Record linkage approaches in pSCANNER

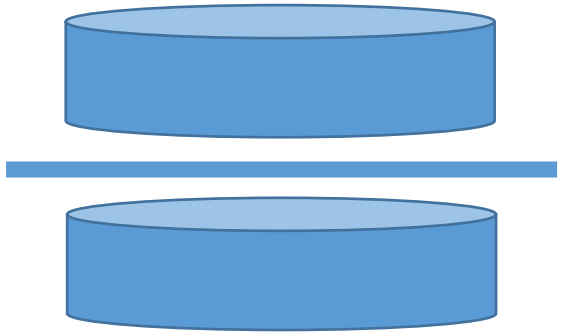
Toan Ong, PhD
Assistant Professor
Department of Pediatrics
University of Colorado, Anschutz Medical Campus

Outline

- Problem
- Challenges
- Record linkage solutions
- Looking ahead

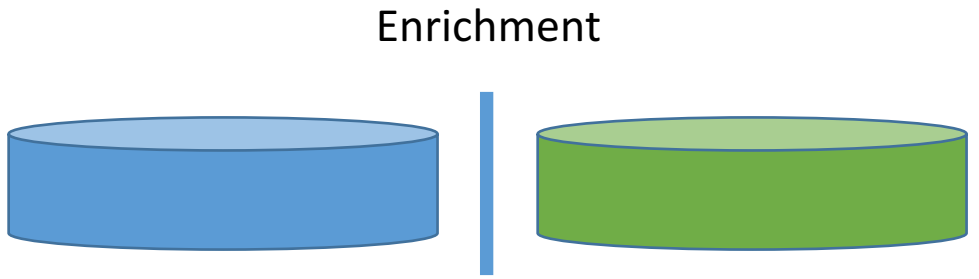
Problem

- Data for analysis are distributed across different institutions
- Horizontally partitioned data



De-duplication

- Vertically partitioned data



Example

- John is a severe chronic asthma patient who received care at both health institution A, B, and C in Colorado
- Mary is a mild asthma patient who received care at only at A
- What is the prevalence of severe asthma among patient with asthma?

prevalence = John + John + John / (3 Johns + Mary) = 75%

Instead of 50%

Definition

- Record linkage: The process of linking records that represent the same entity in one or more databases
  - ❖Objective:
    - Data completeness
    - Data de-duplication
- Privacy-preserving record linkage (PPRL): record linkage without revealing clear-text linkage data using data encryption

Challenges

- A universally shared identifier does not exist
- Clear-text linkage variables (SSN, first and last name, DOB…) are HIPAA-protected information
- Linkage data have errors (e.g., typographical errors)
- Attack to decrypt hashed data
- Lack of gold-standard linked data to test record linkage methods
- Difficult to perform linkage verification

