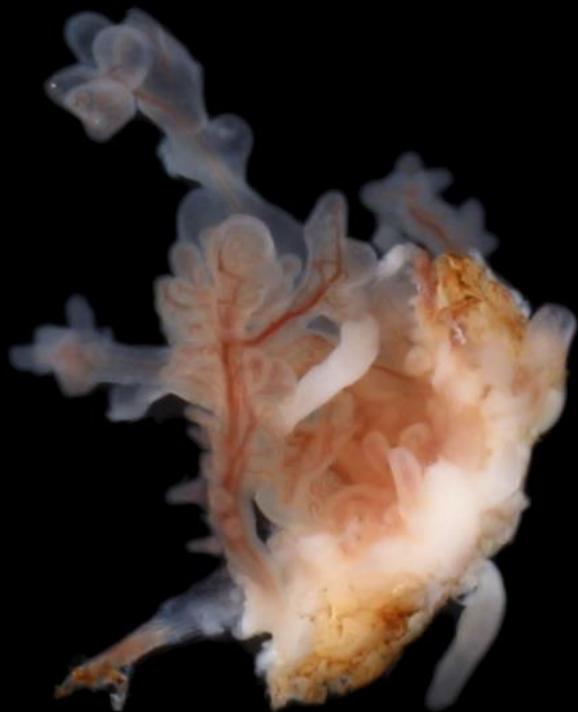


JPL-Caltech Virtual Summer School Big Data Analytics

September 2 – 12, 2014

Thomas Fuchs (JPL, Caltech)
Application: Cancer Research



Bladder



Kidney

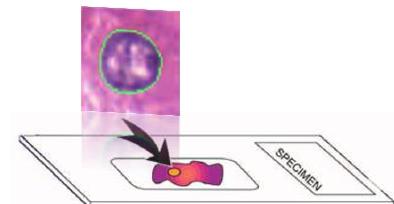


Lymphnodes

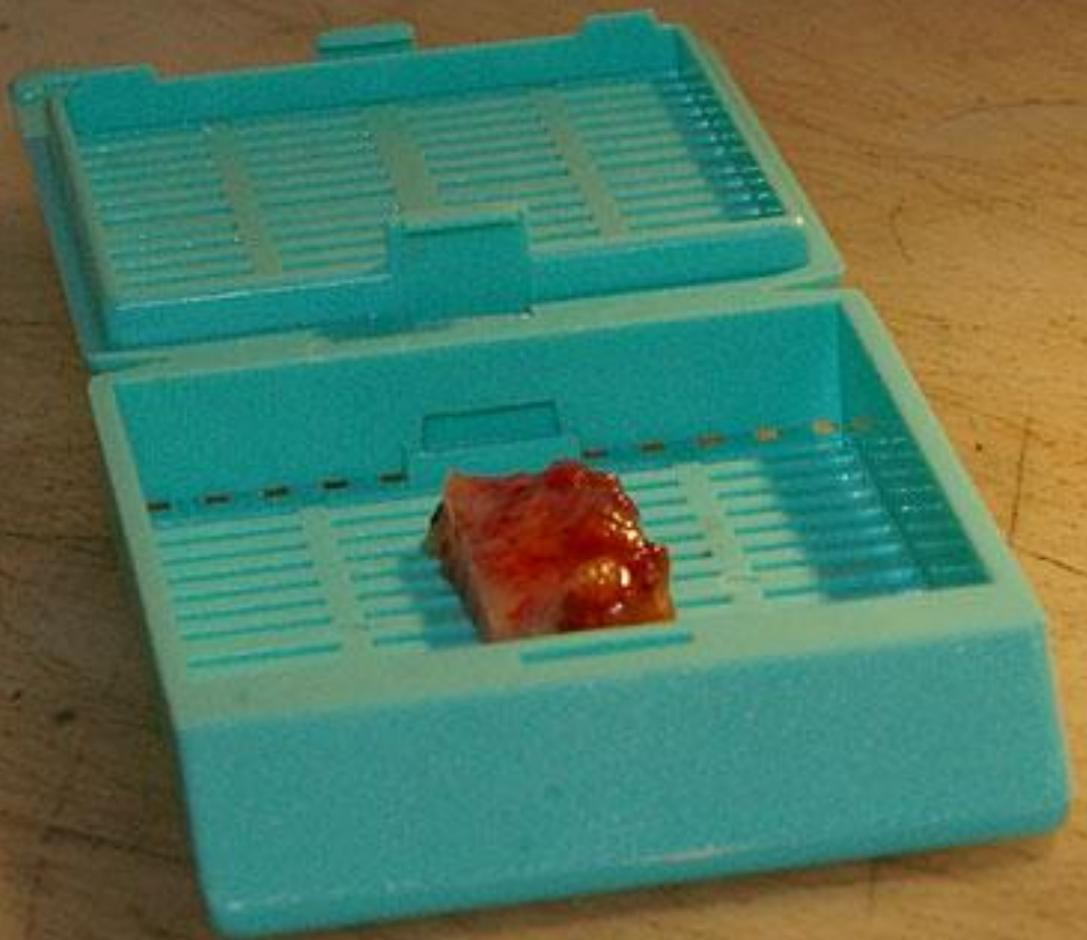
Definition

“**Computational Pathology** investigates a **complete probabilistic treatment** of scientific and clinical workflows in general pathology, i.e. it combines experimental design, statistical pattern recognition and survival analysis within an **unified framework** to answer scientific and clinical questions in pathology.”

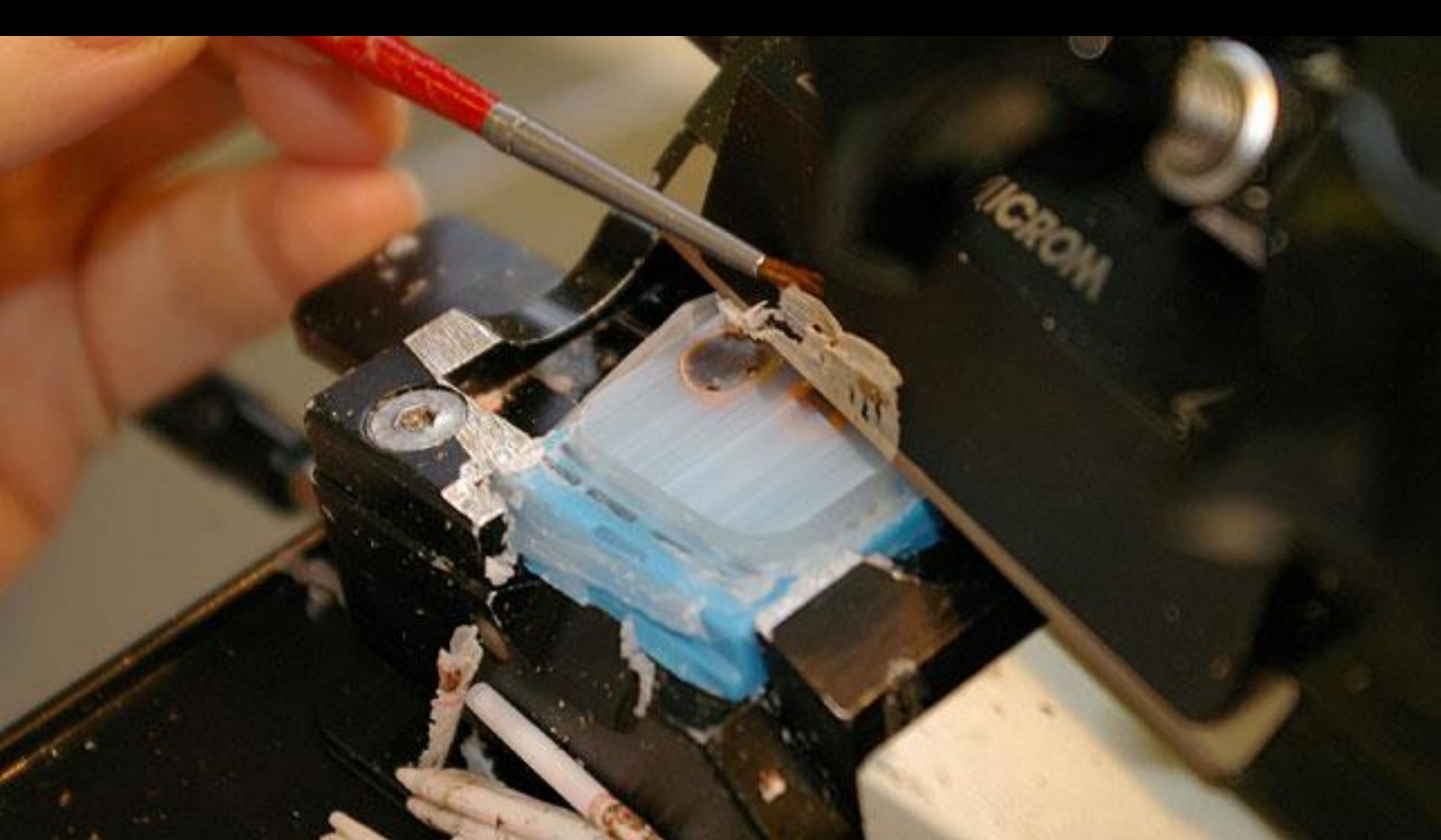
[CMIG 2011]

