Data Science Constellation

Let’s Talk Research Presentation
March 26, 2021

Leads: Melanie Mormile
V.A. Samaranayake

Presenters: Marco Cavaglia
Tie (Tony) Luo
Fiona Nah
Joe Stanley
Don Wunsch

Key to Knowledge Discovery
Outline of Presentation

➢ Why a Data Science Constellation?
➢ Major Goals
➢ A Sample of our Members’ Research
  ▪ Joe Stanley
  ▪ Marco Cavaglia
  ▪ Fiona Nah
  ▪ Tie (Tony) Luo
  ▪ Don Wunsch
Why a Data Science Constellation?

➢ Multiple groups at S&T are conducting research on DS fundamentals & applications
➢ But there is only a limited exchange of research ideas and technologies across these groups
➢ Universities that have established DS centers and research institutes have seen a transformative change with an emergence of a vibrant research community*

*Creating Institutional Change in Data Science: The Moore-Sloan Data Science Environments at New York University, UC Berkeley, and the University of Washington, referenced in How to Encourage Data-Driven Discovery March 4, 2018 | Ed Lazowska, et al | Chronicle of Higher Education
Quotes from the publication

Creating Institutional Change in Data Science – The Moore-Sloan Data Science Environments: New York University, UC Berkeley, and the University of Washington

➢ Data science is the great unifier. Its emergence is bringing about a major cultural shift in universities toward cross-disciplinary scholarship—not just between methodology fields and application fields but across methodology fields and across application fields.

➢ Our collaboration has undertaken the challenge of blazing trails into new methods, new software, new partnerships, new organizational forms, and new types of people.

➢ We are pioneering the development of tools and software environments that are sustainable, reusable, extensible, and translatable across problem domains.

➢ We have started new curricula and new degree options for students in data science to empower the next generation.

➢ We have leveraged institutional commitments and other funding to hire faculty with deep expertise in data science methodologies and a domain science so they can help lead the way through teaching and discovery as well as to support targeted data science projects in a broad range of domains ranging from astronomy and high-energy physics to neuroscience and urban science.
Our Goals

• Establish a collaborative research network that enables its members to engage in the exchange and synthesis of ideas and methods, thereby creating an incubator for transformational research

• Bring together researchers in application and foundational areas of data science

• Create a network of researchers who can respond quickly and effectively to large-scale funding opportunities and challenges

• Within the framework of the big umbrella of data science, create subgroups that work on synergistic areas such as foundational research in data science, high-performance computing, data and society, financial and business analytics, intelligence systems, etc.

• Establish and strengthen connections to research center, signature research areas, and the newly proposed Kummer Institute research centers

• Create resources and opportunities for both undergraduate and graduate students to engage in data science research that will prepare them for the modern workplace or for future research careers

• Establish cross-disciplinary data science emphasis areas in undergraduate and graduate degree programs, including those at the Kummer School of Entrepreneurship, which will integrate data science research with the university’s core mission of education.
A Sample of our Members’ Research

Dr. Joe Stanley
Professor
Computer Engineering

stanleyj@mst.edu
Phone: (573) 341-6896
127 Emerson Electric Company Hall
Digitized Histopathology Image Analysis for Pre-Cervical Cancer Screening

Problem Overview

- Cervical Cancer, fourth most frequent cancer affecting women worldwide.
- WHO 2018 Statistics:
  - Estimated new cases: 570,000.
  - Approx. deaths: 311,000.
  - Low- and middle-income countries account for 90% of deaths.
- The cancer can be prevented if detected early.
- An automated whole slide image analysis tool is necessary to assist pathologists.

