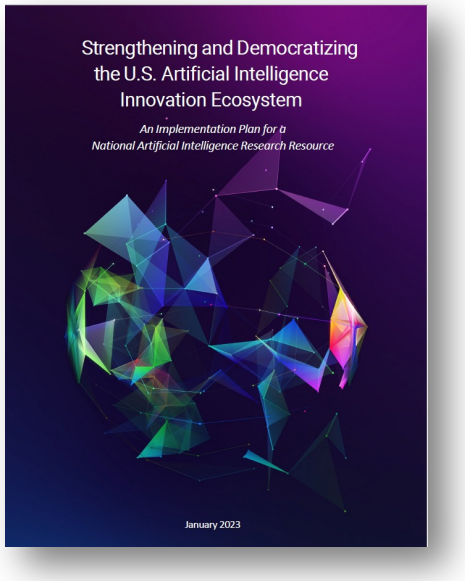
FAIR data for AI and AI for FAIR data



AKA Democratization of Data

Julia Lane, NYU

Terry McConnell, UNESCO

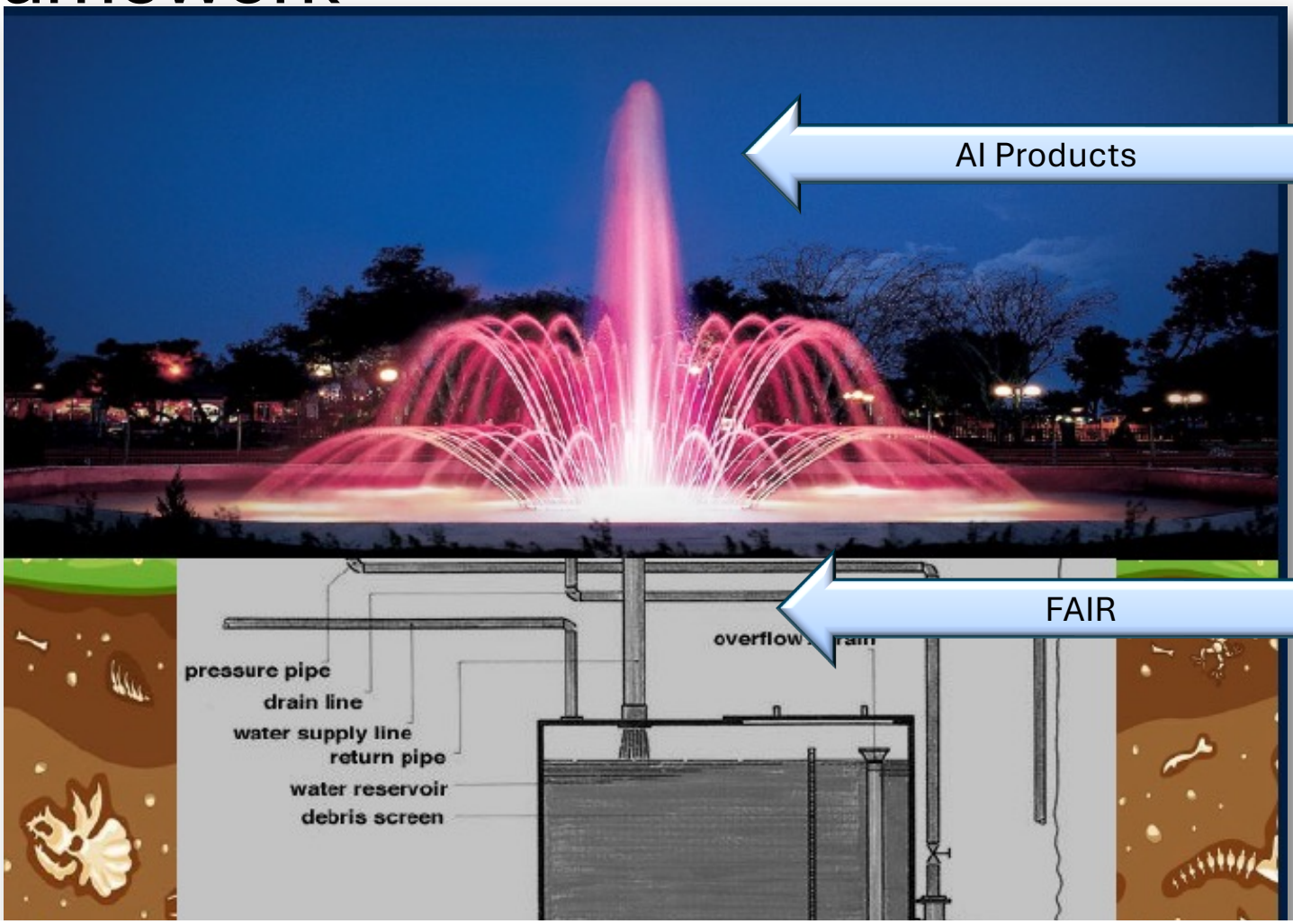
Outline

- Organizing framework
- US experience
- Theory of change
- Incentivize and democratize
- A possible agenda

Outline

- Organizing framework
- US experience
- Theory of change
- Democratization
- A possible agenda

Framework



Democratization

Outline

- Organizing framework
- US experience
- Theory of change
- Incentivize and democratize
- A possible agenda

Motivating Example: NAIRR

Administration The

JANUARY 24, 2023

National Artificial Intelligence Research Resource Task Force Releases Final Report

OSTP › NEWS & UPDATES › PRESS RELEASES

Today, the National Artificial Intelligence Research Resource (NAIRR) Task Force released its final report ↗, a roadmap for standing up a national research infrastructure that would broaden access to the resources essential to artificial intelligence (AI) research and development.

