Profiles Research Networking Software

Pubmed Disambiguation Service

http://profiles.catalyst.harvard.edu

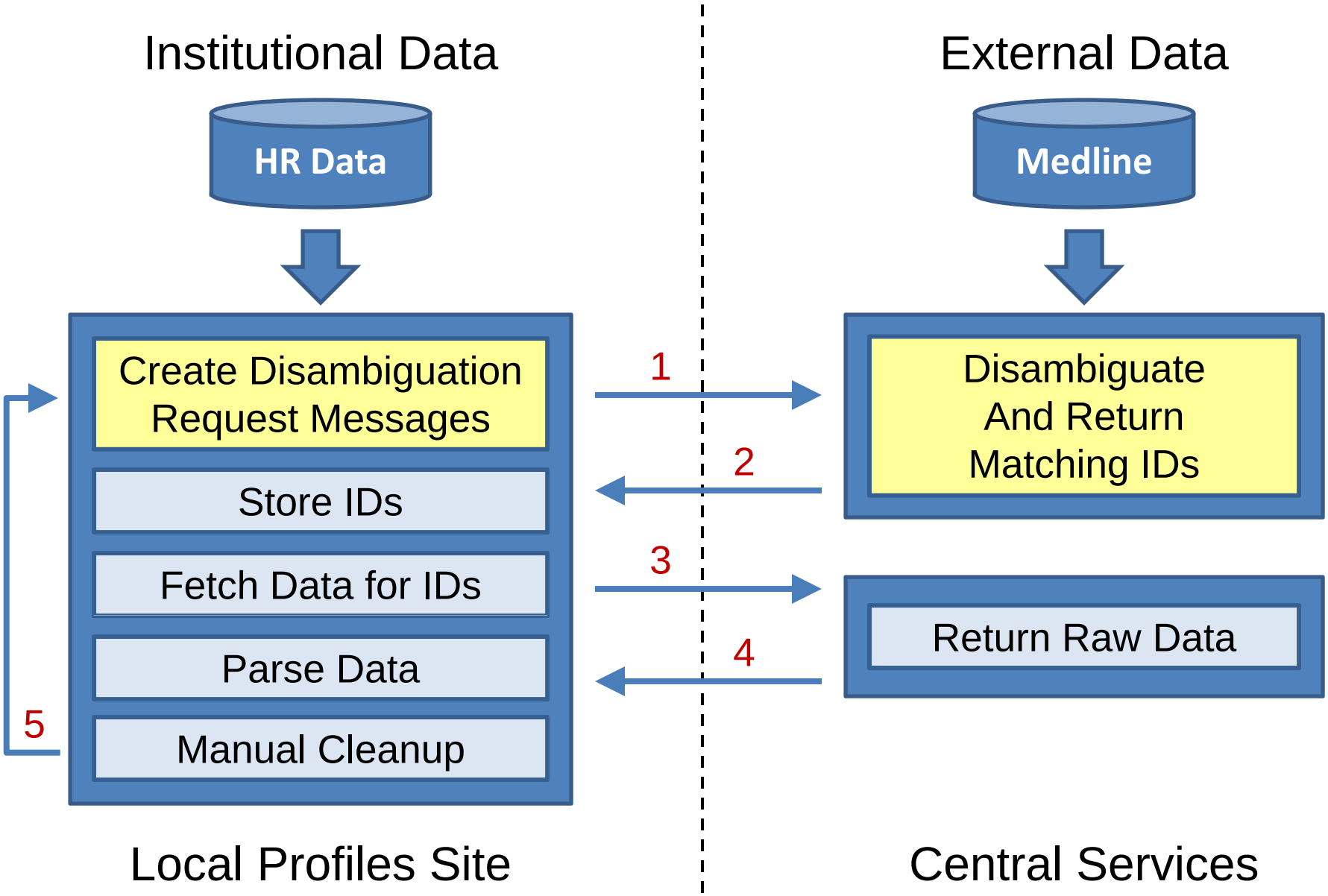
Griffin Weber, MD, PhD
Harvard Medical School
Beth Israel Deaconess Medical Center
January 20, 2012

