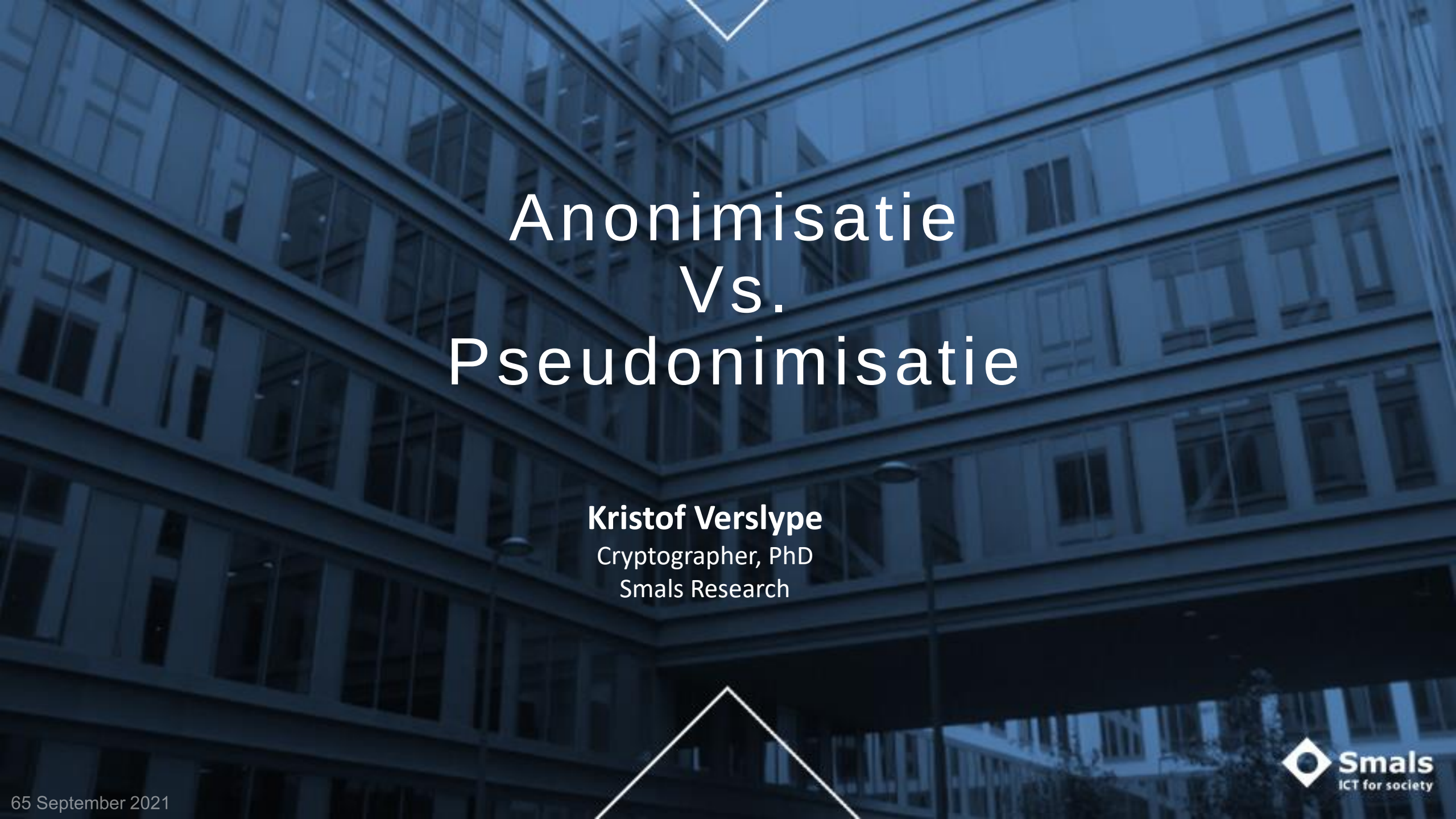


Anonimisatie Vs. Pseudonimisatie

Kristof Verslype
Cryptographer, PhD
Smals Research

Smals
ICT for society

Uitnodiging

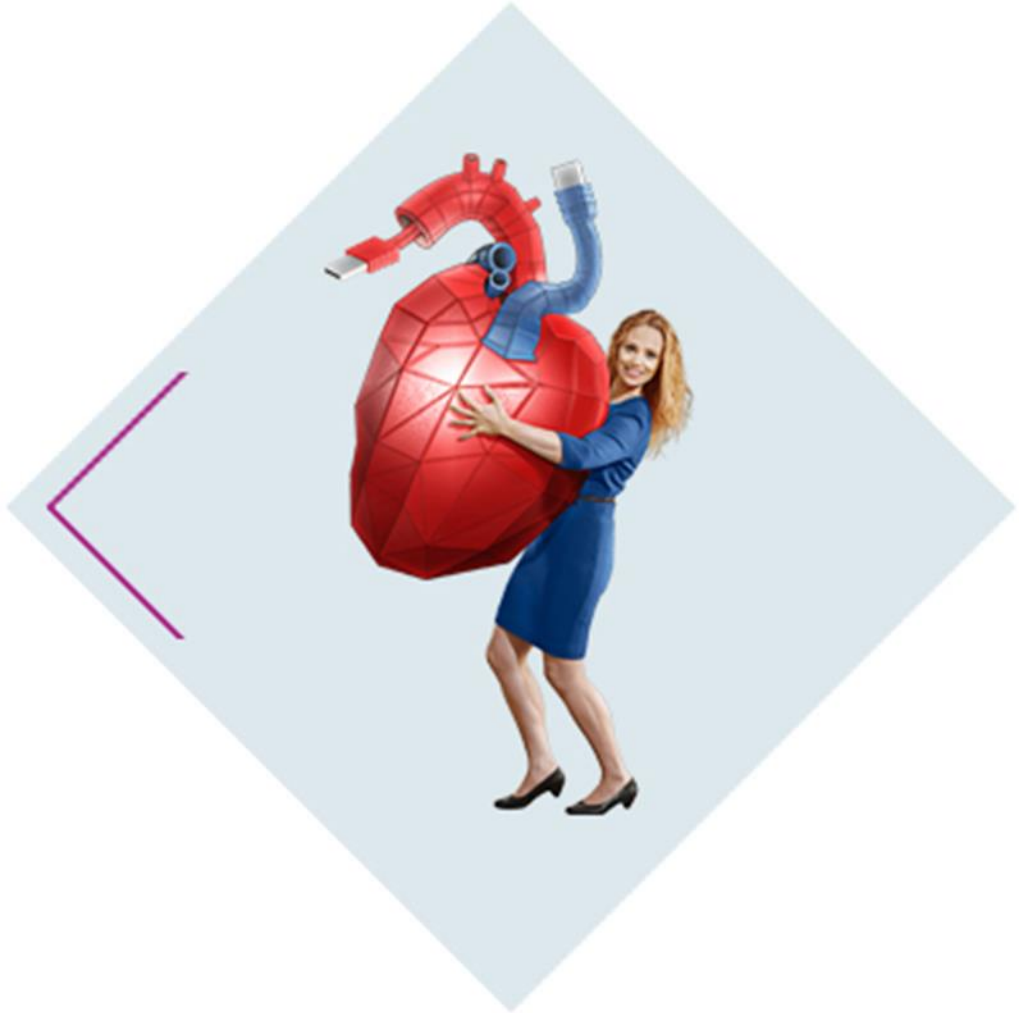
Feedback welkom mocht u het met bepaalde interpretaties oneens zijn

Doel presentatie is ook triggeren van discussie

Smals
ICT for society

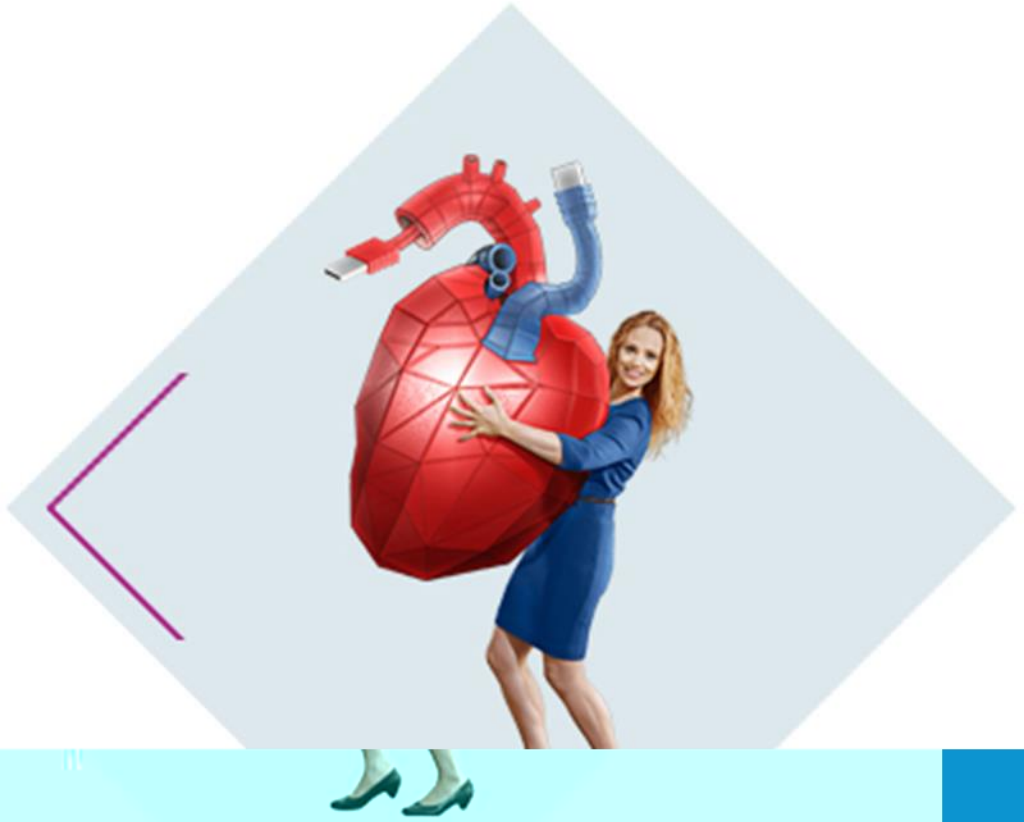
Agenda

- GDPR terminologie
- Quasi-identifiers
- Vervagingstechnieken
- Hoogdimensionale data
- Hashing & encryptie
- Afronding



