École Pour l'Informatique et les Techniques Avancées (EPITA)

MSc – Oct 2021

Course instructor: M Salman Nadeem
Information Security Analyst
ContactOffice/Mailfence – Brussels

salman@mailfence.com

ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

Data Privacy by Design (PbD)

Course schedule (tentative)

Date & Time	No.	Topics	Duration (in hours)
22/10/2021 *	**1**	**Data & its types, Information & knowledge, Introduction to Data Privacy by Design (PbD)**	**3 hours**
29/10/2021 *	2	DPbd Case studies, Data privacy risks & solutions	3 hours
05/11/2021 *	3	Privacy Enhancing Technologies (PET's)	3 hours
12/11/2021 *	4	General Data Protection Regulation (GDPR), PbD and GDPR	3 hours
19/11/2021 *	5	Open session, Putting it all together, Quiz, Final project presentation	3 hours
* Check 'Zeus' for exact timing of each class		***Total Lecture (hours)***	***15***

Evaluation: 10% Class attendance + 10% Class participation + 30% Class/home exercises + 50% Final Evaluation

Notes & Collaboration

- MS Teams Channel:
  ‘Data Privacy by Design – Fall 2021’
  - Course specific channel to collaborate
    - Code to join: **9w023c0** **(you are required to join)**
  - Will be used for:
    - Course related announcements,
    - Share course slides/material,
    - Carry out assignments & quizzes
- Course Mindmap:
  - To allow better organization and easy refreshing
  - Access link (read-only):
    https://www.mindomo.com/mindmap/60f6d856c480464ab9f113f60e2fc986

Lecture 1 Outline

- **Setting up the scene**
  - **Data & its types**
  - ***Class exercise 1***
  - **Data privacy, secrecy & control**
  - **What can be done**

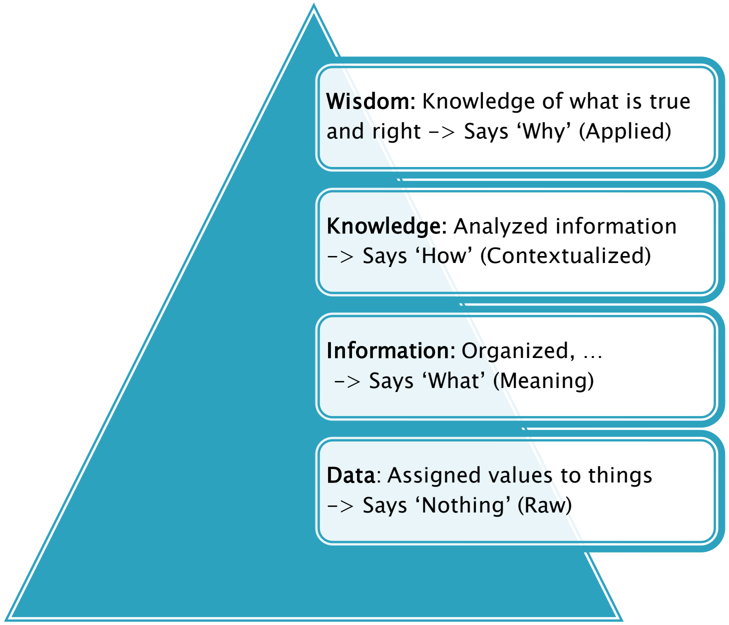
- Introduction to Data Privacy by Design (PbD)
  - An obligation
  - PbD principles, goals & strategies
  - Assumptions & activities
  - Case Studies & *Class exercise 2*
  - Take away!

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)

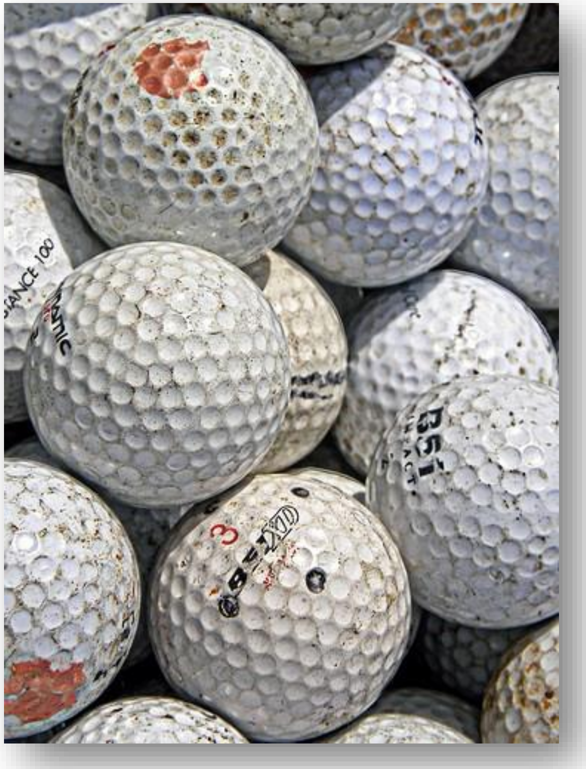
Knowledge: Analyzed information –> Says ‘How’ (Contextualized)

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

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

Let us take an example

- Check picture on the right:
  - What can we say about these?
    - Golf balls -> Sport
    - Category of sport: golf -> Taxonomy
    - No. of Golf balls: ~15
    - More to it?
      - Color: White
      - Condition: Used
      - …



