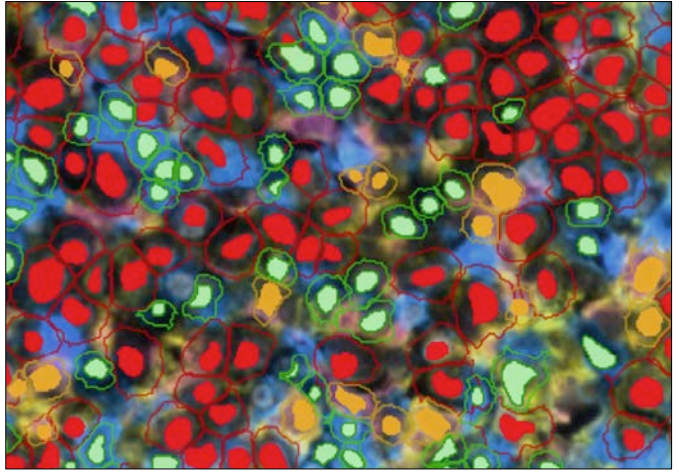
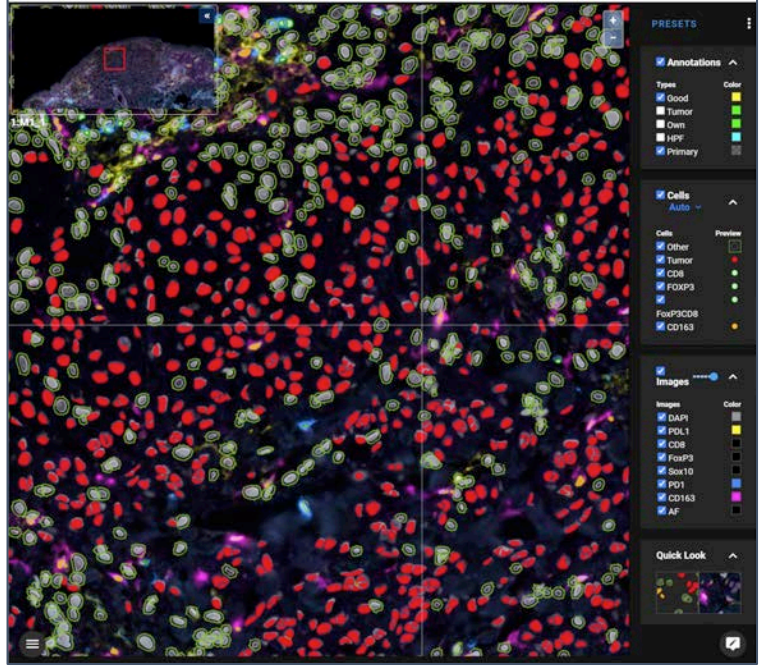




AstroPath:
Astronomy Meets Pathology

Alex Szalay, Janis Taube
and the AstroPath Team

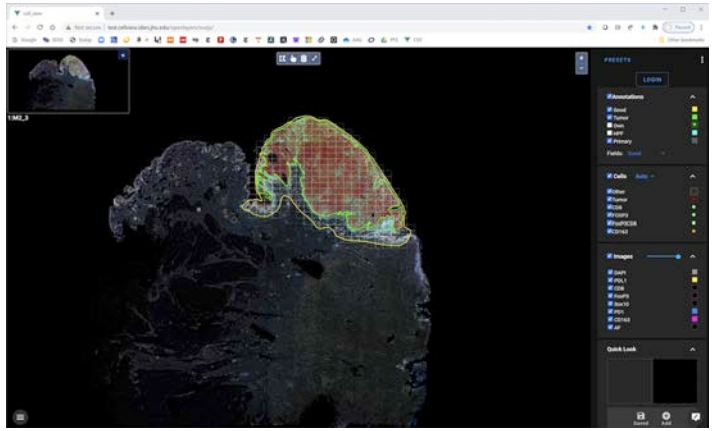
in collaboration with Akoya Biosciences

Institute for Data Intensive Engineering and Science
idies

AKOYA BIOSCIENCES

JOHNS HOPKINS MEDICINE

AstroPath: Atlas of Cancer Cells

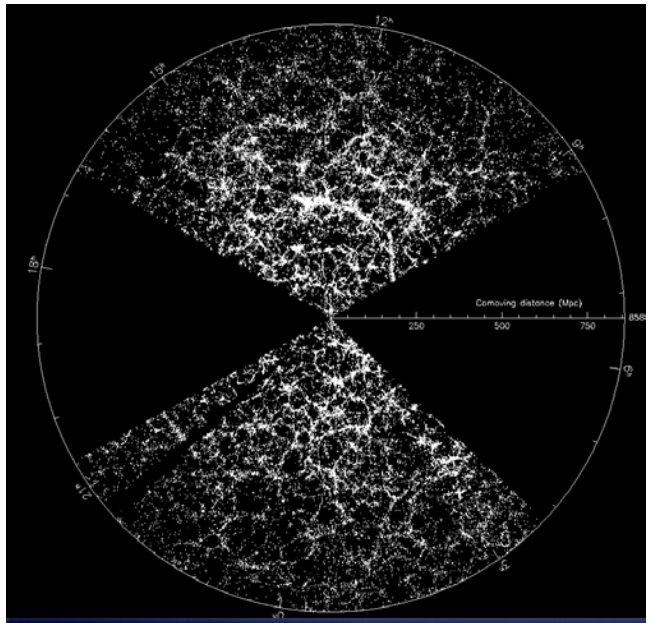


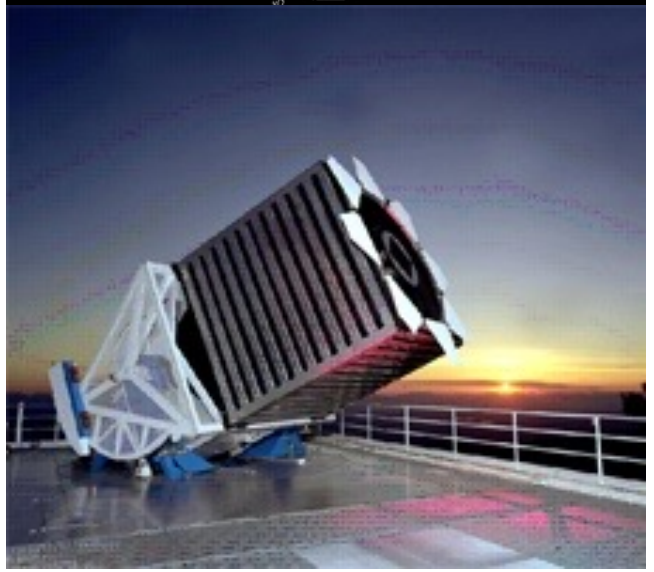
- **Astro**nomy meets **Path**ology
  - Project started by Prof. Janis Taube (JHMI BKI) and Alex Szalay (JHU IDIES)
- Studying the tumor microenvironment to understand cancer immunotherapy
  - Spatial interactions of activated T cells and tumor near the tumor boundaries
- Transitioning to the “industrial revolution”
- Goal: increase data collection by a factor of >1,000
  - 400GB mosaic of 35-band multiplex images/slide (from 10 to 2000 images/slide)
  - 7 markers (lineage + PD-1, PD-L1), more markers via additional panels
  - Use a farm of automated microscopes => 2PB/year
  - Heavy use of parallel processing
- Parallels sky surveys (as of 20 years ago)
  - “Disruptive assistance” from astronomy to pathology
  - Using techniques astronomers learned the hard way (flat field, unwarp, calibrate)
- Tumor boundaries, cell geometries represented as GIS polygons
- Dynamic computation of nearest neighbors, spatial relations
- Interactive viewer like the SkyServer, or Google Maps
- Processing workflows mostly automated
- Working on validating a large enough training set for Deep Learning
- Databases linked to SciServer, collaborative Jupyter, Keras/TensorFlow, R
- Collaboration with Akoya BioSciences (microscopes)

Sloan Digital Sky Survey

"**The Cosmic Genome Project**"

- Started in 1992, finished in 2008
- Data is public
  - 2.5 Terapixels of images => 5 Tpx of sky
  - 10 TB of raw data => 100TB processed
  - 0.5 TB catalogs => 35TB in the end
- Database and spectrograph built at JHU (SkyServer)
- Now SDSS-4, data served from JHU

