Statistical Challenges in Stellar Parameter Estimation from Theory and Data

Joshua S. Speagle (沈佳士)

Banting & Dunlap Postdoctoral Fellow
Department of Statistical Sciences
David A. Dunlap Department of Astronomy & Astrophysics
Dunlap Institute for Astronomy & Astrophysics

j.speagle@utoronto.ca

UNIVERSITY OF TORONTO

Bourses postdoctorales Banting Postdoctoral Fellowships

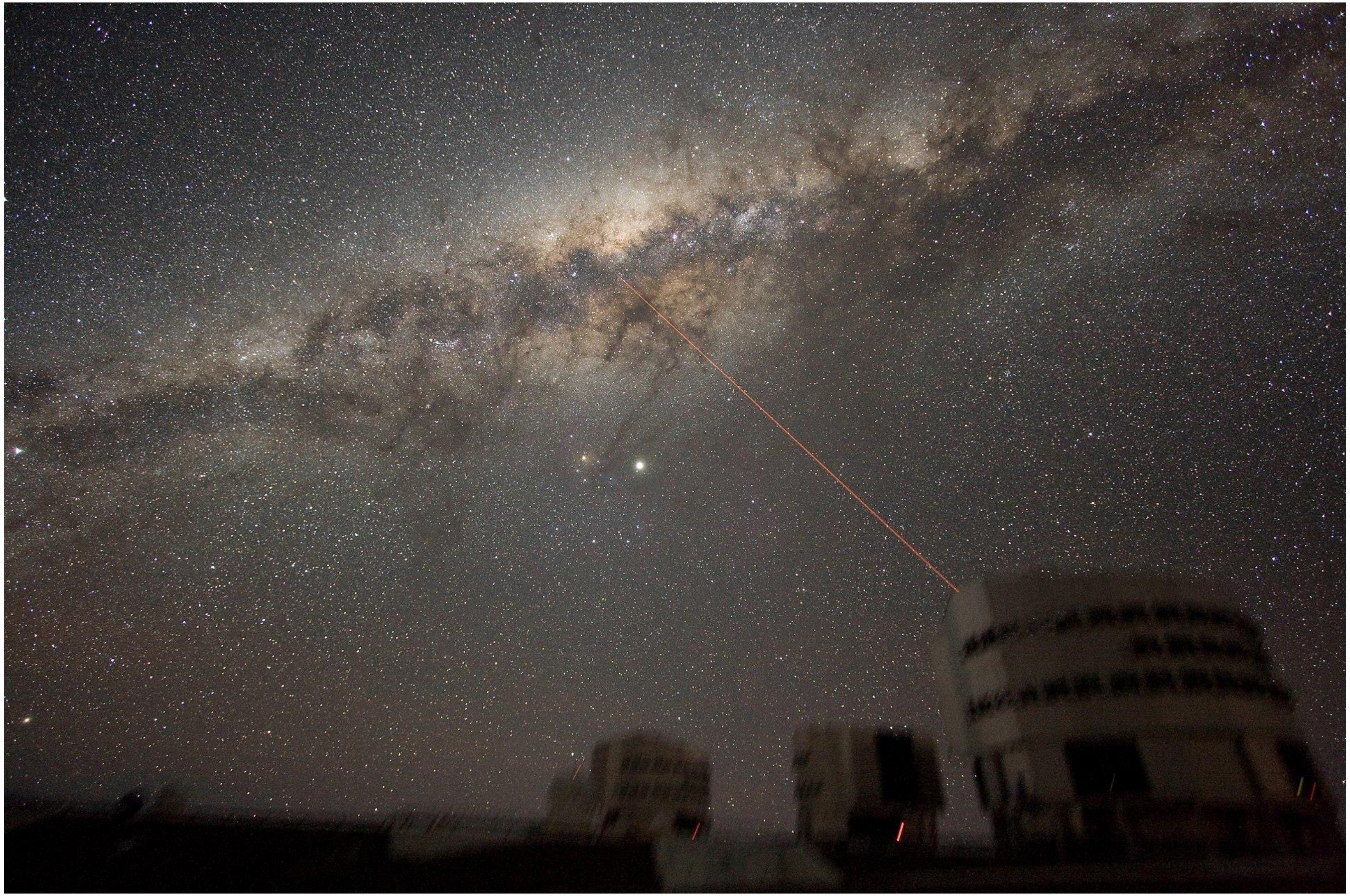


Image credit: ESO /Y. Beletsky (VLT)

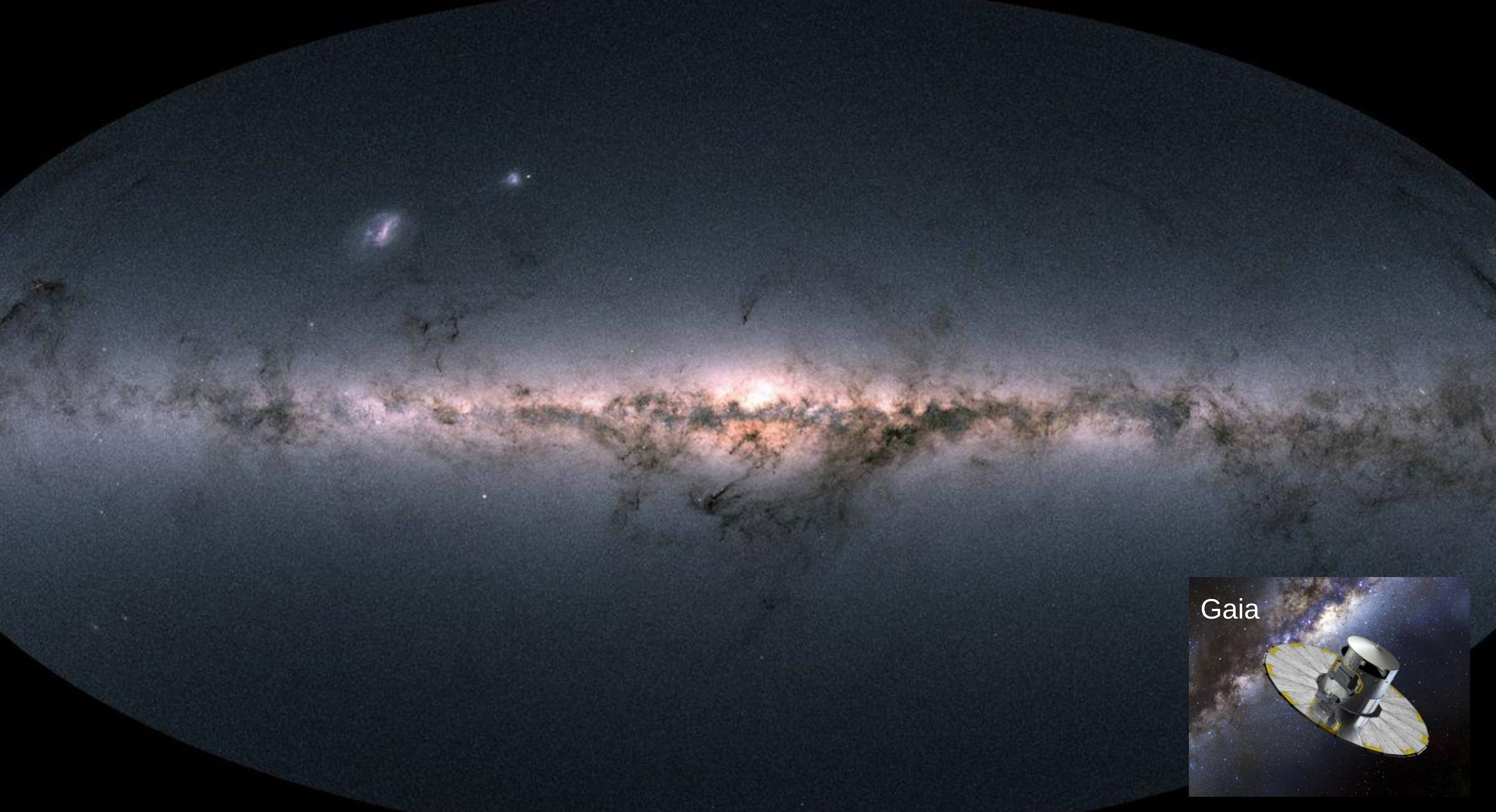
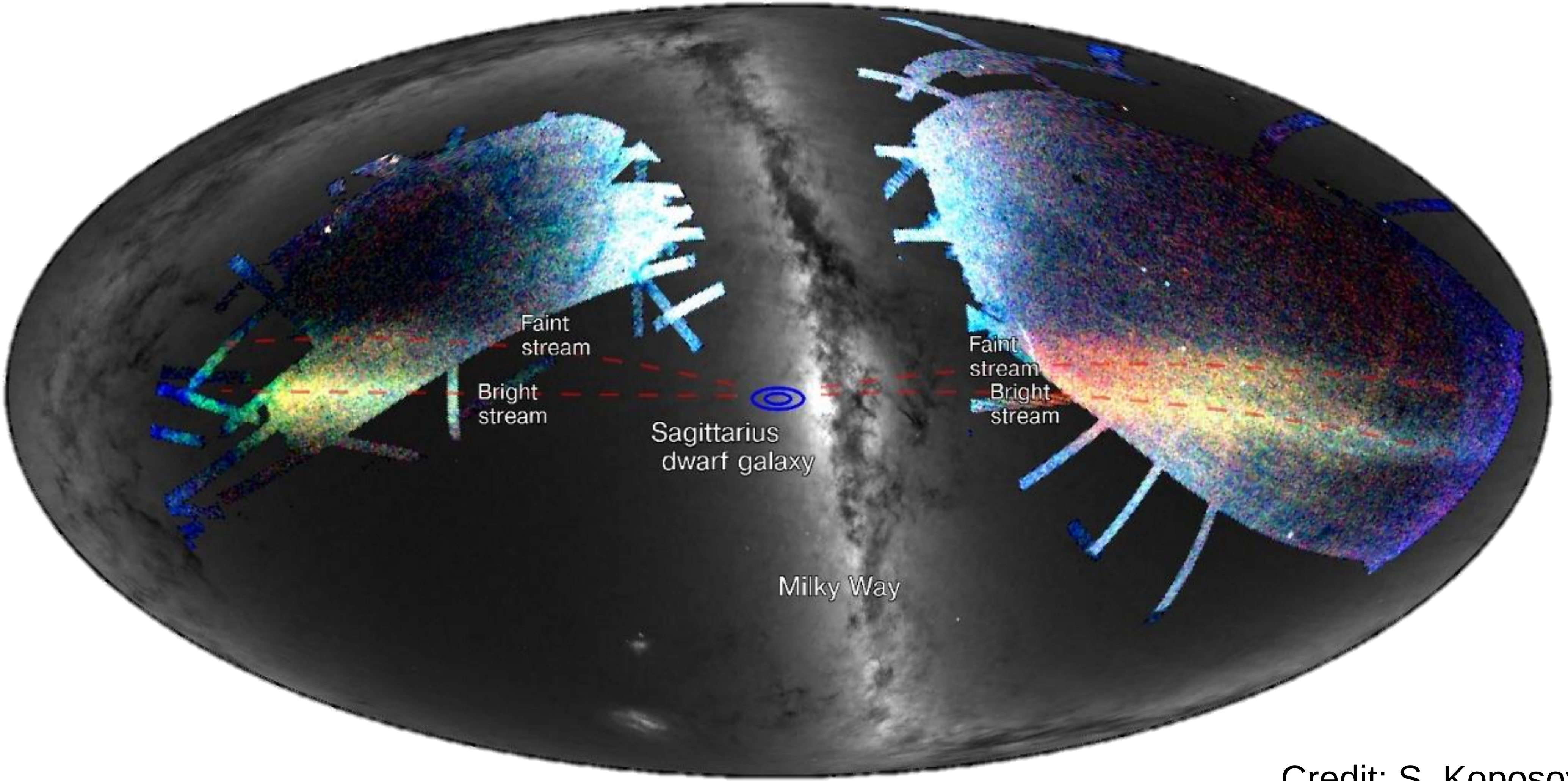


Image credit: ESA / Gaia / DPAC

What can stars tell us about our Galaxy?



Credit: S. Koposov / SDSS III

Understanding the Milky Way

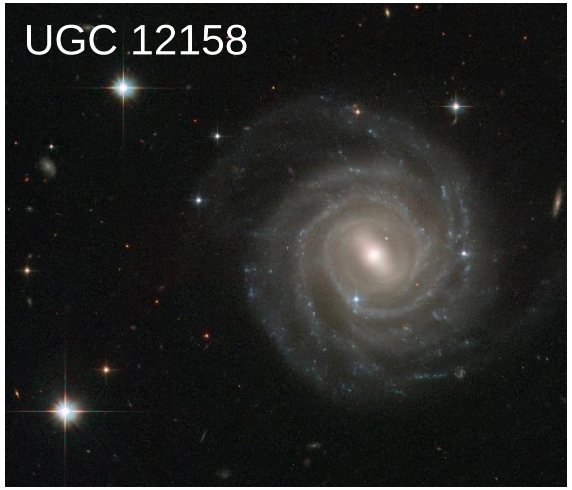
Data (2-D)

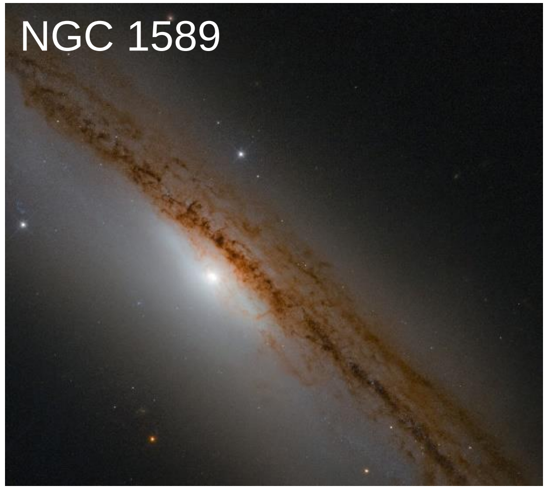
Image credit: ESO /Y. Beletsky (VLT)

Model (3-D)

