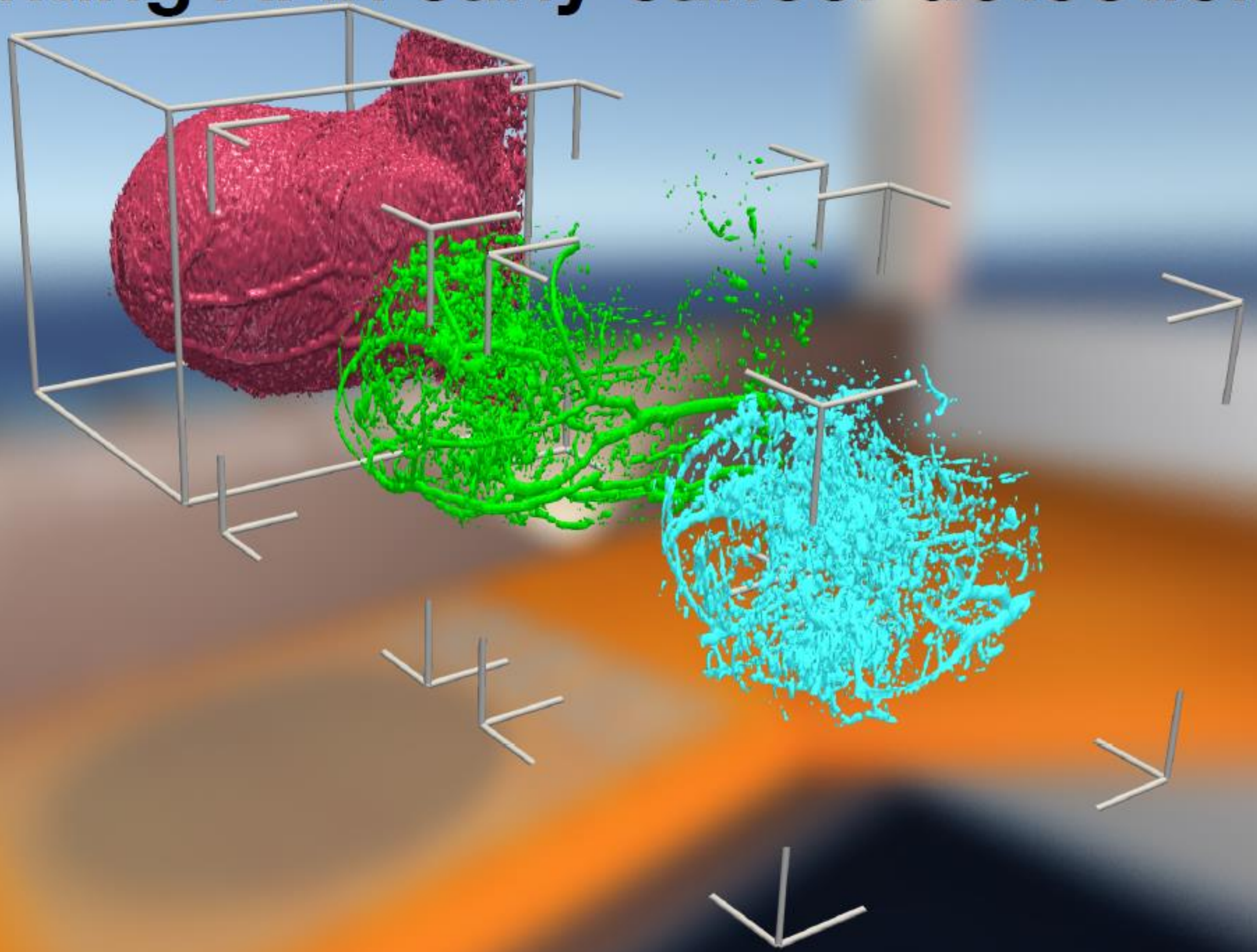
Implementing AI in early cancer detection



Ashish Mahabal,
Center for Data Driven Discovery, Caltech
EDRN, 30 Jun 2020

Problem setup

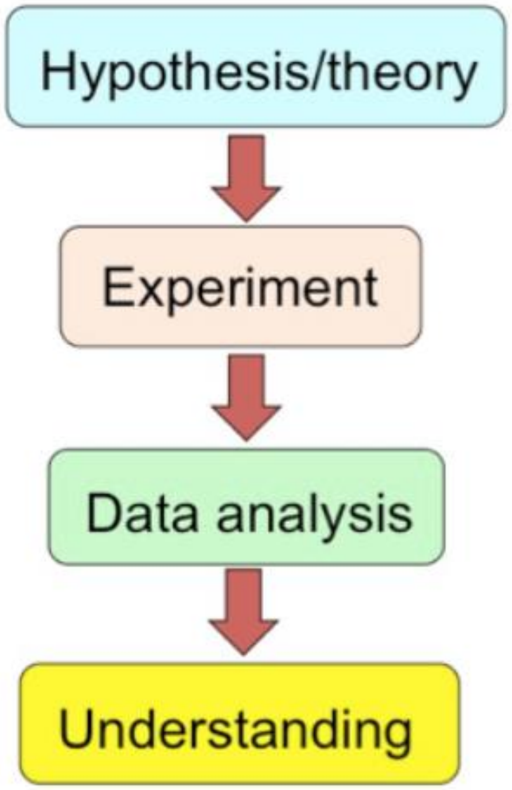
- Number of patients with cancer has been on the rise
- More diverse measurements are possible
    - Imaging (MRI, CT, Mammogram, …)
    - Genomic (mRNA, miRNA, …)
- Finer measurements are available
    - Better resolution

Traditional Machine Learning Flow

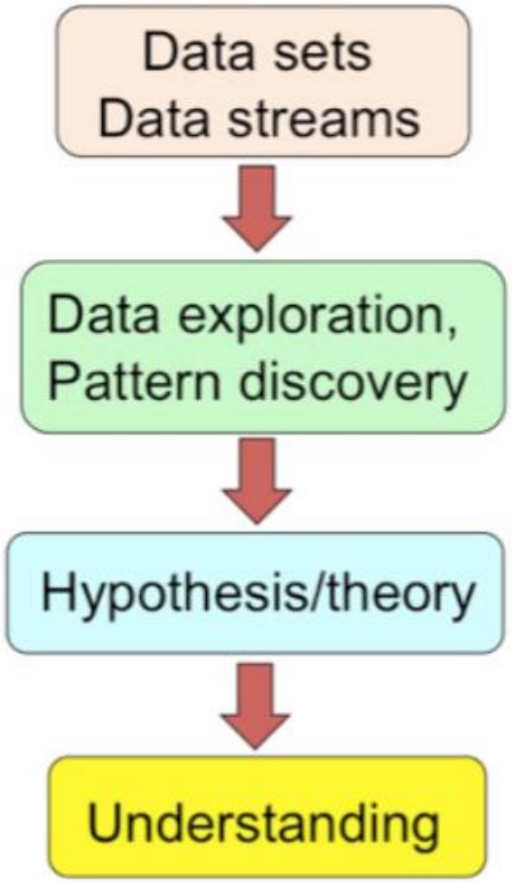
Over diagnosis?

