GIS Data Source and Quality

Dr. Tarendra Lakhankar

Today's Outline

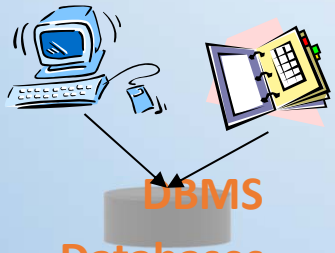
- GIS Data Sources
  - Data acquisition
  - Data digitizing
  - Data transformation
- Data Quality
  - Data accuracy and precision
  - Error measurement
- Research Project

Data Sources

Questions Before Getting Data

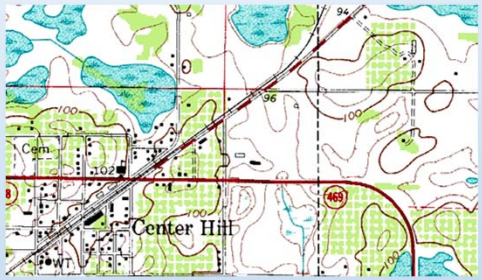
- Where can we get GIS data?
  - Government or commercial
  - Payment or not
  - Collection (new) or existing
- How can we get GIS data?
  - Format conversion
  - Quality/Accuracy
- What is the format?
  - Digital (raster or vector)
  - Non-digital (hardcopy)
  - Projection

Sources of GIS Data



DBMS

Databases

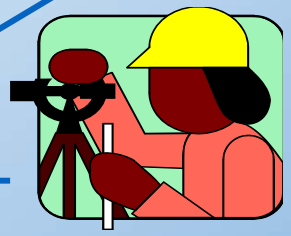


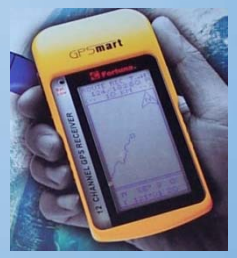
Digital Maps

Maps

Remote Sensing

Data Collection





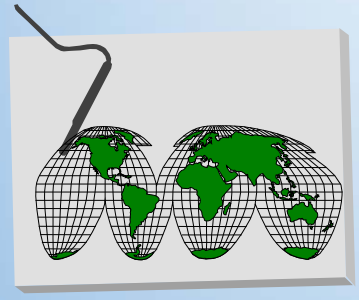
Land Surveys/GPS

Other (photographs, video)

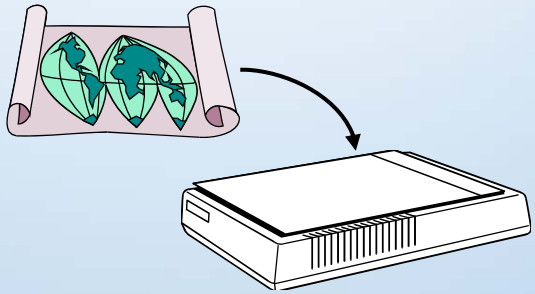
Manual Files

Data Collection is the MOST expensive and time consuming task

Five Main Methods



Manual Digitizing



Scanning

Existing Databases

Digital Photogrammetry

Field Measurements

Mapit GIS

Source: Digital or Hardcopy

- Hardcopy (Analog: A):
  - Maps
  - Survey data
  - Coordinate lists
  - Legal records
  - Aerial photos

- Digital (D)
  - Text files (coordinate lists: ID 1, 2, 3,…)
  - Structured files (coordinates *and attributes*)
  - Images of maps (GeoTIFF, world file,…)
  - COGO (directly measured survey data)
  - GPS (global positioning system)
  - Satellite imagery
  - Scanned aerial photos
  - Ready-made GIS data sets (net, CD-ROM,…):

Where do we get Data?

- Data Collection is the MOST expensive and time-consuming task
- Often, data collection is an end in itself. Almost invariably, it's the costliest element of any project-- > 80%.

- Purchased
- Found from existing sources in digital form
- Captured from analog maps

Finding Map & Digital Data Sources

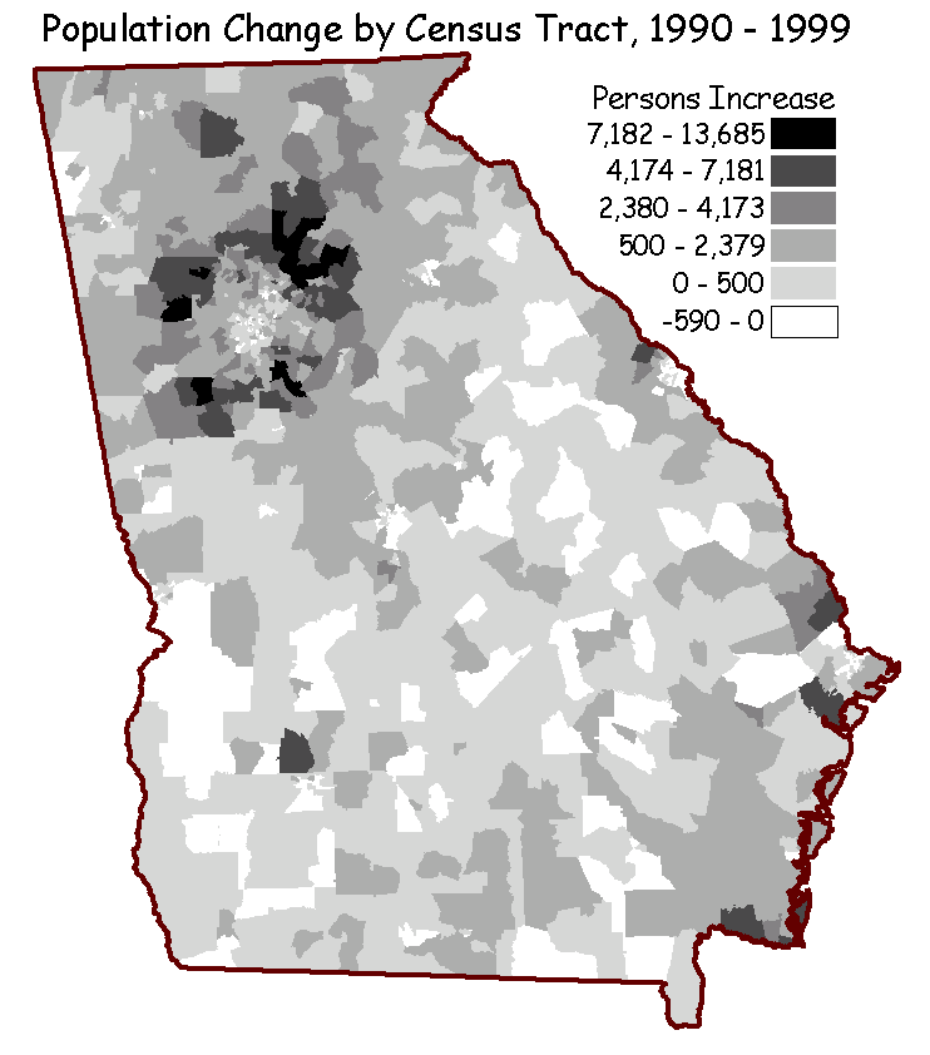
- Map libraries
- Reference books
- State and local agencies, esp. GIS departments/groups
- Federal agencies
- Commercial data suppliers

Finding Map & Digital Data Sources

- Via Web searches (e.g., Google.com)
  Many major data providers make their data available via the WWW
  - http://www.esri.com/software/arcgis/arcgisonline
    => nice basemaps
  - https://en.wikipedia.org/wiki/List_of_GIS_data_sources
    This is best resources listed GIS data websites.
  - https://gisgeography.com/best-free-gis-data-sources-raster-vector/
- GIS vendors package data with products.

Digital Census Data

- TIGER system:
  Topologically **I**ntegrated
  Geographic
  Encoding and **R**eferencing
- Typically, vector polygons with population, income, health, age, etc., attributes.



http://www.census.gov/

Census TIGER/Line files

- Topologically Integrated Geographic Encoding and Referencing files https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html
- The core TIGER/Line Files and Shapefiles do not include demographic data, but they do contain geographic entity codes (GEOIDs) that can be linked to the Census Bureau's demographic data, available on **data.census.gov**.
- TIGER Geodatabases are spatial extracts from the Census Bureau's MAF/TIGER database for use with Esri's ArcGIS. The geodatabases contain national coverage (for geographic boundaries or features) or state coverage (boundaries within state).

Census Bureau website

Interactive Maps

Work with interactive mapping tools from across the Census Bureau.

Census Business Builder

Census Data Mapper

Census Flows Mapper

Data Visualization Gallery

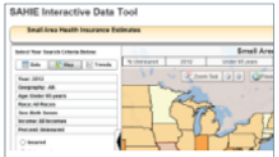
Metro/Micro Thematic Map Viewer

OnTheMap (Employment)

OnTheMap for Emergency Management

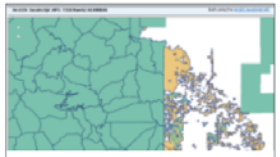
Response Outreach Area Mapper (ROAM)

Rural America: A Story Map



SAHIE Interactive Data Tool (Health Insurance)

SAIPE Interactive Data Tool (Poverty)



TIGERweb

Federal Data Agencies

- USGS: http://www.usgs.gov/ngpo/
- USGS: http://ned.usgs.gov
- USGS: http://water.usgs.gov/maps.html
- USGS: http://earthexplorer.usgs.gov
- USDA: http://datagateway.nrcs.usda.gov/
- Geodata one stop: http://www.data.gov
- NOAA: http://www.climate.gov/maps-data
- EPA: https://www.epa.gov/enviro/data-downloads
- EPA: http://www.epa.gov/geospatial/
- Census Bureau: http://www.census.gov

Federal Data Agencies

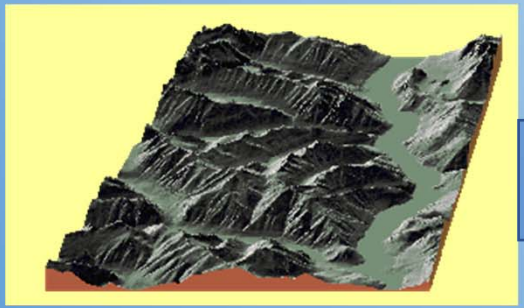
- https://www.ncei.noaa.gov/ National Centers Environ Information
- http://ngdc.noaa.gov National Geophysical Data Center
- http://www.nesdis.noaa.gov satellite and information service
- http://nodc.noaa.gov National Oceanographic Data Center
- http://ncep.noaa.gov National center environmental prediction
- http://earthdata.nasa.gov EOSDIS NASA DAAC centers
- https://highways.dot.gov/
- http://www.gis.fhwa.dot.gov/statepracs.asp Federal Highway Administration
- http://www.cdc.gov/gis Centers for Disease Control
- https://www.mapsofworld.com/ GIS Maps of World

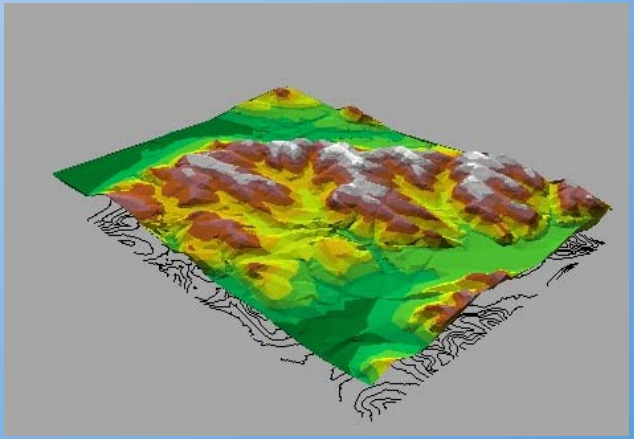
Digital Elevation Models (DEM)

- Elevation data in a raster format
- Acquisition:
  - Traditionally, ground and aerial surveys were used to collect elevation data points
  - Modern techniques:
    - derivation from manual or automatic digitization of mapped contour lines
    - Softcopy photogrammetric methods using digital or digitized aerial photography
    - LIDAR
  - Next, space-based radar

The USGS produces different digital elevation products: **3DEP** 3D Elevation Program

- **1/3-Arc-sec DEM** 10- x 10-meter data spacing
- **1 Arc-Sec DEM** 30- x 30-meter data spacing
- **2 Arc-Sec DEM** 60- x 60-meter data spacing
- **1/9 Arc-sec DEM** 3- x 3-meter data spacing
- **1 meter** 1 x 1-meter data spacing



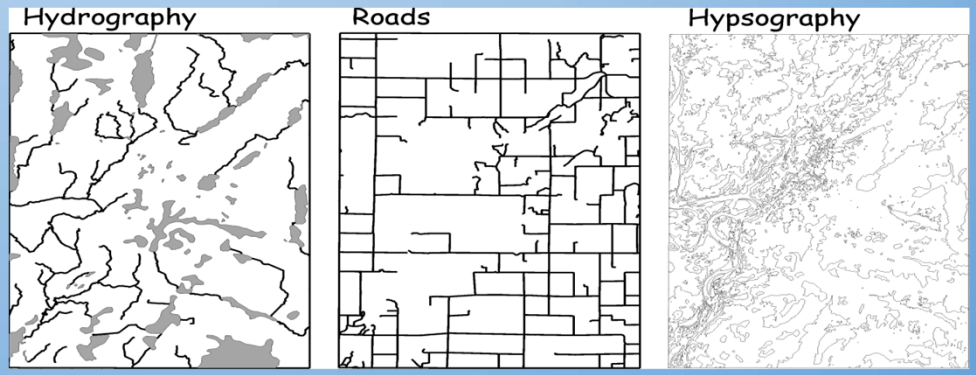


USGS File Types

- The USGS has been a driving force in GIS. Some of the USGS formats date back to the 1980s just when Geographic Information System was emerging from its roots.
- These three USGS file types are:
  - Digital Line Graph (DLG)
  - Digital Raster Graphic (DRG)
  - Digital Orthophotos Quadrangle (DOQ)