Missouri S&T
- Dr. R. Joe Stanley

Stoecker & Associates
- Dr. William V. Stoecker

University of Oklahoma
Health Sciences Center,
Department of Pathology
- Dr. Rosemary Zuna

University of Missouri,
Department of Health and
Anatomical Sciences
- Dr. Shellaine Frazier

PhD Graduate Students
- Dr. Soumya De*
- Dr. Beibei Cheng*
- Dr. Peng Guo*
- Dr. Haidar AlMubarak*
- Dr. Sudir Sornapudi*
- Jason Hagerty
- Akanksha Maurya

* PhD graduate from Missouri S&T with dissertation topic from this project

Project funded by: NIH U.S. National Library of Medicine
Benchmark is the pathologist’s manual process.
Ongoing Research

• Investigation of Deep Learning and data fusion techniques for automated epithelium segmentation and cervical intraepithelial neoplasia classification

• Open source webtool development
  o User whole digitized histology slide processing toolbox available to pathologists, experts, researchers, students, etc.
  o Publicly accessible database
  o Database query tool for dataset creation and sharing
  o Annotated example images for training
A Sample of our Members’ Research

Dr. Marco Cavaglia
Professor
Physics

cavagliam@mst.edu
204 Physics Building
WoU-MMA: Enabling Multi-Messenger Astrophysics with Advanced LIGO: From Detector Calibration to Interpretation of Gravitational-wave Signals

Project overview

LIGO scientific operations: Run astrophysical multi-messenger searches for gravitational-wave burst sources, characterize instrumental data, calibrate the detectors, assist operations on site, improve existing data analysis algorithms.

R&D: Develop new machine learning-based methods for denoising data, extending the reach of astrophysical searches for gravitational-wave signals, and characterizing LIGO’s multi-messenger observations.

Missouri S&T
- Dr. Marco Cavaglia
- Dr. Ryan Quitzow
- James (postdoc)
- Dr. Sudarshan Karki (postdoc)

PhD Graduate Students
- Ms. Dripta Bhattacharjee
- Mr. Sushant Chaudhary
- Sharma
- Mr. Kentaro Mogushi
- Ms. Yanyan Zheng

LIGO Scientific Collaboration
- 1200+ members
- 100+ institutions
- 20+ countries

Undergraduate Students
- Mr. Alex Love
- Mr. Mason Labrot

High school Students
- Ms. Ashini Modi
- Mr. Cole Johnson

Project funded by: (S&T awards PHY-1921006 and PHY-2011334)
Big data in multi-messenger astrophysics
(Sample of) Ongoing Research

- Deep Learning for denoising of gravitational-wave signals and detector data quality prediction.
- Genetic Programming algorithms for prediction of electromagnetic signatures of gravitational-wave signals
- Investigation of machine learning-based techniques for reduction of detector background in gravitational-wave searches
- Recent selected publications:
A Sample of our Members’ Research

Dr. Fiona Nah
Professor
Business & Information Technology

nahf@mst.edu
Phone: (573) 341-6996
101 Fulton Hall
Physiological Indicators of User States and Experience

- Dr. Fiona Nah, Department of Business and Information Technology
  - Behavioral/social science researcher in human-computer interaction/user experience (HCI/UX)
  - BS (Hons) & MS – Computer and Information Sciences
  - PhD in Business with specialization in Information Systems & minor in Psychology
  - Editor-in-chief, AIS Transactions on Human-Computer Interaction

- Data Science questions of interest
  - A challenge in behavioral/social science research as compared to STEM research
    - Measurement can be challenging and subjective (e.g., self-reported or based on observations)
    => in search of more objective measures
  - E.g., user states and experience such as sleepiness, engagement/flow, and anxiety

Brain Mapping: Neural Correlates of Flow Experience using Electroencephalogram (EEG)
Graduate (MS) students: Tejaswini Yelamanchili & Chandana Mallapragada (graduated)
Collaborators: Langtao Chen (BIT), Keng Siau (BIT), & Don Wunsch (ECE)
New Ph.D. student collaborator: Sima Azizi (ECE)

Eye-tracking Metrics for Sleepiness
Graduate student: Debasis Roy
Collaborator: Matt Thimgan (Biology)
Neural Correlates of Flow Experience using EEG

Flow experience
- Cognitive absorption or focused concentration
- Total involvement/immersion
- Lose track of time
- Loss of self-consciousness
- Merging of action & awareness
- Self-rewarding experience