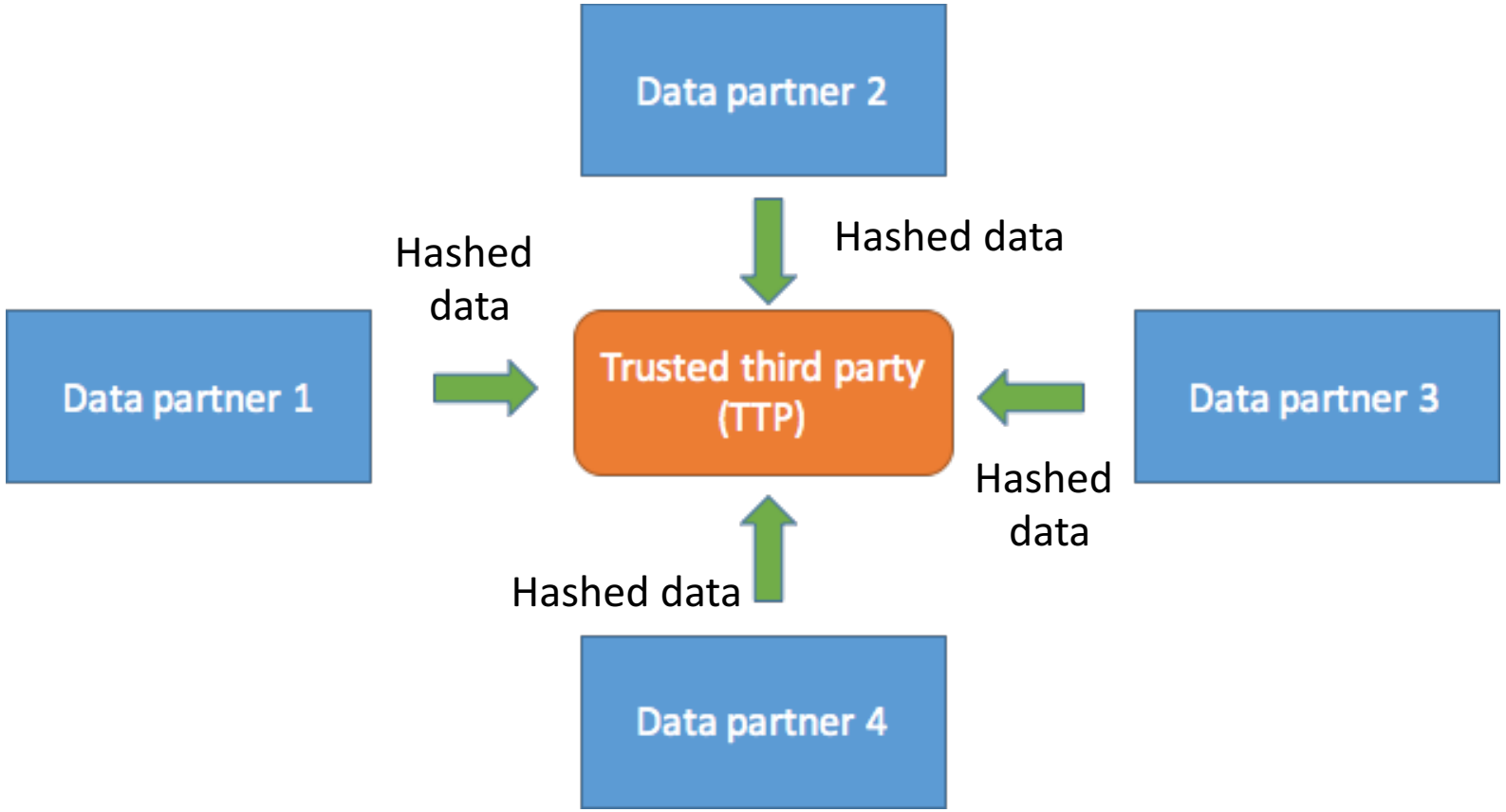
Linkage variables

- Social security number
- First name
- Last name
- Date of birth

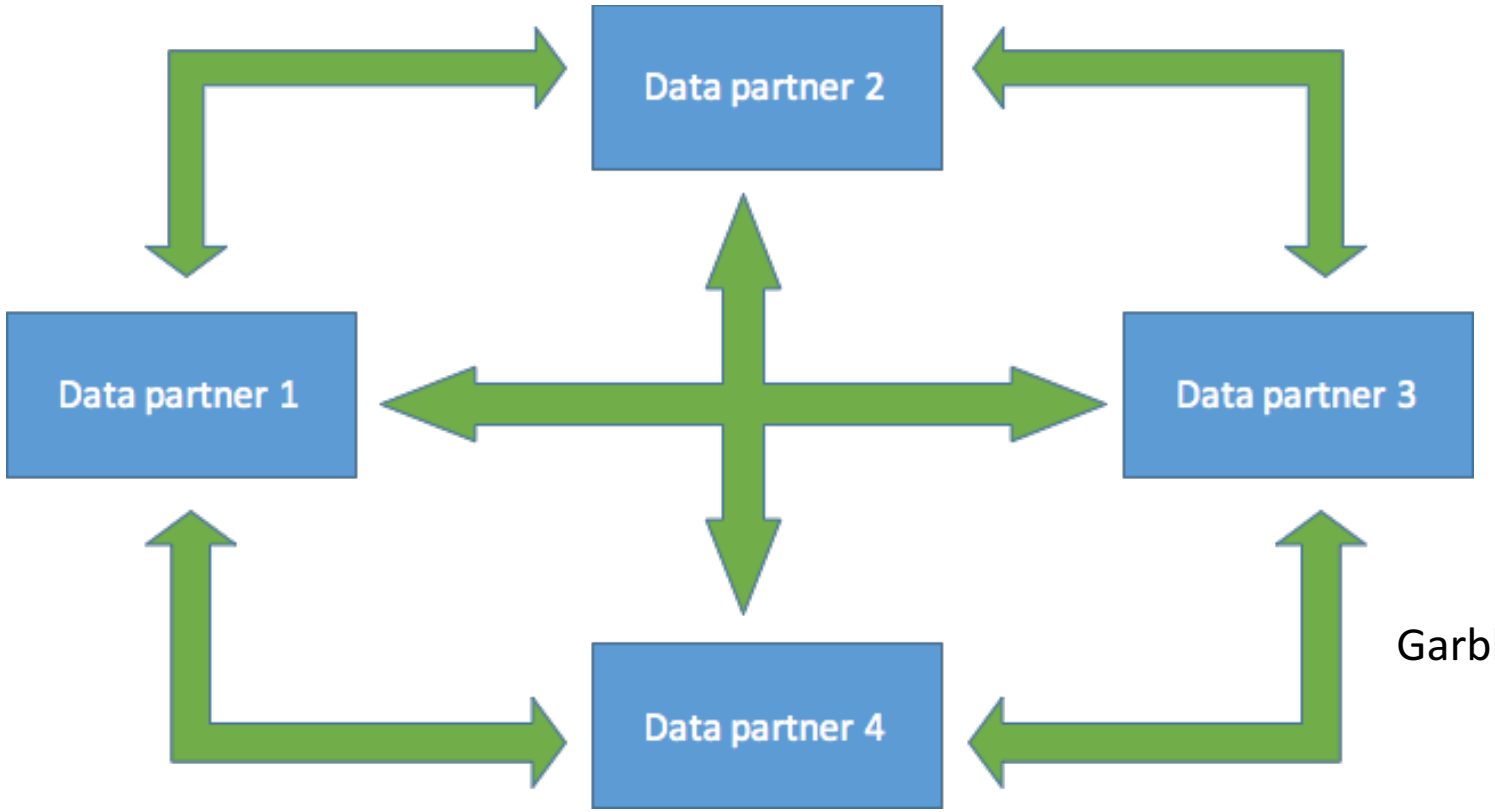
1. Seeded HashID of (First Name + Last Name + Date of Birth),
2. Seeded HashID of (Date of Birth + SSN),
3. Seeded HashID of (Last Name + SSN), or
4. Seeded HashID of (Three Letter First Name + Three Letter Last Name + Soundex First Name + Soundex Last Name + Date of Birth + SSN).

Abel et al., 2015

Record linkage approach



