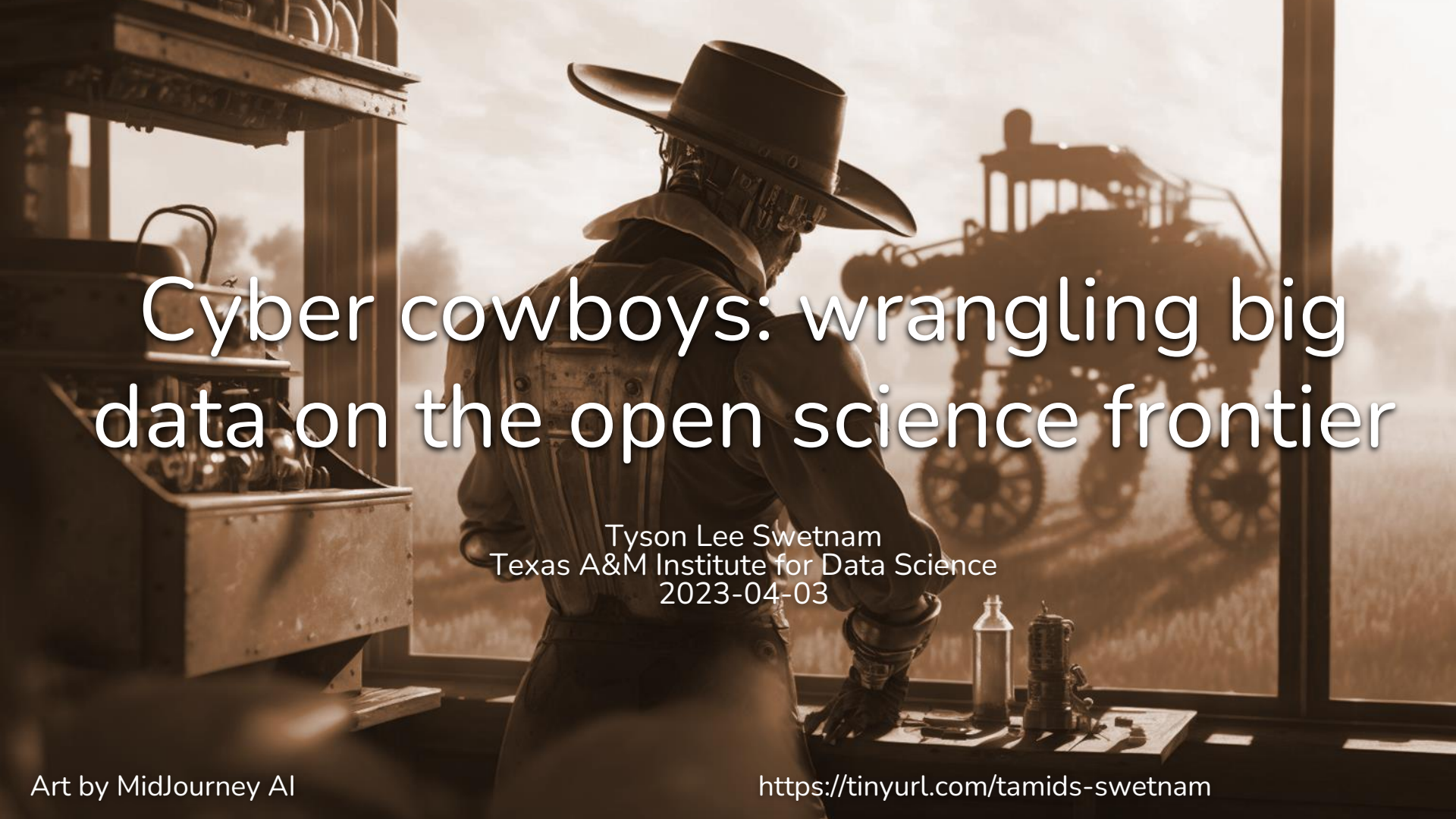


Cyber cowboys: wrangling big data on the open science frontier

Tyson Lee Swetnam
Texas A&M Institute for Data Science
2023-04-03

Art by MidJourney AI

https://tinyurl.com/tamids-swetnam

Overview

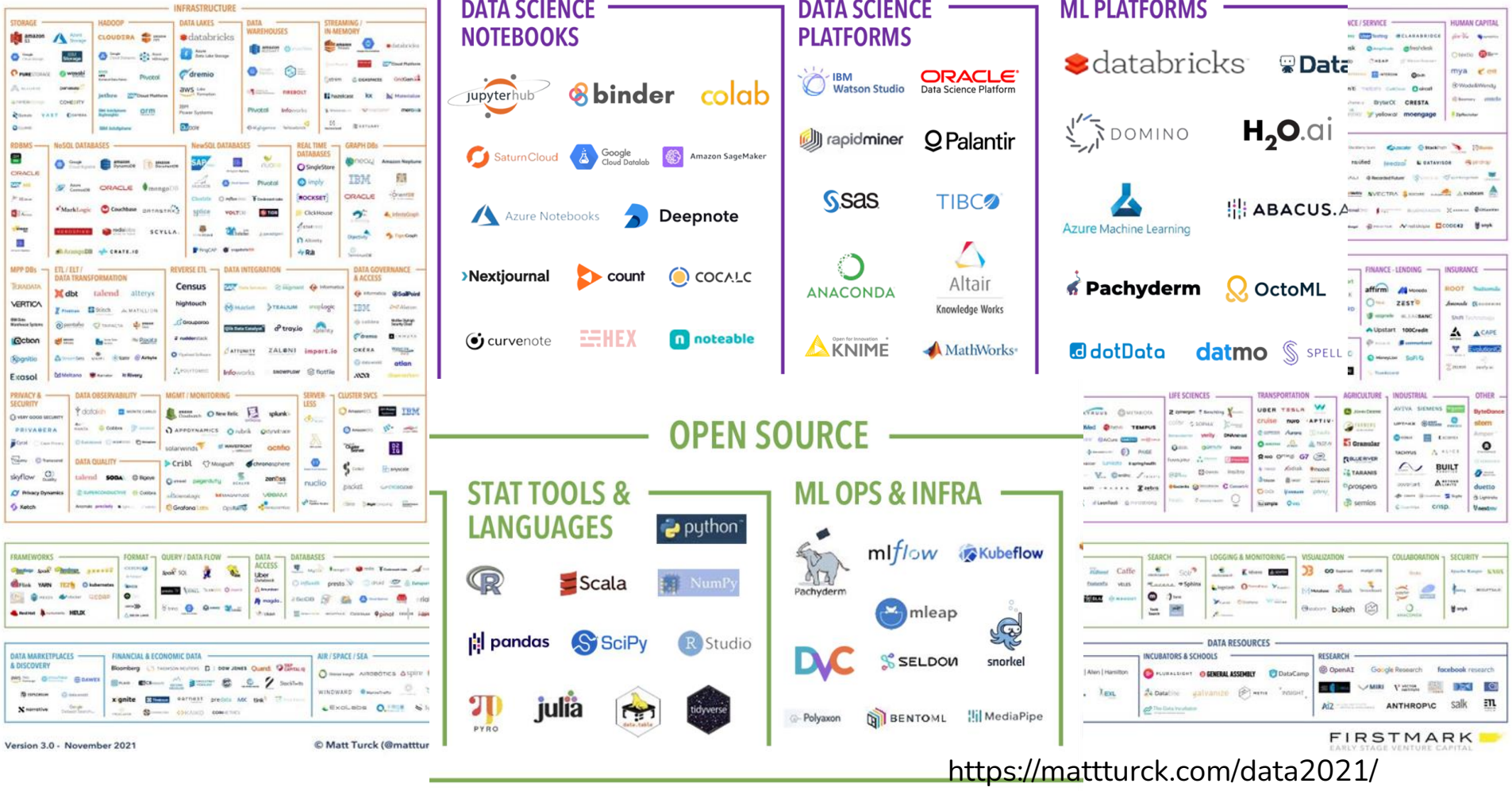
The Big Data Landscape

2023 Year of Open Science

Becoming Cloud-Native

AI in the workplace

CyVerse Science Use Cases



https://mattturck.com/data2021/

The Big Data Landscape is really, really, really, big

- >100 Thousand data scientists in the US[1]
- ~26.9 Million software engineers in the world[2]
- >15 Billion devices connected to the Internet[3]
- ~7.9 Zettabytes (7.9 Billion Terabytes) of data on the internet[4]

1 https://www.bls.gov/ooh/math/data-scientists.htm
2 https://en.wikipedia.org/wiki/Software_engineering_demographics
3 https://www.statista.com/statistics/1290925/internet-access-by-device-us/
4 https://firstsiteguide.com/big-data-stats/

Source: https://www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Dr. Alondra Nelson
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:

1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

Source: https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

How many of you are trained to make data accessible immediately after you analyze it?

document

curate

share

Capacity vs Capability

Compute
Storage
Data
Software

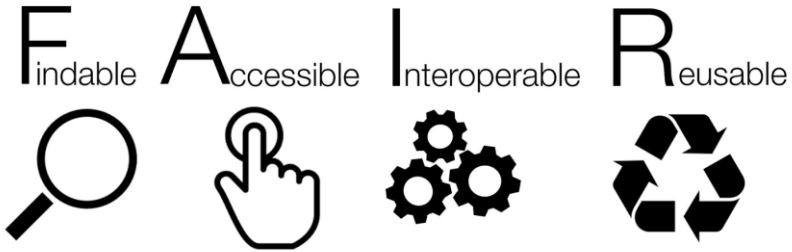
Digital Literacy
Data Science Skills

Source: https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/

2023 - Federal Year of Open Science Initiative

Source Images: NASA TOPS, National Institutes of Health, National Science Foundation

What is Open Science?



Source Images: https://www.go-fair.org/, https://opensource.org/, https://www.openaccess.io/

CYVERSE®

Welcome to Foundational Open Science Skills (FOSS) Spring 2023!

https://foss.cyverse.org

Introduction to Open Science

Learning Objectives

After this lesson, you should be able to:

- Explain what Open Science is
- Explain the components of Open Science
- Describe the behaviors of Open Science
- Explain why Open Science matters in education, research, and society
- Understand the advantages and the challenges to Open Science
- Identify who the practitioners of Open Science are
- Understand the underlying Ethos of Open Science

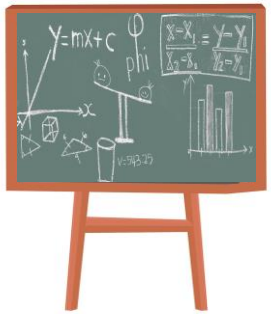
What is Open Science?

If you ask a dozen researchers this question, you will probably get just as many answers.

This means that Open Science isn't necessarily a set of checkboxes you need to tick, but rather a holistic approach to doing science.

