GeoEd' 17 Conference

Jefferson Community & Technical College, Louisville, KY

Wednesday, June 07, 2017, 1:00pm – 1:30pm

Big Data and Impact on Geospatial Education

Dr. Ming-Hsiang Tsou

Email: mtsou@mail.sdsu.edu, Twitter **@mingtsou**

Director of **the Center for Human Dynamics in the Mobile Age**

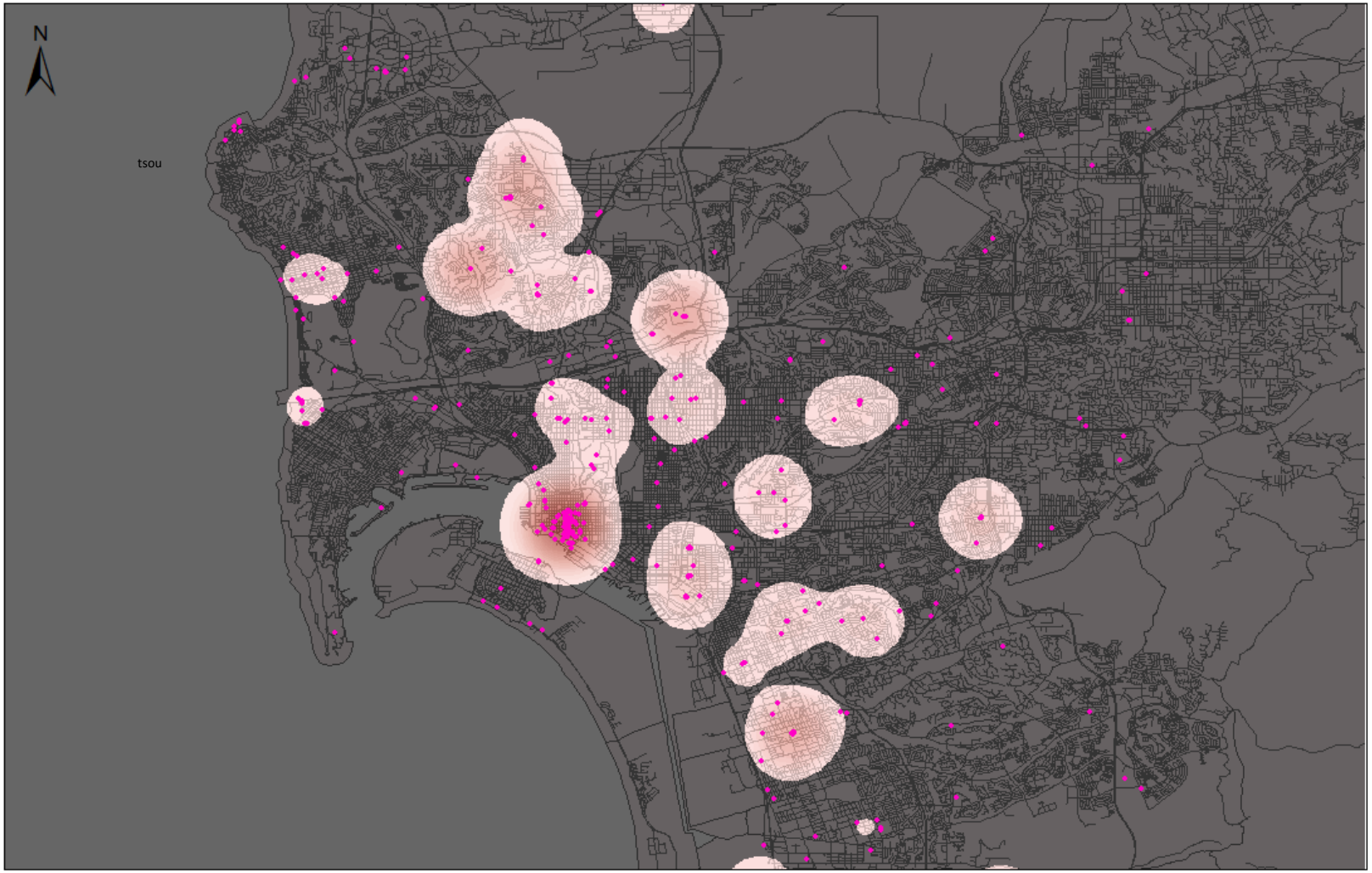
Professor, Department of Geography , San Diego State University

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

SAN DIEGO STATE UNIVERSITY

What is Big Data?

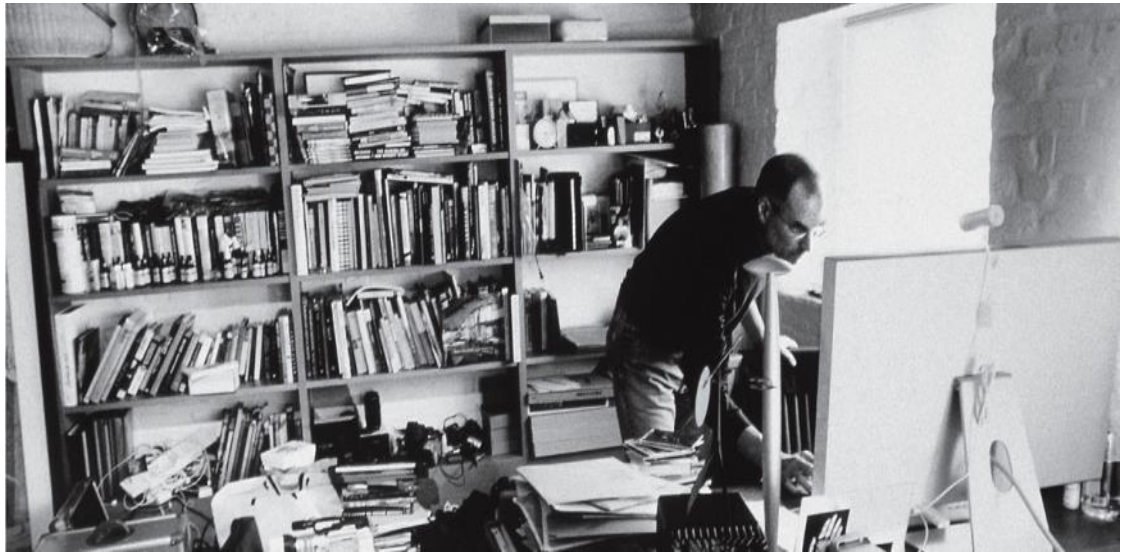
Twitter User Activity during 0:00 to 1:00 in San Diego City for the Weekdays of Month of July, 2015 (Kernel Density)