Agenda

- GDPR terminologie
- Quasi-identifiers
- Vervagingstechnieken
- Hoogdimensionale data
- Hashing & encryptie
- Afronding



GDPR

NEW

Anonieme gegevens

Onlinkbaar
aan natuurlijk persoon

GDPR niet van toepassing

Gepseudonimiseerde persoonsgegevens

Linkbaar met aanvullende gegevens
aan natuurlijk persoon

GDPR van toepassing # bepalingen versoepeld

Geïdentificeerde persoonsgegevens

Linkbaar zonder aanvullende gegevens
aan natuurlijk persoon

GDPR ten volle van toepassing

← (Identificeerbare) Persoonsgegevens →

Tot welke categorie behoren gegevens?

Er zijn grijze zones → interpretatie

***Persoonsgegevens*.** Alle informatie over een geïdentificeerde of identificeerbare natuurlijke persoon („de betrokkene”); als identificeerbaar wordt beschouwd een natuurlijke persoon die **direct of indirect kan worden geïdentificeerd**, met name aan de hand van een identificator zoals een naam, een identificatienummer, locatiegegevens, een online identificator of van **een of meer elementen** die kenmerkend zijn voor de fysieke, fysiologische, genetische, psychische, economische, culturele of sociale identiteit van die natuurlijke persoon.

GDPR – Definities

***Anonieme gegevens*.** Gegevens die geen betrekking hebben op een geïdentificeerde of identificeerbare natuurlijke persoon of op persoonsgegevens die zodanig anoniem zijn gemaakt dat de betrokkene niet of niet meer identificeerbaar is.

Pseudonimisering. Het verwerken van persoonsgegevens op zodanige wijze dat de persoonsgegevens niet meer aan een specifieke betrokkene kunnen worden gekoppeld zonder dat er **aanvullende gegevens** worden gebruikt, mits

1. deze aanvullende gegevens **apart worden bewaard** en
2. **technische en organisatorische maatregelen worden genomen om ervoor te zorgen dat de persoonsgegevens niet aan een geïdentificeerde of identificeerbare natuurlijke persoon worden gekoppeld**.

Gepseudonimiseerde persoonsgegevens die door het gebruik van aanvullende gegevens aan een natuurlijke persoon *kunnen* worden gekoppeld, **moeten als gegevens over een identificeerbare natuurlijke persoon worden beschouwd.**

Om te bepalen of een natuurlijke persoon identificeerbaar is, **moet rekening worden gehouden met alle middelen waarvan redelijkerwijs valt te verwachten dat zij worden gebruikt door de verwerkingsverantwoordelijke of door een andere persoon om de natuurlijke persoon direct of indirect te identificeren.**

Opgelet

Indien resultaat nog steeds m.b.v. publieke informatie te linken aan bekende belg, dan geen (correcte) pseudonimisering

Gegevens zoals id uit records weglaten niet per se voldoende voor correcte pseudonimisering, laat staan anonimisering

Hoe weten we of we met *alle middelen* rekening gehouden hebben? Geregeld foute inschattingen!

Anonieme gegevens & statistische gegevens

***Anonieme gegevens*.** Gegevens die geen betrekking hebben op een geïdentificeerde of identificeerbare natuurlijke persoon of op persoonsgegevens die zodanig anoniem zijn gemaakt dat de betrokkene niet of niet meer identificeerbaar is

Statistische gegevens vallen doorgaans onder “anonieme” gegevens
Maar ook daar opletten met extreme gevallen

Vb. 98% van de bevolking in Zwijndrecht heeft meer dan 100 µg PFOS in het bloed (*).

Over elke burger in Zwijndrecht weten we met quasi zekerheid dat ze te hoge PFOS waarden in hun bloed hebben.

⇒ **Medische geïdentificeerde persoonsgegevens** (met een klein beetje onzekerheid/ruis op)

Open vraag

Wanneer kunnen we spreken van geanonimiseerde gegevens?
Bij 90%? 80%? 50%? 20%? 10%? 5%? 1%?

Concept uit cryptografie

Plausible deniability (plausibele ontkenning / déni plausible)

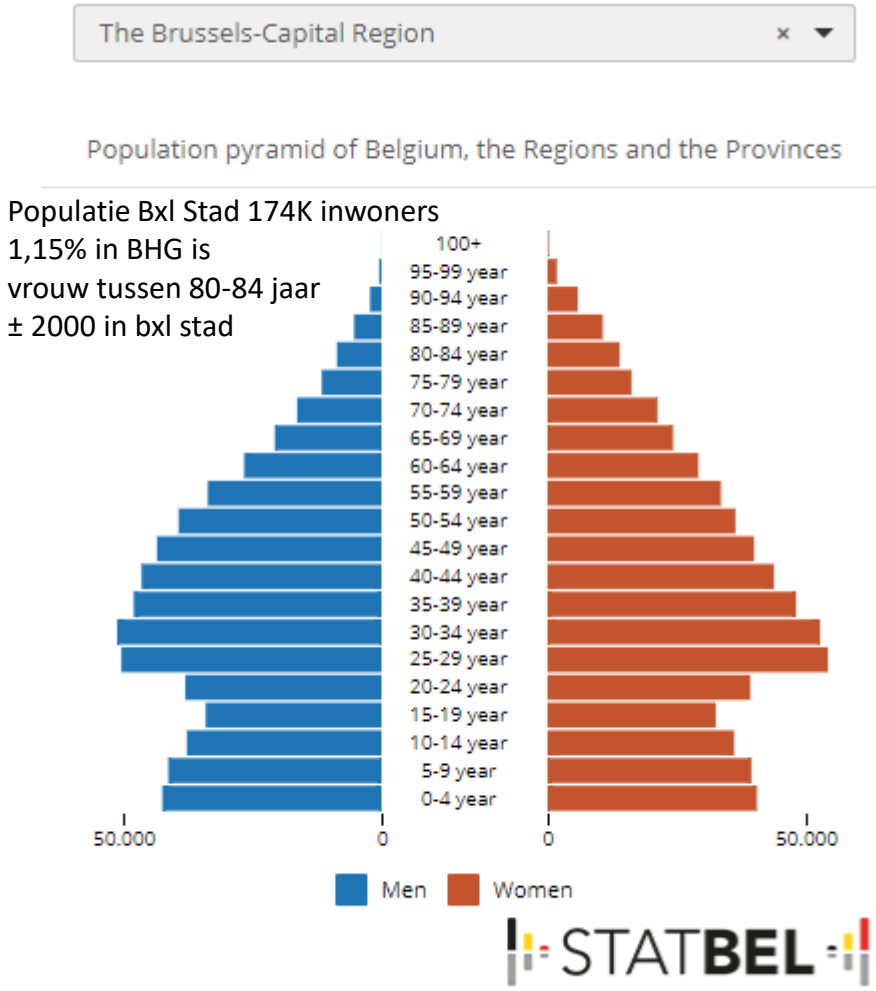
(*) fictieve waarden

Smals
ICT for society

Agenda

- GDPR terminologie
- Quasi-identifiers
- Vervagingstechnieken
- Hoogdimensionale data
- Hashing & encryptie
- Afronding

Verwijderen identifiers



	RRN	DoB	Sex	ZIP	Disease
be	84052010955	20/05/1984	M	3211	Antisocial personality disorder
als	76011320675	13/01/1976	M	9300	Schizofrenie
Calabria	37091100247	11/09/1937	F	1000	Multiple sclerosis
	50111221385	12/11/1950	M	4000	Megalomanie

	RRN	DoB	Sex	ZIP	Disease
[redacted]	[redacted]	20/05/1984	M	3211	Antisocial personality disorder
[redacted]	[redacted]	13/01/1976	M	9300	Schizofrenie
[redacted]	[redacted]	11/09/1937	F	1000	Multiple sclerosis
[redacted]	[redacted]	12/11/1950	M	4000	Megalomanie

Dit zijn geen anonieme gegevens!

