Stellar Label Estimates Through Use of Machine Learning

Mason Beaty and Leigh Harwell
Mentor: Dr. Hayden

NASA SVS | TESS Tunes into an All-sky 'Symphony' of Red Giants

Red Giant Stars

Red giants are stars that exhaust the hydrogen in their core and have now started to burn it in its outer layers. Eventually these stars have a helium flash and burn helium in their core.

Red giants are significant to galactic archaeology.

- They are very bright, allowing us to see them from large distances
- Allow us to map out the structure of the Milky Way
- Key observables in understanding how our Galaxy has formed and evolved

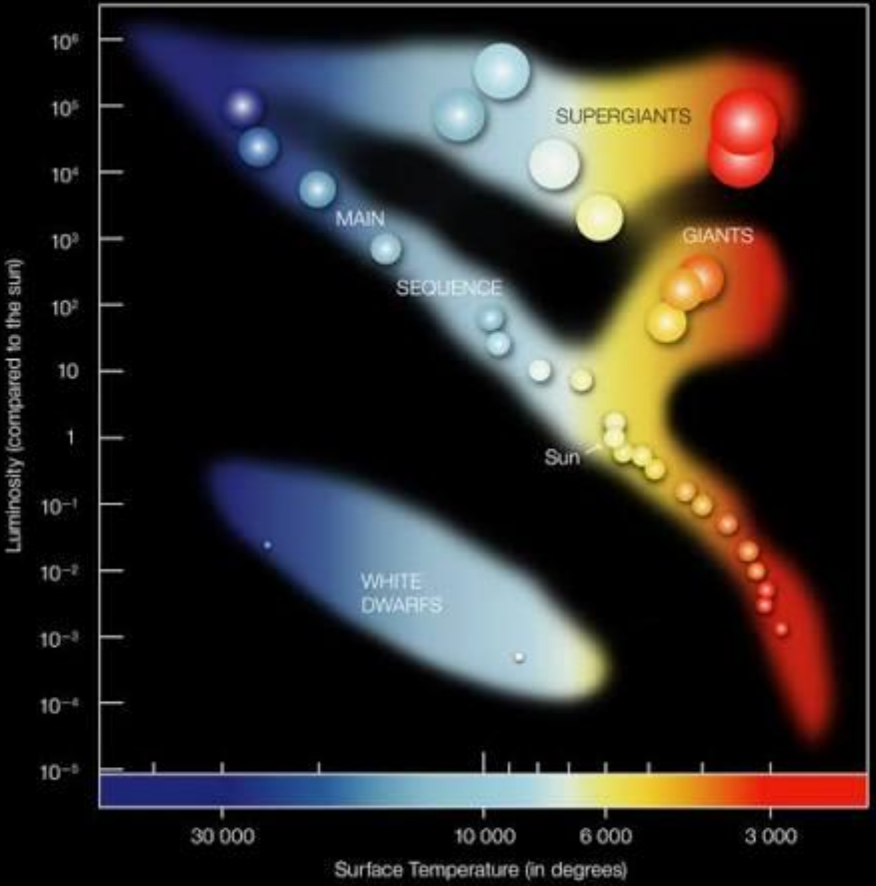


Image credit: [Chandra :: Educational Materials :: Pulsating Variable Stars and the Hertzsprung-Russell Diagram]

Wikipedia contributors. (n.d.). *Red giant*. In *Wikipedia*. Retrieved July 20, 2026, from https://en.wikipedia.org/wiki/Red_giant

Red Giant Stars

Red giants are stars that exhaust the hydrogen in their core and have now started to burn it in its outer layers. Eventually these stars have a helium flash and burn helium in their core.

Red giants are significant to galactic archaeology.

- They are very bright, allowing us to see them from large distances
- Allow us to map out the structure of the Milky Way
- Key observables in understanding how our Galaxy has formed and evolved

However there are two different types of Red Giants.

Wikipedia contributors. (n.d.). *Red giant*. In *Wikipedia*. Retrieved July 20, 2026, from https://en.wikipedia.org/wiki/Red_giant

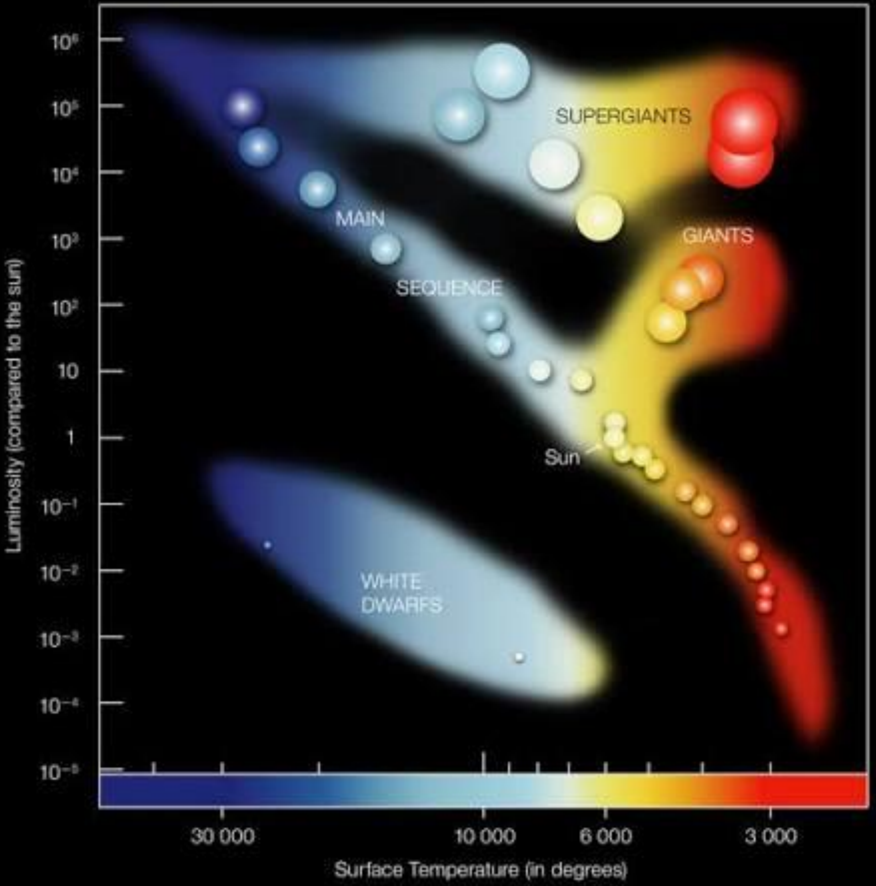


Image credit: Chandra :: Educational Materials :: Pulsating Variable Stars and the Hertzsprung-Russell Diagram

Red Giant Branch vs Red Clump Stars

The stars will pulsate differently. Best classification method is through Asteroseismology

Red Giant Stars

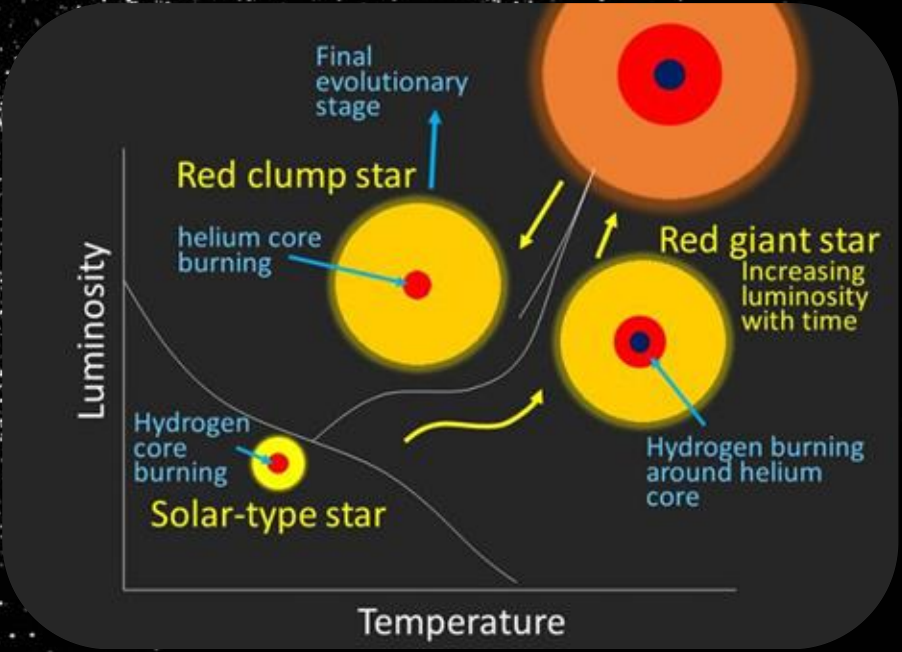
RGB

Hydrogen burning shell
Low period spacing

Red Clump Stars

RC

Star that ignites helium fusion in their cores resulting in a “helium flash”
High period spacing



Hawkins, K., Ting, Y.-S., & Rix, H.-W. (2017). Photospheric diagnostics of core helium burning in giant stars. arXiv e-prints, arXiv:1712.02405 (Accepted to The Astrophysical Journal).

