2018 Education Summit @ Esri UC

Date: Saturday, July 7, 2018, Time: 10:00 AM - 11:00 AM
Room: Marriott - Coronado

SAN DIEGO STATE UNIVERSITY

Teaching Big Data Analytics and Geospatial Data Science

Dr. Ming-Hsiang Tsou

Email: mtsou@mail.sdsu.edu, Twitter **@mingtsou**
Director of **the Center for Human Dynamics in the Mobile Age**
Professor, Department of Geography , San Diego State University

HDMA @SDSU | THE CENTER FOR HUMAN DYNAMICS IN THE MOBILE AGE

GeoTECH GEOTECHCENTER.ORG | National Geospatial Technology Center of Excellence
Empowering Colleges: GROWING THE WORKFORCE

What is "Data Science"?

Major "Knowledge Domain" in Data Science

(from O'Neil, C., & Schutt, R. (2013). Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, Inc.)

- Computer science
- Mathematics
- Statistics
- Machine learning
- Communication and presentation skills
- Data visualization
- Domain expertise

- We need to add "GIScience and Geospatial Technology" in HERE!!!

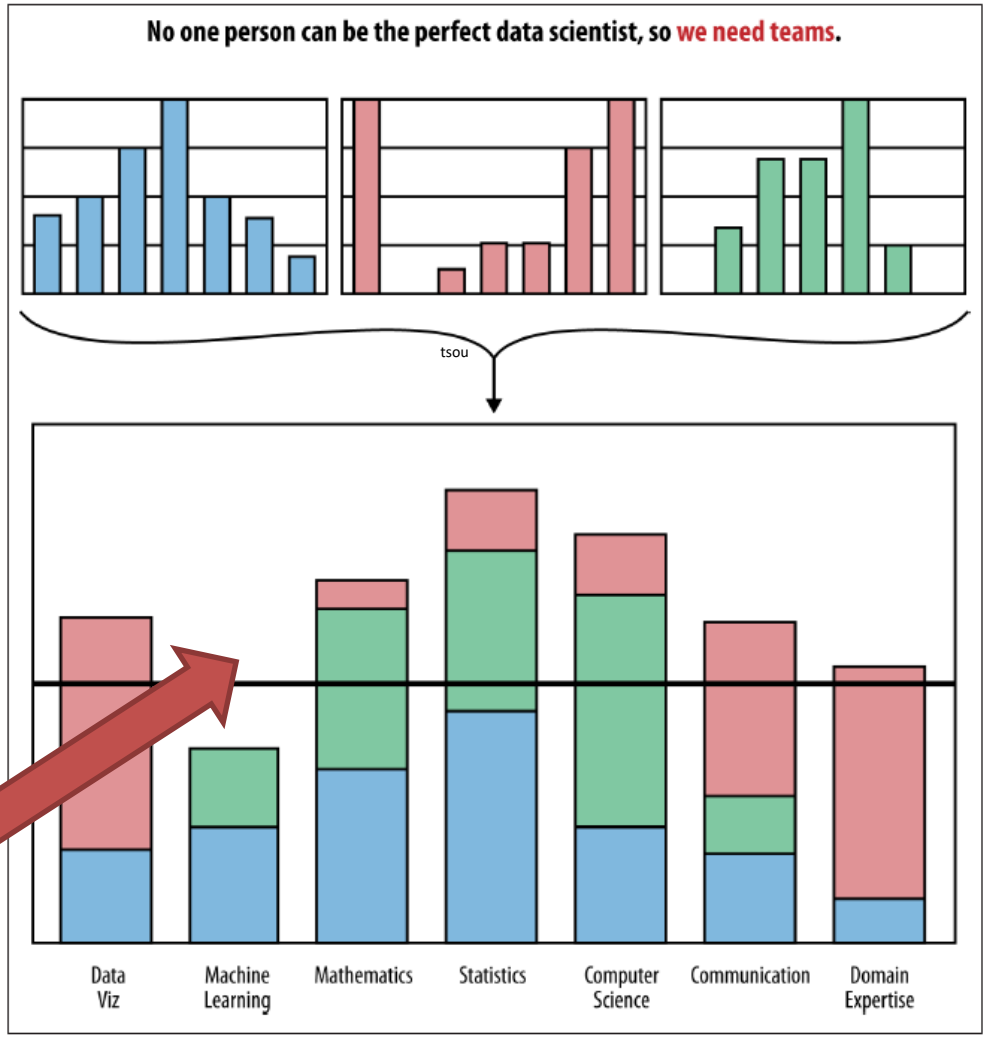
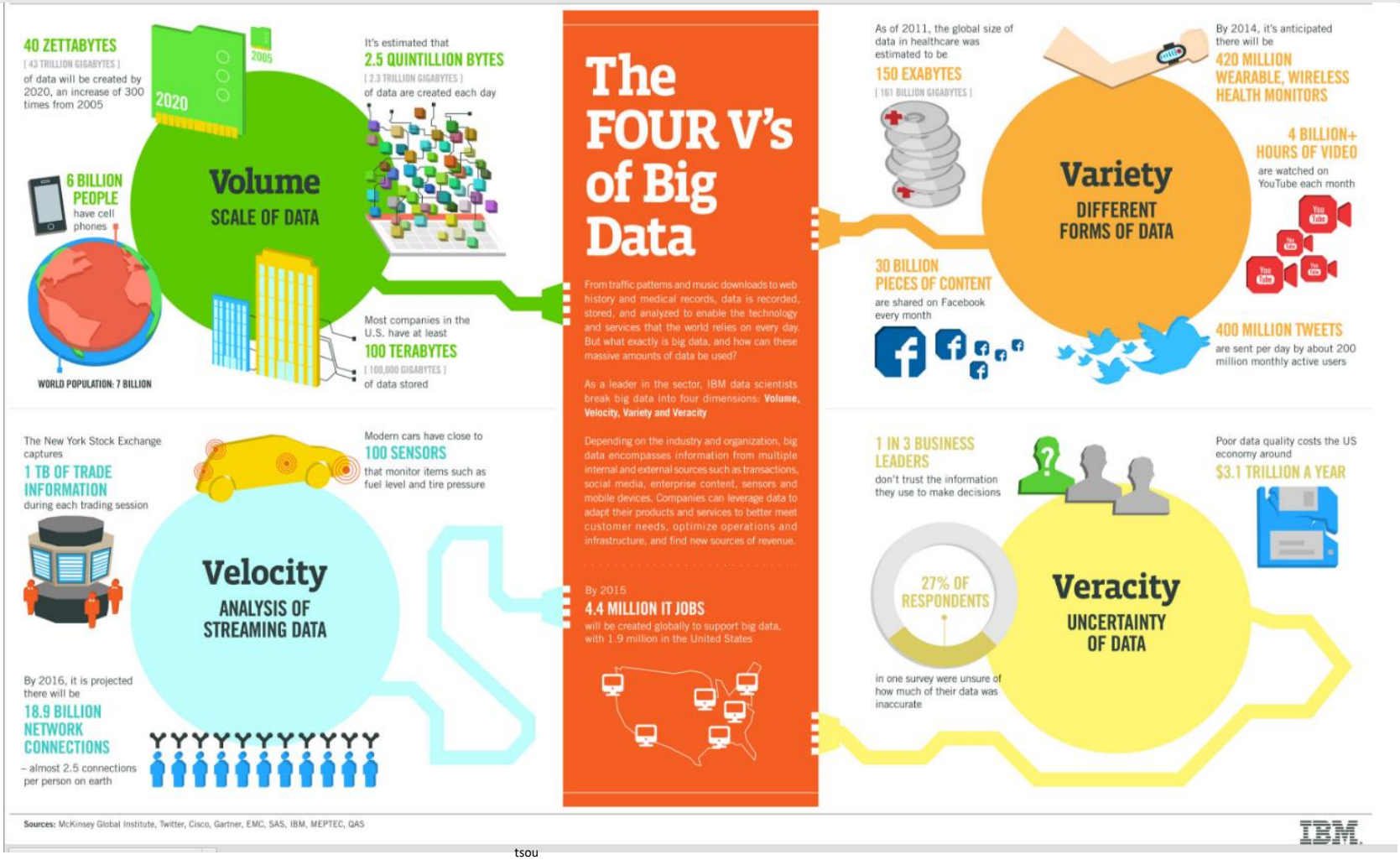


Figure 1-3. Data science team profiles can be constructed from data scientist profiles; there should be alignment between the data science team profile and the profile of the data problems they try to solve

What are "Big Data"?



