AHLA

Is Anonymization of Genomic Information and Other Big Data a Fallacy?

Daniel C. Barth-Jones
Assistant Professor of Clinical Epidemiology
Columbia University, Mailman School of Public Health
New York, NY

Barbara J. Evans
Mary Ann and Lawrence E. Faust Professor of Law, Professor, Electrical and Computer Engineering Director, Center for Biotechnology & Law
University of Houston
Houston, TX

Charles Sawyers
Chair of the Human Oncology and Pathogenesis Program
Memorial Sloan Kettering Cancer Center
New York, NY

Academic Medical Centers and Teaching Hospitals Institute • January 30-31, 2020

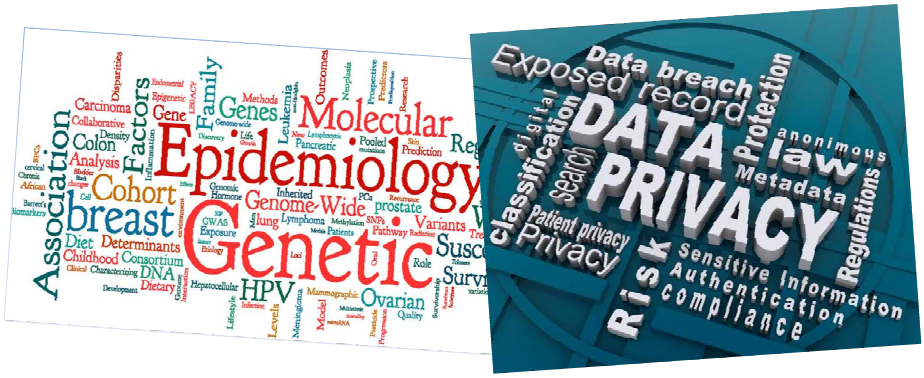
Is Anonymization of Genomic Information and Other Big Data a Fallacy?

Dissecting De-Identification: Myths in the Media

Daniel C. Barth-Jones, M.P.H., Ph.D.
Assistant Professor of Clinical Epidemiology,
Mailman School of Public Health
Columbia University
E-mail: db2431@Columbia.edu

1

A Historic and Important Societal Debate is underway…



Public Policy Collision Course

2

2

The Research Value of De-identified Health Data

3

3

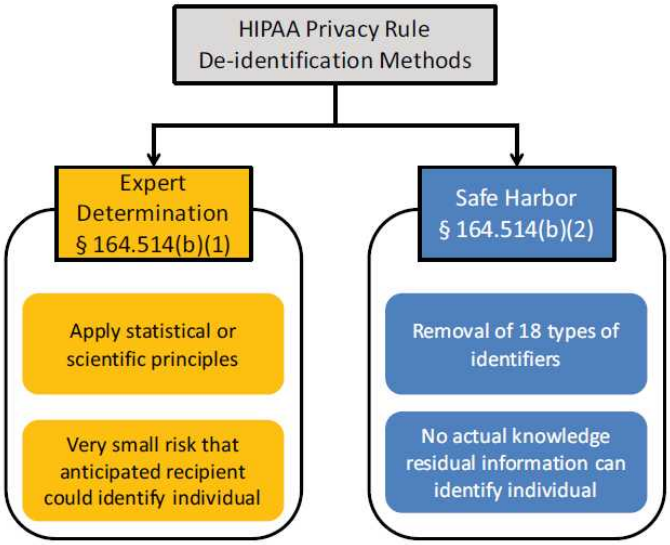
The Societal Value of De-identified Data

- Properly de-identified health data is an ***invaluable "public good"***. *The broad availability of de-identified data is an essential tool for society supporting scientific innovation and health system improvement and efficiency.*
- De-identified data does and can serve as the engine driving forward innumerable essential health systems improvements: quality improvement, health systems planning, healthcare fraud, waste and abuse detection, and medical/public health research (e.g. comparative effectiveness research, adverse drug event monitoring, patient safety improvements and reducing health disparities).
- De-identified health data greatly benefits our society and provides strong privacy protections for the individuals. As the promise of EHRs and Health IT yields richer de-identified clinical data, the progress of our nation's healthcare reform will likely be built on a foundation of such de-identified health data.

4

4

Two Methods of HIPAA De-identification



5

5

HIPAA Expert Determination Conditions

§164.514(b)(1)

- "Risk is ***very small…"***
  - —"that the ***information could be used"…***
  - —"alone or ***in combination with other reasonably available information"…,***
  - —***"by an anticipated recipient"…***
  - —***"to identify an individual"…***

6

6

HIPAA §164.514(b)(2)(i) -18 "Safe Harbor" Exclusions

All of the following must be **removed in order** for the information **to be** considered **de-identified.**

(2)(i) The **following identifiers of the individual or of relatives, employers, or household members** of the individual, are removed:

(A) Names;
(B) All **geographic subdivisions smaller than a State**, including street address, city, county, precinct, zip code, and their equivalent geocodes, **except for the initial three digits of a zip code** if, according to the current publicly available data from the Bureau of the Census: (*1*) The geographic unit formed by combining all zip codes with the same three initial digits contains **more than 20,000 people**; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
(C) **All elements of dates (except year)** for dates directly related to an individual, including **birth date**, **admission date**, **discharge date, date of death**; and **all ages over 89** and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
(D) Telephone numbers;
(E) Fax numbers;
(F) Electronic mail addresses;
(G) Social security numbers;
(H) Medical record numbers;
(I) Health plan beneficiary numbers;
(J) Account numbers;
(K) Certificate/license numbers;
(L) Vehicle identifiers and serial numbers, including license plate numbers;
(M) Device identifiers and serial numbers;
(N) Web Universal Resource Locators (URLs);
(O) Internet Protocol (IP) address numbers;
(P) Biometric identifiers, including finger and voice prints;
(Q) Full face photographic images and any comparable images; and
(R) **Any other unique identifying number, characteristic, or code** except as permitted in §164.514(c)

7

7

57 UCLA LAW REVIEW 1701 (2010)

BROKEN PROMISES OF PRIVACY: RESPONDING TO THE SURPRISING FAILURE OF ANONYMIZATION

Paul Ohm*

Computer scientists have recently undermined our faith in the privacy-protecting power of anonymization, the name for techniques that protect the privacy of individuals in large databases by deleting information like names and social security numbers. These scientists have demonstrated that they can often "reidentify" or "deanonymize" individuals hidden in anonymized data with astonishing ease. By understanding this research, we realize we have made a mistake, labored beneath a fundamental misunderstanding, which has assured us much less privacy than we have assumed. This mistake pervades nearly every information privacy law, regulation, and debate, yet regulators and legal scholars have paid it scant attention. We must respond to the surprising failure of anonymization, and this Article provides the tools to do so.