Profiles Pubmed Disambiguation

http://profiles.catalyst.harvard.edu/software

- Free public web service
- Data updated weekly from Medline
- Does not require Profiles RNS
- Provide service with:
  - Name, affiliation, email, known publications
- Service returns:
  - Pubmed IDs or Pubmed Data
- Algorithms inspired by work done by:
  - Vetle I Torvik and Neil R Smalheiser

Disambiguation Architecture



Disambiguation Algorithm

- Identify “seed” publications
  - Known PMIDs
  - Email matches
  - Author name matches above probability threshold
  - Remove “exclude” PMIDs
- Identify “possible” publications
  - Start with author name matches below probability threshold
  - Calculate probability possible publications were written by the same author as a seed publication
  - Select publications above threshold

Name Matches

- Require exact match of last name
- Require substring match of first name, middle name, and suffix name
- Example: “John D. Smith”
  - Matches:
    - J. D. Smith
    - John David Smith
    - J. Daniel Smith
    - J. Smith
  - Does not match:
    - Jonathan D. Smith
    - John D. Smtih
    - James Smith
    - John D. Smithers

Name Match Probabilities

- Author names matching the person “Griffin M Weber”
  - G Weber (1344 Publications)
  - G M Weber (18 Publications)
  - Griffin Weber (11 Publications)
  - Griffin M Weber (4 Publications)

Name Match Probabilities

- Author names matching the author “G Weber”

G Weber (1344)
G A Weber (8)
G C Weber (3)
G D Weber (1)
G F Weber (46)
G H Weber (18)
G I Weber (2)
G J Weber (5)
G L Weber (8)
G M Weber (18)
G N Weber (4)
G P Weber (1)
G R Weber (5)
G S Weber (1)
G W Weber (22)
G Y Weber (2)
Gabriel Weber (8)
Gabriele Weber (5)
Genevieve L Weber (2)
Georg Weber (7)
Georg F Weber (24)
George Weber (11)
George H Weber (1)
George L Weber (1)
Georges Weber (1)
Geraint J Weber (1)
Gerald Weber (6)
Gerald I Weber (2)
Géraldine Weber (1)
Gerd Weber (2)
Gerhard Weber (20)
Gerhard J Weber (9)
Gerhard W Weber (19)
Gerhard Wilhelm Weber (1)
Geri Weber (1)
Germain Weber (7)
Gerrit Weber (5)
Gert Weber (15)
Gertrud Weber (1)
Giovanna Weber (28)
Gregor G Weber (3)
Gregory Weber (4)
Gregory A Weber (4)
Gregory F Weber (6)
Gregory J Weber (1)
Gregory L Weber (2)
Gregory M Weber (16)
Griffin Weber (11)
Griffin M Weber (4)
Guglielmo Weber (1)
Gunthard Weber (1)
Gunther Weber (2)
Gunther H Weber (8)
Günther Weber (30)
Guy Weber (2)
Gyorgy Weber (9)

- This list must represent many different people because, for example, “George H Weber” is not a name match to “Gerald I Weber”.
- Thus, the probability that an author “G Weber” matches the real person Griffin M Weber is low.

Name Match Probabilities

- Author names matching the author “G M Weber”

G Weber (1344)
G M Weber (18)
Gabriel Weber (8)
Gabriele Weber (5)
Georg Weber (7)
George Weber (11)
Georges Weber (1)
Gerald Weber (6)
Géraldine Weber (1)
Gerd Weber (2)
Gerhard Weber (20)
Geri Weber (1)
Germain Weber (7)
Gerrit Weber (5)
Gert Weber (15)
Gertrud Weber (1)
Giovanna Weber (28)
Gregory Weber (4)
Gregory M Weber (16)
Griffin Weber (11)
Griffin M Weber (4)
Gunther H Weber (8)

- This list must also represent different people because, for example, “Germain Weber” is not a name match to “Gunther H Weber”.
- However, this list is smaller, so it could be fewer people.
- Thus, the probability that an author “G M Weber” matches the real person Griffin M Weber is a little higher.

Name Match Probabilities

- Author names matching the author “Griffin Weber”

G Weber (1344)
G A Weber (8)
G C Weber (3)
G D Weber (1)
G F Weber (46)
G H Weber (18)
G I Weber (2)
G J Weber (5)
G L Weber (8)
G M Weber (18)
G N Weber (4)
G P Weber (1)
G R Weber (5)
G S Weber (1)
G W Weber (22)
G Y Weber (2)
Griffin Weber (11)
Griffin M Weber (4)

- This list must also represent different people because, for example, “G A Weber” is not a name match to “G S Weber”.
- However, this list is even smaller, and could be even fewer people.
- Thus, the probability that an author “Griffin Weber” matches the real person Griffin M Weber is even higher.

