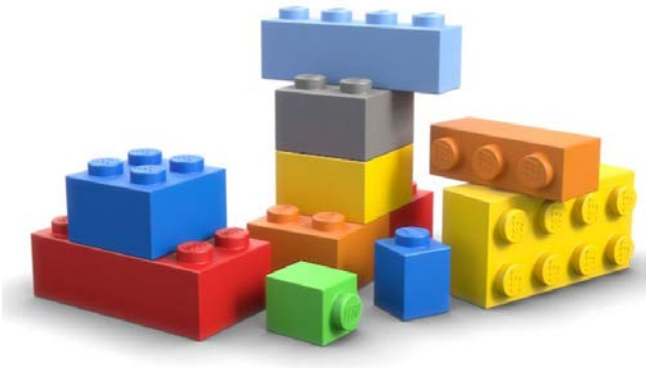
Generating and Managing Spatial Data

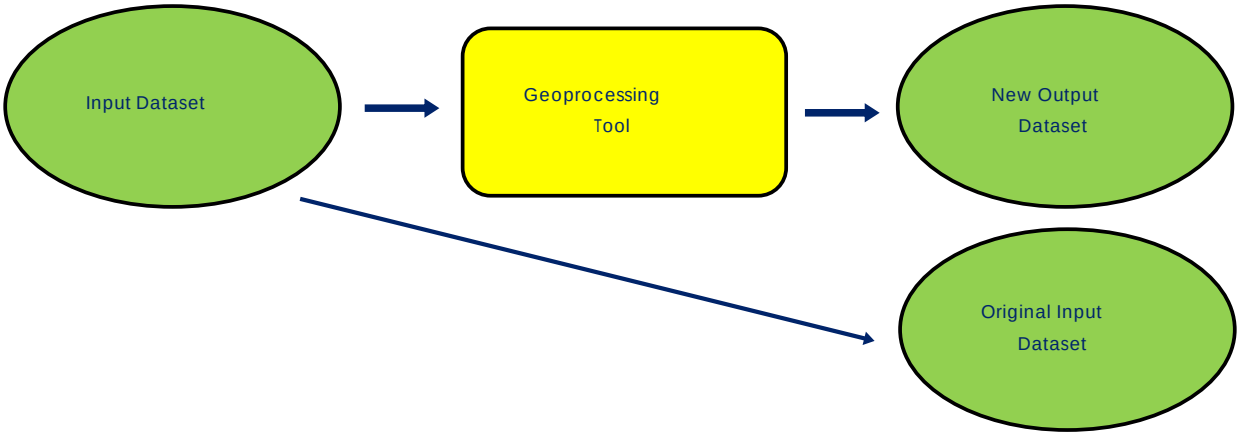
Learning Objectives

- **Introduce the concept of geoprocessing**
- **Explore common geoprocessing tools**
- **Understand the geocoding process**



What are geoprocessing tools

- **Any operation that transforms spatial data**
- **In many cases, geoprocessing tasks create a new dataset**
- **Usually honors active selections**



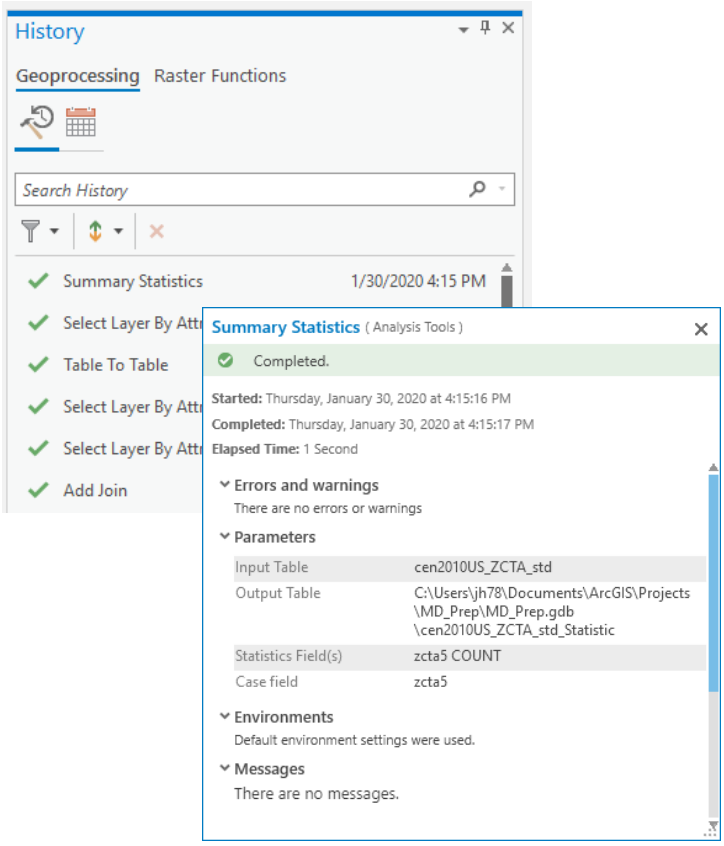
Geoprocessing is a catch-all term for the manipulation of geographic data.

