

# CALIFORNIA HYDROGEN INFRASTRUCTURE TOOL TECHNICAL FORMULATION WEBINAR

---

Primer on GIS and Data Source Concepts

October 09, 2015

For questions or comments, contact:  
Andrew Martinez  
(916) 322-8449  
[andrew.martinez@arb.ca.gov](mailto:andrew.martinez@arb.ca.gov)

# Primer Outline

Purpose: The optional primer is intended as a brief overview of GIS-related concepts pertinent to many aspects of the technical formulation of CHIT

- Brief Introduction and review of CHIT and AB 8 process
- Introduction to GIS fundamentals
- ArcGIS tools utilized in formulation of CHIT
- Creating custom tools in ArcGIS
- Working with geographies in Census and DMV data sources

# Primer Outline

- This primer will answer questions like:
  - What are some of the major built-in ArcGIS tools used in CHIT and what do they accomplish?
  - What are the differences in the spatial structure of the data sources and how are they reconciled?
  - How does ArcGIS assess spatial distribution of data? What statistical analyses are applied?
  - How is a tool like CHIT made in ArcGIS?

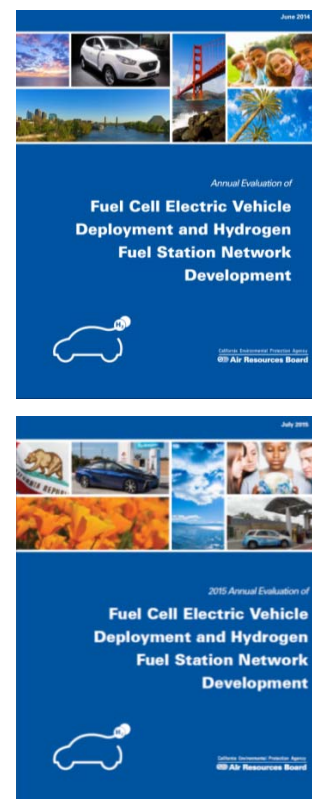
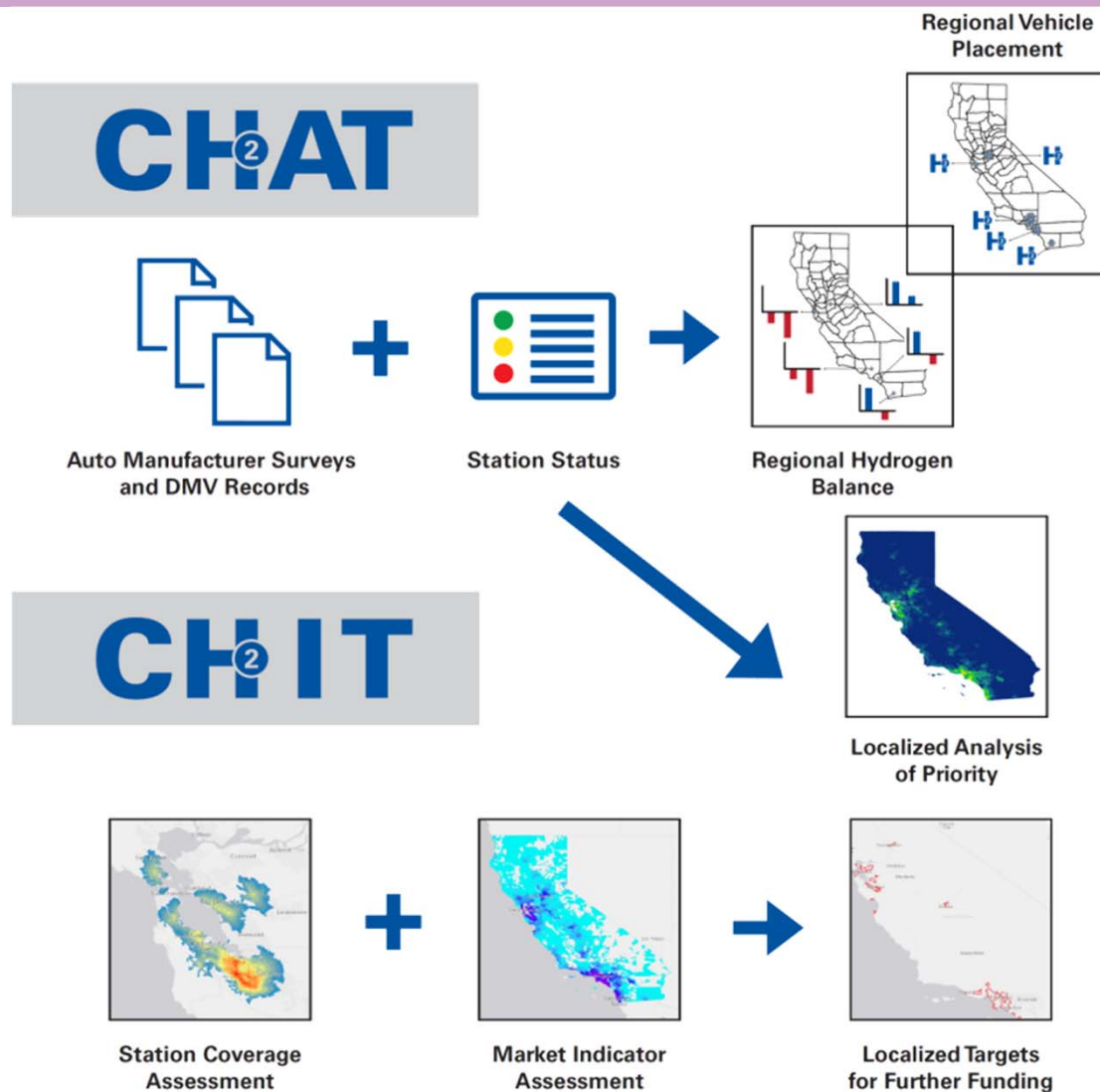


