École Pour l'Informatique et les Techniques Avancées – EPITA

Masters program

Course: Data Privacy by Design

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

Created & Coordinated by: M. Salman Nadeem

This work is licensed under the Creative Commons Attribution 4.0 International License

cc BY

Data Privacy by Design (DPbD)

Course schedule (tentative)

Date	No.	Topics	Duration (in hours)
(check **Zeus** platform for the date & time)	1	**Introduction, DPbD fundamentals with case studies**	3
(check **Zeus** platform for the date & time)	2	Data privacy risks, Crypto. Package, Data masking (Anonymization vs Pseudonymisation)	3
(check **Zeus** platform for the date & time)	3	Privacy Enhancing Technologies (PETs), DPbD and General Data Protection Regulation (GDPR)	3
(check **Zeus** platform for the date & time)	4	Recap, Conclusion	3
(check **Zeus** platform for the date & time)	5	Final evaluation	3
		Total	*15 hours*

GRADING criteria:
Class participation comprising attendance & reactivity: 10%
Class activities (quizzes): 30%
Final project (individual report & group presentation): 60%

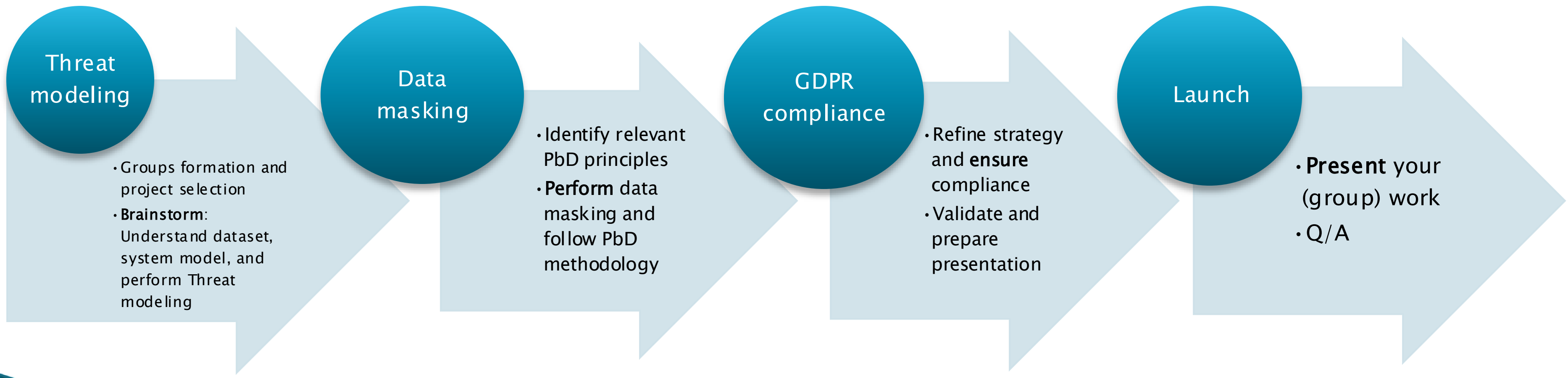
Notes & Collaboration

- MS Teams Channel:
  'Data Privacy by Design'
  - Course specific channel to collaborate
  - Will be used to:
    - Publish course related announcements
    - Provide course slides/material
    - Receive assignments and projects
- Course Mindmap:
  - For better organization and easy refreshing of course topics
  - Access link (read-only):
    https://www.mindomo.com/mindmap/60f6d856c480464ab9f113f60e2fc986

Lecture 1 Outline

1. Data & its types
2. Data privacy
3. Data Privacy by Design (PbD) principles
4. Data PbD goal, strategies & methodology
5. Case study

QUIZ



Threat modeling

- Groups formation and project selection
- **Brainstorm**: Understand dataset, system model, and perform Threat modeling

Data masking

- Identify relevant PbD principles
- **Perform** data masking and follow PbD methodology

GDPR compliance

- Refine strategy and **ensure** compliance
- Validate and prepare presentation

Launch

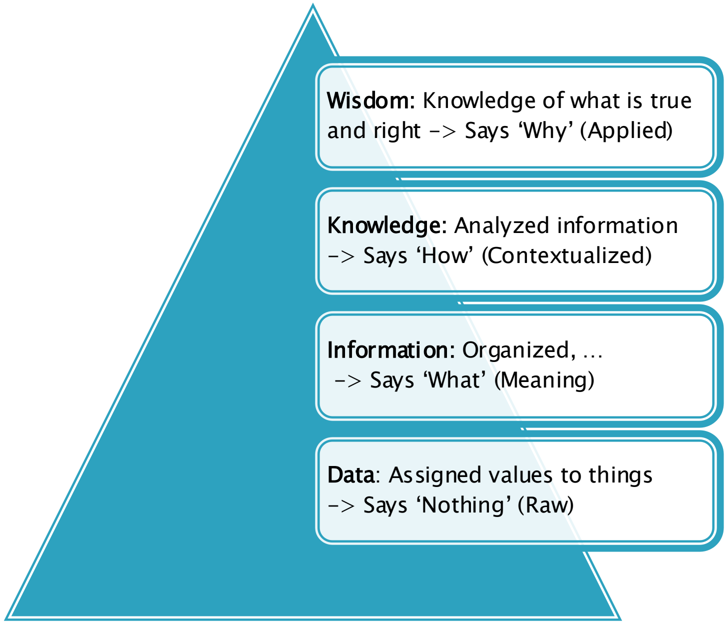
- **Present** your (group) work
- Q/A

Hands-on → Final project

Data

- “Facts and statistics collected together for reference or analysis”

  – Oxford dictionary
- Data is all around us
- Representing Data into Information, Knowledge and Wisdom
  - a.k.a the DIKW pyramid



Wisdom: Knowledge of what is true and right -> Says ‘Why’ (Applied)

