

# PathML

## An Open-Source Software Toolkit for Computational Pathology Research

---

**Pathology Innovation Collaborative Community (Picc)**

Jacob Rosenthal, Renato Umeton, and Team

June 2022



**Weill Cornell  
Medicine**



**Dana-Farber  
Cancer Institute**

# The Team



**Jacob Rosenthal, MS**

Data scientist



**Ryan Carelli**

Data scientist



**Massimo Loda, MD**

Chair of Pathology and  
Laboratory Medicine, Weill  
Cornell Medicine

Pathologist-in-Chief,  
NewYork-Presbyterian  
Hospital



**Renato Umeton, PhD**

Assoc. Dir. AI Operations &  
Data Science Svcs,  
Informatics & Analytics Dept.,  
Dana-Farber Cancer Institute

All dually affiliated with  
**Dana-Farber** and **Weill Cornell**

Additional collaborators:  
David Brundage, Mohamed Omar, Karen Xu, Luigi Marchionni

**& the open-source community – which helped a lot since the project started in 2019!**

# Renato Umeton, Ph.D.



- Bachelor in Computer Science – Summa Cum Laude
- Master in Computer Science – Summa Cum Laude
- Ph.D. in Mathematics and Informatics
  - Thesis defense: Optimization and Ontology for Computational & Systems Biology
  - Advisors: Giuseppe Nicosia & Salvatore Di Gregorio. Adj: C. Forbes Dewey Jr.
  - A journey across: U. of Calabria
  - Microsoft – Synthetic/Computational Biology Group
  - Massachusetts Institute of Technology – Departments: Biological Engineering, Mechanical Engineering. Collaborations: CSAIL, LIDS.

— *fast forward after working in other hospitals, academia, consulting, and industry, in roles ranging from postdoc to director* —



- 2015: Awarded a green card by the United States as Engineer of National Interest
- 2016: Joined **Dana-Farber Cancer Institute** in the Informatics & Analytics department
- 2021:
  - Associate Director of Artificial Intelligence Operations & Data Science Services, reporting to the Chief Data and Analytics Officer
  - Affiliate at MIT, Harvard T.H. Chan School of Public Health, and Weill Cornell Medicine
  - Always contributed to research & development:
    - 110+ scientific works co-authored so far (AI, Cancer research, Machine learning, Data science, Biological Engineering, Computer Science, Immunology, etc.)
    - Co-author of 4 preliminary Innovation Disclosures with DFCI. Prior IP works: 6 Patents (2 now used at Brigham, 1 used in another hospital, 2 licensed)
    - Have been reviewer for various journals (by Nature Publishing Group, Cell Press, IEEE, ACM, Oxford Press, etc.)
    - Have been manager for a group of machine learning professionals that has 80,000+ members
    - Co-Chair for Medical AI at MLCommons – an organization whose sole objective is benchmarking AI and assess "SOTA" across Industries



# Artificial Intelligence Operations & Data Science Services group



TLDR; Professional Services in AI & Machine Learning

## Mission:

- Bridge the gap between Research and the Clinic by designing, implementing, and deploying artificial intelligence (AI) and data science solutions for the Institute
- Assist DFCI faculty by providing customized AI and data science support in laboratories, centers, and departments

## Primary Services and Offering:

- Artificial intelligence, data science, machine learning (ML), automated ML, multi-modal ML, self-supervised learning, federated learning
- **Natural language processing** & natural language understanding (e.g., text as main input data modality)
- **Computer vision** on image data from **pathology, radiology, and radiation oncology** (e.g., pathology slide images such as H&Es, whole-slide images, tissue-micro arrays, and radiology/radiation oncology imaging studies such as MRI, CT and other imaging modalities)
- **Machine learning operations (MLOps)** and ML production deployment (i.e., large scale, reproducible, and hybrid/multi-cloud deployments) for operationalization in the clinic and in the Institute
- Cloud innovation & AI strategy
- AI & machine learning enablement: building the data pipelines, software libraries, and data pre-processing tools, that aim at lowering the barrier of entrance for researchers who want to **transition from "X" to "AI-powered X"** in the context of cancer research
- Client management (in partnership with other I&A client service areas)





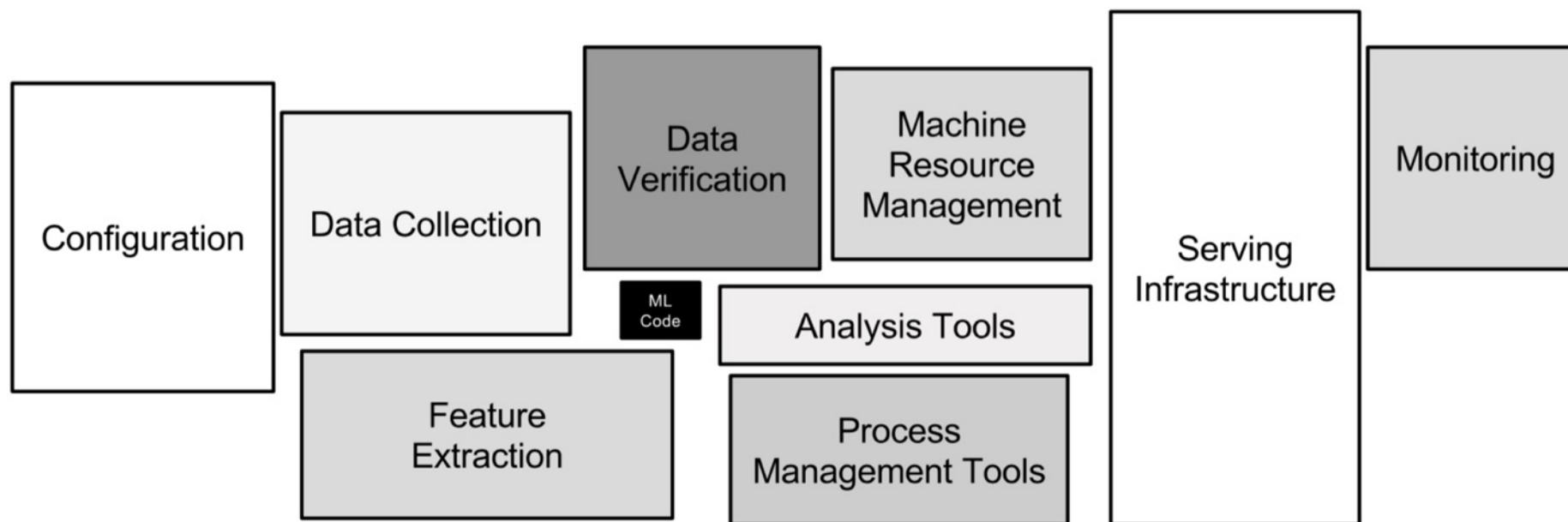
**“I’m applying AI to pathology ...  
I’m going to write a ton of ML Code!”**

**ML  
Code**



“Turns out AI was a tiny percentage of the work...”

## Machine learning systems in the real world

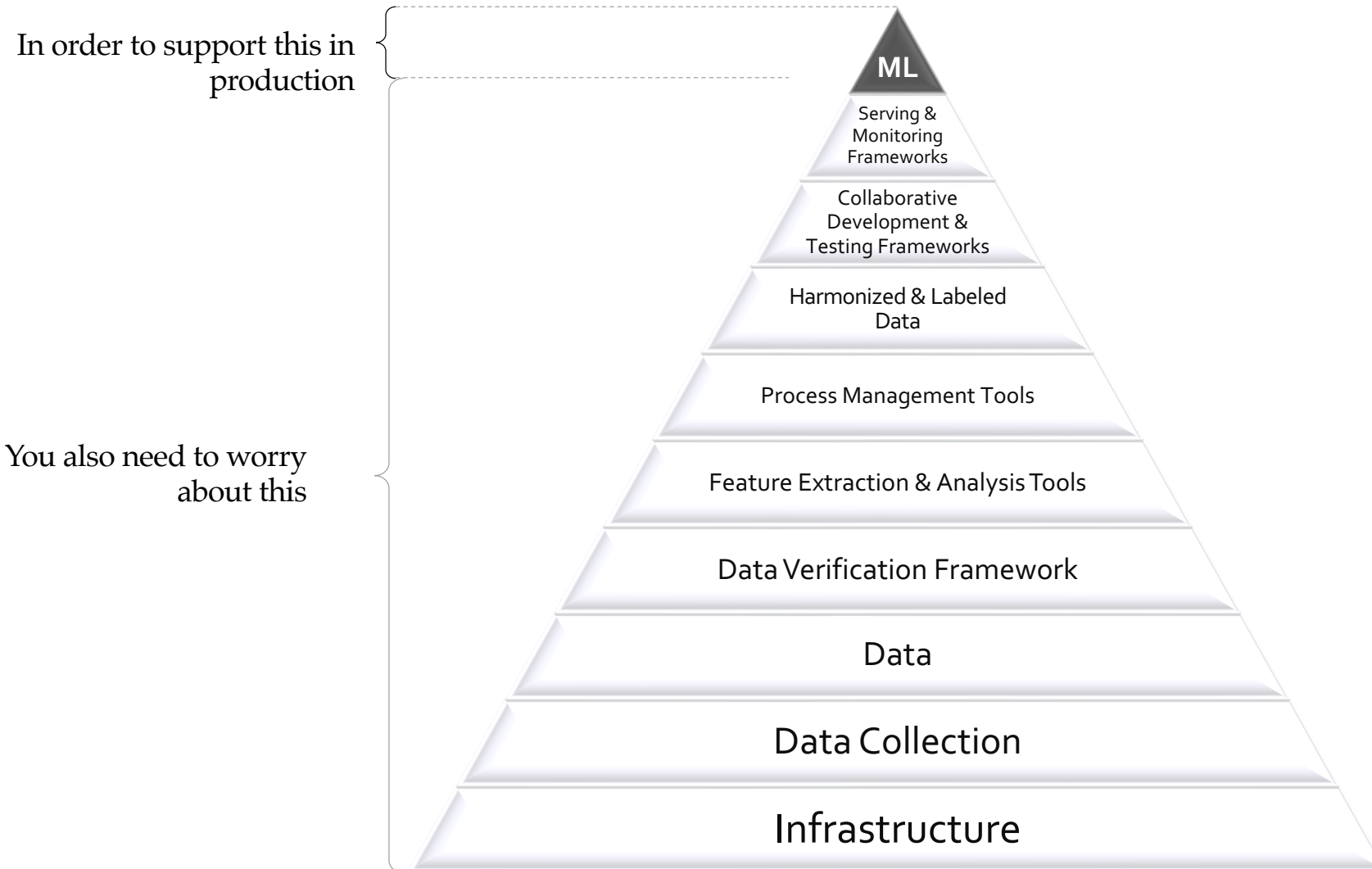


Only a small fraction of real-world machine learning systems actually constitutes machine learning code.

