

Introduction to Big Data and Data Science (CSCE 5300 Section 001)*

Yunhe Feng

Assistant Professor, Department of Computer Science and Engineering

28th August, 2024



About Me

- Assistant Professor at CSE, UNT
- Director of the [Responsible AI Lab](#) (We are looking for talented students!)



About Me - Education Background and Professional Experience

- Education Background:

- Ph.D. in Computer Science, University of Tennessee Knoxville, 2020
- M.E. in Computer Technology, Beijing University of Technology, 2014
- B.E. in Computer Science & Technology, Beijing University of Technology, 2011

- Postdoctoral Training:



- Doctoral Training:



OAK RIDGE NATIONAL LABORATORY

Artificial Intelligence (AI) Initiative

Computing and Computational Sciences Directorate

Computational Sciences and Engineering Division

Physical Sciences Directorate

Materials Science and Technology Division

Contact Me

- Office: F251 CSE, Discovery Park
- Office Hour: 3:00pm-5:00pm Friday
- Email: Yunhe.Feng@unt.edu
- Homepage: <https://yunhefeng.me/>

Contact IA/TAs



Figure 1: Sumera Anjum



Figure 2: Hanzhi Zhang

Contact IA/TAs

• Sumera Anjum

- Office: E247 A, Discovery Park
- Office Hour:
 - 11:30am-3:30pm, Monday
- Email: sumeraanjumsumeraanjum@my.unt.edu

• Hanzhi Zhang

- Office: F221, Discovery Park
- Office Hour:
 - 2:00pm-5:00pm, Tuesday
- Email: hanzhizhang@my.unt.edu

1 Course Syllabus

2 Concepts of Big Data and Data Science

3 Course Overview

4 Assignment

1 Course Syllabus

② Concepts of Big Data and Data Science

3 Course Overview

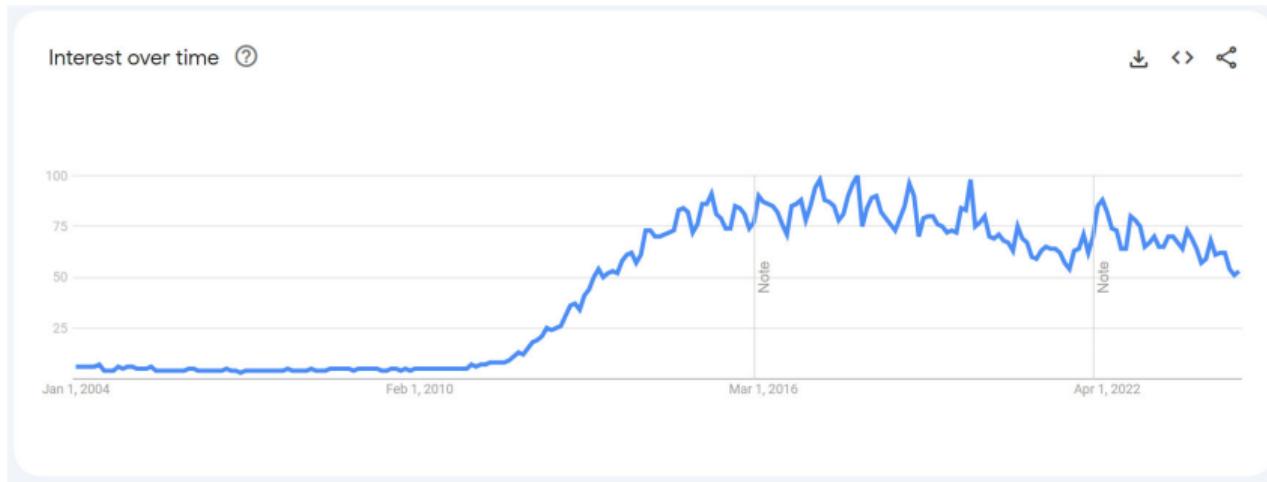
4 Assignment

Understanding Big Data and Data Science

- What is Big Data
- Why Big Data
- What is Data Science
- Why Data Science
- Data Science VS Deep Learning

What is Big Data

- Big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. - Wikipedia.
- The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. - Oracle

