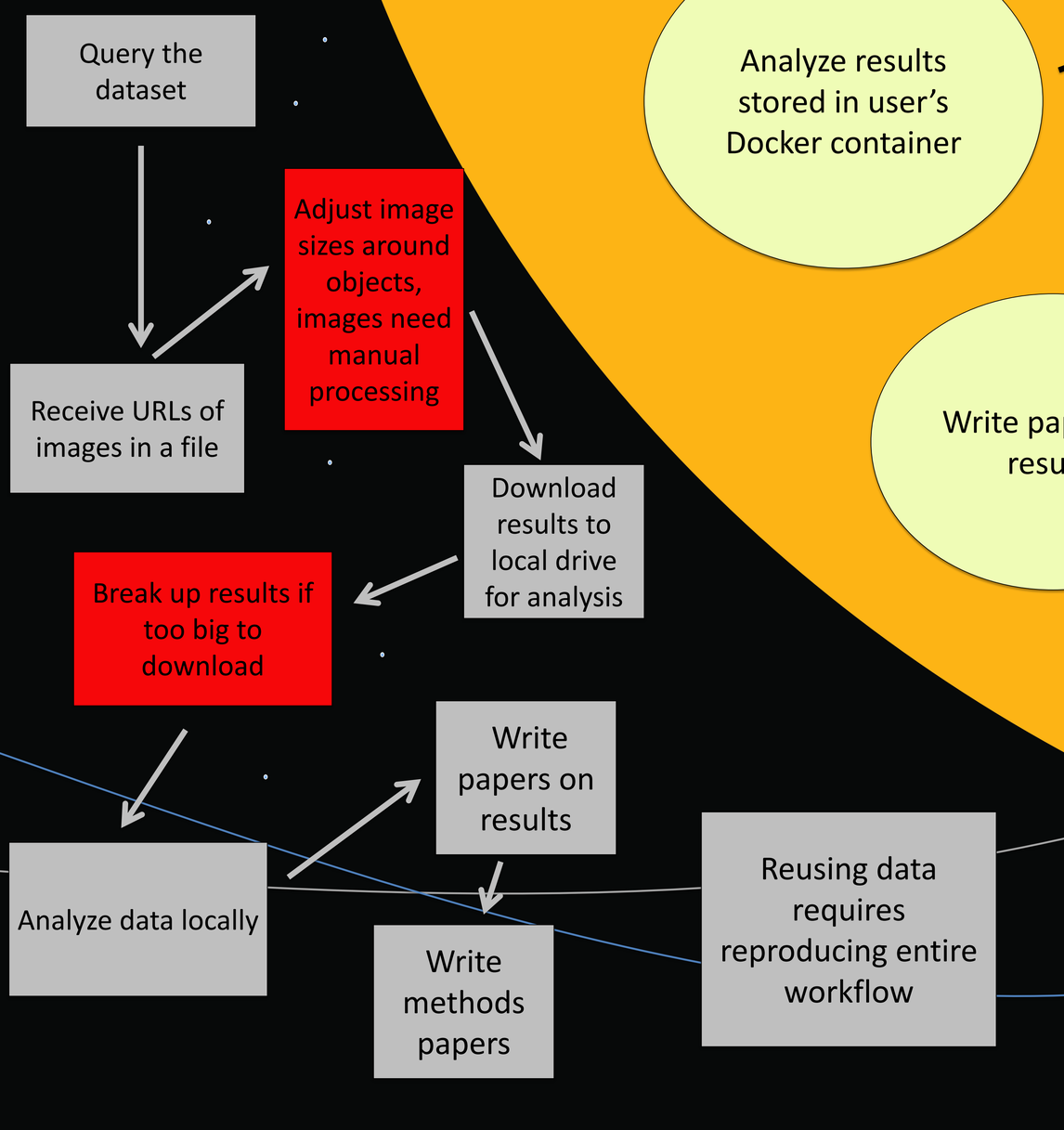
Too Big to Share?
Scaling Up Knowledge Transfer Workflows in Computational Sciences

Bernadette M. Randles, Ashley E. Sands, Christine L. Borgman
UCLA, Information Studies