# ML Dependencies



# ML Dependencies



# AI in pathology



You in the near future, after using AI to solve a biologically relevant problem in the pathology realm



You today





# AI in pathology: data engineering was the biggest obstacle we faced, before AI design could even start!



# Data Eng.

You in the near future, after using AI to solve a biologically relevant problem in the pathology realm



You today

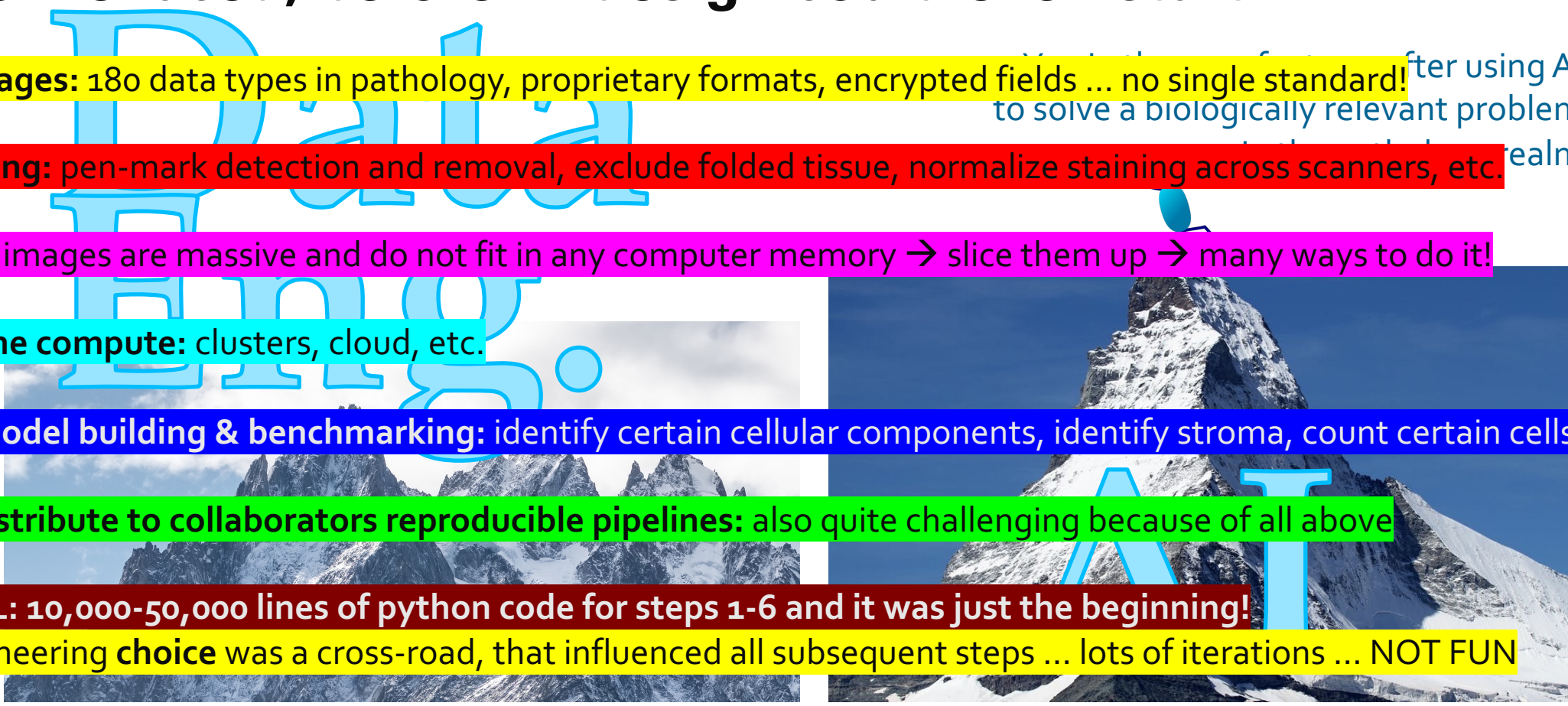


timeline



# AI in pathology: data engineering was the biggest obstacle we faced, before AI design could even start!

- 1. Load the images:** 180 data types in pathology, proprietary formats, encrypted fields ... no single standard!
  - 2. Pre-processing:** pen-mark detection and removal, exclude folded tissue, normalize staining across scanners, etc.
  - 3. Tiling:** these images are massive and do not fit in any computer memory → slice them up → many ways to do it!
  - 4. Distribute the compute:** clusters, cloud, etc.
  - 5. Iterate on model building & benchmarking:** identify certain cellular components, identify stroma, count certain cells, etc.
  - 6. Build and distribute to collaborators reproducible pipelines:** also quite challenging because of all above
- GRAND TOTAL: 10,000-50,000 lines of python code for steps 1-6 and it was just the beginning!**
- Each data engineering choice was a cross-road, that influenced all subsequent steps ... lots of iterations ... NOT FUN



You today



# AI in pathology: **data engineering** was the biggest obstacle we faced, before AI design could even start!



1. **Load the images:** 180 data types in pathology, proprietary formats, encrypted fields ... no single standard!
  2. **Pre-processing:** pen-mark detection and removal, exclude folded tissue, normalize staining across scanners, etc.
  3. **Tiling:** these images are massive and do not fit in any computer memory → slice them up → many ways to do it!
  4. **Distribute the compute:** clusters, cloud, etc.
  5. **Iterate on model building & benchmarking:** identify certain cellular components, identify stroma, count certain cells, etc.
  6. **Build and distribute to collaborators reproducible pipelines:** also quite challenging because of all above
- GRAND TOTAL:** 10,000-50,000 lines of python code for steps 1-6 and it was just the beginning!
- Each data engineering **choice** was a cross-road, that influenced all subsequent steps ... lots of iterations ... NOT FUN

You today

timeline

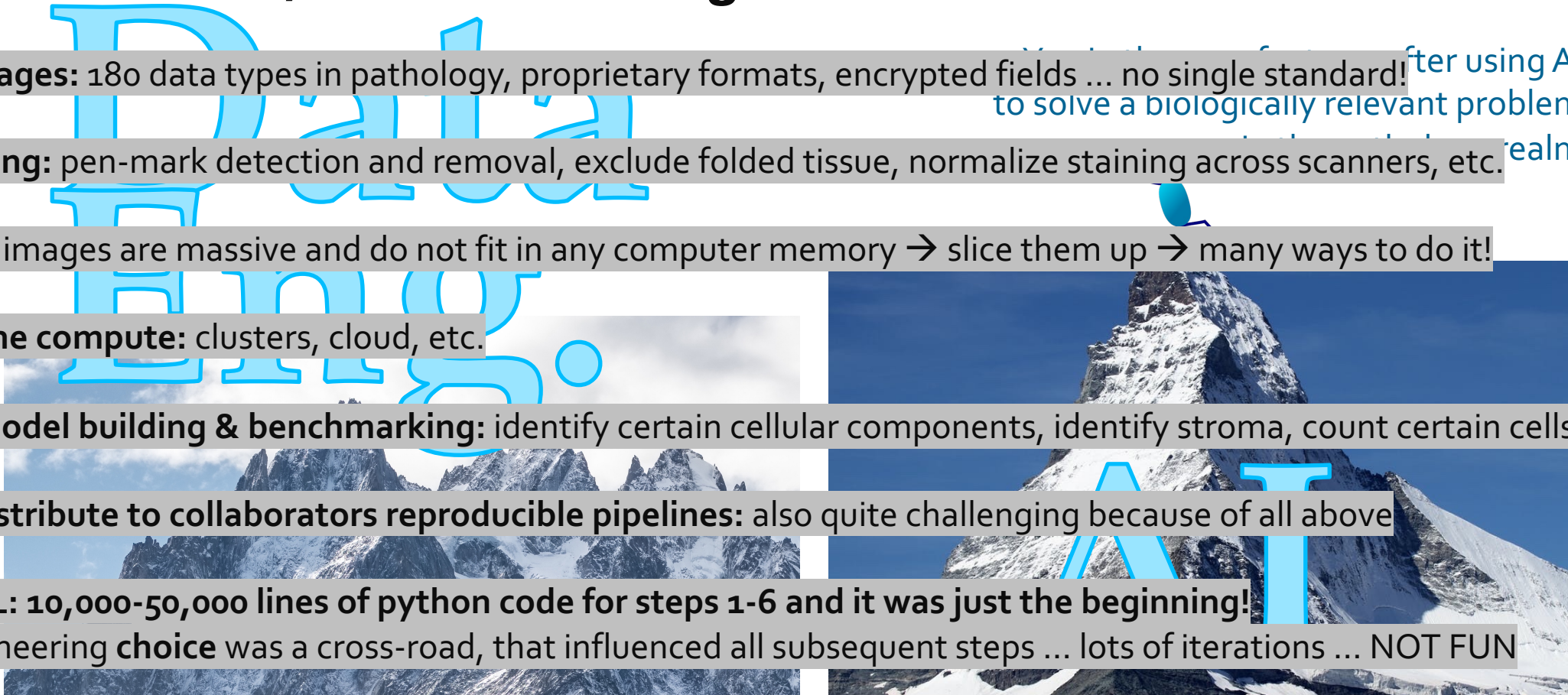
We realized we needed a low-code software library to solve these at scale and reliably for all projects!



# AI in pathology: data engineering was the biggest obstacle we faced, before AI design could even start!



1. **Load the images:** 180 data types in pathology, proprietary formats, encrypted fields ... no single standard!
  2. **Pre-processing:** pen-mark detection and removal, exclude folded tissue, normalize staining across scanners, etc.
  3. **Tiling:** these images are massive and do not fit in any computer memory → slice them up → many ways to do it!
  4. **Distribute the compute:** clusters, cloud, etc.
  5. **Iterate on model building & benchmarking:** identify certain cellular components, identify stroma, count certain cells, etc.
  6. **Build and distribute to collaborators reproducible pipelines:** also quite challenging because of all above
- GRAND TOTAL:** 10,000-50,000 lines of python code for steps 1-6 and it was just the beginning!  
 Each data engineering choice was a cross-road, that influenced all subsequent steps ... lots of iterations ... NOT FUN



You today



We realized we needed a low-code software library to solve these at scale and reliably for all projects!



# PathML

 (started in 2019)

# PathML in a nutshell

The logo for PathML features the word "Path" in a purple, textured font, followed by "ML" in a white, outlined font with a purple glow.

- Python software library
- Released under an open-source license
- Aimed at supporting end-to-end digital pathology work-flows
- It's a "Low-code" type of software library
  
- **Designed by pathologists**
- **Implemented by computer scientists & data scientists**
  
- Makes industry-standard choices for you
- Uses best-in-class algorithms
- Highly modular and extensible
- Let's you **focus on the research** and alleviates the data engineering headaches
  1. e.g., a 20k lines of code processing pipeline in python → **30 lines of code using PathML**
  2. e.g., a CODEX pipeline that used to take 2 days and required 100+ clicks → **1 click and 2h**
  3. e.g., another cell identification pipeline of 10k lines of code → **15 lines using PathML**



# “Big Data” in Pathology

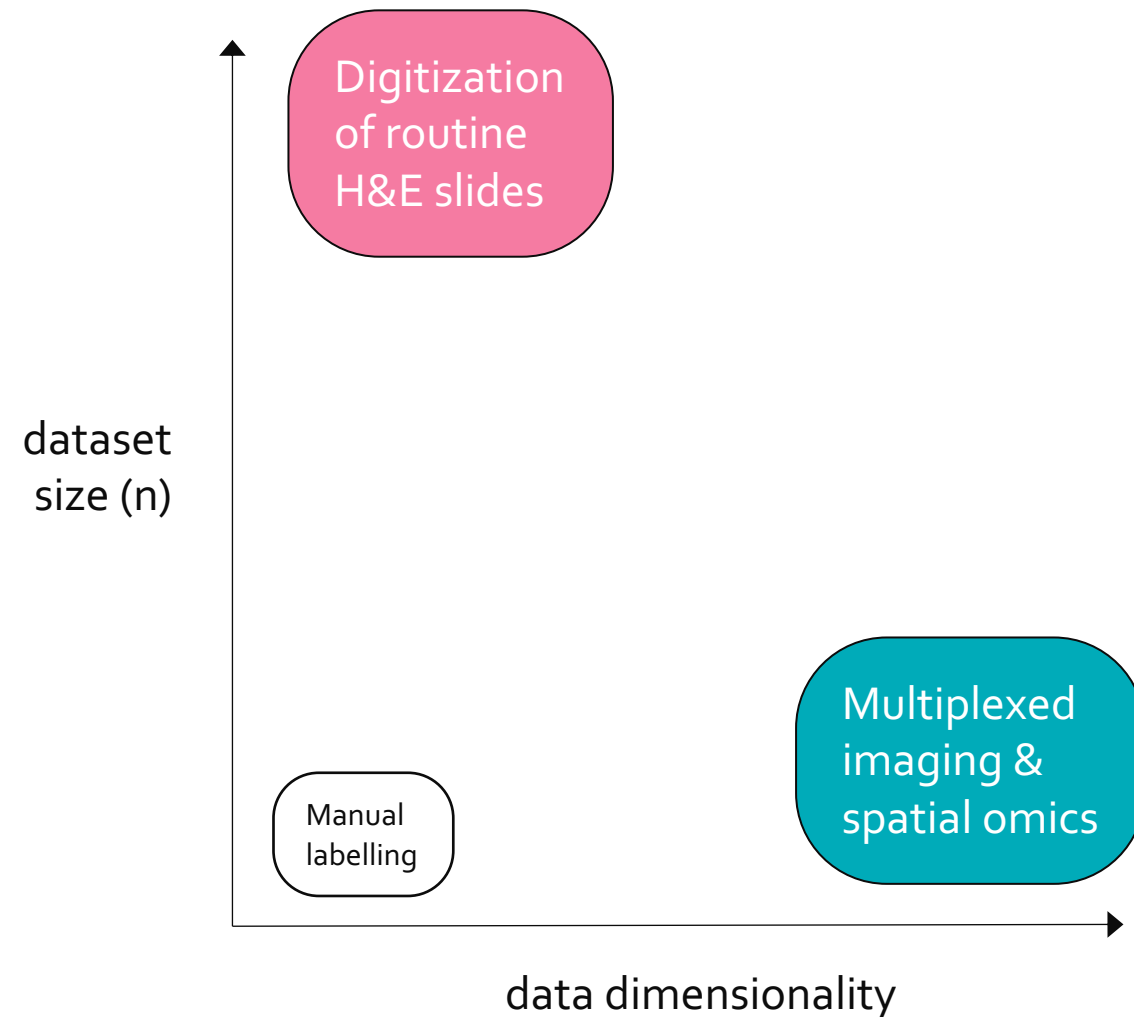
- As pathology goes digital, the size of available data will increase drastically
- Datasets get bigger along at least two axes: dimensionality and dataset size
- Technical challenges drive requirements:

