Data Processing Workshop

Presenter: Buz Galbraith

Outline

- My Background
- Technical Prerequisites
- Motivation
- Case Study
- Questions

My Background

My Background

- Did my undergrad and masters here at NYU.
- Worked this past year as a Data Scientist at the Kaiser Permanente Division of Research.
- Starting my PhD in the fall at Northeastern in CS, working on computational biology.

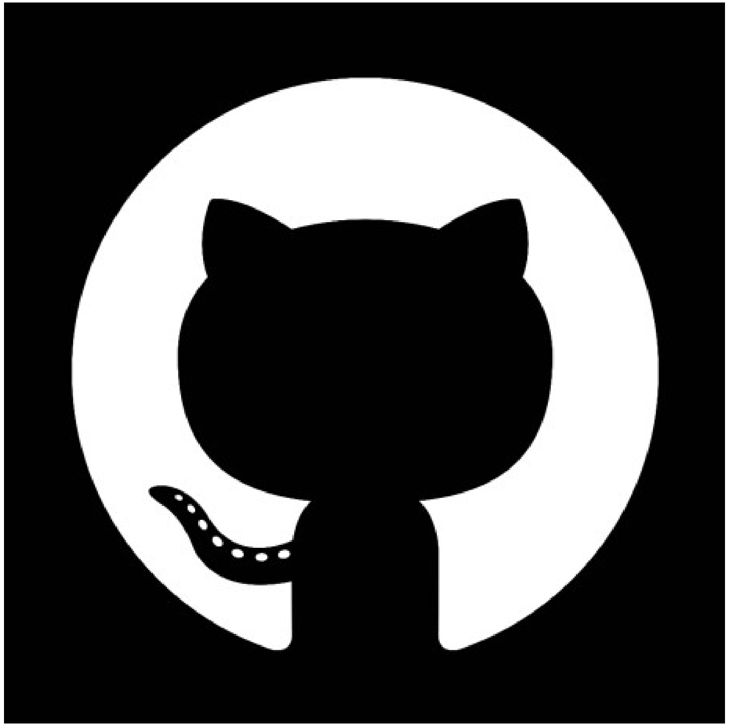
Technical Prerequisites

Technical Prerequisites: A note about virtual environments

- Virtual environments create self contained Python interpreters. This isolation allows the separate installation of python packages without version conflicts between projects.
- There are many tools to do this. Some popular ones are Conda, Hatch, Pipenv and Poetry.
- For this demonstration I will use plain Venv which is provided directly by the Python Foundation.
- I also have the project setup to use poetry as it is the most modern/popular approach.
- I would caution against solely relying on anaconda as they recently left the open source so a lot of companies and labs are moving away from it.

Technical Prerequisites: Git

- Git is a version control system that tracks changes made to files
- Github is a website to remotely store your git repositories, so others can view and collaborate on code.
- All code for this tutorial can be found at this repo: https://github.com/buzgalbraith/data-processing-workshop



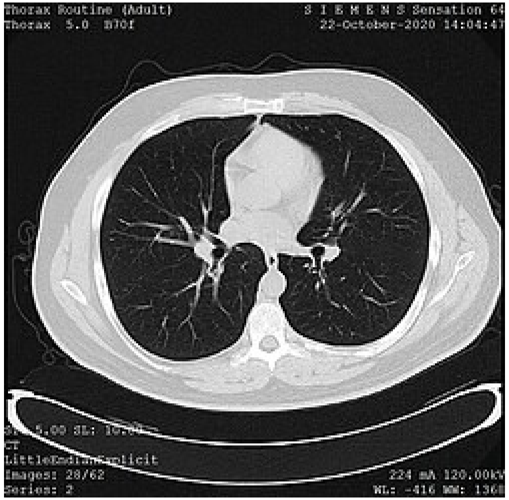
Technical Prerequisites: Polars and Bash

- A lot of this analysis will be done using Bash as well as Polars.
- Bash: the born again shell, is a scripting language and the command line environment for UNIX based systems.
- Polars: is a highly optimized data manipulation library, implemented in Rust and with a Python API highly influenced by Pandas
- Why these tools:
  - They are the fastest for the example we are working on
  - To show we should be flexible in the tooling we work with.