Name Match Probabilities

- Author names matching the author “Griffin M Weber”

  G Weber (1344) G M Weber (18) Griffin Weber (11) Griffin M Weber (4)

- In this list, every name matches every other name.
- There is no evidence that the author “Griffin M Weber” is more than one person.
- Thus, the probability that an author “Griffin M Weber” matches the real person Griffin M Weber is exactly 1.

Name Match Probabilities

- Probabilities author names match “Griffin M Weber”
  - Looking at all authors in Medline

Author Name	Publications	Probability	Threshold > 0.9
G Weber	1344	0.01643	Possible Match
G M Weber	18	0.07575	Possible Match
Griffin Weber	11	0.14865	Possible Match
Griffin M Weber	4	1.00000	Seed Publication

  - Looking only at authors on “Harvard” publications

Author Name	Publications	Probability	Threshold > 0.9
G Weber	1	0.14394	Possible Match
G M Weber	1	0.90909	Seed Publication
Griffin Weber	7	0.18182	Possible Match
Griffin M Weber	3	1.00000	Seed Publication

Matching Possible Pubs to Seeds

- Determine probability that each possible publication was written by the same author as each seed publication.
- If the probability > threshold for any pair, then return that possible publication
- Use the following publication attributes:
  - Journal name, title words, MeSH keywords, coauthors, affiliation words
- What is the probability that two publications were written by the same author, if (for example) they are in the same journal?

Bayes' Theorem

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|\sim A)*P(\sim A)}$$

* ⇨ P(A|B) = P(Same Author | Same Journal)

P(A) = P(Same Author) = P(Name Match) ⇦ ✓

P(B|A) = P(Same Journal | Same Author) ⇦ ?

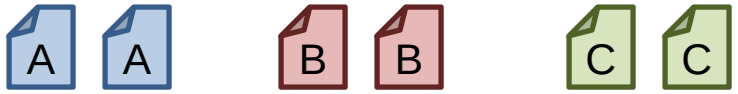
P(B|~A) = P(Same Journal | Different Author) ⇦ ?

Bayes' Theorem

- Create two randomly generated sets of 100,000 pairs of articles:

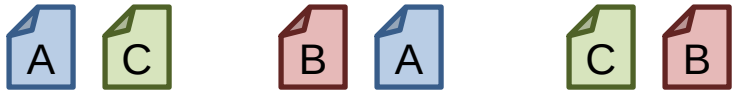
Set #1 – Same Author

Articles known to be written by the same person (based on same email in both affiliations)



Set #2 – Different Authors

Articles known to be written by different people (based on having no author name matches)



- Suppose…

	Same Author	**Different Authors**
Same Journal	10,000	10
Different Journal	90,000	99,990

- Then…

$$P(B|A) = P(\text{Same Journal} \mid \text{Same Author}) = 10{,}000/(10{,}000 + 90{,}000) = 0.1000$$

$$P(B|\sim A) = P(\text{Same Journal} \mid \text{Different Author}) = 10/(10 + 99{,}990) = 0.0001$$

Bayes' Theorem

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|\sim A)*P(\sim A)}$$

$$P(A|B) = P(\text{Same Author} \mid \text{Same Journal})$$

$$P(A) = P(\text{Same Author}) = P(\text{Name Match})$$

$$P(B|A) = P(\text{Same Journal} \mid \text{Same Author}) = 0.1000$$

$$P(B|\sim A) = P(\text{Same Journal} \mid \text{Different Author}) = 0.0001$$

Author Name	Publications	P(Same Author)	P(Same Author | Same Journal)
G Weber	1344	0.01643	0.94352
G M Weber	18	0.07575	0.98795
Griffin Weber	11	0.14865	0.99431

Bayes' Theorem

- For each possible publication / seed publication pair, calculate the following probabilities:

  P(Same Author | Same Journal)

  P(Same Author | Same Title Word)

  P(Same Author | Same Affiliation Word)