## Large N

- Cloud infrastructure
- Distributed computing (e.g. kubernetes)
- Unsupervised or weakly-supervised learning

## High-Dimensional

- Spatial analysis methods
- Interface with single-cell analysis ecosystem
- Visualization tools
- Choosing markers



- Digital Pathology strategy must scale in **both dimensions** to maximize innovation and patient impact

# Computational Pathology Tools are Lacking

Robust ecosystem of open-source tools for general purpose machine learning and computer vision

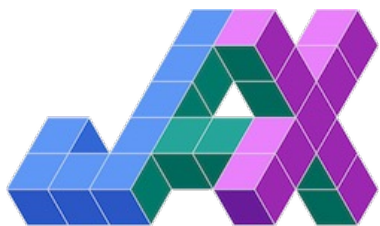
 PyTorch

 TensorFlow

 OpenCV

 OME

 Keras



scikit-image  
image processing in python

 OpenSlide

But these tools are lacking...

1. They assume small natural images (ours are gigapixel!)
2. Minimal support for 'spatial omics' or multiparametric imaging
3. There are few pathology-specific algorithms, and development is difficult

# Analysis Infrastructure

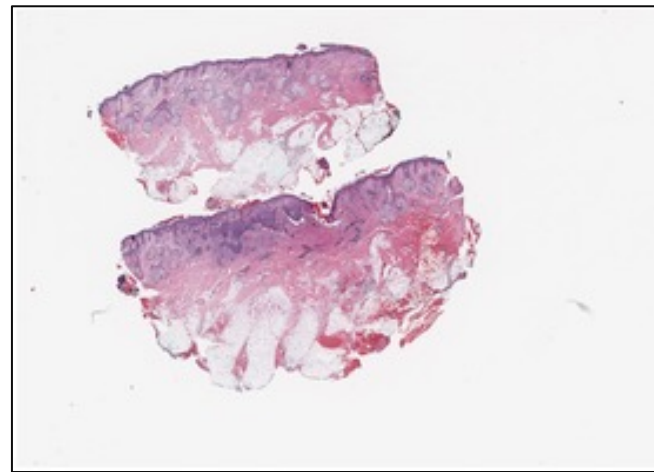
- Data alone is not enough - we also need to provide our researchers with the tools to enable them to analyze it

## Requirements:

- Easy to use for researchers (low barrier to entry)
  - Scale in both axes: large  $n$ , and data dimensionality
  - Support for all commonly used data types and file formats in pathology
  - Support for domain-specific functionality
  - Promote standardized, reproducible workflows
  - Integrate with industry-standard machine learning and analysis tools (e.g. PyTorch, Scanpy)
- There are many vendor solutions and general-purpose machine learning tools, but none that satisfied all of our requirements
  - **So, we built our own: PathML (development began in 2019)**

# Gigapixel Scale Images

- Image size presents a technical challenge
- Whole-slide images are too big to feed into neural network directly



← 100,000 px  
typical WSI →



28 px  
MNIST



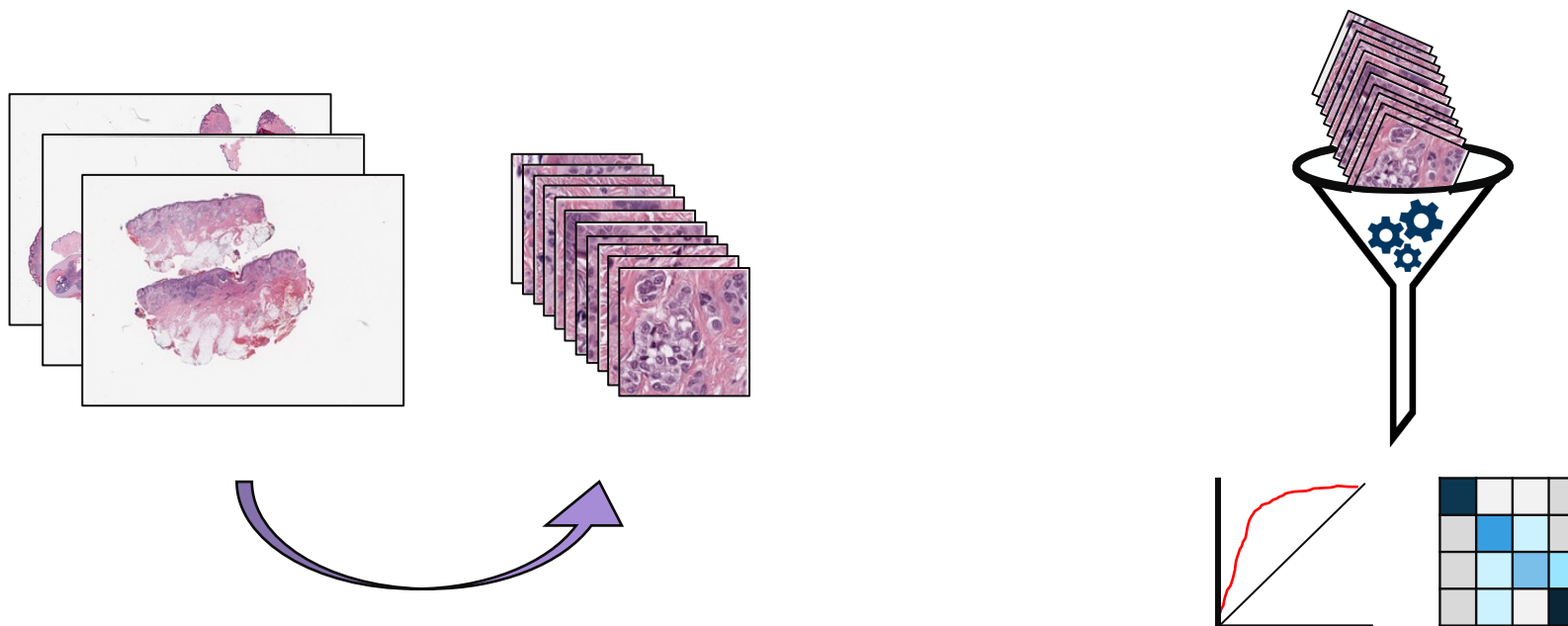
32 px  
CIFAR-10

WSIs are **4-8 orders of magnitude larger**  
compared to images in many other domains  
(number of pixels)

- Similar story for highly-multiplexed immunofluorescence images

# Preprocessing

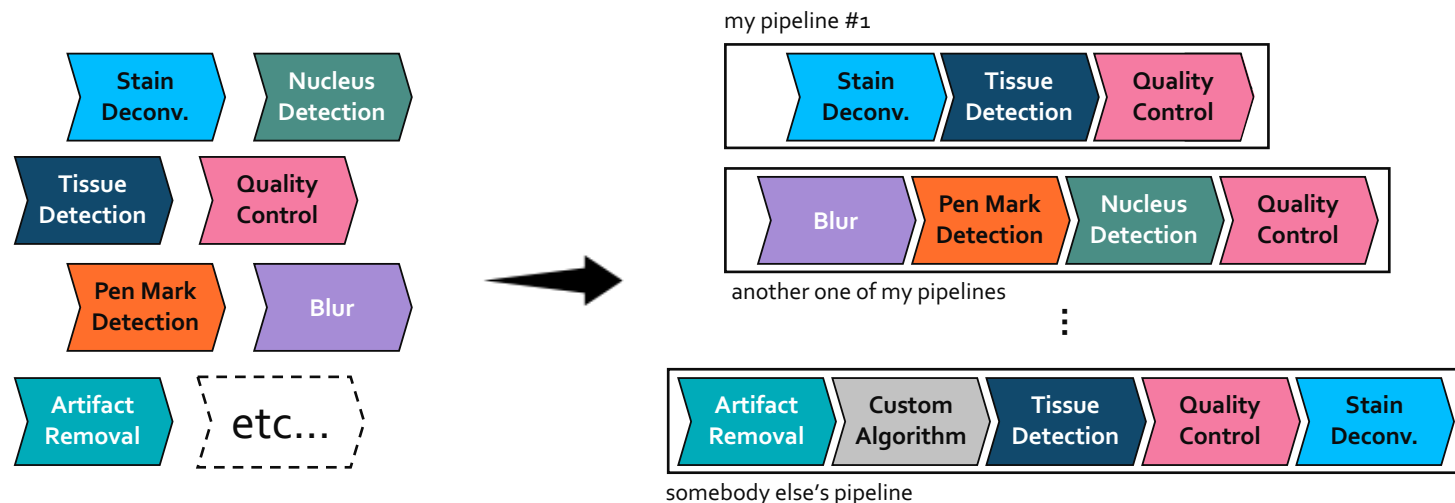
- To overcome the size problem, whole-slide images are typically preprocessed and divided into tiles
- Tiles are then fed into downstream analysis (e.g. neural networks)



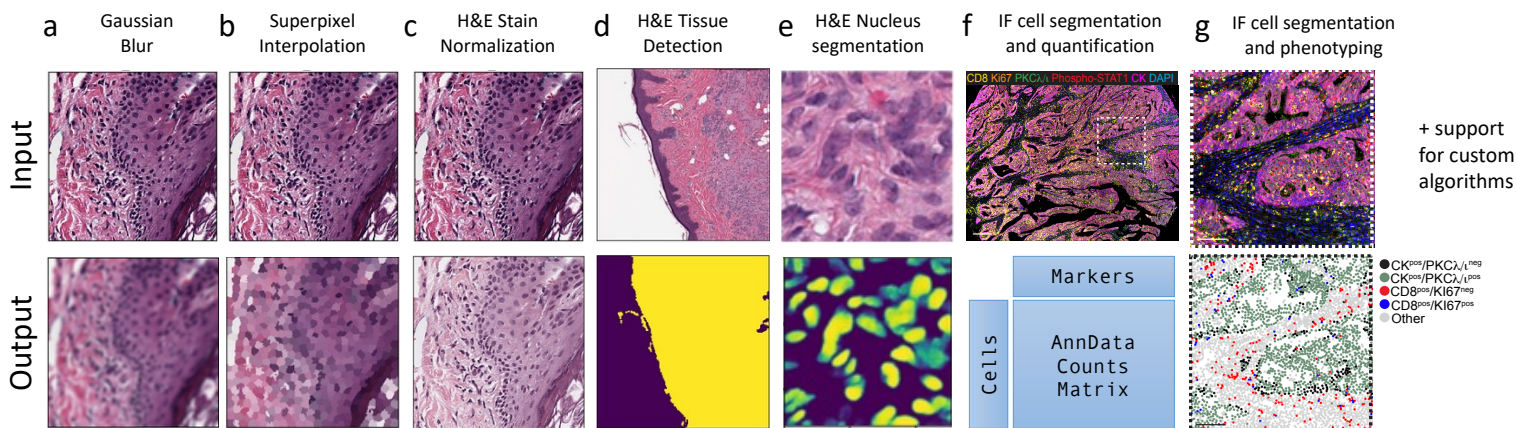


# Designing Preprocessing Pipelines

- Preprocessing pipelines operate at tile-level
- Defined as sequential application of modular transformations
- Mix-and-match custom operations with pre-made
- General framework can be applied to any imaging modality