# INTRODUCTION AND REVIEW OF CHIT

---

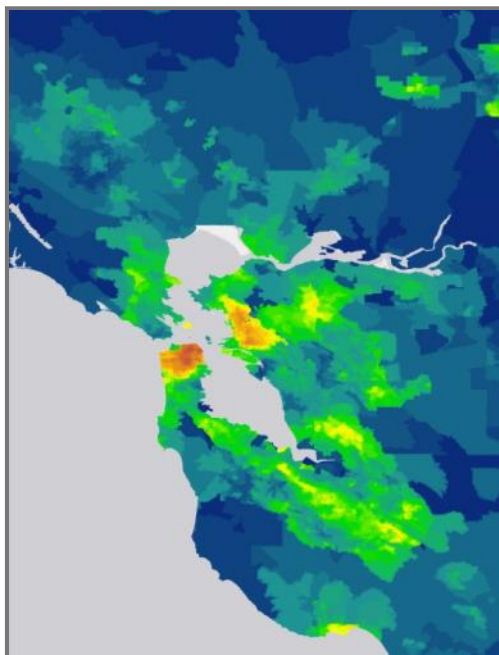
# Introduction

## CHIT/CHAT Tools and AB 8



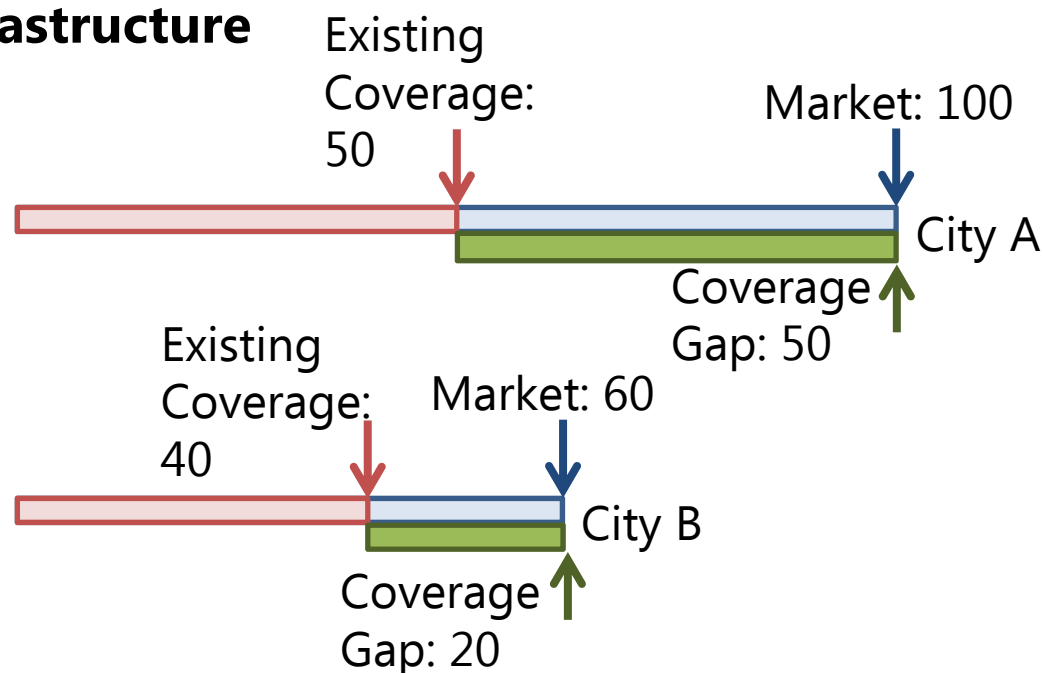
# Introduction

## CHIT: A Coverage and Market Assessment Tool



GIS Network Analysis  
and Station Area  
Planning

- CHIT is a planning tool intended to provide general direction indicating areas of needed infrastructure
- CHIT evaluates relative need for hydrogen infrastructure based on a gap analysis between a projected market and current infrastructure





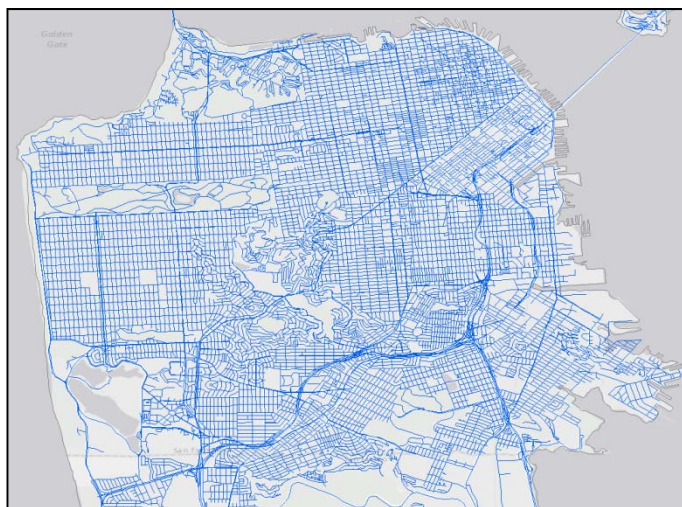
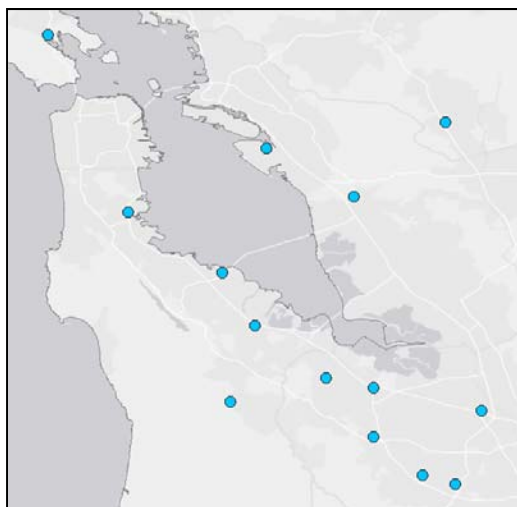
# FUNDAMENTALS OF GEOGRAPHIC INFORMATION SYSTEMS

---

# GIS Fundamentals

## Representing Geospatial Data

- Geospatial data can be conceptualized as objects that are stored, analyzed, compared, edited or transformed
- Objects have geometry (geography) and attributes
- Geometry includes dimension, size (resolution), orientation, location, etc...

