



The MSD-LIVE Data and Computational Platform

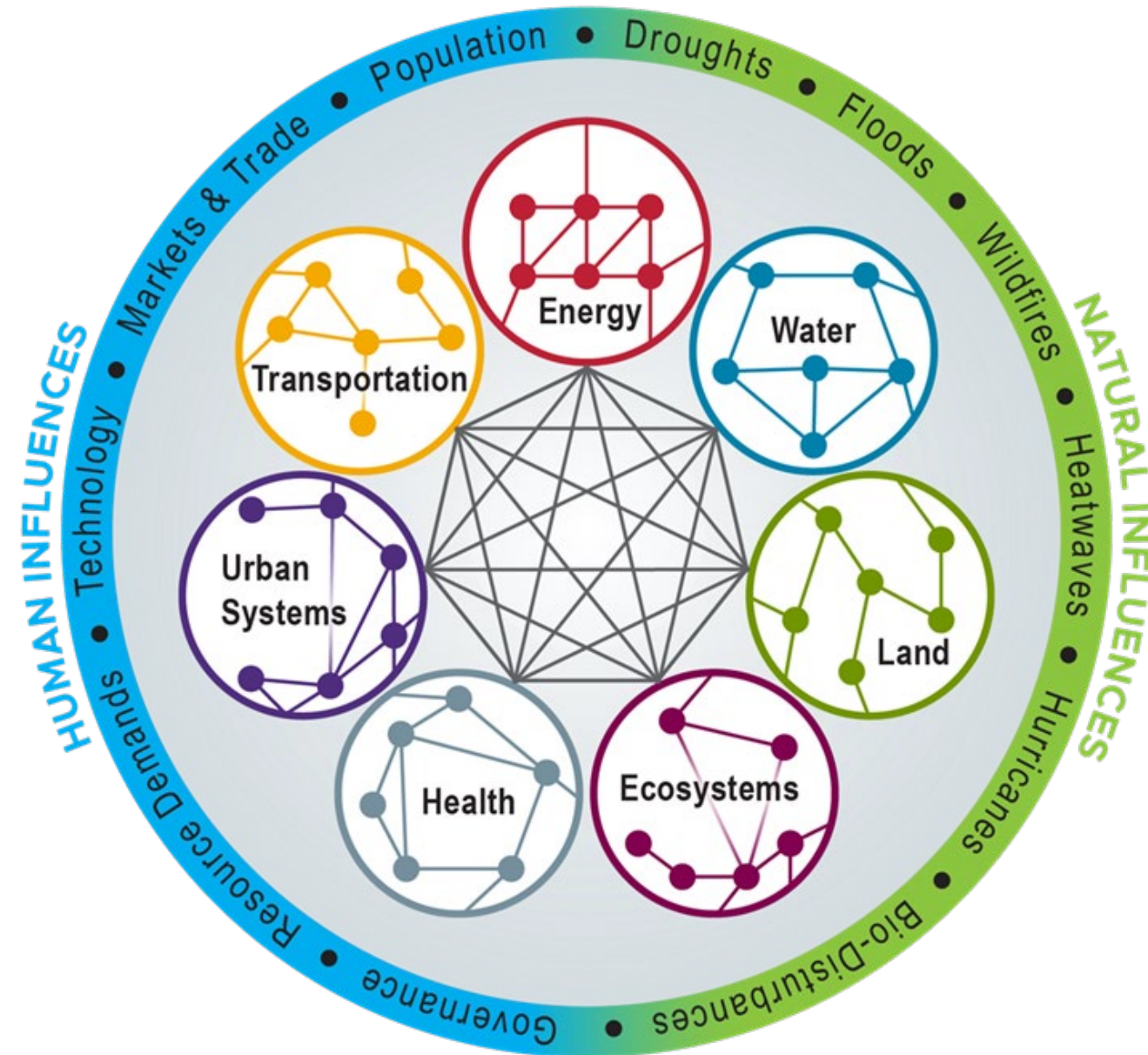
Devin McAllester, Casey Burleyson, Carina Lansing, Zoë Guillen, Matthew Macduff, and Jon Weers