Red Giant Branch vs Red Clump Stars

Best classification method through Asteroseismology (different t=lum and diff ages)

Red Giant Stars

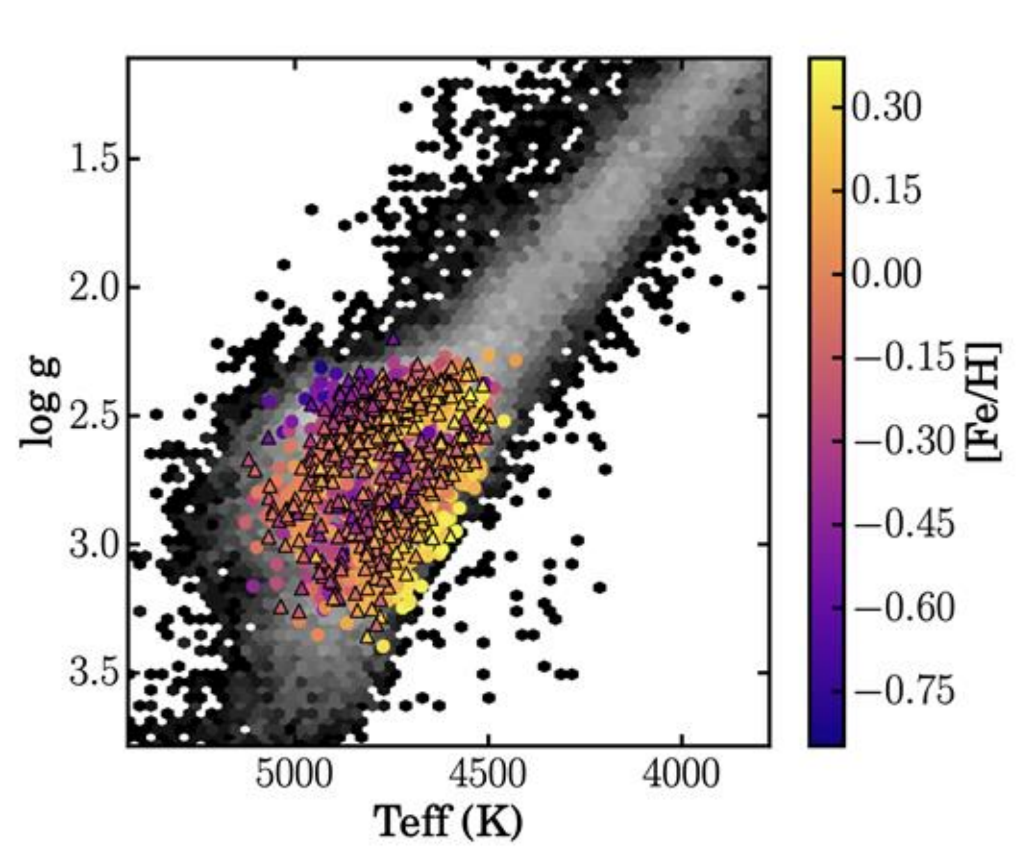
RGB

High mass Hydrogen burning shell
Low period spacing

Red Clump Stars

RC

High-Mass star that ignites helium fusion in their cores resulting in a “helium flash”
High period spacing

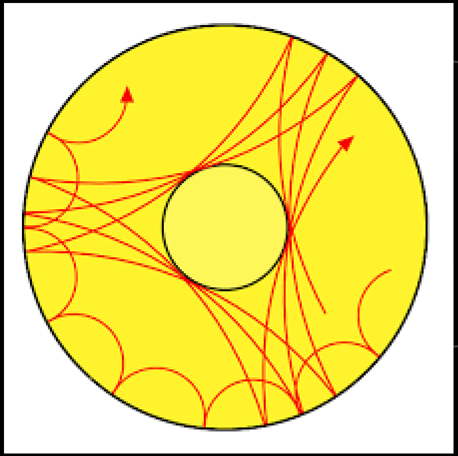


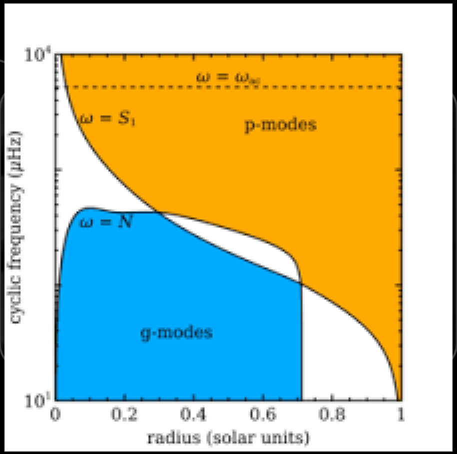
Hawkins, K., Ting, Y.-S., & Rix, H.-W. (2017). Photospheric diagnostics of core helium burning in giant stars. arXiv e-prints, arXiv:1712.02405 (Accepted to The Astrophysical Journal).

Asteroseismology

- Oscillations in stars from near-surface convection can have either or both p-mode or g-mode characteristics
- Stronger coupling between g-mode and p-mode leading to appearance of observable mixed modes in oscillation spectrum
    - This can be caused by RC stars core having lower density than a RGB star

Period spacing are the time intervals between a star's mixed modes (p-mode and g-mode), vibrations that combine the characteristics of sound waves and gravity waves inside the star.





P-mode

Associated with stellar envelope with pressure as a restoring force

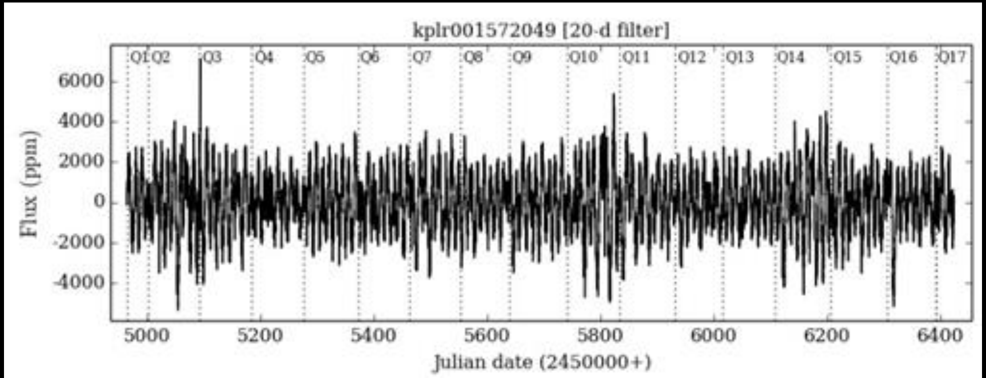
G-mode

Probe its core with buoyancy as a restoring force.

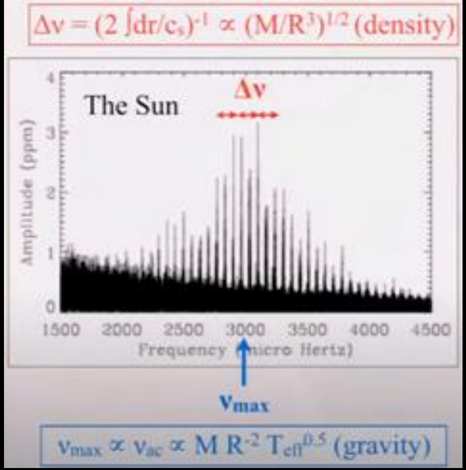
- Bedding Tim, Huber Daniel. "ASTR 5970 Lecture 16-Seismology." https://docs.google.com/presentation/d/1n4gnIVjk0QskgZYAVGrrYui_9yyZ9EoGr3Z9NKOY4KY/edit?slide=id.g2feb84dece6_0_122#slide=id.g2feb84dece6_0_122
  Ness et al. 2016, "Spectroscopic Determination of Masses (and Implied Ages) for Red Giants," *The Astrophysical Journal*, 823:114 https://iopscience.iop.org/article/10.3847/0004-637X/823/2/114/pdf

How We Get Data

- Asteroseismology uses the frequency of stellar oscillations which creates very small changes in the luminosity of a star, probing the interior of the star
- Taking a fourier transform of the light curve provides a power spectrum for oscillations



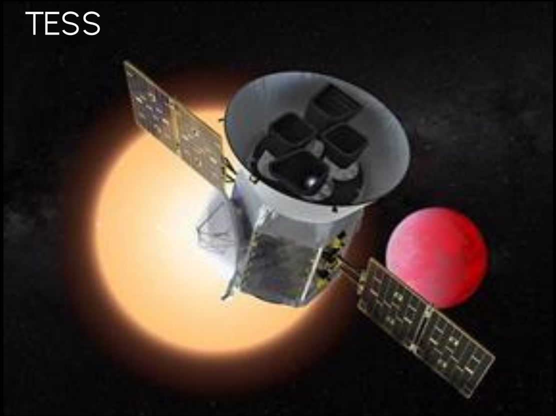
Light Curve for a Kepler Star



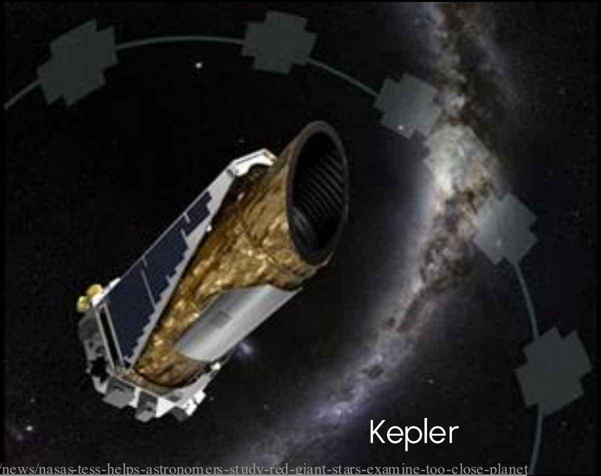
Solar Oscillation Power Spectrum

- Star Clusters: Inside the Universe's Stellar Collections - NASA Science

Kepler and TESS



- Kepler’s ultra-sensitive photometers capture the rhythmic pulses of the stars brightness.
- TESS (Transiting Exoplanet Survey Satellite) has a secondary mission for gathering continues brightness measurements of stars.

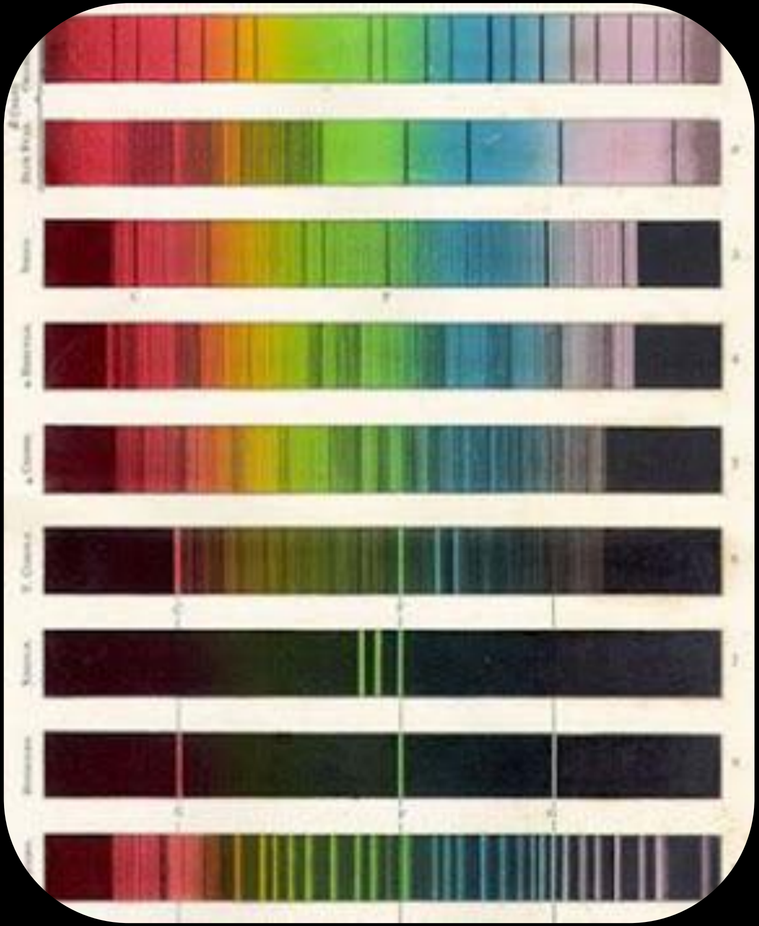


Krapfl, M. (2019, November 19). *NASA's TESS helps astronomers study red-giant stars, examine a too-close planet*. Iowa State University News Service. https://www.news.iastate.edu/news/nasas-tess-helps-astronomers-study-red-giant-stars-examine-too-close-planet
Cermak, A. (2026, January 5). *Kepler / K2*. NASA Science. https://science.nasa.gov/mission/kepler/
Brill, J. (2026, February 10). *Asteroseismology using the Galactic Bulge Time Domain Survey*. NASA Science. https://science.nasa.gov/mission/roman-space-telescope/asteroseismology-using-the-galactic-bulge-time-domain-survey/
Huber, D., Chaplin, W. J., Christensen-Dalsgaard, J., Gilliland, R. L., Kawaler, S. D., Kjeldsen, H., & Working Groups of the Kepler Asteroseismic Science Consortium (KASC). (2013). Asteroseismology with NASA's Kepler mission. *American Astronomical Society Meeting Abstracts*, *221*, 301.01. https://ui.adsabs.harvard.edu/abs/2013AAS...22130101H/abstract