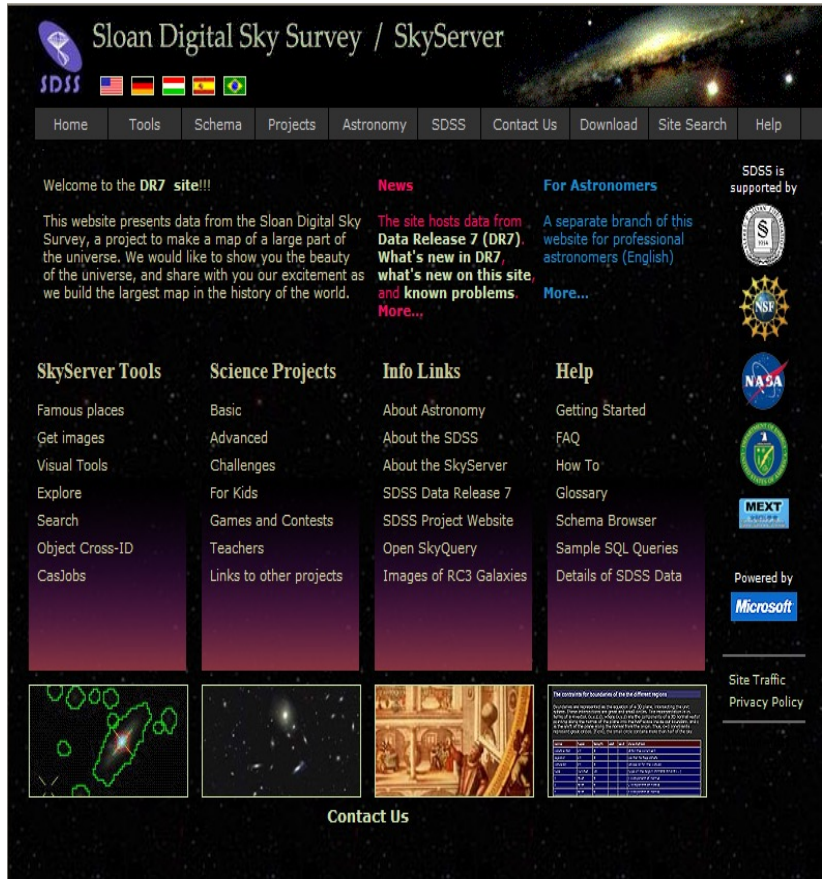


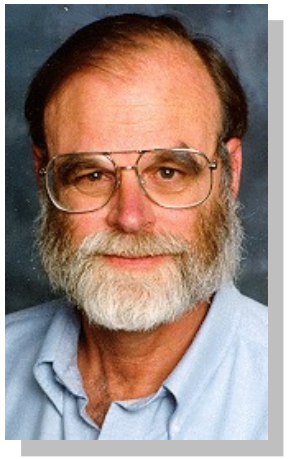


Skyserver

Prototype in 21st Century data access

- 2.8B web hits in 16 years
- 414M external SQL queries
- 7,000 refereed papers and 450K citations
- 5,000,000 distinct users vs. 15,000 astronomers
- The emergence of the "Internet Scientist"
- The world's most used astronomy facility today
- Collaborative server-side analysis done by 9,000 astronomers
- Morphed into the SciServer recently





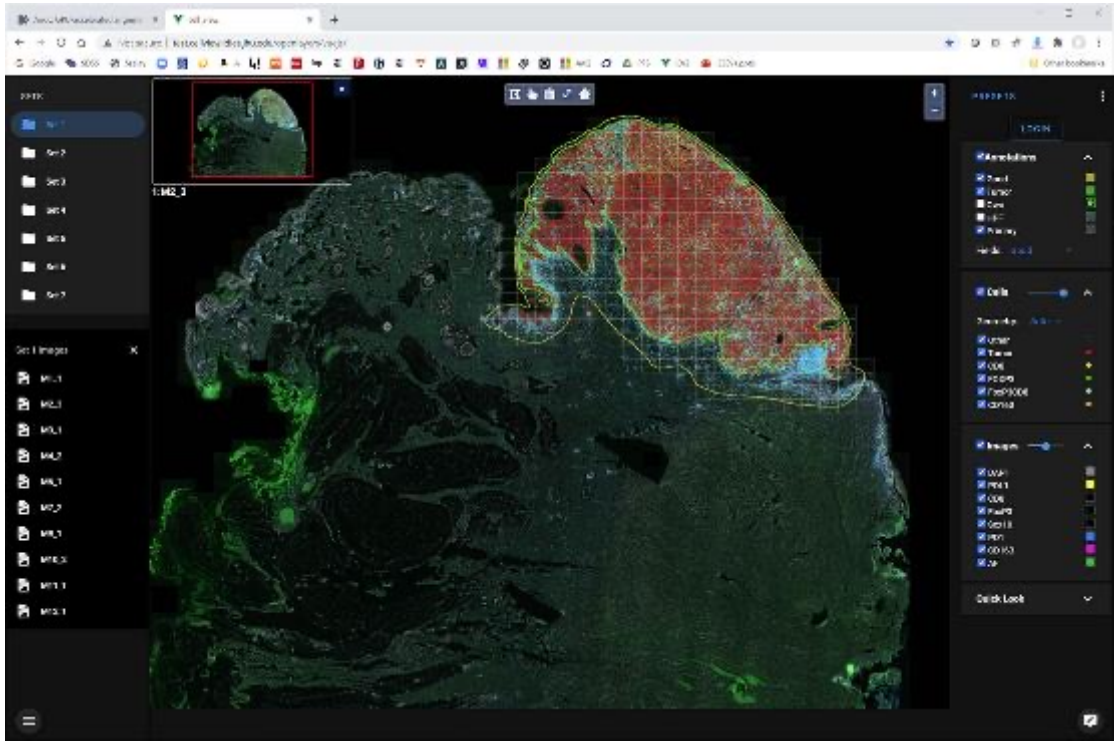
Jim Gray

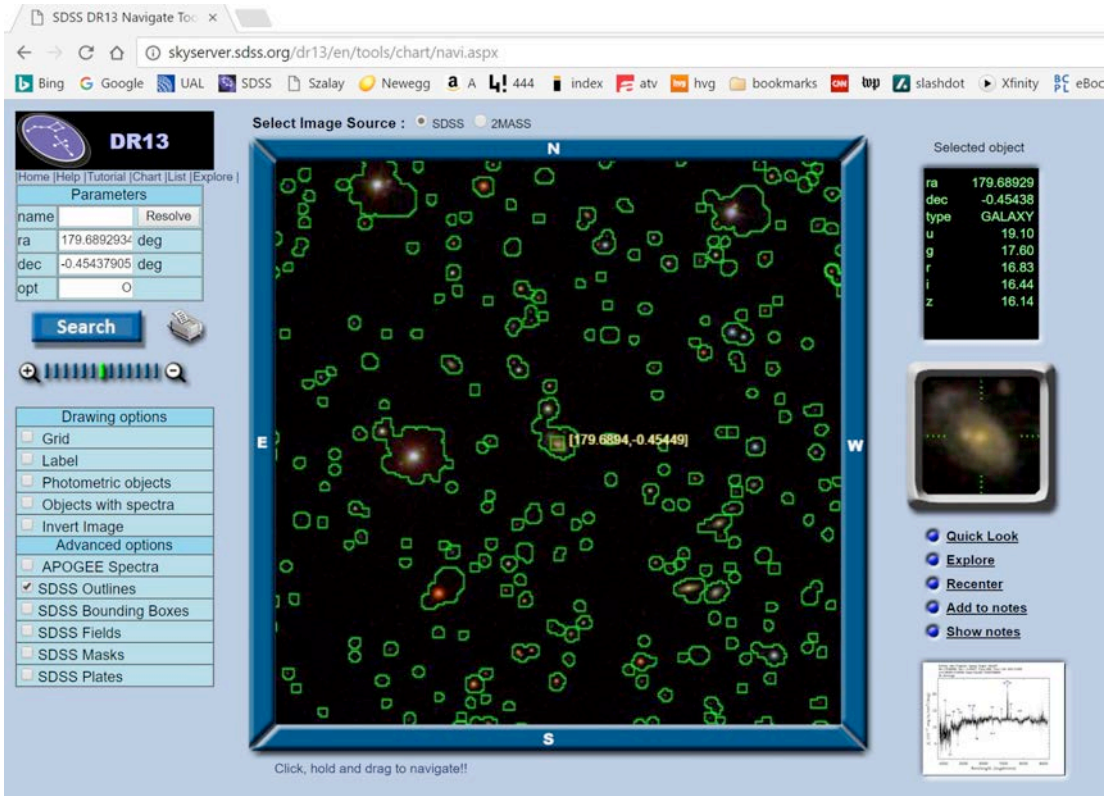
From Stars to Cells

- Strong parallels between medicine today and astronomy 25 years ago
- Stars and galaxies are like the cells in pathology
  - Multicolor photometry, image segmentation, locality
  - Spatial relations, spatial searches, outlines
- Deep links to the raw files
- Astronomy lessons:
  - Statistical analyses and collaboration easier in DB than flat files
  - Find a common processing level that is “good enough” and earn the TRUST of the community
  - Automation is needed for statistical reproducibility at scale
  - Scaling out was much harder than we ever thought
  - Moving many terabytes of data is hard

