

CS 444: Big Data Systems

Chapter 1. Introduction

Chase Wu

Professor, Associate Chair of Data Science

Director of Center for Big Data

New Jersey Institute of Technology

chase.wu@njit.edu

Course Website

- Google “Chase Wu”, go to Dr. Wu’s personal website, click “Teaching Courses” on the left panel, and click the link on the top.
 - https://web.njit.edu/~chasewu/Courses/Spring2025/CS444BigData/CS444_BigData_Spring25.html
 - The slides will be uploaded to the course website AFTER we finish each chapter.
- Check out this course website on a regular basis for homework/project assignments, reading materials, tutorials, etc.

The 1st Class Attendance Check

Two purposes:

- Earn your very first attendance credits;
- Have an opportunity to learn about each other's education/research background or work experience to possibly form a team with common interests for homework or projects.

- **Name**
- **Program/Year**
- **Why do you take this course?**
- **What is the largest data size you've ever personally handled and in what context?**
 - application domain
 - data type
 - file format
 - processing/analysis tools and purposes
 - etc.
- **How many of the following buzzwords have you heard of?**
 - Hadoop
 - MapReduce
 - Spark
 - NoSQL
 - HDFS
 - Pig
 - Voxel
 - HBase
 - YARN
 - Containerization
 - MPI
 - Mahout
 - Naïve Bayes
 - ChatGPT
 - K-means
 - Supercomputer

Order of Magnitude:



About this course

- **Recent Developments and Future Trends on Big Data Computing**
 - Continuum Computing: from Edge to Cloud
 - High-performance Computing: Supercomputer, Cluster, etc.
- **Overview of Big Data Analytics**
- **Big Data Ecosystem**
 - Systems, Platforms, Tools, and Techniques for Big Data Transfer, Storage, Management, Computing, Processing, and Analysis
- **Machine Learning for Big Data**
- **Advanced Topics:**
 - **Big Data Meets Large Models**
 - **Big Data Visualization**
 - **Big Data Transfer**
 - **Big Data Workflows**
 - **Big Data Security**

November 2024

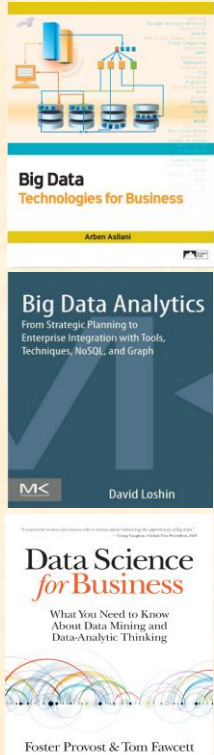
Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	El Capitan - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	Aurora - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	Eagle - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
5	HPC6 - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461

State of the arts about big data:

- **Networking: 100's Gbps (backbone)**
- **Storage: PB/EB**
- **Computing: EFlop/s first EVER!**