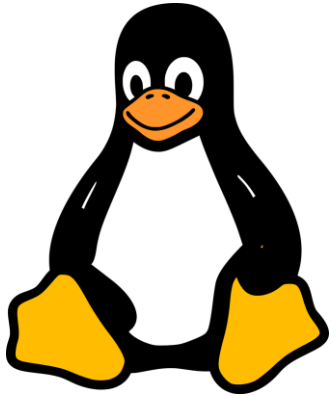
https://foss.cyverse.org

Personal Computing



Personal Computing

macOS

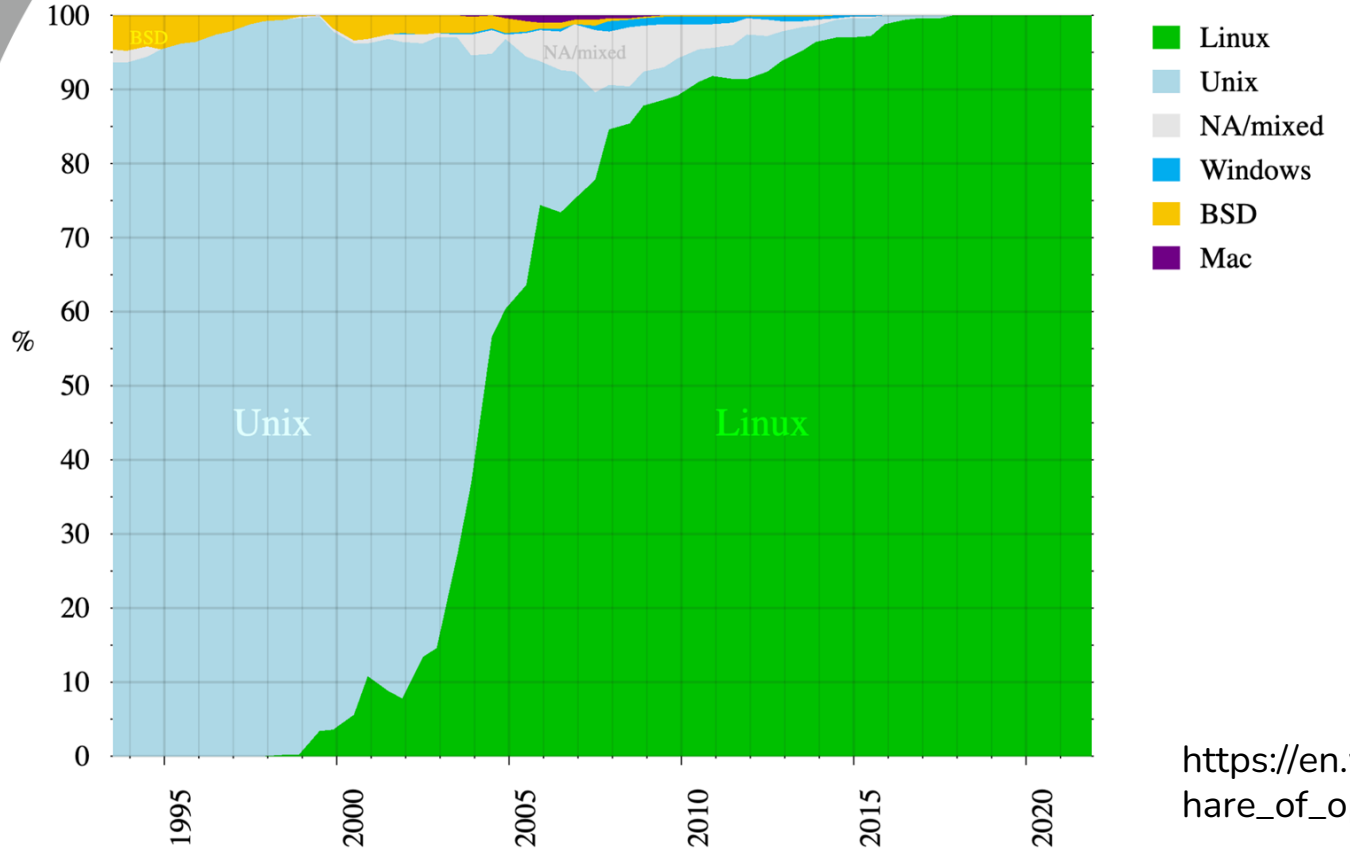


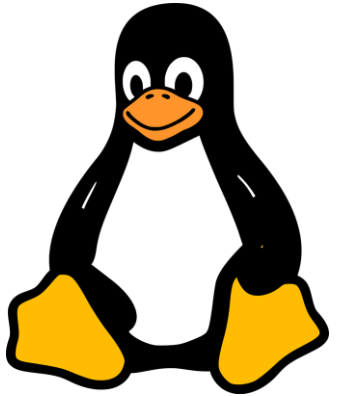
76%

16%

5%

Cloud Computing & SuperComputing

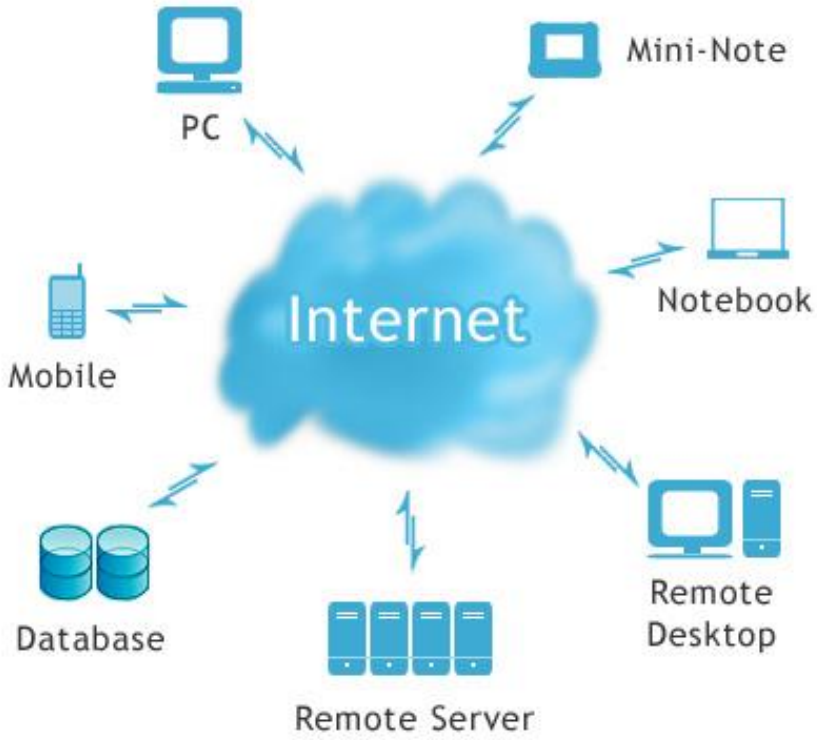




https://en.wikipedia.org/wiki/Usage_share_of_operating_systems

Let’s all use “The Cloud!”

Let’s all use “The Cloud!”



Let’s all use “The Cloud!”

“Free” Cloud services are available almost everywhere for academic educators & researchers

“Starter Tiers” can be good for teaching and getting a feel for things, but in general are not useful for big data analyses

My Daily Drivers

CYVERSE®

Laptop

arm

macOS

Desktop

AMD

ubuntu

HPC & Data Center

NVIDIA

AMD

Rocky Linux™

My Daily Drivers

GitHub

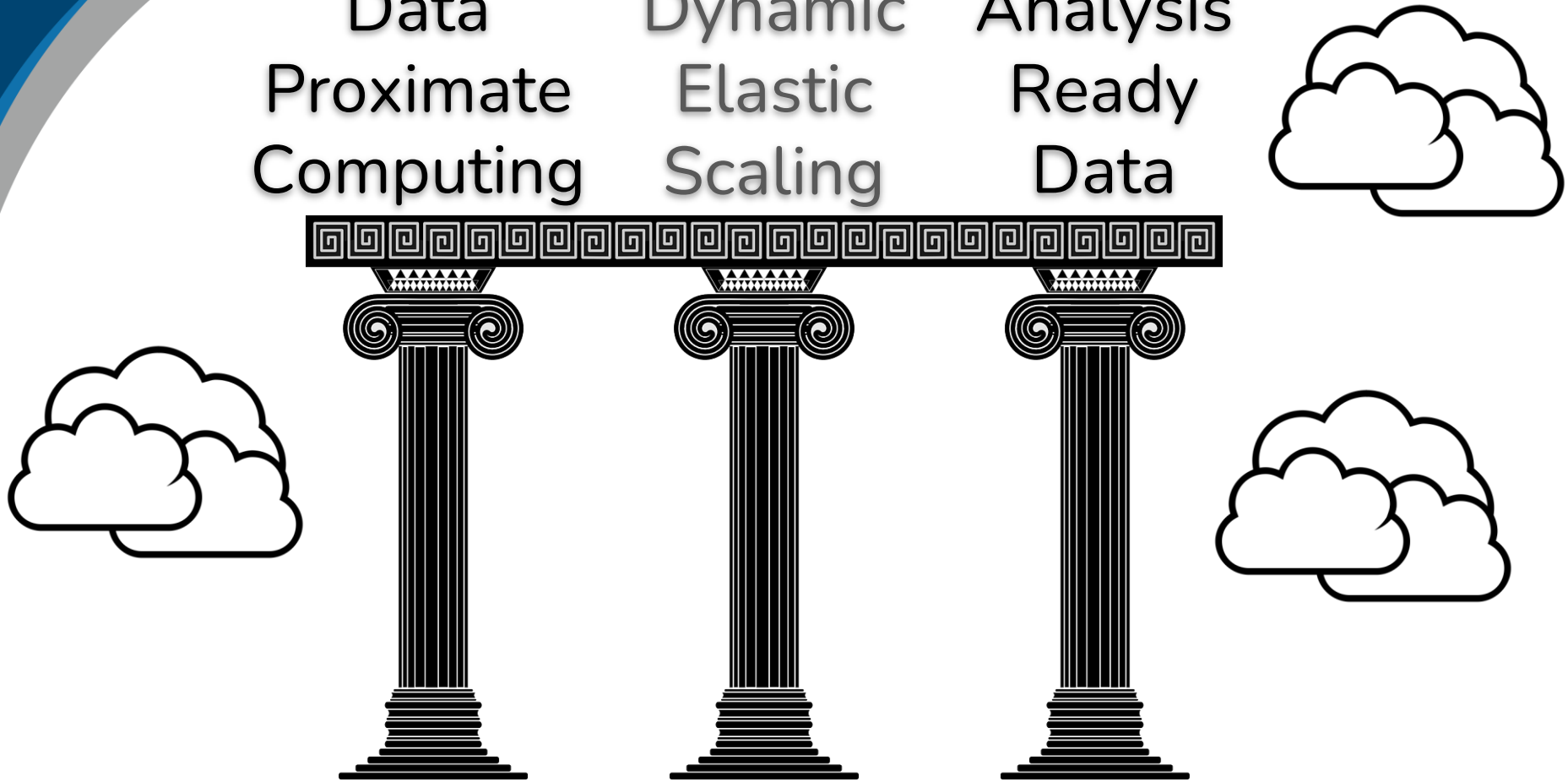
OpenAI

Microsoft

CONDA

YAML

JavaScript



https://doi.org/10.1109/MCSE.2021.3059437

Ryan Abernathey's 2022 PANGEO talk at OpenOceanCloud: https://vimeo.com/670782104

