Deploying KNIME in an Amazon Cloud Environment for High-Throughput Image Analysis

Andries Zijlstra Ph.D.

Vanderbilt University Medical Center
Dept. Pathology, Microbiology and Immunology
Prg. Cancer Biology & Vanderbilt Ingram Cancer Center

The Plan

Intro

Clinical Question

HistoMAP

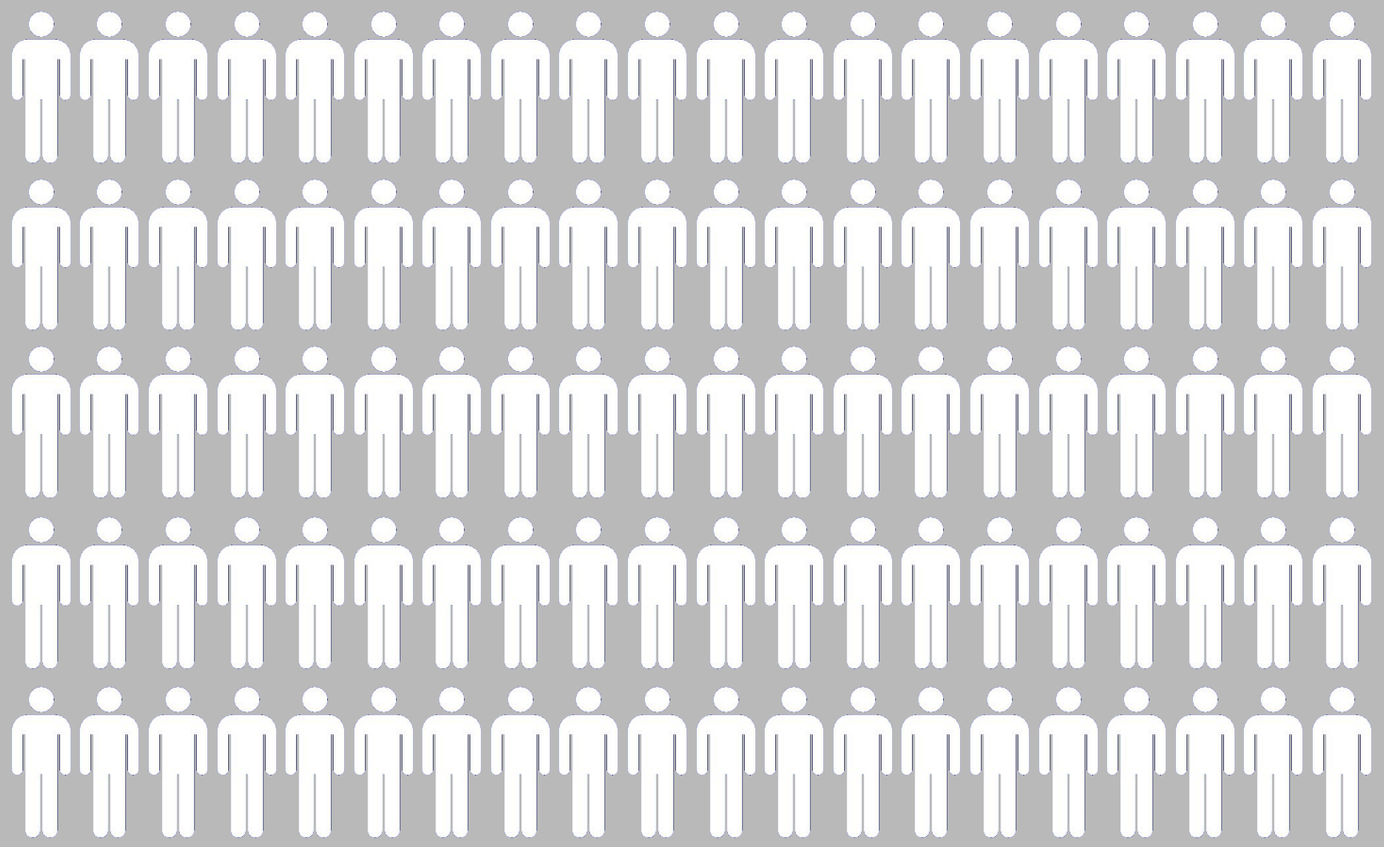
Machine Learning Assisted HistoPathology

[Secure] deployment of KNIME in AWS

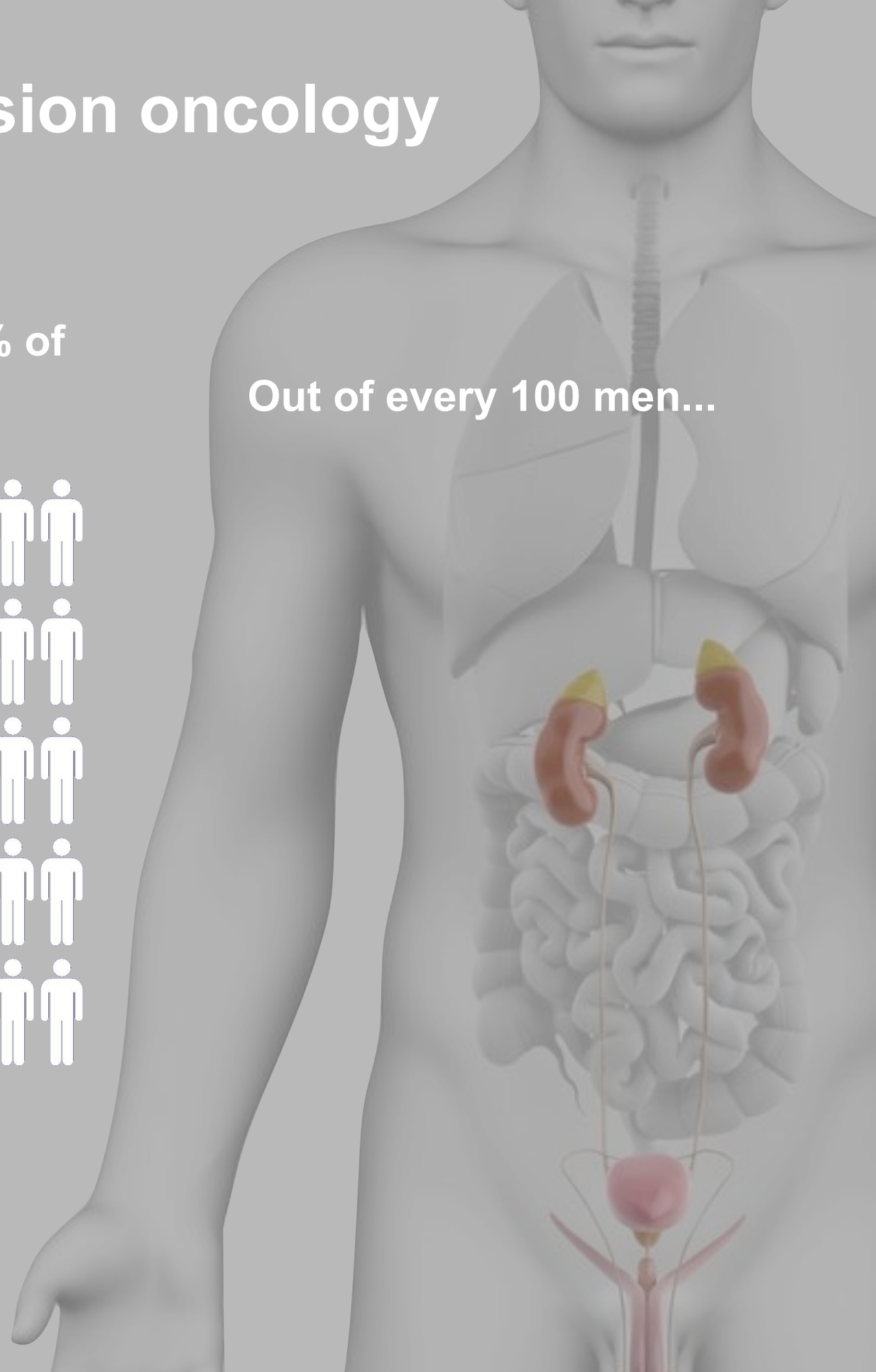
Prediction in Medicine

The need for precision oncology

36% of newly diagnosed cancers, and 10% of all cancer deaths in men

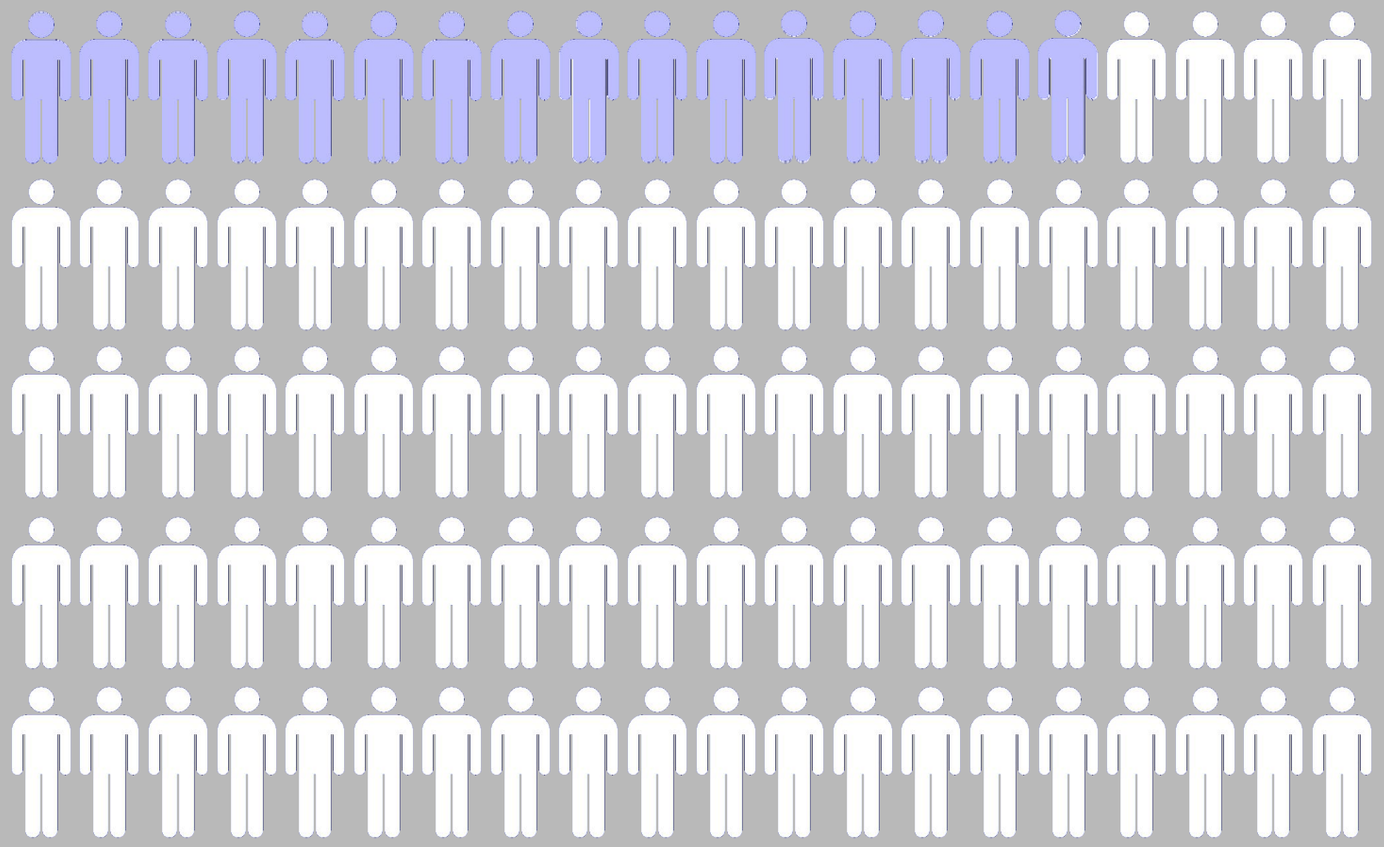


Out of every 100 men...



