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/Spring2024/CS444BigData/CS444_BigData_Fall24.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

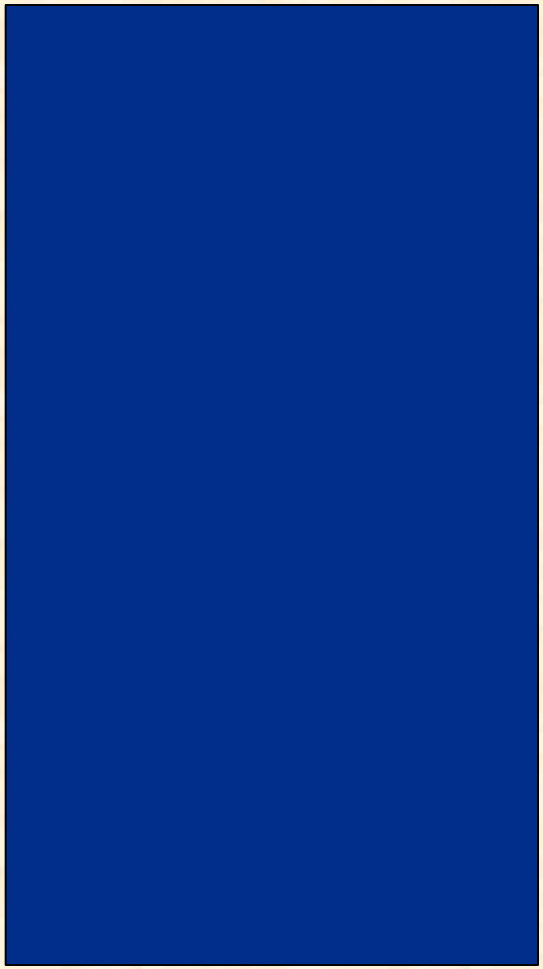
Two purposes:

- 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**
  - **storage format**
  - **processing/analysis purposes**
  - **etc.**
- **How many of the following buzzwords have you heard of?**

• Hadoop	• YARN
• MapReduce	• Containerization
• Spark	• MPI
• NoSQL	• Mahout
• HDFS	• Naïve Bayes
• Pig	• ChatGPT

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**

State of the arts about big data:

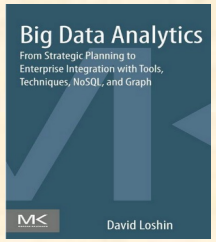
- **Networking: 100's Gbps (backbone)**
- **Storage: PB/EB**
- **Computing: EFlop/s first EVER!**

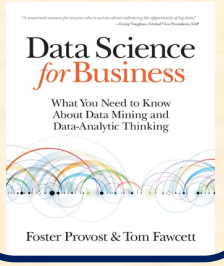
June 2024

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,206.00	1,714.81	22,786
2	**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
3	**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	
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107

