Harnessing Responsible AI for Science: Taming Open Data

Manish Parashar
University of Utah Chief AI Officer
Executive Director, Scientific Computing & Imaging (SCI) Institute
Chair in Computational Science and Engineering
Presidential Professor, Kahlert School of Computing

Multicore World 2026
Christchurch, New Zealand, 02/16/2026

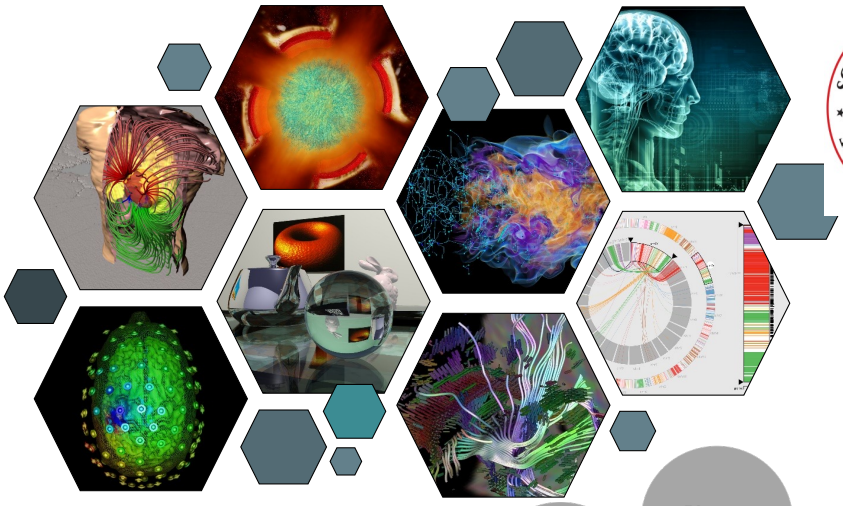
ONE-U RESPONSIBLE AI INITIATIVE *at the*
SCIENTIFIC COMPUTING & IMAGING INSTITUTE

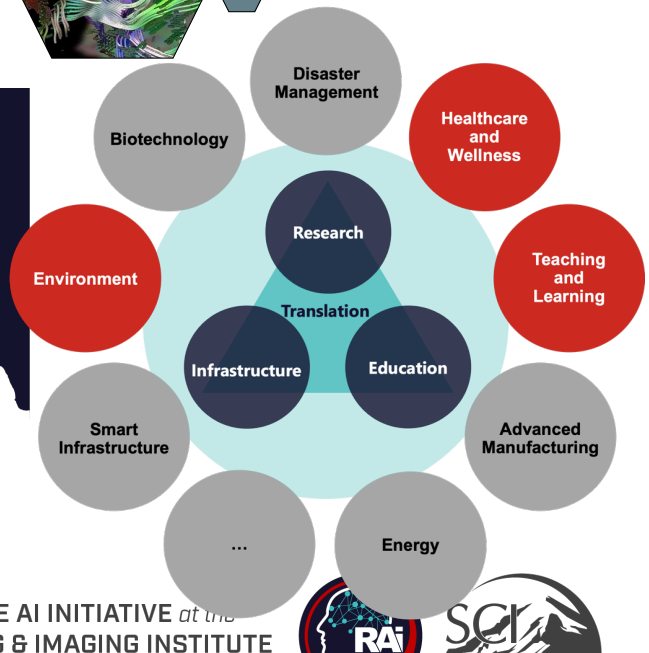
The SCI Institute

www.sci.utah.edu

Mission: Transform science and society through ***translational research and innovation***

- Inter/transdisciplinary, collaborative, convergent
- Core strengths in: Visualization & imaging; Scalable analytics; Advanced computing & data
- Software/system development and distribution are integral to our research processes



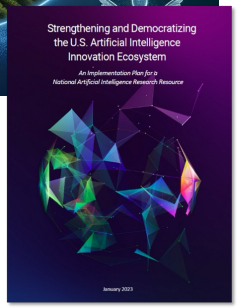


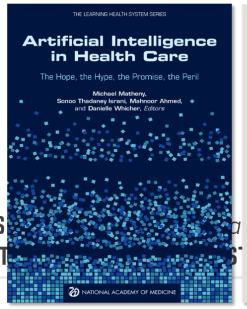
One-U Responsible AI Initiative: Responsibly advance ***translational AI*** for societal good

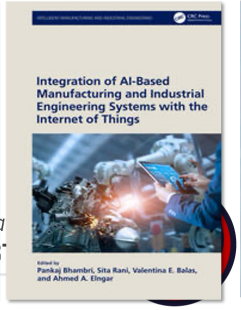
Transcendence of AI:

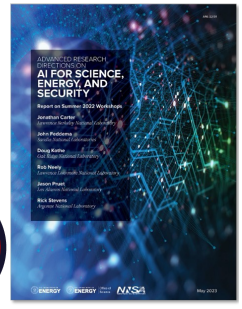
- AI has become an essential engine of innovation, scientific discovery, and economic growth.
  - Impacts span routine daily tasks to societal-level challenges.

Images generated with Microsoft Copilot.









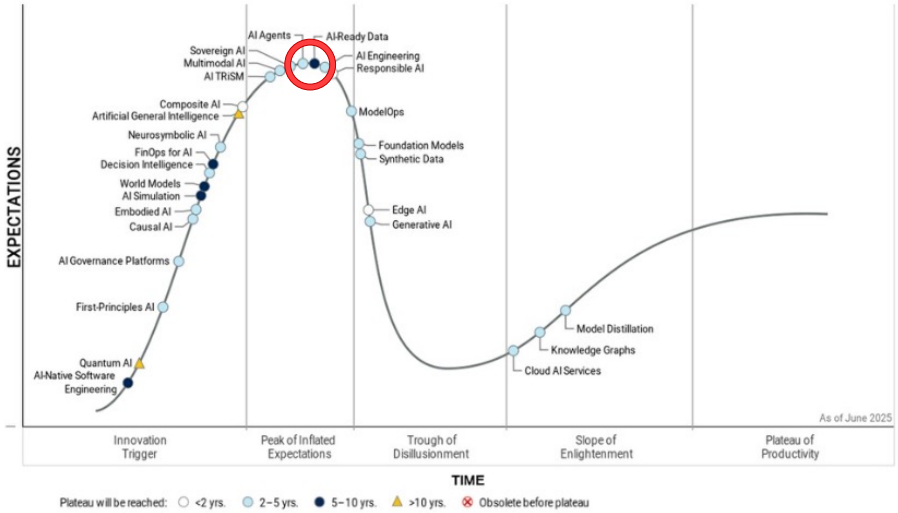
Transcendence of AI

- AI has become an essential engine of innovation, scientific discovery, and economic growth.
  - Impacts span routine daily tasks to societal-level challenges.
- AI has potential negative social, environmental, and economic consequences.
- It is imperative that we responsibly advance and use AI.