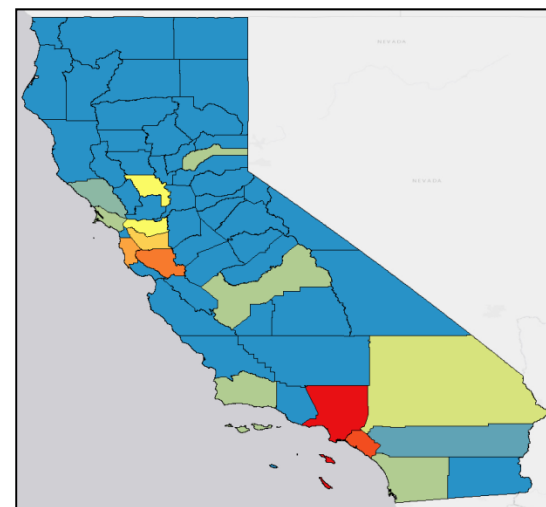
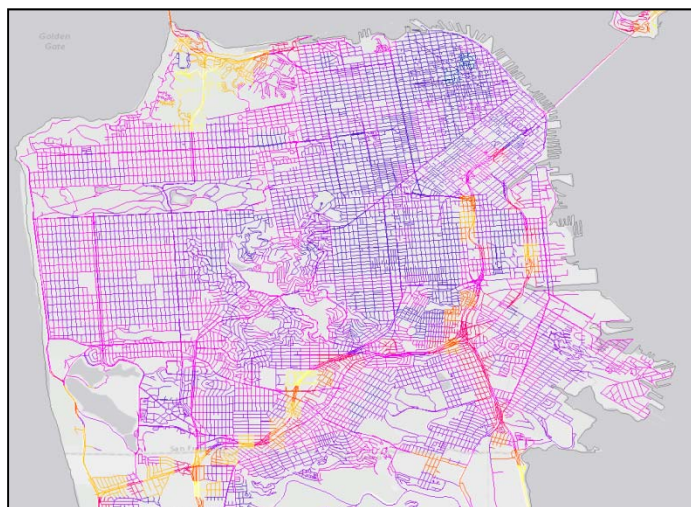
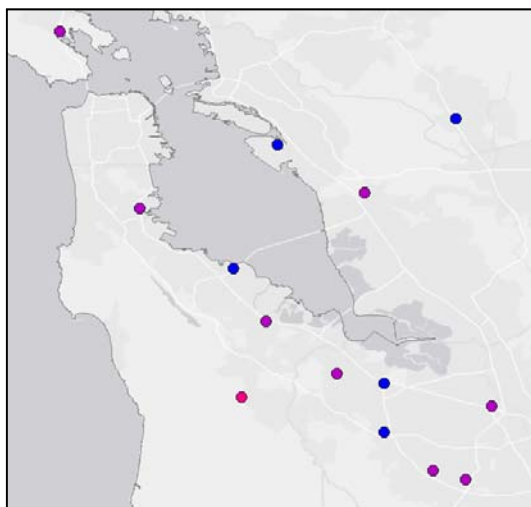




# GIS Fundamentals

## Representing Geospatial Data

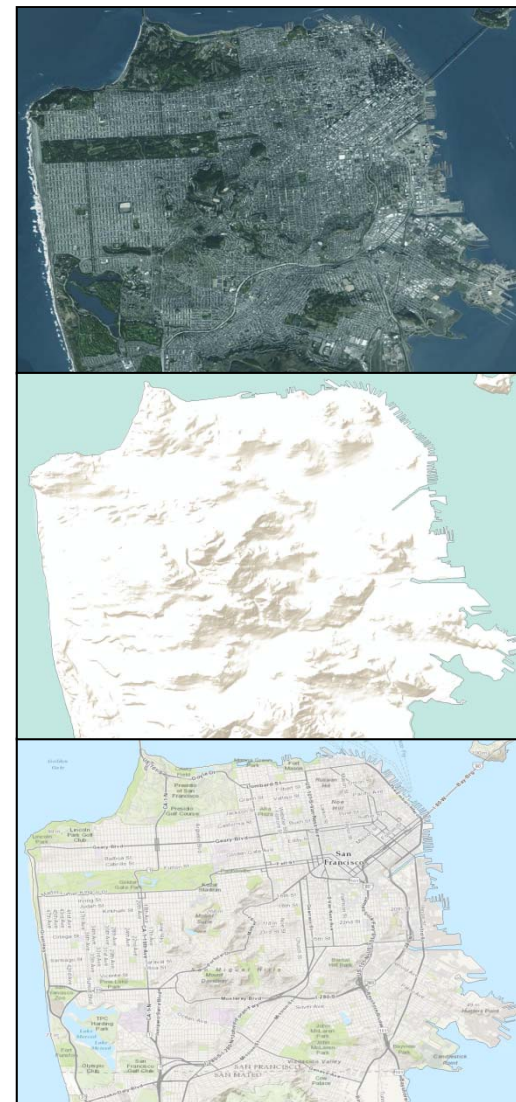
- Attributes are data fields, like temperature reading at a weather station, traffic volume along a stretch of road, or population counts in counties
- Not all data sources are equivalently rich and detailed in geometry and attributes
- May not find a ready-made data source with all required aspects for a given project