The need for precision oncology

36% of newly diagnosed cancers, and 10% of all cancer deaths in men

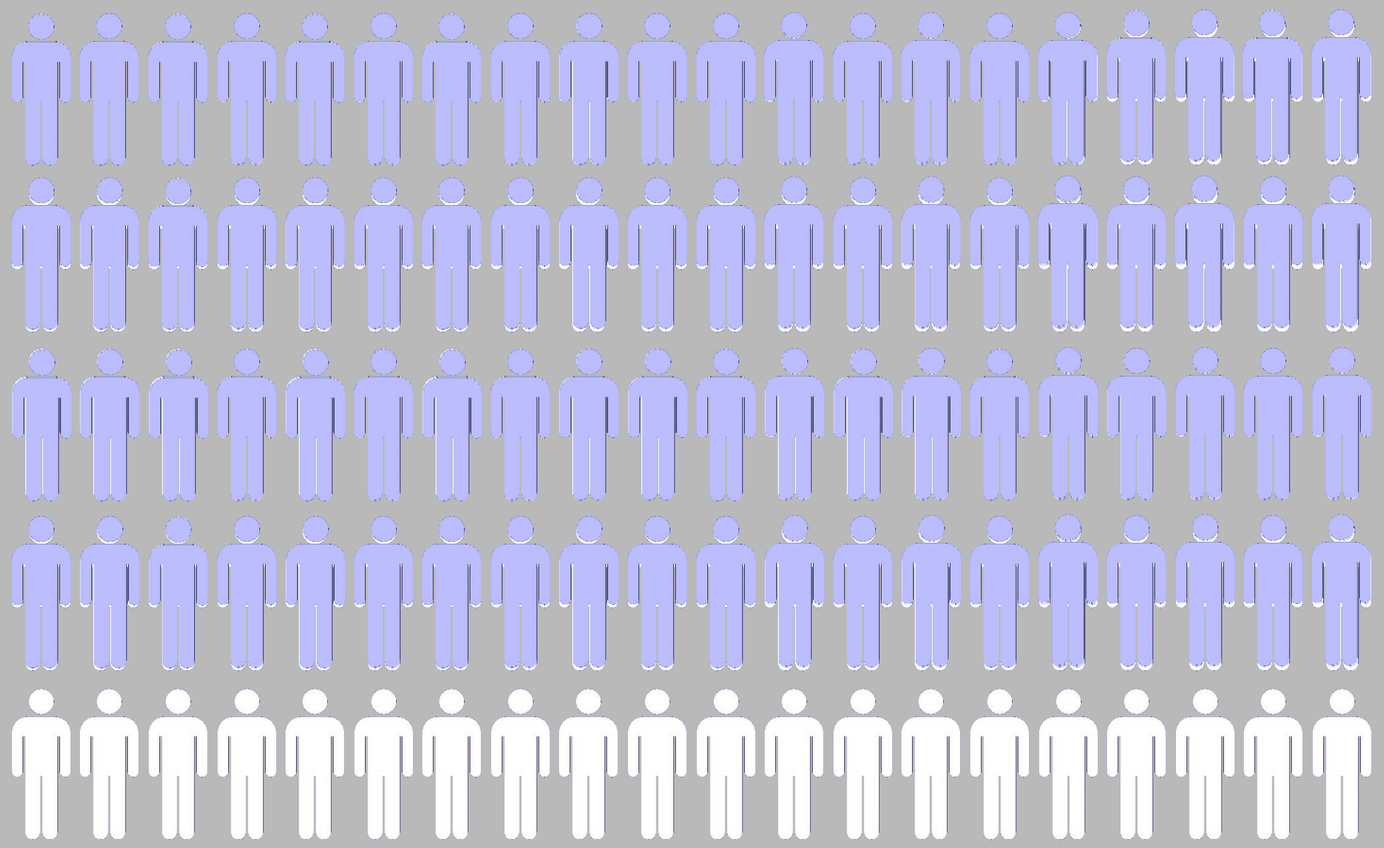


Out of every 100 men...

16 will be diagnosed with prostate cancer in their lifetime

The need for precision oncology

36% of newly diagnosed cancers, and 10% of all cancer deaths in men



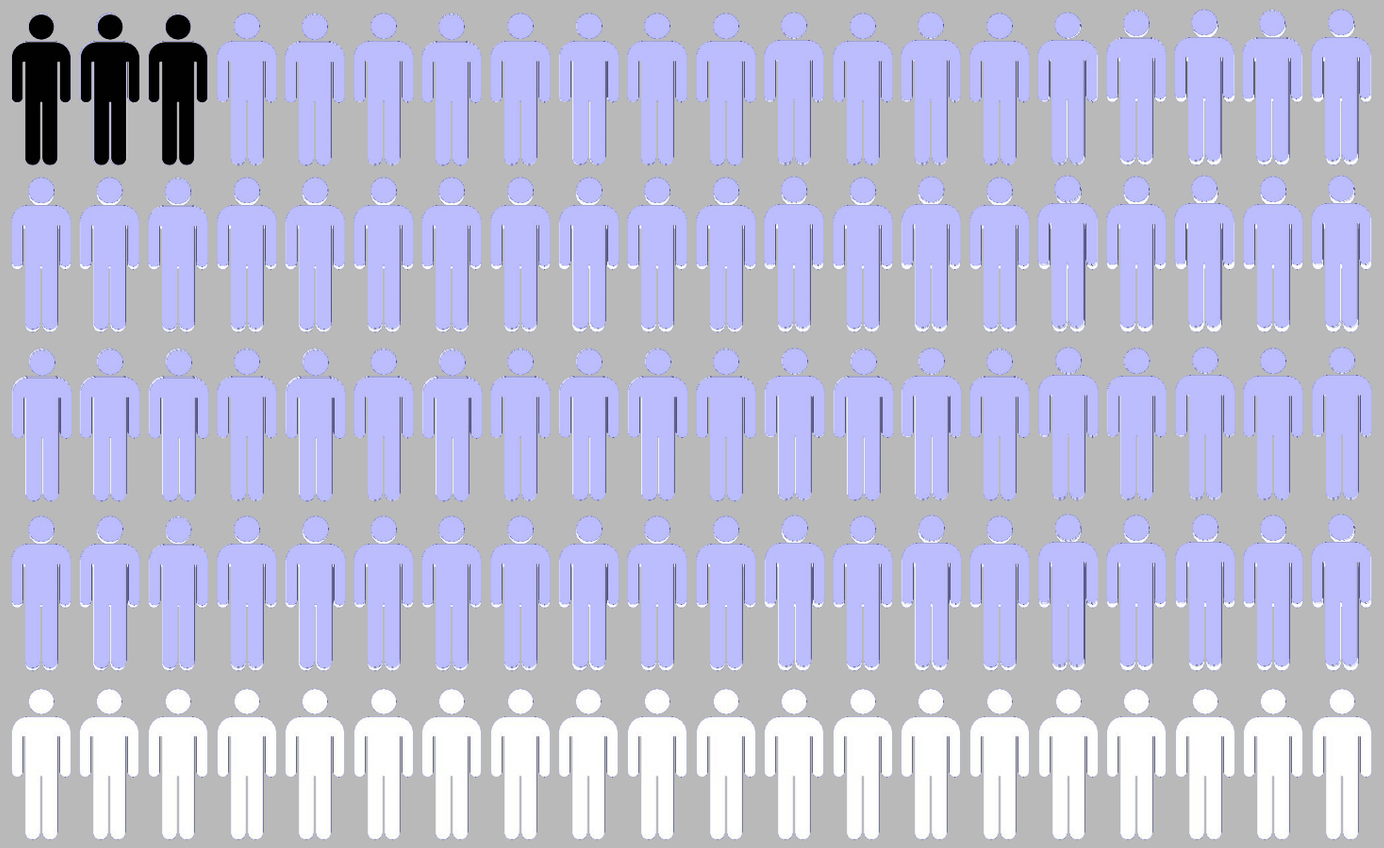
Out of every 100 men...

16 will be diagnosed with prostate cancer in their lifetime

In reality, up to 80 will have prostate cancer by age 70

The need for precision oncology

36% of newly diagnosed cancers, and 10% of all cancer deaths in men



Out of every 100 men...

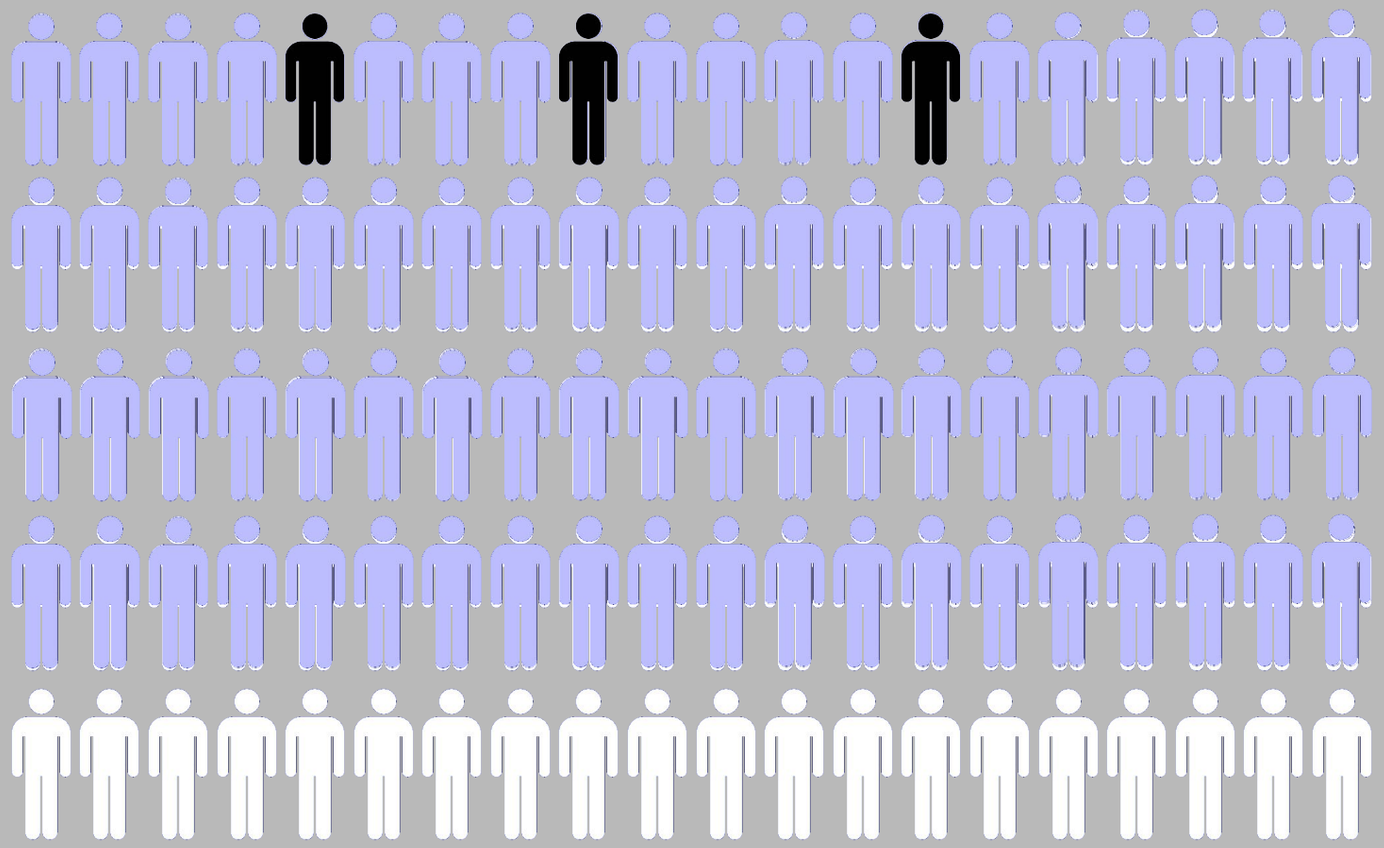
16 will be diagnosed with prostate cancer in their lifetime

In reality, up to 80 will have prostate cancer by age 70

And 3 will die from it.

The need for precision oncology

36% of newly diagnosed cancers, and 10% of all cancer deaths in men



The goal: diagnose patients that have aggressive disease

Out of every 100 men...

