RAIS

Real-time Analytics for Internet of Sports

Marie Curie European Training Network

PERSONAL DATA STORAGE

Ha Xuan Son – ***University of Insubria, Italy***

Outline

1. **Personal Data Storage**
    - Personal data
    - Conditions
    - Examples
2. **Approaches**
    - openPDS
    - Machine learning algorithms
    - Blockchain
3. **Reference**

Personal Data Storage

1. Personal data
2. Conditions
   - Transparency
   - Access
   - Control
   - Transfer
3. Examples
   - MiDATA
   - Smart Disclosure

Personal Data Storage

- **Personal data [4]**
  - Any information that relates to an identified or identifiable living individual.
  - The different pieces of information, which collected together can lead to the identification of a particular person.
  - *Example*:
    - a name and surname;
    - a home address;
    - an email address such as name.surname@company.com;
    - an identification card number;
    - location data (for example the location data function on a mobile phone)*;
    - an Internet Protocol (IP) address;
    - ...
- **Personal data services**
  - It is the services to let an individual store, manage and deploy their key personal data in a highly secure and structured way [1]

Conditions - Transparency

- Require the new approaches that help individuals understand how and when data is being collected, how the data is being used and the implications of those actions.
- The user needs
  - a better understanding of the overall value exchange so that they can make truly informed choices.
  - to exercise choice and control, especially where data uses most affect them.

=> It require that design and usability must lie at the heart of the relationship between *individuals* and the *data* generated by and about them.

- ➢ **Organizations** need to **engage and empower** individuals more effectively and efficiently.

Conditions - Access

- All stakeholders (Organizers – Individuals) must take appropriate steps to secure data from **accidental release**, **theft**, **unauthorized access**, and **misuse**.
- Individuals should be provided with **access to simple tools**
  - Enable them to **either understand/readable [diagram, chart, symbol, color, etc.] (Not raw data)**
  - Enable them to **set the policy** to be applied to the use of data
  - Enable them to **change that selection** over time

Conditions - Control

- The individuals have **full control** over their data (Creatable, Openable, Readable, **Erasable**)

  => There are legitimate reasons why individuals and organizations may want to **delete data**.

  Because the retention of data involves both **costs and risks** including it being breached or misused. => **DELETE**

Conditions - Transfer

- **Transfer of data** to an individual who has the opportunity to forward it to third parties or other providers.
- The data that is seen as **particularly sensitive** in some **contexts** can in other **contexts** be **freed**
- In some cases, **failure to transfer data** (for example, to diagnose a medical condition) can lead to bad outcomes.

=> **"How can protect the personal data when transfer from one context to another one"** --<*system-centric*>

Examples – MiDATA [2]

- MiDATA is a United Kingdom Government initiative.
- Assessing how to give people their **personal data** in a format that is **safe to pass onto third parties**, such as price comparison sites.
- MiDATA gives consumers access to **their personal data** in a **portable and electronic format**.

Examples – MiDATA [2]

- MiDATA seeks to give consumers **access to their transaction data** in a way that is machine readable, portable and secure.
- The **ambition** is to **rebalance the current asymmetry** that exists between **business and consumers**.

HOWEVER

- The companies have been complying with the law in a way that **did not realize** the real **potential value of that right to data**,
  - For example, a citizen could request personal data and it would arrive the mail **weeks later** at a cost of a **few dozen pounds**.

Examples – Smart Disclosure [3]

- Smart Disclosure is a USA Government initiative.
- Smart disclosure typically take the form of providing individual consumers of **goods and services** with ***direct access*** to **relevant information and data sets**.
- Smart disclosure is when a **private company** or **government agency** provides a person with *periodic access* to his or her **own data** in open formats that enable them to **easily put the data to use**.
- **Smart disclosure is "a new tool that helps provide consumers with greater access to the information they need to make informed choices,"**

Approaches

1. openPDS
2. Machine learning algorithms
3. Blockchain

Approaches - openPDS

[5] De Montjoye, Y. A., Wang, S. S., Pentland, A., Anh, D. T. T., & Datta, A. (2012). On the Trusted Use of Large-Scale Personal Data. IEEE Data Eng. Bull., 35(4), 5-8.

[6] De Montjoye, Y. A., Shmueli, E., Wang, S. S., & Pentland, A. S. (2014). openpds: Protecting the privacy of metadata through safeanswers. PloS one, 9(7), e98790.

