

---

# Anonymity and Privacy

---

# Privacy

- ▷ *someone's right to keep their personal matters and relationships secret*
- ▷ *the state of being alone, or the right to keep one's personal matters and relationships secret*
- ▷ *the right that someone has to keep their personal life or personal information secret or known only to a small group of people*

<https://dictionary.cambridge.org/pt/dicionario/ingles/privacy>

# Privacy Types

## ▷ Physical

- ◆ Regarding the properties off the own body

## ▷ Surveillance

- ◆ Regarding the state of being observed

## ▷ Information

- ◆ Regarding the information about the self (how it is handled)

# IEEE Digital Privacy Model

- ▶ Expectations of Privacy (EOP)
  - ◆ When/how users expect privacy
- ▶ EoP is has six characteristics:
  - ◆ Identities
  - ◆ Behaviors
  - ◆ Inferences
  - ◆ Transactions
  - ◆ Confidentiality & Integrity
  - ◆ Access & Observability

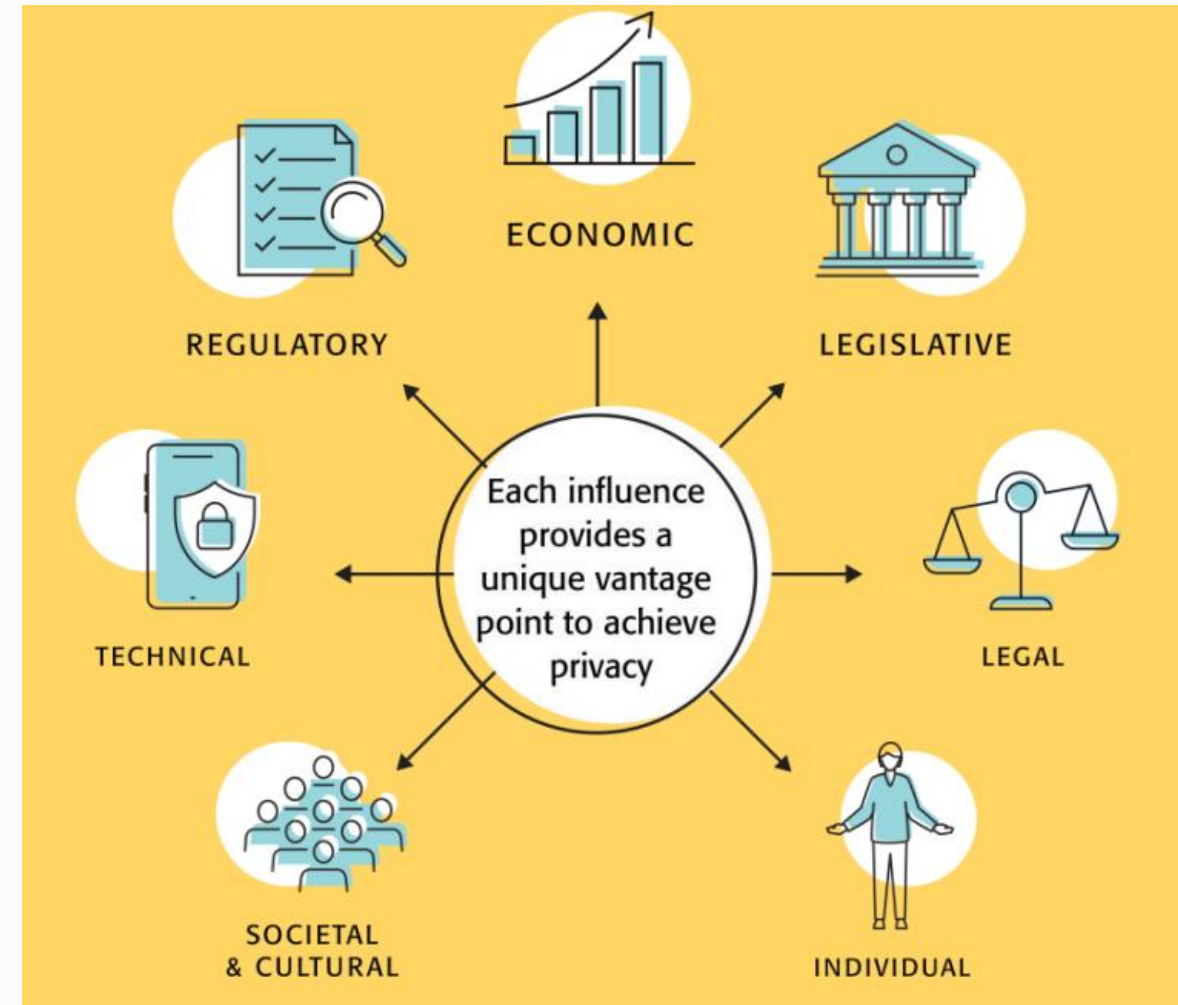


# IEEE Digital Privacy Model

- ▶ **Confidentiality & Integrity** plus **Access & Observability** of data/meta data about individuals'
- ▶ **Identities, Behaviors, Inferences, and Transactions**
- ▶ forms the overall **Expectations of Privacy**

