

# **Medical Data De-ID**

## **A Canadian Perspective**

**Dr. William Parker, MD, BSc, FRCPC, DABR**

---

**Session 3: International session (40 min) Chairperson**

**Founder @ SapienSecure.io**

**University of British Columbia**

**Cardiovascular Radiologist**



# Disclosures

Founder of  sapiensecure

A healthcare intelligence company.

Focused on Medical Data Indexing, Extracting and De-ID.

Many of the examples and tools discussed in this presentation are from using SapienSecure v2 software.

In the news...

Nova Scotia

# Privacy commissioner calls for changes at N.S. Health after staff found snooping



Health authority issues apology, says it plans to accept most recommendations



[Josefa Cameron](#) · CBC News · Posted: Feb 08, 2023 9:25 AM PST | Last Updated: February 8



FASKE

De-  
Take


Hea



JUNE 2, 20

- having all employees, consultants, and sub-contractors sign confidentiality contracts prohibiting data linking and/or re-identification;
- only allowing authorized staff to access and use data on a “need-to-know” basis;
- ensuring all employees, consultants, and sub-contractors working with the data receive adequate privacy and security training;
- developing and maintaining data privacy, security, and usage standard operating procedures that specifically prohibit re-identification;
- developing and maintaining strictly enforced retention, destruction and storage policies;
- developing and maintaining role-based data access policies and processes, which are enforced and periodically audited;
- maintaining records of all signed data-sharing agreements and confidentiality agreements, and making those available to the data custodian on request;
- maintaining a proactive program for monitoring privacy, confidentiality and security polices and procedures, a mandatory and on-going training program for all individuals, and a breach protocol that is regularly updated and tested;
- ensuring that external and internal privacy reviews and audits are regularly conducted and that any identified gaps are mitigated; and
- prohibiting data linking and re-identification.



# Canadian Association of Radiologists White Paper on De-Identification of Medical Imaging: Part I, General Principles

Canadian Association of  
Radiologists' Journal  
2021, Vol. 72(1) 13-24  
© The Author(s) 2020  
Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
DOI: 10.1177/0846537120967349  
[journals.sagepub.com/home/caj](https://journals.sagepub.com/home/caj)  




**William Parker, MD, BMSc<sup>1,2</sup> , Jacob L. Jaremko, MD, PhD, FRCPC<sup>3</sup>,  
Mark Cicero, MD, BEMSc, FRCPC<sup>4,5</sup>, Marleine Azar, MSc<sup>6</sup>,  
Khaled El-Emam, PhD<sup>7</sup>, Bruce G. Gray, MD, FRCPC<sup>8</sup>,  
Casey Hurrell, PhD<sup>9</sup> , Flavie Lavoie-Cardinal, MSc, PhD<sup>10</sup>,  
Benoit Desjardins, MD, PhD, FACR<sup>11</sup>, Andrea Lum, MD, FRCPC, CCPE<sup>12</sup>,  
Lori Sheremeta, LLB, LLM<sup>13</sup>, Emil Lee, MD, FRCPC<sup>14</sup>,  
Caroline Reinhold, MD, MSc<sup>15,16</sup>, An Tang, MD, MSc, FRCPC<sup>17</sup>,  
and Rebecca Bromwich, PhD, LLM, LLB<sup>18</sup>**

## Abstract

The application of big data, radiomics, machine learning, and artificial intelligence (AI) algorithms in radiology requires access to large data sets containing personal health information. Because machine learning projects often require collaboration between



# Canadian Association of Radiologists White Paper on De-identification of Medical Imaging: Part 2, Practical Considerations

**William Parker, MD, BMSc<sup>1,2</sup> , Jacob L. Jaremko, MD, PhD, FRCPC<sup>3</sup>, Mark Cicero, MD, BEMSc, FRCPC<sup>4,5</sup>, Marleine Azar, MSc<sup>6</sup>, Khaled El-Emam, PhD<sup>7</sup>, Bruce G. Gray, MD, FRCPC<sup>8</sup>, Casey Hurrell, PhD<sup>9</sup> , Flavie Lavoie-Cardinal, MSc, PhD<sup>10</sup>, Benoit Desjardins, MD, PhD, FACR<sup>11</sup>, Andrea Lum, MD, FRCPC, CCPE<sup>12</sup>, Lori Sheremeta, LLB, LLM<sup>13</sup>, Emil Lee, MD, FRCPC<sup>14</sup>, Caroline Reinhold, MD, MSc<sup>15,16</sup>, An Tang, MD, MSc, FRCPC<sup>17</sup>, and Rebecca Bromwich, PhD, LLM, LLB<sup>18</sup>**

Canadian Association of  
Radiologists' Journal

2021, Vol. 72(1) 25-34

© The Author(s) 2020

Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/0846537120967345

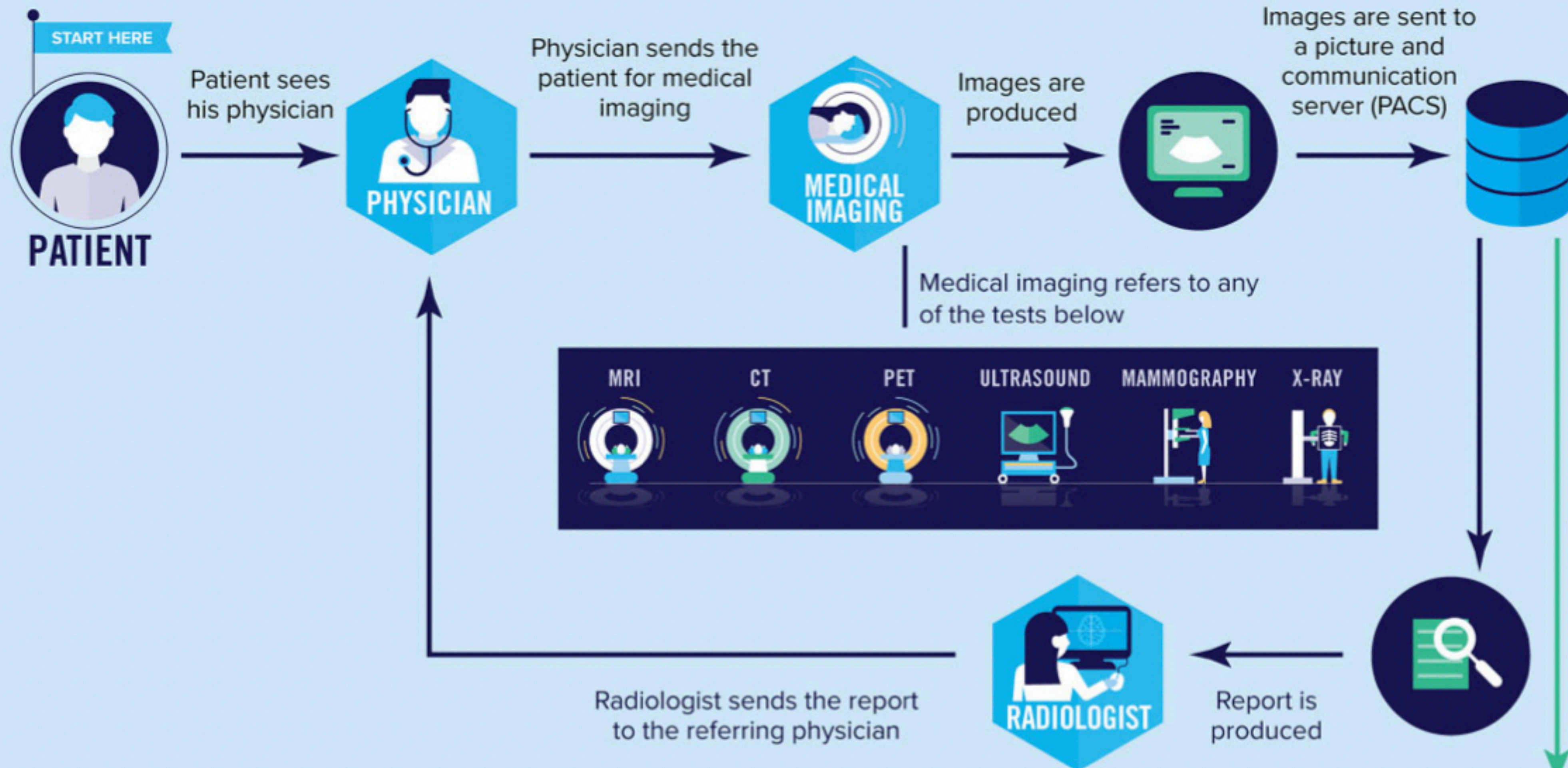
[journals.sagepub.com/home/caj](https://journals.sagepub.com/home/caj)





# Medical Data and the Patient Journey

## A. Patient Journey





# De-identification

Process of transforming direct and indirect identifiers, and possibly implementing additional controls so that the **likelihood** of data subjects being correctly identified from the information is **very small** under the circumstances of use or disclosure.

