Data privacy: Introduction

Vicenç Torra

January 2025

Umeå University, Sweden

V. Torra (2022) A guide to data privacy, Springer (Chapter 1)

Outline

Introduction

Introduction to data privacy

- Chapter 1. Introduction

A context:

Data-driven machine learning/statistical models[1]

[1] This is not the only type of scenario/application context we will discuss

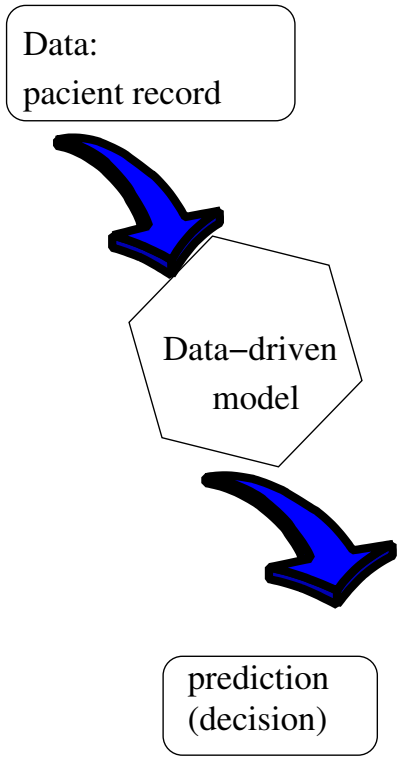
Prediction using (machine learning/statistical) models

- Data is collected to be used
  (otherwise, better not to collect them[2])

[2] Concept: Data minimization

Prediction using (machine learning/statistical) models

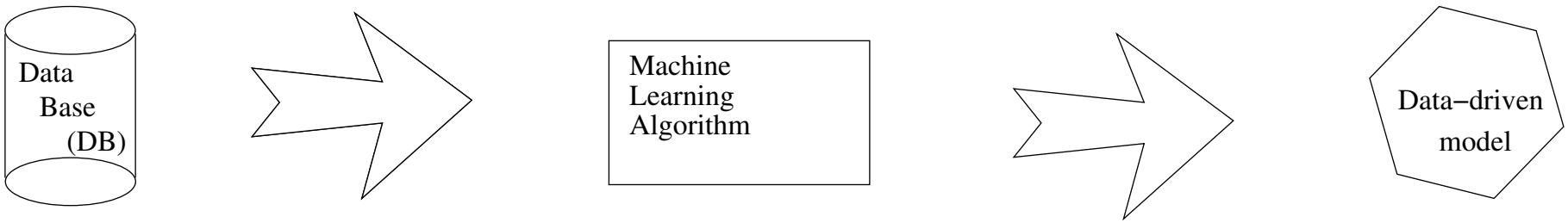
- Application of a model for decision making
  data ⇒ prediction/decision



- Example: predict the length-of-stay at admission

Data-driven machine learning/statistical models

- From (huge) databases, build the "decision maker"
  - Use (logistic) regression, deep lerning, neural networks, . . . classification algorithms, decision trees, . . .



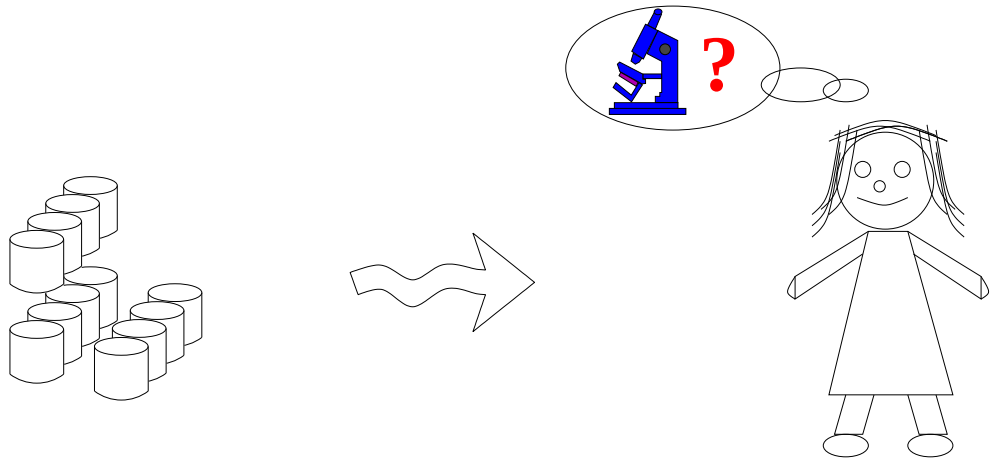
- Example: build a predictor from hospital historical data about length-of-stay at admission

Privacy for machine learning and statistics:

Data-driven machine learning/statistical models

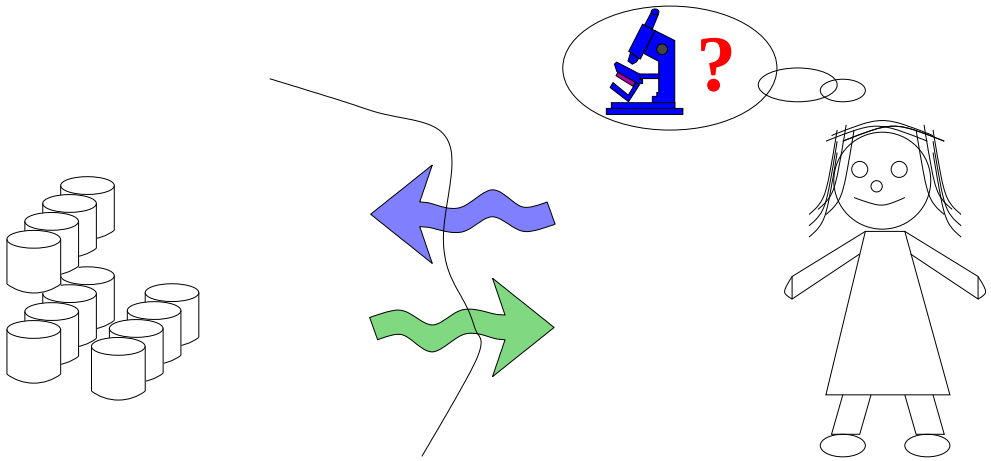
Data is sensitive

- Who/how is going to create this model (this "decision maker")?
- Case #1. Sharing (part of the data)



Data is sensitive

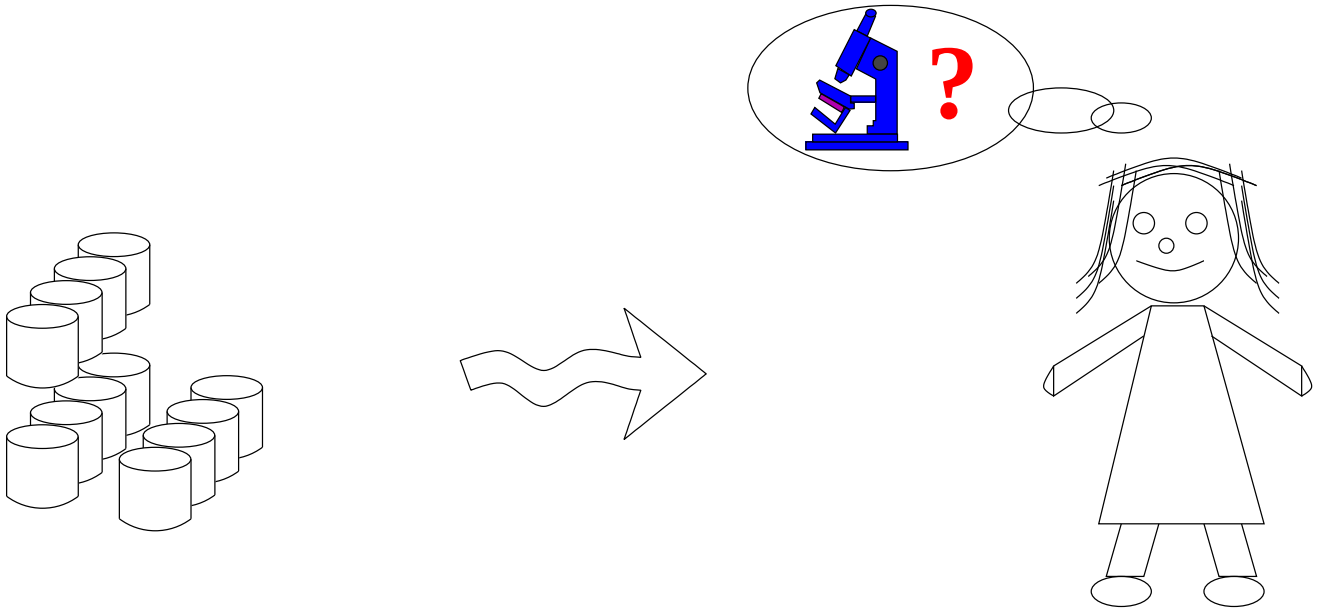
- Who/how is going to create this model (this "decision maker")?
- Case #2. Not sharing data, only querying data



Two motivating examples

Data is sensitive

- Data privacy: core
  - Someone needs to access to data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure.



E.g., you are authorized to compute the average stay in a hospital, but maybe you are not authorized to see the length of stay of your neighbor.

Data is sensitive

- Case #1. Sharing (part of the data)
- Q: How different children ages and diagnoses affect this length of stay? Average length of stay is decreasing in the last years due to new hospital policies?
- Data: Existing database with previous admissions (2010–2019).
  To avoid disclosure a view of the DB restricting records to children born before 2019 and only providing for these records year of birth, town, year of admission, illness, and length of stay.