Images generated with Microsoft Copilot.

Ack: ChatGPT

Gartner AI Hype Cycle: A Focus on Data



https://www.bigdatawire.com/wp-content/uploads/2025/08/AI-hype-cycle.png

David Baker, shared the 2024 Nobel Prize in Chemistry, "for computational protein design."

David Baker on the importance of the Protein Data Bank for his AI-driven breakthroughs: ***"It's not just the methods, it's the data. And there aren't so many places where we have that kind of data."***

- Analysis of 41 M papers (2 M researchers) finds that AI expands individual impact and benefits career trajectories (Science 01/26, 10.1126/science.zc518iv)
  - Scientists adopting AI published 3.02 times as many papers and received 4.84 times as many citations
  - Junior scientists using AI less likely to drop out of academia and more likely to become established research leaders (1.5 years earlier than peers who hadn't)
- But using AI wasn't good for science overall – it narrows collective scientific exploration
  - AI papers covered 4.6% less territory than conventional scientific studies
- A feedback loop: *Popular problems motivate the creation of massive data sets, which make the use of AI tools appealing, and advances made using AI tools attract more scientists to the same problems*
- To realize the promise of AI for, we need to make data more *abundant, usable, useful and used!*

Why is data so challenging to use?

Infrastructure Challenges

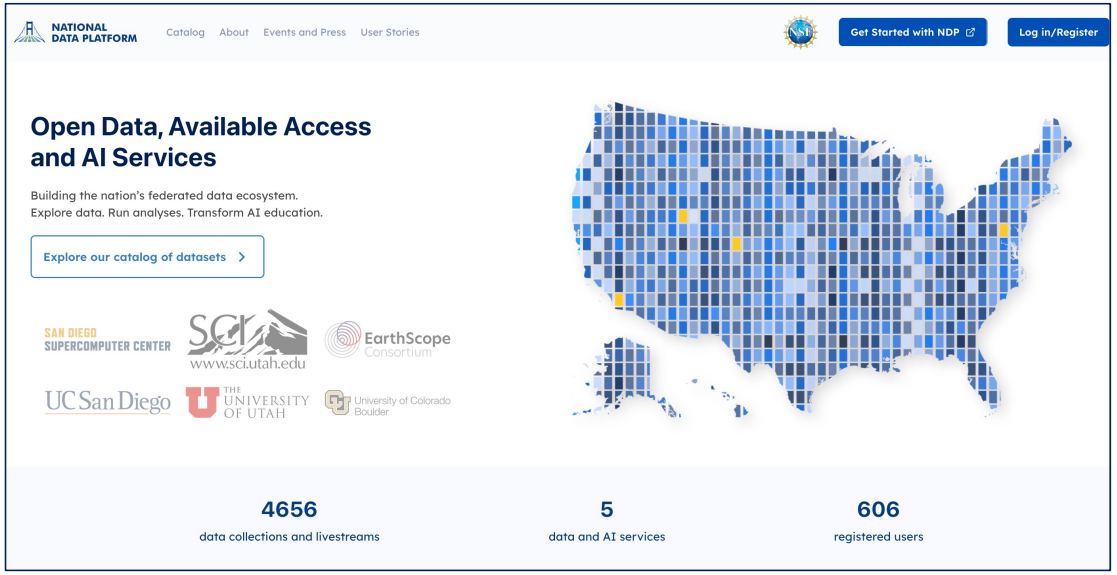
- **Siloed data ecosystems:** Diversity of data sources/repositories that are not interoperable or composable
- **Inconsistent data services:** Disparate, specialized data pipelines built for application domains
- **Difficult to deploy and use:** Data platforms and services require specialized skills to deploy and use
- **Access to computing:** Data source disconnected from computing; access to computing requires funding, training, and local infrastructure

User/Usage Challenges

- **Awareness:** Why is data important? What data is relevant to me? Where do I find the data I need? What do I need to know about it?
- **Ability:** How can I get the skills needed to use data? Where do I go for help?
- **Access:** Where do I get the resources to access, contribute to, and/or use data?
- **Fit for use:** How do I determine what data is appropriate for me to use? How do I build trust in data?
- **Association:** How do I find out who else is using the data and how? How do I become part of a community?

Move beyond Open => Useful, Usable, Used

National Data Platform (NDP)



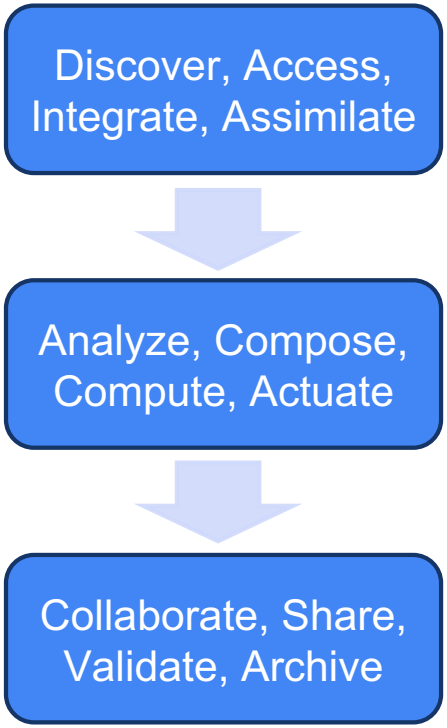
A **broad**, **federated** and **extensible** data ecosystem to promote collaboration, innovation, and customizable use of data on top of existing national infrastructure capabilities.

https://www.nationaldataplatform.org/

What can NDP help with?

