Privacy-Preserving Record Linkage

Elizabeth Ashley Durham
Health Information Privacy Lab
Department of Biomedical Informatics
Vanderbilt University

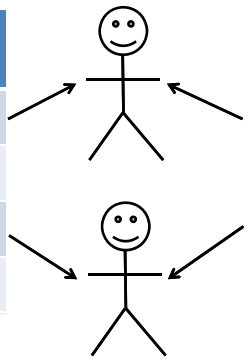
Wednesday, 24 November, 2010

1

Record linkage

Set of records from Vanderbilt

First Name	Last Name
john	**smith**
lucille	ball
bill	**clinton**
hillary	clinton



Set of records from Emory

First Name	Last Name
jon	**smyth**
taylor	swift
william	**clinton**
jon	bon jovi

2

Privacy-preserving record linkage (PPRL)

Set of records from Vanderbilt

First Name	Last Name
john	**smith**
lucille	ball
bill	**clinton**
hillary	clinton

POLICY

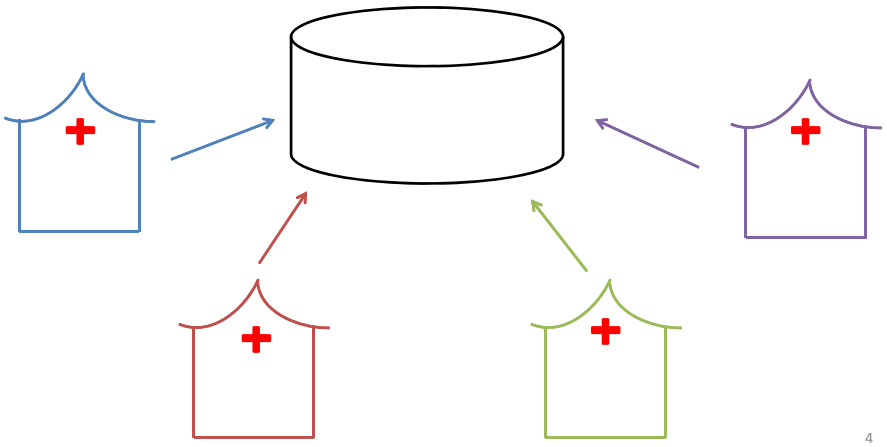
Set of records from Emory

First Name	Last Name
jon	**smyth**
taylor	swift
william	**clinton**
jon	bon jovi

3

PPRL applications in healthcare

sharing patient data for research



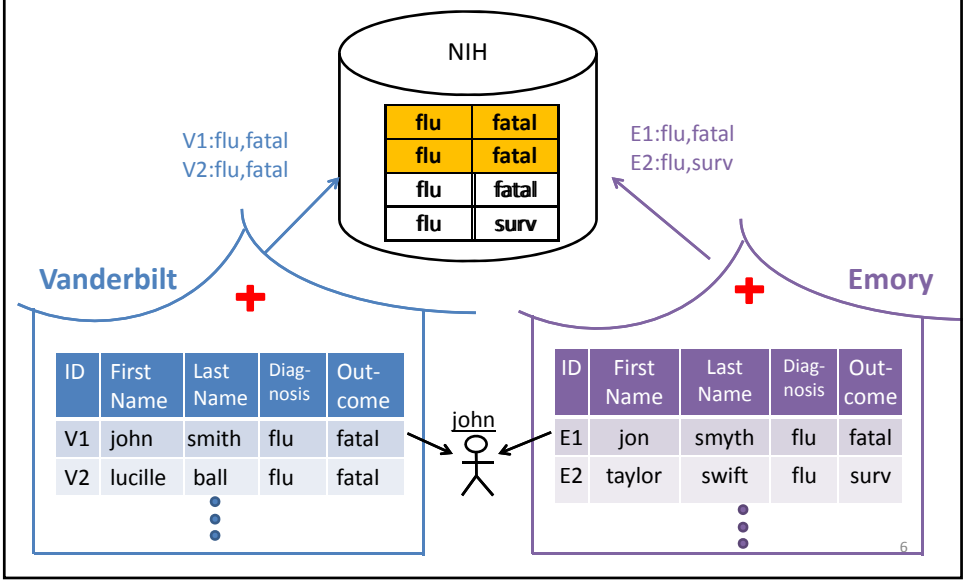
4

The NIH requires researchers share de-identified patient data

- **U.S. National Institutes of Health (NIH) data sharing policy**
  - "Data should be made as widely & freely available as possible"
  - Researchers who receive ≥ $500,000 **must** develop a data sharing plan or describe why data sharing is not possible
  - Derived data must be shared in a manner that is devoid of "identifiable information"
- **NIH supported genome-wide association studies policy**
  - Researchers funded for genome-wide association studies must share data

