Anonymization and Data Privacy

Finse Cyber Security Winter School

KARLSTAD
UNIVERSITY
SWEDEN

Meiko Jensen

Karlstad University

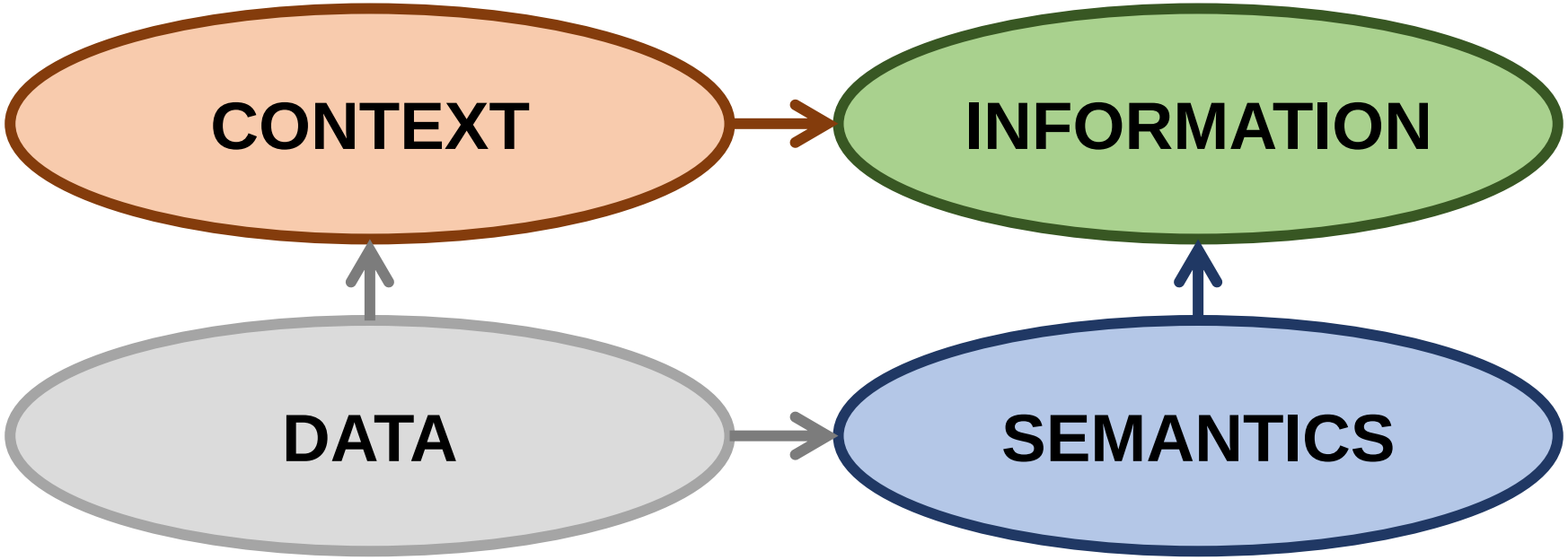
Facts

- 19 000 students
- 265 doctorate/Phd students
- 1 300 employees
- 1,3 billion SEK turnover
- 75 programs
- 750 courses

What is data?
What is information?

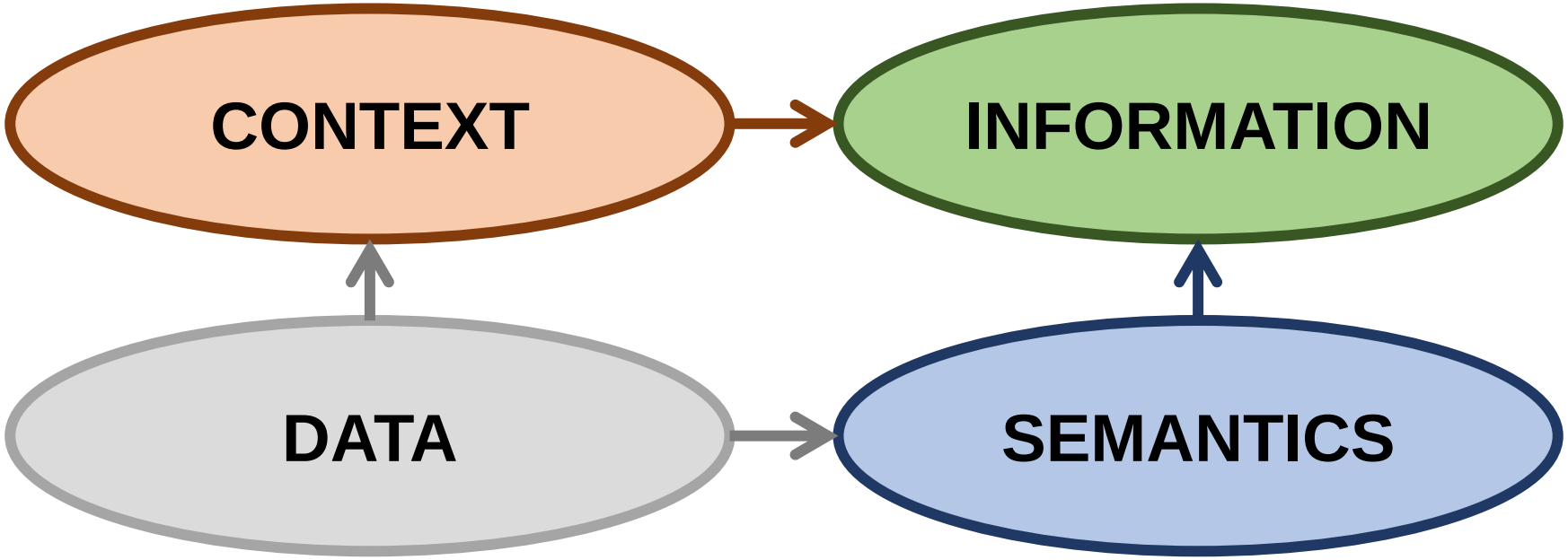
Data vs. Information

- „3!“
- „The number of children I have is: 3!“



Data vs. Information

- „3!“
- „The number of children I have is: 3!“



Data vs. Information

Meiko.Jensen@kau.se
mje@kau.se
student36456@kau.se
info@kau.se
382599341@kau.se
meiko@jensen.name
q853092@nwytg.net

What information can you learn?

AOL publishes „anonymized“ search engine requests of 3 months of 2006

116874	thompson water seal	2006-05-24 11:31:36	1	http://www.thompsonswaterseal.com
116874	express-scripts.com	2006-05-30 07:56:03	1	http://www.express-scripts.com
116874	express-scripts.com	2006-05-30 07:56:03	2	https://member.express-scripts.com/
116874	knbt	2006-05-31 07:57:28		
116874	knbt.com	2006-05-31 08:09:30	1	http://www.knbt.com
117020	naughty thoughts	2006-03-01 08:33:07	2	http://www.naughtythoughts.com
117020	really eighteen	2006-03-01 15:49:55	2	http://www.reallyeighteen.com
117020	texas penal code	2006-03-03 17:57:38	1	http://www.capitol.state.tx.us
117020	hooks texas	2006-03-08 09:47:08		
117020	homicide in hooks texas	2006-03-08 09:47:35		
117020	homicide in bowie county	2006-03-08 09:48:25	6	http://www.tdcj.state.tx.us
117020	texarkana gazette	2006-03-08 09:50:20	1	http://www.texarkanagazette.com
117020	tdcj	2006-03-08 09:52:36	1	http://www.tdcj.state.tx.us
117020	naughty thoughts	2006-03-11 00:04:40	1	http://www.naughtythoughts.com
117020	cupid.com	2006-03-11 00:08:50		

Slides adapted from: M. Hansen

- "fear that spouse is contemplating cheating"
  user no. 7268042
- "how to kill oneself"
  user no. 9486162
- "how to kill your wife"
  user no. 17556639
- "underage lolitas"
  user no. 4797906

Slides adapted from: M. Hansen

school su

safest plac

hand trem

numb fin

taly

or good health

bipolar

verything

The New York Times

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

CAMCORDERS | CAMERAS | CELLPHONES | COMPUTERS | HANDHELDS | HOME VIDEO | MUSIC | PERIPHERALS | WI-

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS
PRINT
SINGLE PAGE
REPRINTS

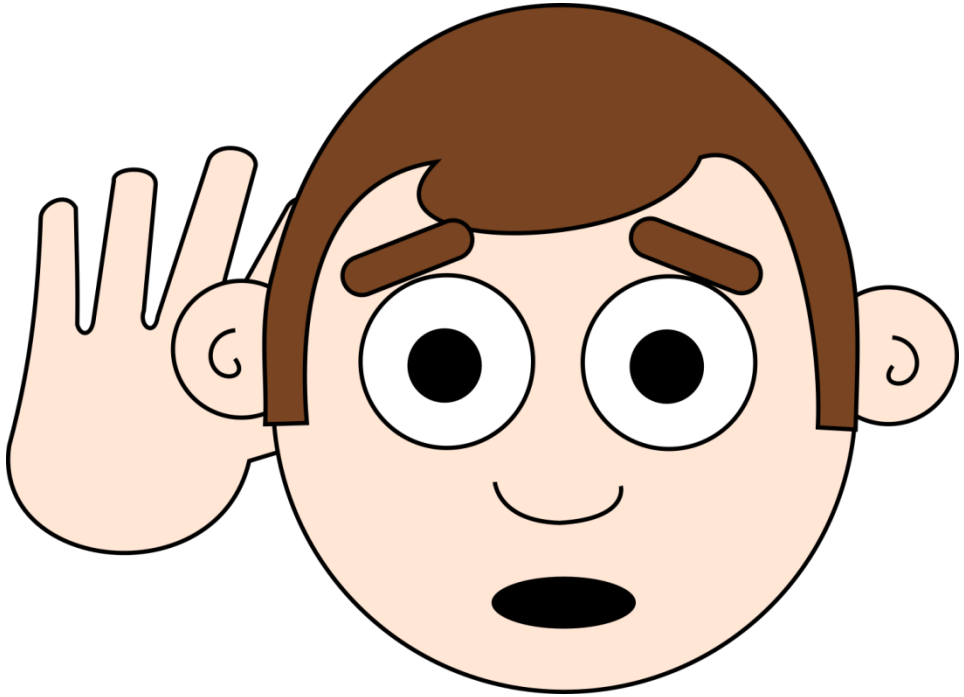
Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

Mrs Arnold said she was shocked that her search queries had been recorded and released to the public by AOL.

