Confidentiality Protection in AI

Xiao-Li Meng
(Panel for AI Day for Federal Statistics, May 2, 2024)

- What does *that* mean?
  - Confidentiality protection of the data used by AI?
  - Confidentiality protection of the users of AI?
  - Confidentiality protection of the AI technologies?

You

Confidentiality Protection in AI

What is the meaning of "Confidentiality Protection in AI"?

ChatGPT

"Confidentiality Protection in AI" generally refers to the measures and techniques employed to ensure that sensitive data used or generated by artificial intelligence (AI) systems is kept secure from unauthorized access, use, or disclosure. This involves implementing various safeguards, such as encryption, access controls, and data anonymization, to protect personal and proprietary information processed by AI technologies.

This concept is crucial because AI systems often handle large volumes of data, some of which may be sensitive or private. Ensuring confidentiality helps in maintaining user trust and complying with legal and ethical standards, such as the General Data Protection Regulation (GDPR) in Europe or other privacy laws worldwide.

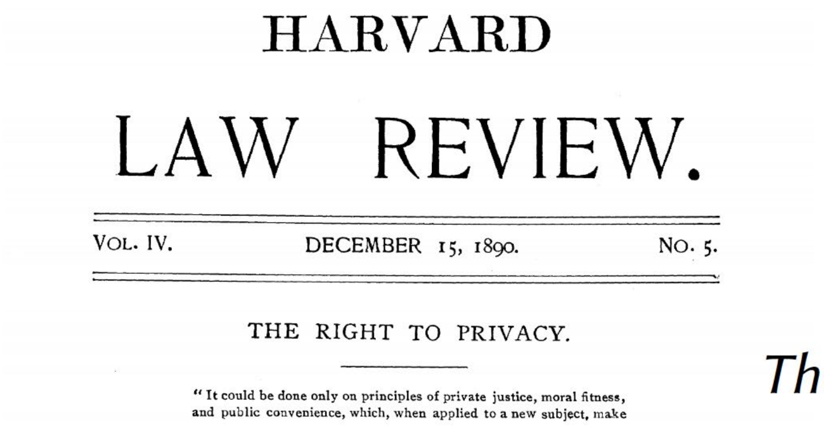
ME **You**

What is the meaning of "Privacy Protection in AI"?

ChatGPT

"Privacy Protection in AI" refers to the practices and methodologies designed to safeguard individual privacy when using artificial intelligence (AI) systems. This concept is crucial in the design, development, and deployment of AI technologies, ensuring that they respect user privacy by:

1. **Data Minimization**: Limiting the collection of personal data to what is directly relevant and necessary to accomplish a specified purpose.
2. **Anonymization**: Removing personally identifiable information from datasets to prevent the identification of individuals, making it harder to link data back to an individual.
3. **Secure Data Storage and Transfer**: Implementing secure protocols for data storage and transfer to protect data from unauthorized access or breaches.
4. **Transparency**: Providing clear information to users about what data is collected, how it is used, and who it is shared with.
5. **Consent Management**: Ensuring that users provide informed consent for the collection and use of their data.
6. **Regulatory Compliance**: Adhering to relevant privacy laws and regulations, such as GDPR, which impose strict requirements on data handling.


HARVARD

LAW REVIEW.

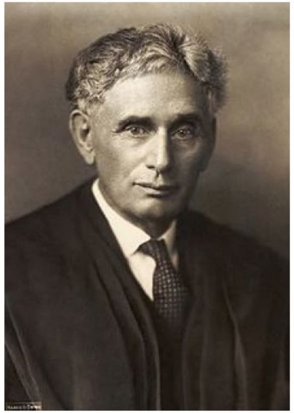
VOL. IV. DECEMBER 15, 1890. NO. 5.

THE RIGHT TO PRIVACY.