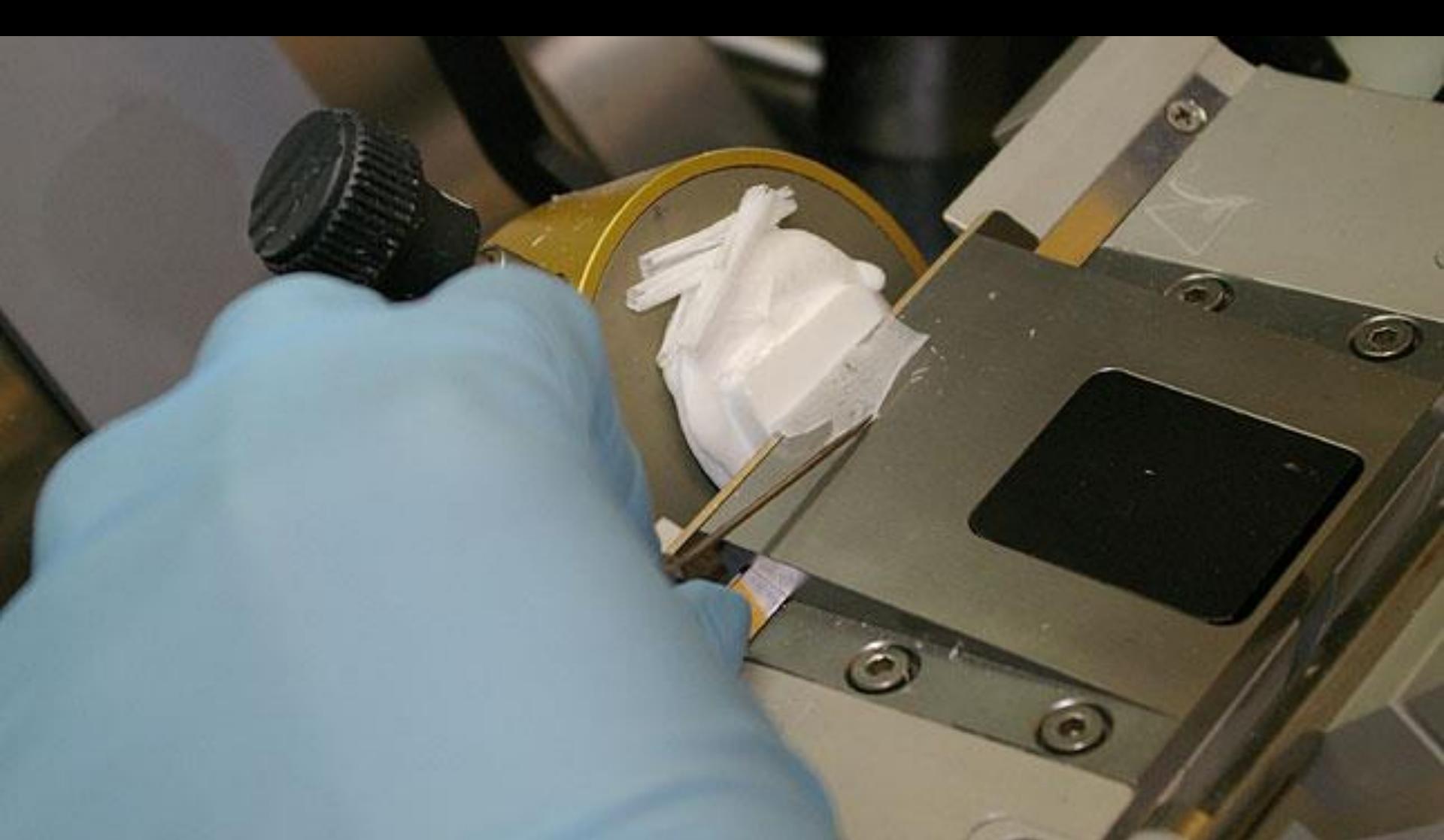


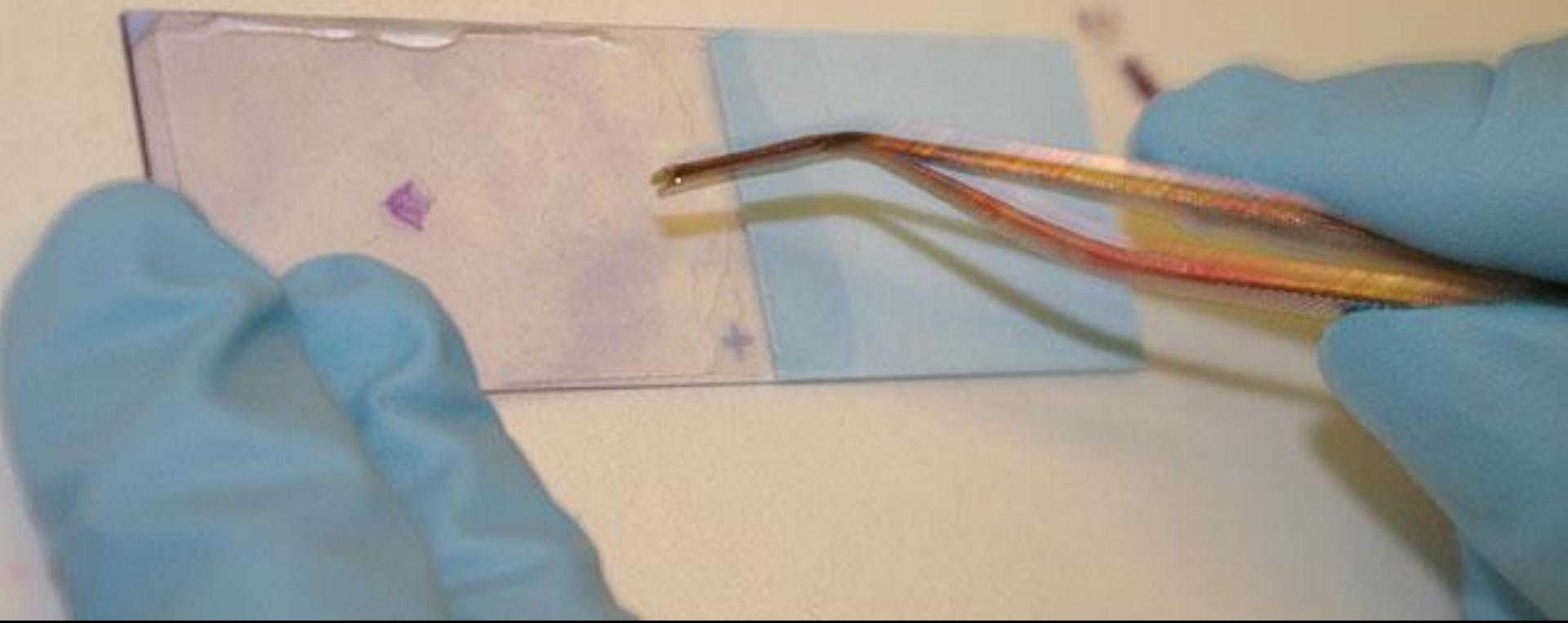














19486/1	19486/1	19486/1	19486/1	19486/1	19486/1
19486/1	19486/1	19486/1	19486/1	19486/1	19486/1
purple	purple	purple	purple	orange	orange
19510- 151	18510/8	18510/8	18510/8	18510/8	18510/8
E6	110	110	110	110	110





