Data Archiving for Exoplanet Science

Angie Wolfgang
Assistant Research Professor
Pennsylvania State University
Center for Exoplanets and Habitable Worlds

Image credit:
NASA/JPL-Caltech/
R. Hurt (SSC-Caltech)

Key Questions for Exoplanets

(from the Exoplanet Science Strategy, NAS, 2018)

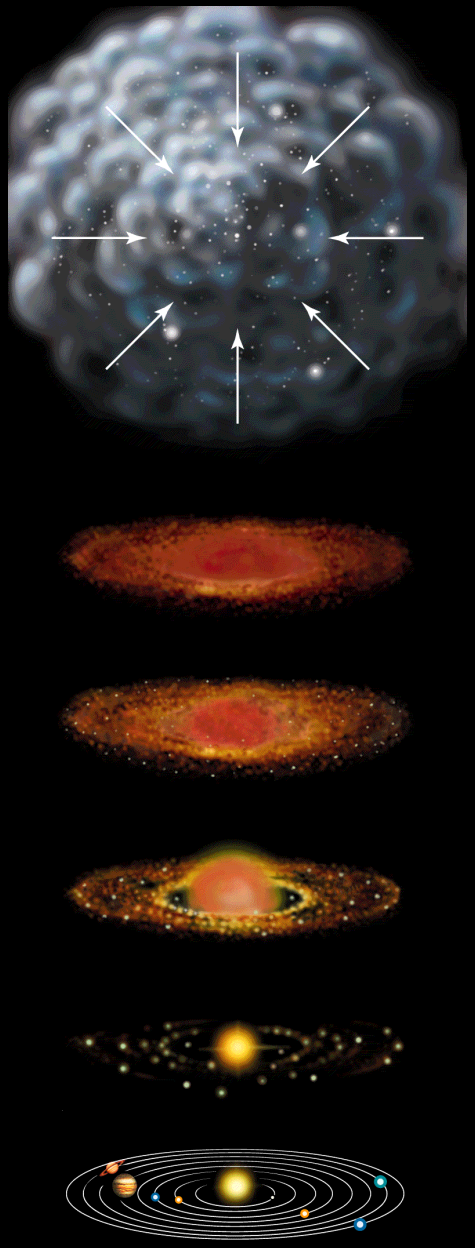
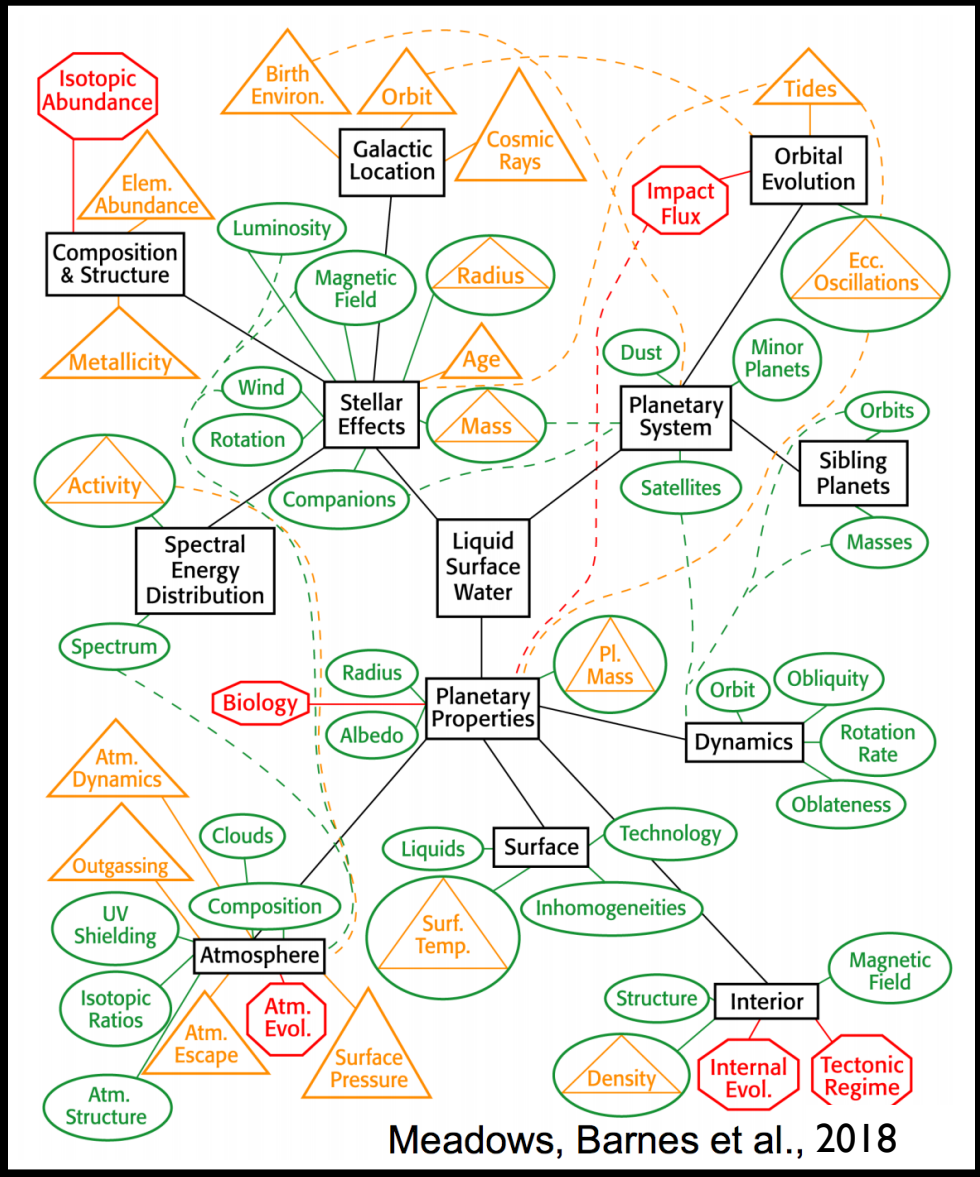


Image credit: astroclock2010.wordpress.com

1) How do planets form and evolve?

2) What characterizes a habitable planet, and do they host life?



Key Questions for Exoplanets

(from the Exoplanet Science Strategy, NAS, 2018)

1) How do planets form and evolve?