"It could be done only on principles of private justice, moral fitness, and public convenience, which, when applied to a new subject, make

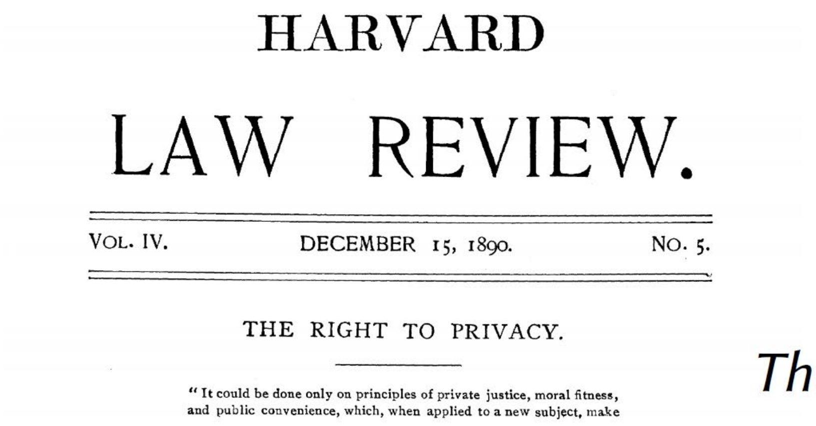
The right to be **let alone.**

Protect my privacy in the 1st place.



Samuel D. Warren II

Louis Brandeis



HARVARD

LAW REVIEW.

VOL. IV. DECEMBER 15, 1890. NO. 5.

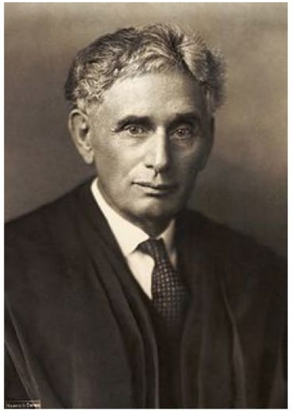
THE RIGHT TO PRIVACY.

"It could be done only on principles of private justice, moral fitness, and public convenience, which, when applied to a new subject, make

The right to be **let alone.**

Protect my privacy in the 1st place.

Samuel D. Warren II



Louis Brandeis

Over 125 years later (GDPR, 2016):

The right to be **forgotten.**

Protect my privacy in the 2nd place.

***HDSR* SPECIAL ISSUE**

Future Shock: Grappling With the Generative AI Revolution

OPEN CALL is now closed

The submission deadline for this special issue on generative AI, **launching May 2024**, has passed. We encourage authors to consider submitting their proposals related to this topic through the regular channel.

Click for more details on how to submit

As an open access platform of the Harvard Data Science Initiative, *Harvard Data Science Review* (*HDSR*) features **foundational thinking, research milestones, educational innovations, and major applications,** with a primary emphasis on reproducibility, replicability, and readability. We aim to publish content that helps define and shape data science as a scientifically rigorous and globally impactful multidisciplinary field based on the principled and purposed production, processing, parsing, and analysis of data. By uniting the strengths of a premier research journal, a cutting-edge educational publication, and a popular magazine, *HDSR* provides a crossroads at which fundamental data science research and education intersect directly with societally-important applications from industry, governments, NGOs, and others. By disseminating inspiring, informative, and intriguing articles and media materials, *HDSR* aspires to be a global forum on ***everything data science and data science for everyone.***

Future Shock:

Did the Generative AI Revolution Trigger an International AI Governance Crisis?

Professor David Leslie

Professor of Ethics, Technology & Society, QMUL

Director of Ethics & Responsible Innovation Research, Turing

What is future shock?

Alvin Toffler coined the term "future shock" to capture the **widespread societal dislocation** effected by the rapid advent of the digital revolution.

Future shock describes the dizzying disorientation brought on by **the premature arrival of the future**.

Image credit: Angela

Did generative AI really trigger future shock?

