Geospatial Data Acquisition and Evaluation

Center for Geographic Analysis

Harvard University

Nicole Alexander, Ph. D.

nalexander@cga.harvard.edu

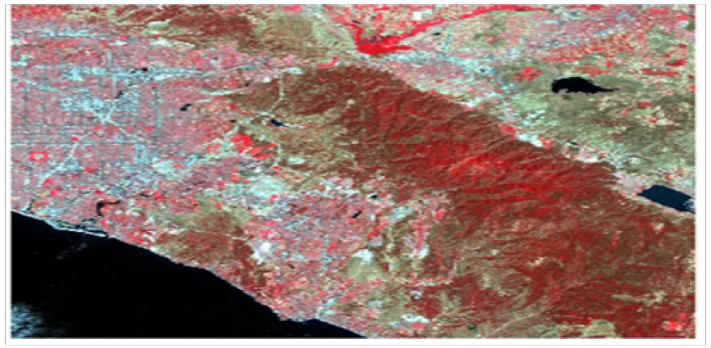
Harvard DataFest 2019

Geospatial Data Acquisition and Evaluation

- Geospatial Data Sources
- Data Transfer
- Metadata
- Big Data Collection and Analytics
- Geocoding Datasets in QGIS using the MMQGIS Plugin

Geospatial Data Acquisition and Evaluation

- How data are captured determines the quality of decisions that can be made from analyzing the data
  - *Primary* sources: obtained through direct measurement
  - *Secondary* sources: derived from other sources
- Data accuracy can more reliably be determined from primary sources

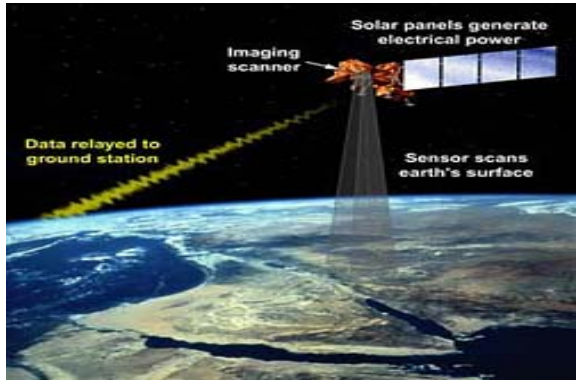


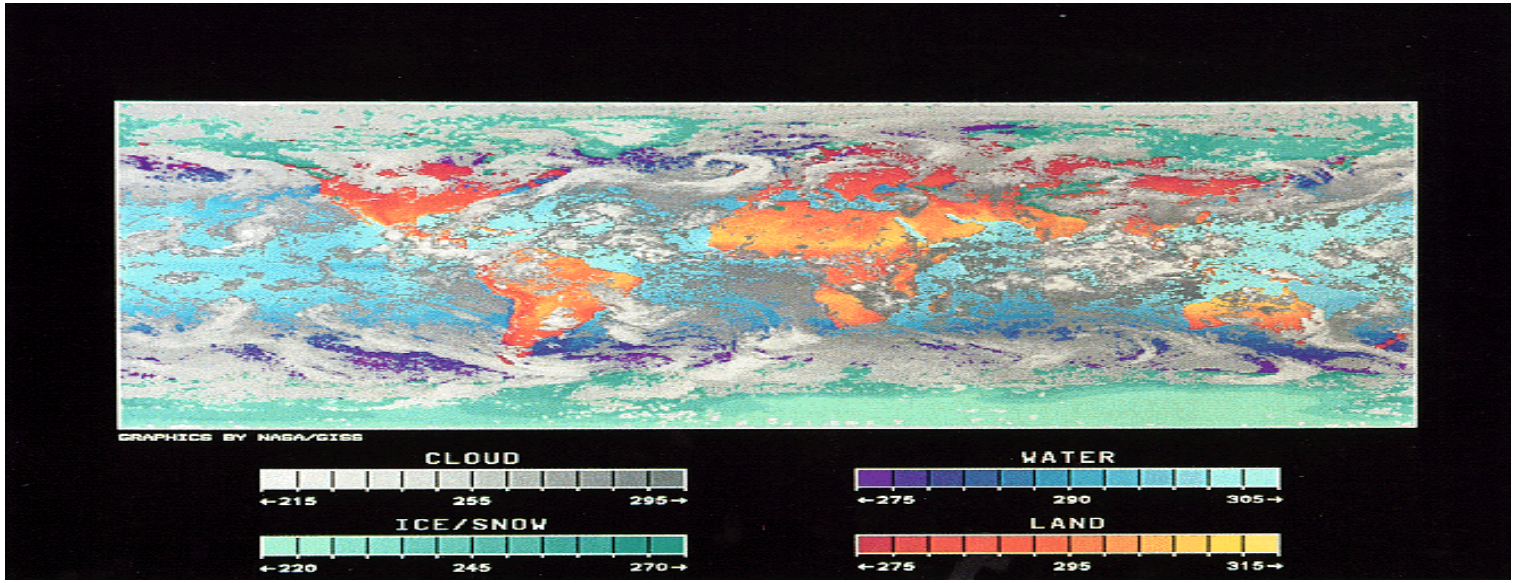
Geospatial Data Sources

	RASTER	VECTOR
Primary	• Digital satellite remote-sensing images • Digital aerial photographs	• GPS measurements • Field survey measurements • LiDAR
Secondary	• Scanned maps and photographs • Digital elevation models from topographic map contours	• Topographic maps • Toponymy (place-name) databases • Geocoding

