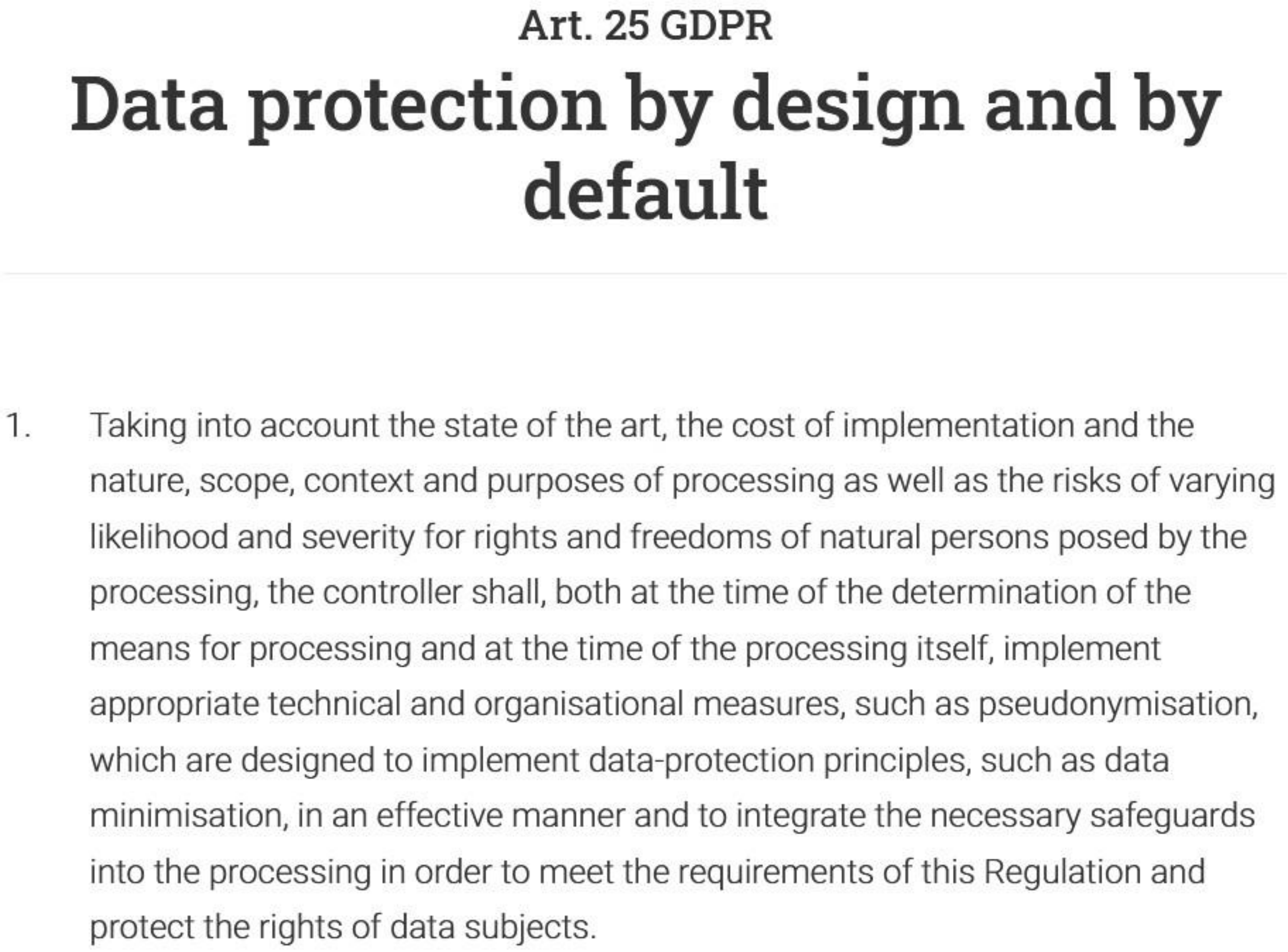
Lecture 12:
Privacy Engineering

CMSC 25910

Spring 2022

The University of Chicago

Privacy by Design (PbD)



