Building a scalable, machine learning-driven anonymization pipeline for clinical trial transparency

PEPR 21
June 11, 2021

David Di Valentino, PhD | Muqun (Rachel) Li, PhD

AI/ML Solutions Lead, Data Scientist

NLP Lead, Senior Machine Learning Engineer

Table of contents

What is clinical trial transparency (CTT)?

A brief history of clinical trial transparency

Clinical trial transparency is a term that covers the **sharing and use of anonymized clinical trial data** through global public registries or sponsor-supported channels.

2015

First year that clinical documents were shared publicly by the *European Medicines Agency* (EMA)

681

Total number of datasets shared to date by EMA, Health Canada, YODA

Goals of clinical trials transparency include …

Building **public trust** in medical, pharmaceutical research

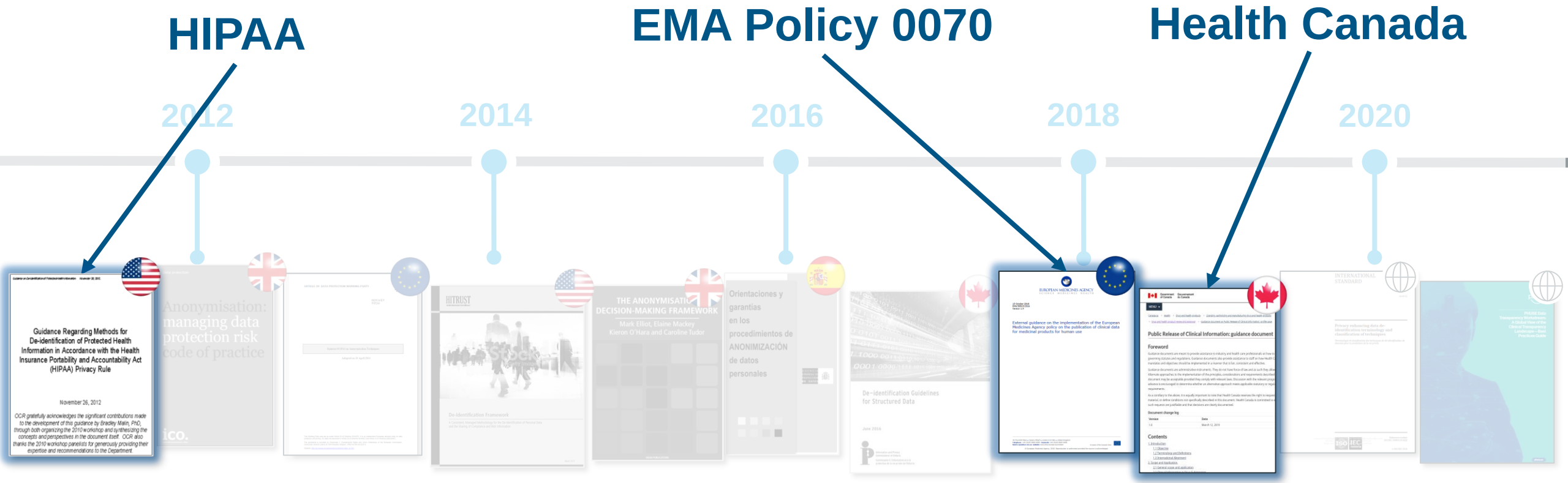
Driving **healthcare outcomes and innovation** through secondary research

PRIVACY ANALYTICS
an IQVIA company

IQVIA

A focus on healthcare data

In practice, our document anonymization methodology centers around regulations covering ***healthcare data privacy***.

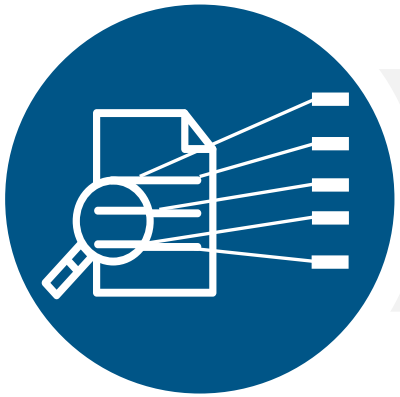


PRIVACY ANALYTICS
an IQVIA company

IQVIA

What is “anonymization” in this context?

Specific guidance **varies between regulators**, but broadly speaking, we can consider **anonymized data** to have the following properties:



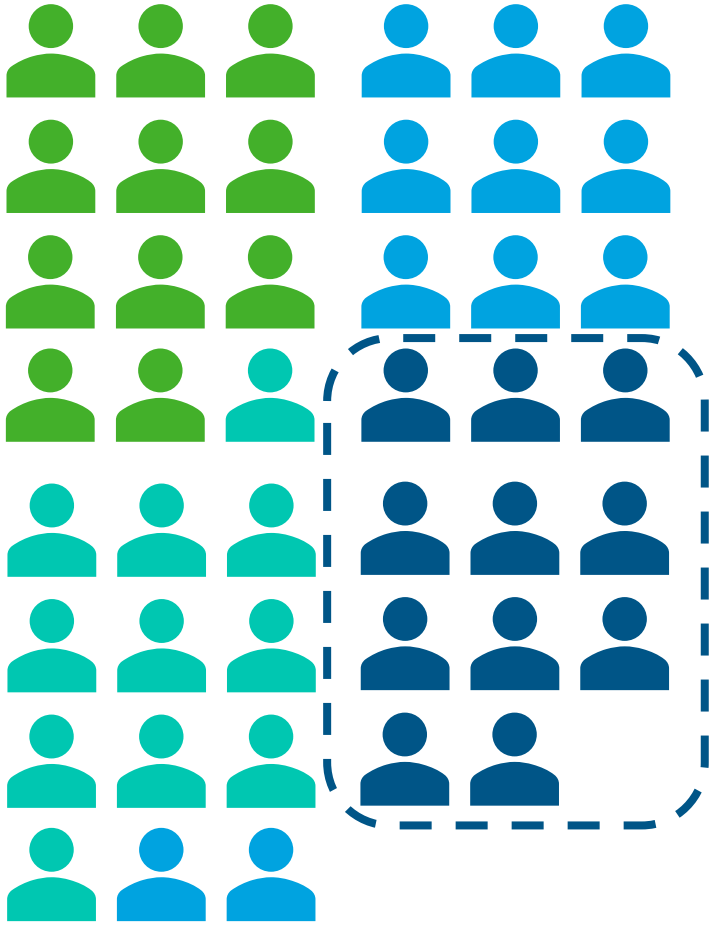
Personal information is captured in the data and transformed in a way that renders subjects **non-identifiable**.

Identifiability can be quantified using a **disclosure risk metric** and determined by an **expert** to be below a specified **threshold**.

PRIVACY ANALYTICS
an IQVIA company

IQVIA

What does it mean for public documents to be “anonymized”?



To facilitate the **public release** of clinical documents, we typically require each data subject to be **statistically similar to least 10 other individuals.**

If this condition is met, we would consider the data **anonymized**.

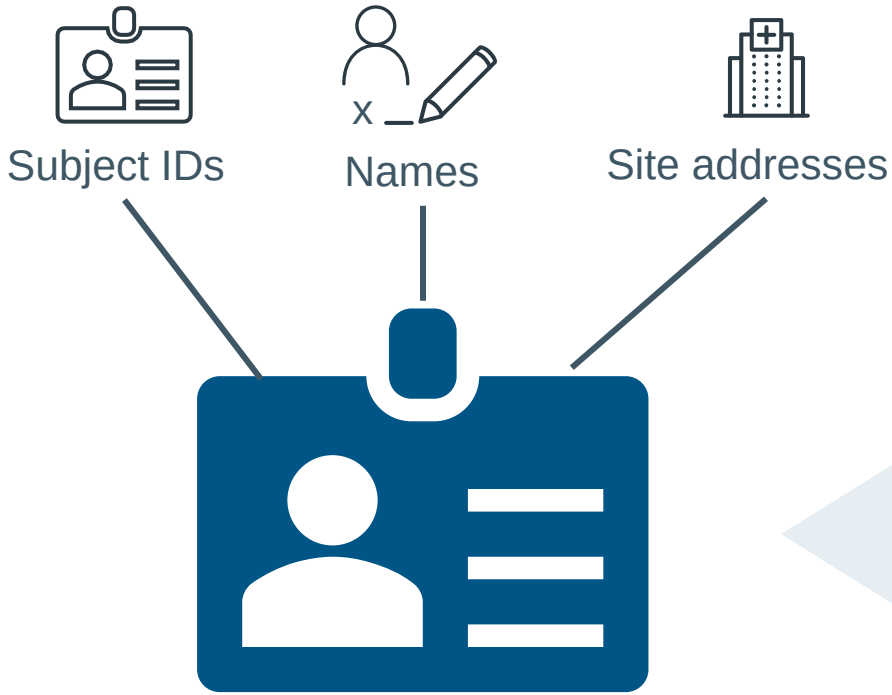
(See EMA/90915/2016 Section 5.4.1, Health Canada “Public Release of Clinical Information: guidance document” Section 6.2)

PRIVACY ANALYTICS
an IQVIA company

IQVIA

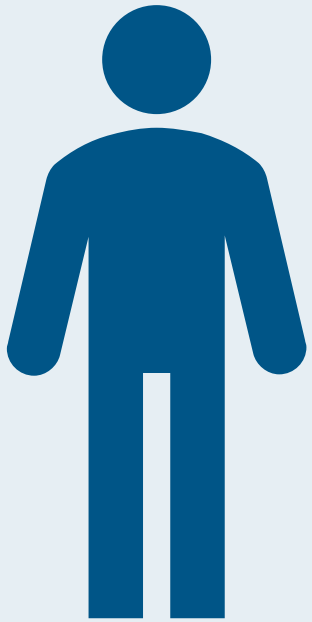
Practical challenges with detecting personal information in clinical documents

What kind of identifying information do we encounter?



Typical indirect identifiers include …

Indirect identifiers



- Age (or date of birth)
- Date of death *
- Vital signs data (e.g., height, weight)
- Medical details
- Notable calendar dates
- Adverse event information

* If applicable

PRIVACY ANALYTICS
an IQVIA company

IQVIA

An example of clinical text

Patient ID:	59810-6001824
Sex:	F
Age:	51
Weight (kg):	88.9
Height (cm):	163
BMI:	33.5
Treatment Start Date:	18FEB99
Treatment End Date:	20FEB99
Start Date of Adverse Event:	20FEB99
End Date of Adverse Event:	23FEB99
Reason for Narrative:	CELLULITIS
Intensity:	SEVERE
Relation to Study Drug:	UNRELATED

Patient 6001824:

Medical History

Term	Body System	Ongoing at Entry/Screening
HYPERTENSION	VASCULAR	YES
ARTHRITIS	ALLERGIC/IMMUNOLOGIC	YES

Adverse Event

This 51 year-old female, was a past history of hypertension and arthritis, complained of influenza-like symptoms which included fever of 100.4ºF, chills, body ache, rhinitis and an infected throat, and was randomized to the study. On Day 2, the patient stopped in to see the doctor and said she felt better, but had a rash on her left ankle. The patient said she had noted the rash the day before entering the study but had failed to mention it. The patient was diagnosed with dermatitis and Diprosone cream was prescribed. On FEB 20 (Day 3), the patient had a fever of 104ºF and was hospitalized and discontinued from study. Cellulitis was diagnosed and the patient received IV fluids and Ancef. The patient was discharged from the hospital on FEB 23 (Day 6) on Keflex 500 mg tid for one week and her usual medications (Tiazac 240 mg daily and Accupril 20 mg daily). Both condition and prognosis at time of discharge was listed as good.