Digital Line Graphs (DLGs)

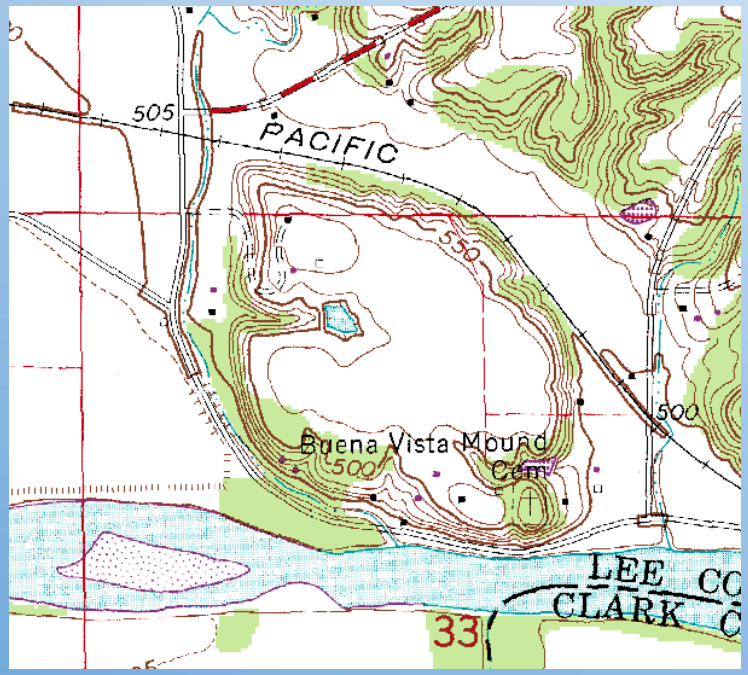
- Vector representations of most features portrayed on a USGS national series map
- Provided as separate themes (9 typically)
  - transportation data
    - (roads, railroads, pipelines, power lines, and airports)
  - hydrography
    - (water features).
  - hypsography
    - elevation contour information
- Three major groups:
  - 1:2,000,000-scale
  - 1:100,000-scale
  - 1:24,000-scale



https://lta.cr.usgs.gov/DLGs

Digital Raster Graphic (DRG)

- A georeferenced raster image of a USGS Map, created by scanning the paper maps (ex: map in georeferencing lab)
- Topographic map georeferenced to the Universal Transverse Mercator projection.

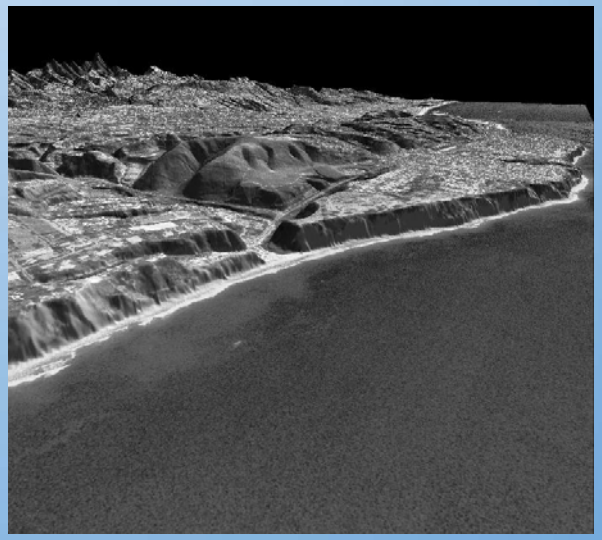


USGS 1:24,000 DRG

Digital Orthophoto Quadrangles (DOQs)

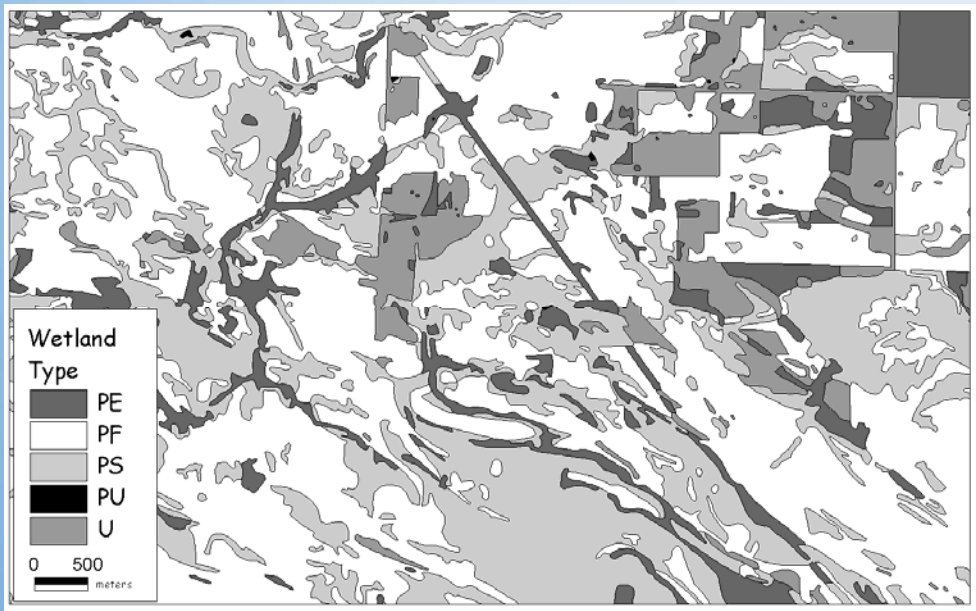
- Aerial photographs
- Scanned
- Corrected for distortions (camera, terrain etc.) to be orthographic
- MMU about 1 m; resolution = 1m/pixel
- Usually false-color IR
- Frequently used for updating land cover, GCP location, spatial changes etc.
- Flown approx. every 2 years
- Lead agency: USGS

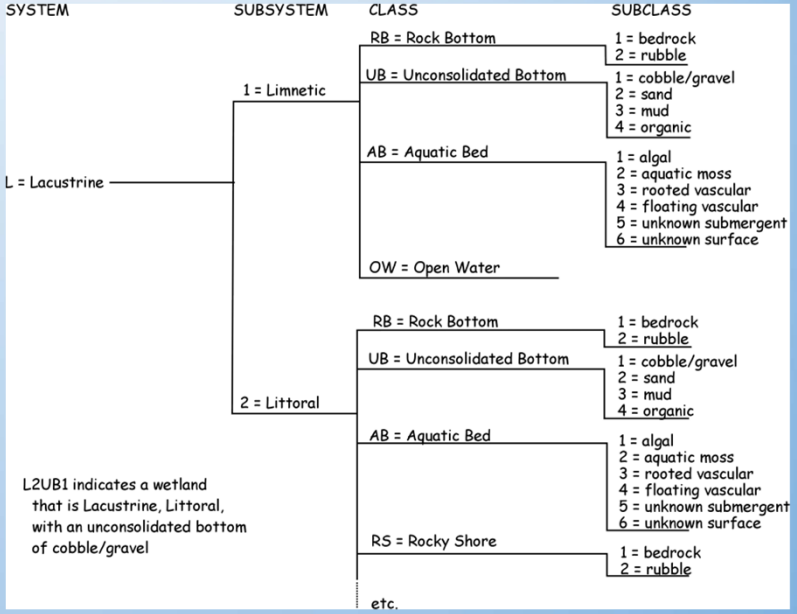
MMU stands for Minimum Mapping Unit. This is the smallest resolvable unit within the data.



U.S. Fish and Wildlife Service (USFWS)

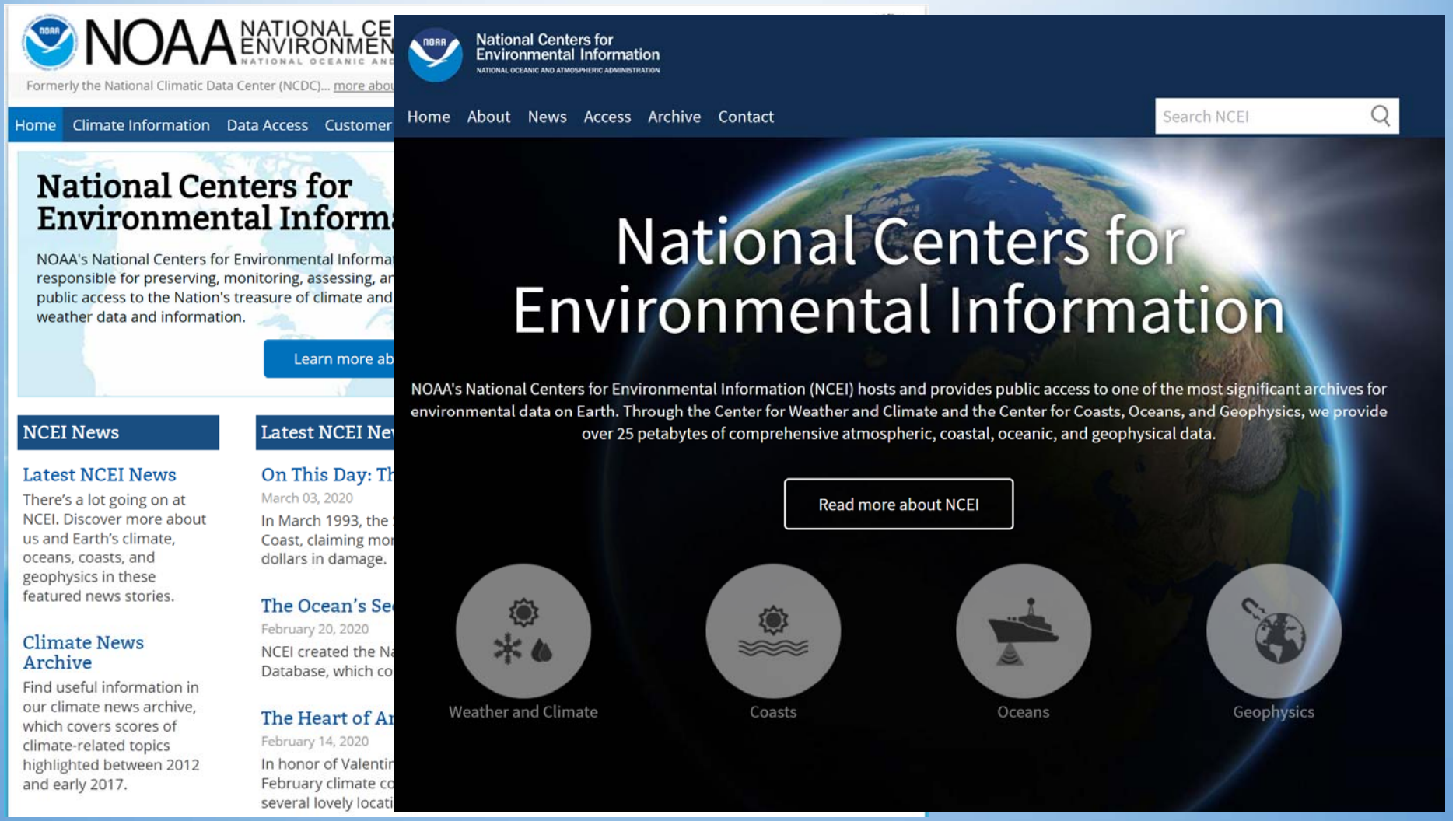
https://www.fws.gov/GIS/





- Data on location and condition of wetlands in the USA
- Produced by US Fish and Wildlife Service
- Decadal updates planned, started in 1970s, not all digital
- Combination of aerial photo interpretation (spring photos) and field visits

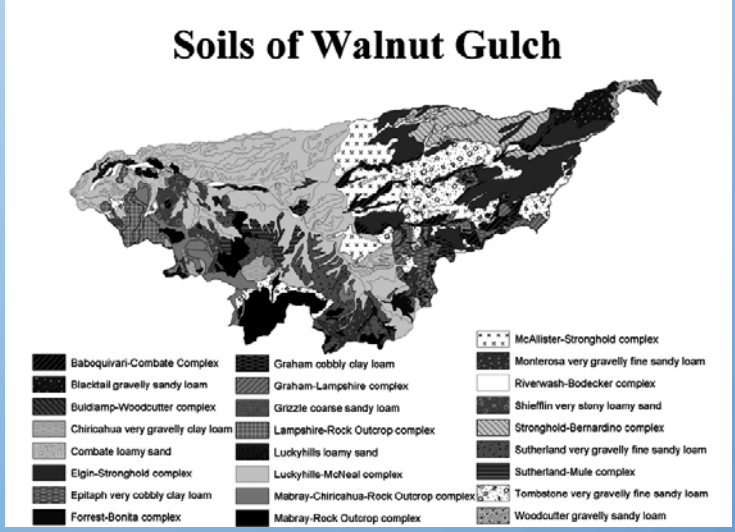
National Centers for Environmental Information



http://www.ncdc.noaa.gov/

Digital Soils Data

- The USDA Natural Resource Conservation Service (NRCS) digital soils datasets: http://datagateway.nrcs.usda.gov/
  - NATSGO – National soil geography; highly generalized
  - STATSGO – State soil geographic data; intermediate scale and resolution, is less detailed map.
  - SSURGO – Soil survey geographic data; most spatial and categorical detail, county level data.
- 1:12,000 to 1:7,500,000-Scale Digital Soils Information.