"My goodness, it's my whole personal life," she said. "I had no idea somebody was looking over my shoulder."

Slides adapted from: M. Hansen



What would people learn about you knowing only your search queries of the last 3 months?

Types of Data/Information

- Volunteered
  - What you reveal *explicitly when asked*
- Observed
  - What you reveal *implicitly by your behaviour*
- Inferred
  - What is derived from other data about you

[World Economic Forum Report
Personal Data: The Emergence of a New Asset Class]

Types of Data/Information

Data	Metadata
The contents of messages	The context of messages
E-Mail Text	E-Mail Sender/Recipient/Date
Can be spoofed / encrypted	Hard to spoof / encrypt

Types of Data/Information

Data	Metadata
The co mess	
E-Ma	
cry	

"We kill people based on metadata."

General Michael Hayden,
former director of the NSA and the CIA

What is anonymity?

Anonymity

derived from the Greek word ἀνωνυμία, *anonymia*, meaning "without a name“ or "namelessness".

Source: Alice Chodura

Slides adapted from: M. Hansen

Anonymous crowd?

 Source: Dennis Jarvis

Slides adapted from: M. Hansen

Anonymous crowd? – No, not for everybody

 Source: Dennis Jarvis

Slides adapted from: M. Hansen

No, really not anonymous

Distinguishable
(and uniquely
identifiable)
via names
or other
identifiers

Source: Tambako the Jaguar

Slides adapted from: M. Hansen

Identifiers

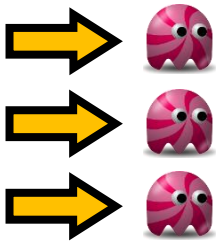
Explicit Identifiers

- Uniquely attributable
  - name
  - phone number
  - address

Alice Kausson

+46 54 7001000

Karlstadsgatan 1



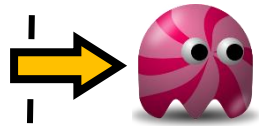
Quasi-Identifiers

- In combination, can uniquely identify
  - birth date
  - gender
  - ZIP code

01.07.80

female

SE 65188

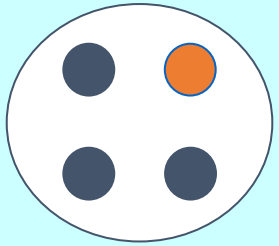


Slides adapted from: L. Martucci

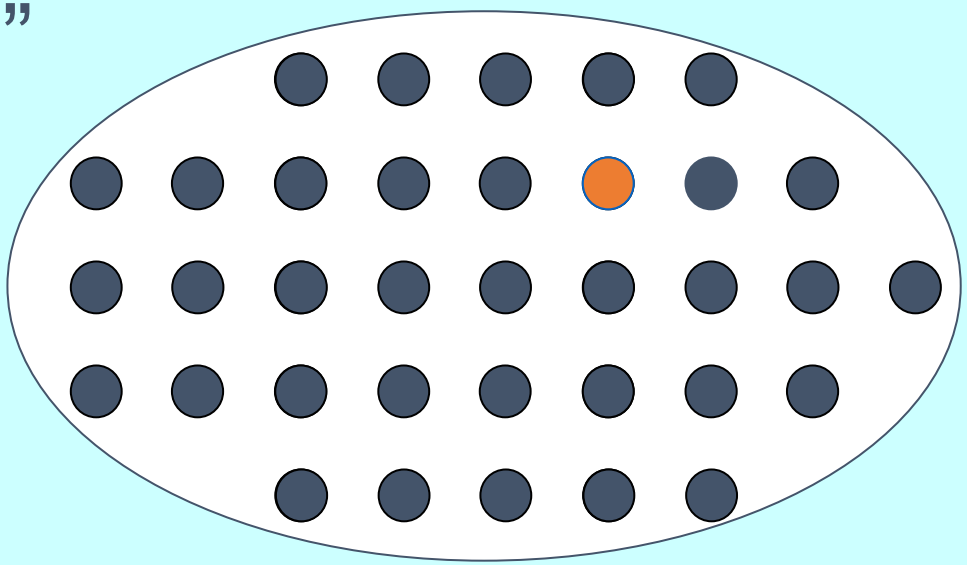
Anonymity Set

- The larger the set of **indistinguishable** entities, the lower the probability of identifying any one of them!

"Hiding in a crowd"



"Less" anonymous (1/4)



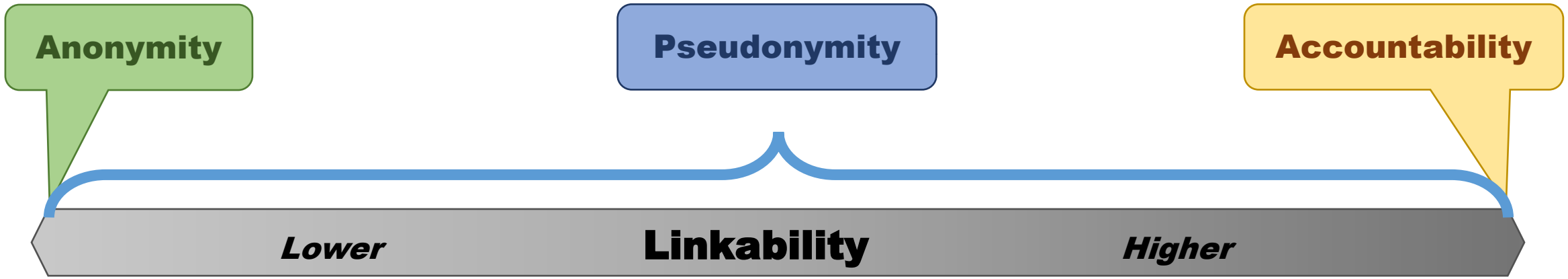
"More" anonymous (1/*n*)

Slides adapted from: M. Hansen

Anonymity vs. Pseudonymity

"Whereas ***anonymity*** and ***accountability*** are the extremes with respect to ***linkability*** to subjects, ***pseudonymity*** is the entire field between and including these extremes. Thus, pseudonymity comprises all degrees of linkability to a subject."

Pfitzmann & Hansen: Anonymity Terminology, 2010

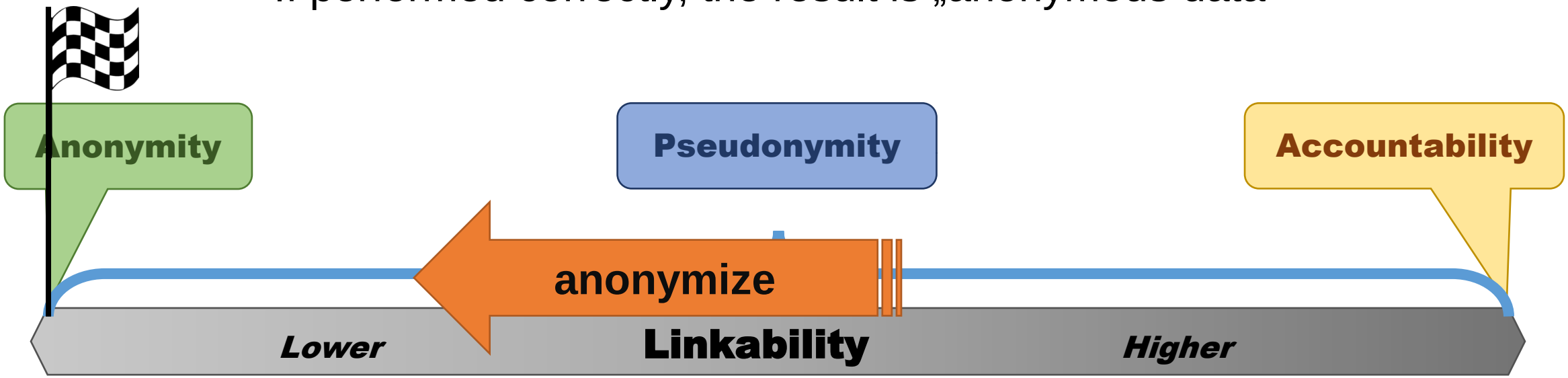


Anonymization

generalize *aggregate* *mask* *blur* *delete* *replace* *perturbate* *synthesize*

Anonymization is the act of processing data in order to **remove** any **linkability** to the subject(s) behind the data

If performed correctly, the result is „anonymous data“



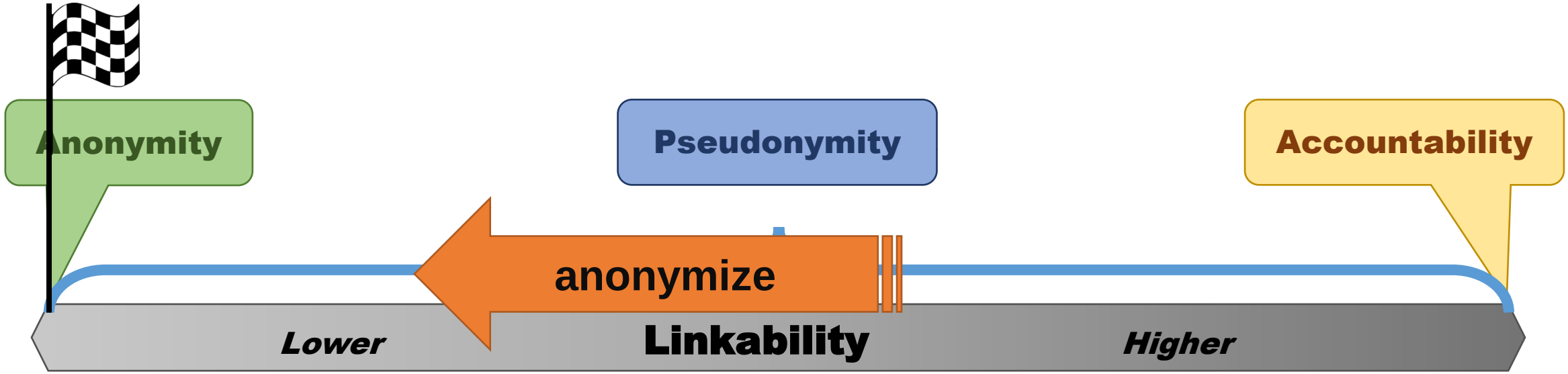
Anonymization

„Once data is truly anonymous

and individuals are no longer identifiable,

the data will not fall within the scope of the GDPR.“

European Data Protection Supervisor



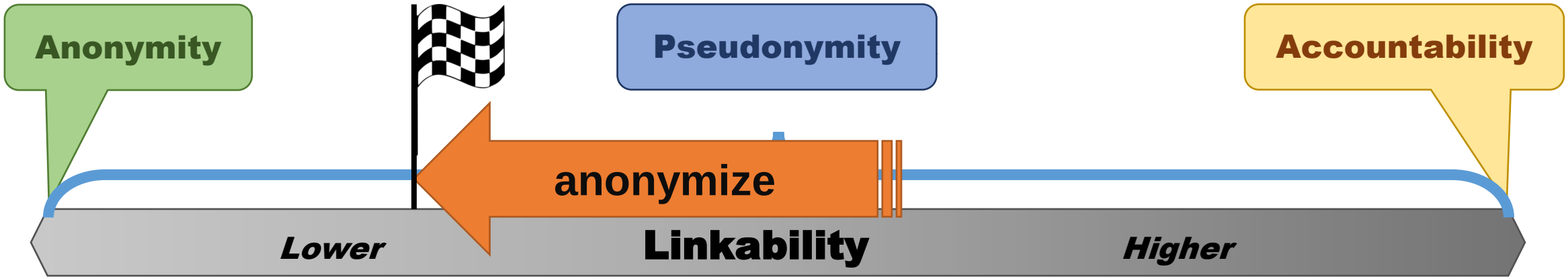
Anonymization

Unfortunately, anonymization techniques **rarely work** reliably!

