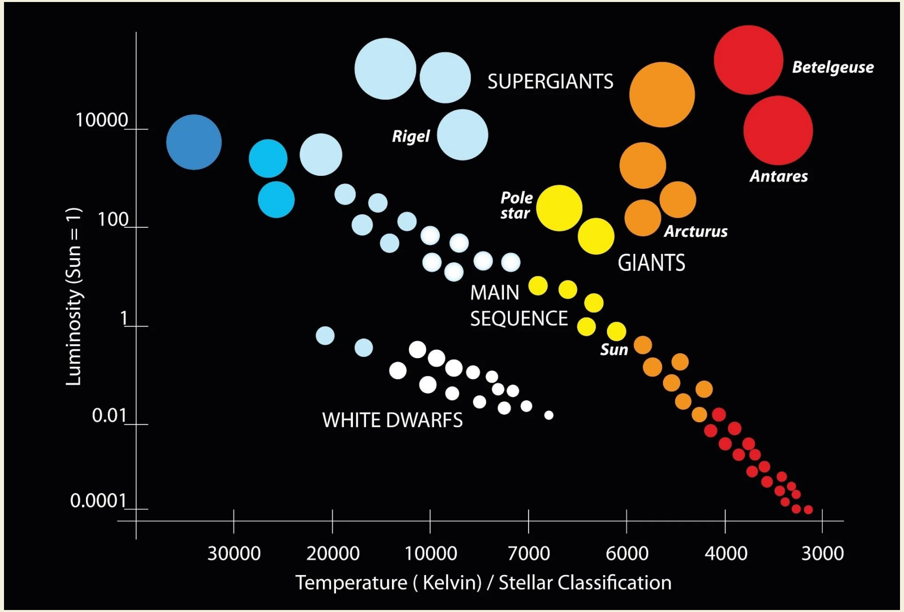
12/08/24

Clustering for Star Classification

Jaxon Fielding
Leo Weimer

Background / Motivation

- HR diagram visually shows types of stars by graphing temperature vs. luminosity.
- Clustering stars based on these variables can provide an algorithm for star classification, eliminating hard boundaries and manual classification.
- Classifying stars is important to know what they will do in the future, and by extension what experiments can be done and what can be revealed by monitoring them.



Data Sources

- Data was originally collected from Gaia (bottom left).
- When Gaia data proved too incomplete to use, we found a source that filled in the blanks using astrophysics equations (bottom right).

#	# RA_ICRS	# DE_ICRS	# Source	# e_RA_ICRS	# e_DE_I
0 – 100.0k (85999.14 - 87999.12 Count: 12,000)	0 – 360	-89.1 – 89.8	38.7b – 6916186185b	0 – 0.8	0
0	44.58901169477	2.19529765977	1306361548360576	0.0655	0.0612
1	35.35803466314	8.98881281137	23700286669971584	0.0658	0.0717
2	44.450766685059996	10.079117671110001	27109837867995776	0.0627	0.0517
3	48.40490939026	15.10591245631	31009771252186752	0.05	0.0421
4	57.092837839350004	11.55092659635	36876009385300352	0.0521	0.0335
5	60.276284204259994	20.47152883876	50654882946632832	0.0286	0.0171
6	51.36671855962	22.840387256210004	62631897466925184	0.0941	0.0852
7	55.654786835390006	20.022519645179997	63139734399542528	0.0464	0.0328
8	28.985641515179996	15.98687562221	90329037750913792	0.0978	0.065
9	26.502509466630002	16.15244743275	91271014273084800	0.1126	0.0788
10	35.52393427788	24.08298492578	102520255176039296	0.0606	0.0548

1	Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute magnitude(Mv)	Star type	Star color	Spectral Class
2	3068	0.0024	0.17	16.12	0	Red	M
3	3042	0.0005	0.1542	16.6	0	Red	M
4	2600	0.0003	0.102	18.7	0	Red	M
5	2800	0.0002	0.16	16.65	0	Red	M
6	1939	0.000138	0.103	20.06	0	Red	M
7	2840	0.00065	0.11	16.98	0	Red	M
8	2637	0.00073	0.127	17.22	0	Red	M
9	2600	0.0004	0.096	17.4	0	Red	M
10	2650	0.00069	0.11	17.45	0	Red	M
11	2700	0.00018	0.13	16.05	0	Red	M
12	3600	0.0029	0.51	10.69	1	Red	M
13	3129	0.0122	0.3761	11.79	1	Red	M
14	3134	0.0004	0.196	13.21	1	Red	M
15	3628	0.0055	0.393	10.48	1	Red	M
16	2650	0.0006	0.14	11.782	1	Red	M
17	3340	0.0038	0.24	13.07	1	Red	M
18	2799	0.0018	0.16	14.79	1	Red	M
19	3692	0.00367	0.47	10.8	1	Red	M
20	3192	0.00362	0.1967	13.53	1	Red	M