# Anonymization

Process of permanently removing all protected health information from the data set.

# Pseudonymization

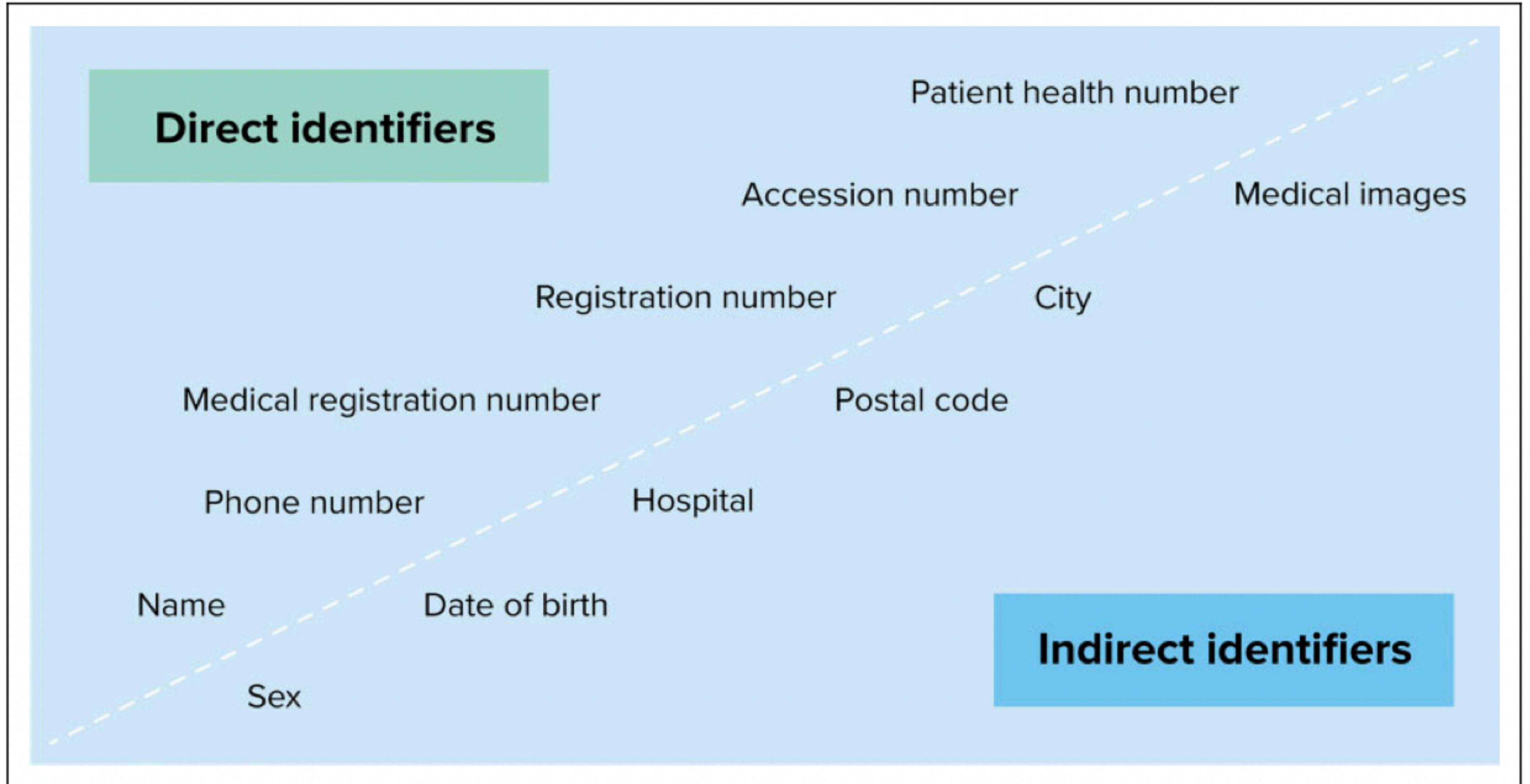
Process of transforming direct identifiers and possibly implementing additional controls so that data subjects **cannot** be correctly identified from the information under the circumstances of the use or disclosure.

# Encryption

Process of using a computer algorithm to obfuscate the patient information, making it unintelligible and random looking. The original information can be recovered using a decryption algorithm.



## Classifying Variables





**Health Insurance Portability and Accountability Act (HIPAA) clause 164.514**

*“risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual.”*



- Name**
- Address (geography smaller than state)**
- Any specific date**
- Telephone numbers**
- Fax number**
- Email address**
- Social Security Number**
- Medical record number**
- Health plan beneficiary number**
- Account number**
- Certificate or license number**
- Vehicle identifiers and serial numbers, including license plate numbers**
- Device identifiers and serial numbers**
- Web URL**
- Internet Protocol (IP) Address**
- Finger or voice print**
- Photographic images**
- Any other characteristic that could uniquely identify the individual**