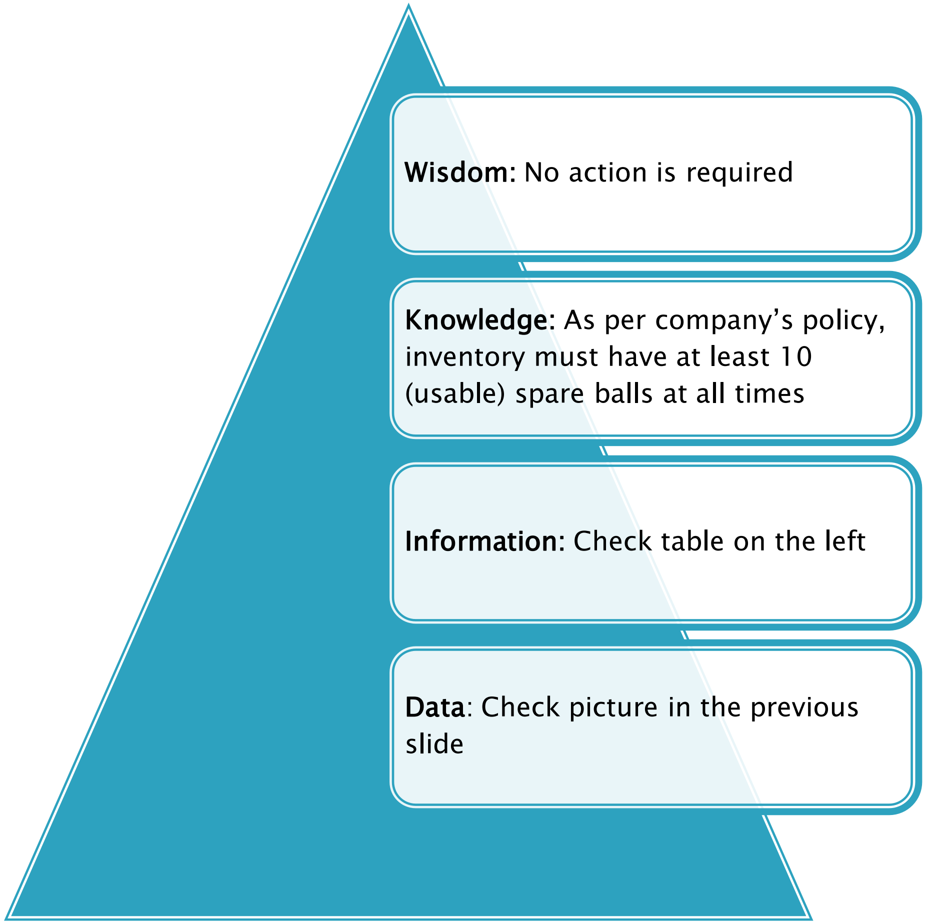
Golf balls (CC) by Kaptain Kobold on Flickr

From Data to Information to Knowledge

- Let’s organize that data:

‘XYZ’ Golf club inventory: Golf balls	
Color	White
Category	Sport
Condition	Used
Diameter	43mm
Price (per ball)	1 EUR
No. of balls	~15

Table form



Wisdom: No action is required

Knowledge: As per company’s policy, inventory must have at least 10 (usable) spare balls at all times