Motivation

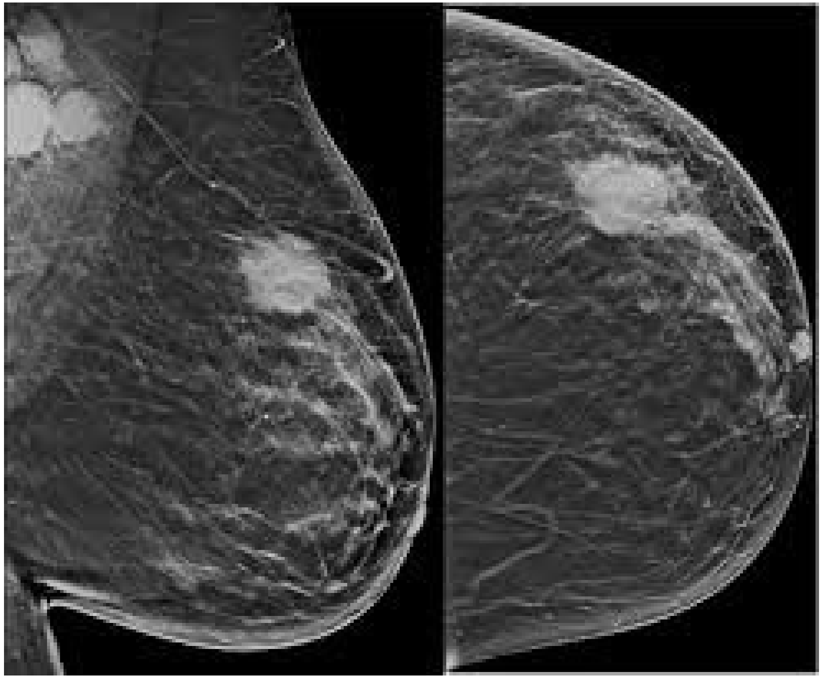
Motivation: Medical Imaging

- Hospitals run thousands of imaging exams of various types every day. Meaning some institutions have collected millions of scans.
- These scans are used by clinicians for a wide range of tasks including:
  - The diagnosis of breast cancer
  - Detecting arrhythmia in the heart
  - Accessing the bone-density of older patients



Motivation: Imaging Research

- As such there is a very active area of research trying to leverage these images to either extract new information or automate existing radiological sans
- However for such projects to take place we need a paradigm for sharing these images that is:
  - Standardized: across health systems, tools and exam types
  - Space efficient: Images are large, we want to send the minimal amount of information from which both meta data about the scan and the resulting image can be extracted
  - Easily de-identifiable: Projects are often collaborations between research teams at multiple health centers as such, we need a way to easily remove PHI (personal health information) from these scans.



Motivation: The DICOM Standard

- The Digital Imaging and Communications in Medicine (DICOM) standard addresses many of these concerns
- DICOM files are composed of two primary data structures:
  - The header: a plain text store of the files meta data (ex exam type) which can be easily used to filter or find files
  - The pixel array: a compressed representation of the pixel data of a given scan from which the original image can be reconstructed

