Lecture 2: Anonymization / Deanonymization

CMSC 25910

Spring 2024

The University of Chicago

THE UNIVERSITY OF CHICAGO

“…a tension that shakes a foundational belief about data privacy: Data can be either useful or perfectly anonymous but never both.” – Paul Ohm

Historical Conceptualizations of Anonymization and Personal Data

Personally Identifiable Information **(PII)**

- Also termed **“personal data”**
- 2010 NIST Special Publication 800-122 Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)
- General Data Protection Regulation (GDPR) in the EU

NIST 800-122 Definitions

- “**PII** is —any information about an individual maintained by an agency, including (1) any information that **can be used to distinguish or trace an individual’s identity, such as name, social security number, date and place of birth, mother‘s maiden name, or biometric records**; and (2) any other information that is **linked or linkable to an individual, such as medical, educational, financial, and employment information**.”

NIST 800-122 PII Examples

- **Name**, such as full name, maiden name, mother‘s maiden name, or alias
- **Personal identification number**, such as social security number (SSN), passport number, driver‘s license number, taxpayer identification number, patient identification number, and financial account or credit card number
- **Address information**, such as street address or email address
- **Asset information**, such as Internet Protocol (IP) or Media Access Control (MAC) address or other host-specific persistent static identifier that consistently links to a particular person or small, well-defined group of people
- **Telephone numbers**, including mobile, business, and personal numbers
- **Personal characteristics**, including photographic image (especially of face or other distinguishing characteristic), x-rays, fingerprints, or other biometric image or template data (e.g., retina scan, voice signature, facial geometry)
- **Information identifying personally owned property**, such as vehicle registration number or title number and related information
- **Information about an individual that is linked or linkable to one of the above** (e.g., date of birth, place of birth, race, religion, weight, activities, geographical indicators, employment information, medical information, education information, financial information).

NIST 800-122 Definitions

- "To **distinguish** an individual is to identify an individual. Some examples of information that could identify an individual include, but are not limited to, name, passport number, social security number, or biometric data. In contrast, a list containing only credit scores without any additional information concerning the individuals to whom they relate does not provide sufficient information to distinguish a specific individual."
- "To **trace** an individual is to process sufficient information to make a determination about a specific aspect of an individual's activities or status. For example, an audit log containing records of user actions could be used to trace an individual's activities."

NIST 800-122 Definitions

- **Linked information** is information about or related to an individual that is logically associated with other information about the individual. In contrast, linkable information is information about or related to an individual for which there is a possibility of logical association with other information about the individual. For example, if two databases contain different PII elements, then someone with access to both databases may be able to link the information from the two databases and identify individuals, as well as access additional information about or relating to the individuals. If the secondary information source is present on the same system or a closely-related system and does not have security controls that effectively segregate the information sources, then the data is considered linked. If the secondary information source is maintained more remotely, such as in an unrelated system within the organization, available in public records, or otherwise readily obtainable (e.g., internet search engine), then the data is considered linkable.

GDPR Definitions (Article 4)

- **‘personal data’** means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

GDPR Definitions (Article 4)

- **‘processing’** means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;
- **‘restriction of processing’** means the marking of stored personal data with the aim of limiting their processing in the future;

GDPR Definitions (Article 4)

- **‘pseudonymisation’** means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;

Example from UChicago IRB

4. * Do you anticipate that the research data will be transferred or transported at any time?

   ○ Yes ● No Clear

5. * Do you plan to store data on a server or cloud service?

   ● Yes ○ No Clear

   a. * Which server or cloud service do you plan to use?

   UChicago Box

6. * Will you collect any identifiers from the research participants (including names, addresses, Social Security Numbers, email and phone contact information, etc.)?

   ○ Yes ● No Clear

7. * What identifying information about research participants will be linked to the data?

   ☐ Data/specimens will be directly labeled with personal identifying information
   ☐ Data/specimens will be labeled with a code that the research team can link to personal identifying information through a crosswalk to the coding system
   ☐ Data/specimens will be labeled with a code but the research team will not have access to the crosswalk that connects the codes to participant identifiers
   ☑ Data/specimens will not be labeled with any identifying information and a coding system will not be used
   ☐ Other

8. If you will be using a coding system, who will have access to the crosswalk that links participant identifiers to the data/specimens and where will you store the crosswalk?