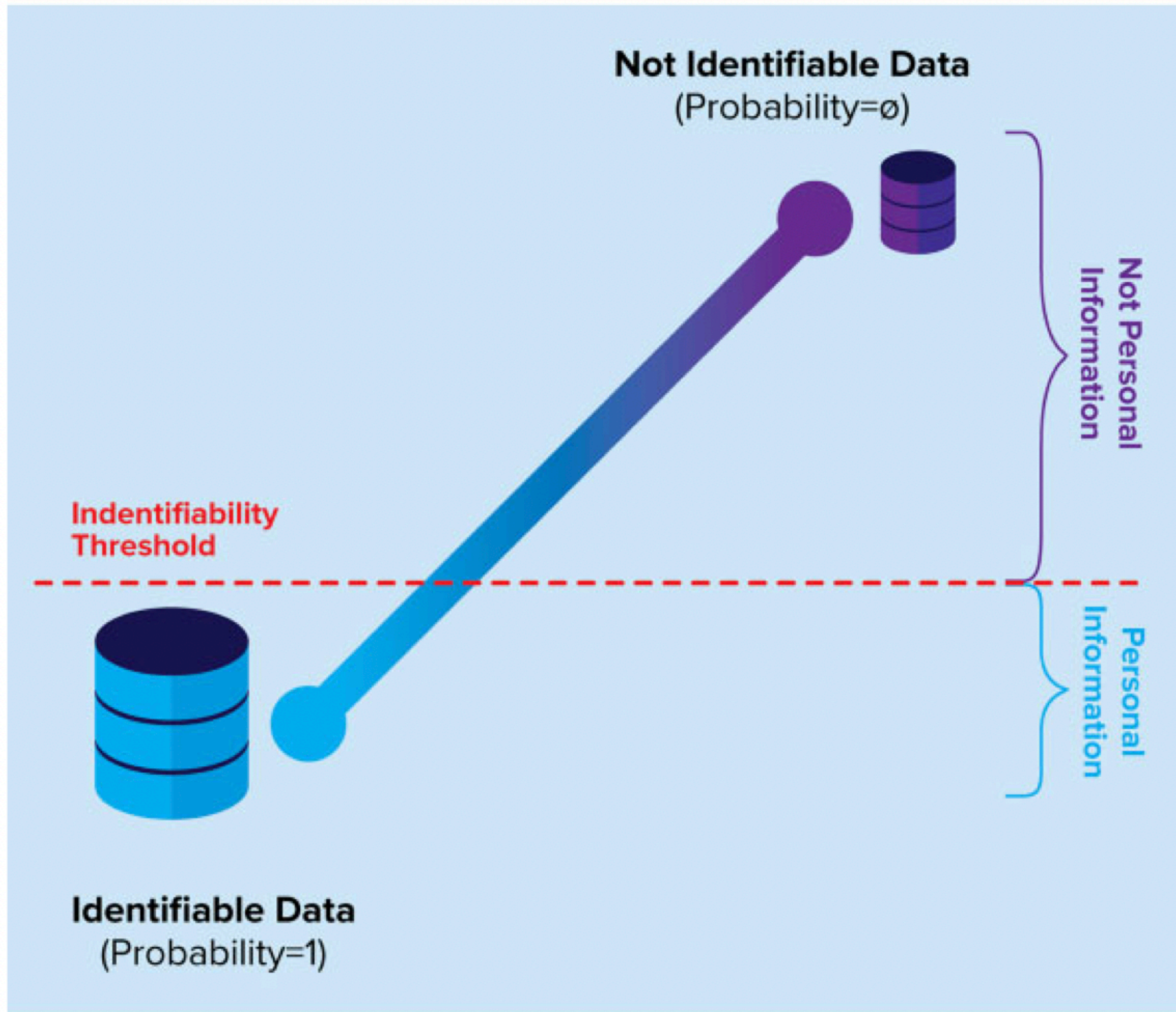
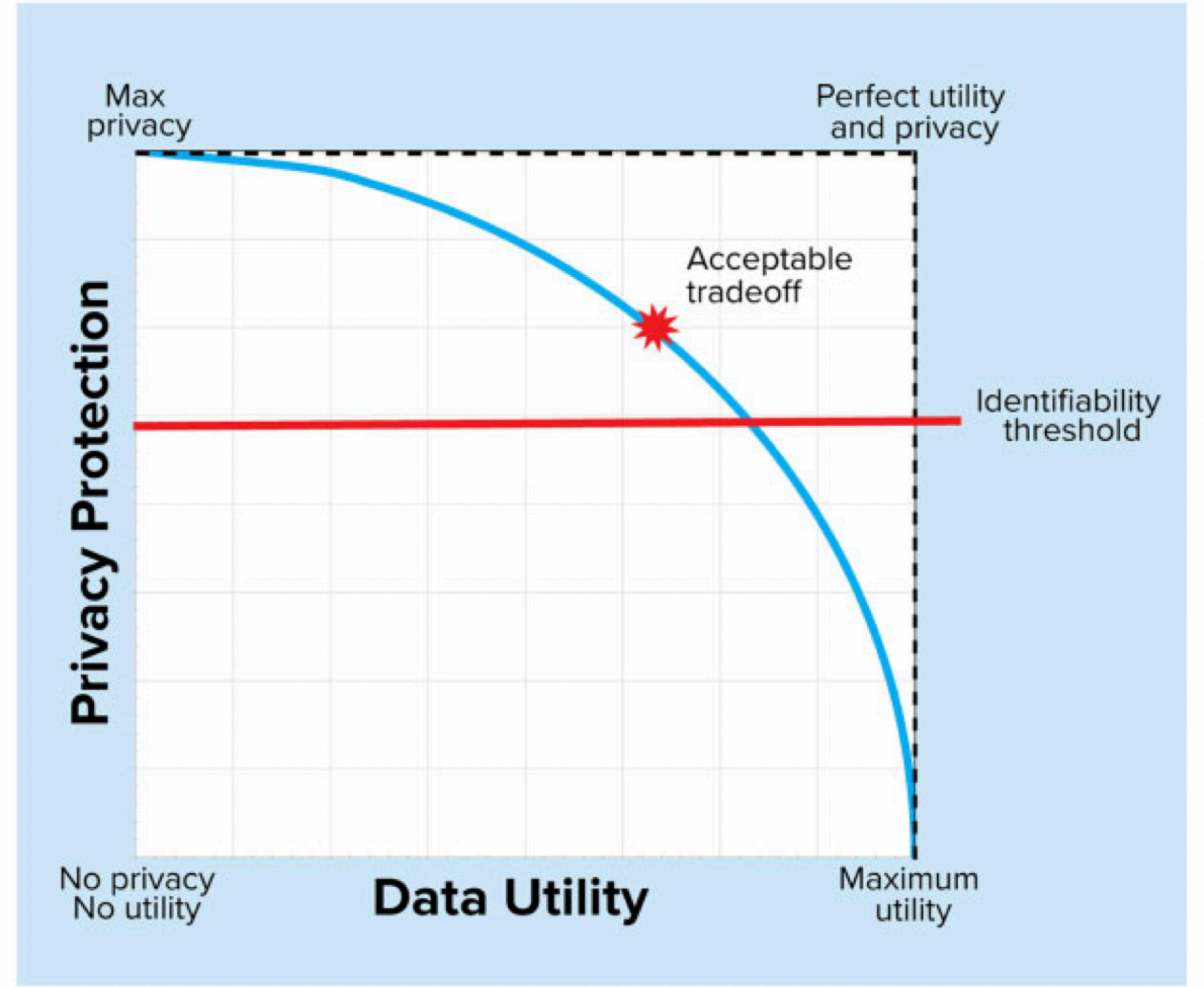
**Personal Information Privacy and Electronic Documents Act (PIPEDA)** applies to private-sector organizations across Canada that collect, use or disclose personal information in the course of a commercial activity.



**Under PIPEDA, personal information includes any factual or subjective information, recorded or not, about an identifiable individual. This includes information in any form, such as:**

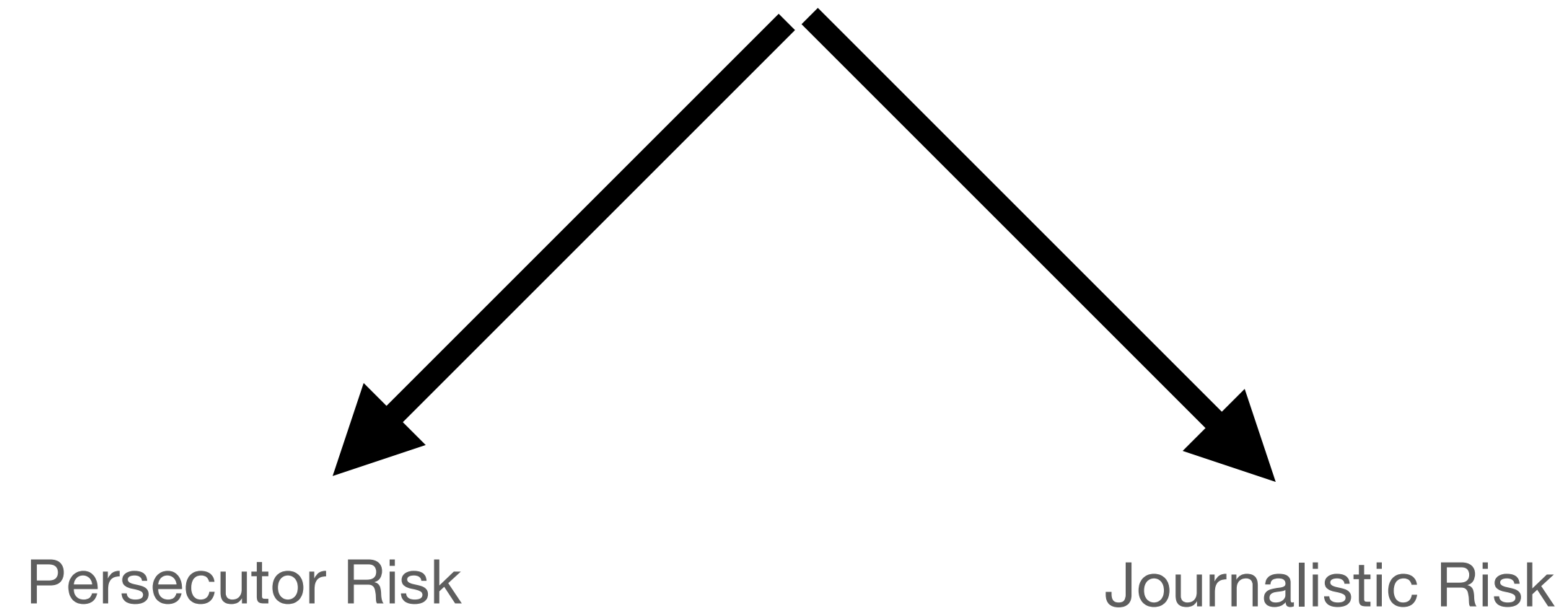
- age, name, ID numbers, income, ethnic origin, or blood type;**
- opinions, evaluations, comments, social status, or disciplinary actions; and**
- employee files, credit records, loan records, medical records, existence of a dispute between a consumer and a merchant, intentions (for example, to acquire goods or services, or change jobs).**



**A****B**



# Assessing Risks of Re-identification



**Prosecutor risk assumes that an adversary is informed of the individuals within a data set.**

**‘For example, if a teenager’s parents know that their child has participated in a survey and the results are to be released in de-identified form, the risk of the parents attempting to re-identify their child’s responses would qualify as prosecutor risk.**