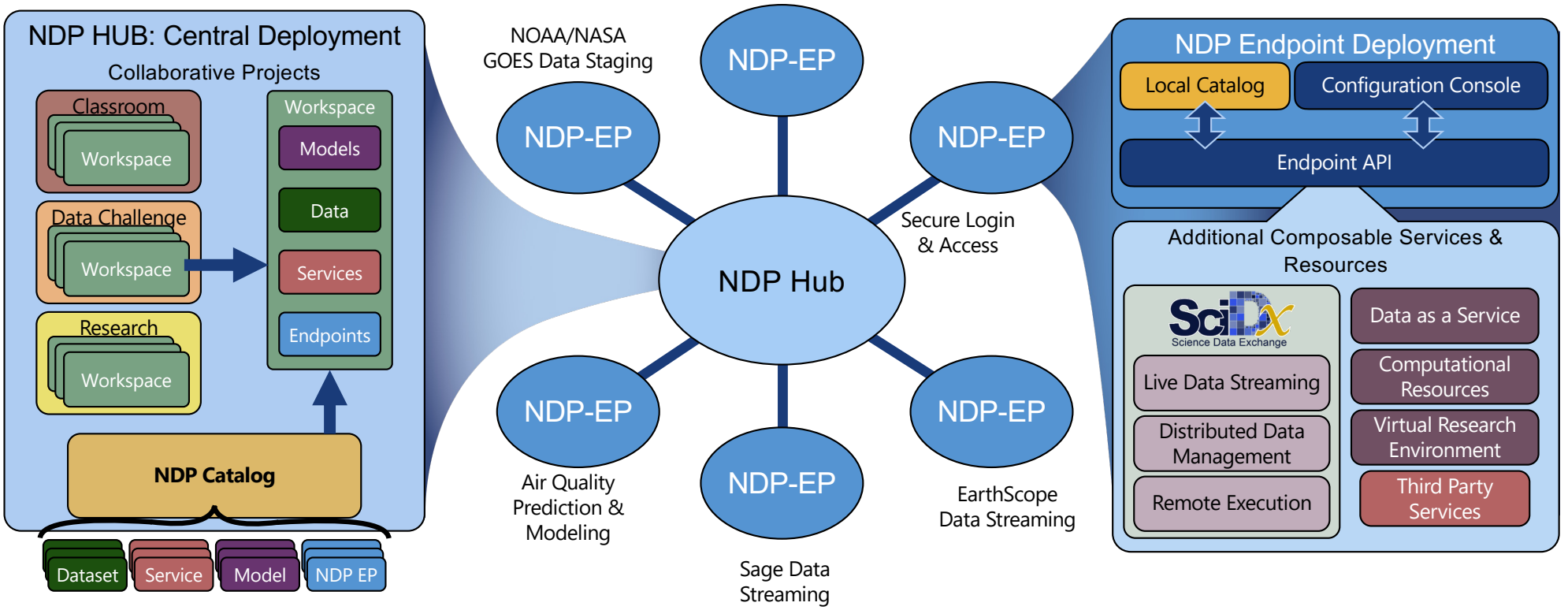
NATIONAL DATA PLATFORM

- Facilitates data cataloging to make data ***FAIR and used at scale***.
- Provides collaborative workspaces and customizable (near-data) services for ***data utilization***.
- Connects users to ***national cyberinfrastructure and Cloud resources***.
- Enables the development and deployment ***of AI-integrated workflows***.
- Offers tools to create classroom and ***data challenge learning experiences***.



NDP Architecture: A Federated Data Ecosystem

NATIONAL DATA PLATFORM



Science Data Exchanges (SciDx) Services

A customizable Data-Pop software stack for in-situ data access & processing

SciDx Staging Services

- Transient resources for in-situ (close to the data) data processing and access
  - High-performance in-memory processing
  - Server-side data transformations (e.g., sub-setting, reduction, user-defined analysis, etc.)
  - Caching/sharing of data, results, and data-products
  - Registration of data-triggers
- Efficient management of data in-motion
  - Streamline workflows; minimize data transfers
  - Perform ETL operations at data source

Science Data Exchanges (SciDx) Services

A customizable Data-Pop software stack for in-situ data access & processing

SciDx Staging Services

- Transient resources for in-situ (close to the data) data processing and access
  - High-performance in-memory processing
  - Server-side data transformations (e.g., sub-setting, reduction, user-defined analysis, etc.)
  - Caching/sharing of data, results, and data-products
  - Registration of data-triggers
- Efficient management of data in-motion
  - Streamline workflows; minimize data transfers
  - Perform ETL operations at data source

SciDx Streaming Service

- Streams registration, curation/archival for discovery and access
- User-defined operations/filters on streaming; containerized execution
- Combine streaming data with archived/playback data
- Mechanism for online data product generation (i.e., new data streams)

EarthScope Early Earthquake Detection Workflow

Contact: Charles Meertens, University of Colorado and Dave Mencin, EarthScope Consortium

Live Streamed & Archived Sensor Data

3D Global Navigation Satellite System (GNSS)

Ground displacement measurements

High temporal rate (1 Hz)

~ 1000 stations (geo-located)

Subscription

NDP Endpoint with SciDx Streaming

Historical data storage

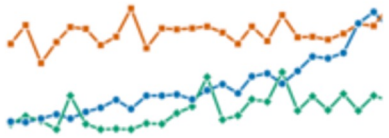
OSDF/Pelican origin

AI Analytics

Trigger HTC workflows

AI model retraining

Real-time visualization of (derived) streams; model-based anomaly detection



Register metadata, derived streams

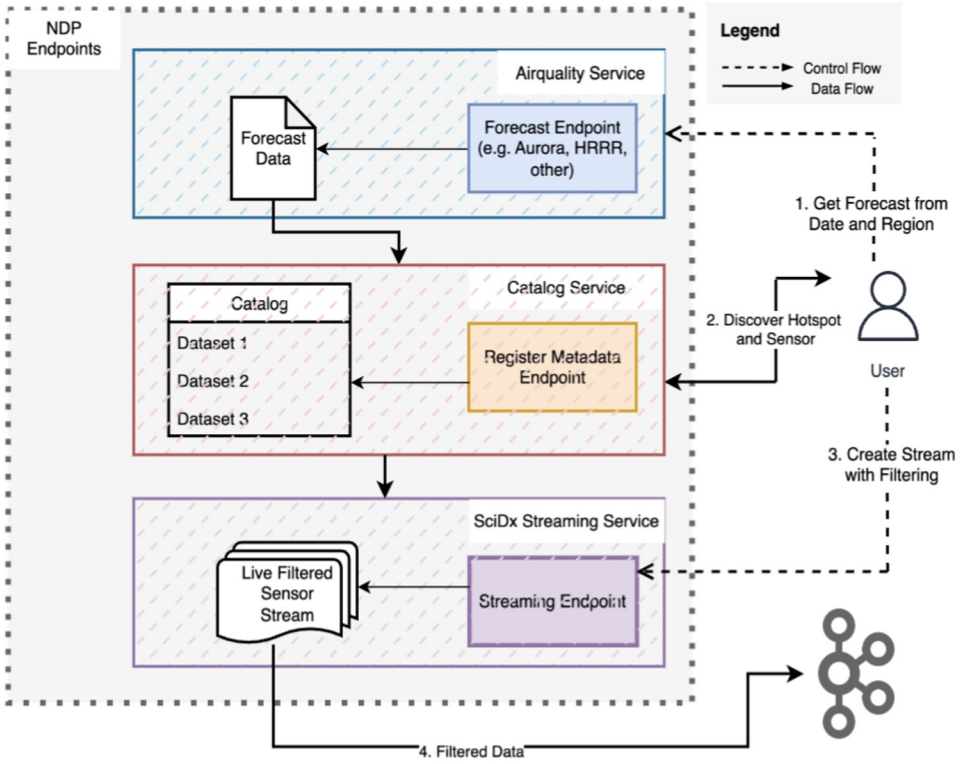
HUB

- Data search & discovery
- Workspace access

- **NDP Endpoint** with SciDX Streaming Service and JupyterHub deployed on AWS close to EarthScope
- Streams are visualized and shown through custom endpoint interfaces (Jupyter notebook or web)
- Derived streams are registered to **NDP Catalog** with metadata and made discoverable
- The streams are also uploaded to OSDF as historical data for use by **AI workflows** on OSG