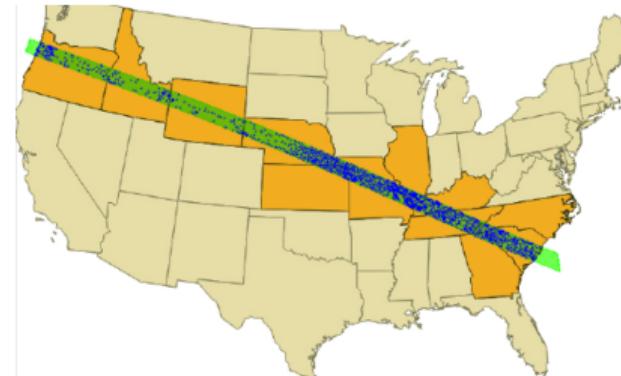
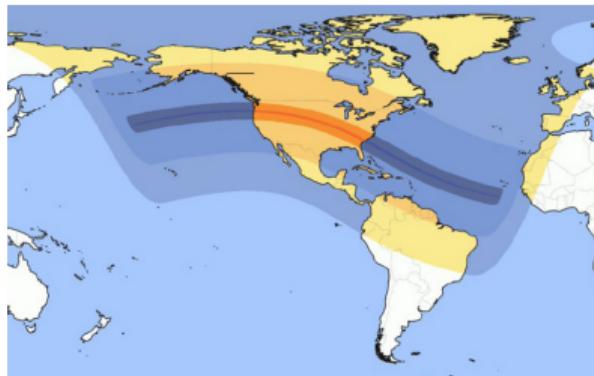


5 Vs of Big Data

- Volume
- Variety
- Velocity
- Veracity
- Value

Examples of 5Vs of Big Data - Chase Eclipse on Twitter: Big Social Data Analytics in Once-in-a-lifetime Events [Feng et al., IEEE GlobeCom'19]

2017 Great American Eclipse - According to The Washington Post¹, nine in ten adults in the U.S. watched this eclipse.



¹B. Guarino, Wildlife fled the sky and bees went silent during the 2017 total solar eclipse, 2018. [Online]. Available: <https://wapo.st/2Ok4G2D>

Examples of 5Vs of Big Data - Chase Eclipse on Twitter: Big Social Data Analytics in Once-in-a-lifetime Events [Feng et al., IEEE GlobeCom'19]

- **Volume:** 5.97 million tweets including original tweets, replies, retweets, and quoted tweets, and 1.17 million unique tweets.
- **Variety:** Structured data (e.g., emojis, emoticons) and unstructured data (e.g., plain text, #hashtags, and geolocations).
- **Velocity:** > 2500 tweets per minute when the eclipse occurred.
- **Veracity:** Remove tweets created by bots to ensure data quality.
- **Value:** The first investigation of how once-in-a-lifetime events influence human behaviors in large scale.

We Are in the Era of Big Data

- **Explosive Data Growth:**

- Global data is doubling every two years.
- Volume of data from various sources: social media, sensors, transactions, etc.

- **Big Data Technologies:**

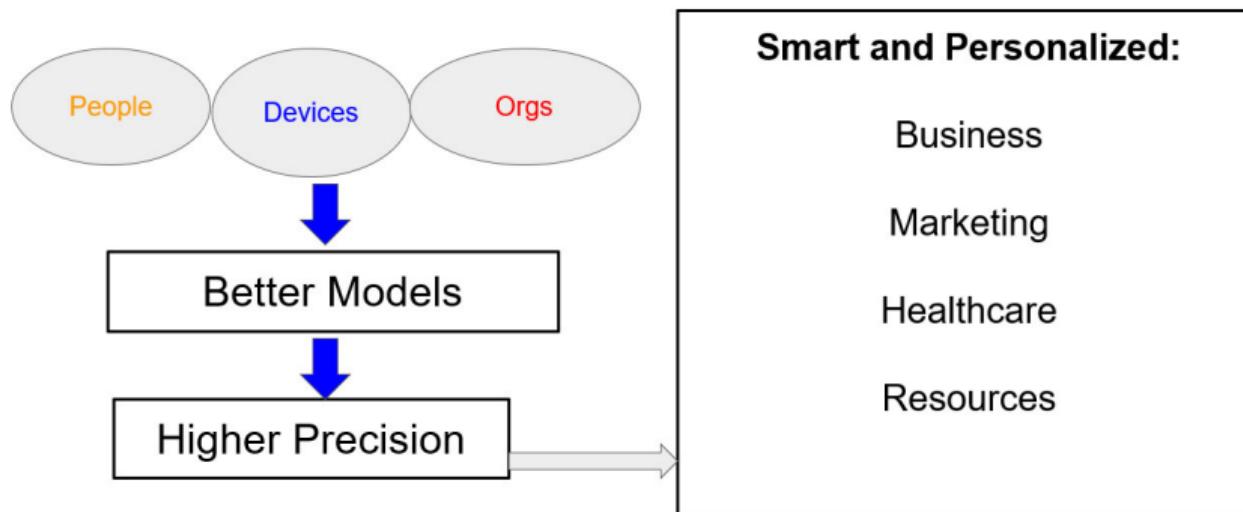
- Distributed storage and processing (e.g., Hadoop, Spark).
- NoSQL for managing unstructured data.
- Data lakes and real-time analytics platforms.