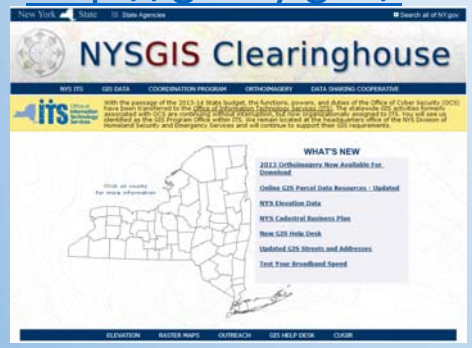


National Land Cover Database

- Earth Resources Observation and Science (EROS) Center is central to the creation of the National Land Cover Database (NLCD), which is updated every five years and stands as the definitive land cover database for the United States.
- NLCD is generated in cooperation with Federal agencies working together to produce current, nationally consistent, land cover products for all 50 states and Puerto Rico.
- NLCD is used for thousands of applications that require information on landcover location or where it is changing. The latest iteration is NCLD 2016.

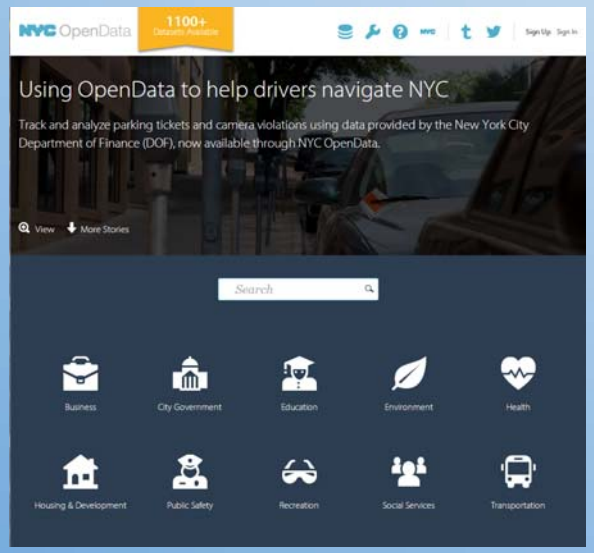
Statewide data also available: New York

http://gis.ny.gov/



http://cugir.mannlib.cornell.edu/

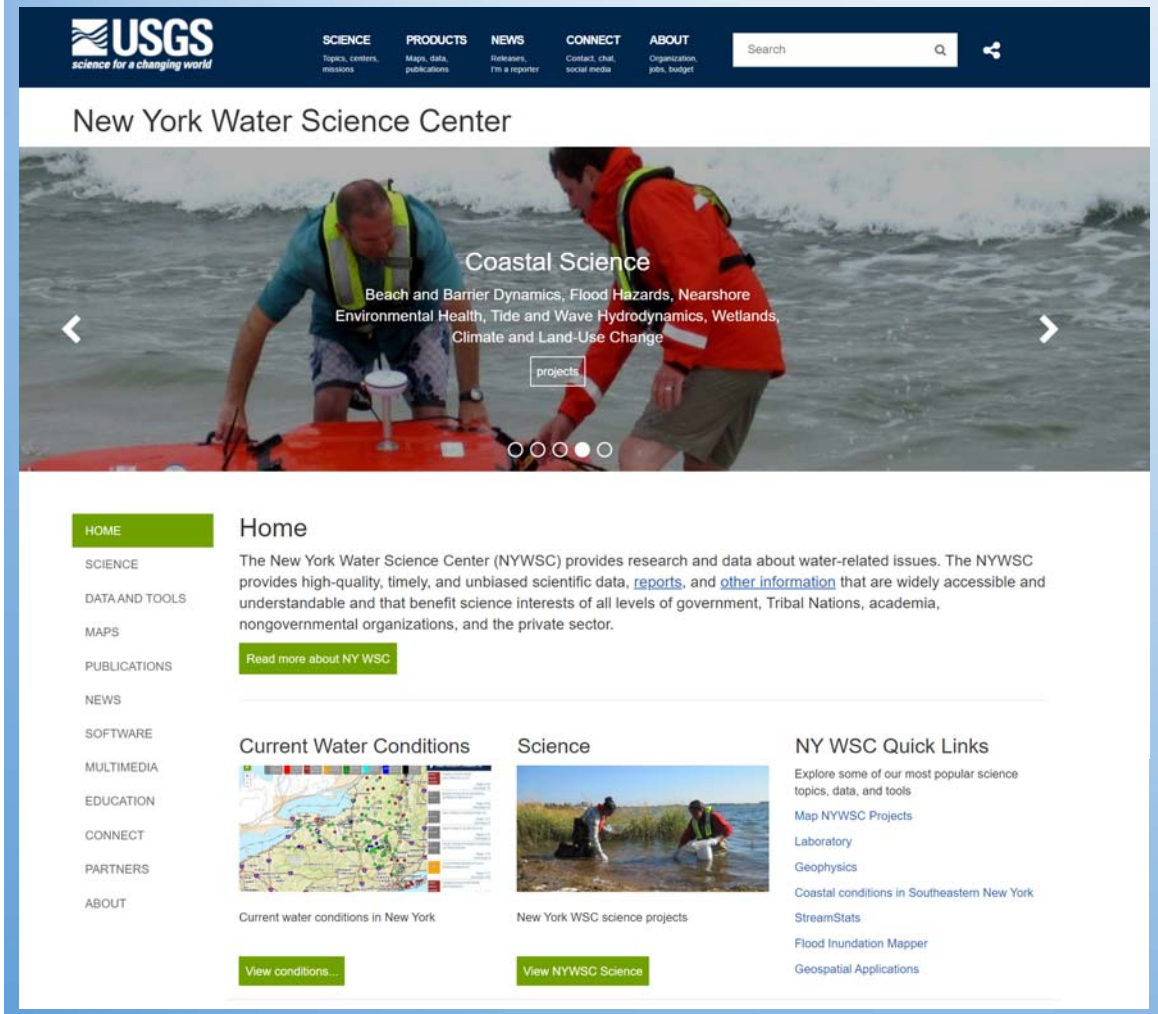
https://data.cityofnewyork.us /

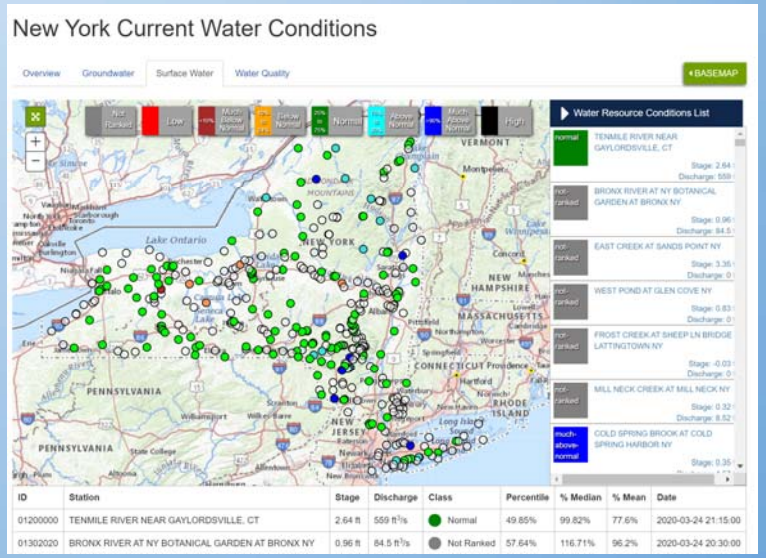


http://www.nan.usace.army.mil/BusinessWithUs/GeospatialDataSystemsActivities.aspx

