Statistical Inference for Complex Data in the Physical Sciences

Ann B. Lee
Department of Statistics & Data Science
Carnegie Mellon University

Statistics --- Close Ties to Many Fields

Physical Sciences
Astronomy, Earth Sciences
Chemistry

Machine learning (ML)
Computer science
Engineering

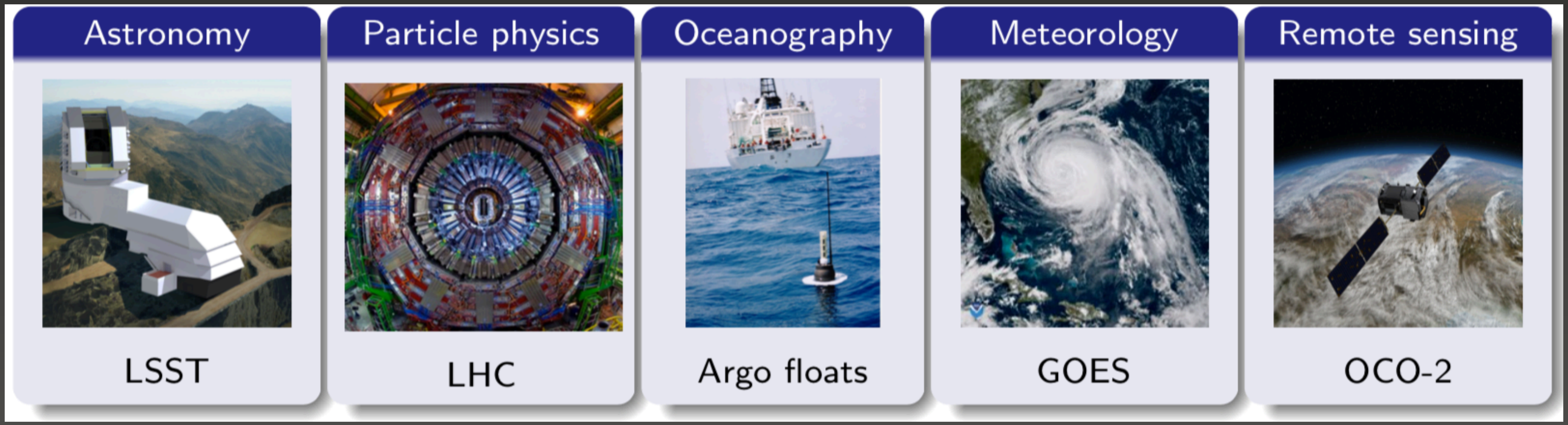
STATISTICS AT CMU

Biology
Genetics, Medicine
Neuroscience (CNBC)

Public policy (Heinz)
Social Sciences
Finance

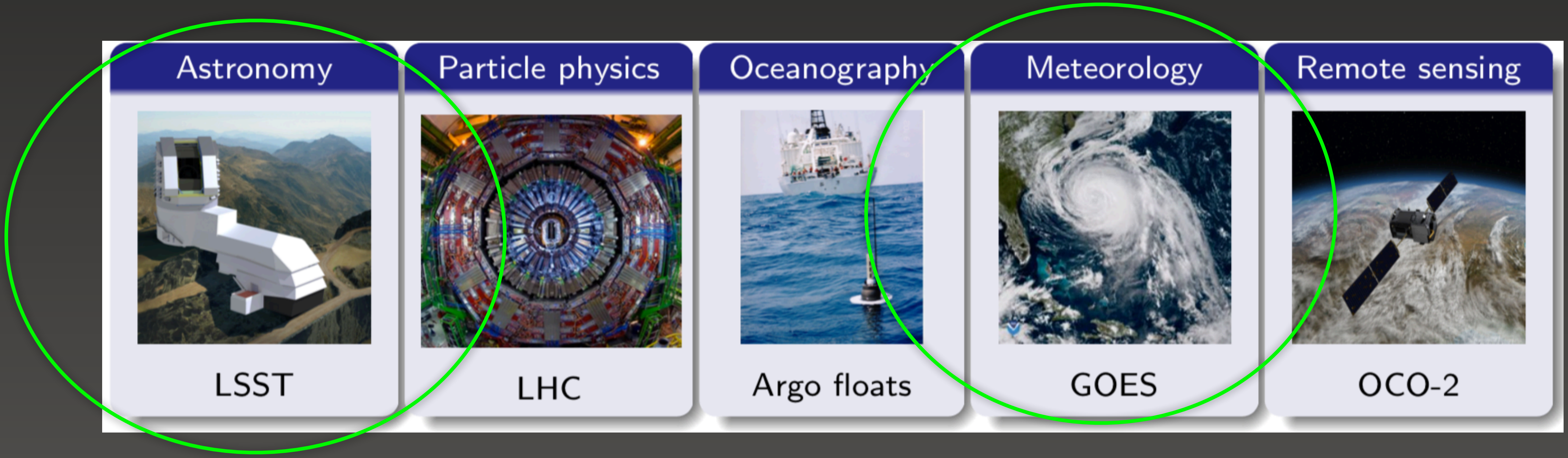
STAMPS@CMU

- In 2018, we started the STAtistical Methods for the Physical Sciences (STAMPS) research group.
- Problems in the physical sciences have similar statistical challenges —- involving massive data sets from different physical probes (e.g. images from satellites), large simulations, and complex measurement errors.

