CS 644: Introduction to Big Data

Daqing Yun
New Jersey Institute of Technology

https://goo.gl/Awz54i

Who, where, and when

- Daqing Yun, dy83@njit.edu
- CKB 313
- R, 2:30 – 5:20 PM EST
- See Moodle page for more detailed syllabus

The 1st Class Attendance Check

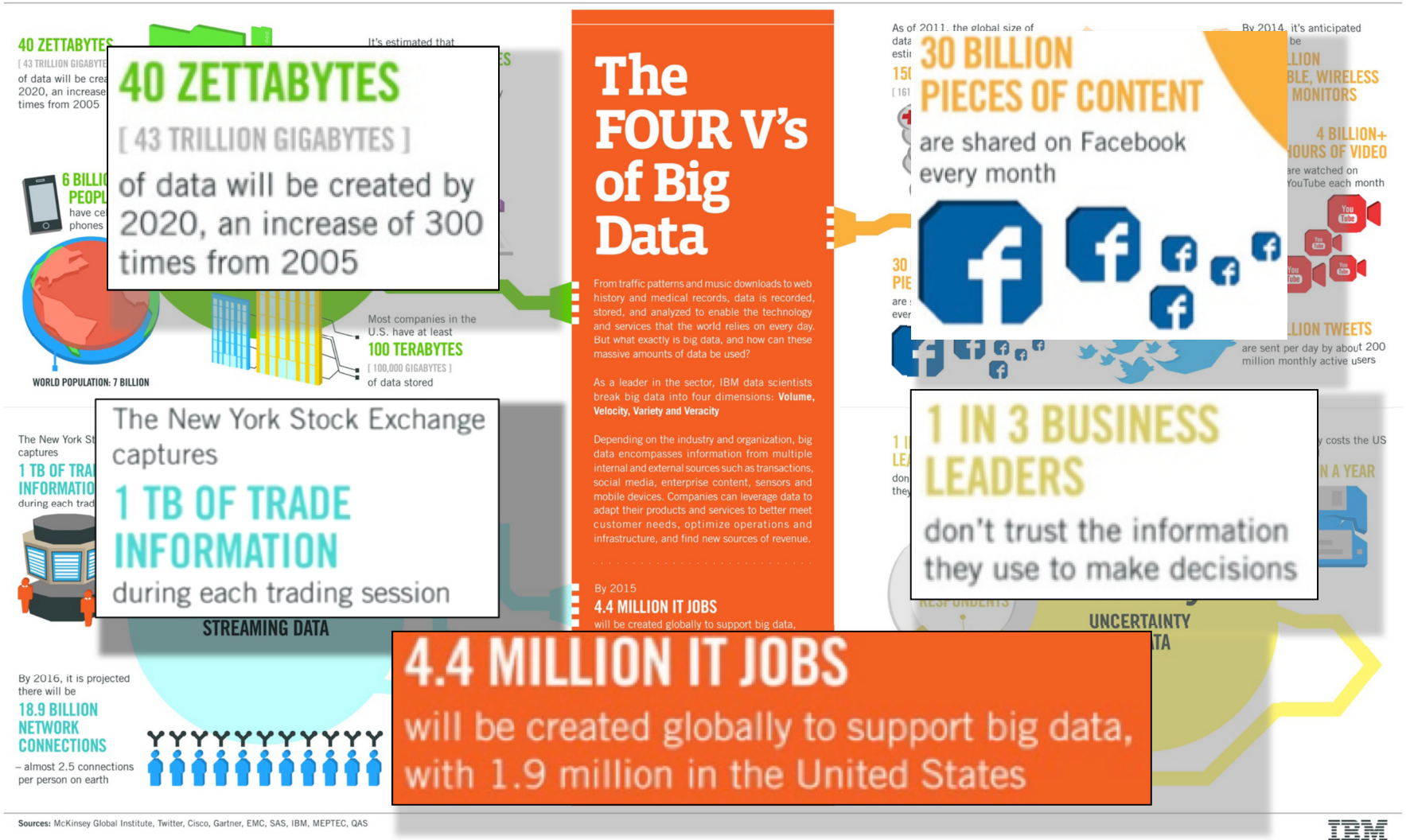
- Name
- Program (MS, Ph.D., etc.)
- Year
- Why do you take this course?
- What is the largest data size you've ever personally handled and in what context (application domain, data type, storage format, etc.)?