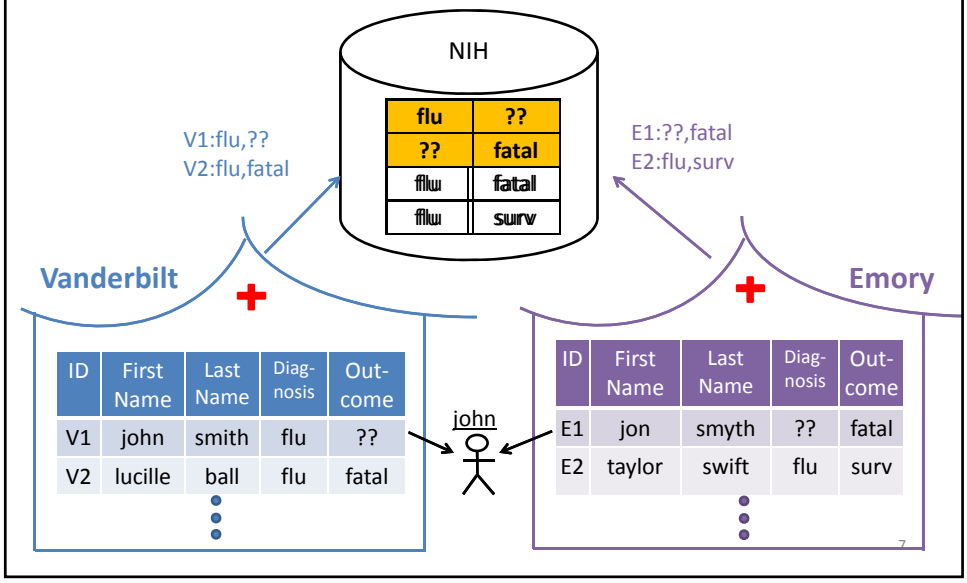
5

Duplicates: a flaw in the current model for sharing de-identified data

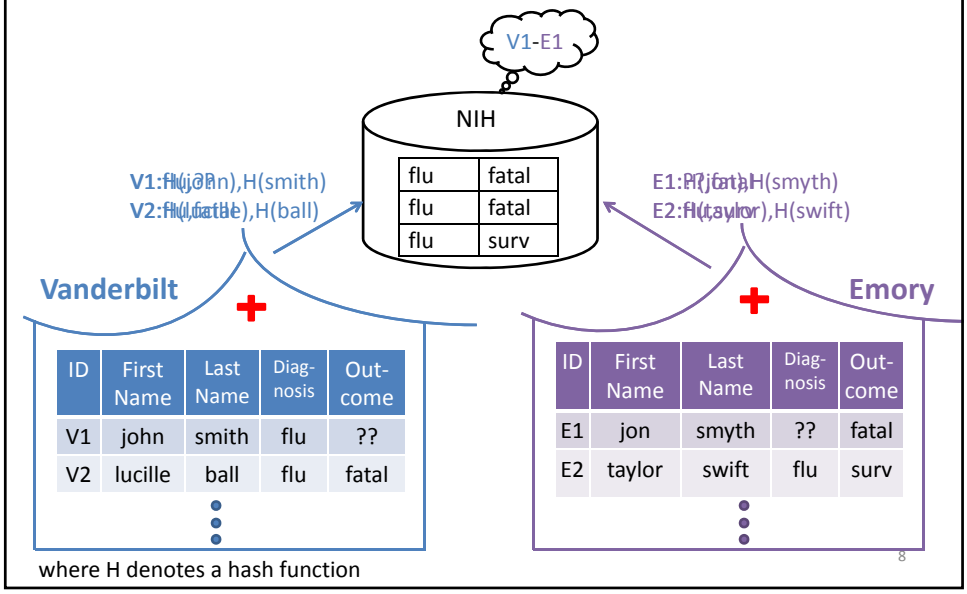


6

Fragmented data: a flaw in the current model for sharing de-identified data



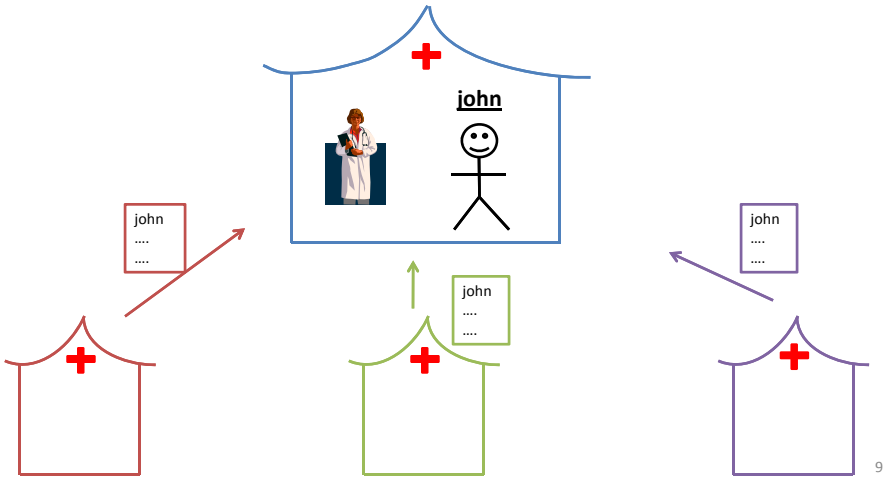
PPRL can improve the model for sharing de-identified data and enable more effective medical research



where H denotes a hash function

PPRL applications in healthcare

improving patient care



9

Other PPRL applications

- Business
- Counter-terrorism efforts

10

Roadmap

- Definition
- Motivation
- Record linkage
- Privacy-preserving record linkage
  - Background
  - Experimental design
  - Experimental results
  - Discussion
  - Open questions in record linkage
  - Conclusion

11

Roadmap

- Definition
- Motivation
- Record linkage
- Privacy-preserving record linkage
  - Background
  - Experimental design
  - Experimental results
  - Discussion
  - Open questions in record linkage
  - Conclusion

