Using healthcare data for research

Alistair Johnson
Scientist, SickKids

SickKids®

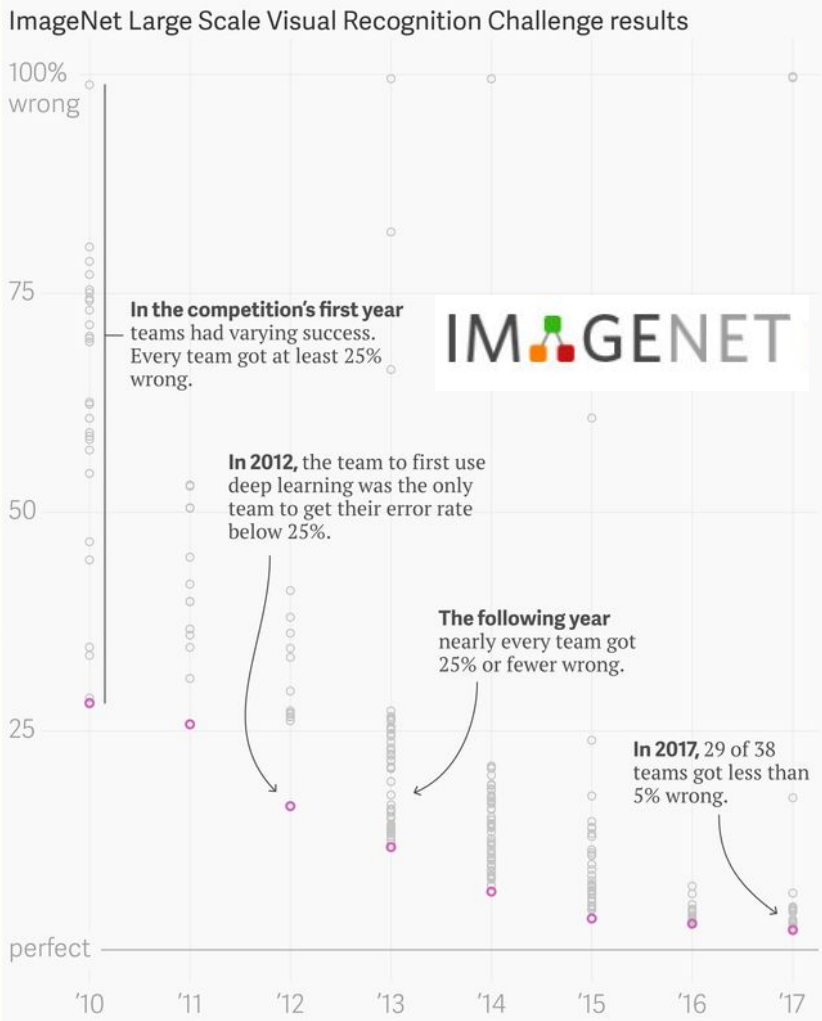
alistair.johnson@sickkids.ca

Computer Vision

2010: Average accuracy 50%

2017: Superhuman performance (<1%)

What about critical care?



David Yanofsky, Quartz, https://qz.com/1034972

Critical care

2016

Box 4. qSOFA (Quick SOFA) Criteria

Respiratory rate ≥22/min

Altered mentation

Systolic blood pressure ≤100 mm Hg

derived from a database of over one million encounters

2019

DeepMind Health

31 JUL 2019

Using AI to give doctors a 48-hour head s life-threatenin

Code availability

We make use of several open-source libraries to conduct our experiments: the machine learning framework TensorFlow (https://github.com/tensorflow/tensorflow) along with the TensorFlow library Sonnet (https://github.com/deepmind/sonnet), which provides implementations of individual model components[58]. Our experimental framework makes use of proprietary libraries and we are unable to publicly release this code. We detail the experiments and implementation details in the Methods and Supplementary Information to allow for independent replication.

Data availability

The clinical data used for the training, validation and test sets were collected at the US Department of Veterans Affairs and transferred to a secure data centre with strict access controls in de-identified format. Data were used with both local and national permissions. It is not publicly available and restrictions apply to its use. The de-identified dataset (or a test subset) may be available from the US Department of Veterans Affairs, subject to local and national ethical approvals.