Goal 1: To understand the formation and evolution of planetary systems as products of the process of star formation, and characterize and explain the diversity of planetary system architectures, planetary compositions, and planetary environments produced by these processes.

To achieve this goal, the exoplanet science community needs to:

- Determine the range of planetary system architectures by surveying planets at a variety of orbital separations and searching for patterns in the structures of multiplanet systems;
- Characterize the diversity of bulk compositions and atmospheric compositions;
- Identify the parameters that determine which stars can form certain types of planetary systems; and
- Identify relationships between the planet formation process and the resulting planetary evolution, bulk composition, and atmospheric properties.

These activities require future advancements in data archiving and the archives' role in scientific analysis.

Key Questions for Exoplanets

(from the Exoplanet Science Strategy, NAS, 2018)

2) What characterizes a habitable planet, and do they host life?

Goal 2: To learn enough about the properties of exoplanets to identify potentially habitable environments and their frequency, and connect these environments to the planetary systems in which they reside. Furthermore, scientists need to distinguish between the signatures of life and those of nonbiological processes, and search for signatures of life on worlds orbiting other stars.

To achieve this goal and support the search for life in the galaxy, the following two areas of interdisciplinary research need to be developed:

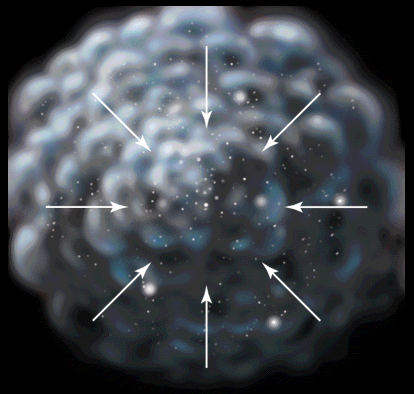
- A multiparameter habitability assessment for target selection; and
- A comprehensive framework for biosignature assessment.

In parallel with these theoretical underpinnings, a sequence of observational milestones need to be achieved to

- Identify and rank exoplanet targets;
- Characterize the environment, including the atmosphere, surface, and interior of the planet, and the spectrum and variability of the star; and
- Search for life in the context of the planetary and stellar environment.

This informs the content of data archives, but not as much their structure or the analysis services they provide.

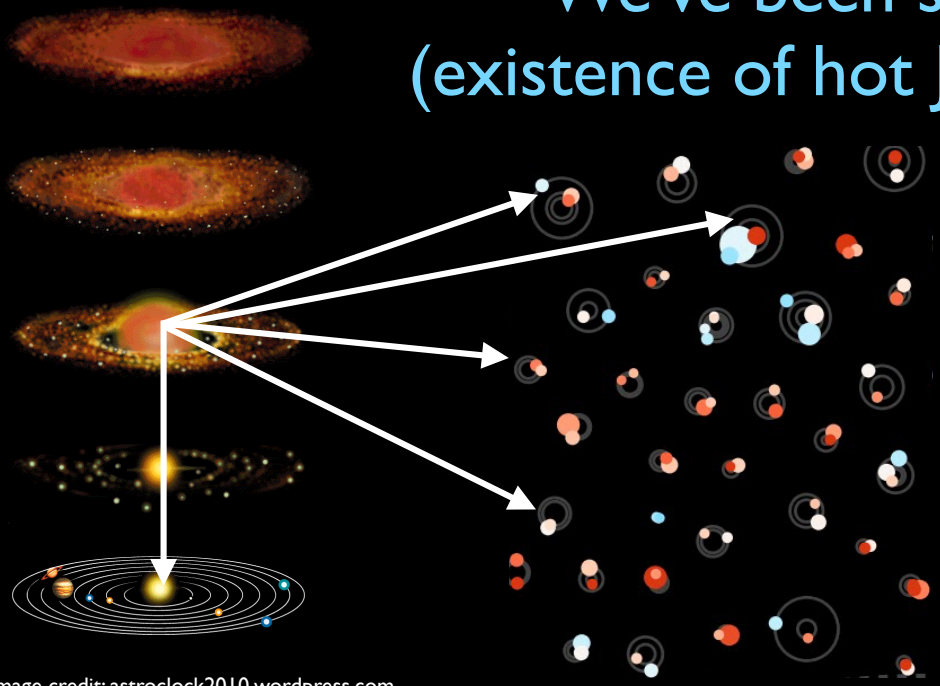
Why does understanding formation require advances in data archives?



Many physical processes are at work (accretion, migration, photoevaporation) and they can be stochastic (i.e. giant impacts).

⇒ We need $N_{sys}>1$ to test these theories!!

We've been surprised by exoplanets before (existence of hot Jupiters, plethora of sub-Neptunes)

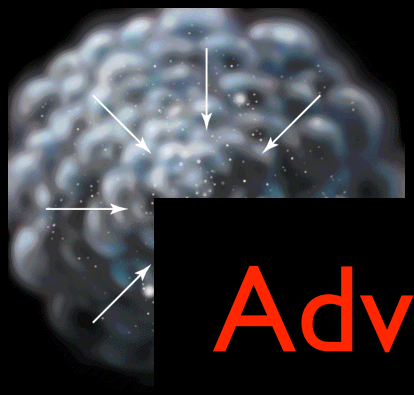


⇒ We must map out the diversity of extrasolar systems!!

Efficiently studying planet populations requires close collaboration with archives.

Image credit: astroclock2010.wordpress.com

Why does understanding formation require advances in data archives?



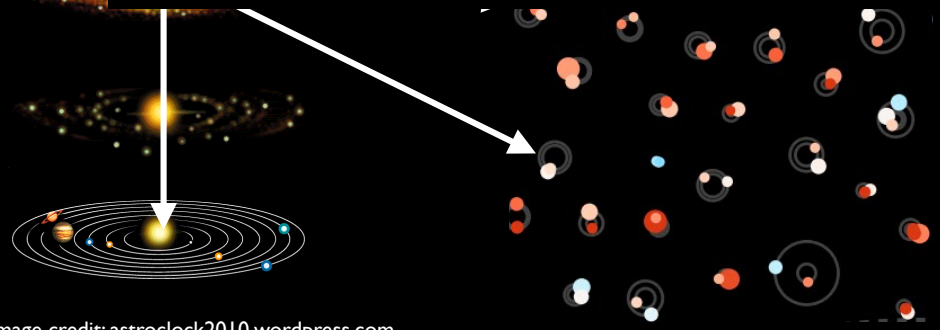
Many physical processes are at work (accretion, migration, photoevaporation) and they can be stochastic (i.e. giant impacts).