In most cases, you don't need to worry about "messing up" your data by running a geoprocessing tool because your input dataset remains intact.

History

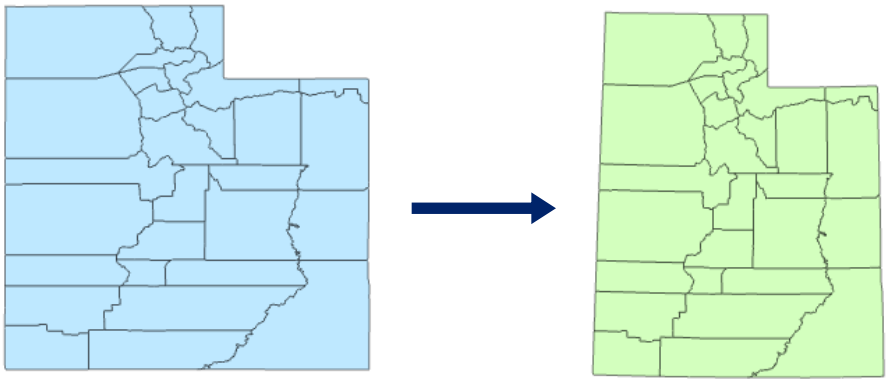
- Log of tools run are saved in the project

- Review input parameters
- Check warning/error messages
- Re-run the analysis



Project

- **Project spatial data from one coordinate system to another**
- **Can also define output coordinate system from the Environments options**



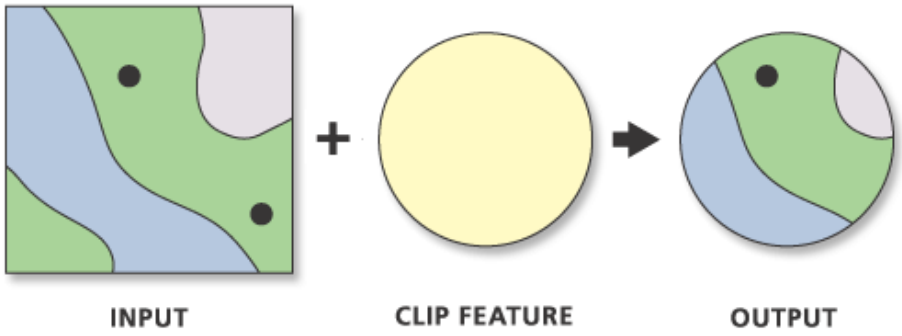
INPUT OUTPUT

The Project tool takes an input dataset, transforms its coordinate system, and returns an output dataset.

Good practice when using geoprocessing tools is for all input feature classes to be in the same coordinate system.