Data is sensitive

- Case #1. Sharing (part of the data)

Year birth	Year Admission	Town	Illness	Length stay (days)
2017	2019	Umeå	a	3
2015	2020	Umeå	b	2
2011	2020	Luleå	c	5
2017	2019	Luleå	a	2
2016	2020	Dorotea	b	4
2016	2020	Holmöns	d	2
2015	2019	Täfteå	e	4
2015	2019	Täfteå	e	4
2015	2018	Täfteå	e	4
2015	2018	Täfteå	e	4

- Is this data safe?

Data is sensitive

- Case #1. Sharing (part of the data)

Year birth	Year Admission	Town	Illness	Length stay (days)
2017	2019	Umeå	a	3
2015	2020	Umeå	b	2
2011	2020	Luleå	c	5
2017	2019	Luleå	a	2
2016	2020	Dorotea	b	4
2016	2020	Holmöns	d	2
2015	2019	Täfteå	e	4
2015	2019	Täfteå	e	4
2015	2018	Täfteå	e	4
2015	2018	Täfteå	e	4

- Is this data safe?

Holmöns 63, Täfteå 1383, Luleå 49123, Umeå 83249, Dorotea 2366

Data is sensitive

- Case #1. Sharing (part of the data). Example 2
  - Q: stress influenced by studies and commuting distance ?
  - Data: DB/view = (where students live, what study, got sick?)

    (Umeå, CS, no)
    (Umeå, CS, yes)
    (Umeå, . . . , . . .)
    . . .
    (Holmsund, CS, no)
    (Holmsund, CS, no)
    (Holmsund, CS, yes)
    (Holmsund, . . . , . . .)
    . . .
    (Norrbyn, XXXX, yes)
  - No "personal data", is this ok ? NO!!

  $\Rightarrow$ We learn that our friend[3] is sick !!

[3] Norrbyn 168 (2023)

Data is sensitive: computation leads to disclosure

- Case #2. Sharing a computation.
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town (Bunyola)?

Data is sensitive: computation leads to disclosure

- Case #2. Sharing a computation.
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town (Bunyola)?
  - Mean income is not "personal data", is this ok ? NO!!:

Data is sensitive: computation leads to disclosure

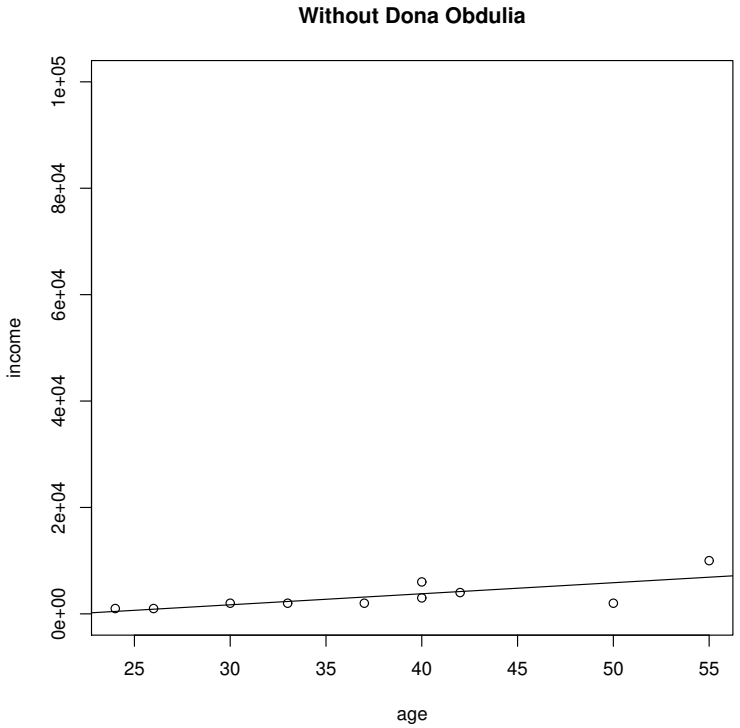
- Case #2. Sharing a computation.
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town (Bunyola)?
  - Mean income is not "personal data", is this ok ? NO!!:
  - Example 1000 2000 3000 2000 1000 6000 2000 10000 2000 4000 $\Rightarrow$ mean = 3300