8

8

Misconceptions about HIPAA De-identified Data:

"It doesn't work…" "easy, cheap, powerful re-identification" (Ohm, 2009 "*Broken Promises of Privacy*")

***Pre-HIPAA** Re-identification Risks {Zip5, Birth date, Gender} able to identify 87%?, 63%, 28%? of US Population* (Sweeney, 2000, Golle, 2006, Sweeney, 2013)

- Reality: HIPAA compliant de-identification provides important privacy protections
  - Safe harbor re-identification risks have been estimated at 0.04% (4 in 10,000) (Sweeney, NCVHS Testimony, 2007)
- Reality: Under HIPAA de-identification requirements, re-identification is expensive and time-consuming to conduct, requires substantive computer/mathematical skills, is rarely successful, and usually uncertain as to whether it has actually succeeded

9

9

Misconceptions about HIPAA De-identified Data:

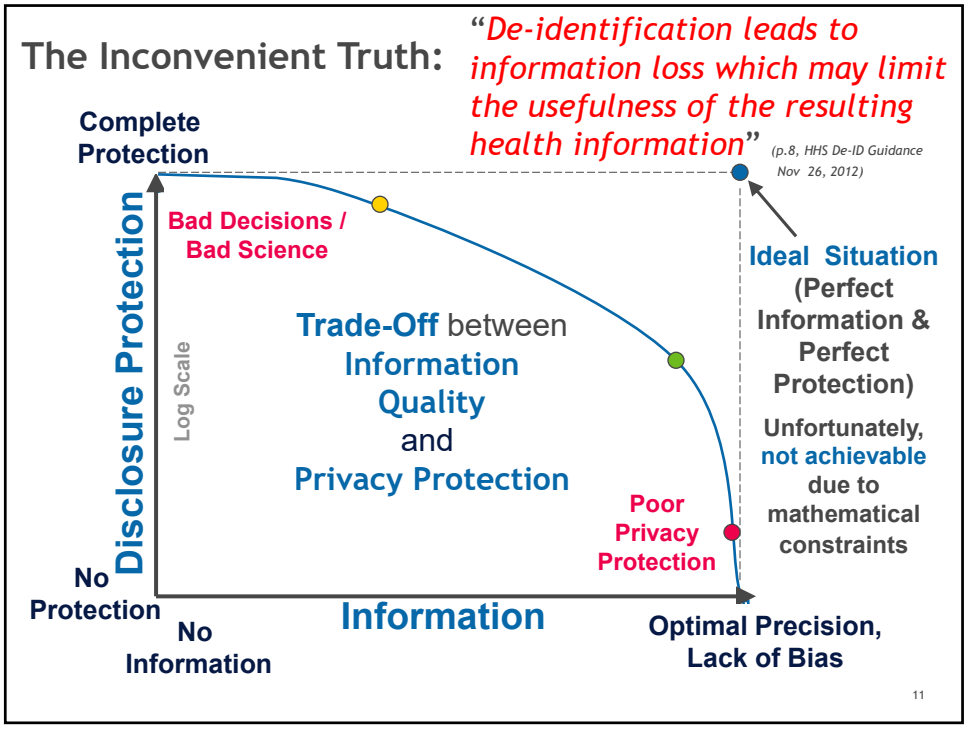
"It works perfectly and permanently…"

- Reality:
  - Perfect de-identification is not possible.
  - De-identifying does not free data from all possible subsequent privacy concerns.
  - Data is never permanently "de-identified"…

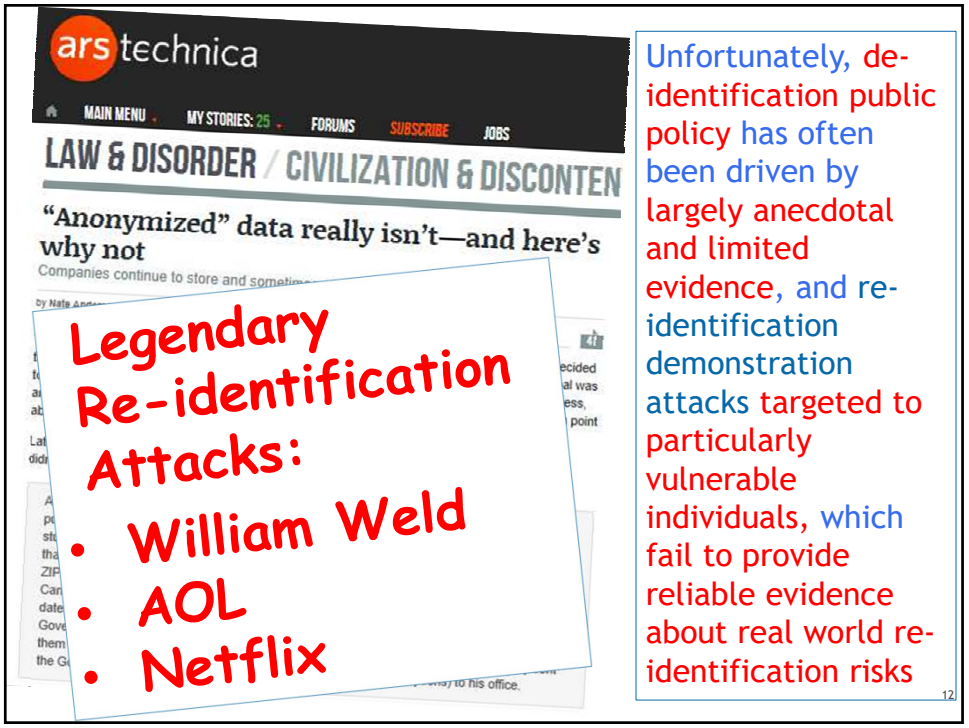
    There is no 100% guarantee that de-identified data will remain de-identified regardless of what you do with it after it is de-identified.

10

10


The Inconvenient Truth:
"De-identification leads to information loss which may limit the usefulness of the resulting health information" (p.8, HHS De-ID Guidance Nov 26, 2012)
Complete Protection
Disclosure Protection
Log Scale
Bad Decisions / Bad Science
Trade-Off between Information Quality and Privacy Protection
Ideal Situation (Perfect Information & Perfect Protection)
Unfortunately, not achievable due to mathematical constraints
Poor Privacy Protection
No Protection
No Information
Information
Optimal Precision, Lack of Bias
11

11


ars technica
MAIN MENU
MY STORIES: 25
FORUMS
SUBSCRIBE
JOBS
LAW & DISORDER / CIVILIZATION & DISCONTEN
"Anonymized" data really isn't—and here's why not
Companies continue to store and sometim
Legendary Re-identification Attacks:
• William Weld
• AOL
• Netflix
Unfortunately, de-identification public policy has often been driven by largely anecdotal and limited evidence, and re-identification demonstration attacks targeted to particularly vulnerable individuals, which fail to provide reliable evidence about real world re-identification risks
12

