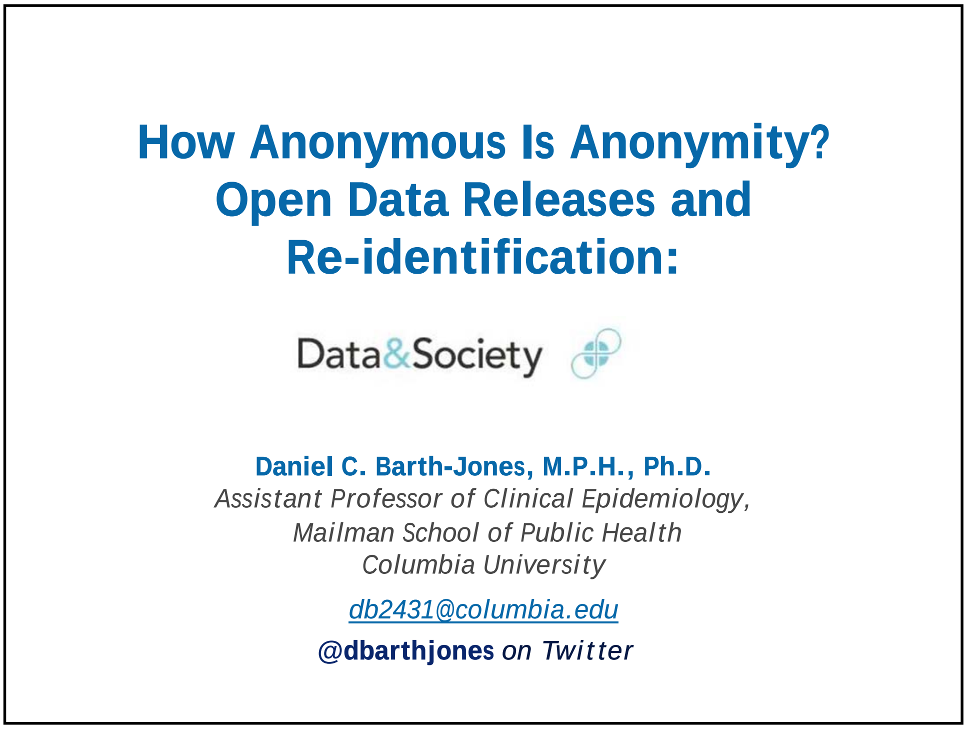


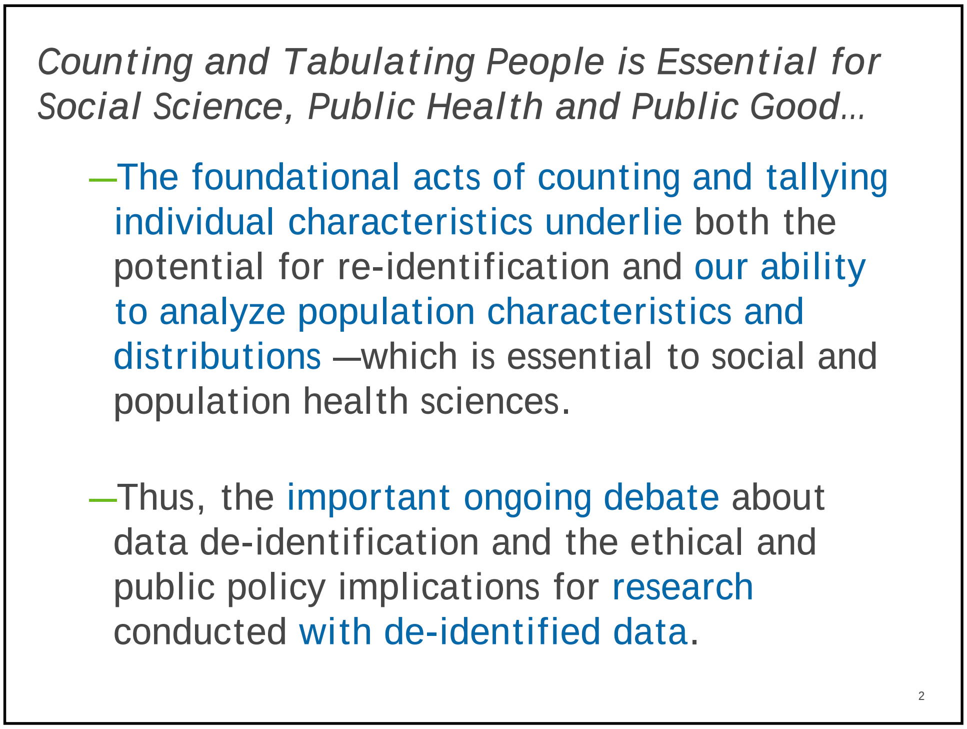
How Anonymous Is Anonymity? Open Data Releases and Re-identification:

Data&Society

Daniel C. Barth-Jones, M.P.H., Ph.D.
Assistant Professor of Clinical Epidemiology,
Mailman School of Public Health
Columbia University

db2431@columbia.edu

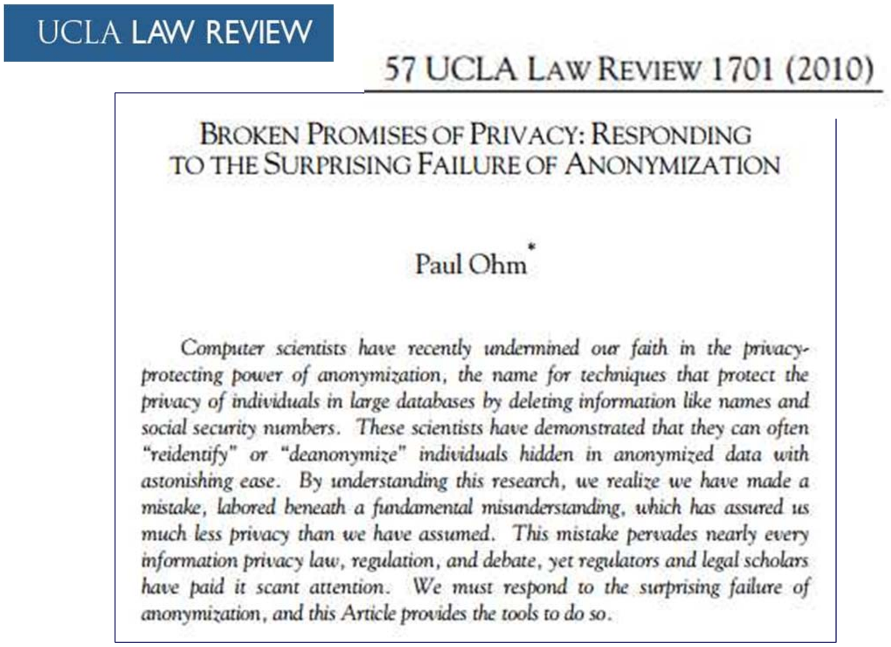
@dbarthjones *on Twitter*



Counting and Tabulating People is Essential for Social Science, Public Health and Public Good…

- —The foundational acts of counting and tallying individual characteristics underlie both the potential for re-identification and our ability to analyze population characteristics and distributions —which is essential to social and population health sciences.