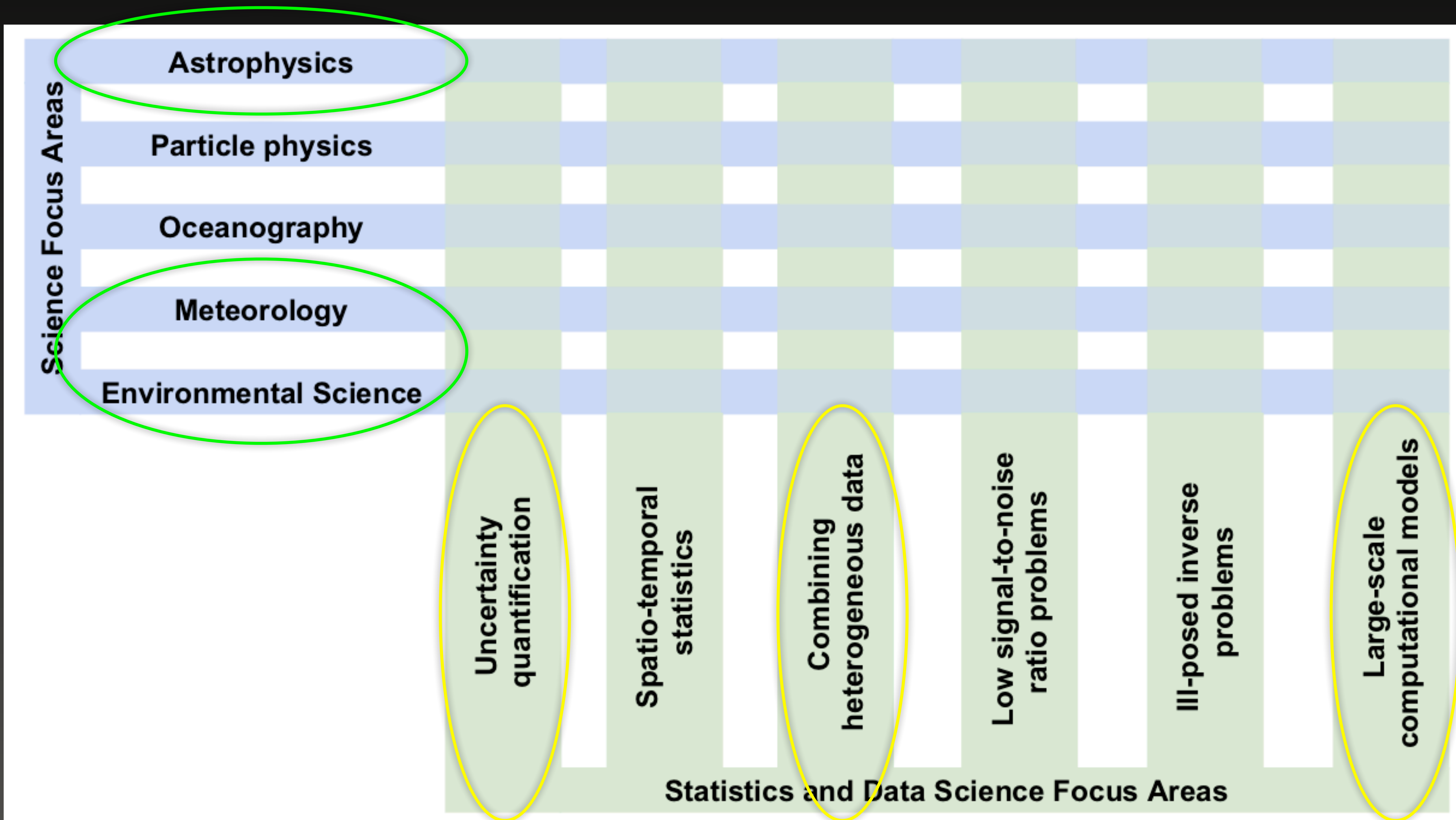


STAMPS@CMU

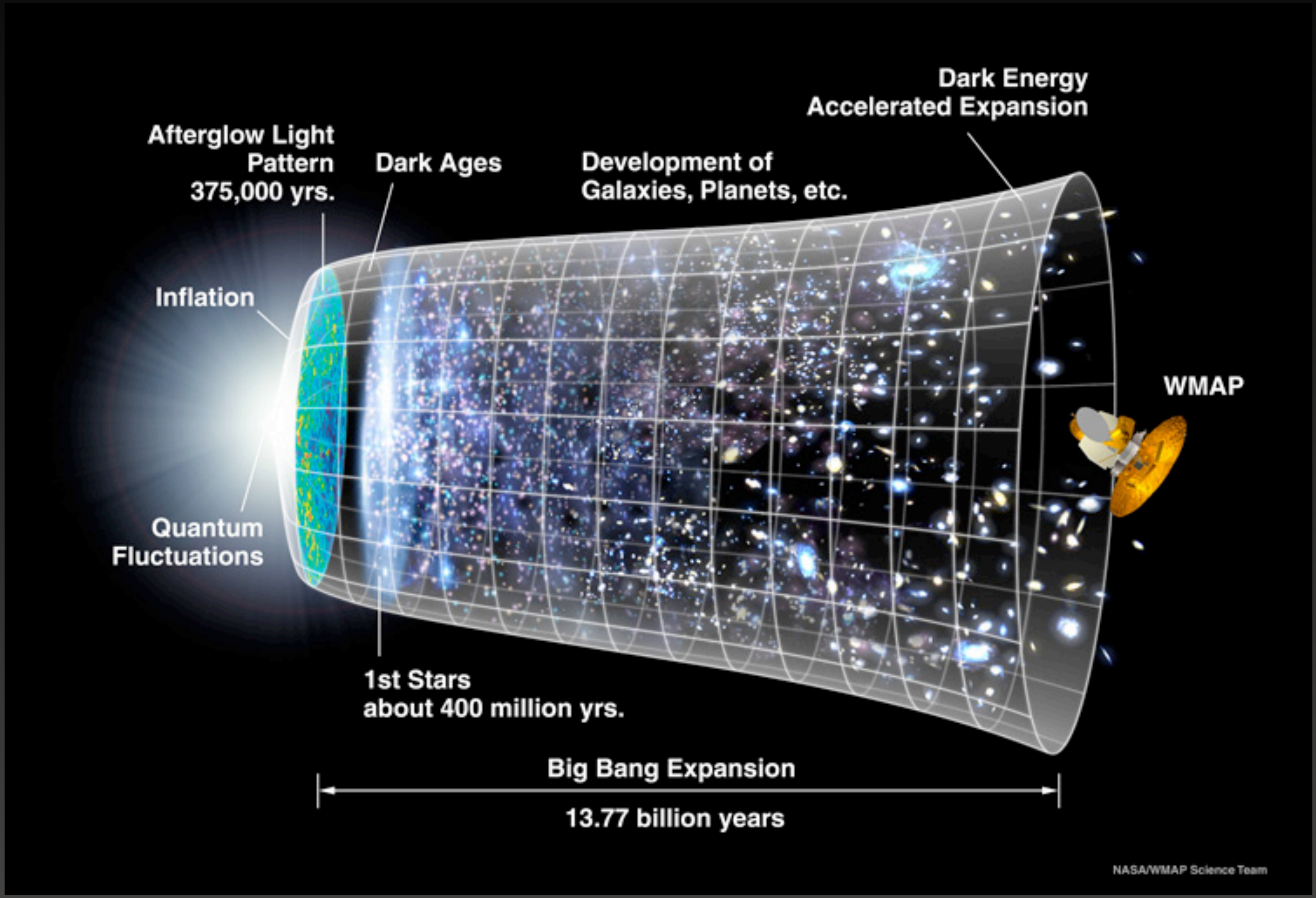
- In 2018, we started the STAtistical Methods for the Physical Sciences (STAMPS) research group.
- Problems in the physical sciences have similar statistical challenges —- involving massive data sets from different physical probes (e.g. images from satellites), large simulations, and complex measurement errors.



Commonalities in Physical Sciences: ``The Matrix'' behind STAMPS…



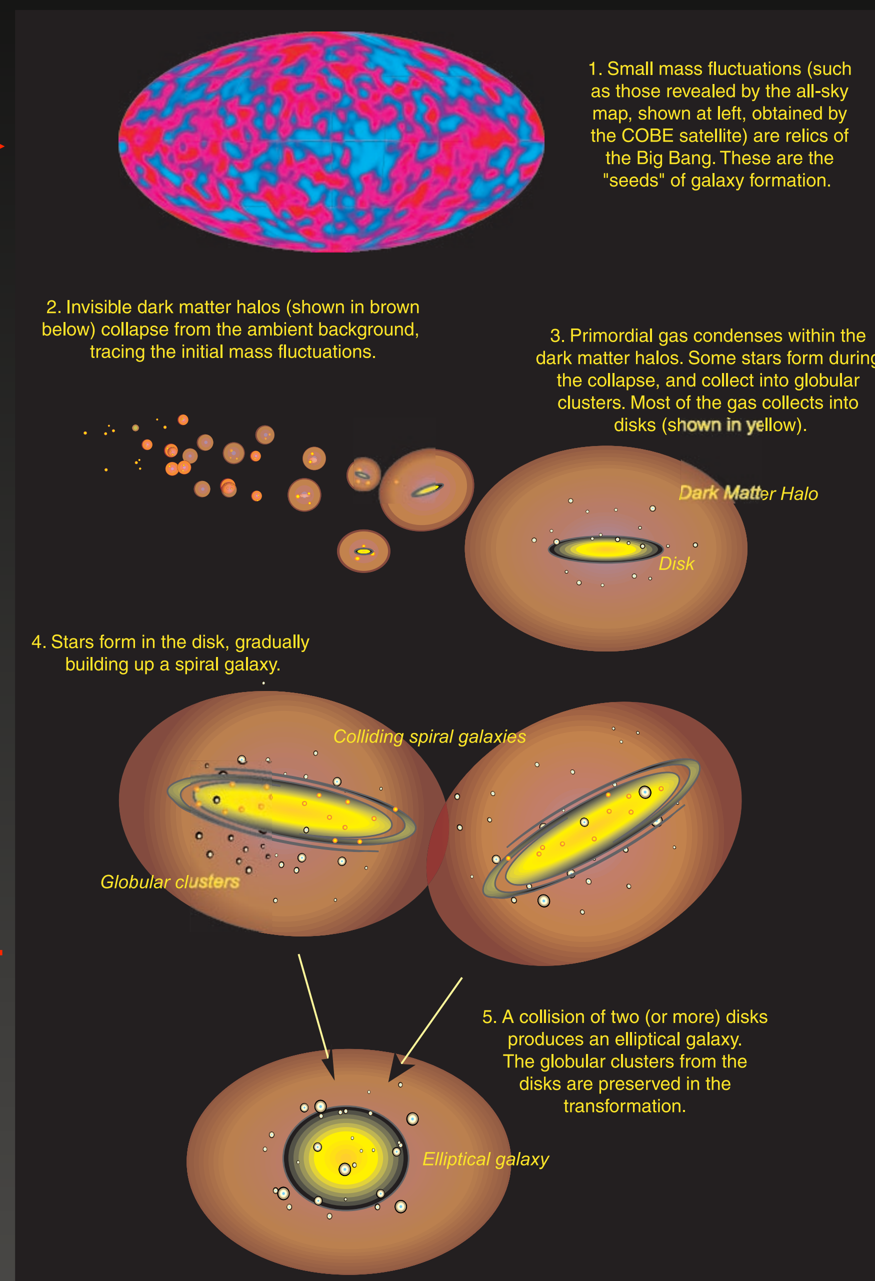
Astronomy/Cosmology Context



- Timeline of the Universe

Project 1 with Nic (4th Year): Likelihood-Free Inference and Validation of Complex Simulation Models

Theory In
(Simulation)



Abraham & van den Bergh (2001)

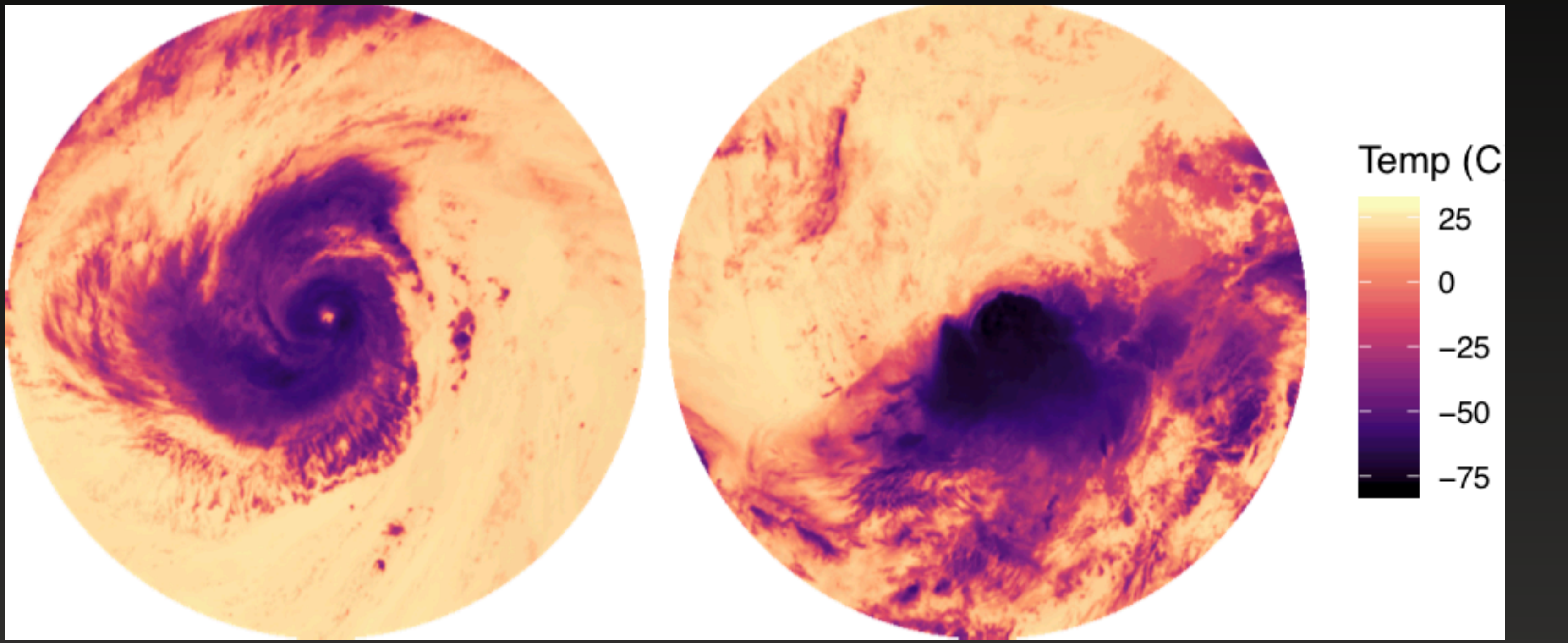
Ensemble of
Galaxies Out

How, Exactly?

