

PP13: Building a scalable, ML-driven anonymization pipeline for fast, efficient clinical document publication

INTRODUCTION

Clinical trial transparency is a term that includes the **public release** of **clinical study documents** through regulators such as European Medicines Agency and Health Canada.

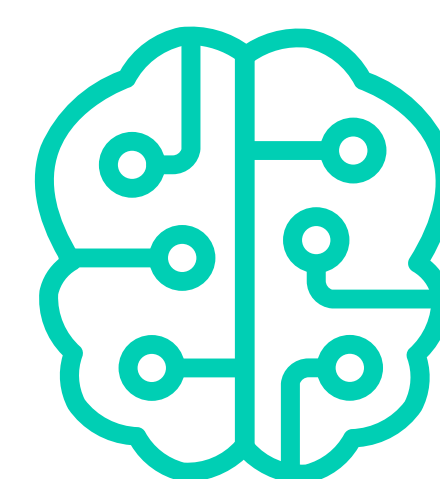
In our process, documents pass through an **anonymization pipeline** where **personal information (PI)** is captured and **anonymized** to protect the privacy of data subjects.

An important early step in the pipeline is **detection of PI in unstructured clinical text** (see right), which can then be transformed to protect privacy.

Several characteristics of clinical documents make **efficient detection of PI at scale** a **challenging problem** to solve:

-  Clinical documents can reach upwards of **100k pages** and submissions can have **tight turnaround times** (e.g., 60 days).
-  **Medical histories** can appear **in many forms and contexts**, making “classical” techniques to find PI (e.g., regular expressions) **difficult to scale**.
-  Some types of personal information (e.g., names) **may only appear a few times** per clinical document, making them **harder to catch**, but **important not to miss**.

Here, we describe **case studies** in which **machine learning (ML)** enabled us to handle these challenges. In each case, we used **different technologies** for **different sizes** of documents to **scale our anonymization pipeline**. While model performance may differ based on the nature of content in the documents, all three case studies achieved **very high scores for automation quality** by using techniques **calibrated to the document size and complexity**.



