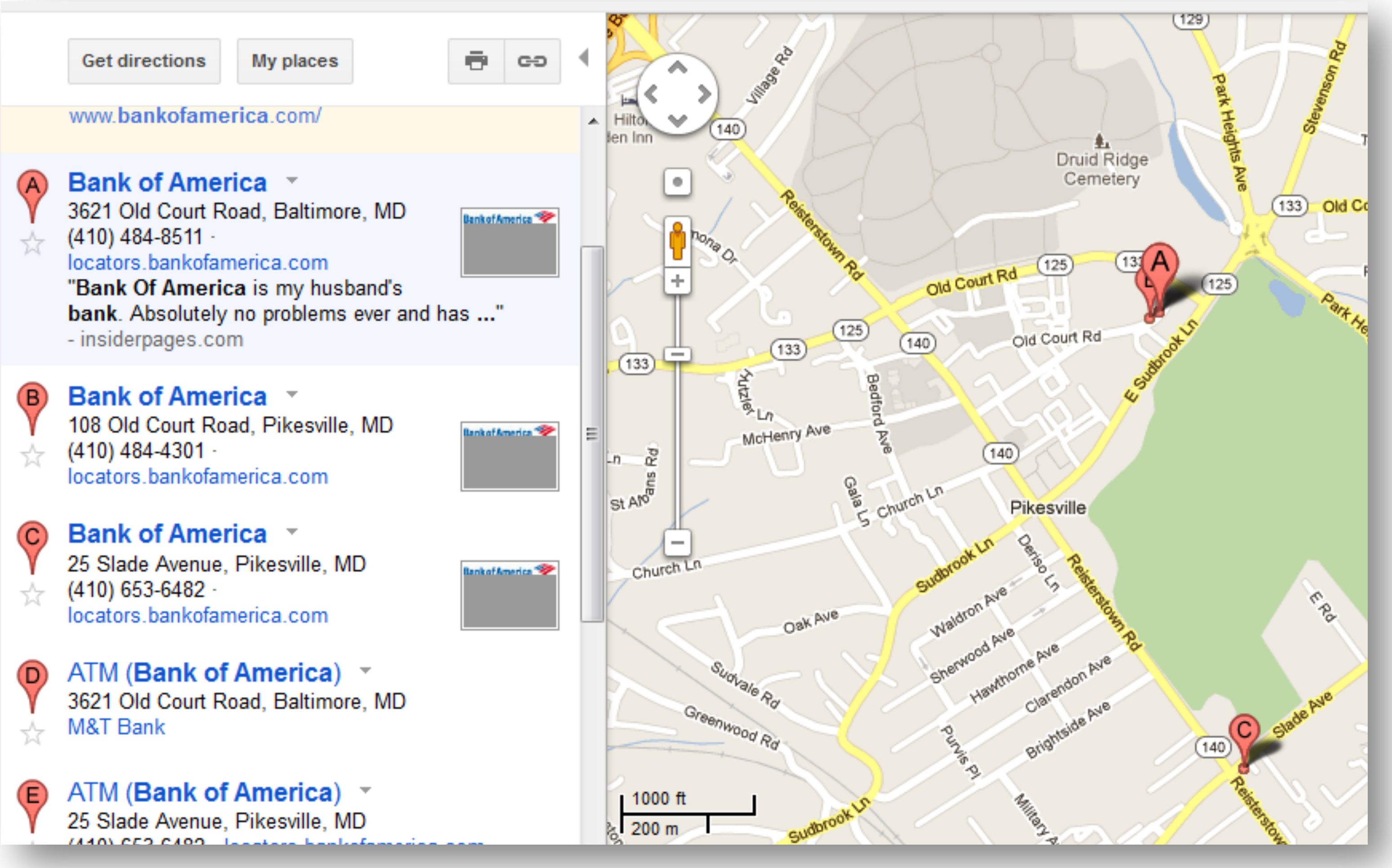
H

Performance AutoMall
1810 Durham-Chapel Hill Boulevard, Chapel Hill, NC
(888) 908-4949 - performanceautomall.com
Category: Car Repair and Maintenance

I

Performance AutoMall
1810 Durham-Chapel Hill Boulevard, Chapel Hill, NC
(919) 942-3191 - performanceautomall.com
Category: Auto Repair

Motivating Example 1: Web



www.bankofamerica.com/

A **Bank of America**
3621 Old Court Road, Baltimore, MD
(410) 484-8511 ·
locators.bankofamerica.com
"**Bank Of America** is my husband's **bank**. Absolutely no problems ever and has ..."
- insiderpages.com

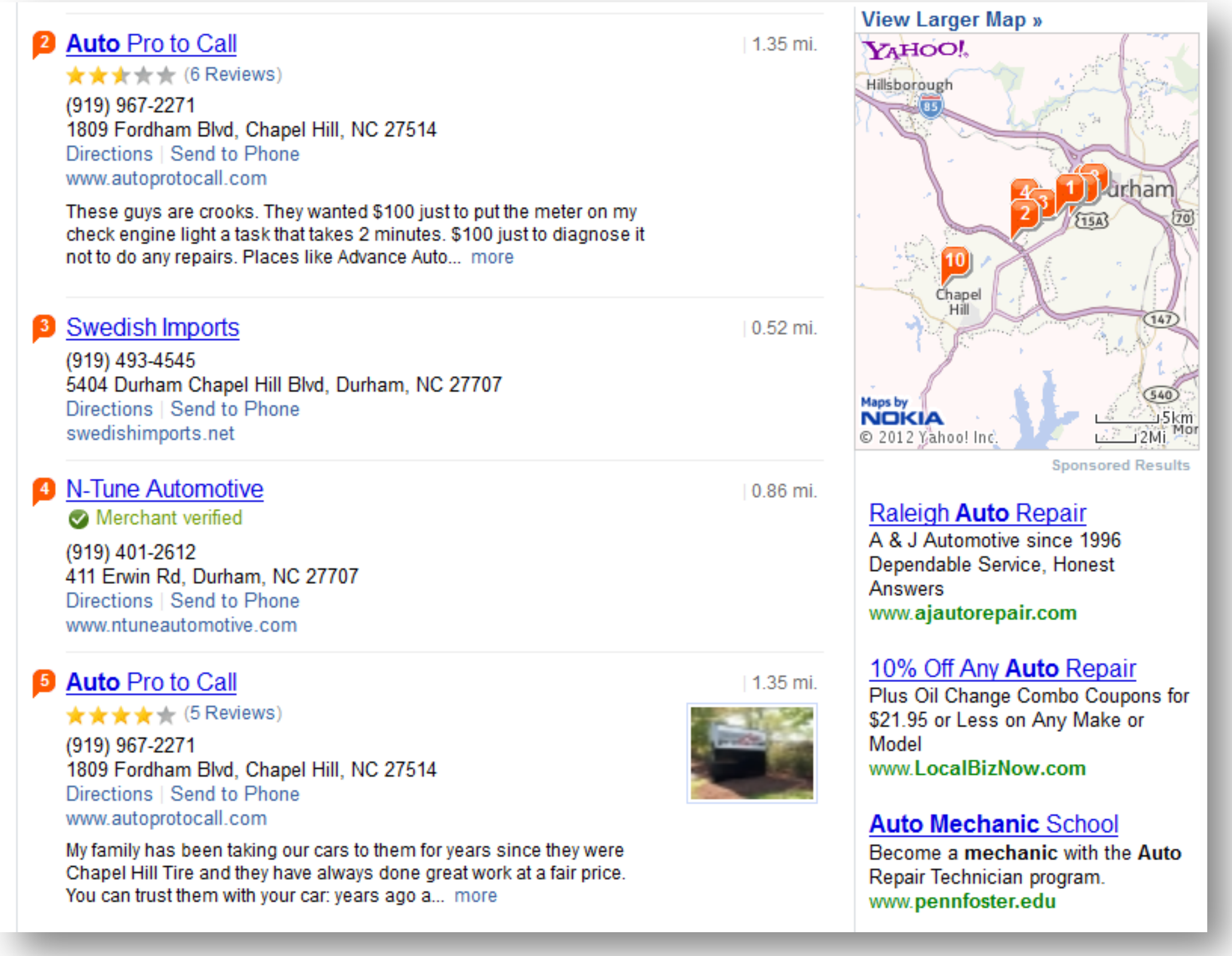
B **Bank of America**
108 Old Court Road, Pikesville, MD
(410) 484-4301 ·
locators.bankofamerica.com

C **Bank of America**
25 Slade Avenue, Pikesville, MD
(410) 653-6482 ·
locators.bankofamerica.com

D ATM (**Bank of America**)
3621 Old Court Road, Baltimore, MD
M&T Bank

E ATM (**Bank of America**)
25 Slade Avenue, Pikesville, MD
(410) 653-6482 · locators.bankofamerica.com

Motivating Example 1: Web



2 **Auto Pro to Call** | 1.35 mi.

