NASEM
Opportunities for Accelerating Scientific Discovery: Realizing the Potential of Advanced and Automated Workflows

Status and Trajectory of Supporting Tools and Systems

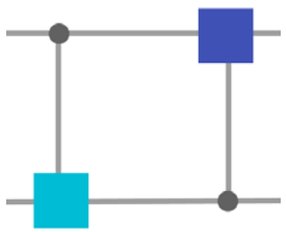
Brian Granger
AWS, AI Platform
Project Jupyter, Co-Founder
Cal Poly SLO, Professor of Physics (on leave)

Quick summary of my background

- Physics
    - 2001: Ph.D. in Theoretical and Computational Physics (CU Boulder)
    - 2001-now: Postdoc → Professor (Harvard, Santa Clara University, Cal Poly SLO).
- Software Engineering (2005-2015)
    - 2005: Started to work on open-source software (IPython) with classmate Fernando Perez.
    - 2011: Wrote the first version of the IPython Notebook over the summer.
    - 2014-current: Co-founder and co-director of Project Jupyter with Fernando Perez
- User Experience (UX) Design and Research (2015-current)
    - Over the last decade, the most challenging and interesting problems for Jupyter have transitioned from the technical (software engineering) to the human (UX design and research).
    - Why? 10s of millions of users, wide range of usage cases, diverse roles and personas, firehose of feedback.
    - My focus has shifted from writing software to UX design and research strategy, process, resourcing, etc.
- Organizational (2015-current)
    - 2015-current: Lots of time spent on the organizational aspects of open source (funding, governance, stakeholders, etc.)
    - 2019-current: On leave with Amazon Web Services, working on tools with an organizational focus (SageMaker)

Which workflows tools/systems have you worked on ?

Open-source community/tools, open-standards, architecture for humans working with code and data.



- `sympy.physics.quantum`: computer algebra system for quantum computing/mechanics

- PyZeroMQ: high performance, distributed message passing for Python

- IPython: interactive computing for Python

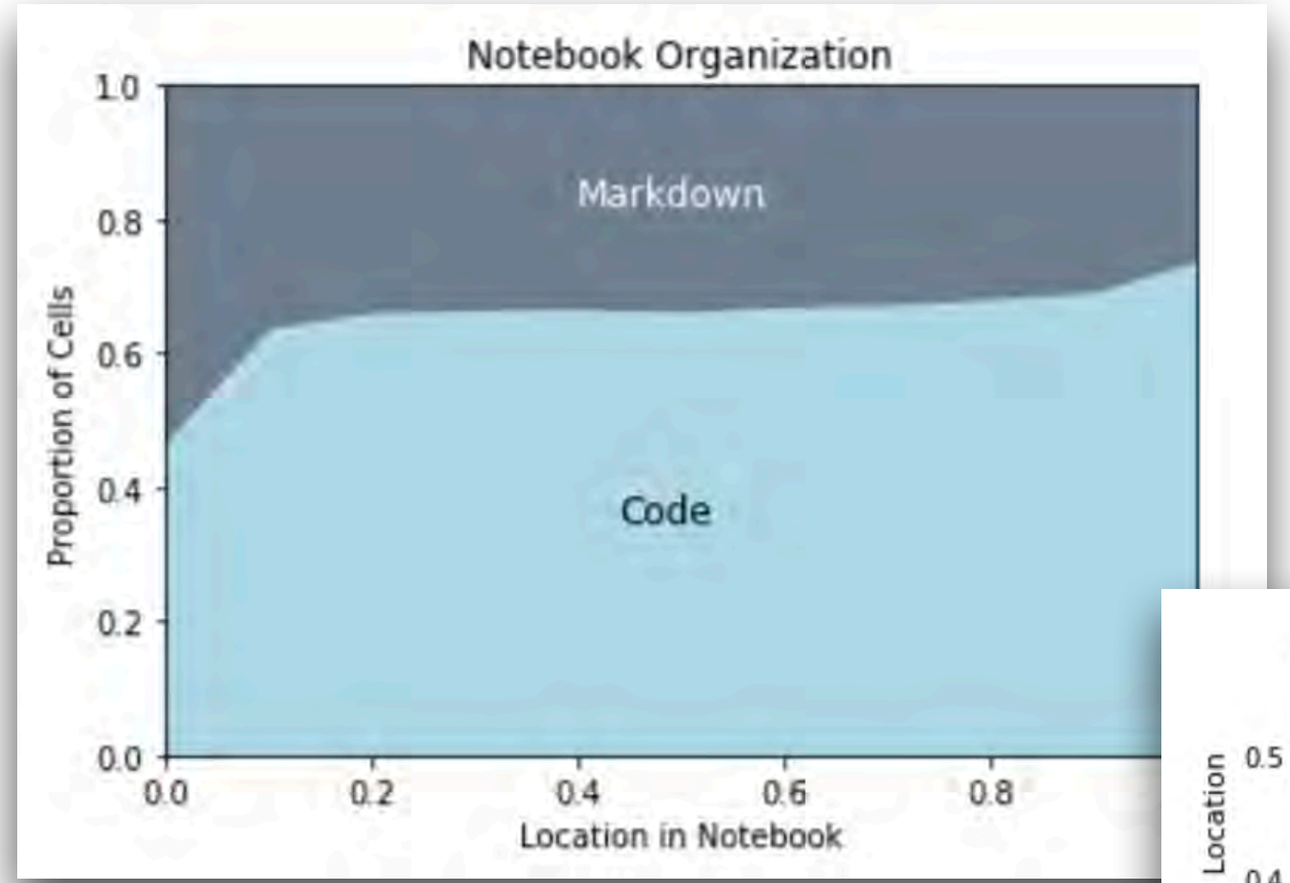
- Jupyter: interactive/reproducible computing

