# De-identified Digital Pathology Training Session

Gabriele Campanella

Brandon Veremis

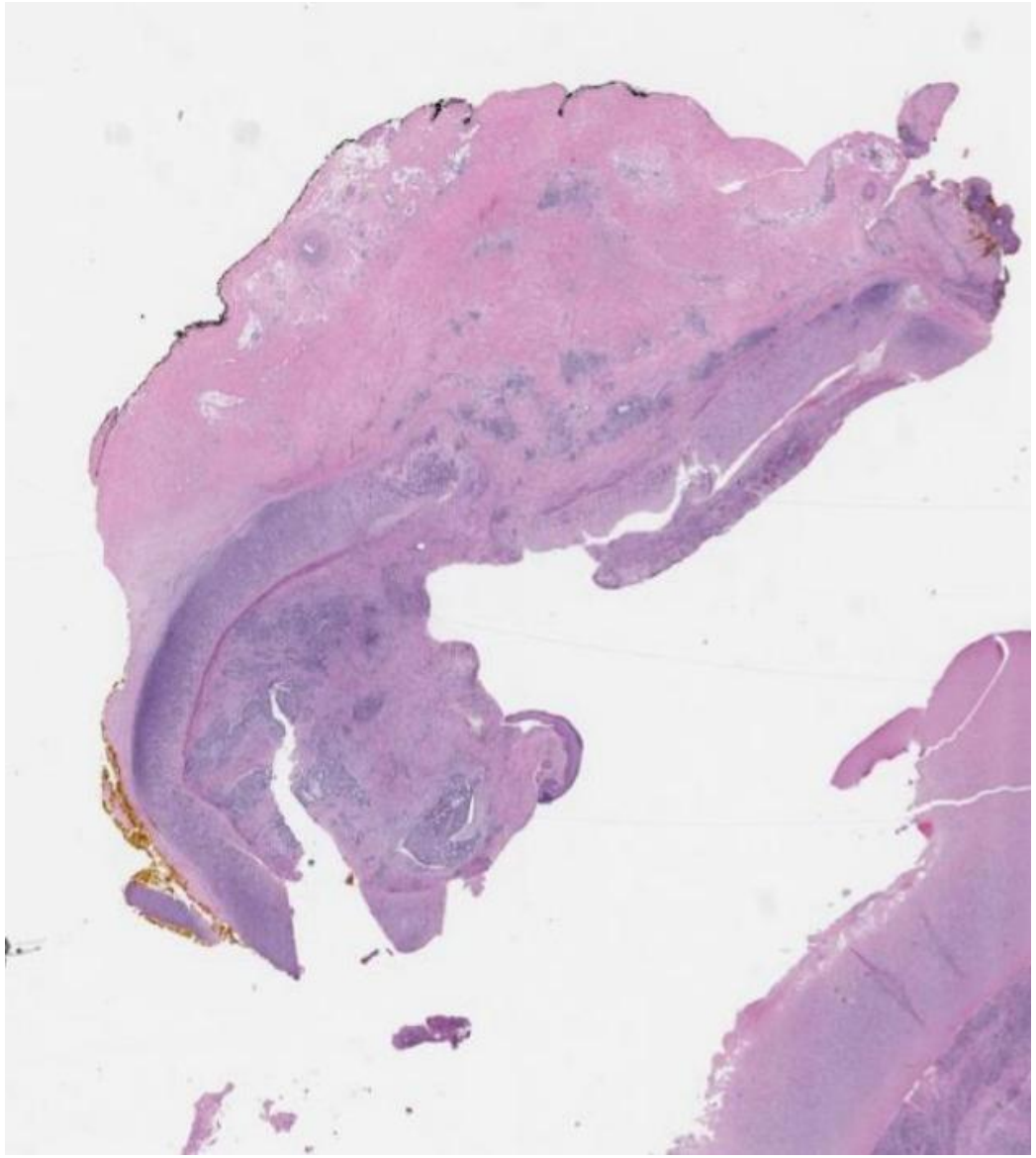
Andrew Deonarine

**November 12, 2026**

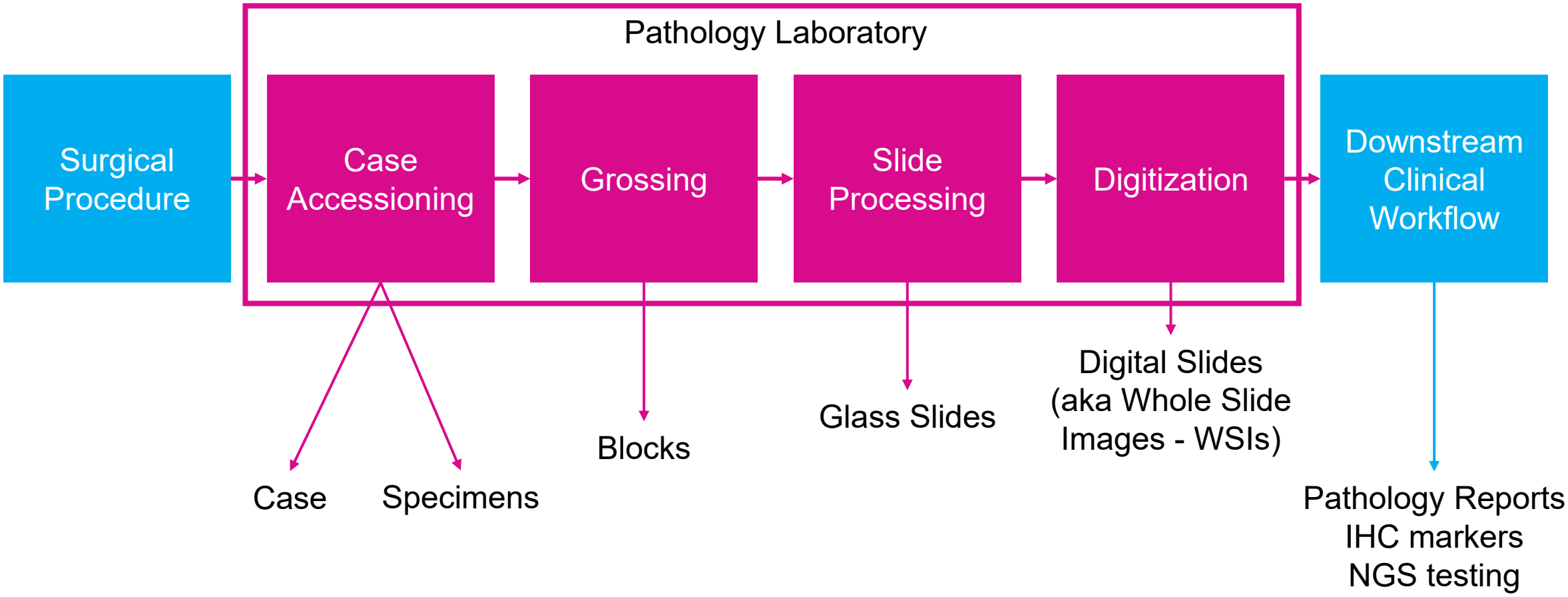
Icahn School of Medicine at Mount Sinai

# Agenda

1) Basic Pathology Asset Structure
2) Overview of Pathology Image Data
3) How To Interact With Images
4) Computational Pathology
5) Case Study
6) Identify Case/Slide/Cohorts
7) AI Ready Mount Sinai (AIR·MS) & Sample Code
8) Summary



# Pathology Lab Workflow and Data



# Overview of Pathology Data

The pathology department produces large amounts of physical and digital data.

**Data Modalities**

- Biological Assets → Stored for a minimum of 20 years per state regulations
  - Tissue blocks
  - Glass Slides
- Digital Assets
  - Pathology Laboratory Information System (LIS) Metadata → Links cases, specimens, blocks, slides to reports. Tracks most aspects of pathology lab
  - Whole Slide Images → Pyramidal .tiff files scanned with Philips WSI scanners at 40x resolution. Available as part of Data Ark
  - Pathology Reports → Free text diagnosis written by the pathologist for the entire case. Can include test results (IHC, FISH, etc.) as free text.

