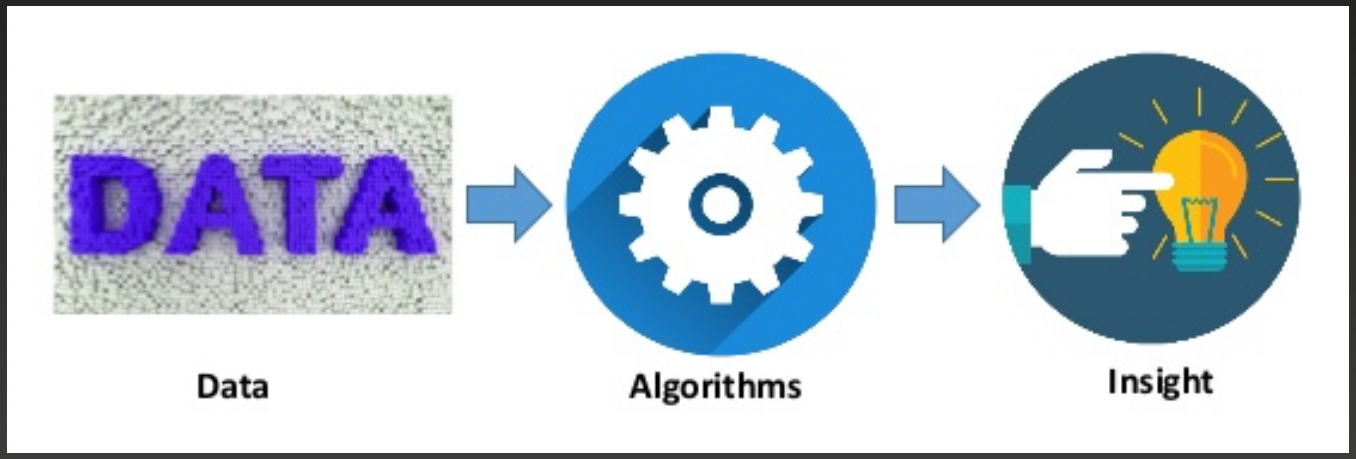
Statistics and Machine Learning for the Physical Sciences

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

What Does a Statistician Do?

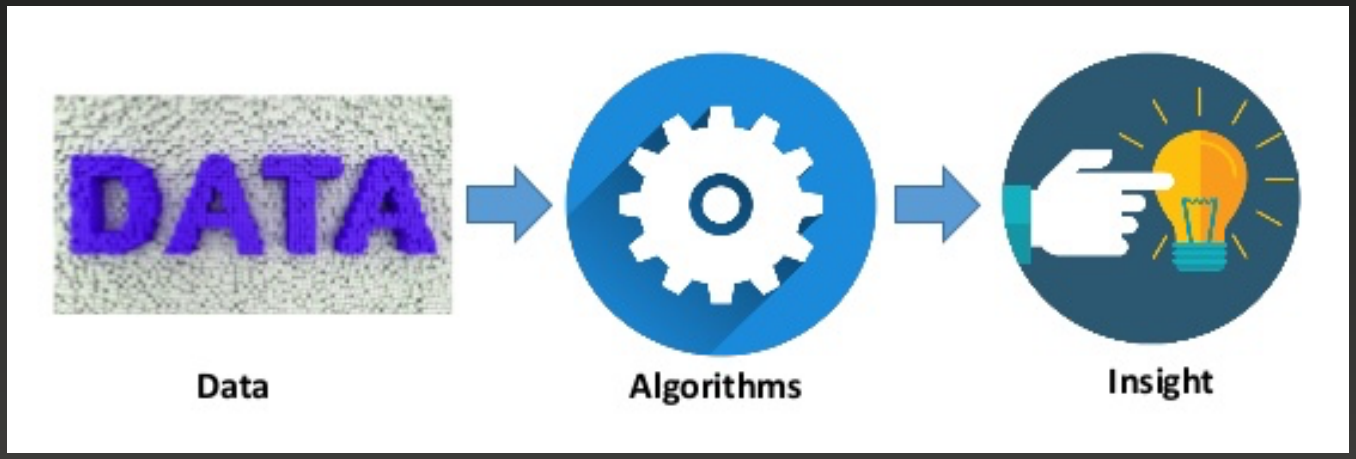
What Does a Statistician Do?

- Short answer: We try to make sense of the world by analyzing data.
- The fundamental problem in statistics:



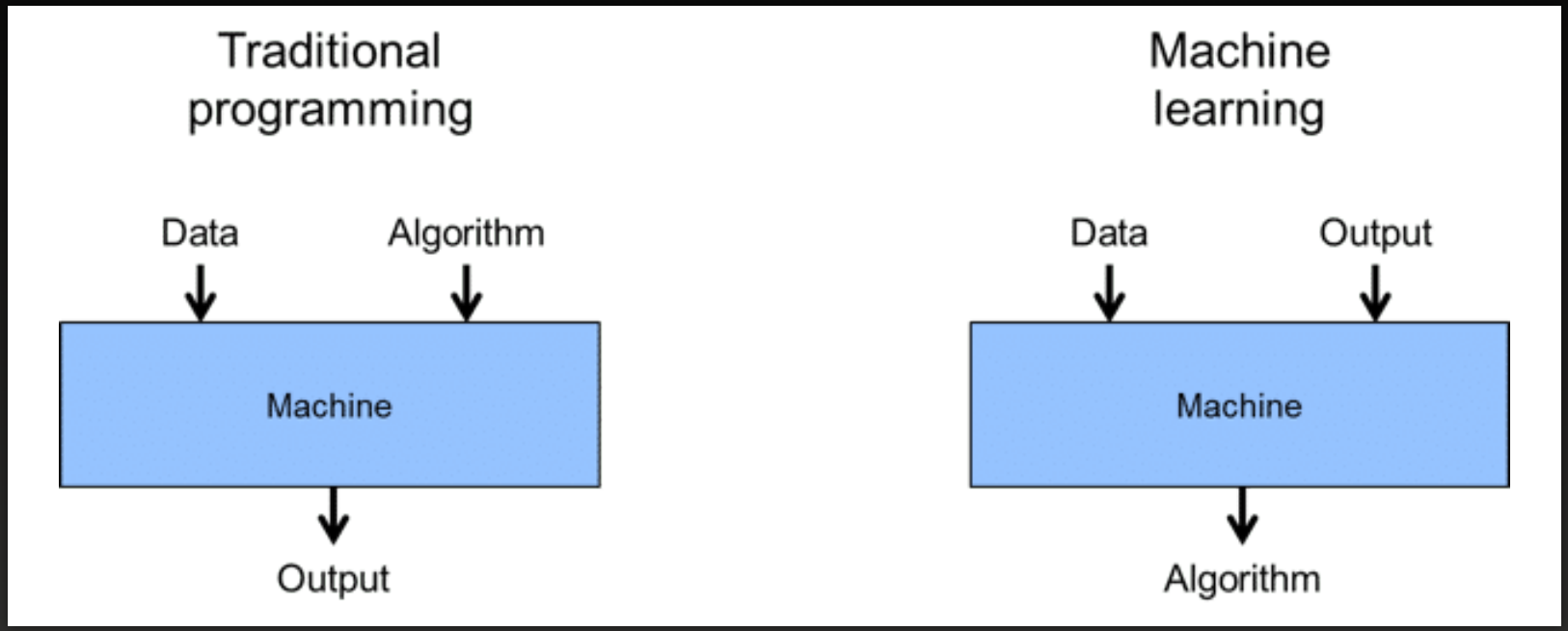
What Does a Statistician Do?

- Short answer: We try to make sense of the world by analyzing data.
- The fundamental problem in statistics:



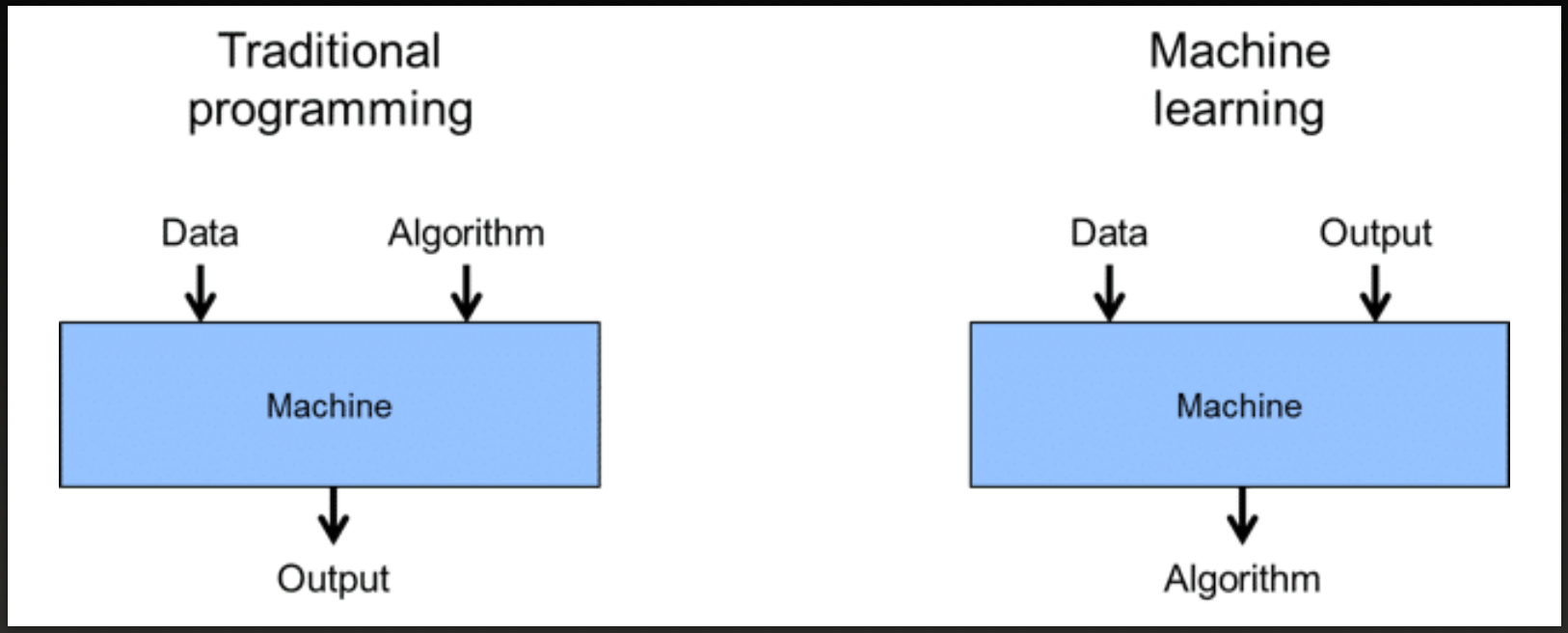
- Machine Learning = science of using algorithms to learn from and make predictions from data

Machine Learning & Statistics



https://dzone.com/articles/what-everyone-should-know-about-machine-learning

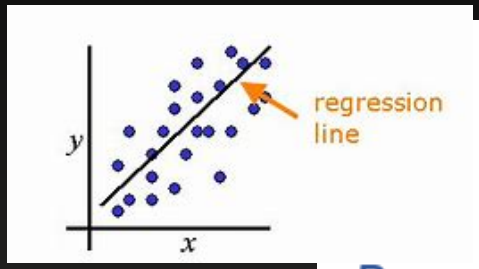
Machine Learning & Statistics



https://dzone.com/articles/what-everyone-should-know-about-machine-learning

- In Statistics/Data Science:
  - Heavy focus on developing insight regarding the physical system that generated the data.
  - Work a lot with probabilistic models. Quantify the uncertainty in our predictions.

There Exists a Whole Range of Different Learning Algorithms for Different Tasks and of Different Complexity



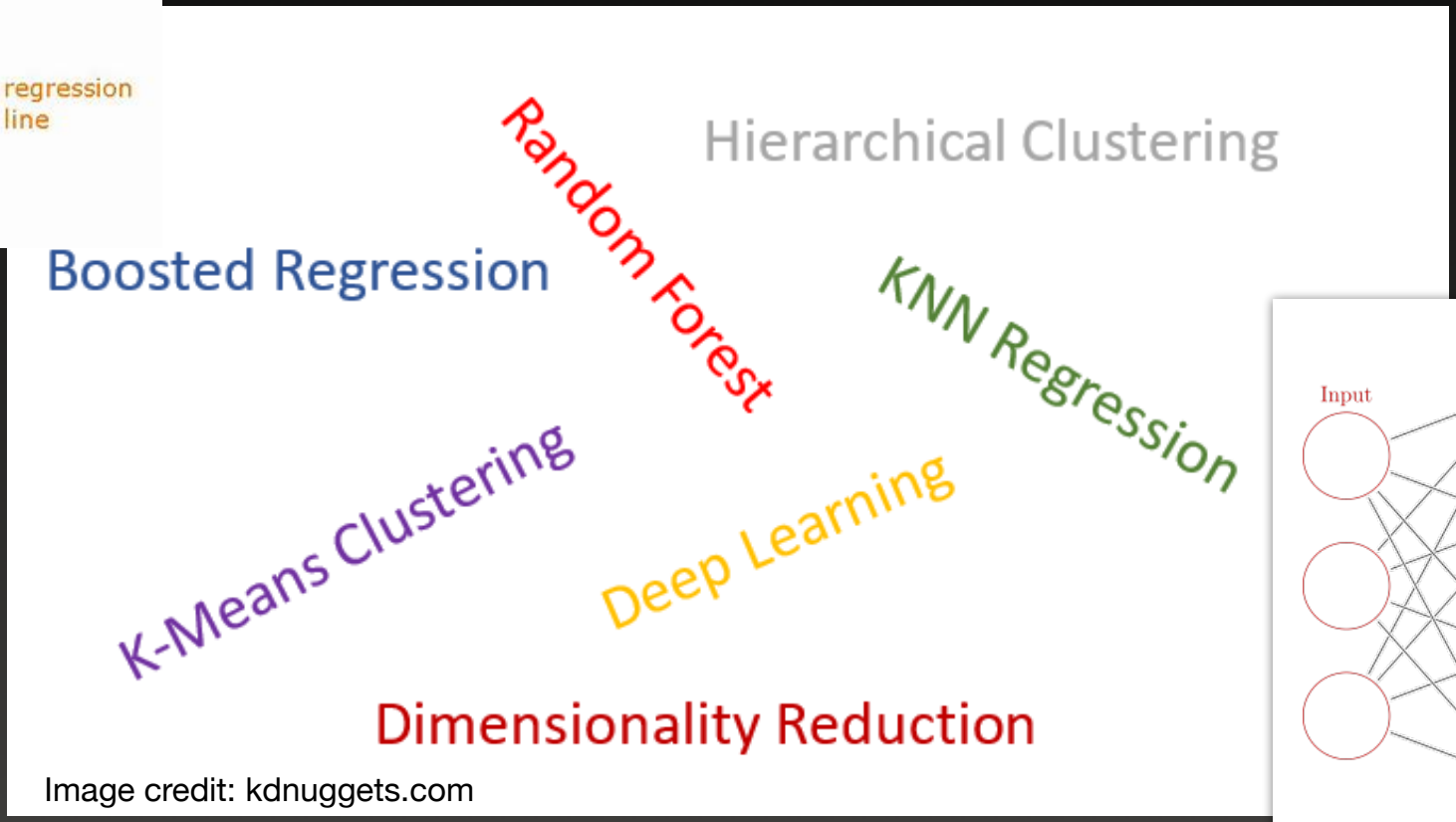
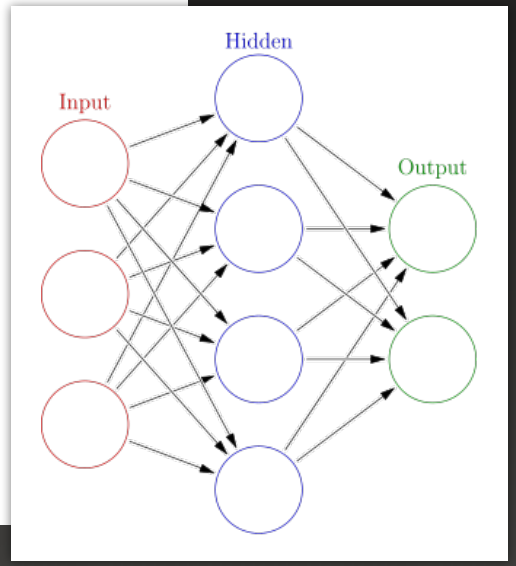


Image credit: kdnuggets.com



- In general: More flexible models require more data (examples) for "training". They also tend to be less interpretable. So there's a trade-off…

So What Kind of Problems Do Statisticians Work On?

The best thing about being a statistician is that you get to play in everyone's backyard.

– John W. Tukey

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)

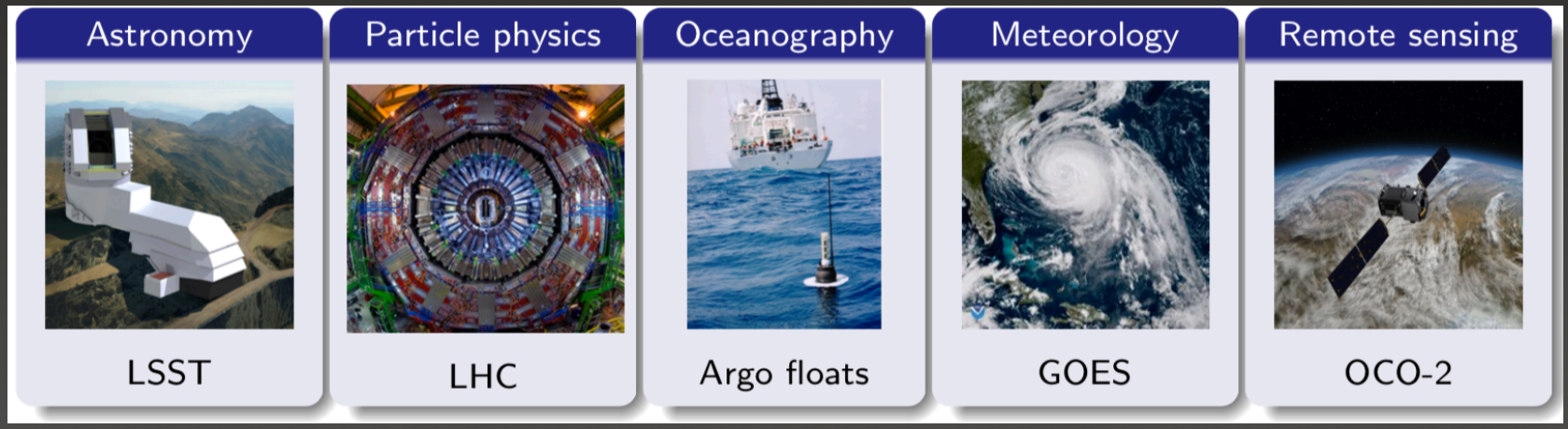
Public policy (Heinz)
Social Sciences
Finance

Astrostatistics at CMU

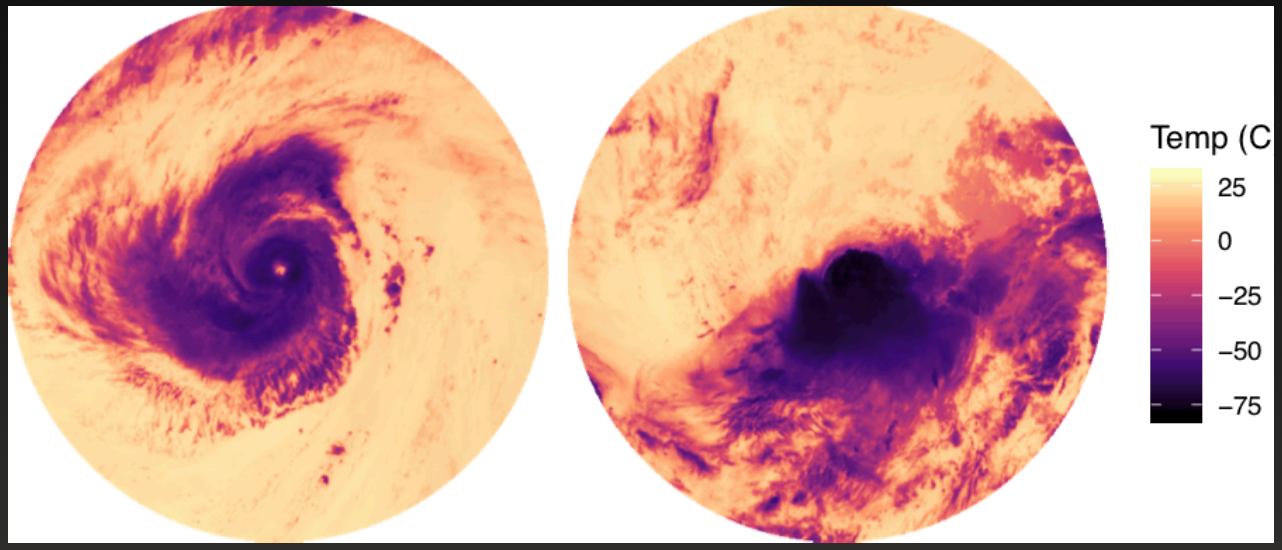
- The CMU Statistics department has been involved in astrostatistics since the late 1990s.
- Our group is one of very few (others are based at Harvard, Imperial College, UC Berkeley, UC Davis and University of Washington)
- We work closely with astronomers at Pitt and the McWilliams Center for Cosmology at CMU.