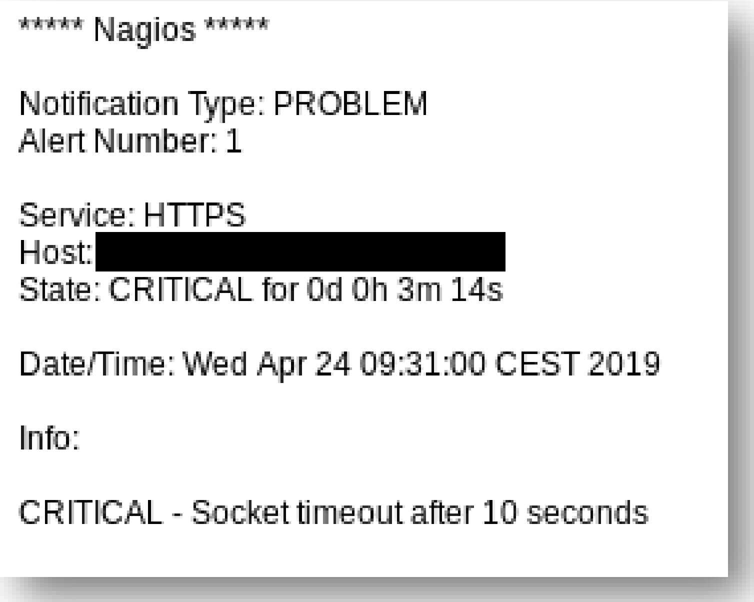
Knowledge: Analyzed information -> Says ‘How’ (Contextualized)

Information: Organized, … -> Says ‘What’ (Meaning)

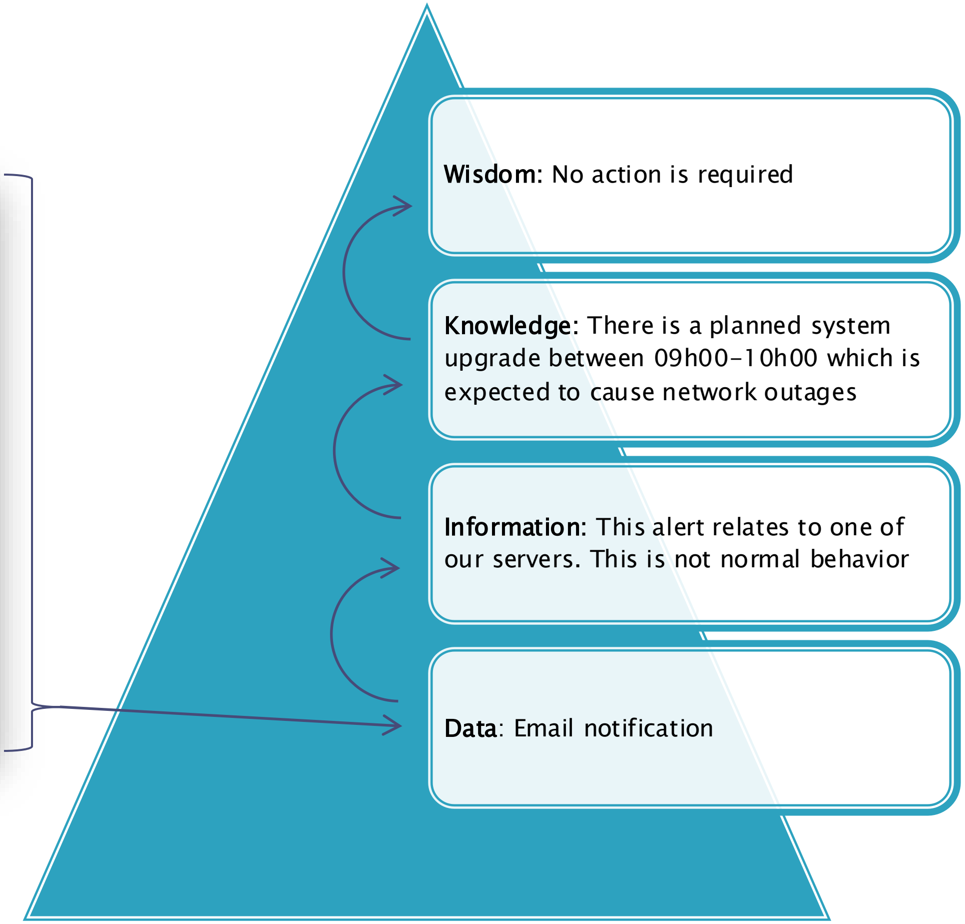
Data: Assigned values to things -> Says ‘Nothing’ (Raw)

From Data to Information to Knowledge (example)

- Let’s have a look at the notification below:



```
***** Nagios *****

Notification Type: PROBLEM
Alert Number: 1

Service: HTTPS
Host: [redacted]
State: CRITICAL for 0d 0h 3m 14s

Date/Time: Wed Apr 24 09:31:00 CEST 2019

Info:

CRITICAL - Socket timeout after 10 seconds
```



Wisdom: No action is required

Knowledge: There is a planned system upgrade between 09h00-10h00 which is expected to cause network outages

Information: This alert relates to one of our servers. This is not normal behavior

Data: Email notification

DIKW Pyramid

Types of data

- Two major data categories are:
  - **Qualitative data:** Description that refers to the quality of something (e.g., color, texture, feel of an item, …)
  - **Quantitative data:** Description of something in numbers (e.g., number, size, price of an item, …)
- Other categories (or sub-categories):
  - **Categorical data** represents categories or groups without any inherent order (e.g., used/unused, yes/no)
  - **Discrete data** consists of distinct, separate values (for instance, count of items, such as the number of books on a shelf which can only be whole numbers:1, 2, 3, etc)
  - **Continuous data** can take any value within a given range and has no gaps between possible values (e.g., measurements like height or weight, which can be measured to any level of precision: 1.23 meters, 65.7 kilograms)

Unstructured vs Structured data (1/3)

- Data for Humans:

  “we have 5 white used golf balls with a diameter of 43mm at 50 cents each”

  → Easy to understand for a human, but not for a machine

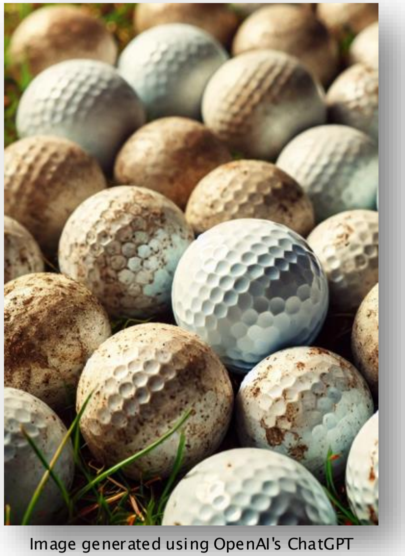


Image generated using OpenAI's ChatGPT

- The above sentence is what we call **unstructured** data
  - No fixed underlying structure

  → Likewise, PDFs and scanned images may contain information which is pleasing to the human-eye as it is laid out nicely, but they are not machine-readable in as-is form

CC BY

Unstructured vs Structured data (2/3)

- Data for machines: Hard to extract information from certain sources that humans find easy
  - E.g., Interpreting text that is presented as an image is a challenging task for a machine
    - It must be able to read and process the data
    - This means data needs to be structured, and presented in a machine-readable (standard) form
- E.g., CSV (Comma Separated Values) format

→ “quantity”, “color”, “condition”, “item”, “category”, “diameter (mm)”, “price per unit (EUR)”

→ 5, “white”, “used”, “ball”, “golf”, 43, 0.5

→ There are many more formats out there that are **structured** and machine readable e.g., https://opendatahandbook.org/guide/en/appendices/file-formats

CC BY

Unstructured vs Structured data (3/3)

- Unstructured DNS server log:

```
06-Jun-2020 07:55:34.142 info: client 192.168.100.105#58985
(_http._tcp.security.ubuntu.com): query: _http._tcp.security.ubuntu.com IN SRV +
(192.168.100.105)
```

…several lines with same format (as above) in a log file

A fix data structure and format might not always be based on a standard data structure and format!