…and you end up with **pseudonymous** data, not anonymous data!

Anonymization = apply technique to move towards anonymity!

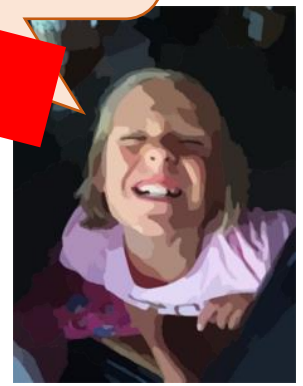
However, reaching anonymity is **not guaranteed!**

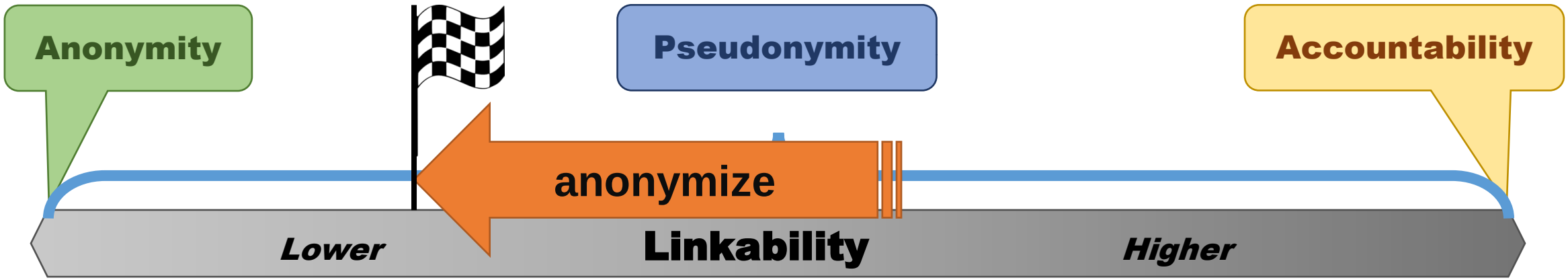


Anonymization

I anonymized the da
so now it is anonym

ushed my teeth,
y they are clean!





Anonymity vs. Pseudonymity

Recall:

AOL published
pseudonymized data,

but claimed it to be
anonymized *data!*

school supplies for Iraq children

the best sea

termites

safest place to live

mature living

nicotine effects on the body

hand tremors

d

60 single men

numb fingers

dog that u

The New York Times

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

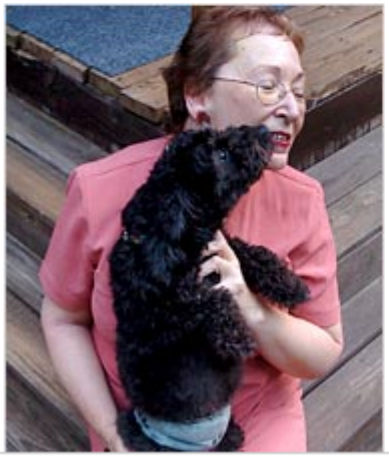
CAMCORDERS | CAMERAS | CELLPHONES | COMPUTERS | HANDHELDS | HOME VIDEO | MUSIC | PERIPHERALS | WI

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS
PRINT
SINGLE PAGE
REPRINTS

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.



No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail

Anonymity vs. Pseudonymity

Recall:

AOL published

pseudonymized data,

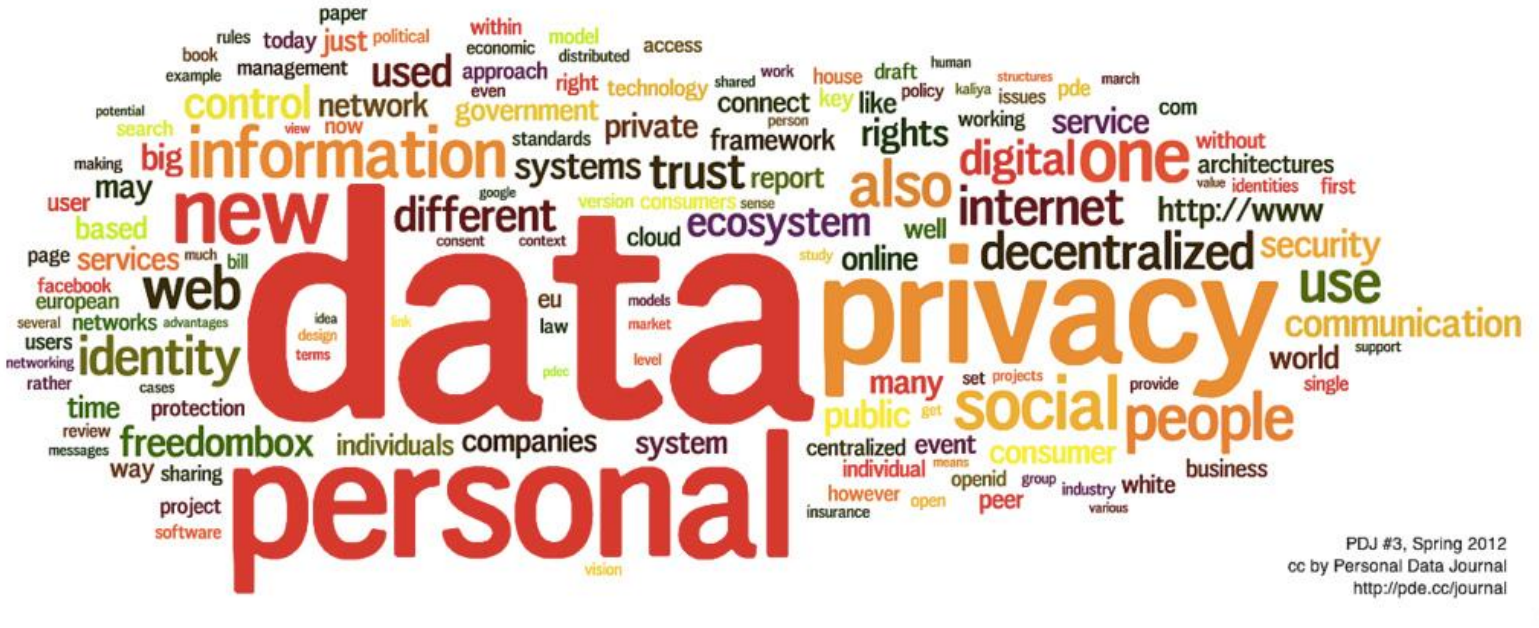
but claimed it to be

anonymized *data!*

Explicit Pseudonym!

116874	th			://www.thompsonswaterseal.com
116874				://www.express-scripts.com
116874	e			ps://member.express-scripts.com/
116874	knbt	2006-05-31 07:57:28		
116874	knbt.com	2006-05-31 08:09:30	1	http://www.knbt.com
117020	naughty thoughts	2006-03-01 08:33:07	2	http://www.naughtythoughts.com
117020	really eighteen	2006-03-01 15:49:55	2	http://www.reallyeighteen.com
117020	texas penal code	2006-03-03 17:57:38	1	http://www.capitol.state.tx.us
117020	hooks texas	2006-03-08 09:47:08		
117020	homicide in hooks texas	2006-03-08 09:47:35		
117020	homicide in bowie county	2006-03-08 09:48:25	6	http://www.tdcj.state.tx.us
117020	texarkana gazette	2006-03-08 09:50:20	1	http://www.texarkanagazette.com
117020	tdcj	2006-03-08 09:52:36	1	http://www.tdcj.state.tx.us
117020	naughty thoughts	2006-03-11 00:04:40	1	http://www.naughtythoughts.com
117020	cupid.com	2006-03-11 00:08:50		

Source: http://www.lunchoverip.com/2006/08/being_user_4417.html



What is pseudonymity?

Pseudonymity

“A **pseudonym** is an **identifier** of a **subject** other than one of the subject’s real names.”

“The subject which the pseudonym refers to is the **holder** of the pseudonym.”

“A subject is **pseudonymous** if a pseudonym is used as identifier instead of one of its real names.“

Pfitzmann & Hansen: Anonymity Terminology, 2010

Pseudonymity

”**Pseudonym**

comes from Greek ***pseudonumon***

meaning ***falsely named***

(*pseudo*: false; *onuma*: name).

Thus, it means a name other than the “real name”.