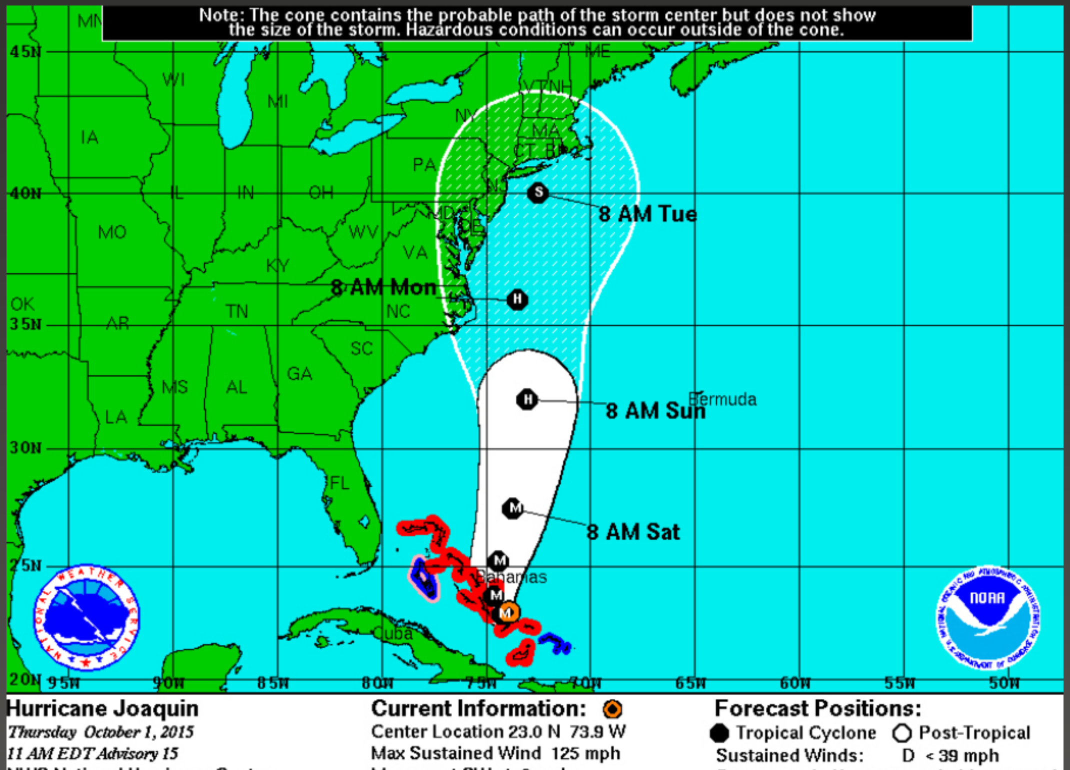
Observed Data

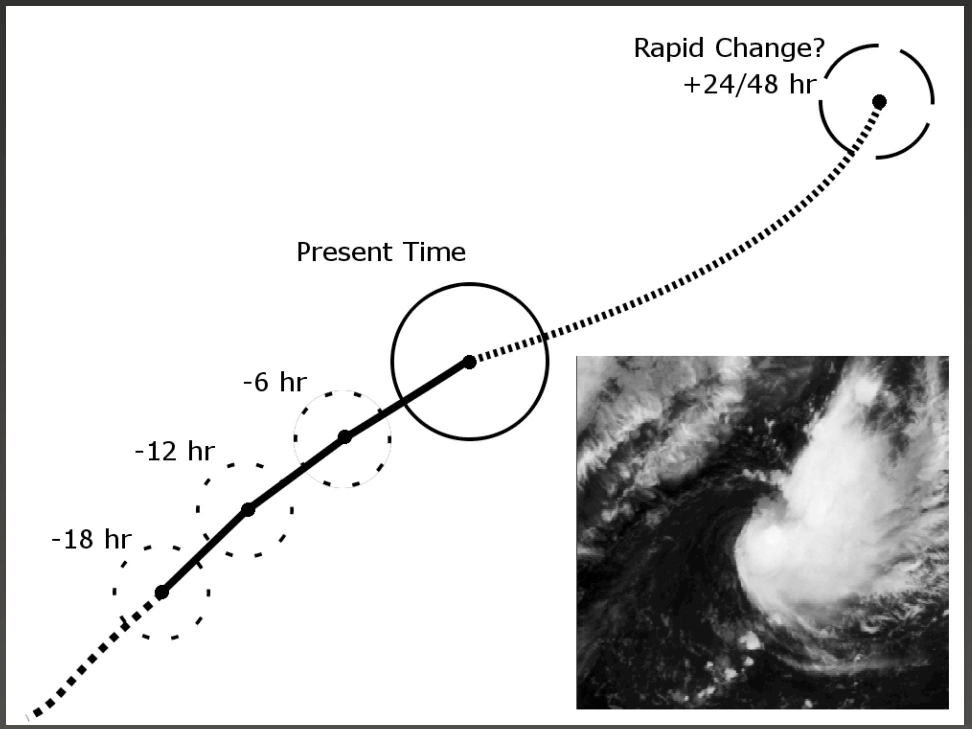
- Comparing distributions in high dimensions [with Ilmun Kim and Jing Lei]
- ``Likelihood-free inference'': Attach meaningful measures of uncertainty to estimates. (Statistical and computational efficiency.)
- Model validation. Assessing "emulators" and approximate likelihood models.

Project 2 with Trey (3rd Year): Modeling Hurricane Intensity Change Using GOES Satellite Imagery



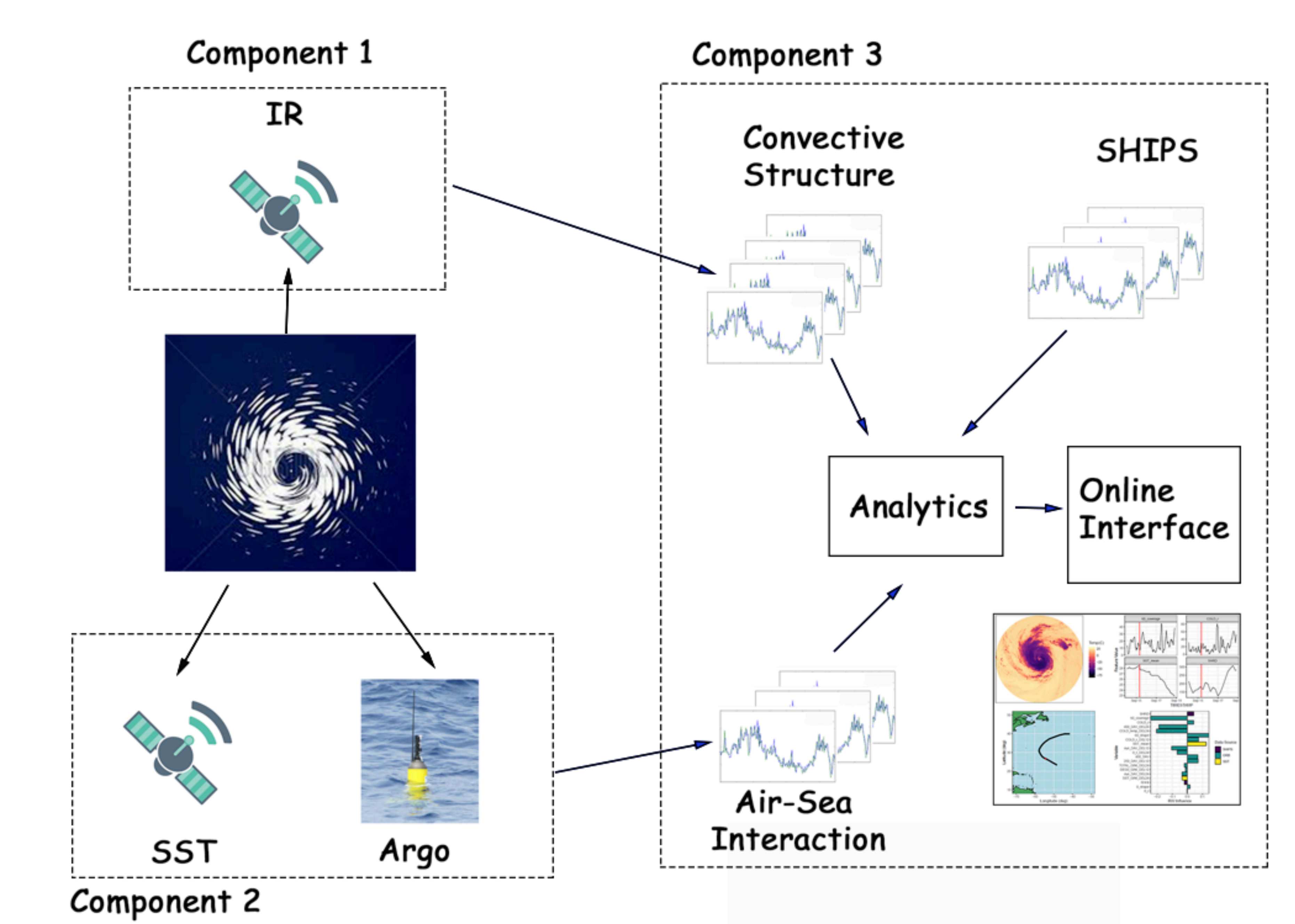
Hurricanes Edouard (2014; 109mph) and Nicole (2016; 54mph)





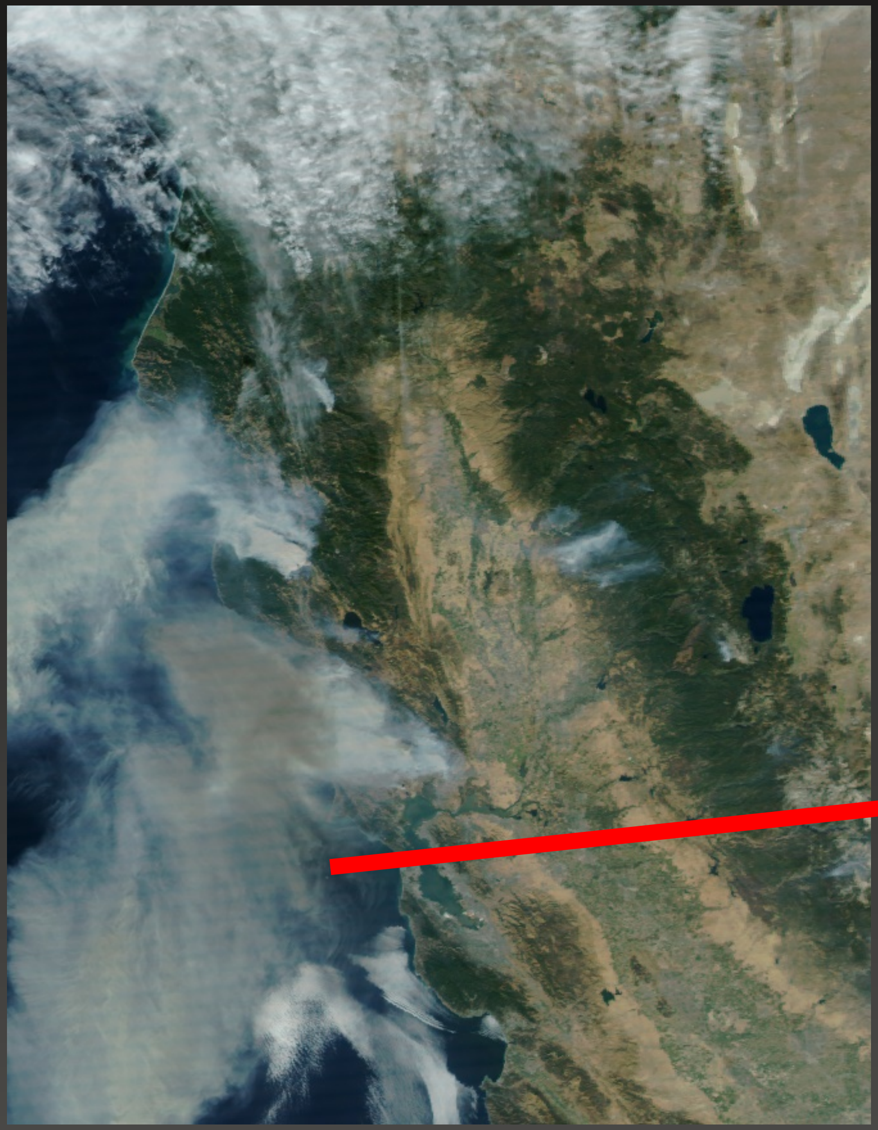
Project 3 with Addison (2nd Year ADA), Mikael and Trey:

Joint Analysis of TC Intensity Change by Integrating Satellite and In Situ Observations

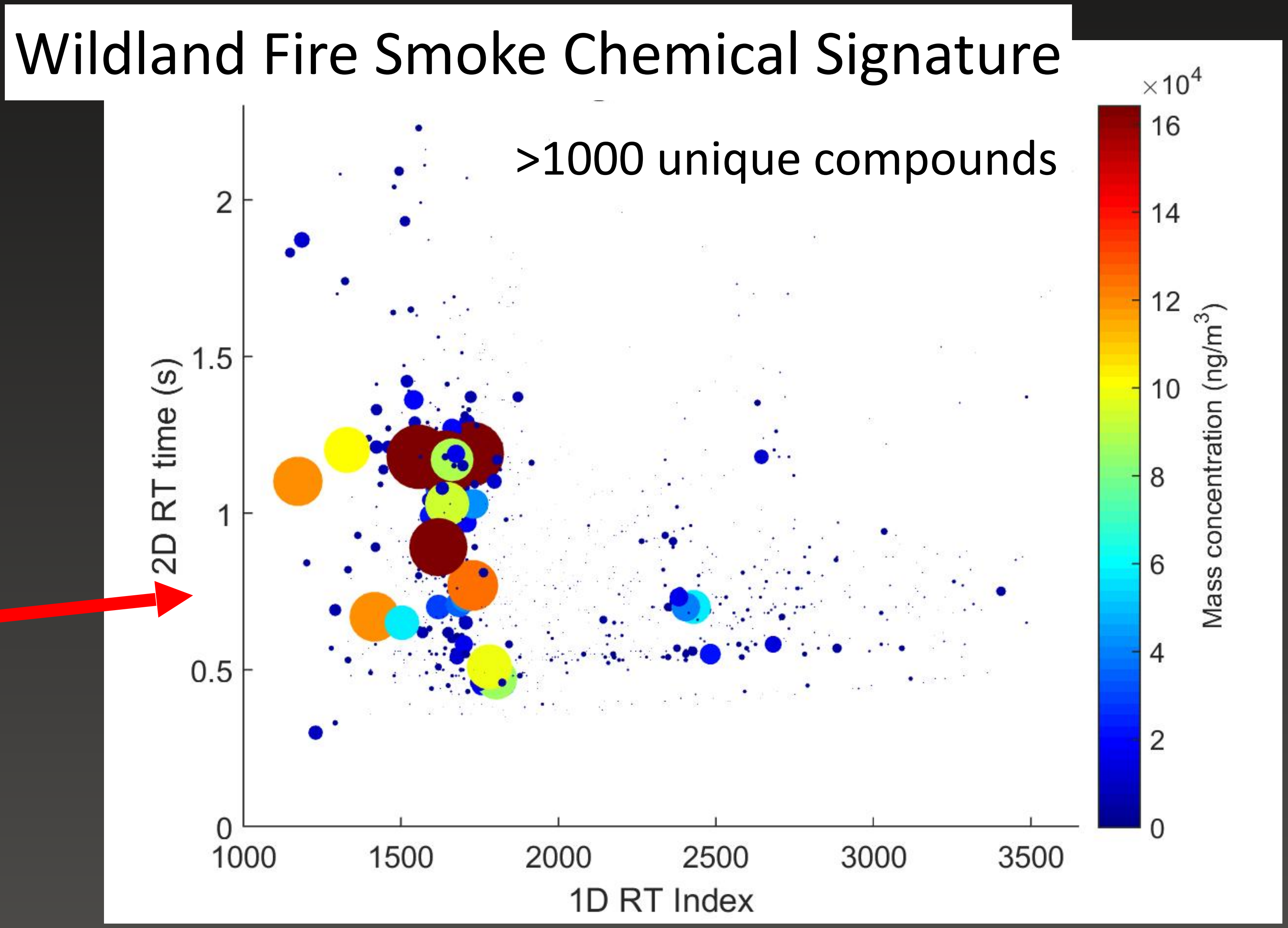


Project 4 with Lorenzo (1st Year ADA) and Coty: Chemical Fingerprinting of Wildfire Smoke

- Knowing what fuels burned during a wildfire is critical to predicting impact of wildfire smoke on air quality



Santa Rosa, CA Fires (Oct 2018), NASA



Key Points

- Tons of data and interesting science/methodology/algorithmic problems in the physical sciences.
- Look on the Arxiv for my recent papers.

Contact: annlee@cmu.edu

References

https://arxiv.org/abs/1911.11089

Unlocking GOES: A Statistical Framework for Quantifying the Evolution of Convective Structure in Tropical Cyclones

TREY MCNEELY* AND ANN B. LEE

Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania

KIMBERLY M. WOOD

Department of Geosciences, Mississippi State University, Mississippi State, Mississippi

DORIT HAMMERLING

Department of Applied Mathematics and Statistics, Colorado School of Mines, Golden, Colorado

ABSTRACT

Tropical cyclones (TCs) rank among the most costly natural disasters in the United States, and accurate forecasts of track and intensity are critical for emergency response. Intensity guidance has improved steadily but slowly, as processes which drive intensity change are not fully understood. Because most TCs develop far from land-based observing networks, geostationary (Geo) satellite imagery is critical to monitoring these storms. Modern high-resolution Geo observations provide an unprecedented scientific opportunity. These complex data are however challenging to analyze by forecasters in real time, whereas off-the-shelf machine learning algorithms have limited applicability due to their "black box" structure. This study presents analytic tools that quantify convective structure patterns in infrared Geo imagery for over-ocean TCs, yielding lower-dimensional but rich representations that support analysis and visualization of how these patterns evolve during a rapid intensity change. The proposed `ORB` feature suite targets the global `O`rganization, `R`adial structure, and `B`ulk morphology of TCs. Combined with a functional basis, the resulting representation of convective structure patterns on multiple scales serves as input to powerful but sometimes hard-to-interpret machine learning methods. This study uses the logistic lasso, a penalized generalized linear model, to relate predictors to rapid intensity change. Using `ORB` alone, binary classifiers identifying the presence (versus absence) of

https://arxiv.org/abs/1905.11505 (to appear in AISTATS'2020)

Validation of Approximate Likelihood and Emulator Models for Computationally Intensive Simulations

Niccolò Dalmasso,[1] Ann B. Lee,[1] Rafael Izbicki,[2] Taylor Pospisil,[3] Ilmun Kim,[1] Chieh-An Lin[4]

Electronic Journal of Statistics
Vol. 13 (2019) 5253–5305
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1648

Global and local two-sample tests via regression

Ilmun Kim, Ann B. Lee, and Jing Lei

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **471,** 3273–3282 (2017) doi:10.1093/mnras/stx1807
Advance Access publication 2017 July 18

Local two-sample testing: a new tool for analysing high-dimensional astronomical data

P. E. Freeman,[★] I. Kim and A. B. Lee
Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

https://arxiv.org/abs/2002.10399

Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting

Niccolò Dalmasso[1] **Rafael Izbicki**[2] **Ann B. Lee**[1]

Abstract

Parameter estimation, statistical tests and confidence sets are the cornerstones of classical statistics that allow scientists to make inferences about the underlying process that generated the observed data. A key question is whether one can still construct hypothesis tests and confidence sets with proper coverage and high power in a so-called likelihood-free inference (LFI) setting; that is, a setting where the likelihood is not explicitly known but one can forward-simulate observable data according to a stochastic model. In this paper, we present ACORE (Approximate Computation via Odds Ratio Estimation), a frequentist approach to LFI that first formulates the classical likelihood ratio test (LRT) as a parametrized classification problem, and then uses the equivalence

1. Introduction

Parameter estimation, statistical tests and confidence sets are the cornerstones of classical statistics that relate observed data to properties of the underlying statistical model. Most frequentist procedure with good statistical performance (e.g., high power) require explicit knowledge of a likelihood function. However, in many science and engineering applications, complex phenomena are modeled by forward simulators that *implicitly* define a likelihood function: For example, given input parameters θ, a statistical model of our environment, climate or universe may combine deterministic dynamics with random fluctuations to produce synthetic data $\mathbf{X}$. Simulation-based inference without an explicit likelihood is called *likelihood-free inference* (LFI).

The literature on LFI is vast. Traditional LFI methods, such as Approximate Bayesian Computation (ABC; Beaumont et al. 2002; Marin et al. 2012; Sisson et al. 2018), estimate posteriors by using simulations sufficiently close to

Statistical Tools and Software for CDE in Python and R

https://doi.org/10.1016/j.ascom.2019.100362

Conditional Density Estimation Tools in **Python** and **R**
with Applications to Photometric Redshifts and Likelihood-Free Cosmological Inference

Niccolò Dalmasso,[1] Taylor Pospisil,[2] Ann B. Lee,[1] Rafael Izbicki,[3] Peter E. Freeman,[1] and Alex I. Malz[4,5]

[1] *Department of Statistics & Data Science, Carnegie Mellon University, USA*
[2] *Google LLC, Mountain View, USA*
[3] *Department of Statistics, Federal University of São Carlos, Brazil*
[4] *German Centre of Cosmological Lensing, Ruhr-Universität Bochum, Germany*
[5] *Center for Cosmology and Particle Physics, New York University, USA*