About this course

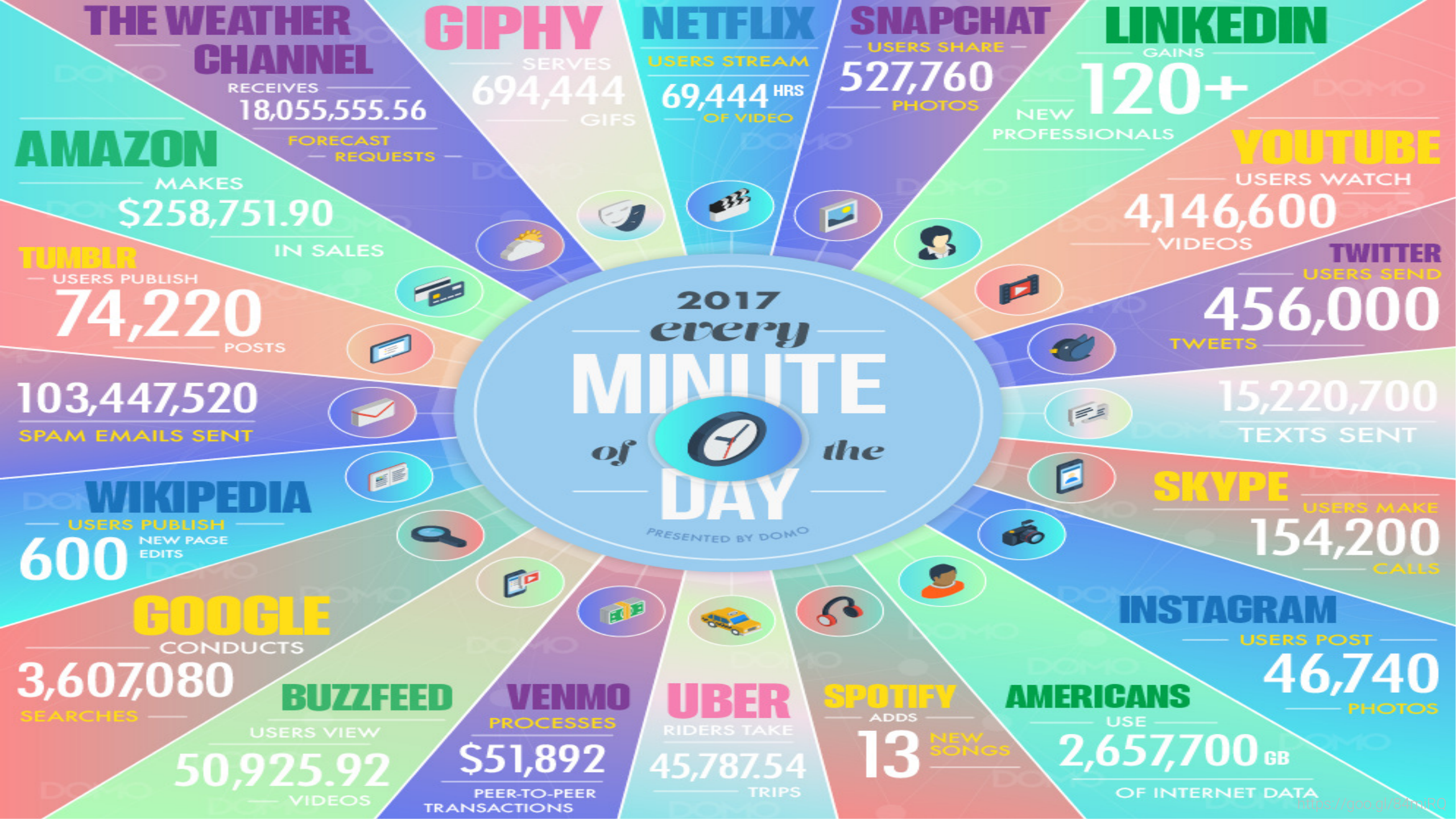
- Recent Developments and Future Trends on Big Data Computing
- Overview of Big Data Analytics
- Platforms, Storages, Databases, and Algorithms for Big Data Analytics
- Advanced Topics:
  - Big-Data Visualization
  - Big-Data Movement
  - Big-Data Security
  - Big-Data Workflows