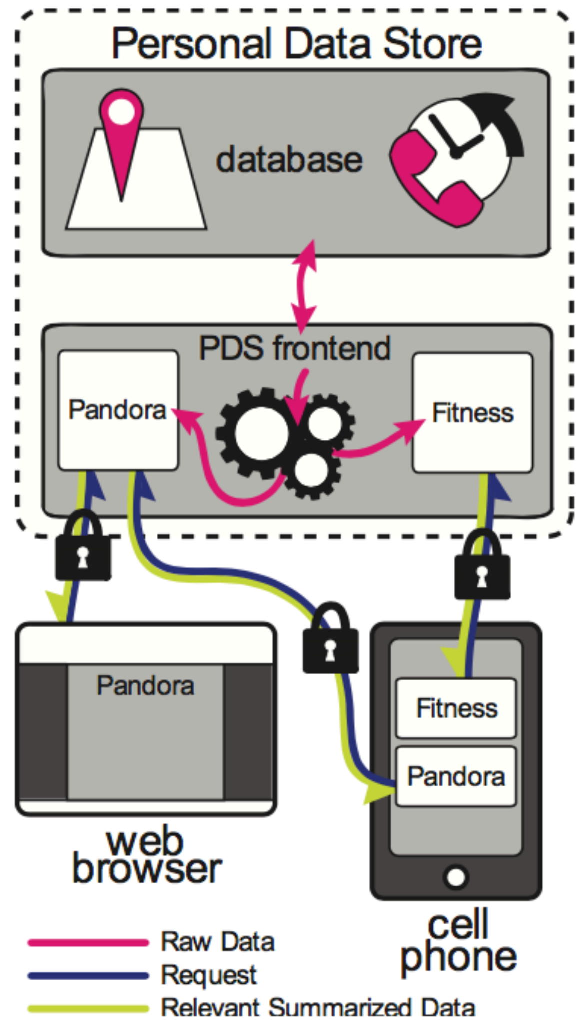
Approaches - openPDS

- **openPDS allows**
  - **User's data**
    - collect, store (Cloud)
    - give fine-grained access
      - by sharing anonymous answers, not raw data
  - **Metadata (SafeAnswers)**
    - collect, store (Cloud)
    - give fine-grained access
      - by asking questions whose answers (instead of anonymizing metadata)

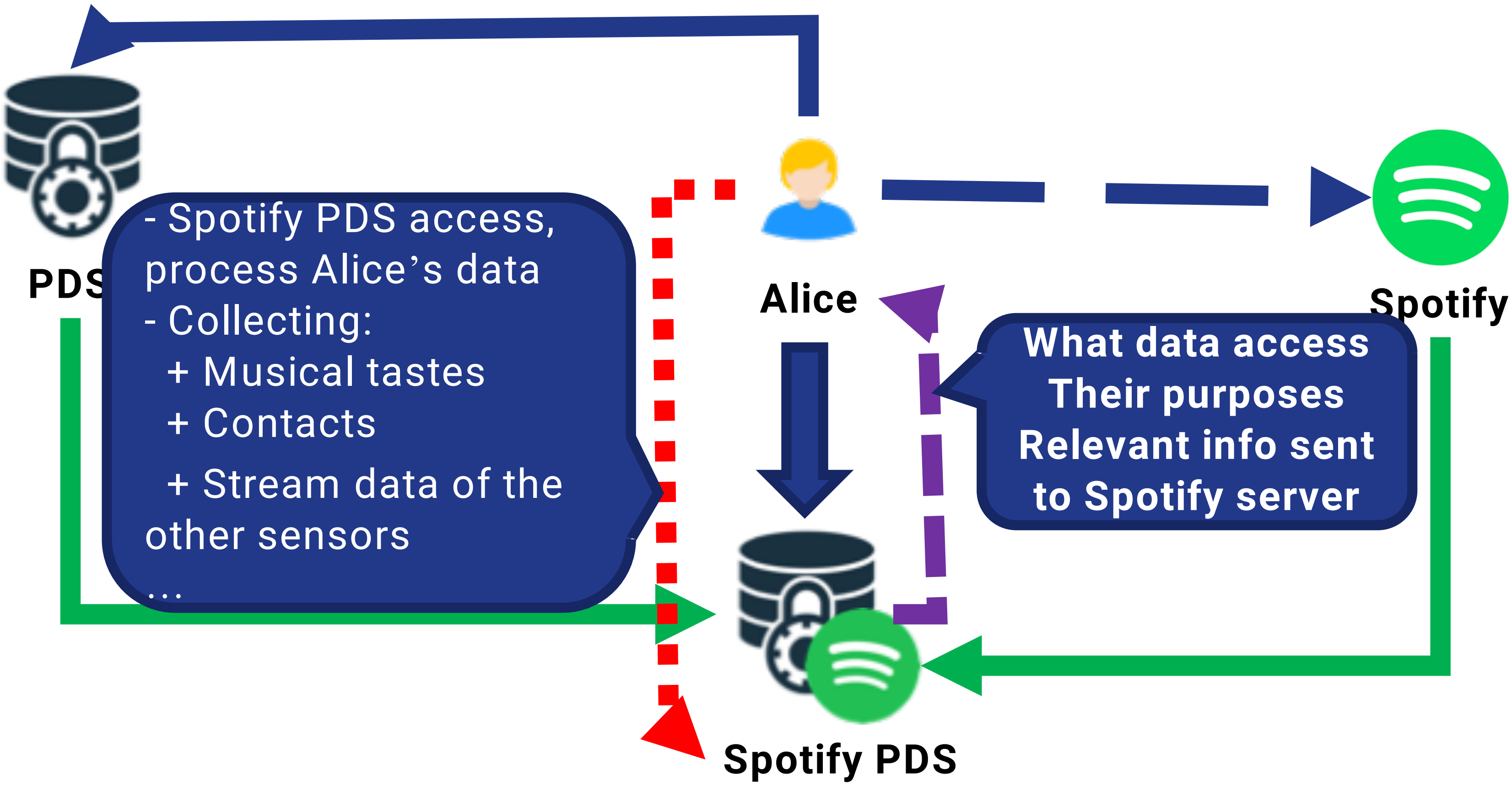
User's Data

- **Tradeoffs**
  - convenience
  - risk
- **openPDS**
  - Provide better data-powered services
- **The key innovation:**
  - compute on user data are performed in PDS
  - describe only the relevant summarized data