Data Proximate Computing

Mainframe — Server — Many Servers

High Density Servers

Terminal — Desktop — Web Hosting

Cloud-enabled Apps

1970s — 1980s — 1990s-2010s

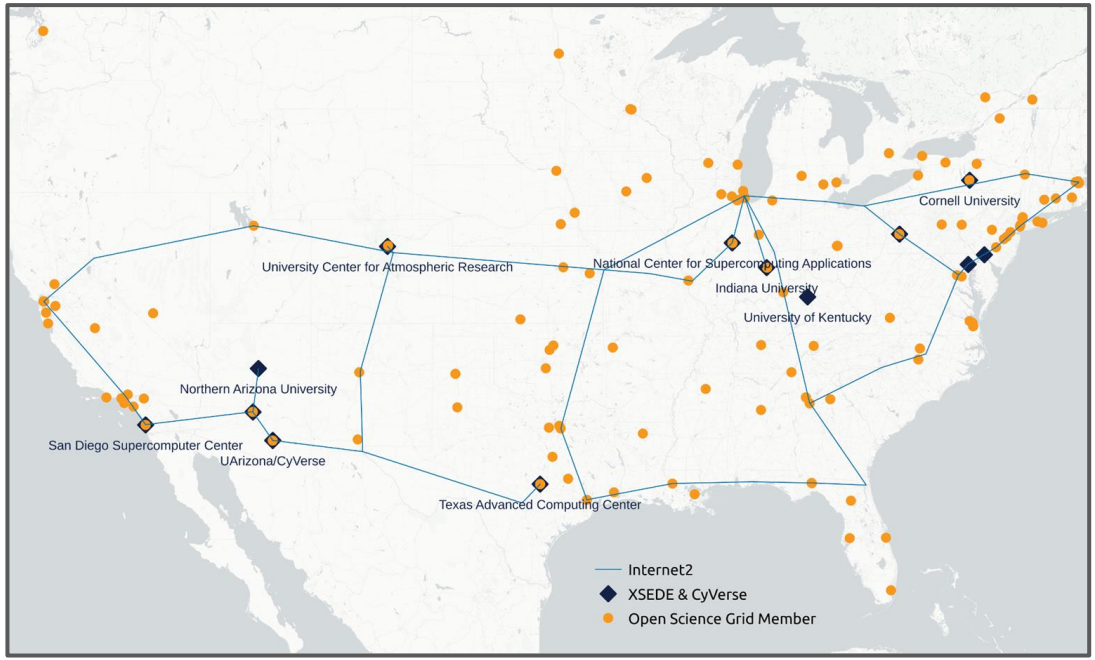
2020+

Data Proximate Computing

Connect to Public Cyberinfrastructure & Commercial Cloud via Internet2

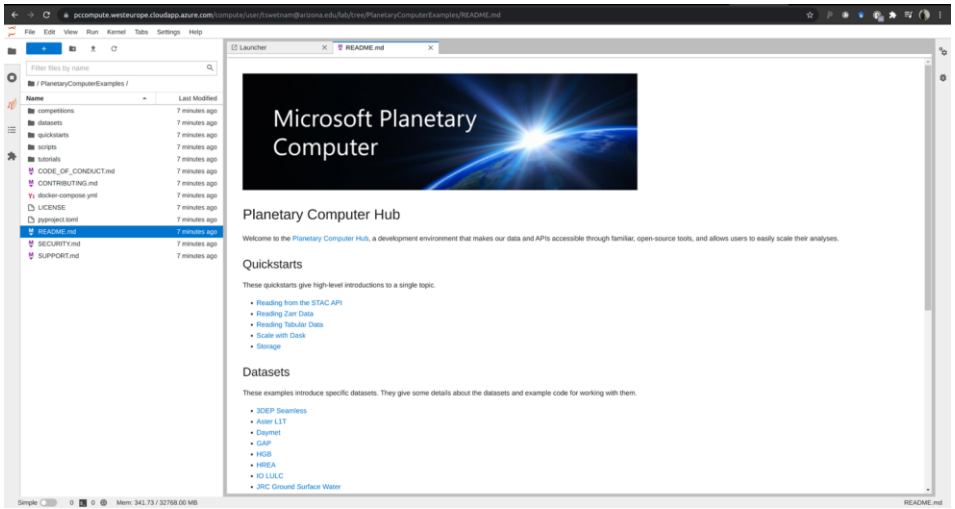
ADVANCE TO ACCESS
MAKING A SUCCESSFUL TRANSITION