Characterization of big data

- Four V's



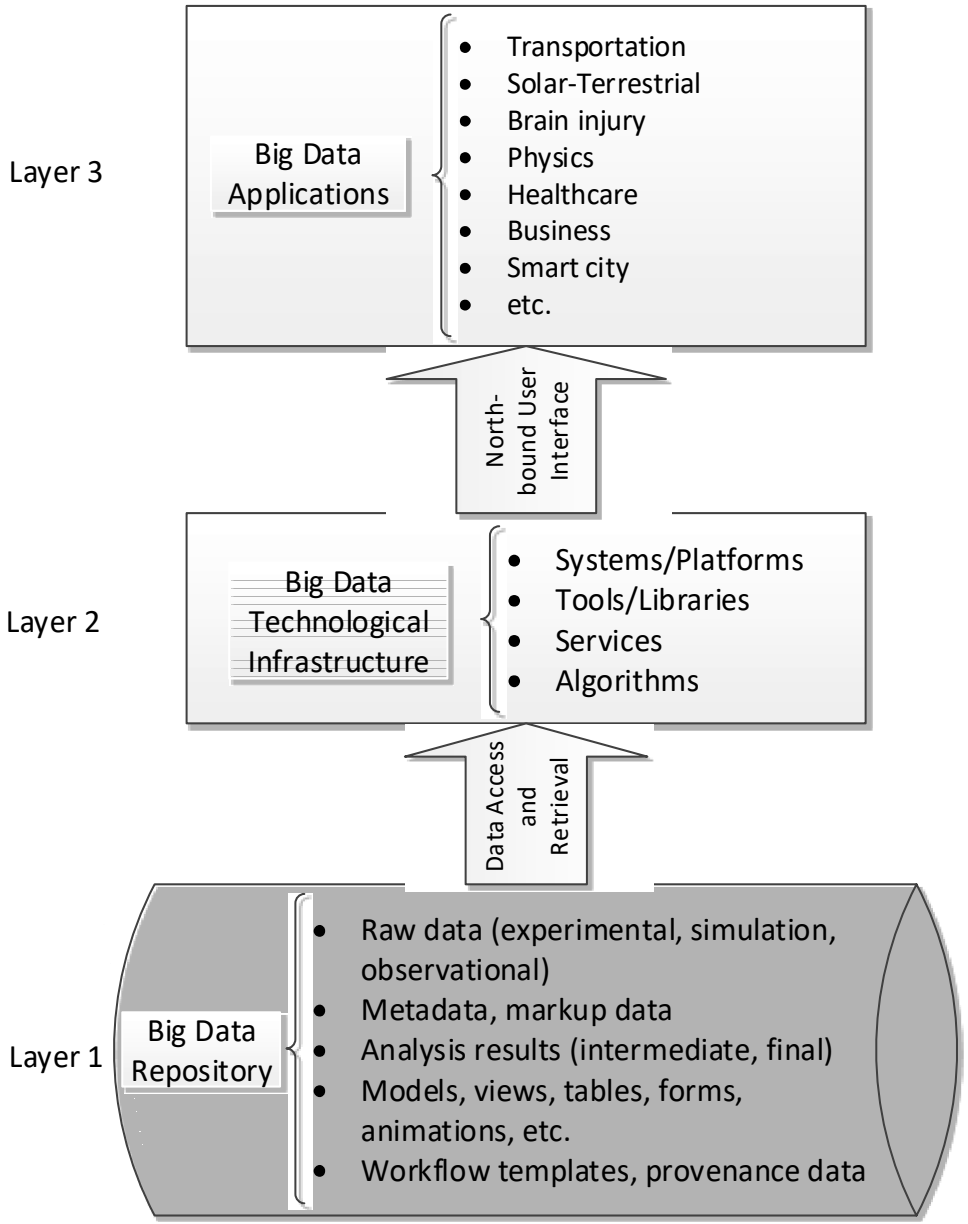
https://www.ibmbigdatahub.com/infographic/four-vs-big-data


2017
every
MINUTE
of the
DAY
PRESENTED BY DOMO
THE WEATHER CHANNEL
RECEIVES
18,055,555.56
FORECAST REQUESTS
GIPHY
SERVES
694,444
GIFS
NETFLIX
USERS STREAM
69,444 HRS
OF VIDEO
SNAPCHAT
USERS SHARE
527,760
PHOTOS
LINKEDIN
GAINS
NEW 120+
PROFESSIONALS
AMAZON
MAKES
$258,751.90
IN SALES
YOUTUBE
USERS WATCH
4,146,600
VIDEOS
TUMBLR
USERS PUBLISH
74,220
POSTS
TWITTER
USERS SEND
456,000
TWEETS
103,447,520
SPAM EMAILS SENT
15,220,700
TEXTS SENT
WIKIPEDIA
USERS PUBLISH
600
NEW PAGE EDITS
SKYPE
USERS MAKE
154,200
CALLS
GOOGLE
CONDUCTS
3,607,080
SEARCHES
INSTAGRAM
USERS POST
46,740
PHOTOS
BUZZFEED
USERS VIEW
50,925.92
VIDEOS
VENMO
PROCESSES
$51,892
PEER-TO-PEER TRANSACTIONS
UBER
RIDERS TAKE
45,787.54
TRIPS
SPOTIFY
ADDS
13
NEW SONGS
AMERICANS
USE
2,657,700 GB
OF INTERNET DATA

Center for Big Data

- Director: Chase Wu, chase.wu@njit.edu
- https://centers.njit.edu/bigdata
- Location: GITC 4111
- Mission Statement:
  - Synergize the strong expertise in various disciplines across the NJIT campus
  - Build a unified 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



- Goals: Advance sciences in various domains
- Tasks: Adapt, customize, and refine application-specific solutions