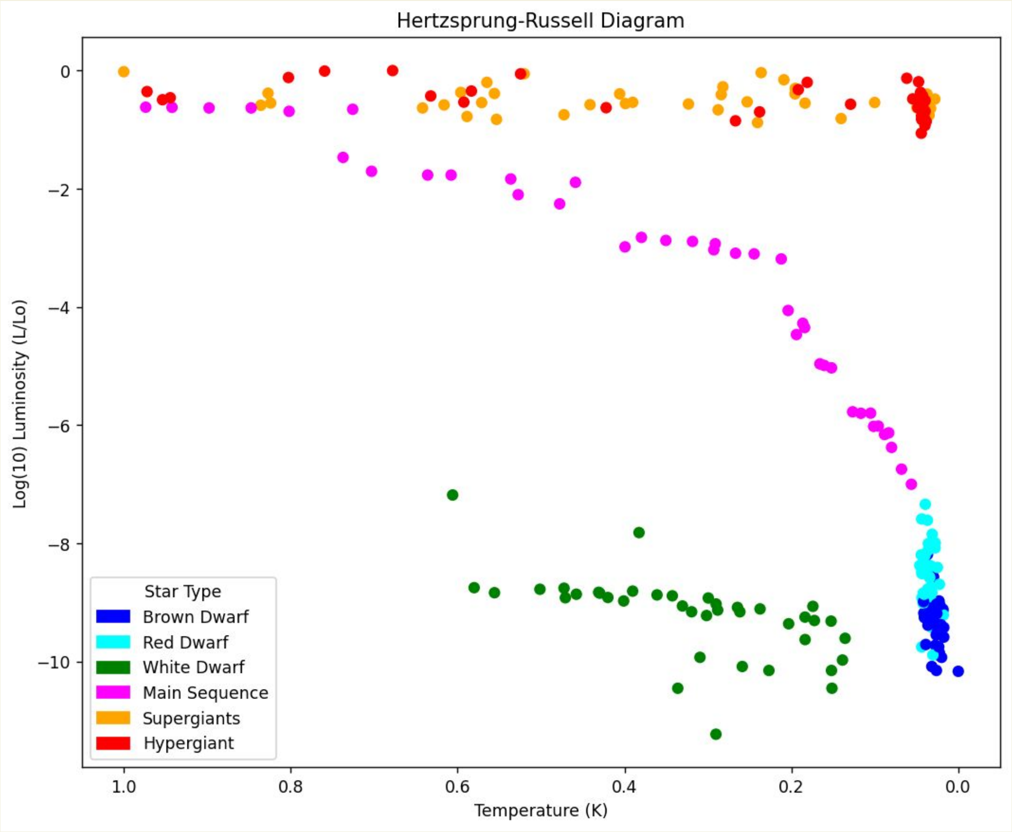
Data Processing

- For Gaia data, a UNIX script (right) was used to remove entries that weren't stars.
- For Gaia and filled in data, we used Python to check for incomplete entries, take the log of the luminosity, and normalize the data.

```bash
#!/bin/bash

# Input file containing the star data
INPUT_FILE="dataGaia2.csv"

# Output file to store the filtered results
OUTPUT_FILE="filteredGaia.csv"

# Write header to the output file
echo "Source,Teff,Lum-Flame,SpType,Pstar,PWD,GMAG" > "$OUTPUT_FILE"

# Process the input file
awk -F, '
    BEGIN { OFS = "," }
    NR > 1 {
        # Check the conditions for Pstar or PWD
        if ($29 == 1 || $30 > 0.99) {
            # Print the selected fields
            print $4, $32, $41, $46, $29, $30, $38
        }
    }
' "$INPUT_FILE" >> "$OUTPUT_FILE"
```

Data with Original Labels

- Stars types in the original data were generated manually or by using estimation algorithms (so they aren't necessarily accurate).
- For our purposes, red and brown dwarfs were merged into one category, as were supergiants and hypergiants.