Primary Raster Data Capture

- Remote sensing
  - Satellite
  - Aircraft
- Image Resolution
  - Spatial
  - Spectral
  - Temporal



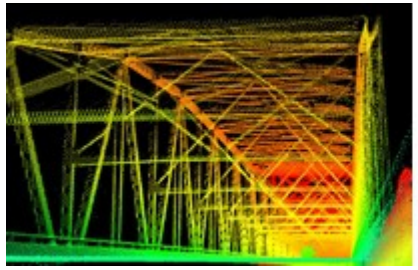


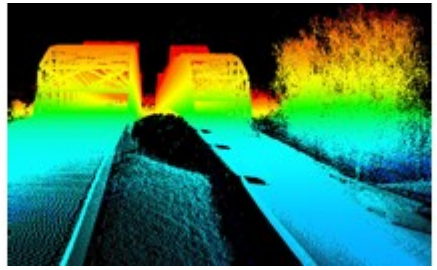
Remote Sensing Data Capture

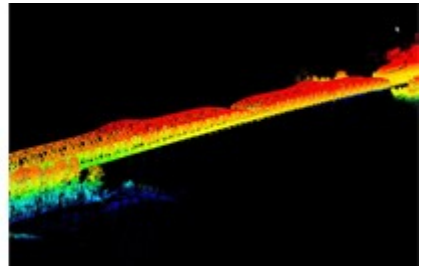
- Captures data over a large geographic areas
  - Total ground coverage range from 9 x 9 km – 200 x 200 km
- Pixel size determines spatial resolution of an image
  - Spatial accuracy of features increases as pixel size decreases
- Satellite systems capture data in the range of 0.5 m – 1 km pixel size
- Camera systems capture data in the range of 0.01 m – 5 m pixel size
- Costly compared with other methods of data capture
- Data volumes can be very large

Primary Vector Data Capture

- Main sources
  - GPS
  - Surveying
- Remote Sensing
  - LiDAR (Light Detection And Ranging)
    - a “cloud” of points that reflects the surface







Primary Vector Data Capture

- GPS
  - Recreational: low precision 6 – 12m
  - Mapping and GIS: medium precision 30cm – 5m
  - Surveying: high precision 5mm – 1cm
- Surveying
  - Used for large scale mapping of small areas and property boundaries
  - Capable of 1 mm accuracy
  - Equipment and crews are very expensive
- LiDAR
  - 30,000 points per second at an accuracy of around 15cm
  - Often rasterized to create DEMs

Secondary Raster Data Capture

- Scanning of hardcopy media
  - Building plans, CAD drawings, property deeds, film, paper maps, aerial photographs, images, etc.
  - Spatial resolution of scanners in the range of 400 – 900 dpi (16 – 40 dots per mm)
- DEM generation from topographic map contours or LiDAR