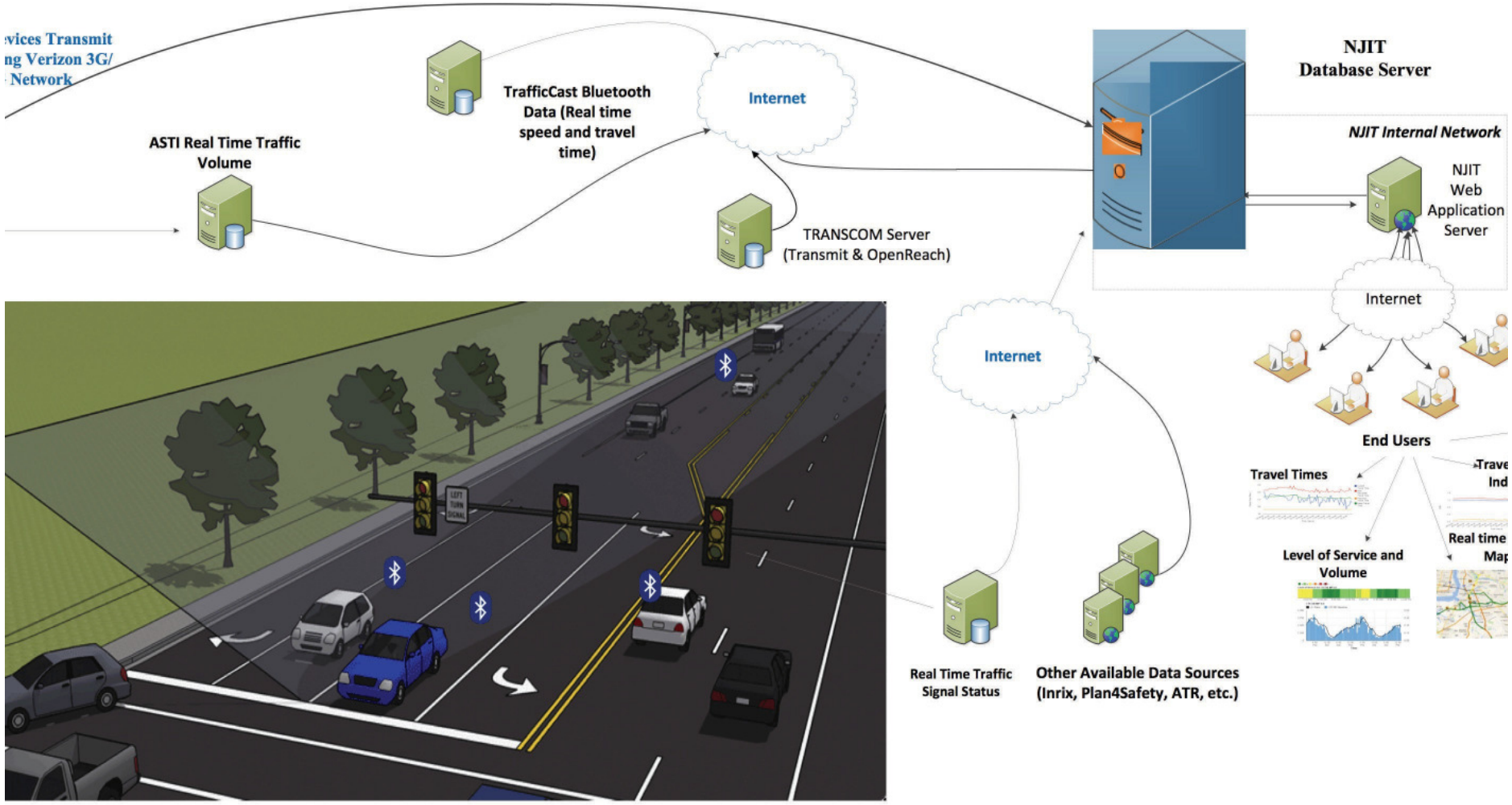
- Goals: Provide generic and special big-data enabling solutions
- Tasks: Investigate, design, develop, implement, and test big data-oriented analytics, visualization, computing, networking, workflow, storage, and retrieval solutions

- Goals: Share data and analysis results for community building
- Tasks: Standardize, categorize and benchmark datasets