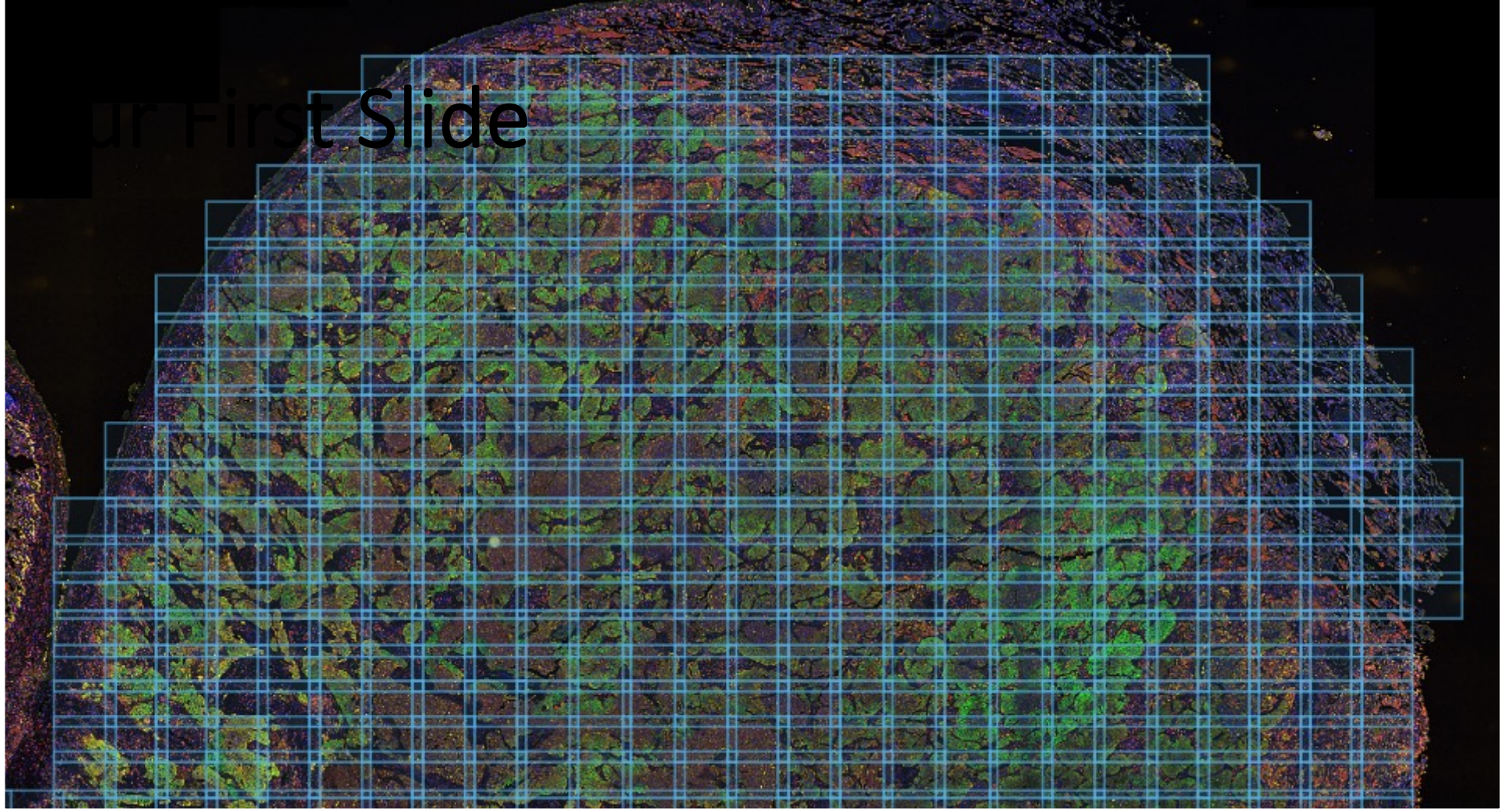
This will require:

- Bespoke approaches to cell segmentation, image analysis, and data management
- Focus on scaling workflow





ur First Slide

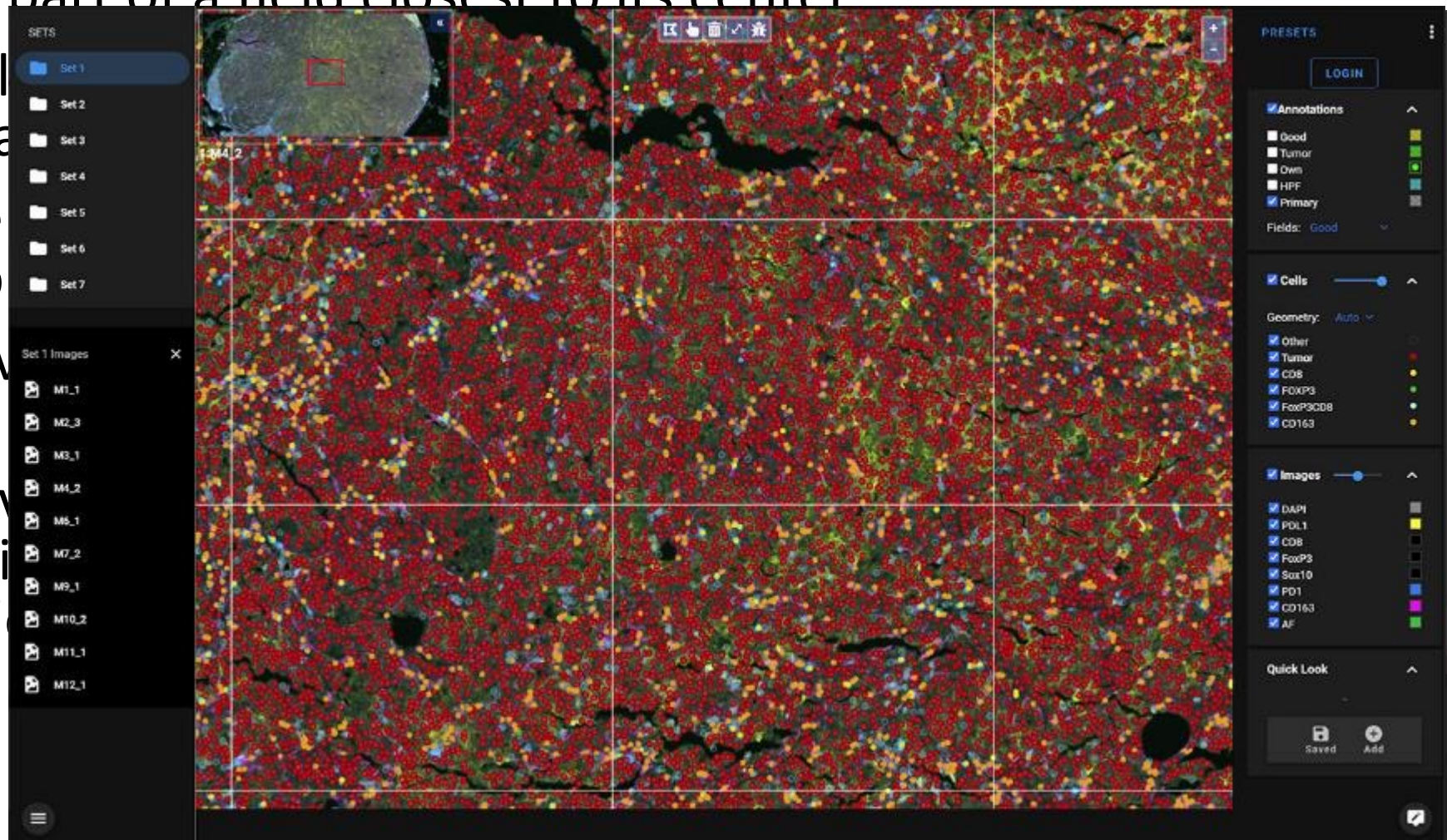


- 1226 High Powered Fields, 200GB / slide, 1344x1004 pixels, at 0.5µ/pix
- 1.6B pixels x 35 narrow band filters = 60B pixels /slide
- Overlaps provide repeated segmentations and measurements for intrinsic validation and quality control
- **800-1000 times more data collected for each slide**

