EMPIRICAL ASSESSMENT OF PRIVACY RISKS IN DATA

JANICE BRANSON, NATHAN GOOD & KHALED EL EMAM

cc: highersights - https://www.flickr.com/photos/65339210@N02

Agenda

Time	Speaker	Topic
11:00 – 11:05	Khaled El Emam	Logistics & Introduction
11:05 – 11:15	Janice Branson	Business context • why is this an area relevant for a company like Novartis • what are the business reasons why motivated intruder tests in general are relevant
11:15 – 11:35	Khaled El Emam	Methodology • an overview of motivated intruder methodology - how it works • literature review
11:35 – 11:55	Nathan Good	Experiences • generalize over multiple experiences doing these tests • are social media big sources of information useful for attacks ? • what is hard and easy ? • what should we do and not do when de-personalizing data ?
11:55 – 12:00	Khaled El Emam	Q&A

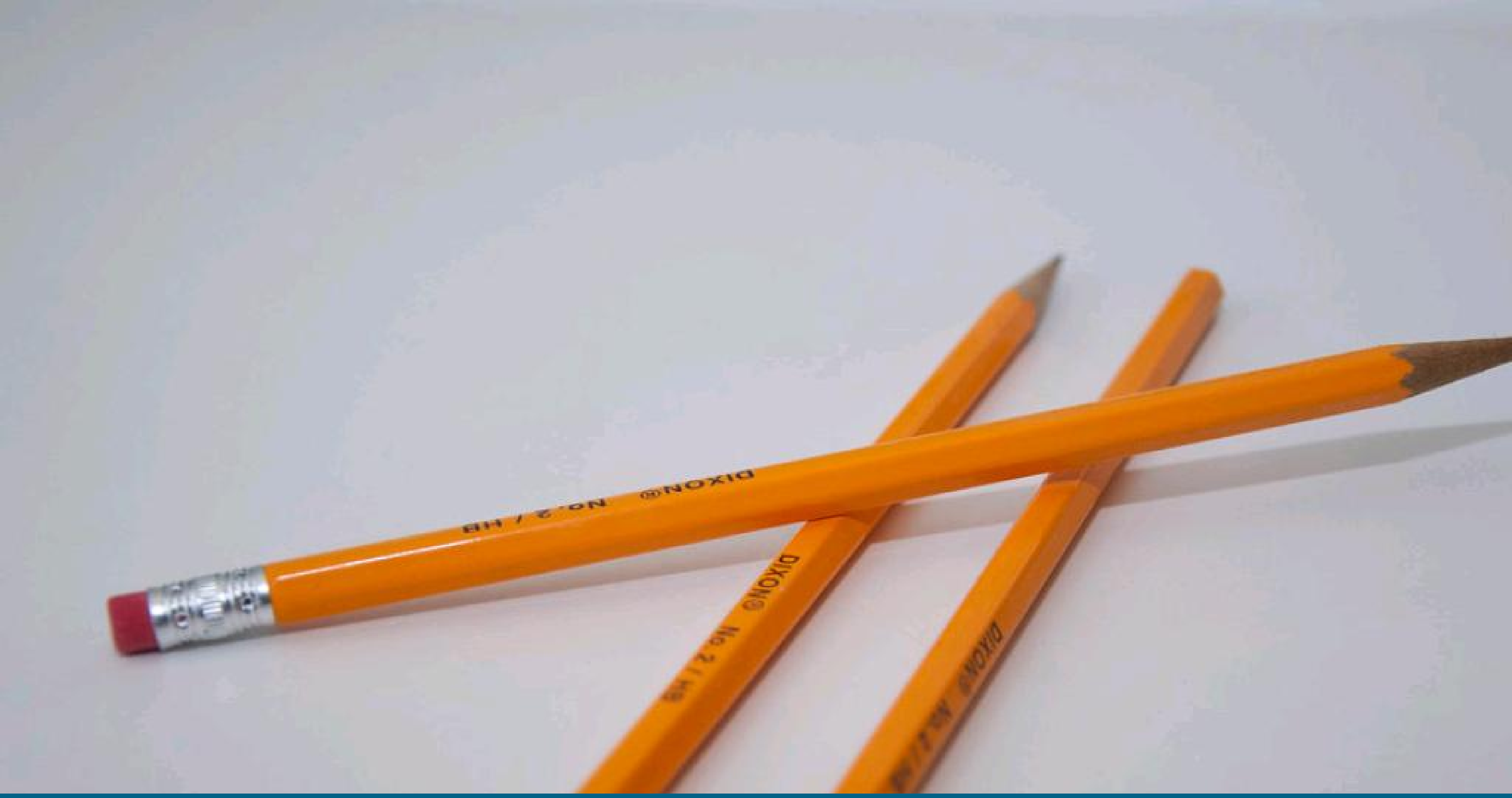
De-Personalized Data

- Two general ways to evaluate de-personalized data:
  1. Models to estimate the probability of matching a record with a real person
  2. Empirically through a motivated intruder test

cc: Kelly Sikkema - https://unsplash.com/@kellysikkema?utm_source=haikudeck&utm_medium=referral&utm_campaign=api-credit

Motivated Intruder Test

cc: David Pennington - https://unsplash.com/@dtpennington?utm_source=haikudeck&utm_medium=referral&utm_campaign=api-credit



1. Motivations
2. Methdology
3. Experiences

cc: taylor.a - https://www.flickr.com/photos/40944554@N04

Clinical Development &
Analytics

Motivated Intruder Attack – why is it relevant for Novartis?

NOVARTIS | Reimagining Medicine

Evolving era of data sharing – where we are today

	Voluntary data sharing PhRMA-EFPIA	EMA Policy 0070 (CDP) & Health Canada
What is shared?	Anonymized data sets + supporting documents for Phase 2-3 studies	Defined sections of CSRs and clinical summary documents
When can it be shared?	After submission is approved by EMA and FDA & trial results are published	After CHMP opinion (positive or negative) or sponsor withdrawal
With whom is it shared?	Researchers with legitimate analysis proposal who sign a terms of use agreement and publish analyses within 12 months post data analyses	Public (register and agree to terms of use). To have print or save access provide passport details and valid EU address
How is it accessed?	Through secure portal CSDR ClinicalStudyDataRequest.com	View via portal

Why was a Motivated Intruder Attack important for us?

- Novartis strives for a framework that
  - Covers all aspects of these 2 types of information and data sharing and