Adaptive Air Quality Sensing

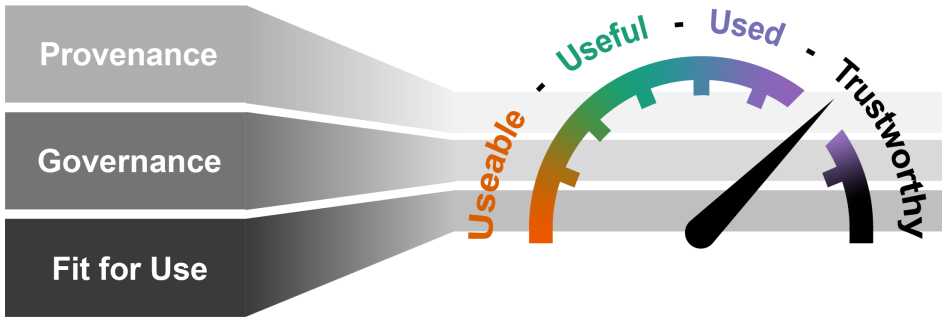
- **Forecast ingestion:** Run Aurora on forecast inputs and catalog the forecast output in NDP.
- **Hotspot detection:** Find PM2.5 hotspots and turn them into persistent trigger events.
- **Adaptive ingestion:** Use triggers and NDP metadata to select nearby sensors and create streams only from those sensors.
- **Dissemination:** Publish activated streams via SciDx Streaming to downstream consumers (alerts/analytics/storage).



In this talk …

- Towards trustworthy data – accessing fit for purpose
- Leveraging agents for data discovery

Towards Trustworthy Data: Dimensions of Trust



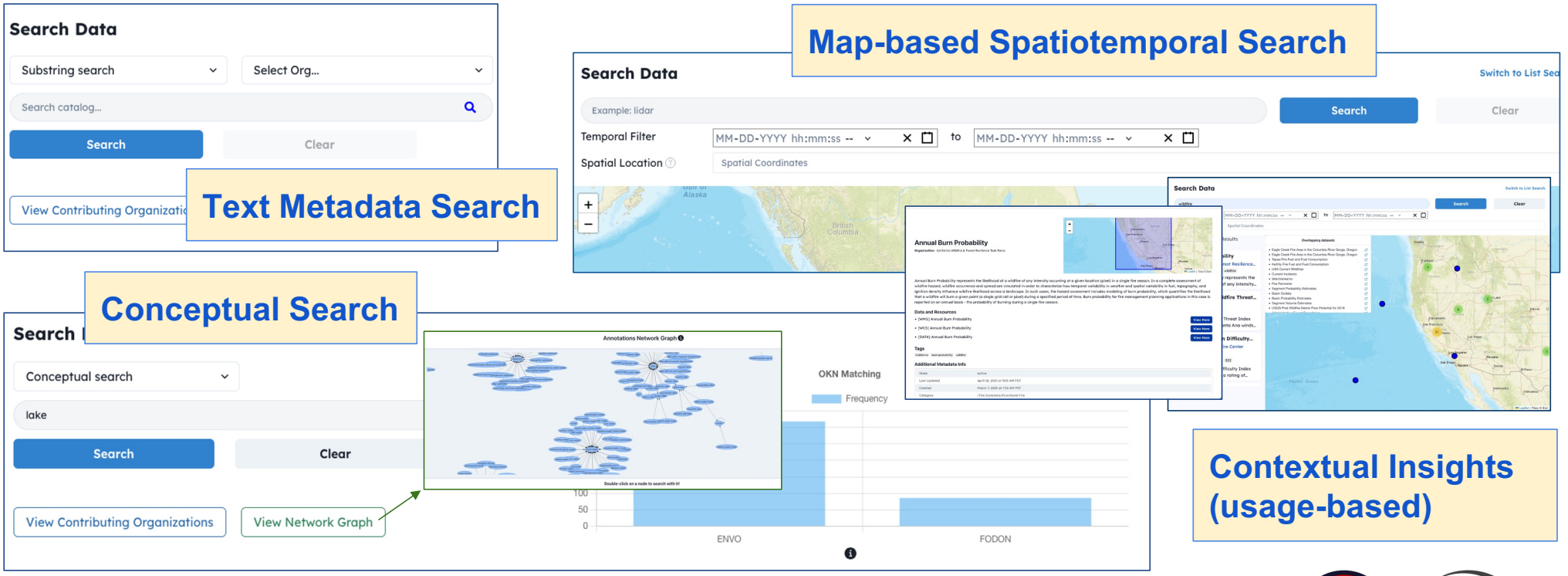
- **Usable:** Accessibility through governance, documentation, openness, and infrastructure.
- **Useful:** Context-dependent alignment with intended use cases; determination of *fit-for-use*.
- **Used:** Verification of usage and impact; user feedback.

Determining *Fit-for-Use*

- Who is using the data and for what research?
- What is the impact of research using the data?
- What data sets are used together, how are they used, and for what research?
- What tools/codes are used for the data?
- …

Data Discovery and Usage Exploration

Multimodal Search and Exploration Interfaces



NDP Contextual Insights

Data Registration

- Leverage LLMs to characterize the dataset using keywords.
- Approve dataset keywords.

Metadata Collection

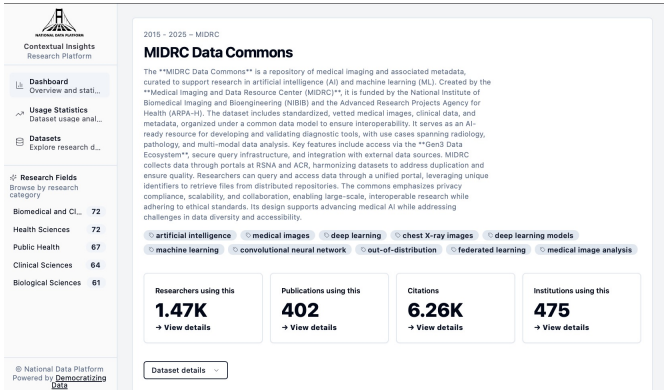
- Leverage publication corpora (e.g., OpenAlex, Dimensions) to generate context metadata knowledge graph about who, what, and how data have been used.
- Validate generated metadata for accuracy.