  P(Same Author | Same MeSH Keyword)

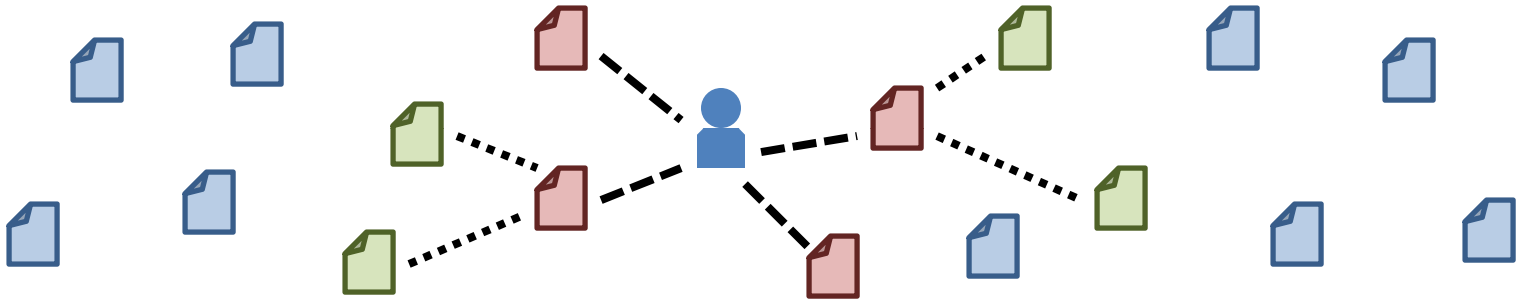
  P(Same Author | Same Co-Authors)

  etc.

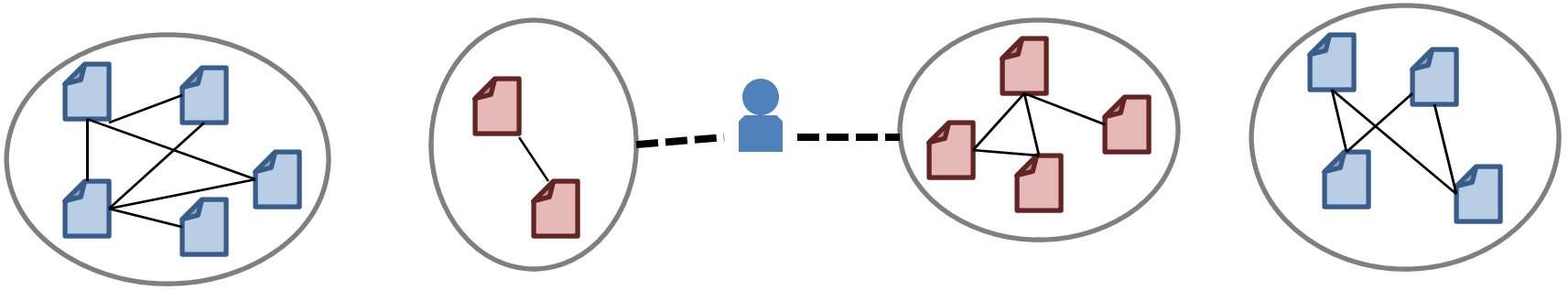
- If any probabilities above threshold, then return the publication.

