Let's Talk Informatics

Medical imaging and machine learning

The long journey to get clinical data into an AI model

- Audience audio and video options have been disabled.
- To interact in the Q & A portion of the presentation, type your question in the chat window **and select the "all panelist" option to direct your question**.
- Today's session is being recorded and registered guests will be emailed a link to access from EventBrite.
- Want to stay informed about future sessions? Get on our mailing list here: **letstalkinformatics@nshealth.ca.**

Acknowledgement

We acknowledge we are gathered today
in Mi'kma'ki (*Mig-***_maw_**-gee), the traditional ancestral
unceded territory of the Mi'kmaq (*Mig-**maw**) people.

Informatics

Informatics utilizes health information and health care technology to enable patients to receive best treatment and best outcome possible.

Let's Talk Informatics Objectives

This series is designed to enable participants to:

- Identify knowledge and skills healthcare providers need in order to use information now, and in the future.
- Prepare health care providers through an introduction to concepts and experiences in Informatics.
- Acquire knowledge to remain current by becoming familiar with new trends, terminology, studies, data and news.
- Collaborate with a network of colleagues to establishing connections with leaders who can provide advice on business issues, best-practice and knowledge sharing.

Let’s Talk Informatics

Medical imaging and machine learning

The long journey to get clinical data into an AI model

Dr. Alex Guida & Jeff Kowalski

October 27, 2022

Conflict of Interest Declaration

Part of our research is sponsored by major industry manufacturing partners and startups in the biomedical field.

Session Specific Objectives

- At the conclusion of this activity, you will be able to:
  - Understand the components and part required to run a machine learning research project in medical imaging.
  - Gain an intuition of the common problems and challenges in such processes.
  - Gain familiarity with the research infrastructures and resources available to our local community.

Abstract

- Running a research project to apply machine learning to medical imaging is not trivial. It takes a whole set of people, skills and tools to get the data into the right shape. This talk will outline common pitfalls that should be considered from the start and how we, at the Biomedical Translational Imaging Centre, are tackling these problems.

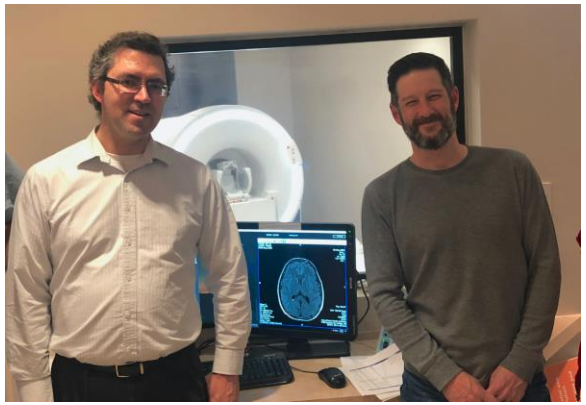
Introduction

https://bioticimaging.ca

BIOTIC (**BIO**medical **T**ranslational **I**maging **C**entre) is a multi-site imaging centre that is embedded in the two leading research and teaching hospitals in Nova Scotia. Our multidisciplinary and cross-functional teams, provide expertise in all facets of imaging research and development, collaborate on commercial development projects with industry partners as well as research and development projects with a number of institutions. Our advanced pre-clinical and clinical imaging equipment are housed in three labs, in two health centres encompassing over 12,000 square feet of lab space.

Introduction

https://bioticimaging.ca



Imaging Modalities we work with:

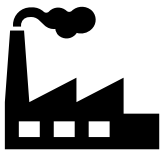
- MRI
- CT
- Ultrasound

Introduction

Academic/Clinical Research

With: Institutions
Goal: Scientific Publication

BIOTIC
Imaging for Life



Industrial Research

With: Commercial partners
Goal: Generate IP

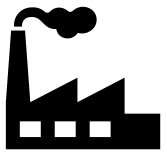
Introduction

Academic/Clinical Research

With: Institutions
Goal: Scientific Publication

Focus of this presentation

BIOTIC
Imaging for Life



Industrial Research

With: Commercial partners
Goal: Generate IP

Introduction

Academic/Clinical Research

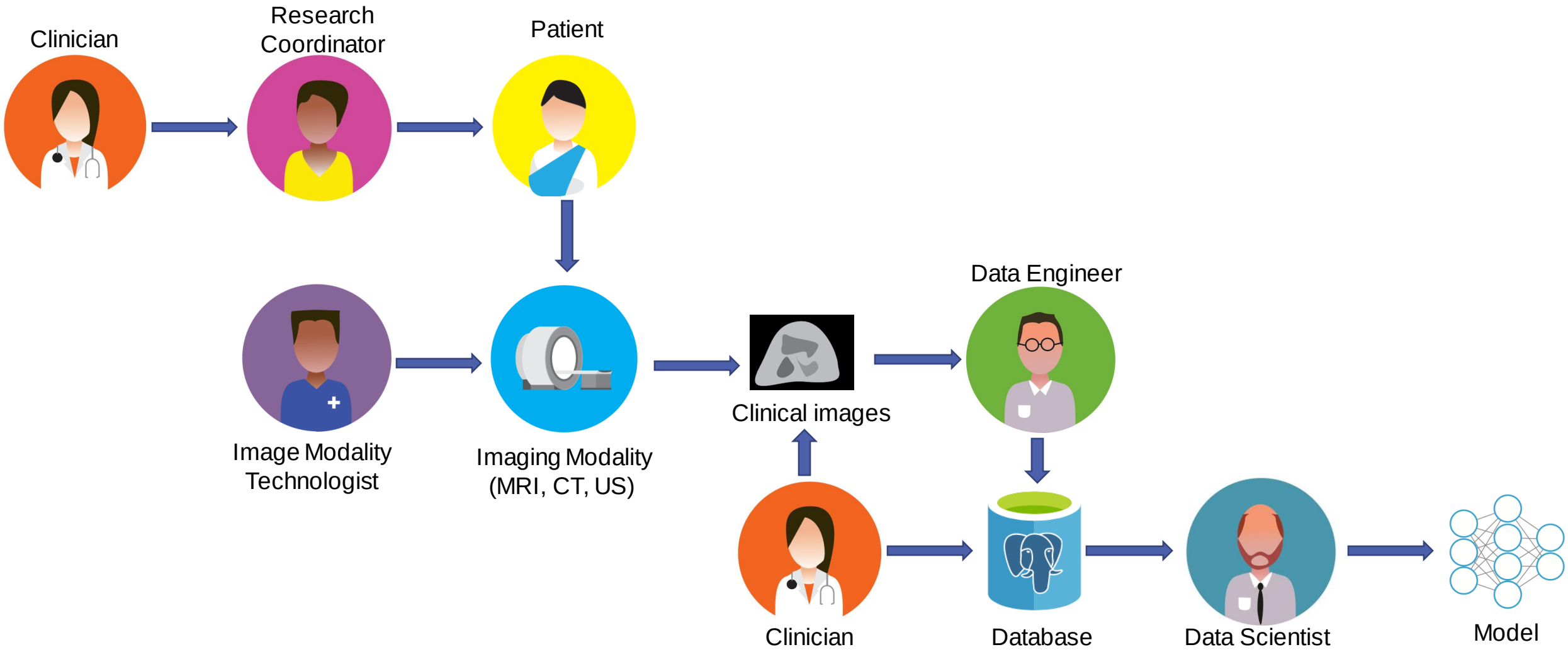
With: Institutions
Goal: Scientific Publication

You would like to apply **machine learning** on **biomedical images** to investigate your research hypothesis

Common questions

- Where do I start?
- Where and how do I get the data?

The journey of the data

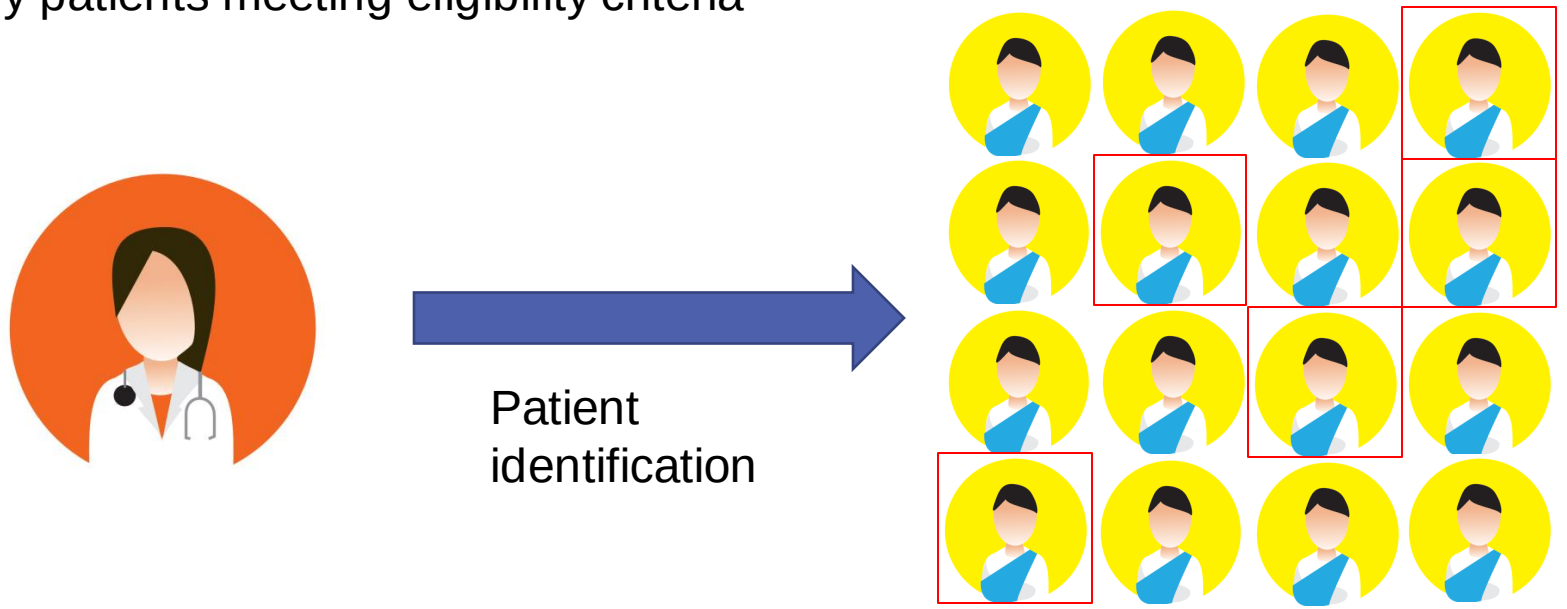


Getting the Data

Retrospective or Prospective study?

