

Spatial Analysis on Histological Images Using Spark

Wei-Yi Cheng and Franziska Mech

Roche Pharma Research and Early Development (pRED)
Informatics, Data Science

Roche Innovation Center New York / Munich






SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE

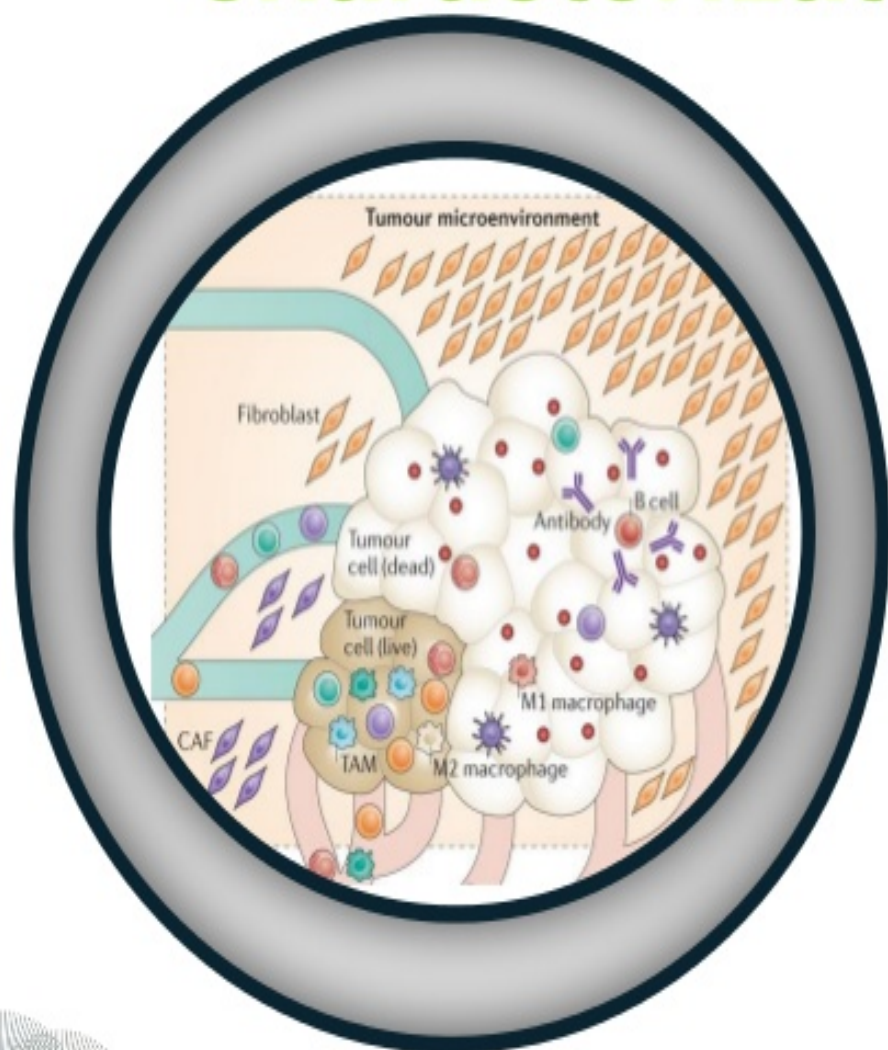
Disclaimer

- This presentation is ...
 - **✗ NOT** about computer vision / image processing
 - **✗ NOT** about drug, biology, or biochemistry
 - **✗ NOT** about new algorithm or infrastructure

Disclaimer

- This presentation is ...
 -  An application of Spark on spatial analysis on biomedical images
 -  A proof of concept of a small module in a complex pipeline
 -  A work in progress

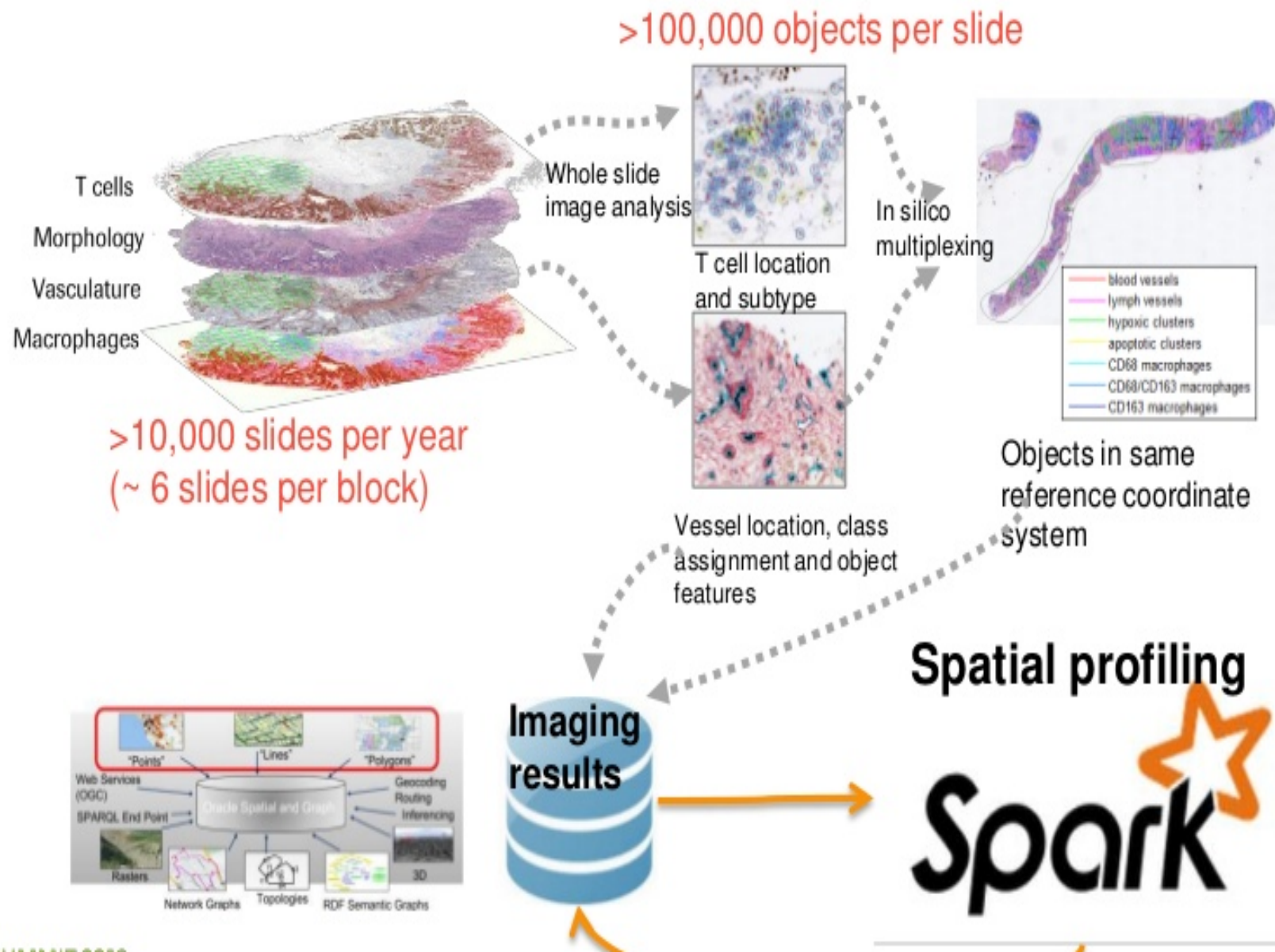
Towards a systematic characterization of tumor context