Profiles RNS vs Other Methods

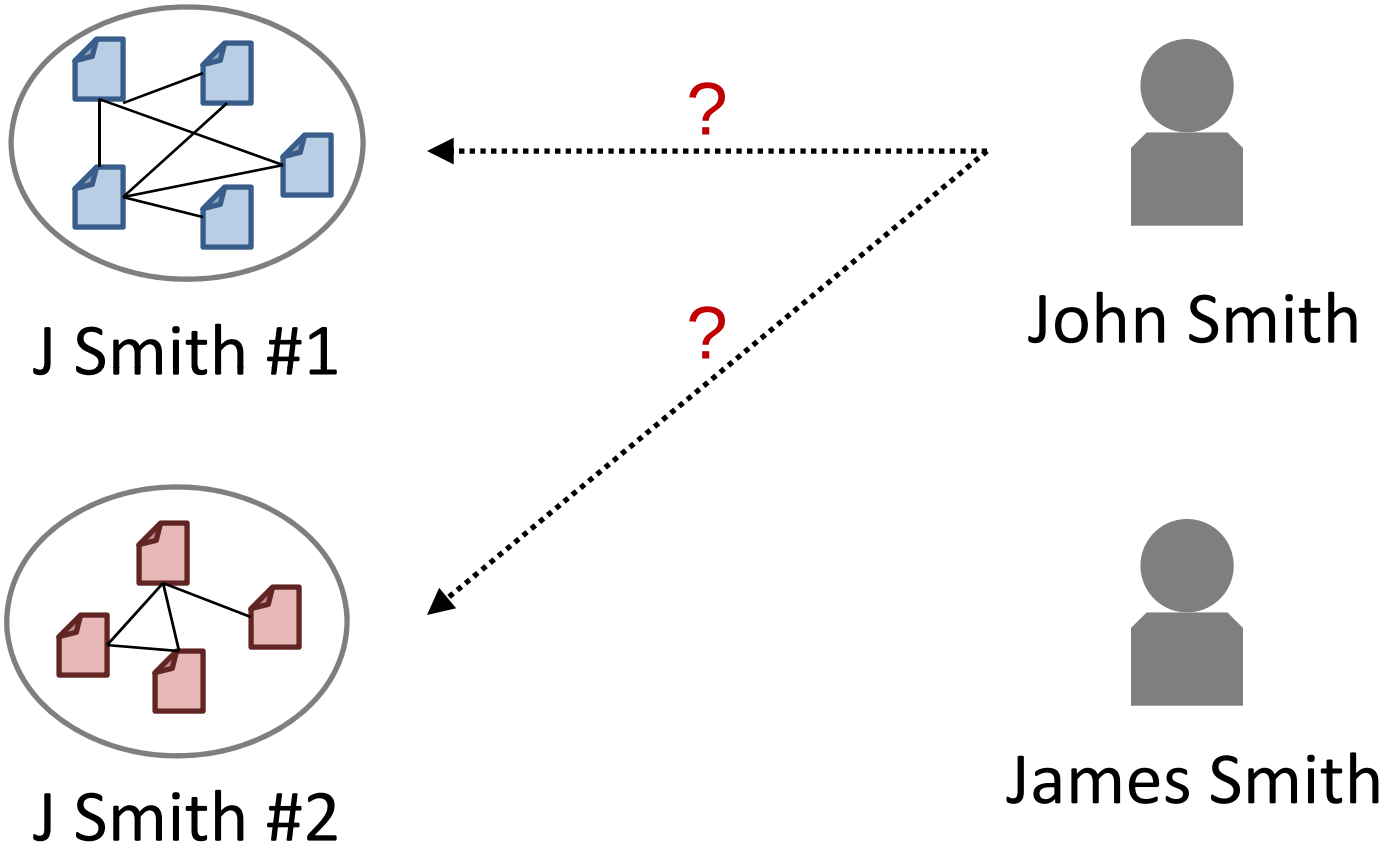
- Profiles RNS Disambiguation
  - Match person to seed publications
  - Match seed publications to possible pubs



- Other methods based on clustering
  - Group all publications into author clusters
  - Match person to clusters