AKOYA BIOSCIENCES®

Geometry: Overlaps and Primary Regions

- Primary area is the part of a field closest to its center
- These form a seaml
  tiling of the whole a
- Cells detected here
  the statistical samp
- Overlaps are observ
  multiple times
- Secondary cells serv
  QA tests to determi
  internal uncertainti

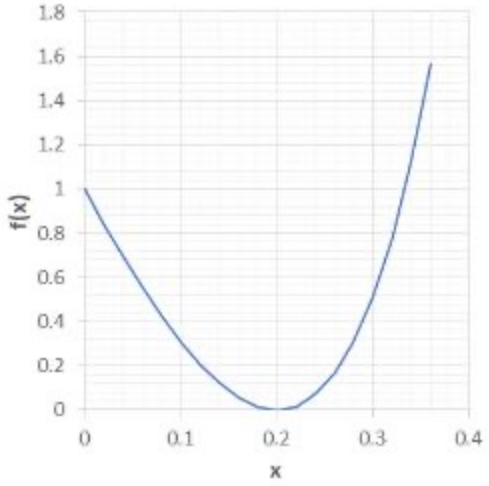


Rationale Behind Overlaps

- Overlaps provide independent photons from the same cells
- Repeated measurements – able to determine uncertainties in individual cell fluxes
  - More overlaps: better signal-to-noise
  - Too much overlap: waste of resources
- Measuring systematic errors
  - Microscope systematic errors largest in the corners: overlaps give information on how to correct them

Questions:

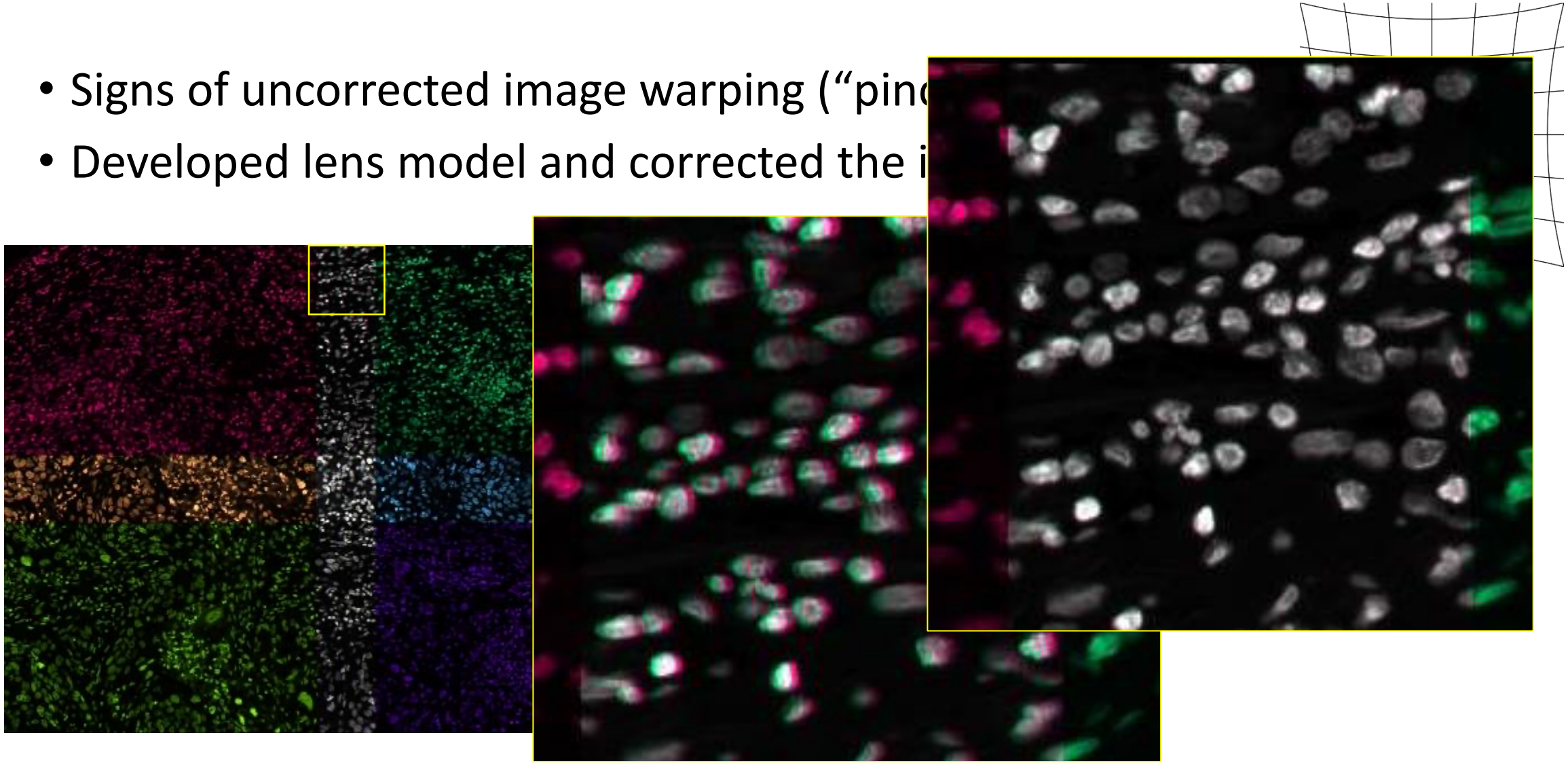
- How can we justify how much overlap should we take
- Overlaps must be big enough to characterize microscope model
- But too much overlap is costly
- Statistical error 1/sqrt(N), good balance between primary and secondary objects
- How can we use it in practice



20% is the optimum!

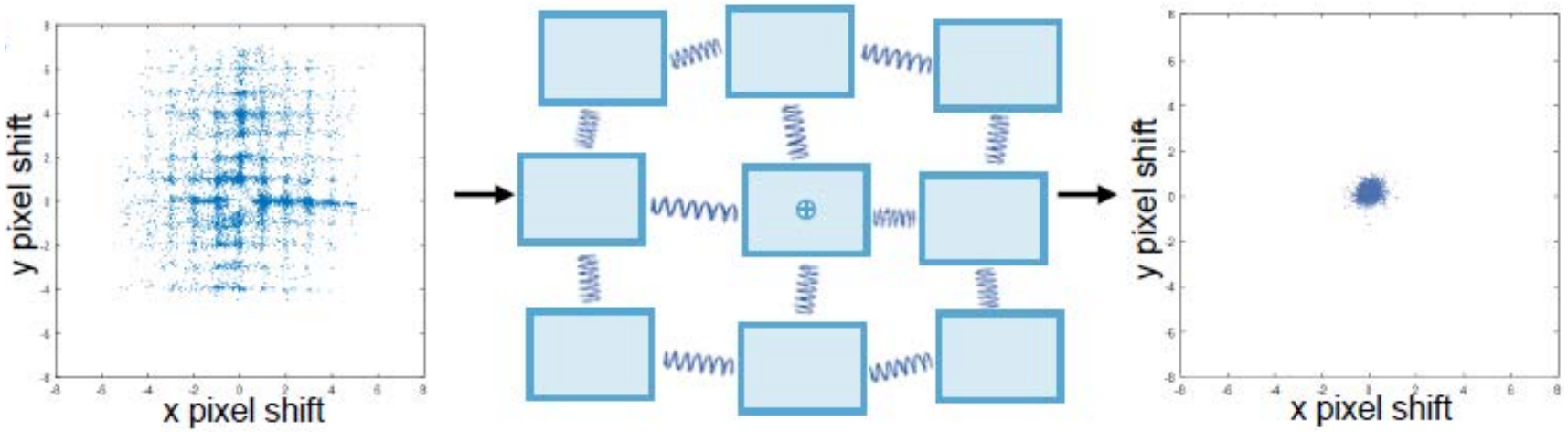
Using the Overlap Areas for Quality Control

- Signs of uncorrected image warping (“pinc
- Developed lens model and corrected the i