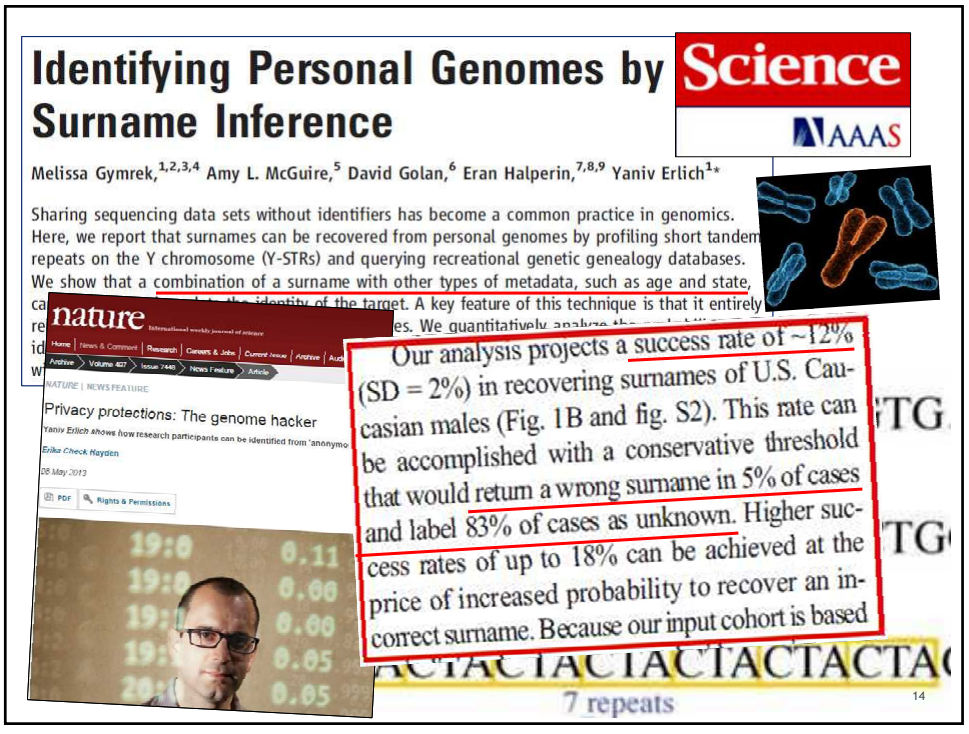
12

Re-identification Demonstration Attack Summary

Re-identification Attacks	Quasi-Identifers (w/ HIPAA Safe Harbor exclusion data in Red)	Vulnerable Subgroup Targeted?	Used Stat. Sampling	Individuals w/ Alleged/Verified Re-identification	At-Risk Sample Size	Notable Headlines & Quotes	Attack Against HIPAA Compliant (or SDL Protected) Data?	Demonstrated Re-identification Risk
Governor Weld [1,2]	Zip5, Gender, DoB	Yes	No	n=1	99,500	*"Anonymized" Data Really Isn't* [27]	No	0.00001
AOL [3]	Free Text from Search Queries w/ Name, Location, etc	Yes	No	n=1	657,000	*A Face is Exposed* [3]	No	0.0000015
Netflix [4]	Movie Ratings & Dates	Yes	No	n=2	500,000	*"...successfully identified 99% of people in Netflix database"* [28]	No	0.000004
ONC Safe Harbor [5]	Zip3, YoB, Gender, Marital Status, Hispanic Ethnicity	No	N/A	n=2	15,000	[Press Did Not Cover This Study]	Yes	0.00013
Heritage Health Prize [6,7,8,9]	Age, Sex, Days in Hospital, Physician Specialty, Place of Service, CPT Code, Days Since First Claim, ICD-9 Diagnosis	Yes	No	n=0	113,000	*To best of my judgment, reidentification is within realm of possibility* [9] *El Emam estimated < 1% of Pts could be re-identified. Narayanan estimated > 12% of Pts were identifiable.* [29]	Yes	0.0
Y-Chromosome STR Surname Inference [10,11] - Simulation Study Part	Y-STR DNA Sequences* Age in Years & State	No	N/A, Simulation	Not Attempted: Simulated Results	~150 Million US Males	*"nice example of how simple it is to re-identify de-identified samples"* [30]	*No? (Safe Harbor vs. Expert Determination) *Safe Harbor Excludes: *Any unique identifying #, characteristic or code*	.12 (For Males Only), after accounting for 30% False Positive Rate
- CEU Attack Part	Age, Utah State, Genealogy Pedigrees & Mormon Ancestry	Yes, Highly Targeted	No	n=5 w/ Y-STR Alone, (but w/ Geneology Amplification n=50)	?	*DNA Hack Could Make Medical Privacy Impossible* [31]		Not Clearly Calculable for CEU Attack
Personal Genome Project [12,13,14]	Zip5, Gender, DoB	No	N/A	n=161	579	*"...re-identified names of > 40% anonymous participants"* [32] *re-identified 84 to 97% of sample of PGP volunteers* [33]	No	0.28 (w/ Embedded Names Excluded)
Washington St. Hospital Discharge [15,16]	Hospital Data w/ Diagnoses, Zip5, Month/Yr of Discharge	Yes	No	n=40 (8 verified) from 81 News Reports	648,384	*"...how new stories about hospital visits in Washington State leads to identifying matching health record 43% of the time"* [34]	No	0.000062
Cell Phone "Unicity" [17]	High Resolution Time (Hours) and Cell Tower Location	No	N/A	Not Attempted	1.5 Million	*"four spatio-temporal points enough to uniquely identify 95%"* [17]	No	0.0
NYC Taxi [18,19]	High Resolution Time (Minutes) and GPS Locations	Yes	No	n=11	173 Million Rides	*How Big Brother Watches You With Metadata* [35]	No	0.0000001
Credit Card "Unicity" [20,21,22,23,24,25,26]	High Resolution Time (Days), Location and Approx. Price	No	N/A	Not Attempted	1.1 Million	*With a Few Bits of Data, Researchers Identify 'Anonymous' People* [36]	No	0.0

- *Publicized attacks are on data without HIPAA/SDL de-identification protection.*
- *Many attacks targeted especially vulnerable subgroups and did not use sampling to assure representative results.*
- *Press reporting often portrays re-identification as broadly achievable, when there isn't any reliable evidence supporting this portrayal.*

13



14

"Y-STR Surname" Attack Headlines

15

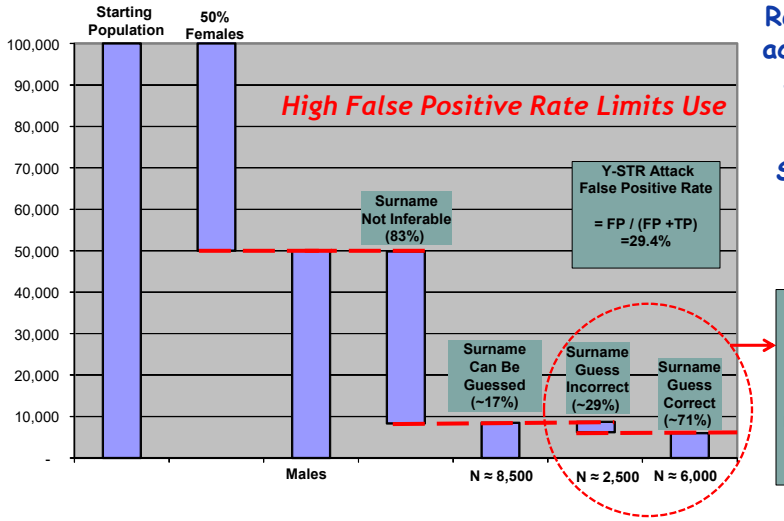
15

Question 1: Is Y-STR Attack Economically Viable?