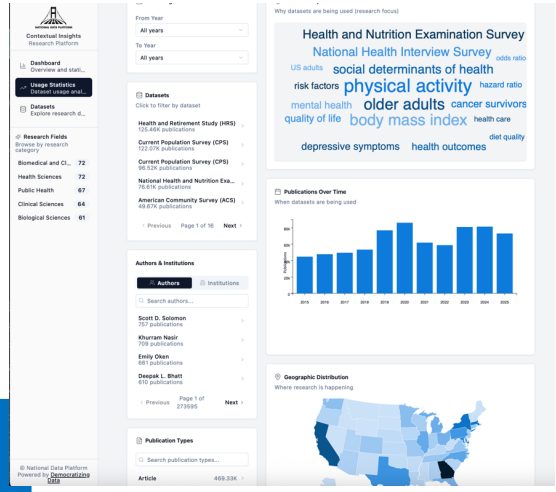
Data Discovery

- Search query returns context metadata for matching datasets, which is summarized for the user.
- The user has the option to launch the portal for a more in-depth exploration.

Data Use

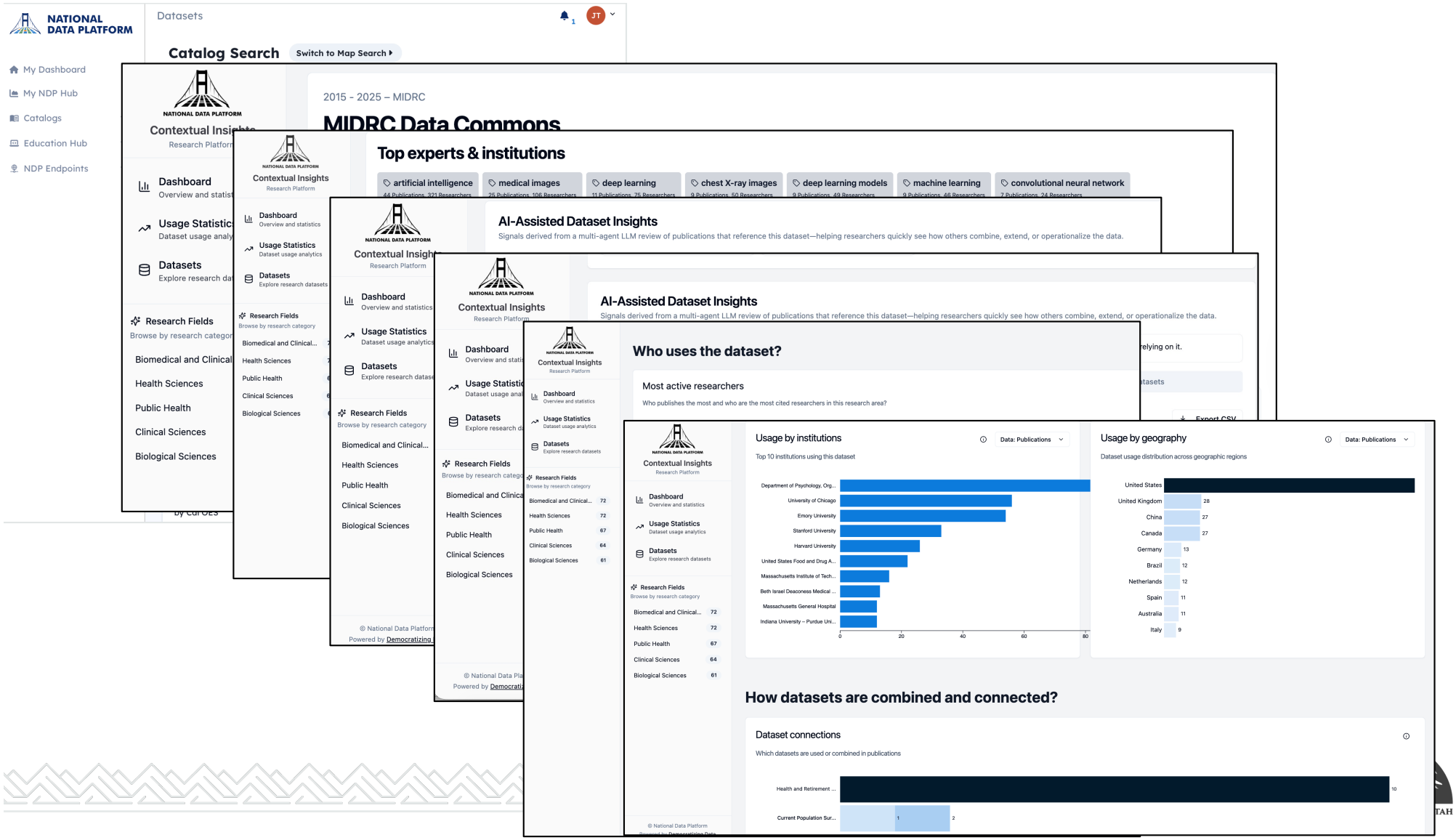
- The user provides feedback on datasets/context metadata and fit for use.
- Context metadata is periodically updated.