Some **common features** of clinical text:

- Multiple **data formats** (i.e., paragraphs, tables)
- Several **distinct types** of personal information present
- Different **representations** of personal information
- **Domain-specific** language

An example of clinical text

Patient ID:	59810-6001824
Sex:	F
Age:	51
Weight (kg):	88.9
Height (cm):	163
BMI:	33.5
Treatment Start Date:	18FEB99
Treatment End Date:	20FEB99
Start Date of Adverse Event:	20FEB99
End Date of Adverse Event:	23FEB99
Reason for Narrative:	CELLULITIS
Intensity:	SEVERE
Relation to Study Drug:	UNRELATED

Patient 6001824:

Medical History

Term	Body System	Ongoing at Entry/Screening
HYPERTENSION	VASCULAR	YES
ARTHRITIS	ALLERGIC/IMMUNOLOGIC	YES

Adverse Event

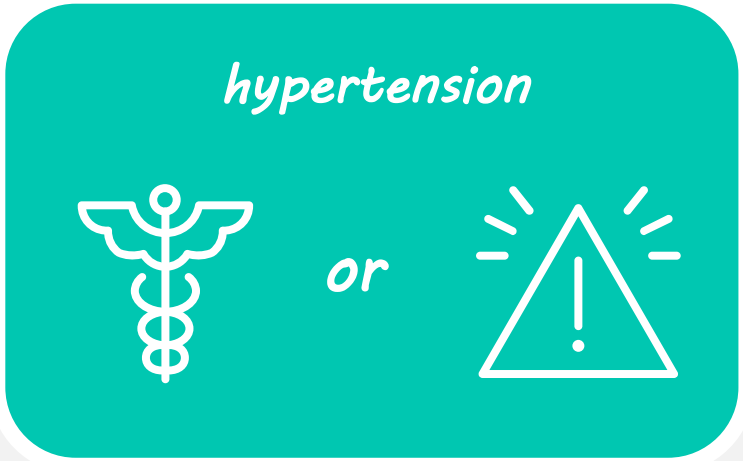
This 51 year-old female, was a past history of hypertension and arthritis, complained of influenza-like symptoms which included fever of 100.4ºF, chills, body ache, rhinitis and an infected throat, and was randomized to the study. On Day 2, the patient stopped in to see the doctor and said she felt better, but had a rash on her left ankle. The patient said she had noted the rash the day before entering the study but had failed to mention it. The patient was diagnosed with dermatitis and Diprosone cream was prescribed. On FEB 20 (Day 3), the patient had a fever of 104ºF and was hospitalized and discontinued from study. Cellulitis was diagnosed and the patient received IV fluids and Ancef. The patient was discharged from the hospital on FEB 23 (Day 6) on Keflex 500 mg tid for one week and her usual medications (Tiazac 240 mg daily and Accupril 20 mg daily). Both condition and prognosis at time of discharge was listed as good.

In this text alone we have representations of:

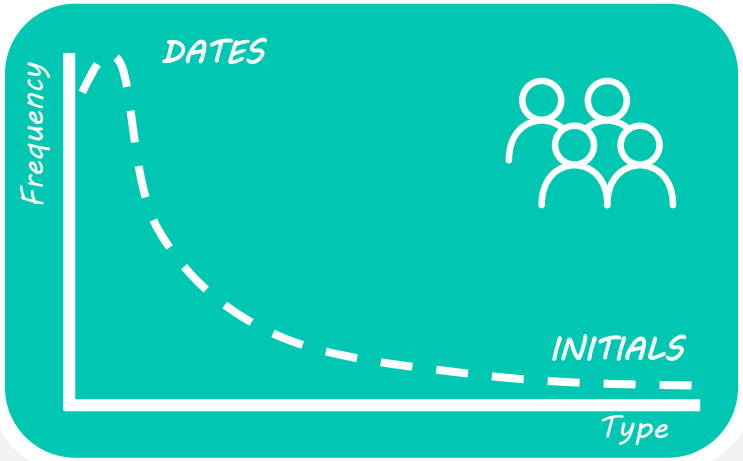
- **Subject ID number**
- **Age**
- **Weight**
- **Height**
- **BMI**
- **Dates**
- **Medical history terms**

Scalability challenges in clinical documents

Clinical documents can reach upwards of **100k pages** of detailed information on **thousands of study participants**.



Clinical language is highly **domain specific**, with **common terminology** spanning multiple data types (e.g., medical histories vs. adverse events).



Personal information tends to be **long tailed**, with several types having only few examples per clinical document.

PRIVACY ANALYTICS
an IQVIA company

IQVIA

Applying AI/ML to detect personal information in unstructured text

AI/ML considerations for our use case

A **complicated workflow** can result in **dirty data** or **missed identifiers**, making privacy protection **more difficult**.

Our tooling is mainly used by **data analysts** who do not necessarily have a background in **machine learning** or **data science**.

To ensure adequate **privacy protection**, the **AI/ML tooling** that we build into our workflow must be,

Efficient at detecting personal information

and

Usable by non-experts

PRIVACY ANALYTICS
an IQVIA company

IQVIA

How to use AI for clinical document anonymization - rule-based Named Entity Recognition (NER)

Patient ID:	59810-6001824
Sex:	F
Age:	51
Weight (kg):	88.9
Height (cm):	163
BMI:	33.5
Treatment Start Date:	18FEB99
Treatment End Date:	20FEB99
Start Date of Adverse Event:	20FEB99
End Date of Adverse Event:	23FEB99
Reason for Narrative:	CELLULITIS
Intensity:	SEVERE
Relation to Study Drug:	UNRELATED

Patient 6001824:

Medical History

Term	Body System	Ongoing at Entry/Screening
HYPERTENSION	VASCULAR	YES
ARTHRITIS	ALLERGIC/IMMUNOLOGIC	YES

Adverse Event

This 51 year-old female, was a past history of hypertension and arthritis, complained of influenza-like symptoms which included fever of 100.4ºF, chills, body ache, rhinitis and an infected throat, and was randomized to the study. On Day 2, the patient stopped in to see the doctor and said she felt better, but had a rash on her left ankle. The patient said she had noted the rash the day before entering the study but had failed to mention it. The patient was diagnosed with dermatitis and Diprosone cream was prescribed. On FEB 20 (Day 3), the patient had a fever of 104ºF and was hospitalized and discontinued from study. Cellulitis was diagnosed and the patient received IV fluids and Ancef. The patient was discharged from the hospital on FEB 23 (Day 6) on Keflex 500 mg tid for one week and her usual medications (Tiazac 240 mg daily and Accupril 20 mg daily). Both condition and prognosis at time of discharge was listed as good.

Rule Based NER

Rules for Dates

- \d{2}[A-Z]{3}\d{2}
- [A-Z]{3} \d{2}
- Day \d{1,}
- ....

Rules for medical history terms

- HYPERTENSION|VASCULAR|ARTHRITIS|...
- (?<=a past history of)\w+(?= and)
- ....

How to use AI for clinical document anonymization - machine learning-based NER

Patient ID:	59810-6001824
Sex:	F
Age:	51
Weight (kg):	88.9
Height (cm):	163
BMI:	33.5
Treatment Start Date:	18FEB99
Treatment End Date:	20FEB99
Start Date of Adverse Event:	20FEB99
End Date of Adverse Event:	23FEB99
Reason for Narrative:	CELLULITIS
Intensity:	SEVERE
Relation to Study Drug:	UNRELATED

Patient 6001824:

Medical History

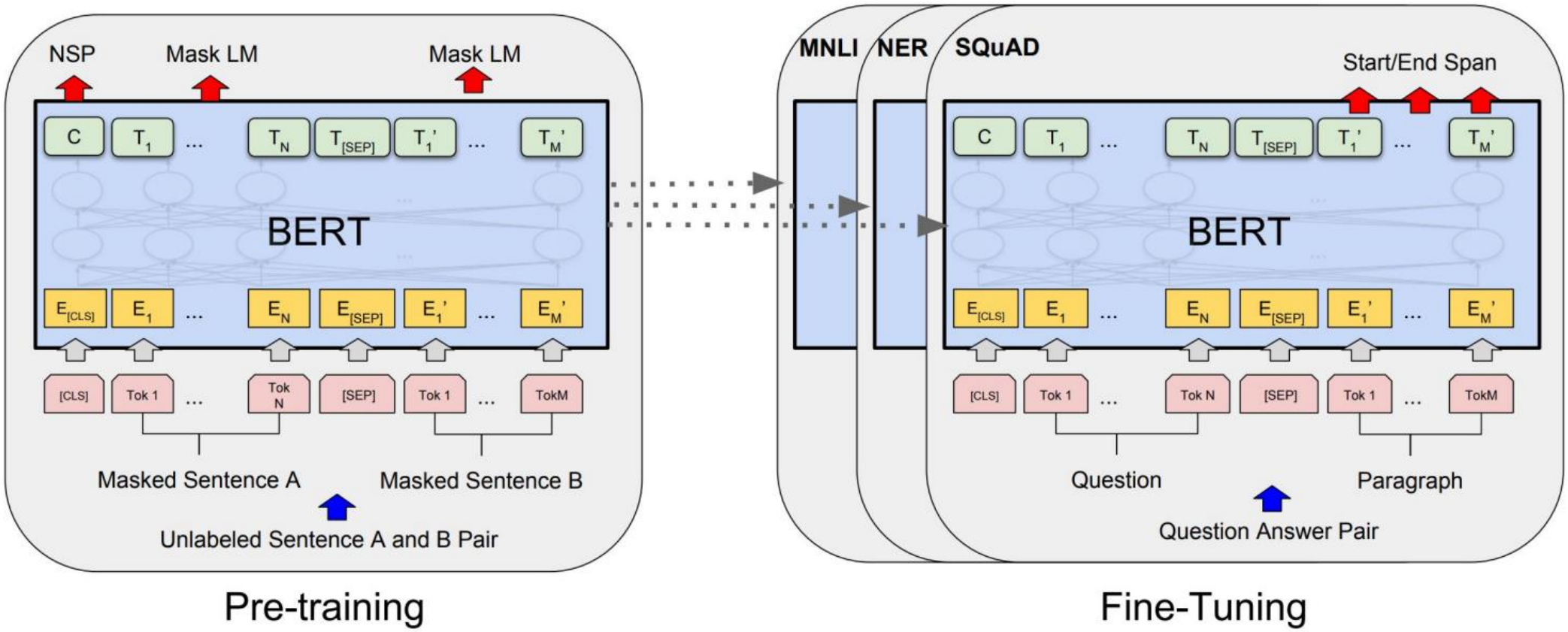
Term	Body System	Ongoing at Entry/Screening
HYPERTENSION	VASCULAR	YES
ARTHRITIS	ALLERGIC/IMMUNOLOGIC	YES

Adverse Event

This 51 year-old female, was a past history of hypertension and arthritis, complained of influenza-like symptoms which included fever of 100.4ºF, chills, body ache, rhinitis and an infected throat, and was randomized to the study. On Day 2, the patient stopped in to see the doctor and said she felt better, but had a rash on her left ankle. The patient said she had noted the rash the day before entering the study but had failed to mention it. The patient was diagnosed with dermatitis and Diprosone cream was prescribed. On FEB 20 (Day 3), the patient had a fever of 104ºF and was hospitalized and discontinued from study. Cellulitis was diagnosed and the patient received IV fluids and Ancef. The patient was discharged from the hospital on FEB 23 (Day 6) on Keflex 500 mg tid for one week and her usual medications (Tiazac 240 mg daily and Accupril 20 mg daily). Both condition and prognosis at time of discharge was listed as good.