Probably not -- unclear whether it eventually could be.

Question 2: Is "De-identification" pointless?

No, removing State, Grouping YoB would help importantly.



Re-ID isn't achieved by Surname Guess.

So what's the Threat Model?

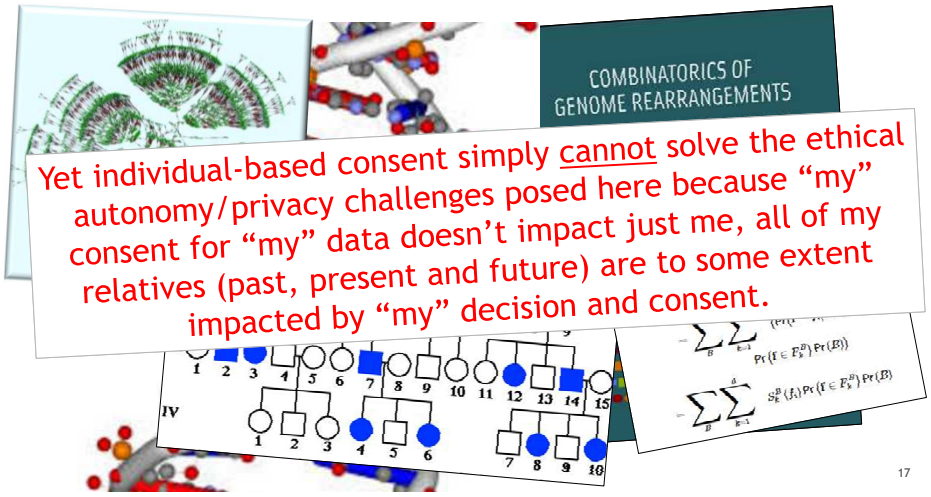
Surname Guess Could Serve as a (Faulty) Quasi-identifier (e.g., w/ YoB & State)

But Will Produce Substantive Re-identification Errors

16

16

Given the inherent extremely large combinatorics of genomic data nested within inheritance networks which determine how genomic traits (and surnames) are shared with our ancestors/descendants, the degree to which such information could be meaningfully "de-identified" are non-trivial.



17

17

Re-identification Demonstration Attack Summary

- For Ohm's famous "Broken Promises" attacks (Weld, AOL, Netflix) a total of n=4 people were re-identified out of 1.25 million.
- For attacks against HIPAA de-identified data (ONC, Heritage*), a total of n=2 people were re-identified out of 128 thousand.
  - ONC Attack Quasi-identifers: Zip3, YoB, Gender, Marital Status, Hispanic Ethnicity
  - Heritage Attack Quasi-identifiers*: Age, Sex, Days in Hospital, Physician Specialty, Place of Service, CPT Procedure Codes, Days Since First Claim, ICD-9 Diagnoses (*not complete list of data available for adversary attack)
  - Both were "adversarial" attacks.
- For all attacks listed, a total of n=268 were re-identified out of 327 million opportunities.

Let's get some perspective on this…

18

18

Obviously, This slide is **BLACK**

So clearly, De-identification Doesn't Work.

19

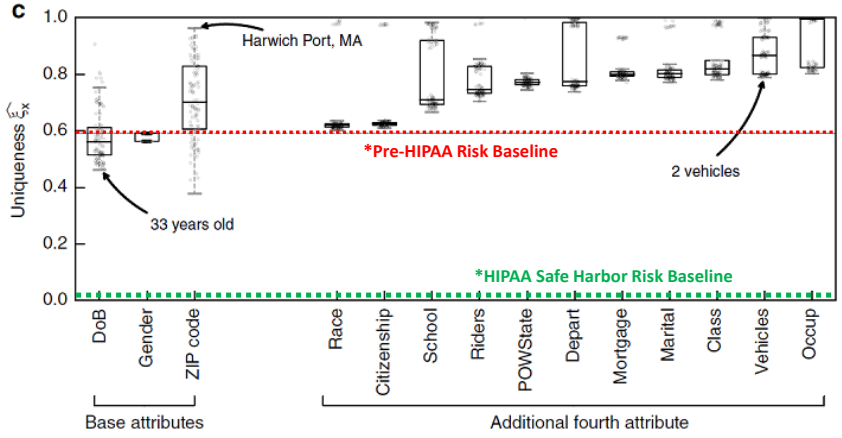
nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-10933-3 OPEN

Estimating the success of re-identifications in incomplete datasets using generative models

Luc Rocher[1,2,3], Julien M. Hendrickx[1] & Yves-Alexandre de Montjoye[2,3]



20

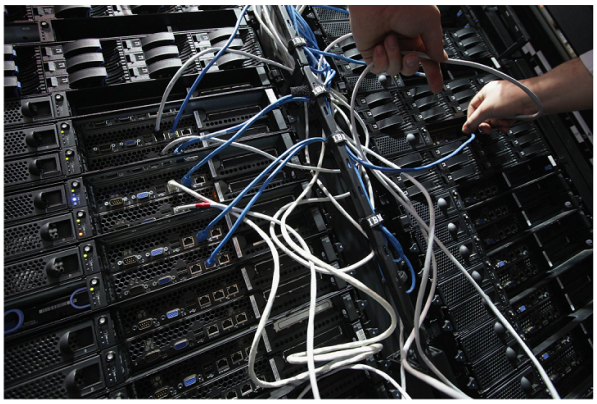
The New York Times

Your Data Were 'Anonymized'? These Scientists Can Still Identify You

By Gina Kolata

July 23, 2019

Computer scientists have developed an algorithm that can pick out almost any American in databases supposedly stripped of personal information.