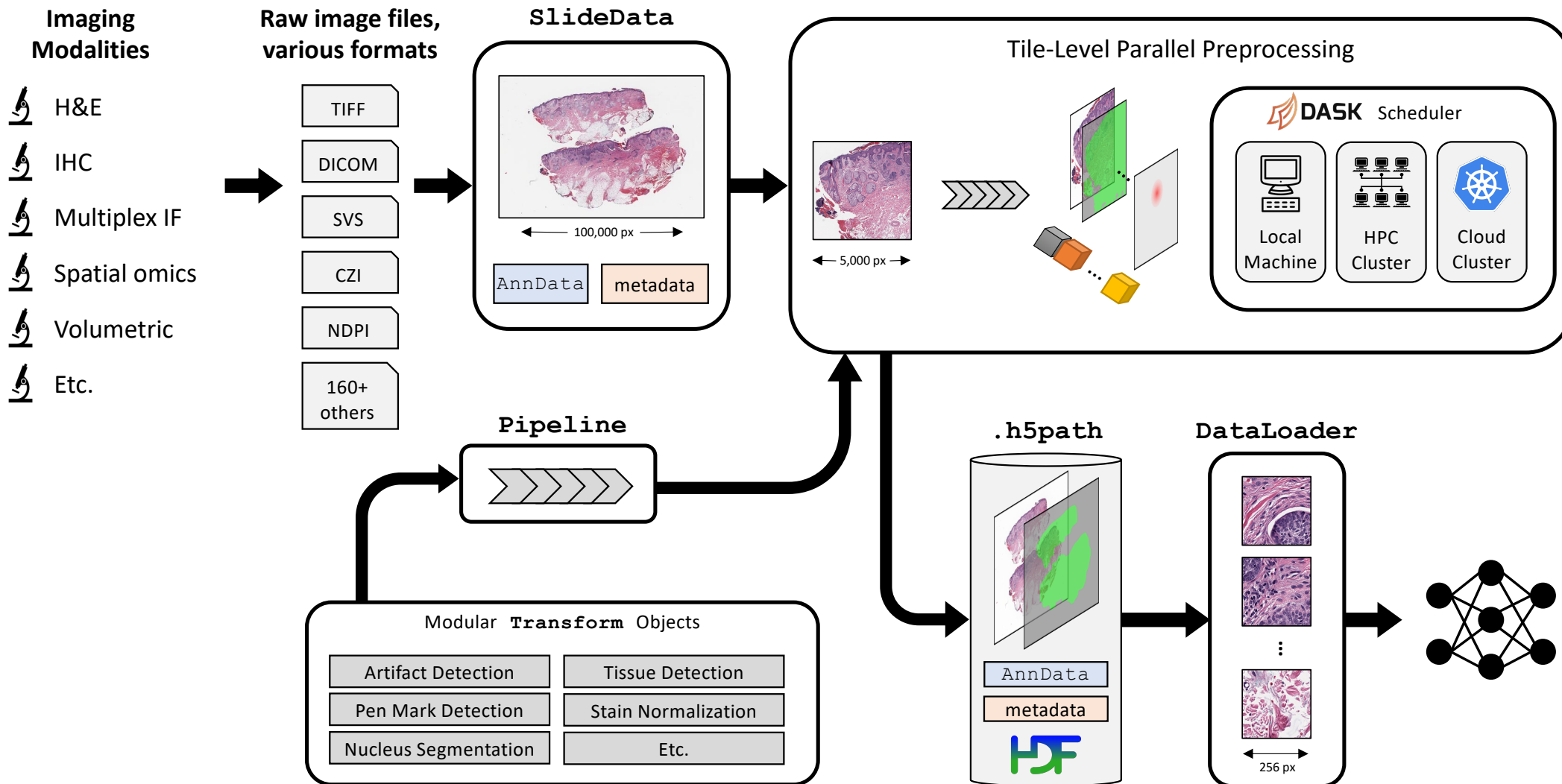


Schematic of mix-and-match pipeline construction



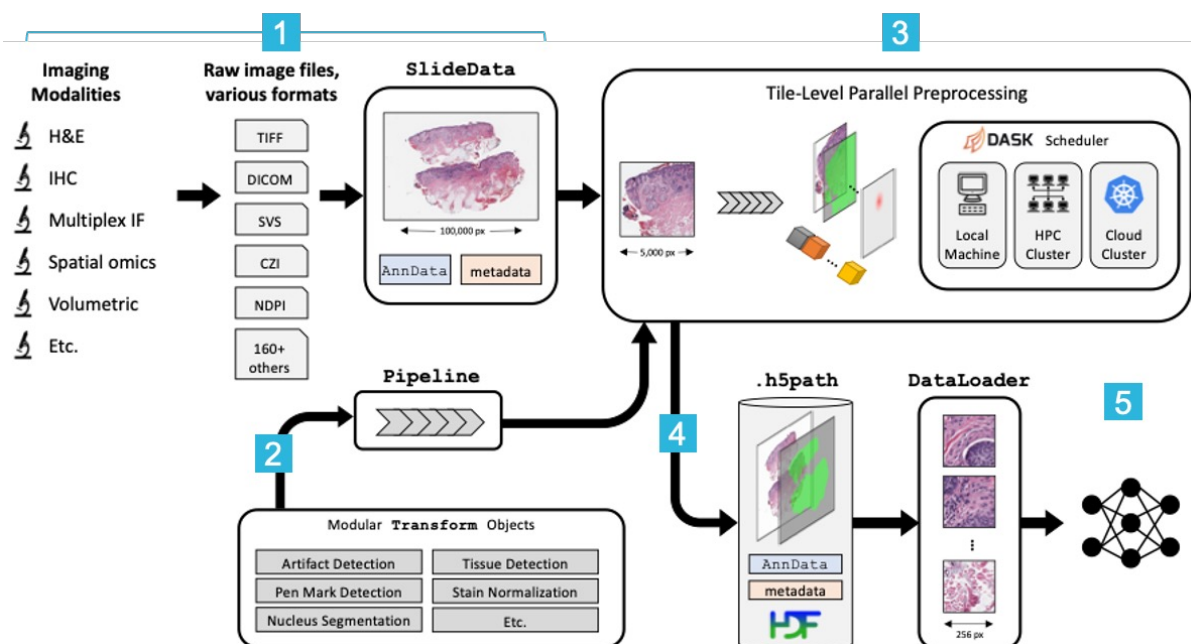
Examples of some transformations that come pre-made in PathML

# Putting it All Together: PathML Preprocessing Framework



# Streamlined Analysis Workflows

- Complete end-to-end pipelines in ~10 lines of code
- Lower barrier to entry for image analysis research
- Enables rapid prototyping/development
- Built-in support for HPC and cloud clusters



```

1 # load the image
slide = CODEXSlide("/path/to/image.ome.tif")

# Define a pipeline
pipe = Pipeline([
    CollapseRunsCODEX(z = 0),
    SegmentMIF(model = "mesmer",
               nuclear_channel = 0,
               cytoplasm_channel = 11,
               image_resolution = 0.5),
    QuantifyMIF("cell_segmentation")
])

3 # run pipeline with distributed computing
slide.run(pipe, tile_size = 1024)

4 # Save output
slide.write("/path/to/image.h5path")

5 # PyTorch DataLoader
dataset = TileDataset("/path/to/image.h5path")
dataloader = DataLoader(dataset, batch_size = 16)

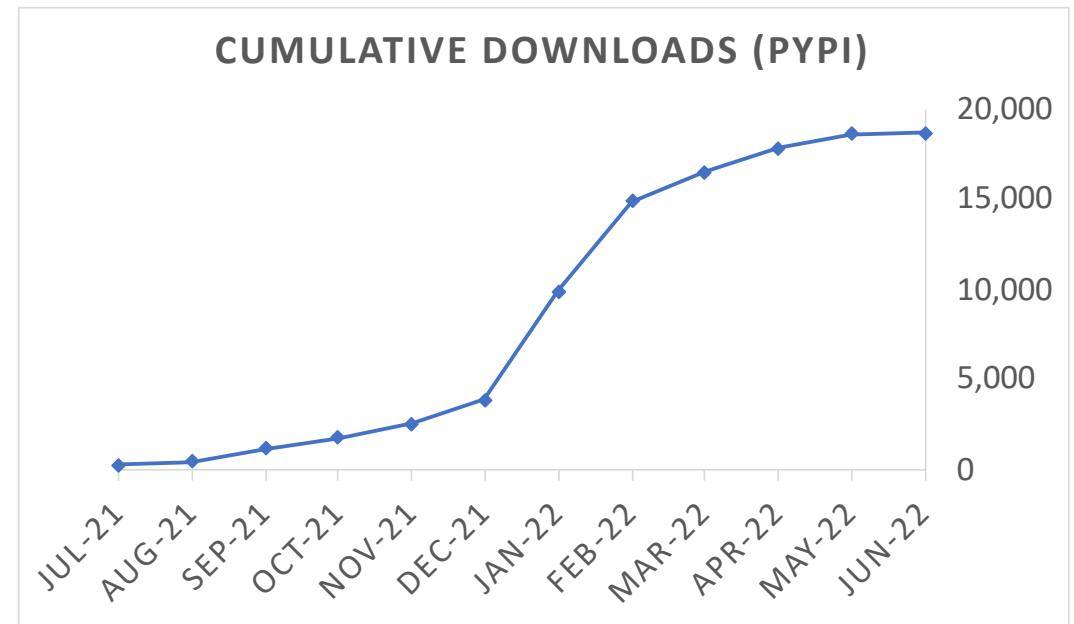
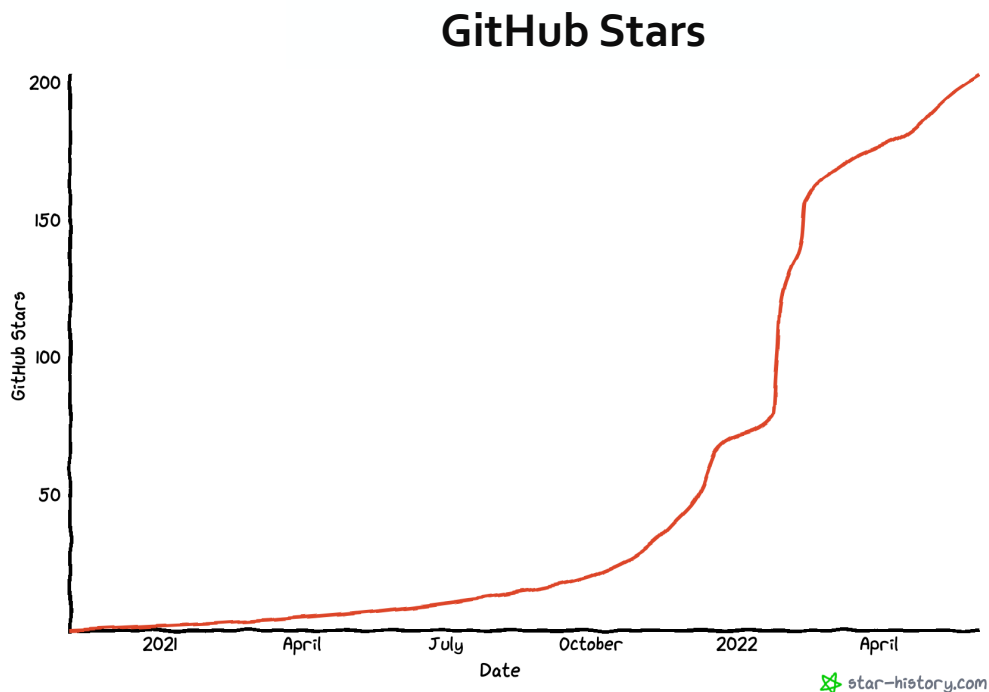
```