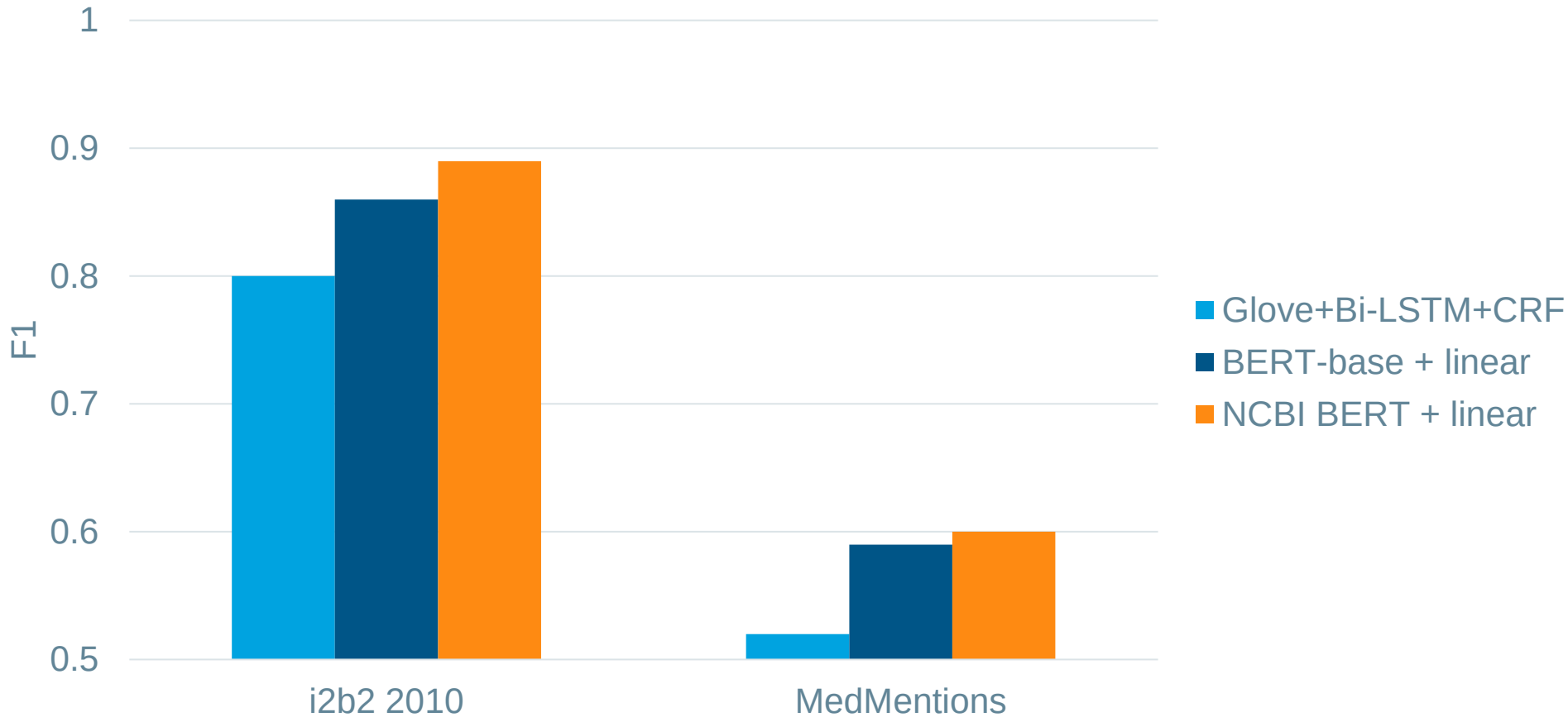
- Patient ID number
- Age
- Weight
- Height
- BMI
- Dates
- Medical history terms

Transfer learning for NER with pretrained language models



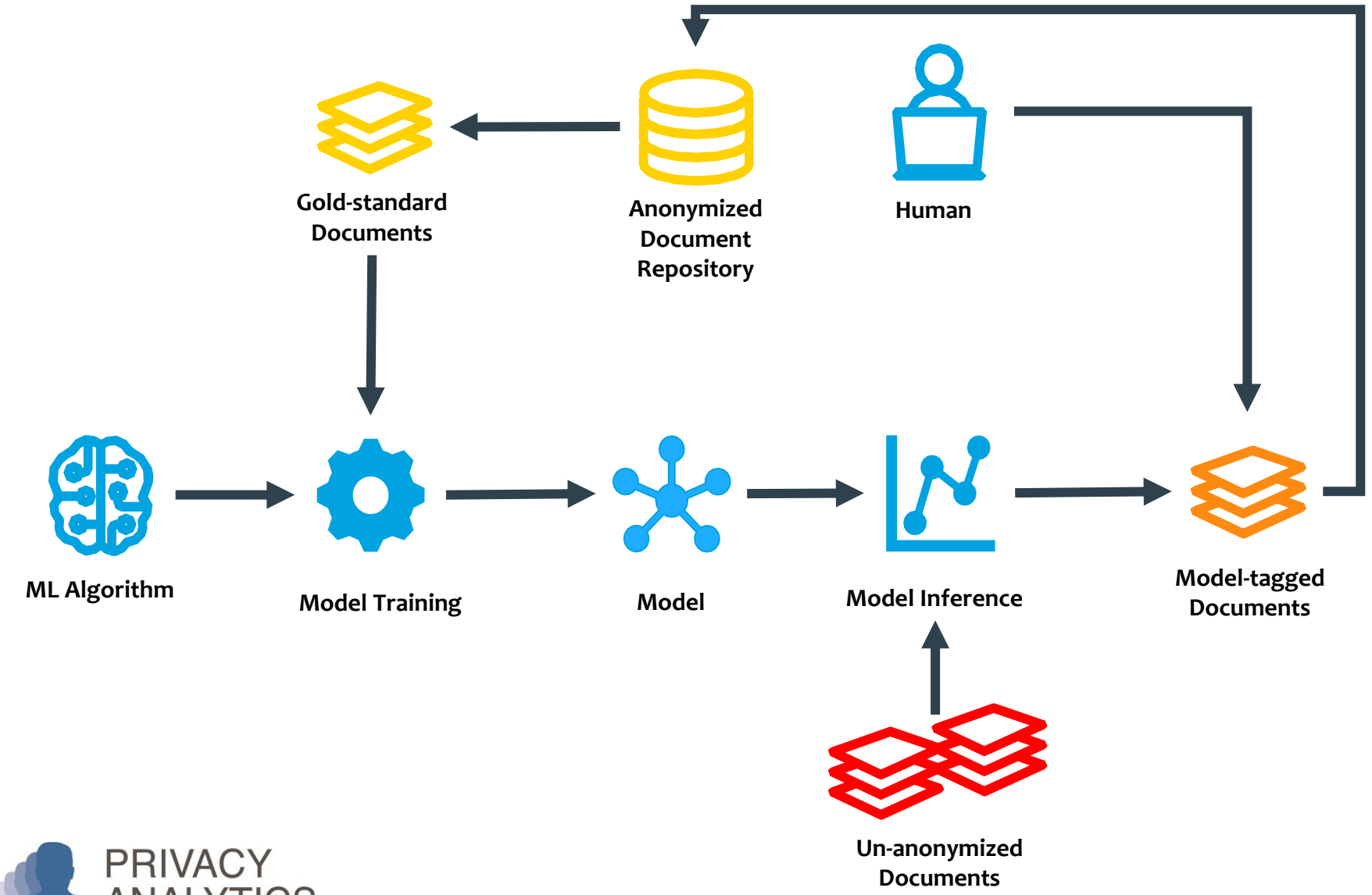
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Recognizing medical entities with domain-specific language models



Nejadgholi, Isar, Kathleen C. Fraser, Berry De Bruijn, Muqun Li, Astha LaPlante, and Khaldoun Zine El Abidine. "Recognizing umls semantic types with deep learning." In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pp. 157-167. 2019.

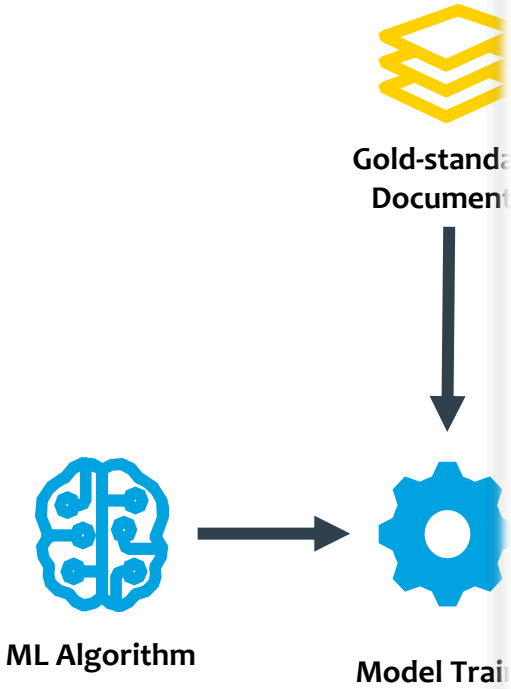
Traditional machine learning-based document anonymization workflow