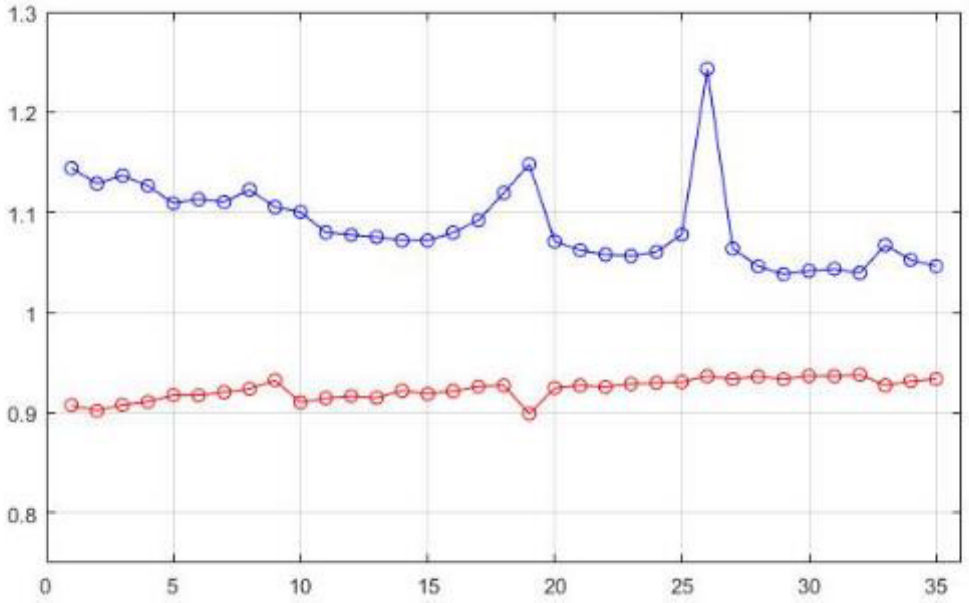
Automated Alignment of Image Mosaics

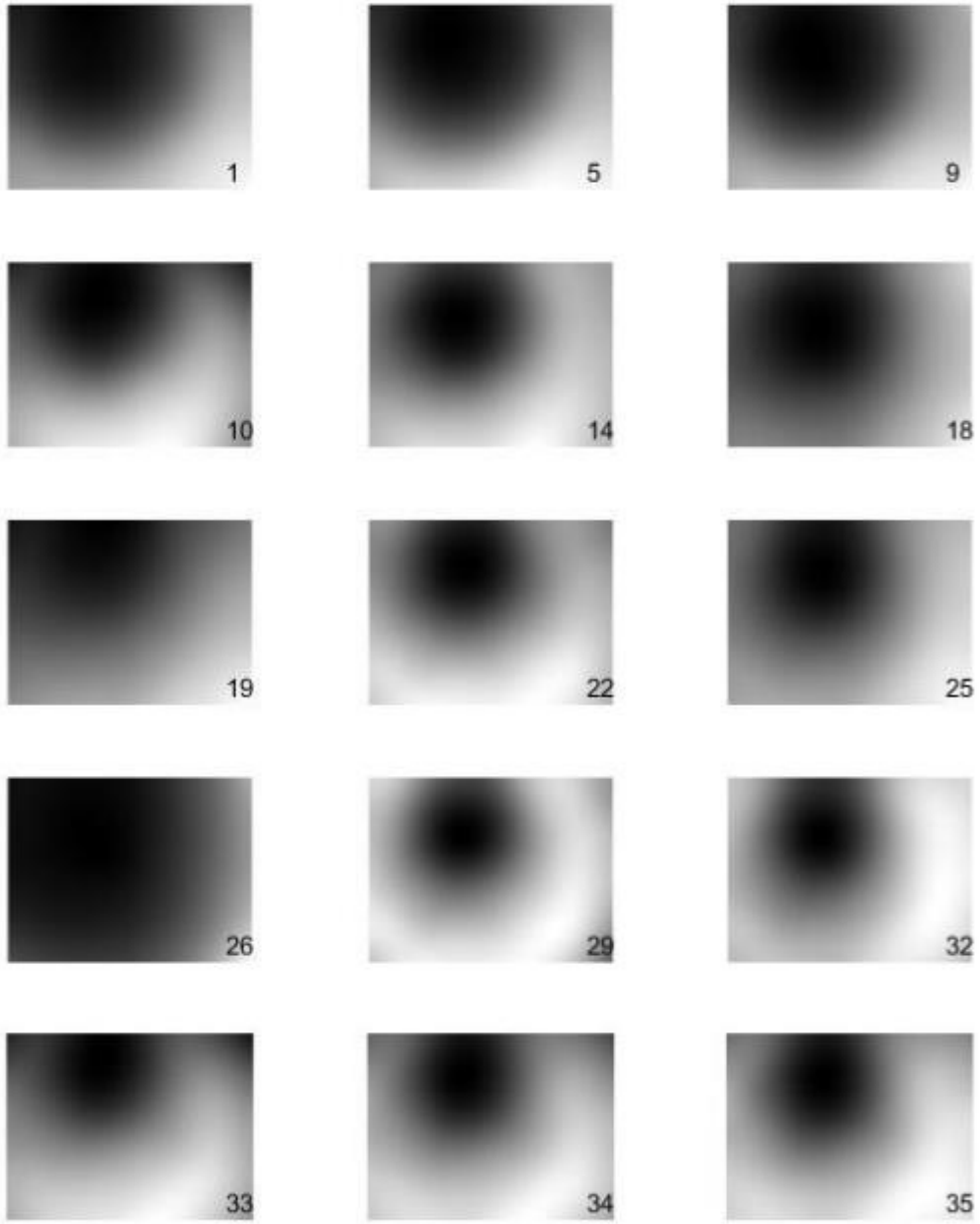
- The Vectra 3 microscope has a positioning “jitter” (3-6 pixels)
- Solved for optimal relative shifts of each pair of overlapping fields
- Consider each shift as a spring
- Pin down center, and let physics work -> equilibrium (minimum energy)



Improve Flat Fielding

- Originally: smoothed stack of 8,000 raw HPF images in the 35 filters
- Range of flat field correction aligns with the broad- and narrow filters