- Bevolking Binkom (ZIP 3211) bedraagt 1459
- Geboortedatum, geslacht en woonplaats Koningin Paola publiek gekend

Smals
ICT for society

Coderen identifiers

Pseudonimisering. Het verwerken van persoonsgegevens op zodanige wijze dat de persoonsgegevens niet meer aan een specifieke betrokkene kunnen worden gekoppeld zonder dat er **aanvullende gegevens** worden gebruikt, mits

1. deze aanvullende gegevens **apart worden bewaard** en
2. **technische en organisatorische maatregelen worden genomen om ervoor te zorgen dat de persoonsgegevens niet aan een geïdentificeerde of identificeerbare natuurlijke persoon worden gekoppeld**."

Naam	RRN	DoB	Sex	ZIP	
Kristof Verslype	84052010955	20/05/1984	M	3211	
Jan Vandesmals	76011320675	13/01/1976	M	9300	Schizofrenie
Paola Ruffo di Calabria	37091100247	11/09/1937	F	1000	Multiple sclerosis
Dick Tatuur	50111221385	12/11/1950	M	4000	Megalomanie

RRN	Code
84052010955	X38LS45
76011320675	X38DI56
37091100247	X38XD12
50111221385	X38MP68

Naam	RRN	DoB	Sex	ZIP	Disease
X38LS45		20/05/1984	M	3211	Antisocial personality disorder
X38DI56		13/01/1976	M	9300	Schizofrenie
X38XD12		11/09/1937	F	1000	Multiple sclerosis
X38MP68		12/11/1950	M	4000	Megalomanie

Dit is geen correcte pseudonimisering!

Identificatie ook mogelijk zonder apart bewaarde aanvullende gegevens

Smals
ICT for society

Quasi-identifiers

Quasi-identifiers are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier. [1]

Typisch voorbeeld

DoB + Sex + ZIP

- Afzonderlijk identificeren deze attributen geen burgers (of misschien uitzonderlijk)
- Gecombineerd uniek voor 87% [2] van de bevolking in de US (en dus een identifier)
- Bijkomend identificatierisico: DoB, Sex, ZIP wijdverspreid zijn
  - VIPs
  - Profielen sociale media
  - Publiek beschikbare CV's
  - Vrienden, familie, kennissen
  - ….

Lager reidentificatierisico naarmate minder entiteiten quasi-ids kunnen koppelen aan natuurlijk persoon

Quid gender X? Quasi-identifier verdient extra aandacht

[1] https://en.wikipedia.org/wiki/Quasi-identifier
[2] Latanya Sweeney . k-Anonymity: A model for protecting privacy. Mei 2001. https://epic.org/privacy/reidentification/Sweeney_Article.pdf

Smals ICT for society

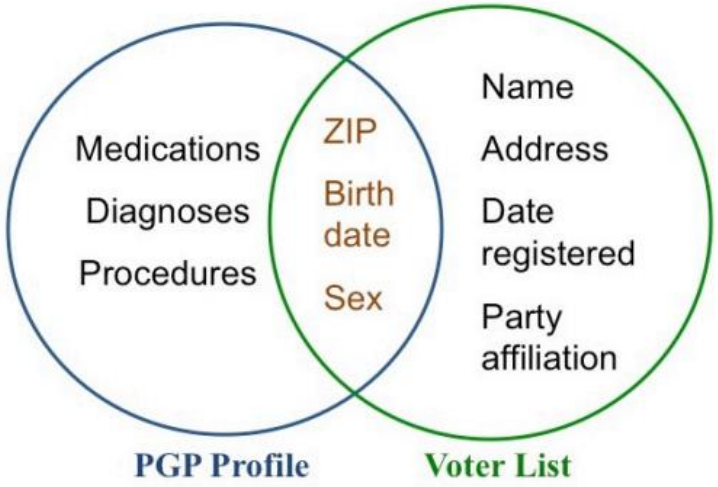
Personal Genome Project (PGP)

± 100 000 volunteers

DNA information, behavioral traits, medical conditions, physical characteristics, and environmental factors

51% of volunteers discosed exact DoB+Sex+ZIP

Identificeerbare persoonsgegevens onder GDPR. Dus niet anoniem!
Geen correcte pseudonimisering



Sample of 1130 records
241 gelinkt aan unieke id
84% / 97% correct

"Our sources for public records was a national sample of voter registrations ("Voter Data") and online access to a public records website ("Public Records"). The voter data was purchased from a third-party data broker and contained a sample of voter registrations for the 5-digit ZIP codes listed in Dataset. "

Sweeney, Latanya, Akua Abu, and Julia Winn. "Identifying participants in the personal genome project by name." Available at SSRN 2257732 (2013).

Smals ICT for society

HIPAA Identifiers

- Names
- **Geographic data** (all geographic subdivisions smaller than state, including street address, city, county, and zip code)
- **Dates related to an individual** (including birthdate, admission date, discharge date, date of death, and exact age if over 89)
- Telephone numbers
- FAX numbers
- Social Security numbers
- Email addresses
- Medical record numbers
- Account numbers
- Health plan beneficiary numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers including license plates
- Web URLs
- Device identifiers and serial numbers
- Internet protocol addresses
- Full face photos and comparable images
- Biometric identifiers (i.e. retinal scan, fingerprints)
- Any unique identifying number or code

The Health Insurance Portability and Accountability Act (HIPAA) is a United States federal statute. It was created primarily to modernize the flow of healthcare information, stipulate how personally identifiable information maintained by the healthcare and healthcare insurance industries should be protected from fraud and theft, and address limitations on healthcare insurance coverage.

One or more of these identifiers turns health information into PHI,

Smals ICT for society

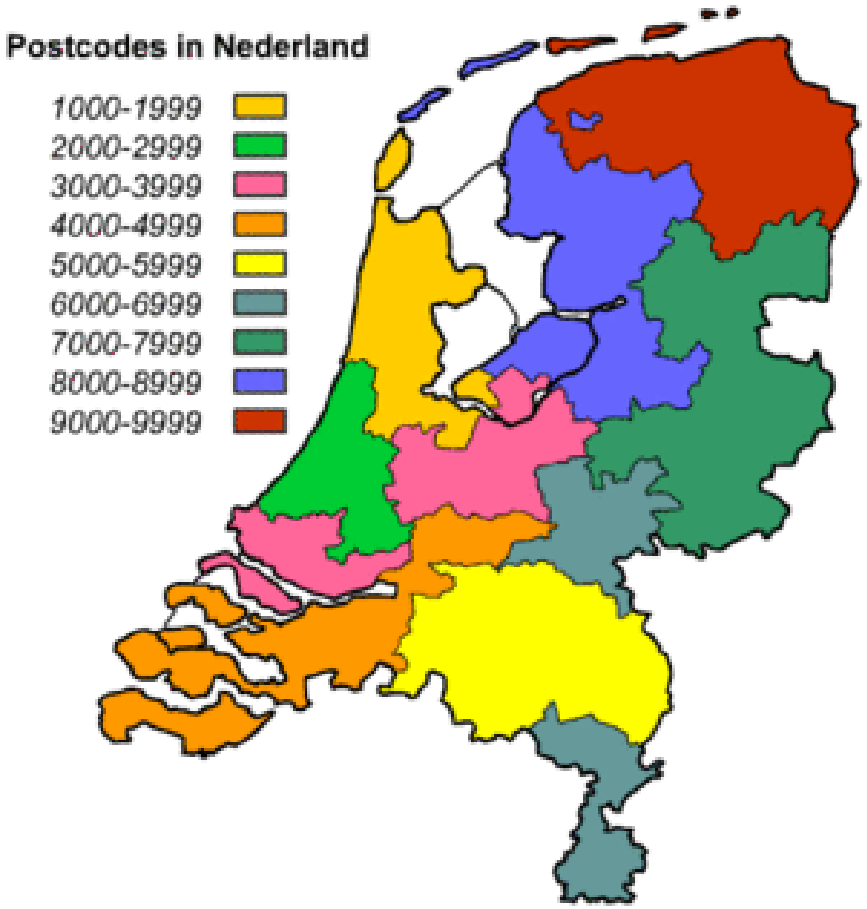
Ziekte	Datum diagnose	Hersteld	Sex	Burgerlijke status	Provincie	Bedrag
Rheumatism	28/02/2007	No	Man	Ongehuwd	Luik	8300
Smallpox	29/04/2014	No	Vrouw	Gehuwd	Waals-Brabant	100
Plague	03/04/2006	No	Vrouw	Gehuwd	Namen	1300
Meningitis	10/11/2001	Yes	Man	Verweduwd	Luik	9700
Necrotizing Fasciitis	25/04/2015	No	Man	Gehuwd	Vlaams-Brabant	7300
Metastatic cancer	16/04/2011	Yes	Man	Verweduwd	Brussels Hoofdstedelijk Gewest	9700
Coronary heart disea	06/07/2017	Yes	Vrouw	Verweduwd	Limburg	8600
Chavia	26/10/2000	Yes	Vrouw	Gehuwd	Luxemburg	5500
Crohn's Disease	23/05/2020	No	Vrouw	Gehuwd	West-Vlaanderen	2900

PHI under HIPAA

Nederland

Vorm: 1234 AB

Straat



analysis is based on registry office data of 2.7 million Dutch citizens, ~16% of the total population. [ ] 15 of 441 Dutch municipalities of varying size.

found that ***67.0% of the sampled population is unambiguously identiable by date of birth and four-digit postal code alone****, and that 99.4% is unambiguously identiable if date of birth, full postal code and gender are known.*

België telt 2,3 x meer inwoners per postcode.