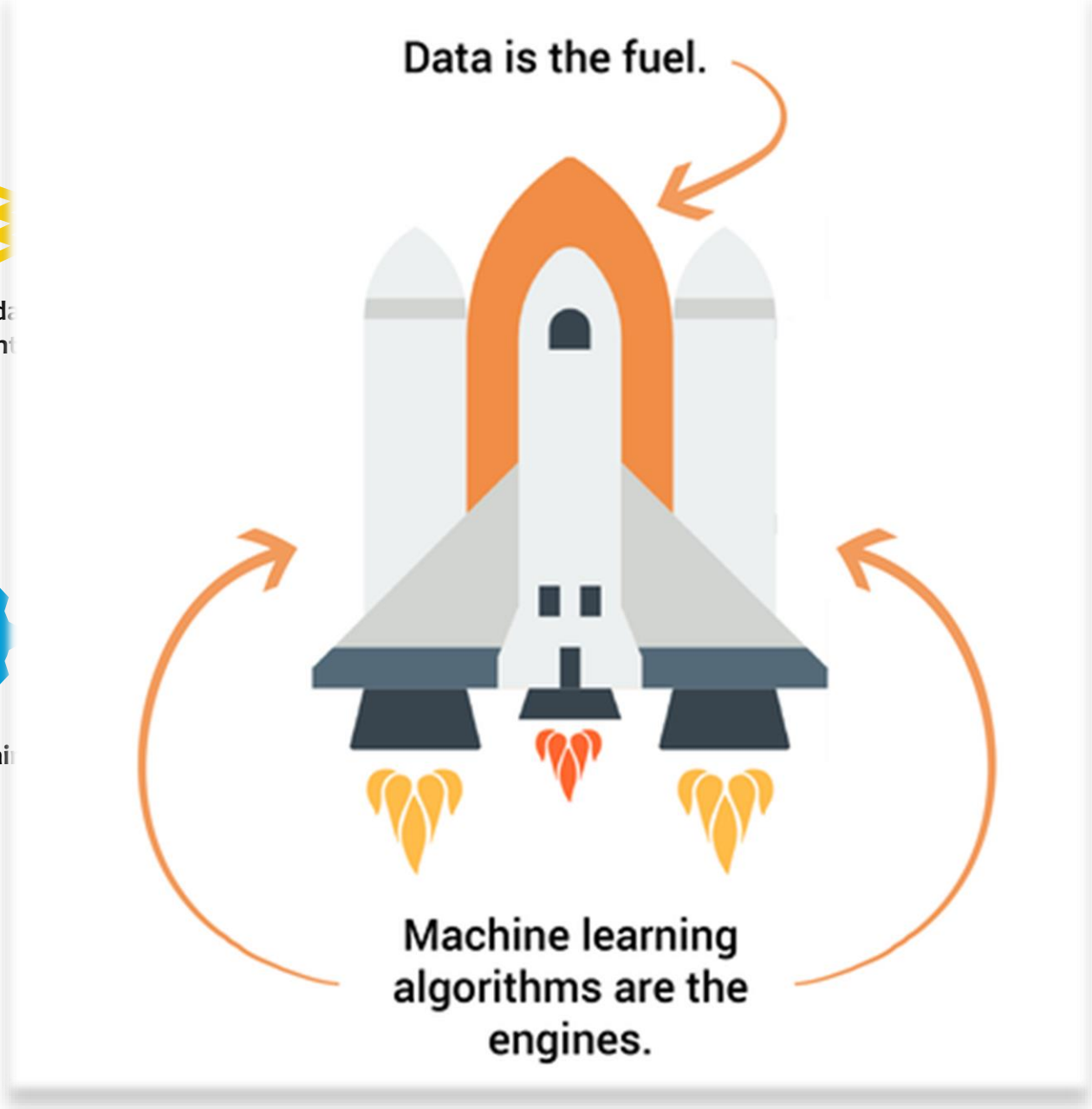


PRIVACY ANALYTICS
an IQVIA company

IQVIA

Traditional machine learning-based document anonymization workflow

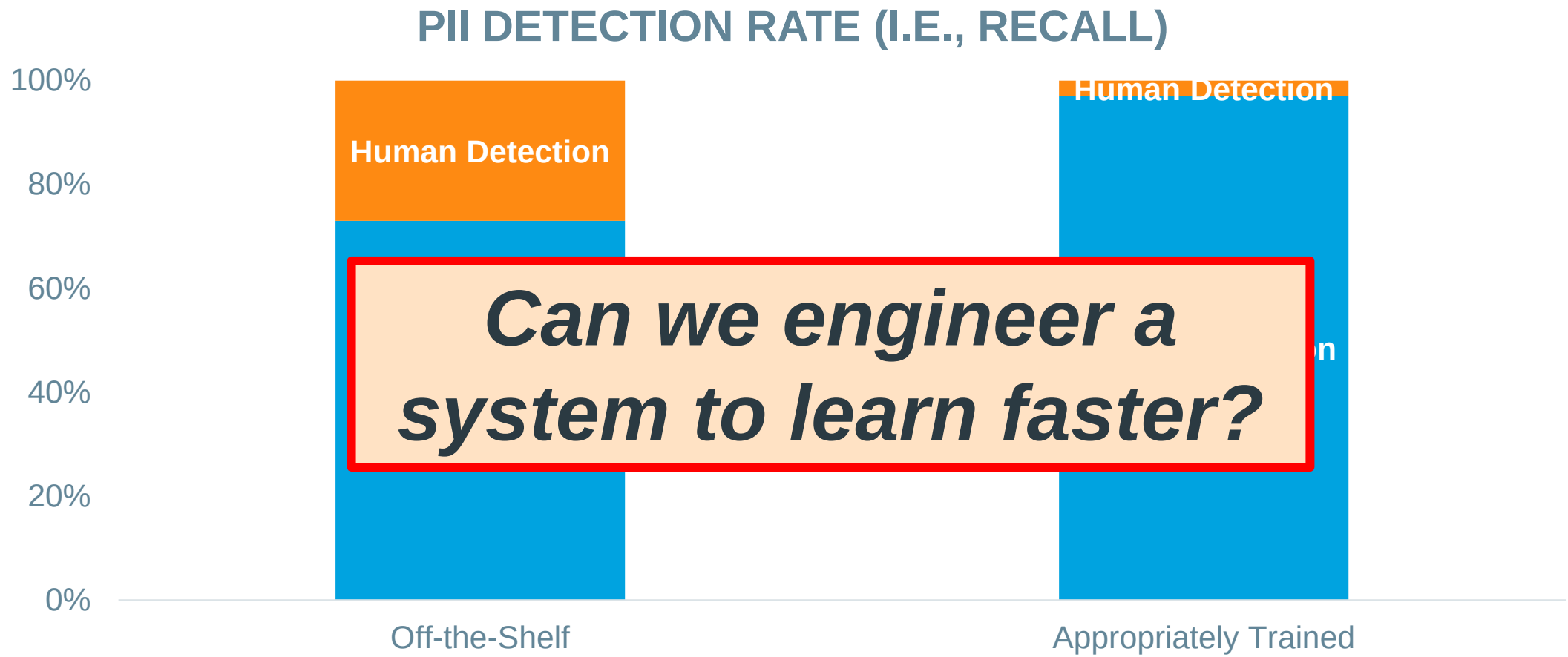




PRIVACY ANALYTICS
an IQVIA company

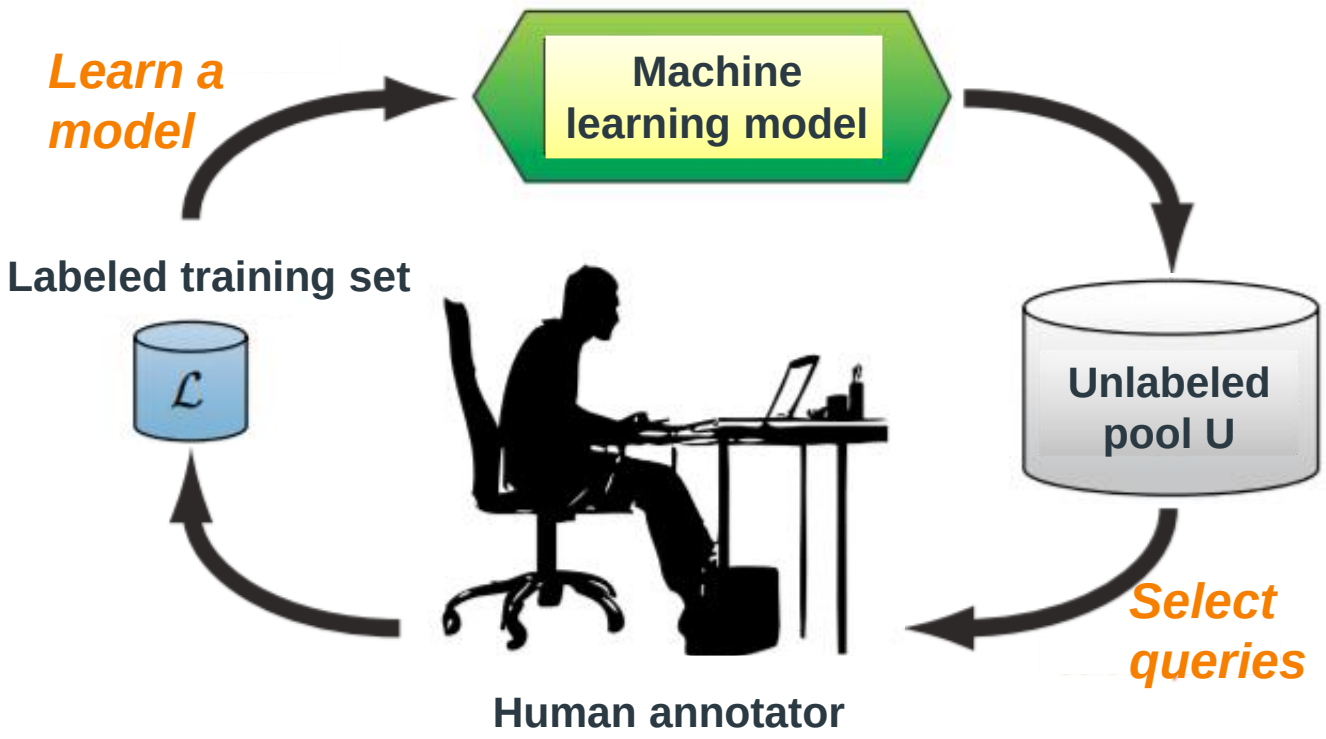
IQVIA

Why NER models cannot be used off-the-shelf and need to be re-trained



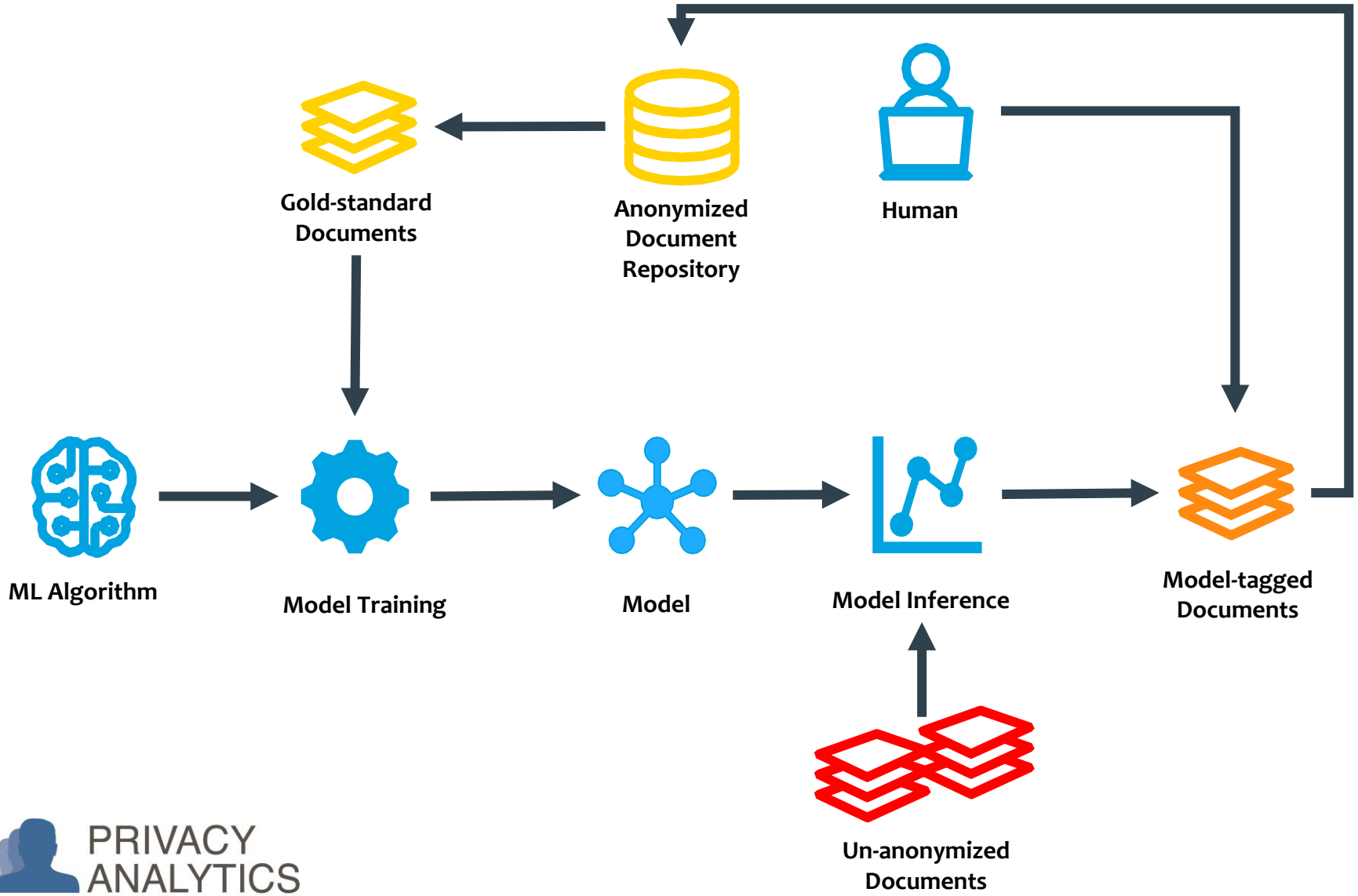
Stubbs, Amber, Michele Filannino, and Özlem Uzuner. "De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1." *Journal of biomedical informatics* 75 (2017): S4-S18.
Dernoncourt, Franck, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. "De-identification of patient notes with recurrent neural networks." *Journal of the American Medical Informatics Association* 24, no. 3 (2017): 596-606.