Image credit: NASA

Stars: The Building Blocks of Galaxies





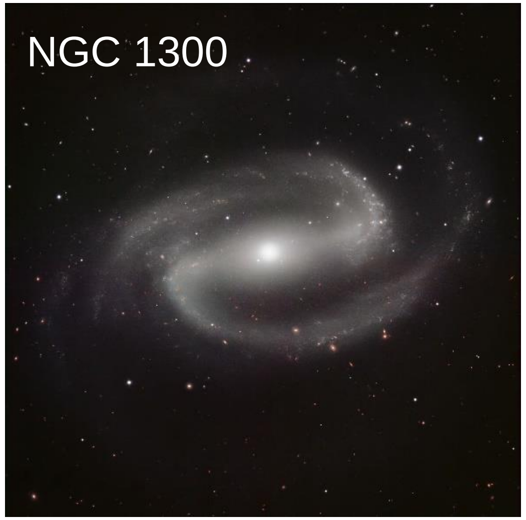
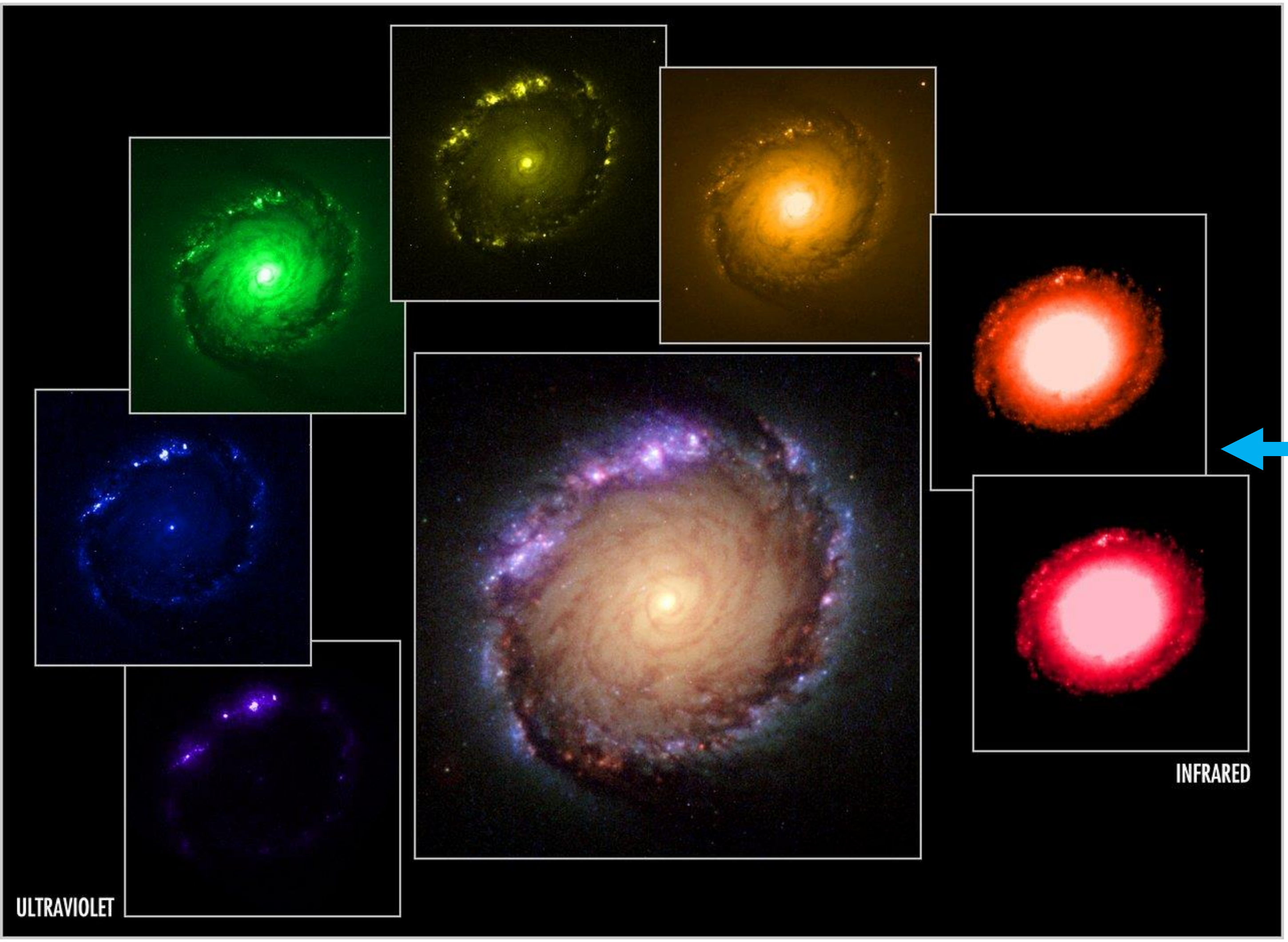
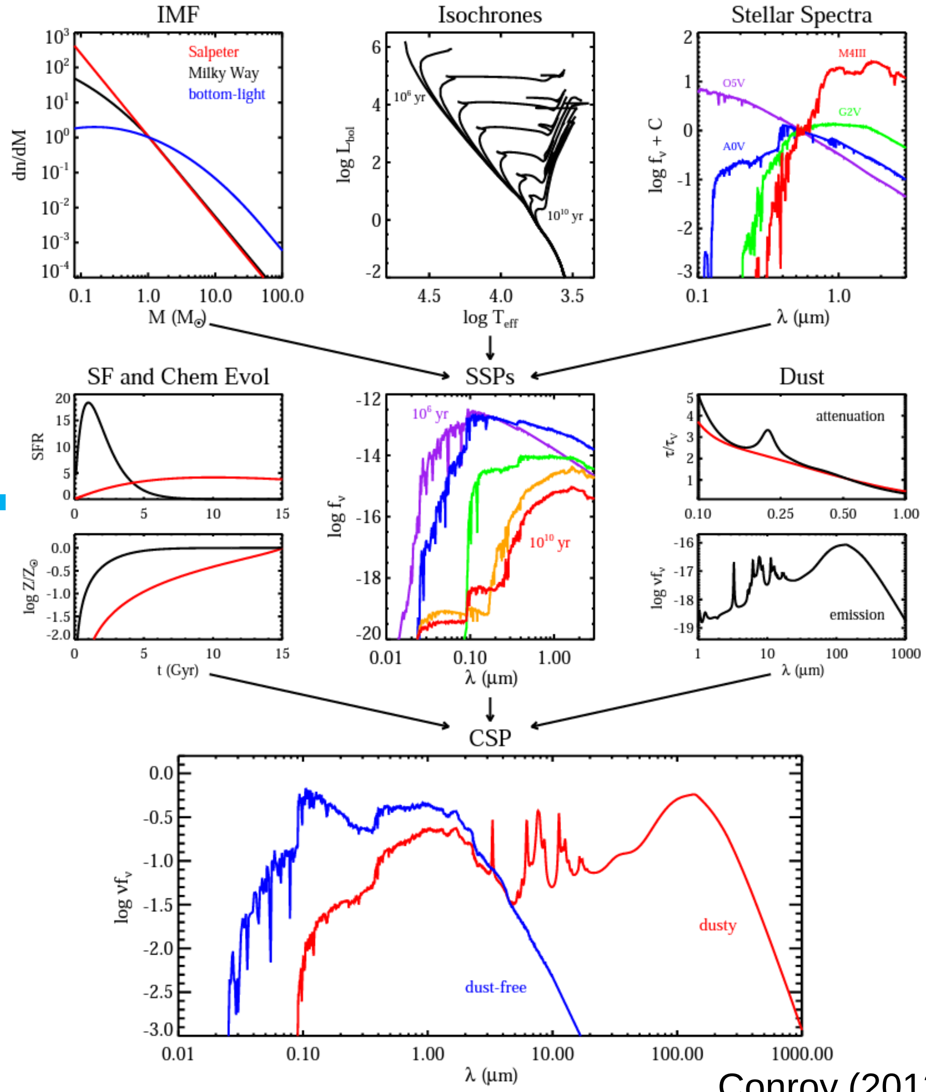


Image credit:
ESA/ESO/Hubble/NASA

Stars: The Building Blocks of Galaxies



NGC 1512 / Picture Credit: HST



Conroy (2013)

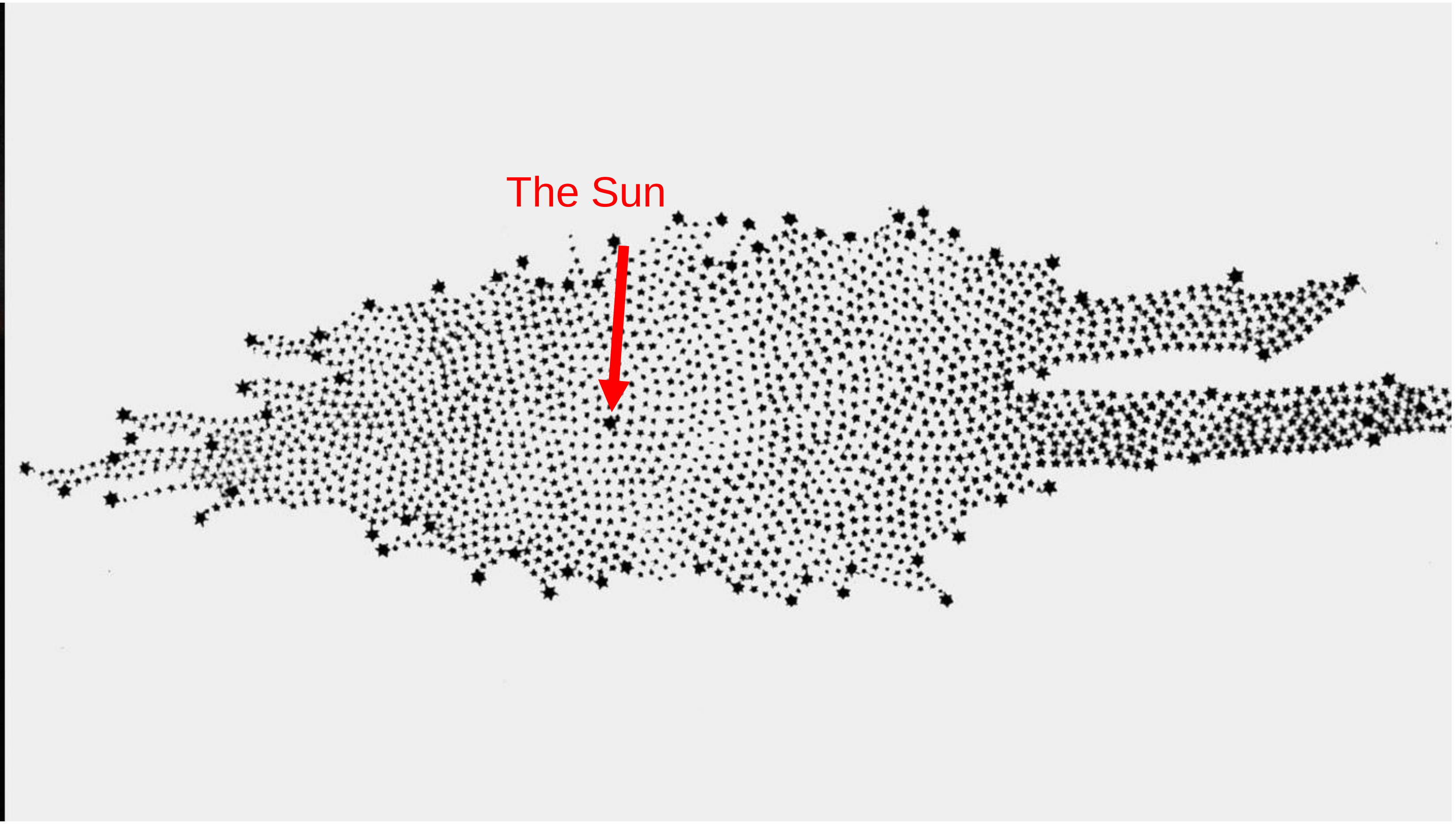
Part 1

A Motivated Historical Example

Milky Way (c. 2010)

Milky Way (c. 1781)

William Herschel



Credit: On the Construction of the Heavens (1785)

How did he do it?

Two primary assumptions:

1. All stars have the **same intrinsic** brightness.

$$B_{\text{obs}} = \frac{B_{\text{int}}}{d^2}$$

2. There is **no material** between us and other stars.

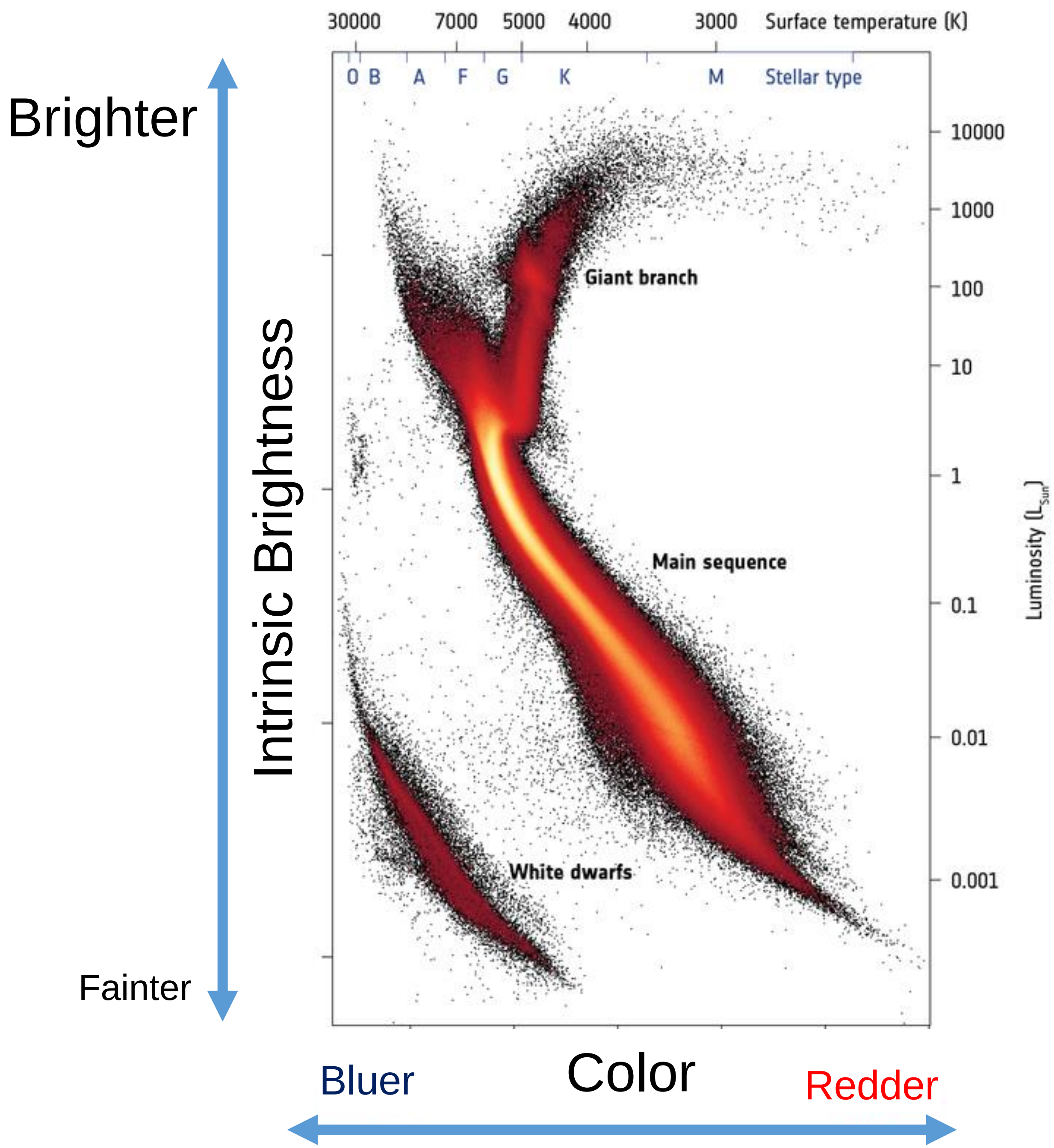
$$\rho_{\text{obs}} \sim \rho_{\text{int}}$$

density

Stellar Populations

- c. 1910: measurements of the **intrinsic brightness** and **colors** of stars by Ejnar Hertzsprung and Henry Norris Russell spanned a wide range of values.