STAMPS@CMU

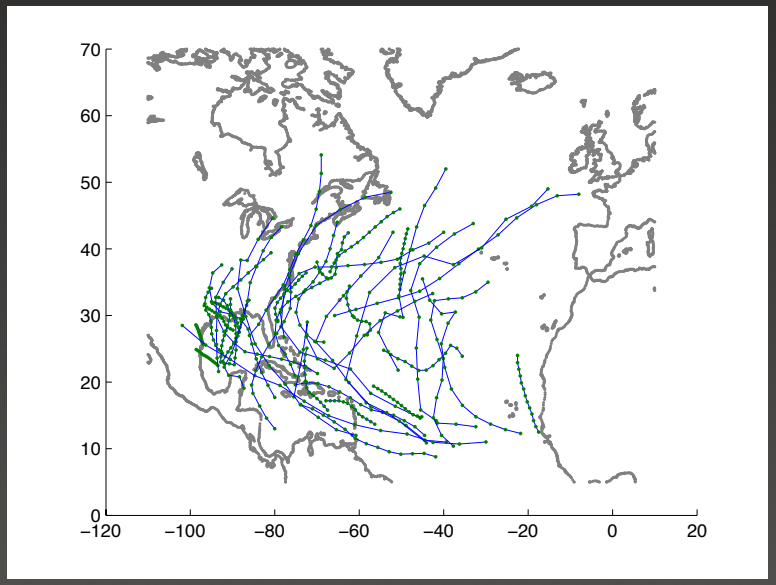
- In 2018, we started the STAtistical Methods for the Physical Sciences (STAMPS) research group.
- Problems in the physical sciences have similar statistical challenges —- involving massive data sets from different physical probes (e.g. images from satellites), large simulations, and complex measurement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

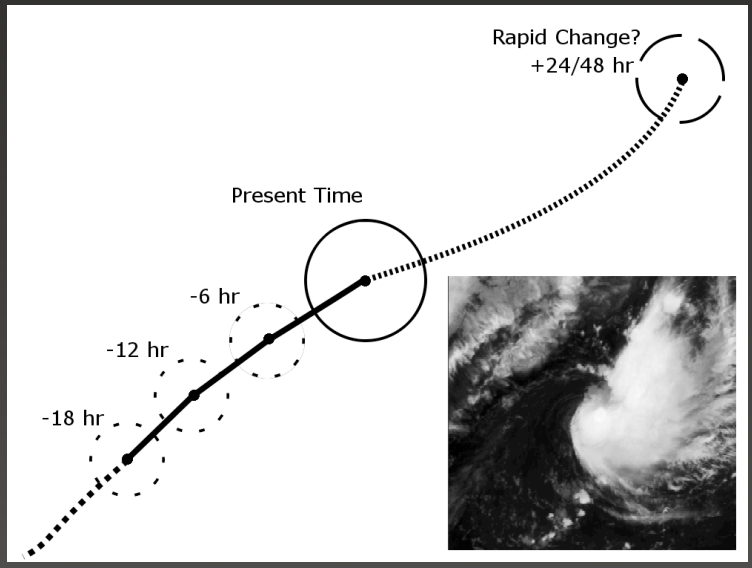


Modeling Hurricane Intensity Change Using GOES Satellite Imagery [McNeely/Lee/Wood/Hammerling]



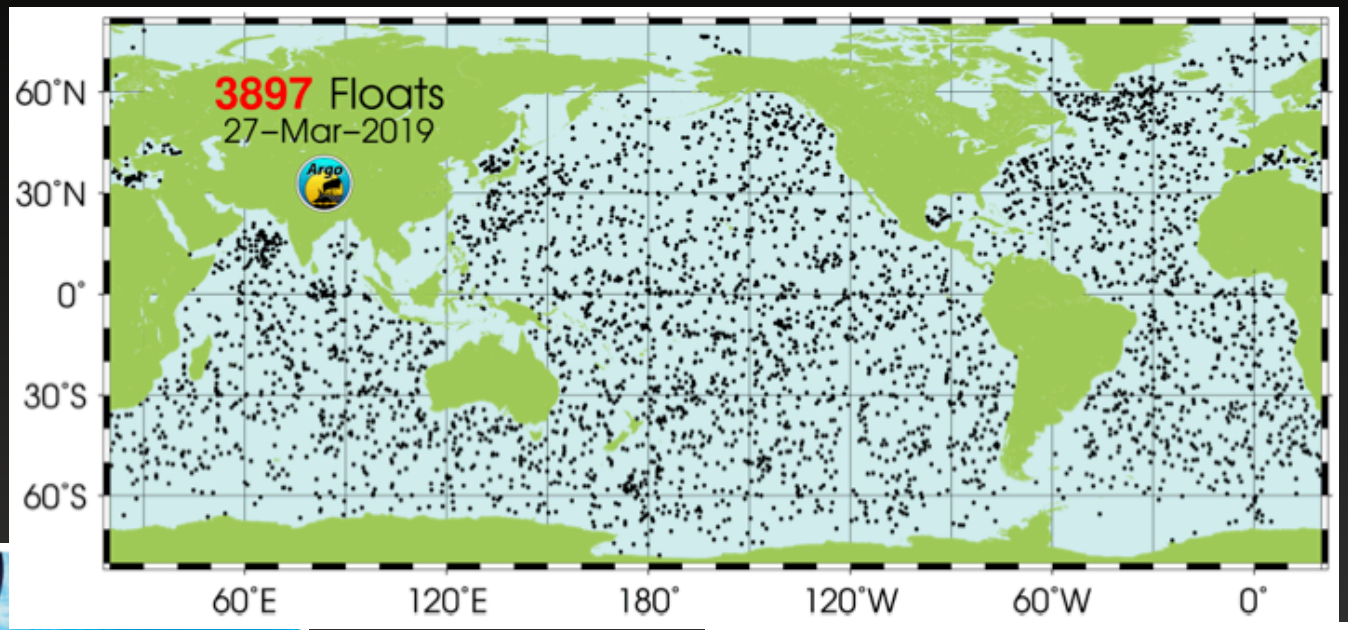
Hurricanes Edouard (2014; 109mph) and Nicole (2016; 54mph)

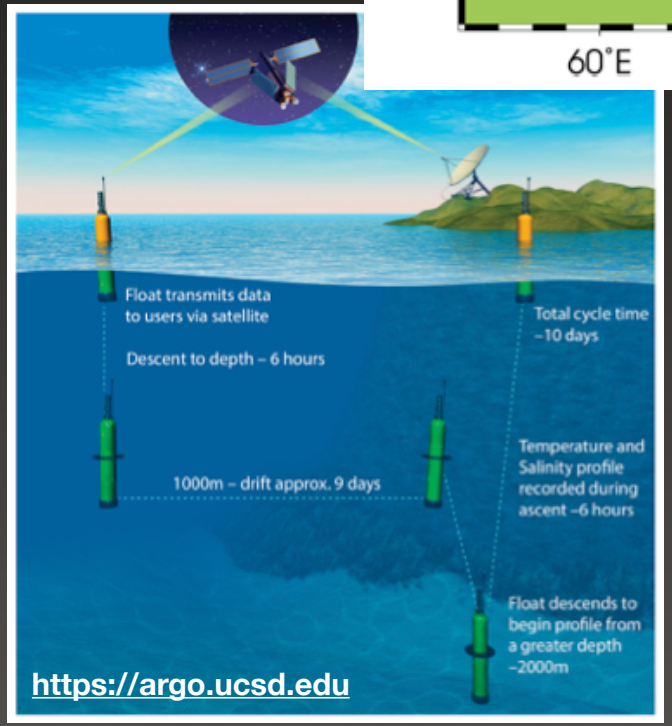


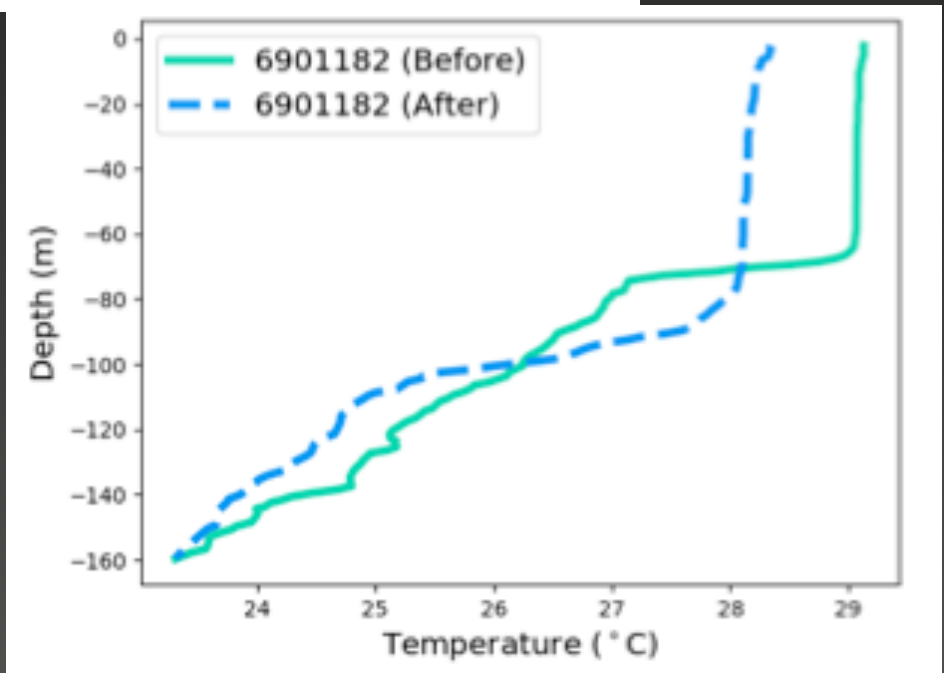


Hurricanes and Ocean Heat Content

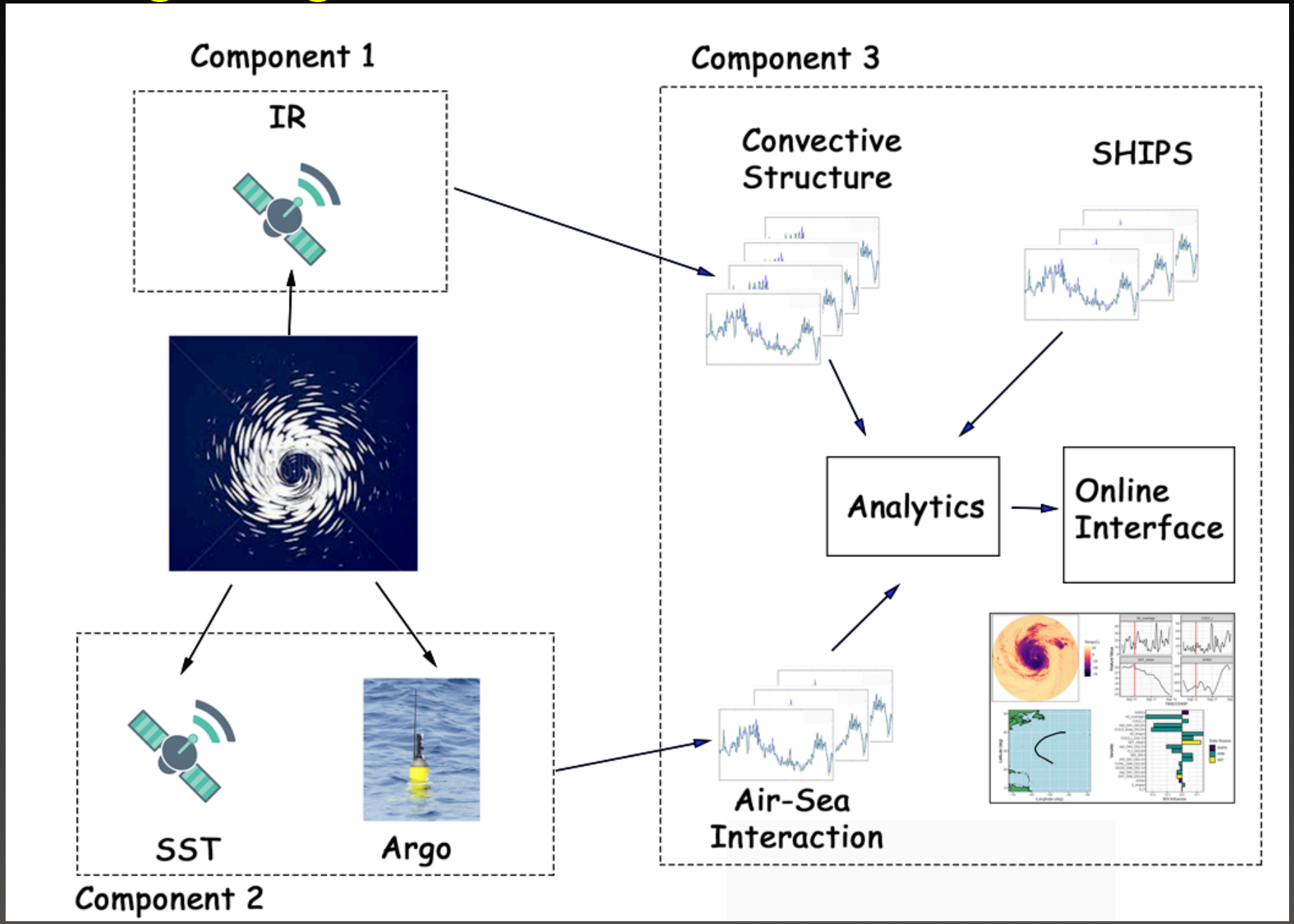
[Hu/Kuusela/Lee/Giglio/Wood]



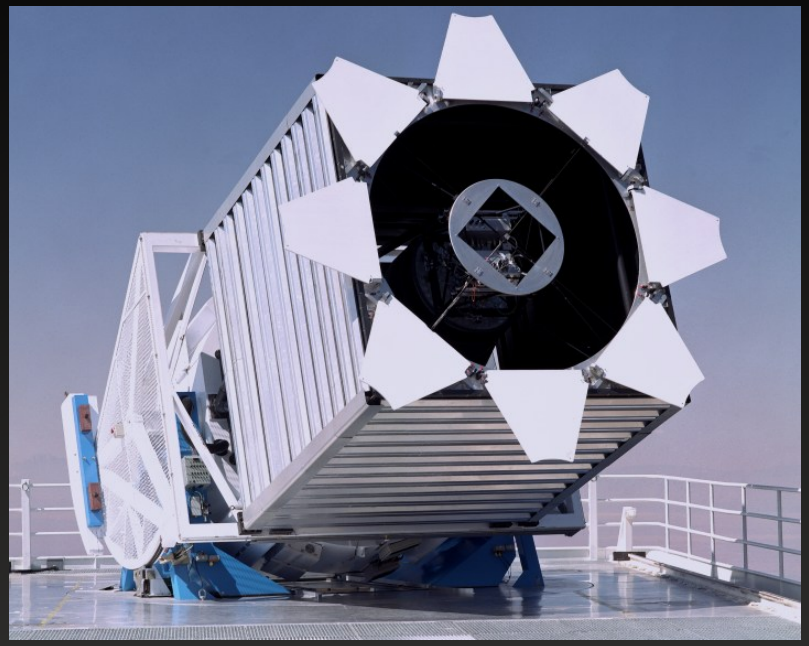




Joint Analysis of Hurricane Intensity Change by Integrating Satellite and In Situ Observations

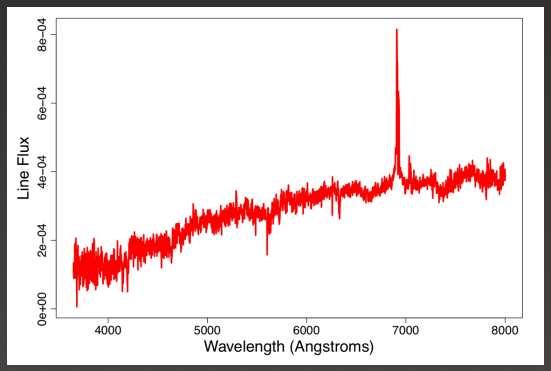


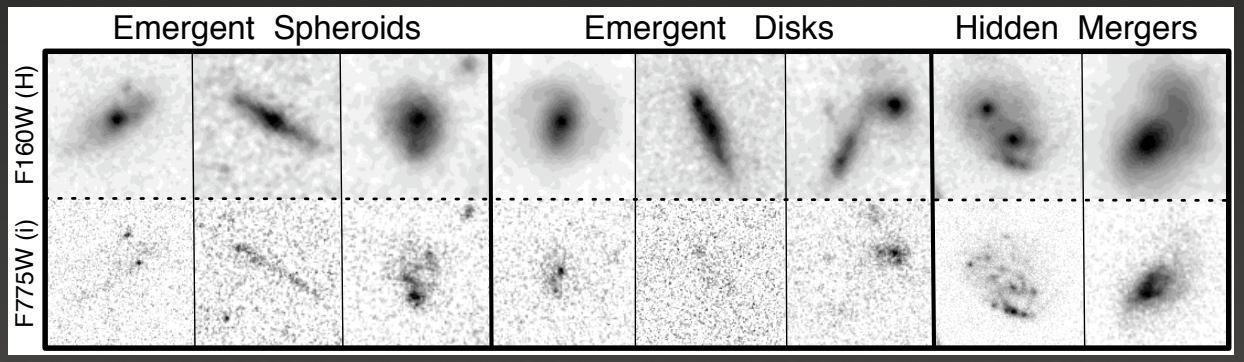
Tons of Data and Exciting Problems in Astronomy…



Sloan Foundation Telescope (2.5-m)