While AI research and development (R&D) in the United States is advancing rapidly, opportunities to pursue cutting-edge AI research and new AI applications are often inaccessible to researchers beyond those at well-resourced companies, organizations, and academic institutions. A NAIRR would change that by providing AI researchers and students with significantly expanded access to computational resources, high-quality data, educational

Administration The Record

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

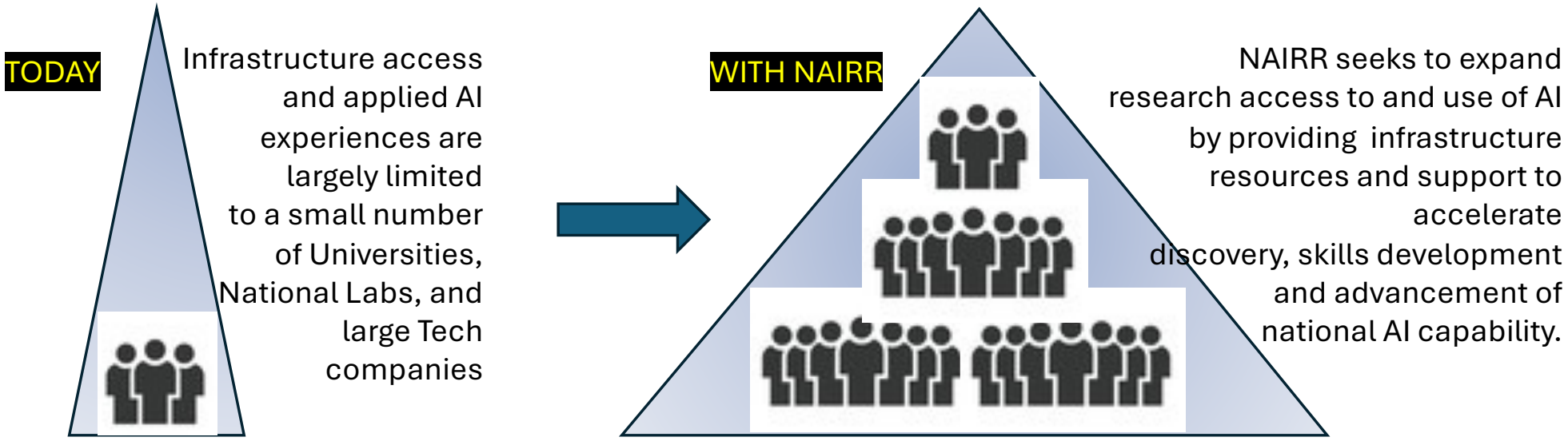
BRIEFING ROOM › PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor

Rationale

The NAIRR was created to expand the base for AI basic and applied research, to build a larger population of skilled people for US jobs, and to expand depth of knowledge about the capabilities of advanced technology.