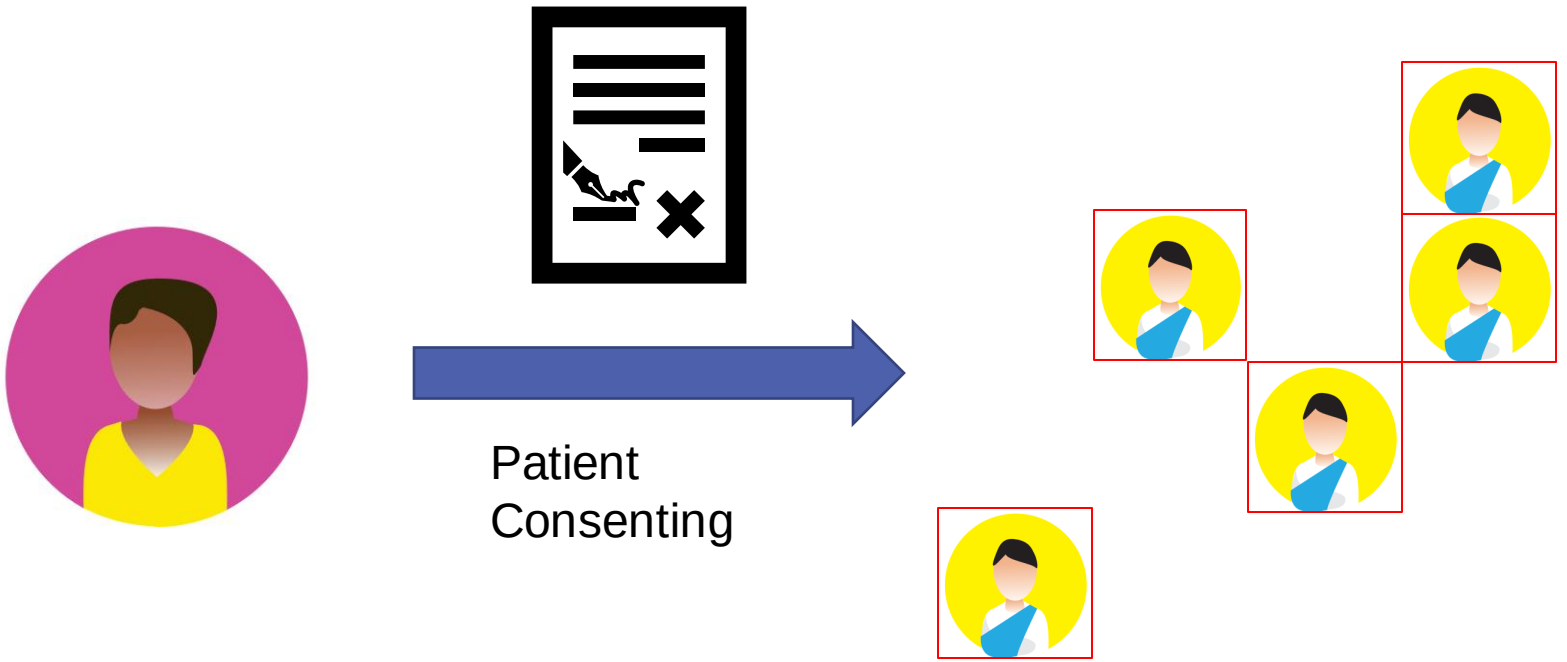
Prospective Study

- Clinicians identify patients meeting eligibility criteria



Getting the Data

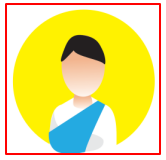
1) RC (Research Coordinator) consents patients



Getting the Data

1) RC (Research Coordinator) consents patients
2) **RC assigns a Unique identifier for each patient**

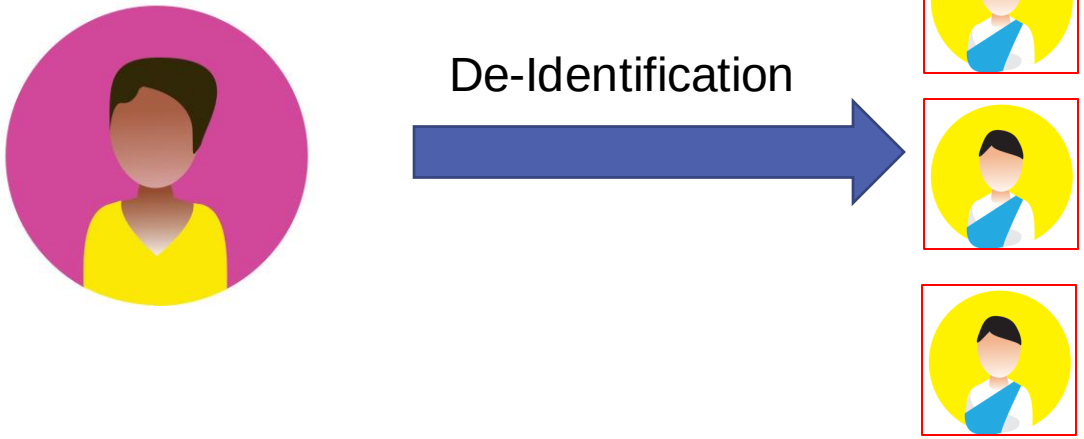
De-Identification



ID
1001
9911
8831
4291
7256

Getting the Data

1) RC (Research Coordinator) consents patients
2) RC assigns a Unique identified for each patient
3) **RC schedules a scan with the patient**



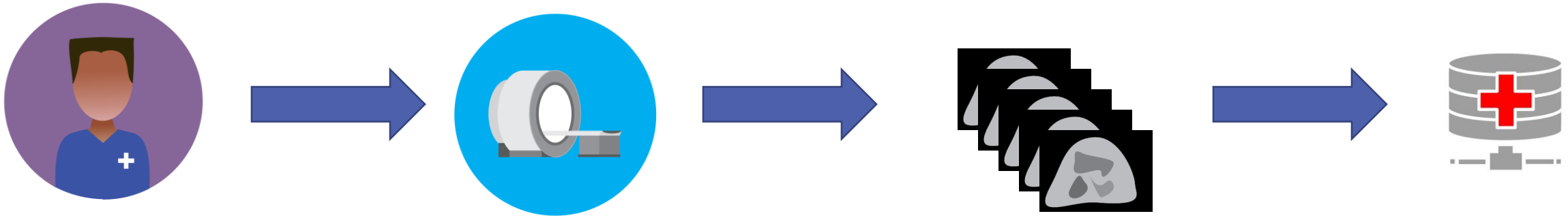
ID	Appointment
1001	11 June 2022
9911	13 Sept 2022
8831	5 Sept 2022
4291	12 Sept 2022
7256	10 Sept 2022

Getting the Data

Image acquisition

- The Technologist scans the patient & acquires the data
- Data are pulled from PACS to the AW Server

ID	Appointment
ID = 1001	11 June 2022



Technologist

Imaging Modality (MRI, CT, US)

Clinical images

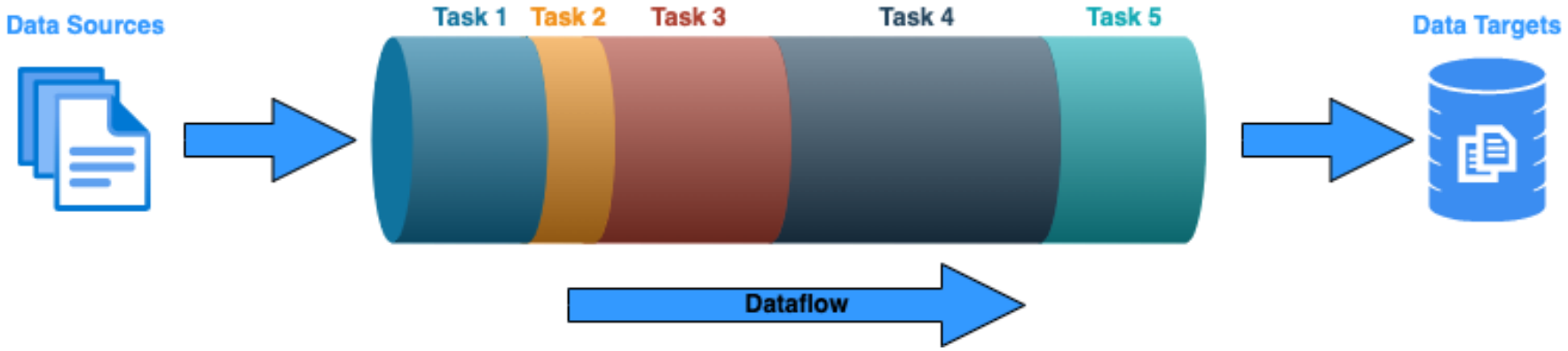
PACS

Meet the Data Engineer

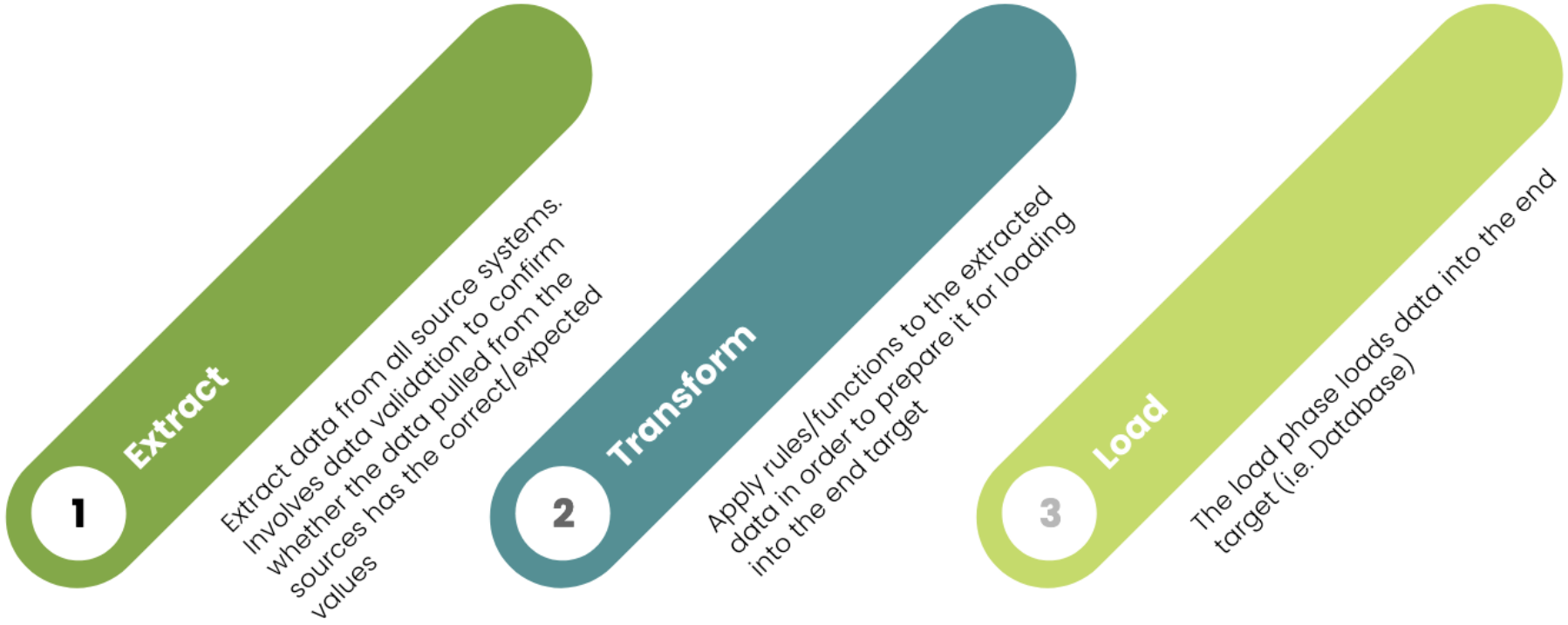
A data engineer is **an IT worker whose primary job is to prepare data for analytical or operational uses**. These software engineers are typically responsible for building data pipelines to bring together information from different source systems. They create the architectures that allow the data to flow to the data scientists who generate insights on the value of the data.