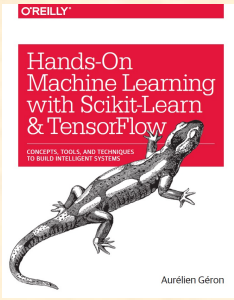
Textbooks and Reference Books

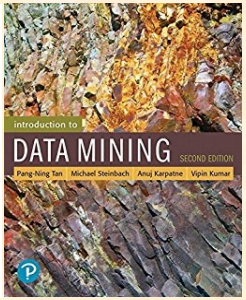
Overview

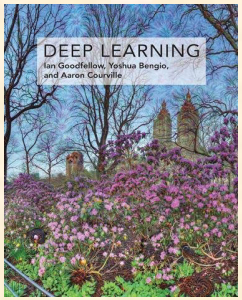


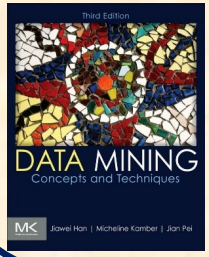


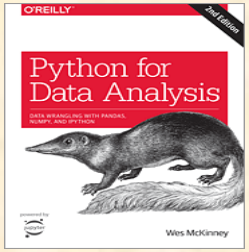
Machine Learning / Data Mining

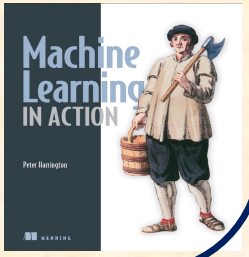




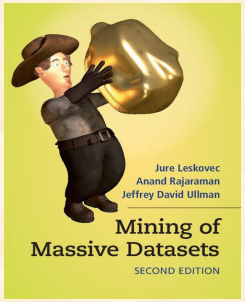


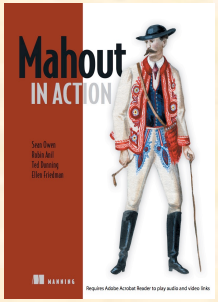


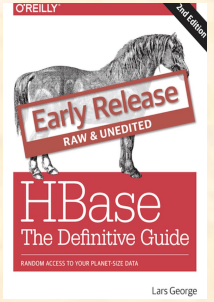


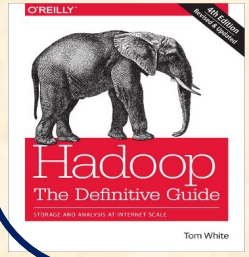


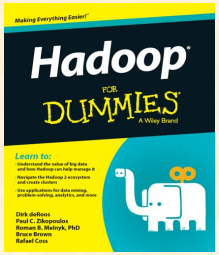
MapReduce / Hadoop

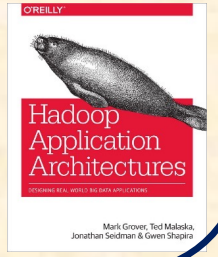












Popular Frameworks

TensorFlow

Caffe2

scikit learn

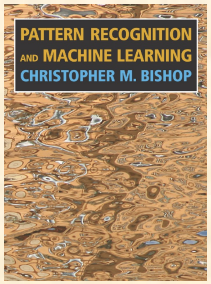
PYTORCH

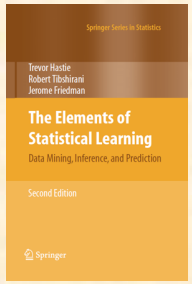
Keras

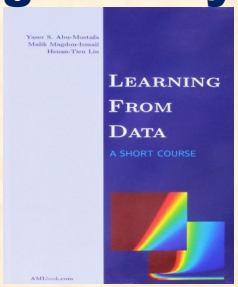
pandas

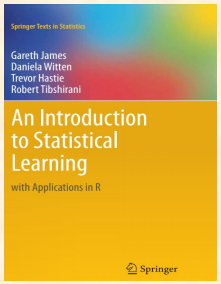
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Learning Theory