# GIS Fundamentals

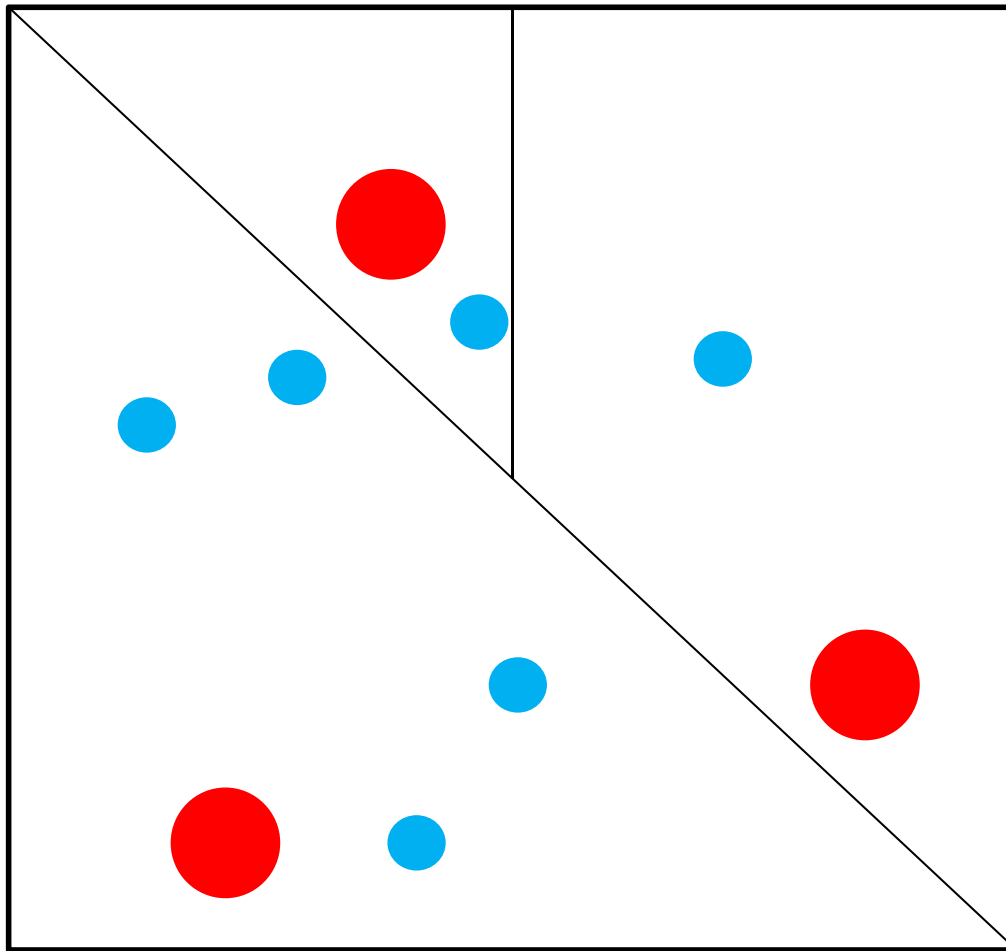
## Representing Geospatial Data

- When working with geospatial data, it is important to remain aware of geometry and attributes and assure the best and most logical choices are used for data analysis, transformation, and representation
- It is also important to always remember that even the most detailed geospatial data is itself *a representation or model* of the world
  - May closely match the real world but all data will have inherent errors, even before any analysis is performed
  - There are potentially many valid models of the real world
- Data quality needs to be constantly monitored and assessed during analyses



# GIS Fundamentals

## Normalization and Representation of Quantitative Data



Target  
Population

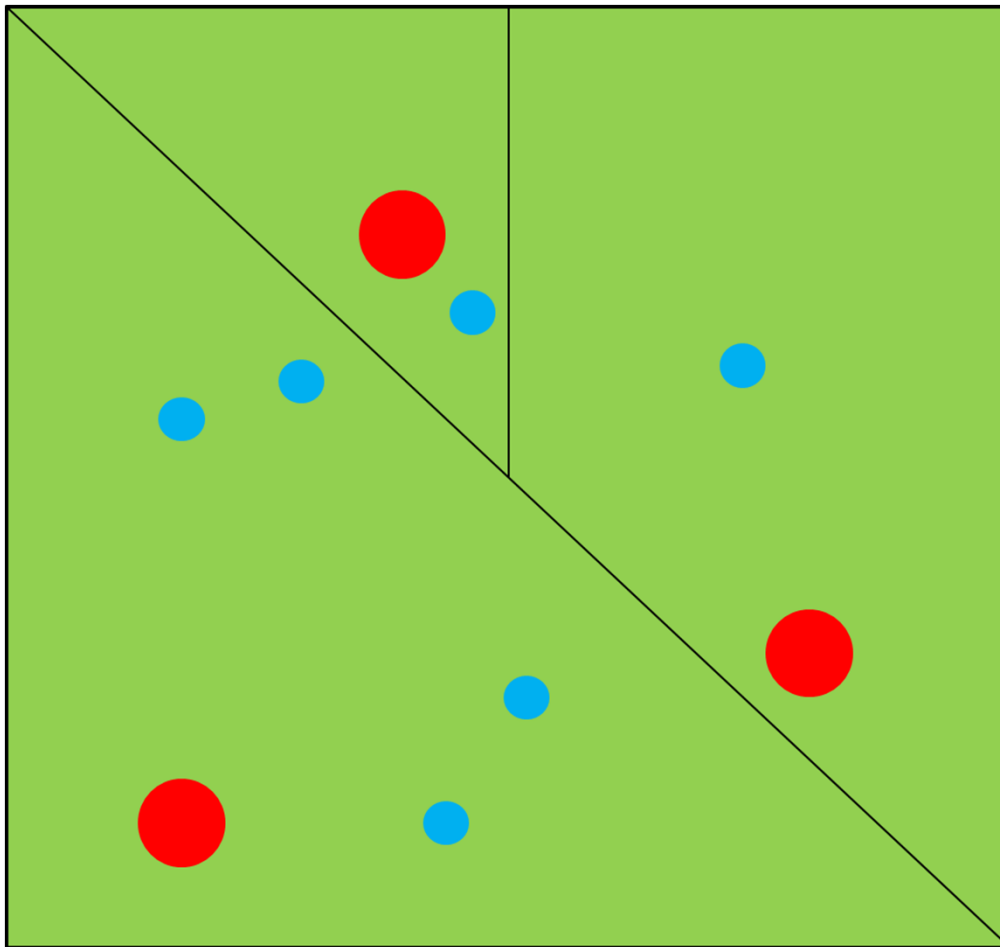
Additional  
Population

How should we relatively weight the count of red dots in each division?

- 1) Independent of containing area and blue dots in same area
- 2) Proportional to total number of dots (red + blue) in area
- 3) Normalized to size of containing area
- 4) Proportional to total and normalized by area

