Datenreduktion vor Herausgabe von Informationen -- der Werkzeugkasten der Kryptographen

KARLSTAD UNIVERSITY SWEDEN

29.06.2022

Prof. Dr.-Ing. Meiko Jensen

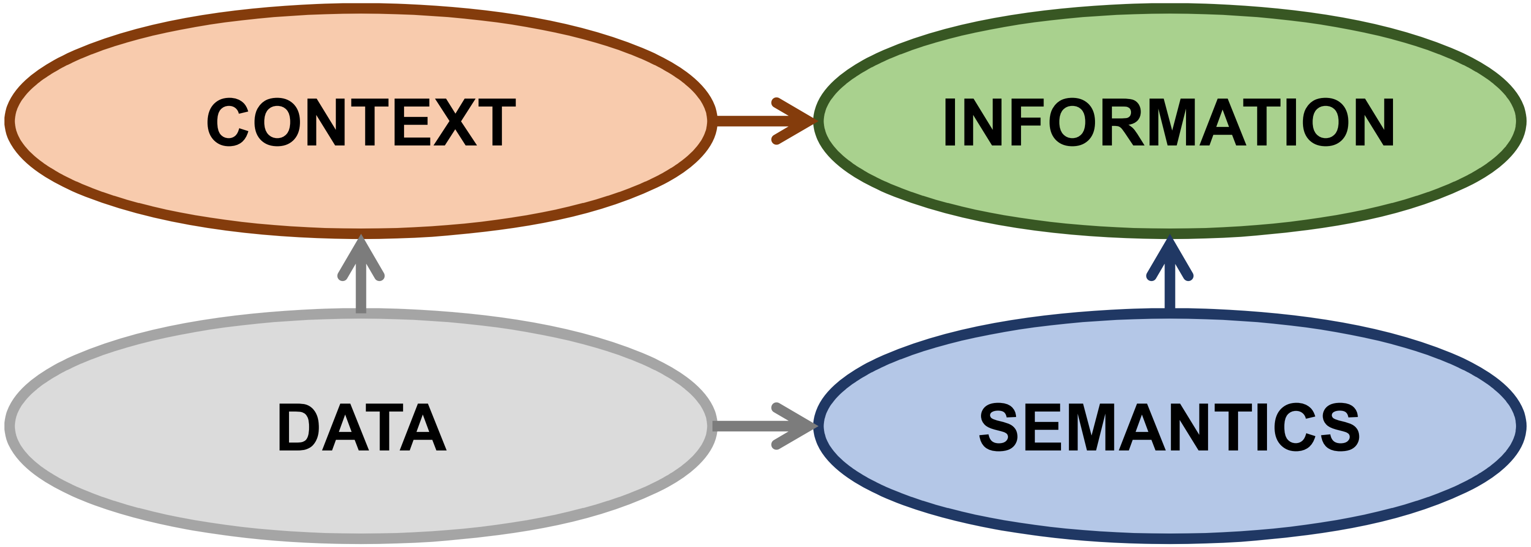
CC-BY-4.0

Agenda

- Data vs. Information
- Issues with information disclosure
- Techniques for information reduction
  - Pseudonymization
  - K-Anonymity
  - Differential Privacy
- Techniques for information documentation
  - Digital Signatures
  - Advanced Digital Signatures
- Summary

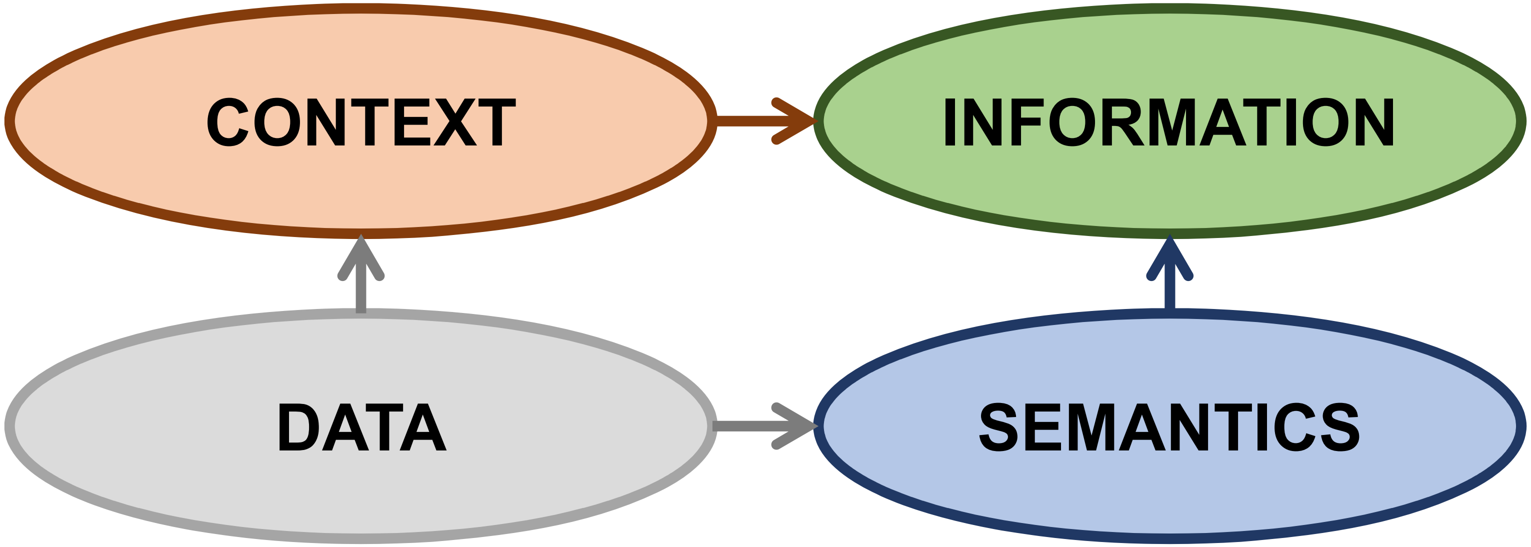
Data vs. Information

- „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

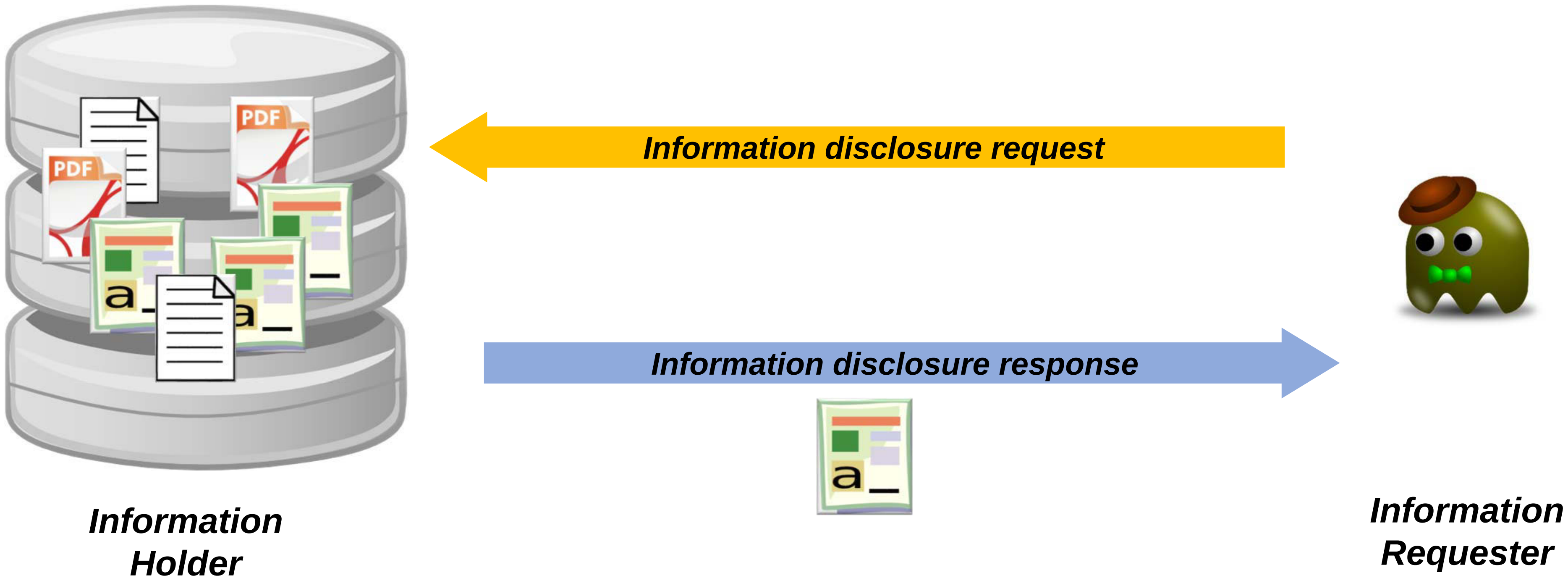
382599341@kau.se

meiko@jensen.name

q853092@nwytg.net

The information disclosure scenario

Information Disclosure Scenario



What could go wrong?

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		

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

school sup | **aly**

safest place | **r good health**

hand tremo | **bipolar**

numb fing | **erything**

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 | WH

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.

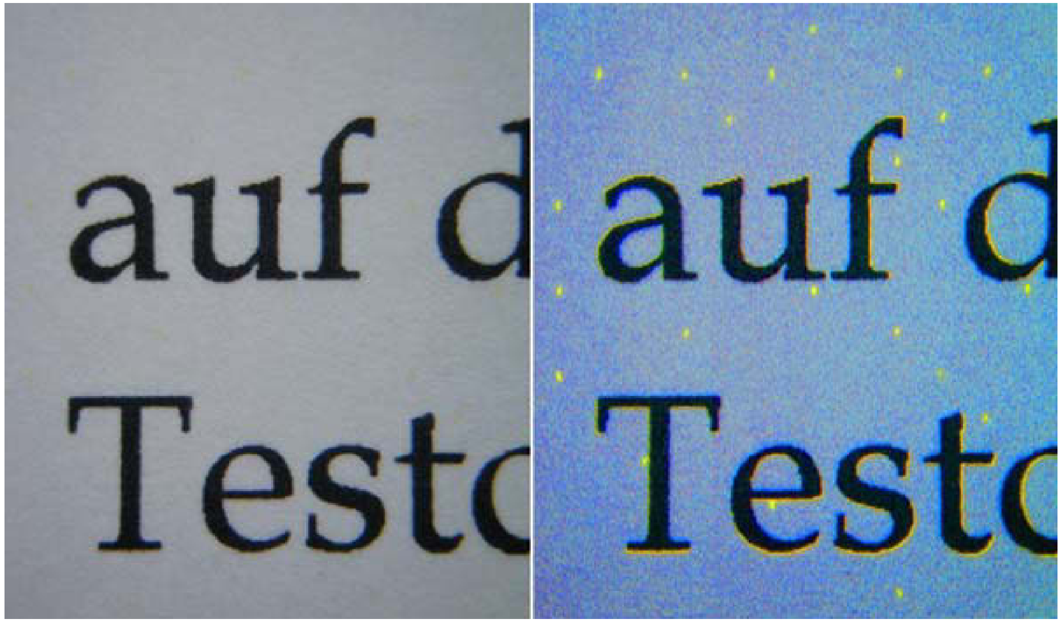
No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

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

Machine Identification Codes

“A Machine Identification Code (MIC) [..] is a digital watermark which certain color laser printers and copiers leave on every printed page, allowing identification of the device which was used to print a document.”



