



# De-id for PHI NLP task

Zhendong Wang

zhw65@pitt.edu

# Term Explain

**De-id:** De-identification, the process used to prevent a person's identity from being connected with information

**PHI:** Protected Health Information (e.g. your health record)

**NLP:** Nature Language Processing

# Agenda

- Why de-id is important
- Challenges in de-id process
- Best practice of de-id process

# Why important?

- Deep Learning bring clinicians the possibility to use NLP+DL to ease their daily job:
  - Linguamatics
  - Talix
  - Health Fidelity
- A lot of DL algorithm needs medical data
- We can't use clinical records directly because it contains your personal information

Talix<sup>™</sup>

(H<sub>F</sub>) HEALTH FIDELITY<sup>®</sup>

 Linguamatics

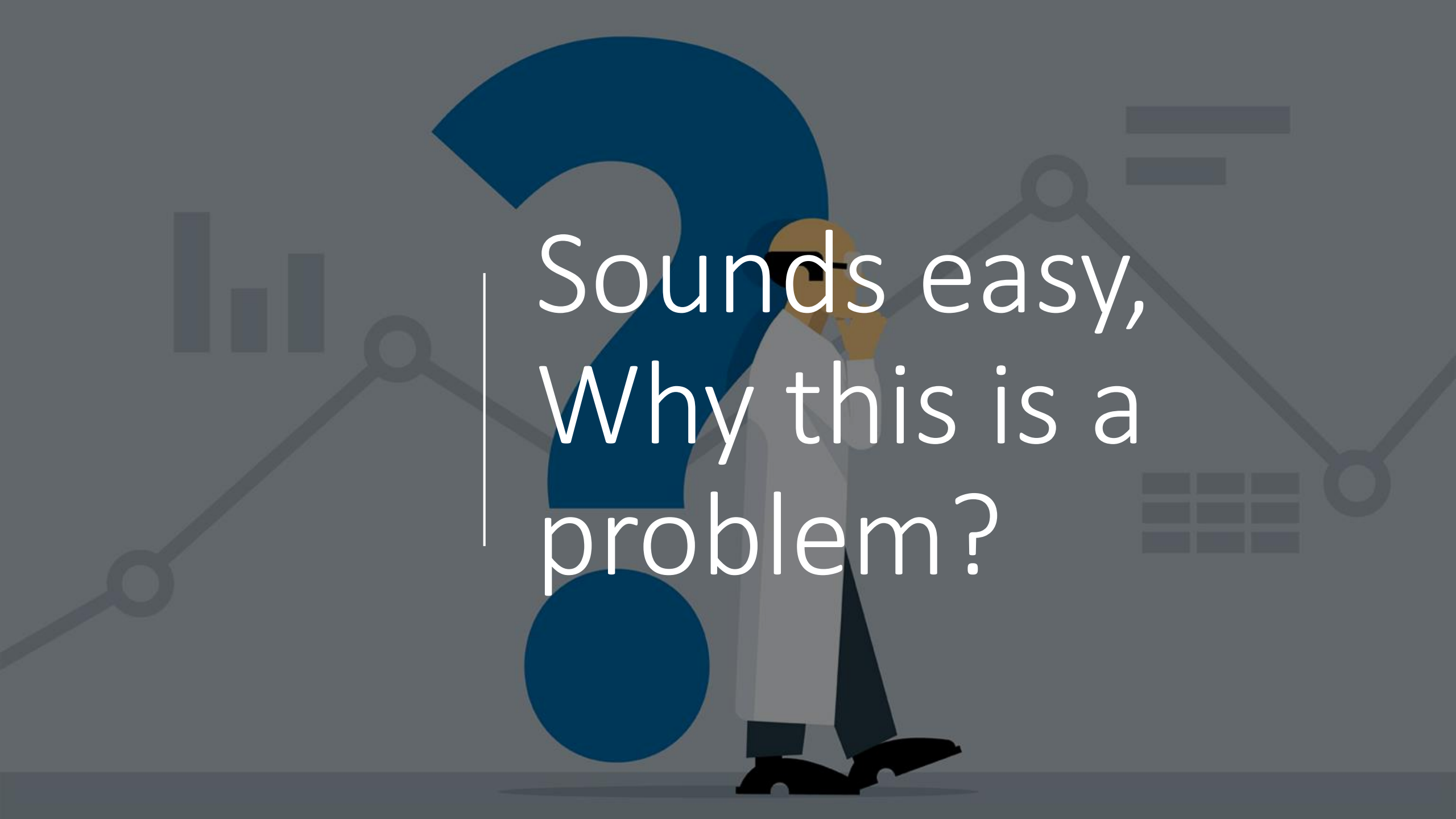
## A simple de-id sample

Before de-id:

“Tony Wang living in Pittsburgh uses Aspirin and feels better now”

After de-id:

“XXX living in XXX uses Aspirin and feels better now”



Sounds easy,  
Why this is a  
problem?

# Challenge 1

- De-id process should try to preserve semantic meaning.
- If we replace everything to “XXX” we lost the semantic information like “Place”, “Name” which is important for NLP task

“<Person> living in <City> uses Aspirin and feels better now”



# Challenge 2

- We need to preserve the co-reference information