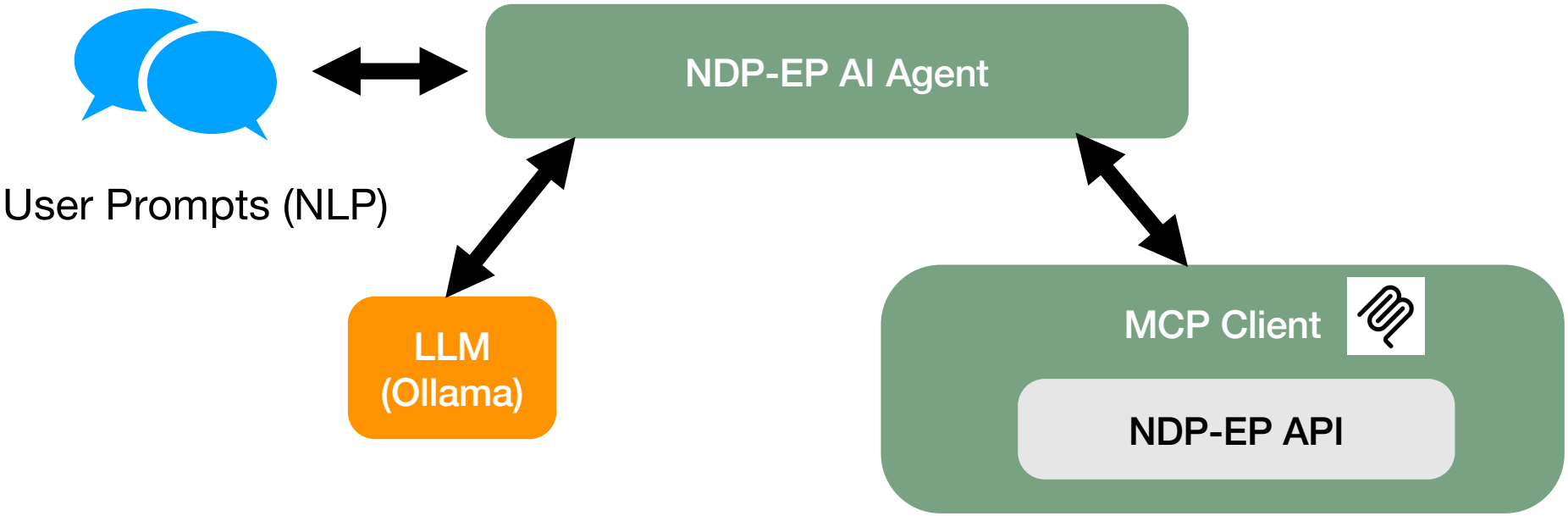


Powered by:

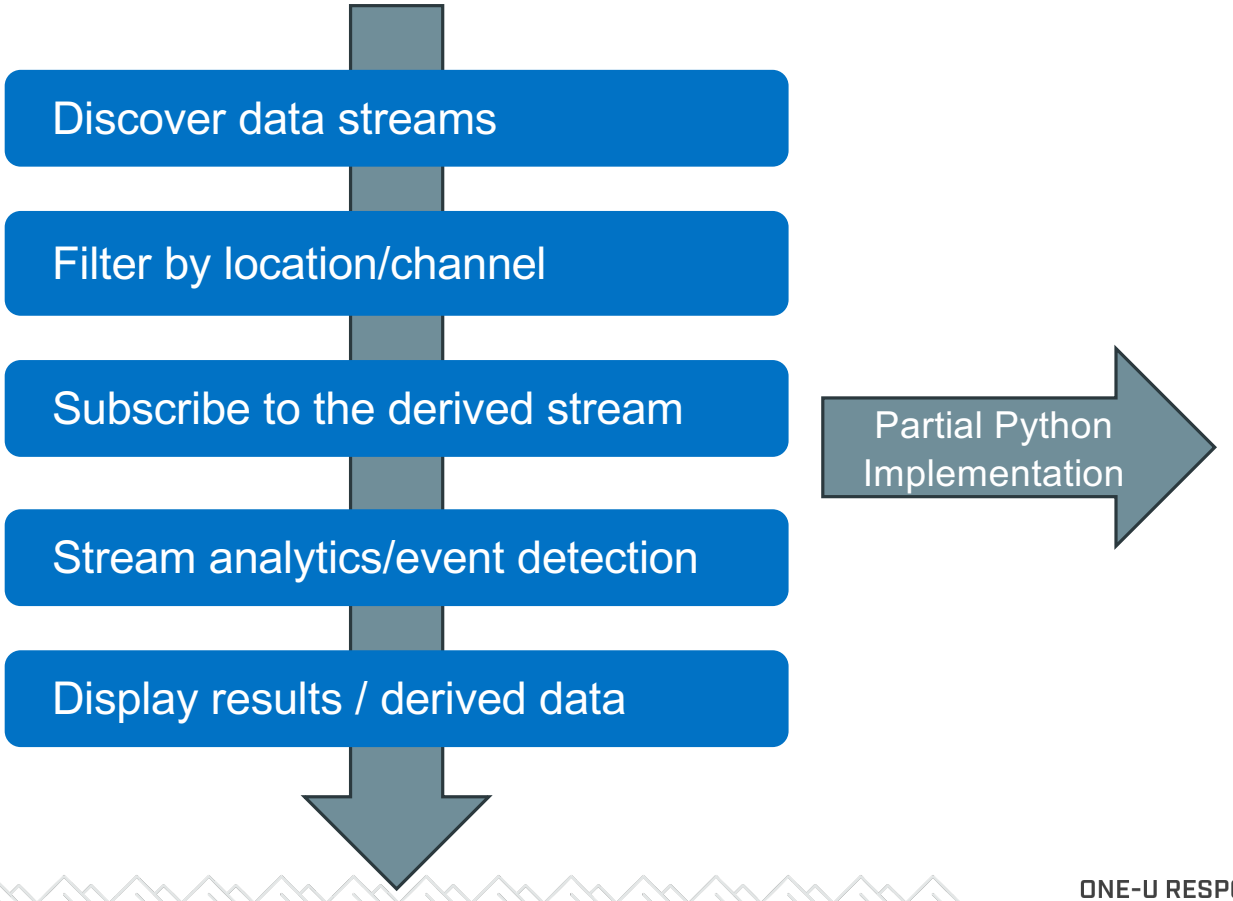
Democratizing data


NATIONAL DATA PLATFORM
Datasets
Catalog Search
Switch to Map Search
My Dashboard
My NDP Hub
Catalogs
Education Hub
NDP Endpoints
2015 - 2025 – MIDRC
MIDRC Data Commons
Top experts & institutions
artificial intelligence
medical images
deep learning
chest X-ray images
deep learning models
machine learning
convolutional neural network
Contextual Insights
Research Platform
Dashboard
Overview and statistics
Usage Statistics
Dataset usage analytics
Datasets
Explore research datasets
Research Fields
Browse by research category
Biomedical and Clinical...
Health Sciences
Public Health
Clinical Sciences
Biological Sciences
AI-Assisted Dataset Insights
Signals derived from a multi-agent LLM review of publications that reference this dataset—helping researchers quickly see how others combine, extend, or operationalize the data.
Who uses the dataset?
Most active researchers
Who publishes the most and who are the most cited researchers in this research area?
Usage by institutions
Top 10 institutions using this dataset
Data: Publications
Department of Psychology, Org...
University of Chicago
Emory University
Stanford University
Harvard University
United States Food and Drug A...
Massachusetts Institute of Tech...
Beth Israel Deaconess Medical ...
Massachusetts General Hospital
Indiana University – Purdue Uni...
Usage by geography
Dataset usage distribution across geographic regions
United States
United Kingdom 28
China 27
Canada 27
Germany 13
Brazil 12
Netherlands 12
Spain 11
Australia 11
Italy 9
How datasets are combined and connected?
Dataset connections
Which datasets are used or combined in publications
Health and Retirement ... 10
Current Population Sur... 1 2