Deep Learning Flow

Hypothesis-driven science



Data-driven science



multi-omics

The two approaches are complementary

Djorgovski

A Modern Scientific Discovery Process

Data Gathering (finstruments, sensor networks, their pipelines…)

↳ **Data Farming:**

Storage/Archiving
Indexing, Searchability
Data Fusion, Interoperability

} Databases
Data grids

↳ **Data Mining**

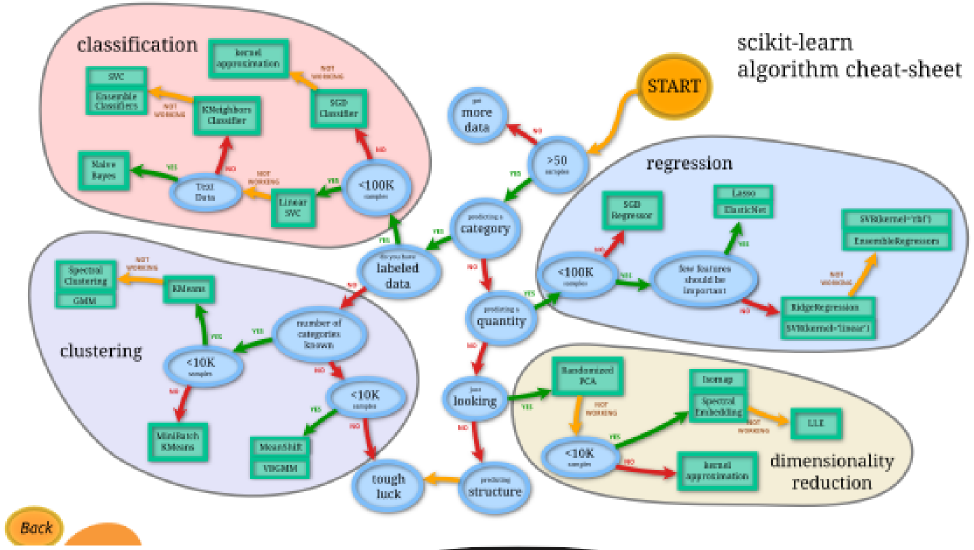
Key Technical Challenges

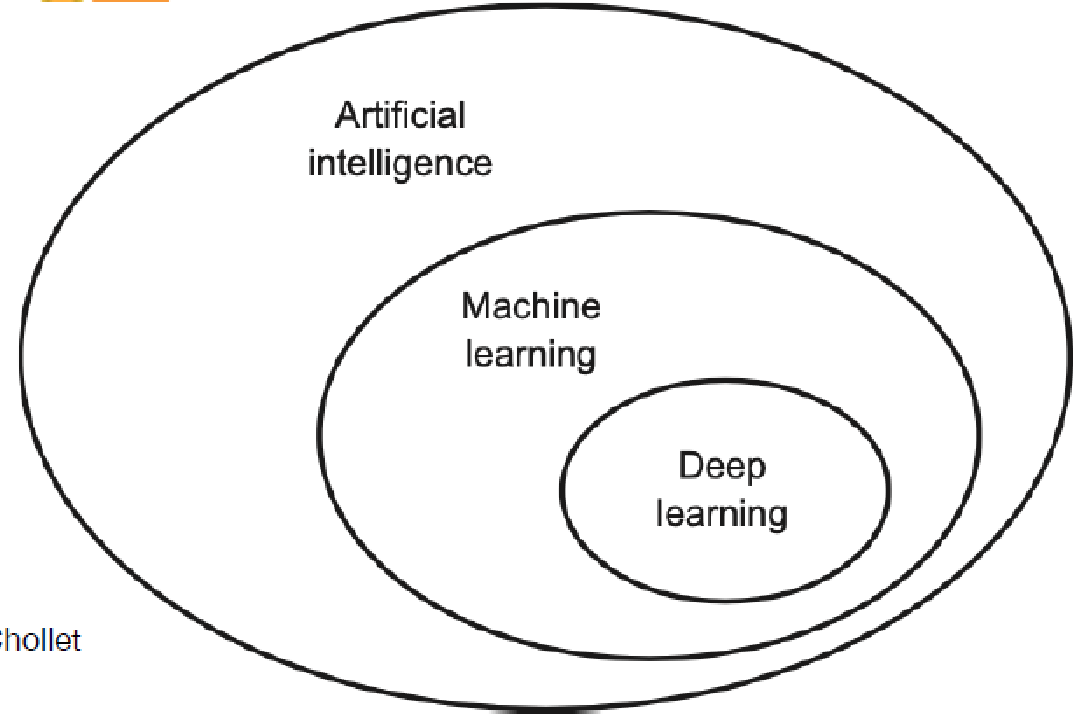
Pattern or correlation search
Clustering analysis, classification
Outlier / anomaly searches
Hyperdimensional visualization