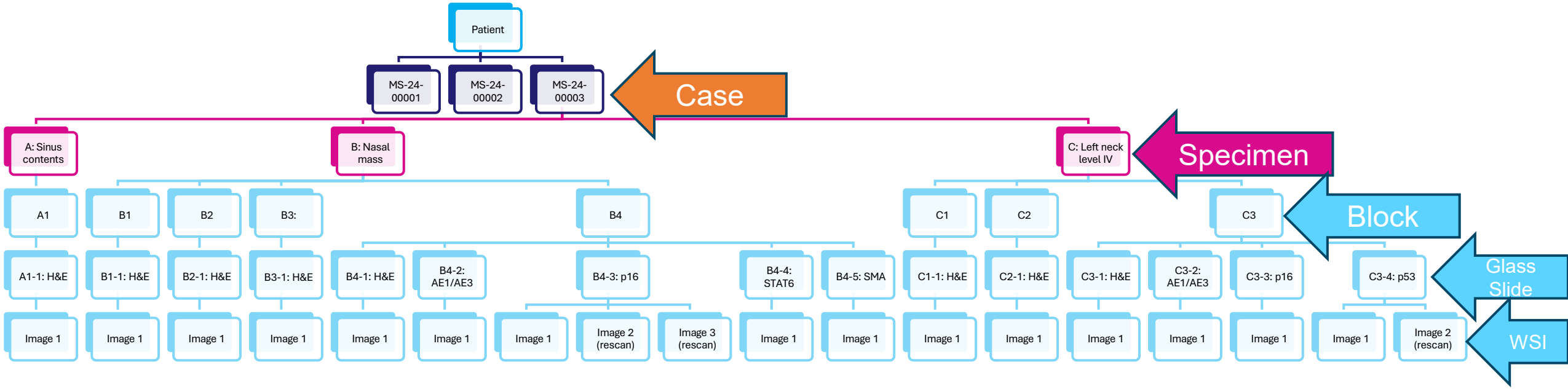
# Basic pathology asset structure

*One to many to many to many to many to many*

- A patient can have many cases
- A case can have many specimens
- A specimen can have many blocks
- A block can have many slides (which can be different stains/orders)
- A slide can have many images (i.e., rescans)
- No designated "best image" for a patient



# Pathology Data Hierarchy



# Digital Pathology Statistics

## Pathology Service (after 2010)

- Slides : 17,712,871
- Cases: 2,632,983
- Patients: 1,477,343

## Digital Pathology (prospective since 2020, select retrospective)

- Slides: 4,961,158
- Cases: 779,495
- Patients: 530,164

## Data Ark

- Slides: 3,133,524

# Overview of Pathology Image Data



Large (easily > 2 GB each)



40x resolution (~0.25 microns/pixel)



Contains all tissue detected by the scanner (usually reliable)



Most images rescanned if QC issues present (e.g., out of focus)