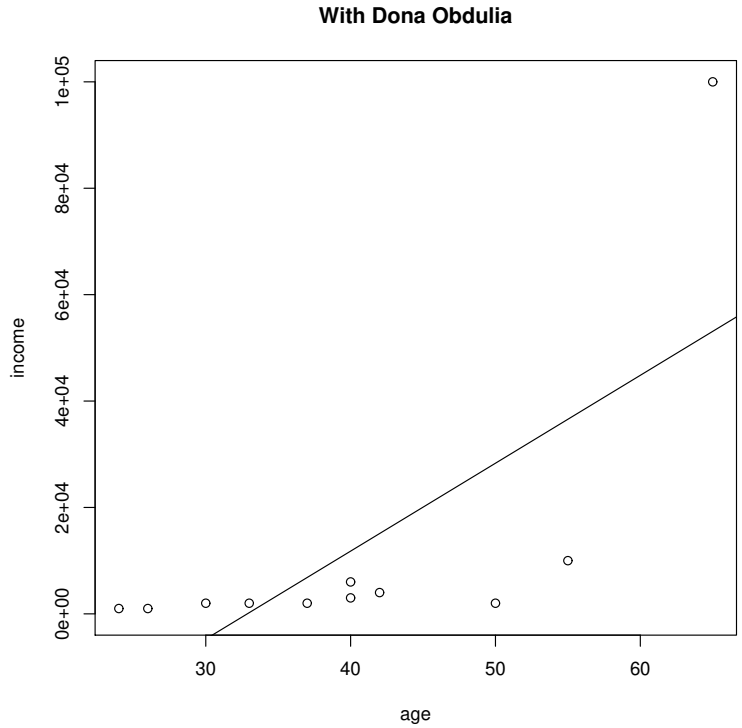
Data is sensitive: computation leads to disclosure

- Case #2. Sharing a computation.
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town (Bunyola)?
  - Mean income is not "personal data", is this ok ? NO!!:
  - Example 1000 2000 3000 2000 1000 6000 2000 10000 2000 4000 $\Rightarrow$ mean = 3300
  - Adding Ms. Rich's salary 100,000 Eur/month: mean = 12090,90 !
    (a extremely high salary changes the mean significantly)
    $\Rightarrow$ We infer Ms. Rich from Town was attending the unit

Data is sensitive: computation leads to disclosure

- Case #2. Sharing a computation. Example 2





- Regression of income with respect to age with (right) and without (left) the record of Dona Obdúlia
  - income = -4524.2 + 207.5 age (without Ms. Rich = Dona Obdúlia)
  - income = -54307 + 1652 age (with Ms. Rich = Dona Obdúlia)

Data privacy: the core

Data is sensitive: Data privacy

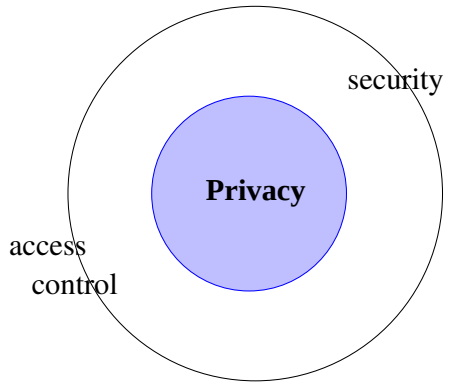
- Privacy leaks
  - Transmissions of unencrypted e-mails: senders, receivers, content
  - Unauthorized access to an account or service: sensitive information
  - AOL[4]: linkage of search log queries
    $\Rightarrow$ User No. 4417749, hundreds of searches over a three-month period including queries 'landscapers in Lilburn, Ga'
    $\longrightarrow$ Thelma Arnold identified!
  - Netflix case[5]: linkage of movie ratings
    $\Rightarrow$ individual users matched with film ratings on the Internet Movie Database.
  - Similar with credit card payments, shopping carts, ...

[4] http://www.nytimes.com/2006/08/09/technology/09aol.html
[5] Narayanan and Shmatikov, 2008

Data is sensitive: Data privacy

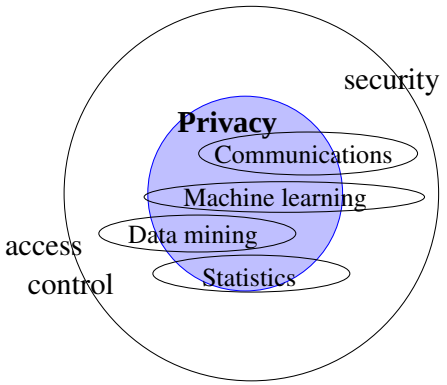
- A personal view of core and boundaries of data privacy: boundaries
  - Database in a computer or in a removable device
    $\Rightarrow$ access control to avoid unauthorized access
    $\Longrightarrow$ Access to address (admissions), Access to blood test (admissions?)
  - Data is transmitted
    $\Rightarrow$ security technology to avoid unauthorized access
    $\Longrightarrow$ Data from blood glucose meter sent to hospital. Network sniffers
    Transmission is sensitive: Near miss/hit report to car manufacturers

- Data privacy: disclosure because of inference



Data is sensitive: Data privacy

- A personal view of core and boundaries of data privacy: core
- Someone needs to access to data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure.
  - data uses / rellevant techniques
    - ▷ Data to be used for data analysis
      - ⇒ compute indices, find patterns, build models
      - ⇒ statistics, machine learning, data mining
    - ▷ Data is transmitted
      - ⇒ communication and inform. retrieval (identities/data)
      - ⇒ communications



Data privacy: difficulties

Difficulties

- Naive anonymization does not work

  Passenger manifest for the Missouri, arriving February 15, 1882; Port of Boston[6]

  Names, Age, Sex, Occupation, Place of birth, Last place of residence, Yes/No, condition (healthy?)

