Astrostatistical Analysis in Solar and Stellar Physics

David C. Stenning

Department of Statistics, University of California, Irvine

Outline

Outline

The Sun

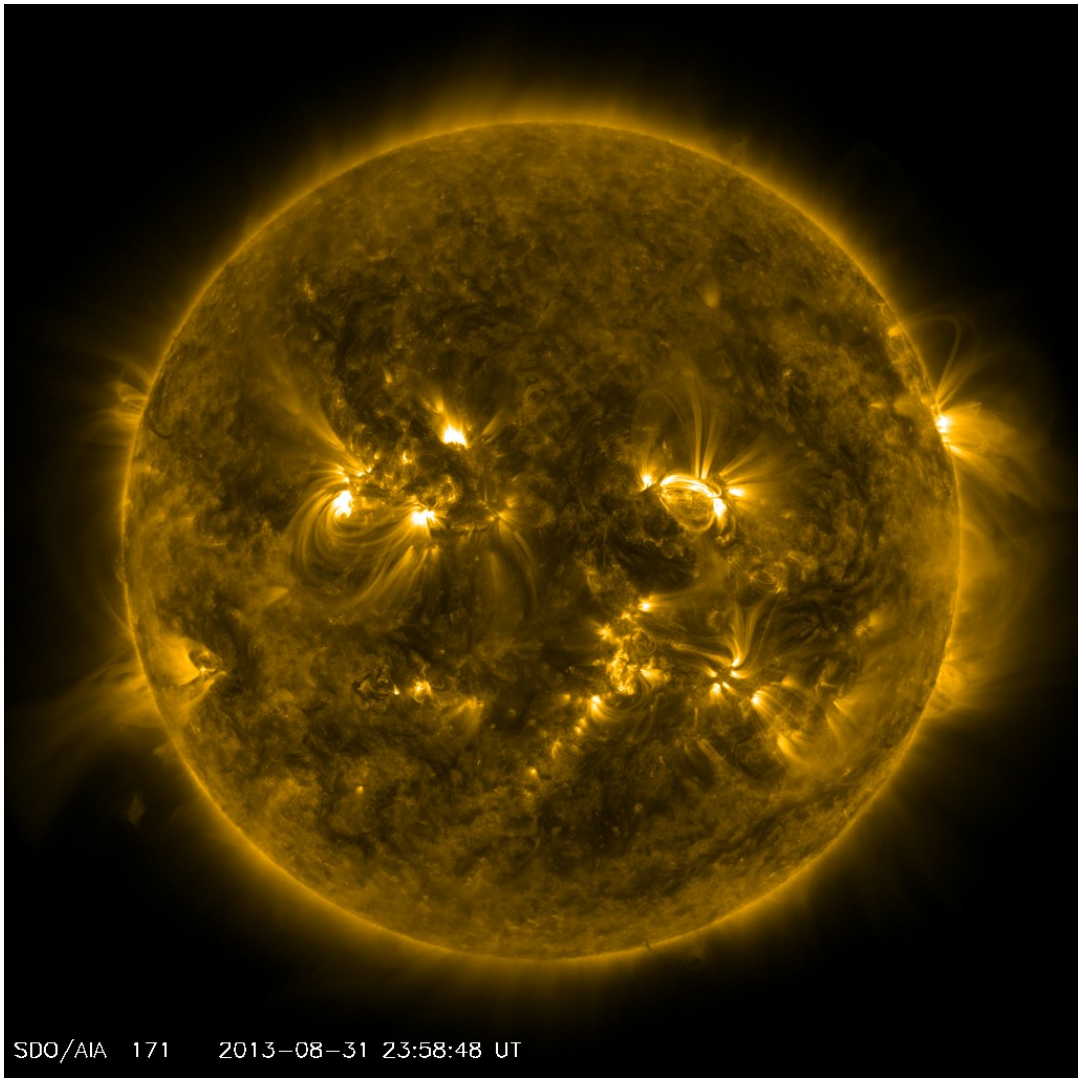


Image Credit: NASA/SDO

- The Sun in extreme ultraviolet light.
- Brigher regions correspond to higher temperature plasma.
- *Coronal loops* trace out the Sun's magnetic field.
- Complex magnetic field configurations are related to volatile events:
  - *solar flares.*
  - *coronal mass ejections.*

Sunspots

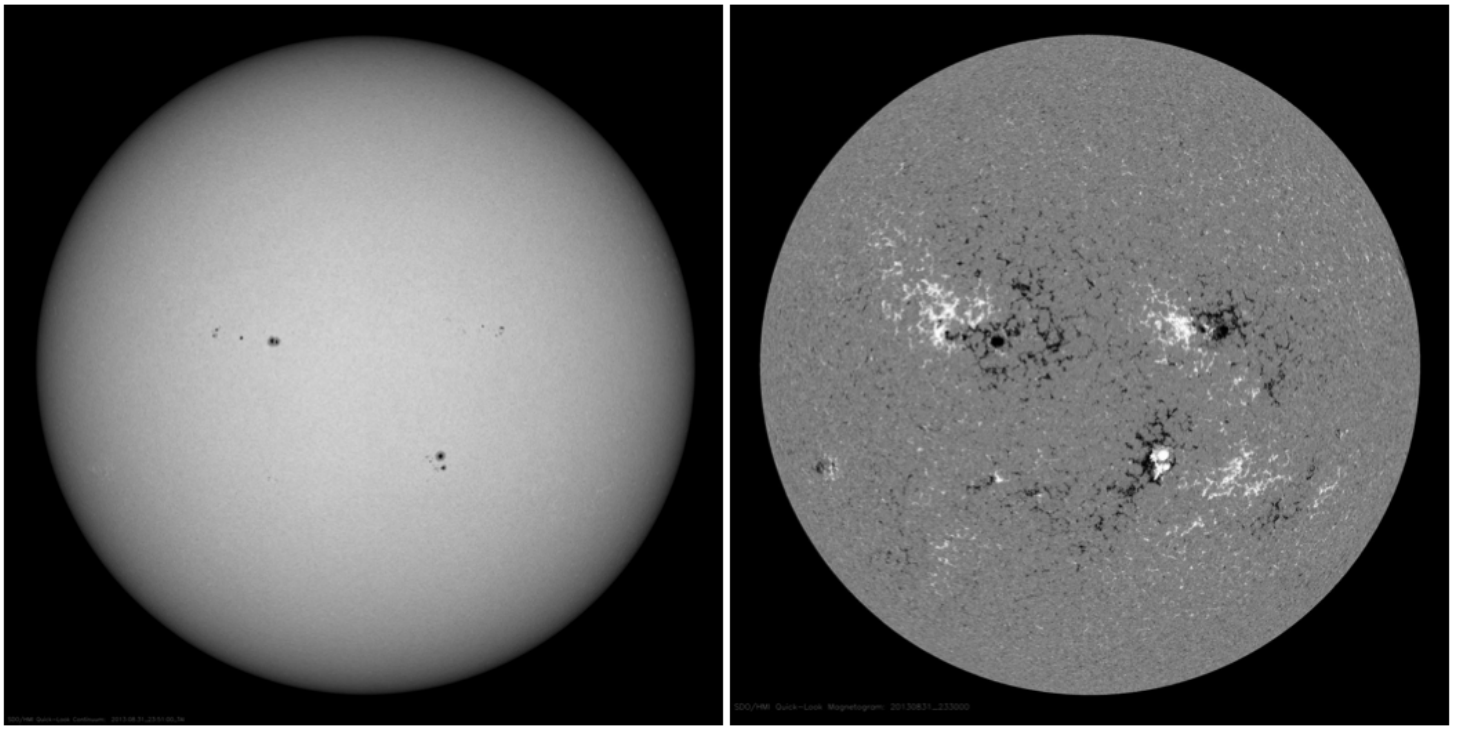
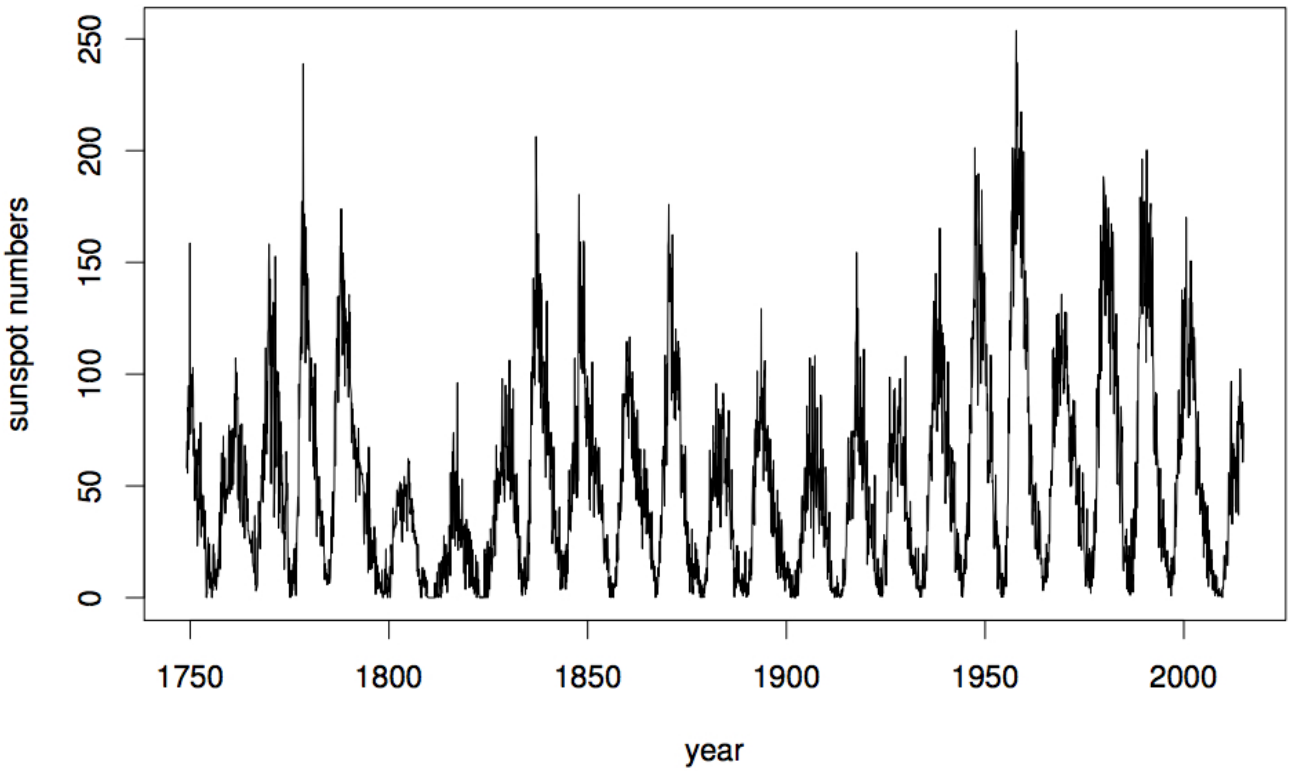


Image Credit: NASA/SDO

- *Sunspots* form when intense magnetic fields inhibit convection.
  - Show up as dark spots on the Sun's photosphere in *white-light images* (left image).
  - Classified based on the complexity of magnetic flux distribution as seen in *magnetograms* (right image).

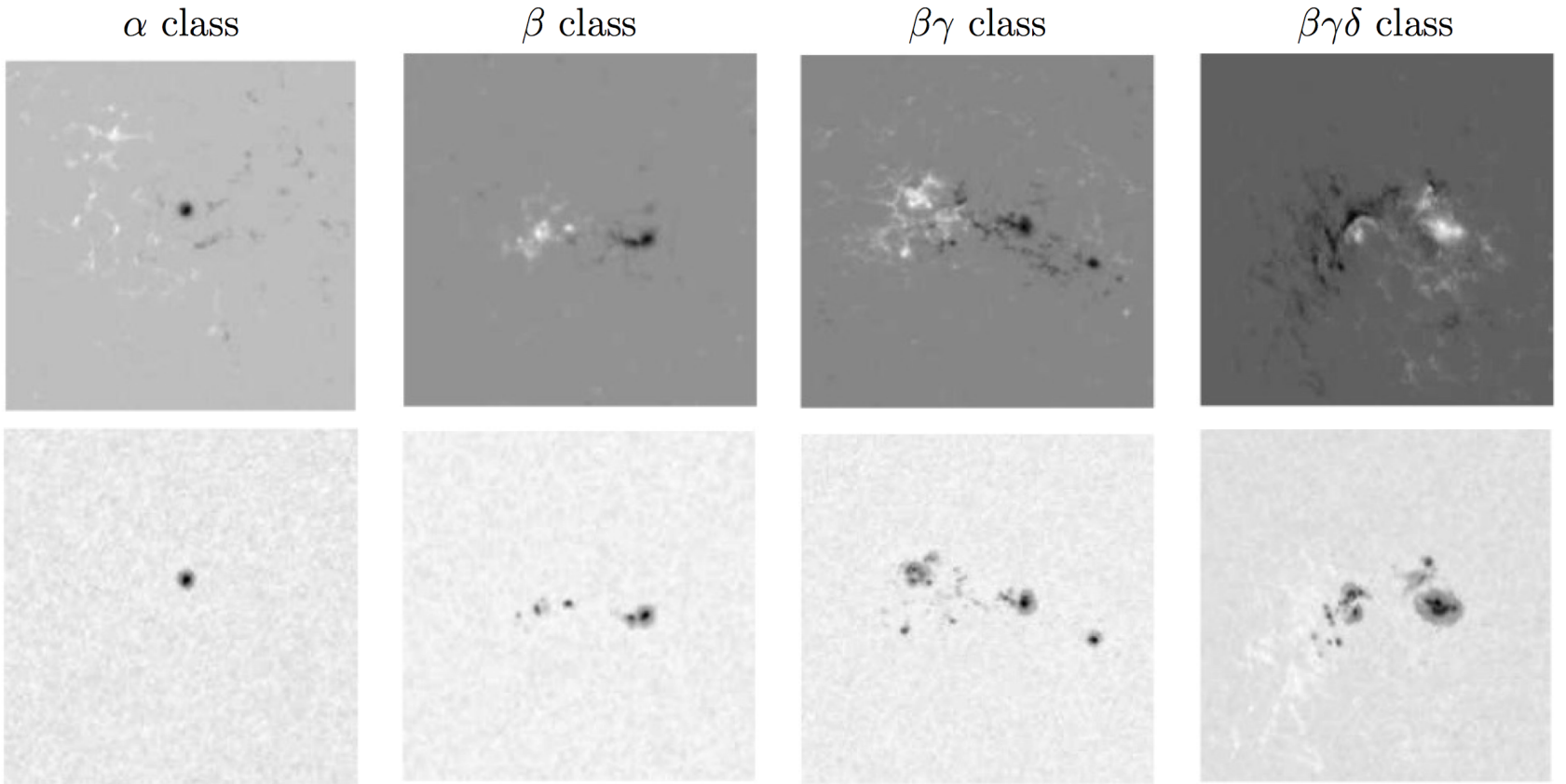
Solar Activity Cycle



- Solar activity follows a roughly 11-year *solar cycle*.
- First noticed in the time series of sunspot numbers (SSNs).
- Other proxies of solar activity (e.g., 10.7cm flux) follow the same underlying cycle.

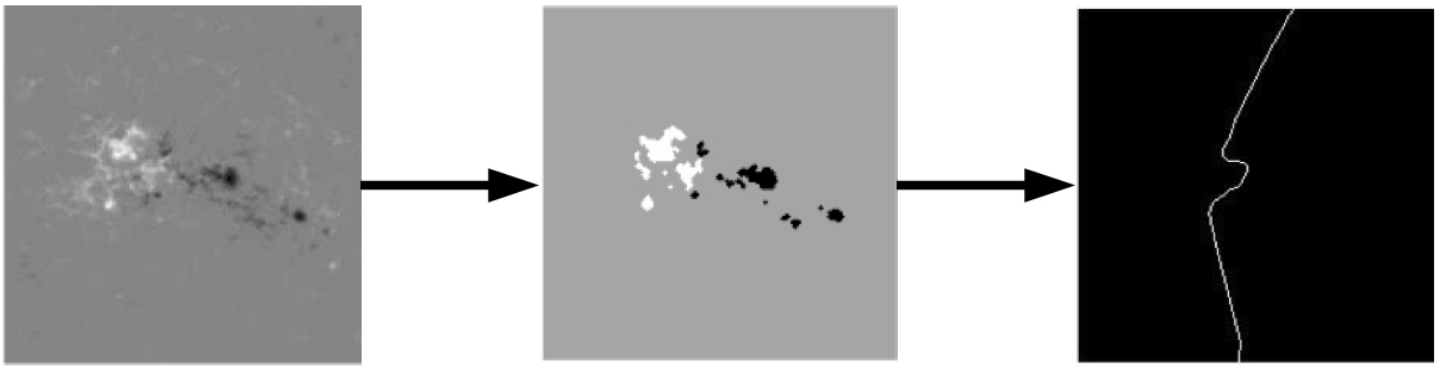
Outline

Mount Wilson Classification



Four broad classes— α, β, $\beta\gamma$, and $\beta\gamma\delta$—based on the complexity of magnetic flux distribution. *Top row:* magnetograms. *Bottom row:* white-light images.