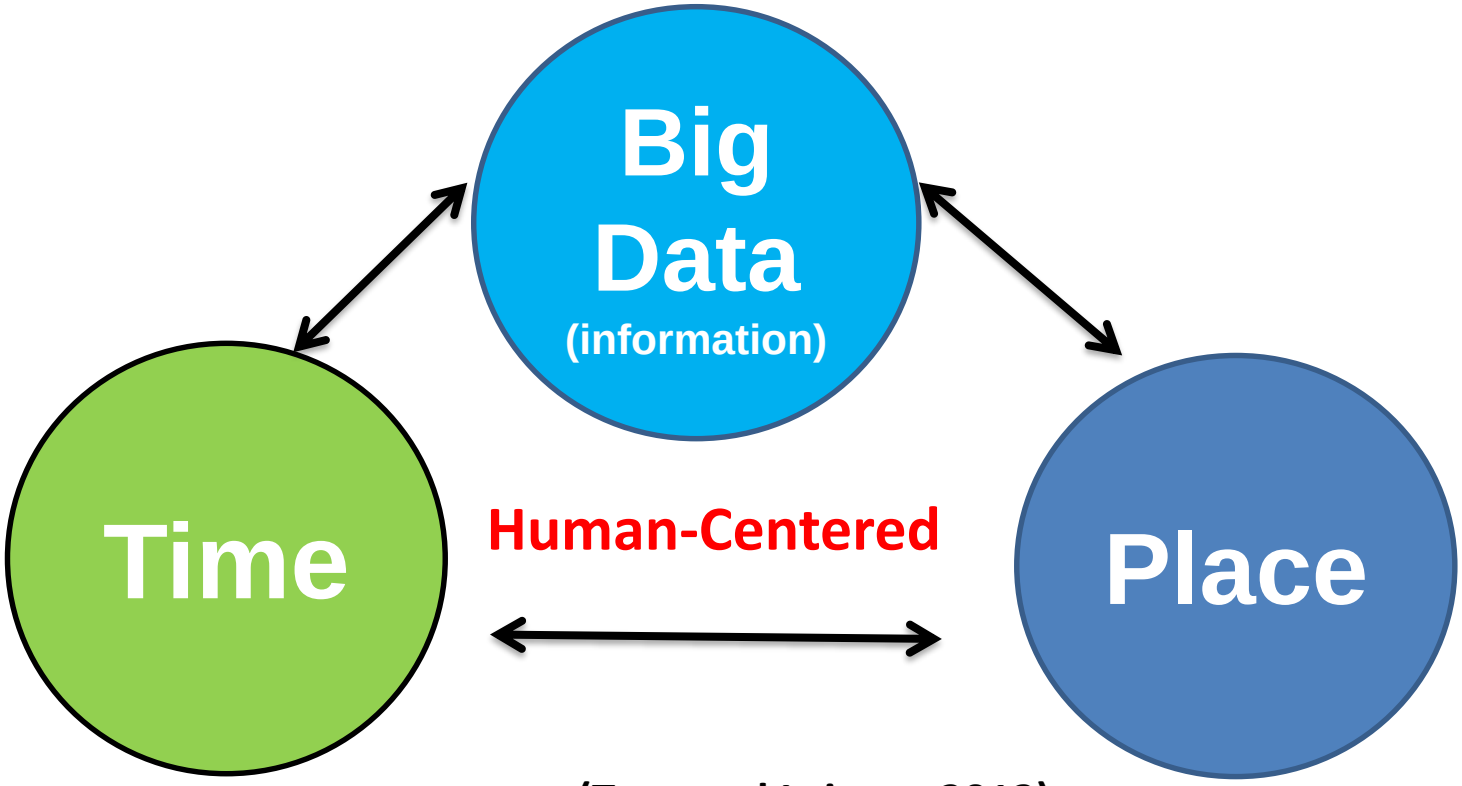
tsou

Veracity (the truth of data) – What is this?