   Not applicable.

Models of Data-Release Stewardship

Scope of Releasing Data

- Release to third parties
- Release to the public
- Release to others within your organization
- Inadvertent release
    - Data breaches
    - Unintentional leakage / inference

Models of Data-Release Stewardship

- *(Note that Blase just made up the terms on this page)*
- **A Release-and-Forget Model:** Try to remove PII and otherwise "deidentify" data, but then provide unrestricted access (e.g., through publicly posting a dataset)
- **A Release-Under-Conditions Model:** Try to remove PII and otherwise "deidentify" data, but then provide restricted access to them (e.g., through data processing covered under contractual obligations and an approval process) and sometimes conditions upon the processing or the release of aggregate data
- **A Managed-Processing Model:** The data steward never releases the data, but will run computation for others and provide aggregate answers

Your Approaches to Redaction / Data Release in Assignment 1 Part A

Problem Setting

First Name	Last Name	Age	Occupation	ZIP Code	Location	Household Income (Dollars)	Number of Children
Blayre	Mercado	32	Miscellaneous Health Technologists a	19947	Georgetown, DE	42591	2
Loughlin	Villalobos	30	Other Entertainment Attendants and R	23882	Stony Creek, VA	64392	0
Elouise	Gentry	20	Software Developers	93635	Los Banos, CA	67919	1
Winnie	Dorsey	37	Underground Mining Machine Operato	25247	Hartford, WV	43184	1
Arielle	Camacho	51	Broadcast, Sound, and Lighting Techn	32448	Marianna, FL	27453	1
Tamara	Nolan	22	Miscellaneous Health Technologists a	73944	Hardesty, OK	54910	1
Harrisson	Howard	40	Explosives Workers, Ordnance Handli	72073	Humphrey, AR	87814	1
Andi	Sellers	43	Web and Digital Interface Designers	50071	Dows, IA	52848	1
Ferdinand	Solis	45	Other Transportation Workers	37343	Hixson, TN	91204	0
Micaiah	Maldonado	56	Athletes and Sports Competitors	44135	Cleveland, OH	62055	0
Thea	Pratt	22	Personal Service Managers, All Other	55401	Minneapolis, MN	72749	0
Aaran	Stuart	39	Commercial and Industrial Designers	78516	Alamo, TX	50006	0
Star	Moon	33	Other Educational Instruction and Libr	29161	Timmonsville, SC	47076	2
Waheed	Cantrell	29	Other Healthcare Practitioners and Te	55387	Waconia, MN	128212	0
Yvonne	Diaz	29	Other Metal Workers and Plastic Work	62521	Decatur, IL	44672	2
Iyvhn	Dennis	37	Healthcare Social Workers	62854	Kinmundy, IL	46856	0
Reimi	Nixon	27	Military Enlisted Tactical Operations	49307	Big Rapids, MI	40821	1
Enija	Hayden	38	Public Safety Telecommunicators	48768	Vassar, MI	73049	1
Daire	Mccall	50	Other Installation, Maintenance, and R	97865	Mount Vernon, OR	29877	0
Eva	Mcintosh	24	Floral Designers	13812	Nichols, NY	81443	5
Kyle-Jay	Levy	60	Other Transportation Workers	55041	Lake City, MN	101603	2
Lucy-Ann	Lindsey	21	Paramedics	49051	East Leroy, MI	63589	0
Manolis	Guerra	61	Earth Drillers, Except Oil and Gas	15734	Dixonville, PA	45705	1
Adhiya	Spence	18	Rehabilitation Counselors	20706	Lanham, MD	82522	2

Techniques You Applied (/51)

- Removing full names (48)
- Removing first names (0)
- Removing last names, keeping first (1)
- Replace names with initials (2)

- Let’s call this **deletion / suppression / omission**

Techniques You Applied (/51)

- Removing ZIP code (27)
- Removing occupations (3)
- Removing location (2)
- Removing age (0)
- Removing income (0)

- Let’s call this **deletion / suppression / omission**

Techniques You Applied (/51)

- Grouping age (32)
- Grouping income (20)
- Grouping number of children (11)
  - Replace number of children with binary “children/none” (4)
  - Create one group for 3+ children (2)
  - Create one group for 5+ children (2)
- Grouping occupation (1 with ChatGPT, but 5 thought about it)

