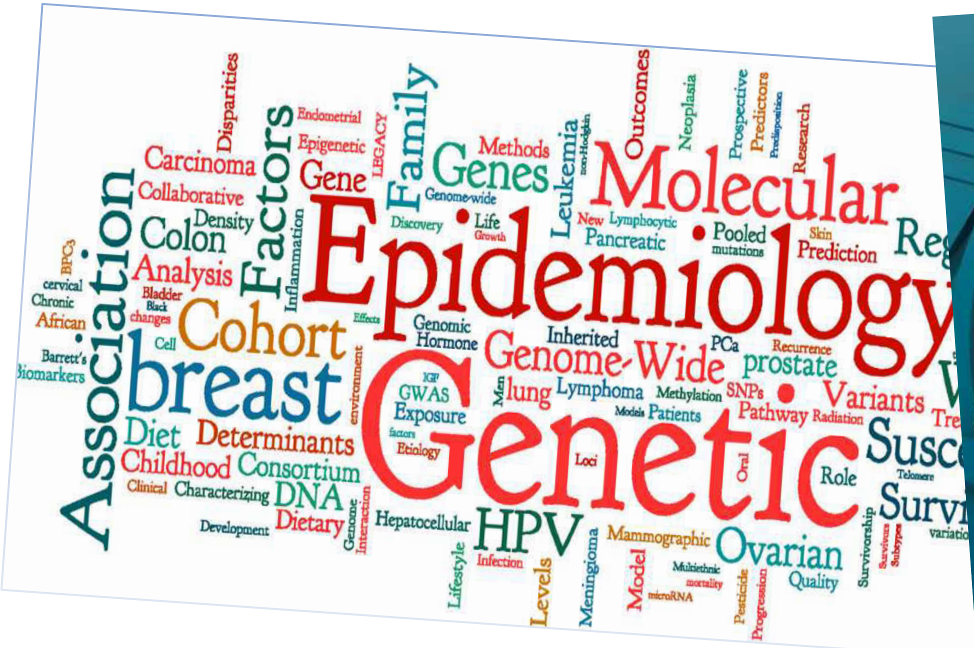
NCVHS Hearing:
De-identification and HIPAA

May 24, 2016

Improving
HIPAA De-identification
Public Policy

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

A Historic and Important Societal Debate is underway…



Public Policy Collision Course

The Research Value of De-identified Data

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)

- **Post-HIPAA 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)
- **Post-HIPAA 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

Misconceptions about HIPAA De-identified Data:

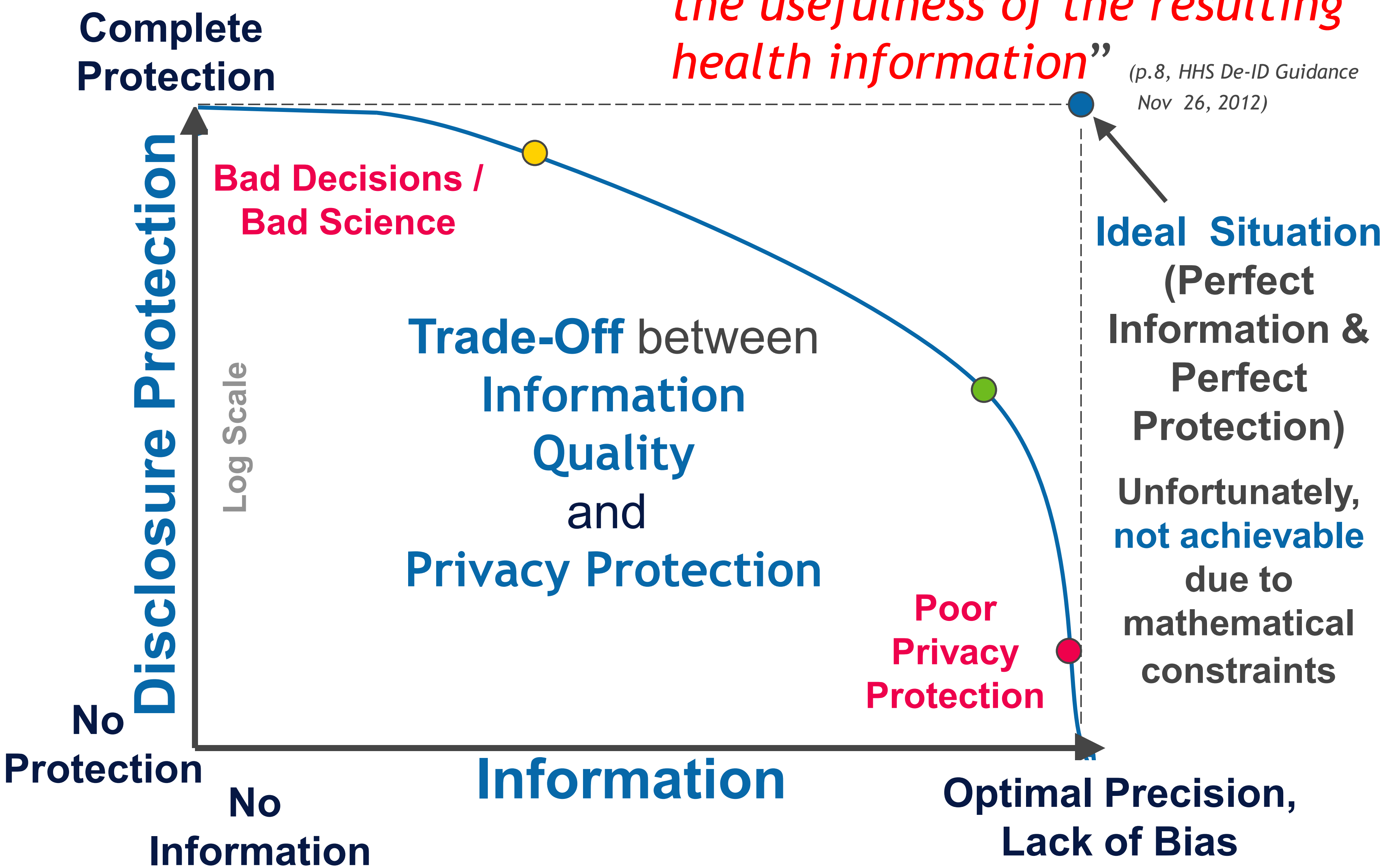
"It works perfectly and permanently…"