Source: Wikipedia / Florian Heise

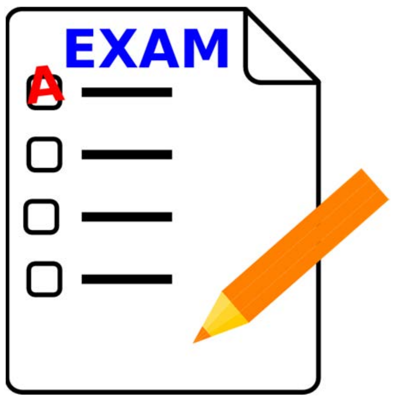
Machine Identification Codes

- Has led to identification and arrest of whistleblower Reality Leigh Winner
- Leaked NSA documents on russian interference with US elections in 2016
- Leaked documents were scanned and published
- Yellow dots found in the scans by the FBI
- Her printer was identified ➔ she was identified

Source: Wikipedia

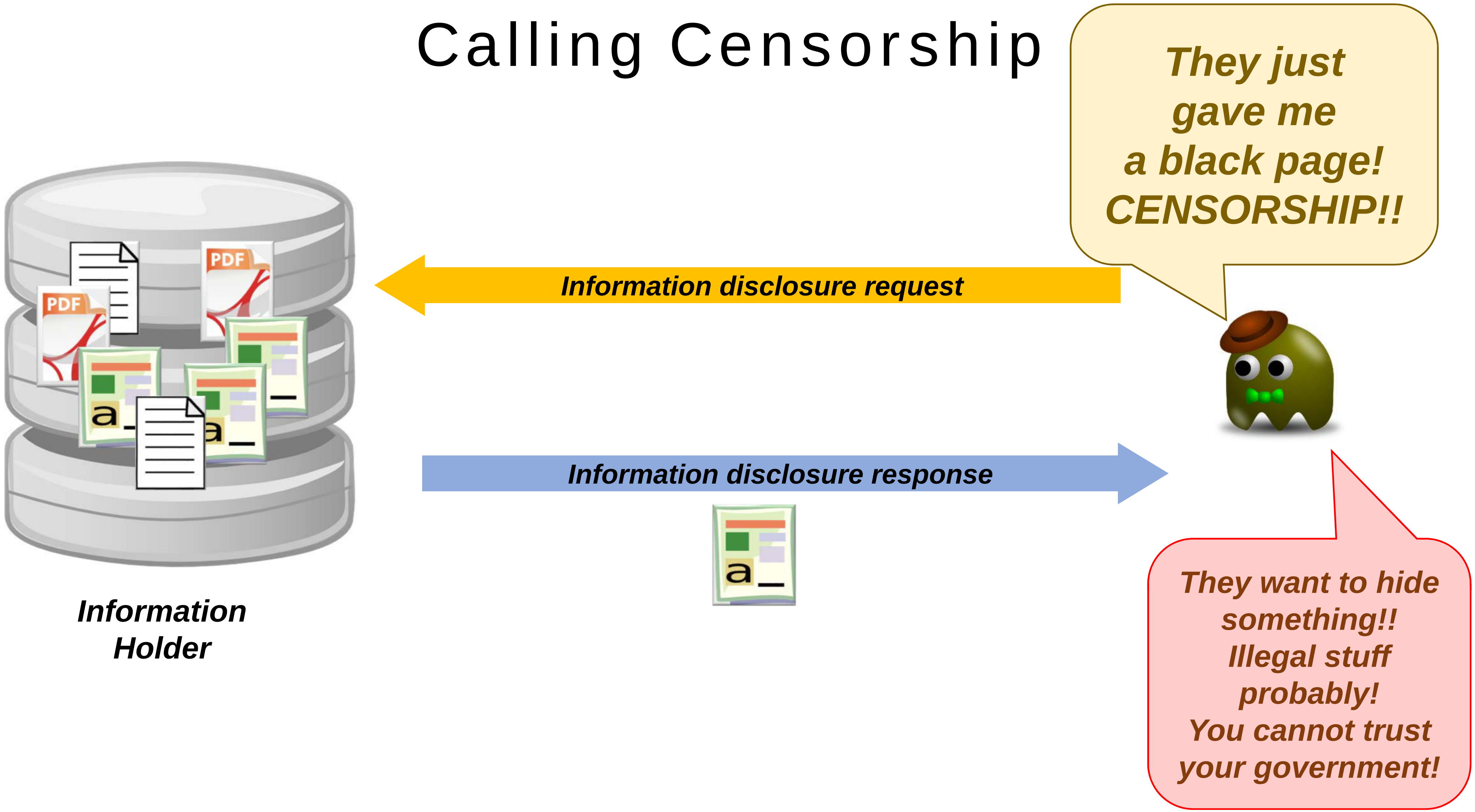
University troubles

- Students demanding access to exam documents before exams are written
- …or to the standard solutions documents
- …or to the grades database files
- …or to the emails of the professor (that may contain the exams)
- Swedish **principle of public access to official documents**:
  - Every human may demand all documents created by Swedish government officials
  - …such as university employees
  - …free of charge, without restriction or fee
  - …unless explicit secrecy is declared
  - …for arbitrary purposes (no “misuse” concept in the law)

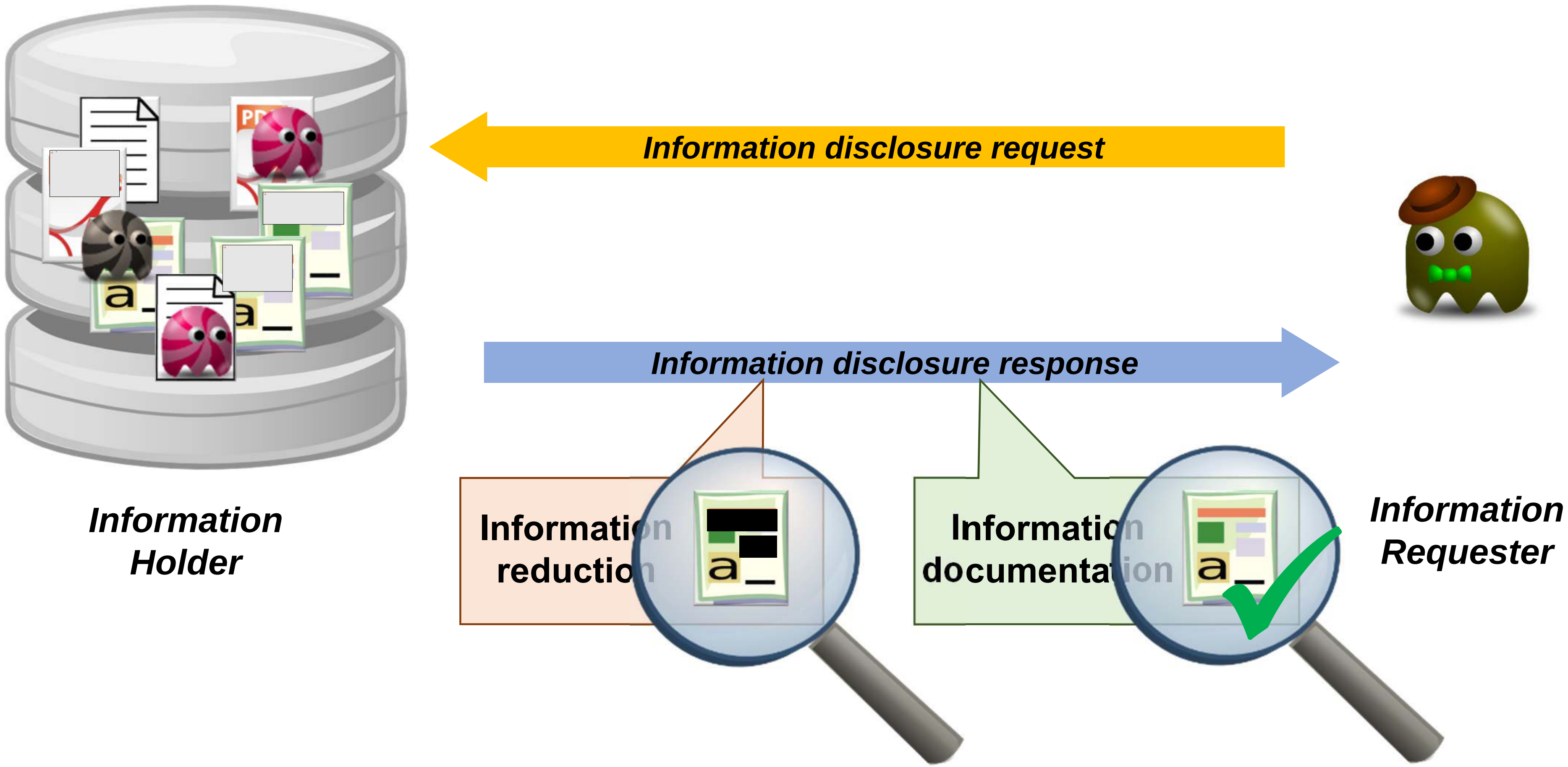


➔ Employees prefer phone/zoom to email (“chilling effect”)

Calling Censorship



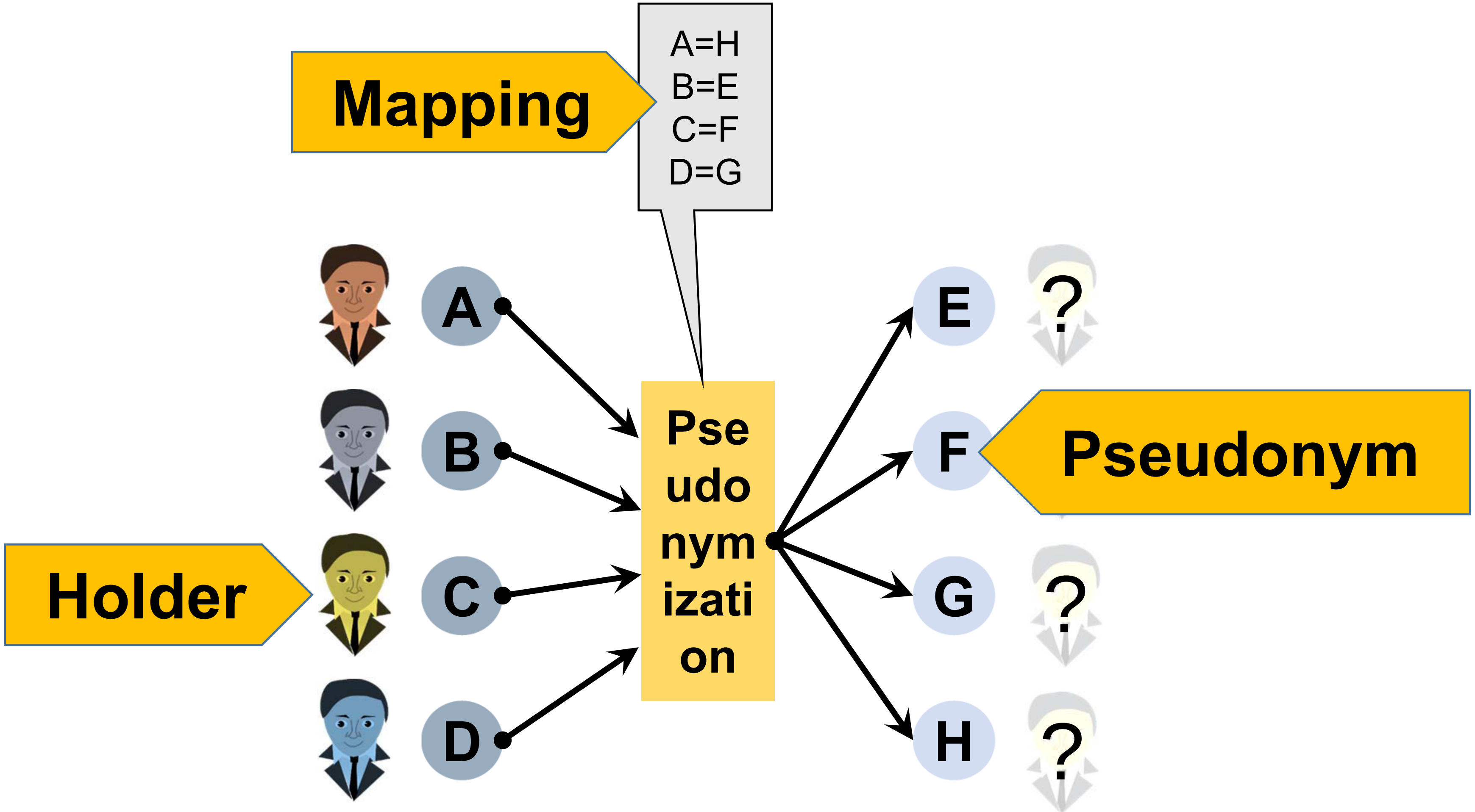
Information Disclosure Scenario



Information reduction

Technique #1: Pseudonymization

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

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

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

Pseudonymization Techniques

- **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!
- **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
- **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 non-identifiers!
(so-called 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!