What is a Data Pipeline

- Data is like oil and natural gas but in another way – it flows through pipelines. A data pipeline ensures the efficient flow of data from one location to the other. A good pipeline allows your organization to integrate new data sources faster, provide patterns that you can replicate, gives you confidence in your data quality, and builds in security.
- A data pipeline is a set of actions (tasks) that ingest raw data from disparate sources and move the data to a destination for storage and analysis. We like to think of this transportation as a pipeline because data goes in at one end and comes out at another location (or several others).



What is an ETL?

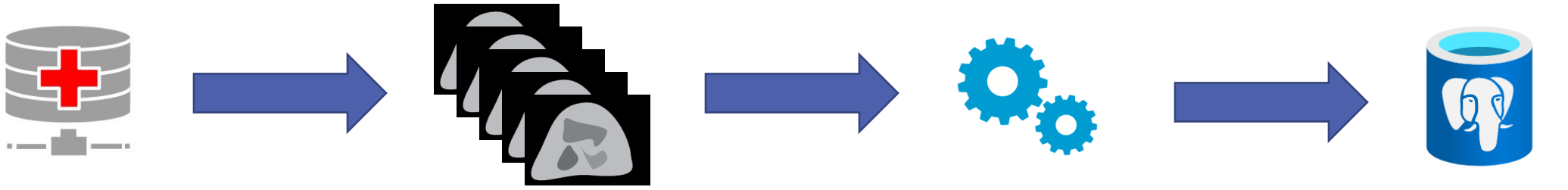


ETL Process: extract data from different sources, transform the data into a usable and trusted resource, and load that data into the systems end-users can access and use downstream to solve business problems.

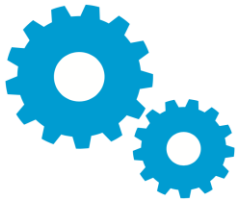
ETL – Extract Transform Load

Data transformation: Anonymization

- Clinical images are pulled from PACS onto the AW Server
- Fully identified imaging datasets are pushed to BIOTIC's anonymization tools
- De-identified datasets loaded into database



Anonymization



- DICOM data follows a file format where a ***Patient*** has ***Study*** that contains ***Series*** that contains ***Instance***.
- Need to ensure that no PHI exists within the file structure
- Anonymization consists in erasing all the tags that are specified in Table E.1-1 from PS 3.15 of the DICOM standard 2008, 2017c or 2021b (default)
- Anonymizer is its own secure (password protected) server within the hospital firewall
- Anonymizer is an *Orthanc Server* using the REST API with Python
  - *Orthanc is a lightweight, open-source DICOM server for medical imaging*
- After the DICOM tags are erased during anonymization, the collected datasets are then further analyzed to ensure no patient data was included in the DICOM Instances.

ORTHANC

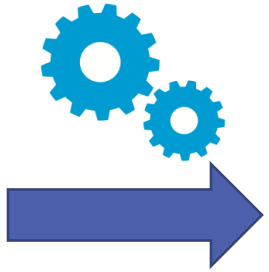
DICOM Anonymizer

docker

Anonymization

An example DICOM file header before and after anonymization

Field Name	Tag	Content
DICOMObject		
MetaElementGroupLength	0002,0000	210
FileMetaInformationVersion	0002,0001	0x0001
MediaStorageSOPClassUID	0002,0002	1.2.840.10008.5.1.4.1.1.6.1
MediaStorageSOPInstanceUID	0002,0003	1.2.840.113619.2.323.5501492376
TransferSyntaxUID	0002,0010	1.2.840.10008.1.2.1
ImplementationClassUID	0002,0012	1.3.6.1.4.1.5962.99.2
ImplementationVersionName	0002,0013	PIXELMEDJAVA001
SourceApplicationEntityTitle	0002,0016	IW10307_11112
ImageType	0008,0008	ORIGINAL\PRIMARY\PEDIATRIC\00
SOPClassUID	0008,0016	1.2.840.10008.5.1.4.1.1.6.1
SOPInstanceUID	0008,0018	1.2.840.113619.2.323.5501492376
StudyDate	0008,0020	20130502
SeriesDate	0008,0021	20130502
ContentDate	0008,0023	20130502
StudyTime	0008,0030	104040.000000
SeriesTime	0008,0031	104040.000000
ContentTime	0008,0033	104607.000000
AccessionNumber	0008,0050	340037
Modality	0008,0060	US
Manufacturer	0008,0070	GE Healthcare
ReferringPhysiciansName	0008,0090	
StationName	0008,1010	
StudyDescription	0008,1030	
ManufacturersModelName	0008,1090	LOGIQS8
PatientsName	0010,0010	Doe, Jane
PatientID	0010,0020	Doe, Jane
PatientsBirthDate	0010,0030	10-27-1986
PatientsSex	0010,0040	F
PatientIdentityRemoved	0012,0062	NO

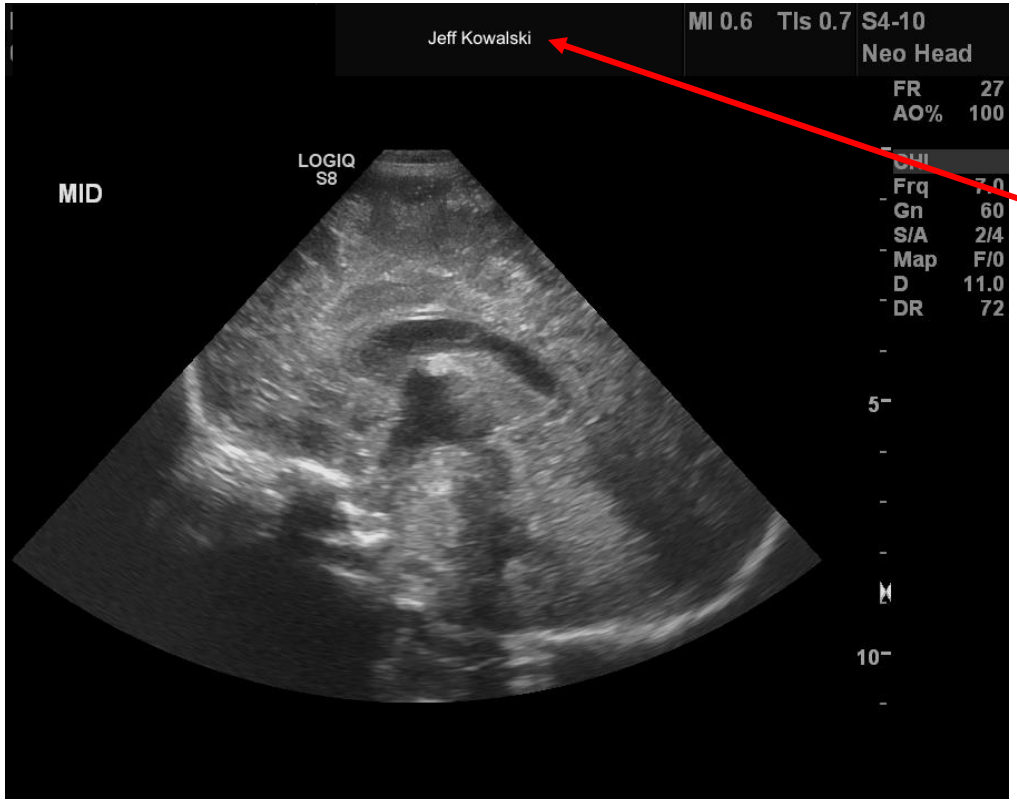


Field Name	Tag	Content
DICOMObject		
MetaElementGroupLength	0002,0000	210
FileMetaInformationVersion	0002,0001	0x0001
MediaStorageSOPClassUID	0002,0002	1.2.840.10008.5.1.4.1.1.6.1
MediaStorageSOPInstanceUID	0002,0003	1.2.840.113619.2.323.550149237622.1367502367.98
TransferSyntaxUID	0002,0010	1.2.840.10008.1.2.1
ImplementationClassUID	0002,0012	1.3.6.1.4.1.5962.99.2
ImplementationVersionName	0002,0013	PIXELMEDJAVA001
SourceApplicationEntityTitle	0002,0016	IW10307_11112
ImageType	0008,0008	
SOPClassUID	0008,0016	1.2.840.10008.5.1.4.1.1.6.1
SOPInstanceUID	0008,0018	1.2.840.113619.2.323.550149237622.1367502367.98
StudyDate	0008,0020	20130502
SeriesDate	0008,0021	20130502
ContentDate	0008,0023	20130502
StudyTime	0008,0030	104040.000000
SeriesTime	0008,0031	104040.000000
ContentTime	0008,0033	104607.000000
AccessionNumber	0008,0050	340037
Modality	0008,0060	US
Manufacturer	0008,0070	GE Healthcare
ReferringPhysiciansName	0008,0090	
StationName	0008,1010	
StudyDescription	0008,1030	RESEARCH PROJECT XYZ
ManufacturersModelName	0008,1090	LOGIQS8
PatientsName	0010,0010	S002
PatientID	0010,0020	S002
PatientsBirthDate	0010,0030	20130502
PatientsSex	0010,0040	
PatientIdentityRemoved	0012,0062	YES

** This person doesn't not exist*

Anonymization

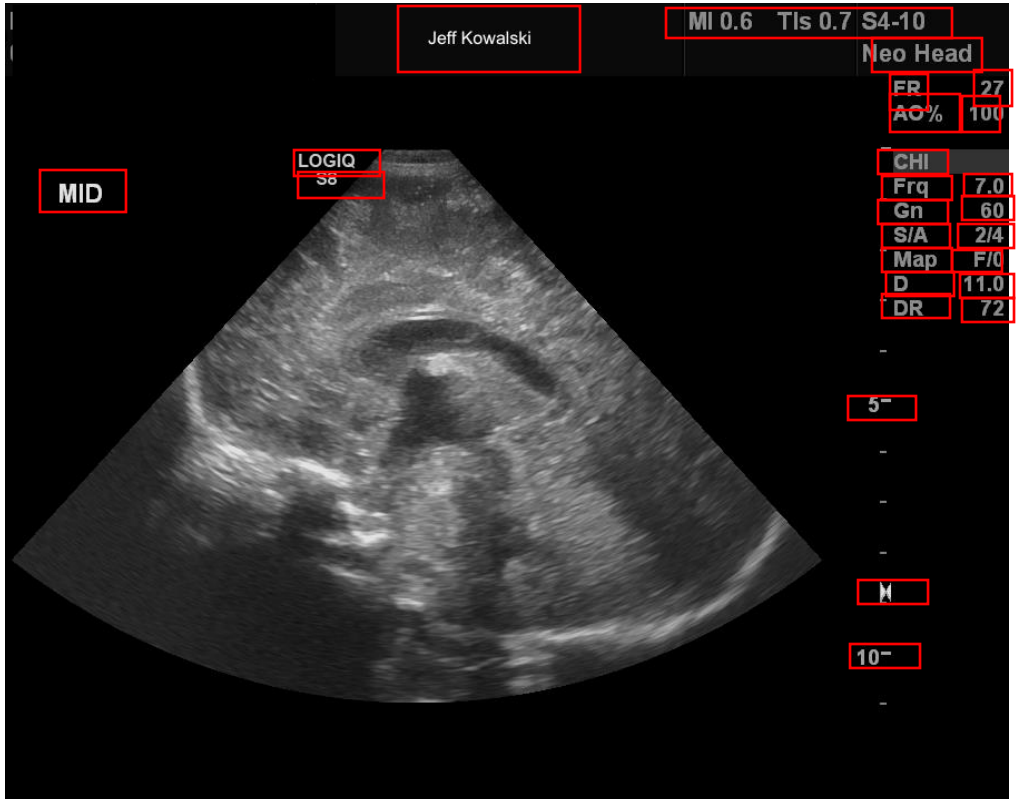
An example anonymizing images from ultrasound



