De-identified Digital Pathology Training Session

Gabriele Campanella

Brandon Veremis

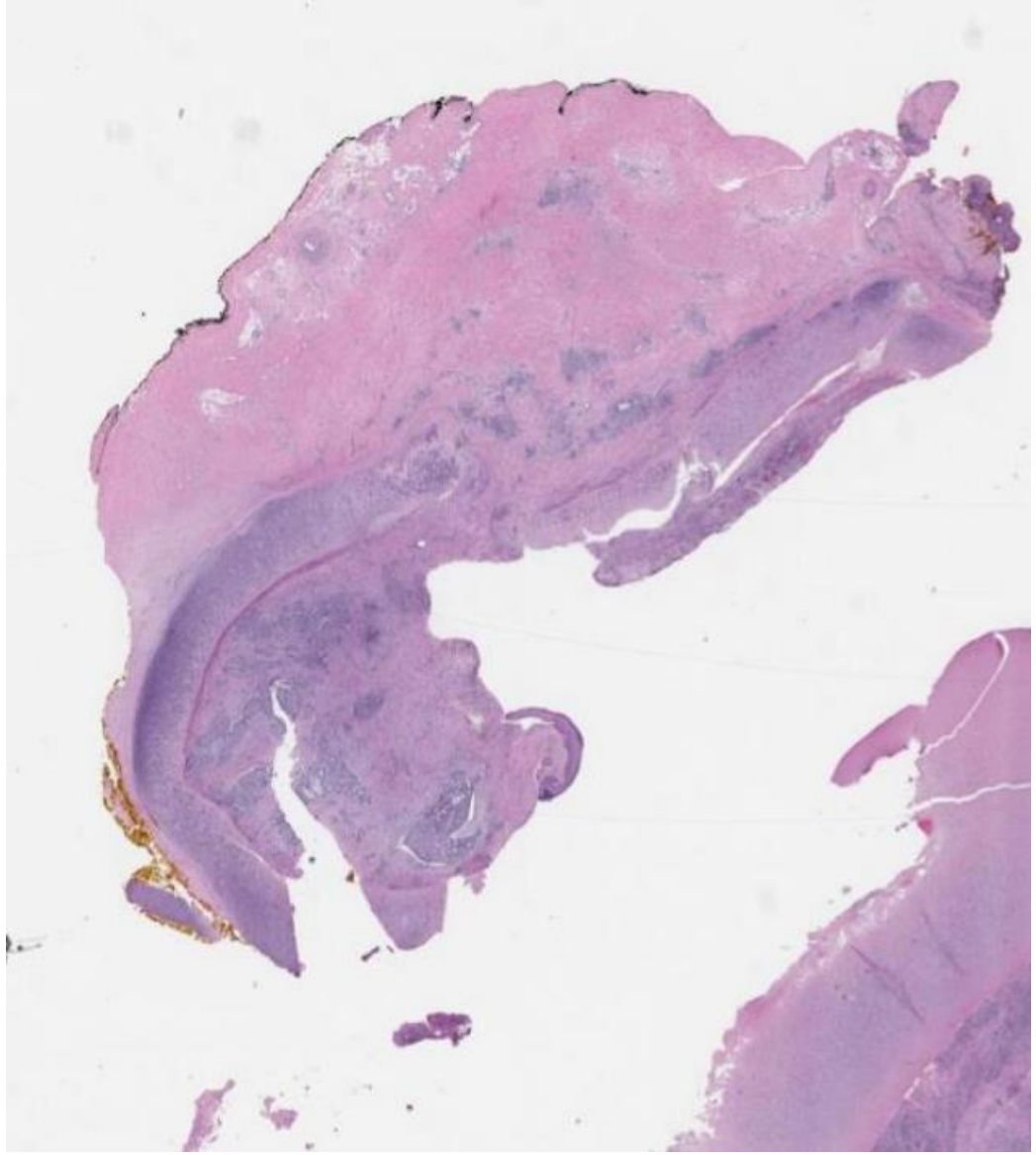
Andrew Deonarine

December 2, 2025

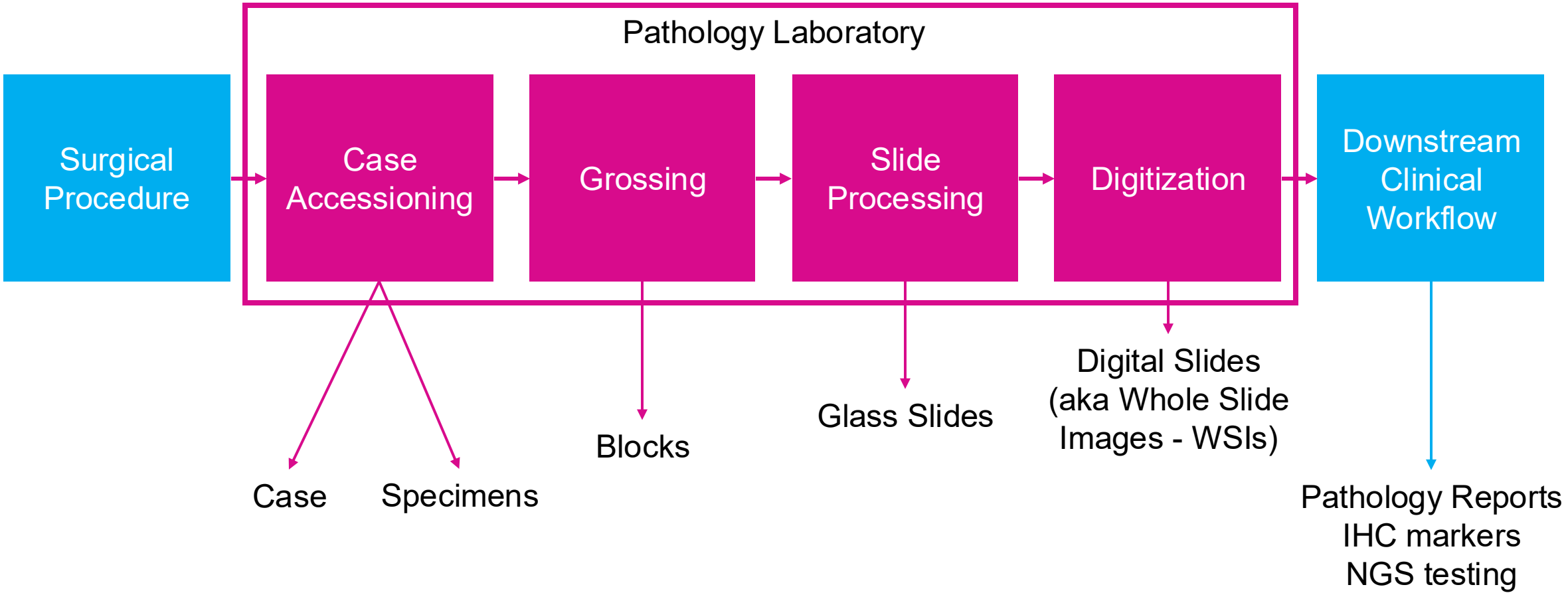
Icahn School of Medicine at Mount Sinai

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) Summary



Pathology Lab Workflow and Data



Overview of Pathology Data

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

Data Modalities

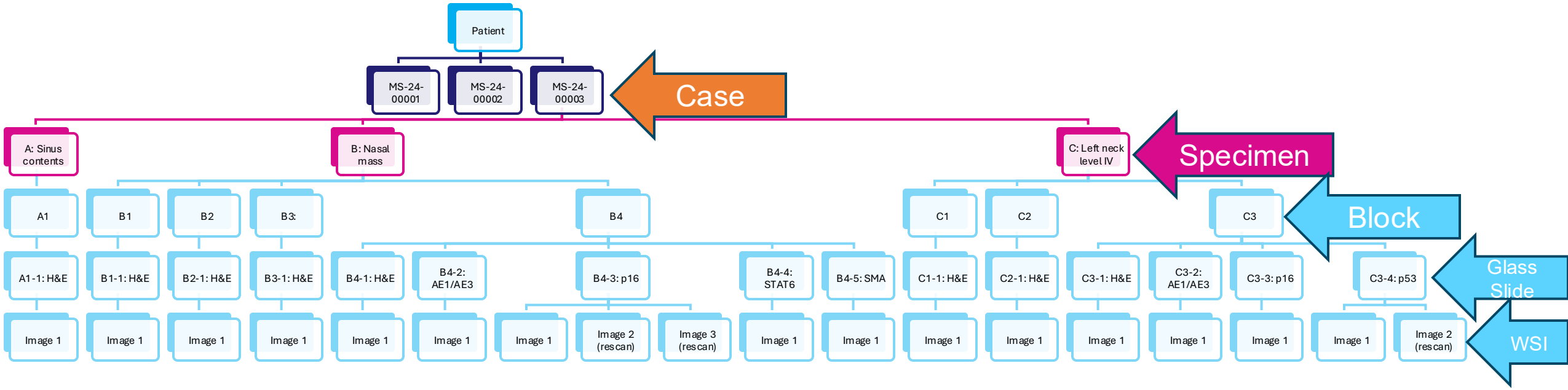
- Biological Assets → Stored for a minimum of 20 years per state regulations
  - Tissue blocks
  - Glass Slides
- Digital Assets
  - Pathology Laboratory Information System (LIS) Metadata → Links cases, specimens, blocks, slides to reports. Tracks most aspects of pathology lab
  - Whole Slide Images → 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 : 16,452,652
- Cases: 2,462,596
- Patients: 1,402,535

Digital Pathology (prospective since 2020, select retrospective)