Animated Image created by the HDMA Center (Hao Zhang).

The Challenges of Big Data Analytics:

Big Data are very Messy, Noisy, and Unstructured!



tsou

Image Source: http://www.contentverse.com/office-pains/10-messy-desks-successful-people/

Require collaboration efforts from linguistics, **geographers (GIS experts),** computer scientists, data mining experts, **statisticians**, physicists, modelers, and domain experts.

The Definitions of Big Data

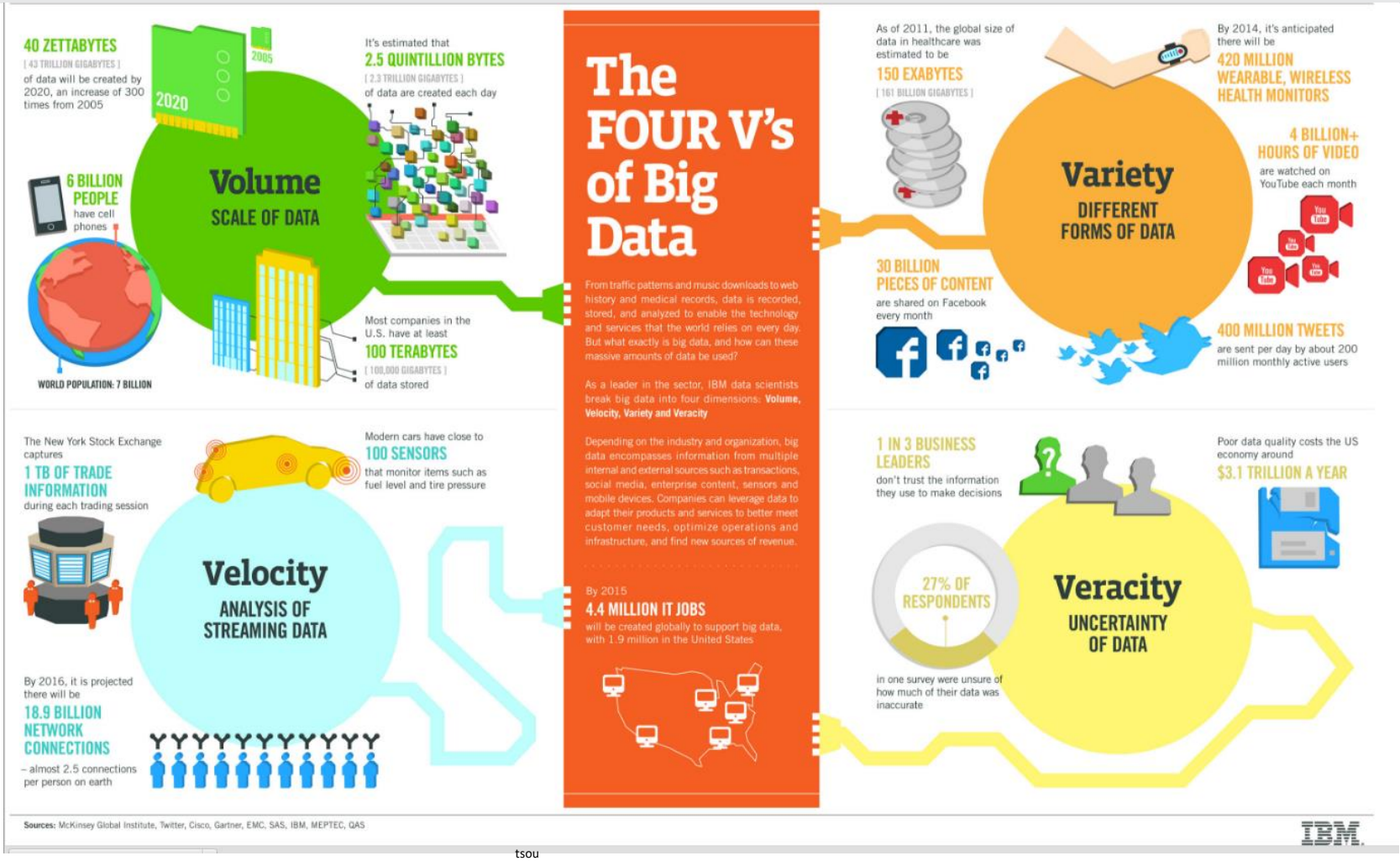
One popular definition of Big Data is that "data is too large, complex, and dynamics of any conventional data tools to capture, store, manage, and analyze (WIPRO 2012). Researchers emphasize three major characteristics of Big Data: large volume, large variety, and high velocity (3Vs) (White 2012; IBM 2012). (4V: adding "Value" or "Veracity").

Wikipedia: "Big data is a term for data sets that are so large or complex that **traditional data processing applications** are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating andinformation privacy." https://en.wikipedia.org/wiki/Big_data

Main problem with these definitions: How to define "data is too large"? How big is "too large"? 100TB? 10,000TB?