UniversitätsSpital
Zürich

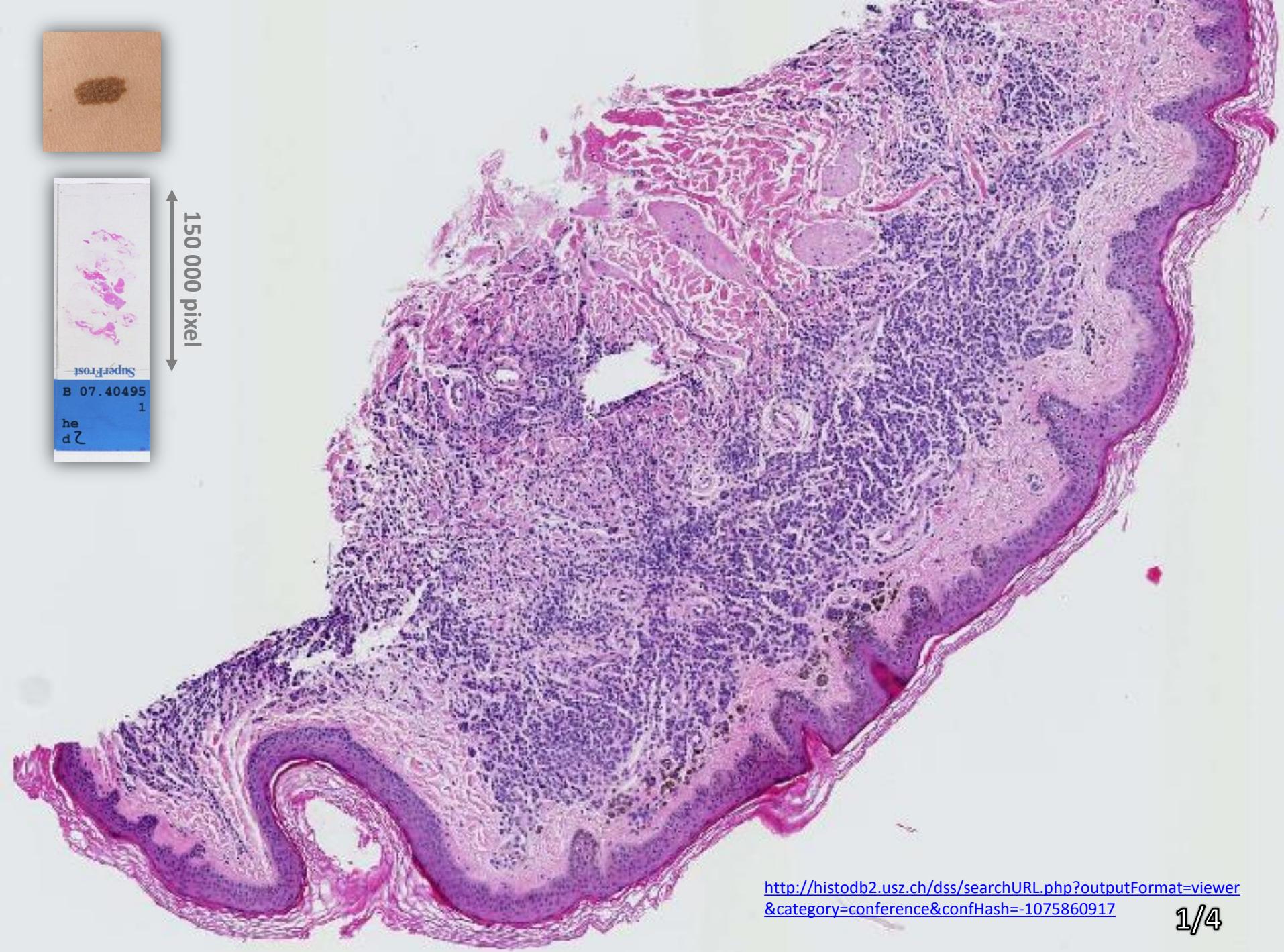


Departement
Pathologie

~67,000 cases per year





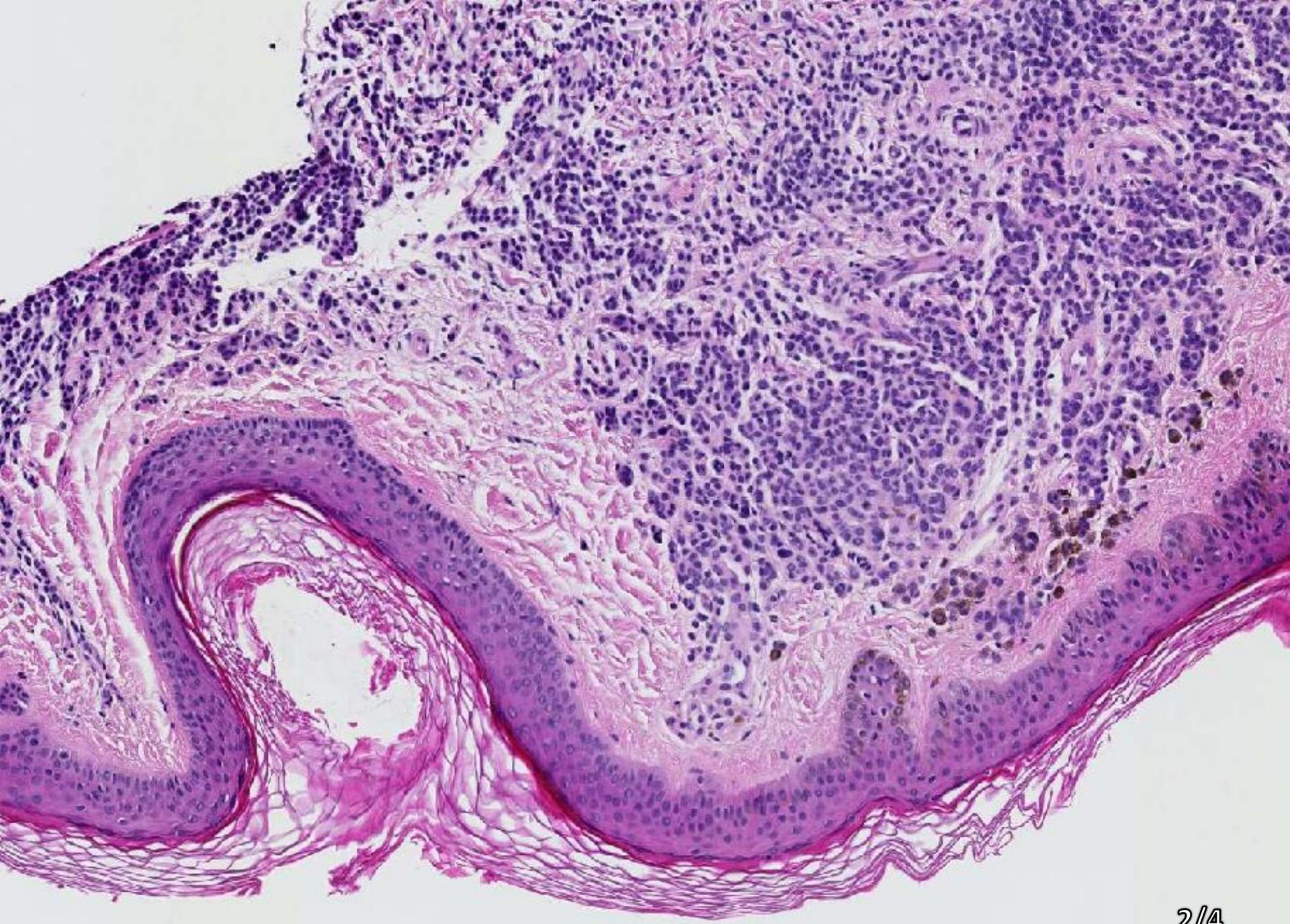


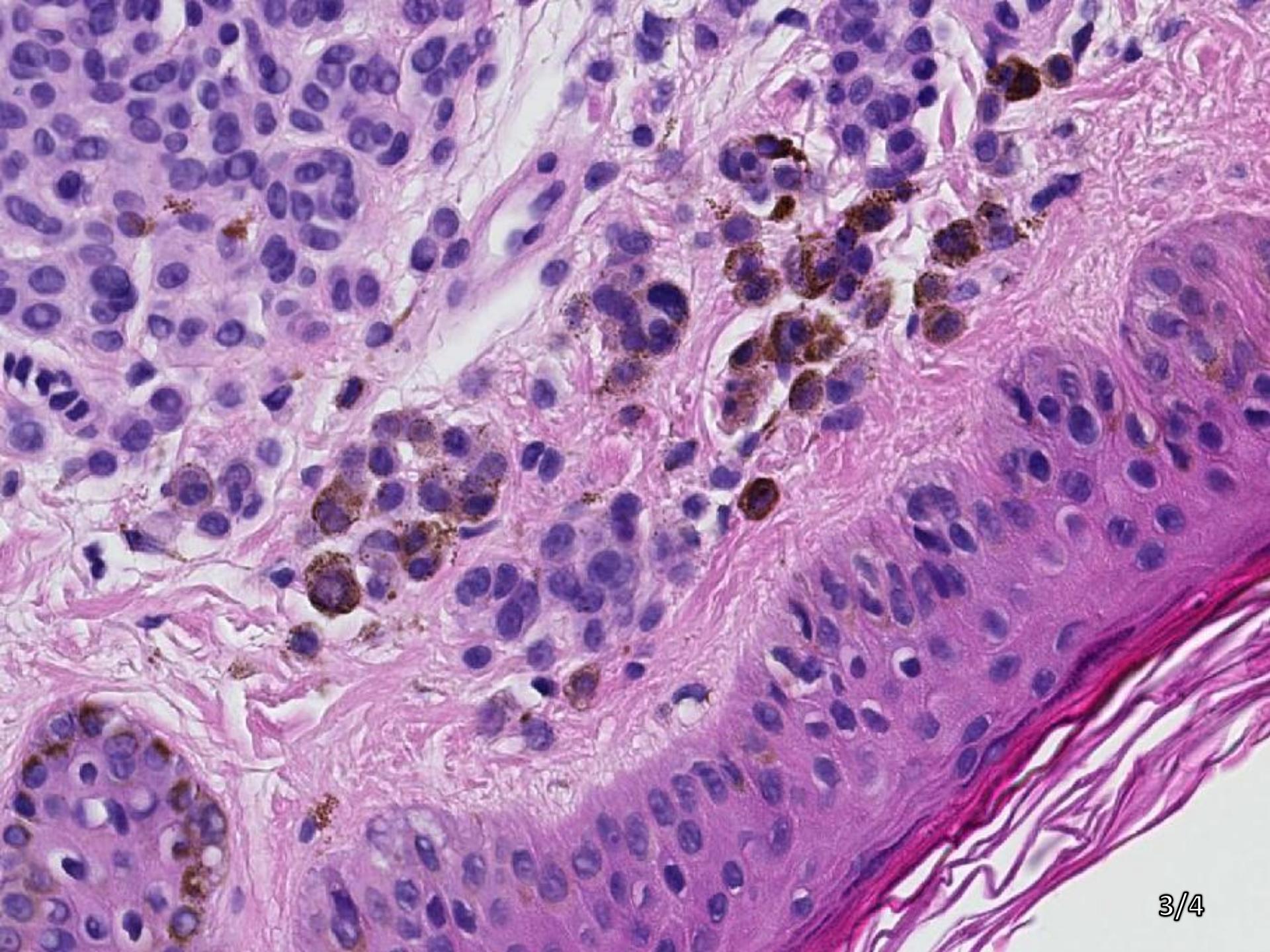
150 000 pixel



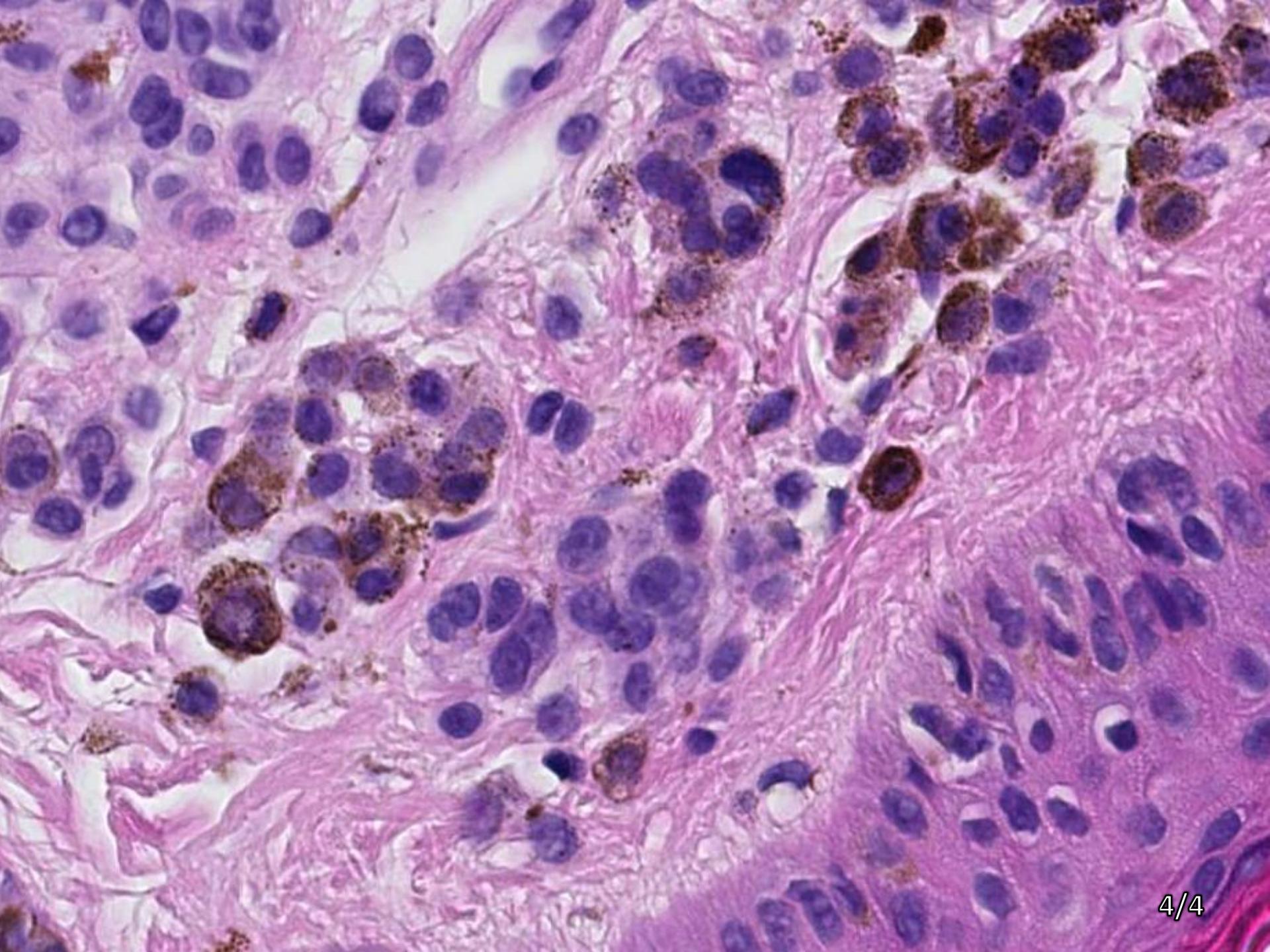
[http://histodb2.usz.ch/dss/searchURL.php?outputFormat=viewer
&category=conference&confHash=-1075860917](http://histodb2.usz.ch/dss/searchURL.php?outputFormat=viewer&category=conference&confHash=-1075860917)