Source: Suzette Kent

Outline

- Organizing framework
- US experience
- Theory of change
- Incentivize and democratize
- A possible agenda

Baker wasn't alone in winning the Nobel Prize for Chemistry. The Royal Swedish Academy of Sciences awarded it to Demis Hassabis, the cofounder and CEO of Google DeepMind, and John M. Jumper, a director at the same company, too. Google DeepMind was awarded for its research on AlphaFold, a tool which can predict how proteins are structured, while Baker was recognized for his work using AI to design new proteins. R
more about it here.

Meanwhile, the physics prize went to Geoffrey Hinton, a computer scientist whose pioneering work on deep learning in the 1980s and '90s underpins all of the most powerful AI models in the world today, and fellow computer scientist John Hopfield, who invented a type of pattern-matching neural network that can store and . Read more about it here.

ARTIFICIAL INTELLIGENCE

A c
No

AI's us

By Me

Octob

ein folding, and the other
aker for tools to help
w proteins.

But there is one problem. AI needs masses of high-quality data to be useful for science, and databases containing that sort of data are rare, says Baker.

The prize is a recognition for the whole community of people working as protein designers. It will help move protein design from the "lunatic fringe of stuff that no one ever thought would be useful for anything to being at the center stage," he says.

Key findings from NAIRTF

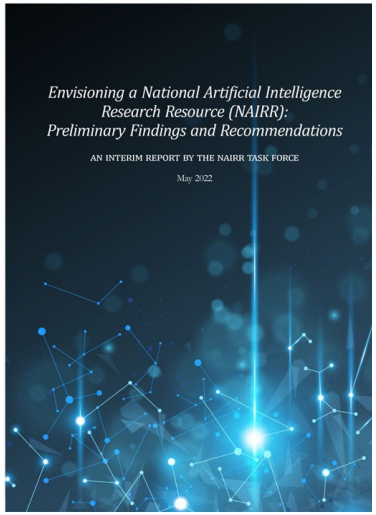
Finding 4-1: Rigorous AI R&D is often not possible without high-quality, trusted, dense, and transparent data resources.

Finding 4-2: There are substantial data quality challenges within and across most research domains.

Finding 4-3: Data curation is a substantial challenge for researchers in all domains.

Finding 4-4: There are substantial costs to combining and linking heterogeneous data.

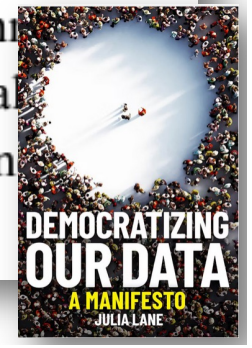
Finding 4-5: There are opportunities to learn from, complement, or build on ongoing activities to leverage data for the public good.



Recommendation: Democratize data

Role of the Operating Entity in Incentivizing and Curating Contributed Datasets and Other Resources

Since the quality of many AI models depends on high-quality training and test data, the Operating Entity should establish a data service that facilitates access to and additional use of existing curated datasets of value and interest to the NAIRR user community. Curation of AI data, models, tools, and workflows should be done by the user community in an AI data commons, facilitated by the NAIRR search and discovery platform. Such a community system, governed by terms of use as well as a review system, would facilitate data sharing and curation by members of the community. In the context of a commons model, researchers who contribute to the comm good through data curation and code sharing, and whose contributions are recognized and val by relevant communities, could be incentivized through high-profile NAIRR recognition an preferential access to NAIRR resources.