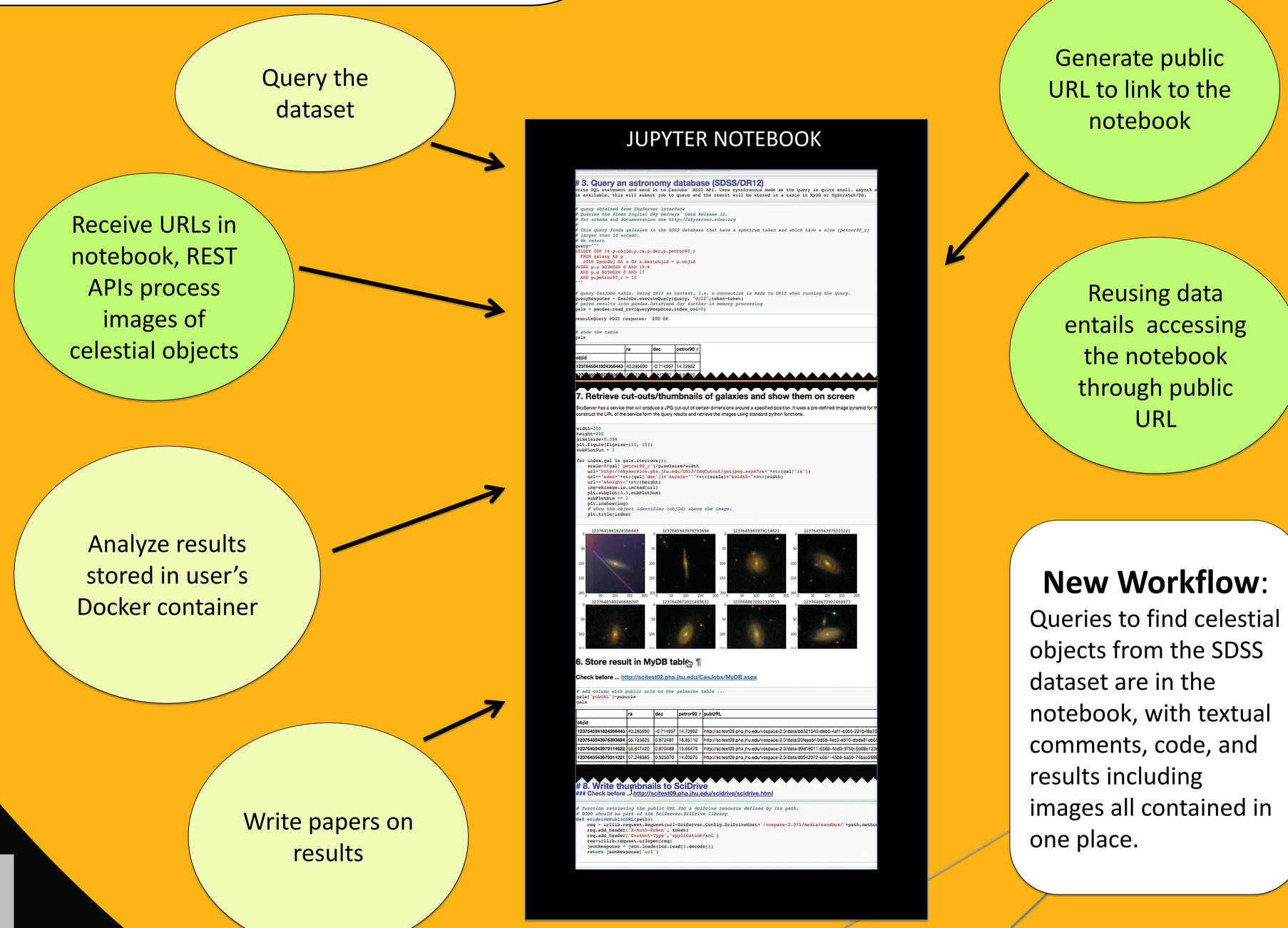
Problem:
Academic researchers experience computational and storage challenges when retrieving and analyzing up to petabyte-scale datasets. Researchers desire infrastructures and tools that support scalable retrieval, storage, and analysis of research data. Adding to these challenges, funders may require the resulting datasets to be cited on a more granular level, be reproducible and publically available, further complicating the traditional research process of documentation.

Case Study:
We compare a traditional astronomy research workflow of querying the Sloan Digital Sky Survey dataset for celestial object images to a new method using a Jupyter notebook. The Jupyter notebooks are the astronomer's main interface tool to remotely access datasets, providing a live computational research document[1] containing functionality to access, analyze, and display results of dataset queries or simulations. This particular (beta) instantiation of the Jupyter notebook platform resides on the server or server cluster where the data lives, keeping the analysis close to the data[2]. Another key element of the new workflow is the Docker container, a lightweight virtual space on the server enabling separate, storage spaces for users to perform larger computations and queries.

Traditional Workflow:
Astronomers query SDSS datasets through an online browser interface



What are Jupyter Notebooks?
Jupyter notebooks are (i)python-based, open source, browser-based documents that support rich text elements, graphs, images, and executable code; supporting over 50 computer languages such as R, Python, and C++. Jupyter notebook sociotechnical benefits include a dedicated user community. Drawbacks emerge in resistance to switch tools and work habits.



New Workflow:
Queries to find celestial objects from the SDSS dataset are in the notebook, with textual comments, code, and results including images all contained in one place.

Findings:
Research processes, code, and results are co-located, enabling others to retrace procedural steps and re-run the same code. This configuration shows potential to address problems of reproducibility and discoverability. Citing notebooks may solve some of the astronomy team's difficulties surrounding citation of code and datasets.

1 Millman, K. J., & Pérez, F. (2014). Developing Open-Source Scientific Practice. *Implementing Reproducible Research*, 149.
2 Raicu, I., Zhao, Y., Foster, I. T., & Szalay, A. (2008, June). Accelerating large-scale data exploration through data diffusion. In *Proceedings of the 2008 international workshop on Data-aware distributed computing* (pp. 9-18). ACM.

Thank you to the Alfred P. Sloan Foundation ("If Data Sharing is the Answer, What is the Question?" Award# 2015-14001. Christine L. Borgman, UCLA, Principal Investigator).
Thank you to the Johns Hopkins University IDIES, SciServer and SDSS teams.
Thank you to the members of the UCLA Knowledge Infrastructures Team, which include: Peter T. Darch, Irene Pasquetto, and Milena Golshan.
http://knowledgeinfrastructures.gseis.ucla.edu @UCLA_KI (and yes, Jupiter does have rings!)

UCLA Center for Knowledge Infrastructures

ALFRED P. SLOAN FOUNDATION 1934