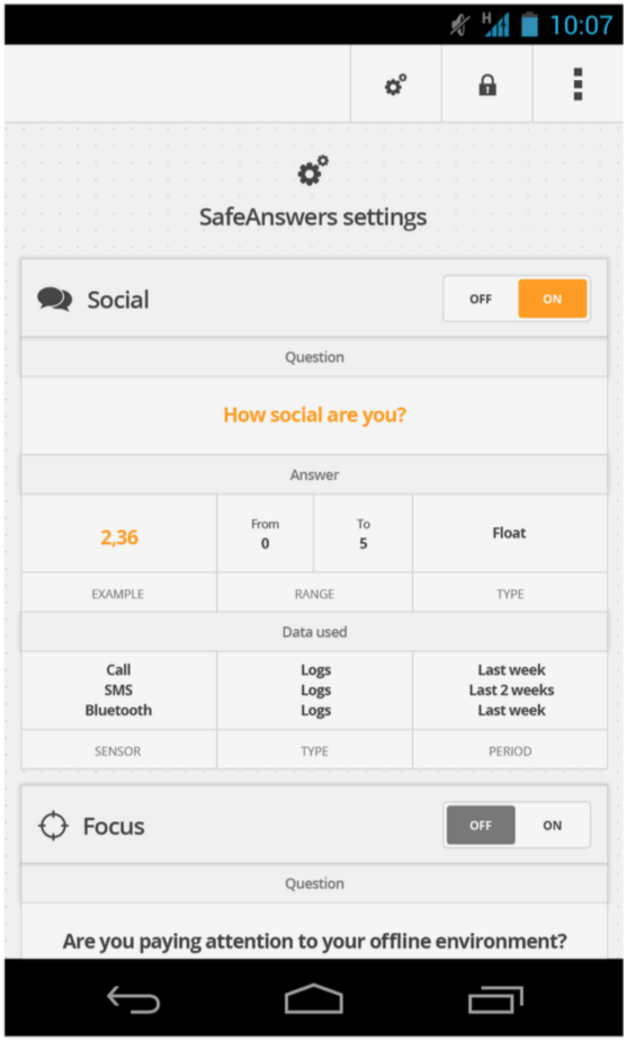
Figure 1: openPDS system's architecture.