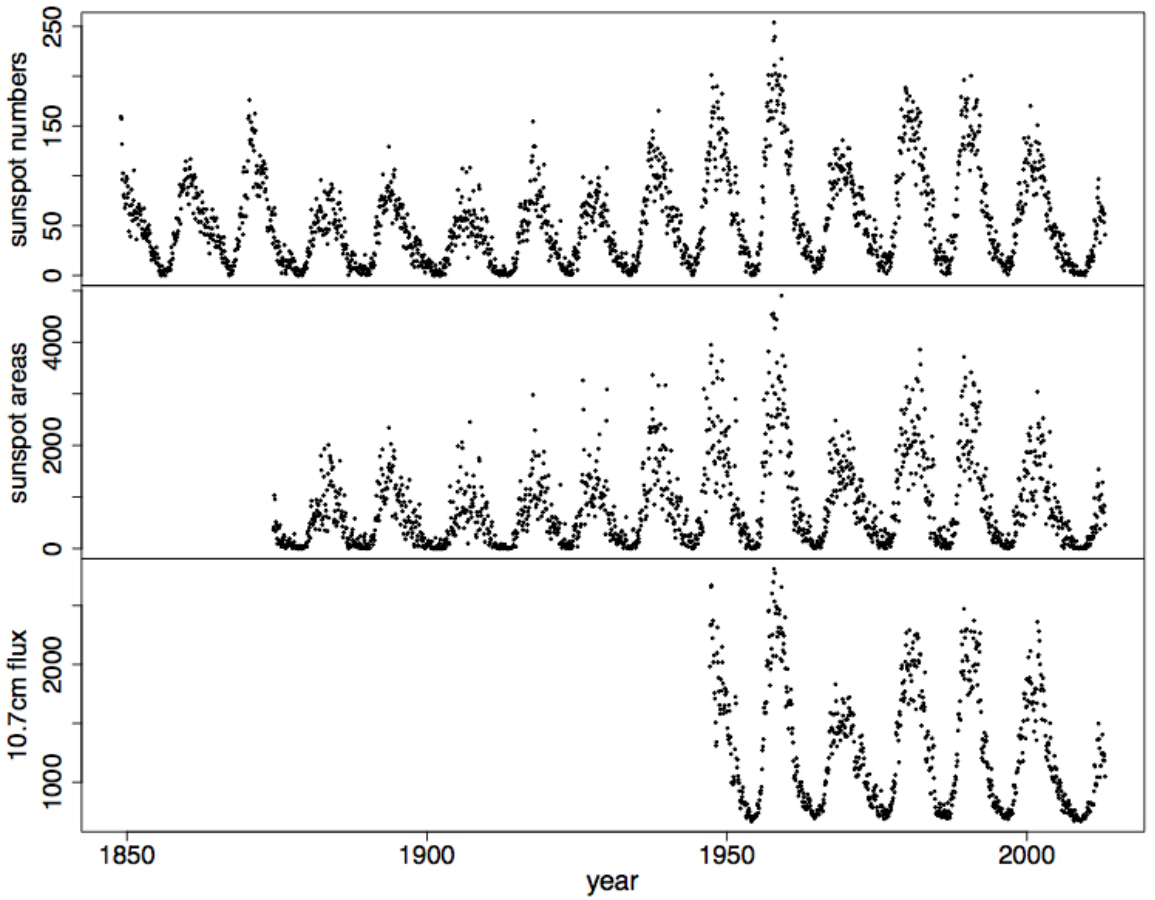
Automatic Classification of Sunspots



- Classification is predictive of solar activity (e.g., solar flares)
- Use Mt. Wilson rules to guide feature selection $\rightarrow$ **science-driven feature extraction**
  - Physically meaningful and interpretable features
- Features from mathematical morphology (e.g., curvature of polarity separating line, amount of scatter per polarity, etc.)
- Encapsulate relevant information in more informative manner than manual classification
- Amenable to statistical analyses: model sunspot evolution

Outline

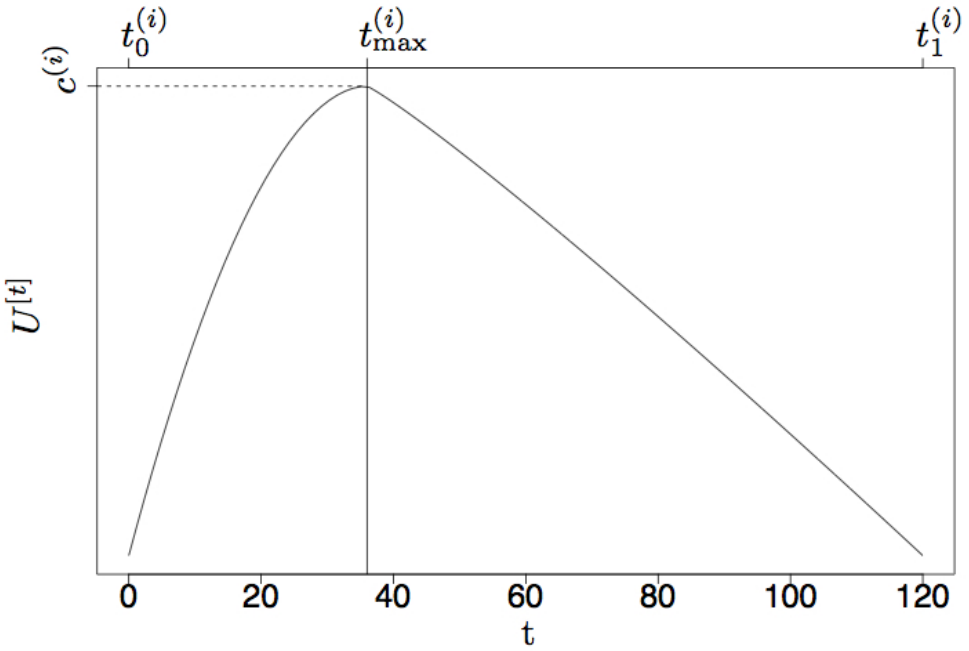
Solar Activity Proxies



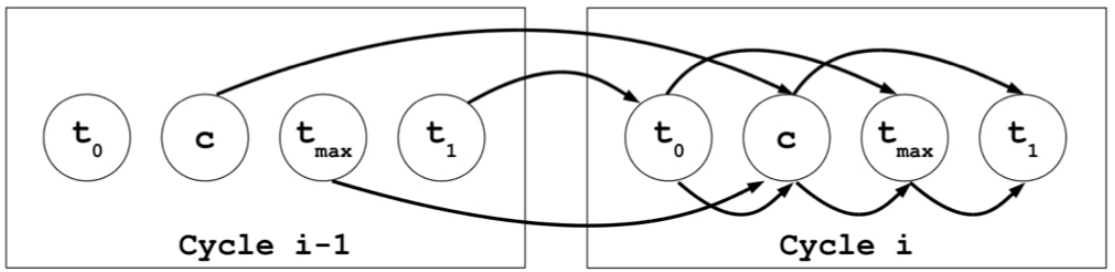
Goal: Model the underlying solar cycle with recently available proxies, while also taking advantage of the long history of SSNs.

Multilevel Model for the Solar Cycle (Yu et al., 2012)

Level One: Modeling cycle i with SSNs



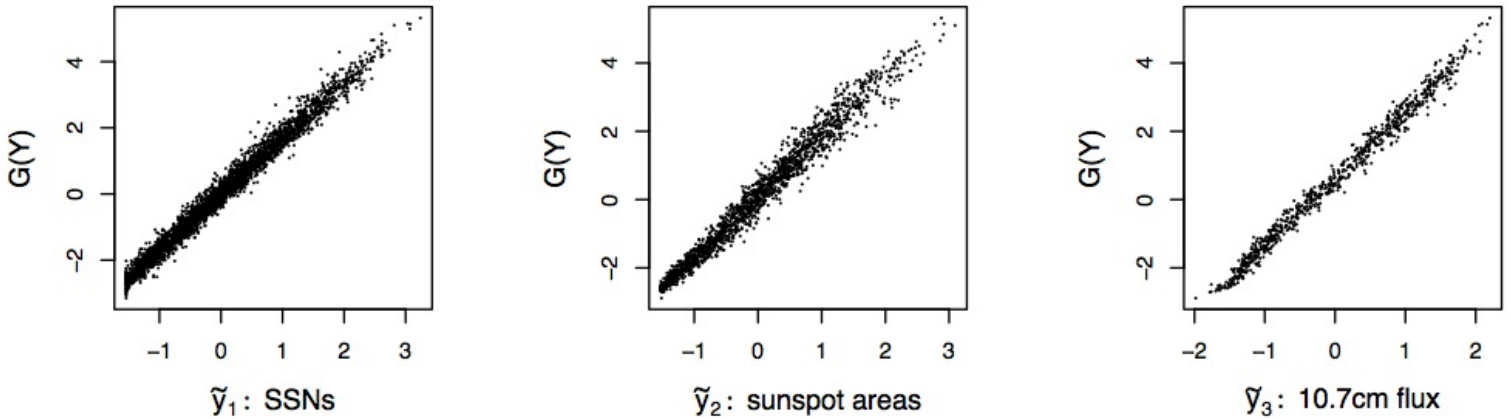
Level Two: Relationships between consecutive cycles



For cycle i:

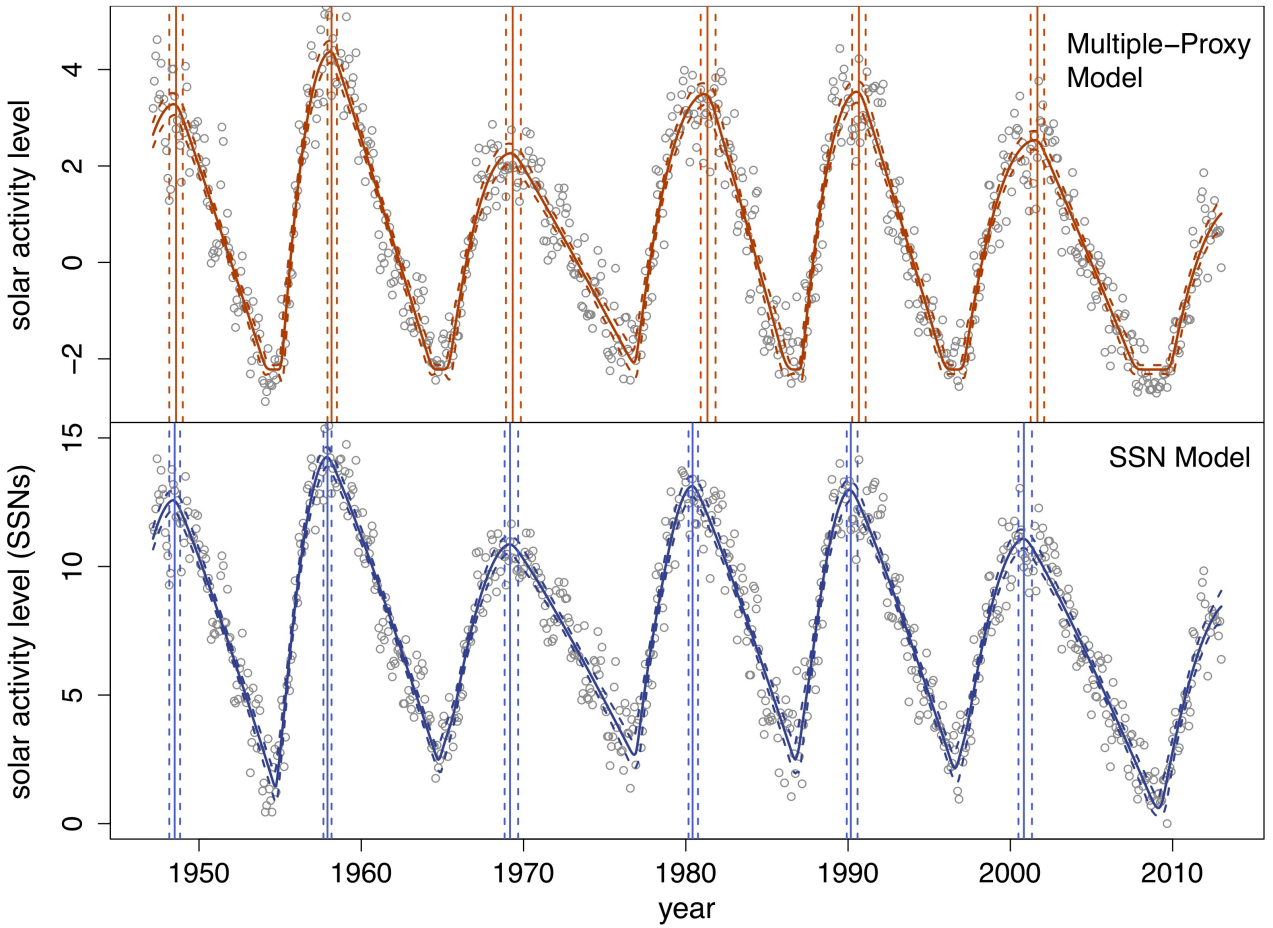
- $t_0^{(i)}$ is the *start time*.
- $t_{\max}^{(i)}$ is the *time of cycle maximum*.
- $t_1^{(i)}$ is the *end time*.
- $c^{(i)}$ is the *amplitude*.
- $U^{[t]}$ is the "average solar activity level" at time t.

Incorporating Multiple Proxies



- Proxies exhibit a *monotone missing data pattern*
  - Specify a simple local missing data model and use multiple imputation
- PCA to project the multivariate time-series data onto the $1-D$ manifold defined by the direction of maximum variance
  - Transformations help stabalize variances and improve linearity
- Fit the solar cycle model with multiple proxies and with the SSNs alone to allow for comparison

The Fitted Solar Cycle



Solid curves are fitted values for the solar activity level and dashed curves their 95% intervals. Solid and dashed vertical lines correspond to fitted values and 95% intervals for the time of solar maximum.