APOGEE

A large-scale, stellar spectroscopic survey conducted in the near infrared portion of the electromagnetic spectrum.

The survey provides spectra that can then be used to derive effective temperature, surface gravity, and metallicity of the observed stars emission and absorption lines.

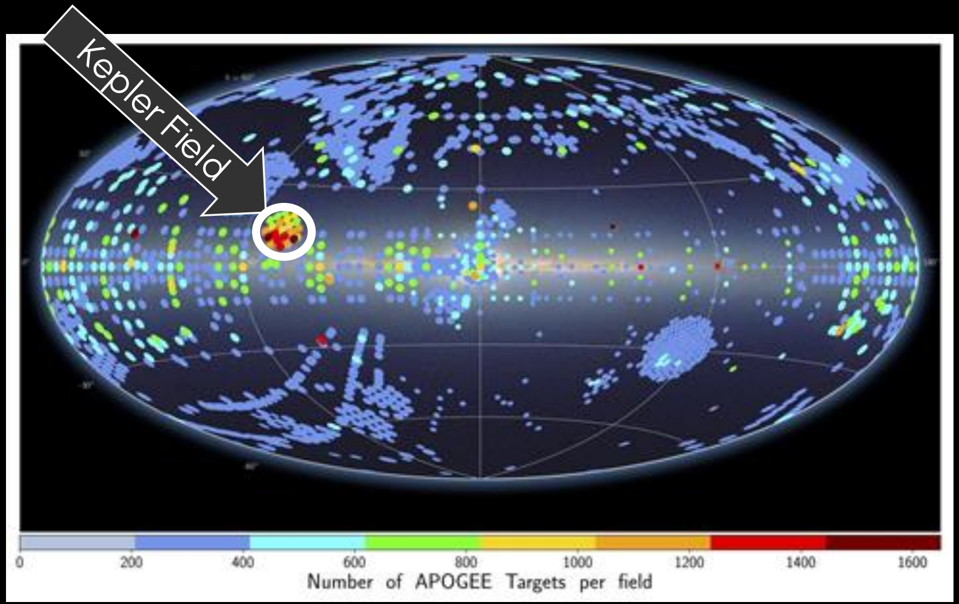


Sloan Digital Sky Survey. (n.d.). *APOGEE*. SDSS Data Release 17. Retrieved July 20, 2026, from

Main Problem

- There are millions of spectral data, with emission and absorption lines containing information about temperature, surface gravity, elemental abundances, and more
  - Goal: link spectra and seismic data

- This however this does not help us when it comes to differentiating between stars in the Red Giant Branch and in the Red Clump of the HR diagram
  - More luminous therefore more data
  - These are two different stages in a giants life cycle.
  - Important for: Stellar ages and Mass loss

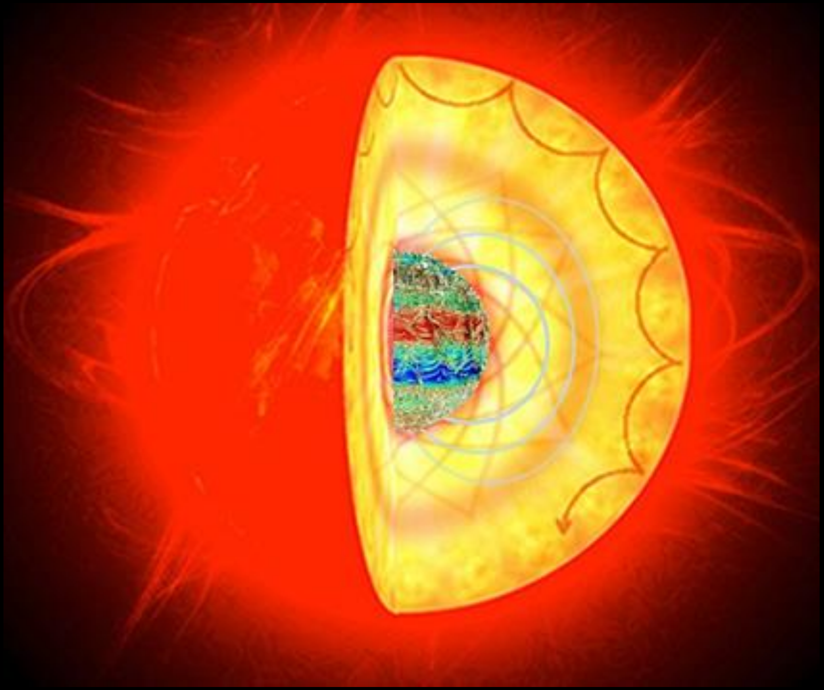


APOGEE observation fields on an all-sky image. Each field is color-coded by the number of targets in that field.

Main Problem

- Mass can be estimated through seismic scaling relations in asteroseismology
- Age is one of the hardest characteristics to determine about a star
- Different masses (along with their evolutionary phases) can provide information on how fast stars burn their fuel, giving a huge constraint on stellar age

Can we use spectra to encode this information?



$$M = \left(\frac{\nu_{\max}}{\nu_{\max,\odot}}\right)^{3}\left(\frac{\Delta\nu}{\Delta\nu_{\odot}}\right)^{-4}\left(\frac{T_{\text{eff}}}{T_{\text{eff},\odot}}\right)^{1.5}.$$

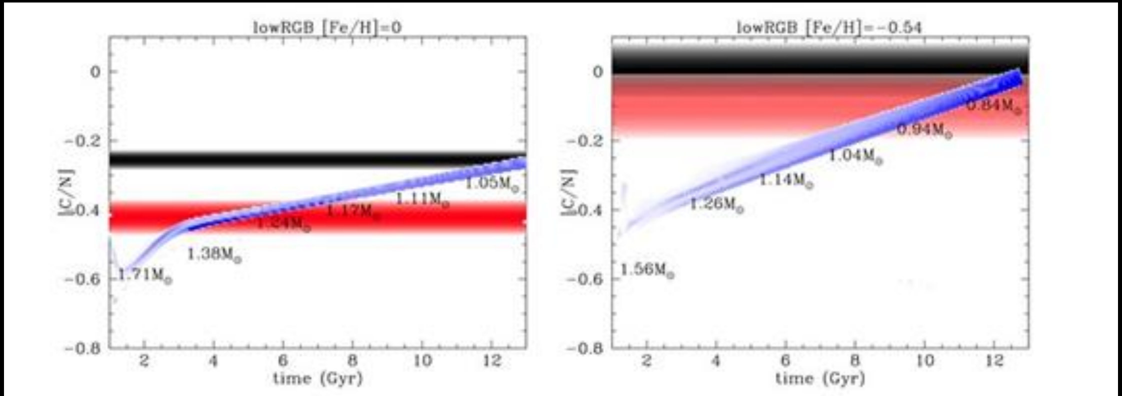
$$\tau \approx 10^{10}\left(\frac{M_{\odot}}{M}\right)^{2.5} yrs$$

- Ness, M., Hogg, D. W., Rix, H.-W., Martig, M., Pinsonneault, M. H., & Ho, A. Y. Q. (2016). Spectroscopic determination of masses (and implied ages) for red giants. The Astrophysical Journal, 823(2), 114.

Image credit: Rafael A.García (SAp CEA),Kyle Augustson (HAO), Jim Fuller (Caltech) & Gabriel Pérez (SMM,IAC), Photograph from AIA/SDO.

Other Indicators and Methods

- Dredge-ups mix CN-processed material to the surface, lowering [C/N] over time
- Different evolutionary states will have different types of CN mixing due to things such as a helium flash
- Age can also be estimated through gyrochronology or isochrone fitting
- Asteroseismology is a better method, reducing uncertainty gaps that are up to 1-2 GYA



- Masseron, T., & Gilmore, G. (2015). "Carbon, nitrogen and α-element abundances determine the formation sequence of the Galactic thick and thin discs." *Monthly Notices of the Royal Astronomical Society*, **453**(2), 1855–1866.

