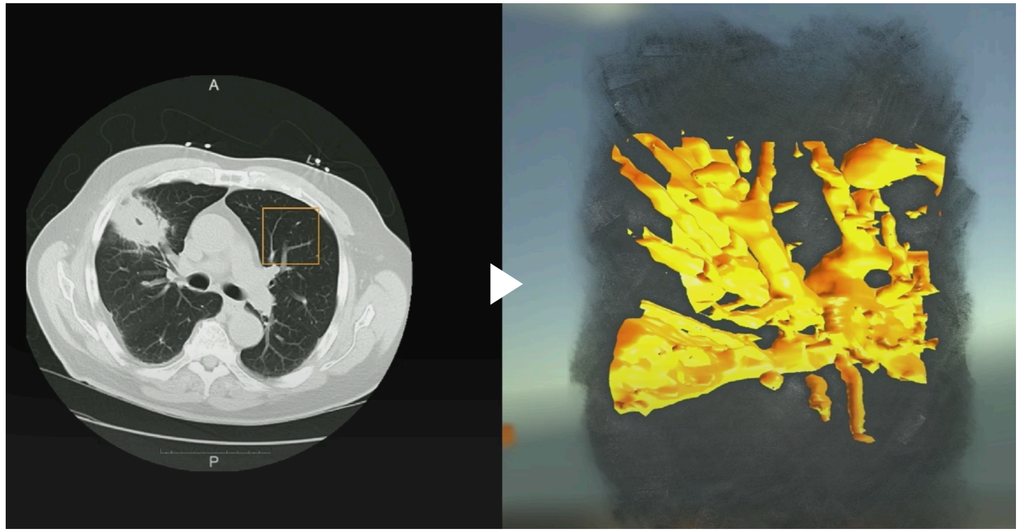
Methodology Transfer in the Era of Huge Datasets



Ashish Mahabal <aam at astro.caltech.edu>
Lead Computational Scientist
Center for Data Driven Discovery, Caltech

2021-02-11

2nd AI/DS Workshop for Earth and Space Sciences, JPL/Caltech

CENTER FOR DATA-DRIVEN DISCOVERY

ZTF

Caltech-JPL Executive Advisory Board on Computational Sciences

JPL
Jet Propulsion Laboratory
California Institute of Technology

Information and Data Science Program (JPL/815):
Program Strategy, Partnerships, Investments, Roadmaps

Center for Data Science and Technology (JPL/3902):
People, Projects, Technologies, Research

Center for Data-Driven Discovery(Caltech):
People, Projects, Technologies, Research

People: JPL and Caltech Data Science Research Affiliates

Projects: Shared NASA, NSF, NIH, DARPA, etc projects

Technologies: Core Caltech and JPL Product Lines and Thrusts

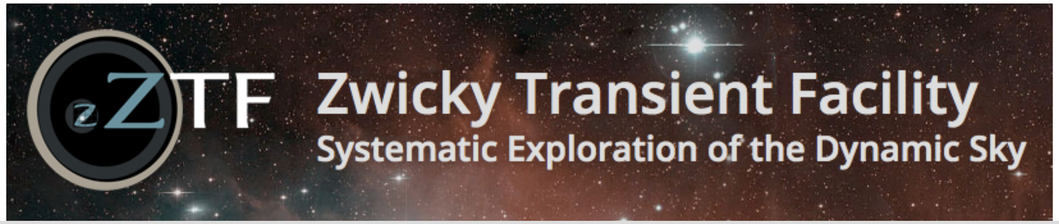
Dan Crichton
Rich Doyle

George Djorgovski

Primary Target: NASA/FFRDC Community, Agencies
Technology Capabilities: Architectures, Cyberinfrastructure, Machine Learning, Visualization
Large-Scale Projects

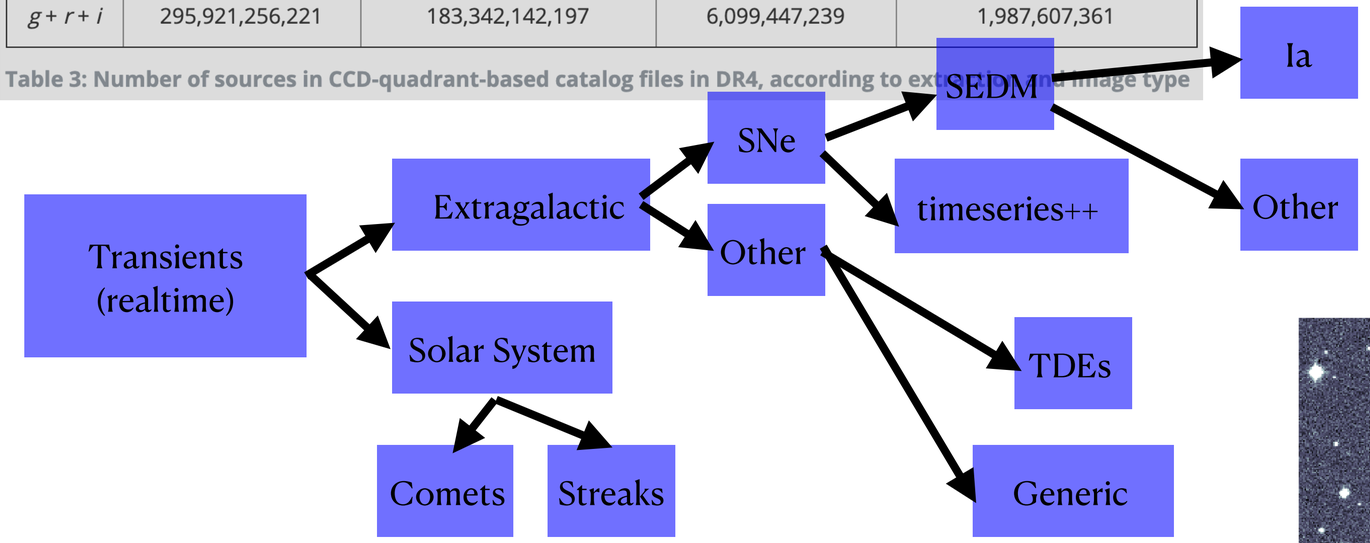
Primary Target: Education, Focused Research, Foundations
Technology Capabilities: Machine Learning, et al

ZTF ML Lead

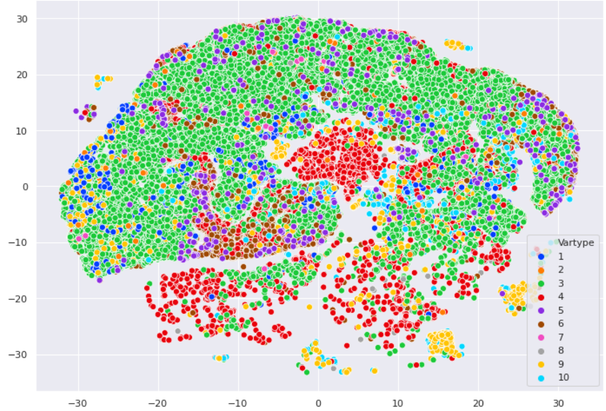


Filter(s)	#PSFcat-*sci* sources	#Aperturecat-*sci* sources	#PSFcat-*ref* sources	#Aperturecat-*ref* sources
g	74,126,685,884	46,764,103,340	2,205,550,406	689,728,722
r	217,553,318,884	134,048,046,583	3,051,221,036	1,040,799,684
i	4,241,251,453	2,529,992,274	842,675,797	257,078,955
g + *r* + *i*	295,921,256,221	183,342,142,197	6,099,447,239	1,987,607,361

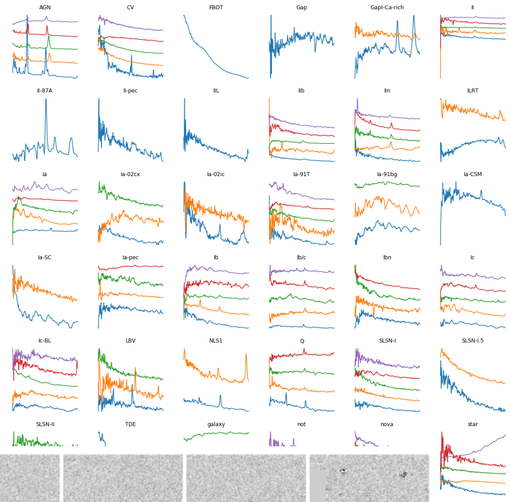
Table 3: Number of sources in CCD-quadrant-based catalog files in DR4, according to ext[illegible] and image type