LIST OF PASSENGERS.

REPORT AND LIST OF PASSENGERS taken on board the S.S. Hooper at London.

NAMES	AGE	SEX	OCCUPATION	PLACE OF BIRTH	Last Place of Residence	Yes	No	CONDITION
[illegible]	11	Male	Gentleman	Aberdeen Eng	Chelmsford Eng	yes		Healthy
[illegible]	24	Lawyer's Clerk	Madras India	Saltwood Eng			no	
[illegible]	21	Tailor	London	London			no	
[illegible]	26	Surgeon	London	London			no	
[illegible]	28	Gentleman	Boston U.S.	Boston U.S.	yes			
[illegible]	30		Scotland	Boston U.S.	yes			
[illegible]	16		London (?)	London	yes			
[illegible]	44		Dover Eng.	Dover Eng.	yes			
[illegible]	28		Alexandria Egypt	Boston U.S.	yes			
[illegible]	28	Cattle Dealer	Boston U.S.	Boston U.S.	yes			

[6] https://www.sec.state.ma.us/arc/arcgen/genidx.htm

Difficulties

- Difficulties: highly identifiable data
  - (Sweeney, 1997) on USA population
    - 87.1% (216 million/248 million) were likely made them unique based on
      5-digit ZIP, gender, date of birth,
    - 3.7% (9.1 million) had characteristics that were likely made them unique based on
      5-digit ZIP, gender, Month and year of birth.

Difficulties

- Difficulties: highly identifiable data and high dimensional data
  - Data from mobile devices:
    $\Rightarrow$ two positions can make you unique (home and working place)
  - AOL[7] and Netflix cases (search logs and movie ratings)
    $\Rightarrow$ User No. 4417749, hundreds of searches over a three-month period including queries 'landscapers in Lilburn, Ga'
    $\longrightarrow$ Thelma Arnold identified!
    $\Rightarrow$ individual users matched with film ratings on the Internet Movie Database.
  - Similar with credit card payments, shopping carts, ...

[7] http://www.nytimes.com/2006/08/09/technology/09aol.html

Difficulties

- Difficulties: highly identifiable data and high dimensional data
  - Ex1: Sickness influenced by studies and commuting distance ?
  - Ex2: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town?
  - Ex3: Driving behavior in the morning
    - ▷ Automobile manufacturer uses (data from vehicles)
    - ▷ Data: First drive after 6:00am
      (GPS origin + destination, time) $\times$ 30 days
    - ▷ No "personal data", is this ok?: NO!!!:
    - ▷ How many cars from your home to your work?
      Are you exceeding the speed limit? Are you visiting a psychiatric clinic every tuesday?

Difficulties

- Difficulties: highly identifiable data and high dimensional data
  - Data from people: In some cases, it is the combination of characteristics that make you unique. (Search logs, Market Basket Analysis)
    - Manga
    - K-pop music
    - Opera
    - IF Björklöven
    - ...

      $\rightarrow |\{ \text{Manga} \cap \text{k-pop} \cap \text{Opera} \cap \text{IF Björklöven} \}| = 1 = 1$

Difficulties

- Data privacy is "impossible", or not? challenging
  - Privacy vs. utility
  - Privacy vs. security
  - Computationally feasible

Privacy and society

Legislation and motivation

- Privacy a fundamental right. (Ch. 1.1)
  - Universal Declaration of Human Rights (UN). European Convention on Human Rights (Council of Europe). General Data Protection Regulation - GDPR (EU). National regulations.
- Regulations
  - EU: General Data Protection Regulation (GDPR)
  - USA: Health Insurance Protability and Accountability Act (HIPAA, 1996), Patriot Act (2001), Homeland Security Act (2002), Children's Online Privacy Protection Act (COPPA, 2000).
  - California: California Consumer Privacy Act (CCPA)

Legislation and motivation

- Enforcement
  - Obligations with respect to data processing
  - Requirement to report personal data breaches
  - Grant individual rights (to be informed, to access, to rectification, to erasure, ...)
  - Data protection officer[8]
  - Data protection impact assessment (DPIA) / privacy impact assessment (PIA) [9]
- Companies own interest.
  - Competitors can take advantage of information.
- Avoiding privacy breach. Several well known cases.

[8] https://edps.europa.eu/data-protection/data-protection/reference-library/data-prote
[9] https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-an

Privacy and society

- Privacy and society
  - Not only a computer science/technical problem
    - Social roots of privacy
    - Multidisciplinary problem
  - Social, legal, philosophical questions
  - Culturally relative?
    I.e., the importance of privacy is the same among all people ?
  - Are there aspects of life which are inherently private or just conventionally so?