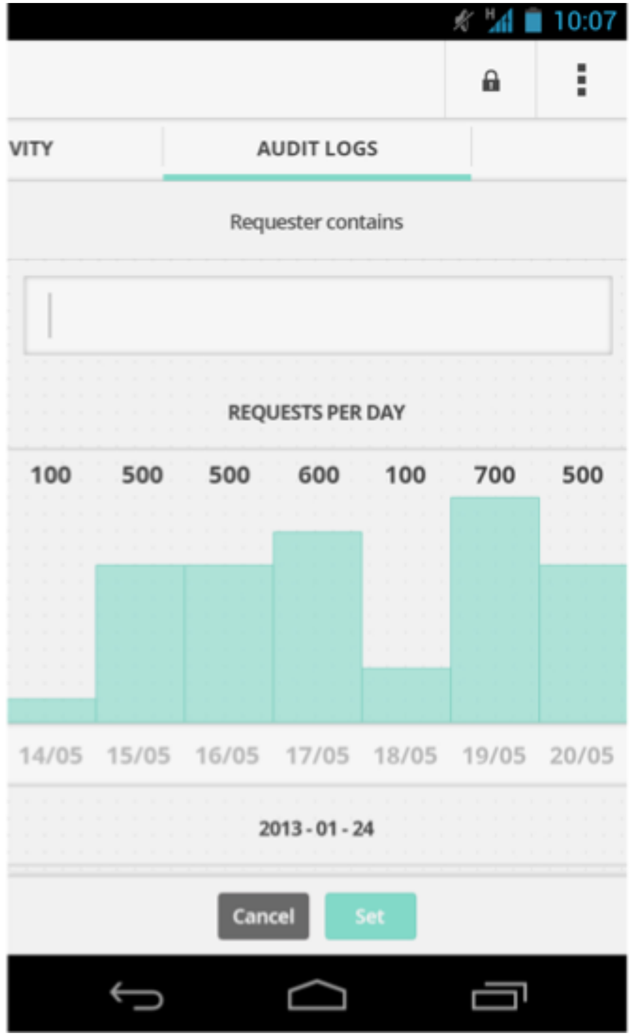


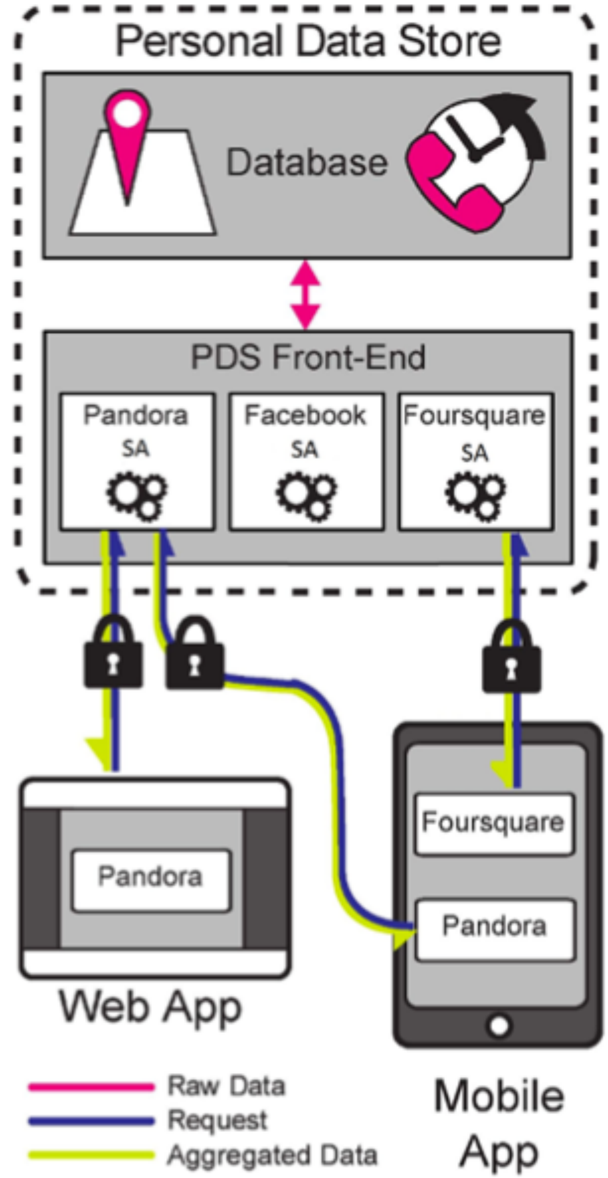
User's Data


PDS
- Spotify PDS access, process Alice's data
- Collecting:
+ Musical tastes
+ Contacts
+ Stream data of the other sensors
...
Alice
Spotify
What data access
Their purposes
Relevant info sent to Spotify server
Spotify PDS

Metadata (SafeAnswers)





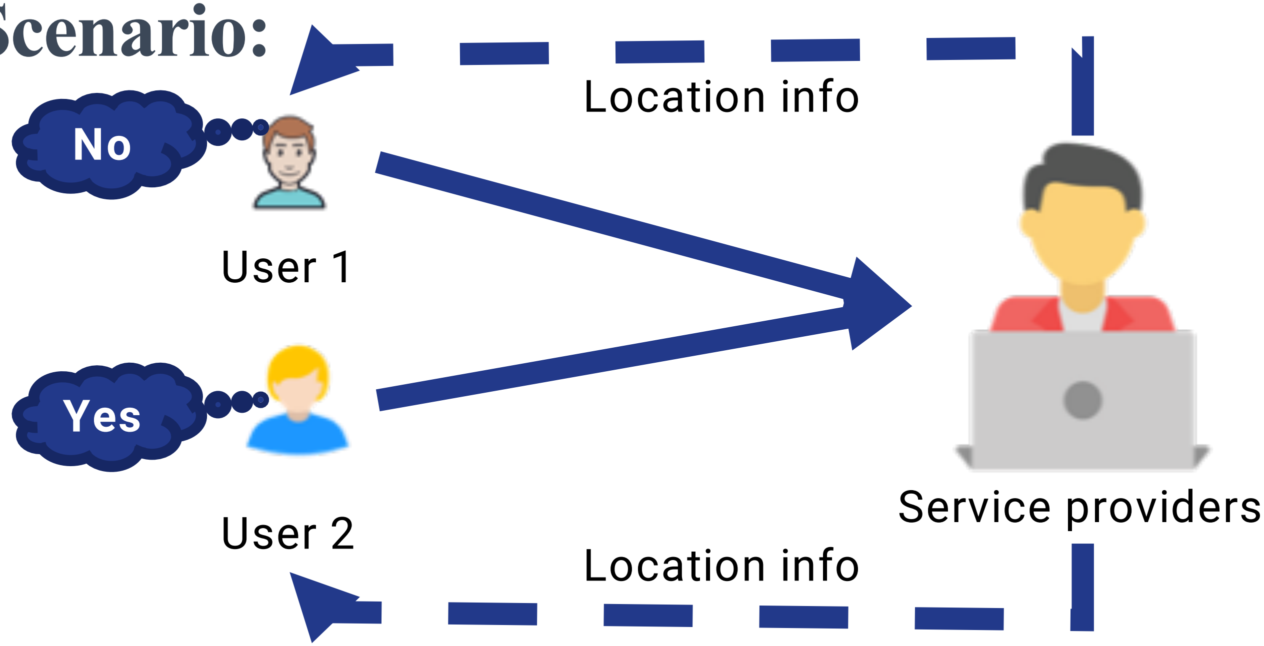


Approaches - Machine learning algorithms

- **Active learning**
  - [7] Singh, B. C., Carminati, B., & Ferrari, E. (2017, June). Learning privacy habits of pds owners. In 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS) (pp. 151-161)
  - [8] Singh, B. C., Carminati, B., & Ferrari, E. (2019). Privacy-aware personal data storage (p-pds): Learning how to protect user privacy from external applications. IEEE Transactions on Dependable and Secure Computing.
- **Distance metrics**
  - [9] Alom, M. Z., Carminati, B., & Ferrari, E. (2019, July). Adapting Users' Privacy Preferences in Smart Environments. In 2019 IEEE International Congress on Internet of Things (ICIOT) (pp. 165-172)
  - [10] Alom, M. Z., Carminati, B., & Ferrari, E. (2019, July). Helping Users Managing Context-Based Privacy Preferences. In 2019 IEEE International Conference on Services Computing (SCC) (pp. 100-107)

Singh et al.

- **Active learning**
  - **Motivation**: designing a Privacy-aware Personal Data Storage (**P-PDS**), that able to automatically take **privacy-aware decisions** on third parties access requests in *accordance with user preferences*.
  - **Scenario:**



Singh et al.

- **Approach**:
  - The authors do a step in this direction by proposing **different active learning algorithms:**
    - *Single-view: Expectation-Maximization (EM) Algorithm*
    - *Multi-view: (Co-EM) Algorithm*
    - *Ensemble learning Algorithm*

  => that allow a **fine-grained learning** of the **privacy aptitudes of PDS owners**.
  - Designing a **privacy-aware PDS** able to **automatically answer** to **service provider requests.**
  - Predicting whether a **new access request** has to be **granted** or **denied**

Singh et al.