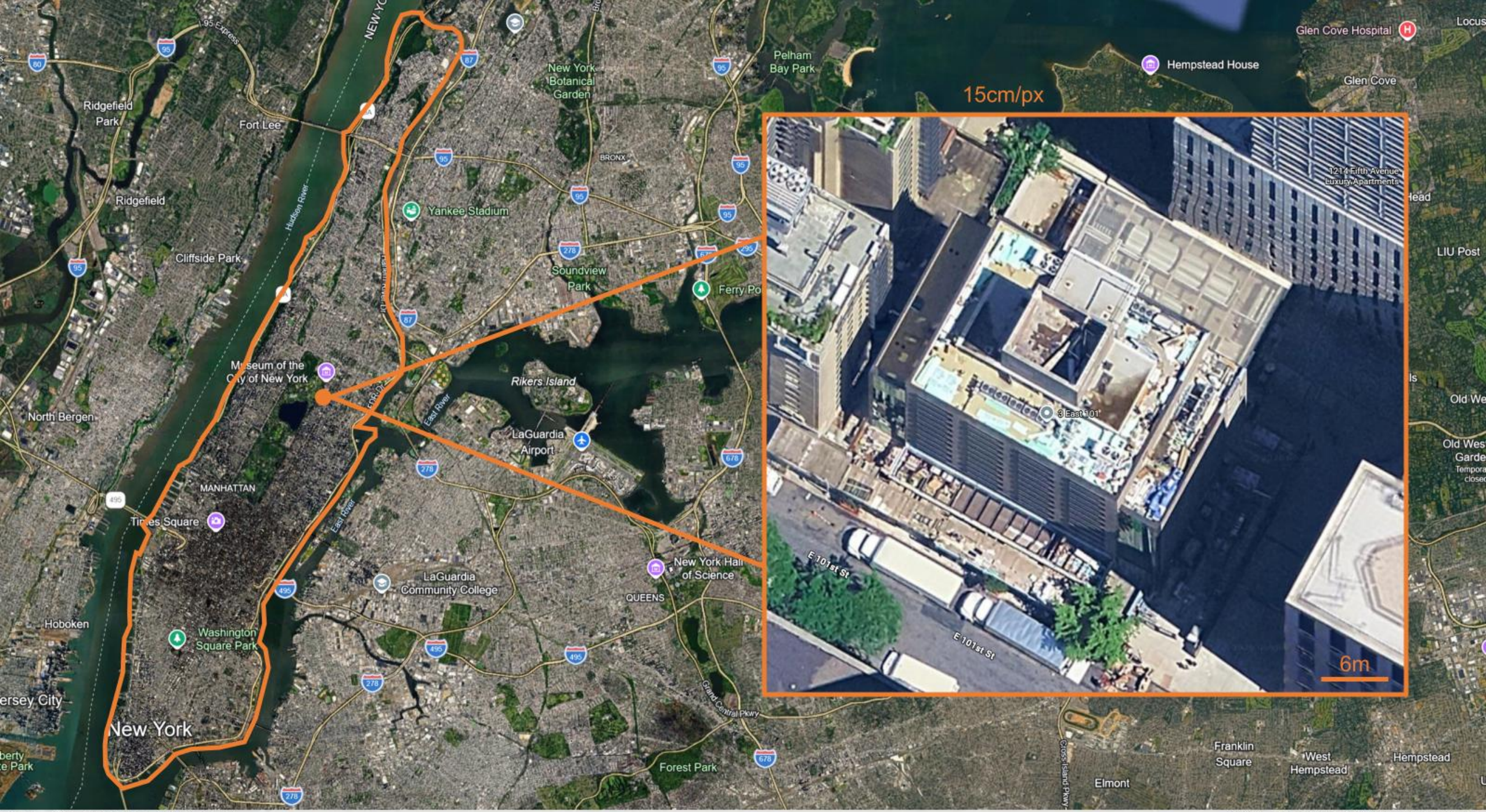


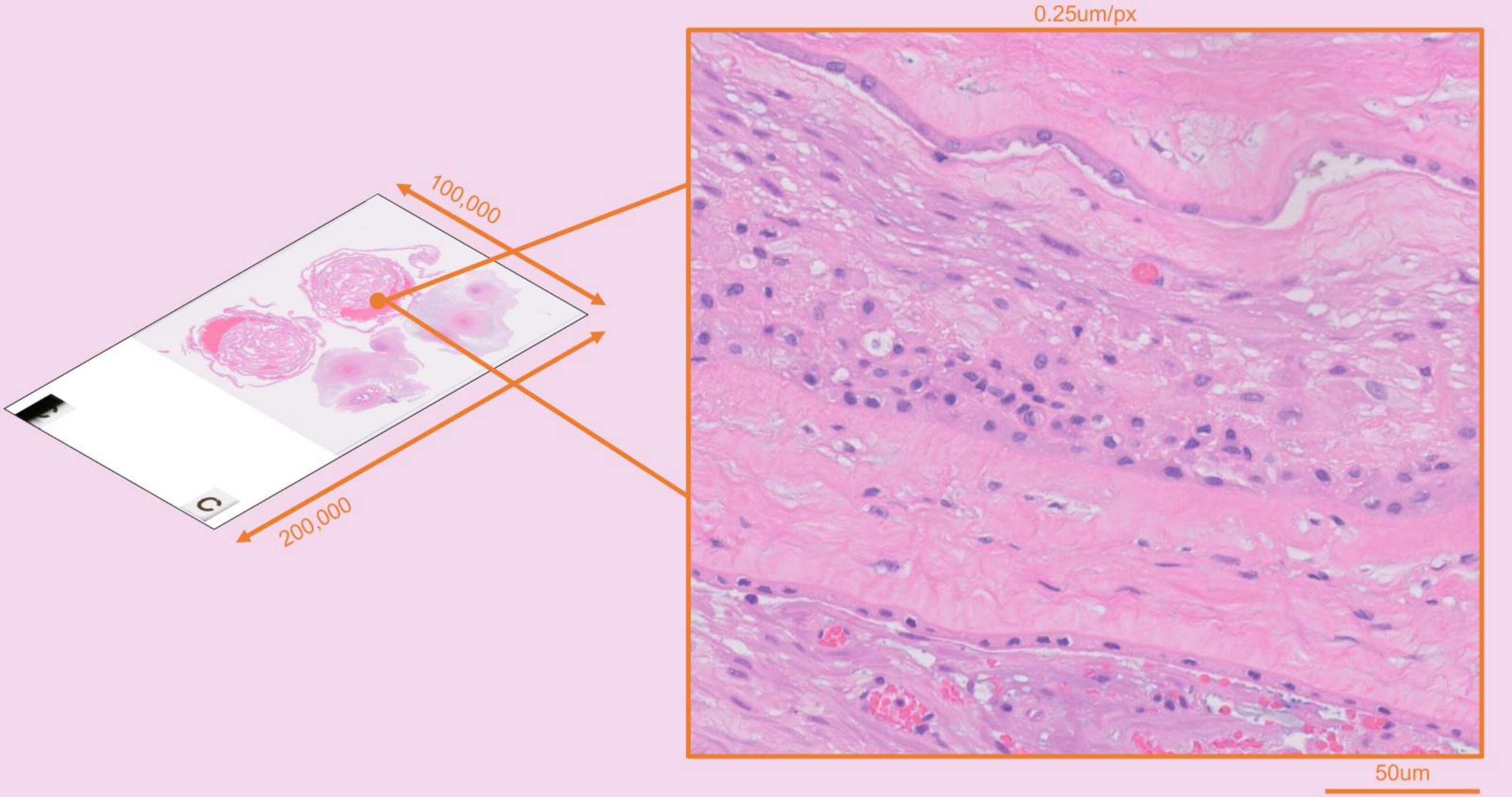
TIFF format

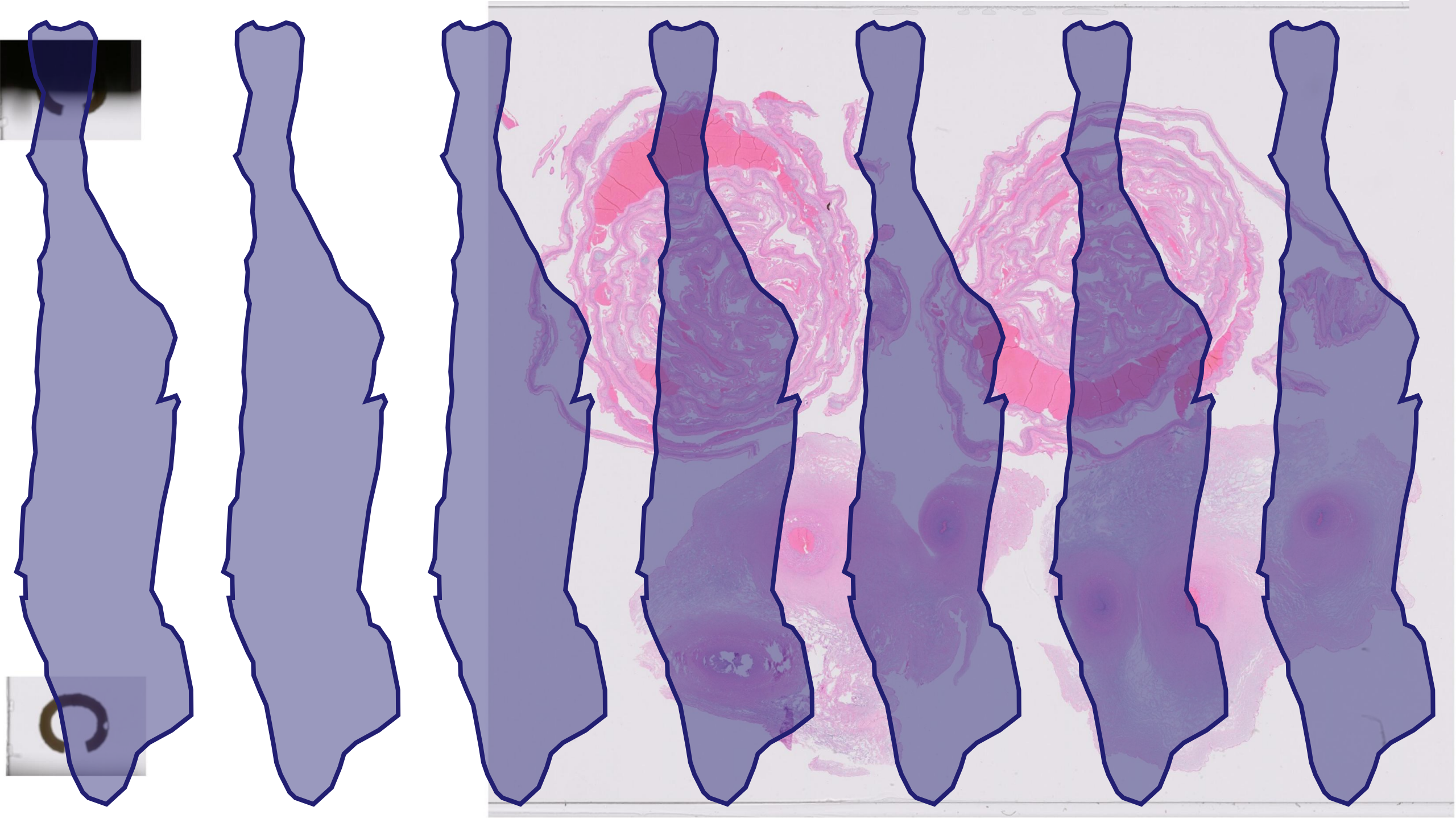


Open-source tools available

186 000 pixel
B 07.40495
1
he
>x300k

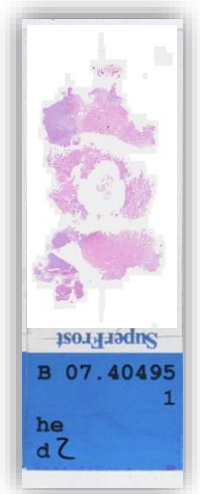

15cm/px
1214 Fifth Avenue Luxury Apartments
3 East 101
E 101st St
E 101st St
6m
Glen Cove Hospital
Hempstead House
Glen Cove
Pelham Bay Park
New York Botanical Garden
BRONX
Yankee Stadium
Soundview Park
Ridgefield Park
Fort Lee
Ridgefield
Cliffside Park
Hudson River
Museum of the City of New York
Rikers Island
East River
LaGuardia Airport
North Bergen
MANHATTAN
Times Square
LaGuardia Community College
New York Hall of Science
QUEENS
Hoboken
Washington Square Park
New York
Forest Park
Grand Central Pkwy
Cross Island Pkwy
Franklin Square
West Hempstead
Hempstead
Elmont
LIU Post
Old Westbury Garden
Google
100%
Data attribution
8/17/2020–5/11/2024
3,000 m
Camera: 40 km
40°47'29.37"N 73°53'01.99"W
6 m