- Reality:
  - Perfect de-identification is not possible.
  - De-identifying does not free data from all possible subsequent privacy concerns.
  - Data is never permanently "de-identif<u>ied"</u>…

    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.

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)*



Balancing Disclosure Risk/Statistical Accuracy

- Balancing disclosure risks and statistical accuracy is essential because some popular de-identification methods (e.g. k-anonymity) can unnecessarily, and often undetectably, degrade the accuracy of de-identified data for multivariate statistical analyses or data mining (distorting variance-covariance matrixes, masking heterogeneous sub-groups which have been collapsed in generalization protections)
- This problem is well-understood by statisticians, but not as well recognized and integrated within public policy.
- Poorly conducted de-identification can lead to "bad science" and "bad decisions".

Reference: C. Aggarwal http://www.vldb2005.org/program/paper/fri/p901-aggarwal.pdf

Percent of Regression Coefficients which changed Significance:

T.S. Gal et al. / Journal of Biomedical Informatics xxx (2014) xxx–xxx

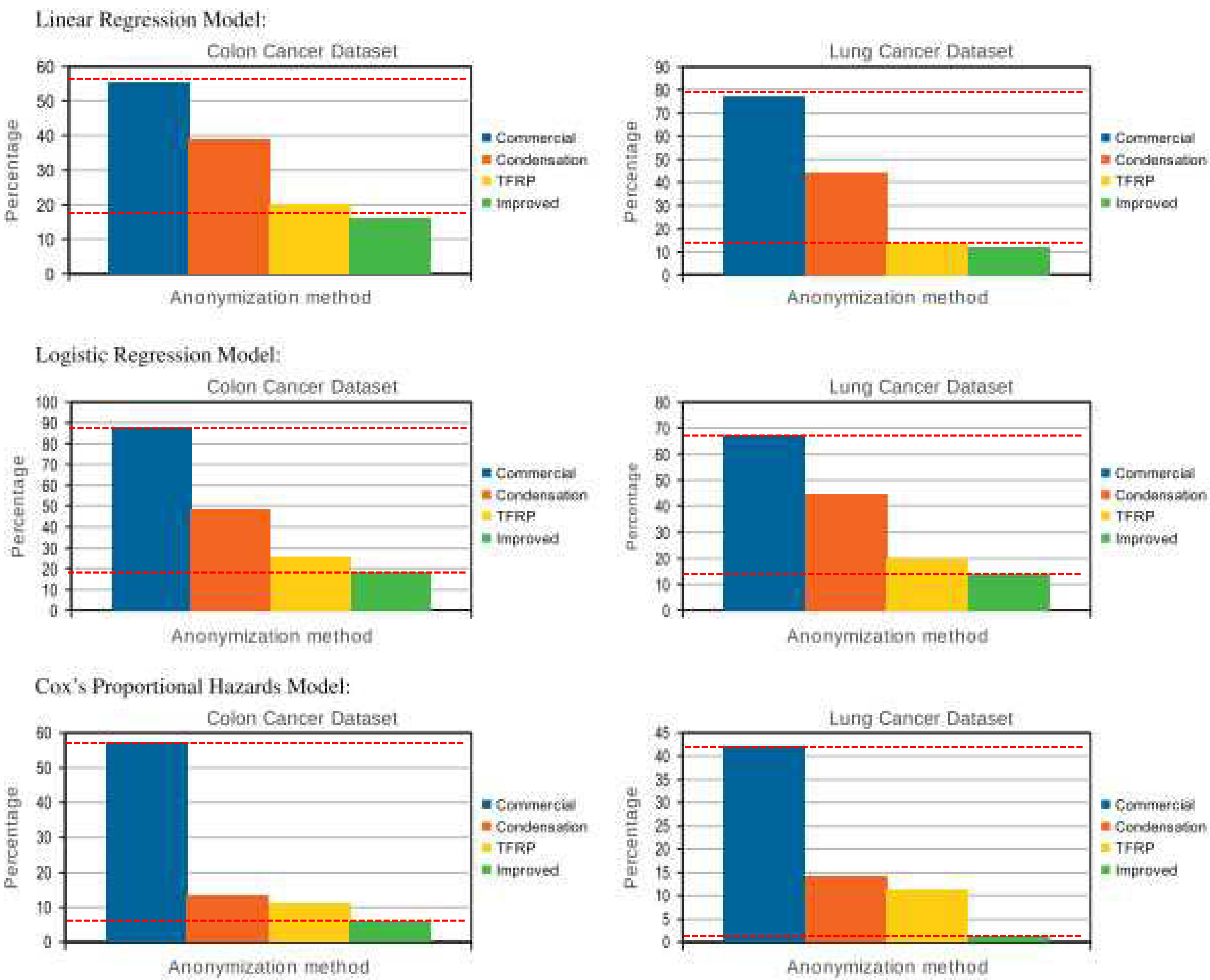


Fig. 1. Coefficients changed significance.

