

De-identified Digital Pathology Training Session

March 21, 2025

Basic pathology asset structure

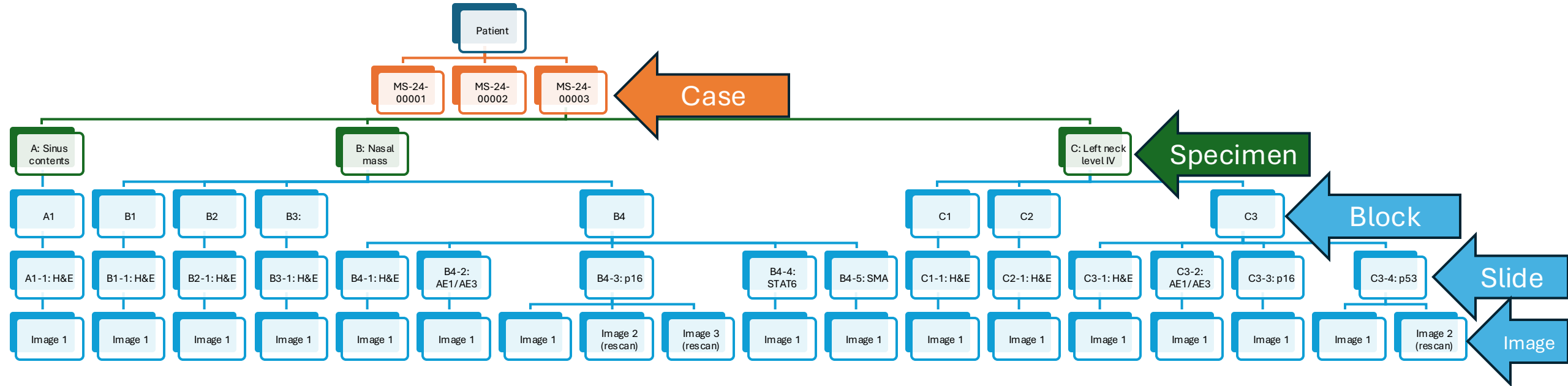
One to many to many to many to many to many

- 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

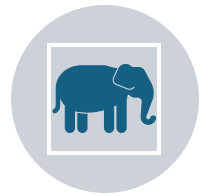


Basic pathology asset structure

One to many to many to many to many to many



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)



