Towards Contextualizing Sensitive Data Detection

Madelon Hulsebos

UNECE Expert Meeting on Statistical Data Confidentiality

October 16, 2025

Talk on the paper:

Detecting Contextually Sensitive Data with AI
by Telkamp, Rabier, Teran, and Hulsebos
UNECE Expert meeting on Statistical Data Confidentiality

Funding:

Funded by the European Union
NextGenerationEU

Team

Liang

Madelon

Researcher in (neural) **tabular AI**: semantics, retrieval, querying, predicting.

Alex

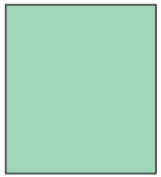
Godfrey

Javier

Jos

Melanie

Metasebya



Nafissatou

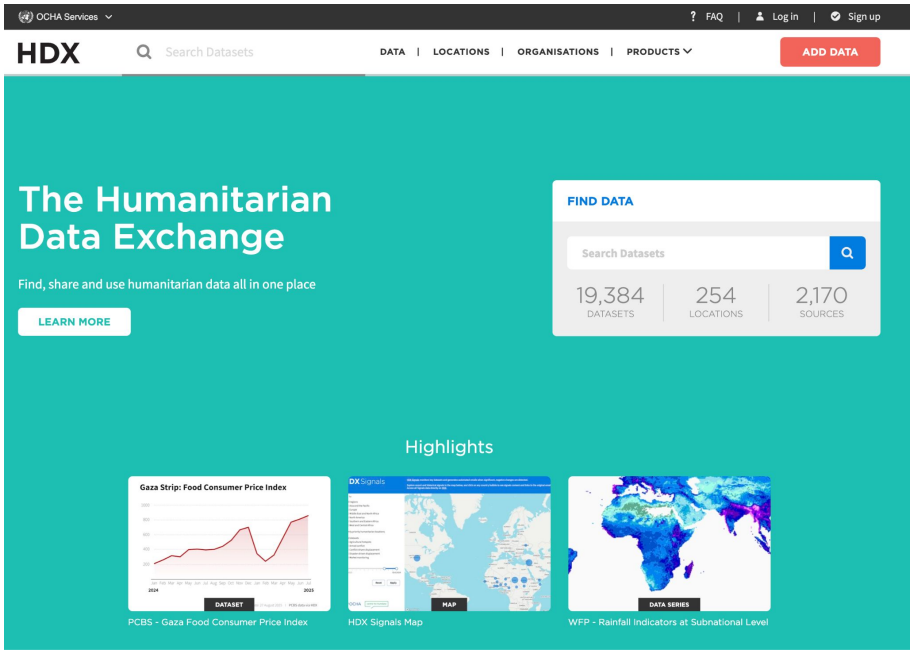
Shah

Sarah

Serban

HDX: The UN's Humanitarian Data Exchange

How to facilitate data sharing, **responsibly**?



humdata.org

How to protect sensitive data? Part 1.

Sensitive data: private financial and personal information, intellectual property and proprietary corporate data, which requires protection from unauthorised access, use, disclosure, interruption or alteration.*

Yet, most focus is on personal identifiable information?

*Inferred from “Information Security” def. in “Dictionary of Privacy, Data Protection, and Information Security” (Elliot et al., 2024).

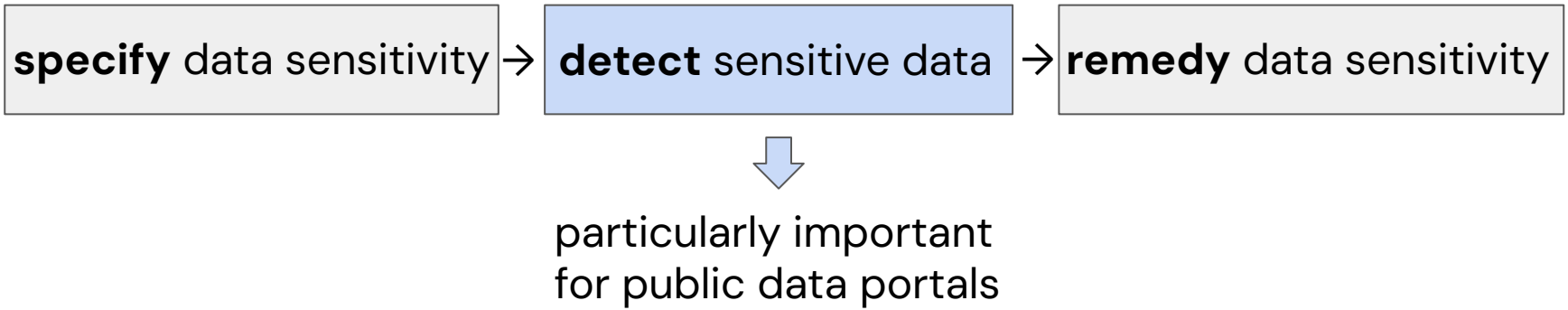
How to protect sensitive data? Part 2.