1/4

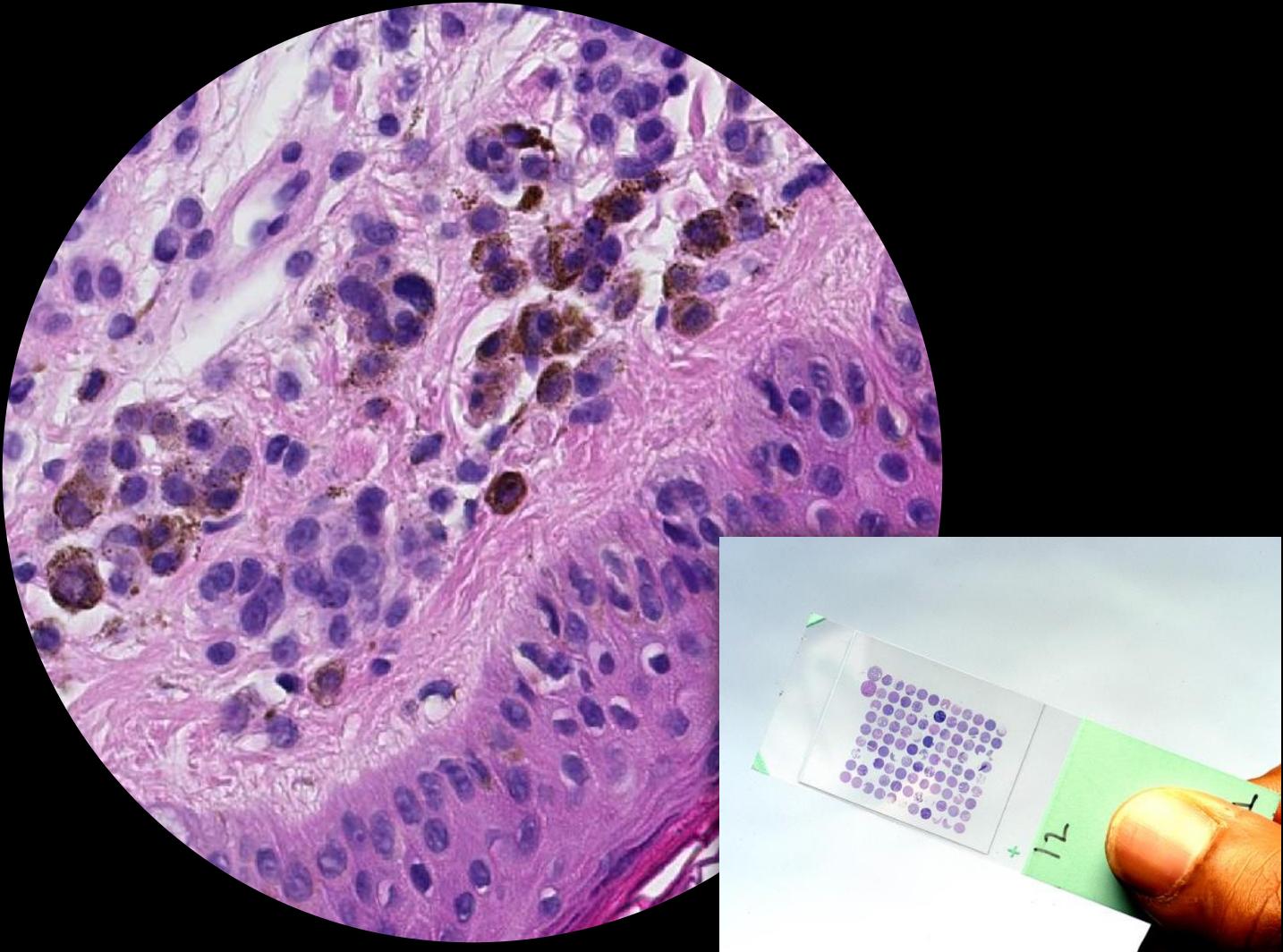




3/4

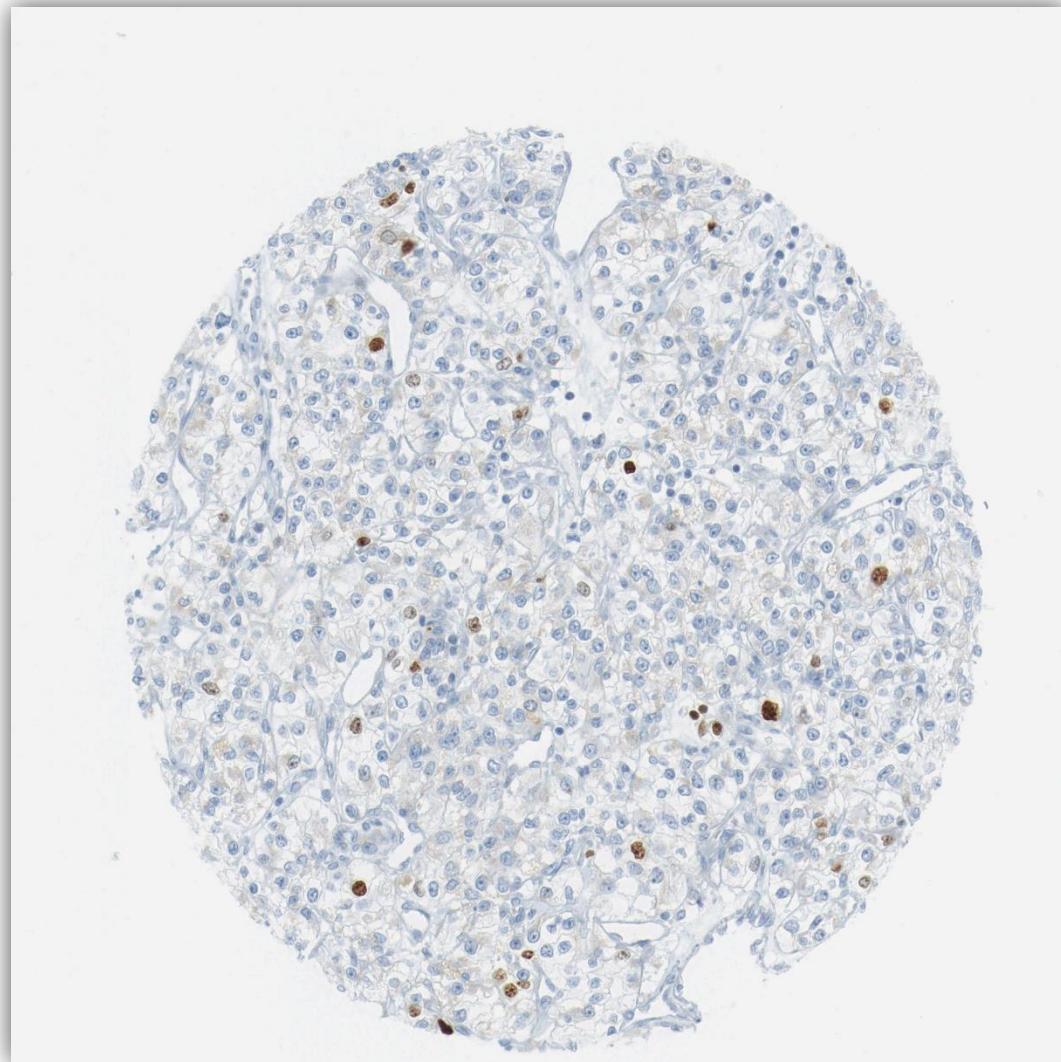
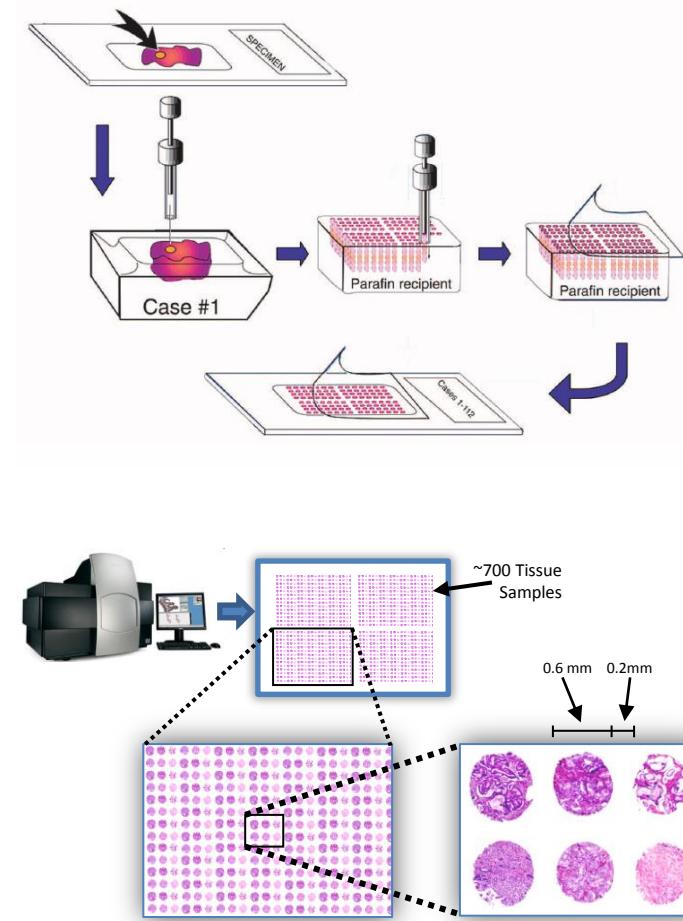


From Subjects to Cohorts



Tissue Microarray (TMA) Spot

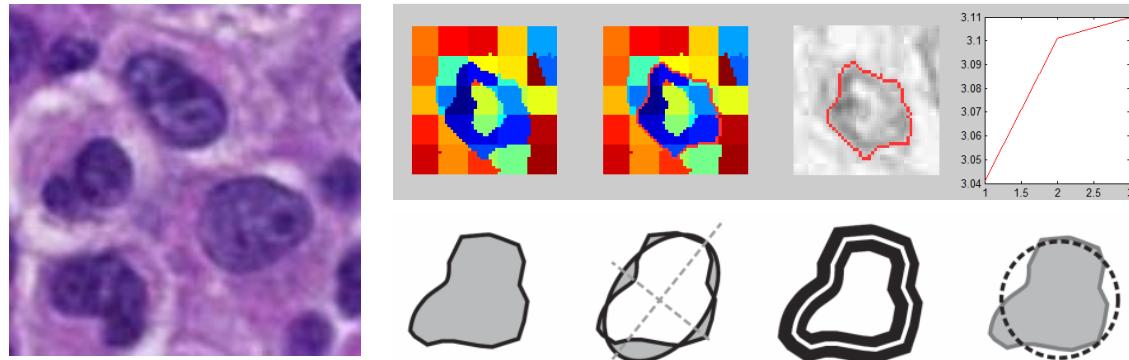
Tissue Micro Array (TMA)