Theory of Change

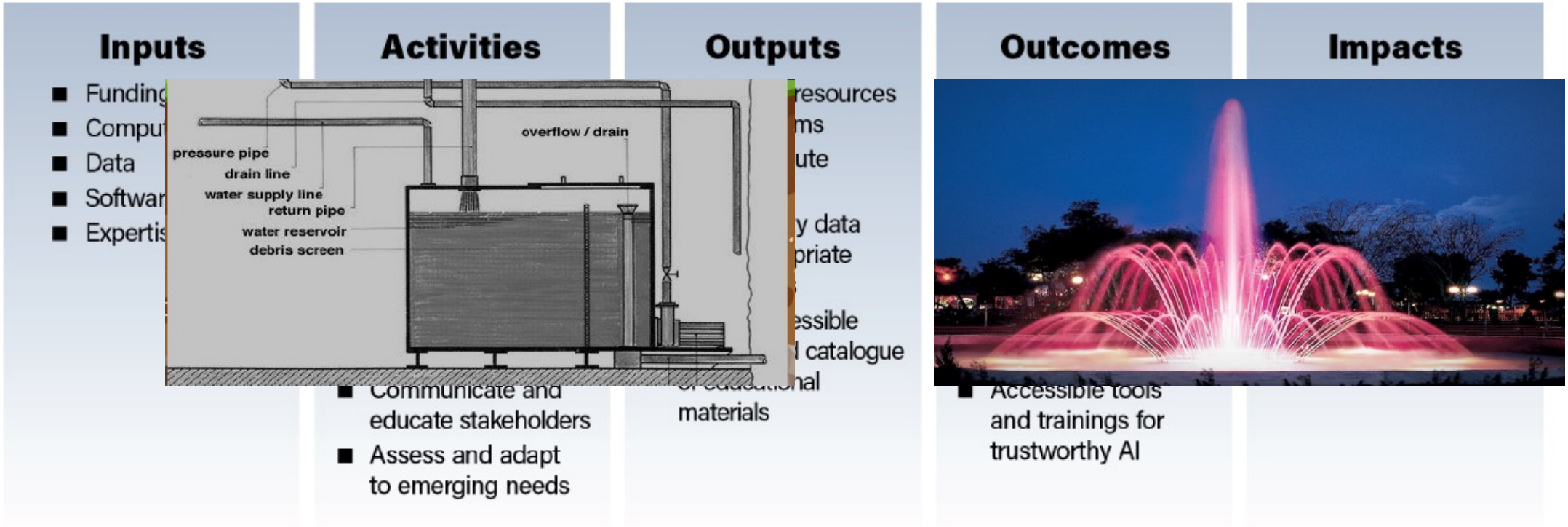


Figure 6. Example Elements of a Theory of Change for the NAIRR