↳ **Data Understanding**

↳ **New Knowledge**

+feedback





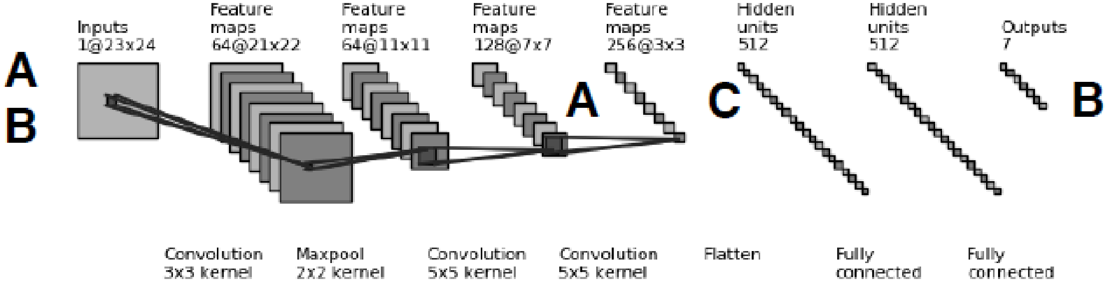
Chollet

Train computers using examples

Ashish Mahabal

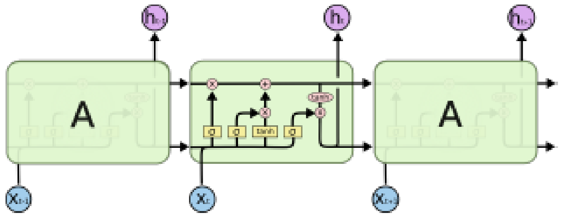
Convolutional neural networks

Big Imaging Data



Recurrent neural networks

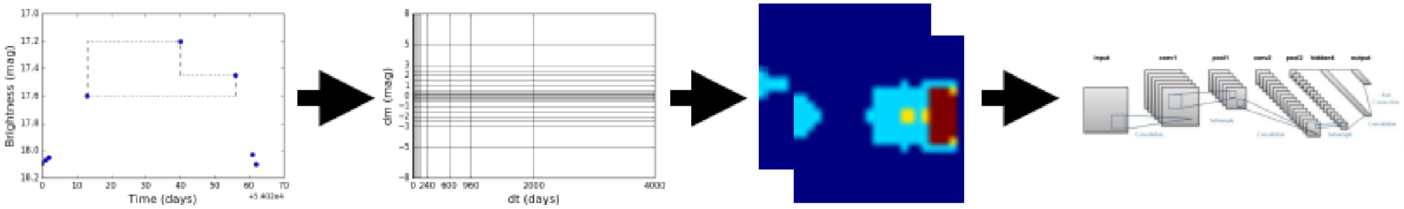
Sequence Data



Random Forest, Support Vector Machines

(Relatively) Small Data

Requirements

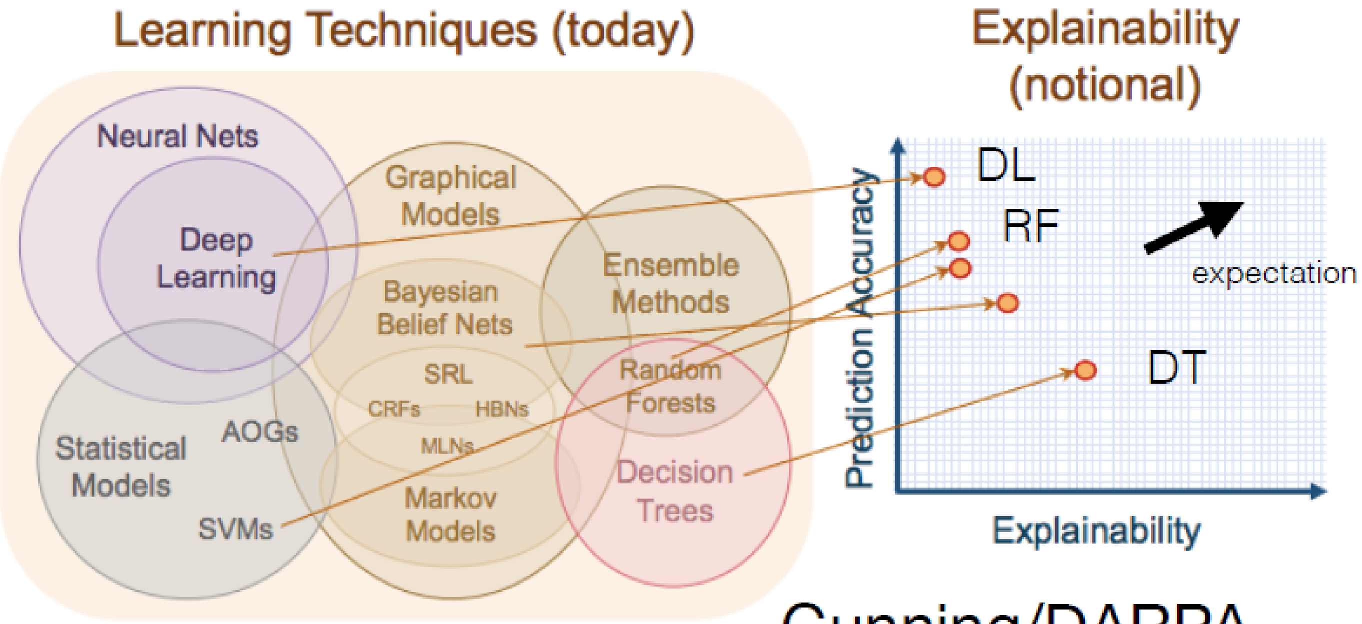


Large samples

Segmentation — Registration

Annotation — Standardization

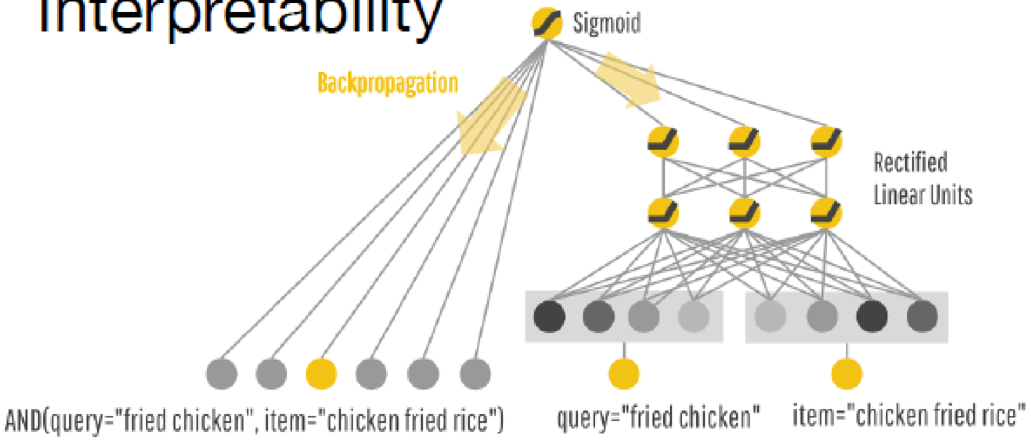
Combining multi-omics data



Gunning/DARPA

Wide networks

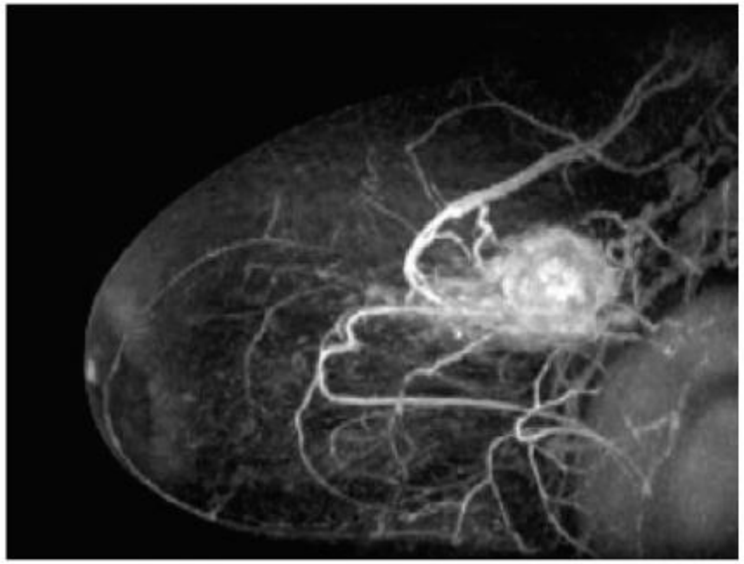
Interpretability

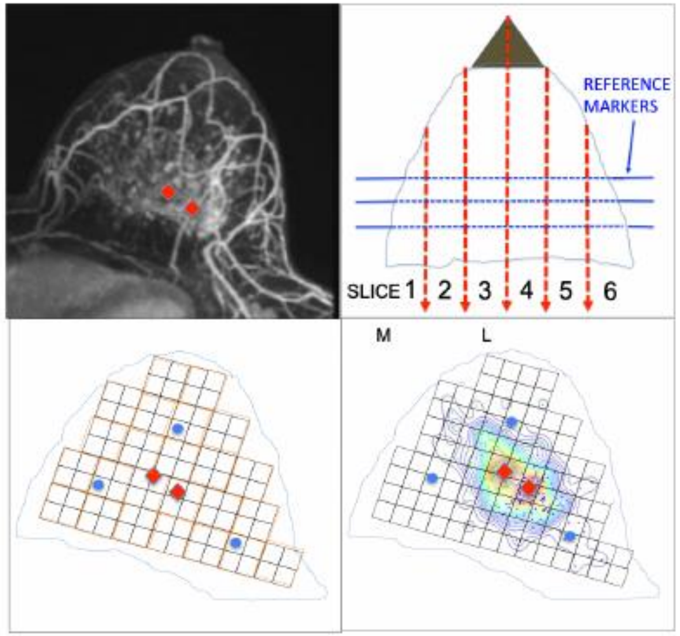


COH Project with Victoria Seewaldt, David Franckhouser, Daniel Crichton, …

- 51% Ashkenazi
- 38% European American
- **29% Hispanic American**
- **26% Asian American**
- **20% African American**
- **15% Palestinian**
- **5.6% Lebanese**

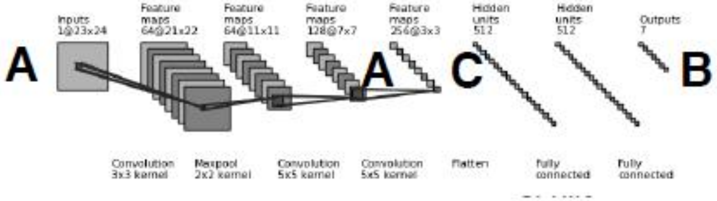
Greenup et al. *Ann Surg Onc* 2013
Jalkh et al. BMC Med Genomics, 2017