Can Big Data represent 100% of our Real World? NO! N ≠ all

[Image source: http://www.ibmbigdatahub.com/infographic/four-vs-big-data](http://www.ibmbigdatahub.com/infographic/four-vs-big-data)

Geography (place and time) is the KEY for Understanding and Integrating Big Data

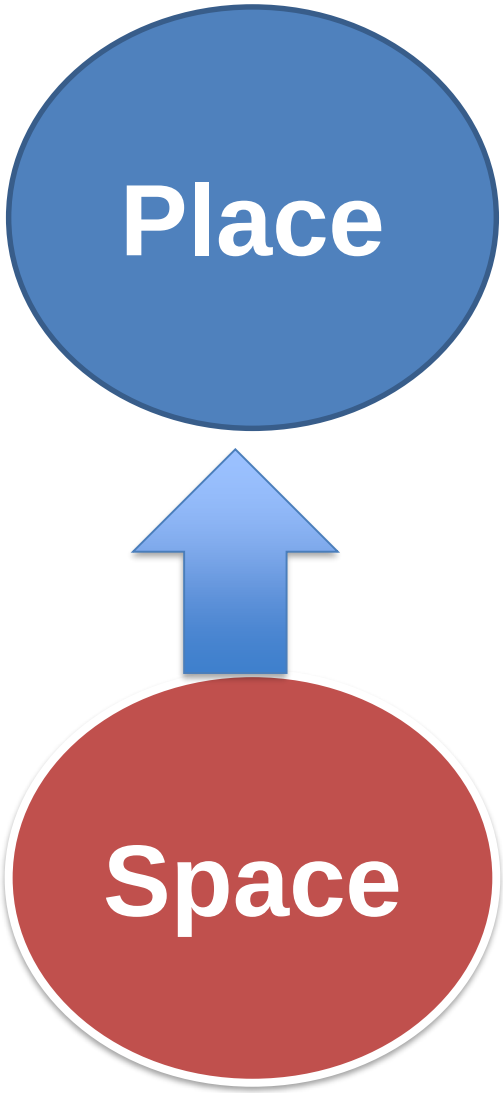


(Tsou and Leitner, 2013)

Knowledge Discovery in Big Data (KDBD) framework

Tsou, M. H. and Leitner, M. (2013). Editorial: Visualization of Social Media: Seeing a Mirage or a Message? In Special Content Issue: "Mapping Cyberspace and Social Media". Cartography and Geographic Information Science. 40(2), pp. 55-60. DOI: 10.1080/15230406.2013.776754

Human-centered Big Data Analytics → "Place" vs. "Space"



= **Personalized locations and dynamic geometry** (**sense of place**), Fuzzy boundary (dynamics), **human-centered** (task-oriented, functional) **London**, my home town, San Diego, UCSD. (**Social Media Content/Conversations**).

= **Basic Geometry** (**point, lines, polygons**) defined by coordinates, precise boundary and locations, **mathematical/ computational, traditional Geographic Information Systems (GIS). San Diego** (lat/long) a point or a polygon (map scale).

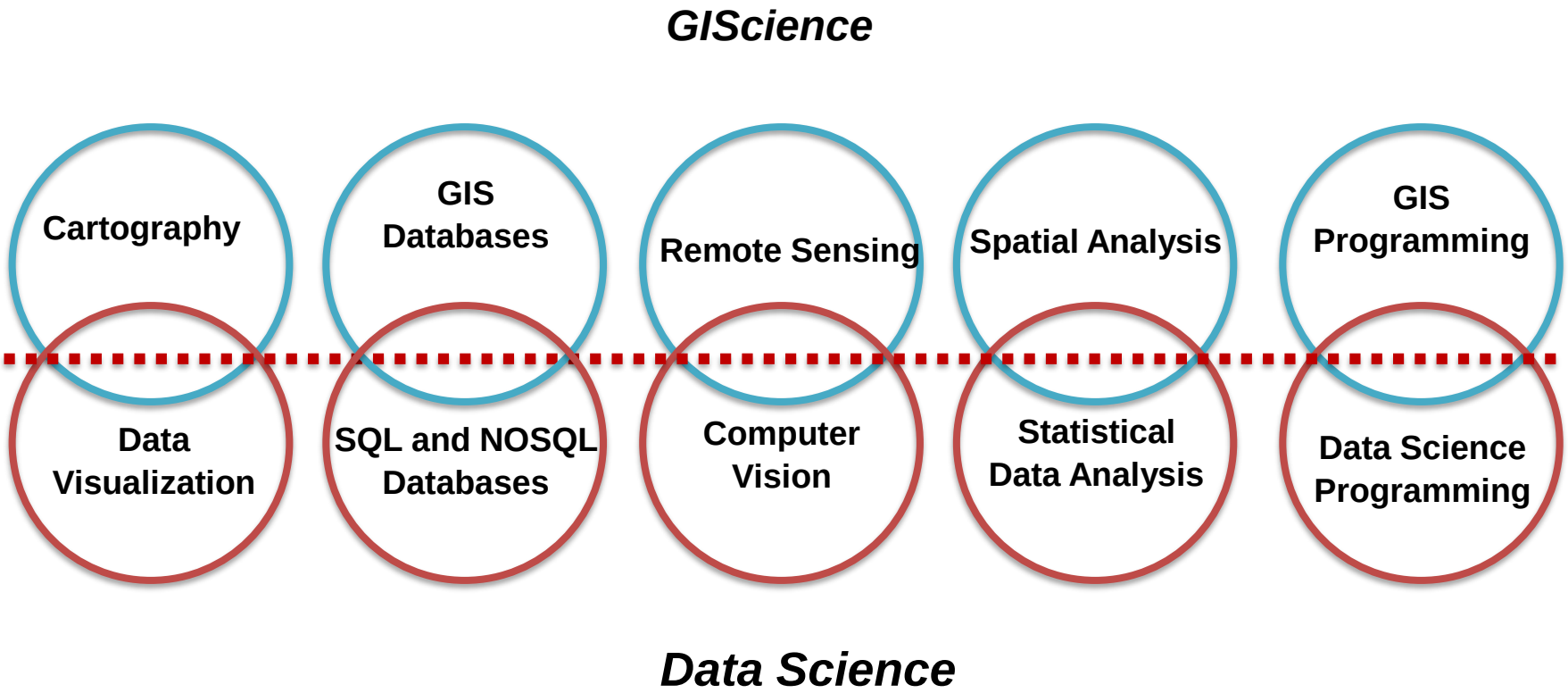
Tuan, Yi-Fu. *Space and place: The perspective of experience*. U of Minnesota Press, 1977.