- Slides: 3,892,209
- Cases: 618,004
- Patients: 440,169

Data Ark

- Slides: 2,593,332

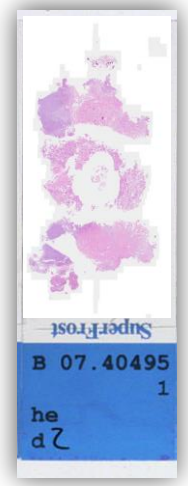
Overview of Pathology Image Data

- 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

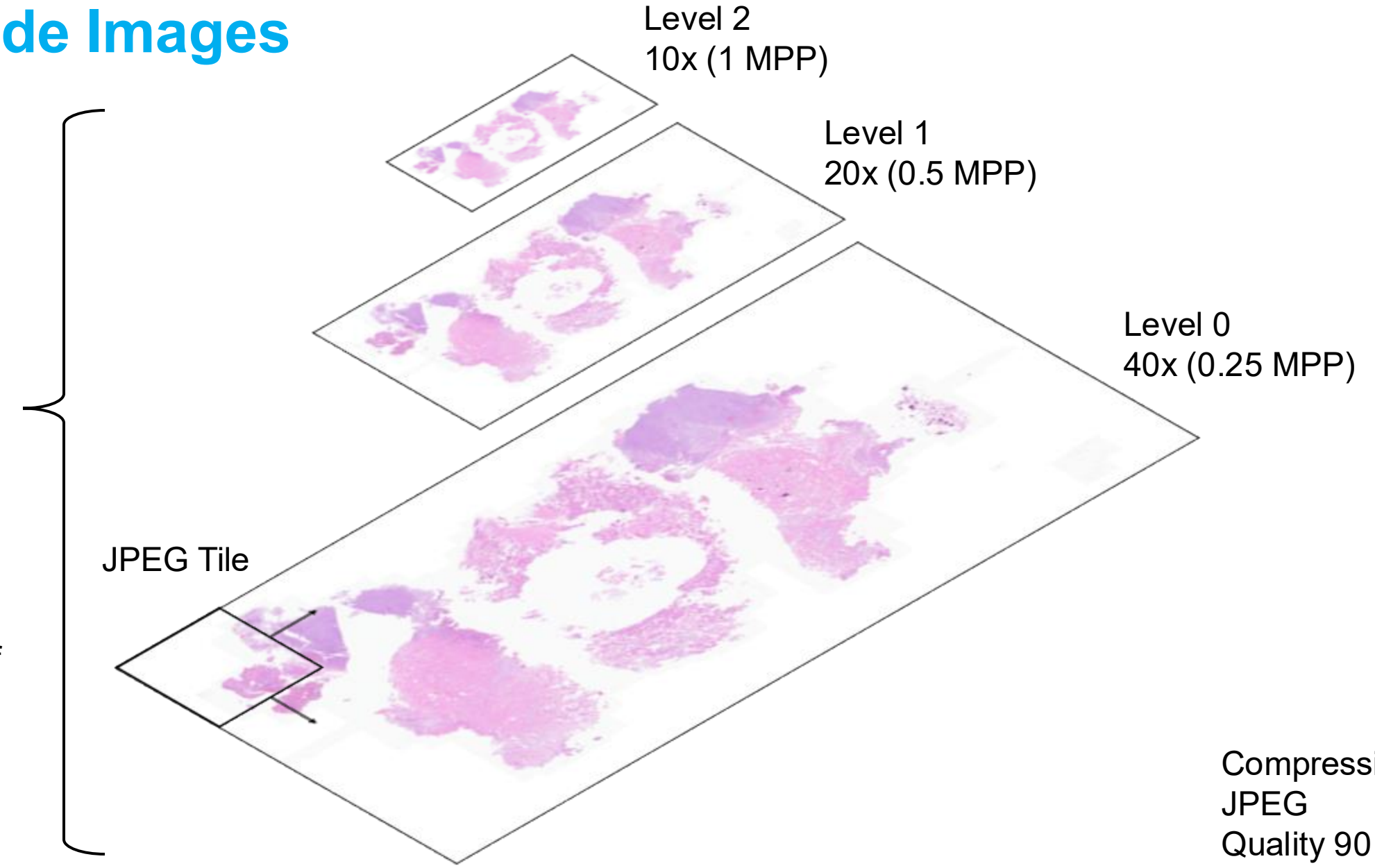
186 000 pixel
B 07.40495
1
he
>x300k

MPP: microns per pixel

Whole Slide Images



Pyramidal .tiff



Compression:
JPEG
Quality 90

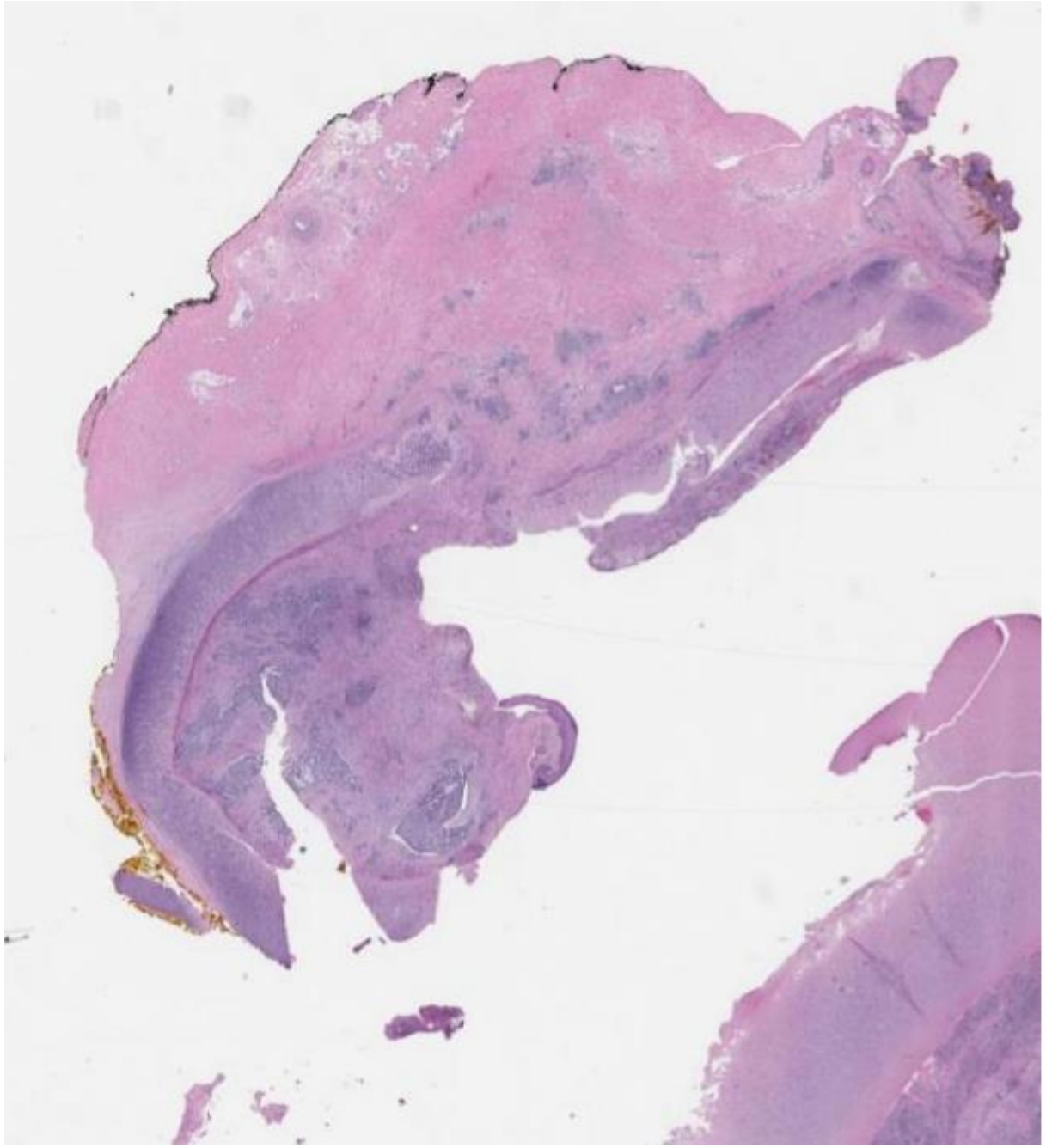
How To Interact With Images