- **Impact on Industries:**

- Transforming sectors like healthcare, finance, retail, and more.
- Enabling data-driven decision-making and automation.



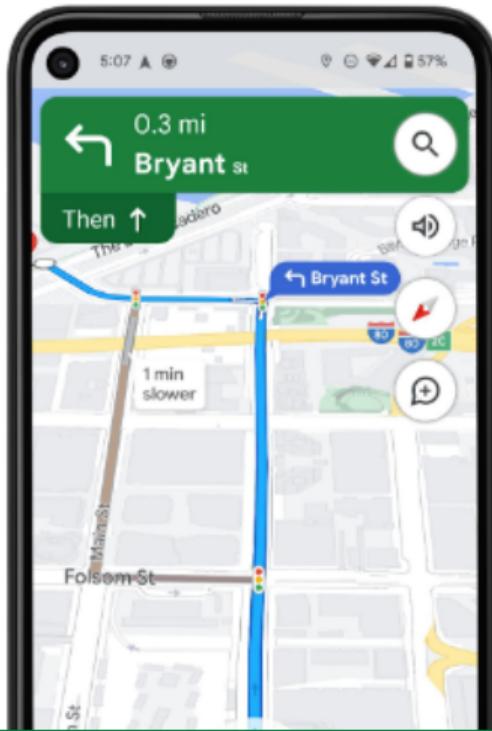
Why Big Data



Why Big Data - Captcha



Why Big Data - Google Map



*The teaching materials are reorganized and reformed based on Prof. Ravi Vadapalli's slides (Ravi.Vadapalli@unt.edu, UNT & University of Miami)

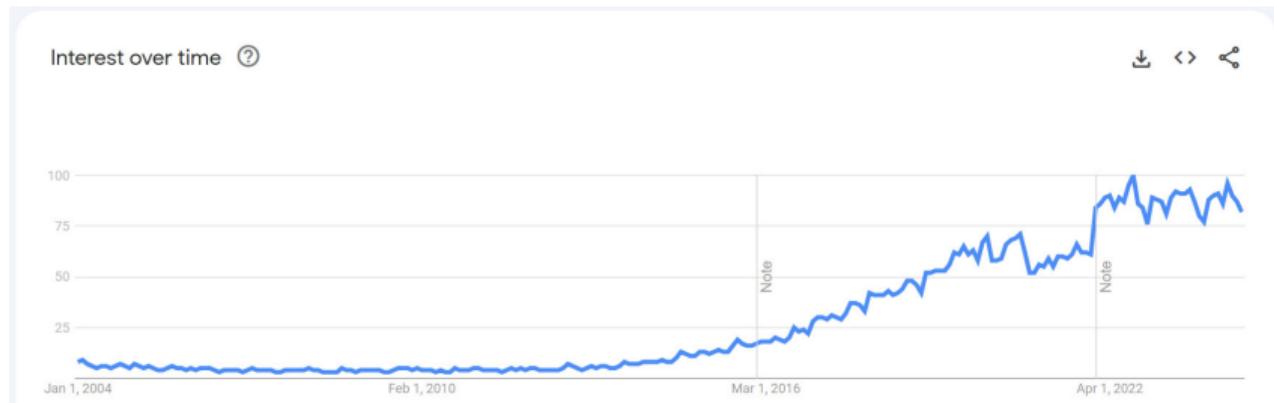
Introduction to Big Data and Data Science (CSCE 5300 Section 001)*

Why Big Data - Large Language Model (LLM) Training

	Training Set (Words)	Training Set (Tokens)	Relative size (Llama 3 = 1)
Recent LLMs			
Llama 3	11 trillion	15T	1
GTP-4	5 trillion	6.5T	0.5
Humans			
Human, age 5	30 million	40 million	10^{-6}
Human, age 20	150 million	200 million	10^{-5}