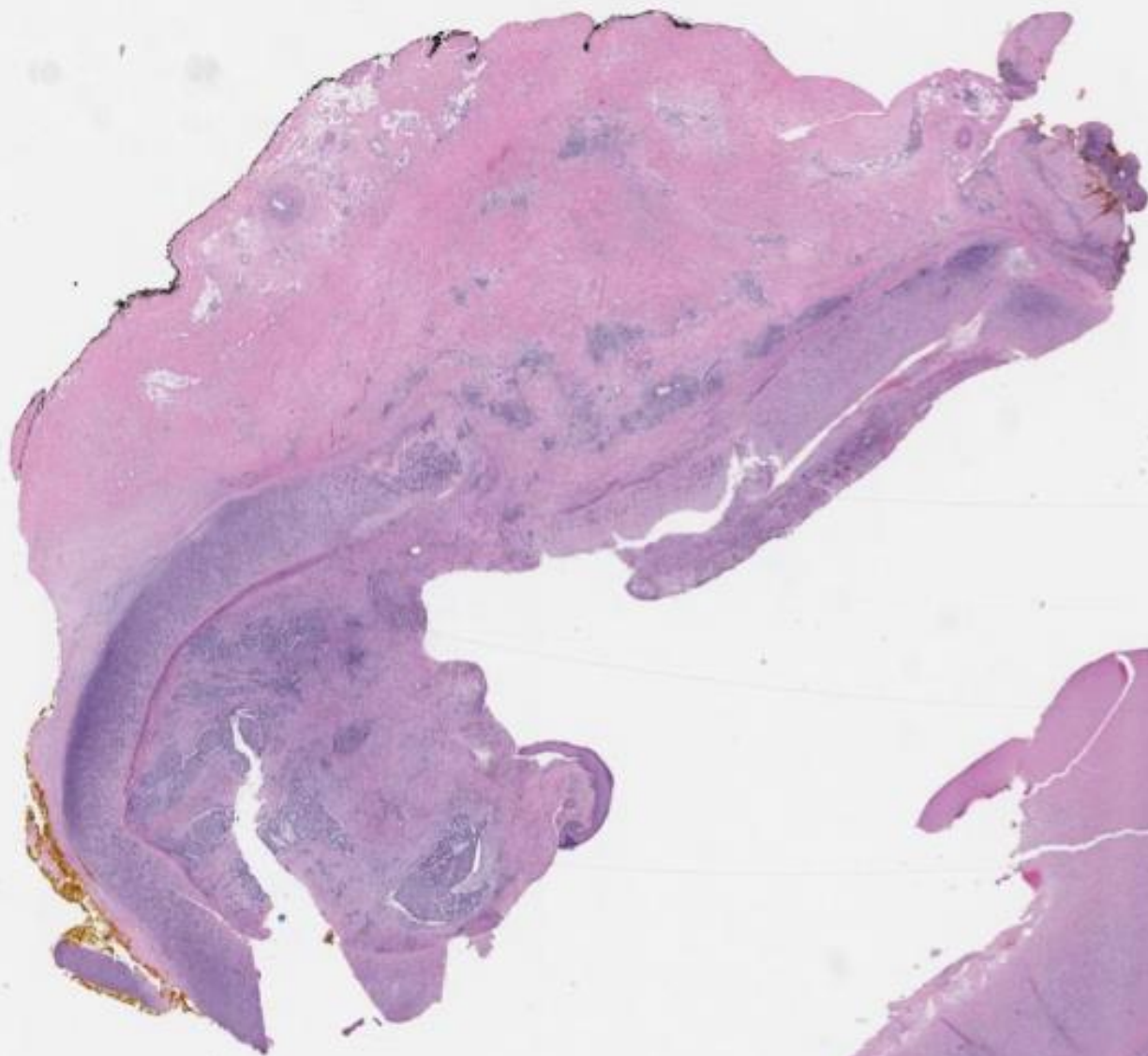
TIFF format



Open-source tools available