- **Definition 1**. *Access request. An access request AR is a tuple (*DC, st, d0, p, o*), where* DC *is the data consumer, that is, the third party requesting data to the PDS,* st *is the type of service provided by* DC, d0 *are the requested data, whereas* p *is the access purpose. If the access is granted,* DC *will provide an additional benefit, called offer, modelled by* o.
- **Algorithms**
  - Expectation- Maximization (EM) [13]
  - co-EM algorithm [14]
  - ensemble approach [15], [16]

Singh et al.

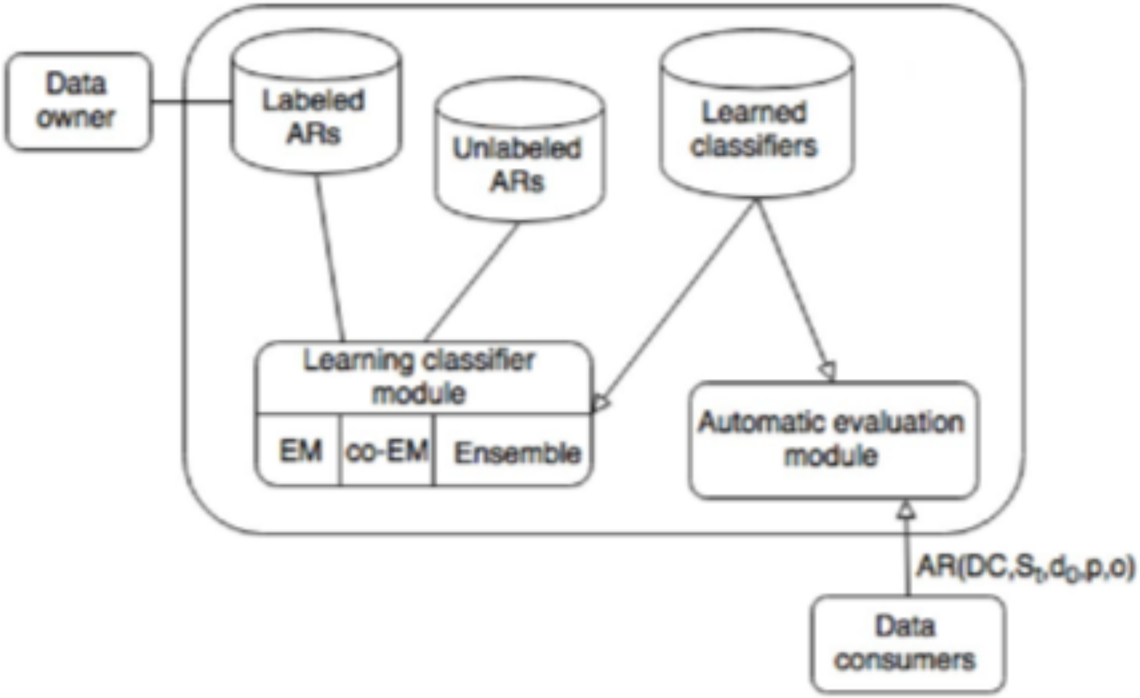


Fig. 1: Privacy-aware PDS

Alom et al.

- Services are context dependent, provide services based on sensing users' **contextual information**

any piece of data that used to define individual's current situation

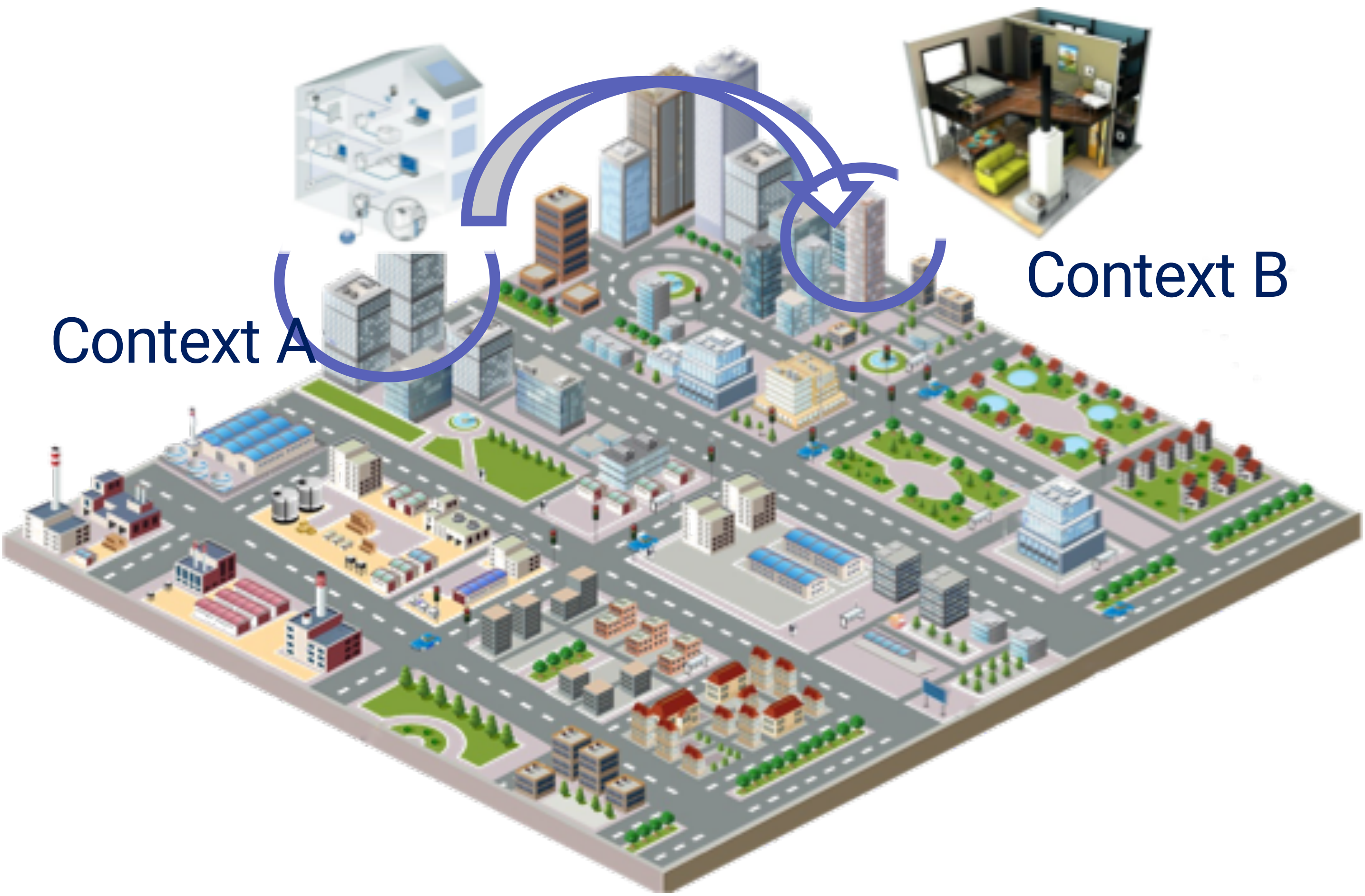
Why Context-based Privacy Preferences ?