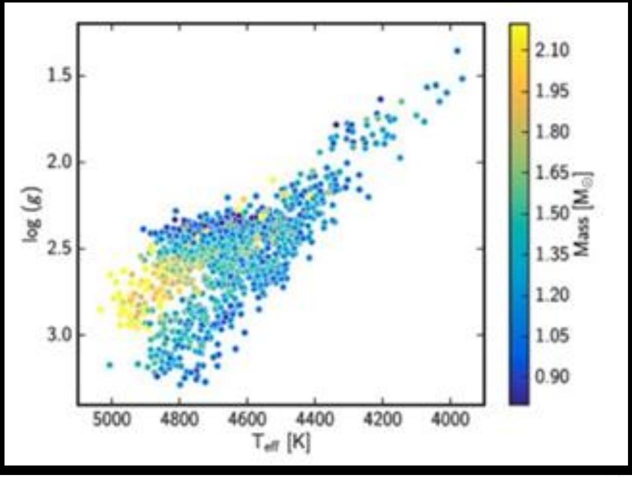
TheCannon

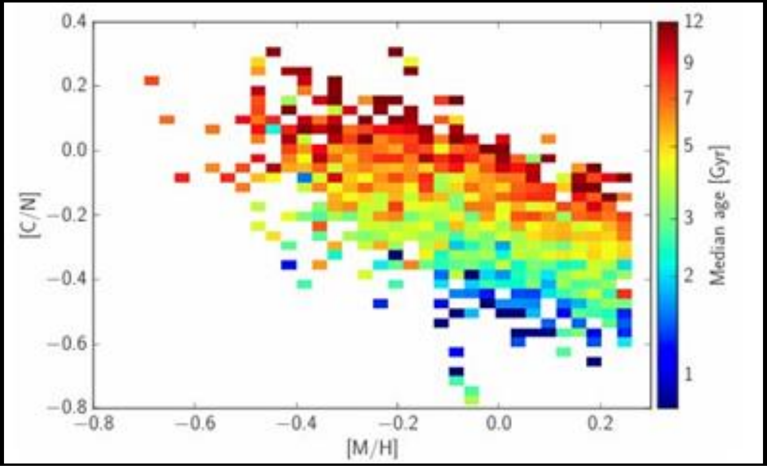
TheCannon is a machine learning tool that can learn the relationship between a star's spectra and its known physical properties
The training set is Kepler seismic data and APOGEE spectra

Steps of use:

1. Get and format data
2. Use reference set to fit a spectral model (training)
3. Make sure the model is sensible, perform a leave-n-out cross-validation
4. Then use it to infer labels from all the other spectra in the dataset (testing)

Once seismic masses are found for a few stars, those stars can then train stars with no mass data





- M. Martig et al. 2015 Red giant masses and ages derived from carbon and nitrogen abundances

Goal of TheCannon

Spectra should therefore be able to encode information regarding the evolutionary states, masses, and ages of stars

Machine Learning Process

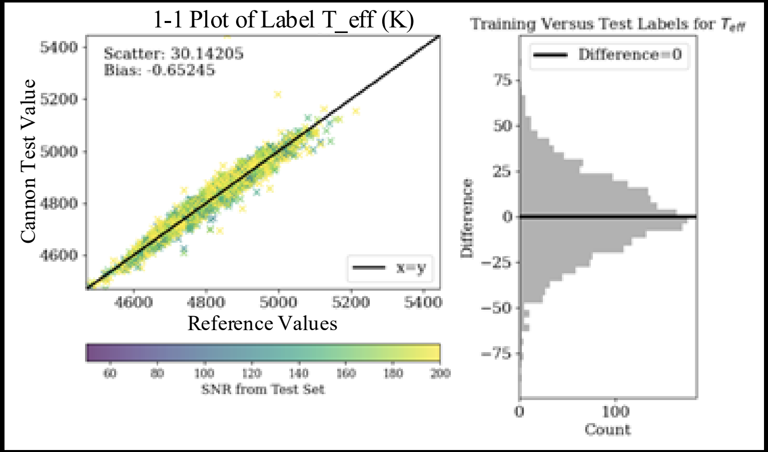
Training data with known spectral and seismic data → TheCannon fits a generative model, learning how each pixel fits each label → A new trained model is produced

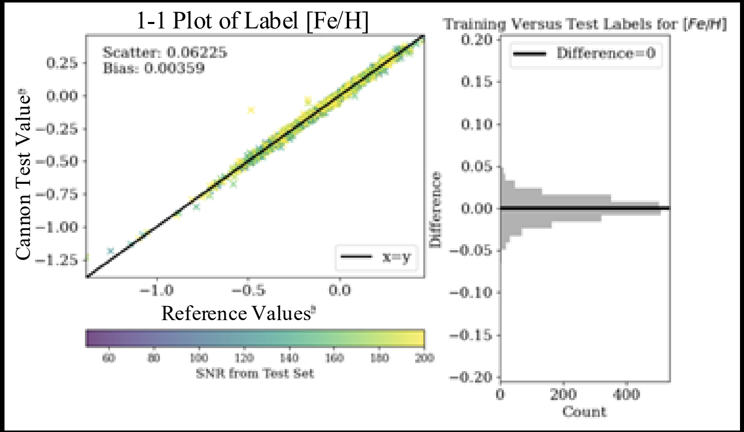
Data with unknown labels is introduced → "Which labels best reproduce this star using the trained model?" → TheCannon infers new labels for this new data

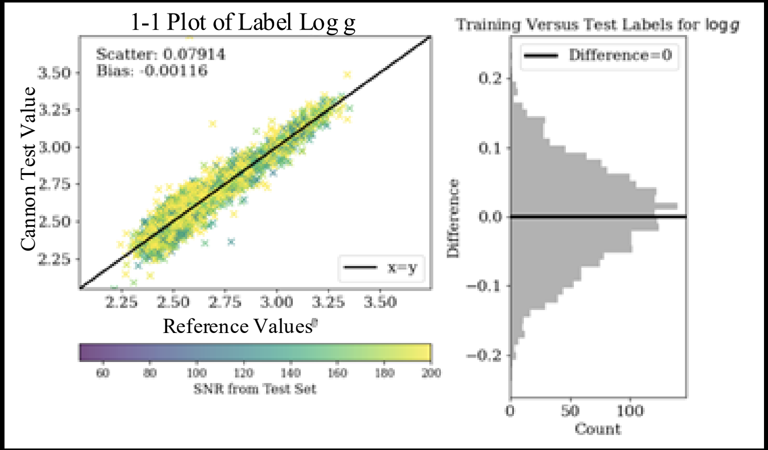
Results

Some of the 1-to-1 plots from TheCannon of labels that I chose to train it on to help constrain the label estimates of period spacing to be more accurate.

Trained on effective temperature, surface gravity (log scaled), and iron and magnesium ratio from the stars spectra. This is how well TheCannon predicted these labels.

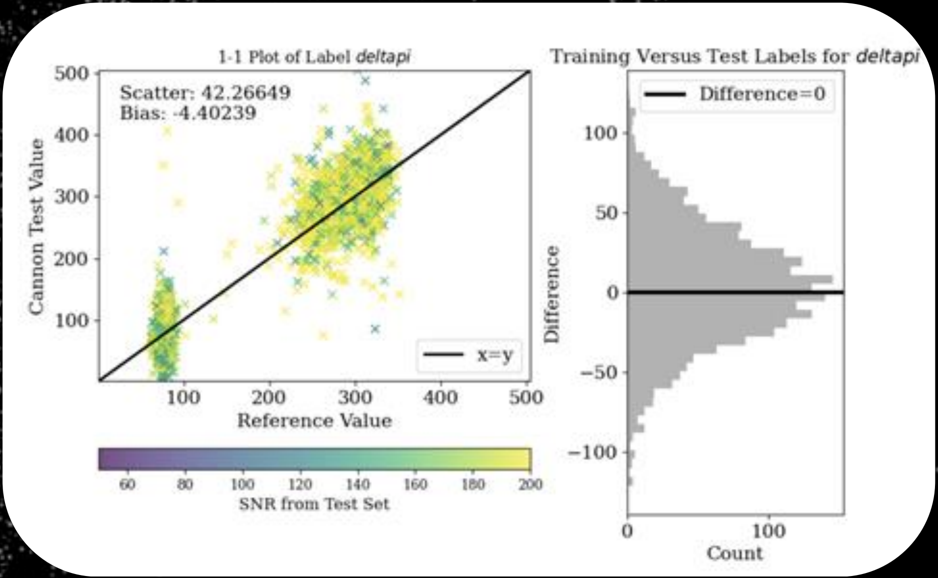






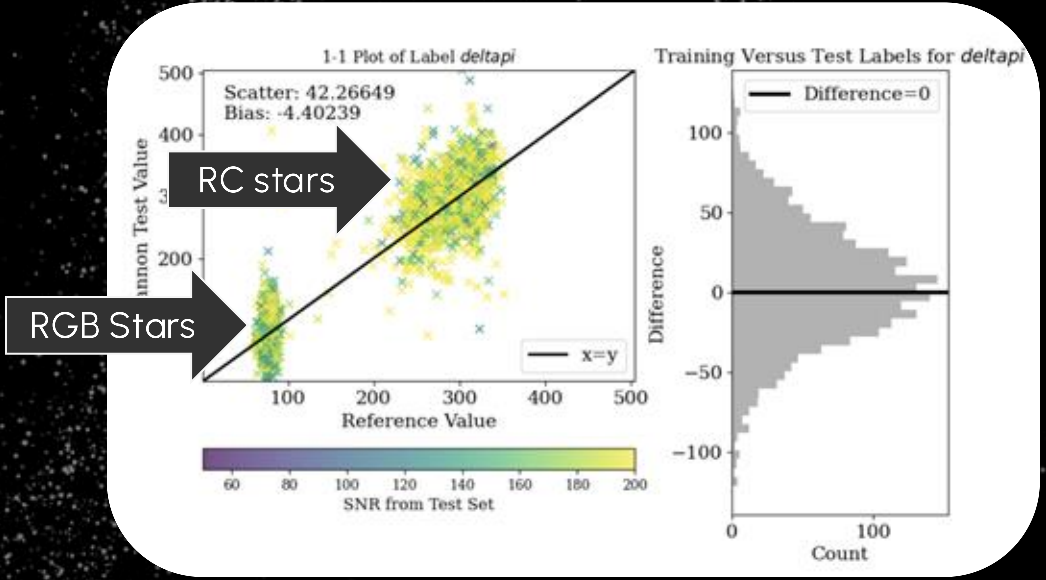
Results

I was able to end up with an accurate model with the original cannon model where the separation between the RGB and RC stars being pretty evident and the histogram presenting like a gaussian. However the scattering of this one is bad.



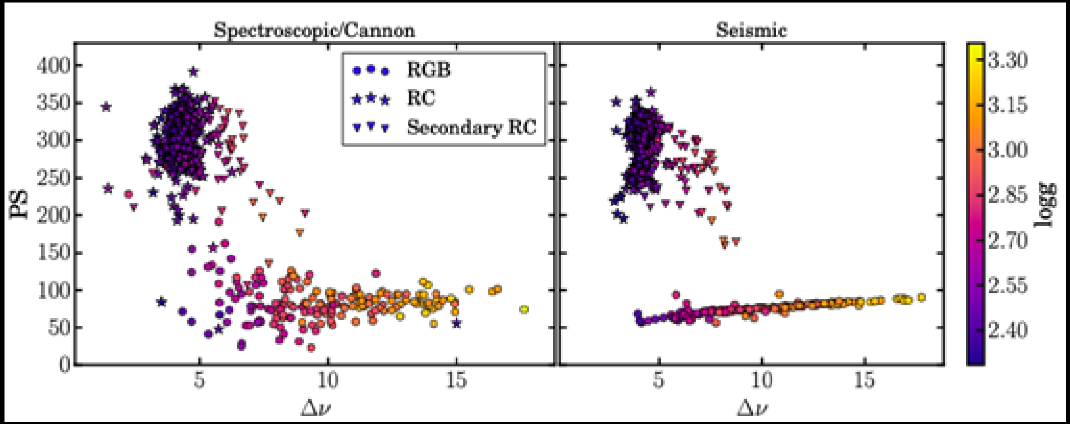
Results

I was able to end up with an accurate model with the original cannon model where the separation between the RGB and RC stars being pretty evident and the histogram presenting like a gaussian. However the scattering of this one is bad.