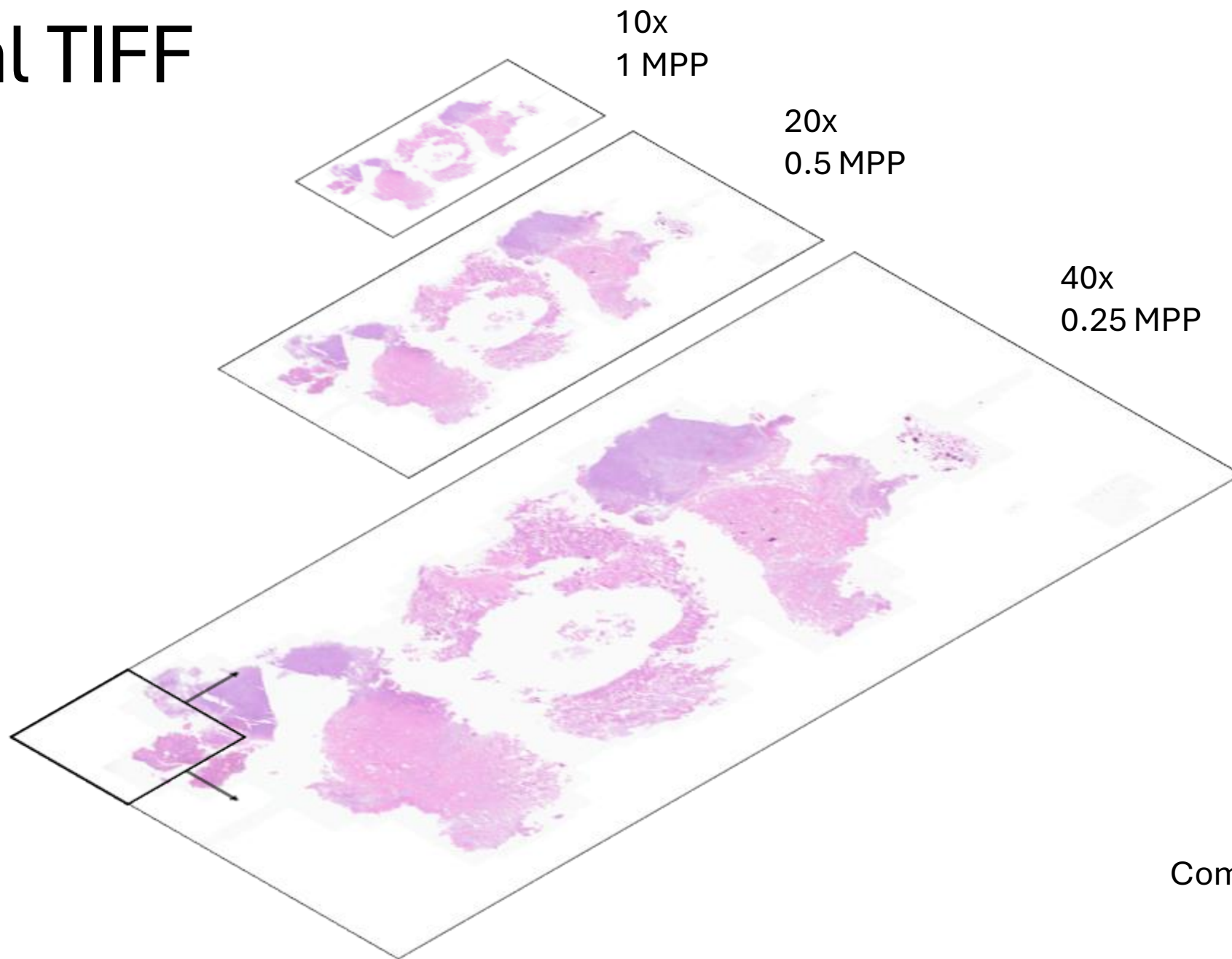
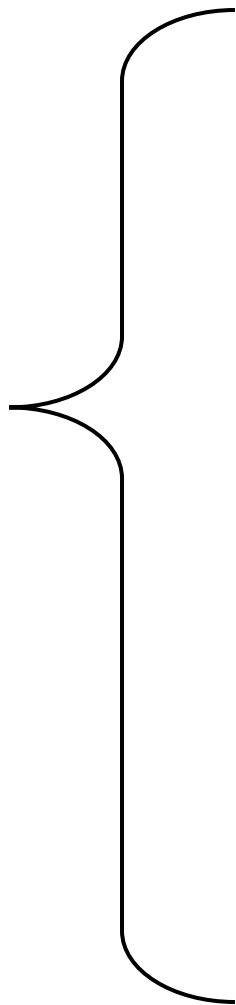
186 000 pixel