This image contains PHI

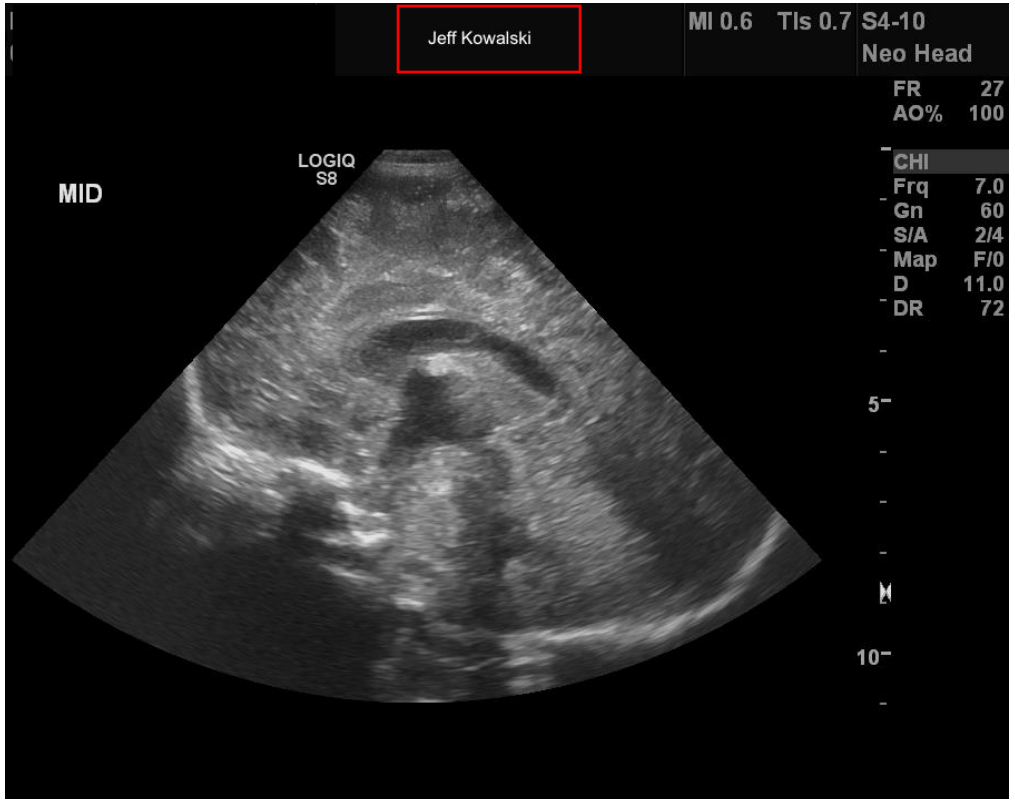
Anonymization

- Text detection - Uses OCR (Optical Character Recognition)



Anonymization

- Text detection
- PHI recognition - NLP (Natural Language Processing)

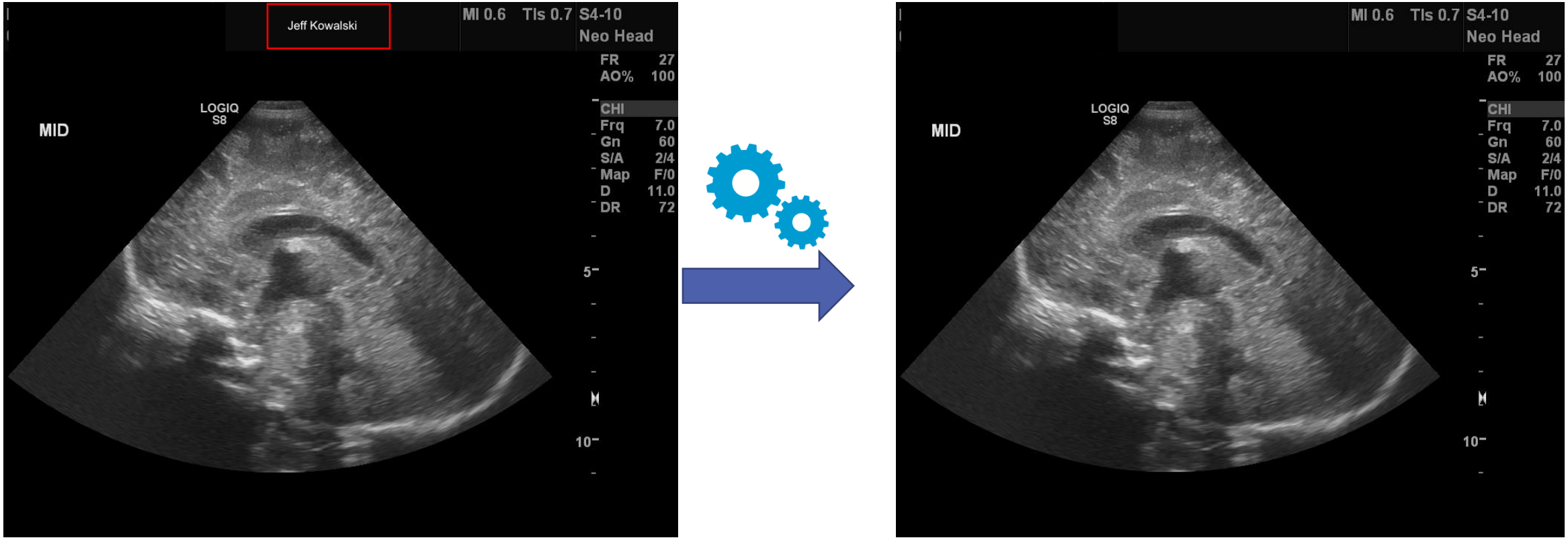


NLP Library will check if any of the text found contains a NAME

54 15 17 CARDINAL
27 33 35 CARDINAL
Jeff Kowalski 0 13 PERSON
CHI 0 3 ORG
7.0 8 11 CARDINAL
60 15 17 CARDINAL
MID 0 3 ORG

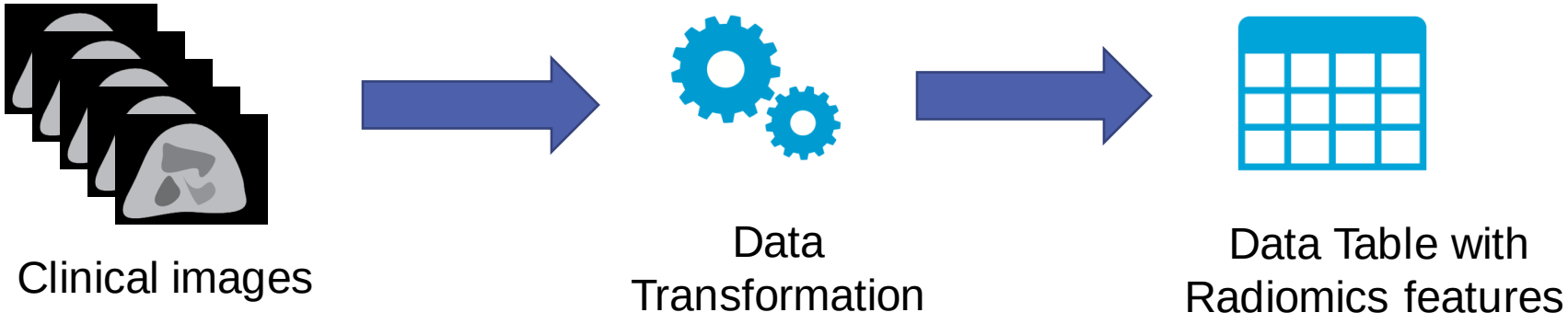
Anonymization

- Text detection
- PHI recognition
- Masking PHI information



Data Transformation

Anonymization is not the only type of data processing. Depending on the project there are several. One classical example is to convert **MRI images** into **radiomics features**



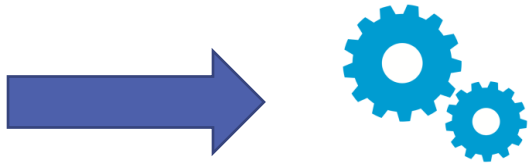
Data Transformation

Anonymization is not the only type of data processing. Depending on the project there are several. One classical example is to convert **MRI images** into **radiomics features**

