A National Discovery Cloud

https://cra.org/ccc/wp-content/uploads/sites/2/2021/04/CCC-Whitepaper-National-Discovery-Cloud-2021.pdf

CCC

Computing Community Consortium
Catalyst

Ian Foster

The University of Chicago

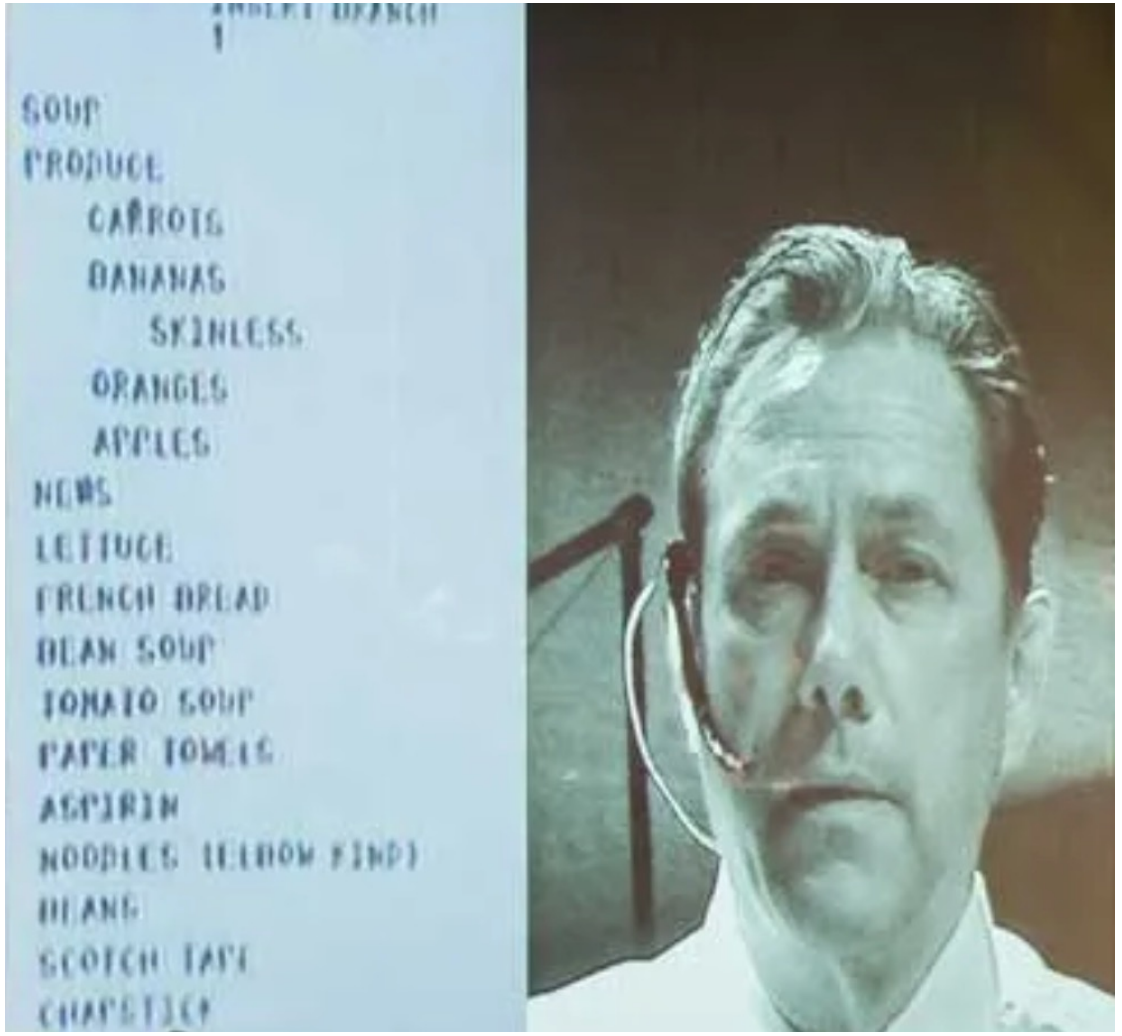
Argonne National Laboratory

Crescat scientia; vita excolatur

globus labs

Tools for augmenting human intellect: 1962

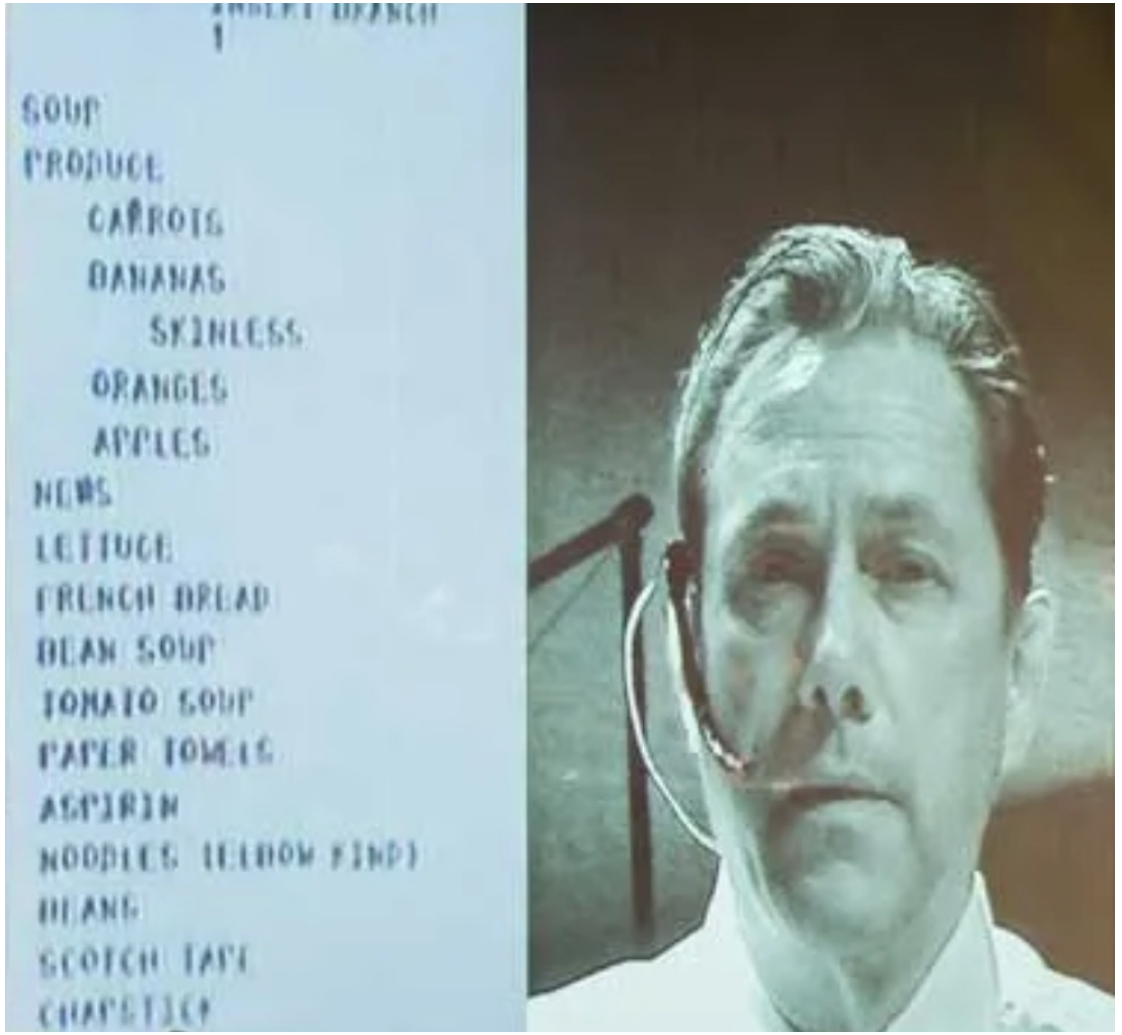
"By 'augmenting human intellect' we mean increasing the capability of a [person] to approach a complex problem situation, to gain comprehension to suit [their] particular needs, and to derive solutions to problems." *



* Doug Engelbart, 1962 -- https://www.dougengelbart.org/content/view/138/

Tools for augmenting human intellect: 1962

"By 'augmenting human intellect' we mean increasing the capability of a [person] to approach a complex problem situation, to gain comprehension to suit [their] particular needs, and to derive solutions to problems." *



"I don't get it - everything you've shown me today I can do on my ASR-33." – prominent prof, as reported by Andries Van Dam +