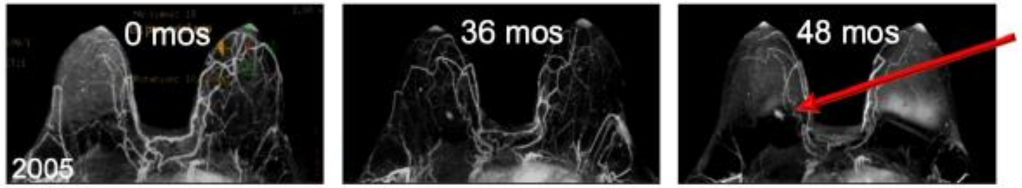




Neovascularization and progression

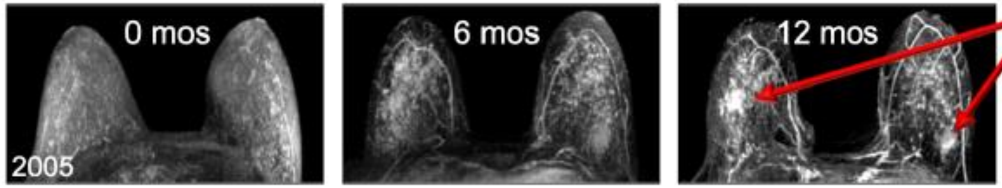
Multi-parametric 3-D portrait
transcript analysis
protein expression
mutational analysis





ER/PR+
T1N0

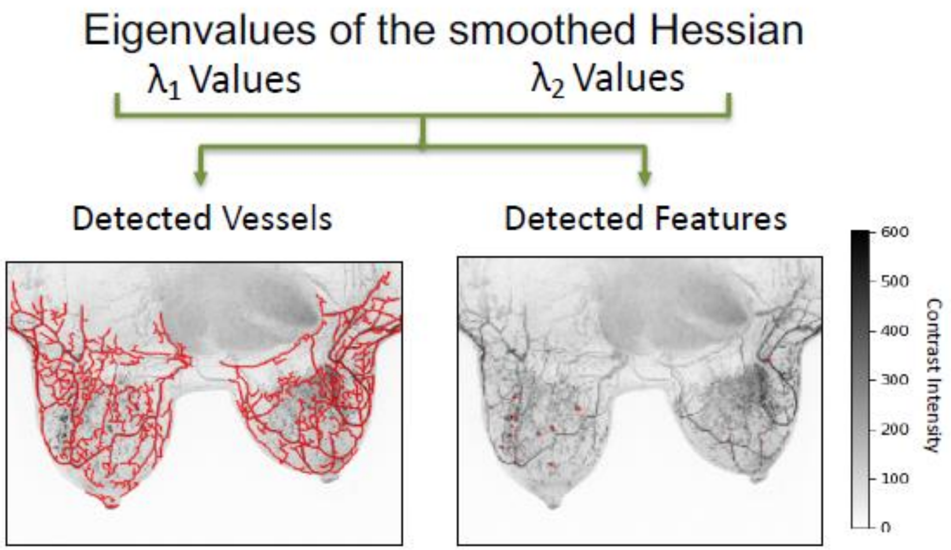
Slow (22/51 women)
Focal, benign biology



TNBC
T1N0

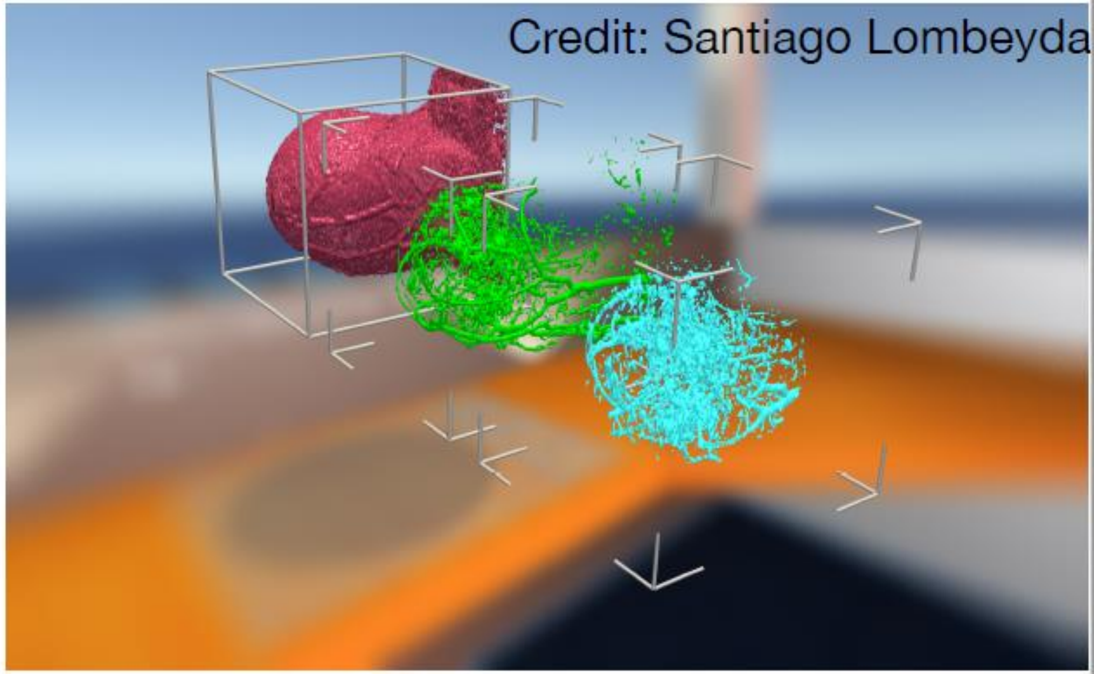
Accelerated (29/51 women)
Diffuse, aggressive biology

COH Project with Victoria Seewaldt and David Franckhouser



Breast Isosurfaces

Credit: Santiago Lombeyda



–Characterize the vascularity
- Quantify vasculature
- Identify neovascularity

–Novel feature identification
- Identify breast cancer development predictors
- Detect vasculature/neovascularity features

–Tumor detection
- Segment tumors
- Track tumor environment

Vascularity
– enhance contrast in filamentary structure
–*Quantify vascularity*

Feature Detection
–Identify ROIs based on contrast washout
–Assess feature dynamics over time

Image Reader Study for Breast Cancer with UCSF

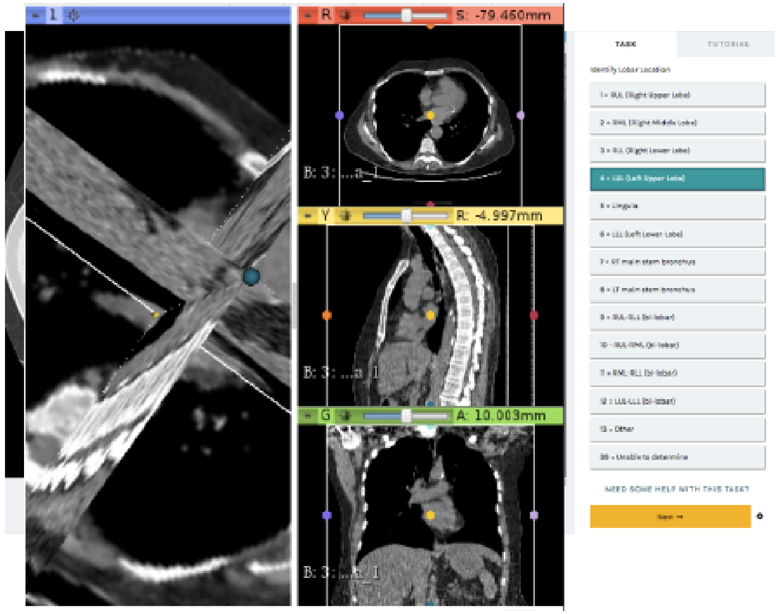
with Amrita Basu, Case Brabham, Laura Esserman, Rita Freimanis