ENISA Reports 2019-2021

- **Terminology**
- **Scenarios**
- **Adversary Model**
- **Techniques**
- **Anonymity vs. Utility**
- **Application Scenarios**
  - IP Address Pseudonymization
  - E-Mail Address Pseudonymization
  - Pseudonymization in Practice
- **Use Case: Medical Data Analytics**
- **Data Custodian Models**

Information reduction

Technique #2: k-anonymity

Types of Identifiers

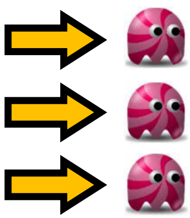
Explicit Identifiers

- Uniquely attributable
  - name
  - phone number
  - address

Alice Kausson

+46 54 7001000

Karlstadsgatan 1



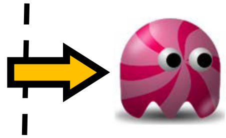
Quasi-Identifiers

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

01.07.80

female

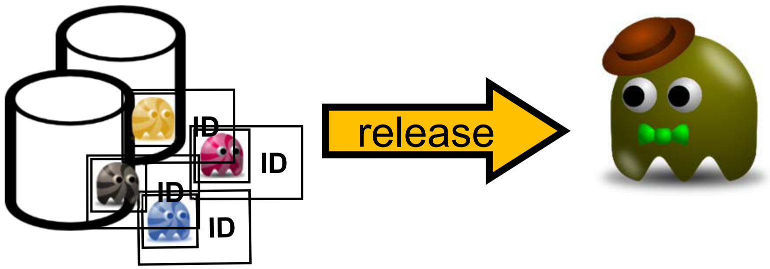
SE 65188



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

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.

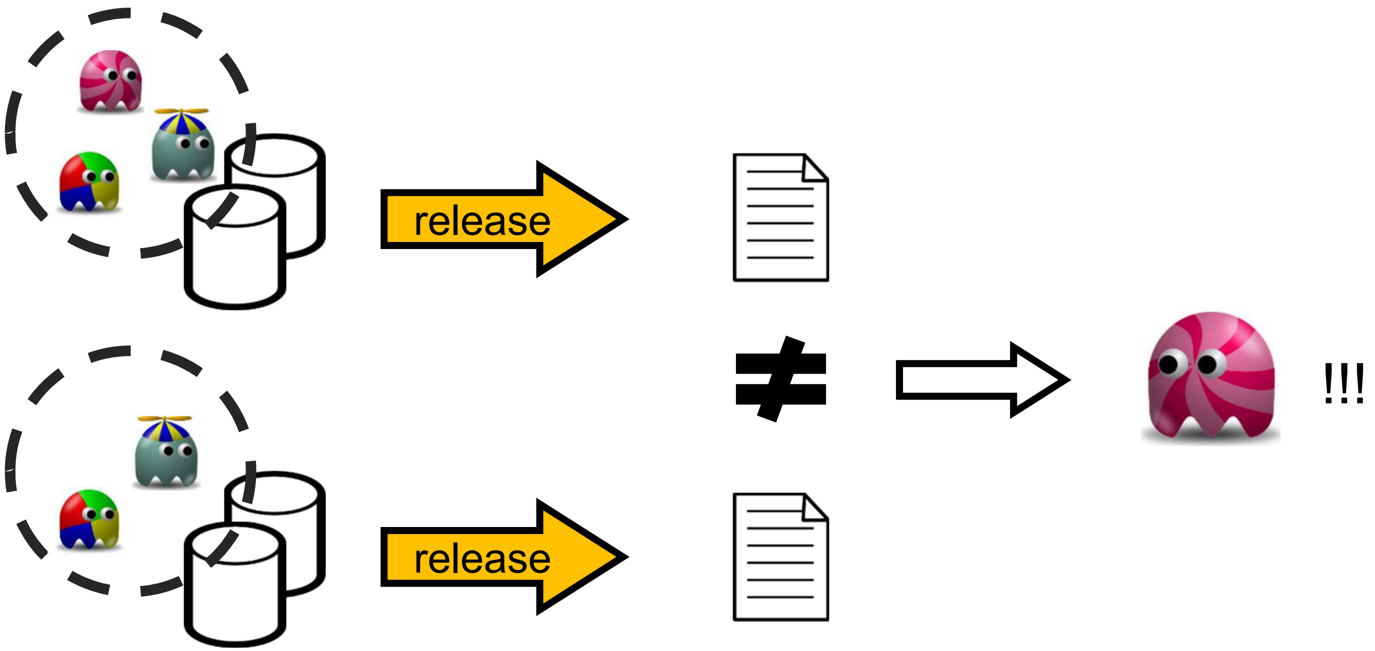
Slides adapted from: L. Martucci

Information reduction

Technique #3: Differential Privacy

Releasing Personal Data

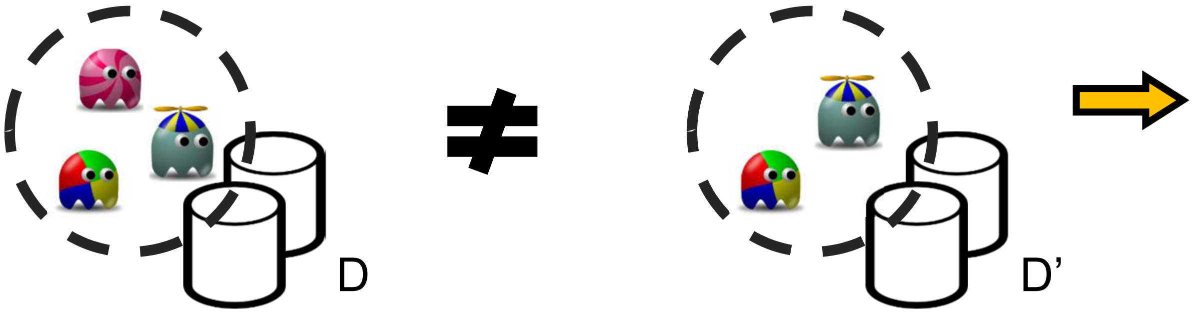
- Looking into two data releases: (from a statistical database)



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



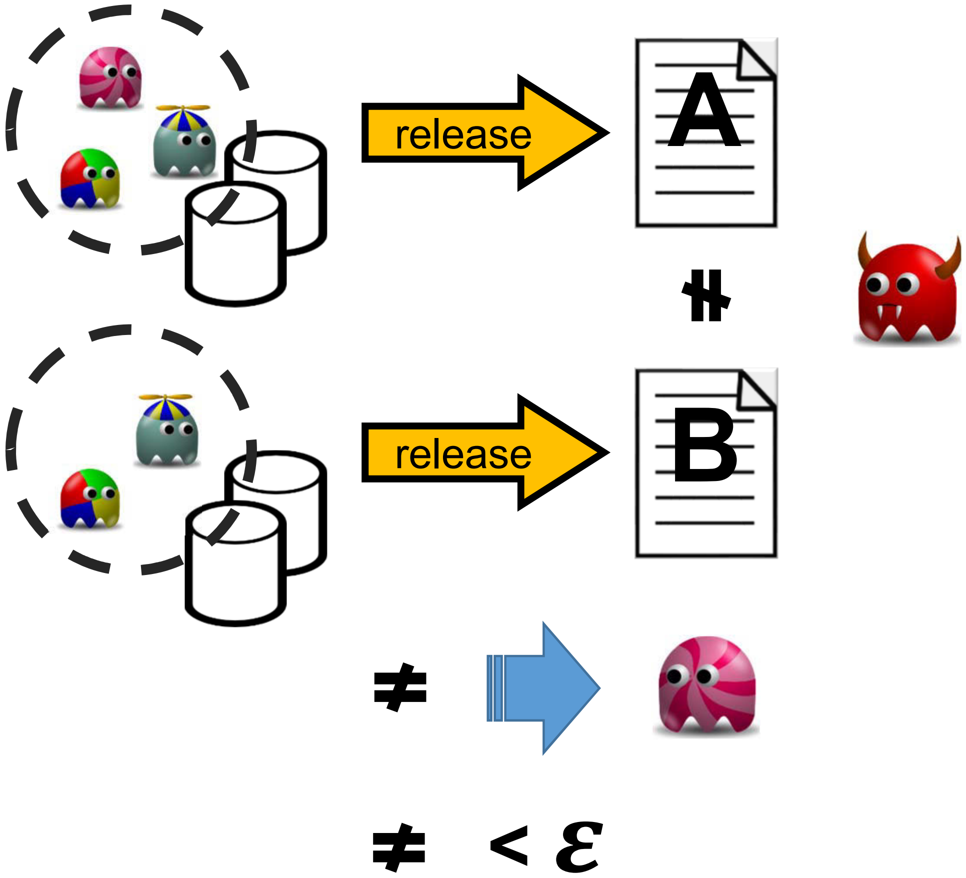
what does one learn?
at most $\exp(\varepsilon)$