Preprocessing Workflow

manual | matlab | dbload

annotate

im3, Annotations

flatwarp

Unpack im3 bytes

Flat-field, Warp

Reinsert into im3

Extract DAPI

segment

Segment im3 InForm/dual

merge

Object list merger

Cell tables in CSV

geom

Tissue Geometry

Cell Geometry

Geom tables in CSV

Slide metadata

Scan geometry

Annotation geometry

prepDB

Read XML annotations

Compute overlaps

Overlap geometry

align

DAPI images

Pairwise shifts

HPF rel shifts

shift

Solve stitching

Final HPF geometry

zoom

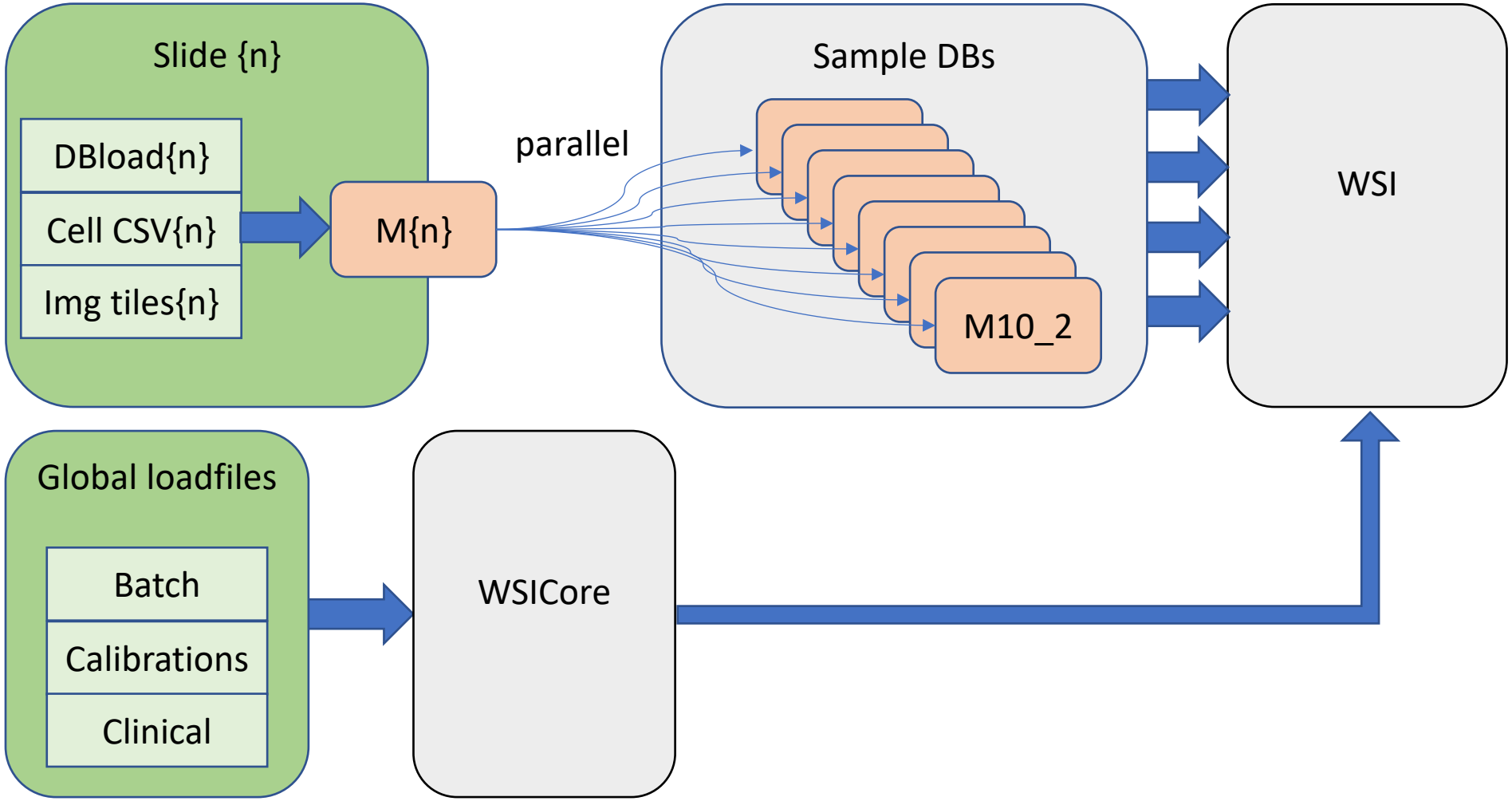
Build mosaic

Build tile pyramid

Tile pyramid

Tile metadata

DB Hiearachy with Two-Phase Load



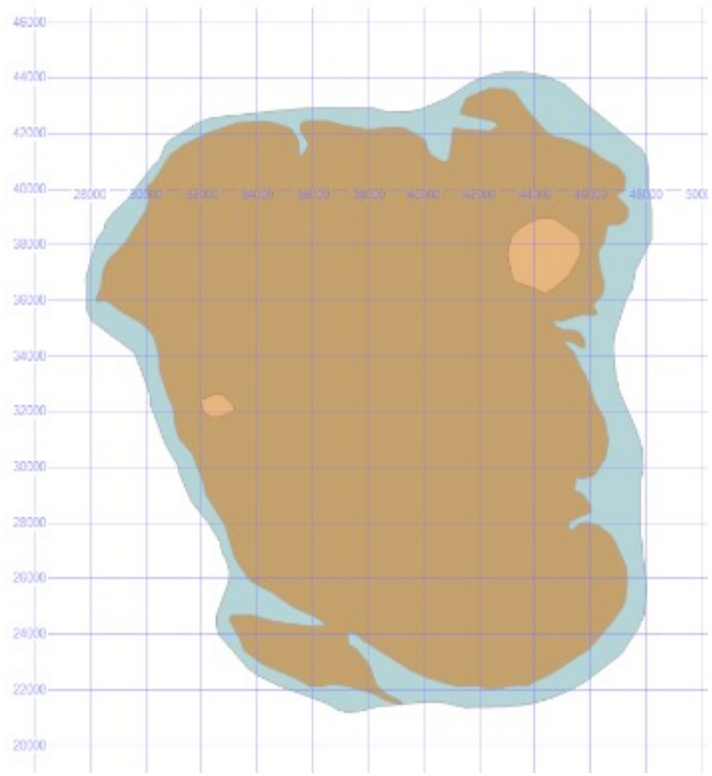
Imaging Improvements

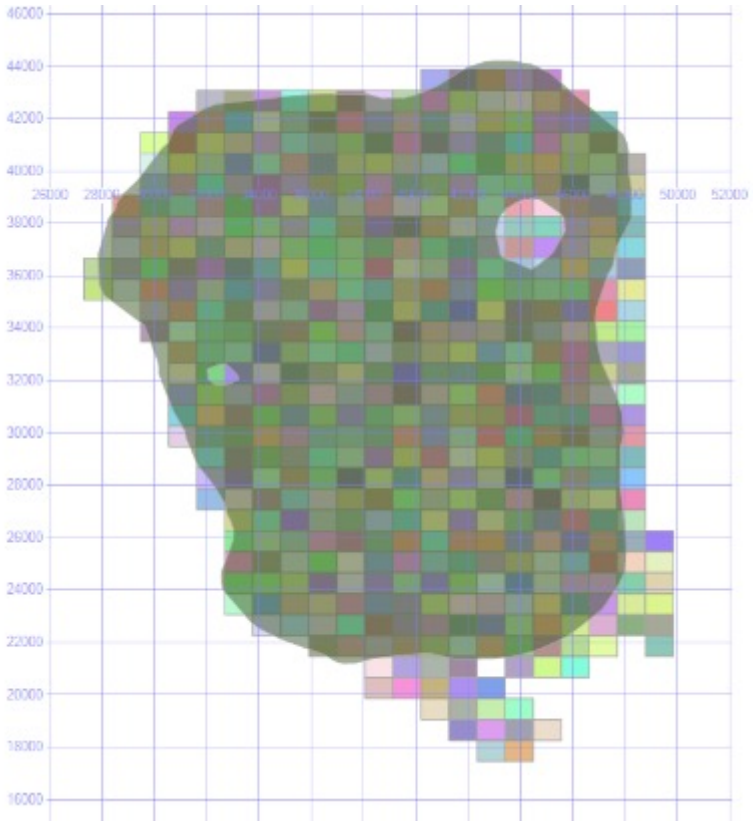
- Switching to Python packages from Matlab
- Better Flatfield model (M. Eminizer)
- Better Warping corrections (M. Eminizer)
- Capture all the Akoya XML metadata (R. Wilton)
- Substantially improve image stitching accuracy (H. Roskes)
- Cross-registration of different imaging modalities (J. Doyle)
- Built whole slide visualization and used it for QA (D. Medvedev)
- Data processing + loading is almost fully automated with multi-level logging, arbitrarily parallel design, increasing use of GPUs (B. Green)

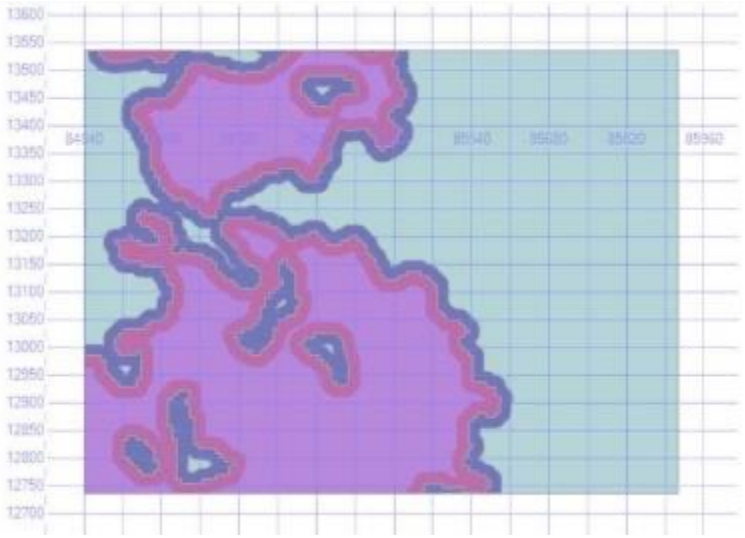
Spatial Features in the DB

- Represented as spatial polygons, using GIS grammar
- Geometries represented
  - HPF outlines and primary regions
  - Manual annotation of good tissue
  - Manual annotation of tumor boundaries
  - Automatic tumor boundaries
  - Membrane outline for each cell
  - Nucleus outline for each cell
- Distances and areas
  - Each cell has its signed distance from tumor boundary computed
  - Areas of different buffers around tumor boundary precomputed
  - Fractional area of each HPF inside good tissue and tumor computed