Outline

Life and Death of a Star

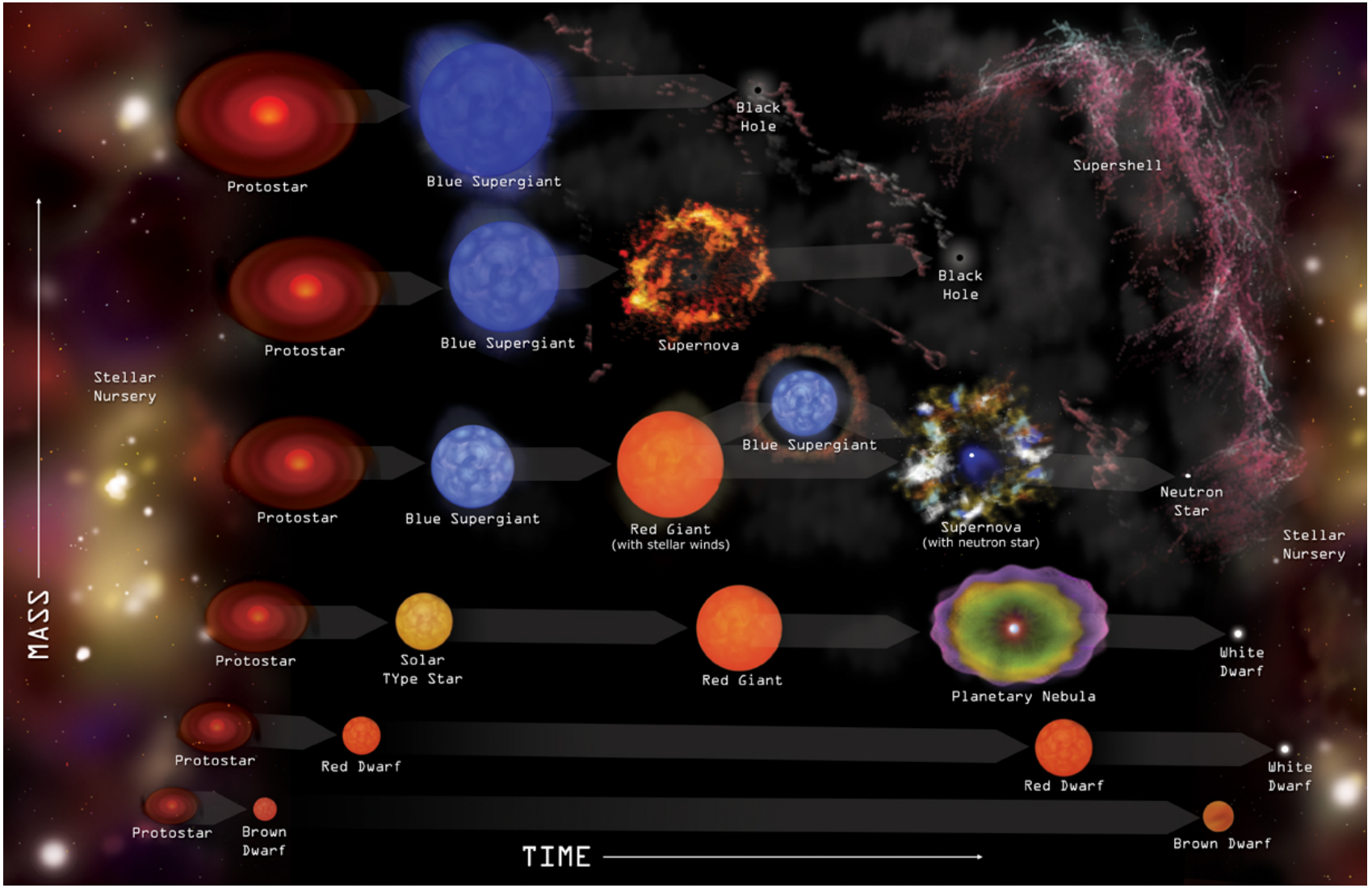
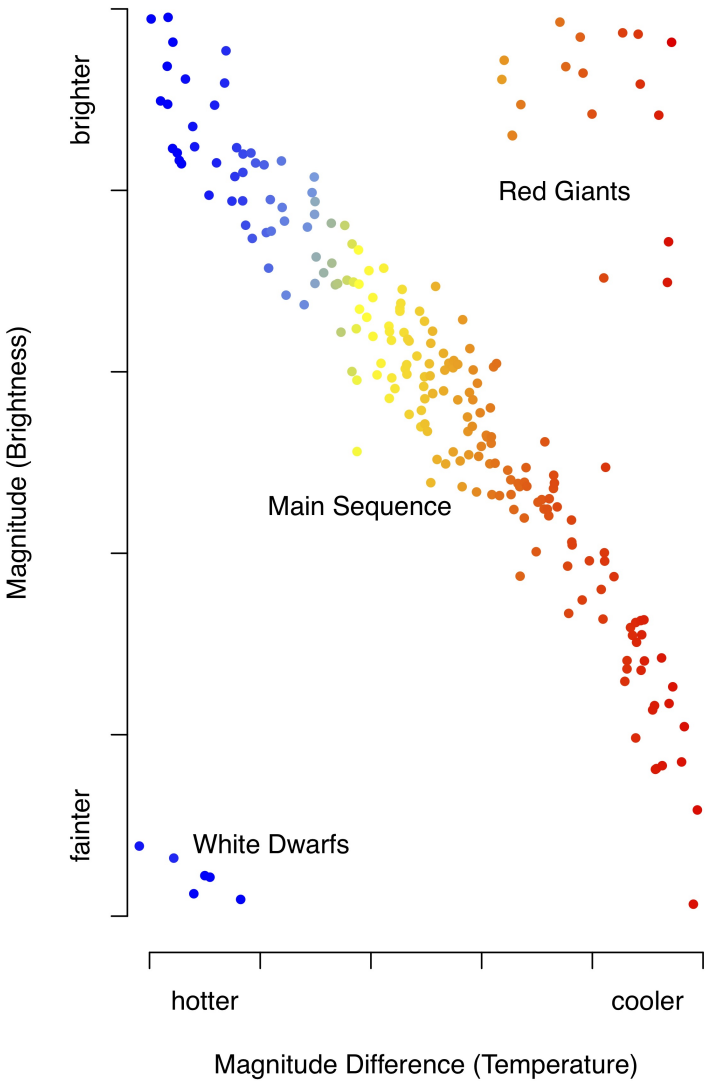


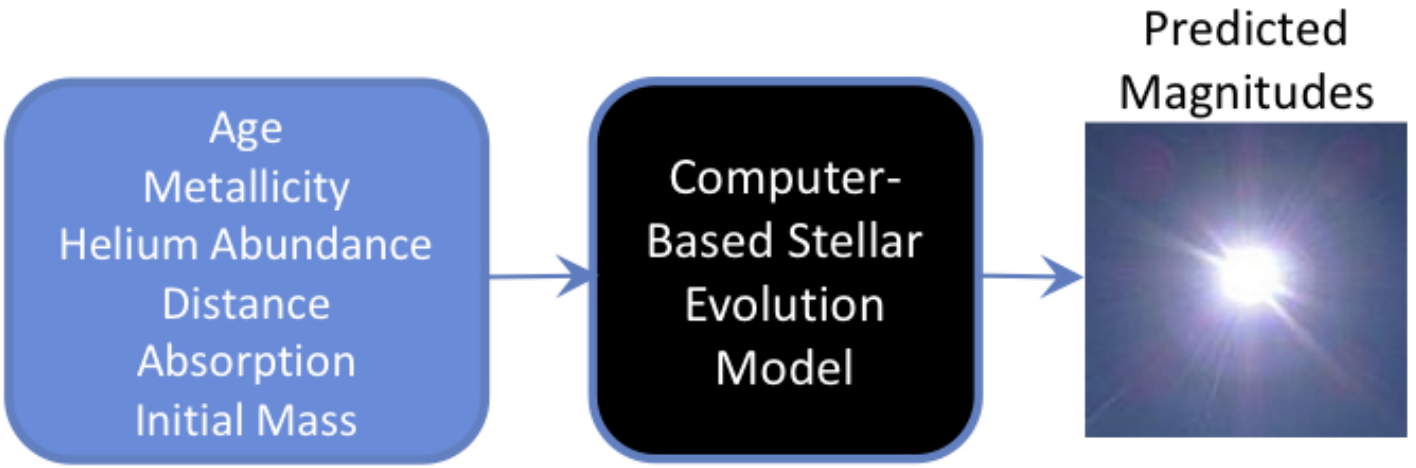
Image Credit: Chandra, NASA

Color Magnitude Diagrams



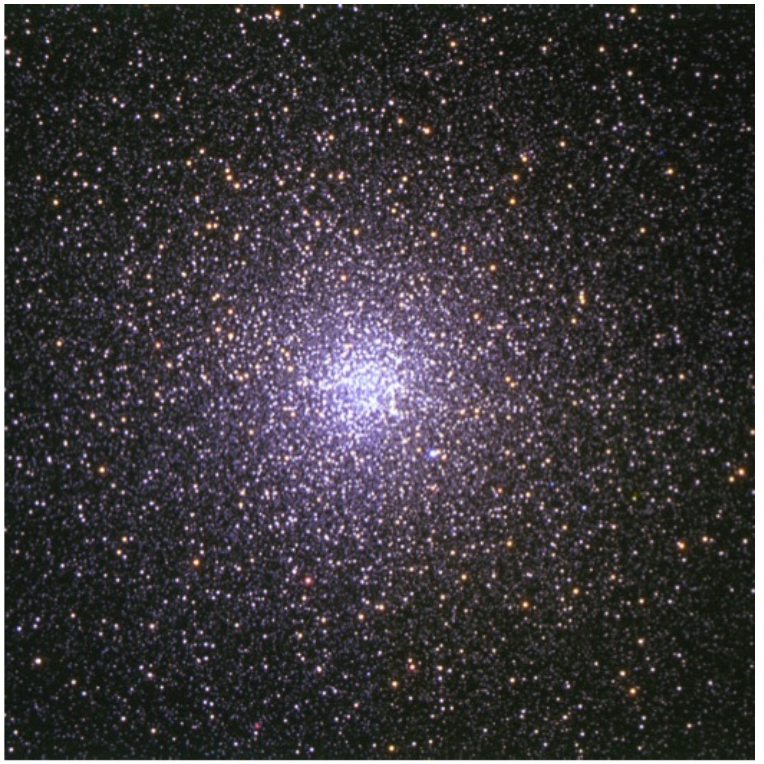
- Color Magnitude Diagrams (CMDs) plot Magnitude Difference vs. Magnitude.
- Used to identify stars at different stages of the lifecycle.
- Provides physical intuition of likely parameter values.
- Typically fit "chi-by-eye" methods.

Computer Model for Stellar Evolution



- We observe a star's *photometric magnitudes*—the apparent brightness of a star in several wide wavelength bands.
- Astrophysicists use computer models to predict the photometric magnitudes of a star given a set of input parameters that describe certain characteristics about the star.

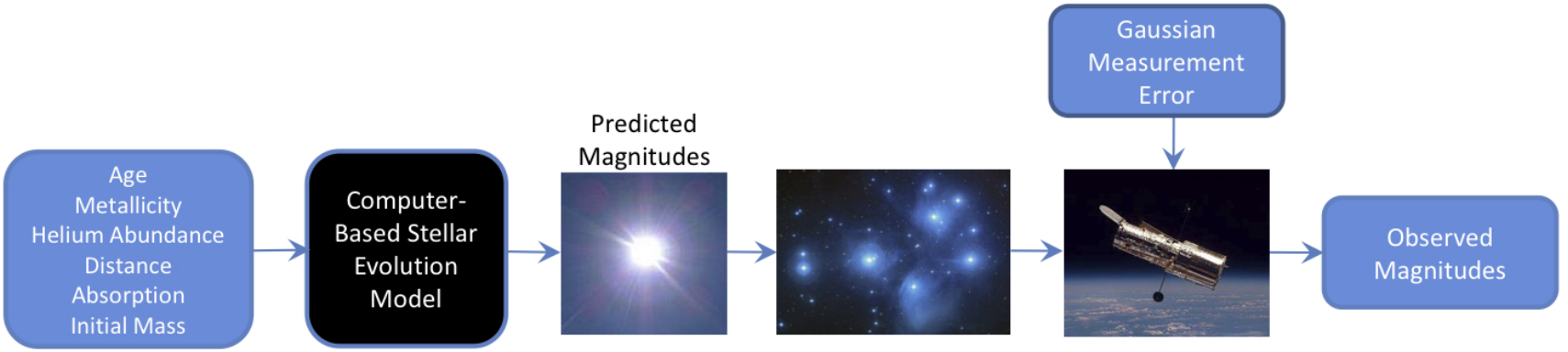
Photometric Magnitudes from Star Clusters

Left: The Hyades open cluster. *Right*: The globular cluster 47 Tucanae. **Image Credit:** NASA

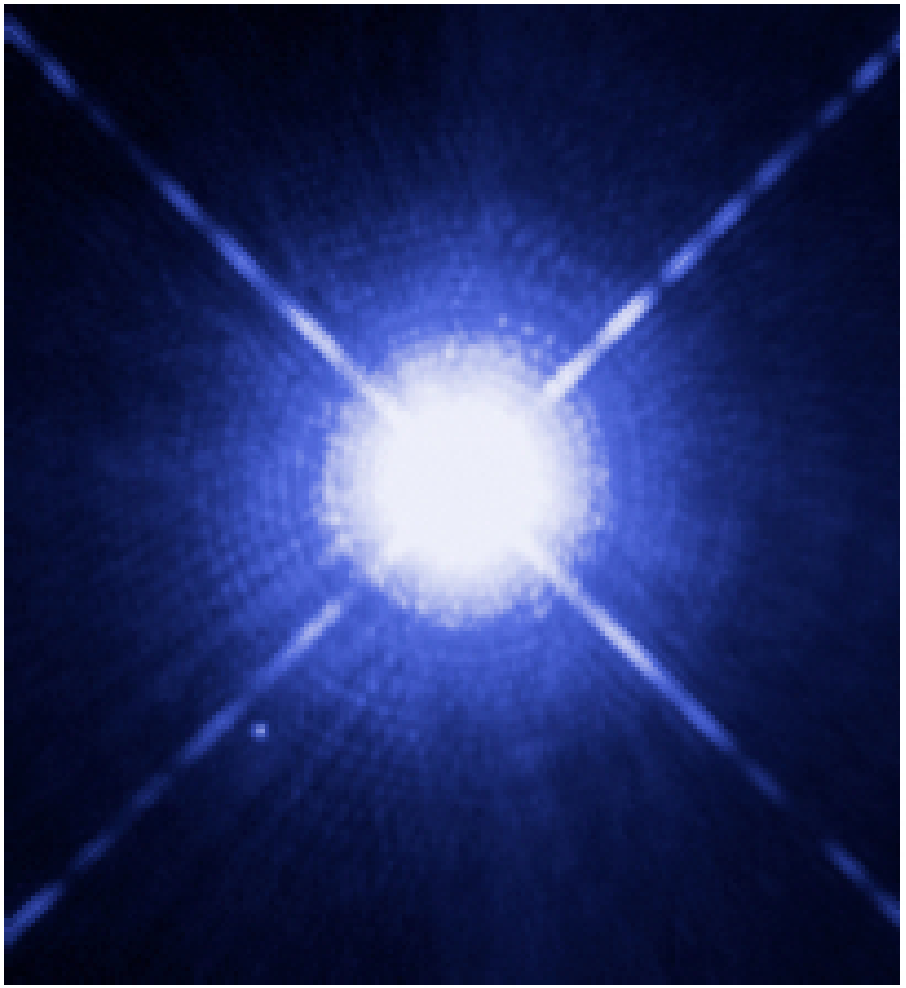
- By working with *star clusters*, we can assume that each star has the same age, metallicity, distance, absorption, and carbon fraction (white dwarfs only). Only initial masses will vary.
- Photometric magnitudes observed with Gaussian measurement error.

Combining Computer Models and Statistical Models



- Observe photometric magnitudes through n different filters per star.
- Model photometric magnitudes as n independent Gaussian random variables.
  - Means involve the computer models for stellar evolution; depend on the stellar evolution parameters.
  - Known Gaussian measurement errors in the covariance matrix.
- Data is contaminated by non-cluster *field stars*.
  - Use a finite mixture model, with field star magnitudes assumed uniform over the range of the data.

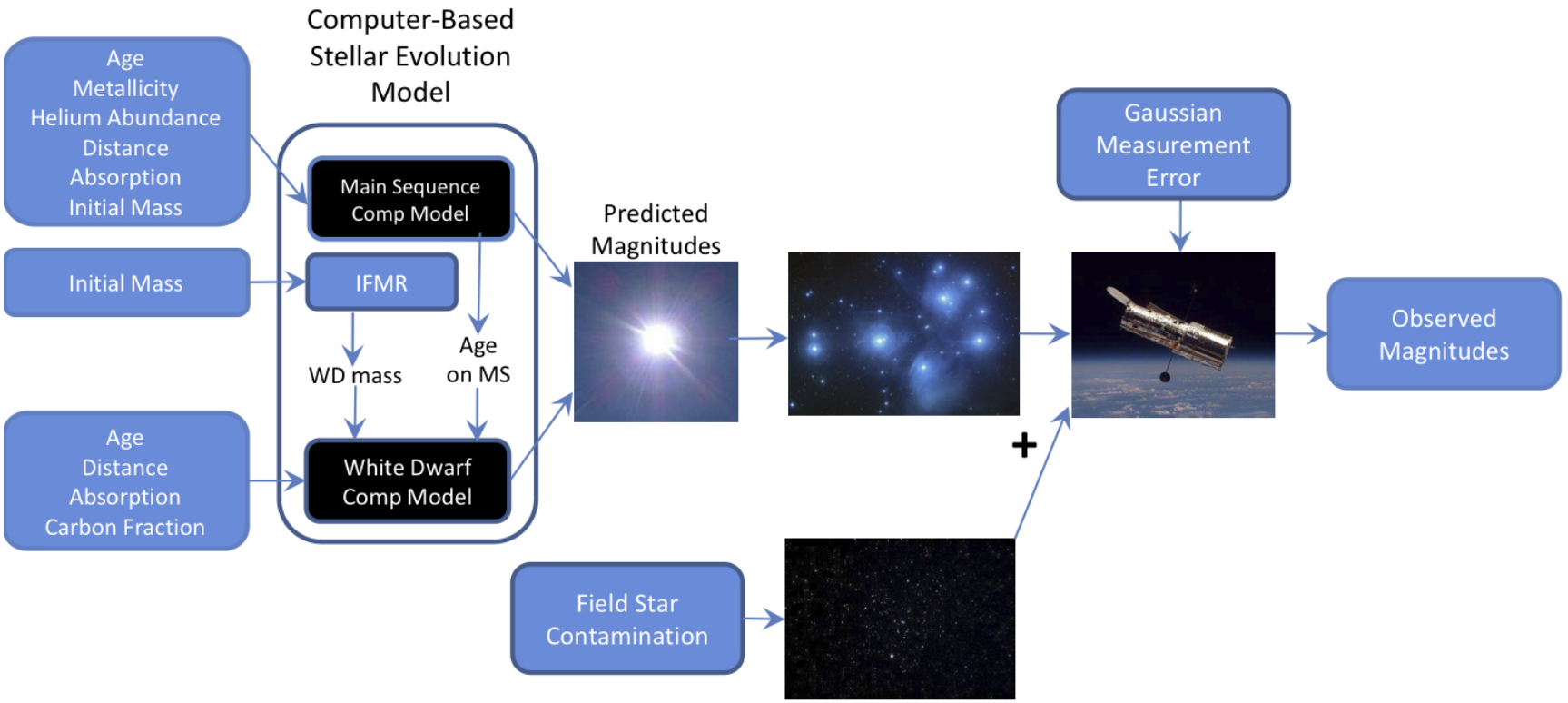
Accounting for White Dwarfs



Sirius A and Sirius B captured by the Hubble Space Telescope. Sirius B is a white dwarf. **Image Credit:** NASA

- Separate computer models are needed to predict photometric magnitudes for white dwarfs (WDs).
- Different models account for different physical processes:
  1. Computer Model for WD cooling
  2. Computer Model for WD atmosphere
  3. Initial Final Mass Relationship
- Combination of 1-3 is the *white dwarf computer model*.

