Introducing the Data Science Mission Office

Arfon Smith, arfon[at]stsci.edu, Josh Peek, peek[at]stsci.edu, Iva Momcheva, momcheva[at]stsci.edu, and Mike Fox, mfox[at]stsci.edu

Abstract

The Data Science Mission Office is approaching its first birthday and so we wanted to take this opportunity to introduce ourselves, our backgrounds and interests, and some of the initial focus areas for the Institute's newest Mission Office.

Introducing the team

Like all Mission Offices at the Institute, we're a relatively small group. In our case, we have a mixed team of astronomers and engineers:

Arfon Smith, Mission Head

Generally considers himself a 'lapsed' academic, since completing his PhD in Astrochemistry in 2006 (Nottingham, UK), he's spent his career building research technology and infrastructure and promoting the role of software in academia. Notable highlights include co-founding the Zooniverse citizen science platform (zooniverse.org) at the University of Oxford and most recently working at GitHub (github.com) to help bring open source software to the academy.

Josh Peek, Project Scientist

Studies the the dusty gas and plasma that suffuses and surrounds galaxies. Josh has a background in big survey astronomy, and leads the Galactic Arecibo L-band Feed Array Survey in Neutral Hydrogen (GALFA-HI). He is interested in advanced image processing and analysis methods in astronomy, and is studying how convolutional neural neworks can advance our understanding of complex images. Josh has worked in MAST since starting at the Institute in 2014, and has led efforts and better integration between MAST and the astronomical literature.

Iva Momcheva, Mission Scientist

Ivelina (Iva) Momcheva received her PhD from University of Arizona in 2009. Following a postdoc at Carnegie Observatories, she joined the 3D-HST Treasury Team as a project manager in 2011. This being her first encounter with *Hubble* observations, she was quickly enamored. Iva arrived at the Institute in 2015 as a support scientist in the WFC3 instrument group where she developed the Drift and Shift (DASH) observing technique, before moving to the Data Science Mission Office in early 2017 as a Mission Scientist. Iva's research focuses on galaxy evolution. She is interested in how environment affects galaxy evolution, particularly in groups and high-redshift proto-clusters. She is passionate about gravitational lenses, spectra, databases and well-documented application-programming interfaces (API)s.

Mike Fox, Senior Infrastructure Engineer

Mike Fox is our Infrastructure Engineer; he is focused on developing next-generation science and data processing systems at the Institute. Mike studied Electrical Engineering at Georgia Tech. He started his career writing device drivers and low-level APIs for custom reconfigurable computing systems. Moving up the software stack, he helped to develop a high-performance email server

which currently delivers more than 25% of the worlds legitimate email. Later he helped to turn this email server into a modern Cloud email service, as Director of the Site Reliability Engineering Team. Outside of work Mike enjoys sailing, kayaking, remote control helicopters, and biking.

Why create a Data Science Mission Office?

Since its beginnings, the Institute has been responsible for turning raw, binary data downloaded from spacecraft into science-ready data. Since 1997, we have served those data through the Mikulski Archive for Space Telescopes (MAST). A suite of data management tasks is involved in processing science data including: developing calibration pipelines that encode complex algorithms; infrastructure for executing these pipelines; high-volume, high-performance storage for archiving and serving data to the astronomical community; and developing tools, services and technologies for interacting with the data.

As the volume and variety of datasets grow (the Pan-STARRS archive is ~2PB in volume), and in an era of large-scale survey missions such as LSST and *WFIRST*, it is clear that the tools, services, and technologies required to support the astronomical community are very different from those needed by a single guest observer using *Hubble.*

We already know from experience that building high-quality services enables vast quantities of 'archival science' (see Figure 1) and it's clear already that these increasingly large datasets will be combined with more sophisticated, algorithmic analyses by the astronomical community. This combination of data volume and compute-intensive analyses means that the infrastructure required to support community science, as well as the tools to visualize and extract their information content, must become a 'mission-level' focus for the Institute.

The Data Science Mission Office aims to bring together the objectives of all the missions of the Institute in a single coherent vision for data processing, archive services, and community software. Through this vision, we aim to fulfuill our mission of maximizing the scientific impact of our missions and archive. Our focus is not just on technologies, but also on engaging and training the astronomy community in software, and working with the broader astronomy community in bringing tools to you that help you do your science.