- Contexts can greatly impact the users' privacy preferences

home at night

office hours at office

Alom et al.



Alom et al.

✓ Most Similar Context

✓ Adapting Privacy Preferences

Alom et al.

- Context CTX_U is a set of pairs $\{(tm, v_{tm}), (lc, v_{lc}), (ac,\ v_{ac}), (sl, v_{sl})\}$

- A context-based priva… where, CTX is a context, and PP…

tm is the time

p is the purpose

d is the data

ret indicates storage time

rec specifies third parties access status

Alom et al.

- Measure the distance to determine how far the new context is from existing contexts
- To do so, they measure the distance of each context-based privacy preference component:

Time distance: time can be expressed as a numerical value

$$D_{tm}\left(CTX_{U_n}.tm,\ CTX_{U_p}.tm\right) = \frac{|CTX_{U_n}.tm - CTX_{U_p}.tm|}{max(CTX_{U_n}.tm, CTX_{U_n}.tm)}$$

Location distance: location can be represented as hierarchy,

$$D_{lc}\left(CTX_{U_n}.lc,\ CTX_{U_p}.lc\right) = 1 - \frac{2 * depth(ccn)}{dis(CTX_{U_n}.lc) + dis\left(CTX_{U_p}.lc\right) + 2 * depth(ccn)}$$

Activity distance: activity can be represented as ontology

$$D_{ac}\left(CTX_{U_n}.ac,\ CTX_{U_p}.ac\right) = 1 - \frac{2 * depth(ccn)}{dis(CTX_{U_n}.ac) + dis\left(CTX_{U_p}.ac\right) + 2 * depth(ccn)}$$

Social distance: social presented as hierarchy

$$D_{sl}\left(CTX_{U_n}.sl,\ CTX_{U_p}.sl\right) = 1 - \frac{2 * depth(ccn)}{dis(CTX_{U_n}.sl) + dis\left(CTX_{U_p}.sl\right) + 2 * depth(ccn)}$$

Fig. 8 Location hierarchy

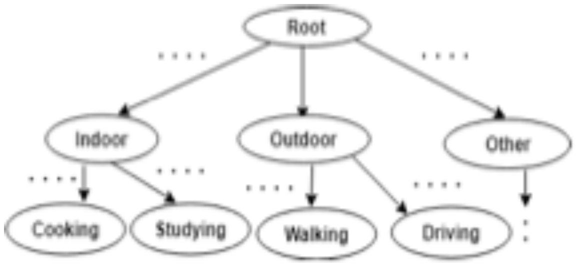


Fig. 9 Activity hierarchy

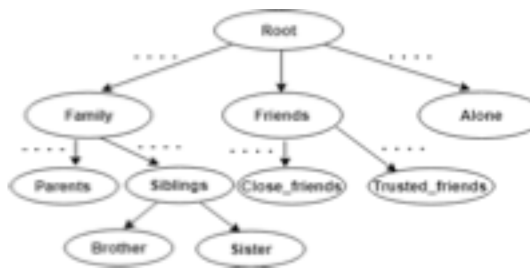


Fig. 10 Social hierarchy

Alom et al.

- CTX_{U_1} and CTX_{U_2} be two contexts for user U. Let w_1 ...... w_2 be the weights associated with each of the four context attributes. The similarity score is defined as follows:

$$Sim_w(CTX_{U1}, CTX_{U2}) = 1 - \frac{\begin{pmatrix} w_1 * D_{tm}(CTX_{U1}.tm, CTX_{U2}.tm) + w_2 * D_{lc}(CTX_{U1}.lc, CTX_{U2}.lc) \\ + w_3 * D_{ac}(CTX_{U1}.ac, CTX_{U2}.ac) + w_4 * D_{sl}(CTX_{U1}.sl, CTX_{U2}.sl) \end{pmatrix}}{4}$$

Alom et al.