Final Combined Computer/Statistical Model



Prior Distributions

- Whenever possible, informative prior distributions are constructed based on previous studies and astrophysical theory.
- We specify a Gaussian prior distribution on the base 10 logarithm of primary mass that is based on the Miller-Scalo Initial Mass Function (ApJS, 1979):

$$P\left(log_{10}\left(M_{i1}\right)\right) \propto exp\left(-\frac{1}{2}\left(\frac{\log_{10}\left(M_i\right)+1.02}{0.677}\right)^2\right)$$

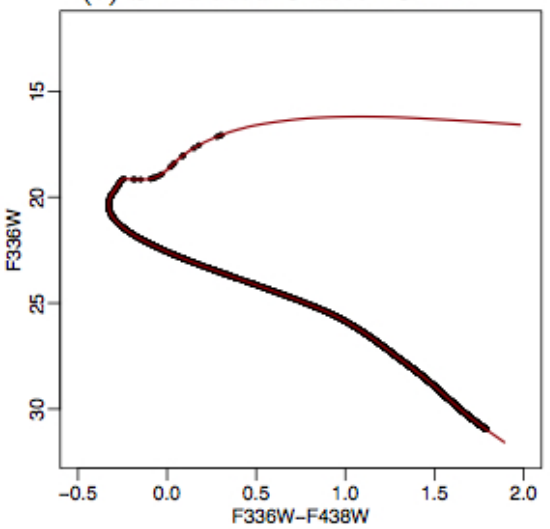
truncated to the range 0.1 to 8.0 solar masses, M_{Sun}.

Statistical Properties of the Parameter Estimates

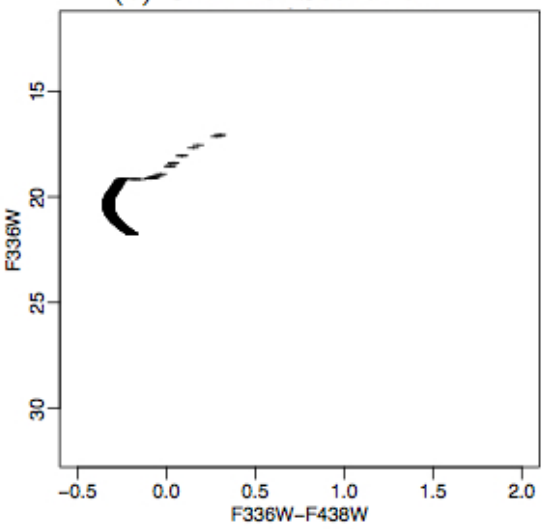
- We explore non-standard asymptotic behavior of the parameter estimates—the influence of the prior distribution does not diminish as the sample size grows.
- Simulate a globular cluster with "average" parameters that are based on previously published studies.
- Consider a partial 2×2 simulation study design: vary data truncation depth and Gassuan measurement error assumed for model fitting.
  - **Scenario 1:** truncation at F275W $= 23$; assumed $\sigma = 0.01$.
  - **Scenario 2:** truncation at F275W $= 23$; assumed $\sigma = 0.03$.
  - **Scenario 3:** truncation at F275W $= 27.5$; assumed $\sigma = 0.03$.
- All scenarios are simulated without any actual errors on the photometric magnitudes.

Simulation Study Design

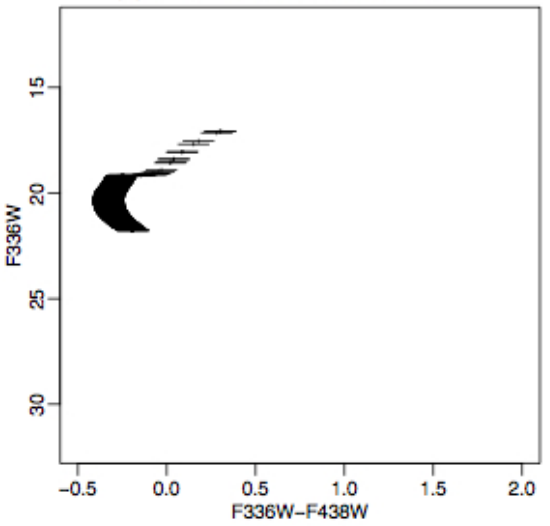
(a) Simulated Cluster CMD



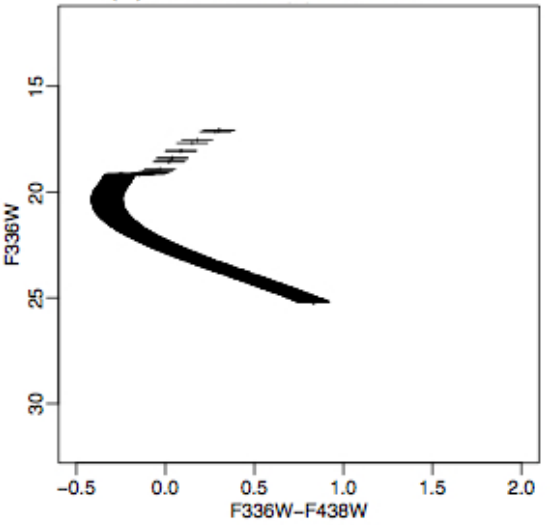
(b) CMD for Scenario 1



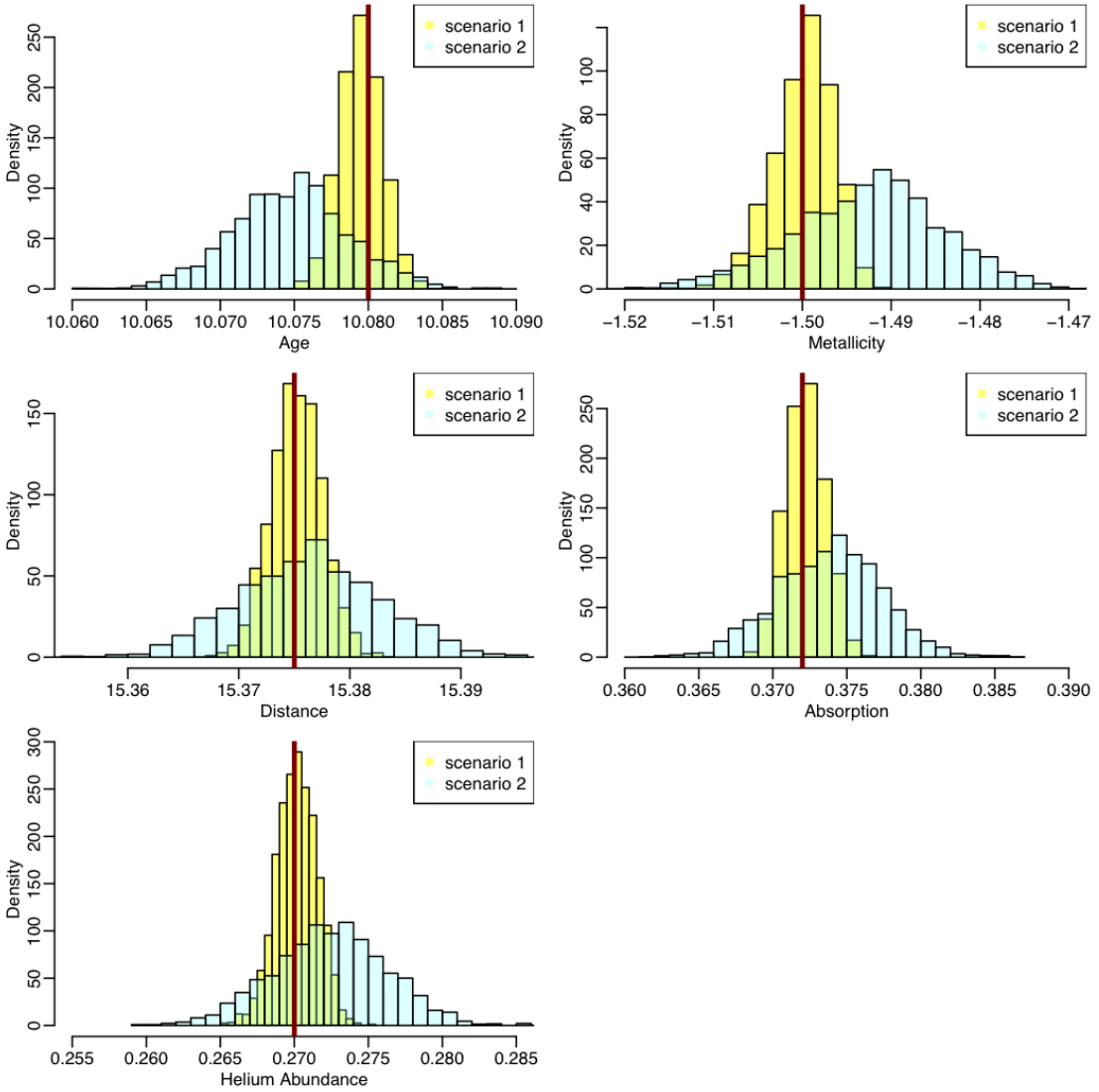
(c) CMD for Scenario 2



(d) CMD for Scenario 3

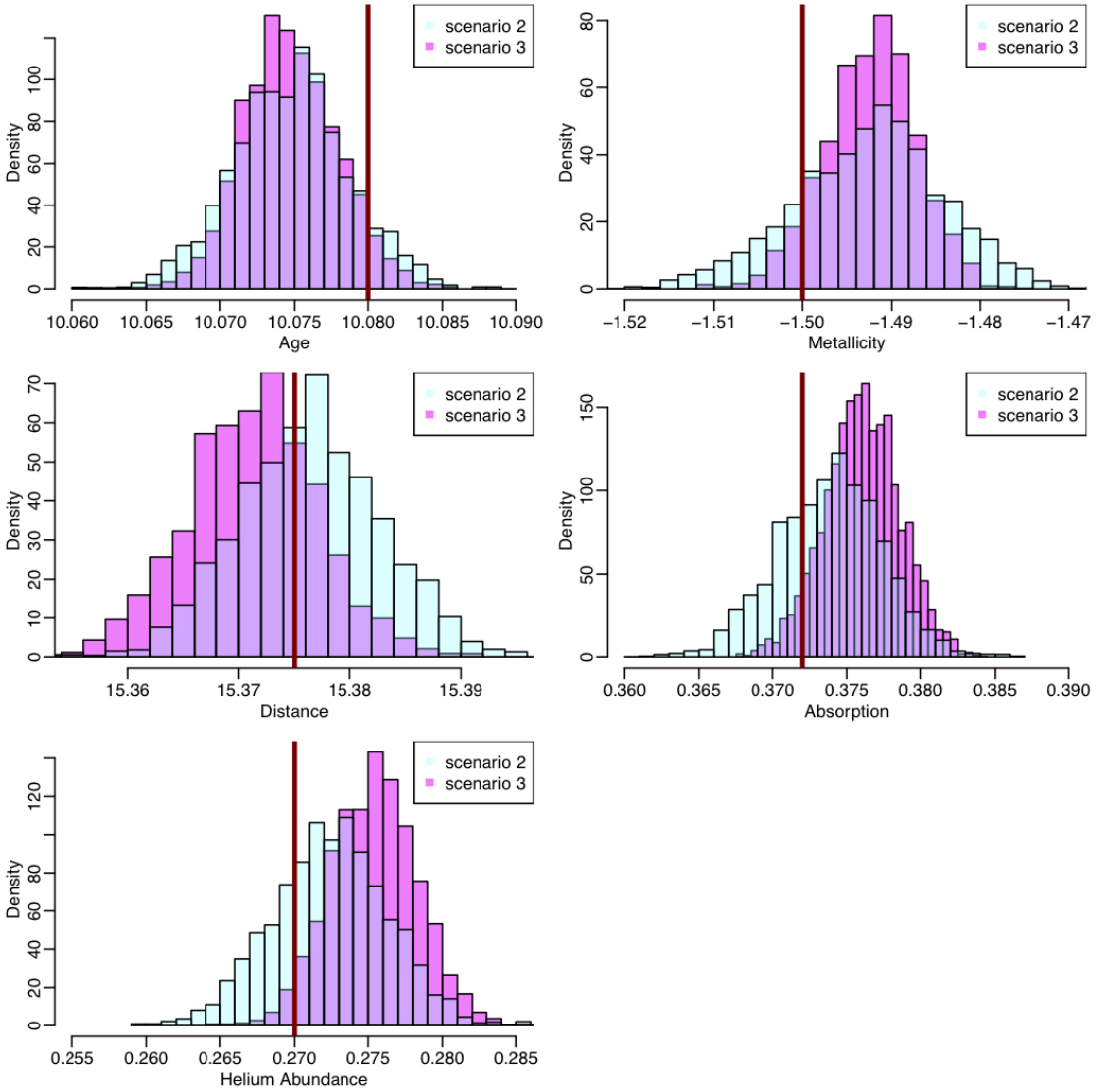


Comparing Scenario 1 and Scenario 2



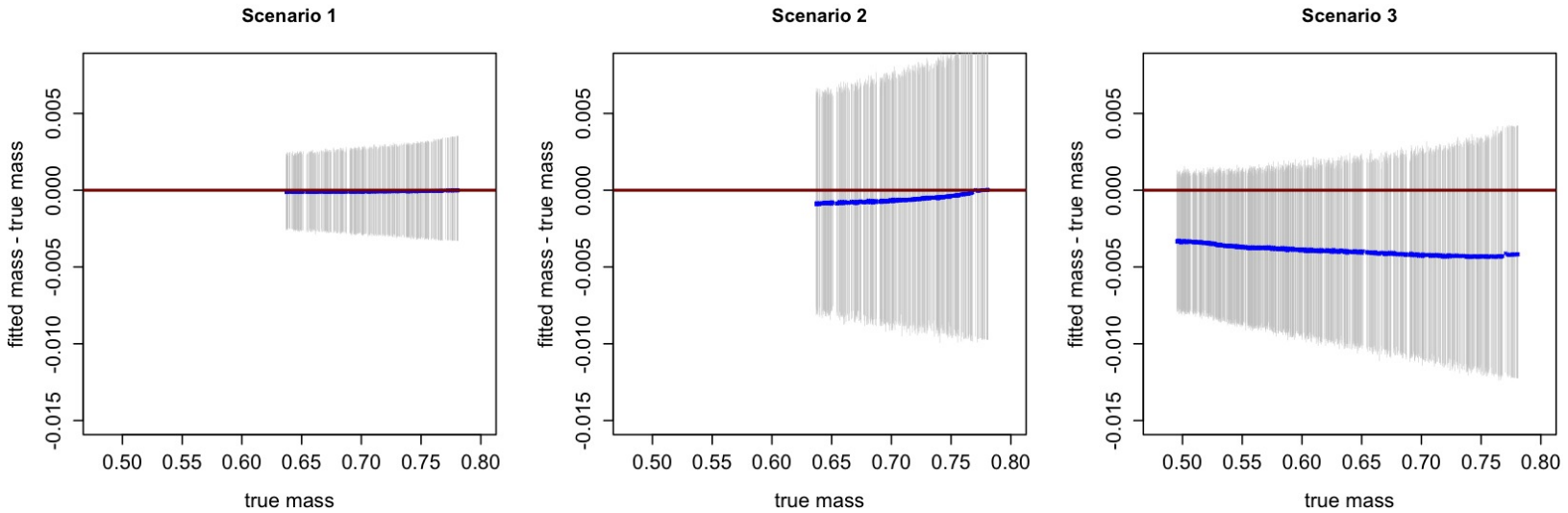
- **Red vertical lines:** true parameter values under the simulation
- **Yellow histograms:** marginal posterior dist'ns under Scenario 1.
- **Blue histograms:** marginal posterior dist'ns under Scenario 2.
- (assumed) σ increases from Scenario 1 to Scenario 2.

