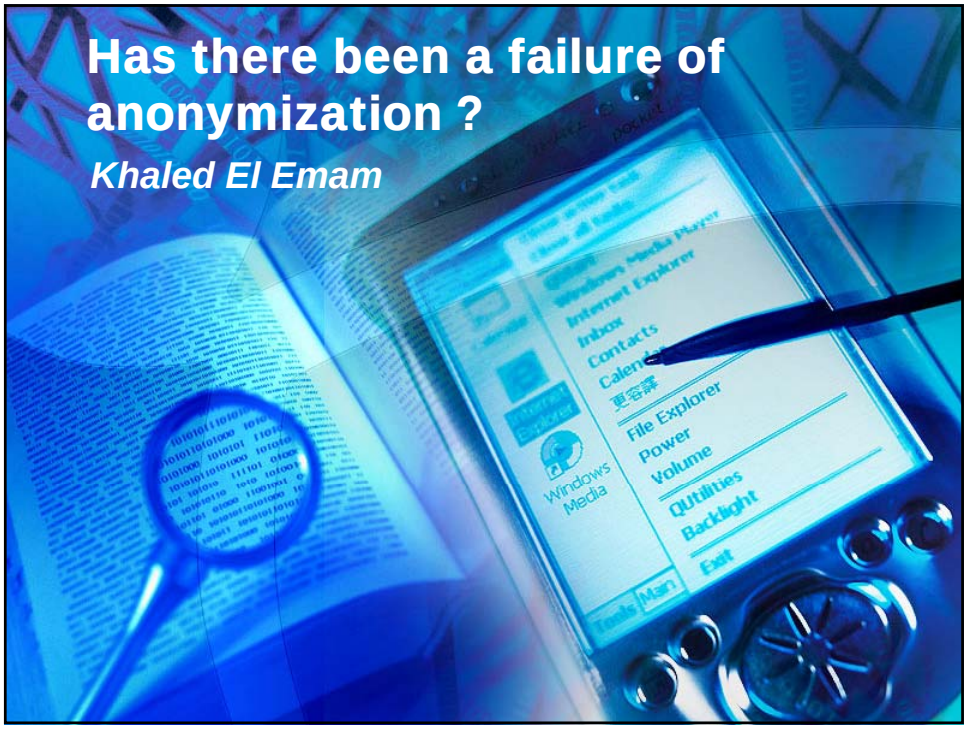
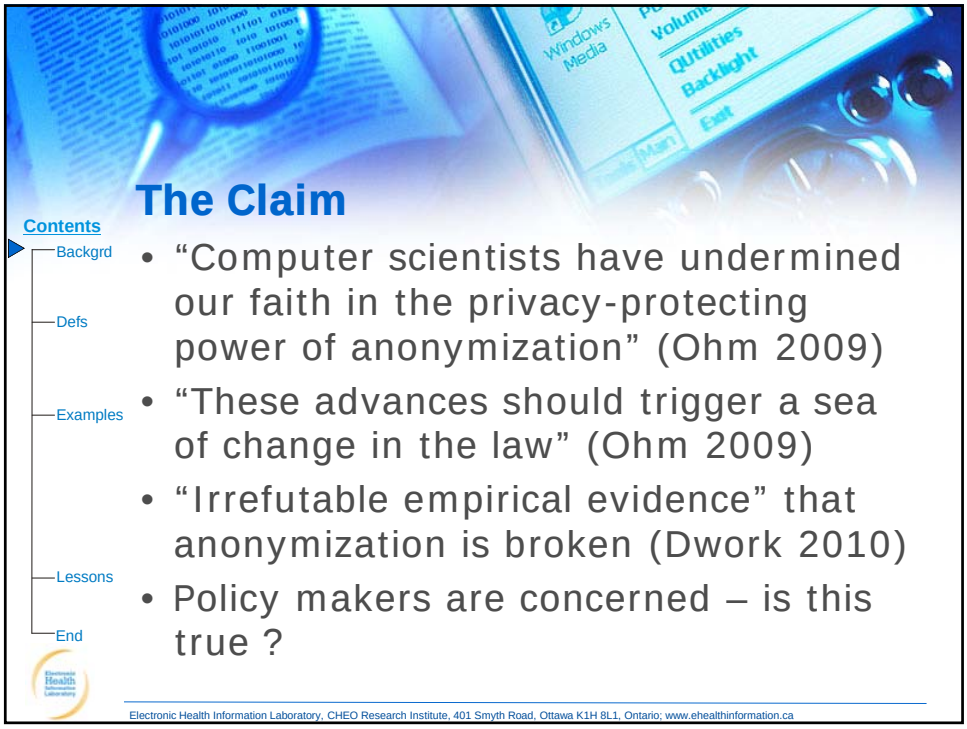
Has there been a failure of anonymization ?