Full code for analysis of mIF images (including ML-powered cell segmentation, marker quantification, PyTorch DataLoader)

# Users Today (June 2022)

- Research community (direct collaborators)
  - 5+ labs/groups (DFCI + Cornell)
  - 2 imaging core facilities
  - 2 institutions

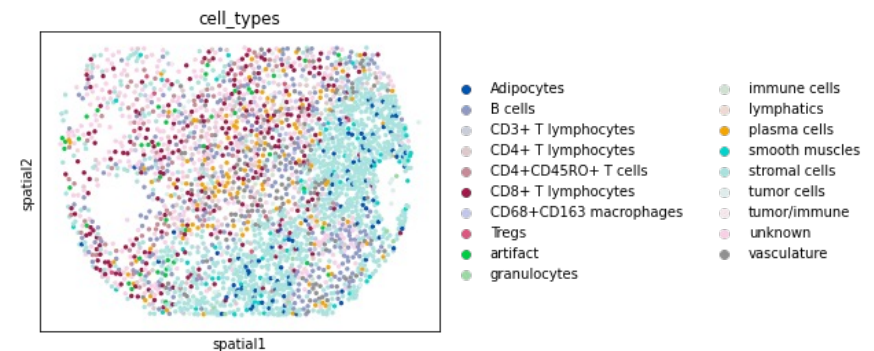
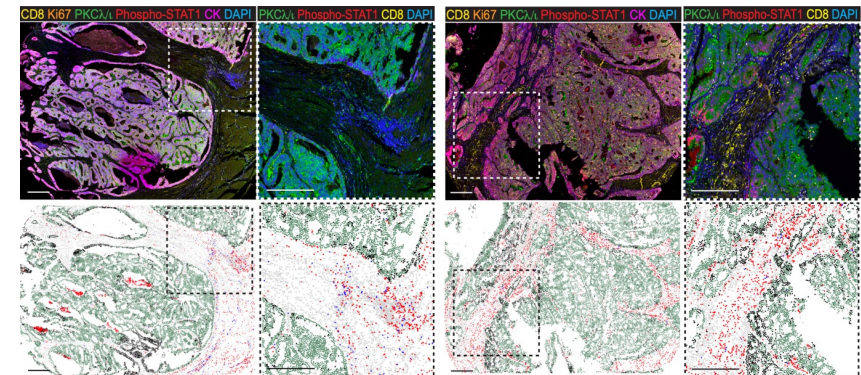
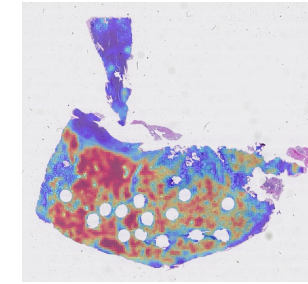
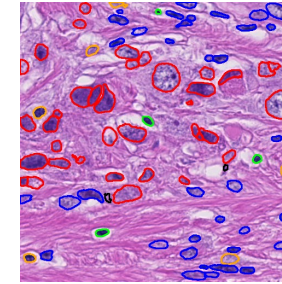
- Open-source community
  - 18,000+ downloads from PyPI
  - 200+ stars on GitHub
  - Users around the globe (Germany, China, Brazil, UK, ...)
  - Part of a growing ecosystem of tools





# How PathML is Being Used Today

- Prostate cancer research, H&E images
  - Multiple instance learning
  - Cell segmentation and classification
- Colorectal cancer research, 7-channel IF images
  - Cell segmentation and rules-based phenotyping
  - Spatial biology
- Production image quantification pipelines, DFCI core facility, CODEX spatial proteomics
  - Improving operational efficiency of key institutional resource





# Conclusions

- PathML toolkit provides a framework to build modular, fully-customizable preprocessing pipelines for gigapixel-scale images
- Unified API for H&E, IHC, multiplex fluorescence, spatial omics, etc.
- Support for 160+ file formats, many different imaging modalities
- Fast dataloaders for integrating with the broader ML ecosystem (PyTorch, Tensorflow, Jax, etc.)
- Integration with broader single-cell analysis ecosystem (Scanpy, Squidpy, etc.) via AnnData standard
- Streamlined, fully-documented workflows lower barrier to entry
- PathML is being used across a number of labs, cores, and institutions for a wide variety of projects

# Thank You

Any Questions?

## DFCI I&A

- Bryan Gass
- Sreekar Reddy Puchala
- Ella Halbert
- Xiaoxuan Liu
- Haoyuan Li
- Jie Sun
- Daniel Waranch
- AIOS Group
- Renato Umeton
- Jason Johnson

## DFCI Med Onc

- Jackson Nyman
- Surya Hari
- Eli Van Allen

## Weill Cornell Pathology

- Ryan Carelli
- Mohamed Omar
- David Brundage
- Karen Xu
- Luigi Marchionni
- Massimo Loda

## DFCI Innovation Office

- Kate Strayer-Benton
- Vladimir Leopard
- Aaron Dy
- Lesley Solomon

- Interested in using PathML in your work?
- Want to contribute to development of new features?

**Get in touch!**

[PathML@dfci.harvard.edu](mailto:PathML@dfci.harvard.edu)

# PathML

- Website: <https://pathml.org>
- Documentation: <https://docs.pathml.org>
- GitHub: <https://github.com/Dana-Farber-AIOS/pathml>
- Manuscript: [doi.org/10.1158/1541-7786.MCR-21-0665.MCR-21-0665](https://doi.org/10.1158/1541-7786.MCR-21-0665.MCR-21-0665)

# Bonus slides

---

# Example: Quantitation of Multiplexed IF with PathML

---

**Molecular Cell**

**PKC $\lambda/1$  inhibition activates an ULK2-mediated interferon response to repress tumorigenesis**

Linares et al., 2021

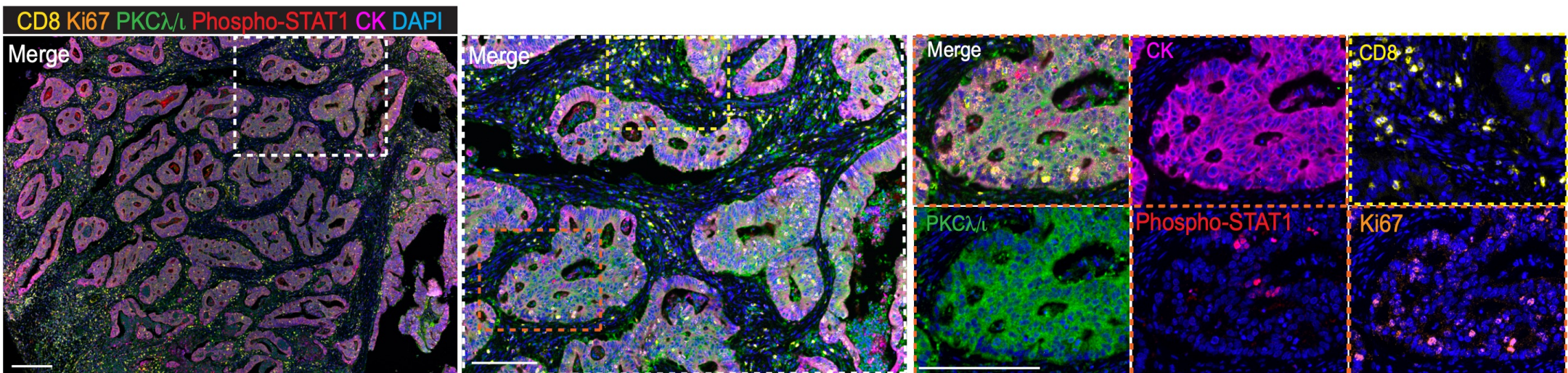
# Example: Quantitation of Multiplexed IF

## Opal Multiplexed Fluorescence IHC



## MARKERS

Nucleus	DAPI
Epithelia	CYTOKERATIN
T cells	CD8
Activation /Proliferation	Ki-67
IFN Response	Phospho-STAT1 PKC $\lambda/\iota$