★★½☆☆ (6 Reviews)

(919) 967-2271

1809 Fordham Blvd, Chapel Hill, NC 27514

Directions | Send to Phone

www.autoprotocall.com

These guys are crooks. They wanted $100 just to put the meter on my check engine light a task that takes 2 minutes. $100 just to diagnose it not to do any repairs. Places like Advance Auto... more

3 **Swedish Imports** | 0.52 mi.

(919) 493-4545

5404 Durham Chapel Hill Blvd, Durham, NC 27707

Directions | Send to Phone

swedishimports.net

4 **N-Tune Automotive** | 0.86 mi.

✓ Merchant verified

(919) 401-2612

411 Erwin Rd, Durham, NC 27707

Directions | Send to Phone

www.ntuneautomotive.com

5 **Auto Pro to Call** | 1.35 mi.

★★★★☆ (5 Reviews)

(919) 967-2271

1809 Fordham Blvd, Chapel Hill, NC 27514

Directions | Send to Phone

www.autoprotocall.com

My family has been taking our cars to them for years since they were Chapel Hill Tire and they have always done great work at a fair price. You can trust them with your car: years ago a... more

View Larger Map »

Sponsored Results

Raleigh Auto Repair

A & J Automotive since 1996 Dependable Service, Honest Answers

www.ajautorepair.com

10% Off Any Auto Repair

Plus Oil Change Combo Coupons for $21.95 or Less on Any Make or Model

www.LocalBizNow.com

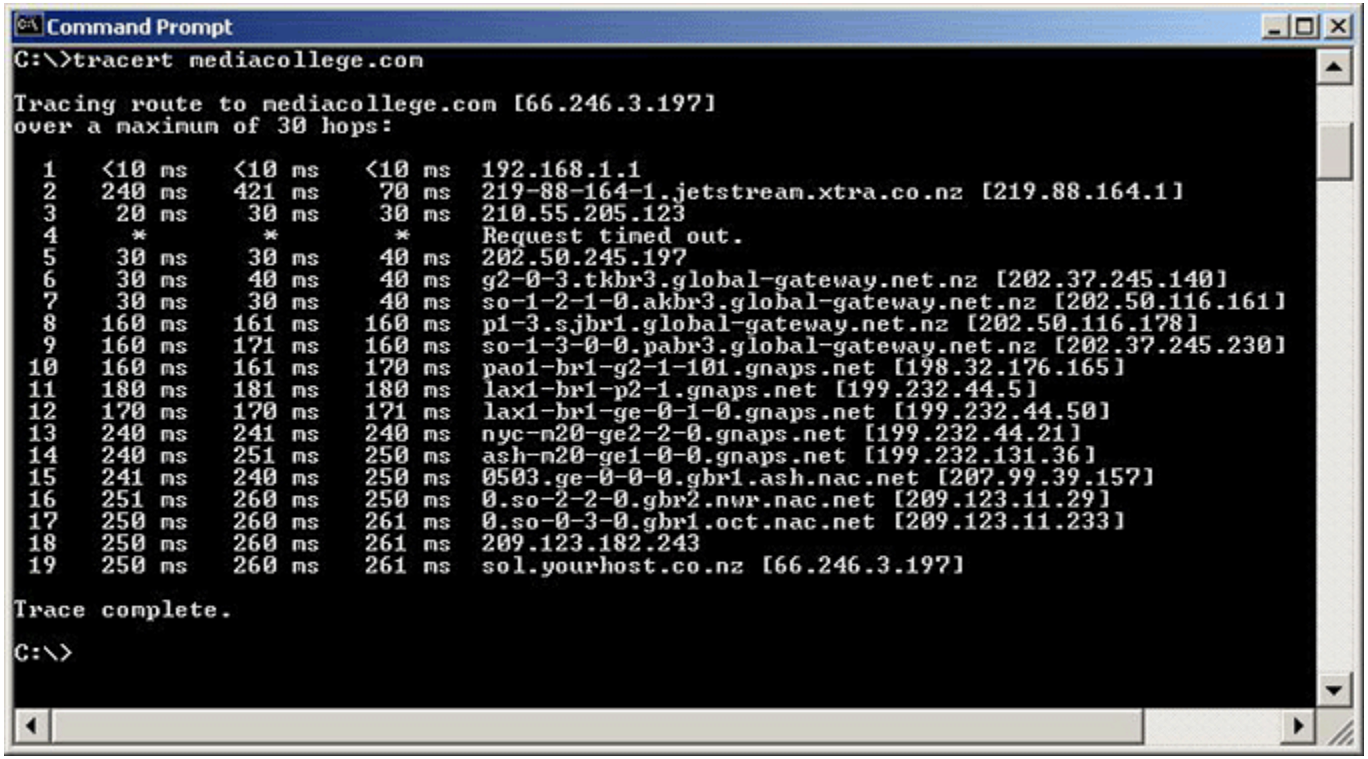
Auto Mechanic School

Become a **mechanic** with the **Auto** Repair Technician program.

www.pennfoster.edu

Motivating Example 2: Network Science

- Measuring the topology of the internet … using `traceroute`



```
C:\>tracert mediacollege.com

Tracing route to mediacollege.com [66.246.3.197]
over a maximum of 30 hops:

  1    <10 ms   <10 ms   <10 ms  192.168.1.1
  2    240 ms   421 ms    70 ms  219-88-164-1.jetstream.xtra.co.nz [219.88.164.1]
  3     20 ms    30 ms    30 ms  210.55.205.123
  4      *        *        *     Request timed out.
  5     30 ms    30 ms    40 ms  202.50.245.197
  6     30 ms    40 ms    40 ms  g2-0-3.tkbr3.global-gateway.net.nz [202.37.245.140]
  7     30 ms    30 ms    40 ms  so-1-2-1-0.akbr3.global-gateway.net.nz [202.50.116.161]
  8    160 ms   161 ms   160 ms  p1-3.sjbr1.global-gateway.net.nz [202.50.116.178]
  9    160 ms   171 ms   160 ms  so-1-3-0-0.pabr3.global-gateway.net.nz [202.37.245.230]
 10    160 ms   161 ms   170 ms  pao1-br1-g2-1-101.gnaps.net [198.32.176.165]
 11    180 ms   181 ms   180 ms  lax1-br1-p2-1.gnaps.net [199.232.44.5]
 12    170 ms   170 ms   171 ms  lax1-br1-ge-0-1-0.gnaps.net [199.232.44.50]
 13    240 ms   241 ms   240 ms  nyc-m20-ge2-2-0.gnaps.net [199.232.44.21]
 14    240 ms   251 ms   250 ms  ash-m20-ge1-0-0.gnaps.net [199.232.131.36]
 15    241 ms   240 ms   250 ms  0503.ge-0-0-0.gbr1.ash.nac.net [207.99.39.157]
 16    251 ms   260 ms   250 ms  0.so-2-2-0.gbr2.nwr.nac.net [209.123.11.29]
 17    250 ms   260 ms   261 ms  0.so-0-3-0.gbr1.oct.nac.net [209.123.11.233]
 18    250 ms   260 ms   261 ms  209.123.182.243
 19    250 ms   260 ms   261 ms  sol.yourhost.co.nz [66.246.3.197]

Trace complete.

C:\>
```