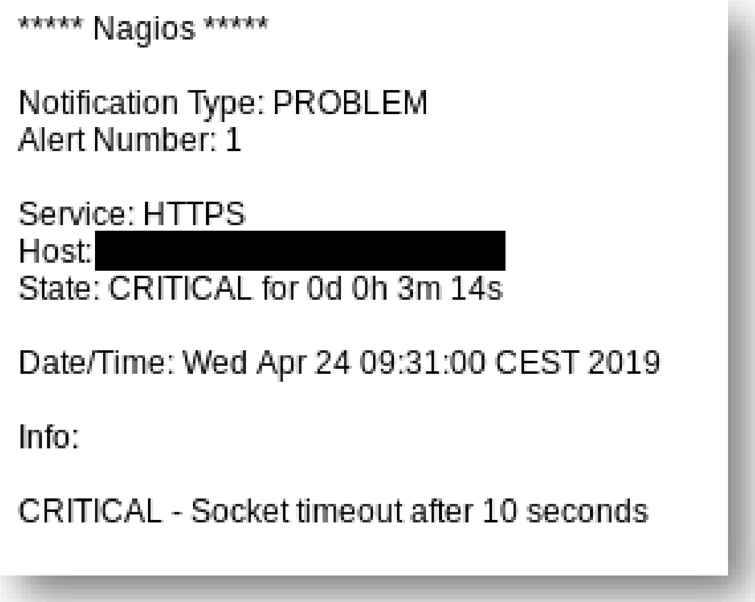
Information: Check table on the left

Data: Check picture in the previous slide

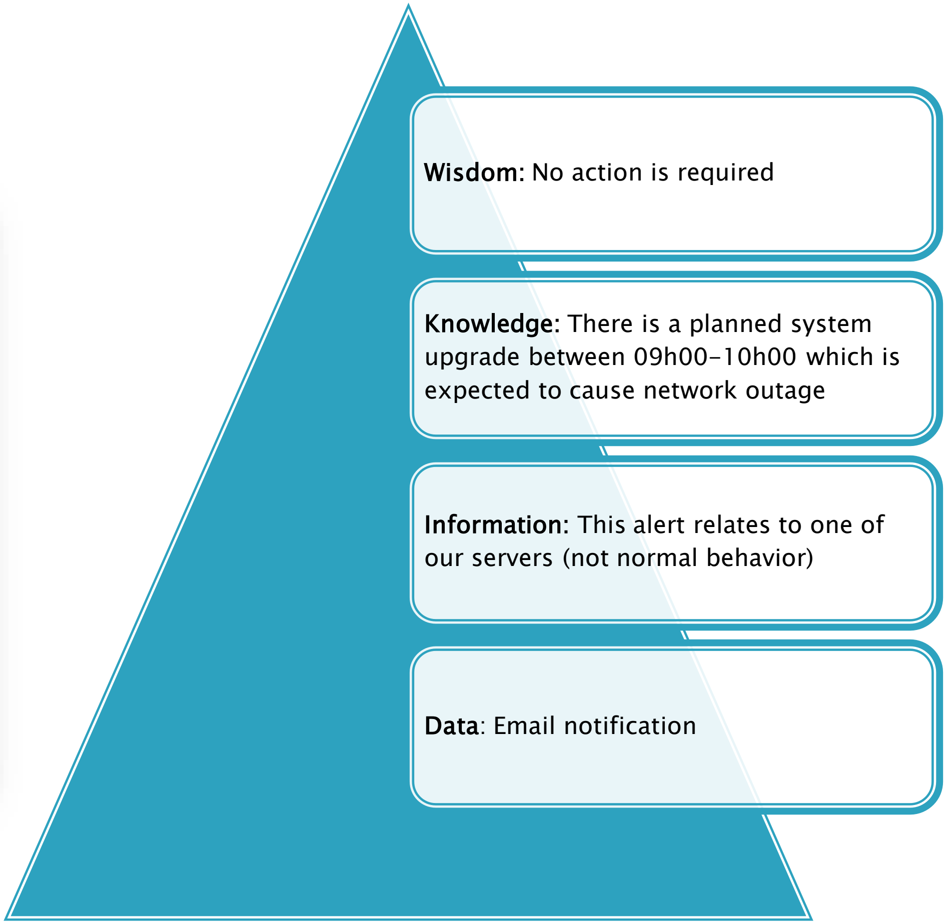
DIKW Pyramid

From Data to Information to Knowledge (another 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 outage

Information: This alert relates to one of our servers (not normal behavior)

Data: Email notification

DIKW Pyramid

Types of data (in light of previous example)

- Two major data categories are:
  - **Qualitative data:** Description that refers to the quality of something (e.g., color, texture of golf ball, …)
  - **Quantitative data:** Description of something in numbers (e.g., number of golf balls, size, price, …)
- Other categories:
  - **Categorical data:** Categorizing items (e.g., used, …)
  - **Discrete data:** Number of data having gaps b/w it (e.g., count of golf balls, it can't be 0.3)
  - **Continuous data:** Numerical data with a continuous range (or no gaps), can be any value (e.g., size of golf balls, …)

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
- 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

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
    - This means it needs to be structured, and presented in a machine-readable form, in order to be processed
- E.g., CSV (Comma Separated Values) format

-> “quantity”, “color”, “condition”, “item”, “category”, “diameter (mm)”, “price per unit (AUD)”
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

Unstructured vs Structured data (3/3)

- Unstructured DNS server log (example)

```
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)
```

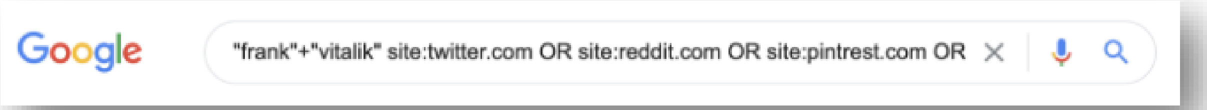
- Structured example in JSON of DNS server log (example)

```
{
"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“
}
```

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

Class exercise: 1(a)

1. Use any search engine of your choice
2. Perform OSINT on **your identity**:
   - Use google dorks (on your name/email ID/GSM no./…) and perform text, image/reverse-image, document, etc based searches



"frank"+"vitalik" site:twitter.com OR site:reddit.com OR site:pintrest.com OR

   - Social media:
     - Check if your username/email is taken, using aggregator service E.g., https://namechk.com/, piple.com, social-searcher
     - Look with your home connection public IP address
     - …
   - Leaked databases (HIBP, …), …
   - Forums: Reddit, 4chan, …
   - Common tools: osintframework.com, magma.lavafeld.org/guide/osint-sources.html
3. Look for information that was put there **without your consent!**
4. **Create a text file** (lastname_firstname.txt) and write down:
   1. Data you found that was put there **without** your consent!
   2. What would you prefer to do now (e.g., ask site owner to put it down? Let it remain online?)

Be creative!

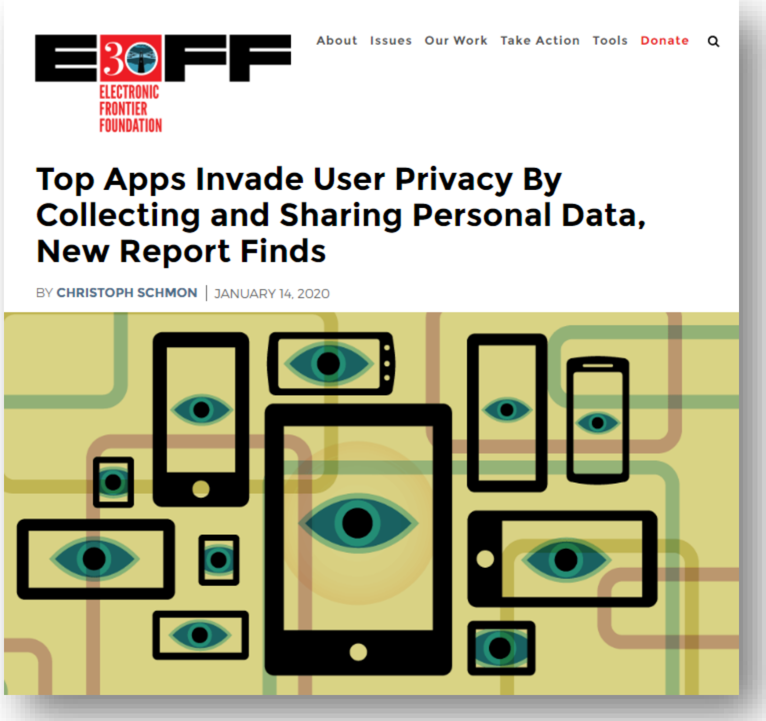
Class exercise: 1(b)

- Go to:
  1. Open *Microsoft Teams -> Data Privacy by Design (Teams) -> Files*
  2. Download the 'Class_exercise_1b'
- **Fill 5 entries** (instant chatting apps) in the table, and answer questions on the next page for each of the entry
- After completing the exercise:
  - Export it as a pdf file, and name it with sender's "lastname_firstname"
  - Upload it to the 'Teams' assignment section (using your EPITA account).

Deadline: See 'Teams' Assignment section

What did you learn?