Clip

- **Use a clip feature to extract portion of input feature**
- **New output has all the same attributes of input**

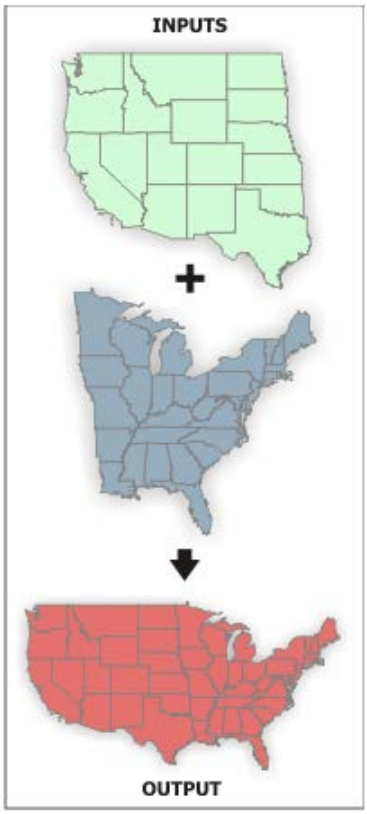


Clip: Use this tool to cut out a piece of one feature class using one or more of the features in another feature class as a cookie cutter.

Merge

- **Combines multiple input datasets into one (new) output dataset**

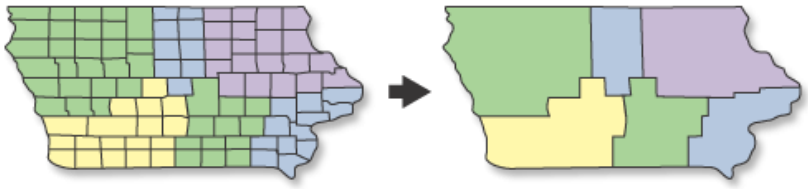
- **Merges both spatial and tabular information**



This tool is helpful if your data comes in multiple parts.

Dissolve

- **Aggregates features based on specified attribute(s).**
- **Dissolves the boundaries between features with the same value for that attribute.**
- **Aggregated tabular data can be summarized or described**

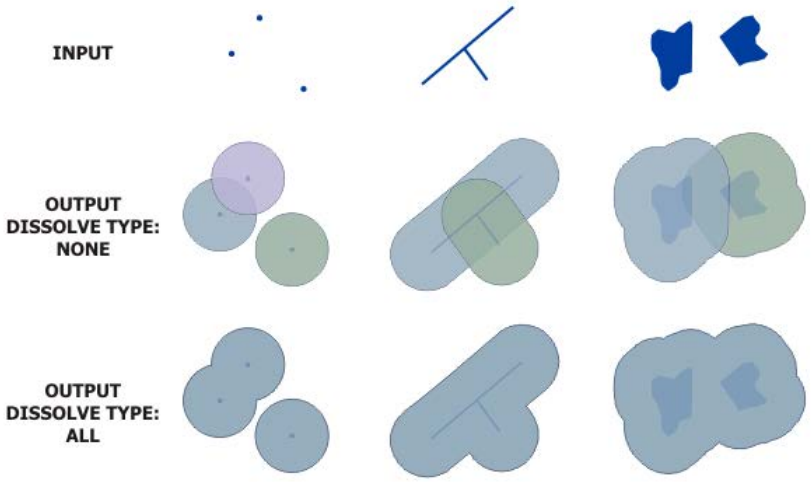


INPUT OUTPUT

For example, if you have counties that each have a code for the health region they belong to, you can dissolve using the health region code to create boundaries of the health regions

Buffer

- **Creates polygons showing the areas within a specified distance of the input features**
- **Input features can be points, lines, or polygons**
- **Output feature will always be polygons**

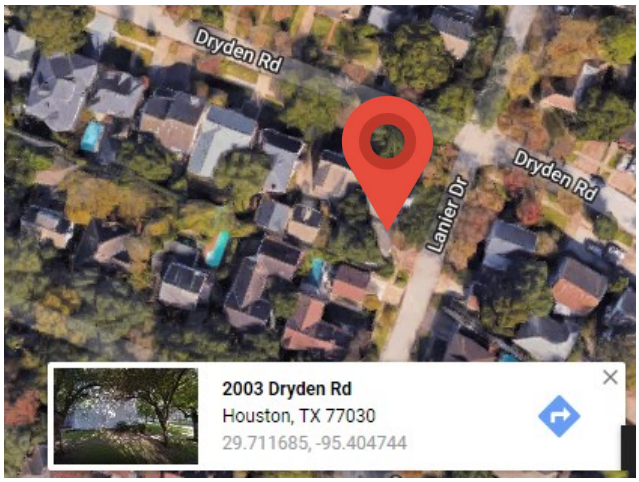


Useful for visualizing distances from points/places of interest

Risk of Re-identification

- **Standard data protections apply to spatial data**
  - **HIPAA**
  - **Institutional Review Boards**
  - **Privacy Laws**

