# NLM Scrubber to de-identify clinical text data and facilitate data sharing

Lightning Talk – MIRL 2023

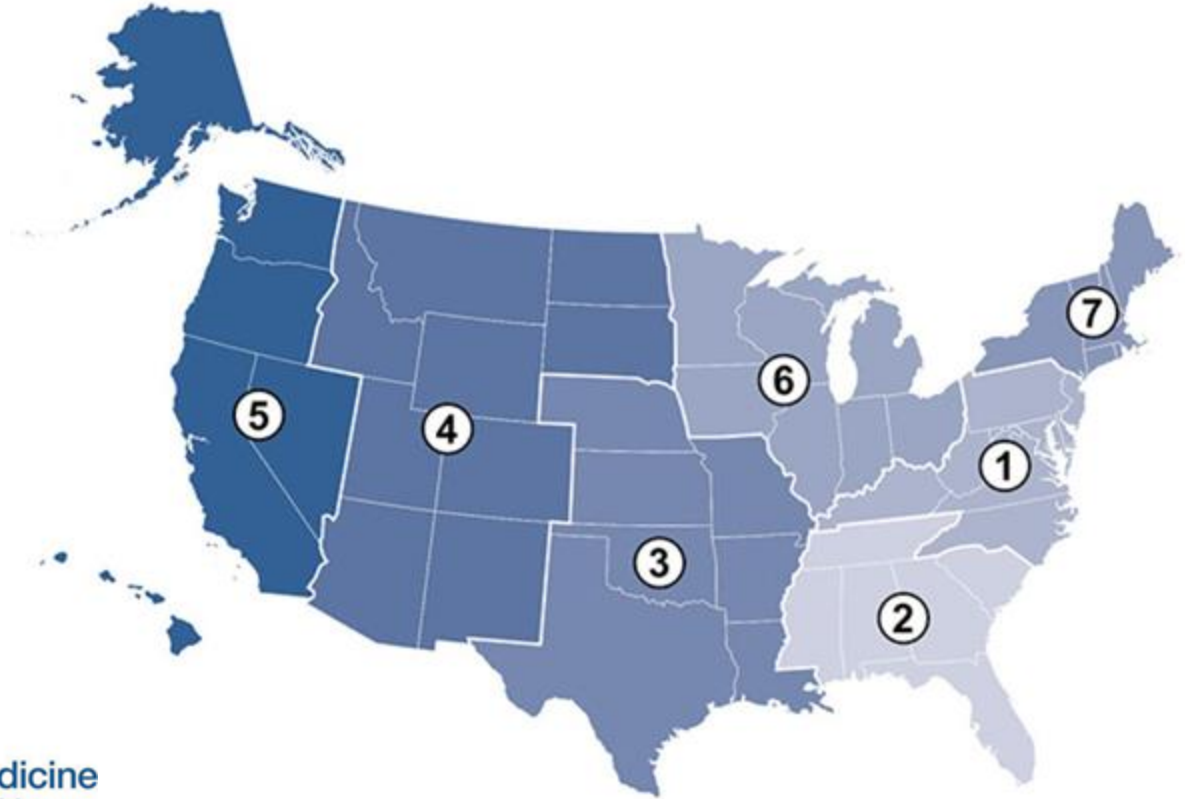
Katie Pierce Farrier, Data Science Strategist