- Structured example in JSON of DNS server log:

```
{
"EventReceivedTime": "2020-06-06 07:55:34",
"SourceModuleName": "dns_queries",
"SourceModuleType": "im_file",
"Date": "12-Mar-2019",
"QName": "example.com",
"QType": "A",
"RFlags": "+E",
"RemoteIP": "127.0.0.1",
"Severity": "info",
"Time": "07:17:09.816",
"EventTime": "2019-01-12 07:17:09“
}
```

Data Privacy (definitions)

- "The claim of individuals to determine for themselves when, how, and to what extent information about them is communicated to others" – Westin (1970)

- "Privacy as contextual integrity" – Nissembaum (2004)
  - Appropriate information flows that conform with contextual information norms

- Legal frameworks:
  - GDPR: transparency, purpose, proportionality, accountability
  - ECHR Art 8: "respect for private and family life, home and correspondence"

An Obligation

- **Users expectations (part of user experience)**
  - Users expect companies to request only the personal data needed to deliver the product or service
  - Users want to know who accesses their data, how and for which purpose
  - Users want their personal data to be handled with care and security

→ In short, they expect to stay in control of their personal data

- **…translated into a law (mandatory compliance)**
  - Article 25 European General Data Protection Regulation (GDPR):

    "the controller shall [...] implement appropriate technical and organisational measures […] which are designed to implement data-protection principles[...] in order to meet the requirements of this Regulation and protect the rights of data subjects."
  - Actually… "**Data Protection by design and by default"**

→ Organizations needs to cover both LAW requirements and USERS' expectations

Privacy by Design (PbD) 'foundational' principles

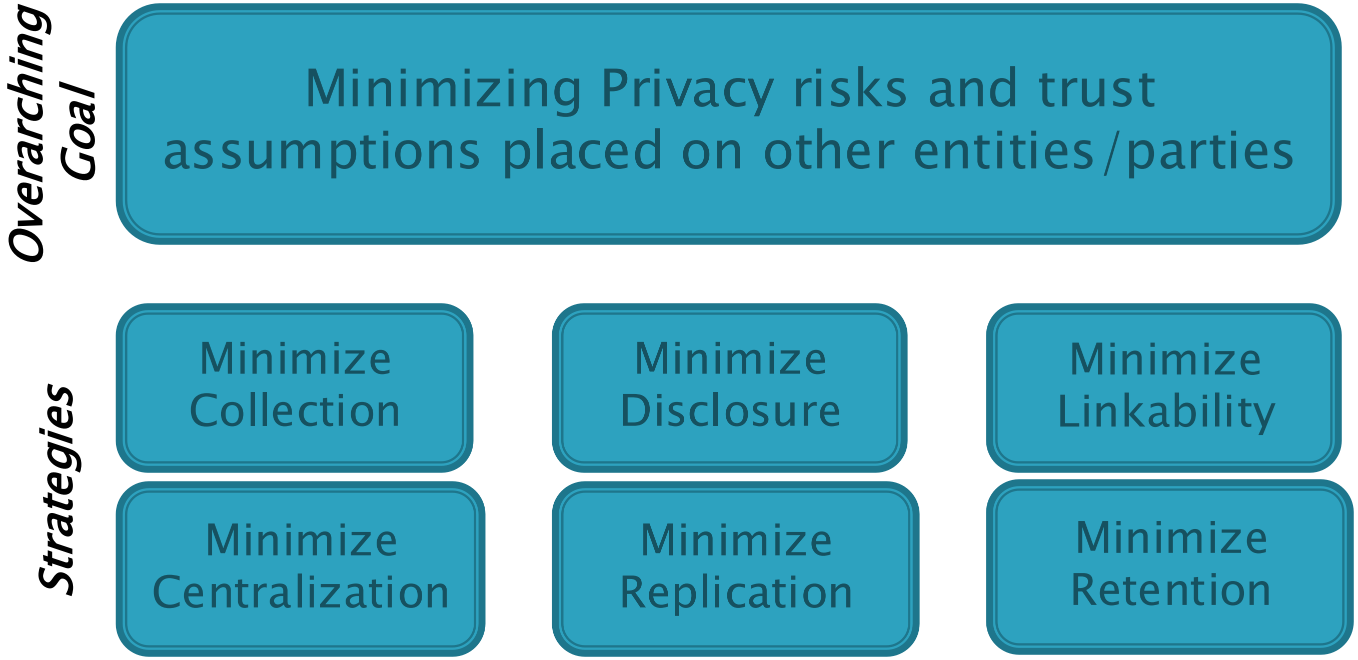
1. Proactive not Reactive; Preventive not remedial
2. Privacy as the default setting
3. Privacy Embedded into Design
4. Full functionality – Positive-Sum, not Zero-sum
5. End-to-end security – Full lifecycle protection
6. Visibility and transparency – keep it open
7. Respect for user privacy – keep it user-centric

Privacy by Design in Law, Policy and Practice
by Ann Cavoukian (1990s) [Former Information and Privacy Commissioner – Ontario, Canada]
Detailed version: https://www.ipc.on.ca/en/media/1826/download
Summarized version: https://www.ipc.on.ca/sites/default/files/legacy/2018/01/pbd-1.pdf

The European Data Protection Supervisor (EDPS) document, titled *Preliminary Opinion on Privacy by Design* (May 2018), outlines how the principles of PbD should be integrated into legal frameworks, systems, and organizational practices in Europe, in alignment with the requirements of the General Data Protection Regulation (GDPR)
Ref. https://www.edps.europa.eu/sites/default/files/publication/18-05-31_preliminary_opinion_on_privacy_by_design_en_0.pdf

Data privacy by design & by default

To answer the question, "Where shall we start?" we could take the following approach:



Overarching Goal

Minimizing Privacy risks and trust assumptions placed on other entities/parties

Strategies

Minimize Collection

Minimize Disclosure

Minimize Linkability

Minimize Centralization

Minimize Replication

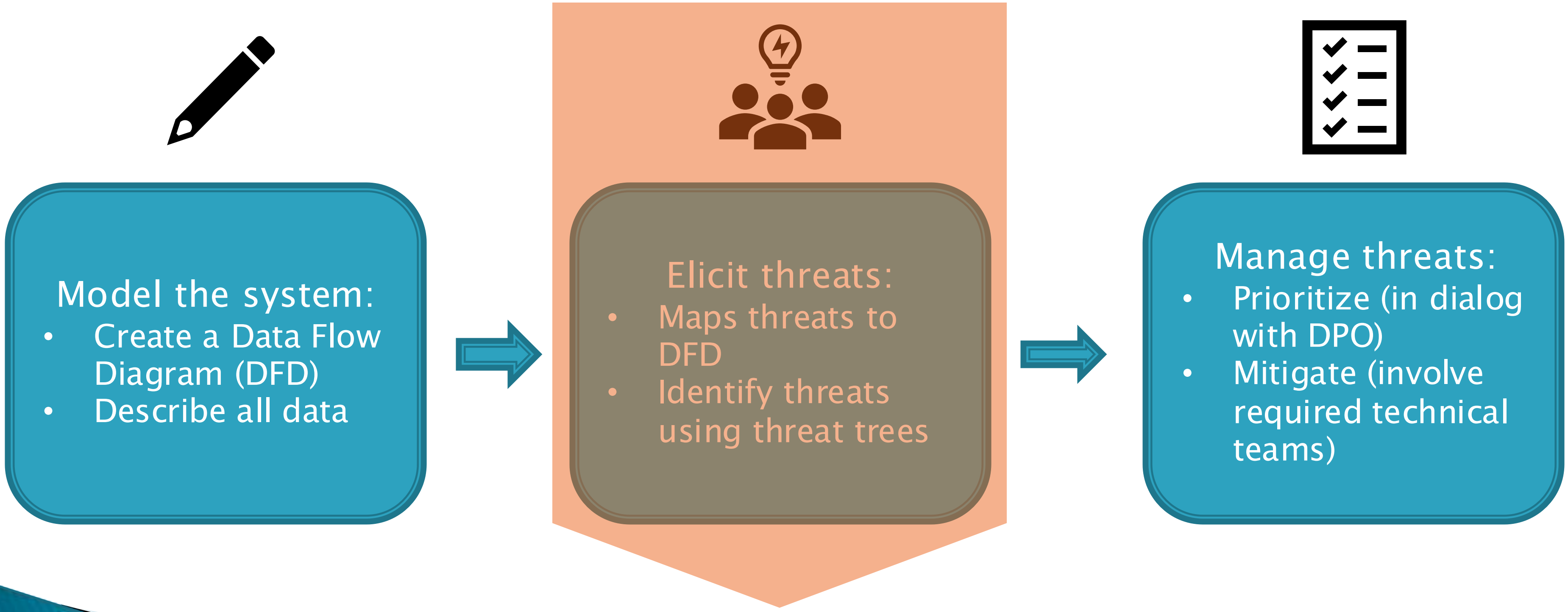
Minimize Retention

(Gurses S., Troncoso C., & Diaz C., 2015)

Great! but... how do we use these strategies?

LINDDUN (https://linddun.org/)

- Systematic threat assessment methodology



Model the system:

- Create a Data Flow Diagram (DFD)
- Describe all data

Elicit threats:

- Maps threats to DFD
- Identify threats using threat trees

Manage threats:

- Prioritize (in dialog with DPO)
- Mitigate (involve required technical teams)

Focus for this course (-> Let's check different categories for now to learn a **taxonomy or common vocabulary for identifying threats**)

LINDDUN – Threat categories (1 / 7)

Linkability (The ability to link two or more pieces of information related to an individual)

Identifiability (The ability to identify an individual from a dataset)

Multiple apples Linking to the same individual…

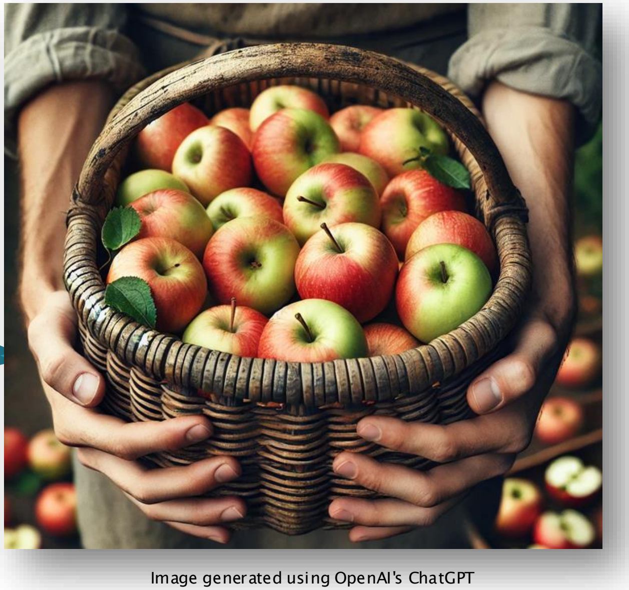


Image generated using OpenAI's ChatGPT

Image generated using OpenAI's ChatGPT

Red apple is easily identifiable without extra information

cc BY

LINDDUN – Threat categories (2/7)

Non-repudiation

(Ensuring that someone cannot deny the validity of their actions or transactions → We usually need the opposite i.e., Deniability, depending on the use-case)

Image generated using OpenAI's ChatGPT

Once you've handed over the item, you can't deny the transaction

Detectability

(The possibility to detect the presence of data or actions without their explicit presence)

Seems like a butterfly!