- **Standard de-identification does not work for spatial data**
  - **Reverse Geocoding**
  - **Tax Assessors Office**
  - **Public Voter Database**
  - **Internet Searches**



We must use caution when dealing with individual-level or other data that is deemed confidential by your organization. In addition to general privacy laws, additional considerations such as HIPAA and IRB rules govern health-related data. All geographical subdivisions smaller than a State are considered confidential data.

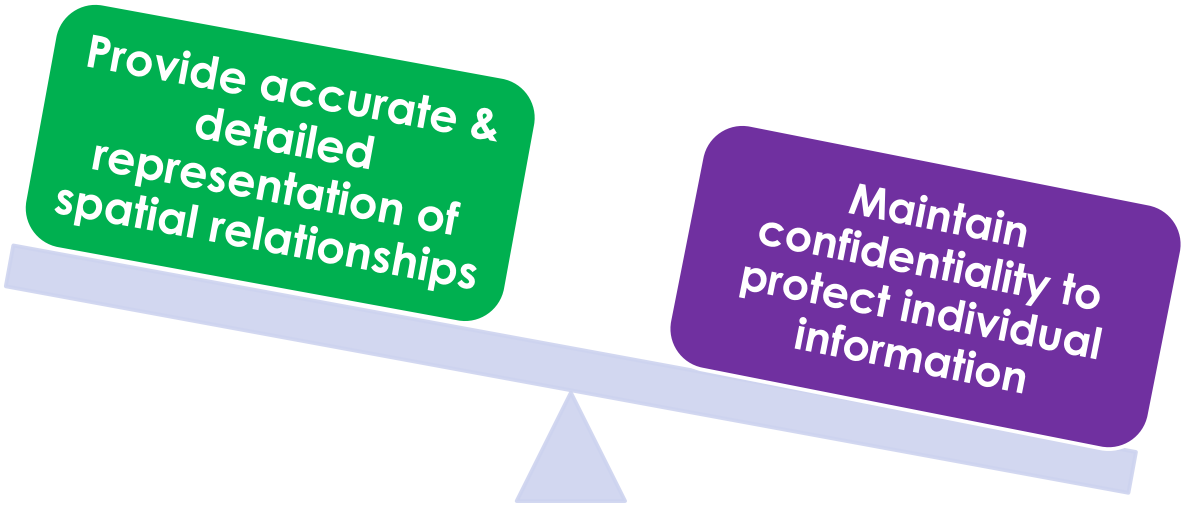
Simply removing name/DOB does not de-identify an individual who has associated geospatial data (lat/long, address, ZIP code).

There is a real risk to individuals if proper measures aren't taken. Latitude and Longitude is powerful information – you can even get this from a PDF file if not carefully protected. With lat/long, you can re-identify an individual and use publicly available datasets to find all kinds of personal information, especially given today's internet resources.

Balancing Act

Masking techniques *reduce* risk of de-identification but also alter your data structure

Example: aggregating individuals to the county level



Although masking protects confidentiality, it degrades spatial resolution and can obscure existing spatial patterns in some cases.

For example, the higher the level of aggregation, the less effective a map will be for identifying geographical and epidemiological trends at the local level. A map with the highest possible spatial resolution (i.e. showing the most detail) should be the most informative and lead to the most efficient use of public health resources. Also, patterns that cross a geopolitical boundary will be obscured by techniques such as aggregation.

Discussion

- **What address data needs geocoding? (ex. program locations, program participants, partner locations)**
- **What data could be made more useful though geoprocessing? (ex. creating new boundaries, clipping data to specific extent)**
- **How can we use buffers to describe distance from resources?**
- **Are there any potential issues regarding confidentiality and privacy?**