Cisco Webex Meetings

12:33

Meeting Info

Speaking: **Khaled El Emam**

Layout

Viewing Khaled El Emam...

An Introduction to Re-identification Risk Measurement

Khaled El Emam
kelemam@ehealthinformation.ca

Mute | Start video | Share | ··· | Participants | Chat | ···

Speaking: **Khaled El Emam**

Disclosures

- Khaled El Emam is co-founder and director of Replica Analytics Ltd, a spinoff company from the University of Ottawa / CHEO Research Institute specializing in the development of data synthesis software for health data. In December 2021 Replica was acquired by Aetion.

Speaking: **Khaled El Emam**

Basic definitions – identity disclosure is when a person's identity is assigned to a record

Quasi-identifiers

Sex	Year of Birth	NDC
Male	1975	009-0031
Male	1988	0023-3670
Male	1972	0074-5182
Female	1993	0078-0379
Female	**1989**	**65862-403**
Male	1991	55714-4446
Male	1992	55714-4402
Female	1987	55566-2110
Male	1971	55289-324
Female	1996	54868-6348
Male	1980	53808-0540

Speaking: **Khaled El Emam**

Basic definitions – identity disclosure is when a person's identity is assigned to a record

Sally

Quasi-identifiers

Sex	Year of Birth	NDC
Male	1975	009-0031
Male	1988	0023-3670
Male	1972	0074-5182
Female	1993	0078-0379
Female	**1989**	**65862-403**
Male	1991	55714-4446
Male	1992	55714-4402
Female	1987	55566-2110
Male	1971	55289-324
Female	1996	54868-6348
Male	1980	53808-0540

Electronic Health Information Laboratory

Basic definitions – generalization means that more than one record can match a person

Sex	Year of Birth	NDC
Male	1970-1979	009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	**1980-1989**	**65862-403**
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	**1980-1989**	**55566-2110**
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540

Attacks can be in two directions – population to sample attack

Sex	Year of Birth	NDC
Male	1970-1979	009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	**1980-1989**	**65862-403**
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	**1980-1989**	**55566-2110**
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540

Attacks can be in two directions – population to sample attack

Sex	Year of Birth	NDC
Male	1970-1979	009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	**1980-1989**	**65862-403**
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	**1980-1989**	**55566-2110**
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540

- medical history
- language
- income range
- ZIPS

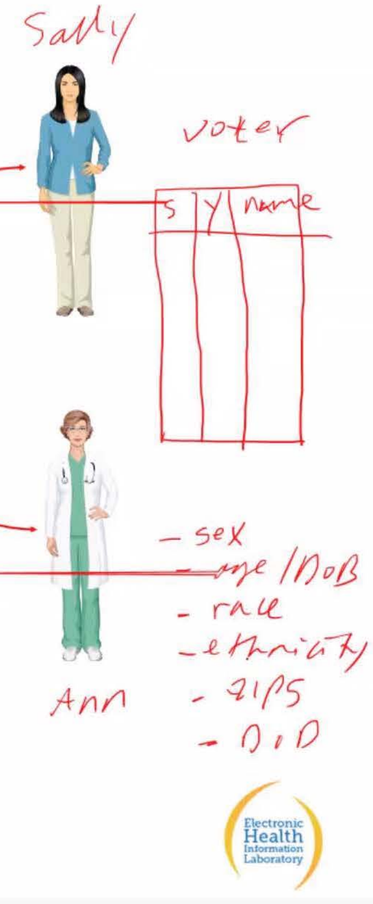
Attacks can be in two directions – sample to population attack

Sex	Year of Birth	NDC
Male	1970-1979	009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	**1980-1989**	**65862-403**
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	1980-1989	55566-2110
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540

Speaking: Khaled El Emam

Attacks can be in two directions – sample to population attack

Sex	Year of Birth	NDC
Male	1970-1979	009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	**1980-1989**	**65862-403**
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	1980-1989	55566-2110
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540



Electronic Health Information Laboratory

Viewing Khaled El Emam...