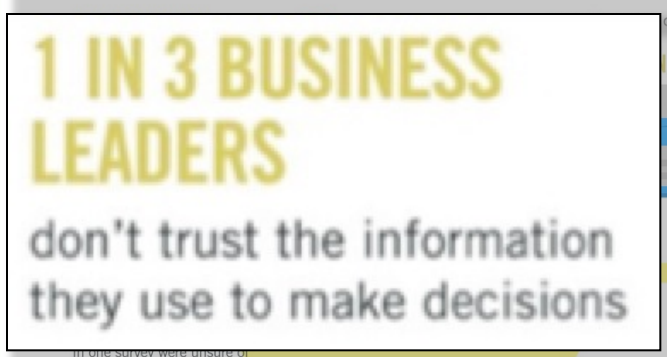








Four V's of Big Data

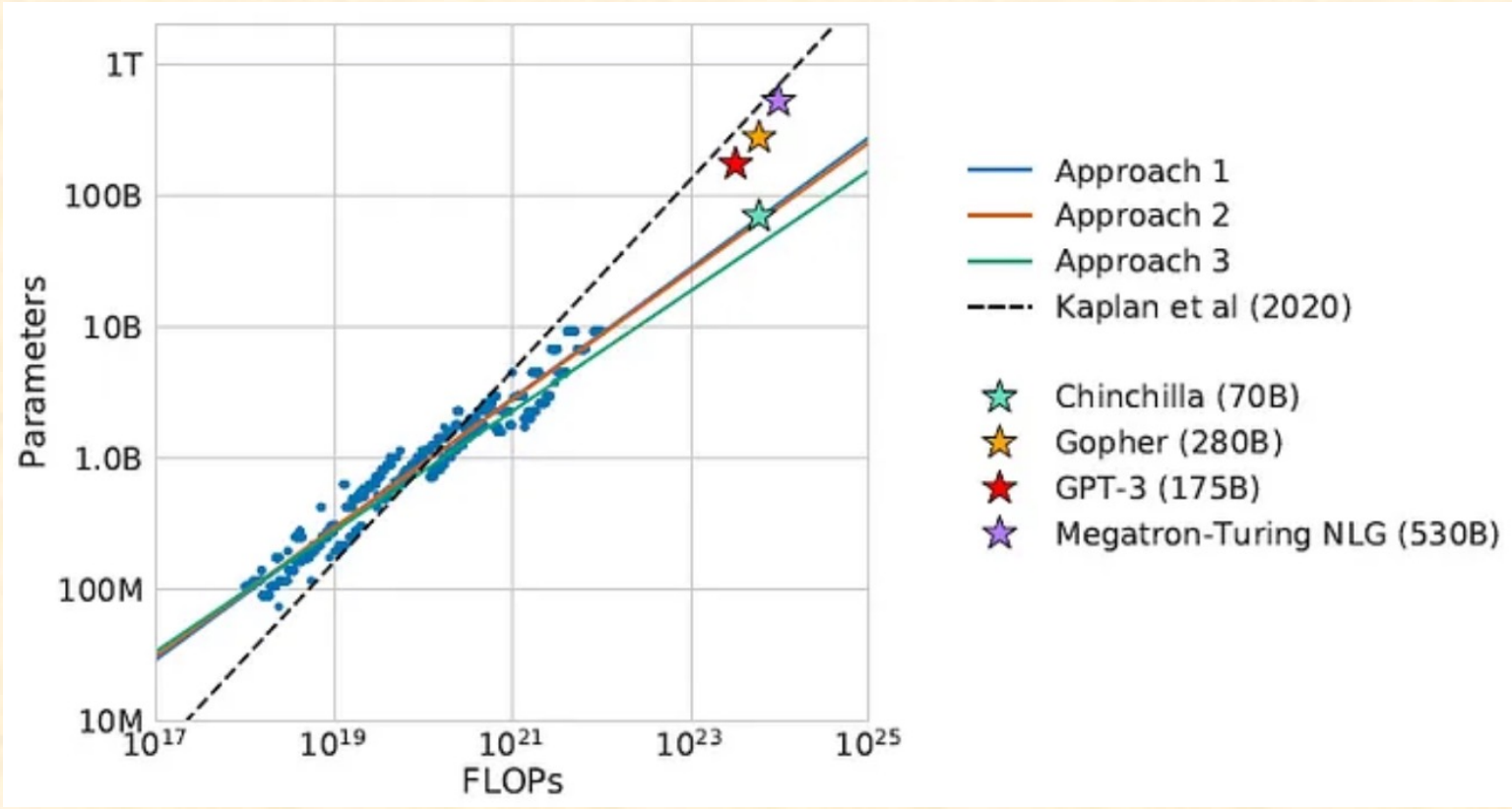


Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

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
Chinchilla	70 Billion	1.4 Trillion