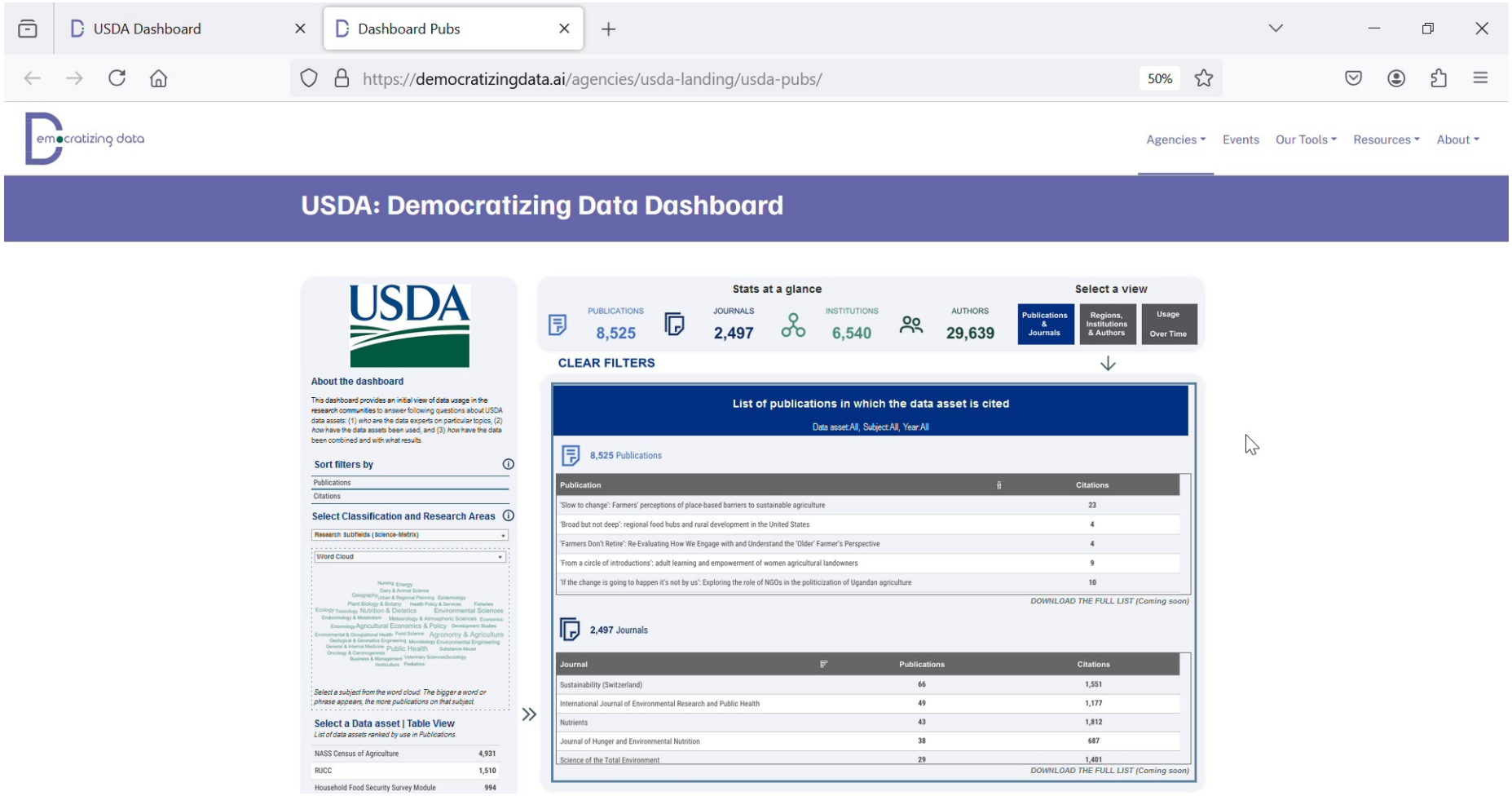
Outline

- Organizing framework
- US experience
- Theory of change
- Incentivize and democratize
- A possible agenda