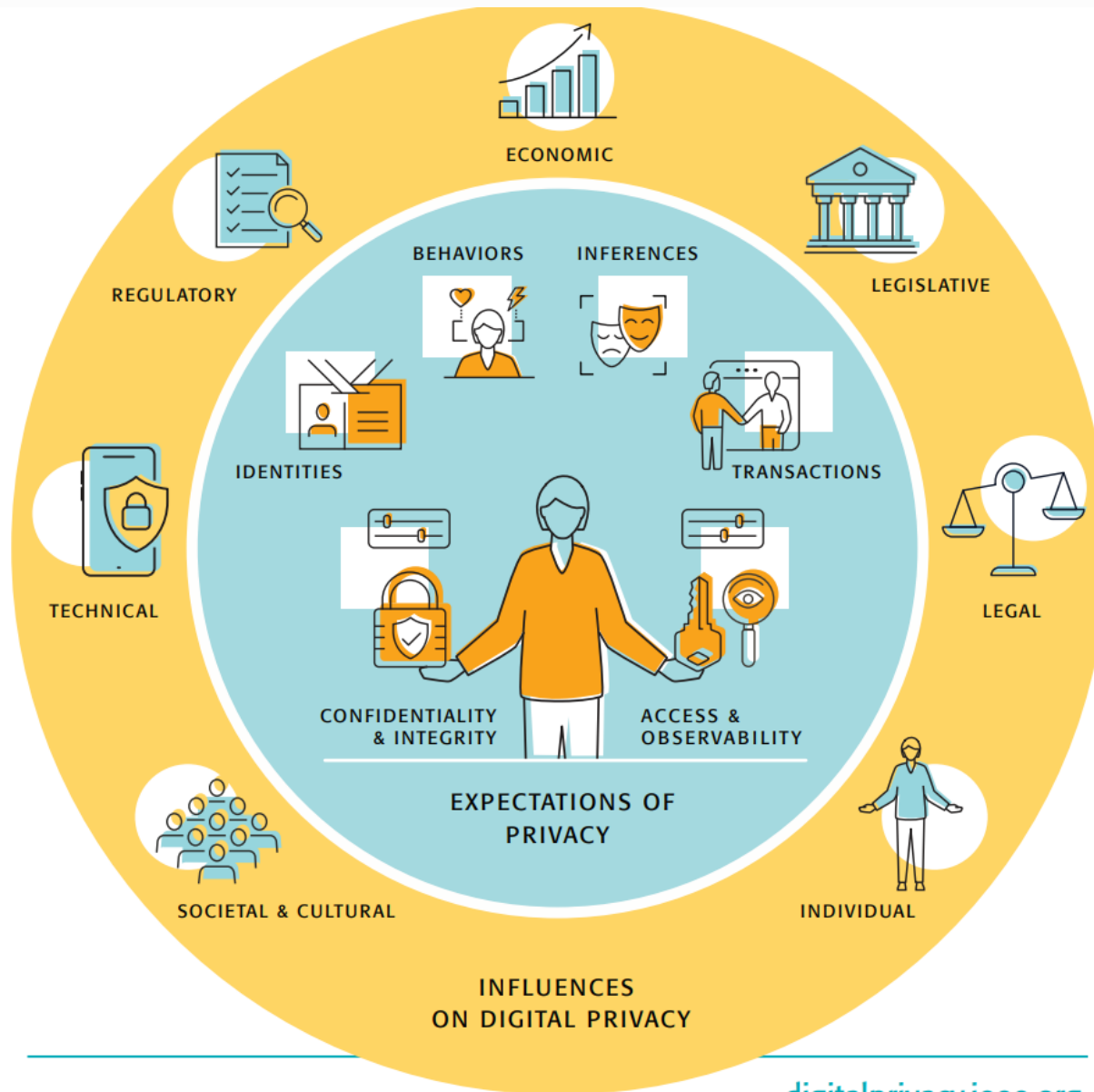
# IEEE Digital Privacy Model

- ▷ Influences on Privacy (IOP)
- ▷ Influences that help shape the overall digital privacy infrastructure in any environment or region.
- ▷ Influences on Privacy are:
  - ◆ Technical
  - ◆ Regulatory
  - ◆ Economic
  - ◆ Legislative
  - ◆ Legal
  - ◆ Individual
  - ◆ Societal & Cultural