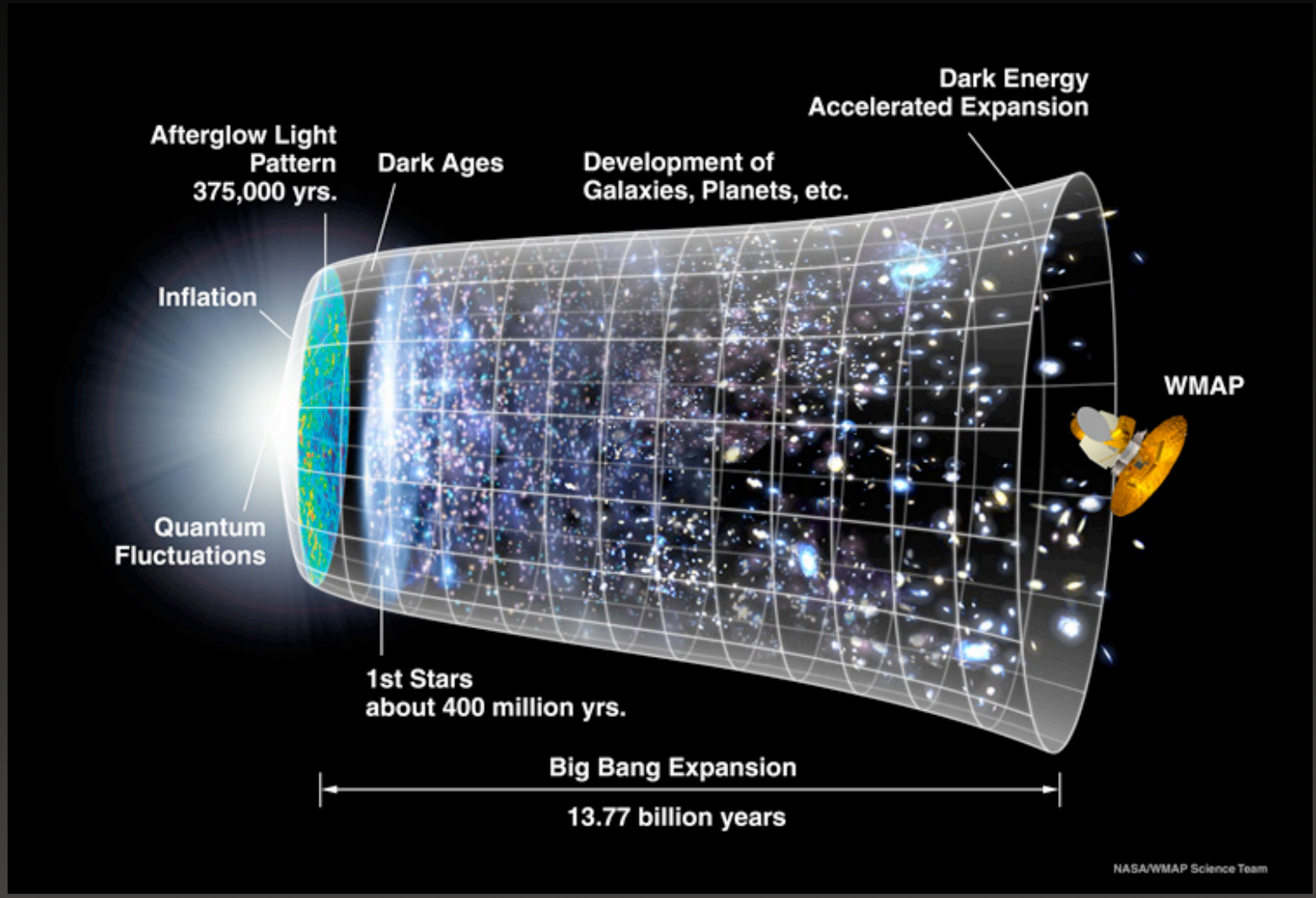
Hubble Space Telescope





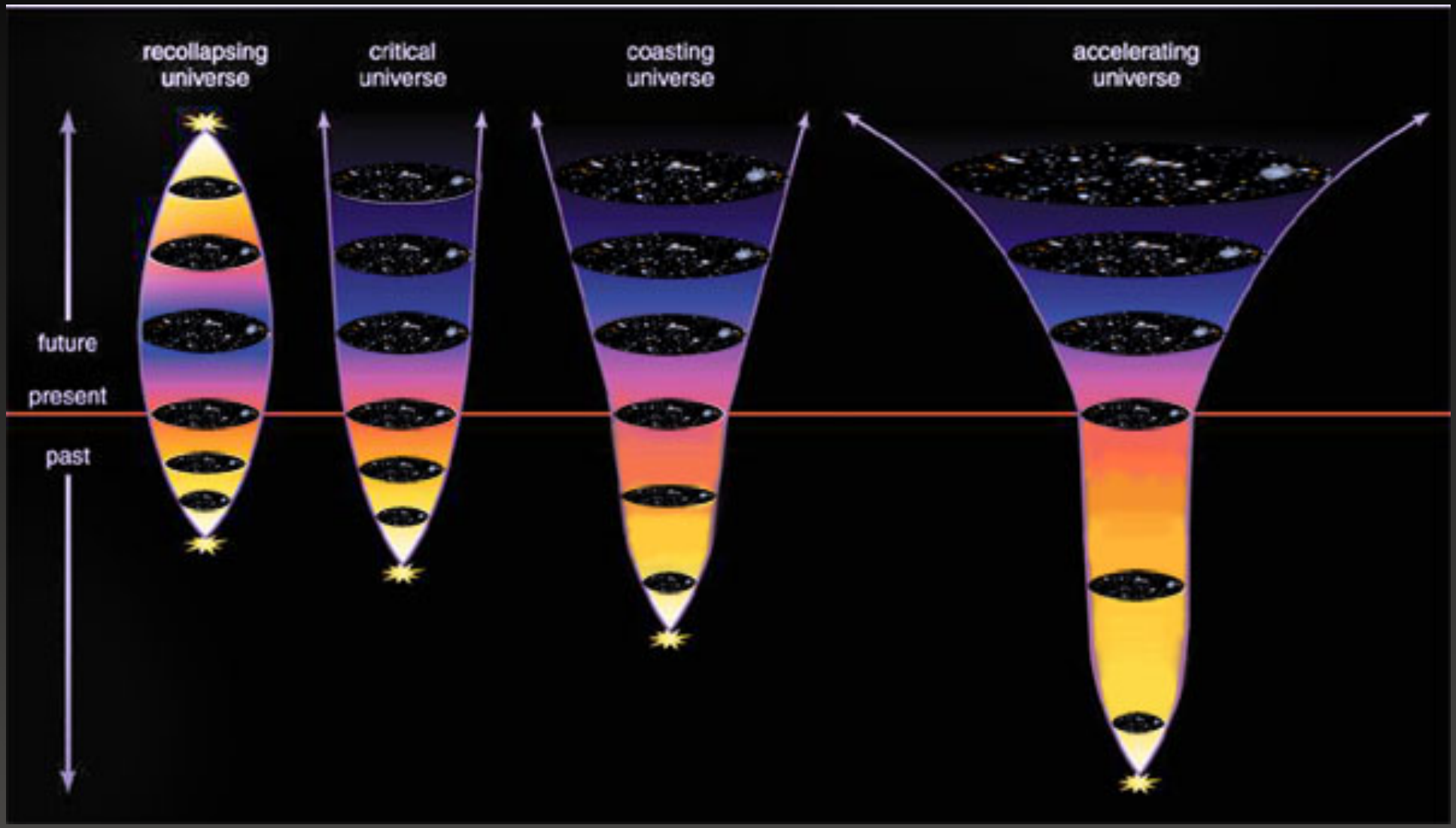
Our main focus at CMU Astrostats is on cosmology — the study of the origin, evolution and structure of the Universe.

First a Brief Cosmology Primer…



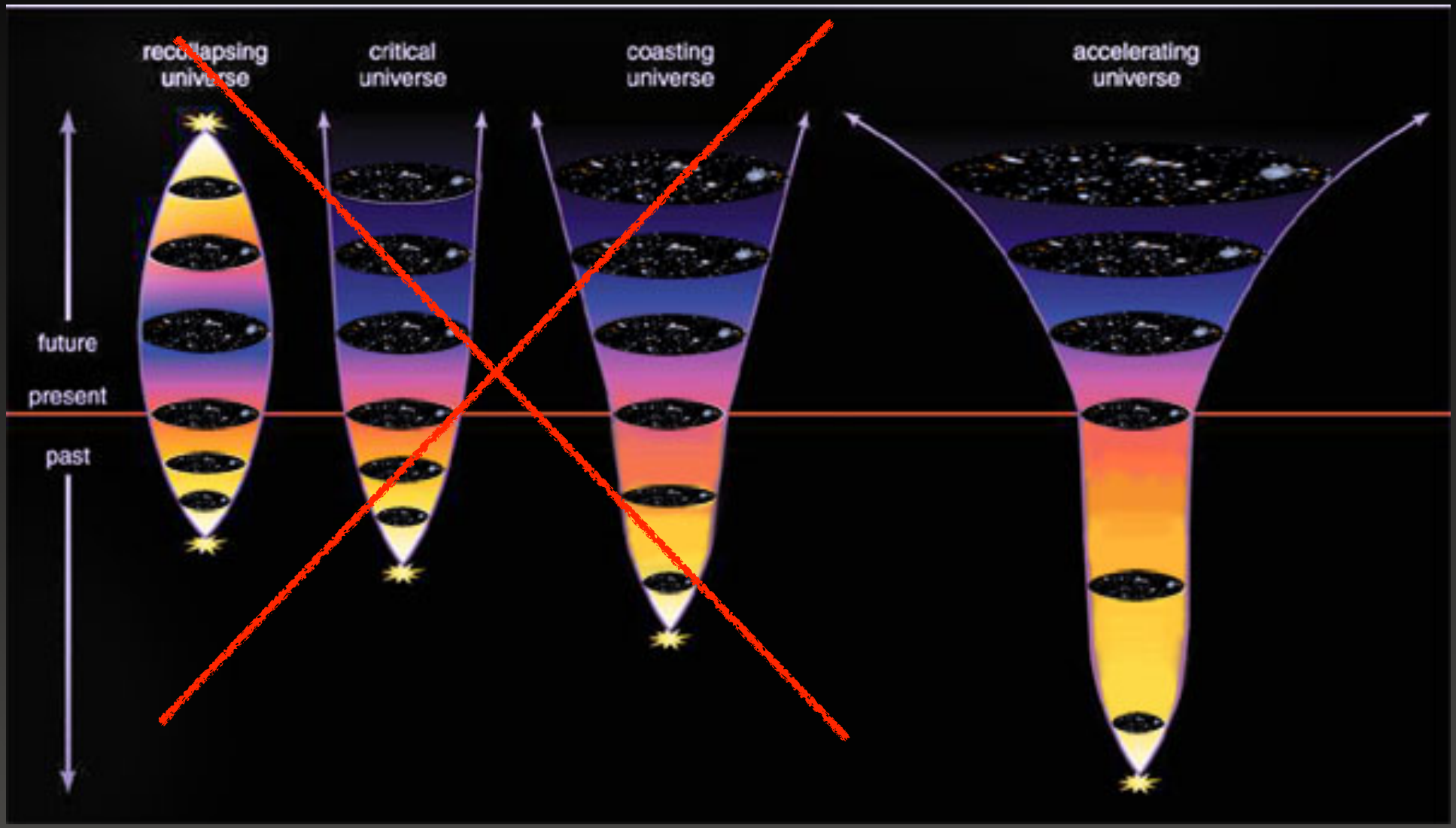
- Go back in time by probing deeper into the sky. Allows us to constrain theories of how the Universe is evolving.

Possible Fates of the Universe



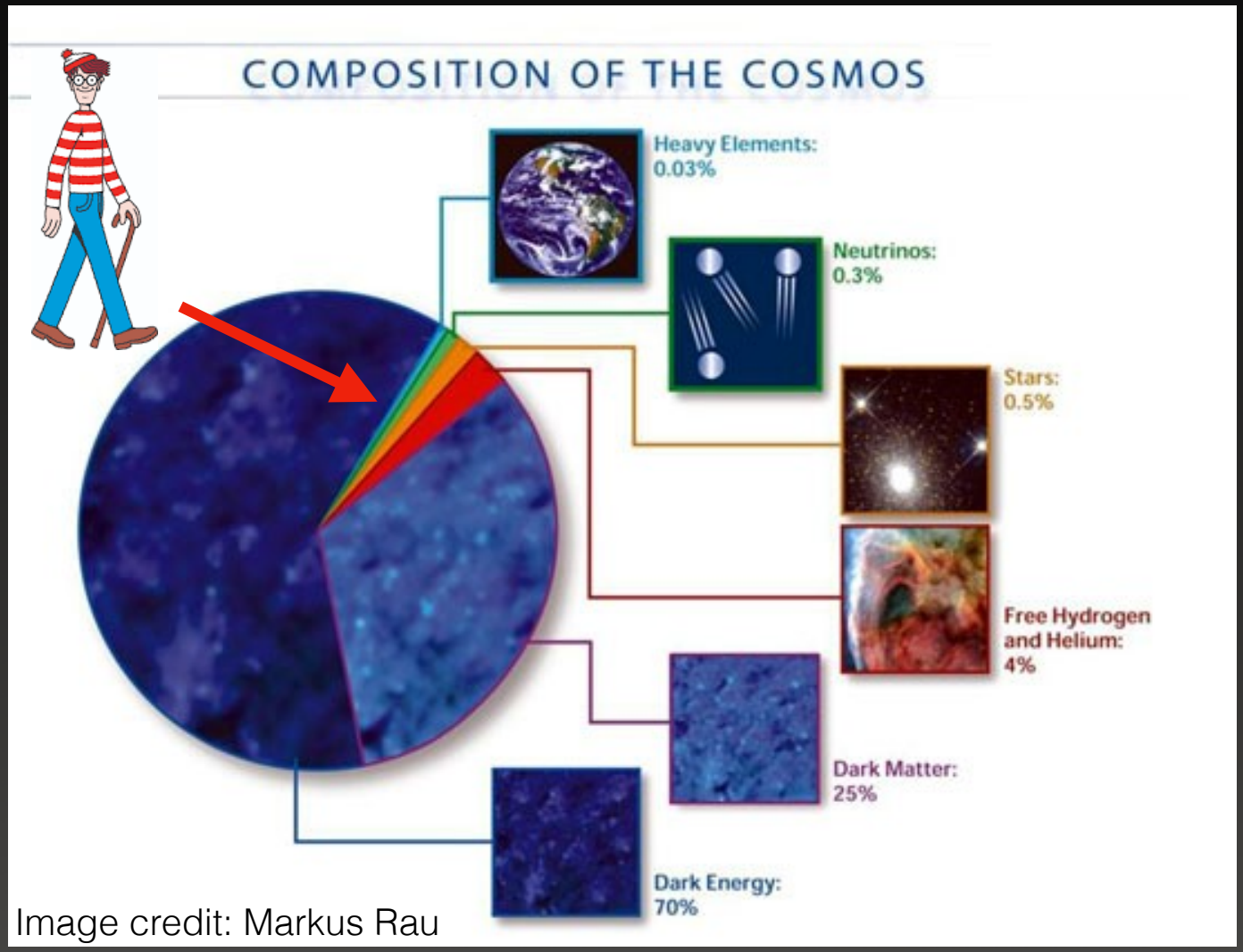
- The ``Big Crunch'' or the ``Big Rip'' or a coasting universe?

Possible Fates of the Universe



- Yes, we live in an accelerating universe!

What is the Universe Made Up of?

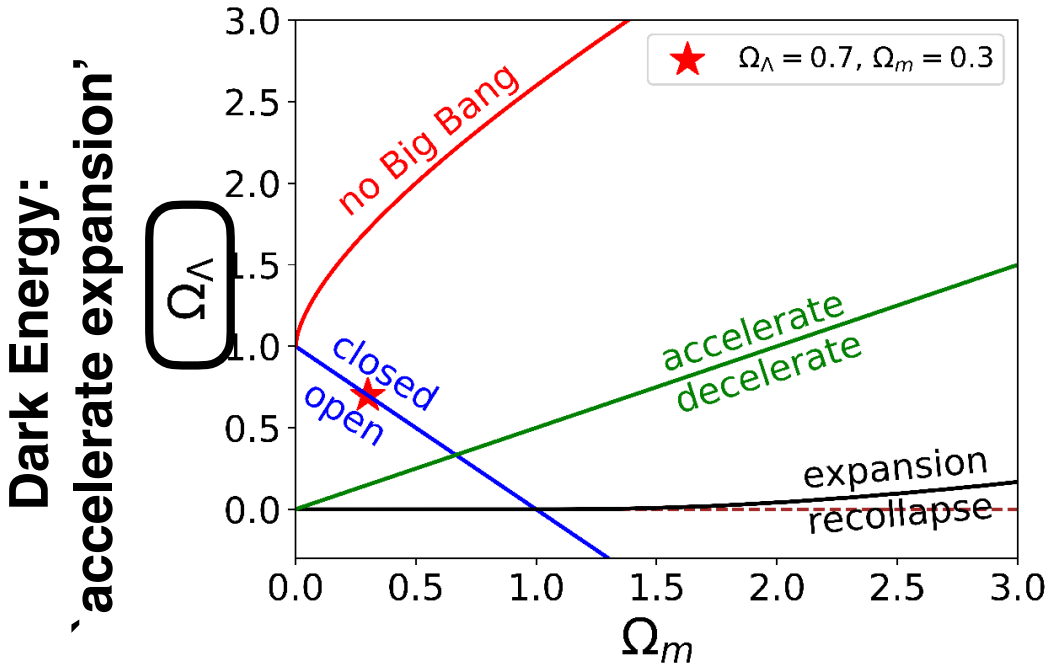


https://www.lsst.org/science/dark-energy

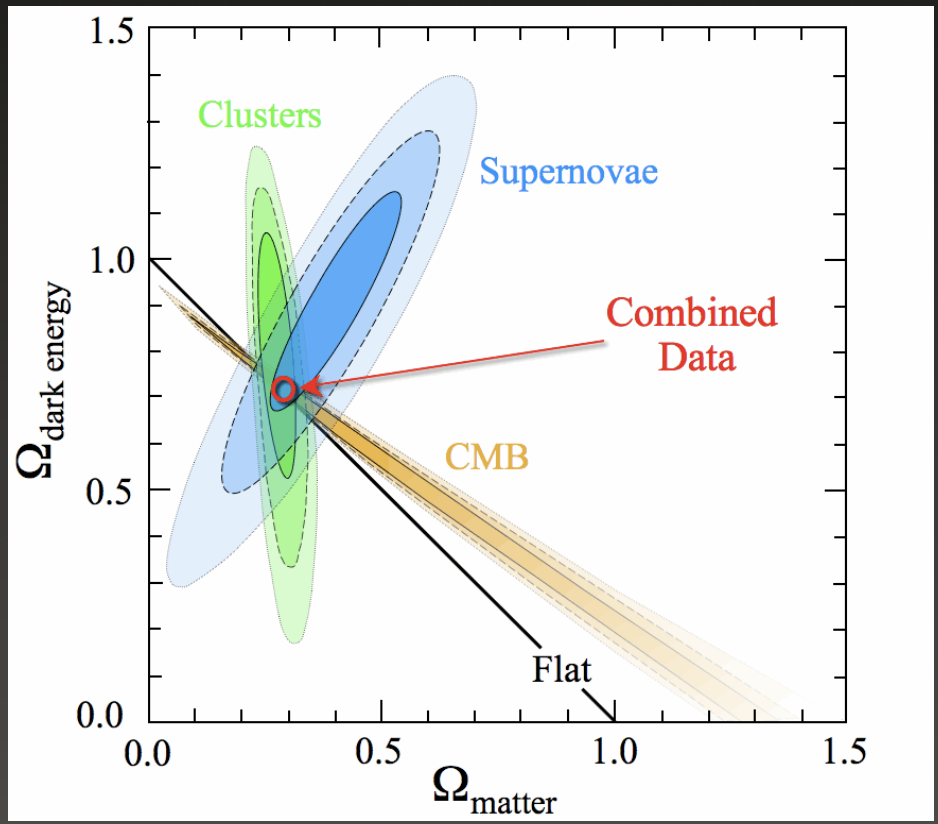
- 70% Dark Energy; 25% Dark Matter; 5% Ordinary Matter

The Standard Model of Big Bang Cosmology (Lambda CDM Model)

- Lambda-CDM is a parametrization of Big Bang cosmology.
- These parameters can be constrained by observations.

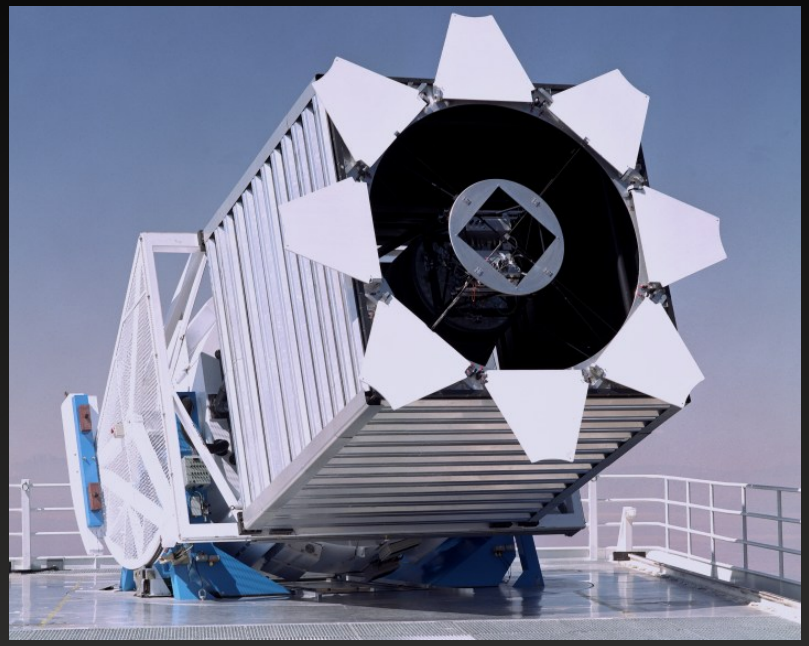


(Dark) Matter:
`pulls universe together'



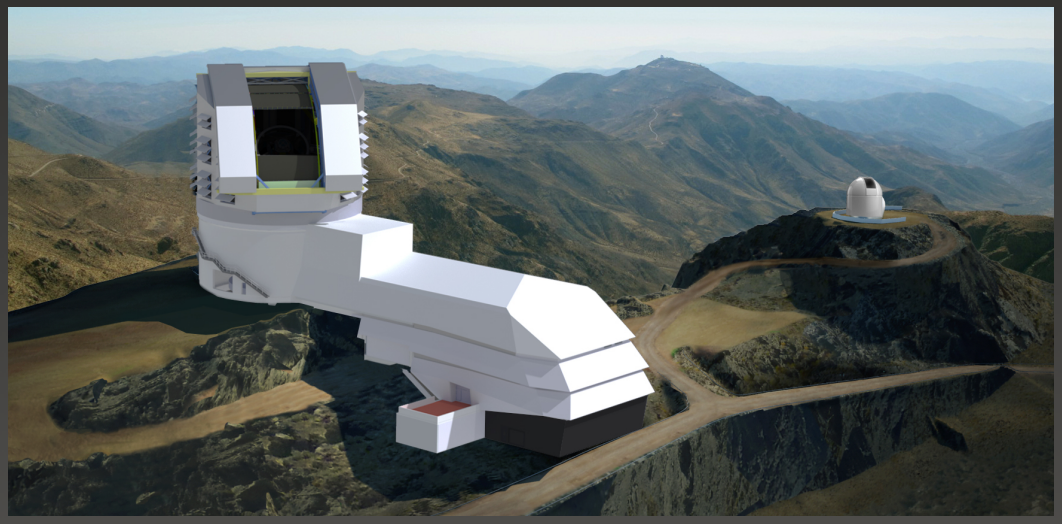
Source: https://www.hep.ucl.ac.uk/darkMatter/