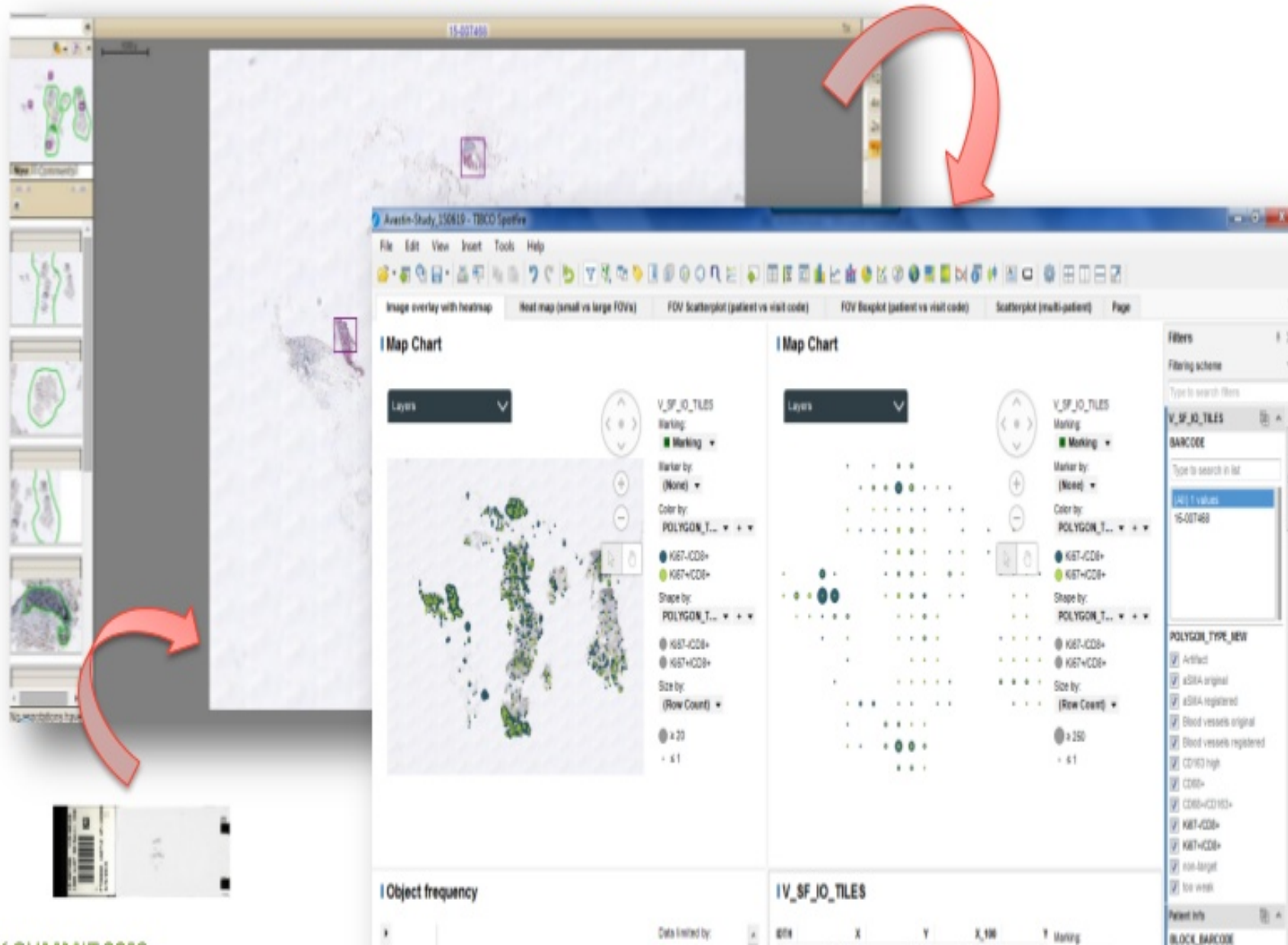


Challenges:

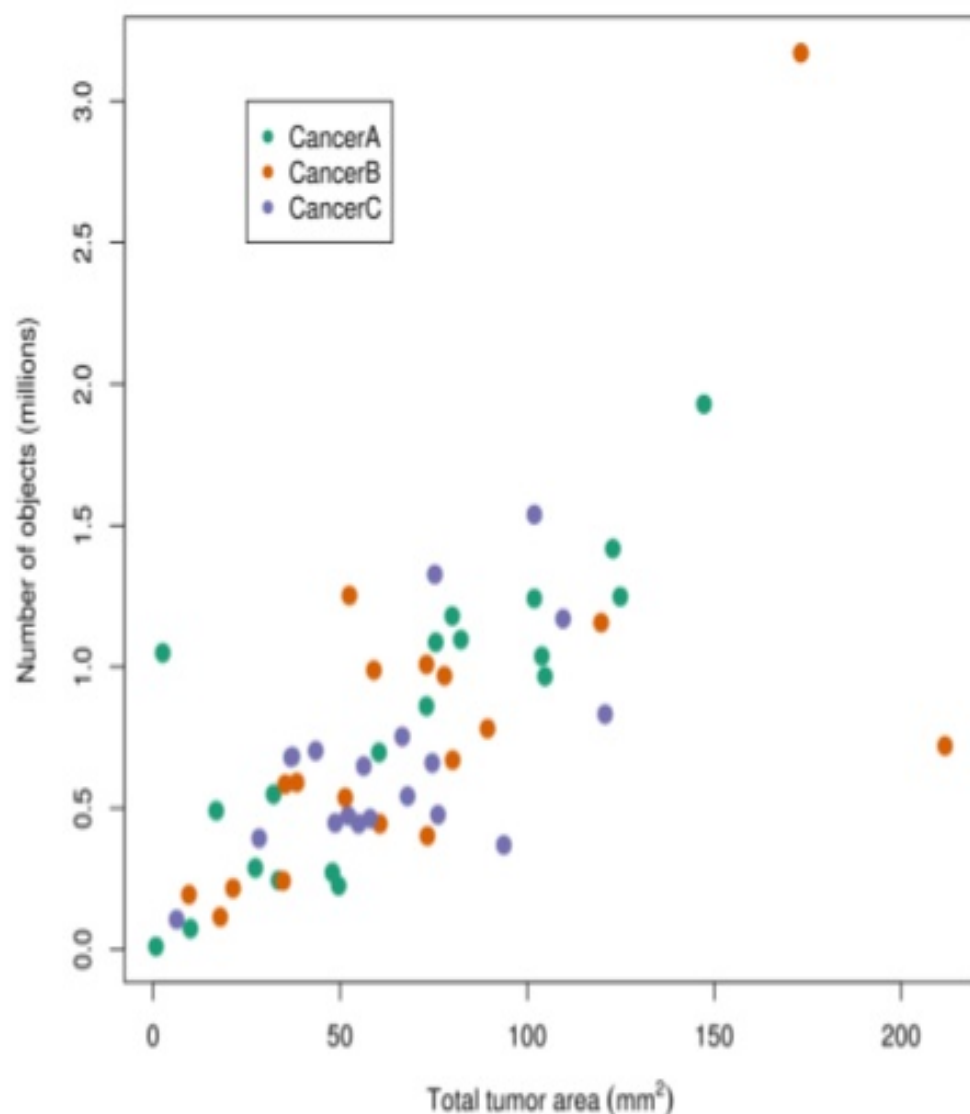
- Location and density of different immune cell populations are associated with patient prognosis and prediction outcome
- Huge variation in immune infiltrates across tumor entities and patients
- Inconsistent data in the literature

Roche pRED Tissue image analysis workflow





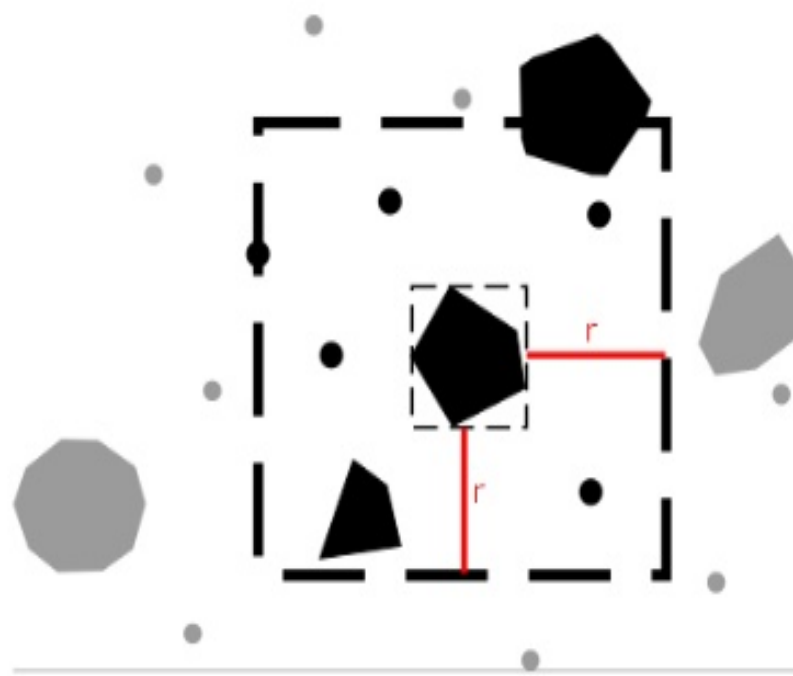
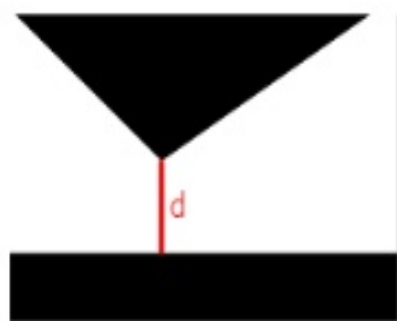
POC data set