12

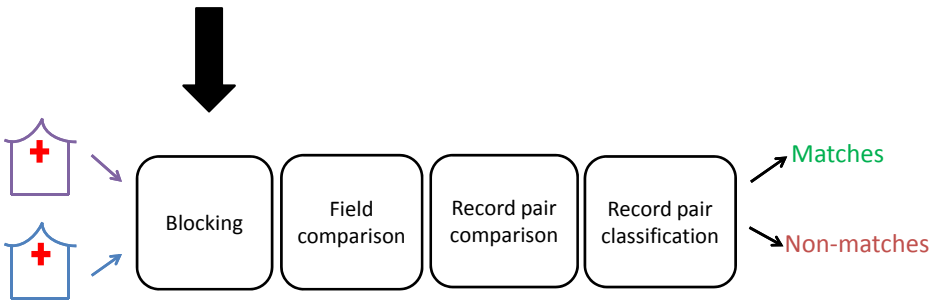
Steps in record linkage

A few assumptions…

1) common schema
2) common method of data standardization
3) records from an institution have been deduplicated (*i.e.*, record linkage has been applied within each institution such that an individual is represented by only a single record within an institution)

13

Steps in record linkage



14

Blocking: sample dataset

Set of records from Vanderbilt

First Name	Last Name
john	smith
lucille	ball
bill	clinton
hillary	clinton

Set of records from Emory

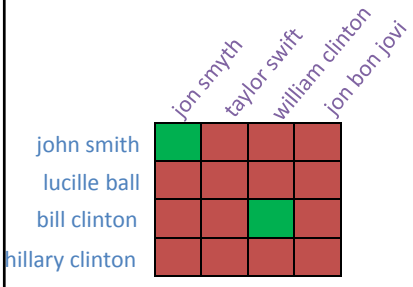
First Name	Last Name
jon	smyth
taylor	swift
william	clinton
jon	bon jovi

15

Blocking

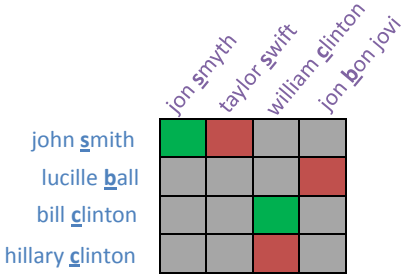
= match
= non-match

no blocking



|Vanderbilt| |Emory| = 16 record pair comparisons

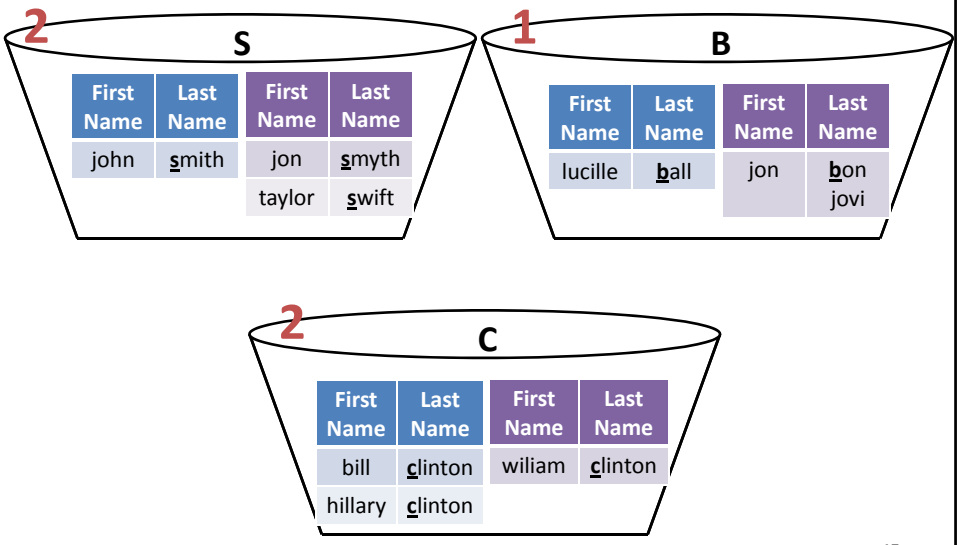
blocking
(first letter of last name)



