More privacy to formalize...

Simson L. Garfinkel

May 24, 2017

Privacy Semester at Simons Institute 2019 workshop

Photos from https://pixabay.com

More privacy to formalize

Abstract:

Differential privacy provides a formal definition of data privacy within a database, but experience has shown that it's hard to apply differential privacy beyond structured sets of tabular data and some limited graph databases. However, there are many kinds of information that require sharing and computation. Simple datatypes include time, geographical, and imagery information. How do you privatize a picture of a crowd? Today practitioners are at a loss for privatizing even many kinds of structured information, such as 3D models or genetic information. In the cybersecurity world, there is a need to privatize netflow data, cyber threat intelligence, and provenance. And then there's text. Even if the world of tabular databases, we still lack tools for applying differential privacy to high-dimensional data. Differential privacy doesn't seem to have a concept of group privacy. Finally, while differential privacy does give us tools for private data publishing, it is silent on the privacy of data users.

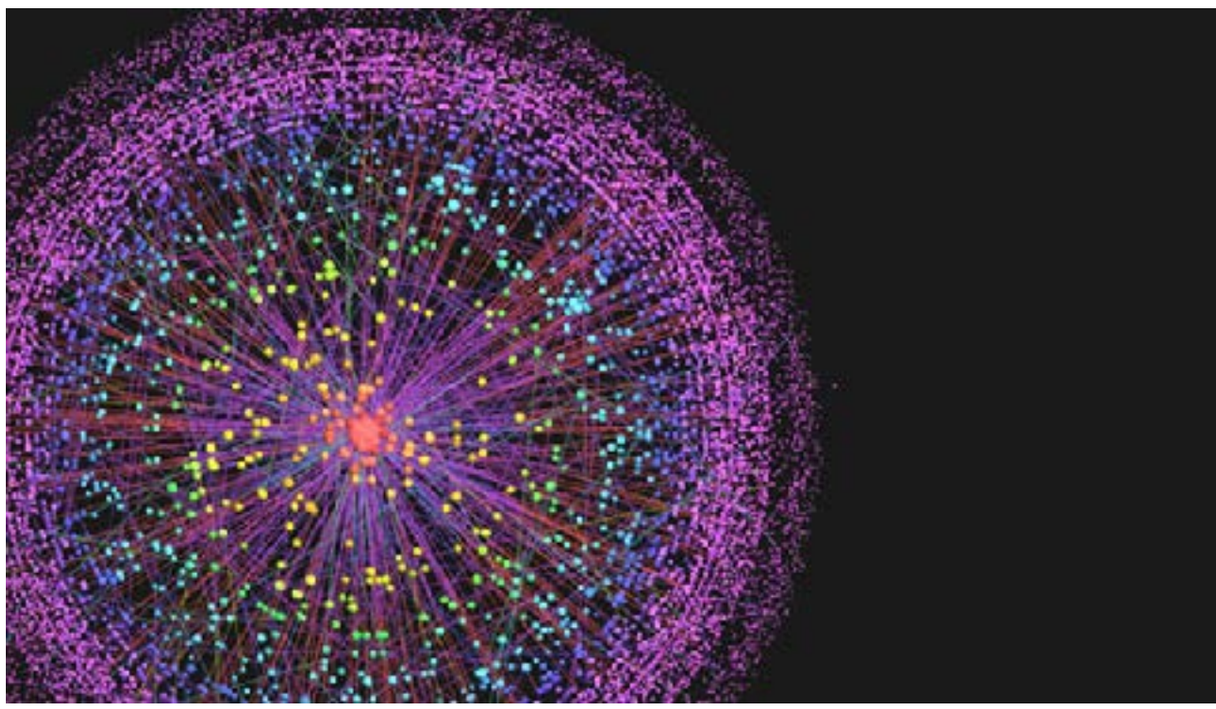
Simson Garfinkel will present a slide for each of these examples, discussing how it would be really neat to privatize this kind of data, but no recommendations on how to addresses these open problems.

Differential privacy and synthetic data

"The Best API is no-API"

Differential privacy was created for interactive queries.
Data scientists want to work with data.

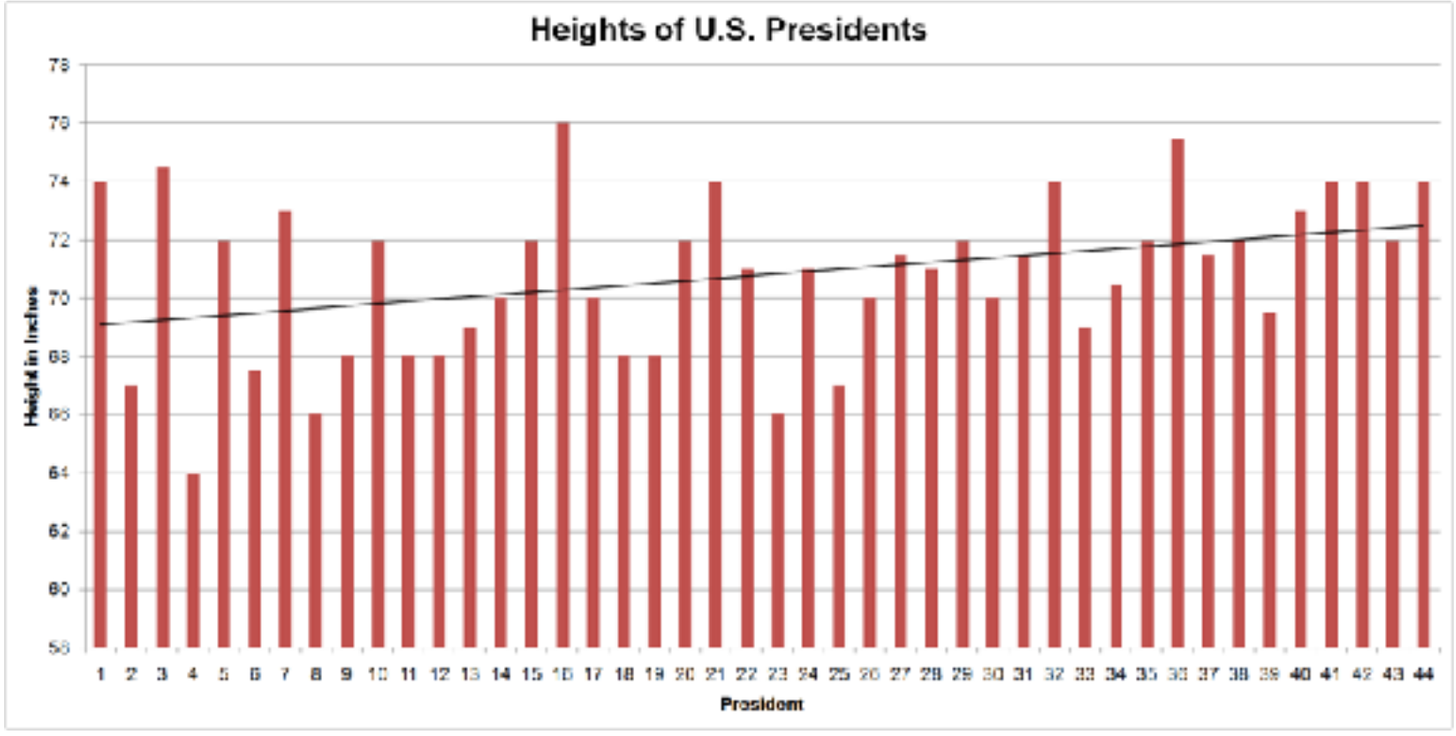
We need better tools for creating high-dimensionality synthetic data.