0.25um/px
100,000
200,000
50um

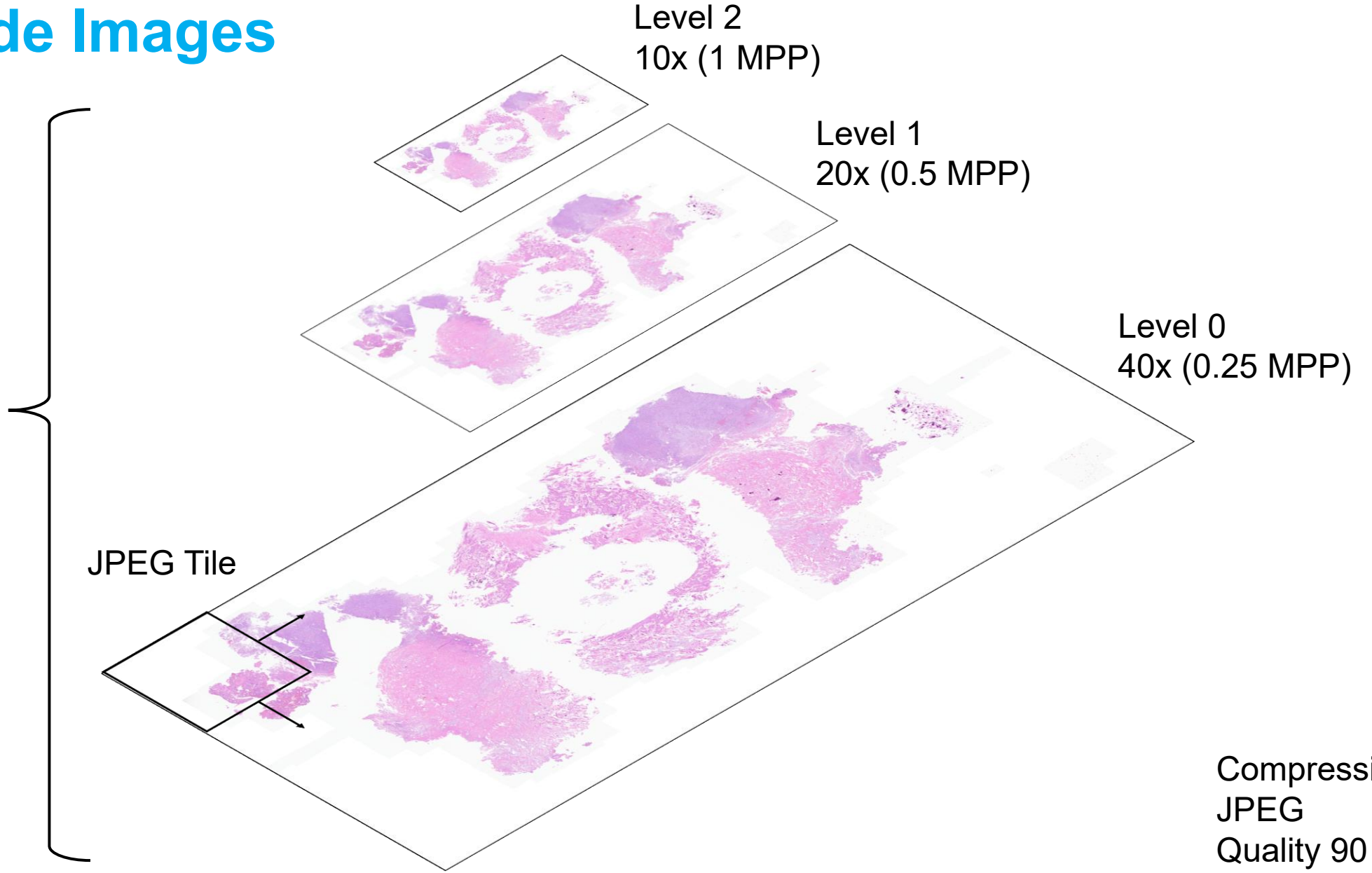


MPP: microns per pixel

# Whole Slide Images




Pyramidal .tiff




Compression:
JPEG
Quality 90

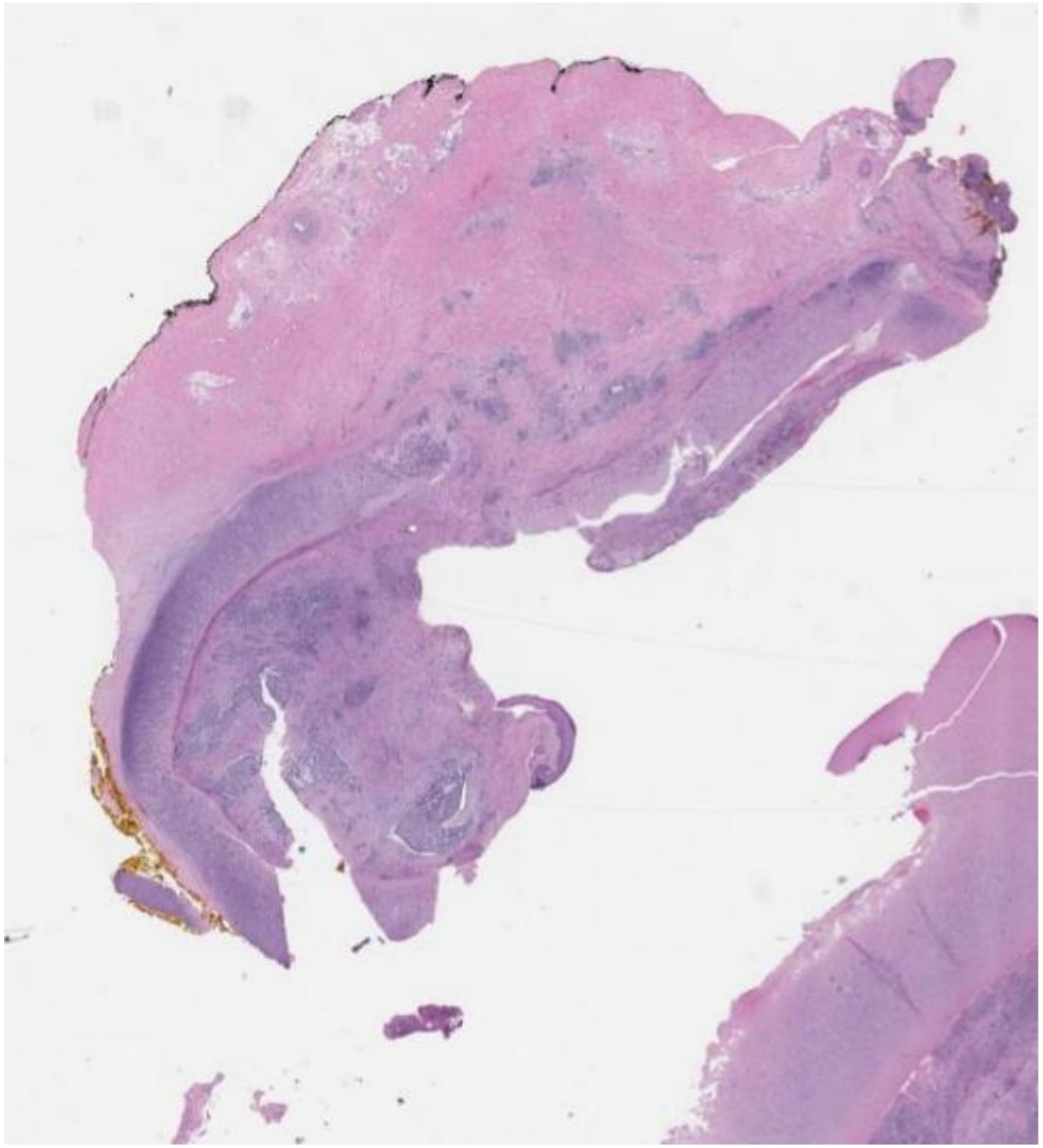
# How To Interact With Images

- Files can’t be loaded into memory all at once
- Image has to be tiled or downsampled for analysis
- Need specialized packages to interact with the files (python)
  - `openslide`
  - `cucim`
- Needs specialized viewers to view the files
  - QuPath (local)
  - Digital Slide Archive (browser app with backend)

# How To Interact With Images

```
import openslide
slide = openslide.OpenSlide('path/to/slide.tiff')
base_mpp = float(slide.properties[openslide.PROPERTY_NAME_MPP_X])
thumb = slide.get_thumbnail((1000,1000))
downsamples = slide.level_downsamples
tile40x = slide.read_region((14356, 24674), 0, (256,256)).convert('RGB')
tile20x = slide.read_region((14356, 24674), 1, (256,256)).convert('RGB')
```