16 will be diagnosed with prostate cancer in their lifetime

In reality, up to 80 will have prostate cancer by age 70

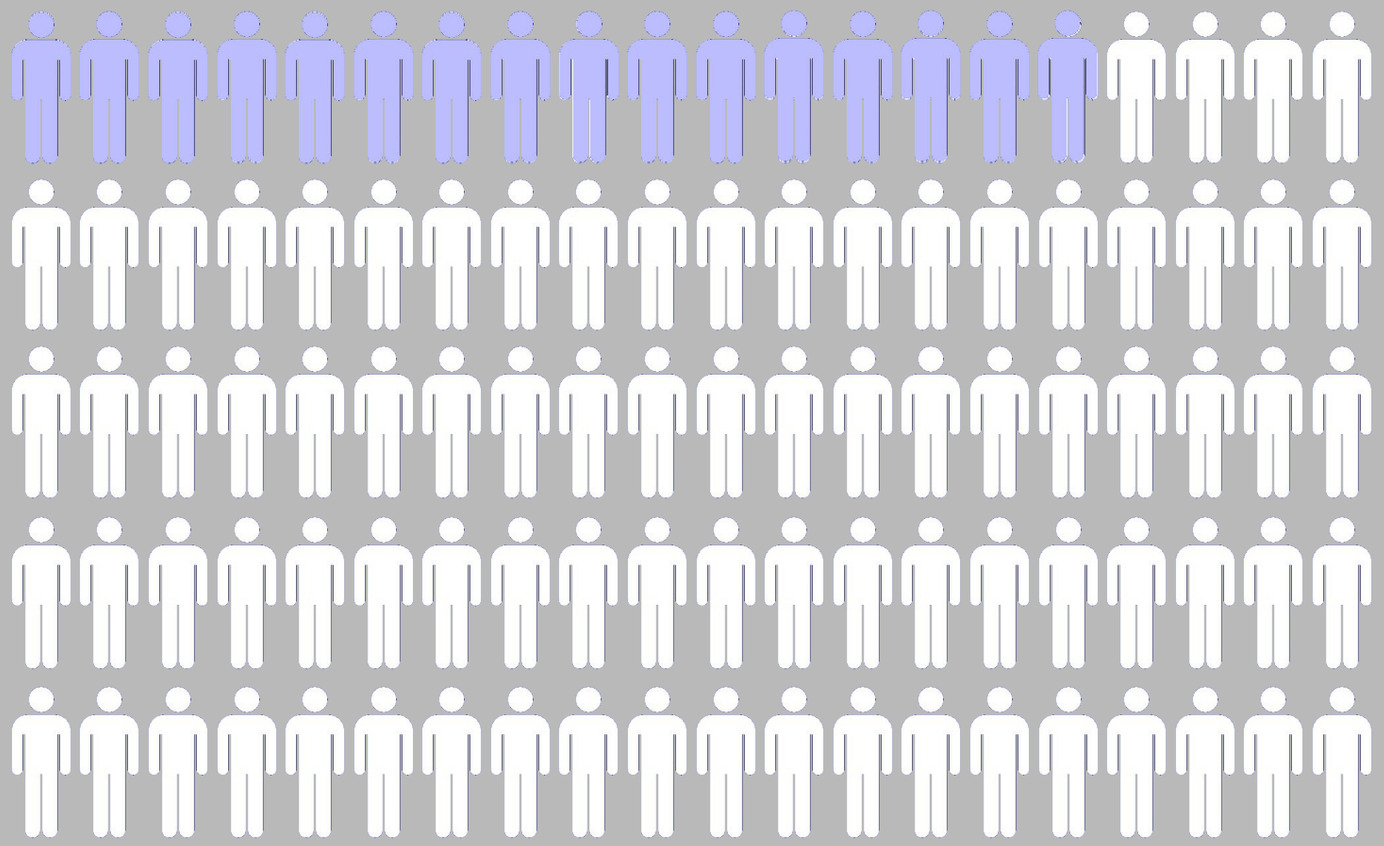
And 3 will die from it.

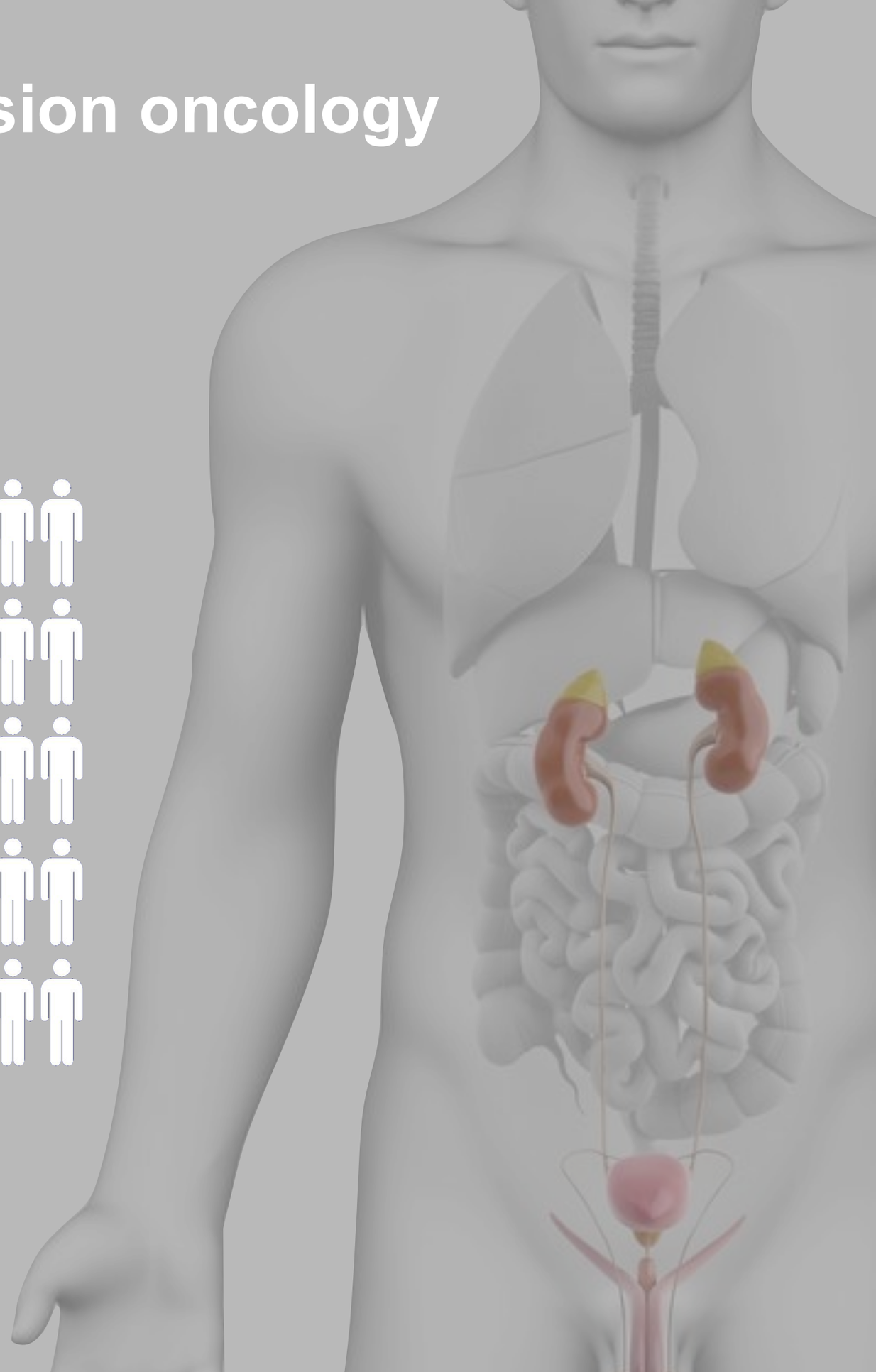
But which 3 ?

In the meantime, we <u>over-treat</u> many patients

The need for precision oncology

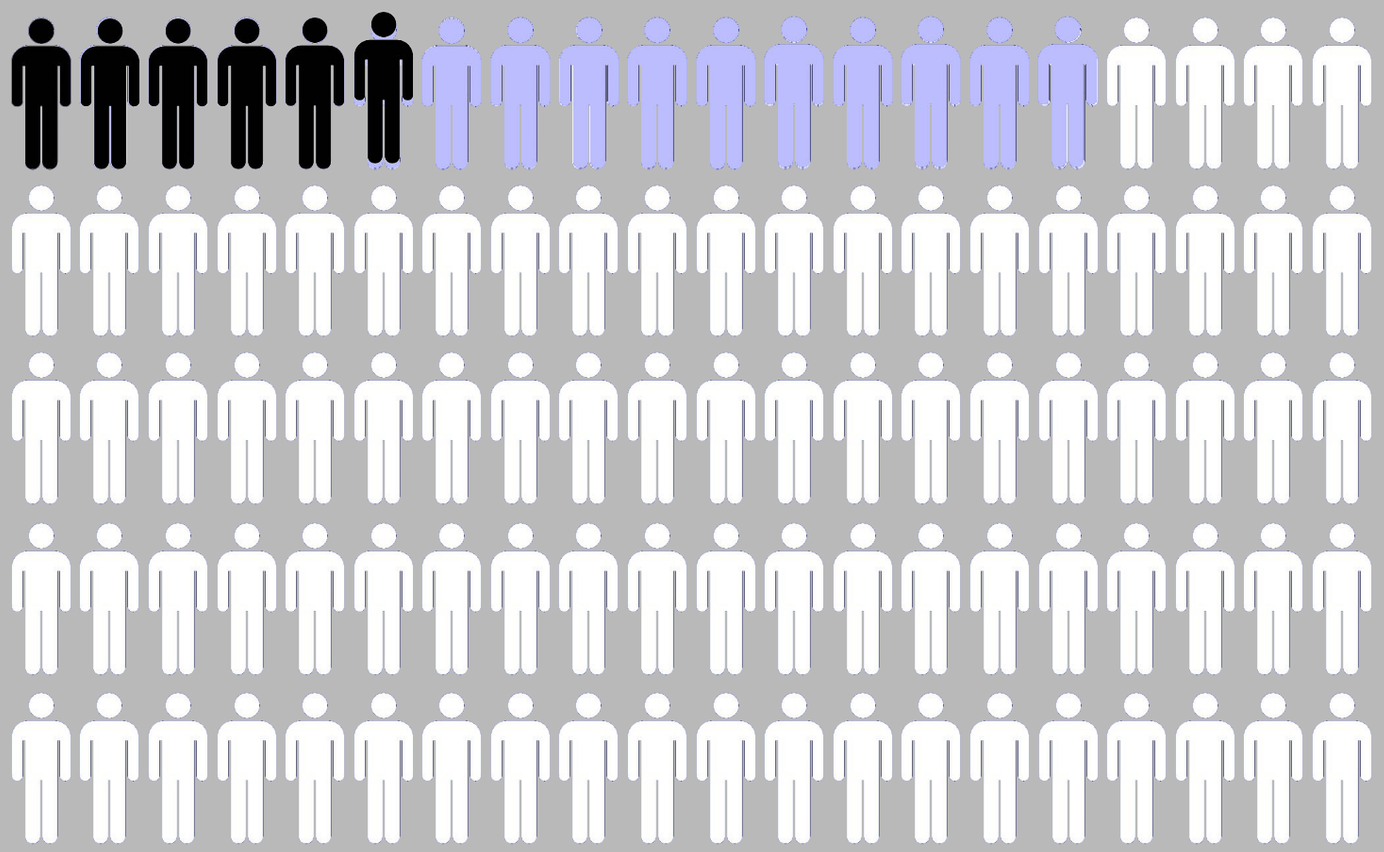
Impact of Prostate Cancer Treatments: Differences in Treatment Responses