- Observations:
  - Companies have a lot of data on you
  - Shared data can be mapped from different data points to induce a more informative and useful result
  - Deciding how much data you should share and with which company is an **important** decision
  - …
- Data therefore is a **commodity,** that fuels the business models of tech industry



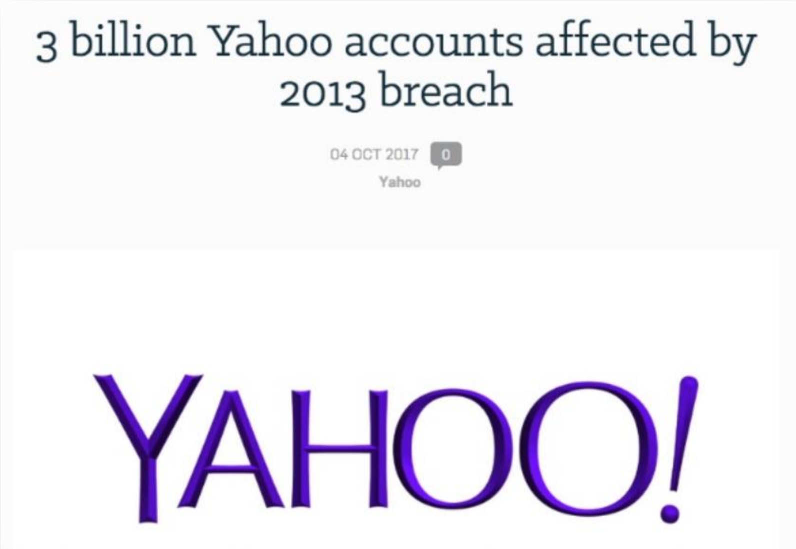
Top Apps Invade User Privacy By Collecting and Sharing Personal Data, New Report Finds

BY CHRISTOPH SCHMON | JANUARY 14, 2020

Source: https://www.eff.org/deeplinks/2020/01/new-report-exposes-adtech-invading-our-privacy

Question: Does it require protection?

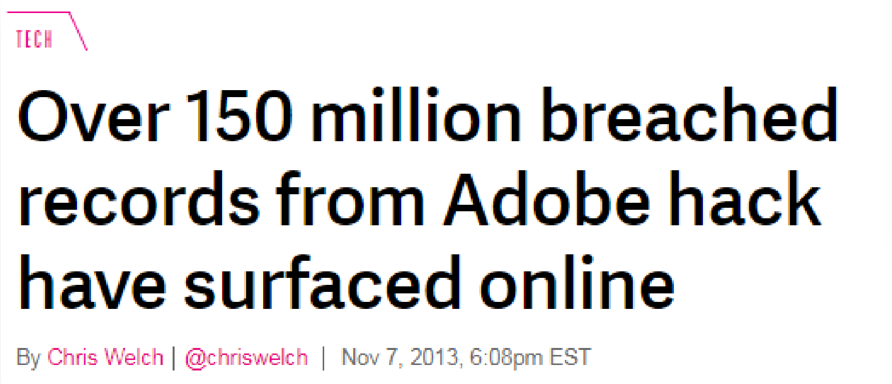
Present situation!



3 billion Yahoo accounts affected by 2013 breach

04 OCT 2017 0

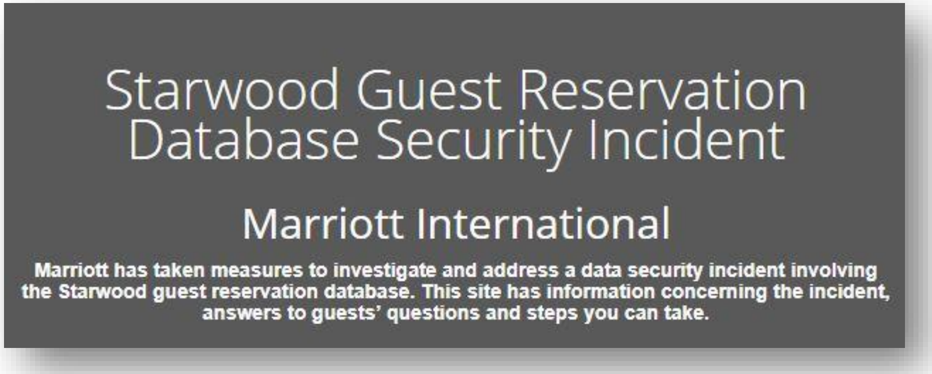
Yahoo



TECH

Over 150 million breached records from Adobe hack have surfaced online

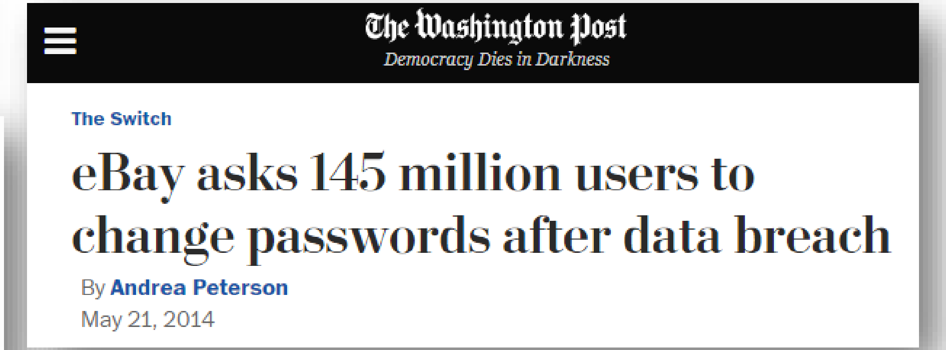
By Chris Welch | @chriswelch | Nov 7, 2013, 6:08pm EST



Starwood Guest Reservation Database Security Incident

Marriott International

Marriott has taken measures to investigate and address a data security incident involving the Starwood guest reservation database. This site has information concerning the incident, answers to guests' questions and steps you can take.