PNNL is operated by Battelle for the U.S. Department of Energy

MultiSector Dynamics (MSD) Overview

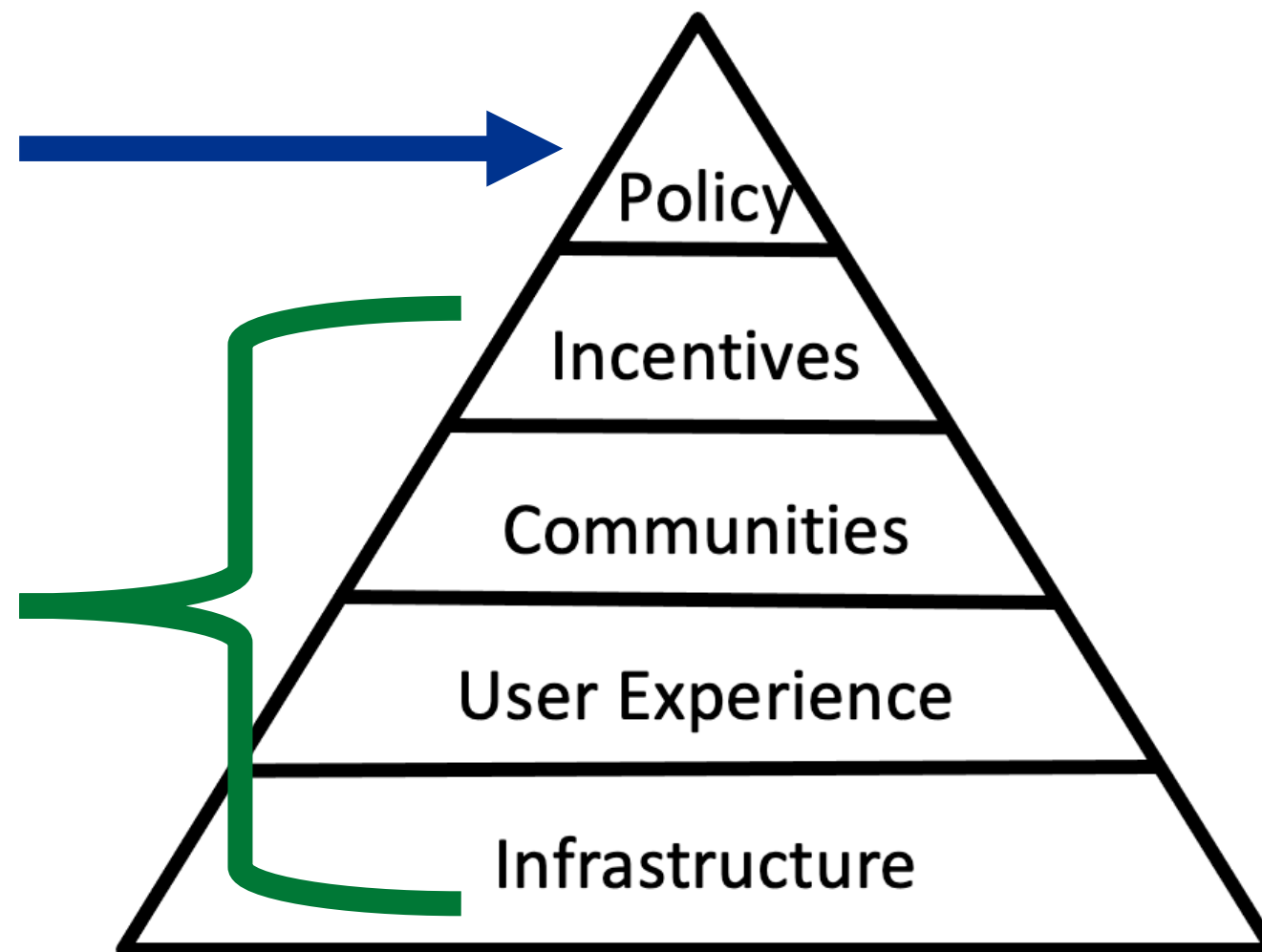
- Multisector dynamics refers to the complex interactions and interdependencies that exist among different sectors of human and natural systems. It involves understanding how changes and developments in one sector can influence and impact other sectors.
- The importance of studying multisector dynamics lies in the recognition that no sector operates in isolation. Economic, social, and environmental systems are interconnected, and changes in one sector can have cascading effects across multiple sectors



Facilitating Open Science with MSD-LIVE

Journals largely
skipped to this end of
the pyramid...