“Accuracy” of Clustering Algorithms



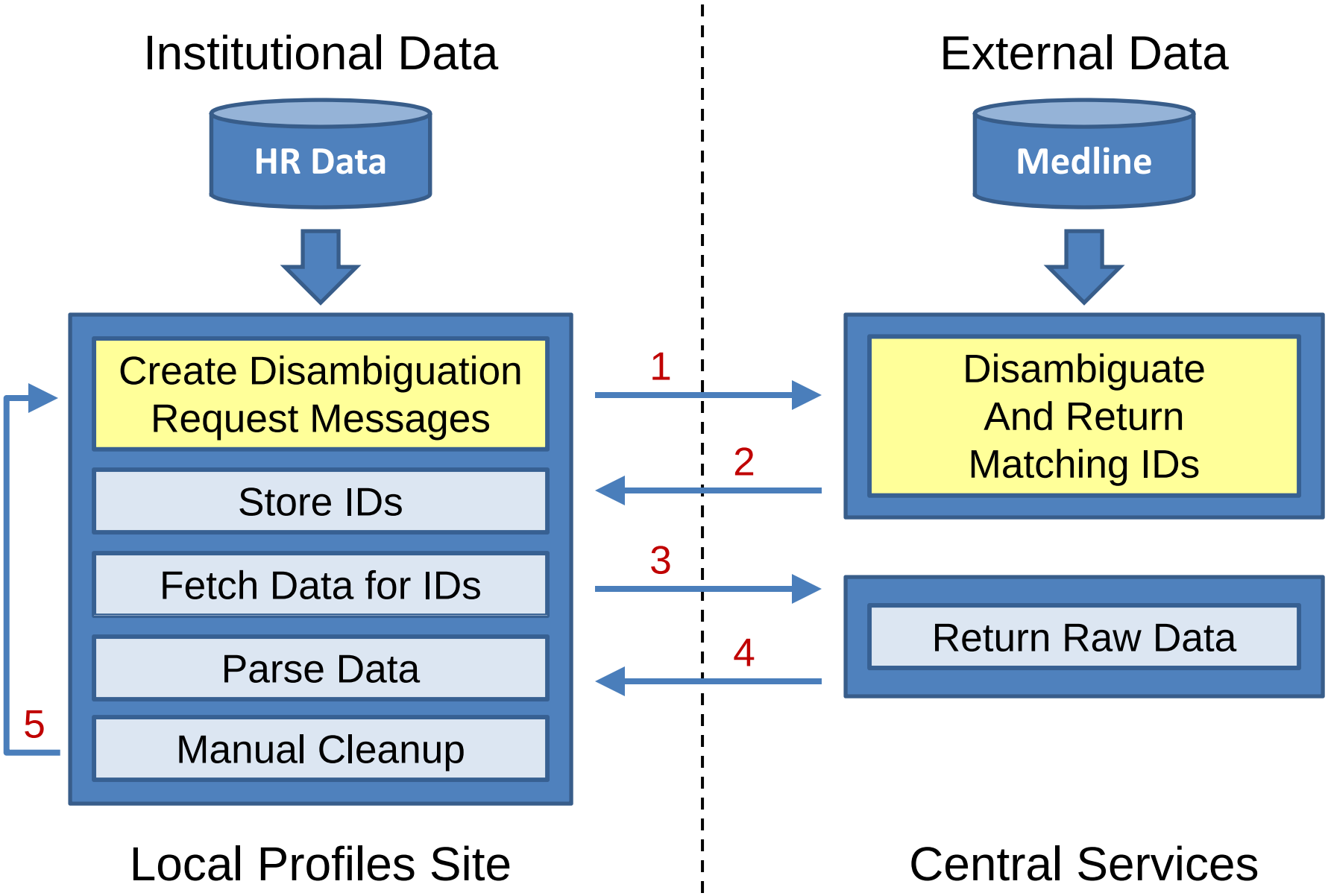
Even if clusters are 100% accurate…

…matching clusters to the correct person here is 50%

Limitations to Automation

- Best automated algorithms can only match ~85% of people. Beyond that always requires manual cleanup.
  - Bad data in Medline
  - Bad data in local HR system
  - Names too common (e.g., Chen)
  - Too little HR data (e.g., first initial only “J Smith”)
  - Too little Medline data (e.g., first names only started in 2002)
  - Hyphenated names
  - Names with foreign characters
  - Maiden vs married names
  - Nicknames (e.g., Bill vs William)
  - Common affiliation name (e.g., “Department of Medicine”)
  - Changed affiliation over time
  - Changed discipline over time
  - Keyword ambiguity (e.g., computer “virus” vs biological “virus”)

Disambiguation Architecture



Create Request Message

- Name
- Email
- Known matches (PMIDAddList)
- Known mismatches (PMIDExcludeList)
- Match threshold
- Affiliation
- Local duplicates
- Require first name

Match Threshold

- Recall
  - Fraction of actual publications that are found
  - Lowering threshold maximizes recall
  - Low recall results in people with missing or no publications
- Precision
  - Fraction of matched publications that are correct
  - Increasing threshold maximizes precision
  - Low precision causes search results & networks to be less useful
- Increasing recall causes precision to decrease
- The default threshold in Profiles RNS is 98%
- High precision allows people to find some collaborator in a field, even though it makes it harder to find all experts