Zooming Out: Why this matters

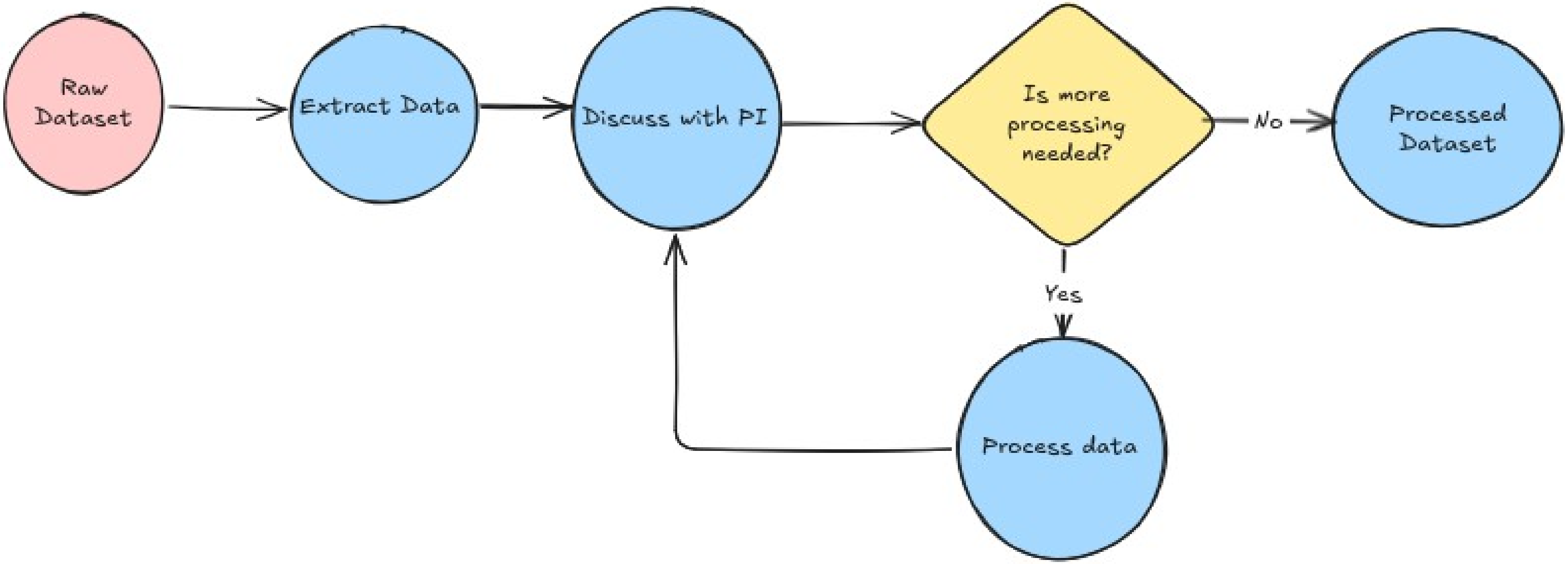
- If you don’t plan to work in imaging or the medical field at all you may wonder how this relates to you.
- I feel going through this example with DICOM files offers a few interesting insights:
  - Illustrates general data processing workflows
  - Shows how domain knowledge can influence how to process data
  - Demonstrate the collaborative nature of multi-disciplinary projects
  - Show that not all data comes in a nice tabular format, and how we can transfer our existing knowledge to new modalities.
  - This is a striped-down version of a real processing workflow I have used at work.

Case Study

Case Study: Setup

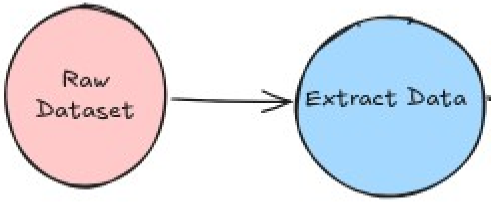
- Suppose you are working on a medical research team with a principal investigator from a pure bio background
- As the first step in a imaging project, relating body fat to cancer outcomes, she has asked you to construct a dataset from an existing set of DICOM files.