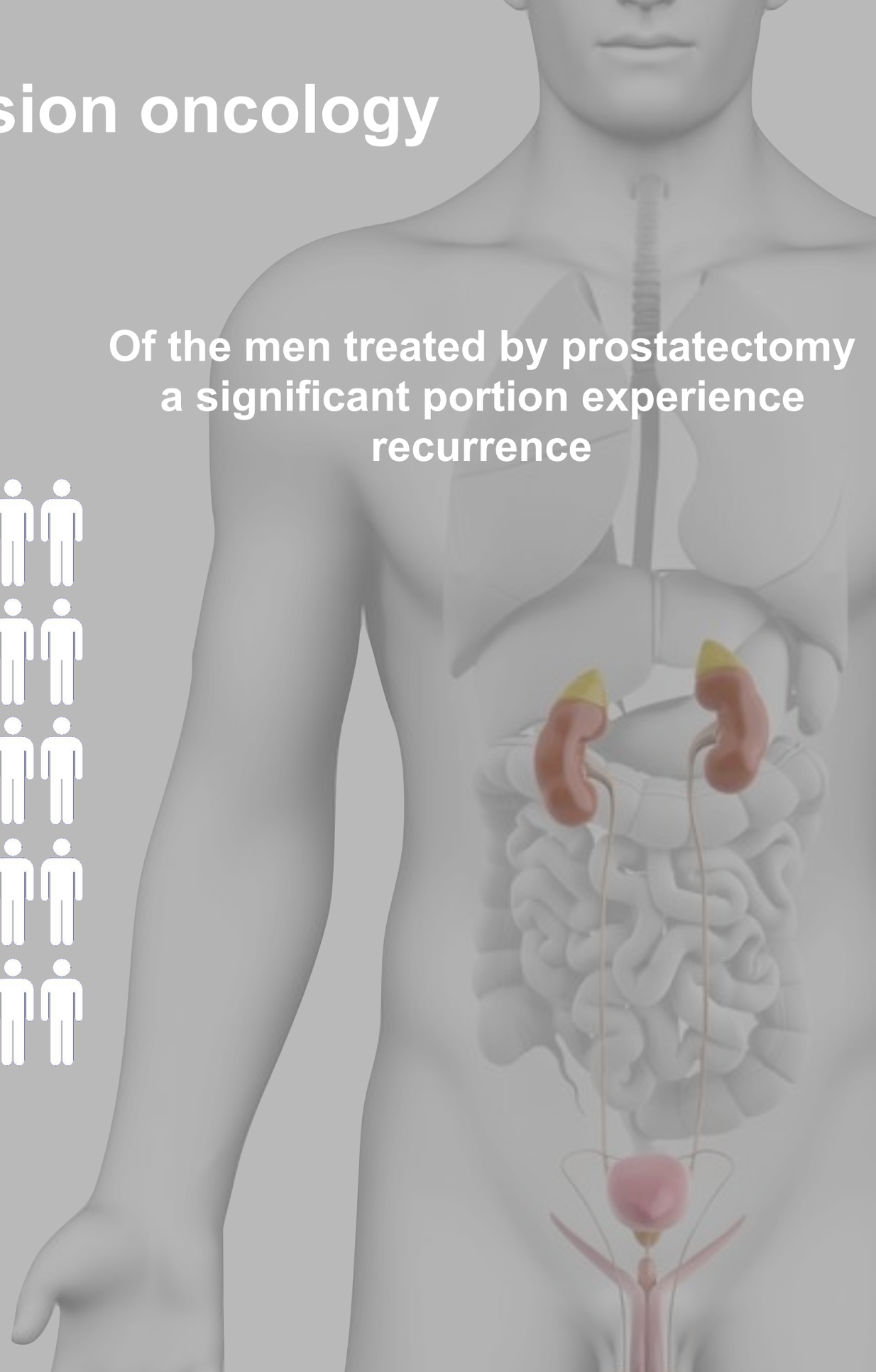


The need for precision oncology

Impact of Prostate Cancer Treatments: Differences in Treatment Responses

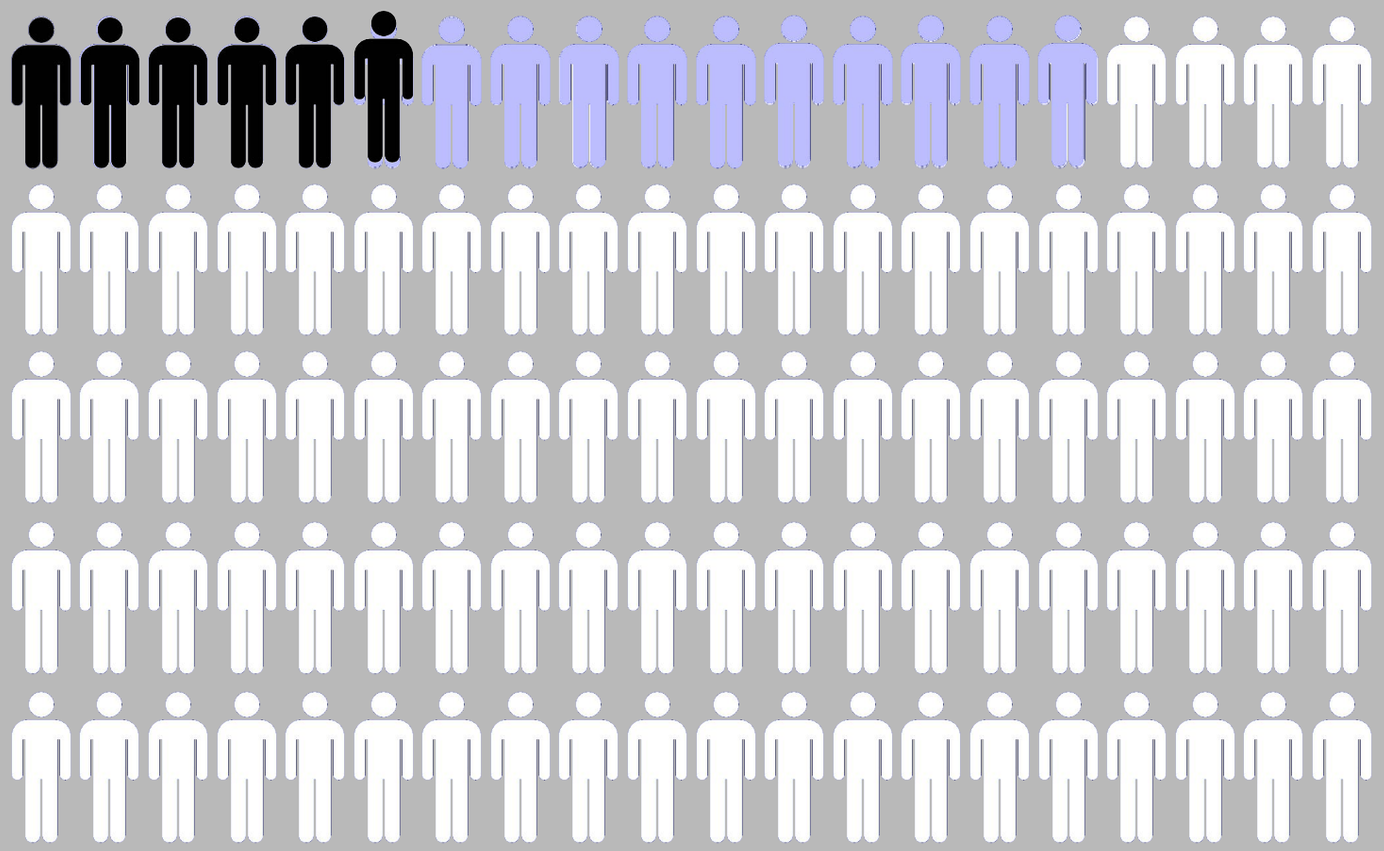


Of the men treated by prostatectomy a significant portion experience recurrence



The need for precision oncology

Impact of Prostate Cancer Treatments: Differences in Treatment Responses

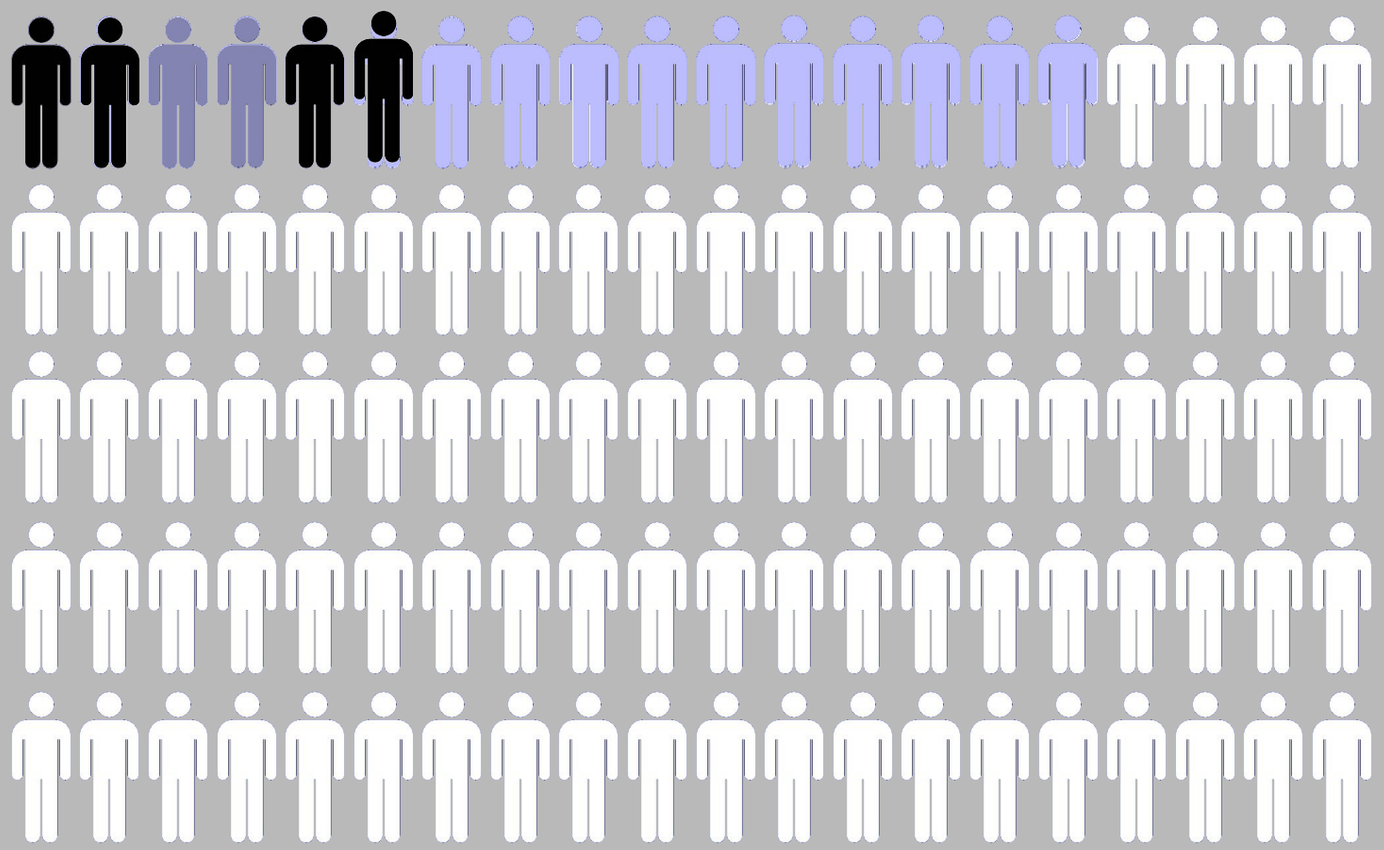


Of the men treated by prostatectomy a significant portion experience recurrence

Multiple treatment options are available for these men

The need for precision oncology

Impact of Prostate Cancer Treatments: Differences in Treatment Responses



The goal: determine which patient will respond to what drug.

Of the men treated by prostatectomy a significant portion experience recurrence

Multiple treatment options are available for these men

On average, the probability responding to treatment is the same, however,…

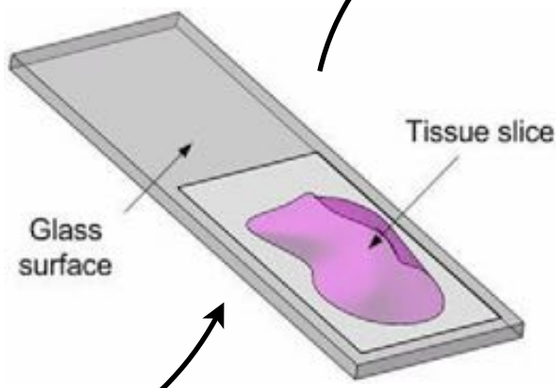
Not every patient will respond equally

Who responds to what treatment?

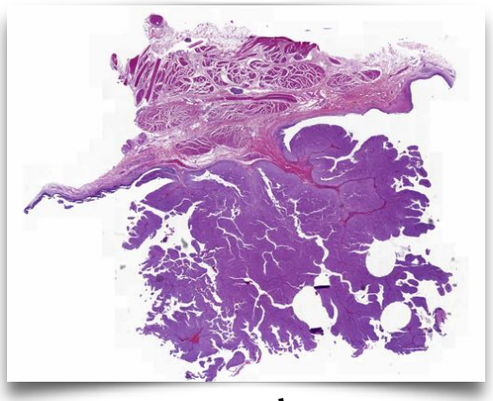
Computer-guided image analysis in pathology