5 record pair comparisons

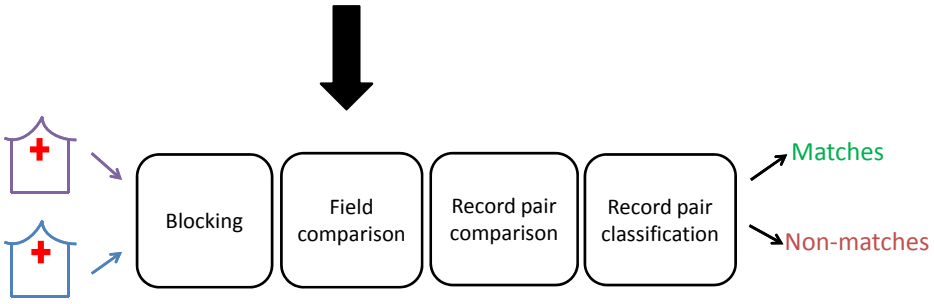
16

Blocking: another perspective



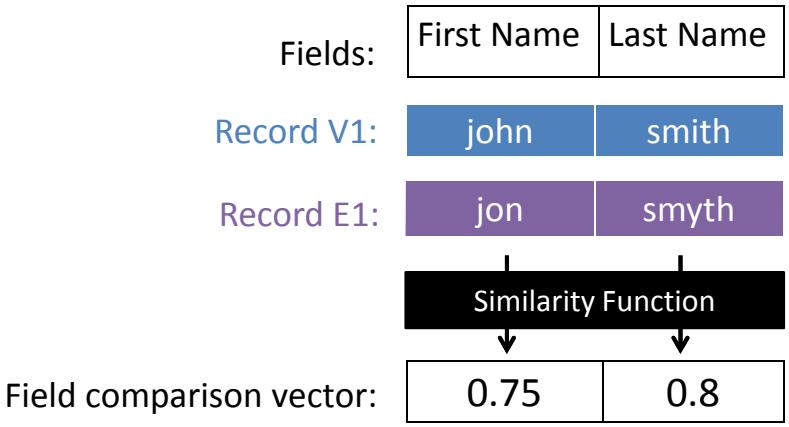
17

Steps in record linkage



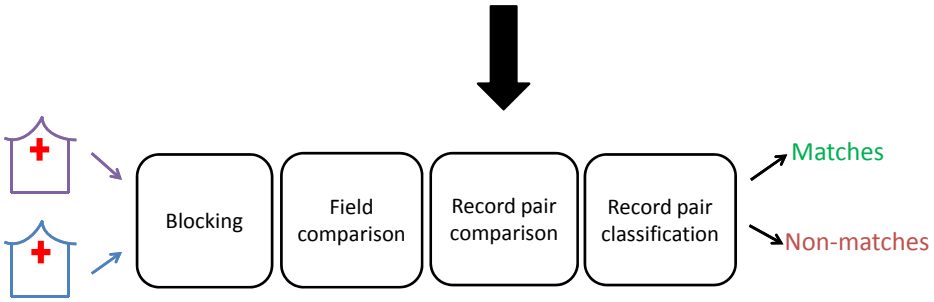
18

The field comparison step of record linkage



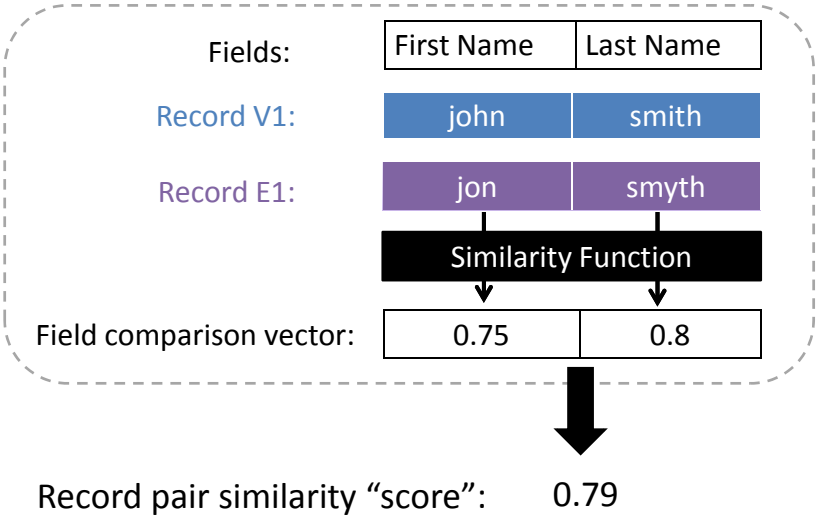
19

Steps in record linkage



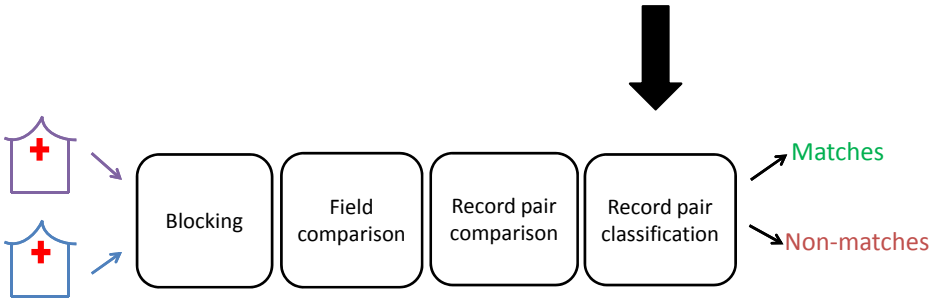
20

The record pair comparison step of record linkage



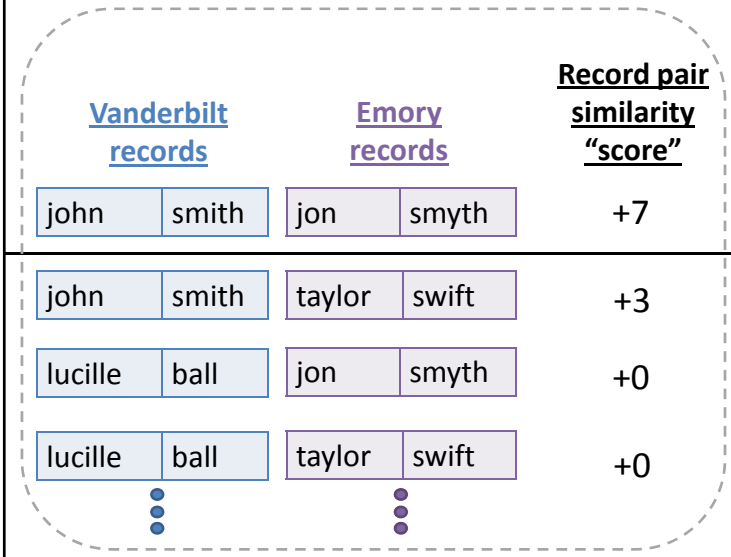
21

Steps in record linkage



22

The record pair classification step of record linkage