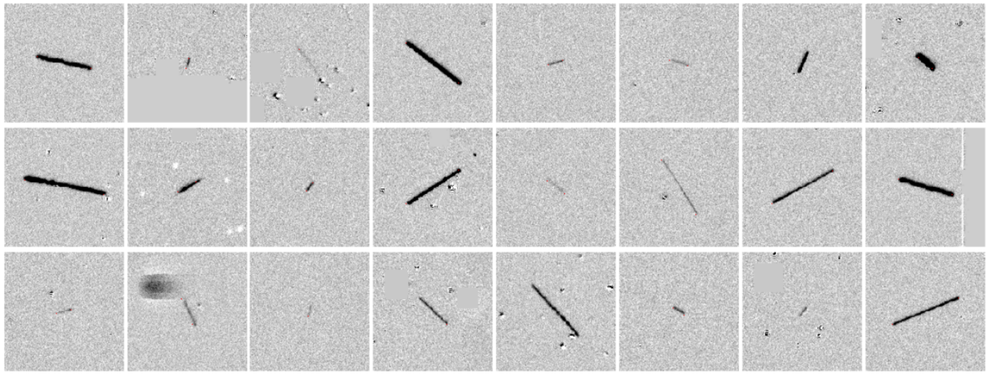


Variables

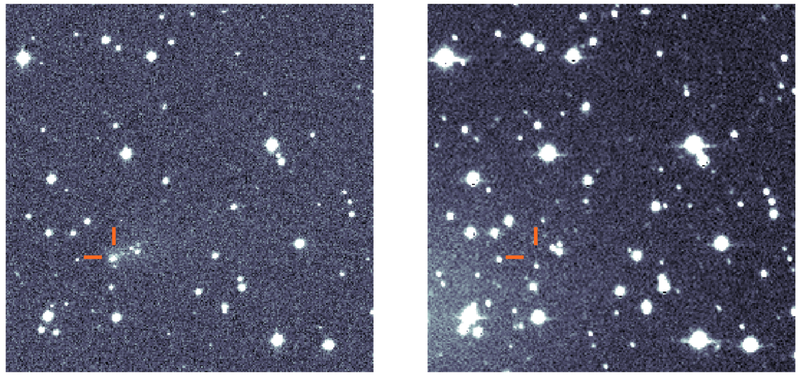


Supernovae

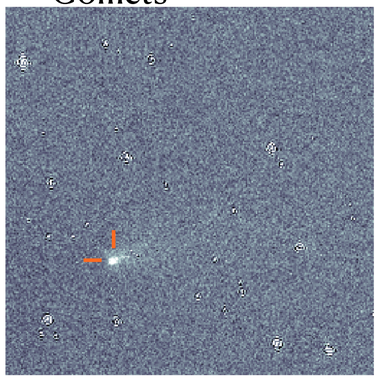




Streaking asteroids

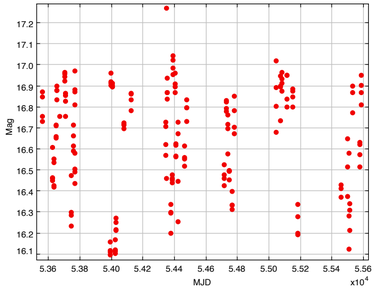


Comets

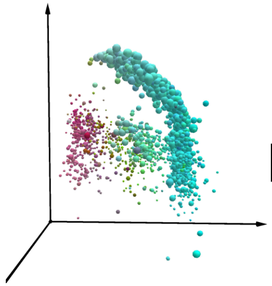


Mahabal et al. 2017 http://arxiv.org/abs/1709.06257v1
Duev et al 2019 http://arxiv.org/abs/1904.05920v2
Duev et al. 2021 NuerIPS
Van Roestel et al. 2021 Submitted
Fremling 2021 In prep. Ashish Mahabal

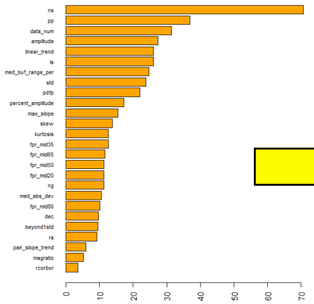
Plus Gravitational Waves, Exoplanets, ...



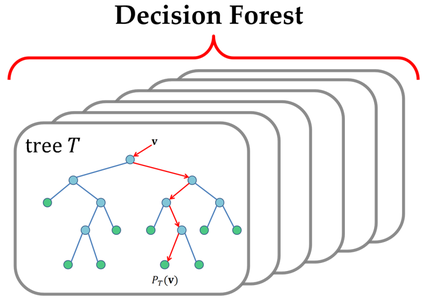
Light curves



Feature vectors



Dimensionality Reduction



Classification

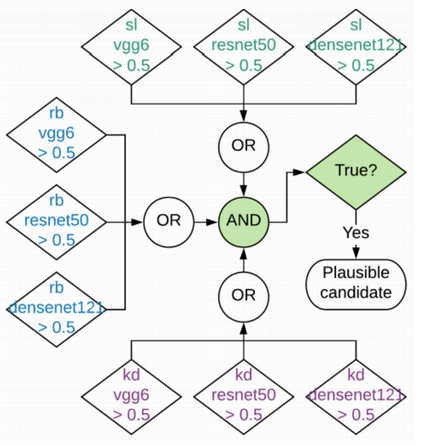
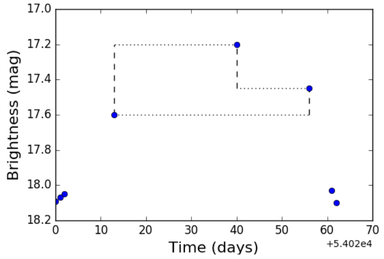
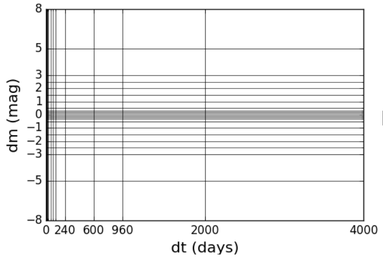


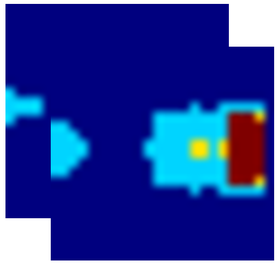
Figure 2. Decision logic used by **DeepStreaks** to identify plausible streaks. The problem is split into three simpler sub-problems, each solved by a dedicated group of classifiers assigning real vs. bogus ("rb"), short vs. long ("sl"), and keep vs. ditch ("kd") scores. At least one member of each group must output a score that passes a pre-defined threshold. See Section 2.1 for details.