PATh PARTNERSHIP to ADVANCE THROUGHPUT COMPUTING

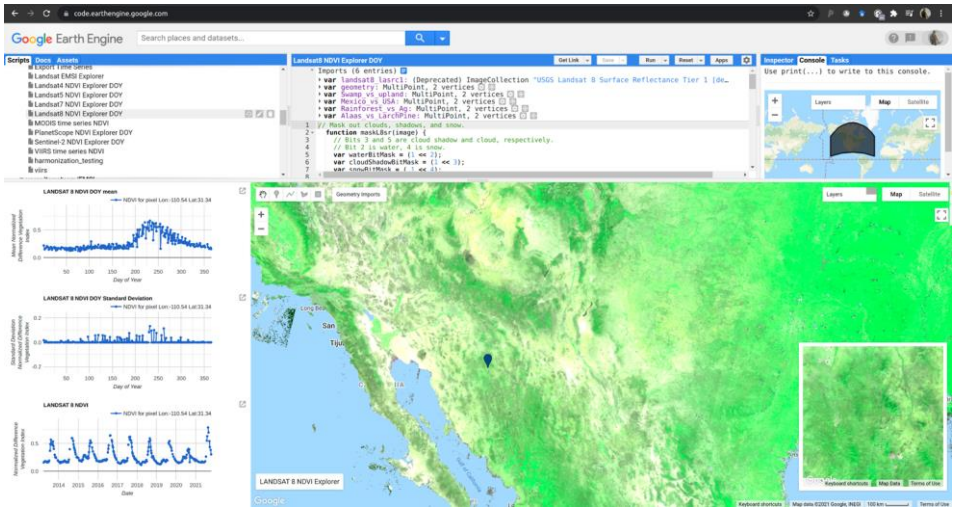


Dynamic and Elastic Scaling

https://planetarycomputer.microsoft.com/



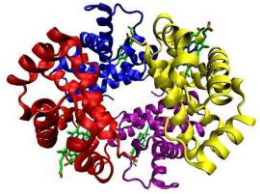
https://code.earthengine.google.com/



https://github.com/microsoft/PlanetaryComputerExamples

https://github.com/giswqs/Awesome-GEE

Analysis Ready Data

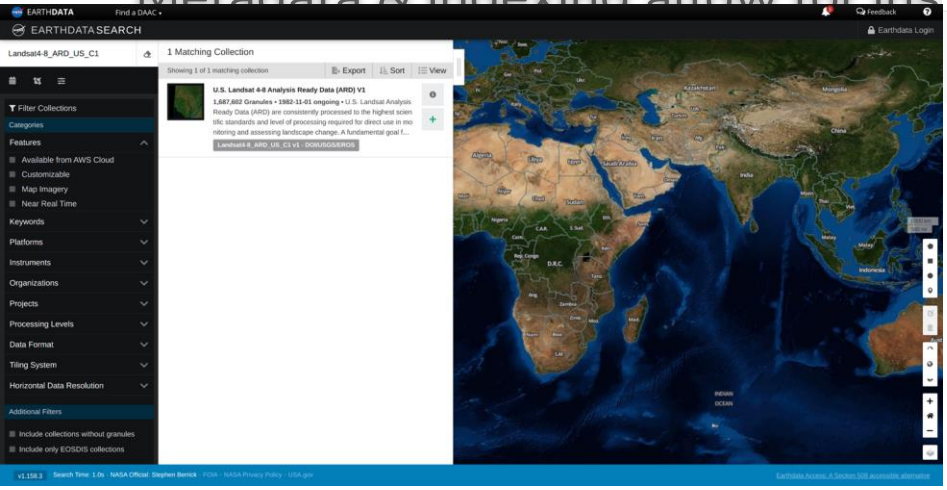


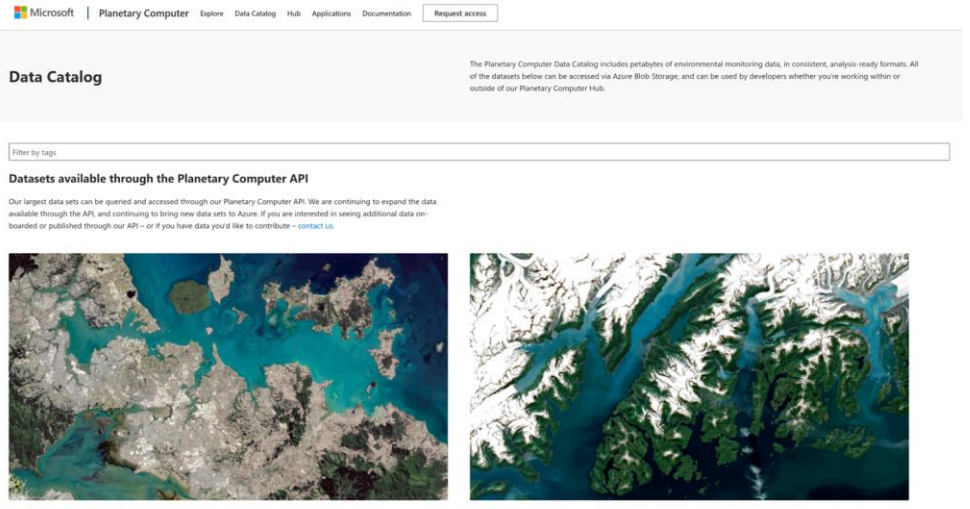
Analysis Ready Data

All major scientific data archives are hosted on the cloud - most are free

Programatically scalable to millions of concurrent requests

Metadata & indexing allow for ins

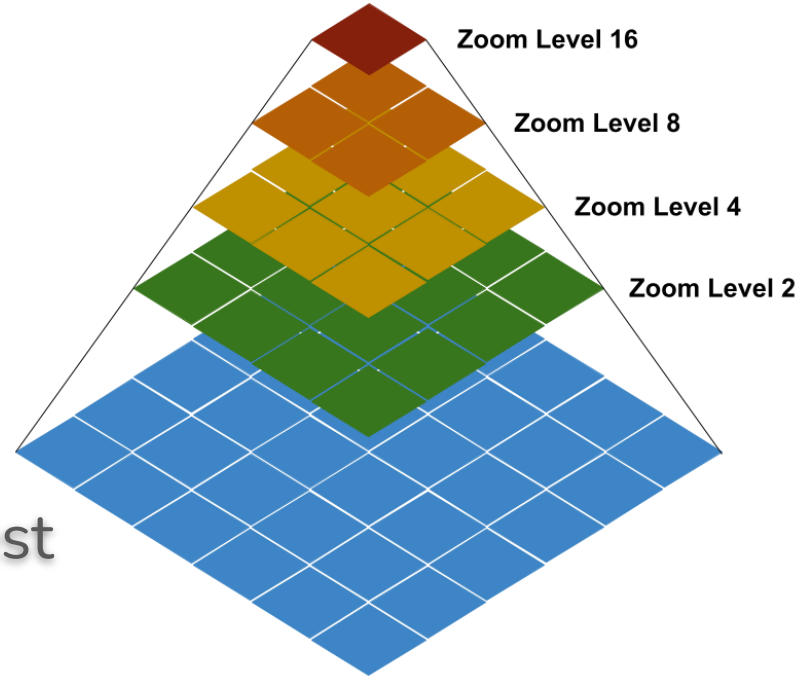


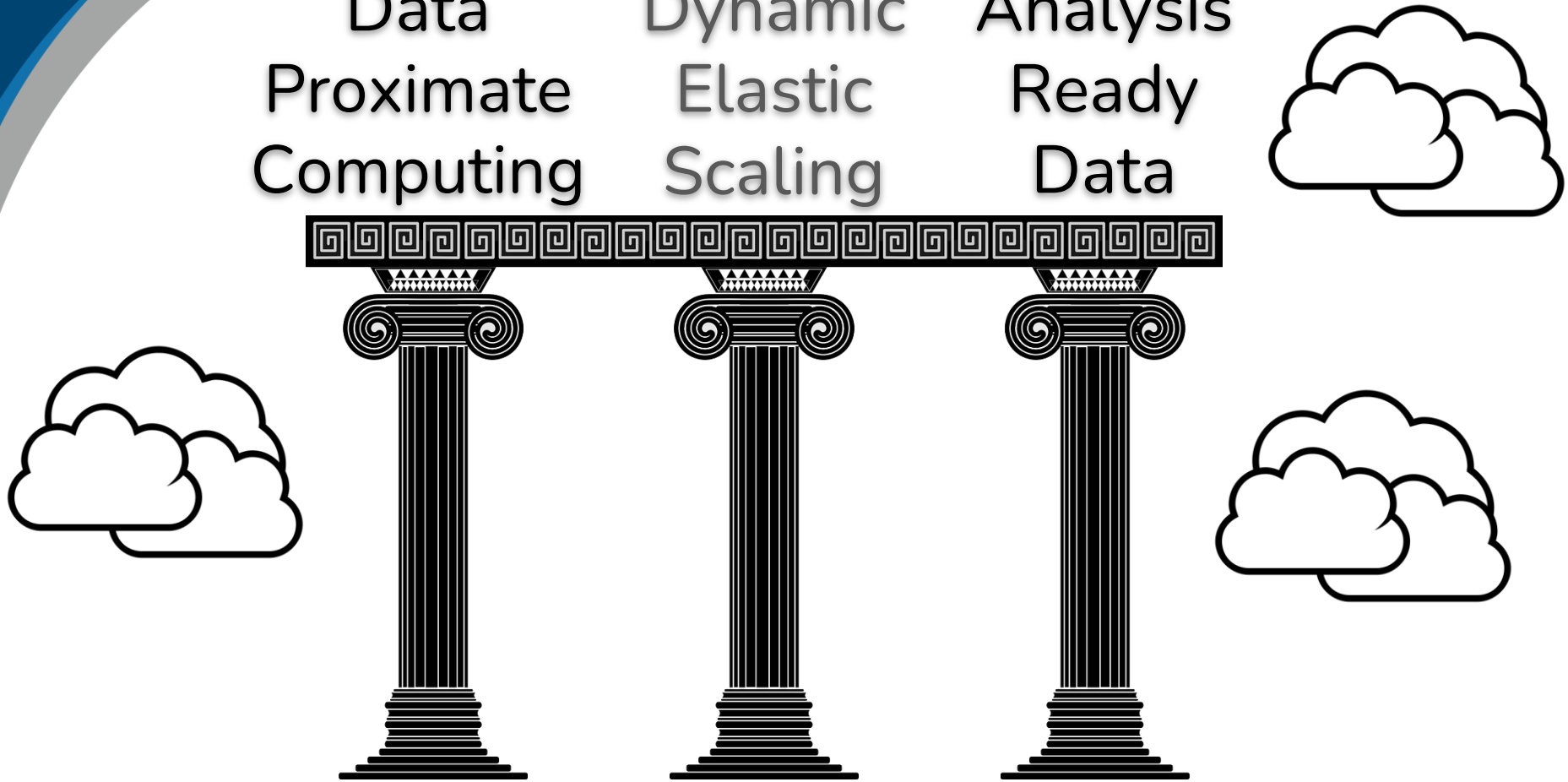


Analysis Ready Data

Why?

- Accessibility
- Preview without downloading
- Cloud Optimized
  - Byte range requests
  - Download only area of interest





https://doi.org/10.1109/MCSE.2021.3059437

Ryan Abernathey's 2022 PANGEO talk at OpenOceanCloud: https://vimeo.com/670782104

Artificial Intelligence & Machine Learning

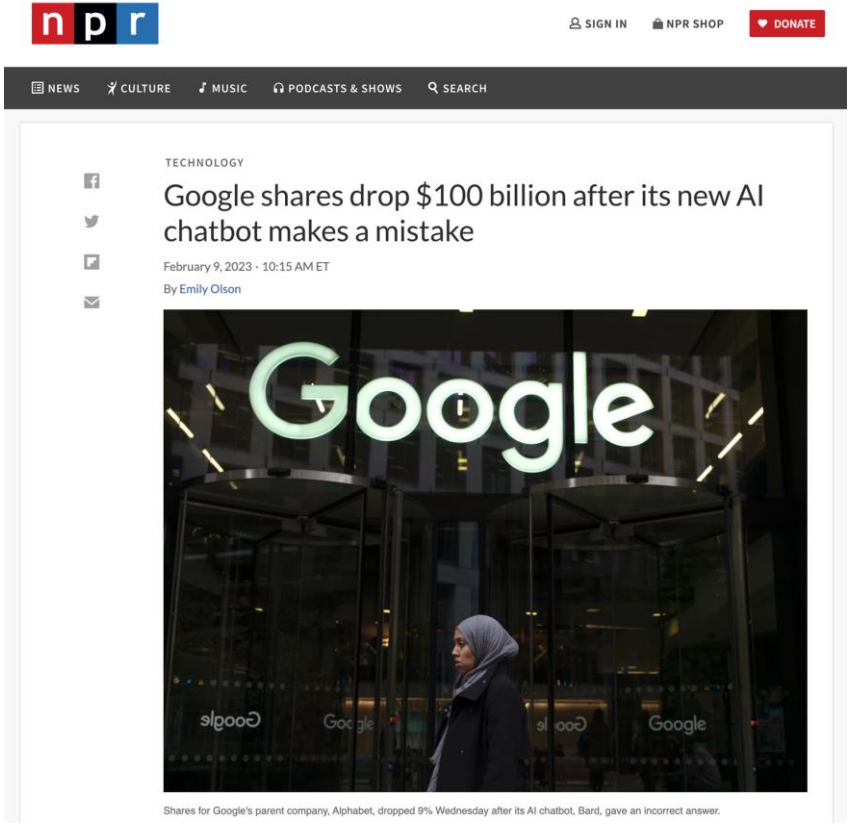
Microsoft

Google Meta AI

amazon web services

AI investments are expected to go from

10 billion $ to >200 billion $ by 2025



npr

SIGN IN NPR SHOP DONATE

NEWS CULTURE MUSIC PODCASTS & SHOWS SEARCH

TECHNOLOGY

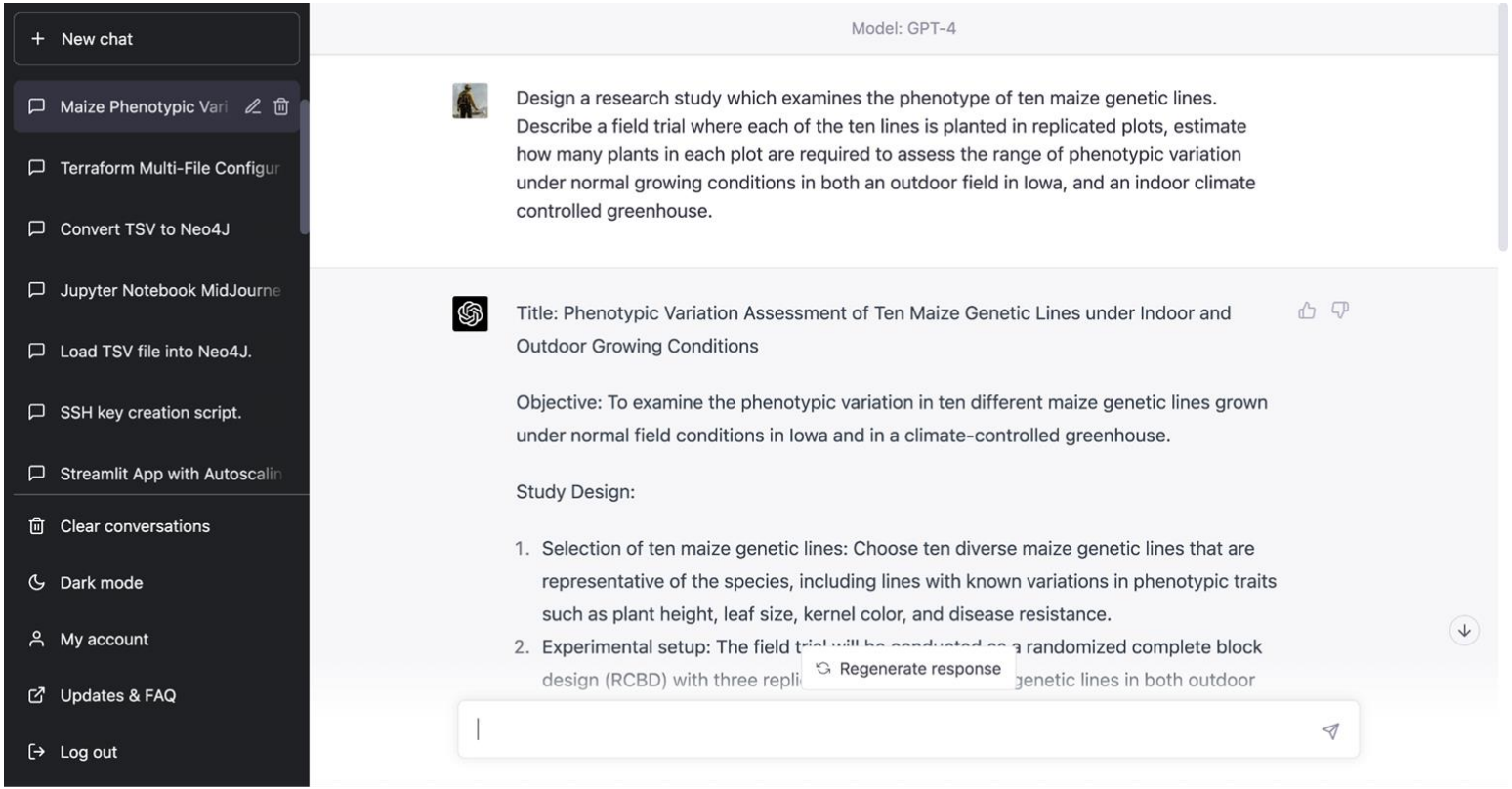
Google shares drop $100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