Khaled El Emam



www.ehealthinformation.ca

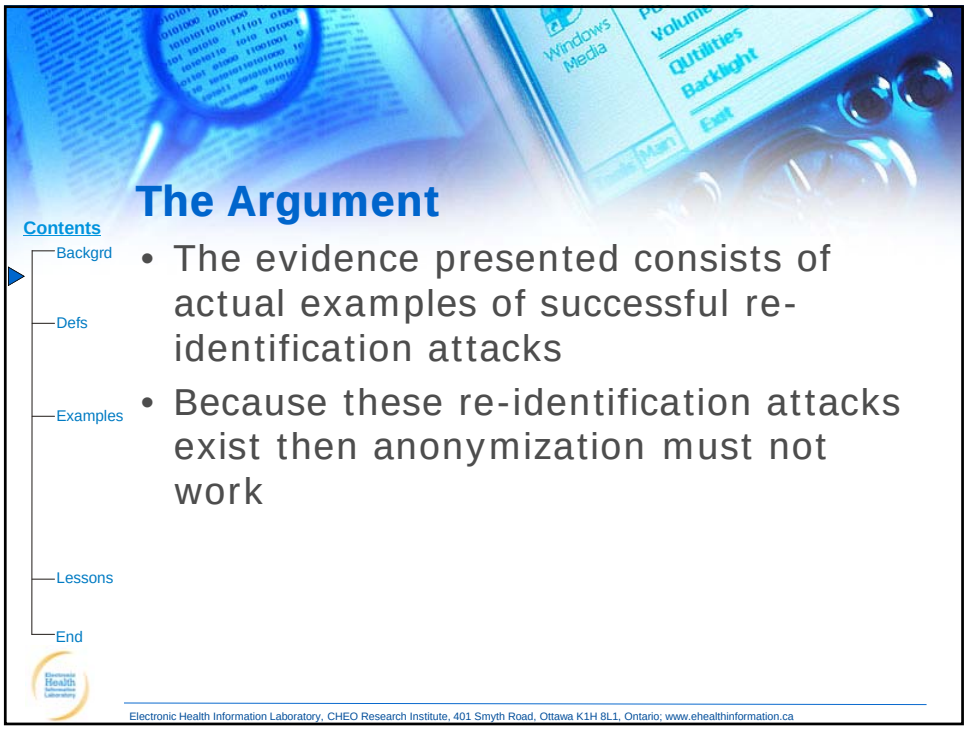
www.ehealthinformation.ca/knowledgebase



The Claim

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- "Computer scientists have undermined our faith in the privacy-protecting power of anonymization" (Ohm 2009)
- "These advances should trigger a sea of change in the law" (Ohm 2009)
- "Irrefutable empirical evidence" that anonymization is broken (Dwork 2010)
- Policy makers are concerned – is this true ?