- —Thus, the important ongoing debate about data de-identification and the ethical and public policy implications for research conducted with de-identified data.



UCLA LAW REVIEW

57 UCLA LAW REVIEW 1701 (2010)

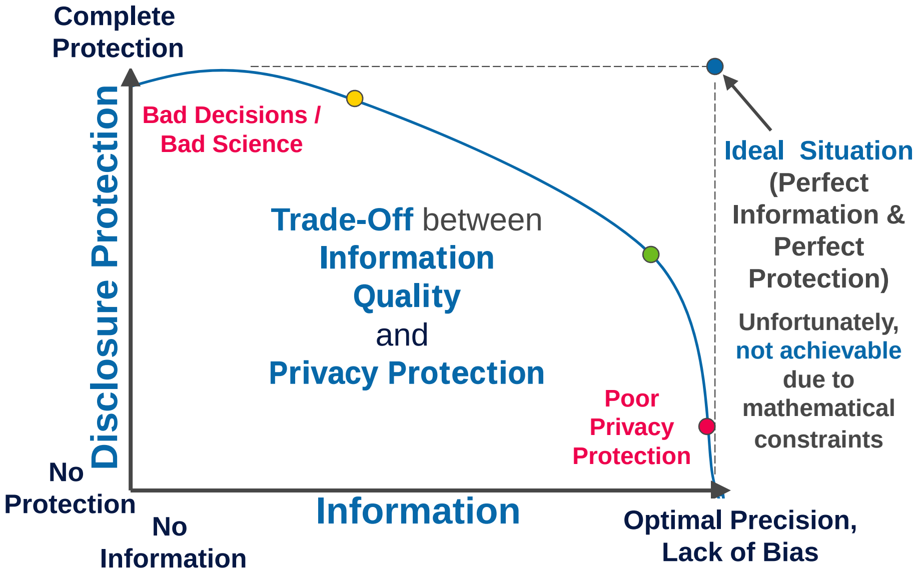
BROKEN PROMISES OF PRIVACY: RESPONDING TO THE SURPRISING FAILURE OF ANONYMIZATION

Paul Ohm*

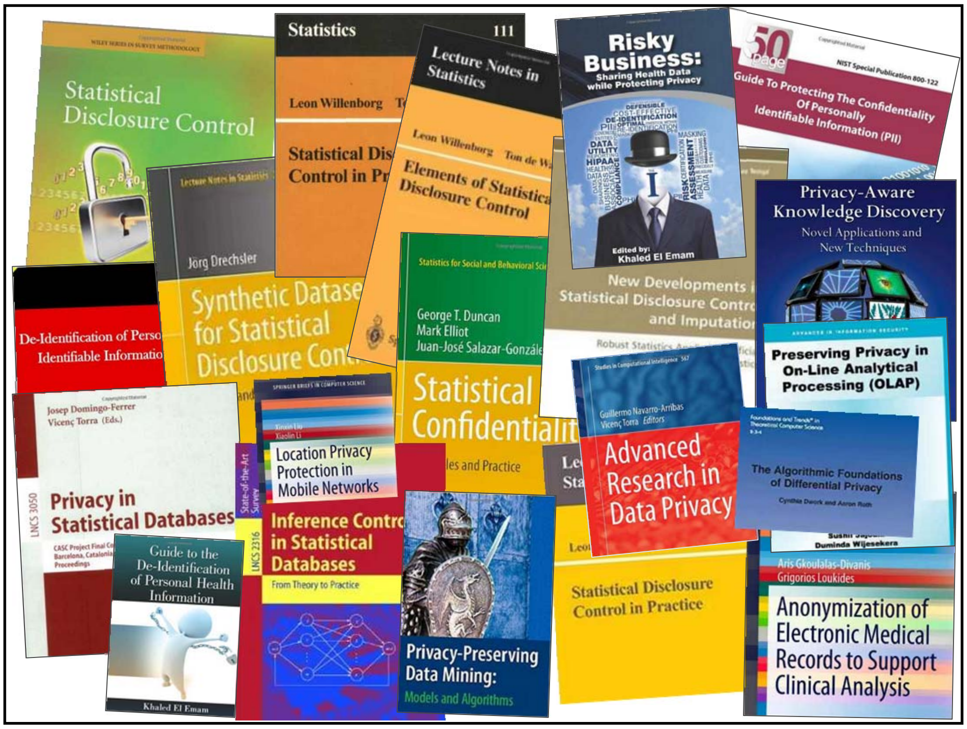
Computer scientists have recently undermined our faith in the privacy-protecting power of anonymization, the name for techniques that protect the privacy of individuals in large databases by deleting information like names and social security numbers. These scientists have demonstrated that they can 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.

3

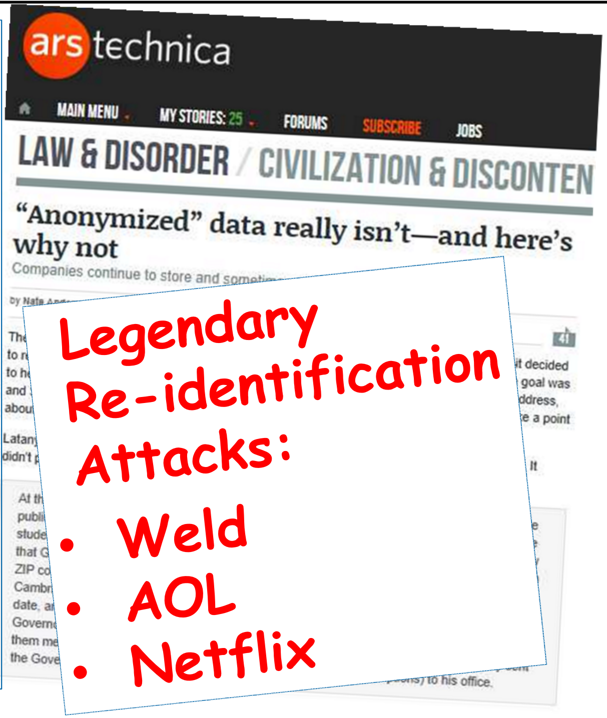
The Inconvenient Truth:


Complete Protection
Disclosure Protection
No Protection
Bad Decisions / Bad Science
Trade-Off between Information Quality and Privacy Protection
Poor Privacy Protection
Ideal Situation (Perfect Information & Perfect Protection)
Unfortunately, not achievable due to mathematical constraints
No Information
Information
Optimal Precision, Lack of Bias

4


Statistical Disclosure Control
Statistics 111
Leon Willenborg
Statistical Disclosure Control in Practice
Lecture Notes in Statistics
Elements of Statistical Disclosure Control
Risky Business: Sharing Health Data while Protecting Privacy
Edited by: Khaled El Emam
Guide To Protecting The Confidentiality Of Personally Identifiable Information (PII)
Privacy-Aware Knowledge Discovery
Novel Applications and New Techniques
De-Identification of Personally Identifiable Information
Jörg Drechsler
Synthetic Datasets for Statistical Disclosure Control
George T. Duncan
Mark Elliot
Juan-José Salazar-González
Statistical Confidentiality
New Developments in Statistical Disclosure Control and Imputation
Preserving Privacy in On-Line Analytical Processing (OLAP)
Josep Domingo-Ferrer
Vicenç Torra (Eds.)
Privacy in Statistical Databases
Location Privacy Protection in Mobile Networks
Guillermo Navarro-Arribas
Vicenç Torra Editors
Advanced Research in Data Privacy
The Algorithmic Foundations of Differential Privacy
Cynthia Dwork and Aaron Roth
Inference Control in Statistical Databases
From Theory to Practice
Guide to the De-Identification of Personal Health Information
Khaled El Emam
Privacy-Preserving Data Mining: Models and Algorithms
Aris Gkoulalas-Divanis
Grigorios Loukides
Anonymization of Electronic Medical Records to Support Clinical Analysis

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



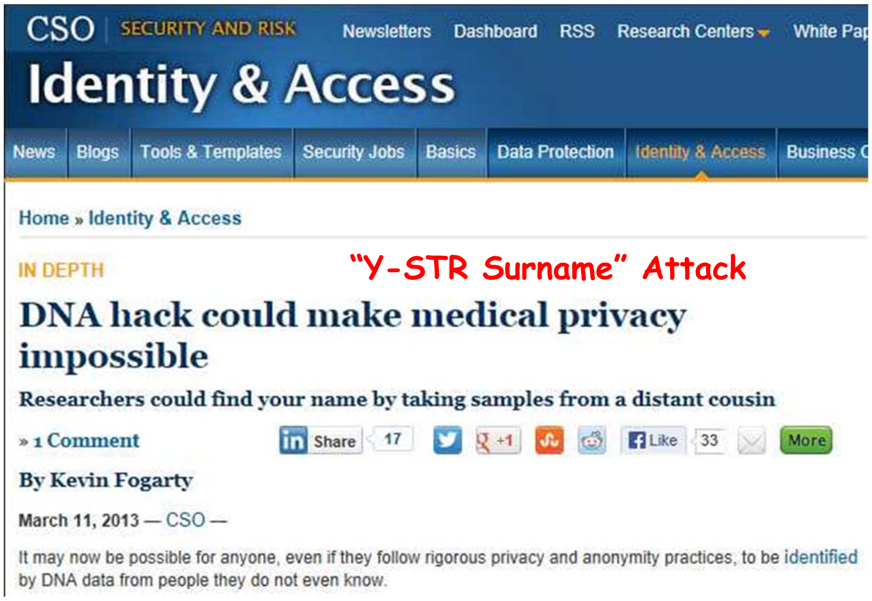
“Anonymized” data really isn’t—and here’s why not

Companies continue to store and someti

Legendary Re-identification Attacks:

- Weld
- AOL
- Netflix

6



CSO | SECURITY AND RISK

Newsletters Dashboard RSS Research Centers White Pap

Identity & Access

News | Blogs | Tools & Templates | Security Jobs | Basics | Data Protection | Identity & Access | Business C

Home » Identity & Access

IN DEPTH

“Y-STR Surname” Attack

DNA hack could make medical privacy impossible

Researchers could find your name by taking samples from a distant cousin

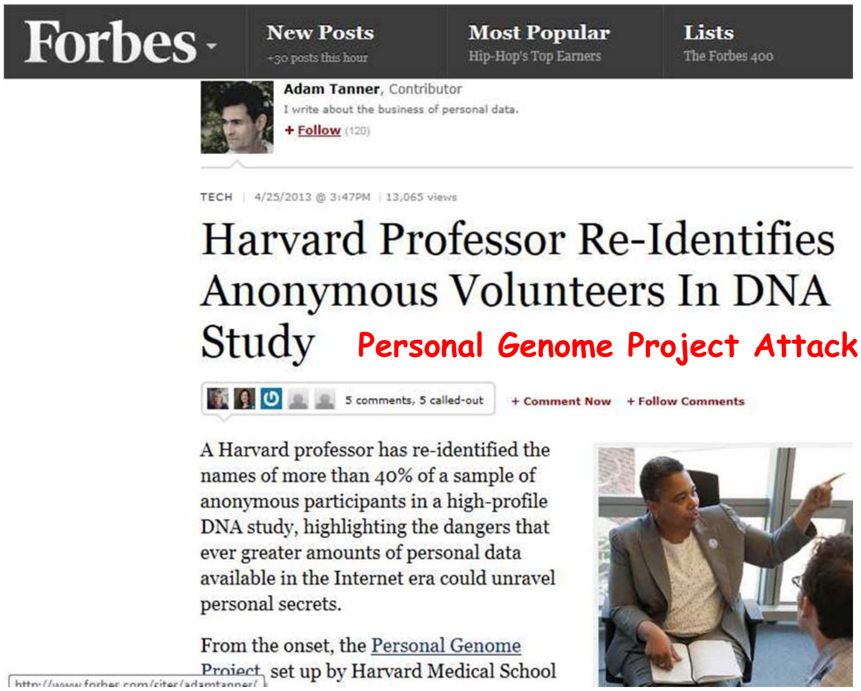
» 1 Comment

Share 17 +1 Like 33 More

By Kevin Fogarty

March 11, 2013 — CSO —

It may now be possible for anyone, even if they follow rigorous privacy and anonymity practices, to be identified by DNA data from people they do not even know.



