Privatsphäre und Datenschutz in Online-Experimenten

Max R. P. Grossmann

25. September 2021

An wen richtet sich dieser Vortrag?

- An Experimentator:innen (zur Ermutigung);
- an Labormanager:innen (zur Erleuchtung);
- an jene, die sich nicht so gut mit Datenschutz auskennen, aber darüber reden (zur Reflektion).

Datenschutz und „Datenschutz“—Zwei Konzepte

- Wikipedia beschreibt Datenschutz als „Schutz vor missbräuchlicher Datenverarbeitung, Schutz des Rechts auf informationelle Selbstbestimmung, Schutz des Persönlichkeitsrechts bei der Datenverarbeitung und auch Schutz der Privatsphäre verstanden. Datenschutz wird häufig als Recht verstanden, dass jeder Mensch grundsätzlich selbst darüber entscheiden darf, wem wann welche seiner persönlichen Daten zugänglich sein sollen“.
- „Datenschutz“ ist der Datenschutz-Cargo-Kult (vergleich Feynman, 1974): Eine auf Verhinderung abzielende Taktik von Bürokraten, die sich kaum mit der geltenden Rechtslage auseinander setzen wollen und/oder ihr Unterpfand—das der Privatsphäre—verkennen.

Die Reflektion: Spielt Datenschutz überhaupt eine Rolle? Kann man unter Einbezug von Technikexpert:innen eine saubere und datensparsame Lösung finden, anstatt zu bremsen?

Warum sind Datenschutz und Privatsphäre für uns wichtig?

- Fehlende Privatsphäre als **Messproblem**:
- Der experimentalökonomische Default ist, dass Entscheidungen *anonym* getroffen werden. Ein Mangel an Privatsphäre kann ein Confound sein; siehe die Forschung zu Chilling Effects.

> *Wie Schauspieler auf der Bühne, so Goffman, kann der Einzelne eine Rolle nur für einen angemessenen Zeitraum spielen, und niemand kann unbegrenzt und ohne Abwechslung die Vielfalt der Rollen spielen, die das Leben verlangt. Es muss Momente „abseits der Bühne“ geben, in denen der Einzelne „er selbst“ sein kann. (Westin, 1968, übers.)*

- Wer „echtes“ Verhalten messen möchte, muss dies auch in den Rahmenbedingungen einrichten.
- Wichtig: Fehlende Privatsphäre kann verschiedene Treatments in verschiedener Weise und Intensität betreffen.

Warum sind Datenschutz und Privatsphäre für uns wichtig?

- Fehlende Privatsphäre als **ethisches Problem**:
- Die Offenbarung privater Daten kann dem Individuum drastische Sonderopfer abverlangen. Das ist offensichtlich.
- Es ist moralisch erforderlich, dass das Experiment nach der Teilnahme beendet ist. Experimentator:innen sind hierfür verantwortlich; sie haften (ethisch und ggf. rechtlich) für alle Schäden, die Versuchspersonen anheim kommen.
- Selbst wenn die Erhebung geschickt gestaltet ist, können Versuchspersonen ggf. deanonymisiert werden. (Wichtigstes?) Beispiel: Das innere Produkt von demografischen Variablen; siehe die „Experimente“ von Bill Weld mit Gesundheitsdaten in Massachusetts (Kearns & Roth, 2019).
- Lösung(?): Weniger demografische Daten erheben; besonders gefährlich sind objektiv nachprüfbare oder anderswo erfasste Variablen. Oder: Bereinigung (De-Individualisierung).

Wann sind Daten anonym?

- Daten sind dann anonym, wenn sie nicht personenbezogen sind.
- Nur personenbezogene Daten fallen unter den Datenschutz.
- Ein guter (aber unverbindlicher!) Maßstab ist der folgende: *Kann ein Gericht die Beschlagnahme von Daten anordnen, und ist dann durch irgendeine Methode eine Zuordnung auch nur eines Datums zu einer natürlichen Person möglich, so enthält der Datensatz personenbezogene Daten.*
- Wichtig: Ein Gericht kann eine Beschlagnahme bei mehreren Personen anordnen.
- Wichtig: Daten, die (sicher) gelöscht wurden, erlauben natürlich keine Zuordnung durch irgendeine Methode mehr!
- Aber: Der datenschutzrechtliche Mindeststandard ist nicht mit einem ethischen Standard gleichzusetzen.

Zwei Grundprinzipien...

1. Daten, die nicht erhoben werden, oder die nicht mehr existieren, können nicht missbraucht werden (Datensparsamkeit).
2. Für Daten, die bereinigt werden müssen, gilt: Automatische Methoden sind manuellen Methoden stets vorzuziehen.

. . . und drei Bewusstseinsstufen

Stufe 0 Ich kümmere mich nicht drum. YOLO/Vergelt's Gott!

Stufe 1 Eine geschickte Erhebung löst alle Probleme.

Stufe 2 Eine geschickte Erhebung löst nicht alle Probleme!