Results Right After Training

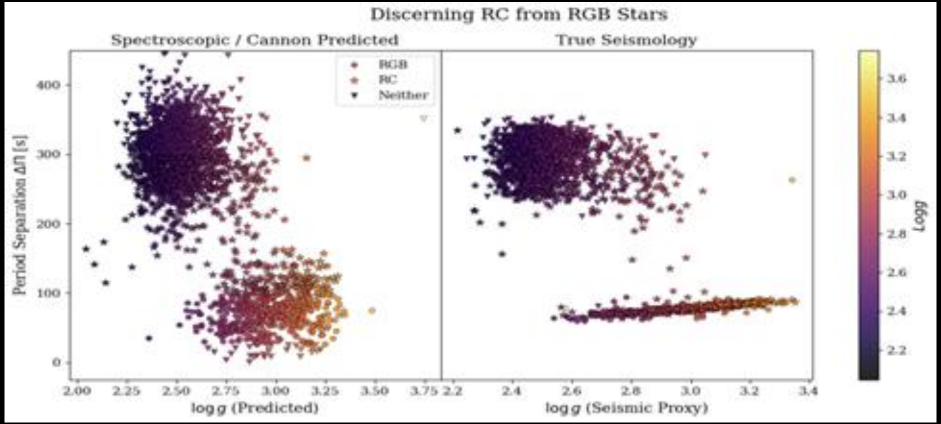
I then tried to recreate a plot from Hawkins to see how well TheCannon model can predict the labels for period spacing. Unlike the the papers, TheCannon models our predictions a little less accurate as shown with how low some of the values for period spacing are.



Predicted RGB: 413 True num RGB: 582

Predicted RC: 944 True num RC: 840

169 under RGB and over 104 for RC

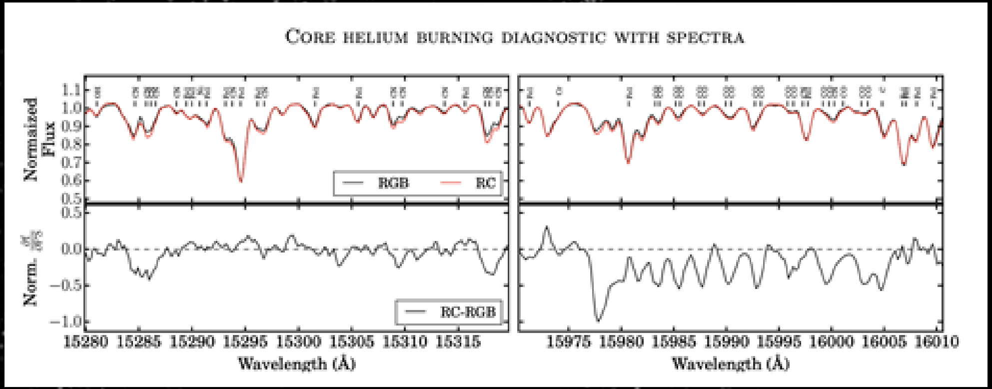


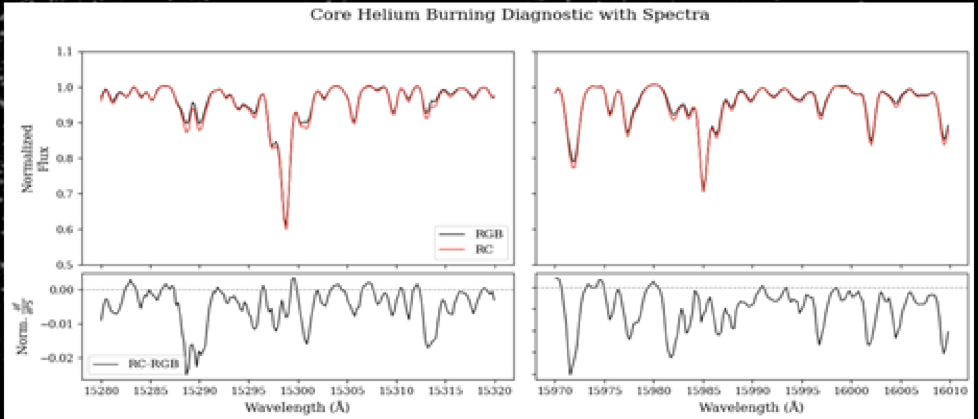
Hawkins, K., Ting, Y.-S., & Rix, H.-W. (2017). Photospheric diagnostics of core helium burning in giant stars. arXiv e-prints, arXiv:1712.02405 (Accepted to The Astrophysical Journal).

Why does this work?

The spectral region area shows that the differences between RGB and RC stars lie mostly around molecular CN and CO line features.

- Likely a result of extra mixing from the red giant branch bump to the helium flash causing RC stars to be lower in their carbon to nitrogen ratio compared to RGB stars of similar parameters.





Hawkins, K., Ting, Y.-S., & Rix, H.-W. (2017). Photospheric diagnostics of core helium burning in giant stars. arXiv e-prints, arXiv:1712.02405 (Accepted to The Astrophysical Journal).

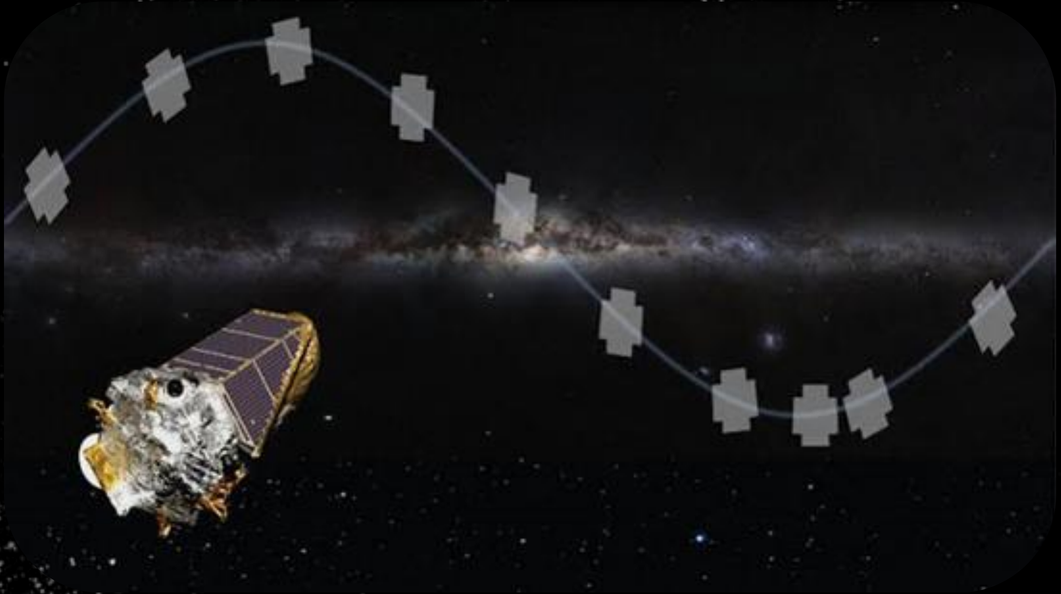
K2 Data

- Kepler mission was repurposed into Kepler 2, using the pressure of sunlight to maintain observations of different fields along the ecliptic plane
- Would greatly increase training set size for asteroseismic data

NASA/Ames Research Center/W. Stenzel/D. Rutter

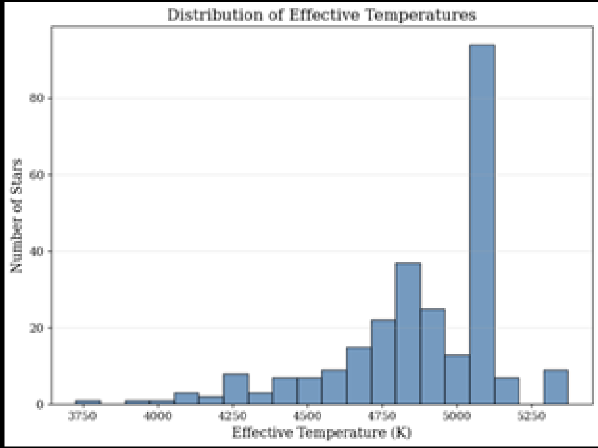
Now going further

We plan to compare TheCannon's predicted labels with the data that the K2 telescope found to see if TheCannon can predict these labels more accurately rather than a short (3 month long) observation mission. We can use our current model to check how well this works!

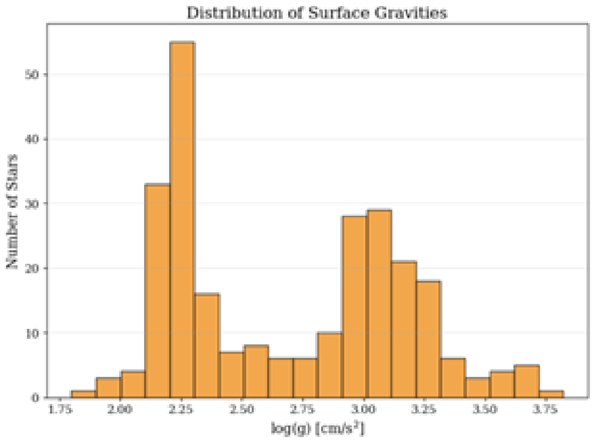


- Cermak, A. (2026, January 5). *Kepler / K2*. NASA Science. https://science.nasa.gov/mission/kepler/

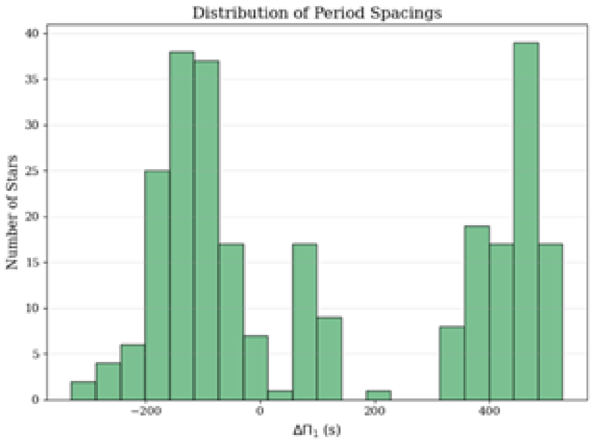
Results with new data



Histogram over effective temperature and the number of stars. This shows us that the model can predict the temperature of the stars and it's what we would expect.