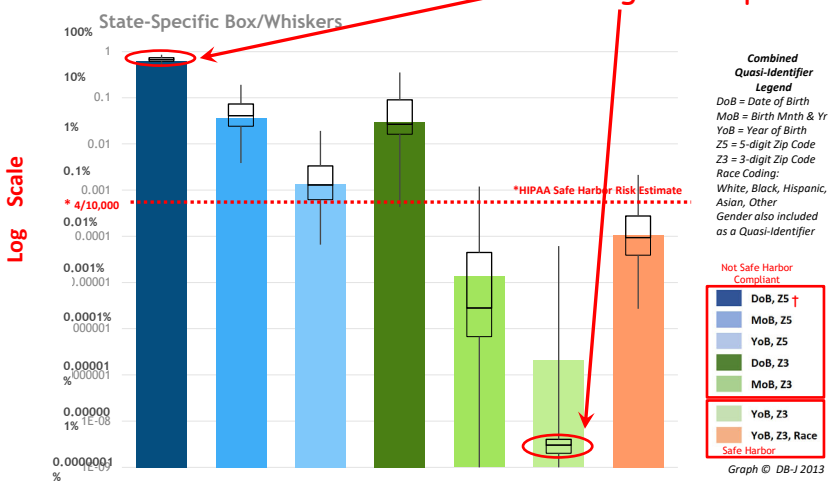
Scientists have found a way to identify virtually any American from any data set with just 15 attributes, like gender, ZIP code or marital status. Sean Gallup/Getty Images

21

21

Re-identification Risks: Population Uniqueness

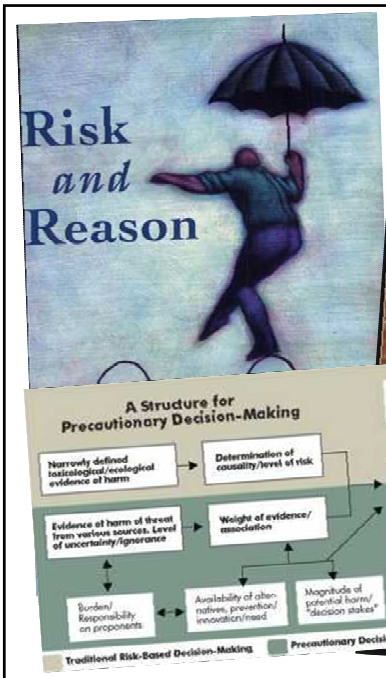
Starting Assumptions?

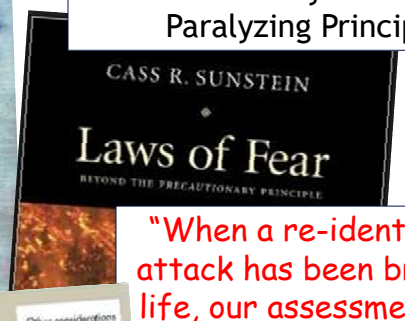


Data Source: 2010 U.S. Decennial Census

† *HIPAA Safe Harbor does not permit any Dates more specific than the year, or Geographic Units smaller than 3-digit Zip Codes (Z3).*

22





Precautionary Principle or Paralyzing Principle?

"When a re-identification attack has been brought to life, our assessment of the probability of it actually being implemented in the real-world may subconsciously become 100%, which is highly distortive of the true risk/benefit calculus that we face." – DB-J

23

Re-identification Demonstration Attack Summary

What can we conclude from the empirical evidence provided by these 11 highly influential re-identification attacks?

- The proportion of *demonstrated* re-identifications is extremely small.
- Which ***does not imply data re-identification risks are necessarily very small*** (*especially if the data has not been subject to Statistical Disclosure Limitation methods*).
- But with only 268 re-identifications made out of 327 million opportunities, Ohm's "Broken Promises" assertion that "*scientists have demonstrated they can often re-identify with astonishing ease*" seems rather dubious.
- It also seems clear that the state of "re-identification science", and the "evidence", it has provided needs to be dramatically improved in order to better support good public policy regarding data de-identification.

24

24

So, How Do We Move Beyond Anecdotes to a Rigorous, Scientific, Evidence-Based Risk Management Approach for Dealing with Re-identification Risks?

25

25

Re-identification Science Policy Short-comings:

6 ways in which "Re-identification Science" has (thus far) typically failed to best support sound public policies:

1. Attacking only trivially "straw man" de-identified data, where modern statistical disclosure control methods (like HIPAA) weren't used.
2. Targeting only especially vulnerable subpopulations and failing to use statistical random samples to provide policy-makers with representative re-identification risks for the entire population.
3. Making bad (often worst-case) assumptions and then failing to provide evidence to justify assumptions.

   Corollary: Not designing experiments to show the boundaries where de-identification finally succeeds.

26

26

Re-identification Science Policy Short-comings:

6 ways in which "Re-identification Science" has (thus far) typically failed to support sound public policies (Cont'd):

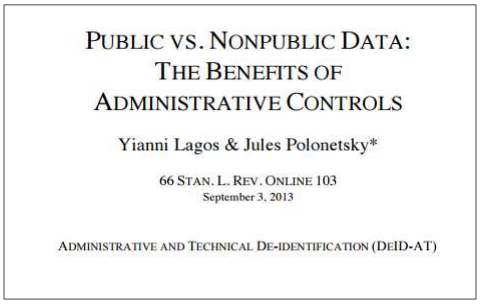
4. Failing to distinguish between sample uniqueness, population uniqueness and re-identifiability (i.e., the ability to correctly link population unique observations to identities).
5. Failing to fully specify relevant threat models (using data intrusion scenarios that account for all of the motivations, process steps, and information required to successfully complete the re-identification attack for the members of the population).
6. Unrealistic emphasis on absolute "Privacy Guarantees" and *failure to recognize unavoidable trade-offs between data privacy and statistical accuracy/utility.*

27

27

Supplementing Technical Data De-identification with Legal/Administrative Controls

However, in many cases, because of the possibility of highly-targeted demonstration attacks, arriving at solutions which will appropriately preserve the statistical accuracy and utility will also require that we supplement our statistical disclosure limitation "technical" data de-identification methods with additional legal and administrative controls.



PUBLIC VS. NONPUBLIC DATA:
THE BENEFITS OF
ADMINISTRATIVE CONTROLS

Yianni Lagos & Jules Polonetsky*

66 STAN. L. REV. ONLINE 103
September 3, 2013

ADMINISTRATIVE AND TECHNICAL DE-IDENTIFICATION (DEID-AT)

28

28

Recommended De-identified Data Use Requirements

Recipients of De-identified Data should be required to:

1) Not re-identify, or attempt to re-identify, or allow to be re-identified, any patients or individuals within the data, or their relatives, family or household members.
2) Not link any other data elements to the data without obtaining determination that the data remains de-identified.
3) Implement and maintain appropriate data security and privacy policies, procedures and associated physical, technical and administrative safeguards to assure that it is accessed only by authorized personnel and will remain de-identified.
4) Assure that all personnel or parties with access to the data agree to abide by all of the foregoing conditions

29

29

We also need…

Comprehensive, Multi-sector Legislative Prohibitions Against Data Re-identification

A BILL

To protect the privacy of potentially identifiable personal information by establishing accountability for the use and transfer of potentially identifiable personal information. [Version 4.4]

SECTION 1. SHORT TITLE.