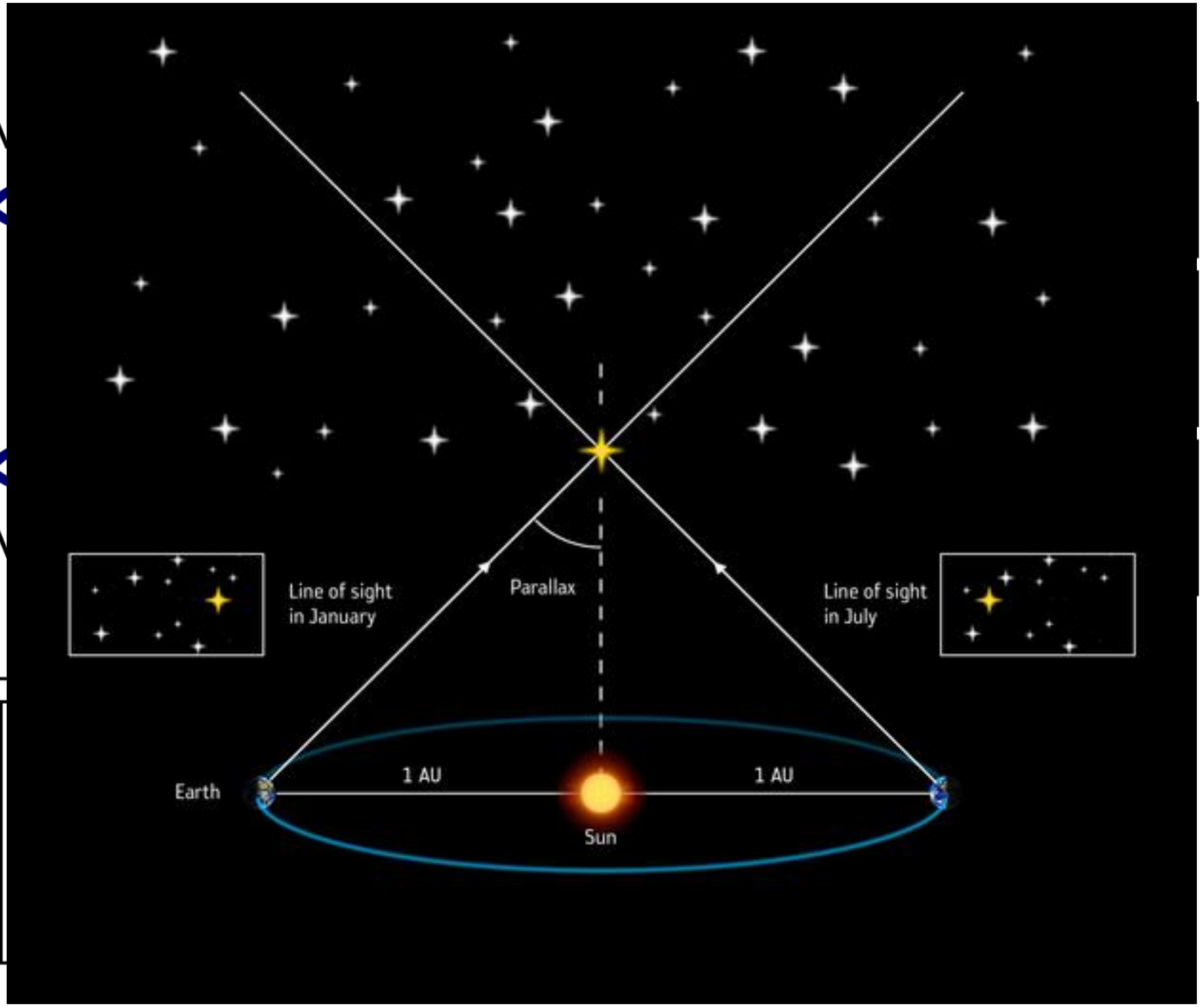
Credit: ESA / Gaia / DPAC



Measuring Distances with Parallax

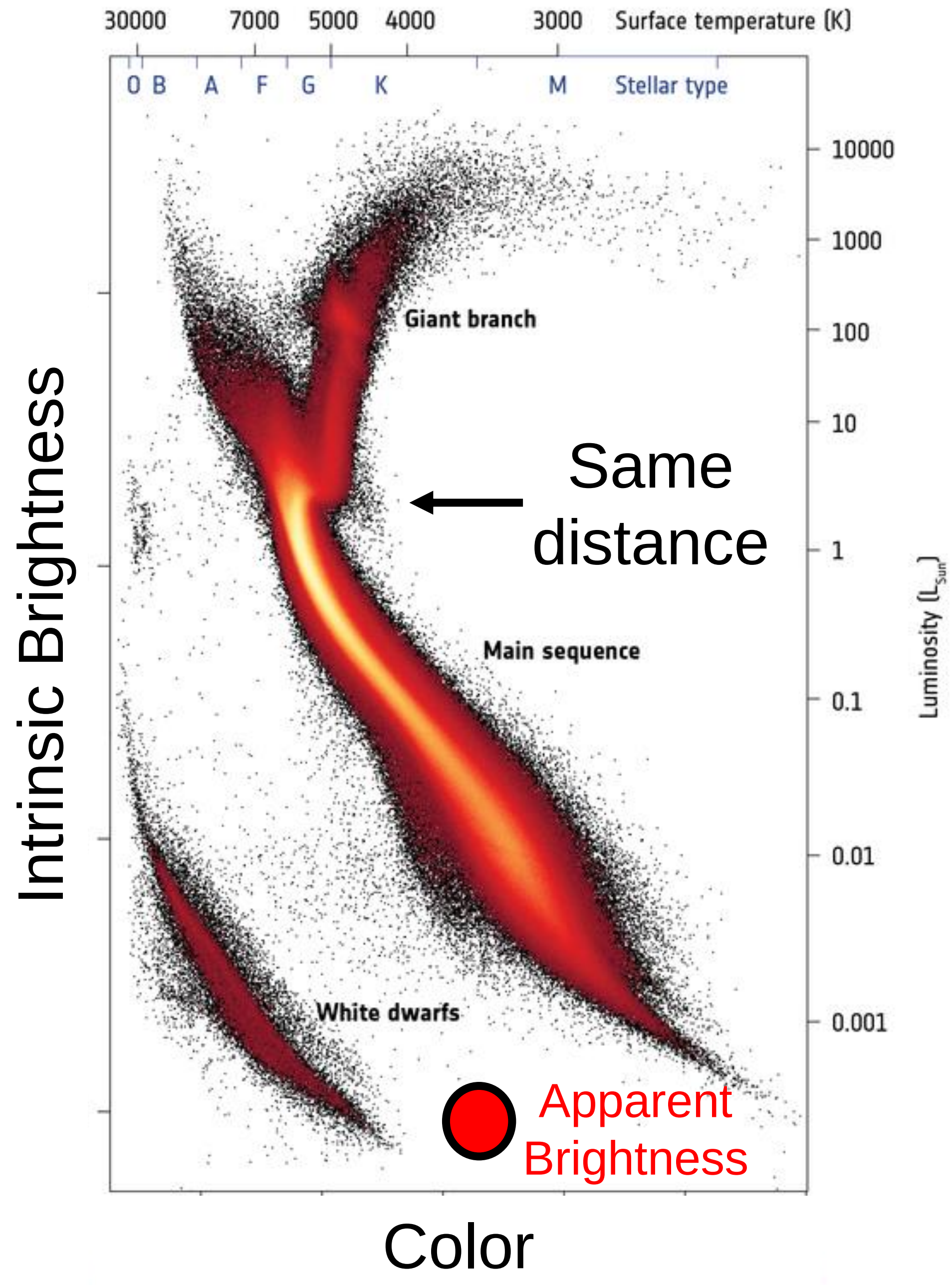
- We **can leverage the Earth's orbit around the Sun** to get an independent measurement of distance.

$$d = 1/\varpi$$



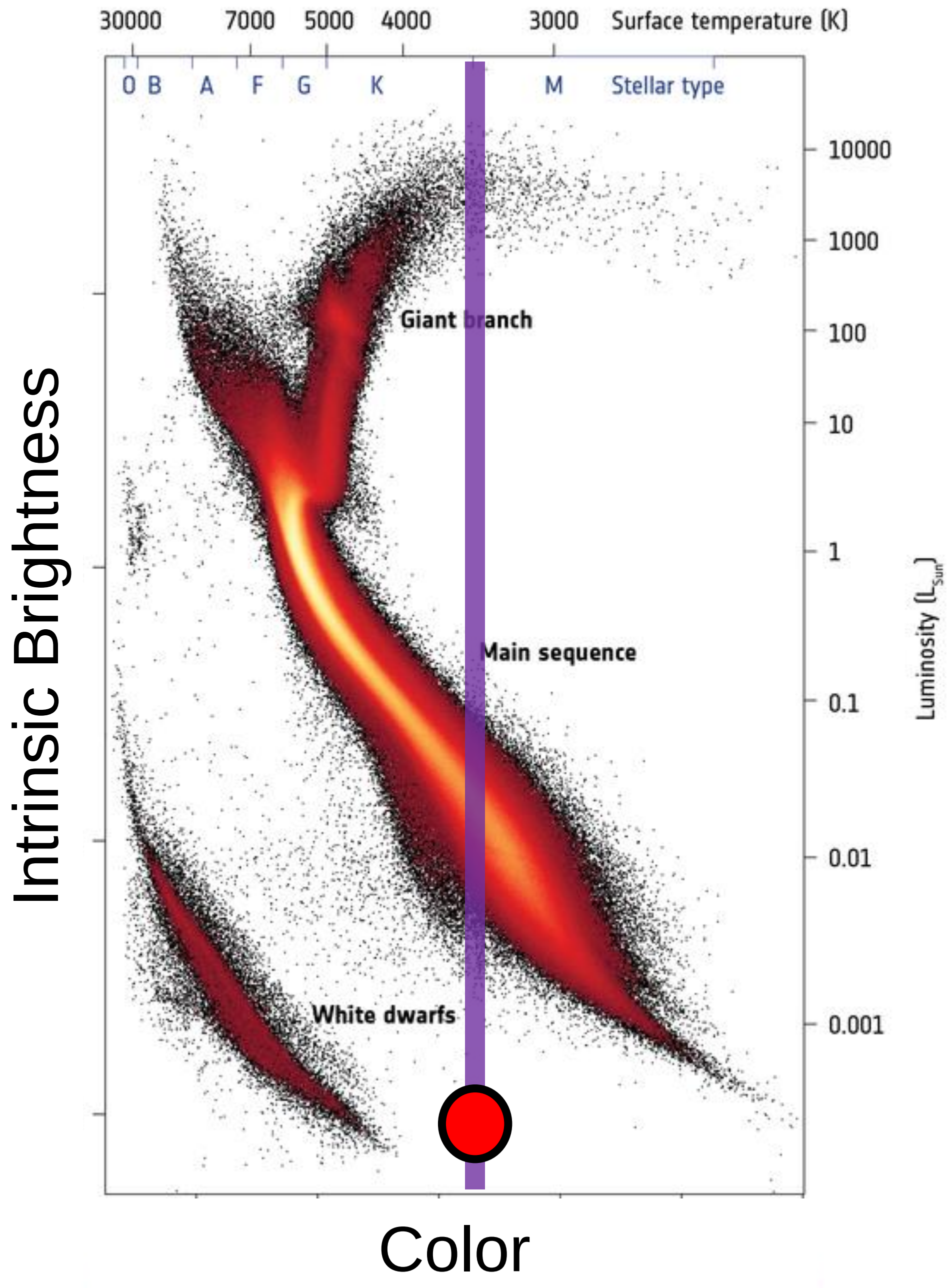
Credit: ESA / Gaia / DPAC

Measuring Distances

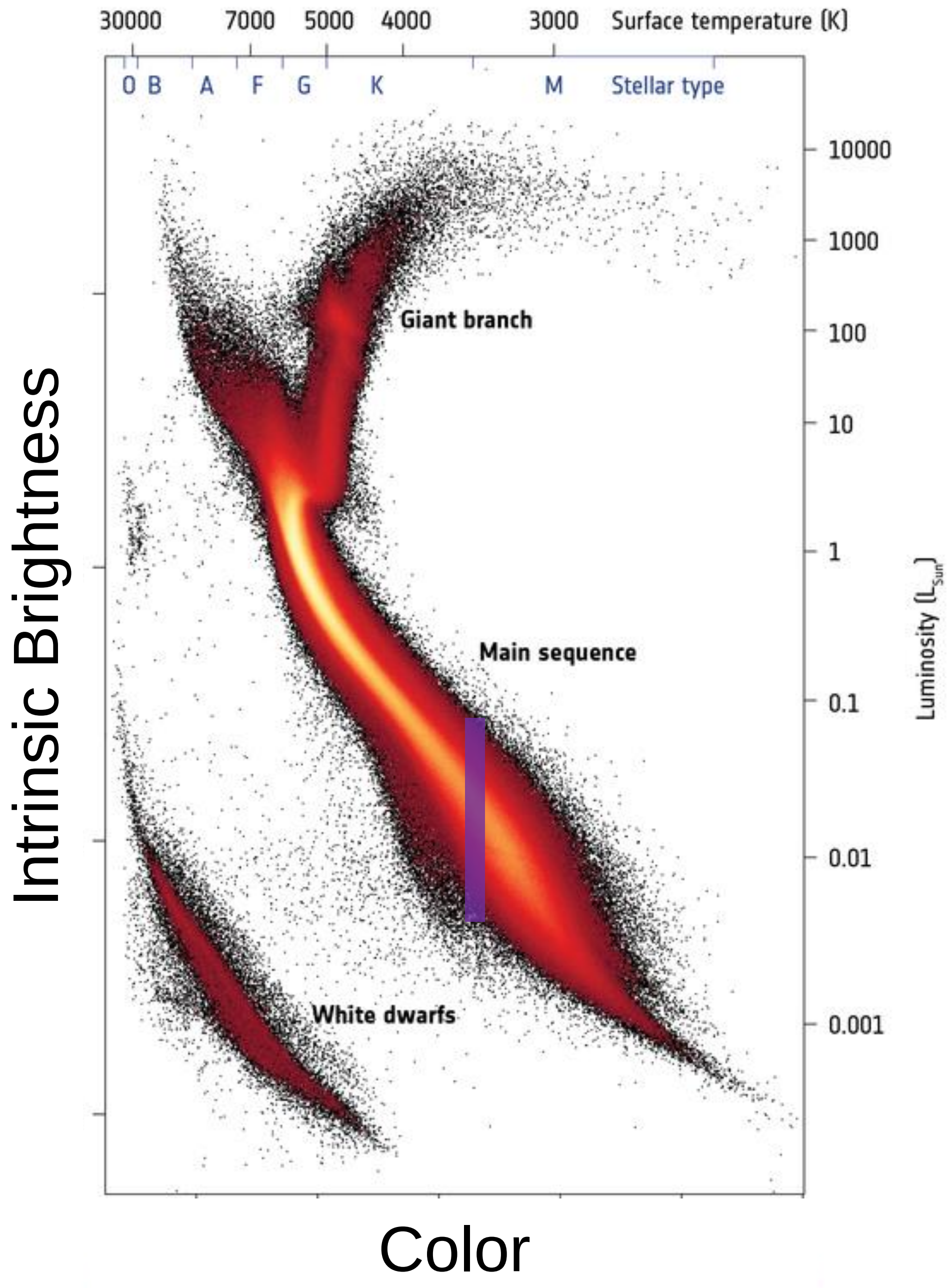


Credit: ESA / Gaia / DPAC

Measuring Distances



Measuring Distances



How did he do it?

Two primary assumptions:

1. All stars have the **same intrinsic** brightness.

$$B_{\text{obs}} = \frac{B_{\text{int}}}{d^2}$$

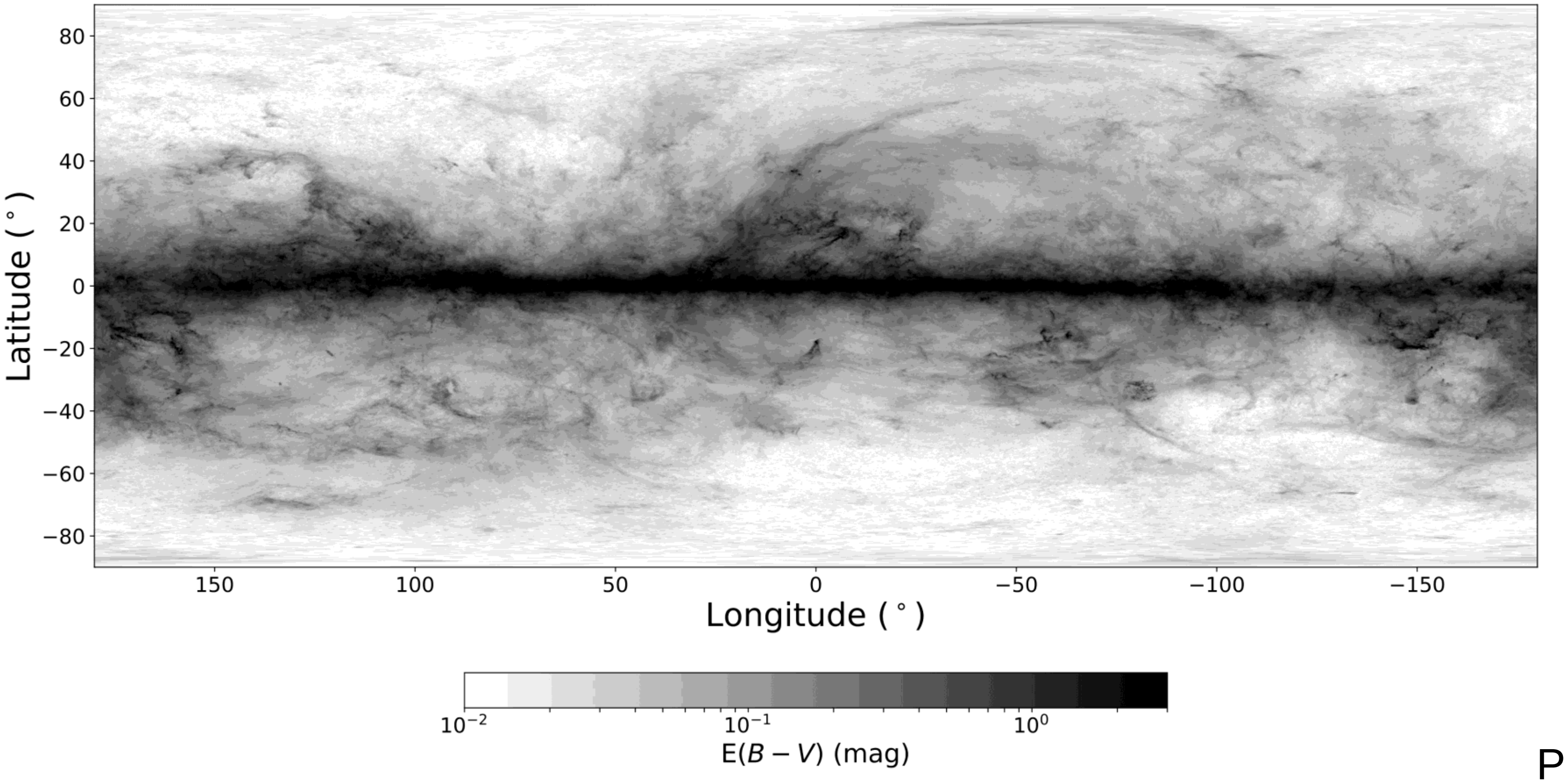
2. There is **no material** between us and other stars.