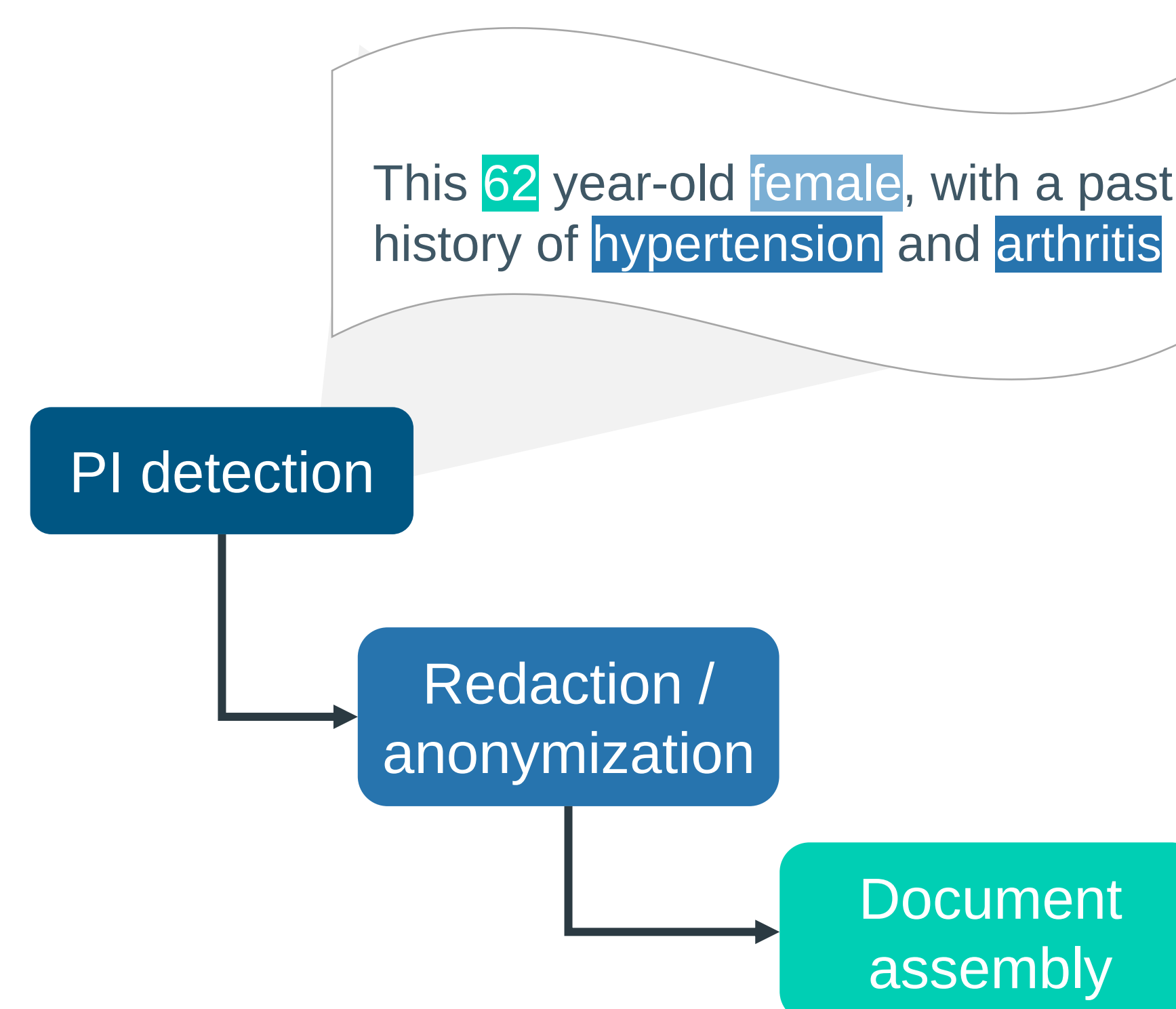
CASE STUDIES

Medium-sized document | Regex + human-in-the-loop PI detection

For medium-sized documents, the use of **regular expressions (regex)** with **active learning** allows **efficient PI detection**, as only **small amounts** of training data annotated by a **subject matter expert** are needed to capture most PI.

 **100+**
pages of narrative text
to be detected

 **200**
data subjects



Model performance & impacts:

97.0 **90.8** **93.8** **3** Iterations until model convergence
Recall Precision F-measure

Large document | Regex + pre-trained language model

For larger documents, **efficient PI detection** can be achieved using **transformer models** which have been trained in-house on **large volumes of data** annotated by **subject matter experts**.

 **4,000+** pages of narrative text to be detected  **1,000+** data subjects

Model performance & impacts:

97.8 **95.4** **96.6** **4.5** Total hours spent on detection
Recall Precision F-measure **7** Seconds per page on QC

Very large document | Tuning a pre-trained language model

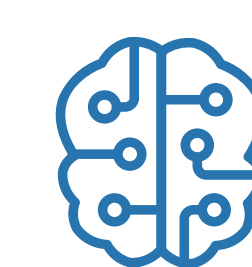
If model performance deteriorates for a **very large document** which is **rich with new forms of PI**, experts can **fine-tune the model** using a small amount of annotated data from the same document to achieve **efficient PI detection**.

 **~ 500** pages of training data used for fine-tuning  **87,000+** pages of narrative text to be detected  **4,300+** data subjects

Model performance & impacts:

95.0 **93.0** **94.0** **25%** Improvement in model recall (relative to pre-trained model) **50%** Reduction in QC time (relative to pre-trained model)
Recall Precision F-measure

CONCLUSIONS



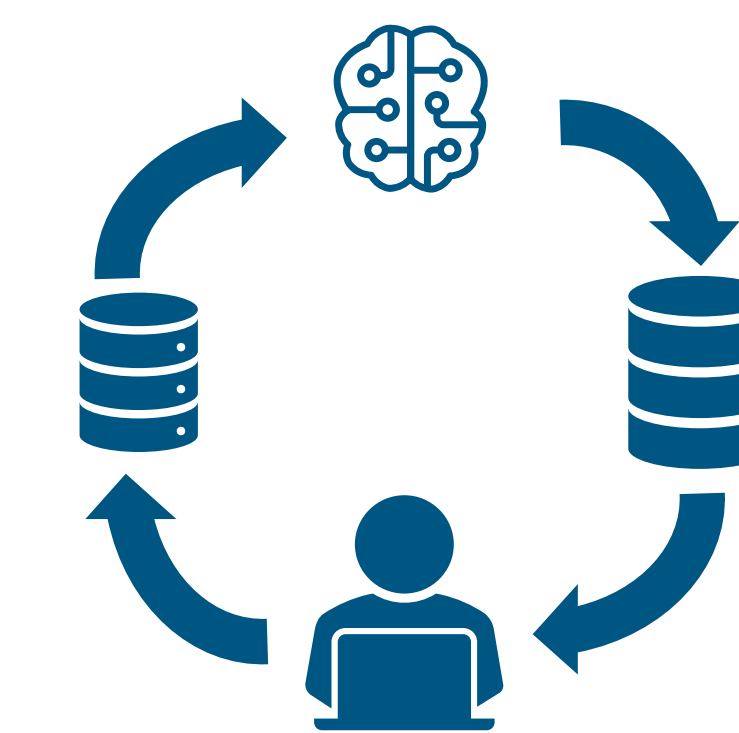
Expert-driven ML can allow for **accurate PI detection** for anonymization with notable **gains in efficiency and throughput**