# IEEE Digital Privacy Model

- ▷ Influences on Privacy (IOP)
- ▷ **Technical, Regulatory, Economic, Legislative, Legal, Individual, and Societal & Cultural**
- ▷ influences determine the overall **Expectations of Privacy** achievable in any Digital Ecosystem



[digitalprivacy.ieee.org](https://digitalprivacy.ieee.org)

© IEEE 2023. All rights reserved.



# Privacy with computing technology

- ▶ Technology usage is problematic for privacy
  - ◆ Technology usage (always) creates information
    - Data: information explicitly created by subjects
    - Metadata: control, addressing and descriptive information resulting from interactions between systems
  - ◆ Computers process information at high speed
    - Damage is done very rapidly and silently, across the globe
  - ◆ Computers may be used to store massive amounts of data
    - Which can be leaked
  
- ▶ Is it possible to use technology without providing personal data?

# Privacy and companies

- ▷ Companies need data to fulfill their business models
  - ◆ Financial, marketing, transactional
  - ◆ Metadata
- ▷ Frequently, privacy is compromised for providing the product
  - ◆ If you are not paying, you are the product (your data also)
  - ◆ If you are paying, you may also be the product
  - ◆ Insufficient controls may enable data leakage
- ▷ Awareness of privacy violation may cause damage
  - ◆ Impact brand and user perception
  - ◆ Legal impact

# Privacy and IAA

- ▷ Privacy is highly connected with the methods and technologies used in IAA
- ▷ IAA deals with users, their attributes and their relations
  - ◆ Also considers context and interactions

# Identification

- ▷ Privacy is relevant as the identification data may reveal additional information
  - ◆ Impact: OSINT tracking, Doxing, ID Theft
- ▷ Identifier reuse
  - ◆ Tracking users across platforms
- ▷ Badly constructed
  - ◆ Reveal real name, location, age, profession
- ▷ Non selective
  - ◆ Presenting identification discloses too much information

# Identification: OSINT

- ▶ Identifier reuse
- ▶ Track users across services
- ▶ Some services allow further deanonymization
  - ◆ Github: profession
  - ◆ Social networks: age
  - ◆ Chess: ?

```
~/sherlock
$ python3 sherlock hackerman1337
[*] Checking username hackerman1337 on:

[+] 9GAG: https://www.9gag.com/u/hackerman1337
[+] AskFM: https://ask.fm/hackerman1337
[+] BitBucket: https://bitbucket.org/hackerman1337/
[+] Chess: https://www.chess.com/member/hackerman1337
[+] Codecademy: https://www.codecademy.com/profiles/hackerman1337
[+] Disqus: https://disqus.com/hackerman1337
[+] Docker Hub: https://hub.docker.com/u/hackerman1337/
[+] FortniteTracker: https://fortnitetracker.com/profile/all/hackerman1337
[+] Freesound: https://freesound.org/people/hackerman1337/
[+] GitHub: https://www.github.com/hackerman1337
[+] Instagram: https://www.instagram.com/hackerman1337
[+] Kik: https://kik.me/hackerman1337
[+] LeetCode: https://leetcode.com/hackerman1337
[+] Lichess: https://lichess.org/@/hackerman1337
[+] Minecraft: https://api.mojang.com/users/profiles/minecraft/hackerman1337
[+] OK: https://ok.ru/hackerman1337
[+] OpenStreetMap: https://www.openstreetmap.org/user/hackerman1337
[+] Pastebin: https://pastebin.com/u/hackerman1337
[+] Periscope: https://www.periscope.tv/hackerman1337/
[+] Pokemon Showdown: https://pokemonshowdown.com/users/hackerman1337
[+] Quizlet: https://quizlet.com/hackerman1337
[+] Redbubble: https://www.redbubble.com/people/hackerman1337
[+] Reddit: https://www.reddit.com/user/hackerman1337

[*] Search completed with 26 results



~/sherlock
$
```