IP Aliasing Problem [Willinger et al. 2009]

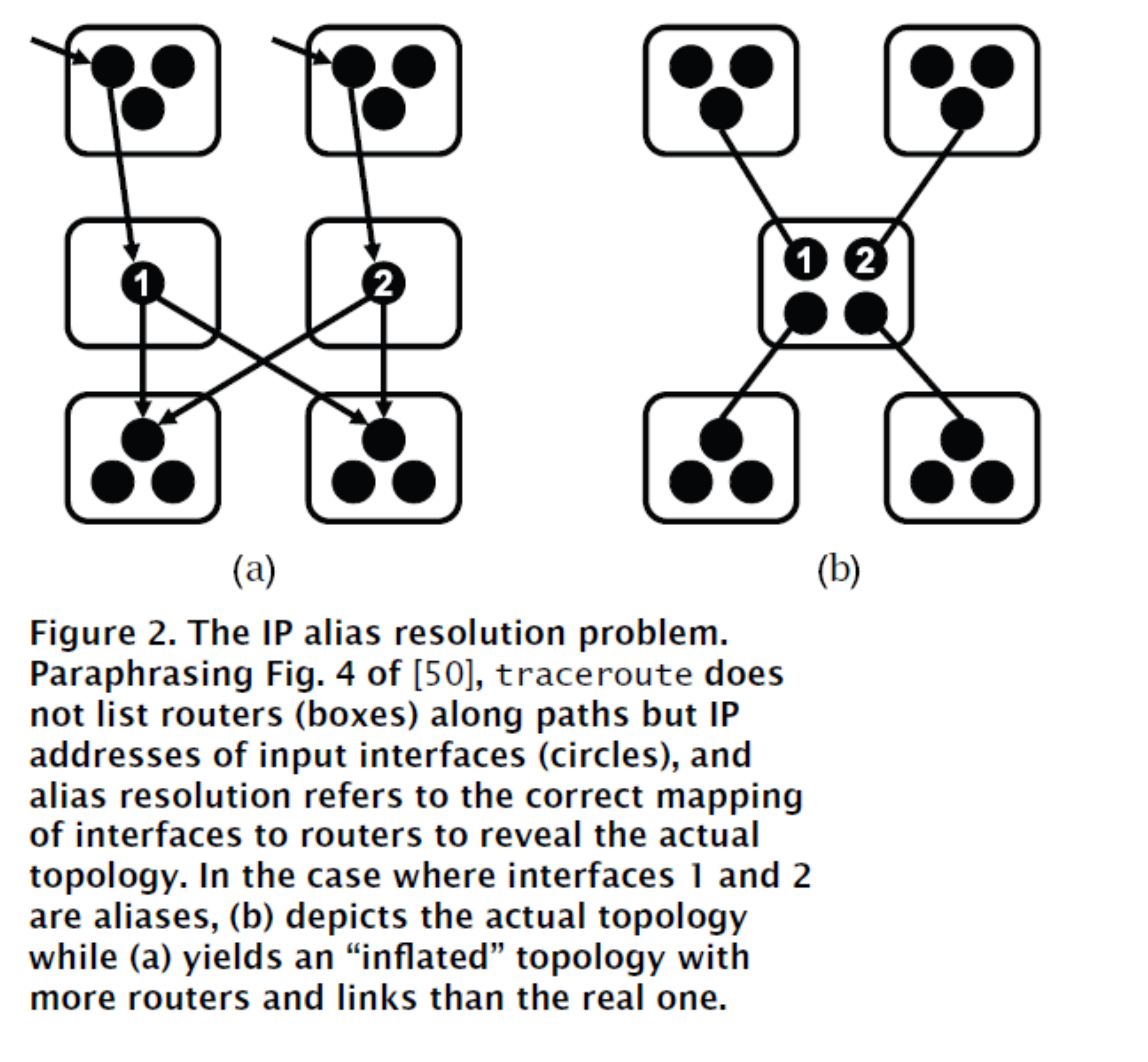


Figure 2. The IP alias resolution problem. Paraphrasing Fig. 4 of [50], `traceroute` does not list routers (boxes) along paths but IP addresses of input interfaces (circles), and alias resolution refers to the correct mapping of interfaces to routers to reveal the actual topology. In the case where interfaces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