Every researcher is struggling with data access

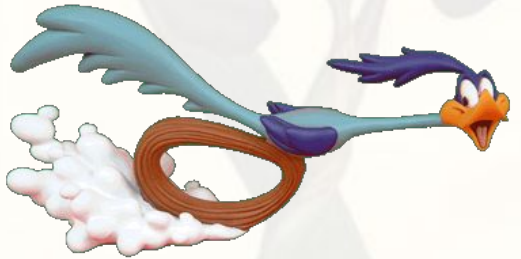
- Unable to reproduce any healthcare studies
- Months to *years* for data access
  - Students come and go before a dataset is available
- Stymies medical research

Comparison with computer vision

- Predicting sepsis for critical care patients
  - One year or more to do extract data from hospital
  - Define sepsis
    - billing code criteria?
    - CDC definition?
    - Sepsis-3?
  - Define time window
    - 3 hours before?
    - 6 hours before?
    - Within 1 hour of ED admission?
  - Extract treatment data, vital signs, labs, ...

Comparison with computer vision

- Detecting objects in an image (ImageNet)
  - Download the images with annotations
  - http://image-net.org/download
  - Build your model



You can share healthcare data.

Really, you can. We did it.

Three requirements

Deidentify the dataset.

Then, to share it, require three steps:

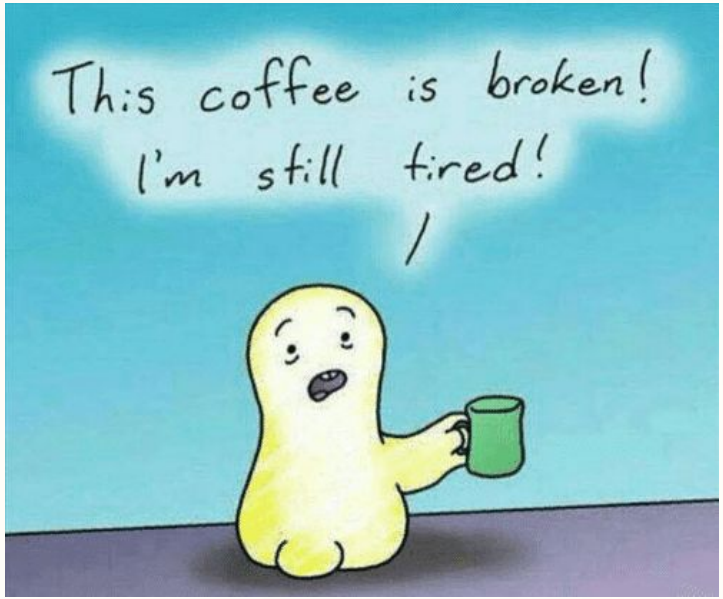
- Registration
- Human participants training
- Data use agreement

Registration

Manually verify that the applicant is a legitimate user

- Check online presence
- Verify affiliations
- Contact references

Resource intensive!



Human participants training

"Data or Specimens Only Research" course

SBE Refresher 1 – Defining Research with Human Subjects (ID: 15029)
SBE Refresher 1 – Privacy and Confidentiality (ID: 15035)
SBE Refresher 1 – Assessing Risk (ID: 15034)
SBE Refresher 1 – Research with Children (ID: 15036)
SBE Refresher 1 – International Research (ID: 15028)

… etc

Can verify online (wish we could do this with TCPS-2!)
https://www.citiprogram.org/verify/?kceff8c59-b392-4f48-a923-095cadcc00ac-27422851

PhysioNet Data Use Agreement

If I am granted access to the PhysioNet Clinical Databases, I agree to the terms and conditions below:

1. I will not attempt to identify any individual or institution referenced in PhysioNet restricted data.
2. I will exercise all reasonable and prudent care to avoid disclosure of the identity of any individual or institution referenced in PhysioNet restricted data in any publication or other communication.
3. I will not share access to PhysioNet restricted data with anyone else.
4. I will exercise all reasonable and prudent care to maintain the physical and electronic security of PhysioNet restricted data.
5. If I find information within PhysioNet restricted data that I believe might permit identification of any individual or institution, I will report the location of this information promptly by email to PHI-report@physionet.org, citing the location of the specific information in question so that it can be investigated and removed if necessary.
6. I have requested access to PhysioNet restricted data for the sole purpose of lawful use in scientific research, and I will use my privilege of access, if it is granted, for this purpose and no other.
7. I have completed a training program in human research subject protections and HIPAA regulations, and I am submitting proof of having done so.
8. I will indicate the general purpose for which I intend to use the database in my application.
9. If I openly disseminate my results, I will also contribute the code used to produce those results to a repository that is open to the research community.
10. This agreement may be terminated by either party at any time, but my obligations with respect to PhysioNet restricted data shall continue after termination.