# Identification: Information reveal

- ▷ admin
- ▷ administrator
- ▷ pedro.afonso
- ▷ mariasilva@xyz.gov.pt
- ▷ rita1987@gmail.com
- ▷ joe420
- ▷ bookworm98
- ▷ coffeefueledCoder
- ▷ sarcastic\_seashell
- ▷ greenpeacewarrior
- ▷ ihatemondays
- ▷ genzgamer
- ▷ londonlarry
- ▷ bigskybiker
- ▷ PT215295386
- ▷ PT501343923

# Identification: Behavior reveal

What's in a name? Ages and names predict the valence of social interactions in a massive online game

[Athanasios V. Kokkinakis](#)<sup>a</sup>, [Jeff Lin](#)<sup>b</sup>, [Davin Pavlas](#)<sup>b</sup>, [Alex R. Wade](#)<sup>a</sup>  

- ▷ Data from League of Legends reveals links between real-world and online personality.
- ▷ Players with antisocial usernames behaved in an antisocial manner within the game.
- ▷ Ages estimated from usernames correlate strongly with ages entered at registration.
- ▷ Online interactions valence correlates positively age. Older players behave better.
- ▷ Gaming usernames provide a useful source of psychological information.

# Identification: Information reveal / non selective





# Identification: Best practices

- ▷ Identifiers should:
  - ◆ Be service specific
  - ◆ Be unique within the limits of usability
    - Alternative: numbers or a digest of a public key
  - ◆ Be kept private within the limits of service operation
  - ◆ Not disclose further information
- ▷ Good practices:
  - ◆ Discord: allows a different profile per user
  - ◆ Reddit: allows random usernames
- ▷ May require application to manage usernames
  - ◆ Key vault

# Authentication

- ▷ Privacy is relevant as the authentication data may reveal additional information
  - ◆ Impact: OSINT tracking, Doxing, ID Theft, medical condition
- ▷ Authentication data has personal information
- ▷ Authentication methods behavioral information

# Authentication: personal information

- ▶ Privacy is relevant as the authentication data may reveal additional information
  - ◆ Impact: OSINT tracking, Doxing, ID Theft, medical condition
- ▶ Authentication data has personal information
- ▶ Authentication methods behavioral information

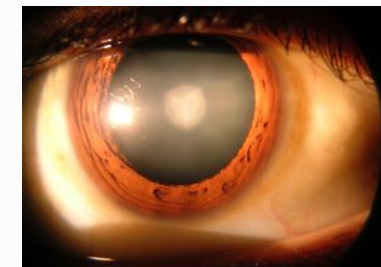
# Authentication: personal information

## Security Questions

Select a security question or create one of your own. This question will help us verify your identity should you forget your password.

Security Question	What is the first name of your best friend in high s <input type="button" value="v"/>
	Please select
Answer	What is the first name of your best friend in high school?
	What was the name of your first pet?
Security Question	What was the first thing you learned to cook?
	What was the first film you saw in a theater?
	Where did you go the first time you flew on a plane?
	What is the last name of your favorite elementary school teacher?
Answer	*****
	<input type="button" value="Save answers"/> <input type="button" value="Cancel"/>

# Authentication: personal information



- ▷ gender, ethnicity, age, medical conditions, tracking/surveillance
- ▷ It is static: cannot be changed

# Authentication: behavioral information

## Login History

User Name	Email Address	Last Login Date/TI...	IP Address	Browser	Successful?
admin	informzqa@gmail.c...	3/6/2019 11:53 AM	194.44.234.160	Mozilla/5.0 (Window...	✓
admin	informzqa@gmail.c...	3/6/2019 11:54 AM	194.44.234.160	Mozilla/5.0 (Window...	✓
admin	informzqa@gmail.c...	3/6/2019 11:54 AM	194.44.234.160	Mozilla/5.0 (Window...	✓
admin	informzqa@gmail.c...	3/6/2019 11:54 AM	194.44.234.160	Mozilla/5.0 (Window...	
admin	informzqa@gmail.c...	3/6/2019 11:54 AM	194.44.234.160	Mozilla/5.0 (Window...	
admin	informzqa@gmail.c...	3/6/2019 11:55 AM	194.44.234.160	Mozilla/5.0 (Window...	
admin	informzqa@gmail.c...	3/6/2019 11:55 AM	194.44.234.160	Mozilla/5.0 (Window...	
admin	informzqa@gmail.c...	3/6/2019 11:56 AM	194.44.234.160	Mozilla/5.0 (Window...	
admin	informzqa@gmail.c...	3/6/2019 11:56 AM	194.44.234.160	Mozilla/5.0 (Window...	