Incer

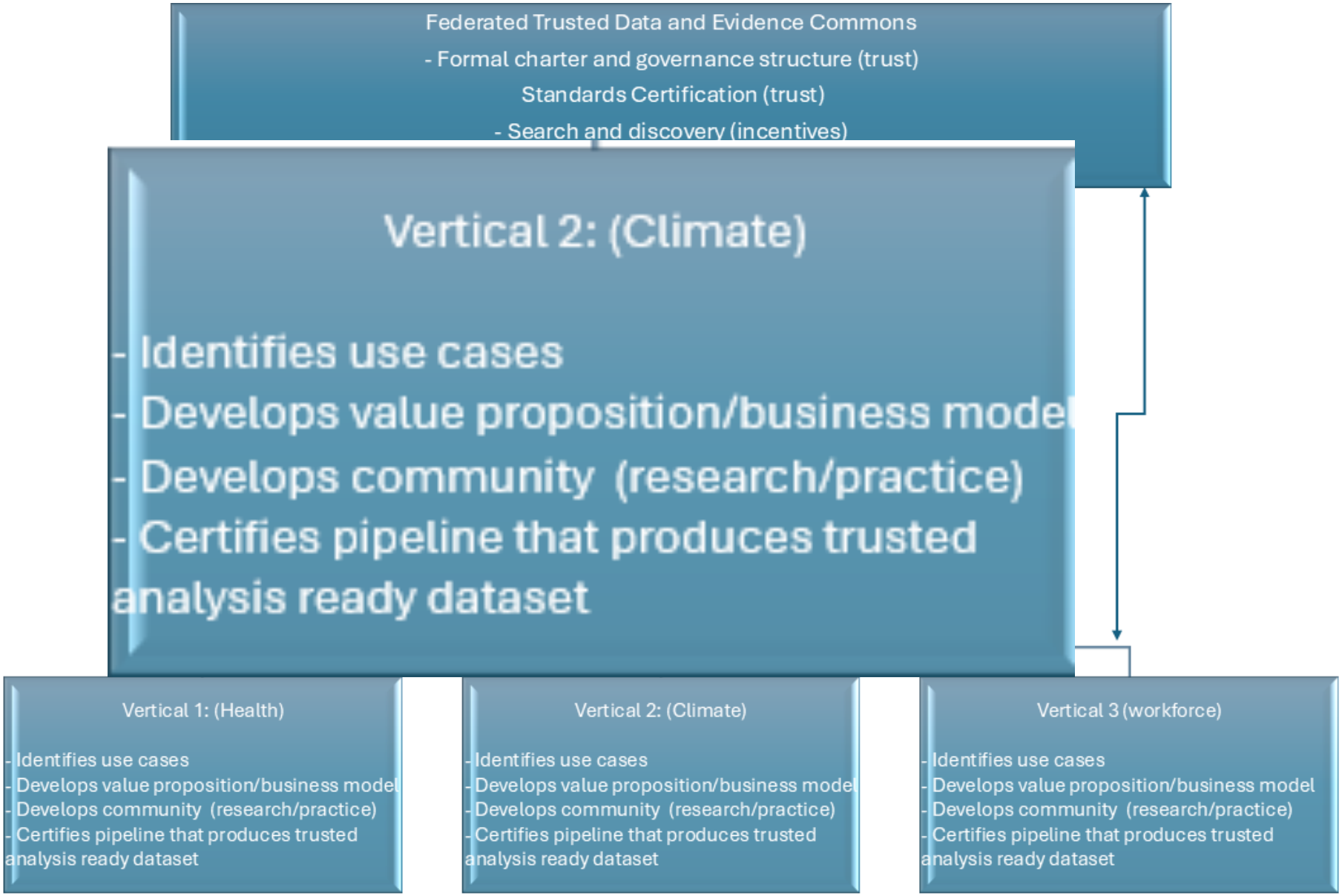
hdsr.mitpress.mit.edu/pub/6blzwgpg/release/4?readingCollection=a6db7dff
HOME
ISSUES
SECTIONS
COLUMNS
COLLECTIONS
PODCAST
SUBMIT
ABOUT
MASTHEAD
Special Issue 4: Democratizing Data
Published on Apr 02, 2024
DOI 10.1162/99608f92.03719804
SHOW DETAILS
CITE
SOCIAL
DOWNLOAD
CONTENTS
An Invisible Hand for Creating Public Value From Data
by Julia Lane, Alfred Spector, and Michael Stebbins
last released
5 months ago
ABSTRACT
In this article we describe how data usage statistics could be designed to reward the contributions of both researchers and government agencies, functionally creating a carrot that rewards engagement. We provide a concrete use case—the Democratizing Data pilot project—and discuss how the results could be used to develop a community-driven incentive structure, or an 'invisible hand,' to generate better evidence for policymakers and lower costs for the taxpayer. We also note the potential to extend the approach and platform to other data-driven initiatives, notably investments in artificial intelligence.