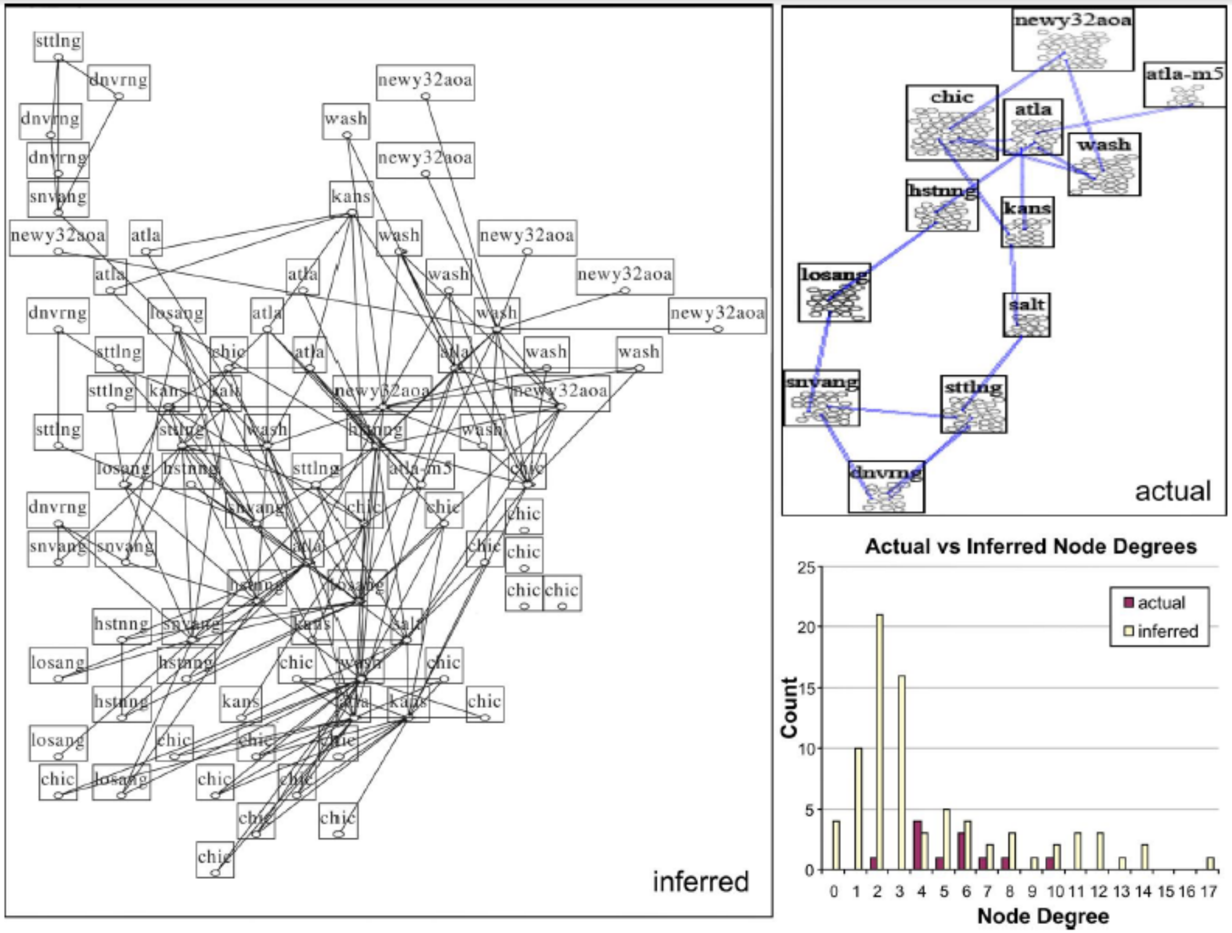


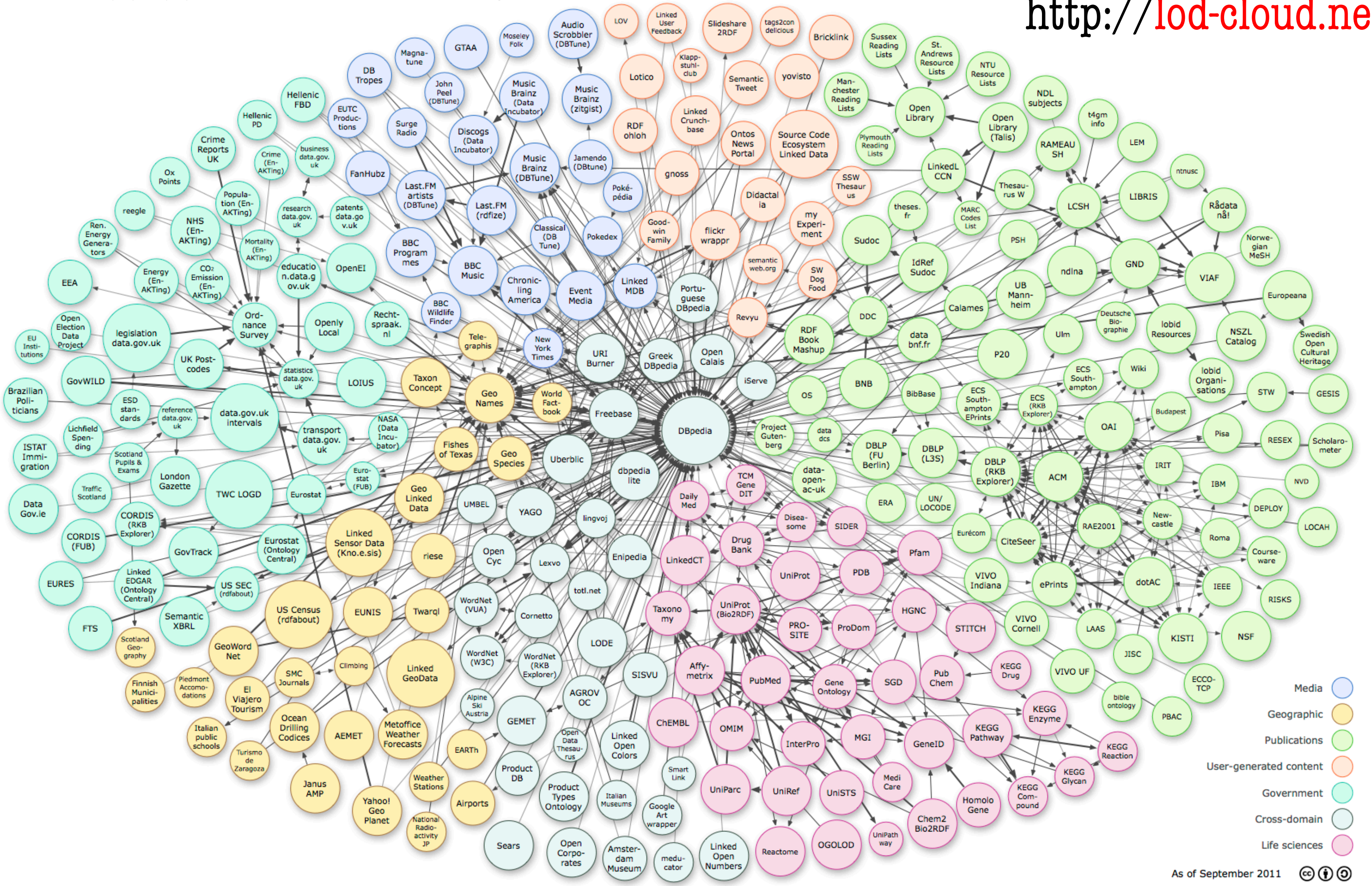
Figure 3. The IP alias resolution problem in practice. This is re-produced from [48] and shows a comparison between the Abilene/Internet2 topology inferred by Rocketfuel (left) and the actual topology (top right). Rectangles represent routers with interior ovals denoting interfaces. The histograms of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

And many many more examples

- *Linking Census Records*
- *Public Health*
- *Medical records*
- *Web search – query disambiguation*
- *Comparison shopping*
- *Maintaining customer databases*
- *Law enforcement and Counter-terrorism*
- *Scientific data*
- *Genealogical data*
- *Bibliographic data*

Opportunity

http://lod-cloud.net/



Back to our example

- Join with the persons table (from govtrack)

- But there is no key to join on …

- What about (firstname, lastname)?

Attempt 1:

```
SELECT w.*, date_part('year', current_date) -
date_part('year', p.birthday) AS age
FROM wallst w, persons p
WHERE w.first_name = p.first_name
        and w.last_name = p.last_name;
```

Problems

- Join condition is too specific
  - Nicknames used instead of real first names

Attempt 2:

- Join on Last name + Age < 100 (senator must be alive)

```
SELECT w.*, date_part('year', current_date) -
date_part('year', p.birthday) AS age
FROM wallst w, persons p
WHERE w.lastname = p.last_name and
date_part('year', current_date) - date_part('year',
p.birthday) < 100;
```

Problem:

- Join condition is too inclusive
  - Many individuals share the same last name.

Surname	Approx #	Rank
Smith	2.4 M	1
Johnson	1.8 M	2
Williams	1.5 M	3
Brown	1.4 M	4
Jones	1.4 M	5

http://www.census.gov/genealogy/www/data/2000surnames/

“Where is Joe Liebermen ?”

- Spelling mistake
  - – Liebermen vs Lieberman

- Need an approximate matching condition!

Levenshtein (or edit) distance

- The minimum number of character **edit** operations needed to turn one string into the other.

LIEBERMAN
LIEBERMEN

  - Substitute A to E. Edit distance = 1

Levenshtein (or edit) distance

- Distance between two string s and t is the shortest sequence of **edit commands** that transform s to t.

Costs can be different

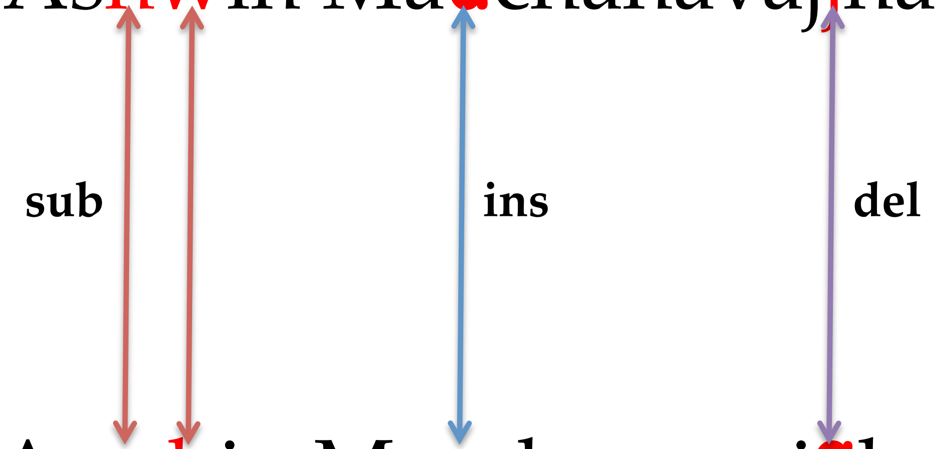
- Commands:
  - Copy character from s to t (cost = 0)
  - Delete a character from s (cost = 1)
  - Insert a character into t (cost = 1)
  - Substitute one character for another (cost = 1)