Proposed Workflow



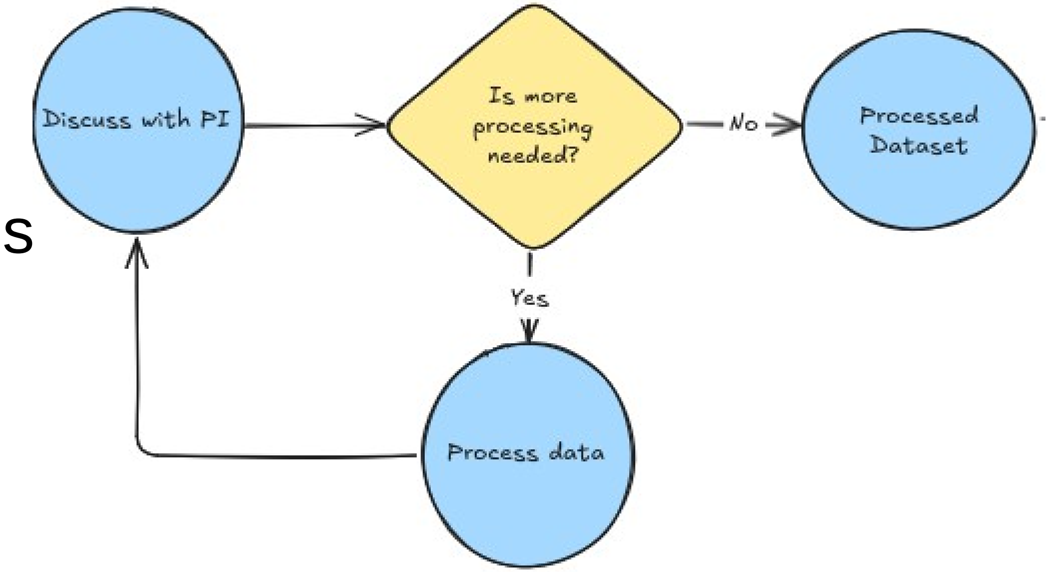
Case Study: Data Extraction

- The first step in this process is finding the relevant files and extracting relevant data from them.
- Once this is done, this is done you can get some simple summary statistics for the data.



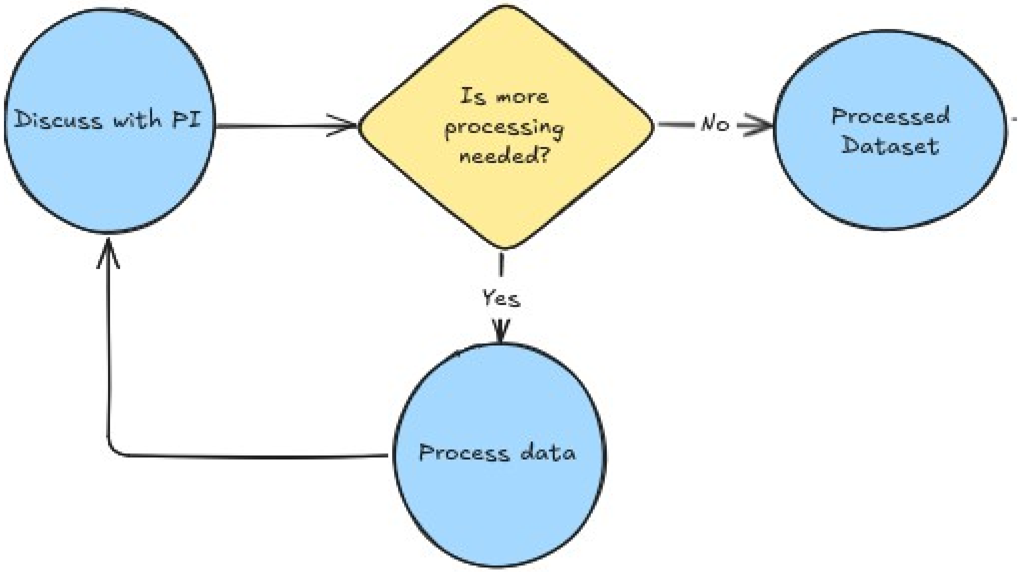
Case Study: Processing Iteration One

- Lets say you brought your findings to the PI.
- In the discussion you identified the following processioning steps to be taken:
  - Adding an absolute path to the file.
  - Removing scans relating to the patient Citizen^Jan