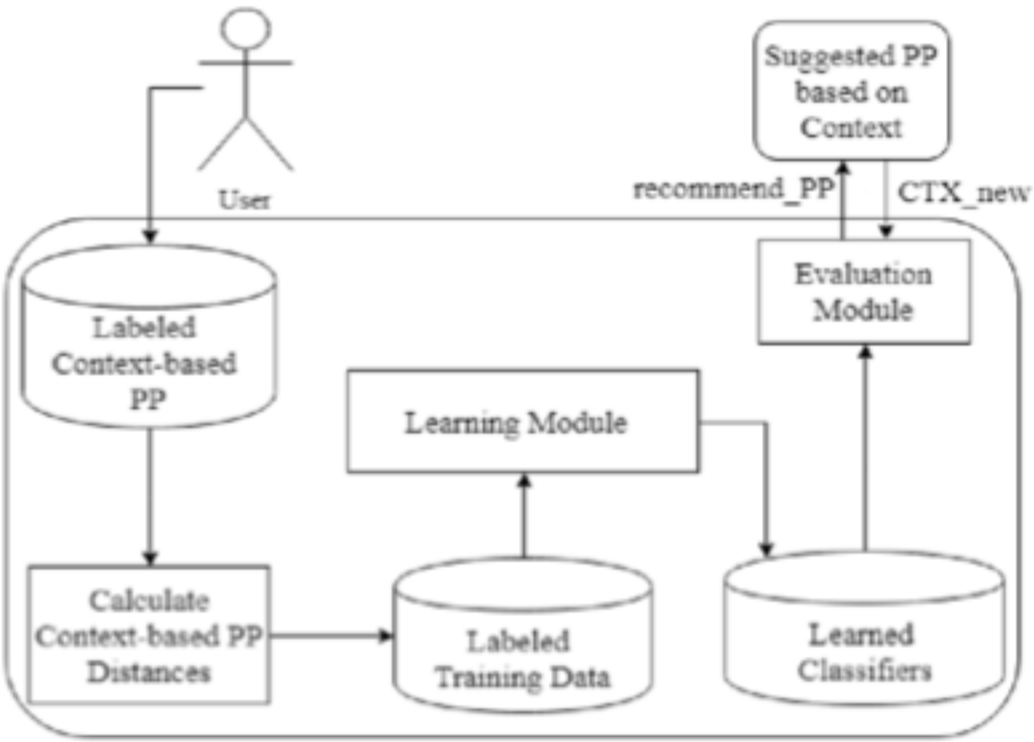


Figure 2: Learning architecture

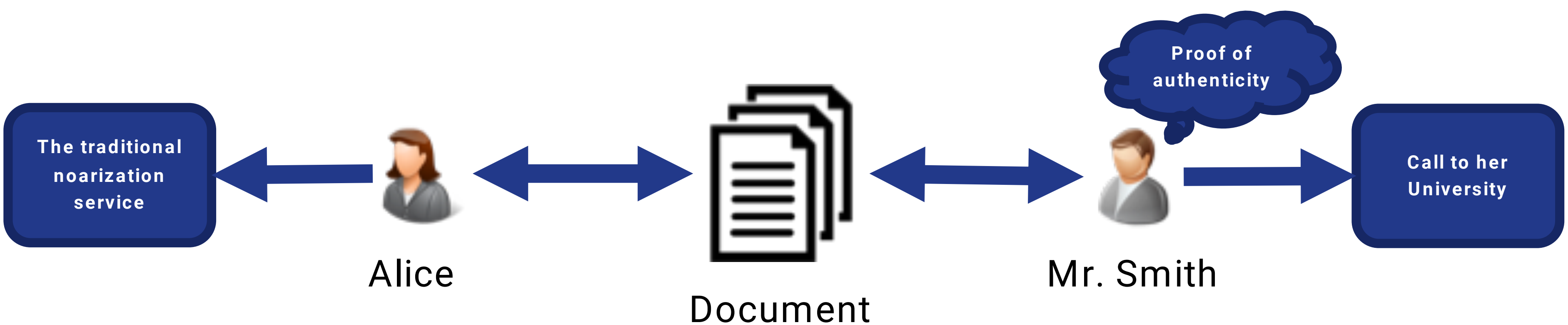
Approaches - Blockchain

[11] Chowdhury, M. J. M., Colman, A., Kabir, M. A., Han, J., & Sarda, P. (2018, August). Blockchain as a notarization service for data sharing with personal data store. In 2018 17th ieee international conference on trust, security and privacy in computing and communications (TrustCom) (pp. 1330-1335).

[12] Alessi, M., Camillo, A., Giangreco, E., Matera, M., Pino, S., & Storelli, D. Make users own their data: A decentralized personal data store prototype based on ethereum and ipfs. In 2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech) (2018, June), (pp. 1-7)

Approaches - Blockchain

- A blockchain is a
  - distributed
  - irreversible
  - tamper resistant
  - …
- Scenario



Approaches - Blockchain

Key(cudstodianID||stutID||docID)
Value(*H* {DSid || DCusid || DR })

- Membership Service
- Data
  - Company
- Data Custodian
  - University
- PDS
  - Data storage
- Context Provider

data-subject id (DSid)
data-custodian id (DCusid)
data resource (DR).