- ▶ Time
- ▶ Location
- ▶ Browser
- ▶ IP



### Activity on this account

This feature provides information about the last activity on this mail account and any concurrent activity. [Learn more](#)

This account does not seem to be open in any other location. However, there may be sessions that have not been signed out.

[Sign out all other sessions](#)

### Recent activity:

Access Type [ ? ] (Browser, mobile, POP3, etc.)	Location (IP address) [ ? ]	Date/Time (Displayed in your time zone)
Browser	* United States (CA) (108.73.11.1)	9:14 pm (0 minutes ago)
IMAP	United States (CA) (108.73.11.1)	9:06 pm (8 minutes ago)
Mobile	United States (CA) (174.254.36.106)	2:41 pm (6 hours ago)
Mobile	United States (CA) (108.73.11.1)	Sep 4 (1 day ago)
Mobile	United States (CA) (108.73.11.1)	Sep 3 (2 days ago)
Browser	United States (CA) (108.73.11.1)	Sep 2 (3 days ago)
Mobile	United States (CA) (108.73.11.1)	Sep 2 (3 days ago)
Browser	United States (CA) (108.73.11.1)	Sep 2 (3 days ago)
Browser	United States (CA) (108.73.11.1)	Sep 2 (3 days ago)
Mobile	United States (CA) (108.73.11.1)	Sep 2 (3 days ago)

**Alert preference:** Show an alert for unusual activity. [change](#)

\* indicates activity from the current session.

# Authentication: Best practices

- ▶ Authentication data may need to be handled under GDPR
  - ◆ Considered to be personal information
  - ◆ Have a justification for every data item recorded
  - ◆ Include authentication data in GDPR data requests
    - Request to be forgotten
    - Personal data request
  
- ▶ Clear data processing practices
  - ◆ DPIA - Data Protection Impact Assessment
  - ◆ Strict management of authentication third parties
  - ◆ User consent

# Anonymity

- ▶ The state of being **not observable** within a set of subjects, the anonymity set
  - ◆ e.g. a particular person among all possible persons
  - ◆ e.g. a particular voter among all possible voters
  - ◆ e.g. a particular address among all possible addresses
- ▶ The anonymity set is the set of all possible subjects
  - ◆ From the attacker's point of view
- ▶ Is a context-dependent concept
  - ◆ The context defines the anonymity set regarding a particular action



# Microdata privacy issues

## ▷ Microdata

- ◆ Information at the level of individual respondents

## ▷ Privacy issues

- ◆ Microdata is often used for several studies
- ◆ How can we share microdata among companies without exposing its source?
  - The identity of the persons that provided it

# Microdata privacy enhancing: Removal of potentially unique IDs

- ▷ Basic strategy
  - ◆ By removing potentially unique IDs we cannot link microdata items from several databases
- ▷ Candidate IDs
  - ◆ Name
  - ◆ National IDs (passport, identity card, etc.)
  - ◆ Social Security ID, Tax ID, etc.
  - ◆ Phone numbers
  - ◆ Car plate numbers
- ▷ Not enough!
  - ◆ A study in the States proved that 87% of its the population could be identified using a link attack using 3 non-unique attributes
    - 5-digit ZIP code, gender and birthday