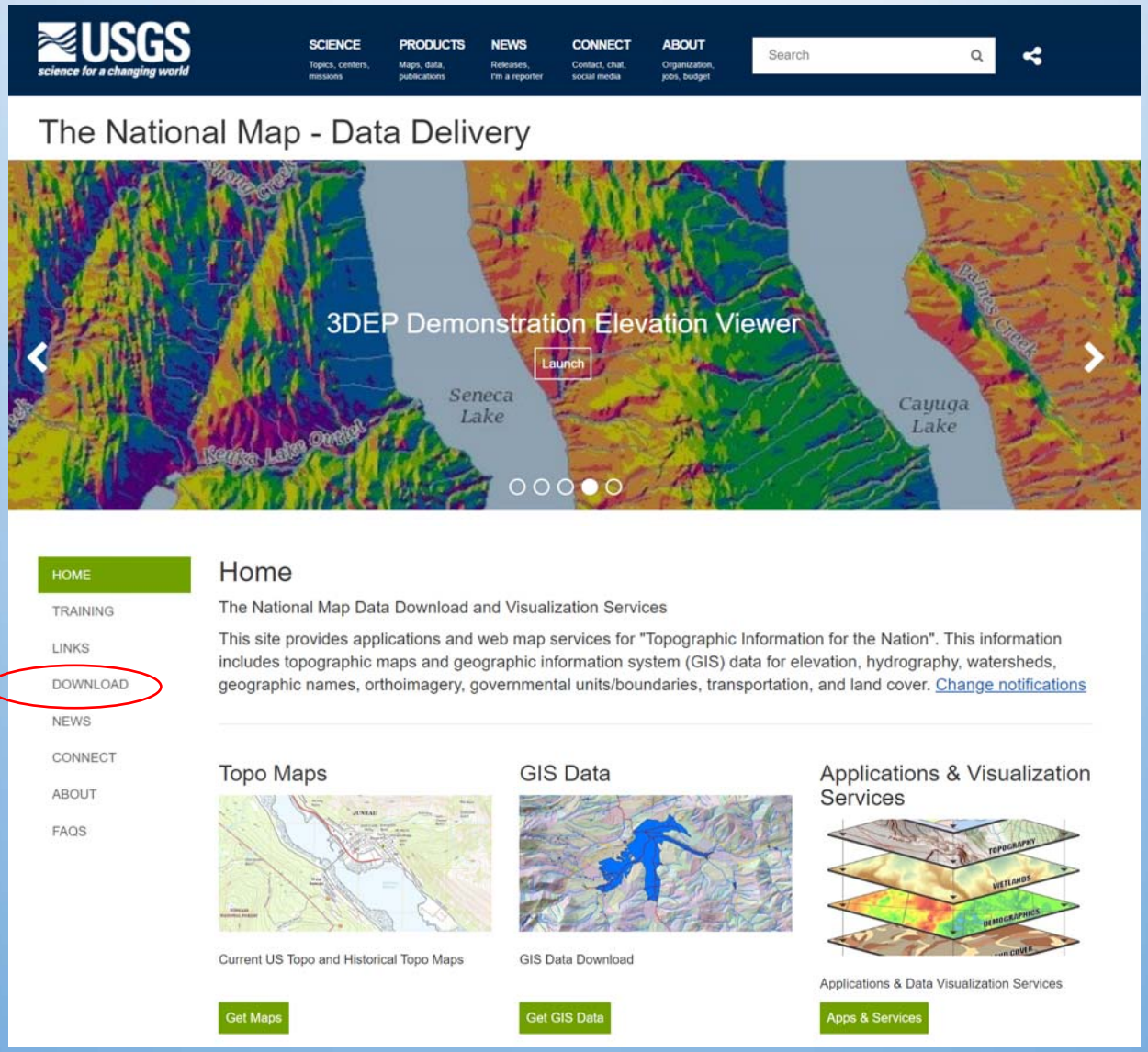
USGS New York Water Science Center

https://www.usgs.gov/centers/ny-water

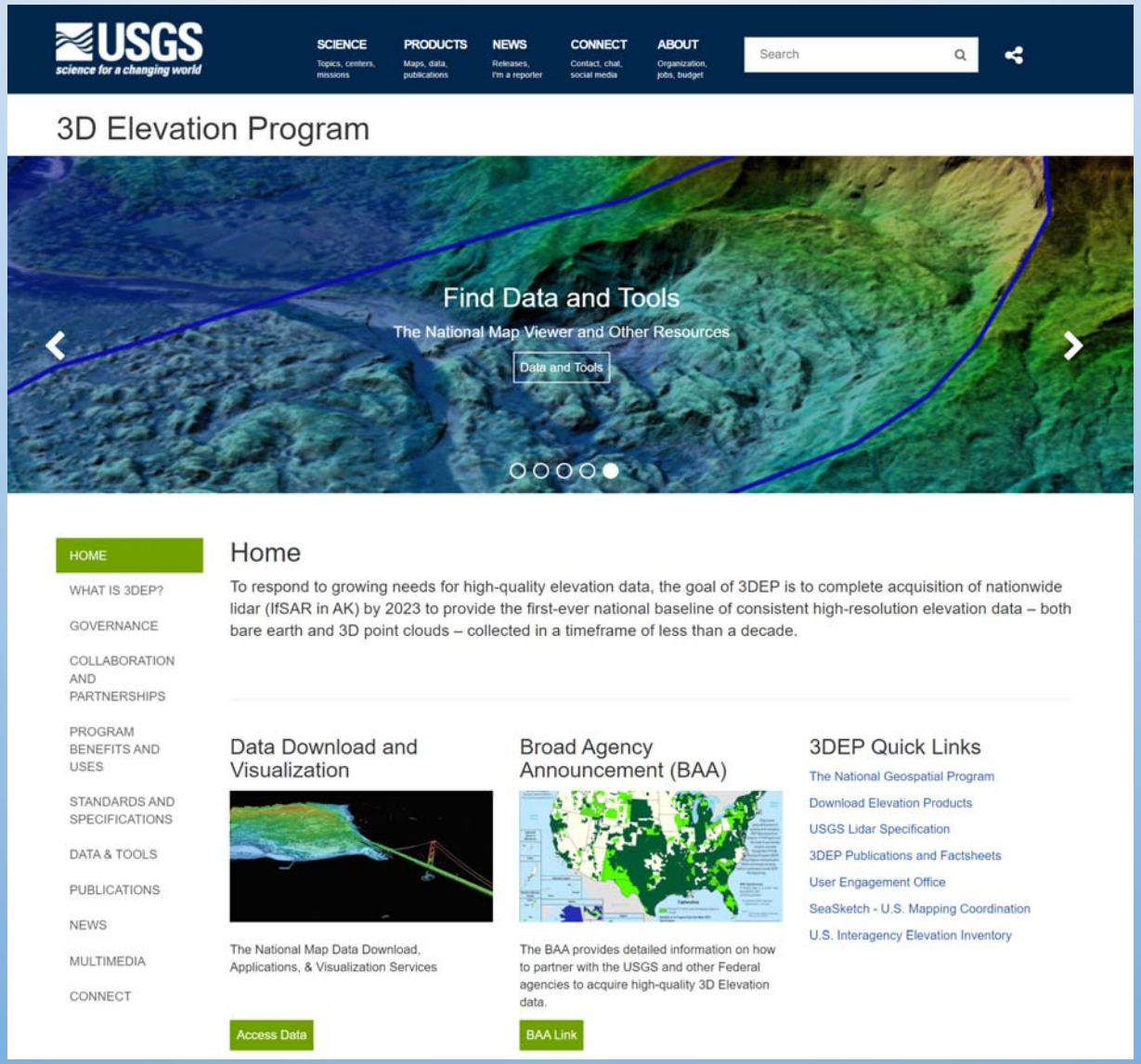




National Map Data



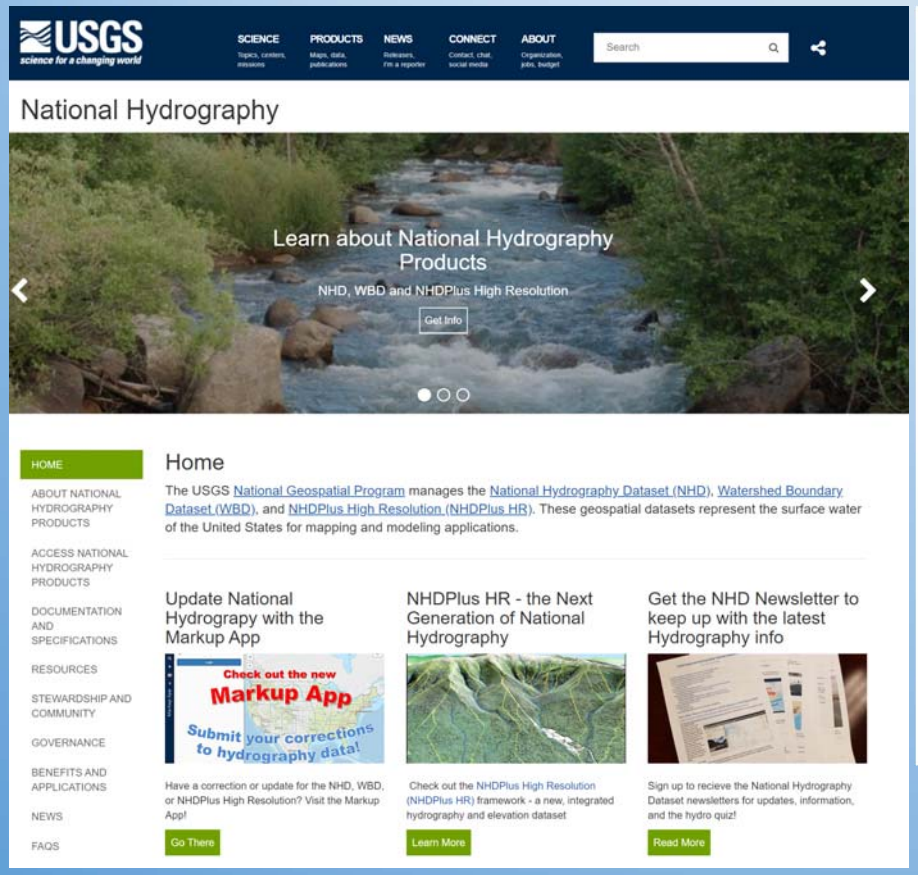
National Elevation Data (NED)

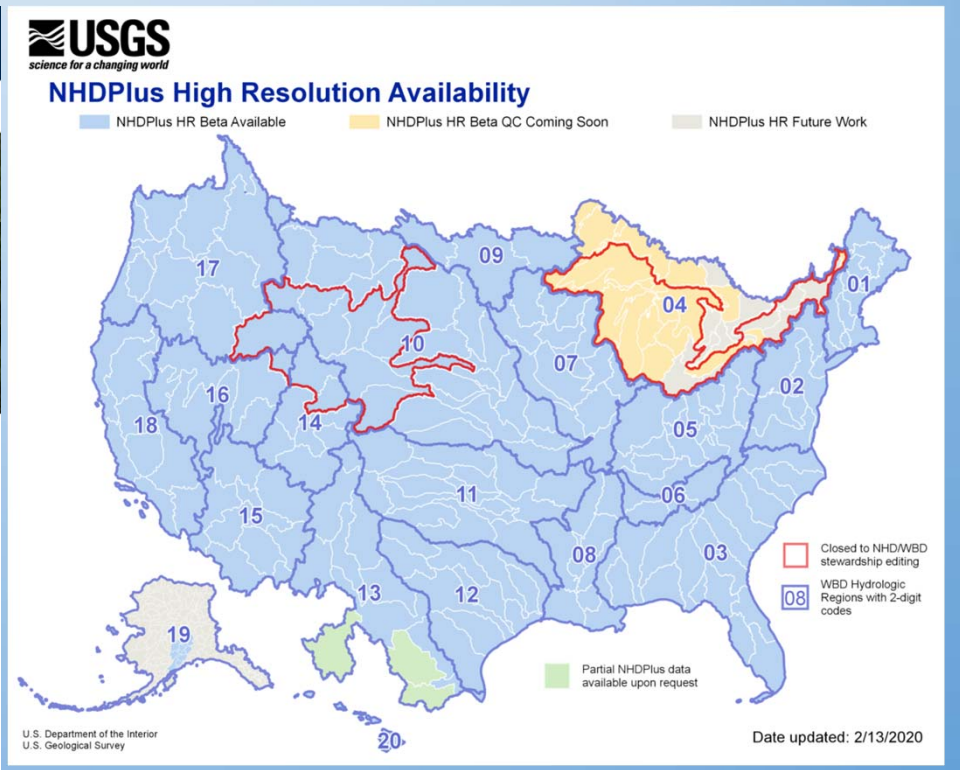


National Hydrography Dataset

- The National Hydrography Dataset (NHD) represents the water drainage network of the United States with features such as rivers, streams, canals, lakes, ponds, coastline, dams, and stream-gages.

https://www.usgs.gov/core-science-systems/ngp/national-hydrography

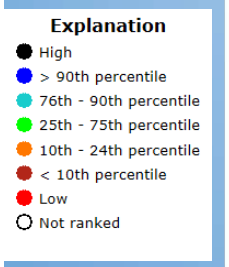


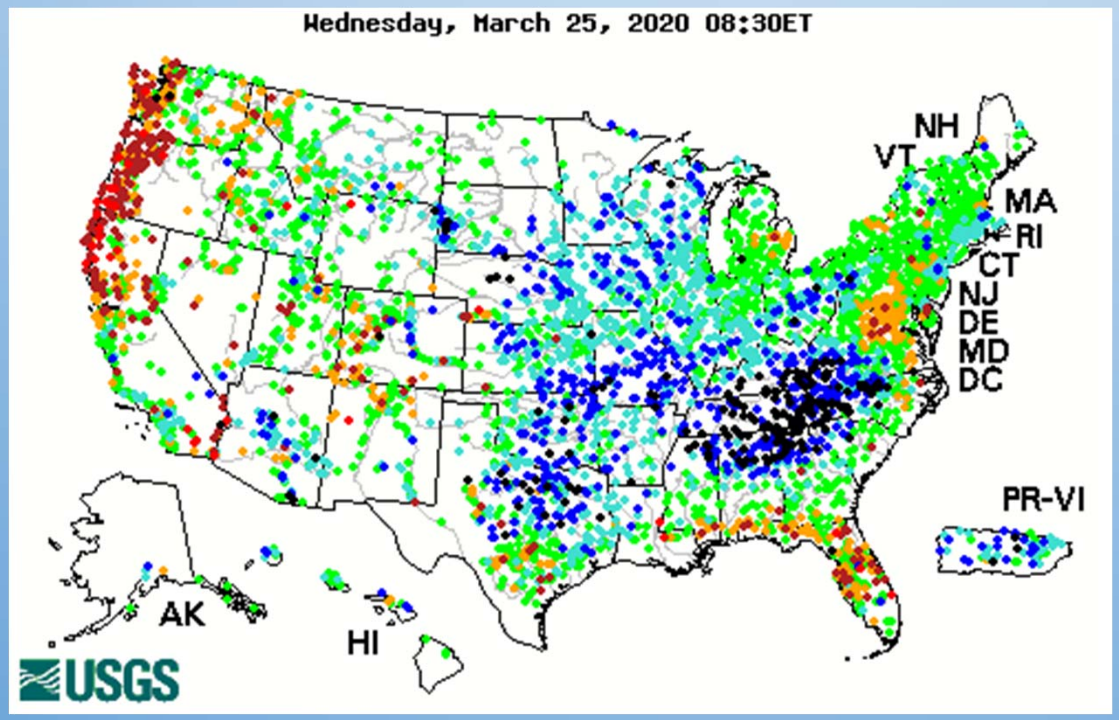


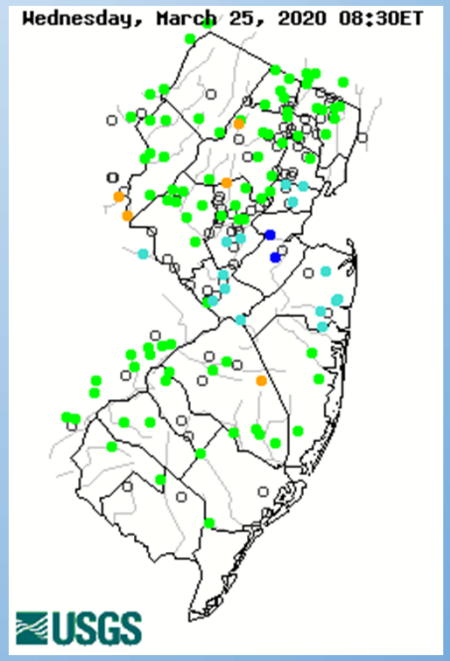
USGS national water datasets

- Surface-water data, such as gage height (stage) and streamflow (discharge), are collected at major rivers, lakes, and reservoirs.
- Groundwater data, such as water level, are collected at wells and springs.