Incentivize and democratize

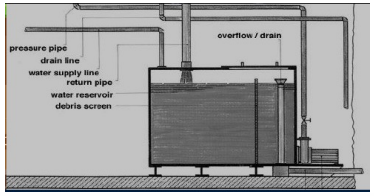


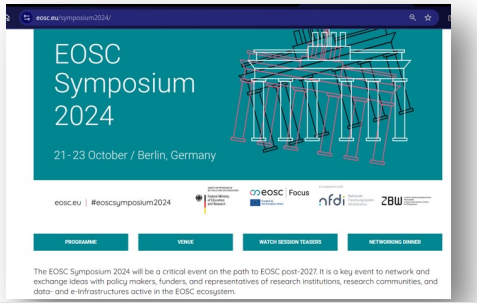
Outline

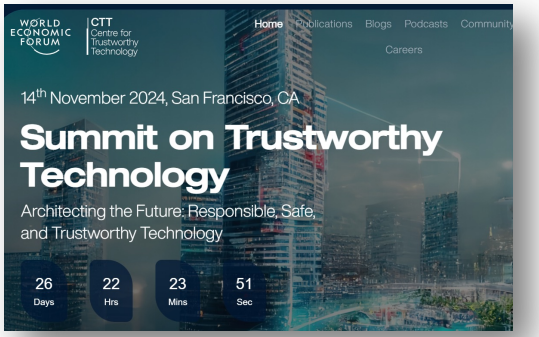
- Organizing framework
- US experience
- Theory of change
- Incentivize and democratize
- A possible agenda

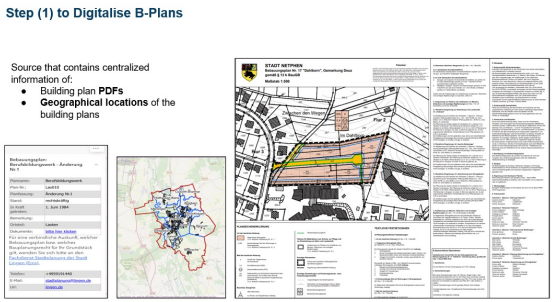

Federated Trusted Data and Evidence Commons
- Formal charter and governance structure (trust)
Standards Certification (trust)
- Search and discovery (incentives)
Vertical 2: (Climate)
- Identifies use cases
- Develops value proposition/business model
- Develops community (research/practice)
- Certifies pipeline that produces trusted analysis ready dataset
Vertical 1: (Health)
- Identifies use cases
- Develops value proposition/business model
- Develops community (research/practice)
- Certifies pipeline that produces trusted analysis ready dataset
Vertical 2: (Climate)
- Identifies use cases
- Develops value proposition/business model
- Develops community (research/practice)
- Certifies pipeline that produces trusted analysis ready dataset
Vertical 3 (workforce)
- Identifies use cases
- Develops value proposition/business model
- Develops community (research/practice)
- Certifies pipeline that produces trusted analysis ready dataset