$$\rho_{\text{obs}} \sim \rho_{\text{int}}$$

density

Dust and the Interstellar Medium

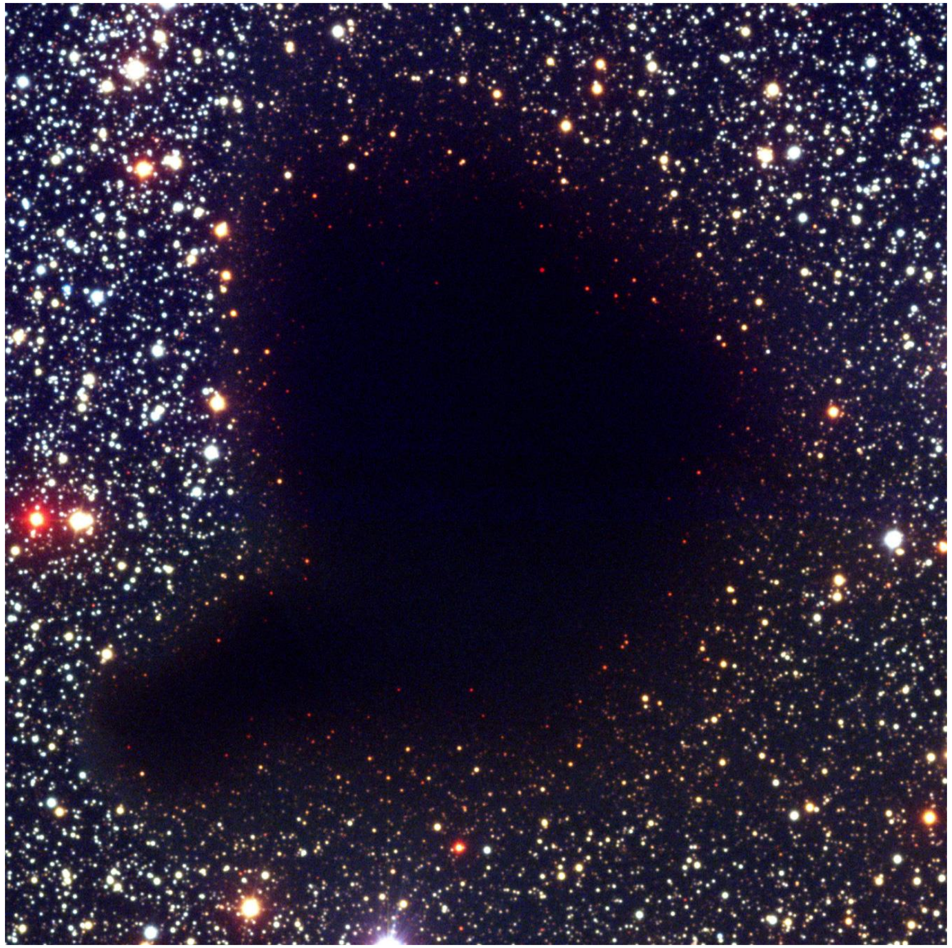
- The space between stars is filled with gas, dust, and other material.
- While dust is ~1% of ISM by mass, it **scatters 30% of all light** (Draine 2003).



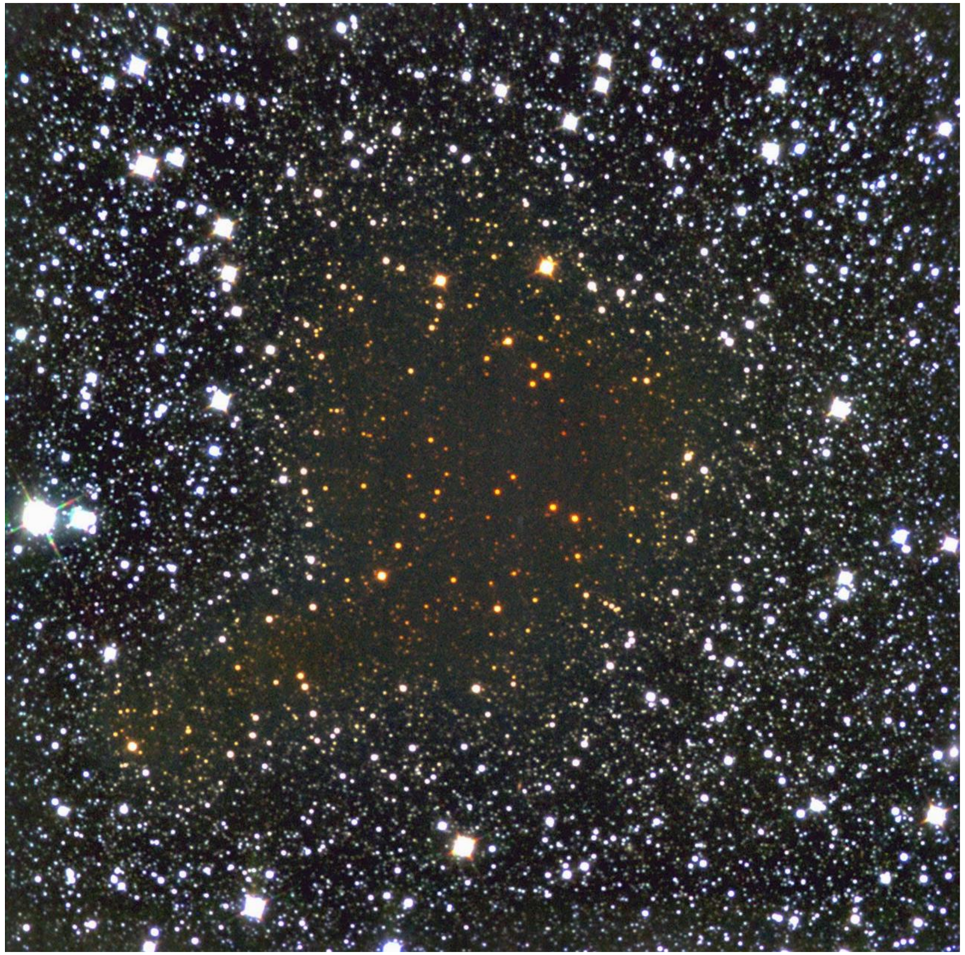
Planck Team (2013)