Today's Big Volume data will become "small data" in Five Years.

The 4 V's of 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

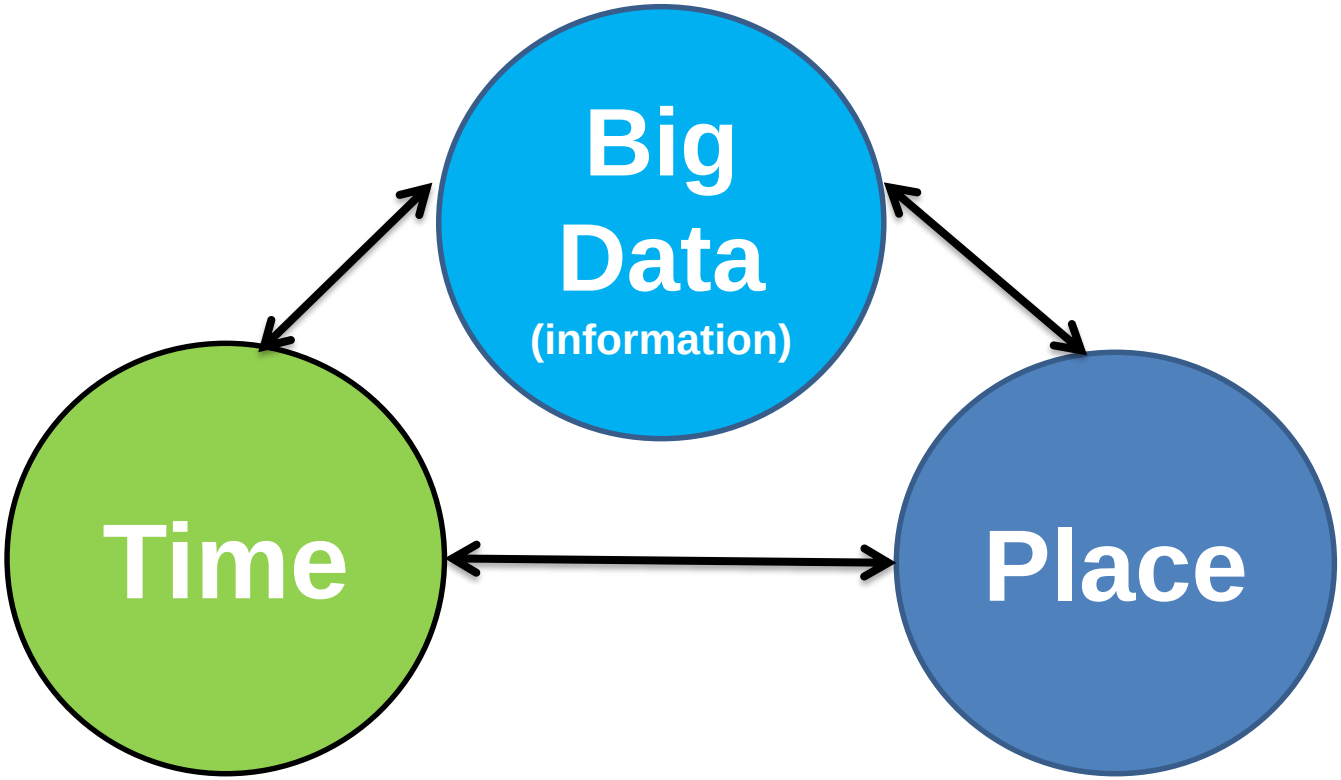
Big Data is Human-Centered Data

*Big Data is a large **dynamic** dataset created by or derived from **human activities, communications, movements, and behaviors**. (Tsou, 2015).*

The term, Big Data, refers to **big ideas, big impacts, and big changes** for our society in addition to a **big volume** of datasets.

Tsou, M. H. (2015). Research challenges and opportunities in mapping social media and Big Data. Cartography and Geographic Information Science, 42:sup1, 70-74. doi: 10.1080/15230406.2015.1059251. http://www.tandfonline.com/doi/full/10.1080/15230406.2015.1059251#.VeCVyPlVhBd

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

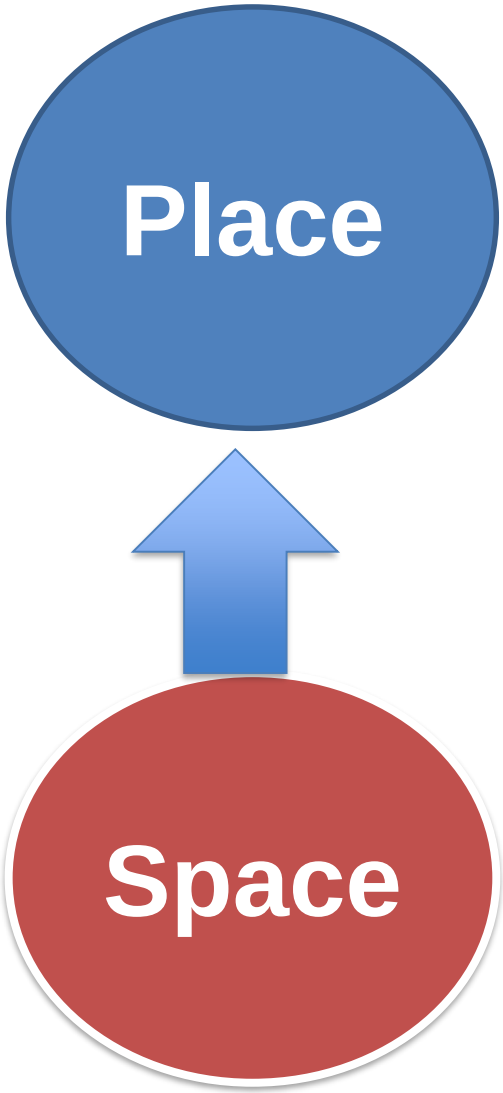


(Tsou and Leitner, 2013)

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

How to re-define and analyze "Place"?