Advances in population analyses

advances in the archives

ınes)

ns!!

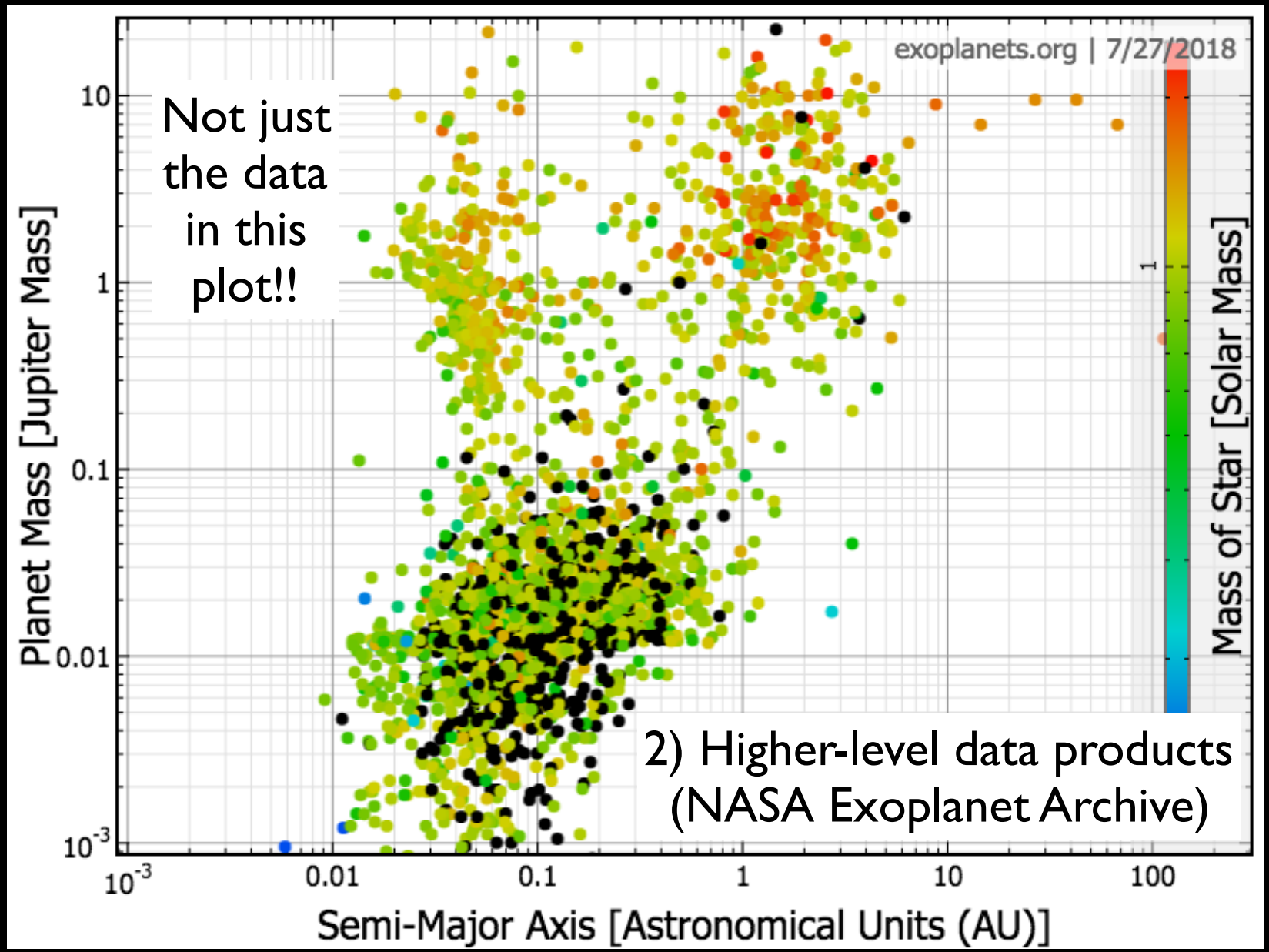


Efficiently studying planet populations requires close collaboration with archives.

Image credit: astroclock2010.wordpress.com

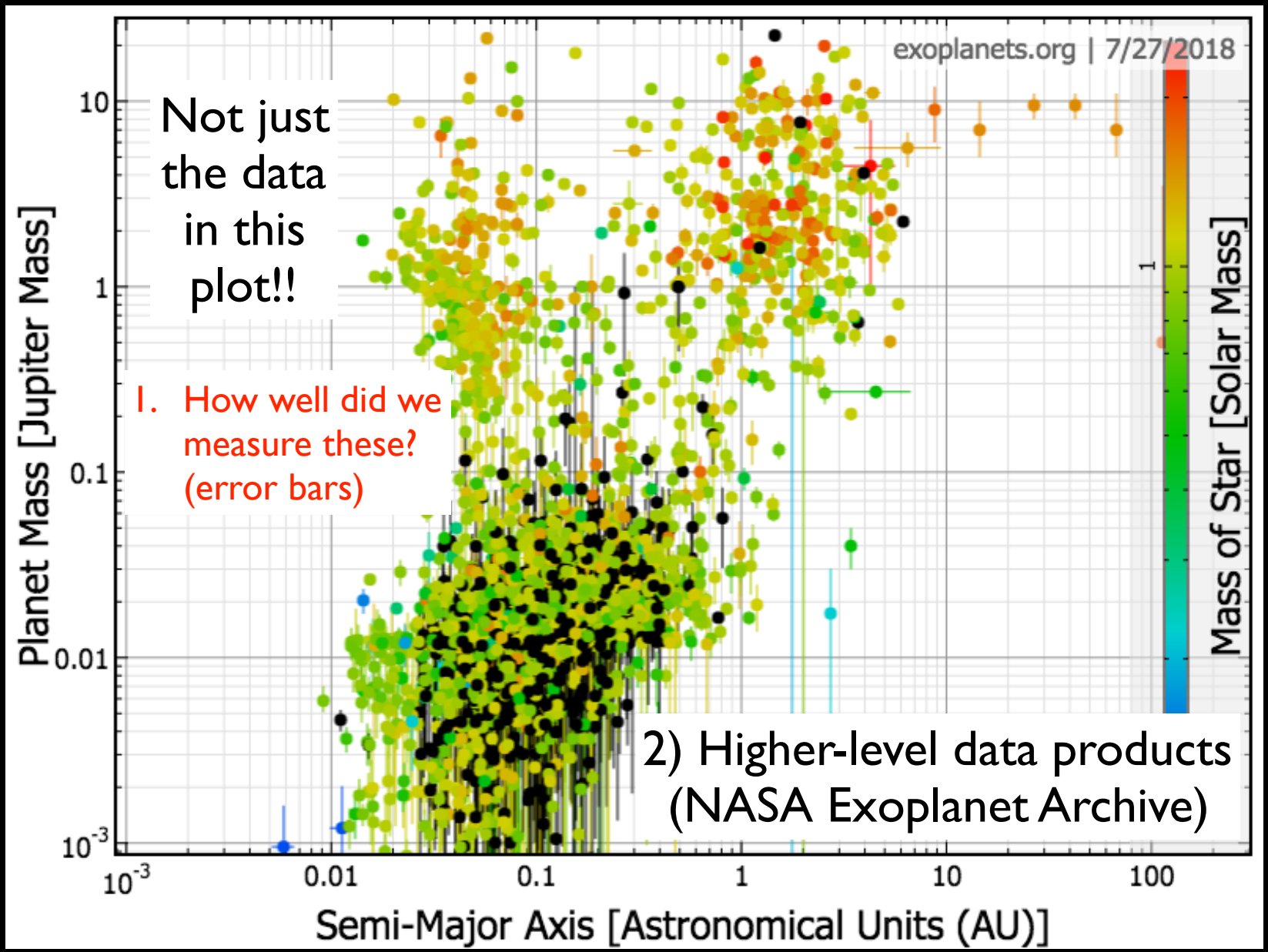
The archives currently provide:

1) Image and lightcurve-level data (MAST)



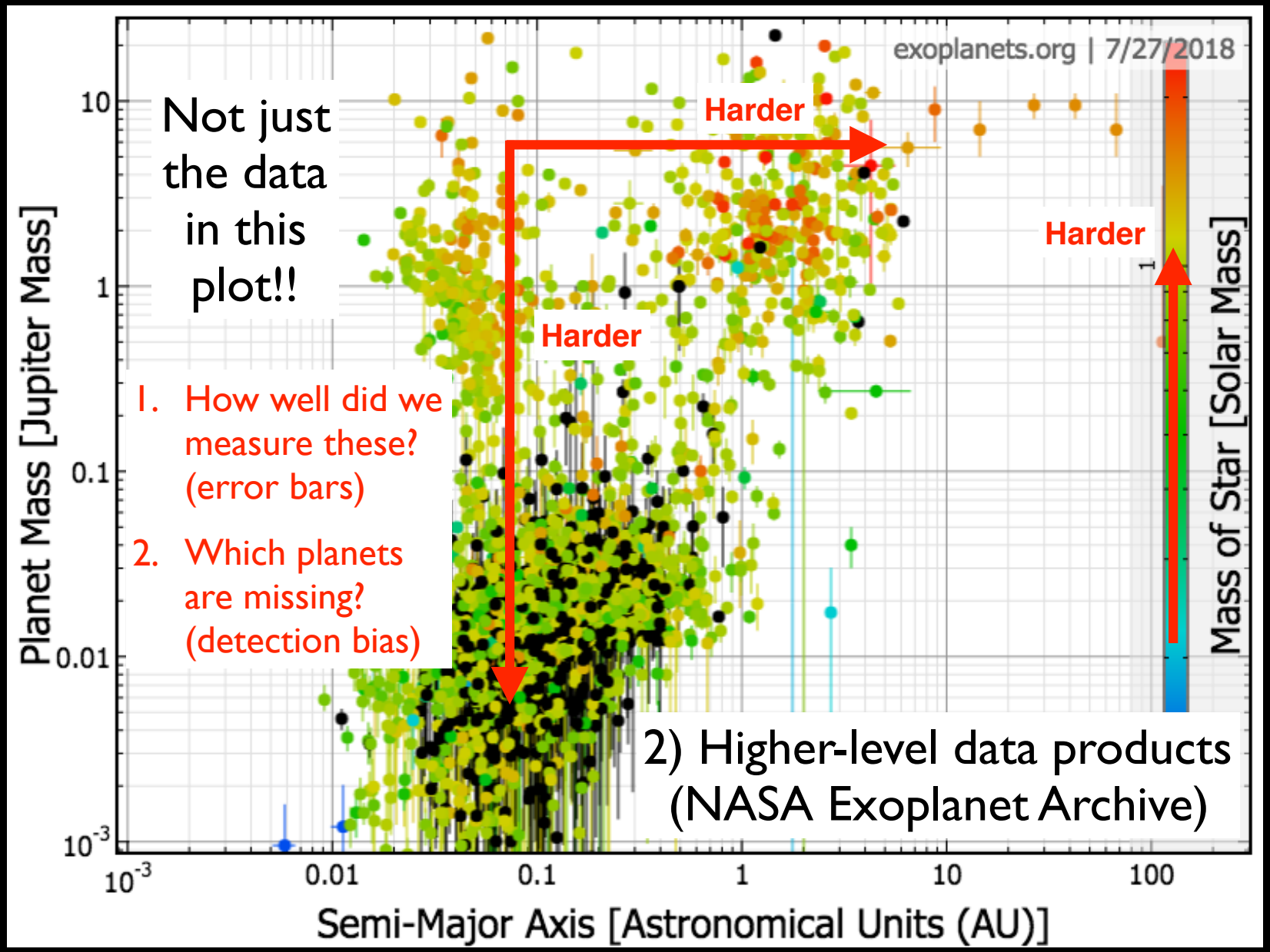
The archives currently provide:

1) Image and lightcurve-level data (MAST)