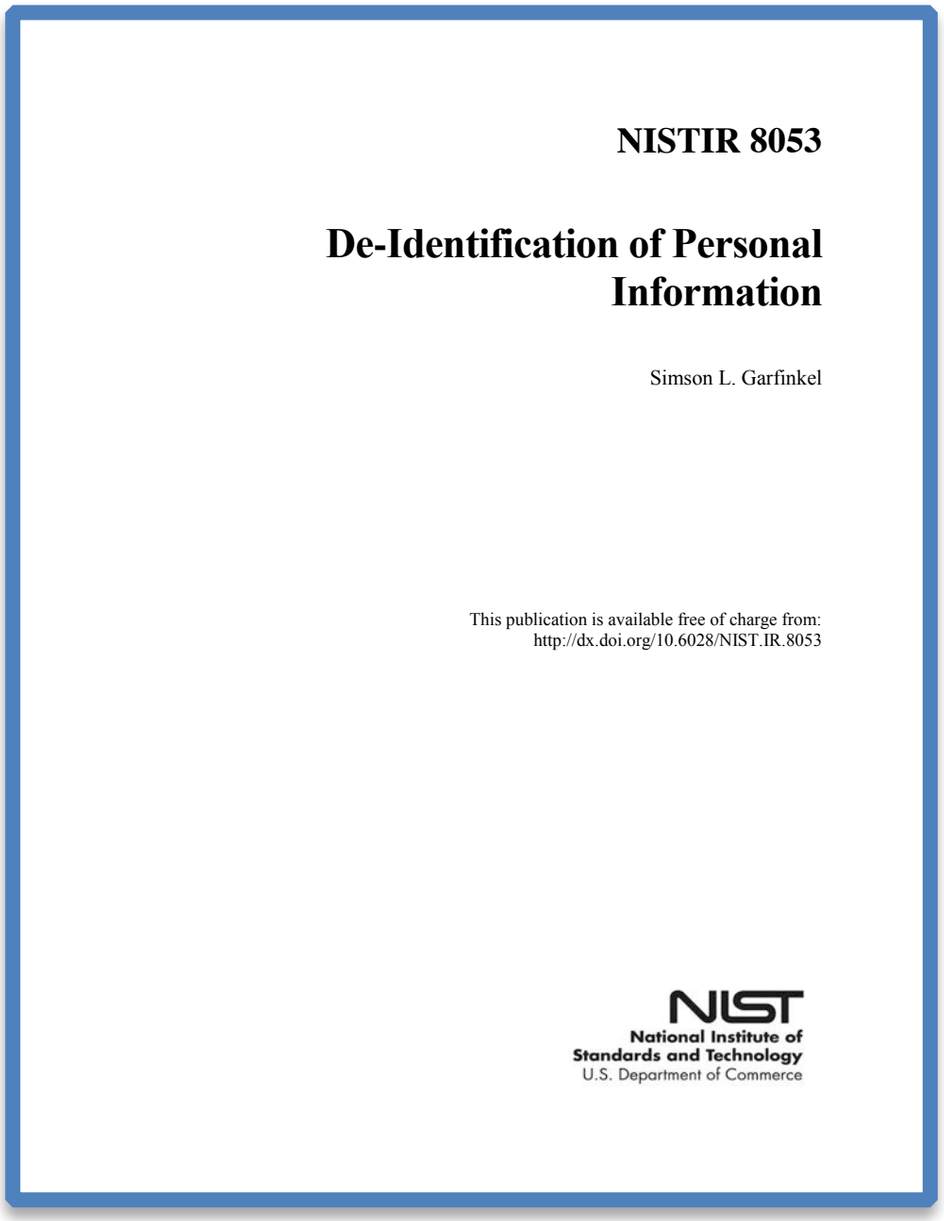
Really, we need just one tool.

Differential Privacy has been (mostly) focused on tabular data.

President	Birth	Date of Inauguration	Age at Inauguration	Height (cm)
George Washington	February 22, 1732	April 30, 1789	57 years, 67 days	188 cm
John Adams	October 30, 1735	March 4, 1797	61 years, 125 days	170 cm
Thomas Jefferson	April 13, 1743	March 4, 1801	57 years, 325 days	189 cm
James Madison	March 16, 1751	March 4, 1809	57 years, 353 days	163 cm
James Monroe	April 28, 1758	March 4, 1817	58 years, 310 days	183 cm
John Quincy Adams	July 11, 1767	March 4, 1825	57 years, 236 days	171 cm
Andrew Jackson	March 15, 1767	March 4, 1829	61 years, 354 days	185 cm
Martin Van Buren	December 5, 1782	March 4, 1837	54 years, 89 days	168 cm



While you are working on differential privacy, the world is pursuing de-identification.

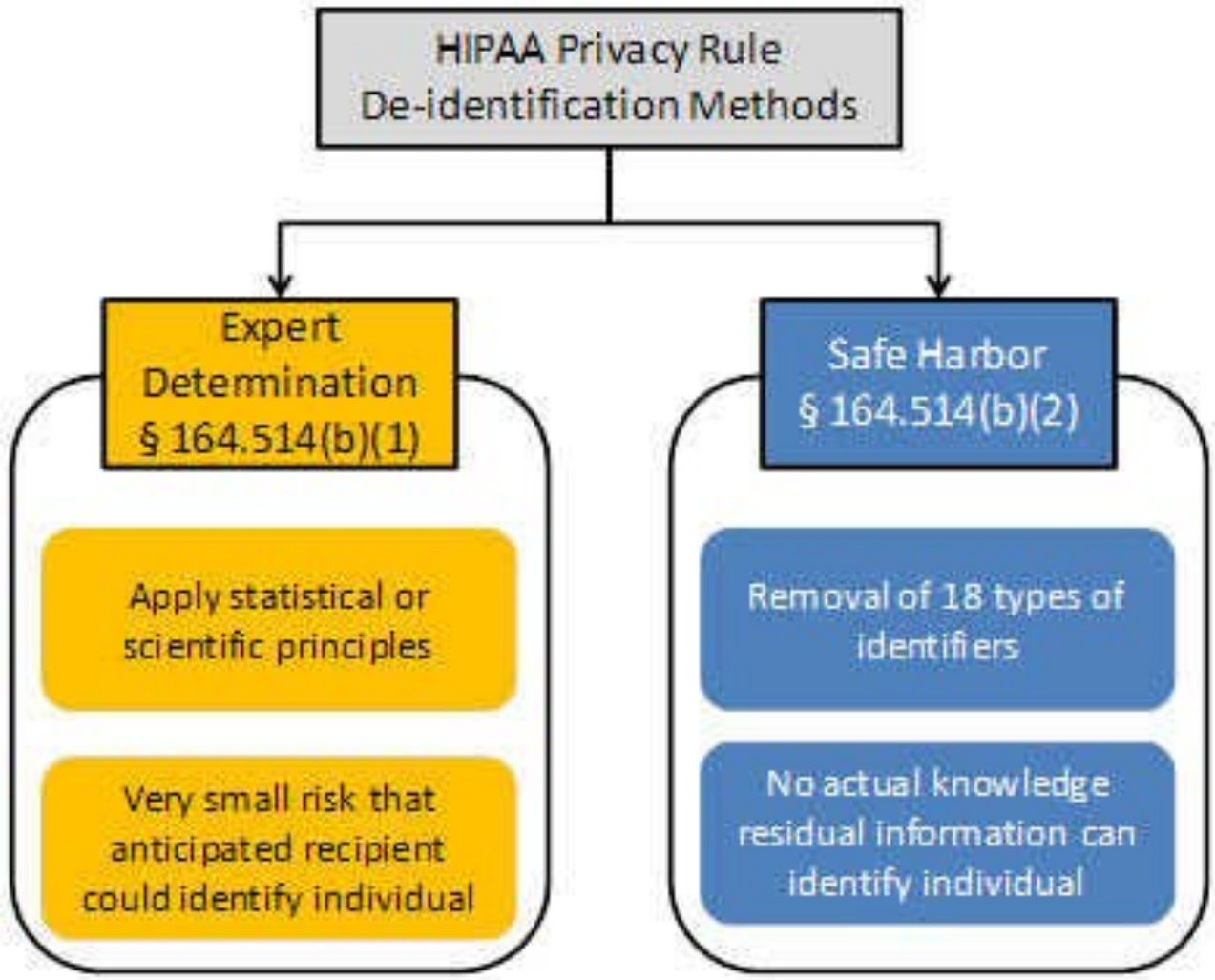

NISTIR 8053

De-Identification of Personal Information

Simson L. Garfinkel

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8053

NIST
National Institute of Standards and Technology
U.S. Department of Commerce



De-identificati is safe. De-identificati is predictable. De-identi tion works.