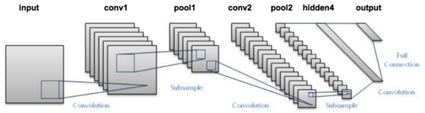
Light curves



Density representation

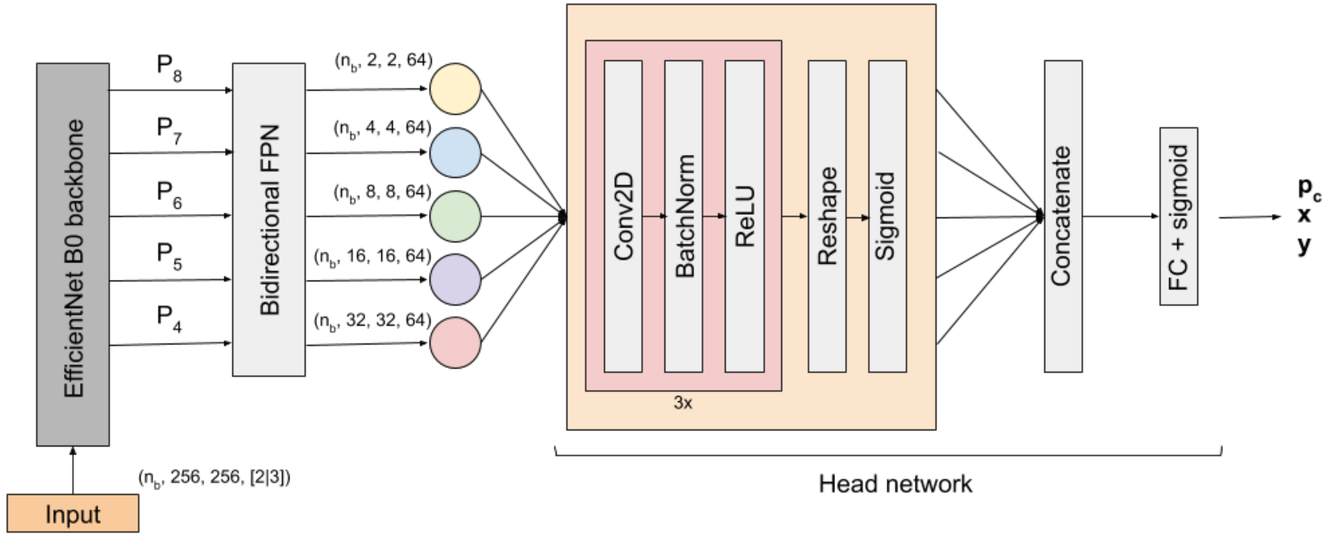


Equi-area images



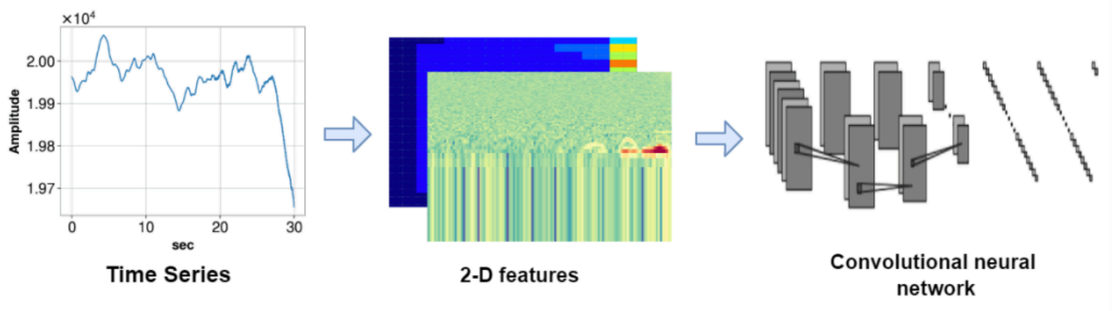
Convolutional Neural Network

Mahabal et al. 2017 http://arxiv.org/abs/1709.06257v1
Duev et al 2019 http://arxiv.org/abs/1904.05920v2
Duev et al. 2021 NuerIPS
Van Roestel et al. 2021 Submitted
Fremling 2021 In prep.

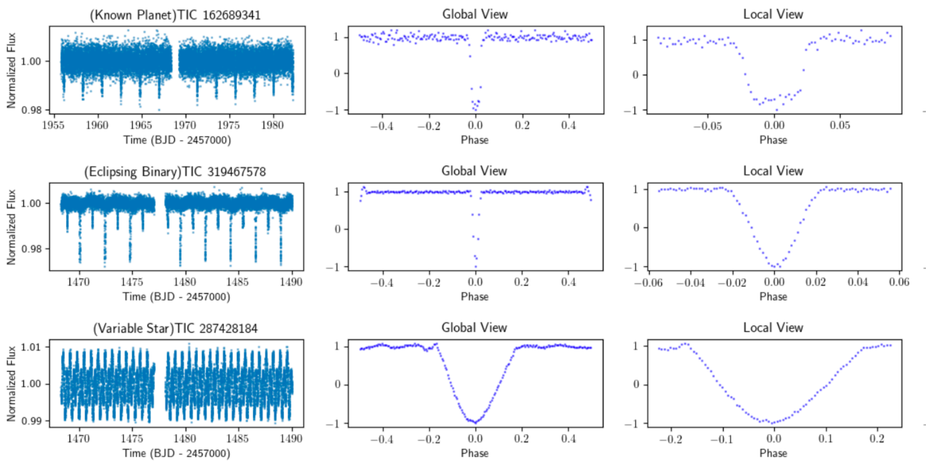


Ashish Mahabal

Other astro applications

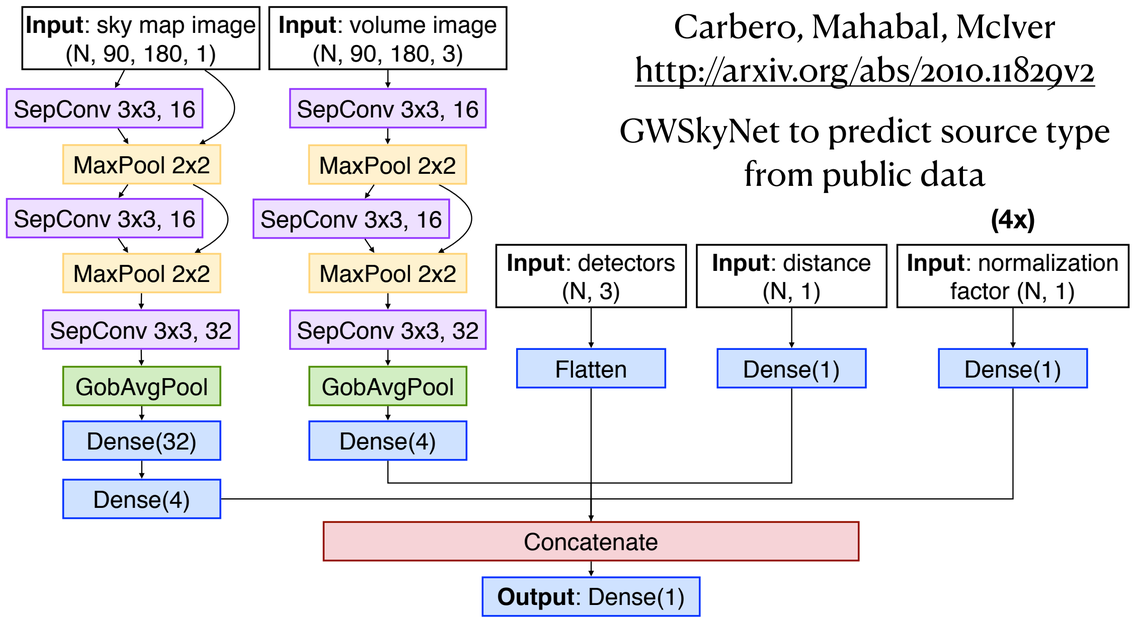


Improving LIGO duty cycle Biswas, Mahabal, McIver



Identifying Exoplanet candidates from TESS data

Rao, Mahabal et al. 2021 http://arxiv.org/abs/2101.09227v1



Carbero, Mahabal, McIver
http://arxiv.org/abs/2010.11829v2

GWSkyNet to predict source type from public data

Manay other astronomy applications
e.g. LIGO data for Gravitational waves,
TESS data for Exoplanets,
Deep-Wide-Fast data for anomalies etc.

JPL Informatics Center
Data Science Team

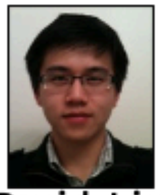
Dan Crichton
NASA/JPL
Principal Investigator