Integrating Agentic AI



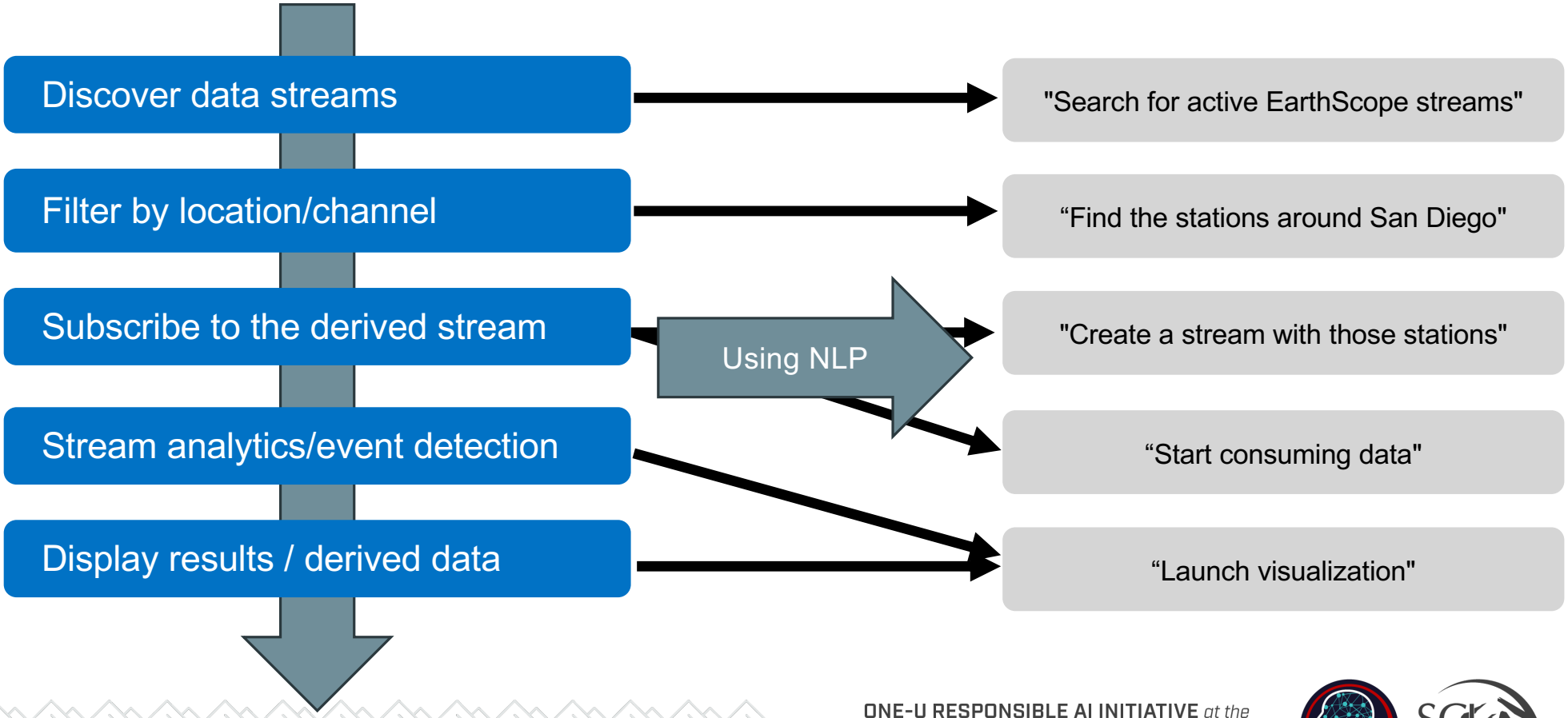
SciDx Data Streaming Pipeline



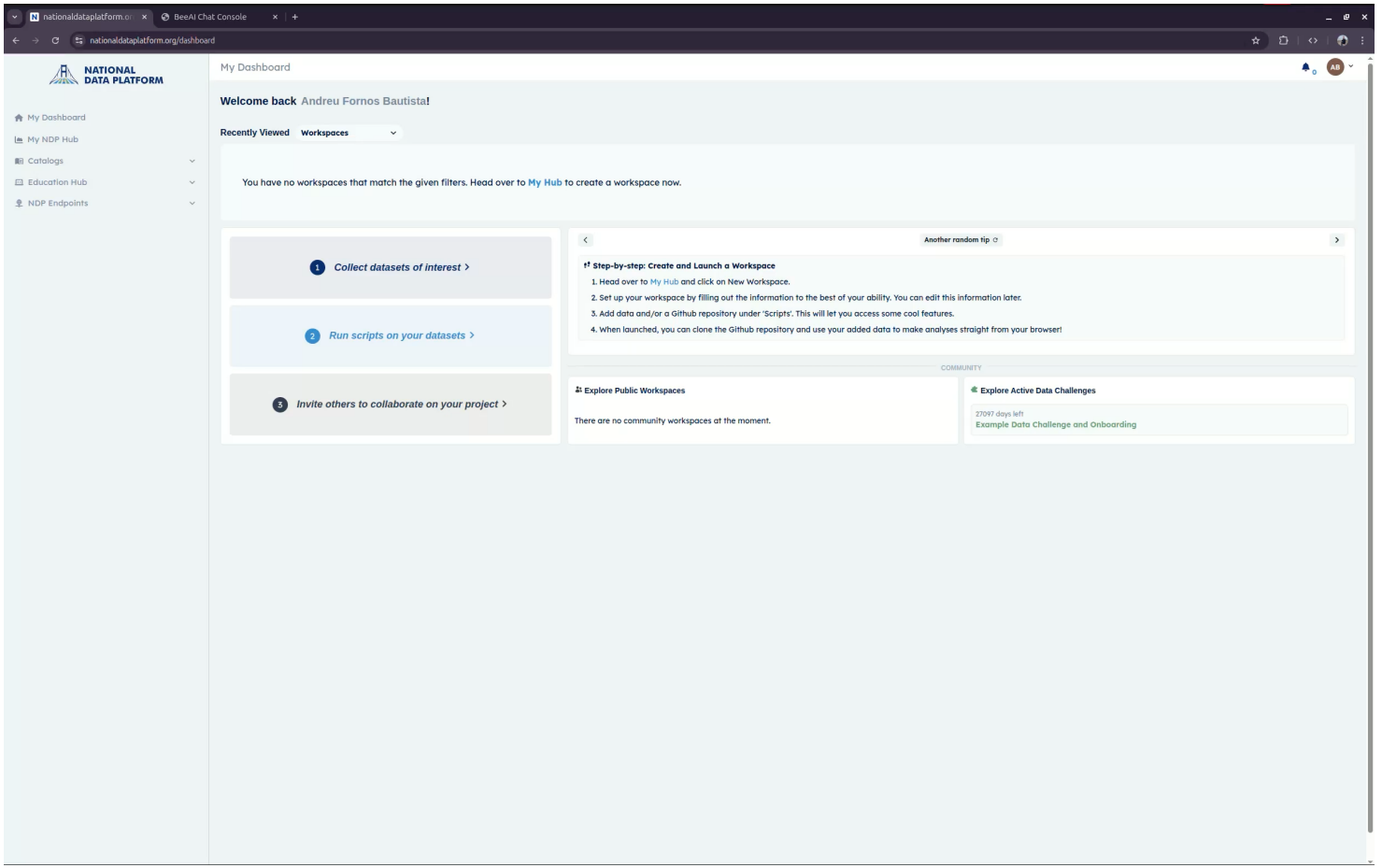
Partial Python Implementation

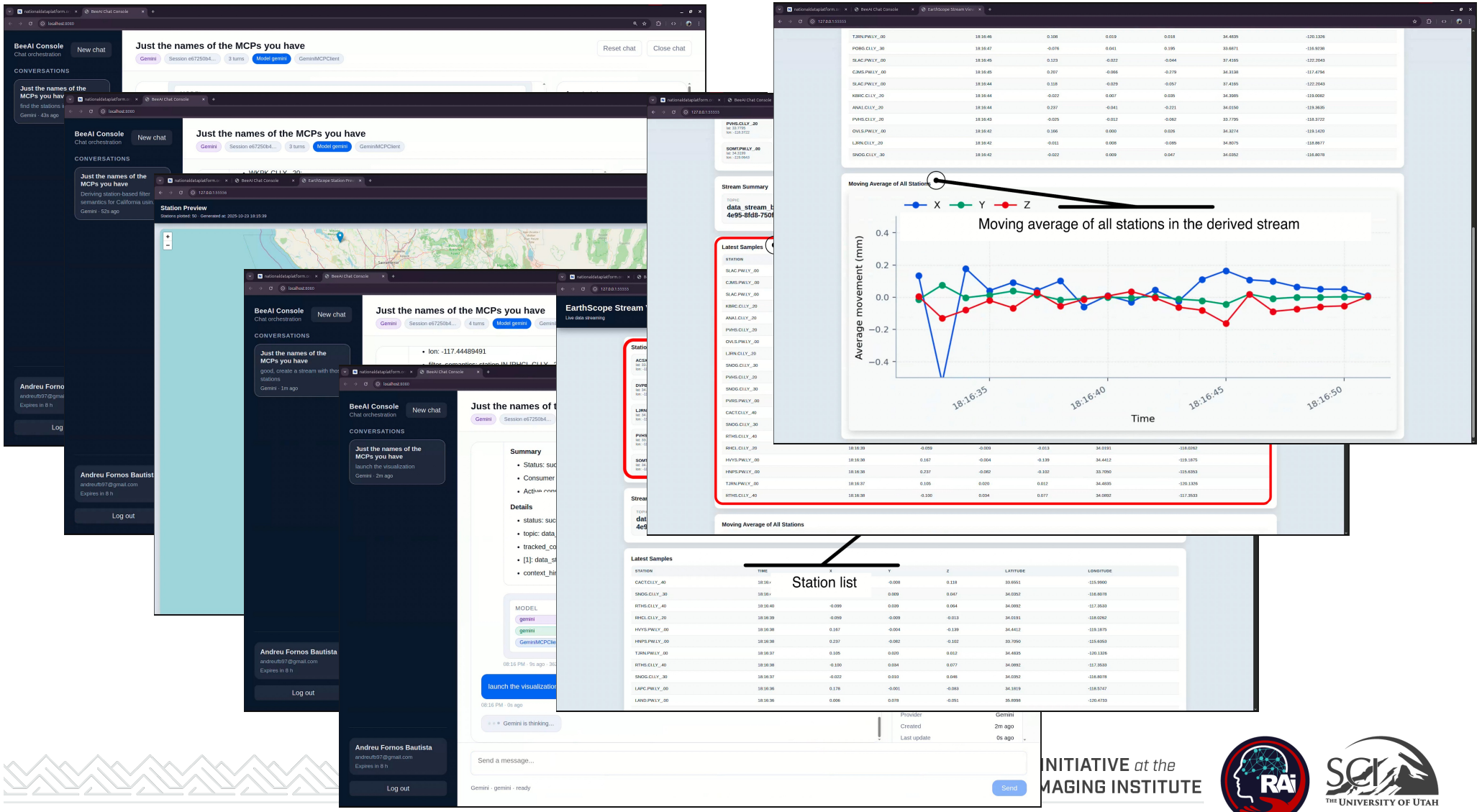
```python
# Discover streams
streaming.search_consumption_methods(
    terms = ["name"], types = ["*"], server = "local"
    )

# Create a Derived Kafka stream with spatial/sensor filtering
stream = await streaming.create_kafka_stream(
    keywords=["sensor_data"], match_all=True, filter_semantics=filters,
    username="username", password="password"
)

# Retrieve the stream's topic name
topic = stream["data_stream_id"]

# Start consuming the filtered Kafka stream
consumer = streaming.consume_kafka_messages(topic)

# Retrieve the actual state of the real-time DataFrame
df = consumer.dataframe

# Apply temporal filtering
df['MA'] = df.rolling(window=5).mean()

# Plot/dispay data

# Pad all lists in each row to the same length
def pad_row(row):
    max_len = max(len(val) if isinstance(val, list) else 1 for val in row)
    return pd.Series({
        col: (val if isinstance(val, list) else [val]) + [np.nan] *
             (max_len - len(val) if isinstance(val, list) else max_len - 1)
        for col, val in row.items()
    })

# Apply padding and explode
df_exploded = df.apply(pad_row, axis=1).explode(df.columns.tolist()).reset_index(drop=True)
df = df_exploded

# Format dataframe for display
total_rows = len(df)
display_rows = min(max_rows, total_rows)

result = f"📊 Consumer Dataframe for topic: {topic}\n"
result += f"📈 Total rows: {total_rows}\n"
result += f"🗒️ Columns: {list(df.columns)}\n"
result += f"🔍 Showing first {display_rows} rows:\n\n"

# Convert dataframe to string representation
result += df.head(max_rows).to_string(index=False)
```

SciDx Data Streaming Pipeline



Demo: Using Agentic-AI for Data Discovery and Use





Thank you!

Manish Parashar

Email: manish.parashar@utah.edu

WWW: manishparashar.org / sci.utah.edu / rai.utah.edu