Secondary Vector Data Capture

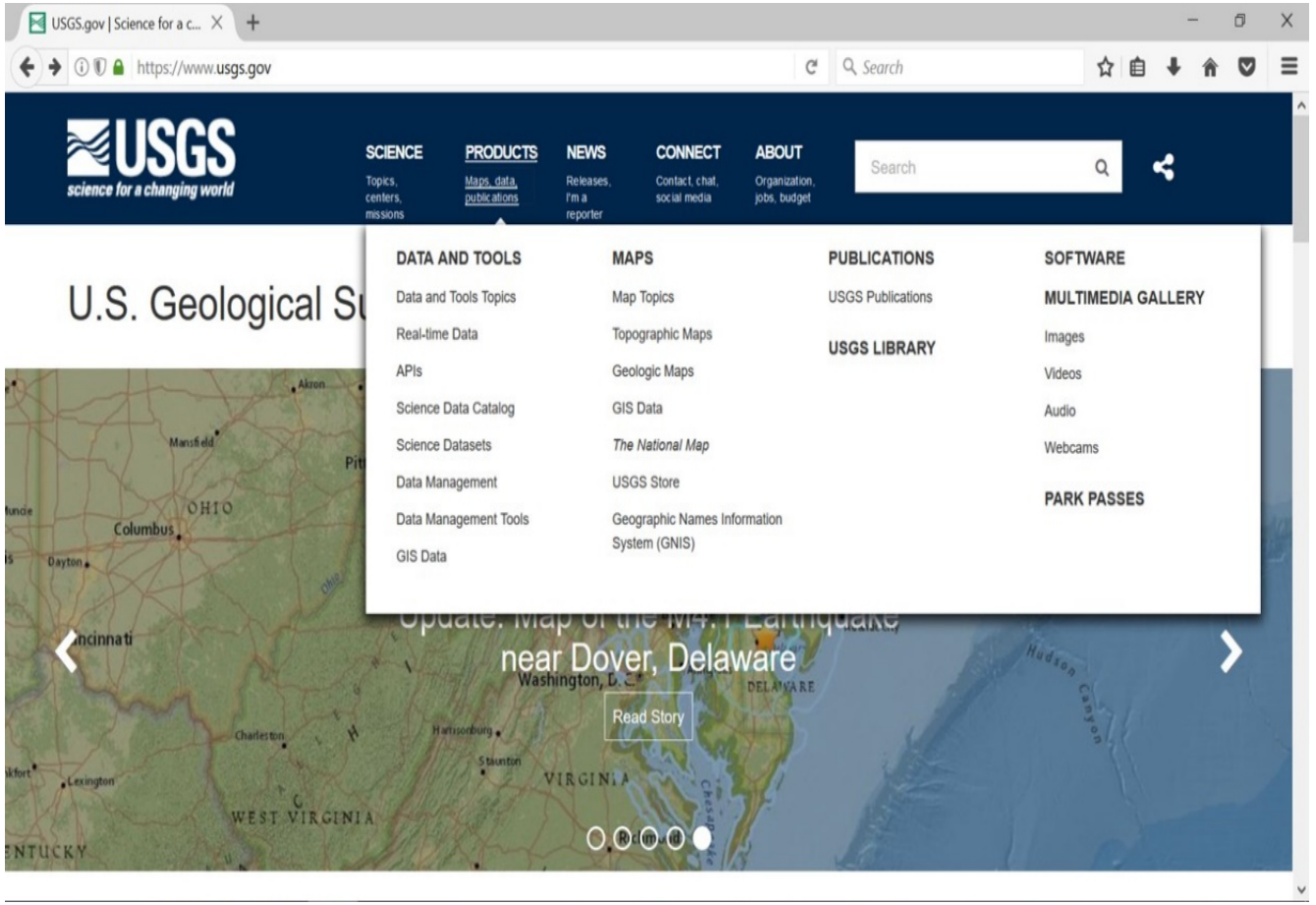
- Vectorizing raster data
- Digitizing
- Geocoding
- Photogrammetry
- COGO – Coordinate Geometry

Data Transfer: Obtaining Data from External Sources

- U.S. Geological Survey
- U.S. Census Bureau
- OpenStreetMap
- GeoNames
- Other Geospatial Data Sites

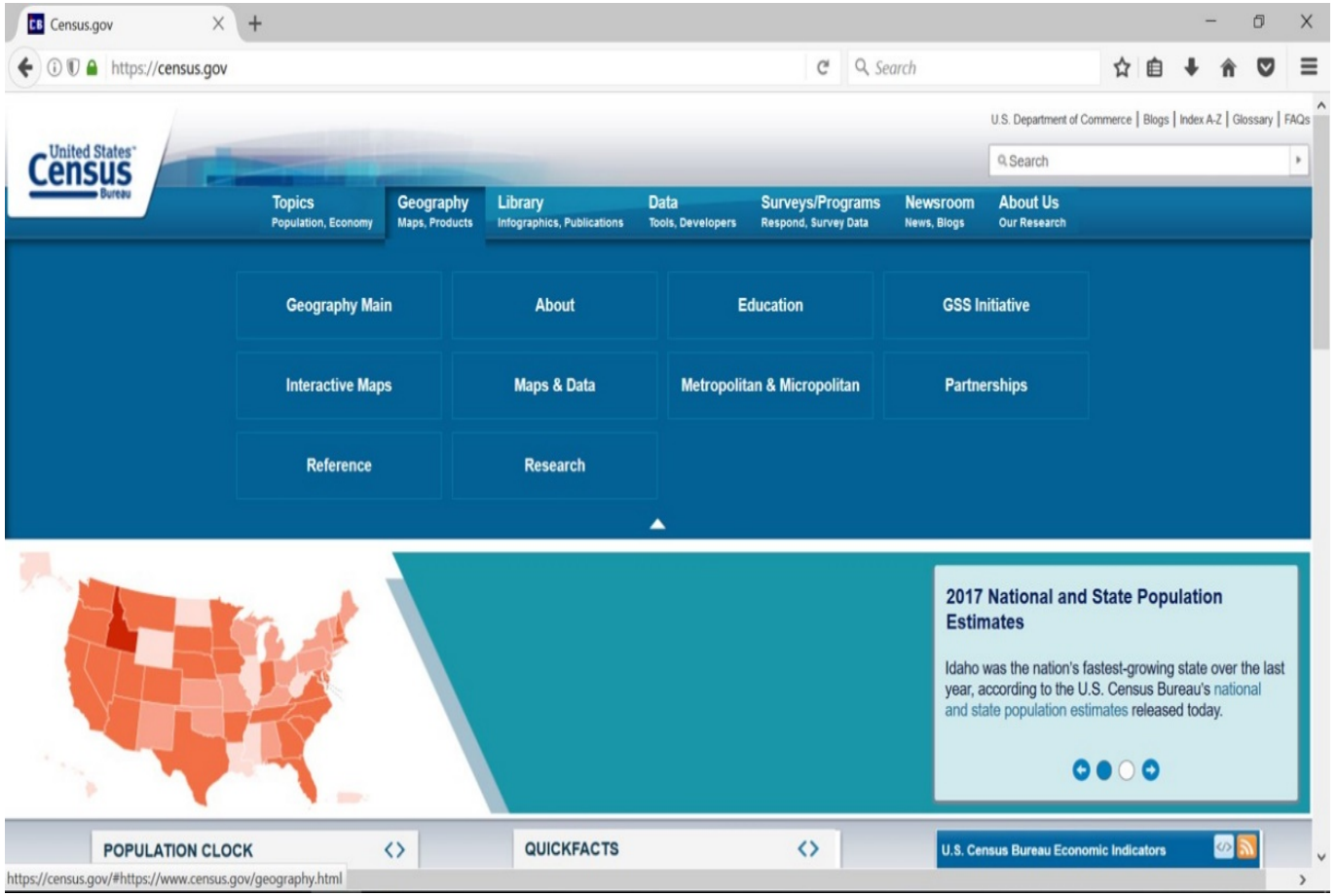
U.S. Geological Survey (usgs.gov)