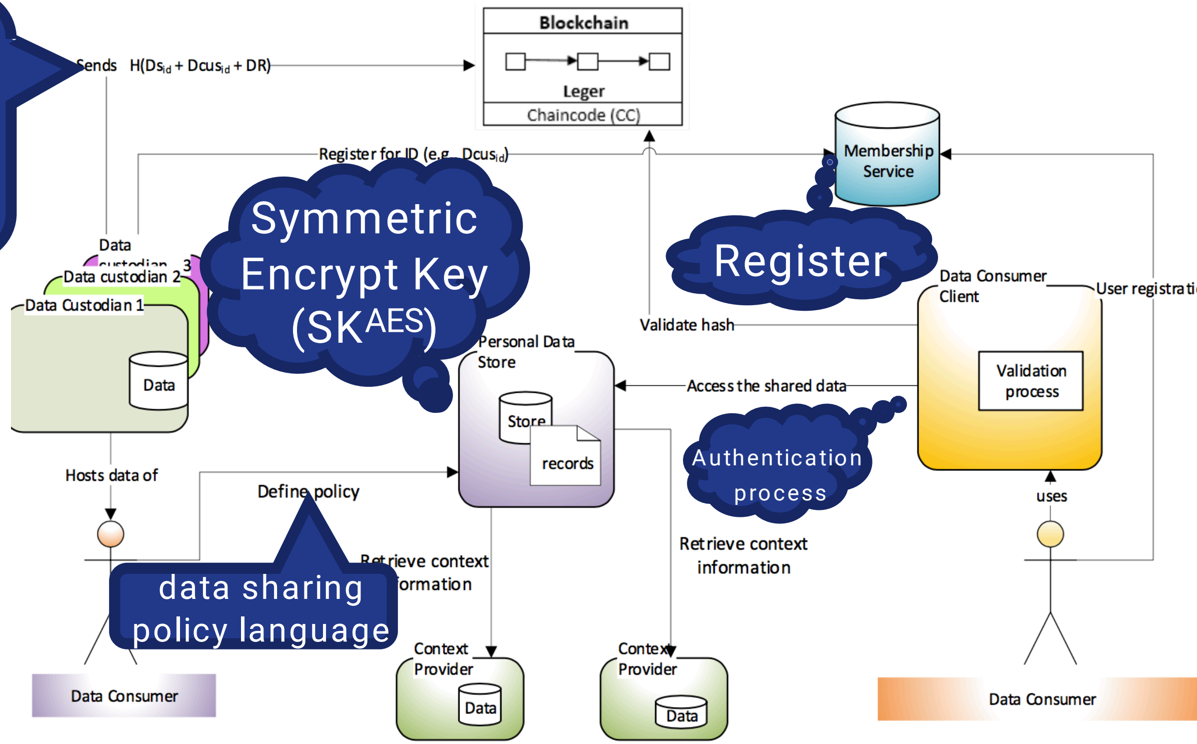


Fig. 1: The system architecture of blockchain based data notarization

Reference

[1]Kalapesi, C. “Unlocking the value of personal data: From collection to usage.” World Economic Forum technical report. 2013.

[2]“Better choices, better deals”, Published 13 April 2011, URL: https://www.gov.uk/government/news/better-choices-better-deals

[3]“Informing Consumers through Smart Disclosure”, Published March 30, 2012, URL: https://obamawhitehouse.archives.gov/blog/2012/03/30/informing-consumers-through-smart-disclosure

[4] Regulation, Protection. "Regulation (EU) 2016/679 of the European Parliament and of the Council." REGULATION (EU) 679 (2016): 2016.

[5] De Montjoye, Y. A., Wang, S. S., Pentland, A., Anh, D. T. T., & Datta, A. (2012). On the Trusted Use of Large-Scale Personal Data. IEEE Data Eng. Bull., 35(4), 5-8.

[6] De Montjoye, Y. A., Shmueli, E., Wang, S. S., & Pentland, A. S. (2014). openpds: Protecting the privacy of metadata through safeanswers. PloS one, 9(7), e98790.

[7] Singh, B. C., Carminati, B., & Ferrari, E. (2017, June). Learning privacy habits of pds owners. In 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS) (pp. 151-161)

[8] Singh, B. C., Carminati, B., & Ferrari, E. (2019). Privacy-aware personal data storage (p-pds): Learning how to protect user privacy from external applications. IEEE Transactions on Dependable and Secure Computing.

Reference

[9]Alom, M. Z., Carminati, B., & Ferrari, E. (2019, July). Adapting Users' Privacy Preferences in Smart Environments. In 2019 IEEE International Congress on Internet of Things (ICIOT) (pp. 165-172)

[10] Alom, M. Z., Carminati, B., & Ferrari, E. (2019, July). Helping Users Managing Context-Based Privacy Preferences. In 2019 IEEE International Conference on Services Computing (SCC) (pp. 100-107)

[11] Chowdhury, M. J. M., Colman, A., Kabir, M. A., Han, J., & Sarda, P. (2018, August). Blockchain as a notarization service for data sharing with personal data store. In 2018 17th ieee international conference on trust, security and privacy in computing and communications (TrustCom) (pp. 1330-1335).

[12] Alessi, M., Camillo, A., Giangreco, E., Matera, M., Pino, S., & Storelli, D. Make users own their data: A decentralized personal data store prototype based on ethereum and ipfs. In 2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech) (2018, June), (pp. 1-7)

[13] S. Borman. “The Expectation Maximization algorithm a short tutorial”, Technical report, 2006

[14] K. Nigam and R. Ghani. “Analyzing the Effectiveness and Applicability of Co-training”, In Proc. of CIKM ’00, pp 86-93, New York, NY, USA, 2000.

[15] M. Sewell. “Emsemble learning”, UCL Research Note,2007

[16] Zhou ZH. “When Semi-supervised learning meets ensemble learning”, In Proc. of MCS 2009, pp 529-538, Raykjavik, Iceland, Jun. 10-12, 2009

Q & A

Thank you for Listening!