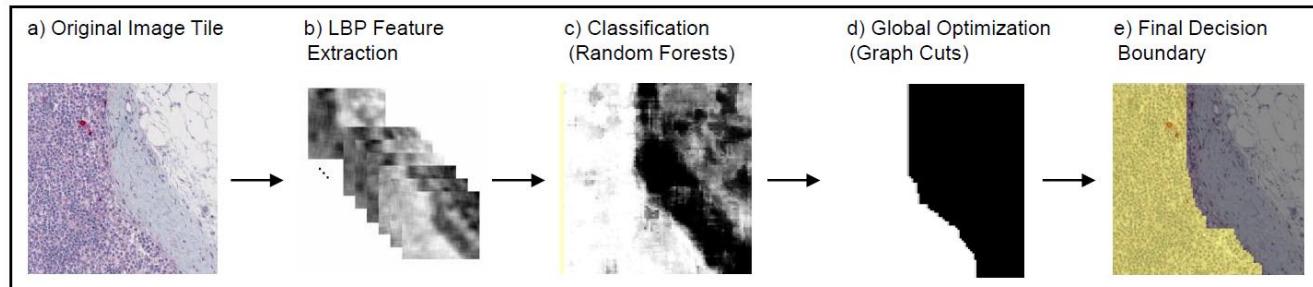
Computer Vision Tasks in Pathology

Nuclei Detection and Classification

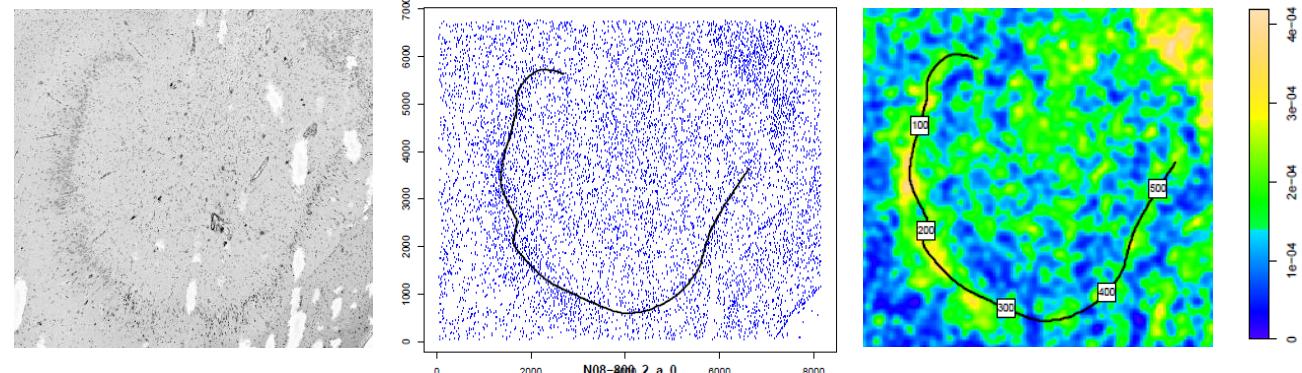
Sub-cellular level



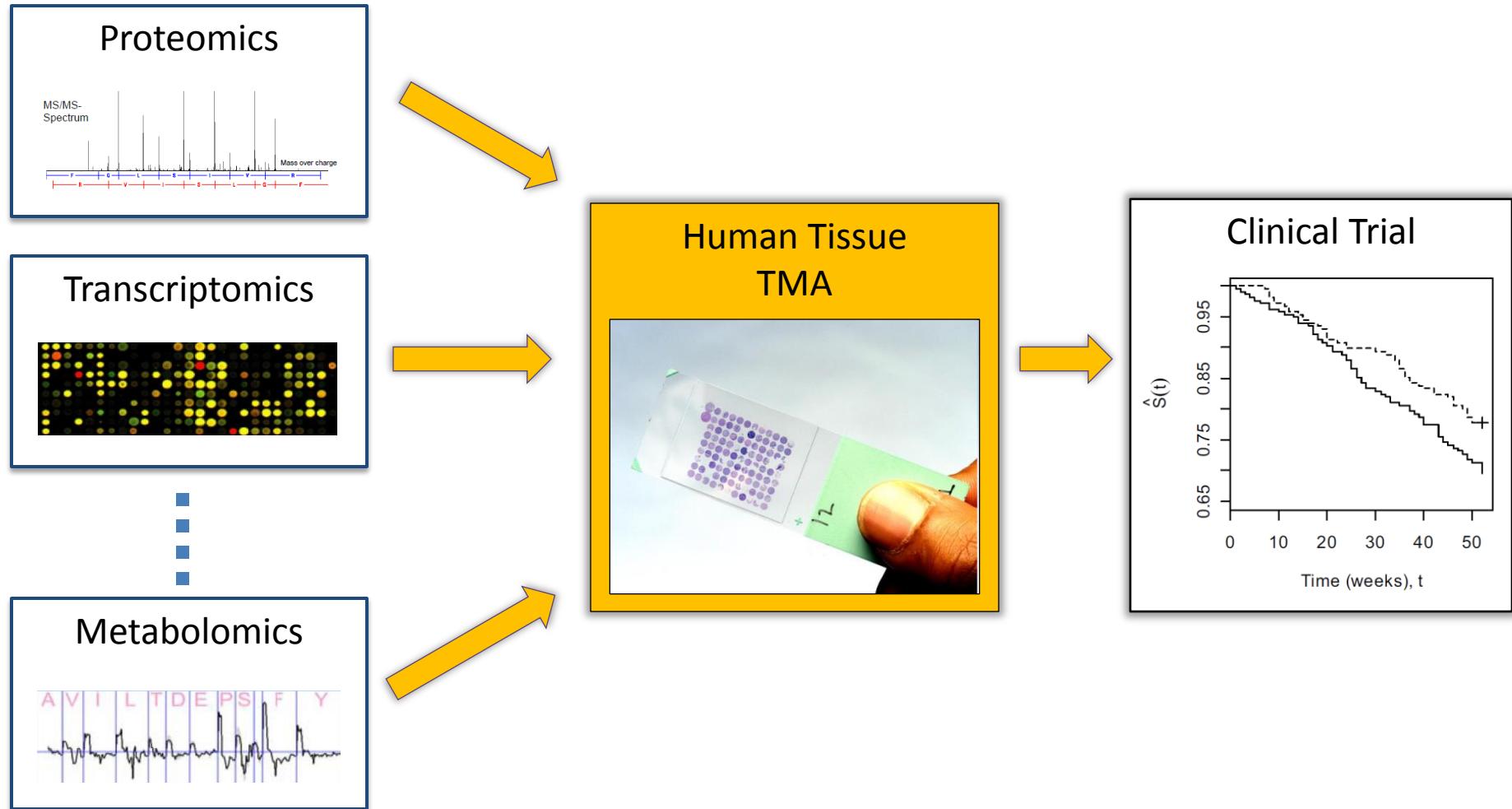
Segmentation



Structure Estimation Morphology

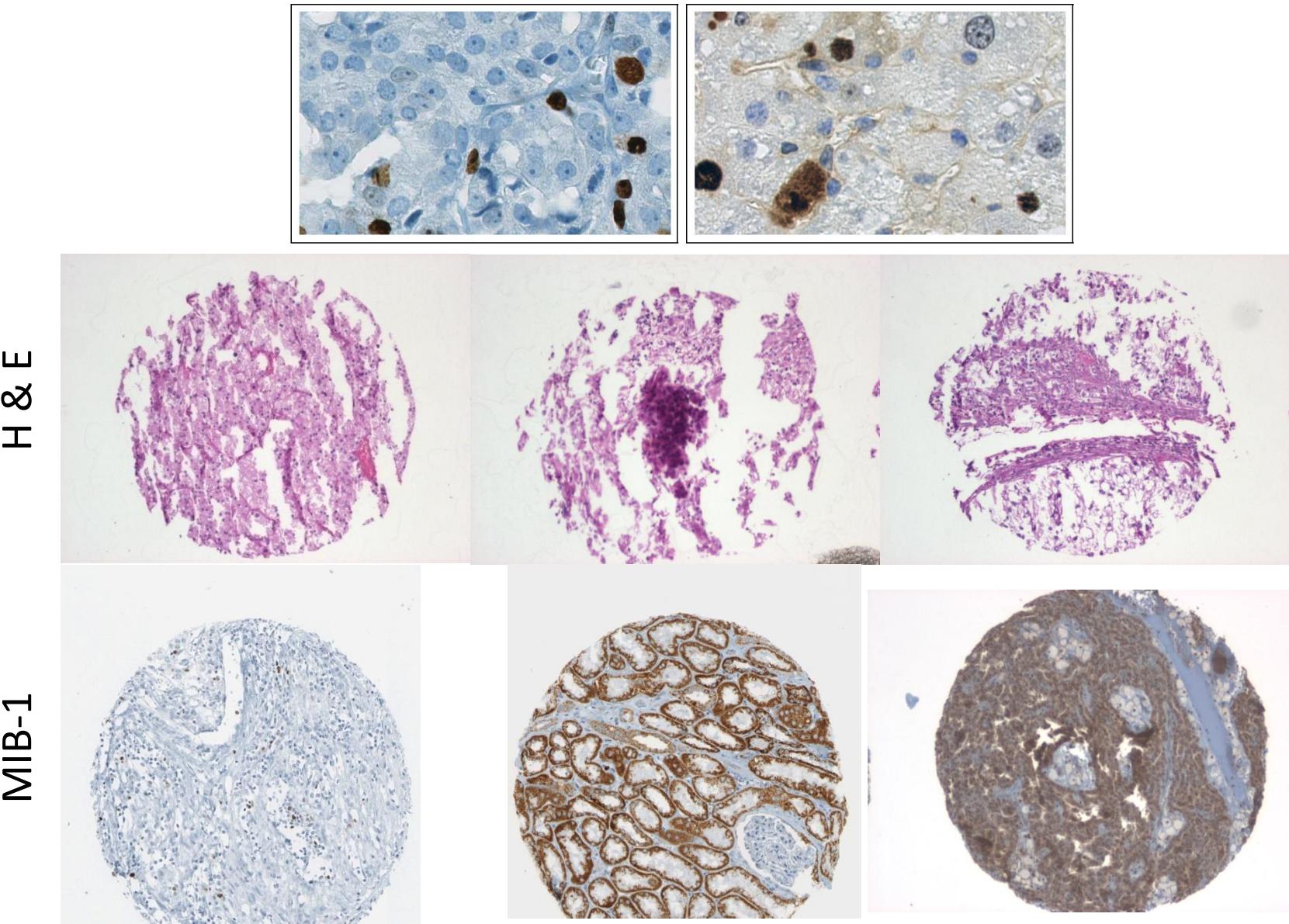


Biomarker Detection & Validation



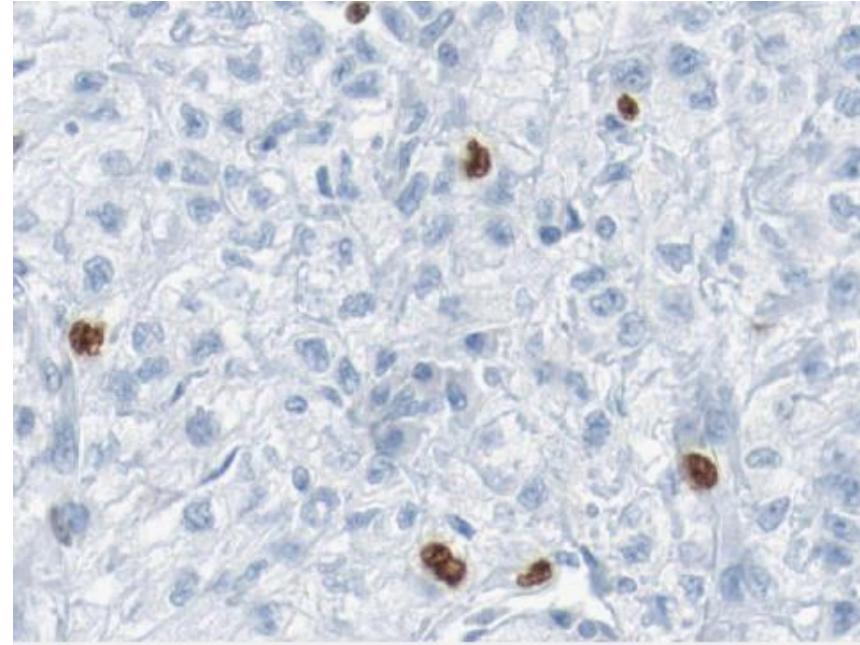
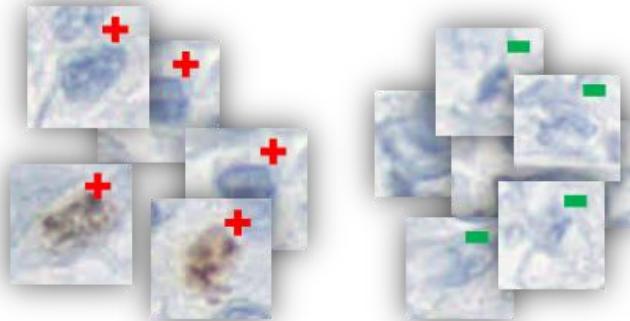
Variability

MIB-1



Ground Truth for Statistical Learning

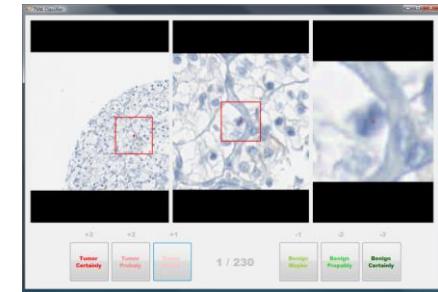
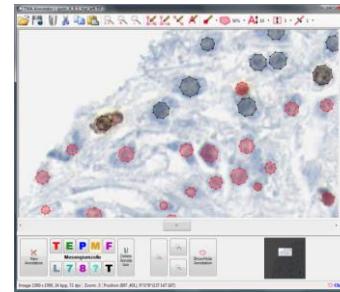
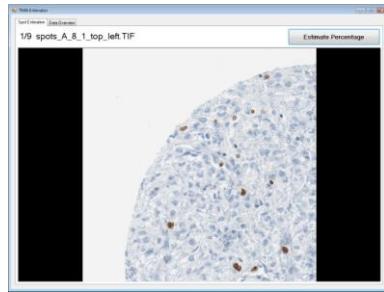
Labeled samples are needed
for training and validation.