- 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

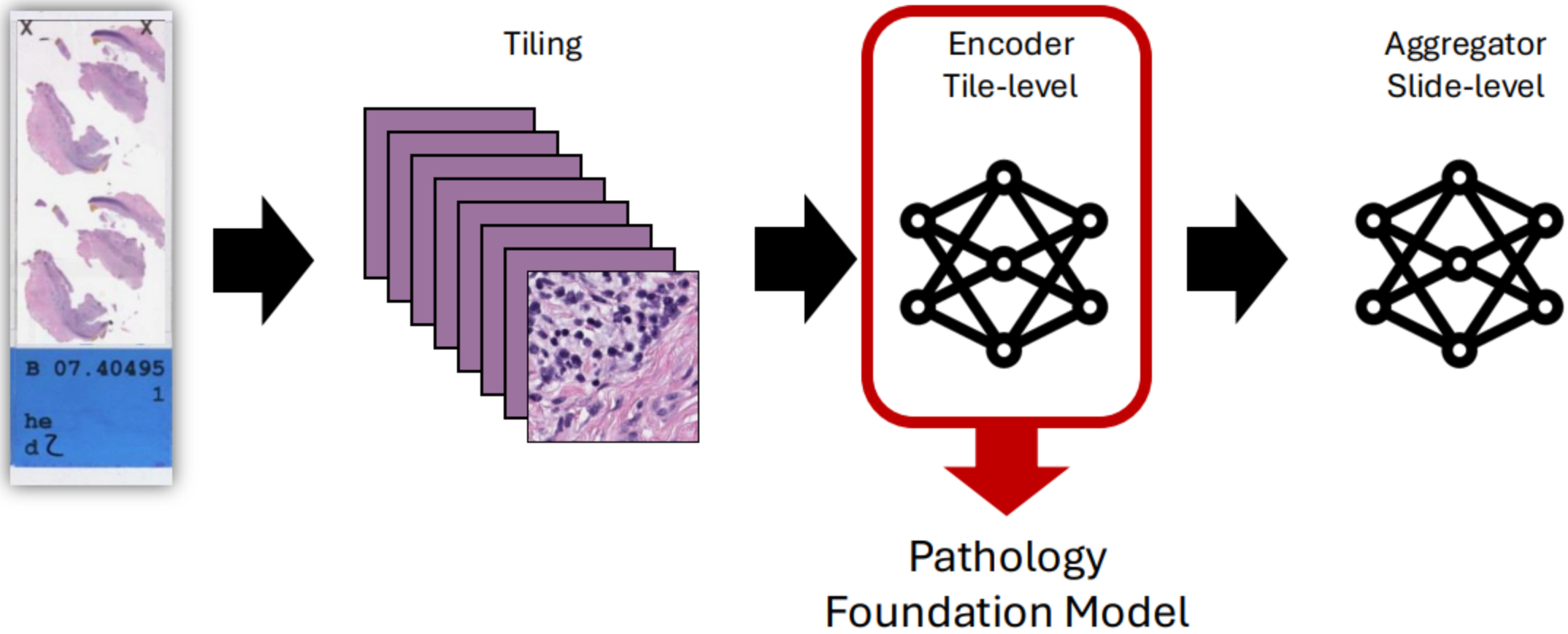
Diagnostic Reporting

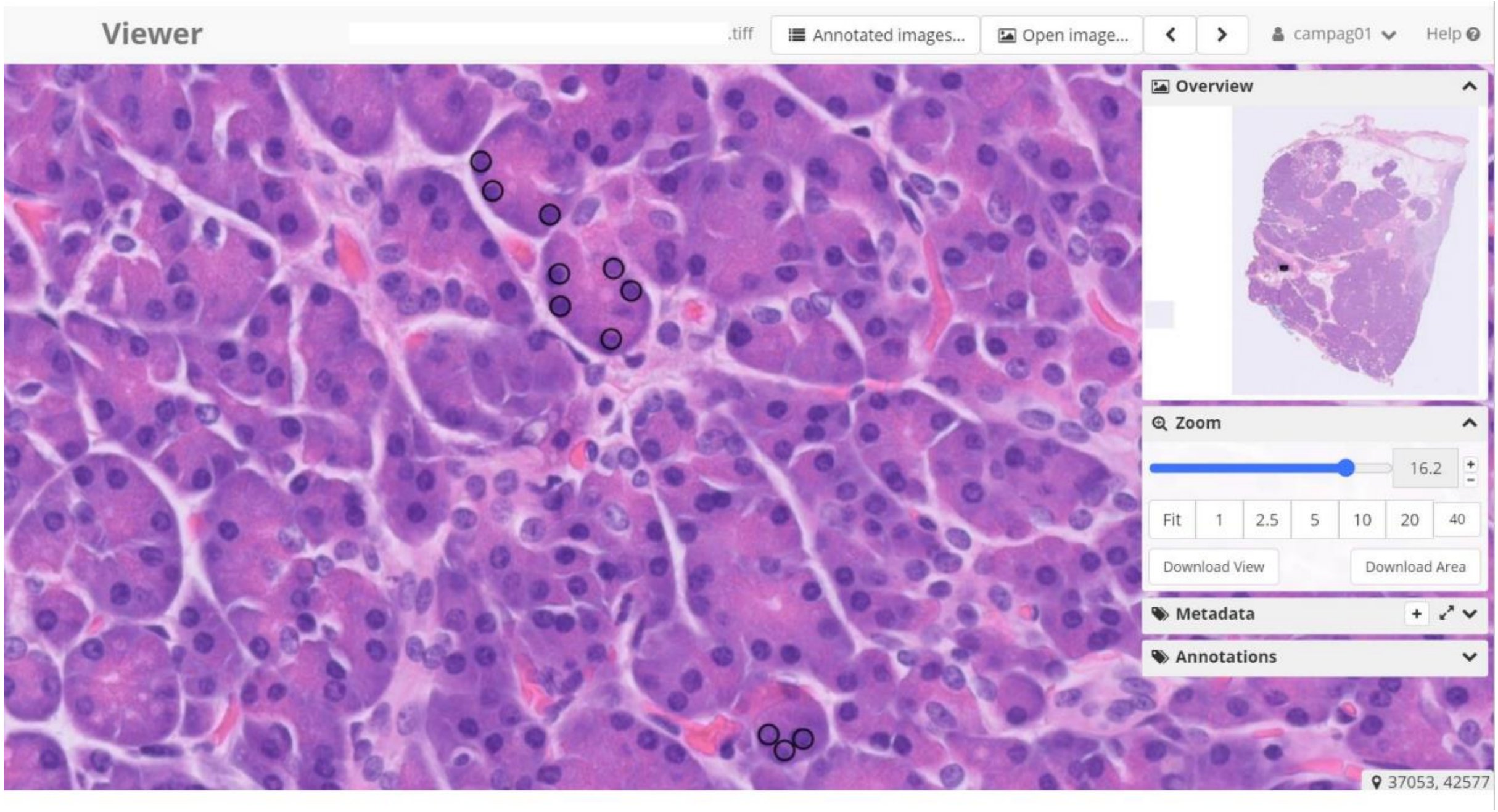
- Tumor detection
- Tumor segmentation
- Tumor staging
- Cell detection
- Slide retrieval