This Act may be cited as the "Personal Data Deidentification Act".

SEC. 2. DEFINITIONS.

As used in this Act:

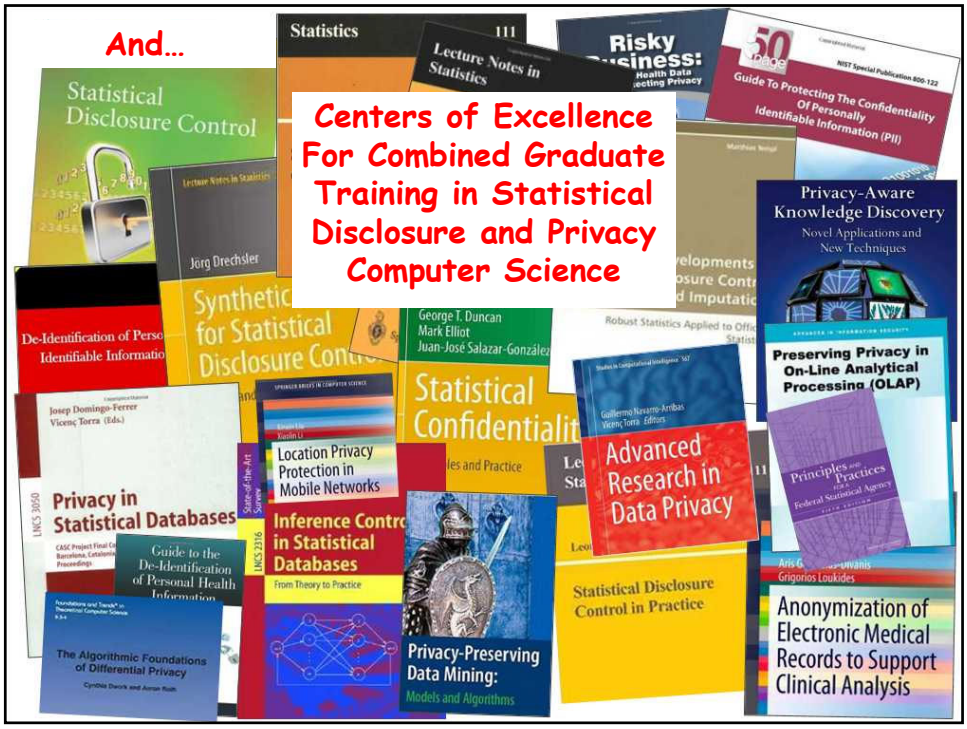
(1) DATA AGREEMENT.—The term "data agreement" means a contract, memorandum of understanding, data use agreement, or similar agreement between a discloser and a recipient relating to the use of personal information.

(2) DATA AGREEMENT SUBJECT TO THIS ACT.—The term "data

Robert Gellman, 2010
https://fpf.org/wp-content/uploads/2010/07/The_Deidentification_Dilemma.pdf

30

30



31

Stay Tuned… Much more on the way

www.nytimes.com/2018/06/27/science/dna-family-trees-cold-cases.html

Genealogists Turn to Cousins' DNA and Family Trees to Crack Five More Cold Cases

Police arrested a D.J. in Pennsylvania and a nurse in Washington State this week, the latest examples of the use of an open-source ancestry site since the break in the Golden State killer case.

By Heather Murphy

June 27, 2018

32

32

References for Re-identification Attack Summary Table

1. Sweeney, L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.
2. Barth-Jones, DC., The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now (July 2012). http://ssrn.com/abstract=2076397
3. Michael Barbaro, Tom Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749. New York Times August 6, 2006. www.nytimes.com/2006/08/09/technology/09aol.html
4. Narayanan, A., Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. Proceeding SP '08 Proceedings of the 2008 IEEE Symposium on Security and Privacy p. 111-125.
5. Kwok, P.K.; Lafky,D. Harder Than You Think: A Case Study of Re-Identification Risk of HIPAA Compliant Records. Joint Statistical Meetings. Section on Government Statistics. Miami, FL Aug 2, 2011. p. 3826-3833.
6. El Emam K, et al. De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset. J Med Internet Res 2012;14(1):e33
7. Valentino-DeVries, J. May the Best Algorithm Win… With $3 Million Prize, Health Insurer Raises Stakes on the Data-Crunching Circuit. Wall Street Journal. March 16, 2011. March 17, 2011 http://www.wsj.com/article_email/SB10001424052748704662604576202392747278936-lMyQjAxMTAxMDEwNTExNDUyWj.html
8. Narayanan, A. An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset. May 27, 2011 http://randomwalker.info/publications/heritage-health-re-identifiability.pdf
9. Narayanan, A. Felten, E.W. No silver bullet: De-identification still doesn't work. July 9, 2014 http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf
10. Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, Yaniv Erlich. Identifying Personal Genomes by Surname Inference. Science 18 Jan 2013: 321-324.
11. Barth-Jones, D. Public Policy Considerations for Recent Re-Identification Demonstration Attacks on Genomic Data Sets: Part 1. Harvard Law, Petrie-Flom Center: Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations. http://blogs.harvard.edu/billofhealth/2013/05/29/public-policy-considerations-for-recent-re-identification-demonstration-attacks-on-genomic-data-sets-part-1-re-identification-symposium/
12. Sweeney, L., Abu, A, Winn, J. Identifying Participants in the Personal Genome Project by Name (April 29, 2013). http://ssrn.com/abstract=2257732