https://waterdata.usgs.gov/nwis







New Jersey Streamflow Table

New Jersey Lake and Reservoir Table

New Jersey Groundwater Table

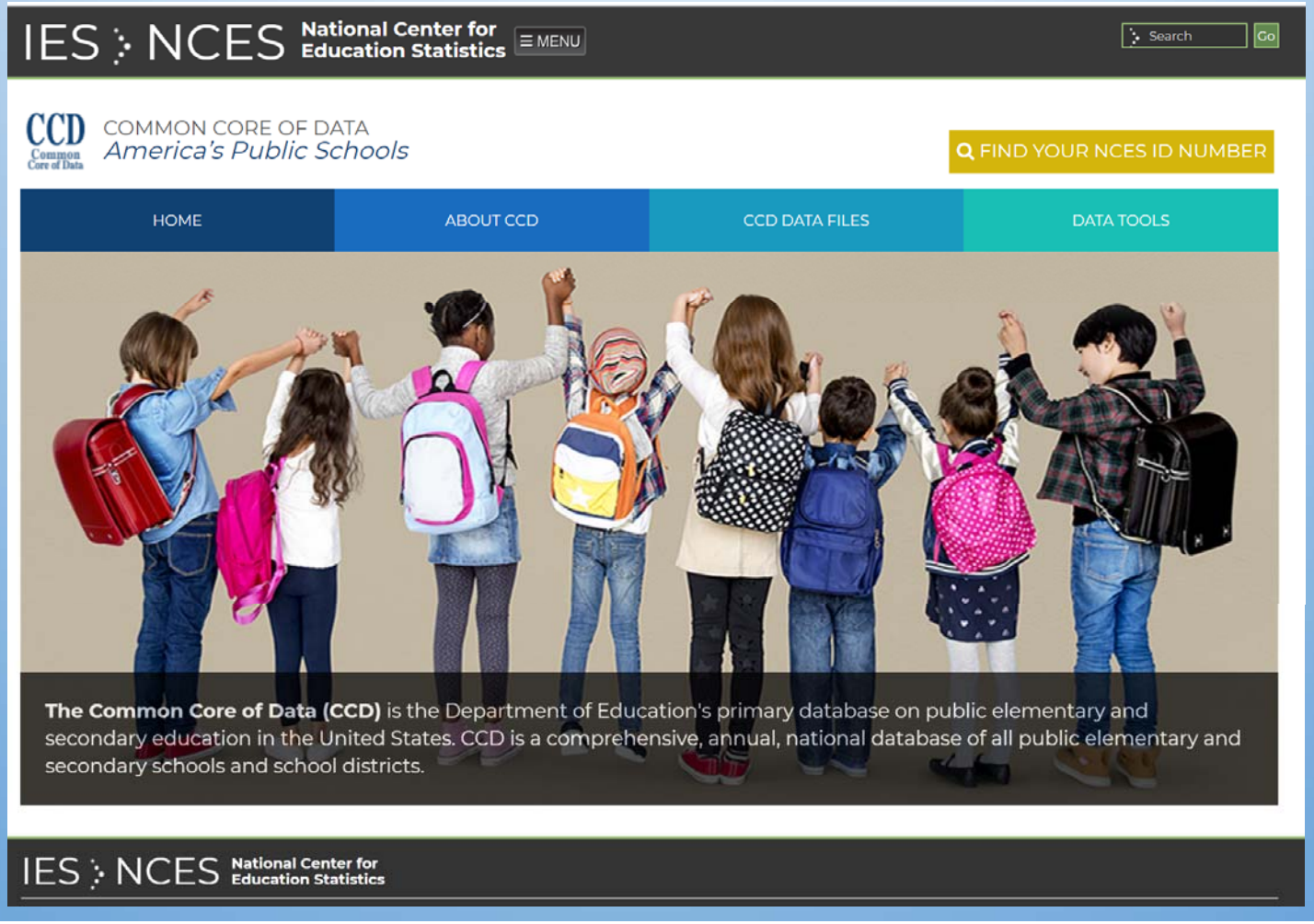
New Jersey Water-Quality Table

New Jersey Weather Table

New Jersey Tide Table

National Center for Education Statistics

- The National Center for Education Statistics (NCES) is the primary federal entity for collecting and analyzing data related to education. https://nces.ed.gov/ccd/



National Historical Geographic Information System (NHGIS)

- The National Historical Geographic Information System (NHGIS) provides easy access to summary tables and time series of population, housing, agriculture, and economic data, along with GIS-compatible boundary files, for years from 1790 through the present and for all levels of U.S. census geography, including states, counties, tracts, and blocks.

Two Kinds of Data Conversions

- Paper Maps to Digital (A➔D)
  - Generally the most complex & expensive
  - Digitizing, "heads-up" on screen or Scanning
  - Automated extraction of layers

ArcScan Extension

- Digital to Digital Conversions (D➔D)
  - Translation between different vendor's GIS formats
  - Spreadsheet software (e.g. Excel) to GIS database
  - Specialized conversion packages for converting between different databases
  - Efforts at standardization, which reduces need for conversions, have had limited success because of competitive pressures
    - FGDC's, Spatial Data Transfer Standard (SDTS), is a federal standard
    - Open GIS Consortium, a vendor and user group, lobbies for standards and non-proprietary approaches to GIS database creation

Data Interoperability Extension

Examples: A->D / D-> D Conversion

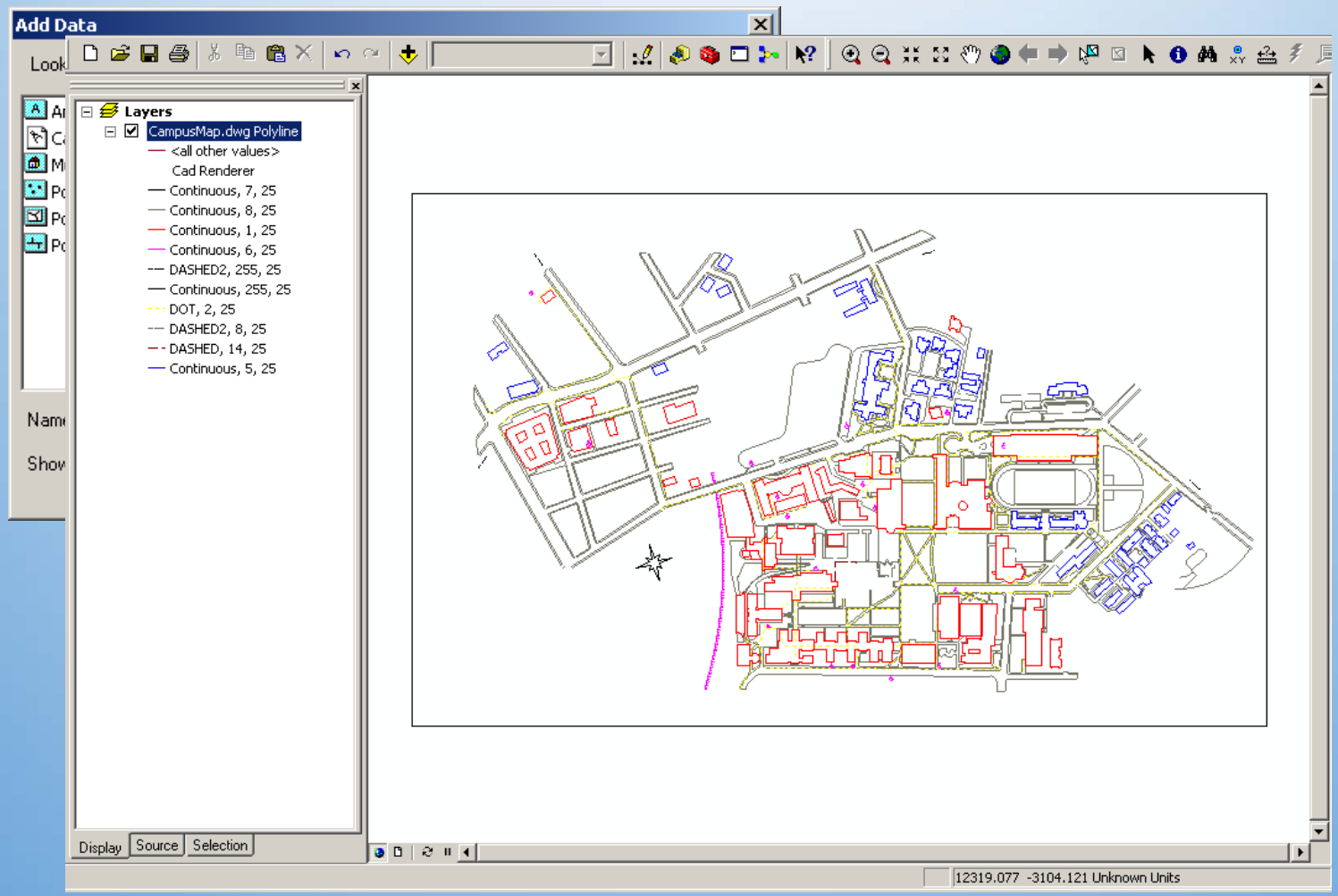
- **Maps and Drawings**
  - digitizing, or
  - scanning then raster to vector conv
- **Aerial Photographs**
  - photogrammetry/photo interpretation to extract features
  - digitizing or scanning to convert to digital
  - rectification and DTM (digital terrain model) to create digital orthos
- **Satellite Imagery**
  - rectification and DTM to create digital orthos (if desired)
- **Records and Documents**
  - OCR (optical char recognition) scanning
  - keyboarding
  - then, same as attribute data bases
- **GIS Data Bases**
  - conversion between proprietary standards (ARC/INFO, Intergraph, AutoCAD, etc.)
  - Spatial Data Transfer Standard
- **Attribute Databases**
  - geocoding
  - conversion between geographic units (e.g. zip codes and census tracts)
  - conversion between different databases
- **CAD Data Bases**
  - translator software (pre-existing or custom-written) needed to convert to required GIS format

Vector data formats

Computer-aided design (CAD) drawings

- CAD software
  - Autodesk, AutoCAD (.dwg)
  - Bentley, Microstation (.dgn, .dxf)
- Often used by engineering companies
- Better digitizing precision

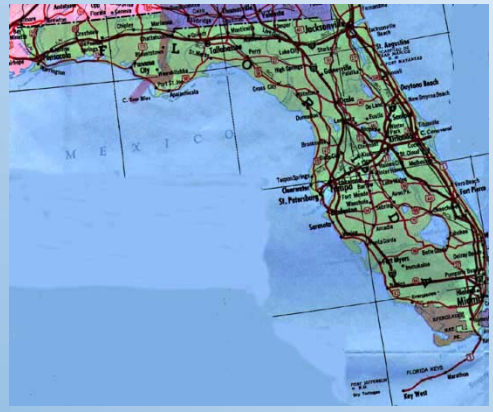
CAD drawings



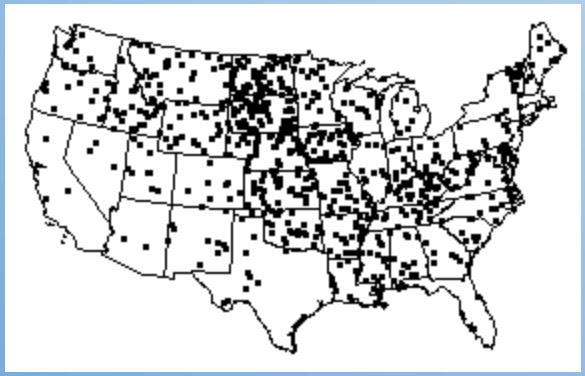
Hard Copy: Issues to Consider and Account for

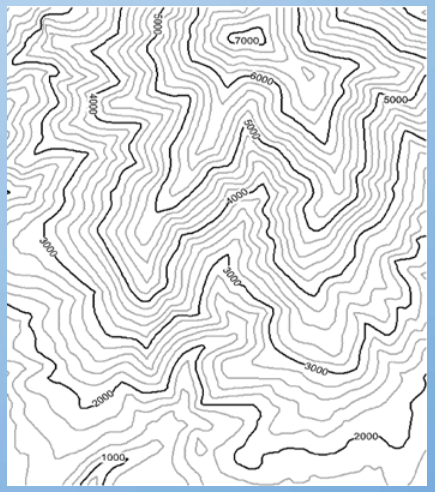
- Map Type
- Map Generalization
- Map Media
- Map Boundaries
- Map Scale

Issues to Consider: Map Type



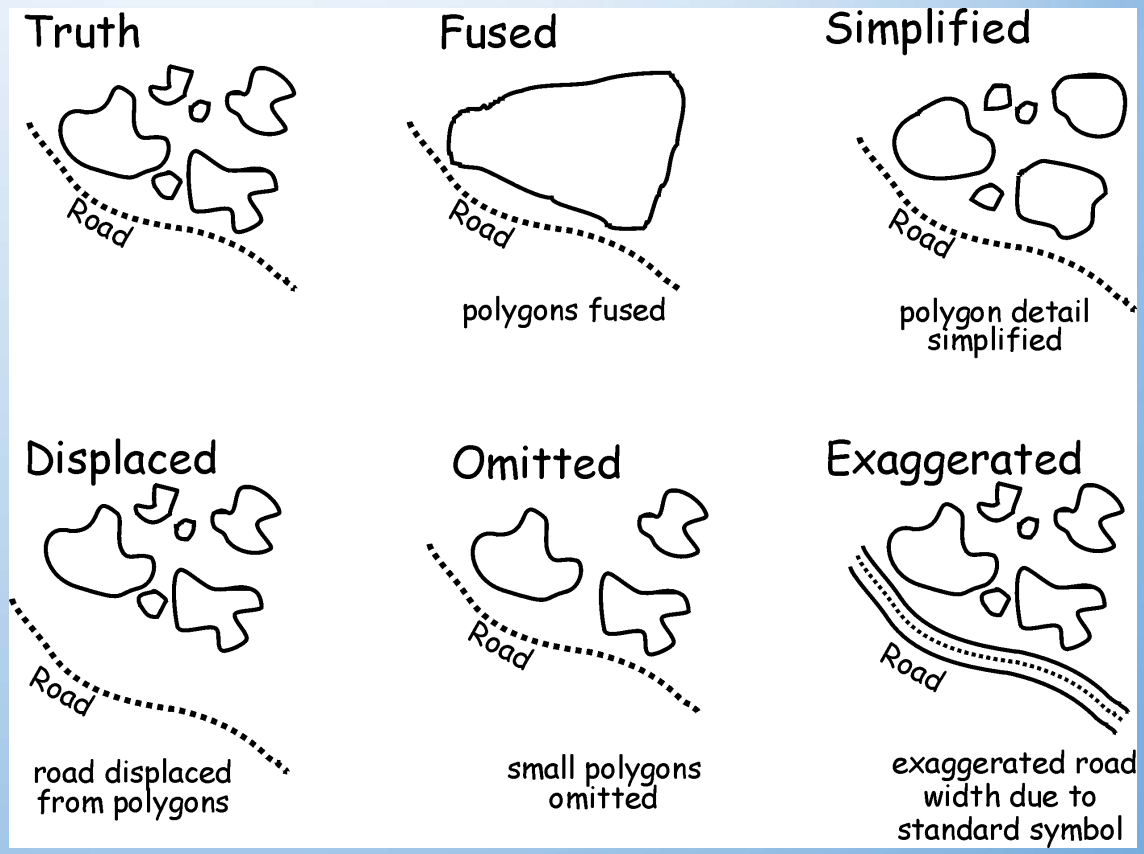
Feature map
Choropleth map
Dot density map
Contour map