The Argument

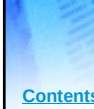
Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- The evidence presented consists of actual examples of successful re-identification attacks
- Because these re-identification attacks exist then anonymization must not work

The Problem with the Argument

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- The evidence often cited does not actually show that:
  - Re-identification was successful, or
  - That the data that was attacked was anonymized
- Therefore, there is actually no empirical evidence that anonymized data can be re-identified



This Presentation

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

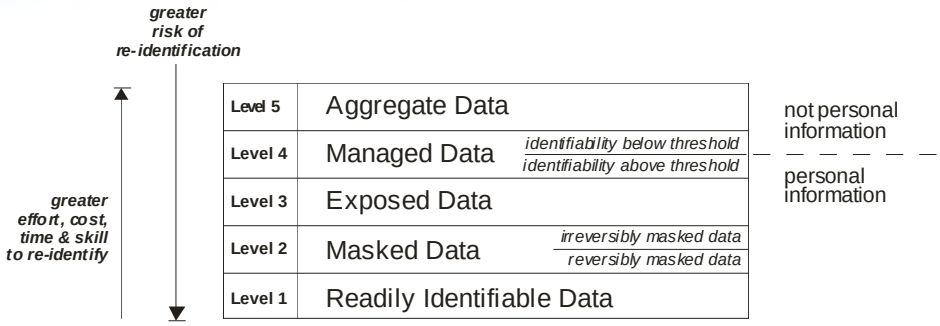
- We will examine 21 empirical studies that looked at re-identification on actual data sets or in real world examples (i.e., not theoretical attacks) to see what we can learn about them
- Focus mostly on demographics that can appear in clinical data (i.e., no genetic re-identification studies)
- Some subset of these 21 studies is often (mis-)cited as evidence

Variable Distinctions

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- Directly identifying
  - Can uniquely identify an individual by itself or in conjunction with other readily available information
- Quasi-identifiers
  - Can identify an individual by itself or in conjunction with other information
- Sensitive variables

Five Levels of Identifiability



Evaluation Criteria - I

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- Risk or re-identification:
  - Some studies evaluate (estimate or measure) the risk of re-identification but do not actually attempt to re-identify a data set
  - Therefore, risk evaluation studies would not be considered successful re-identification attacks

Evaluation Criteria - II

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- Was the data de-identified:
  - This would be a question whether the study was a risk assessment or a re-identification
  - Was the data properly de-identified in a measurable way or was it only “masked data” ?
  - Masked data is not anonymized data – re-identifying masked data is trivial

Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario; www.ehealthinformation.ca

Evaluation Criteria - III

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- Was a re-identification verified:
  - Verifying a re-identification match is necessary to ensure that the match was correct (i.e., confirming that the match of the record to the identity is correct)
  - If a match is not verified then it is not possible to know if the match was a correct one or not
  - The only exception would be a match to a population registry (e.g., voter list)

Example - I

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

Governor William Weld of MA

GIC Case

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- The Group Insurance Commission is responsible for purchasing health insurance for state employees in Massachusetts
- Insurance data on 135,000 state employees and their families was released after being “anonymized”
- Database was matched with the voter list for Cambridge, Massachusetts

William Weld

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

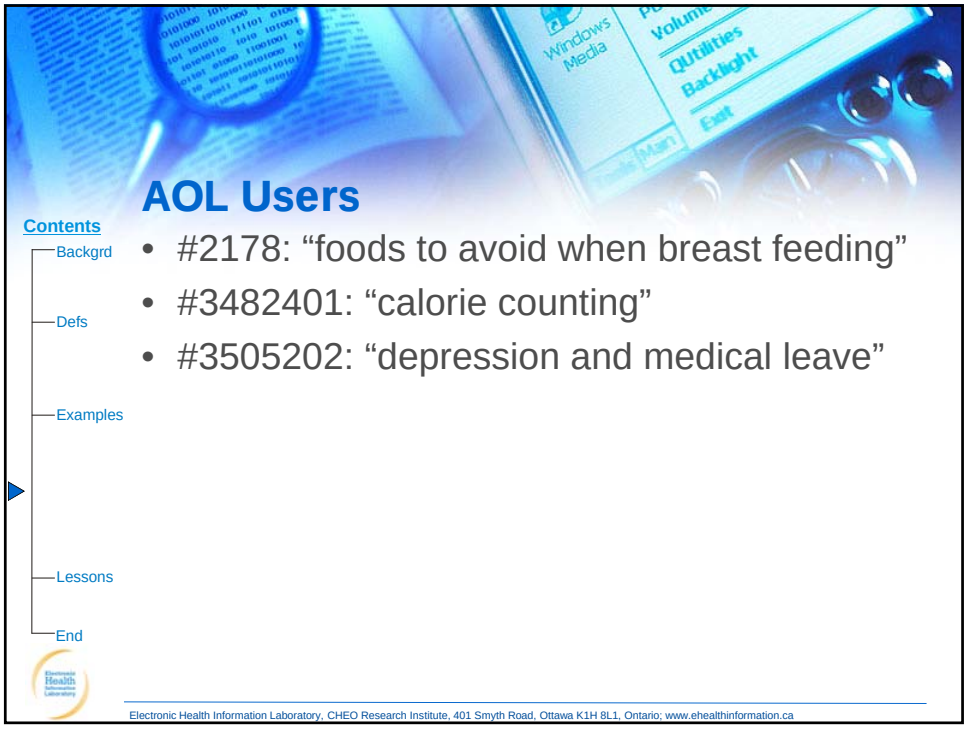
- Six people in the database have the same DoB
- Three are men
- One in his 5 digit zip code
- His insurance record was re-identified
- William Weld was the governor of Massachusetts

Evaluation of Example I

- This was a successful re-identification attack
- The match was not verified but it was a match to a population registry
- The data that was disclosed by GIC was Masked data and not properly de-identified

Example 2 - AOL Case

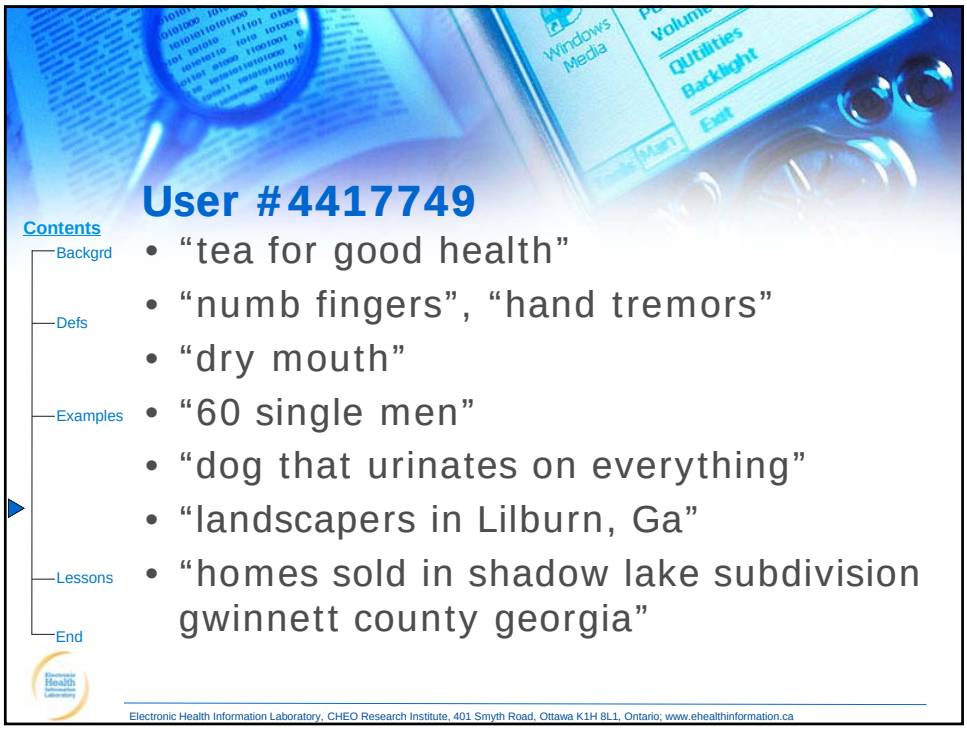
- In the Summer of 2006 AOL released “anonymized” data on ~20 million discrete search queries for >650,000 individuals on a public web site for researchers to use
- The records include date and time of the query and the web site clicked on, as well as a unique identifier for each user so records can be linked to get a user profile



AOL Users

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- #2178: “foods to avoid when breast feeding”
- #3482401: “calorie counting”
- #3505202: “depression and medical leave”



User #4417749

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- “tea for good health”
- “numb fingers”, “hand tremors”
- “dry mouth”
- “60 single men”
- “dog that urinates on everything”
- “landscapers in Lilburn, Ga”
- “homes sold in shadow lake subdivision gwinnett county georgia”

Thelma Arnold

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- 62 year old widow living in Lilburn Ga re-identified by the New York Times
- She has three dogs

Evaluation of Example 2

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- This was a successful re-identification attack
- The match was verified (we have a picture)
- The data that was disclosed by AOL was Masked data and not properly de-identified

Example 3 - Gordon Case

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- The CBC obtained the Adverse Drug Event database from Health Canada through an ATI request
- They identified a 26 year old female who died while on an acne drug
- They found a number of families through the obituaries and contacted them all until they found the family of the girl who died
- They did a story on the drug

Evaluation of Example 3

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- This was a successful re-identification attack
- The match was verified (they did a story)
- The data that was disclosed by Health Canada was Exposed data and not properly de-identified

Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario; www.ehealthinformation.ca

Estimates of Population Uniqueness

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- Sweeney and Golle, using different methodologies, estimated the percentage of the population of the US that is unique on their basic demographics (DoB, gender, and ZIP code)
- These studies assume that masked data would be disclosed
- No actual re-identification is performed in these studies

What Have We Learned ? - I

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- Just under half are risk assessment studies rather than actual re-identification studies (9/21)
- Risk assessment studies are useful but do not take into account practicalities (time, cost and skill needed to do an actual re-identification, and addressing data quality issues)

What Have We Learned ? - II

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- Of the remaining 12 actual re-identification attacks:
  - 1 case included identifiable data
  - 6 cases only removed direct identifiers (Masked data)
  - 1 case removed direct identifiers and attempted to obfuscate (Exposed Data)
  - 3 claimed to have performed unknown de-identification: “factually anonymous”, comply with “Privacy Act”, unknown perturbations
  - 1 case the data was de-identified but the identity of matches were not verified

What Have We Learned ? - III

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- All of the examples cited by the critics of anonymization are of the “Masked” variety (i.e., one of the six cases)
- All risk assessment studies conducted by researchers
- Of the re-identification attacks:
  - 8 by researchers, 1 in a hospital, 2 in courts, 1 by a journalist

Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario; www.ehealthinformation.ca

What Have We Learned ? - IV

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- All risk assessment and re-identification studies are identity disclosure; no actual successful re-identification has demonstrated attribute disclosure
- Of the re-identification attacks:
  - 2 in Canada, 1 in Germany, 1 in the UK, 8 in the US

What Can We Conclude ?

Contents
- Backgrd
- Defs
- Examples
- Lessons
- End

- Masked data can be re-identified
- No evidence exists whereby a properly de-identified data set (i.e., with a known risk level) has been re-identified
- Claims otherwise are quite misinformed or severely exaggerated
- But it is important to properly de-identify the data

www.ehealthinformation.ca

www.ehealthinformation.ca/knowledgebase