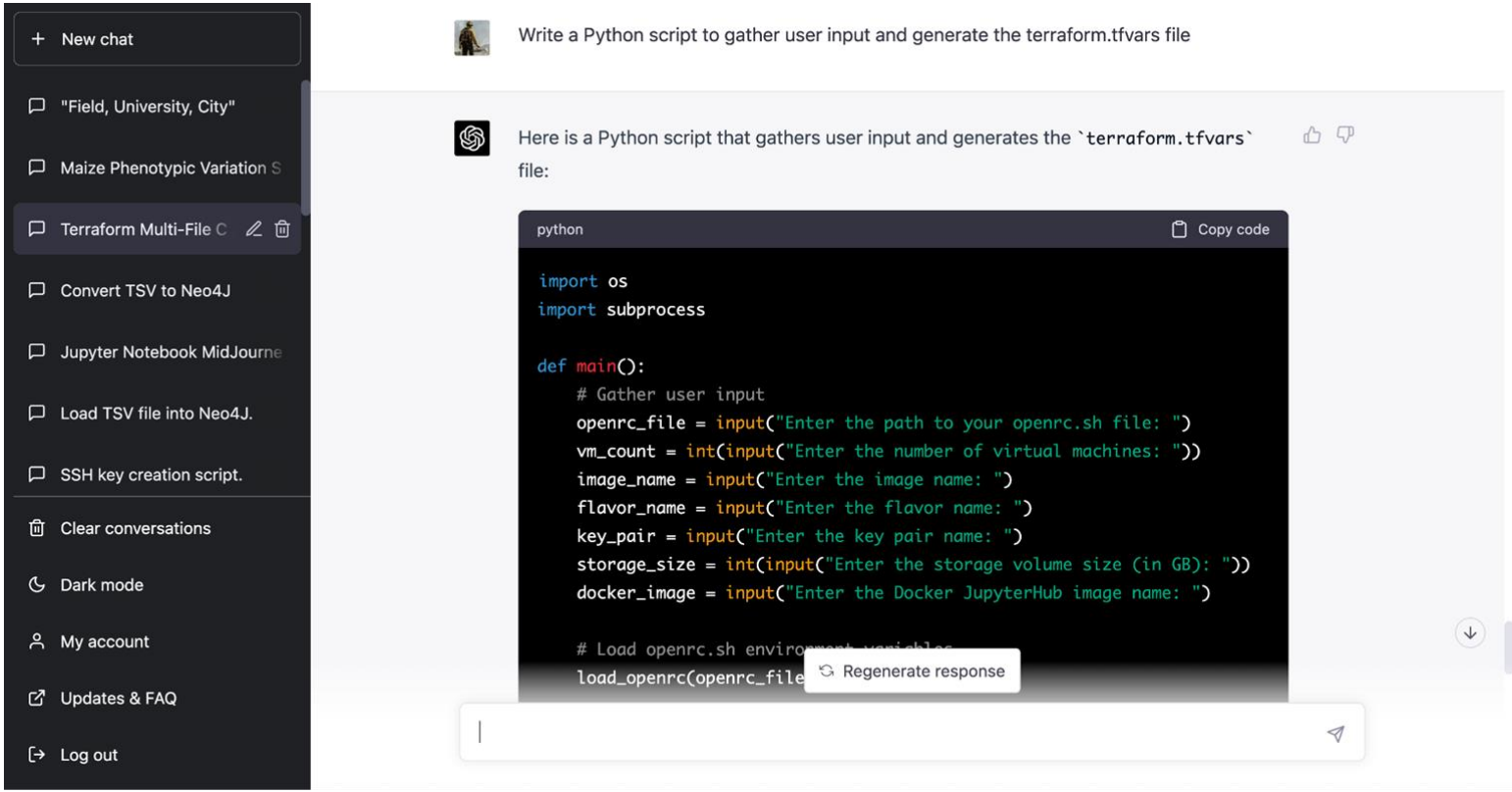
By Emily Olson

Shares for Google's parent company, Alphabet, dropped 9% Wednesday after its AI chatbot, Bard, gave an incorrect answer.

Artificial Intelligence & Machine Learning



Artificial Intelligence & Machine Learning



GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

Tyna Eloundou[1], Sam Manning[1,2], Pamela Mishkin*[1], and Daniel Rock[3]

[1]OpenAI
[2]OpenResearch
[3]University of Pennsylvania

March 20, 2023

Abstract

We investigate the potential implications of Generative Pre-trained Transformer (GPT) models and related technologies on the U.S. labor market. Using a new rubric, we assess occupations based on their correspondence with GPT capabilities, incorporating both human expertise and classifications from GPT-4. Our findings indicate that approximately 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of GPTs, while around 19% of workers may see at least 50% of their tasks impacted. The influence spans all wage levels, with higher-income jobs potentially facing greater exposure. Notably, the impact is not limited to industries with higher recent productivity growth. We conclude that Generative Pre-trained Transformers exhibit characteristics of general-purpose technologies (GPTs), suggesting that as these models could have notable economic, social, and policy implications.

https://arxiv.org/abs/2303.10130

Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence

Shakked Noy MIT
Whitney Zhang MIT

March 2, 2023
Working Paper (not peer reviewed)

Abstract

We examine the productivity effects of a generative artificial intelligence technology—the assistive chatbot ChatGPT—in the context of mid-level professional writing tasks. In a preregistered online experiment, we assign occupation-specific, incentivized writing tasks to 444 college-educated professionals, and randomly expose half of them to ChatGPT. Our results show that ChatGPT substantially raises average productivity: time taken decreases by 0.8 SDs and output quality rises by 0.4 SDs. Inequality between workers decreases, as ChatGPT compresses the productivity distribution by benefiting low-ability workers more. ChatGPT mostly substitutes for worker effort rather than complementing worker skills, and restructures tasks towards idea-generation and editing and away from rough-drafting. Exposure to ChatGPT increases job satisfaction and self-efficacy and heightens both concern and excitement about automation technologies.

https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf

The Impact of AI on Developer Productivity: Evidence from GitHub Copilot

Sida Peng,[1*] Eirini Kalliamvakou,[2] Peter Cihon,[2] Mert Demirer[3]

[1]Microsoft Research, 14820 NE 36th St, Redmond, USA
[2]GitHub Inc., 88 Colin P Kelly Jr St, San Francisco, USA
[3]MIT Sloan School of Management, 100 Main Street Cambridge, USA

*To whom correspondence should be addressed; E-mail: sidpeng@microsoft.com.

Abstract

Generative AI tools hold promise to increase human productivity. This paper presents results from a controlled experiment with GitHub Copilot, an AI pair programmer. Recruited software developers were asked to implement an HTTP server in JavaScript as quickly as possible. The treatment group, with access to the AI pair programmer, completed the task 55.8% faster than the control group. Observed heterogenous effects show promise for AI pair programmers to help people transition into software development careers.

https://arxiv.org/abs/2302.06590

Artificial Intelligence & Machine Learning

GPT v4

(Generative Pre-trained Transformer)

LLM

(Large Language Model)

- ➔ Use well defined, progressive, prompts
- ➔ Create or refine code you are struggling with
- ➔ Ask it about a specific piece of code or error you cannot understand
- ➔ For tasks that are otherwise a drudgery & / or subject to human error


CYVERSE®
The Open Science Workspace
DBI-0735191, DBI-1265383, DBI-1743442 and OAC-1664172

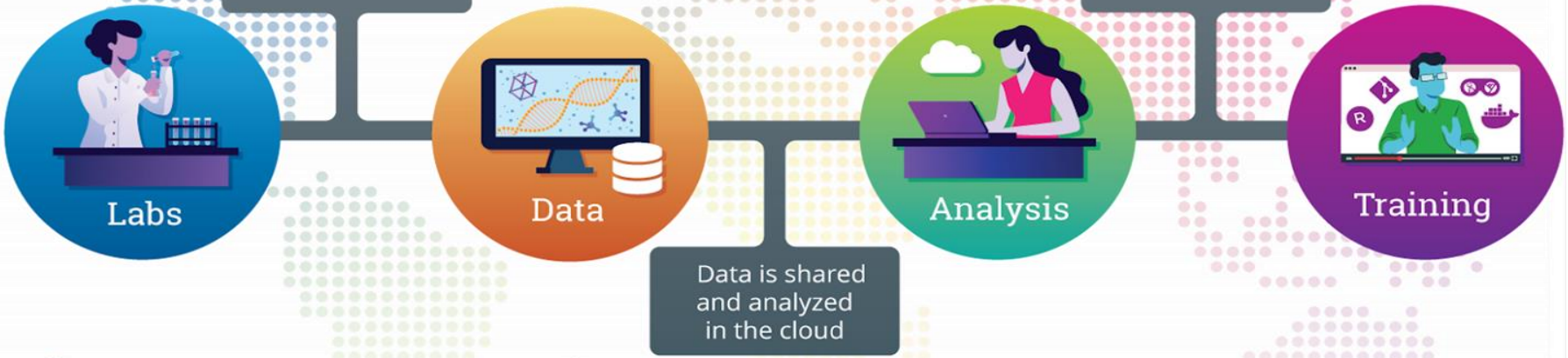
CYVERSE® Public

110,000 users

1500 publications

11 PBs of data

40,000 trained



CYVERSE® Professional

- Federation with local and commercial cloud and high-performance computing
- Integration with local user identity management systems
- Security compliance

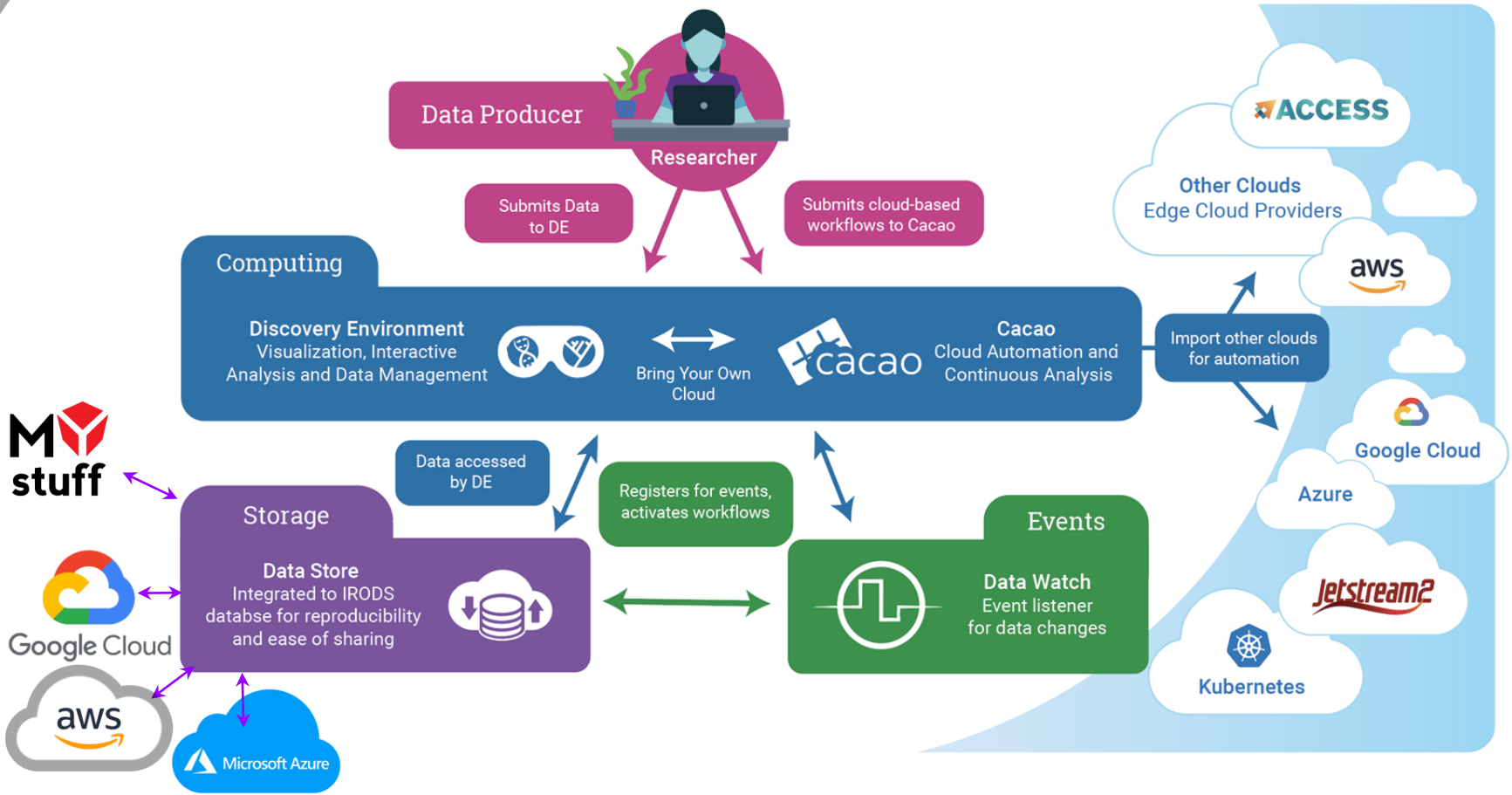
CYVERSE® Health

- Secure Perimeter (VPN)
- HIPAA Compliant Platforms inside of Perimeter
- XNAT
- Training

CYVERSE® Defense

- ITAR Compliance
- Receive data from multiple sensors
- Policy based data visibility and sharing
- Support for multiple teams and data partitioning

How does CyVerse work?