→ Beduidend minder dan 67% burgers uniek identificeerbaar o.b.v. geboortedatum + postcode

Koot M, Noordende G, De Laat C. A study on the re-identifiability of Dutch citizens. In Workshop on Privacy Enhancing Technologies (PET 2010) 2010.

Smals ICT for society

Quasi-identifiers - Conclusie

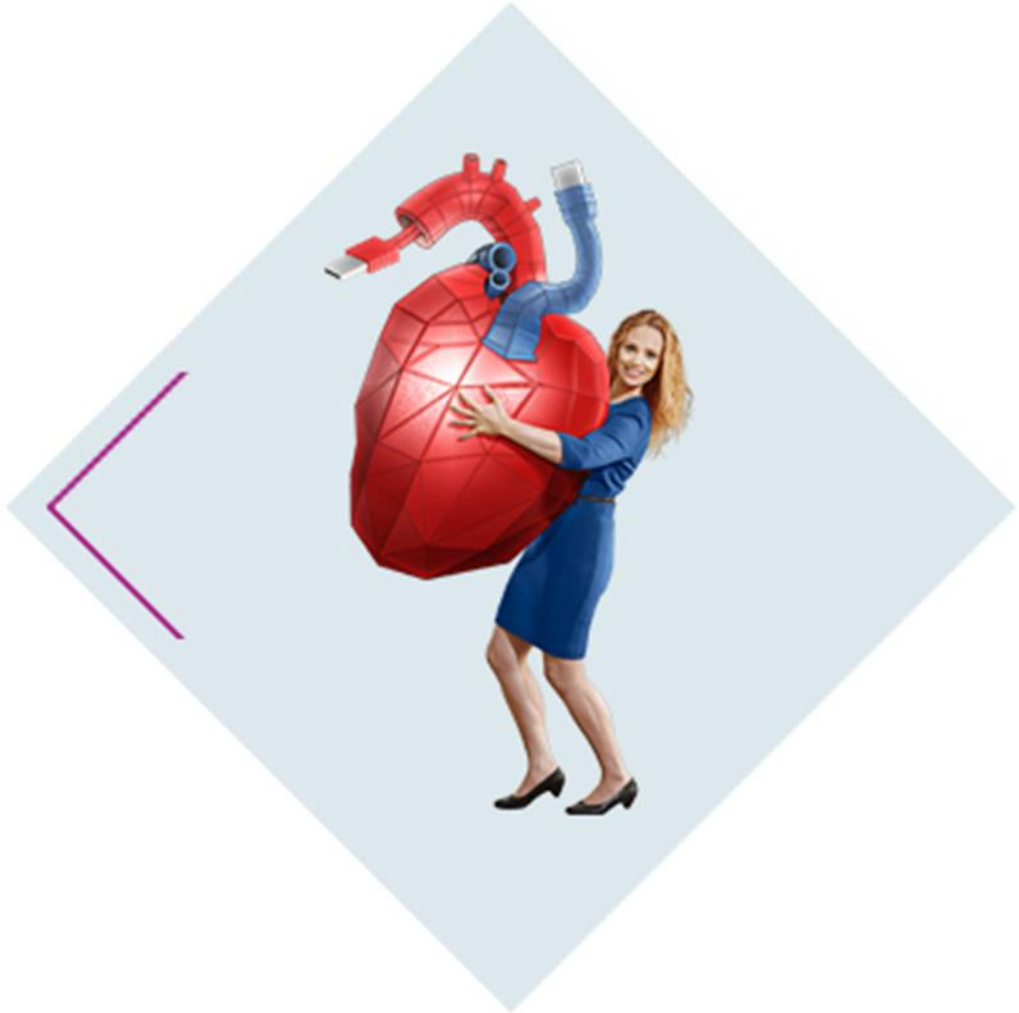
Combinatie van quasi-identifiers kan uniek zijn voor een persoon (en dus een identifier vormen)

Kan veel meer zijn dan DoB, Gender & ZIP

Cruciale vraag bij risicoinschatting: *Wie is in staat specifieke quasi–identifiers aan natuurlijke personen te koppelen?*

Agenda

- GDPR terminologie
- Quasi-identifiers
- Vervagingstechnieken
- Hoogdimensionale data
- Hashing & encryptie
- Afronding



Vervagingstechnieken

Willen een evenwicht vinden tussen privacy en nut door details van de data weg te laten, maar ook weer niet te veel.

Eigenlijk “anonimisatietechnieken” → misleidende term

Generalisatie

Naam	RRN	DoB	Sex	ZIP	Problem
X38LS45		20/05/1984	M	3211	Antisocial personality disorder
X38DI56		13/01/1976	M	9300	Schizofrenie
X38XD12		11/09/1937	F	1000	Multiple sclerosis
X38MP68		12/11/1950	M	4000	Megalomanie

Naam	RRN	Age	Sex	Loc	Problem
X38LS45		32	M	Vl-B	Antisocial personality disorder
X38DI56		39	M	O-Vl	Schizofrenie
X38XD12		77	F	BXL	Multiple sclerosis
X38MP68		64	M	LUIK	Megalomanie

Toevoegen ruis

Naam	RRN	DoB	Sex	ZIP	Problem
X38LS45		20/05/1984	M	3211	Antisocial personality disorder
X38DI56		13/01/1976	M	9300	Schizofrenie
X38XD12		11/09/1937	F	1000	Multiple sclerosis
X38MP68		12/11/1950	M	4000	Megalomanie

Naam	RRN	DoB	Sex	ZIP	Problem
X38LS45		07/07/1984	M	3211	Antisocial personality disorder
X38DI56		28/12/1975	M	9300	Schizofrenie
X38XD12		29/09/1937	F	1000	Multiple sclerosis
X38MP68		07/12/1950	M	4000	Megalomanie

Smals ICT for society

Aggregeren

Naam	RRN	DoB	Sex	ZIP	Problem
X38LS45		20/05/1984	M	3211	Antisocial personality disorder
X38DI56		13/01/1976	M	9300	Schizofrenie
X38XD12		11/09/1937	F	1000	Multiple sclerosis
X38MP68		12/11/1950	M	4000	Megalomanie

DoB	Sex	ZIP	Problem
≥ 1970	…	…	Antisocial personality disorder, Schizofrenie
< 1970	…	…	Megalomanie, Multiple sclerosis

K-anonymity

2-anonymity

SSN	Race	BirthDate	Gender	ZIP	Problem
021-57-1445	black	9/20/1965	male	02141	short of breath
021-77-8034	black	2/14/1965	male	02141	chest pain
107-21-0876	black	10/23/1965	female	02138	painful eye
021-37-1573	black	8/24/1965	female	02138	wheezing
021-54-4229	black	11/7/1964	female	02138	obesity
117-26-3042	black	12/1/1964	female	02138	chest pain
127-91-4819	white	10/23/1964	male	02138	short of breath
270-89-1234	white	3/15/1965	female	02139	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	obesity
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ever
white	2/13/1967	male	02138	117-61-0504	omiting
white	3/21/1967	male	02138	021-668-9440	ack pain

Identifiers | **Quasi-identifiers (*)** | **Sensitive data (*)**

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

Elk record heeft minstens twee mogelijke subjecten

Hogere k-waarde => hogere privacy, lager nut

Geavanceerdere aanpakken: l-diversity, t-closeness, …

* Onderscheid quasi-identifiers en sensitive data lijkt me artificieel

Smals ICT for society

Vervangingstechnieken

Hoe representatief zijn dergelijke voorbeelden met een zeer beperkt aantal attributen per record?

In welke mate kan data werkelijk genanonimiseerd worden zonder de data de facto nutteloos te maken?

Eerder *vervagingstechnieken* dan *anonimisatietechnieken*

→ Kunnen identificatierisico reduceren, ook al is resultaat niet juridisch anoniem

Smals ICT for society

Over het potentieel van “anonimisatie” technieken

Whitehouse.org

“Anonimisatie van een data record lijkt misschien makkelijk te implementeren. Helaas is het steeds makkelijker om anonimisatie teniet te doen met behulp van de technieken die ontwikkeld worden voor legitieme toepassingen van big data.”

“Sommige oudere technologieën, zoals anonimisatietechieken, hebben maar een beperkt toekomstig potentieel, hoewel ze in het verleden wel waardevol waren.”