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

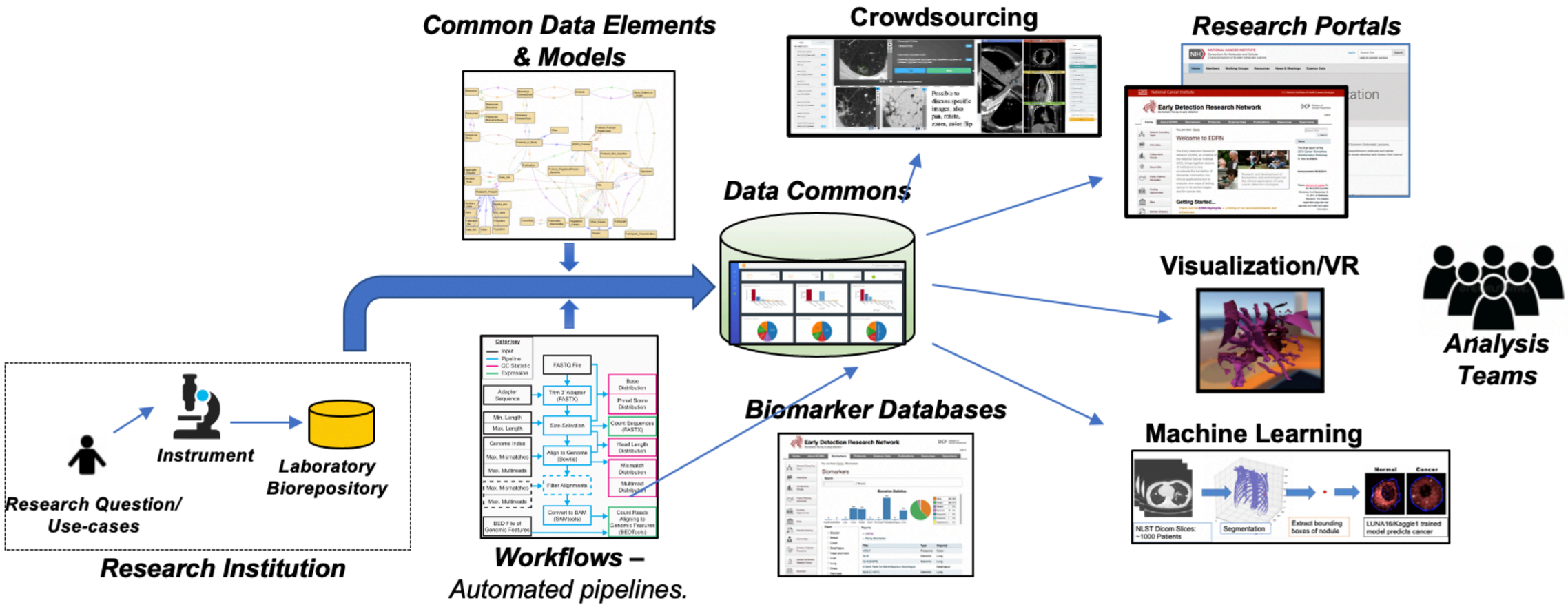
University of North Carolina

System Admin
Cloud Computing ⟹

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

Paul Zimdars
NASA/JPL

PI Dan Crichton

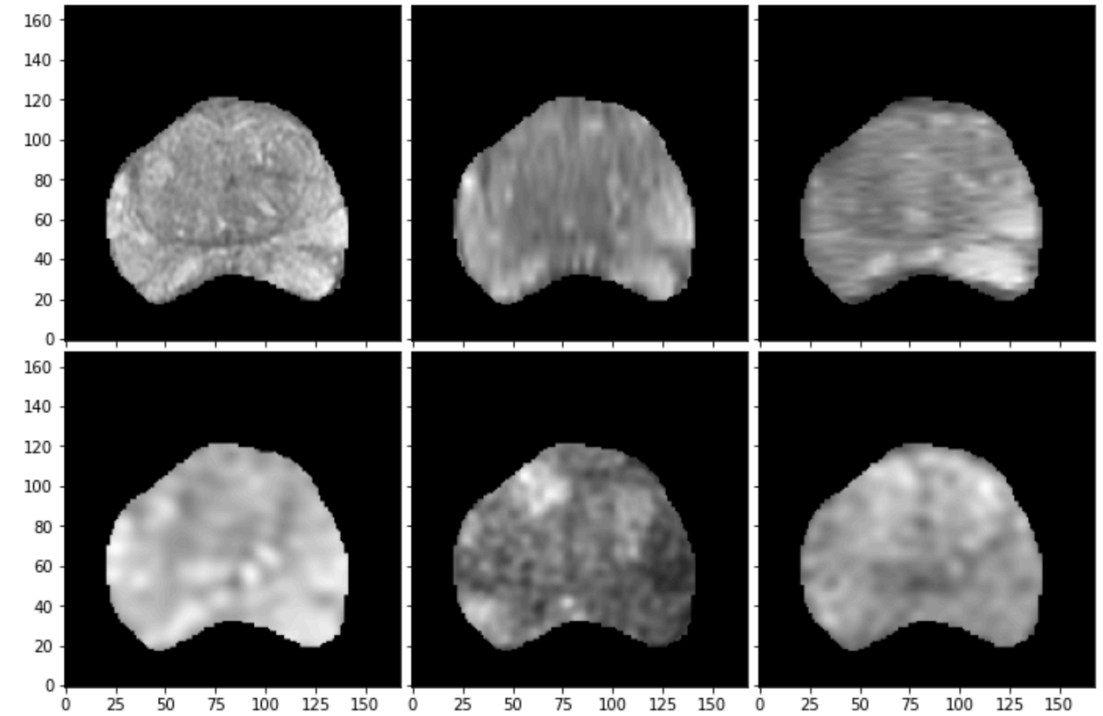


Early-Detection Research Network (EDRN)
Consortium for the Characterization of Screen Detected Lesions (MCL)

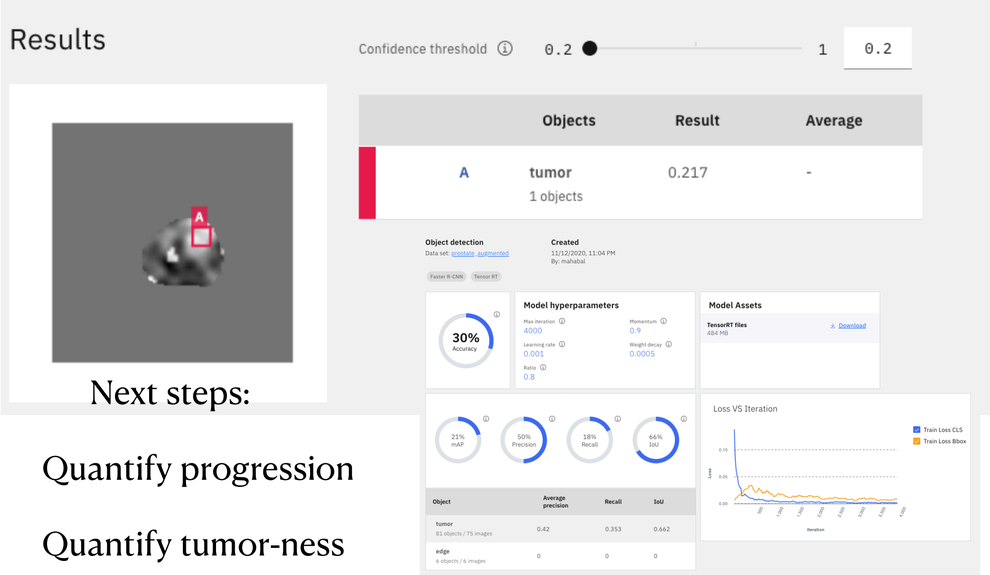
Finding Prostate Cancers

with Stoyanova, Crichton, ...

- ~200 patients, ~300 epcohs total (annual visits)
- Progressive (PR) and Active Surveillance (AS)
- Data cubes (168x168x168)
- Six modes with a mask

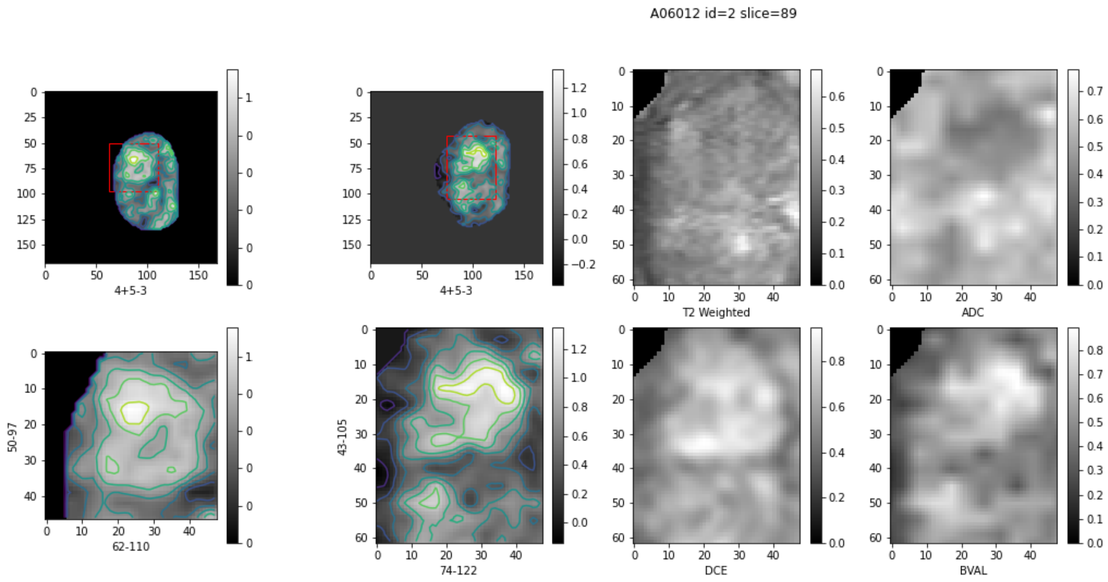


Using IBMs Visual Environment and follow-up local python programs



Next steps:

- Quantify progression
- Quantify tumor-ness

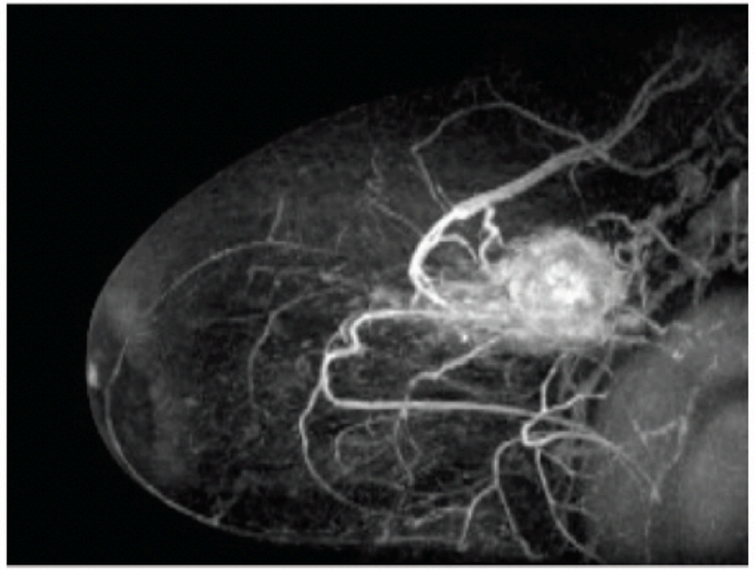


Ashish Mahabal

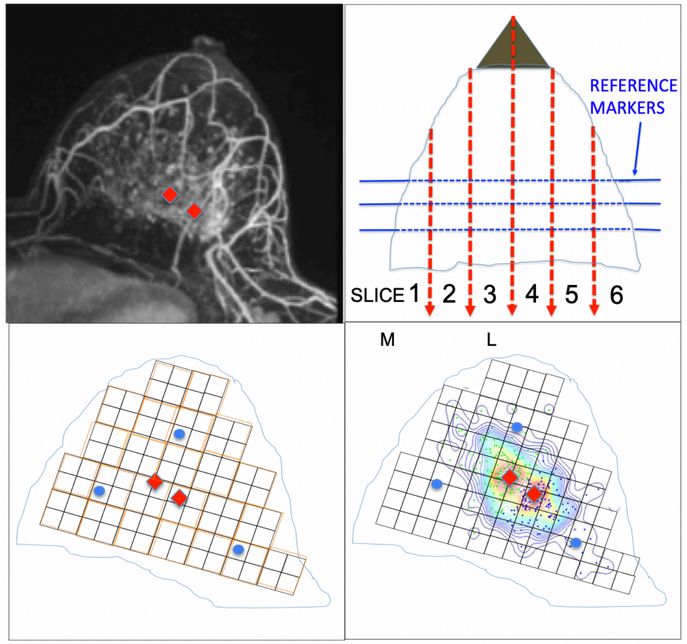
COH Project with Victoria Seewaldt, David Franckhouser, Daniel Crichton, …

- 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

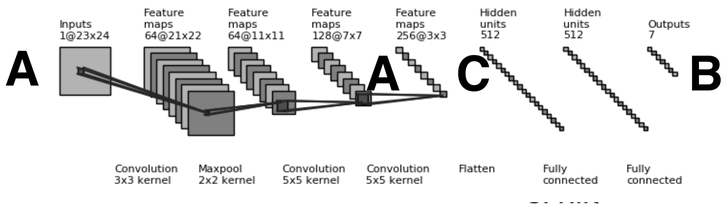


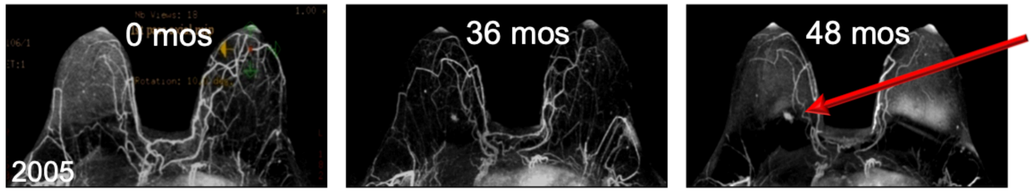
Ashish Mahabal



Neovascularization and progression

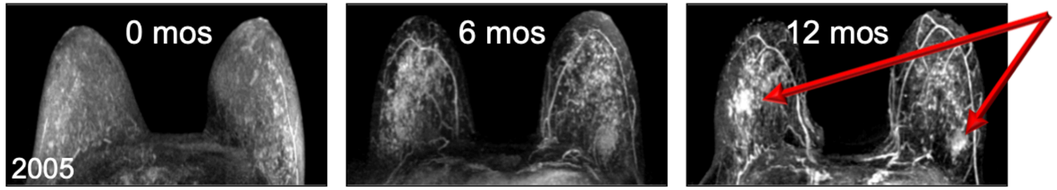
Multi-parametric 3-D portrait
transcript analysis
protein expression
mutational analysis





ER/PR+
T1N0

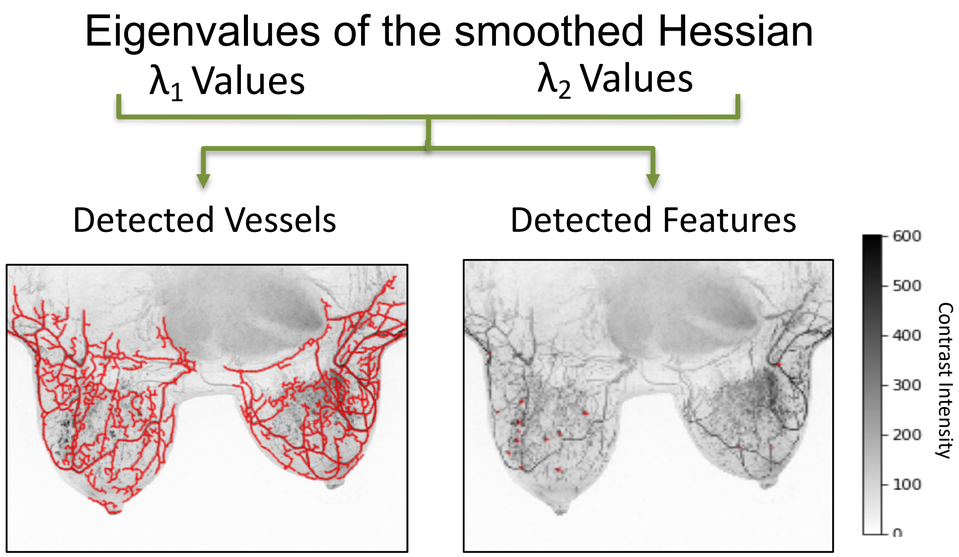
Slow (22/51 women)
Focal, benign biology