The Washington Post

Democracy Dies in Darkness

The Switch

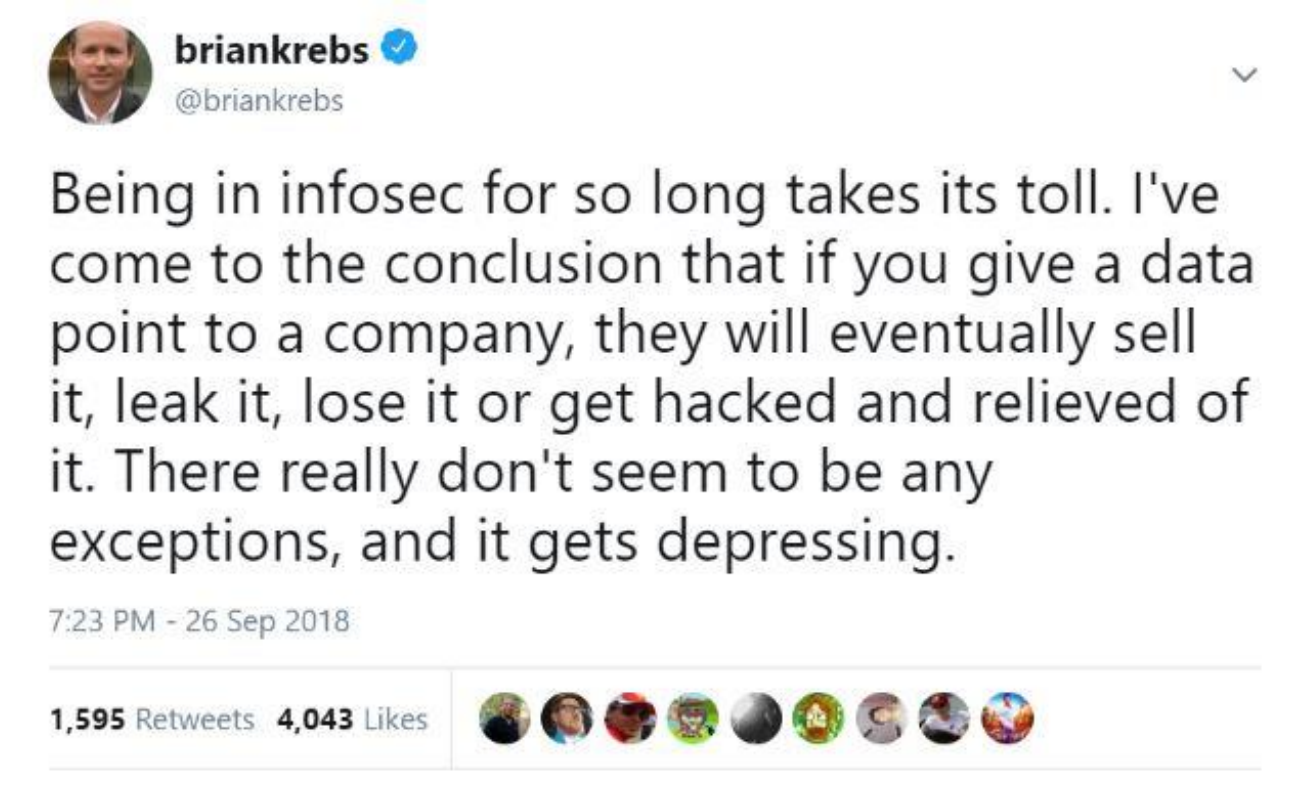
eBay asks 145 million users to change passwords after data breach

By Andrea Peterson

May 21, 2014

For extensive list: **https://en.wikipedia.org/wiki/List_of_data_breaches**

What does that mean?



briankrebs
@briankrebs

Being in infosec for so long takes its toll. I've come to the conclusion that if you give a data point to a company, they will eventually sell it, leak it, lose it or get hacked and relieved of it. There really don't seem to be any exceptions, and it gets depressing.

7:23 PM - 26 Sep 2018

1,595 Retweets 4,043 Likes

Privacy, Secrecy & Control

Privacy ≠ Secrecy

Privacy is not about having something to hide

(a subjective approach)

Privacy measure: Giving full Control to the user

(an objective approach)

What can be done?

- Establish countermeasures that we can take to avoid becoming a victim of data privacy/security incidents
- There is no ‘silver bullet’
  - The goal is to avoid as much risk as possible
  - Thus, any counter-approach should be based on a holistic view e.g., let’s put some hats on:

1. **Protecting yourself** ***(as an end-user)***

2. **Protecting others** ***(as an organization)***

- Spreading awareness *(passing on the message)*

Lecture 1 Outline

- Setting up the scene
  - Data & its types
  - *Class exercise 1*
  - Data privacy, secrecy & control
  - What can be done

- **Introduction to Data Privacy by Design (PbD)**
  - **An obligation**
  - **PbD principles, goals & strategies**
  - **Assumptions & activities**
  - **Case Studies &** ***Class exercise 2***
  - **Take away!**