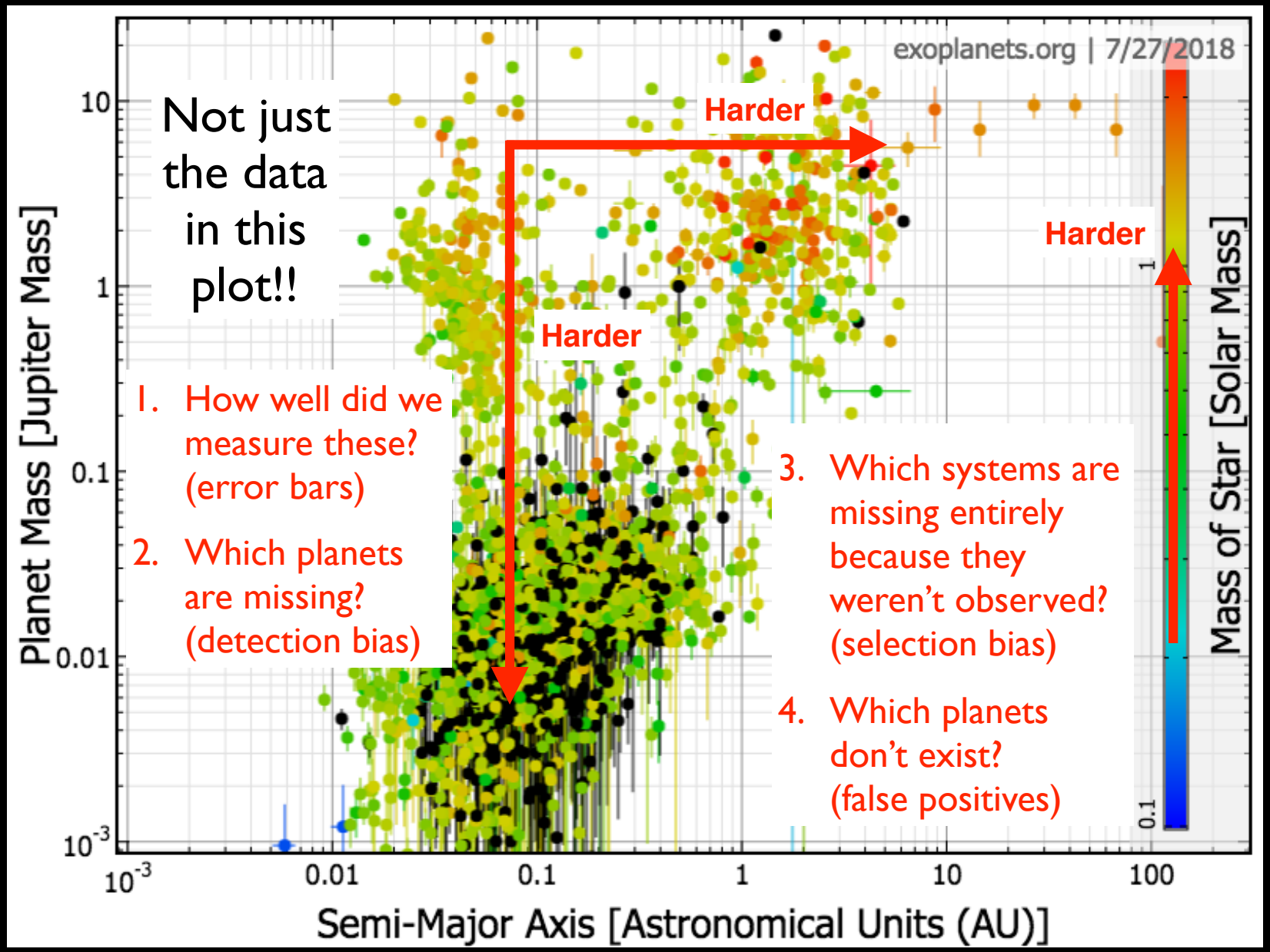


The archives currently provide:

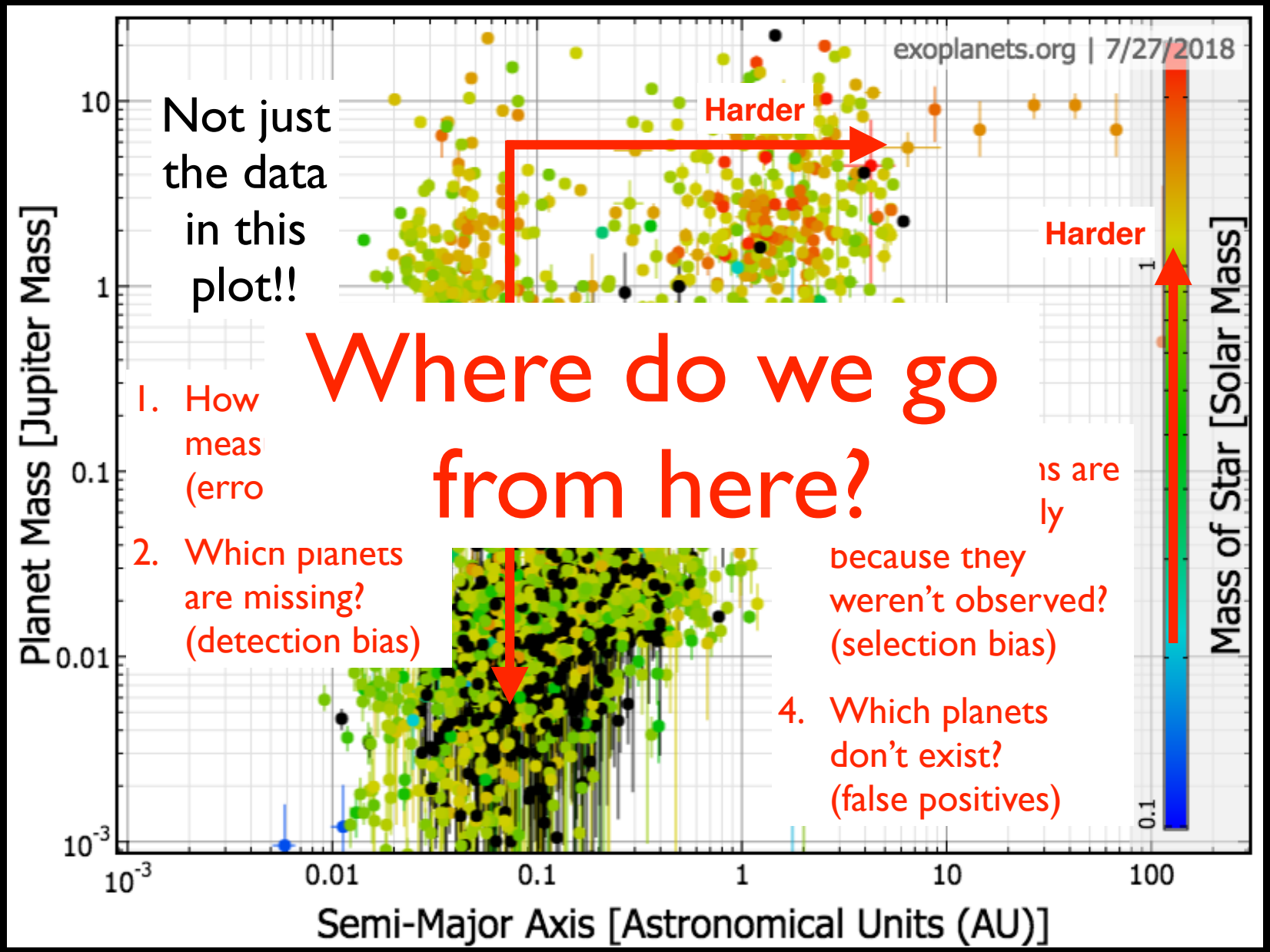
1) Image and lightcurve-level data (MAST)