  - Has a standard and consistent approach which ensures that patient privacy is maintained
- EMA with CDP and then Health Canada require the public sharing of clinical trial reports
- Both agencies have provided guidance for the quantitative anonymization of these clinical reports before they are shared.
- Previously any sharing of information was through Access to Documents EMA Policy 0043 and in general all companies used redaction i.e. blacking out information thought to be identifiable of patients.
- Changing to anonymization rather then redaction coupled with the fact that under CDP these documents are made public then we as a company wanted to gather more empirical data on the effectiveness of anonymization in protecting patient privacy

Why was a Motivated Intruder Attack important for us?

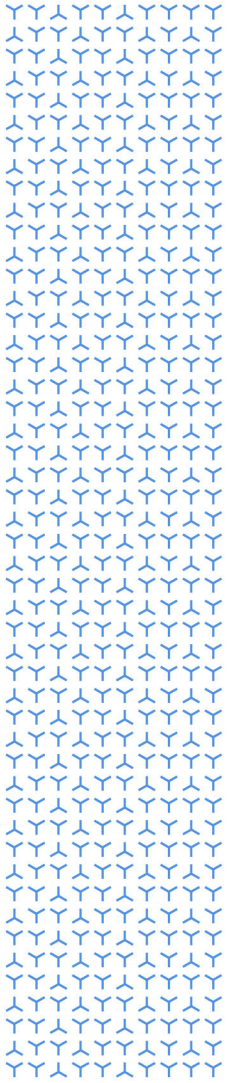
- We focus on risk based anonymization, taking into account the data sharing context and assessing the risk of re-identification
- We want to ensure the probability of re-identification that is computed during the anonymization process is indeed as low as assumed
- Re-identification risk calculations are based on statistical models, and these models make assumptions. The assumptions that we make tend to be conservative, which means that the true re-identification risk might be underestimated
- How can we gain confidence in the anonymization approach and the calculated probability of re-identifying someone? – This was needed for internal decision making in regards to how we implement the policies as well as ensuing data privacy for our patients

Our expected goals from the Motivated Intruder Attack

Ensures that Novartis has an updated understanding of the real risks in some data recipient environments

Provides additional data points to improve the anonymization practices for a particular data release or type of data release

Helps adjust the assumptions that have been made in the re-identification risk measurement (for example, what are the plausible direct and indirect identifiers that can be used in an attack)



Thank you

NOVARTIS | Reimagining Medicine

Motivated Intruder Tests

Methodology

25th March 2020

Replica Analytics

Background

- Many articles have been published examining the ability to correctly map a de-personalized record to a real person
- Important criteria to interpret them:
  - was the data pseudonymous ?
  - was this a statistical or empirical assessment ?
  - was the match rate measured on the sample or the population ?