If this is what we are going to do to our ability to conduct accurate research - then… we should all just go home.

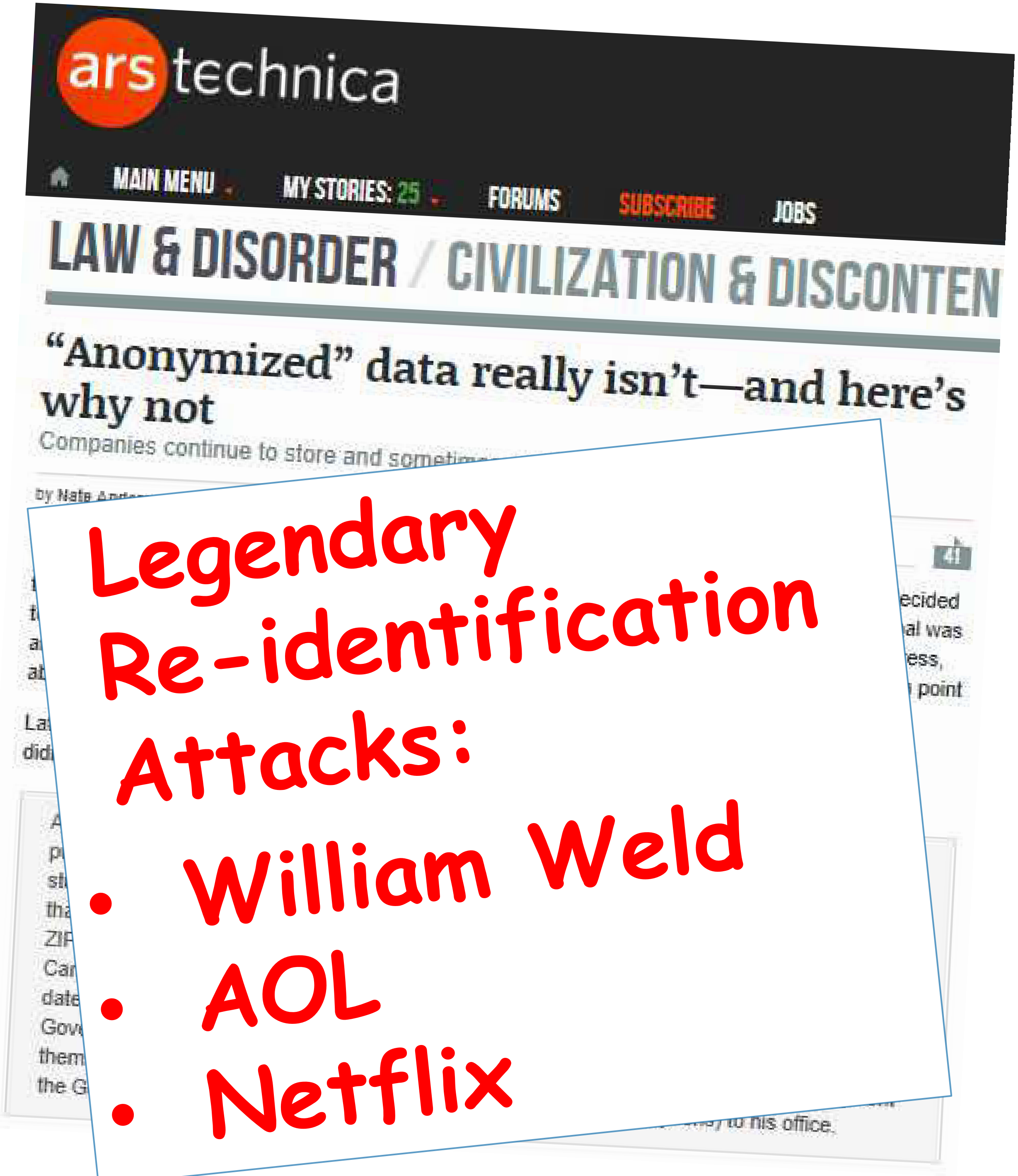
- Although poorly conducted de-identification can distort our ability to learn what is true leading to "bad science/decisions", this does not need to be an inevitable outcome.
- Well-conducted de-identification practice always carefully considers both the re-identification risk context and examines and controls the possible distortion to the statistical accuracy and utility of the de-identified data to assure de-identified data has been appropriately and usefully de-identified.
- But doing this requires a firm understanding/grounding in the extensive body of the statistical disclosure control/limitation literature.

Data Privacy Concerns are Far Too Important (and Complex) to be summed up with Catch Phrases or "Anecdata"

Eye-catching headlines and twitter-buzz announcing *"There's No Such Thing as Anonymous Data"* might draw the public's attention to broader and important concerns about data privacy in this era of "Big Data",

but such statements are essentially meaningless, even misleading, for further generalization without consideration of the specific de/re-identification contexts -- including the precise data details (e.g., number of variables, resolution of their coding schemas, special data properties, such as spatial/geographic detail, network properties, etc.) de-identification methods applied, and associated experimental design for re-identification attack demonstrations.

Good Public Policy demands reliable scientific evidence…

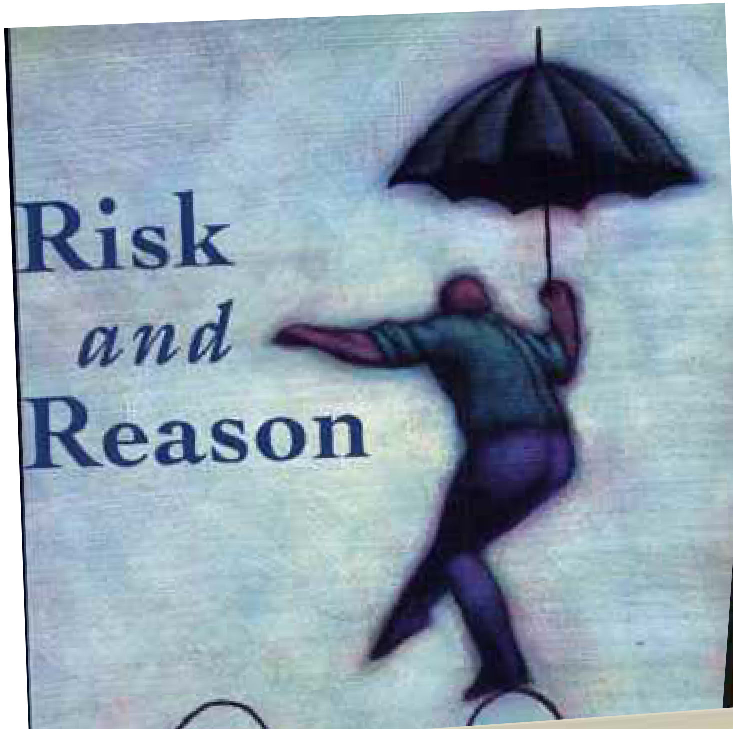


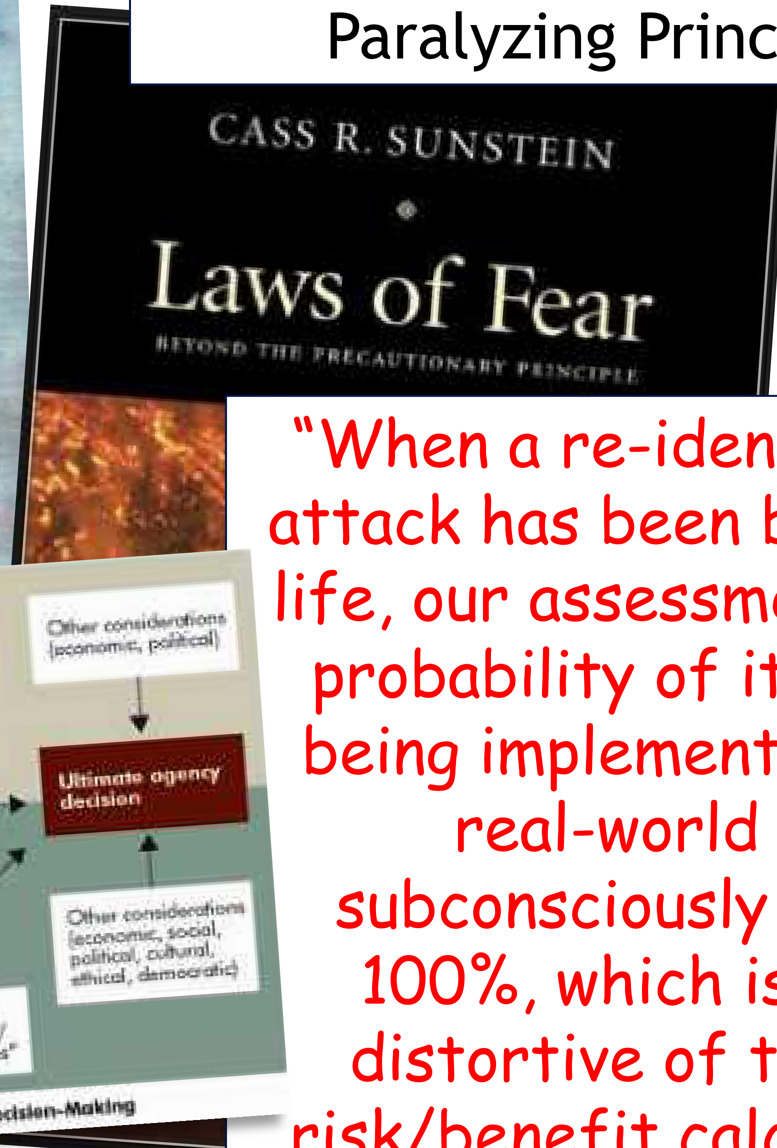
Legendary Re-identification Attacks:

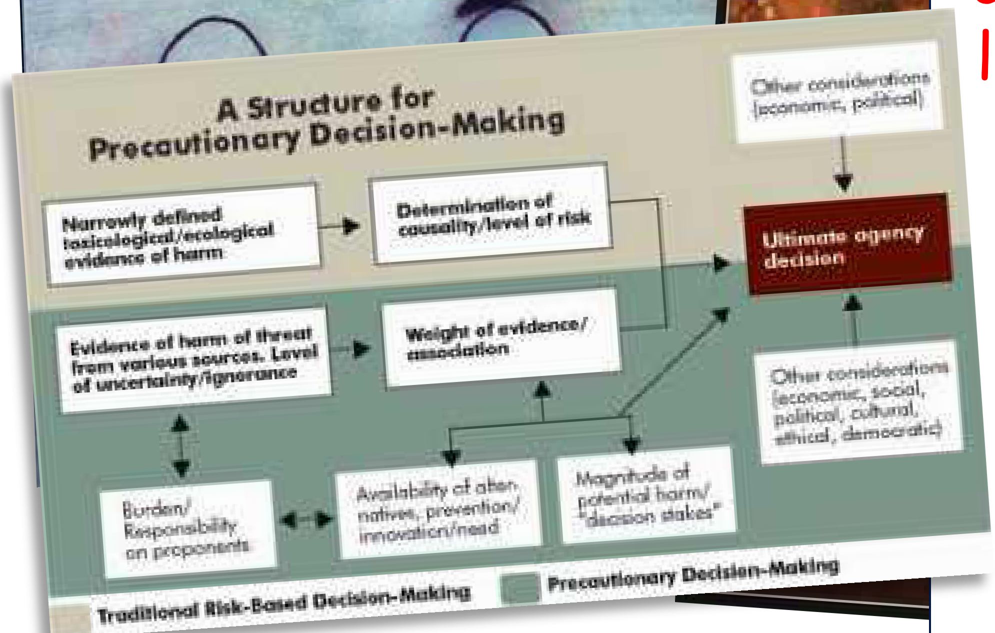
- William Weld
- AOL
- Netflix

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

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

Re-identification Demonstration Attack Summary

Re-identification Attacks	Quasi-Identifers (w/ HIPAA exclusion data marked in Red)	Vulnerable Subgroup Targeted?	Statistical Sampling to Select Targets?	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* [22]	No	0.00001
AOL [3]	Free Text from Search Queries w/ Name, Location, etc	Yes	No	n=1	675,000	*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"* [23]	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
Y-Chromosome STR Surname Inference [6,7] - 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"* [24]	*No?	.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* [25]	*Safe Harbor Excludes: *Any unique identifying #, characteristic or code*	Not Clearly Calculable for CEU Attack
Personal Genome Project [8,9,10]	Zip5, Gender, DoB	No	N/A	n=161	579	*"...re-identified names of > 40% anonymous participants"* [26]	No	0.28 (w/ Embedded Names Excluded)
Washington St. Hospital Discharge [11,10]	Hospitalization News Reports w/ Names,Addresses,Events 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"* [27]	No	0.000062
Cell Phone "Unicity" [12]	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%"* [12]	No	0.0
NYC Taxi [13,14]	High Resolution Time (Minutes) and GPS Locations	Yes	No	n=11	173 Million Rides	*How Big Brother Watches You With Metadata* [28]	No	0.0000001
Credit Card "Unicity" [15,16,17,18,19,20,21]	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* [29]	No	0.0

- *Publicized attacks have been on data without HIPAA 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 reliable evidence supporting this portrayal.*

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.

Re-identification Science Policy Short-comings:

Cont'd: 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 (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*.

Re-identification Science Can Better Inform Policy/Practice

1. Demonstrate re-identification risks on data where modern statistical disclosure control methods have actually been used.
2. Use proper statistical random samples and scientific study designs in order to provide *representative* risk estimates.
3. Use ethically-designed re-identification experiments to better characterize re-identification risks for quasi-identifiers beyond simple demographics
4. Design experiments to show the boundaries where de-identification finally succeeds and provide evidence to justify any data intruder knowledge assumptions.
5. Verify re-identifications and report false-positive rates for supposed re-identifications.
6. Investigate multiple realistic and relevant threats and fully specify these re-identification threat models.
7. Use modern probabilistic uncertainty analyses to examine impact of uncertainties in re-identification experiments.

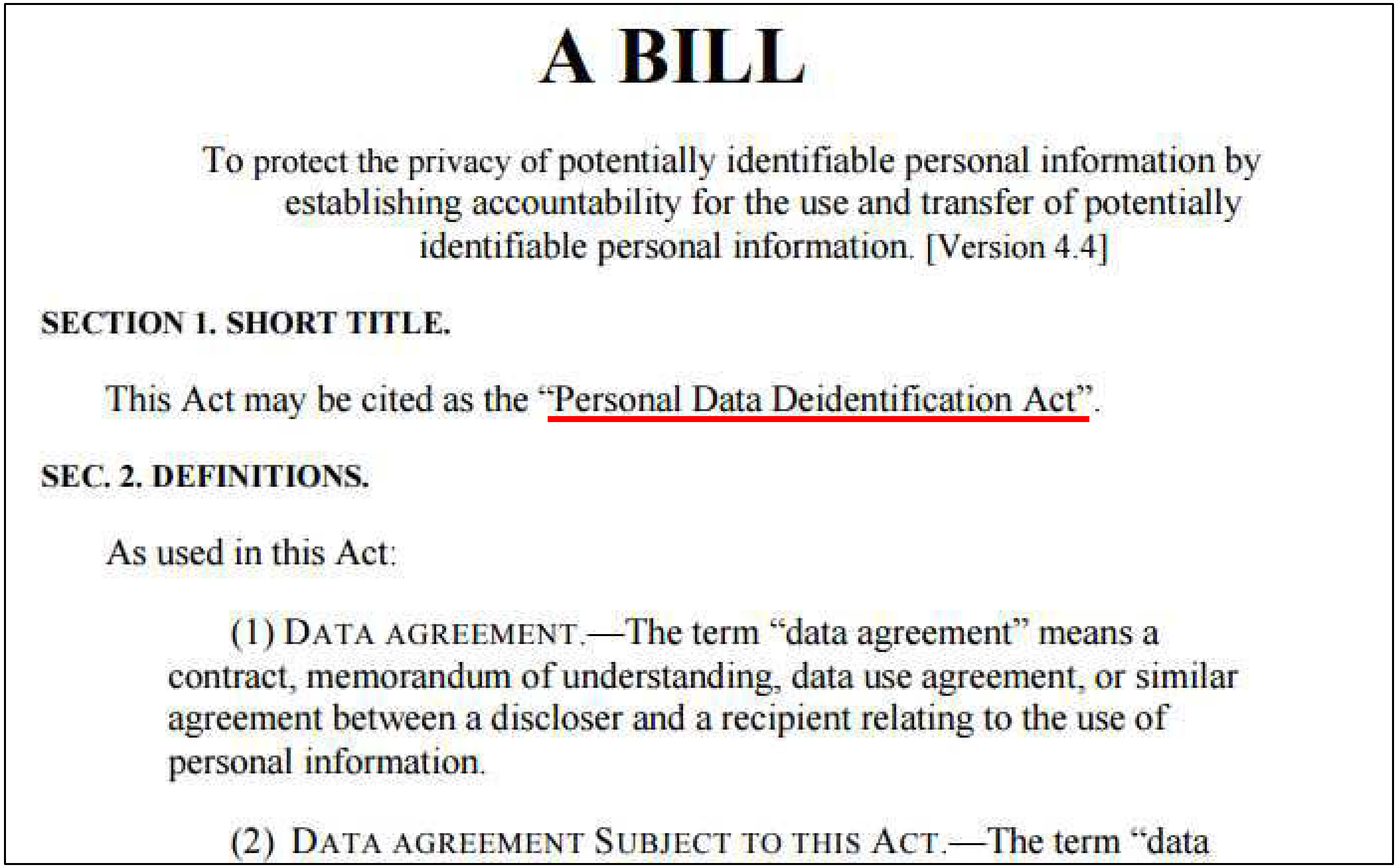
Recommended De-identified Data Use Requirements

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 certification 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.

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.

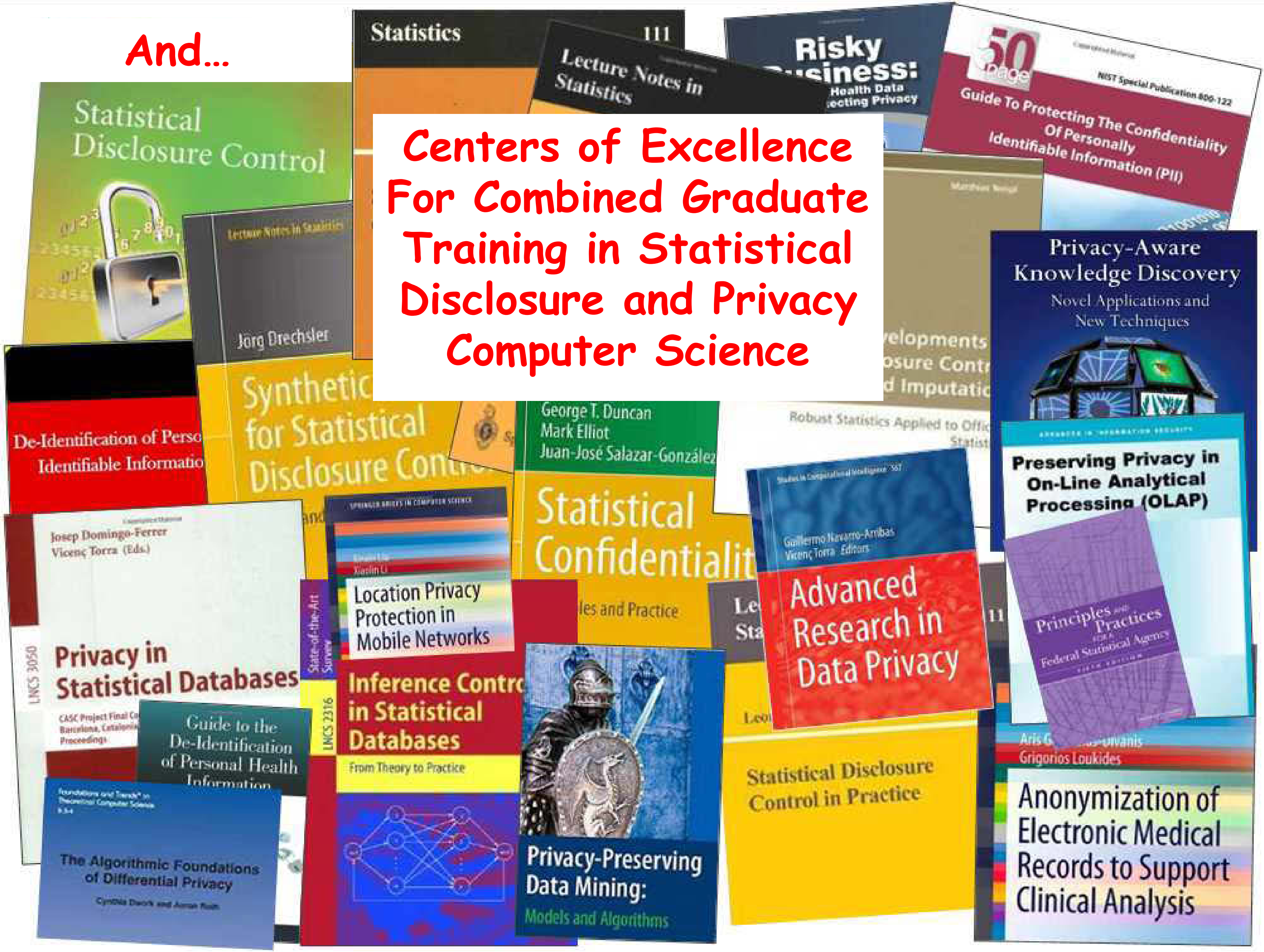
(2) DATA AGREEMENT SUBJECT TO THIS ACT.—The term "data

Robert Gellman, 2010

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

And…

Centers of Excellence For Combined Graduate Training in Statistical Disclosure and Privacy Computer Science



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. Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, Yaniv Erlich. Identifying Personal Genomes by Surname Inference. Science 18 Jan 2013: 321-324.
7. 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/
8. Sweeney, L., Abu, A, Winn, J. Identifying Participants in the Personal Genome Project by Name (April 29, 2013). http://ssrn.com/abstract=2257732
9. 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/
10. 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/
11. Sweeney, L. Matching Known Patients to Health Records in Washington State Data (June 5, 2013). http://ssrn.com/abstract=2289850
12. 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
13. 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/
14. 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/

References for Re-identification Attack Summary Table

15. 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.
16. Barth-Jones D, El Emam K, Bambauer J, Cavoukian A, Malin B. Assessing data intrusion threats. Science. 2015 Apr 10; 348(6231):194-5.
17. de Montjoye, et al. Assessing data intrusion threats—Response Science. 10 Apr 2015: Vol. 348, Issue 6231, pp. 195
18. Jane Yakowitz Bambauer. Is De-Identification Dead Again? April 28, 2015. https://blogs.harvard.edu/infolaw/2015/04/28/is-de-identification-dead-again/
19. 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.
20. 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
21. 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
22. 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/
23. 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
24. 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/
25. 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
26. 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/
27. Sweeney L. Only You, Your Doctor, and Many Others May Know. Technology Science. 2015092903. September 29, 2015. http://techscience.org/a/2015092903
28. 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
29. 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/

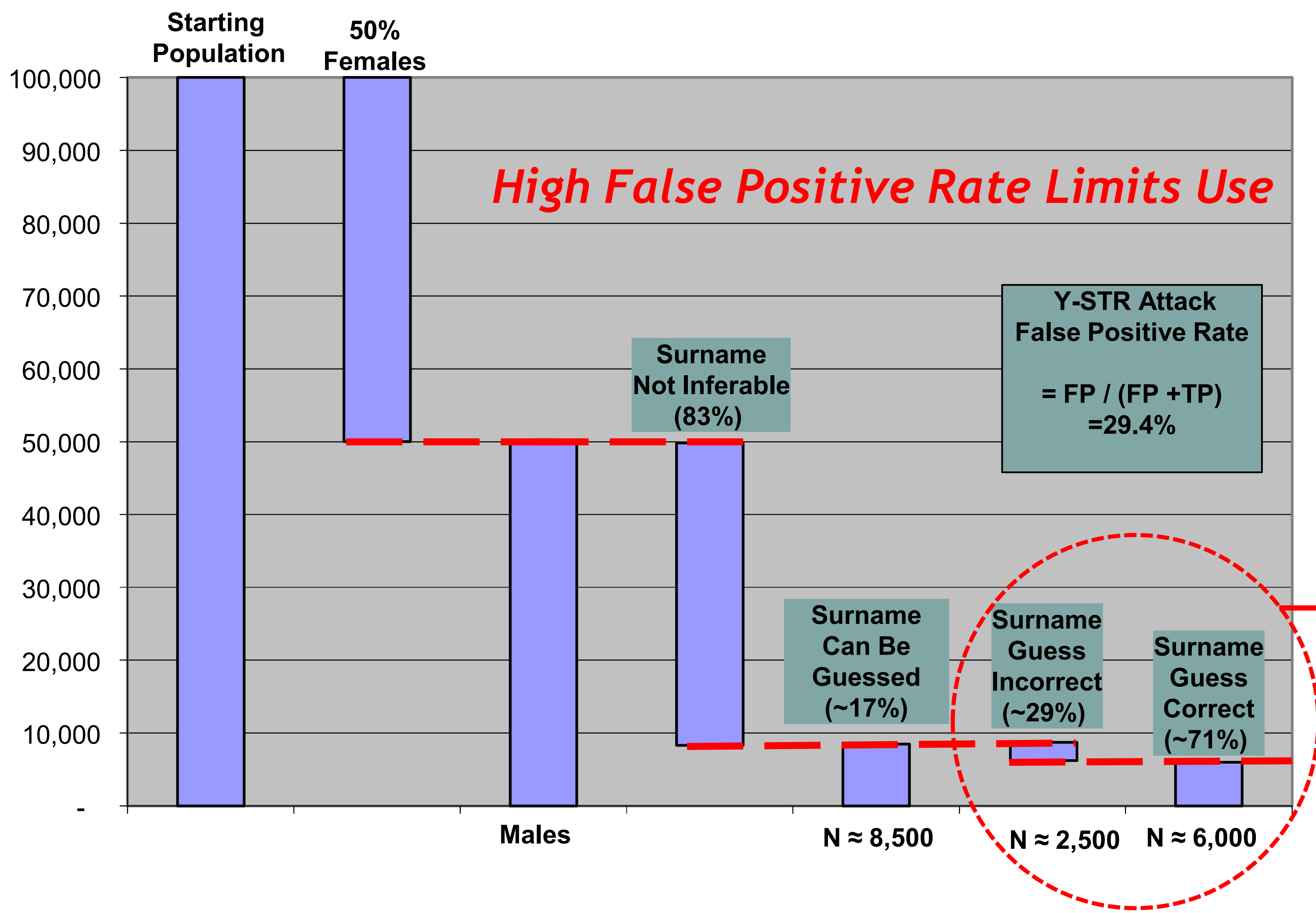
Reserve Slides for Questions

Question 1: Is Y-STR Attack Economically Viable?

Probably not -- unclear whether it eventually could be.

Q2: Is Genomic "De-identification" pointless?

No, removing State, Grouping YoB would help importantly.



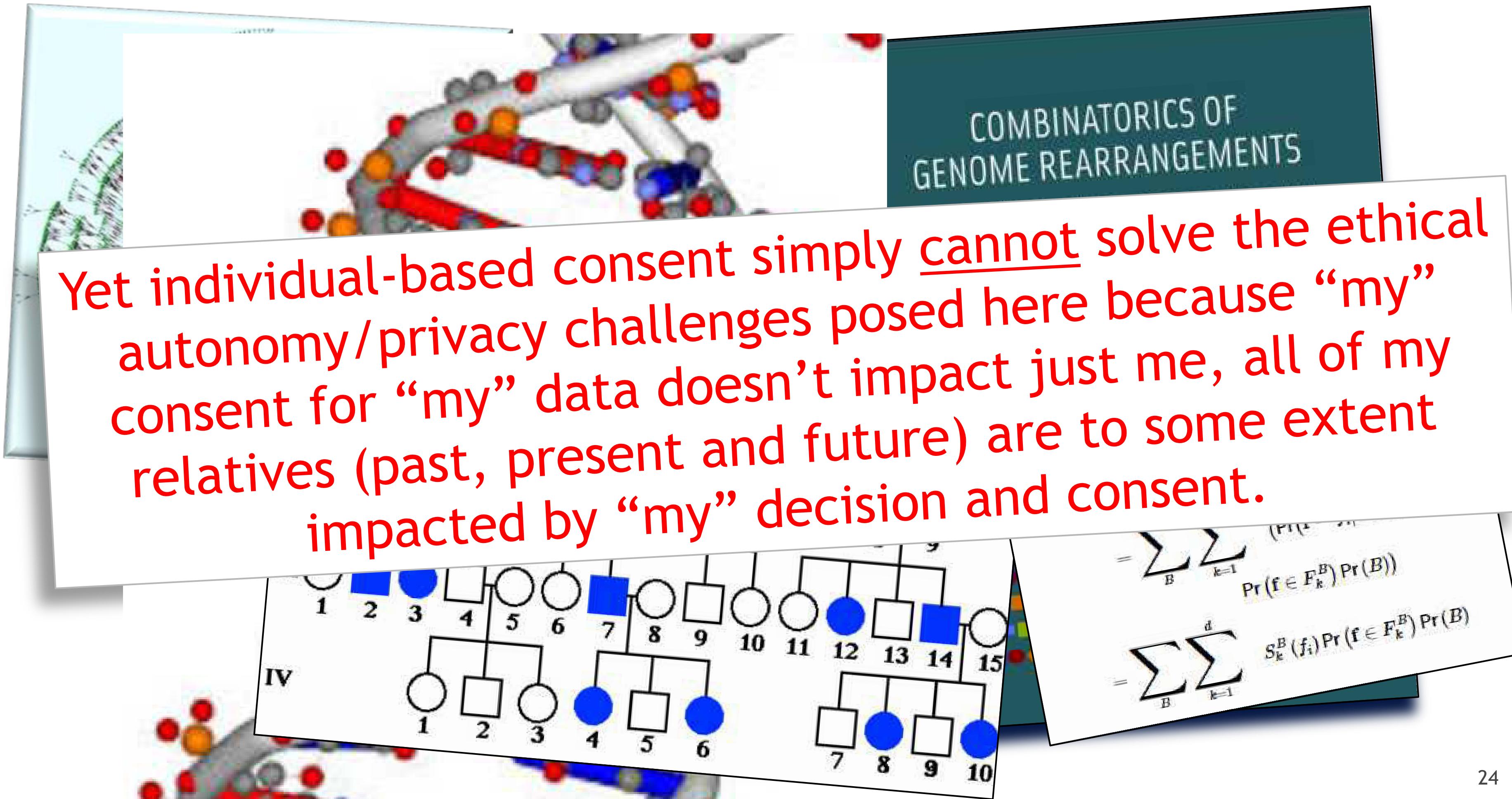
Re-ID isn't achieved by Surname Guess.

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

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.



Yet individual-based consent simply cannot solve the ethical autonomy/privacy challenges posed here because “my” consent for “my” data doesn’t impact just me, all of my relatives (past, present and future) are to some extent impacted by “my” decision and consent.