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

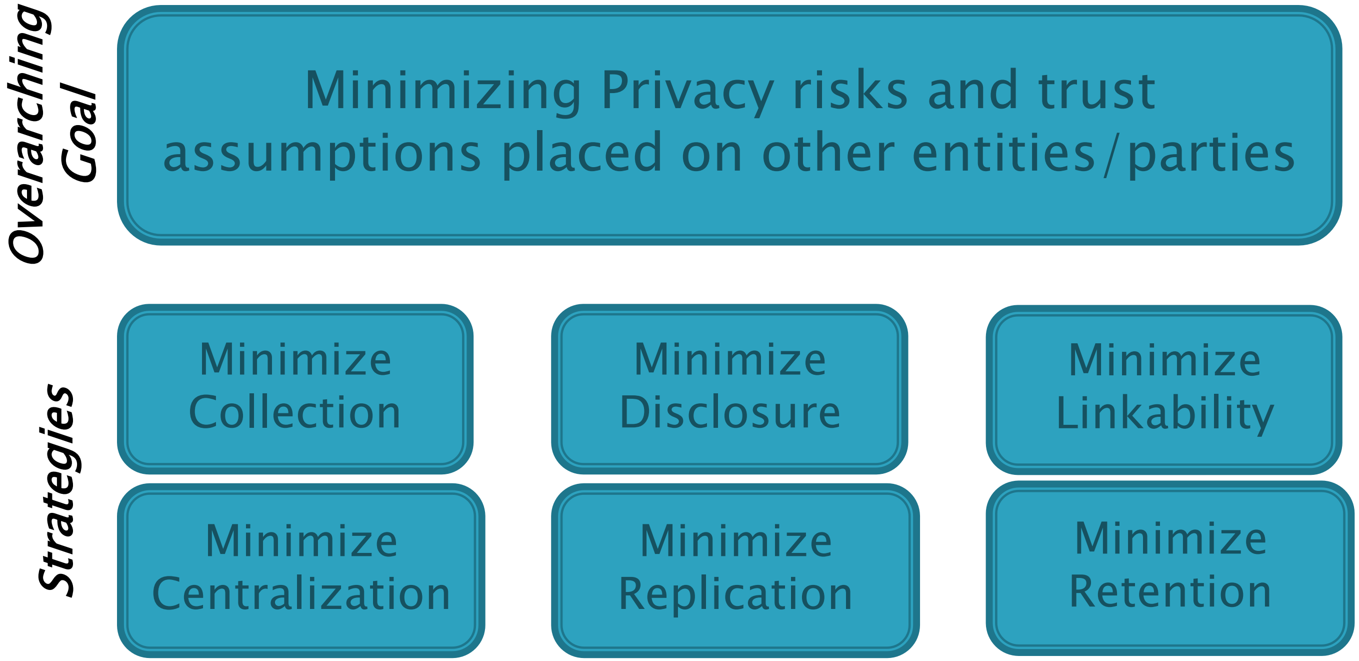
Privacy by Design (PbD) ‘foundational’ principles

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

Referred from the white paper by Ann Cavoukian
– Former Information and Privacy Commissioner – Ontario, Canada

Data privacy by design & by default

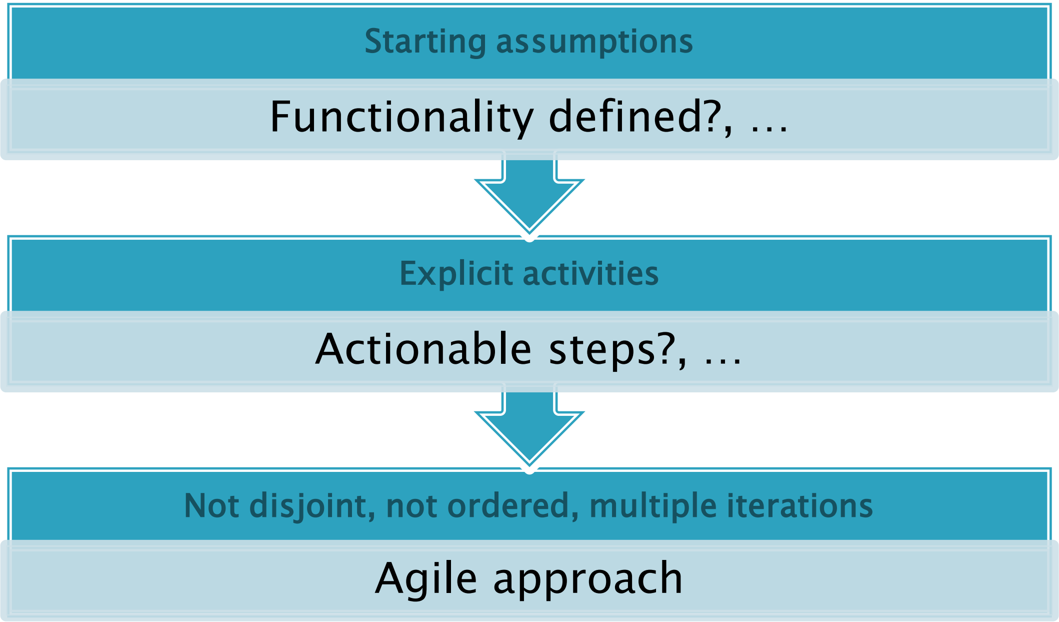
So what is ‘Data by design and by default’ really means?