Higher-level data products are necessary for population science!



Higher-level data products are necessary for population science!



Near-term advancements:

(with input from other Penn State CEHW members)

1) Archive *correlations* between parameter uncertainties, not just +/- error bar.
   - covariance matrices, Fischer Information, bootstrap samples
   - posterior samples (+ priors and likelihood function as metadata)
2) Improve accessibility of documentation
   - logistically (is finding it intuitive? how many clicks from the table?
   - metadata (which columns are derived from other columns, and using which equations? where do host stars' parameters come from? how are default planet parameters chosen?)
   - analysis services (use Jupyter notebooks to it's easier to customize; manage a blog with examples on how to use them)
3) Improve cross-matching with external archives (Gaia, CPS)
4) RV: link "cleaned" data directly to the derived parameters

Other Considerations

(with input from other Penn State CEHW members)

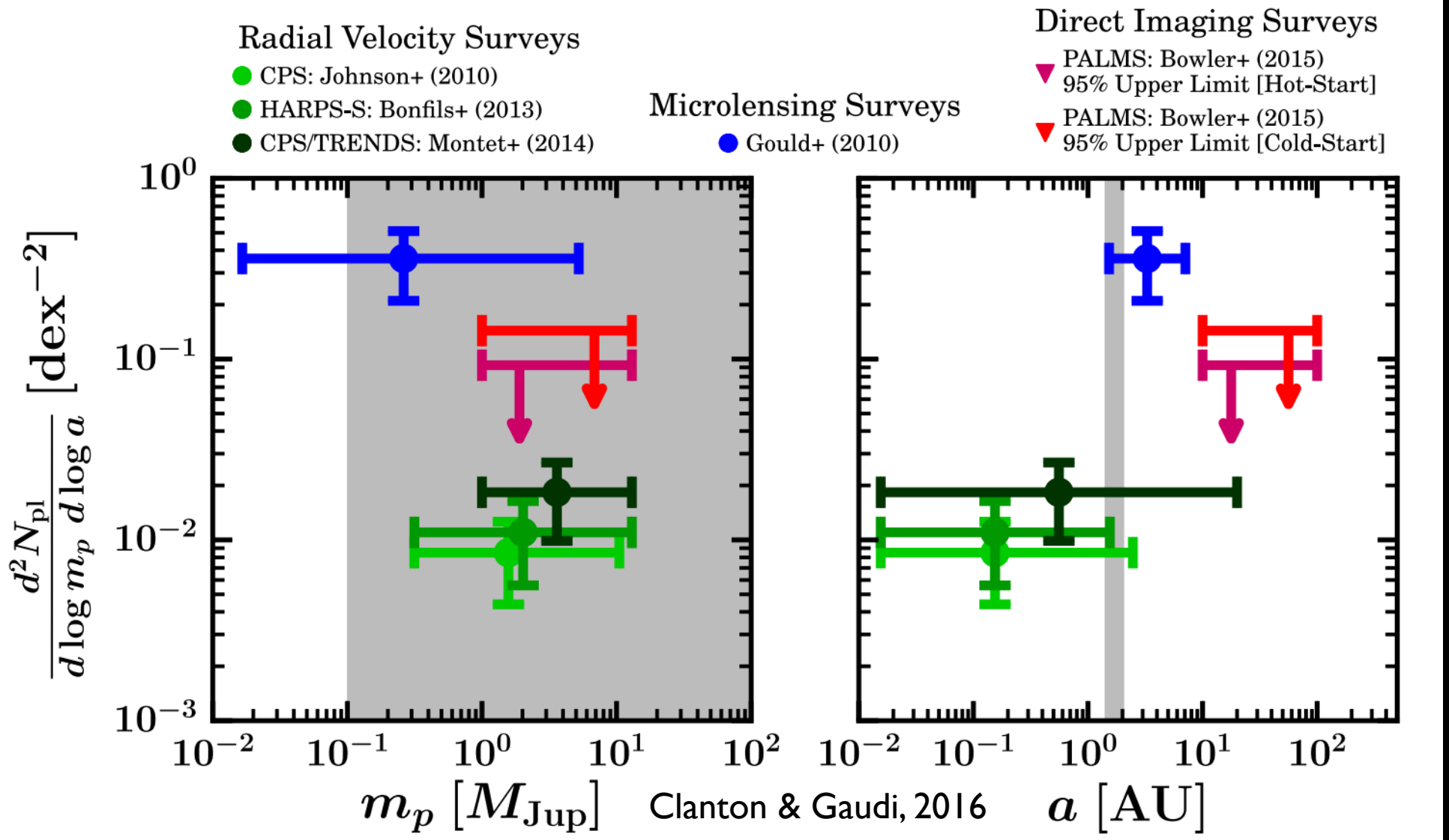
1) Software engineers really improve the product (exoplanets.org)
2) It is very easy to misuse/over-interpret high-level data products - caveats communicated on same page as table?
   (e.g. finding features in planet mass vs. planet-star distance without accounting for heterogenous data/selection effects/completeness)
3) Different people want different things → a range of products
   - Both point estimates *and* full posteriors are useful!
   - For RV, archive parameters from multiple models (# of planets, different parameterizations for stellar activity)
4) Reproducibility is hard ... house snapshots available via API?
   (Current solutions include keeping multiple dated versions of high-level products on local drive, but this is difficult to scale up.)
5) No community consensus on where to upload simulation data (NASA, NSF, institution-specific like PSUScholarSphere, GitHub)