Textbooks and Reference Books

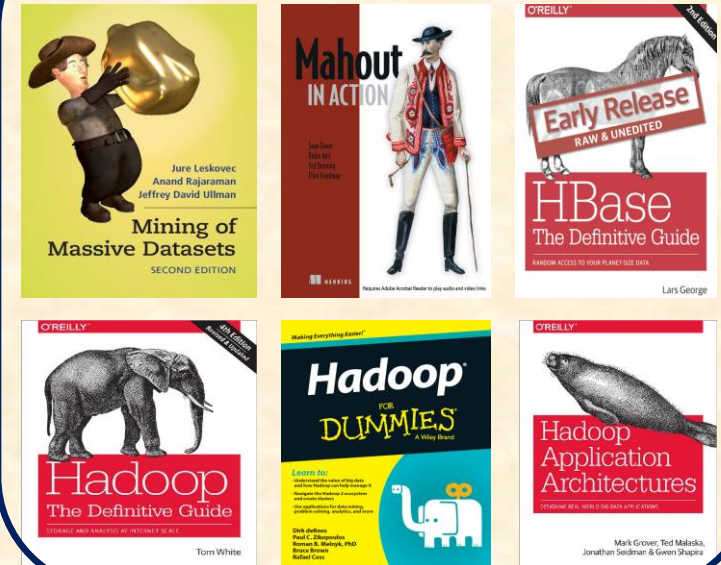
Overview



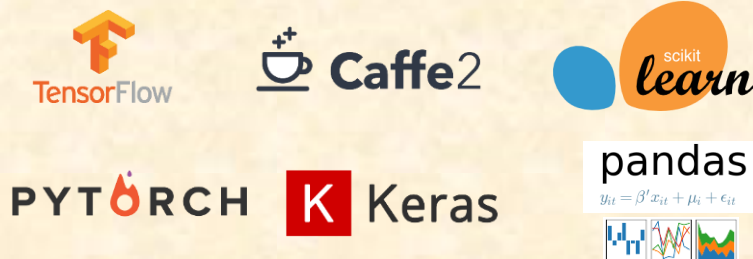
Machine Learning / Data Mining



MapReduce / Hadoop



Popular Frameworks



Learning Theory



Four V's of Big Data

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

WORLD POPULATION: 7 BILLION

100 TERABYTES
[100,000 GIGABYTES]
of data stored

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

there will be
18.9 BILLION NETWORK CONNECTIONS

— almost 2.5 connections per person on earth



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States



4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

30 BILLION PIECES OF CONTENT

are shared on Facebook every month



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

In one survey were unsure of how much of their data was inaccurate

icipated

WIRELESS TORS

BILLION+ OF VIDEO

shed on each month



TWEETS

by about 200 active users

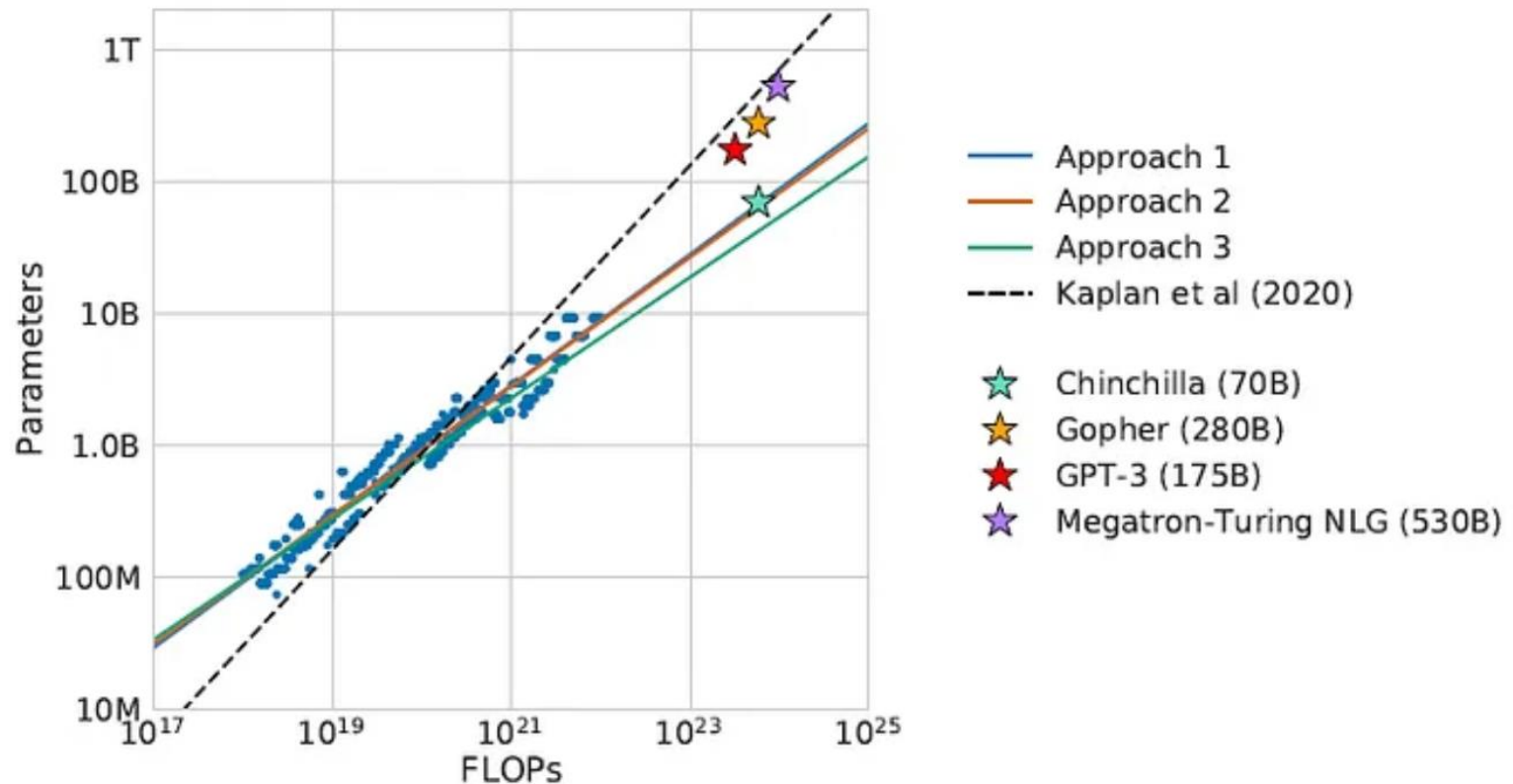
costs the US

A YEAR



Big Data and HPC for LLMs

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
<i>Chinchilla</i>	70 Billion	1.4 Trillion