What is *active* learning and why?



Settles, Burr. "Active learning literature survey." (2009).

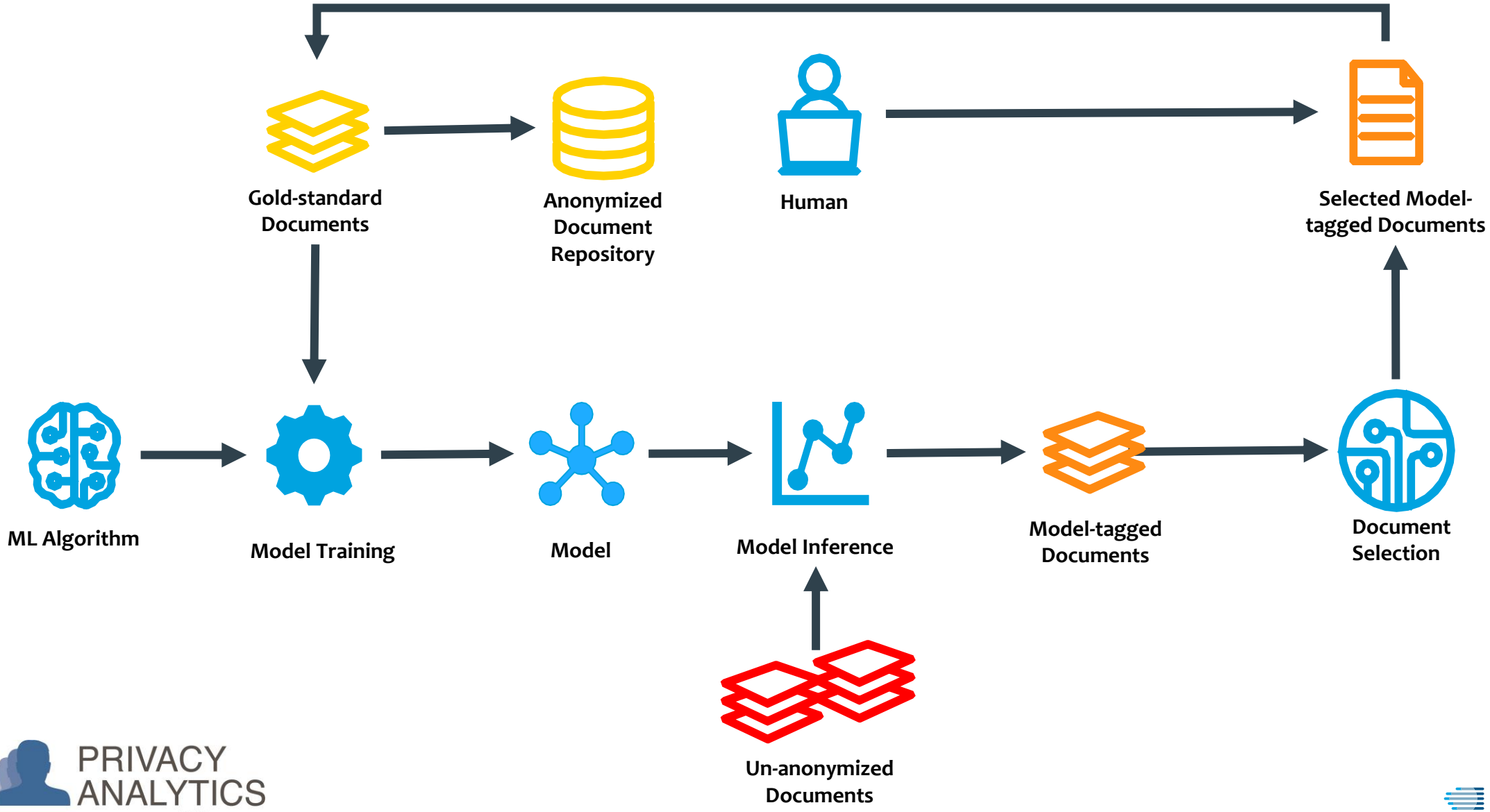
Traditional machine learning-based document anonymization workflow



PRIVACY ANALYTICS
an IQVIA company

IQVIA

Active learning-based document anonymization workflow

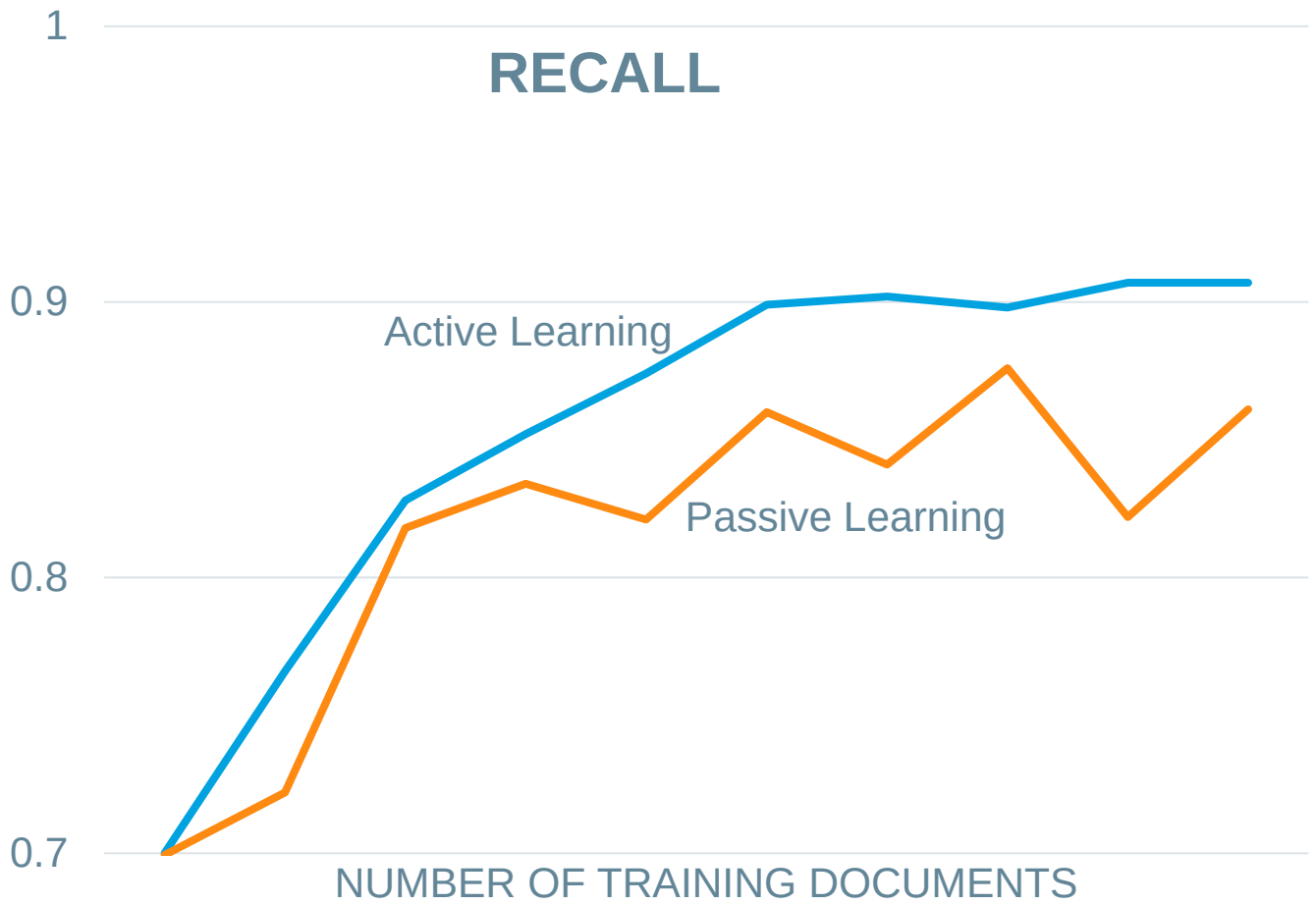


PRIVACY ANALYTICS
an IQVIA company

IQVIA

How fast does active learning learn?

Simulation on a real-world clinical trials dataset



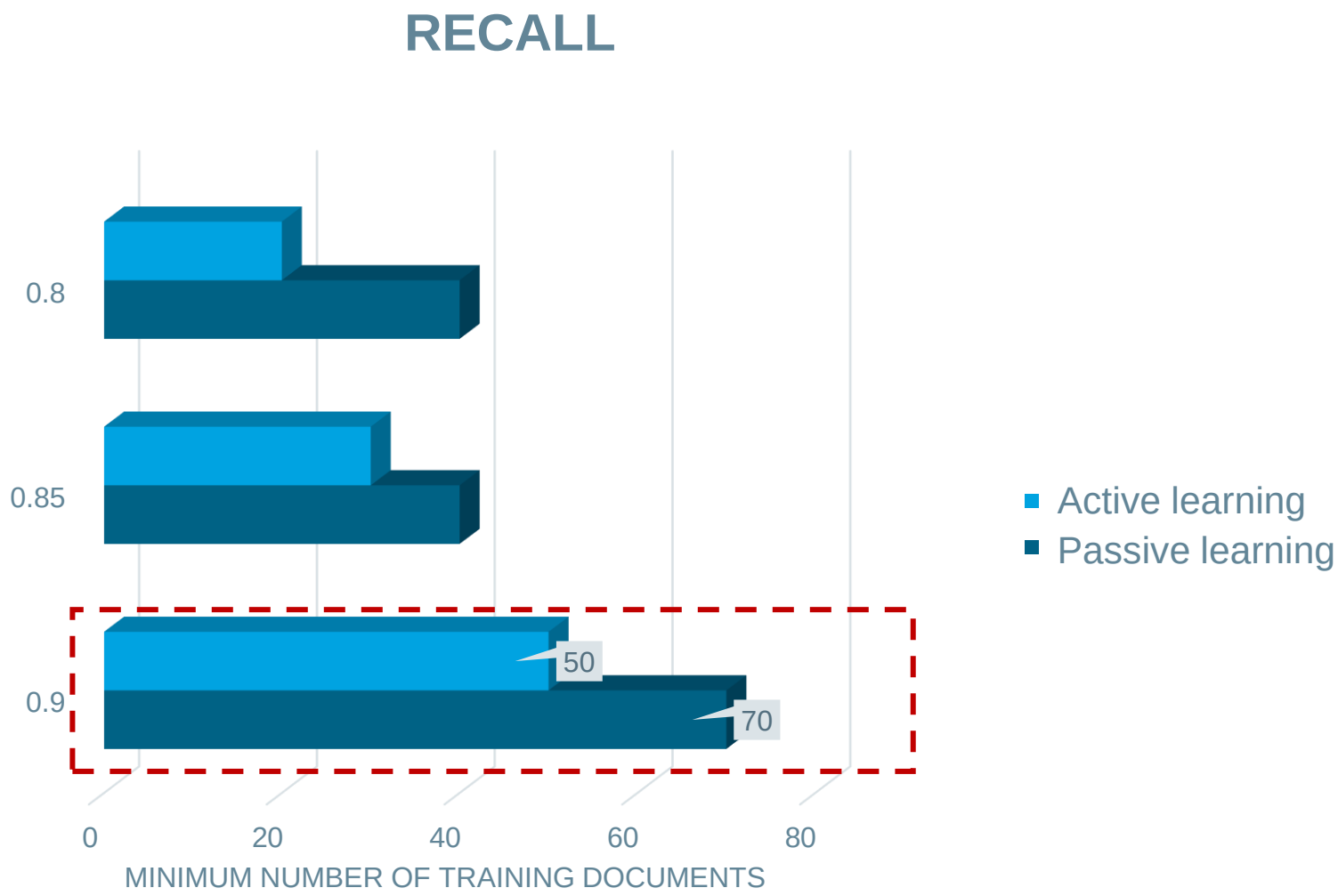
Li, Muqun, Martin Scaiano, Khaled El Emam, and Bradley A. Malin. "Efficient active learning for electronic medical record de-identification." *AMIA Summits on Translational Science Proceedings* 2019 (2019): 462.