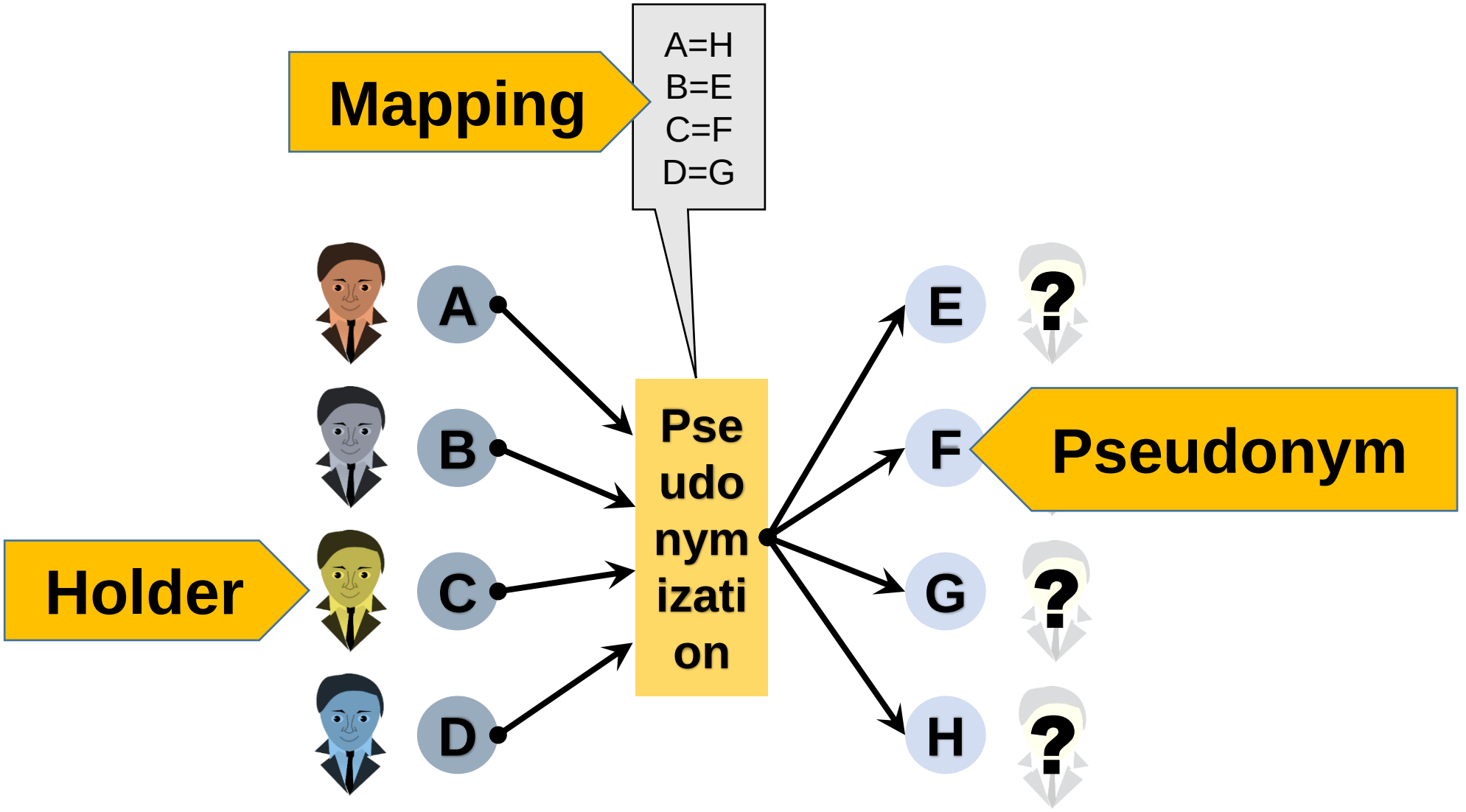
To avoid the connotation of “*pseudo*” = false,

some authors call pseudonyms [..] simply **nyms**.

Pseudonymity is the use of pseudonyms as identifiers.”

Pfitzmann & Hansen: Anonymity Terminology, 2010

Pseudonymization



Example

Name	Study Program	Grade
Aron First	MIE	1.0
Betty Second	MIE	3.3
Carl Third	MIE	2.7
Denise Fourth	INI	2.0
Eddy Fifth	INI	5.0
Fae Sixth	INI	5.0
Gerald Seventh	INI	1.7
Hannah Eigth	BDS	1.3
Igor Ninth	BDS	4.0

Example

Pseudonym

Matriculation Number	Study Program	Grade
9200189	MIE	1.0
9200198	MIE	3.3
9200127	MIE	2.7
9200117	INI	2.0
9200226	INI	5.0
9200228	INI	5.0
9200298	INI	1.7
9200201	BDS	1.3
9200204	BDS	4.0

Pseudonym Types

- **Public pseudonym:**

  The linking between a public pseudonym and its holder may be ***publicly known*** even from the very beginning.
  Example: linking could be listed in public directories such as the entry of a phone number in combination with its owner.

- **Initially non-public pseudonym:**

  The linking between an initially non-public pseudonym and its holder may be ***known by certain parties***, but is ***not public*** at least initially.
  Example: a bank account where the bank can look up the linking may serve as a non-public pseudonym. For some specific non-public pseudonyms, certification authorities acting as identity brokers could reveal the civil identity of the holder in case of abuse.

- **Initially unlinked pseudonym:**

  The linking between an initially unlinked pseudonym and its holder is – at least initially – ***not known to anybody*** with the possible exception of the holder himself/herself.
  Example: (non-public) biometrics like DNA information unless stored in databases including the linking to the holders.

Source: Pfitzmann & Hansen: Anonymity Terminology

Pseudonym Types

- **Person Pseudonym**
  - Bound to **human individual**
  - A person pseudonym is a substitute for the holder's name which is regarded as representation for the holder's civil identity. It may be used in many different contexts, e.g., a number of an identity card, the social security number, DNA, a nickname, the pseudonym of an actor, or a mobile phone number.
- **Role Pseudonym**
  - Bound to the **role** of a human individual in a **context**
  - The use of role pseudonyms is limited to specific roles, e.g., a customer pseudonym or an Internet account used for many instantiations of the same role "Internet user". The same role pseudonym may be used with different communication partners. Roles might be assigned by other parties, e.g., a company, but they might be chosen by the subject himself/herself as well.
- **Relationship Pseudonym**
  - Bound to the **relation of a pair** (or more) **of individuals** in a specific context
  - For each communication partner, a different relationship pseudonym is used. The same relationship pseudonym may be used in different roles for communicating with the same partner. Examples are distinct nicknames for each communication partner.

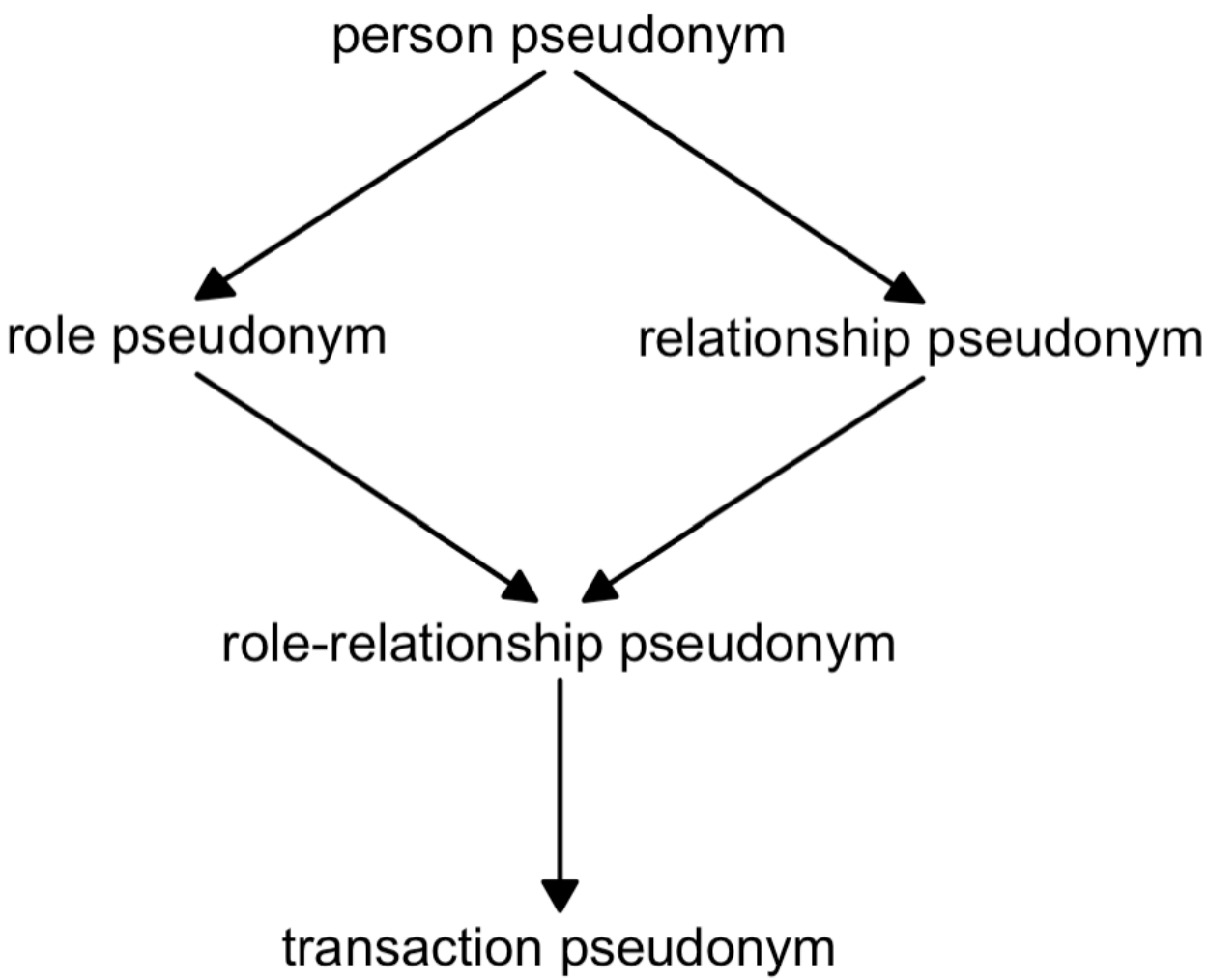
Source: Pfitzmann & Hansen: Anonymity Terminology