# GIS Fundamentals

## Normalization and Representation of Quantitative Data

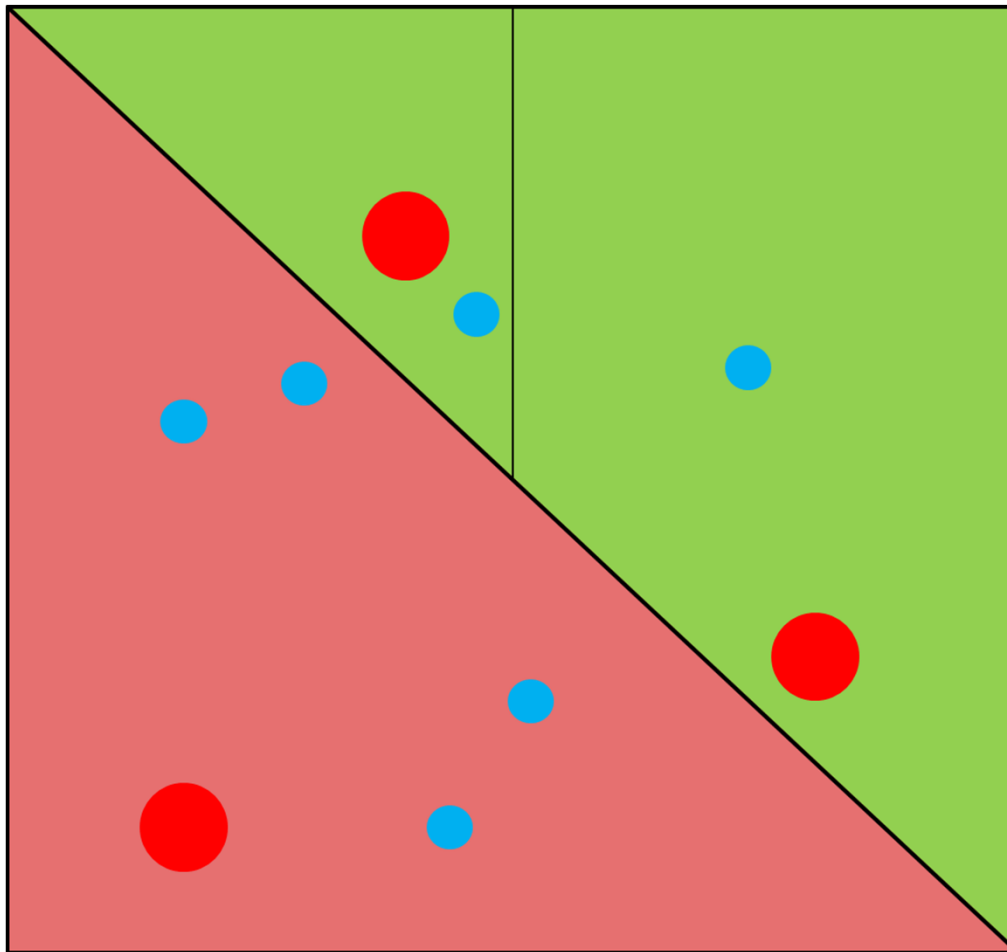


How should we relatively weight the count of red dots in each division?

- 1) Independent of containing area and blue dots in same area
- 2) Proportional to total number of dots (red + blue) in area
- 3) Normalized to size of containing area
- 4) Proportional to total and normalized by area

# GIS Fundamentals

## Normalization and Representation of Quantitative Data

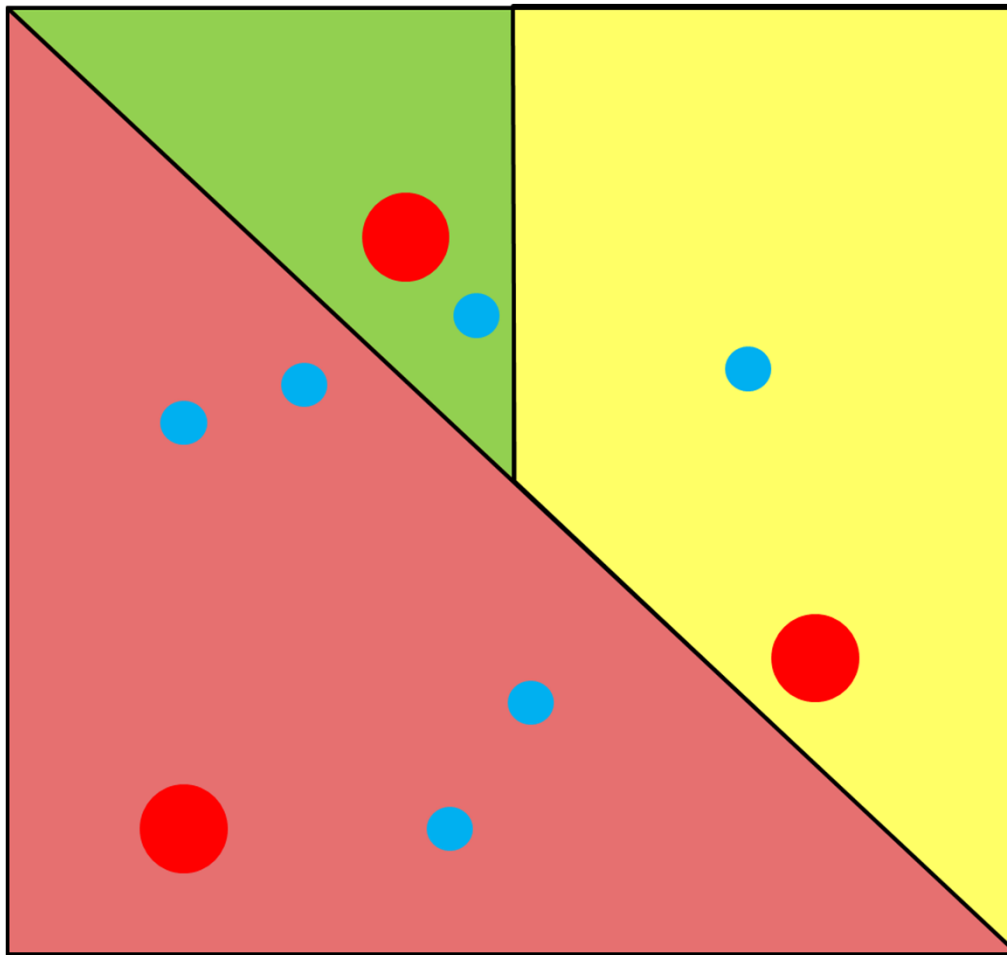


How should we relatively weight the count of red dots in each division?

- 1) Independent of containing area and blue dots in same area
- 2) Proportional to total number of dots (red + blue) in area
- 3) Normalized to size of containing area
- 4) Proportional to total and normalized by area