What is Data Science

- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. - Wikipedia
- Data scientists examine which questions need answering and where to find the related data. - UC Berkeley



Why Data Science

- **Exploration:** “story” of the collected data? data patterns?
 - Visualization and descriptive statistics
- **Prediction:** it involves “guessing” what it means to us! What we wish we knew!
 - Machine Learning and optimization
- **Inference:** it involves establishing a degree of “certainty” of our guess work! Is this information new? Can we quantify?
 - Statistical tests and models

Data Science VS Deep Learning

- Artificial Intelligence: Human Intelligence Exhibited by Machines
- Machine Learning: An Approach to Achieve Artificial Intelligence
- Deep Learning: A Technique for Implementing Machine Learning

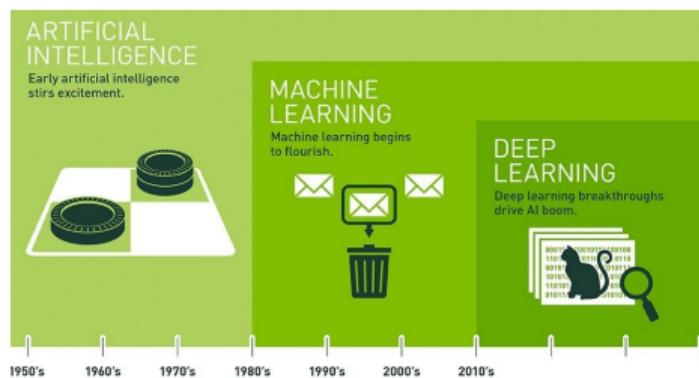


Figure 5: What's the Difference Between Artificial Intelligence, Machine Learning and Deep Learning - Nvidia

Data Science VS Deep Learning

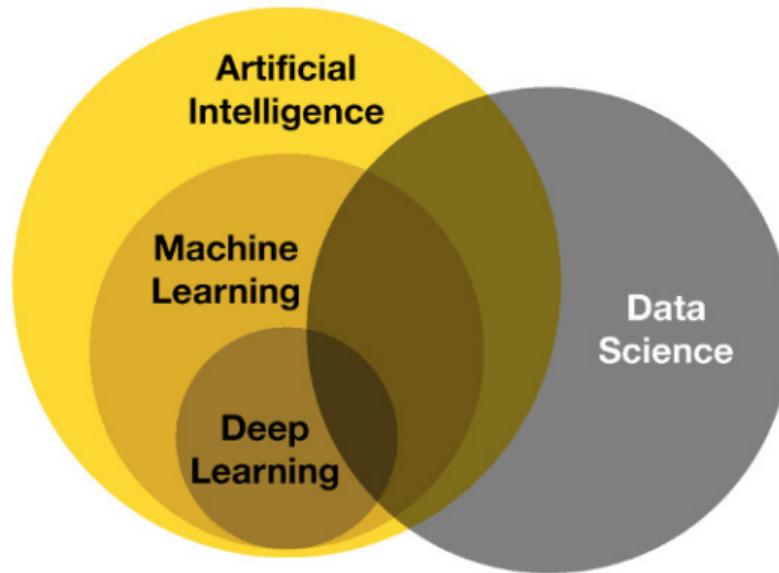


Figure 6: Deep Learning VS Data Science²

²<https://www.deviq.io/insights/artificial-intelligence-vs-machine-learning-vs-data-science> ▶ 🔍 ↻ ↺ ↻ 🔍 ↺ ↻

Data Science VS Deep Learning



Figure 7: ChatGPT - Are Data Science Jobs Now Obsolete?³

³<https://betterdatascience.com/chatgpt/>

ChatGPT - Explain big data in simple terms

YU

Explain big data in simple terms



Big data refers to extremely large and complex sets of data that are difficult to process and analyze using traditional methods. These data sets are often so large and complex that they cannot be easily stored, managed, or analyzed by a single computer or system. Examples of big data include social media activity, online shopping data, and sensor data from Internet of Things devices. The goal of big data technology is to enable organizations to extract valuable insights and knowledge from these large and complex data sets, in order to make better decisions and improve their operations.