Center for Big Data

Director: Chase Wu

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

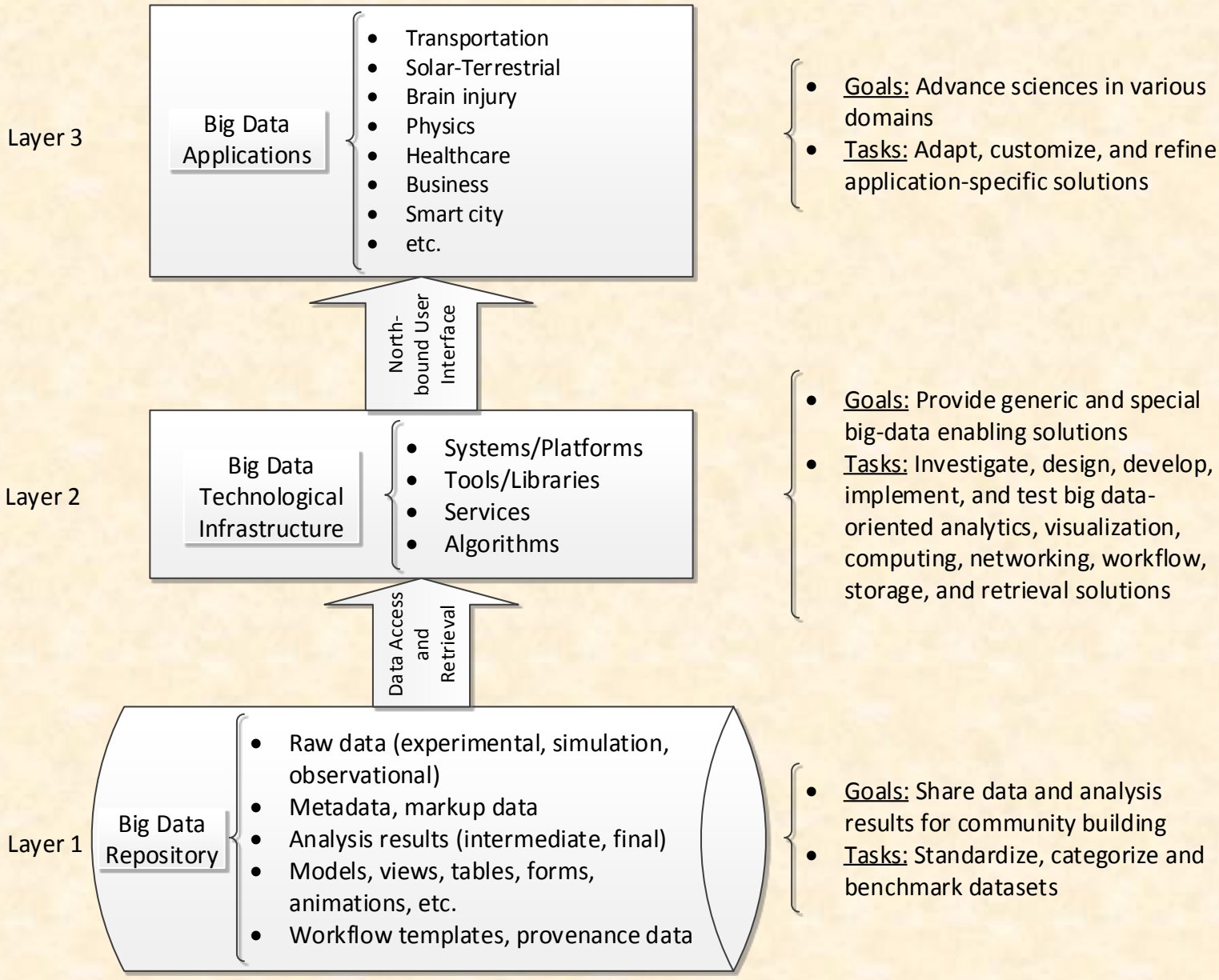
Location: GITC 4416