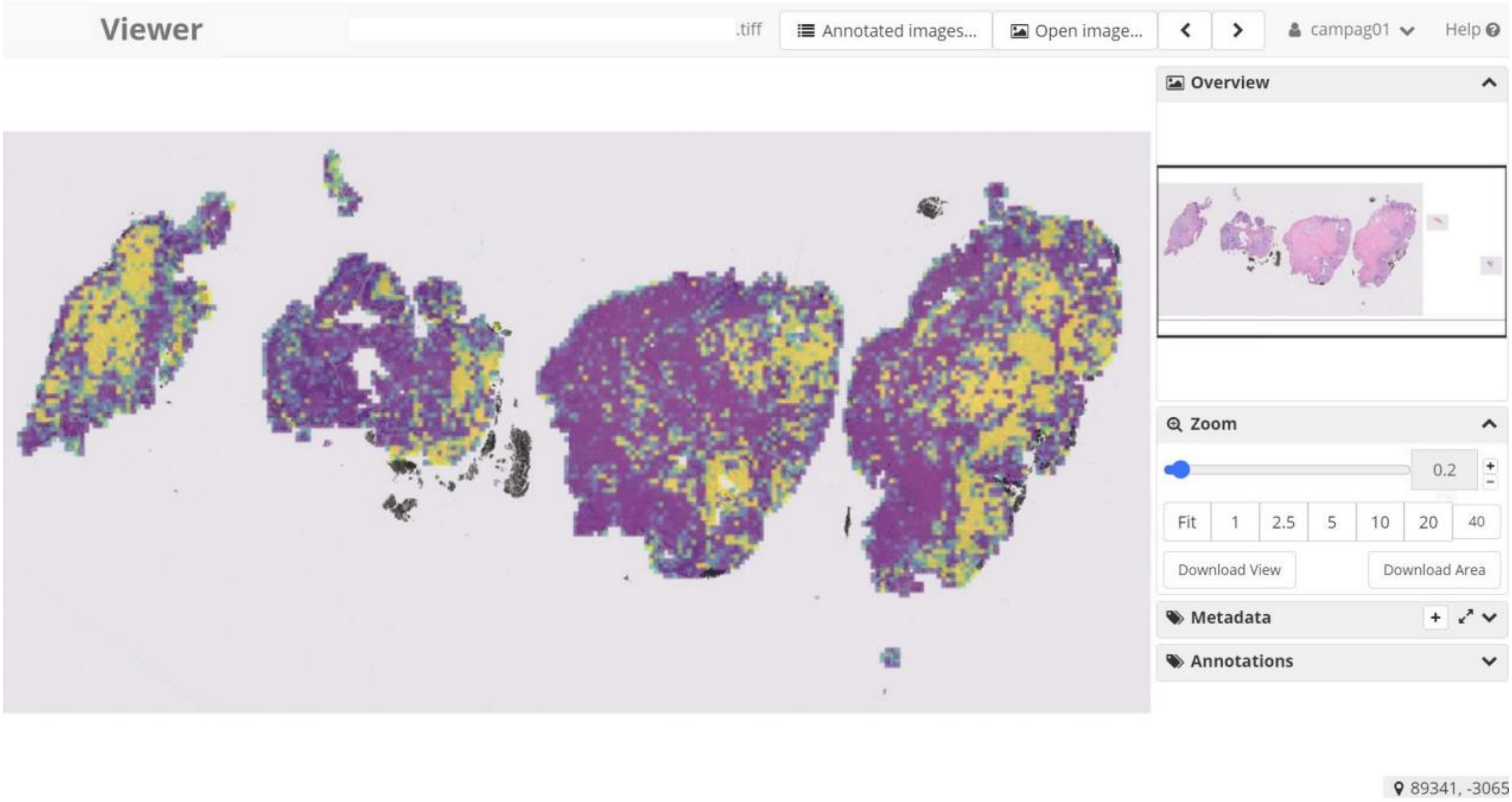
Prognostication

- 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

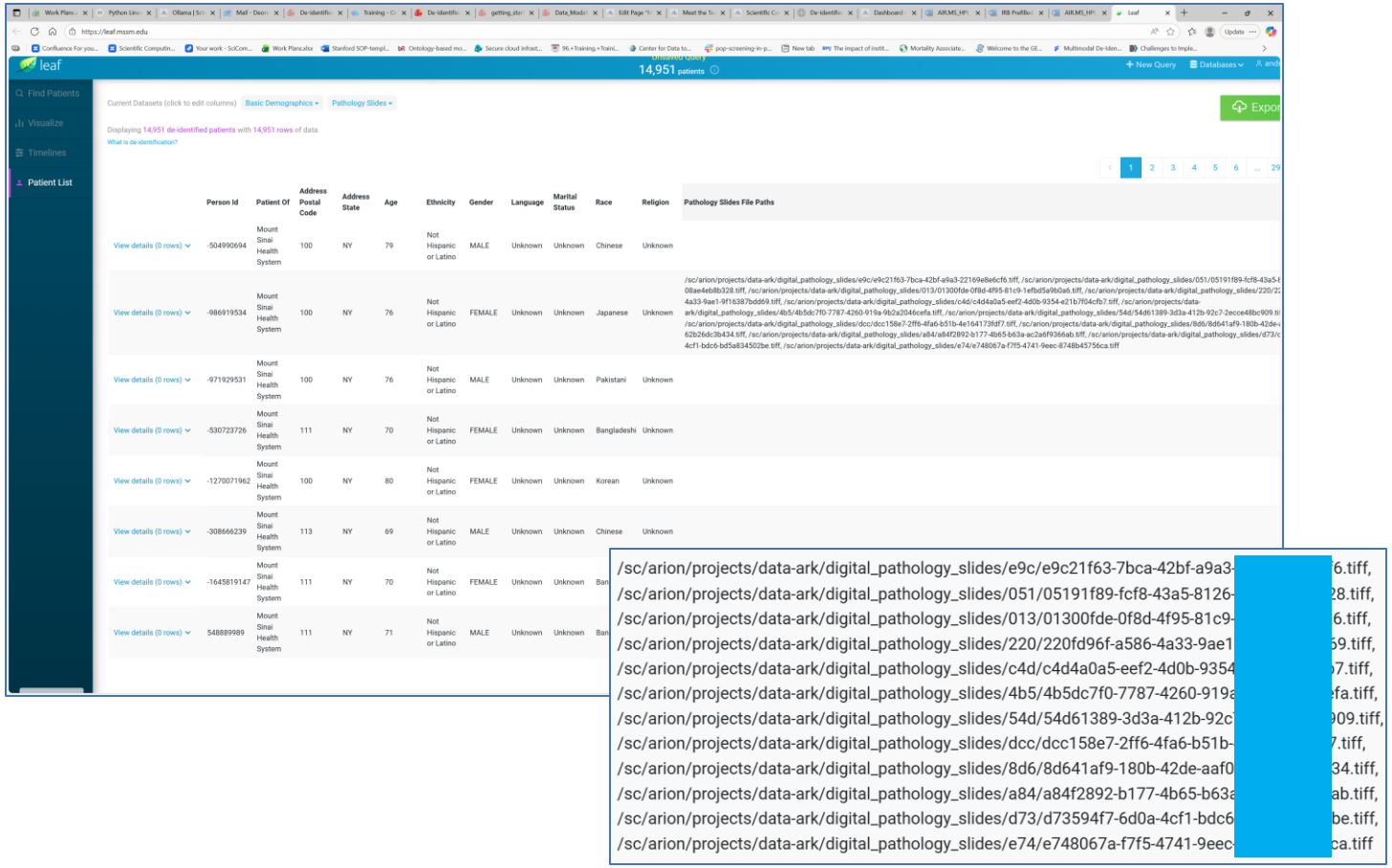
- Preface: pathology report gives specimen level information, not slide level
  - Manual slide review
  - Rely on specialty specific knowledge