- 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

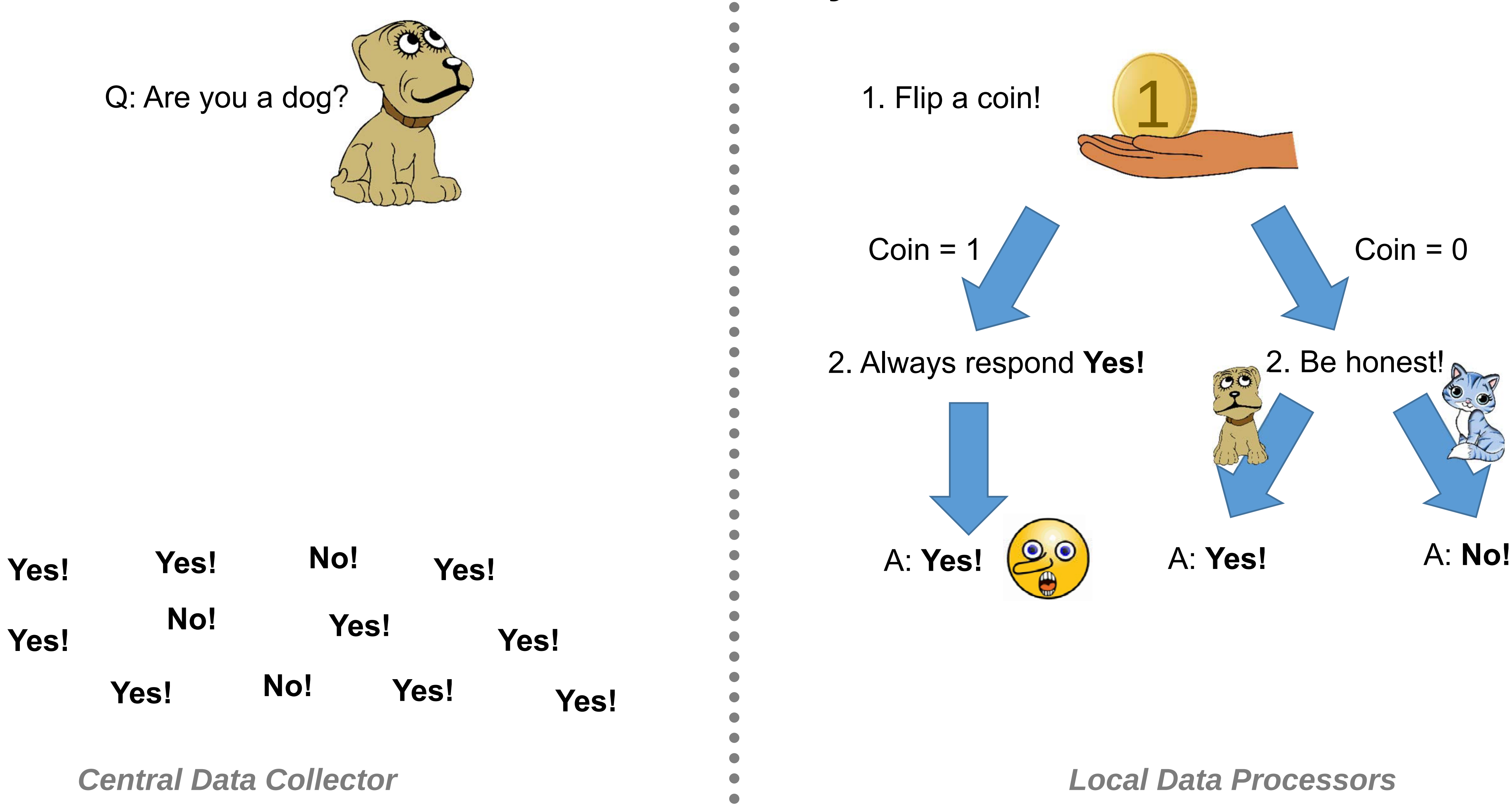
mind the background information!

Slides adapted from: L. Martucci

Differential Privacy in Practice

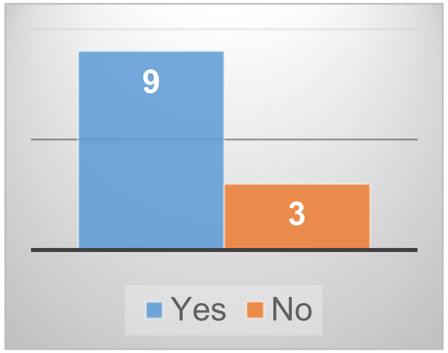
"On the Internet, nobody knows you're a dog."

Differential Privacy in Practice

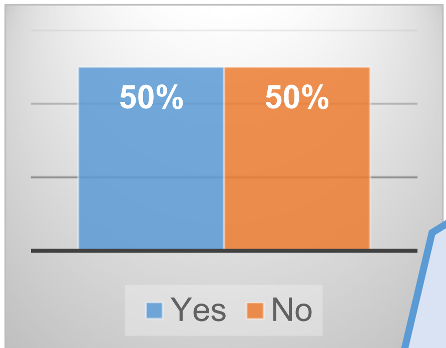


Differential Privacy in Practice

Frequency Analysis

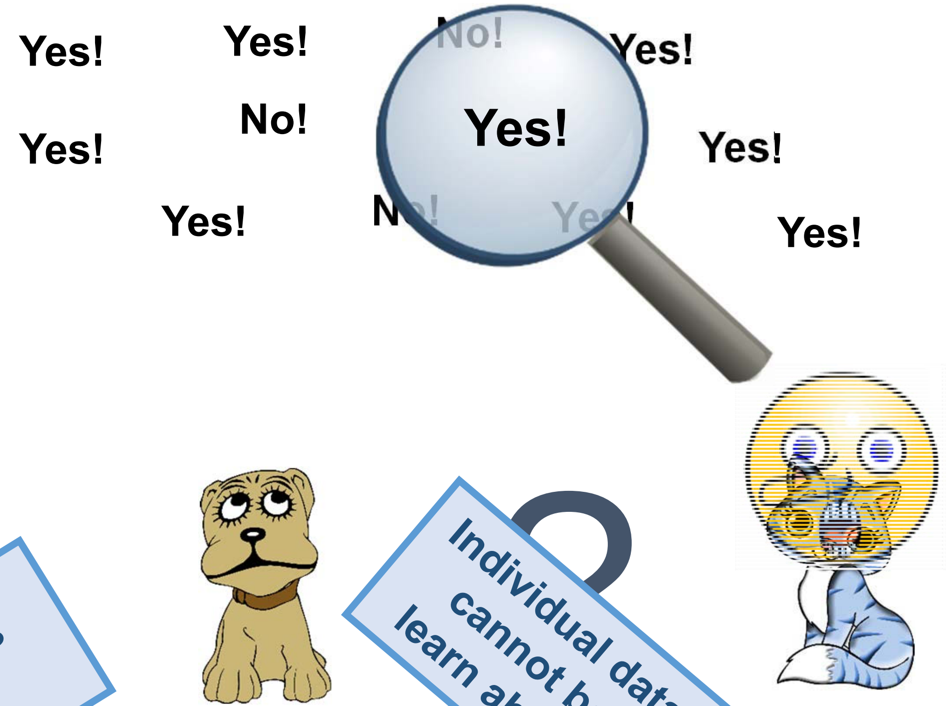


Subtract n/2 from „Yes“, as they were lies…



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…“

Differential Privacy in Practice

- **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“)

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:

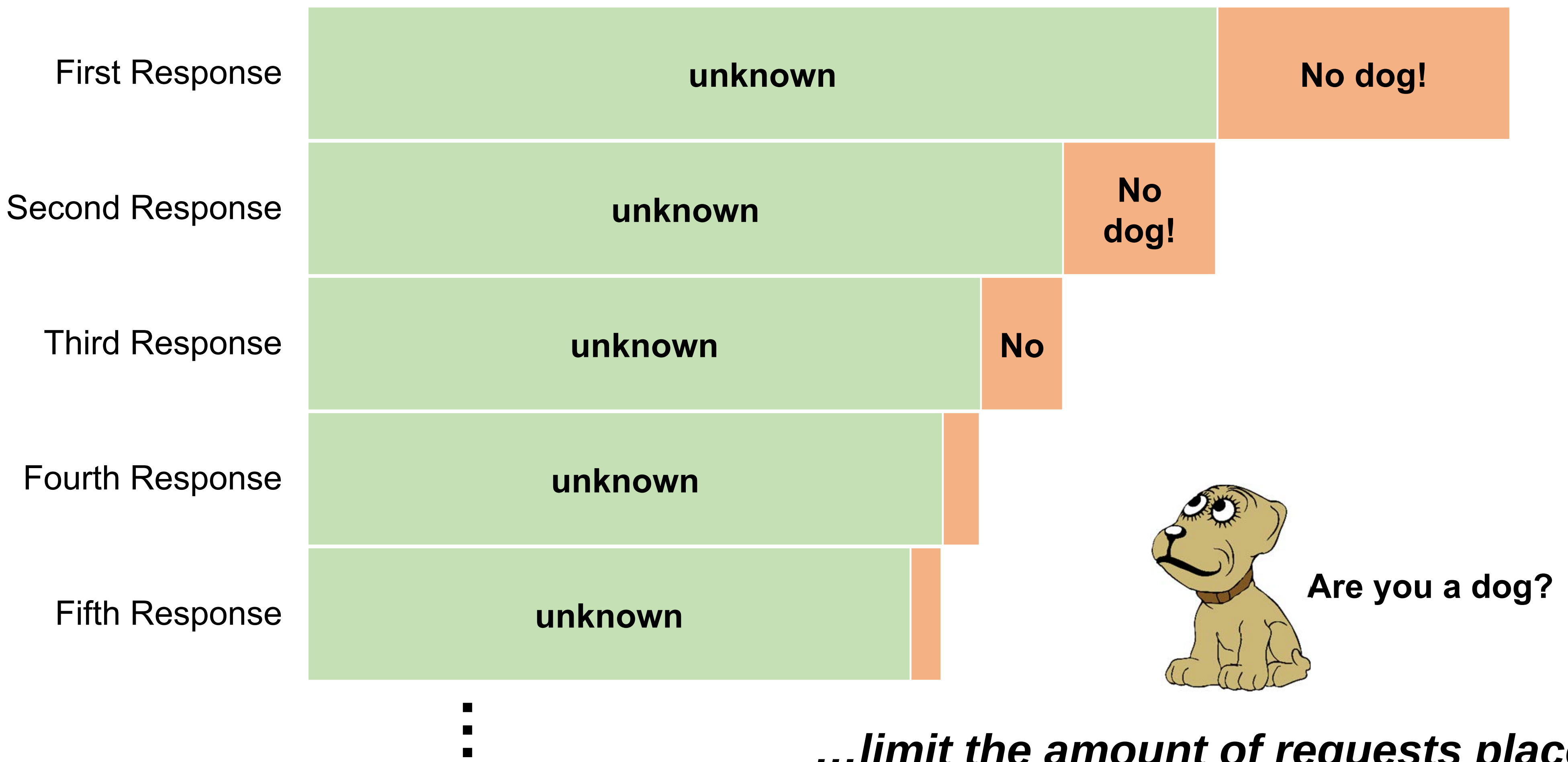
https://security.googleblog.com/2014/10/learning-statistics-with-privacy-aided.html

https://github.com/google/rappor

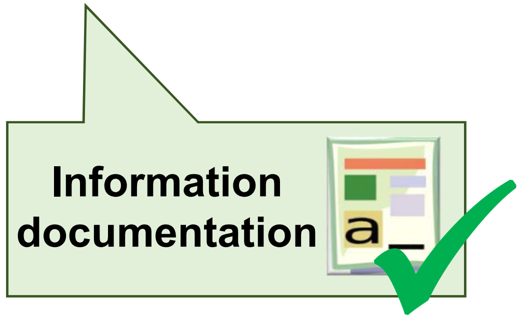
Differential Privacy in Practice

- **Problem**:
  - If you repeat asking the same question to the same person, you learn the correct answer with increasing probability…