Record linkage approach



Garbled circuit

Record linkage methods

- Deterministic:
  - A linkage is determined by exact matching of hash value

  $\Rightarrow$ intolerant to errors in linkage data

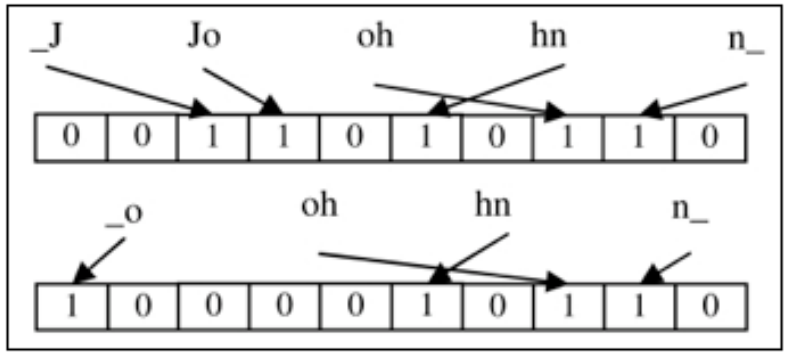
1. Seeded HashID of (First Name + Last Name + Date of Birth),
2. Seeded HashID of (Date of Birth + SSN),
3. Seeded HashID of (Last Name + SSN), or
4. Seeded HashID of (Three Letter First Name + Three Letter Last Name + Soundex First Name + Soundex Last Name + Date of Birth + SSN).