https://openslide.org/api/python/



**Quality Control**

- Detect blurred regions
- Detect tissue folds
- Stain normalization

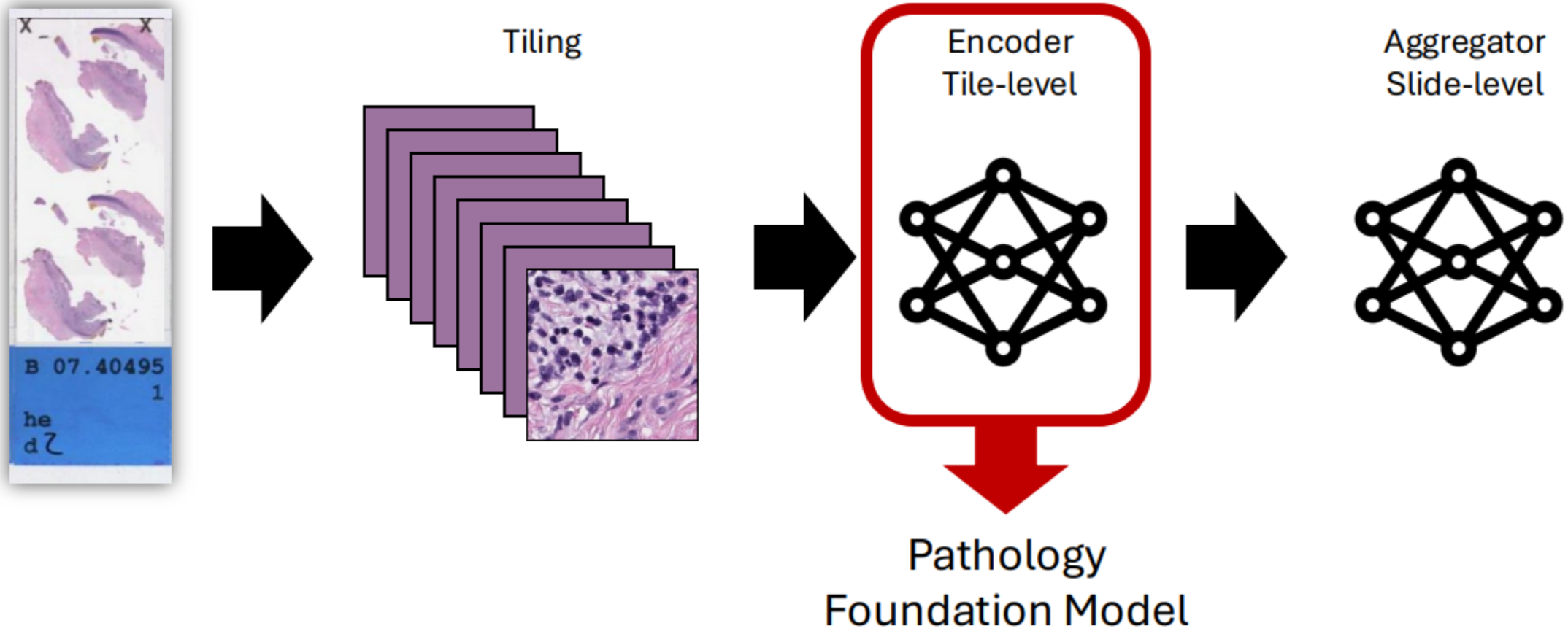
**Diagnostic Reporting**

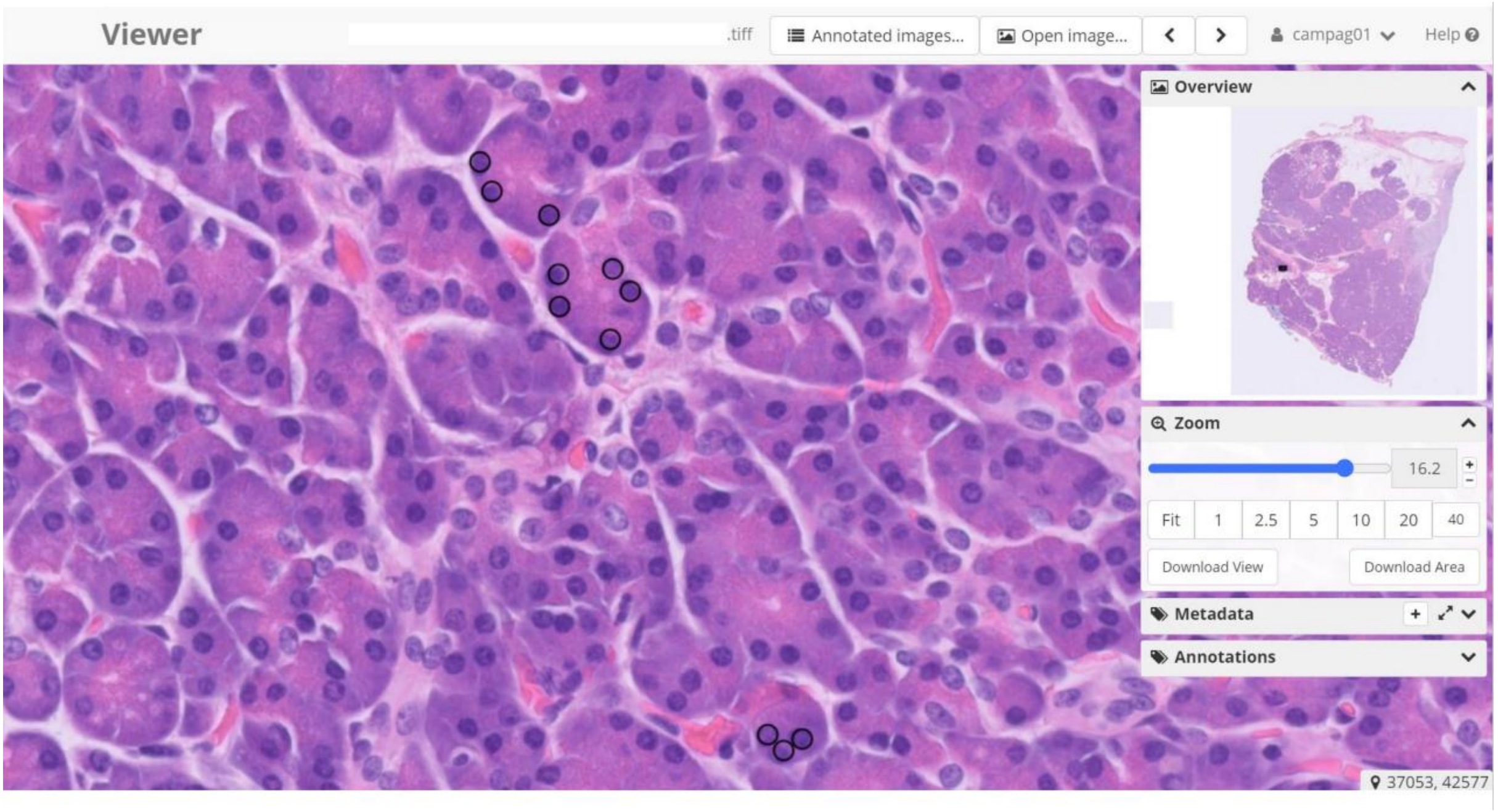
- Tumor detection
- Tumor segmentation
- Tumor staging
- Cell detection
- Slide retrieval

**Prognostication**

- Biomarker prediction
- Response prediction
- Survival analysis

# Computational Pathology




Viewer
.tiff
Annotated images...
Open image...
campag01
Help
Overview
Zoom
16.2
Fit
1
2.5
5
10
20
40
Download View
Download Area
Metadata
Annotations
37053, 42577


Viewer
.tiff
Annotated images...
Open image...
campag01
Help
Overview
Zoom
0.2
Fit
1
2.5
5
10
20
40
Download View
Download Area
Metadata
Annotations
89341, -3065

# Case Study

Goal: Obtain a diagnostic biopsy slide cohort of breast cancer patients with invasive ductal carcinoma and their receptor status.

Tasks:

- Identify the right cases
- Identify representative slide
- *Obtain diagnosis, filter by IDC*
- *Obtain hormone receptor status (ER/PR/HER2)*

# Identify Case

- Manual review
- Rely on pathology Sign-out knowledge
  - Search breast cases where report contains “biopsy” & “ductal carcinoma”
- Cons
  - Will capture DCIS
  - May lose cases where biopsy is not explicitly mentioned
  - Typos
  - Free text fallacy

# Identify Case

- Rely on non-pathology data sources
  - EPIC, cancer registry, etc.
  - Match with collection dates of specimens
- Cons
  - Will not capture biopsies done in outpatient clinic or consults
  - Potential inconsistent MRNs

# Identify Slide

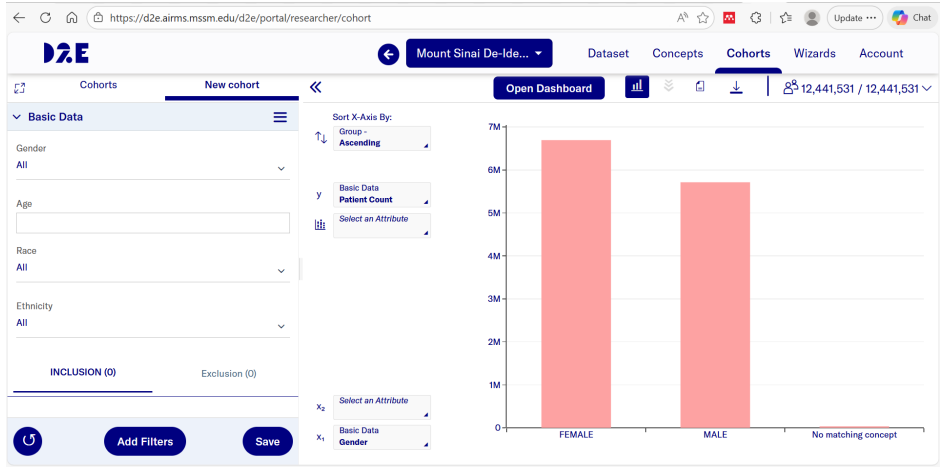
- Preface: pathology report gives specimen level information, not slide level
  - Manual slide review
  - Rely on specialty specific knowledge
    - For breast cancer, only tumor slides receive ER/PR/HER2 testing
- Rely on synoptic reporting
  - Newer synoptic reports include “best” tumor block
  - Cons
    - Implemented only recently
    - No synoptic for biopsies
- AI – vision model

# TLDR

Pathologist collaboration is essential!