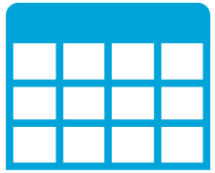
Clinical images

Very high dimensionality

Eg. 512 (height) * 512 (width) * 32 (slices)
= ~ 8,300,000 features per patient



Data Transformation

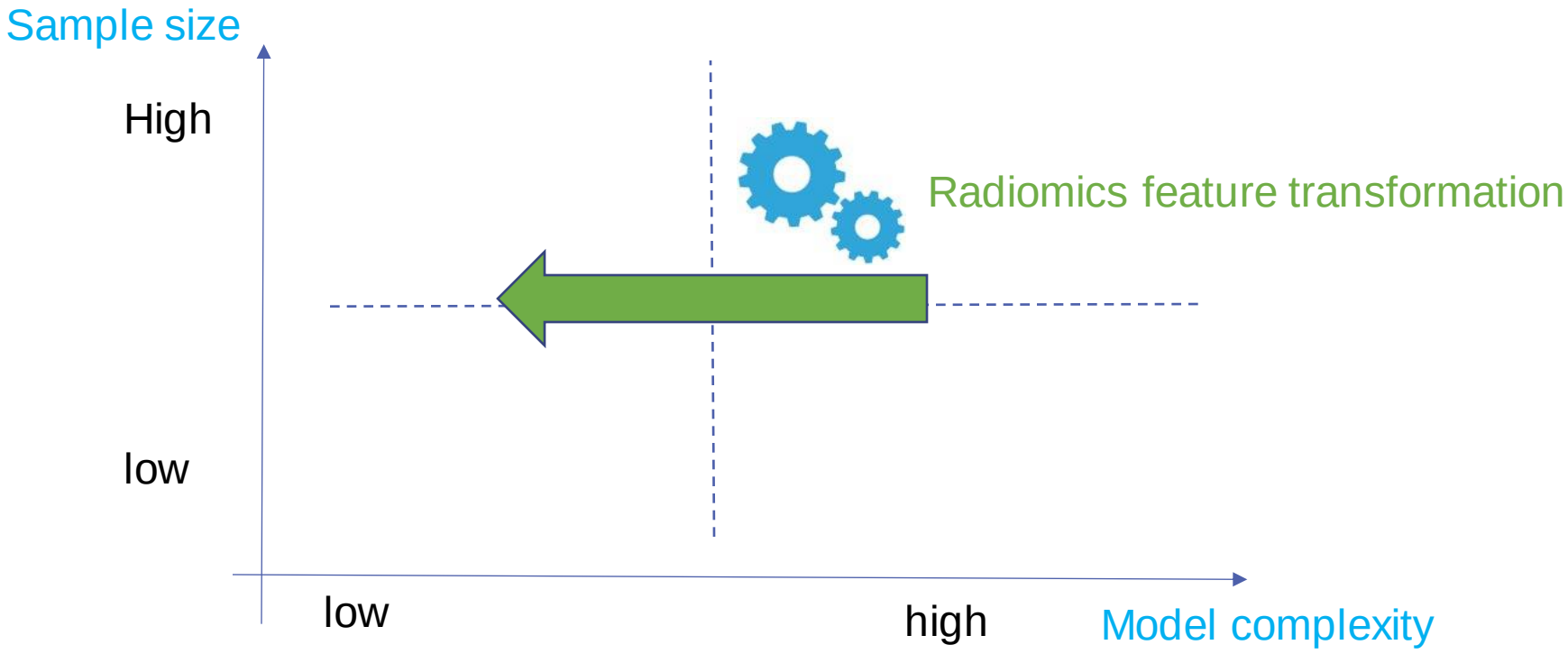


Data Table with Radiomics features

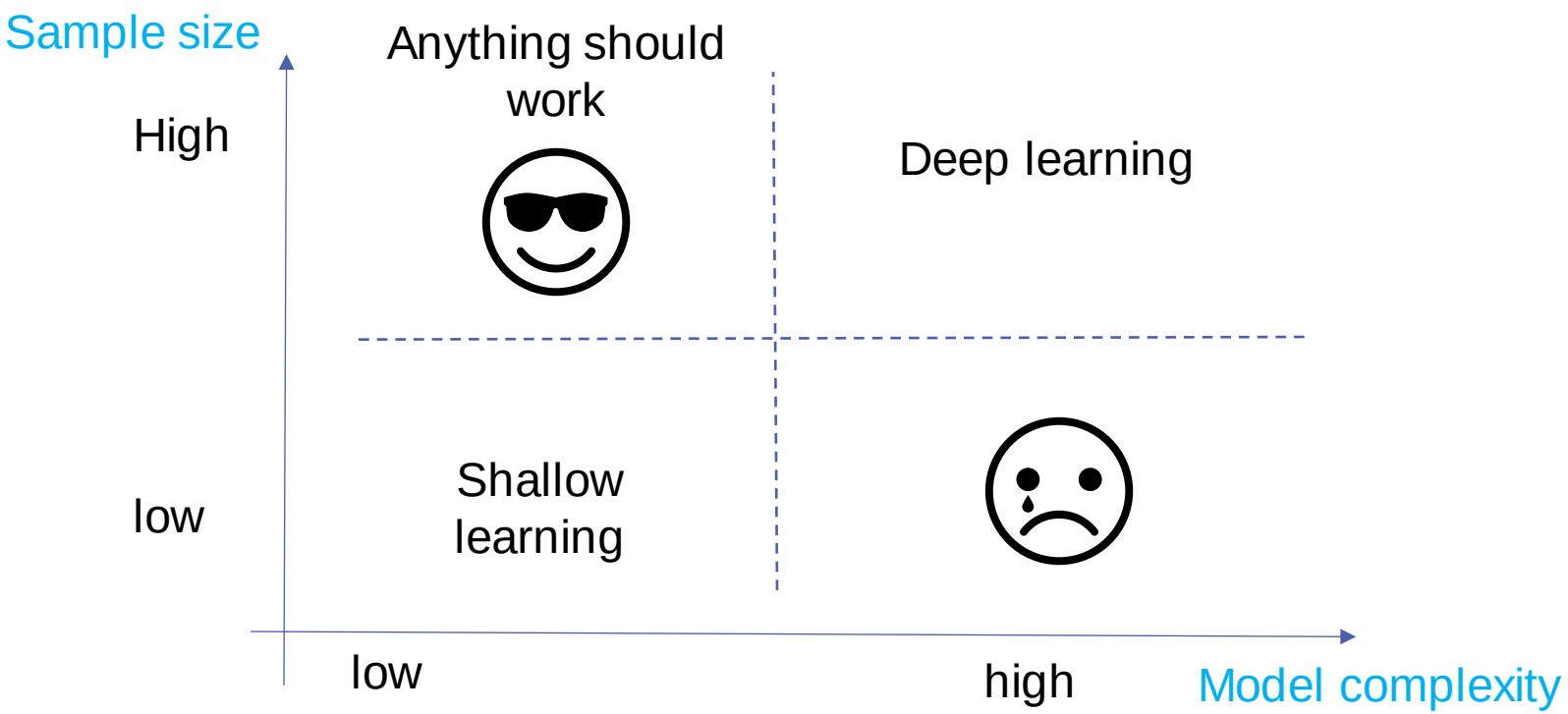
Low dimensionality

~ 100 features per patient

Data Transformation

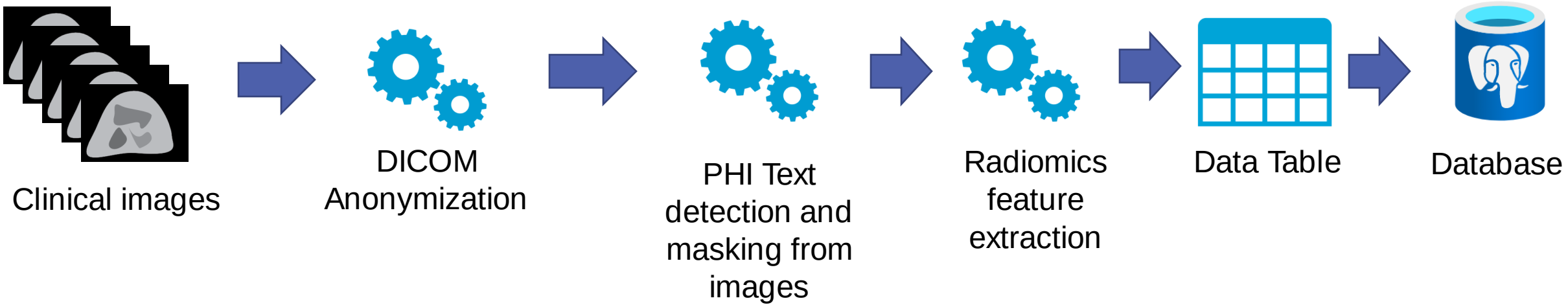


Data Transformation



Data Transformation

Example of a multi-step preprocessing pipeline

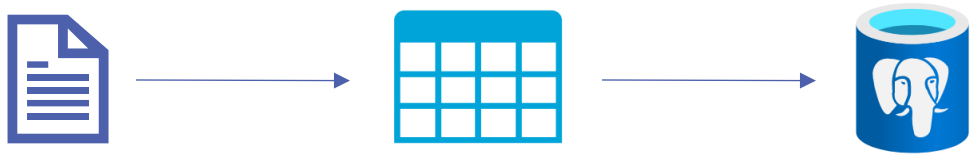


Data Storage

Data are now anonymized and stored on BIOTIC server infrastructure

How are they stored?

- Local filesystem (de-identified raw data)
- Dataframe (.csv)
- Structured database



Important considerations:

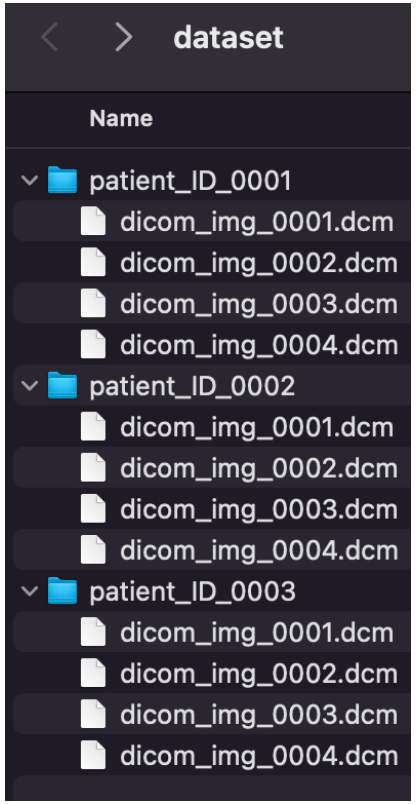
- Data structure
- Data versioning
- Data lineage

Data Storage - Structure

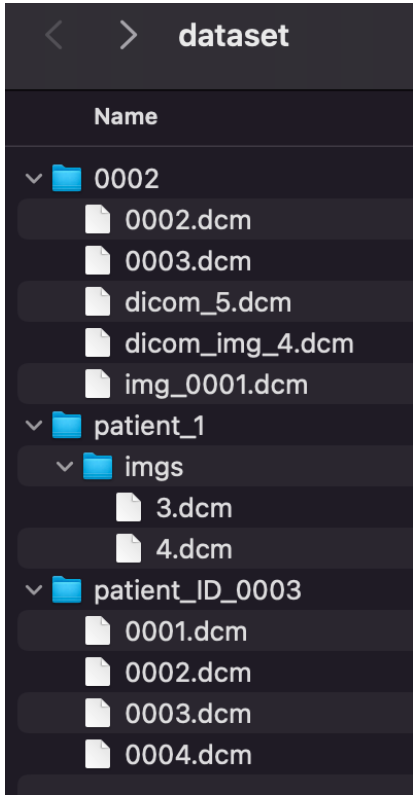
Simplified example to highlight the importance of a structured data format

Clean data structure is easy to parse computationally

Tidy data



Messy data



Error prone

Can cause mislabeling errors and/or a misinterpretation of the data

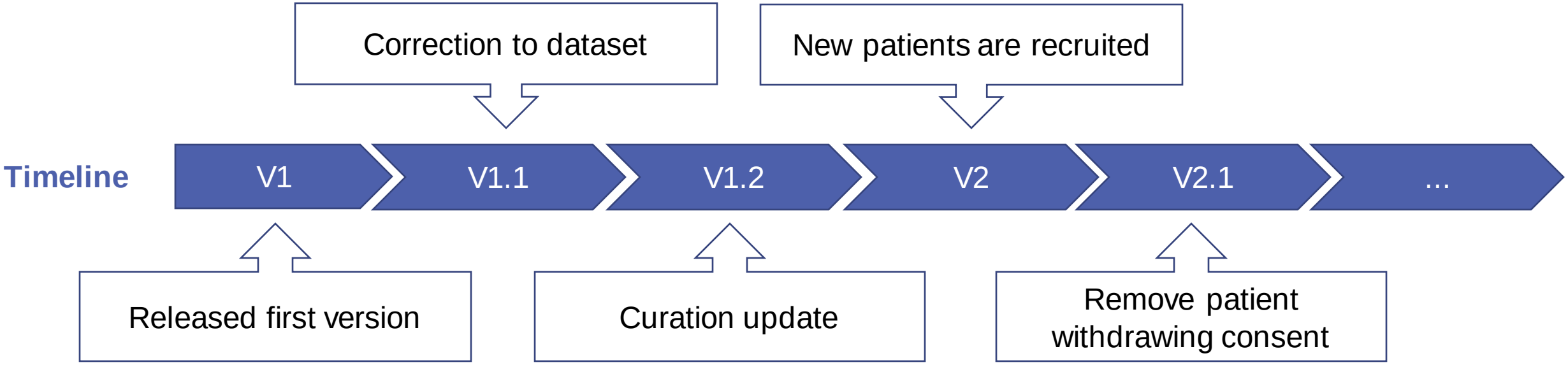
Problems

- No file naming convention
- Missing data
- Extra data or duplication
- No standard folder structure