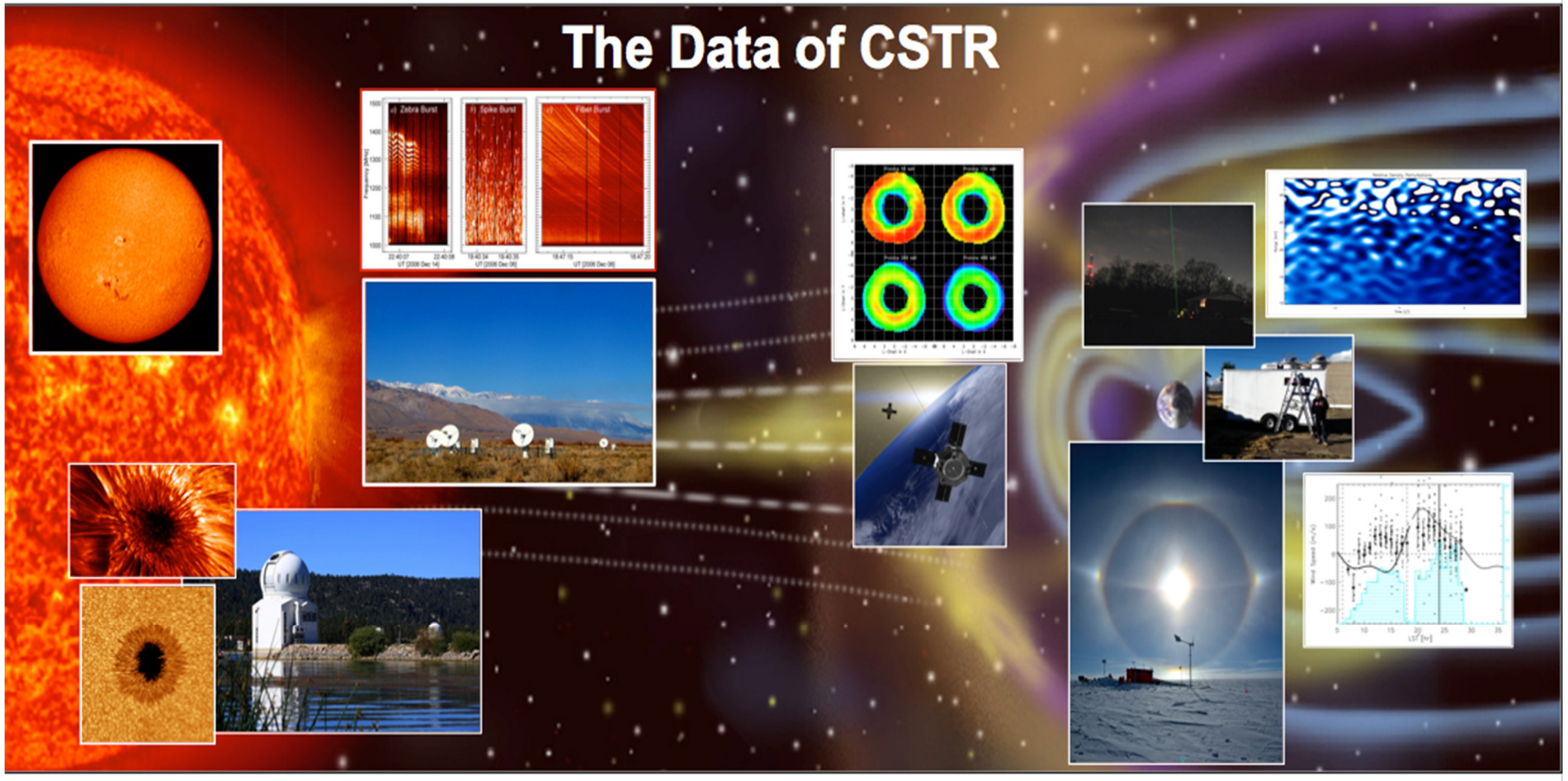
A Three-layer Structure of the CBD

- **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.

Transportation



Solar Terrestrial Research



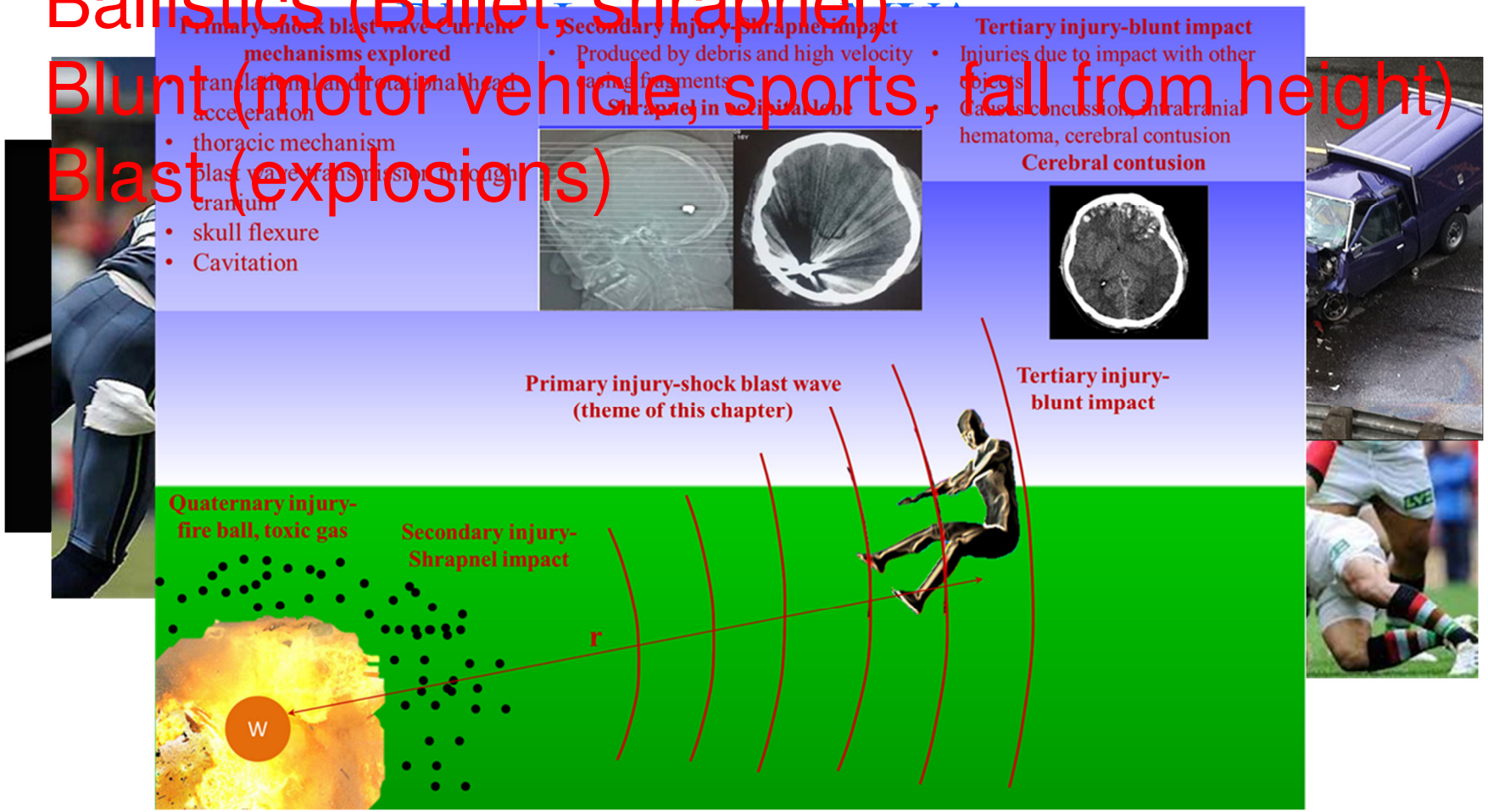
Classification of Traumatic Brain Injury