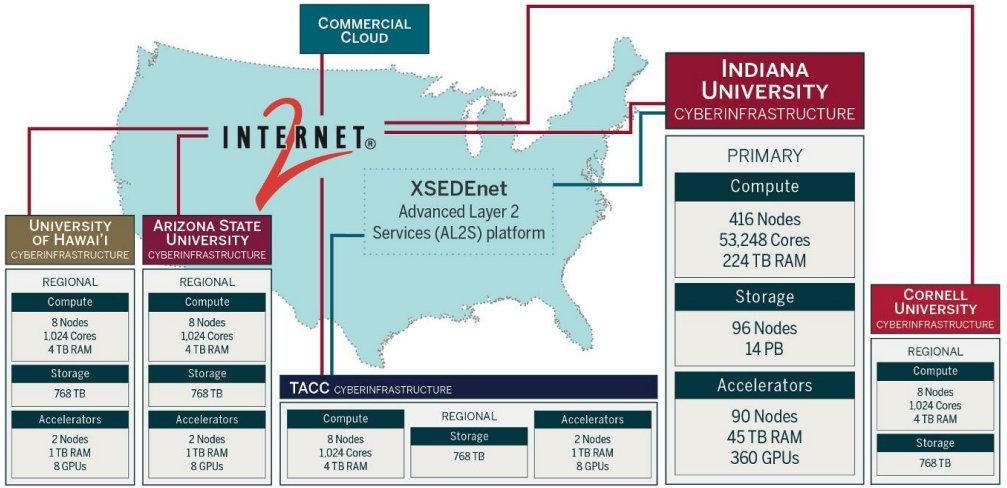
CyVerse Collaborations

ACCESS

Jetstream2

cacao

Cloud Automation & Continuous Analysis Orchestration



NSF

Award # OAC 2005506

CyVerse Collaborations

Organizers

ACCESS

Allocations, resources, projects, ideas

ubuntu

Includes software, environments, data

jupyter

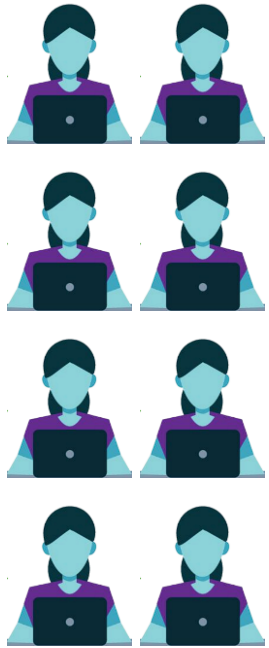
Includes environments, notebooks, data

cacao

Cloud Automation & Continuous Analysis Orchestration

Jetstream2

Template (recipe) selection: number of instances, CPU/ RAM/ Disk, security



Learners

Based on Skidmore, CyVerse Tech Preview Webinar, 2022

CyVerse Use Case # 1

Transforming a Quarter Petabyte of Field Phenomics Data Into Functional Traits

Emmanuel Gonzalez

emmanuelgonzalez@arizona.edu

Duke Pauli Lab
Ph.D. Candidate
University of Arizona

Pauli Lab

Collecting data is becoming easier and cheaper

Robots

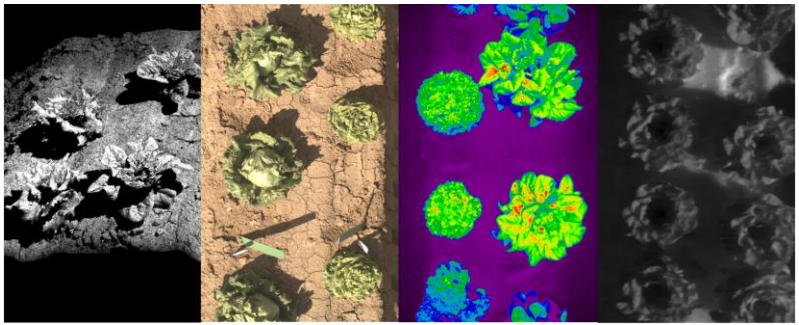
Carts

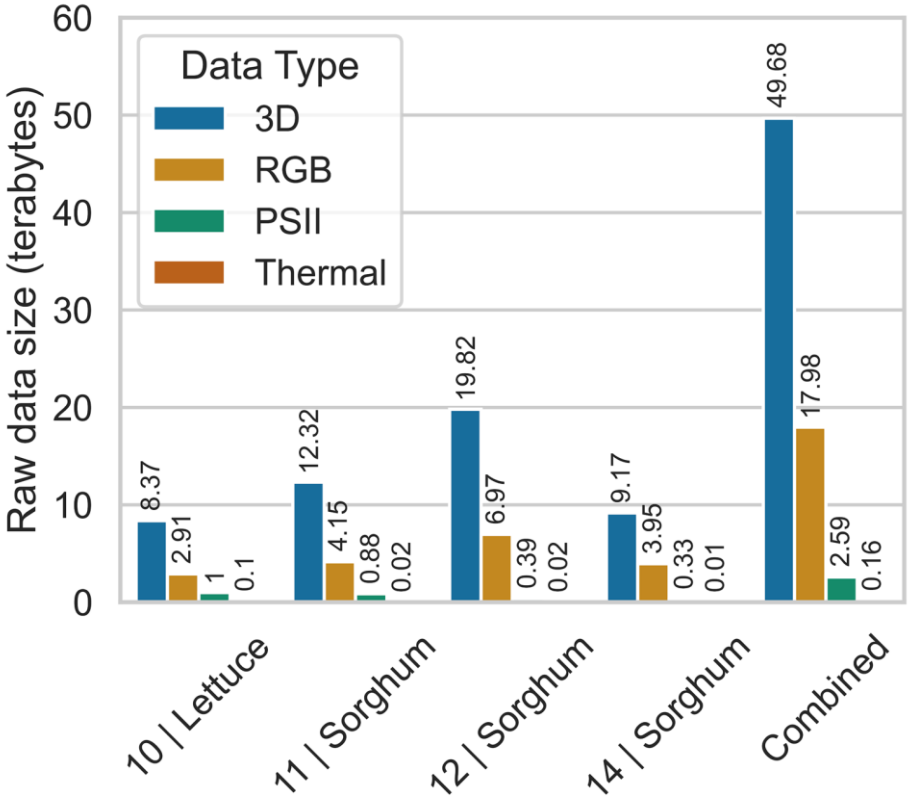
Drones

Phones

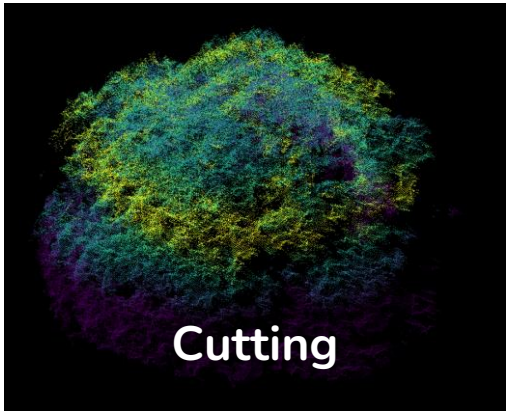
Image: Wall Street Journal, USDA, DJI, and IITA

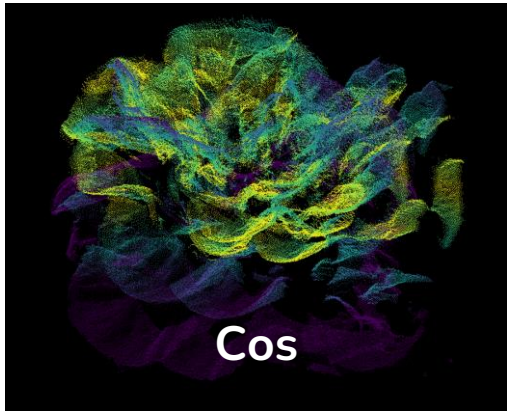
World's Largest Field Robot

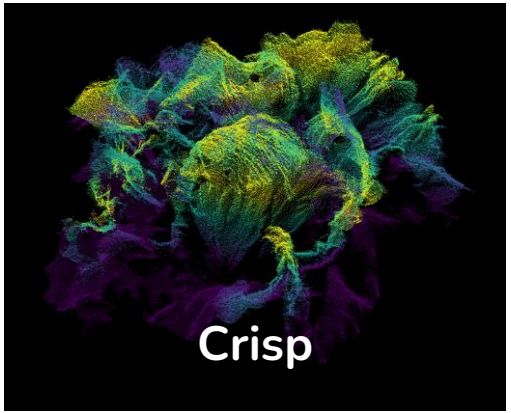


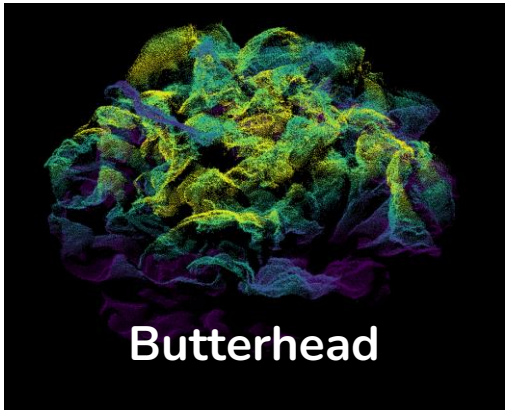


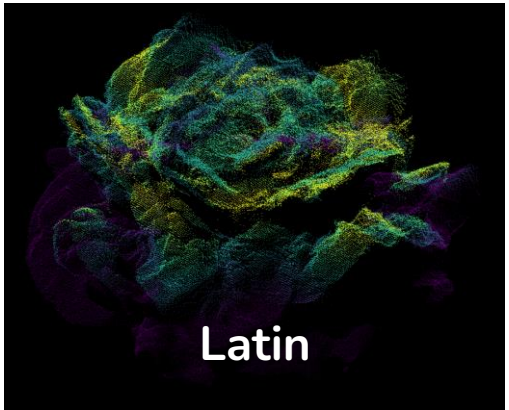
Lettuce exhibits large amounts of phenotypic variation