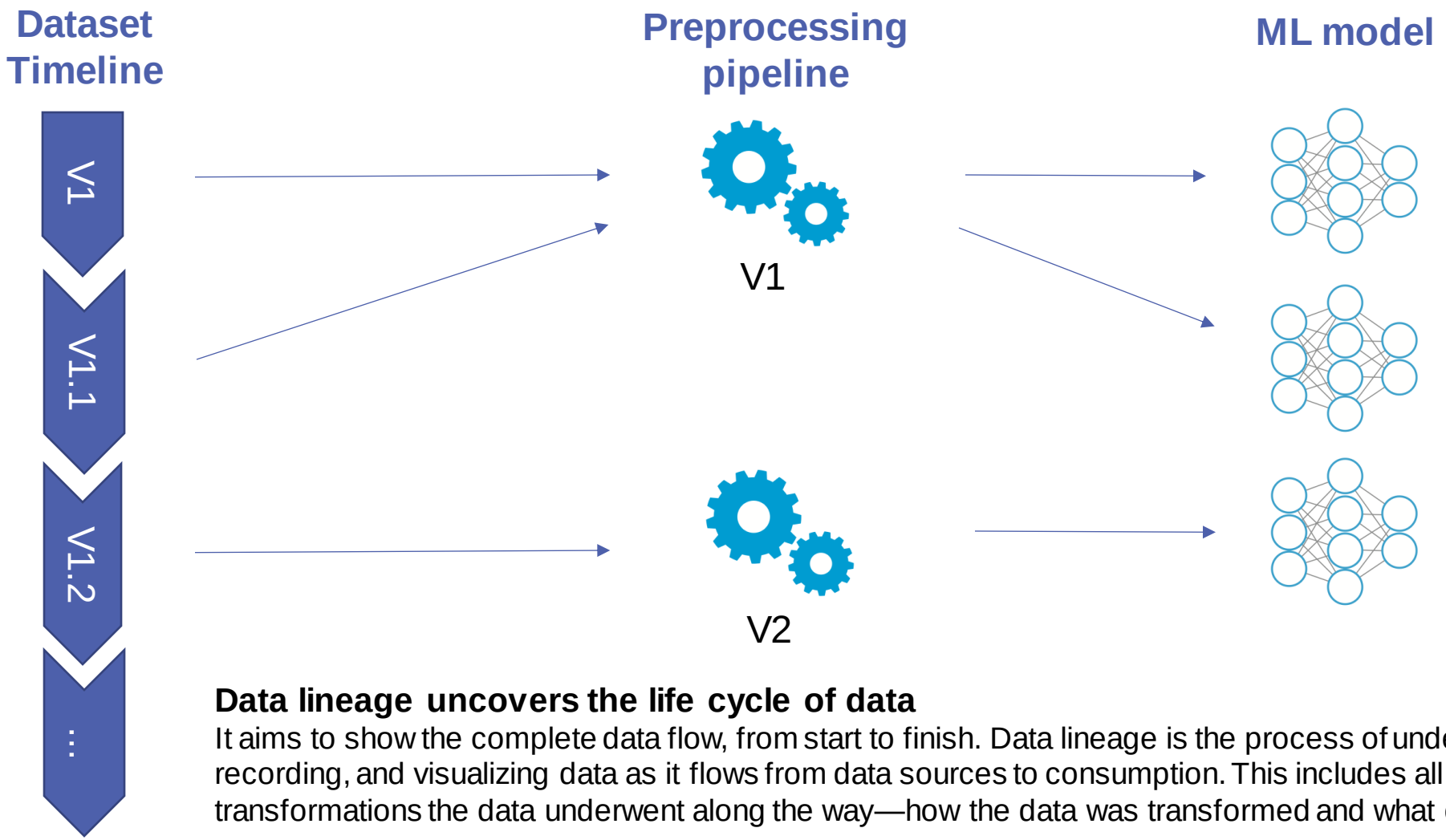
Data Storage - Versioning

Data is not fixed

Multiple changes can occur - a versioning system is necessary to track changes overtime



Data Lineage



Data lineage uncovers the life cycle of data

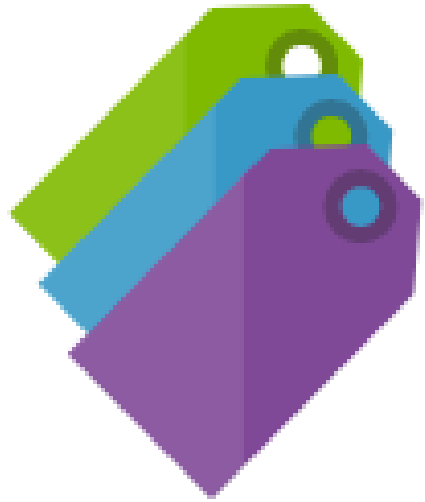
It aims to show the complete data flow, from start to finish. Data lineage is the process of understanding, recording, and visualizing data as it flows from data sources to consumption. This includes all transformations the data underwent along the way—how the data was transformed and what changed.

Data Curation

Two types of data curations:

Specification of **ROI** (Region of interest)

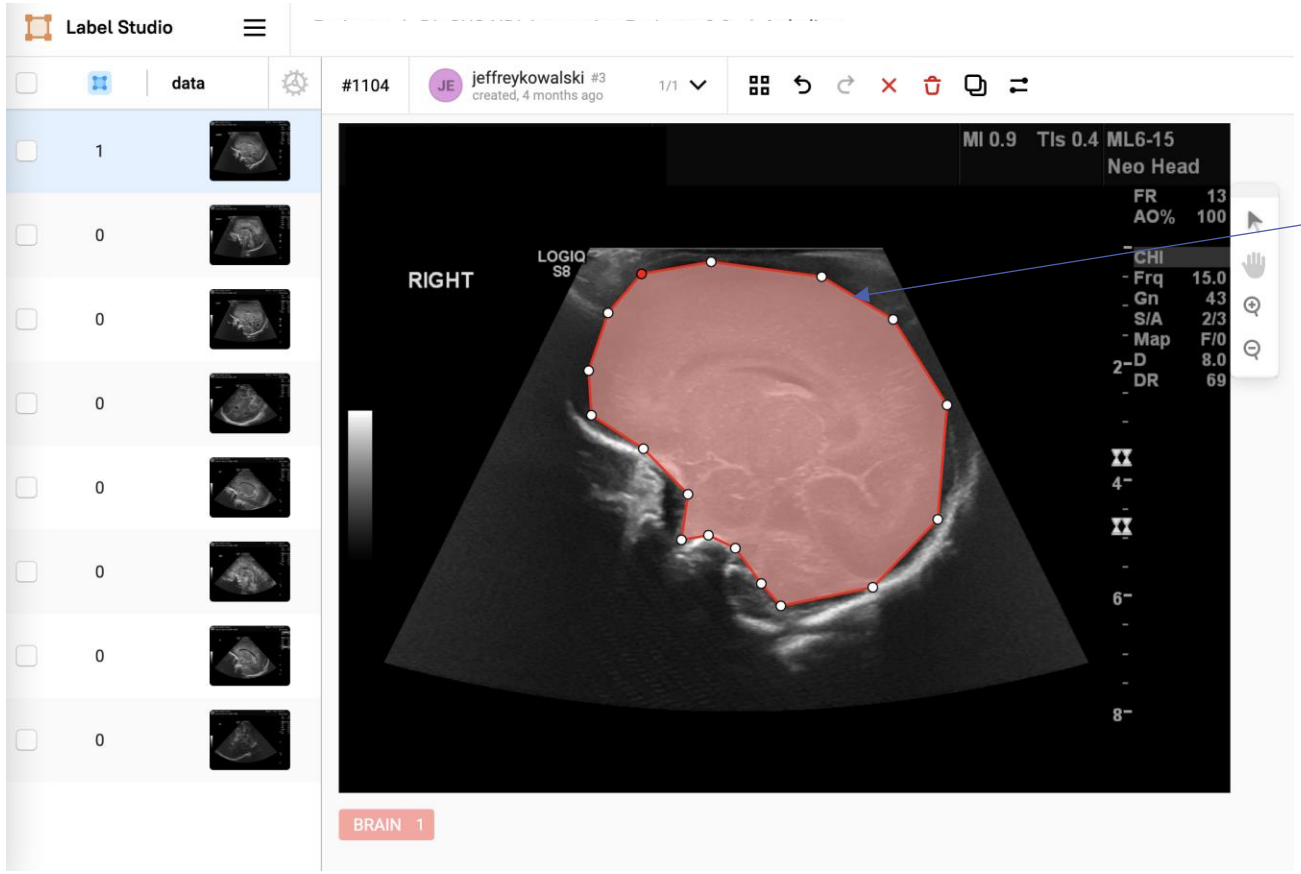
Inclusion of **Labels -** required for the ML task



Data Curation

Example Brain mask from an ultrasound image using Label Studio

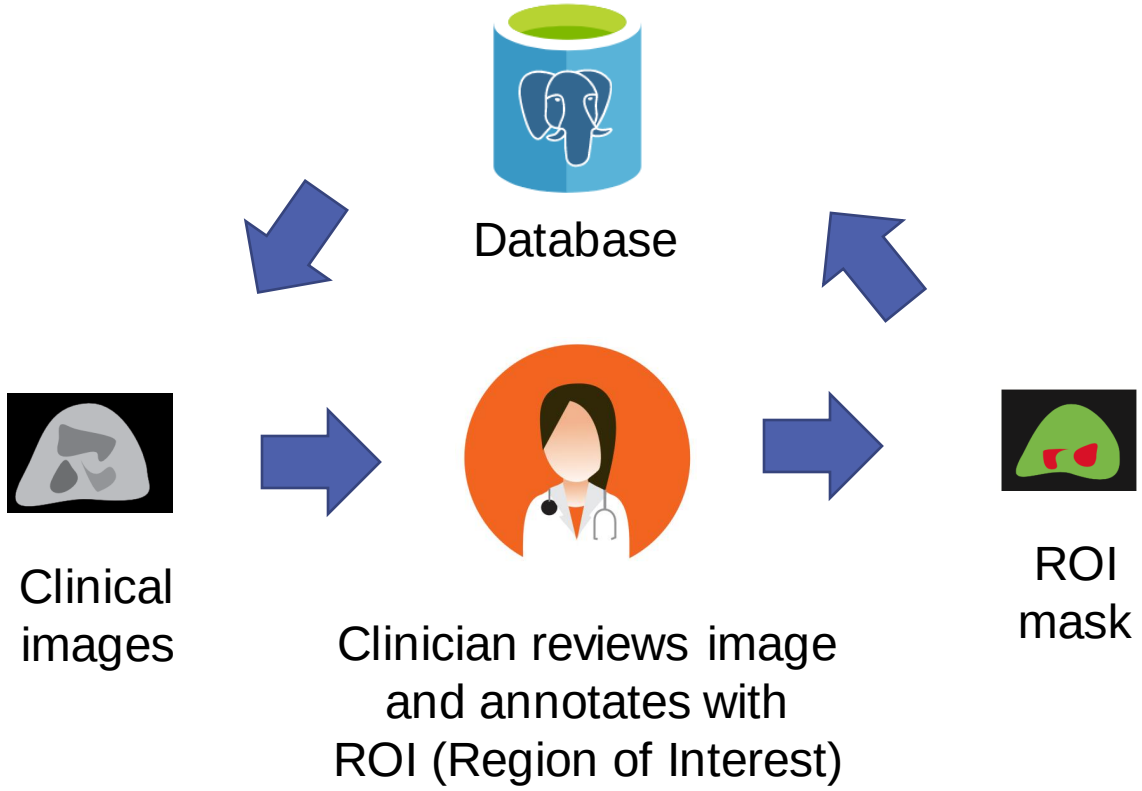
Label Studio
Imaging Labels & Annotations
docker



ROI
Region of interest

Data Curation - ROI

Data flow from database, getting annotated by a clinician, and then going back to the database for storage/versioning