- Let’s call this **binning**

Techniques You Applied (/51)

- Grouping outliers (a few in the context of # children)
- Removing ZIP code and location if fewer than 14 individuals represented (1)
- Removing rare combos of age and occupation (0)
- Removing only people in ZIP for their occupation (0)
- Examining correlations between columns (a few)

- Let’s call this **suppressing rare / infrequent data**

Techniques You Applied (/51)

- Replacing location with state (18)
- Removing parts of ZIP codes (3)
  - Kept first 3 digits (3)
- Replaced ZIP code with county using a library (1)

- Let’s call this **generalization**

Techniques You Applied (/51)

- Hashing location and ZIP (1, but beware!)
- Replacing ZIP codes with pseudonym (2)
- Replacing name, location, occupation with pseudonym (1)

- Let’s call this **pseudonymization** or **replacement**

Techniques You Applied (/51)

- Average numerical categories by demographic (0)

- Let’s call this **aggregation**

Techniques You Applied (/51)

- Shuffled rows (1)
- Added second, random occupation (0)
- +/- Gaussian noise to income (2)
- +/- to age (2)
    - randint(-2,2) (1)
- +/- to number of children (1)

- Let’s call this **perturbation**

Techniques You Applied (/51)

- Inferred gender from first name (2)
- Thought about inferring ethnicity from last name (1)
- Replaced location with approximate town/city size (1)

- Let’s call this **derived data**

General Techniques for Anonymization

Original Data

TABLE 1: Original (Nonanonymized) Data

Name	Race	Birth Date	Sex	ZIP Code	Complaint
Sean	Black	9/20/1965	Male	02141	Short of breath
Daniel	Black	2/14/1965	Male	02141	Chest pain
Kate	Black	10/23/1965	Female	02138	Painful eye
Marion	Black	8/24/1965	Female	02138	Wheezing
Helen	Black	11/7/1964	Female	02138	Aching joints
Reese	Black	12/1/1964	Female	02138	Chest pain
Forest	White	10/23/1964	Male	02138	Short of breath
Hilary	White	3/15/1965	Female	02139	Hypertension
Philip	White	8/13/1964	Male	02139	Aching joints
Jamie	White	5/5/1964	Male	02139	Fever
Sean	White	2/13/1967	Male	02138	Vomiting
Adrien	White	3/21/1967	Male	02138	Back pain

From Paul Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review Vol. 57, p. 1701, 2010.

Suppressing Data

- **Suppression:** Deleting or omitting data

TABLE 2: Suppressing Four Identifier Fields

Race	Complaint
Black	Short of breath
Black	Chest pain
Black	Painful eye
Black	Wheezing
Black	Aching joints
Black	Chest pain
White	Short of breath
White	Hypertension
White	Aching joints
White	Fever
White	Vomiting
White	Back pain

From Paul Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review Vol. 57, p. 1701, 2010.

Generalizing Data

- **Generalization:** Re-code data to be less granular

TABLE 3: Generalized

Race	Birth Year	Sex	ZIP Code*	Complaint
Black	1965	Male	021*	Short of breath
Black	1965	Male	021*	Chest pain
Black	1965	Female	021*	Painful eye
Black	1965	Female	021*	Wheezing
Black	1964	Female	021*	Aching joints
Black	1964	Female	021*	Chest pain
White	1964	Male	021*	Short of breath
White	1965	Female	021*	Hypertension
White	1964	Male	021*	Aching joints
White	1964	Male	021*	Fever
White	1967	Male	021*	Vomiting
White	1967	Male	021*	Back pain

From Paul Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review Vol. 57, p. 1701, 2010.

Aggregating Data

- **Aggregation:** Release summary data rather than raw data

TABLE 4: Aggregate Statistic

Men Short of Breath	2

From Paul Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review Vol. 57, p. 1701, 2010.

The Difficulty of Redaction

How Do You Find Personal Data?

- Example: Microsoft Presidio
  - https://microsoft.github.io/presidio/
- Example: Google's Cloud Data Loss Prevention (DLP) API
  - https://cloud.google.com/dlp/docs/infotypes-reference
- Amazon Macie for Amazon Web Services
  - https://docs.aws.amazon.com/macie/latest/user/what-is-macie.html
  - "Amazon Macie is a fully managed data security and data privacy service that uses machine learning and pattern matching to help you discover, monitor, and protect sensitive data in your AWS environment."
  - "Macie automates the discovery of sensitive data, such as personally identifiable information (PII) and financial data… Macie also provides you with an inventory of your S3 buckets, and it automatically evaluates and monitors those buckets for security and access control."