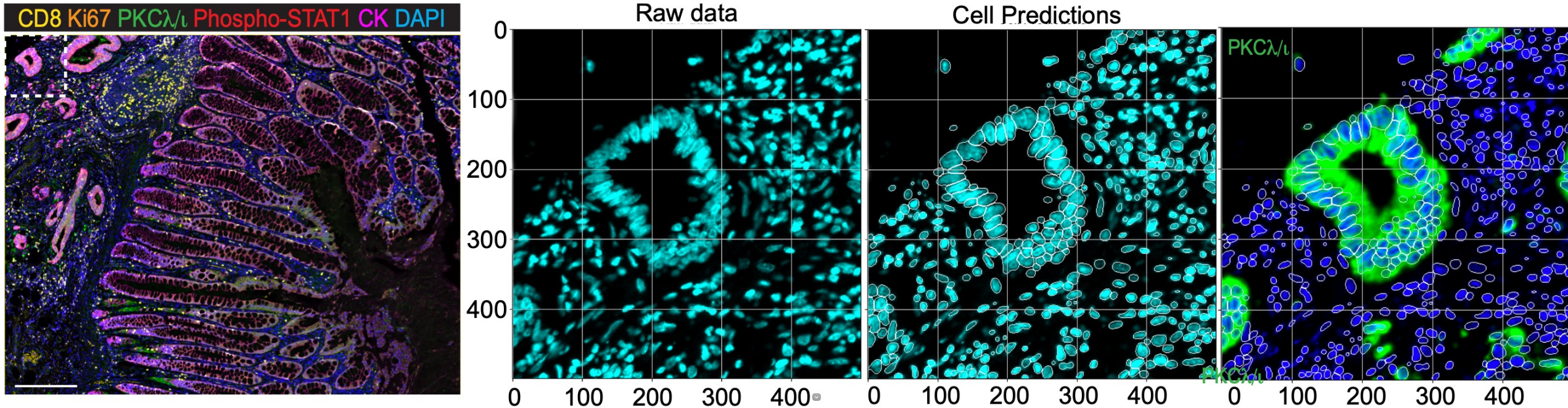
## Analysis Pipeline





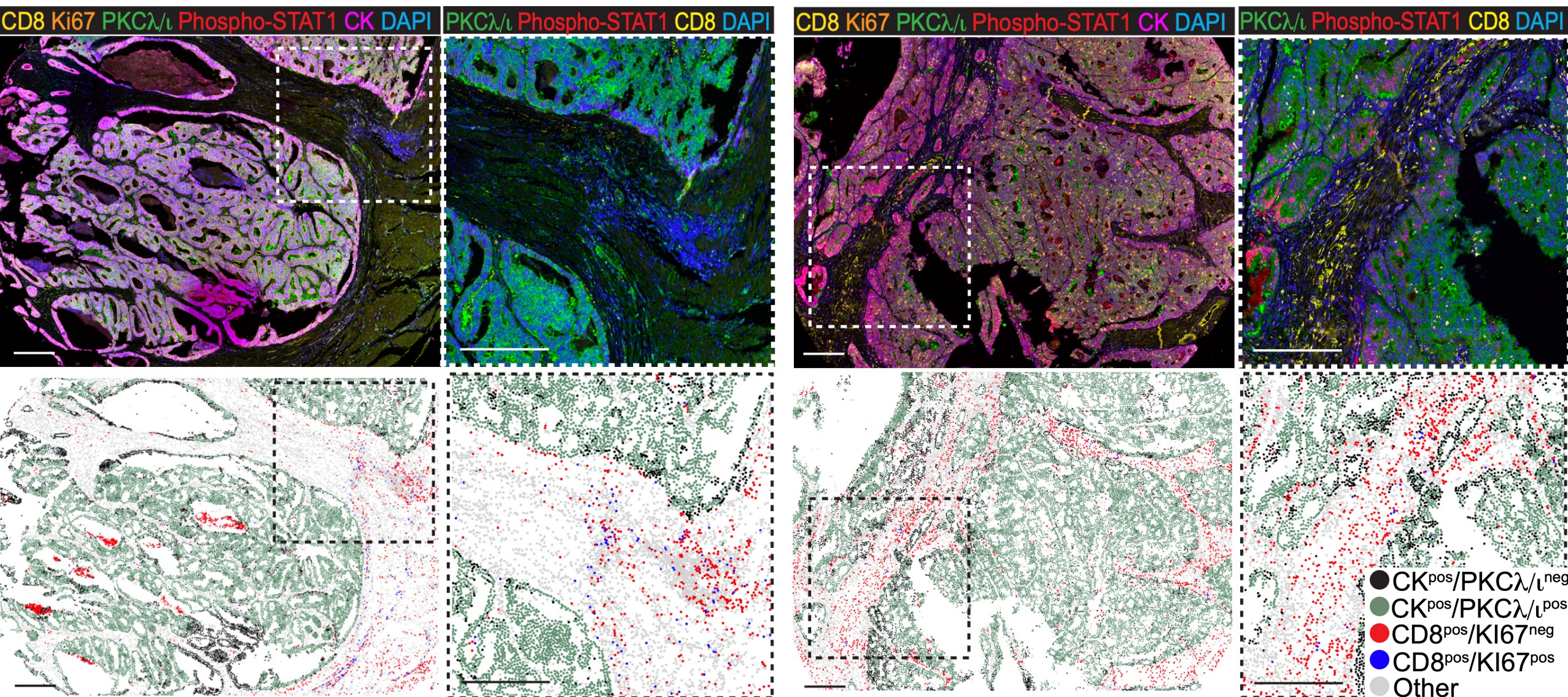
# Example: Quantitation of Multiplexed IF

- PathML implements state of the art deep learning models for segmentation, cell type identification, blur detection, visualization and more





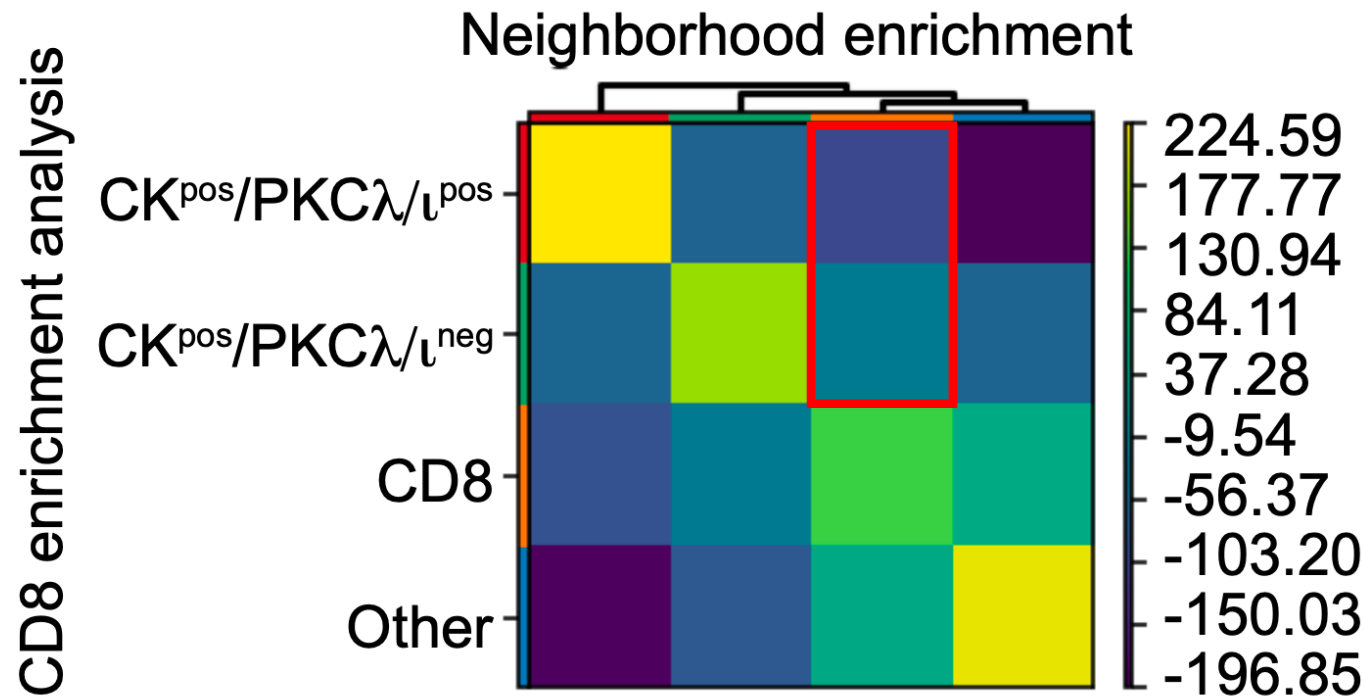
# Example: Quantitation of Multiplexed IF





# Example: Quantitation of Multiplexed IF

- After cell segmentation and phenotyping, spatial biology can be interrogated directly and quantified
- Are T cells differentially enriched in the neighborhood of PKC $\lambda$ /I negative epithelium compared to PKC $\lambda$ /I positive epithelium?

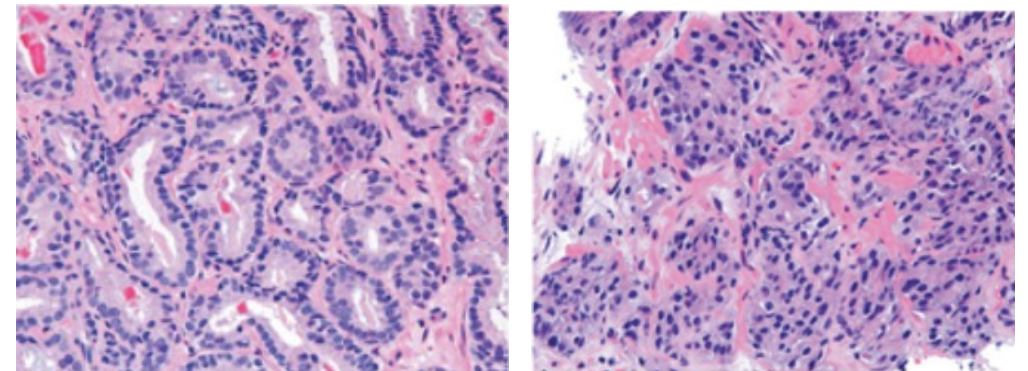




# Example: Deep Learning for H&E Nucleus Segmentation and Classification with PathML

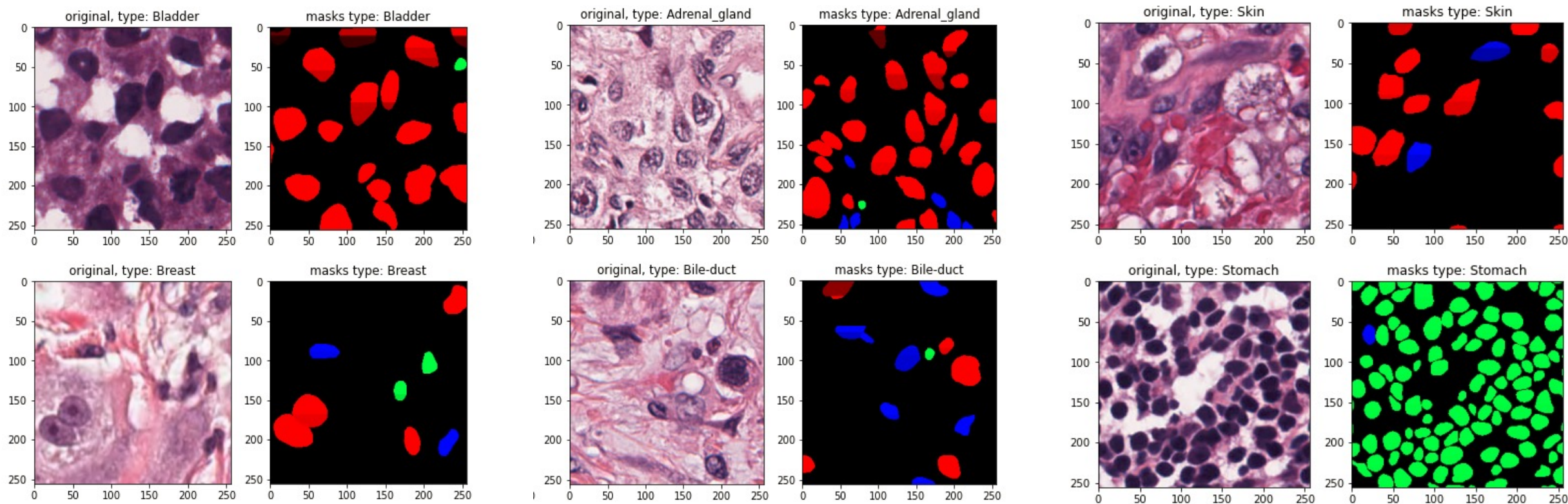
---

Work with Xiaixuan Lui, Daniel Waranch  
(HSPH)



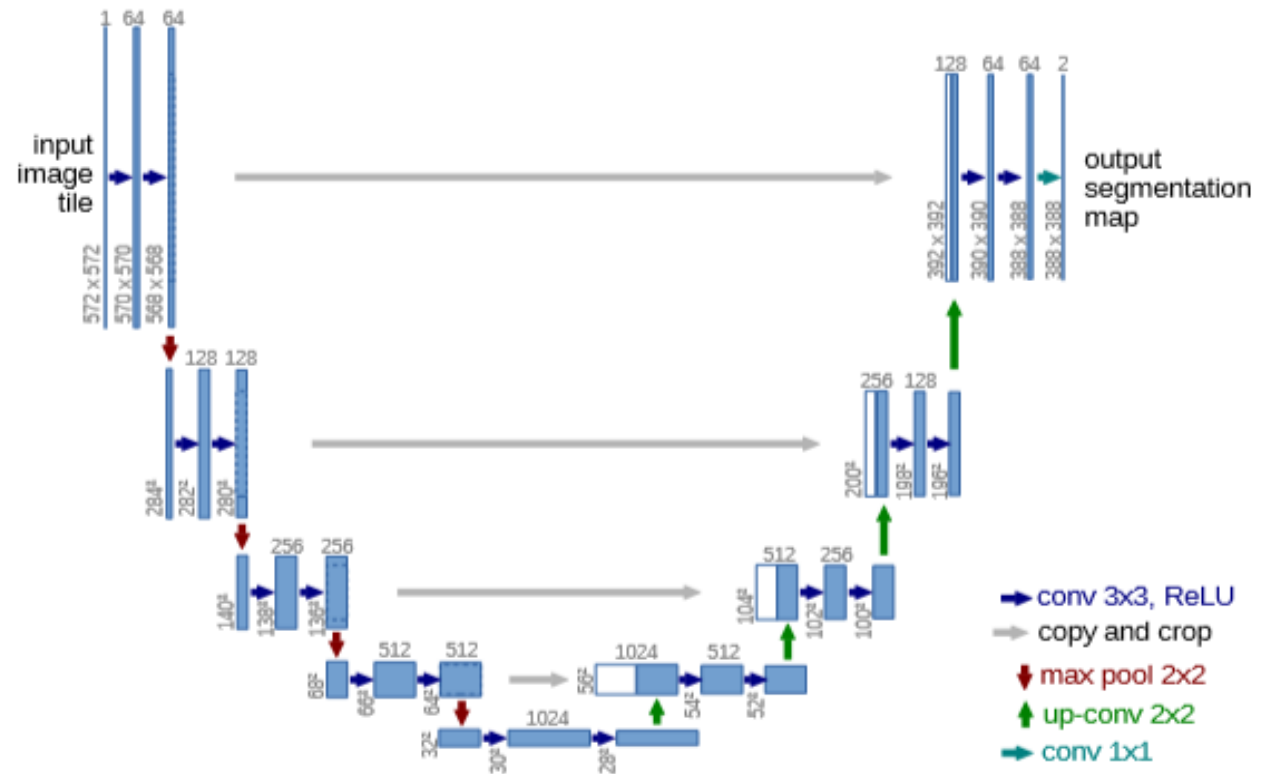
# PanNuke Dataset

- 3 folds for train, validation and test, ~2600 images with masks describing the nuclei and tissue type labels for each image.
- 19 tissue types in total
- Dataset available publicly, integrated in PathML