Levenshtein (or edit) distance

Ashwin Machanavajjhala
Aswhin Maachanavajhala

Levenshtein (or edit) distance

String s: AshwinMaGchanavajjhala



sub

ins

del

String t: Aswhin MaachanavajGhala

Total cost: 4

Computing the edit distance

		A	S	W	A	N
	0	1				
A	1					
S						
W						
H						
I						
N						

Cost of changing
"G" → "A"

Cost of changing
"ASWH" → "AS"

Computing the edit distance

		A	S	W	A	N
	0	1	2			
A	1	0	1			
S	2	1	0			
W	3	2				
H						
I						
N						

Cost of changing “ASW” → “AS”:

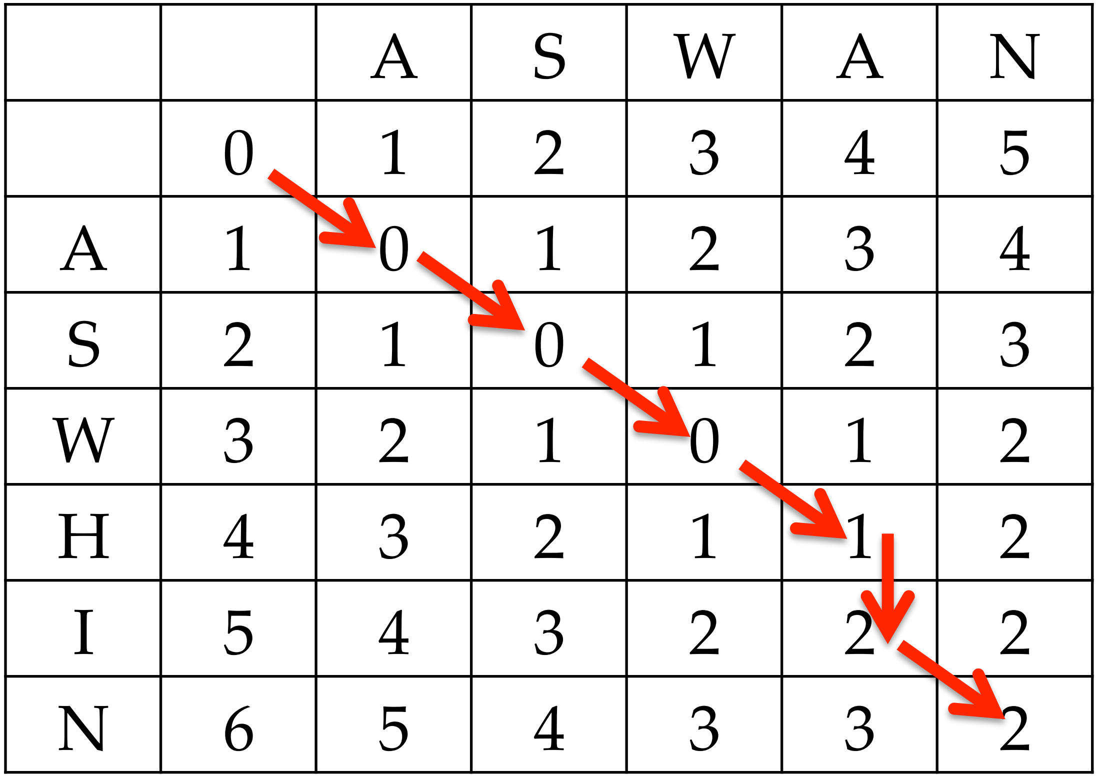
Minimum of:

- Cost of “AS” → “AS” + 1 (delete W)
- Cost of “ASW” → “A” + 1 (insert S)
- Cost of “AS” → “A” + 1 (substitute W with S)

Computing the edit distance

		A	S	W	A	N
	0	1	2	3	4	5
A	1	0	1	2	3	4
S	2	1	0	1	2	3
W	3	2	1	0	1	2
H	4	3	2	1	1	2
I	5	4	3	2	2	2
N	6	5	4	3	3	?

Computing the edit distance



		A	S	W	A	N
	0	1	2	3	4	5
A	1	0	1	2	3	4
S	2	1	0	1	2	3
W	3	2	1	0	1	2
H	4	3	2	1	1	2
I	5	4	3	2	2	2
N	6	5	4	3	3	2

Remember the minimum in each step and retrace your path.

Edit Distance Variants

- Needleman-Munch
  - Different costs for each operation
- Affine Gap distance
  - John Reed vs John Francis “Jack” Reed
  - Consecutive inserts cost less than the first insert.

Back to our example … Attempt 3

```
SELECT w.firstname, w.lastname, w.state, w.party,
p.first_name, p.last_name, date_part('year',
current_date) - date_part('year', p.birthday) AS age
FROM   wallst w, persons p
WHERE levenshtein(w.lastname, p.last_name) <= 1
and date_part('year', current_date) -
date_part('year', p.birthday) < 100;
```

Jaccard Distance

- Useful similarity function for sets
  - (*and for… long strings*).
- Let A and B be two sets
  - Words in two documents
  - Friends lists of two individuals

$$\boxed{\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}}$$

Jaccard similarity for names

- Use character trigrams

LIEBERMAN = {GGL, GLI, LIE, IEB, EBE,
BER, ERM, RMA,MAN, ANG, NGG}

LIEBERMEN = {GGL, GLI, LIE, IEB, EBE,
BER, ERM, RMA,MEN, ENG, NGG}

Jaccard(s,t) = 9/13 = 0.69

Attempt 4:

```
SELECT w.firstname, w.lastname, w.state, w.party,
p.first_name, p.last_name, date_part('year',
current_date) - date_part('year', p.birthday) AS age
FROM  wallst w, persons p
WHERE similarity(w.lastname, p.last_name) >= 0.5
and date_part('year', current_date) -
date_part('year', p.birthday) < 100;
```

Translation / Substitution Tables

- Strings that are usually used interchangeably
  - New York vs Big Apple
  - Thomas vs Tom
  - Robert vs Bob

Attempt 5

```
select w.firstname, w.lastname, w.state,
p.first_name, p.last_name, date_part('year',
current_date) - date_part('year', p.birthday) AS age
from wallst w, persons p
where levenshtein(w.lastname, p.last_name) <= 1
and date_part('year', current_date) -
date_part('year', p.birthday) < 100
and (w.firstname = p.first_name or w.firstname IN
(select n.nickname from nicknames n where
n.firstname = p.first_name));
```

Almost there …

- Tim matches both Timothy and Tim
  - Can fix it by matching on STATE
  - *Homework exercise* ☺

Summary of Similarity Methods

Easiest and most efficient

- Equality on a boolean predicate
- Edit distance
  - Levenstein, 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

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

Evaluating Record Linkage

- Hard to get all the matches to be exactly correct in real world problems
  - As we saw in real examples

- Need to quantify how good the matching is.

Property Testing

- Consider a universe U of objects
  - Documents (in web search)
  - Pairs of records (in record linkage)

- Suppose you want to identify a subset M in U that satisfies a specific property
  - Relevance to a query (in web search)
  - Do the records match (in record linkage)

Property Testing

- Consider a universe U of objects
- Suppose you want to identify a subset M in U that satisfies a specific property

- Let A be an (imperfect) algorithm that guesses whether or not an element in U satisfies the property
  - Let M_A be the subset of objects that A identifies as satisfying the property.