* Doug Engelbart, 1962 -- https://www.dougengelbart.org/content/view/138/

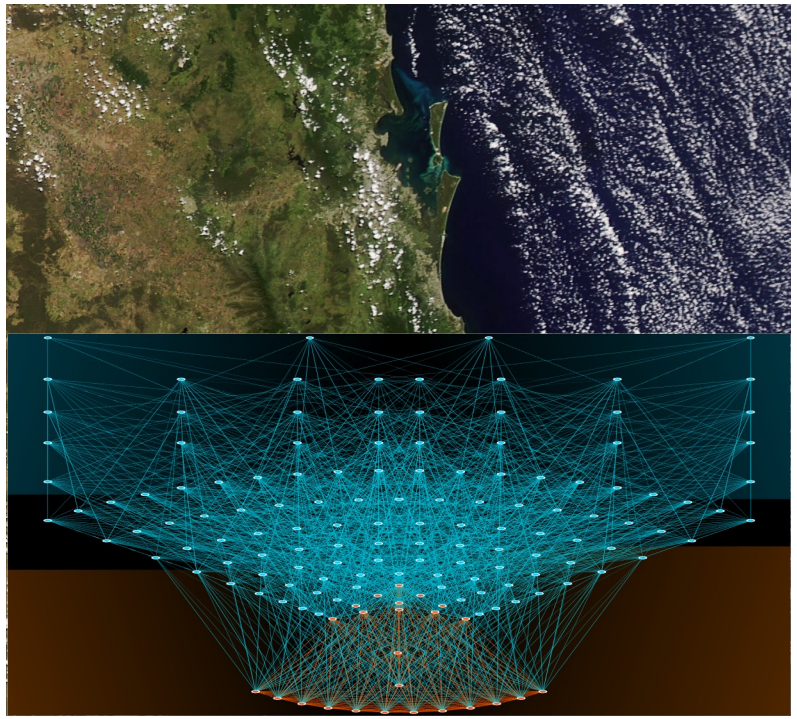
+ https://www.theregister.com/2008/12/11/engelbart_celebration/

2022: Three transformative technologies

Public cloud

A new computing platform enabling new approaches to building, delivering services