Some existing US laws and regulations recognize/require de-identification

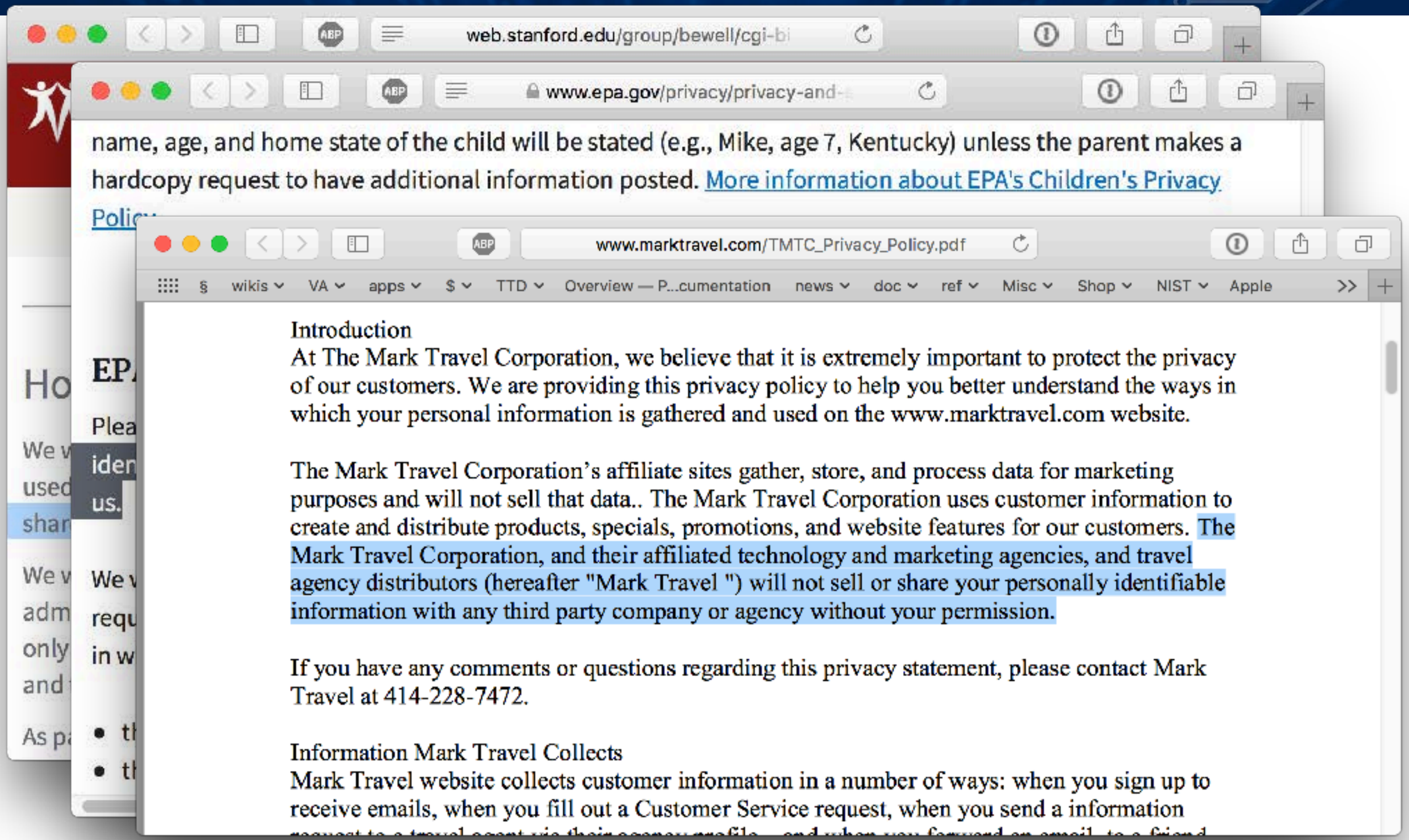
Educational records can be released if de-identified (FERPA).

Medical records can be released if de-identified (HIPAA)

Foodborne Illness Surveillance System allows public release of de-identified aggregate data.

Voluntary safety reports submitted to FAA can be released if the data they contain are de-identified.

De-identification is easy.



More privacy to formalize

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification.

Imagery...

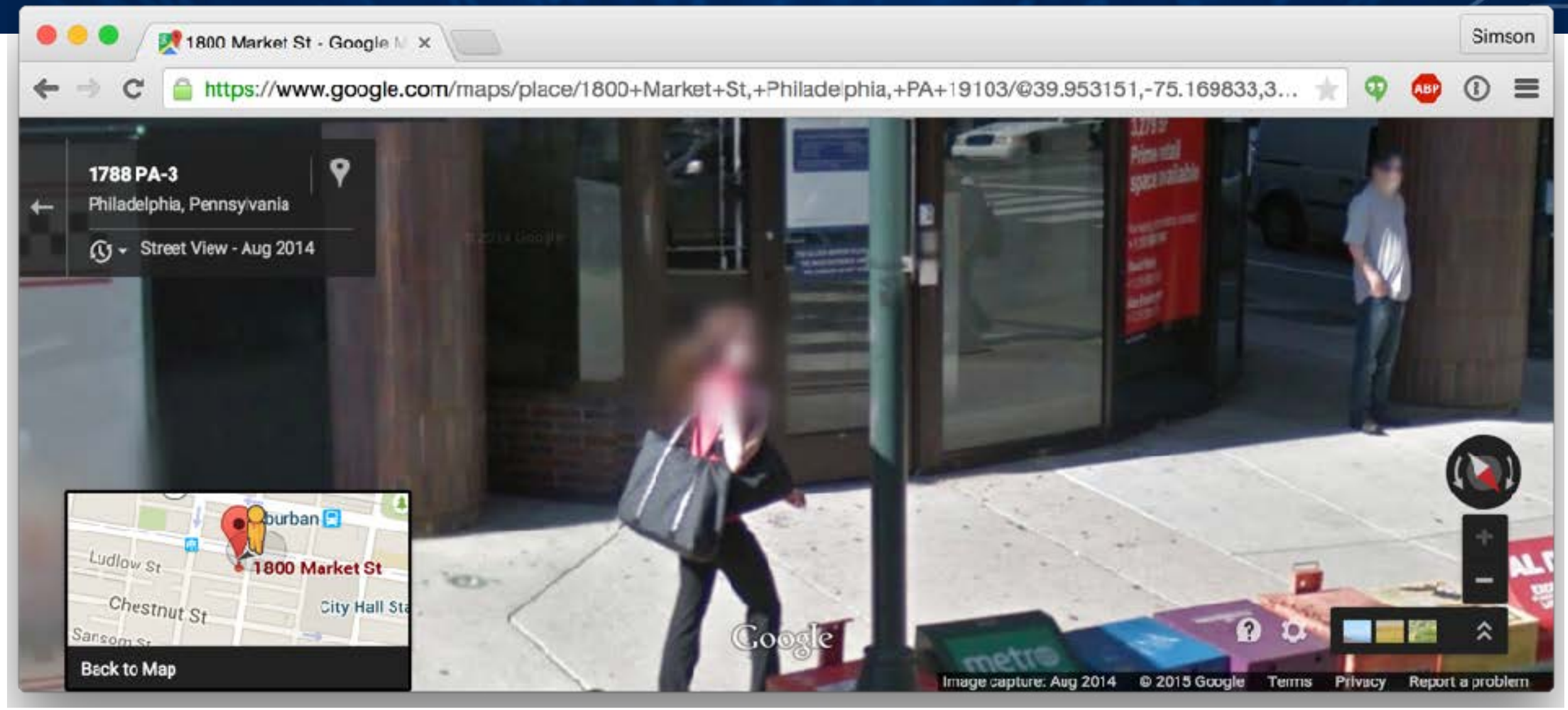
Multimedia de-identification / redaction is an area of growing concern.