Center for Big Data

Director: Chase Wu

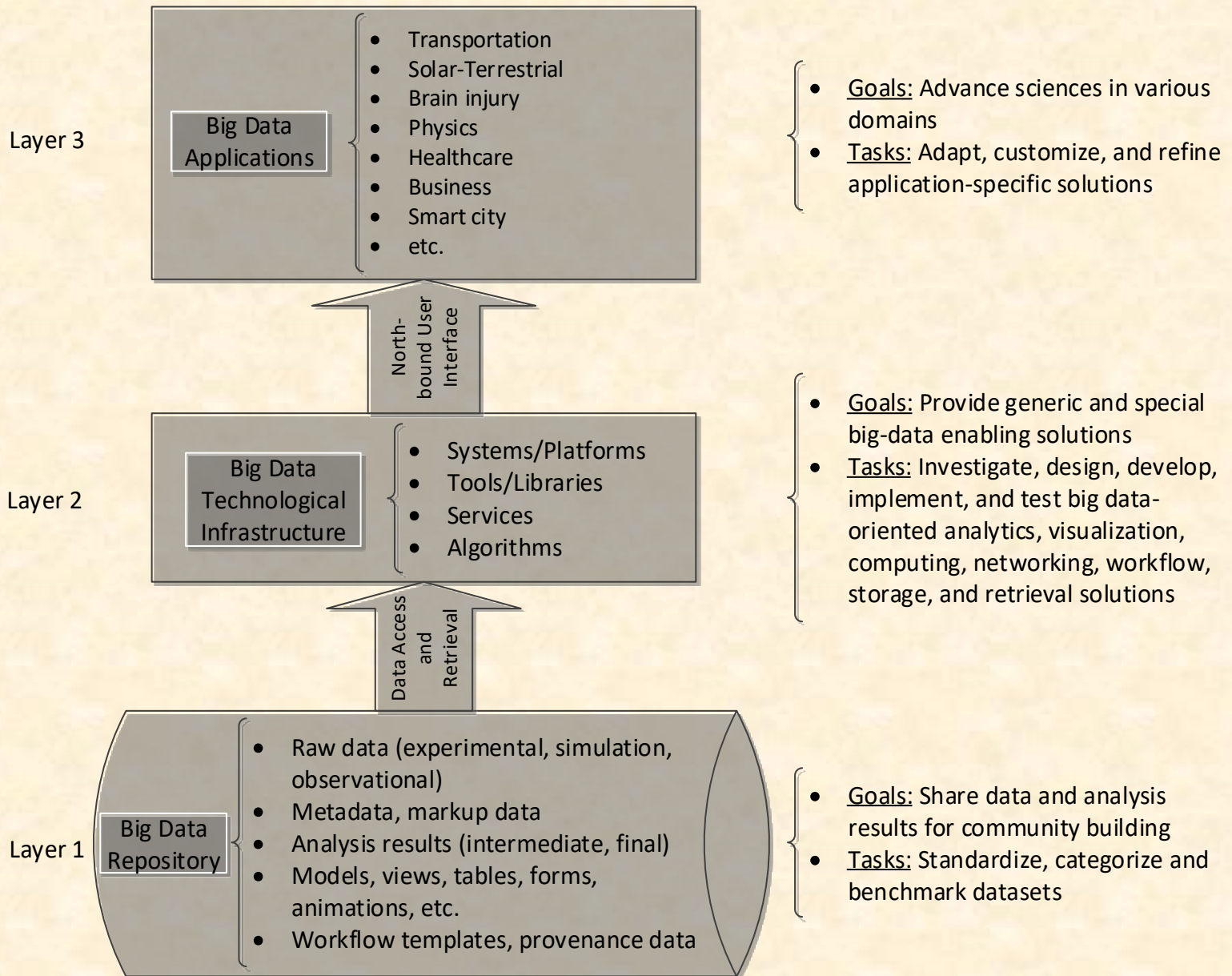
URL: <https://centers.njit.edu/bigdata>

Location: GITC 4416

Mission Statement

- Synergize the strong expertise in various disciplines across the NJIT campus
- Build a unified big data platform that embodies a rich set of big data enabling technologies and services with optimized performance to facilitate research collaboration and scientific discovery
- Investigate, develop, and apply cutting-edge technologies to address unprecedented challenges in big data with high **Volume**, high **Velocity**, high **Variety**, and high **Veracity**,
in order to create high **VALUE!**

A Three-layer Structure of the CBD



– Layer 1: Big Data Repository

- Store, manage, and provide a wide variety of data such as raw data (experimental, simulation, observational, and user-generated content), metadata, markup data, analysis results (intermediate and final) in various forms including models, views, tables, images, and videos, and workflow templates with provenance data.
- Build a dedicated one-stop portal to share research data and analysis results for community building.

– Layer 2: Big Data Technological Infrastructure

- Provide generic and domain-specific big data enabling solutions for data management, movement, and analytics.
- Host and maintain a set of practical technical resources in the form of systems/platforms, tools/libraries, services, and algorithms in various areas including database management, data mining, machine learning, and parallel and distributed computing, which are needed to compose big data solutions in different application domains.

– Layer 3: Big Data Applications

- **Present a common portal to big data applications spanning across a wide spectrum of research fields, including**
 - transportation
 - solar-terrestrial
 - brain injury
 - physics
 - healthcare
 - business
 - smart city
- **Provide researchers powerful and customized big data solutions to advance the frontier of sciences in various application domains.**

Core Faculty of CBD

- **Chase Wu (Director)** Professor, Dept of Data Science
- **Dantong Yu (Co-Director)** Associate Professor, Leir Chair, School of Management
- **Yi Chen** Professor, Leir Chair, School of Management, Dept of Computer Science

