Stats Up AI Alliance

ASA Townhall

The role of statistics for the future of AI

Feb 7th, 2024

Welcome - Ron Wasserstein

Executive Director

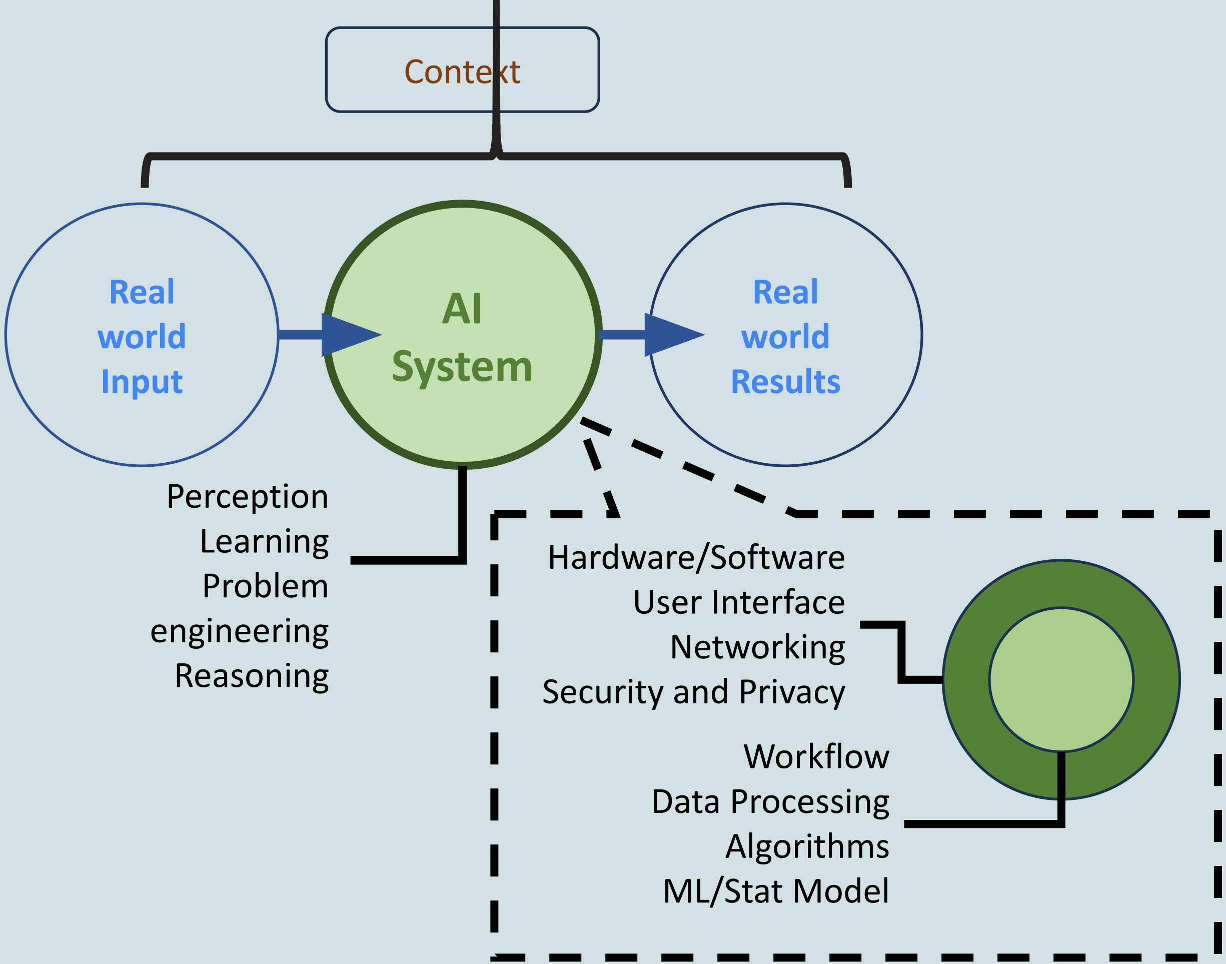
American Statistical Association

Stats Up AI Alliance

Introduction - Tian Zheng (Columbia)

Stats Up AI Alliance

“Where does Statistics fit in?” *versus* “What do Statisticians fit in?”



Development Tasks:

- **Problem identification**
- Problem modulization
- Metric development
- Workflow development
  - Training
  - Deployment
- Data Engineering
- Model development
- Model evaluation
- System development and deployment
- System evaluation and testing

Development resources:

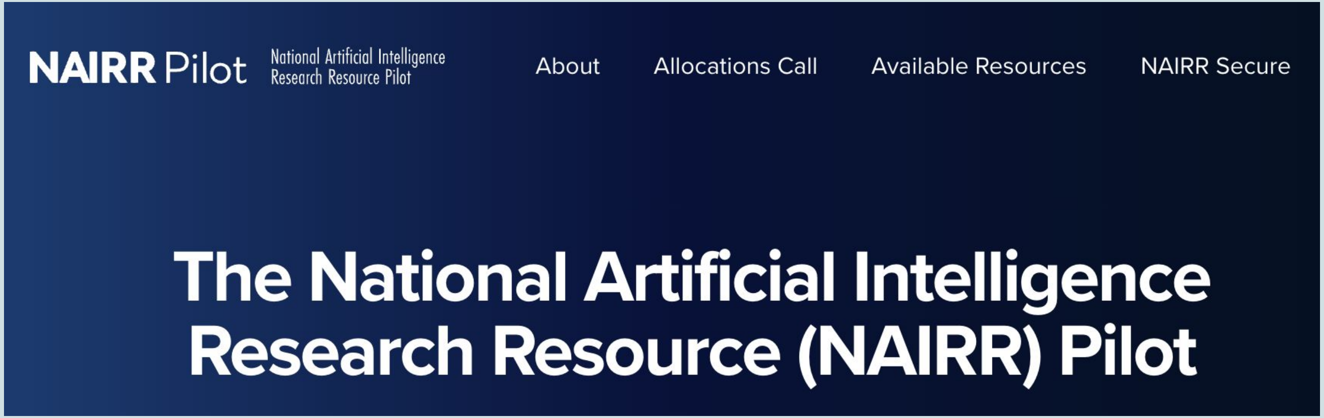
- **Training data**
- Computing infrastructure
- Engineering resources
- **Domain knowledge**
- Data science expertise

Domain Statistics skills
Science Computer science skills

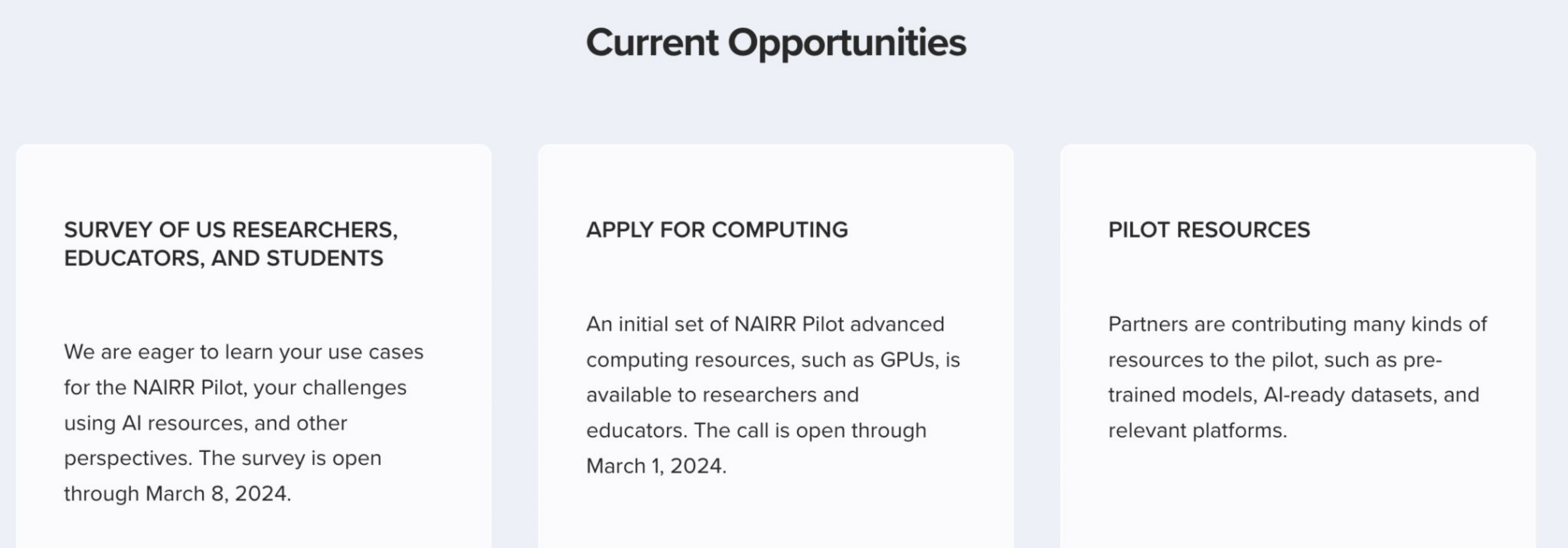
Is Statistics impactful?

- Who should be responsible for plugging “statistical innovations” into real-world AI applications? And make sure it works?
  - Hardware/software
  - Scaling
  - Evaluation and tuning
- Route I: Statisticians.
  - Statisticians are often not bothered by the “CS tasks.”
- Route II: CS and domain scientists pick up ”stat skills”, whenever they are **readily usable**.
  - Statisticians do not often worry about *frictionless adaptation* of their research, or how well their methods work in *real real-world* applications.

Why now?



The National Artificial Intelligence Research Resource (NAIRR) Pilot



Current Opportunities

SURVEY OF US RESEARCHERS, EDUCATORS, AND STUDENTS

We are eager to learn your use cases for the NAIRR Pilot, your challenges using AI resources, and other perspectives. The survey is open through March 8, 2024.

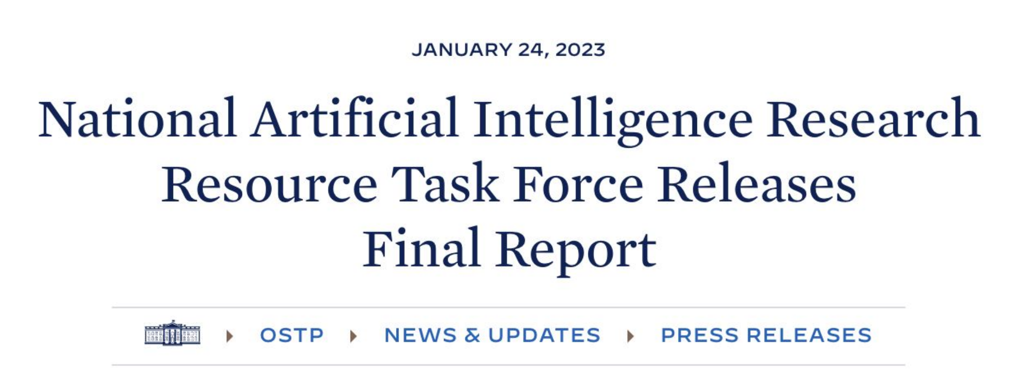
APPLY FOR COMPUTING

An initial set of NAIRR Pilot advanced computing resources, such as GPUs, is available to researchers and educators. The call is open through March 1, 2024.

PILOT RESOURCES

Partners are contributing many kinds of resources to the pilot, such as pre-trained models, AI-ready datasets, and relevant platforms.

Why now?



JANUARY 24, 2023

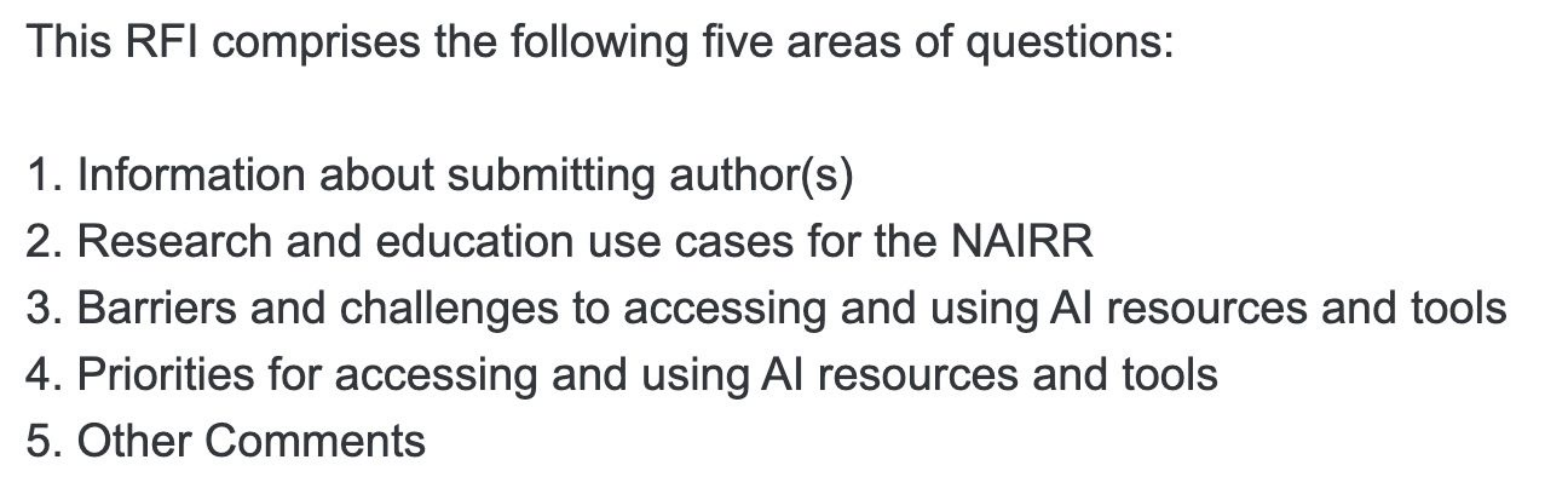
National Artificial Intelligence Research Resource Task Force Releases Final Report

OSTP ▸ NEWS & UPDATES ▸ PRESS RELEASES

Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem

An Implementation Plan for a National Artificial Intelligence Research Resource

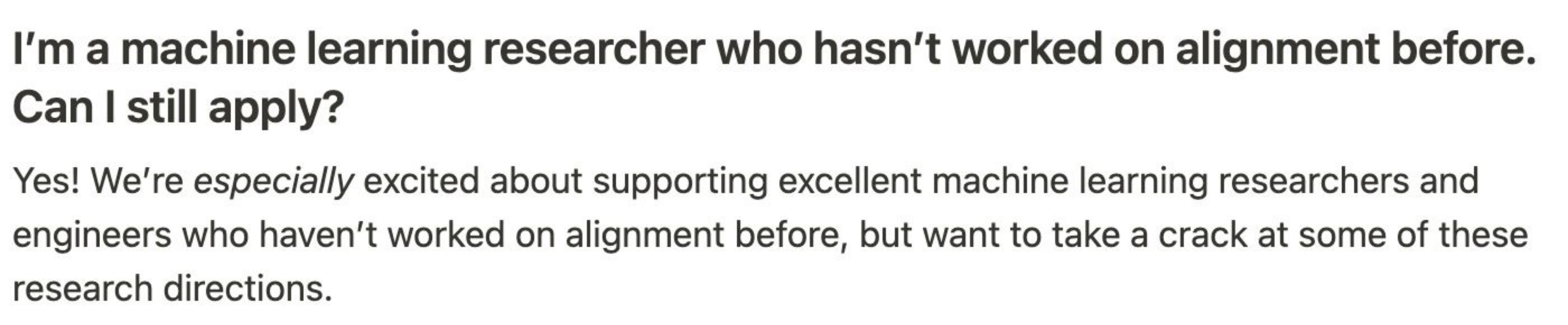
NAIRR Survey Questions



This RFI comprises the following five areas of questions:

1. Information about submitting author(s)
2. Research and education use cases for the NAIRR
3. Barriers and challenges to accessing and using AI resources and tools
4. Priorities for accessing and using AI resources and tools
5. Other Comments

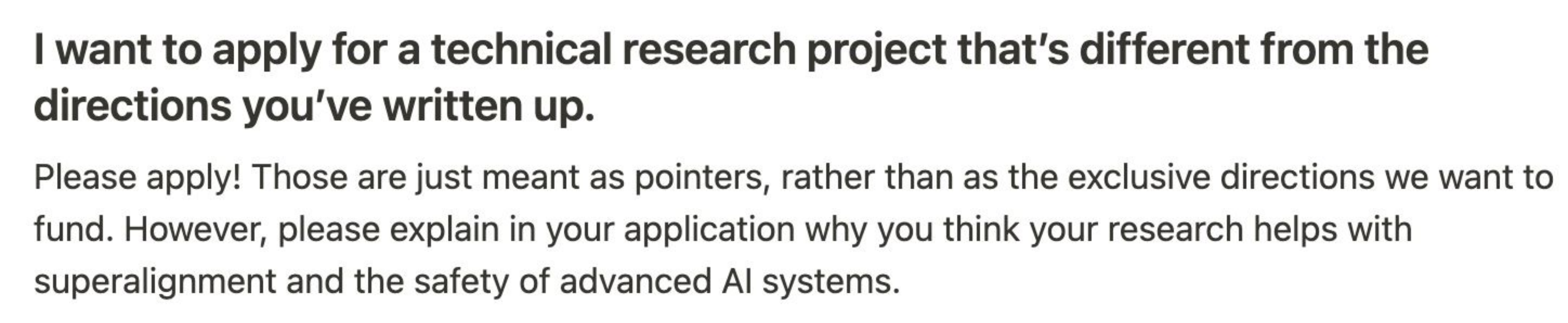
Why now?



I’m a machine learning researcher who hasn’t worked on alignment before. Can I still apply?

Yes! We’re *especially* excited about supporting excellent machine learning researchers and engineers who haven’t worked on alignment before, but want to take a crack at some of these research directions.

systems, including weak-to-strong



I want to apply for a technical research project that’s different from the directions you’ve written up.

Please apply! Those are just meant as pointers, rather than as the exclusive directions we want to fund. However, please explain in your application why you think your research helps with superalignment and the safety of advanced AI systems.

How do we feel about **the door** into AI for Statistics or Statisticians?

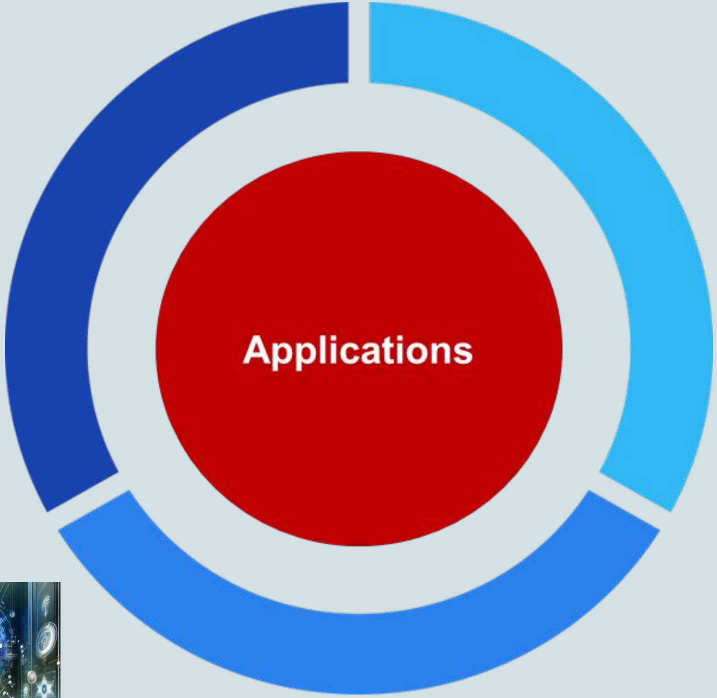
- Does it feel like that it is widening or narrowing?
- Who are driving AI research?
- How much statisticians are involved in AI applications?
- Should we be concerned about our level of participation?
- What should we do as individuals and as a community?

Worrisome Trends! - Hongtu Zhu (UNC)

Stats Up AI Alliance

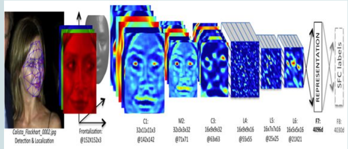
Ecological Layout for AI

Funding Agencies

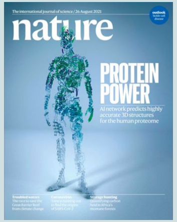


Universities

IMAGENET

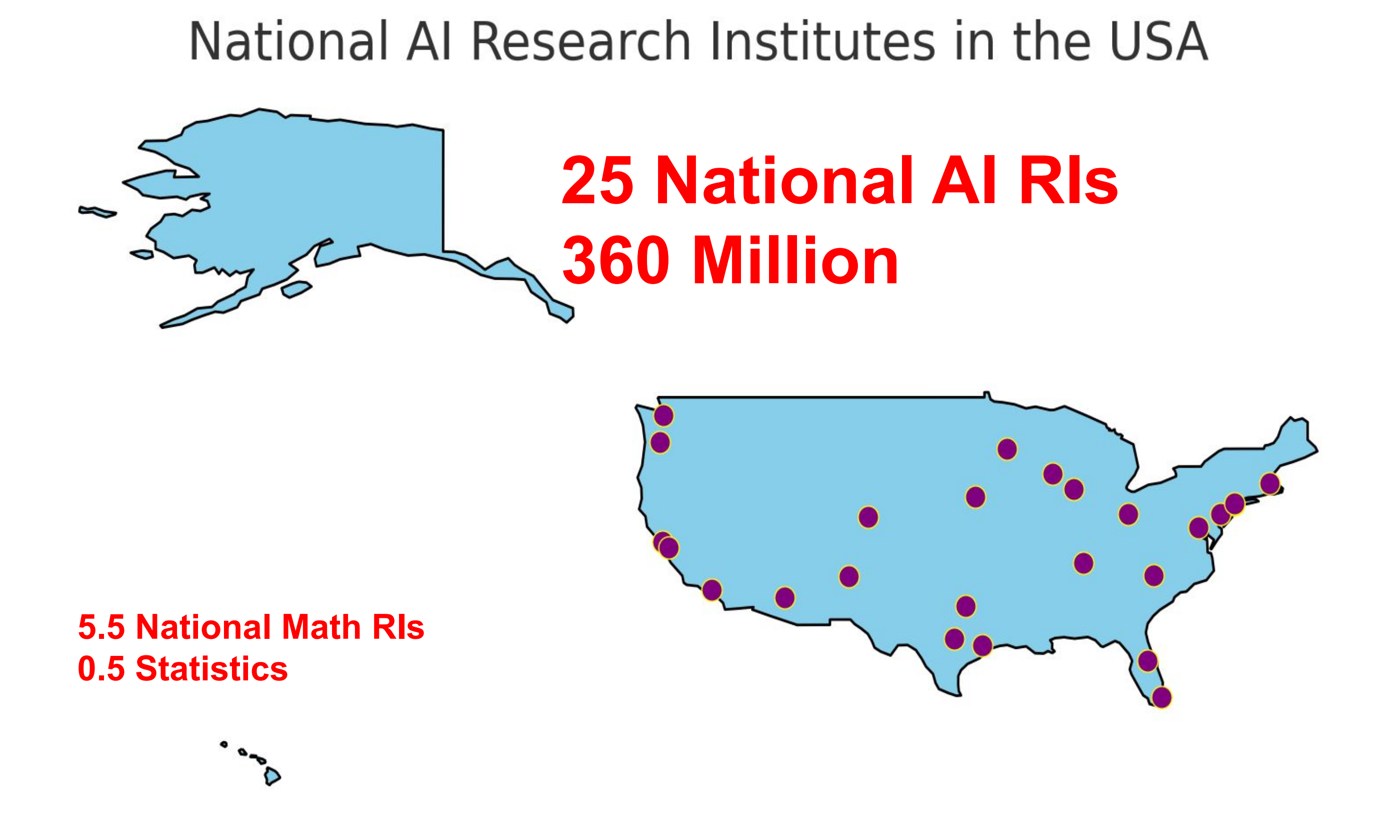


nature
At last — a computer program that can beat a champion Go player
ALL SYSTEMS GO


nature
PROTEIN POWER
AI network predicts highly accurate 3D structures for the human proteome

Private Sectors

National Artificial Intelligence (AI) Research Institutes



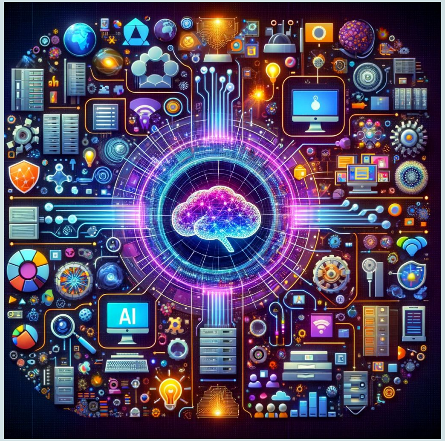
25 National AI RIs
360 Million

5.5 National Math RIs
0.5 Statistics

The National Artificial Intelligence Research Resource

NAIRR is a concept for a shared national research infrastructure that will connect U.S. researchers to **responsible and trustworthy AI resources, as well as the needed computational, data, software, training and educational resources** to fuel AI research and discovery.

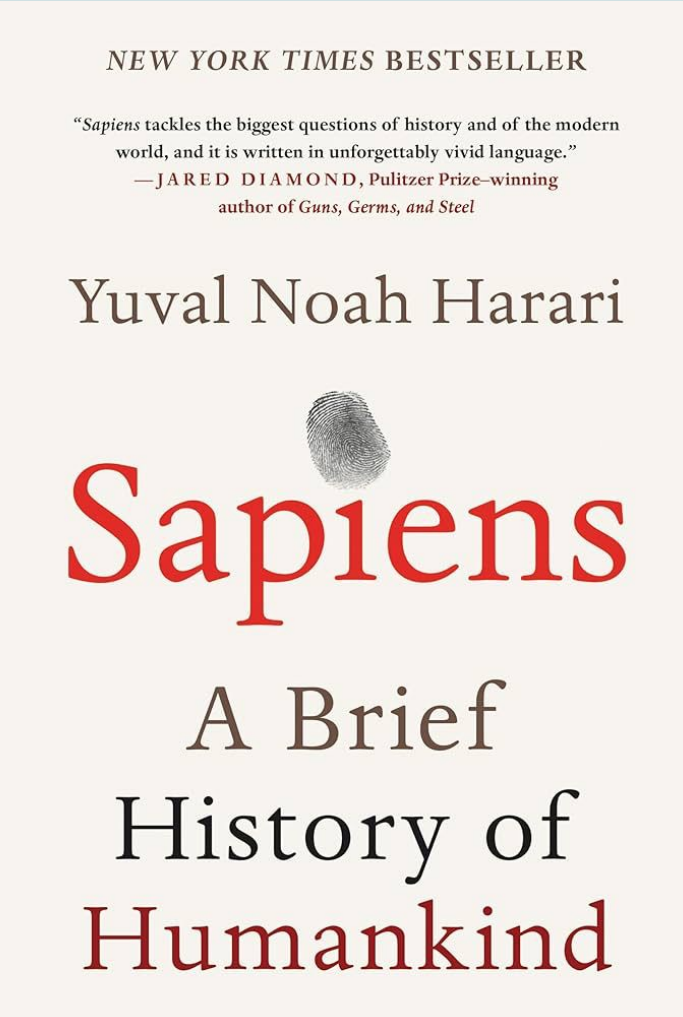
"We are excited about the expanded community of innovation that is emerging from NAIRR, and the pilot convening has been a tremendous success."



- NSF Director Sethuraman Panchanathan

Paradigm Shift! - Haoda Fu (Eli Lilly)

Stats Up AI Alliance



NEW YORK TIMES BESTSELLER

"*Sapiens* tackles the biggest questions of history and of the modern world, and it is written in unforgettably vivid language."
—JARED DIAMOND, Pulitzer Prize–winning author of *Guns, Germs, and Steel*

Yuval Noah Harari

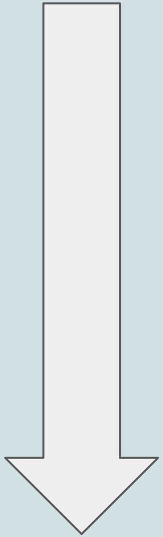
Sapiens

A Brief History of Humankind

From Hunting to Farming

Frictionless Research to Speed up Iterations

1. **Data:** research data sharing (privacy protected data)
2. **Code:** Github + containerization
3. **Problems:** Focus efforts for methodology iterations



Data Scientist

Full Stack Data Scientist

Full Scientist

Be Curious!

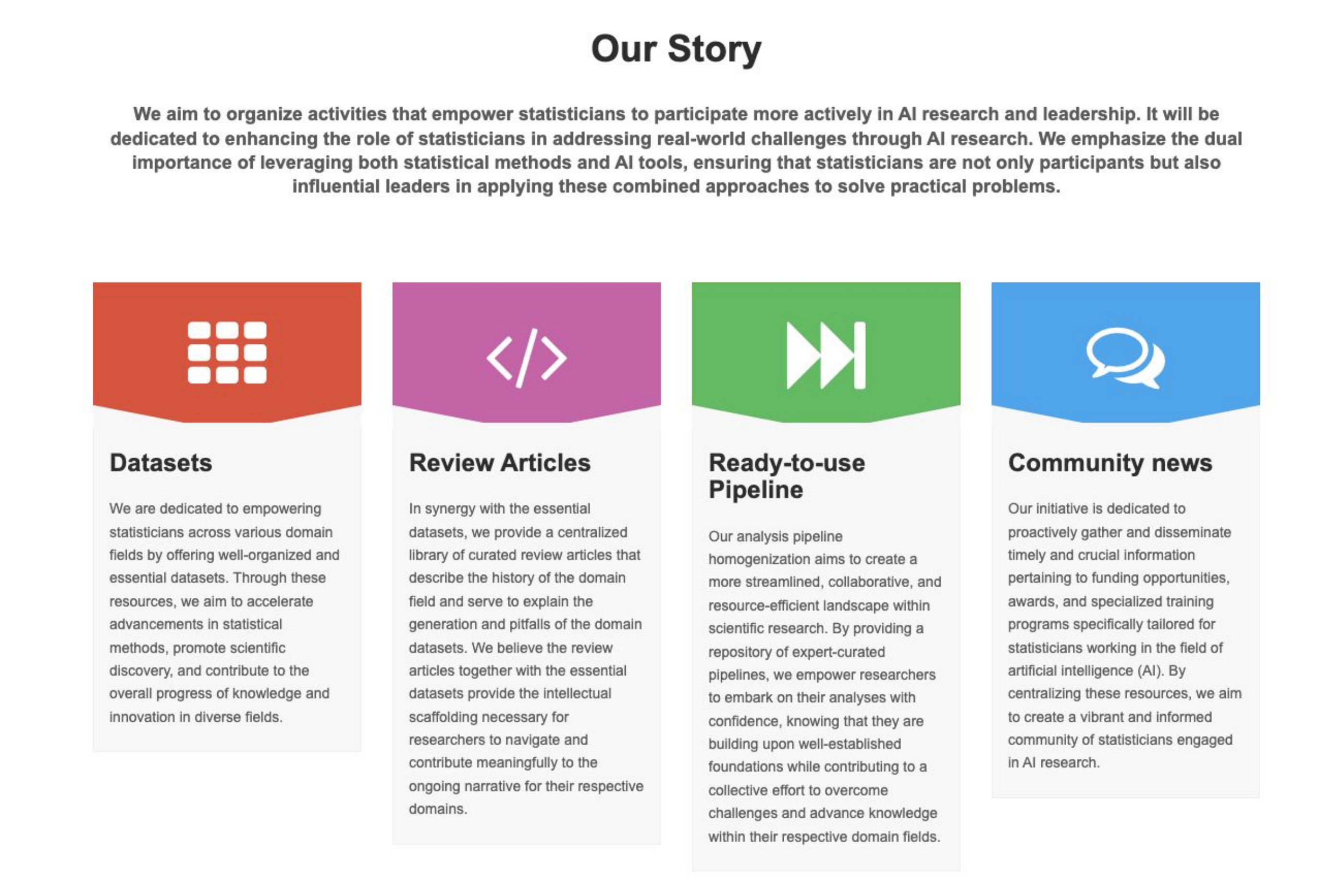
We’ve got talents! - Weijie Su (UPenn)

Stats Up AI Alliance

Actions! - Wenyi Wang (MD Anderson)

Stats Up AI Alliance

Stats Up AI Alliance: https://statsupai.com/index.html



Our Story

We aim to organize activities that empower statisticians to participate more actively in AI research and leadership. It will be dedicated to enhancing the role of statisticians in addressing real-world challenges through AI research. We emphasize the dual importance of leveraging both statistical methods and AI tools, ensuring that statisticians are not only participants but also influential leaders in applying these combined approaches to solve practical problems.

Datasets

We are dedicated to empowering statisticians across various domain fields by offering well-organized and essential datasets. Through these resources, we aim to accelerate advancements in statistical methods, promote scientific discovery, and contribute to the overall progress of knowledge and innovation in diverse fields.

Review Articles

In synergy with the essential datasets, we provide a centralized library of curated review articles that describe the history of the domain field and serve to explain the generation and pitfalls of the domain datasets. We believe the review articles together with the essential datasets provide the intellectual scaffolding necessary for researchers to navigate and contribute meaningfully to the ongoing narrative for their respective domains.

Ready-to-use Pipeline

Our analysis pipeline homogenization aims to create a more streamlined, collaborative, and resource-efficient landscape within scientific research. By providing a repository of expert-curated pipelines, we empower researchers to embark on their analyses with confidence, knowing that they are building upon well-established foundations while contributing to a collective effort to overcome challenges and advance knowledge within their respective domain fields.

Community news

Our initiative is dedicated to proactively gather and disseminate timely and crucial information pertaining to funding opportunities, awards, and specialized training programs specifically tailored for statisticians working in the field of artificial intelligence (AI). By centralizing these resources, we aim to create a vibrant and informed community of statisticians engaged in AI research.

UCI

Hanwen Ye

Annie Qu

U of Michigan

Bangyao Zhao

Jian Kang

Yunnan University

Shan Gao

UNC Chapel Hill

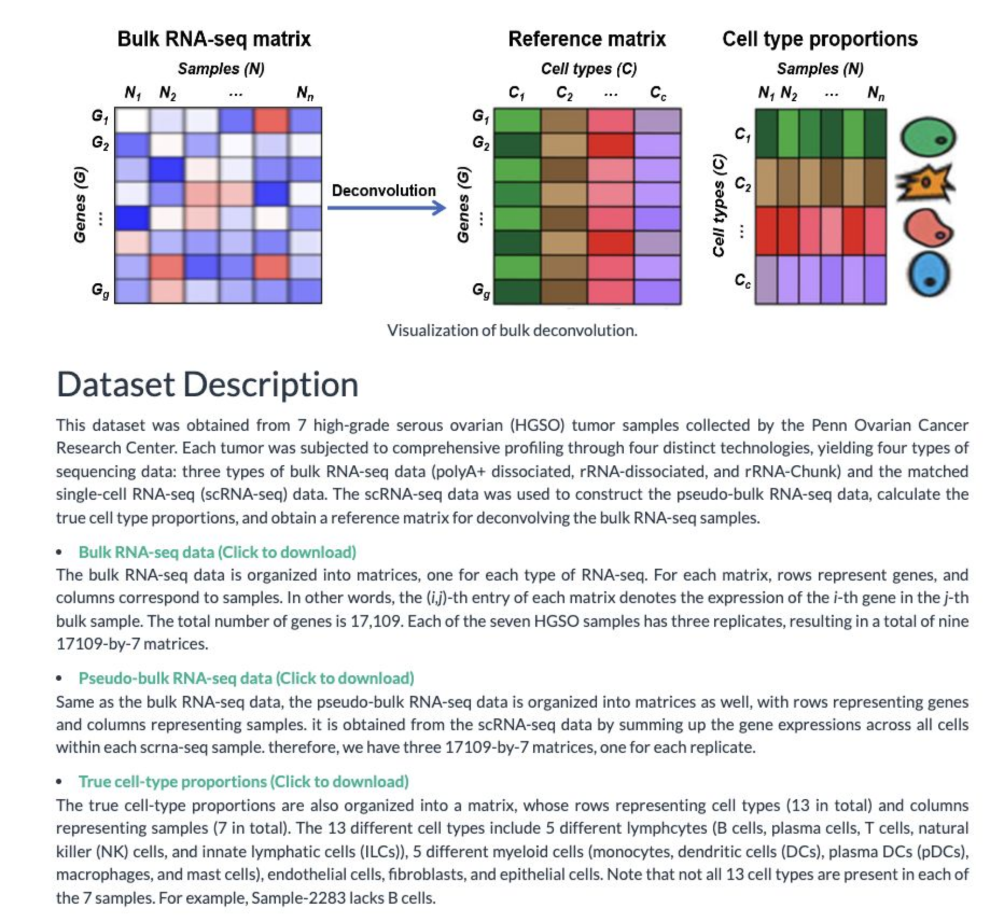
Hongtu Zhu

MD Anderson Cancer Center

Xiaoqian Liu

Wenyi Wang

Curated and organized datasets for methods development



Visualization of bulk deconvolution.

Dataset Description

This dataset was obtained from 7 high-grade serous ovarian (HGSO) tumor samples collected by the Penn Ovarian Cancer Research Center. Each tumor was subjected to comprehensive profiling through four distinct technologies, yielding four types of sequencing data: three types of bulk RNA-seq data (polyA+ dissociated, rRNA-dissociated, and rRNA-Chunk) and the matched single-cell RNA-seq (scRNA-seq) data. The scRNA-seq data was used to construct the pseudo-bulk RNA-seq data, calculate the true cell type proportions, and obtain a reference matrix for deconvolving the bulk RNA-seq samples.

- **Bulk RNA-seq data (Click to download)**

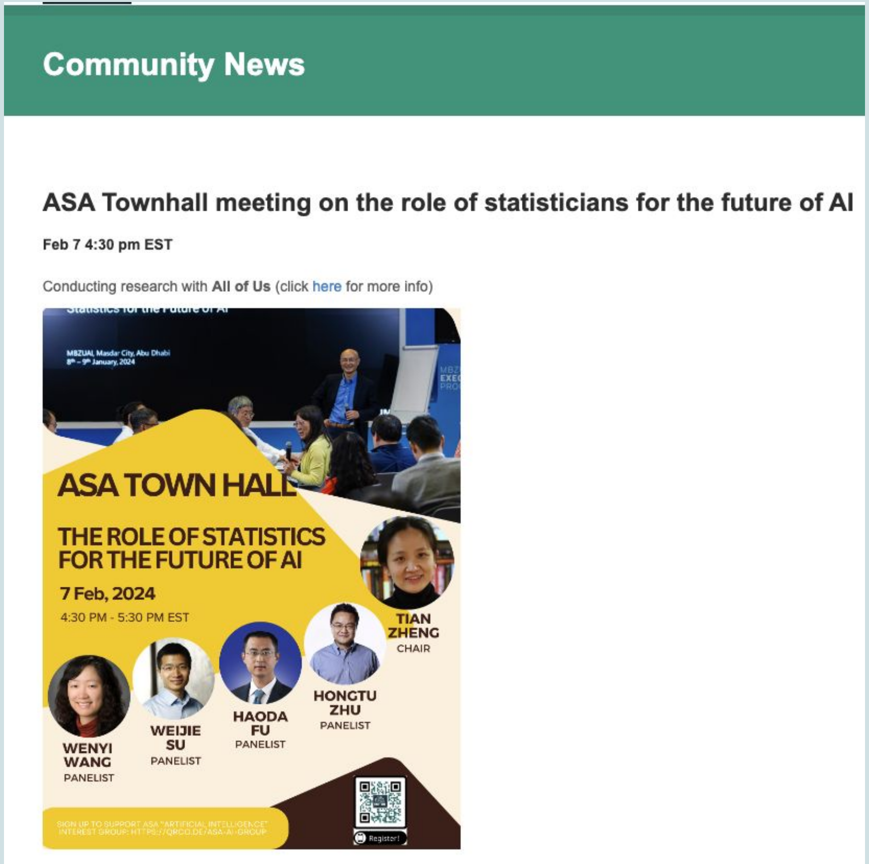
The bulk RNA-seq data is organized into matrices, one for each type of RNA-seq. For each matrix, rows represent genes, and columns correspond to samples. In other words, the (*i*,*j*)-th entry of each matrix denotes the expression of the *i*-th gene in the *j*-th bulk sample. The total number of genes is 17,109. Each of the seven HGSO samples has three replicates, resulting in a total of nine 17109-by-7 matrices.

- **Pseudo-bulk RNA-seq data (Click to download)**

Same as the bulk RNA-seq data, the pseudo-bulk RNA-seq data is organized into matrices as well, with rows representing genes and columns representing samples. it is obtained from the scRNA-seq data by summing up the gene expressions across all cells within each scrna-seq sample. therefore, we have three 17109-by-7 matrices, one for each replicate.

- **True cell-type proportions (Click to download)**

The true cell-type proportions are also organized into a matrix, whose rows representing cell types (13 in total) and columns representing samples (7 in total). The 13 different cell types include 5 different lymphcytes (B cells, plasma cells, T cells, natural killer (NK) cells, and innate lymphatic cells (ILCs)), 5 different myeloid cells (monocytes, dendritic cells (DCs), plasma DCs (pDCs), macrophages, and mast cells), endothelial cells, fibroblasts, and epithelial cells. Note that not all 13 cell types are present in each of the 7 samples. For example, Sample-2283 lacks B cells.



Community News

ASA Townhall meeting on the role of statisticians for the future of AI

Feb 7 4:30 pm EST

Conducting research with **All of Us** (click here for more info)

Statistics for the Future of AI
MBZUAI, Masdar City, Abu Dhabi
8th – 9th January, 2024
ASA TOWN HALL
THE ROLE OF STATISTICS FOR THE FUTURE OF AI
7 Feb, 2024
4:30 PM - 5:30 PM EST
TIAN ZHENG
CHAIR
HONGTU ZHU
PANELIST
HAODA FU
PANELIST
WEIJIE SU
PANELIST
WENYI WANG
PANELIST
SIGN UP TO SUPPORT ASA "ARTIFICIAL INTELLIGENCE" INTEREST GROUP: HTTPS://QRCO.DE/ASA-AI-GROUP
Register!

Videos from January 2024 Workshop

Q&A

Stats Up AI Alliance

Selected questions submitted before the Town Hall

- What do you see as the key step(s) for ensuring that statisticians are more than an afterthought in an AI driven world?
- How or can AI methodologies replace statistical methodologies. Isn't Statistics the foundation of AI?
- I’d like to hear about the update and current challenges for AI and genomics.
- What are new skill sets I should teach students in this AI era?
- What are the most important use cases for AI when it comes to clinical trials and statistics?
- What are your thoughts on the role of uncertainty in AI?
- **What is the unique edge for statisticians during this AI revolution?**