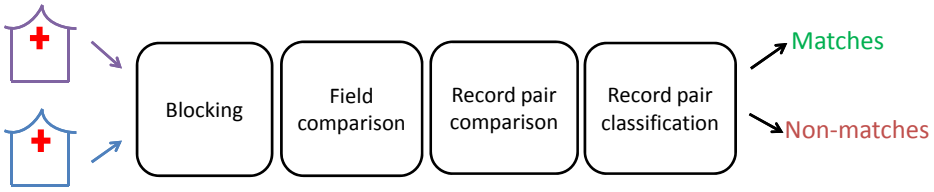
23

Roadmap

- Definition
- Motivation
- Record linkage
- Privacy-preserving record linkage
  - Background
  - Experimental design
  - Experimental results
  - Discussion
  - Open questions in record linkage
  - Conclusion

24

How do we do all of this in a privacy-preserving manner?



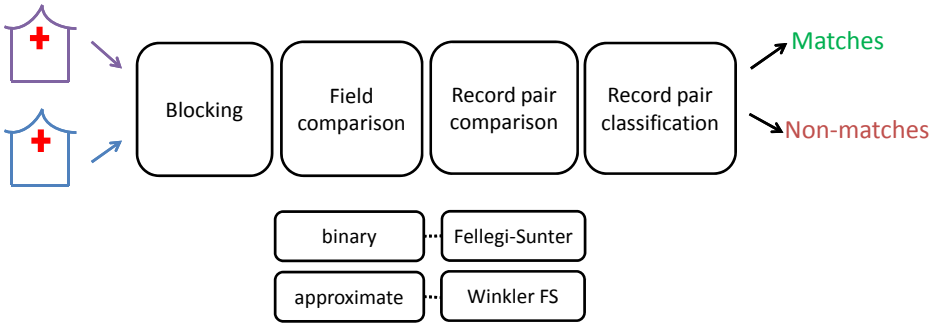
25

Roadmap

- Definition
- Motivation
- Record linkage
- Privacy-preserving record linkage
  - Background
  - Experimental design
  - Experimental results
  - Discussion
  - Open questions in record linkage
  - Conclusion