Differential Privacy in Practice



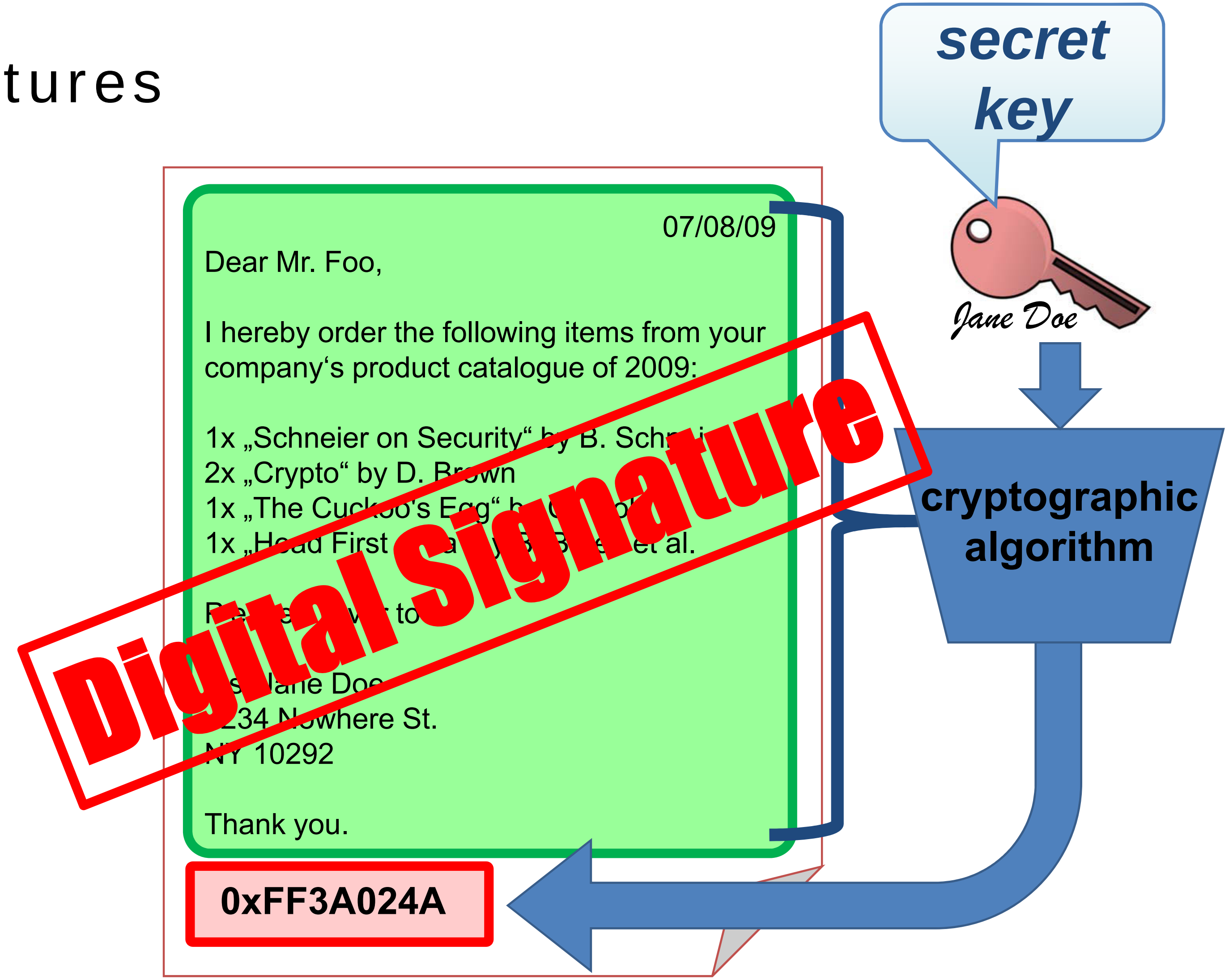
…limit the amount of requests placed!



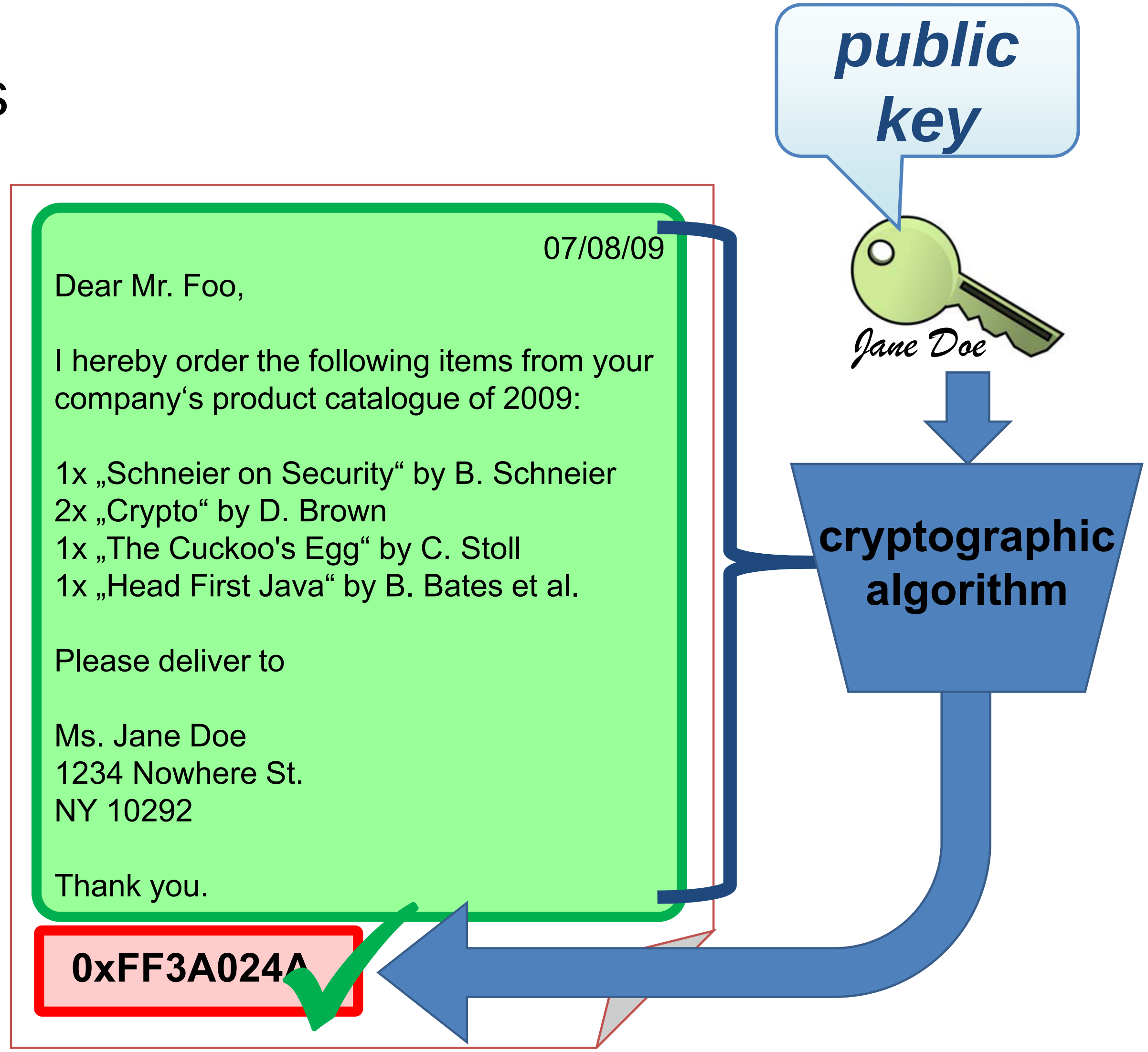
Technique #4: Digital Signatures

Digital Signatures

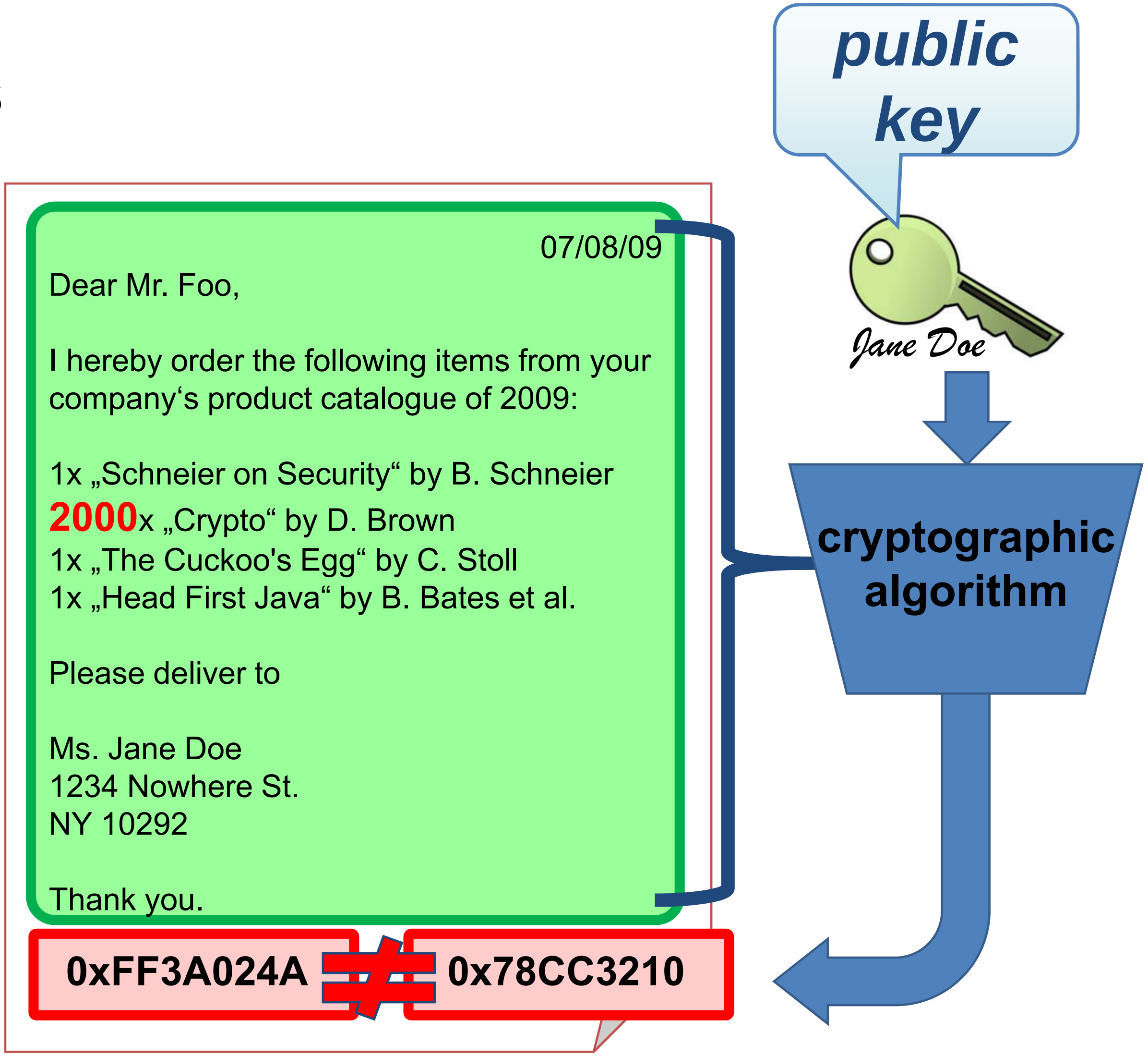
Digital Signatures



Digital Signatures



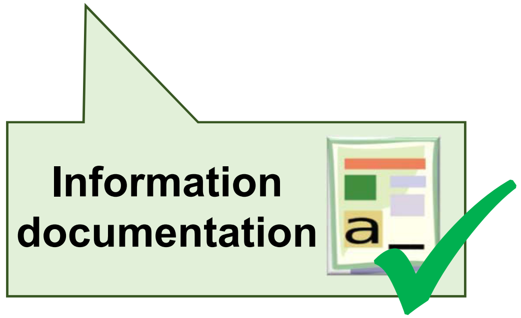
Digital Signatures



Digital Signatures

- A ***valid*** digital signature implies:
  - The corresponding piece of ***data***…
    - (which must be known exactly from the message structure!)
  - …was ***not modified***…
    - (i.e. not a single character was added, deleted, exchanged)
  - …since the ***signing entity*** (or ***signer***)…
    - (e.g. the sender of a message, the contractor of a contract)
  - …had calculated the cryptographic ***signature value***.