Property Testing

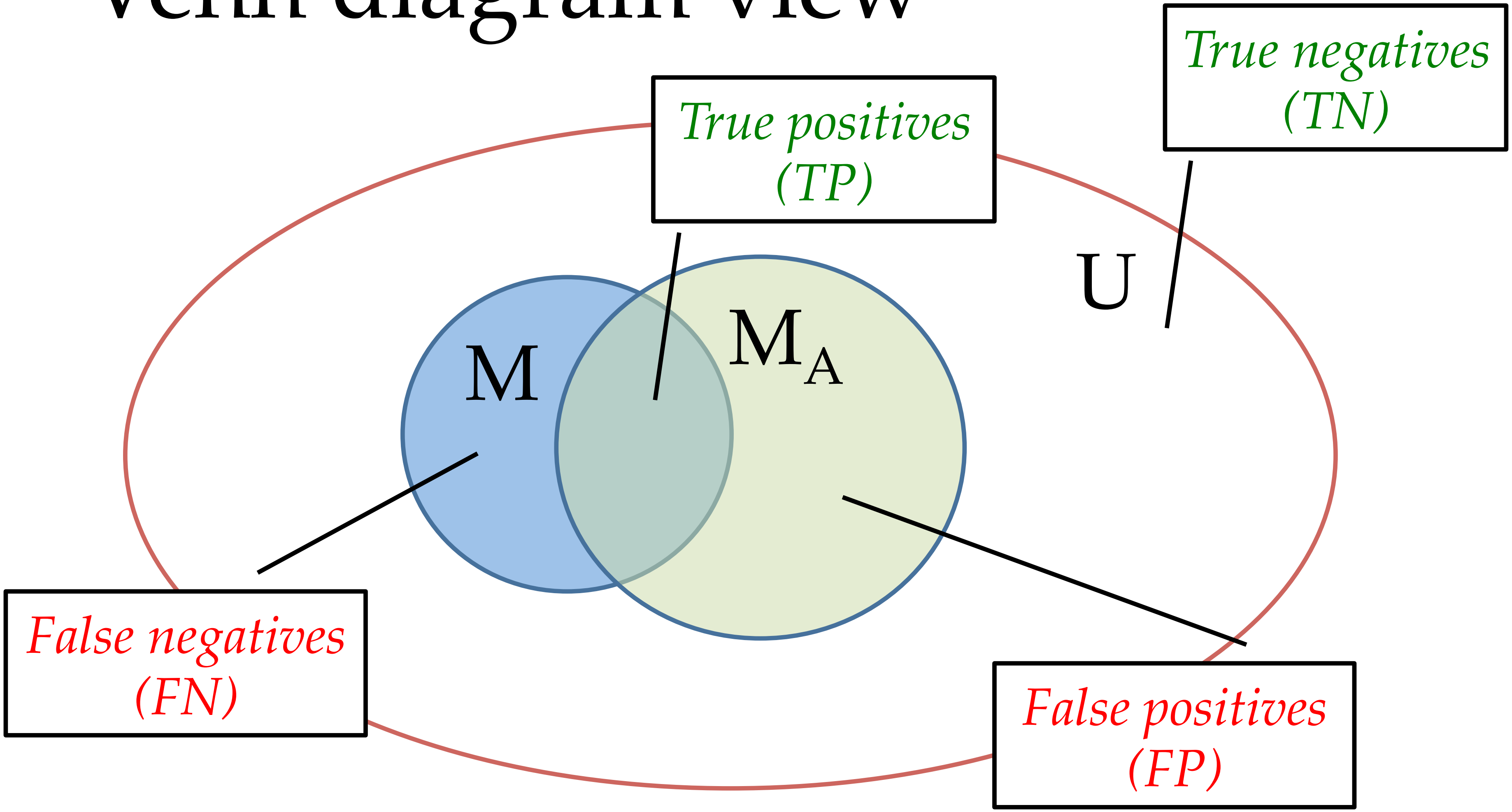
Crying Wolf!

Real World

Algorithm Guess

	Satisfies P	Doesn’t Satisfy P	
Satisfies P	*True positives (TP)*	*False positives (FP)*	M_A
Doesn’t satisfy P	*False negatives (FN)*	*True negatives (TN)*	$U – M_A$
	M	$U - M$	

Venn diagram view



True positives (TP)

True negatives (TN)

U

M

M_A

False negatives (FN)

False positives (FP)

Error: Precision / Recall

$$\text{Precision} = TP/(TP + FP)$$
$$= |M \cap M_A|/|M_A|$$

fraction of answers returned by A that are correct

$$\text{Recall} = TP/(TP + FN)$$
$$= |M \cap M_A|/|M|$$

fraction of correct answers that are returned by A

Error: F-measure

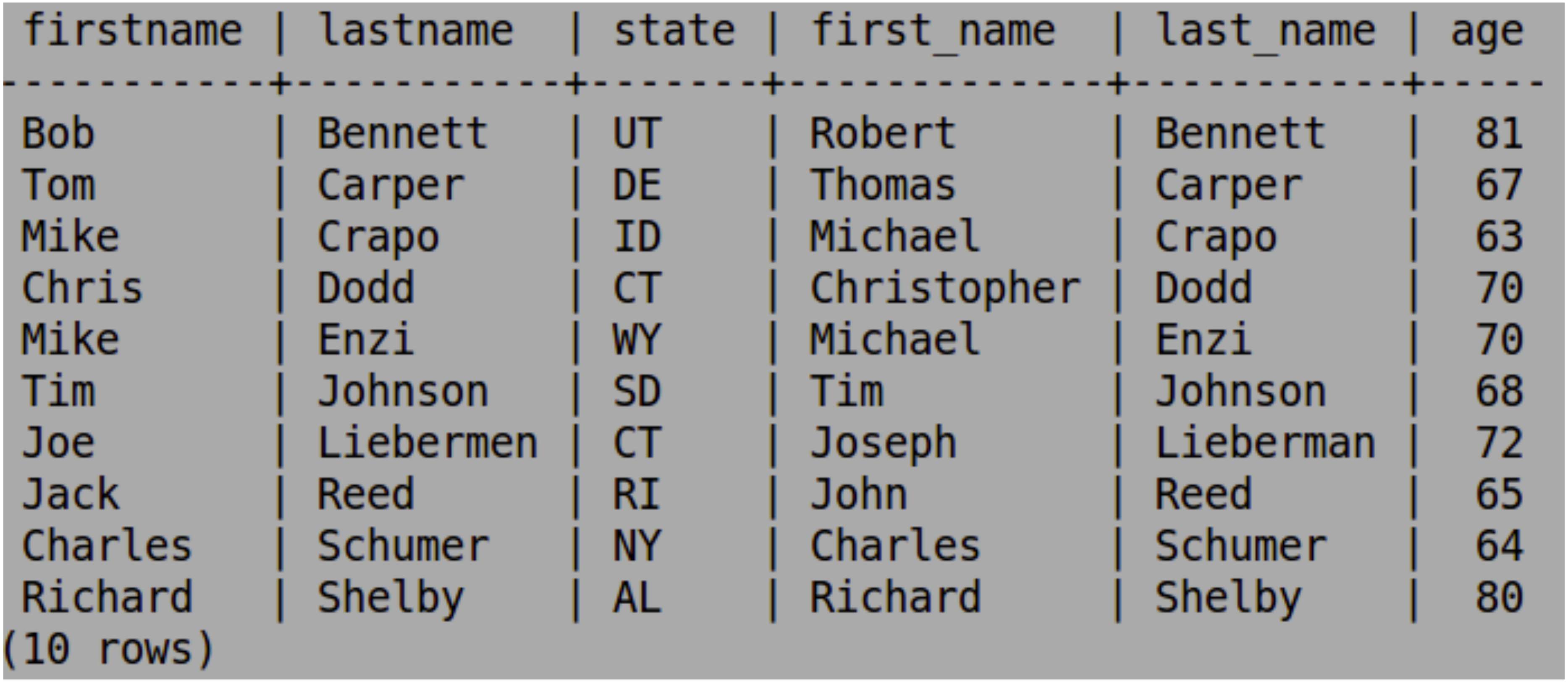
$$\text{Precision} = {|M \cap M_A|}/{|M_A|}$$