= **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).

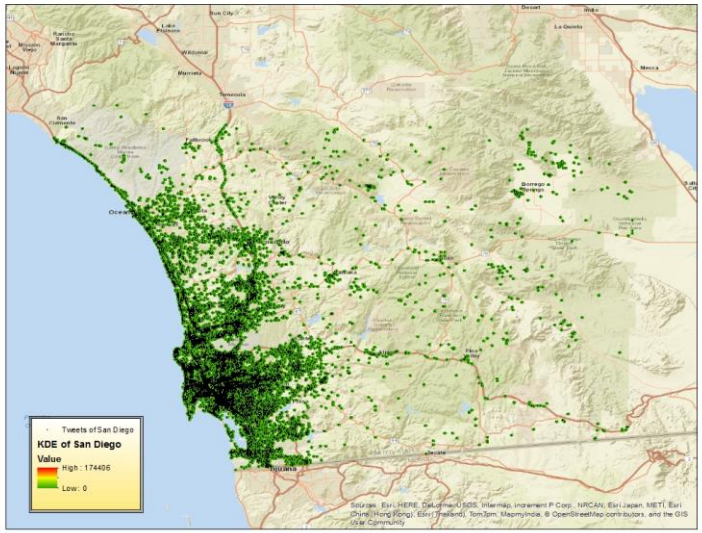
Tuan, Yi-Fu. *Space and place: The perspective of experience*. U of Minnesota Press, 1977.

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

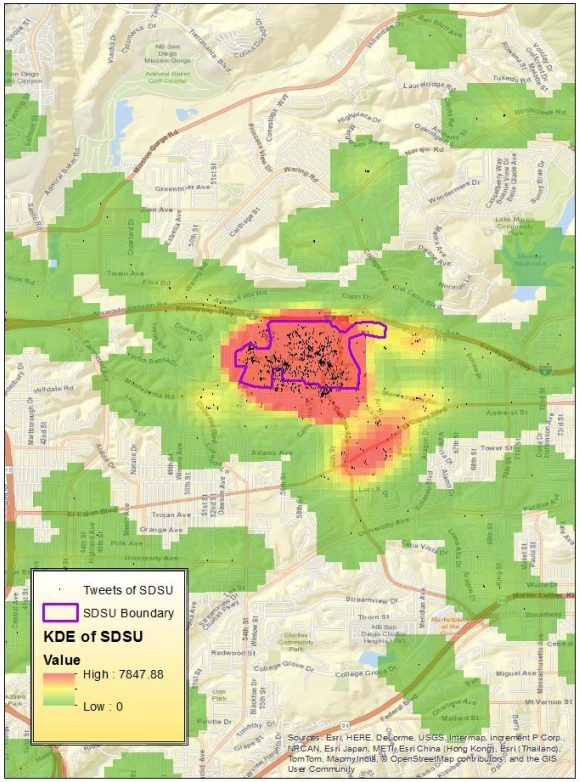
Define “Places” using Social Media

SAN DIEGO STATE UNIVERSITY

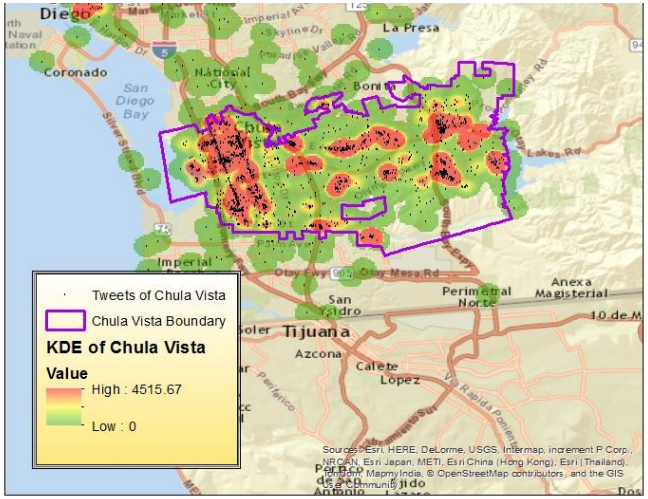
Tweets mentioned “San Diego”



Tweets mentioned “SDSU”



Tweets mentioned “Chula Vista”



We can define “place” by aggregating thousands of geo-tagged tweets mentioned the name of “place” with linguistic analysis (content analysis + word cloud).

How to re-define and analyze "Human Time"?

Time Stamps vs. Human-Centered Time (Defined by Human Activities)

- **Absolute Time:** UTC time stamps (**Coordinated Universal Time** (UTC).
- **Local Time**: converting UTC to local time zone (Twitter Timestamp is UTC time).
- **Human-Centered Time:** (Sleep, working, eating, playing times, Weekend and Weekday).
- Traffic **level of services (LOS) Skims** Time:

NUMBER	DESCRIPTION	BEGIN TIME	END TIME
1	Early	3:00 A.M.	5:59 A.M.
2	A.M. Peak	6:00 A.M.	8:59 A.M.
3	Midday	9:00 A.M.	3:29 P.M.
4	P.M. Peak	3:30 P.M.	6:59 P.M.
5	Evening	7:00 P.M.	2:59 A.M.