# Model 1: U-Net

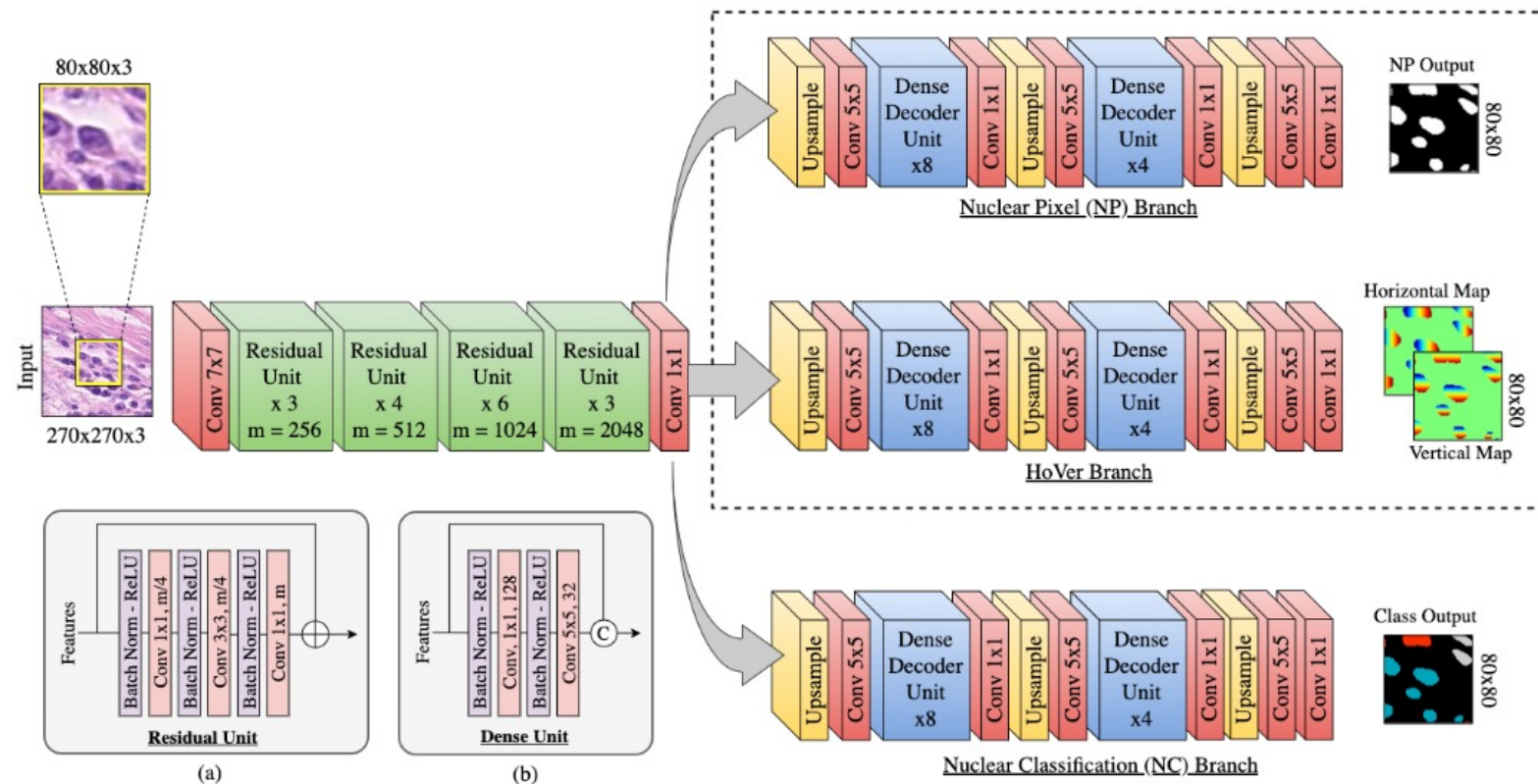
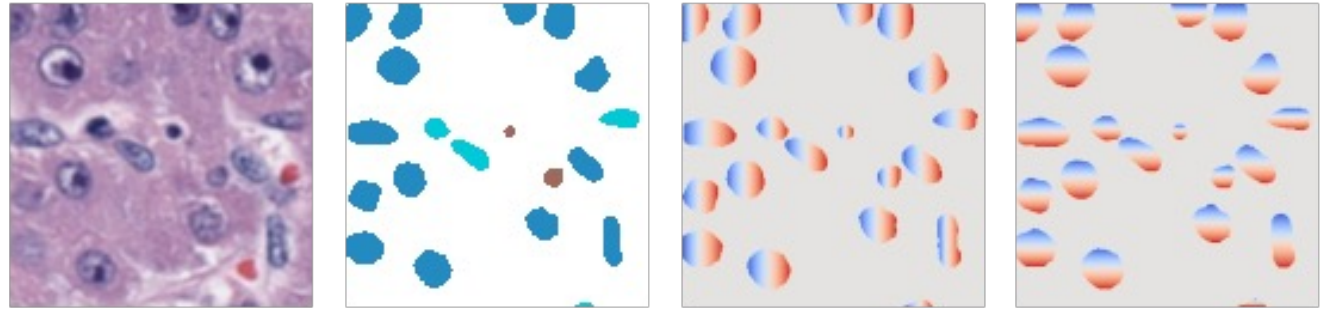
- 4-layer encoder-decoder architecture
- Each layer has two convolutional kernels followed by max-pool/upsampling
- Skip connections transmit information directly between corresponding layers in encoder and decoder





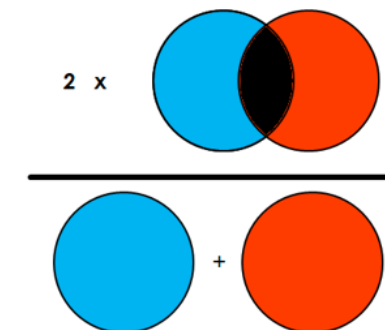
# Model 2: HoVer-Net

- Uses gradient maps to help the network with overlapping/adjacent nuclei
- One encoding branch shared by 3 decoder branches
- Simultaneous segmentation and classification

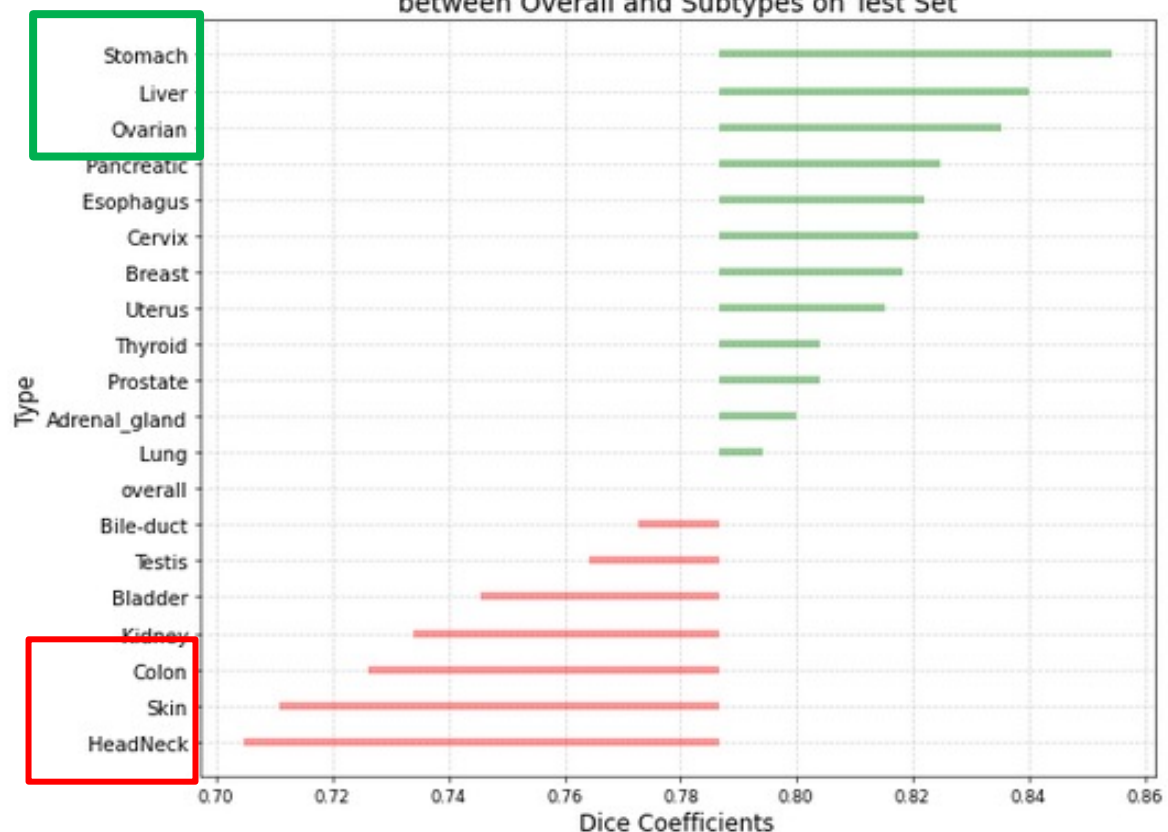


# Results: U-Net

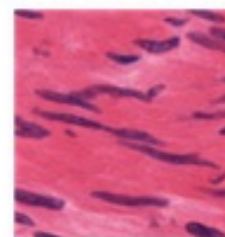
- Results vary by tissue type

$$DICE = \frac{2|X \cap Y|}{|X| + |Y|}$$


U-Net Nucleus Segmentation Dice\_coefficients' Comparison between Overall and Subtypes on Test Set