    - For breast cancer, only tumor slides receive ER/PR/HER2 testing
- Rely on synoptic reporting
  - Newer synoptic reports include “best” tumor block
  - Cons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

TLDR

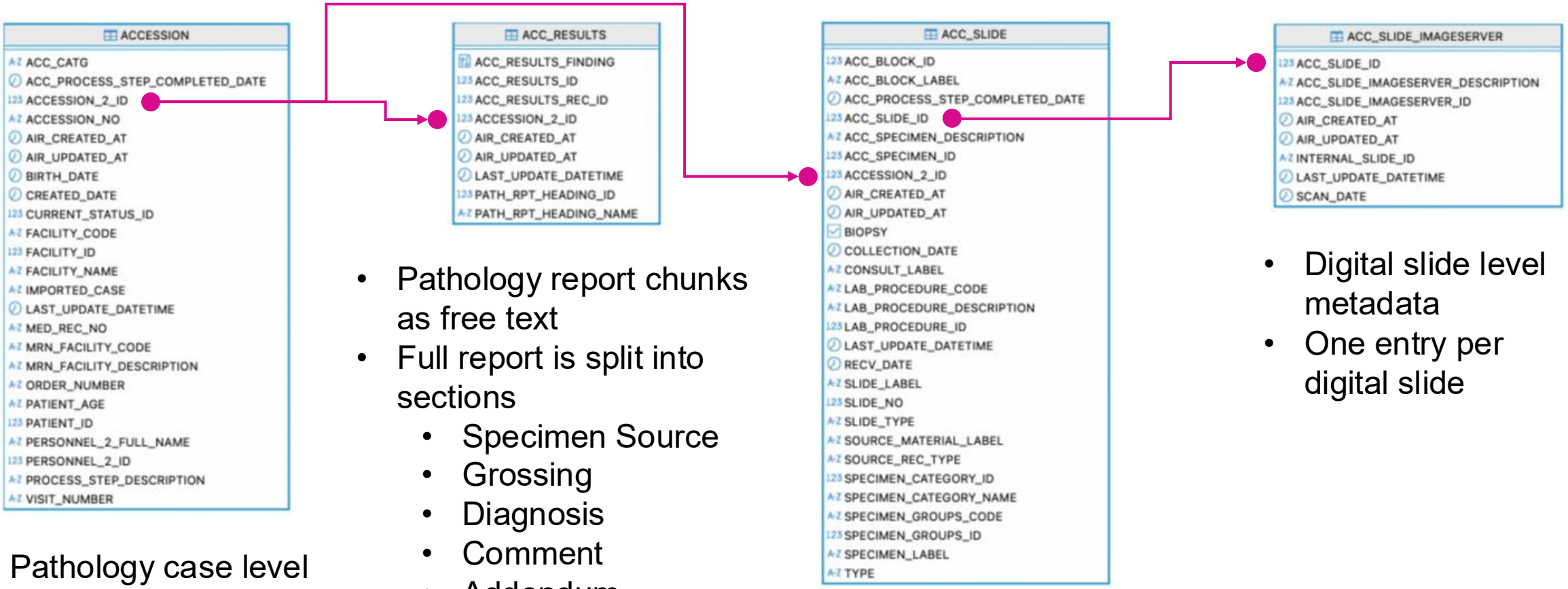
Pathologist collaboration is essential!