***Contextual* sensitive data** reflects information whose sensitivity depends on external factors and requires protection due to how, by whom, and in what context it can be misused.*

Context is key.

*Telkamp & Hulsebos 2025, inspired by Nissenbaum et al. 2004 & Kober et al. 2023.

How to protect sensitive data? Part 3.



We understand and can remedy data sensitivity, but **tools to detect** data sensitivity are limited

① Limited **specificity**:

Existing tools are too generic → overly sensitive → many false alarms.

② Limited **scope**:

Current *tools for detecting* sensitive data are focused on detecting PII.

The solution

context

and a bit of LLM magic

① **limited specificity**: type contextualization

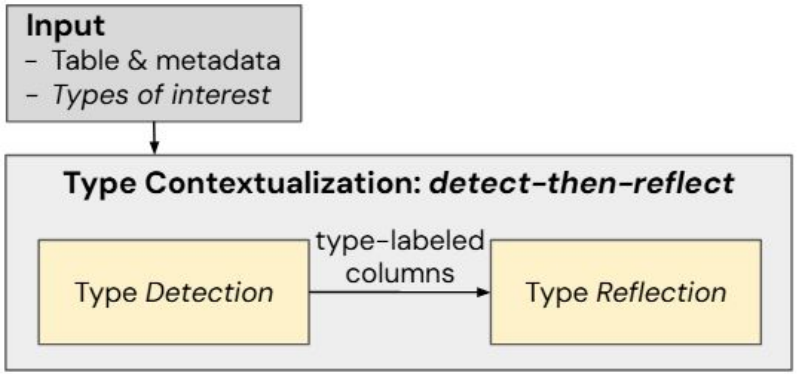
💡 Sensitivity of column types (e.g. PII) depends on **column context**.*

`Names` and `addresses` of humanitarian workers in conflict zones → high risk!
`Names` and office `addresses` of professors → low risk (public info).

Type contextualization mechanism:

Step (1): **Detect** all potential sensitive columns based on their types with LLM.

Step (2): **Reflect** with LLM on sensitivity given *entire* table.



*We focus on tabular data.

Results on PII detection with type-contextualization

Existing tools: **low precision = many FPs**, and **mediocre recall = much PII undetected**

System / Model	No Reflection			With Reflection		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Google DLP	0.531	0.628	0.576	–	–	–
Presidio	0.520	0.618	0.565	–	–	–
Ground-truth PII	0.527	**1.000**	0.690	–	–	–
GPT-4o-mini	0.856	0.639	0.732	**0.938**	0.632	0.755
Gemma 2 9B	0.740	0.819	0.778	0.800	0.792	0.796
Gemma 3 12B	0.487	0.941	0.642	0.753	0.806	0.779
Qwen3 8B	0.742	0.868	0.800	0.749	0.868	0.804
Qwen3 14B	0.565	0.972	0.714	0.732	0.941	0.824
Aya Expanse 8B	0.812	0.674	0.736	0.812	0.674	0.736
Qwen3 8B FT → GPT-4o-mini	–	–	–	0.902	0.861	**0.881**

Precision and recall on PII-annotated real datasets from GitHub

② **limited scope**: domain contextualization

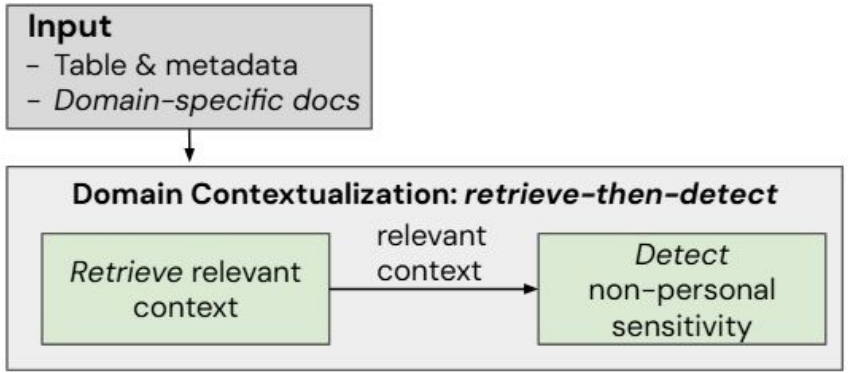
💡 Sensitivity of *non-personal* data often depends on **domain context**.

`Geo-coordinates` from *attacked* hospitals in Gaza → high risk!
`Geo-coordinates` from hospitals in Germany → low risk.