Risk is measured by the group size

Sex	Year of Birth	NDC	Group Size	Risk
Male	1975	009-0031	1	1
Male	1988	0023-3670	1	1
Male	1972	0074-5182	1	1
Female	1993	0078-0379	1	1
Female	**1989**	**65862-403**	**1**	**1**
Male	1991	55714-4446	1	1
Male	1992	55714-4402	1	1
Female	1987	55566-2110	1	1
Male	1971	55289-324	1	1
Female	1996	54868-6348	1	1
Male	1980	53808-0540	1	1

Speaking: Khaled El Emam

When we generalize the group size gets bigger, so the risk decreases

Sex	Decade of Birth	NDC	Group Size	Risk
Male	1970-1979	009-0031	3	0.33
Male	1980-1989	0023-3670	2	0.5
Male	1970-1979	0074-5182	3	0.33
Female	1990-1999	0078-0379	2	0.5
Female	**1980-1989**	**65862-403**	**2**	**0.5**
Male	1990-1999	55714-4446	2	0.5
Male	1990-1999	55714-4402	2	0.5
Female	**1980-1989**	**55566-2110**	**2**	**0.5**
Male	1970-1979	55289-324	3	0.33
Female	1990-1999	54868-6348	2	0.5
Male	1980-1989	53808-0540	2	0.5

Speaking: **Khaled El Emam**

When we generalize the group size gets bigger, so the risk decreases

* public
* k-anonymity
- max = 0.5
strict average
- average ← * non-public
- uricity

Sex	Decade of Birth	NDC	Group Size	Risk
Male	1970-1979	009-0031	3	0.33
Male	1980-1989	0023-3670	2	0.5
Male	1970-1979	0074-5182	3	0.33
Female	1990-1999	0078-0379	2	0.5
Female	**1980-1989**	**65862-403**	**2**	**0.5**
Male	1990-1999	55714-4446	2	0.5
Male	1990-1999	55714-4402	2	0.5
Female	**1980-1989**	**55566-2110**	**2**	**0.5**
Male	1970-1979	55289-324	3	0.33
Female	1990-1999	54868-6348	2	0.5
Male	1980-1989	53808-0540	2	0.5

8

But it is actually the population group size that matters

Sex	Year of Birth	NDC	Group Size	Risk
Male	1970-1979	009-0031	3	
Male	1980-1989	0023-3670	2	
Male	1970-1979	0074-5182	3	
Female	1990-1999	0078-0379	2	
Female	**1980-1989**	**65862-403**	**2**	**0.1**
Male	1990-1999	55714-4446	2	
Male	1990-1999	55714-4402	2	
Female	**1980-1989**	**55566-2110**	**2**	**0.1**
Male	1970-1979	55289-324	3	
Female	1990-1999	54868-6348	2	
Male	1980-1989	53808-0540	2	