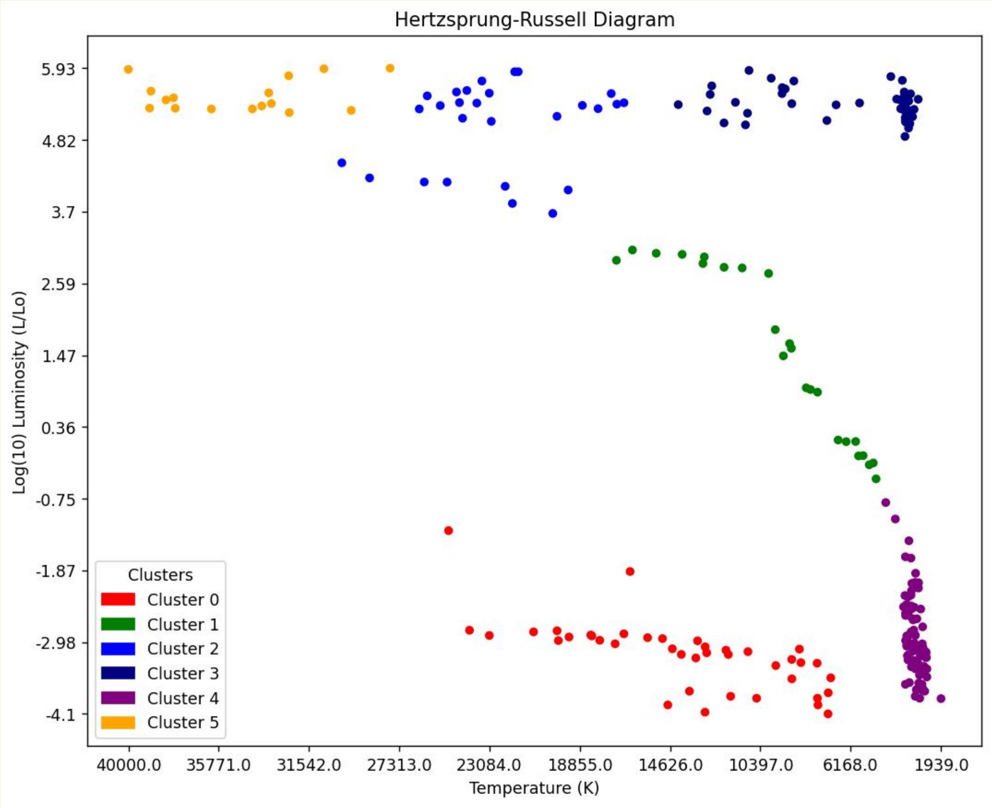


Clustering Methods

- We initially used kmeans, but it didn’t match the actual star classifications very closely.
    - This was because kmeans works best with spherical
- Next we tried Hierarchical clustering (using AgglomerativeClustering in Python)
    - This worked better, but is more computationally expensive than kmeans.
- Clusters needed to be merged to match the large group of supergiants and hypergiants in our data.

Data with Clustering

- Clusters 2, 3, and 5 were merged to most closely match the “supergiants and hypergiants” group from the original data.



Obtaining Errors

- Clusters were manually relabelled to yield the lowest error.
- “Starting error” is error with no relabeling.
- Our clusters agreed with the original data ~81% of the time.

```
Starting Error: 557
Error after relabeling:  46
```

Conclusions

- Our clustering algorithm can be used to quickly and dynamically classify newly found stars and unclassified astronomical objects.
- Classifying stars is important to know what they will do in the future, and by extension what experiments can be done and what can be revealed by monitoring them.
- We used hierarchical clustering and agreed with the original data 81% of the time.
- One area of future work is to use a less strict cleaning algorithm on the Gaia data, which would not filter out objects that have a very slim chance of not being stars. This may result in a more complete data set.

Data References

- Gaia Archive
- Deepraj Baidya (for the filled in data set)
- “Stars and Galaxies” by Seeds and Backman (for the equations used in the filled in data)