Domain contextualization mechanism:

Step (1): **Retrieve** relevant *sensitivity rules* from domain-specific documents (ISPs)

Step (2): **Detect** sensitivity of columns using LLM reasoning on column + retrieved rules



Results on non-PII with domain-contextualization

System / Model	No Domain Knowledge Prec.	Rec.	F1	With Domain Knowledge Prec.	Rec.	F1
All-tables-sensitive (baseline)	0.375	1.000	0.545	–	–	–
GPT-4o-mini	0.474	1.000	0.643	0.692	1.000	**0.818**
Gemma 2 9B	0.375	1.000	0.545	0.429	1.000	0.600
Gemma 3 12B	0.529	1.000	0.692	0.500	1.000	0.667
Qwen3 8B	0.562	1.000	0.720	**0.778**	0.778	0.778
Qwen3 14B	0.562	1.000	0.720	0.643	1.000	0.783
Aya Expanse 8B	0.450	1.000	0.621	0.500	1.000	0.667

Based on expert-informed synthesized humanitarian datasets

Qualitative review by QA officer of context-grounded LLM explanations:

> *"I noticed the LLM explanation column and have to say it already looks great."*

Beyond “classifying” sensitivity: grounded explanation

Reported location on NFIs suppliers							
nfi_supplier_	nfi_supplier_loc- value -from_Iran	nfi_supplier_	nfi_supplier_	nfi_supplier_	nfi_supplier_	nfi_supplier_	nfi_supplier_
NA	0%	0%	0%	100%	0%	0%	0%
NA	0%	0%	0%	33%	0%	67%	0%
NA	0%	0%	0%	100%	0%	0%	0%

GPT-4o-mini

There are no indications in the ISP that suggest this specific information about supplier locations is sensitive, as it does not involve personal data or any identifiable information. ~

*Cited ISP Rule(s): The column aligns with the guidance under \"**low/no sensitivity**\" as it represents operational"* ✔

Gemma-3

*1. The column **"Reported location on NFIs suppliers - nfi_supplier_loc- value -from_Iran" indicates the origin of NFIs suppliers**.*
*2. This information, even with country-level disaggregation, **doesn't reveal personal or sensitive data. It relates to the location of suppliers, which is generally not considered sensitive.*** ✔
*3. It aligns with the ISP's guideline on "Facility data (e.g., health, education, water points) at national or regional level, unless explicitly restricted by clusters." **Since this column is about locations of suppliers and not about individuals or specific operations,** it is categorized as **non-sensitive.*** ✔

Aya Expanse

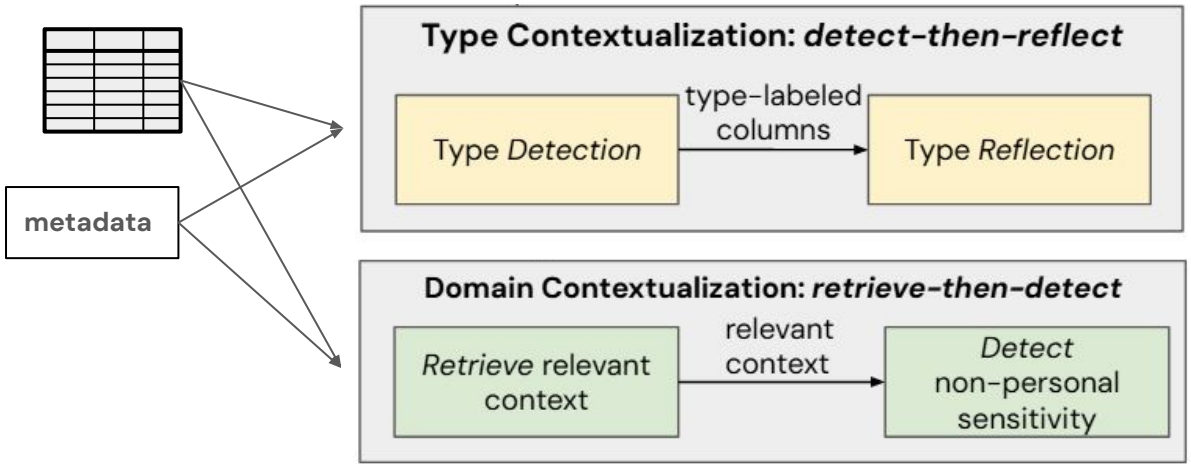
Sensitivity Level: MEDIUM_SENSITIVE ✗

*Context in the table: There's no other information in the table that explicitly flags this data as highly sensitive. **However, the ISP guidance emphasizes the need to be cautious with any data that could indirectly reveal information about aid recipients, beneficiaries,** or …* ✗

Summarizing

How can we **detect *all* sensitive data** of *any* type while **minimizing false positives**? We build on **the contextual definition of sensitive data**.

💡 specific contextual LLM mechanisms >> existing tools and generic LLM calls



94% of personal data detected, compared to **63%** with Google DLP (also less FPs!)

Perfect recall, okeyish precision; but context-grounded reflection *improves human evals in consistency*.

Thank you!

madelon@cwi.nl

https://github.com/**trl-lab/sensitive-data-detection**