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

Assumptions & Activities

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



Starting assumptions

Functionality defined?, …

Explicit activities

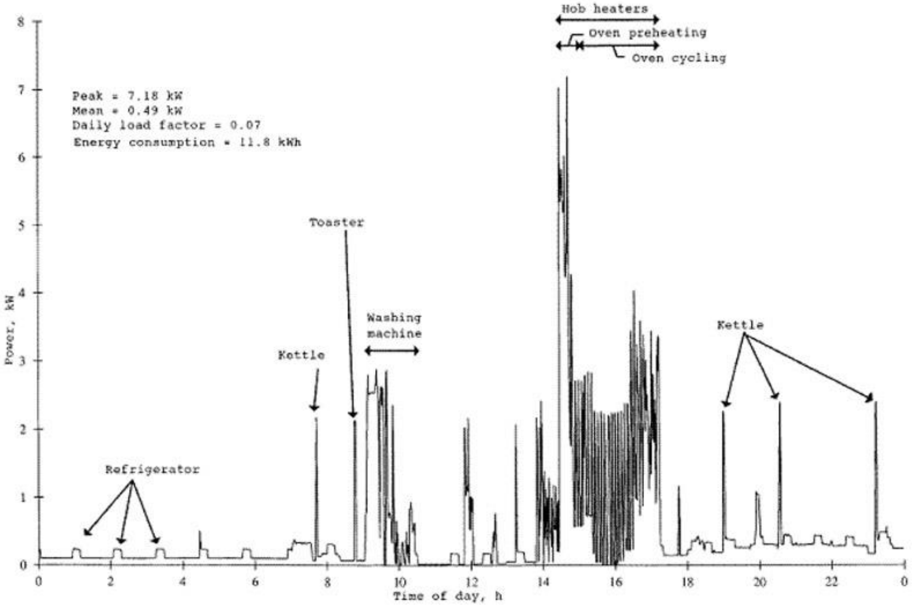
Actionable steps?, …

Not disjoint, not ordered, multiple iterations

Agile approach

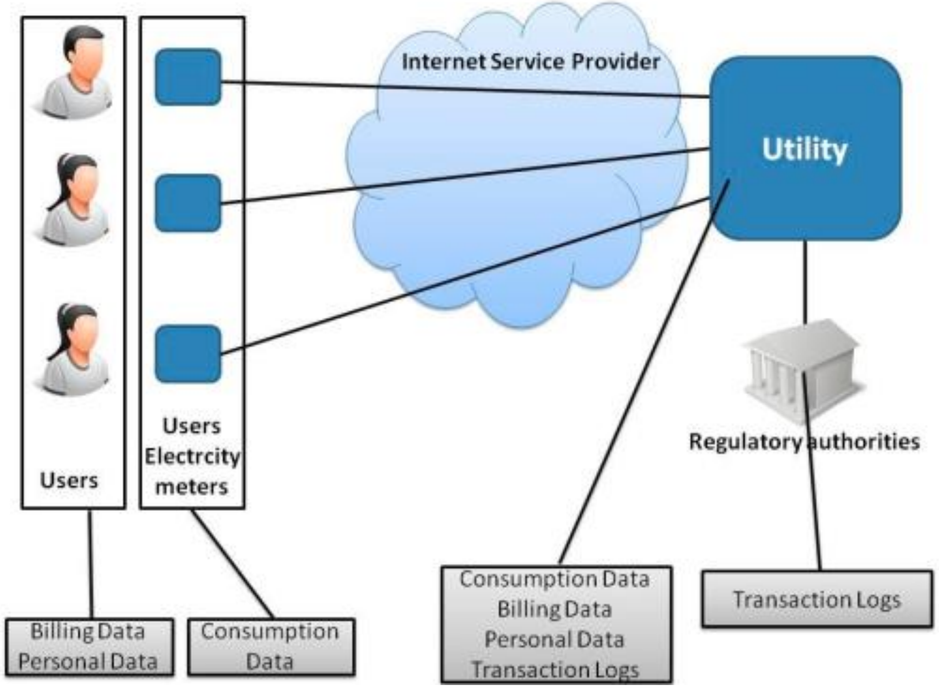
Case study 1: Electricity smart metering system

- 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



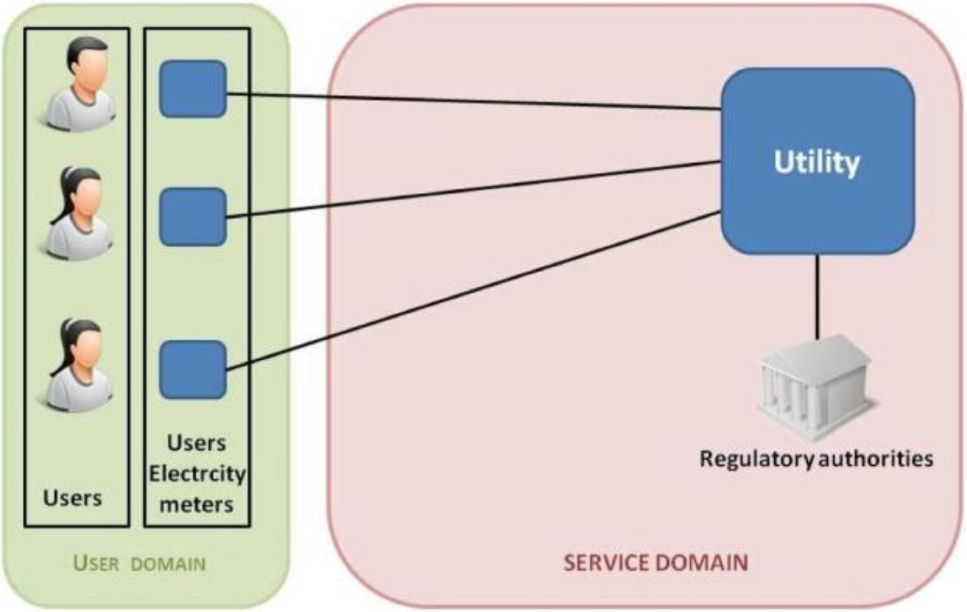
Starting Assumptions

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



Activity 1: Classify Entities in domains

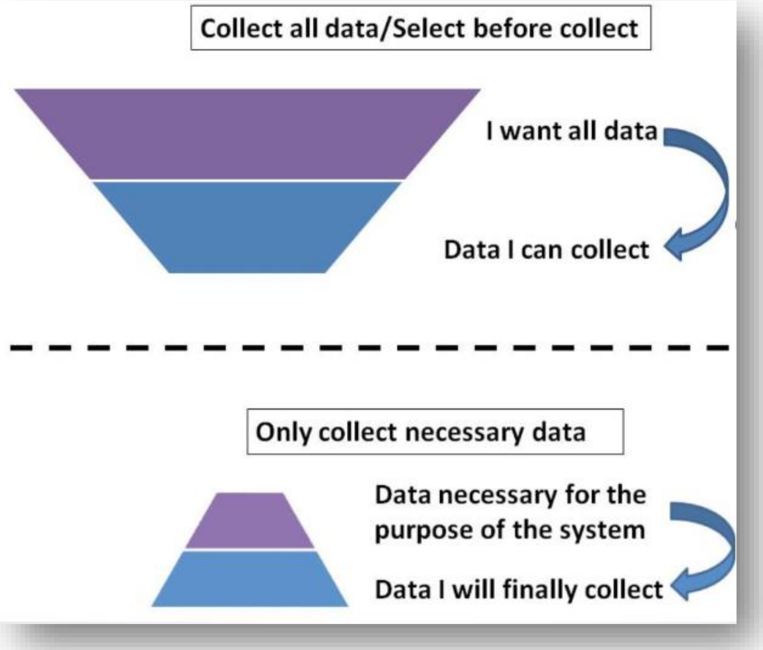
- User domain (trusted):
  - Components under the control of the user, e.g., user devices
- Service domain (non-trusted):
  - Components outside the control of the user, e.g., backend system (at provider side)