Issues to Consider: Map Generalization

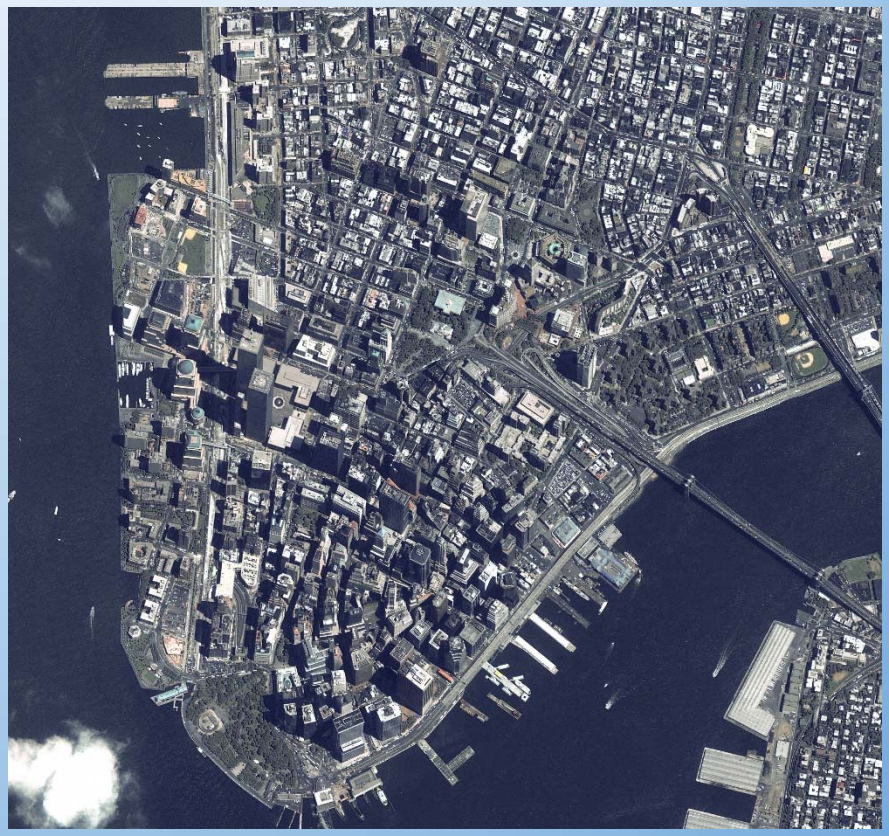
- Reality is abstracted
  - Fusion
  - Simplification
  - Displacement
  - Omission
  - Exaggeration



Driven by:

- Resources (time, $) for original fieldwork & cartography
- Scale/resolution possible or needed by original cartographers

Map generalization



Issues to Consider (continued):

- Map media
  - Shrinking, stretching, folding, etc.
- Map boundaries
  - Matching of boundaries across maps is frequently problematic → "edge matching" is a common technique for the correction of this
- Map scale
  - Affects the minimum mapping unit and the systematic level of map generalization.
  - Affects positional accuracy

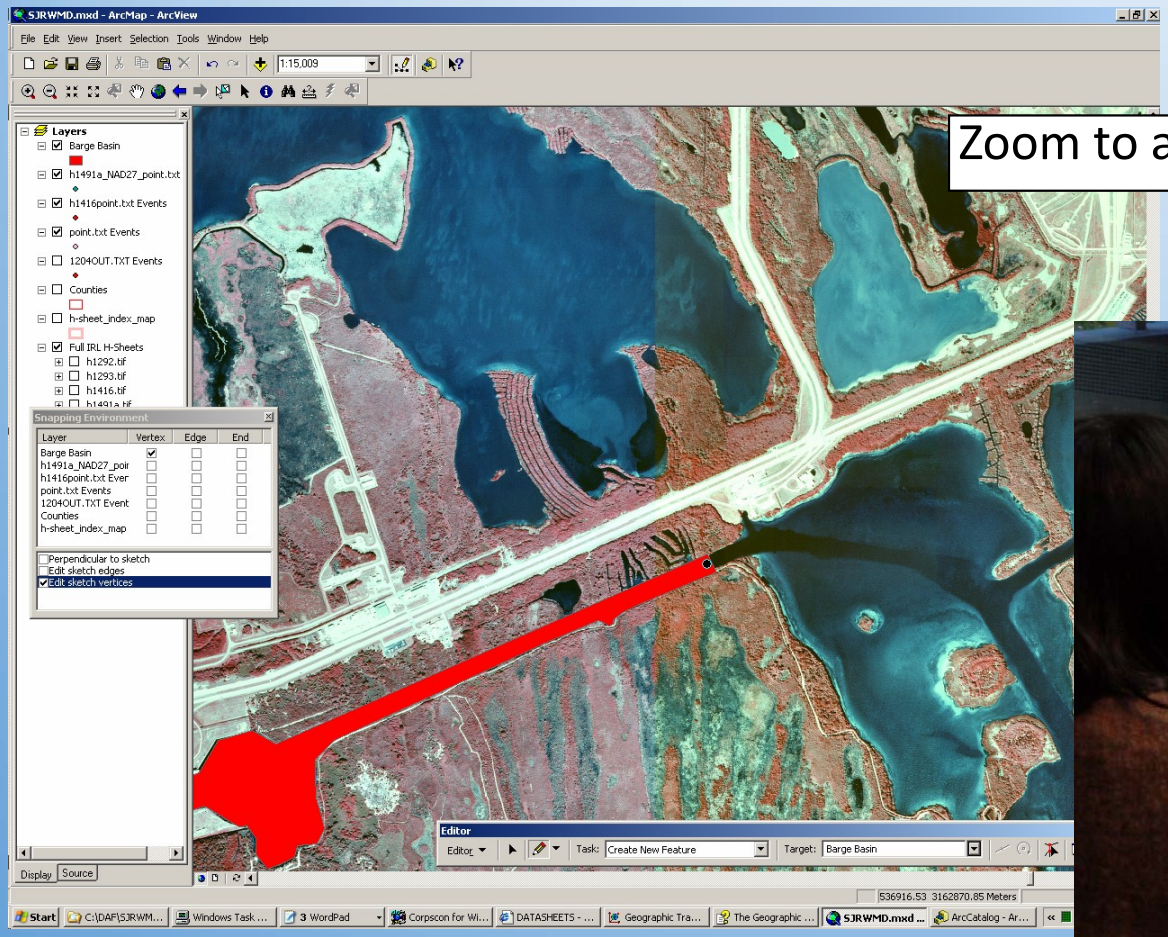
Hardcopy Sources: Digitizing

- Definition: The process of converting map, image or other analog data sources to a digital format.
- Types:
  - Manual digitization
    - Typically for vector data
  - Scan digitization
    - Typically for raster data and images
  - On-screen (head's up) digitization
    - Typically for vector data

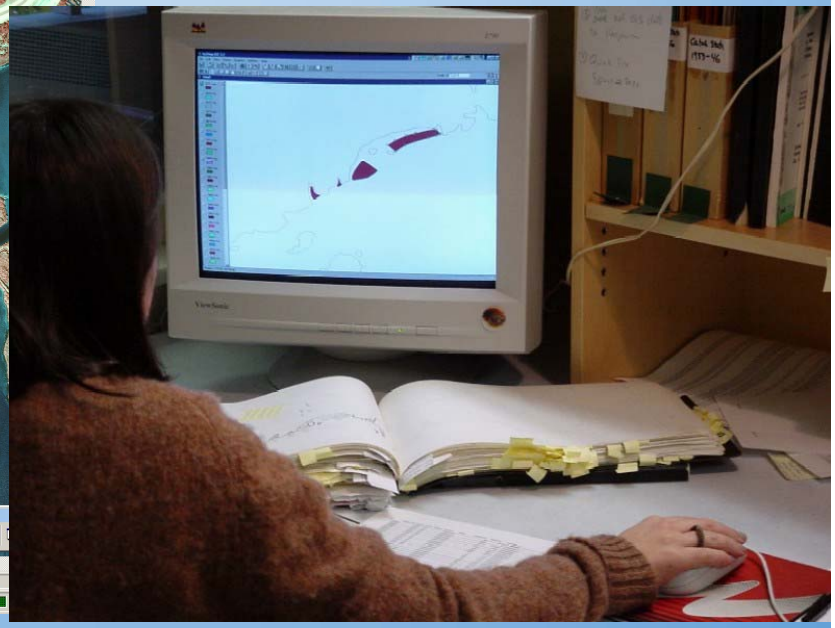
Manual Digitizing

- Point mode
- Stream mode (defined by time or distance)