ChatGPT - Explain the linear regression formula in machine learning

yu

Explain the linear regression formula in machine learning



Linear regression is a statistical method used in machine learning for predicting a continuous target variable based on one or more input variables. It models the relationship between the data-points by fitting a linear equation to the data, i.e $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$

where:

- y is the target variable (also called the dependent variable)
- x_1, x_2, \dots, x_n are the input variables (also called the independent variables)
- $b_0, b_1, b_2, \dots, b_n$ are the coefficients of the equation, which are estimated from the data

The goal of linear regression is to find the values of the coefficients that minimize the difference between the predicted values of y and the actual values of y . This difference is measured using a loss function, such as mean squared error. Once the values of the coefficients are estimated, the linear regression model can be used to make predictions for new data.

ChatGPT - Do you think you could replace a Data Scientist?

yu

Do you think you could replace a Data Scientist?



As a language model, I have been trained on a large dataset of text and can provide useful insights and predictions based on the information I have been trained on. However, data science is a complex field that involves not only the ability to analyze and model data, but also the ability to interpret results, communicate findings, and make decisions based on those findings. Additionally, data scientists often have a deep understanding of the domain in which they are working, which allows them to ask the right questions, identify the most relevant data, and make informed decisions.

While I can assist data scientists with some of their tasks, such as data cleaning, feature engineering, and modeling, I cannot replace the role of a data scientist. A data scientist's role is much more than just writing code or running models, it's a combination of skills, including domain knowledge, data visualization, story-telling and problem solving.

In short, I am a tool that can be used to aid in data analysis and modeling, but I cannot replace the expertise, creativity, and decision-making abilities of a human data scientist.

1 Course Syllabus

2 Concepts of Big Data and Data Science

3 Course Overview

4 Assignment

Intended Topics

- Python Programming
- Data Visualization
- Basic Data Structures
- Machine Learning
- Deep Learning
- Big Data: Hadoop and Spark
- Parallel Computing with Python
- Linux: GCP and VirtualBox VMs
- Research (Concept) Paper or Extra Assignment

Python Programming



Figure 8: The Jupyter Notebook is a web-based interactive computing platform.

Data Visualization

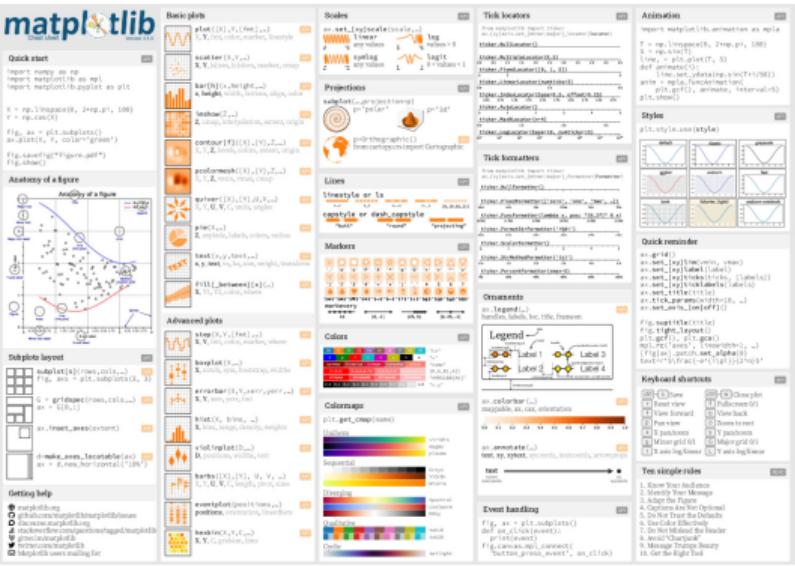


Figure 9: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Basic Data Structures