President’s Council of Advisors on Science and Technology (PCAST). ***Big Data and Privacy: A Technological Perspective***. Mei 2014. http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

Paul Ohm

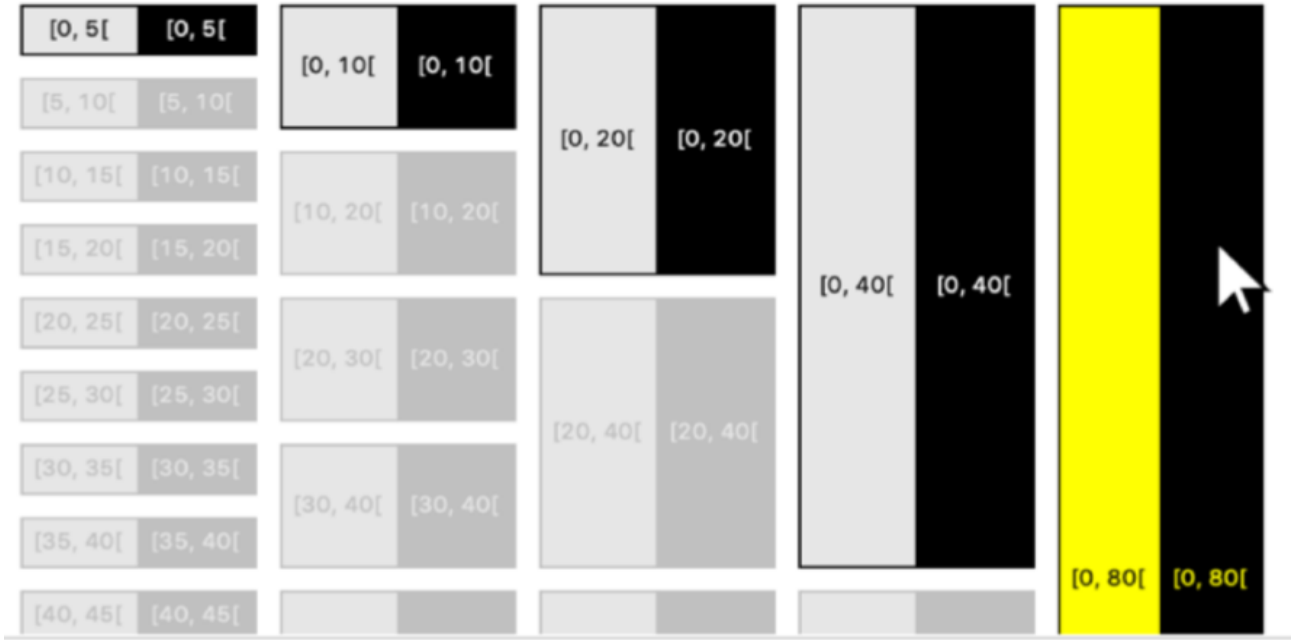
Broken Promises of Privacy

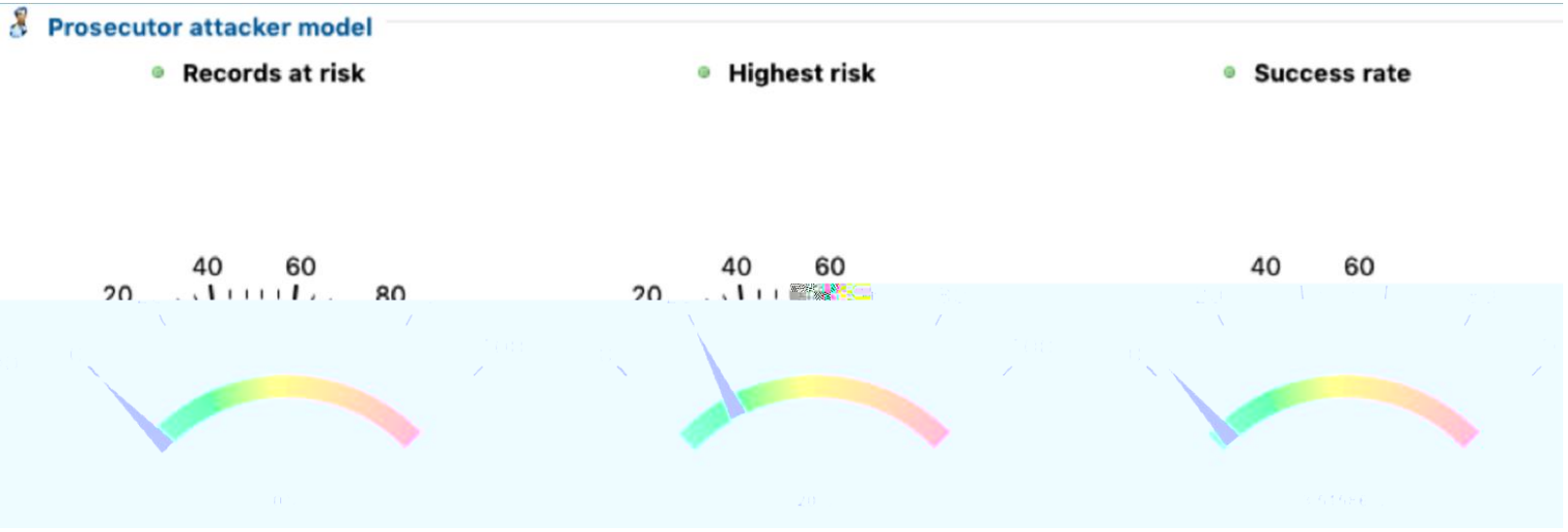
- *“De robuuste anonimisatie-veronderstelling is niet fundamenteel foutief, maar wel diep gebrekkig.”*
- *“We kunnen niet voorspellen tot welke en tot hoeveel externe informatie de aanvaller toegang heeft”*
- *“Persoonsgegevens zijn een steeds groeiende categorie. Tien jaar terug beschouwde bijna niemand filmbeoordelingen als persoonsgegevens.”*
- *“Het aanwasprobleem: Eénmaal een aanvaller twee ‘geanonimiseerde’ databases aan elkaar gelinkt heeft, kan er makkelijker andere informatie aan gelinkt worden, wat kan helpen bij deanonimisatie.”*

Ohm P. Broken promises of privacy: Responding to the surprising failure of anonymization. Ucla L. Rev.. 2009;57:1701.

Smals ICT for society

ARX





ATTACK MODELS

Prosecutor
Adversary can know that the target is in the data set

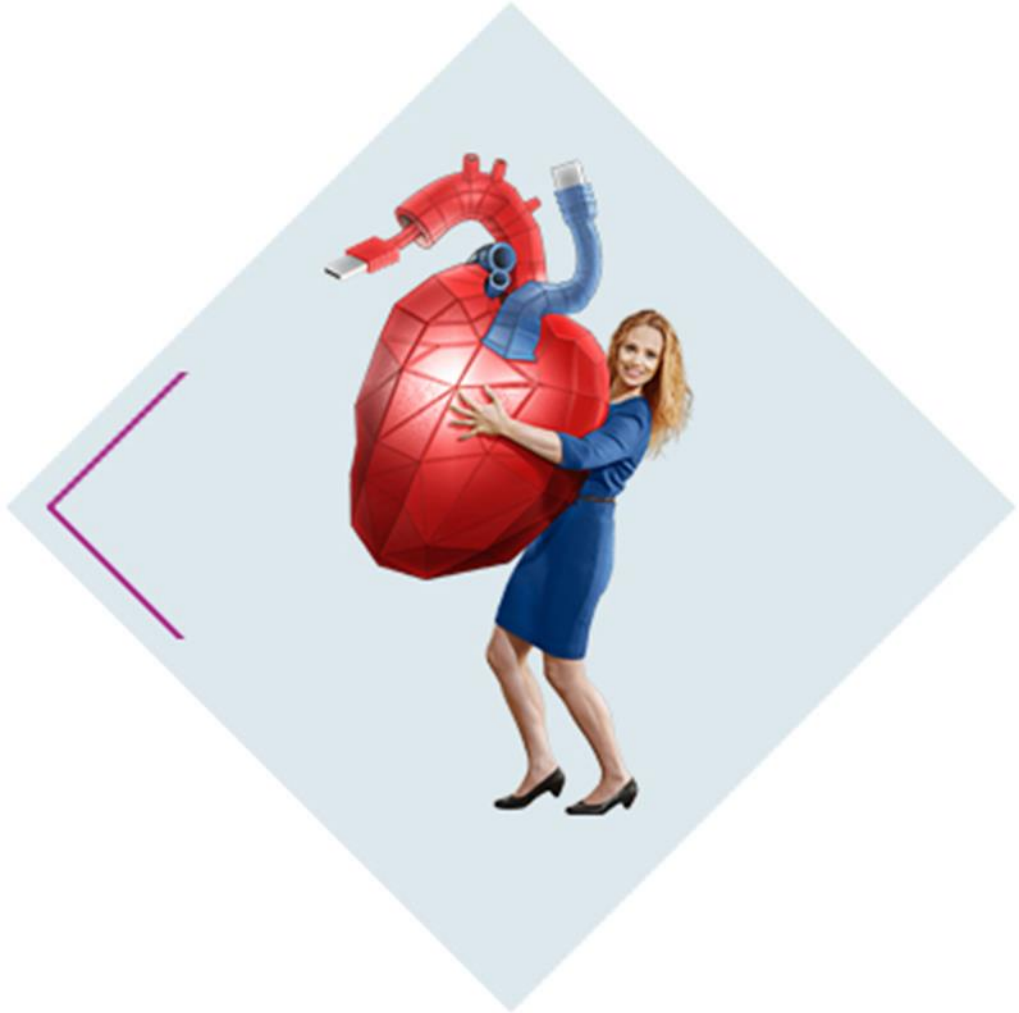
Journalist
Adversary doesn't know for certain that the target is in the data set