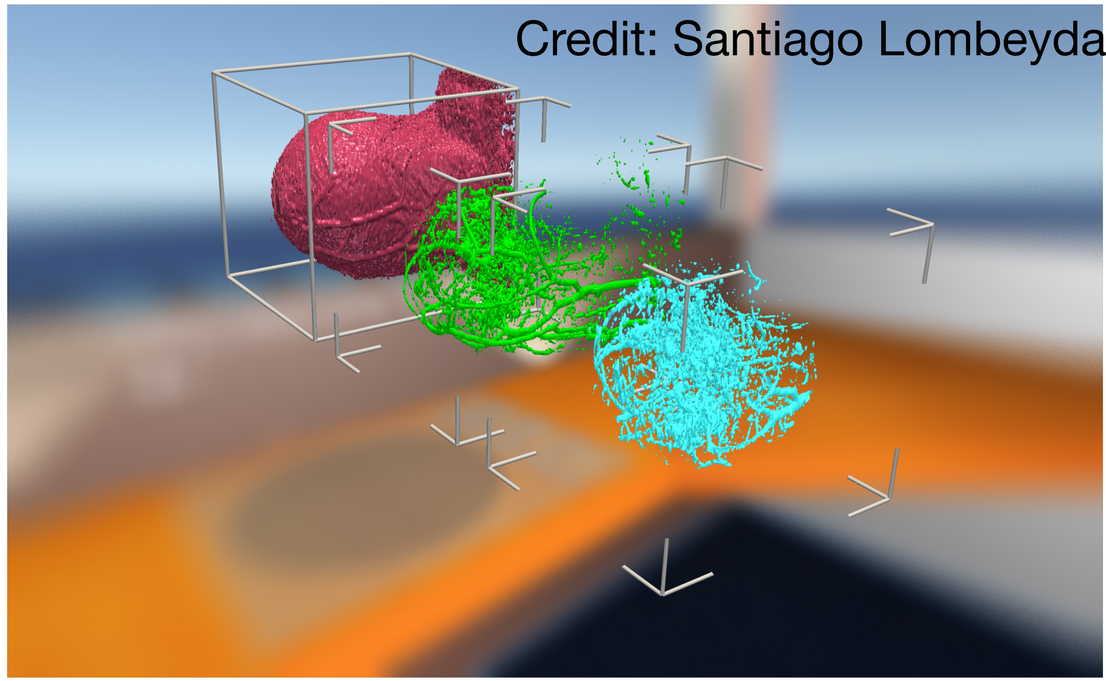
TNBC
T1N0

Accelerated (29/51 women)
Diffuse, aggressive biology

COH Project with Victoria Seewaldt and David Franckhouser



Breast Isosurfaces



–Characterize the vascularity
- Identify neovascularity

–Novel feature identification
- Identify predictors

–Tumor detection
- Segment, and track environment

Vascularity
– enhance contrast and Quantify

Feature Detection
–Identify ROIs, Assess feature dynamics over

REF Frankhauser et al. 2020 https://pubmed.ncbi.nlm.nih.gov/33001587/

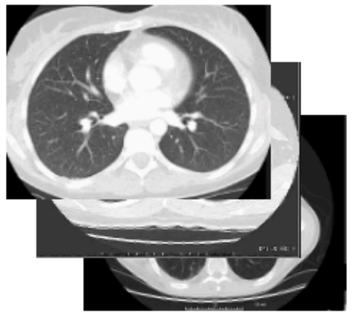
Image Atlas and ML to understand bias and identify smaller/early tumors

With Crichton, Kincaid, Aberle, Schabath,...

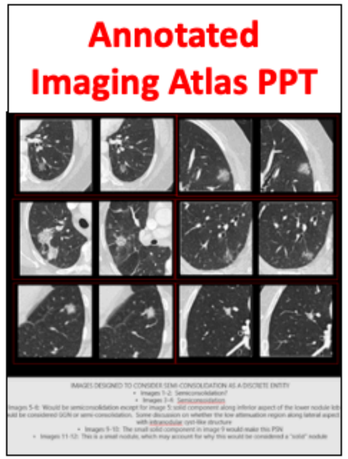
Few labels plus tutorials → Annotations → **Training sample for ML** → ML → **Knowledge**

few tens per category | few thousand | existing and future data

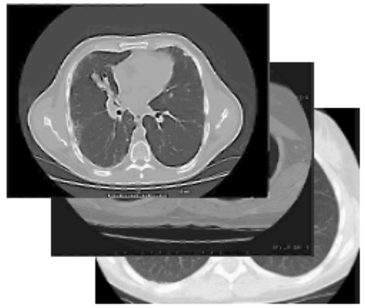
Reader Study



Pre-training read



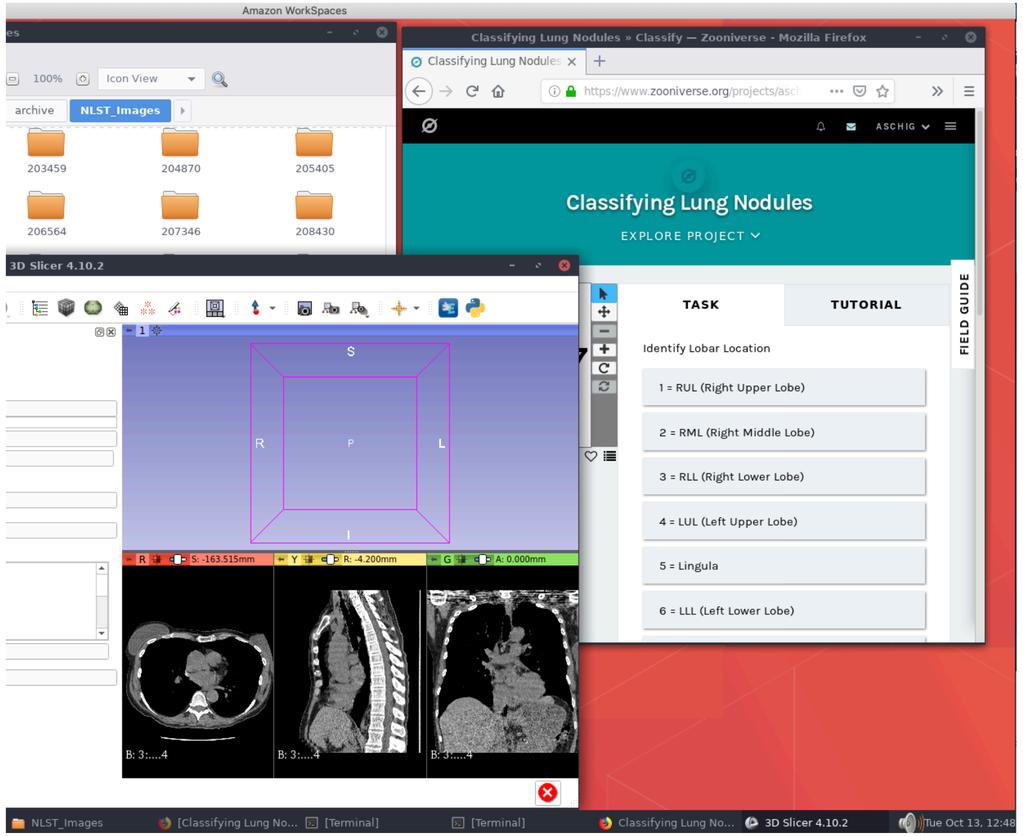
Annotated Imaging Atlas PPT



Post-training read

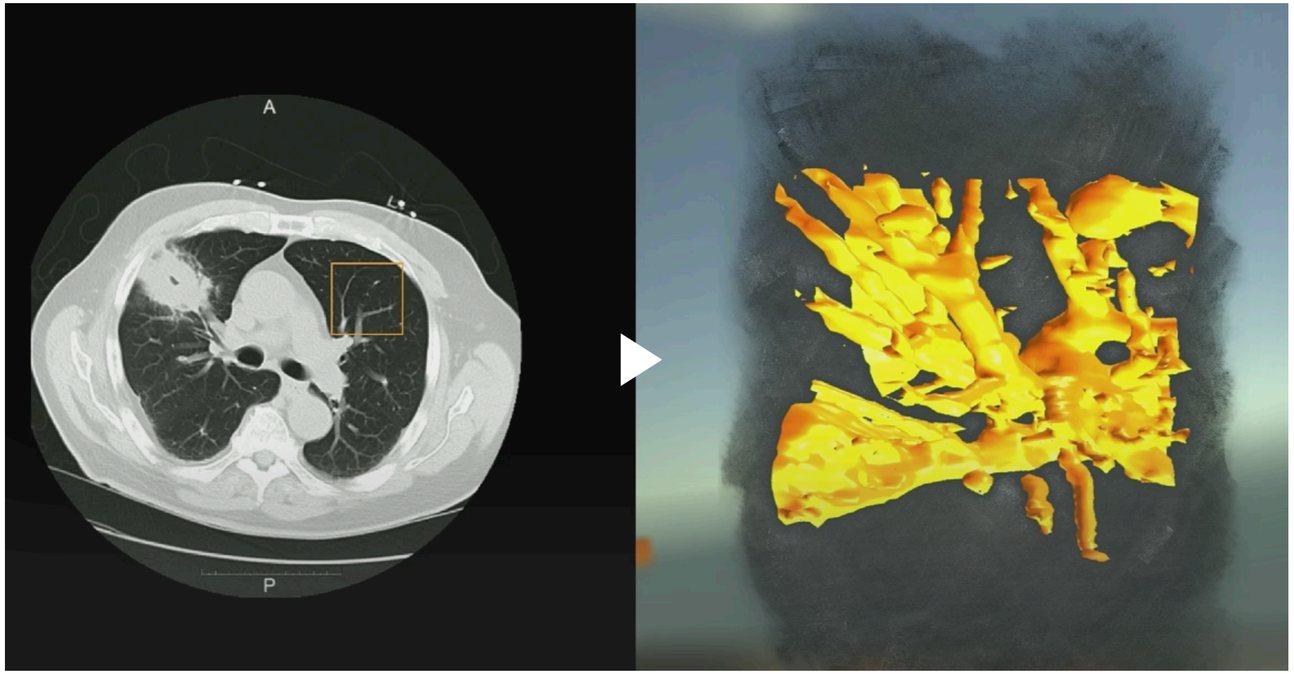
- 8 - 10 naïve radiologists will score nodules from 50 images
- JPL created hosting infrastructure for the reading study (presented next)