Comparing Scenario 2 and Scenario 3



- **Red vertical lines:** true parameter values under the simulation
- **Blue histograms:** marginal posterior dist'ns under Scenario 2.
- **Magenta histograms:** marginal posterior dist'ns under Scenario 3.
- Data depth increases from Scenario 2 to Scenario 3, but deviations do not decrease.
- Influence of prior distribution on initial mass?

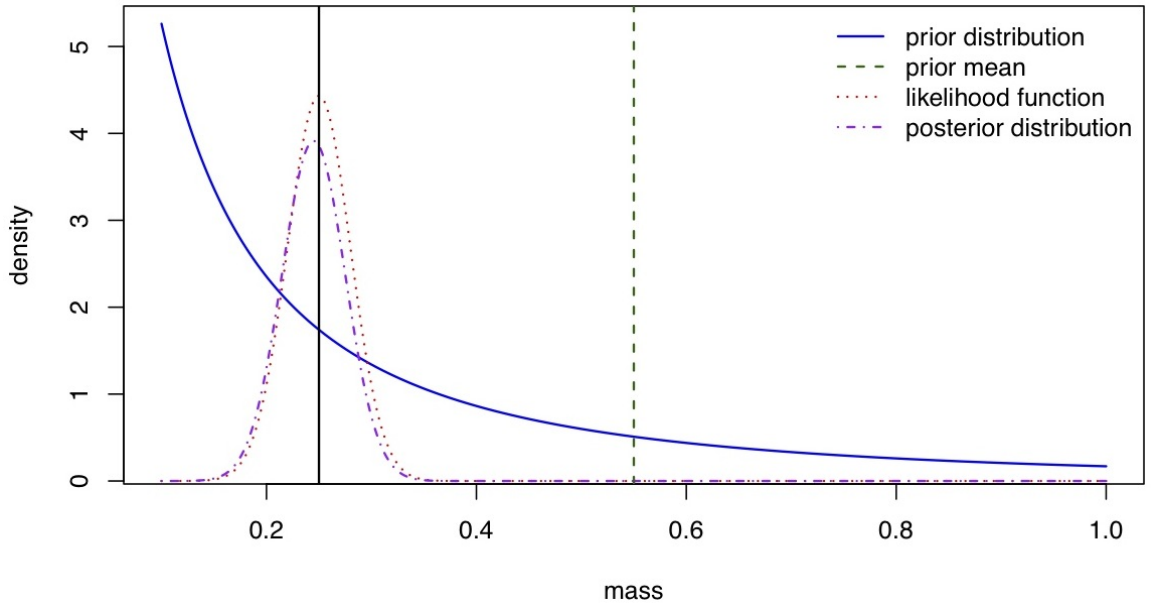
Examining Fitted Mass Values



Blue 'x's denote the difference between the fitted mass and the true mass for each star, and grey vertical bars are 95% intervals.

- Fitted masses are not shrinking towards prior mean (0.55 M_{Sun}).
- *As sample size increases, the influence of the prior distribution on mass does not diminish because there is only one observation (i.e. one star) per mass parameter.*
  - No ability to share strength for these parameters because we do not fit the distribution of the masses.

Toy Example



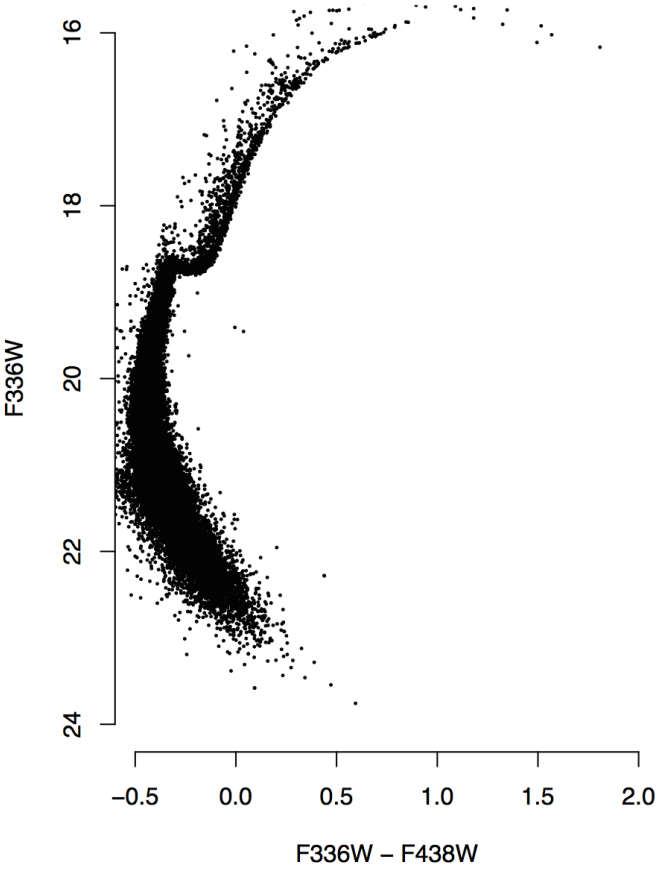
- We model mass, m, as $m \sim N\left(\tilde{m}, \xi^2\right)$.
- Specify a Gaussian prior distribution on the base 10 logarithm of $\tilde{m}$:

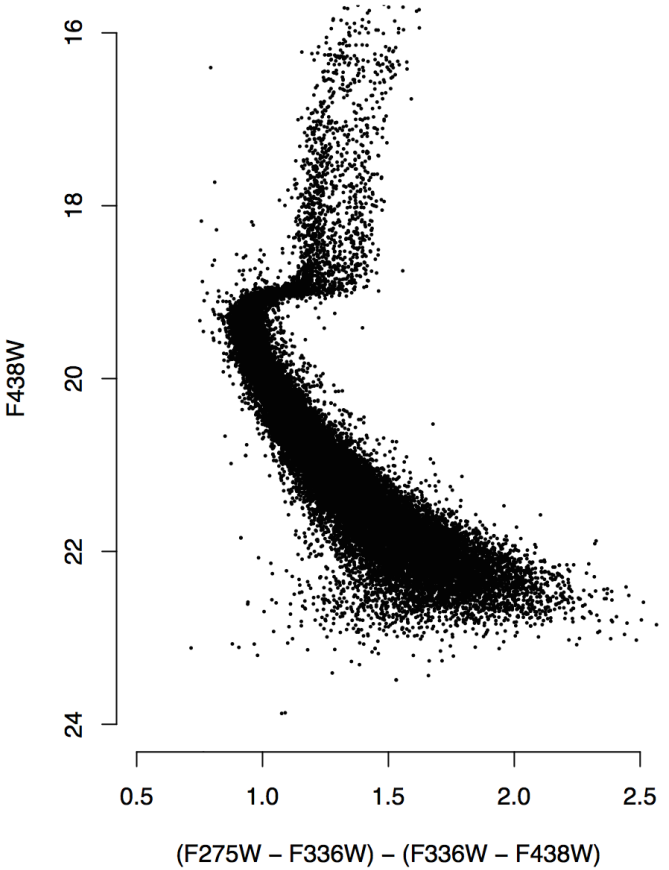
$$P\left(log_{10}\left(\tilde{m}\right)\right) \propto exp\left(-\frac{1}{2}\left(\frac{\log_{10}\left(\tilde{m}\right)+1.02}{0.677}\right)^2\right)$$

  truncated to the range 0.1 to 8.0 M_{Sun}.
- Suppose we observe $m = \tilde{m} = 0.25$, with $\xi = 0.03$.

Outline

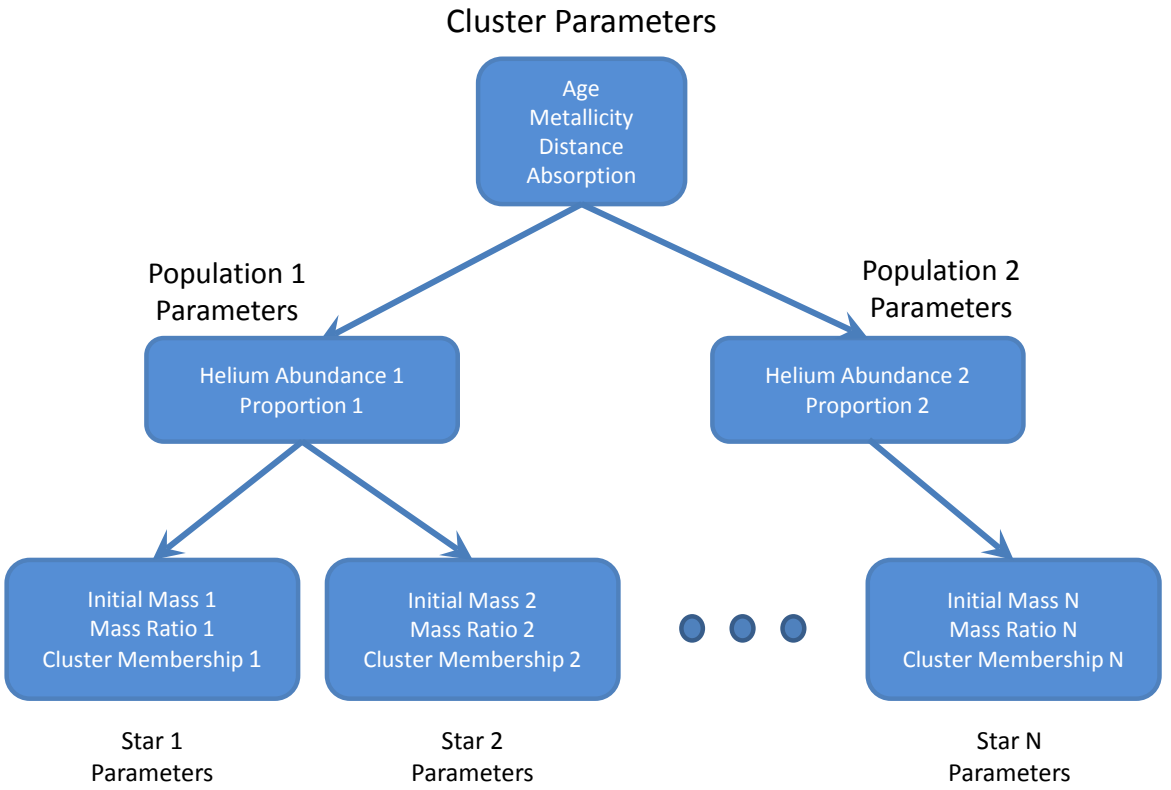
Two CMDs for NGC 5272





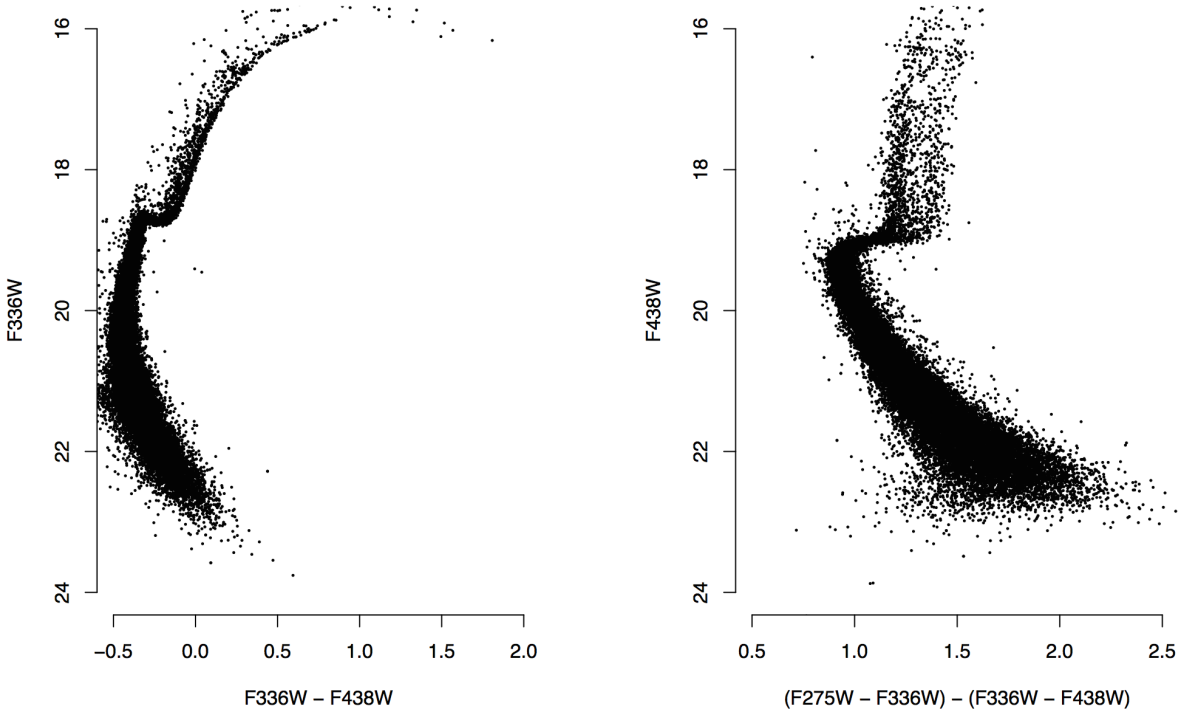
- multiple strands in CMD $\rightarrow$ multiple stellar populations
- multiple stellar populations $\rightarrow$ multiple epochs of star formation

Extending the Hierarchical Model



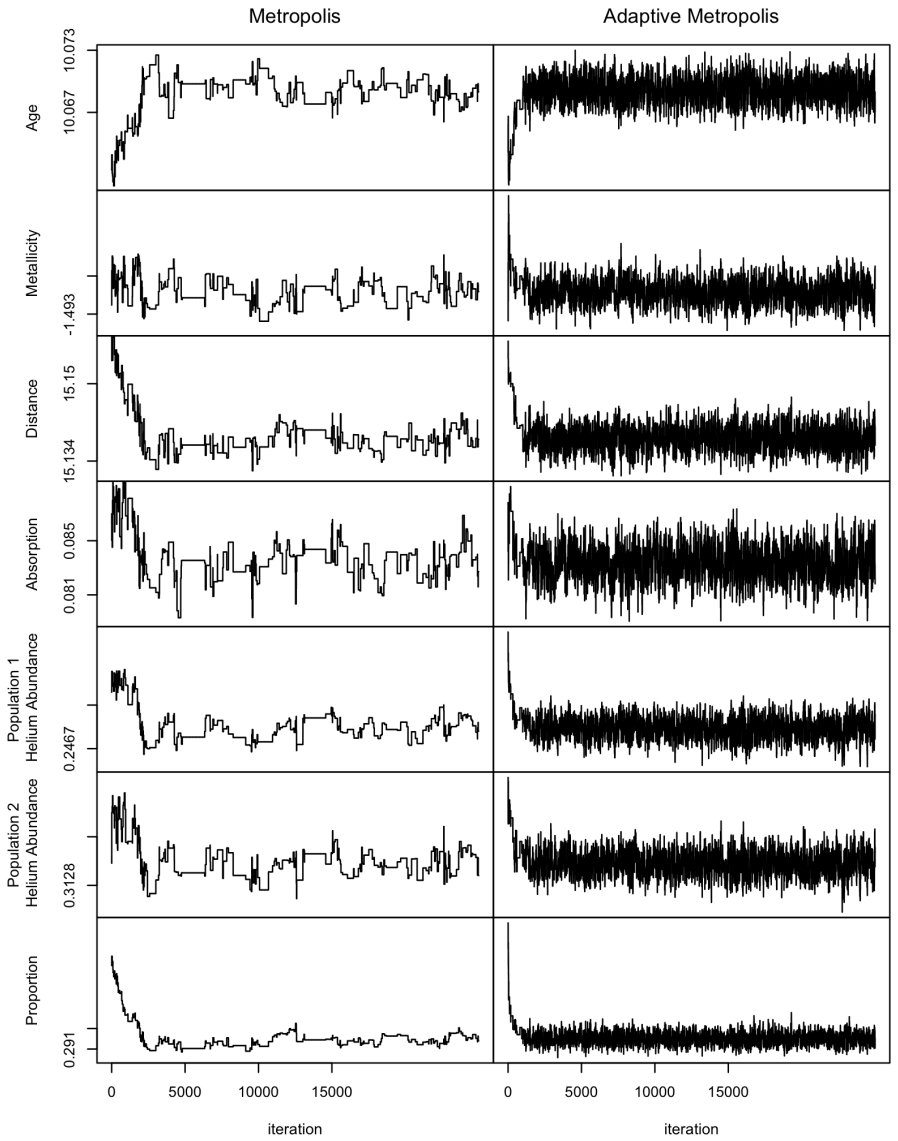
- Preliminary statistical model is based on a hierarchy of n photometric magnitudes within stars and of stars within a cluster.
- Here, we add *population parameters*: helium abundance and the proportion of stars in population k.