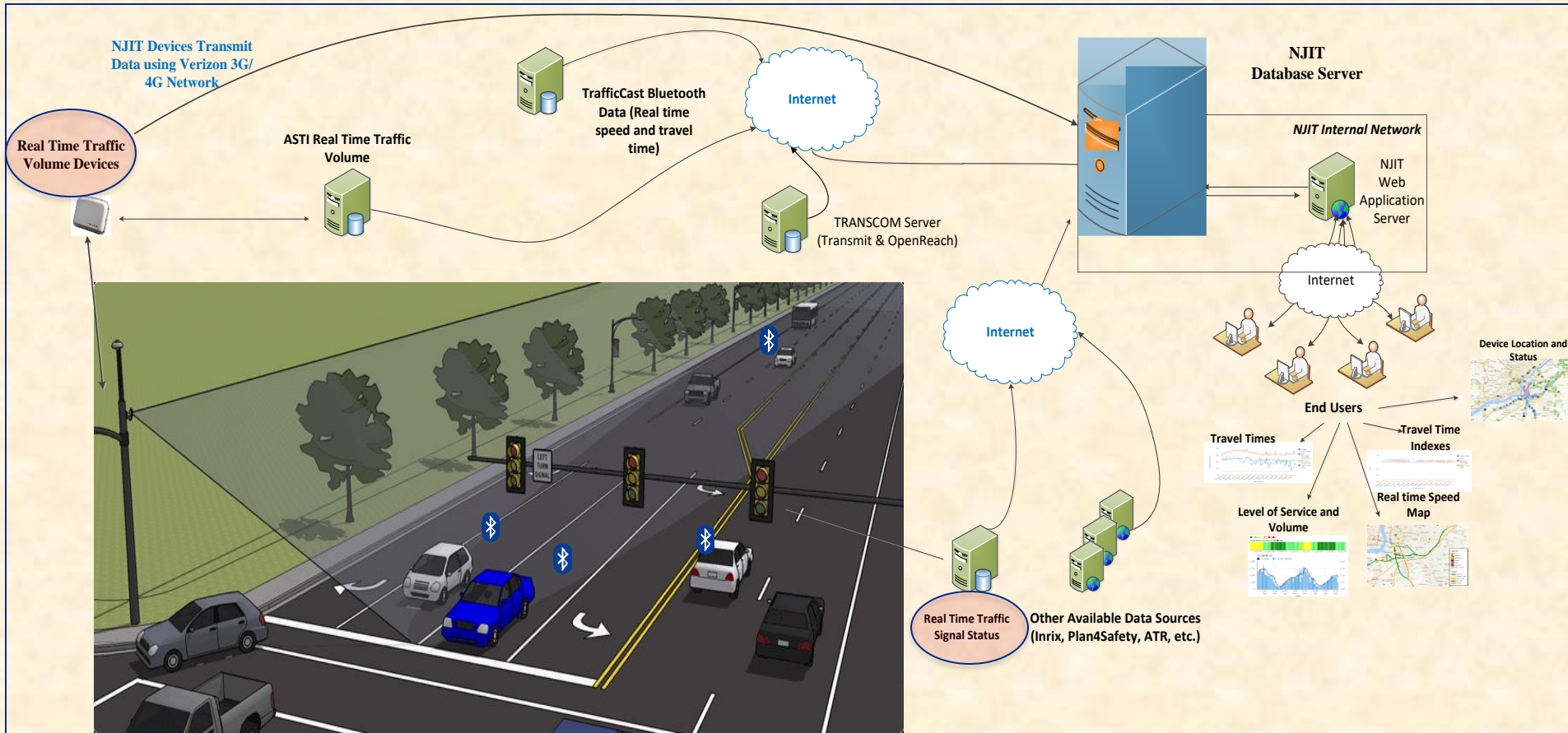
- **Andrew Gerrard** Professor, Dept of Physics, Center for Solar-Terrestrial Research
- **Lazar Spasovic** Professor, Dept of Civil and Environmental Engineering
- **Steven Chien** Professor, Dept of Civil and Environmental Engineering
- **Joyoung Lee** Assistant Professor, Dept of Civil and Environmental Engineering
- **Namas Chandra** Professor, Dept of Biomedical Engineering, Center for Injury Bio-mechanics, Materials and Medicine

- **Jason Wang** Professor, Dept of Computer Science
- **Usman Roshan** Associate Professor, Dept of Computer Science
- **Zhi Wei** Professor, Dept of Computer Science
- **Dimitri Theodoratos** Associate Professor, Dept of Computer Science
- **Vincent Oria** Professor, Dept of Computer Science
- **Senjuti Roy** Associate Professor, Dept of Computer Science
- **Brook Wu** Associate Professor, Dept of Informatics
- **Hai Phan** Assistant Professor, Dept of Data Science

Funded Projects

- DOE: Technologies and Tools for Synthesis of Source-to-Sink High-Performance Flows, DOE Office of Science, Big Data-Aware Terabits Networking.