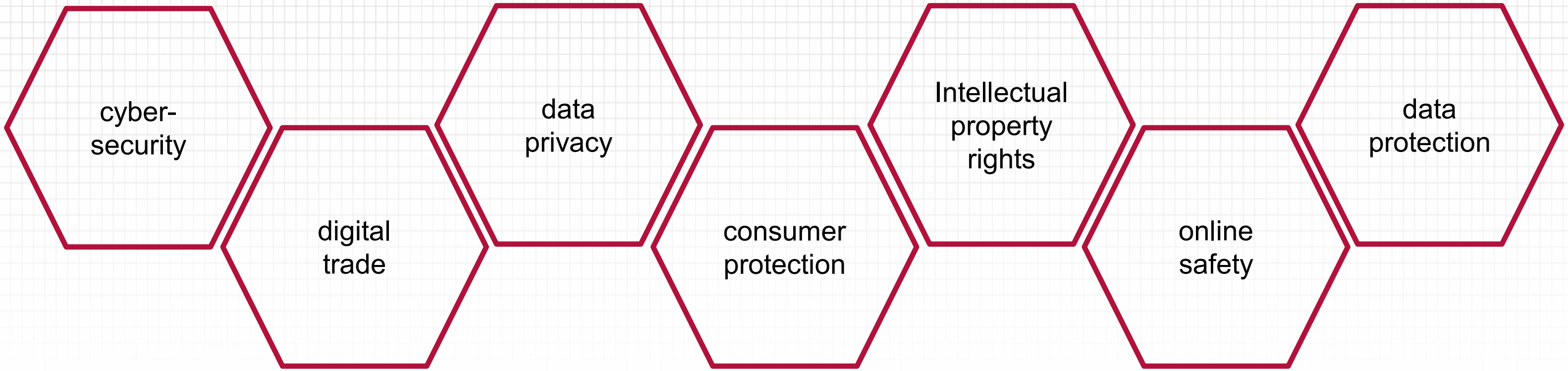
Despite the sudden rise of commercial generative AI technologies causing shockwaves across the digital world, **the extent to which this triggered ‘future shock’ among AI policy and governance communities across the globe remains debatable**.

Several factors should have softened the blow.

Image credit: https://www.searchenginejournal.com/chatgpt-chrome-extensions/485594/

Did generative AI really trigger future shock?

Decades of debate, research, and policy development had yielded standards, good practice protocols, laws, and regulations that formed a robust conceptual basis upon which AI policy and governance communities could draw:



Measures existing before the generative AI ‘revolution’

Concerted efforts had been made to develop **standards**, **policies**, and **governance mechanisms** to ensure the **equitable** production and deployment of AI systems.

Regional treaties, including:

- African Union’s Convention on Cybersecurity and Data Protection (2014)
- Council of Europe’s Convention on Cybercrime (2001)
- Council of Europe’s Convention on Data Protection (1981)

National data protection and privacy laws:

Korea (2011), Japan (2015), Singapore (2012), China (2021), Indonesia (2022), Sri Lanka (2022), Malaysia (2010), the European Union (2018), Egypt (2020), South Africa (2013), Tunisia (2004), Botswana (2018), Ghana (2012), Kenya (2019), Mauritius (2017), Malawi (2019), Nigeria (2023), Tanzania (2022), Uganda (2019), New Zealand (2020), Australia (1988), Bahrain (2019), Qatar (2016), UAE (2021), Argentina (2000), Mexico (2010), Chile (1999), Colombia (2012), and Brazil (2019), among others.

But there are deficits with regard to the enforcement of existing regimes

(Leslie et al, Editorial for Policy Forum, HDSR, 2024)