Heads-up Digitizing

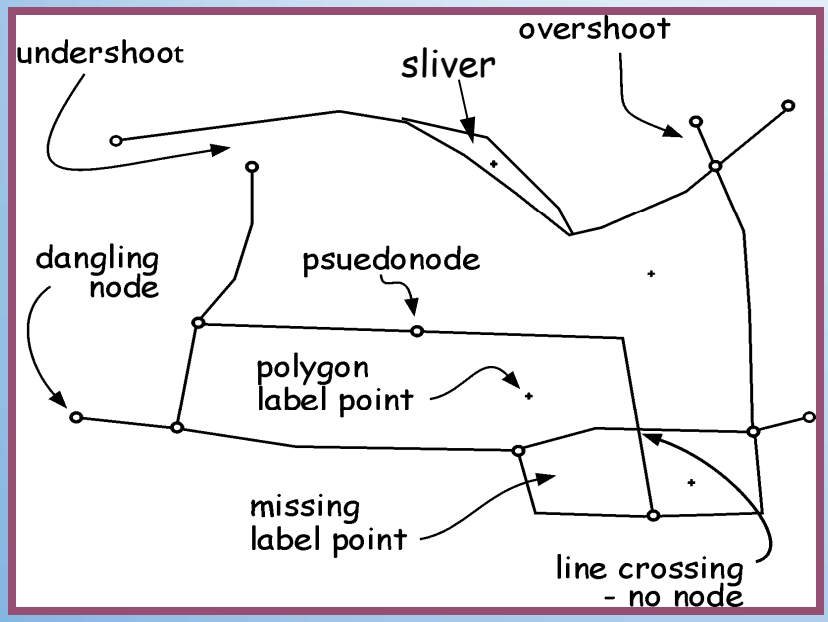


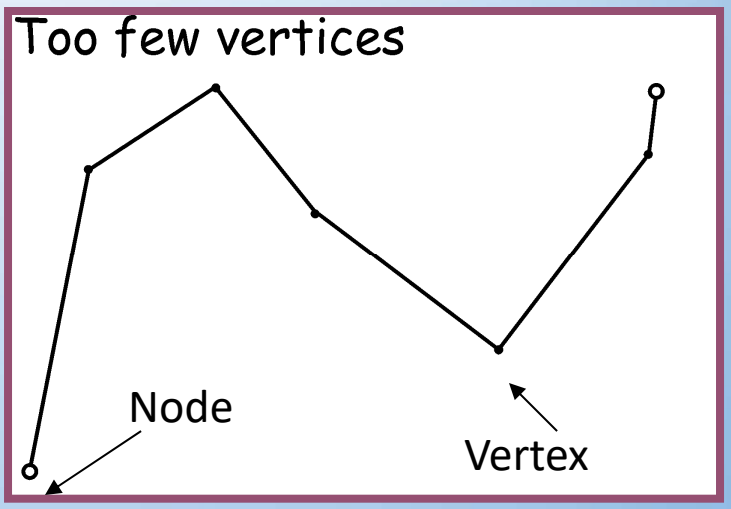
Zoom to appropriate scale then …

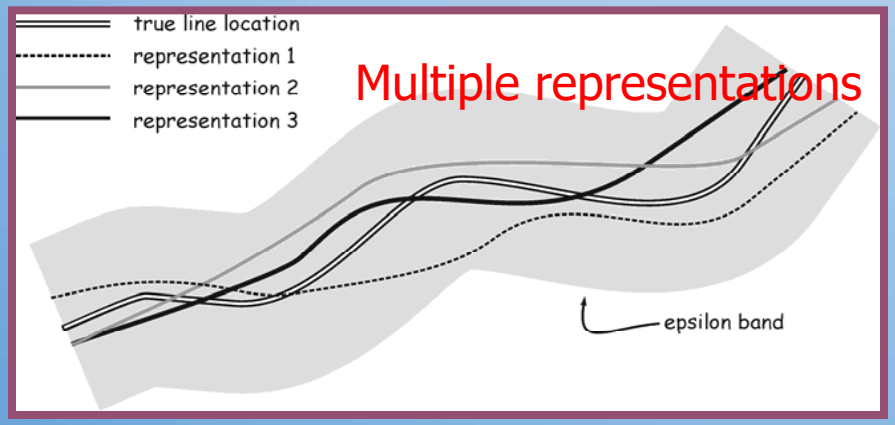


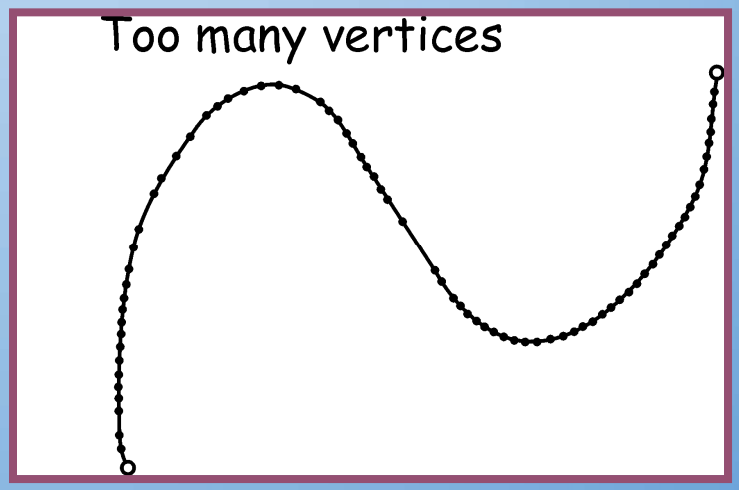
Data Quality and Error

Common Manual Digitizing Errors

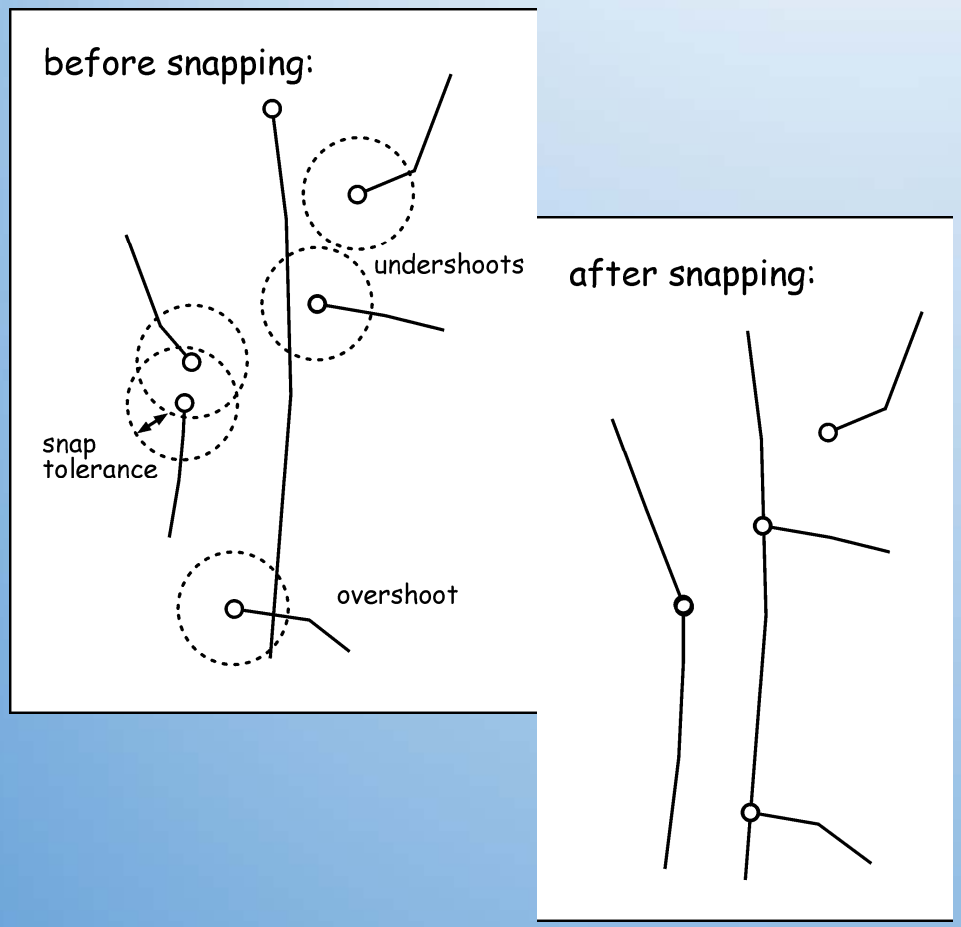








Node and Line Snapping
-overcome undershoot/overshoot problem



Snapping
- Nodes
- Lines

Overcome undershoot/overshoot problem

Line Reshaping

Overcome under-sampling/over-sampling problem

➔ **Overcome undersampling/oversampling problem**

- Smoothing
  - Splines
  - Other polynomials
  - (Grain) tolerance
- Line thinning
  - Vertex thinning
  - (Weed) tolerance

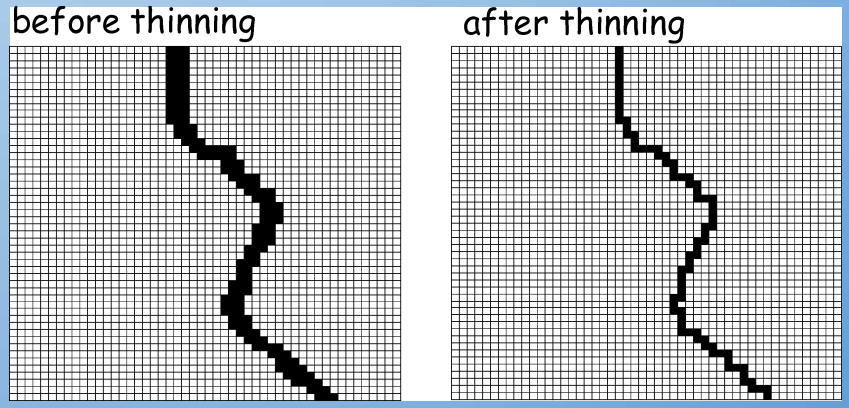
- Splitting
- Adding
- Deleting
- Updating
- Clean with appropriate tolerances

- Use smoothing and thinning functions in combination with tolerances to achieve specified levels of smoothing without overdoing it.

Scan Digitizing

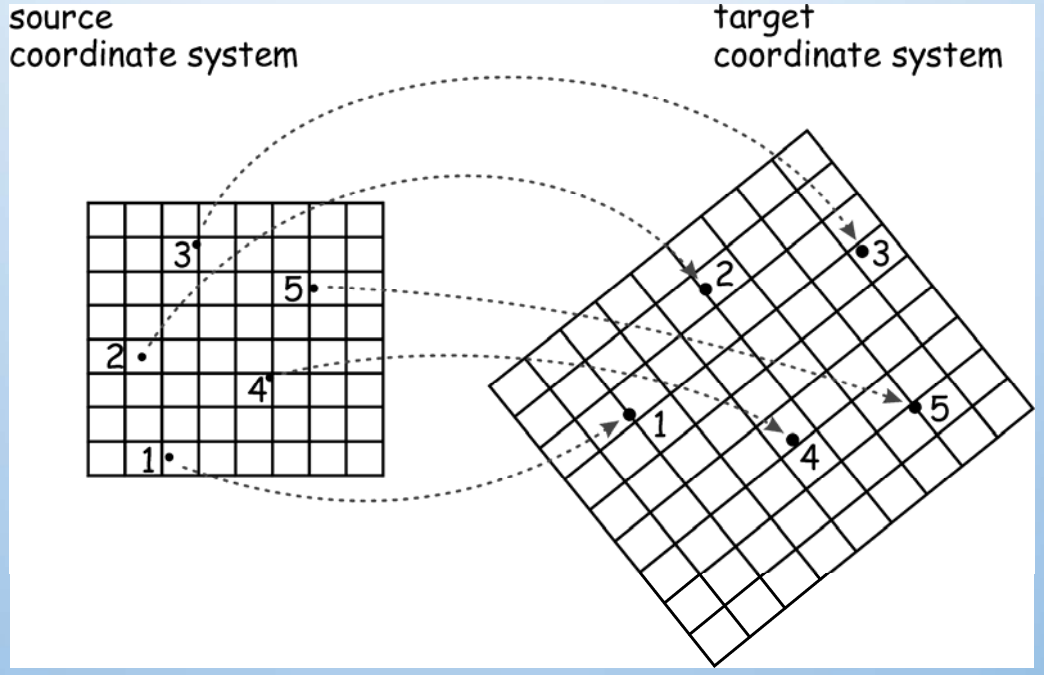
- Errors
  - Spurious feature capture/generation (e.g., creating a line where none exists)
  - Omission of features
  - Wrong scan threshold, overspecification of features (lines are created with 2 dimensions as areas)

- For raster data
  - Thinning may be an iterative process to find the correct scanning thresholds

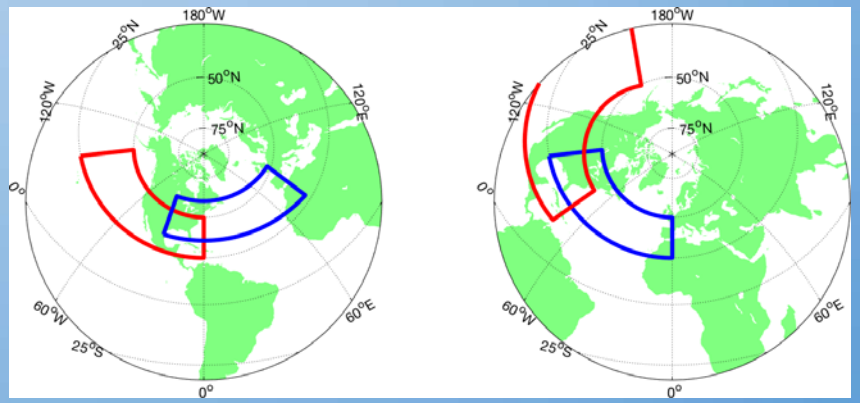


Coordinate Transformation and Georeferencing

The digitized world versus the “real world”

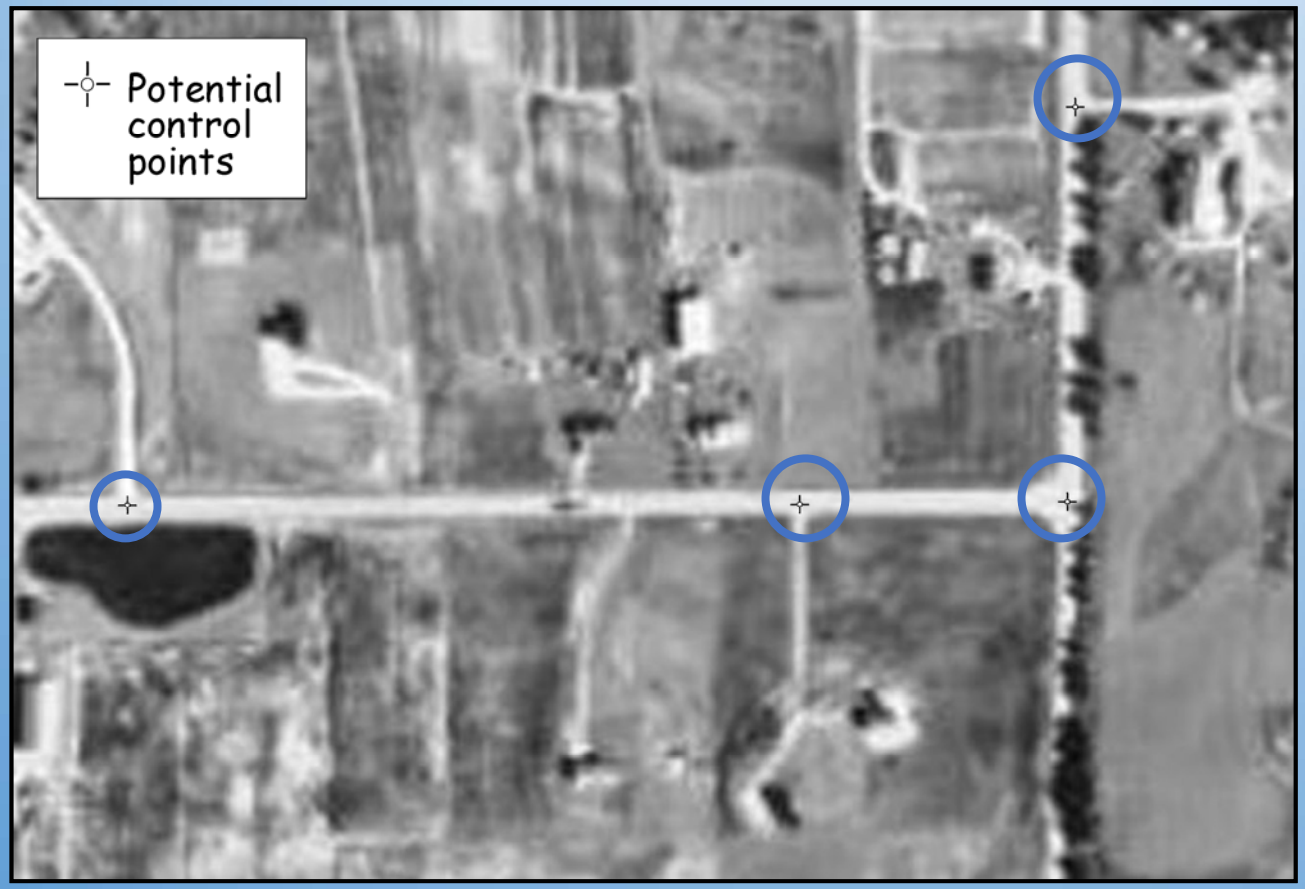


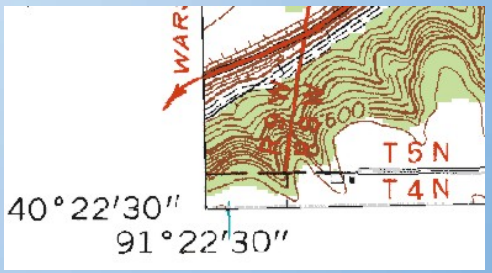
- Airplane heading
- Photo on scanner
- Paper map orientation on digitizing tablet
- Distortions, etc.