- NSF: An Integrated Approach to Performance Modeling and Optimization of Big-data Scientific Workflows, Computer and Network Systems.
- DOE: Towards a Scalable and Adaptive Application Support Platform for Large-Scale Distributed E-Sciences in High-Performance Network Environments, DOE Office of Science, High-Performance Networks for Distributed Petascale Science.
- Google Research Award, Understanding and Processing Subjective Queries on Structured Data
- NSF: CAREER: Analyzing and Exploiting Meta-information for Keyword Search on Semi-structured Data.
- EarthCube IA: Magnetosphere-Ionosphere-Atmosphere Coupling, Abstract #1541009.
- Intelligent Transportation Systems Resource Center - Task: Data Acquisition, Integration, Analysis, and Visualization.

Application 1: Transportation

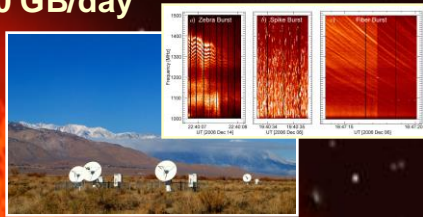


Big Data Challenges:

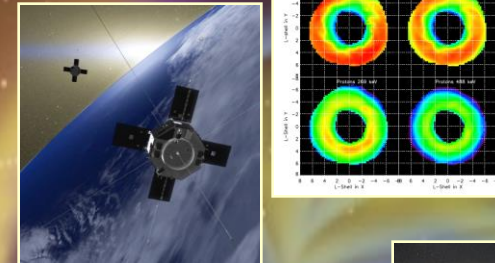
- Standardization of data format
- Accurate modeling
- Clustering and classifying
- Integrating data from independent sources
- Uncovering patterns, correlation, etc.
- Interpretation