➔ If signature ***verification fails***,
at least one of these statements must be wrong!



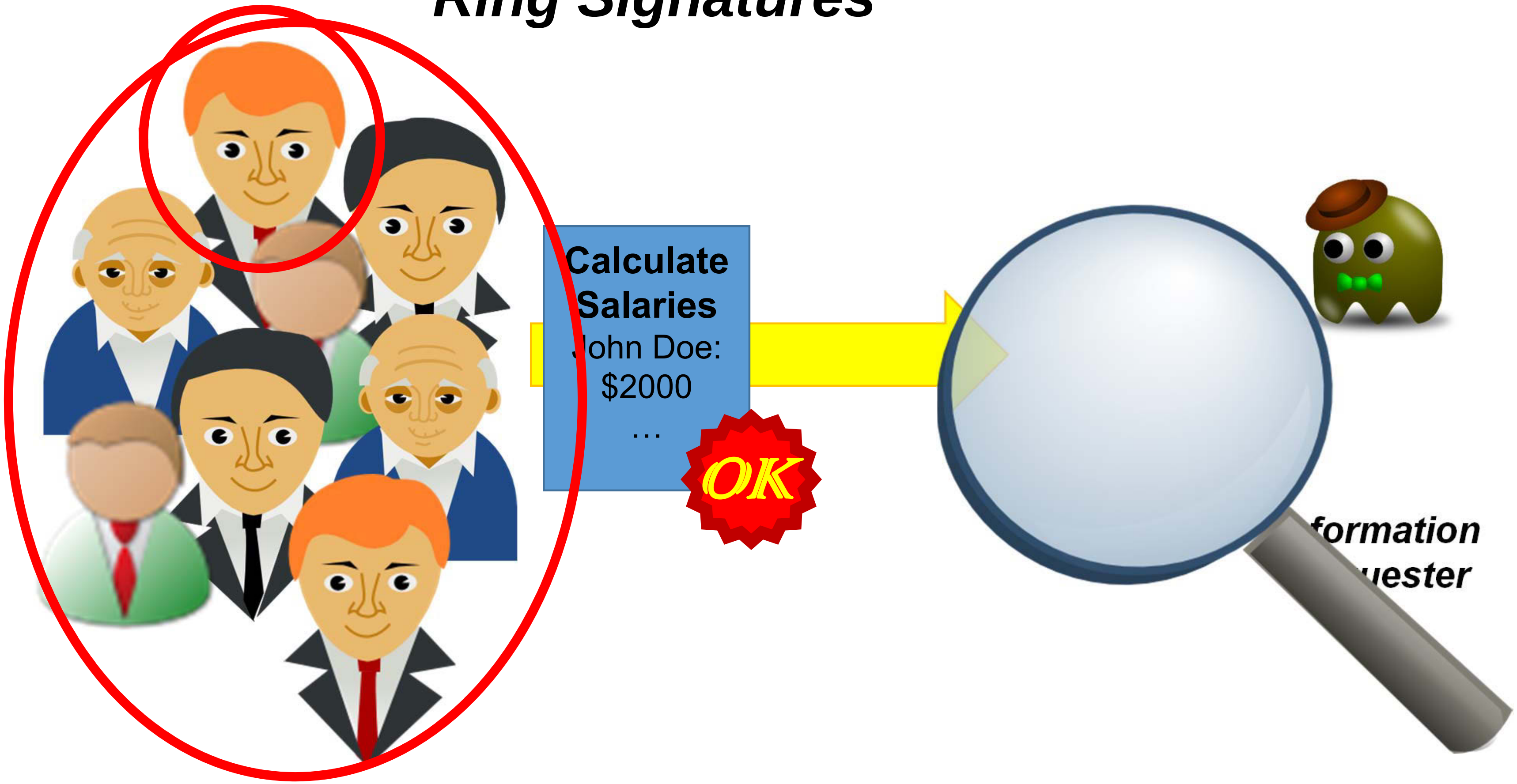
Techniques #5-#9: Advanced Digital Signatures

Advanced Digital Signatures

*[R. L. Rivest, A. Shamir, Y. Tauman: “How to leak a secret”, ASIACRYPT 2001.]

Advanced Digital Signatures

Ring Signatures

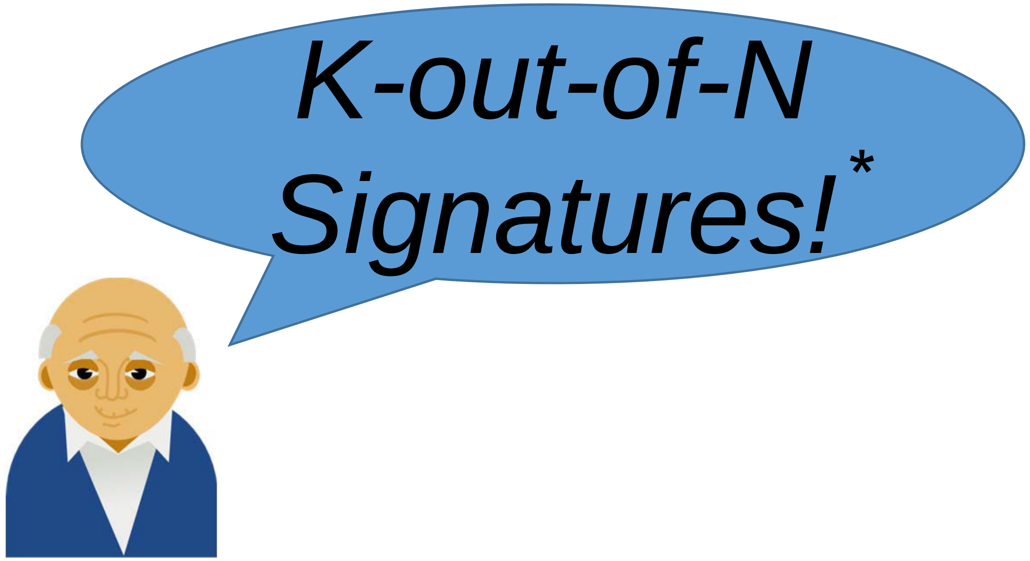


Advanced Digital Signatures

Ring Signatures

- Every group member can sign
- Everybody can verify
- Nobody can determine which group member did sign

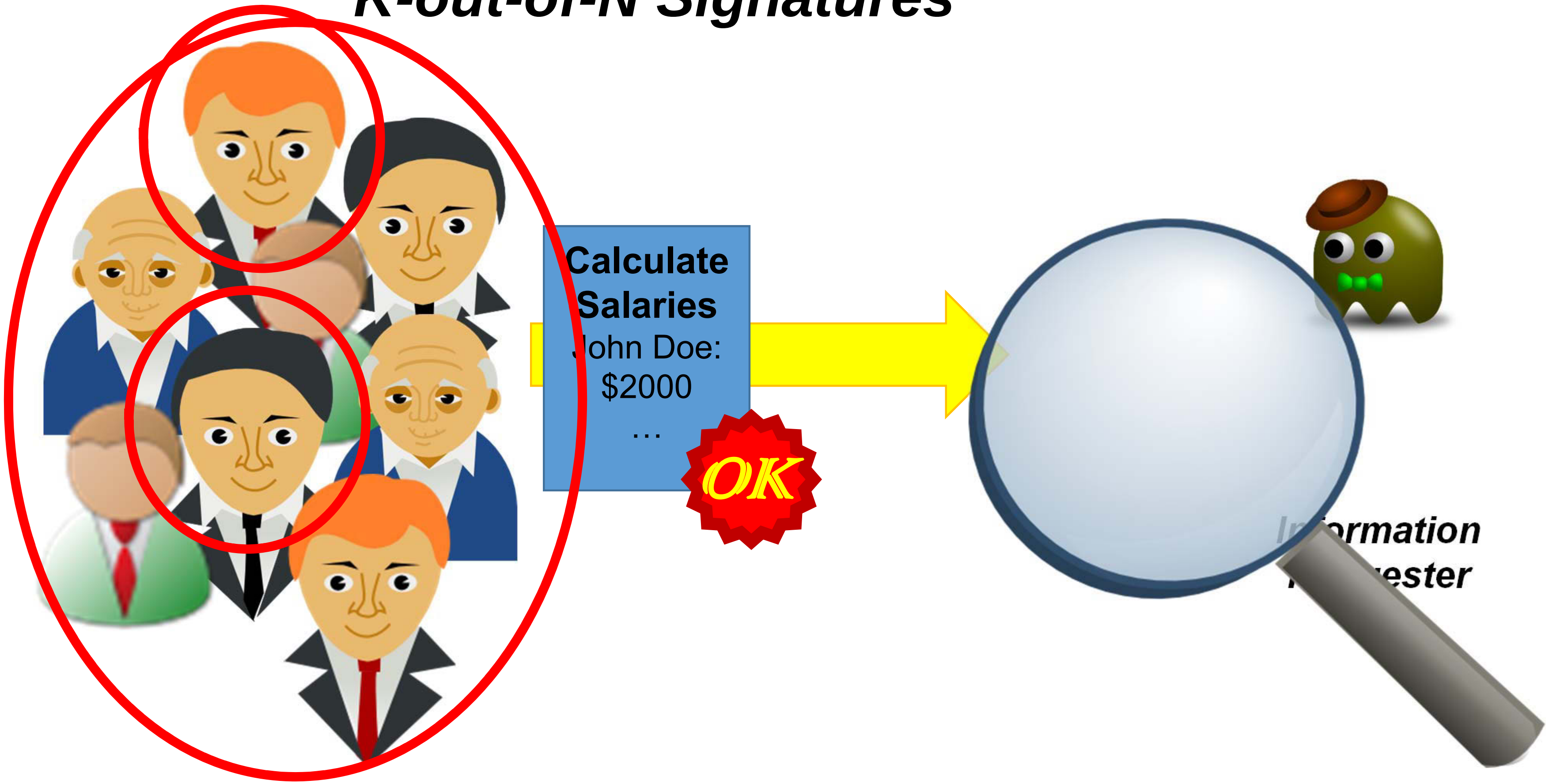
Advanced Digital Signatures



*[Boneh, D., Lynn, B., & Shacham, H.: Short signatures from the Weil pairing. *ASIACRYPT 2001*.]