Christine Nieman Hislop, Data Education Librarian

NIH National Library of Medicine
Network of the National Library of Medicine

# Network of the National Library of Medicine

- Education
- Outreach
- Funding



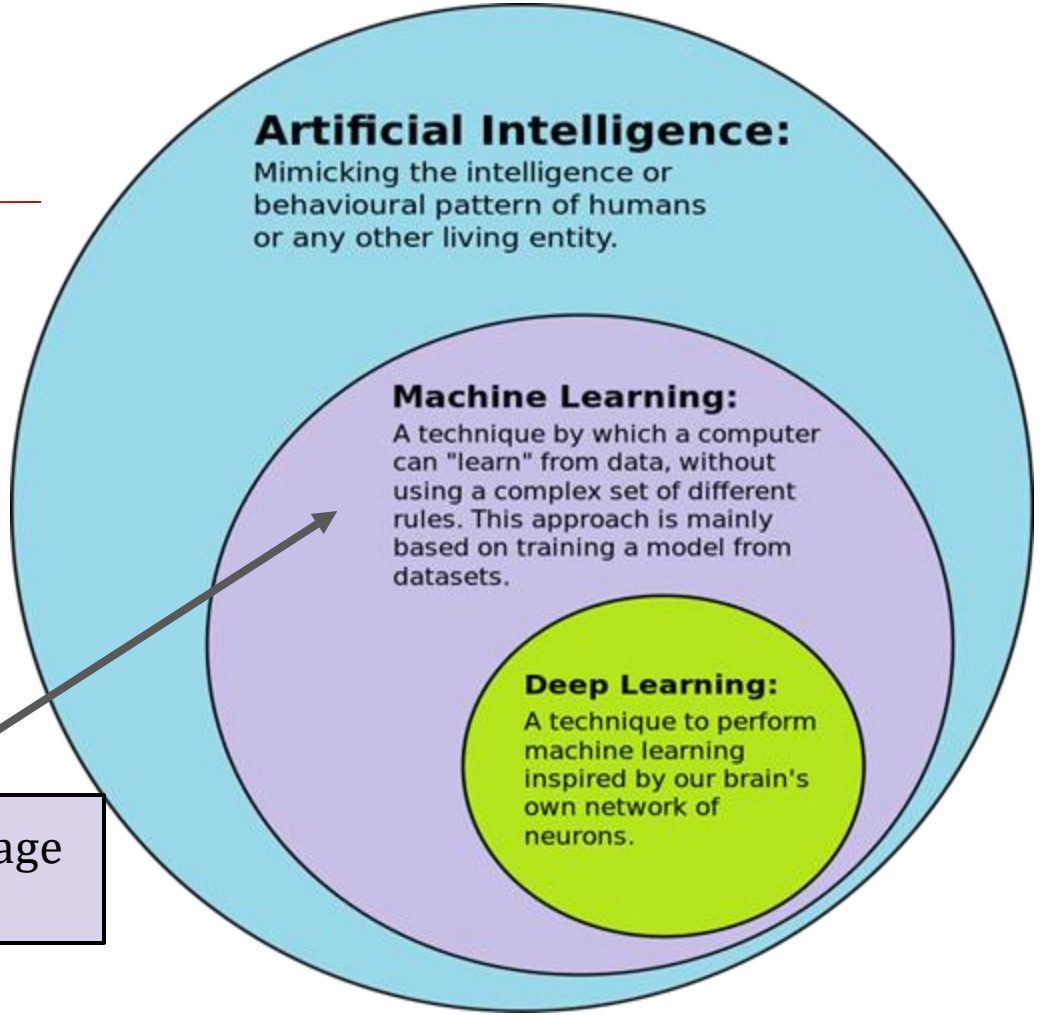

NIH National Library of Medicine
Network of the National Library of Medicine

NNLM.gov

# Defining AI

*No megalomaniac robots involved….for now*

[Photo from Wikimedia Commons]




Natural Language Processing

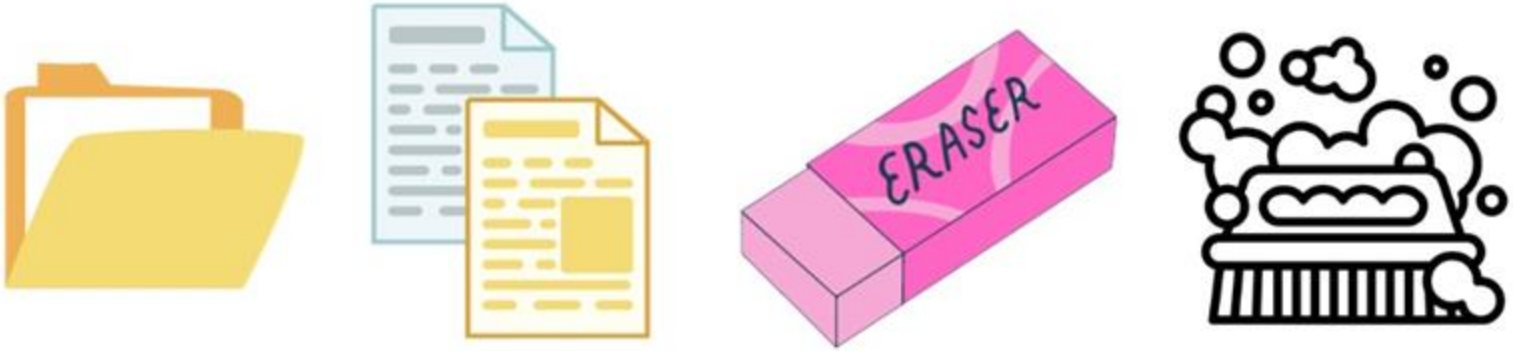
# So what is NLP?

- Languages follow predictable **rules**
- Machine Learning follows **statistics**
- Allows computers to **understand human languages**

