



NSF HDR Summary



[OAC-2117997](#)

Philip Harris (MIT)
A3D3 Deputy Director

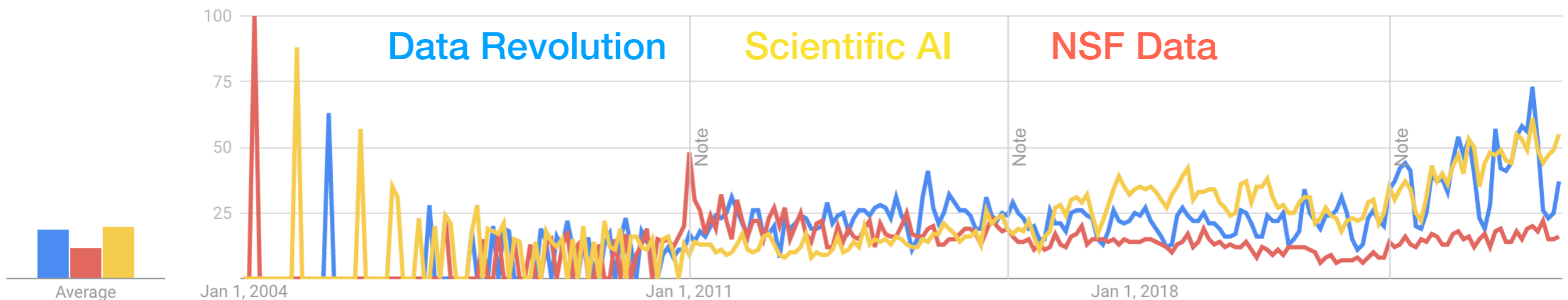


<https://a3d3.ai/>

Harnessing the Data² Revolution

- What is the data revolution?

Interest over time 



The data revolution is a word that is growing with time like AI

I think we all feel it, but here we aim to exploit it

Harnessing the Data³ Revolution

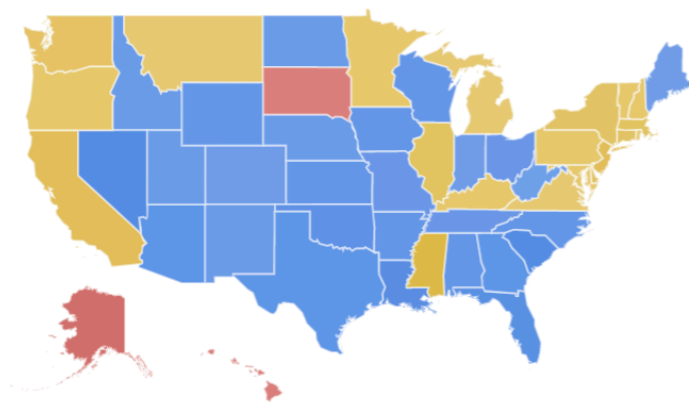
- What is the data revolution?

Compared breakdown by subregion

Subregion ▾



● data revolution ● NSF data ● science artificial intelligence



Sort: Interest for data revolution ▾

1 Nevada



2 South Carolina



3 Louisiana



4 Florida



5 Oklahoma



How has our program evolved?

- We started with a meet and greet meeting in DC
 - PIs focused on trying get their institute grounding

HDR² From Harnessing the Data Revolution to Harvesting the Data Revolution

26 Oct 2022, 10:00 → 27 Oct 2022, 15:00 US/Eastern

Philip Coleman Harris (Massachusetts Inst. of Technology (US)), Shih-Chieh Hsu (University of Washington Seattle (US))

Description With the Data Revolution underway, scientists are quickly harnessing it to lead to greater scientific output. This is no clearer than under [the NSF Harnessing the Data Revolution \(HDR\)](#) Initiative. In this first of five upcoming HDR Principal Investigator meetings, we assemble members of the NSF HDR to talk about how we can work together and eventually go **from Harnessing the Data Revolution to Harvesting the Data Revolution** or HDR².



HDR²

From Harnessing to Harvesting the Data Revolution

How has our program evolved?

- We followed with a collaborative meeting in Denver (2023)
 - Focus was on building cross-HDR collaborations



[Home](#) [People](#) [Research](#) [Papers and Codes](#)

[Conference Home](#) [Keynote Speakers](#) [Agenda](#) [Location & Travel](#) [Posters](#) [Code of Conduct](#) [Conference Summary](#)

NSF Harnessing the Data Revolution (HDR) 2023 Ecosystem Conference
Unites Data-intensive Research Community



Harvesting the Data Revolution

- Now everything starts to come together



Our Posters start to look cooler and cooler thanks to AI

How are universities dealing with this?

THE GRAINGER COLLEGE OF ENGINEERING

Siebel School of Computing and Data Science

computing and data science ecosystem

1

Computer Science

Integration of computing across the UIUC campus

CS+X

- Computer Science + Advertising
- Computer Science + Animal Sciences
- Computer Science + Anthropology
- Computer Science + Astronomy
- Computer Science + Chemistry
- Computer Science + Crop Sciences
- Computer Science + Economics
- Computer Science + Education
- Computer Science + Geography & Geographic IS
- Computer Science + Linguistics
- Computer Science + Mathematics
- Computer Science + Music
- Computer Science + Philosophy
- Computer Science + Statistics

2

Data Science

Integration of data science (w. LAS and iSchool) across the UIUC campus

CS, Math, Stat & iSchool

X+DS

- Accountancy + Data Science
- Astronomy + Data Science
- Business + Data Science
- Finance + Data Science
- Information Sciences + Data Science

3

Integration of computing and data science across The Grainger College of Engineering

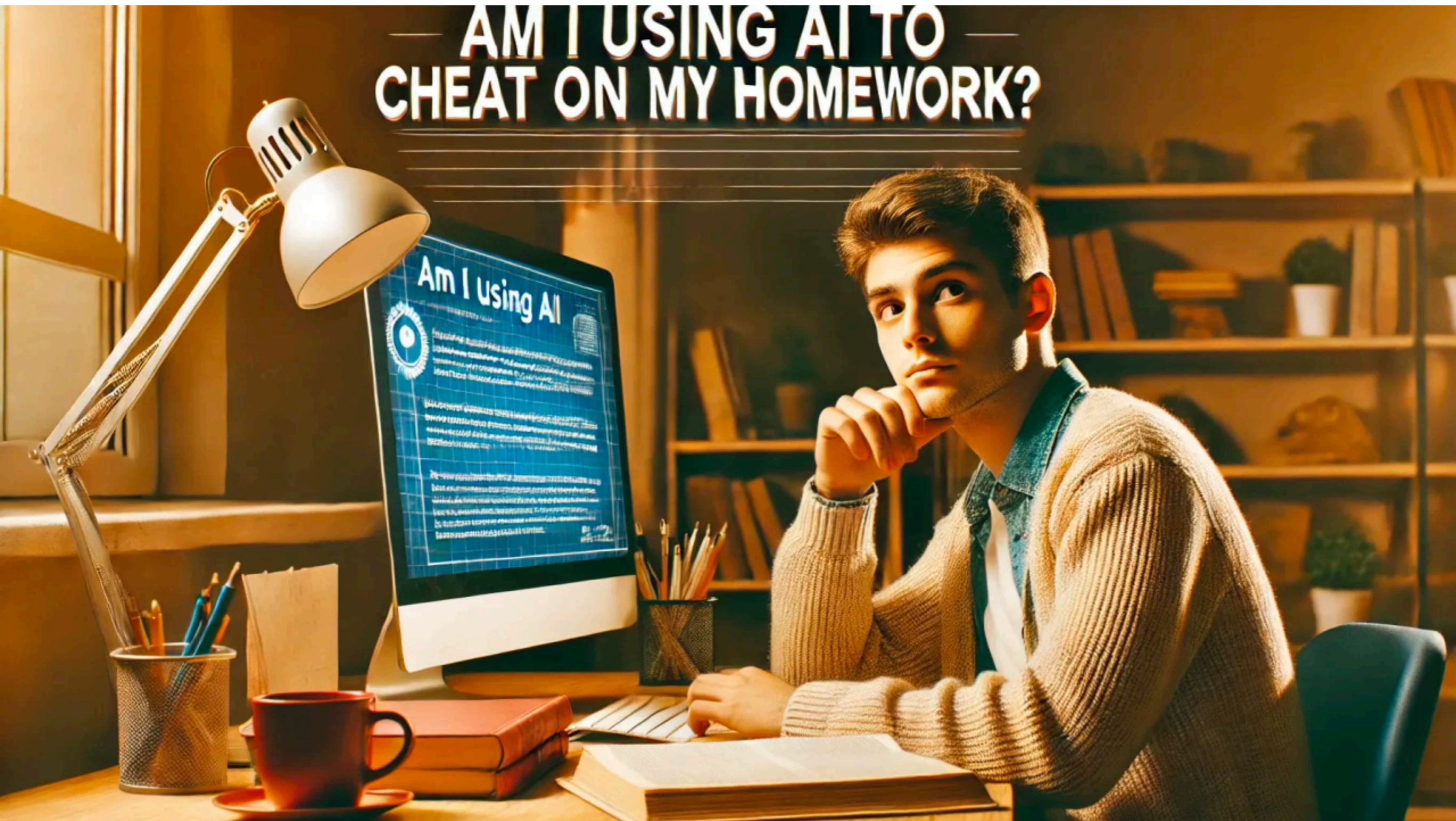
4

UNDER DEVELOPMENT

- NPRE+DS
- MatSE+DS
- ChemE+DS
- PHYS+DS

-  Aerospace Engineering
-  Agricultural & Biological Engineering
-  Bioengineering
CS+Bioengineering
-  Chemical & Biomolecular Engineering
-  Civil & Environmental Engineering
-  Electrical & Computer Engineering
-  Industrial & Enterprise Systems Engineering
-  Materials Science & Engineering
-  Mechanical Science & Engineering
-  Nuclear, Plasma & Radiological Engineering
-  Physics
CS+Physics

Dealing with the information age



How much data in an LLM?

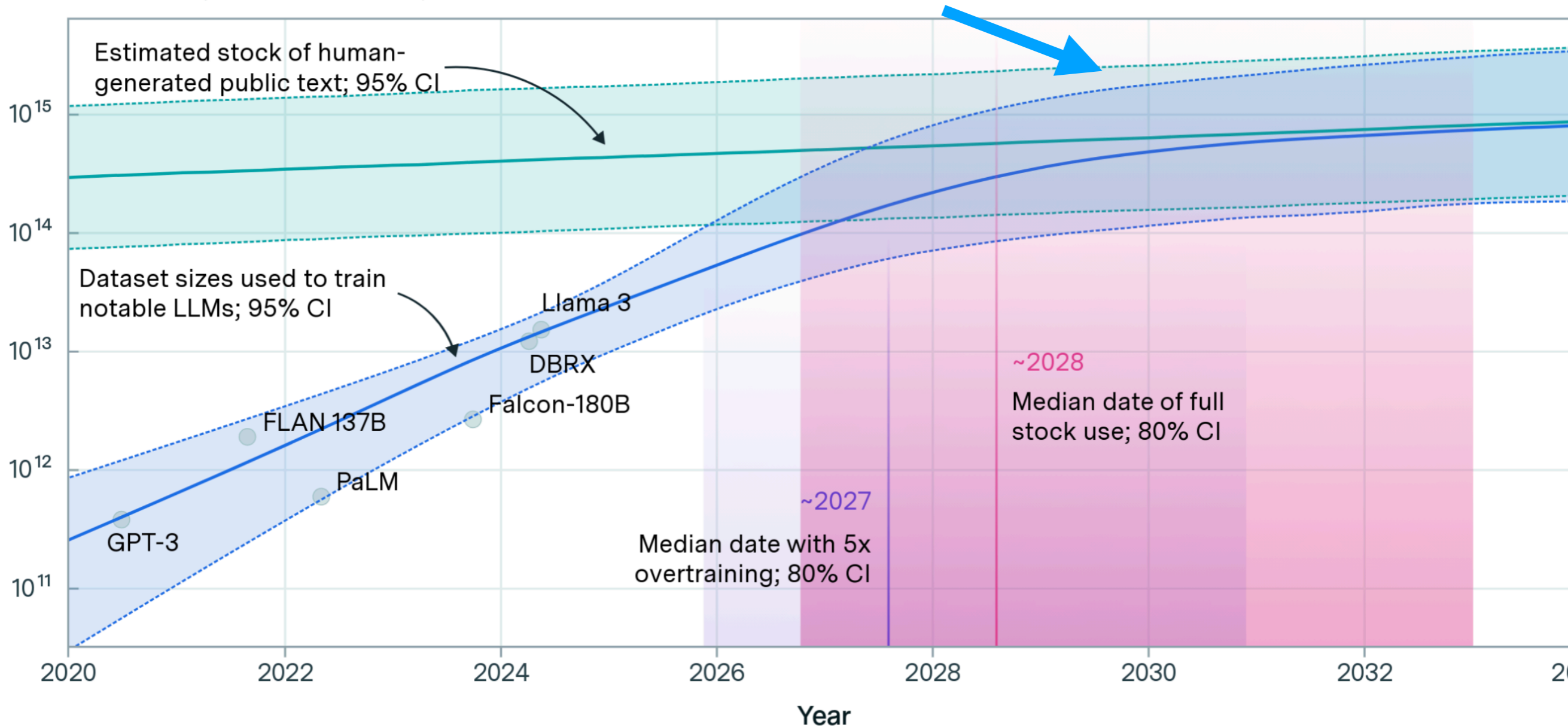
Its about 10 Terabits or so (not so much)

How much data in an LLM?

Its about 10 Terabits or so (not so much)

People seem afraid LLMs will saturate
I think we know better

Effective stock (number of tokens)



Where do we fit in computing landscape

OPPORTUNITIES FOR BUILDING ON HDR

- THIS INITIATIVE HAS LAUNCHED MANY INNOVATIVE IDEAS AND ACTIVITIES: HOW DO WE BUILD ON THAT ENERGY AND THOSE ACTIVITIES, ACROSS ALL SCALES?
- HOW DO WE ENSURE THAT HDR INVESTMENTS SUPPORT NEW OPPORTUNITIES? HOW DO WE BUILD ON THE CROSS-NSF INTEREST, TO ENCOURAGE A WIDER ARRAY OF COMMUNITIES TO COLLABORATE IN A DATA-ENABLED SCIENCE ECOSYSTEM?

Where do we fit in computing landscape

- We have 5 institutes focusing on
 - 5 broad topics covering many different domains

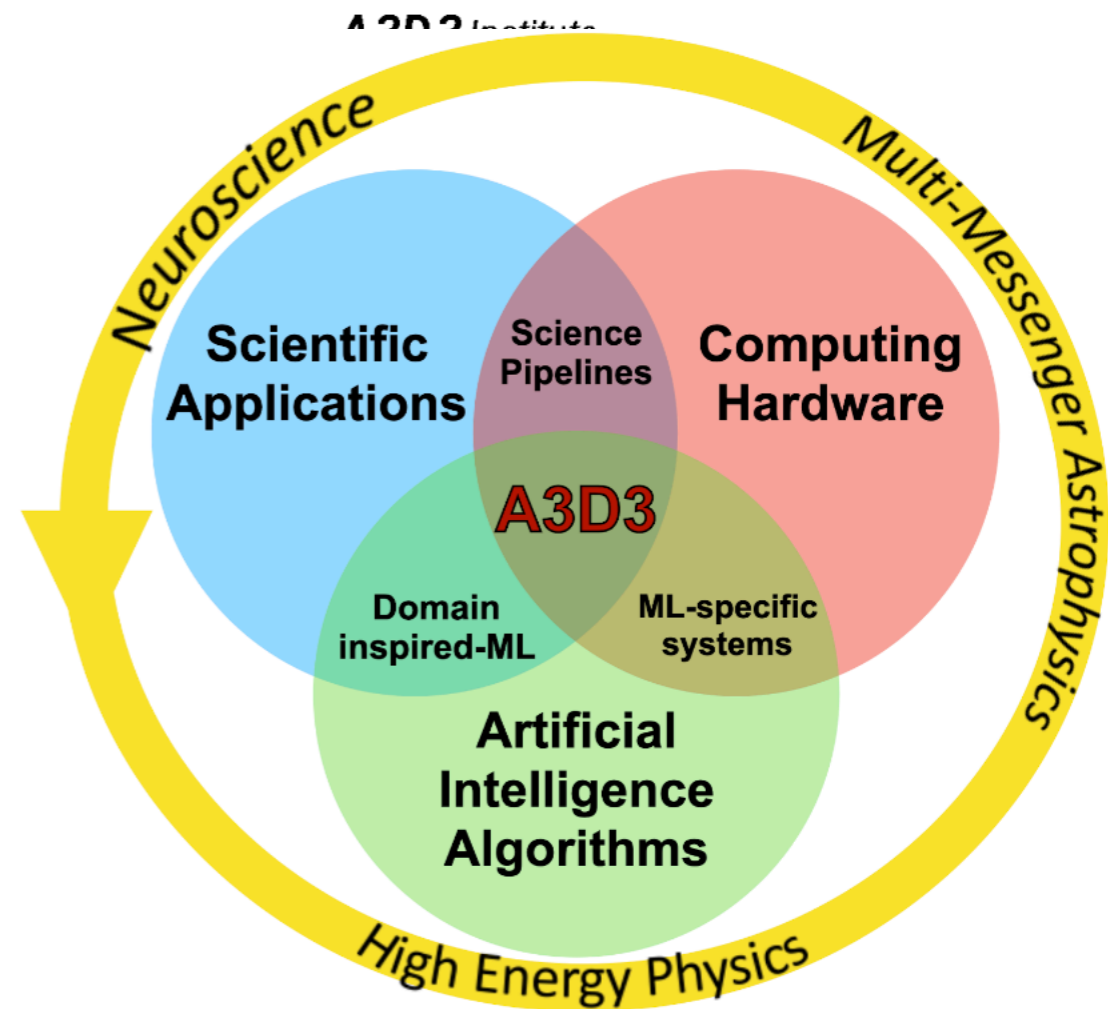
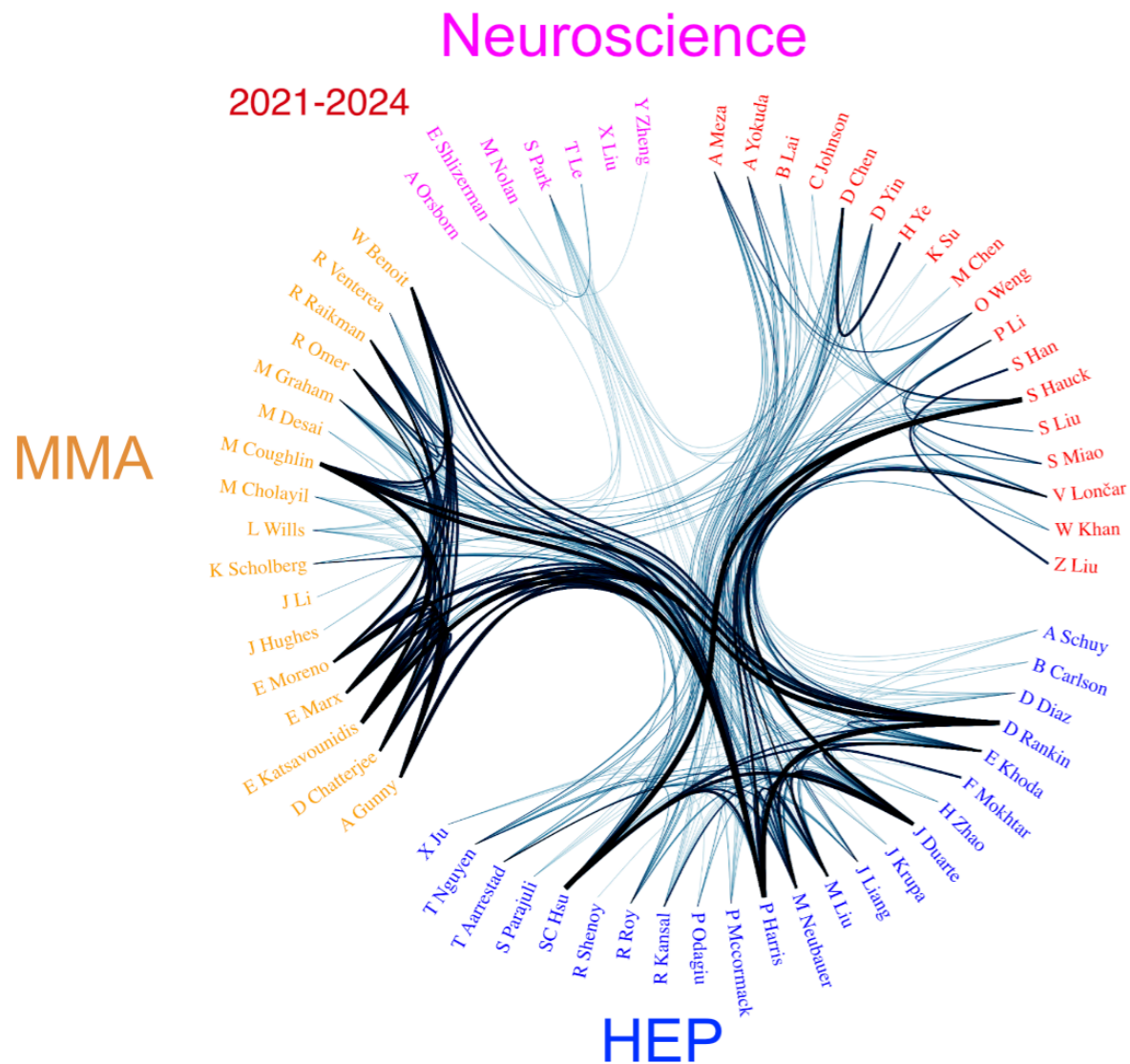




A3D3:

Rapidly connecting domains

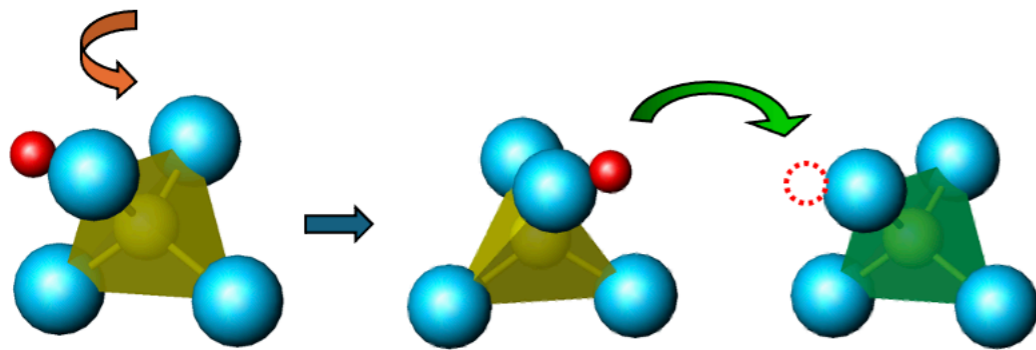
Making software frameworks to enable real-time AI for scientific discovery



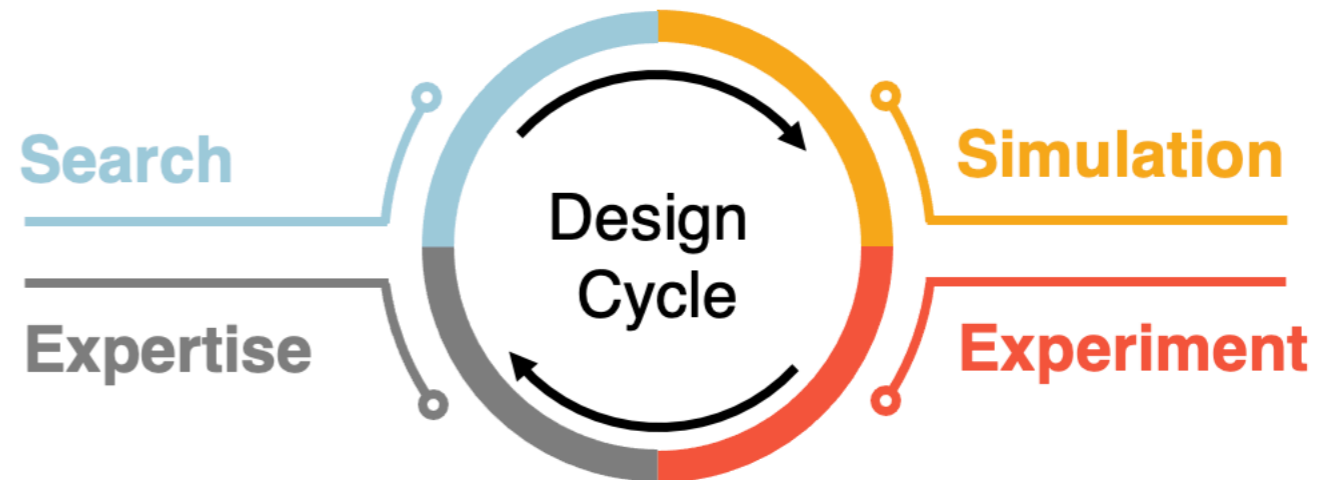
Connecting materials w/Design

Understanding of the micro materials feeds all the way to macro behavior

Ion transport for batteries and fuel cells



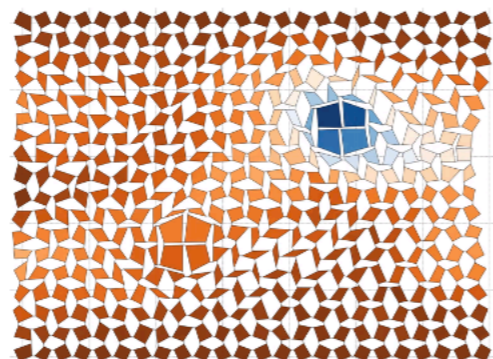
Coordinated dynamics driving proton hopping



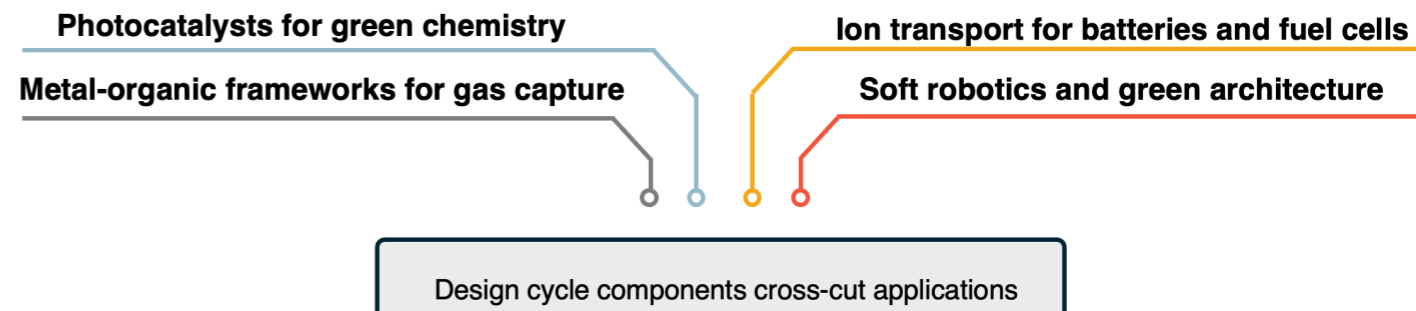
Soft robotics and green architecture



Scaffolding-free assembly

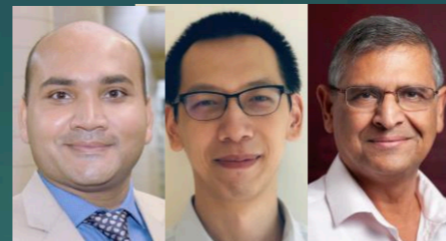


Reprogrammable material





Imageomics: Connecting Bio data with ML



A. Karpatne, X. Jia, V. Kumar
Knowledge-guided Machine Learning: Current Trends and Future Prospects. 2024

**Sparse, imperfect,
heterogeneous data**

Imageomics Data

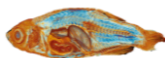
3D Textured Model



Digital Xray



3D mCT w/
Contrast



Videos in the Wild



Curated Images



**Text, Geo,
Molecular...**

Text descriptions

Redbreast sunfish

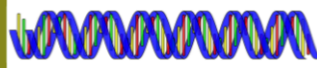
Fish

The redbreast sunfish is a species of freshwater fish in the sunfish family of order Perciformes. The type species of its genus, it is native to the river systems of eastern Canada and the United States. The redbreast sunfish reaches a maximum recorded length of about 30 cm, with a maximum recorded weight of 2.3 lb. Wikipedia

Geospatial



Molecular



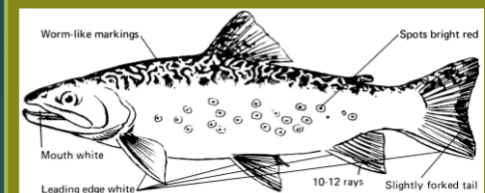
Knowledge-Guided ML

Biological Structures

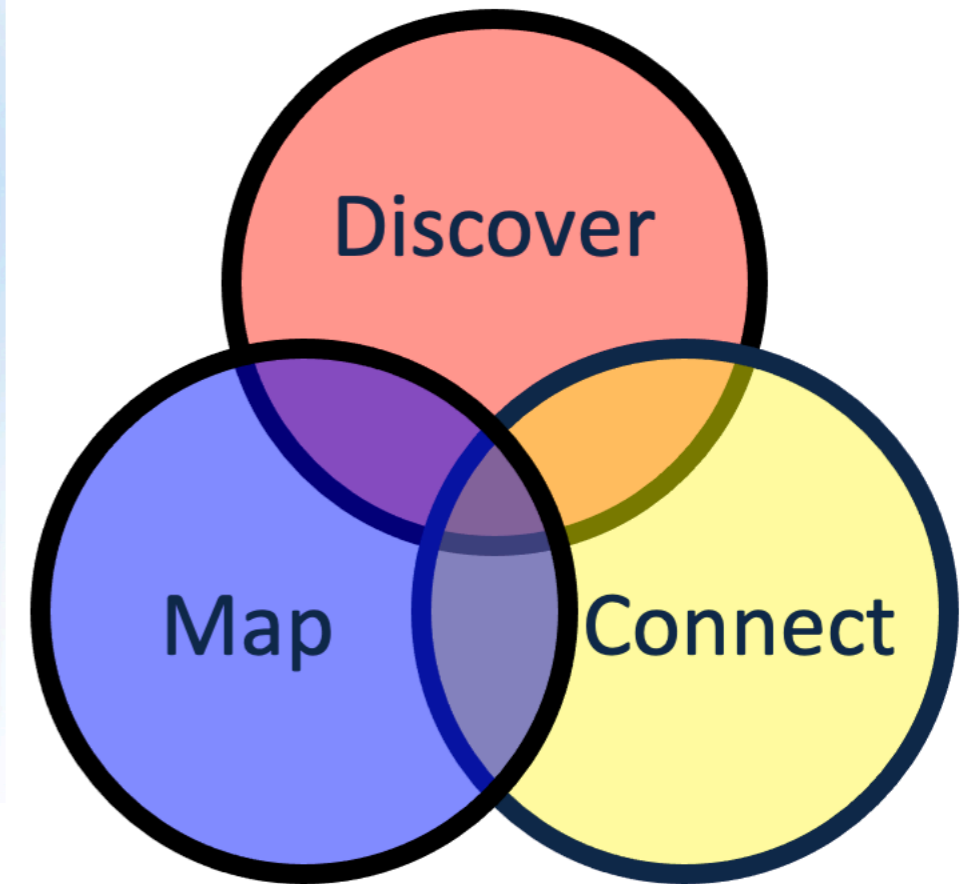
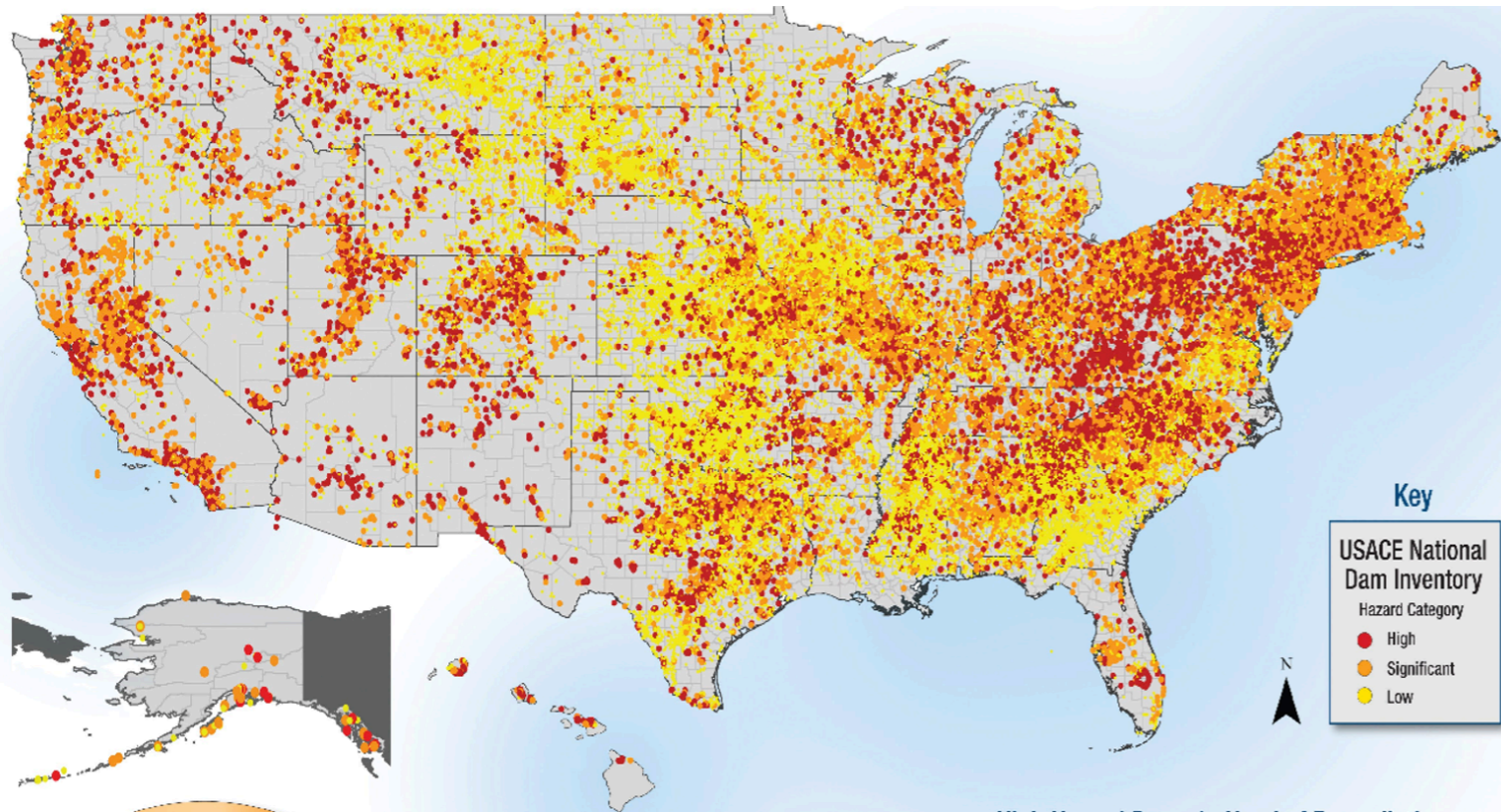
Phylogeny Trait Ontology Ethograms, etc.



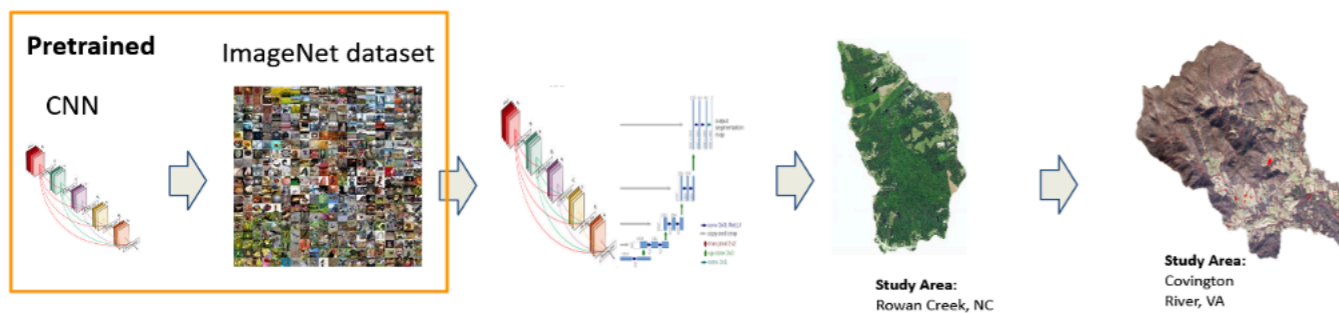
**Explainable
AI**



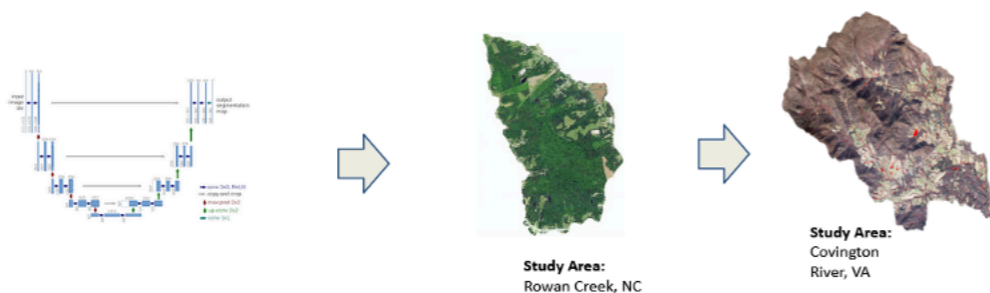
Harnessing the geospatial data revolution



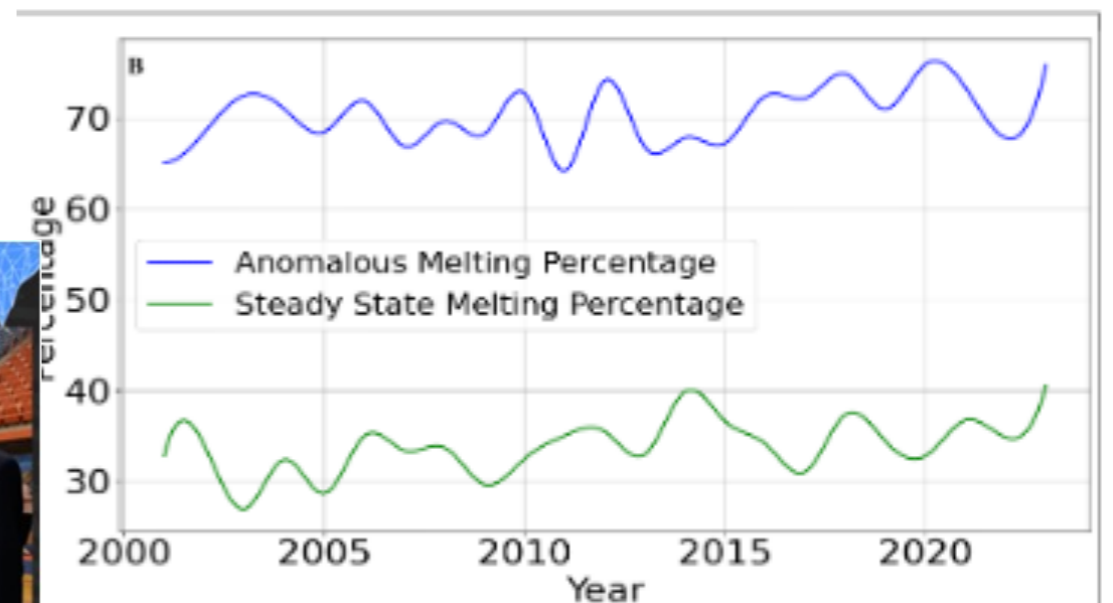
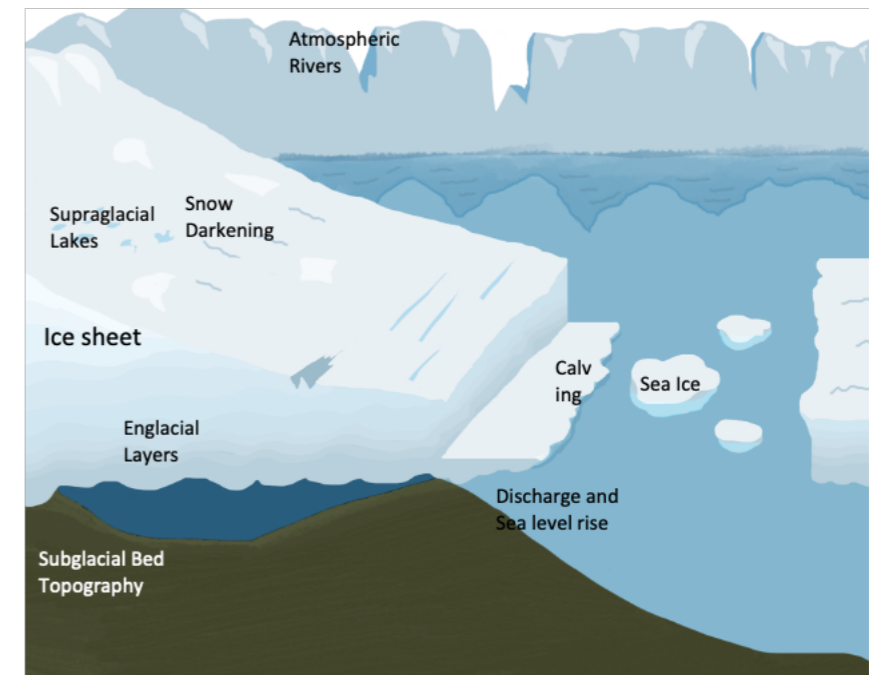
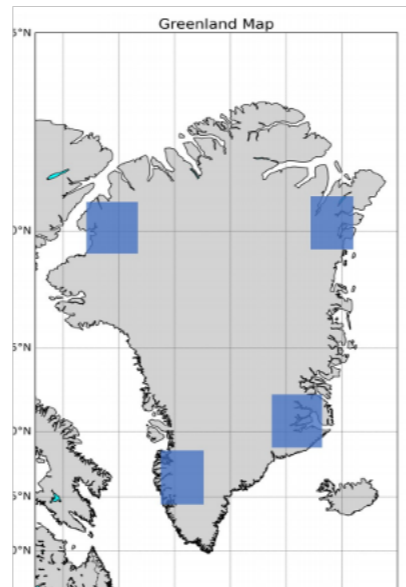
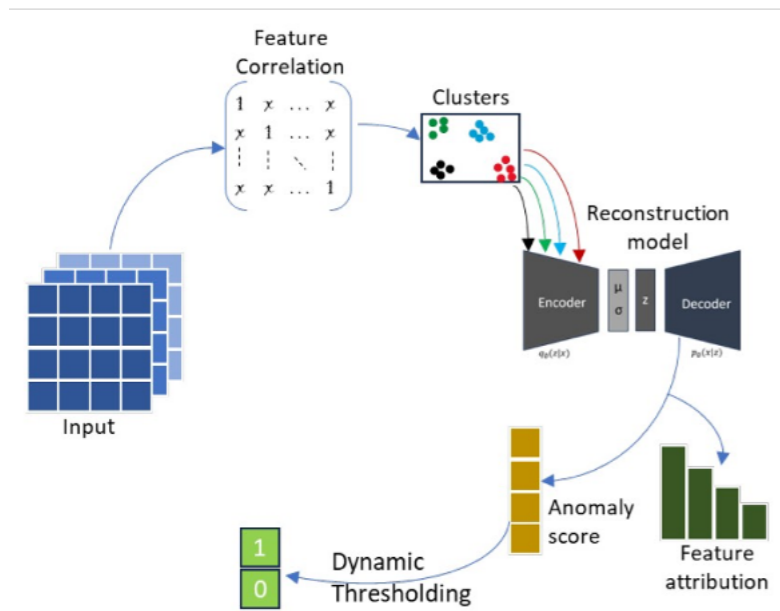
(ImageNet)



Attention U-net model



iHARP: Monitoring the Ice to understand the future



Where does AI in science we go from here?

AI for nature open challenges:

- ▶ Adding domain knowledge (KGML)
- ▶ Focus on the long tail, open set, distribution shift
- ▶ Novelty discovery
- ▶ Quantifying uncertainty
- ▶ Multimodal data analysis
- ▶ Model composition...
- ▶ ...including domain models
- ▶ Human-machine partnership by design

NSF Panel

- ACED call is coming out : builds on many ideas
- How do we connect the HDR activities to connect across HDR
 - Can we find more ways to collaborate
- Take the partnerships and leverage them
- Can we come up with a scheme for data, data preservation
 - Open data strategies
- **We should understand what HDR has produced**
 - Convergence accelerator/Pose/NAIRR pilot/RITEL
 - Don't get distracted by LLMs(shiny toy)
 - Data infrastructure needs some serious work

NSF Panel

The whole is greater than the sum of the parts



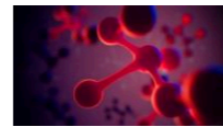
Trans-disciplinary Converge of domains
should be highlighted

Where do we go from here?

Initial NAIRR Pilot Design



AI Researchers



Domain Scientists Applying AI

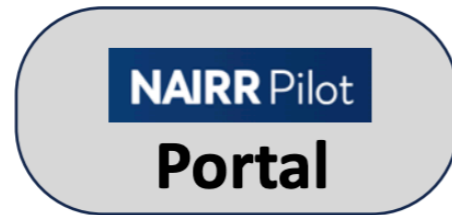


Students and Educators

Pilot Governance



Pilot Users



Community Design Process



Computing, testbeds, datasets, models, software, user support, training

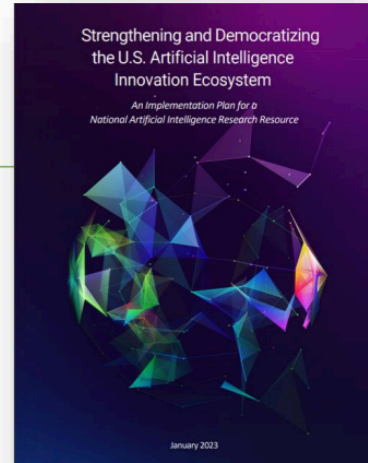


Note: the NAIRR Pilot will provide infrastructure to researchers. The NAIRR Pilot will not fund end-user research.

NAIRR Design for Data

What role does the NAIRR have in supporting data for AI research?

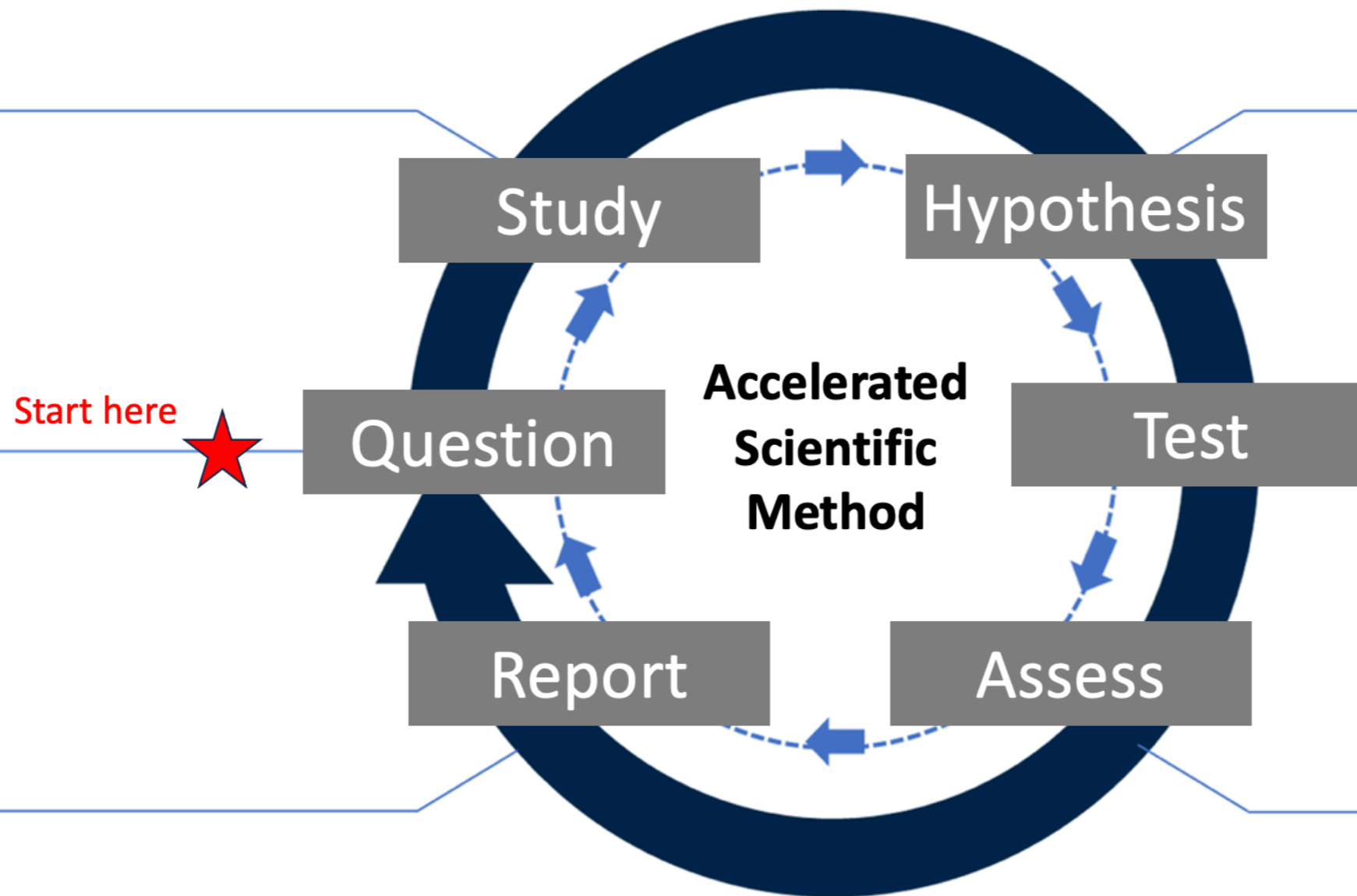
- Set guidelines and criteria for dataset inclusion in the NAIRR
 - Support a data search and discovery service
 - Encourage and incentivize communities to contribute 'analysis-ready' datasets
 - Facilitate an AI data commons
 - Provide access to integrated computing and data platforms
 - Provide access to restricted datasets
 - Provide technical expertise and user support for data communities and community driven curation efforts
- NAIRR will not define dataset standards, which continue to evolve and best defined by communities
 - NAIRR will not fund the collection or creation of specific community datasets



Prototype for a Foundation
model to understand all of
Science

FASST Project

Accelerating Discovery: Accelerating All Steps of Research



Ideally AI will help in all these steps

<https://doi.org/10.1038/s41524-022-00765-z>

Connecting w/Industry can help propel novel heterogeneous HW

AMD Research Center and HACC

1. Four Research Thrusts in the center

- **T1: Compilers and languages – programmability**
 - New high-level synthesis solutions, programming models, and fast compilers.
- **T2: Systems solutions – flexibility, scalability and heterogeneity**
 - Innovative frameworks for unified memory, virtualization, scheduling and resource allocation.
 - Targeting multi-FPGA and multi-accelerator setups.
- **T3: Distributed computing, networking, and storage – parallelization, security and reliability**
 - Smart NICs, cloud RPC, NVMe over fabric, near-data acceleration, security and reliability.
 - P4 and eBPF targeting cloud computing.
- **T4: Applications and algorithms acceleration – performance and efficiency**
 - Tuning, optimizing, and accelerating HPC, ML, bioinformatics, encryption, and other types of applications.
- **Inter-HACC collaboration opportunities**
 - E.g., unified memory work with Coyote of ETH HACC and graph processing with NUS HACC.

2. HACC: AMD Heterogeneous Accelerated Compute Cluster

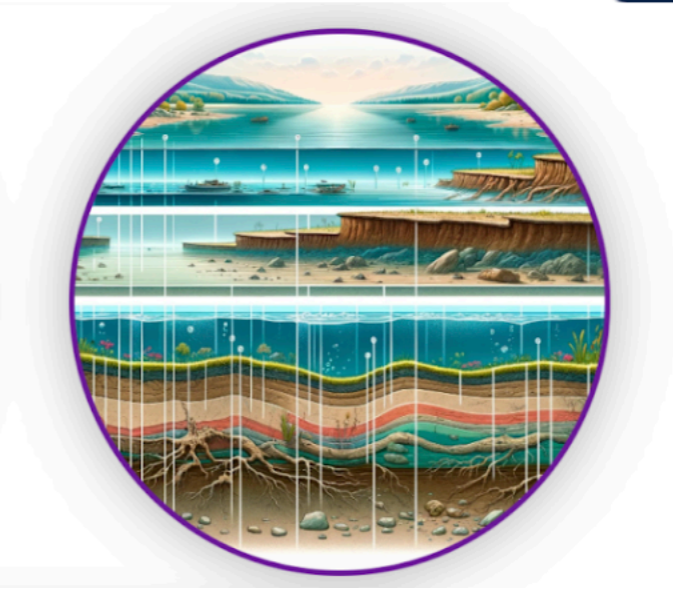
Panel AI&Data Computing

- There is a challenge to making computing accessible to the community
 - A lot of work comes in connecting different domains
 - Models that go across domains can really help
 - Engaging beyond our narrow scopes is something we should think about
 - ML Challenges and broad initiatives exist
 - Harnessing them can help push things forward

ML Challenge



NSF HDR ML Challenge



Data Preparation

FAIR Workflows

- **Full reproducibility was a major effort for this project**
 - Our workflows (not datasets) are all publicly available on github
 - Scoring, submission, data preparation, example models
 - We adhere to a common, public, docker container across all institutes
 - Additional packages (if needed) are installed at submission time
 - A whitelist is enforced to avoid corrupt/exploitive software
- **Our level of reproducibility/FAIRness was not present in other challenges**

A3D3



Imageomics



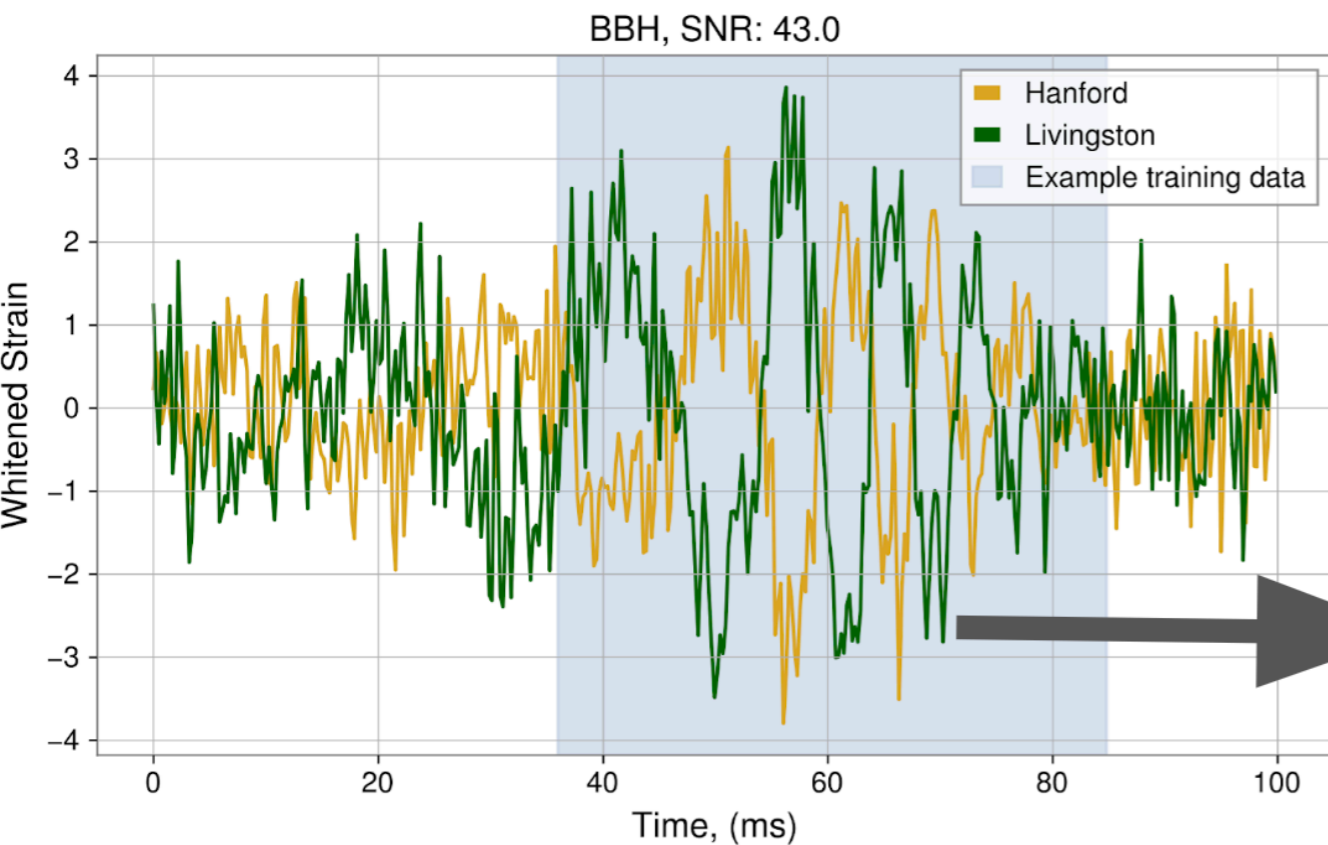
iHARP



A3D3 challenge

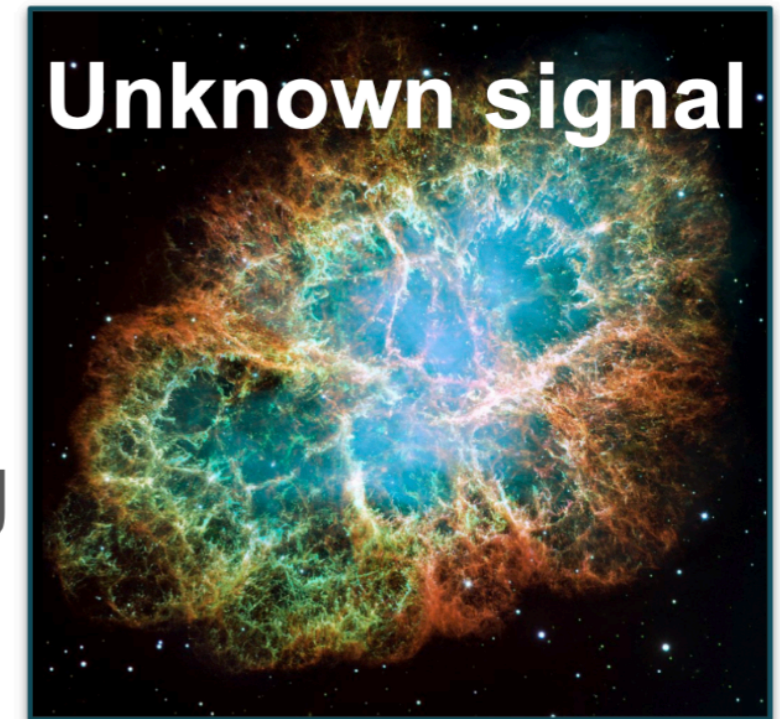
Anomaly detection in LIGO

- Many gravitational wave events are objects we know and understand
 - So far these are the only ones we have observed (Black Holes and Neutron Star Mergers)
- What if something astrophysical happens and we have no idea what it is?



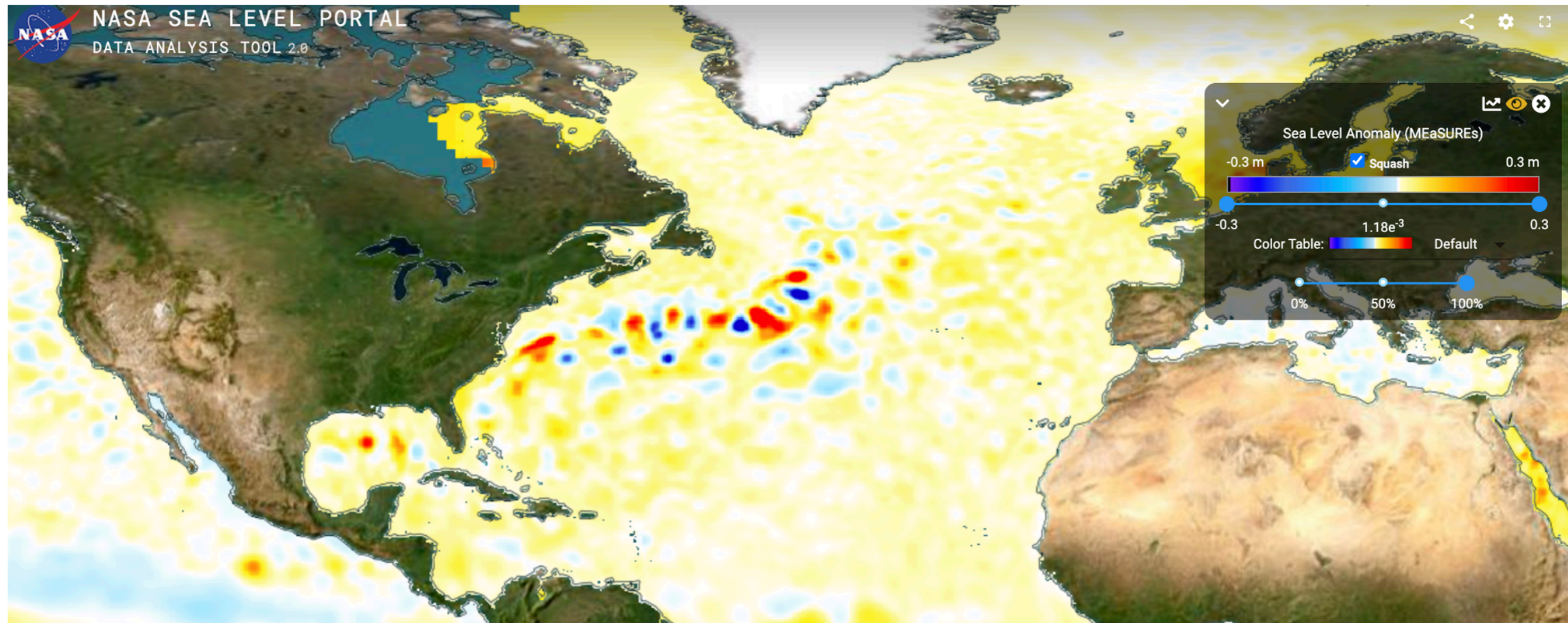
Not this
Something
Different

Core-collapse supernova (CCSN)



iHARP challenge

Machine Learning Challenge: Detect anomalous flooding events from satellite sea level maps



Imageomics challenge

The Challenge: Find the Hybrids

- Among Species A & B, can your algorithm find...
 - Species A signal hybrids?
 - Species A non-signal hybrids?
 - Species B hybrids (mimics of Species A signal hybrids)?

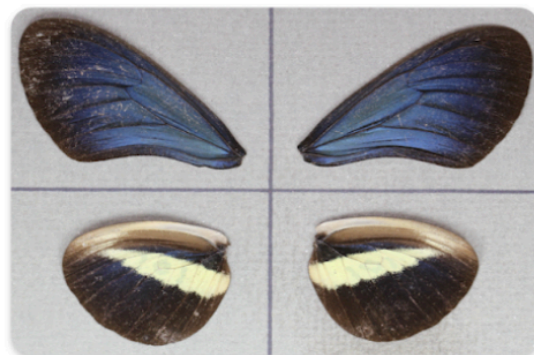
Species A subspecies I



Species A subspecies II



Species A subspecies III



Species A subspecies IV



Species B subspecies II



Species B subspecies I



Cross-HDR resources for education and outreach

Next steps for the repository...

Phase I (Committee-level)

Tasks:

- Created prototype database
- Created data curation form
- Piloted on 5-10 items

Phase II (Inter-HDR)

Tasks:

- Solicit entries from each institute
- Iterate database design and interface
- Expand to datasets

Feedback:

- Are there extraneous or missing metadata fields?
- How do you want to interact with the database? Are you able to?

Phase III (General Public)

Tasks:

- Promote the database to the public
- Iterate as needed

Feedback:

- How easy is it to browse and find new content?
- How easy is it to search for specific content?
- What types of content or data would be useful to share?

Panel on Workforce

- There is a challenge to making computing accessible to the community
 - A lot of work comes in connecting different domains
 - Models that go across domains can really help
- Challenge is:
 - How do we adapt the workforce to the rapid technological change
 - In some ways LLMs/ML makes it easier
 - In many ways it opens new difficult questions

Going to the future

The Promise of LLMs

AI-Driven Behavior Change Could Transform Health Care

**The End Of Originality:
Is AI Replacing Real Artists?**

AI suggested 40,000 new possible chemical weapons in just six hours

Will AI Eventually Replace Doctors?

This AI just figured out geometry – is this a step towards artificial reasoning?

As More AI Tools Emerge in Education, so Does Concern Among Teachers About Being Replaced

As Suresh Pointed out there is huge hype for what is going on

Where are we going ?

The Reality

Airline held liable for its chatbot giving passenger bad advice - what this means for travellers

Lessons From Google's AI Mistakes

Google's new "AI Overview" feature reveals the flawed underbelly of AI

New York lawyers sanctioned for using fake ChatGPT cases in legal brief

**Leading AI Companies Fail to Address Problems With
ation**

**AI Has Become a Technology of
Faith**

Sam Altman and Arianna Huffington told me that they believe generative AI can help millions of suffering people. I'm not so sure.

By Charlie Warzel

At the same time there are frequent set backs to people's realiance

Suggestions from Suresh

Takeaways

- We **must** distinguish the promise from the reality of AI
- We can do this by being scientifically rigorous and using scientific discovery to expand the range of policy options
- We do this through *sociotechnical research* that allows us to
 - *critique*
 - *measure*
 - *evaluate, and*
 - *govern*

Our role as HDR members is to take advantage of science

Where should we do?

how can I make sure the harnessing the data revolution initiative continues beyond 5 years



Ensuring the long-term sustainability of the *Harnessing the Data Revolution (HDR)* initiative or

.....

1. Secure Long-term Funding
2. Institutionalize the Initiative
3. Engage and Build a Supportive Community
4. Demonstrate Value through Impact
5. Leverage Technology and Infrastructure
6. Adapt and Evolve with Changing Trends

Where should we do?

how can I make sure the harnessing the data revolution initiative continues beyond 5 years

1.1. Visionary Leadership

- **Clarity of Purpose:** The initiative's mission and goals are clear and compelling.
- **Adaptability and Innovation:** The initiative is open to new ideas and approaches.
- **Ecosystem Builders:** Trustworthy partners are identified and engaged.

2. Creating a Culture of Collaboration

- **Interdisciplinary:** Encourage cross-pollination of ideas from different fields.
- **Co-creation with Stakeholders:** Involve partners in the design and implementation of the initiative.
- **Engaged Communities:** Build strong relationships with all stakeholders.

3. Commitment to Impact

- **Focus on Tangible Outcomes:** Measure and report on the initiative's impact on science and society.
- **Real-World Problems:** Address pressing challenges that require data-driven solutions.
- **Iterative Progress:** Learn from failures and adapt the initiative's strategy.

4. Agile and Forward-Thinking

- **Invest in Scalable Solutions:** Develop data science programs that can be scaled across different contexts.
- **Data Ethics and Security:** Prioritize the responsible use of data and protect sensitive information.
- **Data Openness:** Share data and findings to maximize the initiative's impact.

5. Storytelling and Advocacy

- **Powerful Narrative:** The ability to tell compelling stories about how data is transforming lives is essential for long-term buy-in. Effective communication can help demystify data science, making it accessible and relatable to decision-makers and the general public.
- **Policy Advocacy:** Building strong relationships with policymakers to secure regulatory and political support is crucial. By demonstrating how the HDR initiative can shape positive outcomes—such as better healthcare, environmental sustainability, or economic growth—you can advocate for supportive policies and public funding.

6. Nurturing Talent and Capacity Building

- **Develop Talent Pipelines:** Partner with universities, offer internships, and build data science programs that create a steady flow of skilled professionals. These talent pipelines will help sustain the initiative by feeding it with future leaders and innovators.
- **Retain Top Talent:** Create a stimulating work environment with opportunities for professional growth, cutting-edge projects, and impact-driven work to retain talented individuals who are passionate about the mission.
- **Democratizing Data Literacy:** Build programs to upskill a broad range of stakeholders in data literacy. This will ensure that a wider audience—beyond data scientists—can engage with and understand the value of the HDR initiative.

7. Sustained Funding with Diversification

- **Creative Financing Models:** Beyond traditional grants, consider creative financing such as impact investments, venture funding for spin-offs, or self-sustaining services that monetize data-driven insights while fueling research.
- **Diversified Partnerships:** The secret lies in not relying too heavily on one type of funder or source of income. Engage a mix of public, private, and philanthropic entities to reduce risk and enhance financial sustainability.
- **Long-term Endowments:** Establishing an endowment fund early in the initiative's life can provide financial security, ensuring that it can continue operations, even if short-term funding sources fluctuate.

8. Global Perspective with Local Relevance

- **Think Globally, Act Locally:** Having a global vision is important, but it's equally essential to solve problems that are deeply relevant to the communities you serve. This dual focus will give the initiative the flexibility to tackle local issues while maintaining a global outlook.
- **Cross-border Collaboration:** Leverage global expertise and datasets by collaborating with international institutions, while ensuring the insights are tailored to local contexts. A global network of partners can sustain momentum and attract international funding.

9. Innovation Culture

- **Fail Fast, Learn Fast:** Cultivate a mindset that encourages experimentation and rapid prototyping, so the initiative is always at the cutting edge of data science. Success often comes from testing bold ideas, failing quickly, and applying lessons learned.
- **Disruption Preparedness:** Foster a culture of innovation where the team is constantly anticipating future technological trends, ensuring that the initiative remains relevant even as data technologies evolve.



Ensuring the long-term sustainability of the *Harnessing the Data Revolution (HDR)* initiative or

can you elaborate on what the secret sauce for success would be




The "secret sauce" for the long-term success of an initiative like *Harnessing the Data Revolution (HDR)* lies in a combination of key principles and strategic actions. While the steps I outlined

Where should we do?


how can I make sure the harnessing the data revolution initiative continues beyond 5 years

 Ensuring the long-term sustainability of the *Harnessing the Data Revolution (HDR)* initiative or

can you elaborate on what the secret sauce for success would be

 The "secret sauce" for the long-term success of an initiative like *Harnessing the Data Revolution (HDR)* lies in a combination of key principles and strategic actions. While the steps I outlined

This answer is just stating the obvious I want something useful can you tell me that

 I understand you're looking for something more insightful, beyond the standard recommendations. Let's get into what really makes a difference—beyond the theoretical, and into the practical nuances of *making it work*. Here's a more pragmatic take on how to ensure the HDR initiative thrives:

Make it Cool and Ubiquitous

Viewpoint

This conference shows that the institutes are producing important results

We are harvesting the data revolution



2024 HDR Ecosystem Conference
Harvesting the Data Revolution
Sep 9-12, 2024 @ UIUC

The 2024 NSF HDR Ecosystem Conference is hosted and supported by:

With gratuitous support from our sponsors:

 **The Grainger College of Engineering**
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Organizing Committee

Tanya Berger-Wolf (OSU)	Mohamed Mokbel (UMN)
Karen Chen (UMBC)	Josephine Namayanja (UMBC)
Julia Gonski (SLAC)	Arnab Nandi (OSU)
Phil Harris (MIT)	Mark Neubauer (UIUC, Chair)
Shih-Chieh Hsu (UW)	Anand Padmanabhan (UIUC)
Vandana Janeja (UMBC)	Eric Toberer (CU Mines)
Paula Mabee (NEON)	Shaowen Wang (UIUC)

Conference Staff

Tonya Ayers (UIUC)
Brandy Koebbe (UIUC)
Menglu Zhang (UW)



<https://go.illinois.edu/hdr2024>

We need to go from Harvesting to Publicizing



We need to go from Harvesting to Publicizing



Something where strength in Numbers will help



Combine ubiquitous and cool features of multi messenger astronomy biological genomics materials science polar science and geospatial data

Something where strength in Numbers will help



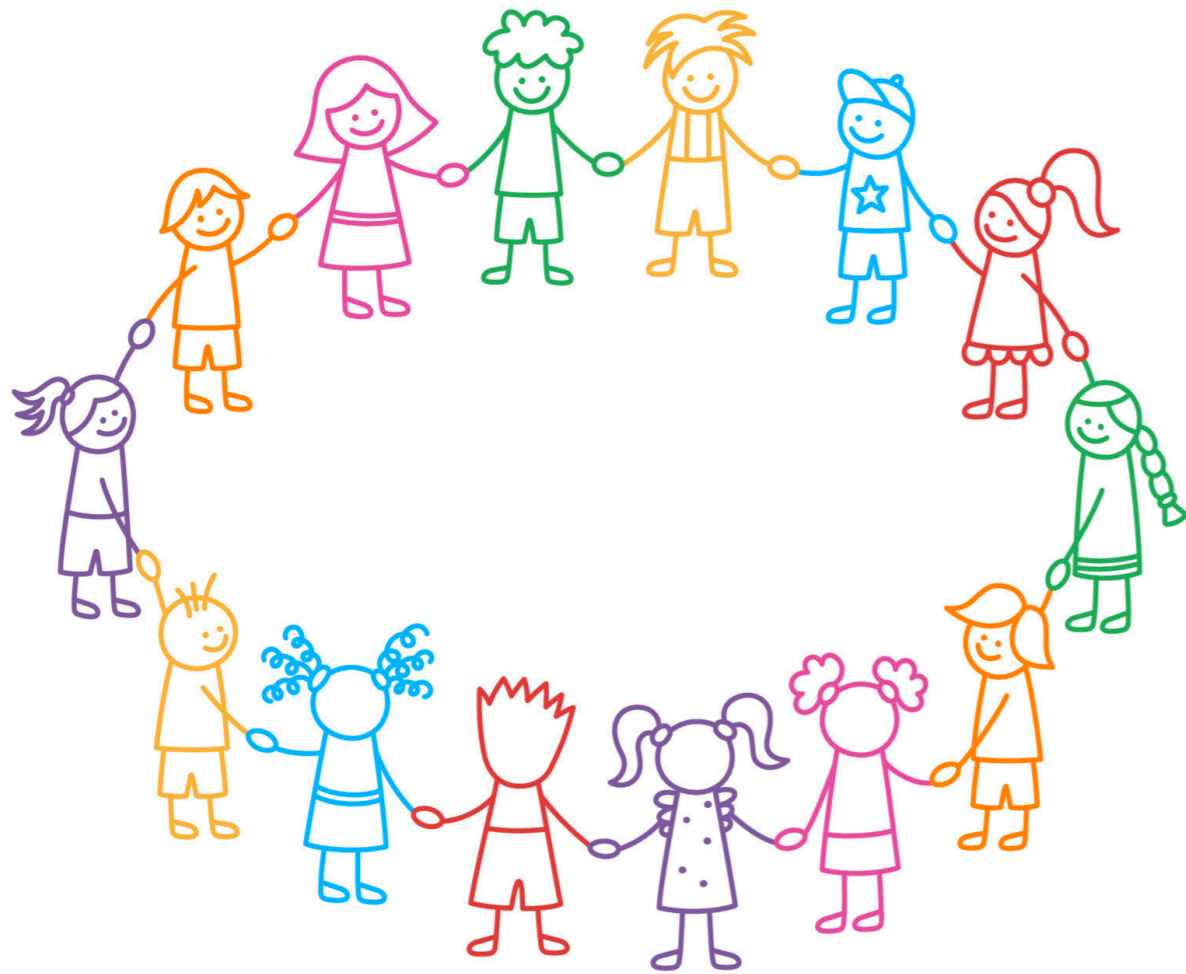
Combine ubiquitous and cool features of multi messenger astronomy biological genomics materials science polar science and geospatial data

Something where strength in Numbers will help



Combine ubiquitous and cool features of multi messenger astronomy biological genomics materials science polar science and geospatial data

Something where strength in Numbers will help



Combine ubiquitous and cool features of multi messenger astronomy biological genomics materials science polar science and geospatial data

The Next Year

- We have seen institutes connecting a broad range of domains
 - These connections are starting to produce significant results



The Next Year

- We have seen institutes connecting a broad range of domains
 - These connections are starting to produce significant results



A common collaboration across the institutes can highlight criticality of our work

Combination of Data and Science can serve to push us forward



Thanks!