“**Tony Wang** living in **Pittsburgh** uses Aspirin and feels better now. **He** should continue use it for the next month.

**Hannah** also living in **Pittsburgh** and uses Aspirin and **she** doesn't like it.”





# Challenge 3

- people names, location can be infinite.  
No explicit rules for them
- Also suffer from the wrongly spelled name



# Challenge 4

- de-id requires several times (usually, you can't do it success in one time)
- More or less there will be data leak and it's not easy to react to this easily



# Challenge 5

- De-id software has limit integration with NLP algorithm
- Usually, if anything wrong happened in de-id process and they need to be re-run, the de-id process will re-run from plain text and the NLP pipeline needs to be totally re-run which is very very slow



# Challenge 6

- Different organization use their own de-  
id format
  - In Mimic III dataset and the one  
from UPMC dataset the format of  
the data is quite different.
  - First they don't contain the same  
data type
  - Second, even for the same data  
type, they are in different data  
format
    - E.g. `[** date **]`, `**DATE`



# Challenge 7

- poor designed de-id format
  - In UPMC dataset, their de-id format is like '\*\*<data type in uppercase>' however, they also has some header data like "\*\*\*\*\*<uppercase word>"
  - Also the half-opened format make it possible to merge de-id format with following text.
    - 120ML
    - \*\*NUMML