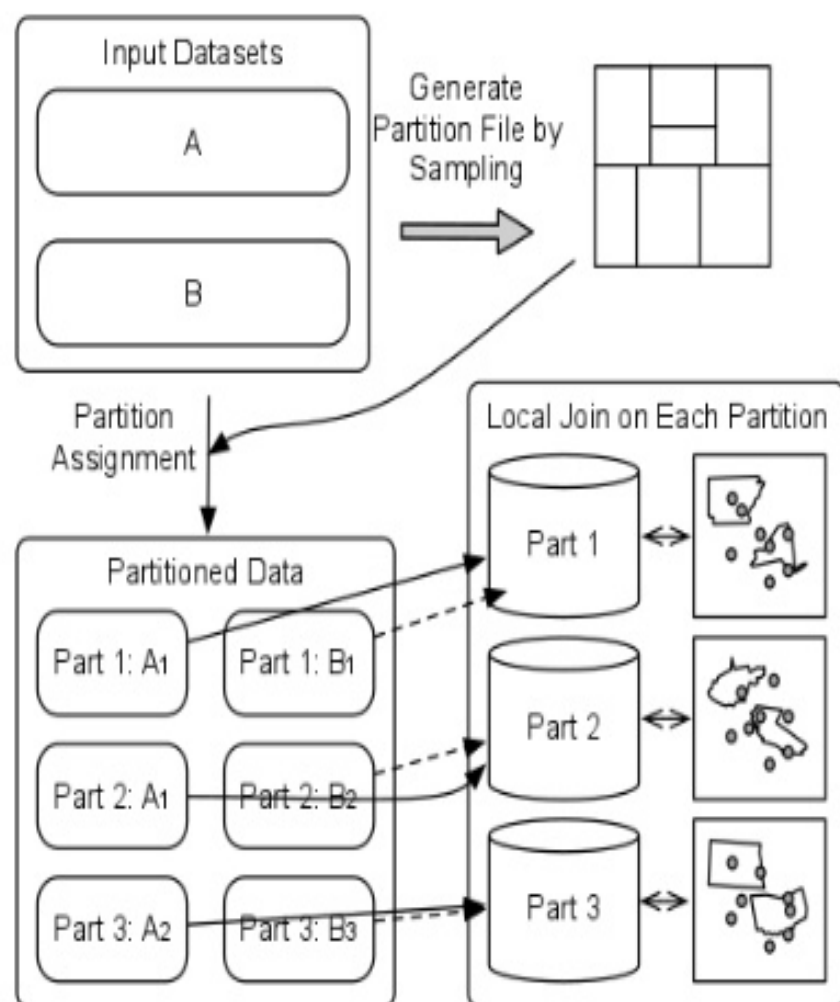
- 57 “blocks” of 3 cancer types
- Each block contains 6 or 7 slides
 - histology stain
 - cancer biomarker
 - microenvironment biomarker
- Each object annotated with object type (T cell, lymph vessels, etc.), shape (point / polygon), and coordinate

Distance calculation

- Basic statistic for distribution, co-localization, spatial clustering, etc.
- Distance: shortest distance between two contours
- Total number of pairwise distances: 5.3 trillion (10^{12}) pairs – prohibitive
- Workaround: only calculate the distance between each object and its “neighbors” within a window r .