- The major provider of geospatial data in the US



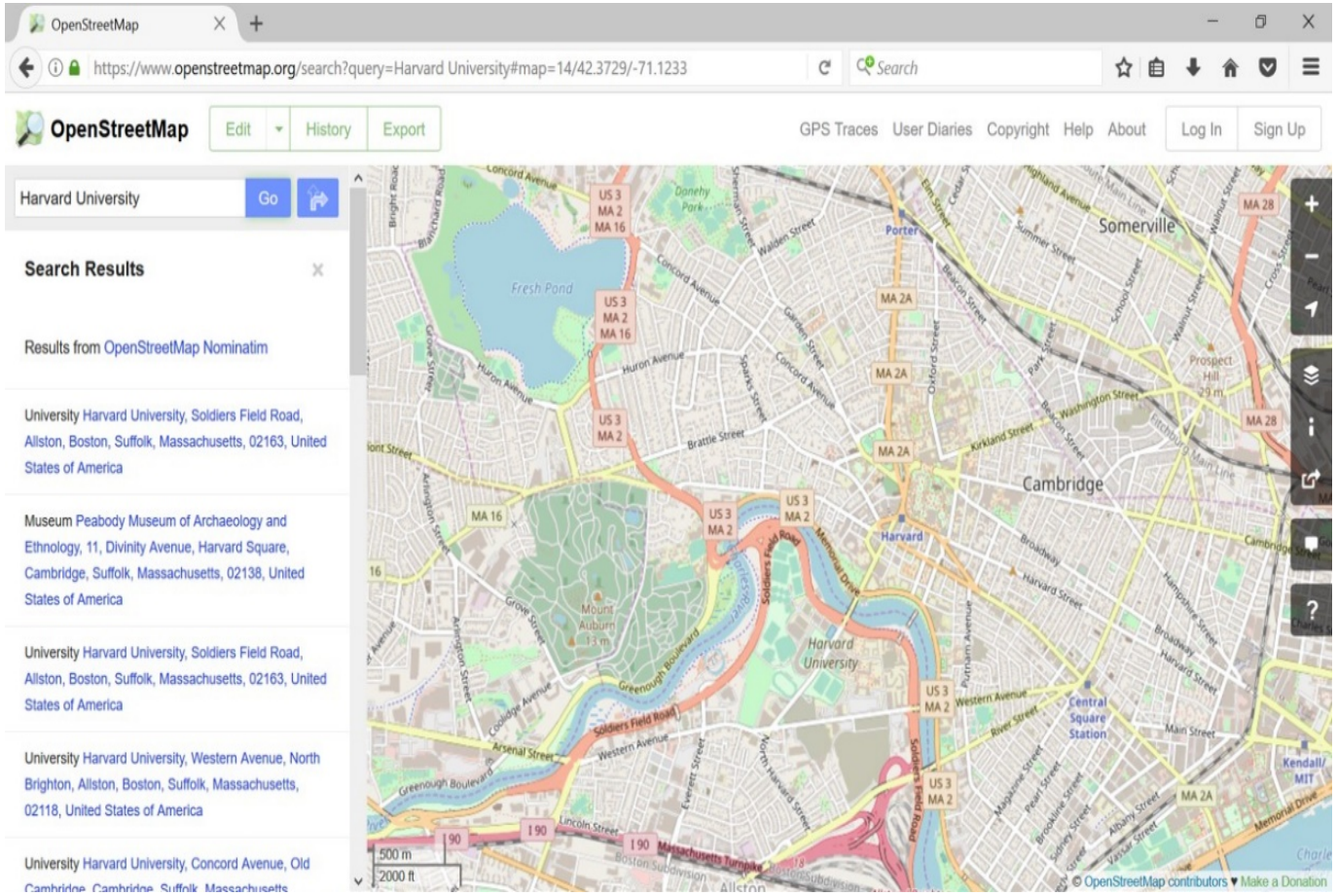
U.S. Census Bureau (census.gov)