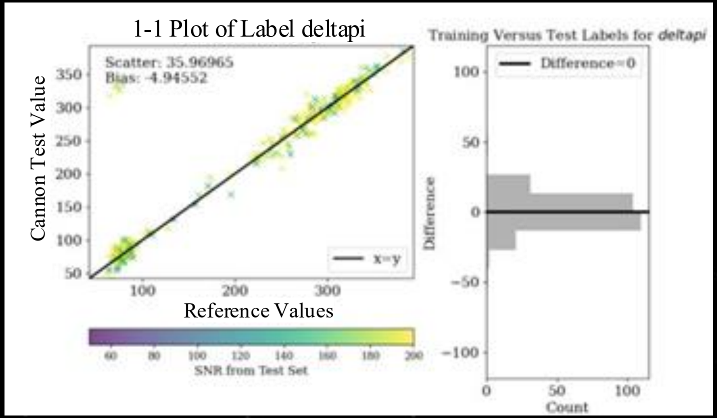


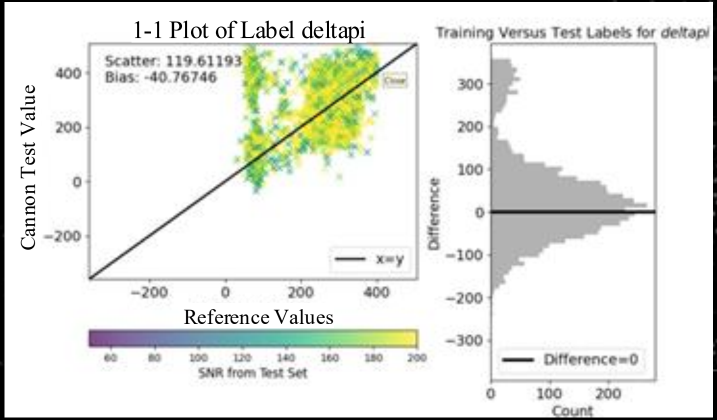
Histogram over surface gravity and the number of stars. This tells us that the model can predict the surface gravity of the star quiet well. (too high, need no more than 3.5~3.4) (plot good in gray and bad in color on an hr plot)

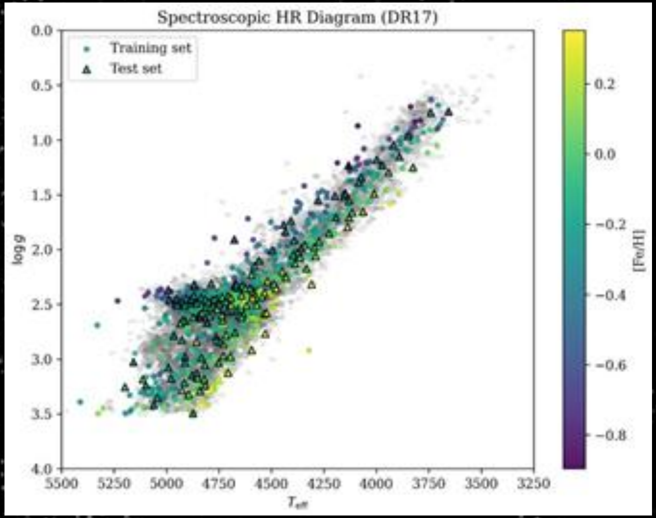


This histogram shows the different period spacing vs the number of stars and with so many stars located in a negative range of period spacing this tells us that this data set does not just contain RGB or RC stars

Issues with Previous Model

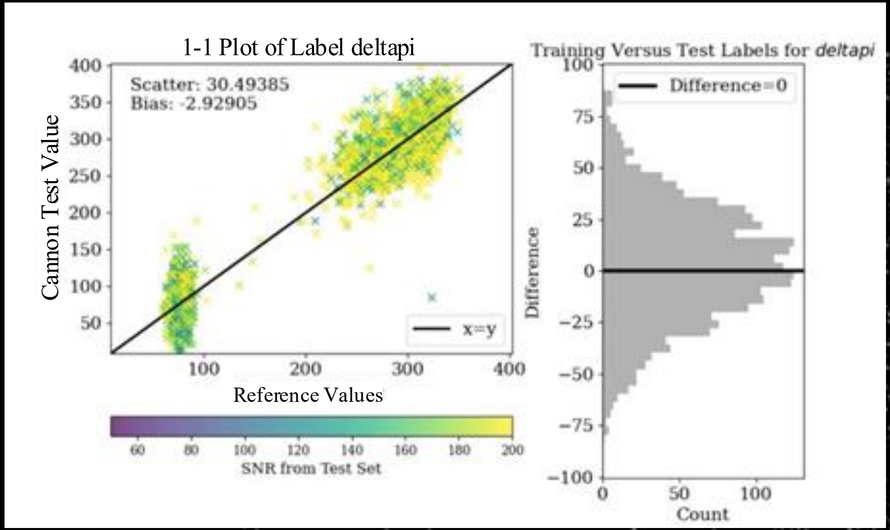






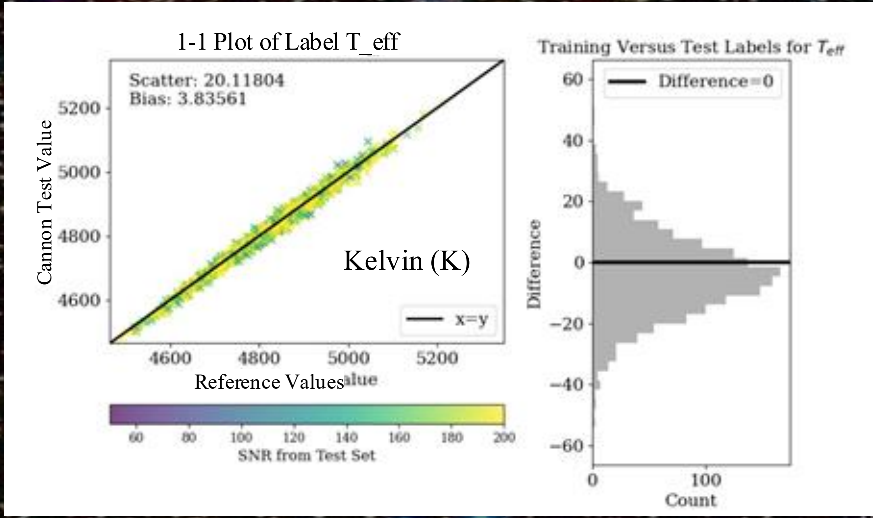
- Model was previously training on the same set of stars
- Fundamental issues when trying to apply the Kepler data
- 20% of each evolutionary class used test set and not used during training and found scatter/bias on it to get an honest estimate

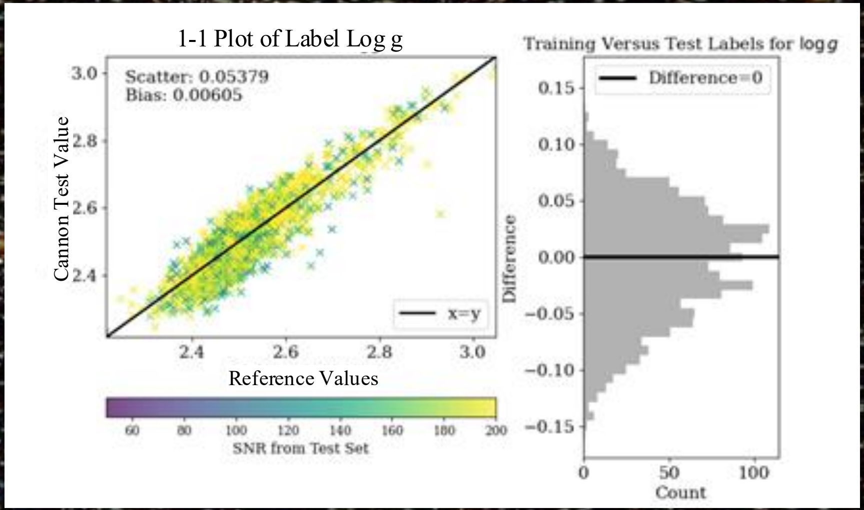
Updated Model

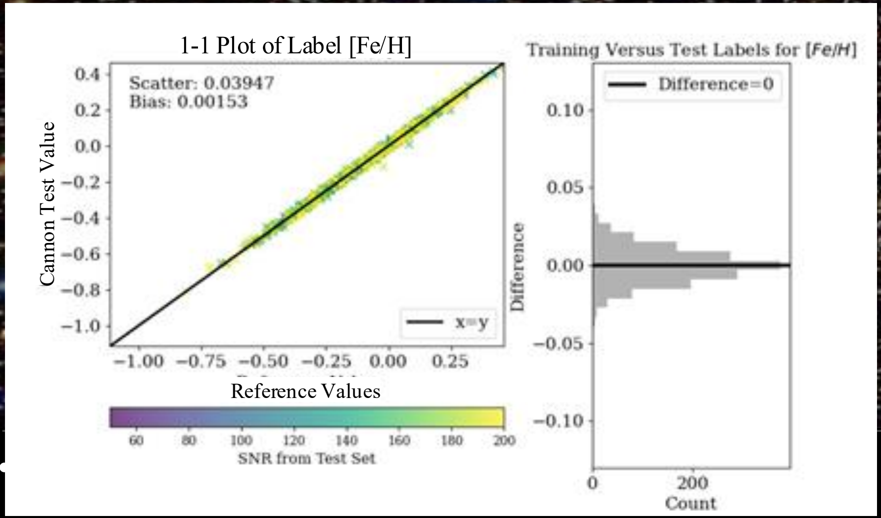


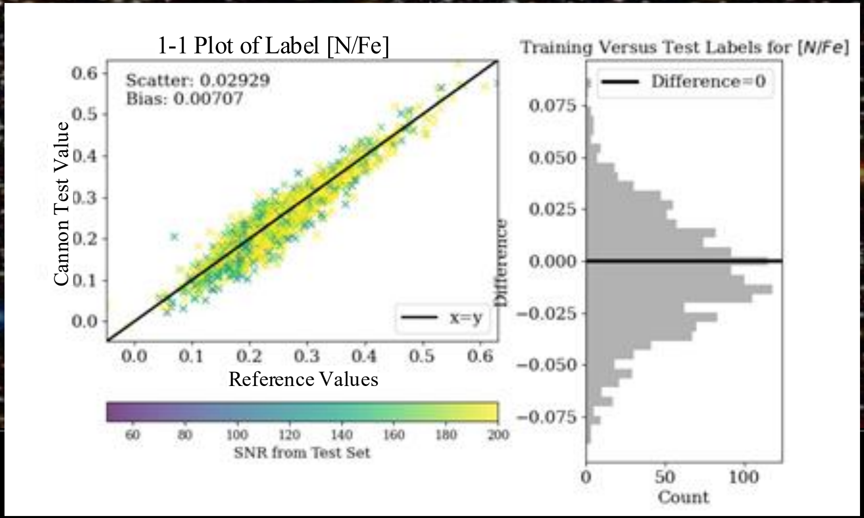
- Fixed SNR mismatching
- Removed data limits
- Stricter normalization step for flux at each pixel ➡ Helps TheCannon's model be more accurate
- Swapped a label matrix ID look up with a data table

Yay!