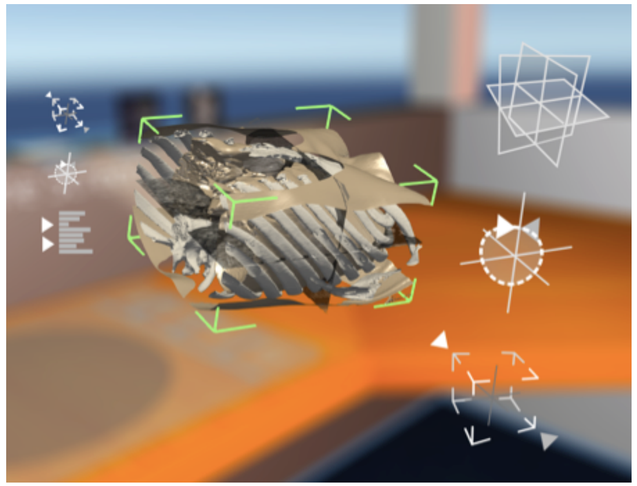
- Same radiologists will score nodules from 50 images (20 images from pre-training)
- Intra- and inter-accuracy will be assessed for each feature

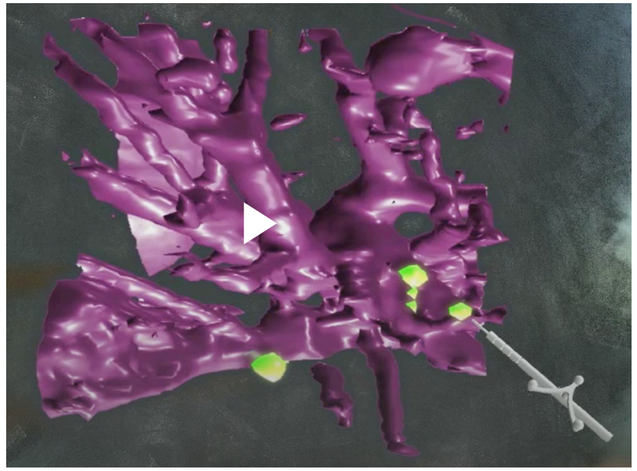


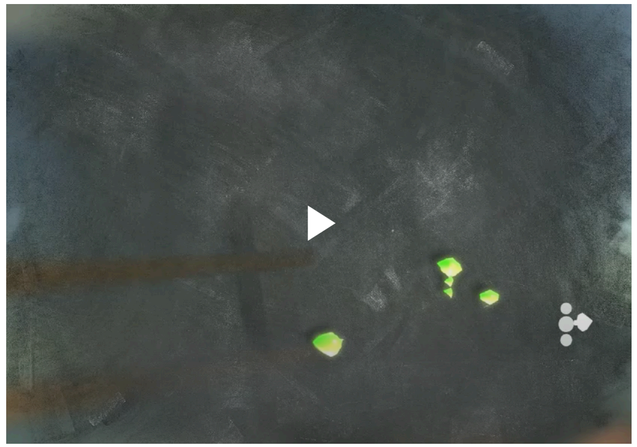
Zooniverse + 3DSlicer on AWS

At SigGraph 2019 with Santiago Lombeyda







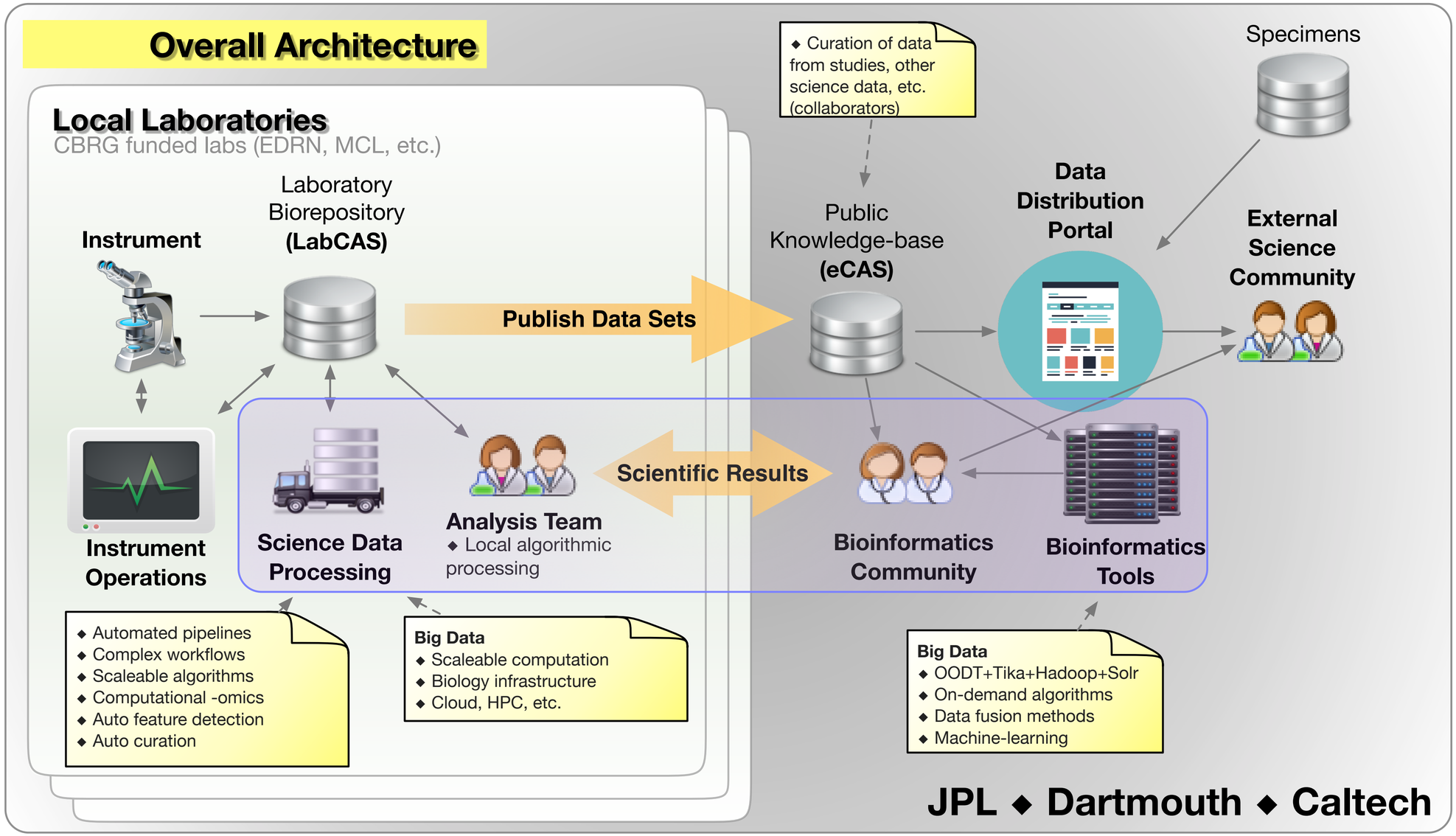


Use of VR for training, marking tumors, and comparisons

REF Lombeyda et al. https://dl.acm.org/doi/10.1145/3306449.3328825

AWS
Google Cloud
IBM

Planetary Science
Cancers
Planetary Protection
Astronomy



Dan Crichton

Methodology transfer to biosciences++

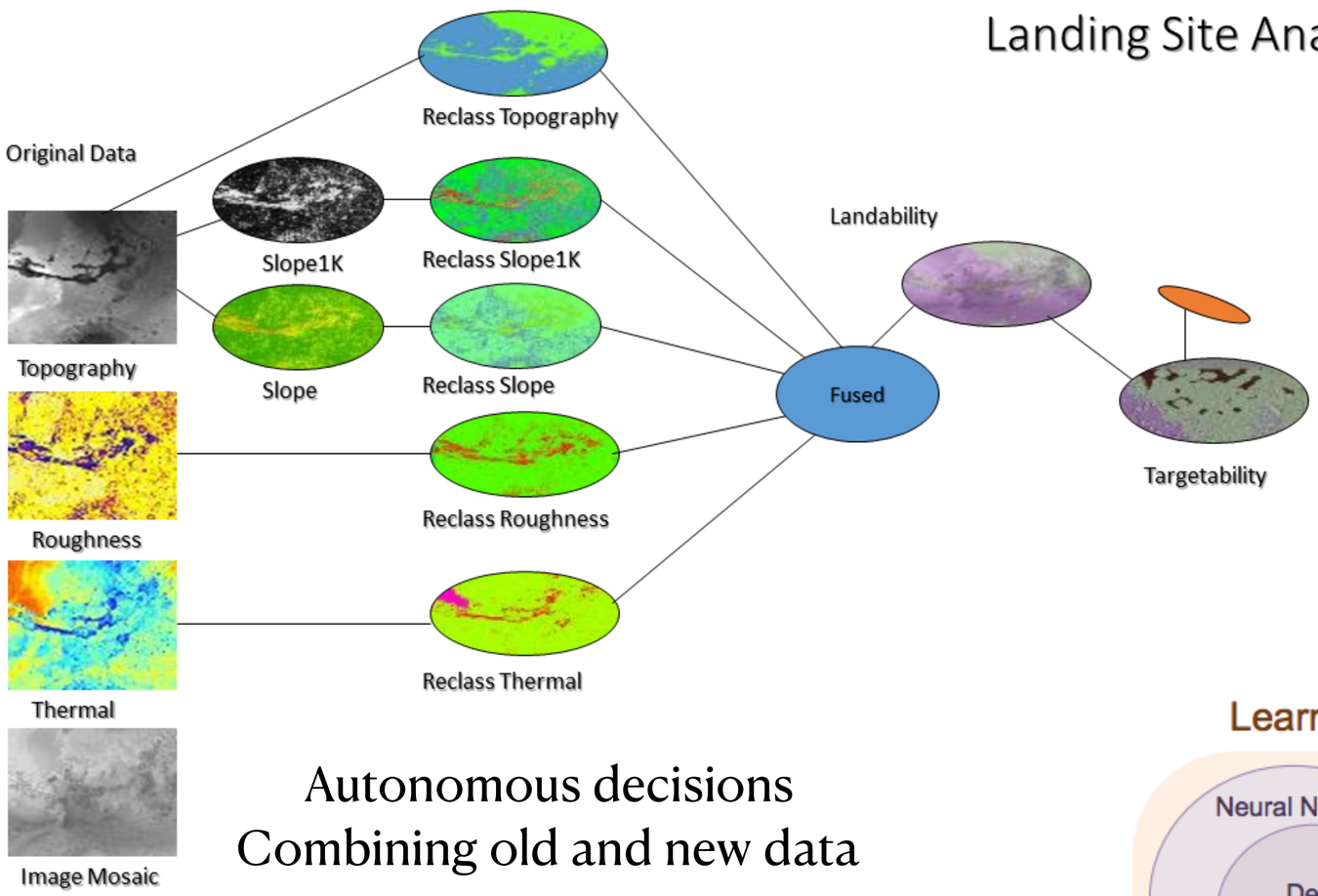
Cancer biomarkers and Big Data: A planetary Science approach

REF Crichton et al. 2020 https://doi.org/10.1016/j.ccell.2020.09.006