Mapping the Sky



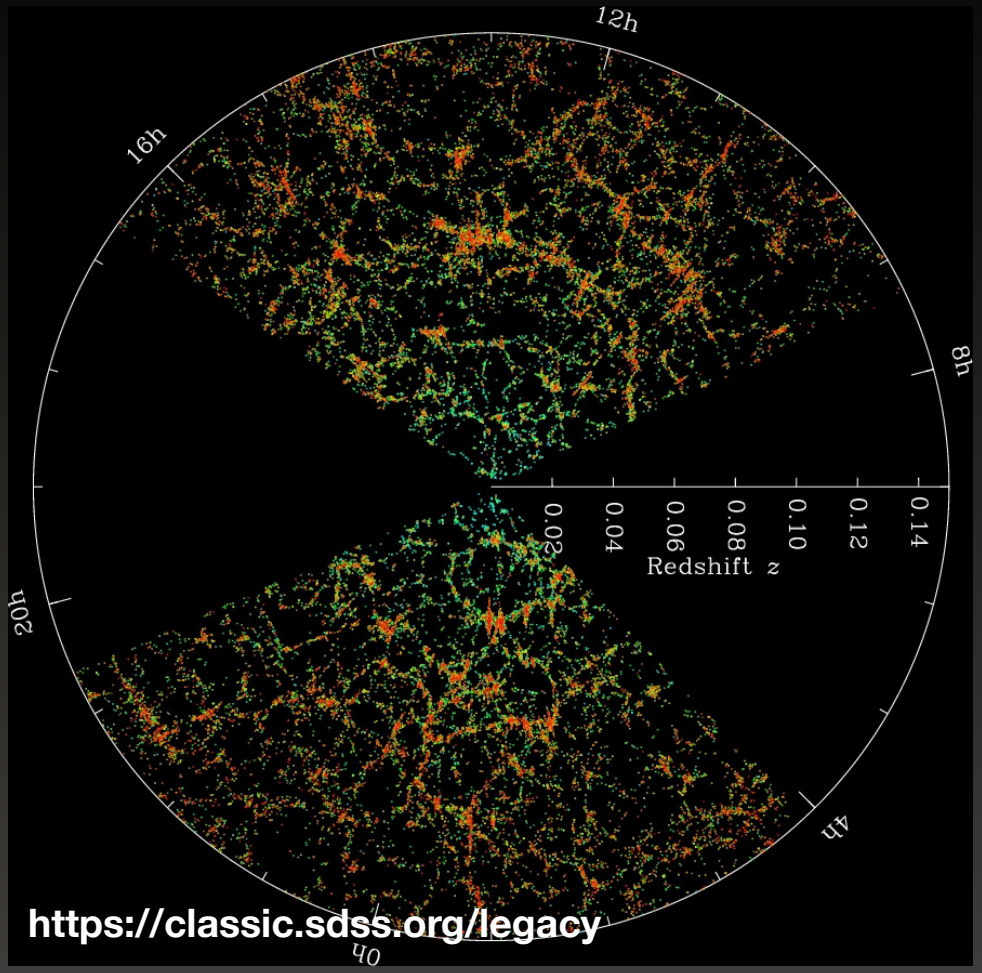
Sloan Foundation Telescope (2.5-m)

Hubble Space Telescope



Large Synoptic Survey Telescope (LSST), 8.4 mirror, ready to launch in 2022

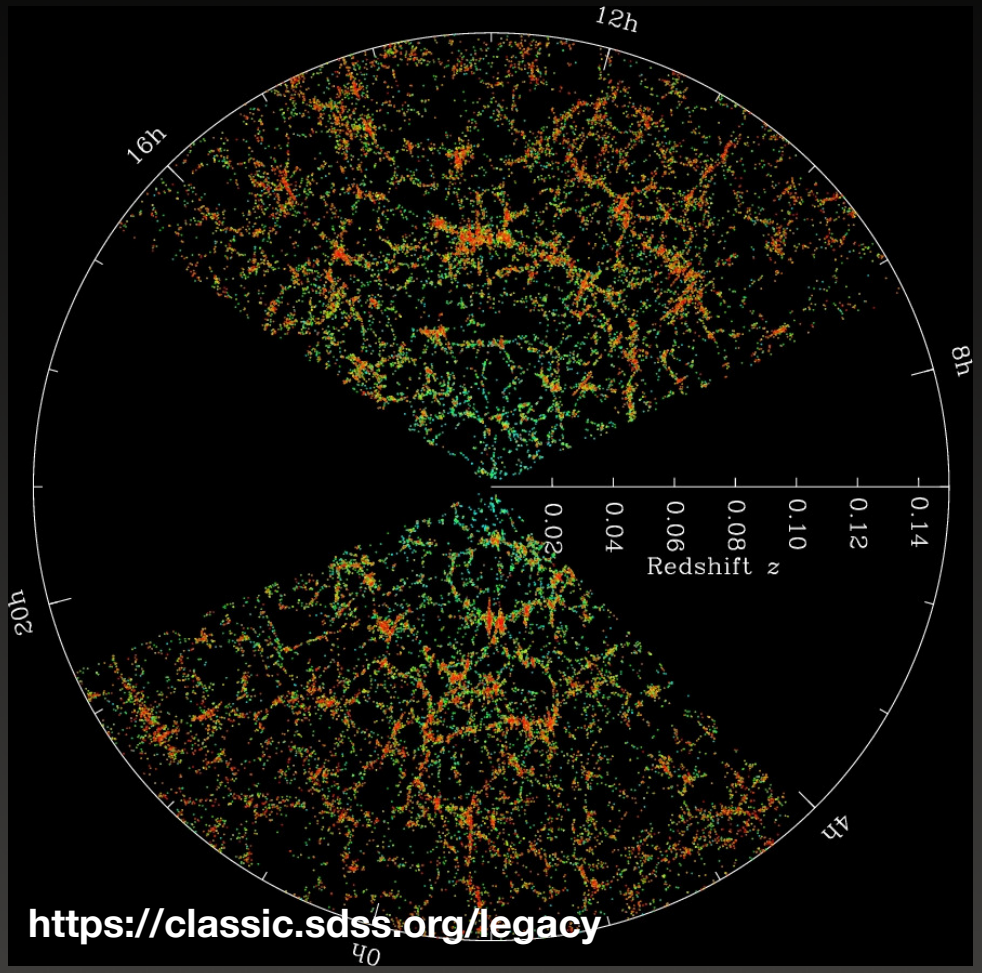
How Much Data Do We Have?



- We have entered the era of ``precision cosmology''.
- We have 200 million galaxy images from ``shallow'' surveys (like SDSS).
- We have 200,000 galaxy images from ``deep'' surveys (like HST CANDELS)

Credit: Peter Freeman

How Much Data Do We Have?



- We have entered the era of ``precision cosmology''.
- We have 200 million galaxy images from ``shallow'' surveys (like SDSS).
- We have 200,000 galaxy images from ``deep'' surveys (like HST CANDELS)

- Beginning in 2022, LSST will begin collecting images of several billion galaxies. (There are thought to be over 200 billion galaxies in the Universe)

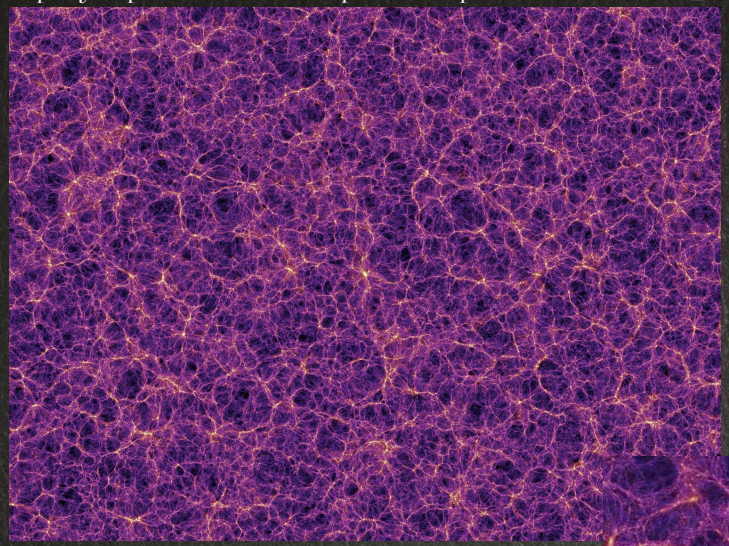
Credit: Peter Freeman

Simulating the Evolution of the Universe

- To understand how the universe evolved into the structures we see today, we can use computer simulations to numerically evolve a representation of some fraction of the universe in time.

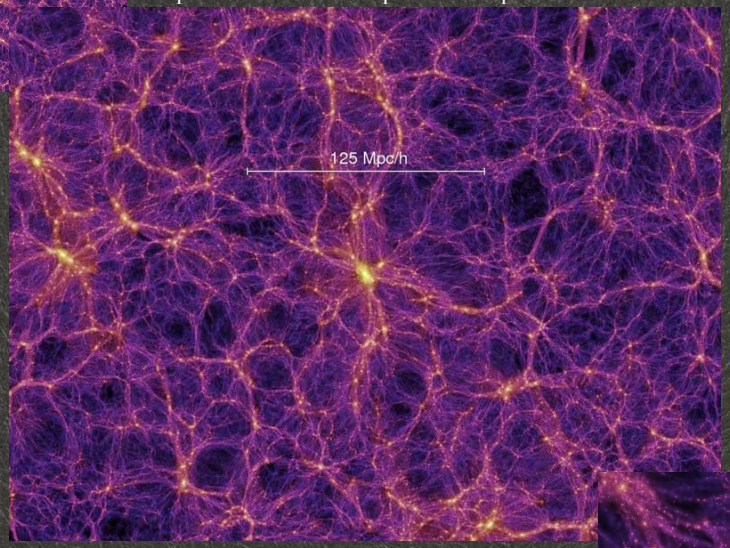
https://www.illustris-project.org

http://cyberpunkswebsite.com/wp-content/uploads/2014/01/cosmic_web_3.jpeg



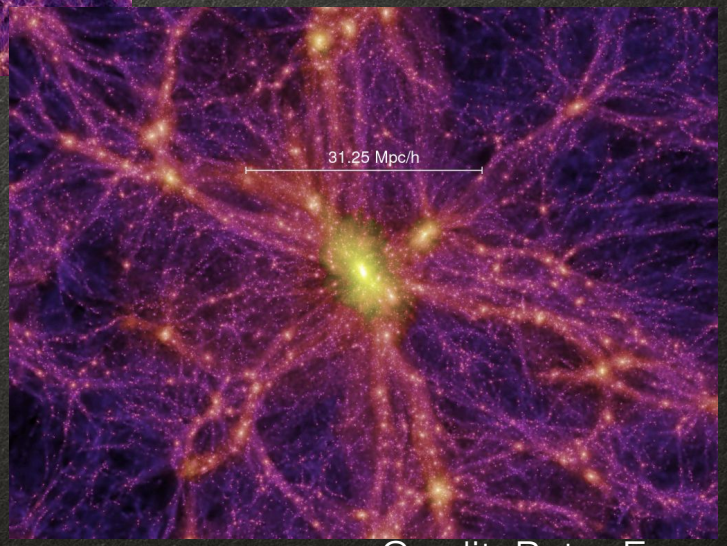
The "cosmic web"

http://astrobites.com/wp-content/uploads/2012/07/cosmic-web.jpg



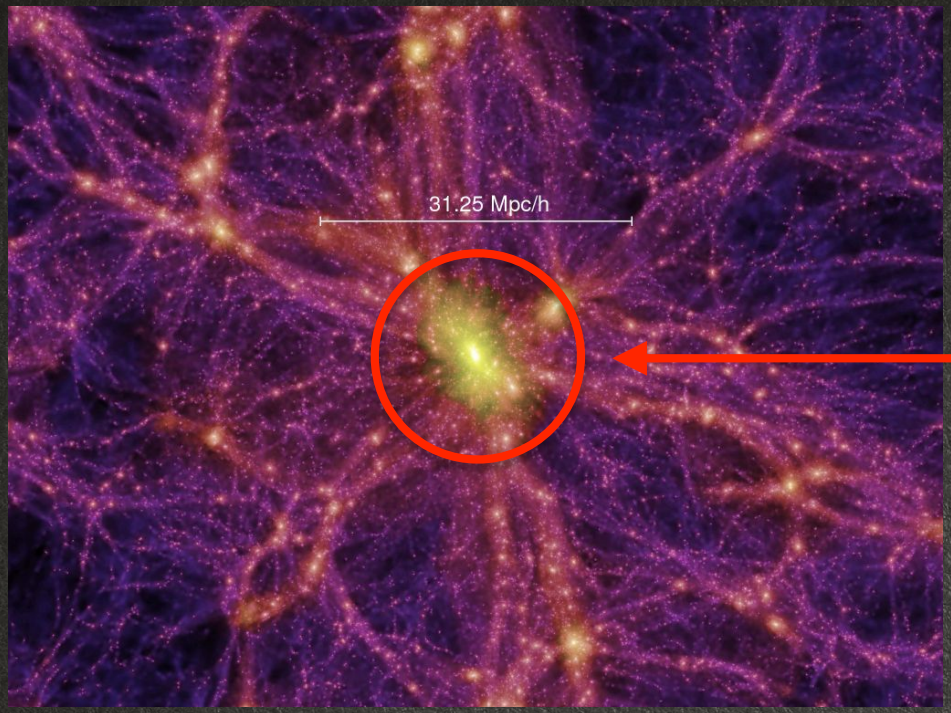
A Simulated Universe

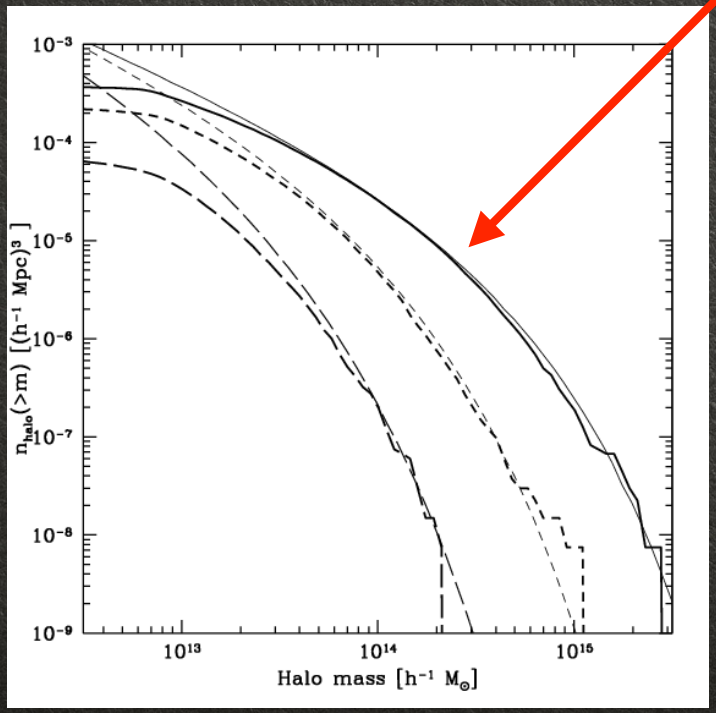
A cluster within the cosmic web (which may contain hundreds of galaxies)



http://www.americaspace.com/wp-content/uploads/2014/01/Astro_Web_Credit_Peter_Freeman.jpg

Credit: Peter Freeman





http://inspirehep.net/record/854412/files/Hallman07f3.png

Key points:

1) Simulation output varies as a function of input theoretical parameters.

2) Simulations create a distribution of clusters that all have different masses.

3) The observed distribution of ***real*** clusters and their masses allow us to constrain the theory that goes into the simulations.

Credit: Peter Freeman

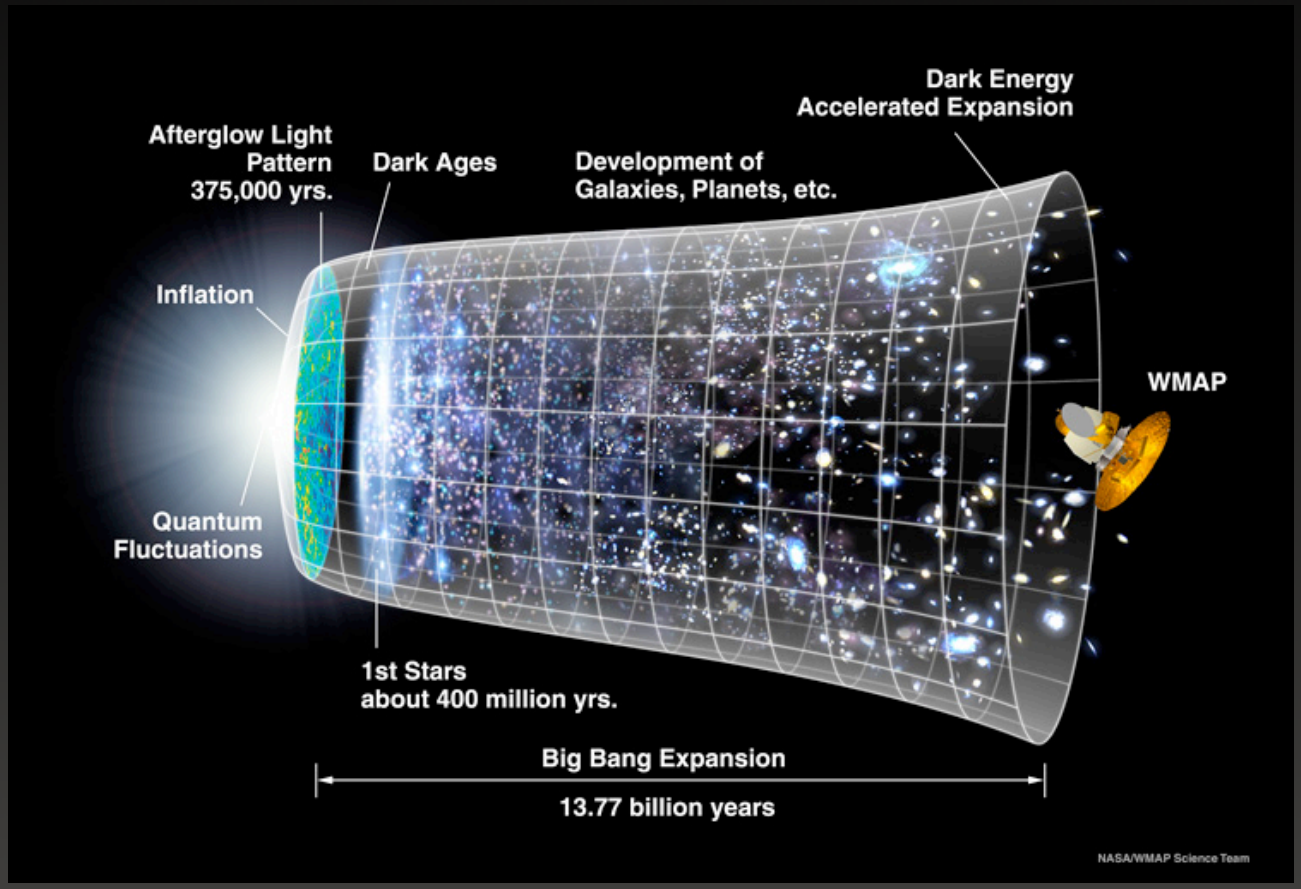
Examples of Things I've Worked On: Three Astrostatistics Problems

1. How do you measure the distance to a galaxy?
2. How do you weigh a galaxy or a galaxy cluster?
3. How do you constrain models of the Universe?

Examples of Things I've Worked On: Three Astrostatistics Problems

1. How do you measure the distance to a galaxy?
2. How do you weigh a galaxy or a galaxy cluster?
3. How do you constrain models of the Universe?

Recall: ``Look-Back Time'" — Go Back in Time by Probing Deeper into the Sky.



- Can't measure distances directly —- but because we live in an expanding universe, more distant galaxies appear to be moving away faster from us than less distant galaxies.