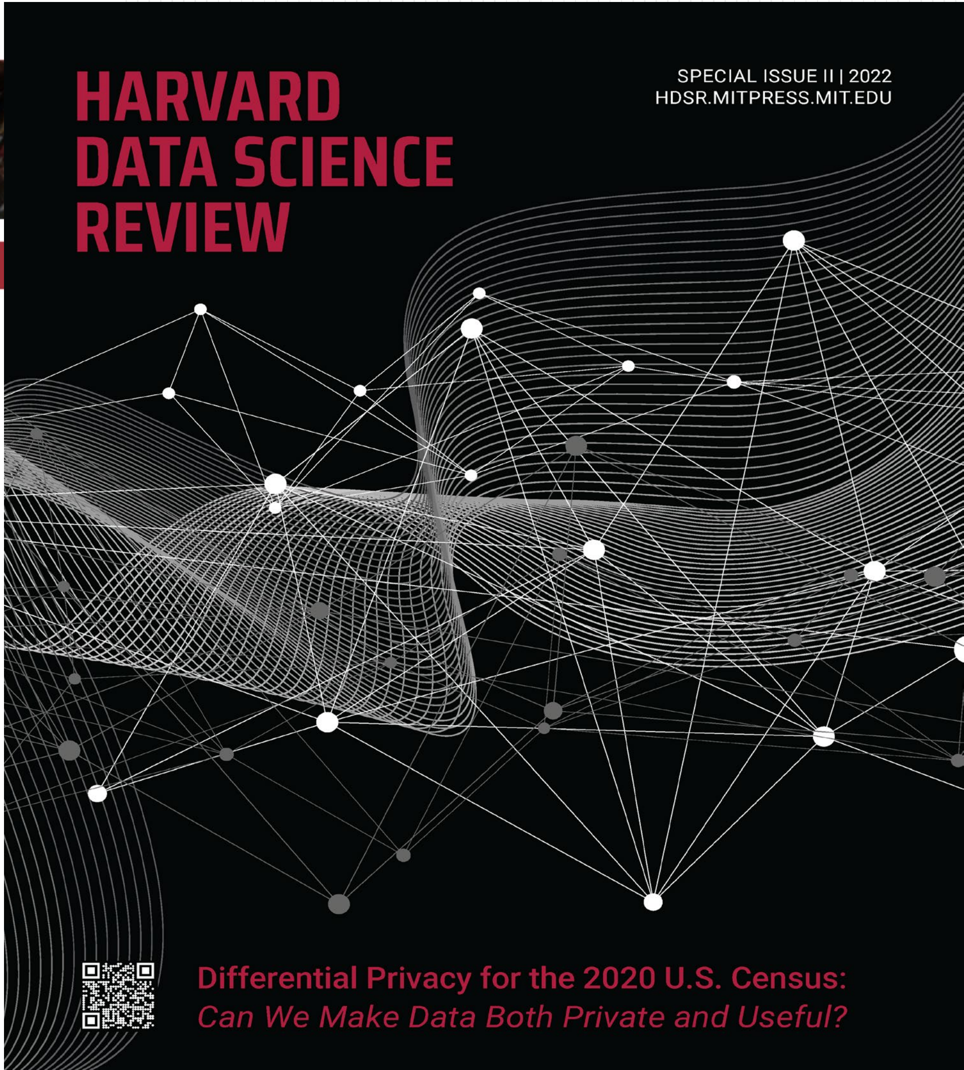
- Significant gaps have arisen over the past several years in the enforcement of existing digital- and data-related legal and regulatory regimes.
- These gaps—combined with deficits in the capacity of regulators to develop the skills and know-how needed to competently confront the novel governance challenges presented by the rapid deployment of largescale AI technologies—have created conditions for regulatory inaction and ineptitude.
- Disparities between legal protections related to digital and data rights and prevalent patterns of unimpeded bad behaviour that transgress such protections can be observed, for instance, in both data protection and cybersecurity law (Kohnke et al., 2021; Lynskey, 2023).

Data leakage and memorization risks (i.e. model intrinsic privacy violation risks)

(Leslie et al, 2024, Editorial for Policy Forum)

- The attitude of "the more data the better" and “scale is all you need” came quickly to prevail as the ‘rule of thumb’ among AI developers (Birhane et al., 2024; Bommasani et al., 2022; Kaplan et al., 2020)
- Unprecedented data scaling required the collection of a magnitude of web-scale data that far outstripped the capacity of human project teams to manually check data quality and source integrity, which has led to widespread risks of data poisoning, memorization, and leakage.
- Web-scale poisoning attacks can be launched inexpensively and with relative ease, making them ‘practical and realistic even for a low-resourced attacker’ (Carlini et al., 2023a).
- the presence of personally identifiable information (PII) (e.g. email addresses and phone numbers) and sensitive documents (e.g. personal medical records) in massive pretraining corpora can yield privacy leaks during model prompting and data extraction attacks (Carlini et al., 2021; Kaddour et al., 2023; Lukas et al., 2023; Mozes et al., 2023).
- Carlini et al. (2023b): ‘memorization in [FMs/LLMs] is more prevalent than previously believed and will likely get worse as models continue to scale, at least without active mitigations’ (p.1).