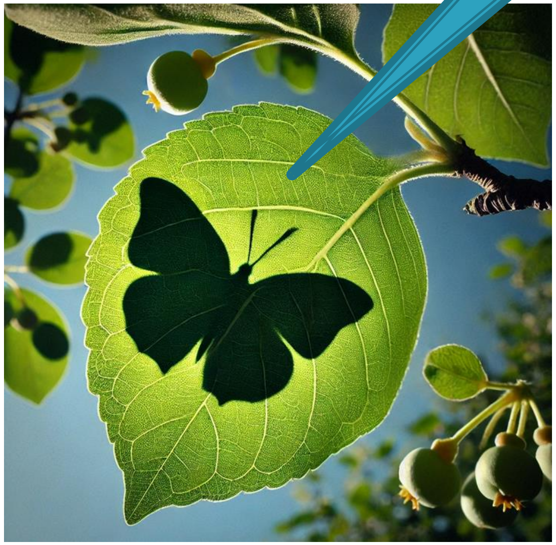
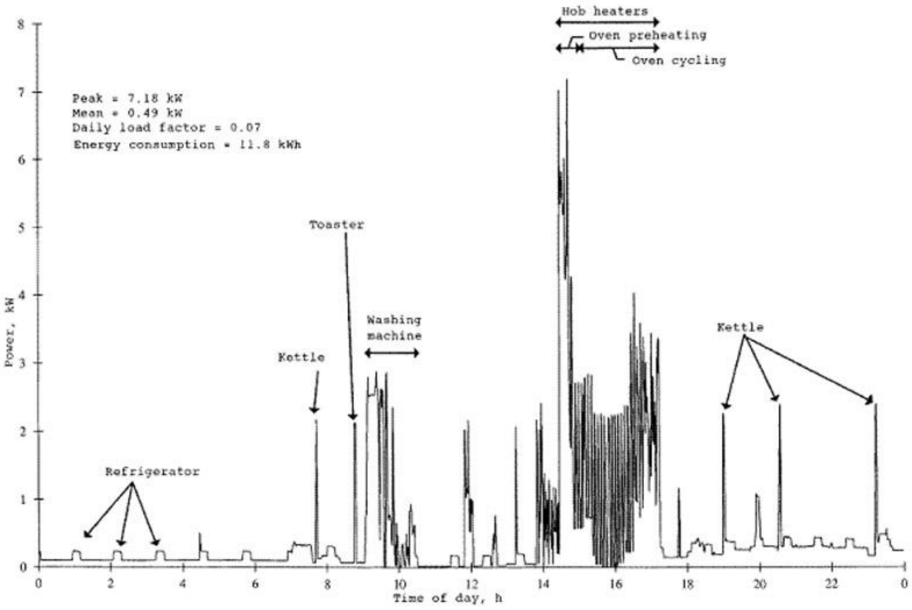


Image generated using OpenAI's ChatGPT

Case study 1:
Electricity smart metering system

- Smart energy meters record household consumption every 30 mins
- Privacy Risks:
  - Inference of sensitive personal attributes. E.g., health, work, …)
- Requirements:
  - Billing should be correct
  - Aggregate statistics per household or group should be available
  - Fraud/tampering detection

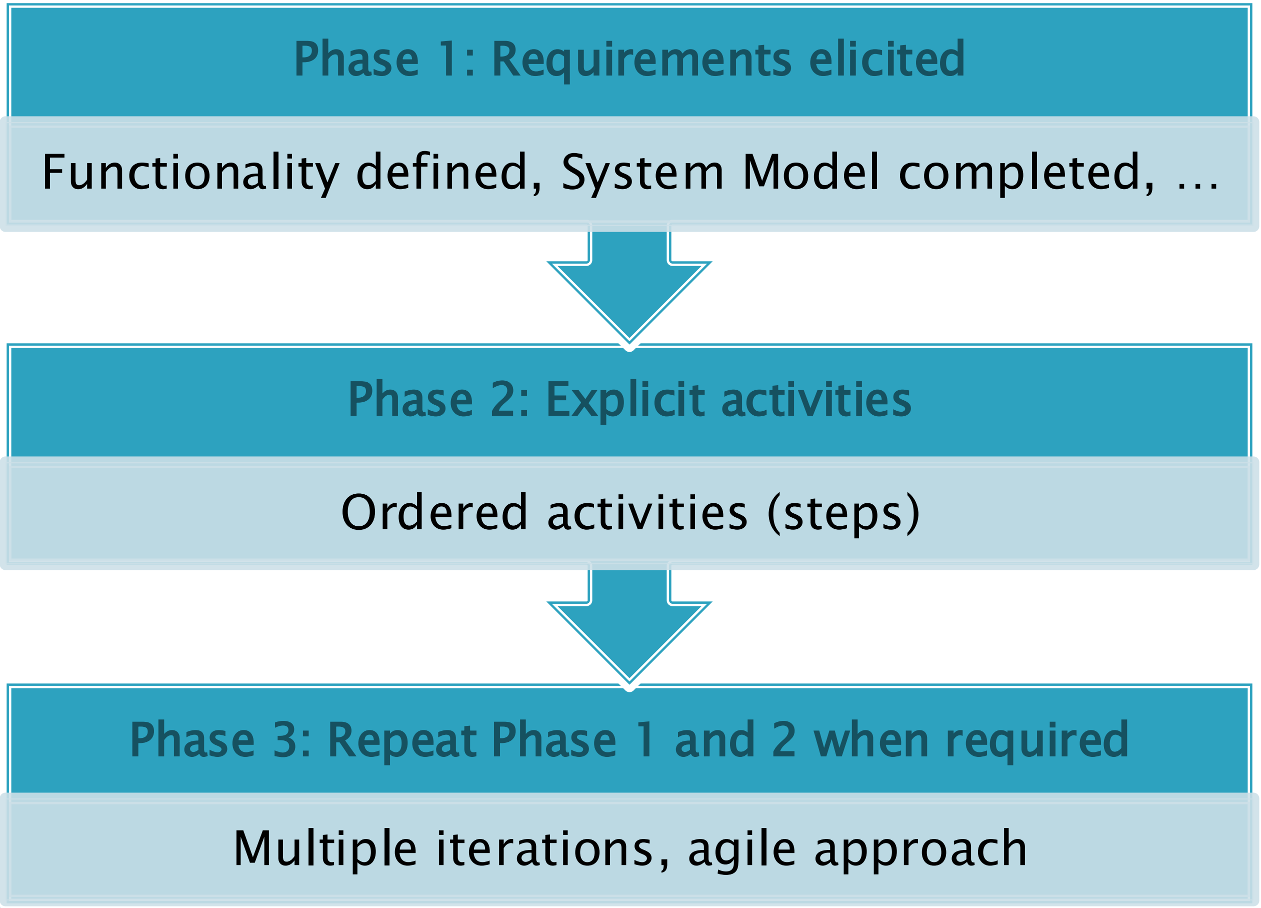


Smart Metering and Privacy: Existing Laws and Competing Policies [QUINN, E. L. – [SSRN eLibrary (2009)]]

Consumption snapshot (example)

Roadmap

- Case study 1: **Electricity smart metering system**



Phase 1: Requirements elicited

Functionality defined, System Model completed, …

Phase 2: Explicit activities

Ordered activities (steps)