PRIVACY ANALYTICS
an IQVIA company

IQVIA

How much training time can active learning save?

Simulation on a real-world clinical trials dataset

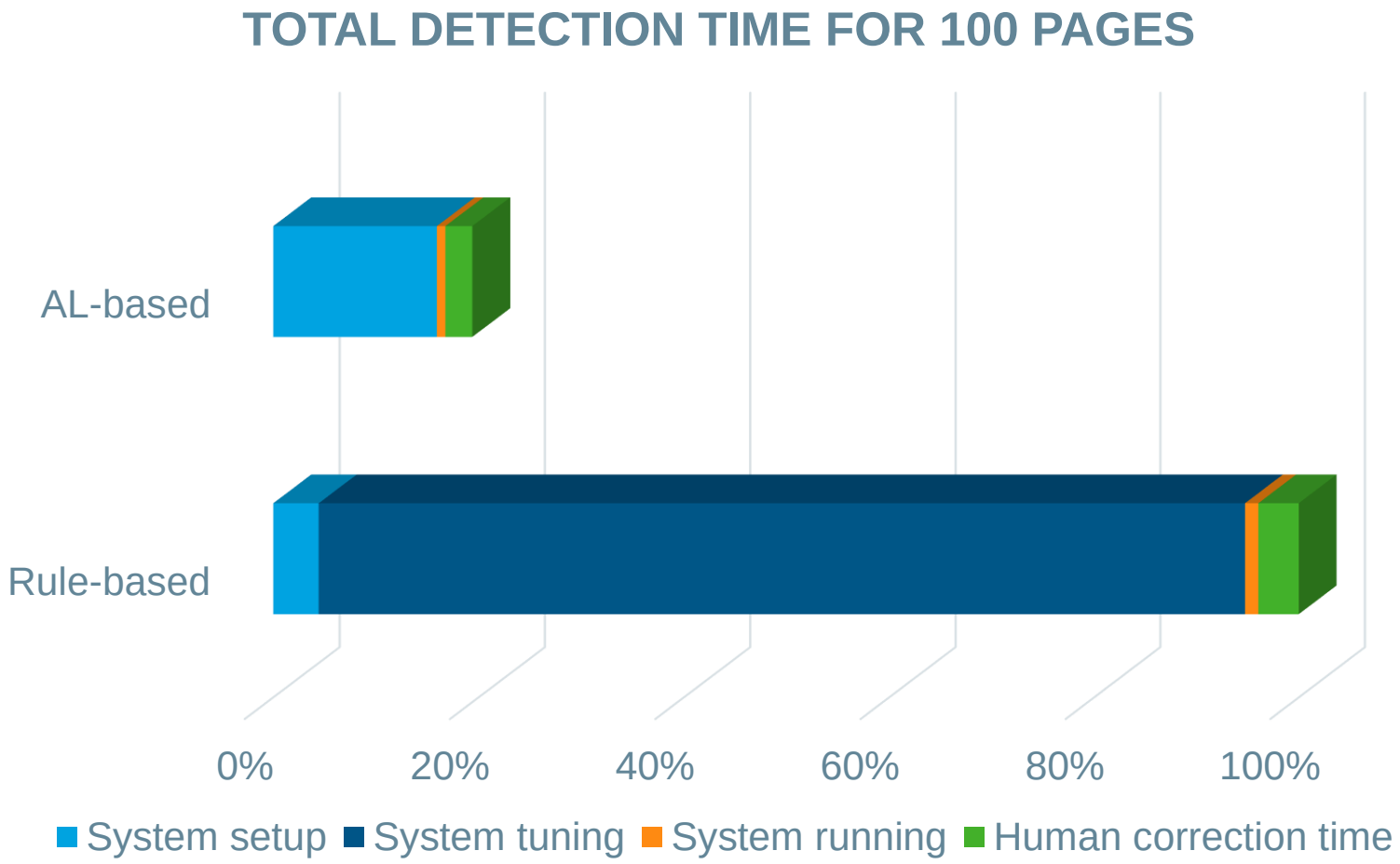


PRIVACY ANALYTICS
an IQVIA company

IQVIA

How much time can be saved in practice (a user study)?

Rule-based vs. active learning-based

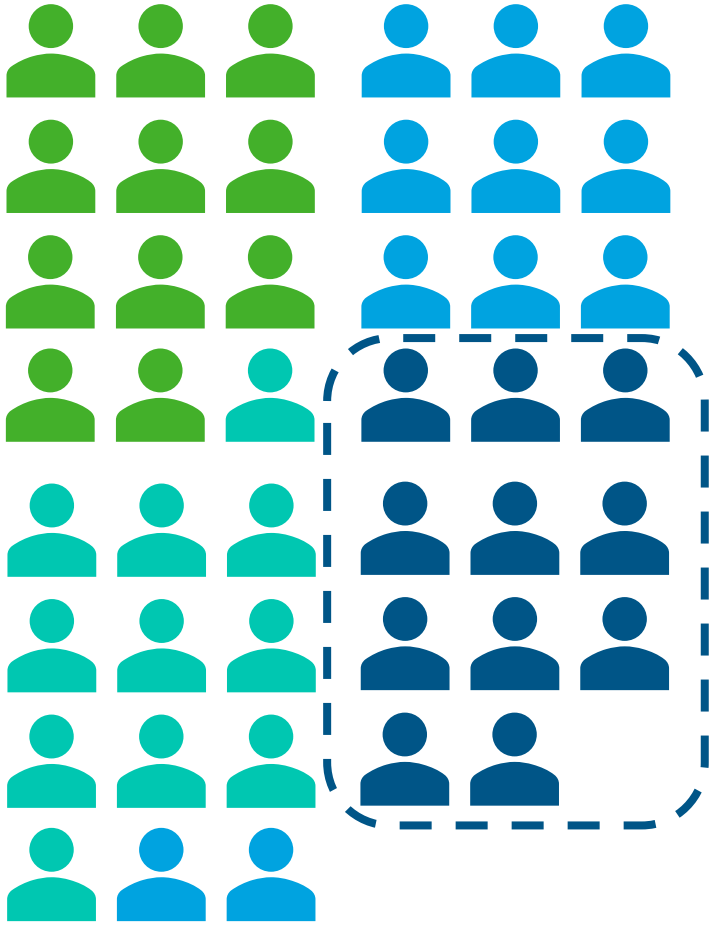


PRIVACY ANALYTICS
an IQVIA company

IQVIA

Evaluating identifiability on unstructured text

Reminder: When are public documents “anonymized”?



For the **public release** of clinical documents, we typically require each data subject to be **statistically similar** to **at least 10 other individuals.**

If this condition is met, we would consider the data **anonymized**.

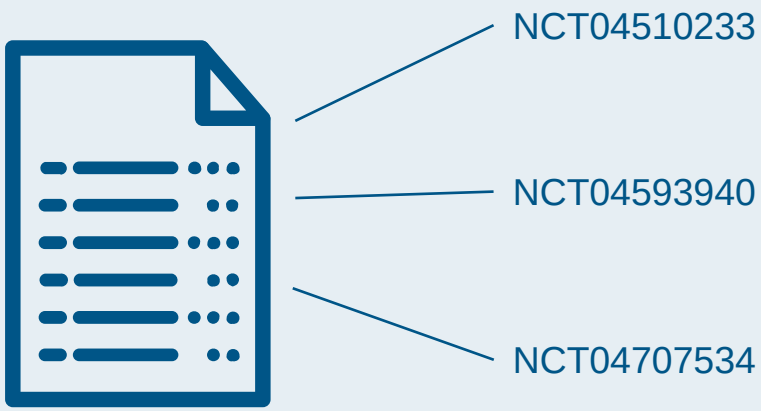
(See EMA/90915/2016 Section 5.4.1, Health Canada “Public Release of Clinical Information: guidance document” Section 6.2)

PRIVACY ANALYTICS
an IQVIA company

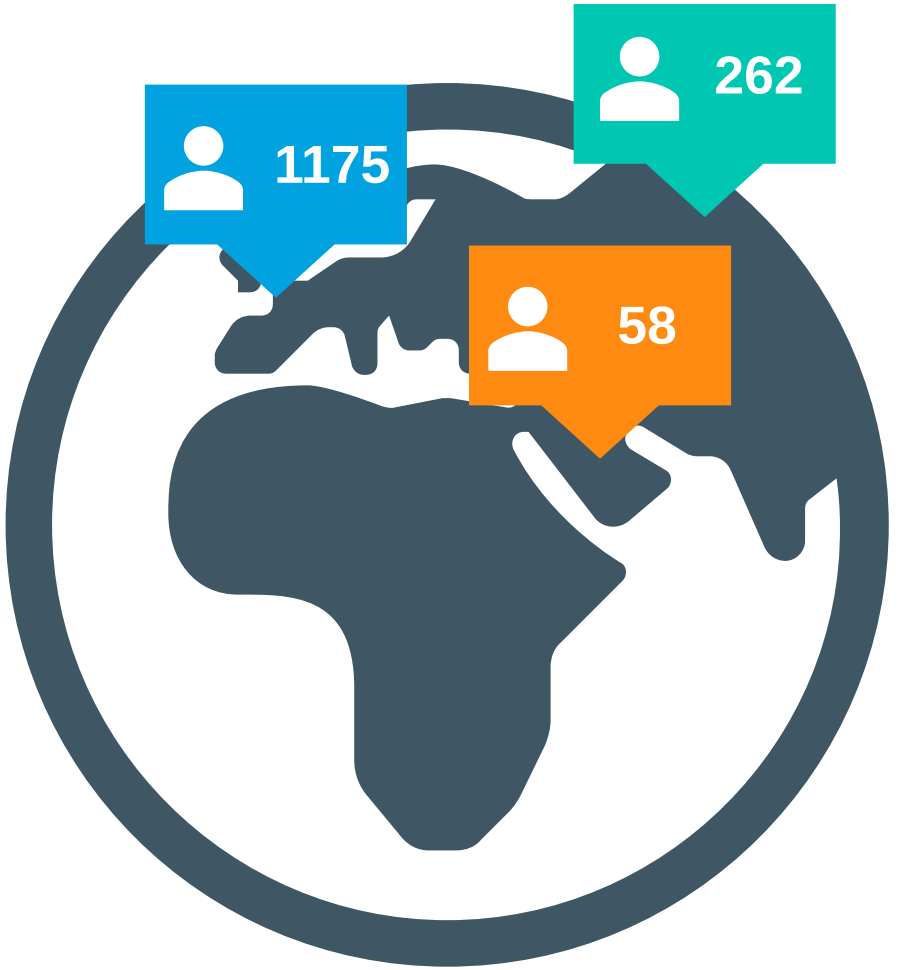
IQVIA

Defining a reference population

We evaluate identifiability with respect to a **reference population** of individuals that ***"look like"*** they could have participated in the clinical trial under consideration.



We leverage **multiple sources** to estimate the size of the reference population, including clinical trial registries like ClinicalTrials.gov.