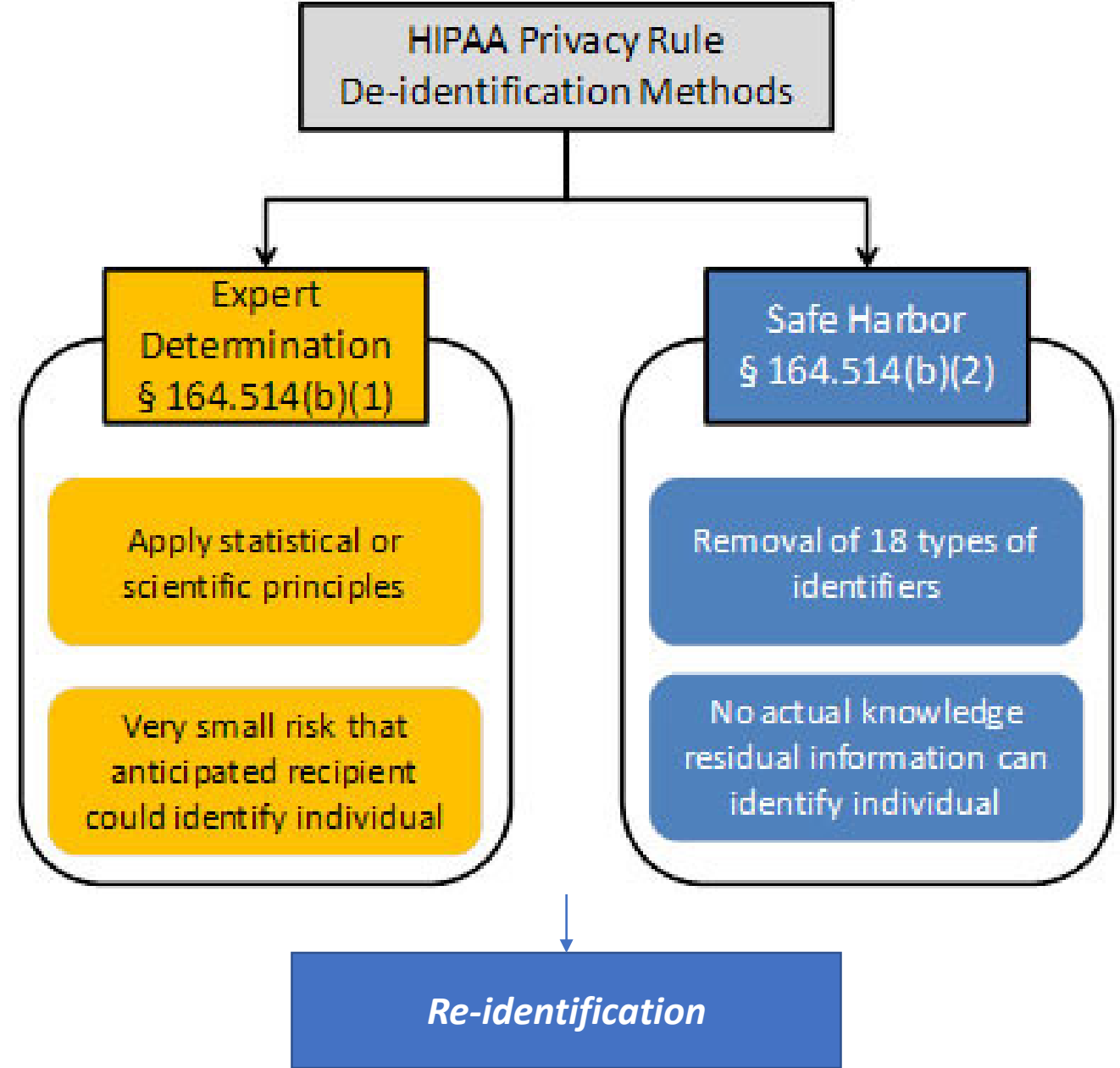


A hand is shown from the bottom right, palm up, holding a glowing lightbulb. The lightbulb is illuminated from within, casting a warm glow. The background is dark and out of focus.

How to solve the challenge?

## Step 1: Follow a guide

- [HIPAA Rules](#) : design the basic workflow for de-id process and the data type
- Didn't provide implementation