Great History of Big Data Processing and Analysis in GIS and Geospatial Analysis Applications

- **U.S. Census data** since **1790** – present (every 10 years).
- **Land use** and land cover survey data (since **1930s** by Ludley Stamp, UK).
- **Remote Sensing and Satellite Imagery** Analysis in 1960s (Cold War) and after.
- Environmental **Sensor** data (1970s, Low-Angle Radar Tracking).
- **GPS** data analysis after 2000 (removing the selective availability signal to improve the accuracy in 2000).
- **Human dynamics data** (social media, mobile devices).

The Overlap in Curricula

Comparing GIScience Programs/Curricula and Data Science/Data Analytics Programs/Curricula



GIScience + Data Science = Geospatial Data Sceince

Geospatial Data Science

- What is "Geospatial Data Science"?
- **Geospatial Data Science as a transdisciplinary field to extracting knowledge and insight from geospatial big data using high performance computing (HPC) resources, spatial and non-spatial statistics, spatiotemporal analysis models, GIS algorithms, machine learning methods, and geovisualization (mapping) tools.**
- The derived **geospatial knowledge and insight** can facilitate spatial thinking and problem solving for various applications, and enable the exploration of new scientific theories.
- (Tsou, et. al. Building Geospatial Data Science Education Resources: the Convergence of GIScience, Geospatial Technology, and Data Science, Paper in preparation, 2018).

Uniqueness of Big Data

(comparing to traditional GIS and RS data)

- Most of them are **points** (due to the collection from sensors and mobile devices, smart phones).
- Most of them have **trajectory data** and **time series analysis** (However, traditional GIS software are lack of spatiotemporal analysis function).
- **Unstructured** data (No-SQL databases, social media data) (traditional GIS data are "relational databases" and "well-structured").
- Multi-level and **dynamic scaling** (how to aggregate point data into meaningful scale level? (census block, zip codes, county, city boundary?) (traditional GIS data are at single scale)

The differences

GIScience

- **Map projection and coordinate systems**
- Remote Sensing Sensors and platforms
- Spatial Analysis (**Buffer, Overlay**, GWR)
- **GIS Software** (ArcGIS, QGIS, OpenLayers).
- Web Map Servers (ArcGIS online)

- Maps and **Visualization**
- Database Management
- **Image Analysis, Identification**, and Recognition
- **Statistical methods** (clustering, classification, hotspot analysis).
- Programming (**Pythons, R**, JavaScripts)
- Web Applications (Mapping Service APIs and Data APIs)

Data Science

- **Text mining** and linguistic analysis (topic modeling, latent Dirichlet allocation (LDA)).
- **Social network Analysis**
- Cloud Platforms (EC2) and HPC (Hadoop and Spark).
- NoSQL databases (MongoDB)
- **Machine learning** (Supervised machine learning vs. Unsupervised machine learning).