N=10

Speaking: **Khaled El Emam**

But it is actually the population group size that matters

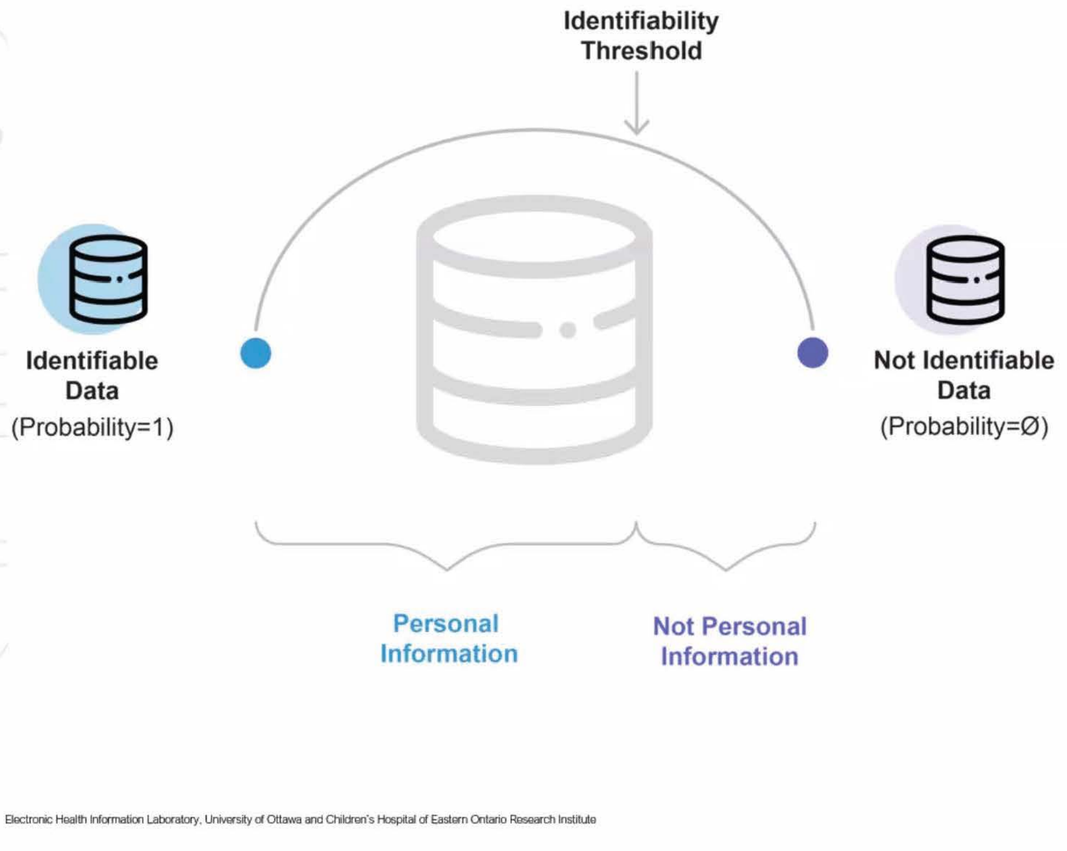
sample

Sex	Year of Birth	NDC	Group Size	Risk
Male	1970-1979	009-0031	3	
Male	1980-1989	0023-3670	2	
Male	1970-1979	0074-5182	3	
Female	1990-1999	0078-0379	2	
Female	**1980-1989**	**65862-403**	**2**	**0.1**
Male	1990-1999	55714-4446	2	
Male	1990-1999	55714-4402	2	
Female	**1980-1989**	**55566-2110**	**2**	**0.1**
Male	1970-1979	55289-324	3	
Female	1990-1999	54868-6348	2	
Male	1980-1989	53808-0540	2	