L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

# Microdata privacy enhancing: Noise

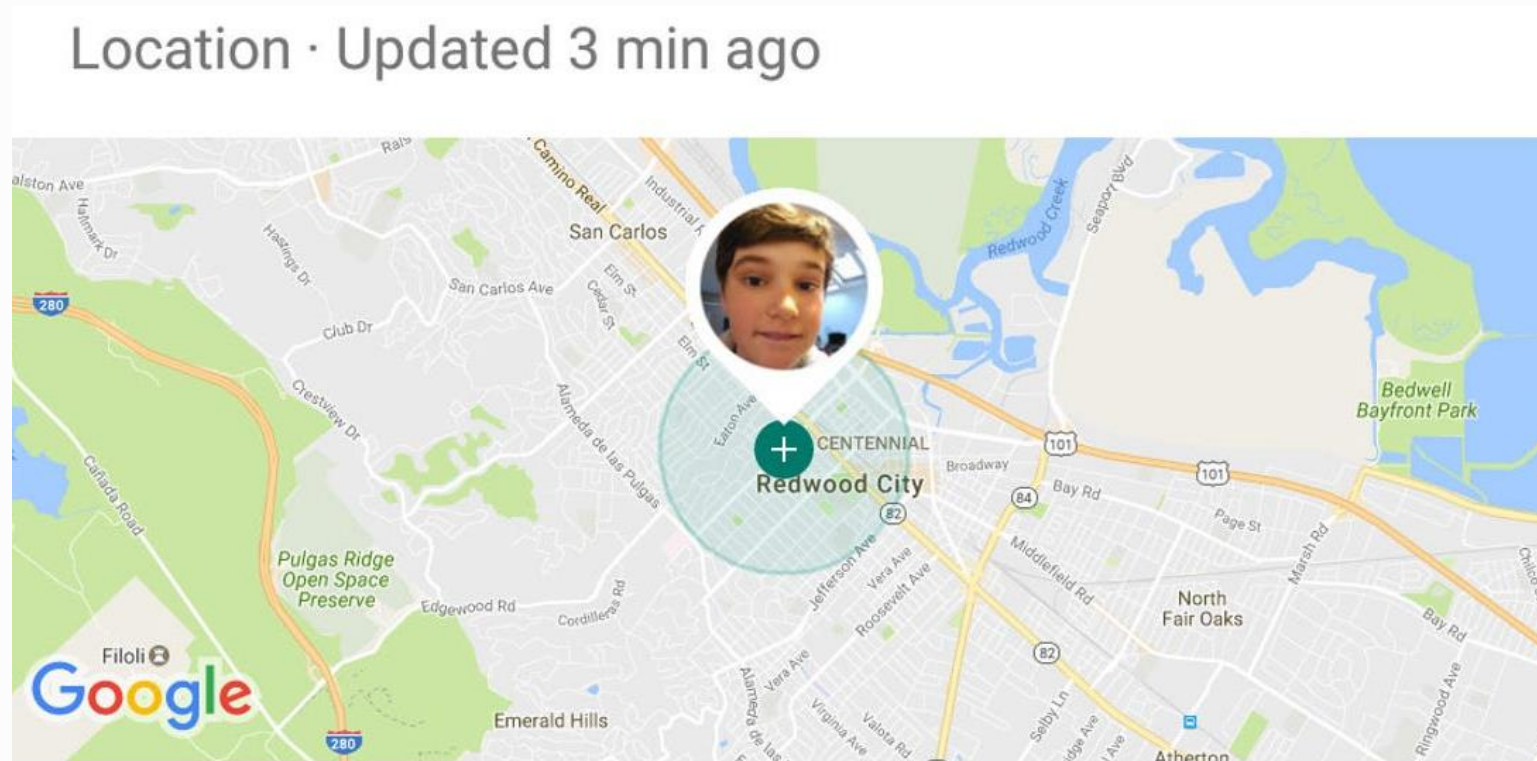
## ▷ Basic strategy

- ◆ Add noise to stored data or to the result of queries

## ▷ Issues

- ◆ Privacy is achieved at the cost of integrity

# Microdata privacy enhancing: Noise



<https://www.cnet.com/tech/mobile/google-accounts-for-kids-5-things-to-know-about-family-link/>

# Microdata privacy enhancing: Truncate

## ▷ Basic strategy

- ◆ Do not provide full data, limiting precision

## ▷ Issues

- ◆ Privacy is achieved at the cost of integrity
- ◆ Difficult to balance usability with privacy
  - Privacy relates to the user providing information, which may not be the user accessing information