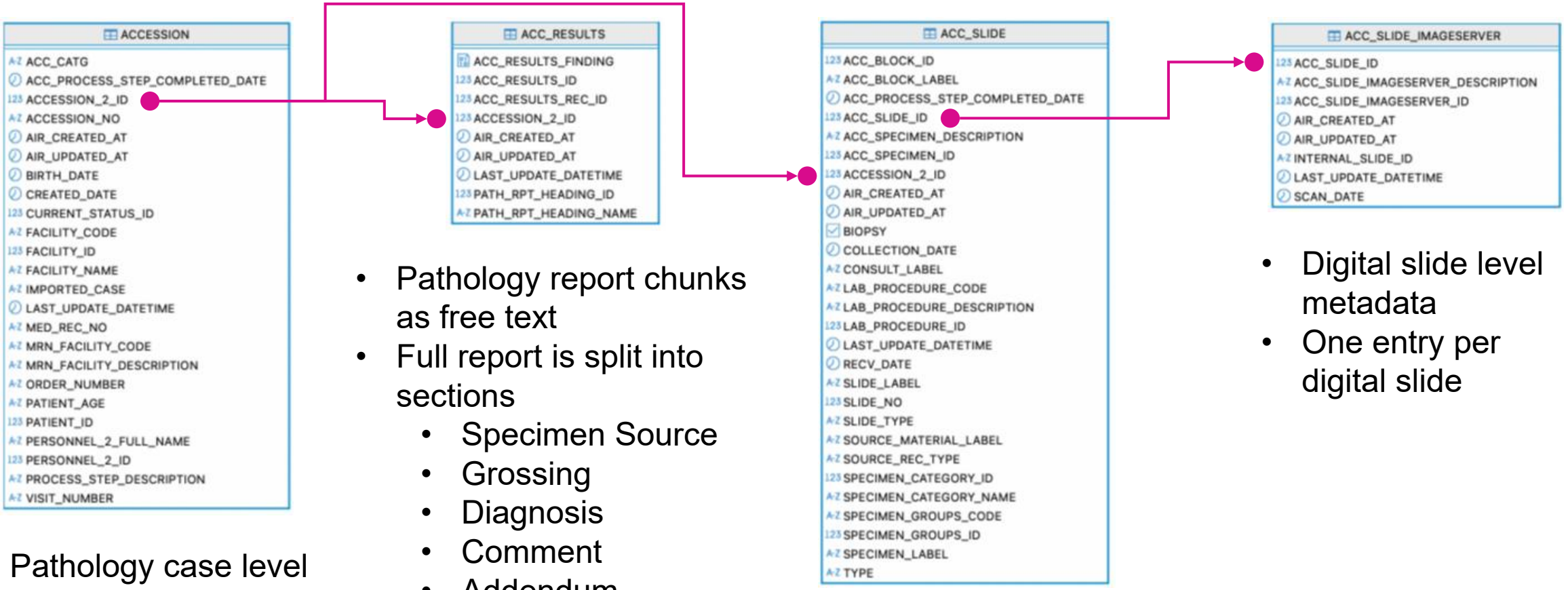
# Using Cohort Query Tools to Find Digital Pathology Images on Data Ark

- Previously this was accessible through Leaf (to the file in Data Ark))
- We are transitioning to Data2Evidence (our new cohort query tool)




https://d2e.airms.mssm.edu/

# Pathology Data in AIR·MS




- Pathology case level metadata
- One entry per case

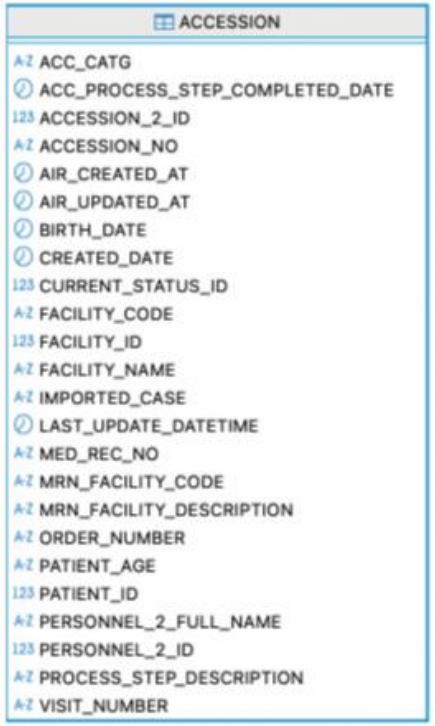
- Pathology report chunks as free text
- Full report is split into sections
  - Specimen Source
  - Grossing
  - Diagnosis
  - Comment
  - Addendum
- One entry per section

- Glass slide level metadata
- One entry per glass slide

- Digital slide level metadata
- One entry per digital slide

AIRMS Schema: CDMPATHOLOGY

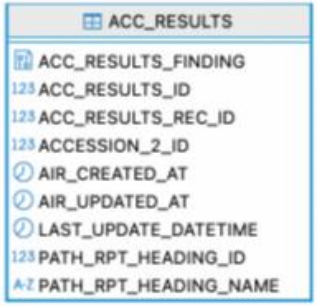
# Case Metadata




Frequently used fields:

- *ACCESSION_2_ID* Linking key to other tables.
- *ACCESSION_NO* Accession number, used to identify cases. Pathologist-friendly.
- *MED_REC_NO* EPIC MRN. Can be used to link to other data modalities.

# Report Text




Frequently used fields:

- *ACCESSION_2_ID* Linking key to case table.
- *PATH_RPT_HEADING_NAME* Name of the section. Section naming is not standardized. Some examples:
  - DIAGNOSIS
  - FINAL DIAGNOSIS
  - SPECIMEN SOURCE
  - NOTE
  - ADDENDUM 1, 2, 3, …
  - COMMENT
  - GROSSING
- *ACC_RESULT_FINDING* Actual report text.

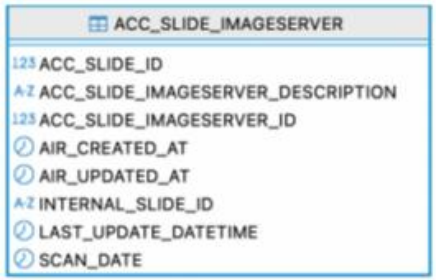
# Glass Slide Metadata




Frequently used fields:

- *ACCESSION_2_ID* Linking key to case table.
- *ACC_SLIDE_ID* Linking key to other tables.
- *SPECIMEN_LABEL* Specimen identifier within the case. Uppercase letter (A, B, …). Matches specimen labels in free text pathology report. Pathologist-friendly.
- *ACC_BLOCK_LABEL* Block identifier within the specimen. Number (1, 2, …). Matches block labels in free text pathology report, usually in the grossing description. Pathologist-friendly.
- *SLIDE_LABEL* Slide identifier within the block. Number (1, 2, …). Not used in free text reports usually. Pathologist-friendly.
- *RECV_DATE* Received date of the specimen. Useful to match cases to other data modalities.
- *LAB_PROCEDURE_DESCRIPTION* Lab order that generated the slide. Usually a stain name. For example: H&E, PDL1, ER, PR, …

# Digital Slide Metadata




Frequently used fields:

- *ACC_SLIDE_ID* Linking key to slide table.
- *INTERNAL_SLIDE_ID* 6-char alphanumeric code that uniquely identifies a glass slide. Not pathologist-friendly. Aka BARCODE. If a barcode exists, the slide was scanned.
- *ACC_SLIDE_IMAGESERVER_DESCRIPTION* long alphanumeric code that uniquely identifies a .tiff file. Unique to a scanning event, not to a glass slide. Same barcode can be associated to multiple .tiff files (e.g., rescanning). Not pathologist-friendly.
- *SCAN_DATE* Timestamp of scanning event.

NOTE:

- Scanning started in 2020 at 40% capacity. 100% prospective digitization was reached in 2024. Most slides are still not scanned.
- *INTERNAL_SLIDE_ID* and *ACC_SLIDE_IMAGESERVER_DESCRIPTION* can be used to identify slides in Data Ark. Mapping is not publicly available.

# Pathology Metadata – Practical Tips

- **How to communicate with the pathologist**
  Accession numbers, specimen/block/slide labels are well understood by pathologists.
  For example: slide “A1-1” from block “A1” in case MS-24-00001.
- **Linking slides to a diagnosis is challenging**
  Reports are written for the entire case. Most granular diagnosis is at the specimen level (unstructured). There is no slide level description. It is important to talk to a pathologist to understand which slides from a case/specimen are relevant to your study.
- **Some metadata can be inaccurate**
  Some metadata fields exposed are used internally in the pathology department for workflow management. They are not for diagnostic purposes and can be inaccurate. Examples include: division, organ, specimen description, ICD codes. Always talk to a pathologist to discuss how to best approach your problem.
- **Data Ark slide links are not publicly available**
  To obtain Minerva paths to a set of slides can be done through MSDW (requires IRB). You will need *INTERNAL_SLIDE_ID* and *ACC_SLIDE_IMAGESERVER_DESCRIPTION* for each slide in your cohort.

# Pathologist collaboration is essential!

- Choosing a single representative image is challenging
  - Generally, requires subspecialty expertise by a pathologist
  - Context-dependent: a selected image for one project is not necessarily the best image for another project

- They can also make sure that the findings reported in the study make sense in the context of pathology

- For assistance in finding a pathologist interested in research, email the director of the Biorepository and Pathology CORE, Rachel Brody, MD, PhD (rachel.brody@mountsinai.org)

# AI Ready Mount Sinai (AIR·MS) & Sample Code

# What is AIR·MS?

Artificial Intelligence-Ready Mount Sinai (AIR·MS) is a research platform that is composed of:

1) A (very fast) integrated database including Mount Sinai Data Warehouse (MSDW), Pathology and Radiology metadata; and the included data is growing.