Privacy and society

- Privacy and society. Is this a new problem? Yes and not
  - No side. See the following:

    Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that "what is whispered in the closet shall be proclaimed from the house-tops." (...)

    Gossip is no longer the resource of the idle and of the vicious, but has become a trade, which is pursued with industry as well as effrontery (...) To occupy the indolent, column upon column is filled with idle gossip, which can only be procured by intrusion upon the domestic circle.

    (S. D. Warren and L. D. Brandeis, 1890)[10]
  - Yes side: big data, storage, surveillance/CCTV, RFID, IoT

[10] https://www.jstor.org/stable/1321160

Concepts

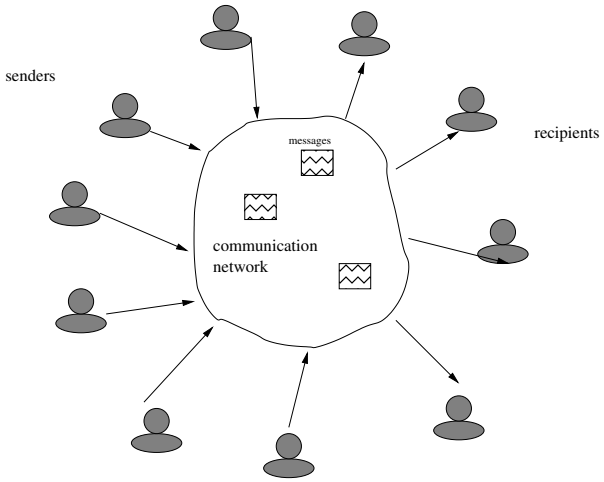
Related fields

- Technical solutions
  - Statistical disclosure control (SDC)
  - Privacy preserving data mining (PPDM)
  - Privacy enhancing technologies (PET)
- Socio-technical aspects
  - Technical solutions are not enough
  - Implementation/management of solutions for achieving data privacy need to have a holistic perspective of information systems
  - E.g., employees and customers: how technology is applied

Terminology

Terminology

- Terminology using as framework a communication network with senders (actors) and receivers (actees)



- Attacker, adversary, intruder
  - the set of entities working against some protection goal
  - increase their knowledge (e.g., facts, probabilities, . . .) on the items of interest (IoI) (senders, receivers, messages, actions)

Terminology

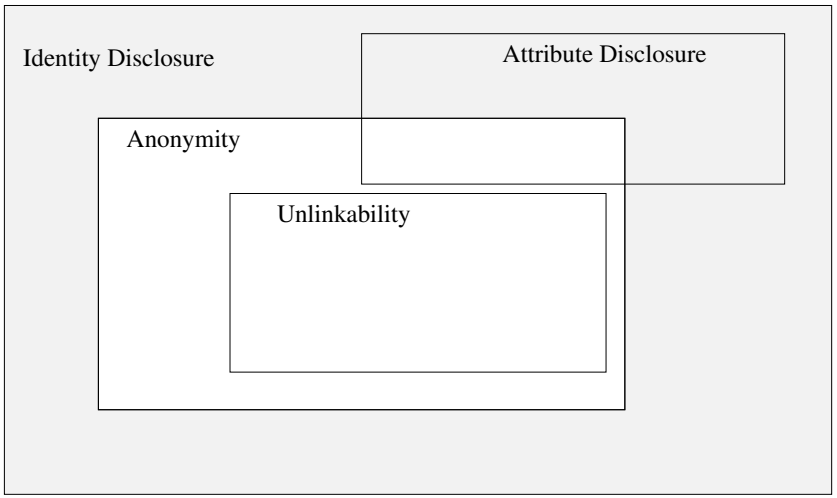
- Anonymity set. Anonymity of a subject means that the subject is not identifiable within a set of subjects, the anonymity set. Not distinguishable!

- Unlinkability. Unlinkability of two or more IoI, the attacker cannot sufficiently distinguish whether these IoIs are related or not.
  $\Rightarrow$ Unlinkability with the sender implies anonymity of the sender.

  - Linkability but anonymity. E.g., an attacker links all messages of a transaction, due to timing, but all are encrypted and no information can be obtained about the subjects in the transactions: anonymity not compromised.
    (region of the anonymity box outside unlinkability box)

Terminology

- Examples of anonymity in communications (definition of IoI):
  - **Sender anonymity.** No link between a message and the sender.
  - **Recipient anonymity.** No link between a message and the receiver.
  - **Relationship anonymity.** No link between a message and both sender and receiver.



Terminology

- Disclosure. Attackers take advantage of observations to improve their knowledge on some confidential information about an IoI.
  $\Rightarrow$ SDC/PPDM: Observe DB, Δ knowledge of a particular subject (the respondent in a database)
  - Identity disclosure (entity disclosure). Linkability. Finding Mary in the database.
  - Attribute disclosure. Increase knowledge on Mary's salary.
    also: learning that someone is in the database, although not found.