Different content ***Same content*** ***Different content***

Improve current GIScience Courses with Data Science and Data Analytics Methods and Tools.

- **Machine learning** and **time series analysis** in **Spatial Analysis courses**
- **R and Pythons with Data analytic libraries** in **GIS programming** courses.
- **Tableau and other Business Intelligent (BI) Software** in **Cartography** courses.
- **NoSQL databases** (MongoDB) in **GIS database** courses
- **Text mining methods** and social network analysis in **GIS application** courses
- **Critical thinking and data privacy issues** in **GIS Design** courses.
- **Mapping APIs (leaflet, MapBox, CartoDB)** and Data APIs (social media) in **Web GIS courses**.

Amazon Educate Starter Account

AWS Account: $100 in credits at member institutions; $40 in credits at non-member institutions. (Need credit card info to create an account…).
AWS Educate Starter Account: $75 in credits at member institutions; $30 in credits at non-member institutions **(Students don't need to provide their credit card info!!!)**

Ask your University's IT staff to apply for "member institution" to get benefit for your students….

When students register the Educate Starter Account using the **University's email address**, they will automatically get the benefit.

Web Exercise 04: Working with Amazon AWS Educate Starter

Due Date: **October 05** (Thursday), 2017. 2:00pm on Blackboard. (Two weeks)

Grade: **15 points.**

1. Create Your AWS Free Account and an EC2 Instance.

Go to the following URL to establish your Amazon AWS Educate Starter account:

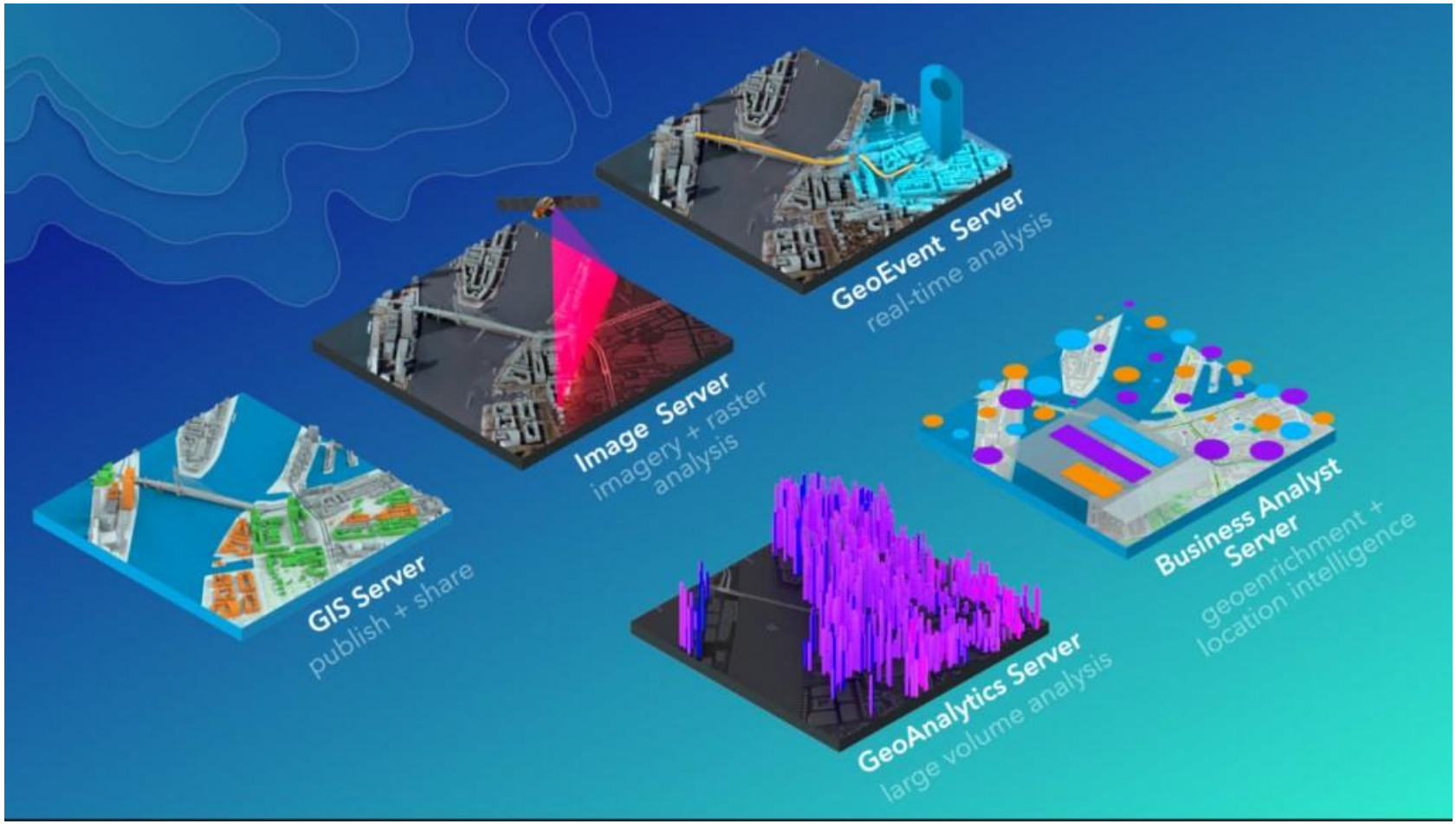
https://aws.amazon.com/education/awseducate/apply/