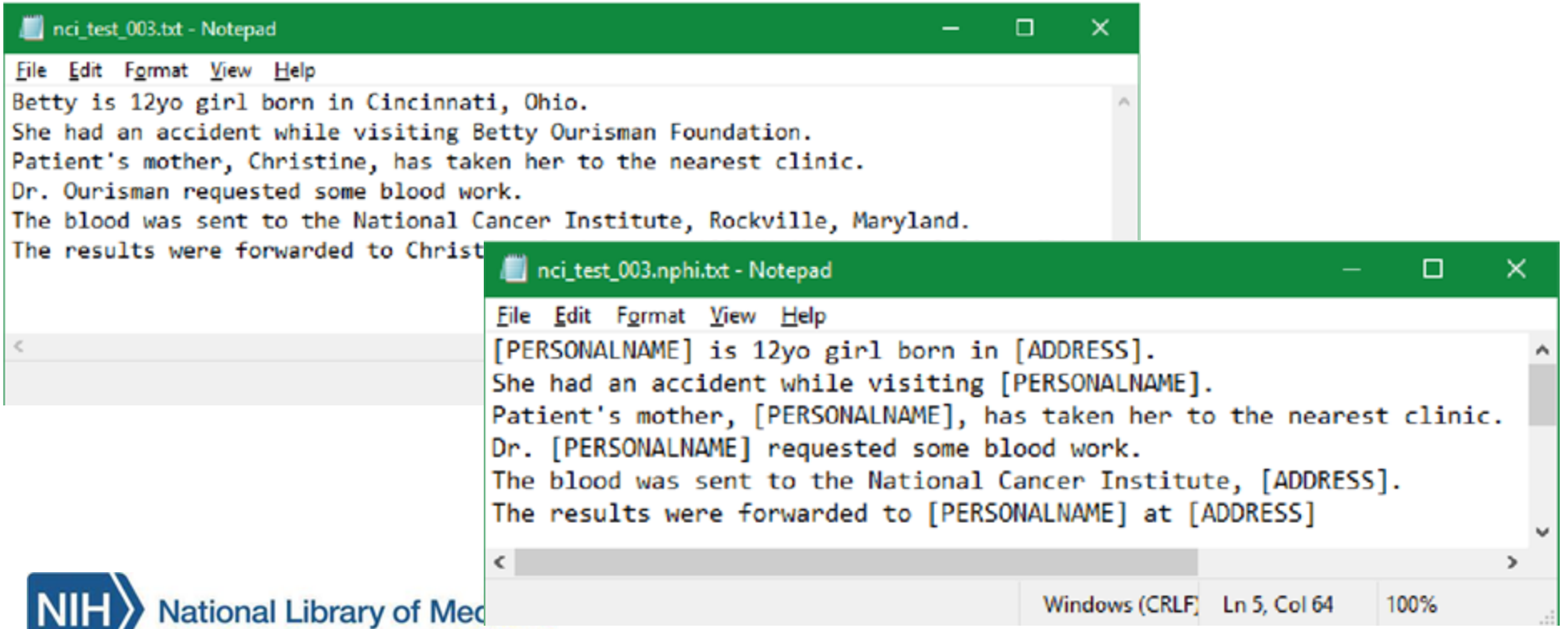
# Examples of NLPs

- Apple Siri
- Amazon Alexa
- ChatGPT
- Customer service chatbots


# NLM Scrubber uses NLP to find and redact PHI ….



# … to protect patient privacy and…



# …. facilitate sharing of clinical de-identified data!


# De-identified Data = Shareable Data

**F**indable

**A**ccessible

**I**nteroperable

**R**eusable

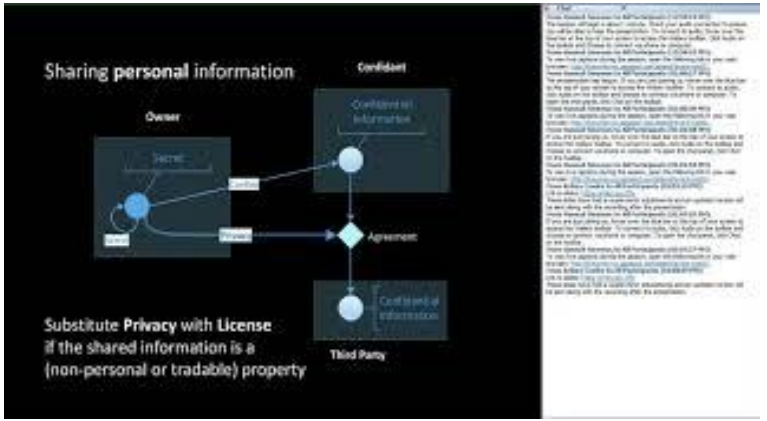
Don't forget **Reproducible!**

- 68% of experiments unable to obtain data
- 39 of 53 papers cited NIH funding (Errington et al., 2021)

NIH National Library of Medicine
*Network of the National Library of Medicine*

# Share with your Institution!

- Try it out!
- Include it in LibGuides
  - John Hopkins Deidentification Tools
  - UNTHSC Data Management LibGuide
- Host a workshop
- Data Privacy Week
  - Jan 21-27




NIH National Library of Medicine
Network of the National Library of Medicine

Privacy Research & Clinical Text Deidentification with NLM-Scrubber, February 27, 2020

# Thank you!

Christine Nieman Hislop
cnieman@hshsl.umaryland.edu
NNLM Region 1

Katie Pierce Farrier
Katie.pierce-farrier@unthsc.edu
NNLM Region 3

NIH National Library of Medicine
Network of the National Library of Medicine

# Resources

NLM Scrubber and its User Manual

IBM: What is Natural Language Processing?

Go FAIR Data Principles

Challenges for assessing replicability in preclinical cancer biology

## Articles About NLM Scrubber

Evaluation of Automated Public De-Identification Tools on a Corpus of Radiology Reports

An atomic approach to the design and implementation of a research data warehouse

A Comparative Analysis of Speed and Accuracy for Three Off-the-Shelf De-Identification Tools