Terminology

- Disclosure. Discussion.
  - Identity disclosure. Avoid.
  - Attribute disclosure. A more complex case. Some attribute disclosure is expected in data mining.

    At the other extreme, any improvement in our knowledge about an individual could be considered an intrusion. The latter is particularly likely to cause a problem for data mining, as the goal is to improve our knowledge. (J. Vaidya et al., 2006, p. 7.)

Terminology

- Identity disclosure vs. attribute disclosure
  - Usually, identity disclosure implies attribute disclosure

    Find record $(HYU, Tarragona, 58)$, learn variable $(Heart\ Attack)$

Respondent	City	Age	Illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS
HYU	Tarragona	58	Heart attack

Terminology

- Identity disclosure vs. attribute disclosure
  - Usually, identity disclosure implies attribute disclosure

Respondent	City	Age	Illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS
HYU	Tarragona	60	Heart attack

  - Identity disclosure without attribute disclosure. Use all attributes *(Tarragona, 60, Heart attack)*.

Terminology

- Identity disclosure vs. attribute disclosure
  - Usually, identity disclosure implies attribute disclosure

Respondent	City	Age	Illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS
HYU	Tarragona	60	Heart attack

  - Identity disclosure without attribute disclosure. Use all attributes $(Tarragona, 60, Heart\ attack)$.
  - Attribute disclosure without identity disclosure. k-anonymity $(ABD, Barcelona, 30)$ not reidentified but learn $Cancer$

Terminology

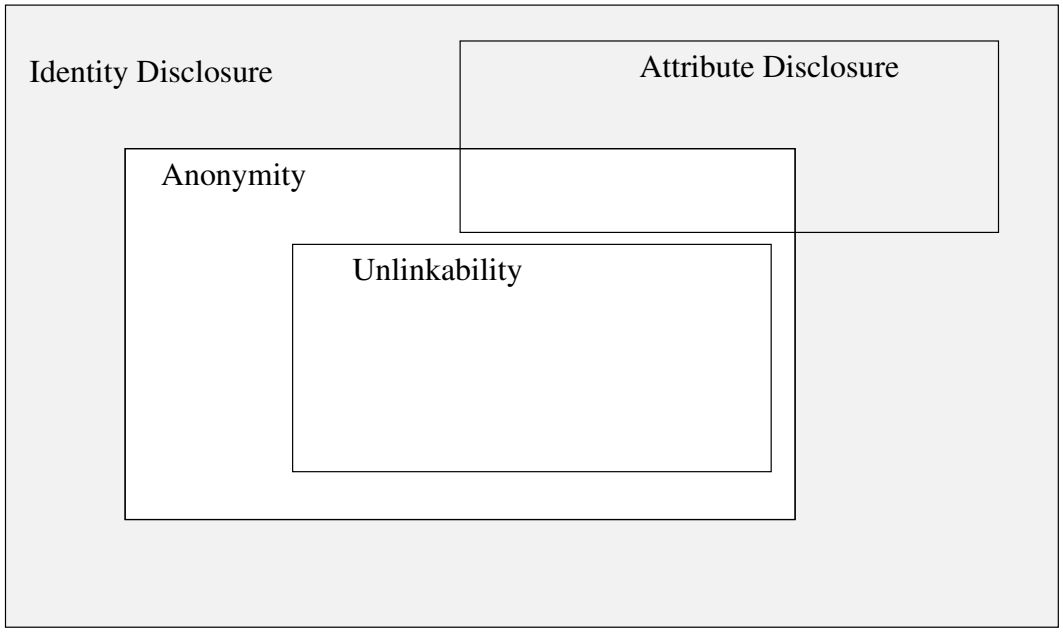
- Identity disclosure vs. attribute disclosure
  - Usually, identity disclosure implies attribute disclosure

Respondent	City	Age	Illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS
HYU	Tarragona	60	Heart attack

  - Identity disclosure without attribute disclosure. Use all attributes $(Tarragona, 60, Heart\ attack)$.
  - Attribute disclosure without identity disclosure. k-anonymity $(ABD, Barcelona, 30)$ not reidentified but learn $Cancer$
  - Neither identity nor attribute disclosure

Terminology

- Identity disclosure and anonymity are exclusive.
  - Identity disclosure implies non-anonymity
  - Anonymity implies no identity disclosure.



Terminology

- So,
  - Identity and attribute disclosure
  - Anonymity, unlinkability
- Privacy try to provide guarantees / algorithms for avoiding disclosure

Terminology

- Undetectability and unobservability
  - Undetectability of an IoI. The attacker cannot sufficiently distinguish whether IoI exists or not.
    E.g. Intruders cannot distinguish messages from random noise
    $\Rightarrow$ Steganography
  - Unobservability of an IoI means
    - ▷ undetectability of the IoI against all subjects uninvolved in it and
    - ▷ anonymity of the subject(s) involved in the IoI even against the other subject(s) involved in that IoI.