# GIS Fundamentals

## Normalization and Representation of Quantitative Data

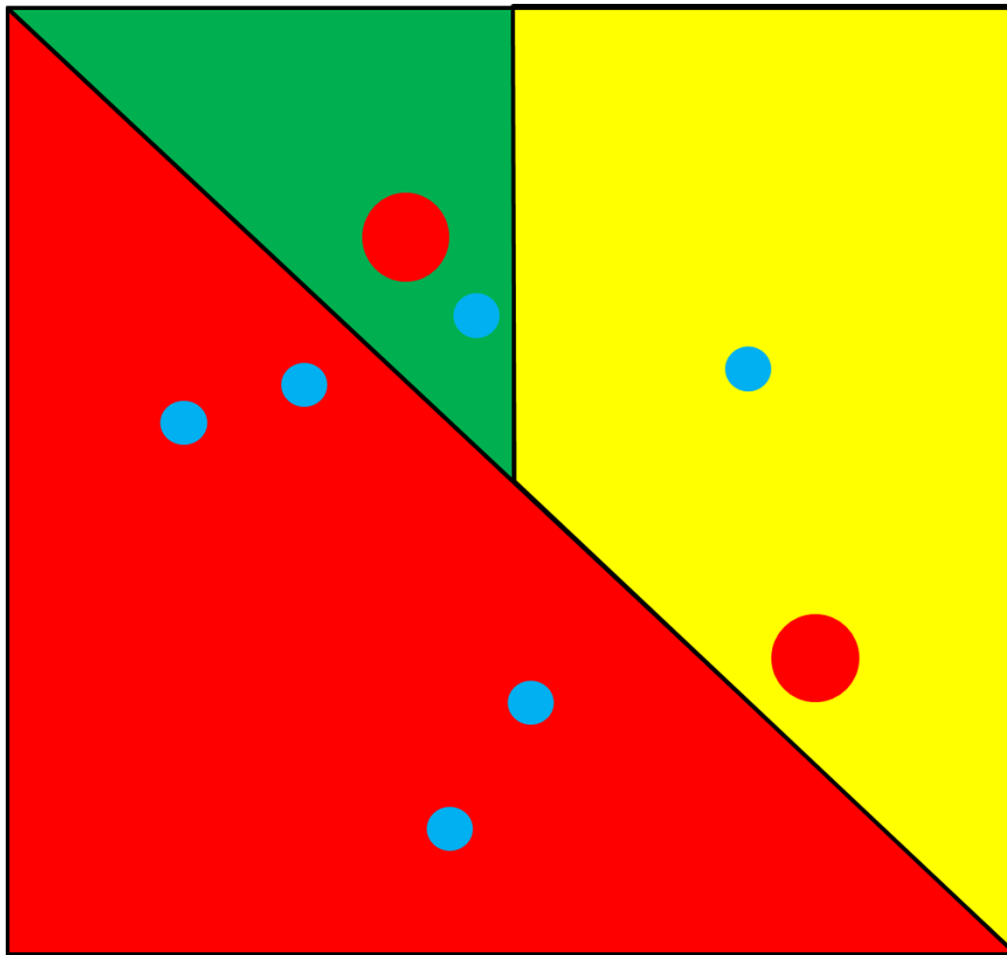


How should we relatively weight the count of red dots in each division?

- 1) Independent of containing area and blue dots in same area
- 2) Proportional to total number of dots (red + blue) in area
- 3) **Normalized to size of containing area**
- 4) Proportional to total and normalized by area

# GIS Fundamentals

## Normalization and Representation of Quantitative Data

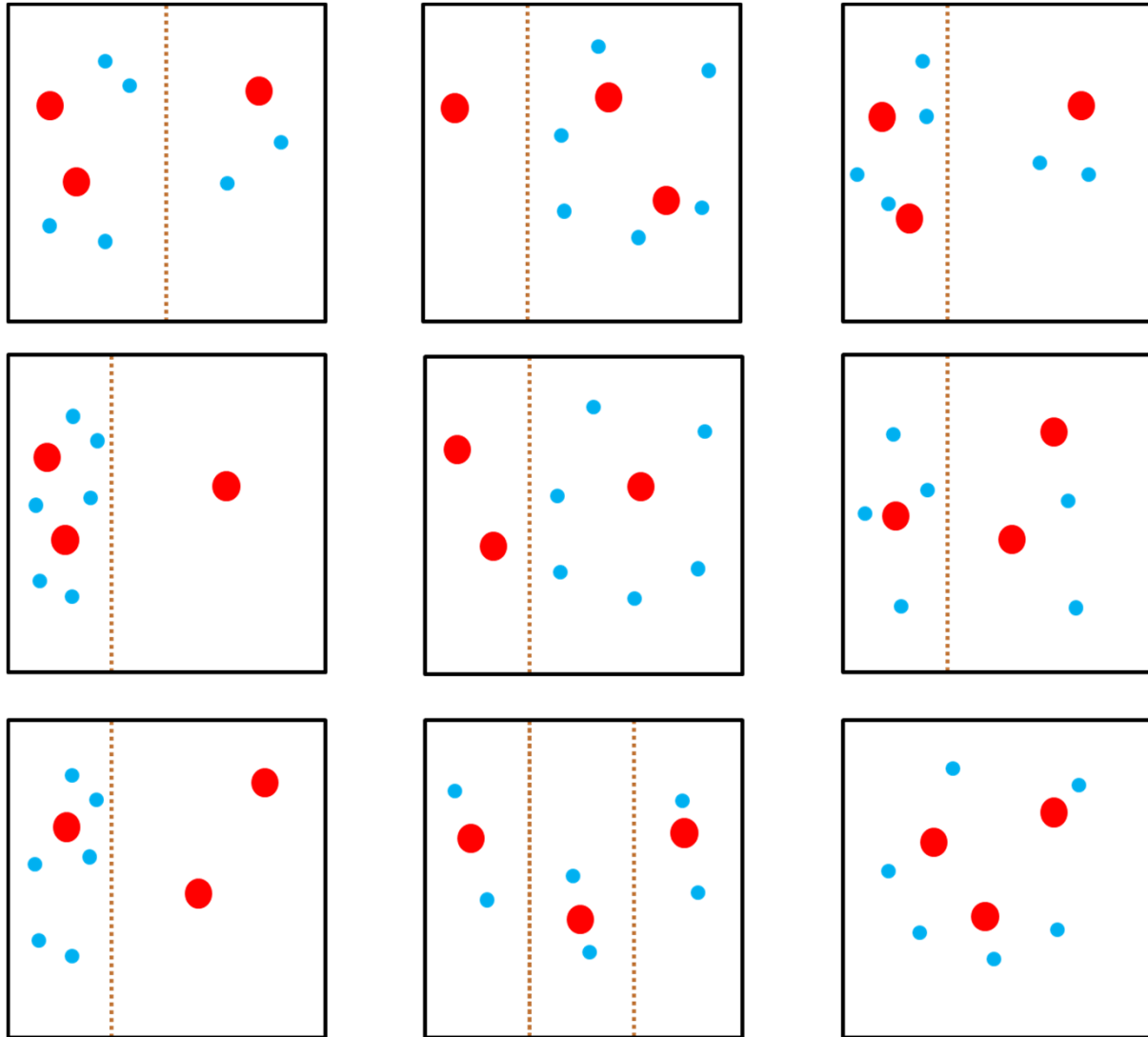


How should we relatively weight the count of red dots in each division?

