According to the inverse square law, the radiant flux (or apparent brightness) of a star depends on:

- its luminosity
- its distance from the Earth

Flux is directly proportional to luminosity and inversely proportional to distance from the Earth.

Astro jargon 🚀

- Luminosity - total amount of energy that a star puts out as light every second.
- Flux - the total amount of energy intercepted by the detector on Earth *(in our case, it is the NASA Kepler space telescope)* divided by the area of the detector.

Reference: https://astronomy.swin.edu.au/cosmos/F/Flux

Intuitively, exoplanet stars are more likely to have greater apparent brightness since they are light-years away from Earth as opposed to stars found in the solar system.

Fun fact 🤔

According to NASA Exoplanet Exploration, the Earth's closest known exoplanet - Proxima Centauri b - is located four light-years away from Earth.

The farthest planet within our solar system is Neptune a magnitude of 4.3 billon kilometers from Earth.

1 light-year = 9.46 × 10^12 km

Let's test this hypothesis!

H0: There is no statistically significant difference in overall radiant flux between exoplanet and non-exoplanet stars.

HA: There exists a statistically significant difference in overall radiant flux between exoplanet and non-exoplanet stars.

```
import pandas as pd
exoTrain = pd.read_csv("exoTrain.csv")
exoTrain["LABEL"] = exoTrain["LABEL"].replace(1, 0)
exoTrain["LABEL"] = exoTrain["LABEL"].replace(2, 1)

exo = exoTrain.loc[exoTrain['LABEL'] == 1]
non_exo = exoTrain.loc[exoTrain['LABEL'] == 0]
```

```
exo.shape
```

```
(37, 3198)
```

```
non_exo.shape
```

```
(5050, 3198)
```

```
exo.head(5)
```

	LABEL	FLUX.1	FLUX.2	FLUX.3	FLUX.4	FLUX.5	FLUX.6	FLUX.7	FLUX.8
0	1	93.85	83.81	20.10	-26.98	-39.56	-124.71	-135.18	-96.27
1	1	-38.88	-33.83	-58.54	-40.09	-79.31	-72.81	-86.55	-85.33
2	1	532.64	535.92	513.73	496.92	456.45	466.00	464.50	486.39
3	1	326.52	347.39	302.35	298.13	317.74	312.70	322.33	311.31
4	1	-1107.21	-1112.59	-1118.95	-1095.10	-1057.55	-1034.48	-998.34	-1022.71

5 rows × 3198 columns

We take average of the flux values for every star to calculate the overall flux values for our hypothesis testing.

```
from statistics import mean
exo['FLUX_AVG'] = exo.iloc[:, 1:3198].mean(axis=1)

print(exo.head(5))
```

```
   LABEL   FLUX.1   FLUX.2   FLUX.3   FLUX.4   FLUX.5   FLUX.6  FLUX.7  \
0      1    93.85    83.81    20.10   -26.98   -39.56  -124.71 -135.18
1      1   -38.88   -33.83   -58.54   -40.09   -79.31   -72.81  -86.55
2      1   532.64   535.92   513.73   496.92   456.45   466.00  464.50
3      1   326.52   347.39   302.35   298.13   317.74   312.70  322.33
4      1 -1107.21 -1112.59 -1118.95 -1095.10 -1057.55 -1034.48 -998.34

    FLUX.8  FLUX.9  ...  FLUX.3189  FLUX.3190  FLUX.3191  FLUX.3192  \
0   -96.27  -79.89  ...    -102.15    -102.15      25.13      48.57
1   -85.33  -83.97  ...     -32.21     -32.21     -24.89      -4.86
2   486.39  436.56  ...      13.31      13.31     -29.89     -20.88
3   311.31  312.42  ...      -3.73      -3.73      30.05      20.03
4 -1022.71 -989.57  ...    -401.66    -401.66    -357.24    -443.76

   FLUX.3193  FLUX.3194  FLUX.3195  FLUX.3196  FLUX.3197   FLUX_AVG
0      92.54      39.32      61.42       5.08     -39.54   9.953857
1       0.76     -11.70       6.46      16.00      19.93  -7.427932
2       5.06     -11.80     -28.91     -70.02     -96.67   8.189087
3     -12.67      -8.77     -17.31     -17.35      13.98  -4.646587
```

```
4    -438.54    -399.71    -384.65    -411.79    -510.54 -14.240660

[5 rows x 3199 columns]
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user
```

```
non_exo['FLUX_AVG'] = non_exo.iloc[:, 1:3198].mean(axis=1)
print(non_exo.head(5))
```

```
    LABEL  FLUX.1  FLUX.2  FLUX.3  FLUX.4  FLUX.5  FLUX.6  FLUX.7  FLUX.8  \
37      0 -141.22  -81.79  -52.28  -32.45   -1.55  -35.61  -23.28   19.45
38      0  -35.62  -28.55  -27.29  -28.94  -15.13  -51.06    2.67   -5.21
39      0  142.40  137.03   93.65  105.64   98.22   99.06   86.40   60.78
40      0 -167.02 -137.65 -150.05 -136.85  -98.73 -103.14 -107.70 -123.19
41      0  207.74  223.60  246.15  224.06  210.77  189.56  172.68  170.31

    FLUX.9  ...  FLUX.3189  FLUX.3190  FLUX.3191  FLUX.3192  FLUX.3193  \
37   53.11  ...     -22.34     -36.23      27.44      13.52      38.66
38    9.67  ...     -38.22     -46.23     -54.40     -23.51     -26.96
39   45.18  ...      -3.03     -30.27     -24.22     -35.10     -39.64
40 -125.65  ...     -79.79     -80.62     -78.22    -105.06     -69.67
41  148.79  ...    -136.92    -174.97    -180.46    -164.01    -126.58

    FLUX.3194  FLUX.3195  FLUX.3196  FLUX.3197   FLUX_AVG
37     -17.53      31.49      31.38      50.03   0.209246
38      -3.95      -0.34      10.52      -7.69   4.212268
39      23.78      23.40      -0.50       0.97   0.891442
40     -90.45     -73.67     -66.71     -66.07   4.850673
41      84.05      63.81     108.36      78.10  33.159481

[5 rows x 3199 columns]
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user
  """Entry point for launching an IPython kernel.
```