- Provides data to support the US decennial census



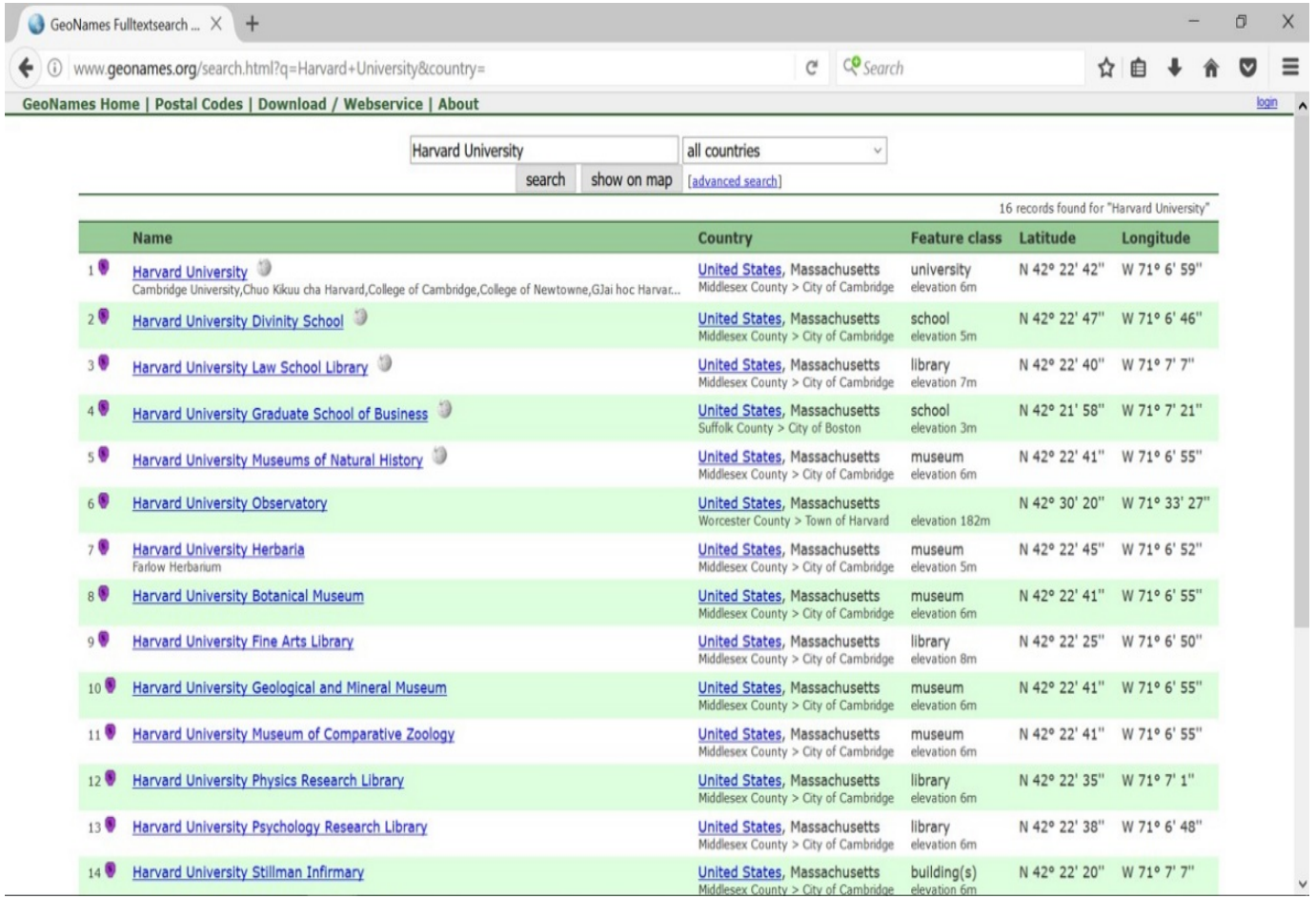
OpenStreetMap (OSM) (openstreetmap.org)

- Map data of the world
- Created and maintained by a community of mappers



GeoNames (geonames.org)