A primary interest is public release of police body cameras:

http://www.cam.ac.uk/research/news/first-scientific-report-shows-police-body-worn-cameras-can-prevent-unacceptable-use-of-force

Other uses:

- Scientific research; privacy preserving surveillance; data retention



"Large-scale Privacy Protection in Google Street View," Frome et al, 2009

Most research has focused on faces and license plates

- Google's Street View — 90% of faces; 95% of license plates

De-identifying photographs and video

Key challenges:

- What to remove?
- Usefulness of de-identified imagery
- Evaluation

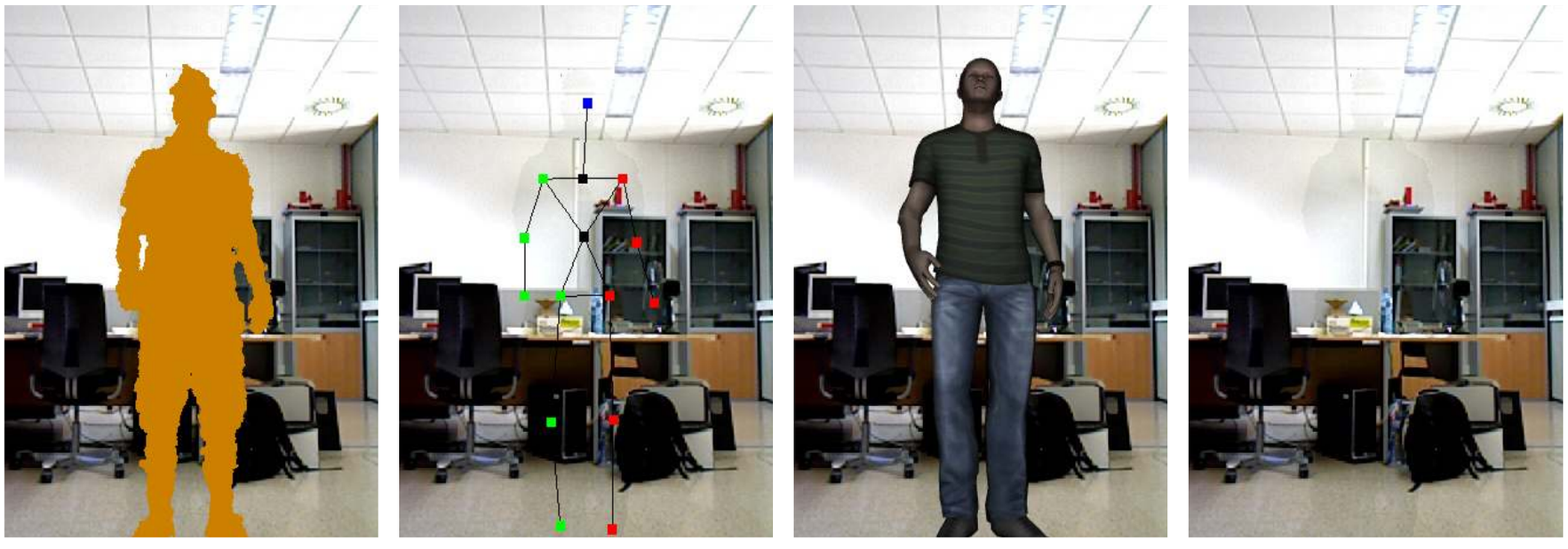
"Face encryption"

Figure 4: Two examples of an encrypted image where the face of the person is considered the sensitive region. Reprinted from Boult (2005).

Medical imagery

http://www.randomhistory.com/photos/2014/scoliosis-xray.jpg

(a) Real image (b) Blur (c) Pixelating (d) Emboss



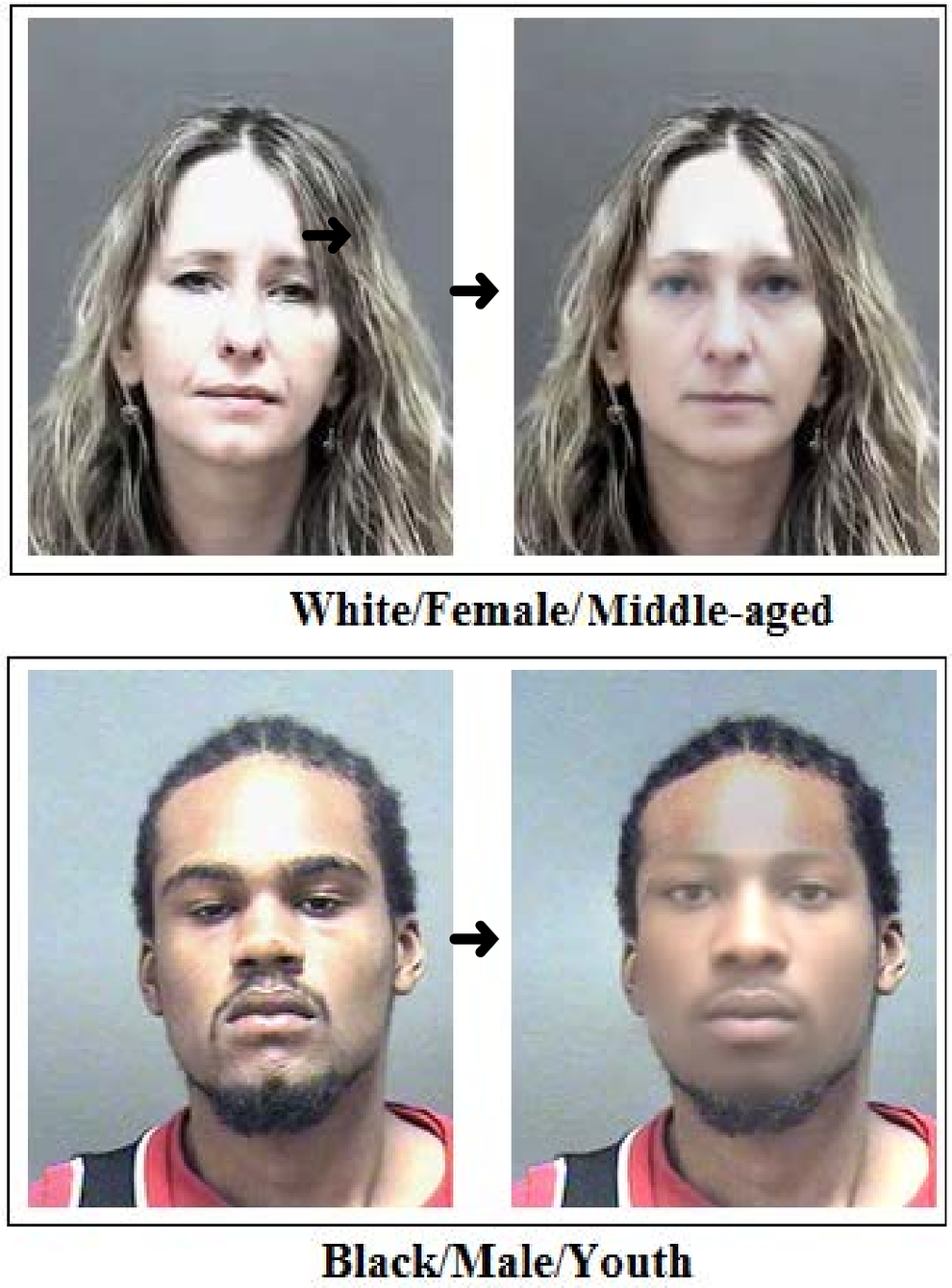
(e) Solid silhouette (f) Skeleton (g) 3D avatar (h) Invisibility