Annotations and Buffer Regions







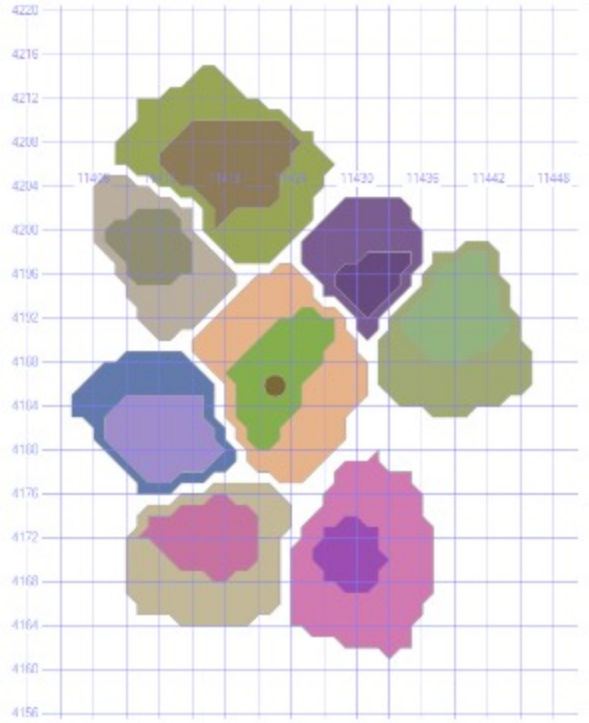
Precomputed Neighbors

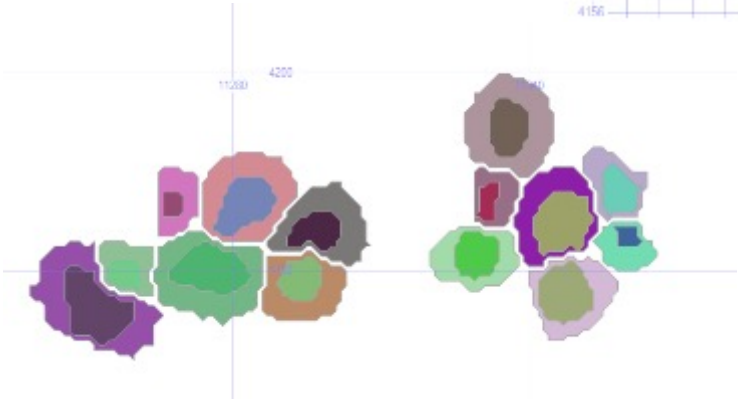
SQLQuery14.sql - B...(WIN\aszalay1 (61))* | SQLQuery13.sql - B...(WIN\aszalay1 (57))* | SQLQuery12.sql - B...(W

```
CREATE TABLE Neighbors (
-----------------------------------------------------------------------
--/T the precomputed neighbors of each cell within 100 pixels
-----------------------------------------------------------------------
    sampleid int NOT NULL,      --/D sampleid
    c1 bigint NOT NULL,         --/D cellid of center
    c2 bigint NOT NULL,         --/D cellid of neighbor
    ptype1 tinyint NOT NULL,    --/D enumerated phenotype of c1 --/E Phenotype
    ptype2 tinyint NOT NULL,    --/D enumerated phenotype of c2 --/E Phenotype
    pexp1 tinyint NOT NULL,     --/D enumerated expression of c1
    pexp2 tinyint NOT NULL,     --/D enumerated expression of c2
    dist float NOT NULL,        --/D centroid distance to neighbor --/U pixels
    qt1 int NOT NULL,           --/D PD-1 markup quantile for c1
    qt2 int NOT NULL,           --/D PD-1 markup quantile for c2
    pt1 int NOT NULL,           --/D PDL-1 markup quantile for c1
    pt2 int NOT NULL,           --/D PDL-1 markup quantile for c2
    tdist1 real NOT NULL,       --/D tumor distance of c1 --/U pixels
    tdist2 real NOT NULL,       --/D tumor distance of c2 --/U pixels
    r bigint NULL               --/D rank of neighbor by increasing distance
)
```

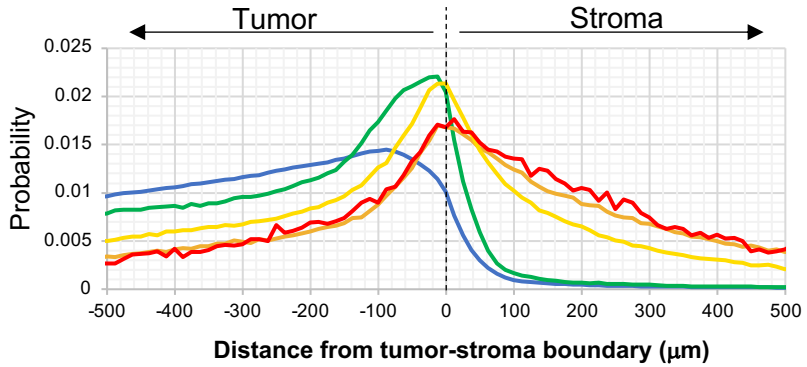
ContactNeighbors

```
insert ContactNeighbors with (tablock)
select n.*
from Neighbors n, CellGeom a, CellGeom b
where n.c1 = a.cellid
  and n.c2 = b.cellid
  and n.dist<=50
  and n.sampleid=@sampleid
  and a.btype=0
  and b.btype=0
  and a.geom.STDistance(b.geom)<2
```





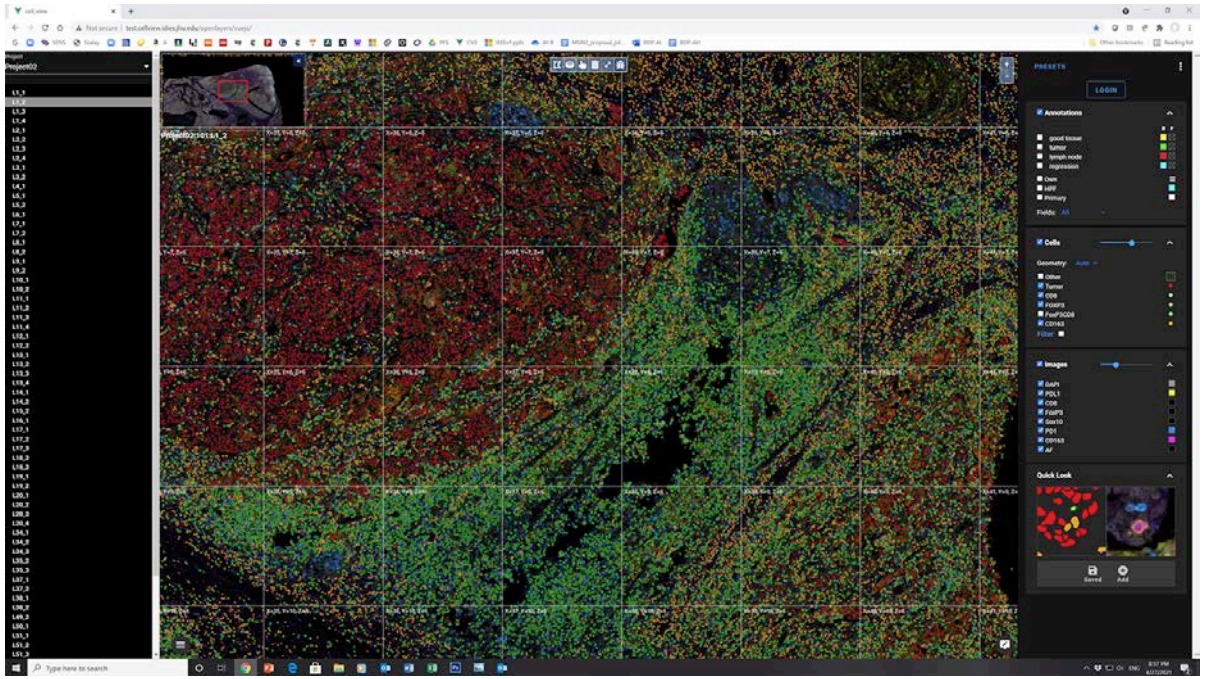
Data Analysis Developments



- Random samples extremely useful for estimating cell density in complex geometries
  - E.g. how far inside the regression area, and how far from tumor
  - Now added precomputed distance from regression boundary
- Perform two queries and a division
  i. Build histogram of distances of the real cell count
  ii. Build identical histogram of random cells with known density
  iii. Their ratio is the density of the cells of interest in each distance bin
- Works with arbitrary geometries
- Working to introduce more advanced spatial statistics and ML tools
  - Correlations, mark-correlations, neighbor statistics, tSNE, UMAP
- Starting to look at genomics integration (w. Alex Baras)