Google Cloud DLP

InfoType	Description
`ADVERTISING_ID`	Identifiers used by developers to track users for *advertising purposes*. These include Google Play Advertising IDs, Amazon Advertising IDs, Apple's identifierForAdvertising (IDFA), and Apple's identifierForVendor (IDFV).
`AGE`	An *age* measured in months or years.
`CREDIT_CARD_NUMBER`	A *credit card number* is 12 to 19 digits long. They are used for payment transactions globally.
`CREDIT_CARD_TRACK_NUMBER`	A *credit card track number* is a variable length alphanumeric string. It is used to store key cardholder information.
`DATE`	A *date*. This infoType includes most date formats, including the names of common world holidays. Note: Not recommended for use during latency sensitive operations.
`DATE_OF_BIRTH`	A *date of birth*. Note: Not recommended for use during latency sensitive operations.
`DOMAIN_NAME`	A *domain name* as defined by the DNS standard.
`EMAIL_ADDRESS`	An *email address* identifies the mailbox that emails are sent to or from. The maximum length of the domain name is 255 characters, and the maximum length of the local-part is 64 characters.
`ETHNIC_GROUP`	A person's *ethnic group*.

Google Cloud DLP

`MALE_NAME`	A common *male name*. Note: Not recommended for use during latency sensitive operations.
`MEDICAL_TERM`	Terms that commonly refer to a person's *medical condition or health*. Note: Not recommended for use during latency sensitive operations.
`ORGANIZATION_NAME`	A name of a *chain store, business or organization*. Note: Not recommended for use during latency sensitive operations.
`PASSPORT`	A *passport number* that matches passport numbers for the following countries: Australia, Canada, China, France, Germany, Japan, Korea, Mexico, The Netherlands, Poland, Singapore, Spain, Sweden, Taiwan, United Kingdom, and the United States.
`PERSON_NAME`	A full *person name*, which can include first names, middle names or initials, and last names. Note: Not recommended for use during latency sensitive operations.
`PHONE_NUMBER`	A *telephone number*.
`STREET_ADDRESS`	A *street address*. Note: Not recommended for use during latency sensitive operations.
`SWIFT_CODE`	A *SWIFT code* is the same as a Bank Identifier Code (BIC). It's a unique identification code for a particular bank. These codes are used when transferring money between banks, particularly for international wire transfers. Banks also use the codes for exchanging other messages.
`TIME`	A *timestamp* of a specific time of day.
`URL`	A *Uniform Resource Locator (URL)*.

Google Cloud DLP

Credentials and secrets

The infoType detectors in this section detect credentials and other secret data.

InfoType	Description
AUTH_TOKEN	An *authentication token* is a machine-readable way of determining whether a particular request has been authorized for a user. This detector currently identifies tokens that comply with OAuth or Bearer authentication.
AWS_CREDENTIALS	Amazon Web Services account access keys.
AZURE_AUTH_TOKEN	Microsoft Azure certificate credentials for application authentication.
BASIC_AUTH_HEADER	A *basic authentication header* is an HTTP header used to identify a user to a server. It is part of the HTTP specification in RFC 1945, section 11.
ENCRYPTION_KEY	An *encryption key* within configuration, code, or log text.
GCP_API_KEY	*Google Cloud API key*. An encrypted string that is used when calling Google Cloud APIs that don't need to access private user data.
GCP_CREDENTIALS	*Google Cloud service account credentials*. Credentials that can be used to authenticate with Google API client libraries and service accounts.
JSON_WEB_TOKEN	*JSON Web Token*. JSON Web Token in compact form. Represents a set of claims as a JSON object that is digitally signed using JSON Web Signature.
HTTP_COOKIE	An *HTTP cookie* is a standard way of storing data on a per website basis. This detector will find headers containing these cookies.
PASSWORD	Clear text *passwords* in configs, code, and other text.