Affiliation Strings

- Goal
  - Choose patterns that will select a small enough set of articles so that the name match probability becomes higher than the threshold, but not so small that you exclude potential seed articles.
- Typical Pubmed Affiliation
  - "Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02215, USA. gweber@bidmc.harvard.edu"
- Good affiliation strings
  - "%Beth Israel%Boston%"
  - "%@bidmc.harvard.edu%"
  - "%Harvard%02215%"
- Bad affiliation strings
  - "%Beth Israel%" (Many hospitals with that name)
  - "Beth Israel Deaconess Medical Center" (No wild-card characters)
  - "%BID%" (Dept of Morbid Anatomy; Biddleford, Maine)
  - Only "%@bidmc.harvard.edu%" (Too specific)

Local Duplicates

- If N people are known to have the same name, then the match probability should be divided by N.
- Harvard Medical School has:
  - 11 sets of 3 people with the same name (N = 3)
  - 134 sets of 2 people with the same name (N = 2)
  - 22,836 people with unique names (N = 1)
- Calculated automatically from Profiles database, but you might want to customize the equation if you have a list of past/expanding faculty/staff/students.

Require First Name

- Percent of author names in Medline that include full first name instead of an initial:

2011 - 84.7%	2007 - 79.3%	2003 - 75.3%	1999 - 0.1%
2010 - 84.1%	2006 - 78.0%	2002 - 74.2%	1998 - 0.1%
2009 - 82.1%	2005 - 77.5%	2001 - 0.6%	1997 - 0.1%
2008 - 80.2%	2004 - 76.3%	2000 - 0.1%	1996 - 0.1%

- If Require First Name is set to true, then:
  - Reduces recall (mostly for articles before 2002)
  - Increases precision (reduces fall positives)
- Harvard uses Require First Name when:
  - Post-docs and fellows (they have few pre-2002 publications)
  - Non-biomedical faculty (e.g., School of Engineering)

FAQ

Why does the disambiguation service not find any PMIDs for “C. William Smith”? I know he has many publications. When I search for “Smith CW” in Pubmed, it returns 100 articles. I even lowered there threshold from 98% to 70% and included some PMIDAdds, and it still does not find the articles.

FAQs

- Authors matching authors matching “C William Smith”:

C Smith (1487)
C Smith Jr (11)
C A Smith (980)
C Allen Smith (1)
C B Smith (297)
C B Smith Jr (1)
C Beebe Smith (2)
C Blake Smith (1)
C Bruce Smith (1)
C C Smith (331)
C C Smith Jr (1)
C Christopher Smith (11)
C Cory Smith (1)
C D Smith (358)
C D Smith Jr (1)
C Dahlem Smith (2)
C Dan Smith (1)
C Daniel Smith (75)
C Douglas Smith (10)
C E Smith (566)
C E Smith Jr (6)
C F Smith (112)
C G Smith (175)
C Giles Smith (1)
C Gilling Smith (1)
C H Smith (324)
C H Smith Jr (2)
C Hord Smith (1)
C I Smith (417)
C J Smith (534)

Plus 500 more!

- Probabilities that author names match “C William Smith” even within his own institution:

Author Name	Publications	Probability
C W Smith	92	0.004129
C William Smith	46	0.130883
Caitlin Smith	1	0.039193
Carrie Smith	1	0.039193
Charnetta Smith	1	0.039193
Chris Smith	1	0.034223
Clay Smith	1	0.039016
Colum Smith	1	0.039193

FAQs

- Full names work much better than initials.
- Your Dean's name is not unique. There are 5+ million authors in Medline, and many have her same name.
- Only 5% of the author names on your institution's articles are people in your instance of Profiles. The rest are former faculty or external collaborators that you have never heard about.
- You only checked the most successful people at your institution. You know that Dean C Smith should have a lot of publications, but the disambiguation service cannot distinguish Dean C Smith from Resident C Smith. You did not check Resident C Smith to see if she has no publications.

Future

- Disambiguation Service
  - Support more types of partial name matches
  - Allow date ranges on affiliation strings
  - Add citation data into algorithm
  - Add discipline into the algorithm
- Create disambiguation request message
  - Call service multiple times for people with known name variations (e.g., maiden vs married name)
  - Use different affiliation strings based on a person's institution, department, or division
  - Adjust parameters based on faculty rank and titles