- 1) Independent of containing area and blue dots in same area
- 2) Proportional to total number of dots (red + blue) in area
- 3) Normalized to size of containing area
- 4) Proportional to total and normalized by area

# GIS Fundamentals

## Normalization and Representation of Quantitative Data





# GIS Fundamentals

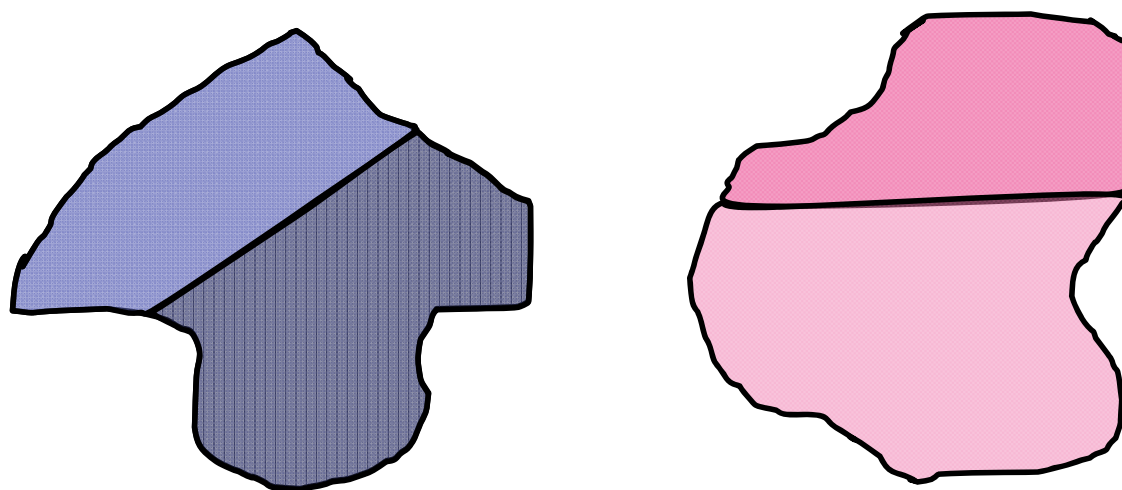
## Normalization and Representation of Quantitative Data

- Best practice in GIS to consider area-based normalization
  - Provides consistent basis of intensity measure for objects of varying sizes
- Considering proportion of target market to total population can be influenced by intent of analysis
  - Do the blue dots (not target market) matter in our calculation?
- Normalization by population accounts for population type clustering, but causes issues of representing large absolute differences
- Normalization by both population and area does not eliminate this problem

# GIS Fundamentals

## Overlay Actions

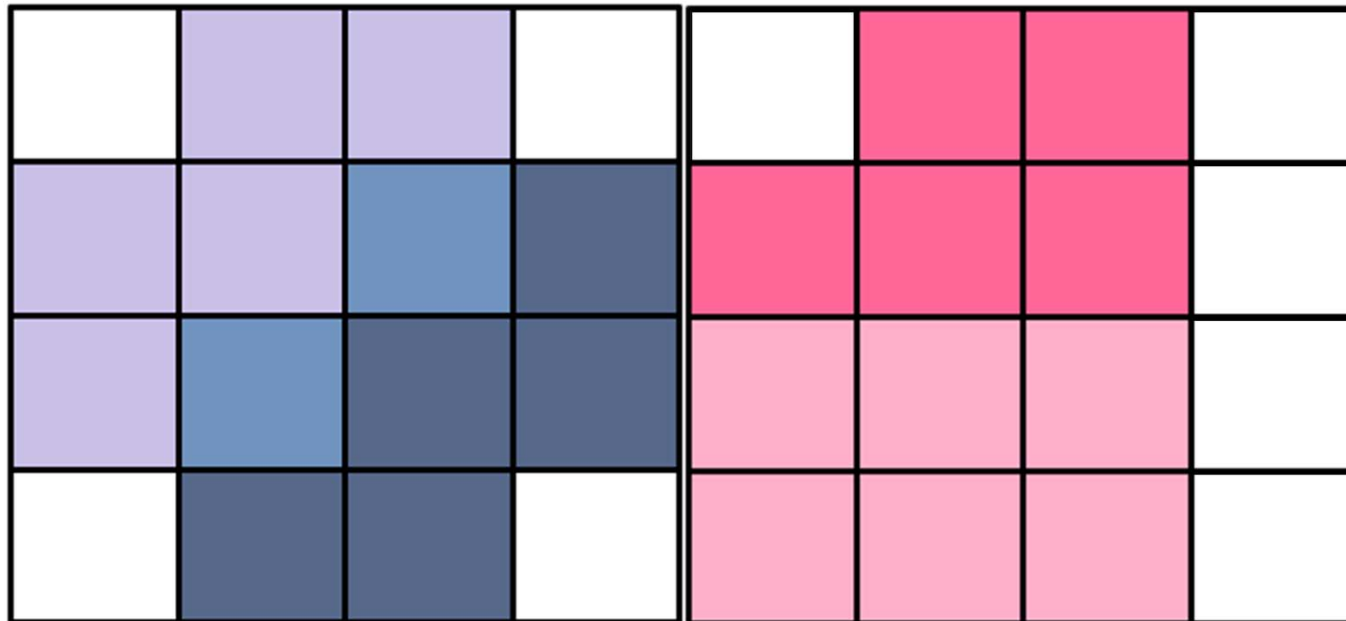
- Overlay actions are one of the most common analysis processes within GIS
- Used to compare attributes from varying data sets and possibly of different geometries
- Overlaying geometries can lead to creation of undesirable features and rapidly growing numbers of objects




# GIS Fundamentals

## Overlay Actions

- Can alleviate by using a structured grid representation of the input data
- Grid resolution must be carefully considered as compared to input data geometry