Lab experimental study
- Within-subjects design
  - Tetris game
- Retrospective processing tracing
  - Identify (EEG) segments corresponding to flow experience
    - Compare with other states such as resting state
- Machine learning to identify brain mapping and activities
Eye-tracking Metrics for Sleepiness

- Lab experimental study (within-subjects design)
  - Subjects volunteer to come to the lab for eye-tracking sessions
    - when they are deeply deprived of sleep (e.g., no or little sleep the night before)
    - when they are in a fully alert and wakeful state (i.e., with at least 8 hours of sleep the night before)
  - Subjects carry out various tasks in the lab, including
    - Psychomotor Vigilance Task (PVT) that captures reaction time
    - Reading and Comprehension Task (RCT)
- Eye-tracking data are captured, and tests will be conducted to assess the reliability and validity of various metrics in detecting sleepiness
  - Blink: frequency, duration, interval (i.e., between blinks), and re-opening time
  - Fixation: duration, rate
  - Saccade (eye movement between fixations): accuracy, latency, curvature/amplitude, (peak) velocity
  - Smooth pursuit (eye movement to follow a moving target): velocity gain, accuracy
  - Eye(lid) closure: percent, speed
  - Gaze: stability, direction, entropy/dispersion
  - Pupil: diameter/size, latency to constriction
A Sample of our Members’ Research: Data Science for IoT and Cybersecurity

Dr. Tie (Tony) Luo
Associate Professor
Computer Science

tluo@mst.edu
Phone: (573) 341-4788
325 Computer Science Building
Data Science for IoT (1):
Distributed Anomaly Detection [Luo & Nagarajan, ICC’18]

• Goal: To achieve fast and accurate anomaly detection in resource-constrained wireless sensor networks
• Propose: a lightweight autoencoder neural network and a distributed two-part algorithm
  - Autoencoder: unsupervised learning that tackles the challenge of no labelled data (anomalies)
  - Two-part algo: separates training and inference to minimize communication and edge computation
• Result: real-time detection at sensor nodes with low misdetections and false alarms

Sensor Data Collection
AUC > 0.8 in most cases
Data Science for IoT (2):
Anomaly Detection in Edge Computing [Ngo & Luo, AAAI’20]

- Existing AD solutions are “one-size-fits-all”:
  - Algorithm runs either at edge or in cloud
  - At edge: suffers from low accuracy
  - In cloud: suffers from long latency
- Can we make adaptive choices (based on the complexity of input data)?
- Construct 3 AD models w/ different complexity, host them in a 3-layer hierarchical edge computing architecture
- Use a policy network (a reinforcement learning technique) to decide which layer to perform AD
- Achieves both high accuracy and low latency for all data
Data Science for Cybersecurity: MQTT Attack with Adversarial BERT [Wong & Luo, KDD’20]

- We design a **Man-in-the-Middle attack** targeting MQTT-based networks
- Man-in-the-Middle attack is more harmful than commonly-seen attacks (e.g., eavesdropping, DoS, jamming, Sybil, spoofing) because it performs direct Read & Write actions
- We use **BERT** (a deep learning technique in **NLP**) to generate adversarial messages to substitute the original benign MQTT messages
- Our designed attack successfully evades a wide range of common defense mechanisms (LR, RF, KNN, SVM, MLP)
- Obtained consistent results w/ different NLP transformers: DistilBERT, Vanilla BERT, RoBERTa, GPT-2
A Sample of our Members’ Research

Dr. Don Wunsch
Mary K. Finley Missouri Distinguished Professor
Applied Computational Intelligence Laboratory

dwunsch@mst.edu
Phone: (573) 341-4521
131 Emerson Electric Company Hall
Mixed-Modality Learning for Lifelong Learning Machines

M2L for L2M — Phase 2

- L2M Challenge, keep improving
- M2L solution UL->RL->SL (w Teledyne)

Context 1

Context 2

Context 3

Context 4

Context Features

Context Label
Questions?