Ballistic (bullet)

Blunt Injury-most prevalent

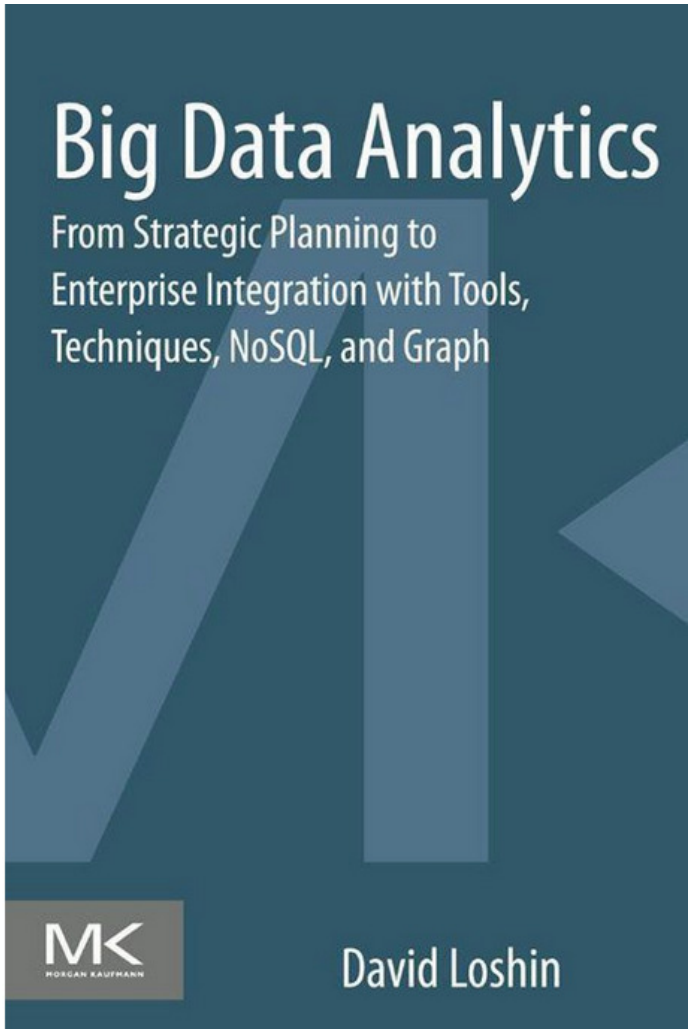
Blast (military)

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



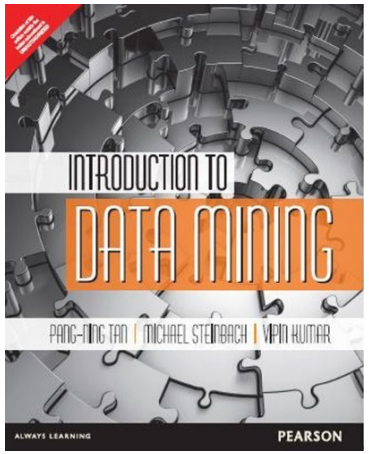
Reading materials

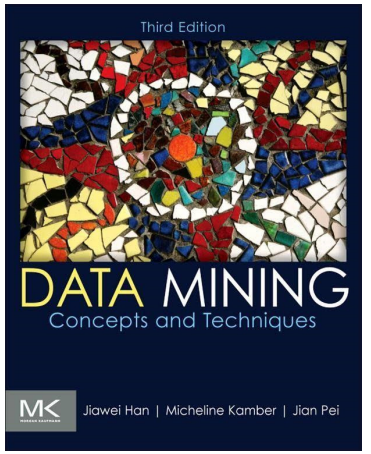
- Chapter 1: Market and Business Drivers for Big Data Analysis
- Chapter 2: Business Problems Suited to Big Data Analytics
- Chapter 3: Achieving Organizational Alignment for Big Data Analytics
- Chapter 4: Developing a Strategy for Integrating Big Data Analytics into the Enterprise
- Chapter 5: Data Governance for Big Data Analytics: Considerations for Data Policies and Processes
- Chapter 6: Introduction to High-Performance Appliances for Big Data Management
- Chapter 7: Big Data Tools and Techniques
- Chapter 8: Developing Big Data Applications
- Chapter 9: NoSQL Data Management for Big Data
- Chapter 10: Using Graph Analytics for Big Data
- Chapter 11: Developing the Big Data Roadmap