Auszahlungen in Online-Experimenten

- In Laborexperimenten...
  - wird Bargeld gezahlt.
  - werden die Unterlagen kurz aufgehoben (oft: 1 Jahr).
  - sind diese Unterlagen analog und nicht leicht durchsuchbar.
  - sind diese Unterlagen oft uneindeutig.
- In Online-Experimenten...
  - werden Zahlungen durch die *überwachtesten Kanäle der Menschheit* geleitet (SEPA, PayPal, etc.).
  - werden Zahlungsunterlagen *für Jahrzehnte* aufbewahrt.
  - sind diese Unterlagen *leicht durchsuchbar* und können mit Verhaltensdaten gematcht werden.
  - sind diese Unterlagen nicht widerlegbar. Sie haben Beweiskraft.

Beispiel für Denken auf Stufe 0

- Ich erhebe einfach die IBAN im Datensatz! In oTree ganz leicht:
- `page_sequence = [..., KontoDetails]`
- `IBAN = models.StringField()`
- Bitte nich… erfordert manuellen Cleanup und erlaubt Missbrauch durch Experimentator:in oder andere.
- Sollte an allen Laboren streng verboten sein. Diese „Methode“ entbehrt jeglicher Rechtfertigung.
- Hinweis: Den Laborregeln kommt auch eine kommunikative Wirkung zu. Auch wenn ein *de jure* Verbot bereits besteht, sollte auf diese konkrete Anwendung explizit hingewiesen werden.

Schon besser: Stufe 1: AnonPay mit CPY und TDY

- Mein eigenes Projekt (AnonPay) erlaubt es, IBANs und andere Daten so zu erheben, dass sie dann getrennt vom Verhaltensdatensatz gespeichert werden.
- Vollkommen freie Software, keine Zitationspflicht. Verfügbar für oTree und z-Tree. Völlig automatisch. Keine Fallstricke.
- Mehr Infos: https://gitlab.com/gr0ssmann/AnonPay
- Hinweis: Die Module CPY und TDY erfüllen den Mindeststandard der getrennten Erhebung.
- Prinzip 2 erfüllt; Stufe 1 erreicht!

Die Erleuchtung: Experimentelle Software kann vieles—eine kreative Nutzung ihrer Features ist hier besonders lohnend.

Wenn AnonPay nicht verfügbar ist...

- Experimentator:in generiert *kurzlebigen Code* C_i, zeigt ihn der Versuchsperson und speichert ihn mit π_i
- Versuchsperson geht auf Laborwebsite und gibt C_i mit IBAN, P_i, ein
- Experimentator:in sendet $[(\pi_1, \ldots, \pi_n), (C_1, \ldots, C_n)]$ an das Labor
- Labor nutzt sein Wissen über $\Theta = [(P_1, \ldots, P_n), (C_1, \ldots, C_n)]$ um $p = \text{rnd}([(P_1, \ldots, P_n), (\pi_1, \ldots, \pi_n)]) \perp C_i$ zu erstellen, welches an Experimentator:in geht
- Eine Methode mit einigen Nachteilen (arbeitsintensiv), löst nicht alles
- Aber: Gute Kredibilität, kann überall angewandt werden
- Hinweis: Die Verhaltensdaten sind erst dann anonym, wenn das Labor Θ löscht $\rightarrow$ Matching sonst bei Kollusion möglich

Der Vorbote von Stufe 2...

- **Die Crux**: Für eindeutige Auszahlungsbeträge ist stets eine Deanonymisierung durch Matching möglich (Grossmann, 2021; Fleder & Shah, 2020).
- Bei seltenen Auszahlungsbeträgen ist sie meist möglich. Eine gute Privatsphäre bieten nur Experimente mit identischen Auszahlungen!

Stufe 1.8??? — NUN und SPA in AnonPay

- AnonPay hat auch die optionalen Module NUN und SPA.
- Diese machen Auszahlungsbeträge uneindeutig. NUN macht dies grundsätzlich, und SPA für ein besonders geschütztes Attribut (z.B. Art. 9 DSGVO).
- Diese Algorithmen sind heuristisch und ändern die Auszahlungsbeträge; damit gehen Einbußen in der Anreizkompatibilität einher, insbesondere bei Spielen mit großen Diskrepanzen im Auszahlungsbetrag.
- Macht Matching viel schwieriger, aber „löst“ das Problem nicht.
- Wenig Nutzung dieser Algorithmen (10x?). CPY und TDY wurden hingegen schon tausende Male eingesetzt. Aufgrund Punkt 3 keine positive Rezeption durch Ökonomen.