Tissue Block

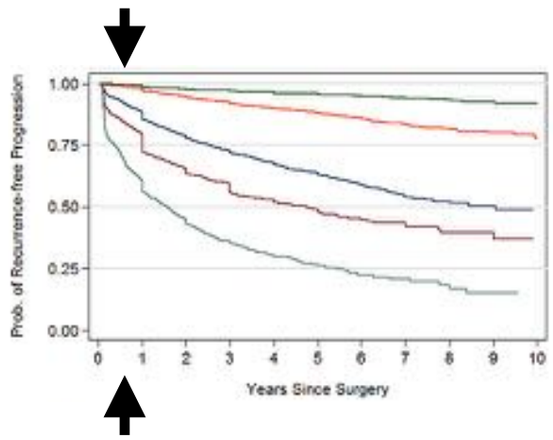
Tissue Section



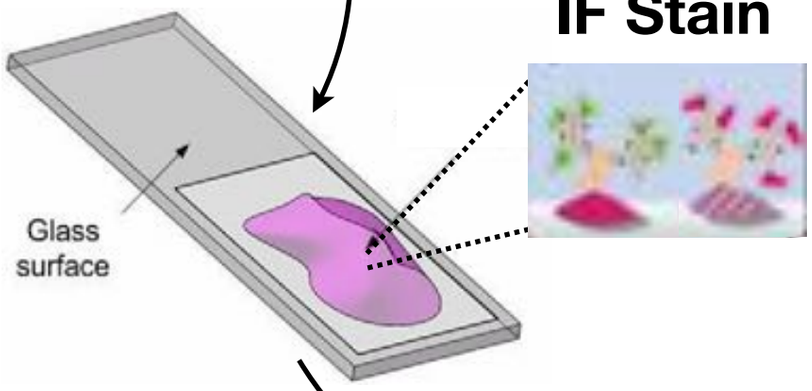
H&E Stain



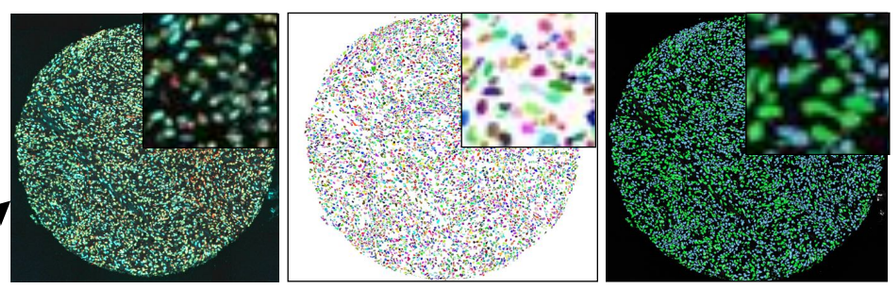
Pathology Review

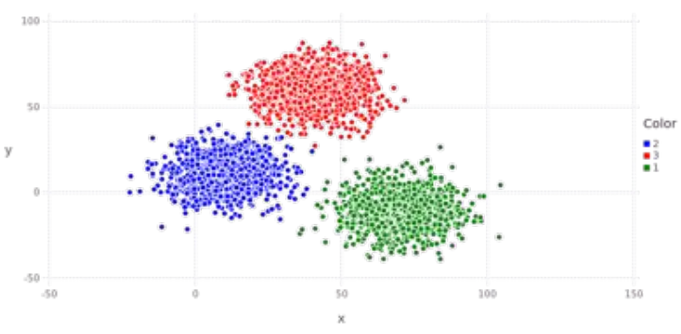


IF Stain



Computer-assisted segmentation, feature extraction and classification





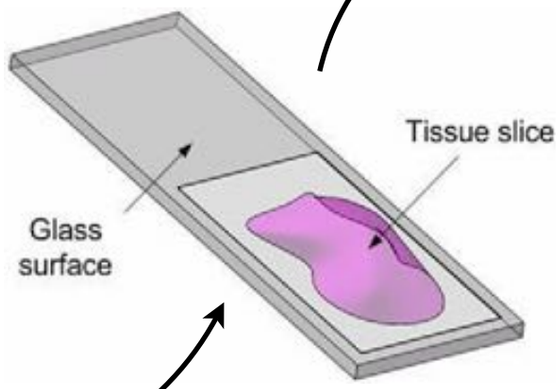
Immunofluorescent Stain

Segmentation/Classification/Clustering

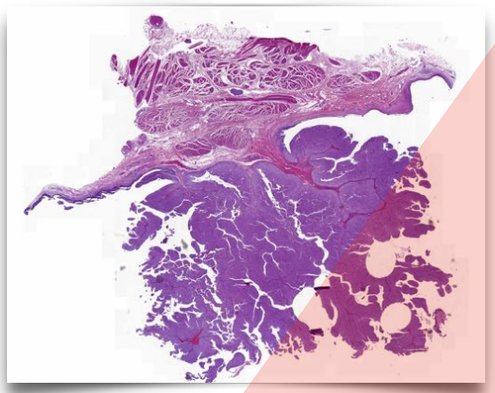
Computer-guided image analysis in pathology