Statistical Model for Two-Population Cluster



- Likelihood function for a single-population cluster can be viewed as a mixture of two populations: cluster stars and field stars.
- We now model each cluster star with a two-component finite mixture of n-dimensional multivariate Gaussian distrubutions.
- Results in three stellar populations: field stars and two cluster pops.

Statistical Computing

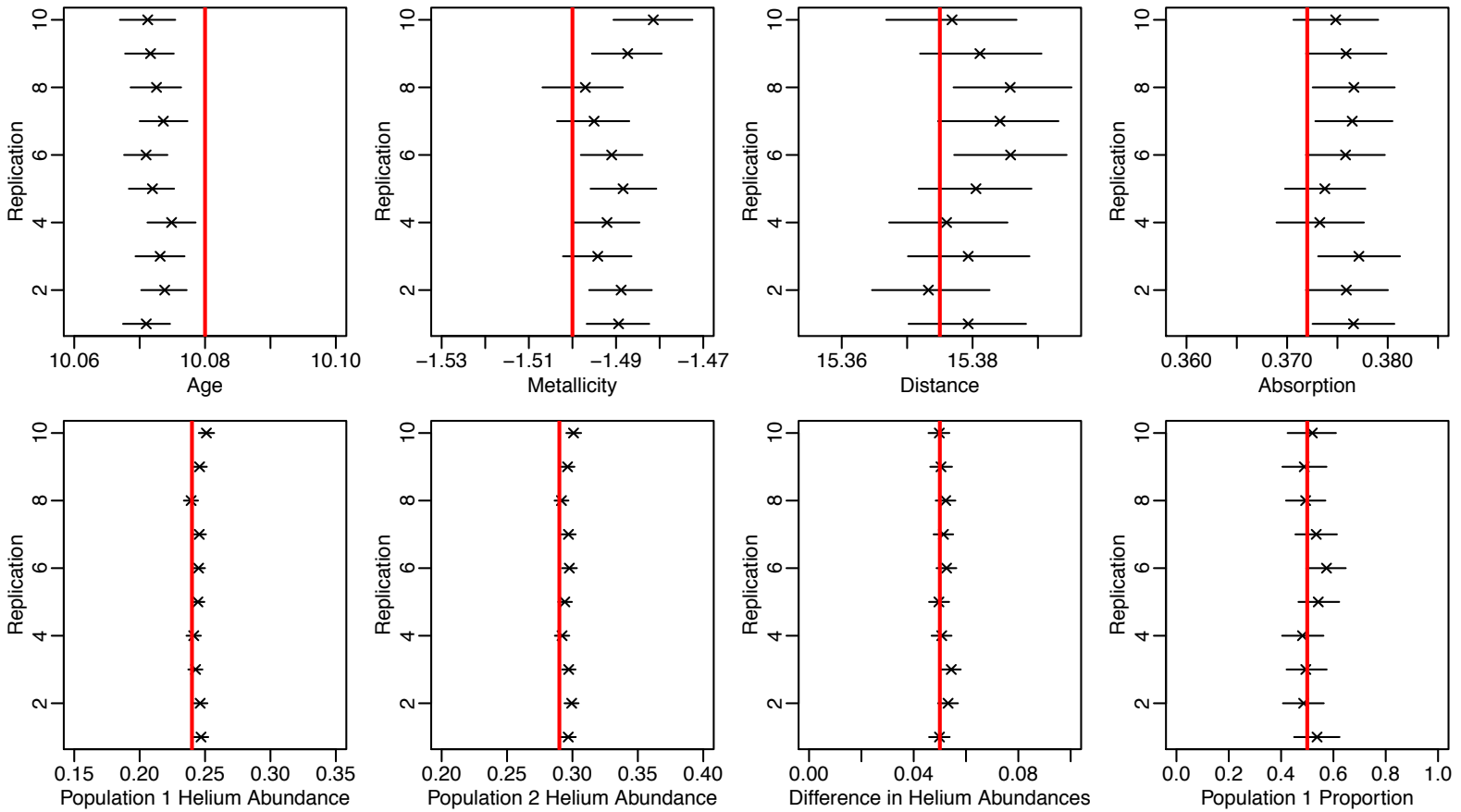


- Marginalize over all the star-specific parameters.
- Explore the marginal posterior distribution using an *Adaptive Metropolis* algorithm.
- Improved efficiency and convergence compared to non-adaptive Metropolis implementation.

Numerical Study

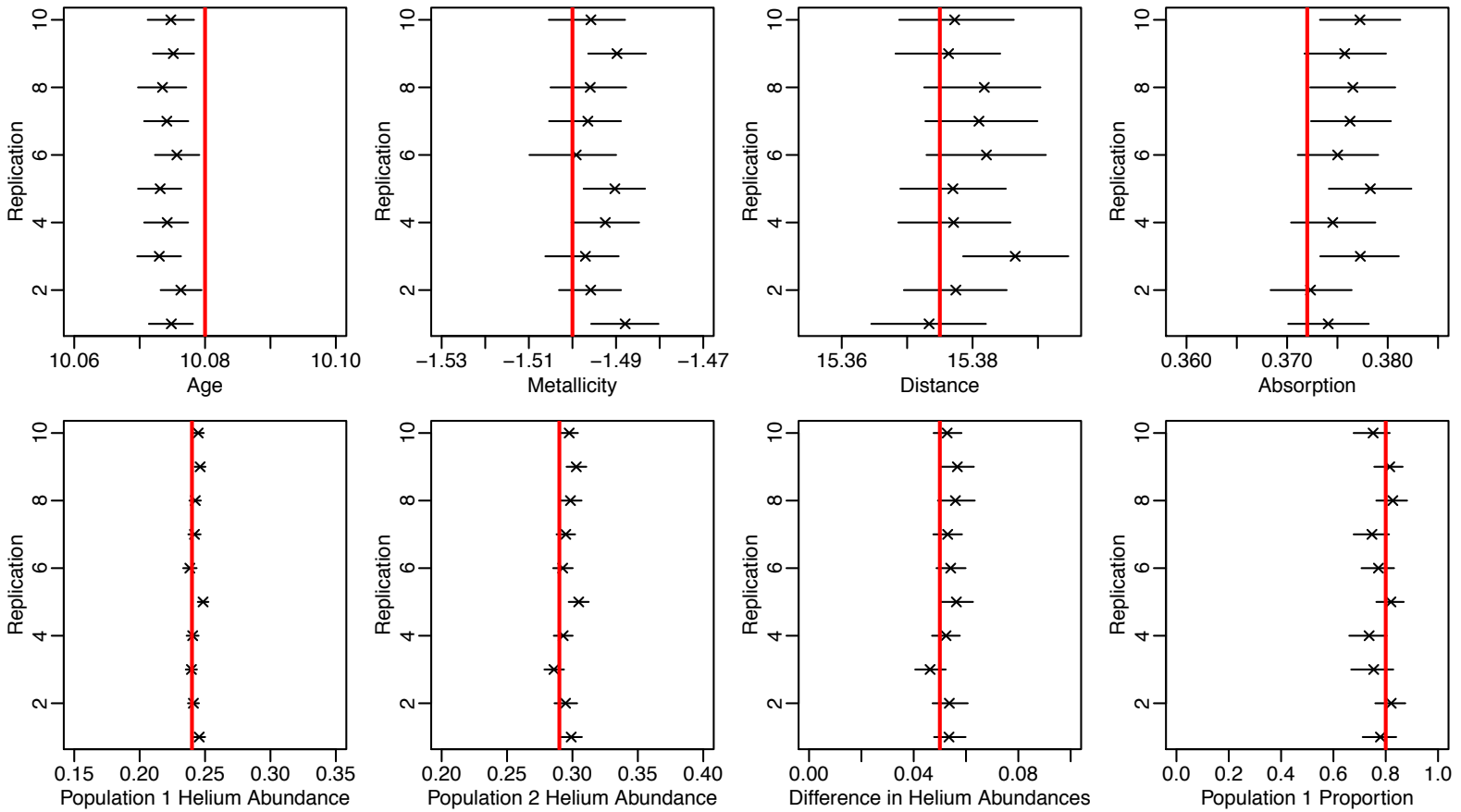
- We simulate two-population globular clusters under three scenarios, with 10 replicate clusters per scenario.
- Scenarios differ in the % of stars belonging to population 1: 50%, 80%, and 100% for scenarios 1, 2, and 3, respectively.
- Each cluster is simulated with "average" cluster parameters.
- Each cluster is also simulated with $\phi_{Y1} = 0.24$ and $\phi_{Y2} = 0.29$.

Scenario 1: 50% of Stars in Population 1



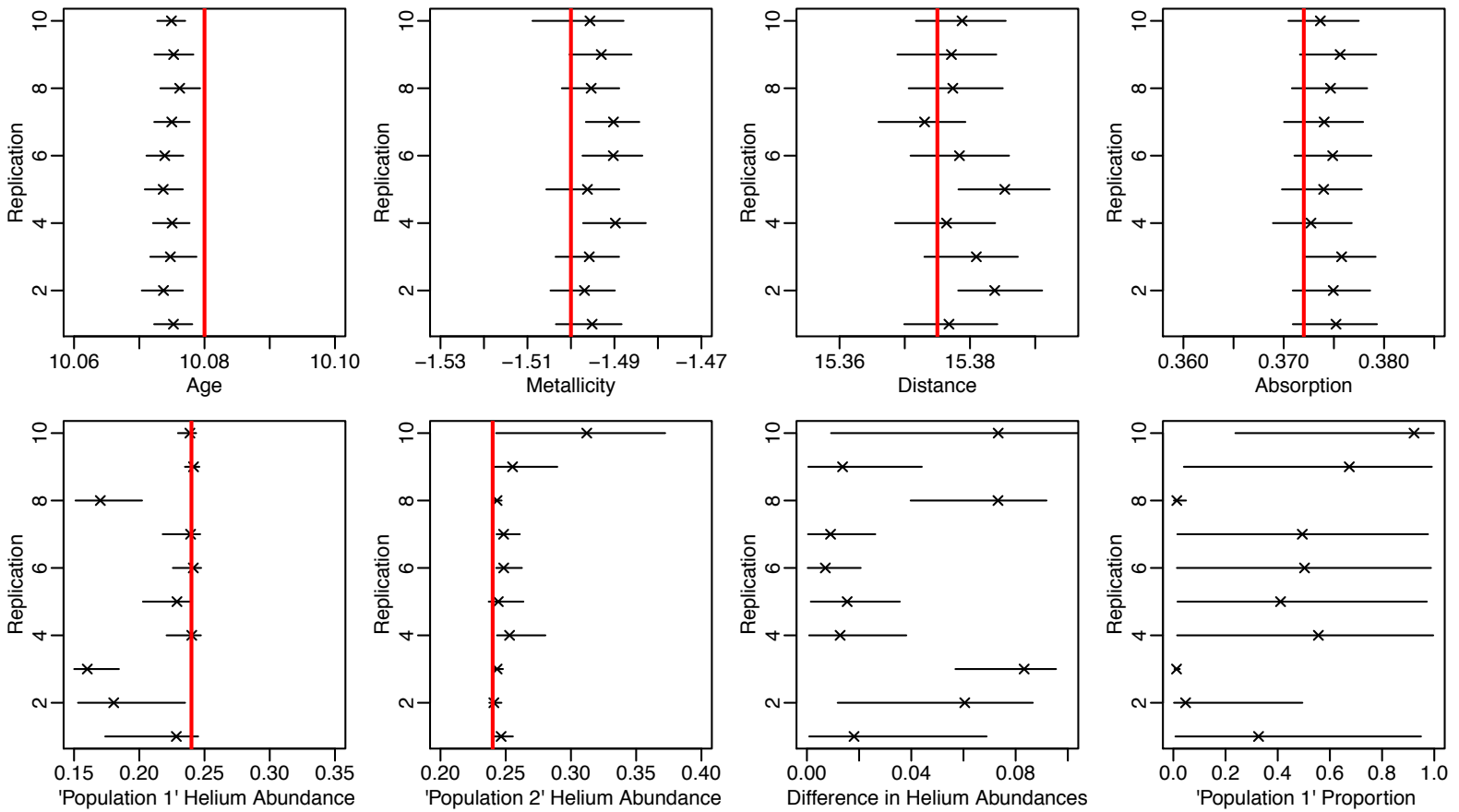
- red vertical lines: true parameter values under the simulation
- 'x's are posterior means and horizontal bars are 95% intervals

Scenario 2: 80% of Stars in Population 1



- red vertical lines: true parameter values under the simulation
- 'x's are posterior means and horizontal bars are 95% intervals

Scenario 3: 100% of Stars in Population 1



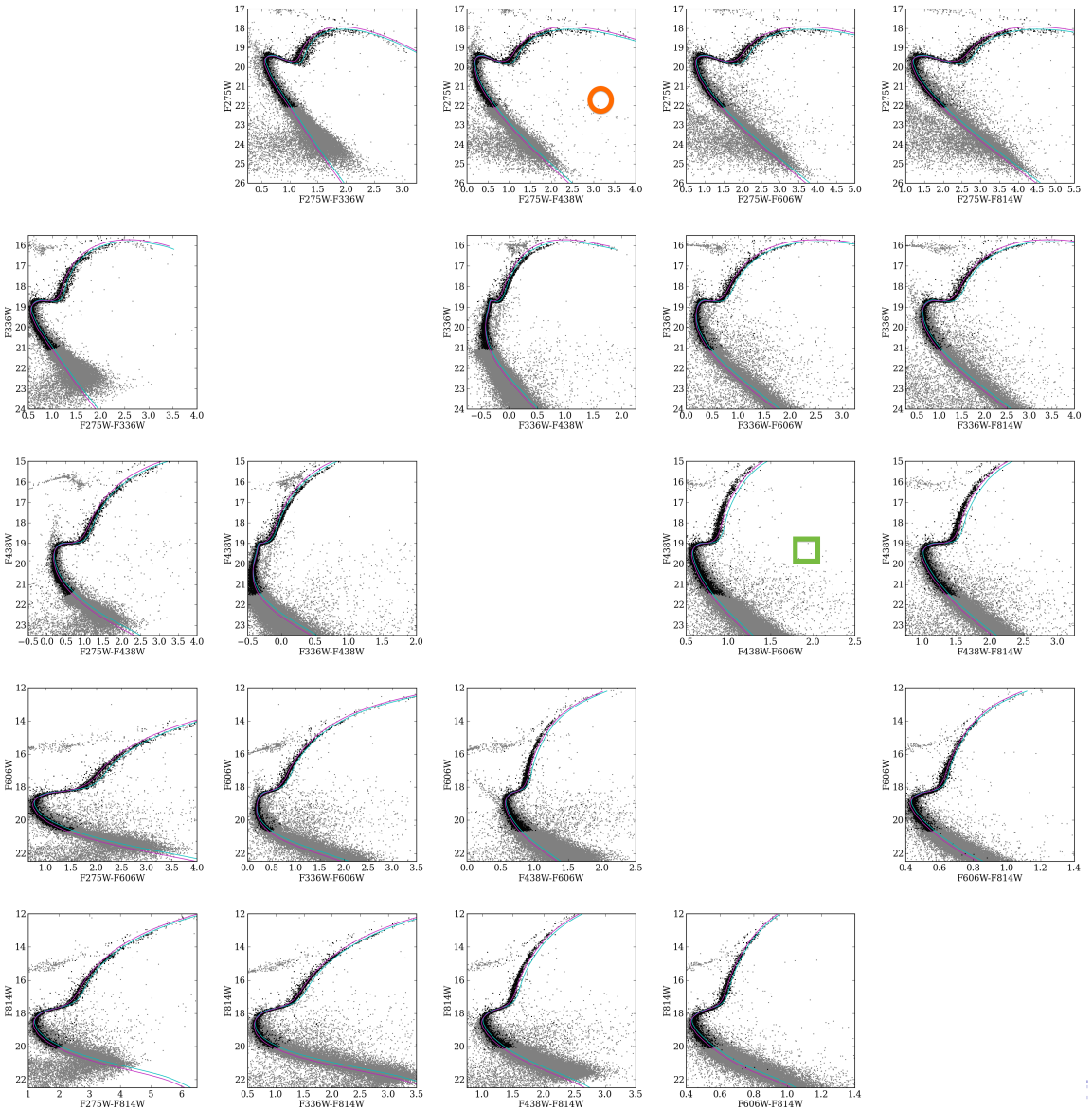
- red vertical lines: true parameter values under the simulation
- 'x's are posterior means and horizontal bars are 95% intervals
- Can diagnose model misspecification!