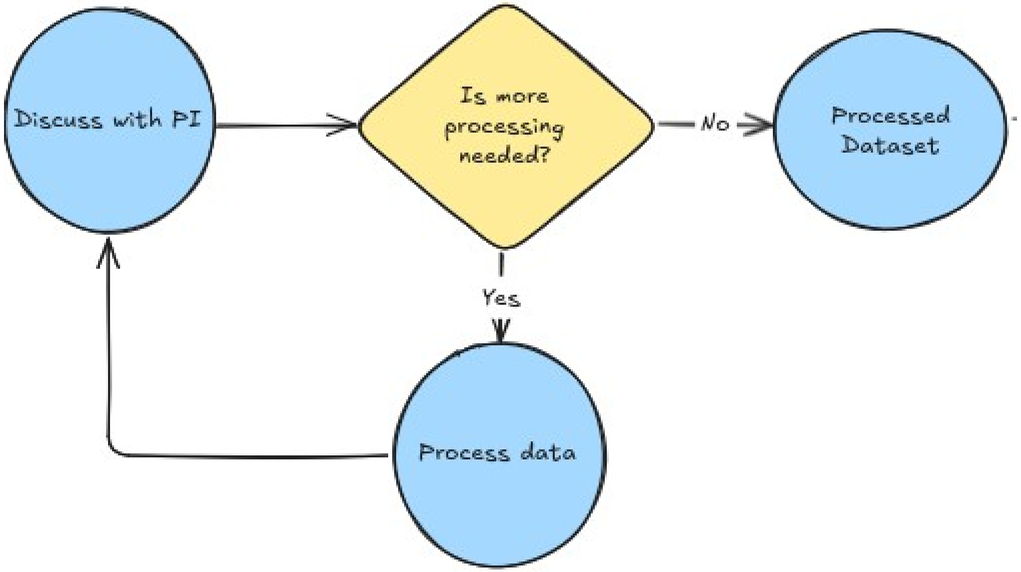
Case Study: Processing Iteration Two

- Lets say you do those processing steps and have another meeting with the PI.
- In this meeting you identify the following further processing steps to take:
  - Filter the data the for only MRI and CAT scans.



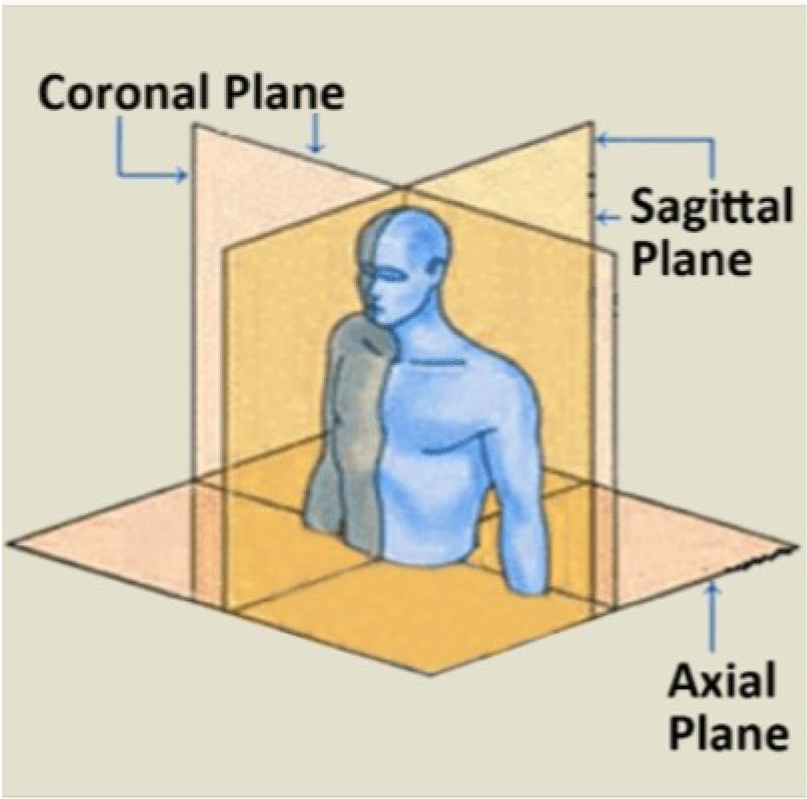
Case Study: Processing Iteration Three

- Lets say you do that and have another meeting with the PI.
- In this meeting you identify the following further processing steps to take:
  - Filter the data for only axial scans.