Extinction and Reddening

Visible Wavelengths

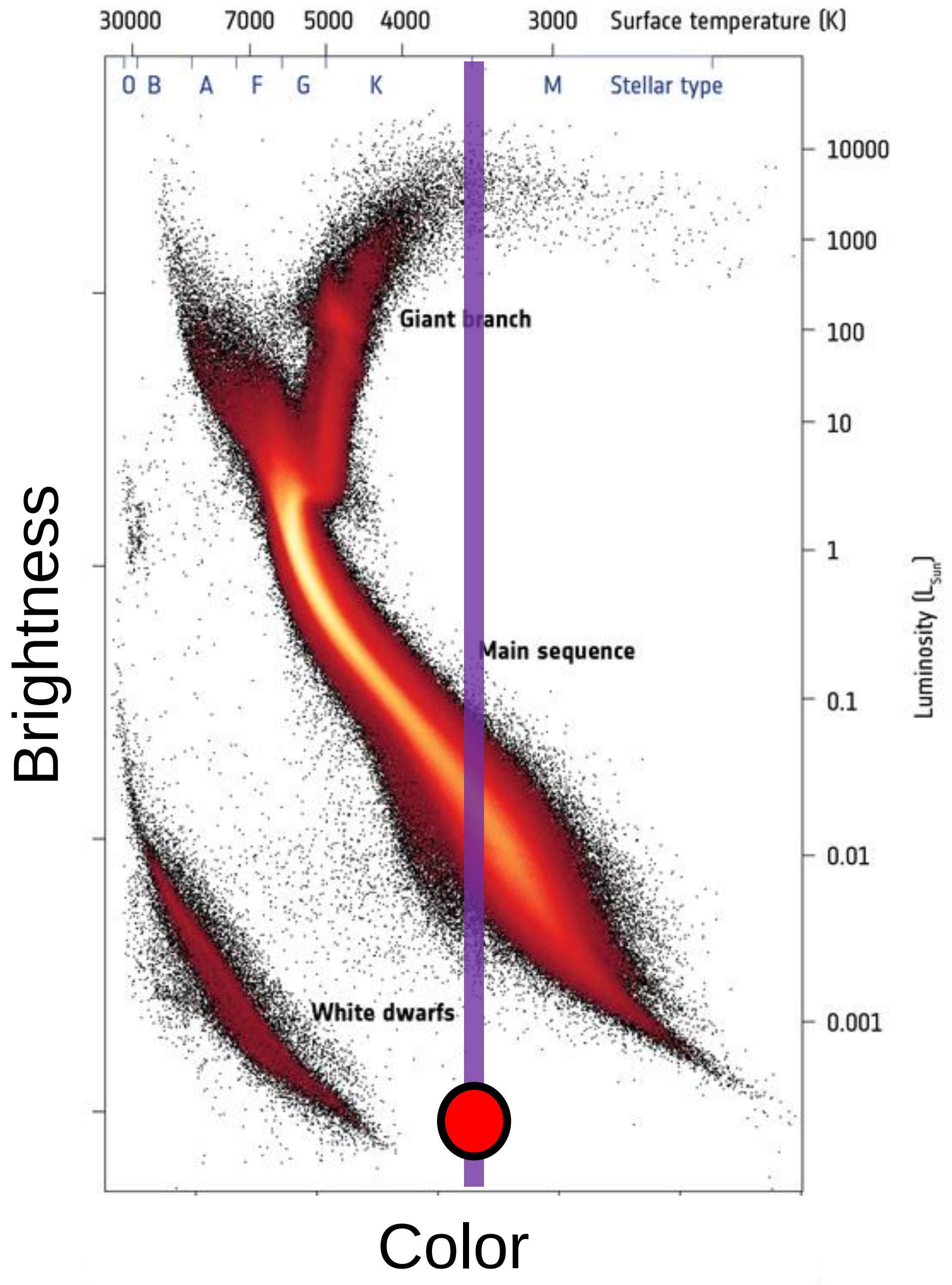


Infrared Wavelengths



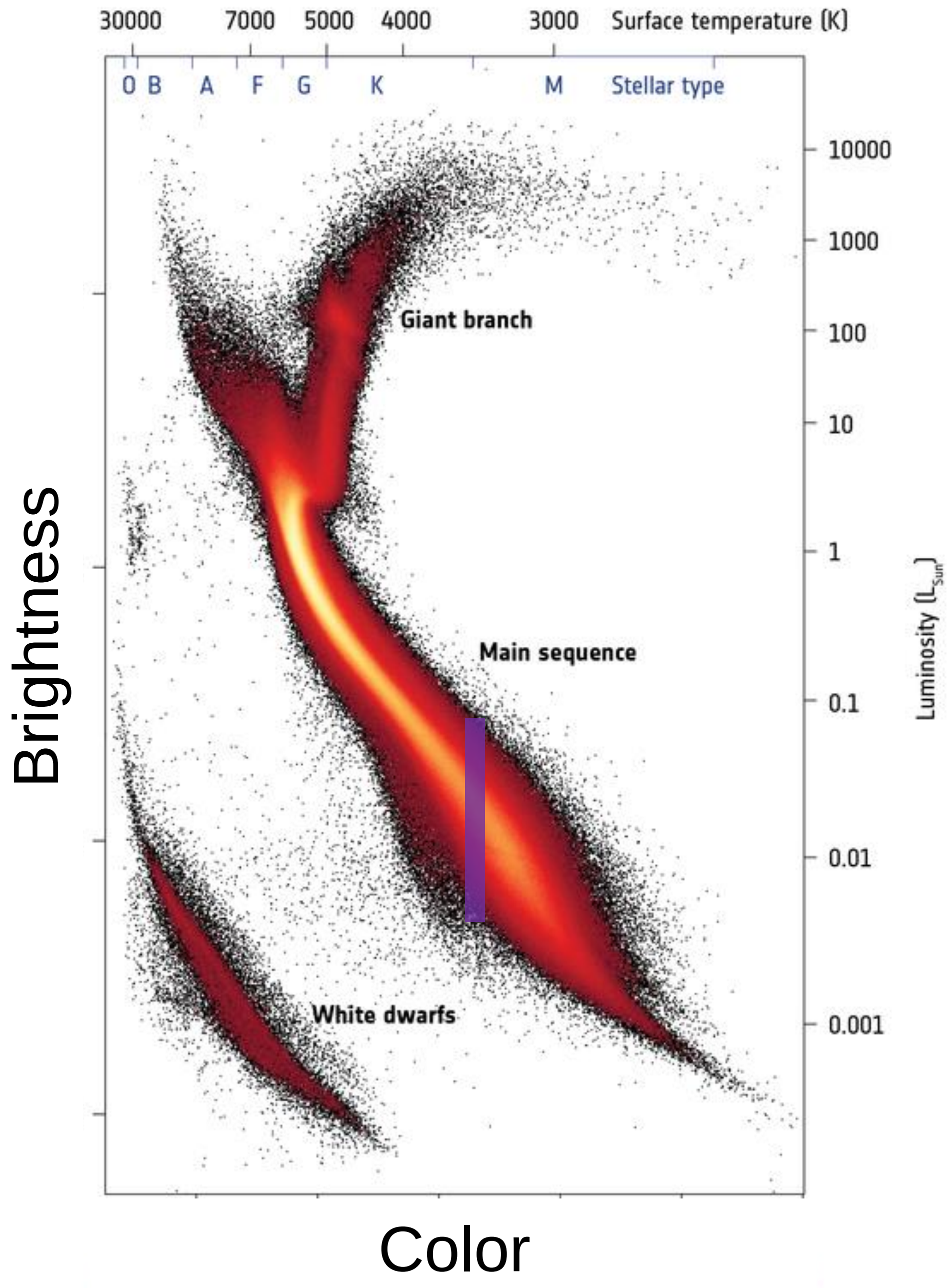
Impact of Dust

Credit: ESA / Gaia / DPAC



Credit: ESA / Gaia / DPAC

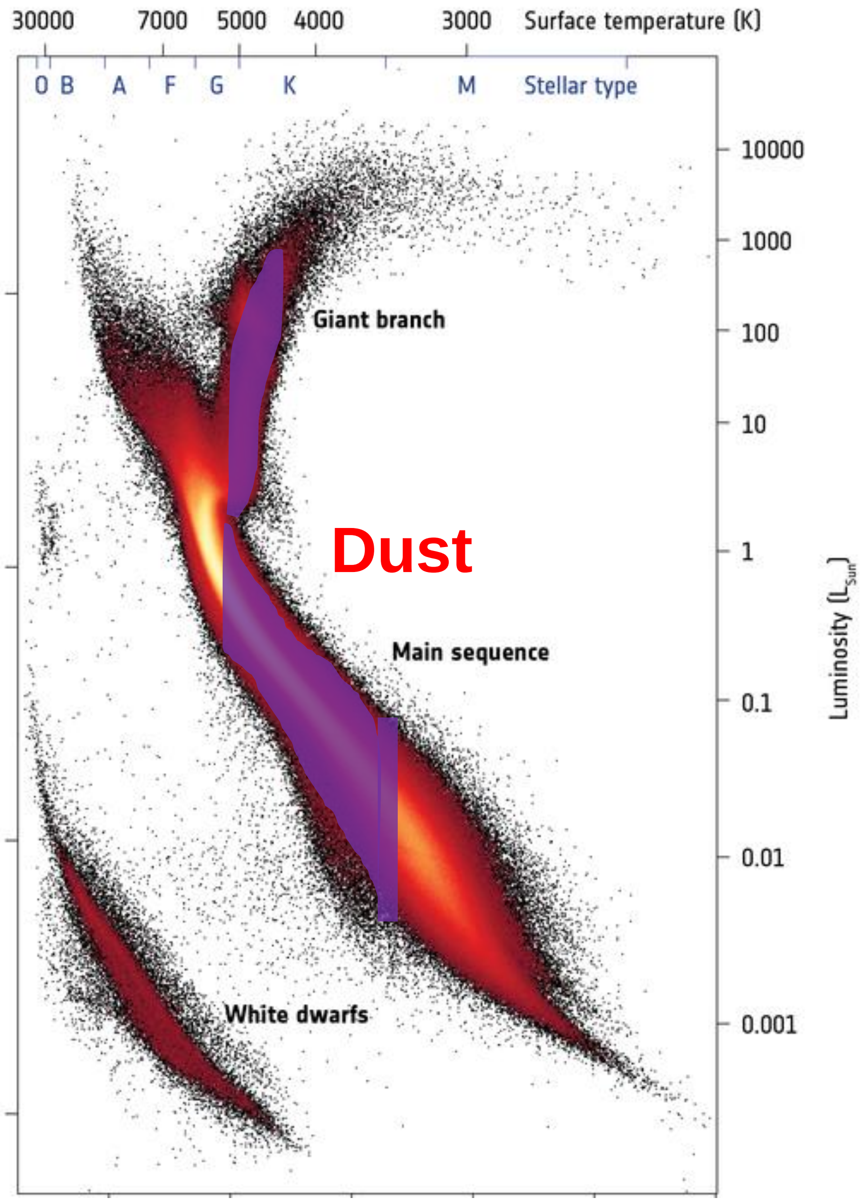
Impact of Dust



Credit: ESA / Gaia / DPAC

Impact of Dust

Brightness

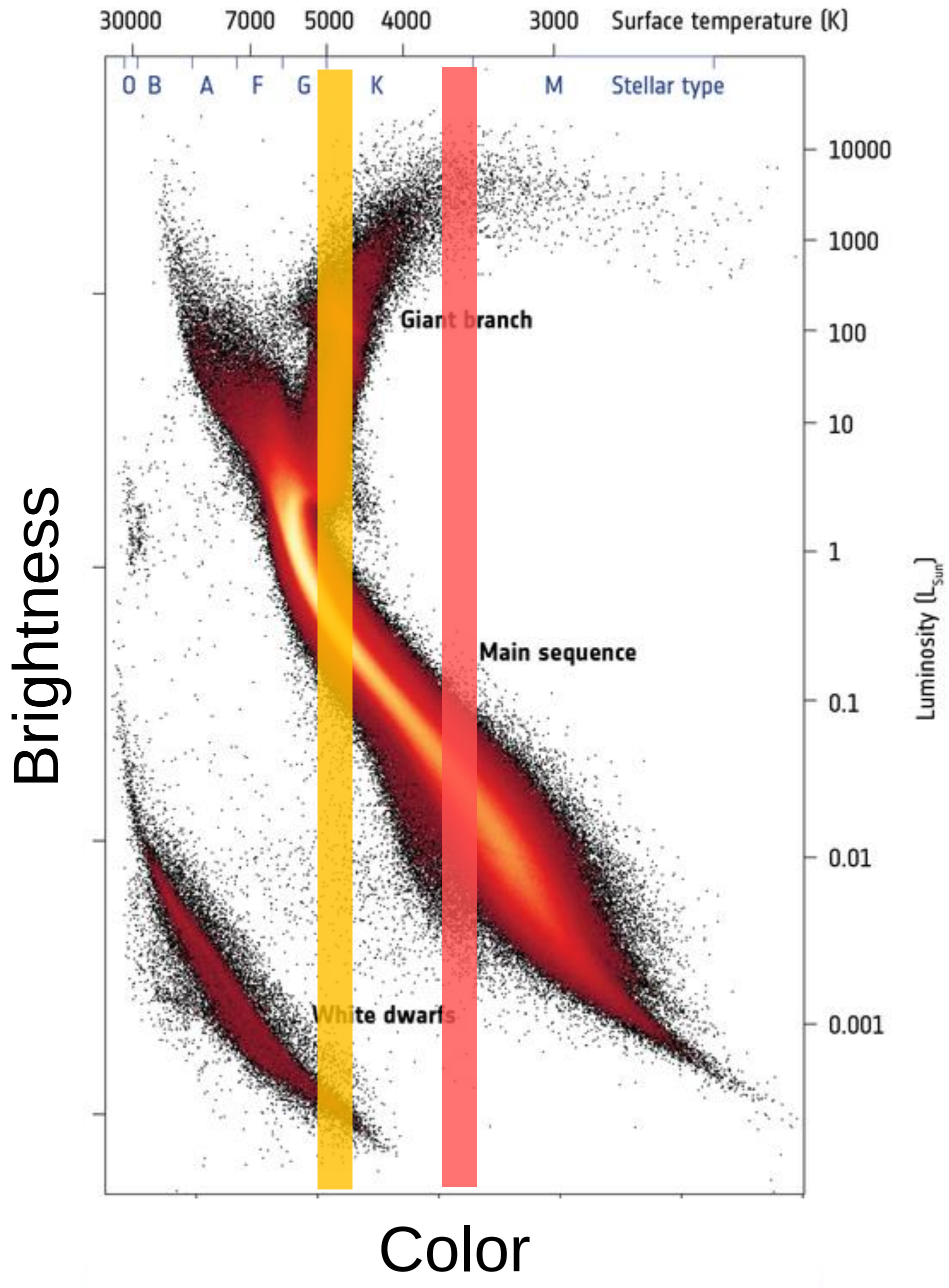


Color

Stellar Confusion

- **Low-mass solution**
  - Less reddening
  - Closer
- **High-mass solution**
  - More reddening
  - Further

Credit: ESA / Gaia / DPAC



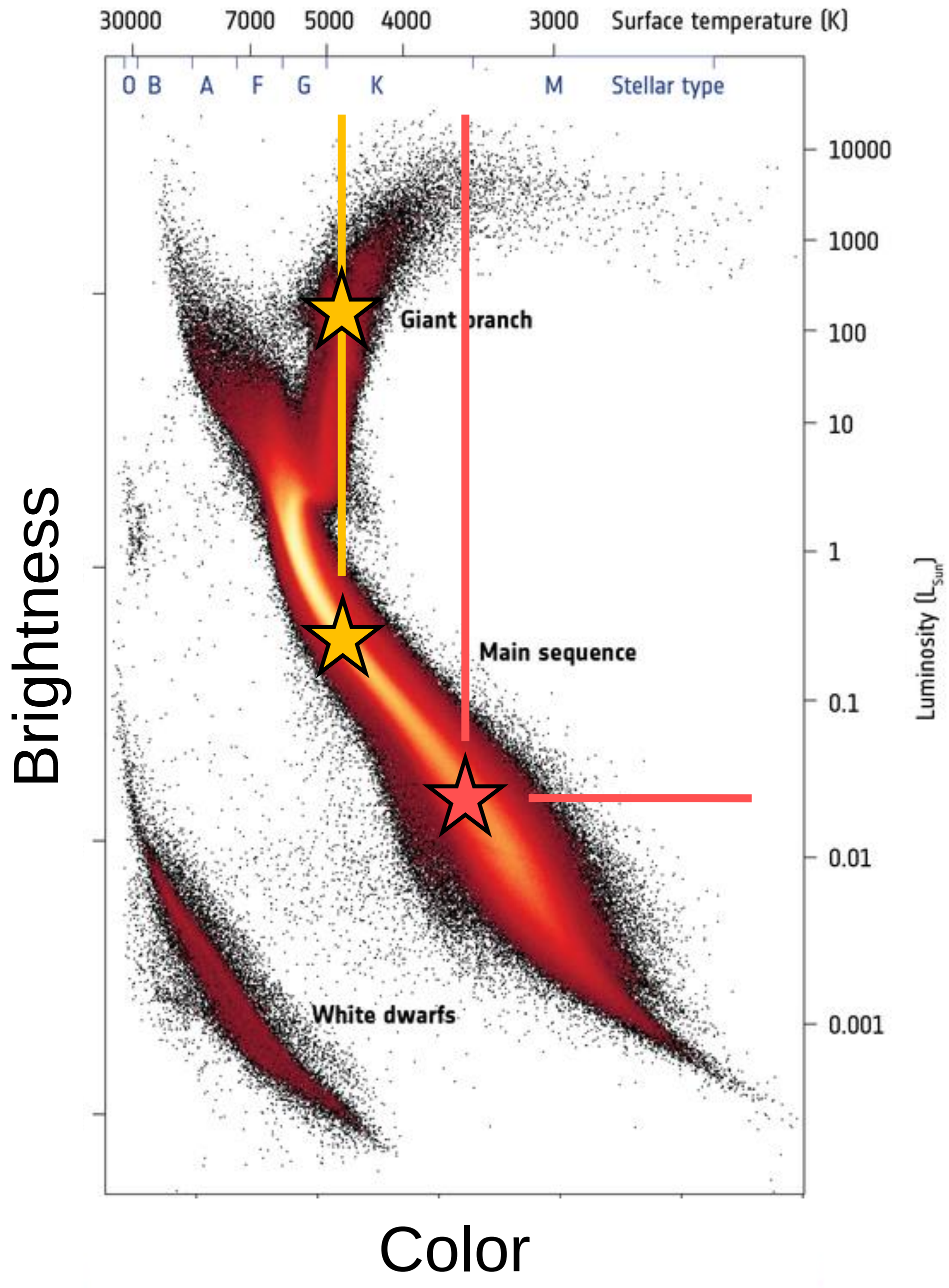
Credit: ESA / Gaia / DPAC

Stellar Confusion

- **Low-mass solution**
  - Less reddening
  - Closer

- **High-mass solution**
  - More reddening
  - Further
  - **Dwarf vs giant?**

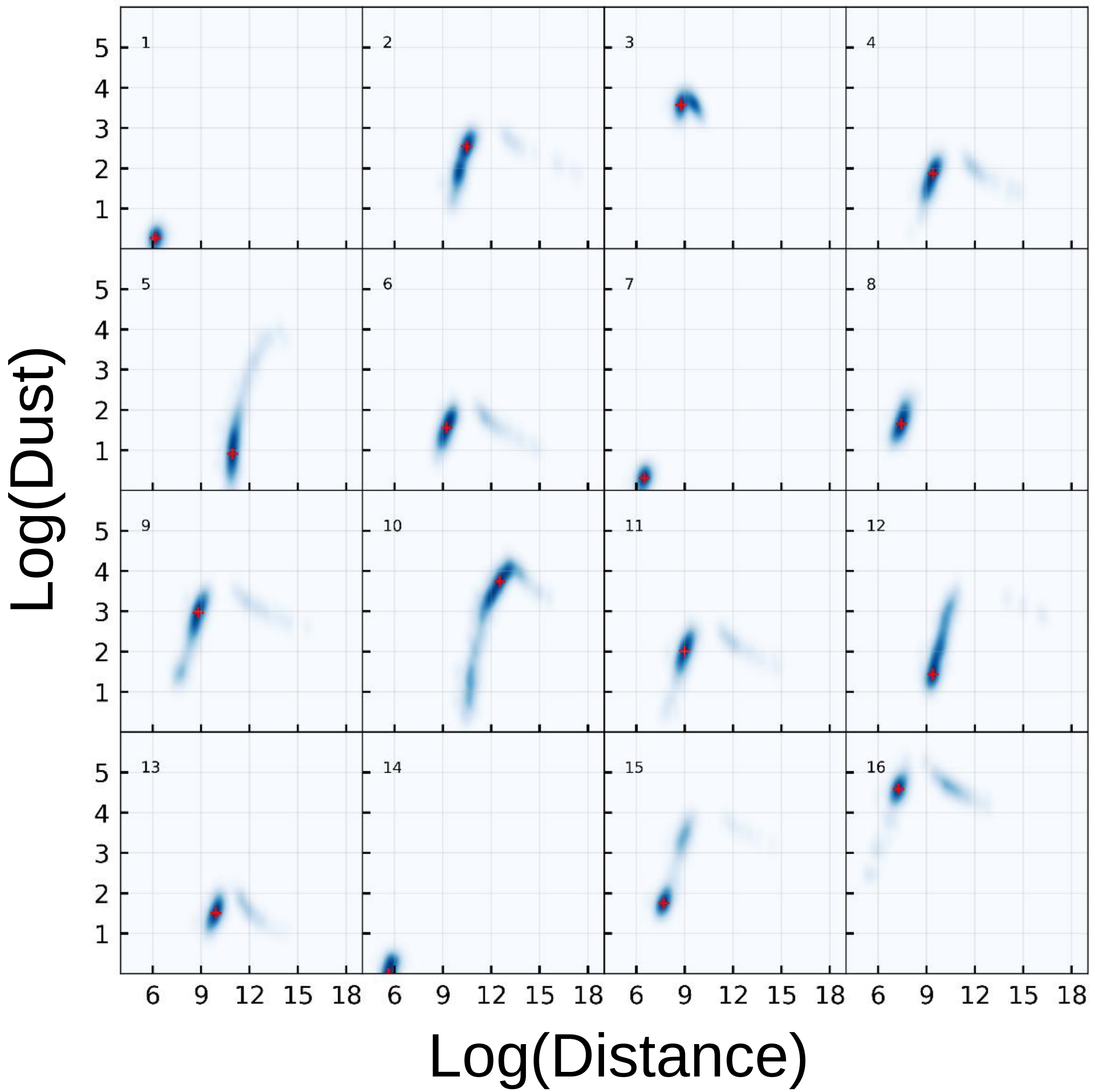
    (Sun) (Betelgeuse)



The Result?

- Inference must cope with:
  - Large uncertainties
  - Complex degeneracies
  - Multi-modal distributions

- Widespread application therefore requires:
  - Larger sample sizes
  - Improved inference
  - Significant computing power



Zucker & Speagle et al. (2019)

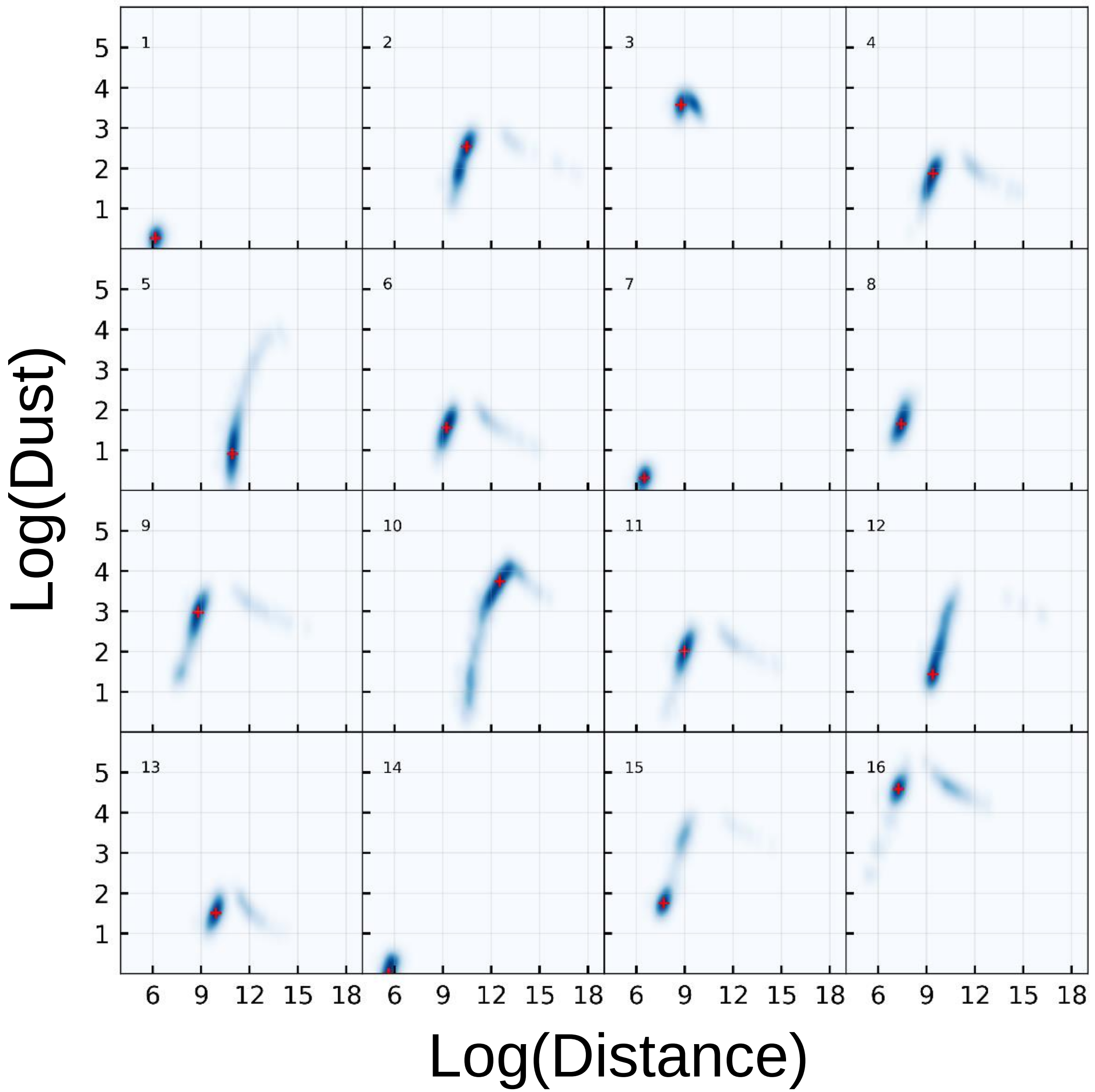
Part 2
Stellar Inference with Photometry and Astrometry

Requirements

Need an approach that:

1. Can characterize **multi-modal**, complex parameter degeneracies robustly.

2. Can perform inference in **seconds** per object to scale up to >100M stars.



Zucker & Speagle et al. (2019)

Stellar Inference

- Use a **Bayesian approach** to simultaneously estimate:
  - **Intrinsic** parameters of a star (θ) such as **intrinsic brightness**
  - **Extrinsic** parameters of a star (ϕ) such as **distance and dust**

Measured parallax

$$p(\theta, \phi | m, \varpi) \propto L(\theta, \phi | m) \times L(\phi | \varpi) \times \pi(\theta, \phi)$$

“Galactic model” prior

Observed brightness from
images at different wavelengths

Speagle et al. (2022b, subm.)