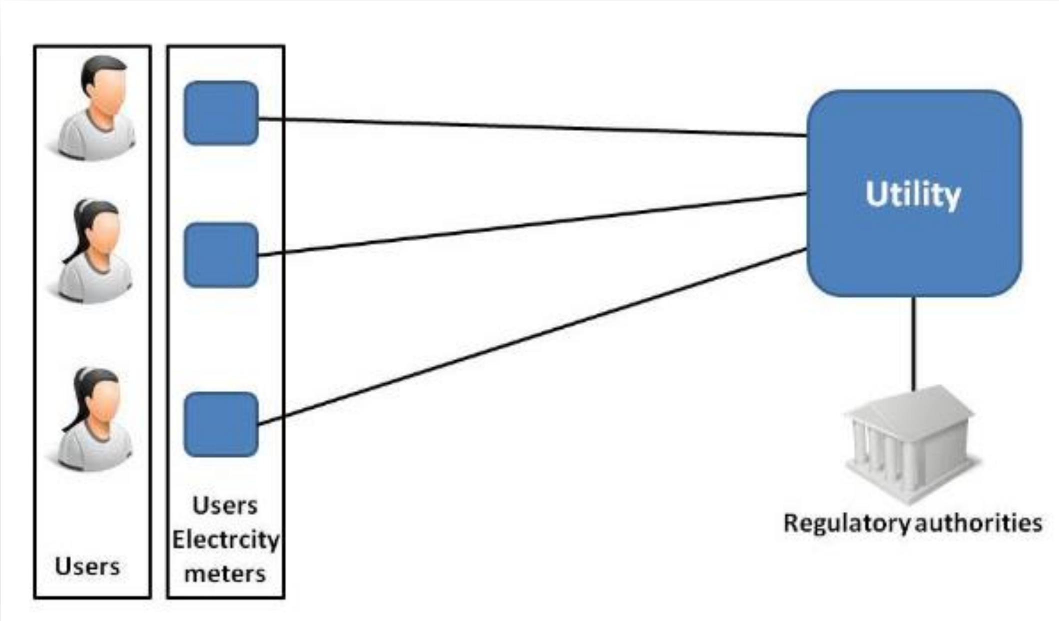
Phase 3: Repeat Phase 1 and 2 when required

Multiple iterations, agile approach

Phase 1: Prerequisites

- ✓ Functionality defined
- ✓ Basic system model(s)
- ✓ Service integrity requirements elicited

…



"Figure 1: Electricity smart metering system – Reference abstract architecture" (Gurses S., Troncoso C., & Diaz C., 2015)

LINDDUN – Threat categories (3/7)

Disclosure of information (Accidental or unintended disclosure of sensitive information)

Unawareness (Lack of awareness regarding the rights of individuals or the availability/sensitivity of data)

OF FRUITS
FREE BASKET
FREE BASKET
FREE BASKET OF FRUITS
IF YOU FILL AN ONLINE FORM
FREE

Image generated using OpenAI's ChatGPT

Really free?

Oops!

Image generated using OpenAI's ChatGPT

Or from a Distant country?

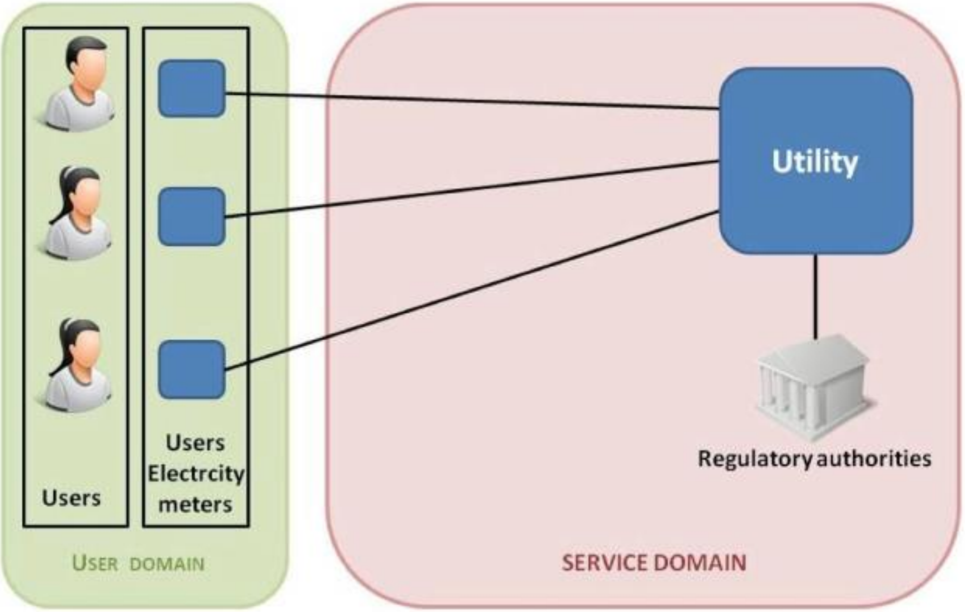
Image generated using OpenAI's ChatGPT

Non-compliance (Failing to comply with data privacy regulations)

Phase 2 → Activity 1: Classify Entities in domains

TEAM ACTIVITY

- User domain (trusted):
  - Components under the control of the user, e.g., user devices
- Service domain (non-trusted):
  - Components outside the control of the user, e.g., backend system (at provider side)

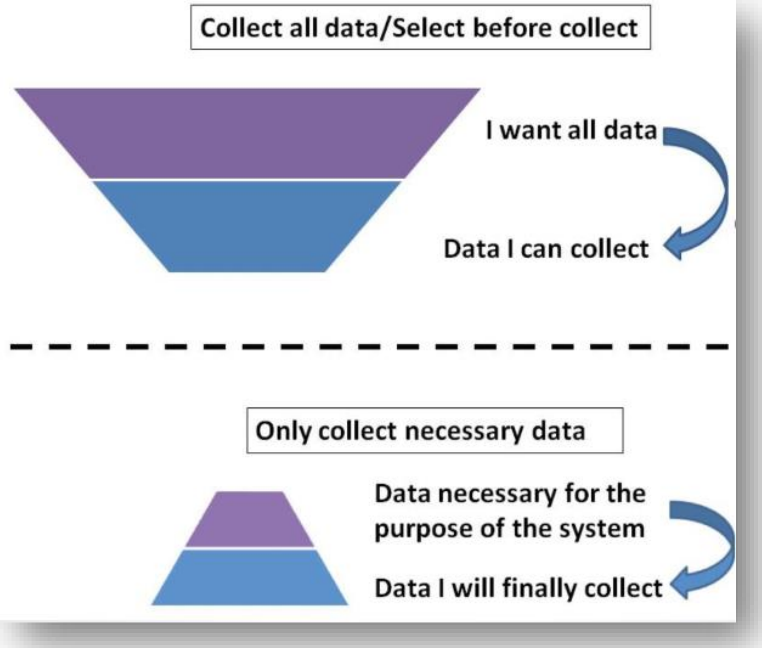


"Figure 2: Electricity smart metering system – User and Service Domains" (Gurses S., Troncoso C., & Diaz C., 2015)