Google Cloud DLP

US_PASSPORT	A *United States passport number*.
US_PREPARER_TAXPAYER_IDENTIFICATION_NUMBER	A *United States Preparer Taxpayer Identification Number (PTIN)* is an identification number that all paid tax return preparers must use on US federal tax returns or claims for refund submitted to the US Internal Revenue Service (IRS).
US_SOCIAL_SECURITY_NUMBER	A *United States Social Security number (SSN)* is a 9-digit number issued to US citizens, permanent residents, and temporary residents. This detector will not match against numbers with all zeroes in any digit group (that is, 000-##-####, ###-00-####, or ###-##-0000), against numbers with 666 in the first digit group, or against numbers whose first digit is 9.
US_STATE	A *United States state name*.
US_TOLLFREE_PHONE_NUMBER	A *US toll-free telephone number*.
US_VEHICLE_IDENTIFICATION_NUMBER	A *vehicle identification number (VIN)* is a unique 17-digit code assigned to every on-road motor vehicle.

Uruguay

InfoType	Description
URUGUAY_CDI_NUMBER	A *Uruguayan Cédula de Identidad (CDI)*, or identity card, is used as the main identity document for citizens.

Google Cloud DLP

★ **Important:** Built-in infoType detectors are not a 100% accurate detection method. For example, they can't guarantee compliance with regulatory requirements. You must decide what data is sensitive and how to best protect it. Google recommends that you test your settings to make sure your configuration meets your requirements.

Global

InfoType	Description
`ADVERTISING_ID`	Identifiers used by developers to track users for *advertising purposes*. These include Google Play Advertising IDs, Amazon Advertising IDs, Apple's identifierForAdvertising (IDFA), and Apple's identifierForVendor (IDFV).

Modeling Personal / Private Data

Categories Implying Sensitivity	% of Participants
Files containing the participant's PII	62%
Files containing PII of other than the participant	31%
Files with intimate or embarrassing content	30%
Files with original or creative content	84%
Files with proprietary information	23%
Categories Implying Usefulness	**% of Participants**
Files stored for future referencing	96%
Files with content of sentimental value	87%
Files which serve as backup	91%

Table 5: The percentage of participants who reported having files in categories implying they might be sensitive or useful.

Category	Collection Method	List of Features
Metadata	Google Drive/Dropbox API	account size, used space, file size, file type (img, doc, etc.), extension (jpg, txt, etc.), last modified date, last modifying user, access type (owner, editor, etc.), sensitive filename, sharing status
Documents	Local text processing	bag of words for top 100 content keywords, LDA topic models, TF-IDF vectors, word2vec representations, table schemas for spreadsheets
Images	Google Vision API [20]	image object labels, adult, racy, medical, violent, logos, dominant RGB values, average RGB value
Sensitive Identifiers	Google DLP API [18]	***counts*** of the following identifiers in a file: name, gender, ethnic group, address, email, date of birth, drivers license #, passport #, credit card, SSN, bank account #, VIN

Table 3: A list of the features we automatically collected for each file using multiple APIs and custom code.

Khan et al. “Helping Users Automatically Find and Manage Sensitive, Expendable Files in Cloud Storage.” In *Proc. USENIX Security*, 2021.

Can You Screw Up Data Releases?

- Yes!

Case Study 1:
ZIP Code, DOB, Sex

Massachusetts Health Data

- Mid 1990s: Group Insurance Commission (GIC)
- Upon request, GIC will release records with 100 attributes for every state employee’s hospital visits
- Latanya Sweeney, “Uniqueness of Simple Demographics in the U.S. Population”:
  - 87%: ZIP code + full Date of Birth + Sex is uniquely identifying
  - 53%: *City* + full Date of Birth + Sex is uniquely identifying
  - 18%: *County* + full Date of Birth + Sex is uniquely identifying
- William Weld (Governor of Massachusetts) deanonymized when Sweeney purchased voter rolls from the city of Cambridge
  - Sweeney sent the governor’s records (diagnoses/prescriptions) to him

Case Study 2:
AOL Search Data

AOL Search Data Release

- AOL Research released 20,000,000 search queries for 650,000 users of AOL's search engine (3 months)
- Suppressed AOL username and IP address
  - Replaced them with unique, pseudonymous identifiers

AOL Search Data Release (Aftermath)

The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

AOL Search Data Release (Aftermath)

- “…User 4417749’s identity in queries such as “‘landscapers in Lilburn, Ga,’ several people with the last name Arnold and ‘homes sold in shadow lake subdivision gwinnett county georgia.’” They quickly tracked down Thelma Arnold, a sixty-two-year-old widow from Lilburn, Georgia who acknowledged that she had authored the searches, including some mildly embarrassing queries such as “numb fingers,” “60 single men,” and “dog that urinates on everything.”