>x300k



Pyramidal TIFF Files



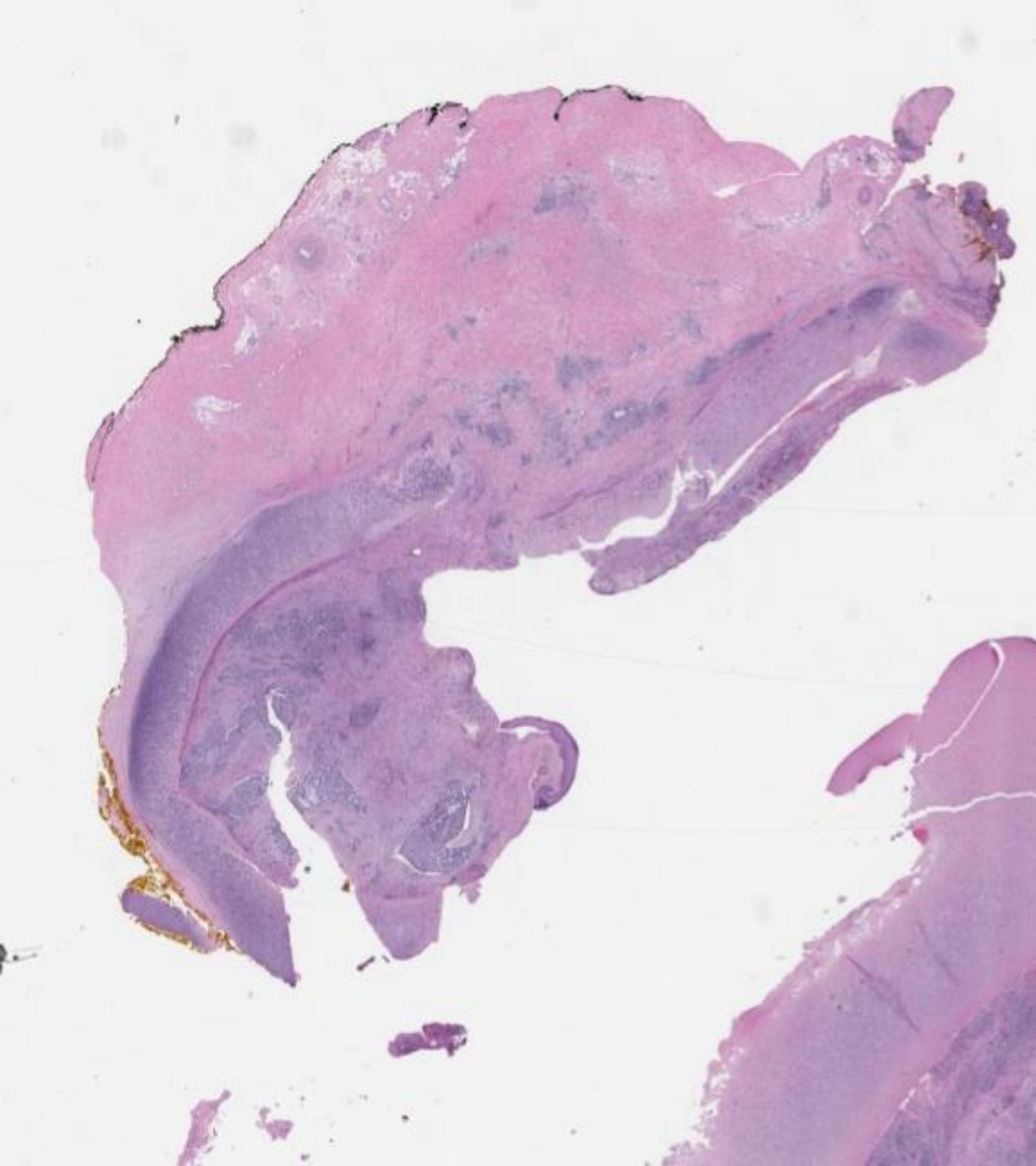
Compression:
JPEG
Quality 90

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')
```

Quality Control

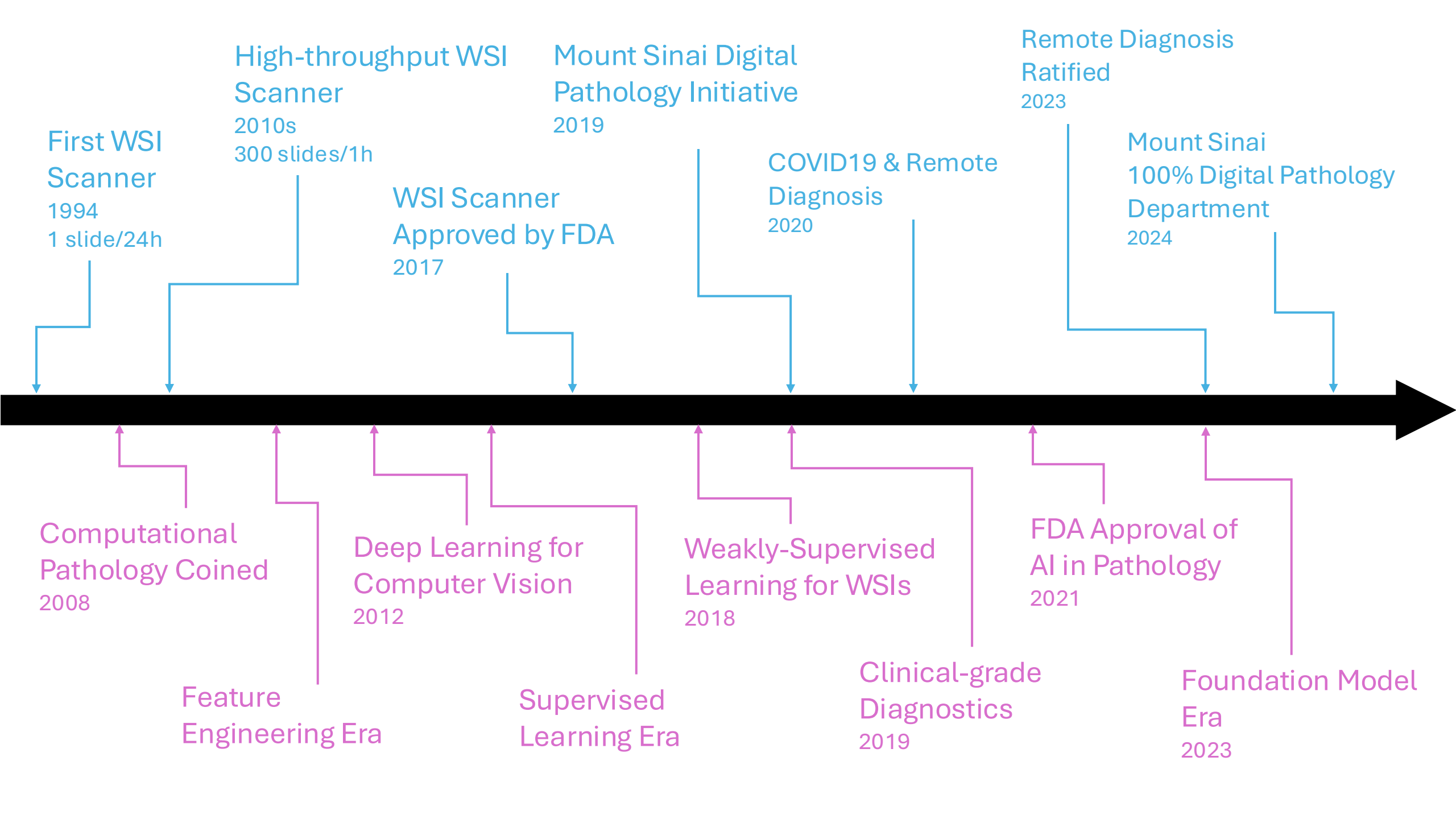
- Detect blurred regions
- Detect tissue folds
- Stain normalization

Diagnostic Reporting

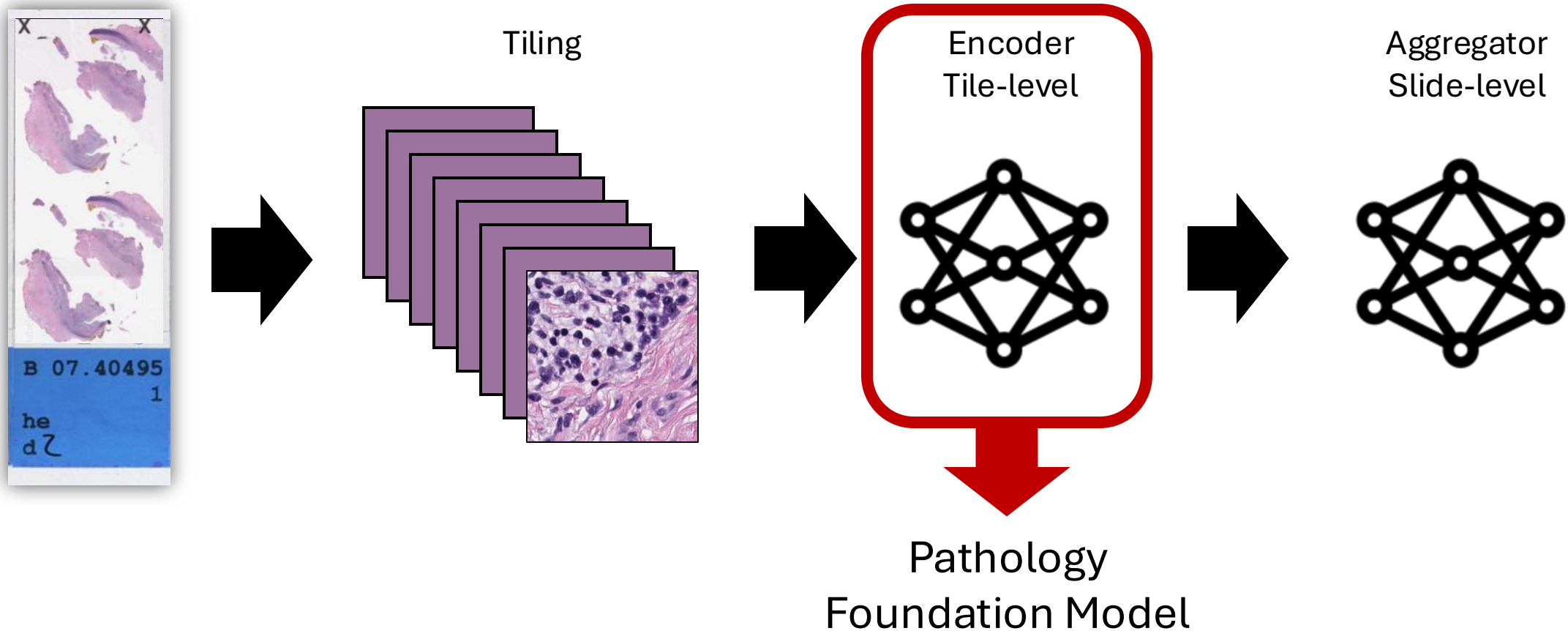
- Tumor detection
- Tumor segmentation
- Tumor staging
- Cell detection
- Slide retrieval

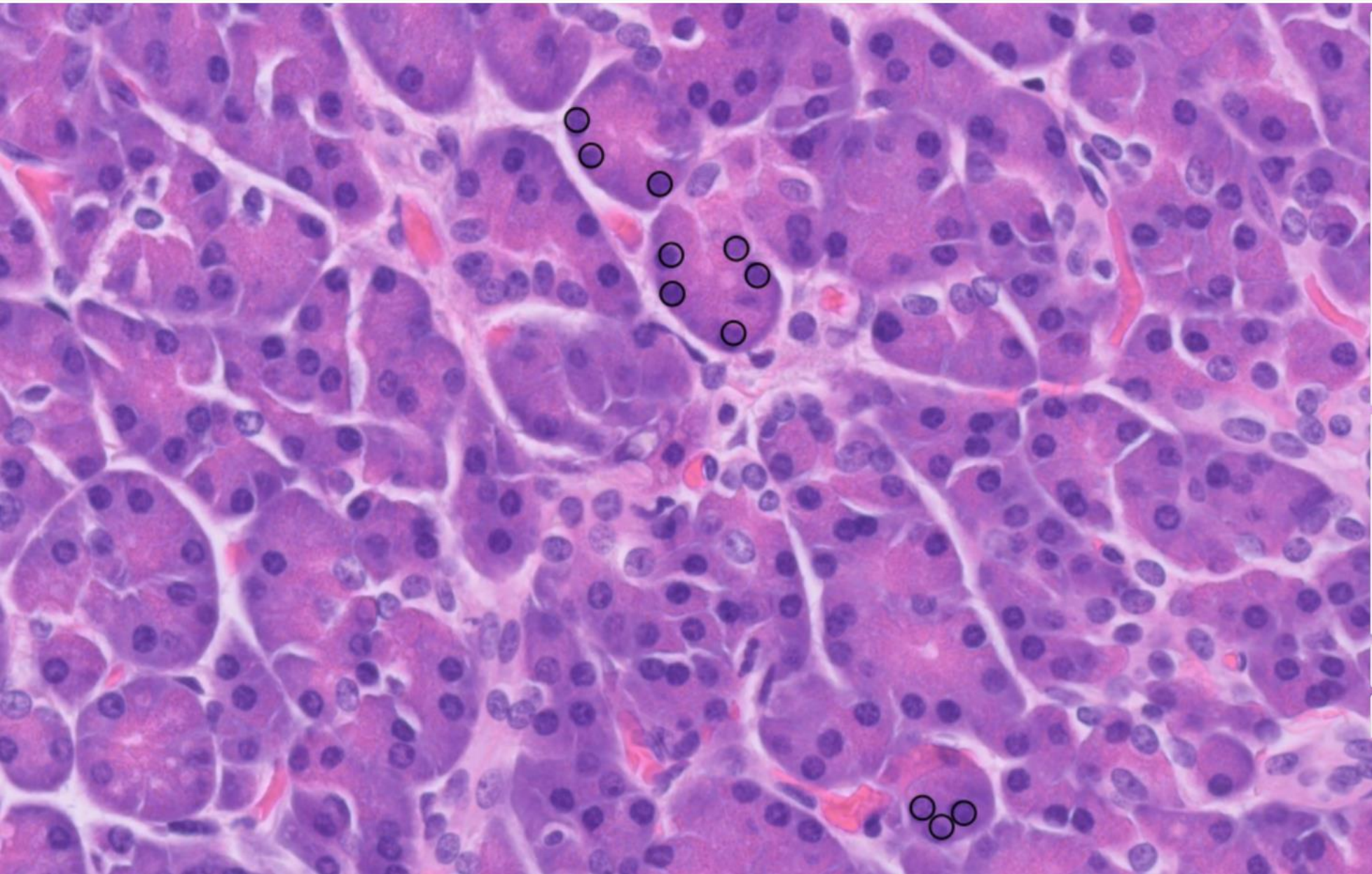
Prognostication

- Biomarker prediction
- Response prediction
- Survival analysis



Computational Pathology





Overview

A small thumbnail image showing the entire tissue section, providing a broader context for the current view. The thumbnail shows the overall shape and distribution of the tissue, with a small black rectangle indicating the area currently displayed in the main viewer.

Zoom

16.2

Fit 1 2.5 5 10 20 40

Download View Download Area

Metadata

Annotations

Viewer

.tiff

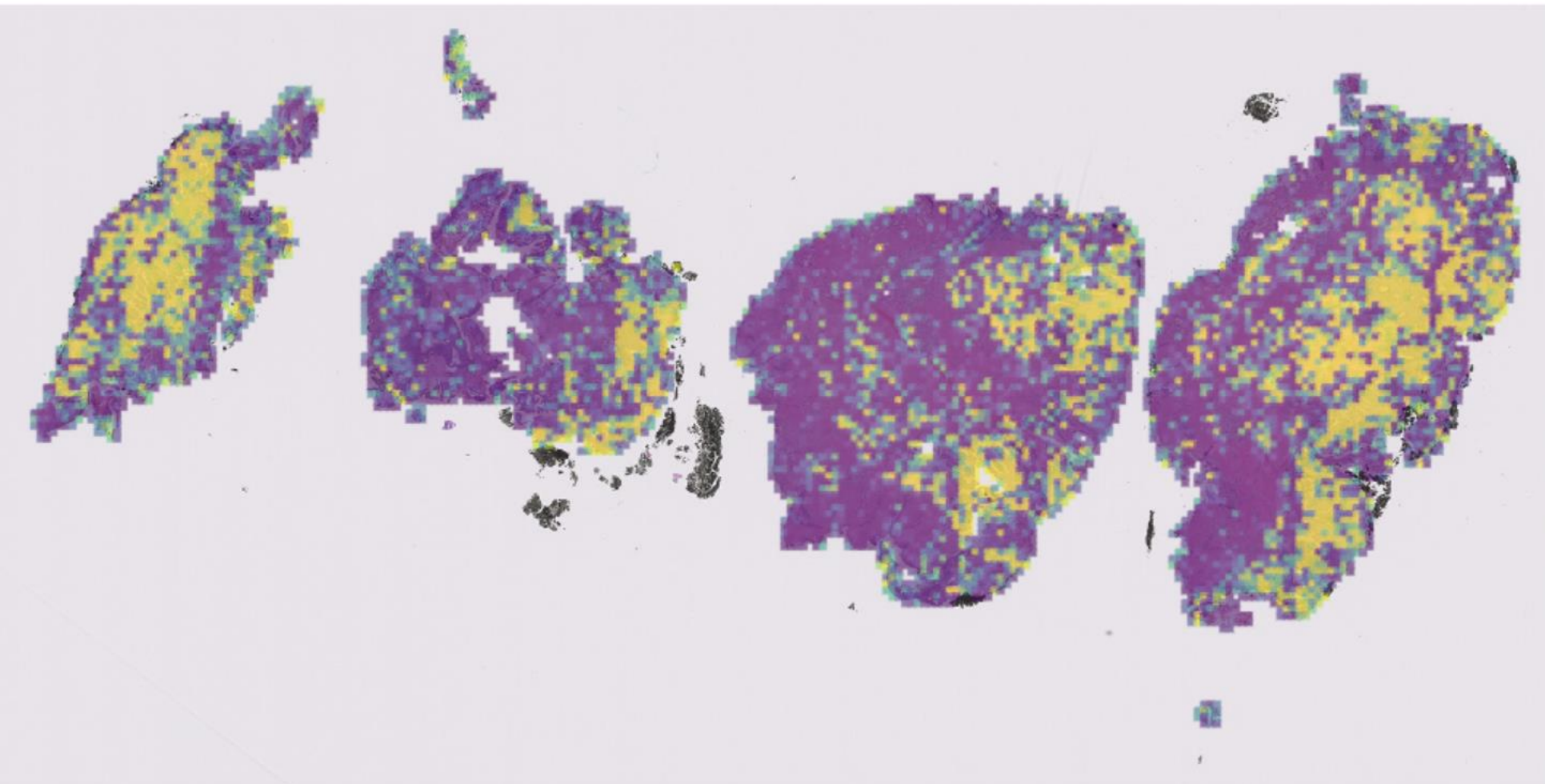
☰ Annotated images...

🖼️ Open image...

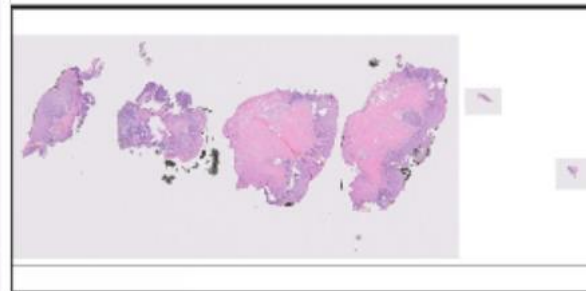


👤 campag01 ▾

Help ⓘ



🖼️ Overview



🔍 Zoom



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!