We Measure Distances to Distant Galaxies by Estimating their ``Redshift''

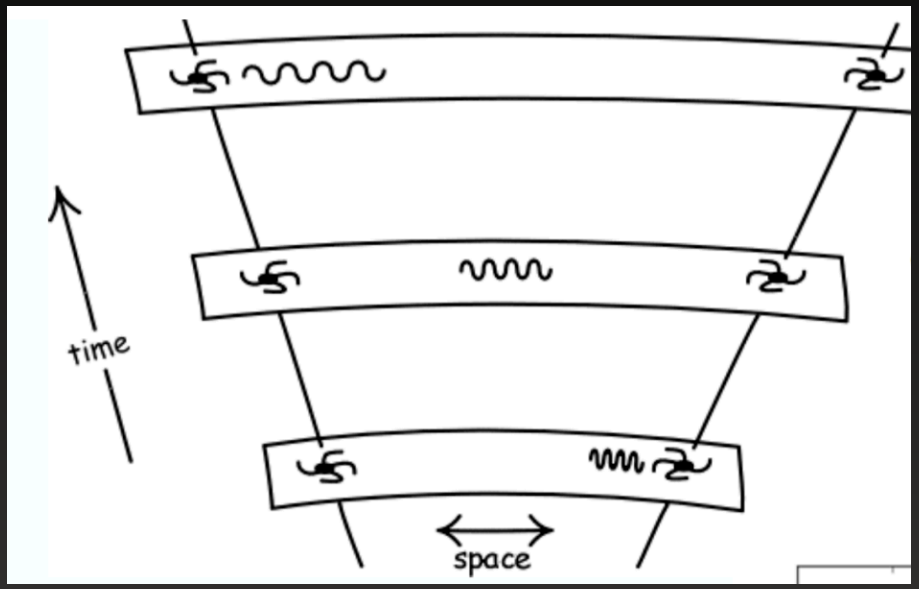


Image credit: Markus Rau

- Background universe expands and stretches light waves.
- Observed spectra shifts to longer wavelengths.
- Redshift (z) = proxy for distance

We Measure Distances to Distant Galaxies by Estimating their ``Redshift''

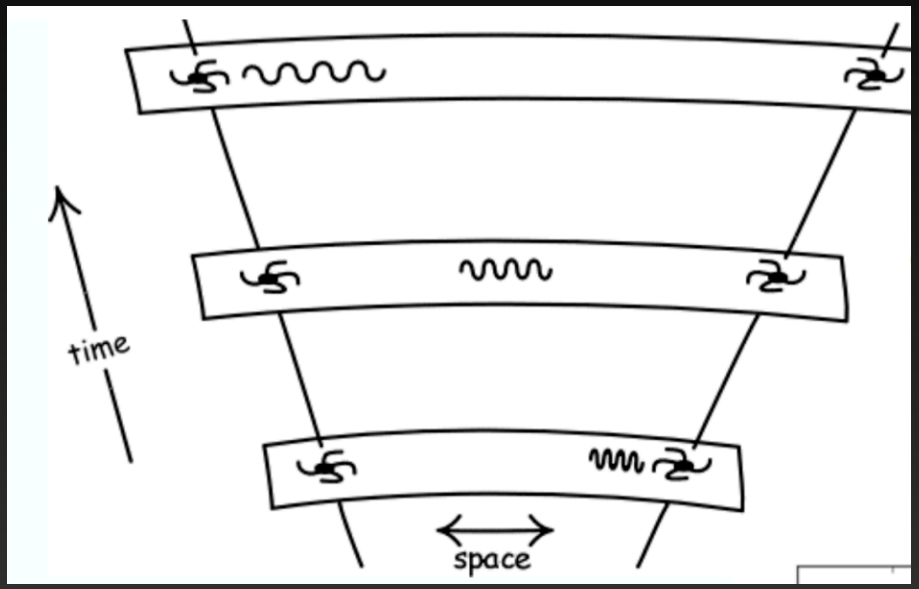
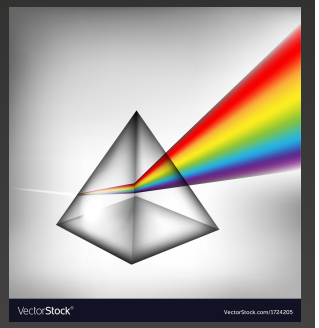
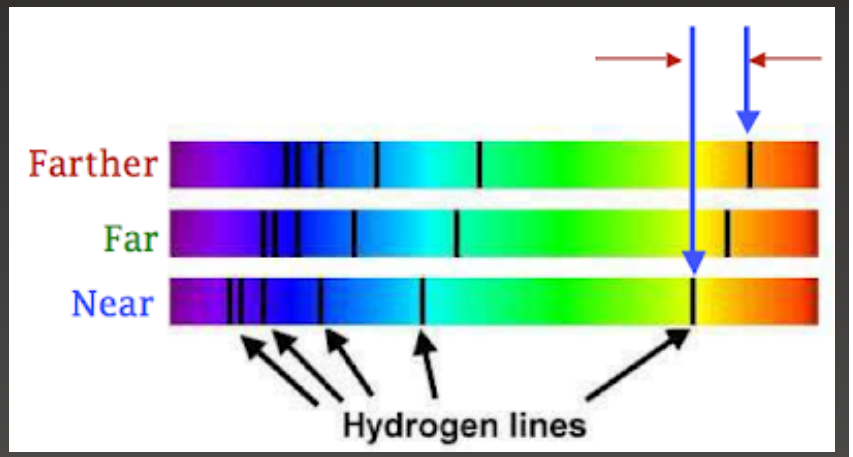


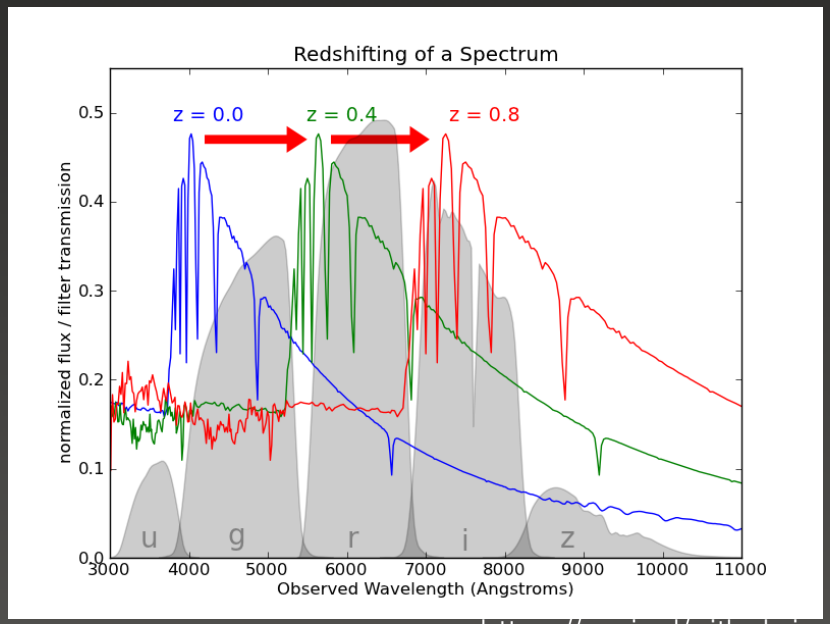
Image credit: Markus Rau

- Background universe expands and stretches light waves.
- Observed spectra shifts to longer wavelengths.
- Redshift (z) = proxy for distance



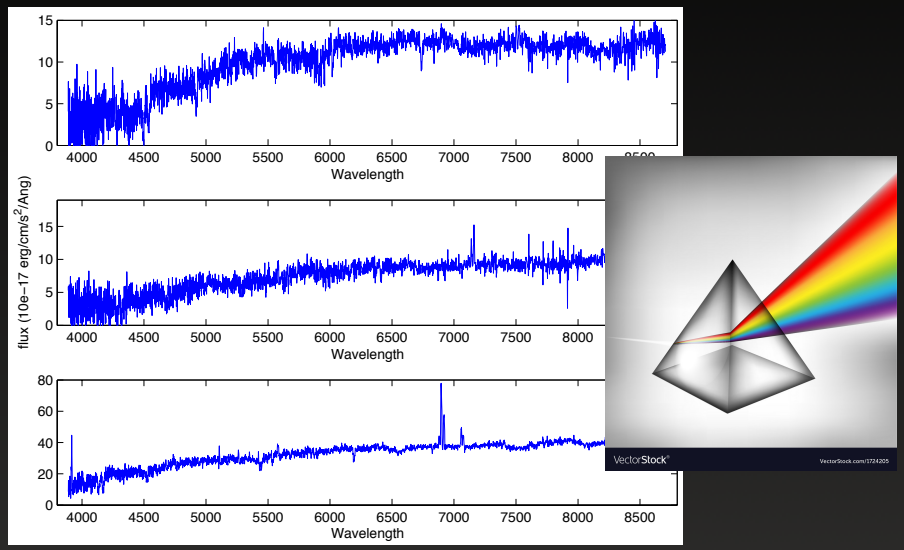


$$z = \frac{\lambda_{\rm obsv} - \lambda_{\rm emit}}{\lambda_{\rm emit}}$$



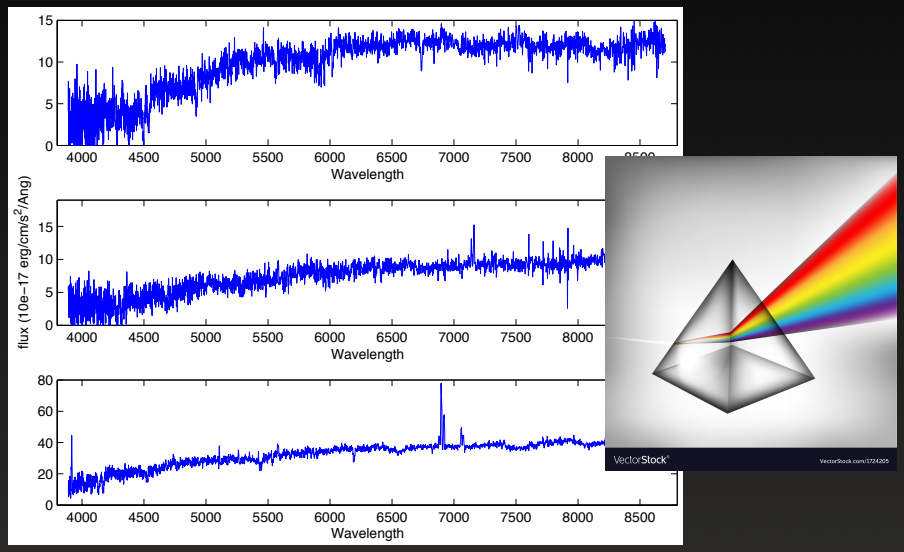
https://ogrisel/github.io

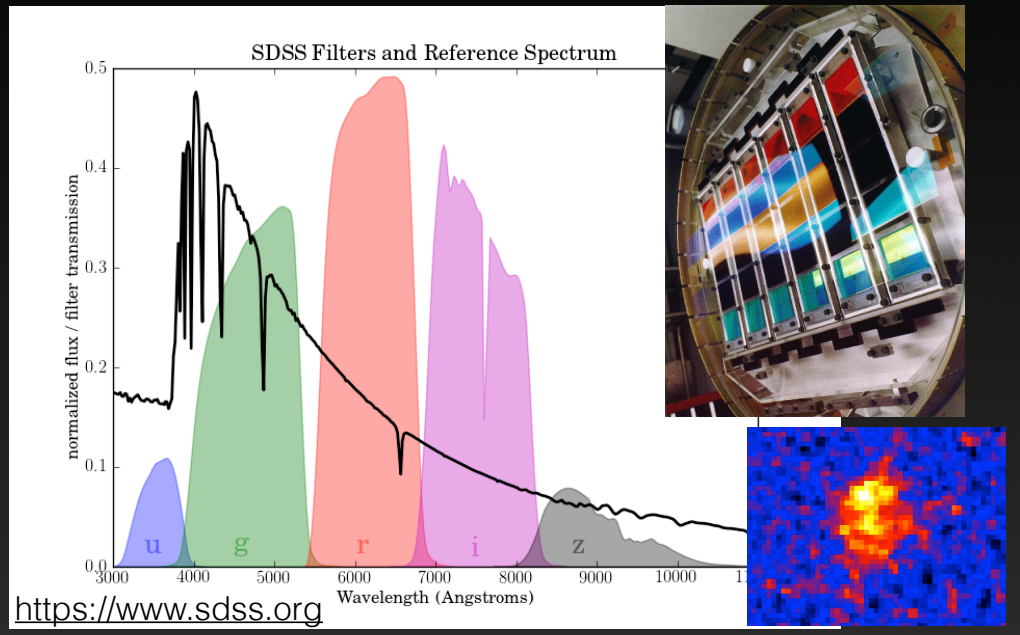
Redshift Estimation from Photometry



- Left: High-resolution galaxy spectra from spectroscopy.
- Spectroscopy resource intensive ⇒ More than 99 percent of today's galaxy observations are instead from photometry.
- Right: Photometry (broad-band filters)
- Challenge -- estimate redshift using photometric "colors"

Redshift Estimation from Photometry



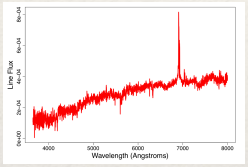


- Left: High-resolution galaxy spectra from spectroscopy.
- Spectroscopy resource intensive ⇒ More than 99 percent of today's galaxy observations are instead from photometry.
- Right: Photometry (broad-band filters)
- Challenge -- estimate redshift using photometric "colors"

Main Steps in "Photo-z Estimation":

Obtain a Calibration catalog

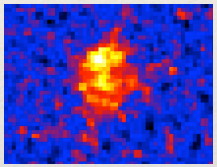
Spectroscopic observations



Accurate Distance Information

Spatially match

Image (photometric) observations



`Target' catalog

Smaller calibration catalog with **both** photometric information and accurate distance information

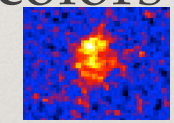
Slide by Markus Rau

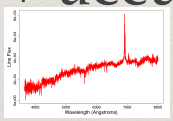
Main Steps in "Photo-z Estimation":

Applications of Machine learning

Match Calibration data from spatially overlapping region of
spectroscopic survey and photometric survey

matched Catalog
colors + accurate distances

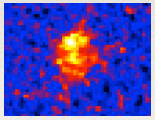




`Train'

Known
accurate Distance

Only photometric
information
colors



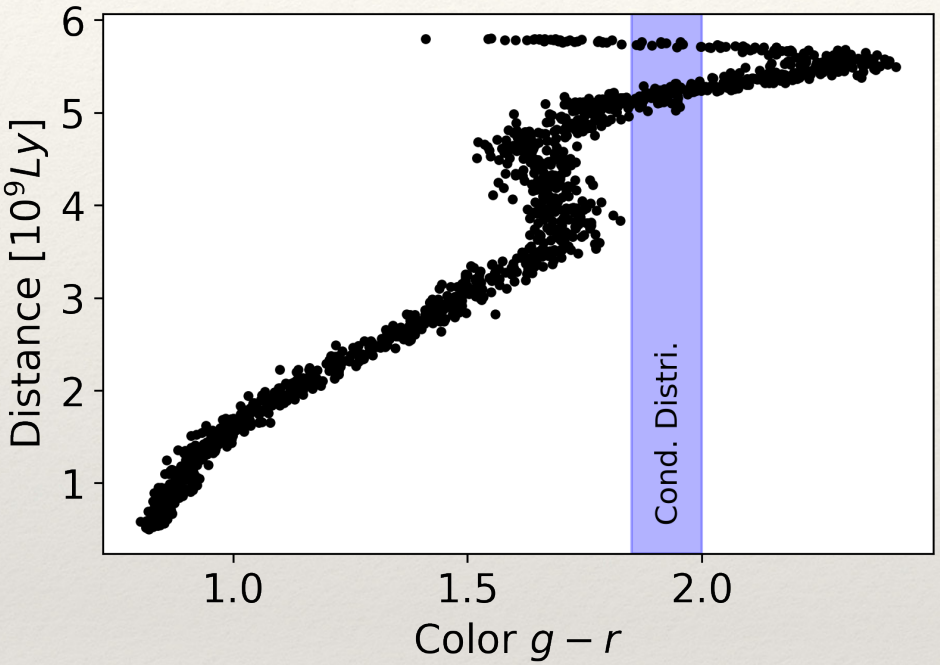
Apply

Predict Distance

Obtain distances for all galaxies in the (photometric) sample

Slide by Markus Rau

However, multiple widely different distances (redshift) can be consistent with the observed colors of a galaxy…



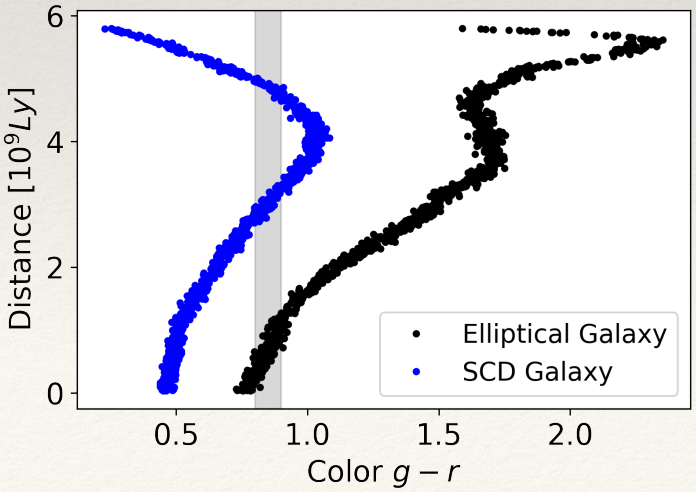
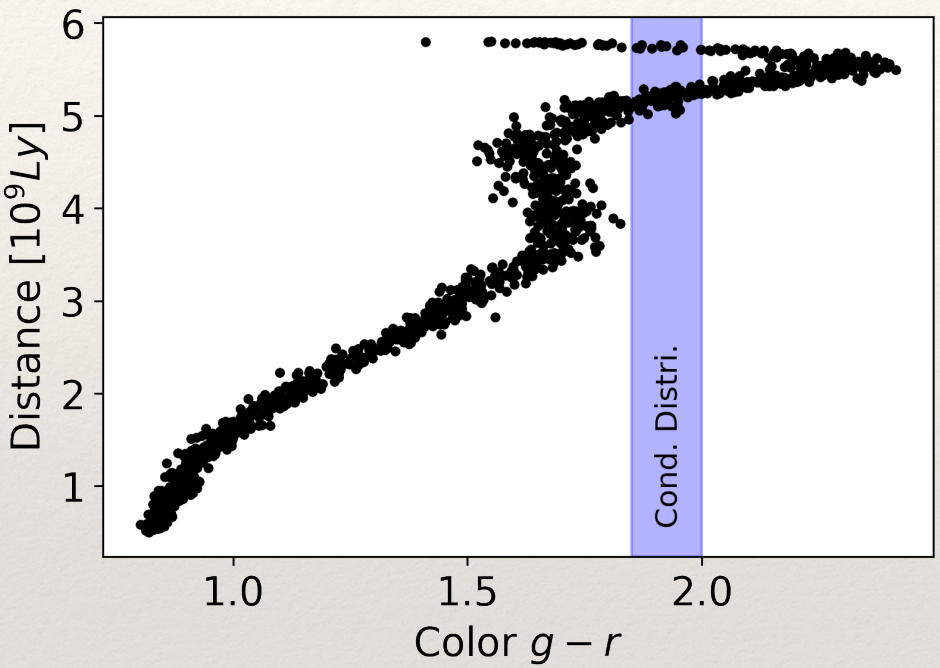
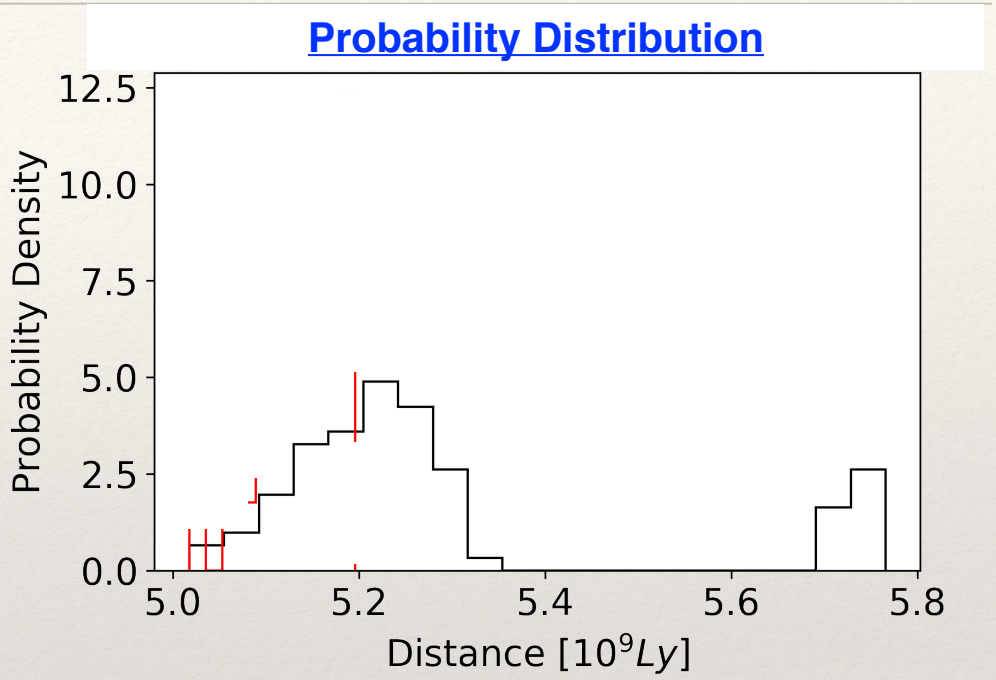


Image credit: Markus Rau

However, multiple widely different distances (redshift) can be consistent with the observed colors of a galaxy…





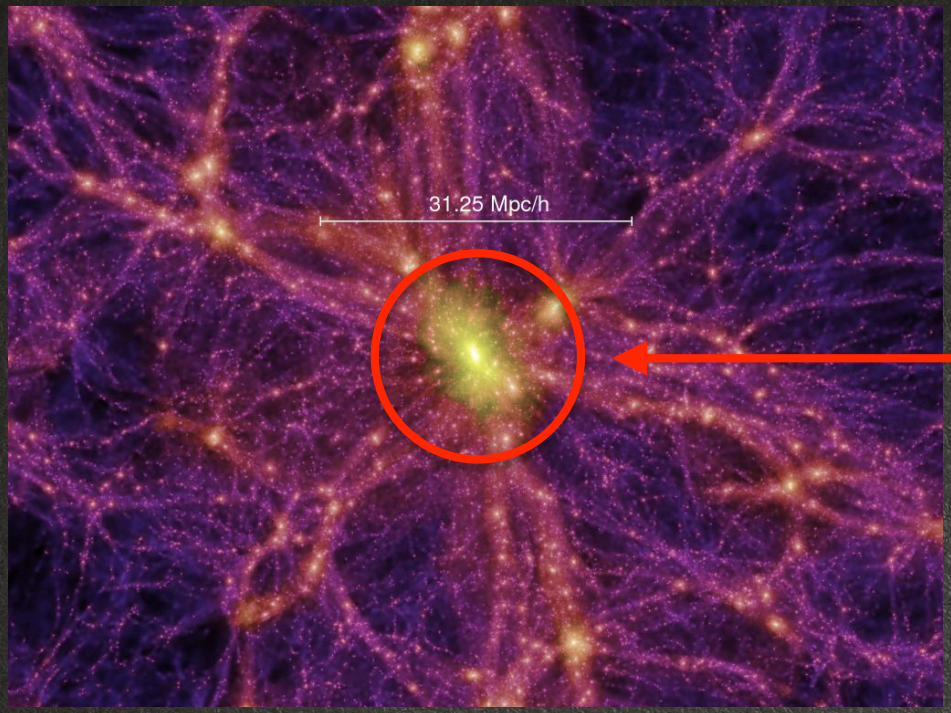
Hence, astronomers are more interested in algorithms that return a **probability distribution function** (PDF) over possible distances (given observed colors of a galaxy) than a "single best guess" of what the distance to that galaxy is.

