island health

Target Information Architecture (TIA) – to satisfy twin task demands:

1. Supply a working blueprint [architecture] & roadmap for Health Service Analytics Innovation

2. Set requirements for data de-identification framework and standard operating procedures that operationalize that framework

PART II
"Real-world" de-identification of transactional data extracted from electronic health records – breaking the curse of dimensionality

Adapted from: presentation to the 20th Annual Privacy and Security Conference (Reboot), February 7, 2019

July 16, 2019

Kenneth A. Moselle, PhD, R.Psych.
Director, Applied Clinical Research Unit, Island Health
Adjunct Assoc. Professor, University of Victoria

This presentation represents Part II of a set of documents concerned with two related challenges:

1. Part I - Target Information Architecture - derivation of products from health service data that are sufficiently statistically/methodologically robust and targeted that they at least warrant consideration as candidates for translation back to a health service system.
2. Part II – Data De-Identification - disclosure/access management of source health service data to those parties/team who are likely to possess the requisite combinations of clinical content domain knowledge and statistical/analytically expertise required to generate useful/usable products.

In effect, Part I sets out the requirements for Part II – the methodology covered in Part II must scale out to the types of datasets required to generate the products covered in Part I.

Streamlined, Privacy-Protected Access to Data – the Ultimate Quest and Grand Challenge

A body of person-level, real-world transactional (and other) health data, extracted from a jurisdictionally-heterogeneous array of clinical information systems, ***pre-authorized*** by multiple stewards for disclosure –under a well- specified set of conditions)

This is in the real world!

My kingdom for ***consensus*** on a data de-identification methodology that scales out to real-world high-dimensional health datasets

- A Thorny Problem - when data linkage, ***de-identificatio***n and access are centrally administered, how can this distributed array of source-data stewards know the privacy risk profile of the ***linked data***? If they don’t know this – how can they sign off on a disclosure – assuming that “risk” and “legislative/regulatory compliance” has something to do with “data contents”. Might this slow down approval processes???
- Required –an explicitly articulated data disclosure privacy risk model, clear ***operational*** definitions of key constructs such as “i***dentifiable***” or “***anonymized***” or “***limiting disclosure***” – or “***de-identified***” or “***risk”*** – and a set of standard operating procedures keyed to that model.
- No model = no shared understanding or consensus at the level of SOPs.
- With these SOPs recapitulated at every point and level within the data ecology (a ‘fractal’ data access management architecture) we can implement distributed and ***proportionate data de-identification*** procedures.
- With proportionate data de-identification in place, we can then implement collective ***proportionate governance*** –calibrating level of oversight, review and data protection to risk.

Our method must be able to adjudicate among a variety of options

1. No de-identification – disclose with unique identifiers
2. Nominal de-identification – remove "obvious" identifiers.
3. *Ad hoc* rule-of-thumb approaches plus #2, e.g., coarsen Postal Codes and Dates of Birth.
4. Documented & validated heuristic approaches (e.g., Safe Harbor 18 categories of re-identification "risk carriers")
5. Statistical disclosure control (SDC)-based methods – e.g., k-anonymization.
6. Data simulation approaches (?? how far can these go??).
7. No disclosure, even if judged to be in the public good.

For related work (with details) see: El Emam, K. & Hassan, W. (2013) The De-identification Maturity Model. Privacy Analytics, Inc. http://waelhassan.com/wp-content/uploads/2013/06/DMM-Khaled-El-Emam-Wael-Hassan.pdf

How about Option #2? We'll just jettison the primary care data, invoke provisions that allow disclosures of personal information for research purposes, and get on with it.

35 (1)A public body may disclose personal information in its custody or under its control for a research purpose, including statistical research [subject to a specified set of conditions that include approval from the head of the public body…] (from BC Freedom of Information and Protection of Privacy Act, current as of Jan 2, 2019).

- **Question**: Why don't we just invoke 35(1) and implement robust technical controls?
- **Answer #1:** General limiting principles in privacy laws/codes; "unreasonable invasion of privacy"; recognition of "mosaic effect" associated with "seemingly innocuous" bits of [linkable] information:

  In the US HIPAA Privacy Rule, there is a similar "minimal necessary" requirement. Therefore, as a starting point, the application of such limiting principles or minimal necessary criteria requires that de-identification be considered for all collection, use, and disclosure of health information. (El Emam, Jonker, & Fineberg. The Case for De-Identifying Personal Health Information (January 18, 2011).
- **Answer #2**: Legislative compliance is not the same as due diligence.
- **Answer #3**: Free floating generalized data disclosure anxiety states (or traits) – GDDAS – not in ICD9/10 or DSM-V)

NOTE: GDDAS is often secondary to absence of organizational standard operating procedures ***and documentation of methods and justification of those methods*** for meeting "limiting disclosure" requirements (see e.g. CFR164.514 or GDPR re: requirements around documenting procedures).

How about Option #5 – we'll just de-identify using classic k-anonymization

- k-anonymization – the industry-standard, most basic tool for implementing the US Privacy Rule [statistical] expert determination method for data de-identification – for "limiting disclosure" in a methodologically transparent fashion that translates across data disclosure scenarios.
- It works by rendering cases INDISTINGUISHABLE on the basis of any information available in the world that could be used for re-identification purposes – compress dimensions in space so that distinguished cases are re-located to essentially exactly the SAME place in data space.
- Sounds great. What's the problem?

Cavoukian and Castro concede that de-identification is inadequate for high dimensional data. But nowadays most interesting datasets are high-dimensional. High-dimensional data consists of numerous data points about each individual, enough that every individual's record is likely to be unique, and not even similar to other records. Cavoukian and Castro admit that: "In the case of high-dimensional data, additional arrangements may need to be pursued, such as making the data available to researchers only under tightly restricted legal agreements." Of course, restrictive legal agreements are not a form of de-identification. Rather, they are a necessary protection in cases where de-identification is unavailable or is likely to fail.

Arvind Narayanan & Ed Felten, No Silver Bullet: De-Identification Still *Doesn't Work,* July 9, 2014. But see also El Emam (numerous) for a response.

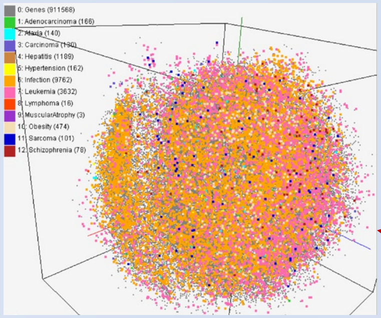
You can’t get there from here – how high dimensionality breaks k-anonymization

Everybody distinguishable – in practically uncountably many different ways

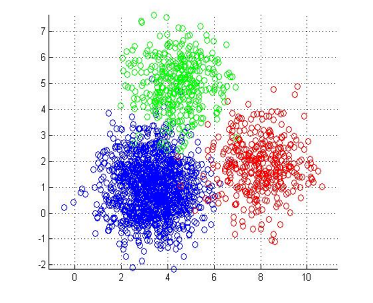
Inherently Problematic

Not so Problematic

People starting to “clump” together into groups where members are similar – and more privacy-protected.



The more similar people are in a dataset, the more difficult it is to re-identify anybody with any “reasonable” degree of confidence.
You want to be able to get from this state to that state,
while preserving the fitness of the data.

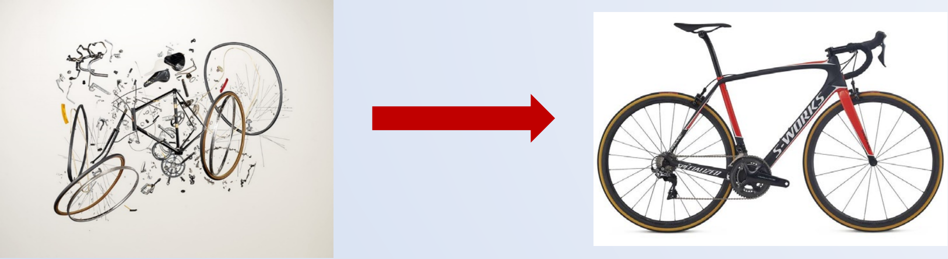


So sorry! Distinguishability is another name for “distance”. k-anonymization measures distance in the quintessentially most simple way – exact correspondence on one or more risk-carrying attributes. But being able to measuring distance (i.e., similarity vs difference) between people is one of the most basic curse(s) of High Dimensional Data – and essentially everybody in a high dimensional dataset is likely to be as different as their fingerprints.

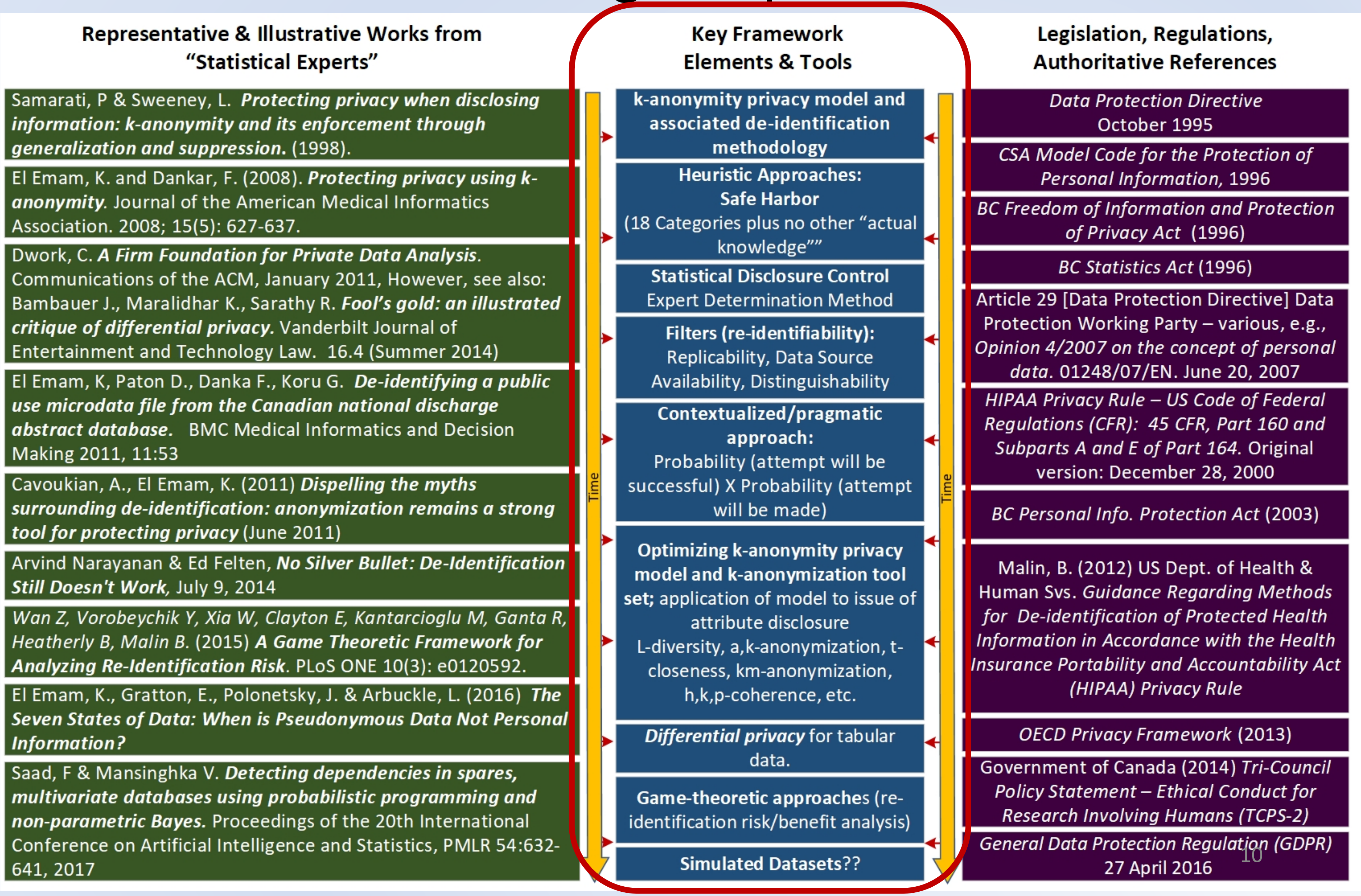
You can inject statistic noise into the dataset. That does nothing about distinguishability, but it does obscure the relationship between the data and reality - that detracts from the fitness of the data. And, the amount of noise you have to inject increases multiplicatively as you add dimensions. So this is not a solution, at least not for research with a real-world applied focus.

Meta-k-anonymization

- Risk is related to how distinguishable people are in a dataset.
- Distinguishability is related to how “far apart” people are in the data ‘space’ created by the variables in the dataset
- Unfortunately, you cannot even meaningfully measure “distance” in these high dimensional spaces, particularly when the bits and pieces in this space are very different (e.g., one acute care admission of 178 days is not the same as one visit to the cast clinic?
- You have only one option: find some ***reasonably*** defensible way of dramatically reducing dimensionality and marking off vast portions of this space as “irrelevant” to the disclosure at hand.
- ***Then*** measure risk – within the context of a method that also factors in context, and supplies tools for at least coarsely operationalizing elusive constructs such as “reasonably likely” means for re-identifying data.
- Use this same method to supply a reference standard for key constructs such as “limiting disclosure” or “low risk”.
- Can we assemble that method out of “off-the-shelf” components?



Yes - nothing new under the sun – assembling the framework from existing components



Representative & Illustrative Works from "Statistical Experts"

- Samarati, P & Sweeney, L. ***Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.*** (1998).
- El Emam, K. and Dankar, F. (2008). ***Protecting privacy using k-anonymity***. Journal of the American Medical Informatics Association. 2008; 15(5): 627-637.
- Dwork, C. ***A Firm Foundation for Private Data Analysis***. Communications of the ACM, January 2011, However, see also: Bambauer J., Maralidhar K., Sarathy R. ***Fool's gold: an illustrated critique of differential privacy.*** Vanderbilt Journal of Entertainment and Technology Law. 16.4 (Summer 2014)
- El Emam, K, Paton D., Danka F., Koru G. ***De-identifying a public use microdata file from the Canadian national discharge abstract database.*** BMC Medical Informatics and Decision Making 2011, 11:53
- Cavoukian, A., El Emam, K. (2011) ***Dispelling the myths surrounding de-identification: anonymization remains a strong tool for protecting privacy*** (June 2011)
- Arvind Narayanan & Ed Felten, ***No Silver Bullet: De-Identification Still Doesn't Work***, July 9, 2014
- *Wan Z, Vorobeychik Y, Xia W, Clayton E, Kantarcioglu M, Ganta R, Heatherly B, Malin B.* (2015) ***A Game Theoretic Framework for Analyzing Re-Identification Risk.*** PLoS ONE 10(3): e0120592.
- El Emam, K., Gratton, E., Polonetsky, J. & Arbuckle, L. (2016) ***The Seven States of Data: When is Pseudonymous Data Not Personal Information?***
- Saad, F & Mansinghka V. ***Detecting dependencies in spares, multivariate databases using probabilistic programming and non-parametric Bayes.*** Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54:632-641, 2017

Key Framework Elements & Tools

- **k-anonymity privacy model and associated de-identification methodology**
- **Heuristic Approaches: Safe Harbor** (18 Categories plus no other "actual knowledge"")
- **Statistical Disclosure Control** Expert Determination Method
- **Filters (re-identifiability):** Replicability, Data Source Availability, Distinguishability
- **Contextualized/pragmatic approach:** Probability (attempt will be successful) X Probability (attempt will be made)
- **Optimizing k-anonymity privacy model and k-anonymization tool set;** application of model to issue of attribute disclosure L-diversity, a,k-anonymization, t-closeness, km-anonymization, h,k,p-coherence, etc.
- ***Differential privacy*** for tabular data.
- **Game-theoretic approaches** (re-identification risk/benefit analysis)
- **Simulated Datasets??**

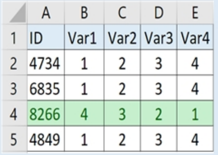
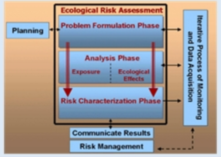
Legislation, Regulations, Authoritative References

- *Data Protection Directive* October 1995
- *CSA Model Code for the Protection of Personal Information, 1996*
- *BC Freedom of Information and Protection of Privacy Act* (1996)
- *BC Statistics Act* (1996)
- Article 29 [Data Protection Directive] Data Protection Working Party – various, e.g., *Opinion 4/2007 on the concept of personal data*. 01248/07/EN. June 20, 2007
- *HIPAA Privacy Rule – US Code of Federal Regulations (CFR): 45 CFR, Part 160 and Subparts A and E of Part 164*. Original version: December 28, 2000
- *BC Personal Info. Protection Act* (2003)
- Malin, B. (2012) US Dept. of Health & Human Svs. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*
- *OECD Privacy Framework* (2013)
- Government of Canada (2014) *Tri-Council Policy Statement – Ethical Conduct for Research Involving Humans (TCPS-2)*
- *General Data Protection Regulation (GDPR)* 27 April 2016

A model assembled from those (and a few other) features

Data De-Identification – In Four Movements

The Four Components of Data Disclosure Privacy Risk

	Existential Identifiability of People and their Attributes	**Mathematical Distinguishability** of Cases in a Dataset	**Theoretical Re-identifiability** of Cases in the World	**Pragmatic/ Contextualized Risk for Re-Identification** of People's Data
Components	The Natural Person(s) (the "Person", the "Population", the "Cohort") in the World → 	The Data in the Dataset (the "Data") → 	The Dataset in the World (the "World") → DATA PER DAY PROCESSED BY GOOGLE THE WORLD OF DATA 24 PETABYTES TWEETS PER DAY 50 MILLION	Data and Data Users in Data Disclosure Environments (the "Context")  Ecological Risk Assessment Planning Problem Formulation Phase Analysis Phase Exposure Ecological Effects Risk Characterization Phase Iterative Process of Monitoring and Data Acquisition Communicate Results Risk Management
The Ecology of Data Re-Identification	**Distinguishability of Persons in the World** • Unique Identifiers associated with the Person • Between/beyond the Person and Unique Identifiers – knowable & distinguishing attributes of Person/Population • Affiliation with what Communities of Interest	**Distinguishing characteristics of data in the dataset** • Ignoring everything but the dataset – what is the ***mathematically quantifiable*** data re-identification worst-case scenario? • 'Baked in' assumptions about the World. • Data contents identify Communities of Interest.	**Existence of distinguishing *plus* identifying data in the World** • Is it ***theoretically possible*** to reconnect a distinct Person in the World to a distinguished case in the Dataset? • Is it possible to make a statistically "educated" guess about level of theoretical risk?	**The Data User and Data in an Ecological Context:** • What actions are feasible? • Is risk-actualization ***reasonbly likely***? • What if assumptions about the person or the context are wrong? • Differential sensitivity of different data contents – potential harm.
Quantifiables	**Population prevalence** Prevalence of distinguishing characteristics – how many People share the same profile of distinguishing characteristics?	**Distinguishable cases** • Range of metrics based on numbers of cases with the same profiles on multiple linked variables in Dataset. • "Zero risk" and "disinguishable cases" do not go together.	**Computed estimates of re-identifiability probabilities** • Statistical estimates of *theoretical* risk for re-identification • Probabilities conditioned on assumptions about ***existence/linkability*** of Data in the World	**Game-theoretic estimates of risk** • How many plausible scenarios can be generated where benefits outweigh the risks? • **"Zero risk" is meaningful** in this measurement paradigm.
Caveats – 'Hard' vs 'Soft' Metrics	**Coarse** Estimates • Sampling will effect estimates; 'poor' sampling yields 'poor' estimates • Case definitions for population estimates may not match cohort definitions in dataset. • Out-of-date estimates • Estimates may not exist at all	**'Hard' Computed Values** • 'Hard' values under a set of worst-case assumptions that may or may not be reasonable • Greatest risk for inflated estimates of risk – greatest potential for putting data integrity at risk	**'Soft' 'but seemingly 'hard' estimates of re-identifiability** • Probabilities conditioned on difficult-to-validate assumptions about Data in the World • Quantifiables require assumptions about what Data User could possibly know.	**Taking the measure of the World** • Metrics conditioned on assumptions about predictability of behaviour and robustness of technical controls. • Risk/cost vs benefit not yet well-recognized in world of 'hard' data de-identification (statistical disclosure control)

	A	B	C	D	E
1	ID	Var1	Var2	Var3	Var4
2	4734	1	2	3	4
3	6835	1	2	3	4
4	8266	4	3	2	1
5	4849	1	2	3	4

Four components that collectively specify the risk profile of a candidate data disclosure

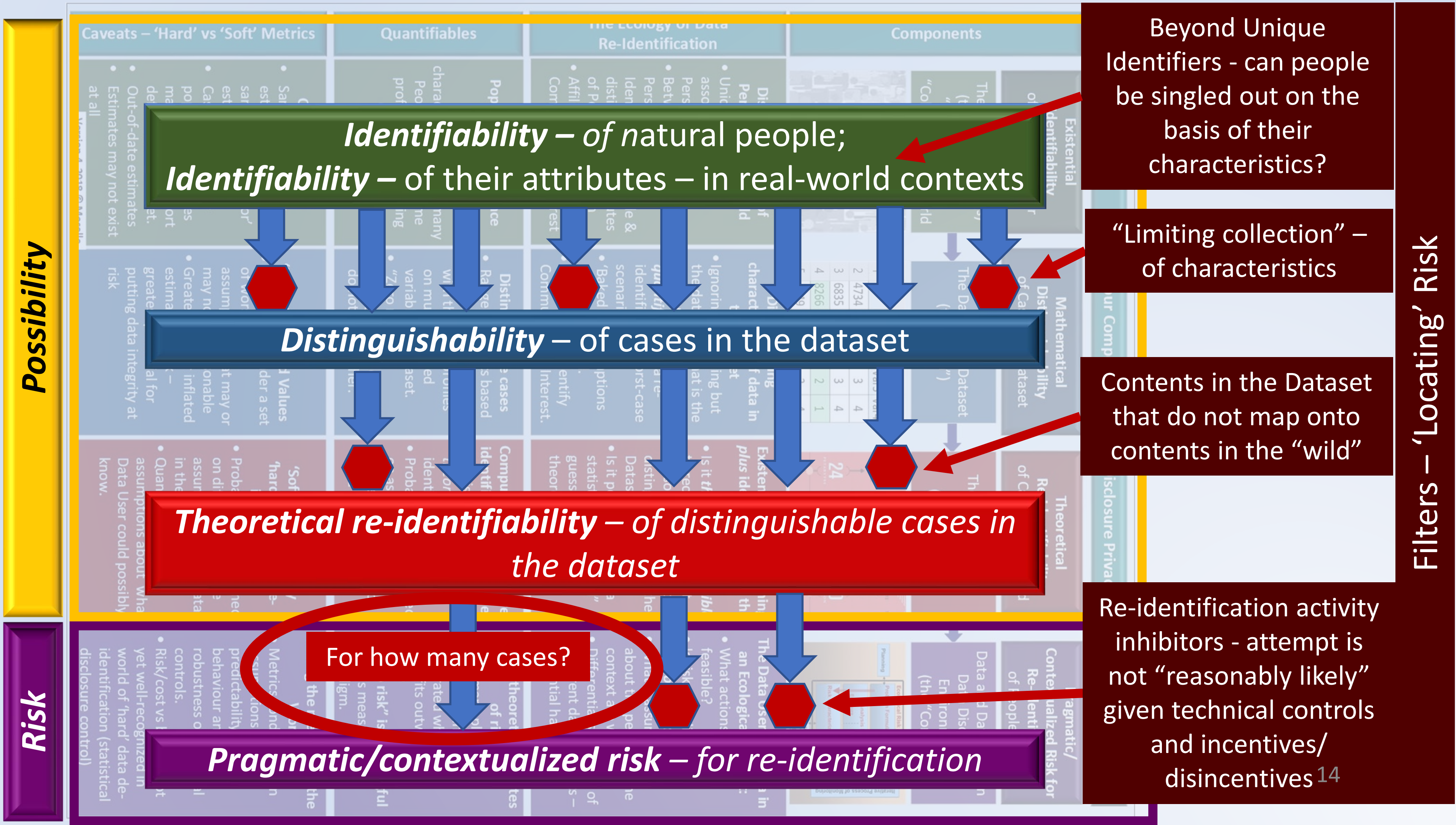
- **Component #1 - People in the world** – with attributes that need to be preserved (e.g., response to treatment) in the data as disclosed, while preserving the privacy of the people associated with those attributes.
- **Component #2 – Mathematical distinguishability of cases** in the Dataset – without which there is no privacy risk associated with the disclosure – the data '"space".
- C**omponent #3 – Dataset in "data space" meets data in the "real-world"** – from "distinguishability" to "theoretical re-identifiability".
- **Component #4 – Logistical/pragmatic features of the disclosure** – how feasible and likely is it that someone will perform the actions required to transform theoretically re-identifiable contents into re-identified contents?

The basic logic of the model and method:

(a) contextualized [lowest-reasonable] assessment of re-identifiability;

(b) dimensional reduction based on multiple methods, including ‘hard’ (statistical) and ‘soft’ (contextual) assessment;

(c) de-identify data based on this complete contextual assessment.



Possibility
Risk
Identifiability – *of natural people;*
Identifiability – of their attributes – in real-world contexts
Distinguishability – of cases in the dataset
Theoretical re-identifiability – *of distinguishable cases in the dataset*
For how many cases?
Pragmatic/contextualized risk – *for re-identification*
Beyond Unique Identifiers - can people be singled out on the basis of their characteristics?
“Limiting collection” – of characteristics
Contents in the Dataset that do not map onto contents in the “wild”
Re-identification activity inhibitors - attempt is not “reasonably likely” given technical controls and incentives/ disincentives
Filters – ‘Locating’ Risk

Unpacking the model

A bunch of indistinguishable things
(no risk for re-identification without a microscope)

People in the World – Possessing Attributes

Idiosyncratic [uniquely distinguishing] features of Natural Person A

Idiosyncratic [uniquely distinguishing] features of Natural Person B

Idiosyncratic [uniquely distinguishing] features of Natural Person C

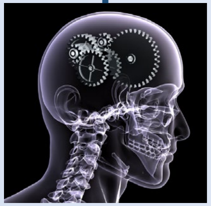
Various real-world contexts (potential data re-identification “seemingly innocuous” “slicers and dicers”

Cohort Characteristics (features shared by Natural Persons A, B and C)

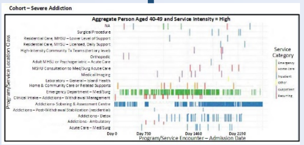
Distal (non-medical) determinants of health profile

Proximal determinants of health (health risk behaviours)

Treatment response



Clinical/functional status over time



Transactions with Health Service System

Idiosyncratic Features – We can mask these in the dataset.

Idiosyncratic

Truly Unique (and of necessity meaningless) Identifier d8f8s2j

Person A

Idiosyncratic

Truly Unique (and of necessity meaningless) Identifier g9u6q3

Person B

Idiosyncratic

Truly Unique (and of necessity meaningless) Identifier g9u6q3

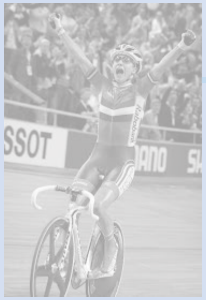
Person C

Various real-world contexts (potential data re-identification "seemingly innocuous" "slicers and dicers"

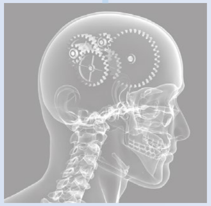
Cohort Characteristics (features shared by Natural Persons A, B and C)

Distal (non-medical) determinants of health profile

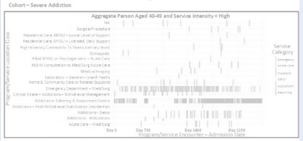
Proximal determinants of health (health risk behaviours)



Treatment response



Clinical/functional status over time



Transactions with Health Service System

Cohort Characteristics-of-Analytical Interest. We need to preserve these.

Idiosyncratic [uniquely distinguishing] features of Natural Person A

Idiosyncratic [uniquely distinguishing] features of Natural Person B

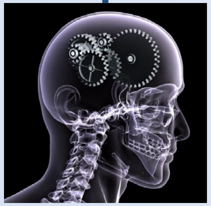
Idiosyncratic [uniquely distinguishing] features of Natural Person C

Various real-world contexts (potential data re-identification "seemingly innocuous" "slicers and dicers"

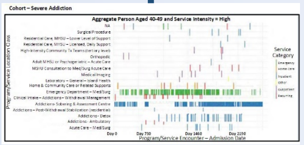
Cohort Characteristics (features shared by Natural Persons A, B and C)

Distal (non-medical) determinants of health profile

Treatment response



Clinical/functional status over time



Transactions with Health Service System

Proximal determinants of health (health risk behaviours)

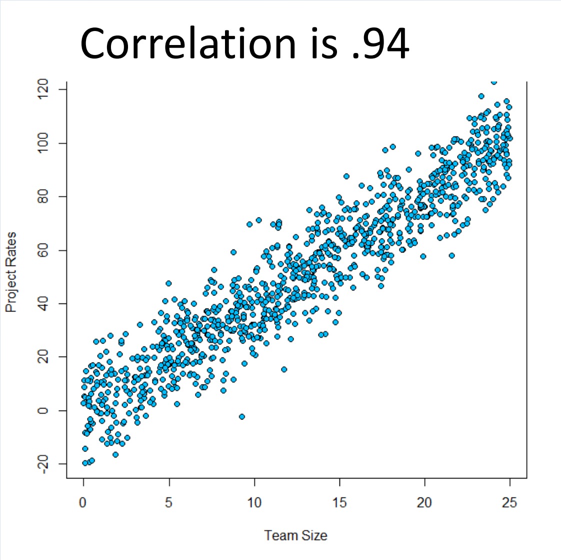
At the point that we disclose the data to the researcher or QA/QI analyst – we want people in the dataset to look ***sort of*** the same – but not really – and the “not really” should be the features of analytical interest and the location of the analytical “public good”.

The challenge –retaining the properties of “me” while masking “me” in the data

- The data in the clinical information system has all of these different types of information all linked to the patient/client identifier.
- We need to mask/suppress/delete the information that only functions to identify the patient/client.
- We need to keep ALL of the information that relates to the cause-effect relationships of interest.
- But SOME of that information may be distinguishing, and may therefore carry re-identification risk, even if some of that “risky” information is NOT “seemingly innocuous” – i.e., analytical integrity of the data depends on retaining that information.
- Sometimes “fit-for-analysis” or “fit-for-purpose” also translates into “fit for successful re-identification attempt”.

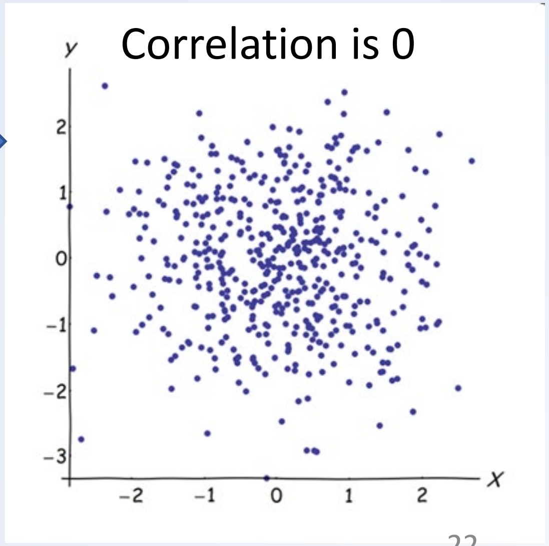
Dataset Fitness – Preserving Analytical Integrity of De-Identified Data

- Fit for analysis
- Fit for purpose-of-use (of analytical products
- Fit as in fit - capable of yielding findings that are as statistically robust or “buff” as the same findings that would be obtained if the data were analyzed in their pristine form.
- What we do NOT want is “fit for re-identification”


Correlation is .94

Before de-identification

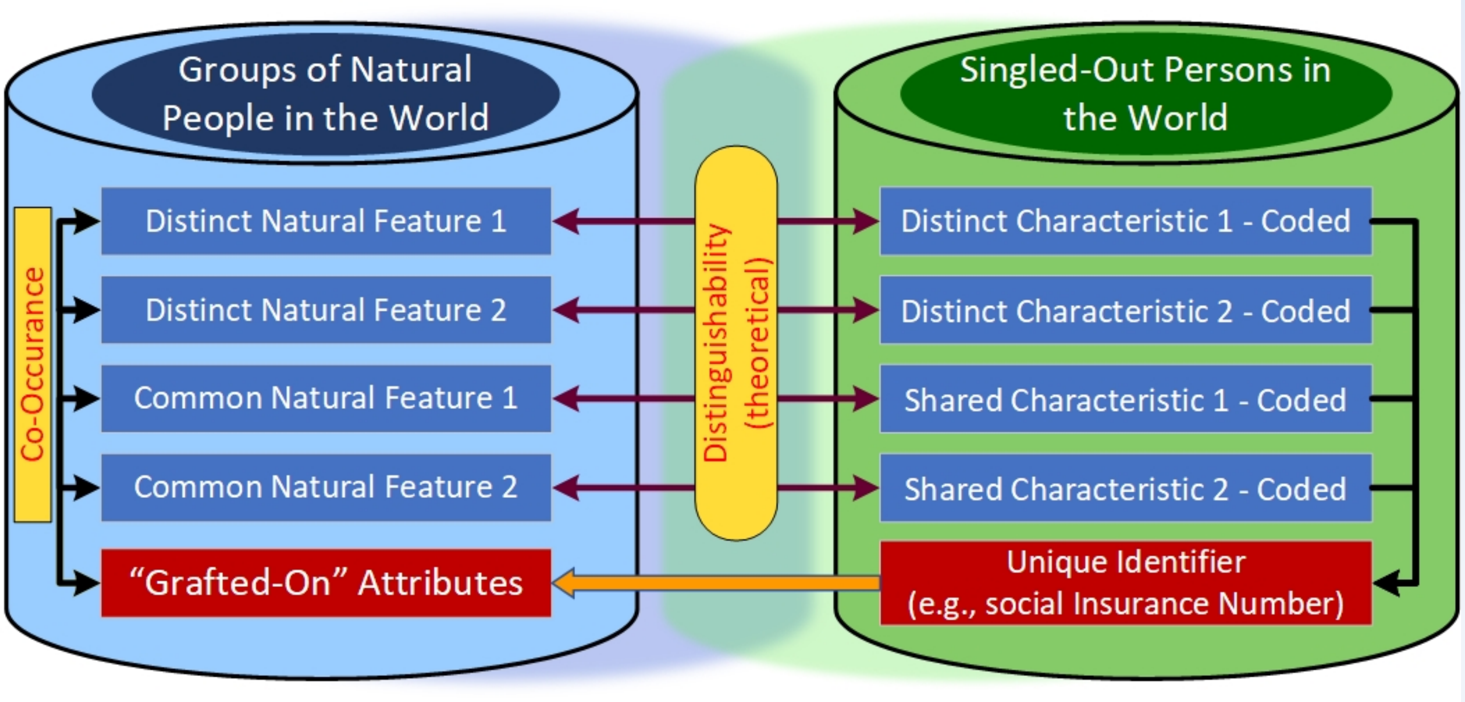
Very **unfit** de-identification process


Correlation is 0

After de-identification

Component #1 – Identifiability of people and their attributes

Component	Ecology of Data Re-Identification	Quantifiables	Caveats – 'Hard' vs 'Soft' Metrics
The Natural Person(s) •The "Person" •The Cohort(s) to which the person belongs •The Population from which the Cohort(s) are drawn	Distinguishability of Persons in the World •Unique Identifiers associated with the Person •Between/beyond the Person and Unique Identifiers – knowable & distinguishing attributes of Person/Population •Affiliation with what Communities of Interest	Population Prevalence Prevalence of distinguishing characteristics – how many People share the same profile of distinguishing characteristics?	Coarse Estimates •Sampling effects estimates •Case definitions for population estimates may not match cohort definitions in dataset. •Out-of-date estimates •Estimates may not exist at all

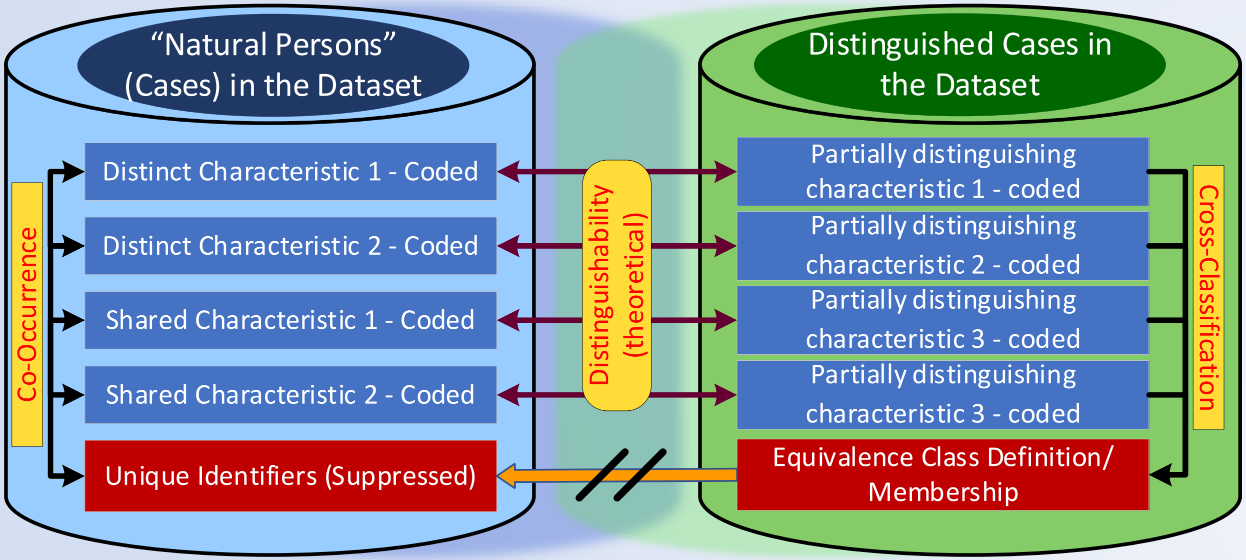

Groups of Natural People in the World
Co-Occurance
Distinct Natural Feature 1
Distinct Natural Feature 2
Common Natural Feature 1
Common Natural Feature 2
"Grafted-On" Attributes
Distinguishability (theoretical)
Singled-Out Persons in the World
Distinct Characteristic 1 - Coded
Distinct Characteristic 2 - Coded
Shared Characteristic 1 - Coded
Shared Characteristic 2 - Coded
Unique Identifier (e.g., social Insurance Number)

What emerges out of this first component – Natural People in the World?

- Estimates of prevalence of attributes within a cluster of people (e.g. a population or cohort).
- Statistical properties of distinct attributes and relationships (ideally cause-effect relationships) between attributes – these must be preserved in the data post de-identification.
- Differential sensitivity of different data contents (e.g., ICD9 303.01 “Acute alcoholic intoxication in alcoholism, continuous” vs 845.00 “Sprain of ankle”)
- Real-world contextual factors (e.g. lives in small community) and other factors (e.g. visibly young or old) that contribute to re-identification risks.
- We can think of these contextual factors as re-identification “slicers and dicers”.
    - Sometimes referred to as “seemingly innocuous” pieces of information that contribute to the “mosaic” effect.
    - They can “partition” definitely not-innocuous bodies of information (potential privacy invaders) in such a way that re-identification risk is increased – possibly substantially.

Component #2 – Distinguishability of cases in the dataset

Component	Ecology of Data Re-Identification	Quantifiables	Caveats – 'Hard' vs 'Soft' Metrics
The Dataset (the "Data") If a dataset has meaningful and useful content, it must reflect distinguishing characteristics of persons in the World.	**Distinguishing characteristics of data in the Dataset** • Ignoring everything but the dataset – what is the mathematically quantifiable data re-identification worst-case scenario? • 'Distinguishability' of cases reflects strong assumptions about data in the World • Data contents identify Communities of Interest.	Distinguishable Cases • Range of metrics - based on numbers of cases with the same profiles on multiple linked variables in Dataset. • "Zero risk" = no information content; "meaningful content = distinguishable cases"; therefore **"zero risk" for a dataset with meaningful content makes no sense.**	**"Hard" Computed Values** • 'Hard' values under a set of worst-case assumptions that may or may not be reasonable • Greatest risk for inflated estimates of risk – greatest potential for putting data integrity at risk



"Natural Persons" (Cases) in the Dataset
Co-Occurrence
Distinct Characteristic 1 - Coded
Distinct Characteristic 2 - Coded
Shared Characteristic 1 - Coded
Shared Characteristic 2 - Coded
Unique Identifiers (Suppressed)
Distinguishability (theoretical)
Distinguished Cases in the Dataset
Partially distinguishing characteristic 1 - coded
Partially distinguishing characteristic 2 - coded
Partially distinguishing characteristic 3 - coded
Partially distinguishing characteristic 3 - coded
Equivalence Class Definition/ Membership
Cross-Classification

What emerges out of this second component – Distinguishability of Cases in the Dataset

- Computed estimates of number of cases that could (in theory) be re-identified – with quantitative estimate estimates of the certainty of re-identification.
- Measures of distinguishability provide MAXIMUM estimates of number or proportion of cases that could be re-identified at different levels of certainty – on the basis of worst-case assumptions that are almost invariably going to be counterfactual.
- Beware of arguments based on counterfactual premises*, e.g., "if wishes were horses then beggars would ride" or the"myth of the perfect population register"** or the "omniscient data adversary" assumption:

*…when considering link disclosure attackers, one has to define what external resources are available to them. As it happens in cryptography, the most recommendable option (in order to ensure privacy even in the **worst case**) is to **assume** that the attacker has obtained some information on **all original records** in T, and then he uses this information in order to infer links between protected records in T' and original records in T. [emphasis added]****

*Goodman, N. The problem of counterfactual conditionals. The Journal of Philosophy, vol 44, No 5, February 1947, pp. 113-128.

**Barth-Jones, D. The "Re-identification" of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risks and Privacy Protections, Then and Now. Pre-publication draft – working paper, June 18, 2012.

***J Herranz, J Nin, P Rodriguez & T Tassa. Revisiting distance-based record linkage for privacy-preserving release of statistical datasets. Data & Knowledge Engineering 100 (2015) 78-93

Conflating Distinguishability and Risk: Appealing Mathematical Precision, but Blind to Context

Hah! Is that a hard and pointy proactive audit? That should keep us safe. Why bother with other controls!

There are a LOT of cases, so if there WERE a breach, it would be REALLY bad.

Doesn't feel like a zero risk to me!

Feels to me like a registry of lab results. If such a thing were publicly available – with unencrypted unique identifiers – then that would make our data very unsafe. I am not prepared to deem this animal not-dangerous.

That feels kind of like Anthony Tockar's re-identification attack on the NY Taxicab dataset. All he had to do was view uncountably many photos of famous people on the net, and crack the hashing algorithm, and he re-identified Jessica Alba and Bradley Cooper. So what if someone did the same kind of thing with our dataset???

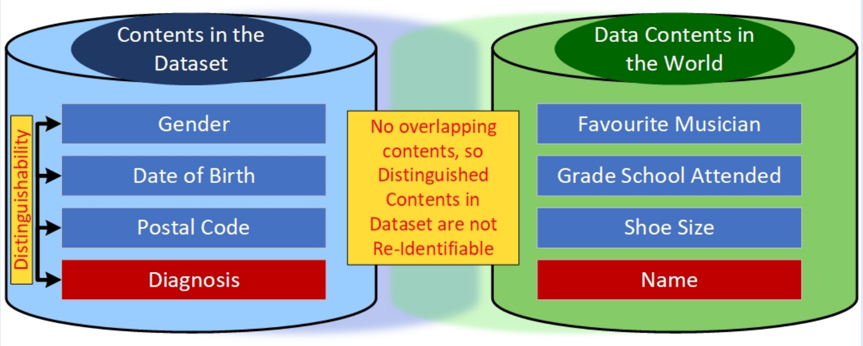
Well, that is clearly not one of the 18 unsafe parts of an elephant as per 45 CFR, Part 160 and Subparts A and E of Part 164 (HIPAA Privacy Rule), so I can safely assume we are Safe.

Yes, I know, there are 4,000,000 records, but this researcher's neighbour has high blood pressure and COULD be in the dataset, and the researcher ***could*** stumble across that record and uncover other information – or even search out the neighbour on purpose – so I think we have to coarsen the 14,000 ICD 9 diagnoses up to the level of 19 chapters – THEN we can disclose the data. No problem!

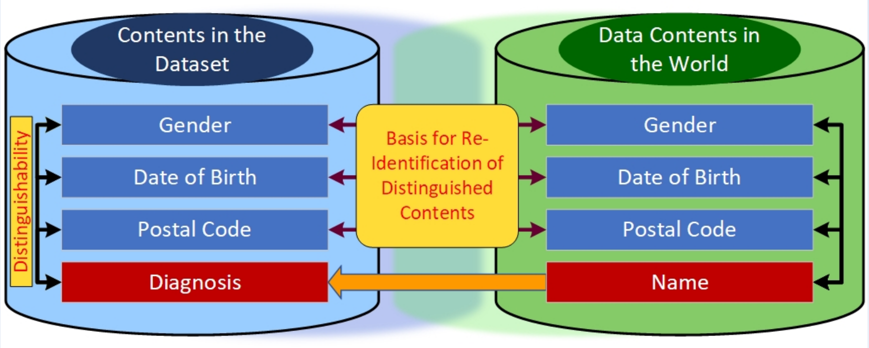
We'll just randomly swap some values on variables in the dataset – and Bob's your Uncle – according to the dataset – even though he really isn't. Isn't that the Differential Privacy way? What? You are studying heritability of Condition X? Oh, that could be a problem.

Component #3 – ***Re-identifiability*** of Cases

Component	Ecology of Data Re-Identification	Quantifiables	Caveats – 'Hard' vs 'Soft' Metrics
The World of Data Do data contents that distinguish cases in the nominally de-identified Dataset also exist in the world – together with identifiers for the distinguished cases?	**Existence of distinguishing *plus* identifying data in the World** • Is it ***theoretically possible*** to reconnect a distinct Person in the World to a distinguished case in the Dataset? • Is it possible to make a statistically "educated" and quantifiable guess about the level of theoretical risk – for distinguished cases or for distinguished groups of cases in the dataset?	**Computed estimates of re-identifiability probabilities** • Statistical estimates of theoretical risk for re-identification • Probabilities conditioned on assumptions about **existence/linkability** of Data in the World	**"'Soft' 'but seemingly 'hard' estimates of re-identifiability** • Probabilities conditioned on difficult-to-validate assumptions about Data in the World • Quantifiables require assumptions about what Data User could possibly know.



Dataset contents are distinguishable but not re-identifiable



Dataset contents are both distinguishable and re-identifiable

*In general terms, a natural person can be considered as "identified" when, within a group of persons, he or she is "distinguished" from all other members of the group. Accordingly, the natural person is "identifiable" when, although the person has not been identified yet, **it is possible to do it (that is the meaning of the suffix "-able")*** [emphasis added]**.**

Article 29 Data Protection Working Party, Opinion 4/2007 ***On the concept of personal data***. 01248/07/EN. June 20, 2007

Component #3 – Dataset meets the real world

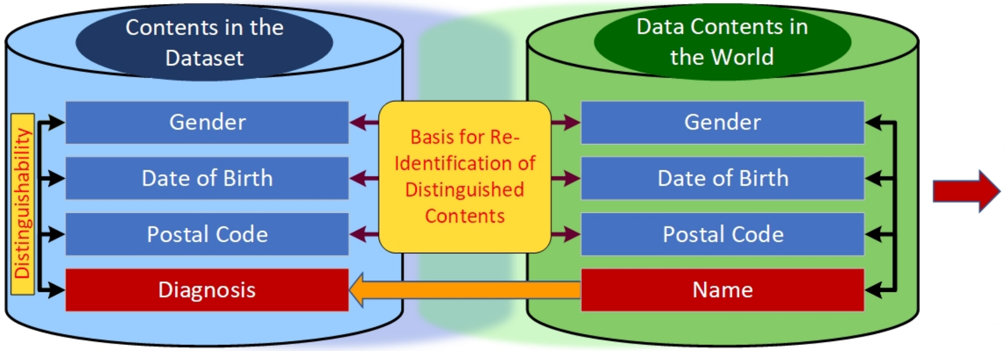
- Estimates of Distinguishability qualified by evaluation of contents in the “real world”.

- We are trying to operationalize the construct in the literature known as “adversary power”.
- We can do this quantitatively – and generate numbers that appear to be “hard”
- However – “hard” (estimates of Distinguishability) divided by “soft” (estimates of adversary power expressed numerically) may be quantitative- and appear precise- but they are not – since they may vary widely based on assumptions.
- Re-identifiability (of data) and “power” of “adversaries” are necessary conditions for the ***possibility*** of re-identification – they are only partially contextualized measures of risk.

Component #4 – Pragmatic/Contextualized ***Risk*** for Re-Identification

Component	Ecology of Data Re-Identification	Quantifiables	Caveats – ‘Hard’ vs ‘Soft’ Metrics
Data & Data Users in Data Disclosure Environments (the Context) • Is it logistically feasible for re-identifiable data contents to be re-identified by parties with authorized access to the data. • What are potential costs and benefits to a data user who attempts to re-identify the data?	**The Data User and Data in an Ecological Context** • What actions are feasible? • Is risk-actualization reasonbly likely? • What if assumptions about the person or the context are wrong? • Differential sensitivity of different data contents – potential harm.	**Game-theoretic estimates of risk** • How many plausible scenarios can be generated where benefits outweigh the risks? • “Zero risk” is meaningful in this measurement paradigm.	**Taking the Measure of the World** • Metrics conditioned on assumptions about predictability of behaviour and robustness of technical controls. • Risk/cost vs benefit not yet well-recognized in world of ‘hard’ data de-identification (statistical disclosure control)

From “Theoretically Possible and Logistically Not-Impossible” to “Real-World/Real-People Risk”



Technical controls against unauthorized access or unauthorized data import/export/ linkage – in the environment where the dataset is accessed and/or in the world

Legislation, regulations, penalties

Enforceable agreements

Re-identification attack - incentives/disincentives

Financial costs associated with re-identification attack

Technical skills required to execute re-identification attack

Typical data analytical workflows – related to accidental re-identification

Audit controls

Differential sensitivity of different data contents

↓

Risks for successful re-identification vs benefits of re-identification

↓

Contextualized *Risk*

Component #4 - Being "reasonable"

The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used ....The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person....[GDPR - ***General Data Protection Regulation****, 27 April 2016]*

"anonymization" is a de-identification process that removes or transforms all direct and indirect identifiers in a record for which there is a reasonable expectation that the identifiers could be used, either alone or with other information, to identify an individual - An anonymized record no longer contains personal information; therefore, the privacy protection provisions contained in Part 3 of the Freedom of Information and Protection of Privacy Act or other applicable legislation no longer apply.
[BC Ministry of Health – Policy – ***Access to Health Data for Research****, September 1, 2018]*

It is important to be realistic and consider plausible attacks*, especially when there are data use agreements that prohibit re-identification, linking to other data, and sharing without permission. Besides the standards which give direction on the selection of identifiers and precedents for acceptable levels of risk, an evaluation or re-identification risk can be limited to the amount of information that an adversary can realistically know (the "attacker's power"). [S. Garfinkel,* ***De-Identification of Personal Information****, National Institute of Standards & Technology, U.S. Dept. of Commerce, October 2015]*

Whatever this "being reasonable" thing is – **it is powerful enough to render data "anonymized"** and free of legislated/regulatory constraints associated with "identifiable" information – as per GDPR, Article 29 Data Protection Working Party (EU), BC Ministry of Health, Office of the Information Commissioner Queensland, others.

What do we get from Component #4

- What technical challenges would need to be addressed to access data in the world required to re-identify data in the dataset? What would it cost? How much technology and technical skills would be involved?
- What incentives/disincentives are at place?
- A "reasonable" method for calibrating assumptions about data availability in the world.
- A "reasonable" method for operationalizing key constructs about people, such as "adversary power", or "risk for doing X, Y or Z".
- An important change in measurement focus: from various measures related to proportion of cases that are distinguishable (k-anonymization or related metrics) to game-theoretic measures based on all possible combinations of risks and benefits associated with re-identification attempts.
- If there are no reasonably envisioned scenarios where benefits outweighs risks or costs, then from a game-theoretic vantage point – there is zero risk!*

*Wan, Zhiyu; Vorobeychik, Yevgeniy; Xia, Weiyi; Clayton E. Katarcioglu M and Malin B. Expanding Access to Large-Scale Genomic Data While Promoting Privacy: A Game Theoretic Approach The American Journal of Human Genetics, 02/2017, Volume 100, Issue 2 Published online 2017 Jan 5. doi: [10.1016/j.ajhg.2016.12.002]

Preliminary Take-Away Messages

- Possibility and risk are related but they are not the same.
- In the real world, in general, Identifiability will be greater than Distinguishability will be greater than Re-identifiability will be greater than Pragmatic/Contextualized Risk.
- Risk is a joint function of those FOUR characteristics.
- If you do not take all four into account, you will typically over-estimate risk (and over-protect/harm the data).
- Some (e.g., El Emam & associates) build all four into their privacy risk models. Most do not.
- The model provides a model for a model – an example of a data disclosure risk model – and a working vocabulary – for access adjudicators who must explain the basis for decisions.
- A data de-identification framework and methodology must supply tools for estimating the magnitude of key quantities (e.g., distinguishability) and the pragmatic likelihood (risk/benefit) of the actions required to translate the possibility of re-identification into re-identification.
- In reference to the four features of this model – when a person or policy or procedure uses the term “risk” in relationship to a data disclosure - to which (combination) of the four components does the term refer?
- You “calibrate” your data disclosure risk analysis and associated protections in relationship to each of the four components – four sets of activities, and four sets of documentation.

The final ascent – from privacy risk model (principles, guidelines) to standard operating procedures

- Full documentation of model, including references to legislation, regulations or policies; math/stat or computer science publications, official opinions or directives.
- Operational definitions of key constructs.

- Questions keyed to each of the four components.
- Methods for answering the questions.
- Templates for registering answers to the questions.
- Benchmarks or cutoffs for evaluating quantitative risk metrics (where applicable or illuminating).
- Scenario-based methodology – standard framework for characterizing data disclosures – to "stress test" candidate data access management models – including tests against the possibility that assumptions made in any given candidate data disclosure turn out not to be correct.
- While we are at it – a working ***target information architecture*** for health services – to supply a working answer to a basic question: "So what data are we talking about pre-authorizing??"

Kenneth A. Moselle, PhD, R.Psych.
Director, Applied Clinical Research Unit
Island Health
British Columbia
kenneth.moselle@viha.ca