MIMIC - Critical Care Dataset

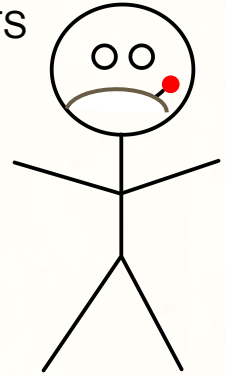
https://mimic.mit.edu

INPUTEVENTS_CV
INPUTEVENTS_MV

LABEVENTS

PRESCRIPTIONS

OUTPUTEVENTS



NOTEEVENTS

CHARTEVENTS

SCIENTIFIC DATA

Altmetric: 45 Views: 3,285 More detail »

Data Descriptor | OPEN

MIMIC-III, a freely accessible critical care database

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi & Roger G. Mark

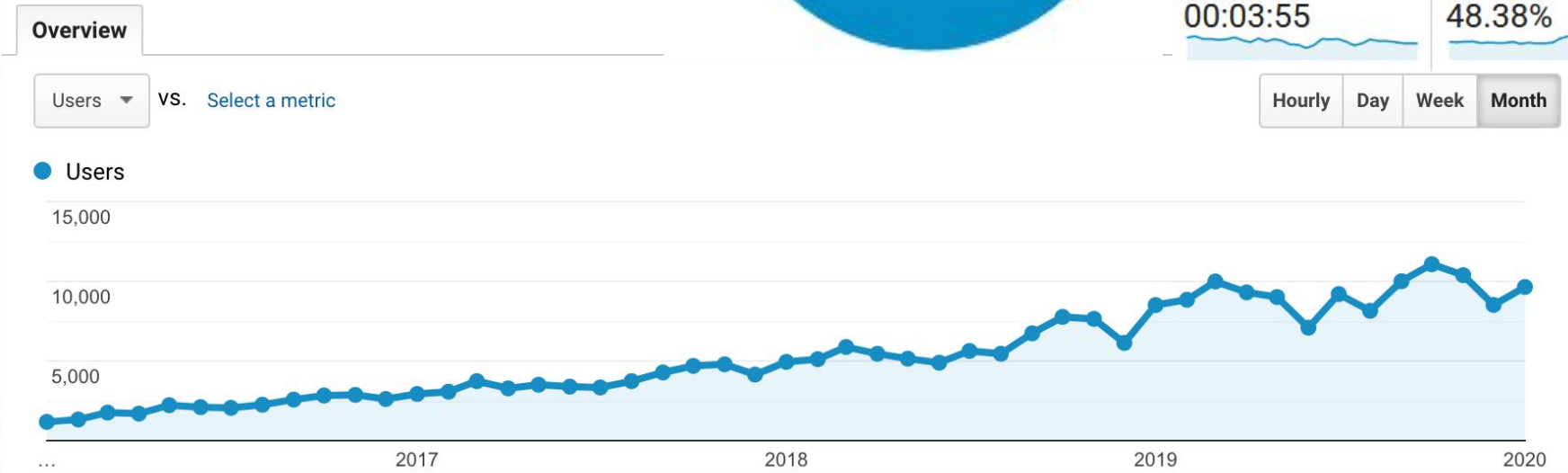
Scientific Data **3**, Article number: 160035 (2016)
doi:10.1038/sdata.2016.35