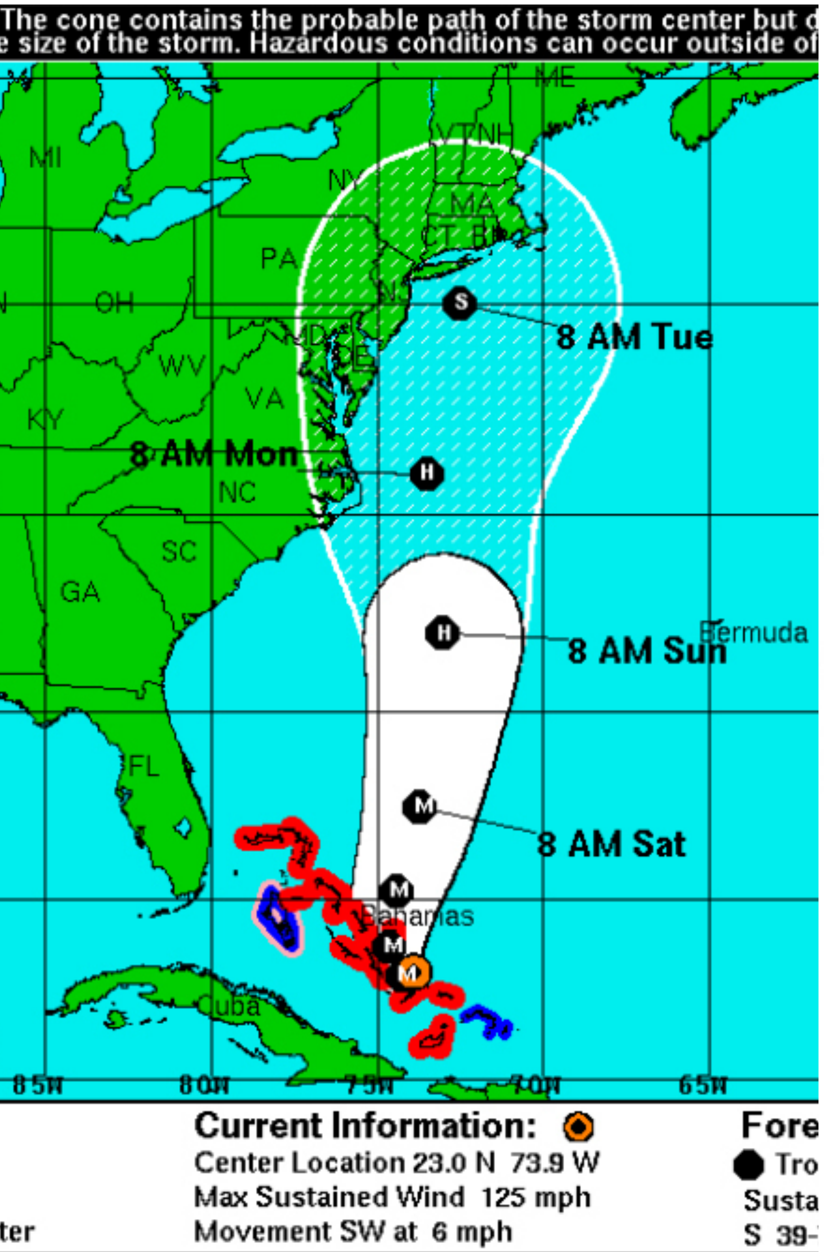
- Estimate $p(\theta|x)$, where $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^p$ ($d \leq 3$, p large)

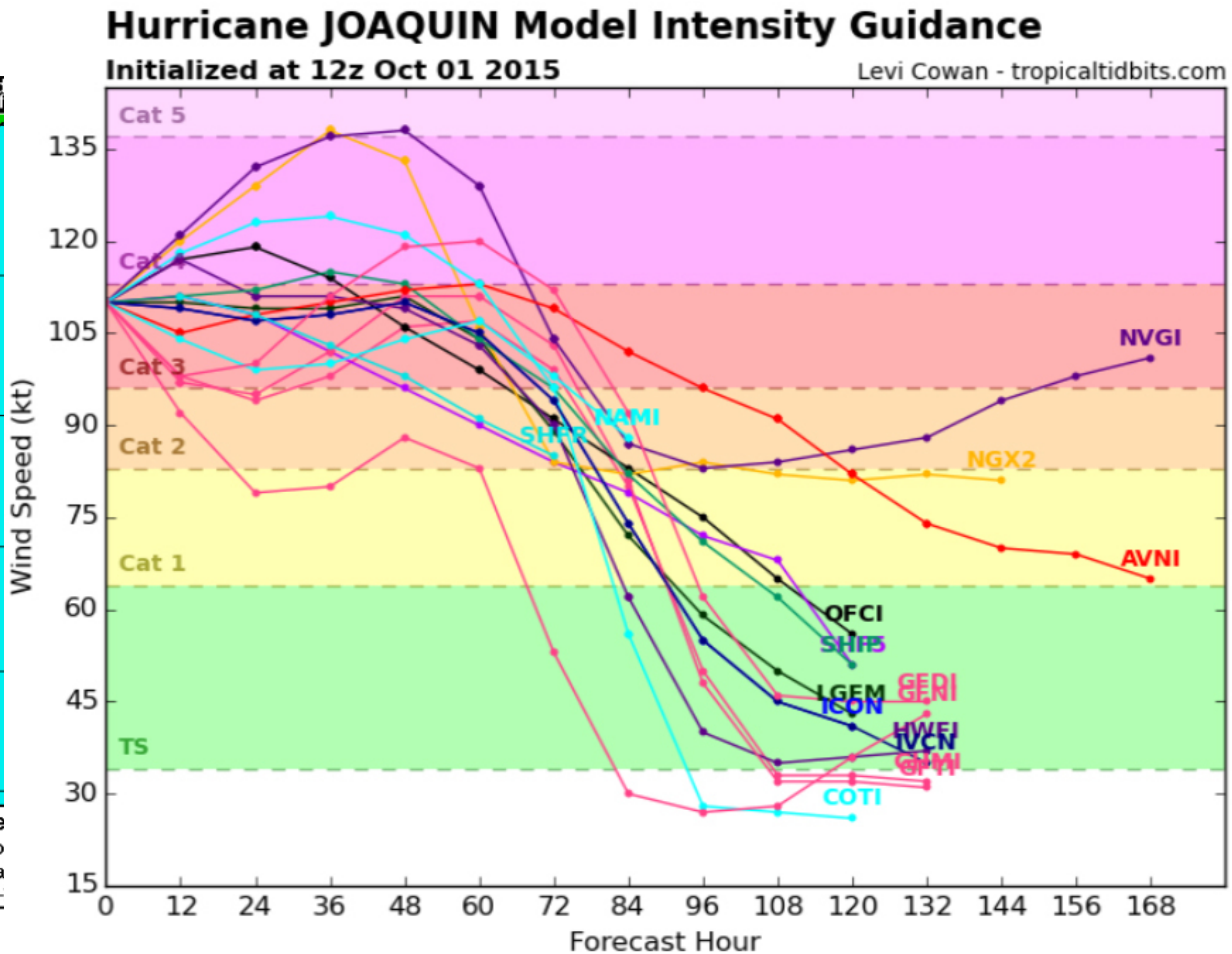
Table 1. Comparison of CDE methods in terms of training capacity and compatibility with multivariate response and different types of covariates. Capacities are roughly estimated based on input with around 100 features, and a standard i5/i7/quad-core processor with 16GB of RAM.

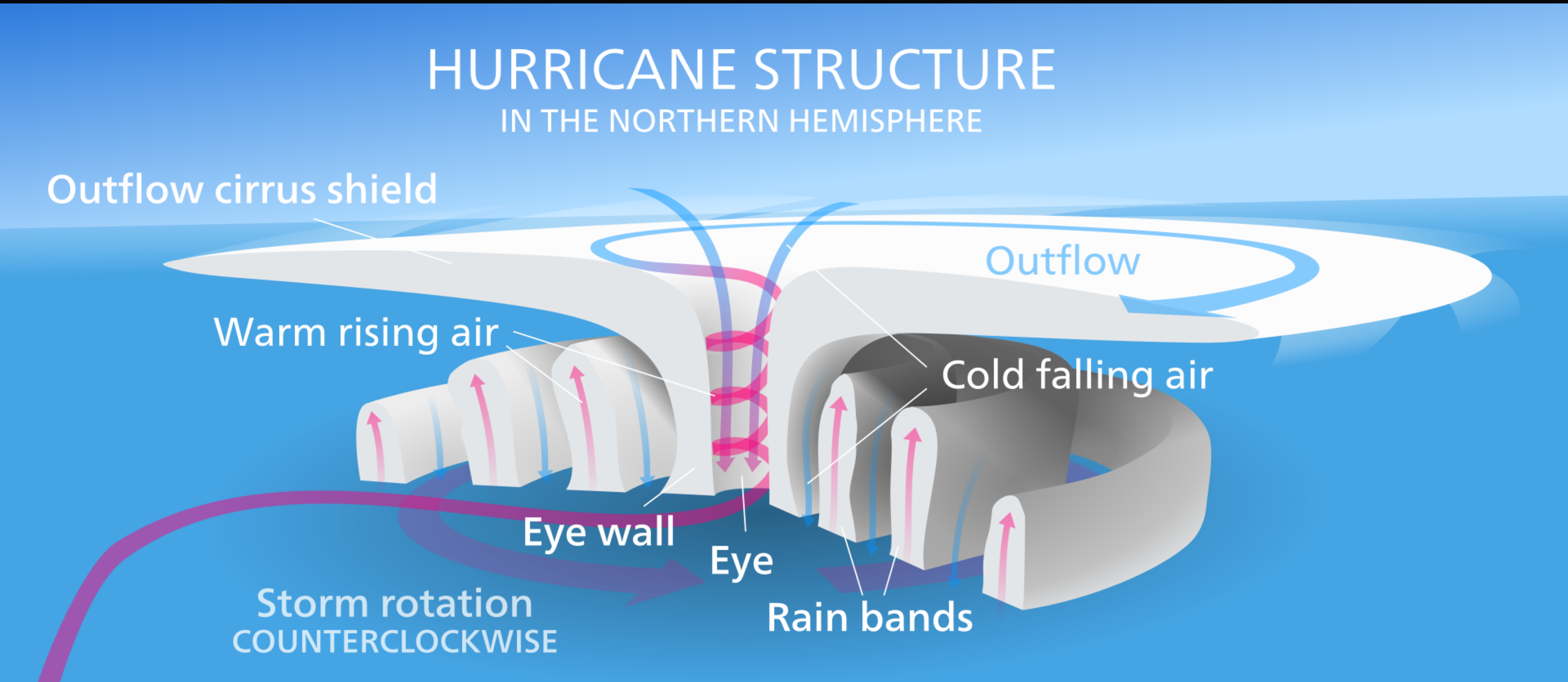
Method Complexity ↓	Method	Capacity (# Training Pts)	Multivariate Response	Functional Covariates	Image Covariates
	NNKCDE	Up to $\sim 10^5$	✓		
	(f)RFCDE	Up to $\sim 10^6$	✓	✓	
	FlexCode	Up to $\sim 10^6$		✓	
	DeepCDE	Up to $\sim 10^8$		✓	✓

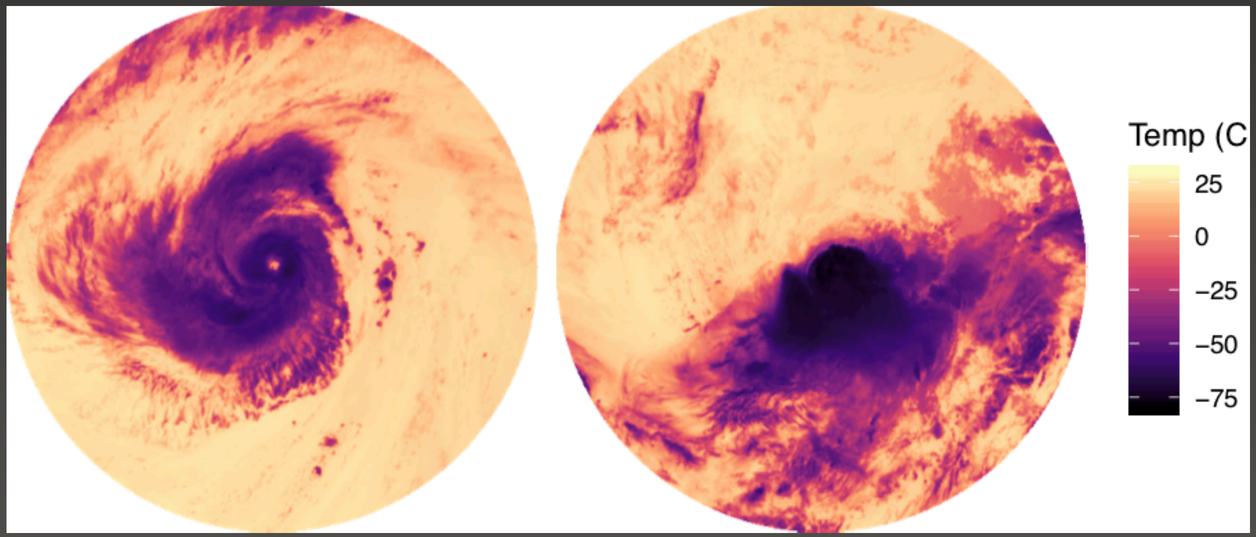
Extra Slides Start Here

TC intensity forecasts have fallen behind trajectory forecasts.