Varieté: Technische Fundamentals

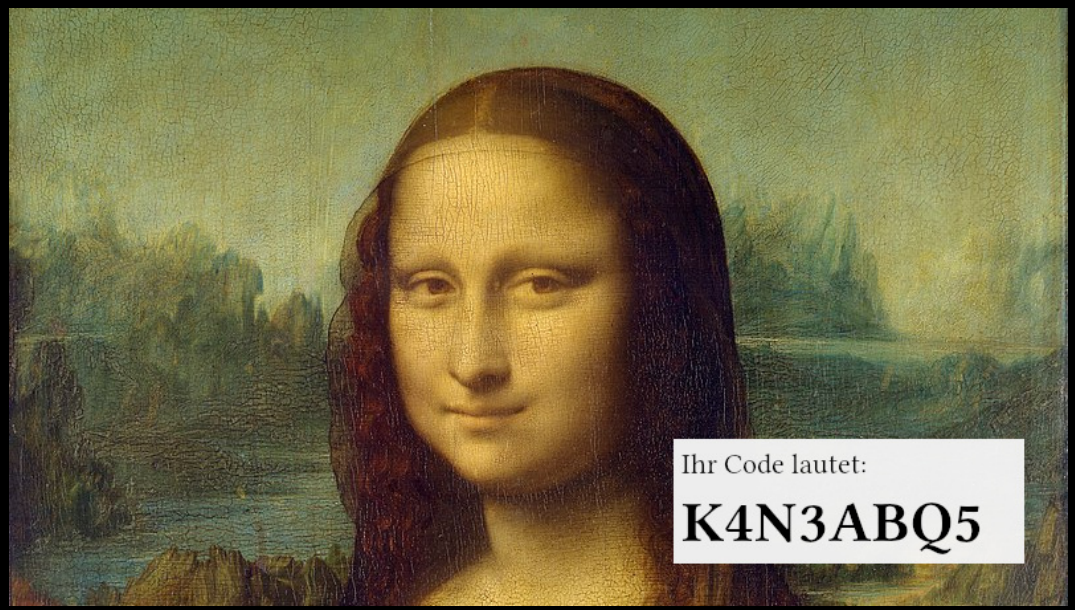
- In meiner Erfahrung tendieren Informatiker:innen dazu, lieber zu viel zu loggen als zu wenig. Dafür gibt es aber in der Regel keine Rechtsgrundlage und auch kein Erfordernis.
- Jegliche Software ist so einzustellen, dass möglichst wenig, und auch nur ganz kurz, geloggt wird. Das Loggen von IPs ist ohne Einwilligung so tabu wie eine Durchführung über HTTP ohne Transport Layer Security! Cookies rechtskonform einsetzen.
- Es bietet sich an, von Anbieter:innen wie Heroku und Cloudprovidern, die fremdem Recht unterliegen, unabhängig zu werden. Ich helfe beim Aufsetzen von oTree-, z-Tree-unleashed- und LimeSurvey-Servern.
- Beim Löschen sensibler Daten unter unixartigen Betriebssystemen kann `shred -u` genutzt werden; unter Windows z.B. die Gutmann (1996) Methode.
- Wenig Cloud nutzen; statt E-Mail lieber Signal.

Varieté: Multi-part-Experimente

- Wie kann man Daten aus (z.B.) Diary-Surveys miteinander matchen?
- Oft genutzt: Methoden, die stabile Codes unter Verwendung des Familienstammbuchs generieren.
- Experimentator:in neulich zu mir: „Ich nutz' das, die Codes sind IMMER unique!!!!“
- Und genau da liegt das Problem (d.h. Matchbarkeit über Studiengrenzen hinweg). Wie wäre es mit dem Folgenden?

Varieté: Multi-part-Experimente

„Bitte nehmen Sie Ihr Smartphone und fotografieren Sie das Bild unten. Bitte löschen Sie das Bild nicht; in ein paar Wochen fragen wir Sie wieder nach dem Code.“



Zusätzlich: Speicherung eines Cookies.

Ich wette: Fast alle Versuchspersonen können das Bild dann in der Smartphone-Galerie wiederfinden (oder haben den Cookie); und wenn nicht, sollten wir das mit den Experimenten vielleicht lassen. . .

Varieté: Andere nützliche Werkzeuge

- **JavaScript**: Zur Durchführung (sensibler) Operationen auf dem Rechner der Versuchsperson und der Übermittlung nur un(ter)kritischen Daten an den Server.
- **Datentreuhänder:innen**: Wenn sie wirklich unabhängig sind.
- **Hashing**: Ermöglicht Commitments über Informationen und ihre selektive Offenbarung.
- **Differential Privacy/Homomorphe Verschlüsselung**: Für das Veröffentlichen „bereinigter“ Datensätze. Löst (u.a.) das Problem mit dem inneren Produkt demografischer Variablen.

Die Ermutigung: (Fast) jedes unserer Desiderata ist mit *Kreativität* sauber und rechtlich und ethisch einwandfrei umzusetzen.

Literatur

- Feynman, R. P. (1974). Cargo cult science. *Engineering and Science*, 37(7), 10-13.
- Fleder, M., & Shah, D. (2020). I Know What You Bought At Chipotle for $9.81 by Solving A Linear Inverse Problem. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(3), 1-17.
- Grossmann, M. R. P. (2021). AnonPay: Enhancing participant privacy in online experiments. Work in progress.
- Gutmann, P. (1996). Secure deletion of data from magnetic and solid-state memory. *Proceedings of the Sixth USENIX Security Symposium, San Jose, CA* (Vol. 14, pp. 77-89).
- Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design.* Oxford University Press.