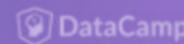
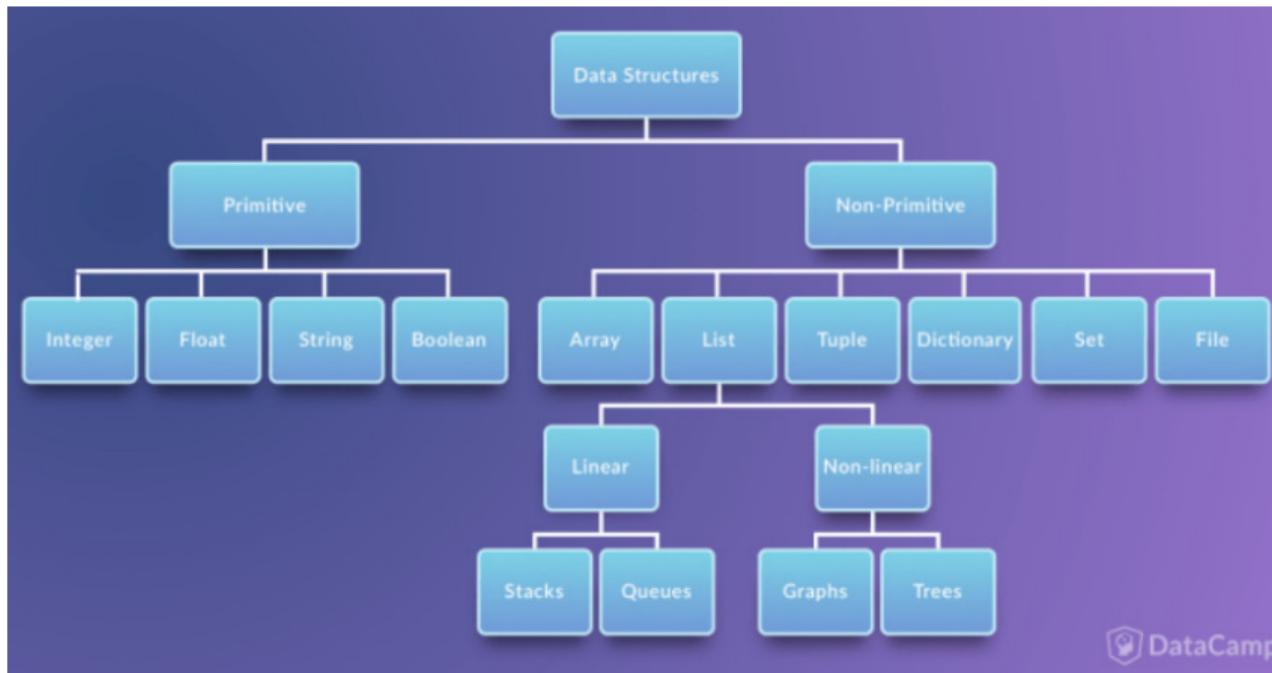


Figure 10: Python data structures.

Machine Learning and Deep Learning

- Concepts
- Classification
- Regression
- Convolutional neural network (CNN, or ConvNet)

Hadoop and Spark

WHAT IS



WHAT IS



Big data processing engine

- o Hadoop Distributed File System (HDFS)
- o MapReduce Programming Model
- o YARN

Data Analytics Engine

- o Spark Core
- o Spark SQL
- o Spark Streaming

net solutions

Guest Lecturer - Learning Apache Spark with Python



- American Express: Sr. Manager-Data Engineer, Former Director of AI Engineering
- Author of online free book (487 pages) – **Learning Apache Spark with Python**
- University of Minnesota IMA Data Science Fellow
- Ph.D. in Computational Mathematics

Figure 11: Dr. Wenqiang Feng

Parallel Computing with Python: MPI4py

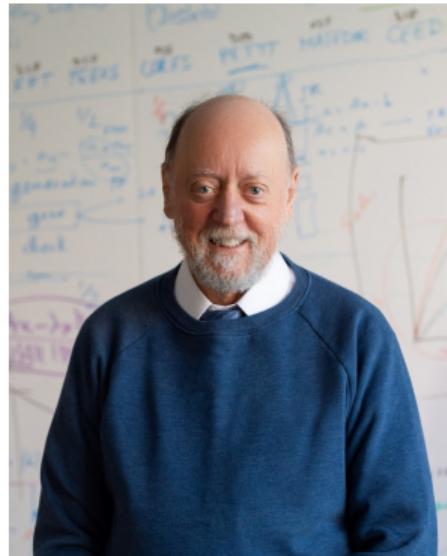
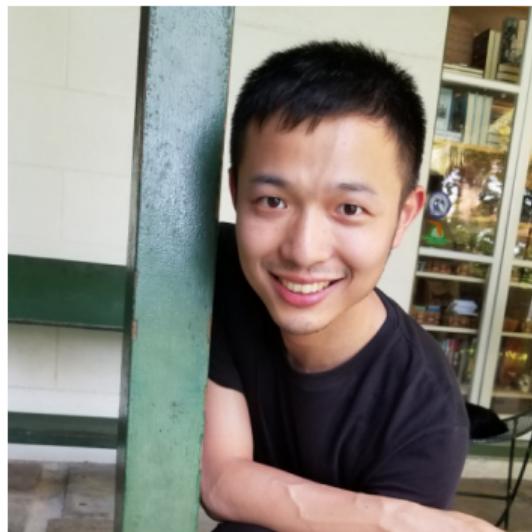


Figure 12: Dr. Jack Dongarra (Turing Award Winner 2022, Professor in University of Tennessee), along with Tony Hey, and David W. Walker, put forward a preliminary draft proposal, **MPI1**, in November 1992.