Identify Cohorts with Digital Pathology Slides on Minerva Using Leaf

The screenshot shows the Leaf web application interface. The top navigation bar includes the Leaf logo, a search bar, and a status bar indicating "Unsaved Query" and "3,058 patients". The left sidebar contains navigation options: "Find Patients", "Visualize", "Timelines", and "Patient List". The main content area displays a query builder with a "Limit to" dropdown and a "Save Query" button. The query builder is structured as follows:

Patients Who	And	And
Anytime	Anytime	Anytime
At Least 1x	At Least 1x	At Least 1x
BREAST	POSITRON EMISSION TOMOGRAPHY (PET)	
In the Same Encounter	In the Same Encounter	

The left sidebar also shows a list of concepts under "All Concepts":

- Cancer Registry
 - Cancer Stage
 - ACCESSORY SINUS
 - ADRENAL GLANDS
 - ANAL CANAL & ANUS
 - APPENDIX
 - BASE OF TONGUE
 - BLOOD, BONE MARROW, & HEMATOPOIETIC SYS
 - BONES & JOINTS (EXCL SKULL AND FACE)
 - BONES OF SKULL AND FACE
 - BRAIN, & CRANIAL NERVES, & SPINAL CORD, (EXCL. VENTRICLE, CEREBELLUM)
 - BREAST
 - STAGE 1
 - STAGE 2
 - STAGE 3
 - STAGE 4
 - STAGE NOT AVAILABLE
 - CEREBELLUM
 - CERVIX UTERI
 - COLON

Cohort inclusion Criteria

- 1) Breast cancer (any stage)
- 2) PET Scan in IRW 2.0