Most galaxy image surveys provide catalogs of such PDFs

Image credit: Markus Rau

Examples of Things I've Worked On: Three Astrostatistics Problems

1. How do you measure the distance to a galaxy?
2. How do you weigh a galaxy or a galaxy cluster?
3. How do you constrain models of the Universe?

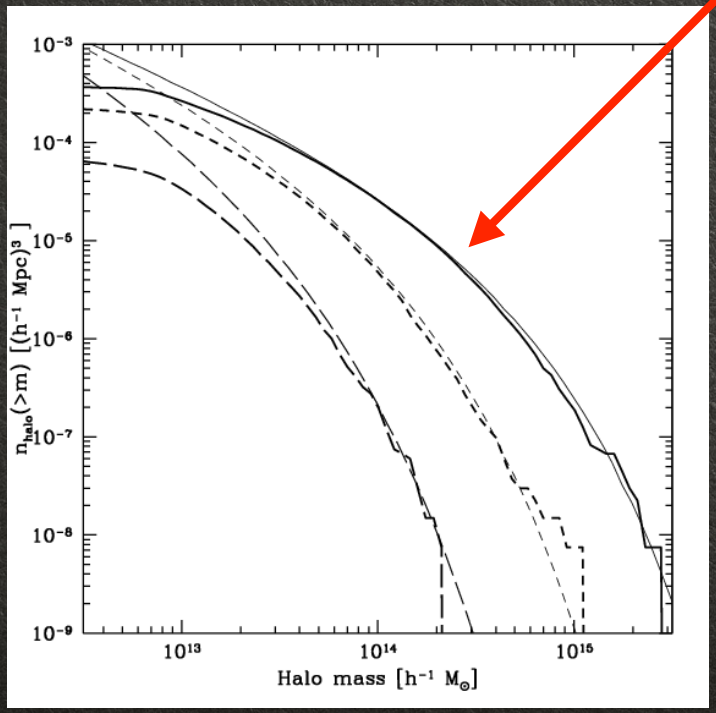


Recall:

1) Simulation output varies as a function of input theoretical parameters.

2) Simulations create a distribution of clusters that all have different masses.

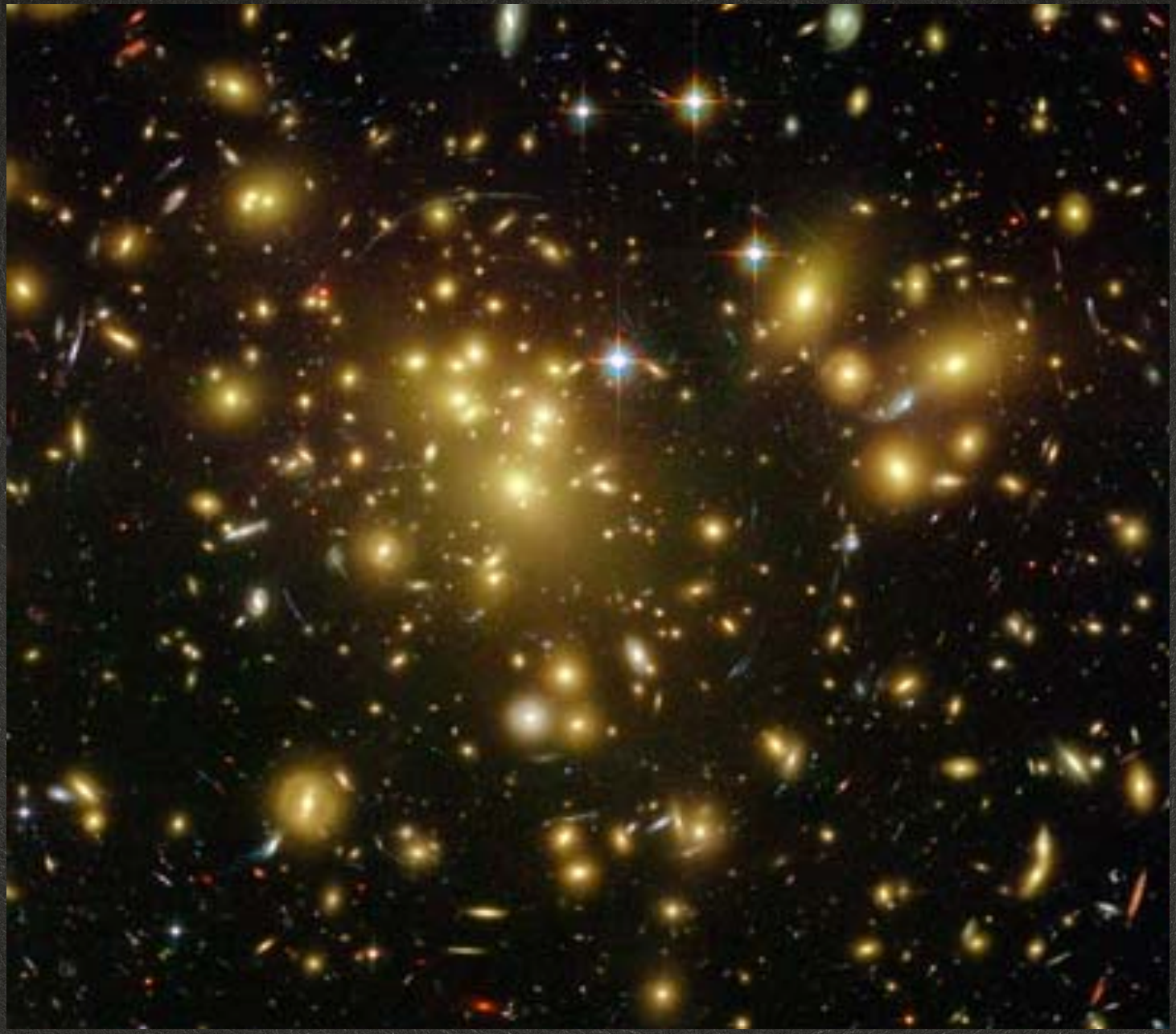
3) The observed distribution of ***real*** clusters and their masses allow us to constrain the theory that goes into the simulations.



http://inspirehep.net/record/854412/files/Hallman07f3.png

Credit: Peter Freeman

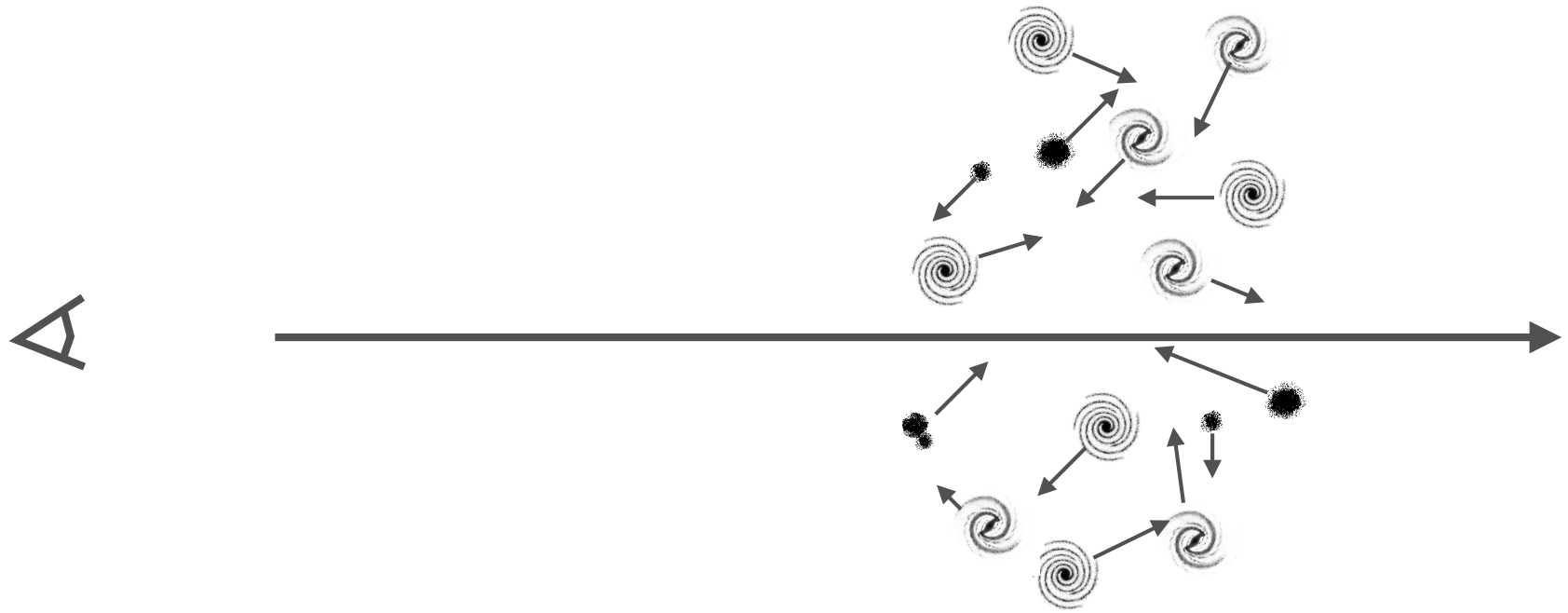
http://www.astronomynotes.com/galaxy/abell1689-gravlens.jpg



But…(small problem)…we cannot "weigh" real clusters.

Credit: Peter Freeman

Cluster galaxy velocities are affected by the cluster mass. Generally: the higher the mass, the greater ***spread*** in velocities. (But we only see the velocities projected onto a line-of-sight.)



Credit: Peter Freeman

CMU Summer Undergraduate Research (SURE) Project

The project: can you calibrate the relationship between cluster galaxy velocities and cluster mass?

A MACHINE LEARNING APPROACH FOR
DYNAMICAL MASS MEASUREMENTS OF GALAXY CLUSTERS

M. Ntampaka[1], H. Trac[1], D.J. Sutherland[2], N. Battaglia[1, 3], B. Póczos[2], J. Schneider[2]

Draft version October 6, 2014

ABSTRACT

We present a modern machine learning approach for cluster dynamical mass measurements that is a factor of two improvement over using a conventional scaling relation. Different methods are tested against a mock cluster catalog constructed using halos with mass $\geq 10^{14}$ $M_\odot h^{-1}$ from Multidark's publicly-available N-body MDPL halo catalog. In the conventional method, we use a standard $M(\sigma_v)$ power law scaling relation to infer cluster mass, M, from line-of-sight (LOS) galaxy velocity dispersion, σ_v. The resulting fractional mass error distribution is broad, with width $\Delta\epsilon \approx 0.86$ (68% scatter), and has extended high-error tails. The standard scaling relation can be simply enhanced by including higher-order moments of the LOS velocity distribution. Applying the kurtosis as a linear correction term to $\log(\sigma_v)$ reduces the width of the error distribution to $\Delta\epsilon \approx 0.74$ (15% improvement). Machine learning can be used to take full advantage of all the information in the velocity distribution. We employ the Support Distribution Machines (SDMs) algorithm that learns from distributions of data to predict single values. SDMs trained and tested on the distribution of LOS velocities yield $\Delta\epsilon \approx 0.41$ (52% improvement). Furthermore, the problematic tails of the mass error distribution are effectively eliminated.

Subject headings: cosmology: theory—dark matter—galaxies: clusters: general—galaxies: kinematics and dynamics—gravitation—large-scale structure of universe—methods: statistical

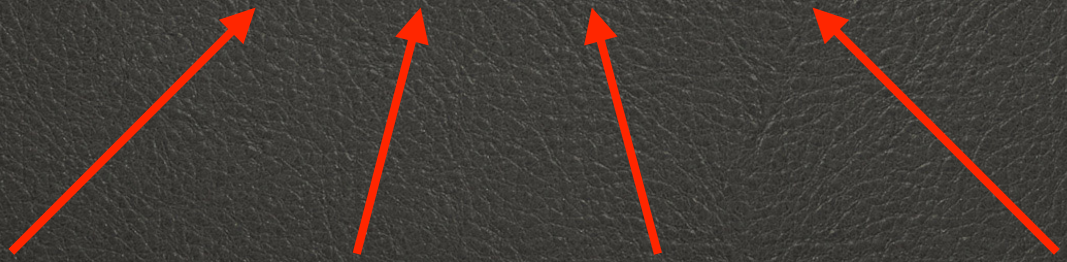
You will attempt this using data from the Multidark simulation supplied by Ntampaka et al.

Credit: Peter Freeman

CMU Summer Undergraduate Research (SURE) Project

```
84302646 0 14.0007 132.709
84302646 0 14.0007 72.7086
84302646 0 14.0007 -52.6014
84302646 0 14.0007 -542.311
84302646 0 14.0007 -37.1414
84302646 0 14.0007 -395.691
84302646 0 14.0007 -24.5914
84302646 1 14.0007 109.843
84302646 1 14.0007 765.153
84302646 1 14.0007 -89.0867
84302646 1 14.0007 -431.837
84302646 1 14.0007 577.253
84302646 1 14.0007 202.013
```

Example of the data



simulated cluster ID

line-of-sight ID

cluster mass (in solar masses, log base 10)

projected velocity (km/s)

Credit: Peter Freeman

CMU Summer Undergraduate Research (SURE) Project

Regression analysis example:

Summarize ($v_1, v_2, \ldots, v_n$) along each line-of-sight and in each cluster via sample standard deviation…

$$s = \frac{1}{n-1} \Sigma_{i=1}^{n} (v_i - \bar{v})^2$$

…to build up k pairs (M,s). Then regress s onto M…

Credit: Peter Freeman

CMU Summer Undergraduate Research (SURE) Project

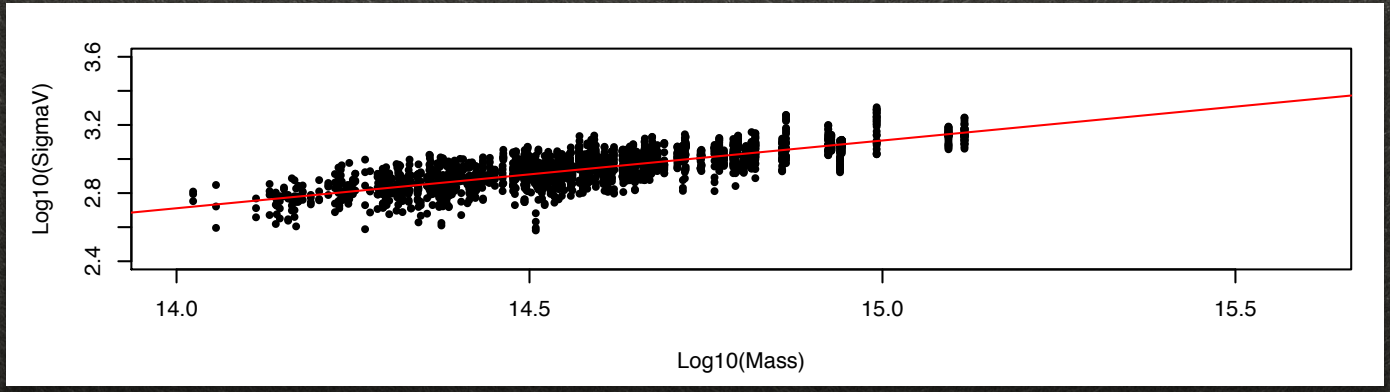
Regression analysis example:

Summarize $(v_1, v_2, \ldots, v_n)$ along each line-of-sight and in each cluster via sample standard deviation…

$$s = \frac{1}{n-1} \Sigma_{i=1}^{n} (v_i - \bar{v})^2$$

…to build up k pairs (M,s). Then regress s onto M…

$$\log_{10} s = \beta_0 + \beta_1 \log_{10} M$$



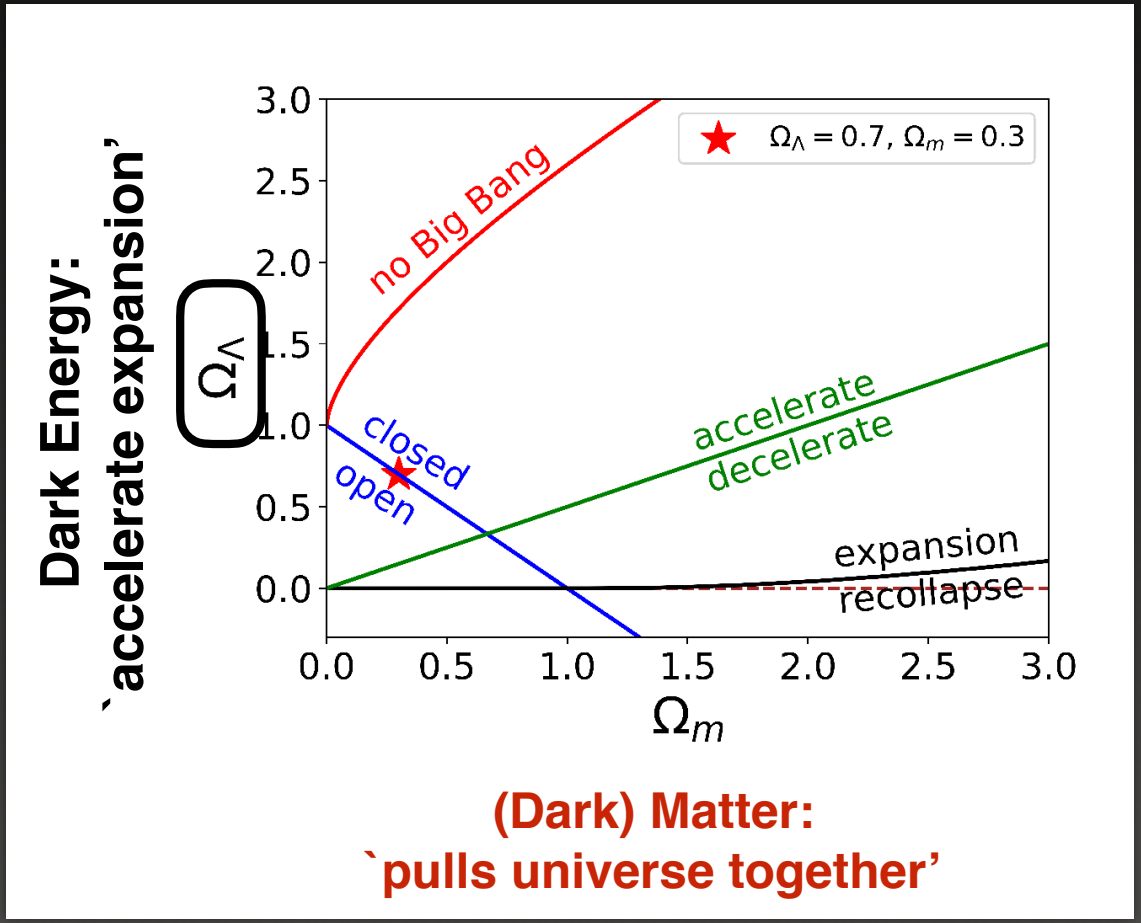
Credit: Peter Freeman

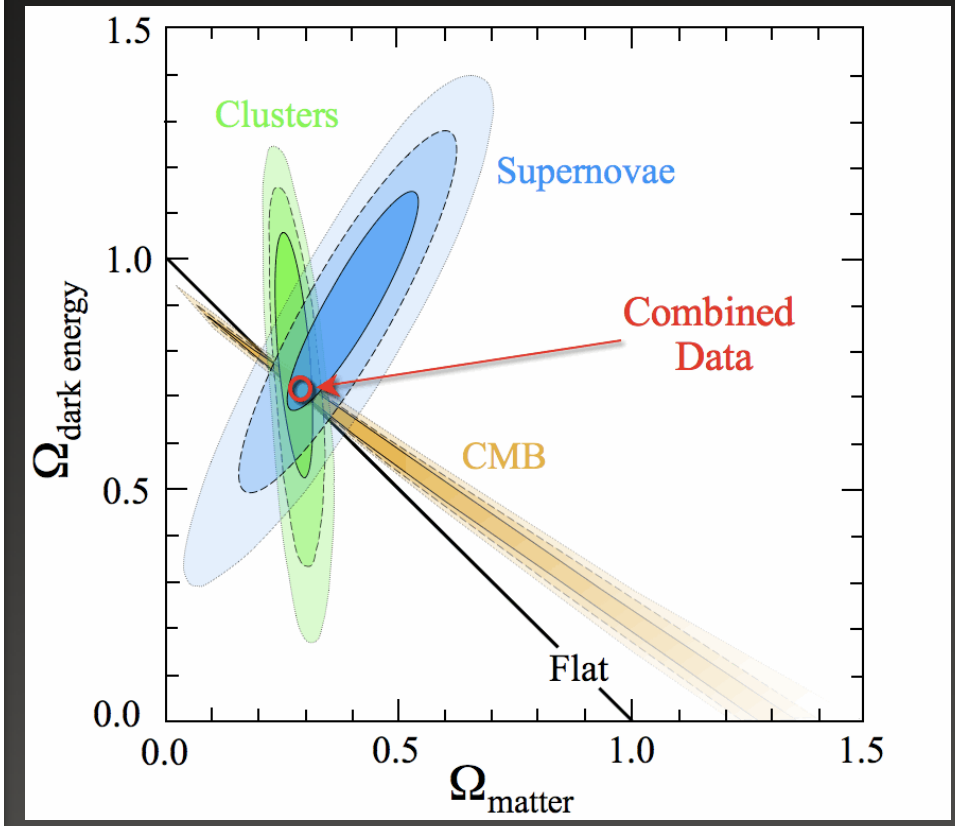
Examples of Things I've Worked On: Three Astrostatistics Problems

1. How do you measure the distance to a galaxy?
2. How do you weigh a galaxy or a galaxy cluster?
3. How do you constrain models of the Universe?