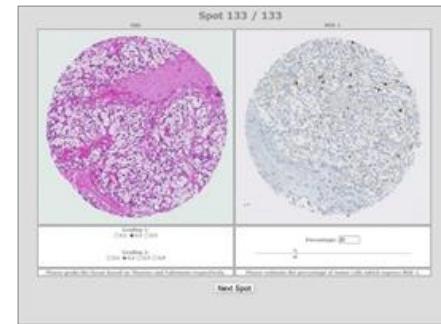
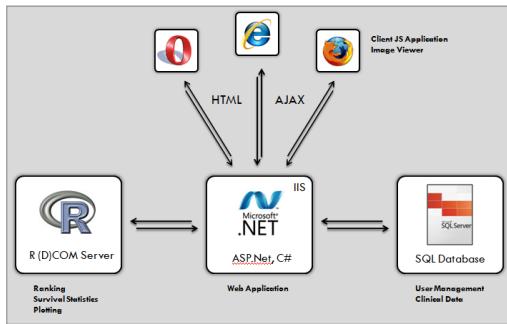
What is the „Ground Truth“?

Expert & Crowd Sourcing

Past

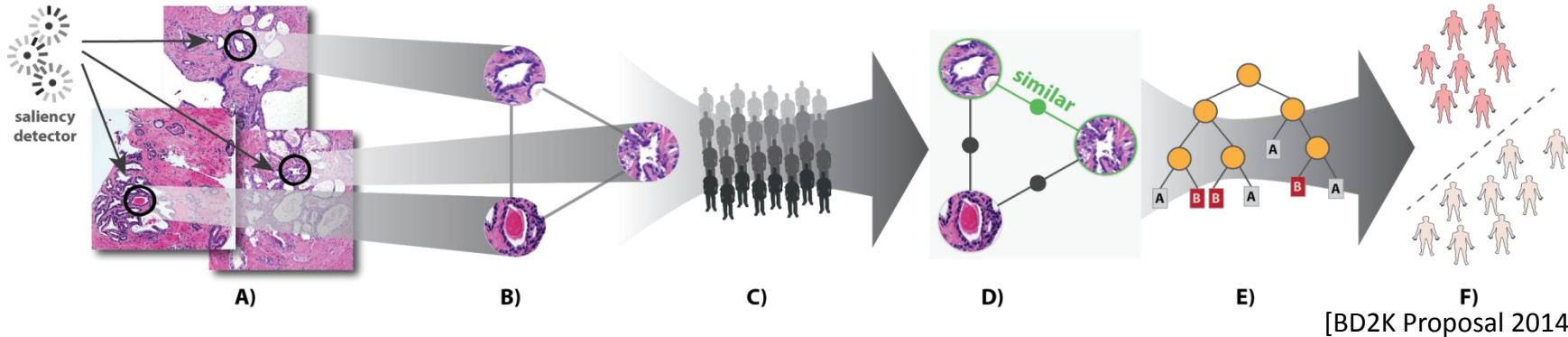


Present



Rank	Thinner	Follicles	Proliferation Factor
1.			AlGPAX $p = 0.329$
2.			johida $p = 0.481$
3.		AlGPAT $p = 0.482$	johida $p = 0.482$
4.	AlGPAT $p = 0.482$	johida $p = 0.482$	johida $p = 0.482$
5.	johida $p = 0.482$	johida $p = 0.482$	AlGPAT $p = 0.482$
6.	johida $p = 0.482$	johida $p = 0.482$	johida $p = 0.482$
7.	johida $p = 0.482$	johida $p = 0.482$	johida $p = 0.482$

Future



[BD2K Proposal 2014]

Staining Estimation Pipeline

5/5 pathologists agreed on +3



283	302	739	1104	1454	1501	1689
4	3	3	3	3	3	3

4/5 pathologists agreed on ± 3

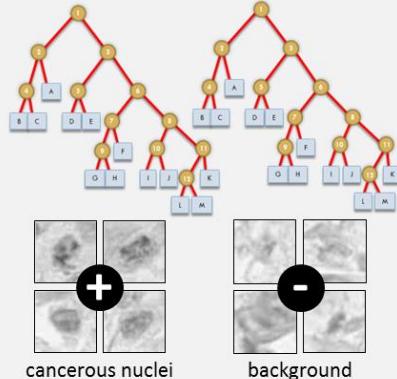


138	209	386	605	1635	1936	2052	2125	2337
2	2	2	2	2	2	2	2	2

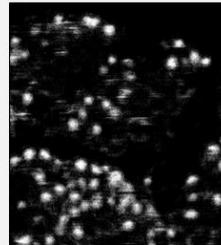
Image Analysis

Learning

Relational Detection Forest

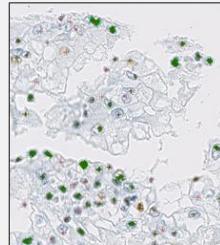


Classifying every pixel of a spot results in a probability map



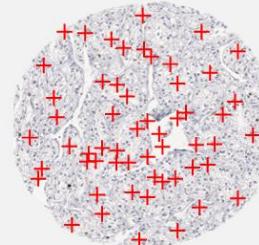
Prediction

Nuclei centers are found by applying mean shift clustering

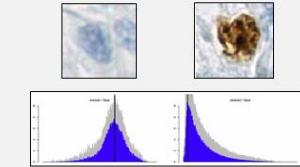


Estimation

Several hundred nuclei are detected on each Image of a TMA spot



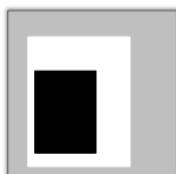
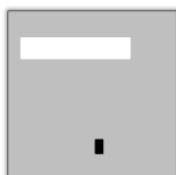
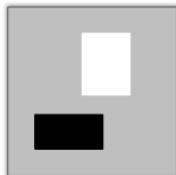
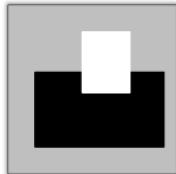
Overall staining per patient is Calculated by assessing the staining of all detected nuclei



x%



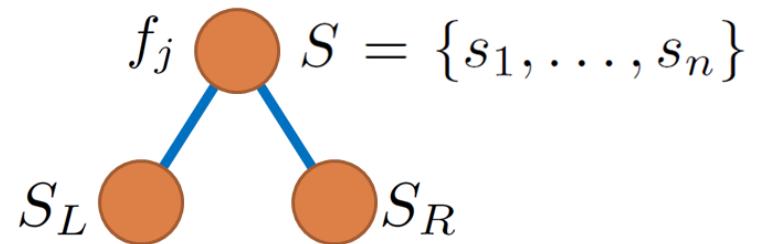
Relational Detection Forest



Procedure LearnTree

Input: set of samples $S = \{s_1, s_2, \dots, s_n\}$
Input: depth d
Input: max depth d_{max}
Input: features to sample $mTry$

- 1 **Init:** $\widehat{label} = null; g = -\inf$
- 2 **Init:** $N_{left} = null; N_{right} = null$
- 3 **if** ($d = d_{max}$) OR ($isPure(S)$) **then**
- 4 $\widehat{label} = \arg \max_{l \in \{true, false\}} \sum_{i|s_i=l} 1$
- 5 **else**
- 6 **for** $i = 0, i < mTry, i + +$ **do**
- 7 $f_i = \text{SampleFeature}()$
- 8 $S_L = \{s_j | f_i(s_j) = true\}$
- 9 $S_R = \{s_j | f_i(s_j) = false\}$
- 10 $g_i = \widehat{\Delta G}(S_L, S_R)$
- 11 **if** $g_i > g$ **then**
- 12 $f_{best} = f_i; g = g_i$
- 13 **end**
- 14 **end**
- 15 $N_{left} = \text{LearnTree}(\{s_i | f_{best}(s_i) = true\})$
- 16 $N_{right} = \text{LearnTree}(\{s_i | f_{best}(s_i) = false\})$
- 17 **end**



Gini Index:

$$\widehat{G}(S) = 2 \frac{N_{false}}{|S|} \left(1 - \frac{N_{false}}{|S|} \right)$$

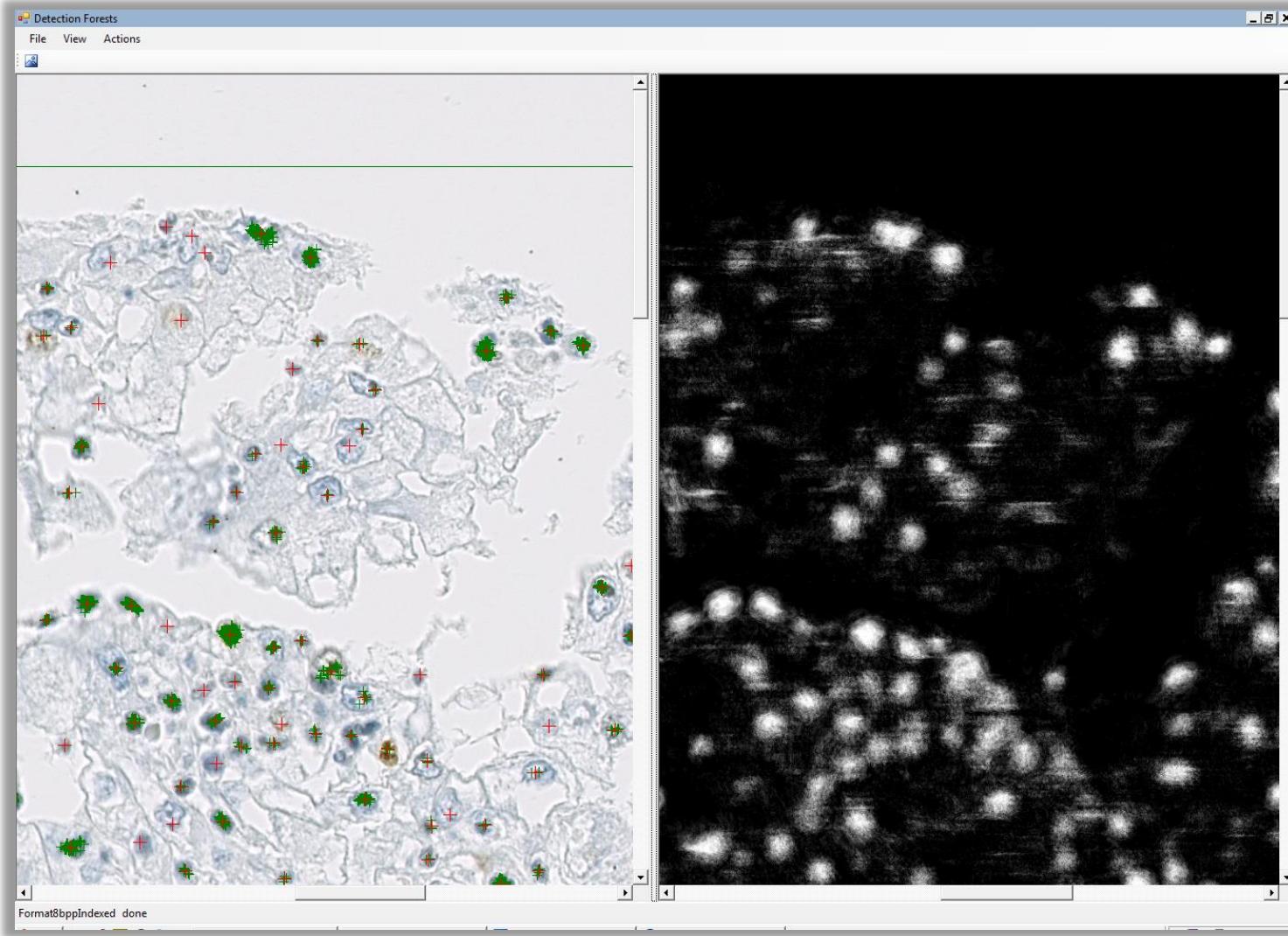
$$N_{false} = \sum_{s_i} I(f_j(s_i) = false)$$