Phase 2 → Activity 2: Identification of Necessary data

TEAM ACTIVITY

- User domain:
  - Personal data
  - Billing data
  - Consumption data
- Service domain:
  - Personal data
  - Billing data
  - Consumption data
  - Transaction logs



Collect all data/Select before collect
I want all data
Data I can collect
Only collect necessary data
Data necessary for the purpose of the system
Data I will finally collect

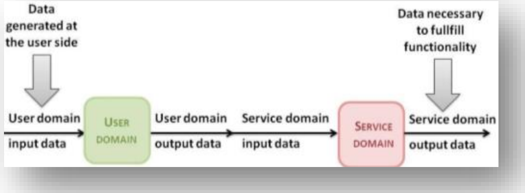
"Figure 3: Identification of necessary data: Typical vs. Experts approaches" (Gurses S., Troncoso C., & Diaz C., 2015)

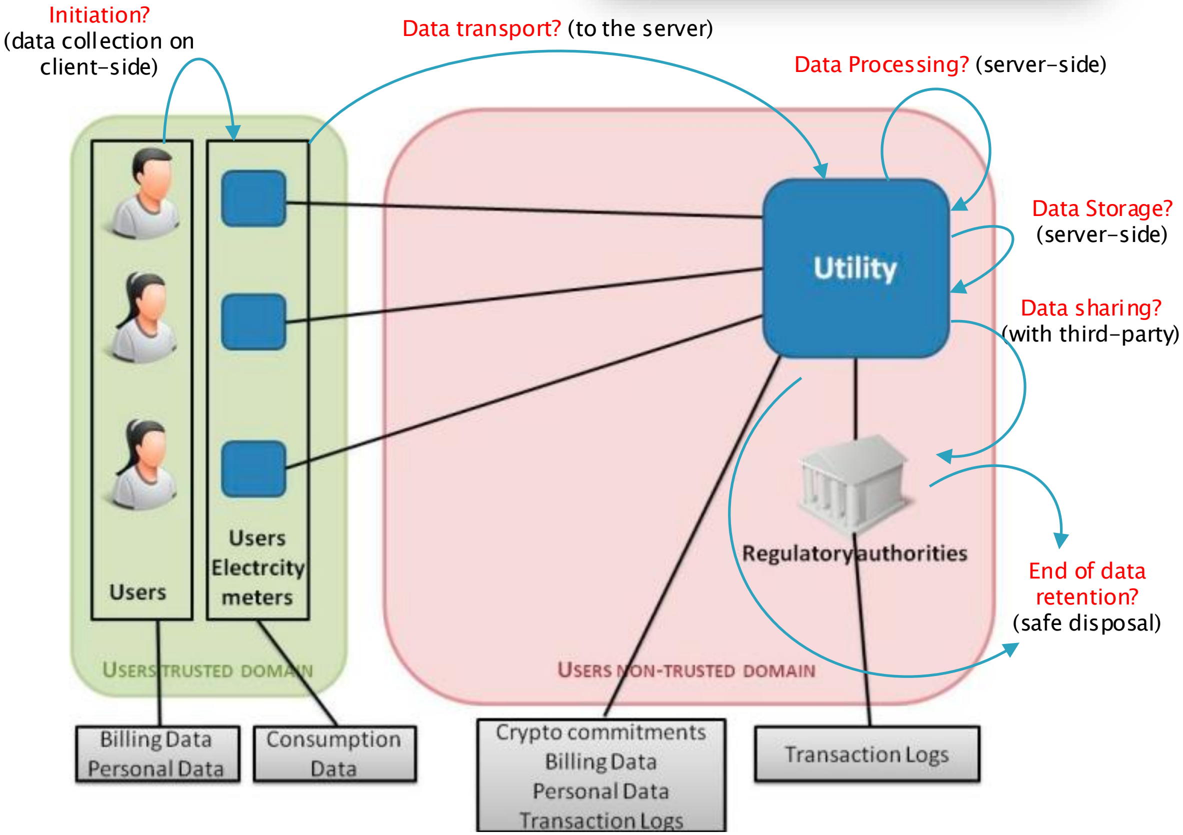
Changing the approach!

cc BY

Phase 2 → Activity 3: Distribution of data in the architecture

→ For One Full cycle


Data generated at the user side
Data necessary to fullfill functionality
User domain input data
USER DOMAIN
User domain output data
Service domain input data
SERVICE DOMAIN
Service domain output data


Initiation? (data collection on client-side)

Data transport? (to the server)

Data Processing? (server-side)

Data Storage? (server-side)

Data sharing? (with third-party)

End of data retention? (safe disposal)

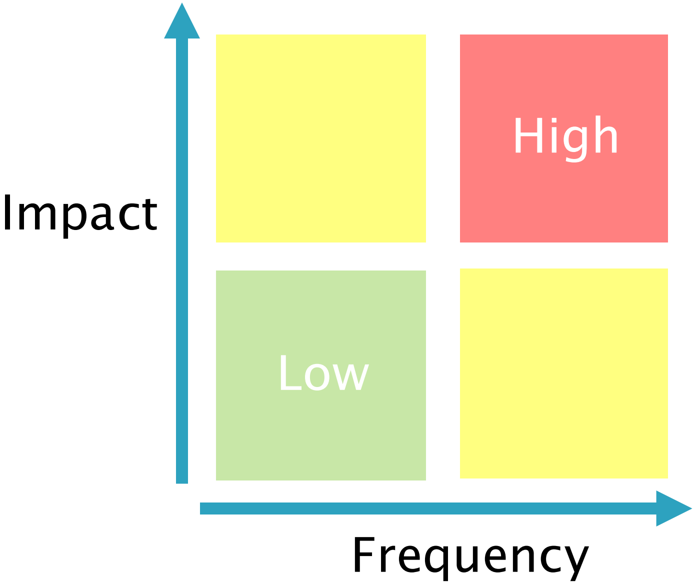
Figure 2 and 5: Electricity smart metering system - Data distribution in domains" (Gurses S., Troncoso C., & Diaz C., 2015)

TEAM ACTIVITY

3.1 Threat modeling

- Systematically thinking about negative scenarios Or thinking about 'What can go wrong?'
- Other factors: data lifecycle, maintenance, etc.