HDSR
Search
Dashboard
Login or Signup
HOME
ISSUES
SECTIONS
COLUMNS
COLLECTIONS
PODCAST
SUBMIT
ABOUT
MASTHEAD
Differential Privacy for the 2020 U.S. Census: Can We Make Data Both Private and Useful?
Special Issue 2
FROM THE EDITORS
Harnessing the Known Unknowns: Differential Privacy and the 2020 Census
by Ruobin Gong, Erica L. Groshen, and Salil Vadhan
Published: Jun 24, 2022
Special Issue 2: Differential Privacy for the 2020 U.S. Census
CENSUS: IMPORTANCE, HISTORY, AND TECHNICAL CHANGES
Coming to Our Census: How Social Statistics Underpin Our Democracy (and Republic)
by Teresa A. Sullivan
Published: Jan 31, 2020
CONNECTIONS
Disclosure Protection in the Context of Statistical Agency Operations: Data Quality and Related Constraints
by John L. Eltinge
Published: Jun 24, 2022


HARVARD DATA SCIENCE REVIEW
SPECIAL ISSUE II | 2022
HDSR.MITPRESS.MIT.EDU
Differential Privacy for the 2020 U.S. Census:
Can We Make Data Both Private and Useful?

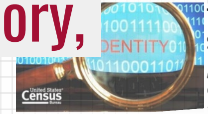
Differential Privacy for the 2020 U.S. Census: Can We Make Data Both Private and Useful?

1. From the Editors

Harnessing the Known Unknowns: Differential Privacy and the 2020 Census

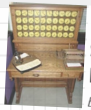
by Ruobin Gong, Erica L. Groshen, and Salil Vadhan

2. Census: Importance, History, and Technical Changes



The 2020 Census Disclosure Avoidance System TopDown Algorithm

by John Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kif and 6 more



Coming to Our Census: How Social Statistics Underpin Our Democracy (and Republic)

by Teresa A. Sullivan

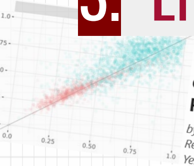
Implementing Differential Privacy: Seven Lessons From the 2020 United States Census

by Michael B. Hawes

Disclosure Protection in the Context of Statistical Agency Operations: Data Quality and Related Constraints

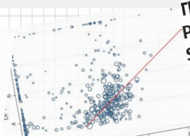
by John Eltinge

3. Empirical Evaluations



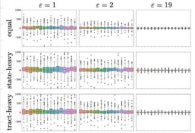
Assessing the Impact of Differential Privacy on Measures of Population and Racial Residential Segregation

by Brian Asquith, Brad Hershbein, Tracy Kugler, Shane Reed, Steven Ruggles, Jonathan Yesiltepe



The Effect of Differentially Private Noise Injection on Sampling Efficiency and Funding Allocations: Evidence From the 1940 Census

by Quentin Brummet, Edward Mulrow, and Kirk Wolter



Private Numbers in Public Policy: Census, Differential Privacy, and Redistricting

by Aloni Cohen, Moon Duchin, JN Matthews, and Bhushan Suwal

4. Commentary and Critique

Transparent Privacy Is Principled Privacy

by Ruobin Gong

What Will It Take to Get to Acceptable Privacy-Accuracy Combinations?

by Ori Heffetz

5. Broader Perspective

A Chronicle of the Application of Differential Privacy to the 2020 Census

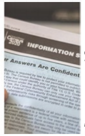
eph Hotz and Joseph Salvo

Differential Privacy and Social Science: An Urgent Puzzle

by Daniel L. Oberski and Frauke Kreuter

Disclosure Avoidance and the 2020 Census: What Do Researchers Need to Know?

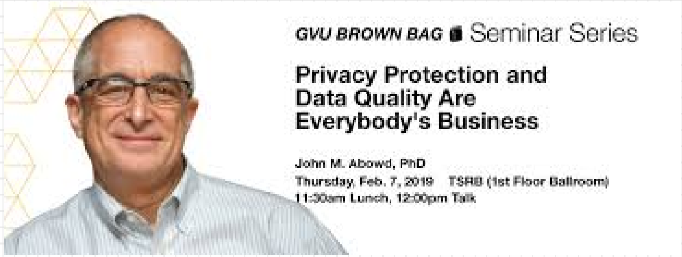
by Erica L. Groshen and Daniel Goroff



Differential Perspectives: Epistemic Disconnects Surrounding the U.S. Census Bureau's Use of Differential Privacy

by danah boyd and Jayshree Sarathy

Differential Privacy for 2020 Census



- ϵ-DP for a random map $R(D)$: If for all $\boldsymbol{S} \subset \boldsymbol{\Omega}$

$$\frac{\mathbf{Pr}\left(\boldsymbol{R}\left(\boldsymbol{D}_{\{-\mathbf{1}\}}\right) \in \boldsymbol{S}\right)}{\mathbf{Pr}(\boldsymbol{R}(\boldsymbol{D}) \in \boldsymbol{S})} \leq \boldsymbol{e}^{\epsilon}$$

where $\boldsymbol{D}_{\{-\mathbf{1}\}}$ and $\boldsymbol{D}$ differ by one individual.

ϵ – privacy loss budget: reducing ϵ by 1 ≈ reducing 1.5 questions in a 20-question game

Protecting Individua Privacy against All Adversaries

– Is It possible?

Theorem (Bailie, Gong & Meng, 2023)

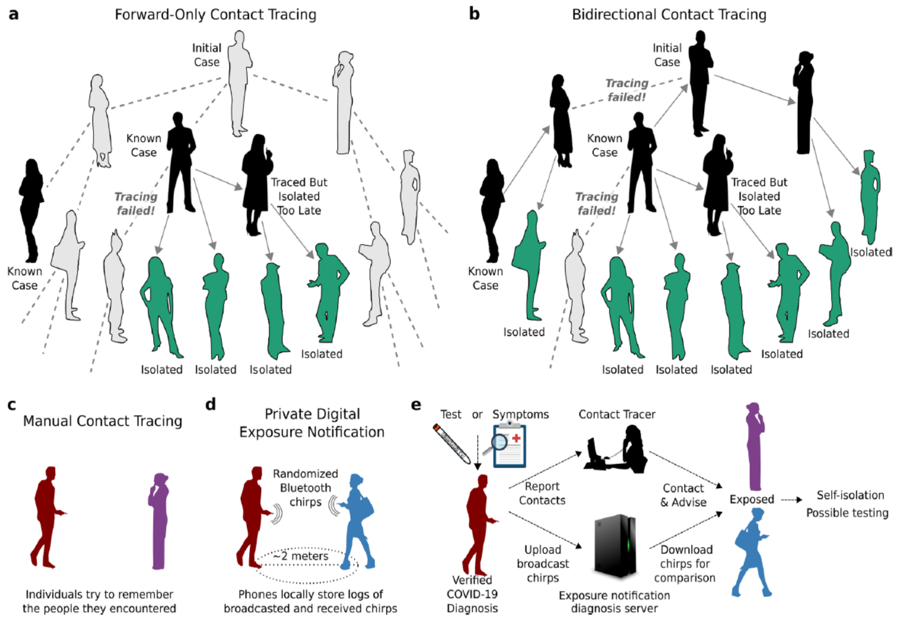
A random map M delivers ϵ-DP under Hamming distance if and only if for every prior π on $\mathcal{D}$, every sub-σ field $\mathcal{F}$ of the corresponding full σ-field $\sigma_\pi(\mathcal{X})$, every $B \in \mathscr{B}\left(\mathbb{R}^d\right)$, every i, and every $A \in \mathscr{B}(\Theta_i)$, where Θ_i is the state space of x_i, we have

$$e^{-c_i\epsilon}\pi(X_i \in A \mid \mathcal{F}) \leq \Pr\left(X_i \in A \mid M \in B; \mathcal{F}\right) \leq e^{c_i\epsilon}\pi(x_i \in A \mid \mathcal{F}), \tag{1}$$

where $\pi(x_i|\mathcal{F})$ is the marginal prior for X_i (conditional on $\mathcal{F}$), $\Pr$ is the marginal posterior for X_i, and c_i is the size of the minimal information chamber (MIC) for X_i.

- $MIC = C_{-i} \cup \{X_i\}$: $C_{-i} \subset \boldsymbol{X}_{-i}$ is the *Markov boundary* for X_i, that is, the smallest subset of $\boldsymbol{X}_{-i}$ such that
$$\pi(X_i|\boldsymbol{X}_{-i}, \mathcal{F}) = \pi(X_i|C_{-i}, \mathcal{F}).$$
- MIC is the X_i's "information family" – knowing any one of them will provide information about X_i, in addition to public knowledge coded into $\mathcal{F}$.