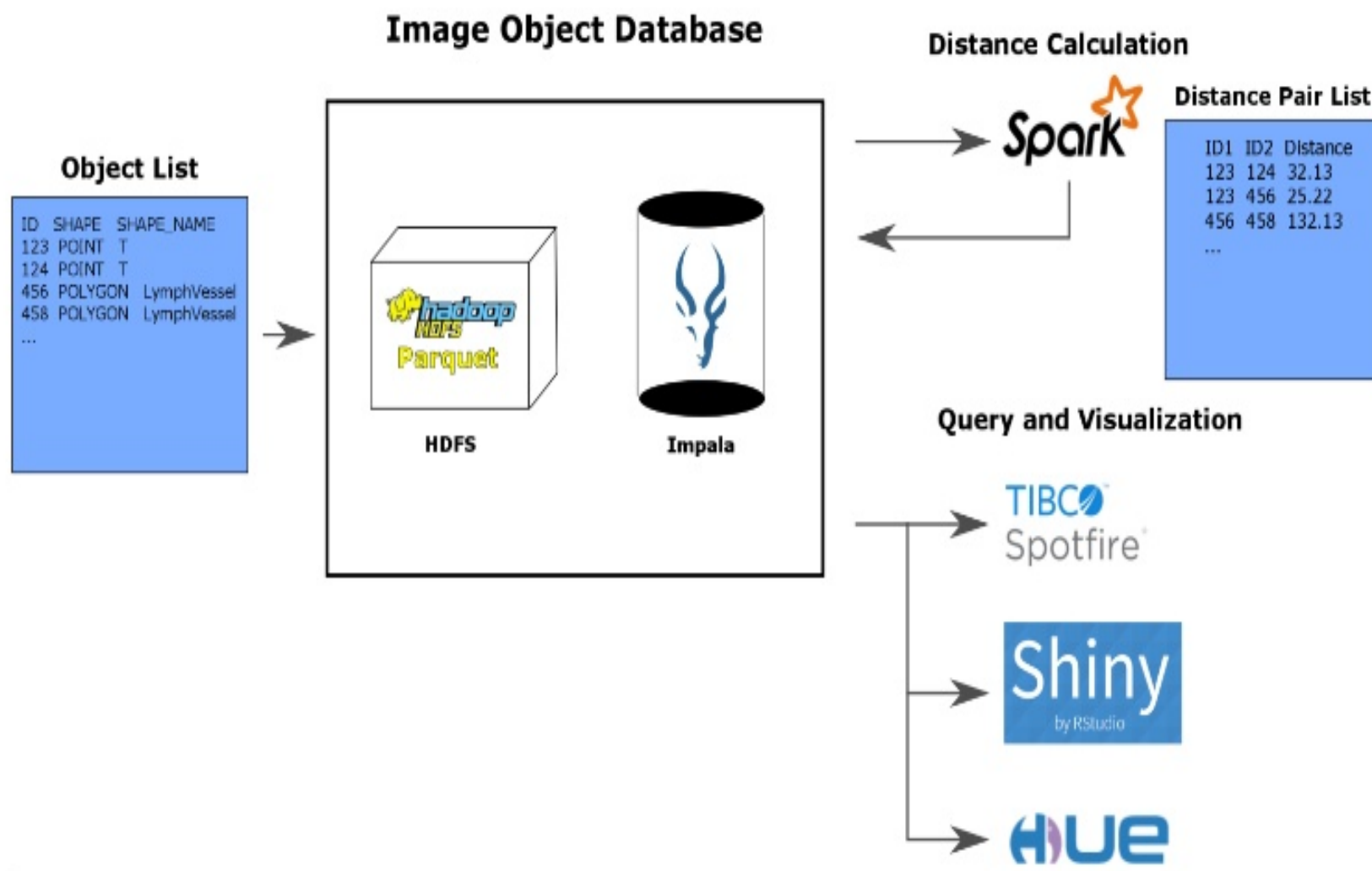
Spatial Spark



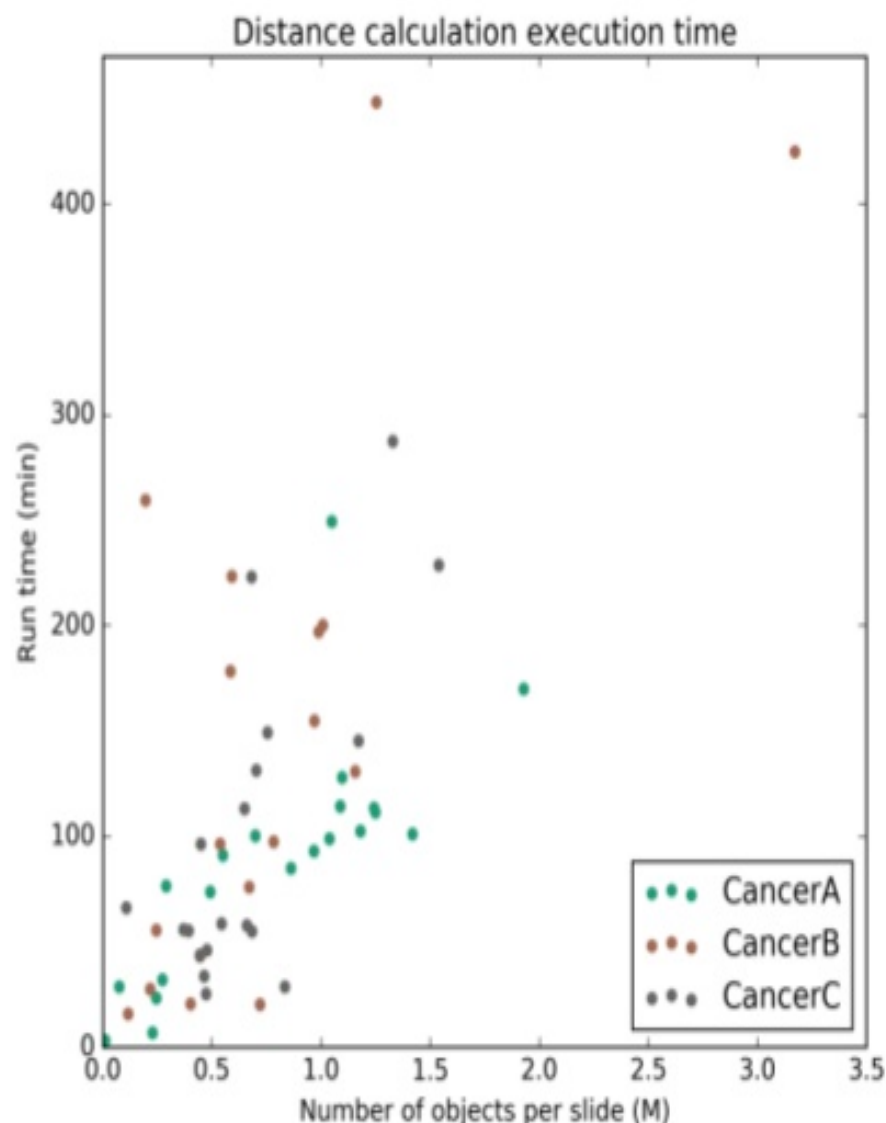
<http://simin.me/projects/spatialspark/>

- An open source library developed by Dr. Simin You and Prof. Jianting Zhang from CUNY.
- Divide-and-conquer, following similar designs of HadoopGIS.
- Support multiple spatial partitioning methods: sort-tile partition, binary-split partition, fixed-grid partition