Aim: Classify data/images
e.g. cancer/non-cancer,
or progression/no progression

Ingredients: (lots of) Data
e.g. Mammograms (2D)
MRI (3D)

DCIS active surveillance
progression/no-progression
MRI at least for invasive ones

Steps:
Obtain labels for training (tools radiologists routinely use)
Classify (Machine Learning)



Parallels our work on lungs

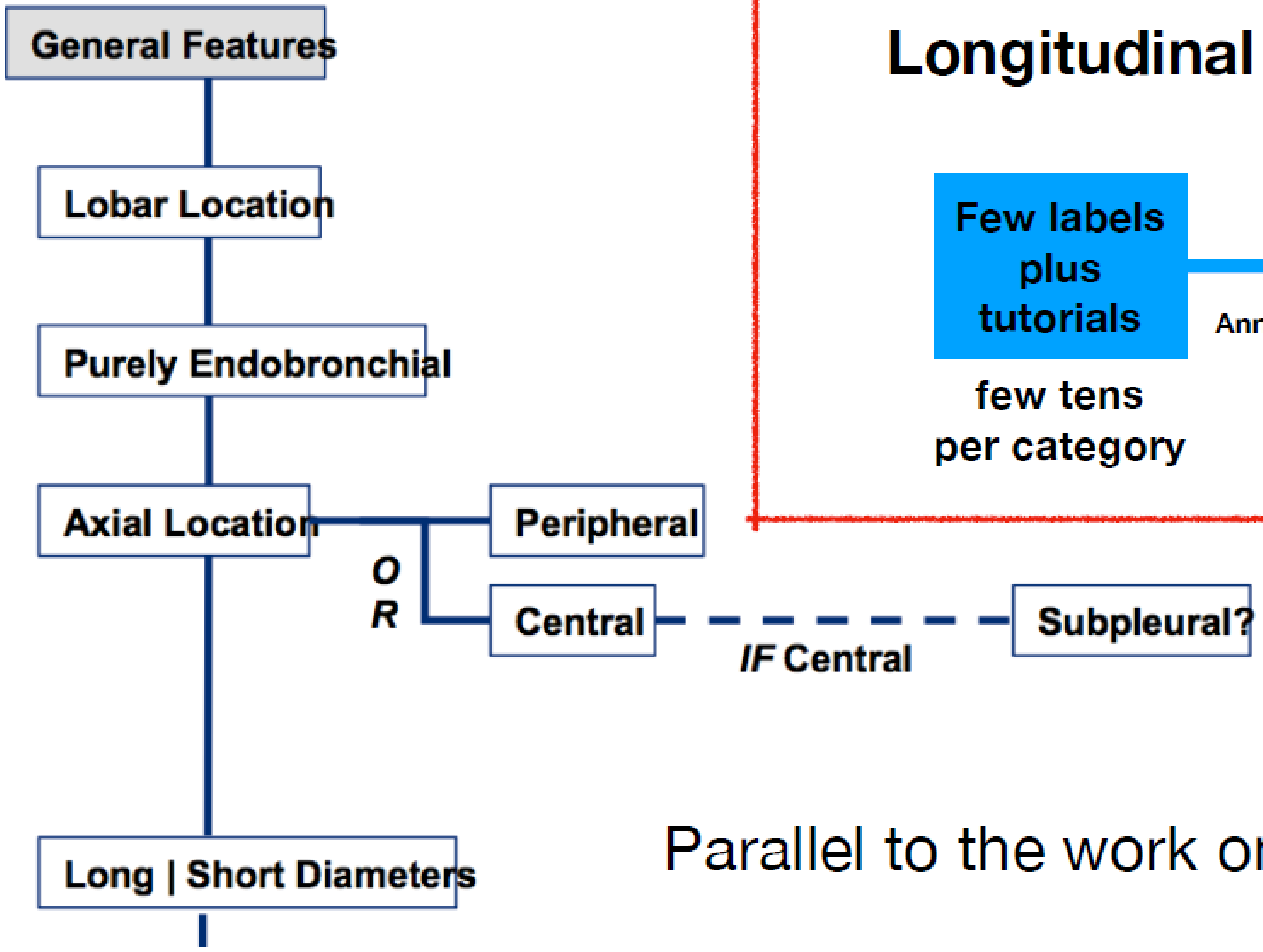
Machine Learning:

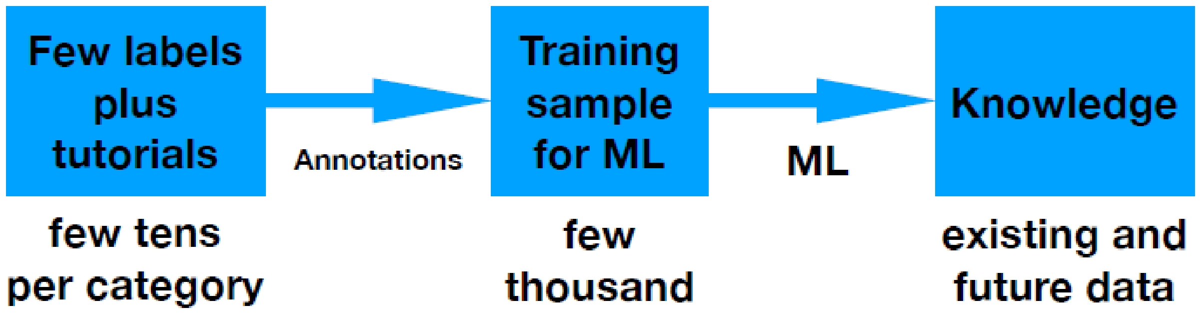
Lots of data: deep learning (convolutional networks)

Moderate data: traditional learning (Random forest, support vector machine)

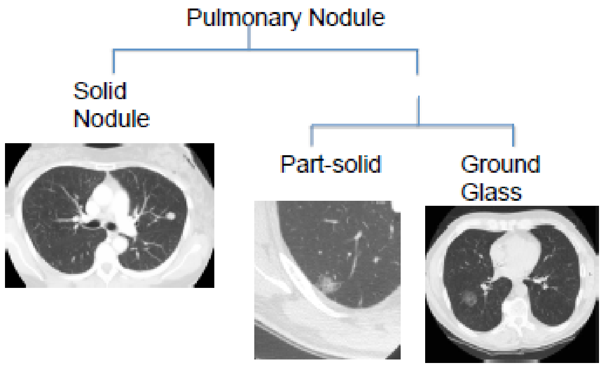
The more data there are, the better it is.

Longitudinal ideal - it will make training easier.