Data Integration / Data Fusion

Explore their **spatiotemporal relationships** in both **network space (cyberspace)** and **geographical space (real world).**

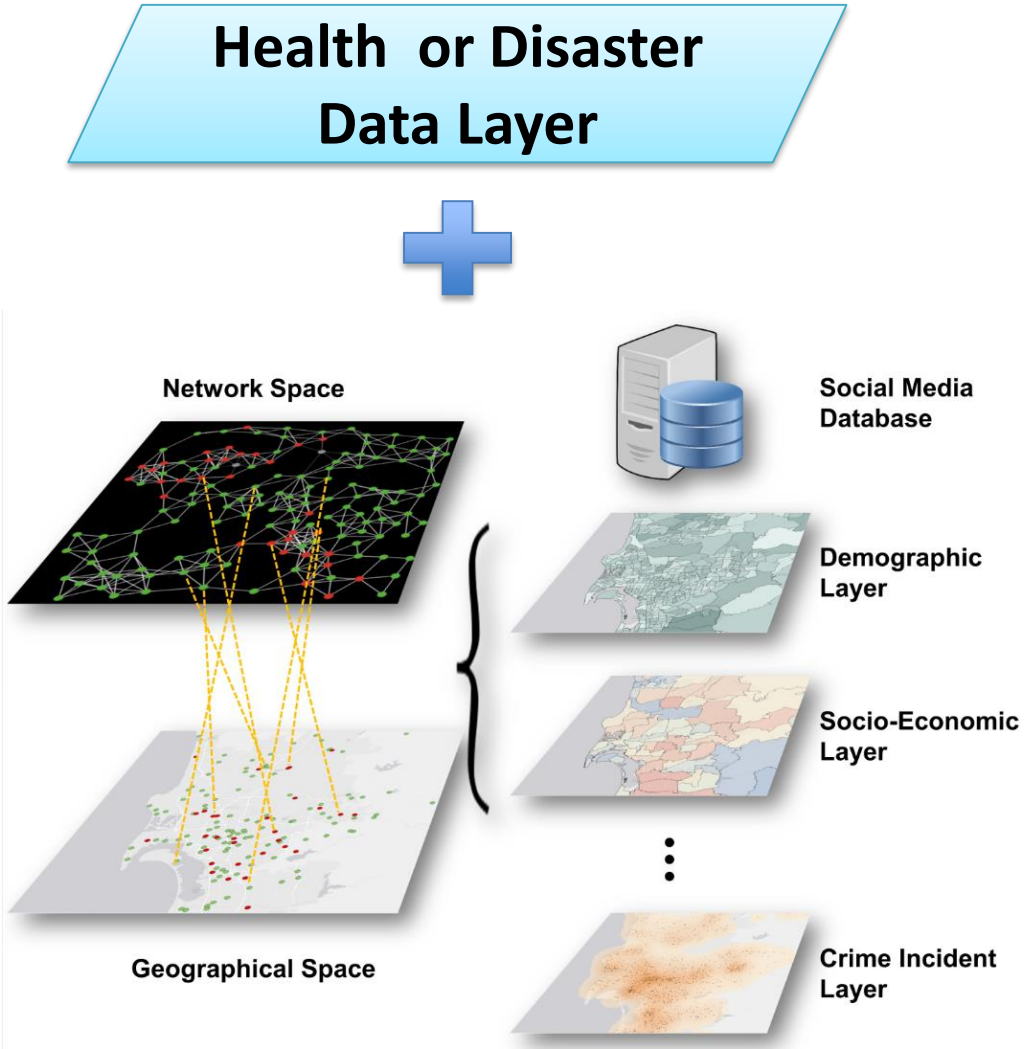


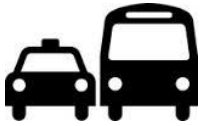
Image provided by Dr. Atsushi Nara (Associate Director of HDMA Center).

Big Data Category (Tsou, 2015).

Social life data: social media services (**Twitter**, Flickr, Snapchat, YouTube, Foursquare, etc.), online forums, online video games, and web blogs.

Health data: electronic medical records (**EMR**) from hospitals and health centers, **cancer registry data**, disease outbreak tracking and epidemiology data.

Business and commercial data: credit card transactions, online business reviews (such as Yelp and Amazon reviews), supermarket membership records, shopping mall transaction records, credit card fraud examination data, enterprise management data, and marketing analysis data.



Transportation and human traffic data: GPS tracks (from taxi, buses, **Uber**, bike sharing programs, and mobile phones), traffic censor data (from subways, trolleys, buses, bike lanes, highways), and mobile phone data (from data transmission records and cellular network data).

Scientific research data include earthquakes sensors, weather sensors, satellite images, crowd sourcing data for biodiversity research, volunteered geographic information, and census data.

Geography (place and time) is the KEY for understanding Big Data!