Gini Gain:

$$\widehat{\Delta G}(S_L, S_R) = \widehat{G}(S) - \left(\frac{|S_L|}{|S|} \widehat{G}(S_L) + \frac{|S_R|}{|S|} \widehat{G}(S_R) \right)$$

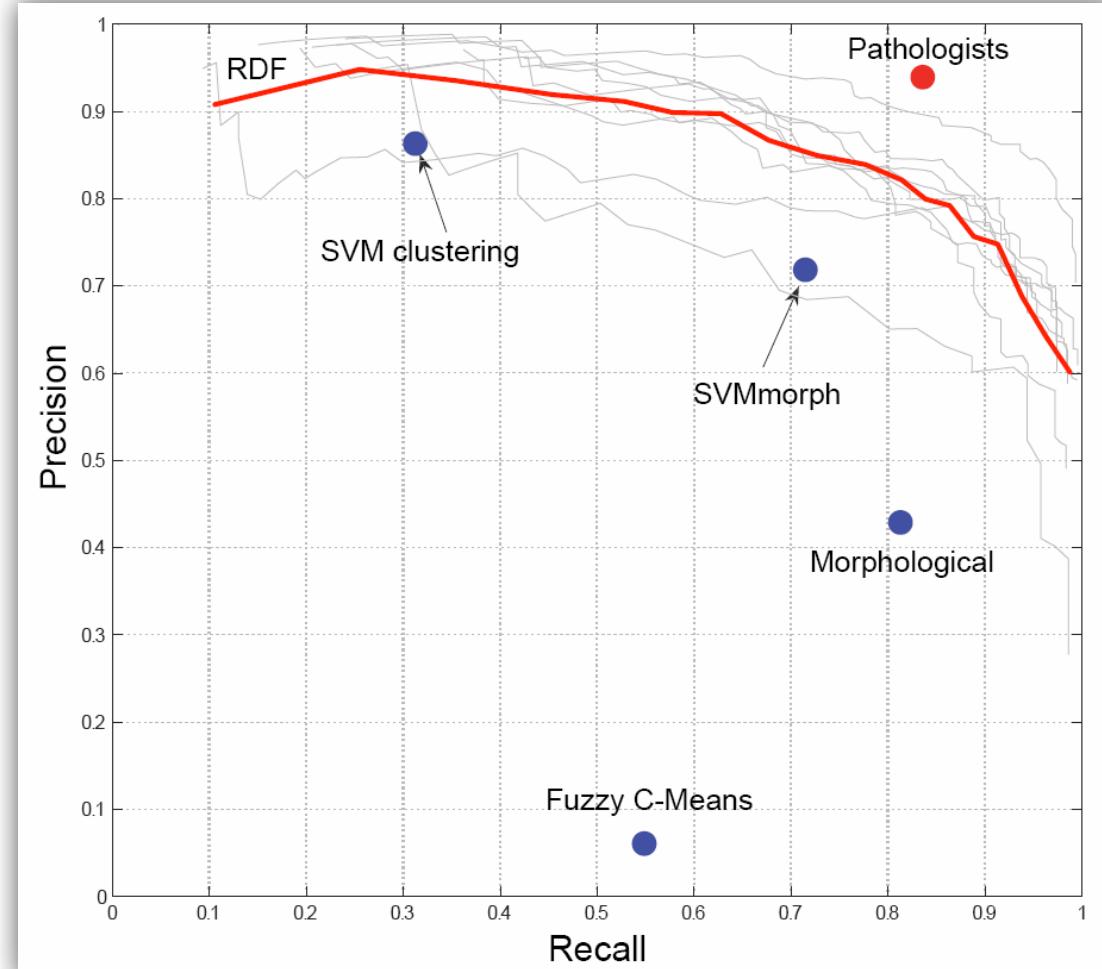
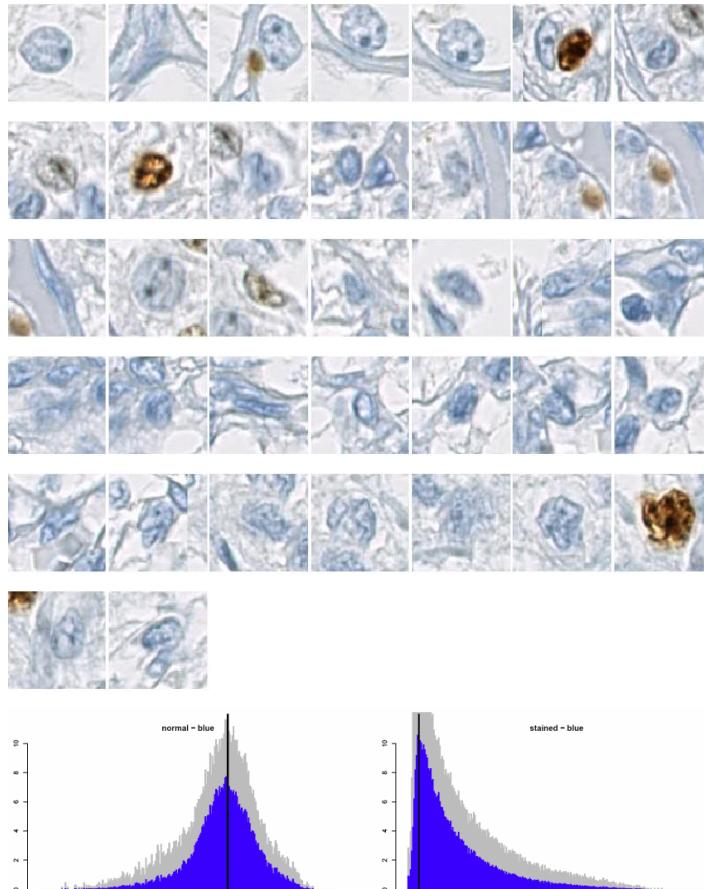


Relational Detection Forest





Relational Detection Forest



Nucleus Based Analysis

$$\text{NucleusIntensity}(n) = \frac{1}{|n|} \sum_{x \in n} x$$

$$\text{InnerIntensity}(n) = \frac{1}{|n|} \sum_{x \in [n \setminus \epsilon_B(n)]} x$$

$$\text{OuterIntensity}(n) = \frac{1}{|n|} \sum_{x \in [\delta_B(n) \setminus n]} x$$

$$\text{InnerHomogeneity}(n) = \text{std}(x \in [n \setminus \epsilon_B(n)])$$

$$\text{OuterHomogeneity}(n) = \text{std}(x \in [\delta_B(n) \setminus n])$$

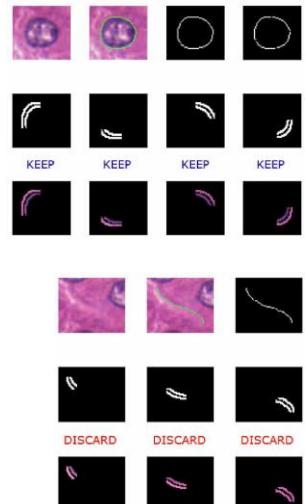
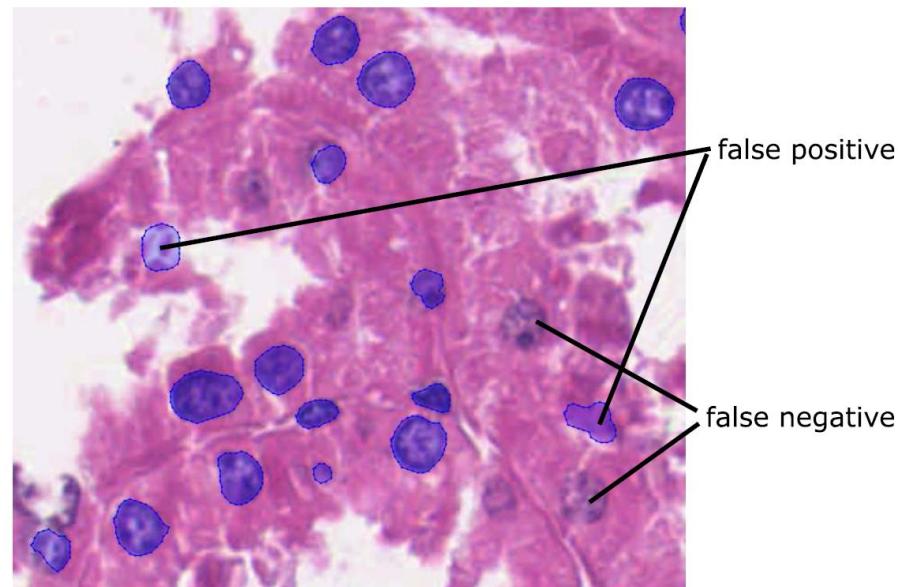
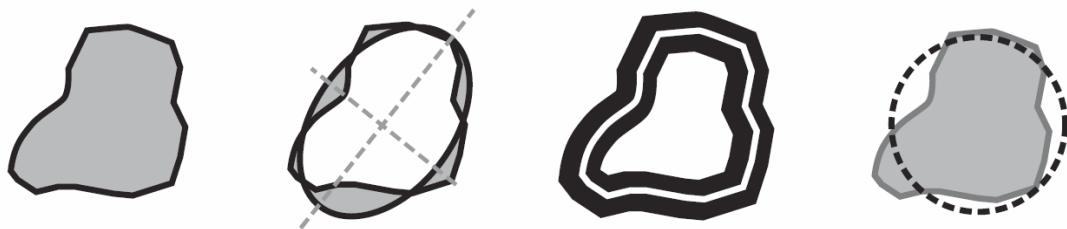
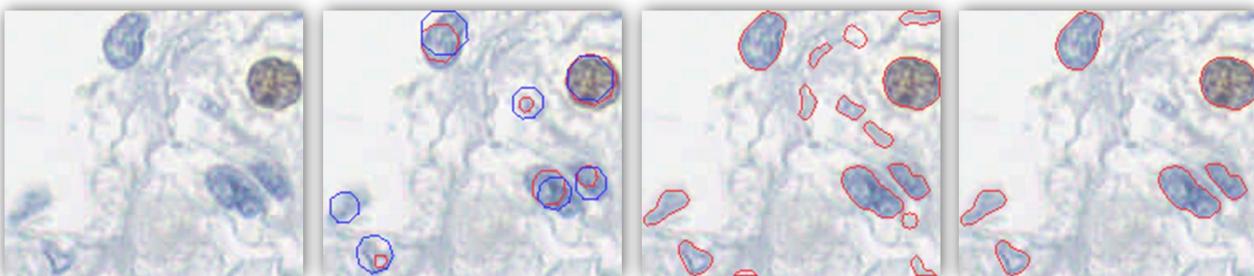
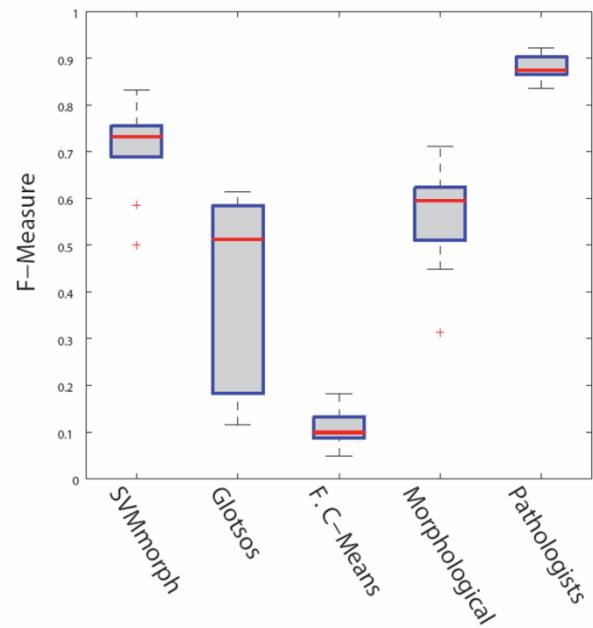
$$\text{IntensityDifference}(n) = \frac{1}{|n|} \sum_{x \in (\delta_B(n) \setminus n)} x \cdot \left(\frac{1}{|n|} \sum_{x \in (n \setminus \epsilon_B(n))} x \right)^{-1}$$

$$\text{Size}(n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(|n| - \mu)^2}{2\sigma^2}\right), \quad \mu = 600, \quad \sigma = 300$$

$$\text{Ellipticity}(n) = \frac{|n_{\text{Ellipse}}|}{|n_{\text{Ellipse}}| + |(n_{\text{Ellipse}} \cup n) \setminus (n_{\text{Ellipse}} \cap n)|}$$

$$\text{ShapeRegularity}(n) = \frac{2\pi\sqrt{\frac{n_{\text{Area}}}{\pi}}}{n_{\text{Perim}}}$$

$$F = 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall}).$$



Computational Pathology

Data Generation

Image Aquisition: X

Sample preparation and TMA construction. MIB-1 staining on clear cell renal cell carcinoma