Recall: The Lambda-CDM Model is a Parametrization of Big Bang Cosmology

- These parameters can be constrained by observations.





Source: https://www.hep.ucl.ac.uk/darkMatter/

But we still have much to do…

Cosmological parameters

$$\Omega_m, \Omega_b, \Omega_{\rm r}, \Omega_\Lambda \quad w_0, w_a \quad H_0 \quad (A_s/\sigma_8), n_s$$

Components of the universe
Dark Matter
Baryonic Matter
Radiation (`Light')
Dark Energy

Equation of state of Dark Energy

Speed of expansion

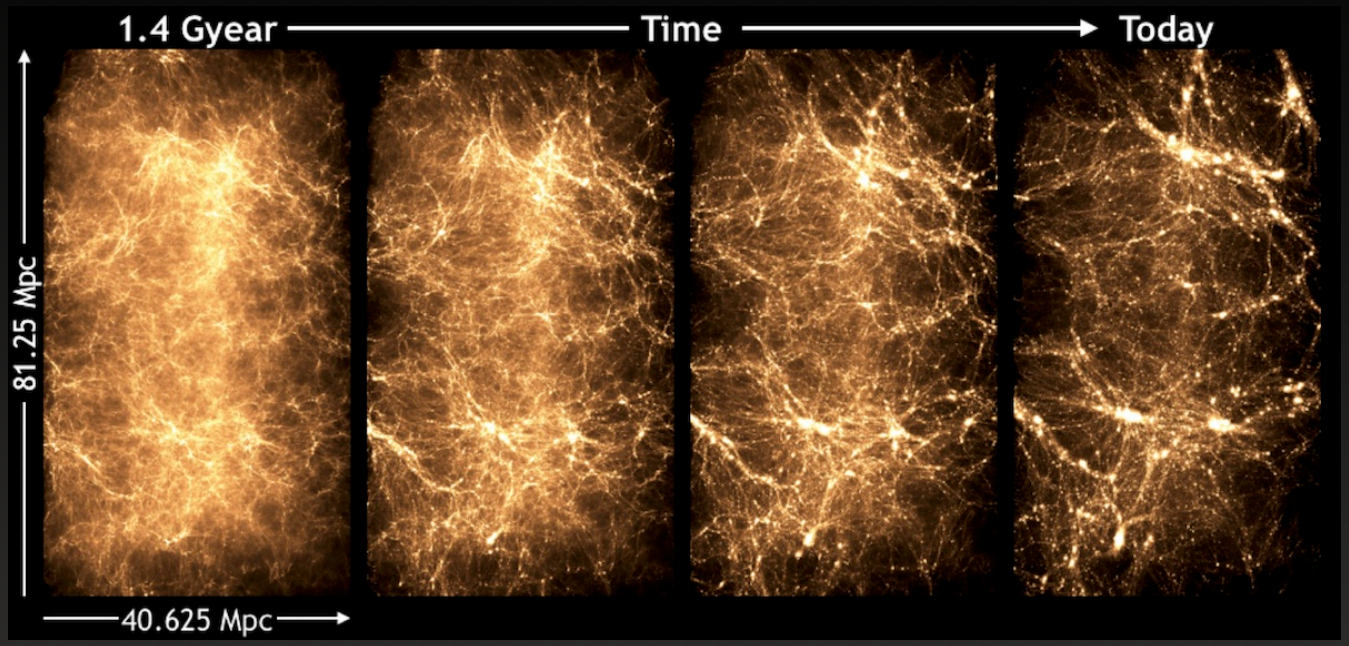
`Initial conditions'
Early Universe
Inflation

Image credit: Markus Rau

- There are ~10 key cosmological parameters. With higher precision data (from next-generation surveys like LSST), we need more advanced statistical techniques than what is currently being used to jointly constrain these key parameters.

Constraining Models of the Universe

Theory In (Simulation with Tunable Parameters)

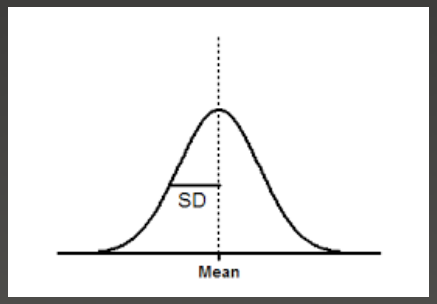


Structures Out

Observed Data

https://devblogs.nvidia.com/gpu-accelerated-cosmological-analysis-titan-supercomputer/

- The so-called likelihood function L(x; θ) connects underlying parameters of interest θ with observable data x. If you have the likelihood, you can estimate θ once you have measured x.



- Current state-of-the art in cosmology analysis is to assume a Gaussian likelihood; the mean and (co)variance of the Gaussian are fit using simulated data.

Can We Improve on Our Statistical Model?

- We still have a lot to do in terms of surveys…
- But it is believed that at the level of precision of future surveys (like LSST), assumptions made in current analyses will become questionable.

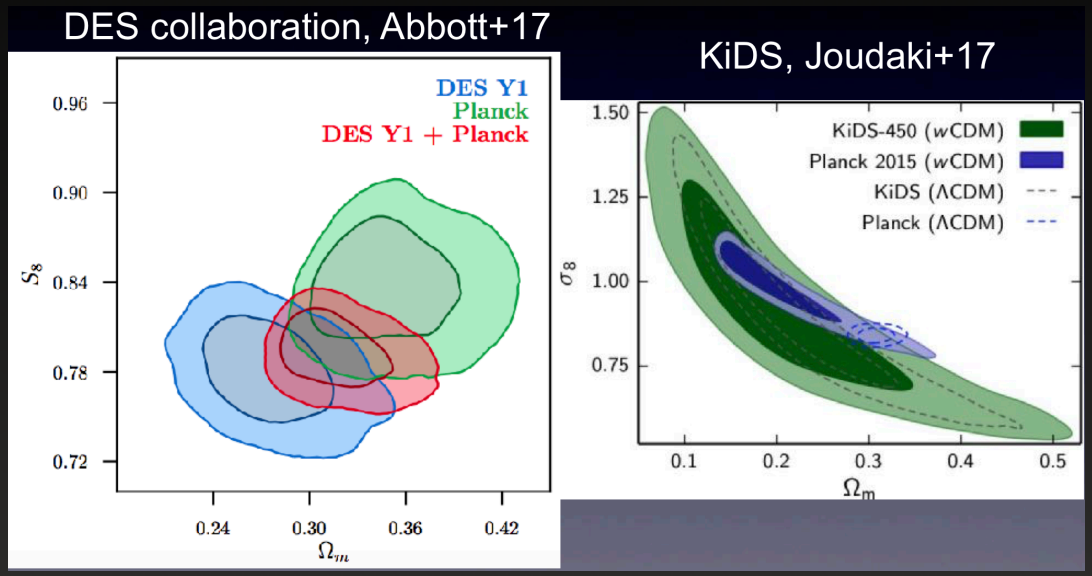
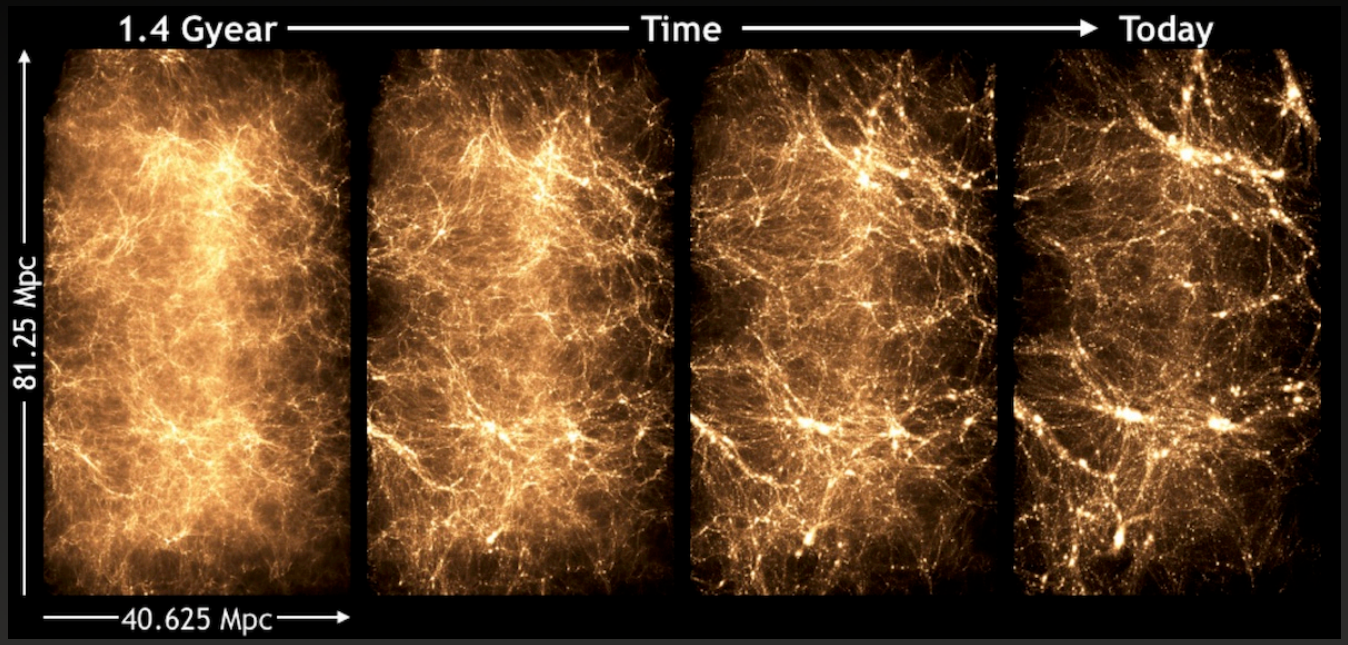


Image credit: Rachel Mandelbaum (Cosmo21)

- Question: Can we use simulations to build a better statistical model $L(x;\theta)$ for the relationship between parameters of interest and observable data? How about not assuming a particular analytic form for the likelihood and instead just repeatedly simulate data under different parameter settings?

Basic Idea of "Likelihood-Free Inference"

Theory In (Simulation with Tunable Parameters)



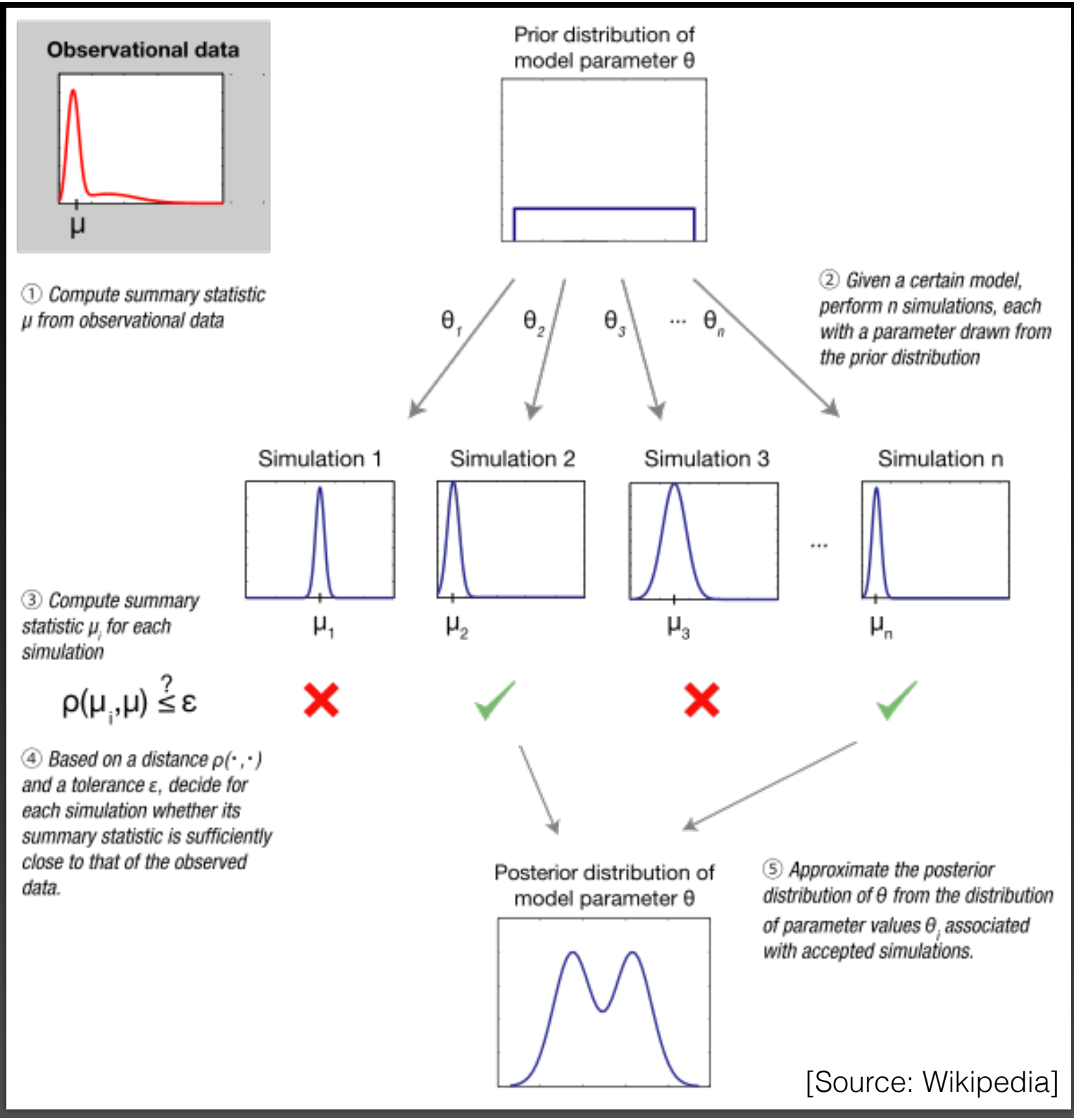
Structures Out

↕

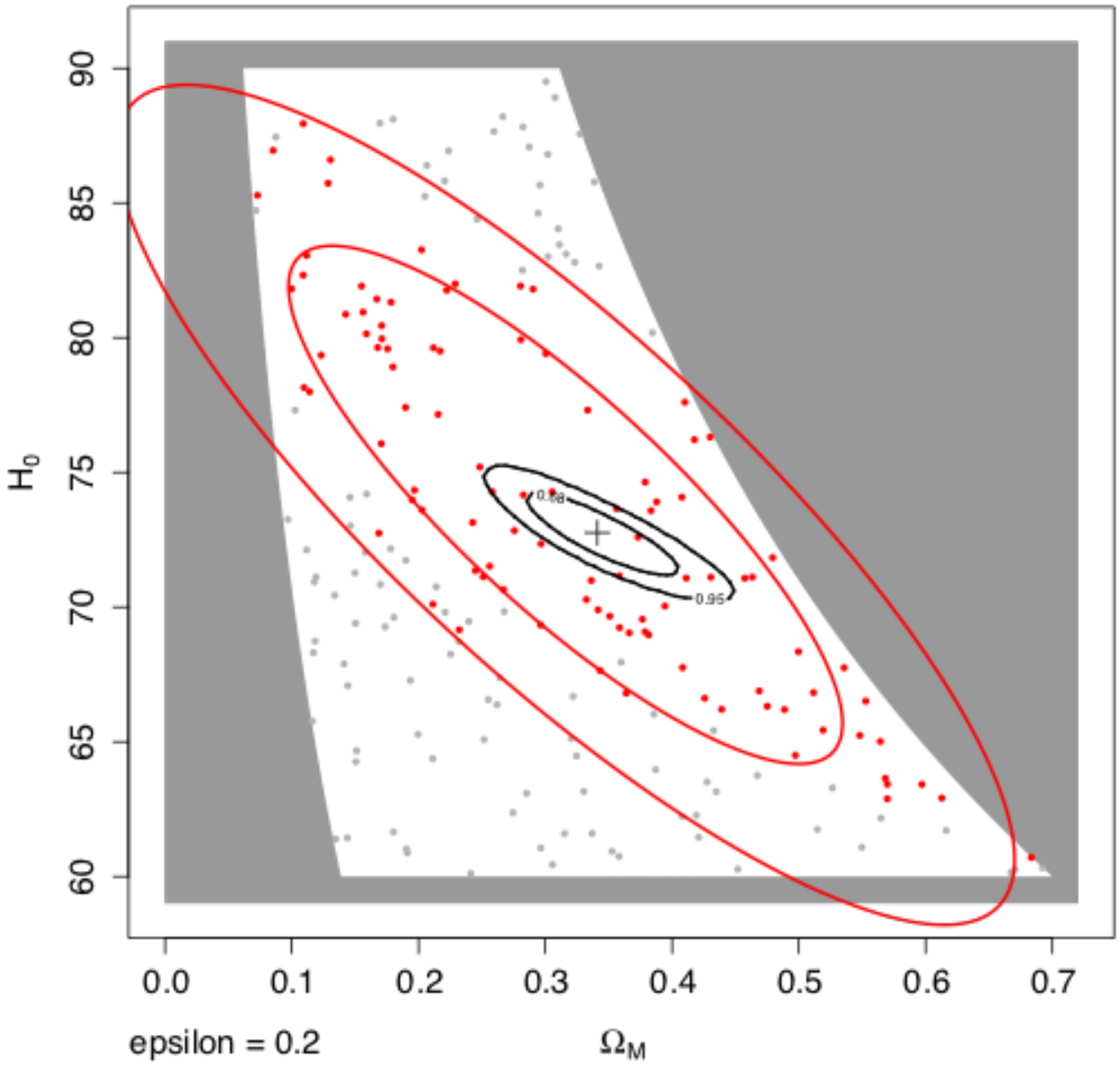
Observed Data

https://devblogs.nvidia.com/gpu-accelerated-cosmological-analysis-titan-supercomputer/

1. Forward-simulate observable data under different parameter settings.
2. Compare the output with actually observed data.
3. Let the parameters consistent with observed data define a "plausible" distribution of parameters.