To get an idea of the new types of scientific investigations that are possible when you bring data, methods, and technology together take a look at the Insitute's report from 2016 "Big Data @ STScI" (https://archive.stsci.edu/reports/BigDataSDTReport_Final.pdf).
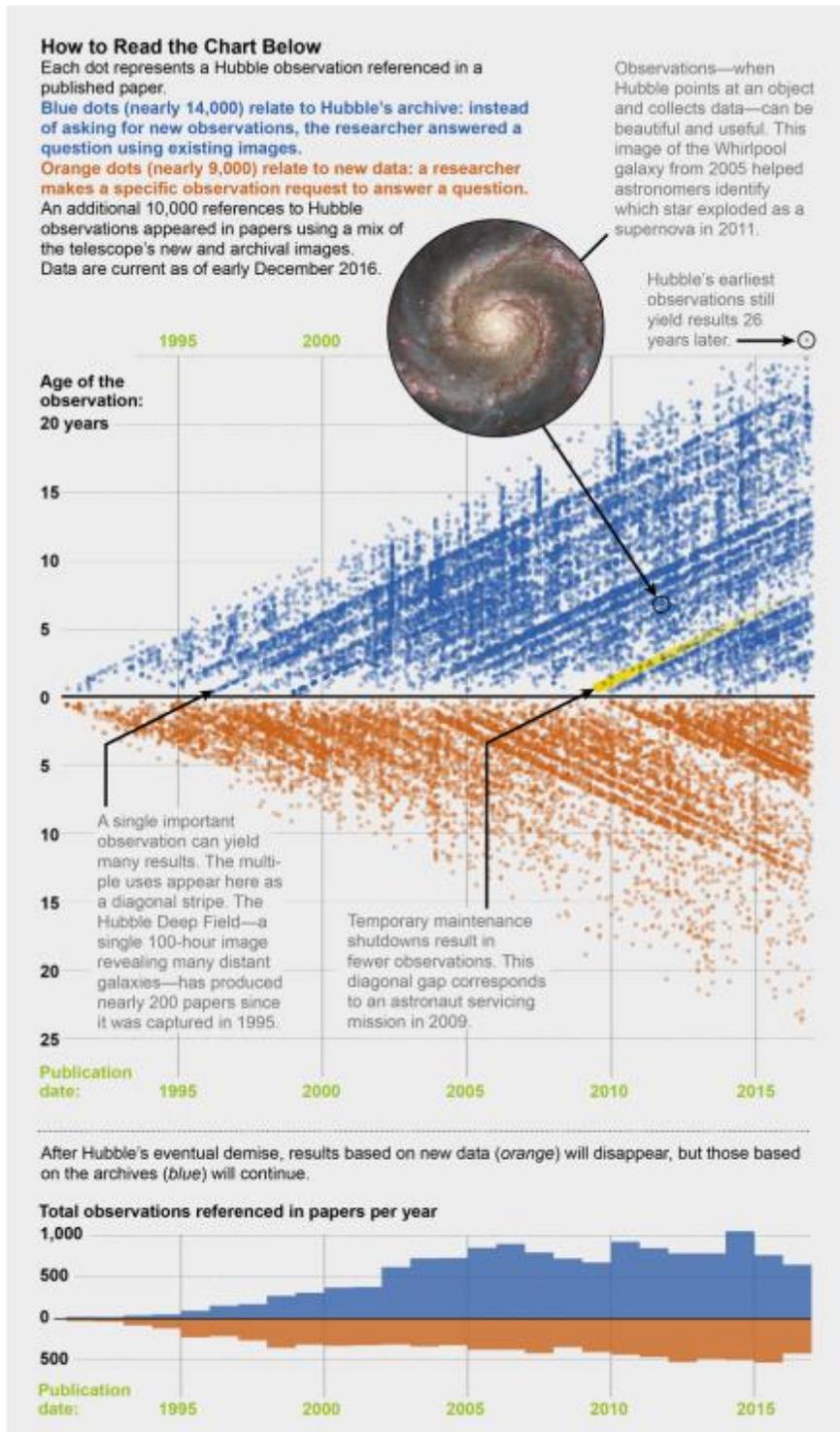
How to Read the Chart Below
Each dot represents a Hubble observation referenced in a published paper.
Blue dots (nearly 14,000) relate to Hubble's archive: instead of asking for new observations, the researcher answered a question using existing images.
Orange dots (nearly 9,000) relate to new data: a researcher makes a specific observation request to answer a question.
An additional 10,000 references to Hubble observations appeared in papers using a mix of the telescope's new and archival images.
Data are current as of early December 2016.

Observations—when Hubble points at an object and collects data—can be beautiful and useful. This image of the Whirlpool galaxy from 2005 helped astronomers identify which star exploded as a supernova in 2011.

Hubble's earliest observations still yield results 26 years later. ⟶

Age of the observation:
20 years

A single important observation can yield many results. The multiple uses appear here as a diagonal stripe. The Hubble Deep Field—a single 100-hour image revealing many distant galaxies—has produced nearly 200 papers since it was captured in 1995.

Temporary maintenance shutdowns result in fewer observations. This diagonal gap corresponds to an astronaut servicing mission in 2009.

Publication date: 1995   2000   2005   2010   2015

After Hubble's eventual demise, results based on new data (*orange*) will disappear, but those based on the archives (*blue*) will continue.

Total observations referenced in papers per year
1,000
500
0
500
Publication date: 1995   2000   2005   2010   2015

Figure 1: Publications resulting from *Hubble* observations over the last 26 years. Originally published at https://www.scientificamerican.com/article/how-old-observations-are-building-hubbles-legacy/. Credit: Katie Peek.