(c) Copyright 2019-2020 Replica Analytics Ltd.

Principles

- Effort and cost are important in deciding whether a match is reasonably likely or not
- Code of conduct:
  - Ethical behavior / Misrepresentation
  - No criminal behavior
  - Informing the controller
- Questions (?):
  - Contact individuals and acquaintances

(c) Copyright 2019-2020 Replica Analytics Ltd.

The Process

(c) Copyright 2019-2020 Replica Analytics Ltd.

PLANNING | MATCHING | EVALUATING | REPORTING

- **Which dataset to evaluate ?**
- **When to evaluate ?**
- **Third party motivated intruder test**
- **External databases and costs**
- **Skills of the analysts**
- **Authority to identify records**
- **Ethical reviews**

(c) Copyright 2019-2020 Replica Analytics Ltd.

PLANNING | MATCHING | EVALUATING | REPORTING

- **Verification**
- **Caps on resources**
- **Levels of matching**
- **Learning something new**
- **Direction of attack**

(c) Copyright 2019-2020 Replica Analytics Ltd.

Replica Analytics

Budget Item	Number of Data Subjects	Number of Hours	Hourly Rate	Total
Base effort on population to sample attack				
Max effort on population to sample attack				
Base effort on sample to population attack				
Max effort on sample to population attack				
Commercial Databases				
Preparation effort: • Custom tools & scripts, pre-processing • Enhancing target dataset				
Report Writing				
Total				

(c) Copyright 2019-2020 Replica Analytics Ltd.

Replica Analytics

PLANNING | MATCHING | EVALUATING | REPORTING

Introduction

- This provides the main purpose of the report

Methods

- Description of the target dataset
- Parameters for the motivated intruder test
- Background and expertise of the attackers
- External sources of data examined
- Outline of the sample to population and population to sample attacks
- How famous people were identified
- Scripts developed
- Search methodologies
- Verification methodology, if any
- Special actions, e.g., advertising

Numeric Results and their interpretation

Conclusions

Appendices

- List of all individuals suspected or verified matched, including all of the variables that they were matched on and if anything new was learned

(c) Copyright 2019-2020 Replica Analytics Ltd.

Experiences in performing Motivated Intruder Analysis

good RESEARCH

Good Research

We are an qualified team of privacy professionals, with expertise in privacy consulting, user research, software engineering, data science, and technology ethics.

- We help build respectful and trusted relationships with customers by taking a proactive, holistic, and user centric approach to Privacy and Security.
- We have conducted **motivated intruder tests** for companies across multiple sectors including pharmaceuticals, manufacturing, and logistics.

Experiences in performing MIAs

Sources of Information for an MIA

1. Contextual data:
   - Clinical Reports, Hospital discharge records, death records
   - Data analysis
2. Social media
   - Facebook, twitter, etc.
   - Online forums, reddit, etc.
3. Purchasing general population datasets
   - Voter registration records
4. FOIAs
   - FDA, DOT, etc.
5. Using a Recruiter

Best-practices in anonymization

good RESEARCH

Sources of Information: **1. Contextual Data**

Data specific to the particular industry and domain.

This can include metadata of processes, related outcome data, or specific ways to process the information for garnering particular insights.

Examples: Clinical Reports; Hospital discharge records; Death records; Data analysis on the initial dataset

Pros:

- **Quantity:** With sufficient resources and time, it tends to yield the most results
- **Quality:** The results can be highly accurate
- **Generative:** Results lead to other results that can help initiate a recursive discovery of resources

Cons:

- **Costly**: needs a significant time investment, in some case the physical deployment resources to talk to people or visit locations
- **High barrier of entry**: more fruitful investigations need more domain knowledge

good RESEARCH

Sources of Information: 2. Social Media

Users may generate their own data that may help identify them in our target dataset. Depending of the dataset, different platforms will contain the specific traits from a user's online fingerprint useful for re-identification.