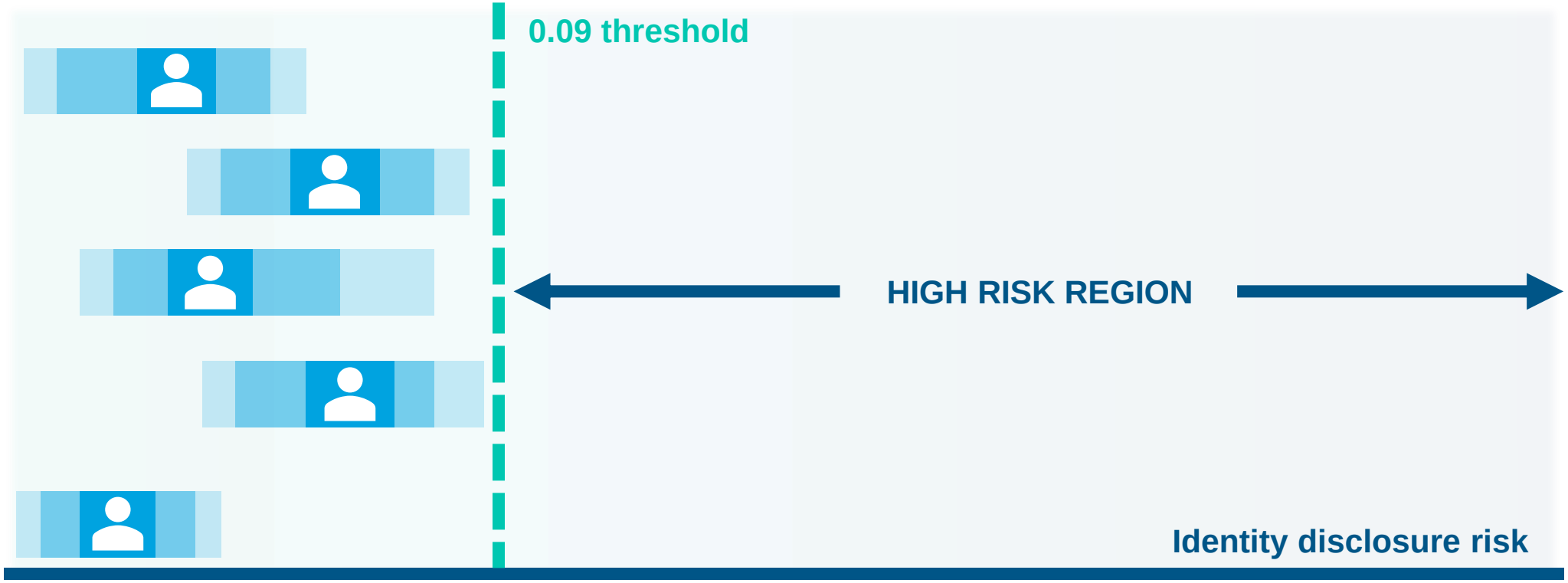


(See EMA/796532/2018 Section 4.2, Health Canada "Public Release of Clinical Information: guidance document" Section 6.2)

PRIVACY ANALYTICS
an IQVIA company

IQVIA

Maximum identity disclosure risk measurement



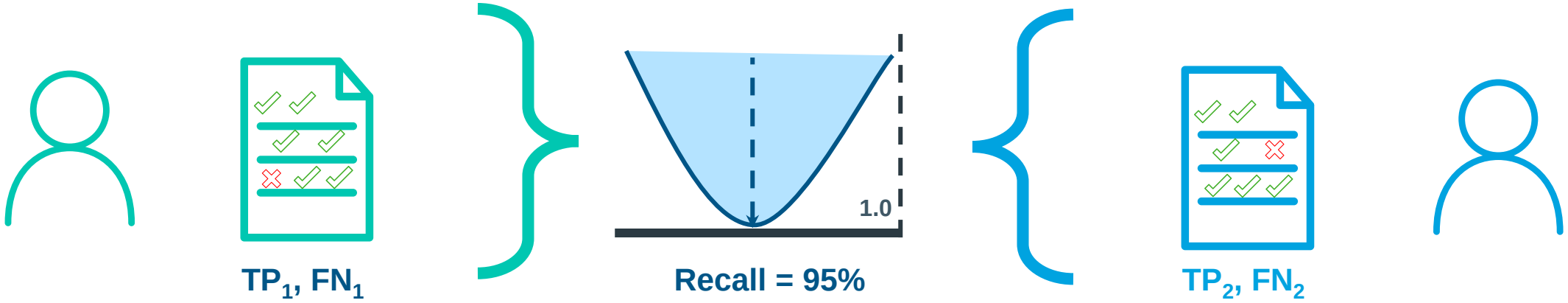
Commonly we translate our “cell size” requirement into an **identity disclosure risk measurement** by requiring that the risk of re-identification of each participant must be **less than 0.09** (or 1/11).

(See EMA/90915/2016 Section 5.4.1, Health Canada “Public Release of Clinical Information: guidance document” Section 6.2)

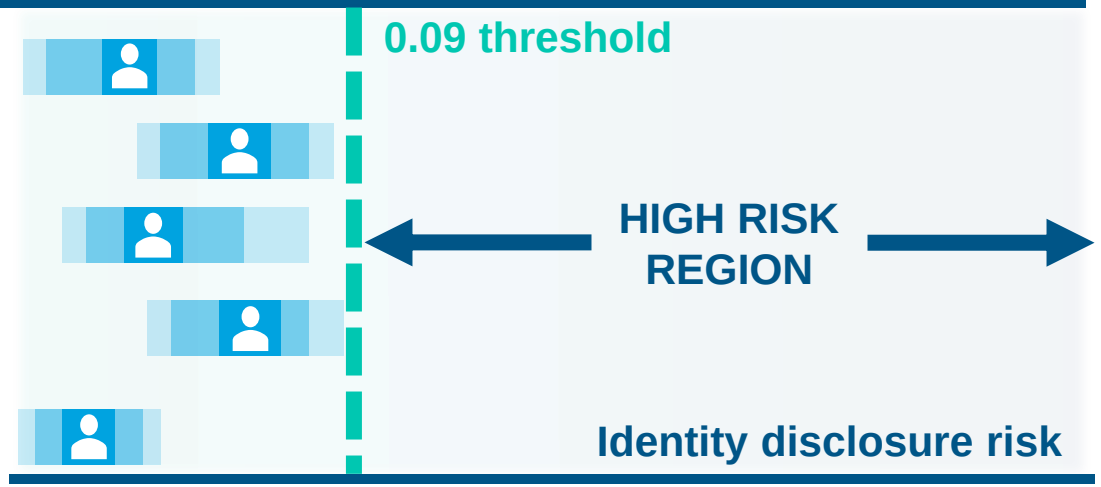
Estimating the effect of leaked identifiers

Scaiano, M., Middleton, G., Arbuckle, L., Kolhatkar, V., Peyton, L., Dowling, M., Gipson, D.S. and El Emam, K., 2016. A unified framework for evaluating the risk of re-identification of text de-identification tools. *Journal of biomedical informatics*, *63*, pp.174-183.

To estimate the effect of **missed (or leaked) identifiers** on identifiability, two analysts **annotate a sample of pages** by hand to create a **gold standard**.



This allows us to estimate the **upper bound of identifiability** of each data subject.



PRIVACY ANALYTICS
an IQVIA company

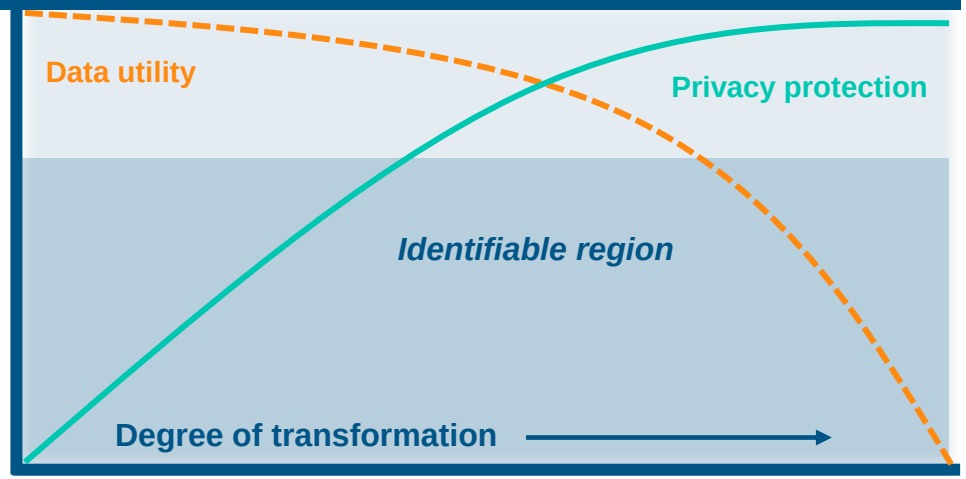
IQVIA

Transforming data to mitigate disclosure risk

This **51** year-old female, with a past history of **hypertension** and **arthritis**, complained of **influenza**-like symptoms which included **fever of 100.4 F**, **chills**, **body ache**, **rhinitis** and **infected throat**, and was randomized to the study. […] On **FEB 20** (Day 3), the patient had a fever of 104 F and was hospitalized and discontinued from study.

This **47** year-old female, with a past history of **[MEDICAL HISTORY]** and **rheumatoid arthropathies**, complained of influenza-like symptoms which included **fever of 100.4 F**, **chills**, **body ache**, **[MEDICAL HISTORY]** and **[MEDICAL HISTORY]**, and was randomized to the study. […] On **JAN 8** (Day 3), the patient had a fever of 104 F and was hospitalized and discontinued from study.

Our goal is to transform the data in a way that optimizes **privacy protection** and **data utility.**



PRIVACY ANALYTICS
an IQVIA company

IQVIA

Wrapping up

Document anonymization requires an approach tailored to the **format, content, and knowledge domain** of the unstructured text

Machine learning and AI can be used to **enable scalability** in document anonymization systems, with **humans-in-the-loop** as a critical element

A **statistical approach** to disclosure control can be used to balance **identifiability** and **data utility**

PRIVACY ANALYTICS
an IQVIA company

IQVIA

Thank you!
Questions?