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

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study

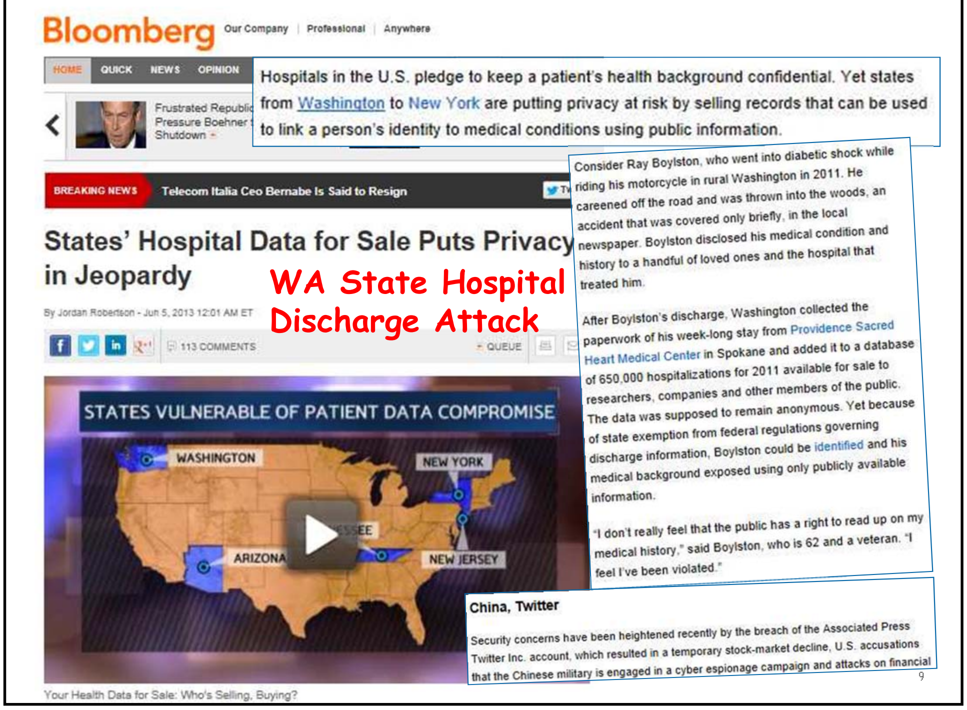
Personal Genome Project Attack

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

http://www.forbes.com/sites/adamtanner/



Bloomberg Our Company | Professional | Anywhere

HOME QUICK NEWS OPINION

Frustrated Republic Pressure Boehner Shutdown

BREAKING NEWS Telecom Italia Ceo Bernabe Is Said to Resign

Hospitals in the U.S. pledge to keep a patient's health background confidential. Yet states from Washington to New York are putting privacy at risk by selling records that can be used to link a person's identity to medical conditions using public information.

States' Hospital Data for Sale Puts Privacy in Jeopardy

By Jordan Robertson - Jun 5, 2013 12:01 AM ET

113 COMMENTS QUEUE

WA State Hospital Discharge Attack

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."

China, Twitter

Security concerns have been heightened recently by the breach of the Associated Press Twitter Inc. account, which resulted in a temporary stock-market decline, U.S. accusations that the Chinese military is engaged in a cyber espionage campaign and attacks on financial

Your Health Data for Sale: Who's Selling, Buying?

HARVARD LAW PETRIE-FLOM CENTER

Bill of Health

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

VE RI TAS

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/

SCIENTIFIC REPORTS

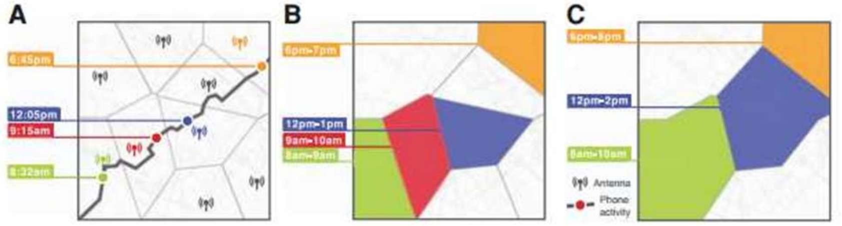
Published
25 March 2013

Unique in the Crowd: The privacy bounds of human mobility

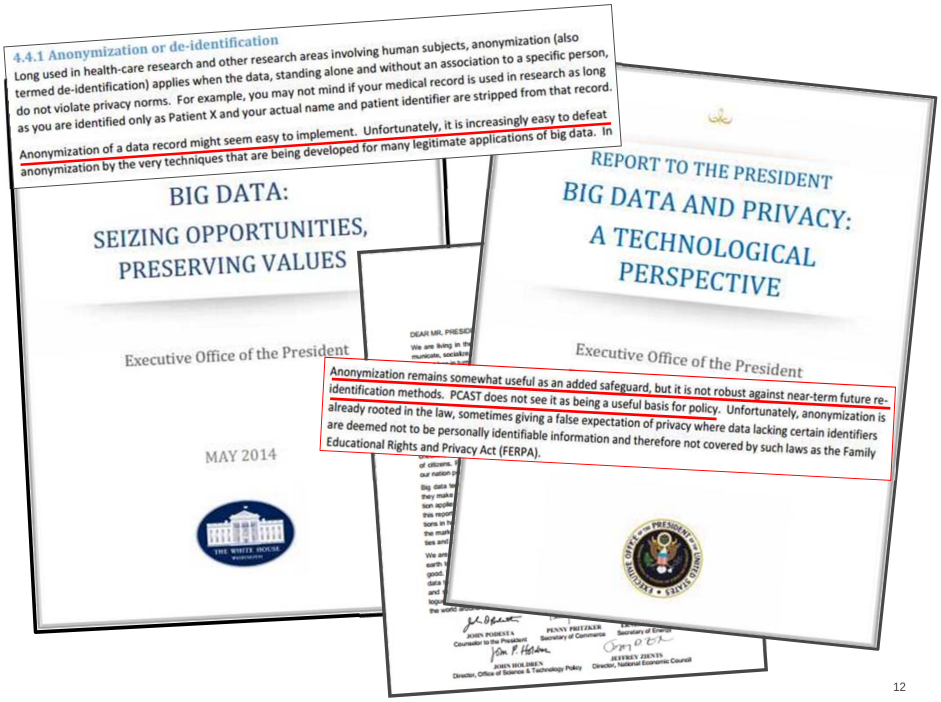
Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

Cell Data Uniqueness

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a formula for the uniqueness of human mobility traces given their resolution and the available outside information. This formula shows that the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity. These findings represent fundamental constraints to an individual's privacy and have important implications for the design of frameworks and institutions dedicated to protect the privacy of individuals.