$$\text{Recall} = {|M \cap M_A|}/{|M|}$$

$$\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Example

- M:



```
 firstname | lastname  | state | first_name  | last_name | age
-----------+-----------+-------+-------------+-----------+-----
 Bob       | Bennett   | UT    | Robert      | Bennett   |  81
 Tom       | Carper    | DE    | Thomas      | Carper    |  67
 Mike      | Crapo     | ID    | Michael     | Crapo     |  63
 Chris     | Dodd      | CT    | Christopher | Dodd      |  70
 Mike      | Enzi      | WY    | Michael     | Enzi      |  70
 Tim       | Johnson   | SD    | Tim         | Johnson   |  68
 Joe       | Liebermen | CT    | Joseph      | Lieberman |  72
 Jack      | Reed      | RI    | John        | Reed      |  65
 Charles   | Schumer   | NY    | Charles     | Schumer   |  64
 Richard   | Shelby    | AL    | Richard     | Shelby    |  80
(10 rows)
```

Example:

Algorithm A:

```
select * from wallst w, persons p
where w.lastname = p.last_name and
date_part('year', current_date) - date_part('year',
p.birthday) < 100
and (w.firstname = p.first_name or w.firstname IN
(select n.nickname from nicknames n where
n.firstname = p.first_name));
```

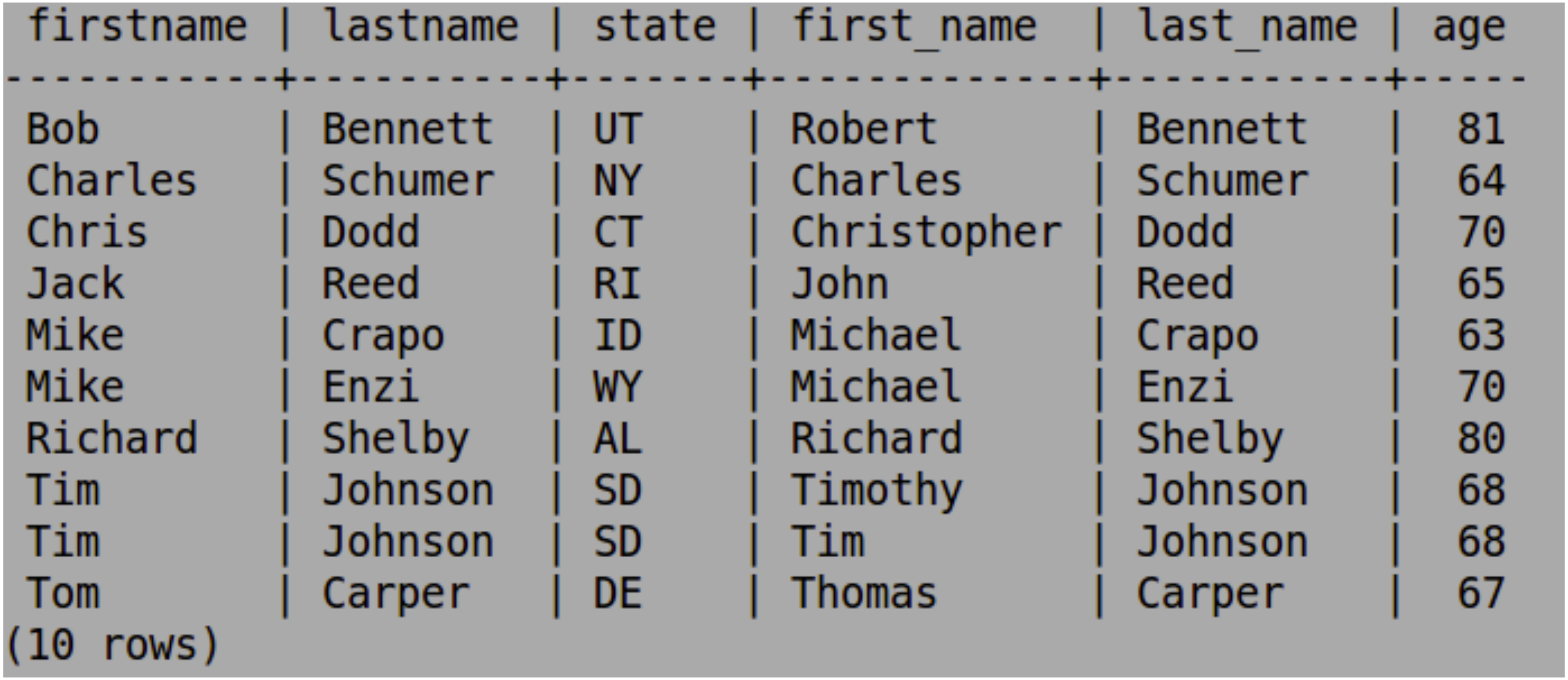
Exact match on last name

Age < 100

First name is same or a nickname

Example

- M_A:



```
 firstname | lastname | state | first_name  | last_name | age
-----------+----------+-------+-------------+-----------+-----
 Bob       | Bennett  | UT    | Robert      | Bennett   |  81
 Charles   | Schumer  | NY    | Charles     | Schumer   |  64
 Chris     | Dodd     | CT    | Christopher | Dodd      |  70
 Jack      | Reed     | RI    | John        | Reed      |  65
 Mike      | Crapo    | ID    | Michael     | Crapo     |  63
 Mike      | Enzi     | WY    | Michael     | Enzi      |  70
 Richard   | Shelby   | AL    | Richard     | Shelby    |  80
 Tim       | Johnson  | SD    | Timothy     | Johnson   |  68
 Tim       | Johnson  | SD    | Tim         | Johnson   |  68
 Tom       | Carper   | DE    | Thomas      | Carper    |  67
(10 rows)
```

Example

$$\text{Precision} = {}^{|M \cap M_A|}/_{|M_A|}$$

$$= {}^{9}/_{10} = 0.9$$

$$\text{Recall} = {}^{|M \cap M_A|}/_{|M|}$$

$$= {}^{9}/_{10} = 0.9$$

$$\text{F1 score} = 2\ \frac{0.9 \times 0.9}{0.9 + 0.9} = 0.9$$

Summary

- Many interesting data analyses require reasoning across different datasets

- May not have access to keys that uniquely identify individual rows in both datasets

Summary

- Use combinations of attributes that are approximate keys (or **quasi-identifiers**)

- Use similarity measures for fuzzy or approximate matching
  - **Levenshtein** or **Edit** distance
  - **Jaccard** Similarity

- Use translation tables

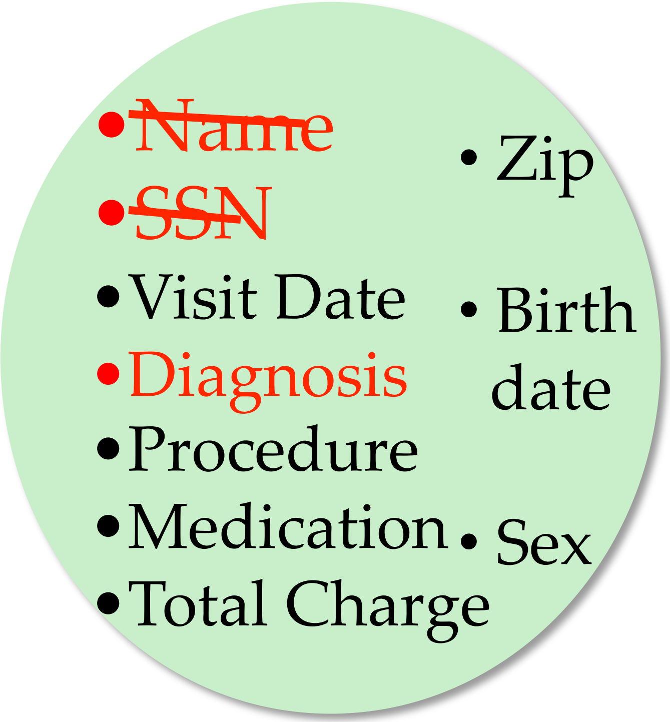
Summary

- Record Linkage is rarely perfect
  - Missing attributes
  - Messy data errors
  - …

- **Precision/Recall** is used to measure the quality of linkage.