Advanced Digital Signatures

K-out-of-N Signatures



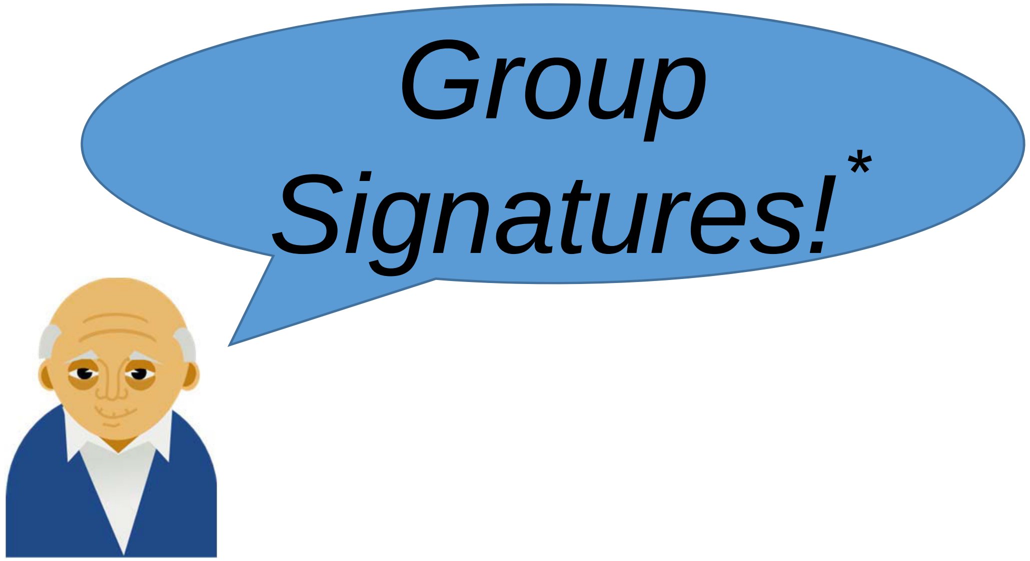
Advanced Digital Signatures

K-out-of-N Signatures (or Threshold Signatures)

- No single group member can sign
- Every subgroup of at least K group members can sign

  (e.g. four-eyes principle)
- Everybody can verify
- Nobody can determine which group member(s) did sign

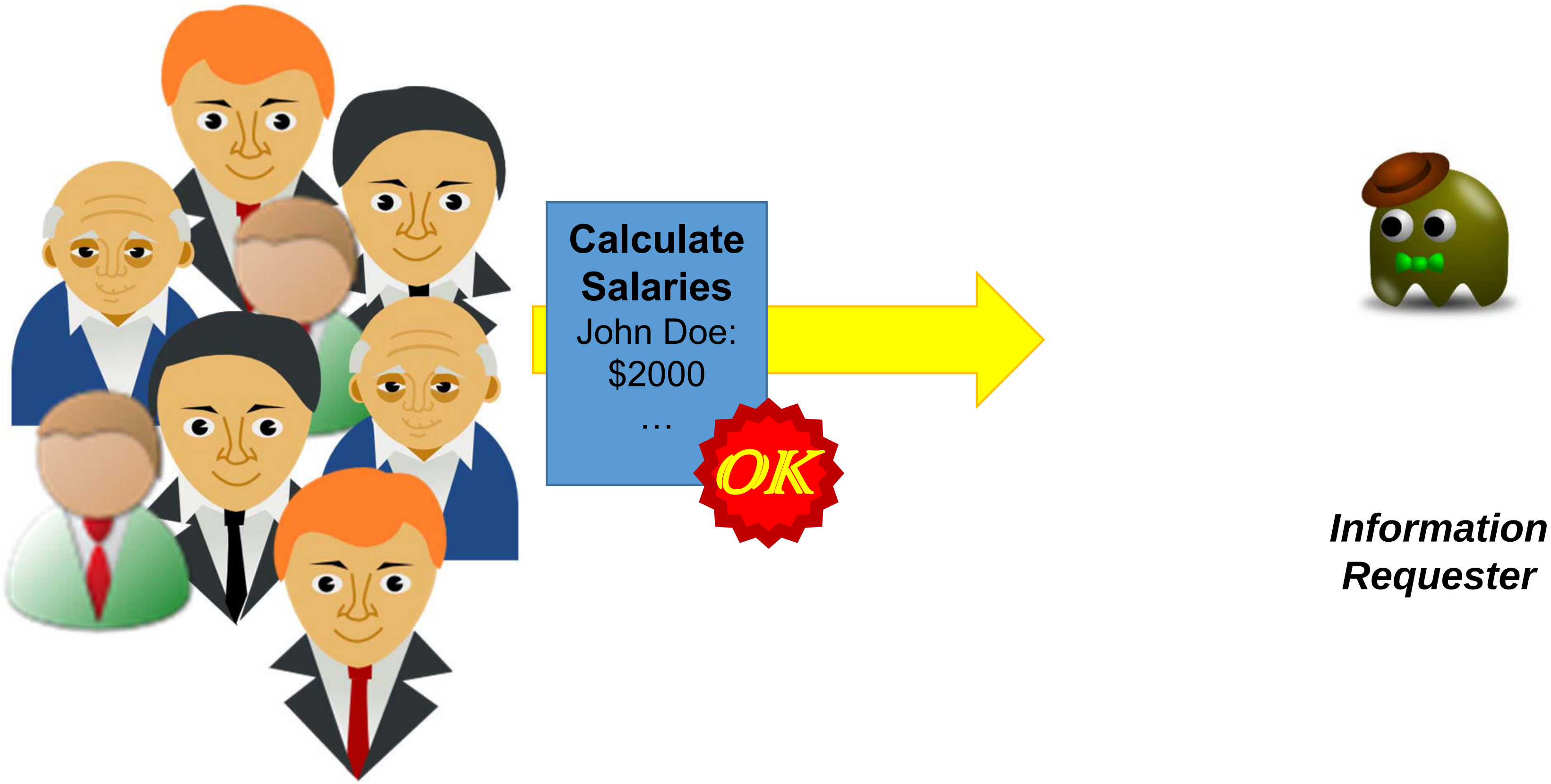
Advanced Digital Signatures



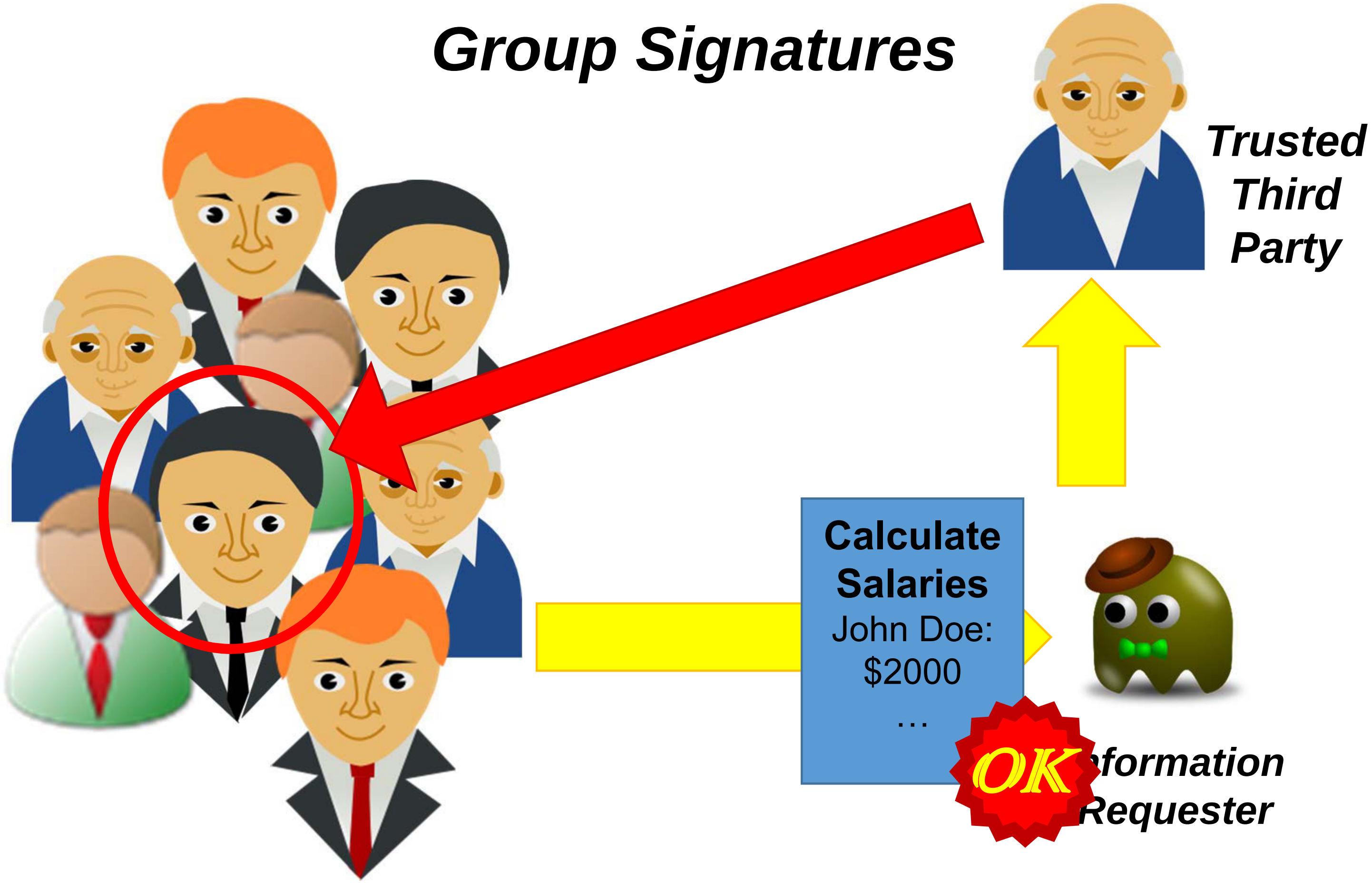
*[D. Chaum, E. van Heyst: “Group signatures”, EUROCRYPT 1991]