Model Likelihood: Photometry (Images)

- A combination of non-linear and linear dependencies with Normally-distributed measurement errors.

Intrinsic stellar brightness

$$m = m_{\text{int}}(\theta) + \beta(\theta) \cdot \phi + \epsilon_m \qquad \epsilon_m \sim N(0, \Sigma_m)$$

Linear coefficients

Speagle et al. (2022b, subm.)

Model Likelihood: Astrometry (Parallaxes)

- Errors are Normally-distributed.

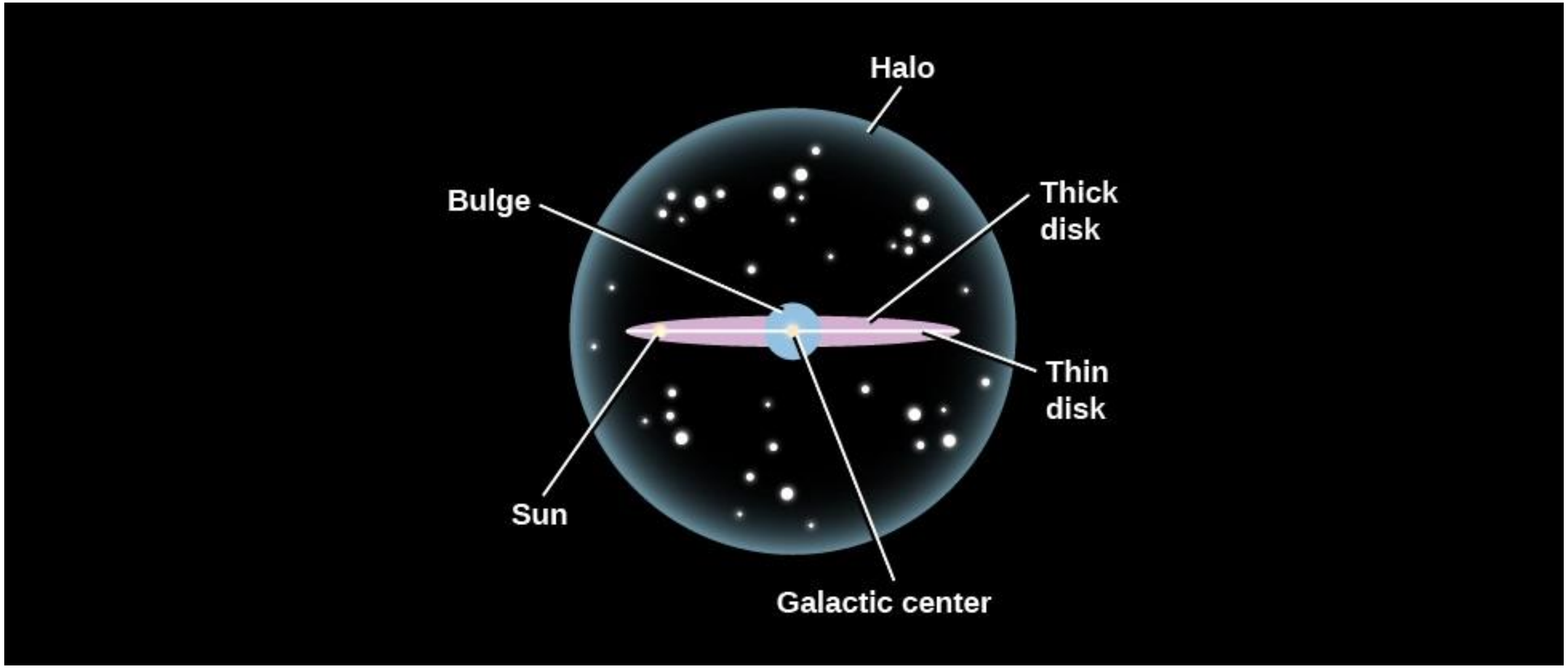
$$\varpi = \underbrace{\frac{1}{d(\phi)}}_{\text{Distance}} + \epsilon_{\varpi}$$

$$\epsilon_{\varpi} \sim N(0, \sigma_{\varpi})$$

Speagle et al. (2022b, subm.)

Model Prior: The “Galactic Model”

- Assumes the distribution of stars and their properties follows a three-component model with a “thin disk”, “thick disk”, and “halo”.



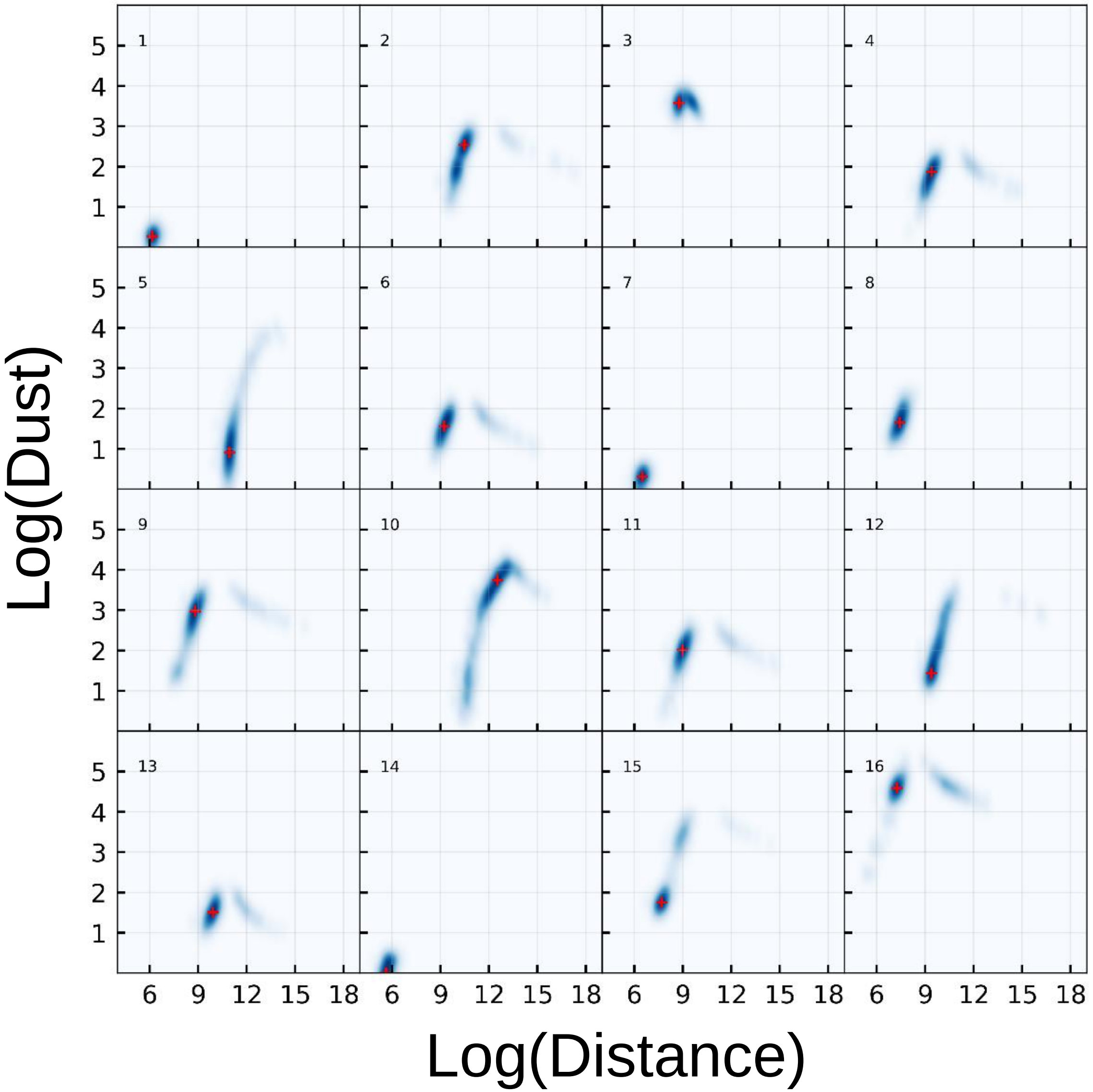
Picture Credit: Lumen Learning

Speagle et al. (2022b, subm.)

Sampling the Posterior

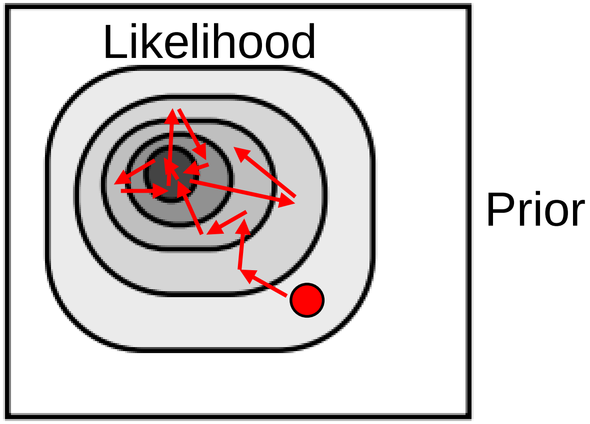
- Typical MCMC methods struggle with efficiently exploring these distributions.

- **Solution 1: Nested Sampling**



Zucker & Speagle et al. (2019)

Nested Sampling Motivation



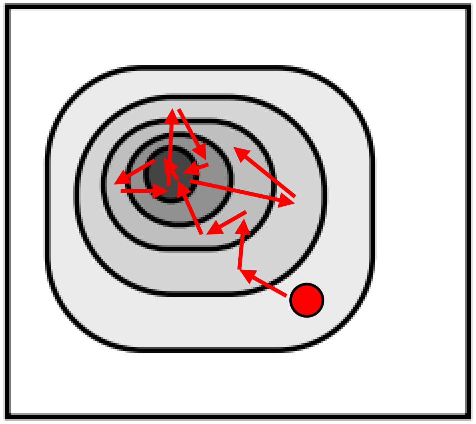
Sampling directly from the likelihood $\mathcal{L}(\mathbf{\Theta})$ is **hard**.

MCMC: Solving a Hard Problem once.

(Markov Chain Monte Carlo)

Pictures adapted from [this 2010 talk] by John Skilling.

Nested Sampling: Motivation



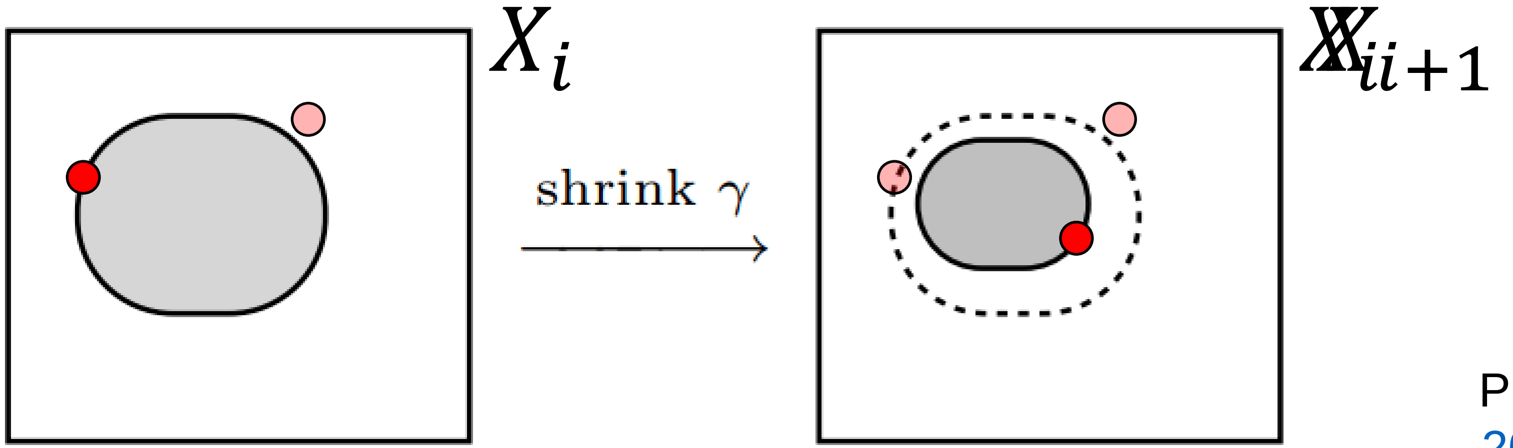
MCMC: Solving a Hard Problem once.

vs

Nested Sampling: Solving an Easier Problem many times.

Sampling $\boxed{\text{uniformly within}}$ $\mathcal{L}(\boldsymbol{\Theta}) > \lambda$ bound is **easier**.

If you have a **prior transform** that converts your priors to look uniform, then this case is equivalent.



Pictures adapted from this 2010 talk by John Skilling.

Nested Sampling

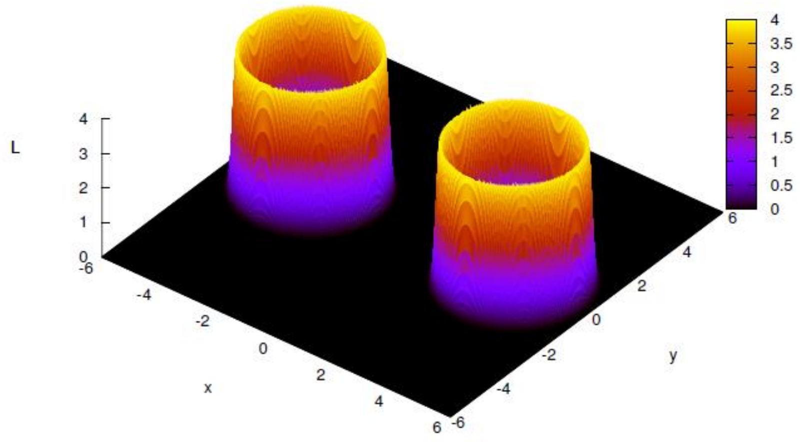
- Method originally designed to estimate the **marginal likelihood** (i.e. Bayesian evidence).

$$\mathcal{Z} \equiv \int_{\Omega_{\boldsymbol{\Theta}}} \mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})d\boldsymbol{\Theta}$$

Nested Sampling

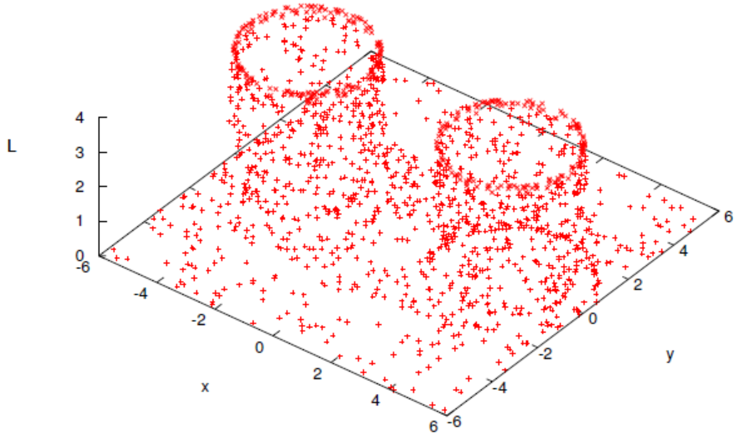
- Method originally designed to estimate the **marginal likelihood** (i.e. Bayesian evidence).
- Allows simultaneous **exploration of widely-separated modes**.