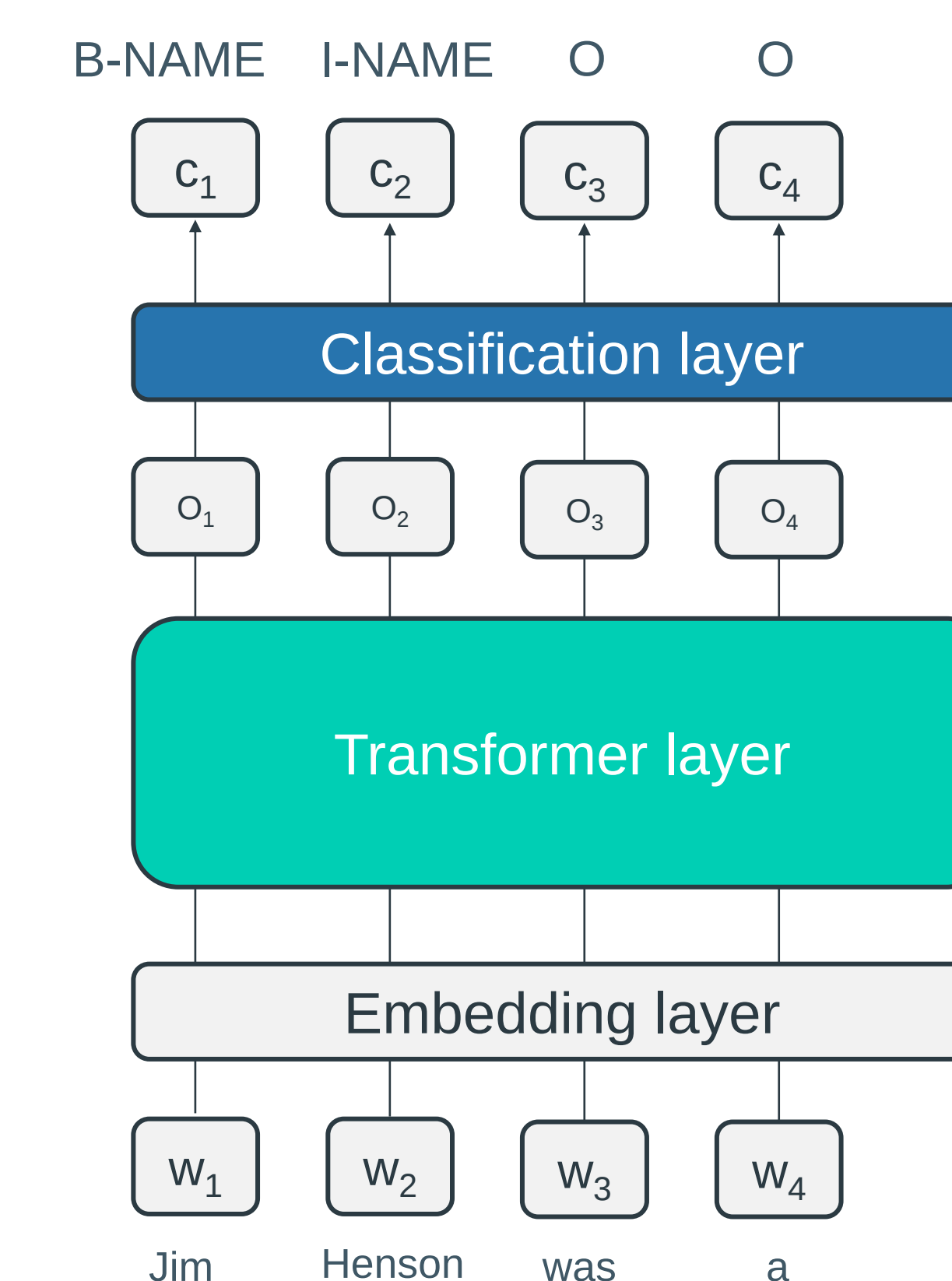
Subject matter expertise is **important** to guide automated systems towards **accurate PI detection** to **ensure maximum transparency**



Efficient PI detection requires ML solutions that are **adaptable** to the **size and complexity** of the documents to be anonymized



In **active learning**, the model training algorithm iteratively presents a human expert with data it is unsure about. The expert **improves the model** by correcting its mistaken predictions for the data.



An example transformer architecture for detecting **personal information** in **unstructured text**.



David Di Valentino
AI/ML Solutions Lead
ddivalentino@privacy-analytics.com



Muqun (Rachel) Li
NLP Lead
rli@privacy-analytics.com