Received: 18 February 2016
Accepted: 25 April 2016
Published online: 24 May 2016

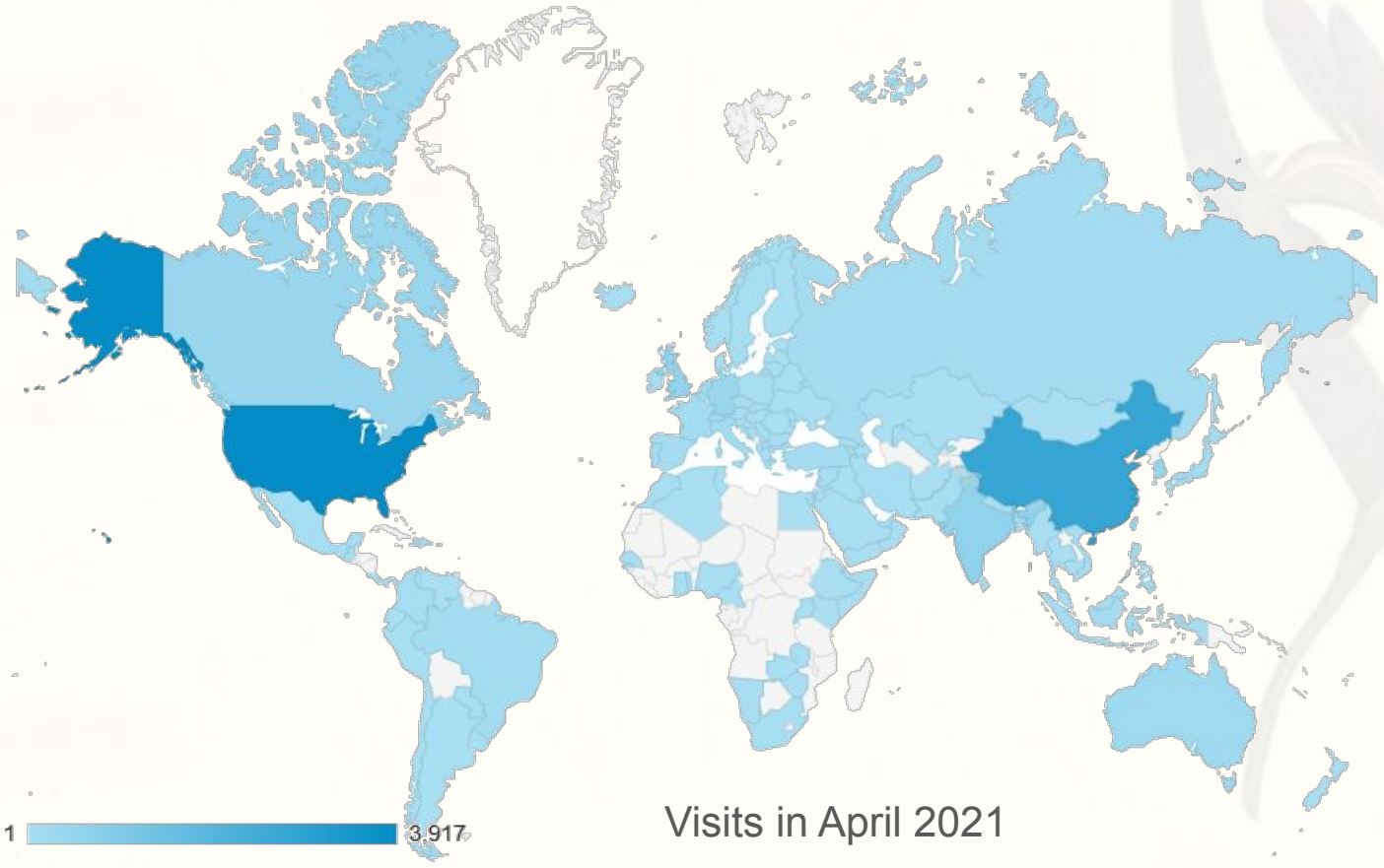
Impact of MIMIC

Snapshot of
April 2021

Users 12,502	New Users 10,536
Sessions 23,402	Number of Sessions per User 1.87
Pageviews 77,217	Pages / Session 3.30
Avg. Session Duration 00:03:55	Bounce Rate 48.38%



Impact of MIMIC



Visits in April 2021

Why release data using this mechanism?

Enables exploratory research.