Art. 25 GDPR

Data protection by design and by default

1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

https://gdpr-info.eu/art-25-gdpr/

Potential Goals of Privacy Engineering

- Compliance with laws
  - GDPR, CCPA, etc.
- Compliance with reasonable consumer expectations and making accurate public statements
  - See, e.g., FTC’s mandates
- Engender trust and goodwill among your users
  - “Competing” based on privacy-protectiveness
- Protecting privacy as a societal value / “it’s the right thing”

Mechanisms

- Thinking carefully and being selective about data collection
  - Data minimization, immediate de-identification/pseudonymization
- Not retaining data
- Thoughtful applications of cryptographic tools
- Thoughtful architectures for computer systems
  - Access control (locked down? open with audit logs?), data storage
- Public statements and user communication (e.g., in UIs)
- Appropriate socio-technical processes, audits, and reviews

Privacy Impact Assessments (PIAs)

- “A PIA is an **analysis of how personally identifiable information is collected, used, disseminated, and maintained**. It examines how the Department has **incorporated privacy concerns throughout its development, design, and deployment** of a technology, program, or rulemaking. “Personally identifiable information” is defined as any information that permits the identity of an individual to be directly or indirectly inferred, including any other information which is linked or linkable to that individual”

https://www.dhs.gov/sites/default/files/publications/privacy_pia_guidance_june2010_0.pdf

Privacy Impact Assessments (PIAs)

- “The purpose of a PIA is to demonstrate that program managers and system owners have **consciously incorporated privacy protections throughout the development life cycle** of a system or program. This involves making certain that privacy protections are **built into the system from the initiation of development**, not after the fact when they can be far more costly or could affect the viability of the project. The PIA process requires that candid and forthcoming communications occur between the program manager, system owner, the component’s Privacy Officer, and the Privacy Office to ensure appropriate and timely handling of privacy concerns. Addressing privacy issues publicly through a PIA builds citizen trust...”

https://www.dhs.gov/sites/default/files/publications/privacy_pia_guidance_june2010_0.pdf

Key Privacy Organizations / Nonprofits

- International Association of Privacy Professionals
  - “IAPP” https://iapp.org
  - Provide certifications like CIPM, CIPP/US, and CIPT
- Future of Privacy Forum
  - “FPF” https://fpf.org/
  - Think tank and advocacy group
- Electronic Frontier Foundation
  - “EFF” https://eff.org/
  - Nonprofit advocacy group defending digital privacy

Privacy Engineering Practice & Respect

- USENIX Conference on Privacy Engineering & Respect (PEPR)
- https://usenix.org/conference/pepr22

PEPR '22

2022 USENIX Conference on Privacy Engineering Practice and Respect

JUNE 23–24, 2022
SANTA CLARA, CA, USA

usenix
THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Some Additional Technical Tools

Hashing

- Function that maps arbitrarily sized data to a fixed output
- Desired properties
  - Maps inputs relatively **uniformly** to the space of possible outputs
  - **Efficiently** computable (but not desirable for password storage!)
  - **Deterministically** always maps a given input to the same output
- **Cryptographic hash functions** are one-way functions (very hard to invert)
- Simply hashing PII seems like it would provide privacy…
  - … but you can simply enumerate and hash possible inputs of interest!
  - In *some* cases, you may want to salt inputs and then discard the salt

Bloom Filters

- Probabilistic data structure for *set membership*
  - False negatives are *impossible*
    - Bloom filter returns “no” → True answer is “no”
  - False positives are *possible*
    - Bloom filter returns “yes” → True answer is probably “yes,” but might be “no” with some probability (that you can calculate)
- Define an array of **m** bits and **k** different hash functions