Slide scanning and tiling of TMA into spots

Label Aquisition: Y

Gold standard: samples of nuclei via labeling experiments

Background objects through Voronoi sampling

Training samples: cancerous nuclei
background

Image Analysis

Learning

Relational Detection Forest

+ cancerous nuclei
- background

Prediction

Classifying every pixel of a spot results in a probability map

Nuclei centers are found by applying mean shift clustering

Estimation

Several hundred nuclei are detected on each Image of a TMA spot

Overall staining per patient is Calculated by assessing the staining of all detected nuclei

x%

Survival Statistics

Patient Cohort

Application to TMA spots of 133 RCC patients.

MIB-1 estimations

Estimation (%)	Number of Patients (%)
0.4%	12.1%
0.8%	17.4%
1.5%	23.6%
12.1%	1.5%
17.4%	0.8%
23.6%	0.4%

Staining Estimation

Estimation from the domain expert and prediction from the algorithm for each patient in the cohort

Density
N = 133 Bandwidth = 1.01

Pathologist
Algorithm

Subgroup Analysis

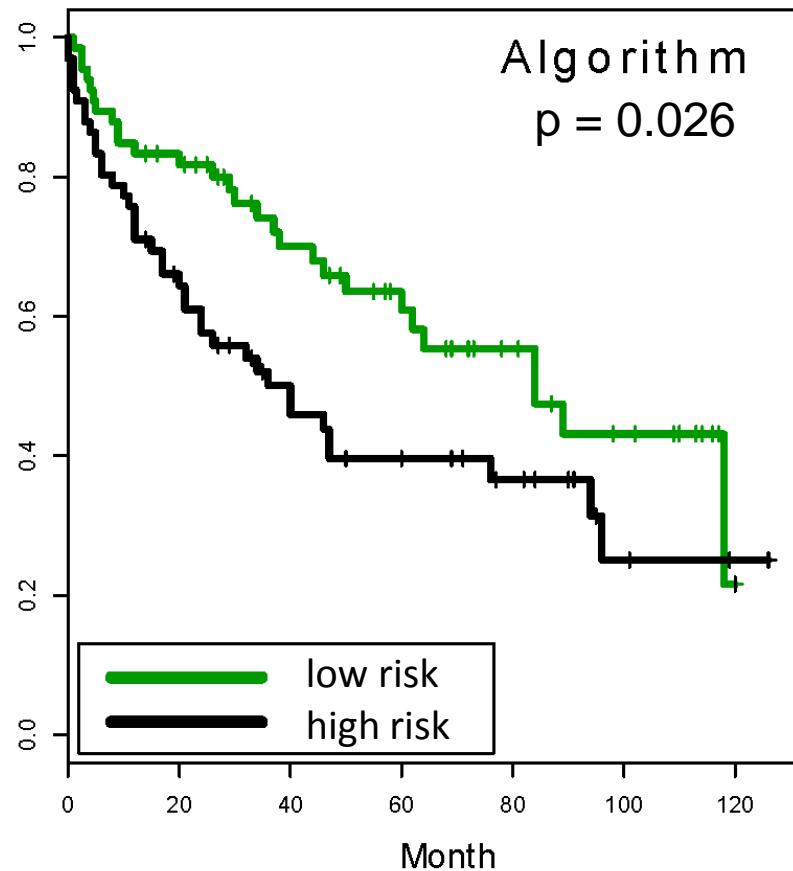
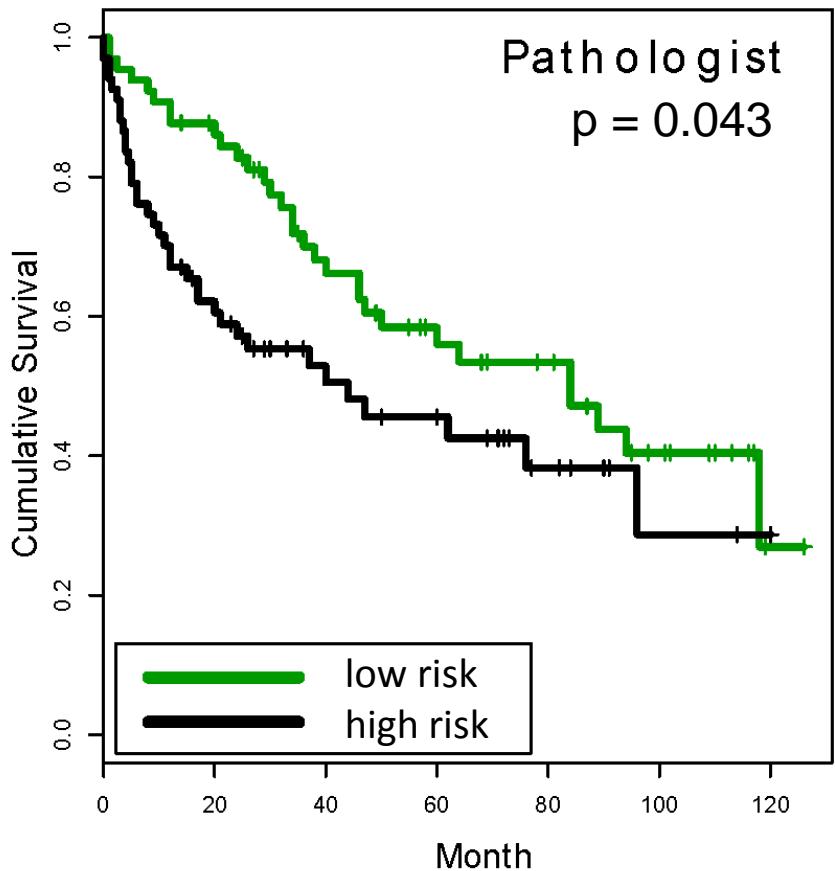
Kaplan-Meier estimators for subgroups of patient with high and low MIB-1 expression.

Cumulative Survival
Month

Pathologist p = 0.043
Algorithm p = 0.026

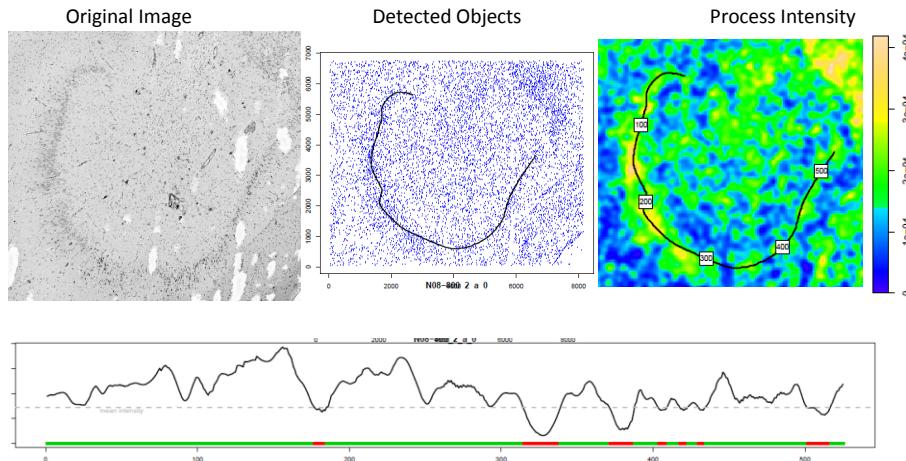
1st Quantile
2nd Quantile

Survival Analysis

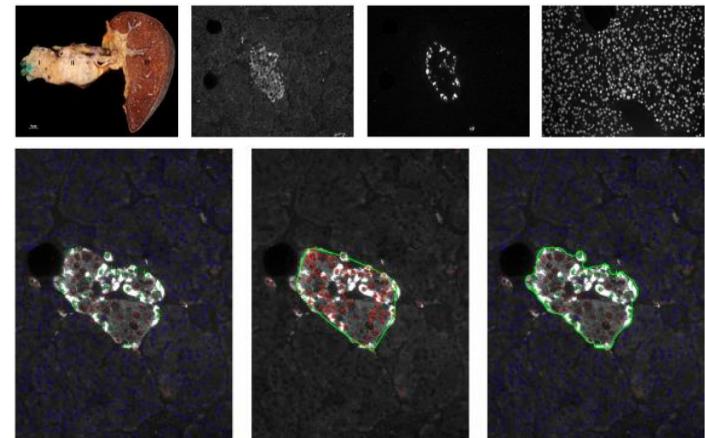


Applications of the Framework

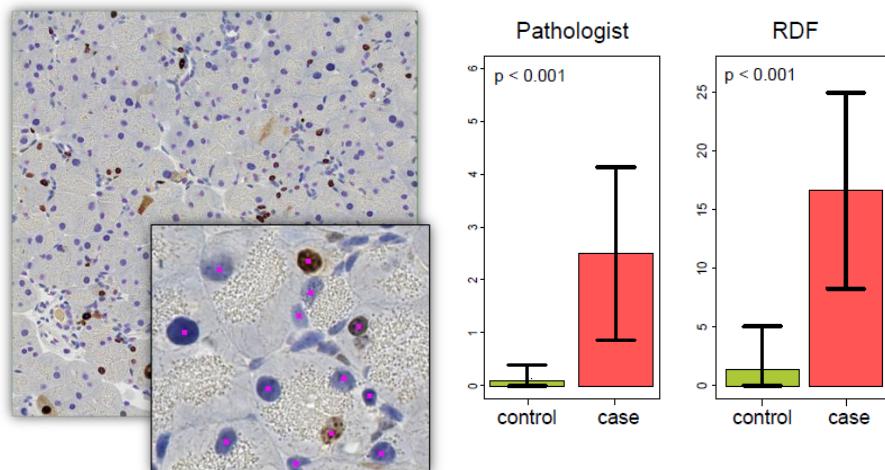
Spatial Processes for Hippocampal Sclerosis



Pancreatic Islet Segmentation for T2 Diabetes



Counting of Mouse Liver Hepatocytes



Detection in IHC Stained Cell Cultures