Guest Lecturer - Introduction to Parallel Computing



- Intel: Middleware Development Engineer (2022-now)
- Stanford-SLAC National Accelerator Laboratory - Associate Staff Scientist (2020-2022)
- Thesis: Optimization of MPI Collective Communication Operations

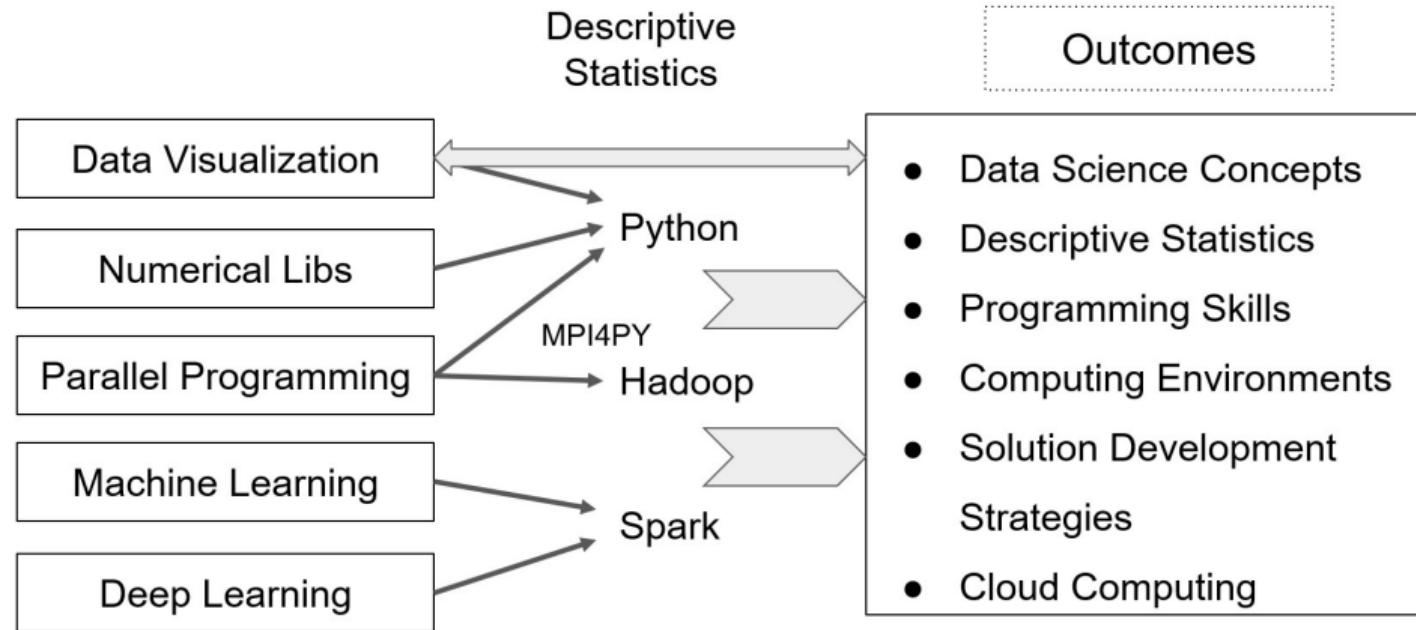
Figure 13: Dr. Xi Luo

Linux: VirtualBox VMs, GCP, or Physical Linux OS



Google Cloud Platform

Course Workflow & Outcomes



1 Course Syllabus

2 Concepts of Big Data and Data Science

3 Course Overview

4 Assignment

Assignment-0 (1.0 pt.)

- Install Jupyter Notebook (0.5 pts.)
- Install Linux VM and FileSharing using VirtualBox (0.5 pts.)
- Due time: 11:59:59 PM, August 28, 2024 (End of Next Wednesday)
- How to submit your assignment will be announced on Canvas