- Global geographical database of place names



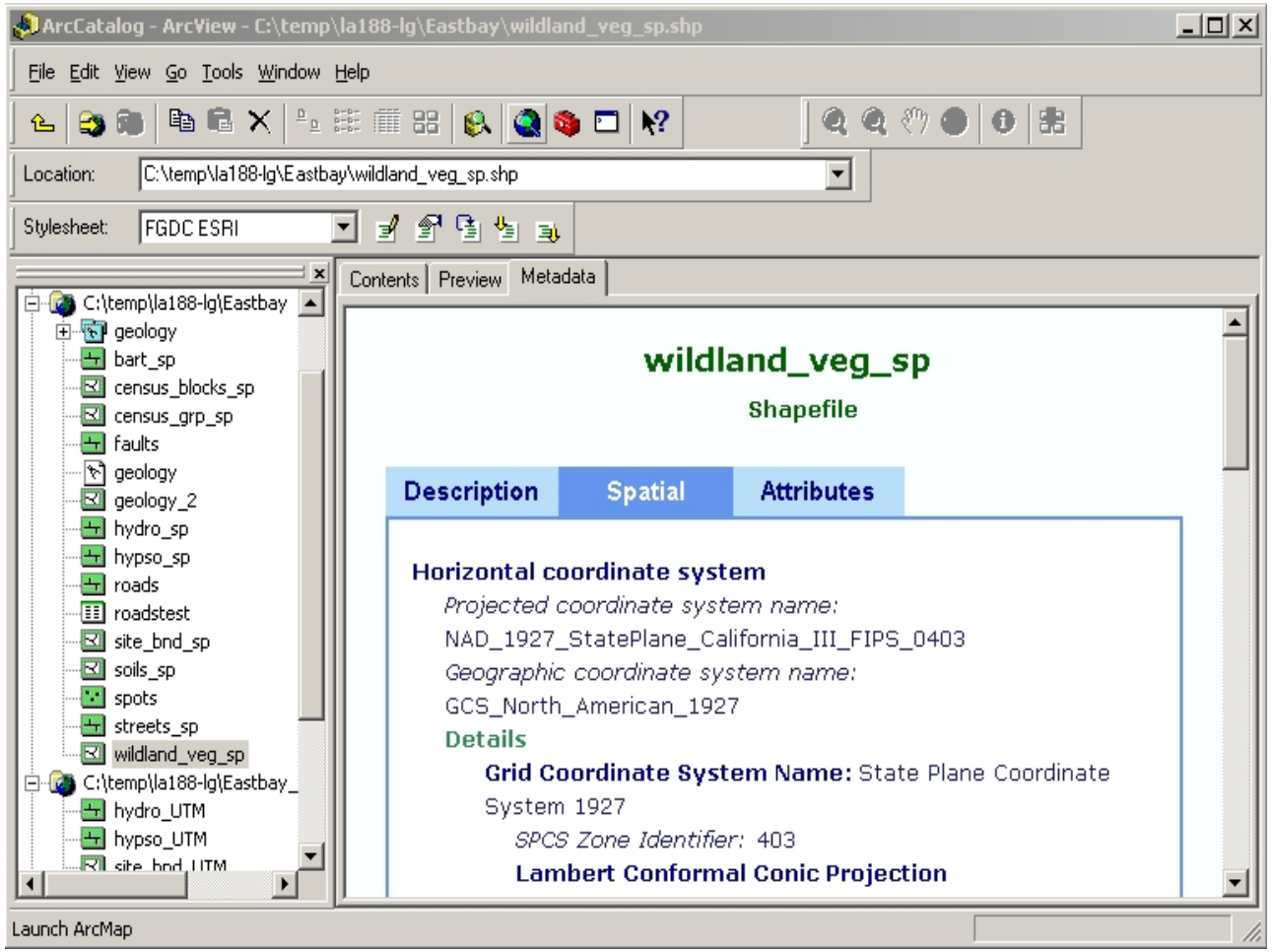
	Name	Country	Feature class	Latitude	Longitude
1	Harvard University Cambridge University,Chuo Kikuu cha Harvard,College of Cambridge,College of Newtowne,GJai hoc Harvar...	United States, Massachusetts Middlesex County > City of Cambridge	university elevation 6m	N 42° 22' 42"	W 71° 6' 59"
2	Harvard University Divinity School	United States, Massachusetts Middlesex County > City of Cambridge	school elevation 5m	N 42° 22' 47"	W 71° 6' 46"
3	Harvard University Law School Library	United States, Massachusetts Middlesex County > City of Cambridge	library elevation 7m	N 42° 22' 40"	W 71° 7' 7"
4	Harvard University Graduate School of Business	United States, Massachusetts Suffolk County > City of Boston	school elevation 3m	N 42° 21' 58"	W 71° 7' 21"
5	Harvard University Museums of Natural History	United States, Massachusetts Middlesex County > City of Cambridge	museum elevation 6m	N 42° 22' 41"	W 71° 6' 55"
6	Harvard University Observatory	United States, Massachusetts Worcester County > Town of Harvard	elevation 182m	N 42° 30' 20"	W 71° 33' 27"
7	Harvard University Herbaria Farlow Herbarium	United States, Massachusetts Middlesex County > City of Cambridge	museum elevation 5m	N 42° 22' 45"	W 71° 6' 52"
8	Harvard University Botanical Museum	United States, Massachusetts Middlesex County > City of Cambridge	museum elevation 6m	N 42° 22' 41"	W 71° 6' 55"
9	Harvard University Fine Arts Library	United States, Massachusetts Middlesex County > City of Cambridge	library elevation 8m	N 42° 22' 25"	W 71° 6' 50"
10	Harvard University Geological and Mineral Museum	United States, Massachusetts Middlesex County > City of Cambridge	museum elevation 6m	N 42° 22' 41"	W 71° 6' 55"
11	Harvard University Museum of Comparative Zoology	United States, Massachusetts Middlesex County > City of Cambridge	museum elevation 6m	N 42° 22' 41"	W 71° 6' 55"
12	Harvard University Physics Research Library	United States, Massachusetts Middlesex County > City of Cambridge	library elevation 6m	N 42° 22' 35"	W 71° 7' 1"
13	Harvard University Psychology Research Library	United States, Massachusetts Middlesex County > City of Cambridge	library elevation 6m	N 42° 22' 38"	W 71° 6' 48"
14	Harvard University Stillman Infirmary	United States, Massachusetts Middlesex County > City of Cambridge	building(s) elevation 6m	N 42° 22' 20"	W 71° 7' 7"