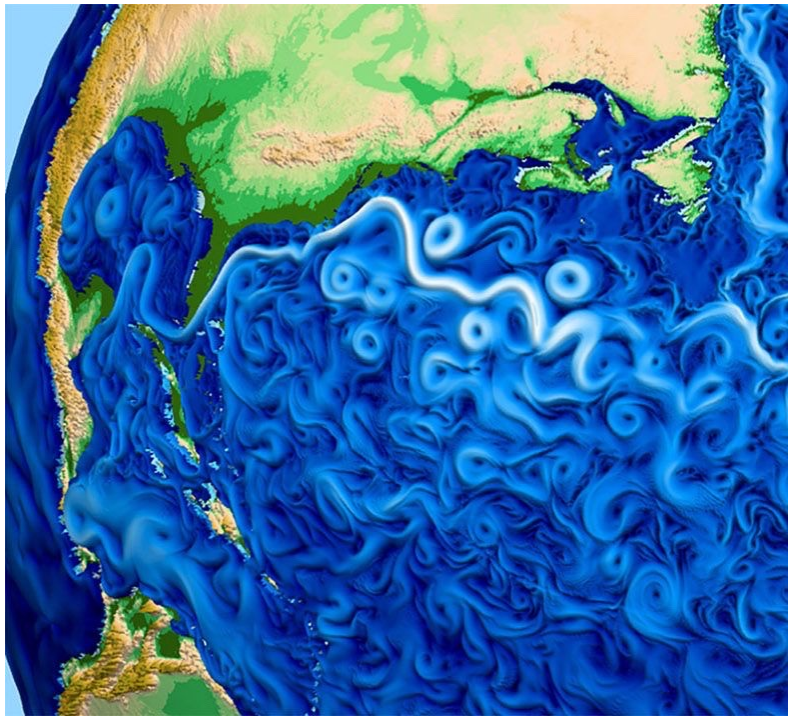
X



Sensors, data, ML

Powerful methods for generating, and extracting information from, huge data

X



Numerical simulation

A scientific method on par with, and sometimes exceeding, experiment

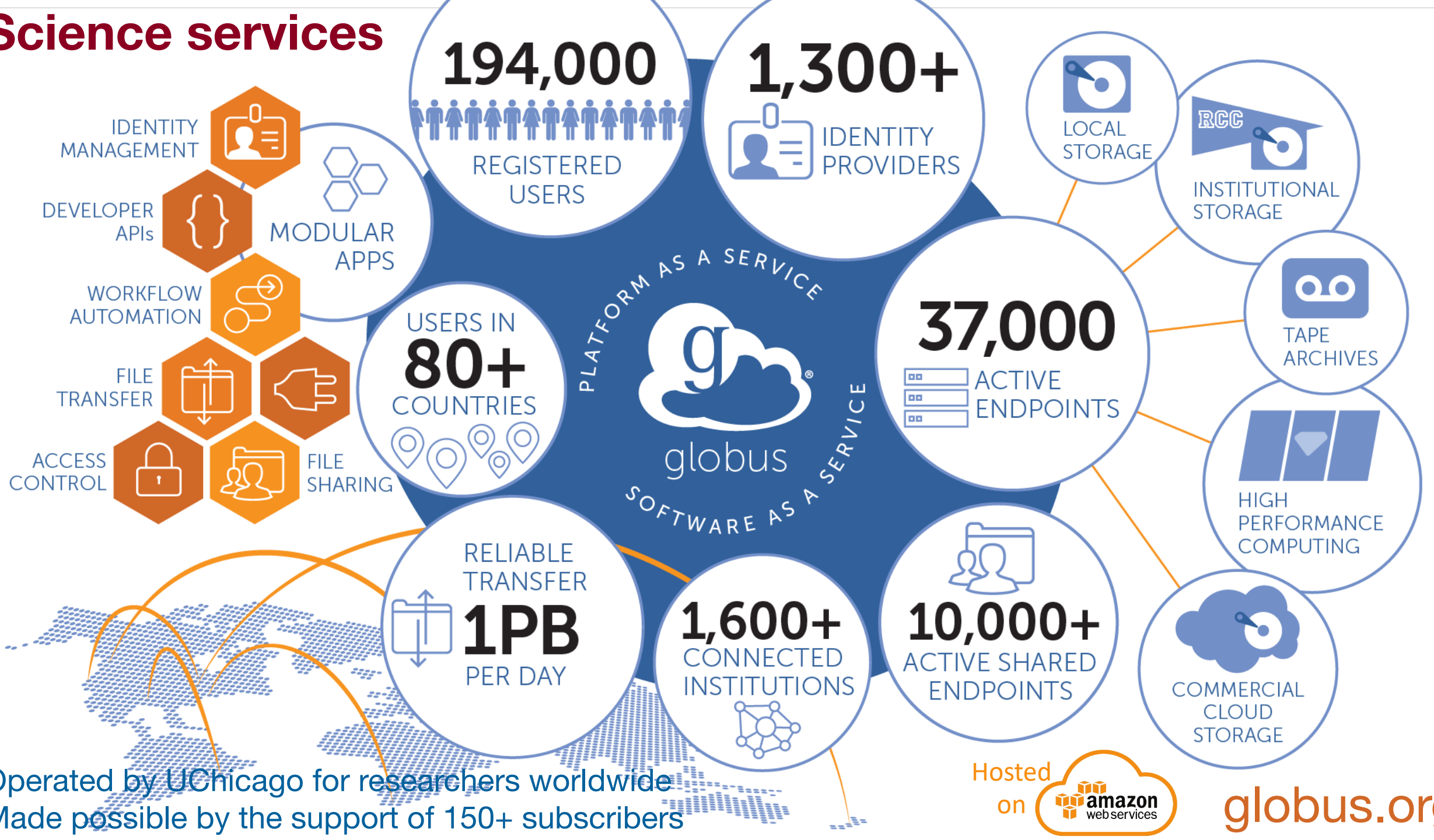
Sources: servecentric.com, visibleearth.nasa.gov, quantamagazine.org, e3sm.org

Challenge and opportunity in 2022: Create new tools for augmenting human intellect

- Vast **curated collections** of observational, experimental, and simulated data, plus associated **ML models**
- A **global knowledge graph** linking publications, data, models, and more—constantly updated by computational agents
- **Digital twins** of complex physical, biological, & social systems, running on powerful computers, plus constantly updated **ML surrogates**
- Rich set of **science services**, with infrastructure to simplify operations and incentives to sustain operations