Farther-term advancement

(with input from other Penn State CEHW members)

1) Instead of bringing data to code, take the code to data (science platforms)
   - Future planet detection will occur on image-level data
   - Will be helpful for population analyses on higher-level data once full posteriors are housed at the archives.
2) Efficient data exploration when visualization services live closer to the data → potential for new science?
   (TESS full-frame image portal + future planet detections = exploration of spatial clustering of planet occurrence??)
3) What should the archives do about data from theoretical simulations?
   (planet population synthesis, N-body output)
4) Construct efficient structures for combining high-level data products from multiple missions

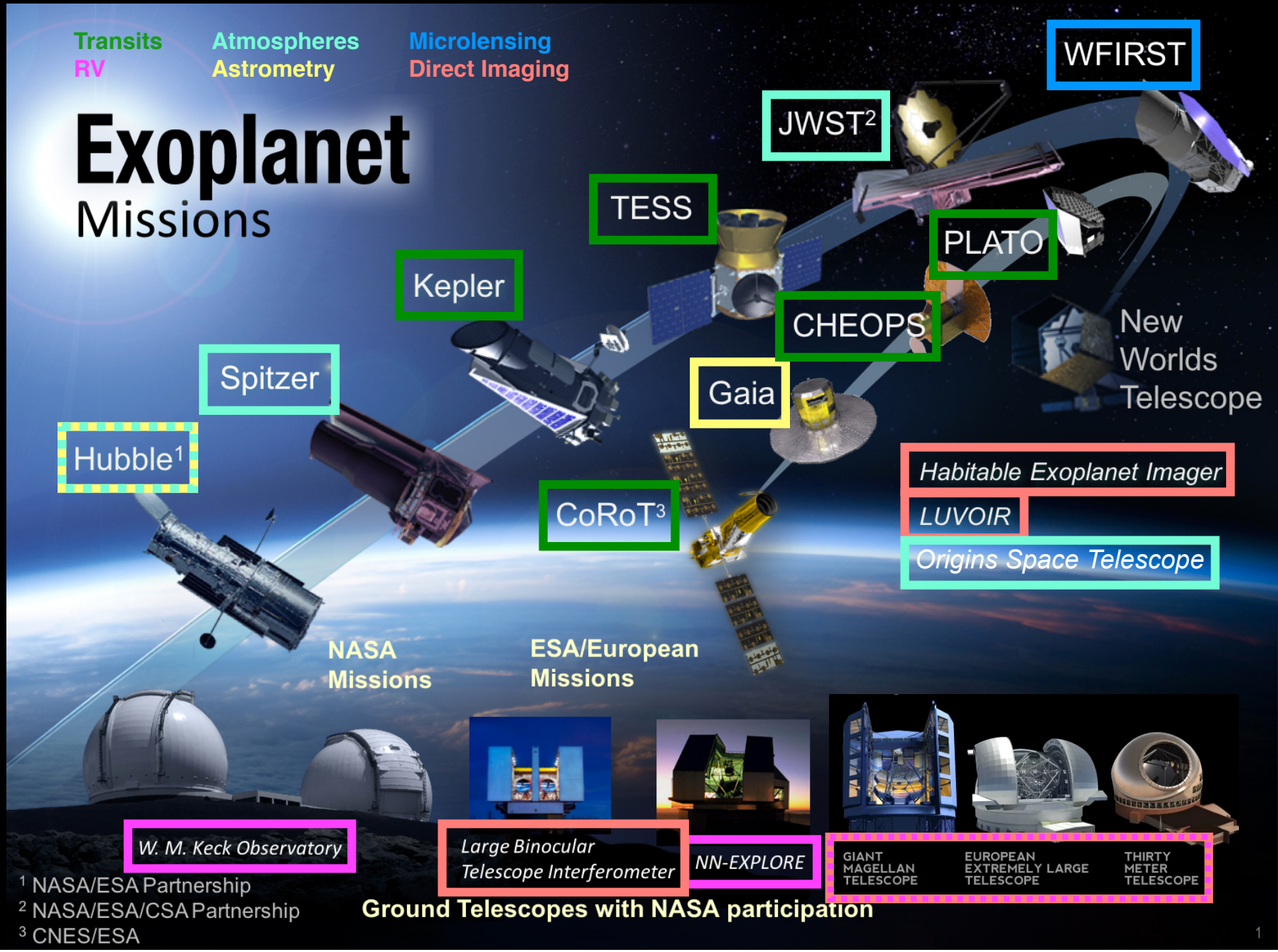
Synthesizing results from many surveys



Clanton & Gaudi, 2016

First look based on published occurrence rate values (highest level product); future work will require incorporating disparate data sources in archives.