Sample Unique ≠ Re-identifiable



4.4.1 Anonymization or de-identification
Long used in health-care research and other research areas involving human subjects, anonymization (also
termed de-identification) applies when the data, standing alone and without an association to a specific person,
do not violate privacy norms. For example, you may not mind if your medical record is used in research as long
as you are identified only as Patient X and your actual name and patient identifier are stripped from that record.
Anonymization of a data record might seem easy to implement. Unfortunately, it is increasingly easy to defeat
anonymization by the very techniques that are being developed for many legitimate applications of big data. In

BIG DATA:
SEIZING OPPORTUNITIES,
PRESERVING VALUES

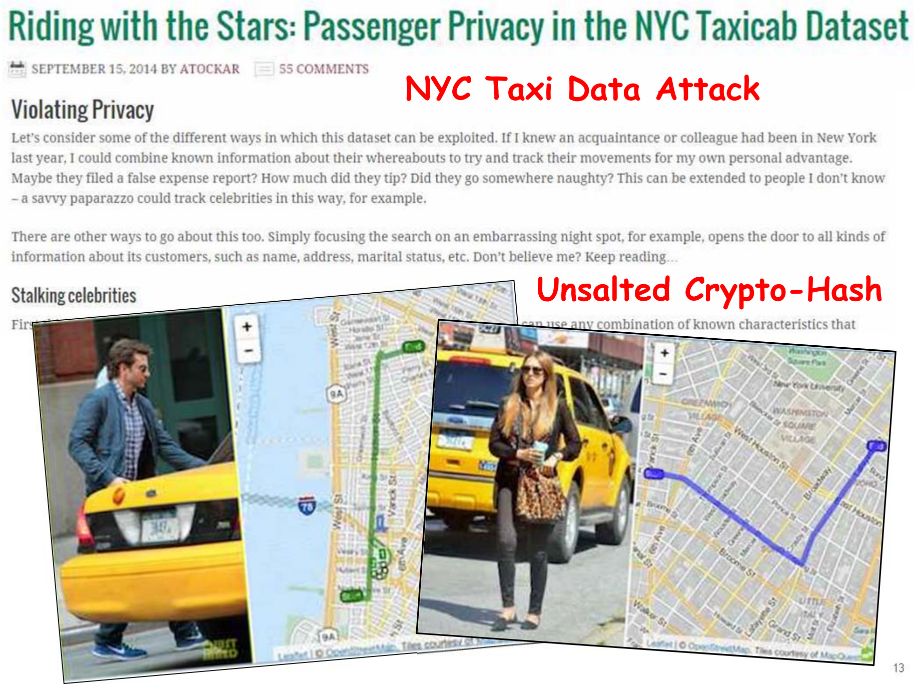
Executive Office of the President

MAY 2014

REPORT TO THE PRESIDENT
BIG DATA AND PRIVACY:
A TECHNOLOGICAL
PERSPECTIVE

Executive Office of the President

Anonymization remains somewhat useful as an added safeguard, but it is not robust against near-term future re-
identification methods. PCAST does not see it as being a useful basis for policy. Unfortunately, anonymization is
already rooted in the law, sometimes giving a false expectation of privacy where data lacking certain identifiers
are deemed not to be personally identifiable information and therefore not covered by such laws as the Family
Educational Rights and Privacy Act (FERPA).



Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset

SEPTEMBER 15, 2014 BY ATOCKAR 55 COMMENTS

NYC Taxi Data Attack

Violating Privacy

Let's consider some of the different ways in which this dataset can be exploited. If I knew an acquaintance or colleague had been in New York last year, I could combine known information about their whereabouts to try and track their movements for my own personal advantage. Maybe they filed a false expense report? How much did they tip? Did they go somewhere naughty? This can be extended to people I don't know – a savvy paparazzo could track celebrities in this way, for example.

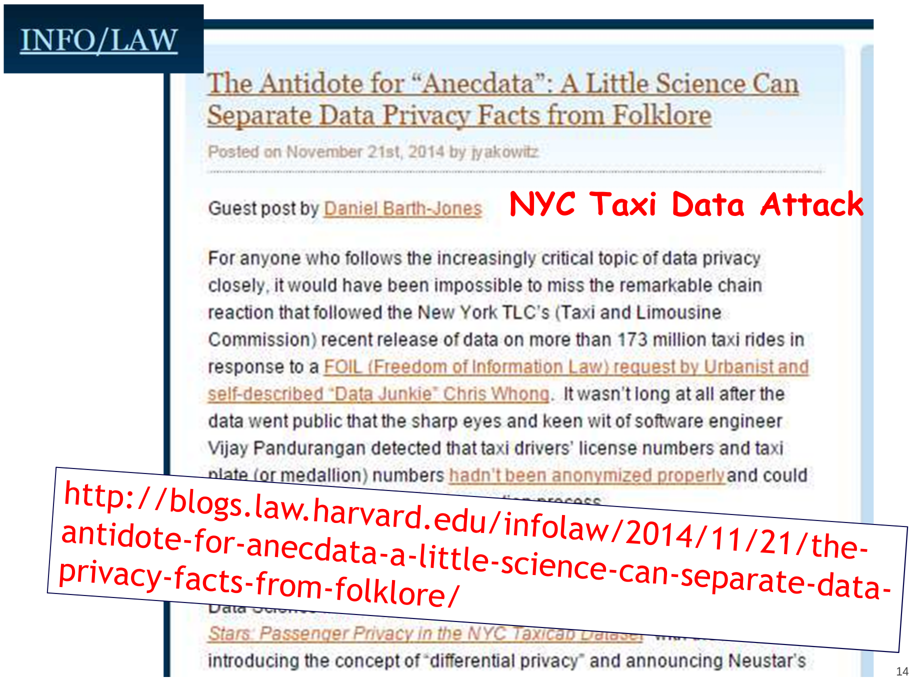
There are other ways to go about this too. Simply focusing the search on an embarrassing night spot, for example, opens the door to all kinds of information about its customers, such as name, address, marital status, etc. Don't believe me? Keep reading...