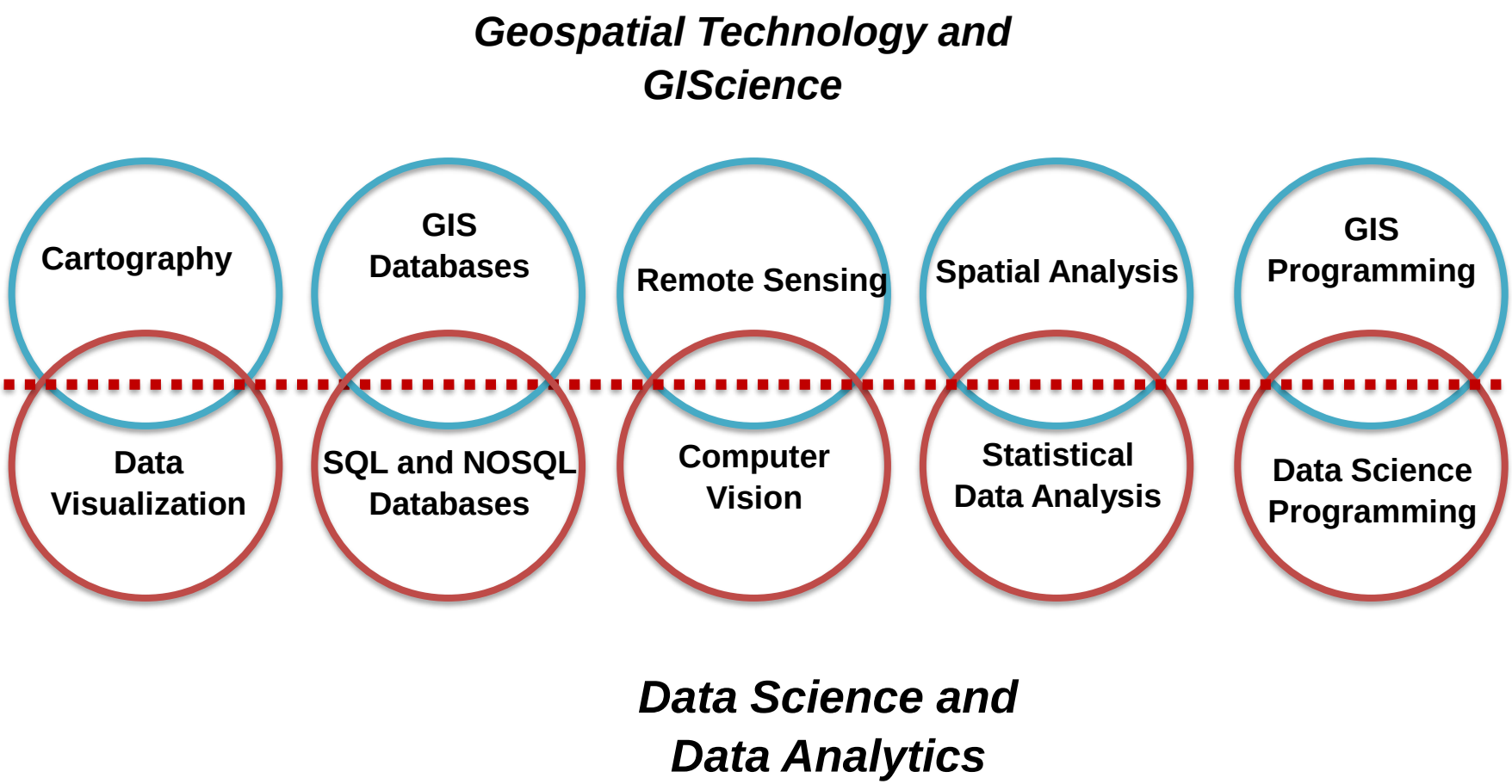
We are born to deal with Big Data!

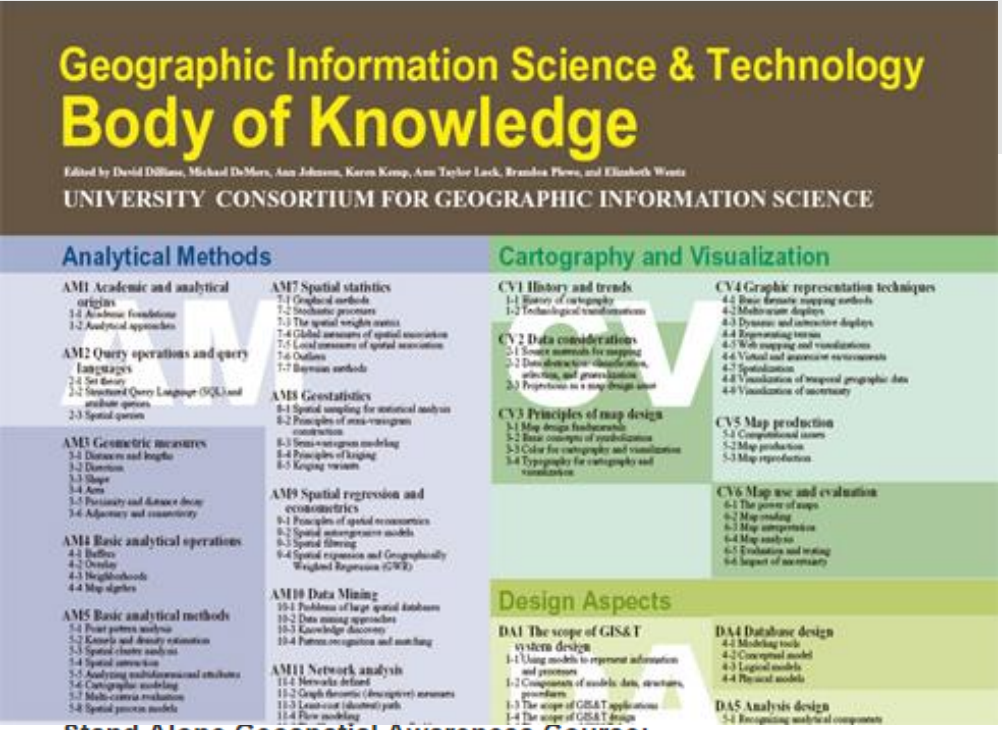
Great History of Big Data Processing and Analysis in GIS and Geospatial Analysis Applications

- **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).