Case Study 3:
Netflix Prize

Deanonymizing the Netflix Prize

- Netflix released 100,000,000 records from 500,000 users
  - December 1999 to December 2005
  - Assigned a unique pseudonymous identifier to each user
- Each record included the pseudonymous identifier, the movie watched, the rating (1-5 stars), and rating’s date

Deanonymizing the Netflix Prize

- Narayanan and Shmatikov correlated with IMDb
- Ratings on IMDb are public
- Databases are not perfect subsets of each other
- What can be leaked from knowing which movies an identified user watched?

Deanonymizing the Netflix Prize

$$\mathsf{Sim}(r_1, r_2) = \frac{\sum \mathsf{Sim}(r_{1i}, r_{2i})}{|\mathsf{supp}(r_1) \cup \mathsf{supp}(r_2)|}$$

Definition 3 (De-anonymization) *An arbitrary subset $\hat{D}$ of a database D can be (θ, ω)-deanonymized w.r.t. auxiliary information* Aux *if there exists an algorithm A which, on inputs $\hat{D}$ and* $\mathsf{Aux}(r)$ *where $r \leftarrow D$*

- *If $r \in \hat{D}$, outputs r' s.t.* $\Pr[\mathsf{Sim}(r, r') \geq \theta] \geq \omega$
- *if $r \notin \hat{D}$, outputs $\perp$ with probability at least ω*

Deanonymizing the Netflix Prize

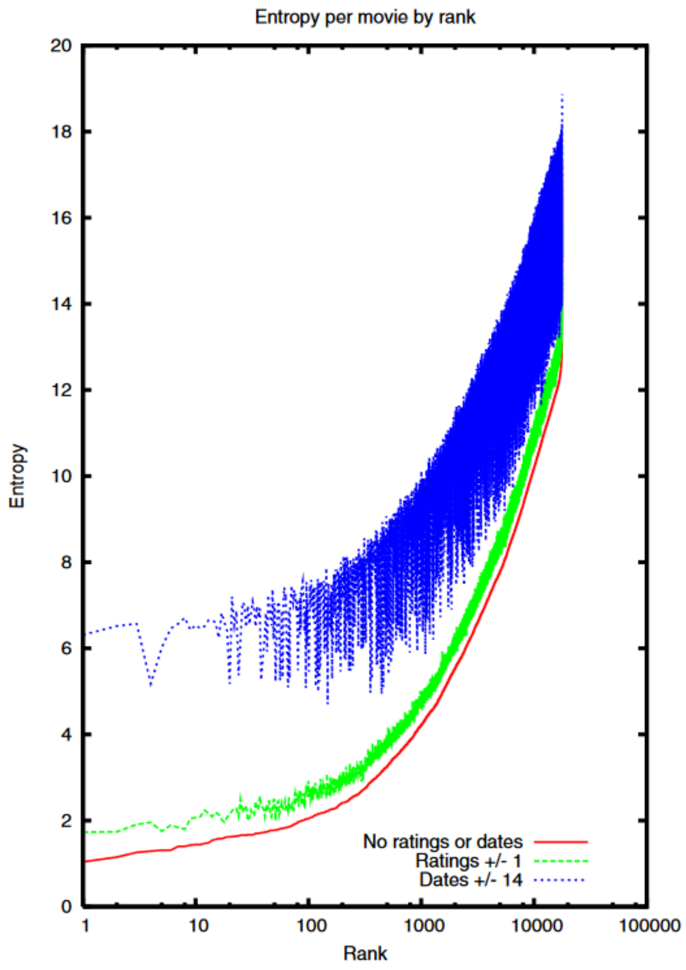

Figure 7. Entropy of movie by rank

Recap

The Surprising Success of Deanonymization

- The use of **auxiliary information**
  - Extremely hard to control
- Errors suppressing data
- Personal data showing up in unexpected places
- It’s hard to reason about what is/is not identifiable
- **Thinking only about personal data / PII is not sufficient**