Unobservability pressumes undetectability but at the same time it also pressumes anonymity in case the items are detected by the subjects involved in the system. From this definition, it is clear that unobservability implies anonymity and undetectability.

Terminology

- Plausible deniability
  - I have nothing to do with this database, model, etc
  - Is this statement credible?
- For a database
  - at record level: This record is not mine!
  - at database level: I am not in this database!

Terminology

- Plausible deniability
  - I have nothing to do with this database, model, etc
  - Is this statement credible?
- For a database
  - at record level: This record is not mine!
  - at database level: I am not in this database!
- We will see that some privacy models provide guarantees for plausible deniability

Terminology

- Connections between plausible deniability and anonymity set
  - Plausible deniabilty: perspective of the individual
  - Anonymity set: perspective of the intruder

Transparency

Transparency

- Transparency
  - DB is published: give details on how data has been produced. Description of any data protection process and parameters
  - Positive effect on data utility. Use information in data analysis.
  - Negative effect on risk. Intruders use the information to attack.

- The transparency principle in data privacy[11]

  Given a privacy model, a masking method should be compliant with this privacy model even if everything about the method is public knowledge. (Torra, 2017, p17)

[11] Similar to the Kerckhoffs's principle (Kerckhoffs, 1883) in cryptography: a cryptosystem should be secure even if everything about the system is public knowledge, except the key

Interim discussion

Discussion

- Privacy and disclosure
  - Identity disclosure appears when information about individuals make them unique: avoid identity disclosure
    - A few attributes, very distinctive attributes
      $\Rightarrow$ highly identifiable data
    - A large number of attributes, not distinctive
      $\Rightarrow$ high dimensional data
  - Same about "interests", multifaceted individuals
    - FC Barcelona, Umeå IK
    - Vegetarian, beef lover
    - etc
  - Attribute disclosure: some attribute disclosure is usually expected, difficult to define/protect

Privacy by design

Privacy by design

- Privacy by design (Cavoukian, 2011)
  - Privacy "must ideally become an organization's default mode of operation" (Cavoukian, 2011) and thus, not something to be considered a posteriori. In this way, privacy requirements need to be specified, and then software and systems need to be engineered from the beginning taking these requirements into account.
  - *In the context of developing IT systems, this implies that privacy protection is a system requirement that must be treated like any other functional requirement. In particular, privacy protection (together with all other requirements) will determine the design and implementation of the system* (Hoepman, 2014)

Privacy by design

- Privacy by design principles (Cavoukian, 2011)

1. Proactive not reactive; Preventative not remedial.
2. Privacy as the default setting.
3. Privacy embedded into design.
4. Full functionality – positive-sum, not zero-sum.
5. End-to-end security – full lifecycle protection.
6. Visibility and transparency – keep it open.
7. Respect for user privacy – keep it user-centric.

Privacy by design

- Data minimization: a cornerstone for privacy
  - *The principle of "data minimisation" means that a data controller should limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose. They should also retain the data only for as long as is necessary to fulfil that purpose. In other words, data controllers should collect only the personal data they really need, and should keep it only for as long as they need it.*

    The data minimisation principle is expressed in Article 5(1)(c) of the GDPR and Article 4(1)(c) of Regulation (EU) 2018/1725, which provide that personal data must be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed".[12]

[12] https://edps.europa.eu/data-protection/data-protection/glossary/d_en

Summary

Summary

- Concepts
  - What is data privacy?
  - Multidisciplinary problem and socio-technical aspects to be considered
  - Difficulties of data privacy: naive annonymization does not work
  - Linkability and anonymity set
  - Identity and attribute disclosure
  - Plausible deniability
  - Transparency
  - Privacy by design
  - Data minimization

References

References

- Torra, V. (2022) Guide to data privacy, Springer.
- Cavoukian, A. (2011) Privacy by design. The 7 foundational principles in Privacy by Design. Strong privacy protection - now, and well into the future.
- Narayanan, A., Shmatikov, V. (2008) Robust De-anonymization of Large Sparse Datasets, Proc. of the 2008 IEEE Symposium on Security and Privacy (SP '08), 111-125.
- Sweeney, L. (1997) Simple Demographics Often Identify People Uniquely, Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.
- Dalenius, T. (1977) Towards a methodology for statistical disclosure control, Statistisk Tidskrift 5 429-444.
- D'Acquisto, G., Domingo-Ferrer, J., Kikiras, P., Torra, V., de Montjoye, Y.-A., Bourka, A. (2015) Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics, ENISA Report.
- V. Torra, G. Navarro-Arribas (2016) Big Data Privacy and Anonymization, Privacy and Identity Management 15-26. https://doi.org/10.1007/978-3-319-55783-0_2 (open access)