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.**

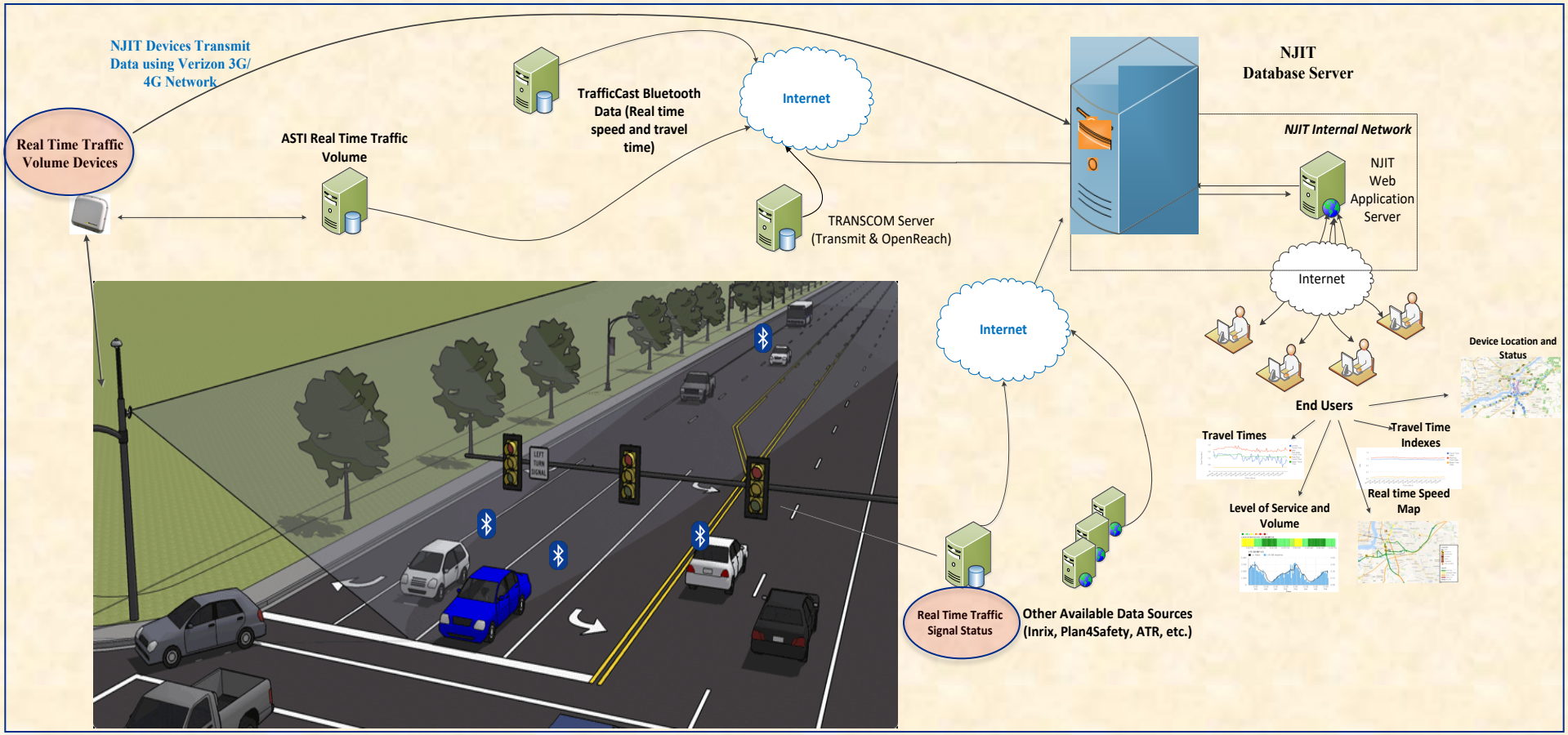
Core Faculty of CBD

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.