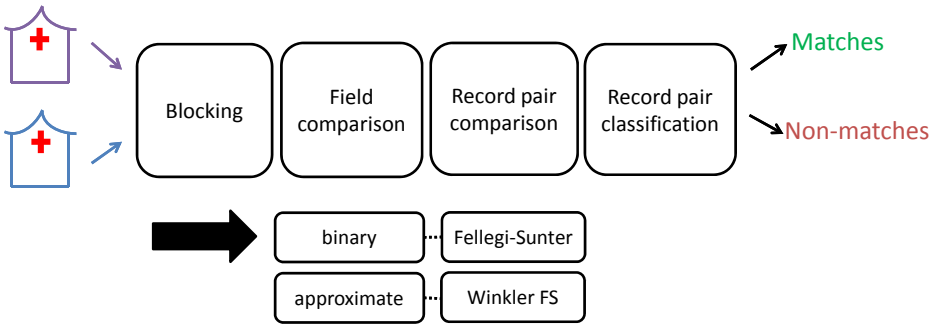
26

Background



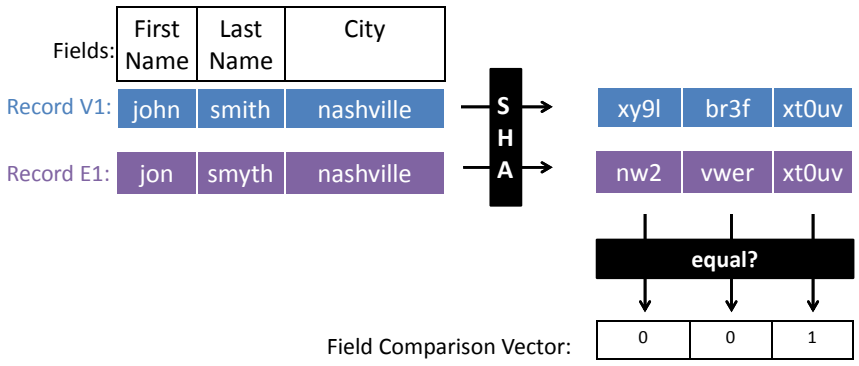
27

Background



28

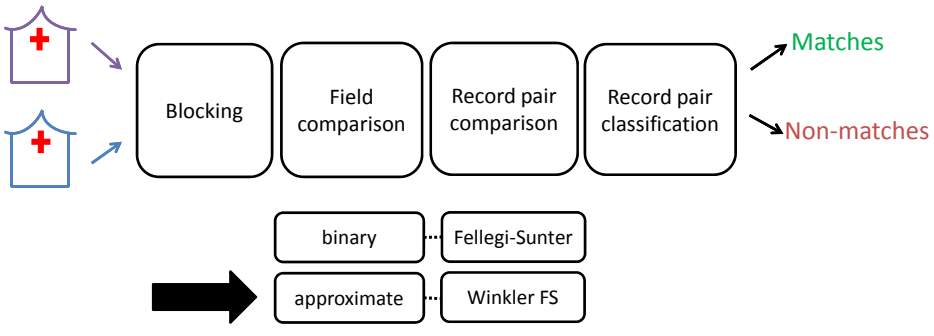
Binary Field Comparison



where SHA refers to the Secure Hash Algorithm

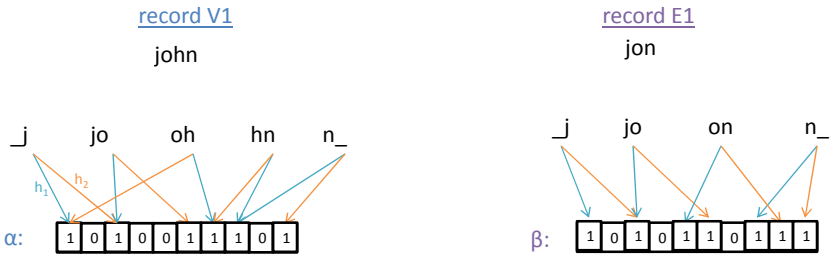
29

Background



30

Approximate Field Comparison

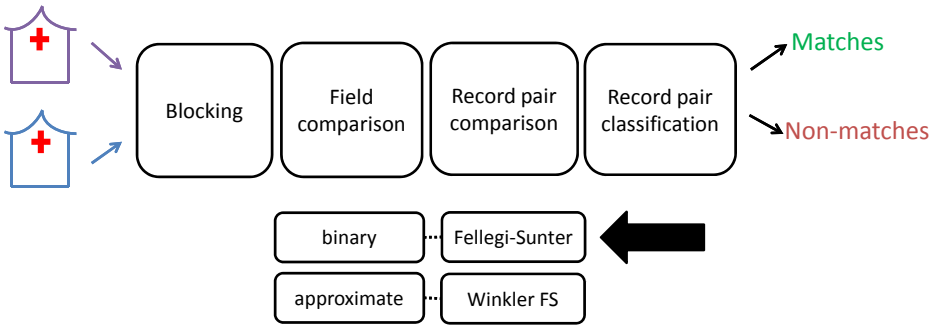


$$Dice\ coefficient(\alpha,\beta) = 2\left(\frac{|\alpha \cap \beta|}{|\alpha| + |\beta|}\right) = \frac{2*5}{13} = 0.77$$

Schnell 2009

31

Background



32

Fellegi-Sunter (FS)

Conditional probability vectors:

Fields:	First Name	Last Name
Match:	0.8	0.9
Non-match:	0.05	0.02

$\log \frac{0.9}{0.02}$

Weight vectors:

Fields:	First Name	Last Name
Agreement:	1.2	1.95
Disagreement:	-0.68	-1

$\log \frac{1 - 0.9}{1 - 0.02}$

Fellegi 1969

33

Fellegi-Sunter (FS)

Fields:	First Name	Last Name
Agreement weights:	1.2	1.95
Disagreement weights:	-0.68	-1

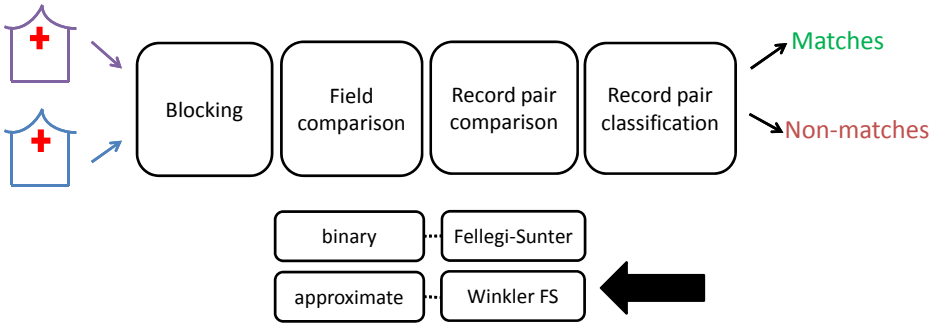
Fields:	First Name	Last Name
Record V1:	john	smith
Record E1:	jon	smyth
Field comparison vector:	0	0
Weight vector:	-0.68	-1

Record pair similarity score: -1.68 (Σ)

Fellegi 1969

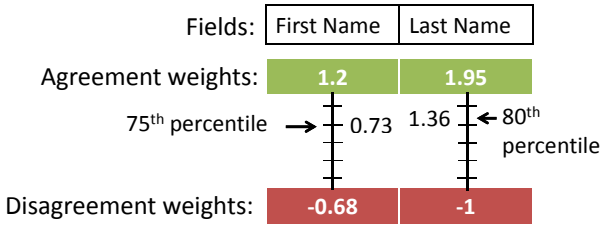
34

Background



35

Winkler's Modificaton to Fellegi-Sunter



Fields:	First Name	Last Name
Record V1:	john	smith
Record E1:	jon	smyth
Field comparison vector:	0.75	0.8
Weight vector:	**0.73**	1.36
Record pair similarity score:	2.09	