Stomach dice coef: 0.8183



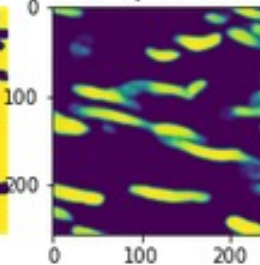
H&E

background



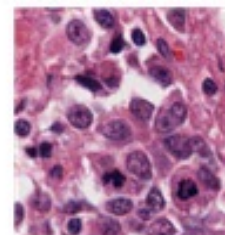
Ground Truth

test prediction

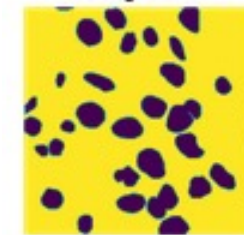


Predictions

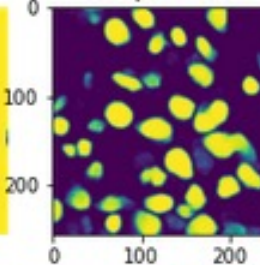
Stomach dice coef: 0.8597



background



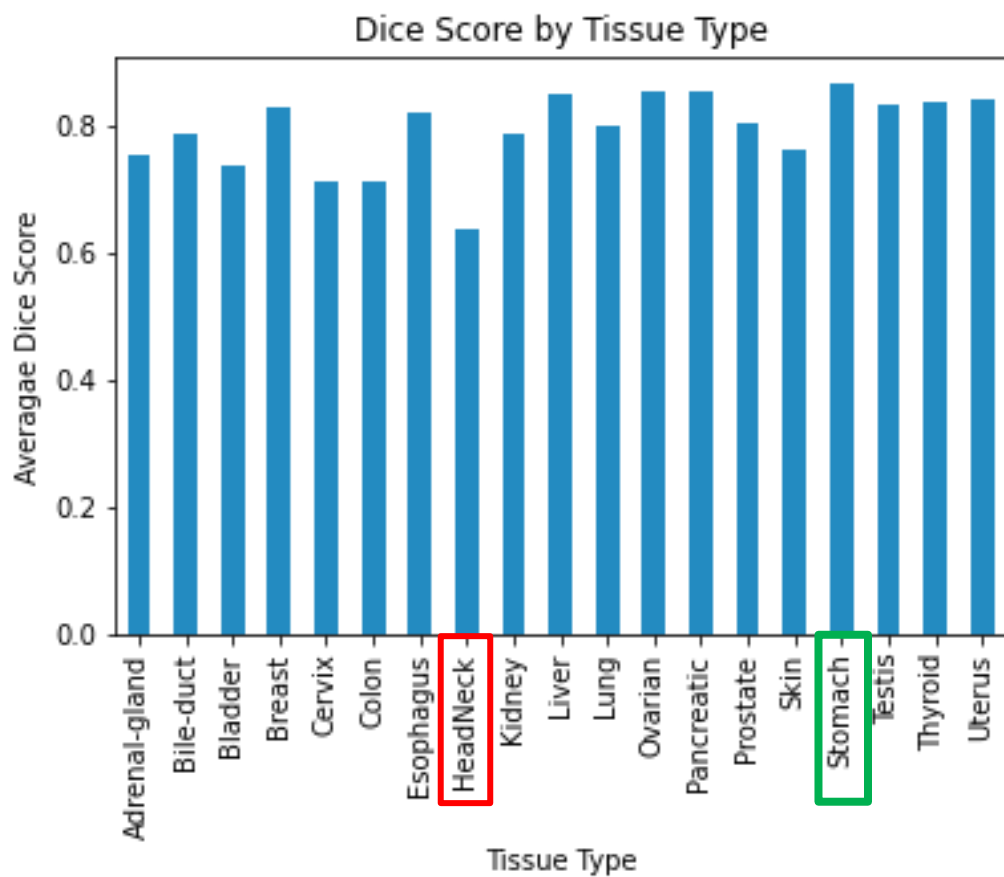
test prediction





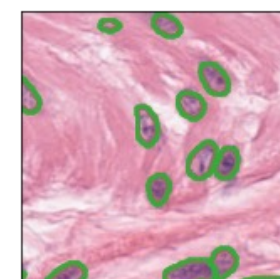
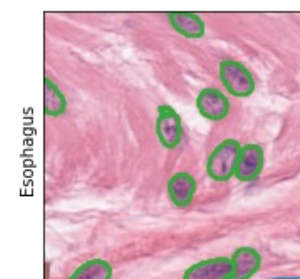
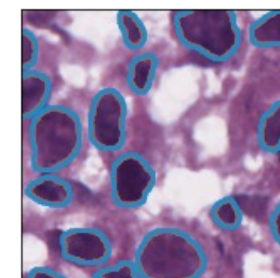
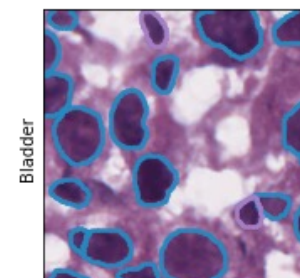
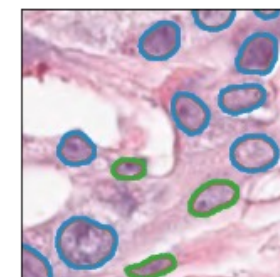
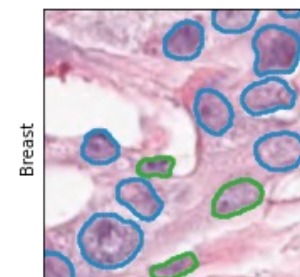
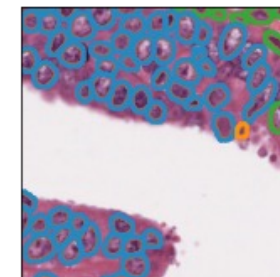
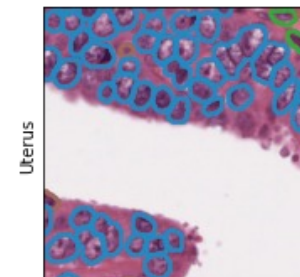
# Results: HoVer-Net

- Results are consistent with U-Net



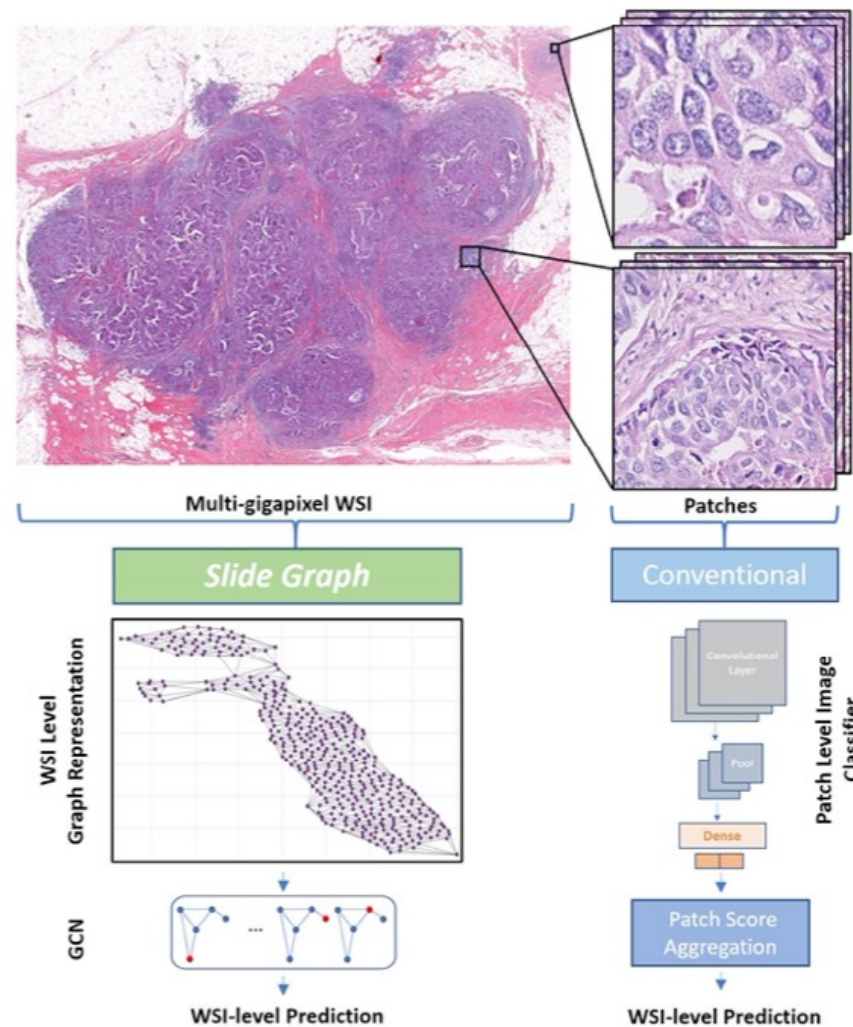
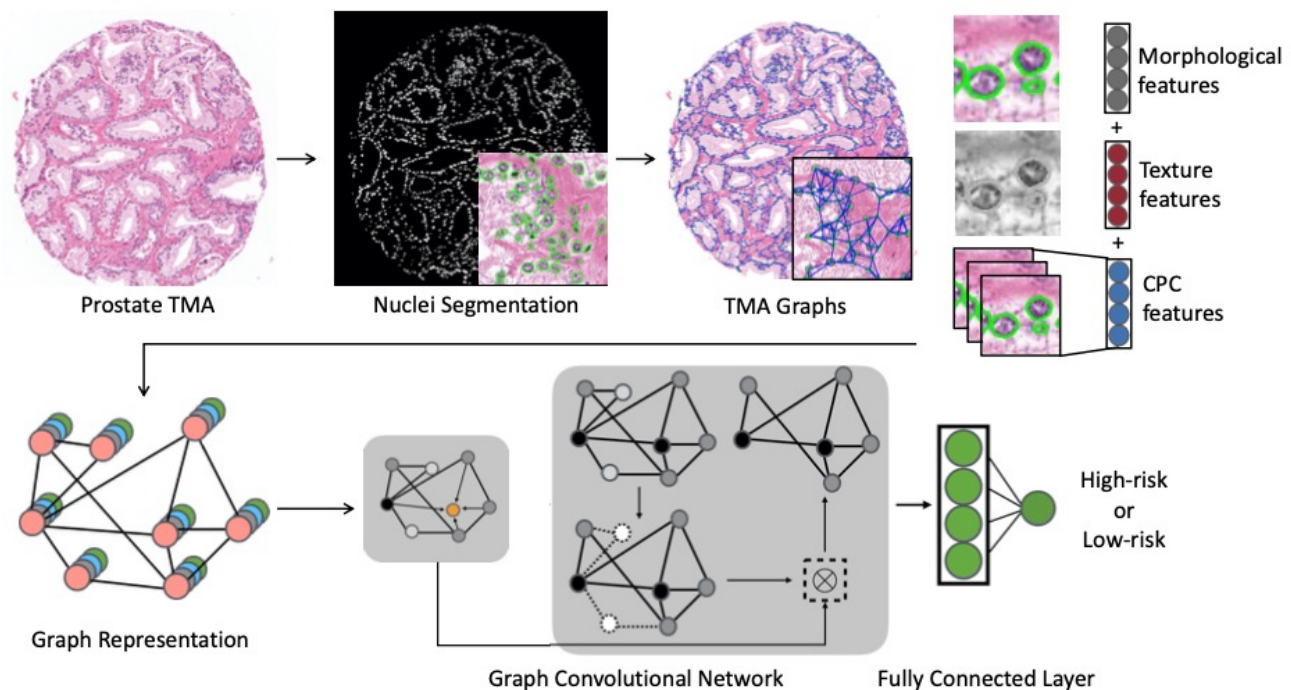
Ground Truth

Predictions



# Downstream Applications

- Models are available for use in PathML
- Segmented nuclei can be used to construct graph representations, then used for Graph Convolutional Networks (GCN)



# Thank You

Any Questions?

## DFCI I&A

- Bryan Gass
- Sreekar Reddy Puchala
- Ella Halbert
- Xiaoxuan Liu
- Haoyuan Li
- Jie Sun
- Daniel Waranch
- AIOS Group
- Renato Umeton
- Jason Johnson

## DFCI Med Onc

- Jackson Nyman
- Surya Hari
- Eli Van Allen

## Weill Cornell Pathology

- Ryan Carelli
- Mohamed Omar
- David Brundage
- Karen Xu
- Luigi Marchionni
- Massimo Loda

## DFCI Innovation Office

- Kate Strayer-Benton
- Vladimir Leopard
- Aaron Dy
- Lesley Solomon

- Interested in using PathML in your work?
- Want to contribute to development of new features?

**Get in touch!**

[PathML@dfci.harvard.edu](mailto:PathML@dfci.harvard.edu)

# PathML

- Website: <https://pathml.org>
- Documentation: <https://docs.pathml.org>
- GitHub: <https://github.com/Dana-Farber-AIOS/pathml>
- Manuscript: [doi.org/10.1158/1541-7786.MCR-21-0665.MCR-21-0665](https://doi.org/10.1158/1541-7786.MCR-21-0665.MCR-21-0665)