Spatial profiling workflow

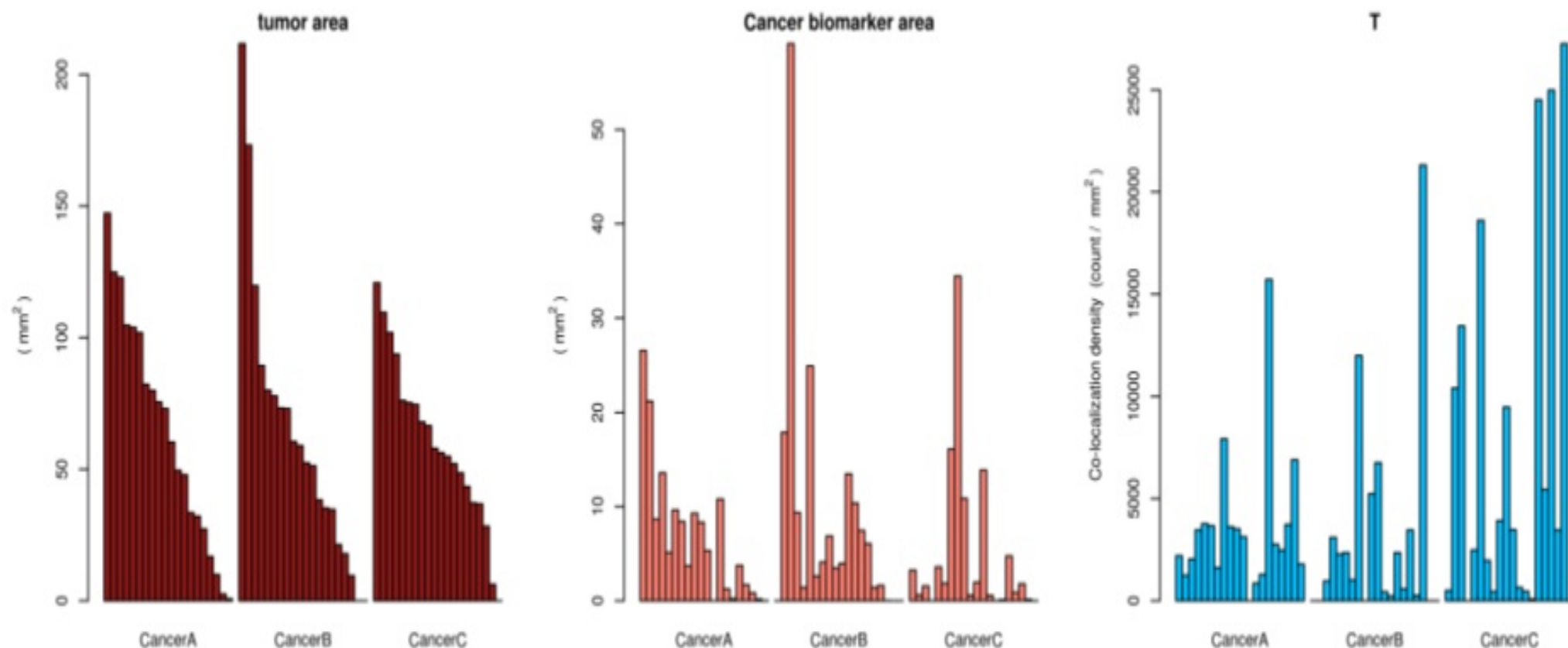


Distance calculation: run time



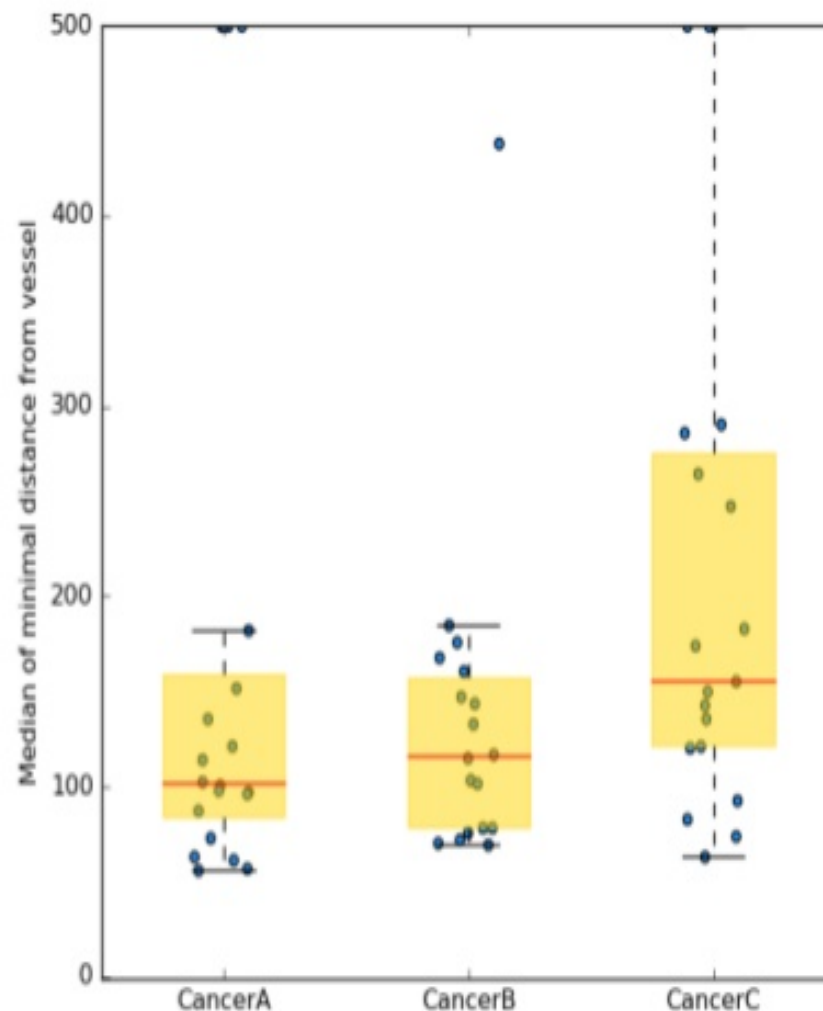
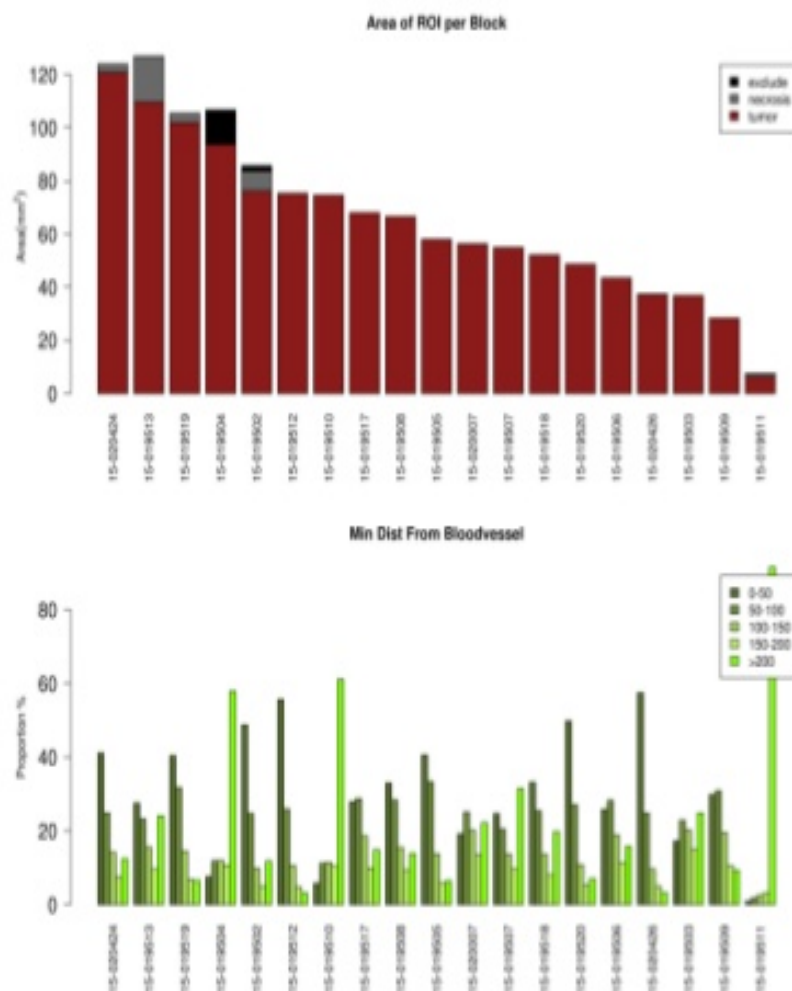
- Using 18 cores, 4 threads / core (72X parallelization).
- Radius = 232.5 μm
- Under shared test environment
- Execute time is roughly linear to number of object (theoretical time complexity).

Co-localization



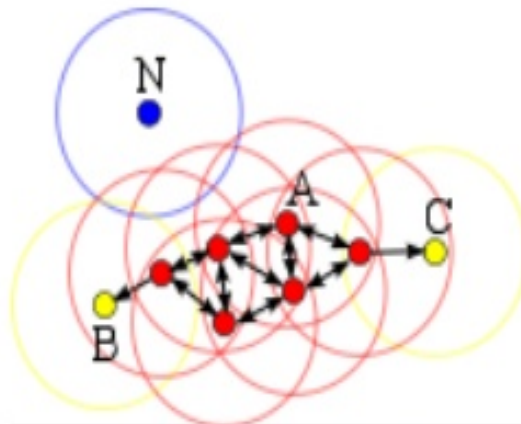
- Enables profiling of between-indication variation and between-tumor variation.