33

References for Re-identification Attack Summary Table

13. Jane Yakowitz. Reporting Fail: The Reidentification of Personal Genome Project Participants May 1, 2013. https://blogs.harvard.edu/infolaw/2013/05/01/reporting-fail-the-reidentification-of-personal-genome-project-participants/
14. Barth-Jones, D. Press and Reporting Considerations for Recent Re-Identification Demonstration Attacks: Part 2. Harvard Law, Petrie-Flom Center: Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations. http://blogs.harvard.edu/billofhealth/2013/10/01/press-and-reporting-considerations-for-recent-re-identification-demonstration-attacks-part-2-re-identification-symposium/
15. Sweeney, L. Matching Known Patients to Health Records in Washington State Data (June 5, 2013). http://ssrn.com/abstract=2289850
16. Robertson, J. States' Hospital Data for Sale Puts Privacy in Jeopardy. Bloomberg News June 5, 2013. https://www.bloomberg.com/news/articles/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy
17. Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, Vincent D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports 3, Article number: 1376 (2013) http://www.nature.com/articles/srep01376
18. Anthony Tockar. Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset. September 15, 2014. https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/
19. Barth-Jones, D. The Antidote for "Anecdata": A Little Science Can Separate Data Privacy Facts from Folklore. https://blogs.harvard.edu/infolaw/2014/11/21/the-antidote-for-anecdata-a-little-science-can-separate-data-privacy-facts-from-folklore/
20. de Montjoye, et al. . Unique in the shopping mall: On the reidentifiability of credit card metadata. Science. 30 Jan 2015: Vol. 347, Issue 6221, pp. 536-539.
21. Barth-Jones D, El Emam K, Bambauer J, Cavoukian A, Malin B. Assessing data intrusion threats. Science. 2015 Apr 10; 348(6231):194-5.
22. de Montjoye, et al. Assessing data intrusion threats—Response Science. 10 Apr 2015: Vol. 348, Issue 6231, pp. 195
23. Jane Yakowitz Bambauer. Is De-Identification Dead Again? April 28, 2015. https://blogs.harvard.edu/infolaw/2015/04/28/is-de-identification-dead-again/
24. David Sánchez, Sergio Martínez, Josep Domingo-Ferrer. Technical Comments: Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata". Science. 18 Mar 2016: Vol. 351, Issue 6279, pp. 1274.
25. Sánchez, et al. Supplementary Materials for "How to Avoid Reidentification with Proper Anonymization"- Comment on "Unique in the shopping mall: on the reidentifiability of credit card metadata". http://arxiv.org/abs/1511.05957
26. de Montjoye, et al. Response to Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata" Science 18 Mar 2016: Vol. 351, Issue 6279, pp. 1274

34

References for Re-identification Attack Summary Table

27. Nate Anderson. "Anonymized" data really isn't—and here's why not. Sep 8, 2009 http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/
28. Sorrell v. IMS Health: Brief of Amici Curiae Electronic Privacy Information Center. March 1, 2011. https://epic.org/amicus/sorrell/EPIC_amicus_Sorrell_final.pdf
29. Ruth Williams. Anonymity Under Threat: Scientists uncover the identities of anonymous DNA donors using freely available web searches. The Scientist. January 17, 2013. http://www.the-scientist.com/?articles.view/articleNo/34006/title/Anonymity-Under-Threat/
30. Kevin Fogarty. DNA hack could make medical privacy impossible. CSO. March 11, 2013. http://www.csoonline.com/article/2133054/identity-access/dna-hack-could-make-medical-privacy-impossible.html
31. Adam Tanner. Harvard Professor Re-Identifies Anonymous Volunteers in DNA Study. Forbes. Apr 25, 2013. http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/
32. Adam Tanner. The Promise & Perils of Sharing DNA. Undark Magazine. September 13, 2016. http://undark.org/article/dna-ancestry-sharing-privacy-23andme/
33. Sweeney L. Only You, Your Doctor, and Many Others May Know. Technology Science. 2015092903. September 29, 2015. http://techscience.org/a/2015092903
34. David Sirota. How Big Brother Watches You With Metadata. San Francisco Gate. October 9, 2014. http://www.sfgate.com/opinion/article/How-Big-Brother-watches-you-with-metadata-5812775.php
35. Natasha Singer. With a Few Bits of Data, Researchers Identify 'Anonymous' People. New York Times. Bits Blog. January 29, 2015. http://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people/

Additional Re-identification Attack Review References

1. Khaled El Emam, Jonker, E.; Arbuckle, L.; Malin, B. A systematic review of re-identification attacks on health data. PLoS One 2011; Vol 6(12):e28071.
2. Jane Henriksen-Bulmer, Sheridan Jeary. Re-identification attacks - A systematic literature review. International Journal of Information Management, 36 (2016) 1184-1192.

35

Bill of Health

Examining the intersection of law and health care, biotech & bioethics
A blog by the Petrie-Flom Center and friends

Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations

- http://blogs.law.harvard.edu/billofhealth/2013/05/29/public-policy-considerations-for-recent-re-identification-demonstration-attacks-on-genomic-data-sets-part-1-re-identification-symposium/
- https://blogs.law.harvard.edu/billofhealth/2013/10/01/press-and-reporting-considerations-for-recent-re-identification-demonstration-attacks-part-2-re-identification-symposium/
- http://blogs.law.harvard.edu/billofhealth/2013/10/02/ethical-concerns-conduct-and-public-policy-for-re-identification-and-de-identification-practice-part-3-re-identification-symposium/