Data Analysis: NGC 5272

Image Credit: "Hewholooks" via wikipedia

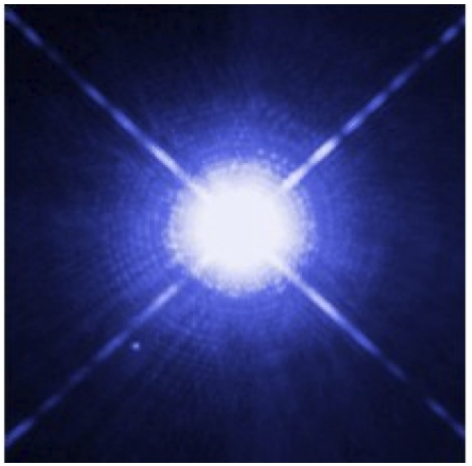
- **Observed data:** photometric magnitudes in 5 wavelength bands
  - two visual wavelength bands
  - three "magic trio" UV wavelength bands (Piotto et al., 2015)
- New stellar evolution models to predict UV photometric magnitudes

CMD Matrix with Fitted Model



Outline

Scientific Motivation



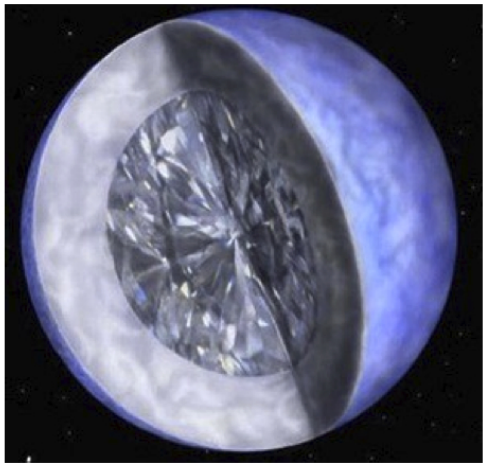


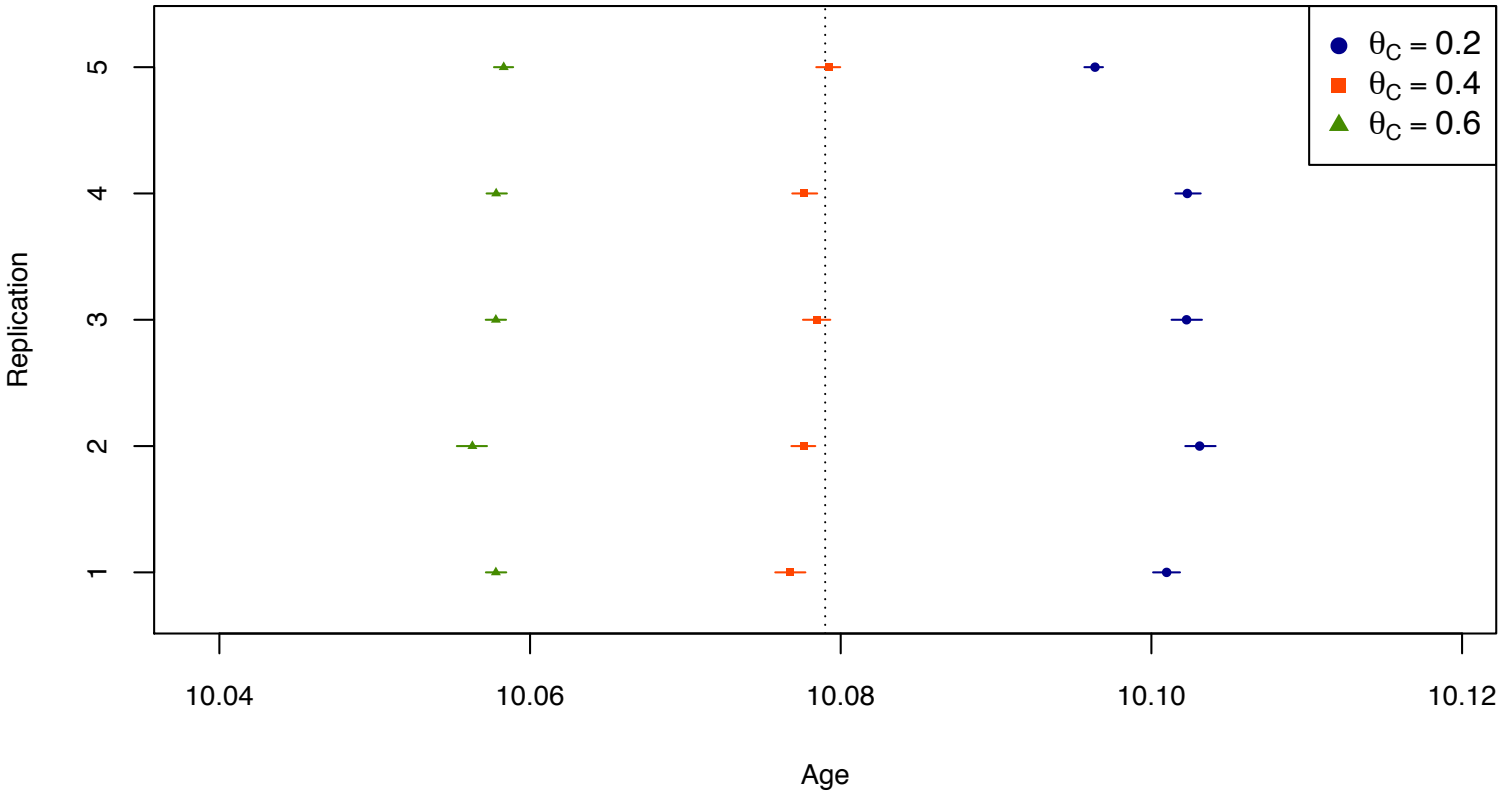
Image Credit: NASA and CfA

- The carbon fraction in WDs depends on experimentally uncertain nuclear reaction rates $\rightarrow$ provides insight into fundamental physics.
- Affects the WD cooling rate, and therefore the implied age of any WD.
- Test of stellar evolution models, and fills in a gap in our understanding of late-stage stellar evolution.

Numerical Study: Simulation Design

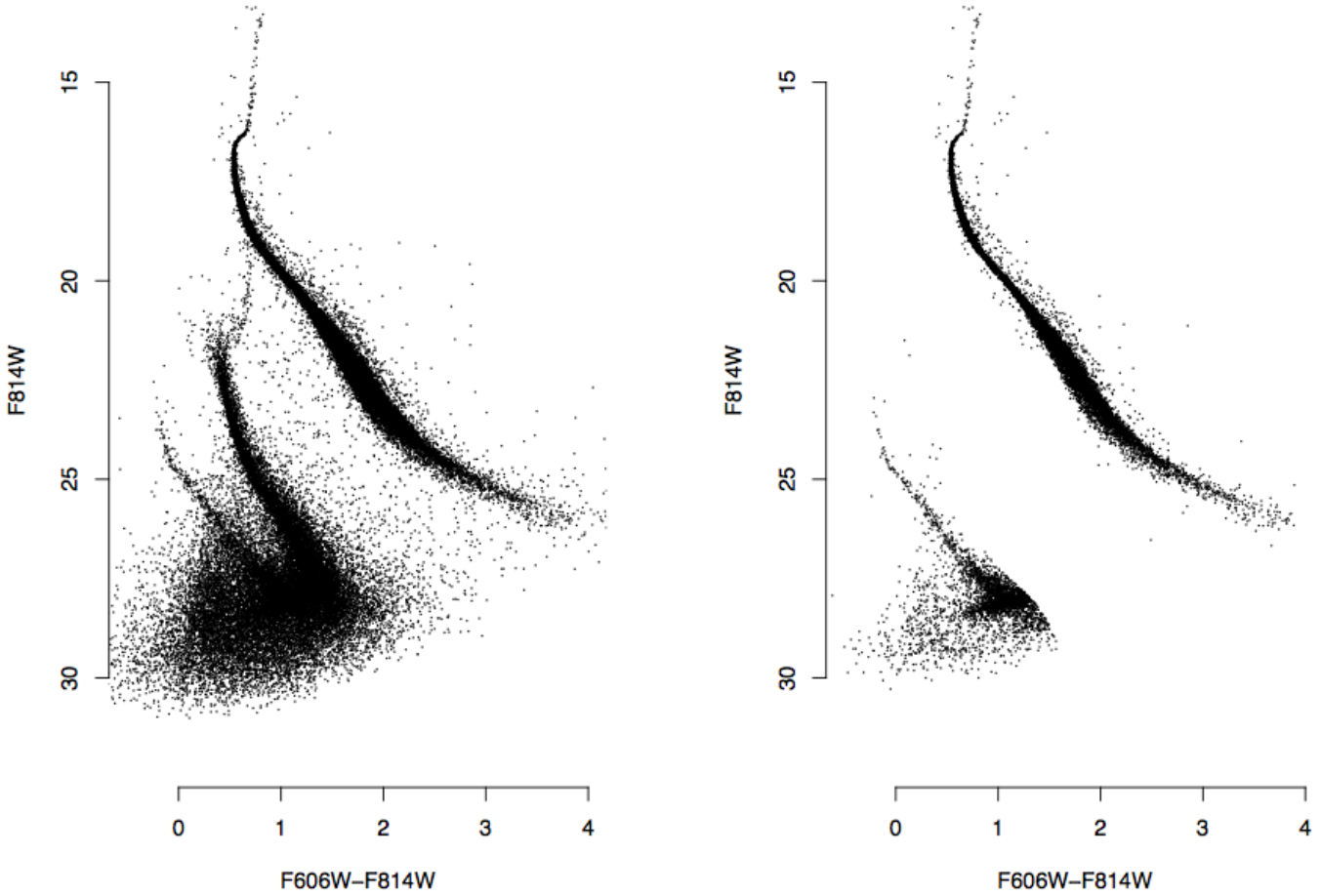
- We demonstrate with a numerical study the potential folly of determining the age of a star cluster based on its WDs (a.k.a. *cosmochronometry*) when the carbon fraction is misspecified.
- We simulate five replicate clusters with carbon fraction = 0.2, five replicate clusters with carbon fraction = 0.4, and five replicate clusters with carbon fraction = 0.6.
- Every cluster is simulated with the same values for the other parameters. In particular, age = 10.079 for each simulated cluster.
- We fix carbon fraction = 0.4 for model fitting.

Numerical Study: Results



- Estimates of age are too low/high when the carbon fraction value assumed for model fitting is incorrect.
  - Small bias when the carbon fraction is correctly specified.

Data Analysis: 47 Tuc

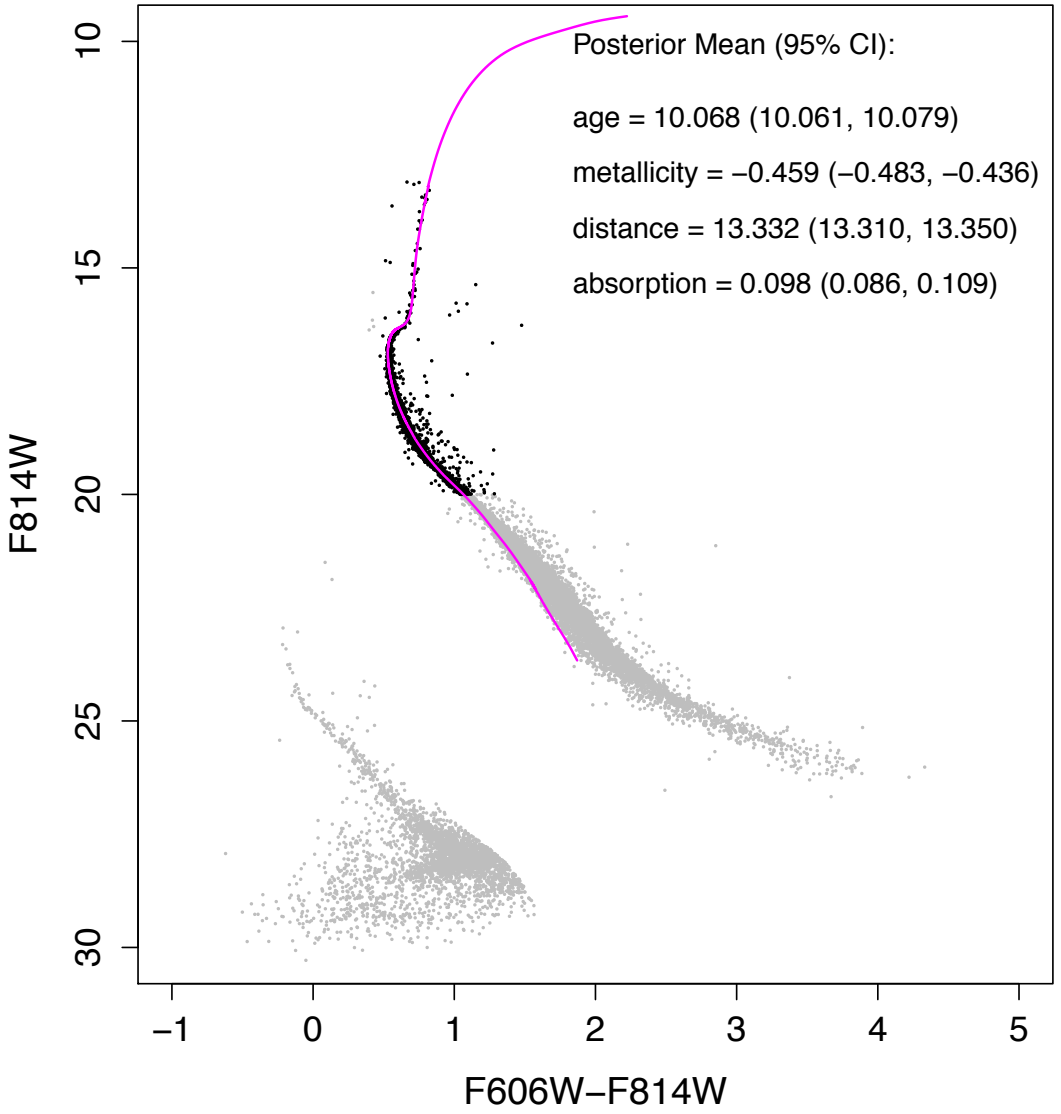


- Observed photometric magnitudes in two wavelength bands
- Contamination from the *Small Magellanic Cloud* (SMC)