Remember ImageNet? Improvements obtained...

by **looking at the data!**

- **Image Examples** (Confusing label)

Predict:
1 *carton*
2 *packet*
3 *toilet tissue*
4 *vending machine*
5 *crate*

Ground Truth:
sunscreen

http://image-net.org/challenges/talks/2016/Trimps-Soushen@ILSVRC2016.pdf

Remember ImageNet? Improvements obtained...

by **looking at the data!**

- **Image Examples** (Non-obvious main object)

Predict:
1 *dock*
2 *submarine*
3 *boathouse*
4 *breakwater*
5 *lifeboat*

Ground Truth:
paper towel

http://image-net.org/challenges/talks/2016/Trimps-Soushen@ILSVRC2016.pdf

Remember ImageNet? Improvements obtained...

by **looking at the data!**

- **Image Examples** (Label may wrong)

Predict:
1 *pencil box*
2 *diaper*
3 *bib*
4 *purse*
5 *running shoe*

Ground Truth:
sleeping bag

http://image-net.org/challenges/talks/2016/Trimps-Soushen@ILSVRC2016.pdf

Discovery of new risk factors

Medications on Admission:
BENADRYL 25MG--Take 2 by mouth at bedtime
BENZAMYCINPAK 3-5%--Apply twice a day to face for acne
CELEXA 20MG--Take one by mouth at bedtime
COLACE 100MG--1-2 tabs by mouth every day as needed
DOX
Dox
RIS

CHEST

Original Research

CRITICAL CARE

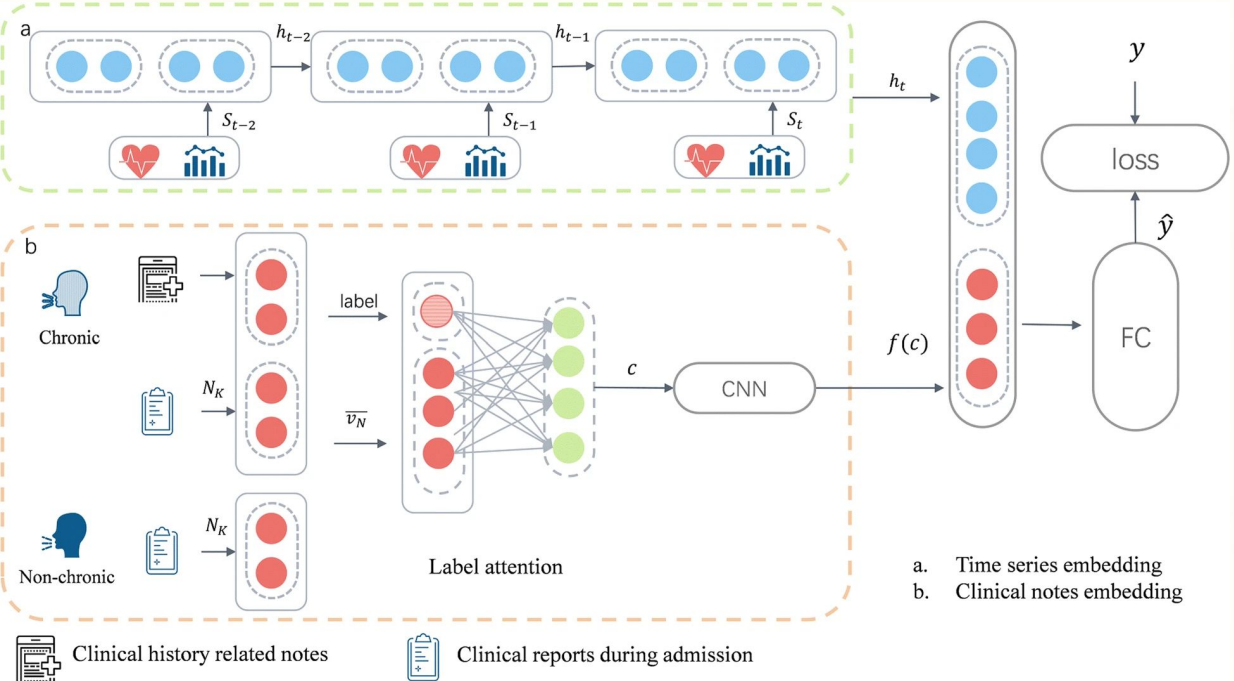
Leveraging a Critical Care Database

Selective Serotonin Reuptake Inhibitor Use Prior to ICU Admission Is Associated With Increased Hospital Mortality

Marzyeh Ghassemi, MS; John Marshall, PharmD; Nakul Singh, MS; David J. Stone, MD; and Leo Anthony Celi, MD, MPH

Development of new machine learning methods

- Health care requires processing sparse, heterogenous, multimodal, unevenly sampled observations which give an incomplete picture of health



Yang H, Kuang L, Xia F. Multimodal temporal-clinical note network for mortality prediction. Journal of Biomedical Semantics. 2021 Dec;12(1):1-4.