Stalking celebrities

Unsalted Crypto-Hash

Fir

can use any combination of known characteristics that



INFO/LAW

The Antidote for "Anecdata": A Little Science Can Separate Data Privacy Facts from Folklore

Posted on November 21st, 2014 by jyakowitz

Guest post by Daniel Barth-Jones

NYC Taxi Data Attack

For anyone who follows the increasingly critical topic of data privacy closely, it would have been impossible to miss the remarkable chain reaction that followed the New York TLC's (Taxi and Limousine Commission) recent release of data on more than 173 million taxi rides in response to a FOIL (Freedom of Information Law) request by Urbanist and self-described "Data Junkie" Chris Whong. It wasn't long at all after the data went public that the sharp eyes and keen wit of software engineer Vijay Pandurangan detected that taxi drivers' license numbers and taxi plate (or medallion) numbers hadn't been anonymized properly and could

http://blogs.law.harvard.edu/infolaw/2014/11/21/the-antidote-for-anecdata-a-little-science-can-separate-data-privacy-facts-from-folklore/

Stars: Passenger Privacy in the NYC Taxicab Dataset

introducing the concept of "differential privacy" and announcing Neustar's

Harvard Business Review

REGULATION

There’s No Such Thing as Anonymous Data

January 2015

DATA PRIVACY DAY

Gauging the allure of designer drugs p. 409
Blown-up brains for a better inside view pp. 474 & 543
Single-crystal perovskite solar cells pp. 519 & 522
Science
$10
30 JANUARY 2015
sciencemag.org
AAAS
SPECIAL ISSUE
The End of
PRIVACY

ENT TEXT SIZE PRINT

About a decade ago, a hacker said to me, flatly, “Assume every card in your wallet is compromised, and

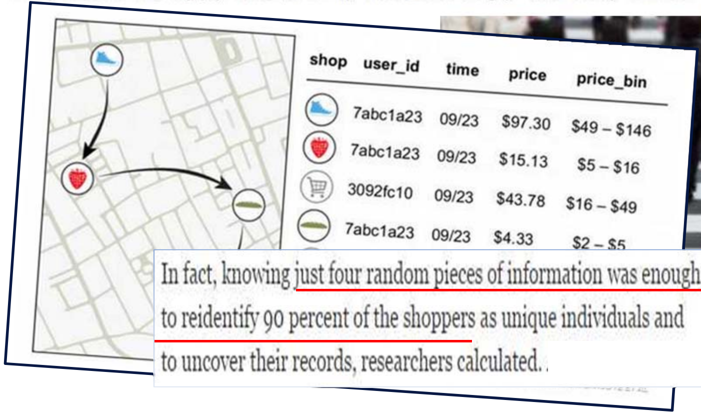
For scientists, the vast amounts of data that people shed every day offer great new opportunities but new dilemmas as well. New computational techniques can identify people or trace their behavior by combining just a few snippets of data. There are ways to protect the private information hidden in big data files, but they limit what scientists can learn; a balance must be struck. Some medical researchers acknowledge that keeping patient data private is becoming almost impossible;

IDENTITY AND PRIVACY

Credit Card
Data Uniqueness

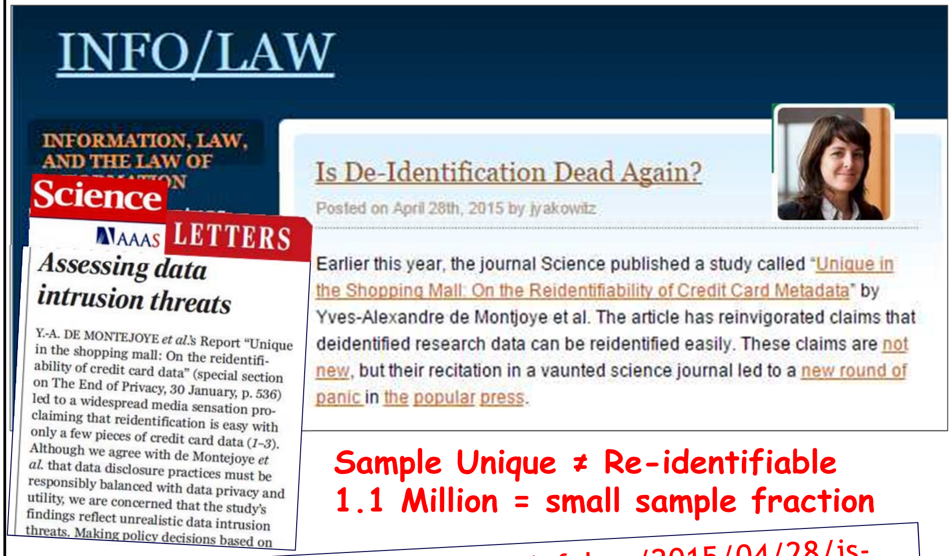
Unique in the shopping mall: On the reidentifiability of credit card metadata

Yves-Alexandre de Montjoye,[1*] Laura Radaelli,[2] Vivek Kumar Singh,[1,3] Alex “Sandy” Pentland[1]



shop	user_id	time	price	price_bin
👟	7abc1a23	09/23	$97.30	$49 – $146
🍓	7abc1a23	09/23	$15.13	$5 – $16
🛒	3092fc10	09/23	$43.78	$16 – $49
🥒	7abc1a23	09/23	$4.33	$2 – $5

In fact, knowing just four random pieces of information was enough to reidentify 90 percent of the shoppers as unique individuals and to uncover their records, researchers calculated. .



INFO/LAW

INFORMATION, LAW, AND THE LAW OF INFORMATION

Science

AAAS LETTERS

Assessing data intrusion threats

Y.-A. DE MONTEJOYE *et al.*'s Report "Unique in the shopping mall: On the reidentifiability of credit card data" (special section on The End of Privacy, 30 January, p. 536) led to a widespread media sensation proclaiming that reidentification is easy with only a few pieces of credit card data (*1*–*3*). Although we agree with de Montejoye *et al.* that data disclosure practices must be responsibly balanced with data privacy and utility, we are concerned that the study's findings reflect unrealistic data intrusion threats. Making policy decisions based on

Is De-Identification Dead Again?