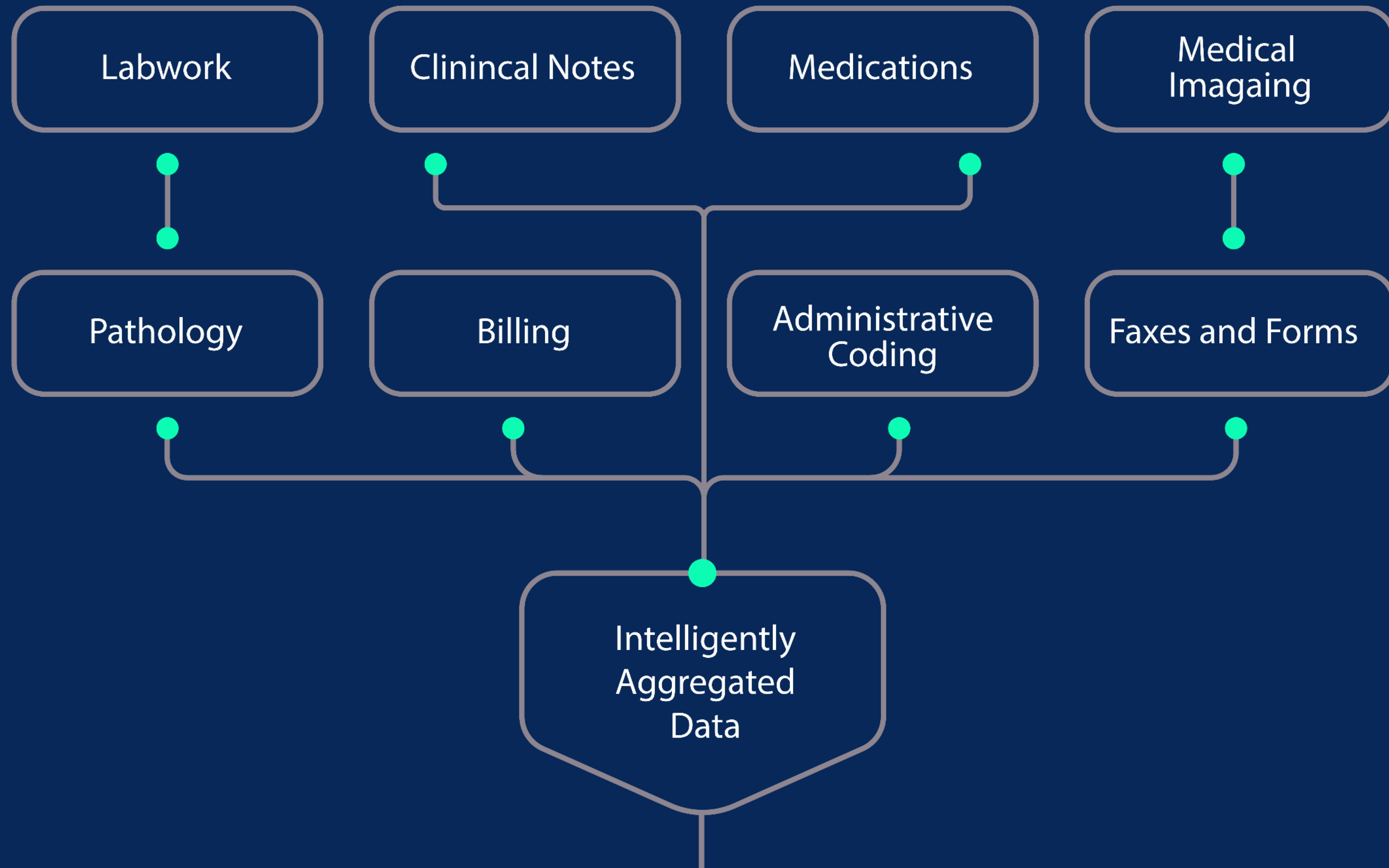
**Journalist risk assumes that the adversary does not know whose data are in the data set, thus the risk of this assessment is significantly decreased.**

**For example, “if only a sample of de-identified rows from an original data set is released, this would qualify as journalist risk.”**

*Because terminology has been somewhat inconsistent across statutes and in general use within jurisdictions, it is most useful to focus on the concept of identification risk as follows. This risk consists of 2 criteria: (a) the ability to correctly match a record to a real person, and (b) by doing so learn something new about that person.*



# Step 1.





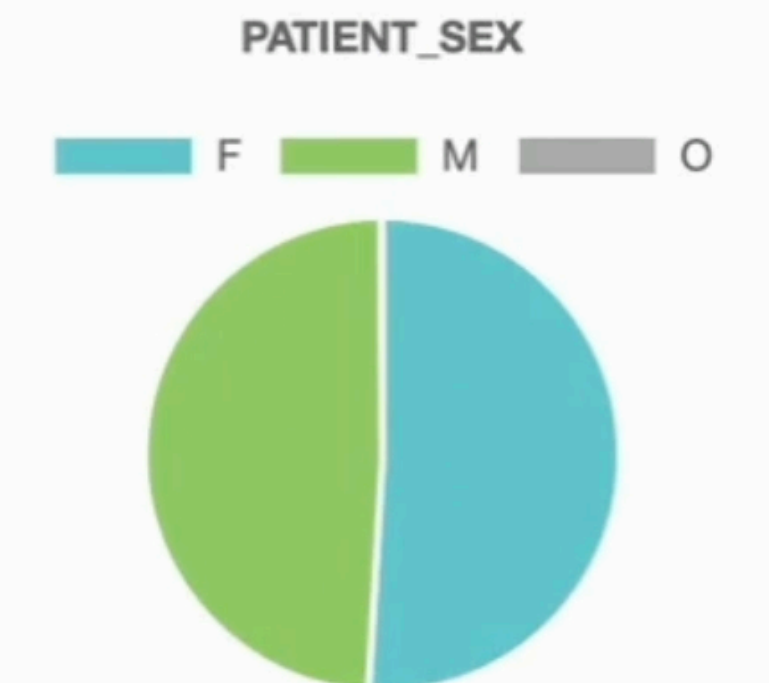
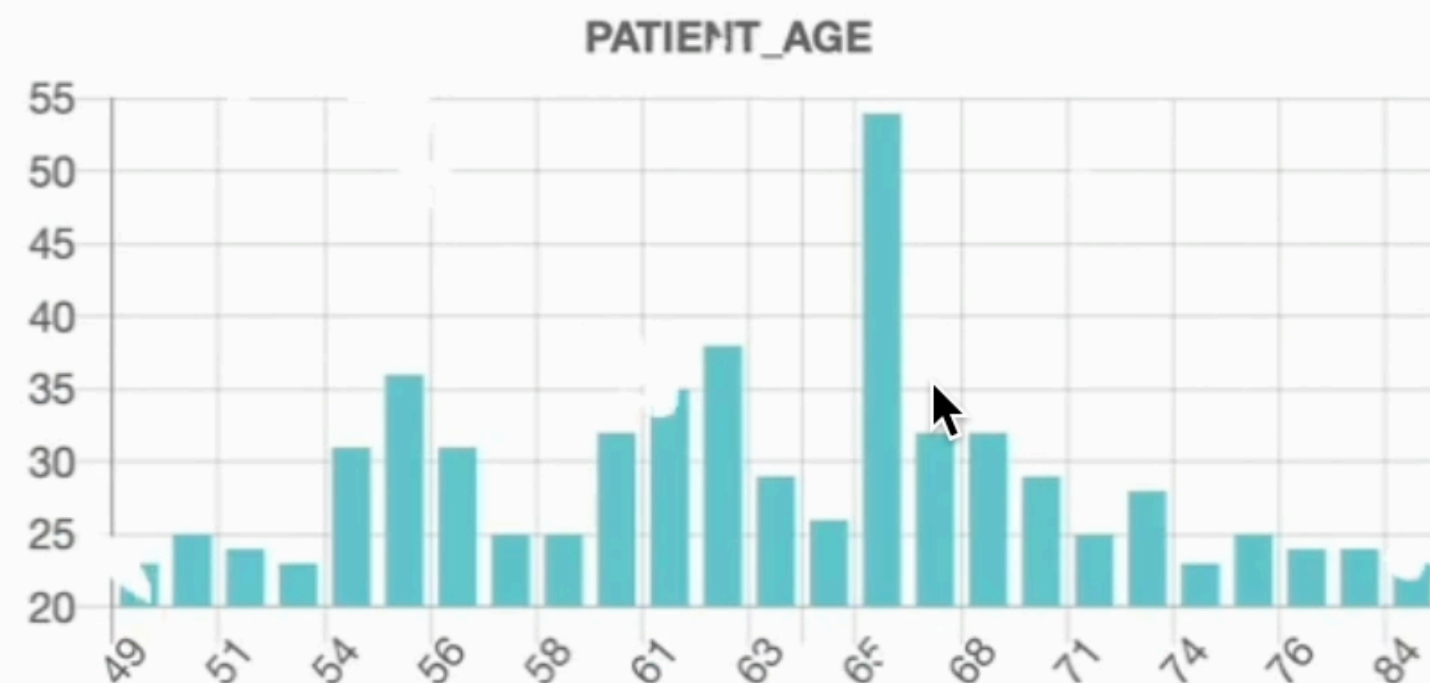
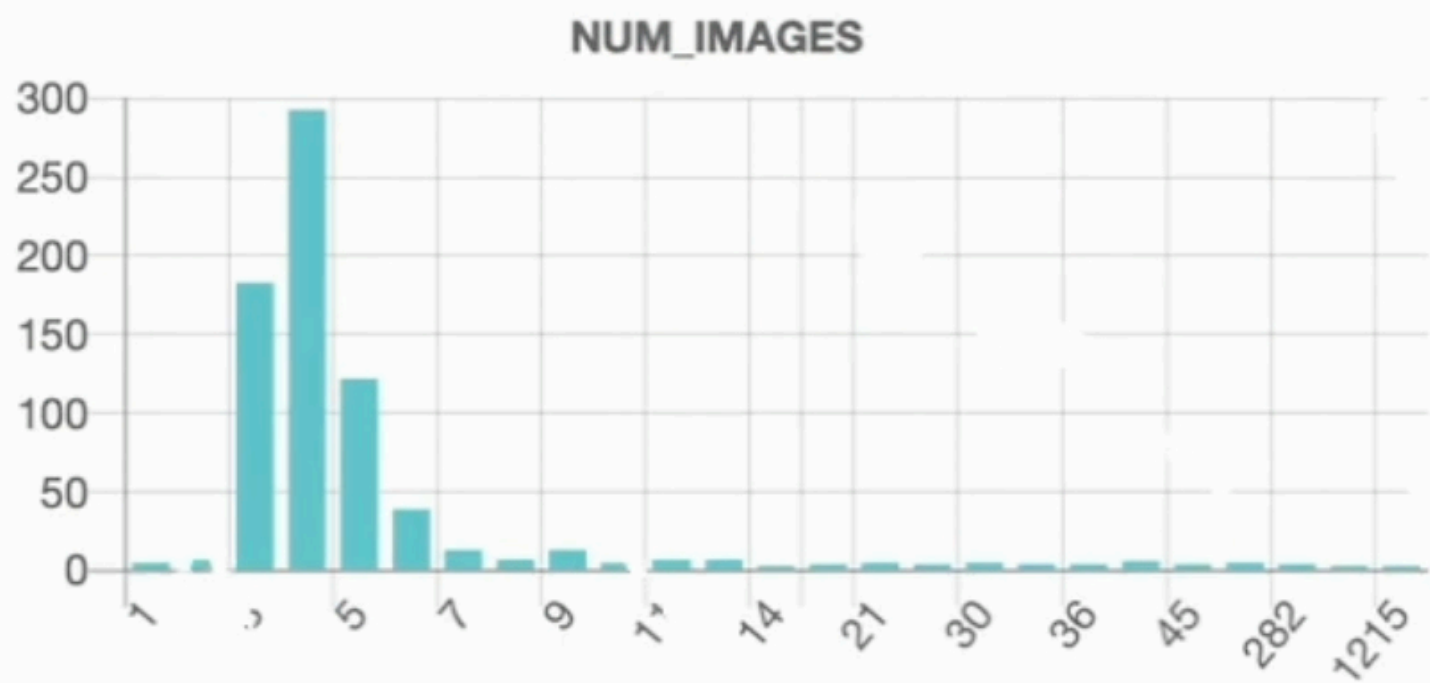
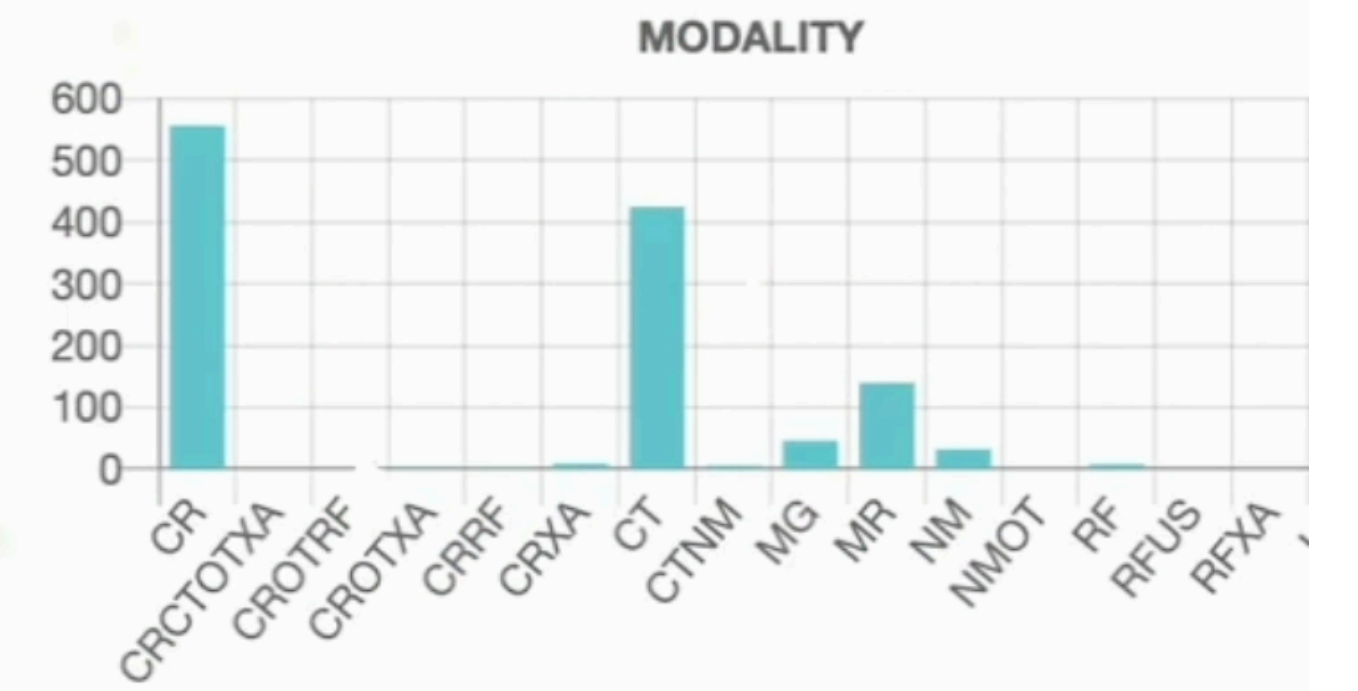
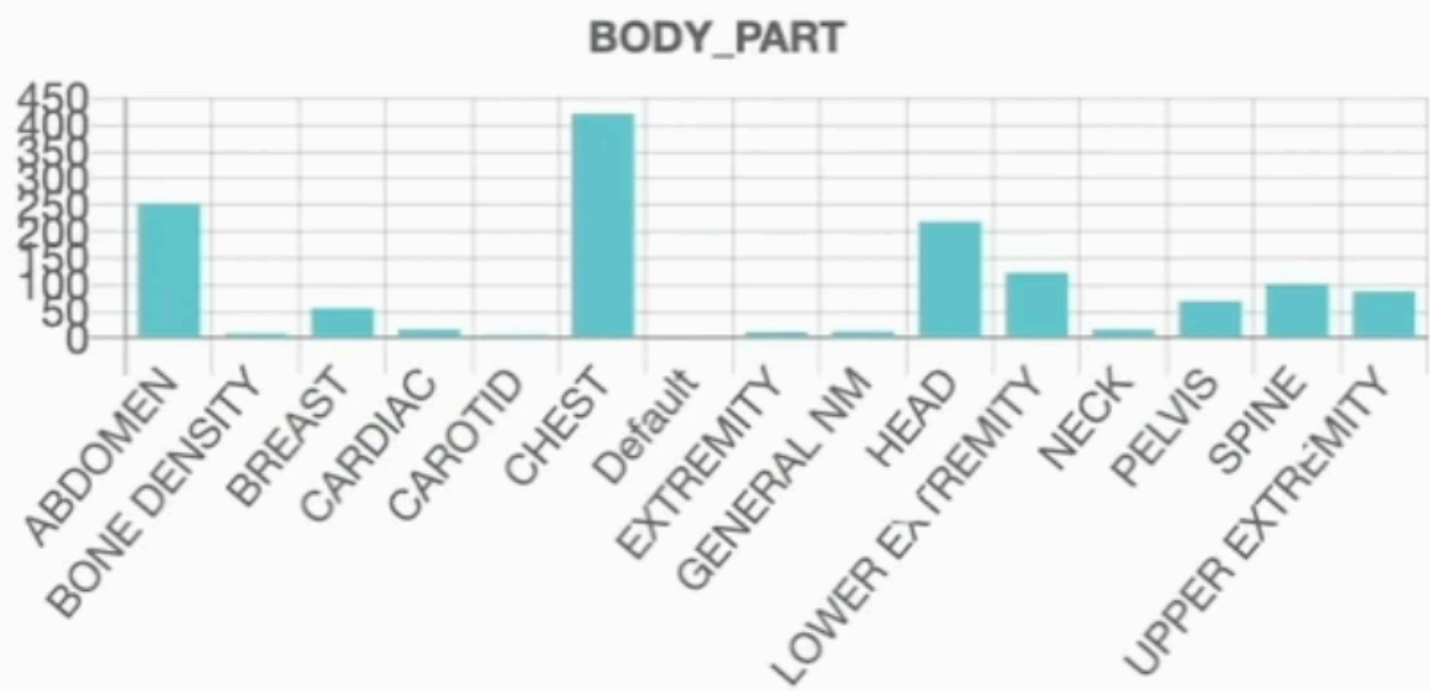
Filters

INSTITUTION\_NAME

Research Clinic



Select a Field...



STUDY DATE

STUDY DESCRIPTION

report Anatomy



# Step 2.

## De-Identification Processes

Hand-writing  
Redaction

Image Defacing

Date Shift  
Consistency

Image Burnt-in  
Text Redaction

Medical Record  
Consistency

Differential Privacy

PHI NLP

K- Anonymity

Forms OCR  
Redaction

DICOM De - ID



# Medical Record Value Consistency

**The unique identifiers, dates, and locations are not usually important specifically in research. But in relationship to other records in the dataset they can be essential to the success of the project. Medical Record Number or Date consistency software can help create safe datasets that retain value.**

PatientID	Gender	Age	Zip code	Test
55998	M	19	15723	Negative
88557	F	35	15674	Positive
55868	F	35	15674	Positive
44551	M	45	15623	Negative
58524	M	45	15623	Negative
25584	F	61	15633	Negative
58744	F	61	15643	Positive
87524	M	19	15762	Positive
87384	M	19	15762	Negative
17583	F	19	15762	Positive

M: male; F: female

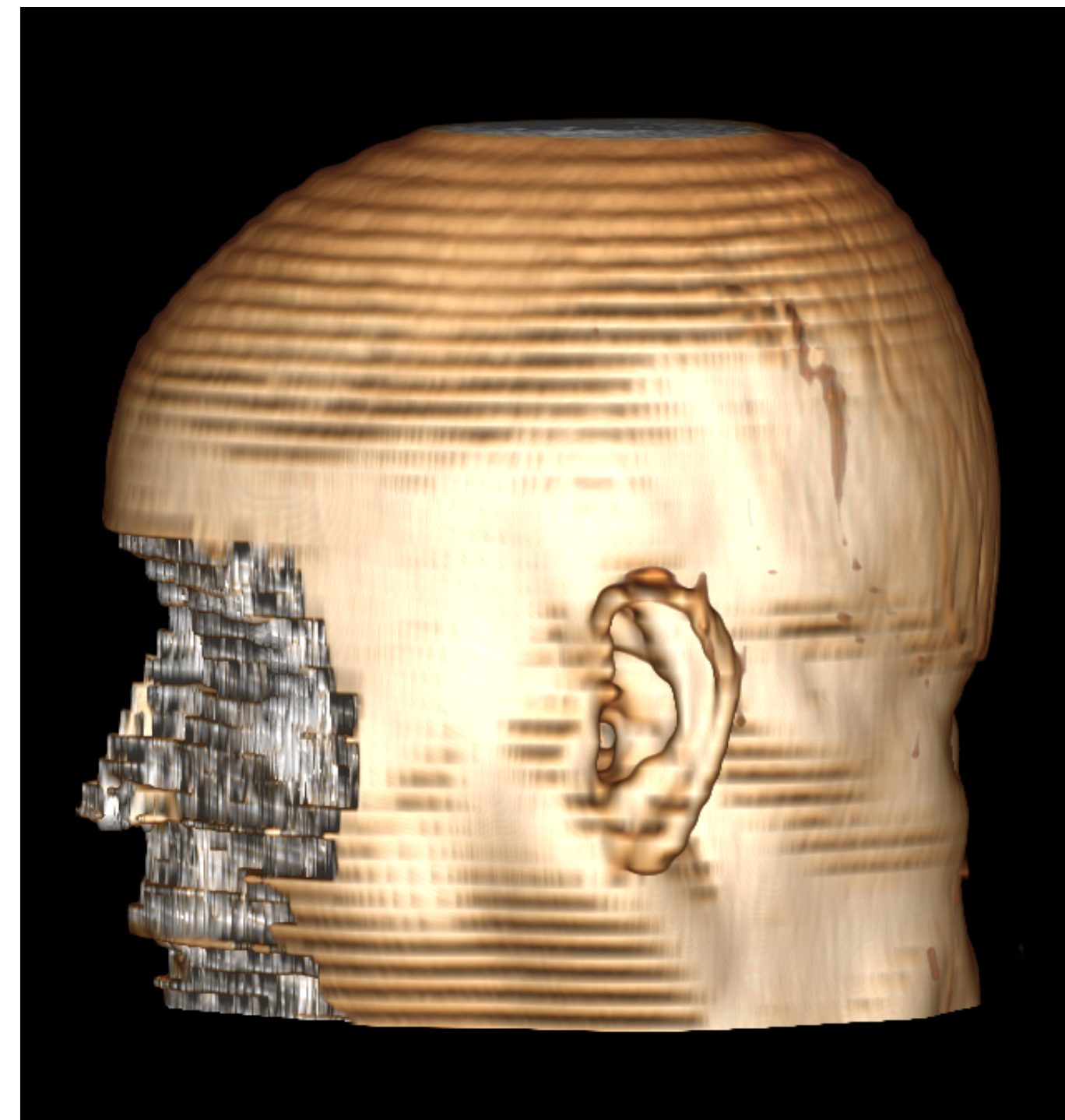


# Image De-Facing

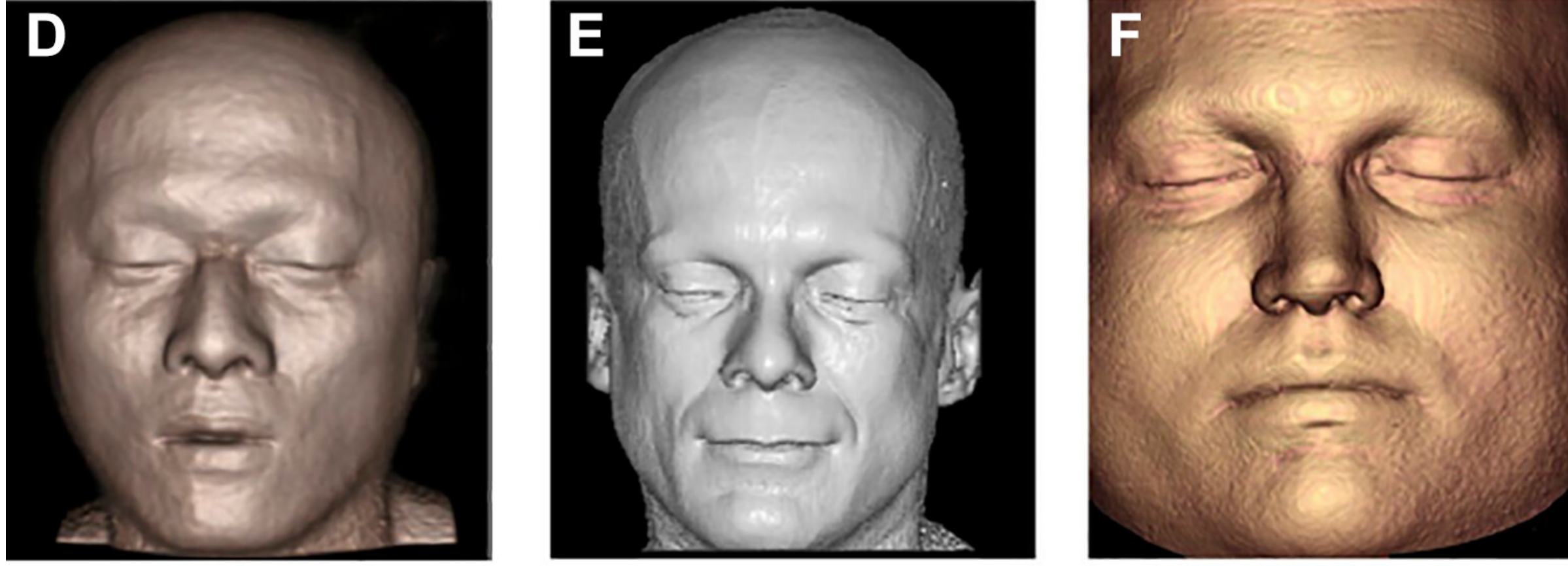
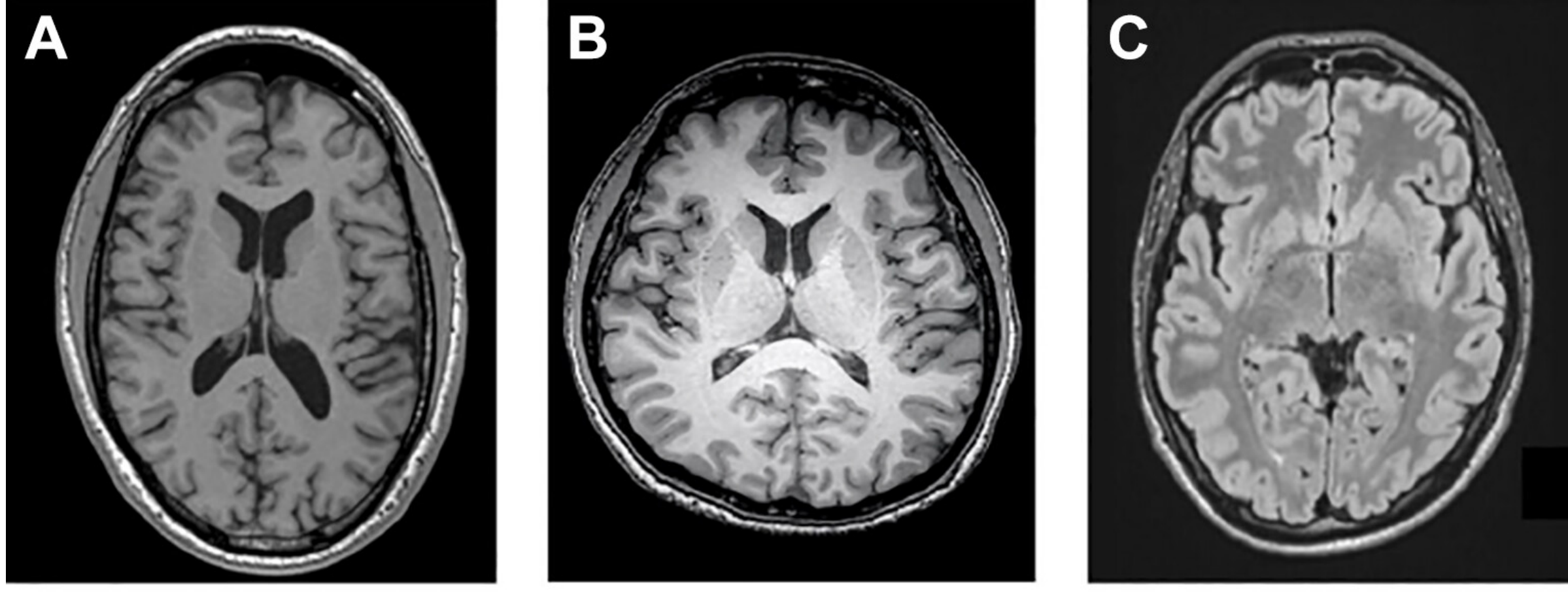
**Multi-planar (CT and MRI) hold enough spacial information to reconstruct the face of the patient.**

**Many big data projects are focused on neuro-based problems and require imaging.**

**Simple defacing algorithms exist open source, and from businesses that will automate this.**



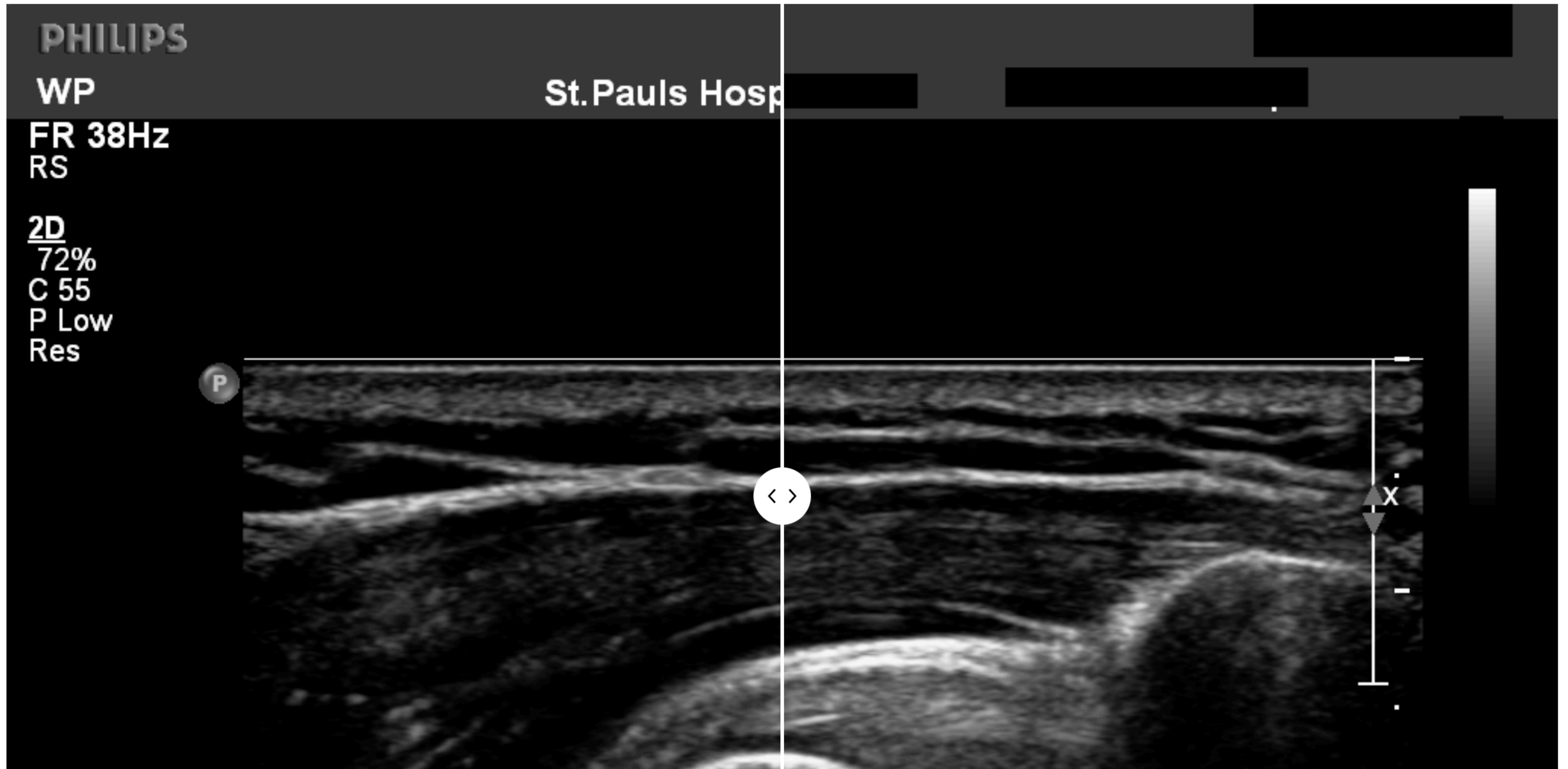




Answers: 1) a-f-g. 2) b-d-i. 3) c-e-h.



# Burnt-in Text Redaction



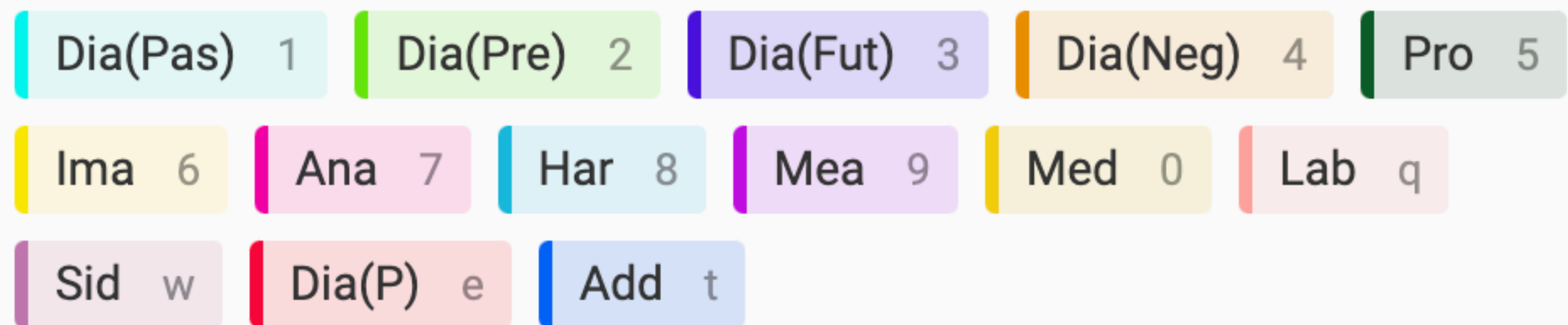
# PHI NLP

**Medical Reports are treasure troves of medical information that can be game changing for a big-data project.**

**Extraction of information from reports should not compromise privacy.**

**PHI NLP is becoming more and more common, and with transformers, can be extremely accurate.**

**Can be locally trained and re-trained at institutions to improve accuracy over time.**



Catheter over guidewire advancement of a 5-French soft tip guiding catheter then followed, with digital subtraction biplane angiograms performed in the right common carotid and right external carotid artery territories.

Sample Name

Discharge Summary - Mesothelioma — 1

Record date 2093 0113,DavidHale,MD ,

DATE

Name : Hendrickson, OraMR

#7194334

NAME

ID

Description

Mesothelioma, pleural effusion, atrial fibrillation, anemia, ascites, esophageal reflux, and history of deep venous thrombosis

(Medical Transcription Sample Report)

PRINCIPAL DIAGNOSIS

Mesothelioma

SECONDARY DIAGNOSES

Pleural effusion, atrial fibrillation, anemia, ascites, esophageal reflux, and history of deep venous thrombosis

PROCEDURES

1 On August 24, 2007 , decortication of the lung with pleural biopsy and transpleural

DATE

fluoroscopy

1 August 20, 2007 , thoracentesis

DATE

1 August 31, 2007 ,

Port

A Cath placement

DATE

LOCATION

HISTORY AND PHYSICAL.

patient is a 41 year old female and was born in Vietnam She has had a nonproductive

AGE

LOCATION

ugh that started last week and right sided chest pain radiating to her back with fever starting

erday She has a history of pericarditis and pericardectomy in May 2006 and developed

DATE

DATE

ugh with nght sided chest pain, and went to an urgent care center Chest x ray revealed



# Differential Privacy

Enough Indirect Identifiers in Combination  
Can Directly Identify an Individual

Name	Postal code	Age	Sex	Dx
Jacquie	S7A 4D1	23	F	Pneumonia
Patricia	S7A 2X2	27	F	Pneumonia
George	A4T 1D7	18	M	COVID-19
Albert	A4T 4E3	13	M	Pneumonia
Sienna	S7A 1G7	25	F	COVID-19



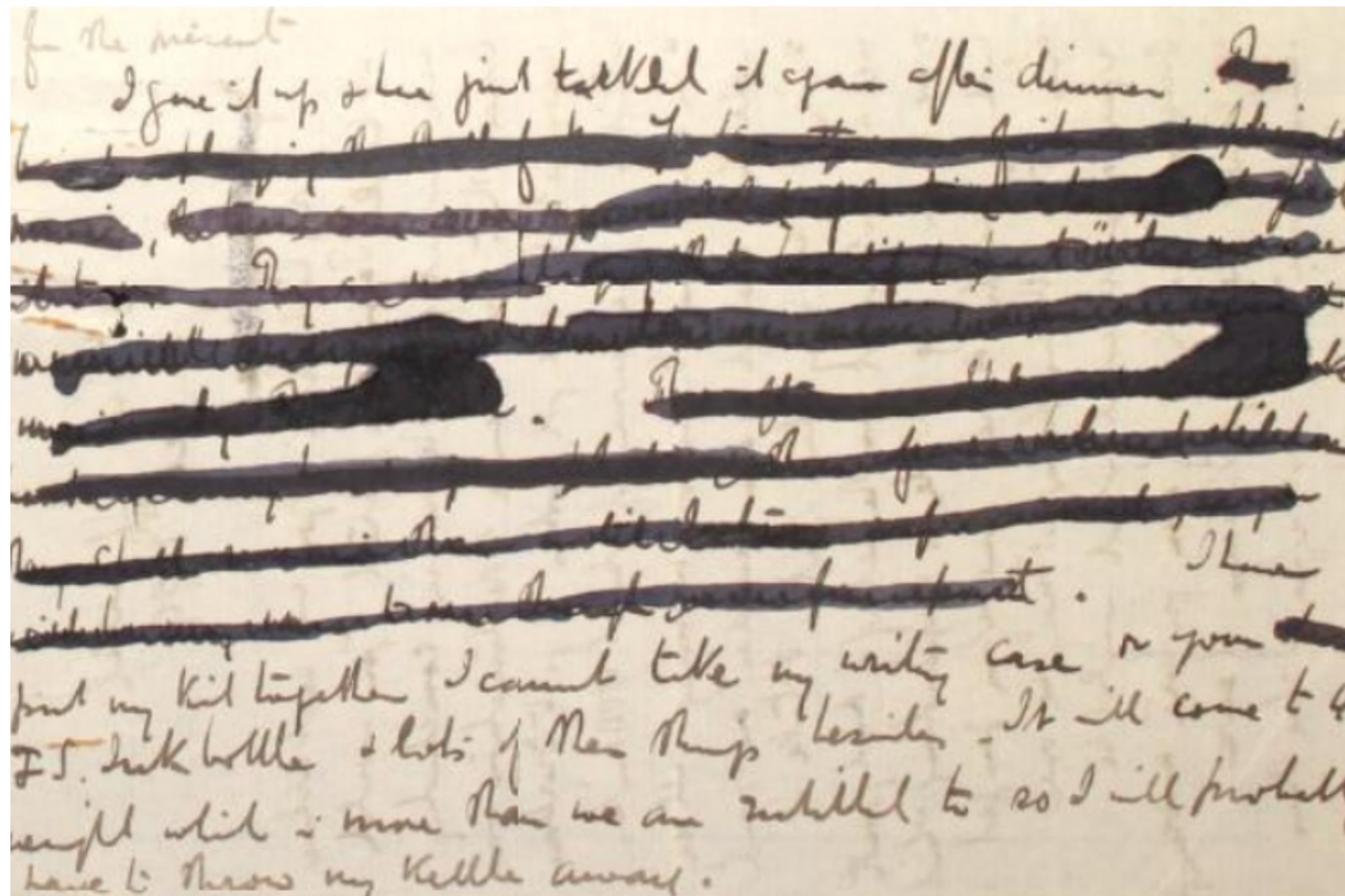


# Hand-writing Redaction

**Hand writing is one of the more challenging data-mineable sources, and littered with personal information.**

**My research projects focus on identifying and removing this data.**

**There are simple software tools that can analyze files and detect the presence of handwriting.**





# Final words...

- True healthcare improvements will occur with increased mobility and movement of medical data.
- Patient privacy does NOT need to be compromised.
- Technology exists to export, aggregate and de-ID medical data almost instantly, and more organizations need to adopt these tools to empower innovation.

# THANK YOU

# Questions?

**Dr. William Parker, MD, BSc, FRCPC, DABR**

---

**Session 3: International session (40 min) Chairperson**

**Founder @ SapienSecure.io**

**University of British Columbia**

**Cardiovascular Radiologist**