Using Leaf to Locate Digital Pathology Images on Data Ark

- Example of file linkages in Leaf (to the file in Data Ark)
- These are linked at the patient level, so you will get a list of files (working on refining this linkage)
- We are in the process of incorporating this into AIRMS



Pathology Data in AIR·MS



- Pathology case level metadata
- One entry per case

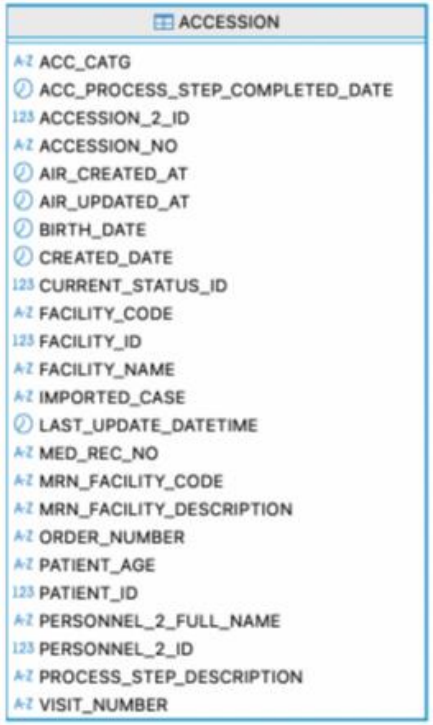
- Pathology report chunks as free text
- Full report is split into sections
  - Specimen Source
  - Grossing
  - Diagnosis
  - Comment
  - Addendum
- One entry per section

- Glass slide level metadata
- One entry per glass slide

- Digital slide level metadata
- One entry per digital slide

AIRMS Schema: CDMPATHOLOGY

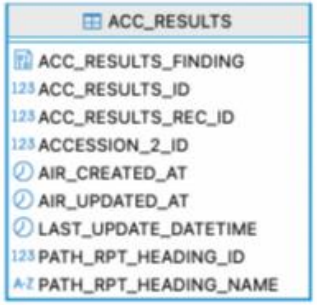
Case Metadata



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

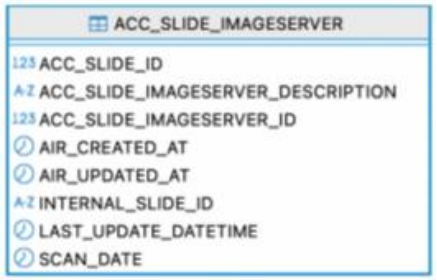
ACC_SLIDE

- ACC_BLOCK_ID
- ACC_BLOCK_LABEL
- ACC_PROCESS_STEP_COMPLETED_DATE
- ACC_SLIDE_ID
- ACC_SPECIMEN_DESCRIPTION
- ACC_SPECIMEN_ID
- ACCESSION_2_ID
- AIR_CREATED_AT
- AIR_UPDATED_AT
- BIOPSY
- COLLECTION_DATE
- CONSULT_LABEL
- LAB_PROCEDURE_CODE
- LAB_PROCEDURE_DESCRIPTION
- LAB_PROCEDURE_ID
- LAST_UPDATE_DATETIME
- RECV_DATE
- SLIDE_LABEL
- SLIDE_NO
- SLIDE_TYPE
- SOURCE_MATERIAL_LABEL
- SOURCE_REC_TYPE
- SPECIMEN_CATEGORY_ID
- SPECIMEN_CATEGORY_NAME
- SPECIMEN_GROUPS_CODE
- SPECIMEN_GROUPS_ID
- SPECIMEN_LABEL
- TYPE

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. <u>Mapping is not publicly available.</u>

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)

Summary

Summary & Conclusions

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

- We are in the process of developing more comprehensive links to the digital pathology files

- The slide viewer is under development

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