Posted on April 28th, 2015 by jyakowitz

Earlier this year, the journal Science published a study called "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata" by Yves-Alexandre de Montjoye et al. The article has reinvigorated claims that deidentified research data can be reidentified easily. These claims are not new, but their recitation in a vaunted science journal led to a new round of panic in the popular press.

Sample Unique ≠ Re-identifiable

1.1 Million = small sample fraction

https://blogs.law.harvard.edu/infolaw/2015/04/28/is-de-identification-dead-again/

"Re-identification Science" Policy Short-comings:

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

1. Attacking only trivially de-identified data, where modern statistical disclosure control methods (like HIPAA) weren't used.
2. Targeting 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.

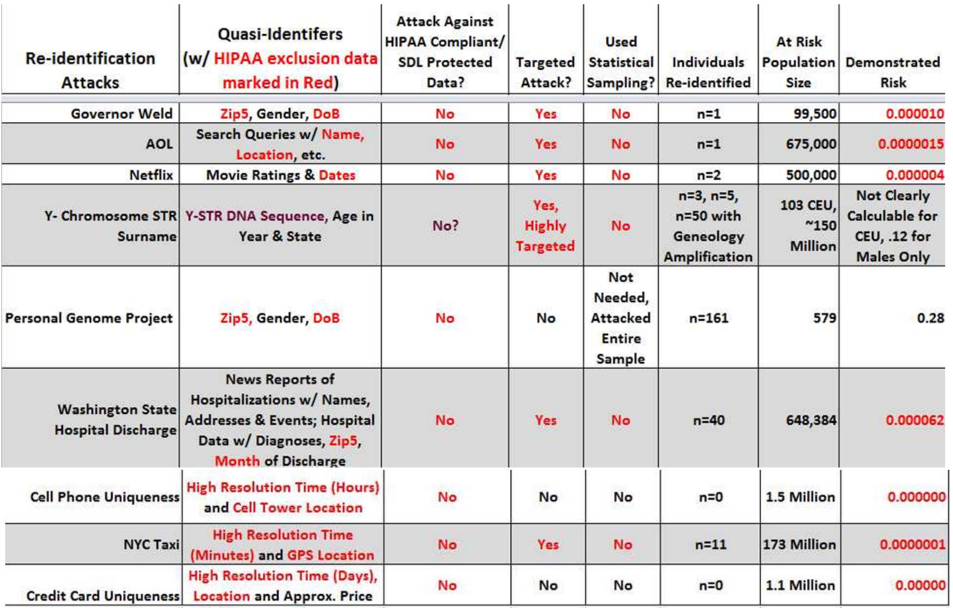
Corollary: Not designing experiments to show the boundaries where de-identification finally succeeds.

"Re-identification Science" Policy Short-comings:

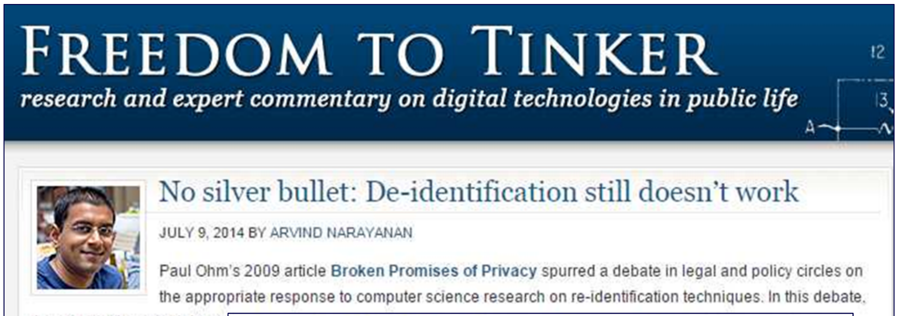
6 ways in which "Re-identification Science" has thus far typically failed to support sound public policies:

4. Failing to distinguish between sample uniqueness, population uniqueness and re-identifiability (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.

Re-identification Demonstration Attack Summary



Re-identification Attacks	Quasi-Identifers (w/ HIPAA exclusion data marked in Red)	Attack Against HIPAA Compliant/ SDL Protected Data?	Targeted Attack?	Used Statistical Sampling?	Individuals Re-identified	At Risk Population Size	Demonstrated Risk
Governor Weld	Zip5, Gender, DoB	No	Yes	No	n=1	99,500	0.000010
AOL	Search Queries w/ Name, Location, etc.	No	Yes	No	n=1	675,000	0.0000015
Netflix	Movie Ratings & Dates	No	Yes	No	n=2	500,000	0.000004
Y- Chromosome STR Surname	Y-STR DNA Sequence, Age in Year & State	No?	Yes, Highly Targeted	No	n=3, n=5, n=50 with Geneology Amplification	103 CEU, ~150 Million	Not Clearly Calculable for CEU, .12 for Males Only
Personal Genome Project	Zip5, Gender, DoB	No	No	Not Needed, Attacked Entire Sample	n=161	579	0.28
Washington State Hospital Discharge	News Reports of Hospitalizations w/ Names, Addresses & Events; Hospital Data w/ Diagnoses, Zip5, Month of Discharge	No	Yes	No	n=40	648,384	0.000062
Cell Phone Uniqueness	High Resolution Time (Hours) and Cell Tower Location	No	No	No	n=0	1.5 Million	0.000000
NYC Taxi	High Resolution Time (Minutes) and GPS Location	No	Yes	No	n=11	173 Million	0.0000001
Credit Card Uniqueness	High Resolution Time (Days), Location and Approx. Price	No	No	No	n=0	1.1 Million	0.00000



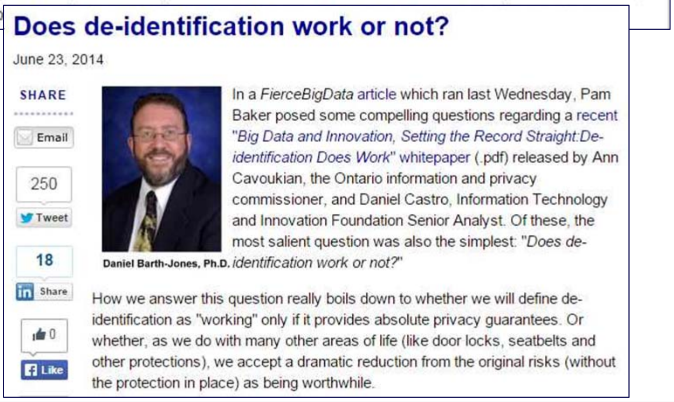
FREEDOM TO TINKER

research and expert commentary on digital technologies in public life

No silver bullet: De-identification still doesn't work

JULY 9, 2014 BY ARVIND NARAYANAN

Paul Ohm's 2009 article **Broken Promises of Privacy** spurred a debate in legal and policy circles on the appropriate response to computer science research on re-identification techniques. In this debate, the empirical research has o



Does de-identification work or not?