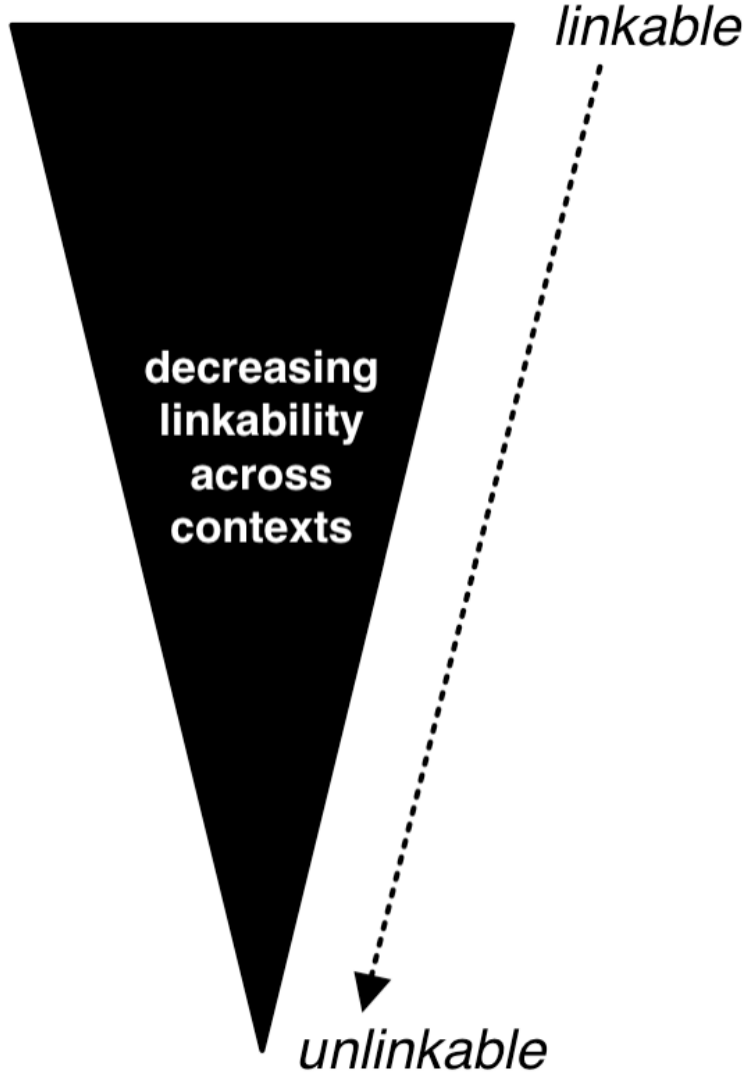
Pseudonym Types

- **Role-Relationship Pseudonym**
  - Bound to all **role-relation-combinations** in a set of individuals in a specific context
  - For each role and for each communication partner, a different role-relationship pseudonym is used. This means that the communication partner does not necessarily know, whether two pseudonyms used in different roles belong to the same holder. On the other hand, two different communication partners who interact with a user in the same role, do not know from the pseudonym alone whether it is the same user.
- **Transaction Pseudonym**
  - Bound to each **single transaction** (or interaction) between any individuals in any roles in a specific context
  - For each transaction, a transaction pseudonym unlinkable to any other transaction pseudonyms [..] is used, e.g., randomly generated transaction numbers for online-banking. Therefore, transaction pseudonyms can be used to realize as strong anonymity as possible.

Source: Pfitzmann & Hansen: Anonymity Terminology

Pseudonym Types





Source: Pfitzmann & Hansen: Anonymity Terminology

Pseudonym Creation

- **Self-chosen Pseudonym**

  Arbitrary sequence of characters chosen by yourself („nickname“)
  - "Mike-O"
  - „FinseRulez2022"
- **Self-created Pseudonym**

  Still created by yourself, but follows a fixed data format / creation algorithm
  - Random number picked yourself
  - Public key of keypair used in Blockchains
- **Centrally Assigned Pseudonym**

  Assigned to you by a central pseudonym creation authority
  - Customer-ID
  - Taxation-ID
  - Student Matriculation Number

Pseudonym Creation

- **Issues with Self-chosen/-defined Pseudonyms:**
  - **Accidential Collisions**
    ("picked the same pseudonym")
  - **Linkage / Information via Pseudonym Text**
    ("likes Finse")
  - **Context Escape**
    ("google the pseudonym, learn the identity")
  - **New Attack Vector: Intentional Collision**
    ("I am Brian!"–"No, I am Brian!")

Pseudonymization Techniques

- **Random Number / Pseudonym Assignment**

  Choose a (truly random) number / pseudonym per identity

  - Make sure different identities are mapped to different numbers / pseudonyms
  - Make sure same identities are mapped to same numbers / pseudonyms

- **Increasing Counter Number Assignment**

  Assign numbers from a counter that is increased with every new pseudonym issued

  - E.g. customer ID‘s, session ID‘s
  - Automatically assigns different pseudonyms to different identities
  - Same identities might get mapped to different pseudonyms!

- **Hashing**

  Map identity to hash value of identity

  - `pseudonym = hash(identity)`
  - Automatically assigns same pseudonyms to same identities
  - Different identities might get mapped to same pseudonyms *(hash collision)!*

…all of these have their issues!

Attacks on Pseudonymization

Matriculation Number	Study Program	Grade
9200189	MIE	1.0
9200198	MIE	3.3
9200127	MIE	2.7
9200117	INI	2.0

Learn identity from quasi-identifiers!

Attacks on Pseudonymization

Matriculation Number	Study Program	Grade
9200189	MIE	1.0
9200198	MIE	1.0
9200127	MIE	5.0

Learn identity from background knowledge!

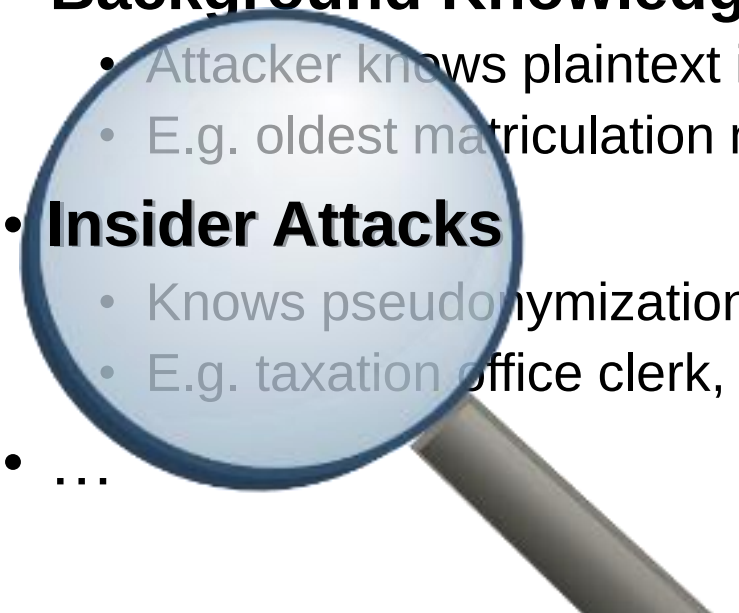
Attacks on Pseudonymization

Matriculation Number	Study Program	Grade
9189726	MIE	1.0
9200198	MIE	3.3
9200127	MIE	2.7
9200117	INI	2.0
9200226	INI	5.0
9200228	INI	5.0
9200298	INI	1.7
9200201	BDS	1.3
9200204	BDS	4.0

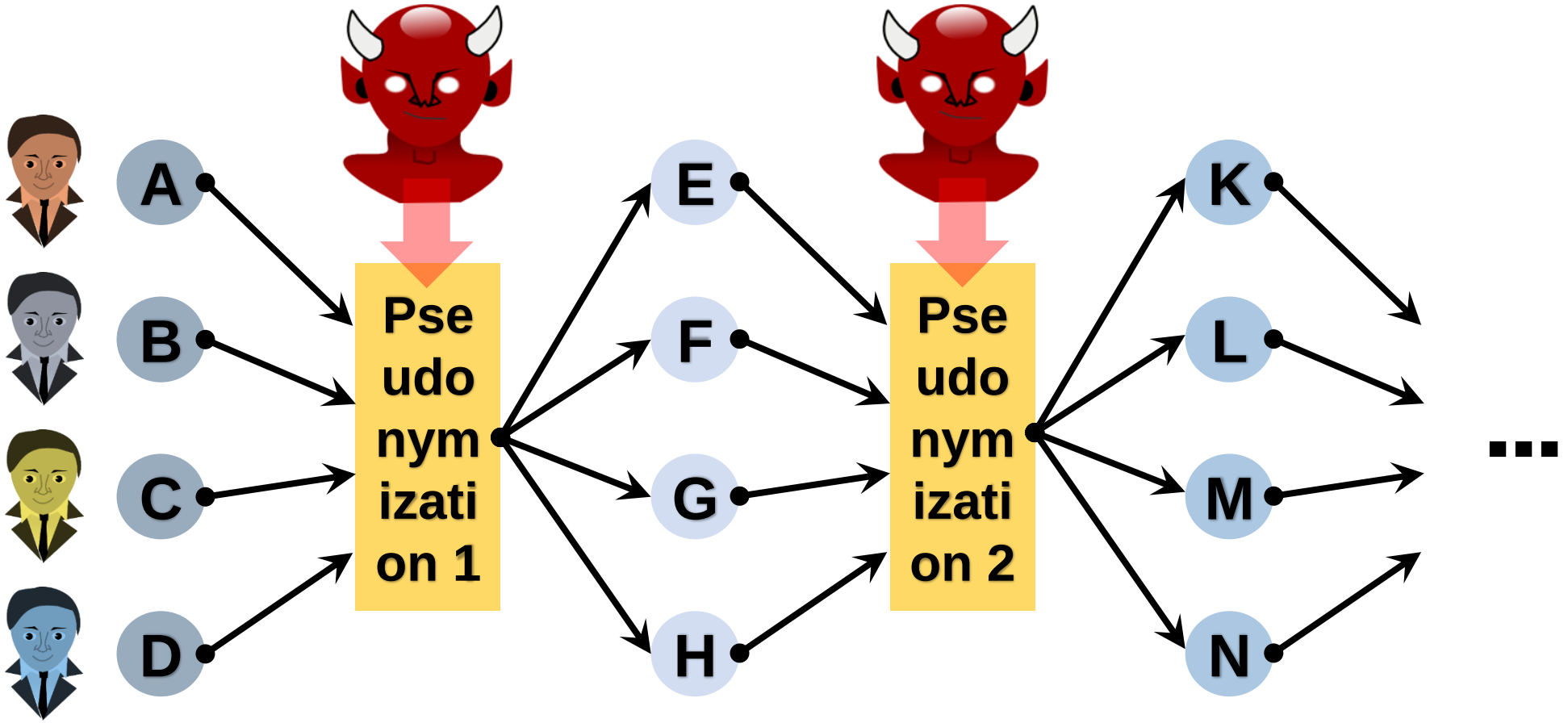
Learn identity from background knowledge!