Choosing the right statistical test to use

- *One-way ANOVA* is used to detect statistically significant differences between the means of three or more independent groups.
- *Student's t-test* is a specific type of ANOVA used when we only have two population means to compare.

- *Welch's t-test* is a non-parametric version of the Student's t-test that does not require the variances of the two groups to be equal. It is also known as the unequal variances t-test for this reason.

Testing for unequal variances of both groups

```
import scipy.stats as stats
import numpy as np

import random as rd
rd.seed(100)
non_exo_flux = non_exo['FLUX_AVG'].sample(n=37)
```

Welch's t-test does not require the sample sizes of the groups to be equal but the number of non-exoplanet stars is exceptionally larger than the count of exoplanet stars. So, we check for equal variances with downsampled non-exoplanets too.

With non-exoplanet star data downsampling:

```
import scipy.stats as stats
import numpy as np

import random as rd
rd.seed(100)
non_exo_flux = non_exo['FLUX_AVG'].sample(n=37)

exo_flux = exo['FLUX_AVG'].tolist()
non_exo_flux = non_exo_flux.tolist()

#printing the variances of both data groups
print(np.var(exo_flux), np.var(non_exo_flux))
```

```
41467.0849229546 1556.058853466069
```

Without non-exoplanet star data downsampling:

```
import scipy.stats as stats
import numpy as np

exo_flux = exo['FLUX_AVG'].tolist()
non_exo_flux = non_exo['FLUX_AVG'].tolist()
```

```
#printing the variances of both data groups
print(np.var(exo_flux), np.var(non_exo_flux))
```

```
41467.0849229546 39651521.98648239
```

Upon getting these results, we can confidently go ahead and test our hypothesis using the Welch's t-test as the variances are widely different.

Welch's t-test

Since we have unequal sample variances and unequal sample sizes for our star groups, we carry out the Welch's t-test.

Using downsampled non-exoplanet star data:

```
exo_flux = exo['FLUX_AVG'].tolist()
non_exo_flux = non_exo_flux.tolist()

#welch's t-test
print(stats.ttest_ind(exo_flux, non_exo_flux, equal_var = False))
```

```
Ttest_indResult(statistic=-1.702158077171895, pvalue=0.09494829953262572)
```

Setting the signficance level as 0.05...

Since our p-value (= 0.09) greater than 0.05, we cannot reject the null hypothesis.

Using original non-exoplanet star data:

```
exo_flux = exo['FLUX_AVG'].tolist()
non_exo_flux = non_exo['FLUX_AVG'].tolist()

#welch's t-test
print(stats.ttest_ind(exo_flux, non_exo_flux, equal_var = False))
```

```
Ttest_indResult(statistic=-1.8844773219290087, pvalue=0.05967580795193132)
```

Since our p-value (= 0.06) greater than 0.05, we cannot reject the null hypothesis.

Inference

Hence, there is no significant difference in the overall flux for exoplanet and non-exoplanet stars.

Limitations of this approach

- Taking the average of radial flux values for every star may not be the best representation of overall flux.
- Another important variable - luminosity - was not taken into consideration while testing out this hypothesis.