MSD-LIVE is
about tackling
these foundational
elements of the
pyramid...



Make it required

Make it rewarding

Make it normative

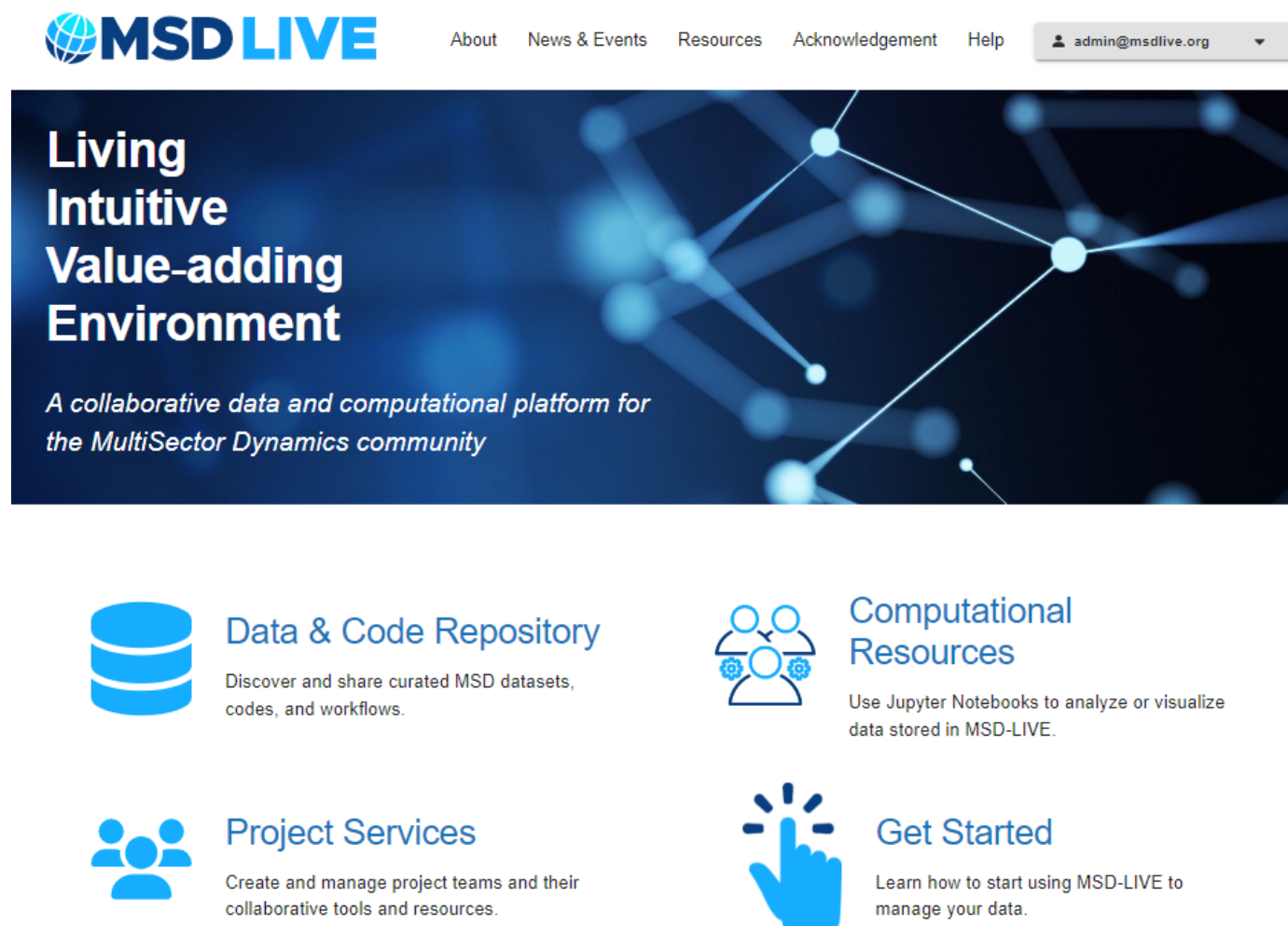
Make it easy

Make it possible

*Conceptual diagram from Brian Nosek of the University of
Virginia and the Center for Open Science*