The Overlap in Curricula

Comparing Geospatial Technology Programs/Curricula and Data Science/Data Analytics Programs/Curricula





Stand Alone Geospatial Awareness Course:

GST 100 – Exploring Our World Fundamentals of Geospatial Science – Syllabus and Description

Certificate Module Course:

GeoTECH GEOTECHCENTER.ORG

National Geospatial Technology Center of Excellence

Empowering Colleges:
EXPANDING THE GEOSPATIAL WORKFORCE

GST 101 – Introduction to Geospatial Technology – Syllabus and Description
GST 102 – Spatial Analysis - Syllabus and Description
GST 103 – Data Acquisition & Management - Syllabus and Description
GST 104 – Cartographic Design - Syllabus and Description
GST 105 – Introduction to Remote Sensing - Syllabus and Description
GST 106 – Introduction to Geospatial Programming - Syllabus and Description
GST 107 – Geospatial Web Applications and Development - Syllabus and Description
GST 108 – Capstone in Geospatial Technology - Syllabus and Description
GST 109 – Internship in Geospatial Technology - Syllabus and Description

Data Science Curricula Sample

UC-Berkeley MIDS (Master of Information and Data Science)

MIDS is designed to be completed in 20 months, but other options are available to complete the program on an accelerated basis. The 27 units of courses are listed below:

Part A: Foundation Courses

- **Research design** and applications for **data and analysis**
- Exploring and analyzing data
- Storing and **retrieving data**
- Applied **machine learning**
- **Visualizing** and communicating data

Part B: Advanced Courses

- Field experiments
- Legal, policy, and ethical considerations and statistics
- Scaling up! Really big data

Part C: Capstone Course

- **Synthetic capstone course**

Stanford: M.S. in Statistics: Data Science

Requirement 1 : Foundational (12 units)

CME 302 Numerical Linear Algebra 3
CME 305 Discrete Mathematics and Algorithms 3
CME 307 Optimization 3
CME 308 Stochastic Methods in Engineering 3
or CME 309 Randomized Algorithms and Probabilistic Analysis 3

Requirement 2 : Data Science Electives (12 units)

STATS 200 Introduction to Statistical Inference 3
STATS 203 Introduction to Regression Models and Analysis of Variance 3
or STATS 305A Introduction to Statistical Modeling
STATS 315A Modern Applied Statistics: Learning 2-3
STATS 315B Modern Applied Statistics: **Data Mining** 2-3

Requirement 3 : Specialized Electives (9 units)

BIOE 214 Representations and Algorithms for Computational Molecular Biology 3-4
BIOMEDIN 215 Data Driven Medicine 3
BIOS 221/STATS 366 Modern Statistics for Modern Biology 3
CS 224W **Social and Information Network Analysis** 3-4
CS 229 **Machine Learning** 3-4
CS 246 Mining Massive Data Sets 3-4
CS 347 **Parallel and Distributed Data Management** 3
CS 448 Topics in Computer Graphics 3-4
ENERGY 240 **Geostatistics** 2-3
OIT 367 **Business Intelligence from Big Data** 3

Team Works in 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
- ?? (GIScience and Geospatial Technology)?

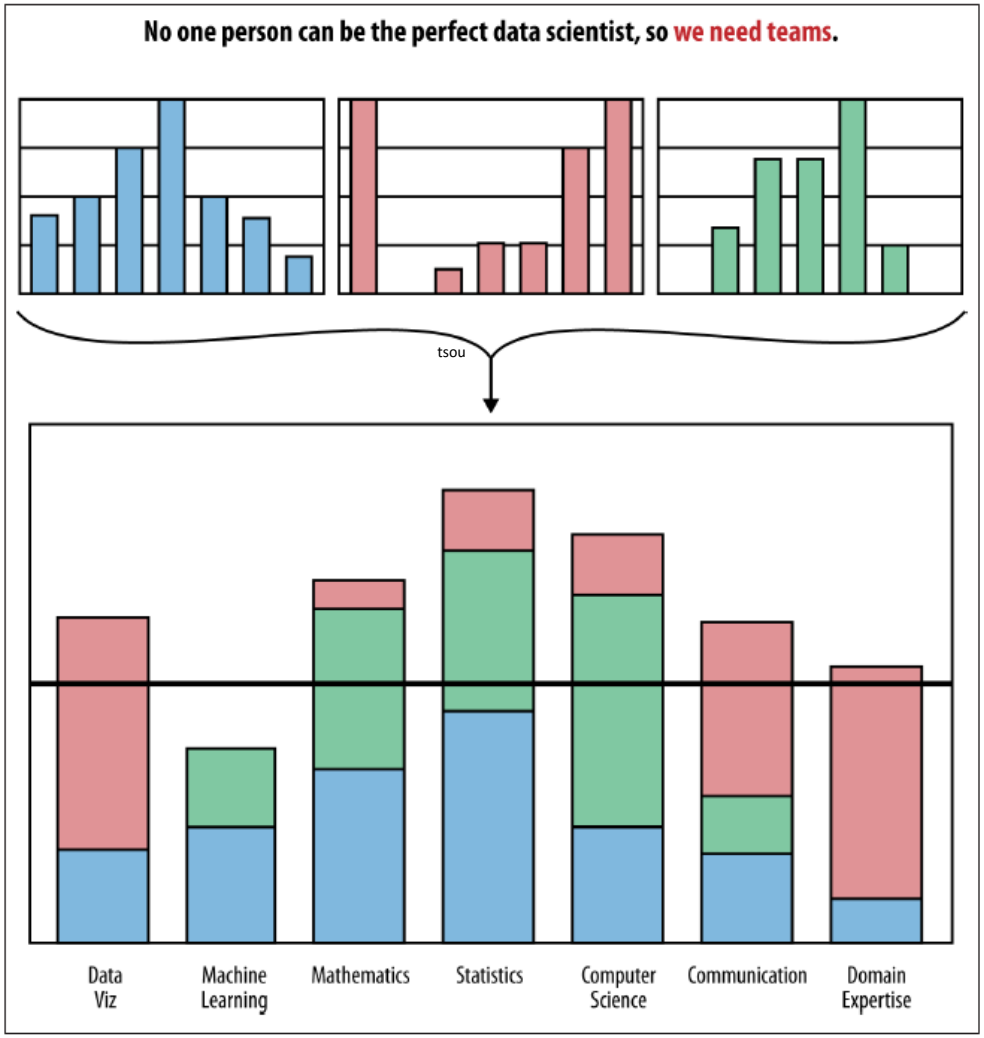


Figure 1-3. Data science team profiles can be constructed from data scientist profiles; there should be alignment between the data science team profile and the profile of the data problems they try to solve