Underlying distribution



(a)

Final samples

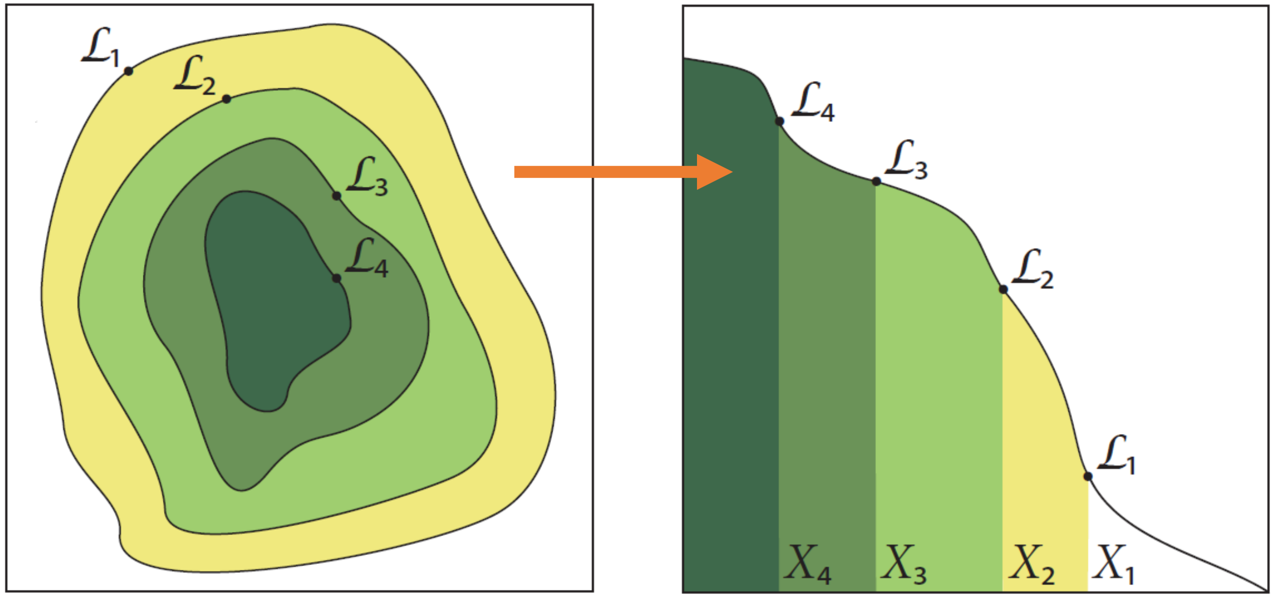


(b)

Feroz et al. (2013)

Nested Sampling

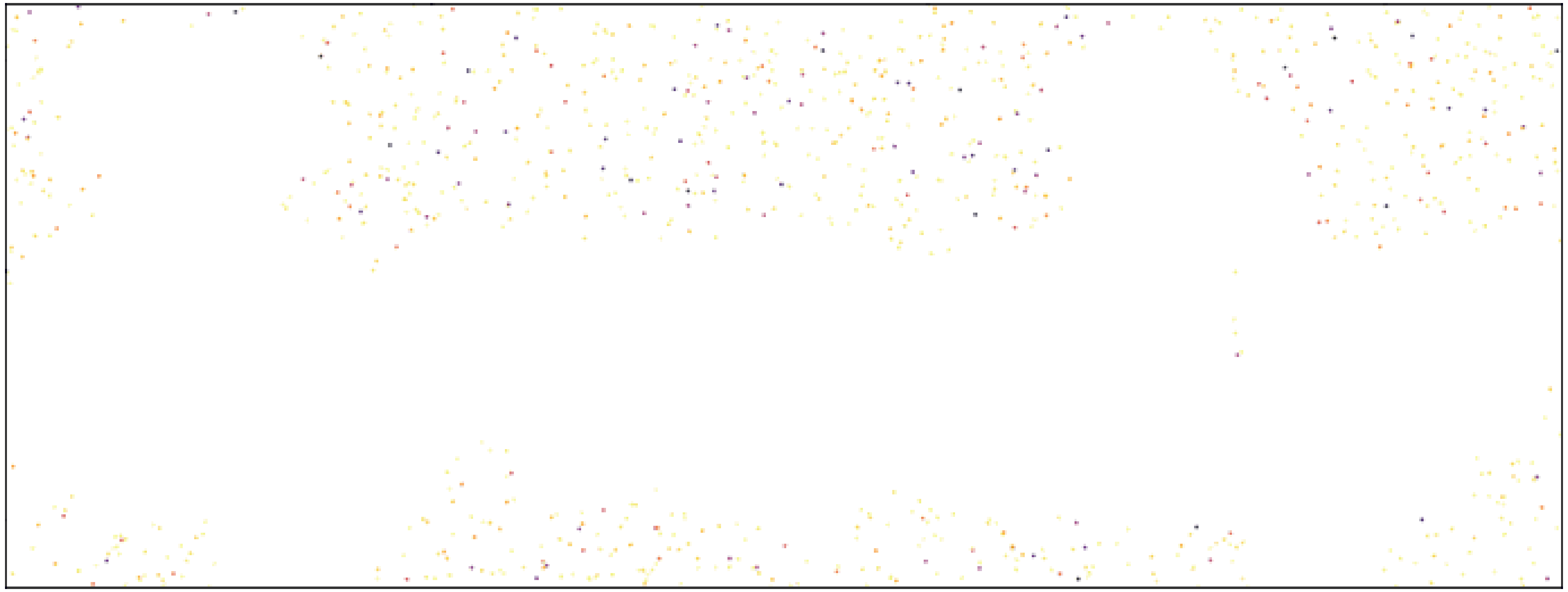
- Method originally designed to estimate the **marginal likelihood** (i.e. Bayesian evidence).
- Allows simultaneous **exploration of widely-separated modes**.
- Leverages **order statistics** to sample distribution, giving **well-behaved error properties**.



Feroz et al. (2013)

Implementation

- Nested Sampling implemented in open-source Python package **dynesty**.



@joshspeagle/dynesty on GitHub

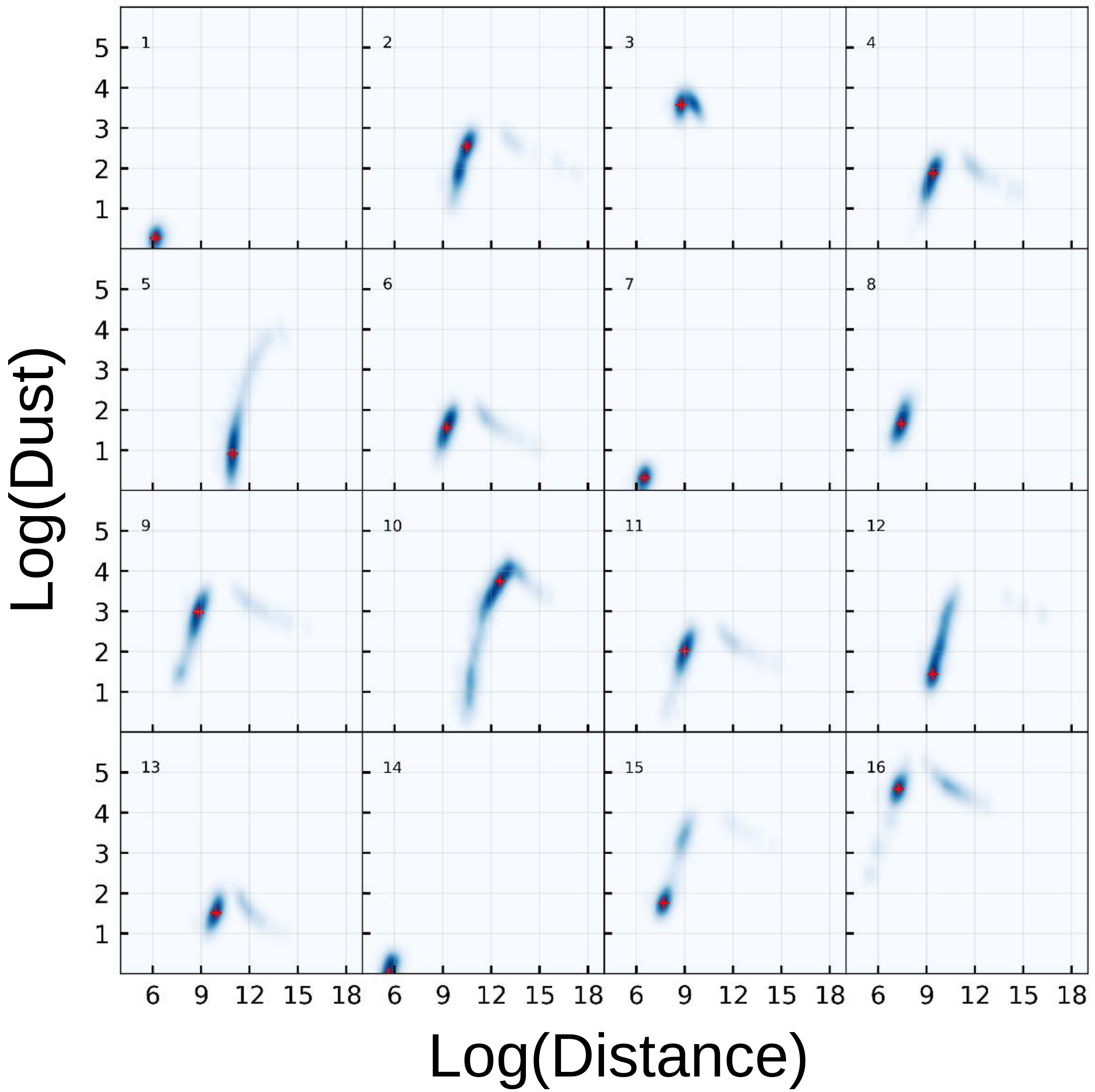
Speagle (2020)

Stellar Inference

Need an approach that:

1. Can characterize **multi-modal**, complex parameter degeneracies robustly.

2. Can perform inference in **seconds** per object to scale up to >100M stars.

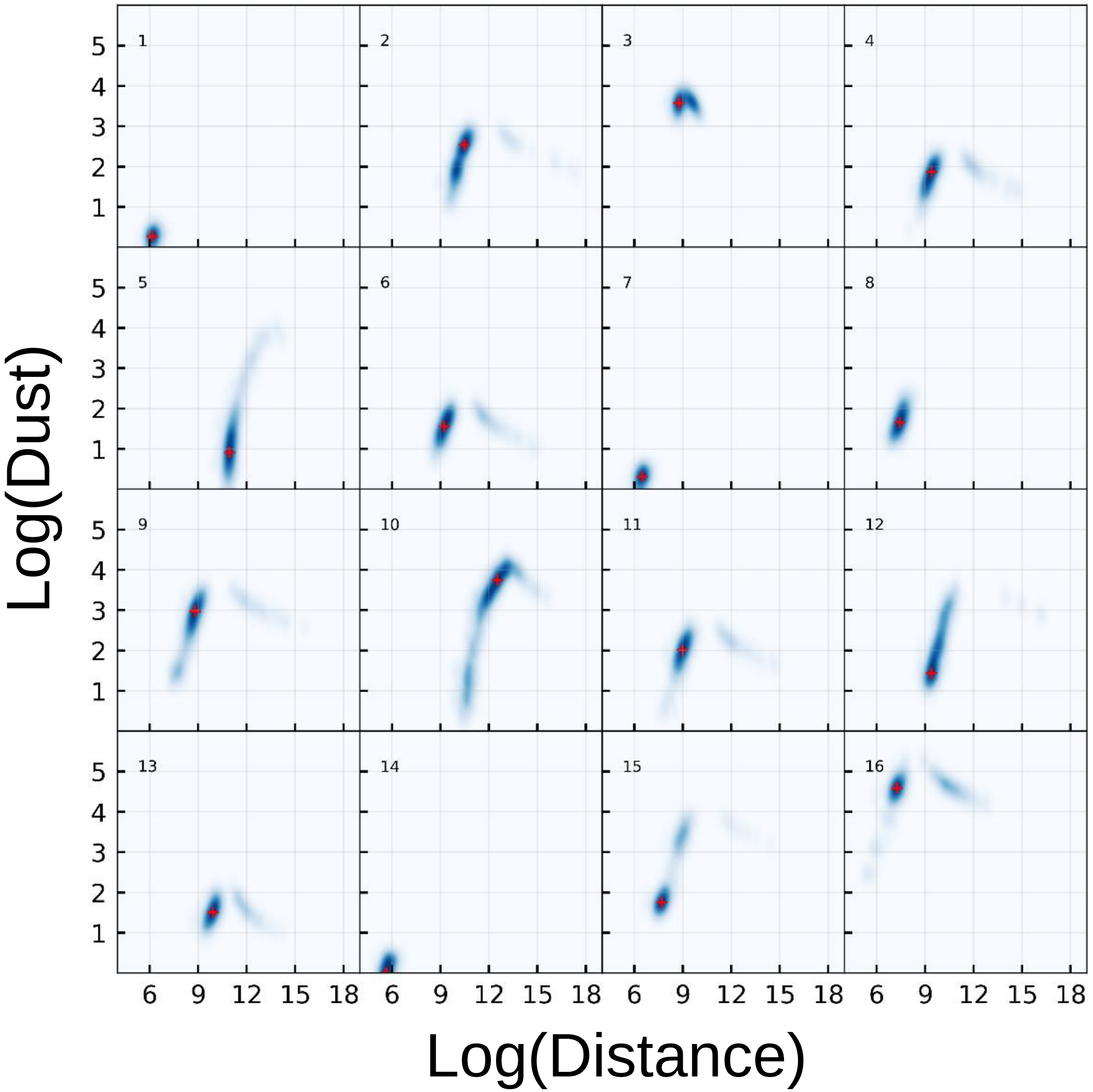


Zucker & Speagle et al. (2019)

Sampling the Posterior

- Typical MCMC methods struggled with efficiently exploring distributions.

- **Solution 2: tailored approach**



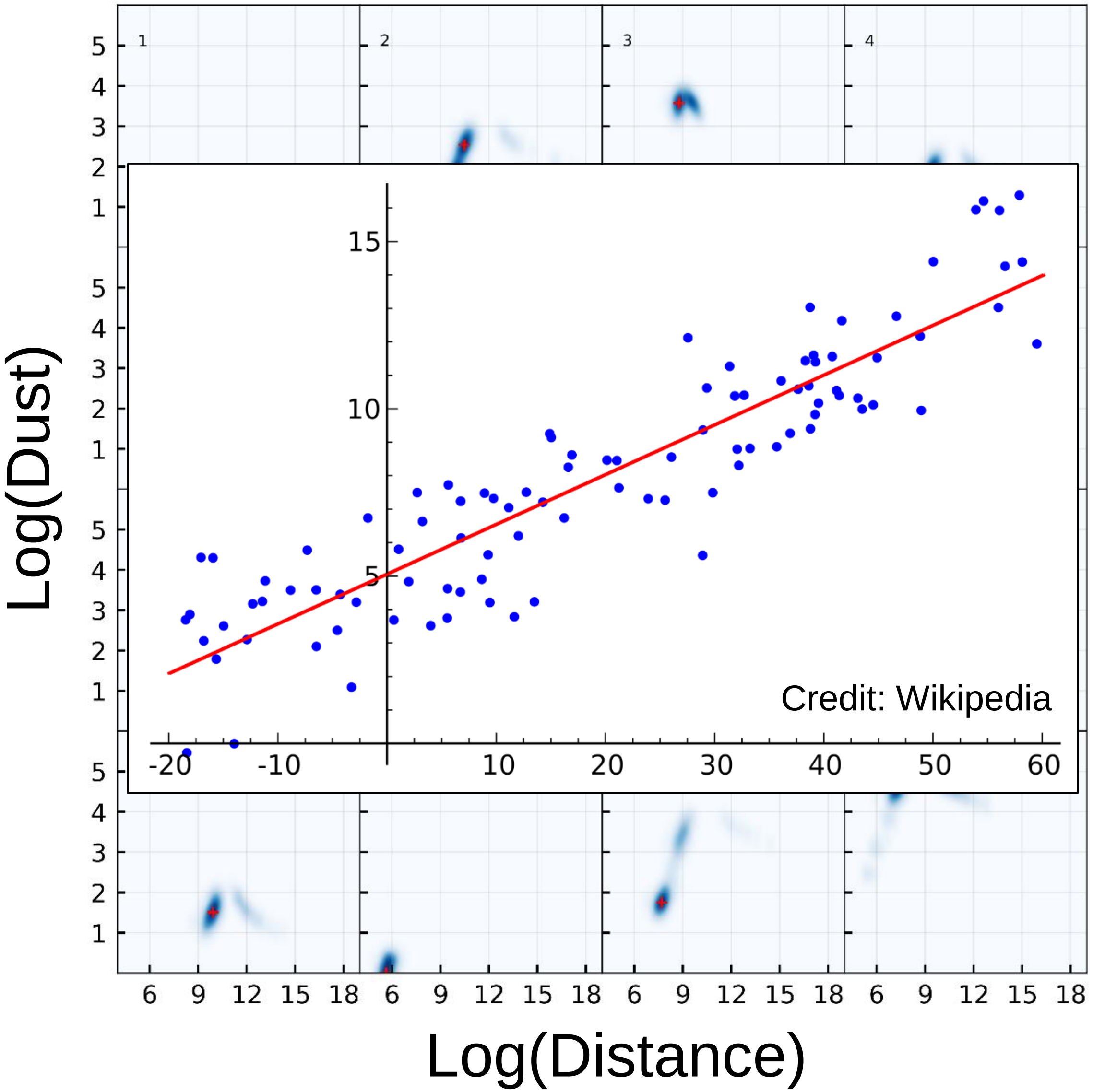
Speagle et al. (2022b, subm.)