Porter 1997

36

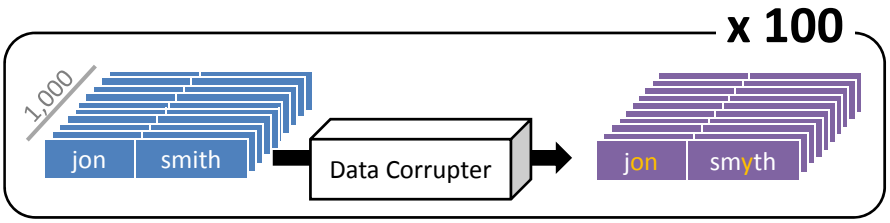
Roadmap

- Definition
- Motivation
- Record linkage
- Privacy-preserving record linkage
  - Background
  - Experimental design
  - Experimental results
  - Discussion
  - Open questions in record linkage
  - Conclusion

37

Experimental design

- **Dataset: ~**6 million records from the North Carolina Voter Registration dataset



- **Fields:**

First name	Middle name	Last name	Birth state	Sex	Race	Street name	Street type	Street suffix	City	State

- **Computational resources:** 2 GHz dual core PC with 4GB of memory

Pudjijono 2009

38

Roadmap

- Definition
- Motivation
- Record linkage
- Privacy-preserving record linkage
  - Background
  - Experimental design
  - Experimental results
  - Discussion
  - Open questions in record linkage
  - Conclusion

39

Experimental results: accuracy

results withheld pending publication

40

Experimental results: run time

results withheld pending publication

41

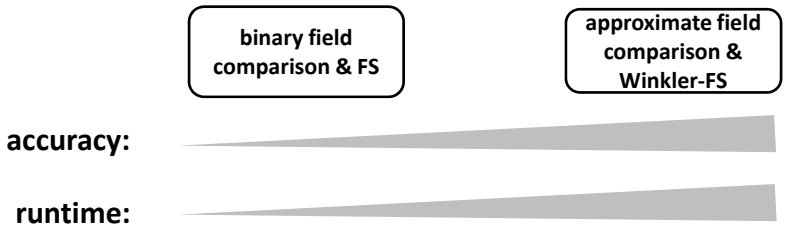
Roadmap

- Definition
- Motivation
- Record linkage
- Privacy-preserving record linkage
  - Background
  - Experimental design
  - Experimental results
  - Discussion
  - Open questions in record linkage
  - Conclusion

42

Discussion

Discussion



Limitations

- Controlled environment

43

Roadmap

- Definition
- Motivation
- Record linkage
- Privacy-preserving record linkage
  - Background
  - Experimental design
  - Experimental results
  - Discussion
  - Open questions in record linkage
  - Conclusion

44

Open questions in record linkage

1. Enforcing one-to-one linkage
2. Decentralized record linkage

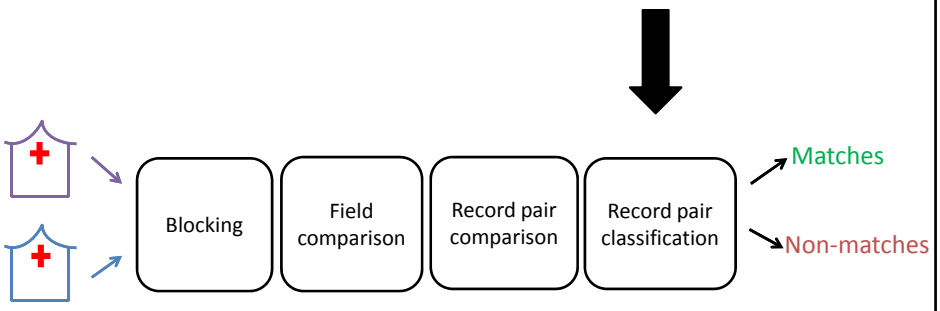
45

Open questions in record linkage

➡ 1. Enforcing one-to-one linkage
2. Decentralized record linkage

46

One-to-one linkage



47

Reminder: the record pair classification step of record linkage

Vanderbilt records		Emory records		Record pair similarity "score"	Record pair classification
john	smith	jon	smyth	+7	Match
john	smith	taylor	swift	+3	Non-match
lucille	ball	jon	smyth	+0	Non-match
lucille	ball	taylor	swift	+0	Non-match

48

One-to-one linkage: sample dataset

Set of records from Vanderbilt

First Name	Last Name	City
john	smith	nashville
bill	clinton	washington dc
hillary	clinton	washington dc

Set of records from Emory

First Name	Last Name	City
jon	smyth	nashville
taylor	swift	nashville
william	clinton	washington dc

49

One-to-one linkage

set of records from Vanderbilt			set of records from Emory			score	classification
bill	clinton	washington dc	william	clinton	washington dc	+2	Match
hillary	clinton	washington dc	william	clinton	washington dc	+2	Match
john	smith	nashville	william	clinton	washington dc	+0	Non-match
bill	clinton	washington dc	taylor	swift	nashville	+0	Non-match
hillary	clinton	washington dc	taylor	swift	nashville	+0	Non-match
john	smith	nashville	taylor	swift	nashville	+1	Non-match
bill	clinton	washington dc	jon	smyth	nashville	+0	Non-match
hillary	clinton	washington dc	jon	smyth	nashville	+0	Non-match
john	smith	nashville	jon	smyth	nashville	+1	Non-match