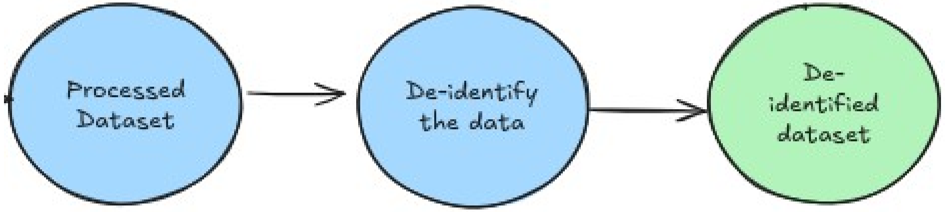
Case Study: Axial Scans

- You identify that the image orientation patient field
- Reading the DICOM documentation you find:
  - Image Orientation (Patient) (0020,0037) specifies the direction cosines of the first row and the first column with respect to the patient. These Attributes shall be provide as a pair. Row value for the x, y, and z axes respectively followed by the Column value for the x, y, and z axes respectively.
- Your know an axial scan moves straight up the body.
- In other words you know:
  - the rows of an axial scan should fully contain information on the lateral (ie left right)
  - The columns should only contain information on the depth (ie front or back) axis
- Thus you can deduce the proper value for an axial scan is [1.0, 0.0, 0.0, 0.0, 1.0, 0.0]

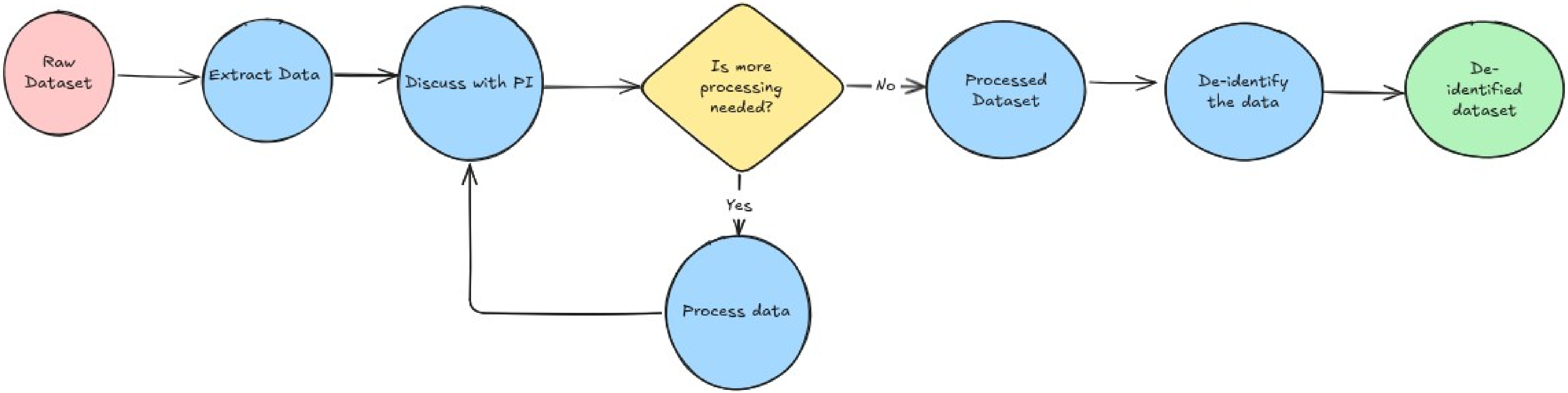


Case Study: De-identification

- Suppose your PI is happy with the processing work we have done with the dataset, but wants to work with a collaborator at another hospital to fit deep learning models to the data.
- As such you need to:
  - De-identify the data, ie create copies of the data with all identifying information removed.
  - Save a key of how to re-identify the data, if you ever need to.
  - Extract images (pngs) from the DICOM files to use in the models



Case Study: Full Workflow



Questions?