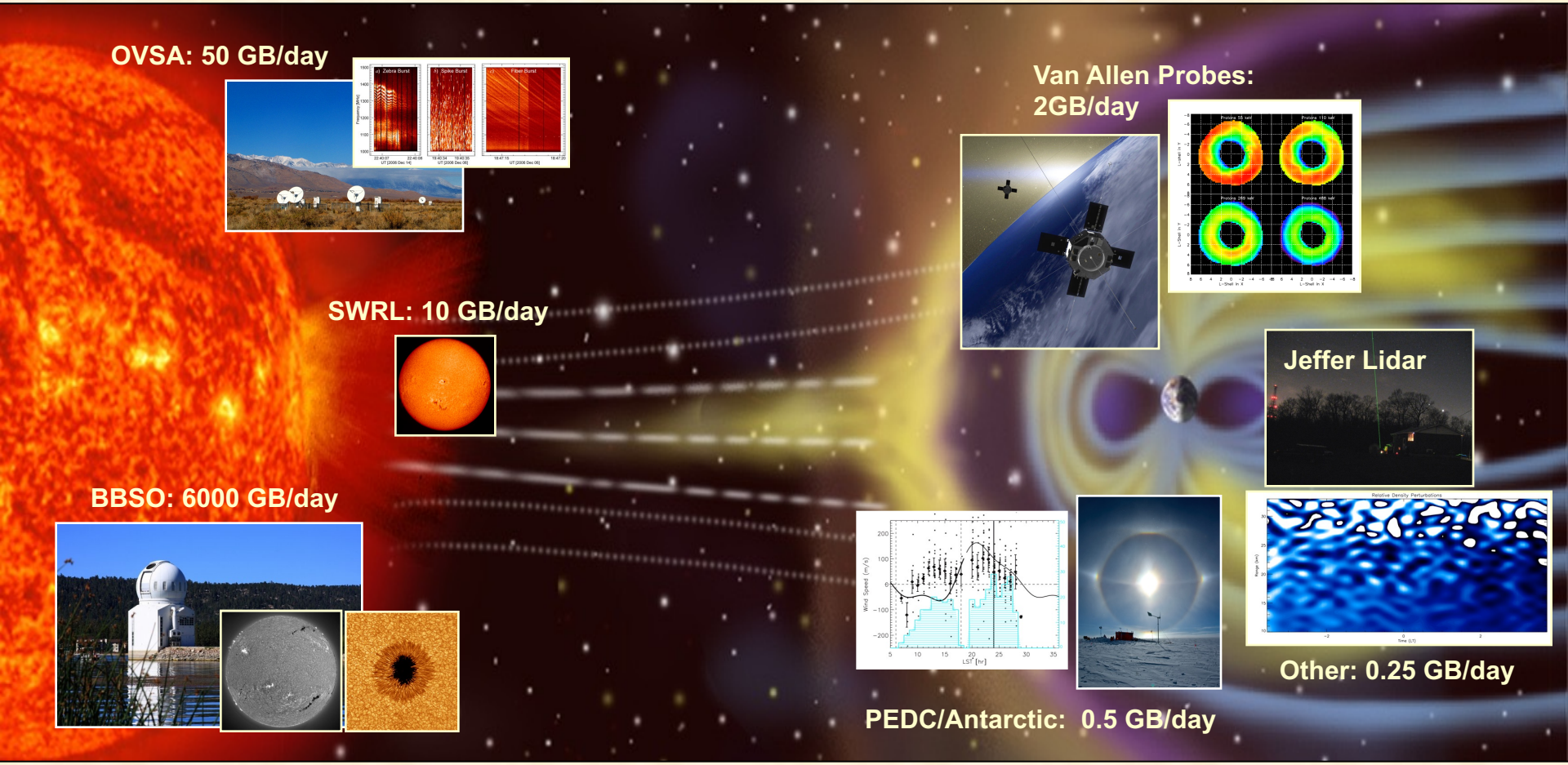
Application 1: Transportation



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



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

Blunt Impacts>> MVA, fall,

Blast (military)

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

Exascale Computing and Big Data

By Daniel A. Reed and Jack Dongarra

Communications of the ACM, July 2015

https://vimeo.com/129742718

Thanks! ☺

Questions ?