# Microdata privacy enhancing: Truncate

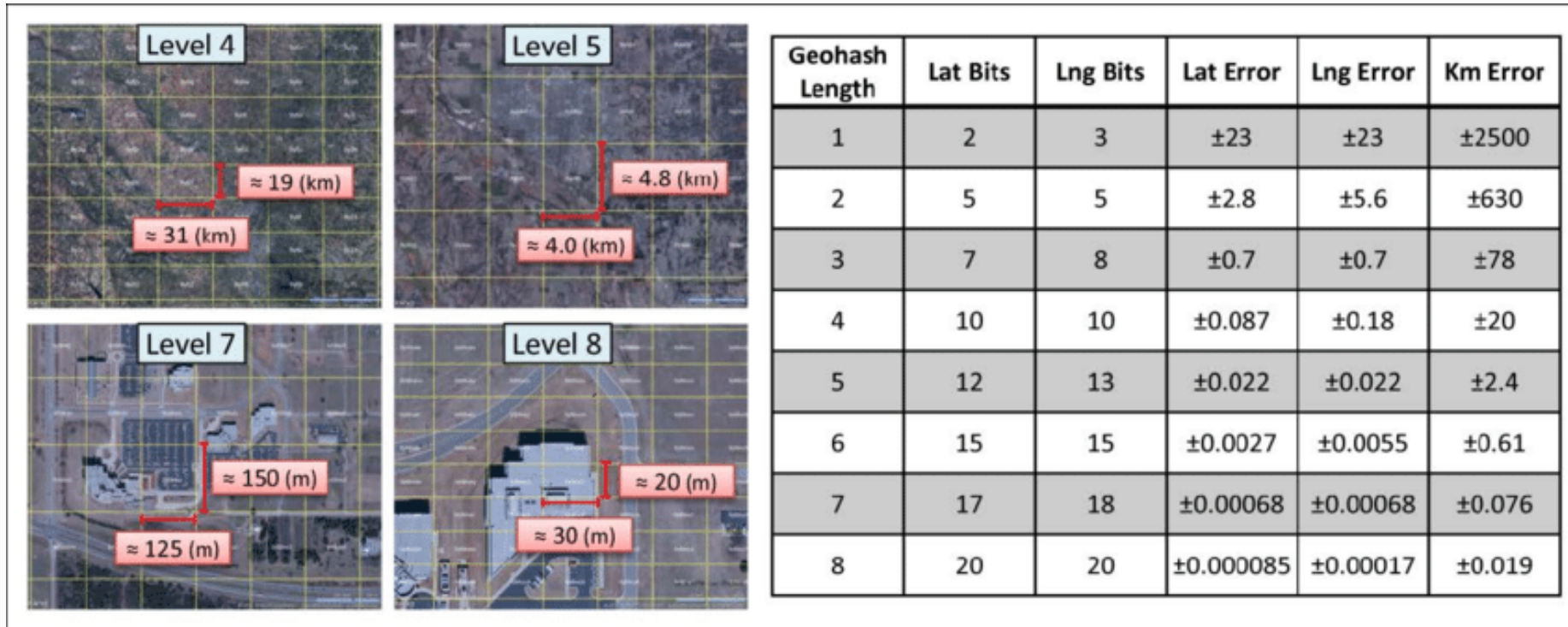


Image from: Yuan, May & Nara, Atsushi. (2015). Space-Time Analytics of Tracks for the Understanding of Patterns of Life. 10.13140/2.1.3690.7525.

# Microdata privacy enhancing: K-anonymity

L. Sweeney, "K-anonymity: A Model for Protecting Privacy", Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 2002

## ▷ Definition

- ◆ No query can deliver an **anonymity set** with less than **k** entries

## ▷ Privacy-critical attributes

- ◆ (Unique) identifiers
- ◆ Quasi-identifiers
  - When combined can produce unique tuples
- ◆ Sensitive attributes
  - Potentially unique per subject
  - Disease, salary, crime committed

# K-anonymity:

## Implementation approaches

- ▷ **Suppression** of quasi-identifiers
  - ◆ Simple to perform
  - ◆ Information loss
- ▷ **Generalization** of quasi-identifiers
  - ◆ Transformation of quasi-identifiers in other ones less specific
    - e.g. 7-digit ZIP → 4-digit ZIP
    - e.g. ages w/ 1 year granularity → 5 or 10 year granularity
  - ◆ There is not a complete loss of information
    - But the generalization should not potentiate wrong data interpretations
  - ◆ We must ensure that there are at least **k** entries with equal generalized quasi-identifiers



# Example

Name	Age	Sex	Zip Code	Illness
Sam	29	M	43102	Diabetes
Gloria	38	F	43102	Breast cancer
Adam	51	M	43102	Colon cancer
Eric	29	M	43102	Diabetes
Tanisha	34	F	43102	HIV
Don	51	M	43102	Heart disease

# Example:

## Identifiers

Name	Age	Sex	Zip Code	Illness
Sam	29	M	43102	Diabetes
gloria	38	F	43102	Breast cancer
Adam	51	M	43102	Colon cancer
Eric	29	M	43102	Diabetes
Tanisha	34	F	43102	HIV
Don	51	M	43102	Heart disease

# Example:

## Quasi identifiers

Name	Age	Sex	Zip Code	Illness
Sam	29	M	43102	Diabetes
gloria	38	F	43102	Breast cancer
Adam	51	M	43102	Colon cancer
Eric	29	M	43102	Diabetes
Tanisha	34	F	43102	HIV
Don	51	M	43102	Heart disease