Application 2: Solar Terrestrial Research

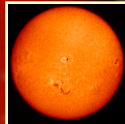
OVSA: 50 GB/day



Van Allen Probes:
2GB/day



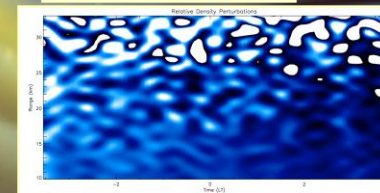
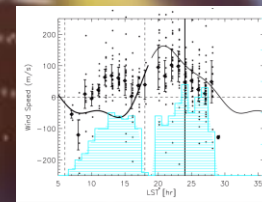
SWRL: 10 GB/day



BBSO: 6000 GB/day



Jeffer Lidar



Other: 0.25 GB/day

PEDC/Antarctic: 0.5 GB/day

Big Data Challenges:

- Complex Process: Plasma Physics + Fluid Dynamics
- Expensive Equipment: Remote Sensing/Instrumentation
- Data Reduction and Inversion
- Modeling and Prediction (sunspot cycle, solar flare)

Application 3: Brain Injury Research

Ballistic (bullet) Blunt Injury-most prevalent Blast (military)

Blunt Impacts >> MVA, fall,

- **Ballistics (Bullet, shrapnel)**
- **Blunt (motor vehicle, sports, fall from height)**
- **Blast (explosions)**

Primary-shock blast wave (Cannon mechanisms explored)

- translational and rotational head acceleration
- thoracic mechanism
- blast wave transmission through cranium
- skull flexure
- Cavitation

Secondary injury-Shrapnel impact

- Produced by debris and high velocity casing fragments

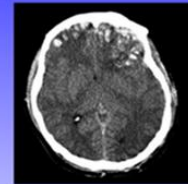
Shrapnel in occipital lobe



Tertiary injury-blunt impact

- Injuries due to impact with other objects.
- causes contusion, intracranial hematoma, cerebral contusion

Cerebral contusion

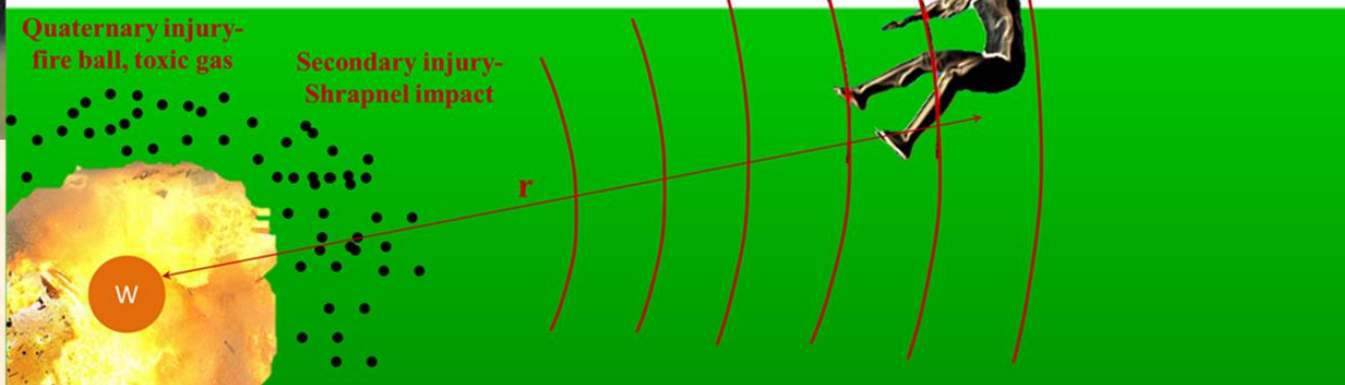


Primary injury-shock blast wave (theme of this chapter)

Tertiary injury-blunt impact

Quaternary injury-fire ball, toxic gas

Secondary injury-Shrapnel impact



Exascale Computing and Big Data

By Daniel A. Reed and Jack Dongarra

Communications of the ACM, July 2015

<https://vimeo.com/129742718>



Thanks! ☺

Questions ?