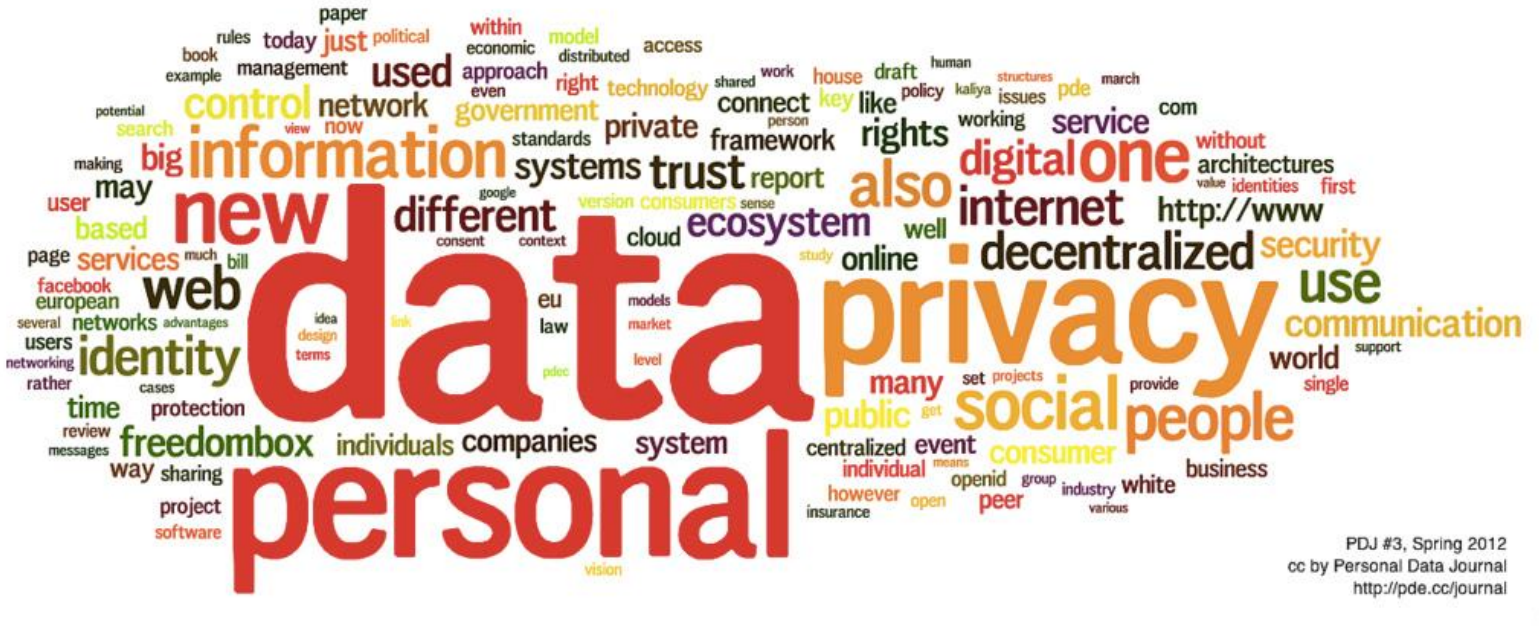
Attacks on Pseudonymization

- **Dictionary Attack**
  - Generate all possible pseudonyms for most likely inputs
  - E.g. hash values of english words as possible passwords
- **Brute Force / Rainbow Tables**
  - Generate all possible pseudonyms for all possible inputs
  - E.g. hash values of all possible IP addresses
- **Background Knowledge Attacks**
  - Attacker knows plaintext information linked to pseudonym
  - E.g. oldest matriculation number
- **Insider Attacks**
  - Knows pseudonymization mapping!
  - E.g. taxation office clerk, or professor
- …



Pseudonymization Chains



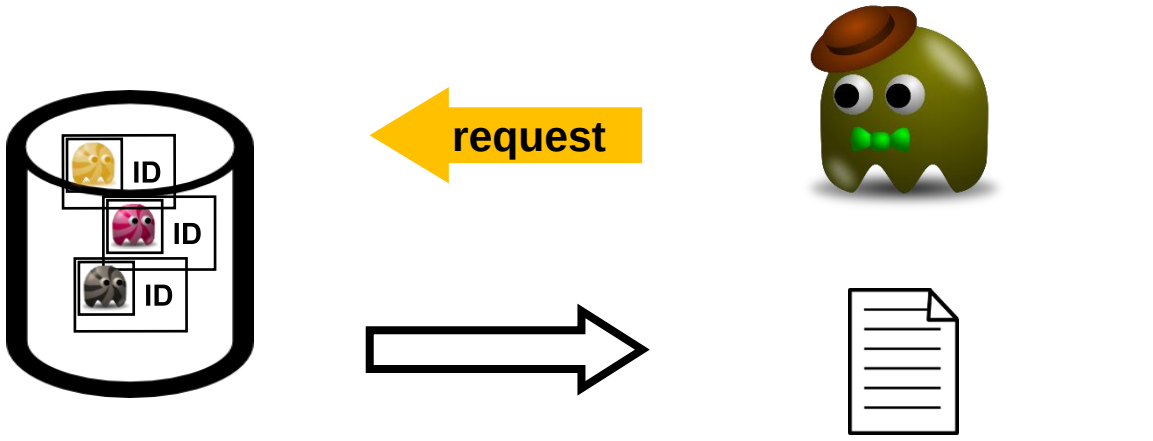


What is k-anonymity?

k-anonymity

- How to use a database that has personal data stored…

…and NOT disclose personal data?



no personal data!

Slides adapted from: L. Martucci

Types of Identifiers

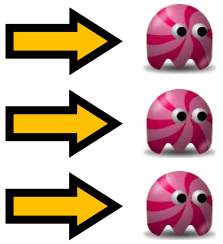
Explicit Identifiers

- Uniquely attributable { name, phone number, address }

Alice Kausson

+46 54 7001000

Karlstadsgatan 1



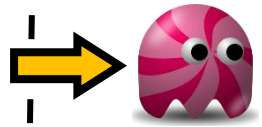
Quasi-Identifiers

- In combination, can uniquely identify { birth date, gender, ZIP code }

01.07.80

female

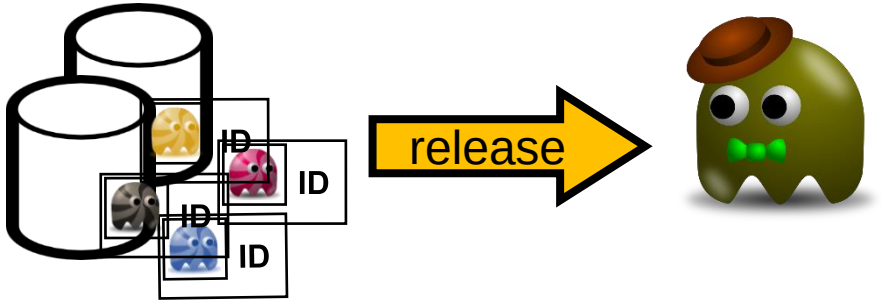
SE 65188



Slides adapted from: L. Martucci

k-anonymity

- Goal: to prevent re-identification of individuals when releasing data



- k-anonymity property:

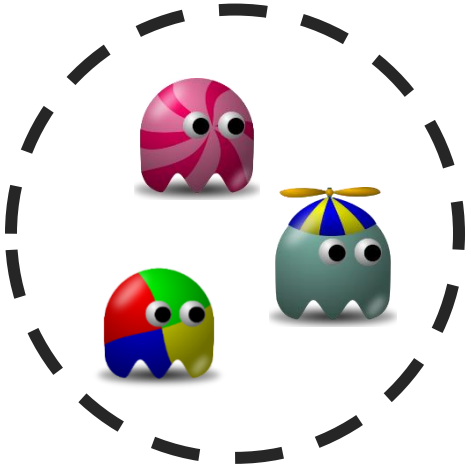
on data release, information about a subject **cannot be distinguished from at least k-1 other individuals**

Slides adapted from: L. Martucci

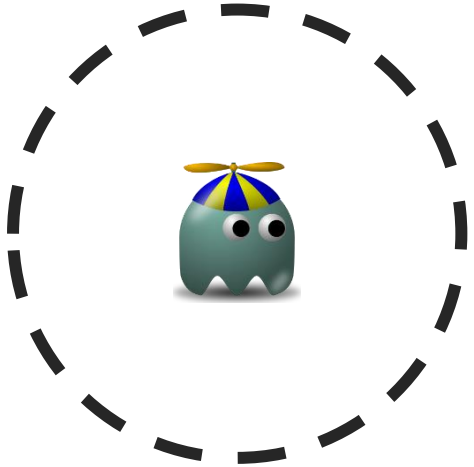
k-anonymity

- Measure for the anonymity set

where min(k) = 2 (k = 1 means NO anonymity)



k = 3



k = 1

no anonymity!