2) A Research Environment that allows interactions with the AIR·MS database from Python or R.

3) An Application Tier to host a growing number of applications – including cohort building tools and annotation apps.



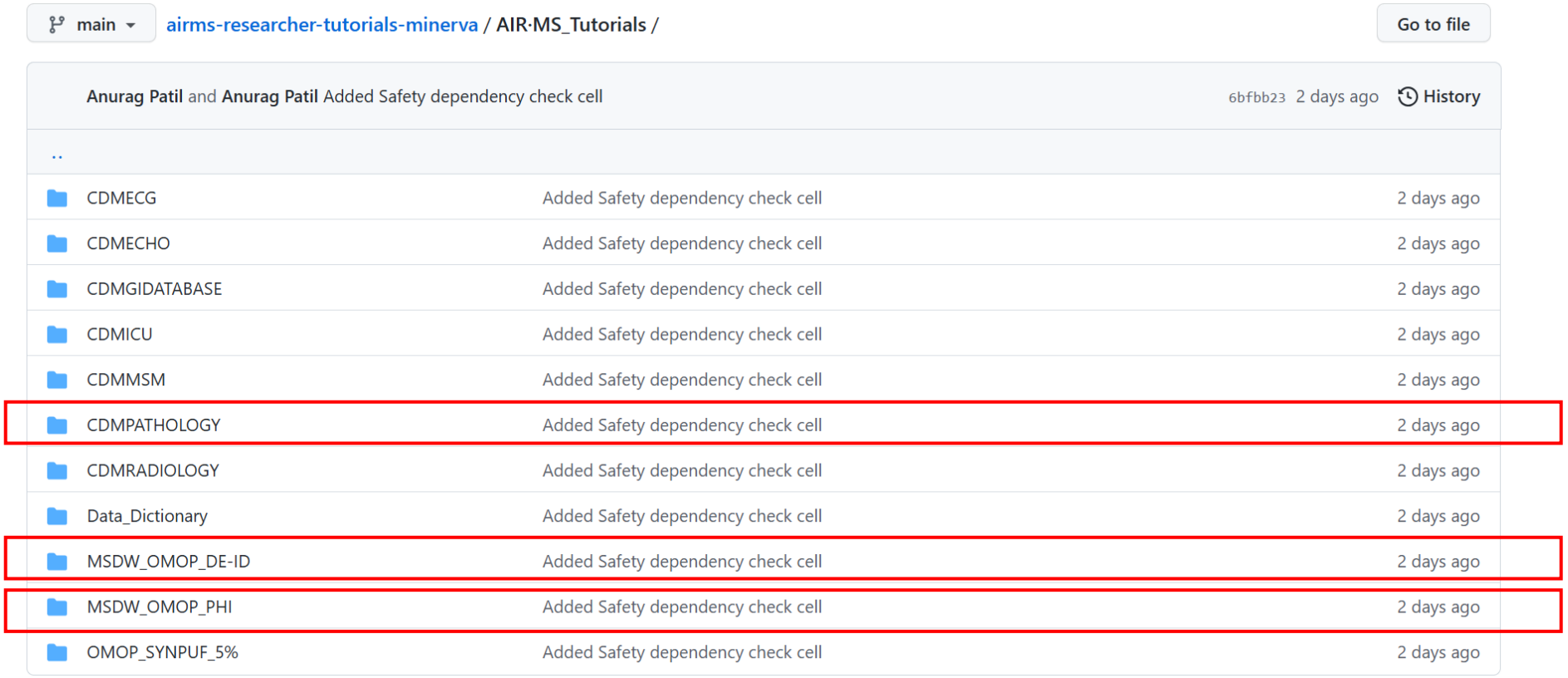
Get Started Here:

https://labs.icahn.mssm.edu/minervalab/ air%E2%80%A7ms-getting-started/

# Sample  Notebooks to Get You Started

Tutorials for Various Schemas (Python, R):

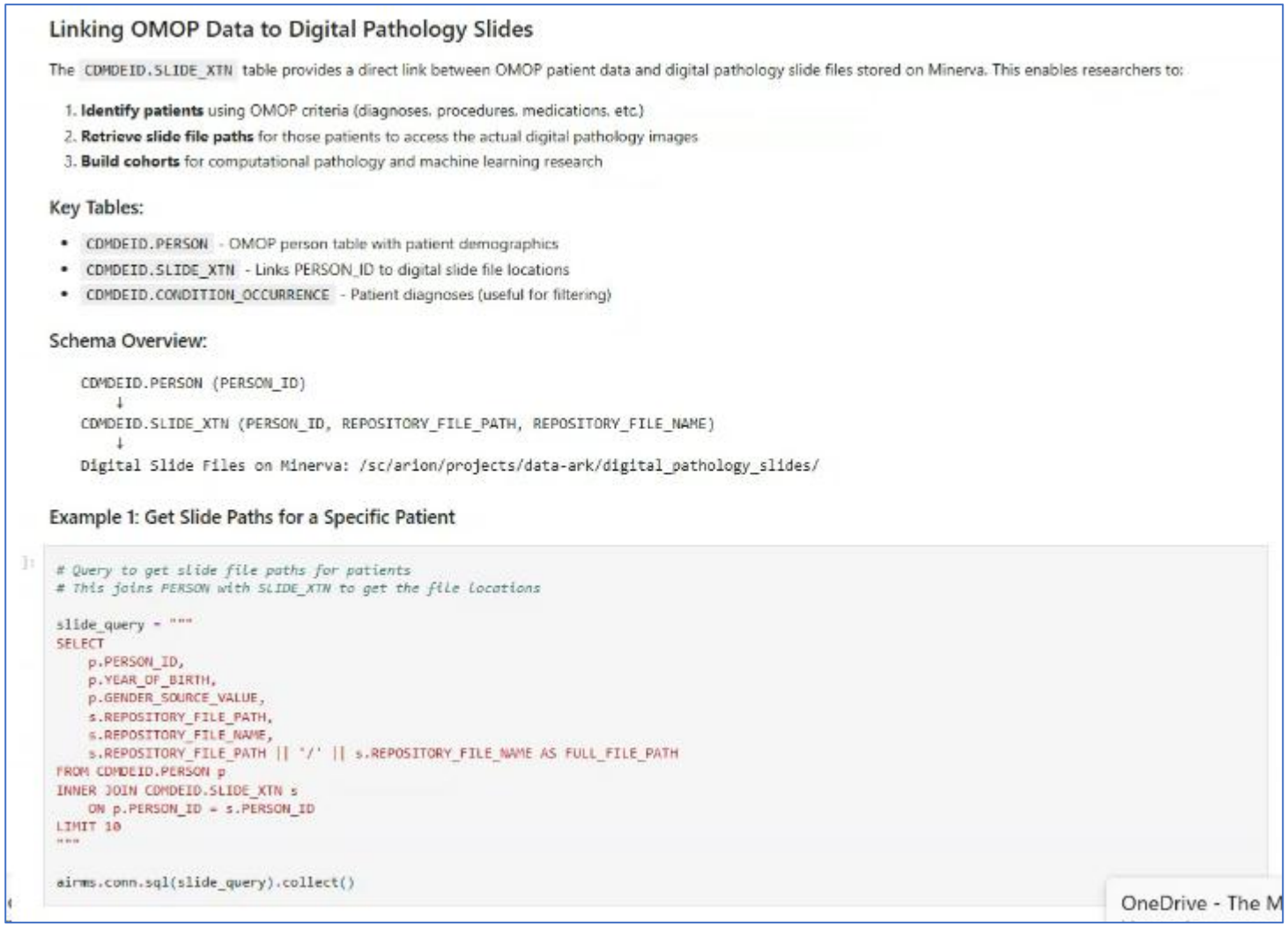
https://github.mountsinai.org/AIRMS/airms-researcher-tutorials-minerva




Explore CDMPATHOLOGY

OMOP DE-ID and PHI Examples

# Sample  Notebooks to Get You Started



## Linking OMOP Data to Digital Pathology Slides

The `CDMDEID.SLIDE_XTN` table provides a direct link between OMOP patient data and digital pathology slide files stored on Minerva. This enables researchers to:

1. **Identify patients** using OMOP criteria (diagnoses, procedures, medications, etc.)
2. **Retrieve slide file paths** for those patients to access the actual digital pathology images
3. **Build cohorts** for computational pathology and machine learning research