TEAM ACTIVITY

Activity 2: Identification of Necessary data

- 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

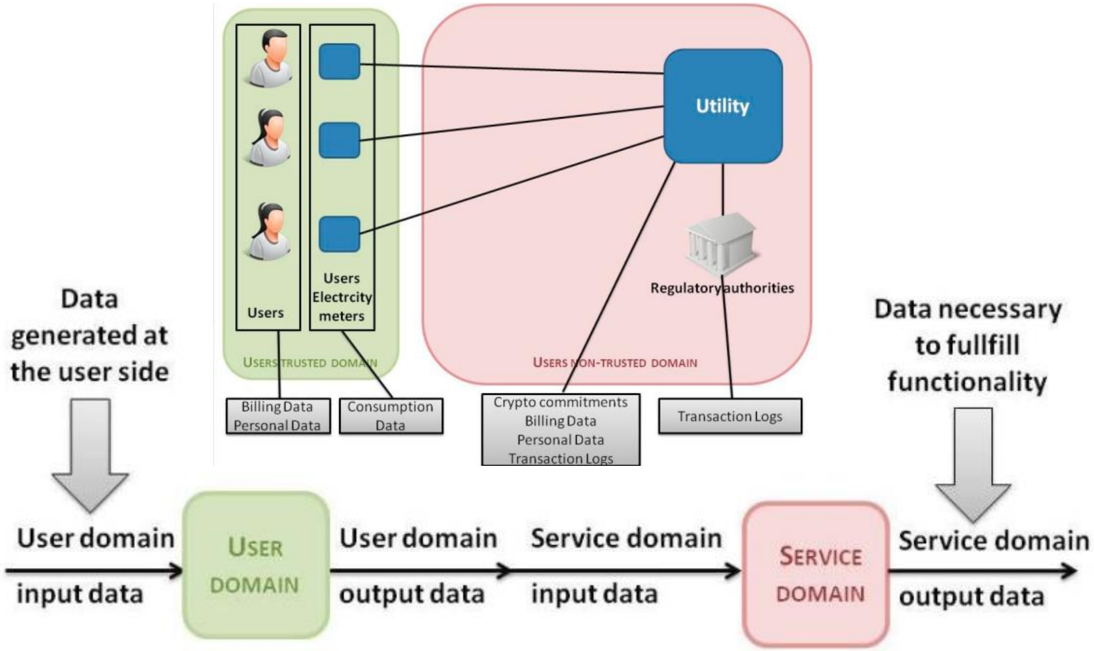
Data necessary for the purpose of the system

Data I will finally collect

Changing the approach!

TEAM ACTIVITY

Activity 3: Distribution of data in the architecture



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

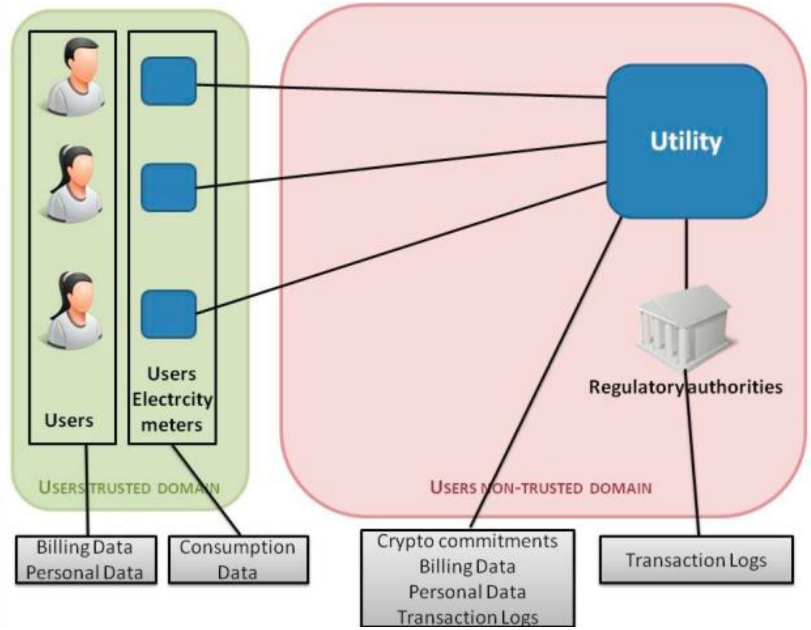
Data consumption

- **Threat modeling**
  - Systematically thinking about negative scenarios
  - Techniques:
    - STRIDE (for Security)
    - STRIPED (for Privacy & Security)
    - LINDDUN (for Privacy)
    - Other factors: data lifecycle, maintenance, etc.
- **Risk analysis**
  - Likelihood vs Impact

TEAM ACTIVITY

Activity 4: Select technological solutions (Patterns)

- Keeping as much data as possible out of the service domain while satisfying service integrity requirements e.g.,
  - Not sending the data (local computations)
  - Encrypting the data
  - Advanced privacy-preserving protocols
  - Obfuscate the data
  - Anonymize the data
  - …
- The considered strategies are to minimize collection, disclosure, linkability, centralization, replication, and retention



TEAM ACTIVITY

Quiz time!

- N/A

Case Study 2: European Electronic Toll Service (EETS)

- Defined functionality:
  - Pay according to road use: time, distance, road type, …
- Requirements:
  - Privacy & integrity risks to be mitigated:
    1. Third party access to traffic/location data of driver.
    2. Abuse of traffic data by authority performing the billing (location data cannot be easily anonymized).
  - The provider needs to know the final fee to charge
  - The provider must be reassured that this fee is correctly computed and users cannot commit fraud

Note: Location as a means to compute above points -> not intrinsic

Class exercise 2: activity

<u>Form basic information model & perform the 4 activities:</u>

1. **Classify entities in domains**
2. **Identify necessary data for providing the service**
3. **Distribute data in architecture**
4. **Select technological solutions**

Upload it to the 'Teams' assignment section (using your EPITA account).
Deadline: See 'Teams' Assignment section

Lecture 1 ends here

- Course Slides:

Open **Microsoft Teams -> Data Privacy by Design (Teams) -> Files**

- Email your questions, concerns and assignments to:

salman@mailfence.com

- Thank You!