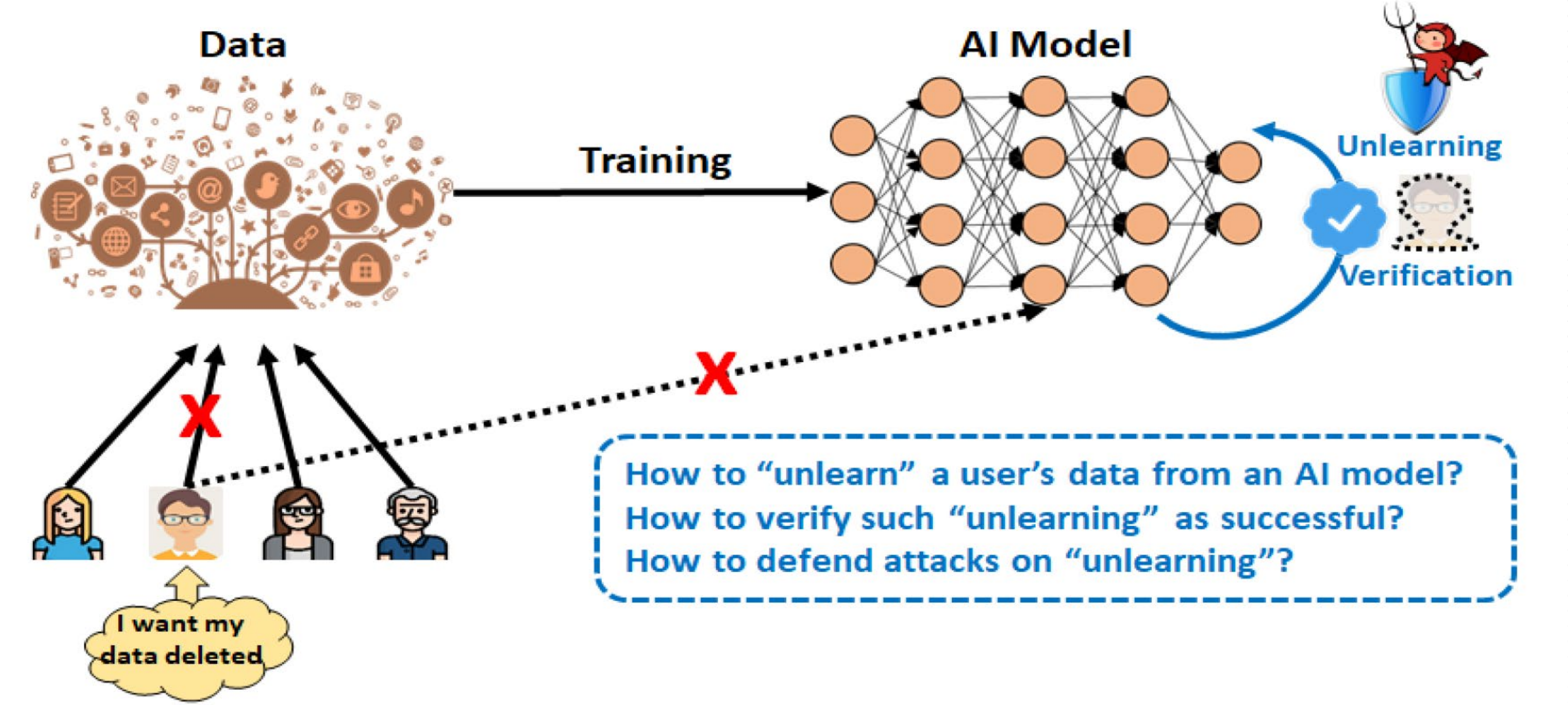
Information spreads like a virus — we need to quarantine not only the infected individual but also everyone they've come into contact with.



But assessing dependence among individuals requires context and statistical modeling.

Image from Bradshaw, W.J., Alley, E.C., Huggins, J.H. et al. Bidirectional contact tracing could dramatically improve COVID-19 control. Nat Commun 12, 232 (2021). https://doi.org/10.1038/s41467-020-20325-7

The dependence issue is critical for *machine unlearning*



Qu, Yuan, Ding, Ni, Rakotoarivelo, and Smith (2023) Learn to unlearn: A survey on machine unlearning.

- Stigler's Law of Privacy:
  The only way to ensure privacy is to make sure that no one cares about you.

- Stigler's Law of Eponymy:
  No scientific discovery is named after its original discoverer.