June 23, 2014

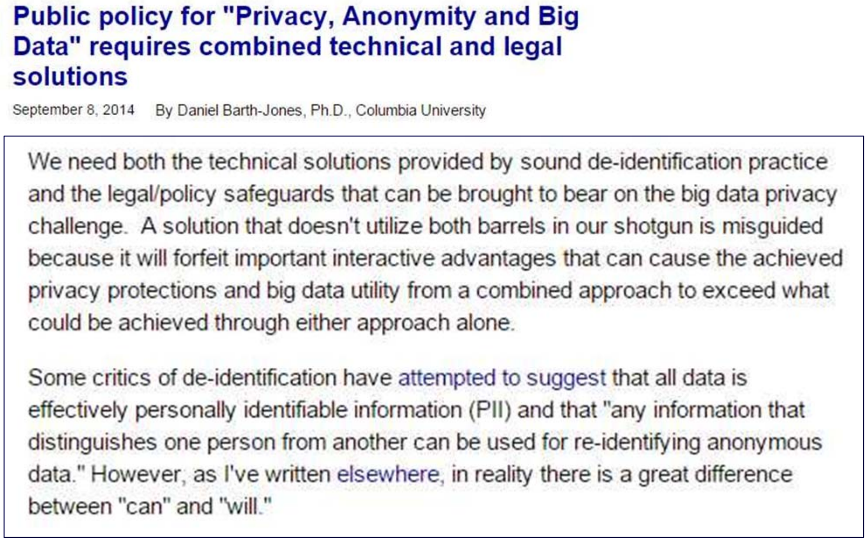
Daniel Barth-Jones, Ph.D.

In a *FierceBigData* article which ran last Wednesday, Pam Baker posed some compelling questions regarding a recent *"Big Data and Innovation, Setting the Record Straight:De-identification Does Work"* whitepaper (.pdf) released by Ann Cavoukian, the Ontario information and privacy commissioner, and Daniel Castro, Information Technology and Innovation Foundation Senior Analyst. Of these, the most salient question was also the simplest: *"Does de-identification work or not?"*

How we answer this question really boils down to whether we will define de-identification as "working" only if it provides absolute privacy guarantees. Or whether, as we do with many other areas of life (like door locks, seatbelts and other protections), we accept a dramatic reduction from the original risks (without the protection in place) as being worthwhile.

Unrealistic Insistence on Absolute Privacy Guarantees?

"Not having a Silver Bullet is not a good reason for dumping all your Ammo"



Public policy for "Privacy, Anonymity and Big Data" requires combined technical and legal solutions

September 8, 2014 By Daniel Barth-Jones, Ph.D., Columbia University

We need both the technical solutions provided by sound de-identification practice and the legal/policy safeguards that can be brought to bear on the big data privacy challenge. A solution that doesn't utilize both barrels in our shotgun is misguided because it will forfeit important interactive advantages that can cause the achieved privacy protections and big data utility from a combined approach to exceed what could be achieved through either approach alone.

Some critics of de-identification have attempted to suggest that all data is effectively personally identifiable information (PII) and that "any information that distinguishes one person from another can be used for re-identifying anonymous data." However, as I've written elsewhere, in reality there is a great difference between "can" and "will."

Reserve Slides for Questions

§164.514(b)(2)(i) -18 Safe Harbor Exclusion Elements

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;

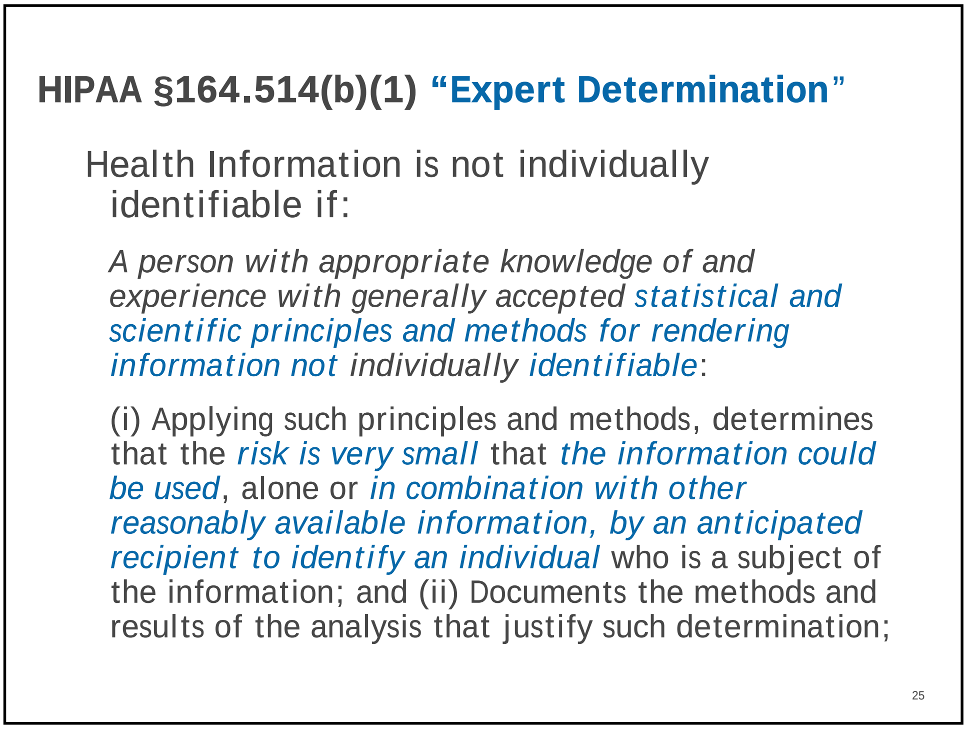
(N) Web Universal Resource Locators (URLs);

(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)



HIPAA §164.514(b)(1) "Expert Determination"

Health Information is not individually identifiable if:

A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

(i) Applying such principles and methods, determines that the *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* who is a subject of the information; and (ii) Documents the methods and results of the analysis that justify such determination;