The Vision for MSD-LIVE

- Cloud-based flexible and scalable data management system and advanced computing platform
- Custom-built to meet the needs of the MSD program area within DOE's Office of Science
- Intuitive user experience at scale:
 - Archive and share data
 - Share software and multi-model workflows
 - Run models and analysis tools



The image shows a screenshot of the MSD LIVE website. At the top is the MSD LIVE logo and a navigation bar with links: About, News & Events, Resources, Acknowledgement, Help, and a user dropdown menu showing 'admin@msdlive.org'. Below the navigation bar is a large banner with the text 'Living Intuitive Value-adding Environment' and a subtitle 'A collaborative data and computational platform for the MultiSector Dynamics community'. The banner background features a network diagram with blue nodes and lines. Below the banner are four service tiles, each with an icon and text:

- Data & Code Repository**: Discover and share curated MSD datasets, codes, and workflows. (Icon: database cylinder)
- Computational Resources**: Use Jupyter Notebooks to analyze or visualize data stored in MSD-LIVE. (Icon: people with gears)
- Project Services**: Create and manage project teams and their collaborative tools and resources. (Icon: three people)
- Get Started**: Learn how to start using MSD-LIVE to manage your data. (Icon: hand pointing up)

Why Not Use an Existing Solution?

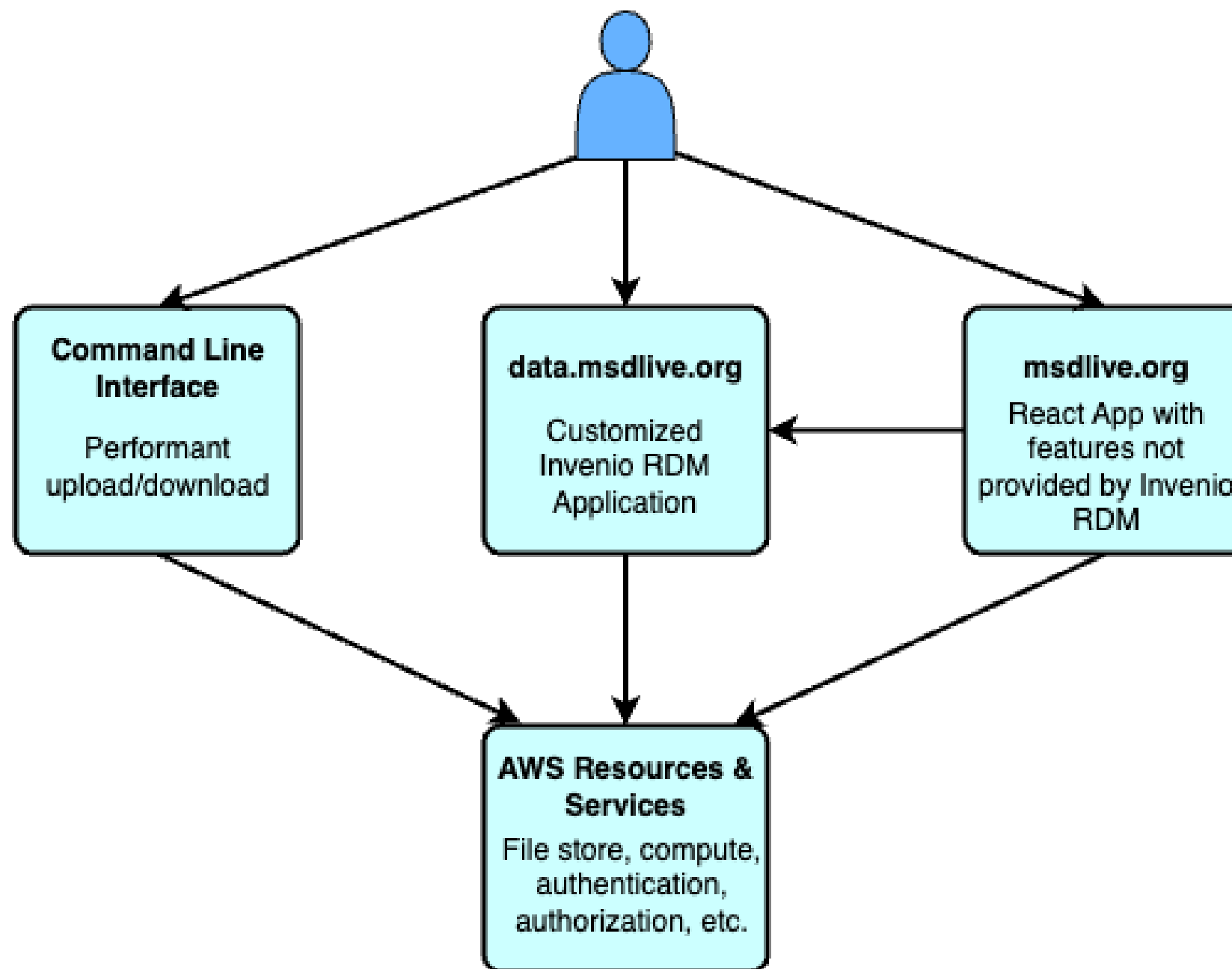
Common issues with existing scientific data management systems:

- Not tailored to the MSD users (e.g., generic metadata)
- Poor performance and reliability when uploading/downloading large datasets
- Lack direct access to data for compute platform
- Folder hierarchies not supported
- Bad user experience: new users struggle to setup and use the system

The MSD-LIVE Approach

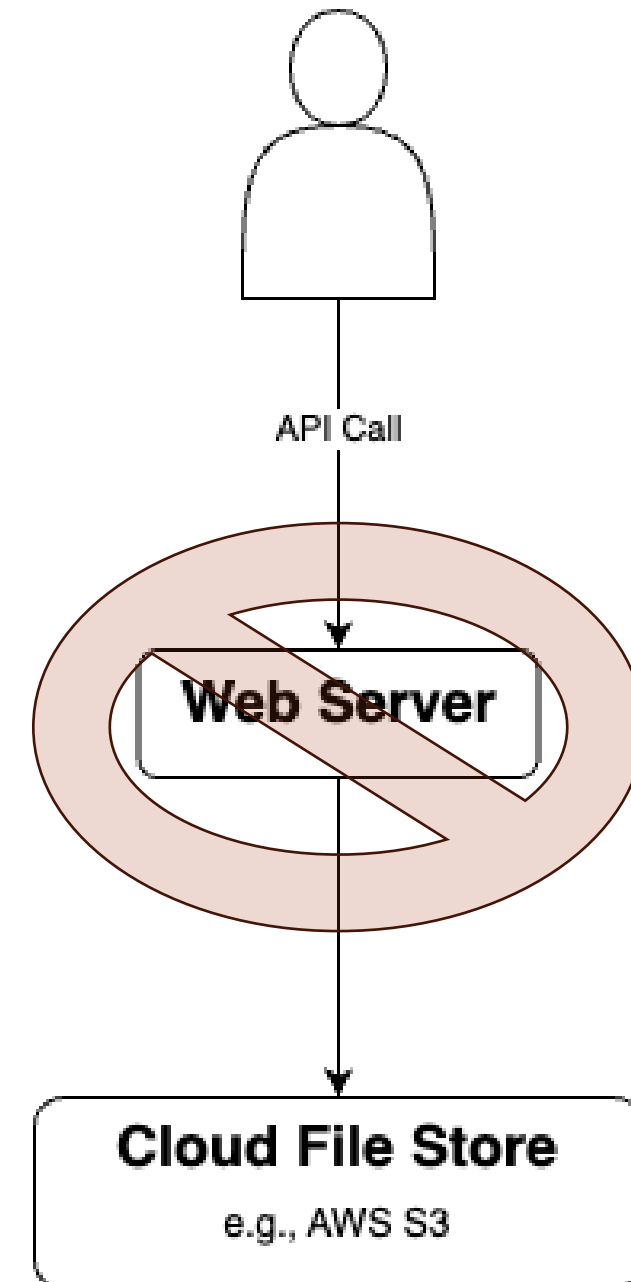
- Start with an open-source solution (Invenio RDM)
- Customize and extend it to achieve our goals
 - Customize metadata schema
 - Integrate with OSTI to mint DOIs
 - Add projects
 - Customize UI for seamless user experience
- Build everything on the cloud (AWS)
- Provide direct access to the data from Jupyter Notebook containers running in AWS
- Use AWS CloudFormation and AWS Cloud Development Kit to create version controlled and reusable infrastructure