Other mechanisms have serious limitations

- The “process the data a bit” approach
  - E.g. extract diagnoses from free-text, release those diagnoses

1. *unremarkable* cardiomediastinal silhouette

2. diffuse reticular pattern, which can be seen with an atypical infection **or** chronic fibrotic change. *no* focal consolidation.

3. *no* pleural effusion or pneumothorax

4. mild degenerative changes in the lumbar spine and old right rib fractures.

Observation	Labeler Output
No Finding	
Enlarged Cardiom.	0
Cardiomegaly	
Lung Opacity	1
Lung Lesion	
Edema	
Consolidation	0
Pneumonia	u
Atelectasis	
Pneumothorax	0
Pleural Effusion	0
Pleural Other	
Fracture	1
Support Devices	

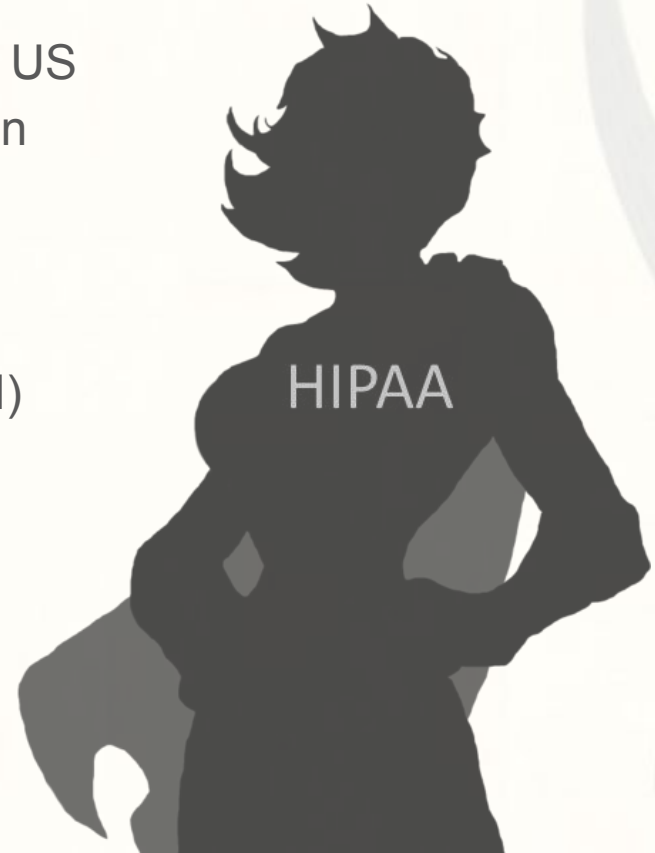
The above task is **hard.**

What enabled us to share the data?

Deidentification saves the day!

Health Insurance Portability and Accountability Act

- HIPAA was a landmark act in the US stipulating how patient information may be processed and handled
- HIPAA provides a **prescriptive definition** of what constitutes protected health information (PHI)
- HIPAA prevents linkage attacks
  - 5 digit zip codes are not allowed
  - Date of birth must be removed
  - Real dates must be removed



Deidentification according to HIPAA

- Remove the 18 identifiers stipulated in the Safe Harbor provision
- Data may now be made broadly available to researchers
- Everyone benefits!

No.	PHI Type
1	Names
2	All geographic subdivisions smaller than a state
3	Dates
4	Telephone Numbers
5	Vehicle Identifiers
6	Fax Numbers
7	Device Identifiers and Serial Numbers
8	Emails
9	URLs
10	Social Security Numbers
11	Medical Record Numbers
12	IP Addresses
13	Biometric Identifiers
14	Health Plan Beneficiary Numbers
15	Full-face photographic images and any comparable images
16	Account Numbers
17	Certificate/license numbers
18	Any other unique identifying number, characteristic, or code.