Marketer
Adversary attempts to re-identify as many subjects in the data set as possible

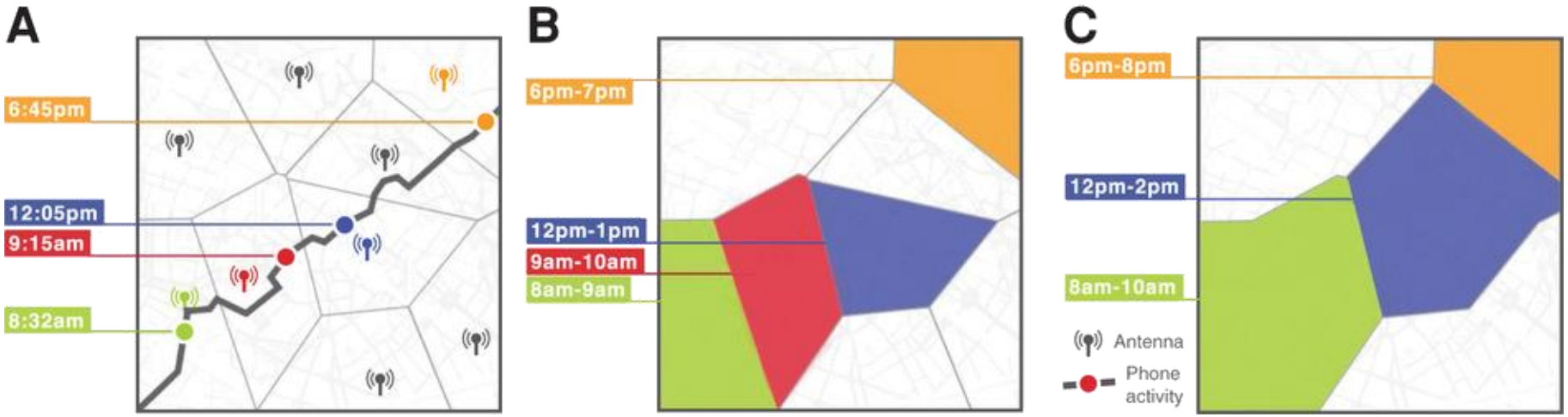
https://www.smalsresearch.be/download/review/quick_review/qr_arx.pdf

Agenda

- GDPR terminologie
- Quasi-identifiers
- Vervagingstechnieken
- Hoogdimensionale data
- Hashing & encryptie
- Afronding



Hoogdimensionale data – Location tracking



Vervagen data (tijd, locatie) heeft weinig impact op uniekheid
Maar maakt dat wel veel minder bruikbaar

Abstract

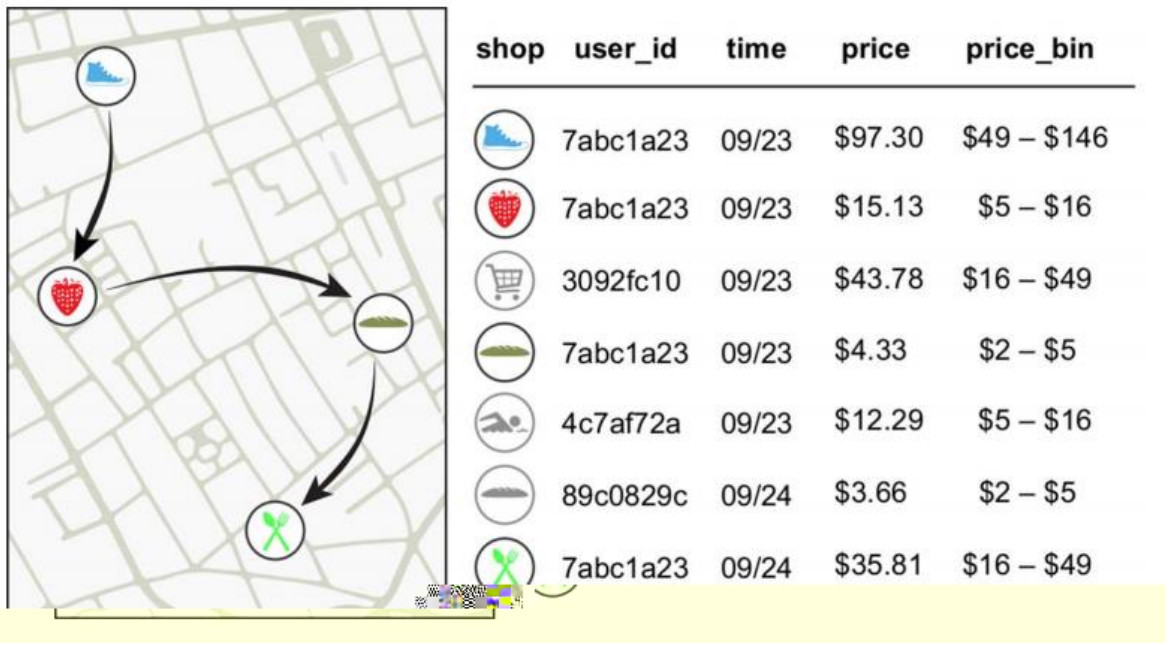
- Fifteen months of human mobility data for one and a half million individuals
- Human mobility traces are highly unique. Four spatio-temporal points are enough to uniquely identify 95% of the individuals. (point: lication data provided by per hour, per antenna)
- The uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity.
- These findings represent fundamental constraints to an individual's privacy

[M13] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel. ***Unique in the Crowd: The privacy bounds of human mobility***. 25 maart 2013. Scientific Reports 3, Article number: 1376.
http://www.nature.com/srep/2013/130325/srep01376/fig_tab/srep01376_F1.html

27

Hoogdimensionale data – Aankoopgedrag

Data
- Kredietkaarttransacties van 1,1 miljoen mensen in 10 000 handelszaken in één county
- Namen, kredietkaartnummers, adressen winkels en exacte tijdstippen van transactie
- Wat overbleef: gespendeerde bedrag, winkeltype (vb. Restaurant, fitness, kruidenier) en code per persoon



shop	user_id	time	price	price_bin
	7abc1a23	09/23	$97.30	$49 – $146
	7abc1a23	09/23	$15.13	$5 – $16
	3092fc10	09/23	$43.78	$16 – $49
	7abc1a23	09/23	$4.33	$2 – $5
	4c7af72a	09/23	$12.29	$5 – $16
	89c0829c	09/24	$3.66	$2 – $5
	7abc1a23	09/24	$35.81	$16 – $49

Four spatiotemporal points* (shop, date) *are enough to uniquely reidentify 90% of individuals. *We show that knowing the price of a transaction increases the risk of reidentification by 22%, on average. Finally, we show that even data sets that provide coarse information at any or all of the dimensions provide little anonymity and that* ***women are more reidentifiable than men*** *in credit card metadata.*

Science. *Credit card study blows holes in anonymity*. 30 januari 2015
http://www.sciencemagazinedigital.org/sciencemagazine/30_january_2015?folio=468#pg16
De Montjoye YA, Radaelli L, Singh VK, Pentland AS. *Unique in the shopping mall: On the reidentifiability of credit card metadata*. Science. 2015 Jan 30;347(6221):536-9.

Smals ICT for society

Uniek ≠ Identificeerbaar

Uniekheid zegt iets over identificatierisico

Niet-verwaarloosbaar identificatierisico → geen anonieme gegevens

Smals ICT for society

Identificaties “Geanonimiseerde” hoogdimensionale data

- 2006: publicatie 20 miljoen “geanonimiseerde” zoekopdrachten (650.000 users, 3 maand).
- The New York Times achterhaalde identiteit van meerdere gebruikers
- Queries s.a.: “*landscapers in Lilburn, Ga*,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.”
- Ontslag CTO, #57 in CNNs “*101 Dumbest Moments in Business*”

- 2007: Publicatie “geanonimiseerde” records met filmratings 500.000 gebruikers
- Identificatie gebruikers in combinatie met publieke IMDB data
- Voor de rechter door gebruikers

- 2014: Publicatie “geanonimiseerde” info over 173 miljoen ritten: tijden, routes & tarieven
- Gelinkt aan celebrities met time-stamped foto’s van in- en utistappen de celebrities (d.m.v. website voor celebrity-spotter bloggers)
- We weten de route en betaalde bedrag per geïdentificeerde rit.