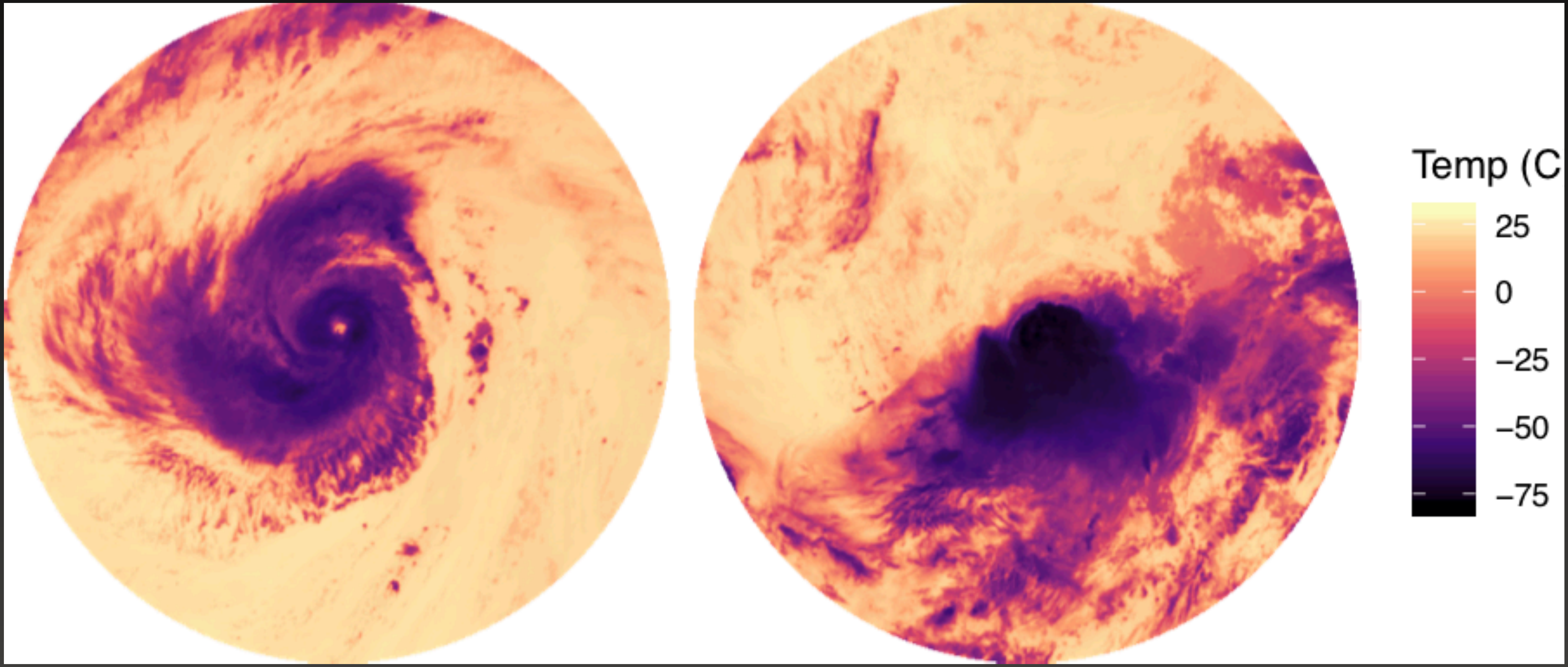





HURRICANE STRUCTURE
IN THE NORTHERN HEMISPHERE
Outflow cirrus shield
Outflow
Warm rising air
Cold falling air
Eye wall
Eye
Rain bands
Storm rotation
COUNTERCLOCKWISE


Temp (C
25
0
−25
−50
−75

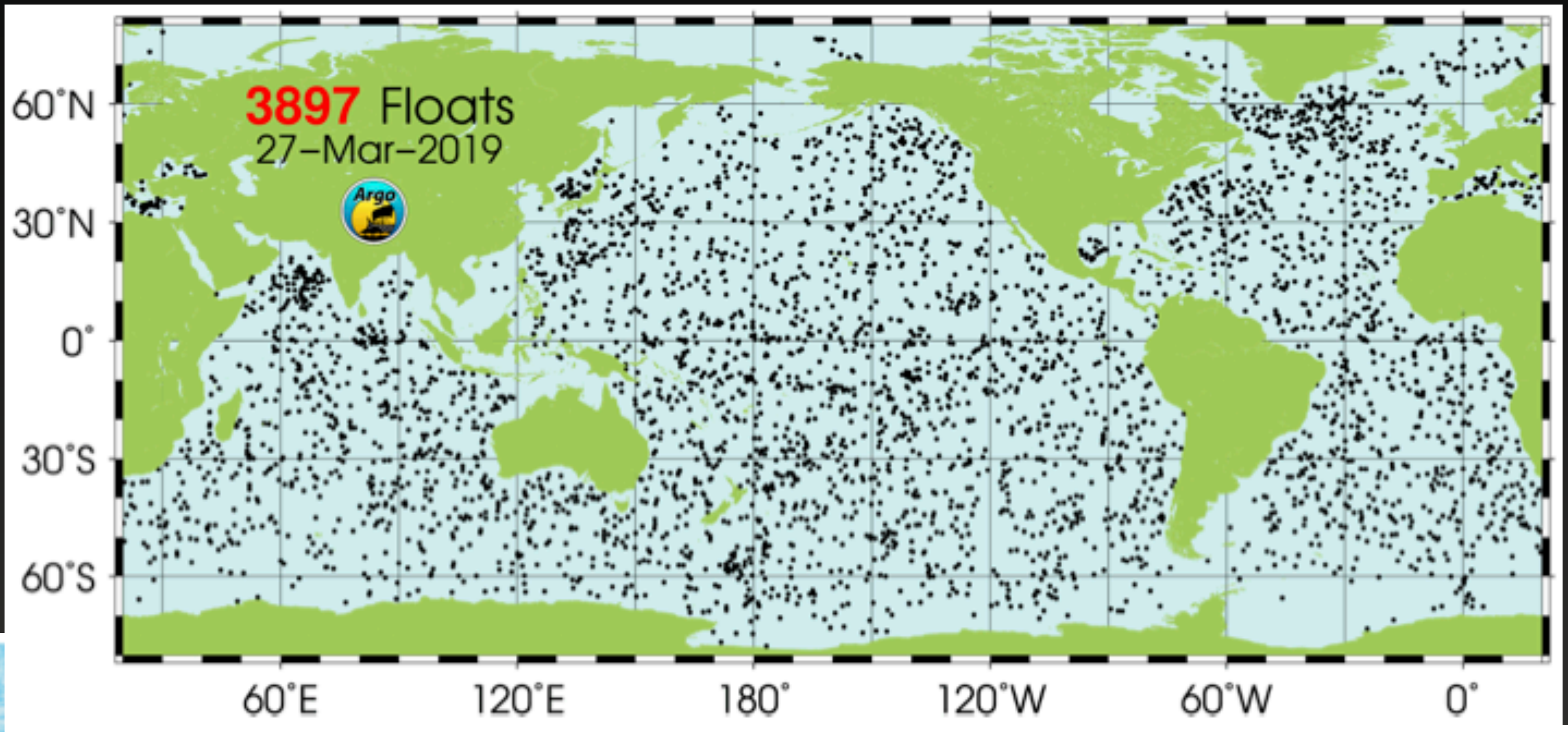
Leverage the axisymmetric structure of a mature, intense storm

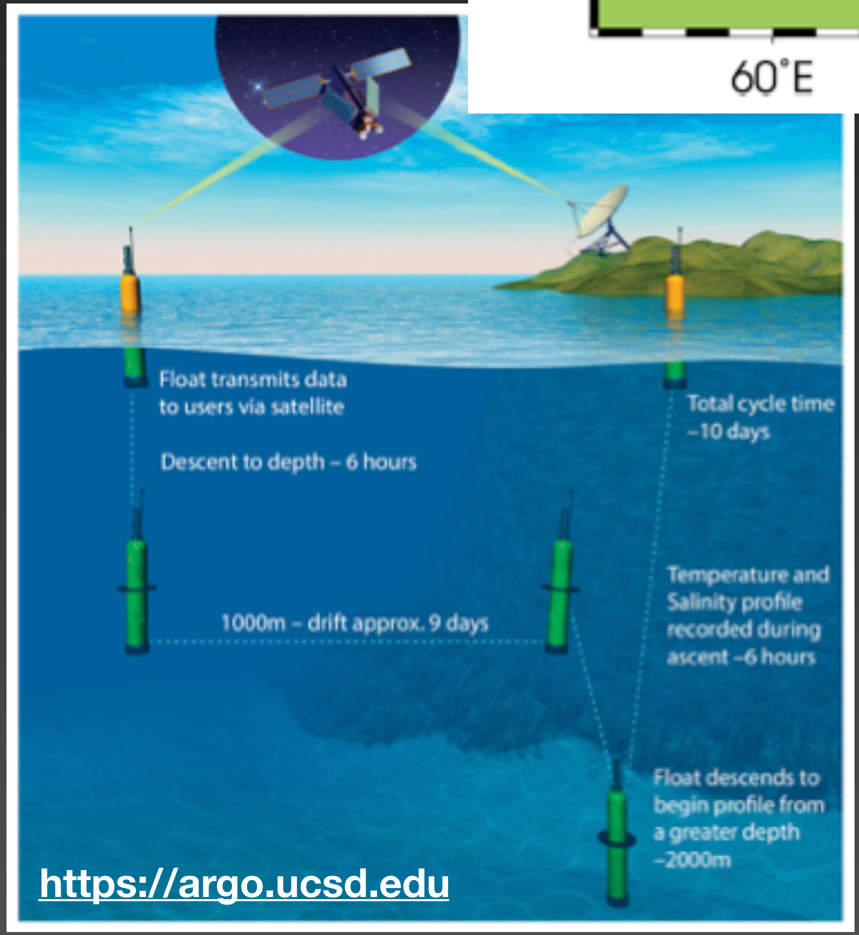


- Left: Hurricane Edouard (95 kt) at 18 UTC 16 Sept 2014
- Right: Hurricane Nicole (~47 kt) at 1 UTC 9 Oct 2016