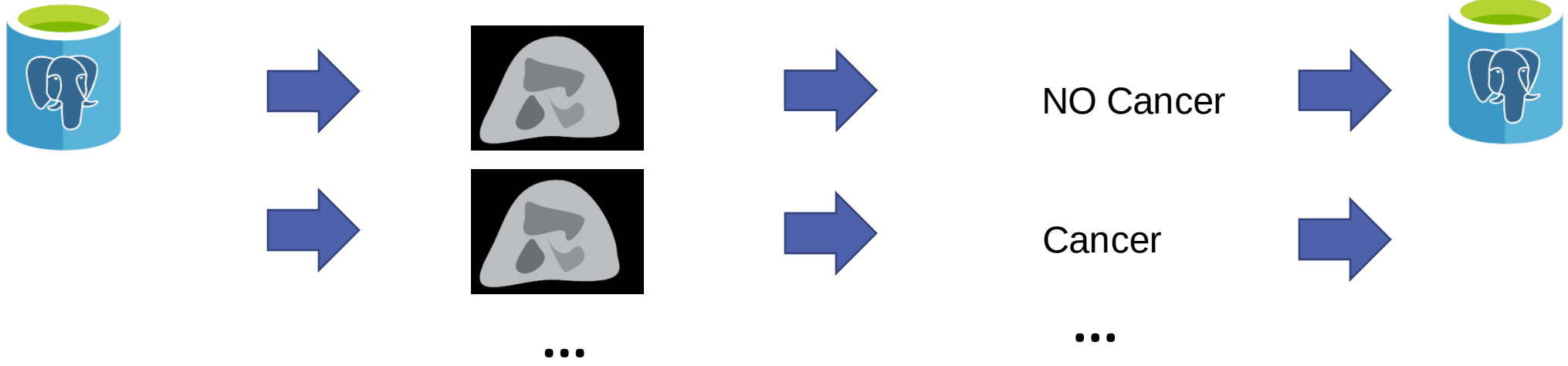
Data Curation - Labels

To address the problem with a machine learning framework we need to have our data **labeled** accordingly

ML task
Binary classification

Input

Label



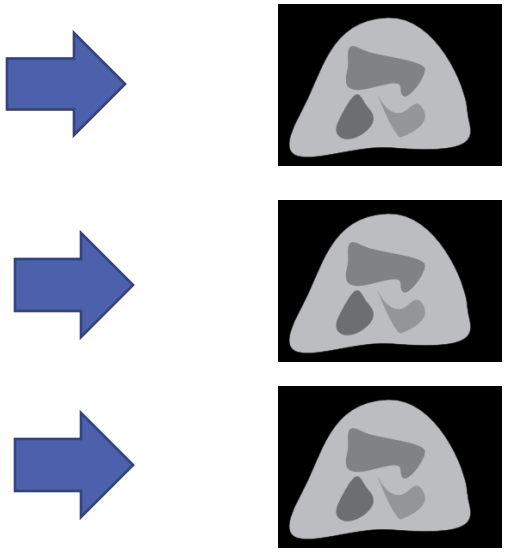
Data Curation - Labels

To address the problem with a machine learning framework we need to have our data **labeled** accordingly

ML task
Multiclass classification

Input

Label



Subtype 1

Subtype 2

Subtype 3

...

...

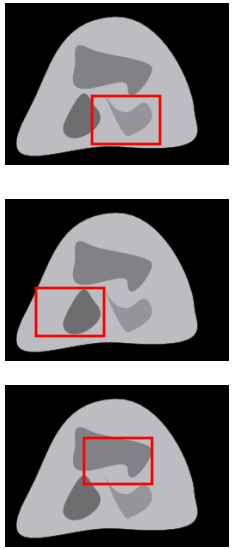
Data Curation - Labels

To address the problem with a machine learning framework we need to have our data **labeled** accordingly

ML task
Detection

Input

Label



Subtype 1 + Bounding box

Subtype 2 + Bounding box

Subtype 3 + Bounding box

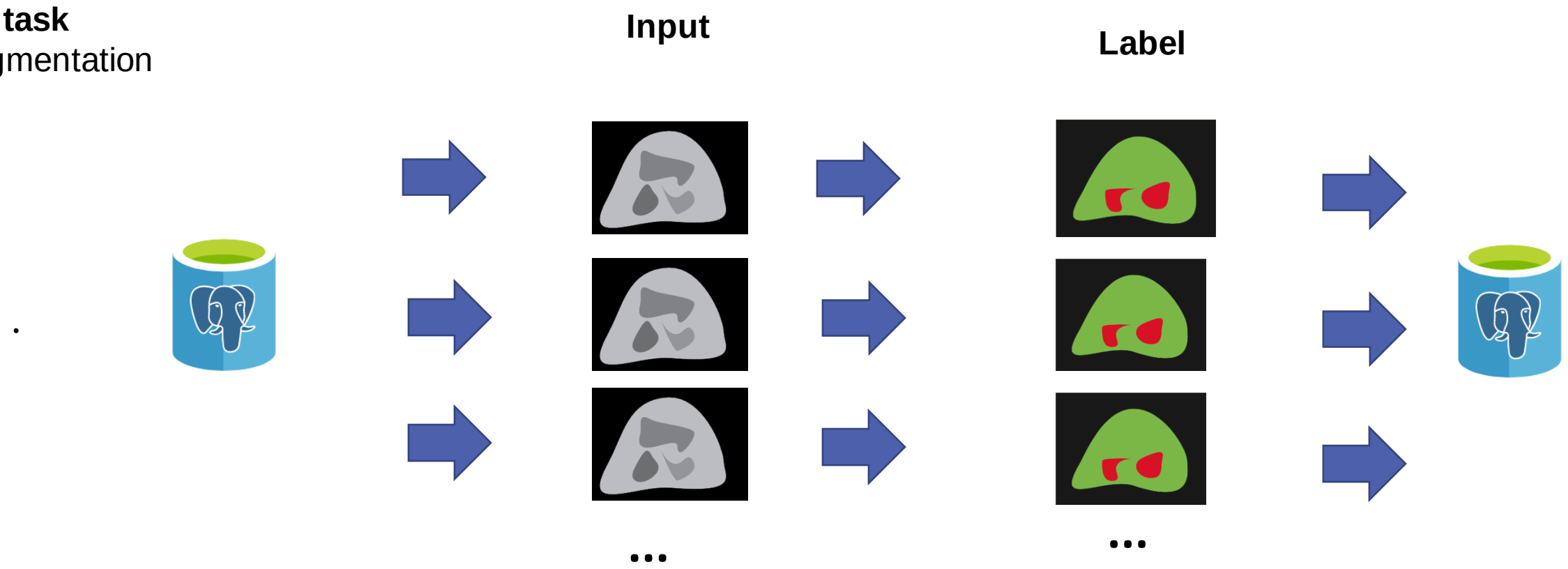
...

...

Data Curation - Labels

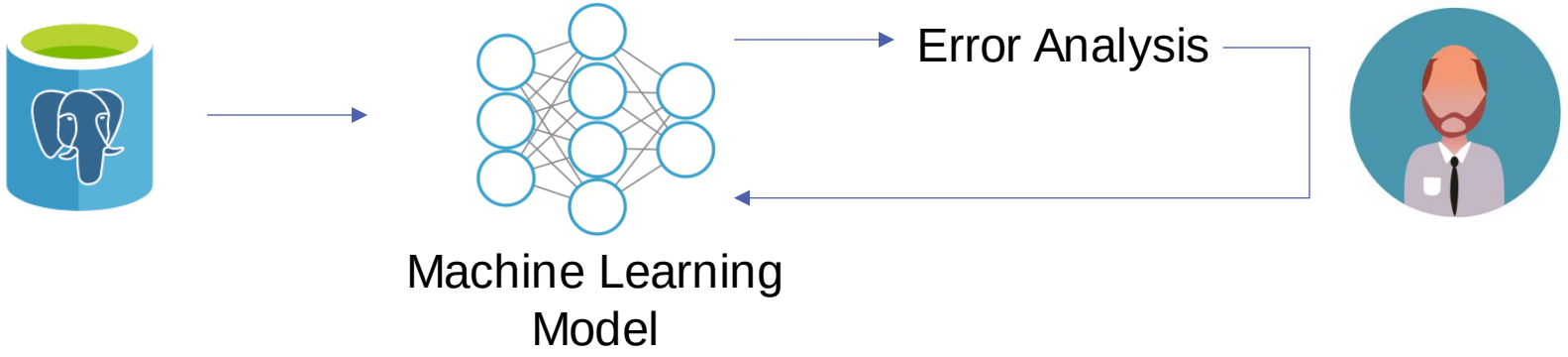
To address the problem with a machine learning framework we need to have our data **labeled** accordingly

ML task
Segmentation



Data Extraction

Data scientist queries SQL database, extracts the data and trains the model



Model centric approach – Data are fixed, model changes

Data Extraction

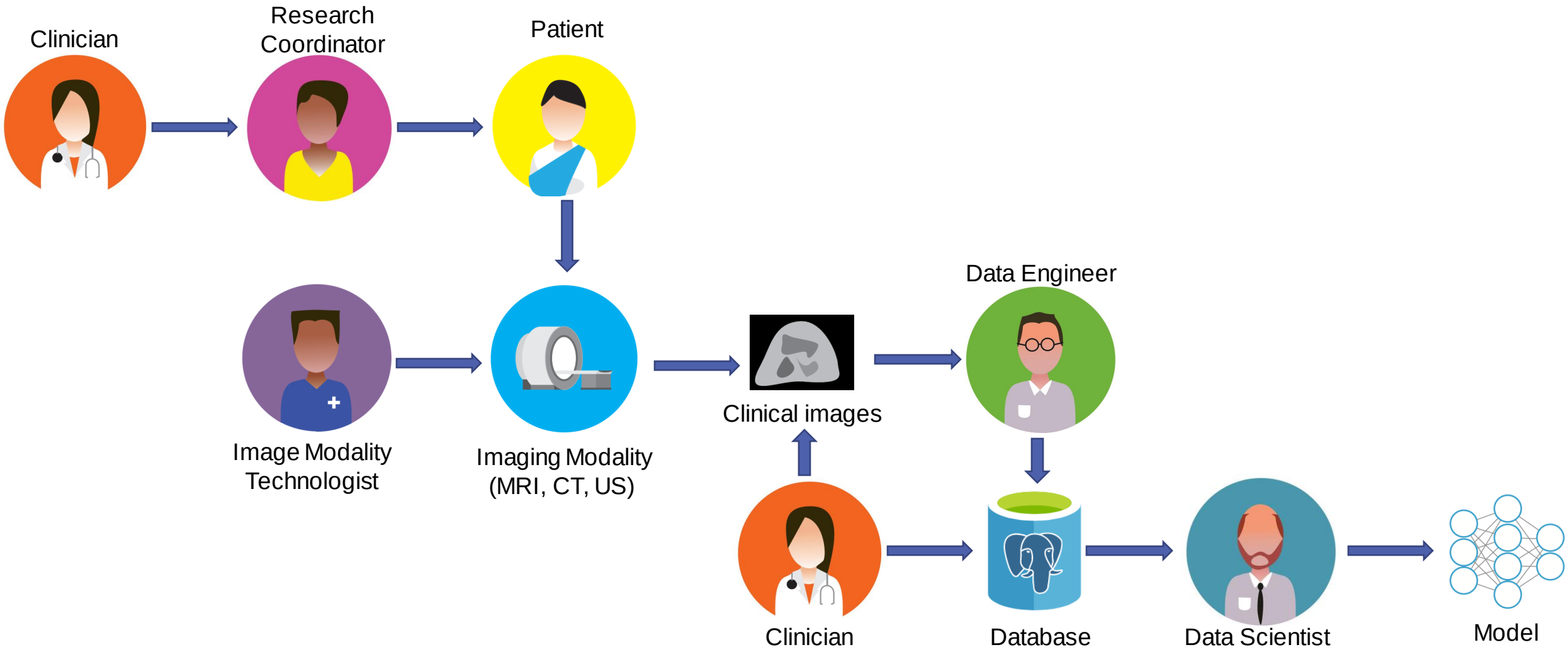
Data scientist queries SQL database, extracts the data and trains the model