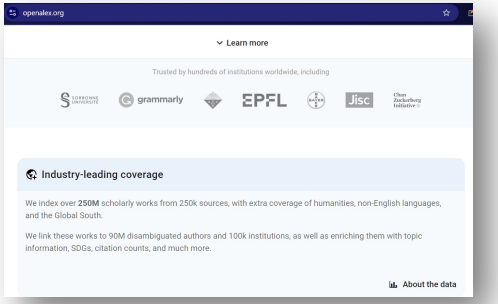
Internationalize

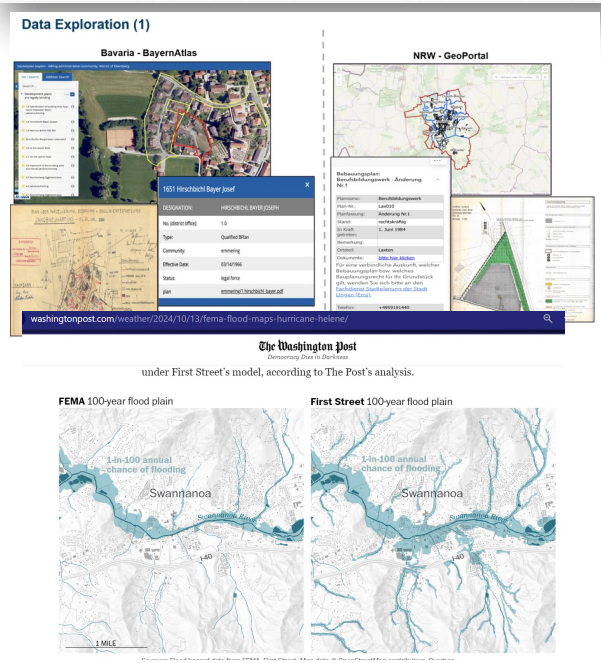




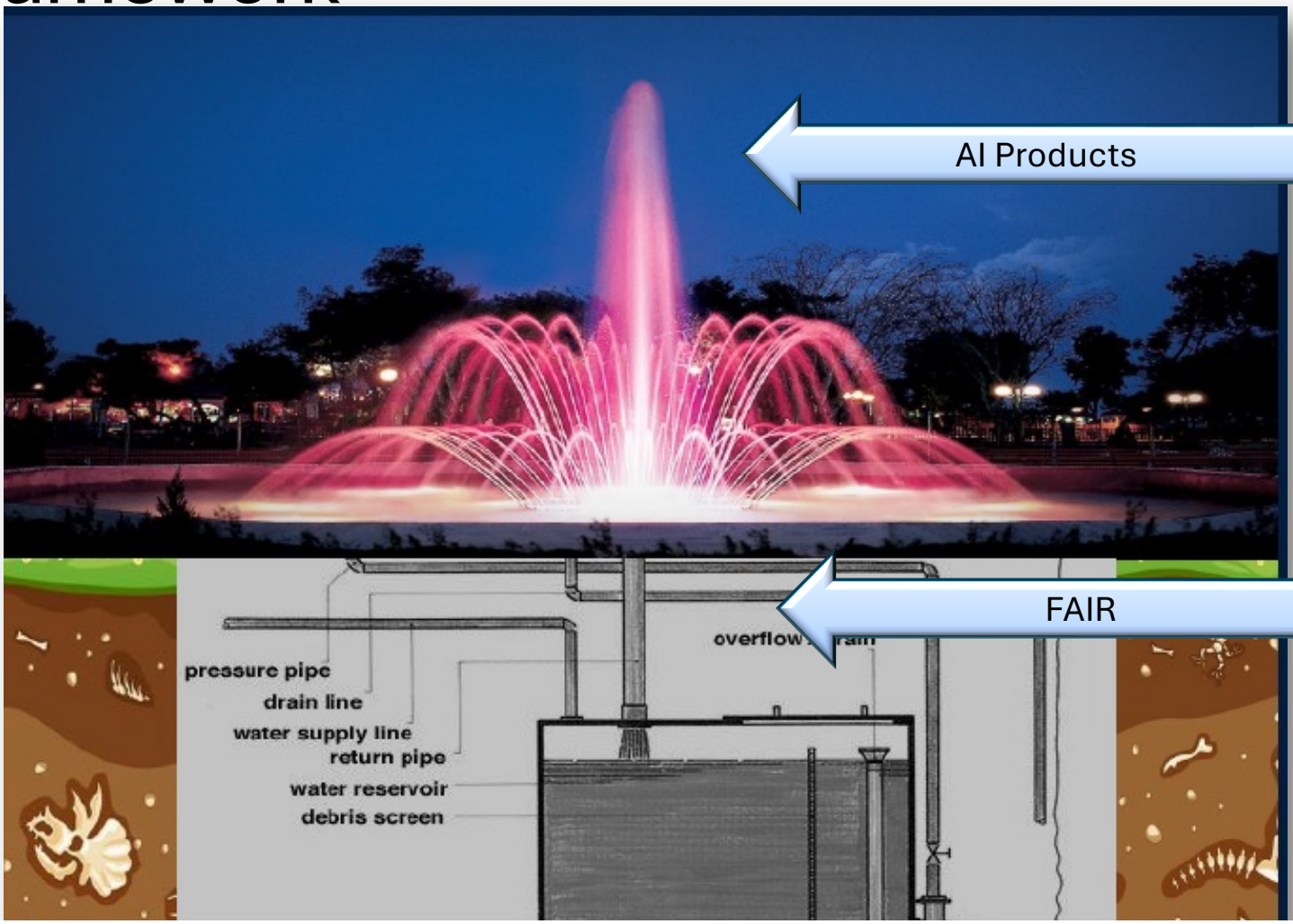








Framework



Democratization

Comments and questions

- Julia Lane
- Julia.lane@nyu.edu
- https://julialane.org/
- https://www.linkedin.com/in/julia-lane/