Data Analysis: 47 Tuc

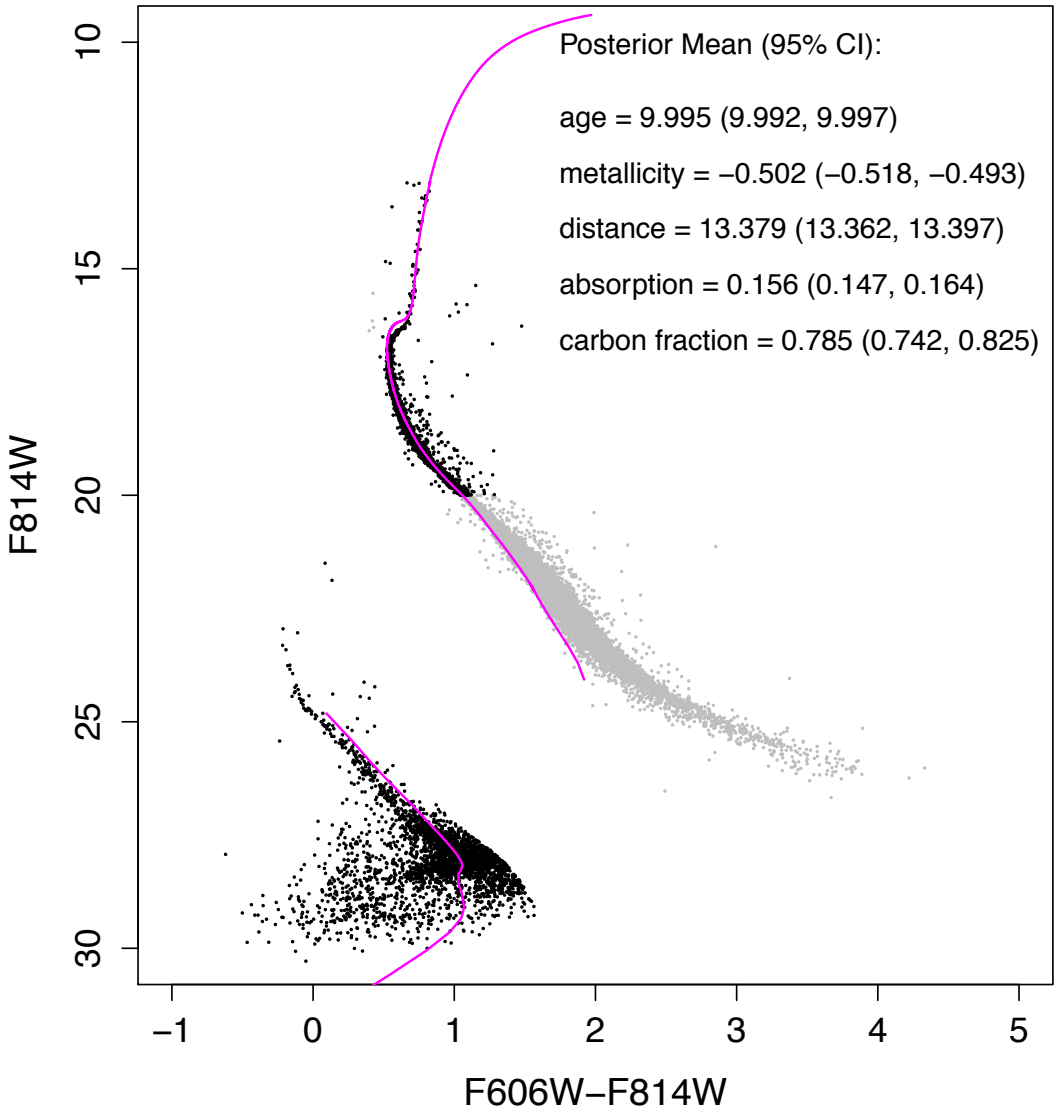
- We proceed through a series of model fits under different constraints.
  - First fit the model without any WDs to establish "baseline" parameters that are determined by main-sequence and red-giant stars (MS/RG stars) alone.
  - Then, include WDs and impose different conditions when fitting the model.
  - Always fitting the carbon fraction when WDs are included.

Fit 1



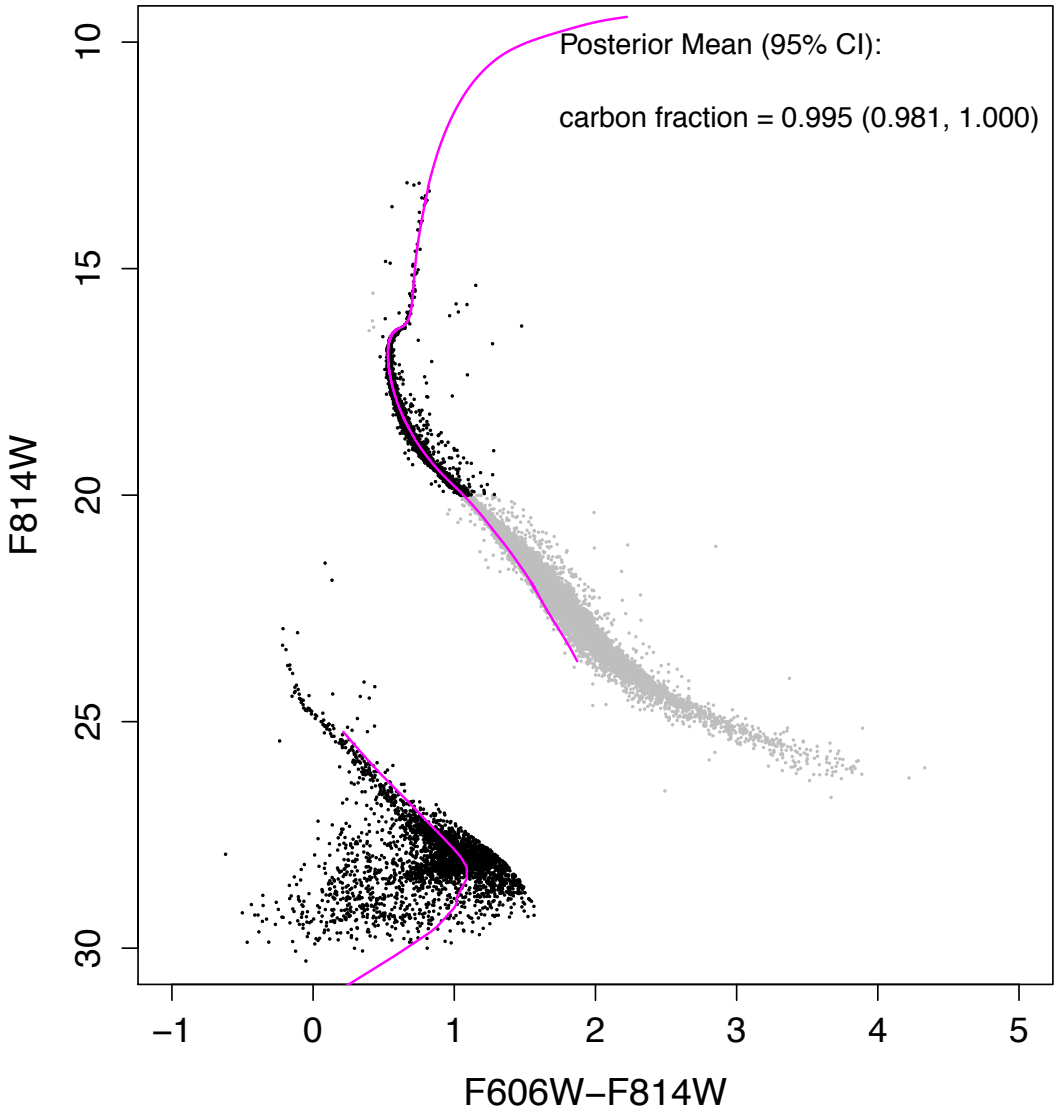
- **Fit 1:** baseline model fit
  - no observed WDs included
  - parameters constrained by MS/RG stars alone
- Magenta curve is the fitted computer model (i.e. fitted *isochrone*)
- Fitted values for the cluster parameters are reasonable

Fit 2



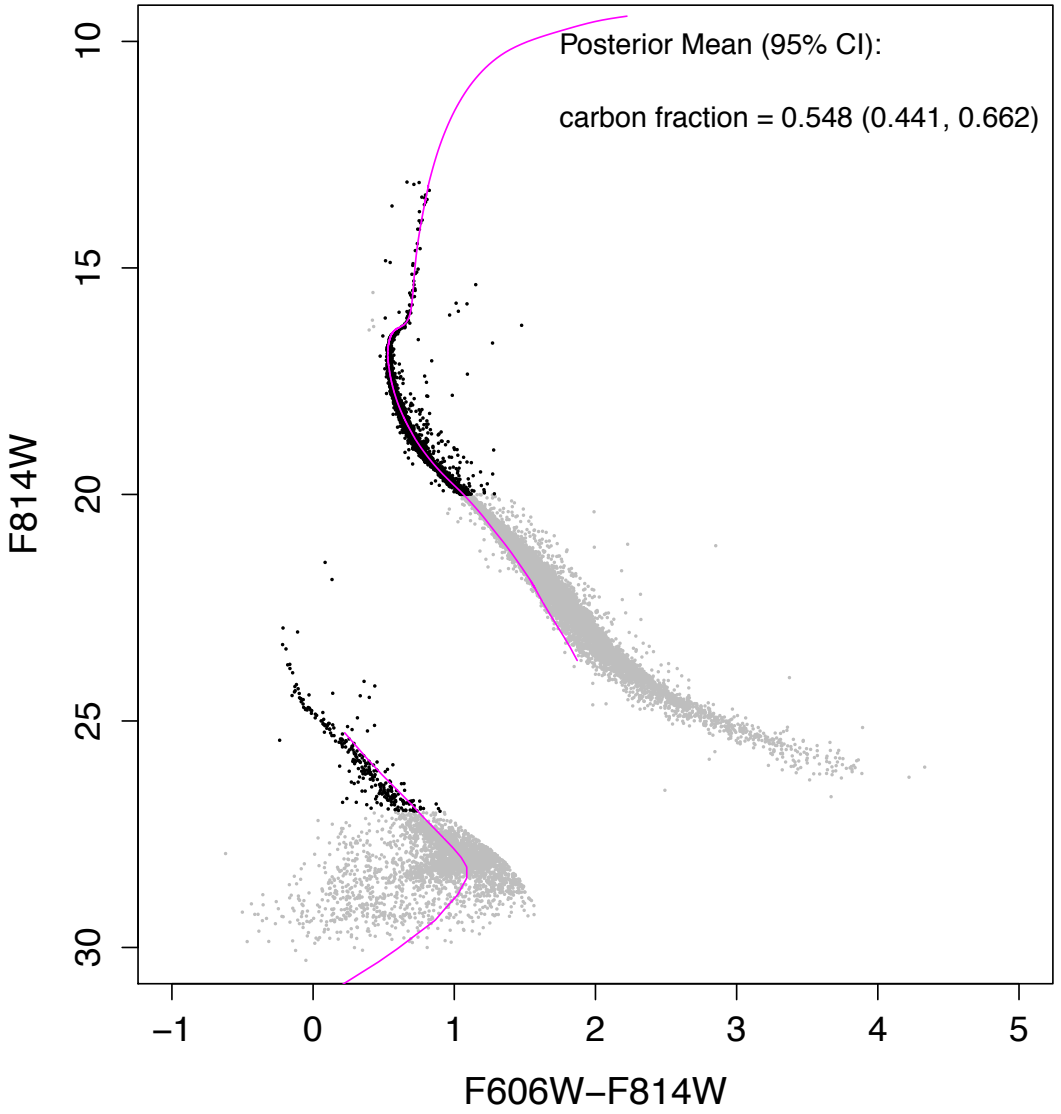
- **Fit 2:** all observed WDs included
- Fitted value for age is below what astrophysicists consider reasonable
- Observed WDs are "bluer" than fitted computer model

Fit 3



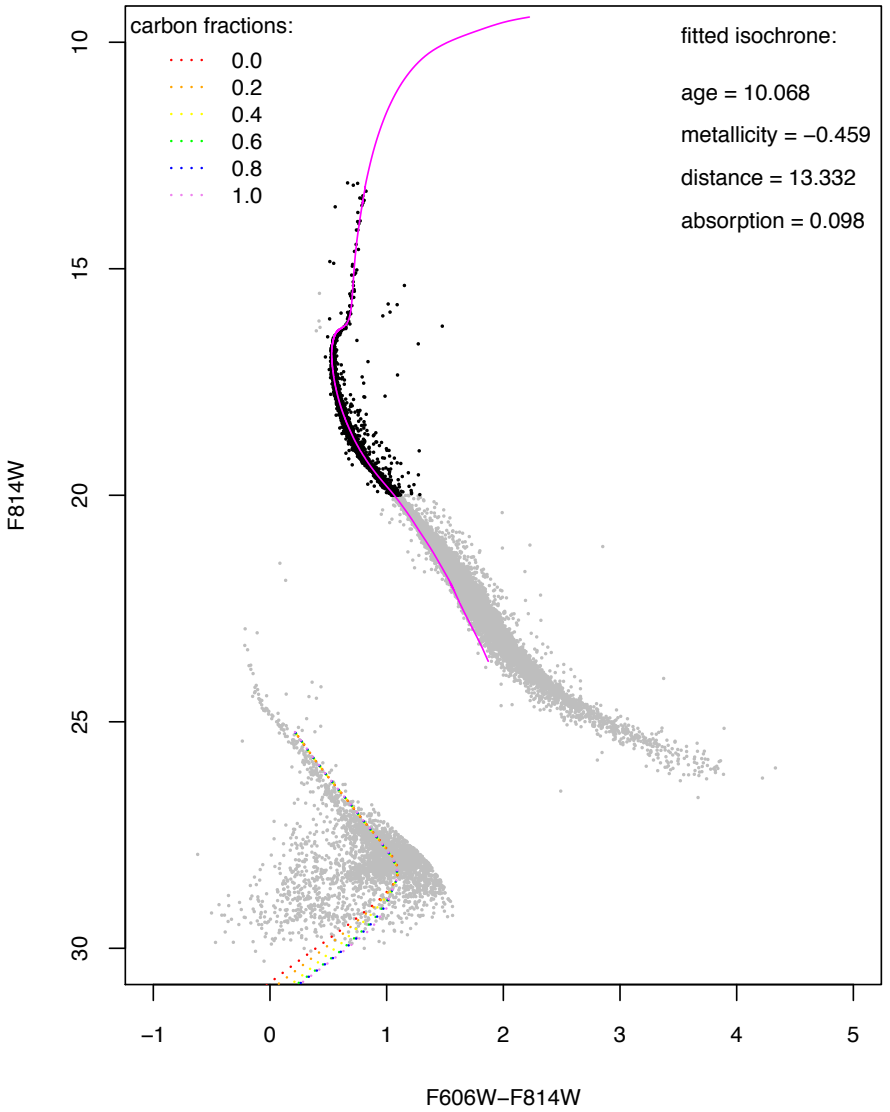
- **Fit 3:** all observed WDs included
  - parameters excluding carbon fraction fixed to their posterior mean under Fit 1
- Posterior mode for carbon fraction at 1.0; not reasonable
- Observed WDs still "bluer" than fitted computer model

Fit 4



- **Fit 4:** WDs truncated at F814W = 27 b/c of SMC contamination
  - parameters excluding carbon fraction fixed to their posterior mean under Fit 1
- Reasonable fitted value for carbon fraction
- Observed WDs still "bluer" than fitted computer model

Computer Model Disagreement



- Disagreement between separate computer models for MS/RG stars and WDs
- Fitted values under the baseline model fit (not involving WDs) are reasonable
- Fitted computer model for WDs is "bluer" than observed WDs, regardless of carbon fraction
- Useful feedback for designing computer models that better represent the observed data

Thanks!

Modeling Solar Activity:

- David A. van Dyk
- Vinay Kashyap
- Thomas C.M. Lee
- Yaming Yu
- C. Alex Young

Bayesian Analysis of Stellar Evolution:

- David A. van Dyk
- Ted von Hippel
- Nathan Stein
- Rachel Wagner-Kaiser
- Elliot Robinson
- Elizabeth Jeffery
- William H. Jefferys
- Steven DeGennaro

And the CHASC International Astro-Statistics Collaboration!

For Further Reading I


Stenning et al.
Morphological Image Analysis and Its Application to Sunspot Classification.
Statistical Challenges in Modern Astronomy V, Springer, 2012.


Stenning et al.
Morphological Feature Extraction for Statistical Learning with Applications to Solar Image Data.
Statistical Analysis and Data Mining, August, 2013.


Yu et al.
A Bayesian Analysis of the Correlations Among Sunspot Cycles.
Solar Physics, 2012.


Stenning et al.
A Bayesian Analysis of the Solar Cycle Using Multiple Proxy Variables.
Current Trends in Bayesian Methodology with Applications (Editors: S. Upadhyay, D.K. Dey, U. Singh and A. Loganathan), Chapman & Hall/CRC Press, 2015. *In Press.*

For Further Reading II

van Dyk et al.
Statistical Analysis of Stellar Evolution.

The Annals of Applied Statistics, 2009.

Stein et al.
Combining computer models to account for mass loss in stellar evolution.
Statistical Analysis and Data Mining, January, 2013.