ddivalentino@privacy-analytics.com

rli@privacy-analytics.com

https://privacy-analytics.com/

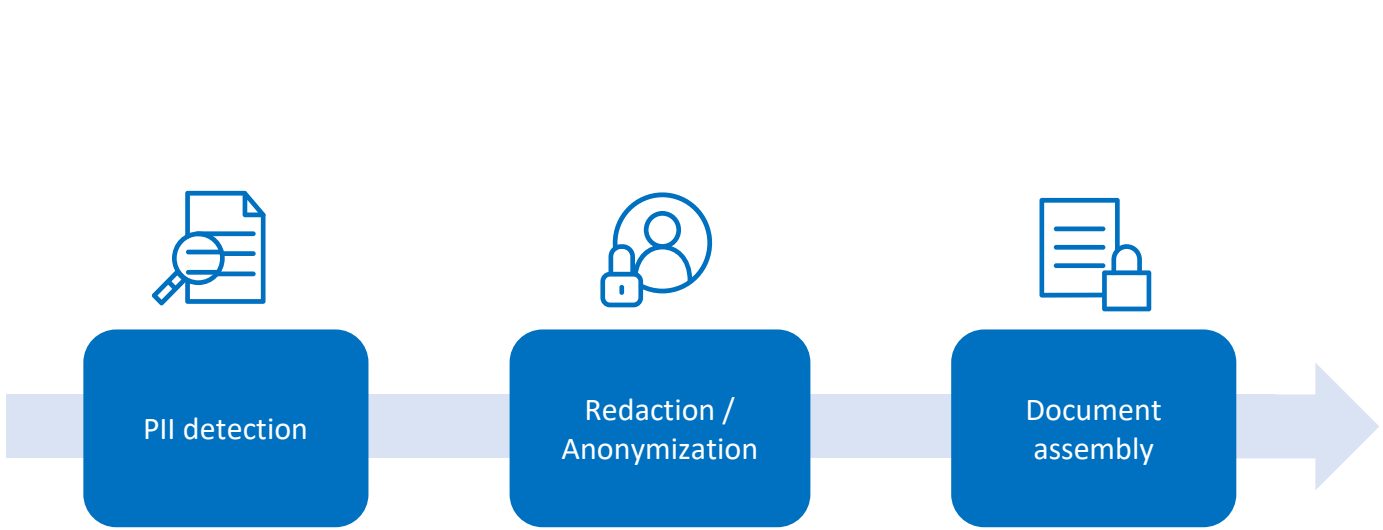
@privacyanalytic

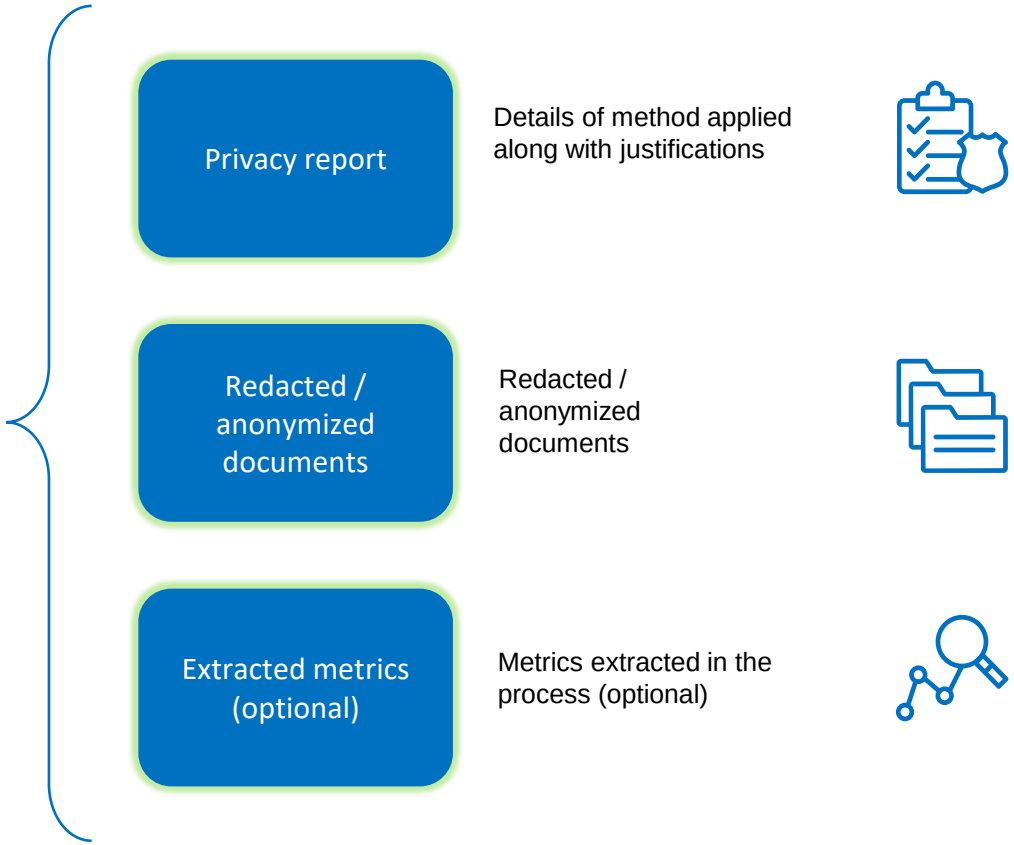
https://www.youtube.com/user/PrivacyAnalytics

https://www.linkedin.com/company/privacy-analytics-inc-/

Extra slides

The anonymization process





PRIVACY ANALYTICS
an IQVIA company

IQVIA

Practical challenges with assessing identifiability

Oftentimes the detected **personal information** about a given data subject is distributed across **several clinical documents**.

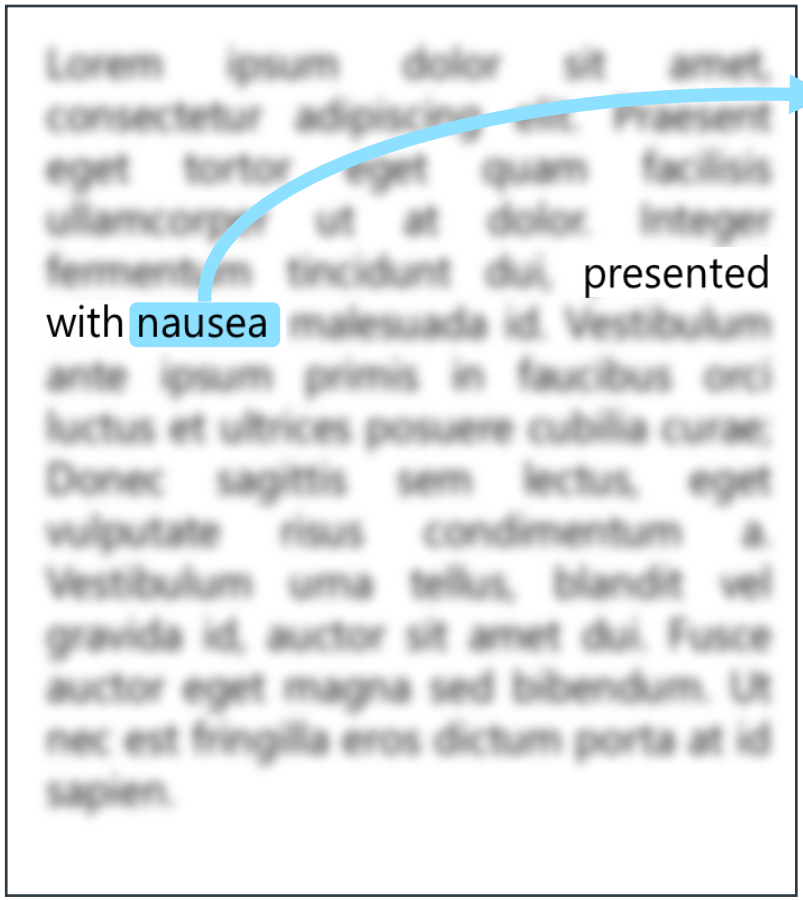
Therefore, we first need to aggregate the **unique set of personal information** for that data subject to properly **assess identifiability**.

PRIVACY ANALYTICS
an IQVIA company

IQVIA

(**Note:** MedDRA = “Medical Dictionary for Regulatory Activities”)

Practical challenges with assessing identifiability

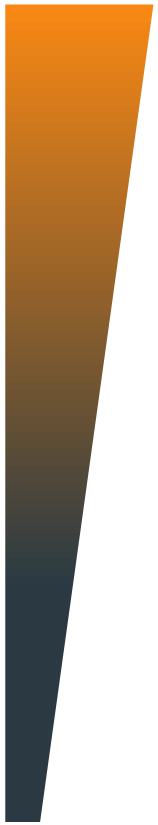


Verbatim Term
Nausea

High Level Term
Nausea and vomiting symptoms

High Level Group Term
Gastrointestinal signs and symptoms

System Organ Class
Gastrointestinal disorders



Identifiability

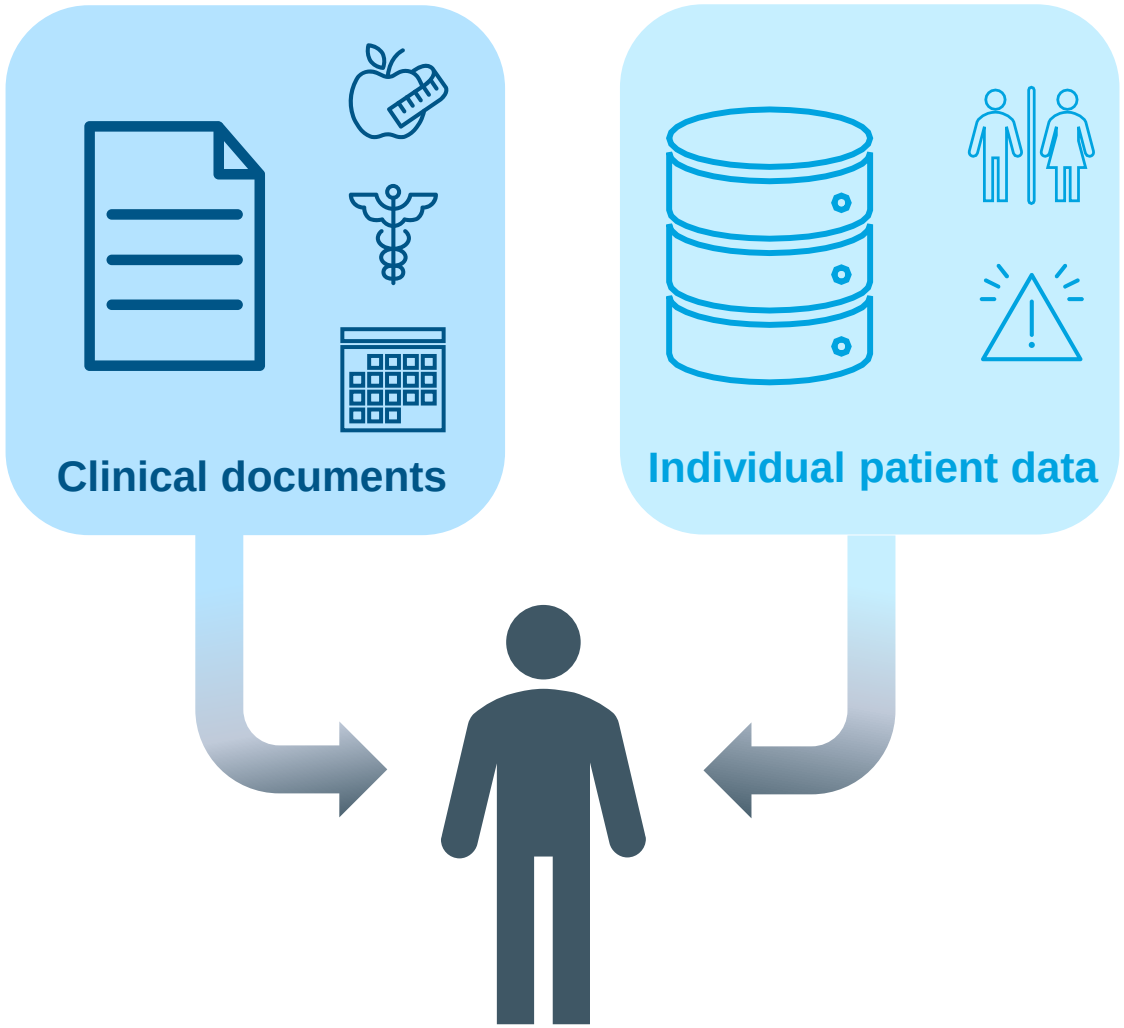
Identifiers such as medical history terms often appear as **verbatim descriptions** of a patient’s conditions.

So we must first **standardize** them in order to properly assess how **identifiable** they are (using e.g., MedDRA dictionary).

PRIVACY ANALYTICS
an IQVIA company

IQVIA

Practical challenges with assessing identifiability



Some information (such as **patient gender**) is generally too **error prone** to be reliably detected in unstructured text.

In such cases, we typically pull information from the corresponding **structured individual patient data** to assess identifiability.

PRIVACY ANALYTICS
an IQVIA company

IQVIA

Practical challenges with assessing identifiability

In cases of **ultra-rare diseases**, or patients in countries with little **clinical trial representation**, we may find that the data must be **significantly transformed** to ensure non-identifiability.

Simulation on a real-world clinical trials dataset