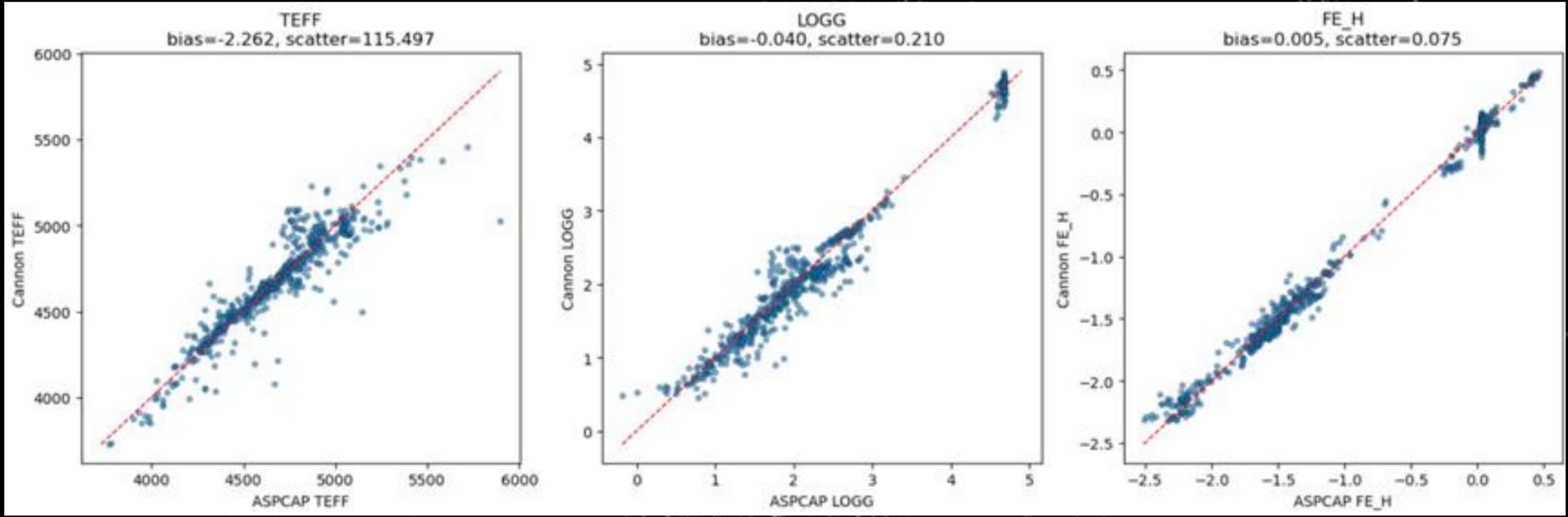

1-1 Plot of Label T_eff
Scatter: 20.11804
Bias: 3.83561
Kelvin (K)
x=y
Cannon Test Value
Reference Values
SNR from Test Set
Training Versus Test Labels for T_{eff}
Difference=0
Difference
Count


1-1 Plot of Label Log g
Scatter: 0.05379
Bias: 0.00605
x=y
Cannon Test Value
Reference Values
SNR from Test Set
Training Versus Test Labels for $\log g$
Difference=0
Difference
Count


1-1 Plot of Label [Fe/H]
Scatter: 0.03947
Bias: 0.00153
x=y
Cannon Test Value
Reference Values
SNR from Test Set
Training Versus Test Labels for [Fe/H]
Difference=0
Difference
Count

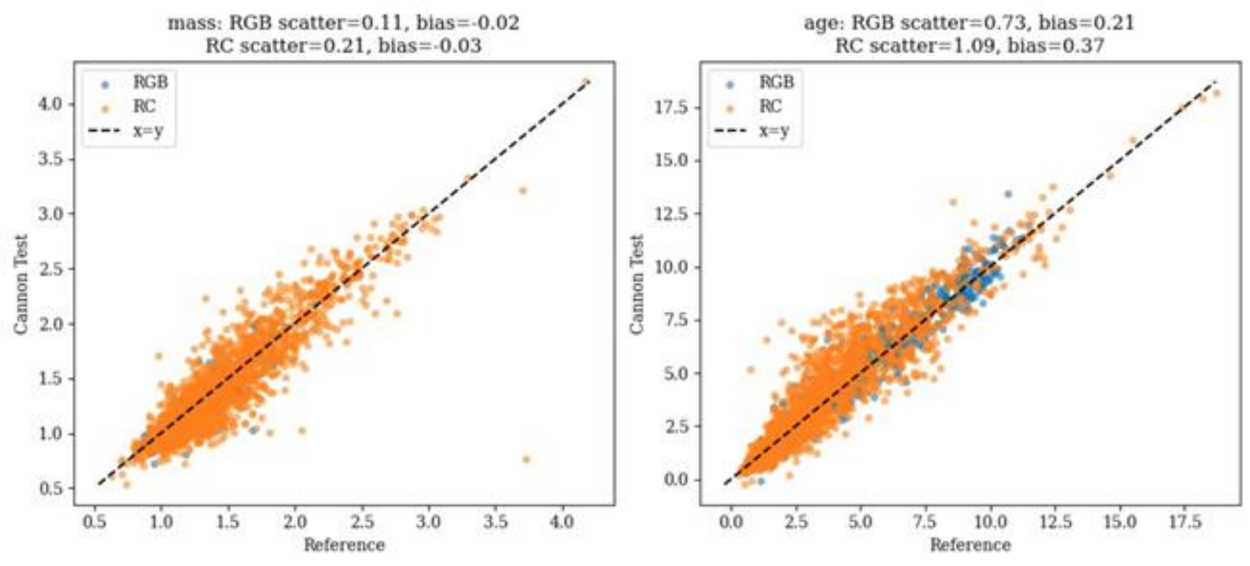

1-1 Plot of Label [N/Fe]
Scatter: 0.02929
Bias: 0.00707
x=y
Cannon Test Value
Reference Values
SNR from Test Set
Training Versus Test Labels for [N/Fe]
Difference=0
Difference
Count

TheCannon 2

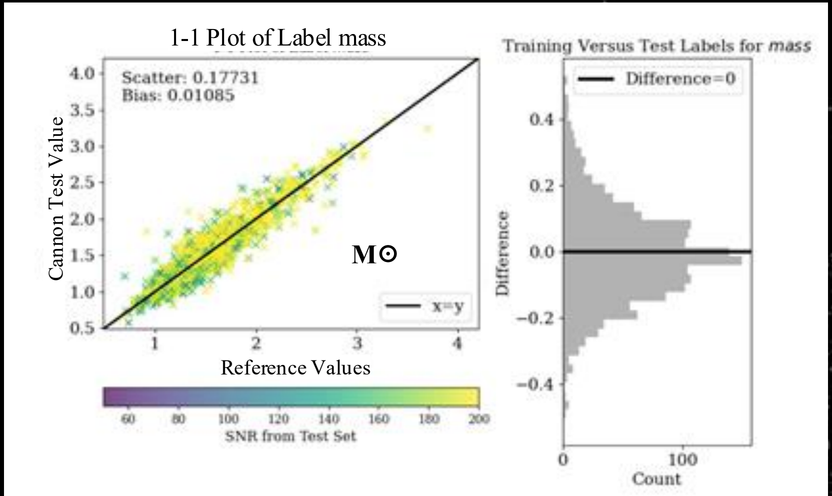


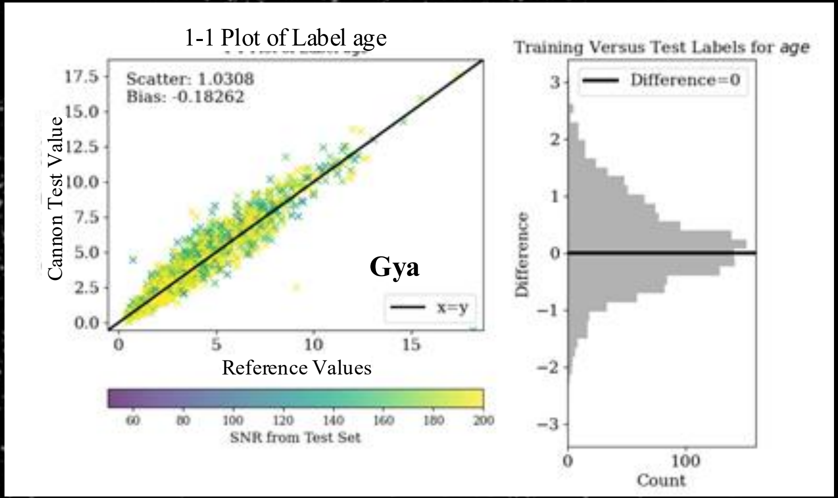
- Written by Andy Casey
- Able to predict more labels without overfitting
- Only uses a coefficient if it improves the overall fit via regularization

Training Mass and Age



- Mass and age data encoded in the asteroseismic data
- Created an evolutionary split to separate RC and RGB stars
- Incorporated split into data mask
- Train with separate models, [C/N] trends for RC and RGB are not the same, so mixing the sample could increase scatter