Advanced Digital Signatures

Group Signatures



Advanced Digital Signatures

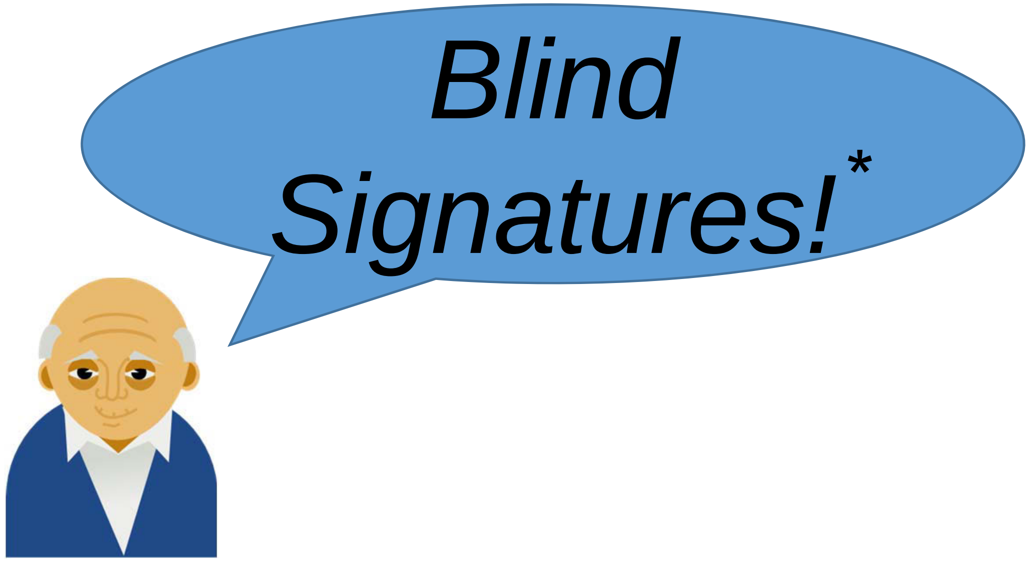


Advanced Digital Signatures

Group Signatures

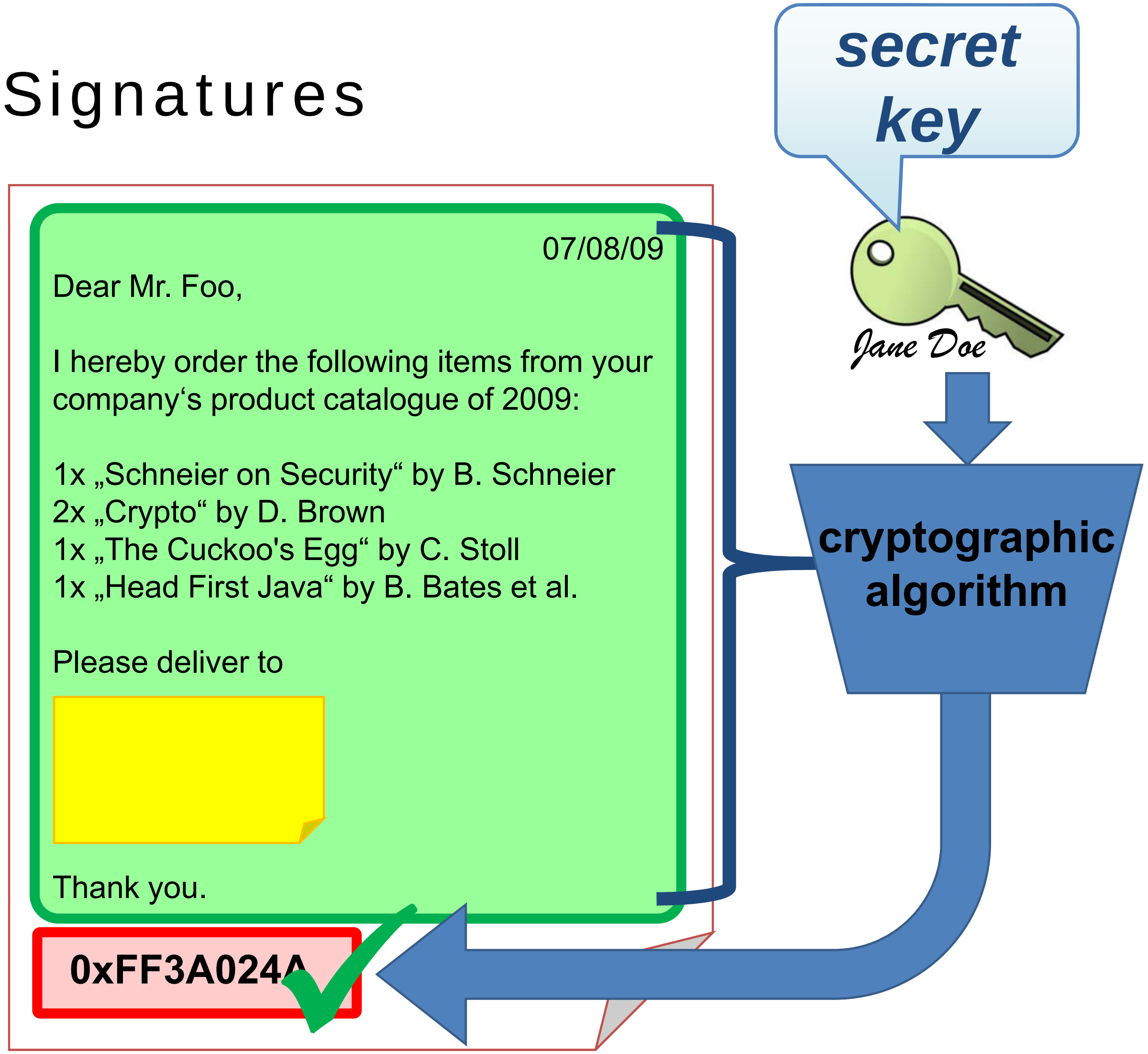
- Every group member can sign
- Everybody can verify
- Only a dedicated trusted third party can determine
  which group member did sign

Advanced Digital Signatures



*[Chaum, David: "Blind signatures for untraceable payments". Advances in Cryptology, 1983]

Advanced Digital Signatures



Advanced Digital Signatures

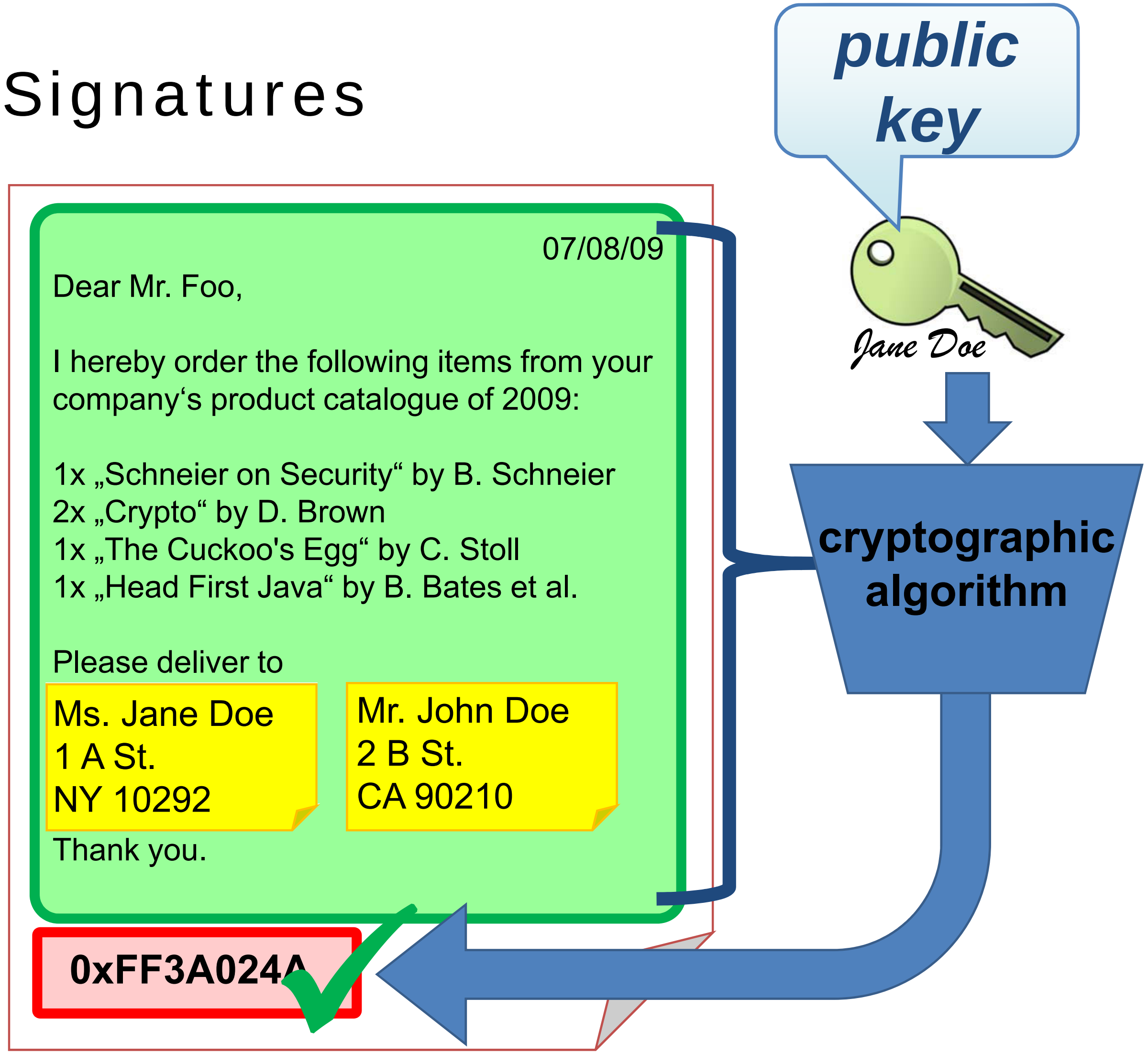
Blind Signatures

- Signer can sign the whole document
- Signer cannot see parts of the document
- Everybody can verify, as long as they know the whole document
- Applications e.g. in election systems, digital cash

Advanced Digital Signatures

*[Ateniese, G., Chou, D.H., de Medeiros, B., Tsudik, G.: Sanitizable Signatures. ESORICS 2005.]

Advanced Digital Signatures



Advanced Digital Signatures

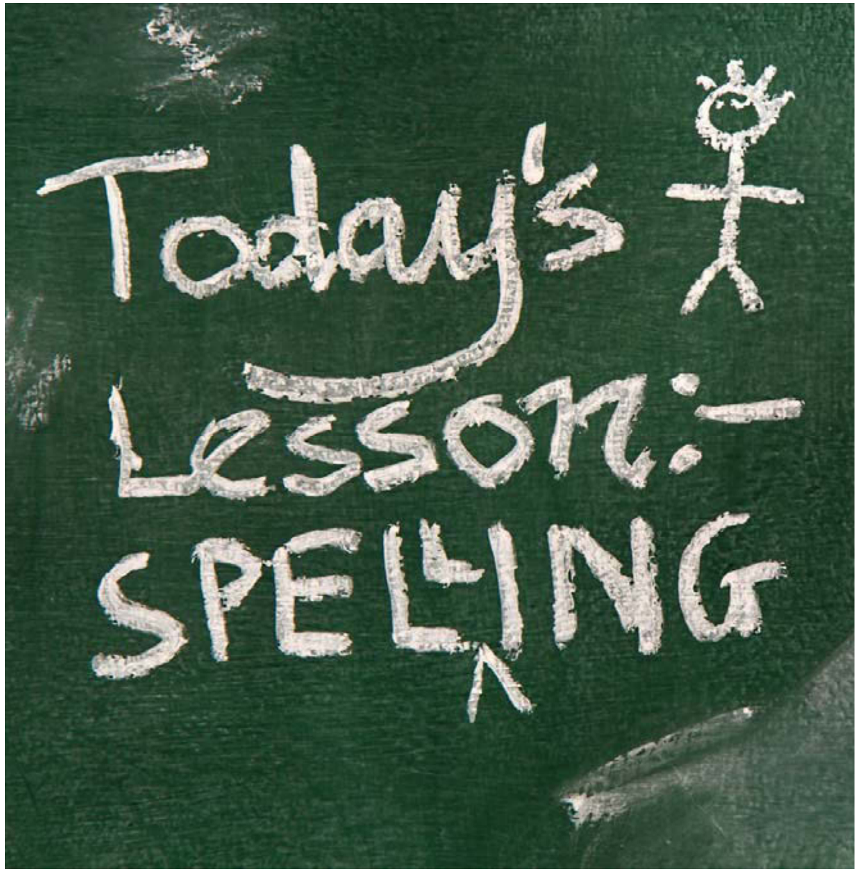
Sanitizable Signatures

- Signer can sign two (or more) alternatives for part of document
- Signer explicitly names all allowed alternatives
- Subsequent processors can replace one alternative with another

  ➔ Signature stays correct
- If other parts of the document are changed ➔ Signature invalid
- Everybody can verify
- Applications in recognizable censorship

Summary

- Techniques for information reduction
  - Pseudonymization
  - K-Anonymity
  - Differential Privacy
- Techniques for information documentation
  - Digital Signatures
  - Advanced Digital Signatures
- Apply techniques whenever reasonable!
- Mind the hidden information!
- Mind the background knowledge!



CC-BY-4.0

Thank you!

Danke!

Tack!

Datenreduktion vor Herausgabe von Informationen
- der Werkzeugkasten der Kryptographen

Prof. Dr.-Ing. Meiko Jensen