Hybrid Approach for De-identification of Texts

Jean-Philippe Goldman, Christian Lovis

WHAT IS DE-IDENTIFICATION ?

Objectives: hide personal data of directly **identified** people (*Mr Johnny Grizzly*) or **identifiable** people (*the woman aged 102 in this village, la ville du bout du lac*) people for data sharing

Reasons: protective + legal + ethical
Make the difference between
- ➔ identifying data / sensitive data (political, racial, medical, legal…)
- ➔ direct identifiers / quasi-identifiers
- ➔ pseudonymisation, anonymisation (irrevocable)

Main goal here:
- **Combine approaches for better scores**
- **Extend to personal health identiers (PHI)**

REGULATIONS

HIPAA - US Health Insurance Portability and Accountability Act regulation (US)
RGPD - Règlement Général sur la Protection des Données (EU)
CNIL Commission nationale de l'informatique et des libertés (F)
LPD - Loi fédérale sur la protection des données (CH)

According to **HIPAA**, 18 types of PHI (protected health information):
- **Names**
- All **geographic** subdivisions smaller than a state (street, city, county, precinct, ZIP code, equivalent geocodes)
- All elements of **dates** (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age
- **Telephone** & Fax numbers
- **Device** identifiers and serial numbers
- **Email** addresses
- Social security **numbers**
- Internet Protocol **(IP) addresses**
- **Medical record** numbers
- **Biometric** identifiers, including finger and voice prints
- **Health plan** beneficiary numbers

HOW TO DE-IDENTIFY ?

1. Find all PHIs (Personal Health Identifiers)

This task includes NER (named entity recognition),
i.e. ***names, organizations, locations***
e.g: *Paris Hilton, Hotel Hilton, Paris*

but also finding : ***time expressions, monetary values, quantities, percentages, symptom, pathology, medical exam, dosage…***

How ? **Several possible approaches :**

1. knowledge-based
(we use some a priori knowledge, e.g in metadata)
+ right-to-the-point
- infrastructure-dependant

2. rule-based
(we create recognizable patterns)
+ high precision
+ correctable, reproducible and shareable rules
+ explainable rules and results
- language & document dependent
- time-consuming development and support

3. AI-based (we use a lot of data to train neural networks)
+ state-of-the-art, best performances
+ few manual work
- reproducible
- explainable
- needs a lot of data (and time) to train or fine-tune

4. combined approach (we use results from previous)
- Combine the **best of** each approach (better precision)
- Combine all approaches (better recall)

2. Replace PHIs

- **masking identifiers** ⇒ replace by
  - credible surrogate information (pseudos John Doe, time shifting)
  - or IDs (id_12)
  - or placeholders (<Person>)
  - or sanitize (ie. redact, blacken, XXX)
- **generalizing quasi-identifiers** (k-anonymity 1998, L-diversity 2007)
  - reduce precision by ranges
  - aggregate marginal values (e.g 90+ y.o.) or suppress or blur/noisen

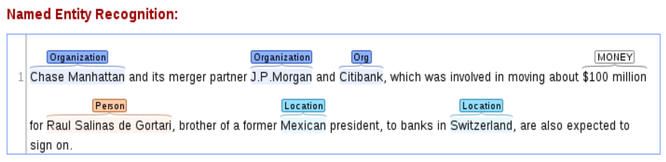
MANUAL ANNOTATION

Manual annotation is crucial to build reference data to **evaluate** the various approaches and to **train** Machine-Learning based techniques.

It is time consuming, needs Ethical Commitee approval and a strong methodology (unambiguous instructions, good and motivated annotators, organizing and monitoring parallel annotations)

Example with brat interface:

Named Entity Recognition:



STUDY

3 annotated corpus
- **Medina** (58k words – various sources)
- **WiNer** (281k words – wikipedia)
- **Simed1050** (891k words - clinical notes)

7 tools
- **spacy, flair, stanza** (AI-based, popular, open-source)
- **deid.unitex** (rule-based, developped at HUG)
- **deid.kb** (knowledge-based, developped at HUG)
- **best-of , maxi-plus** (hybrid approach, developped at HUG)

Results:

Names ➔ **stanza** performed best (0.92) on names for **medina** and **winer** corpus, but as expected **deid.kb** outperforms for **simed1050** corpus (0.95)
Places ➔ **deid.unitex** performed best (0.70) on **medina**, but spacy is 1st on winer (0.94)
Dates ➔ only recognized by **deid.unitex** (0.96 on **medina**, 0.82 on **winer**, 0.71 on **Simed1050**)

Results vary a lot according to corpus and tool. Combining all approaches (**best-of, maxi-plus**) give the best precision and recall, respectively.

SIMED@HUG SIMED@UNIGE

UNIVERSITÉ DE GENÈVE FACULTÉ DE MÉDECINE — Hôpitaux Universitaires Genève — SIMED