Toward the Future



[1] NASA/ESA Partnership
[2] NASA/ESA/CSA Partnership
[3] CNES/ESA

Future: Full Exoplanet Census

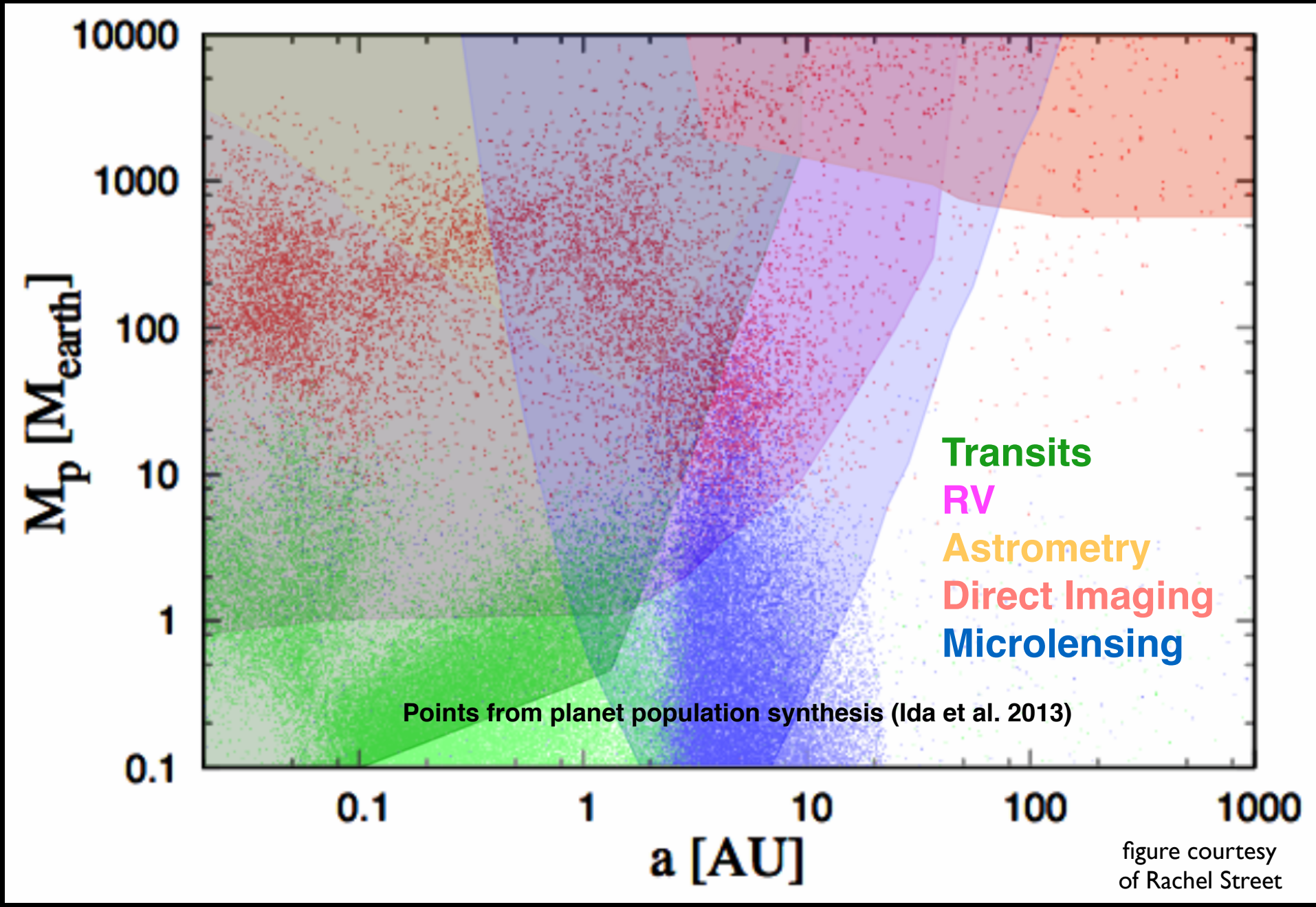


figure courtesy of Rachel Street

Future: Full Exoplanet Census

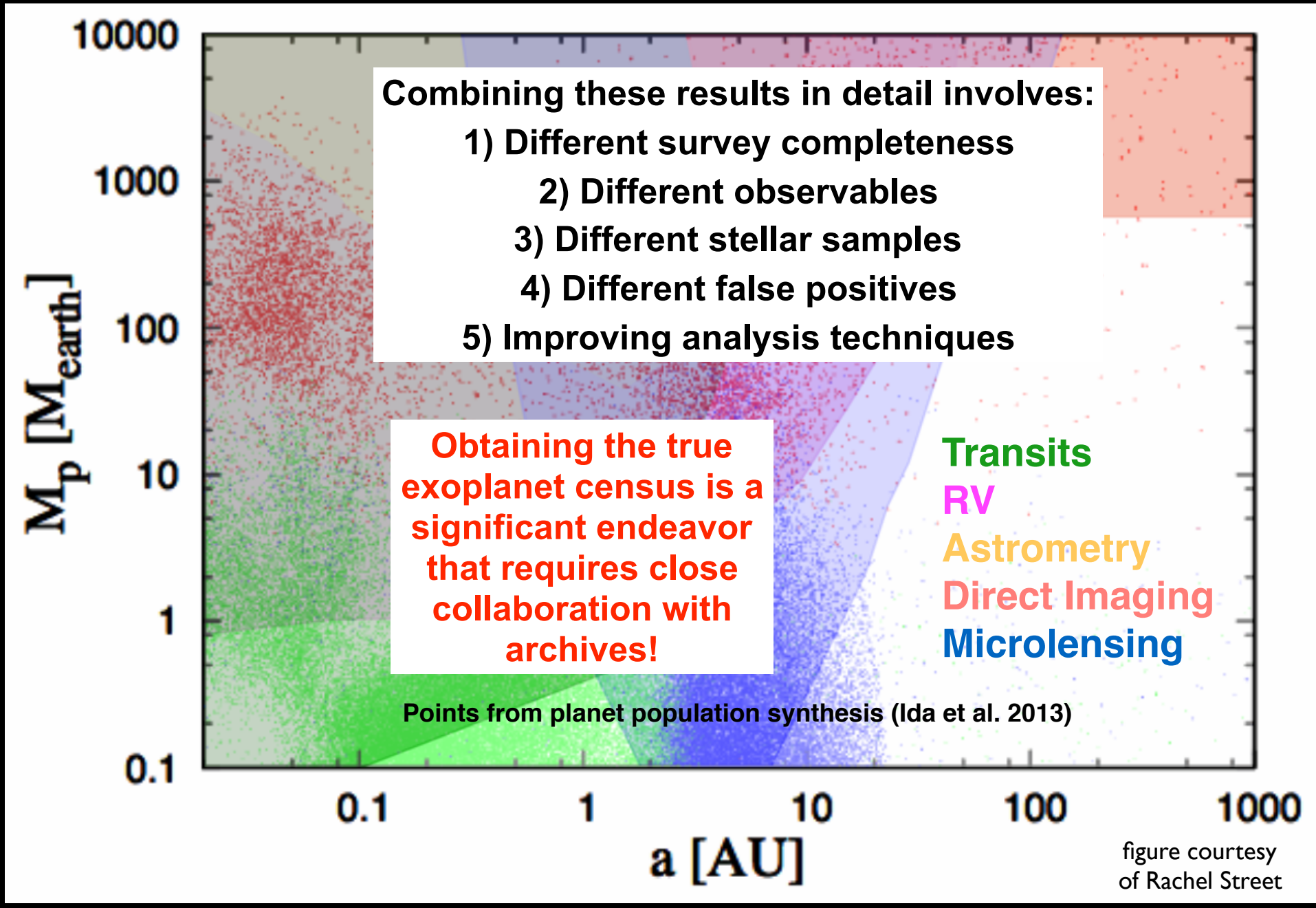


figure courtesy of Rachel Street