Parallel to the work on lungs

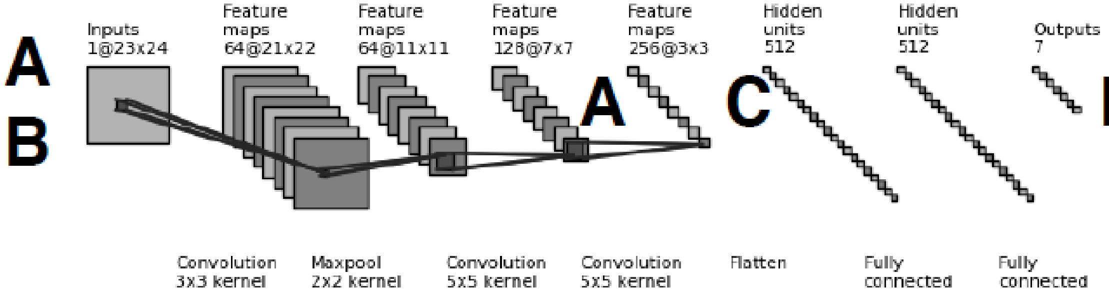


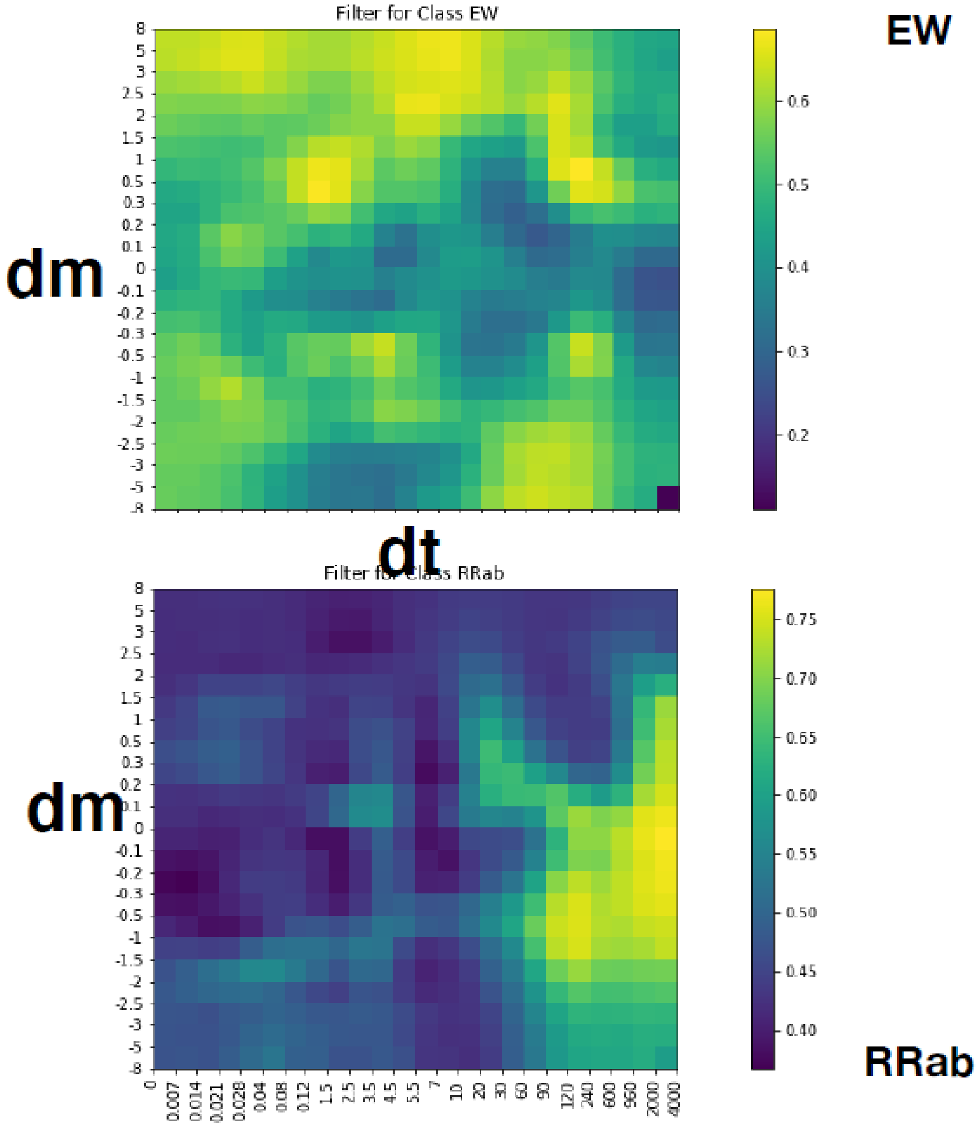
Visualization for interpretability

A. Activation Maximization
  - Initial layer filters easy to visualize
  - Generate input image that activates later filters

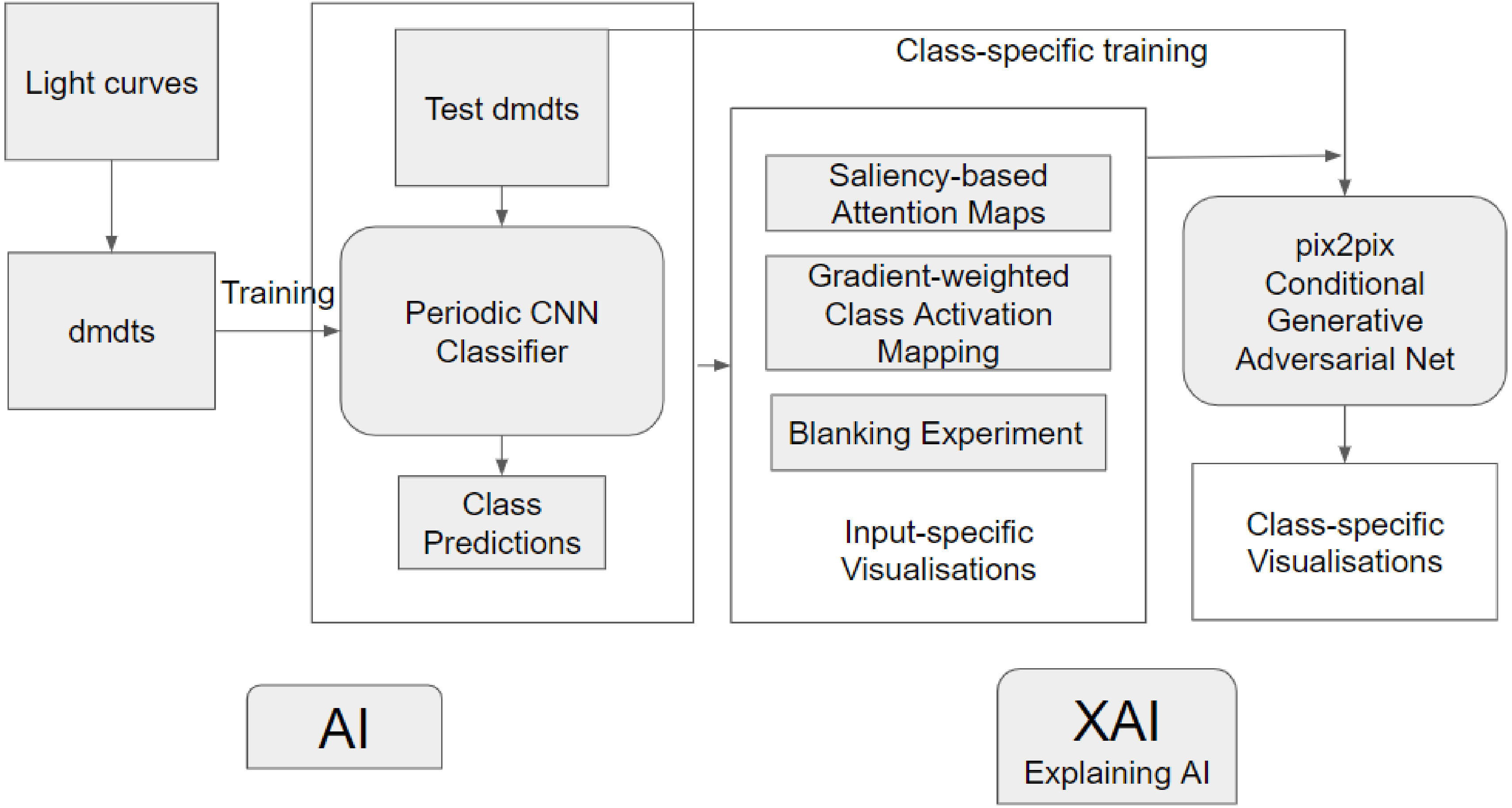
B. Saliency Maps
  - Gradient of o/p category wrt input image
  - Understanding attention of the classifier