Machine Learning Model

Error Analysis

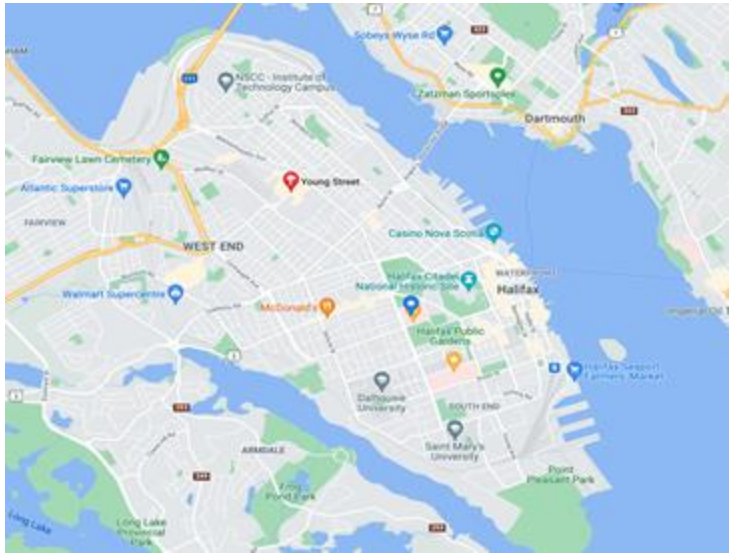
Updates to the dataset

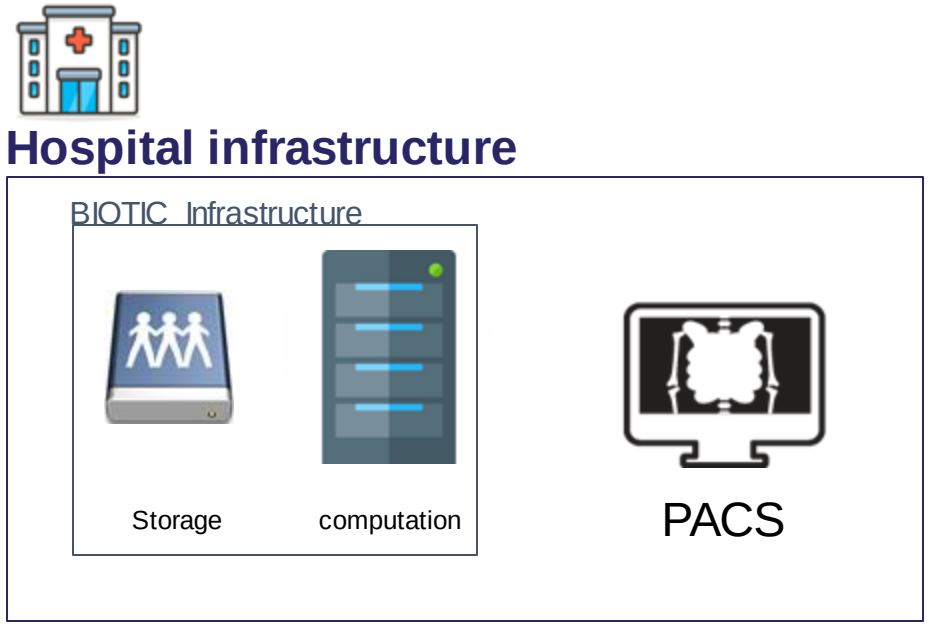
The journey of the data



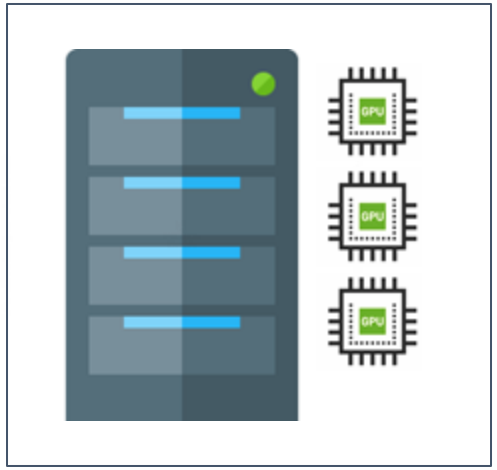
BIOTIC AI Platform

- BIOTIC computational infrastructure sits behind hospital firewall
- No data with PHI going outside firewall
- 3 Servers available
  - 4 GPUs
  - 1 server for model development
  - 1 server for putting model into production with one high performance GPU





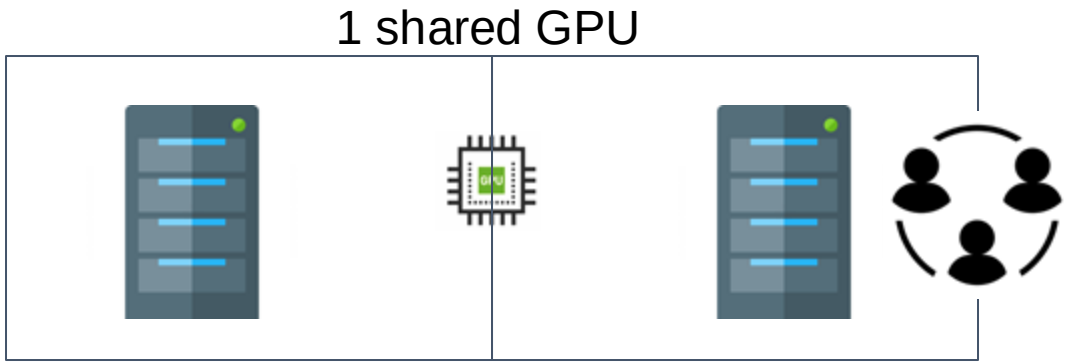
BIOTIC AI Platform



biotic.nshealth.ca

24 CPUs
200 Gb RAM

3 Nvidia Tesla P100 (32 Gb) GPUs



biotic1.nshealth.ca

16 CPUs
64 Gb RAM

1 Nvidia Tesla P100 (16 Gb) GPUs

biotic2.nshealth.ca

16 CPUs
64 Gb RAM

1 Nvidia Tesla P100 (16 Gb) GPUs

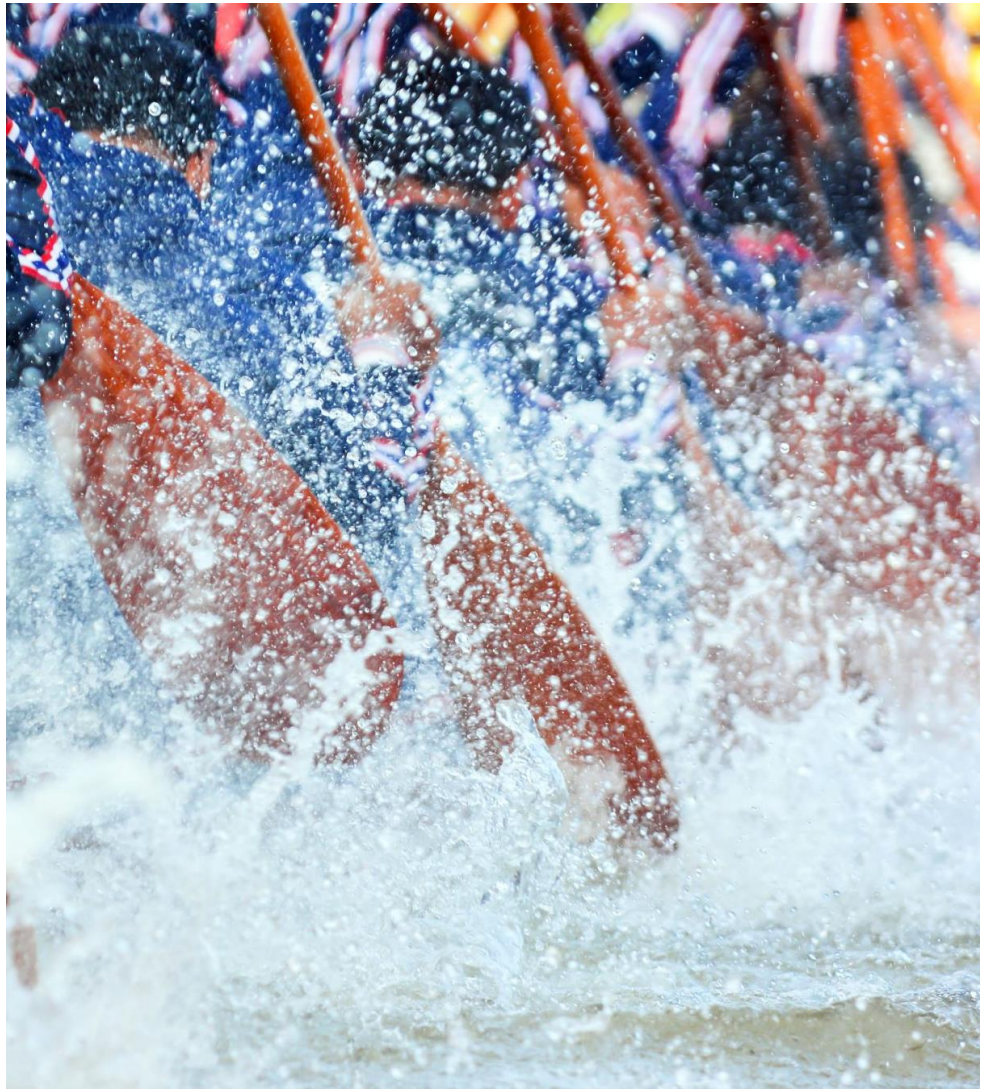
Intended for external collaborations

Moving forward

We are working on setting up a **collaboration platform** to run machine learning projects.

The plan is to have this platform installed in our network that will allow:

- Principal Investigator to select patients from a retrospective updatable, queryable database
- The entire workflow will be streamlined into **one single tool**, from data selection, to processing and curation.
- The platform will support federal learning, and allow cross side research studies without the images leaving the hospital



Conclusions

Successful ML projects need the right tools and people:

- Compute infrastructure
- Programming expertise and knowledge of cutting-edge technology

Our hope is that you can now better appreciate the different components and the number of people required to apply machine learning on biomedical images to investigate a research hypothesis

- A lot of care is required to collect data in the right way without causing loss of efficiency in downstream processing
- PHI confidentiality needs to be maintained at all steps of the workflow

Conclusions

If you have any project ideas, we are happy to chat.

BIOTIC
Imaging for Life

Email:

- Jeff Kowalski: Richard.Kowalski@nshealth.ca
- Alex Guida: Alex1.guida@nshealth.ca

Thank you

Need More Info?

letstalkinformatics@nshealth.ca

Let's Talk Informatics Certifications

- **Digital Health Canada** - participants can claim 1CE hour for each presentation attended.
- **College of Family Physicians of Canada and Nova Scotia Chapter** - participants can earn one Mainpro+ credit by providing proof of content aimed at improving computer skills applied to learning and access to information.
- **Canadian College of Health Information Management** - approves 1 CPE credit per hour for this series for professional members of Canada's Health Information Management Association (CHIMA).