Control Points

- **Georeferencing**: The process involves identifying a series of ground **control points**—known x,y coordinates—that link locations on the raster dataset with locations in the spatially referenced data.
- **Control points** are locations that can be accurately identified on the raster dataset and in real-world coordinates.



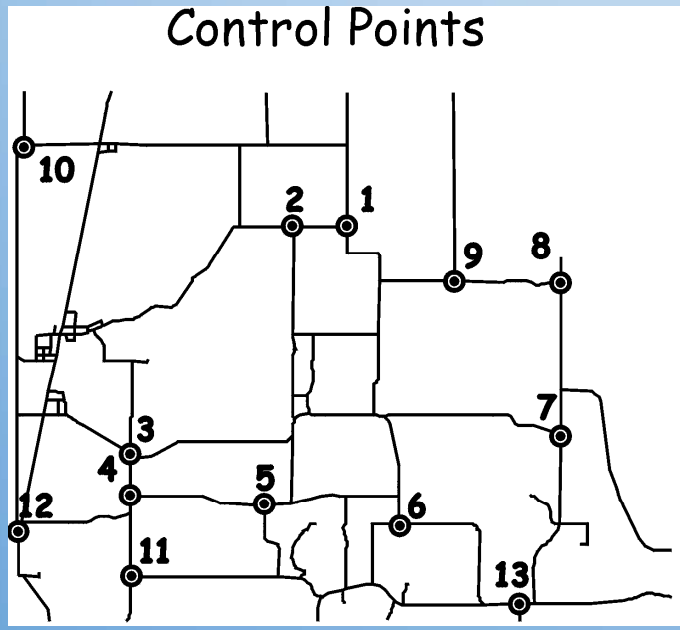


Georeferencing Example

Historic map georeferenced to the basemap.

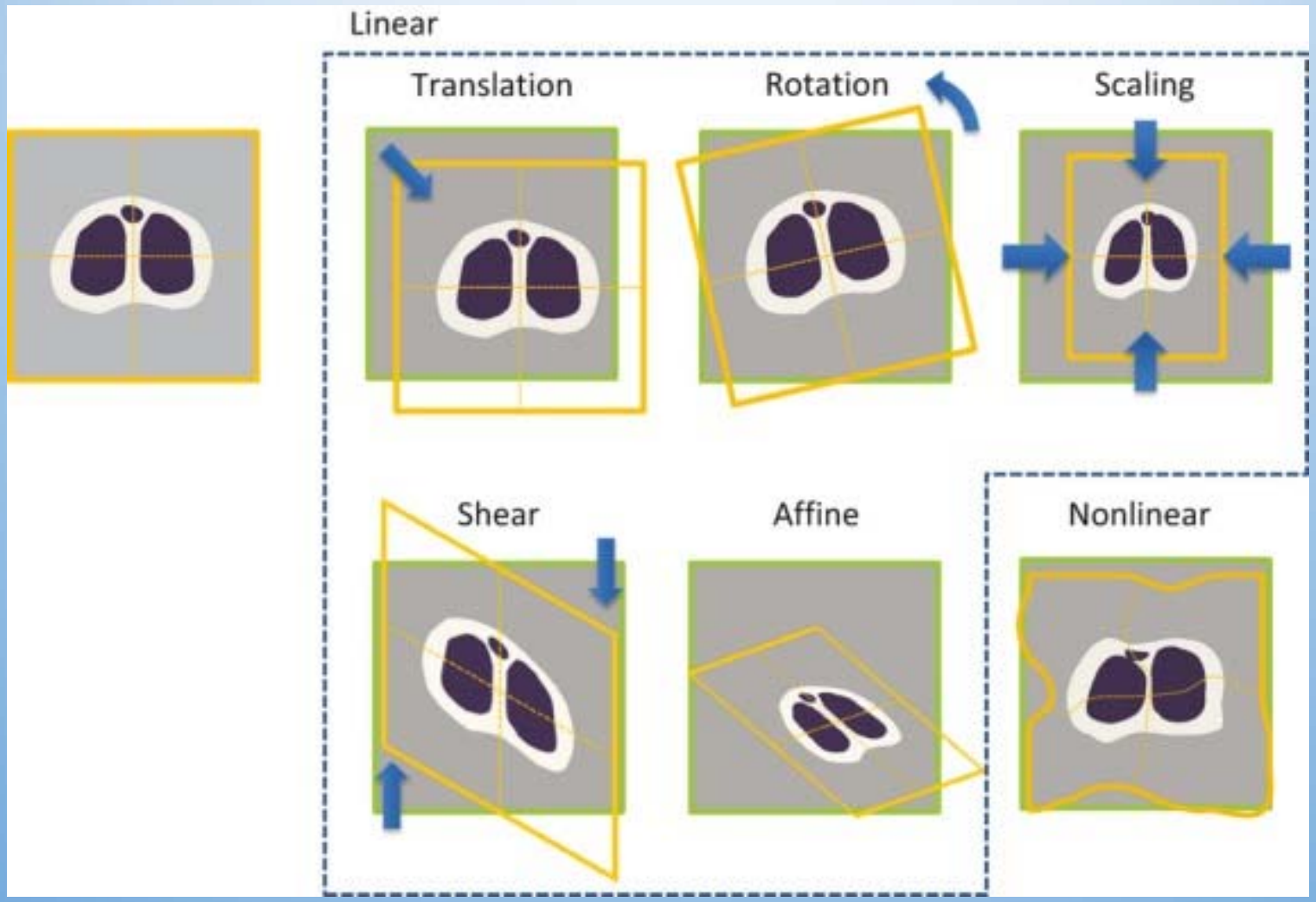
Control Points

- Two sets of known X-Y coordinates (e.g., image pixels or digitizer inches and UTM meters)
- Evenly distributed as possible throughout data area
- GIS applies algorithms to transform coordinate pairs from one system to the other



Digitizer Coordinates			Projection Coordinates (UTM)	
ID	x	y	E	N
1.0	103.0	-100.1	500,083.4	5,003,683.5
2.0	0.8	-69.1	504,092.3	5,002,499.5
3.0	-20.0	-69.0	504,907.5	5,002,499.5
4.0	-60.0	-47.0	506,493.3	5,001,673.5
5.0	-102.0	-47.2	508,101.3	5,001,651.0
6.0	-101.7	10.8	508,090.1	4,999,384.0
7.0	-86.0	75.8	507,475.9	4,996,849.0
8.0	-40.0	45.7	505,689.2	4,998,022.0
9.0	11.0	36.8	503,679.2	4,998,368.0
10.0	63.0	34.0	501,657.9	4,998,479.5
11.0	63.0	17.7	501,669.1	4,999,116.0
12.0	63.0	64.3	501,680.3	4,997,296.0
13.0	106.0	47.7	500,005.3	4,997,943.5

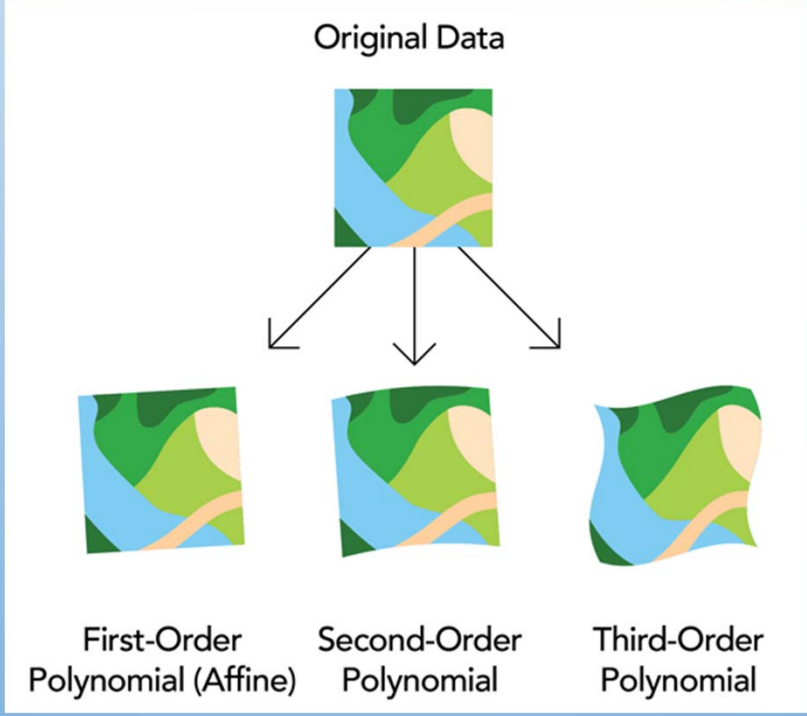
Geometric transformation functions for image registration



Affine transformation is a combination of translation, rotation, scaling, and shear.

Polynomial Transformations

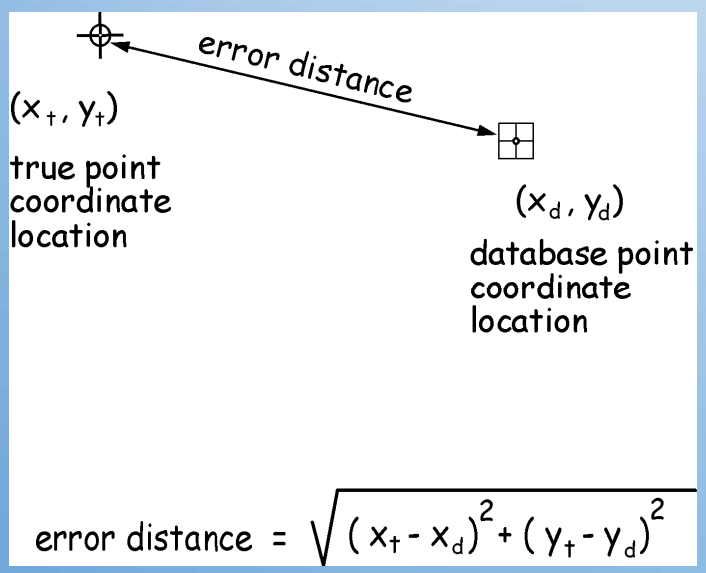
- Higher order for more complex distortions
- Can introduce distortion away from GCPs
- Best with many GCPs, distributed over map, especially away from center

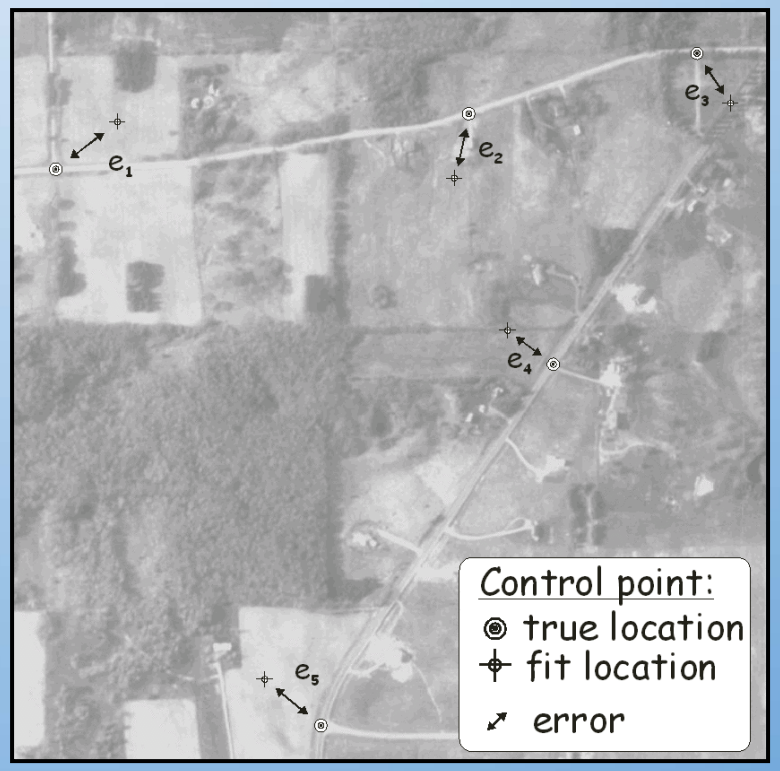


Root Mean Square Error

- Add/delete GCPs
- Move GCPs
- Try different model
- Check RMS error on the fly; note direction of errors

$$RMSE = \sqrt{\frac{(e_1^2 + e_2^2 + \ldots + e_n^2)}{n}}$$





Super Important Tip for Georeference

- When adding GCP pairs in ArcGIS, always first click on raster image, then click on the referenced spatial dataset

Your Project update

Homework 1

- Retrieve Covid-19 data from online source (Try to have latest data).
- You can choose world country data or USA statewide data.
- Process and map any 2 sub-variables the Covid-19 data
- Provide 500 words analysis of your maps produced.
- Homework Due before next class. Upload on google drive link.

- Tips:
  You can use base map provided in Assignment 1 or 2.