3.2 Risk analysis



Phase 2 → Activity 4: Select technological solutions

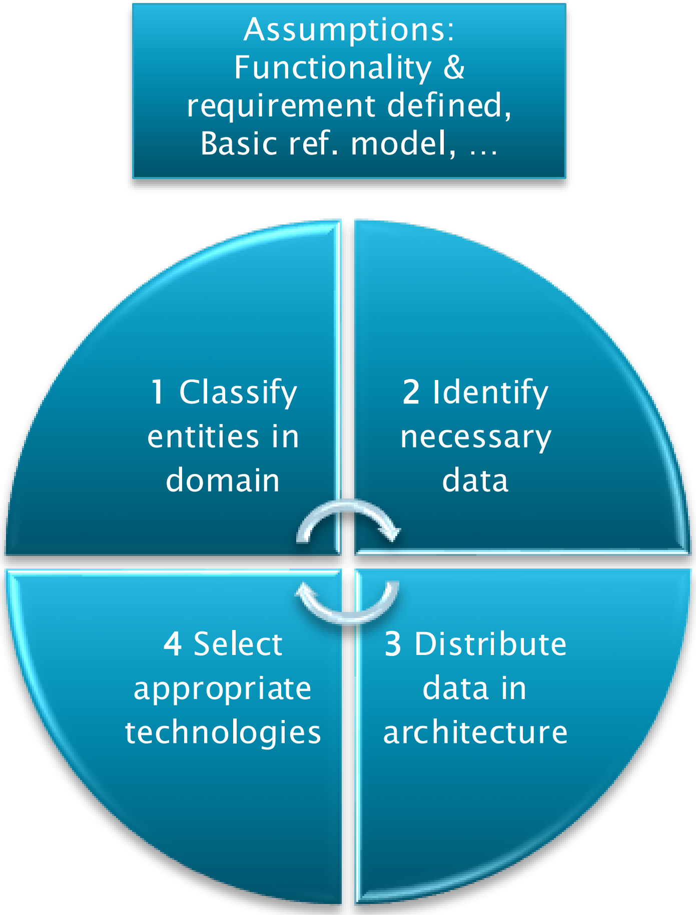
TEAM ACTIVITY

- Address threats or risks with high priority first!
- Use Privacy Enhancing Technologies (PETs)
  - A set or a combination of tools and techniques designed to protect personal data and ensure privacy e.g.,
    - SSL/TLS for protection of data-at-transit
    - Encryption for data-at-rest (storage)
    - Masking of data (anonymization and pseudonymization)
    - Advanced privacy-preserving protocols, e.g., homomorphic encryption, zero-knowledge proofs (ZKP), multi-party computation (MPC), and trusted execution environments (TEE), etc.
- Focus is to: Apply the six strategies i.e., minimize collection, disclosure, linkability, centralization, replication, and retention of data

cc BY

Phase 3: Repeat when required

Remember the Overall Goal, and 6 strategies!


Assumptions: Functionality & requirement defined, Basic ref. model, …
1 Classify entities in domain
2 Identify necessary data
3 Distribute data in architecture
4 Select appropriate technologies

Covering full life-cycle using Agile approach

QUIZ…

No fixed methodology

- Take **privacy considerations** and perform **risk management** at all levels of project management
- Assert data subject rights and integrate **appropriate controls to mitigate privacy risks** at all stages of development
  - E.g., requirements, specification, implementation, testing, deployment, maintenance
- Focus on raising **Transparency** of service/product:
  - Make your Privacy policy/Terms of service easy-to-understand
  - Take clear user consent (e.g. no pre-ticked boxes) with no shady tactics (e.g., clickbait)
  - …

Project: Kick-off (1)

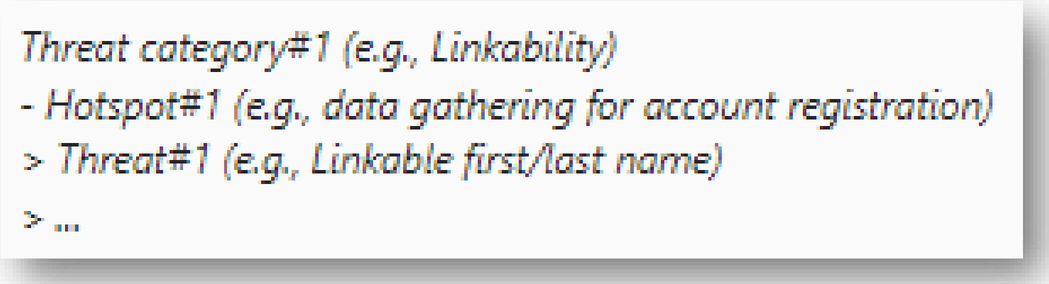
1. Form groups (minimum 3, maximum 5)
2. Select project after reading specifics and viewing corresponding dataset (MS Teams group → Files → Projects)
3. Initiate **individual report**
   - **Section 1**: Explain dataset (1-page max.) e.g., data types (per column), attributes (properties, values), format and other specifications (direct vs indirect identifiers?)
   - **Section 2**: Create a (high-level) system reference diagram [or a Data Flow Diagram (if you prefer)]
     - No need to go into details, a high-level diagram covering all major system components would suffice

NOTE: Collaborate on brainstorming as a group, but avoid copying and pasting your group members' work.

Project: Kick-off (2)

Individual report

- **Section 3:** Perform threat modeling using LINDDUN threat categories (https://www.linddun.org/linddun-go-categories)
  - Use your System diagram as a reference
  - For each LINDDUN category, write an applicable threat for every hotspot (i.e., basic data processing operation, such as ‘collection’, ‘disclosure’, ‘usage’)
  - Format:



Threat category#1 (e.g., Linkability)

- Hotspot#1 (e.g., data gathering for account registration)

> Threat#1 (e.g., Linkable first/last name)

> ...

  - You must list down **at least** 10 data privacy threats!

NOTE: Collaborate on brainstorming as a group, but avoid copying and pasting your group members' work.

Deadline: See ‘Teams’ Assignment section

Lecture 1 ends here

- Course Slides: Go to MS Teams:
  ‘Data Privacy by Design’
  → “Files section”

- Send your questions by email OR via direct message using MS Teams

- Thank You!

CC BY