We will be creating an Amazon Web Services (**AWS**) account associated with San Diego State University, so click the "**Students: [Apply for AWS Educate for Students]**" button (the right one):

Future Plan:

Adopting ArcGIS Enterprise for Big Data Analytics

GeoEvent Server (real time data + IOT), **GeoAnalytics Server** (Big Data Visualization), **Insight for ArcGIS (Data Analytics and Visualization).**



GEOG 594 Big Data Science and Analytics Platform (at San Diego State University)

GEOG 594 Big Data Science and Analytics Platform (Fall 2017)

The Center for Human Dynamics in the Mobile Age
San Diego State University.

HOME | Instructor | Syllabus | Lectures | Labs

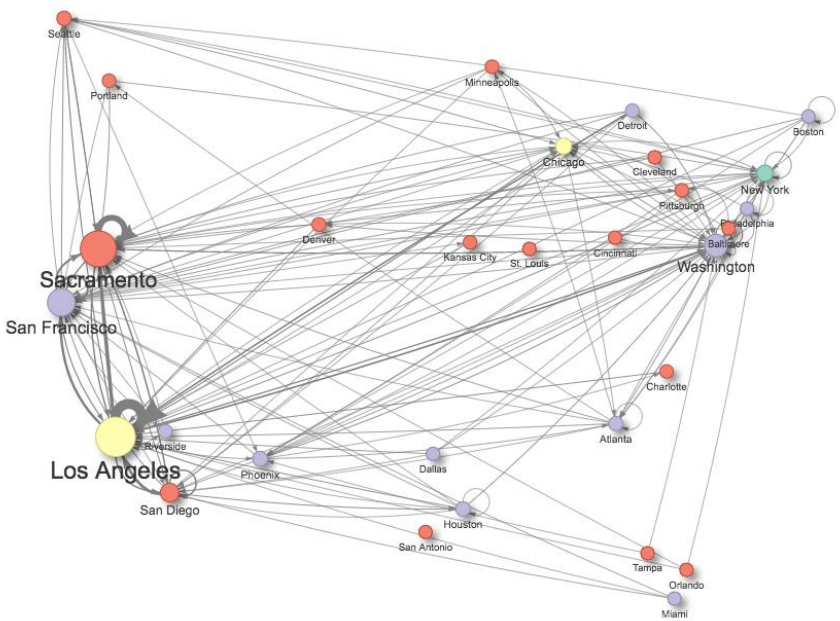
Instructor: **Dr. Ming-Hsiang Tsou** Office: Storm Hall 313C Email: mtsou@mail.sdsu.edu TEL: 619-5940205	Office Hour: Tuesday 10:30am-12:00pm

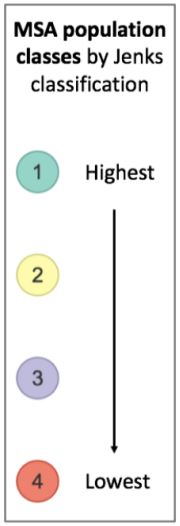
Home / Instructor / Syllabus / Lectures / Labs / Group Projects