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)
- Different **geocoding** needs (city names, neighborhoods, rather than using street addresses).
- **Data uncertainty**: Sampling and representation (Twitter's 1% public data feed).
- Data Privacy and locational privacy protection methods
- Content-rich data and linked data (cross linked by usernames, geolocation, time).
- Data ownership problems. (Private Companies: Facebook, Twitter, Flickr)

The differences

Geospatial Technology

- **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)

Different content

- 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)

Same content

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

Machine Learning

- **Supervised machine learning (labeled training data)**:
  - **kNN (k Nearest Neighbors)**
  - **Linear Regression**
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machines
  - Random Forests
  - Time Series Analysis (Forecasting)
- **Unsupervised machine learning (describe hidden structure from unlabeled data)**:
  - Clustering (**k-means**, DBSCAN, etc…)
  - Factor analysis (PCA, ….)
  - Topic Models

How to Enhance Geospatial Technology Education with Big Data / Data Science?

1. **Add New Data Science Courses into Geospatial Technology and GIS Programs/Curriculum**
   - *GIS 510: Introduction to Big Data*
   - *GIS 520: Common Technologies for Big Data Science and Analytics*
   - *GIS 530: Methods and Key Concepts in Data Science and Data Analytics*

2. **Improve Current Geospatial Technology Courses with Data Science and Data Analytics Methods and Tools.**

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

1. Add Data Science Courses into Geospatial Technology Programs

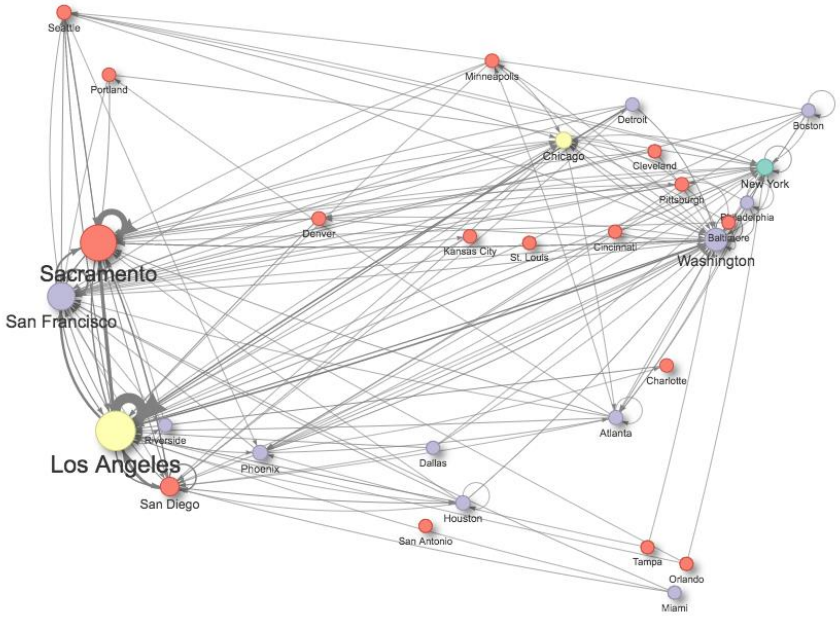
- ***GIS 510 Introduction to Big Data***
  - Big Data Collection Methods
  - Sampling and Re-Sampling in Big Data, Dealing with Biased Data and Missing Data problems, Noise Filtering and Remove.
  - Social Media APIs (Twitter, Facebook, Instagram, etc.)
  - GeoJSON and other data formats (CSV, Excels, Texts, etc.) in social science and public health, Examples in social science and public health.
- ***GIS 520 Common Technologies for Big Data Science and Analytics***
  - Cloud Computing and High Performance Computing
  - Amazon EC2 and other Cloud platform examples, Hadoop and Spark
  - Software Packages (R, Tablueu, Pythons, etc).
  - Database management and integration for Big Data, NoSQL databases (MongoDB)
  - Applications and Case Studies
- ***GIS 530 Methods and Key Concepts in Data Science and Data Analytics***
  - Tools and Software for Data Science, Statistical Inference (R software)
  - Machine Learning (Algorithm) (scikit-lean)
  - Social Network Analysis (Gephi and Crypth..)
  - Computational Linguistic Analysis (WISD), Data Processing and Noise Filtering
  - Critical Thinking in Data Science and Data Analytics.

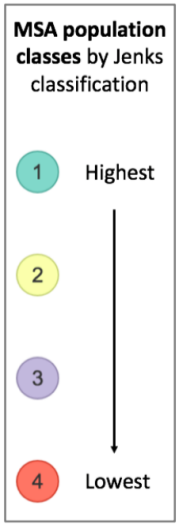
2. Improve current Geospatial Technology Courses with Data Science and Data Analytics Methods and Tools.

- **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**.

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).





Education Goals

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

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

- Traditional GIS emphasize on using **GIS software** to convert "data" to "information". With data science, GIS analysis will utilize more software, more methods and more techniques to convert "data" to "information" and to "knowledge" (actionable).

Geographic Information Science vs. Geographic Data Science? or Geospatial Data Science (GDS)?

Final Remark: Big Data = Transdisciplinary

Geospatial Technology is important for Big Data Science.

We will transform Science and 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

- 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