The Ugly side of Record Linkage

[Sweeney IJUFKS 2002]



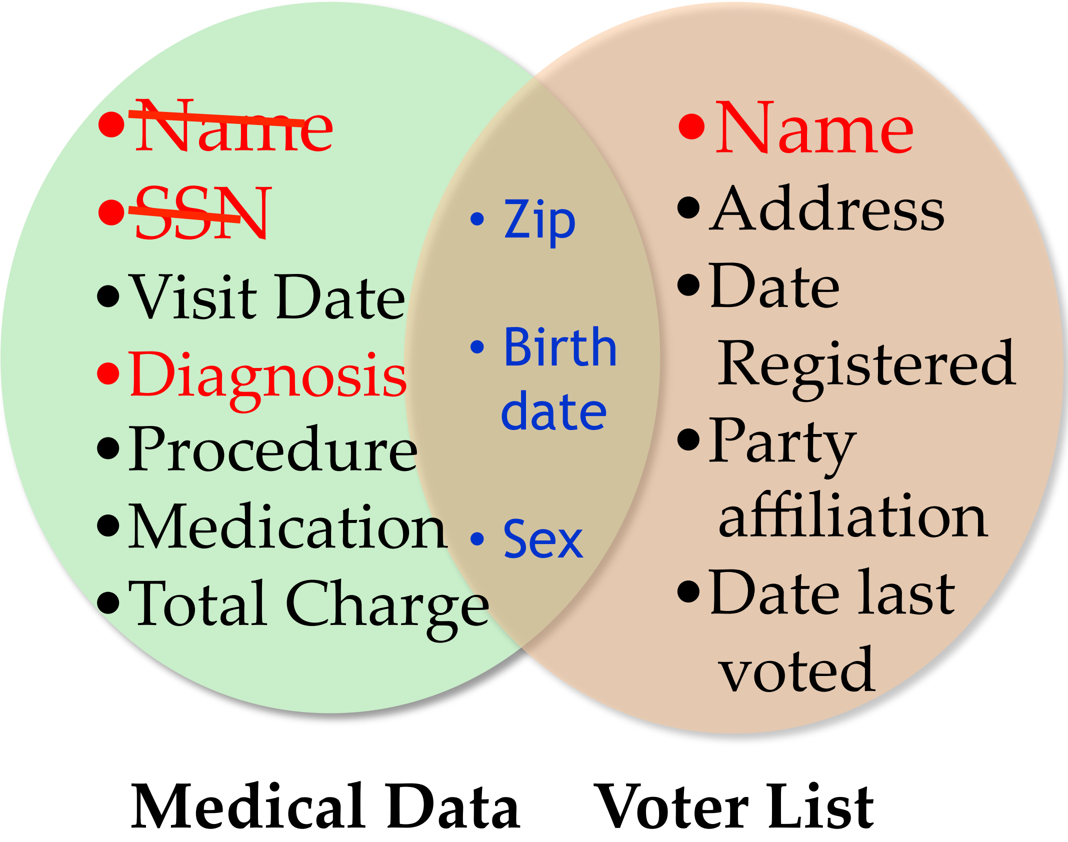
- ~~Name~~
- ~~SSN~~
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

- Zip
- Birth date
- Sex

Medical Data

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

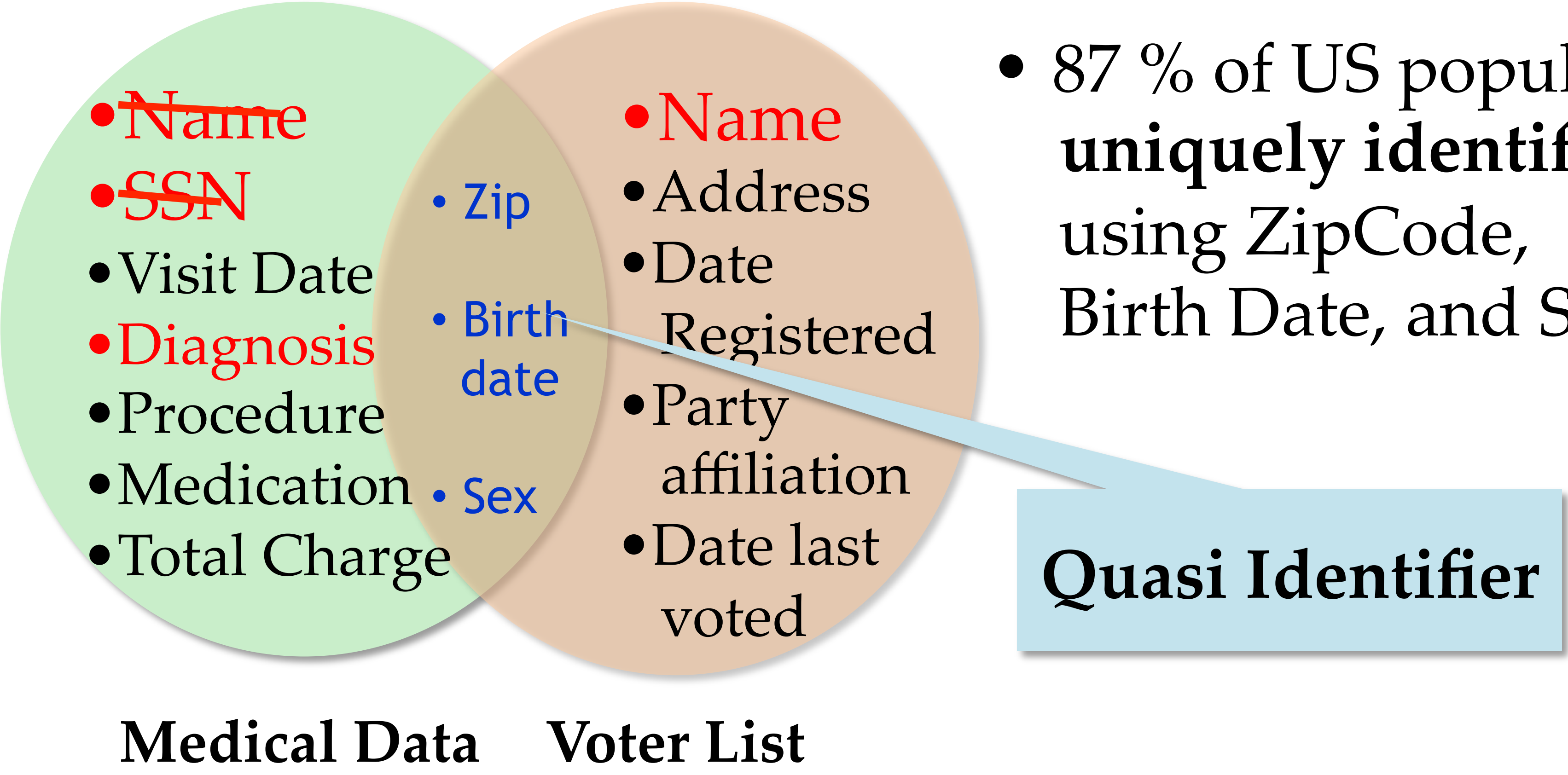
Medical Data **Voter List**

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

Name linked to Diagnosis

The Ugly side of Record Linkage

[Sweeney IJUFKS 2002]



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