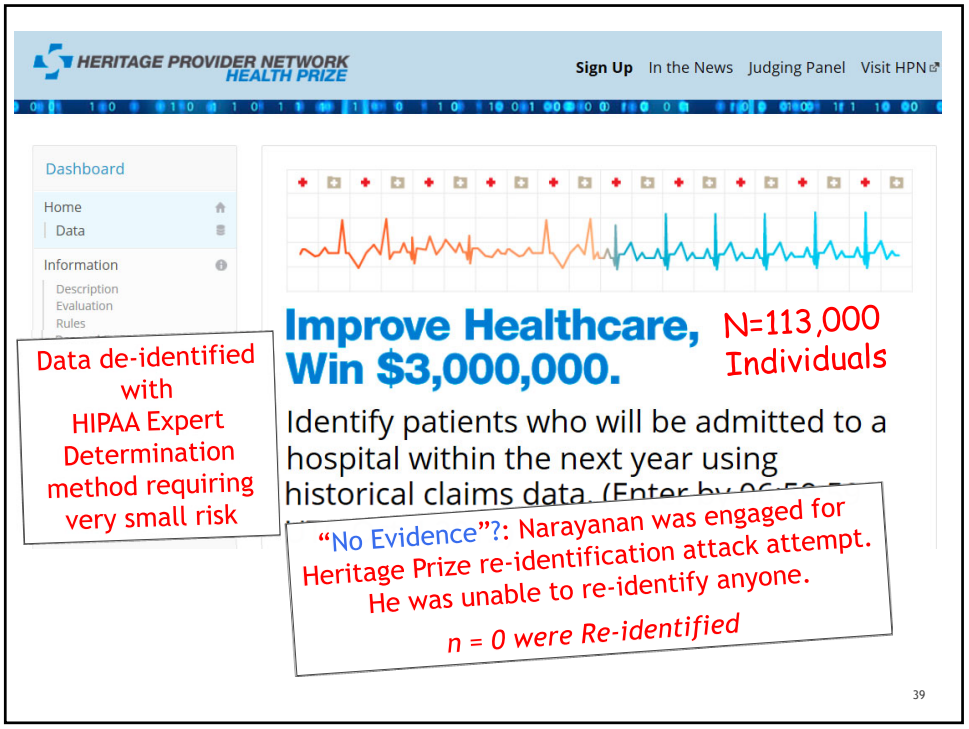
36


Bloomberg
Our Company | Professional | Anywhere
HOME QUICK NEWS OPINION MARKET DATA PERSONAL FINANCE TECH POLITICS
Frustrated Republicans Pressure Boehner to End Shutdown
Shutdown Jokes, Day 3: Letterman, Colbert, Stewart
BREAKING NEWS Telecom Italia Ceo Bernabe Is Said to Resign
States' Hospital Data for Sale Puts Privacy in Jeopardy
By Jordan Robertson - Jun 5, 2013 12:01 AM ET
113 COMMENTS
QUEUE
STATES VULNERABLE OF PATIENT DATA COMPROMISE
WASHINGTON
NEW YORK
ARIZONA
NEW JERSEY
Bloomberg
Your Health Data for Sale: Who's Selling, Buying?
Consider Ray Boylston, who went into diabetic shock while riding his motorcycle in rural Washington in 2011. He careened off the road and was thrown into the woods, an accident that was covered only briefly, in the local newspaper. Boylston disclosed his medical condition and history to a handful of loved ones and the hospital that treated him.
After Boylston's discharge, Washington collected the paperwork of his week-long stay from Providence Sacred Heart Medical Center in Spokane and added it to a database of 650,000 hospitalizations for 2011 available for sale to researchers, companies and other members of the public. The data was supposed to remain anonymous. Yet because of state exemption from federal regulations governing discharge information, Boylston could be identified and his medical background exposed using only publicly available information.
"I don't really feel that the public has a right to read up on my medical history," said Boylston, who is 62 and a veteran. "I feel I've been violated."
WA State Hospital Discharge Attack
40/648,384
= 1/16,200
37

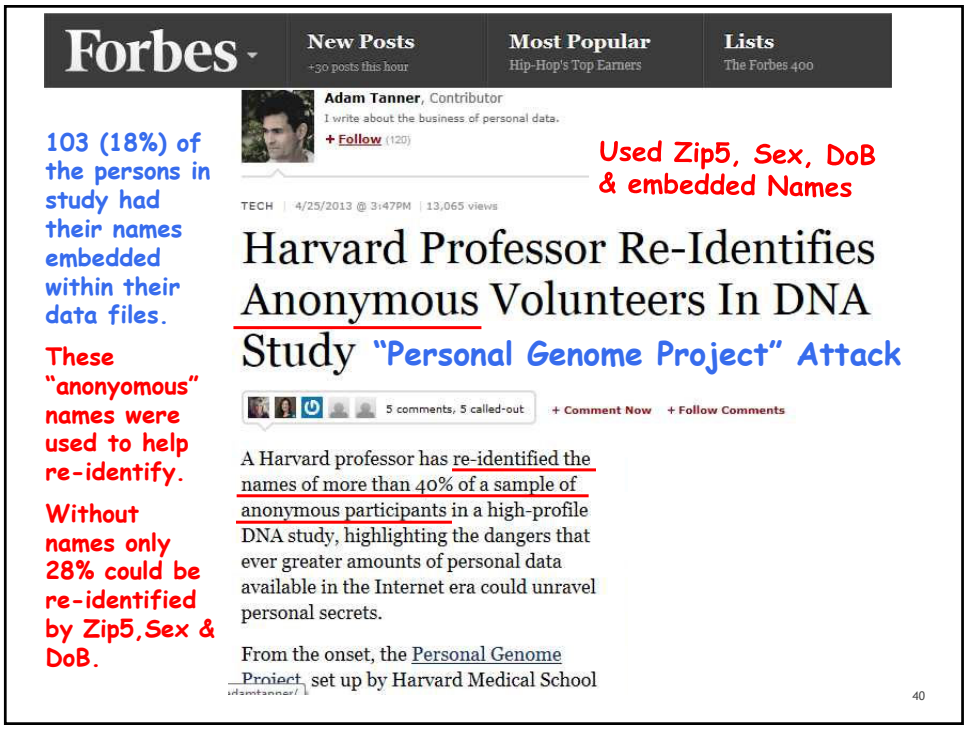
37


THE SPOKESMAN-REVIEW
Topics Times Places Media
October 23, 2011 in City
Man, 61, thrown from motorcycle
A 61-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle.
Raymond E. Boylston was riding his 2003 Harley-Davidson north on Highway 25, about 16 miles north of Davenport, when he failed to negotiate a curve to the left, the Washington State Patrol said in a news release. His motorcycle left the road, becoming airborne before it landed in a wooded area. Boylston was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident, the WSP said.
He was taken to Lincoln Hospital, where his condition was unavailable Saturday night.
1 Washington state news articles were searched for the word "hospitalized." Most of these articles included the person's name, age, town of residence and reason for hospitalization.
2 The person's name, age and residence is searched online. Several online sites will reveal ZIP codes associated with the search terms.
3 Taking the newly learned ZIP code, plus the patient's age[1], approximate date and location of hospitalization[2], a match can be found in the health record dataset purchased from the state.
4 Within the electronic record is private patient information including: physician diagnoses, procedures and payment information.
white pages
Find People Find a Business Reverse Phone Address & Neighbors
Raymond Boylston washington
1 Result for Raymond Boylston in WA
Raymond E Boylston
Soap Lake, WA 98851-1435
MEDICAL RECORD
RECORD: 502855338
admitend: 10/25/11
STAYTYPE: 1
HOSPITAL: 162
COUNTYRES: 13
AGE_MONTH: 725
ZIPCODE 98851
Charges: $71708.47
DescDIAG1
80843: closed fracture of other specified part of pelvis; pelv fx-clos/pelv disrup
DescDIAG2
5185: pulmonary insufficiency following trauma & surgery; post traum pulm insuffic
Raymond E. Boylston, the man identified in the news article, is linked to "anonymous record" #502855338.
How Someone Can Re-identify Your Medical Records
38

38


HERITAGE PROVIDER NETWORK HEALTH PRIZE
Sign Up
In the News
Judging Panel
Visit HPN
Dashboard
Home
Data
Information
Description
Evaluation
Rules
Improve Healthcare, Win $3,000,000.
N=113,000 Individuals
Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by
Data de-identified with HIPAA Expert Determination method requiring very small risk
"No Evidence"?: Narayanan was engaged for Heritage Prize re-identification attack attempt. He was unable to re-identify anyone.
n = 0 were Re-identified
39

39


Forbes
New Posts
+30 posts this hour
Most Popular
Hip-Hop's Top Earners
Lists
The Forbes 400
Adam Tanner, Contributor
I write about the business of personal data.
+ Follow (120)
TECH | 4/25/2013 @ 3:47PM | 13,065 views
Used Zip5, Sex, DoB & embedded Names
Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study
"Personal Genome Project" Attack
5 comments, 5 called-out
+ Comment Now
+ Follow Comments
A Harvard professor has re-identified the names of more than 40% of a sample of anonymous participants in a high-profile DNA study, highlighting the dangers that ever greater amounts of personal data available in the Internet era could unravel personal secrets.
From the onset, the Personal Genome Project set up by Harvard Medical School
103 (18%) of the persons in study had their names embedded within their data files.
These "anonyomous" names were used to help re-identify.
Without names only 28% could be re-identified by Zip5,Sex & DoB.
40

40