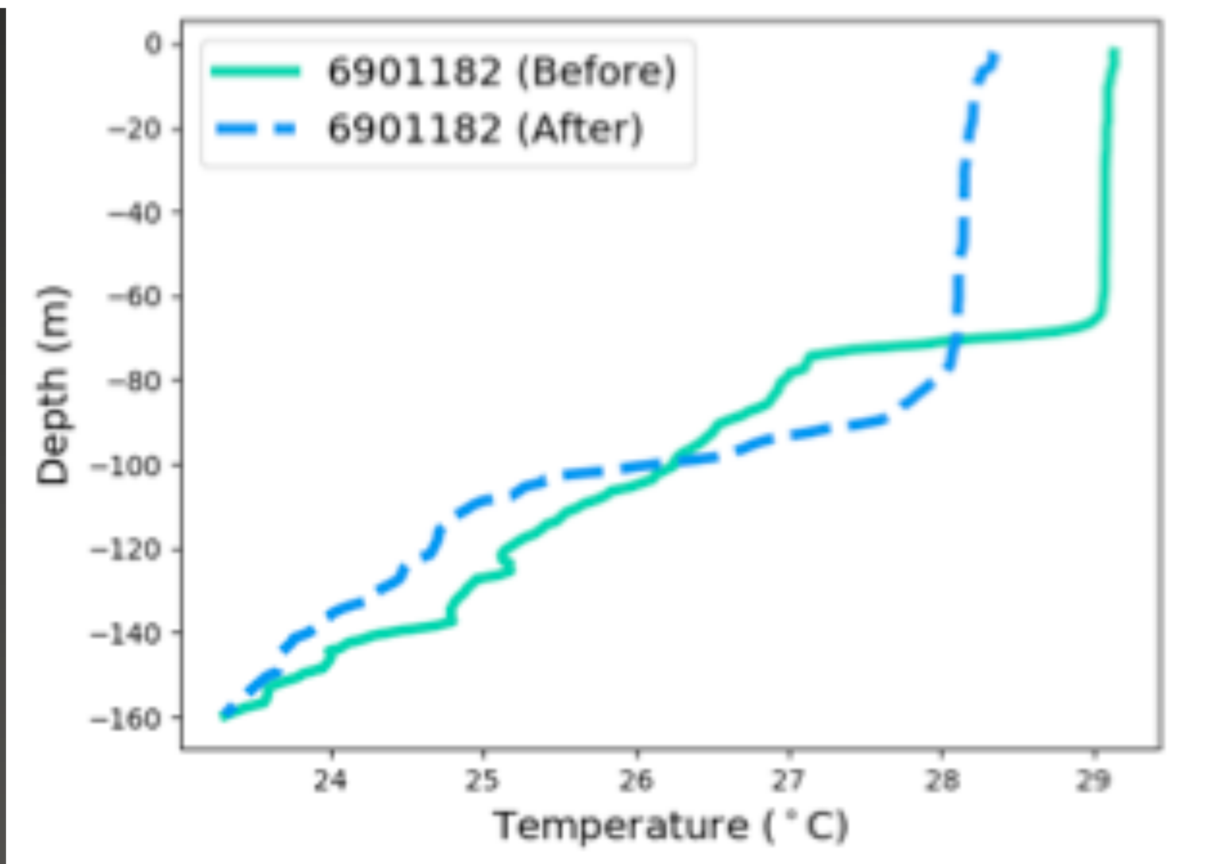
Hurricanes and Ocean Heat Content

[Hu/Kuusela/Lee/Giglio/Wood]





https://argo.ucsd.edu



Statistical Tools for Comparing and Analyzing Distributions of Images

[Freeman/Kim/Lee 2017, Kim/Lee/Lei 2018, Dalmasso et al 2019]

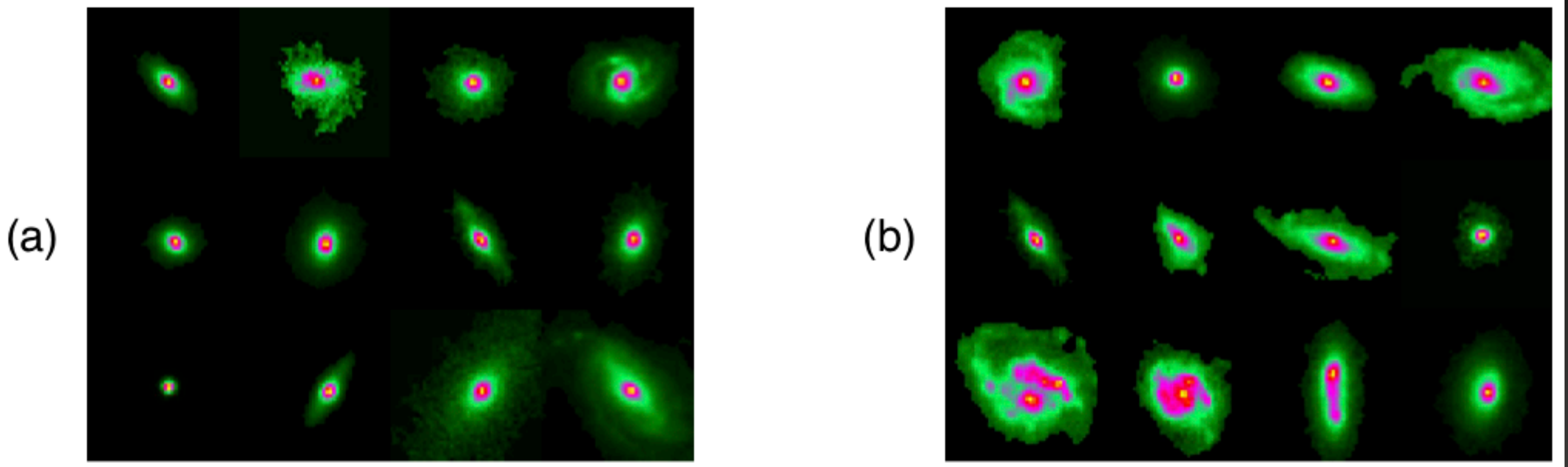


Figure 7: Examples of galaxies from (a) the low-SFR sample $\mathcal{S}_0$ versus (b) the high-SFR sample $\mathcal{S}_1$.

- Can we answer the question if, and if so, how two populations are different without just looking at histograms of just a few individual features?

Statistical Tools for Comparing and Analyzing Distributions of Images

[Freeman/Kim/Lee 2017, Kim/Lee/Lei 2018, Dalmasso et al 2019]

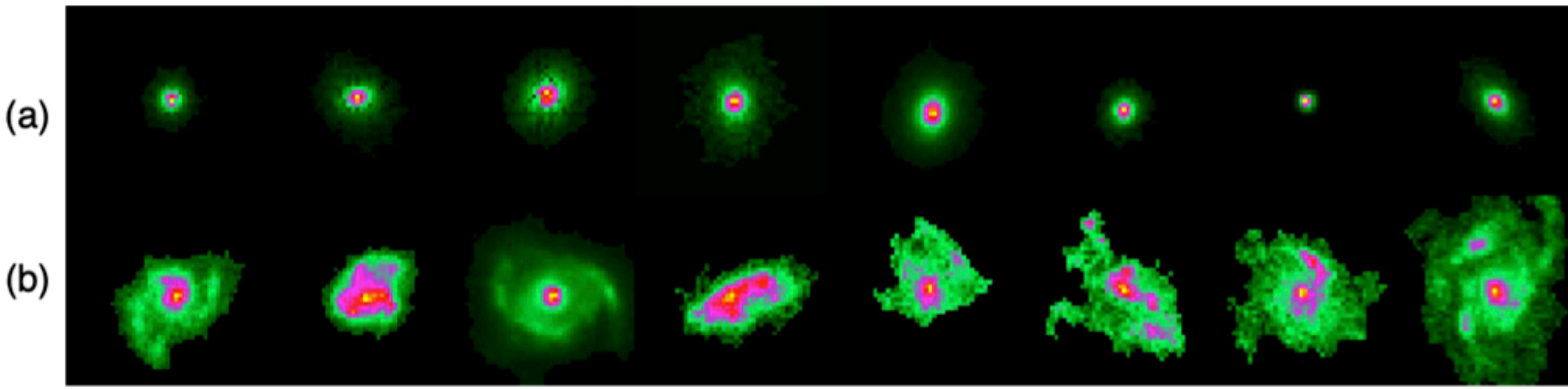


Figure 8: Galaxies in the test set with the highest significant difference $|\widehat{m}(\mathbf{x}) > \widehat{\pi}_1|$ according to our local test in feature space. (a) Galaxies that are more representative of the low-SFR sample $\mathcal{S}_0$, and (b) galaxies more representative of the high-SFR sample $\mathcal{S}_1$. The first group of galaxies presents undisturbed and concentrated morphologies, while the latter galaxies appear more extended. This is in line with what is expected by astronomers when comparing actual low-SFR and high-SFR galaxies.

- We have developed methods that — in an automated way — can identify differences that are statistically significant (that is, unlikely to occur by chance).