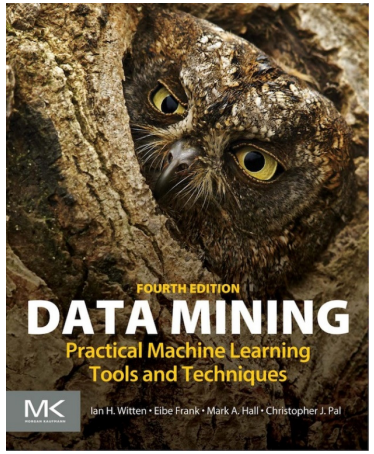


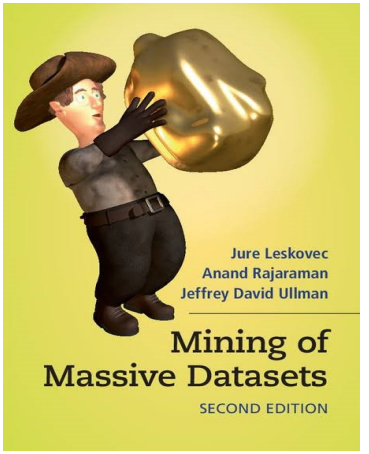
Reading materials

- Introduction to Data Mining, by P. Tan, M. Steinbach, and V. Kumar, ISBN: 978-93-325-1865-0
- Data Mining: Concept and Techniques (3rd Ed.), by J. Han, M. Kamber, and J. Pei, ISBN-13: 978-9380931913
- Data Mining: Practical Machine Learning Tools and Techniques (4th Ed.), by I. Witten, E. Frank, M. Hall, and C. Pal, ISBN: 978-0-12-804291-5
- Mining of Massive Datasets, by J. Leskovec, A. Rajaraman, and J. Ullman









If you are really interested...

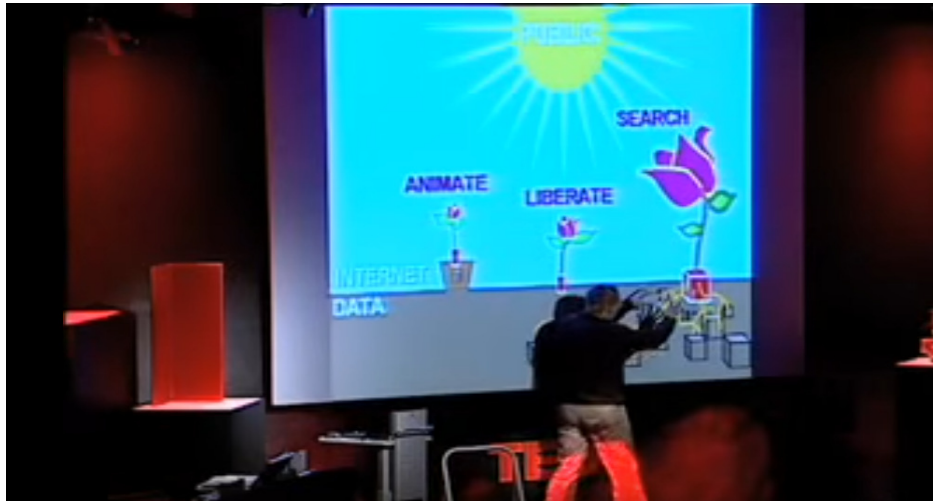
Exascale Computing and Big Data

- By Daniel A. Reed and Jack Dongarra, July 2015, Communications of the ACM
- https://vimeo.com/129742718

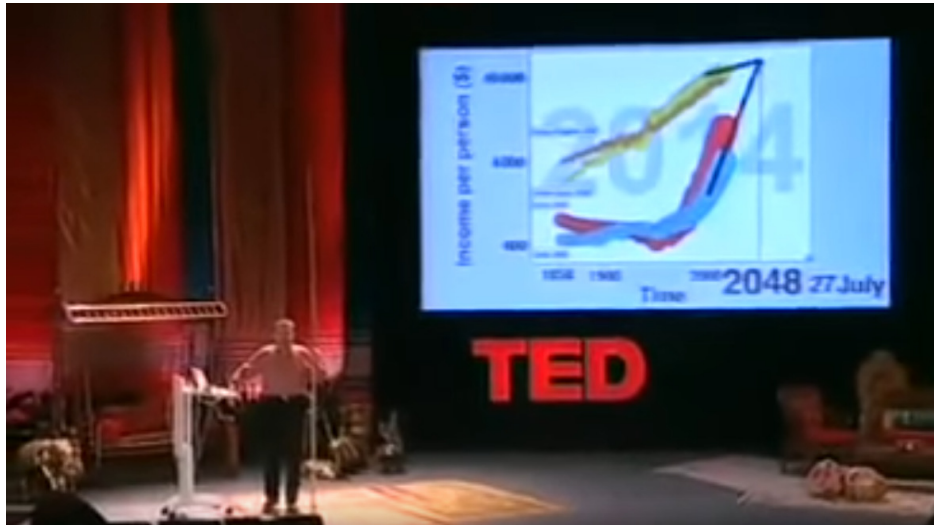
Future of Data Mining

https://www.youtube.com/watch?v=DS310JMdu2s
By Bill Weir, C. Michael Kim, David Miller, Justin Bare & Mark Monroy | This Could Be Big – Tue, Dec 4, 2012