See DOI: 10.1126/science.1110411, 2005, substituting “cloud” for “grid”

Science services



Operated by UChicago for researchers worldwide
Made possible by the support of 150+ subscribers

Hosted on amazon web services

globus.org

MG-RAST

metagenomics analysis server

cite us

version 4.0.3

486,803 metagenomes containing 2,109 billion sequences and

311.88 Tbp processed for 36,497 registered users.

for programmatic access visit our API site

We added some additional resources to process your inbox jobs.

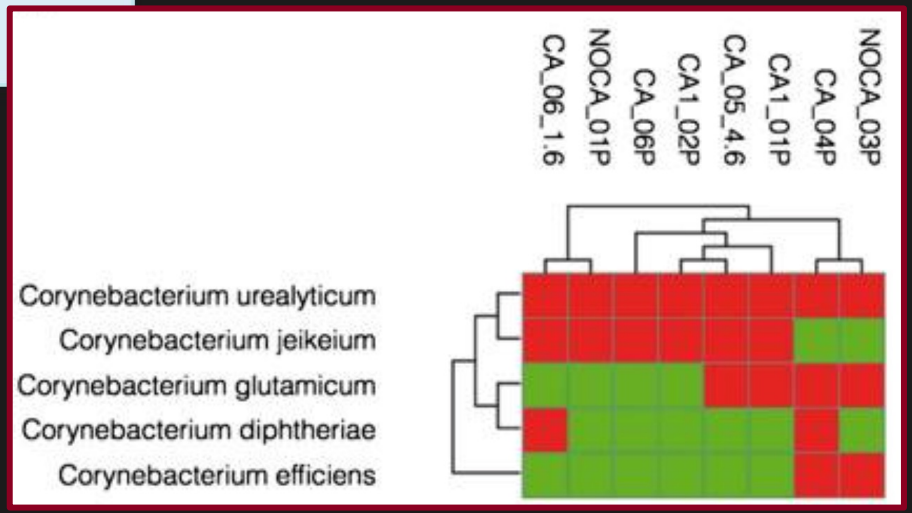
search string e.g. mgp128 or mgm4447970.3 search

upload download analyze

https://www.mg-rast.org

doi: 10.1093/bib/bbx105

Heatmap and clustering of the occurrence of *Corynebacteria* in study mgp128



A National Discovery Cloud requires new capabilities

- The definition, creation, and curation of large **reference datasets** to fuel new data-driven models of the natural world, economy, human physiology, healthcare system, manufacturing processes, etc.
- A **discovery cloud platform** to enable the collaborative development of value-added services that support NDC-powered scholarship and education
- New **educational programs and curricula** to prepare a generation for whom programming and using NDC capabilities is second nature
- Substantial **computing, storage, and network resources** to host and compute over enormous datasets, and to host and operate discovery cloud services that enhance the value of datasets
- Innovative integrations of NDC capabilities with high-performance computers, automated laboratories, and other elements of a 21st century **discovery and innovation ecosystem**
- **Privacy and security** designed in from the beginning, rather than added post facto, and with integrated assurances and audit capabilities so that the NDC advances rather than hinders computing in the public interest

Open issues and challenges include …

- Weaving diverse capabilities just listed into a **coherent whole** that US R&D enterprise can harness for discovery, innovation, and workforce
- Balancing needs for **persistent resources** to support R&E communities vs. supporting **innovation** by those communities
- Enabling research at **lower levels of the 'stack'** (Touch's Law: *The lowest level at which research is permitted in a testbed is also the highest level at which it can occur*)
- **Privacy and security**: Balancing "free and open" vs. "private and secure" in data and services
- Building an NDC that contributes to **environmental sustainability**
- Appropriate balance between bespoke and private sector data centers

Summary: Let’s not underestimate public cloud

- An **elastic** source of computing and storage capacity – **sure**
- A **cheap** source of computing and storage capacity – **maybe/not?**
- A **new technology** to study and engineer – **yes**

- An immensely powerful platform for delivering **scalable, reliable, and democratizing digital services** – **absolutely**!

- Our opportunity and challenge is a **top-to-bottom rethink** of what computing means for research and education