### Key Tables:

- `CDMDEID.PERSON` - OMOP person table with patient demographics
- `CDMDEID.SLIDE_XTN` - Links PERSON_ID to digital slide file locations
- `CDMDEID.CONDITION_OCCURRENCE` - Patient diagnoses (useful for filtering)

### Schema Overview:

```
CDMDEID.PERSON (PERSON_ID)
    ↓
CDMDEID.SLIDE_XTN (PERSON_ID, REPOSITORY_FILE_PATH, REPOSITORY_FILE_NAME)
    ↓
Digital Slide Files on Minerva: /sc/arion/projects/data-ark/digital_pathology_slides/
```

### Example 1: Get Slide Paths for a Specific Patient

```
# Query to get slide file paths for patients
# This joins PERSON with SLIDE_XTN to get the file locations

slide_query = """
SELECT
    p.PERSON_ID,
    p.YEAR_OF_BIRTH,
    p.GENDER_SOURCE_VALUE,
    s.REPOSITORY_FILE_PATH,
    s.REPOSITORY_FILE_NAME,
    s.REPOSITORY_FILE_PATH || '/' || s.REPOSITORY_FILE_NAME AS FULL_FILE_PATH
FROM CDMDEID.PERSON p
INNER JOIN CDMDEID.SLIDE_XTN s
    ON p.PERSON_ID = s.PERSON_ID
LIMIT 10
"""

airms.conn.sql(slide_query).collect()
```

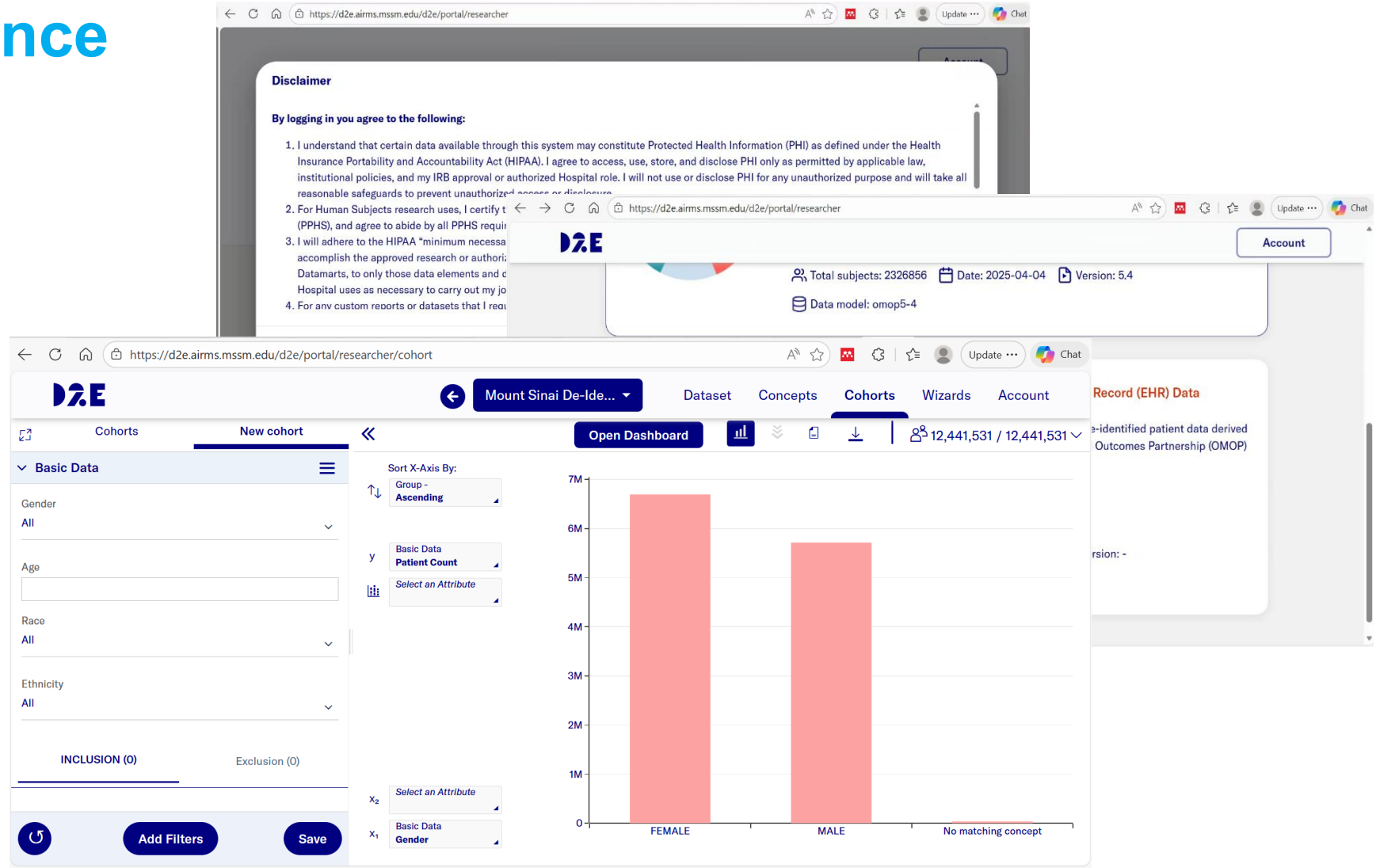
OneDrive - The M

There's code examples for accessing both de-identified and protected health information versions of the digital pathology data

You will need to get an account on github.mountsinai.org

# D2E: Data2Evidence

- Data2Evidence (D2E) is a cohort query tool that allows you to build sophisticated cohorts using a graphical user interface
- Currently undergoing testing and finalizing
- This tool will eventually replace Leaf, ATLAS
- We will announce a training session soon




https://d2e.airms.mssm.edu/

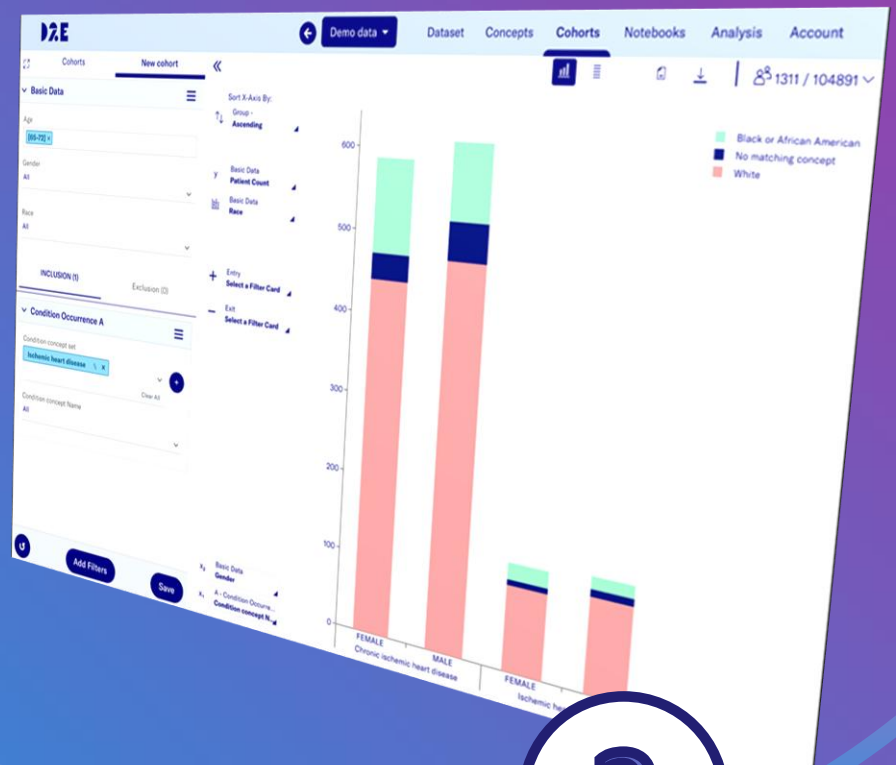
AIR·MS
AI Ready
Mount Sinai

# D2E DATA2EVIDENCE.

## The New **Cohort Query Tool** for Research is Launching Soon!

**Go from Question → Insights with Just a Few Clicks**

D2E's simple point and click interface will make it easier to find cohorts and results




*Formulate a research question or query*

1

2

D2E *does the work*

3

*Get graphs, numbers, and dashboards in seconds!*

Scan the QR Code:
https://d2e.airms.mssm.edu

Icahn School of Medicine at Mount Sinai

# Summary

# Summary & Conclusions

- Find more resources on the Minerva site: https://labs.icahn.mssm.edu/minervalab/

- Linked Digital Pathology data is now available on AI Ready Mount Sinai (AIRMS) and code examples can be found here: https://github.mountsinai.org/AIRMS/airms-researcher-tutorials-minerva

- The slide viewer is under development

- If you have questions, you can submit a ticket here: https://hpims.atlassian.net/servicedesk/customer/portal/67