Abel et al., 2015

Record linkage methods

- Probabilistic PPRL:

Bloom filters



Johnson et al., 2010

Similarity score

Dice coefficient = $\frac{2|X \cap Y|}{|X| + |Y|}$

Probabilistic

- Effective to link data with errors
- Compatible with both TTP or pair-wise approach
- Efficiency can be improved by effective data blocking strategies

Examples of data errors

- Findings from verifying real data:
  - Typos in the value of the linkage variables
  - Nick name
  - Middle name
  - Maiden name included in last name (two-word names)
  - Prefixes and suffixes

Linkage performance (synthetic data)

- Synthetic datasets:
  - 10K records each
  - Corrupted data
  - 6K overlapping records

Approach	Method	# TP	# FP	# FN	Run time (s)
TTP	Deterministic	4,607	0	1,393	47
TTP	Probabilistic	5,757	1	243	1,038
Pairwise (GC)	Deterministic	4,067	0	1,393	13,647
Pairwise (GC)	Probabilistic	5,948	16	52	30,285

Linkage performance (synthetic data)

- Probabilistic pair-wise

Blocking variable	TP	FP	TN	Run time
LN	4643	15	1357	7,978
YOB	4842	6	1158	16,275
MOB+DOB	5407	3	593	6,030
Combined	5948	16	52	30,285

Testing on real data

Site ID	Patients	Site ID	1	2	3	4	5	6	7	8
1	22639	1	NULL	29	25	27	12	10	12	436
2	111743	2	29	NULL	754	85	50	36	64	93
3	75167	3	25	754	NULL	74	43	17	47	65
4	95217	4	27	85	74	NULL	898	186	641	185
5	86514	5	12	50	43	898	NULL	901	264	68
6	78655	6	10	36	17	186	901	NULL	232	43
7	80268	7	12	64	47	641	264	232	NULL	122
8	104286	8	436	93	65	185	68	43	122	NULL
9	107734	9	24	105	41	52	28	20	42	66
10	72295	10	50	129	58	351	165	54	167	393
11	279329	11	54	167	109	424	210	164	524	319
12	108557	12	17	33	22	65	50	41	46	81
13	73248	13	7	51	37	44	27	18	32	37
14	92804	14	11	24	22	109	73	76	189	57
15	238683	15	48	120	57	140	122	73	138	171
16	425518	16	55	134	92	112	109	66	122	241
17	171553	17	40	100	61	67	57	46	74	162
18	208741	18	57	221	112	76	47	35	85	194
19	125578	19	24	92	31	97	82	47	89	105
20	385862	20	83	940	566	223	223	126	193	322
21	42050	21	8	19	19	28	15	2	20	15
22	72934	22	11	705	916	30	28	16	30	63

Progress

- Methods
  - Deterministic PPRL
  - Probabilistic PPRL
  - Deterministic garbled circuit
  - Probabilistic garbled circuit
- Conferences
  - Challenge workshop at the Academy health concordium
  - AMIA record linkage panel
- Grant
  - PCORI letter of intent submitted

Next steps

- Test on real data
  - Using VA datasets (IRB protocol approved)
  - Using USC data (IRB protocol approved)
- Establish pSCANNER protocol for expert determination on record linkage methods
- Link data based on practical use cases
  - Linkage among pSCANNER sites
  - CDRN-PPRN linkage

Team

- Daniella Meeker, Ph.D.
- Lucila Ohno-Machado, MD, Ph.D.
- Xiaoqian Jiang, Ph.D.
- Feng Chen, Ph.D.
- Jason Doctor, Ph.D.
- Michael Kahn, MD, Ph.D.
- Lisa Schilling, MD, MSPH
- Michael E. Matheny, MD, MS, MPH
- Jaideep Vaidya, Ph.D.
- Shuang Wang Ph.D.
- Ibrahim Lazrig, Ph.D. candidate
- Dax Westerman, MS
- Tara Knight, Ph.D.

- Thank you. Questions.