The Magic of Data

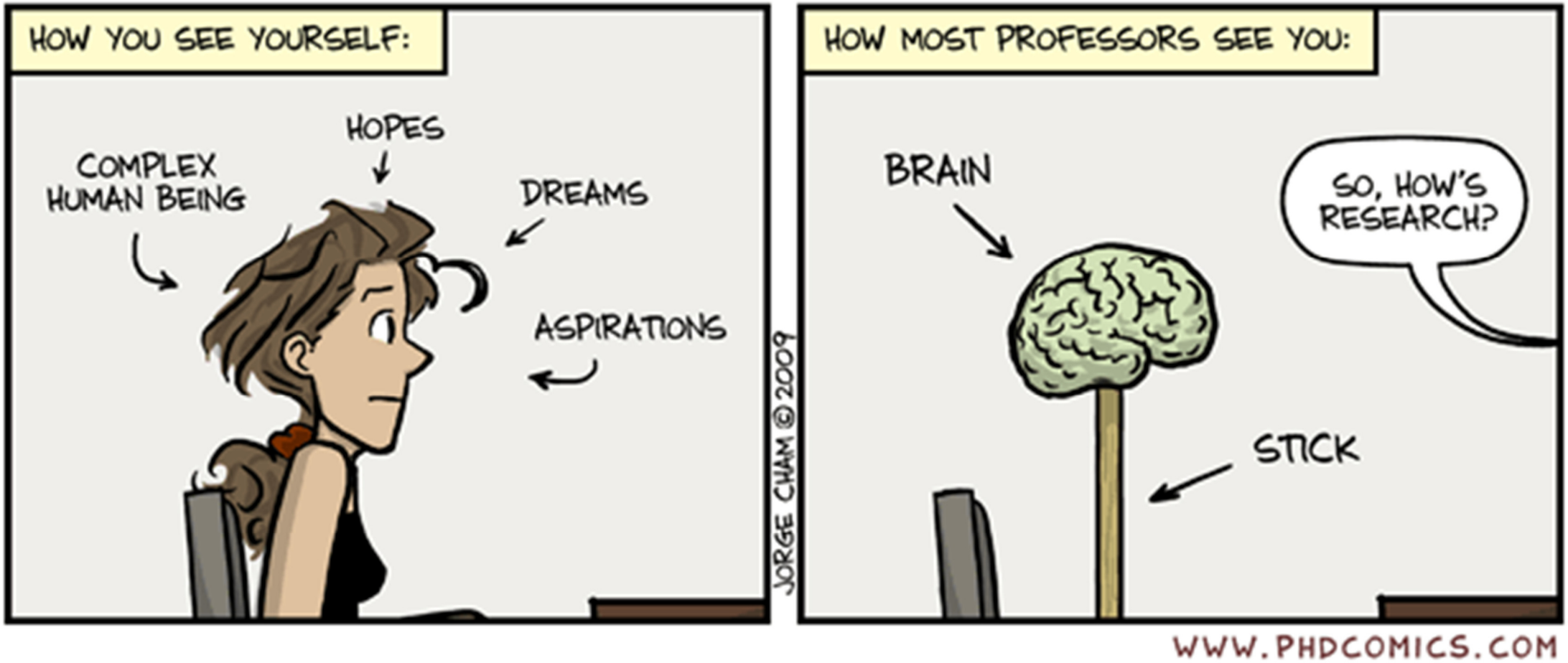


https://www.youtube.com/watch?v=RUwS1uAdUcI
Hans Rosling: Debunking third-world myths with the best stats you've ever seen



https://www.youtube.com/watch?v=fiK5-oAaeUs
Hans Rosling: Asia's rise - how and when

I know you have other things to do, but...



The first rule of CS 644 is…

- *Don't Plagiarize in CS 644*

- The second rule of CS 644 is…
  - *Don't Plagiarize in CS 644*

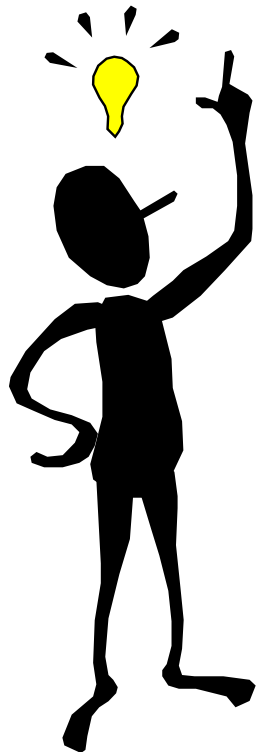
- There are plagiarism-detection systems
  - Work pretty well, seriously
  - Academic integrity tutorial: https://goo.gl/ZX9oBe

- The *penalty* would be: 0.0 in this course

Reminder

- HW 1 is assigned, due by next lecture
- See Moodle for details

Hope you will enjoy…



Thanks ! ☺

Questions ?