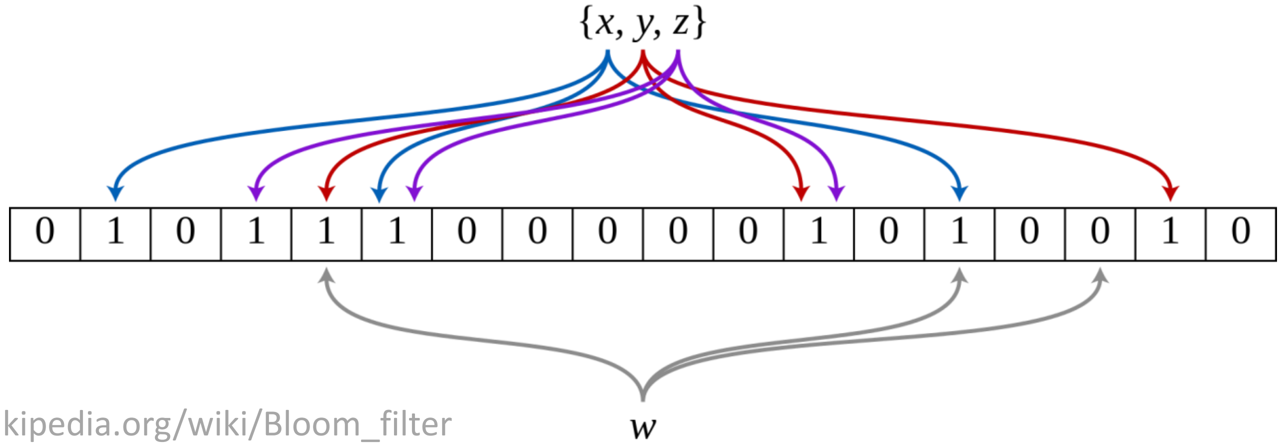


Image taken from https://en.wikipedia.org/wiki/Bloom_filter

Secure Multiparty Computation (MPC)

- Subfield of cryptography
- Multiple people jointly compute a function on inputs provided by everyone *while keeping those inputs private from each other*
- Adversarial models:
  - Semi-honest / honest-but-curious: Assume participants follow the protocol, but want to learn the private values
  - Malicious: Assume participants may cheat
- Interesting cryptography (you can learn about in other courses)
- Techniques like Private Set Intersection (PSI) are used, for instance, in Apple’s password breach monitoring service

Simple MPC Example

- Three people want to compute the average # of succulent plants they have without (shamefully) admitting how many they have
- Each person creates three shares of their own count, distributing them securely to the three people (including themselves)
  - Blase has 150. Makes “shares” 120, -20, 50 (sum to 150)
  - Weijia has 20. Makes “shares” 100, -120, 40 (sum to 20)
  - Kevin has 10. Makes “shares” 50, -60, 20 (sum to 10)
- Blase ends up with (120, 100, 50); Weijia with (-20, -120, -60); Kevin with (50, 40, 20).
- Averages: Blase (90); Weijia (-66.7); Kevin (36.7) = **60**

Secret Sharing

- (Adi) Shamir's Secret Sharing scheme is based on polynomial interpolation over finite fields
- Insight: k points uniquely determine a polynomial of degree k-1
    - 2 points uniquely define a line, 3 points uniquely define a polynomial of degree 2 ($ax^2 + bx + c$)
- Approach: pick a polynomial of appropriate degree such that the y-intercept is the "secret" and give everyone a point on this polynomial as their "share"

Case Studies

1: In-store Tracking

- **Setting:** BlaseMart wants to track customers as they move through the store to:
  - Understand which areas of the store are most popular
  - Optimize the relative location of different displays by understanding which sections are highly correlated among consumers’ visits
  - Provide targeted discounts to specific consumers about specific products

2: Maps App

- **Setting:** Blaze (like Waze, but better) wants to provide a privacy-protective alternative to Google Maps
  - Real-time traffic info
  - Accident/closure reporting
  - Convenient history-based recommendations for users

3: Browser

- **Setting:** The Blaze (like Brave, but better) web browser wants to determine what malware websites its users have visited
  - Caveat: The malware sites are only identified after the fact

4: Voice Assistant

- **Setting:** Blazezon wants to improve the speech recognition of Alexa on the Blazezon Echo