Tissue Block

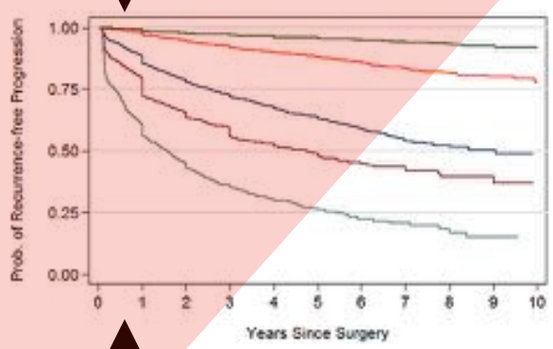
Tissue Section



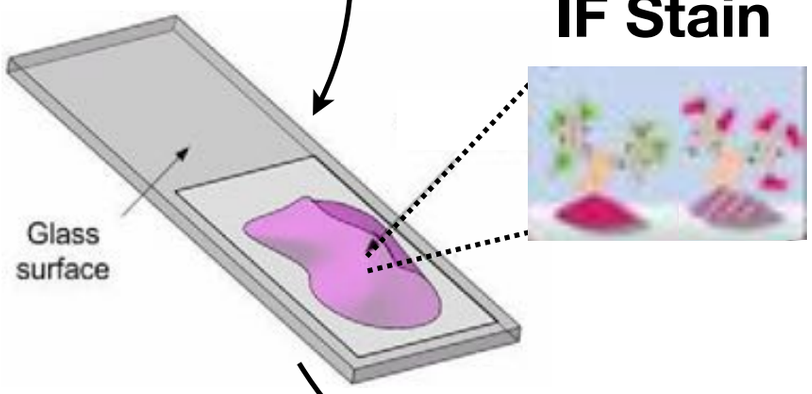
H&E Stain



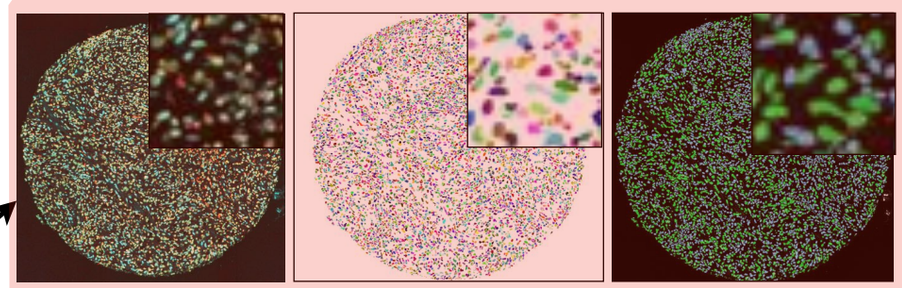
Pathology Review

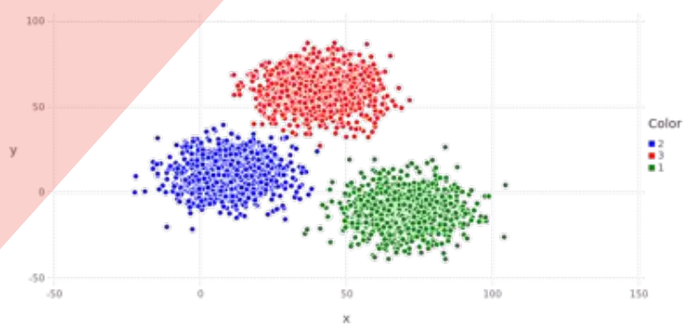


IF Stain



Computer-assisted segmentation, feature extraction and classification



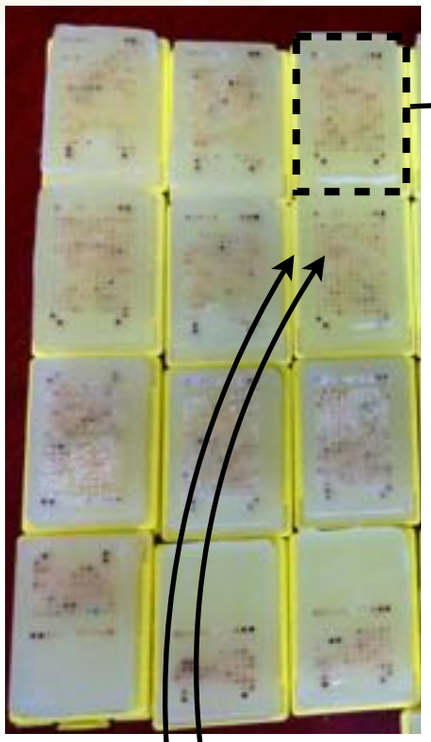


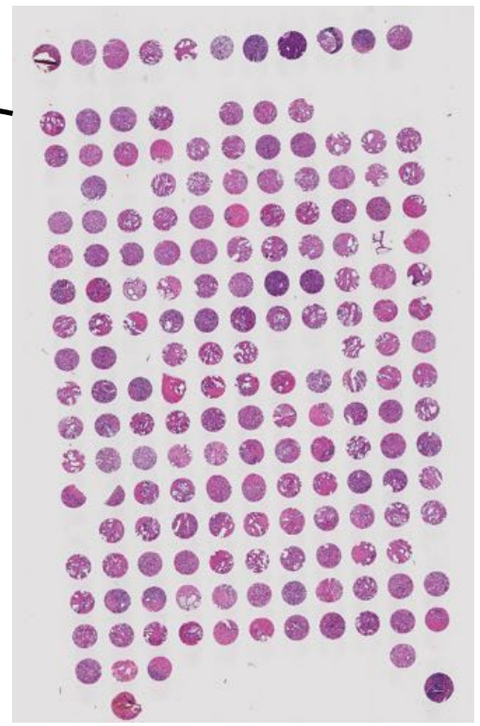
Immunofluorescent Stain

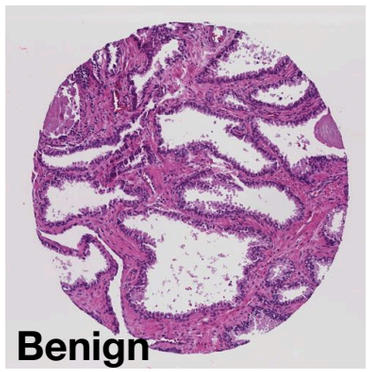
Segmentation/Classification/Clustering

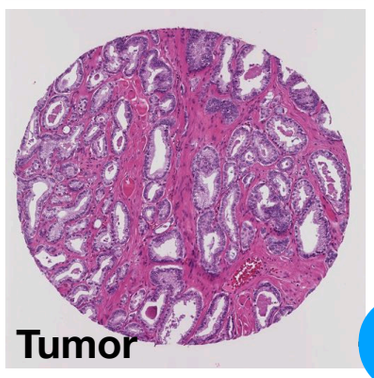
Source Data

The Source Data

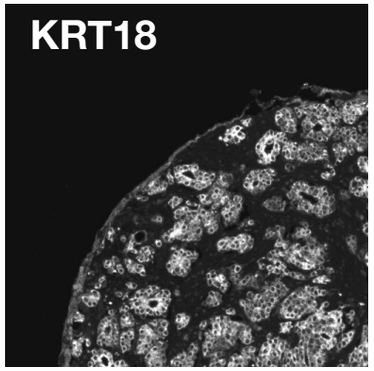


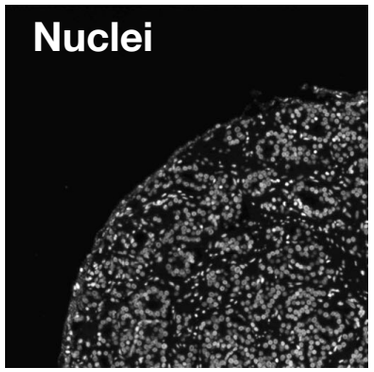


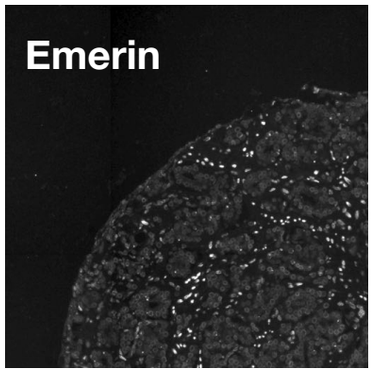


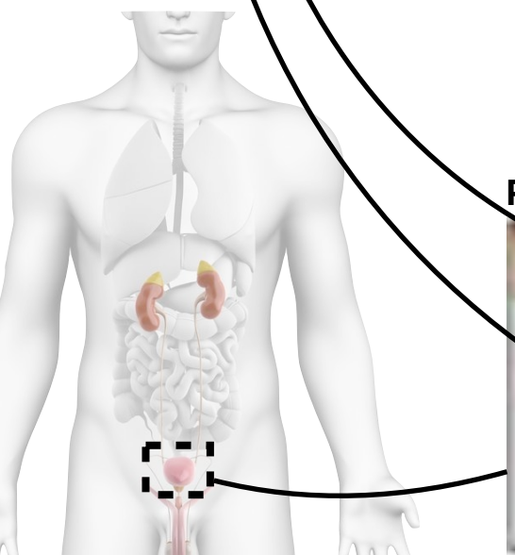


Histological stains



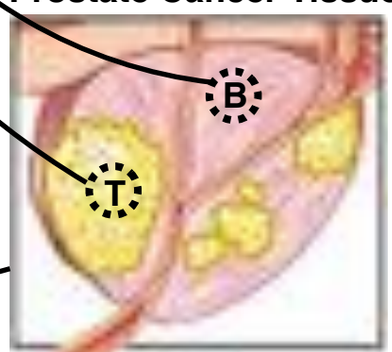






556 Men w. Prostatectomies

Prostate Cancer Tissue

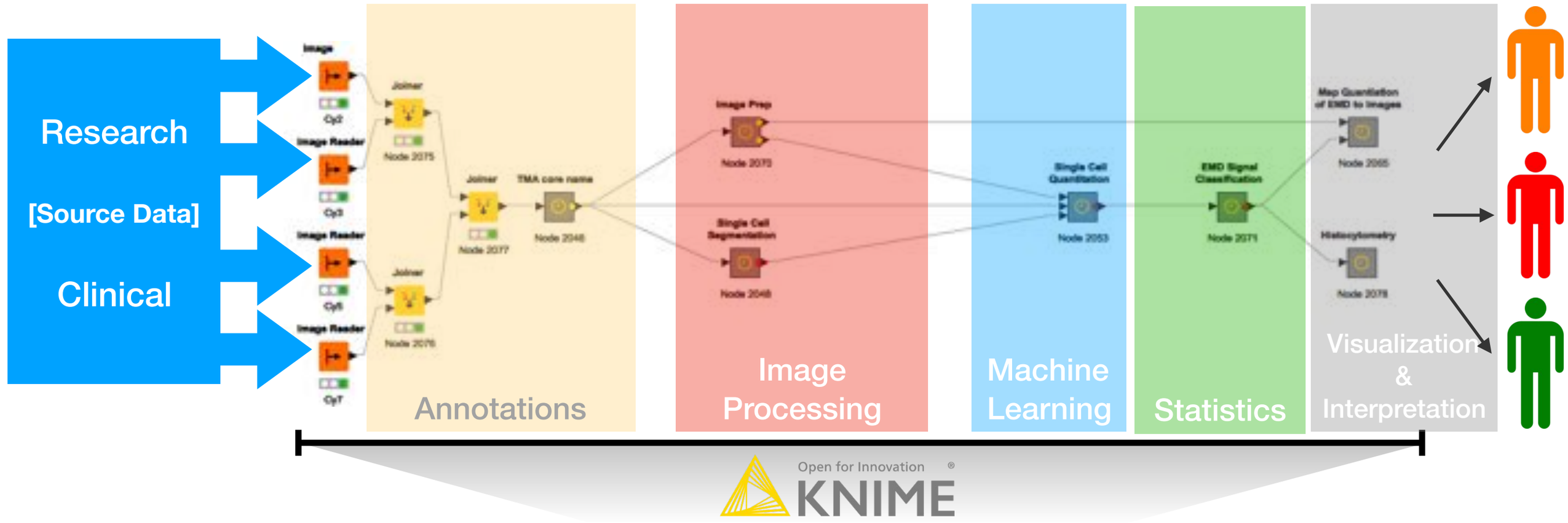


Tissue Annotation		Clinical Data		Image Data
Patient # Tissue Type Tissue origin Slide # Core #	+	Demographics Pathological Staging Treatment Outcome	+	Staining type Image

Why use KNIME?

Why use KNIME

Establish an open computational environment to provide direct access across all relevant expertise.



Why use KNIME

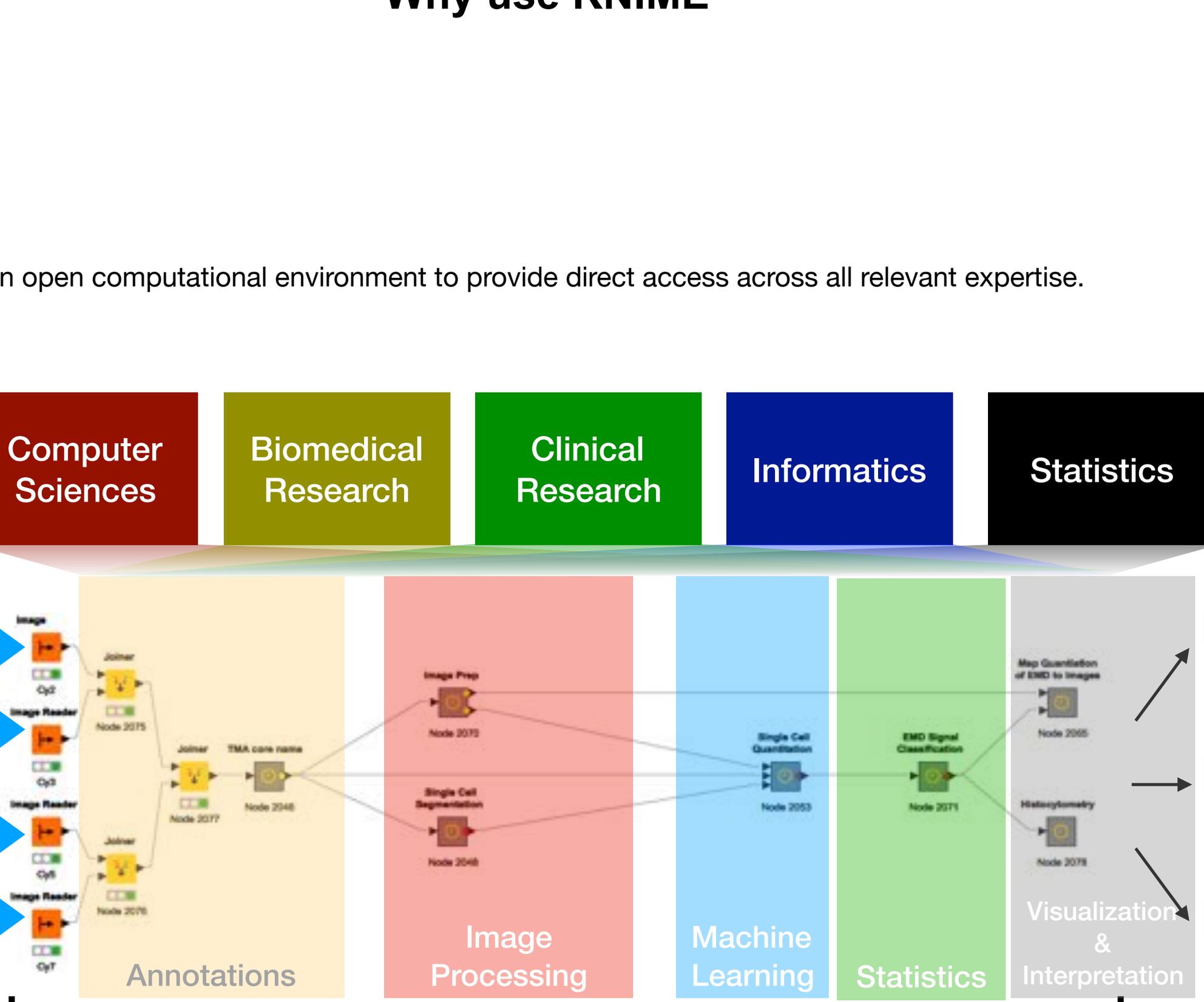
Establish an open computational environment to provide direct access across all relevant expertise.

Computer Sciences | Biomedical Research | Clinical Research | Informatics | Statistics

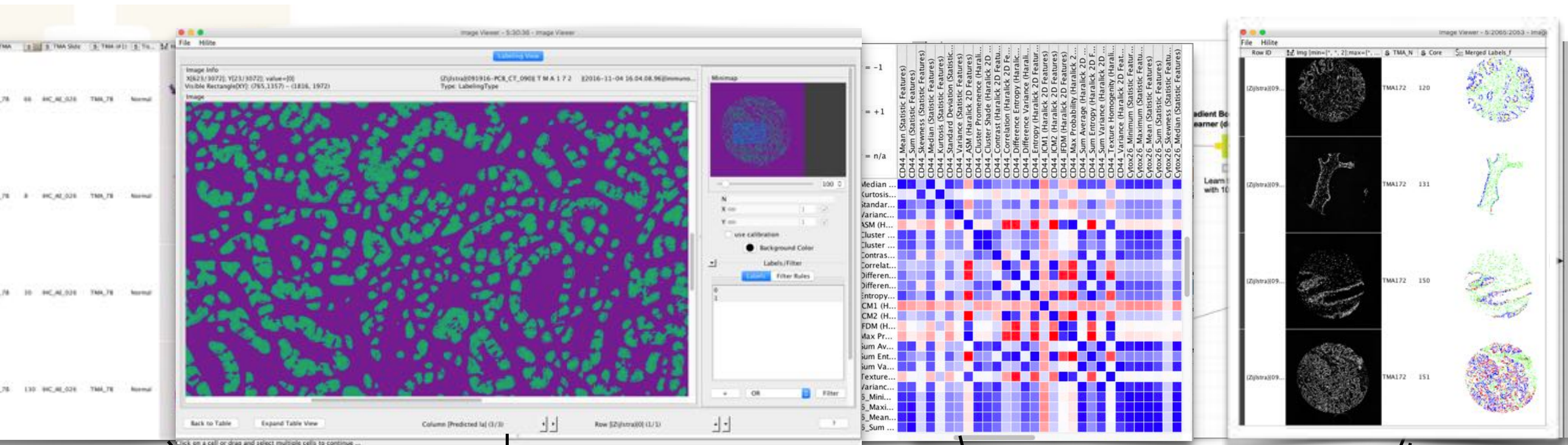
Research

[Source Data]

Clinical



Open for Innovation ® KNIME



Computer Sciences

Biomedical Research

Clinical Research

Informatics

Statistics

Research

[Source Data]

Clinical

Annotations

Image Processing

Machine Learning

Statistics

Visualization & Interpretation

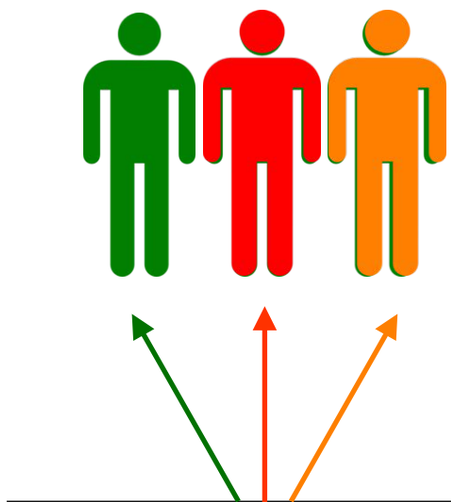
Open for Innovation ®

KNIME

Why go to the “Cloud”

Scalability
Accessibility
Security
Cost

The challenges and opportunities of single-cell analysis



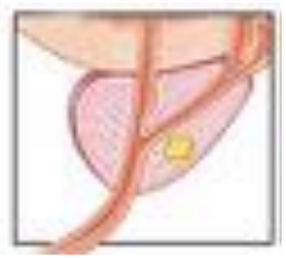
Predict

Cluster/Train

18 billion
data points

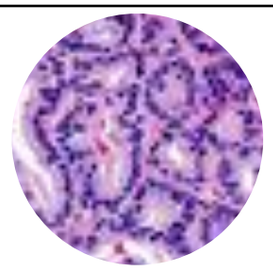
Assay Annotations

Patient Data



Prostate

286 features/cell



1 patient

4 pieces of tissue

556 patients

2,224 specimens

6 stains

13,344 images

5000 cells
per specimen

67 million rows data

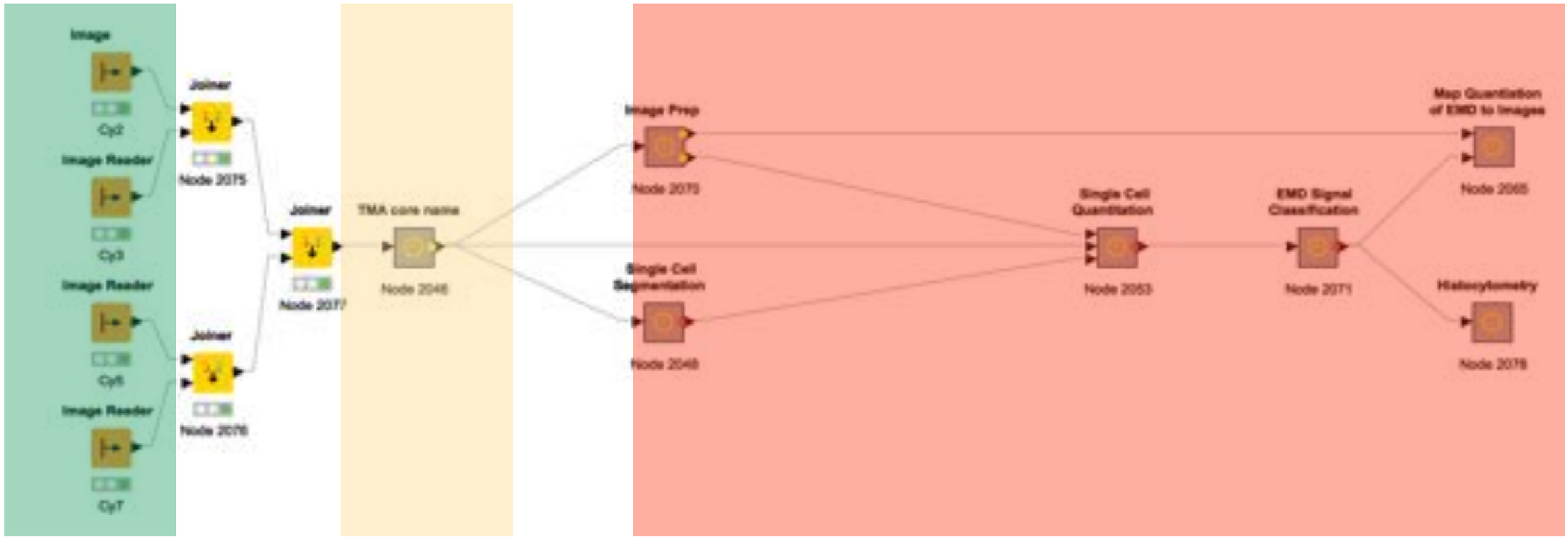
Project scale:

Configuration:

KNIME 3.2
MacPro
12-core
64GB RAM
1TB SSD M2 HD (scratch)
4TB Spindle HD
4-20mb/sec transfer

Project	Scenario	Per image	# images	Scratch (GB)	Time (hr)	Problem
A	1 slide	9.5 megapixel	800	250	12	Slow
	8 slides	9.5 megapixel	6400	2,000	96	Out of space
B	1 slide	16 megapixel	800	500	X	No go
	8 slides	16 megapixel	6400	4,000	X	No go

200 x 9.5 megapixel x 4 -> 12 gb-> - [20X - expansion] -> 256-500 gb



HistoMAP Server

EC2 Instances Types

General Purpose

Compute Optimized

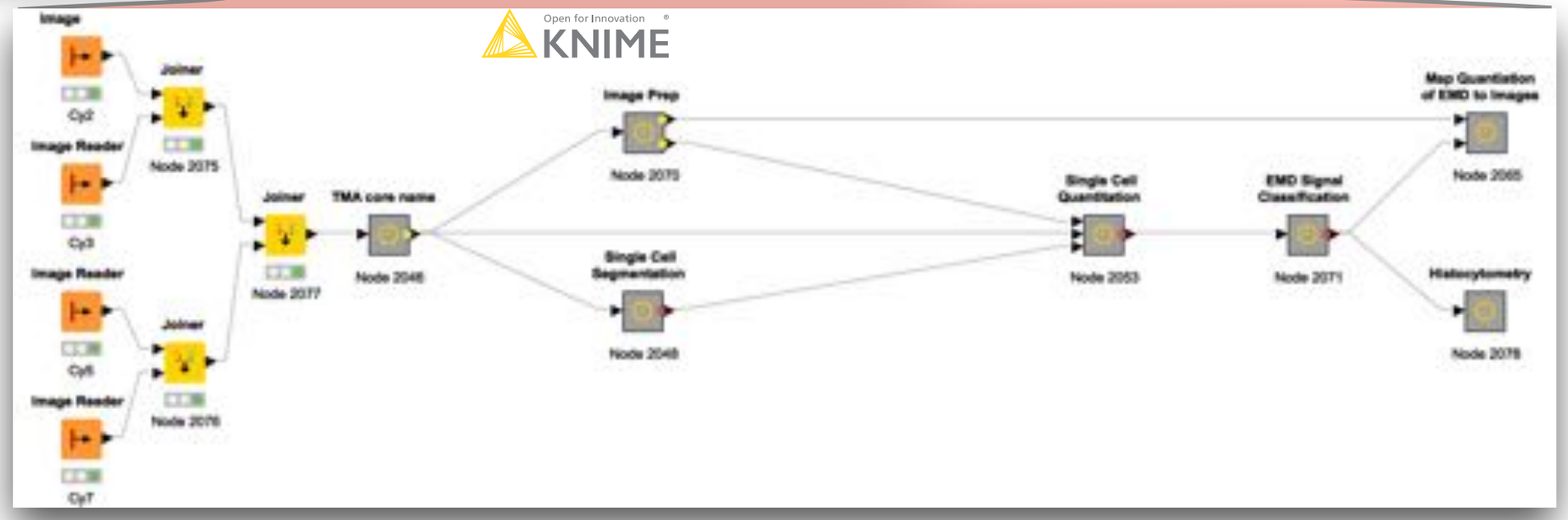
Memory Optimized “r”

Accelerated Computing “p”

Storage Optimized

Instance Features

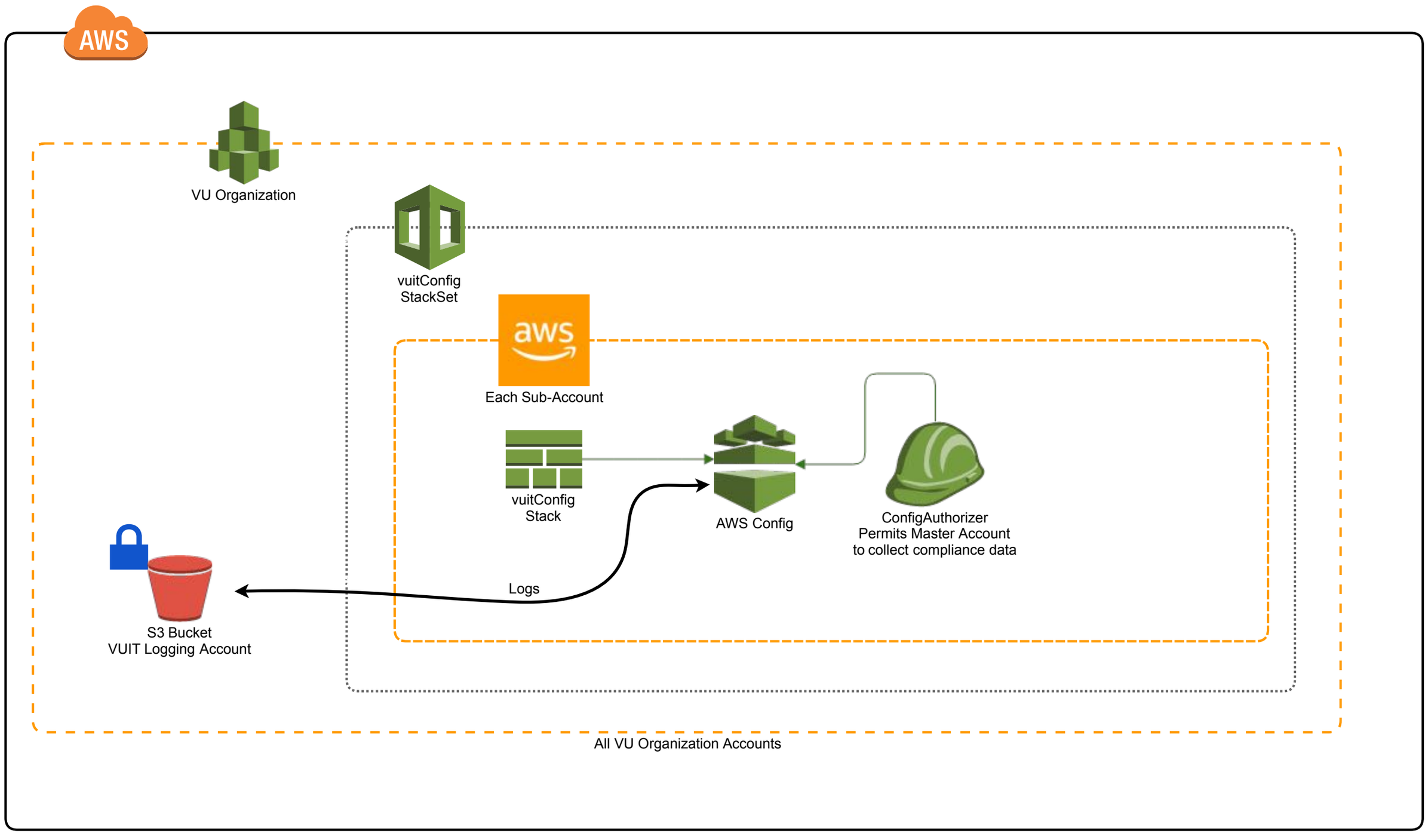
Measuring Instance Performance



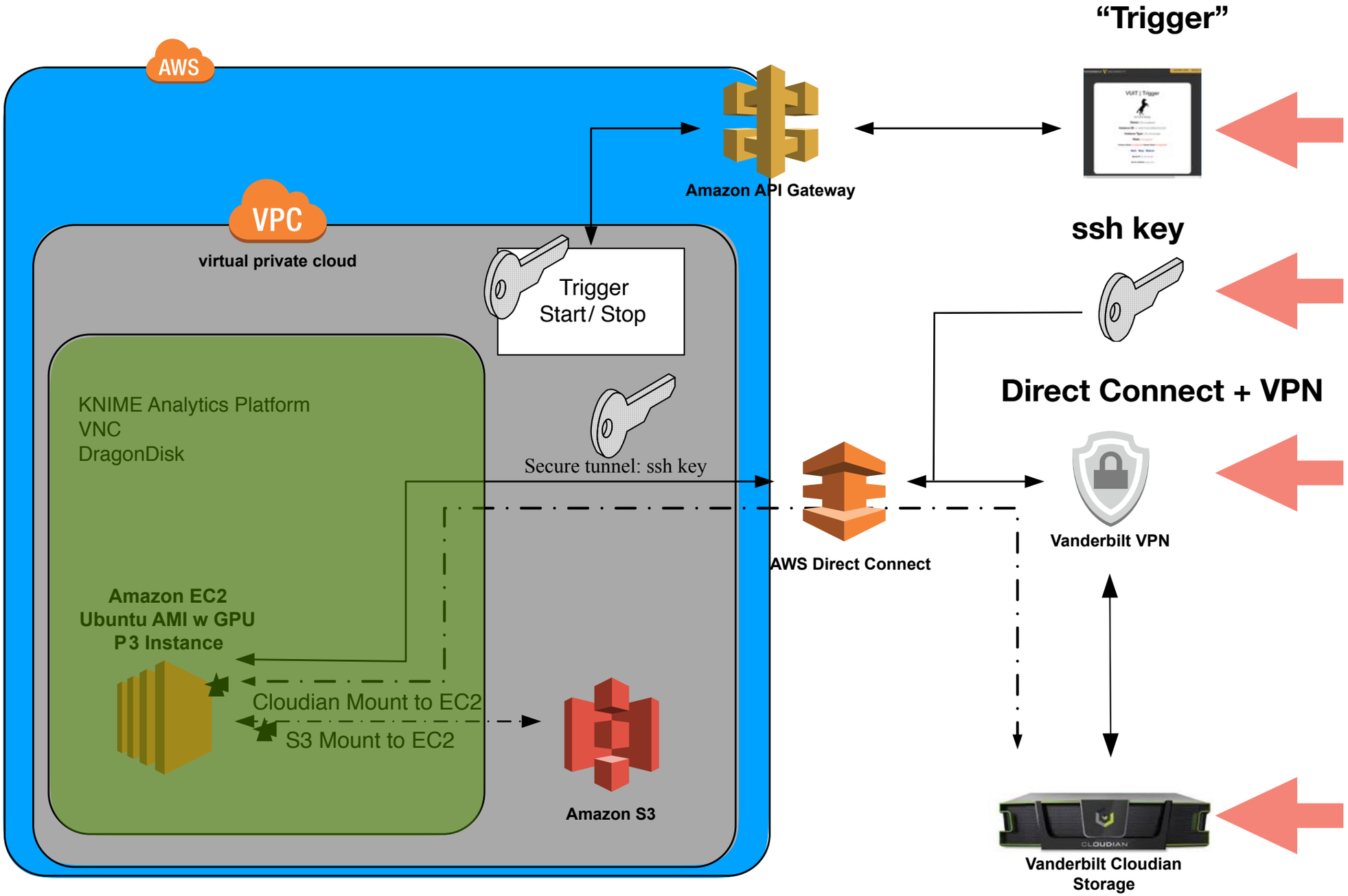
Cloud Requirements

Requirement	Solution
Highly Scalable	Multiple configuration, eg.: r4x4 vs R4x2; Data repository on S3
Remotely Accessible	Enable VPN
Highly Secure	4 part: a) server-less trigger, b) ssh, c) Direct Connect, d) primary data lives on local S3 bucket
Manageable Cost	Easy “on/off” + automated reporting on costs
Ease of Use	Server-less trigger for start/stop, VPN remote access