Other Geospatial Data Sites

- Harvard University
  - CGA: http://gis.harvard.edu/resources/data
  - Harvard Geospatial Library: http://hgl.harvard.edu
  - Harvard WorldMap: http://worldmap.harvard.edu
  - Harvard Map Collection: http://hcl.harvard.edu/libraries/maps/collections/digital.html#overview
- Local
  - MassGIS: http://www.mass.gov/mgis
  - City of Boston: https://data.boston.gov/dataset?groups=geospatial
  - Metro Boston Data Common: http://www.metrobostondatacommon.org/
- National
  - US Federal Government: http://data.gov
  - US Geological Survey: http://viewer.nationalmap.gov/viewer/
- Global
  - The ESRI Data and Maps: http://bit.ly/NBoQzQ
  - ArcGIS Online Services from ESRI: http://www.arcgis.com/home/

Metadata

- Data about the geospatial data:
  - Identification
  - Data quality
  - Coordinate system
  - Attributes, etc.
- Especially important when using public data



What is Big Data?

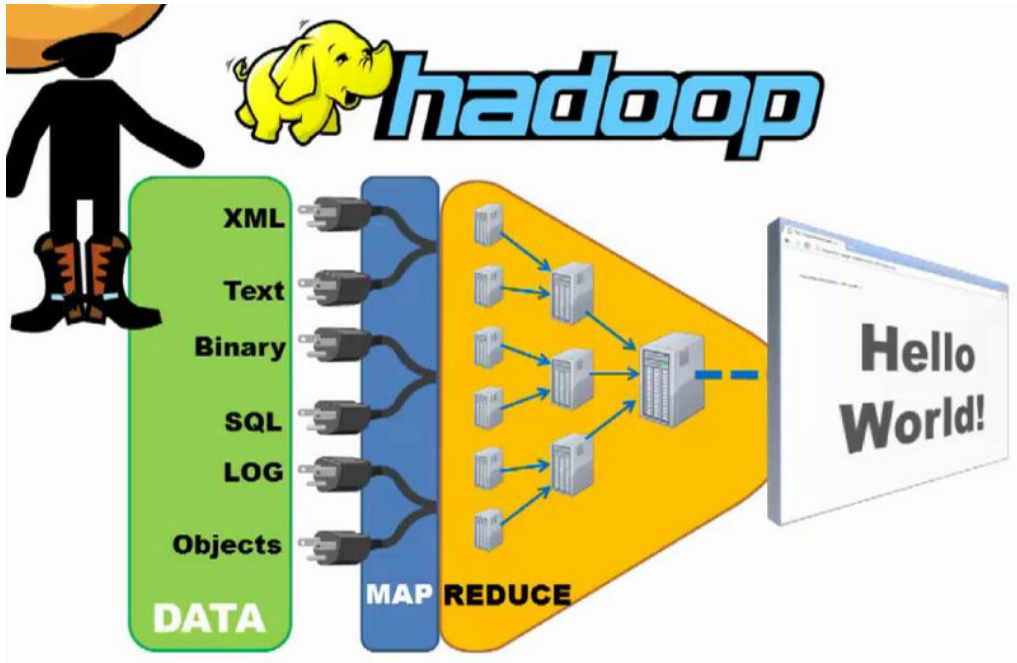
- Volume
  - The amount of data generated in seconds
- Velocity
  - Speed at which the data is generated, stored, and analyzed
- Variety
  - The different data formats including txt, email, photos, video, audio, PDFs, SMSes etc.

Big Data Collection Devices and Applications

- Computers
- Smart/wearable devices
- GPS/sensors
- Social media
- Retail systems
- Healthcare systems
- Government systems
- Networks