Table 1: Personal identifiers defined by HIPAA Safe Harbor legislation.

What about Ontario?

Personal Health Information Protection Act (PHIPA)

Mostly focused on rules for handling information for healthcare organizations

Research (Section 44) - to share **personal health information**, you must:

- Apply with a research plan
- Have REB approval for that research plan
- Agree to the rules set out by whoever has the data

… but what if it's deidentified?

IPC Office Guidance

Information and Privacy Commissioner of Ontario
Commissaire à l'information et à la protection de la vie privée de l'Ontario

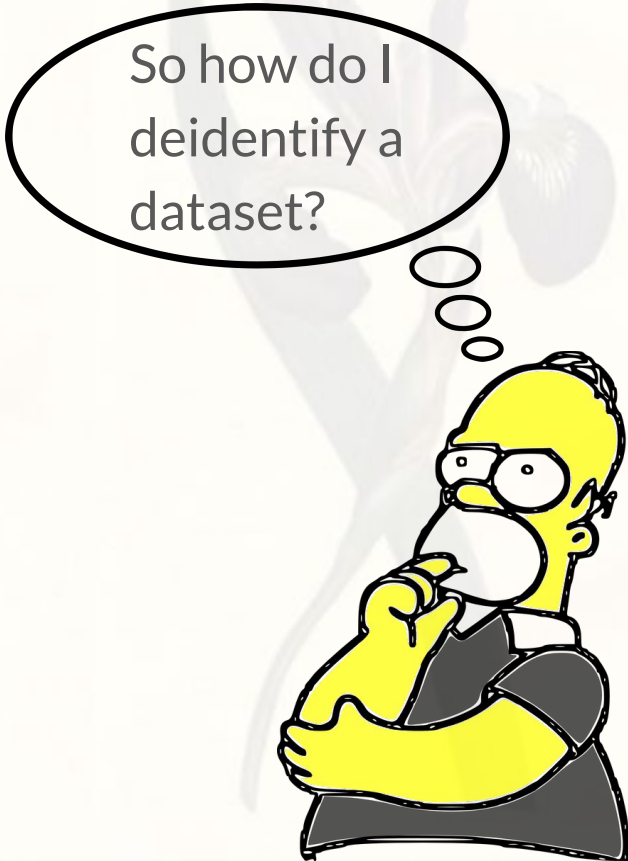
So how do I deidentify a dataset?

If a data set does not contain personal information, its use or disclosure cannot violate the privacy of individuals. Accordingly, the privacy protection provisions of the Freedom of Information and Protection of Privacy Act (FIPPA) and the Municipal Freedom of Information and Protection of Privacy Act (MFIPPA) would not apply to de-identified information.

https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf

PHIPA

DO



PHIPA

No.	PHI Type
1	Names
2	All geographic subdivisions smaller than a state
3	Dates
4	Telephone Numbers
5	Vehicle Identifiers
6	Fax Numbers
7	Device Identifiers and Serial Numbers
8	Emails
9	URLs
10	Social Security Numbers
11	Medical Record Numbers
12	IP Addresses
13	Biometric Identifiers
14	Health Plan Beneficiary Numbers
15	Full-face photographic images and any comparable images
16	Account Numbers
17	Certificate/license numbers
18	Any other unique identifying number, characteristic, or code.

Table 1: Personal identifiers defined by HIPAA Safe Harbor legislation.

So what now?

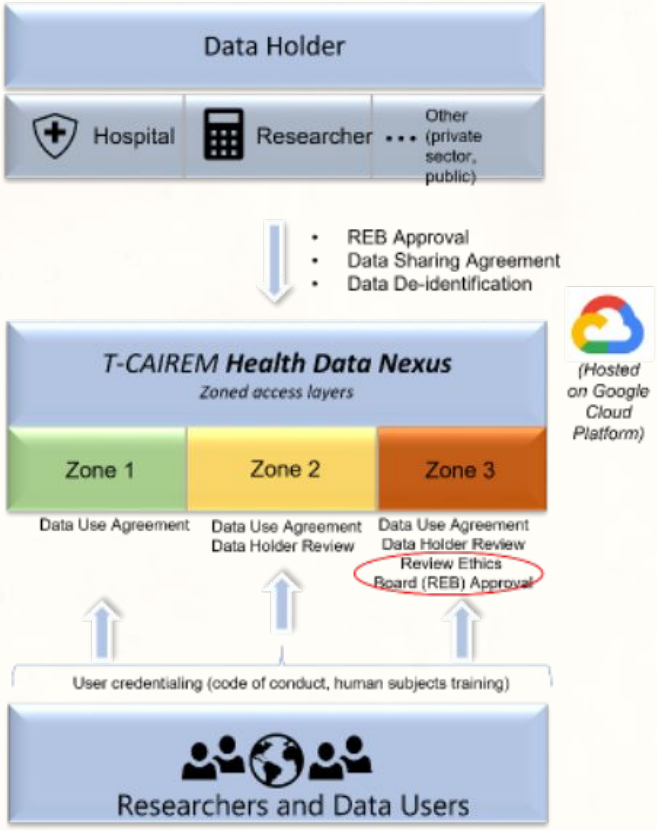
Share your data!

MIMIC has inspired other initiatives

Amsterdam
Medical Data Science
Connecting Healthcare and Data Science

HiRID - high time resolution ICU data set

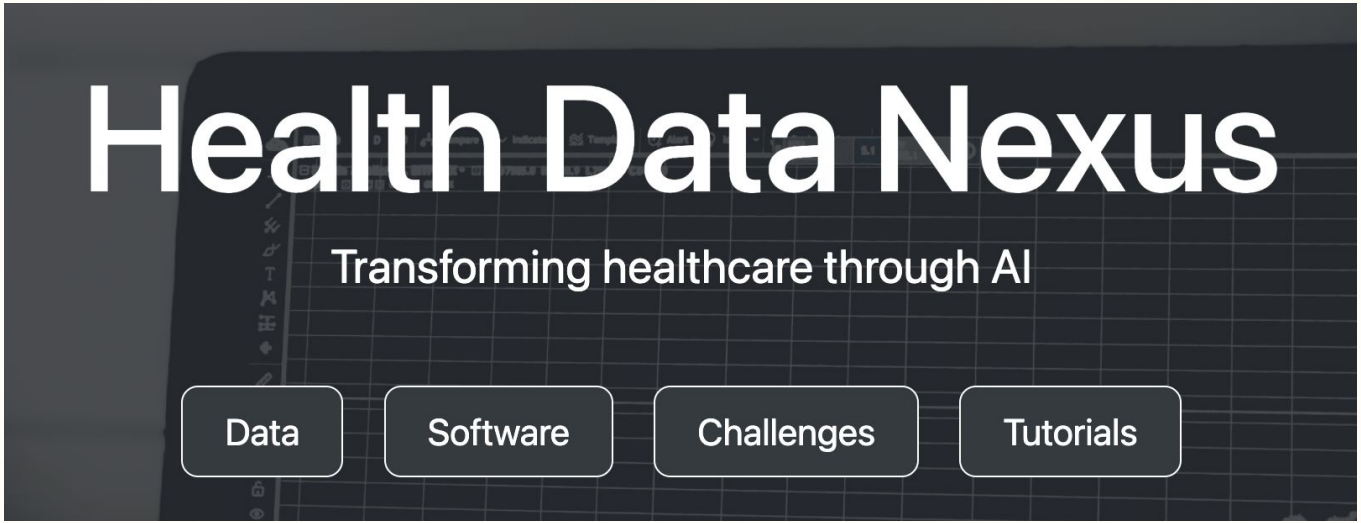
Health Data Nexus



Principles of Health Data Nexus

- **Remove risk** by only storing deidentified data and **prohibiting data egress** (technically & legally)
- **Accelerate research** through a transparent, streamlined, and equitable access policy
- **Scale compute** by integrating commercial cloud vendor services (initially GCP)

https://dev.healthdatanexus.ai



Requires UTORID currently - expanding to all Canadian universities soon!

Thanks to T-CAIREM and Temerty family donation!

https://tcairem.utoronto.ca/ contact@healthdatanexus.ai