Visualizing the Results

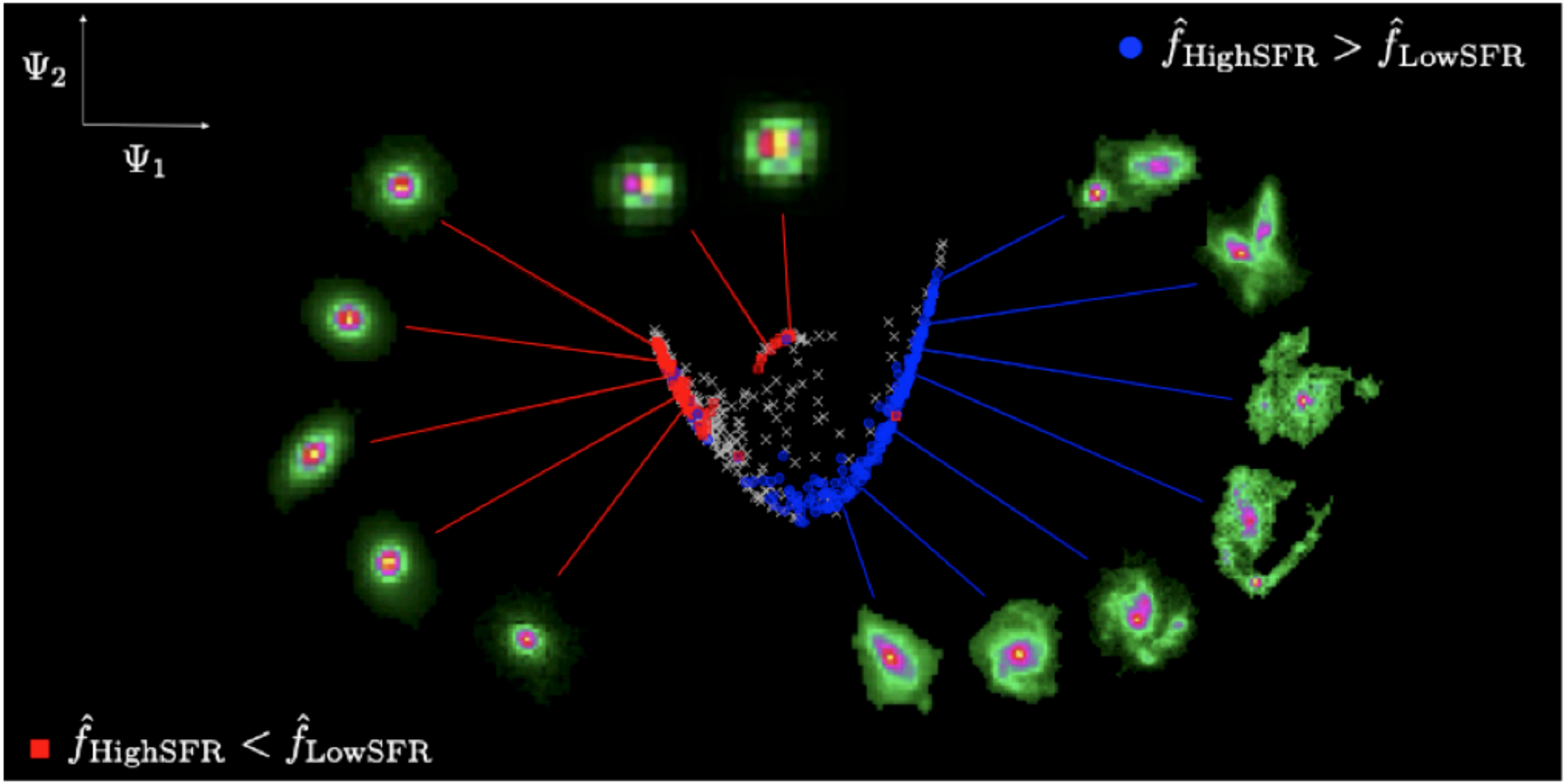
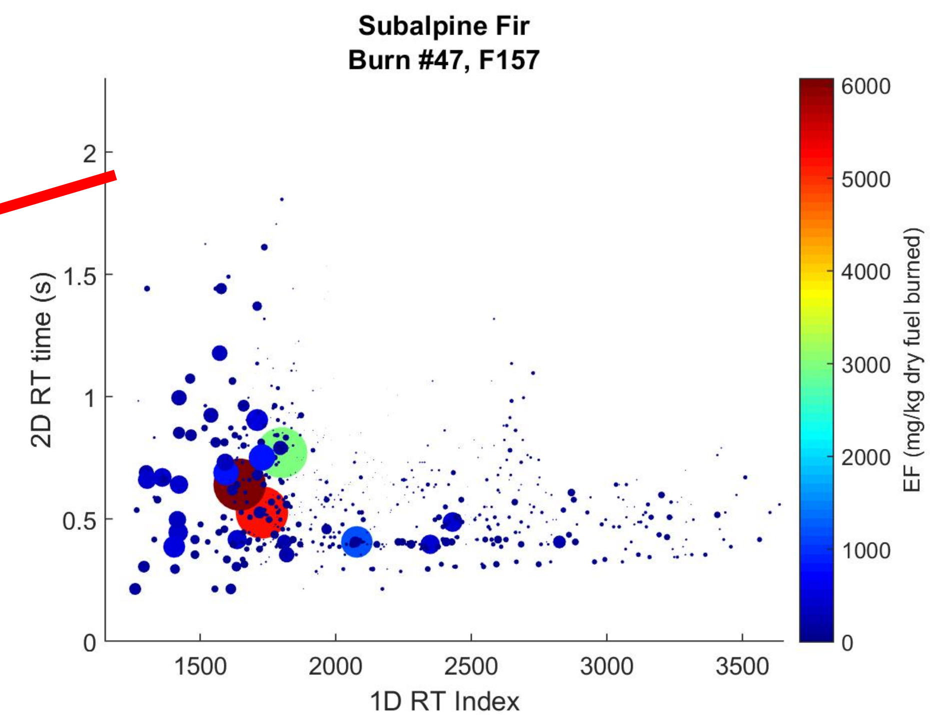
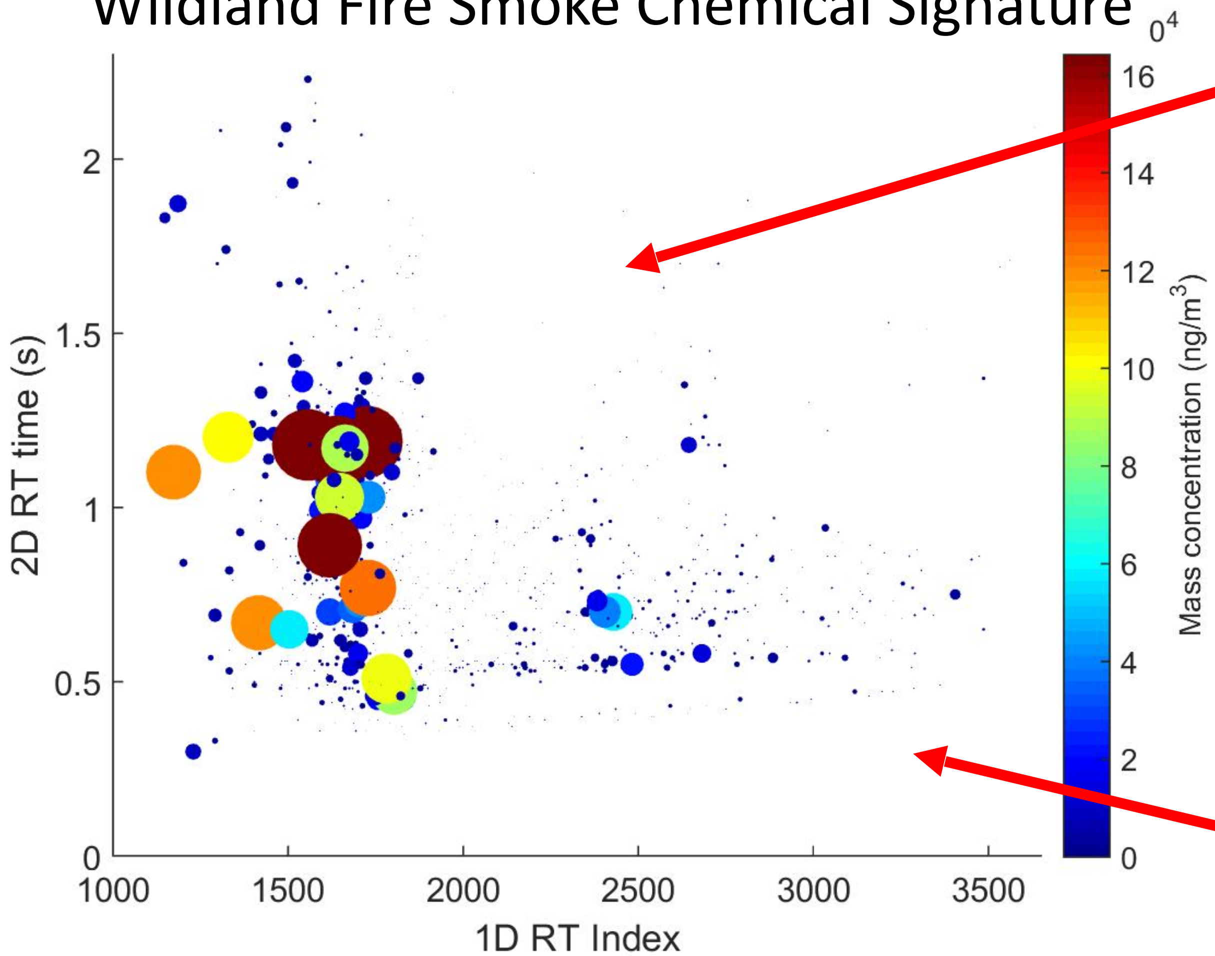


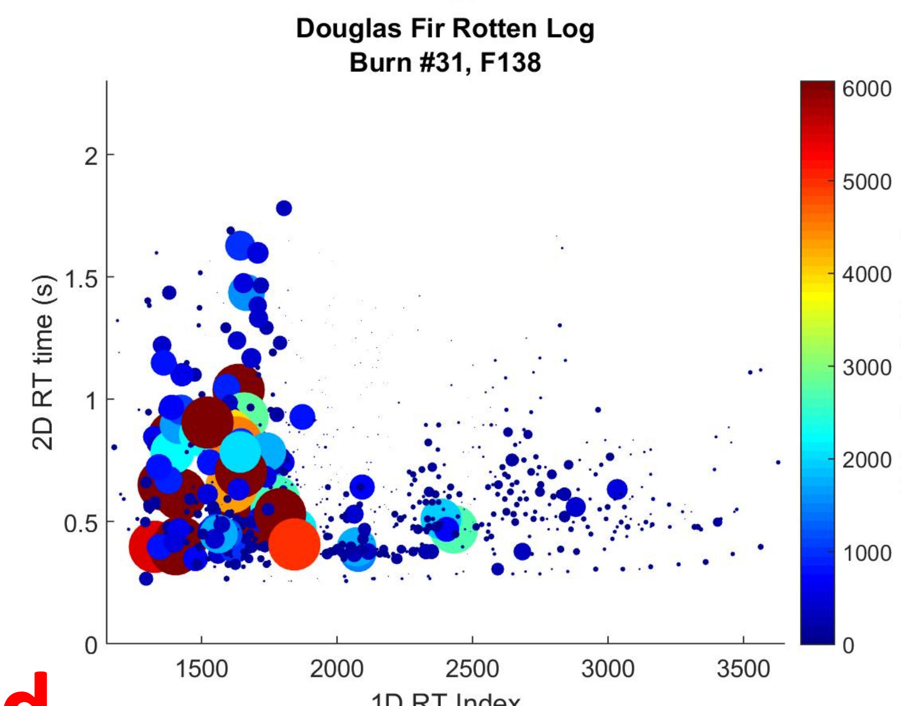
Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from [49].

Individual Fuel Burns



Wildland Fire Smoke Chemical Signature





Goals:

1) Identify the types of fuels that burned
2) Ballpark the amount of fuel that burned