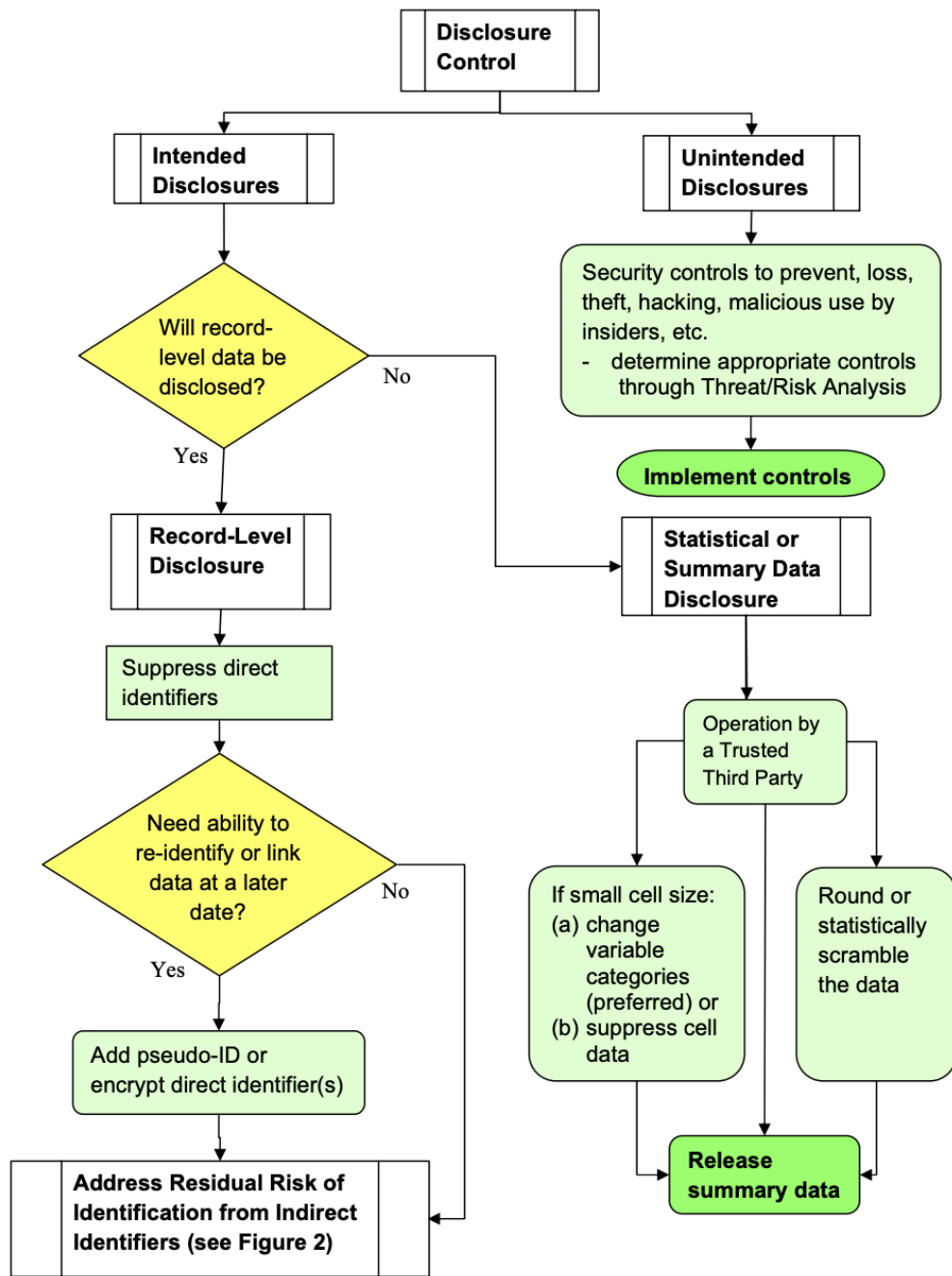


Figure 1 Applying Tools for Disclosure Control

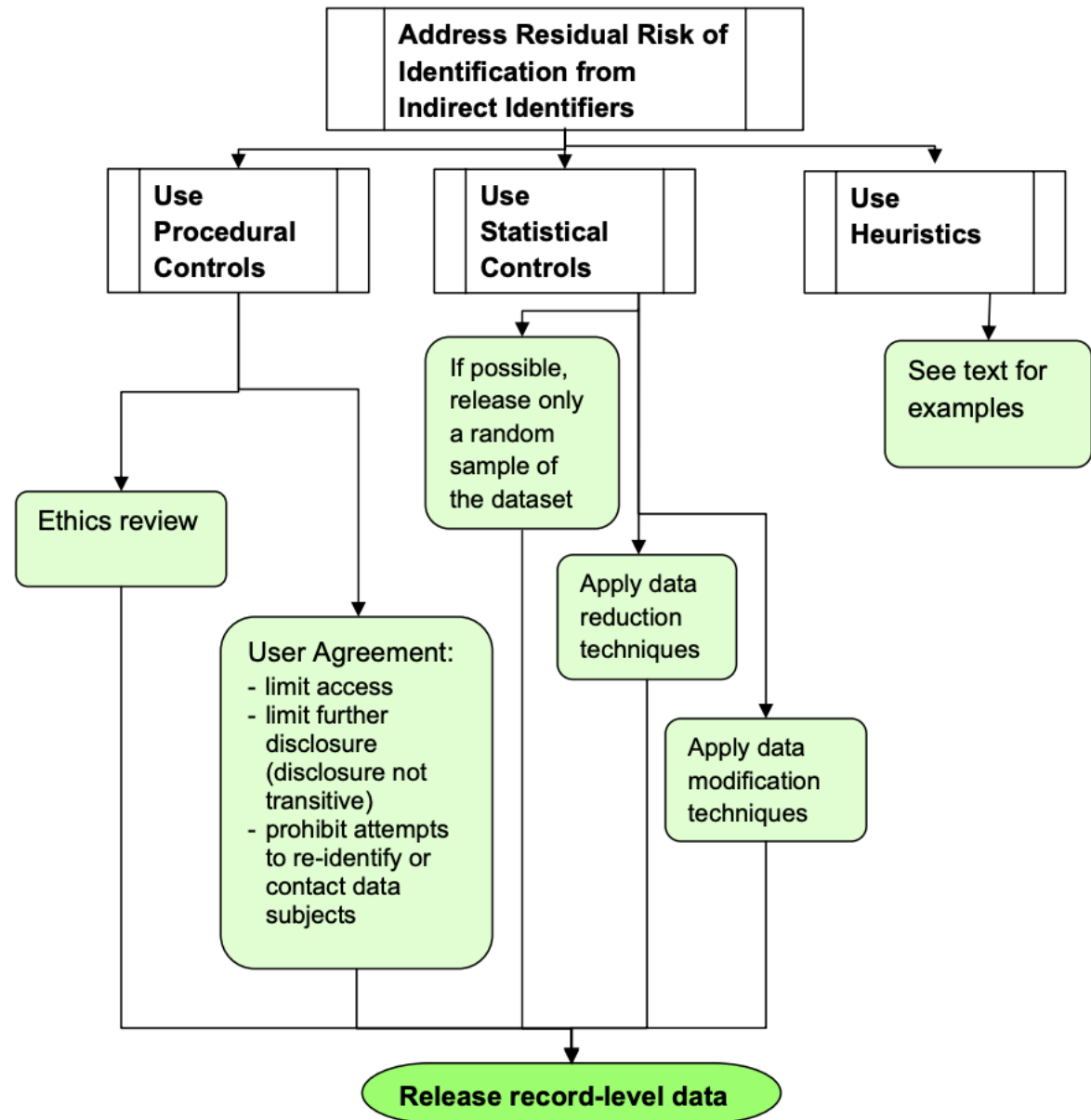


Figure 2 Applying Tools for Disclosure Control (Indirect Identifiers)



# HIPAA also defines a formal sets of data type

- The geographic unit formed by combining all ZIP
- dates
- Telephone numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Fax numbers
- Device identifiers and serial numbers
- Email addresses
- Web Universal Resource Locators (URLs)
- Social security numbers
- Internet Protocol (IP) addresses
- Medical record numbers
- Biometric identifiers, including finger and voice prints
- Health plan beneficiary numbers
- Full-face photographs and any comparable images
- Account numbers
- Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section “Re-identification”]; and
- Certificate/license numbers

# Disclosure of Record-Level Data

## Data Reduction

- Suppression of direct identifiers
- Reduction in detail
- Sampling

## Data Modification

- Random addition of “noise” to the data
- Randomization of data values
- Data swapping

Data Suppression

Pseudonymisation

# Disclosure of Aggregate Data

- Restriction-Based Methods
  - Cell suppression
  - Changing the classification scheme

**Changing the Classification Scheme by Collapsing Cells To Combine Data Ranges:**

Gender variable	Bottom-coding			Top-coding	
	Under 12	12 to 15	16 to 19	20+	Total
Males	23	20	18	19	80
Females	2	5	7	6	20
Total	25	25	25	25	100
<b>Collapse cells to eliminate small cell size:</b>					
	Under 15	16 to 19	20+	Total	
Males	43	20	15	80	
Females	7	5	5	20	
Total	50	25	20	100	

Table 2 Changing the Classification Scheme to Eliminate Small Cell Sizes by Changing Data Ranges

**Changing the Classification Scheme by Changing the Cut-Points for Data Ranges**

Gender variable	Bottom-coding			Top-coding	
	Under 12	12 to 15	16 to 19	20+	Total
Males	23	20	18	19	80
Females	2	5	7	6	20
Total	25	25	25	25	100
<b>Change cut points to eliminate small cell size:</b>					
	Under 13	13 to 16	17 to 20	21+	Total
Males	26	20	19	15	80
Females	5	5	5	5	20
Total	31	25	24	20	100

Step 2:  
Establish  
standard de-  
id format

---



# What kind of format is good?

- Easy to distinguish in the text
  - Better to be closed on both word side.
  - You can special utf-8 char as the boundary char
- Use the format to preserve information as much as possible
  - E.g. `**DATE:2555**` is better than `***DATE*`
- Format should support coreference
  - E.g. if the same name appear a lot inside original text they can use the same id  
`**NAME:<id>**`

# Step 3: build automation tool

The screenshot shows the 'De-Identify Messages' application window. It features a menu bar with 'FILE' and 'TOOLS', and a toolbar with 'View Example' and 'De-identify' buttons. The interface is divided into several sections:

- Fields:** A list of X-Path expressions for de-identification, including `/ClinicalDocument/recordTarget/patientRole/patient/name/...`. A callout '1. Load XML documents' points to the top of this section.
- Value Generator:** A configuration panel in 'Advanced Mode' for generating test data from an Excel file. It includes fields for 'File', 'Worksheet', and 'Column', and options for 'Generate' (Random values or Sequential list). A callout '2. Add & edit de-identification rules' points to the 'Add...' button.
- Message Examples:** A comparison of 'Original' and 'De-identified' XML messages. The 'Original' message shows a name 'Adam' and family 'Everyman'. The 'De-identified' message shows the name changed to 'Kelvin' and family to 'Rosario'. A callout '3. Click to check result in the Message tab' points to the 'View Example' button, and another callout '4. Click to process all XML documents selected at step 1' points to the 'De-identify' button.

A 'De-identify Field' dialog box is also visible, showing a 'Find' button and 'Ctrl+F' shortcut.

# Algorithm we could use

Dictionary Look up

Regular Expression

Rule Based Engine

NLP+DL model

- Name Entity Recognition model:  
<https://arxiv.org/abs/1603.01360>
- Using attention model which can discover new name entity which not appear in dictionary
- DL Coreference model:  
<http://nlp.seas.harvard.edu/papers/corefmain.pdf>

Step 4:  
integration

---





# Integration



Integration with NLP  
algorithm pipeline

Generate  
tokenized  
text instead  
of plain text



Ready for  
information leak and  
re-run de-id process

Only run de-  
id to word  
with problem



Improve the  
performance

Use index  
wisely

# Reference

- <https://nvlpubs.nist.gov/nistpubs/ir/2015/nist.ir.8053.pdf>
- <https://healthitsecurity.com/news/de-identification-of-data-breaking-down-hipaa-rules>
- <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>
- <https://arxiv.org/abs/1603.01360>
- <http://nlp.seas.harvard.edu/papers/corefmain.pdf>
- <https://www.talix.com/>
- <https://healthfidelity.com/>
- <https://www.linguamatics.com/>