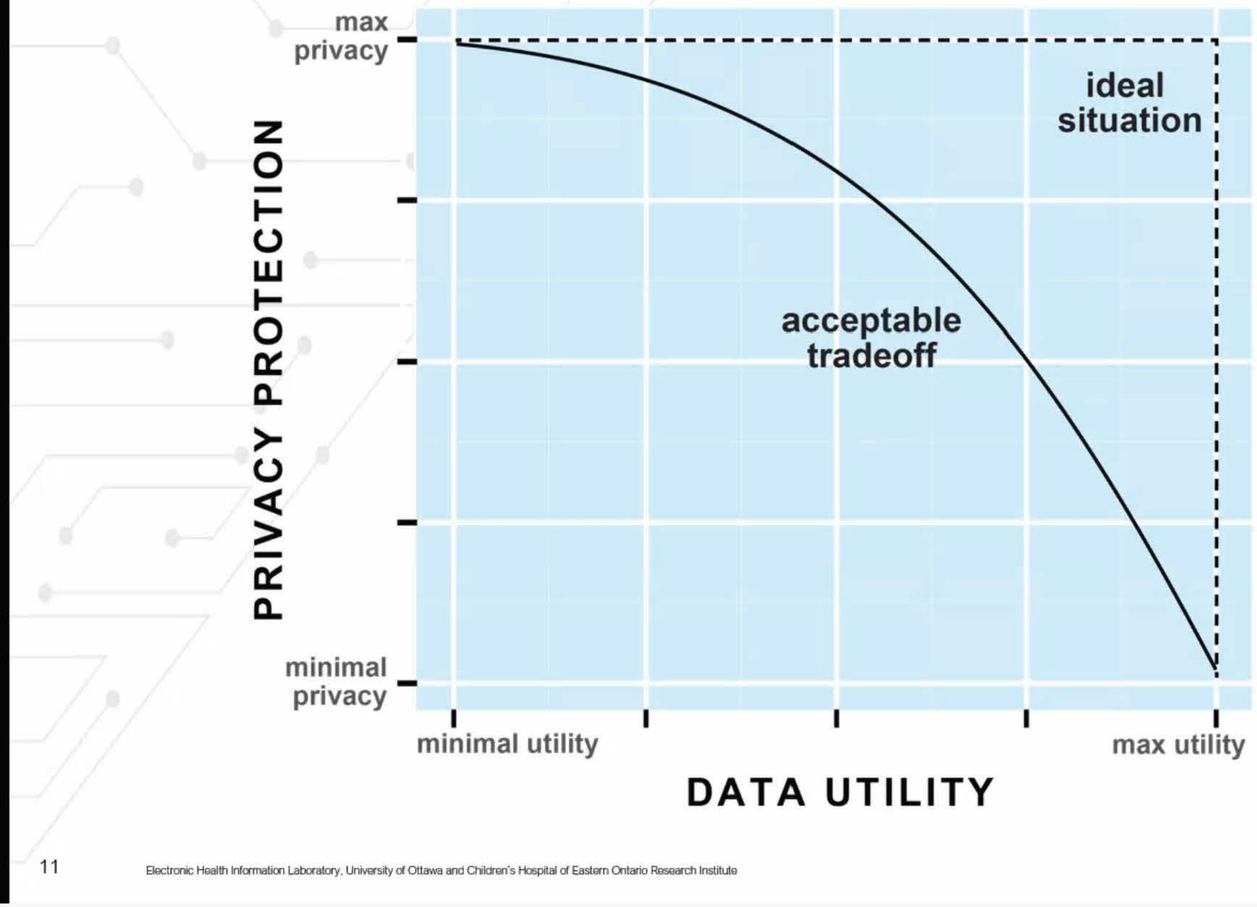
N=10

Electronic Health Information Laboratory

Identifiability spectrum and risk thresholds



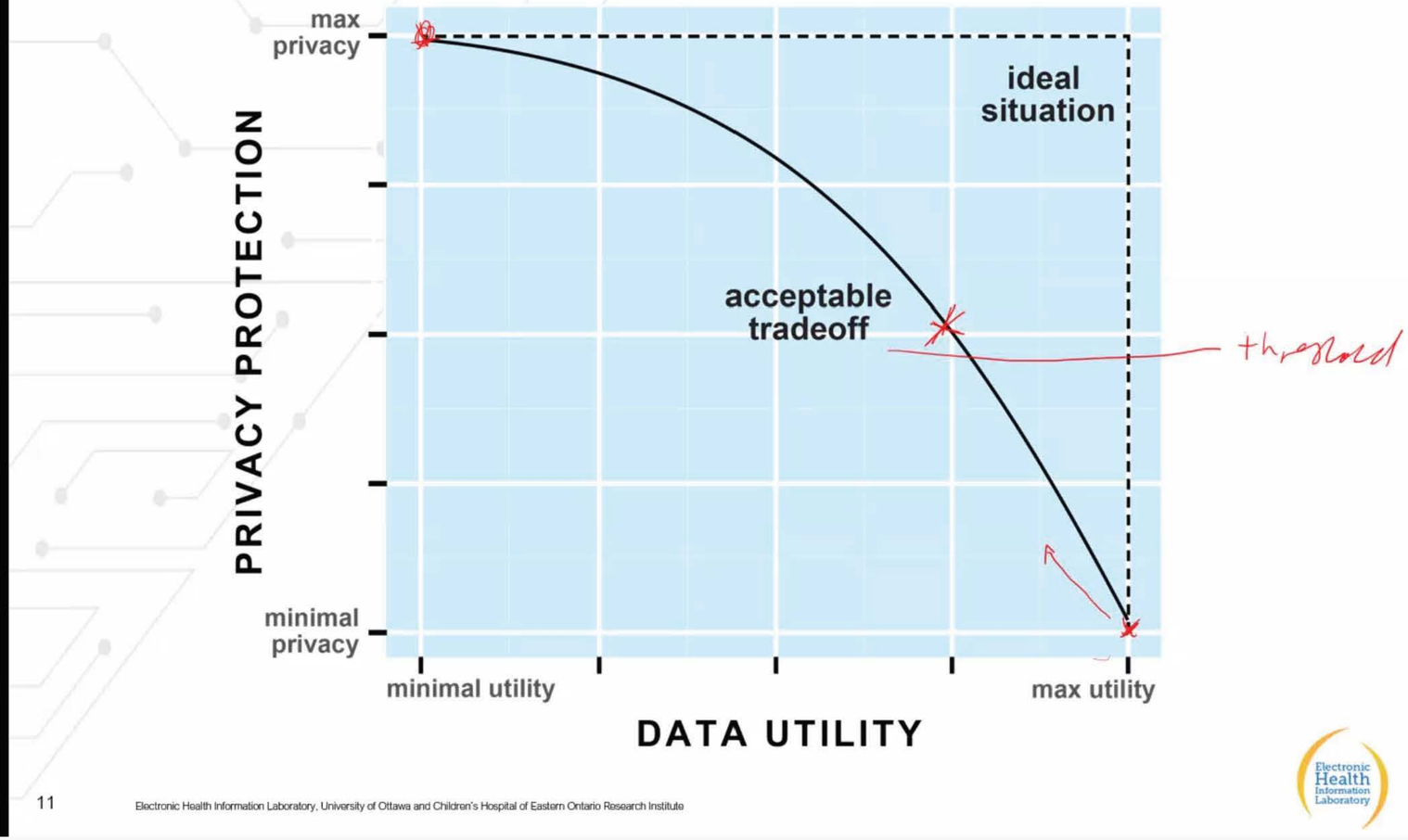
Privacy-Utility Trade-off



Electronic Health Information Laboratory

Speaking: Khaled El Emam

Privacy-Utility Trade-off



Electronic Health Information Laboratory

Viewing Khaled El Emam...

A common approach that has worked well in practice is risk-based anonymization

Data Transformations

- Generalization
- Suppression
- Addition of noise
- Microaggregation

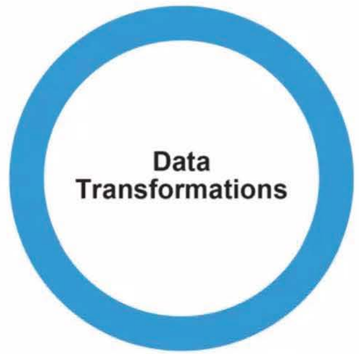
+

Controls

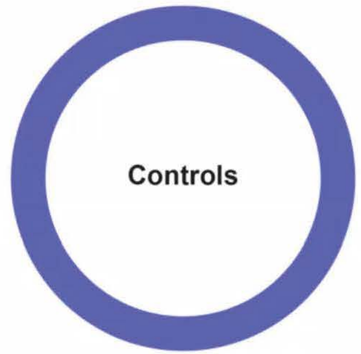
- Security controls
- Privacy controls
- Contractual controls

Non-public

A common approach that has worked well in practice is risk-based anonymization



- Generalization
- Suppression
- Addition of noise
- Microaggregation



- Security controls
- Privacy controls
- Contractual controls