Slides adapted from: L. Martucci

Example: building a k=2 release

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	11.03.79	male	1072	married	1	A
	17.03.79	male	1276	married	7	B
	01.07.80	female	1073	single	2	B
	07.09.84	female	1077	single	0	C
	02.07.89	male	1016	single	2	D
	21.09.91	female	1267	it‘s complicated	4	E
	24.12.98	female	1268	it‘s complicated	4	A

Slides adapted from: L. Martucci

Example: building a k=2 release

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	11.03.79	male	1072	married	1	A
	17.03.79	male	1276	married	7	B
	01.07.80	female	1073	single	2	B
	07.09.84	female	1077	single	0	C
	02.07.89	male	1016	single	2	D
	21.09.91	female	1267	it‘s complicated	4	E
	24.12.98	female	1268	it‘s complicated	4	A

Explicit Identifier

Quasi-Identifiers

Released data

Slides adapted from: L. Martucci

Remove Name Field

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	11.03.79	male	1072	married	1	A
	17.03.79	male	1276	married	7	B
	01.07.80	female	1073	single	2	B
	07.09.84	female	1077	single	0	C
	02.07.89	male	1016	single	2	D
	21.09.91	female	1267	it‘s complicated	4	E
	24.12.98	female	1268	it‘s complicated	4	A

Slides adapted from: L. Martucci

Generalize Birth date to Range

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1072	married	1	A
	1970‘s	male	1276	married	7	B
	1980‘s	female	1073	single	2	B
	1980‘s	female	1077	single	0	C
	1980‘s	male	1016	single	2	D
	1990‘s	female	1267	it‘s complicated	4	E
	1990‘s	female	1268	it‘s complicated	4	A

Slides adapted from: L. Martucci

The Gender Field

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1072	married	1	A
	1970‘s	male	1276	married	7	B
	1980‘s	female	1073	single	2	B
	1980‘s	female	1077	single	0	C
	1980‘s	male	1016	single	2	D
	1990‘s	female	1267	it‘s complicated	4	E
	1990‘s	female	1268	it‘s complicated	4	A

NOT k=2 here

Slides adapted from: L. Martucci

Generalize Gender Field

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1072	married	1	A
	1970‘s	male	1276	married	7	B
	1980‘s	**ghost**	1073	single	2	B
	1980‘s	**ghost**	1077	single	0	C
	1980‘s	**ghost**	1016	single	2	D
	1990‘s	female	1267	it‘s complicated	4	E
	1990‘s	female	1268	it‘s complicated	4	A

Slides adapted from: L. Martucci

OR Suppress Information

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
👻	1970‘s	male	1072	married	1	A
👻	1970‘s	male	1276	married	7	B
👻	1980‘s	female	1073	single	2	B
👻	1980‘s	female	1077	single	0	C
*	*	*	*	*	*	*
👻	1990‘s	female	1267	it‘s complicated	4	E
👻	1990‘s	female	1268	it‘s complicated	4	A

Slides adapted from: L. Martucci

Generalize ZIP data

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1***	married	1	A
	1970‘s	male	1***	married	7	B
	1980‘s	ghost	10**	single	2	B
	1980‘s	ghost	10**	single	0	C
	1980‘s	ghost	10**	single	2	D
	1990‘s	female	12**	it‘s complicated	4	E
	1990‘s	female	12**	it‘s complicated	4	A

Slides adapted from: L. Martucci

Civil Status Field is k=2!

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1***	married	1	A
	1970‘s	male	1***	married	7	B
	1980‘s	ghost	10**	single	2	B
	1980‘s	ghost	10**	single	0	C
	1980‘s	ghost	10**	single	2	D
	1990‘s	female	12**	it‘s complicated	4	E
	1990‘s	female	12**	it‘s complicated	4	A

Slides adapted from: L. Martucci

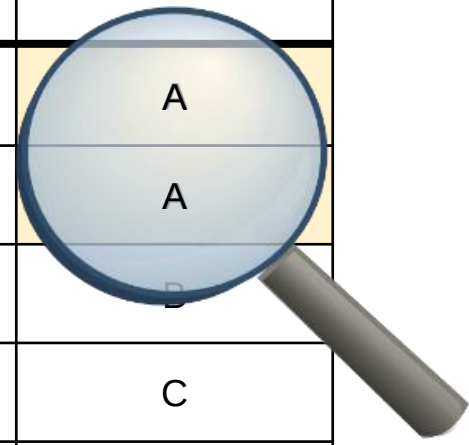
Homogeneity Attack on k-anonymity

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1***	married	1	**A**
	1970‘s	male	1***	married	7	**A**
	1980‘s	ghost	10**	single	2	B
	1980‘s	ghost	10**	single	0	C
	1980‘s	ghost	10**	single	2	D
	1990‘s	female	12**	it‘s complicated	4	E
	1990‘s	female	12**	it‘s complicated	4	A

Slides adapted from: L. Martucci

Homogeneity Attack on k-anonymity

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1***	married	1	A
	1970‘s	male	1***	married	7	A
	1980‘s	ghost	10**	single	2	B
	1980‘s	ghost	10**	single	0	C
						D
						E
						A



 **is from the 1970‘s ➔**  **has Diagnosis A!**

Slides adapted from: L. Martucci

l-diversity and t-closeness

Small L, not large i

l-diversity

- Addresses two attacks on k-anonymity
  - Homogeneity attack
  - Background knowledge attack

BUT

- Difficult, sometimes unnecessary
- Insufficient to prevent attribute disclosure
- it does not consider overall data distribution
- it does not consider semantics

t-closeness

- Addresses l-diversity limitations
- Metric is the attacker's information gain

BUT

- No computational procedure
- Limitations on the utility of data releases

Slides adapted from: L. Martucci

If you want to know more

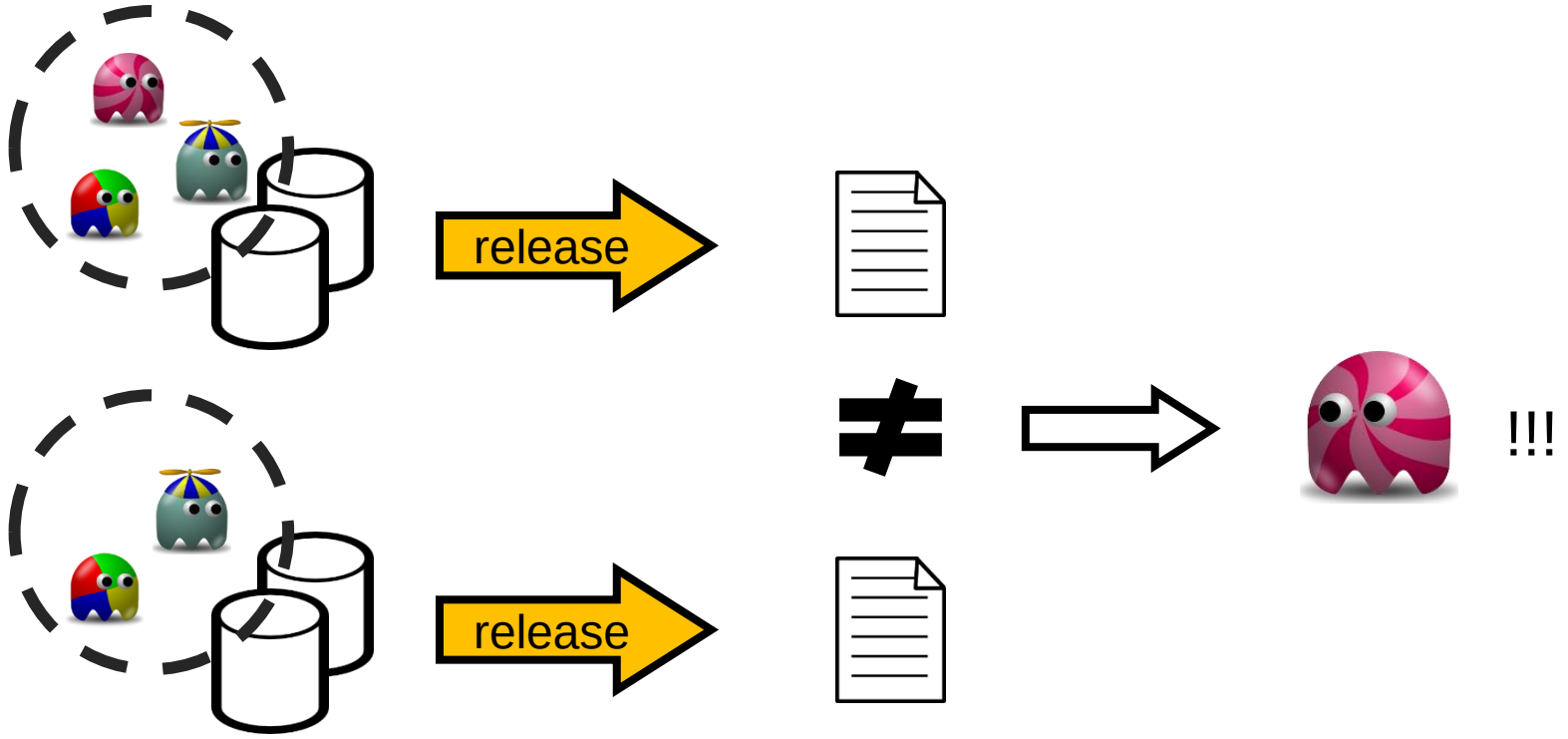
- Sweeney, L.: k-Anonymity: a Model for Protecting Privacy. Int. J. Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
- Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: Int Conf Data Engineering, ICDE 2006.
- Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Int Conf Data Engineering, ICDE 2007.

Slides adapted from: L. Martucci

What is differential privacy?

Releasing Personal Data

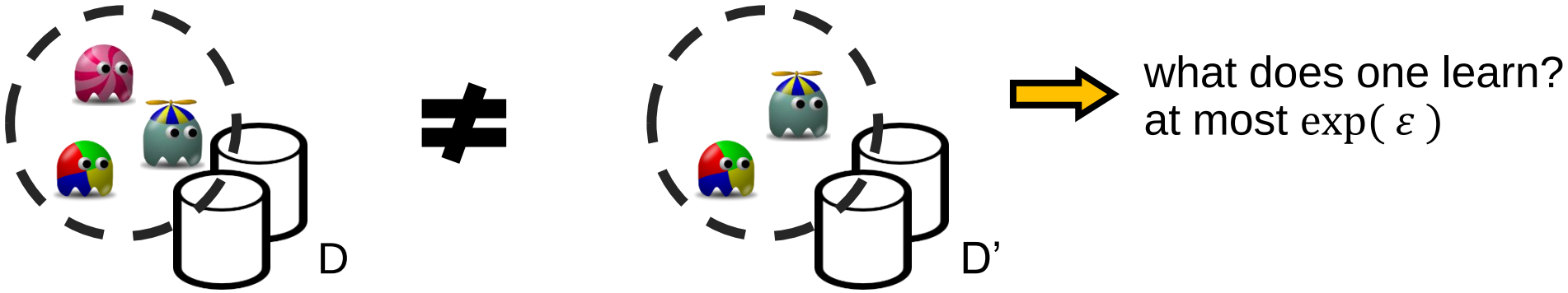
- Looking into two data releases: (from a statistical database)



Slides adapted from: L. Martucci

Differential Privacy

- Quantify the difference in what might be learned about any individual ( ) from a database with or without said individual