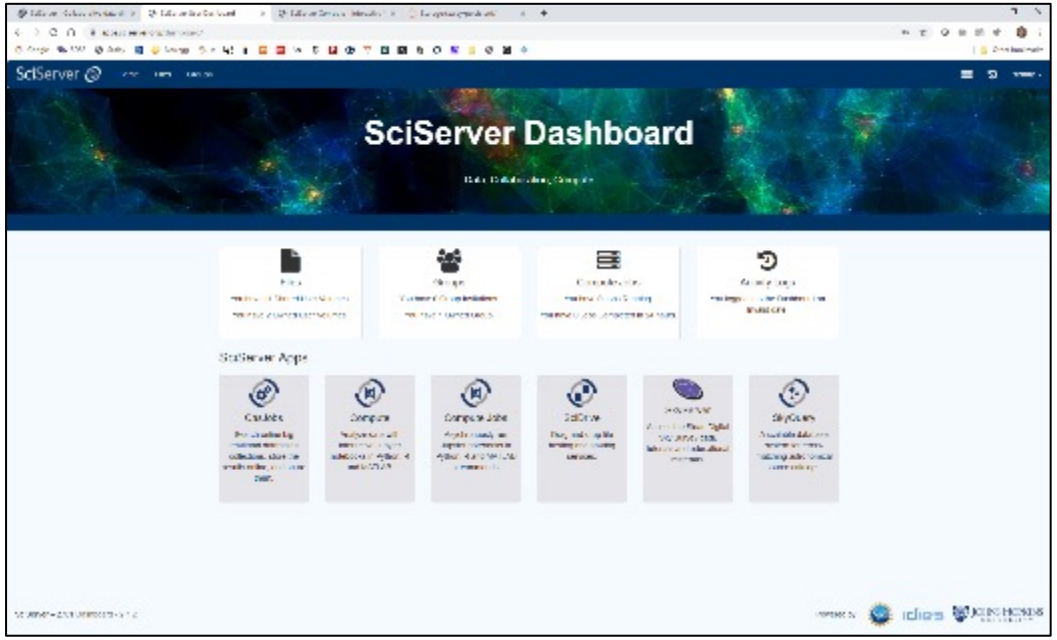
CellView: Designed for Speed

- Images built from hierarchical DeepZoom tiles (256x256)
- What is different from others that we use a lightweight client, all heavy lifting is done server-side
- We store the tiles of the original 8 component layers
- They are mixed on the server, using a user-defined color mixing
- These can be saved as presets



SciServer Integration

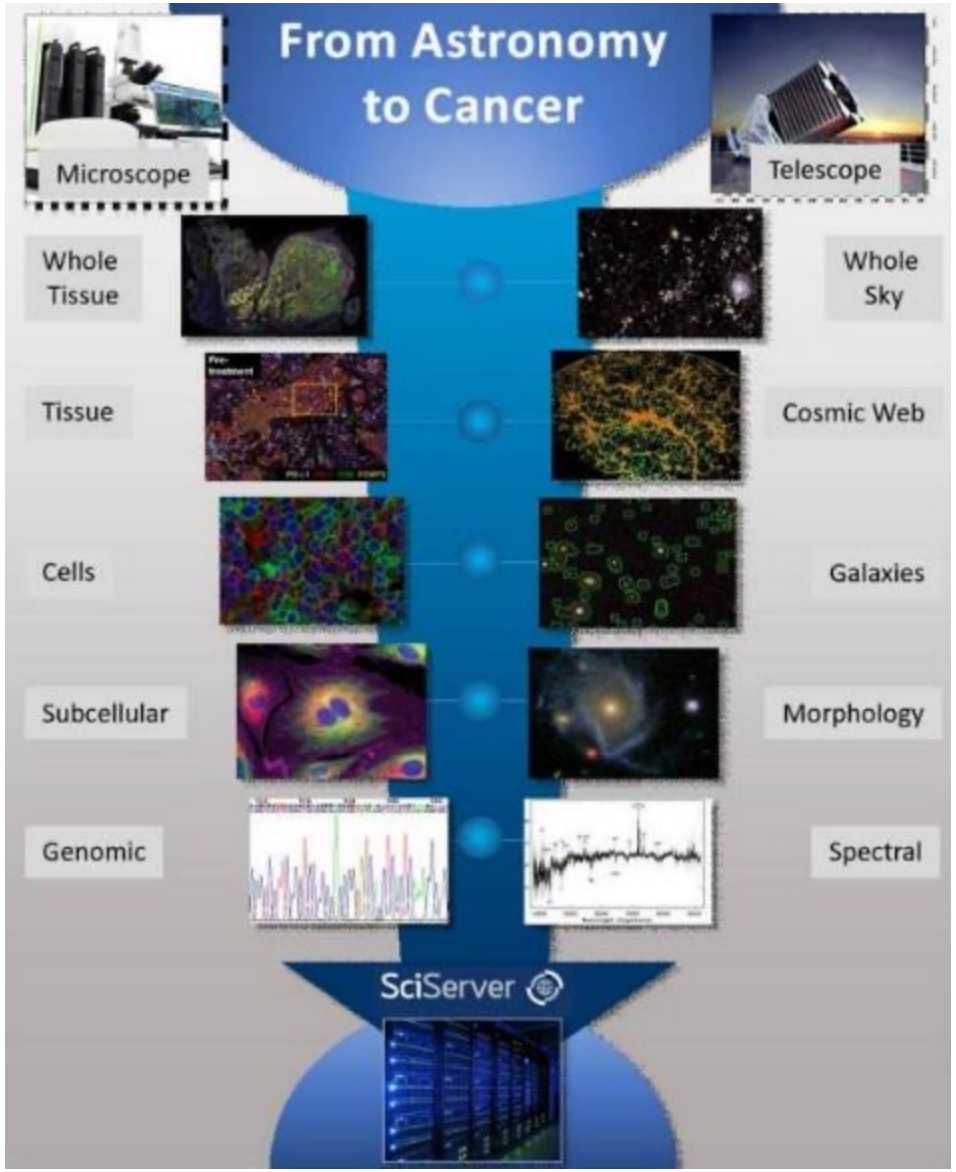
- The database is now linked to the SciServer (JHU data analytics platform)
- Collaborative sharing
- Enables easy data aggregation
  - With genomics etc
- Each user can have their own DB for value added data, linked to main database
- Various options:
  - CasJobs/MyDB (SQL access)
  - Compute (Python, R)
  - Compute Jobs (queues)
  - Preconfigured containers with AI
    - PyTorch
    - Tensorflow
    - Choice of Python2, Python3
    - Geo (spatial tools)



sciserver.org

Current data in the database

- 3 Cohorts, 235 slides
- 84,320 High Powered Fields
- 226M detected cells
- 97M unique cells
- 3.5B neighboring cell pairs precomputed
- 8.7 trillion pixels (whole SDSS was 6.5 Tpixels!)
- Additional 200+ slides already scanned
  with multiple tumor types, processing in
  various stages



Conclusions

- Early results indicate that mIF assays are reproducible
- Found a predictive biomarker for immuno-therapy using AstroPath (Science, June 11, 2021)
- Next generation of tissue-based biomarkers are likely to be identified using large, well-curated datasets
- Established a standardized protocol to process thousands of tissue samples per year on many microscopes
- Developed a scalable facility to produce petabytes of robust tissue imaging data on par with large sky surveys
- Working towards an Open Cancer Cell Atlas with many billions of cells

Acknowledgements

Taube Lab members

Sneha Berry, MS

Nicholas Giraldo, MD, PhD

Benjamin Green

Tricia Cottrell, MD, PhD

Liz Engle, MS

Haiying Xu

Aleksandra Ogurtsova

BKI collaborators

Drew Pardoll, MD, PhD

Robert Anders MD, PhD

Suzanne Topalian, MD

Evan Lipson, MD

Astronomy/IDIES

Heshy Roskes, PhD

Maggie Eminizer, PhD

Richard Wilton, MD

Joshua Doyle, MD

Sahil Hamal, CS

Dmitry Medvedev, CS

Josh Rabichaud (UG, Physics)

Nate Eisenberg (UG, math)

AI/Computer Vision

Alan Yuille, PhD

Seyoun Park, PhD

Yixiao Zhang

Akoya collaborators

Cliff Hoyt, MS

Chi Wang

BMS collaborators

Robin Edwards, MD


The Mark Foundation®
for Cancer Research


Melanoma
Research Alliance


HARRY J. LLOYD
CHARITABLE TRUST


SIDNEY KIMMEL COMPREHENSIVE CANCER CENTER
BLOOMBERG~KIMMEL INSTITUTE
FOR CANCER IMMUNOTHERAPY
JOHNS HOPKINS
MEDICINE


Emerson
Collective


Institute for Data Intensive Engineering and Science
idies


NIH
NATIONAL
CANCER
INSTITUTE


Bristol-Myers Squibb