DART

Data Analytics that are Robust and Trusted

Year 3 All Hands Meeting
April 20, 2023

NSF ARKANSAS EPSCoR

Project Overview

9:30 – 10:30 AM

Jennifer Fowler, PI
Arkansas Economic Development Commission

Jackson Cothren, Co-PI
University of Arkansas, Fayetteville

This material is based upon work supported by the National Science Foundation under Award No. OIA-1946391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The motivation for DART

WIRED

So, Arkansas Is Leading the Learn to Code Movement

Arkansas may be one of the last states that comes to mind when you think of major hubs of tech talent.

How Arkansas implemented its computer science education program

Tyson

Walmart

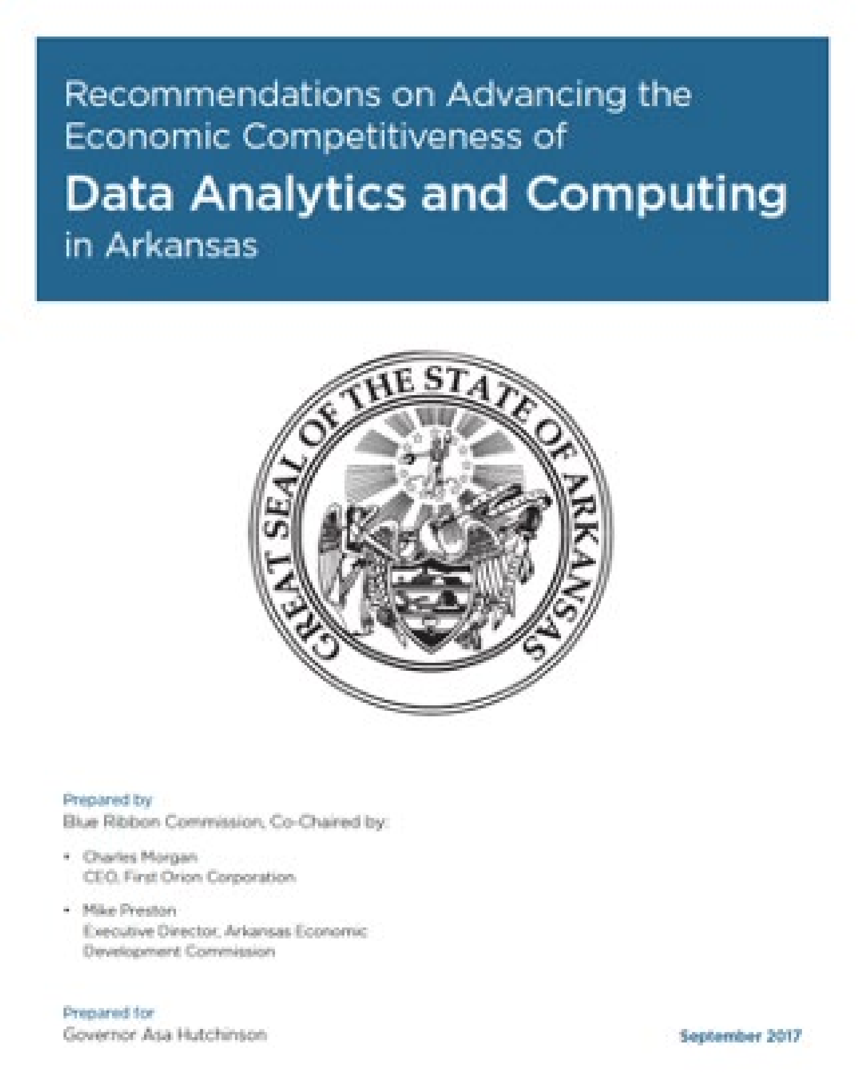
Stephens Inc.

acxiom

J.B. HUNT

FIS

MURPHY OIL CORPORATION



Recommendations on Advancing the Economic Competitiveness of

Data Analytics and Computing

in Arkansas

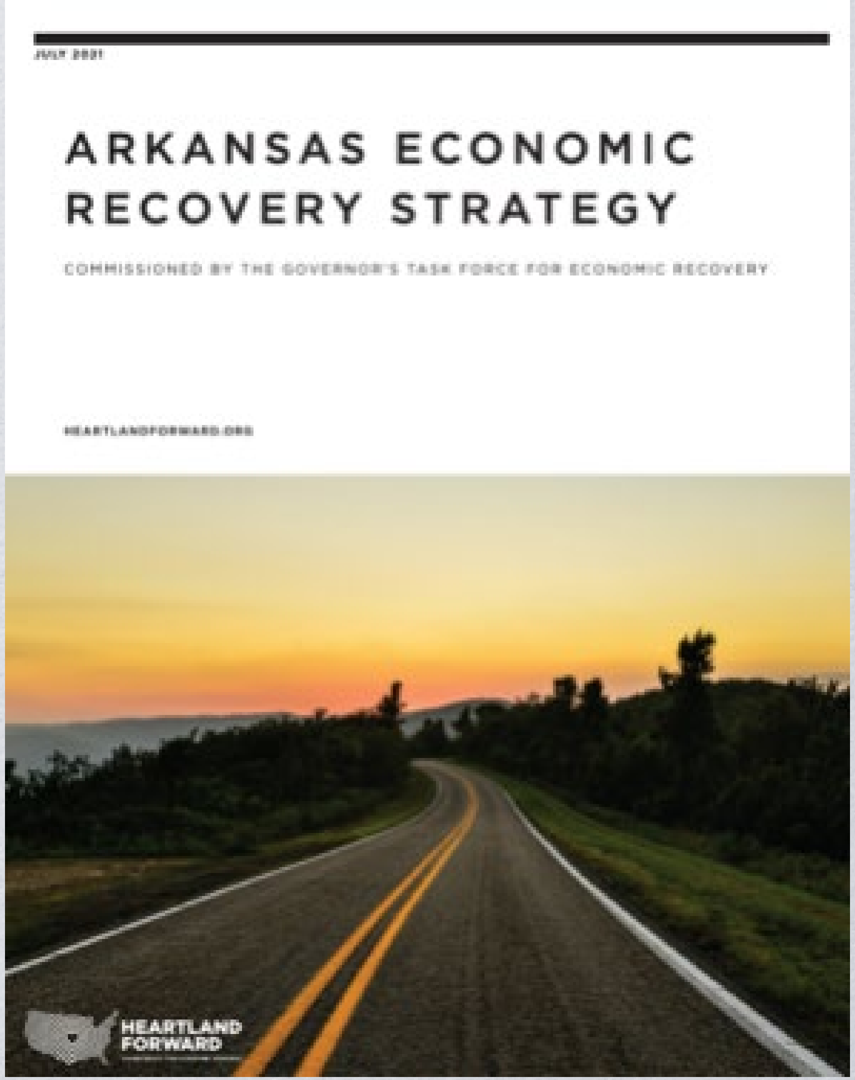
GREAT SEAL OF THE STATE OF ARKANSAS

Prepared by
Blue Ribbon Commission, Co-Chaired by:

- Charles Morgan
  CEO, First Orion Corporation
- Mike Preston
  Executive Director, Arkansas Economic Development Commission

Prepared for
Governor Asa Hutchinson

September 2017



JULY 2021

ARKANSAS ECONOMIC RECOVERY STRATEGY

COMMISSIONED BY THE GOVERNOR'S TASK FORCE FOR ECONOMIC RECOVERY

HEARTLANDFORWARD.ORG

HEARTLAND FORWARD

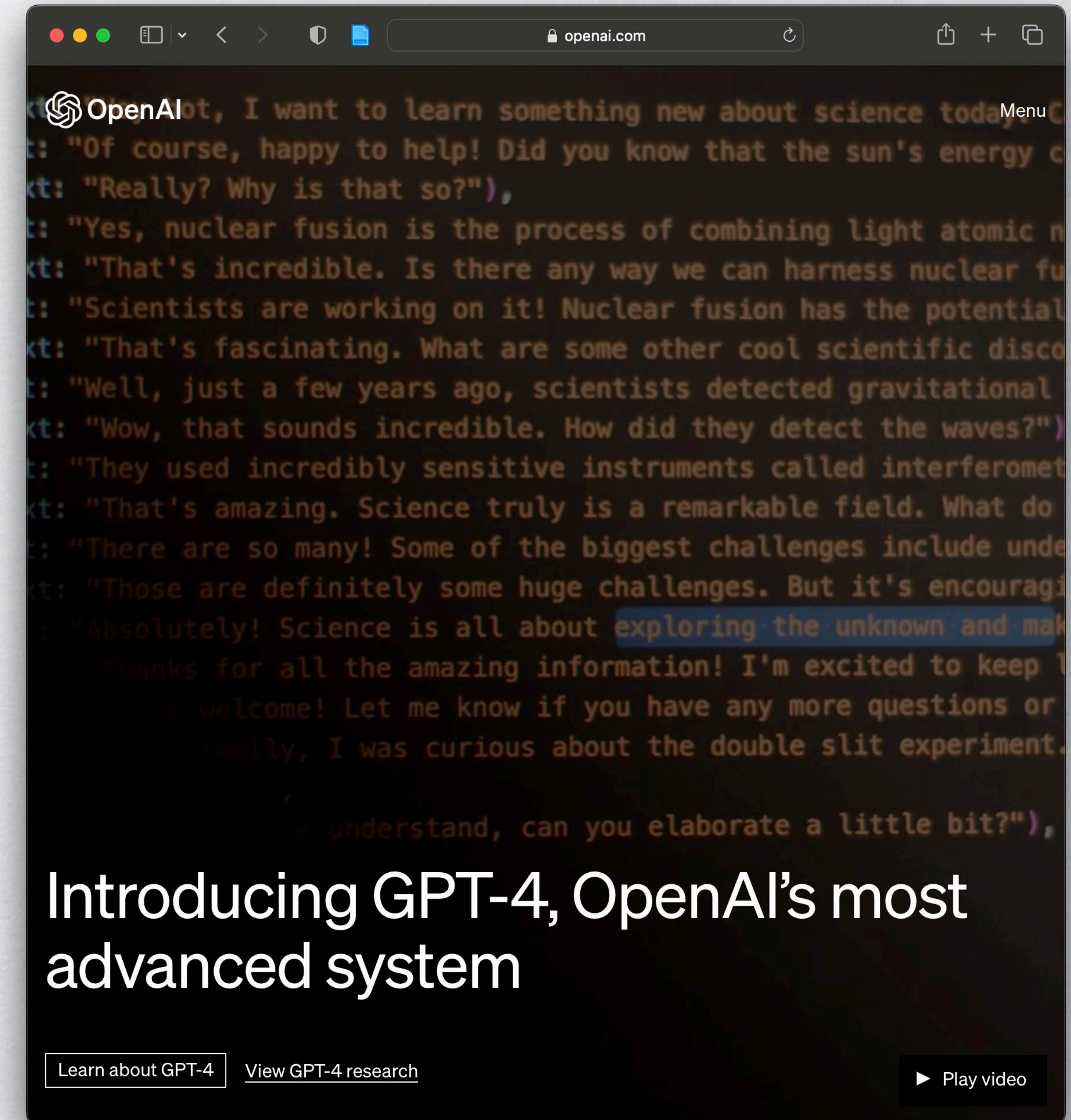
Vision

The Arkansas research community - academic, government, and industry - collaborating often and easily on a shared computing platform through a modern cyberinfrastructure which facilitates cutting-edge data science research.

Mission

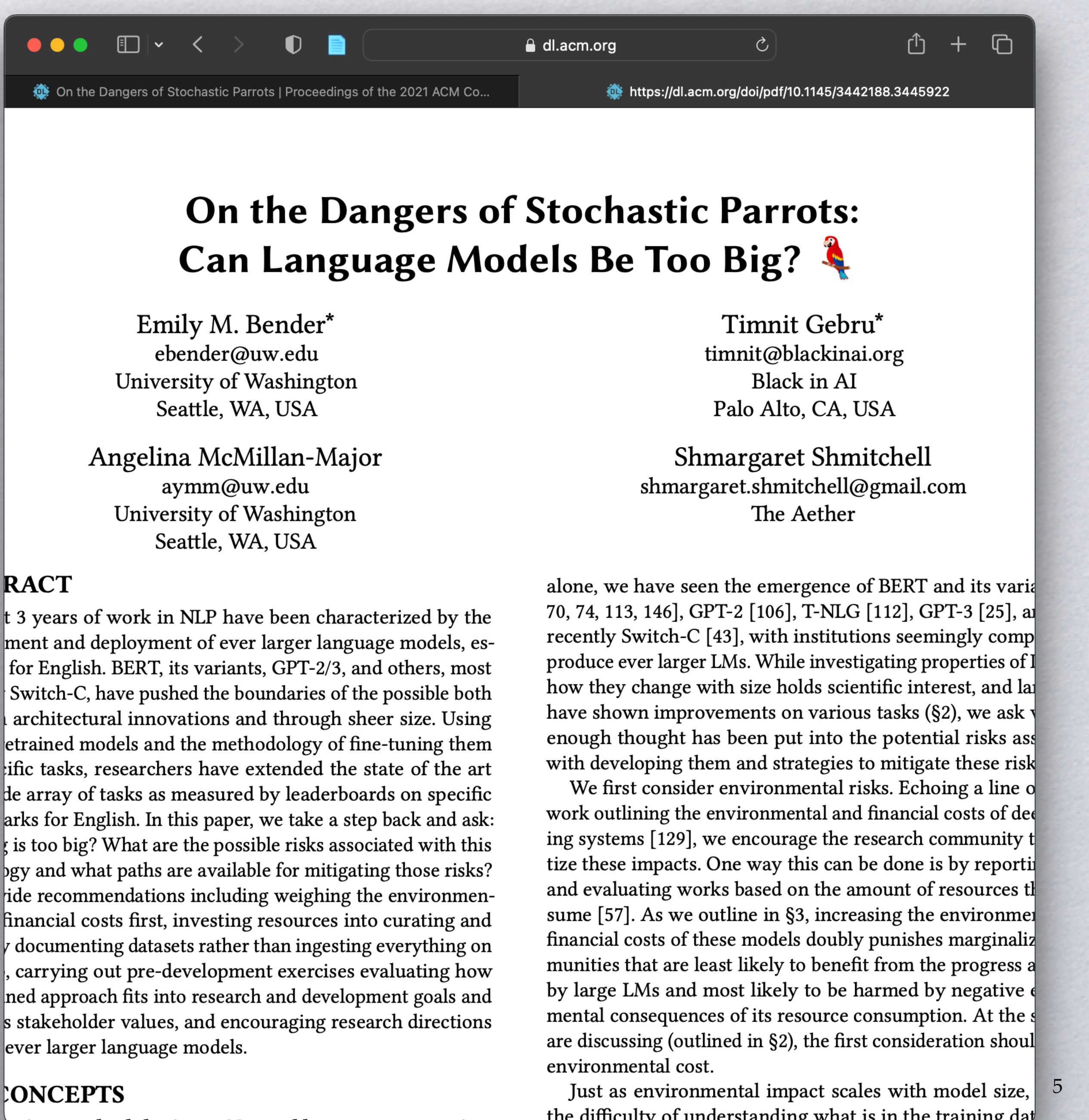
The mission of DART is to improve research capability and competitiveness in Arkansas by creating an integrated statewide consortium of researchers and educators working to establish a synergistic, statewide focus on excellence in data analytics research and training.

The continued (growing) relevance of DART



```
t, I want to learn something new about science toda
t: "Of course, happy to help! Did you know that the sun's energy c
xt: "Really? Why is that so?"),
t: "Yes, nuclear fusion is the process of combining light atomic n
xt: "That's incredible. Is there any way we can harness nuclear fu
t: "Scientists are working on it! Nuclear fusion has the potential
xt: "That's fascinating. What are some other cool scientific disco
t: "Well, just a few years ago, scientists detected gravitational
xt: "Wow, that sounds incredible. How did they detect the waves?")
t: "They used incredibly sensitive instruments called interferomet
xt: "That's amazing. Science truly is a remarkable field. What do
t: "There are so many! Some of the biggest challenges include unde
xt: "Those are definitely some huge challenges. But it's encouragi
 "Absolutely! Science is all about exploring the unknown and mak
  "Thanks for all the amazing information! I'm excited to keep l
     welcome! Let me know if you have any more questions or
        ally, I was curious about the double slit experiment.

          understand, can you elaborate a little bit?"),
```

Introducing GPT-4, OpenAI’s most advanced system


dl.acm.org
On the Dangers of Stochastic Parrots | Proceedings of the 2021 ACM Co...
https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

RACT

t 3 years of work in NLP have been characterized by the
ment and deployment of ever larger language models, es-
for English. BERT, its variants, GPT-2/3, and others, most
Switch-C, have pushed the boundaries of the possible both
architectural innovations and through sheer size. Using
etrained models and the methodology of fine-tuning them
ific tasks, researchers have extended the state of the art
de array of tasks as measured by leaderboards on specific
arks for English. In this paper, we take a step back and ask:
is too big? What are the possible risks associated with this
gy and what paths are available for mitigating those risks?
ide recommendations including weighing the environmen-
financial costs first, investing resources into curating and
documenting datasets rather than ingesting everything on
, carrying out pre-development exercises evaluating how
ned approach fits into research and development goals and
s stakeholder values, and encouraging research directions
ever larger language models.

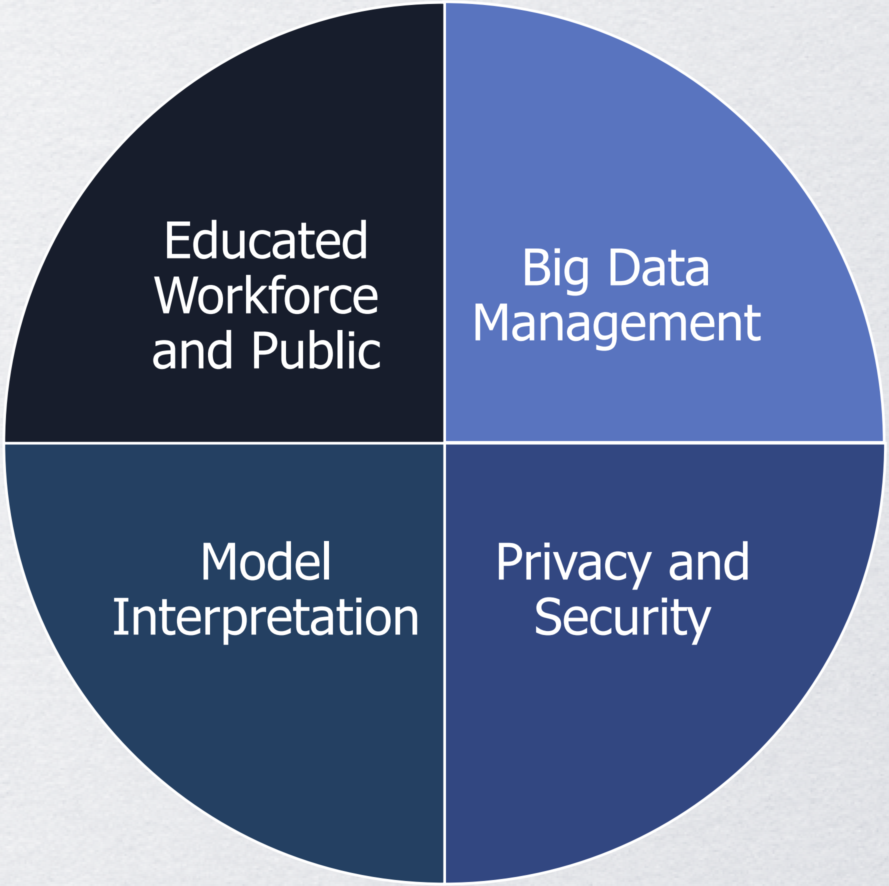
CONCEPTS

alone, we have seen the emergence of BERT and its varia
70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], a
recently Switch-C [43], with institutions seemingly comp
produce ever larger LMs. While investigating properties of
how they change with size holds scientific interest, and la
have shown improvements on various tasks (§2), we ask
enough thought has been put into the potential risks ass
with developing them and strategies to mitigate these risk

We first consider environmental risks. Echoing a line o
work outlining the environmental and financial costs of de
ing systems [129], we encourage the research community t
tize these impacts. One way this can be done is by reporti
and evaluating works based on the amount of resources t
sume [57]. As we outline in §3, increasing the environme
financial costs of these models doubly punishes marginaliz
munities that are least likely to benefit from the progress a
by large LMs and most likely to be harmed by negative e
mental consequences of its resource consumption. At the s
are discussing (outlined in §2), the first consideration shoul
environmental cost.

Just as environmental impact scales with model size,
the difficulty of understanding what is in the training dat

How DART is organized



Educated Workforce and Public

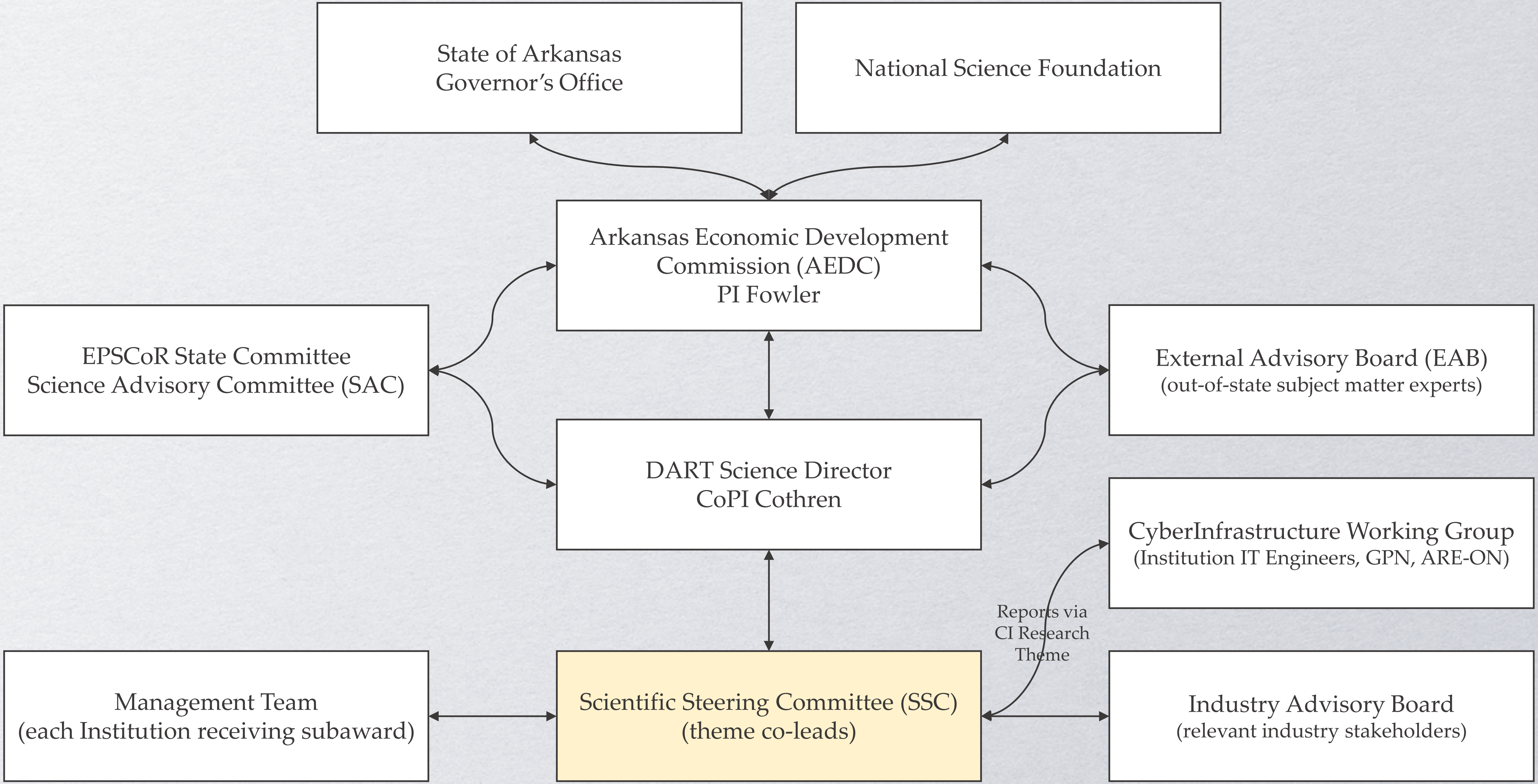
Big Data Management

Model Interpretation

Privacy and Security

- +Addressing barriers to effective and accepted integration of data science into our lives.
- +Building a shared research cyberinfrastructure in Arkansas
- +Creating a state-wide, latticed data science education program
- +Fostering a collaborative research community in Arkansas

Project Management



State of Arkansas
Governor's Office

National Science Foundation

Arkansas Economic Development Commission (AEDC)
PI Fowler

EPSCoR State Committee
Science Advisory Committee (SAC)

External Advisory Board (EAB)
(out-of-state subject matter experts)

DART Science Director
CoPI Cothren

CyberInfrastructure Working Group
(Institution IT Engineers, GPN, ARE-ON)

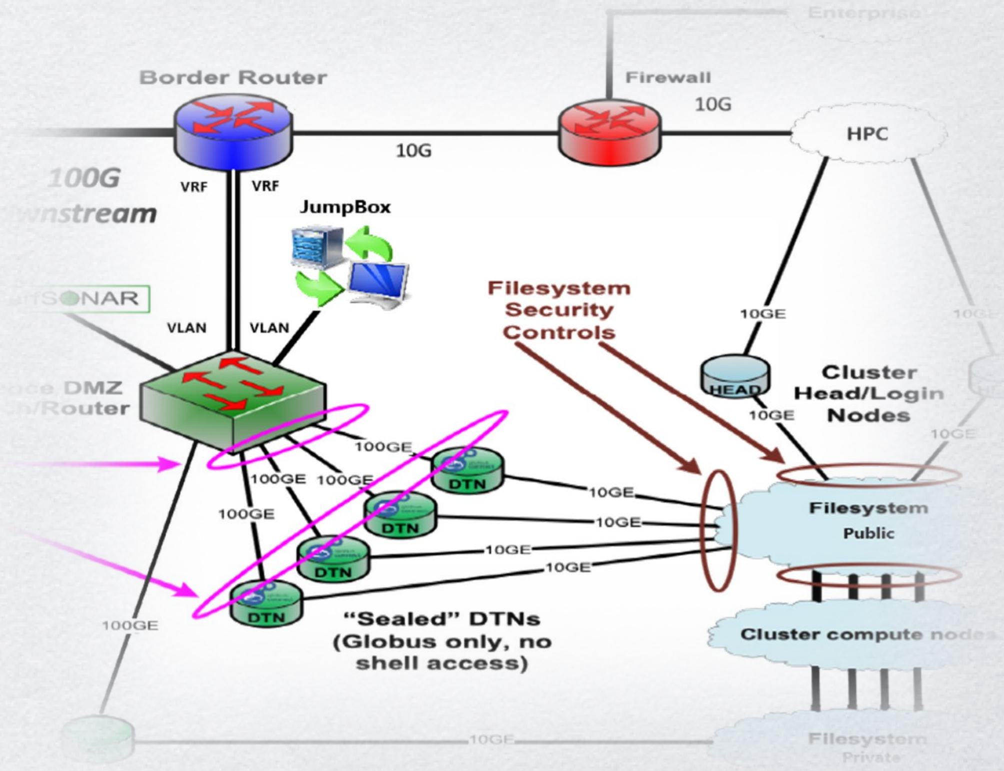
Reports via
CI Research
Theme

Management Team
(each Institution receiving subaward)

Scientific Steering Committee (SSC)
(theme co-leads)

Industry Advisory Board
(relevant industry stakeholders)

Significant Challenges



Enhanced cybersecurity footprint

- Increased oversight
- Off-campus access to research computing more controlled
- ScienceDMZ architecture reconfigured
  - 15-month planning process
  - Limited access to cluster and storage
  - Limited access to Enterprise GitLab

Solution: Build enduring relationships between campus IT and research communities

Significant Challenges

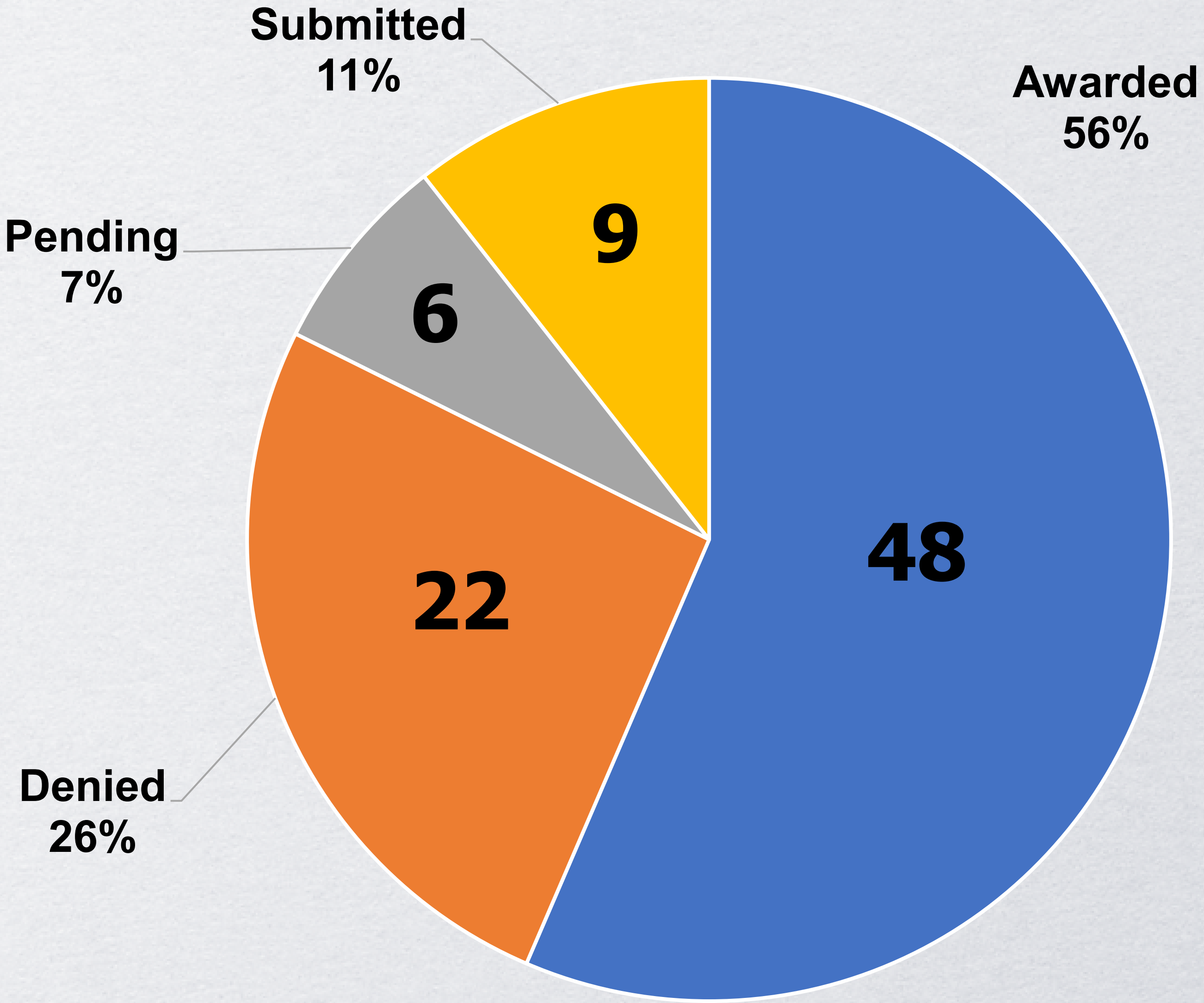
```
        .color = colors        .mags_dt[i]
    def on_key_press(self, symbol, modifiers):
        """Delegate key press"""
        if self.context_index == -1:
            if symbol == key.UP and not self.active_index == 0:
                self.menu_labels[self.active_index].color = [255, 255, 255, 255]
                self.active_index -= 1
                self.mags_dt = self.get_act_color_mag()
            elif symbol == key.DOWN and not self.active_index == 3:
                self.menu_labels[self.active_index].color = [255, 255, 255, 255]
                self.active_index += 1
                self.mags_dt = self.get_act_color_mag()
            elif symbol == key.ENTER:
                if self.active_index == 3:
                    pyglet.app.exit()
                else:
                    self.context_index = self.active_index
            elif symbol == key.ESCAPE:
                if self.context_index == -1:
                    pyglet.app.exit()
                else:
                    self.context_index = -1
        elif self.context_index == 1:
            if symbol == key.ESCAPE:
                self.context_index = -1
            else:
                self.cur_game.on_key_press(symbol, modifiers)

        else:
            if symbol == key.ESCAPE:
                self.context_index = -1
```

Research themes identified a need for additional programming support

- several projects need to scale
- deploying code on ARP clusters and commercial cloud
- 1 FTE currently supporting this need, more talent is needed
- Keeping talent, competing with industry, is hard

Solution: Reallocate program funds to support diverse programming requirements.

Significant Accomplishments – Grants & Proposals



$ 24,389,199.41
Awarded to Date

Significant Accomplishments – Grants & Proposals

Project Team(s)	Amount Awarded
Cyberinfrastructure (CI)	$ 884,318
Data Curation (DC)	$ 7,176,848
Learning & Prediction (LP)	$ 478,763
Learning & Prediction (LP), Cyberinfrastructure (CI)	$ 895,000
Learning & Prediction (LP), Data Curation (DC)	$ 155,199
Learning & Prediction (LP), Data Curation (DC), Social Awareness (SA)	$ 1,450,003
Social Awareness (SA)	$ 1,463,789
Social Awareness (SA), Workforce Development (WD)	$ 49,999
Social Media (SM)	$ 11,835,280

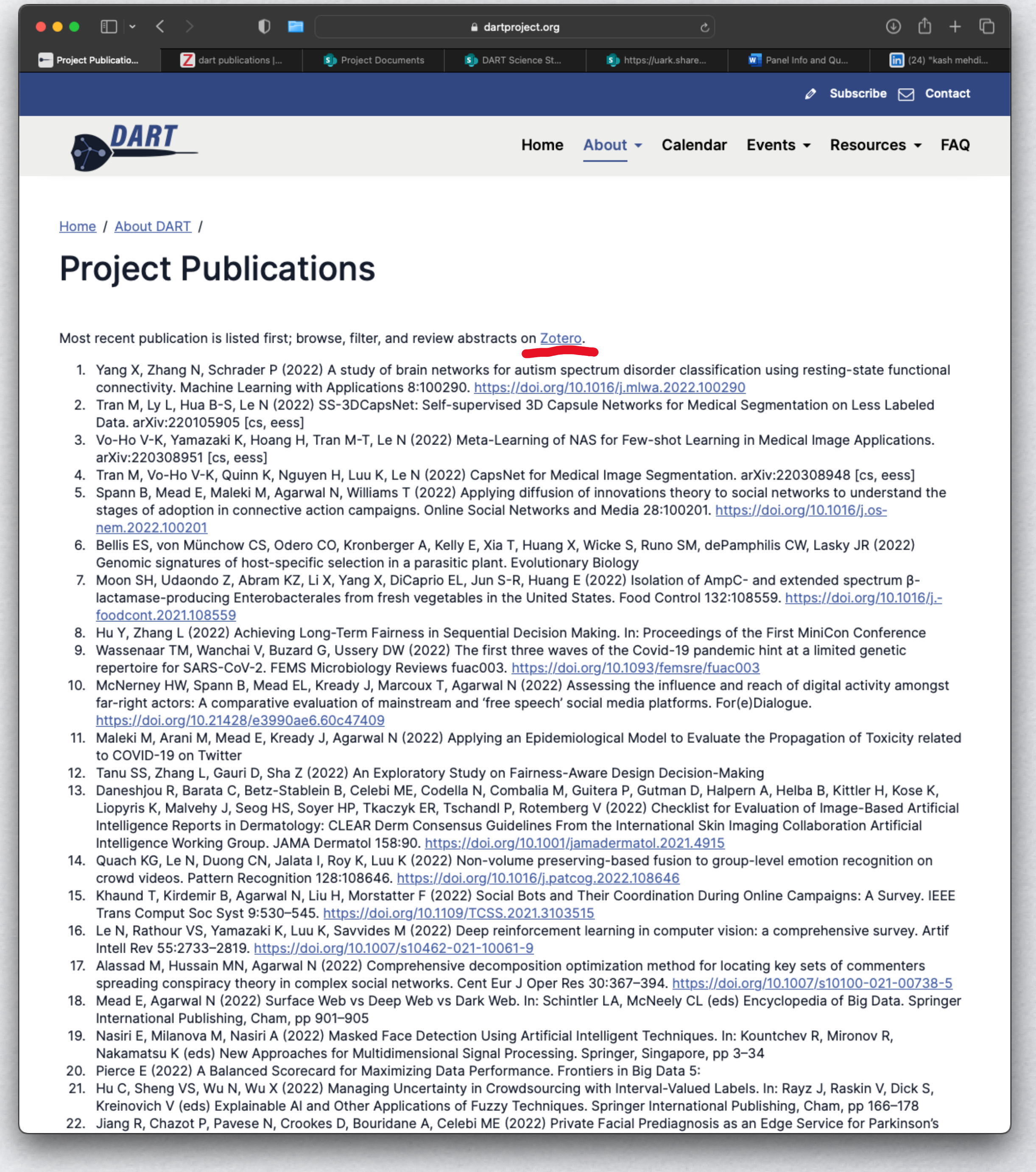
Significant Accomplishments – Grants > $500,000

Title	Investigators	Institutions	Award Amount	Funding Agency
Multi-Level Models of Covert Online Information Campaigns	Nitin Agarwal	UALR	$ 4,965,214	Department of Defense
Developing Rapid Response Capabilities to Evaluate Emerging Social Cyber Threats	Nitin Agarwal	UALR	$ 3,773,526	U.S. Office of Naval Research
Fusing Narrative and Social Cyber Forensics to Understand Covert Influence	Nitin Agarwal	UALR	$ 2,500,000	Department of Defense
Targeting heat shock protein 72 to improve renal function after transplantation	Se-Ran Jun	UAMS	$ 2,461,707	NIDDK
Center for studies of host response to cancer therapy	Se-Ran Jun	UAMS	$ 2,280,000	NIGMS
RII Track-2 FEC: Artificial Intelligence on Sustainable Energy Infrastructure Network (AI SUSTEIN) and Beyond towards Industries of the Future	Haitao Liao, Xintao Wu, Xiao Liu	UARK	$ 1,450,003	NSF
Assessment of antibiotic resistance in fresh vegetables from farm to fork	Se-Ran Jun	UAMS	$ 1,000,000	USDA
Photogrammetry Services, Task Order	Jackson Cothren, Chase Rainwater	UARK	$ 800,000	Department of Energy
FAI: A novel paradigm for fairness-aware deep learning models on data streams	Xintao Wu	UARK	$ 628,789	NSF
IUCRC Phase I The University of Arkansas: Center for Infrastructure Trustworthiness in Energy Systems (CITES)	Qinghua Li	UARK	$ 525,000	NSF

Significant Accomplishments - Publications

Publication Status	Project Total
Accepted	42
Published	140
Submitted	25

Project Team	Total Published	Journal Articles	Books	Conference Proceedings
Cyberinfrastructure (CI)	2	2	0	0
Cyberinfrastructure (CI), Data Curation (DC)	2	1	1	0
Data Curation (DC)	29	19	2	6
Education (ED)	1	0	0	0
Learning & Prediction (LP)	37	18	1	18
Learning & Prediction (LP), Cyberinfrastructure (CI)	1	1	0	0
Social Awareness (SA)	30	5	5	25
Social Media (SM)	23	11	0	11

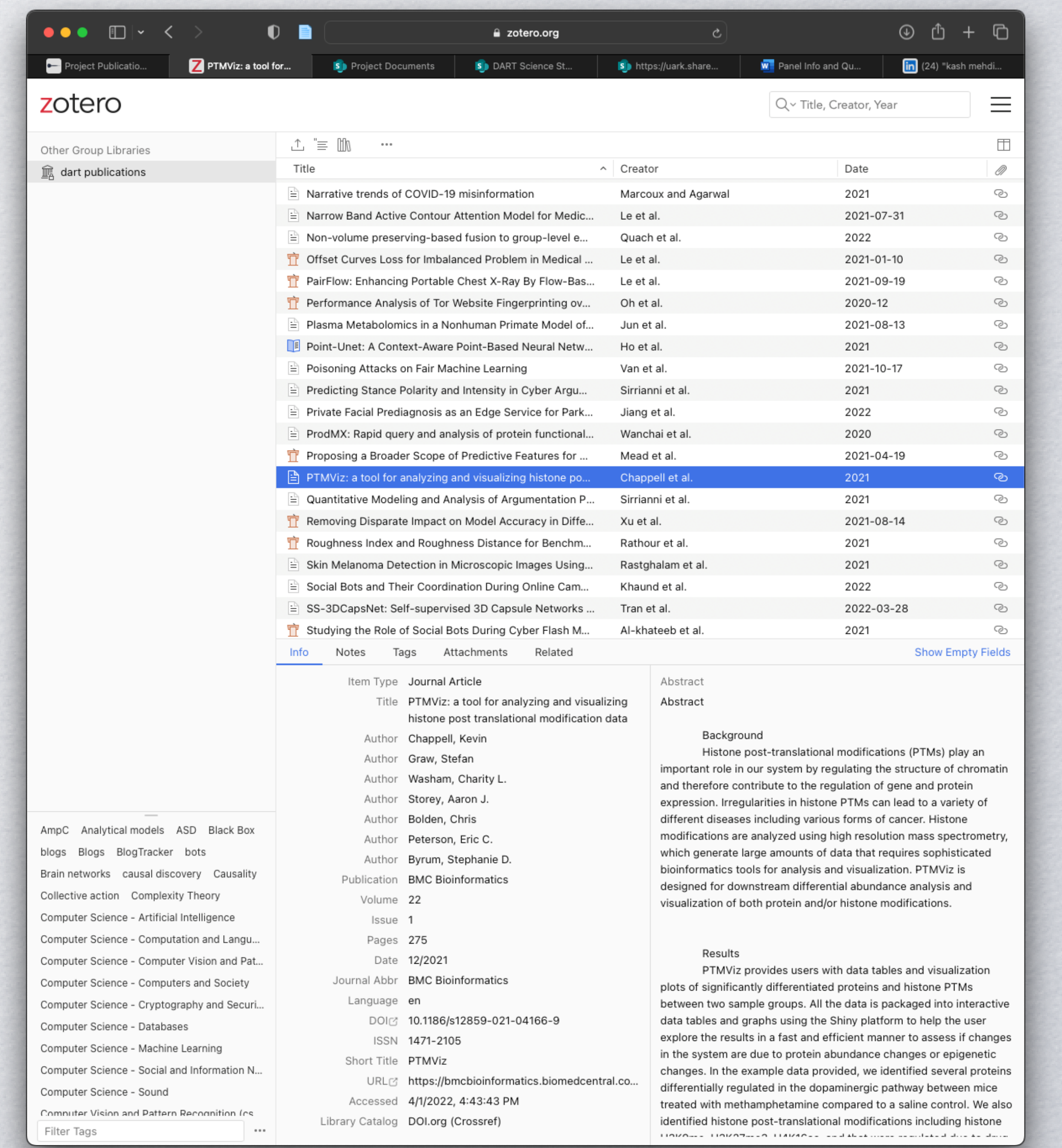


Home / About DART /

Project Publications

Most recent publication is listed first; browse, filter, and review abstracts on Zotero.

1. Yang X, Zhang N, Schrader P (2022) A study of brain networks for autism spectrum disorder classification using resting-state functional connectivity. Machine Learning with Applications 8:100290. https://doi.org/10.1016/j.mlwa.2022.100290
2. Tran M, Ly L, Hua B-S, Le N (2022) SS-3DCapsNet: Self-supervised 3D Capsule Networks for Medical Segmentation on Less Labeled Data. arXiv:220105905 [cs, eess]
3. Vo-Ho V-K, Yamazaki K, Hoang H, Tran M-T, Le N (2022) Meta-Learning of NAS for Few-shot Learning in Medical Image Applications. arXiv:220308951 [cs, eess]
4. Tran M, Vo-Ho V-K, Quinn K, Nguyen H, Luu K, Le N (2022) CapsNet for Medical Image Segmentation. arXiv:220308948 [cs, eess]
5. Spann B, Mead E, Maleki M, Agarwal N, Williams T (2022) Applying diffusion of innovations theory to social networks to understand the stages of adoption in connective action campaigns. Online Social Networks and Media 28:100201. https://doi.org/10.1016/j.osnem.2022.100201
6. Bellis ES, von Münchow CS, Odero CO, Kronberger A, Kelly E, Xia T, Huang X, Wicke S, Runo SM, dePamphilis CW, Lasky JR (2022) Genomic signatures of host-specific selection in a parasitic plant. Evolutionary Biology
7. Moon SH, Udaondo Z, Abram KZ, Li X, Yang X, DiCaprio EL, Jun S-R, Huang E (2022) Isolation of AmpC- and extended spectrum β-lactamase-producing Enterobacterales from fresh vegetables in the United States. Food Control 132:108559. https://doi.org/10.1016/j.foodcont.2021.108559
8. Hu Y, Zhang L (2022) Achieving Long-Term Fairness in Sequential Decision Making. In: Proceedings of the First MiniCon Conference
9. Wassenaar TM, Wanchai V, Buzard G, Ussery DW (2022) The first three waves of the Covid-19 pandemic hint at a limited genetic repertoire for SARS-CoV-2. FEMS Microbiology Reviews fuac003. https://doi.org/10.1093/femsre/fuac003
10. McNerney HW, Spann B, Mead EL, Kready J, Marcoux T, Agarwal N (2022) Assessing the influence and reach of digital activity amongst far-right actors: A comparative evaluation of mainstream and 'free speech' social media platforms. For(e)Dialogue. https://doi.org/10.21428/e3990ae6.60c47409
11. Maleki M, Arani M, Mead E, Kready J, Agarwal N (2022) Applying an Epidemiological Model to Evaluate the Propagation of Toxicity related to COVID-19 on Twitter
12. Tanu SS, Zhang L, Gauri D, Sha Z (2022) An Exploratory Study on Fairness-Aware Design Decision-Making
13. Daneshjou R, Barata C, Betz-Stablein B, Celebi ME, Codella N, Combalia M, Guitera P, Gutman D, Halpern A, Helba B, Kittler H, Kose K, Liopyris K, Malvehy J, Seog HS, Soyer HP, Tkaczyk ER, Tschandl P, Rotemberg V (2022) Checklist for Evaluation of Image-Based Artificial Intelligence Reports in Dermatology: CLEAR Derm Consensus Guidelines From the International Skin Imaging Collaboration Artificial Intelligence Working Group. JAMA Dermatol 158:90. https://doi.org/10.1001/jamadermatol.2021.4915
14. Quach KG, Le N, Duong CN, Jalata I, Roy K, Luu K (2022) Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. Pattern Recognition 128:108646. https://doi.org/10.1016/j.patcog.2022.108646
15. Khaund T, Kirdemir B, Agarwal N, Liu H, Morstatter F (2022) Social Bots and Their Coordination During Online Campaigns: A Survey. IEEE Trans Comput Soc Syst 9:530–545. https://doi.org/10.1109/TCSS.2021.3103515
16. Le N, Rathour VS, Yamazaki K, Luu K, Savvides M (2022) Deep reinforcement learning in computer vision: a comprehensive survey. Artif Intell Rev 55:2733–2819. https://doi.org/10.1007/s10462-021-10061-9
17. Alassad M, Hussain MN, Agarwal N (2022) Comprehensive decomposition optimization method for locating key sets of commenters spreading conspiracy theory in complex social networks. Cent Eur J Oper Res 30:367–394. https://doi.org/10.1007/s10100-021-00738-5
18. Mead E, Agarwal N (2022) Surface Web vs Deep Web vs Dark Web. In: Schintler LA, McNeely CL (eds) Encyclopedia of Big Data. Springer International Publishing, Cham, pp 901–905
19. Nasiri E, Milanova M, Nasiri A (2022) Masked Face Detection Using Artificial Intelligent Techniques. In: Kountchev R, Mironov R, Nakamatsu K (eds) New Approaches for Multidimensional Signal Processing. Springer, Singapore, pp 3–34
20. Pierce E (2022) A Balanced Scorecard for Maximizing Data Performance. Frontiers in Big Data 5:
21. Hu C, Sheng VS, Wu N, Wu X (2022) Managing Uncertainty in Crowdsourcing with Interval-Valued Labels. In: Rayz J, Raskin V, Dick S, Kreinovich V (eds) Explainable AI and Other Applications of Fuzzy Techniques. Springer International Publishing, Cham, pp 166–178
22. Jiang R, Chazot P, Pavese N, Crookes D, Bouridane A, Celebi ME (2022) Private Facial Prediagnosis as an Edge Service for Parkinson's



Title	Creator	Date
Narrative trends of COVID-19 misinformation	Marcoux and Agarwal	2021
Narrow Band Active Contour Attention Model for Medic...	Le et al.	2021-07-31
Non-volume preserving-based fusion to group-level e...	Quach et al.	2022
Offset Curves Loss for Imbalanced Problem in Medical ...	Le et al.	2021-01-10
PairFlow: Enhancing Portable Chest X-Ray By Flow-Bas...	Le et al.	2021-09-19
Performance Analysis of Tor Website Fingerprinting ov...	Oh et al.	2020-12
Plasma Metabolomics in a Nonhuman Primate Model of...	Jun et al.	2021-08-13
Point-Unet: A Context-Aware Point-Based Neural Netw...	Ho et al.	2021
Poisoning Attacks on Fair Machine Learning	Van et al.	2021-10-17
Predicting Stance Polarity and Intensity in Cyber Argu...	Sirrianni et al.	2021
Private Facial Prediagnosis as an Edge Service for Park...	Jiang et al.	2022
ProdMX: Rapid query and analysis of protein functional...	Wanchai et al.	2020
Proposing a Broader Scope of Predictive Features for ...	Mead et al.	2021-04-19
PTMViz: a tool for analyzing and visualizing histone po...	Chappell et al.	2021
Quantitative Modeling and Analysis of Argumentation P...	Sirrianni et al.	2021
Removing Disparate Impact on Model Accuracy in Diffe...	Xu et al.	2021-08-14
Roughness Index and Roughness Distance for Benchm...	Rathour et al.	2021
Skin Melanoma Detection in Microscopic Images Using...	Rastghalam et al.	2021
Social Bots and Their Coordination During Online Cam...	Khaund et al.	2022
SS-3DCapsNet: Self-supervised 3D Capsule Networks ...	Tran et al.	2022-03-28
Studying the Role of Social Bots During Cyber Flash M...	Al-khateeb et al.	2021

AmpC Analytical models ASD Black Box
blogs Blogs BlogTracker bots
Brain networks causal discovery Causality
Collective action Complexity Theory
Computer Science - Artificial Intelligence
Computer Science - Computation and Langu...
Computer Science - Computer Vision and Pat...
Computer Science - Computers and Society
Computer Science - Cryptography and Securi...
Computer Science - Databases
Computer Science - Machine Learning
Computer Science - Social and Information N...
Computer Science - Sound
Computer Vision and Pattern Recognition (cs

Item Type	Journal Article
Title	PTMViz: a tool for analyzing and visualizing histone post translational modification data
Author	Chappell, Kevin
Author	Graw, Stefan
Author	Washam, Charity L.
Author	Storey, Aaron J.
Author	Bolden, Chris
Author	Peterson, Eric C.
Author	Byrum, Stephanie D.
Publication	BMC Bioinformatics
Volume	22
Issue	1
Pages	275
Date	12/2021
Journal Abbr	BMC Bioinformatics
Language	en
DOI	10.1186/s12859-021-04166-9
ISSN	1471-2105
Short Title	PTMViz
URL	https://bmcbioinformatics.biomedcentral.co...
Accessed	4/1/2022, 4:43:43 PM
Library Catalog	DOI.org (Crossref)

Abstract

Abstract

Background

Histone post-translational modifications (PTMs) play an important role in our system by regulating the structure of chromatin and therefore contribute to the regulation of gene and protein expression. Irregularities in histone PTMs can lead to a variety of different diseases including various forms of cancer. Histone modifications are analyzed using high resolution mass spectrometry, which generate large amounts of data that requires sophisticated bioinformatics tools for analysis and visualization. PTMViz is designed for downstream differential abundance analysis and visualization of both protein and/or histone modifications.

Results

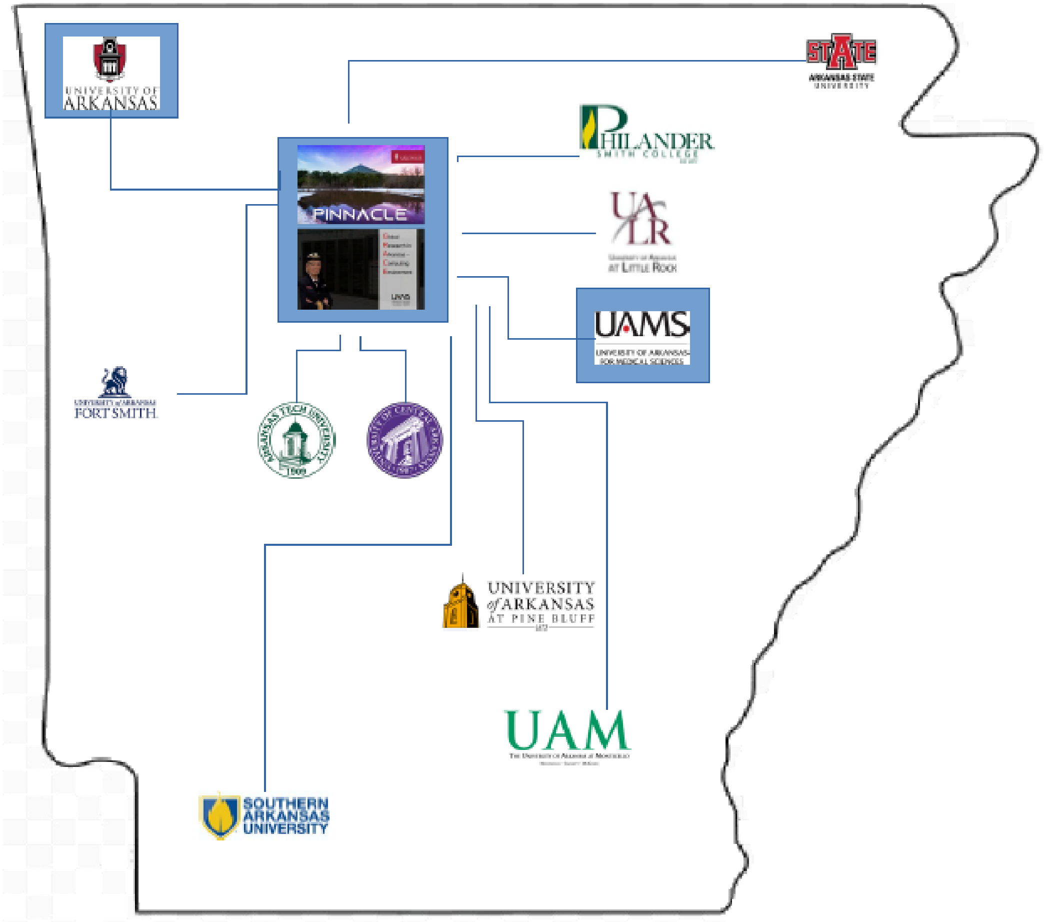
PTMViz provides users with data tables and visualization plots of significantly differentiated proteins and histone PTMs between two sample groups. All the data is packaged into interactive data tables and graphs using the Shiny platform to help the user explore the results in a fast and efficient manner to assess if changes in the system are due to protein abundance changes or epigenetic changes. In the example data provided, we identified several proteins differentially regulated in the dopaminergic pathway between mice treated with methamphetamine compared to a saline control. We also identified histone post-translational modifications including histone

Arkansas Research Platform: Computational Resources

DART

Data Analytics that are Robust and Trusted

Arkansas Research Platform

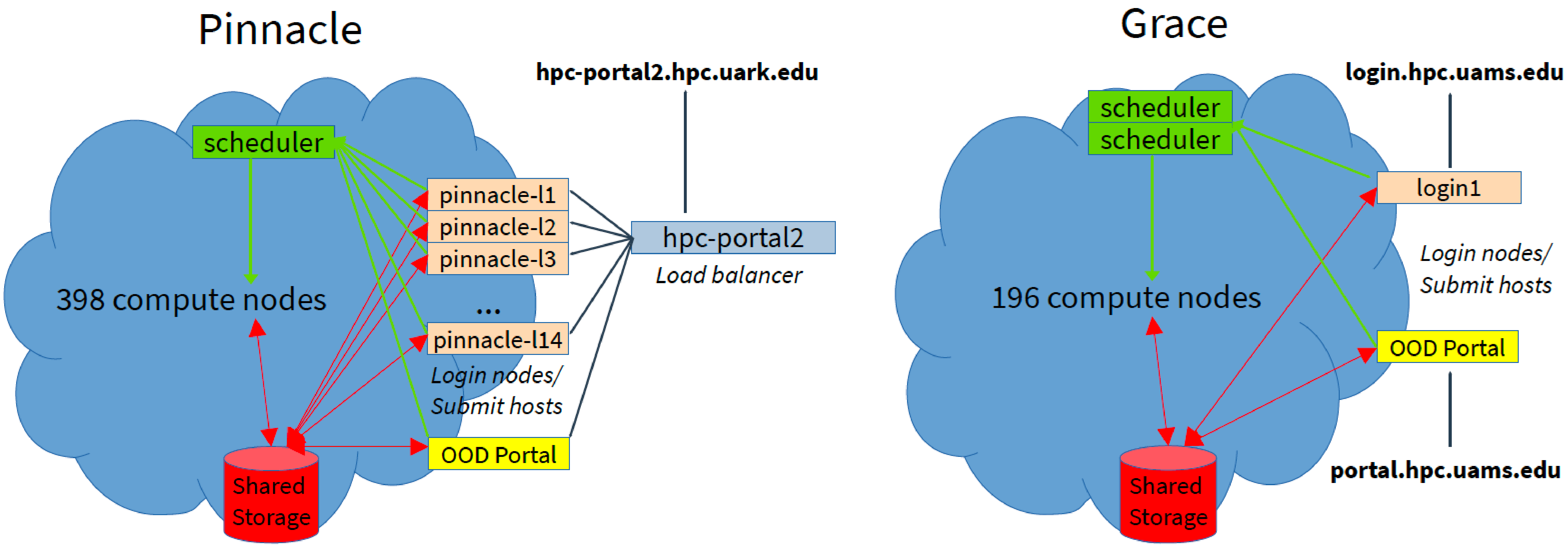

UNIVERSITY OF ARKANSAS
PINNACLE
STATE ARKANSAS STATE UNIVERSITY
PHILANDER SMITH COLLEGE
UALR University of Arkansas at Little Rock
UAMS UNIVERSITY OF ARKANSAS FOR MEDICAL SCIENCES
UNIVERSITY of ARKANSAS FORT SMITH
ARKANSAS TECH UNIVERSITY 1909
UNIVERSITY OF CENTRAL ARKANSAS
UNIVERSITY of ARKANSAS AT PINE BLUFF
UAM
SOUTHERN ARKANSAS UNIVERSITY

Arkansas Research Platform is a collaboration of higher education institutions within the state of Arkansas which provides computing resources and data storage to all its members (free of charge to students, faculty and staff). The core components are the **Pinnacle cluster** managed by the Arkansas High Performance Computing Center (UAF, Fayetteville) and the **Grace cluster** managed by UAMS High Performance Computing Center (UAMS, Little Rock).

supported by UAMS High Performance Computing AHPCC Arkansas High Performance Computing Center

ARKANSAS RESEARCH PLATFORM ARP

ARP Clusters – functional diagrams



scheduler – coordinates all of the nodes in the queue (SLURM queueing system)
submit hosts – nodes from which users submit jobs to the queues
login hosts – nodes accessible from the external network
compute nodes – nodes that assigned to queues and run jobs

Arkansas Research Platform - ARP

Deployed a point-to-point virtual research network between UAMS and UAF, extensible to other ARE-ON members

- Forms the backbone of the ARP

Installed 100Gbps connection between North Little Rock and Fayetteville ARE-ON backbone locations to enable higher throughput between UAMS and UAF.

Installed dedicated research 100Gbps connection to the UAMS campus for the UAMS Science DMZ

ARP Workshops

Began in Fall 2022

- UAMS
- UALR

Offered again Fall 2023

Getting Started with the

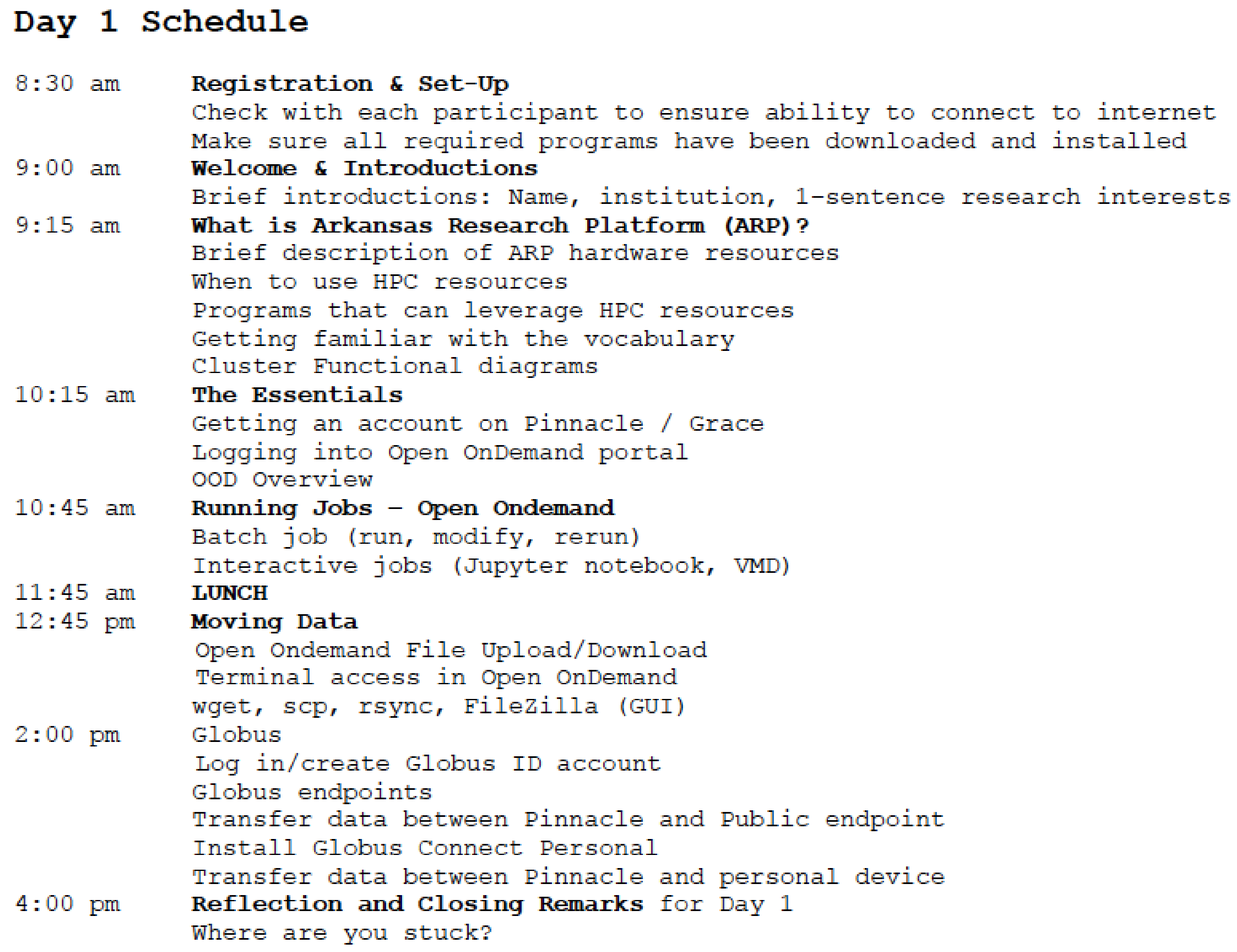
Arkansas Research Platform

UAMS Campus
September 29-30

Day 1

supported by UAMS High Performance Computing AHPCC Arkansas High Performance Computing Center

ARKANSAS RESEARCH PLATFORM ARP



Day 1 Schedule

8:30 am **Registration & Set-Up**
Check with each participant to ensure ability to connect to internet
Make sure all required programs have been downloaded and installed

9:00 am **Welcome & Introductions**
Brief introductions: Name, institution, 1-sentence research interests

9:15 am **What is Arkansas Research Platform (ARP)?**
Brief description of ARP hardware resources
When to use HPC resources
Programs that can leverage HPC resources
Getting familiar with the vocabulary
Cluster Functional diagrams

10:15 am **The Essentials**
Getting an account on Pinnacle / Grace
Logging into Open OnDemand portal
OOD Overview

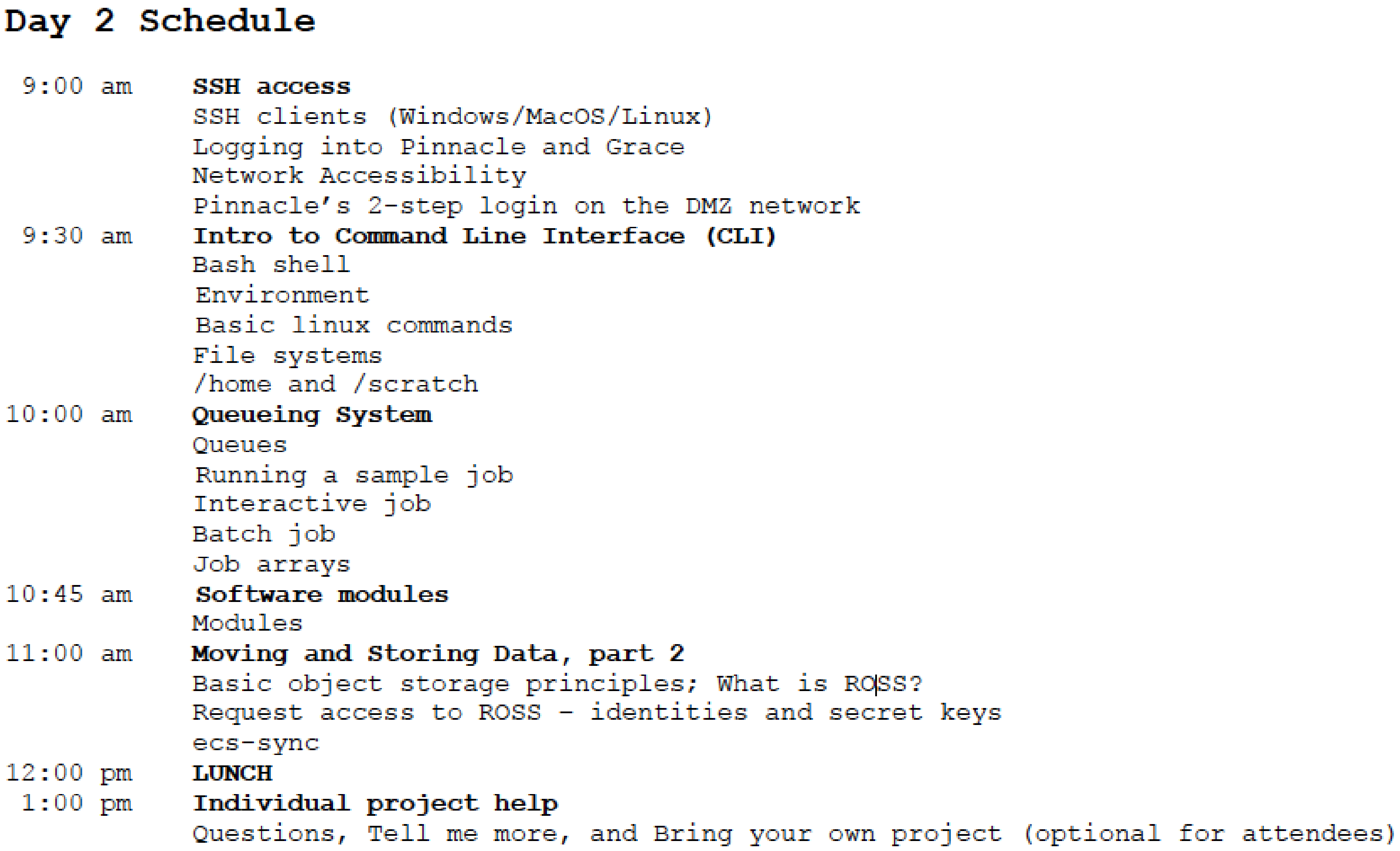
10:45 am **Running Jobs – Open Ondemand**
Batch job (run, modify, rerun)
Interactive jobs (Jupyter notebook, VMD)

11:45 am **LUNCH**

12:45 pm **Moving Data**
Open Ondemand File Upload/Download
Terminal access in Open OnDemand
wget, scp, rsync, FileZilla (GUI)

2:00 pm Globus
Log in/create Globus ID account
Globus endpoints
Transfer data between Pinnacle and Public endpoint
Install Globus Connect Personal
Transfer data between Pinnacle and personal device

4:00 pm **Reflection and Closing Remarks** for Day 1
Where are you stuck?



Day 2 Schedule

9:00 am **SSH access**
SSH clients (Windows/MacOS/Linux)
Logging into Pinnacle and Grace
Network Accessibility
Pinnacle's 2-step login on the DMZ network

9:30 am **Intro to Command Line Interface (CLI)**
Bash shell
Environment
Basic linux commands
File systems
/home and /scratch

10:00 am **Queueing System**
Queues
Running a sample job
Interactive job
Batch job
Job arrays

10:45 am **Software modules**
Modules

11:00 am **Moving and Storing Data, part 2**
Basic object storage principles; What is ROSS?
Request access to ROSS – identities and secret keys
ecs-sync

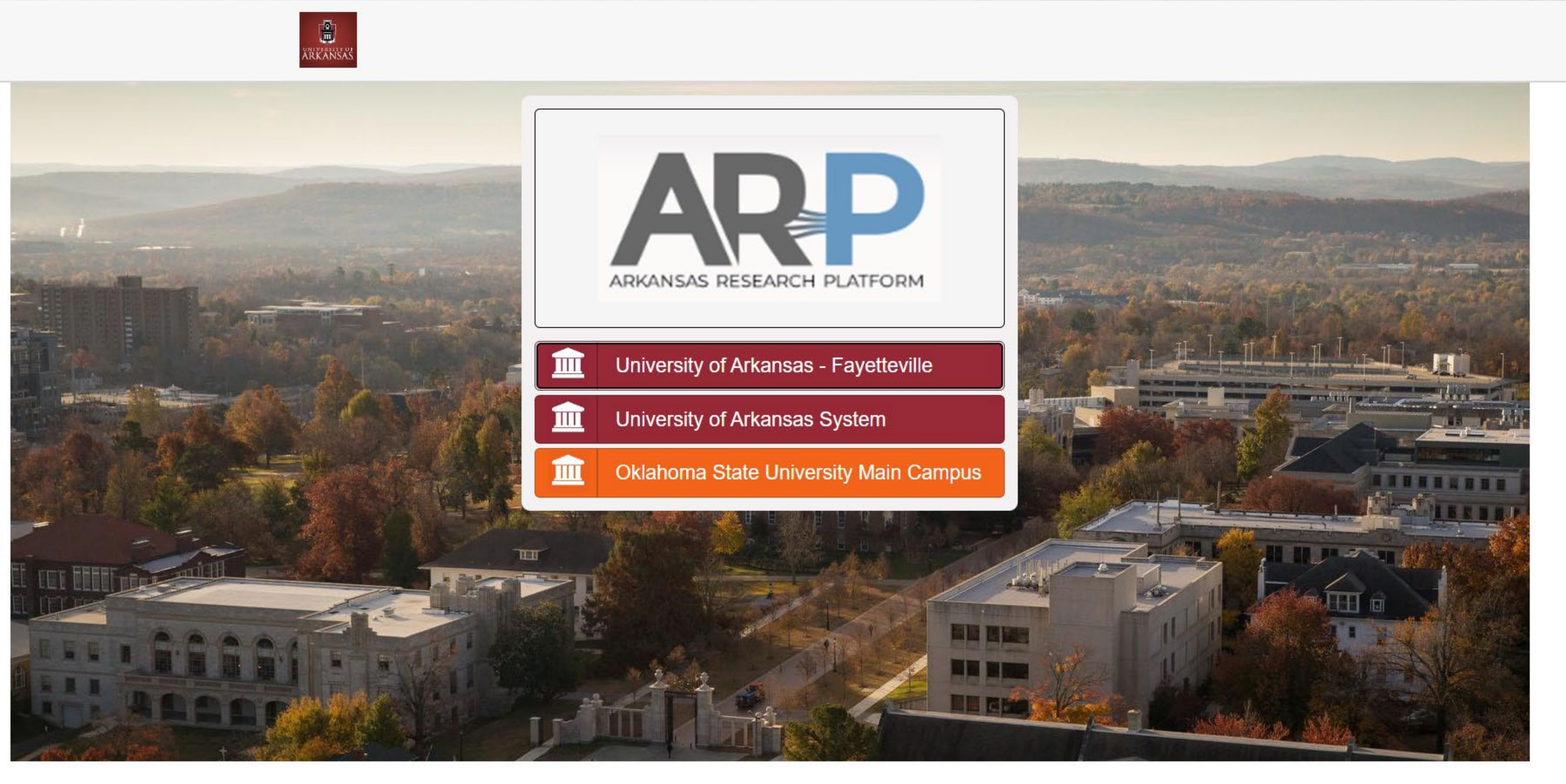
12:00 pm **LUNCH**

1:00 pm **Individual project help**
Questions, Tell me more, and Bring your own project (optional for attendees)

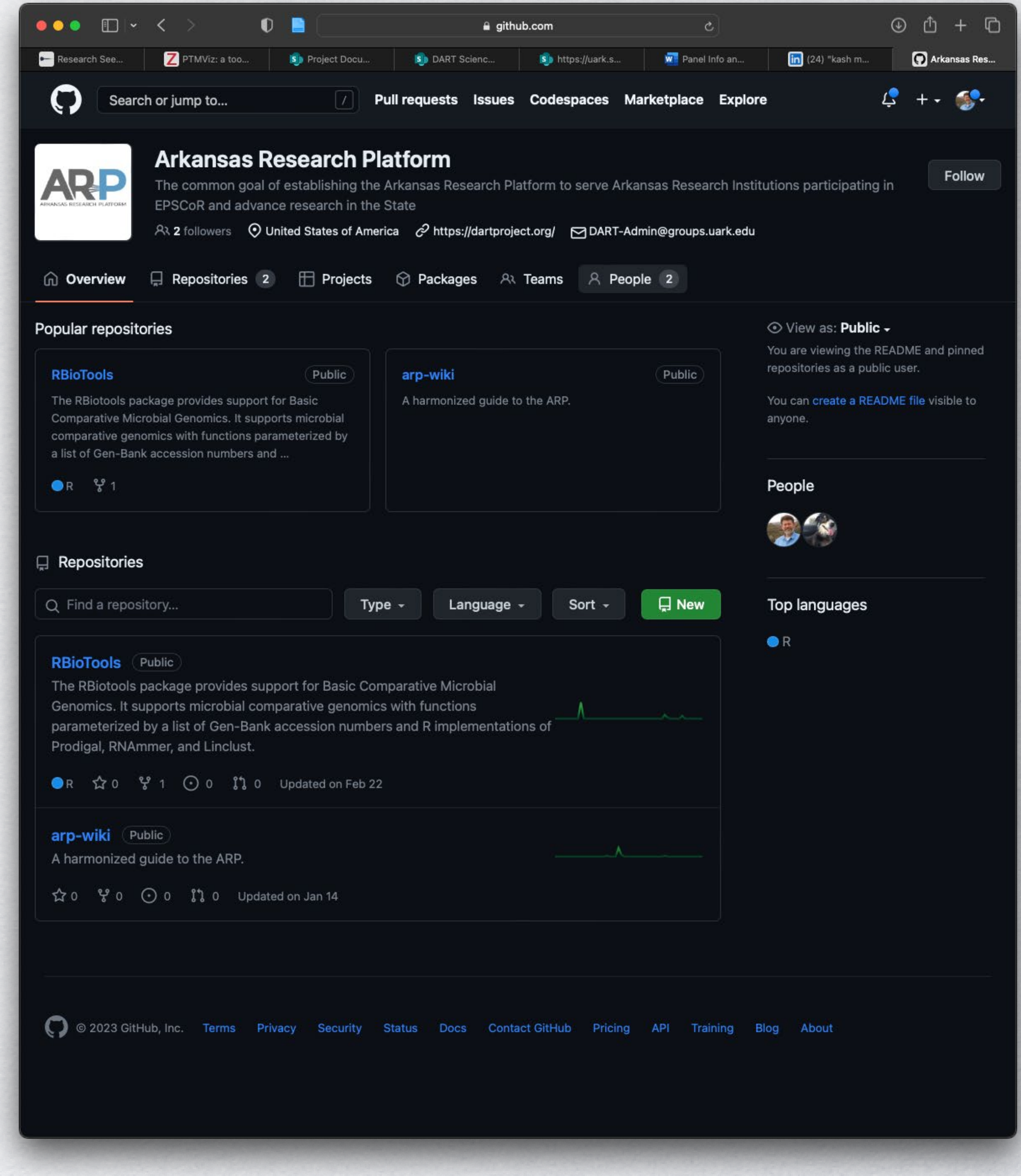
supported by UAMS High Performance Computing AHPCC Arkansas High Performance Computing Center

ARKANSAS RESEARCH PLATFORM ARP

And, in the coming months…


UNIVERSITY OF ARKANSAS
ARP
ARKANSAS RESEARCH PLATFORM
University of Arkansas - Fayetteville
University of Arkansas System
Oklahoma State University Main Campus

…federated identity access



Arkansas Research Platform

The common goal of establishing the Arkansas Research Platform to serve Arkansas Research Institutions participating in EPSCoR and advance research in the State

2 followers · United States of America · https://dartproject.org/ · DART-Admin@groups.uark.edu

Popular repositories

RBioTools Public

The RBiotools package provides support for Basic Comparative Microbial Genomics. It supports microbial comparative genomics with functions parameterized by a list of Gen-Bank accession numbers and ...

R · 1

arp-wiki Public

A harmonized guide to the ARP.

View as: Public

You are viewing the README and pinned repositories as a public user.

You can create a README file visible to anyone.

People

Repositories

Find a repository... · Type · Language · Sort · New

RBioTools Public

The RBiotools package provides support for Basic Comparative Microbial Genomics. It supports microbial comparative genomics with functions parameterized by a list of Gen-Bank accession numbers and R implementations of Prodigal, RNAmmer, and Linclust.

R · 0 · 1 · 0 · 0 · Updated on Feb 22

arp-wiki Public

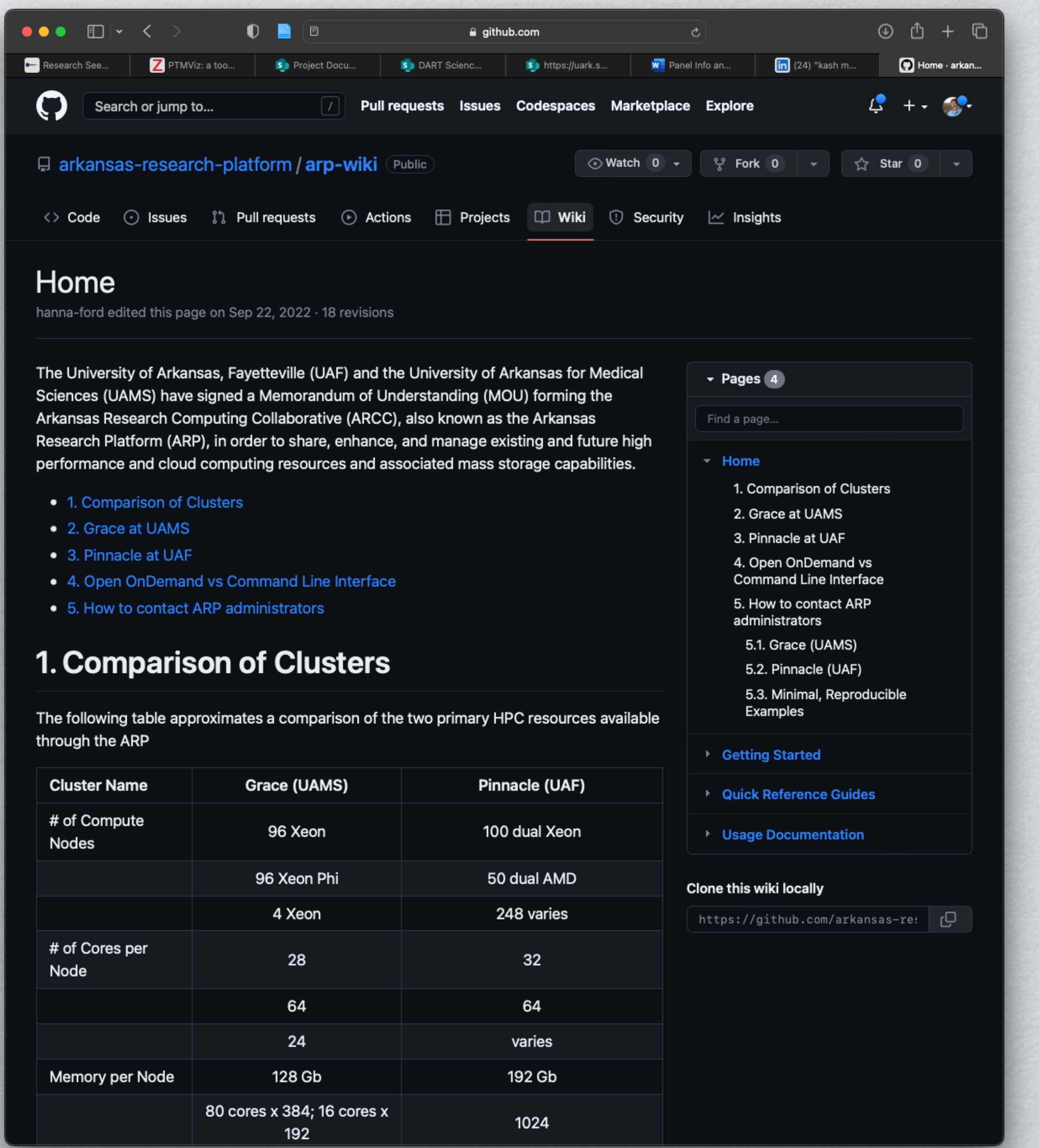
A harmonized guide to the ARP.

0 · 0 · 0 · 0 · Updated on Jan 14

Top languages

R

© 2023 GitHub, Inc. Terms Privacy Security Status Docs Contact GitHub Pricing API Training Blog About



arkansas-research-platform / arp-wiki Public

Home

hanna-ford edited this page on Sep 22, 2022 · 18 revisions

The University of Arkansas, Fayetteville (UAF) and the University of Arkansas for Medical Sciences (UAMS) have signed a Memorandum of Understanding (MOU) forming the Arkansas Research Computing Collaborative (ARCC), also known as the Arkansas Research Platform (ARP), in order to share, enhance, and manage existing and future high performance and cloud computing resources and associated mass storage capabilities.

- 1. Comparison of Clusters
- 2. Grace at UAMS
- 3. Pinnacle at UAF
- 4. Open OnDemand vs Command Line Interface
- 5. How to contact ARP administrators

1. Comparison of Clusters

The following table approximates a comparison of the two primary HPC resources available through the ARP

Cluster Name	Grace (UAMS)	Pinnacle (UAF)
# of Compute Nodes	96 Xeon	100 dual Xeon
	96 Xeon Phi	50 dual AMD
	4 Xeon	248 varies
# of Cores per Node	28	32
	64	64
	24	varies
Memory per Node	128 Gb	192 Gb
	80 cores x 384; 16 cores x 192	1024

Pages 4

Find a page...

Home
1. Comparison of Clusters
2. Grace at UAMS
3. Pinnacle at UAF
4. Open OnDemand vs Command Line Interface
5. How to contact ARP administrators
5.1. Grace (UAMS)
5.2. Pinnacle (UAF)
5.3. Minimal, Reproducible Examples

Getting Started

Quick Reference Guides

Usage Documentation

Clone this wiki locally

https://github.com/arkansas-re:

Cyberinfrastructure Plan

Leveraged two NSF CC* funded projects:

GPN CyberTeam (NSF #1925681):

- CI working group formed, led by Great Plains Network executive director and CyberTeam PI's

UAF SHARP CCI (2021) (NSF #2126108):

- ARP working group composed of all DART campus CIO's and research leads to plan for sustainable ARP

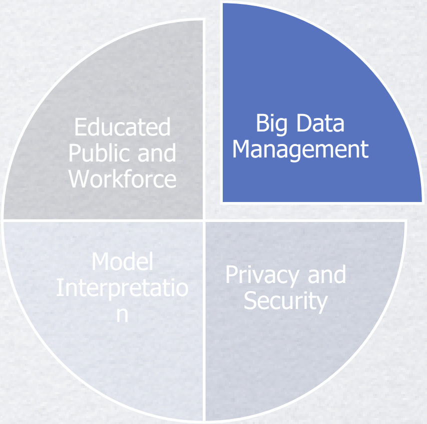
- State-wide gap analysis underway
- Reallocated funds from Globus Data Management service to fund federated identify solutions
- Developed MISP and SSP templates for managing CUI
- HIPAA-compliant storage at UAMS
- Engaged with TrustedCI and EPOC teams at Indiana University

Significant Accomplishments:
CI, LP, DC, SM, SA

DART

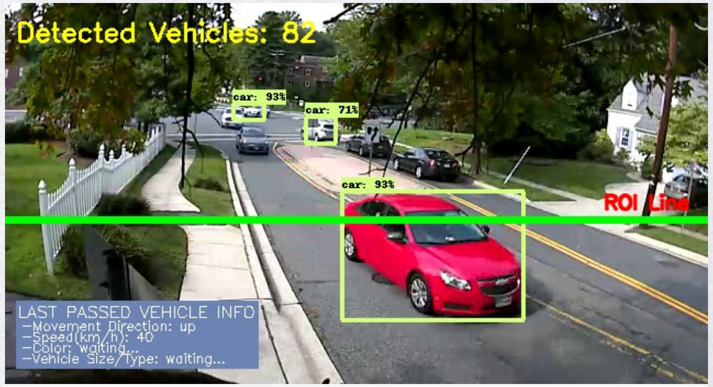
Data Analytics that are Robust and Trusted

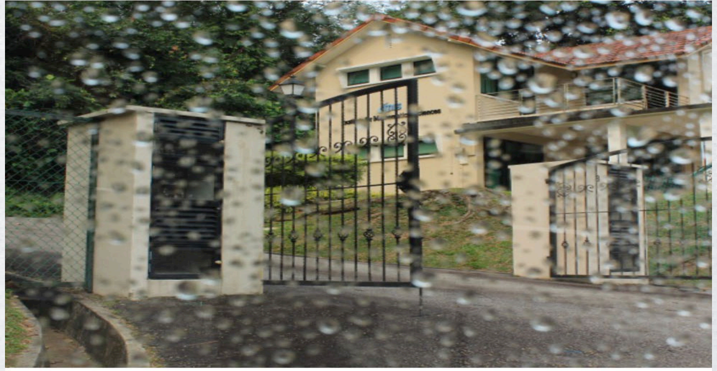
Significant Accomplishments



- Significant progress implementing the Arkansas Research Platform & Computing Collaborative for scaling algorithms and applications
- Integrated research and IT staff statewide working groups consolidating resources and expanding access
- Positive data control concepts and implementations being tested by an industrial partner
- Developed an autoencoder method that improved unsupervised and self-supervised deep learning methods' ability to manage and label much larger datasets.
- Much more sophisticated, coordinated approach to managing controlled or restricted data

Auto-annotation of multimedia data





Developing novel methods and algorithms

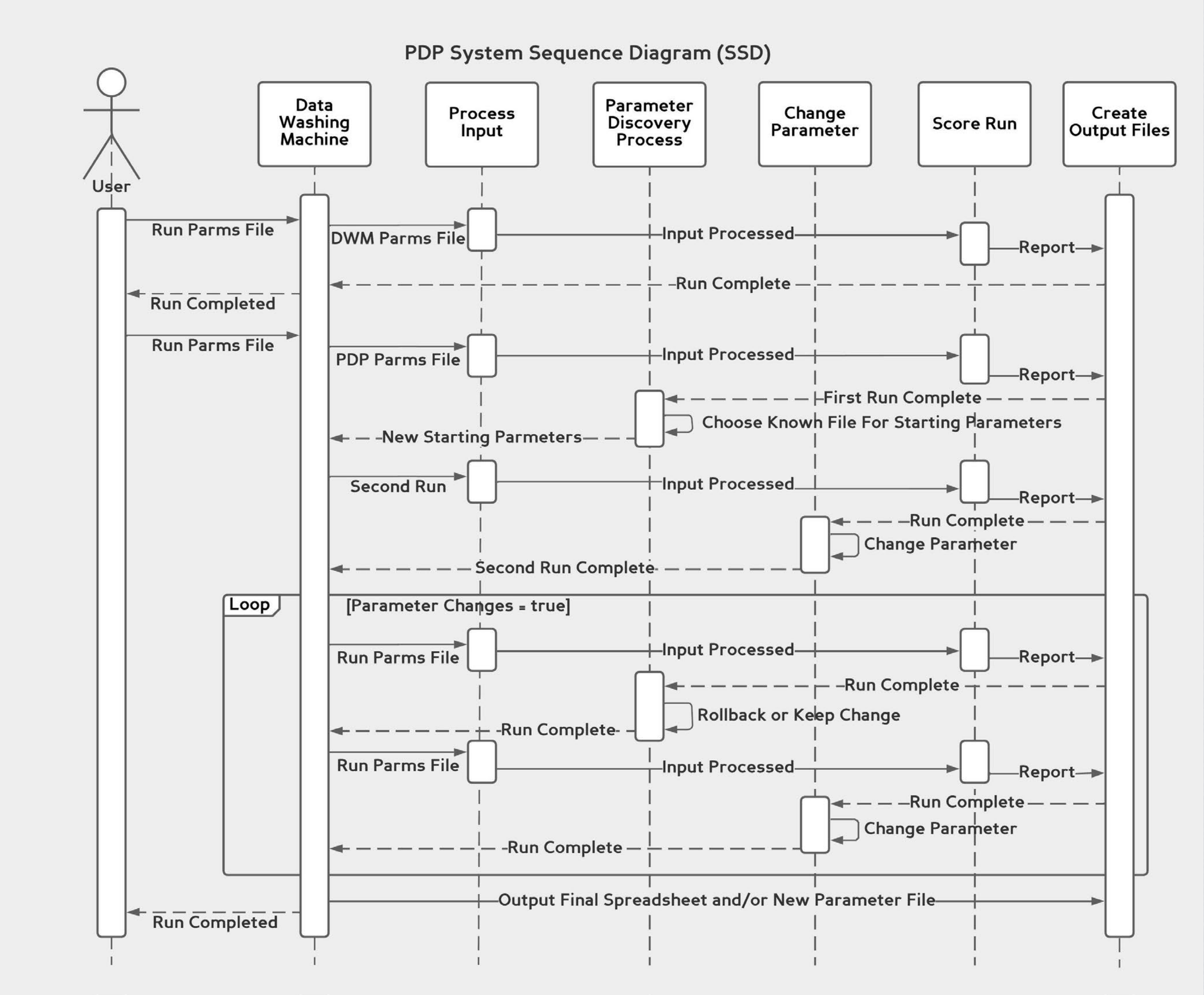
- for image and video analysis
  - Multimedia data characteristics towards target applications have been identified.
  - Applications have been defined around the team’s research into better informing disaster response with social media (SM4).
- vehicle speed estimation which has tested for road condition assessment
- for rain drop removal using GANs network to improve accuracy in using visual data captured in rainy conditions

Dr. Serhan Dagtas, UALR,

GA: Sharafat Hossain, UALR

Parameter Discovery Process (PDP) (DC1)

- Make the **Data Washing Machine (DWM)** truly “unsupervised”
- Automatically sets 14 DWM input parameters
- Given a dataset with redundant (duplicate) records, the PDP process
    - Reads the dataset and generates a set of token statistics
    - Compares the generated token statistics to the statistics from benchmark datasets with known optimal parameters
    - Starts with the nearest known optimal parameters
    - Iterates over the dataset to refine the parameter settings to be optimal for the input dataset



PDP System Sequence Diagram (SSD)

Hadoop-based Data Washing Machine (HDWM)

- The current design of the DWM is single-threaded:
  - Unable to go beyond a few thousands of records
    - Python version – up to 100, 000 records
    - Java version – up to 1, 000, 000 records
  - Goal – cluster more than a million (hundreds of million records)

- HDWM is a highly scalable version of the current Data Washing Machine designed with Hadoop MapReduce

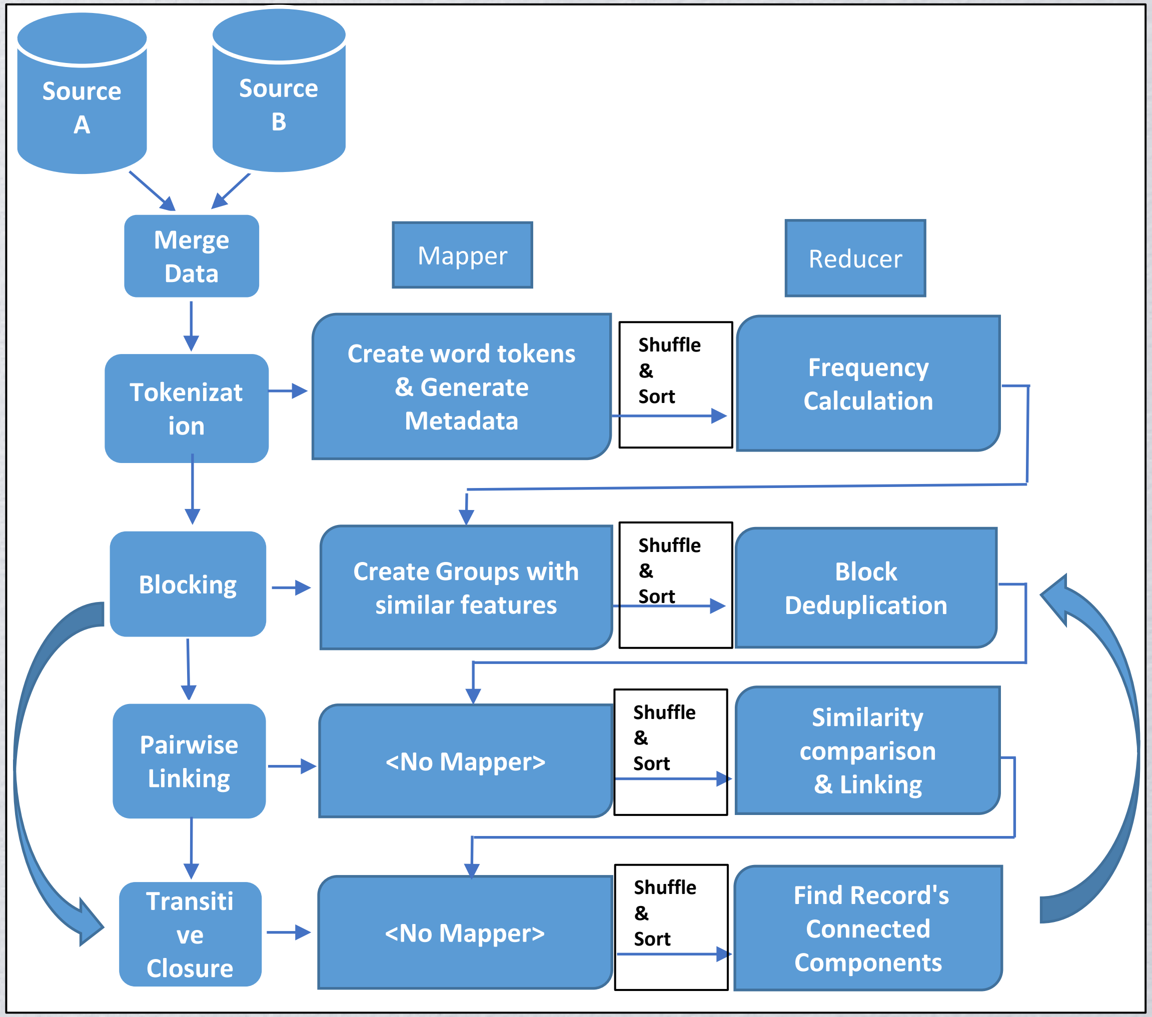
Hadoop-based Data Washing Machine (HDWM)

It solves the linear-based processing of DWM by:

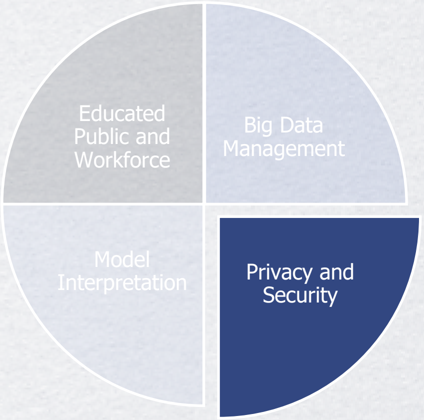
- Using MapReduce programming model
- Ability to scale and cluster billions of records using HDFS

HDWM handles the "out-of-memory" problem caused by the creation of shared memory tables/dictionaries

- Carries intrinsic metadata alongside tokens
- E.g. "token: {refID, token, position, frequency}"



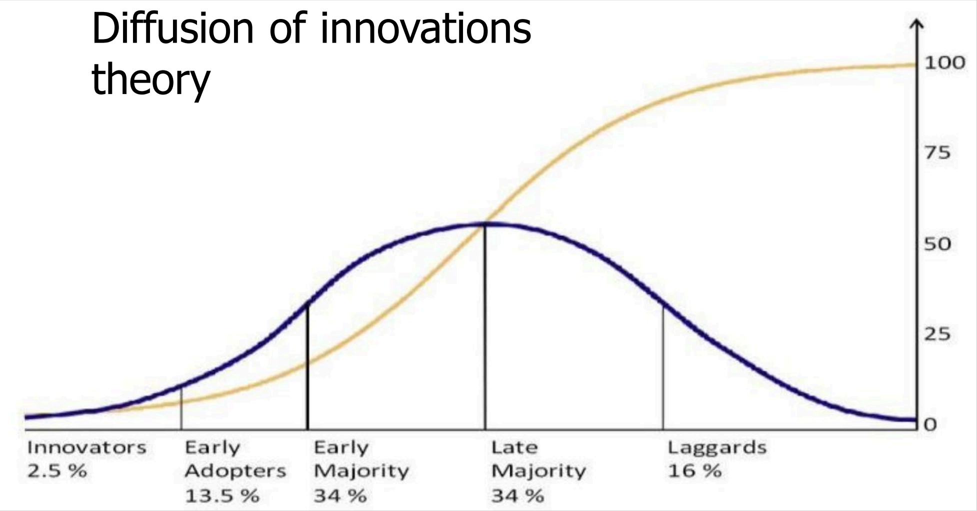
Significant Accomplishments

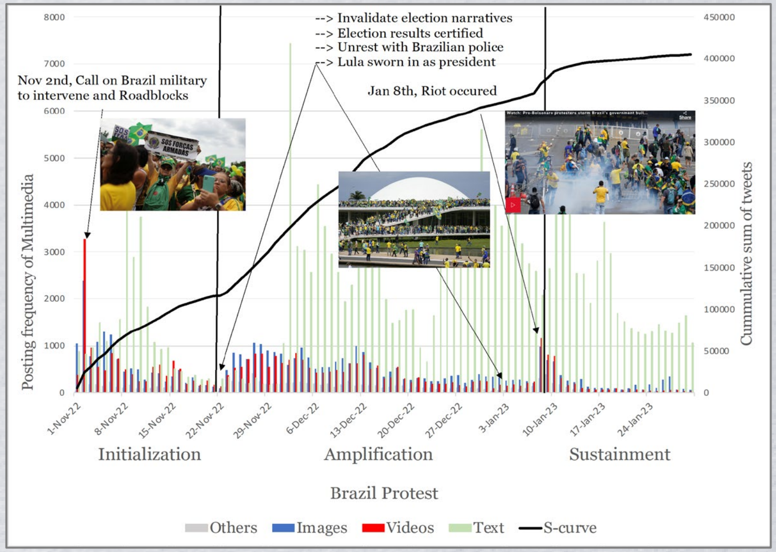


- Developed Master Information Security and System Security Plans for ARP resources to allow for eventual work with CUI data (NIST 800.171)
- Demonstrated a foundation technology for detecting in real time websites disseminating illegal contents
- A positive data curation prototype demonstrated the ability to control both access and metadata reporting for data operations in the Hadoop environment.
- DART researchers continue to develop and refine a suite of novel algorithms (differential privacy preserving multi-party learning, fair and robust learning under sample selection bias or attacks, uncertainty award crowdsourcing, fraud and hate detection in cyberspace, user-centric data sharing in cyberspace, and privacy-preserving analytics in health and genomics).

Role of Multimedia in Social Movement Mobilization

Studies of online Brazil and Peru political protests show that visual content on Twitter exhibits a greater degree of user engagement than text-only posts.

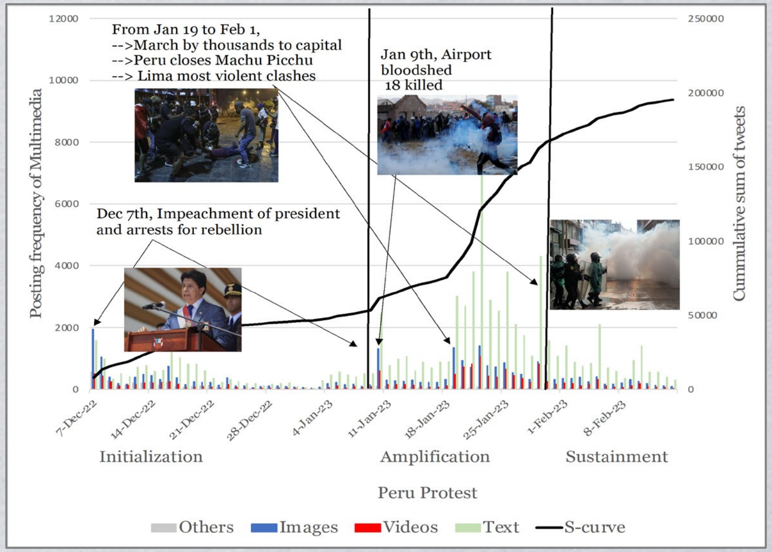

Diffusion of innovations theory


Nov 2nd, Call on Brazil military to intervene and Roadblocks

--> Invalidate election narratives
--> Election results certified
--> Unrest with Brazilian police
--> Lula sworn in as president

Jan 8th, Riot occured

Brazil Protest


From Jan 19 to Feb 1,
-->March by thousands to capital
-->Peru closes Machu Picchu
--> Lima most violent clashes

Jan 9th, Airport bloodshed 18 killed

Dec 7th, Impeachment of president and arrests for rebellion

Peru Protest

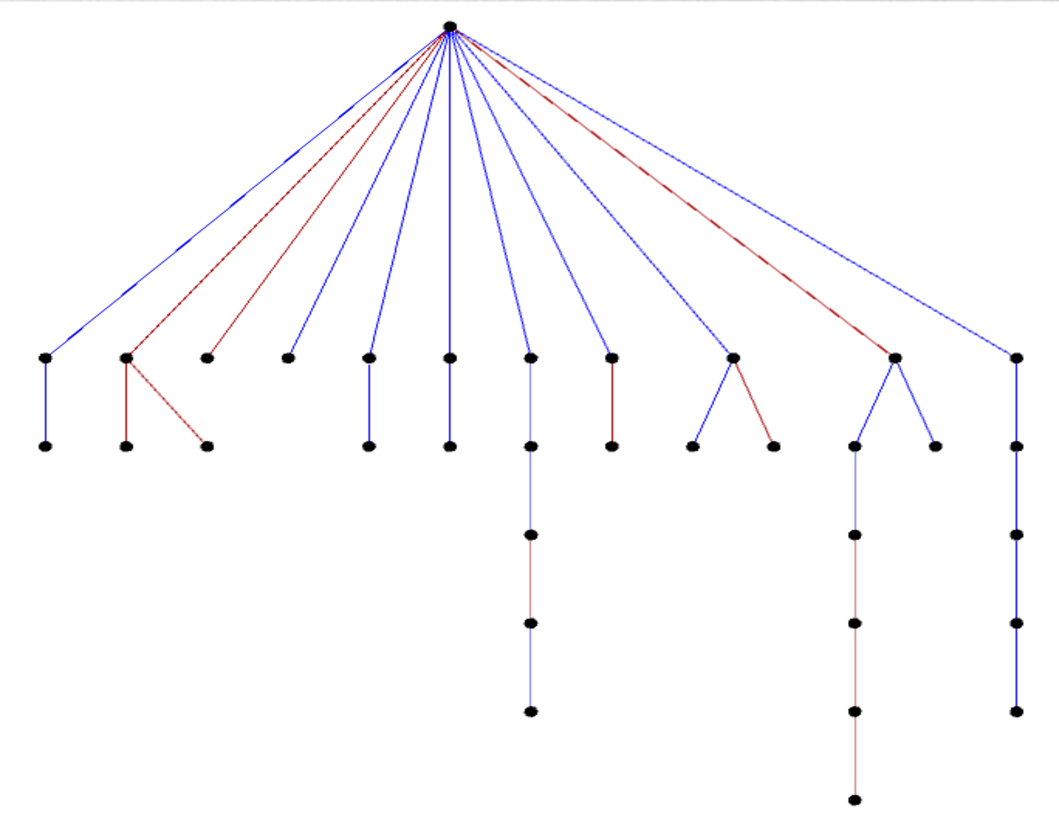
Diffusion of Innovations (DOI) theory (Rogers, 1962) is a social science theory that explains how new ideas, products, and technologies spread through society over time.

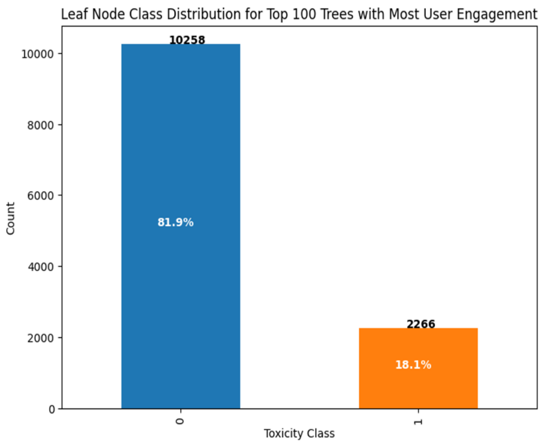
In this research, we apply DOI theory to online networks to evaluate the emerging adoption of information campaigns influenced by text, images, and video.

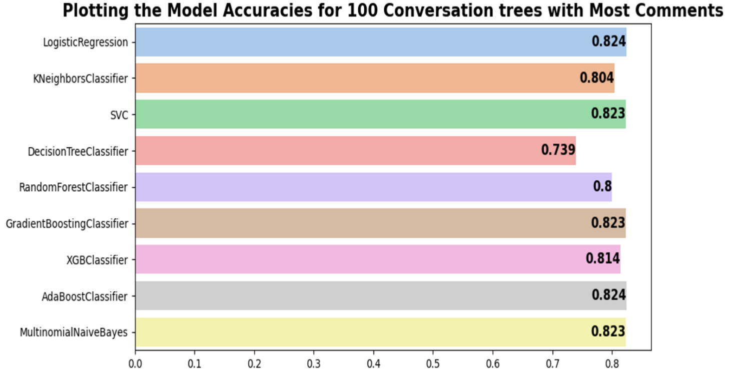
Toxicity and Community Health

- We generate conversation trees for threads with > 50% toxicity and use machine learning to predict the leaf nodes’ toxicity.
- The model predicts if the next reply will be toxic or not based on the structural characteristics of the conversation tree.
- Toxic conversation threads tend to have wider, deeper, and larger conversation branches.
- Those conversations are more likely to *end* with toxic comments.

Subreddit Post with comments-replies
Red = Toxic
Blue = non-Toxic




Leaf Node Class Distribution for Top 100 Trees with Most User Engagement
Count
0
2000
4000
6000
8000
10000
10258
81.9%
2266
18.1%
0
1
Toxicity Class



COSMOS
Collaboratorium for Social Media and Online Behavioral Studies

Deep Learning for Preventing Discrimination and Hate Speech on Social Media

Design and implementation of robust hate speech detection models via mitigating spurious correlations.

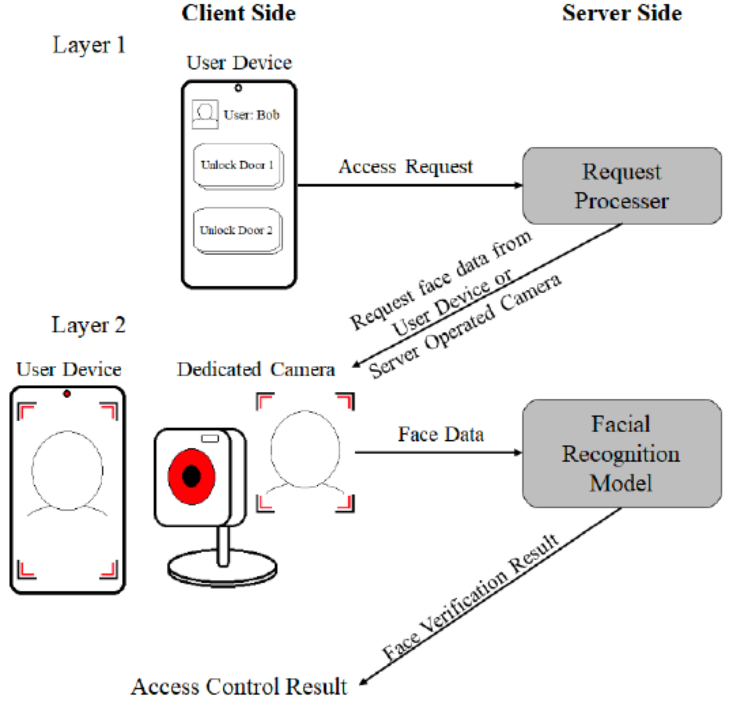
- Automatic filtrations of hateful content have been deployed to prevent hate speech, but intentionally modifying hate words may bypass auto-detection systems.
- The research team developed a robust hate speech detection model by formulating a causal structure to represent the causal relationship among different variables and integrating the causal strength into a regularized cross-entropy loss for removing the spurious correlation.

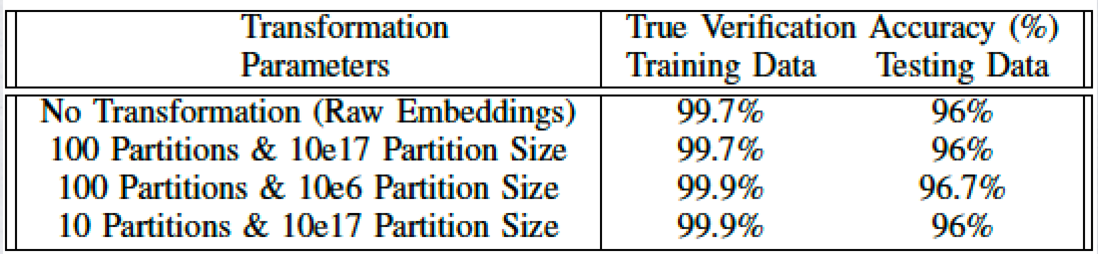
(Lu Zhang)

Cryptography-Assisted Secure and Privacy-Preserving Learning

Privacy-Preserving Face Recognition Access Control

- Goal: the backend server never sees the original face image or features
- Key idea: transform facial embeddings to a different space in a partially order-preserving way
- Transformed facial embeddings of the same user will still stay clustered together
- Transformed facial embeddings of different users will be relatively far away from each other
- Face matching model trained and queried based on transformed facial embeddings





Transformation Parameters	True Verification Accuracy (%) Training Data	Testing Data
No Transformation (Raw Embeddings)	99.7%	96%
100 Partitions & 10e17 Partition Size	99.7%	96%
100 Partitions & 10e6 Partition Size	99.9%	96.7%
10 Partitions & 10e17 Partition Size	99.9%	96%

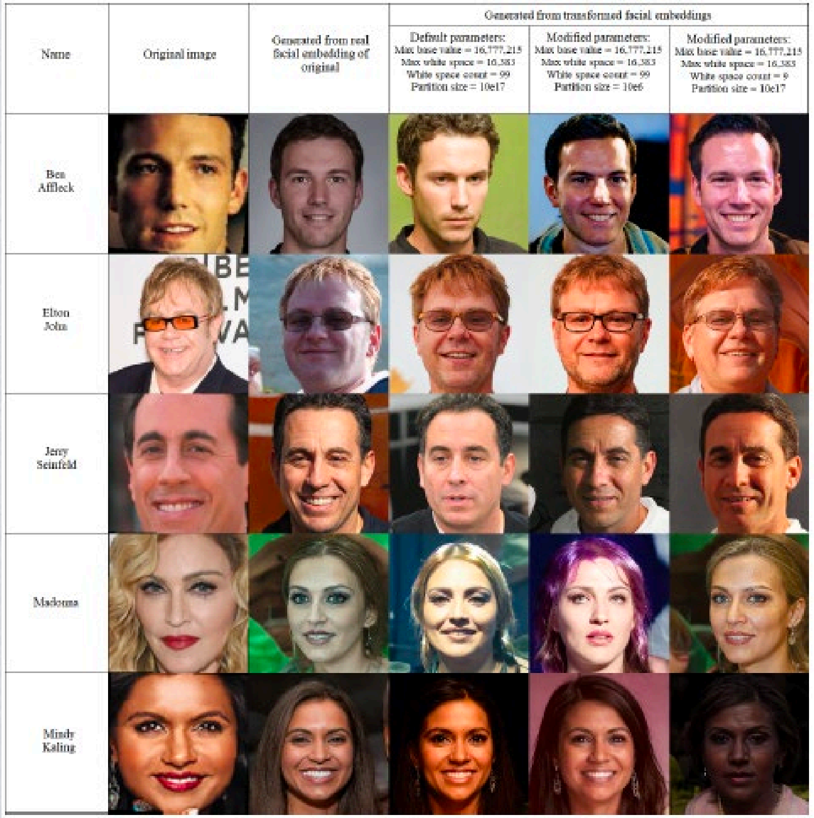
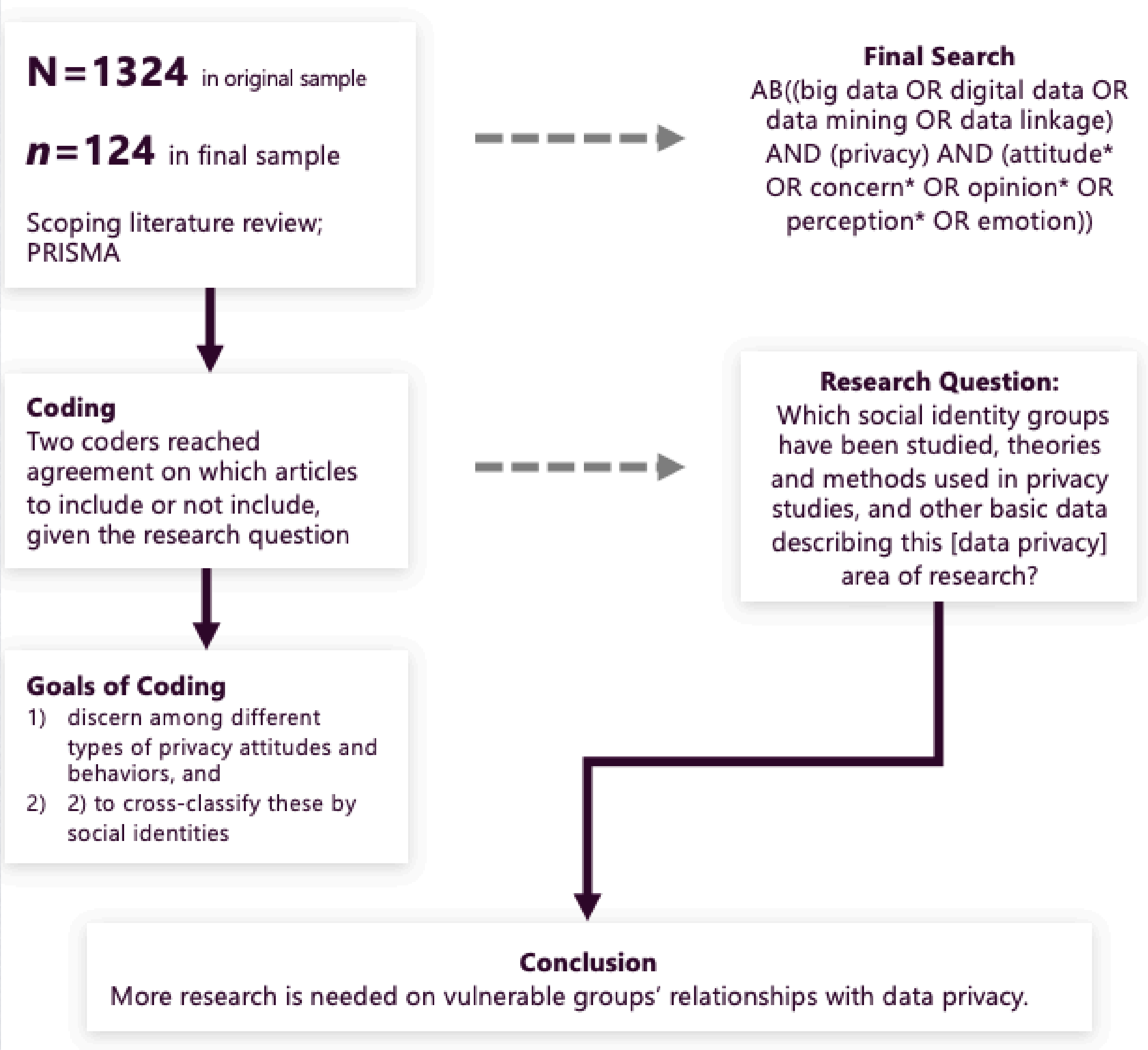

Name
Original image
Generated from real facial embedding of original
Generated from transformed facial embeddings
Default parameters: Max base value = 16,777,215 Max white space = 16,383 White space count = 99 Partition size = 10e17
Modified parameters: Max base value = 16,777,215 Max white space = 16,383 White space count = 99 Partition size = 10e6
Modified parameters: Max base value = 16,777,215 Max white space = 16,383 White space count = 9 Partition size = 10e17
Ben Affleck
Elton John
Jerry Seinfeld
Madonna
Mindy Kaling

Figure 2: Facial Reconstruction vs Proposed Method

The interface between data privacy and user’s social identities

- Systematic literature review of empirical research examining the intersection of social identities and privacy concerns within the big data context.
- Our goal is to better understand how users’ social identities inform privacy attitudes and behaviors concerning digital technologies and big data.
- Context matters for privacy concerns, attitudes, and behaviors. Many of these studies did not report how privacy concerns relate to social identities. Although age and gender were frequently examined, many of these studies lacked a theoretical framework to generate a detailed explanation of their findings.



N=1324 in original sample

***n*=124** in final sample

Scoping literature review; PRISMA

Final Search
AB((big data OR digital data OR data mining OR data linkage) AND (privacy) AND (attitude* OR concern* OR opinion* OR perception* OR emotion))

Coding
Two coders reached agreement on which articles to include or not include, given the research question

Research Question:
Which social identity groups have been studied, theories and methods used in privacy studies, and other basic data describing this [data privacy] area of research?

Goals of Coding

1) discern among different types of privacy attitudes and behaviors, and
2) 2) to cross-classify these by social identities

Conclusion
More research is needed on vulnerable groups’ relationships with data privacy.

Mining cyberargumentation data for opinion evolution

Deploy new version of cyberargumentation platform

- Collect postings from 100s of University of Arkansas students debating 5 sociological topics (e.g., guns on campus) for future analysis.

Develop expertise in transformer-based natural language processing (e.g., BERT) and apply it to two problems:

- Sentiment analysis (IMDB movie reviews)
  - Publication: G. Nkhata, U. Anjum and J. Zhan, “Movie Reviews Sentiment Analysis Using BERT,” The Fifteenth International Conference on Information, Process, and Knowledge Management (eKNOW 2023), to appear.
- Hate speech detection
  - Publication: X. Guo, U. Anjum and J. Zhan, "Cyberbully Detection Using BERT with Augmented Texts," 2022 IEEE International Conference on Big Data (Big Data), pp. 1246-1253.

Project leaders: Gauch, Adams

Sentiment Analysis Architecture

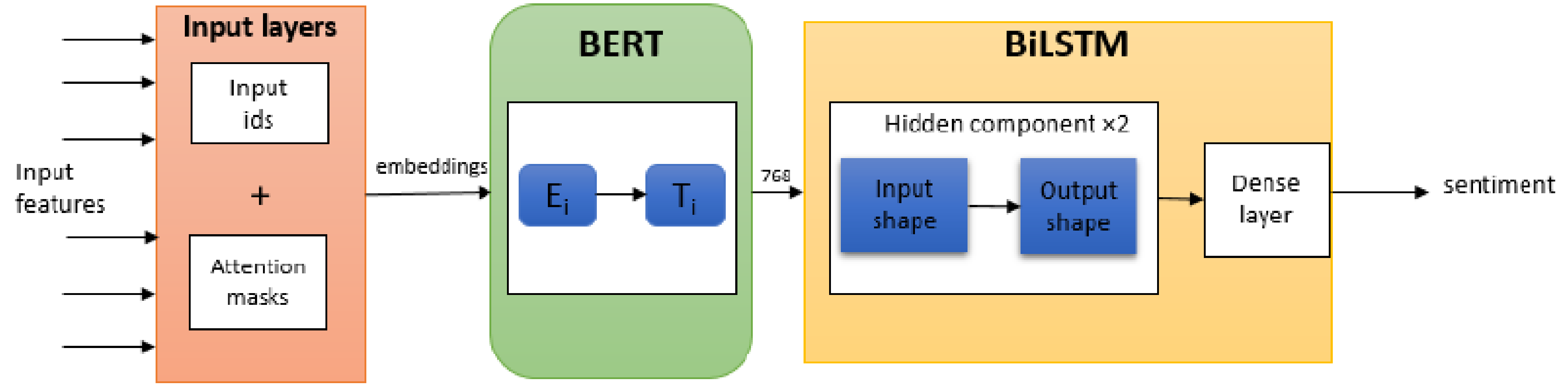
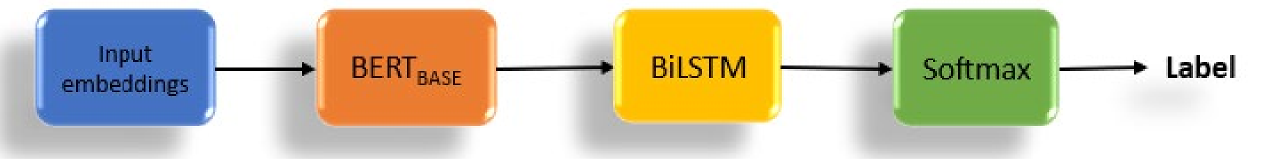


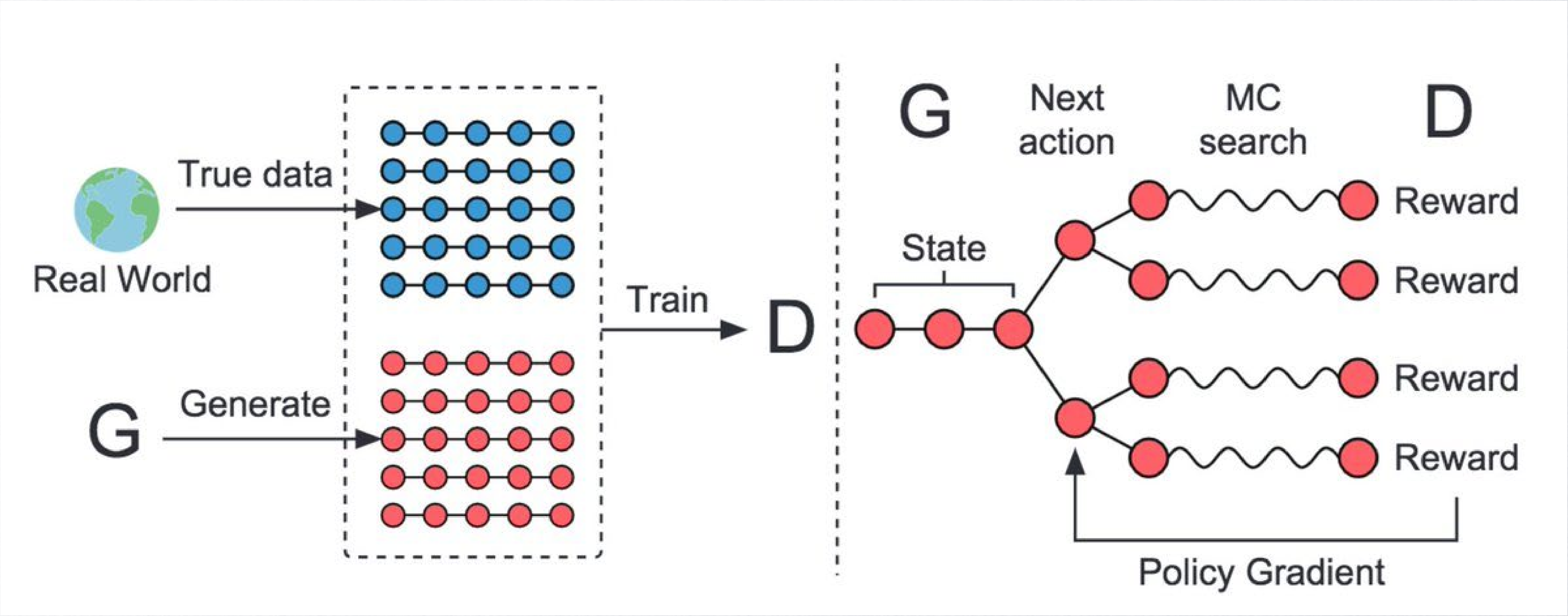
Fig. 2. Fine-tuning of the model



To the best of our knowledge, this is the first work to couple BERT with BiLSTM

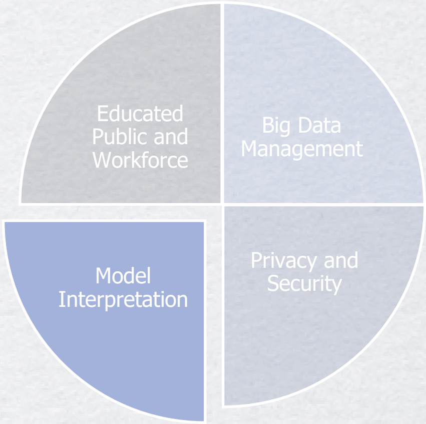
Results exceed state of the art accuracy on multiple datasets.

Cyberbully Detection Architecture



Outperformed state of the art on smaller and/or unbalanced data sets

Significant Accomplishments



- Successfully integrated CNN approaches on tabular data which incorporates neighborhood effects novel and potentially powerful ways

- Demonstrated a causal inference framework which can explain model predictions using unstructured data

- Developed a Boost-R solution to model stochastic event processes with heterogeneous features

- Bijective Maximum Likelihood image segmentation method achieved state-of-the-art image segmentation performance with no pixel-independent assumption in a tractable and invertible solution.

Learning & Prediction (LP1): Statistical Learning – Random Forests for Recurrent Event Analytics led by Xiao Liu

- Developed a new, tree-based model selection algorithm which focuses on interaction among regressors outperforms by multiple measures other selection strategies for linear regression and a logistic regression model of repeat events

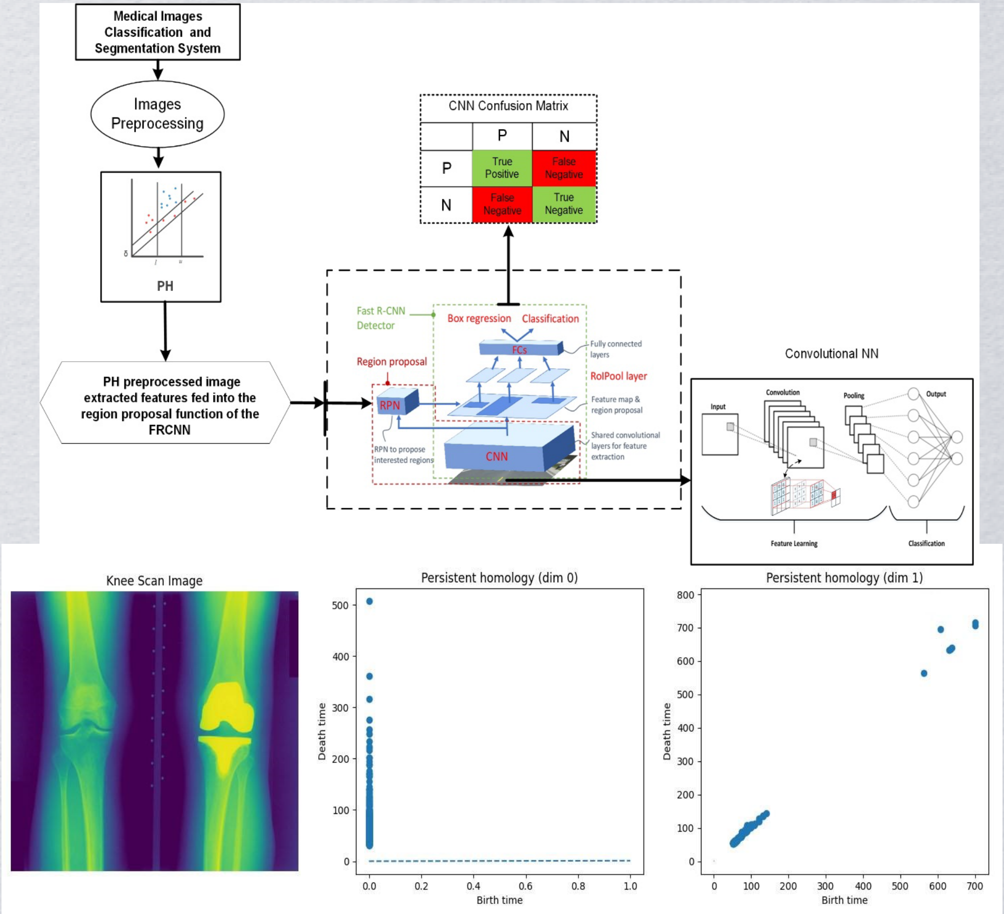
- Created a Random-Forest-Based algorithm and a Gradient-Boosted-Tree-Based algorithm for learning and prediction of recurrent event processes

- Contributed open-source code/tools for implementing the algorithms (available on GitHub)

November 2022
Volume 64, Number 4
Technometrics
A Journal of Statistics for the Physical, Chemical, and Engineering Sciences
ASQ
Published Quarterly by the American Society for Qualit
and the
American Statistical Associat
ISSN: 0022-4065
VOLUME 55 | ISSUE 2 | 2023
JOURNAL OF QUALITY TECHNOLOGY
A Journal of Methods, Applications, and Related Topics
ASQ
Excellence Through Quality
Taylor & Francis
Taylor & Francis Group

Learning & Prediction (LP3): Deep Learning – Novel Approaches led by Md Karim

- Developed PH-Net, a hybrid image processing model based on Persistent Homology (PH) and Fast Regional Convolutional Neural Network (FRCNN)
- Developed methods for image localization and image preprocessing using topological data analysis
- Performed an empirical study to determine dataset-specific usability of topological features
- Developed a new Self-supervised Domain Adaptation method in crowd counting
- Contributed a new Fairness Domain Adaptation (FREDOM) approach to semantic scene understanding
- Developed a new Self-supervised Spatiotemporal Transformers (SPARTAN) approach to group action recognition

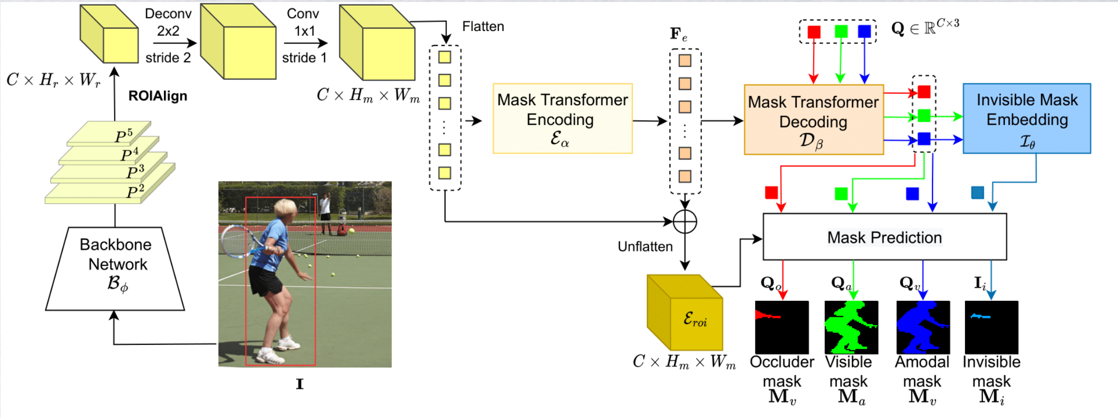


CNN Confusion Matrix

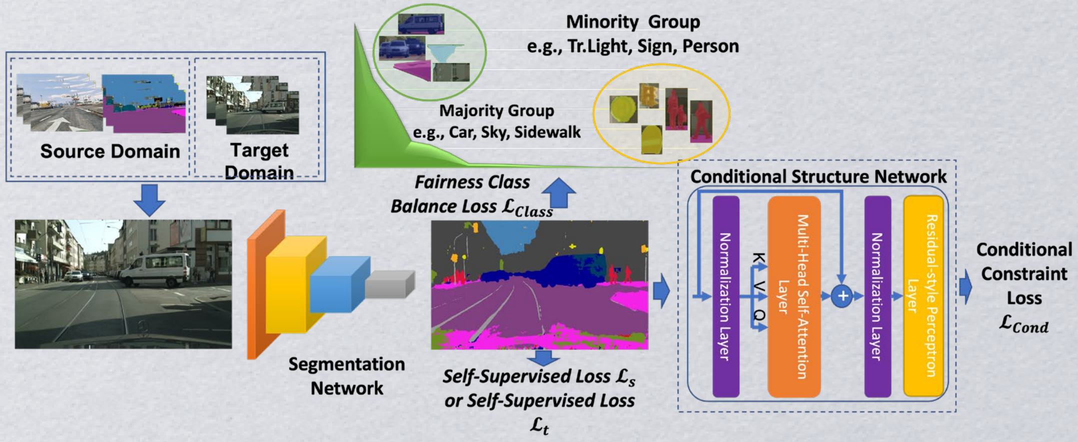
	P	N
P	True Positive	False Negative
N	False Negative	True Negative

Learning & Prediction (LP4): Deep Learning – Efficiency and Specification led by Khoa Luu

- Developed self-supervised 3D capsule networks for medical segmentation on less labeled data
- Developed a new self-supervised domain adaptation deep learning method to deal with limited training data
- Developed a new equipollent domain adaptation approach for image deblurring
- Proposed new fairness metrics to analyze the complexity and bias training data in deep learning



Deconv 2x2 stride 2
Conv 1x1 stride 1
Flatten
$C \times H_r \times W_r$
ROIAlign
$C \times H_m \times W_m$
P^5
P^4
P^3
P^2
Backbone Network $\mathcal{B}_\phi$
$\mathbf{I}$
Mask Transformer Encoding $\mathcal{E}_\alpha$
$\mathbf{F}_e$
$\mathbf{Q} \in \mathbb{R}^{C\times 3}$
Mask Transformer Decoding $\mathcal{D}_\beta$
Invisible Mask Embedding $\mathcal{I}_\theta$
Unflatten
Mask Prediction
$\mathcal{E}_{roi}$
$C \times H_m \times W_m$
$\mathbf{Q}_o$
$\mathbf{Q}_a$
$\mathbf{Q}_v$
$\mathbf{I}_i$
Occluder mask $\mathbf{M}_v$
Visible mask $\mathbf{M}_a$
Amodal mask $\mathbf{M}_v$
Invisible mask $\mathbf{M}_i$



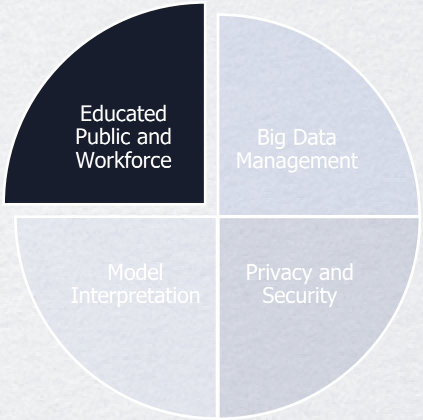
Minority Group
e.g., Tr.Light, Sign, Person
Majority Group
e.g., Car, Sky, Sidewalk
Source Domain
Target Domain
Fairness Class Balance Loss $\mathcal{L}_{Class}$
Conditional Structure Network
Normalization Layer
Multi-Head Self-Attention Layer
K V Q
Normalization Layer
Residual-style Perceptron Layer
Conditional Constraint Loss $\mathcal{L}_{Cond}$
Segmentation Network
Self-Supervised Loss $\mathcal{L}_s$ or Self-Supervised Loss $\mathcal{L}_t$

Significant Accomplishments: ED, Workforce, Broadening Participation

DART

Data Analytics that are Robust and Trusted

Significant Accomplishments



- In 2019, Arkansas had no 4-year DS degrees
- We now have programs implemented with students enrolled at the University of Arkansas, the University of Central Arkansas, and Arkansas State University.
- Two-year college students and faculty are using shared computing resources (in ARP) for the first time
- Arkansas Summer Research Institute keeps growing
- Certificate proposals for 2-year campuses coming soon

Education Theme – Year 3 Accomplishments

- **4-year Hubs for Common B.S. Data Science 8-semester Plan**
  - UAF is teaching all four years as of this year
  - UCA is evolving its curriculum to match the 8-semester plan
  - A-State is evolving its curriculum to match the 8-semester plan

- **ACTS for DASC 1003 Intro to Data Science has been approved**
  - DASC 1104 and 1223 will be proposed to be developed next

- **NorthArk and UAF are close to finalizing the model 2+2 plan**
  - This will facilitate +2 with UAF, UCA, and A-State

- **Study Abroad Program with the University of Nicosia is nearly complete**
  - Finalizing details for Year 2 Fall and/or Year 2 Spring
  - Designed to work for any institution following the 2+ or 8-semester plan
  - All courses our students would take are taught in English

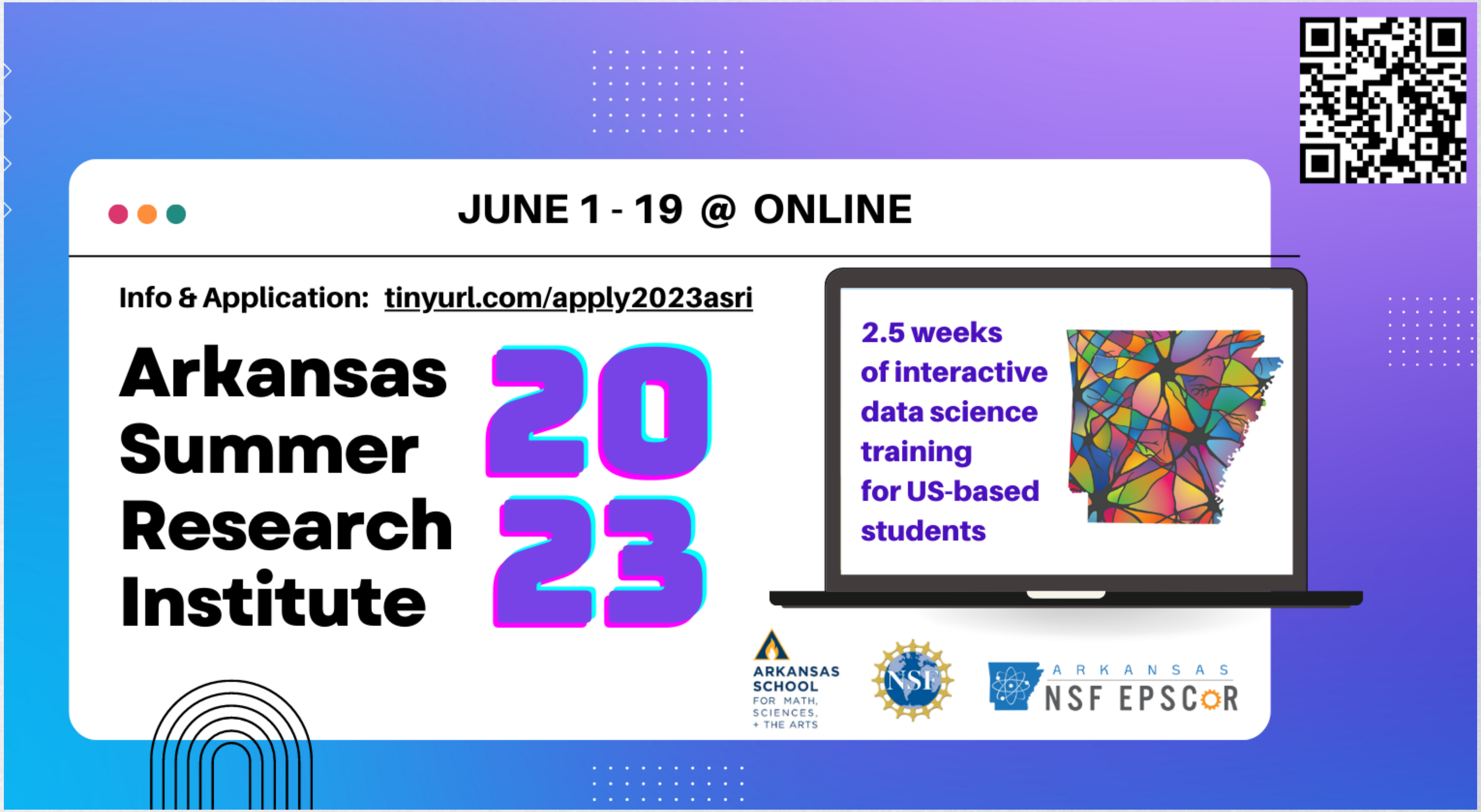
Education Theme – Year 3 Accomplishments (cont.)

- **Participated in many state-wide Data Science Workshops by ADHE**

- **Conference Papers / Presentations Submitted**
  - ACC Annual Fall 2023 Conference: "Data Science Careers Start at Community College"
    - Theme: Workforce Development
    - Presenter: Laura Berry, Interim Dean of Health Professions, North Arkansas Community College
      - Co-authors: Christine Davis, NWACC; Tina Moore, ADHE; Karl Schubert, UAF

  - ASEE 2023 Annual Conference:
    - Expanding & Improving a Multi-College Interdisciplinary B.S. Data Science Program with Concentrations
    - Theme: ASEE MULTI Division
    - Authors: Karl Schubert, UAF; Lee Shoultz, UAF, Shantel Romer, UAF
    - Note: builds on the previous notes on the Ed-Theme project
    - ASEE 2024 Paper submission will include co-leads & those active to cover the entire Ed-Theme at time of submission

Broadening Participation



JUNE 1 - 19 @ ONLINE

Info & Application: tinyurl.com/apply2023asri

Arkansas Summer Research Institute 2023

2.5 weeks of interactive data science training for US-based students

ARKANSAS SCHOOL FOR MATH, SCIENCES, + THE ARTS

NSF

ARKANSAS NSF EPSCoR

~300 applicants so far from all over the US

New sessions, new presenters, and improved flow

Extended by additional 3 days

Any US residents can apply (targeted for undergrads or beginners)

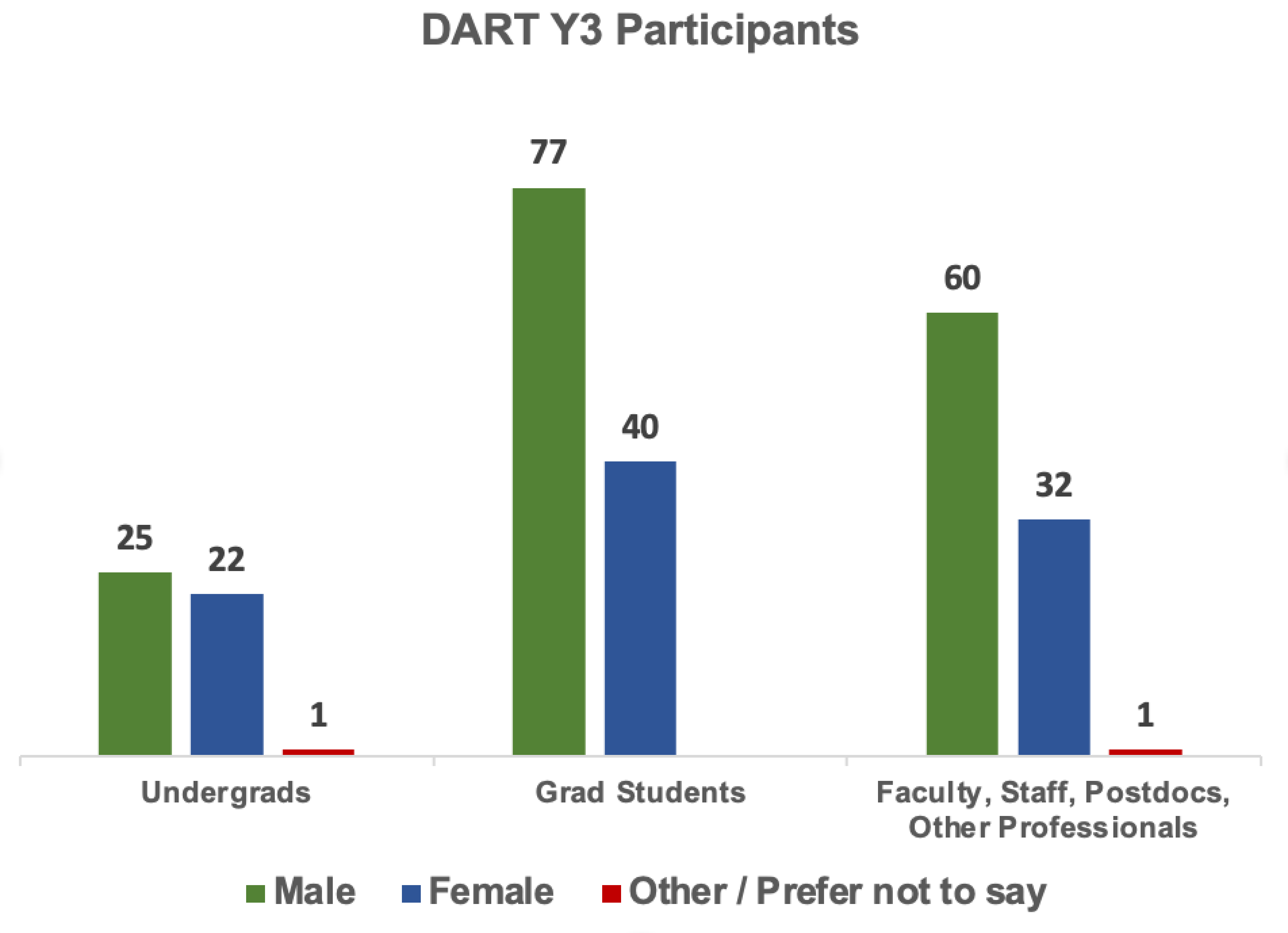
Broadening Participation

SURE Program- Summer Undergraduate Research Experience funding available, contact Brittany @ Brittany.Hillyer@arkansasedc.com

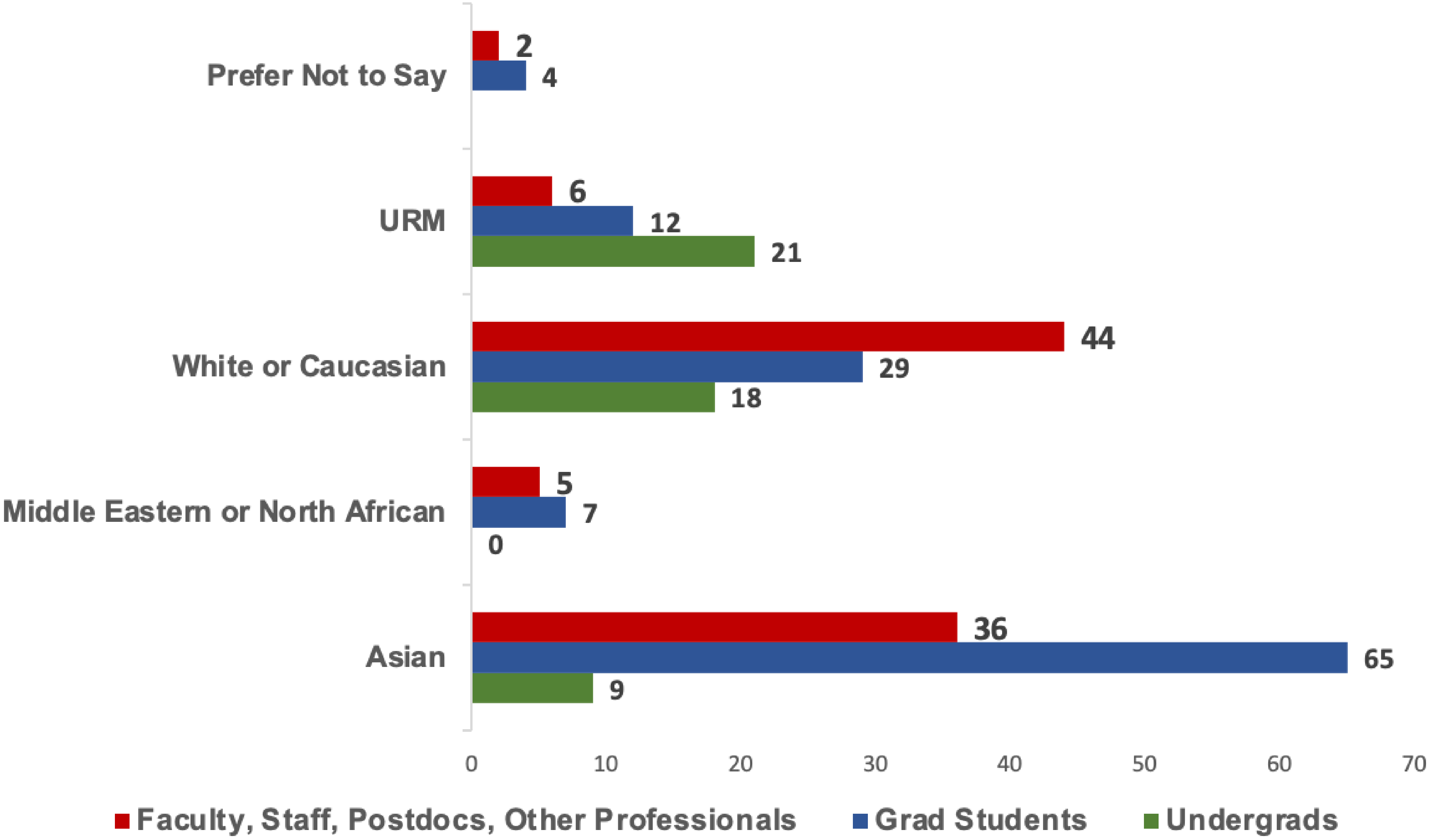
Minigrants up to $5,000 to support professional development, special activities, and other related projects available. Check website for details

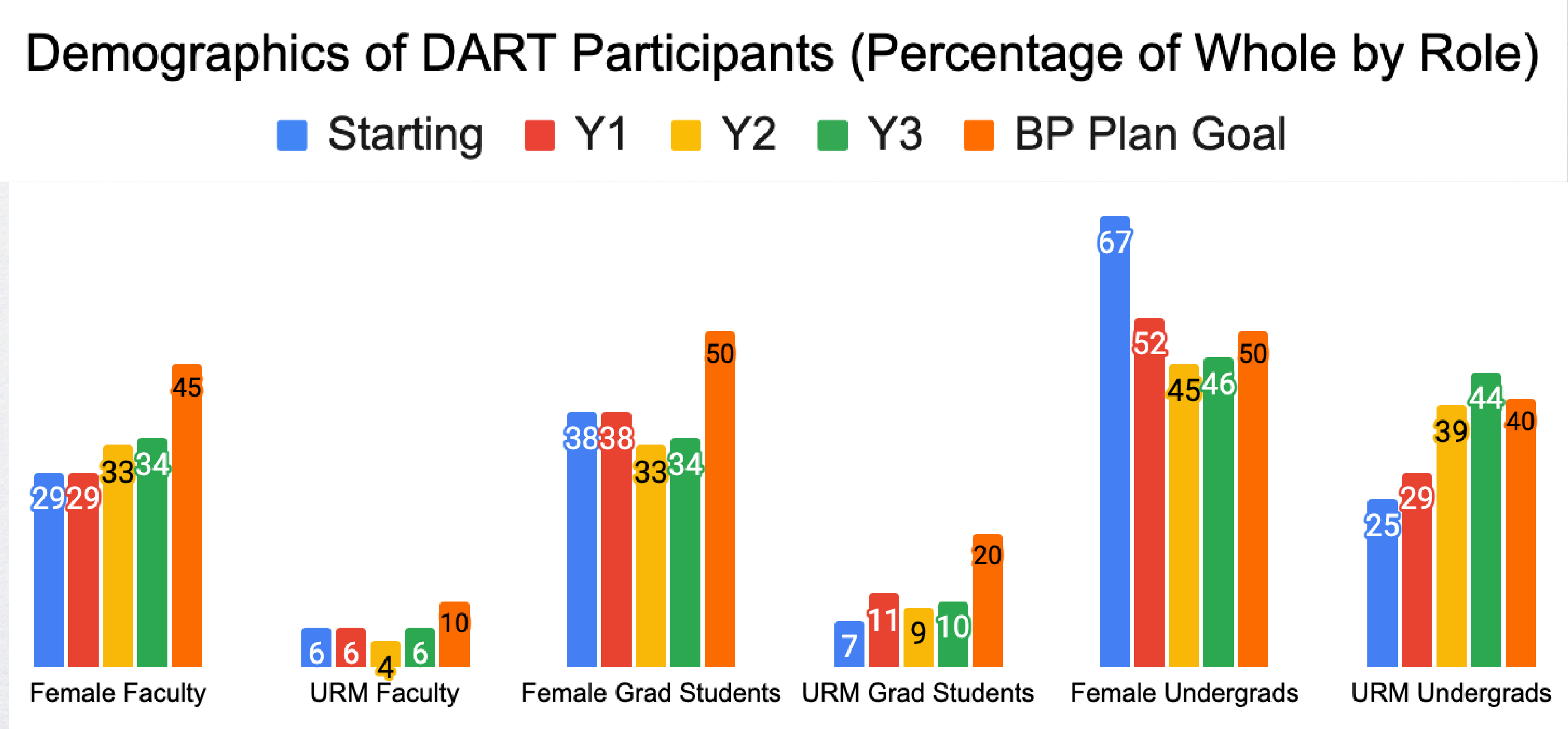
DART Participants: Year 3

AEDC	**6**
Arkansas State University	**14**
Arkansas Tech University	**8**
North Arkansas College	**1**
Philander Smith College	**12**
SAU Tech	**1**
Shorter College	**3**
Southern Arkansas University	**17**
University of Arkansas Division of Agriculture	**1**
University of Arkansas Fayetteville	**79**
University of Arkansas Little Rock	**79**
University of Arkansas Medical Sciences	**21**
University of Arkansas Pine Bluff	**7**
University of Central Arkansas	**11**


DART Y3 Participants
25
22
1
77
40
60
32
1
Undergrads
Grad Students
Faculty, Staff, Postdocs, Other Professionals
🟩 Male
🟦 Female
🟥 Other / Prefer not to say

DART Y3 Participants





Demographics of DART Participants (Percentage of Whole by Role)

Summary

- On track to meet strategic plan milestones and objectives
- Submitting a revision to strategic plan soon
- Made significant progress and have impactful results and collaborations
- Even more great things to share soon

DART

Data Analytics that are Robust and Trusted

This material is based upon work supported by the National Science Foundation under Award No. OIA-1946391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.