Visual Privacy Protection Methods: A Survey, Jose Ramon Padilla-Lopez, Alexandros Andre Chaaraoui, Francisco Florez-Revuelta, 2015

Obscuring with synthetic faces: preserves context, prevents automated identification

These techniques can preserve:

- Gender
- Race
- Age



Effectiveness:

- Stops automated face identification.
- Humans can still identify people they know

(a) Real image

(b) Modified image

Figure 6: An example of a people removal method where the person has been manually selected in the real image (a), and then automatically removed in the second image (b) by filling the region concerning the person using an exemplar-based image inpainting method. Reprinted from Criminisi et al. (2004).

More privacy to formalize

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification.
2. Imagery

Geospatial information

Everything happens somewhere.

Some locations can be highly identifying

- A farm house on the prairie.

Some locations are not identifying in 2D but highly identifying in 3D:

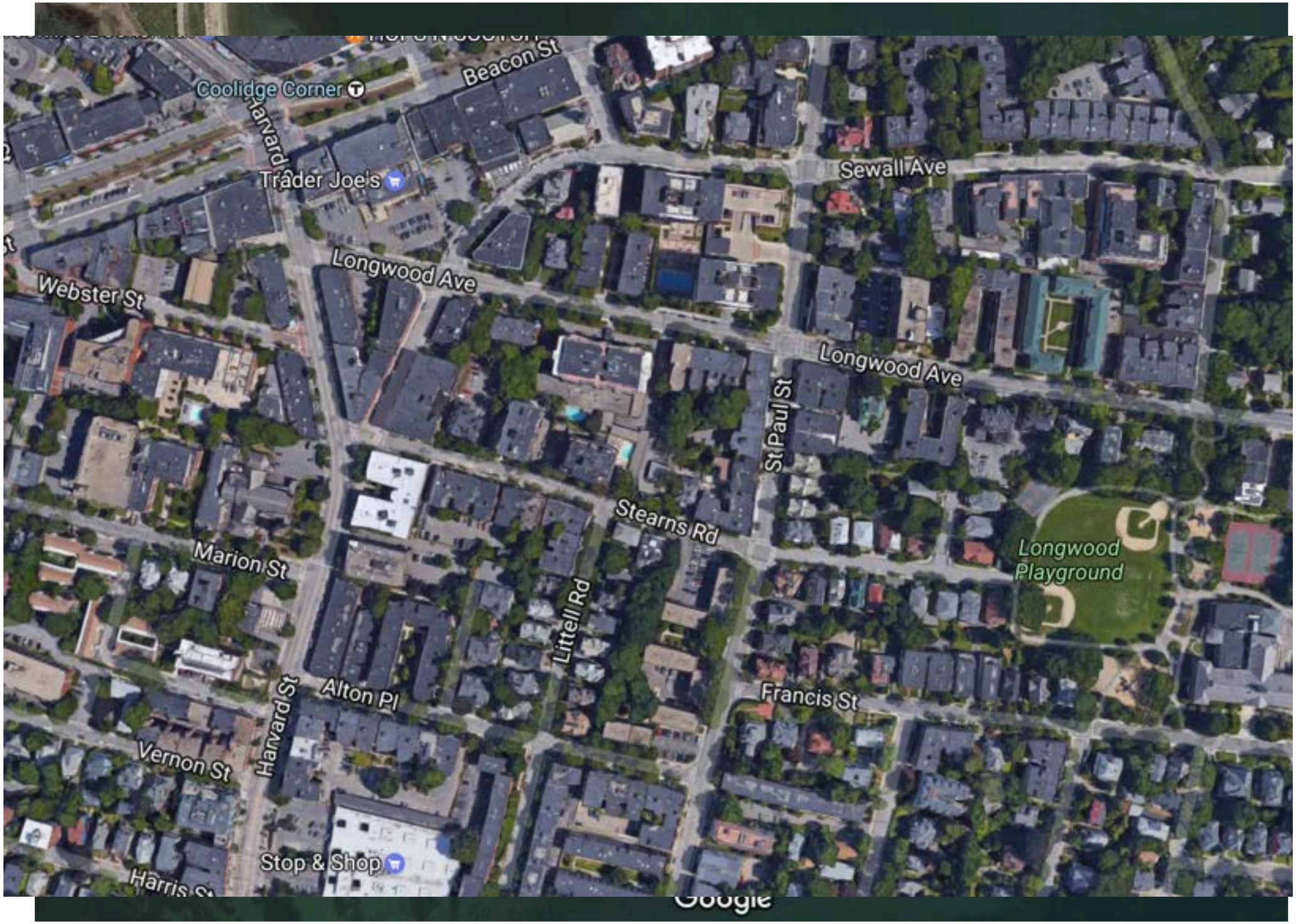
- A 1-bedroom apartment in a high-rise apartment building.

Some locations are identifying if you have temporal information:

- The speaker’s podium at a public event.

And most locations aren’t identifying at all.

Noise infusion doesn't work for geospatial



1000 ft

Geospatial privacy literature search

National Research Council (2007). **Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data.** M. P. Gutmann and P. C. Stern, eds. Committee on the Human Dimensions of Global Change, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.http://books.nap.edu/catalog.php?record_id=11865 Accessed 2 March 2011.

Wang, H. and Reiter, J. P. (2012), **Multiple imputation for sharing precise geographies in public use data**, Annals of Applied Statistics, 6, 229 - 252.

T. Paiva, A. Chakraborty, J. P. Reiter, and A. E. Gelfand, (2014) **Imputation of confidential data sets with spatial locations using disease mapping models**, Statistics in Medicine, 33, 1928 - 1945

H. Quick, S. H. Holan, C. K. Wikle, and J. P. Reiter. (2015) **Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography**, Spatial Statistics, 14, 439 - 451

Anonymisation of geographical distance matrices via Lipschitz embedding, International Journal of Health Geographics, https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-015-0031-7

Richardson, Douglas B., Mei-Po Kwan, George Alter & Jean E. McKendry (2015) **Replication of scientific research: addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research**, Annals of GIS, 21:2, 101-110, DOI: 10.1080/19475683.2015.1027792

“Geo-privacy beyond coordinates” http://geog.ucsb.edu/~jano/agile2016p.pdf

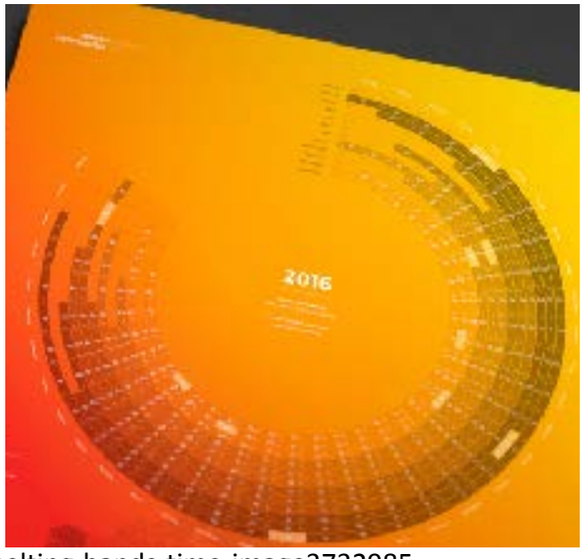
"A multiscale masking method for point geographic data," Keith C. Clarkea, International Journal of Geographical Information Science 30:2, 2016, pp. 300-315 DOI:10.1080/13658816.2015.1085540

More privacy to formalize...