- Bound the risk to a factor of ε

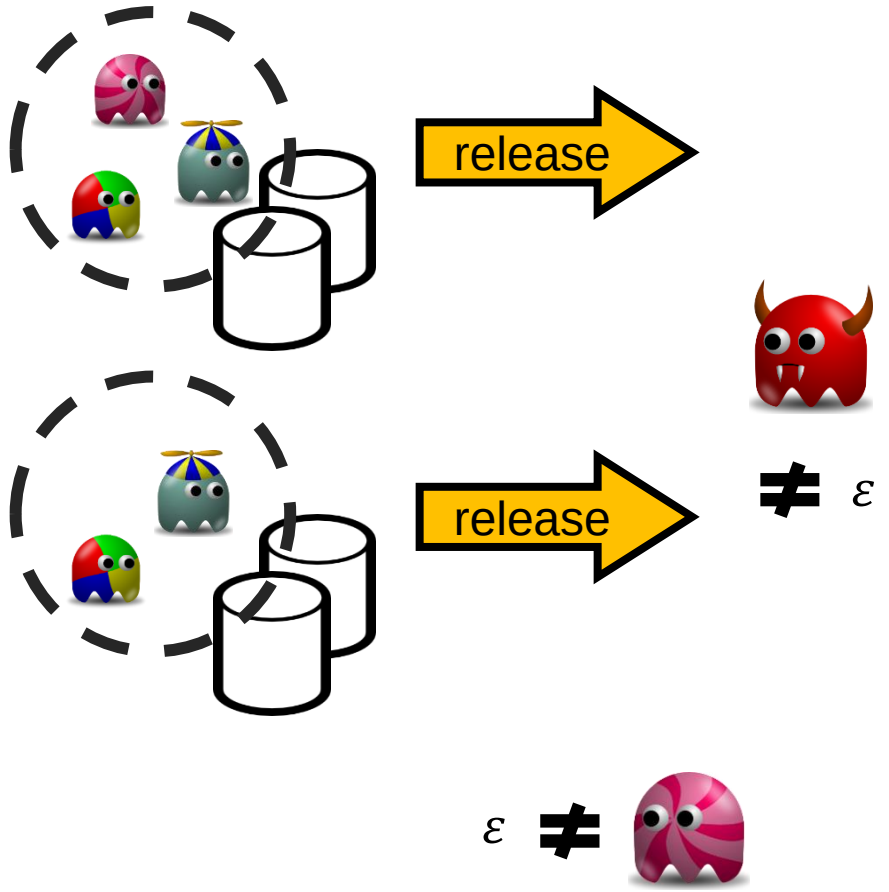
See
- Cynthia Dwork: Differential Privacy.
  In: 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006). Springer, Juli 2006
- Cynthia Dwork, Frank McSherry, Kobbi Nissim, Adam Smith: *Calibrating Noise to Sensitivity in Private Data Analysis*.
  In: Shai Halevi, Tal Rabin (Hrsg.): *Theory of Cryptography*. Springer, 2006, ISBN 978-3-540-32731-8,

Slides adapted from: L. Martucci

Differential Privacy

- Meaning:

an attacker ( ) is not able to learn any additional information that she could not learn if the participant had opted out.



Slides adapted from: L. Martucci

How to do it?

- Add noise to the query result ⇨ how? it depends on...
  - the mechanism design
  - and the type of data.

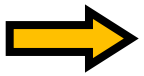
exponential mechanism ⇨ categorical data

Laplace mechanism ⇨ numerical data

Slides adapted from: L. Martucci

Limitations

- Differential Privacy does not mean that  learns nothing about  from the results

 mind the background information!

Slides adapted from: L. Martucci

RAPPOR

- **RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response**
  by Úlfar Erlingsson, Vasyl Pihur, Aleksandra Korolova (Google, USC)
- Built into Google Chrome browser
  - Detection of malicious websites
  - Problem:
    - Community wants to learn which websites are hosting Malware
    - Individual does not want to reveal which websites it has visited

Details:

https://security.googleblog.com/2014/10/learning-statistics-with-privacy-aided.html

https://github.com/google/rappor

RAPPOR

"On the Internet, nobody knows you're a dog."

RAPPOR

Q: Are you a dog?

Yes! **Yes!** **No!** **Yes!**

Yes! **No!** **Yes!** **Yes!**

Yes! **No!** **Yes!** **Yes!**

Central Data Collector

1. Flip a coin!

Coin = 1

2. Always respond **Yes!**

A: **Yes!**

Coin = 0

2. Be honest!

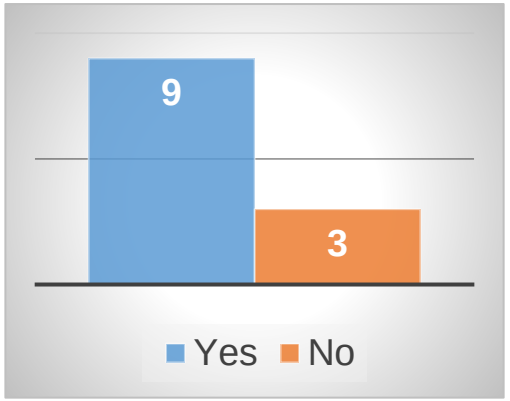
A: **Yes!**

A: **No!**

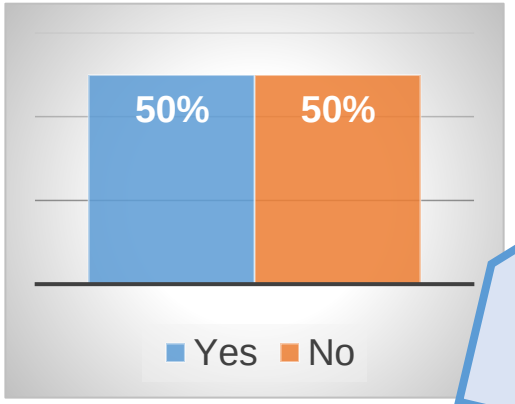
Local Data Processors

RAPPOR

Frequency Analysis

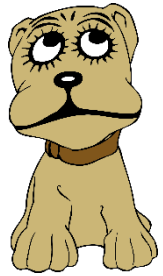


Subtract n/2 from „Yes“, as they were lies…



Statistical analysis is feasible!

„On the Internet, half of all users are dogs!“



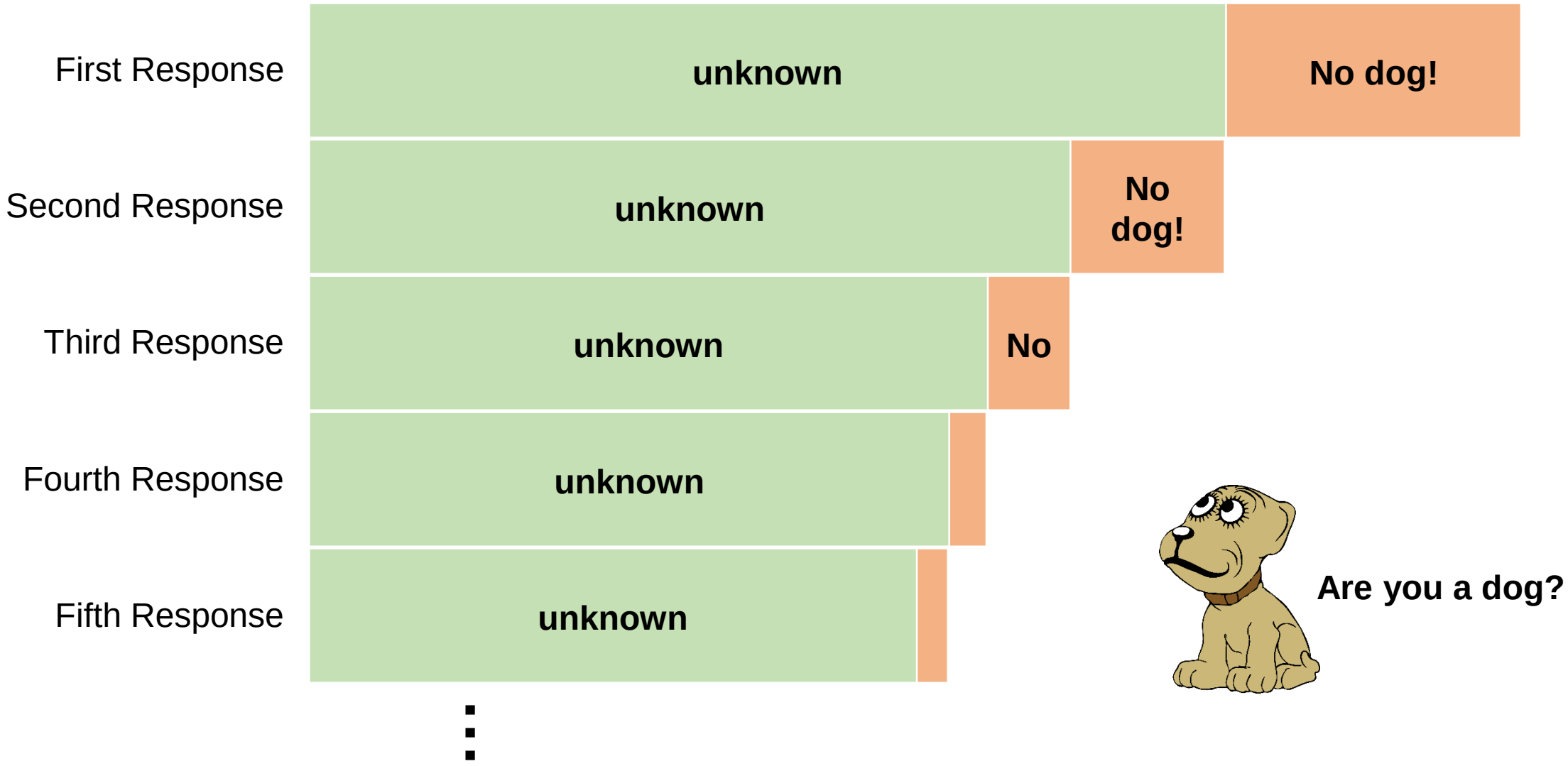
Individual data entries cannot be used to learn about persons!

„You might or might not be a dog…“

RAPPOR

- **In general**:
  - Add random noise to the statistical dataset
    - at the individual data sensors
    - Prior to sending the data to the collector
  - Aggregated dataset then does not contain the noise-free individual data
  - ε-differential privacy, with $\varepsilon = \ln(0.75 / (1 - 0.75))$
  - Can be extended to other types of queries
    (e.g. scaled queries like „give a 5-star rating“)
- **Problem**:
  - If you repeat asking the same question to the same person,
    you learn the correct answer with increasing probability

RAPPOR


First Response
unknown
No dog!
Second Response
unknown
No dog!
Third Response
unknown
No
Fourth Response
unknown
Fifth Response
unknown
Are you a dog?

How about more complex data?