Using Leaf to Locate Digital Pathology Images on Data Ark

The screenshot shows the Leaf application interface. At the top, there's a blue header with the 'leaf' logo, 'Unsaved Query 3,058 patients', and user information 'sharon.nirenberg'. A sidebar on the left contains navigation options: 'Find Patients', 'Visualize', 'Timelines', and 'Patient List' (which is highlighted). The main area displays 'Current Datasets' with tabs for 'Basic Demographics' and 'Pathology Slides' (the latter is selected and highlighted with a blue box). Below the tabs, it says 'Displaying 2,878 de-identified patients with 2,878 rows of data'. A table lists patient information with columns: Person Id, Patient Of, Address Postal Code, Address State, Age, Ethnicity, Gender, Language, Marital Status, Race, Religion, and 'Pathology Slides File Paths' (highlighted with a blue box). The first row shows a patient with 'No matching concept'. The second row shows a patient with a file path: '/sc/arion/projects/data-ark/digital_pathology_slides/485/4856a4df-73c1-4aac-a300-497ab01fdc02.tiff'. A purple starburst callout points to this path with the text 'File location of the de-identified pathology image on Data Ark'. The bottom of the screen features an orange banner with text about a training session.

Person Id	Patient Of	Address Postal Code	Address State	Age	Ethnicity	Gender	Language	Marital Status	Race	Religion	Pathology Slides File Paths
1118470934	Mount Sinai Health System	100	NY	75	No matching concept	FEMALE	Unknown	Unknown	White	Unknown	
2081909695	Mount Sinai Health System	112	NY	54	Not Hispanic	FEMALE	Unknown	Unknown	African	Unknown	/sc/arion/projects/data-ark/digital_pathology_slides/485/4856a4df-73c1-4aac-a300-497ab01fdc02.tiff, /sc/arion/projects/data-
-311036902	Mount Sinai Health System	112	NY	48	Not Hispanic or Latino	FEMALE	Unknown	Unknown	White	Unknown	ark/digital_pathology_slides/d70/d70e9885-8d2b-45bf-b963-9ac9bfe48e.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/df1/df16dd4d-ed11-4818-8e14-441bed381.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/3c3/3c352a0d-6892-40f4-9c8e-3294aa0815a.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/62/62f835ec79-50b3-4f17-9add-b7f4fc2ba7.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/90/90f509-45da-bffa-40b652a36b.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/e0b/e0b00000-0000-0000-0000-000000000000.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/0bc/0bcda190-0000-0000-0000-000000000000.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/d41/d417c72f-0000-0000-0000-000000000000.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/10c/10c00000-0000-0000-0000-000000000000.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/22f/22f720-0000-0000-0000-000000000000.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/839/8399b6dd-d8c1-42e2-8000-000000000000.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/479/4795da29-1bb3-4aef-0000-000000000000.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/f54/f5465bf5-184f-4f31-0000-000000000000.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/020/02081662-7fc5-4ed8-ac61-651390c73c95.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/f65/f65b0f47-1f9c-4158-8114-a6ba874a3712.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/2ab/2ab8368d-8fd2-4b7c-b152-bd3238ab839c.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/0d3/0d378adb-8556-467b-b38f-966ab5bb3d69.tiff, /sc/arion/projects/data-

Learn how to access the Digital Pathology Images on **Minerva Data Ark Training Session**

Thursday, March 27, 1-2PM

Icahn School of Medicine building (1425 Madison Ave) Room L3-36 and Zoom

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, PhD (rachel.brody@mountsinai.org)