# Example:

## Sensitive attributes

Name	Age	Sex	Zip Code	Illness
Sam	29	M	43102	Diabetes
gloria	38	F	43102	Breat cancer
Adam	51	M	43102	Colon cancer
Eric	29	M	43102	Diabetes
Tanisha	34	F	43102	HIV
Don	51	M	43102	Heart disease

# K-anonymity 1st step:

## Remove unique identifiers

Age	Sex	Zip Code	Illness
29	M	43102	Diabetes
38	F	43102	Breast cancer
51	M	43102	Colon cancer
29	M	43102	Diabetes
34	F	43102	HIV
51	M	43102	Heart disease

# K-anonymity 2nd step: Generalization

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breast cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease

# K-anonymity :

## 2-anonymity possible results

Two pairs

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breast cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease

# K-anonymity :

## 3-anonymity possible results

None

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breast cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease



# Issue:

## Sensitive attribute disclosure

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breast cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease

# L-Diversity

Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data (TKDD), 1.1, 2007.

- ▷ K-anonymity is not enough!
- ▷ Homogeneity attack
  - ◆ The attacker knows the generalized Quase-Identifiers of a target
  - ◆ A query reveals the exact same sensitive attributes
  - ◆ The attacker gets the sensitive attribute of the target
  - ◆ Issue: lack of diversity in the results
- ▷ Background knowledge attack
  - ◆ The attacker can filter out query results using known information

# Solution:

## l-diverse k-anonymity

- ▶ Results from a k-anonymity result of a query must contain l different values for each sensitive attribute

Illness
Diabetes
Breast cancer
Colon cancer
Diabetes
HIV
Heart disease

# I-diversity :

## 2-anonymity 1-diversity results

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breat cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease

# I-diversity :

## 2-anonymity 2-diversity results

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breat cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease

# k-anonymity and l-diversity have flaws

- ▷ **k-anonymity:** each equivalence class has at least  $k$  records to protect against identity disclosure.
  - ◆ k-anonymity is vulnerable to homogeneity attacks and background knowledge attacks.

# Attacks on k-anonymity

## ▷ homogeneity attack

- ◆ Bob is a 27-year old man living in zip code 47678 and Bob's record is in the table.
- ◆ So Bob corresponds to one of the first three records and must have heart disease.

## ▷ background knowledge attack

- ◆ Carl is a 32-year old man living in zip code 47622. Therefore he is in the last equivalence class in Table 2.
- ◆ If you know that Carl has a low risk for heart disease then you can conclude that Carl probably has cancer.

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

**Table 1. Original Patients Table**

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

**Table 2. A 3-Anonymous Version of Table 1**

# k-anonymity and l-diversity have flaws

- ▷ • l-diversity: distribution of a sensitive attribute in each equivalence class has at least l “well represented” values to protect against attribute disclosure.
- ▷ • l-diversity is vulnerable to skewness attacks and similarity attacks.
  - ◆ Skewness: keeping diverse groups may change statistical properties
  - ◆ Similarity: similar concepts are not handled



# Attacks on I-diversity

## ▷ Similarity Attack:

- ◆ Table 4 anonymizes table 3. Its sensitive attributes are Salary and Disease.
- ◆ If you know Bob has a low salary (3k-5k) then you know that he has a **stomach related** disease.
- ◆ This is because I-diversity takes into account the diversity of sensitive values in the group,
- ◆ but does not take into account the semantical closeness of the values.

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

**Table 3. Original Salary/Disease Table**

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

**Table 4. A 3-diverse version of Table 3**

# Attacks on I-diversity

- ▶ 10,000 records about a virus that affects 1% of the population.
- ▶ **Skewness attack:** with 2-diversity we have an equal number of positive and negative records.
- ▶ This gives everyone in this equivalence class a 50% chance of having the virus, which is much higher than the real distribution.

	Zip Code	Age	Salary	Disease
1	476**	2*	3k	negative
2	476**	2*	4k	negative
3	476**	2*	5k	negative
4	476**	2*	6k	negative
5	4790*	>=40	7k	negative
6	4790*	>=40	8k	positive
7	4790*	>=40	9k	negative
8	4790*	>=40	10k	positive
9	476**	3*	11k	positive
10	476**	3*	12k	positive
11	476**	3*	13k	positive
12	476**	3*	14k	negative
13	4770*	4*	15k	negative
...	...	...	...	...
10,000	488**	>=60	16k	negative