Sampling the Posterior

- Typical MCMC methods struggled with efficiently exploring distributions.

- **Solution 2:** "Intrinsic" parameters "Extrinsic" parameters
  - Conditioned on θ, solving for ϕ is a **linear regression** problem! Can analytically marginalize over ϕ.

$$m = m_{\text{int}}(\theta) + \beta(\theta) \cdot \phi + \epsilon_m$$



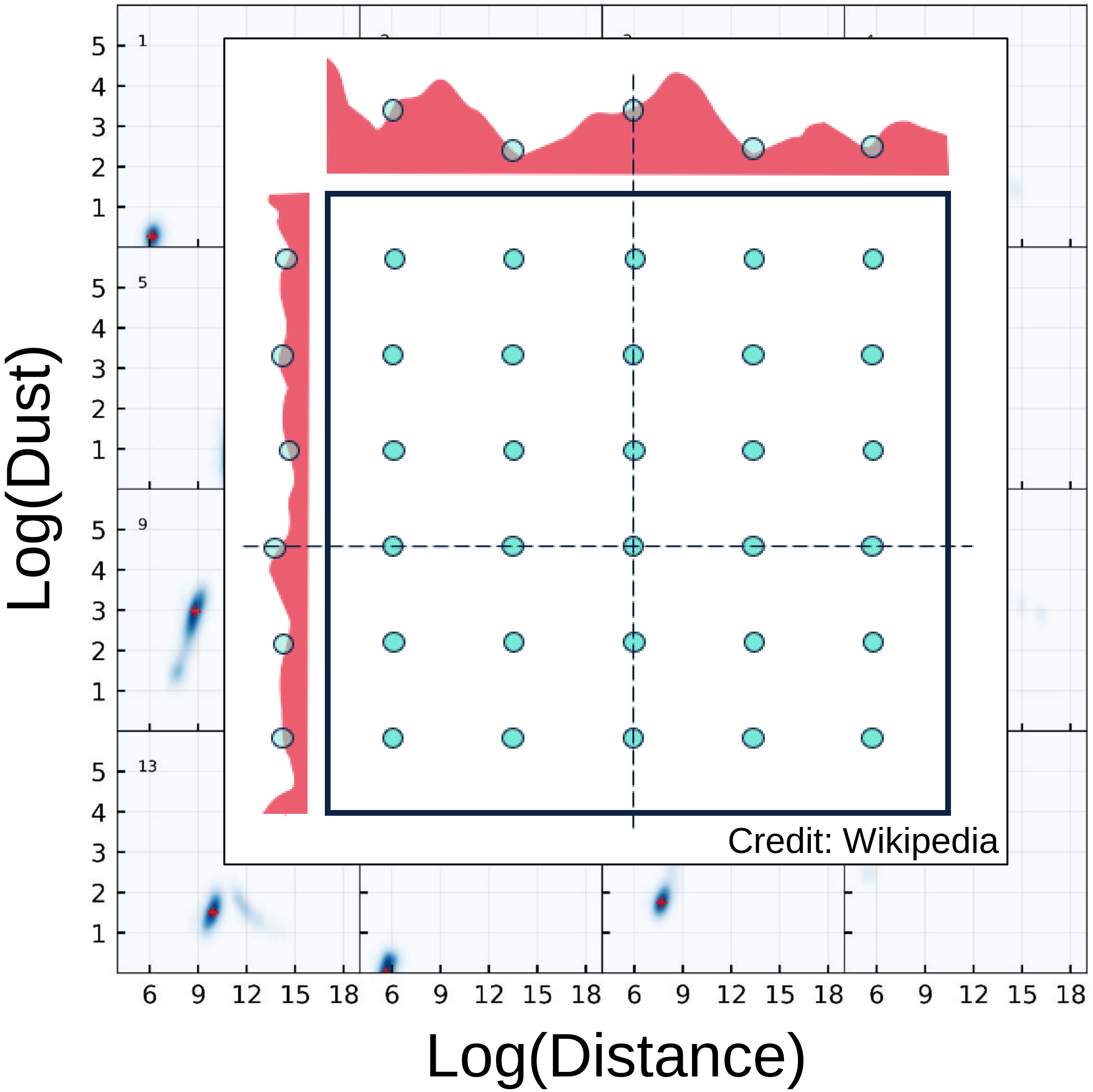
Speagle et al. (2022b, subm.)

Sampling the Posterior

- Typical MCMC methods struggled with efficiently exploring distributions.

- **Solution 2:** "Intrinsic" parameters "Extrinsic" parameters
  - Conditioned on θ, solving for ϕ is a **linear regression** problem! Can analytically marginalize over ϕ.
  - Since θ is low-dimensional (≤ 4), can explore efficiently using a **brute-force** grid.

$$m = m_{\mathrm{int}}(\theta) + \beta(\theta) \cdot \phi + \epsilon_m$$



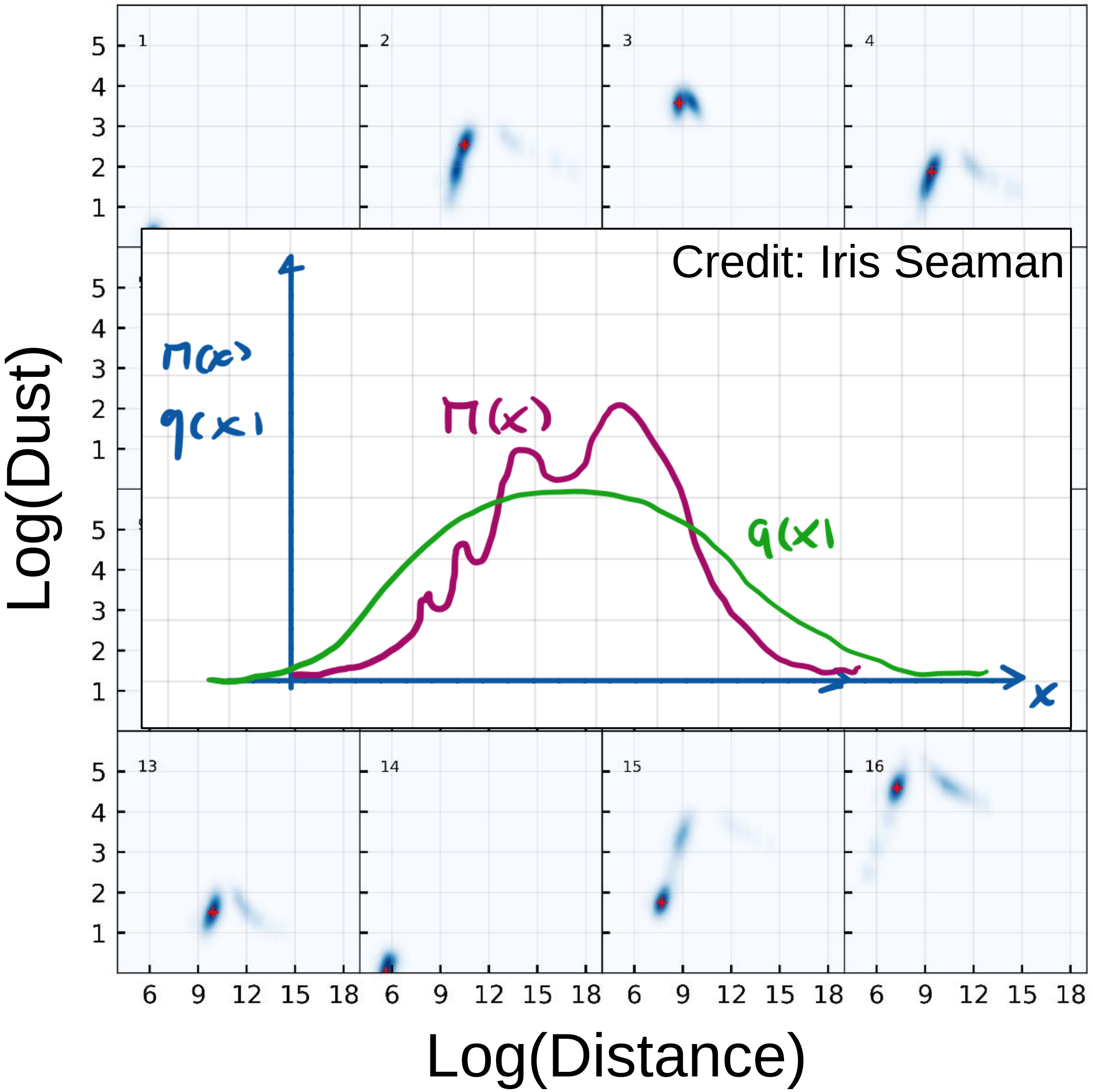
Speagle et al. (2022b, subm.)

Sampling the Posterior

- Typical MCMC methods struggled with efficiently exploring distributions.

- **Solution 2:** "Intrinsic" parameters "Extrinsic" parameters
  - Conditioned on θ, solving for ϕ is a **linear regression** problem! Can analytically marginalize over and sample for ϕ.
  - Since θ is low-dimensional (≤ 4), can explore efficiently using a **brute-force** grid.
  - Add data from parallax and prior using **importance sampling** to "reweight" points.

$$L(\theta, \phi | m) \times \boxed{L(\phi | \varpi) \times \pi(\theta, \phi)}$$



Speagle et al. (2022b, subm.)

Implementation

- Open-source Python package.
- Includes built-in plotting utilities.
- Release planned in coming weeks!

Credit: Catherine Zucker

Visualizing Results

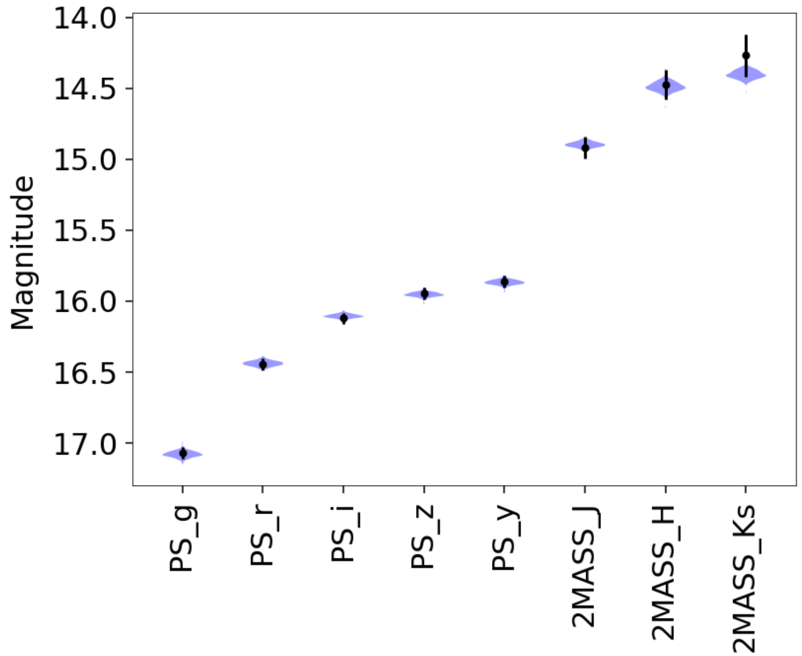
brutus contains a number of utilities for visualizing results in the plotting module. We illustrate a few below.

In [113]:
```
from brutus import plotting as bplot
```

One useful check is to examine the **posterior predictive**, which in this case just means the probabilistic SED derived from the posterior.

In [114]:
```
# pick an object
i = 3

# plot SED (posterior predictive)
fig, ax, parts = bplot.posterior_predictive(models_mist,  # stellar model grid
                                            idxs_mist[i],  # model indices
                                            reds_mist[i],  # A(V) draws
                                            dreds_mist[i],  # R(V) draws
                                            dists_mist[i],  # distance draws
                                            data=phot[i], data_err=err[i],  # data
                                            data_mask=mask[i],  # band mask
                                            offset=off_mist,  # photometric offsets
                                            psig=2.,  # plot 2-sigma errors
                                            labels=filt,  # filters
                                            vcolor='blue',  # "violin plot" colors for the posteriors
                                            pcolor='black')  # photometry colors for the data
```



@joshspeagle/brutus on GitHub

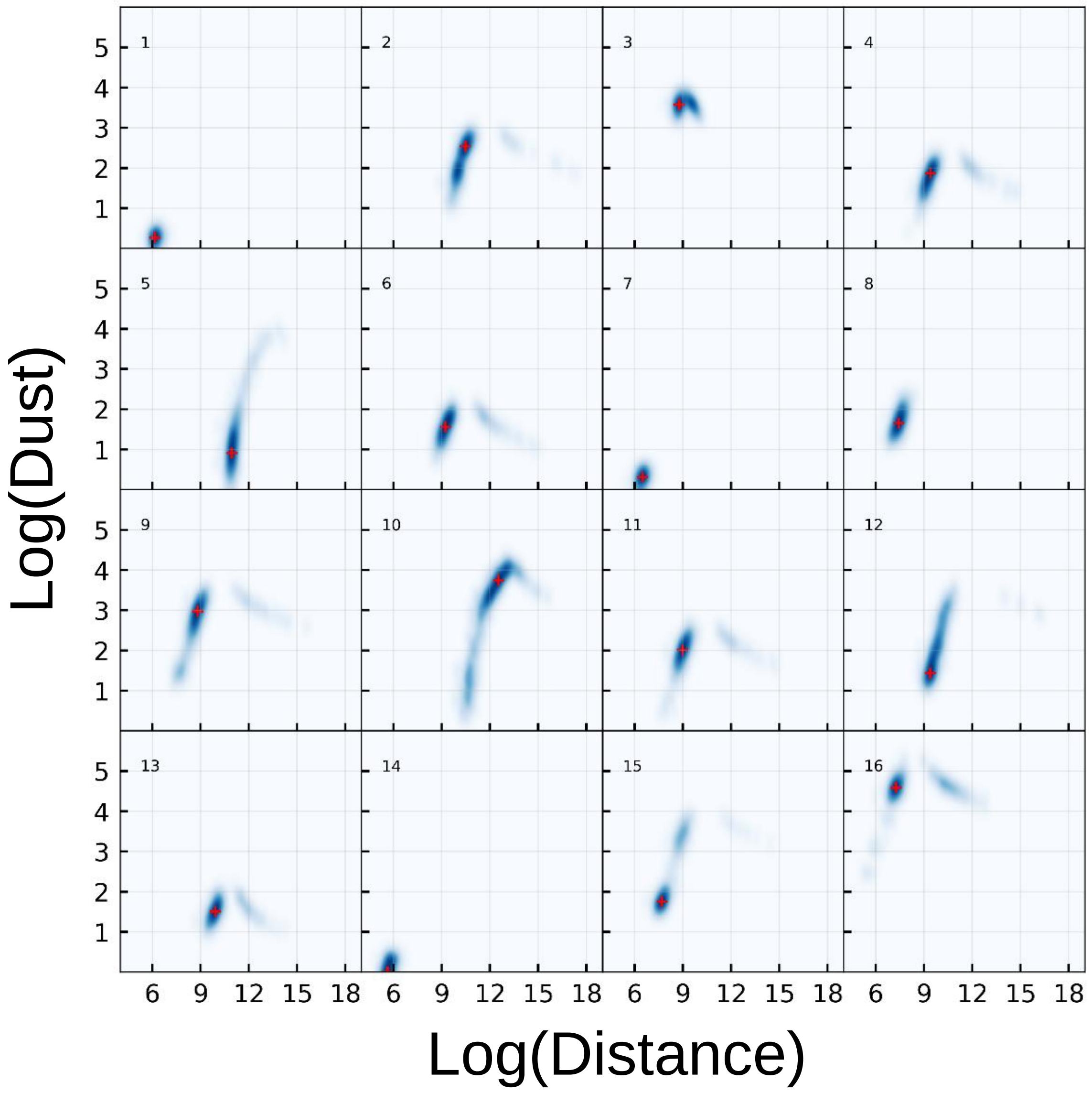
Speagle et al. (2022b, subm.)

Stellar Inference

Need an approach that:

1. Can characterize **multi-modal**, complex parameter degeneracies robustly.

2. Can perform inference in **seconds** per object to scale up to >100M stars.



Zucker & Speagle et al. (2019)

Part 3
A Few Applications

~170M objects/~700k CPU hours later…

Interactive visualization in **allsky**

http://allsky.s3-website.us-east-2.amazonaws.com

@joshspeagle/allsky on GitHub

Speagle et al. (2022a)