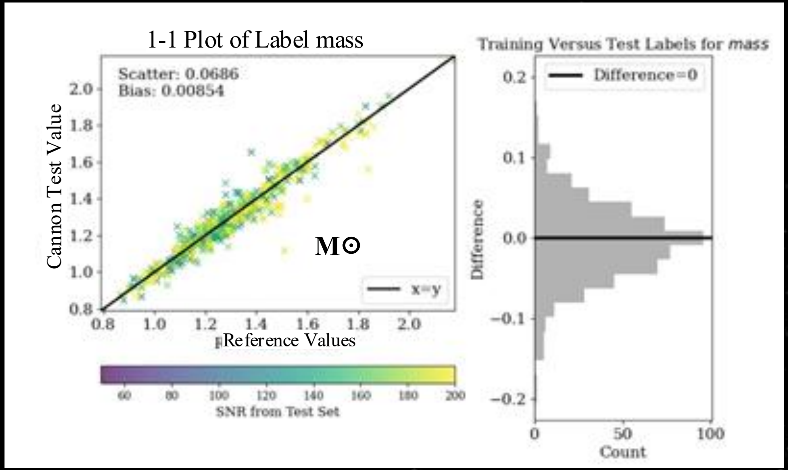


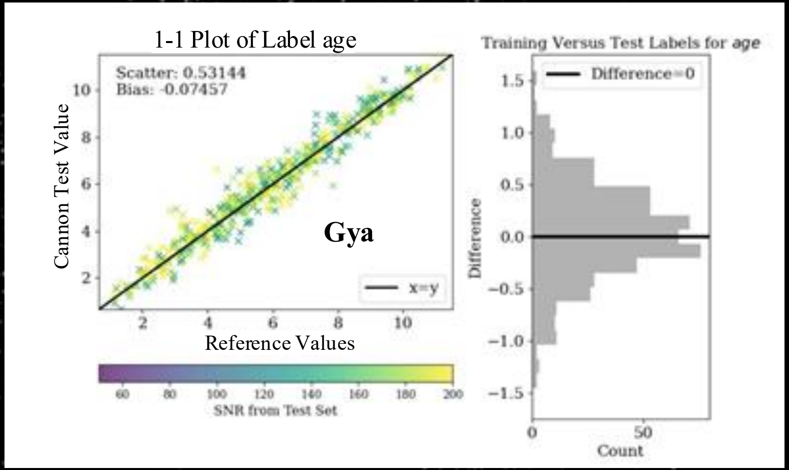


RC Model

- Mass scatter dropped 19%
- Age scatter dropped 6%

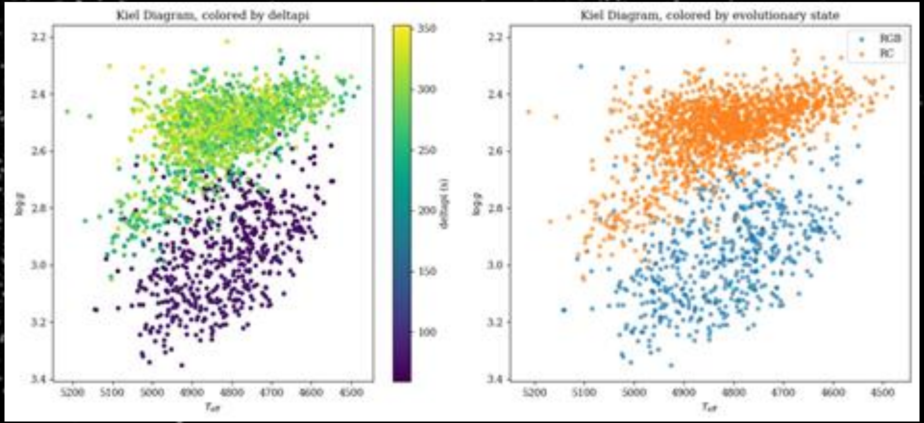
Splitting into separate RC and RGB models reduced scatter for both populations





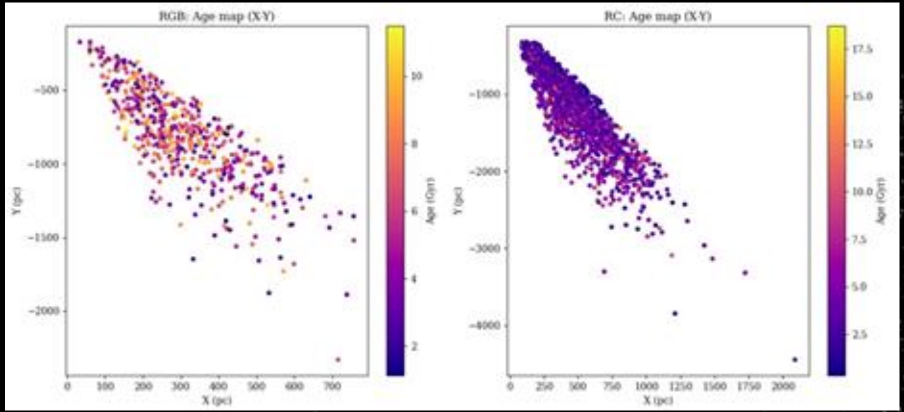
RGB Model

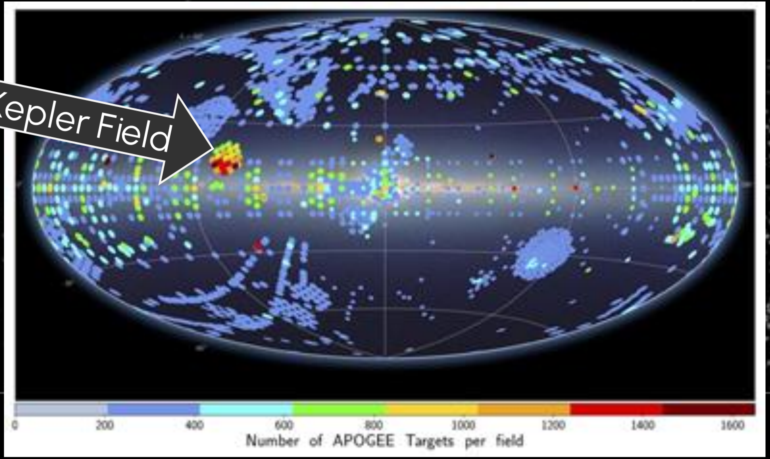
- Mass scatter dropped 46%
- Age scatter dropped 27%



Kiel diagram shows similarities of RC and RGB stars in logg and temperature space but easily separable by distinguishing deltapi

Stellar Age Map

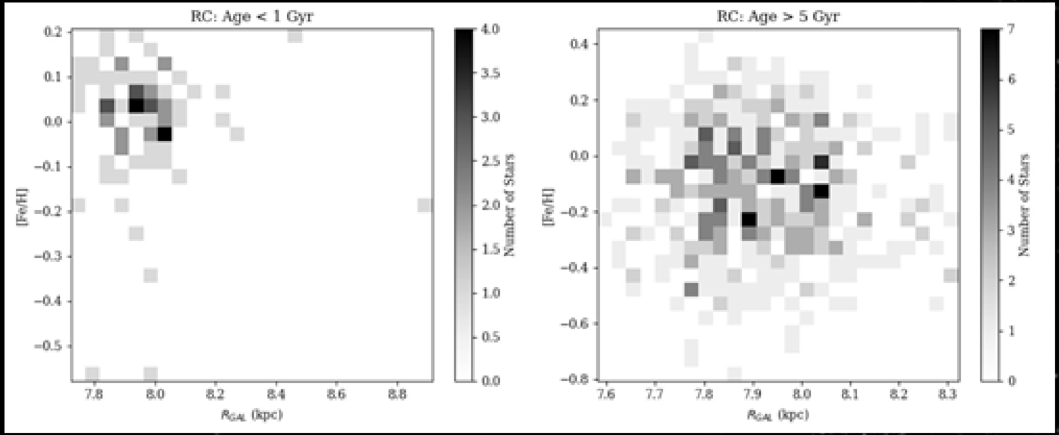


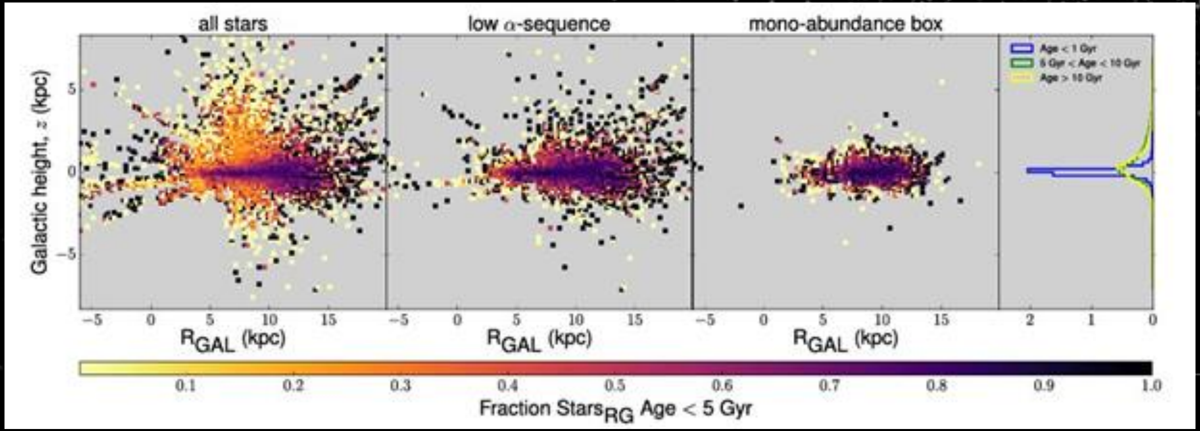


- Using the stars positions (right ascension, declination, and distance via parallax)
- Does not reveal too much for this specific dataset

Kepler stares at a small fixed patch of sky, while the stars all sit in a narrow cone pointing in one direction from the Sun

Expansion into the Milky Way





- Multiple fields adds more asteroseismic giants, directly reducing the noise
- Broader sky coverage = Galactic mapping
- Kepler data combined with all APOGEE stars can turn a single-field plot into an actual map of galactic features in relation to age with 5 times the data than before

Ness, M., Hogg, D. W., Rix, H.-W., Martig, M., Pinsonneault, M. H., & Ho, A. Y. Q. (2016). Spectroscopic determination of masses (and implied ages) for red giants. The Astrophysical Journal, 823(2), 114.

Summary

- Created a working model of TheCannon
- Applied asteroseismic data from Kepler to APOGEE spectra
- Distinguished Red Giant Branch and Red Clump stars based on period spacing
- Predicted whether stars are RGB or RC stars
- Predicted mass and age data by training separate an RGB and RC model
- Created age and spatial maps

- Moving forward, applying Kepler 2 data and the larger APOGEE catalog to our trained model

Questions?