Data transparency and anonymization when sharing clinical study reports (CSRs)

An exploration of Natural Language Processing, Knowledge Graphs and Statistical Disclosure Control

Nastazja Laskowski

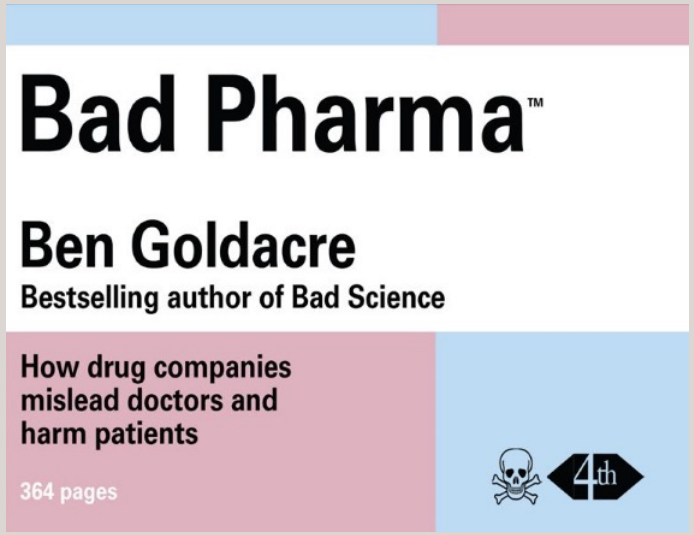
PhD Research Fellow
Data Privacy & Sharing Pod
Data Strategy & Delivery, PD Data Sciences

Supervised by Alex Hughes

Table of contents

Introduction

Introduction



Roche

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

Vivli
CENTER FOR GLOBAL CLINICAL RESEARCH DATA

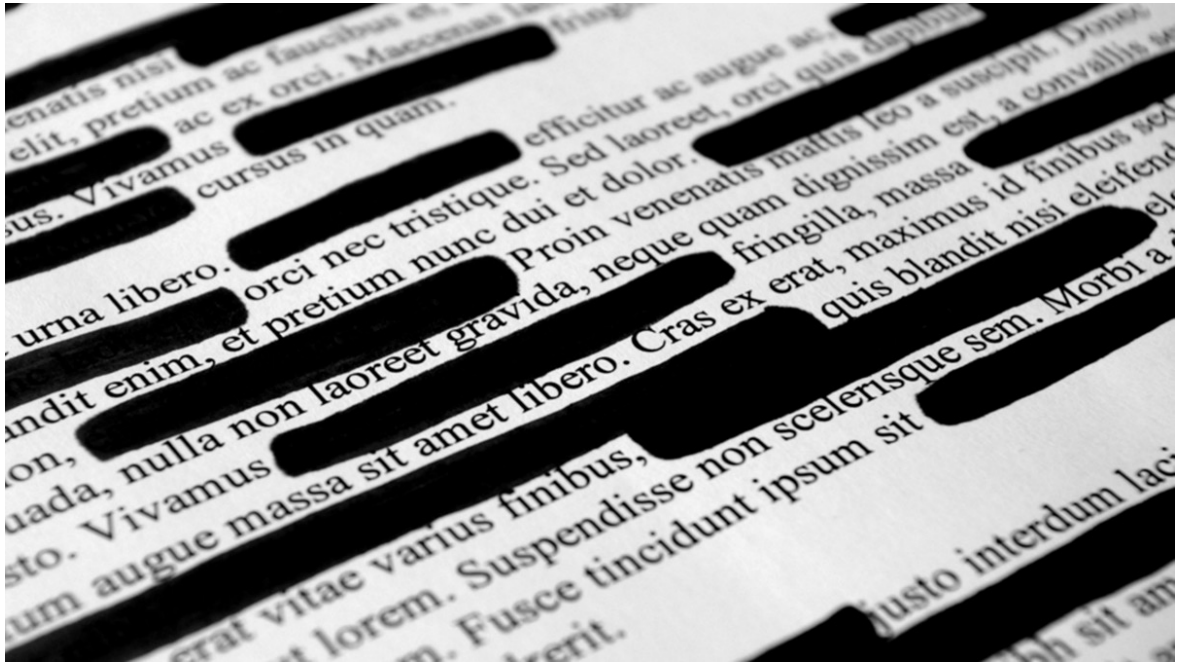
GDPR

HIPAA

PIPEDA

Moreover,

In that same vein,



The opposing forces of data transparency and privacy are constantly shifting… There is a need for a scalable automation solution to share data proactively…