Goal: repeatable, fine-scale phenotype extraction

3D

20,000 plants * 32 scans * 4 phenotypes

= 2.6M individual plant phenotypes

RGB

20,000 plants * 36 scans * 1 phenotype

= 720K individual plant phenotypes

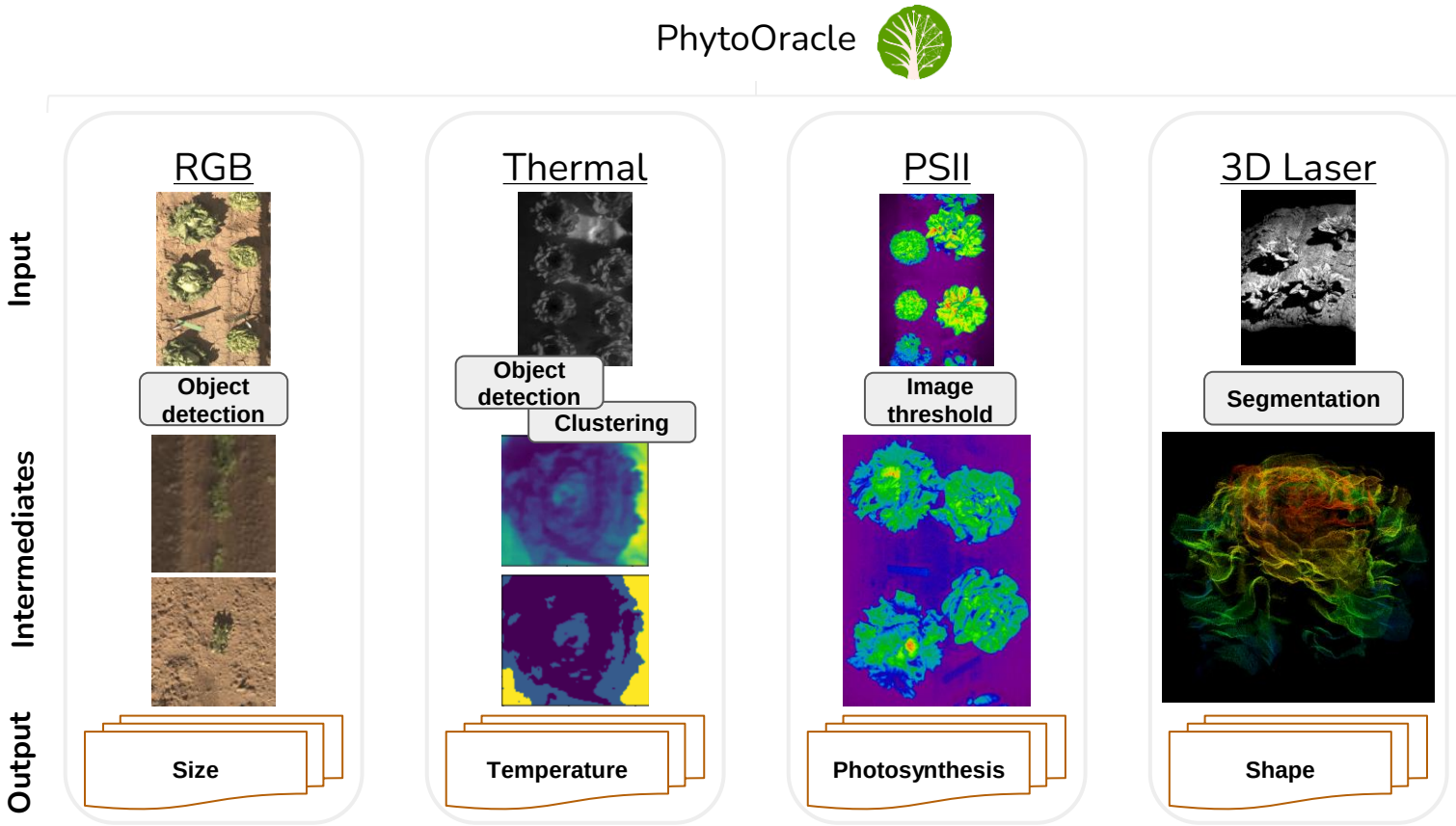
Thermal

20,000 plants * 36 scans * 4 phenotypes

= 2.9M individual plant phenotypes

Total: 6.2M time-series individual plant phenotypes

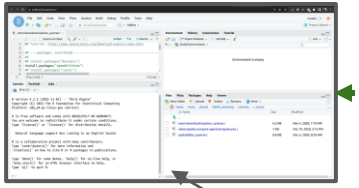
Requires modular & scalable workflows



Data transfer and computation

Algorithms

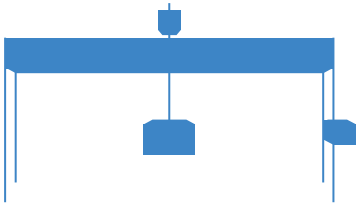
- Prototyping
- Development
- Testing
- Collaboration/Publication

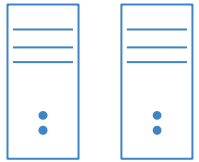


Collaborators and Public

iRODS

Makeflow





Cache Server

UA HPC

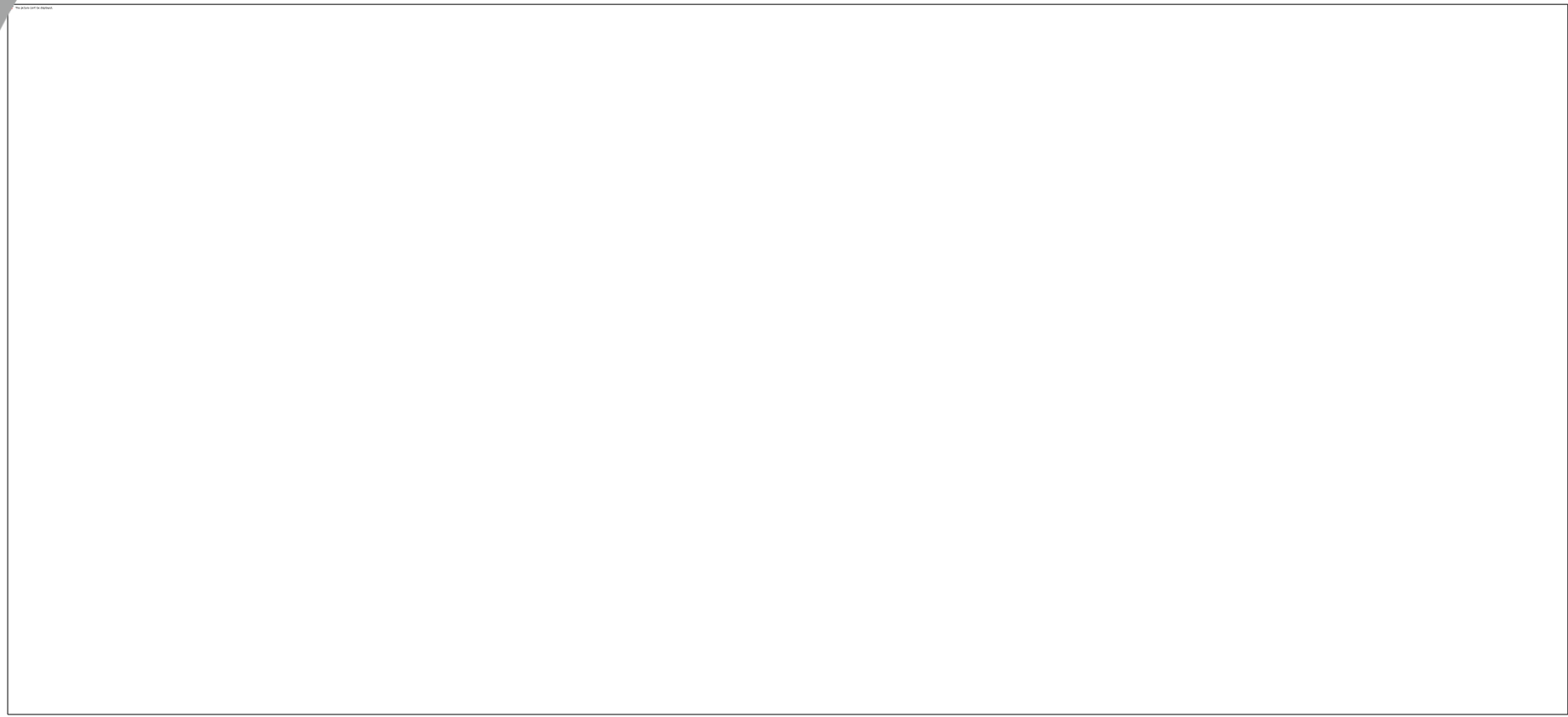
- Collect data
- 1-10TB/day

- Compression
- Checksums

- Storage
  (raw + processed)

- HPC
- Workflows

Data transfer and computation



METHODS article

Front. Plant Sci., 06 March 2023
Sec. Technical Advances in Plant Science
Volume 14 - 2023 |
https://doi.org/10.3389/fpls.2023.1112973

This article is part of the Research Topic
High-Throughput Field Phenotyping to Advance Precision Agriculture and Enhance Genetic Gain, Volume II
View all 6 Articles >

PhytoOracle: Scalable, modular phenomics data processing pipelines

Emmanuel M. Gonzalez[1], Ariyan Zarei[2], Nathanial Hendler[1], Travis Simmons[1], Arman Zarei[3], Jeffrey Demieville[1], Robert Strand[1], Bruno Rozzi[1], Sebastian Calleja[1], Holly Ellingson[4], Michele Cosi[1,5], Sean Davey[6], Dean O. Lavelle[7], Maria José Truco[7], Tyson L. Swetnam[5,8], Nirav Merchant[4,5], Richard W. Michelmore[7,9], Eric Lyons[1,4,5] and Duke Pauli[1,4*]

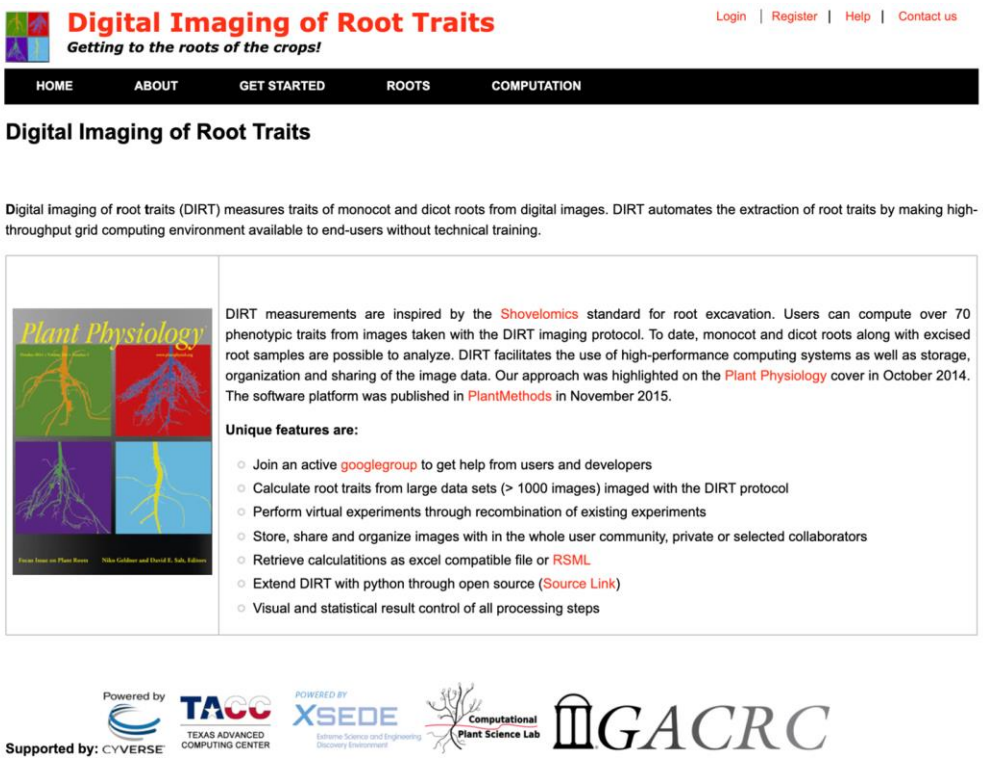
[1] School of Plant Sciences, University of Arizona, Tucson, AZ, United States
[2] Department of Computer Science, University of Arizona, Tucson, AZ, United States
[3] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[4] Data Science Institute, University of Arizona, Tucson, AZ, United States
[5] BIO5 Institute, University of Arizona, Tucson, AZ, United States
[6] Department of Cellular and Molecular Medicine, University of Arizona, Tucson, AZ, United States
[7] The Genome and Biomedical Sciences Facility, University of California, Davis, Davis, CA, United States
[8] School of Natural Resources and the Environment, University of Arizona, Tucson, AZ, United States
[9] Department of Plant Sciences, University of California, Davis, Davis, CA, United States

https://doi.org/10.3389/fpls.2023.1112973

CyVerse Use Case # 2

Alex Bucksch

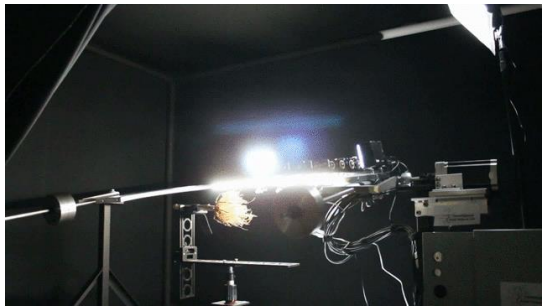
https://www.computational-plant-science.org