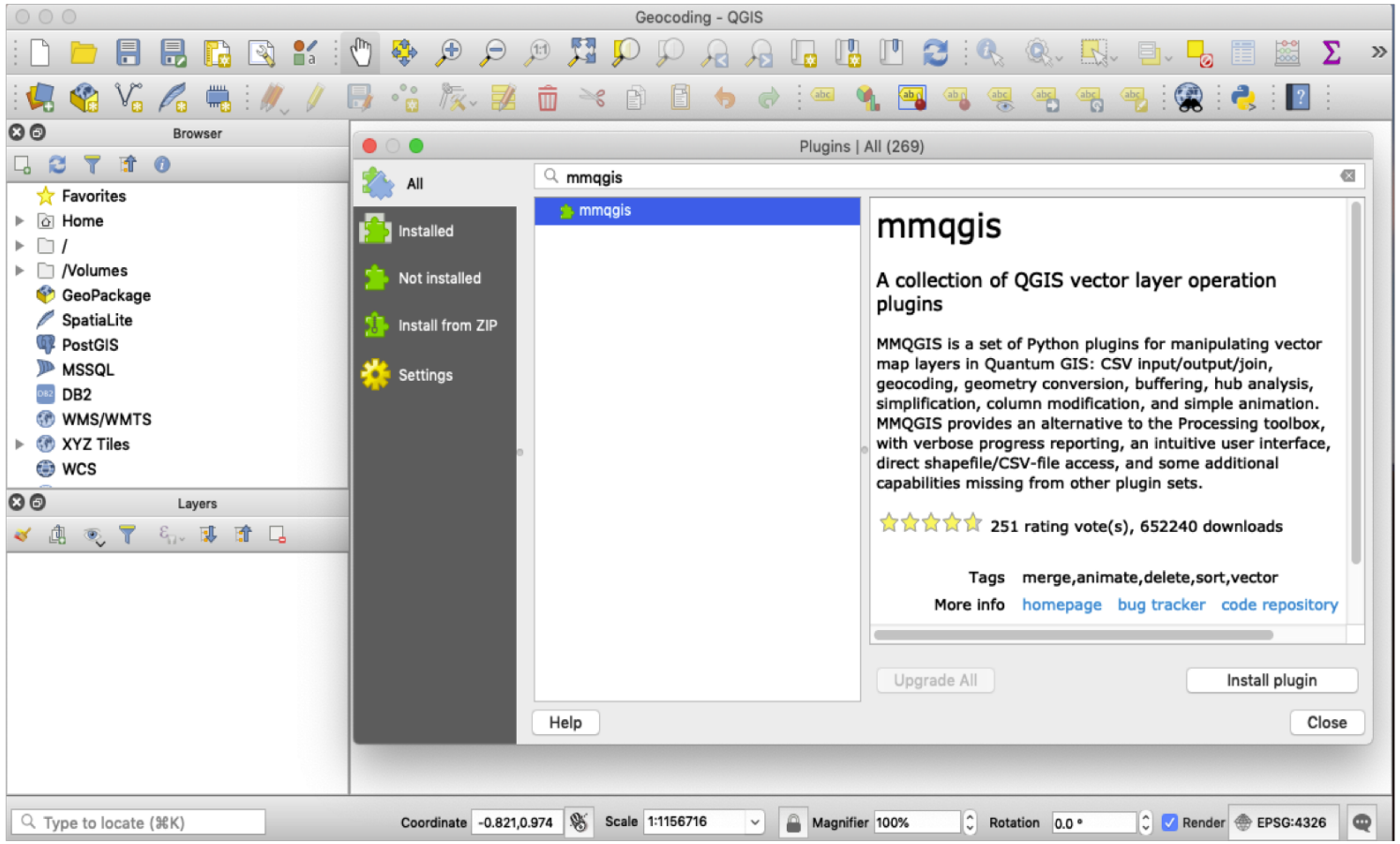
Big Data Analytics using Apache Hadoop

- Google and Facebook use Hadoop
- Hadoop open source software library
  - A framework that allows for distributed processing of large datasets across clusters of computers
  - **HDFS**: A distributed file system
  - **MapReduce**: A distributed data processing model and execution environment



Geocoding Datasets in QGIS using the MMQGIS Plugin

- Datasets containing place-name or address information can be geocoded to perform spatial analysis
- Kaggle.com provides CSV files of data science datasets



Demo: Geocoding a CSV of Addresses in QGIS using the MMQGIS Plugin

Geospatial Data Acquisition and Evaluation Summary

- Geospatial data collection can be expensive and time-consuming
  - Main techniques
    - Primary
      - Raster – e.g. remote sensing
      - Vector – e.g. GPS, field survey and LiDAR
    - Secondary
      - Raster – e.g. scanning
      - Vector – e.g. digitizing and geocoding
  - Conversion of existing data and online data options available
- Big Data
  - High volume, velocity and variety
  - Hadoop: HDFS and MapReduce
  - Spatial analysis is important in Big Data analytics

References

- Longley, P. A., M. F. Goodchild, D. J. Maguire, D. W. Rhind (2010). Geographic Information Systems & Science (3rd Ed). John Wiley & Sons, Inc.
- Clarke, K.C. (2010). Getting Started with GIS (5th Ed). Prentice-Hall, Inc., London.
- Chang, K-T. (2010). Introduction to Geographic Information Systems (5th Ed). McGraw-Hill.
- Center for Geographic Analysis (CGA) Harvard: GIS Training Materials.
- QGIS Demo: https://www.gislounge.com/how-to-geocode-addresses-using-qgis/
- Hadoop: https://www.youtube.com/watch?v=9s-vSeWej1U

Q&A