C. Class Activation Maps
  - Gradients based on first dense layer
  - Spatial information still intact



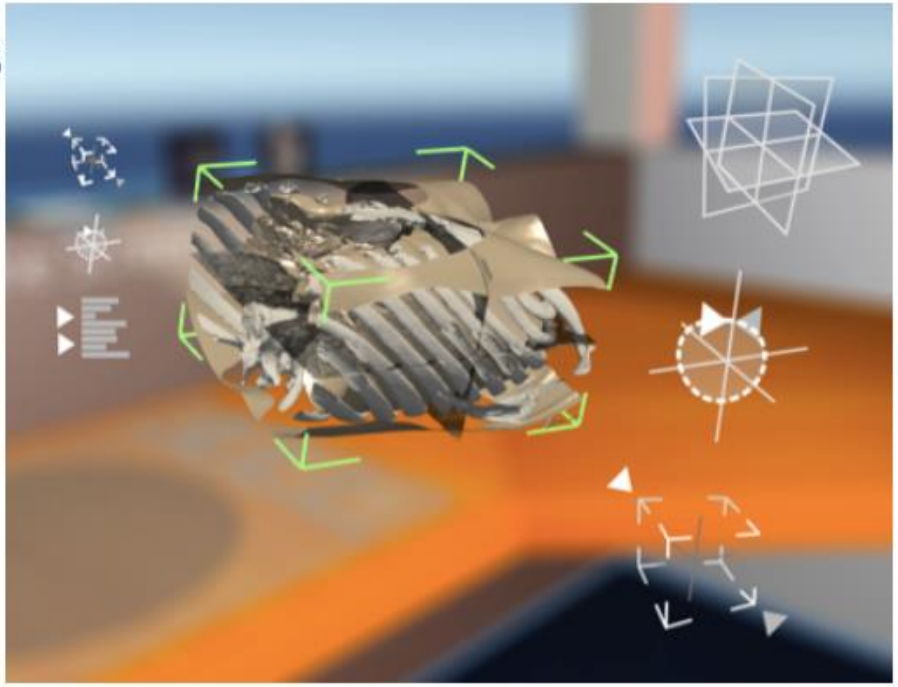


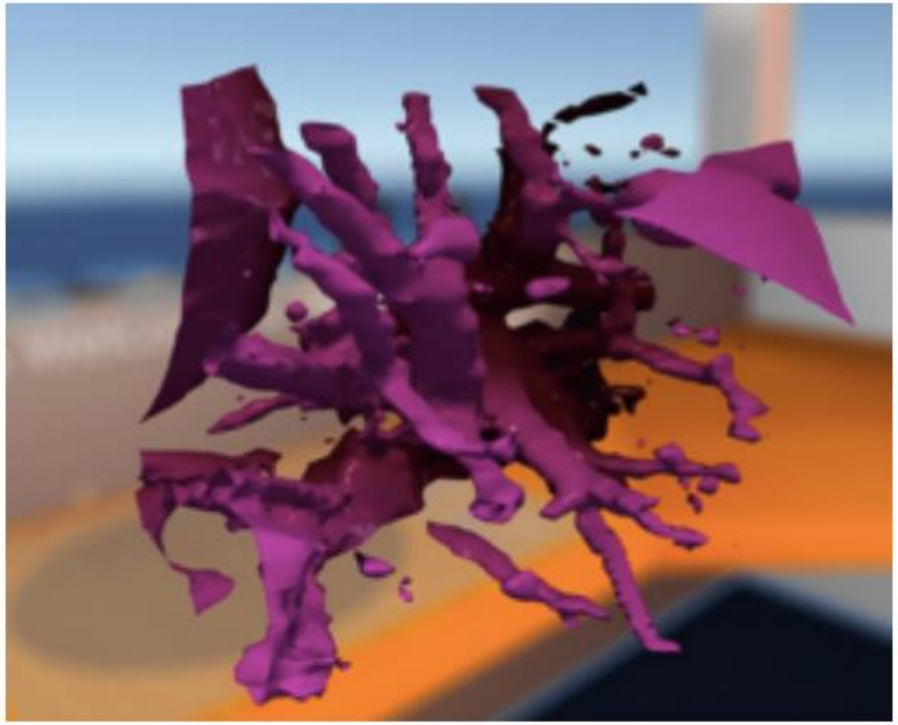
https://raghakot.github.io/keras-vis/



with Meet Gandhi

The (very near) future includes Virtual Reality explorations



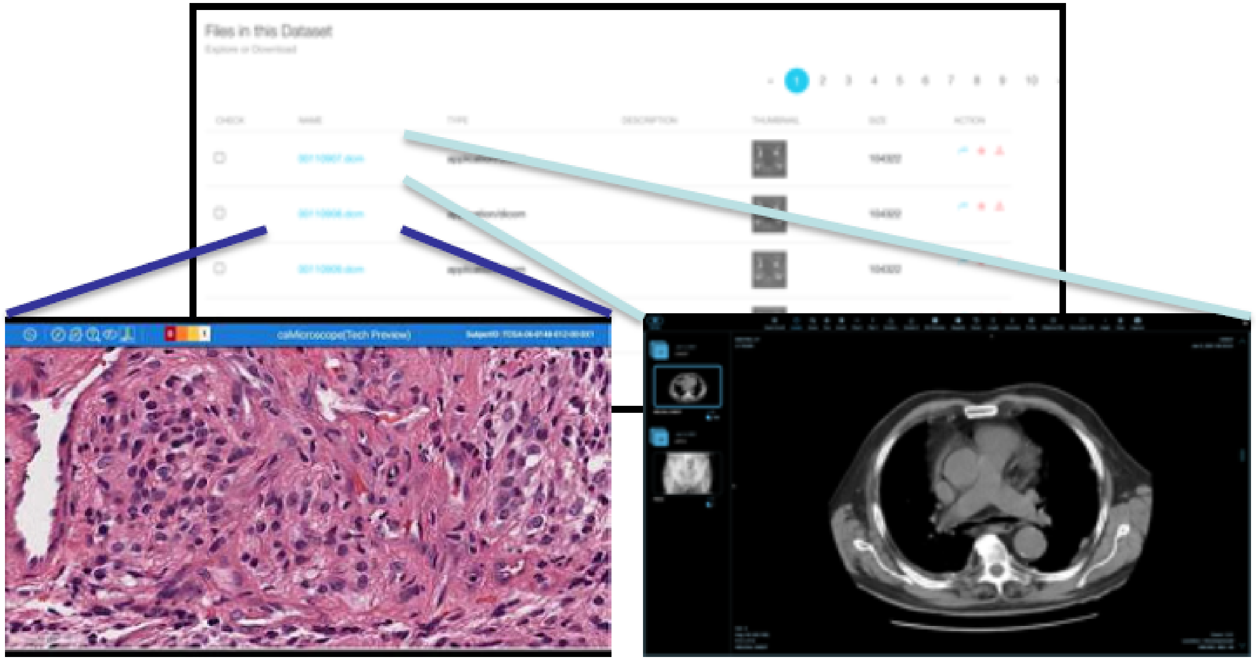


Use of VR for training, marking tumors, and comparisons

At SigGraph 2019

with Santiago Lombeyda

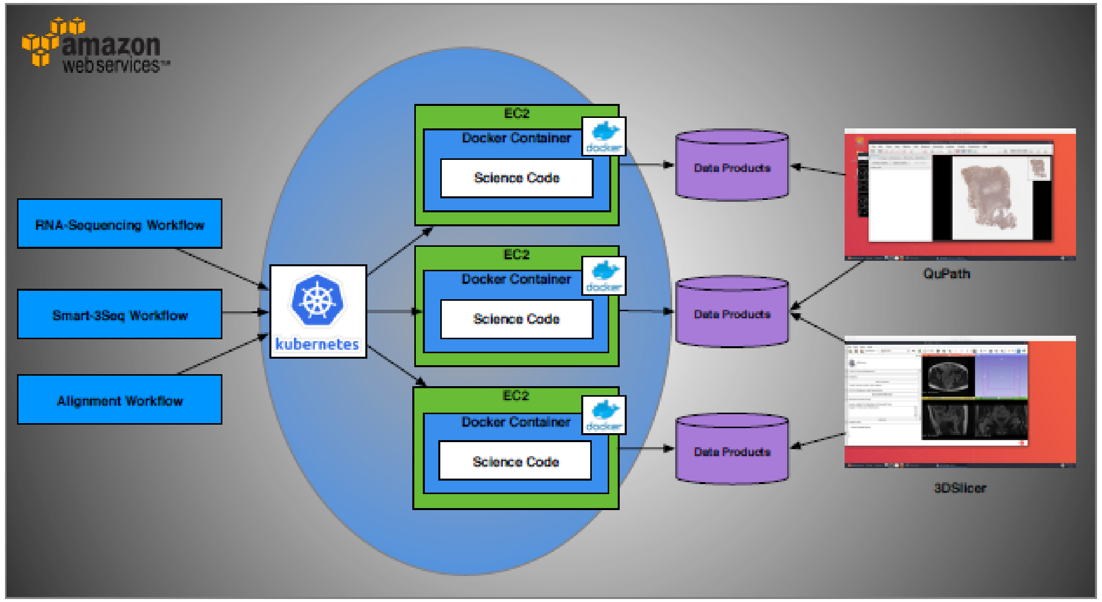
LabCAS structure and relevant holdings at JPL

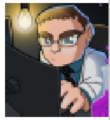


Multi-omics

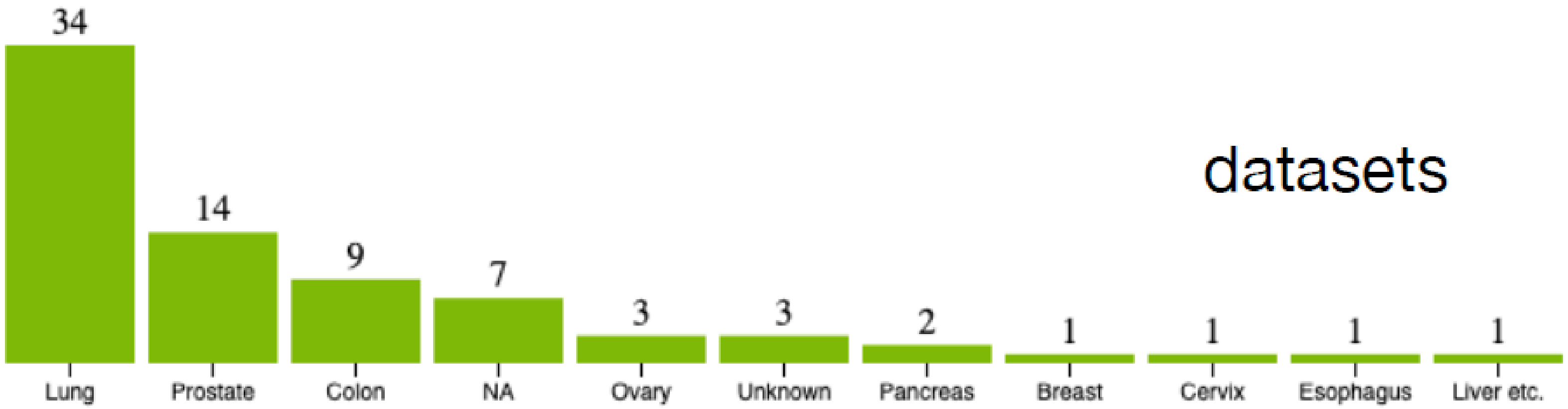
PI Dan Crichton

ML at scale



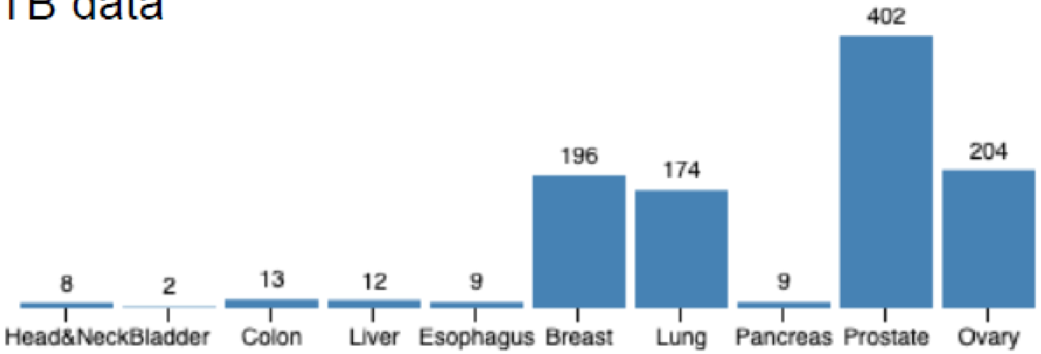


Increasing importance of biomarkers

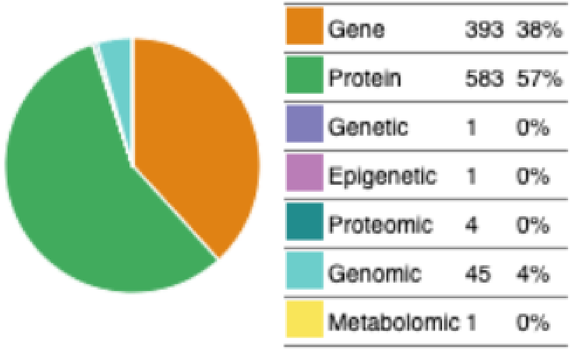


datasets

A biomarker data infrastructure consisting of ~1000 biomarkers, ~200 protocols, 1500 publications, 150 TB data



Biomarker Statistics:



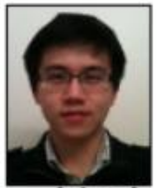
JPL Informatics Center
Data Science Team

Dan Crichton
NASA/JPL
Principal Investigator

Data Architecture →

- Luca Cinquini NASA/JPL
- Asitang Mishra NASA/JPL

Portal
System Engineering →



Sean Kelly **David Liu**
NASA/JPL

Project Management
Data Coordination →

Heather Kincaid
NASA/JPL

Machine Learning
Visualization →

- Alphan Altinok NASA JPL
- Santiago Lombeyda Caltech

Ashish Mahabal
Caltech

Bioinformatics →
Biomarker
Curation/CDEs

Kristen Anton **Maureen Colbert**
University of North Carolina

System Admin →
Cloud Computing

- Susan Neely NASA/JPL
- Rojeh Yaghoobi NASA/JPL

Paul Zimdars
NASA/JPL

Take-away points

Many fundamental science problems can be attacked by data science

Many tools/libraries available

Computing has become cheaper and faster

One universal requirement is good data

Large amounts of data used to be a requirement but …

There is enough for everyone to do. We are still plucking low-hanging fruits

Holistic analysis through abstractions and some domain knowledge is the key

Scope for cross-fertilization of ideas between bio- and physical sciences

Ashish Mahabal aam at astro.caltech.edu