Examples: Facebook, Twitter and other social media platforms; Reddit and other online forums

Pros:

- **Low barrier of entry:** Social media platforms are easy to use and require no specific knowledge
- **Extendible and repeatable:** Use and build tools to perform analysis at scale and that could be reused for other MIAs

Cons:

- **Needle in a haystack:** Vast volume of data to sift through to find the specific relevant information
- **Confidence:** Given that this is a search on a sizable population, the confidence of correct identification tends to be lower

good RESEARCH

Sources of Information: 3. General Population Datasets

General population datasets can be purchased for population-to-sample attacks, and are one of the most common demographic enhancements attackers use.

Examples: Voter Registry List, Transactional data

Pros:

- **Reliable:** Usually considered the "ground-truth" of the actual population.
- **Demographic-rich:** This data usually comes with several demographic information from each subject

Cons:

- **Cost can escalate easily**: Most voter information data purchasing services have per-person pricing, which makes
- **Need for demographic data in the target**: If not there will be very little information to try to

good RESEARCH

Sources of Information: 4. FOIAs

Data specific to the particular industry and domain of the data. This can include from metadata of processes, related outcome data, or specific ways in which to process the information for garnering particular insights relevant to the problem at hand.

Examples: Clinical Reports, Hospital discharge records, Death records, Data analysis

Pros:

- **Repeatable:** A process can be set in place so as to perform the relevant FOIAs for a specific MIA at the start of the exercise. These processes can be fairly consistent across government agencias
- **Free or low cost:** The FOIA request is always free, but the agencies may charge for the time it took to perform the processing (usually tens of $)

Cons:

- **Long time frames**: some FOIAs can take up to several months until the request is fulfilled
- **Laborious analysis**: the information obtained may not be machine-readable or easy to perform scalable analysis on

good RESEARCH

Sources of Information: 5. Using a Recruiter

An attacker may try to perform custom subject recruitment for interviews or other user analysis by imitating some of the restrictions in order to encounter some of the subjects in the target dataset.

Pros:

- **Self-identification:** Most of the labor of obtaining the matches is performed by the users (or companies that provide these services)
- **Interaction with matches can lead to further matches**

Cons:

- **Not always possible**: Depending on the target dataset, it may not be possible or legal to perform subject recruitment
- **High economic costs**: Performing subject recruitment can be pricey (up to thousands of $)

good RESEARCH

Best Practices in De-Personalization of Datasets

- **Identifiers** Do **NOT** strip the main identifiers (name, address, etc.) and call it a day…
- **Aggregation** Do consider the possibility of aggregation… and when aggregating:
    - Think of the amount of people (within the dataset) that fall in the bucket, how varied the information is for these individuals, and the amount of general population that would fall in this bucket. (k-anonymity, k-map, l-diversity, delta-presence, etc.)
    - Consider using non-exclusive semi-random aggregation groups
    - Consider adding potential noise (when possible) to the aggregated results
- Consider what information, aside from individuals, can be obtained and inferred from the dataset: places people frequent, companies' business clients, trade secrets, businesses running BAU vs high-capacity
- **Look for Outliers** Look for outliers in your data:
    - Why are they outliers? what information do they tell? Should you remove/clamp outliers?
    - How are you measuring outliers? what other dimensions are in the data?
- Consider removing dates (e.g. events, DOB), or providing only wide-range date intervals, to try to defend against social media searches (although these attacks can still be successful even without dates)
- **All de-personalization isn't equal -** There is no one size fits all and Different anonymization techniques can be applied incorrectly, so be careful how you do it and what your risk profile is

Perform a Motivated Intruder Attack.

motivatedintruder.com

good RESEARCH

QUESTIONS

cc-ar-untrained-cc-https://www.flickr.com/photos/[illegible]

You will receive

- The materials from this webinar
- We organize monthly webinars on privacy and privacy enhancing technologies – we will send you information about these events
- We will be making our content available through online courses (general to advanced audiences) and will let you know about these

Contacts

Janice Branson: janice.branson@novartis.com

Nathan Good: nathan@goodresearch.com

Khaled El Emam: kelemam@replica-analytics.com

QUESTIONS

cc-ar-untrained-cc-https://www.flickr.com/photos/[illegible]