“Geanonimseerde” hoogdimensionale data blijken toch persoonsgegevens te zijn

http://shawndra.pbworks.com/f/A+Face+Is+Exposed+for+AOL+Searcher+No.+4417749+-+New+York+T.pdf

Narayanan A, Shmatikov V. How to break anonymity of the netflix prize dataset. arXiv preprint cs/0610105. 2006 Oct 18.

https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn

Hoogdimensionale data

Type	Voorbeeld	Dimensies
Sociale netwerken 	Gemiddeld 100 vrienden op facebook	
Reviews 	Gemiddelde Netflix gebruiker in dataset: 213 ratings	426 (tijd+rating)
Locatie tracking 	Elke 2 uur locatie + tijd	720 per maand
Aankoop-gedrag 	4 keer per maand 20 producten aankopen	160 per maand
Surfgedrag	…	…
Medisch dossier	…	…
Genetische data	…	1.000.000
…	…	…

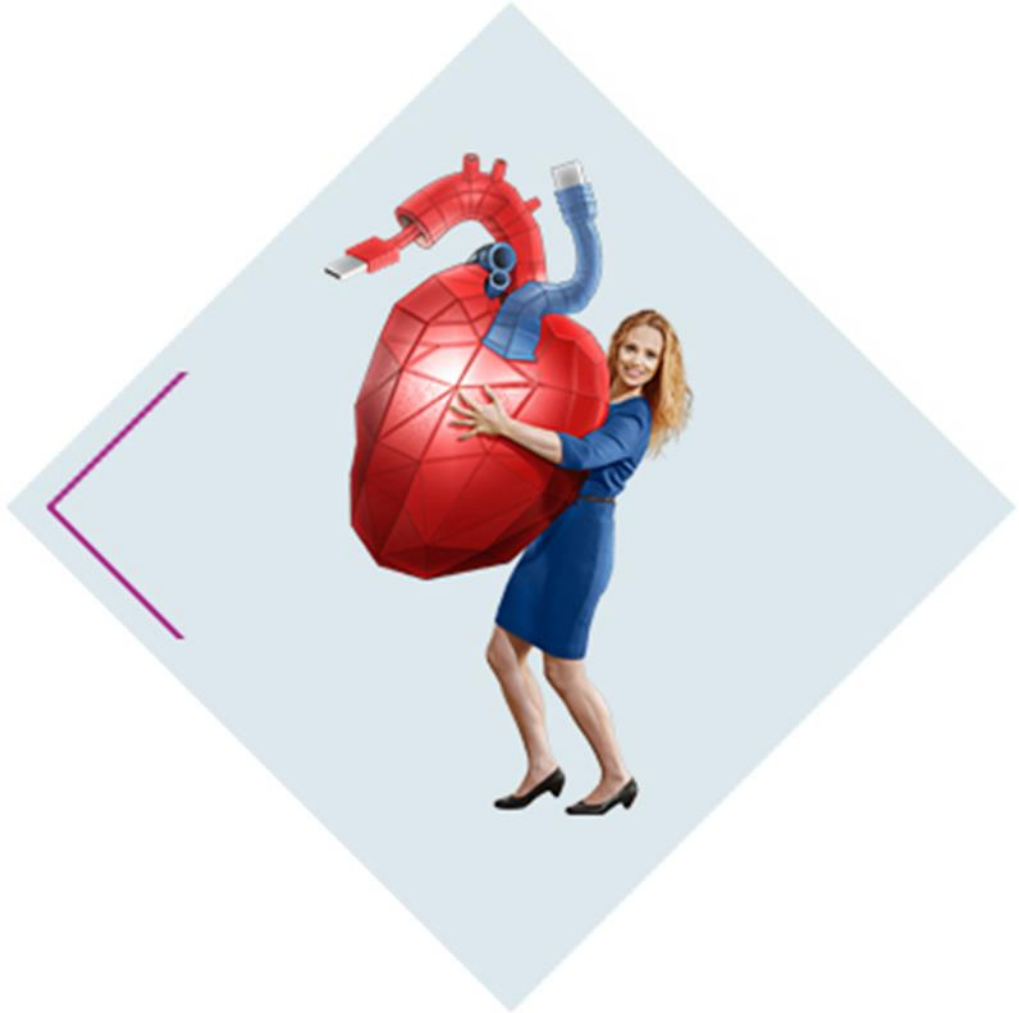
Zeer snel unieke combinatie (en dus identificatierisico)

Smals ICT for society

Quasi onmogelijk om hoogdimensionale data werkelijk te anonimiseren

Agenda

- GDPR terminologie
- Quasi-identifiers
- Vervagingstechnieken
- Hoogdimensionale data
- Hashing & encryptie
- Afronding



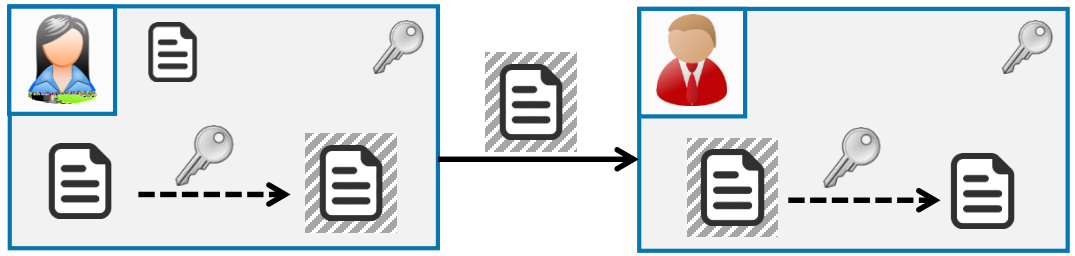
Encryptie

Doel

Confidentialiteit = vertrouwelijkheid

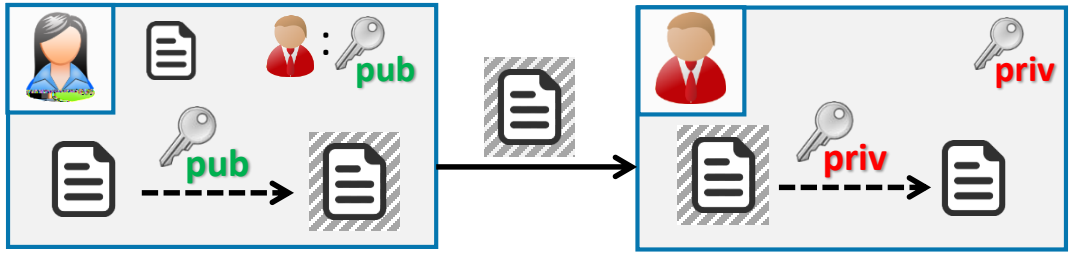
Symmetrische encryptie

- Sneller
- vb. AES



Publieke sleutel encryptie

- Trager
- Vb. RSA, ElGamal



Vercijferen van persoonsgegevens

Pseudonimisering. het verwerken van persoonsgegevens op zodanige wijze dat de persoonsgegevens niet meer aan een specifieke betrokkene kunnen worden gekoppeld zonder dat er **aanvullende gegevens** worden gebruikt, mits

1. deze aanvullende gegevens **apart worden bewaard** en
2. **technische en organisatorische maatregelen worden genomen om ervoor te zorgen dat de persoonsgegevens niet aan een geïdentificeerde of identificeerbare natuurlijke persoon worden gekoppeld**."

Vercijferde persoonsgegevens zijn gepseudonimiseerde persoonsgegevens

De (symmetrische of private) sleutel is de aanvullende gegegevens

GDPR van toepassing. Dus passende maatregelen vereist.
→ Voldoende beschermen van (geheime of private) sleutel kan volstaan

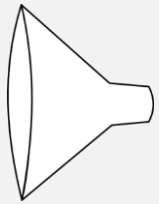
Het verwijderen van de sleutel kan volstaan om de cijfertekst als geanonimiseerd te beschouwen

Smals
ICT for society

Cryptografische hashfunctie

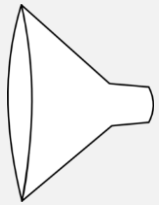
- Integriteit
- Zeer courant gebruikt (vb. elektronische handtekeningen, bestanden, blockchain)
- Vb. SHA1, SHA2, RIPE-MD

```
5e 50 6e 82 7f d5 50 ec 4e 08 8e e7 75 8f 34 b3
a6 8e 34 93 d5 89 98 52 97 48 f0 c6 c1 70 f3 3c
```



```
5f 3b fa 41 9c 63 be 2a 3a 09 ad bd 06 30 c5 1f
64 5e b0 3a ba fc d5 f2 ad 39 63 7a 30 6b 41 77
```

“Hell0 world!”



```
c0 5a 70 4b ef b2 9f f4 9a f0 19 00 04 70 f8 87
d1 d2 4b 20 a2 d0 5b f6 f8 8b 2d f9 97 d1 85 10
```

Fixed-length output | One-way | Collision resistant

De hash van persoonsgegevens is
{anonieme, gepseudonimiseerde, geïdentificeerde} gegevens?

Hashen van identifiers

Hash van rijksregisternummer

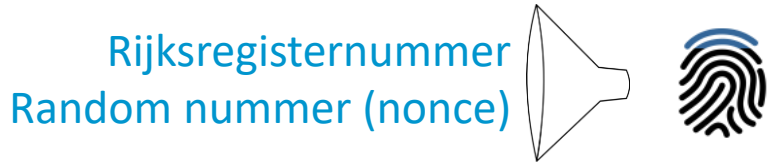
AANVAL

- Ongeveer 80M strings die structuur hebben van INSZ (BIS & NIS)
- 80M strings als input geven aan hash, tot aanvaller een matchende hash gevonden heeft
- Geen additionele informatie nodig (enkel wat rekenkracht)

Hash van identifier is nog steeds identifier

In record vervangen van rijksregisternummer door zijn hash is geen correcte pseudonimisering

Hash van rijksregisternummer met nonce



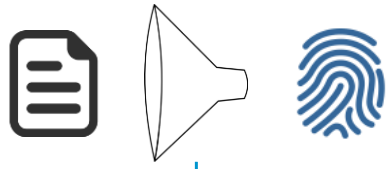
3 scenario's

- Nonce gekend door 1 of enkele partijen:
  → Gepseudonimiseerd
- Nonce publiek beschikbaar
  → Geïdentificeerd
- Nonce wordt (overal) verwijderd / vergeten,
  → Anoniem

In record vervangen van rijksregisternummer door hash(rijksregisternummer, nonce) kan een stap zijn in correcte pseudonimisering

Nonce: willeurige gekozen, voldoende lang getal, slechts 1x gebruikt.