50

One-to-one linkage

predicted

First Name	Last Name	City
john	smith	nashville
bill	clinton	washington dc
hillary	clinton	washington dc

First Name	Last Name	City
jon	smyth	nashville
taylor	swift	nashville
william	clinton	washington dc

actual

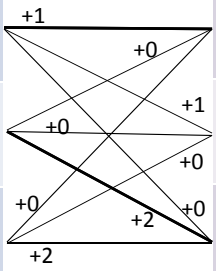
First Name	Last Name	City
john	smith	nashville
bill	clinton	washington dc
hillary	clinton	washington dc

First Name	Last Name	City
jon	smyth	nashville
taylor	swift	nashville
william	clinton	washington dc

51

One-to-one linkage

First Name	Last Name	City
john	smith	nashville
bill	clinton	washington dc
hillary	clinton	washington dc



First Name	Last Name	City
jon	smyth	nashville
taylor	swift	nashville
william	clinton	washington dc

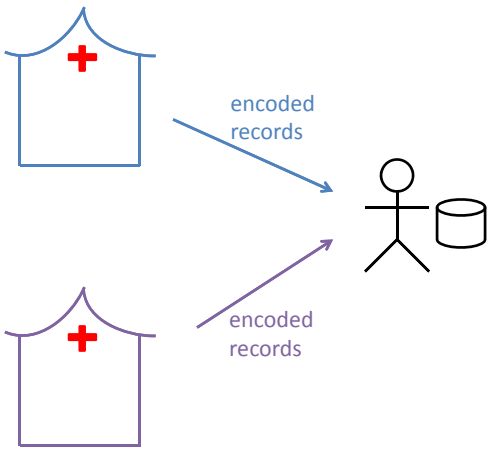
52

Open questions in record linkage

1. Enforcing one-to-one linkage
2. Decentralized record linkage

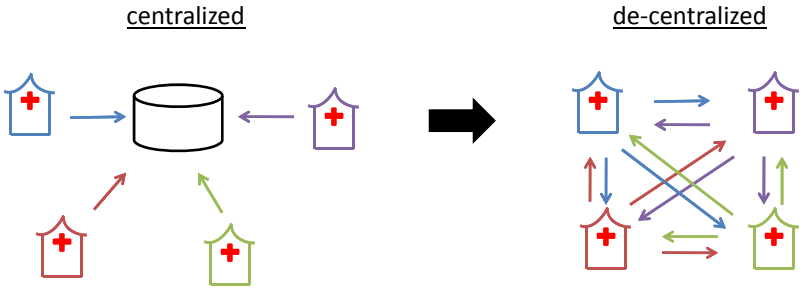
53

Centralized framework



54

De-centralization of record linkage



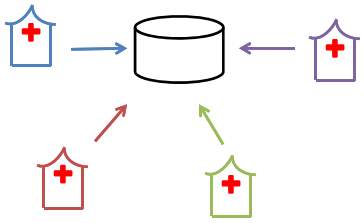
55

Roadmap

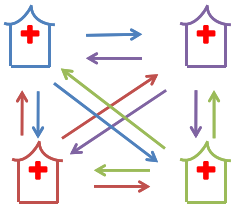
- Definition
- Motivation
- Record linkage
- Privacy-preserving record linkage
  - Background
  - Experimental design
  - Experimental results
  - Discussion
  - Open questions in record linkage
  - Conclusion

56

Conclusion



Privacy-preserving record linkage can inform and improve medical research



Privacy-preserving record linkage can improve patient care

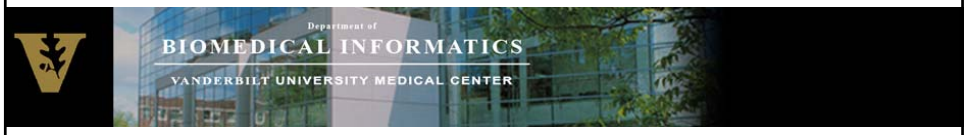
57

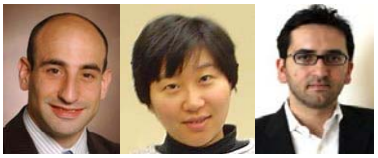
References

- Christen P, Pudjijono A. Accurate Synthetic Generation of Realistic Personal Information. *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. 2009.
- Fellegi I, Sunter A. A theory for record linkage. *J Amer Stat Assoc.* 1969; 64: 1183–210.
- Porter E, Winkler W, Approximate string comparison and its effect on an advanced record linkage system, Research Report RR97/02, U.S. Census Bureau, 1997.
- Schnell R, Bachteler T, and Reiher J. Privacy-preserving record linkage using Bloom filters. *BMC Medical Informatics and Decision Making* (9). 2009.

58

Acknowledgements





U.S. National Library of Medicine grant 2-T15LM07450-06
U.S. National Institutes of Health R01 LM009989

59

Thank You

Contact:

ea.durham@vanderbilt.edu

http://hiplab.mc.vanderbilt.edu/

Publications:

- E Durham, M Kantarcioglu, Y Xue, and B Malin. Private medical record linkage with approximate matching. *Proceedings of the American Medical Informatics Association*. 2010 November.

60