Ashish Mahabal

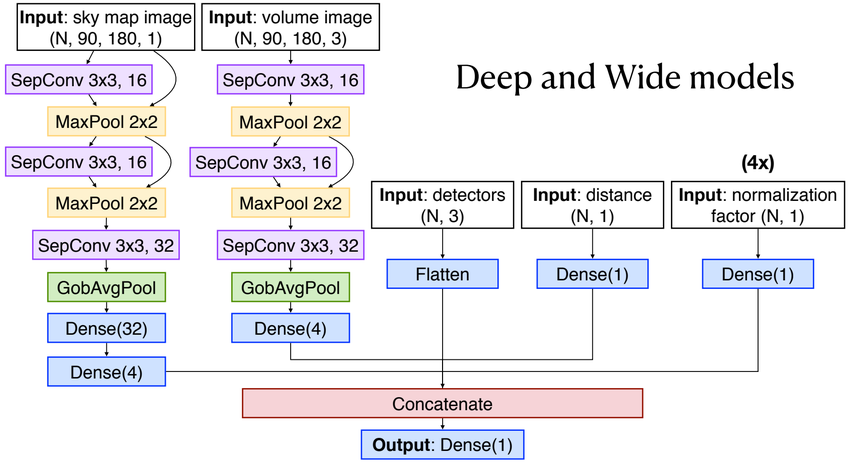
Interpretability, Explainability

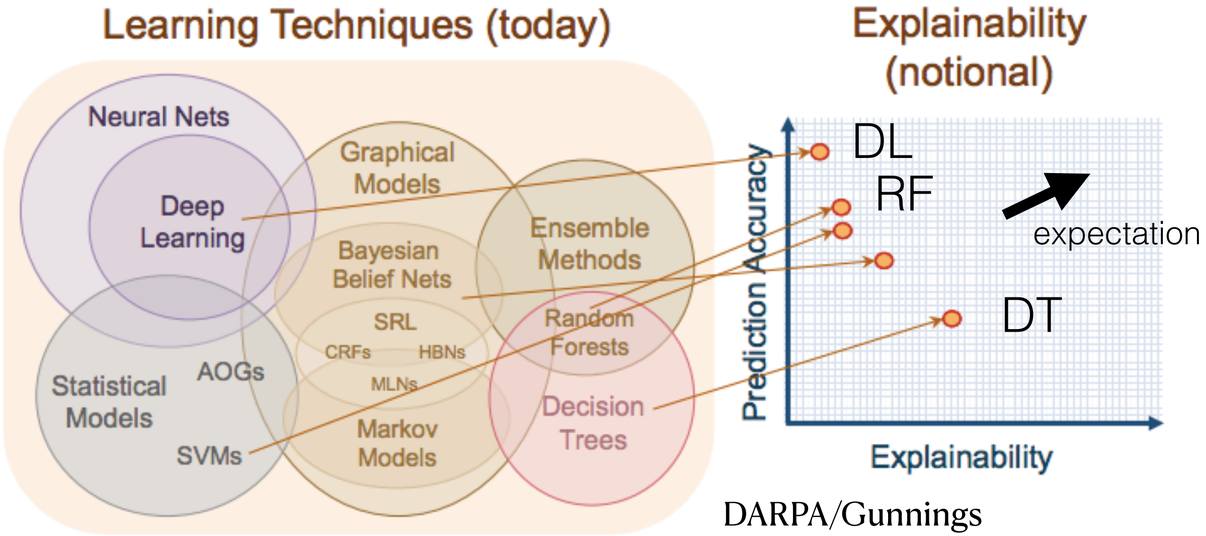
Landing Site Analysis



Autonomous decisions
Combining old and new data

Deep and Wide models





DARPA/Gunnings

From Data Fusion for In-space Nebulae
Mahabal, Hare, Fox, Hallinan 2020
REF: https://tinyurl.com/spacedatafusion
Based on a Keck Institute for Space Studies workshop

Ashish Mahabal

Summary

- We are working on several related projects in computational oncology
- Similarities exist between the nature of abstracted data and hence applicable techniques
- The diversity in data require several techniques to be brought together
- The current datasets are the tips of vast icebergs
- Interpretability and explainability being built in during methodology transfer

aam at astro.caltech.edu