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification
2. Imagery
3. Geospatial information

Time, time series and synthetic data

President	Birth	Date of Inauguration	Age at Inauguration	Height (cm)
George Washington	February 22, 1732	April 30, 1789	57 years, 67 days	188 cm
John Adams	October 30, 1735	March 4, 1797	61 years, 125 days	170 cm
Thomas Jefferson	April 13, 1743	March 4, 1801	57 years, 325 days	189 cm
James Madison	March 16, 1751	March 4, 1809	57 years, 353 days	163 cm
James Monroe	April 28, 1758	March 4, 1817	58 years, 310 days	183 cm
John Quincy Adams	July 11, 1767	March 4, 1825	57 years, 236 days	171 cm
Andrew Jackson	March 15, 1767	March 4, 1829	61 years, 354 days	185 cm
Martin Van Buren	December 5, 1782	March 4, 1837	54 years, 89 days	168 cm



https://www.dreamstime.com/royalty-free-stock-photo-melting-hands-time-image3732985

Time series synthetic data

2017

2019

2017

? ? ? ?

2019

How do we create formally private synthetic data with persistent identifiers from year-to-year?

Synthetic datasets ... how do we find them?

Synthetic population housing and person files for the United States

- https://zenodo.org/record/556121
- http://doi.org/10.5281/zenodo.556121

Synthetic Survey of Income and Program Participation:

- https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html

Synthetic Longitudinal Business Database:

- https://www.census.gov/ces/dataproducts/synlbd/

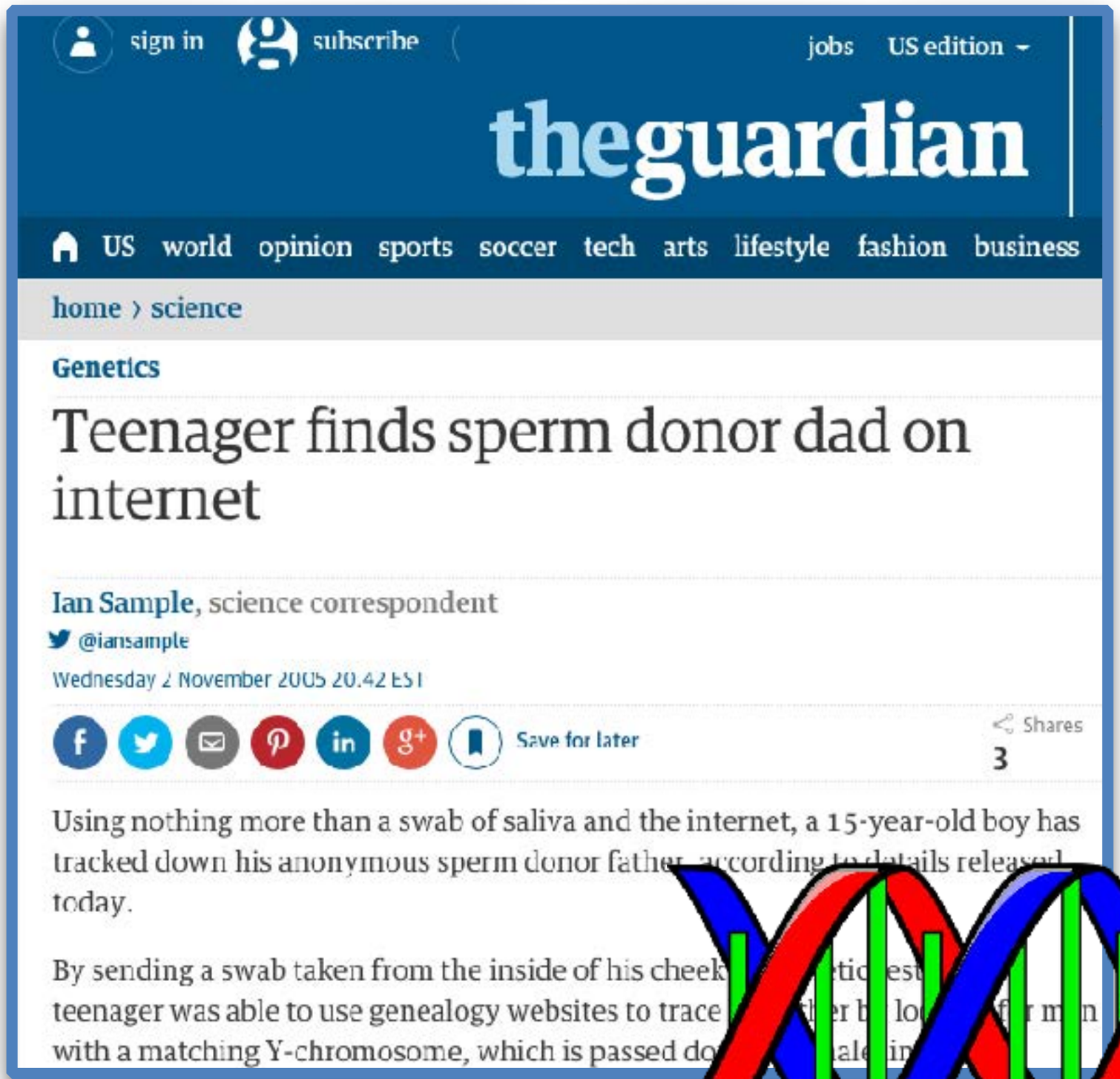
Virtual RDC@Cornell:

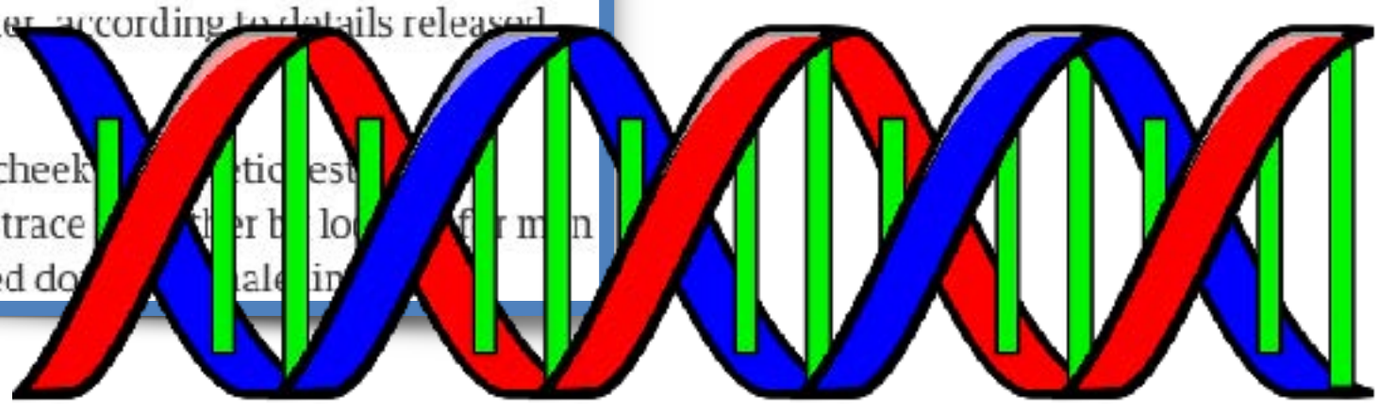
- https://www2.vrdc.cornell.edu/news/synthetic-data-server/

More privacy to formalize

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification
2. Imagery
3. Geospatial information
4. Time and time Series

Formally private genetic information?





Journal of Genetic Counseling, Vol. 4, No. 2, 1995

Huntington Disease: A Case Study Describing the Complexities and Nuances of Predictive Testing of Monozygotic Twins

Audrey Heimler[1,3] and Andrea Zanko[2]

When a candidate for predictive testing for the Huntington disease gene is a monozygotic twin, confidentiality of the co-twin's diagnosis and autonomy of participation are among the critical genetic counseling issues. Predictive testing can proceed when twins voluntarily and simultaneously request counseling and evaluation in an HD testing program. This case describes a young man referred for predictive testing to an HD testing site on the East Coast of the United States. Family history revealed a twin brother of unknown zygosity who resided on the West Coast of the United States. The genetic counselors on opposite coasts collaborated to provide genetic counseling and evaluation for voluntary, informed predictive testing of the twins, protecting their rights while observing national protocol guidelines.

KEY WORDS: Huntington disease; predictive testing; twins; confidentiality; autonomy.

http://simson.net/ref/1995/Huntington_Disease_Twins_Heimler_Zanko.pdf

Arch Neurol. 2005 Jun;62(6):995-7.

Monozygotic twins discordant for Huntington disease after 7 years.

Friedman JH1, Trieschmann ME, Myers RH, Fernandez HH.

BACKGROUND:
Huntington disease (HD) has only rarely been identified in identical twins. All described twins have had disease onset within 1 year of each other, suggesting that disease onset is determined solely by genetic influences.

OBJECTIVE:
To describe a unique set of monozygotic twins in whom clinical HD onset is at least 7 years apart.

DESIGN:
A 71-year-old woman was diagnosed as having HD based on medical history, physical examination results consistent with HD, and a CAG trinucleotide repeat number of 39 in the HD gene on chromosome 4. Her onset was 6 years earlier. Her genetically confirmed identical twin, carrying the same number of CAG repeats, was neurologically healthy when examined the next year. Only the HD-manifest twin had chronic bronchitis, rheumatoid arthritis, type 2 diabetes mellitus, and chronic anemia. Both had hypertension.

CONCLUSIONS:
To our knowledge, this is the first report of monozygotic twins discordant for HD by more than 2 years. The onset of HD symptoms in a patient with 39 triplet repeats at least 7 years earlier than her identical twin suggests the possibility that the disease may be initiated (or delayed) by environmental factors. We have identified increased cigarette use and longer exposure to various industrial toxins as potential explanations for the earlier onset in one twin.

PMID: 15956172 DOI: 10.1001/archneur.62.6.995

BRCA genetic issue

Forward

Journal List › Curr Oncol › v.22(4); 2015 Aug › PMC4530819

CURRENT ONCOLOGY

Curr Oncol. 2015 Aug; 22(4): e233–e236.
doi: 10.3747/co.22.2527
PMCID: PMC4530819

Is it time to offer *BRCA1* and *BRCA2* testing to all Jewish women?

K.A. Metcalfe, RN PhD,[*†] A. Eisen, MD,[‡§] J. Lerner-Ellis, PhD,[‖] and S.A. Narod, MD[†]

Author information ► Copyright and License information ►

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4530819/

More privacy to formalize

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification
2. Imagery
3. Geospatial information
4. Time and time series
5. Genomic information

Medical text — de-identifying medical narratives

Challenges:

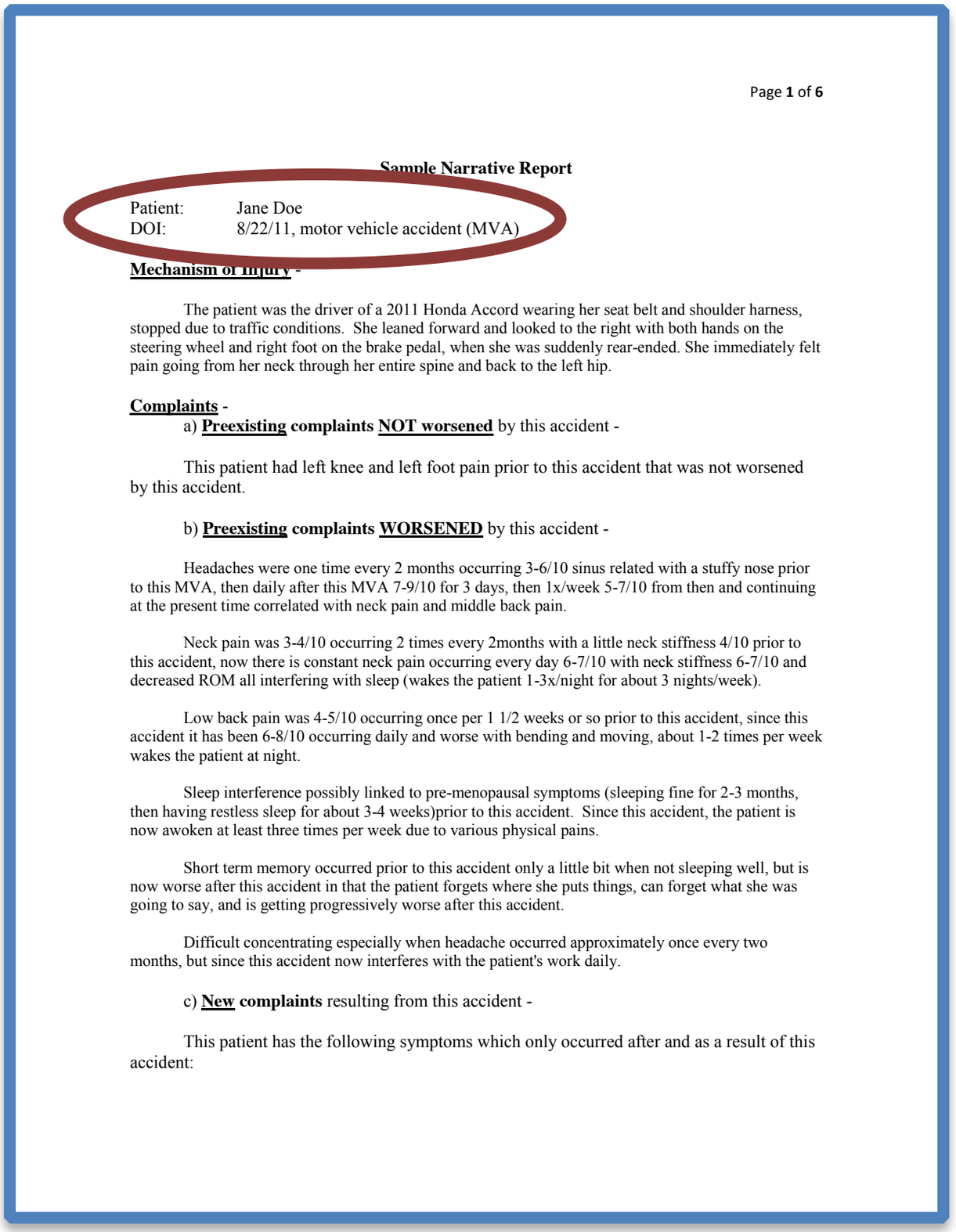
- Finding the direct identifiers
- Not removing important medical information like eponyms.
  (e.g. “Addison’s Disease”)

NL Approaches:

- Rule-based (e.g. regex)
- Statistical machine learning.

Several evaluations.

Success rate ≈ 95%



Sample Narrative Report

Patient: Jane Doe
DOI: 8/22/11, motor vehicle accident (MVA)

Mechanism of Injury -

The patient was the driver of a 2011 Honda Accord wearing her seat belt and shoulder harness, stopped due to traffic conditions. She leaned forward and looked to the right with both hands on the steering wheel and right foot on the brake pedal, when she was suddenly rear-ended. She immediately felt pain going from her neck through her entire spine and back to the left hip.

Complaints -

a) **Preexisting complaints NOT worsened** by this accident -

This patient had left knee and left foot pain prior to this accident that was not worsened by this accident.

b) **Preexisting complaints WORSENED** by this accident -

Headaches were one time every 2 months occurring 3-6/10 sinus related with a stuffy nose prior to this MVA, then daily after this MVA 7-9/10 for 3 days, then 1x/week 5-7/10 from then and continuing at the present time correlated with neck pain and middle back pain.

Neck pain was 3-4/10 occurring 2 times every 2months with a little neck stiffness 4/10 prior to this accident, now there is constant neck pain occurring every day 6-7/10 with neck stiffness 6-7/10 and decreased ROM all interfering with sleep (wakes the patient 1-3x/night for about 3 nights/week).

Low back pain was 4-5/10 occurring once per 1 1/2 weeks or so prior to this accident, since this accident it has been 6-8/10 occurring daily and worse with bending and moving, about 1-2 times per week wakes the patient at night.

Sleep interference possibly linked to pre-menopausal symptoms (sleeping fine for 2-3 months, then having restless sleep for about 3-4 weeks)prior to this accident. Since this accident, the patient is now awoken at least three times per week due to various physical pains.

Short term memory occurred prior to this accident only a little bit when not sleeping well, but is now worse after this accident in that the patient forgets where she puts things, can forget what she was going to say, and is getting progressively worse after this accident.

Difficult concentrating especially when headache occurred approximately once every two months, but since this accident now interferes with the patient's work daily.

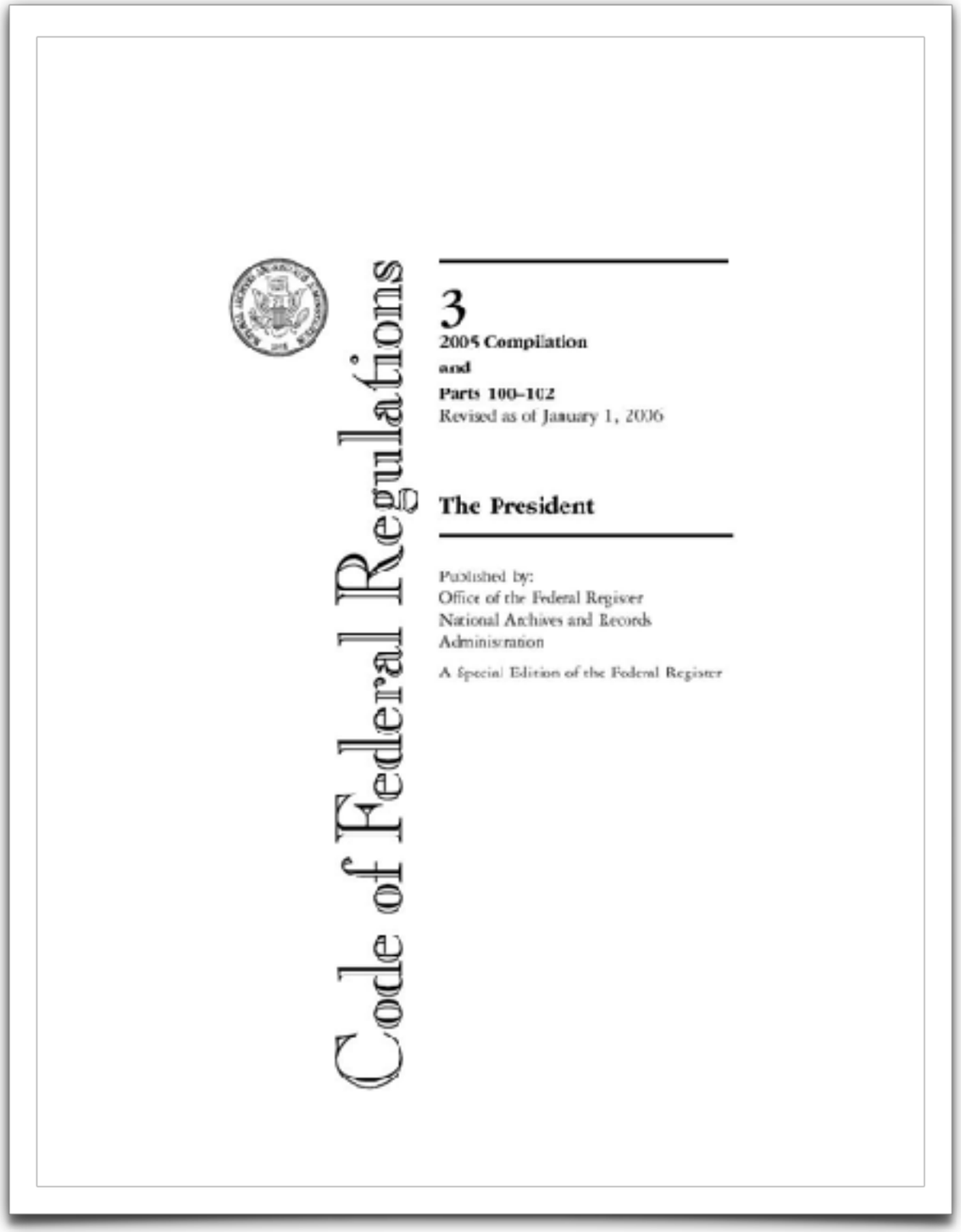
c) **New complaints** resulting from this accident -

This patient has the following symptoms which only occurred after and as a result of this accident:

More privacy to formalize

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification
2. Imagery
3. Geospatial information
4. Time and time series
5. Genomic information
6. Narrative text

Modeling of administrative controls.
Formalizing use limitations.



People who model re-identification risk take into account the ability, resources and motivation of the data intruder.

General public — anyone who has access to the data.

Expert — A computer scientist skilled in re-identification.

Insider — A member of the organization that produced the dataset

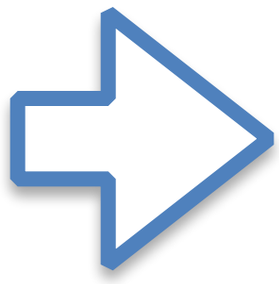
Insider recipient — A member of the organization that received the data and has more background information than the general public.

Information broker — An organization that systematically collects both identified and de-identified information to re-identify.

Nosy Neighbor — Friend or family member with specific info.

Cryptography's success required moving beyond perfect secrecy.

Diffie-Hellman
RSA
DES & 3DES
Certificates & PKI
PGP
S/MIME



Key Escrow
Identity-Based Encryption
Password reset by email
Secure web mail
Security questions

More privacy to formalize

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification
2. Imagery
3. Geospatial information
4. Time and time series
5. Genomic information
6. Narrative text
7. Attacks and controls

TIME'S UP!

More privacy to formalize

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification
2. Imagery
3. Geospatial information
4. Time and time series
5. Genomic information
6. Narrative text
7. Attacks and controls

Thank you.

- These slides available at: https://simson.net/ref/2017/2017-05-23_Formal_Privacy.pdf