Das, A., Schneider, H., Burridge, J. et al. Digital imaging of root traits 2015 (DIRT): a high-throughput computing and collaboration platform for field-based root phenomics. Plant Methods 11, 51 https://doi.org/10.1186/s13007-015-0093-3

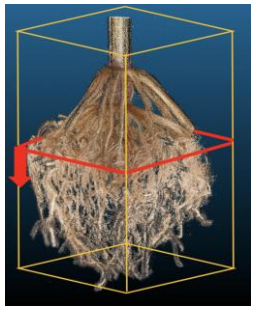
DIRT/3D measures root systems from the field

I. Maize roots in the field

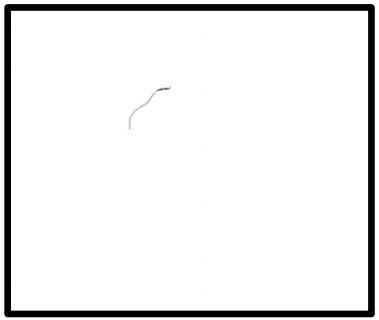


II. 3D root scanner

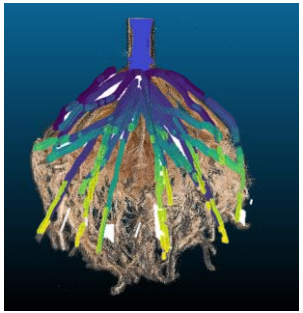
III. Root images data



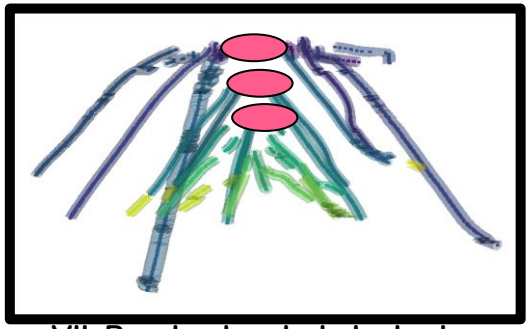
IV. 3D root model



V. Individual root tracking



VI. Computed root structure



VII. Resolved occluded whorl locations

Suxing Liu

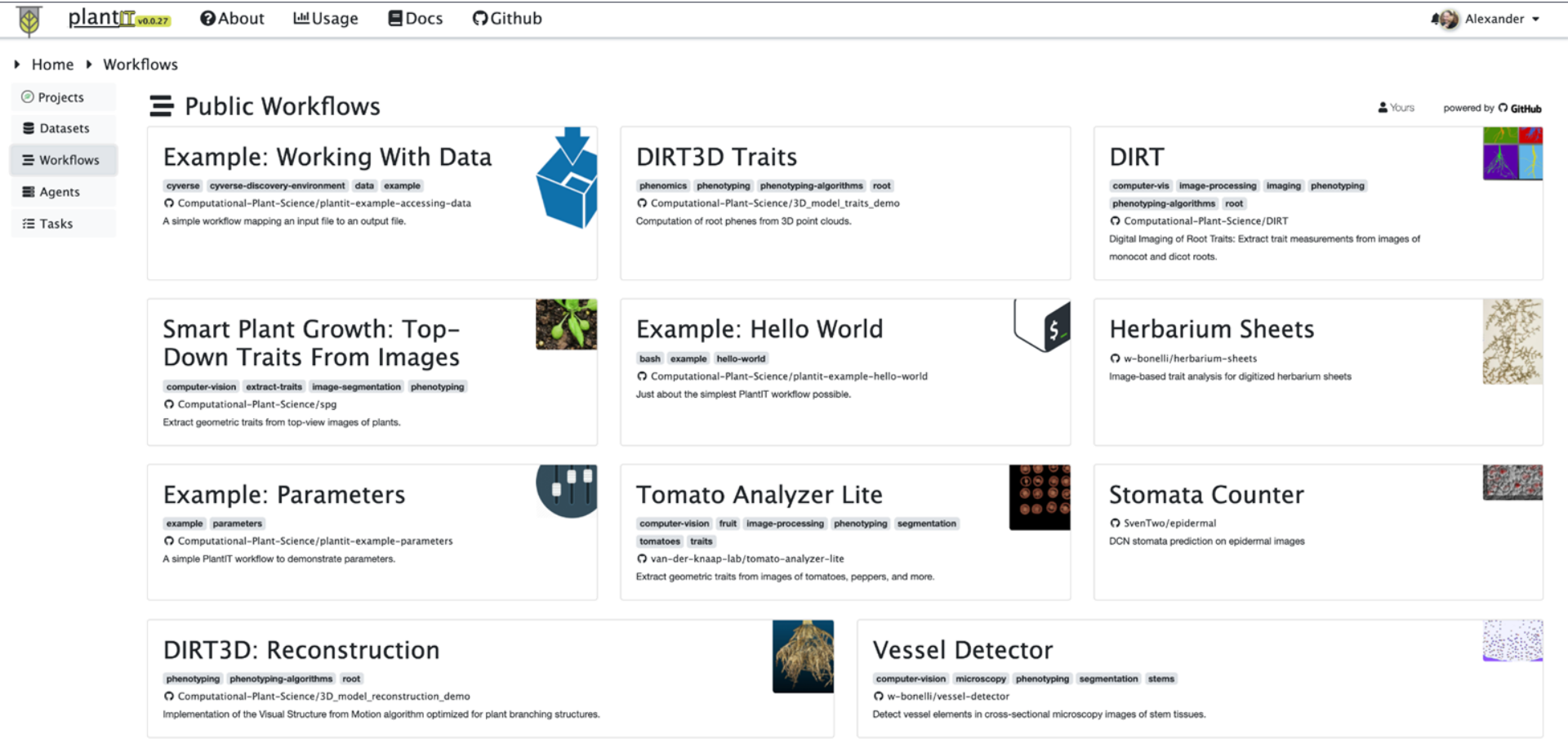
DIRT/3D measures root systems from the field

Dense data sets are giant data sets

1 root = 1 GiB

1 experiment = 1,000 roots

Liu, S. et al. 2021 DIRT/3D: 3D root phenotyping for field-grown maize mays. Plant Physiology 187:2 https://doi.org/10.1093/plphys/kiab311



https://plantit.cyverse.org/

THE UNIVERSITY OF ARIZONA

Three PostDoc positions in root phenomics

Algorithms - Experiments - Plants

Apply: www.bit.ly/RootPhenotyping

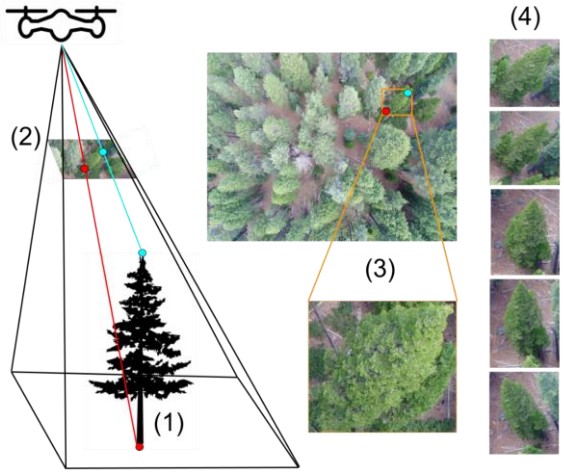
- Discover unknown phenotypes with computers
- Address climate change in drought & stress enviroments
- Share open-source tools with the plant community
- Characterize function in real agricultural settings
- You have worked with plants and trait data
- You know python, imaging, and experimental design

Shape the future of plant phenotyping & receive benefits and $60.000+ (dependent on experience)

More info: www.computational-plant-science.org
Questions: bucksch@arizona.edu

CyVerse Use Case # 3

Open Forest Observatory



Michael J. Koontz
mikoontz@gmail.com
University of Colorado Boulder
Earth Lab

UC DAVIS
UNIVERSITY OF CALIFORNIA

Derek J. Young
djyoung@ucdavis.edu

Award # DBI 2152671
Award # DBI 2152672
Award # DBI 2152673

Jeffrey K. Gillan
jgillan@arizona.edu

Collection Pipeline

RTK Collection

Jpeg drone images (TBs)

CYVERSE Data Store

raw imagery

Imagery products

Jetstream2

geojson.io

COG CLOUD OPTIMIZED GEOTIFF

COPC.io Cloud Optimized Point Cloud

Zarr

Art by MidJourney AI

Analysis Pipeline

Jetstream2

Photogrammetry

Worker VMs

Server VM

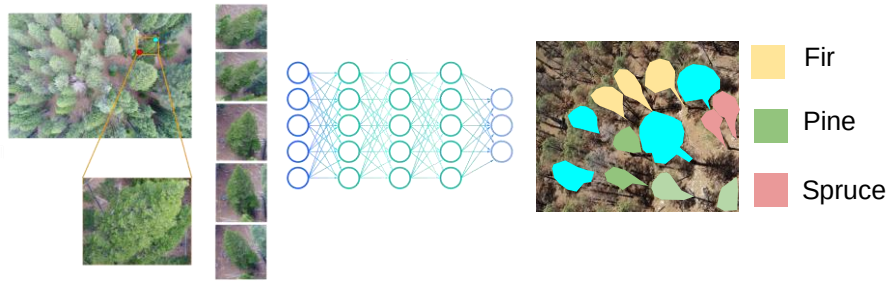
Metashape

Metashape

Metashape

Metashape

Individual Tree Detection
ML Species Classification



https://github.com/open-forest-observatory/automate-metashape

Art by MidJourney AI

Cloud Optimized Formats

Vectors

geojson.io

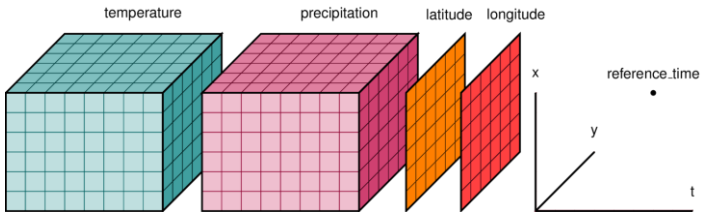
Rasters

STAC
SpatioTemporal
Asset Catalog

COG
CLOUD OPTIMIZED
GEOTIFF

COPC.io
Cloud Optimized Point Cloud

Zarr



Conclusion 😱

The Big Data Landscape - 😵 not an exaggeration

2023 Year of Open Science - ⏳ 12/2025 🔒

Becoming Cloud-Native - 🌥️ not an option

AI in the workplace - 🤖 already changing the world

DBI-0735191, DBI-1265383, DBI-1743442 and OAC-1664172


Thanks for your time!
Art by MidJourney AI