Distance distribution from immune cell to blood vessel

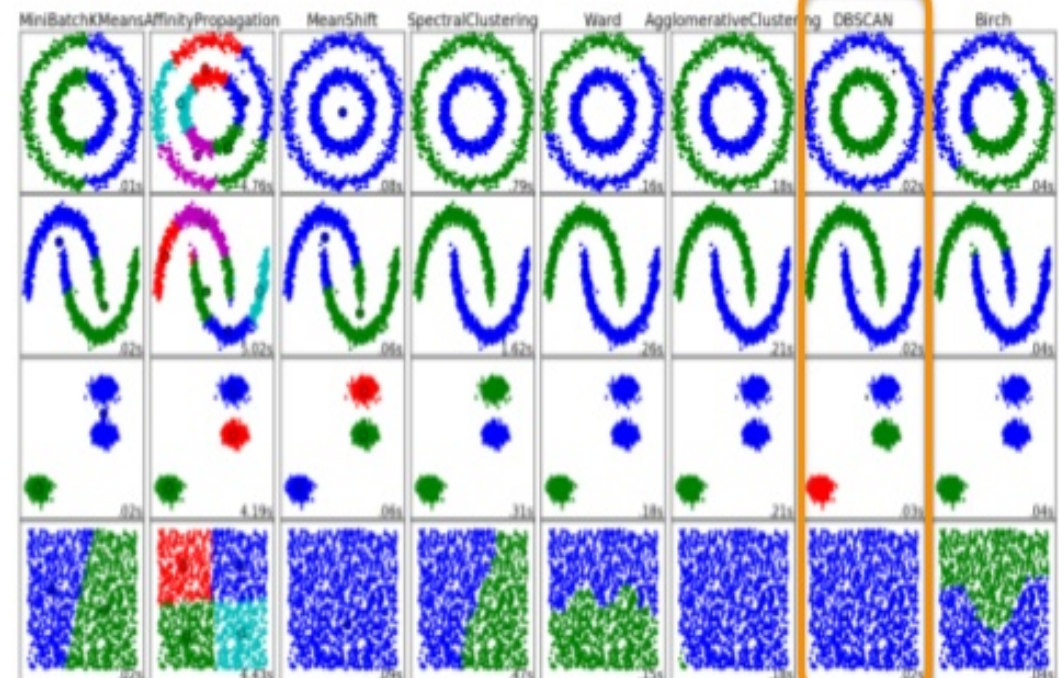


Spatial clustering of objects: DBSCAN

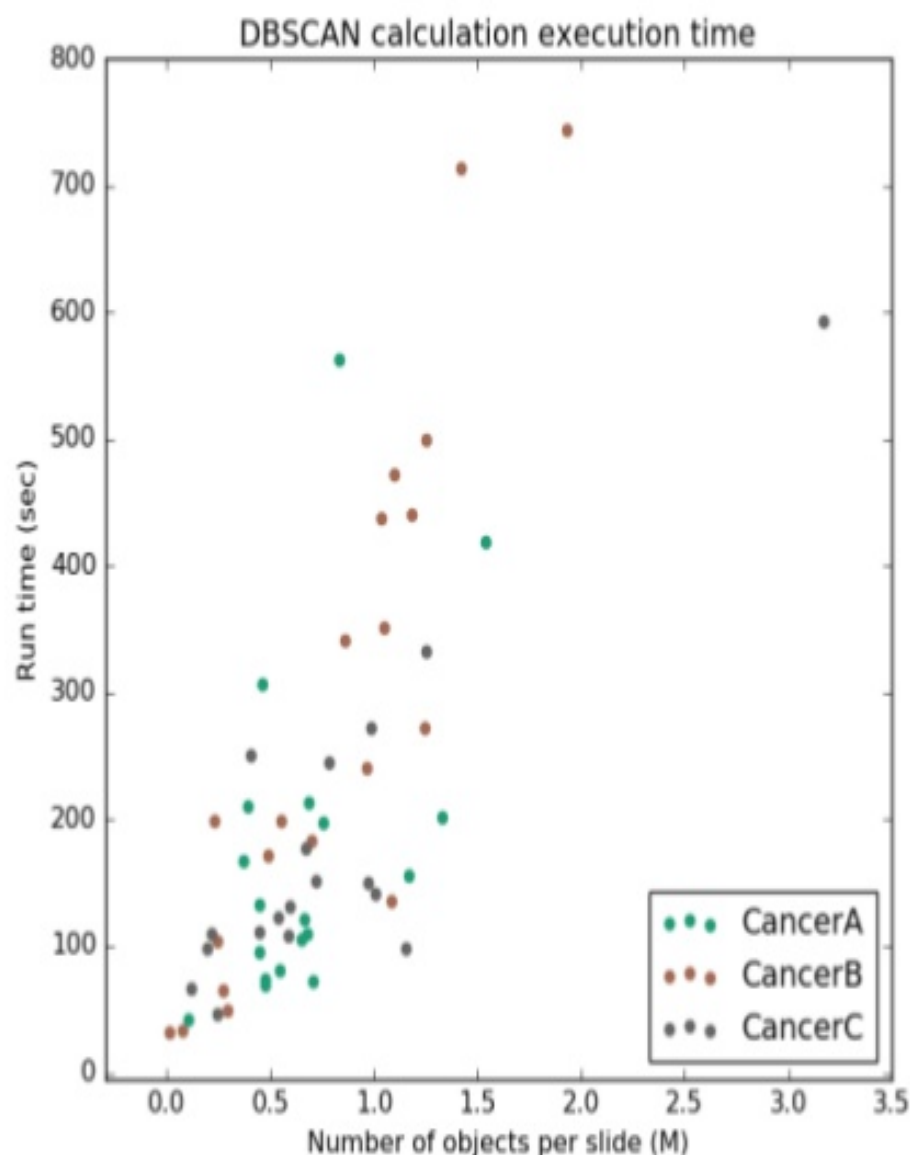
- **Density-based spatial clustering of applications with noise (DBSCAN)**
clusters closely connected points into the same cluster, marking high-density regions.
- Parameters:
 - minPts: minimum number of neighbors
 - Eps: maximum distance to define a neighbor



<https://en.wikipedia.org/wiki/DBSCAN>



DBSCAN: run time



- Directly load distance from parquet.
- Run time is linear to the number of objects (given already calculated distances).

Future works

- Clinical information
- Genomic data
- Scale up and upstream integration
- UI integration

Acknowledgement



- Spark exploratory / support
 - Sittichoke Saisanit (*Roche pREDi*)
 - Xing Yang (*Roche pREDi*)
 - Padmanabha Udupa (*Roche pREDi*)
 - Ivan San Antonio Martinez (*Roche IDW*)
 - Zayed Albertyn (*Novocraft*)
- Tumor image spatial analysis research
 - Angelika Fuchs (*Roche pREDi*)
 - Gerlind Herberich (*Roche pREDi*)
 - Jurriaan Brouwer (*Roche pREDi*)
- Spatial Spark
 - Jianting Zhang (*CUNY*)
 - Simin You (*CUNY*)



Doing now what patients need next

THANK YOU.

Wei-Yi Cheng

wei-yi.cheng@roche.com



SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE