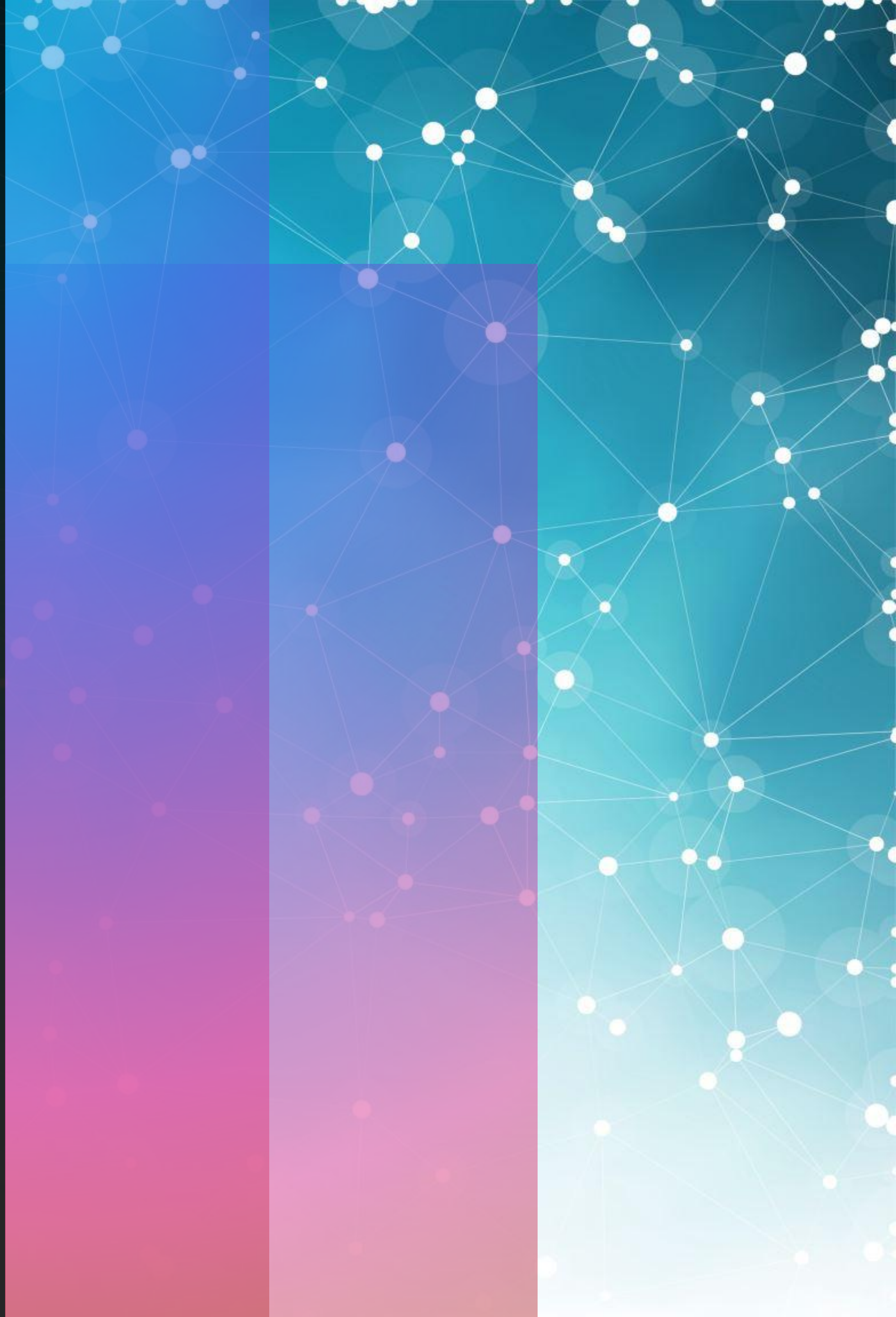
Research Data Management: Student Perspectives

Soma Mukherjee, University of Texas Rio Grande Valley

BRICCS RESEARCH DATA MANAGEMENT CONFERENCE

ALEXANDRIA, VA

07/11/2025



The University of Texas Rio Grande Valley (UTRGV) is a High Research Activity (R2) institution

$86.9 million in total research expenditures for Fiscal Year 2024

68 doctoral degrees in 2024

Age of data explosion in all fields

- Genomic Data: Experts predicted that by 2025, genome sequencing could produce up to 40 exabytes (40 billion gigabytes) of data per year.

- The Vera C. Rubin Observatory is expected to generate approximately 20 terabytes (TB) of data every night. The raw image data collected throughout the survey will amount to approximately 60 petabytes (PB).

- LIGO generates 30 MB per second.

- Square Kilometer Array (SKA) is still in the construction phase and is designed to be the largest radio telescope ever built. The first phase (SKA1) is expected to generate 160 terabytes of data per second, resulting in 5 zettabytes of data per year.

- As of 2025, the global datasphere, which includes data generated by businesses, is projected to reach approximately 181 zettabytes.

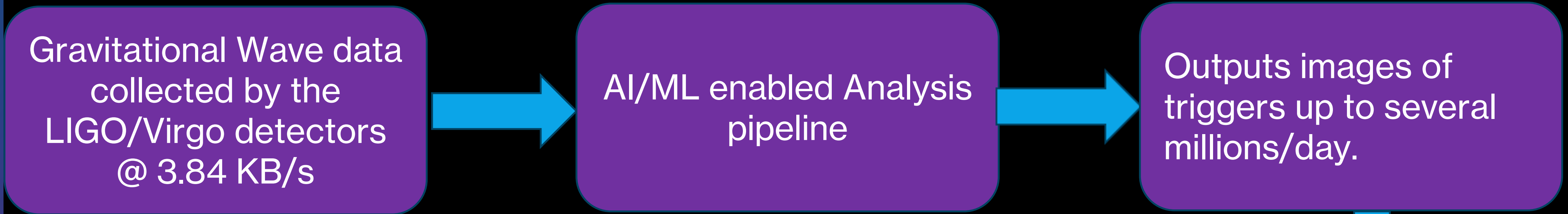
Why is AI/ML data science education essential for students?

Empowers researchers to analyze vast datasets, identify patterns, and generate insights at a speed and scale unattainable through traditional methods.

Fosters collaboration across fields, enabling researchers to apply computational intelligence to complex, cross-cutting problems.

Future-Readiness and Competitiveness: AI knowledge ensures they capable of leading in an increasingly data-driven scientific landscape.

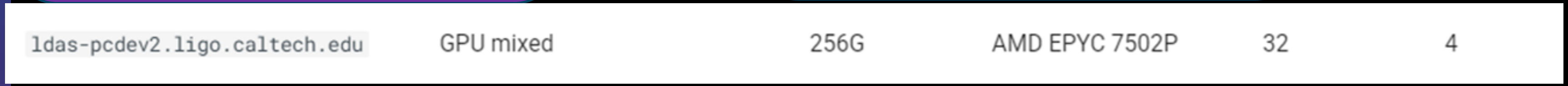
Large-scale data analysis and training in data management: An example



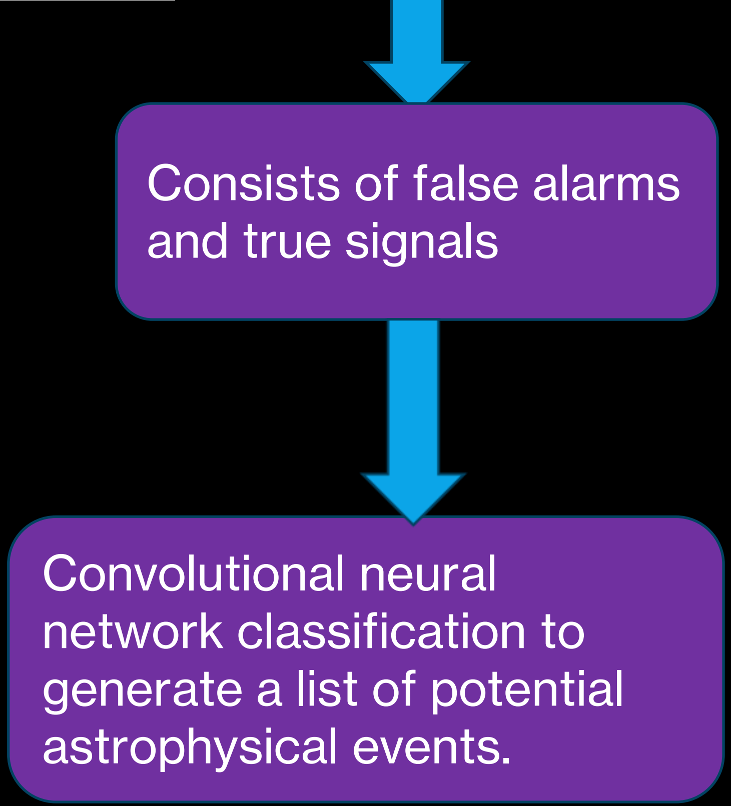
Gravitational Wave data collected by the LIGO/Virgo detectors @ 3.84 KB/s

AI/ML enabled Analysis pipeline

Outputs images of triggers up to several millions/day.



ldas-pcdev2.ligo.caltech.edu	GPU mixed	256G	AMD EPYC 7502P	32	4



Consists of false alarms and true signals

Convolutional neural network classification to generate a list of potential astrophysical events.

In LIGO, the files are stored in an integrated secure database in the LDAS system.

Store metadata and data analysis products: LDAS manages and archives important metadata associated with the LIGO data and the results of various analyses performed.

Support data processing and retrieval: The database facilitates the storage and retrieval of data as it moves through the different stages of the LDAS processing pipelines, ultimately making reduced data products available to users.

continued ...

Naming convention: The files in the database follow a standardized naming convention that includes source, GPS times, trigger parameters and other essential detector information. Specialized scripts can retrieve required data.

Resource limitation: If stored beyond a designated time period, the results repository is typically sent to “cold”, or the job priority can become lower.

Decision insight: Make educated decision what can be archived long-term and how to retrieve data quickly in case of urgency.

Building an AI-Aware Community at UTRGV

STARTER* (**South Texas AI Research, Training, and Education Resources**) is a project by the University of Texas Rio Grande Valley (UTRGV) aimed at establishing a strong foundation for AI-powered research and education. We want to empower faculty and students to become experts in AI technologies and enable them to conduct their own AI projects.

By creating an AI-aware community, STARTER will drive sustainable growth in AI-powered research and education, not just within the university but also beyond its boundaries.

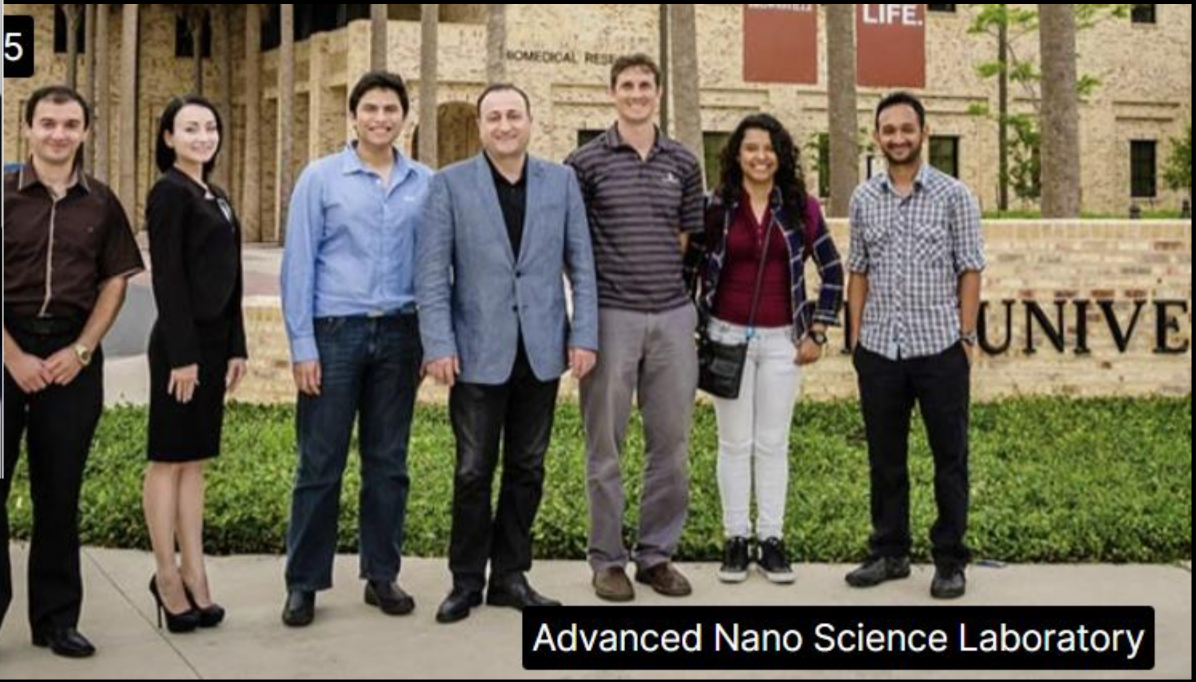
The students have taken the leap toward the future and have begun their research in AI.

STARTER-AI Workshop 2024, supported by NSF, NIH, and NVIDIA

*This project is supported by the National Science Foundation (NSF). NSF IIS-2334389. "CAP: STARTER: South Texas AI Research, Training, and Education Resource."

3 / 5

Center of Advanced Manufacturing Innovation & Cyber Systems


5

Advanced Nano Science Laboratory

4 / 5

Cyber Security Research Lab

An AI approach to the detection of Gravitational Waves from core collapse supernovae
MI@UTRGV

5 / 5

Machine Intelligence Lab (MI@UTRGV)

AI resources at UTRGV

Micro-Credential Format Courses

Micro-credential courses have been developed by our senior personnel and consist of Computer Science, Cyber Security, and programming courses.

AI Enhanced Courses

Existing courses in Physics, Computer Science, and Cyber Security are enhanced with AI-related content. These enhancements provide training in practical use of AI tools, prepare students for participating in research projects, and motivate students to further study AI-related topics.

New PhD Courses

TACC Lonestar6 Cluster

UTRGV faculty and students working under STARTER can also request time allocations (at no monetary cost) on the Lonestar6 cluster at the Texas Advanced Computing Center (TACC).

Access-Controlled Workstations

Access-controlled computer labs in both Brownsville and Edinburg campuses will be made available to students. Each workstation has a GPU and has access to the UTRGV Cradle Cluster and be able to smoothly migrate data. All GPU workstations are managed by UTRGV IT staff

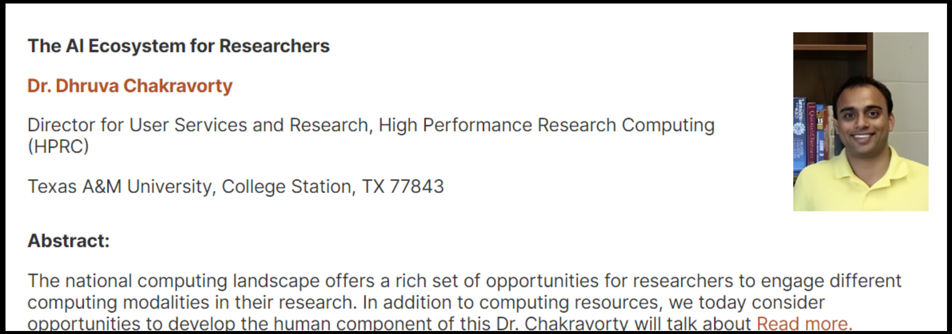
Tutorials & Hands-On Training

Our faculty personnel deliver tutorials and hands-on training sessions on various aspects of AI and machine learning using course material developed both in-house and by NVIDIA to all attending participants.

"Bring Your Own Problems"

"Bring your own problems" sessions are organized and advertised before workshops to inform and attract researchers interested in using AI and machine learning.

These sessions aim to identify startup projects within the community and provide them with the adequate support.



The AI Ecosystem for Researchers

Dr. Dhruva Chakravorty

Director for User Services and Research, High Performance Research Computing (HPRC)

Texas A&M University, College Station, TX 77843

Abstract:

The national computing landscape offers a rich set of opportunities for researchers to engage different computing modalities in their research. In addition to computing resources, we today consider opportunities to develop the human component of this Dr. Chakravorty will talk about Read more.

External Speakers

External speakers are invited to talk about AI technologies and their impacts in diverse fields, including ethical aspects of AI technology. Besides gaining insights into real-world applications of AI, interactions with external speakers open collaboration opportunities for UTRGV faculty and students.

At UTRGV, the students are not yet trained on data management. We will prioritize this topic for our Fall workshop.

Students need to learn:

- How to use HPC: Using command line interface
- How to manage large amounts of output data
- How to set up workflow to post-process
- Archival: Understanding what to archive
- Retrieval: How to retrieve data with low latency
- How to manage limited resources

Going forward

Tremendous student interest in AI/ML training programs from a wide range of disciplines in STEM and beyond.

Focus on growing the computing and data storage resources.

Prioritize data management issues in our next training workshop.

Encourage recognition of data management as a core research activity.

Explore interdisciplinary research where STEM problems drive computational research and data management growth.