High Level Architecture

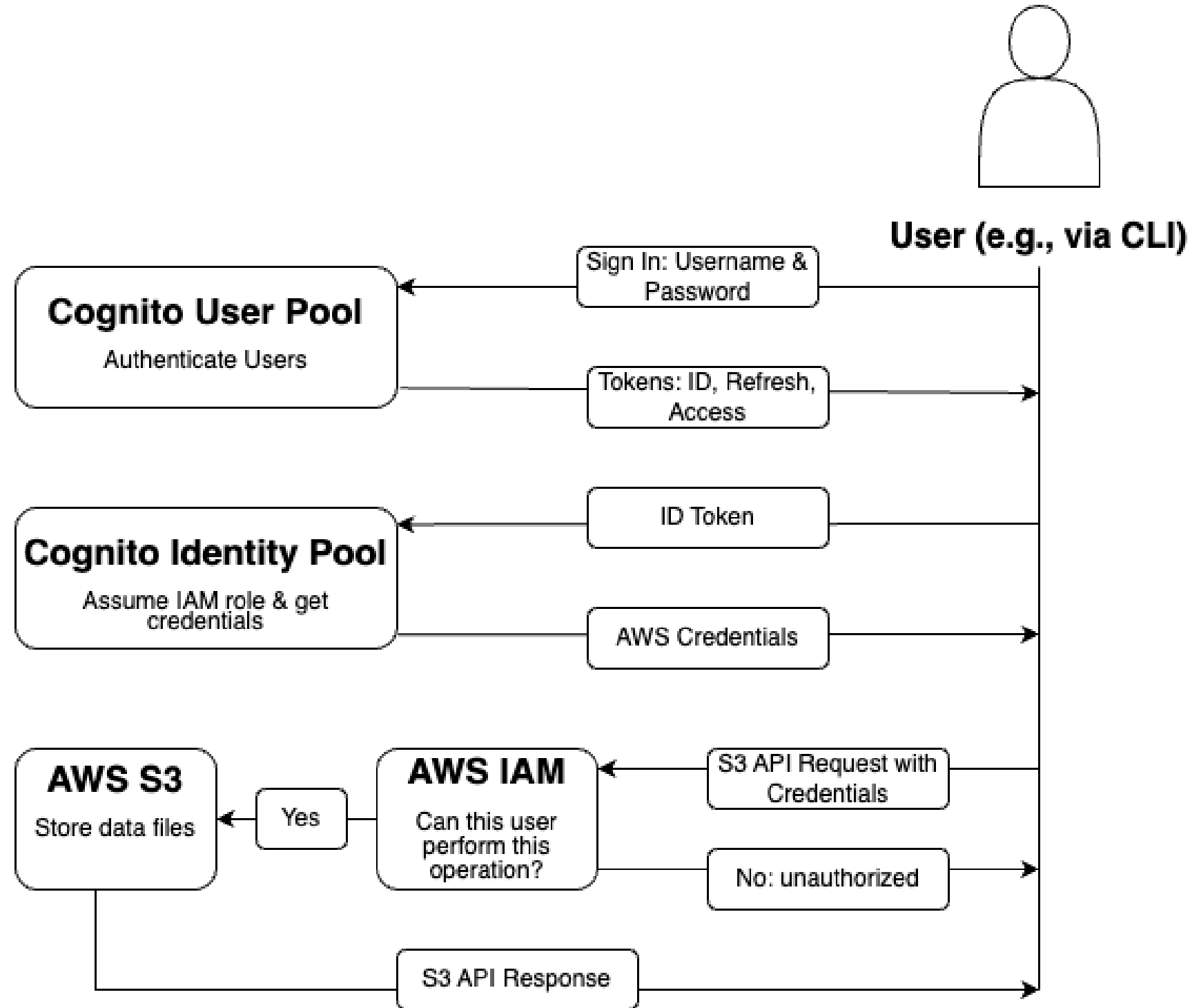


Direct Data Access: Why?