Hashen van persoonsgegevens



Input heeft lage entropie

Makkelijk te raden

- Vb. Een paar honderd miljard mogelijkheden
- Vb. Gestructureerd document met enkel naam & datum als vrije velden

Geïdentificeerde data

Input heeft hoge entropie

De facto onmogelijk te raden

- Vb. Vrije, voldoende lange tekst
- Meer rekenkracht, efficiëntere algoritmes

Gepseudonimiseerde data

Waar ligt scheidingslijn?
(we weten wel dat die opschuift)

Smals ICT for society

Hashen van persoonsgegevens

Nonce gekende door 1 of enkele partijen

Gepseudonimiseerde data

Nonce publiek beschikbaar

Zie vorige slide

Nonce wordt (overal) verwijderd / vergeten

Anonieme data

Hangt af van verspreiding nonce

Nonce moet voldoende lang zijn & willekeurig gekozen
(vb. 256 bit entropie)

Hashing & Encryptie - Conclusie

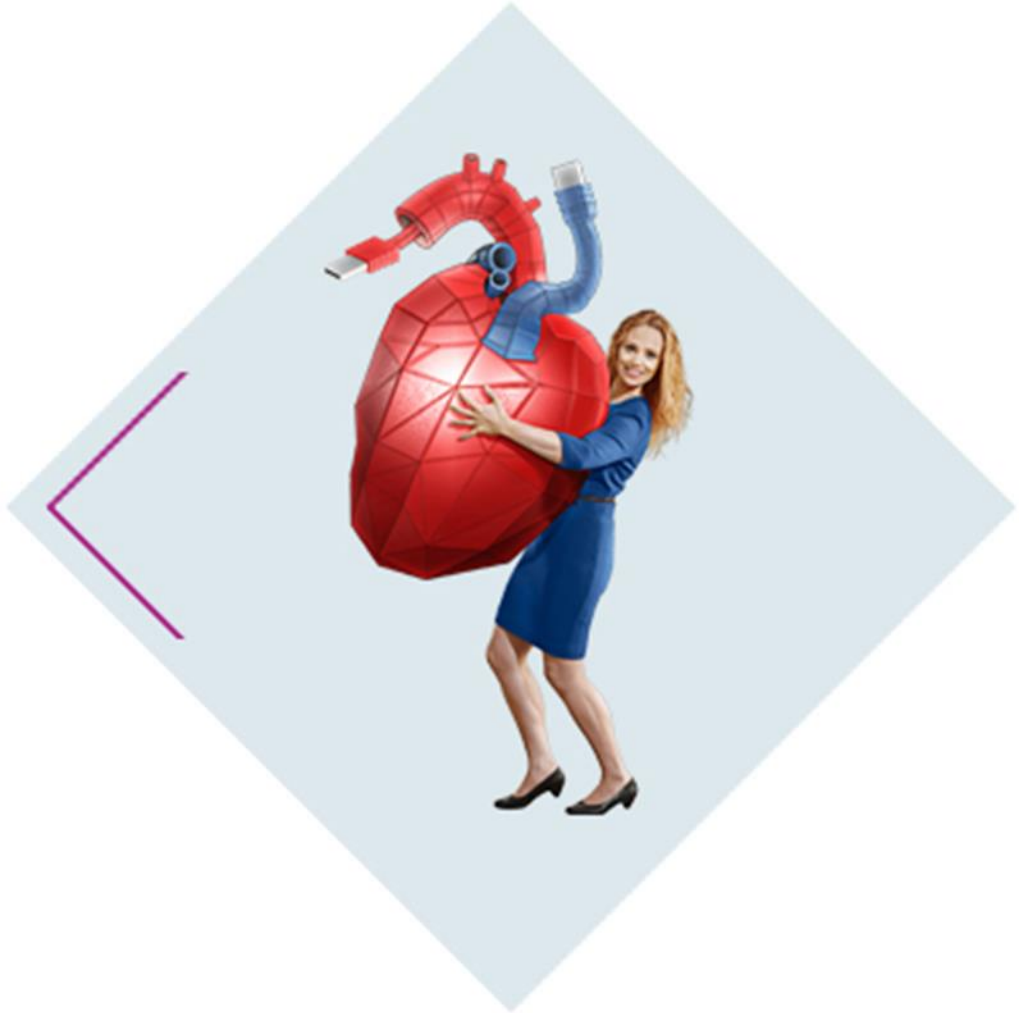
Hash van persoonsgegevens

Vercijfering van persoonsgegevens

Niet per se anonieme data
GDPR blijft veelal van toepassing

Agenda

- GDPR terminologie
- Quasi-identifiers
- Vervagingstechnieken
- Hoogdimensionale data
- Hashing & encryptie
- Afronding



Anonimisatie

- Niet te snel spreken over anonimisatie of anonieme data
- Werkelijk anonimiseren niet steeds mogelijk
- Anonimisatietechnieken is een misleidende term

Pseudonimisatie

- Pseudonimisatie moet correct uitgevoerd wordt
- Vervangen identifier door code/pseudoniem niet steeds voldoende

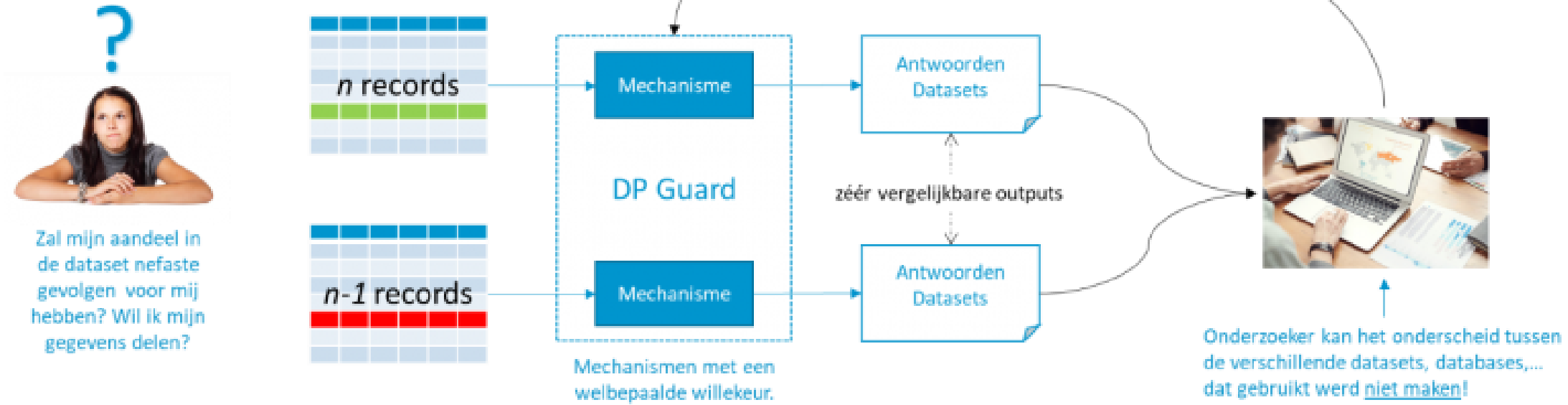
Breng de data naar de berekeningen

Breng de berekeningen naar de data

Smals ICT for society

Differential privacy

Idee
Onderzoeker krijgt geen toegang tot ruwe dataset
Kan wel één of meerdere queries uitvoeren
Query / query's bevatten ruis om privacy (bewijsbaar) te beschermen



https://www.smalsresearch.be/differential-privacy/

GDPR

NEW

Anonieme gegevens

Onlinkbaar
aan natuurlijk persoon

GDPR niet van toepassing

Gepseudonimiseerde persoonsgegevens

Linkbaar met aanvullende gegevens
aan natuurlijk persoon

GDPR van toepassing # bepalingen versoepeld

Geïdentificeerde persoonsgegevens

Linkbaar zonder aanvullende gegevens
aan natuurlijk persoon

GDPR ten volle van toepassing

← (Identificeerbare) Persoonsgegevens →

Tot welke categorie behoren gegevens?

Er zijn grijze zones → interpretatie

***Persoonsgegevens*:** alle informatie over een geïdentificeerde of identificeerbare natuurlijke persoon („de betrokkene”); als identificeerbaar wordt beschouwd een natuurlijke persoon die **direct of indirect kan worden geïdentificeerd**, met name aan de hand van een identificator zoals een naam, een identificatienummer, locatiegegevens, een online identificator of van **een of meer elementen** die kenmerkend zijn voor de fysieke, fysiologische, genetische, psychische, economische, culturele of sociale identiteit van die natuurlijke persoon;

Kristof Verslype

Cryptographer, PhD

Smals Research

Smals
ICT for society

kristof.verslype@smals.be

www.smals.be
www.smalsresearch.be
www.cryptov.net (personal)