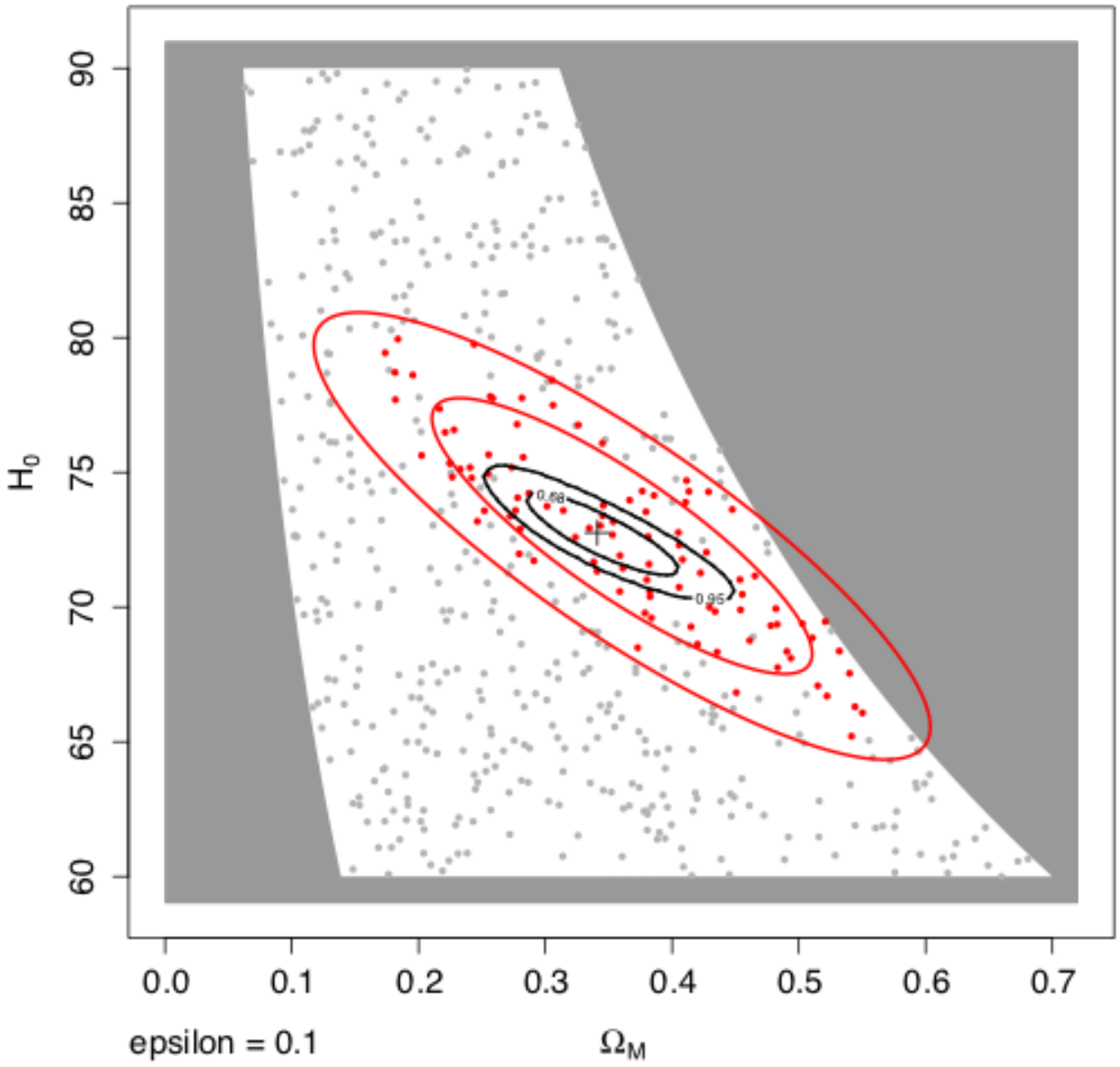


[Source: Wikipedia]



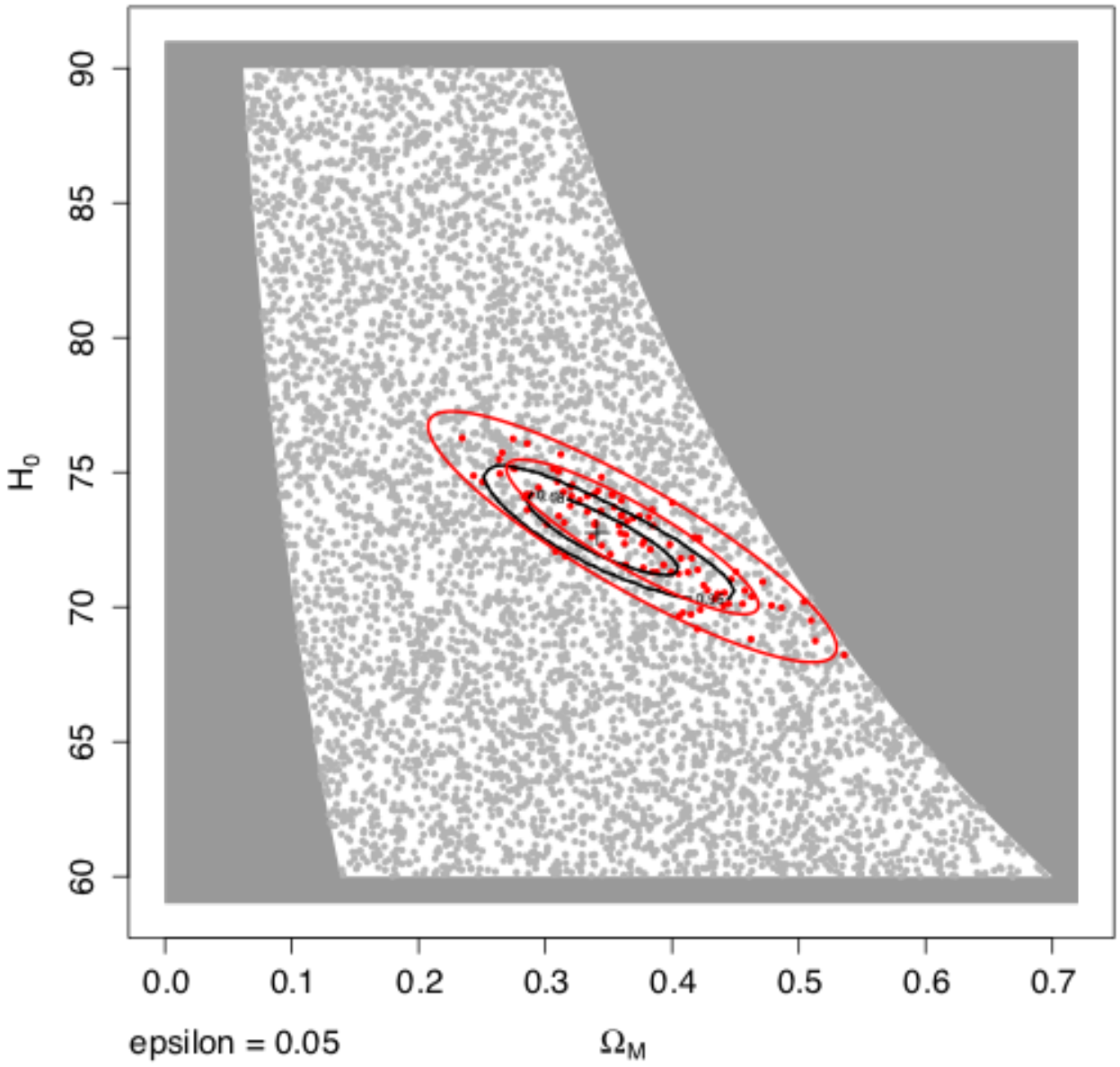
Basic rejection approach applied to SNe data

[Credit: Chad Schafer (Weyant/Schafer/Wood-Vasey 2013)]



Basic rejection approach applied to SNe data

[Credit: Chad Schafer]



Basic rejection approach applied to SNe data

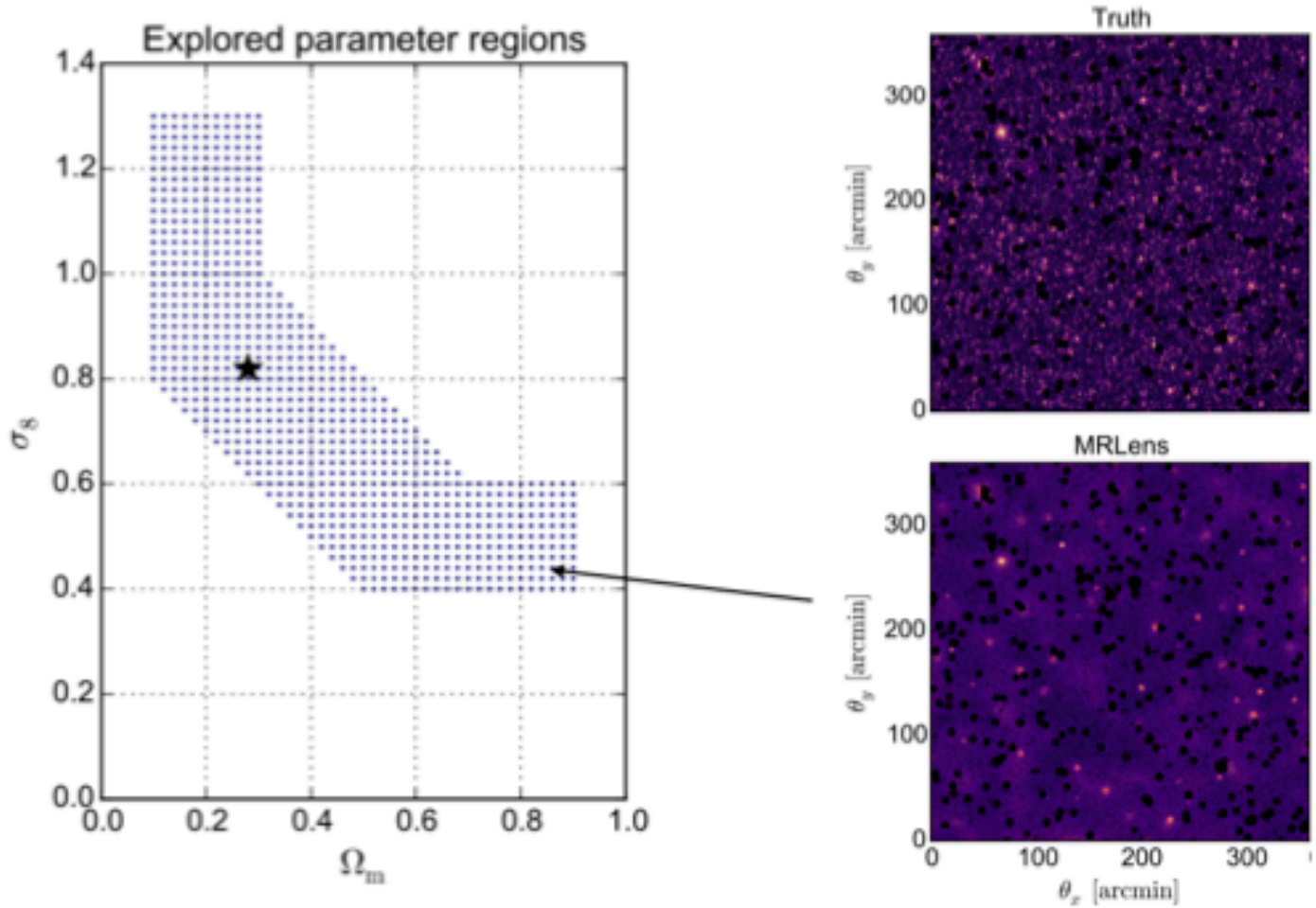
[Credit: Chad Schafer]

Statistical Challenges in Likelihood-Free Inference: What We are Working on Now

- Cosmological simulations are often very slow; common practice to fit faster "emulators". How do you calibrate and validate these emulators?
- How do you compare distributions/populations of simulated and observed data, or simulated and "emulated" data?
  - This is a statistical problem as you can't compare individual images.

Forward Simulators Example - CAMELUS Simulator [1]

In cosmology, weak lensing data simulations can be used to provide constraints on parameters of the ΛCDM model.



[1] Credits: Lin et al., 2018

Statistical Tools for Comparing and Analyzing Distributions of Images

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

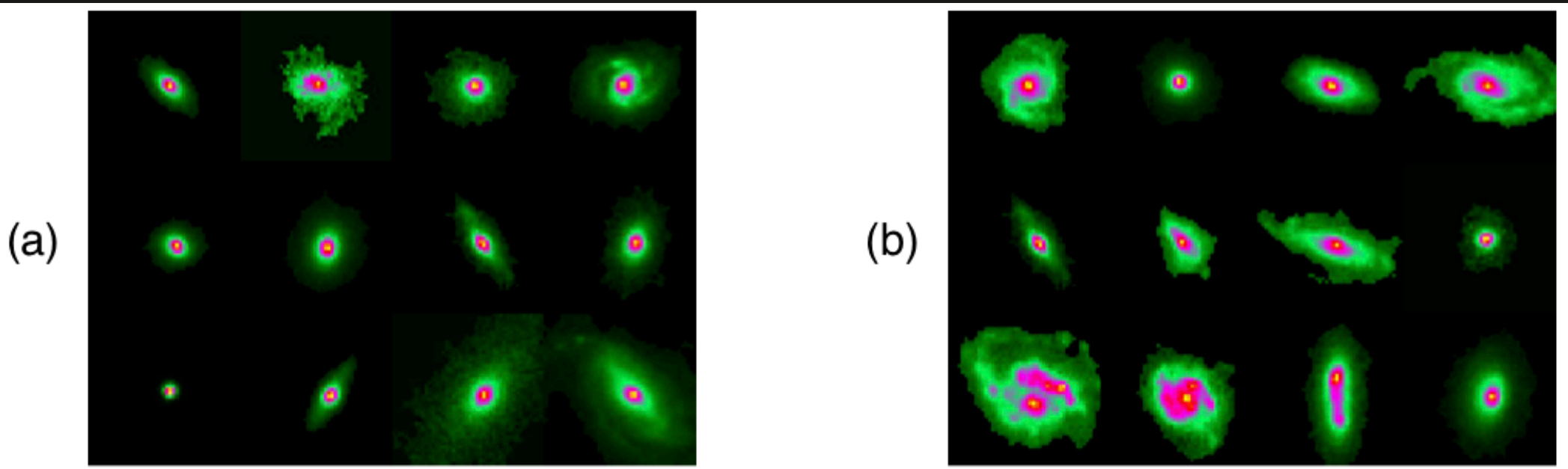


Figure 7: Examples of galaxies from (a) the low-SFR sample $\mathcal{S}_0$ versus (b) the high-SFR sample $\mathcal{S}_1$.

- 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]

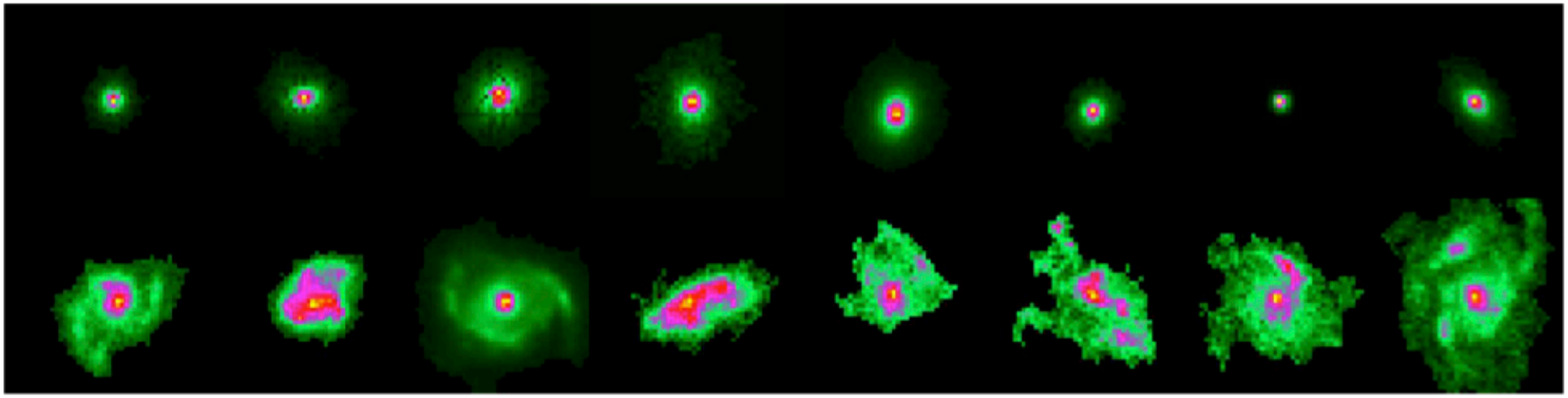


Figure 8: Galaxies in the test set with the highest significant difference $|\widehat{m}(\mathbf{x}) > \widehat{\pi}_1|$ according to our local test in feature space. (a) Galaxies that are more representative of the low-SFR sample $\mathcal{S}_0$, and (b) galaxies more representative of the high-SFR 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

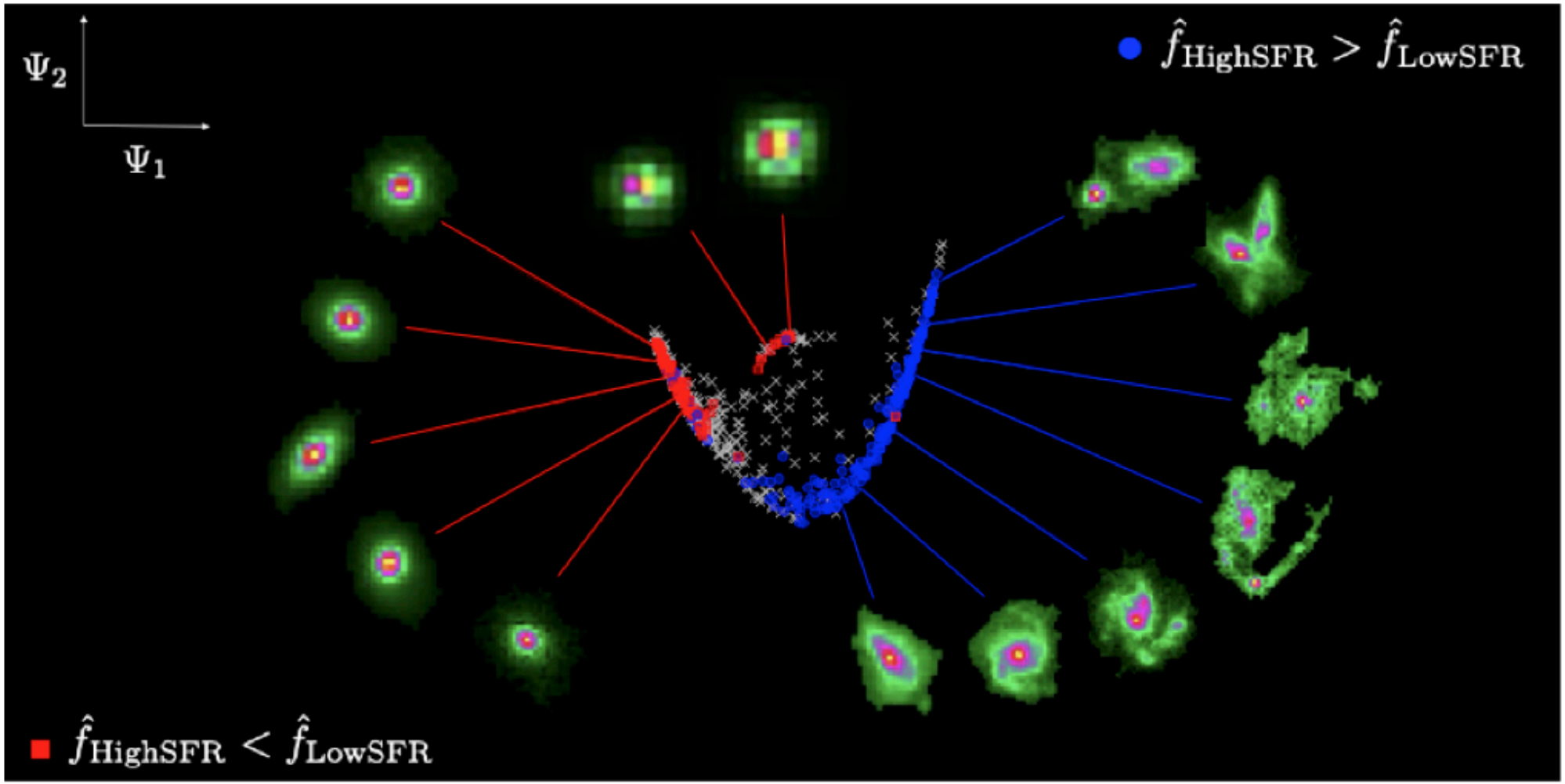


Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR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

Recap: Examples of Astrostatistics Problems

- How do you measure the distance to a galaxy?
  - By photometric redshift estimation
- How do you weigh a galaxy or a galaxy cluster?
  - By dynamical velocity measurements
- How do you constrain models of the Universe?
  - By comparing output of simulation/theoretical models to actual observed data

Acknowledgements


Jeff Newman


Chien-Ah Lin


Rachel Mandelbaum


Markus Rau


Dorit Hammerling


Kim Wood