Natural Language Processing

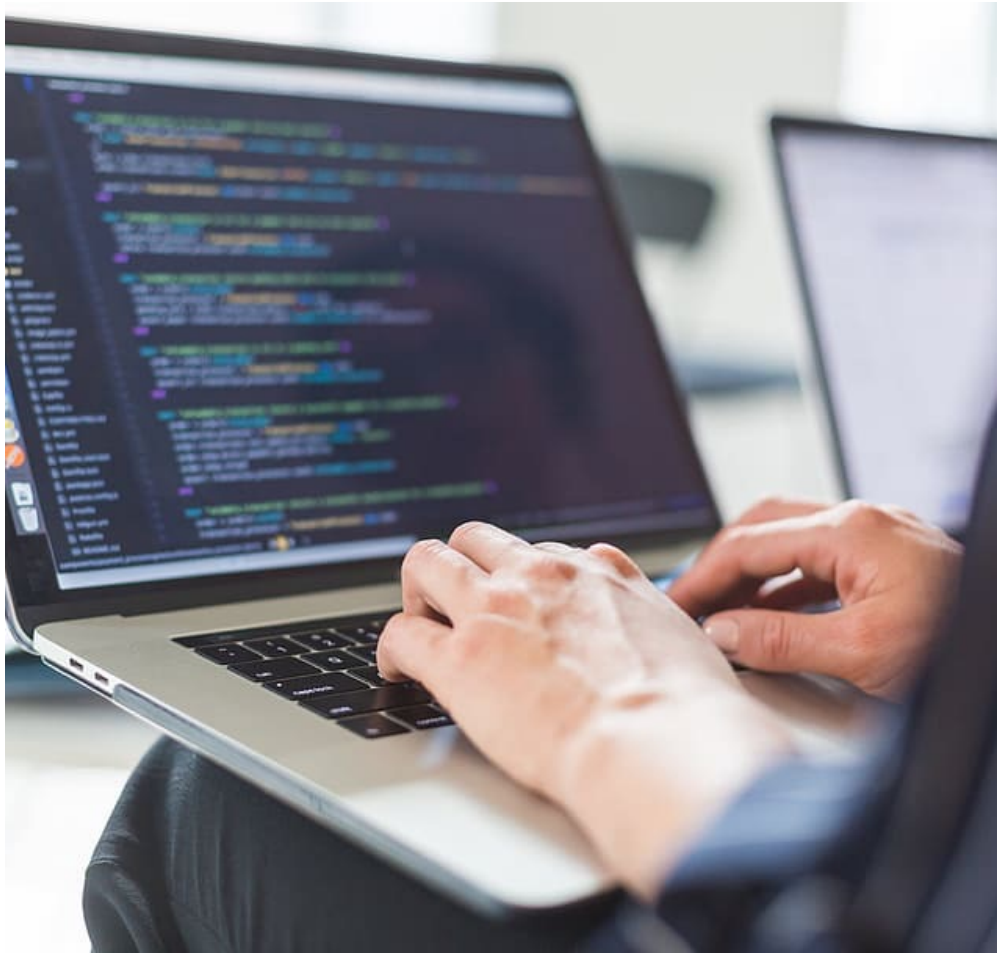
Natural Language Processing of patient narratives

An exploration of the current state-of-the-art

- Text pre-processing: tokenization, stop word removal, stemming/ lemmatization, sentence segmentation
- Part-of-speech tagging (POS)
- Named Entity Recognition (NER) pre-trained model or regular expressions
- Dependency parsing

- Metrics: precision, recall, F1 and F2

- Challenge: ambiguity and variability of human language



Statistical disclosure control

Statistical Disclosure Control of Clinical Study Reports

K-anonymity and L-diversity

- **K-anonymity** ensures that each individual in a dataset is indistinguishable from at least k other individuals.

- **L-diversity** is an extension of k-anonymity which provides more protection against attribute disclosure.

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Li, N., Li, T. and Venkatasubramanian, S., 2006, April. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering* (pp. 106-115). IEEE.

Knowledge Graphs

Storing data in a Knowledge Graph (KG)

An exploration of the current state-of-the-art

- Import structured data (e.g. from CSV or database)
- Define nodes (entities) and edges (relationships) manually or using NER libraries
- Add properties (e.g. source of data)
- Visualise the graph
- Query the nodes

Summary

Summary

- Complex research question in a complex ecosystem
- Convergence of two/three research fields
- Methodological and philosophical challenges

Thanks for listening…

Nastazja.Laskowski@roche.com

Doing now what patients need next