Overview: This course introduces state-of-the-art computational platforms, tools, and skills for big data science and big data analytics with numerous real-world case studies. The big data field provides untapped potential for discovering and analyzing complex problems faced by humankind, including business analytics, disease outbreaks, traffic patterns, urban dynamics, and environmental changes. This class will introduce big data platforms (Amazon EC2) and key concepts (cloud computing, virtualization, information privacy, and crowd sourcing). Students will learn to how to use Amazon EC2, MongoDB, R, Gephi, ArcGIS Online, and Tableau to conduct big data analytics. The course will provide basic introduction to big database management related to NoSQL databases, Hadoop and MongoDB. This course will have both the hands-on training of analytics tools and computer skills, as well as the fundamental concepts for big data science with critical thinking. Students will have the opportunity to create their own big data platform on Amazon EC2 virtual servers, manage their own databases in MongoDB, and access and collect big data from sources of their choosing (e.g. Twitter data and business datasets).

3. Develop New Courses for Both Geospatial Technology and Data Science Programs.

- **Spatiotemporal analysis and trajectory analysis** of point data (GPS and social media data), clustered data, and sensor data.
- **Spatial social network analysis** (combing spatial analysis and social network analysis).





HDMA @SDSU | THE CENTER FOR HUMAN DYNAMICS IN THE MOBILE AGE

A new Big Data Analytics program at SDSU in Fall 2019 https://big.sdsu.edu

SAN DIEGO STATE UNIVERSITY

- **Integrating GIScience with Data Science**
- **Focus on real-world data analytics and case studies.**

SAN DIEGO STATE UNIVERSITY

Big Data Analytics Program (Master of Science)

Center for Human Dynamics in the Mobile Age

HOME | ADMISSIONS | CURRICULUM | FACULTY | RESEARCH

Faculty

Ming-Hsiang Tsou, Ph.D.
Program Director
SH 313C
(619) 594-0205
mtsou@mail.sdsu.edu

Dr. Tsou is a professor in the Department of Geography and the director of Center for Human Dynamics in the Mobile Age (HDMA). He received a B.S. (1991) from National Taiwan University, an M.A. (1996) from the State University of New York at Buffalo, and a Ph.D. (2001) from the University of Colorado at Boulder, all in Geography. His research interests include Human Dynamics, Social Media, Big Data, Visualization, Internet Mapping, Web GIS, Mobile GIS, Cartography, and K-12 GIS education.

Curriculum

The Master of Science degree in Big Data Analytics consists of the completion
minimum grade point average of 3.00 in these courses and no less than a C in
will be required to take courses in preparation for the Master's degree that do
warrant.

− Required Core Courses (12 units)

All students must complete the following courses:

- BDA 500 / GEOG 594 Big Data Science and Analytics Platforms (3)
- BDA 572 / LING 572 Python Scripting for Social Science (3)
- BDA 686 / MIS 686 Enterprise Data Management (3)
- BDA 623 / BA 623 Statistical Analysis (3)

+ Electives (6-12 units)

+ Research (3-6 units)

+ Capstone and Culminating Experience (3-6 units)

Education Goals

Geographic Information Science → Geospatial Data Science

- Data (raw materials)
- Information (processed, human readable)
- **Knowledge and Insight** (Actionable – decision making)

- **We need to provide the education training to teach students how to convert data to information, and information to knowledge and insight.**

Geospatial Data Science = Transdisciplinary

We will transform Data Science and Information Technology in the age of Big Data -- from isolated "instruments" (disciplines) into an epic "orchestra" (collaboration).

Image source: wikipedia.org

http://humandynamics.sdsu.edu/

Thank You Q & A

Director: Dr. Ming-Hsiang (Ming) Tsou
mtsou@mail.sdsu.edu
Twitter **@mingtsou**

Funded by

- **The National GeoTech Center 3.0: DUE 1700496** **http://www.geotechcenter.org/**
- NSF Interdisciplinary Behavioral and Social Science (IBSS) Program, Award #1416509 ($1 million (PI: Tsou, 2014-2019). "Spatiotemporal Modeling of Human Dynamics Across Social Media and Social Networks". http://socialmedia.sdsu.edu/
- NSF IMEE program. Award#: 1634641, **Integrated Stage-based Evacuation with Social Perception Analysis and Dynamic Population Estimation. $449,202**, PI: **Tsou**, **2016-2019**. http://decisionsupport.sdsu.edu