Vanderbilt University IT AWS Configuration



Vanderbilt University IT AWS Secure Access



Performance

VUIT AWS Performance

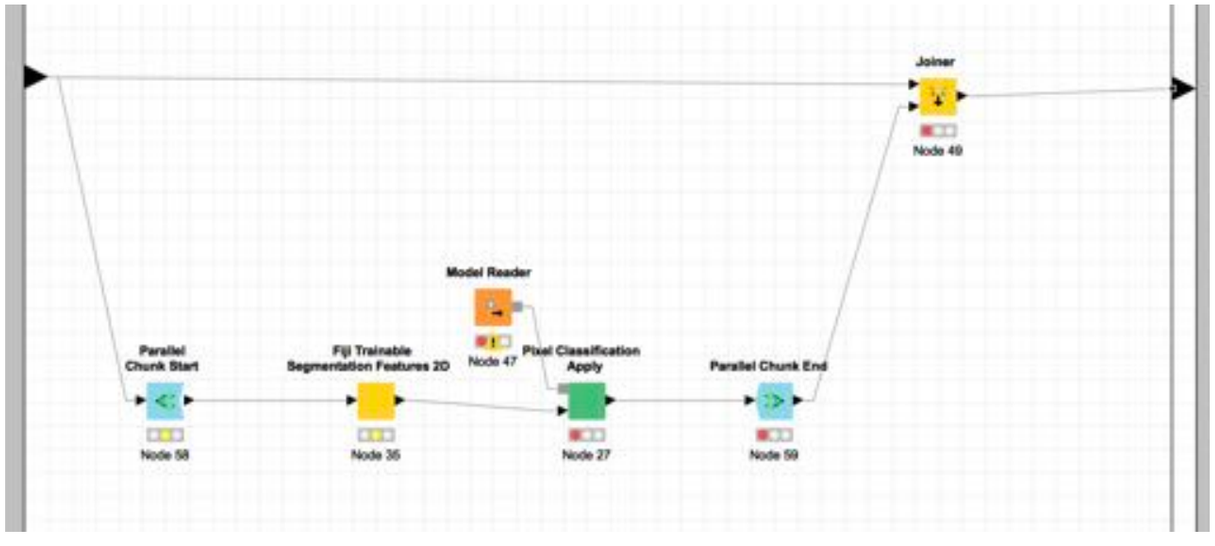
Configuration:
KNIME 3.2
MacPro
12-core
64GB RAM
1TB SSD M2 HD (scratch)
4TB Spindle HD
4-20mb/sec transfer

Configuration:
KNIME 3.2
Ubuntu EC2 r4x4xLarge
16-vCPU
122GB RAM
1TB SSD (scratch)
1001 TB S3
35-400mb/sec transfer

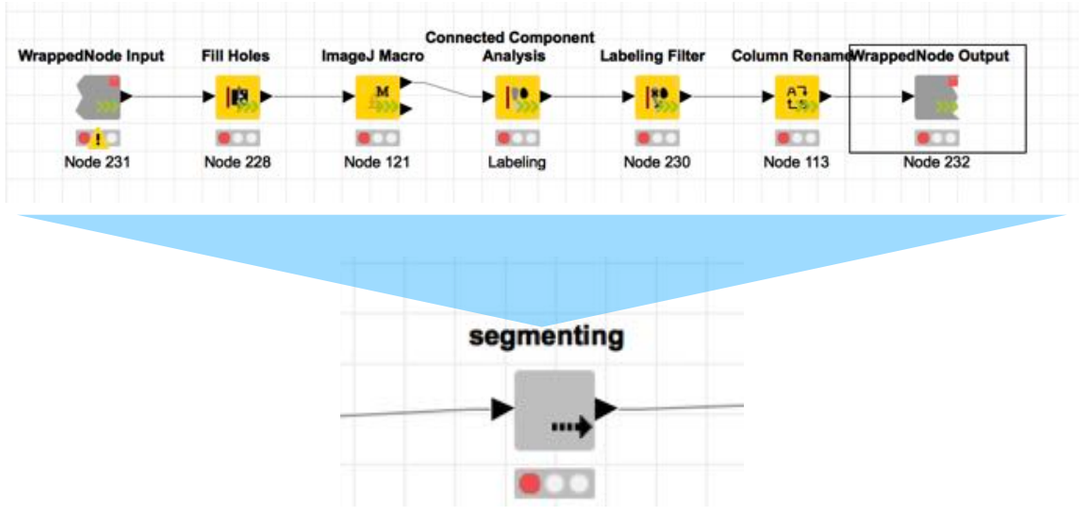
				Local			AWS	
Project	**Scenario**	**Per image**	**# images**	**Scratch (GB)**	**Time (hr)**	**Problem**	**Scratch (GB)**	**Time (hr)**
A	**1 slide**	9.5 megapixel	800	250	12	Slow	20	1
	8 slides	9.5 megapixel	6400	2,000	96	Out of space	200	8
B	**1 slide**	16 megapixel	800	500	X	No go	100	4
	8 slides	16 megapixel	6400	4,000	X	No go	800	32

Improved performance by a factor of 8-10!!!

Parallel Chunks

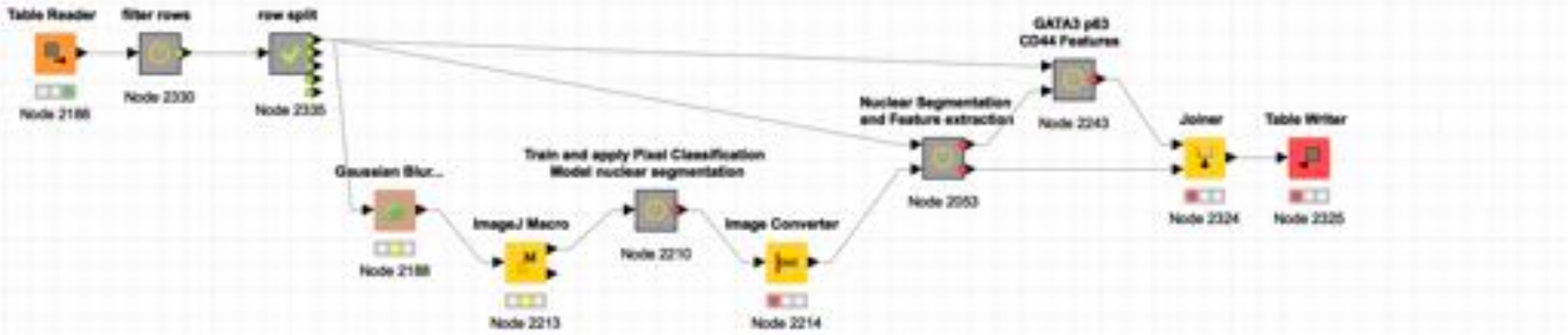


Streaming in wrapped metanodes

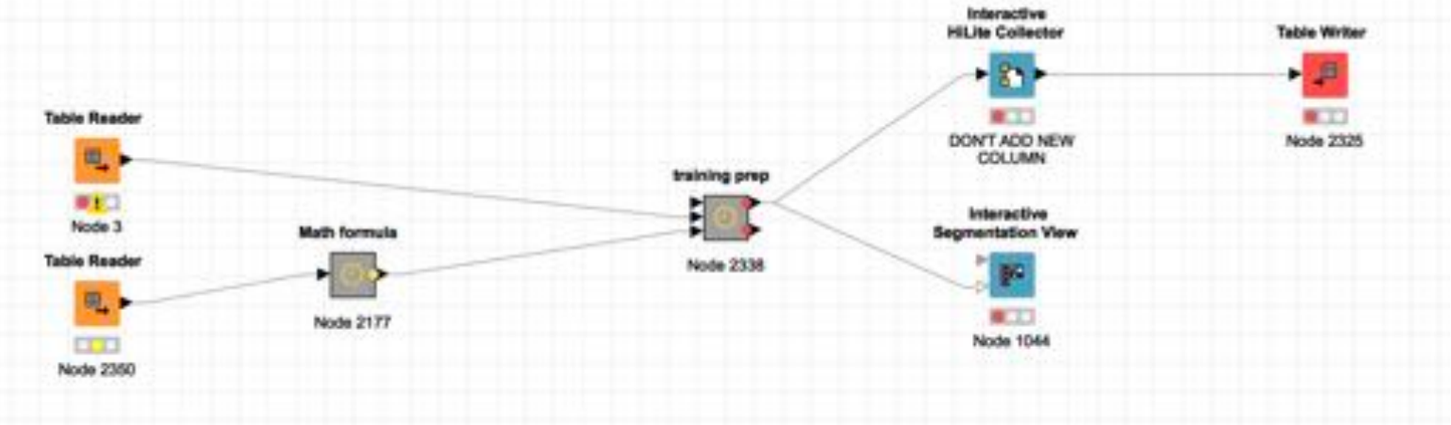


Production Workflows for Large Scale Image Analysis

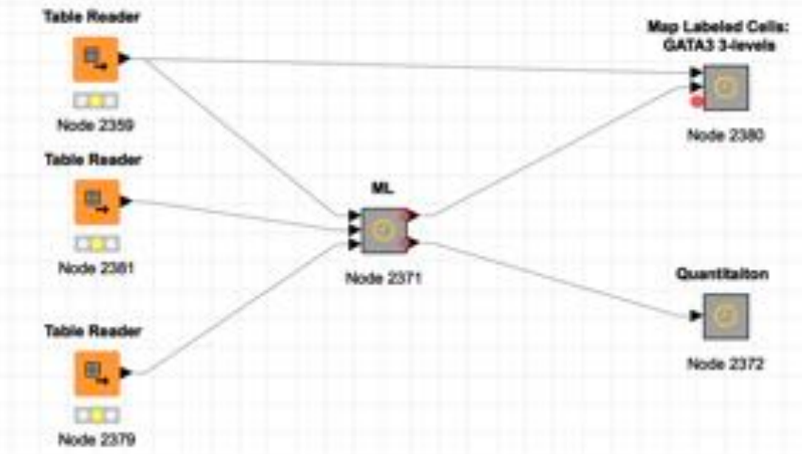
Image Segmentation

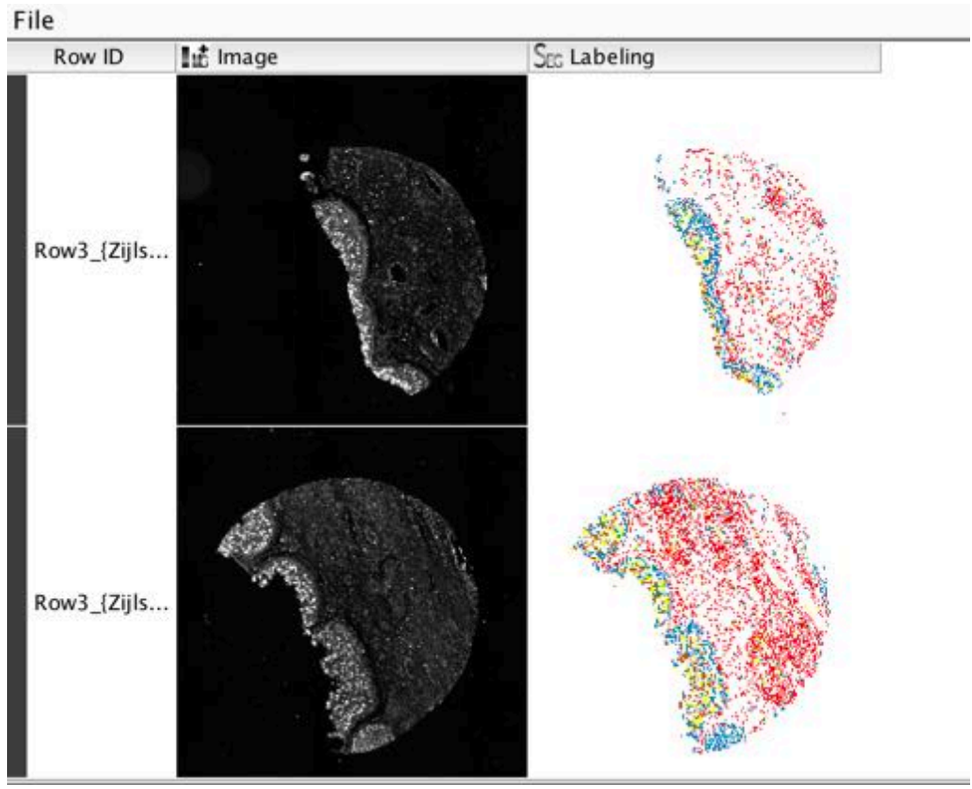


Training & Validation Set Selection



Machine Learning and Classification





Why go to the “Cloud”

- Make the processing of large images possible
- Accelerate analysis of any size image (factor of 10)
- Improve security
- Facilitate access by team members
- Enable remote access
- Facilitate monitoring
- Manage costs

Zijlstra Lab

Shanna Arnold-Egloff
Will Ashby
Adel Eskaros
Amanda Hansen
Joep Houkes
Celestial Jones-Paris
Charlotte Sandford-Sharp
Tatiana Ketova
Elizabeth Li
Ariana Von Lersner
Tatiana Novitskaya
Fabiane Fernandes
Chase Taylor
Lu Zheng

Collaborators

APCaRI
John Lewis
Robert Paproski
Konstantin Stoletov

Cedars Sinai
Dolores Di Vizio
Michael Freeman

Lasergen Inc.
Aparna Krishnan

KNIME
Christian Dietz

VUMC Collaborators
Peter Clark
Christina Derleth

VUIT Collaborators
Julie Catellier
Jacob Margason
Ed Wisdom

Epithelial Biology Core
Joseph Roland

Funding:
NIH/NCI Microenvironmental Influences in Cancer Training Program T32CA009592-24, NIH/NCI 1F31CA189764 (KH), NIH/NCI R01CA143081, R01CA218526, VA IK2BX002498, AIHS

NIH NATIONAL CANCER INSTITUTE