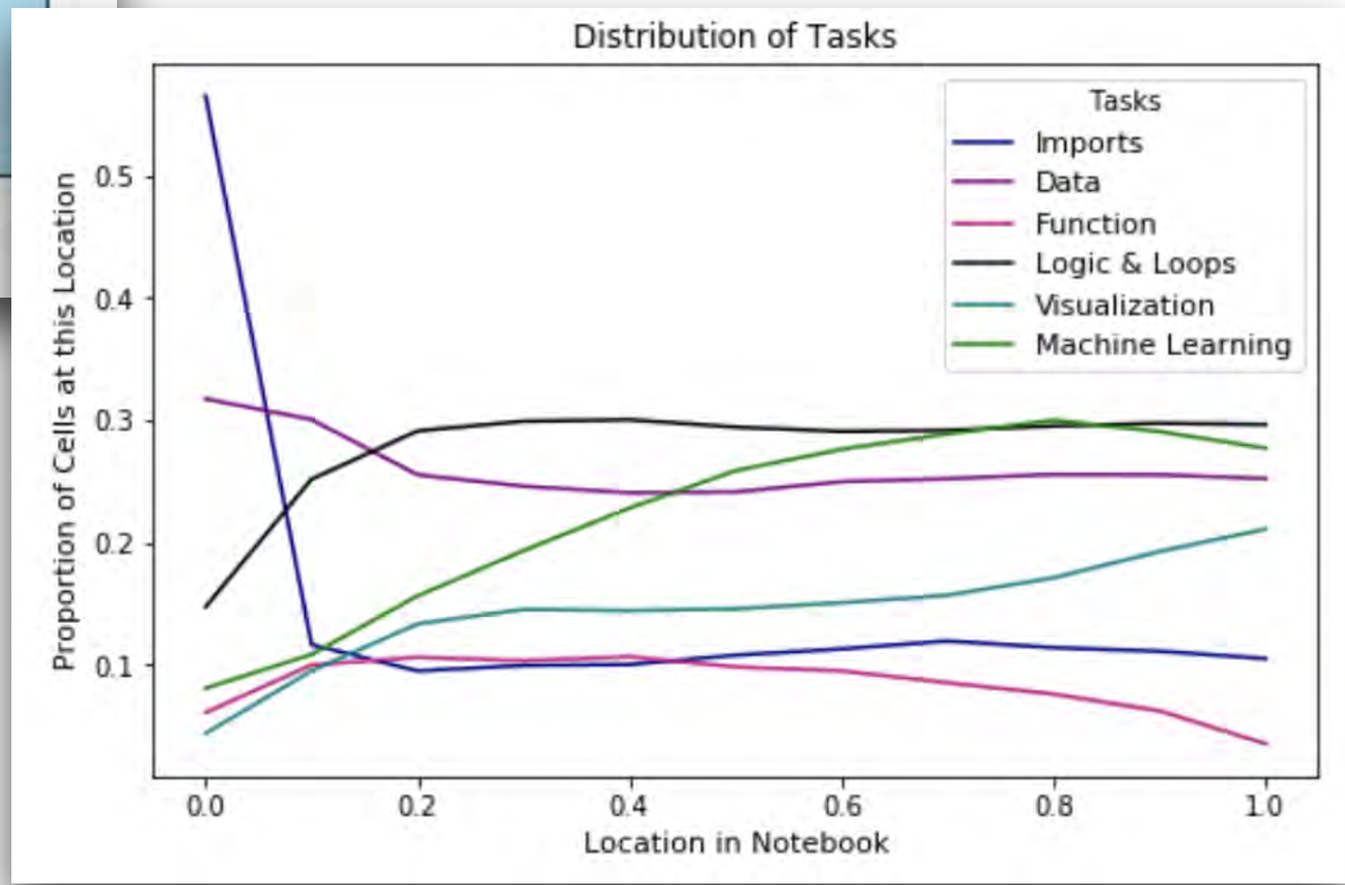
- Altair: statistical visualization for Python based on Vega/Vega-Lite

Is there a case study you can share that includes AI and data-driven scientific workflows?

- Every Jupyter Notebook is a case study along these lines.
- There are currently around 7 million public notebooks on GitHub:
  - https://github.com/parente/nbestimate
- Adam Rule (UCSD DesignLab): Let's look at 1 million Jupyter Notebooks to understand AI and data-driven scientific workflows:
  - Blog post describing the research and paper
- Jenna Landy (Cal Poly, AWS): Continued Adam's work to look at 4 million notebooks and explore deeper questions:
  - https://github.com/jupyter-resources/notebook-research

What can you learn about data-driven workflows by looking at 4 million notebooks





https://github.com/jupyter-resources/notebook-research

Questions

- technical < UX design < organizational
    - *How can we leverage the practices of UX design and research to build tools that accelerate AI and data-driven science?*
    - *How can we structure research funding, departments, universities, research collaborations, companies, etc. to produce tools that accelerate AI/data-driven science?*
- Rich Context
    - Collaboration with Julia Lane (NYU) that tackles some of the UX design and organizational aspects of building tools for AI/data-driven science (funded by Schmidt, Sloan)
        - Commenting and annotation on datasets, notebooks.
        - Metadata about data to understand "who has work with data, what methods and code were used, and what results were produced"
    - *How can we design tools that help diverse roles in organizations to work collaboratively with data in AI/data-driven science?*