- Direct access: data and requests do NOT go through an intermediary web server
- Jupyter Notebooks can access data on the cloud in a secure way (i.e., with project-based access control)
- Allows us to use AWS open-source libraries on the client (e.g., the CLI and in the browser)
 - We don't reinvent the wheel
 - Code is well tested and performant



Direct Data Access: How it Works



Demo Video



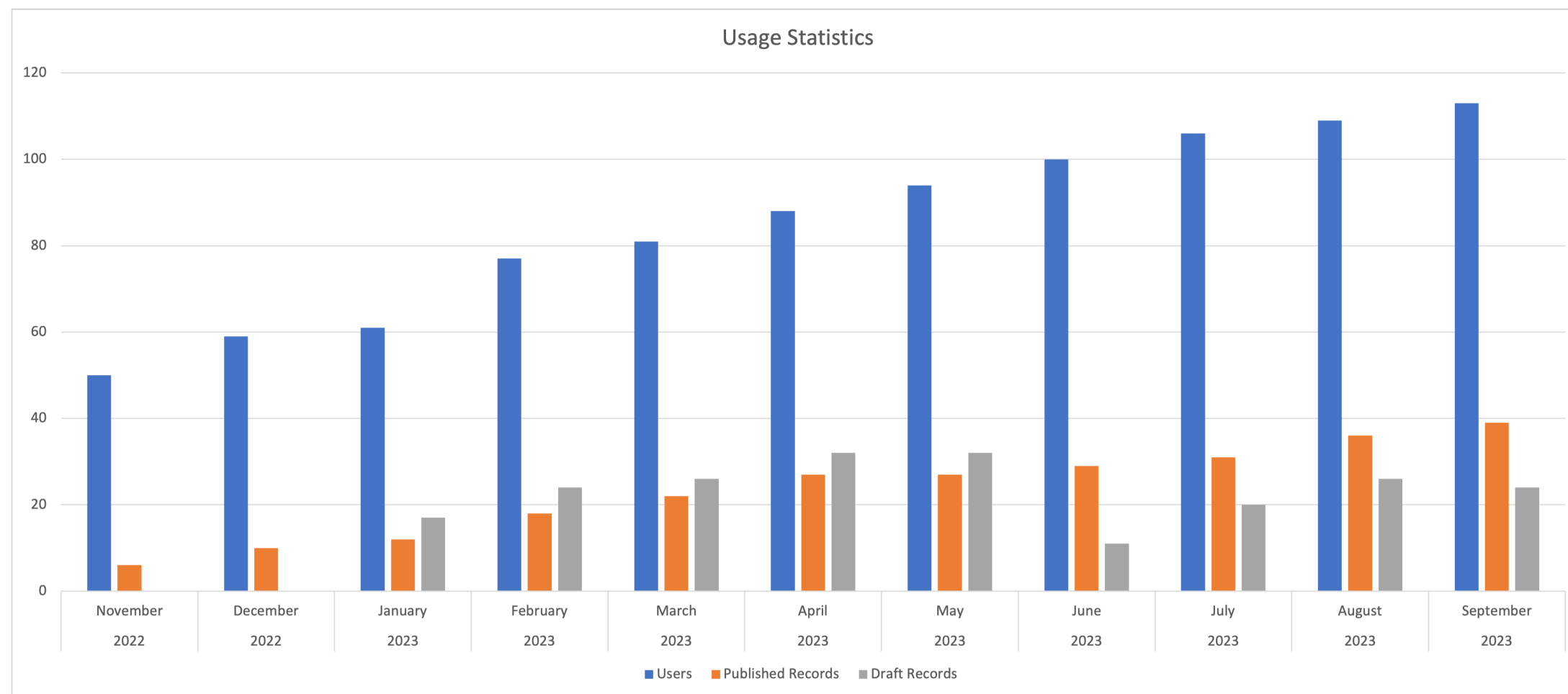
- Demo: [Using MSD-LIVE's CLI to upload files](#)

Current Notebooks

- Interactive model training notebooks:
 - IM3 - Uncertainty Characterization eBook: <https://uc-ebook.org/>
 - GCIMS - Xanthos: <https://xanthos.msdlive.org>
 - GCIMS - Hector: <https://hector.msdlive.org>
 - GCIMS - GCAM Wrapper: <https://gcamwrapper.msdlive.org>
 - GCIMS - Stitches: <https://stitches.msdlive.org>
 - IM3 - StateModify: <https://statemodify.msdlive.org>
- Data dashboard:
 - PCHES - <https://lafferty-sriver-2023-downscaling-uncertainty.msdlive.org/user-redirect/lab/tree/dashboard.ipynb>

Usage Statistics

- 9 projects
- 113 users from 15 states and 4 countries
- 39 published datasets and 24 open draft datasets
- 187+ Tb of data
- 518 uses of Jupyter notebook feature last month



Thank you!

- Questions?