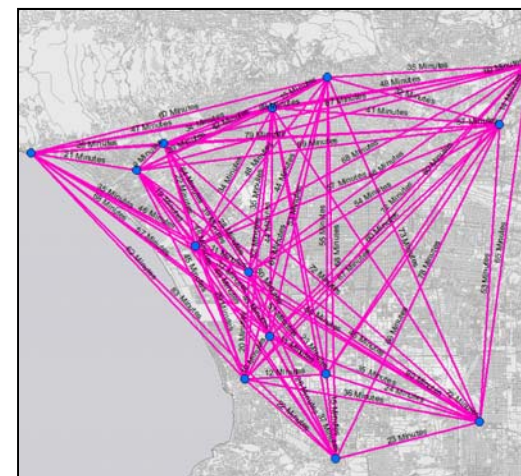
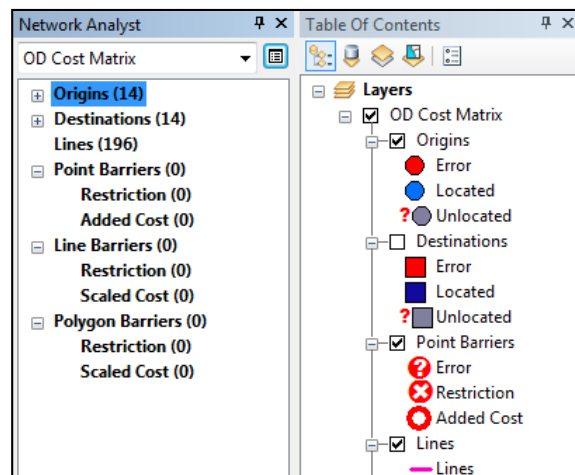
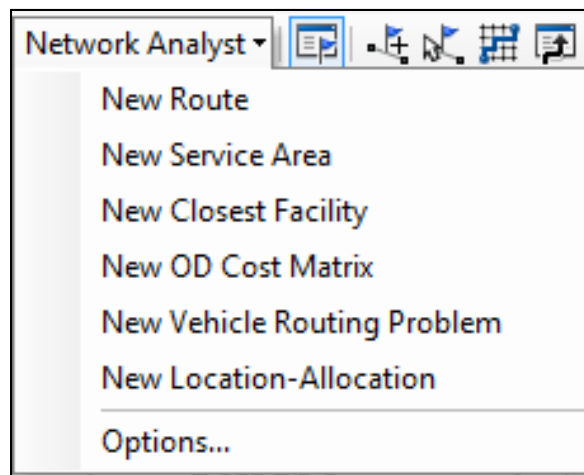
# ARCGIS TOOLS USED IN DEVELOPMENT OF CHIT

---

# ArcGIS Functions

## Network Analysis and Service Areas

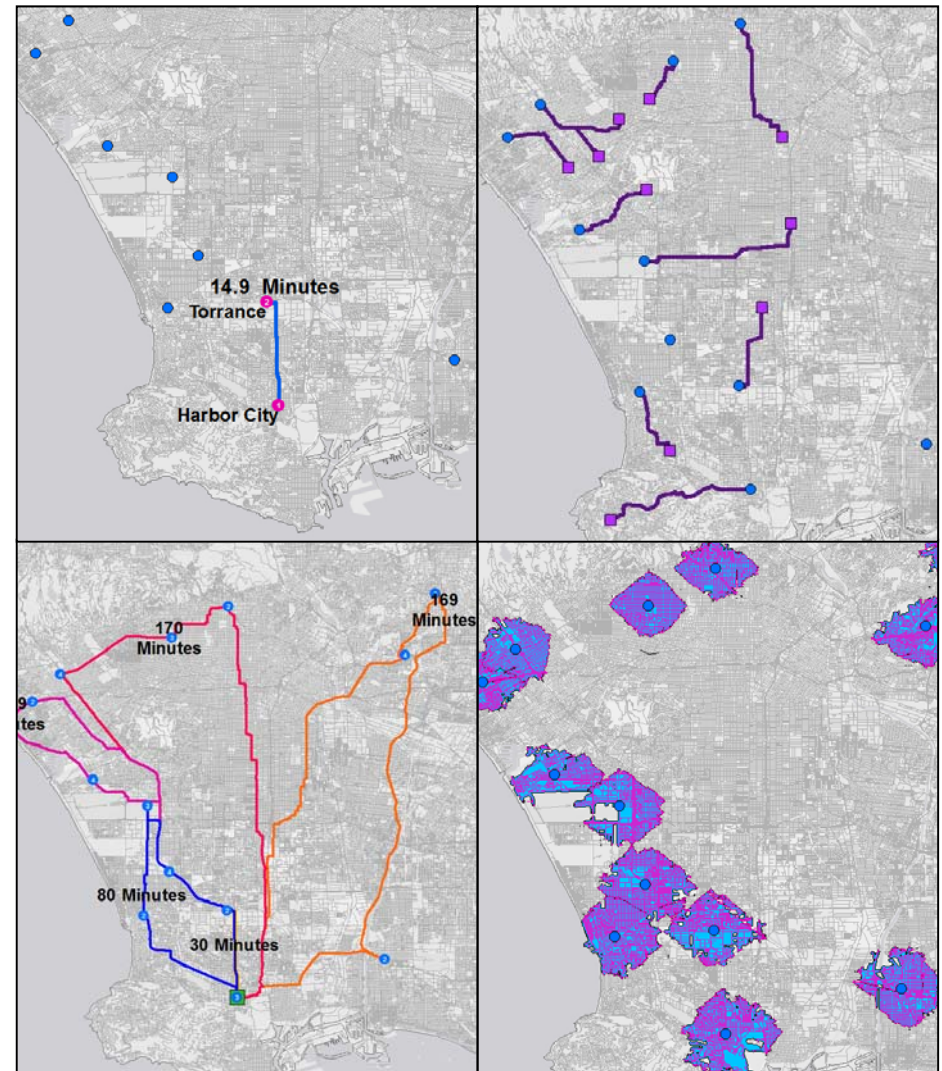
- If you've used a mapping application to find directions to a destination, you have used a network analysis tool
- Network analysis works with data sets having:
  - Geometry representing the system of roadways
  - Attributes representing travel metrics like velocity
- Representation of the roadway network has significant impact on the quality and validity of network analyst solutions



# ArcGIS Functions

## Network Analysis and Service Areas

- Network analysis can be used to:
  - Find the fastest route between an origin and destination (as in the fastest route from work to home when there's heavy highway traffic)
  - Find the closest facility to an origin from a set of available options (as in finding the closest pizza restaurant to your house)
  - Optimize the order of visiting multiple locations on a route (as in optimizing a delivery truck's daily route)
  - Find the boundary of the area that can be reached from a starting point within a given constraint (as in finding all the wineries you can reach by driving at most 20 minutes from downtown Napa)



# ArcGIS Functions

## Network Analysis and Service Areas

- Network analyses provide options for highly detailed considerations, if the data are sufficiently known:
  - Elevations of roads, to determine sections that are over/underpasses
  - U-turn capabilities of vehicles and locations of allowable U-turns
  - Curb approaches allowed for vehicles to various locations
  - Delays encountered at traffic lights and in making turns
- Degree of detail to consider depends heavily on data and computing resources, extent of the network analysis, and desired goals of the analysis

# ArcGIS Functions

## Statistical Analysis and Spatial Distributions

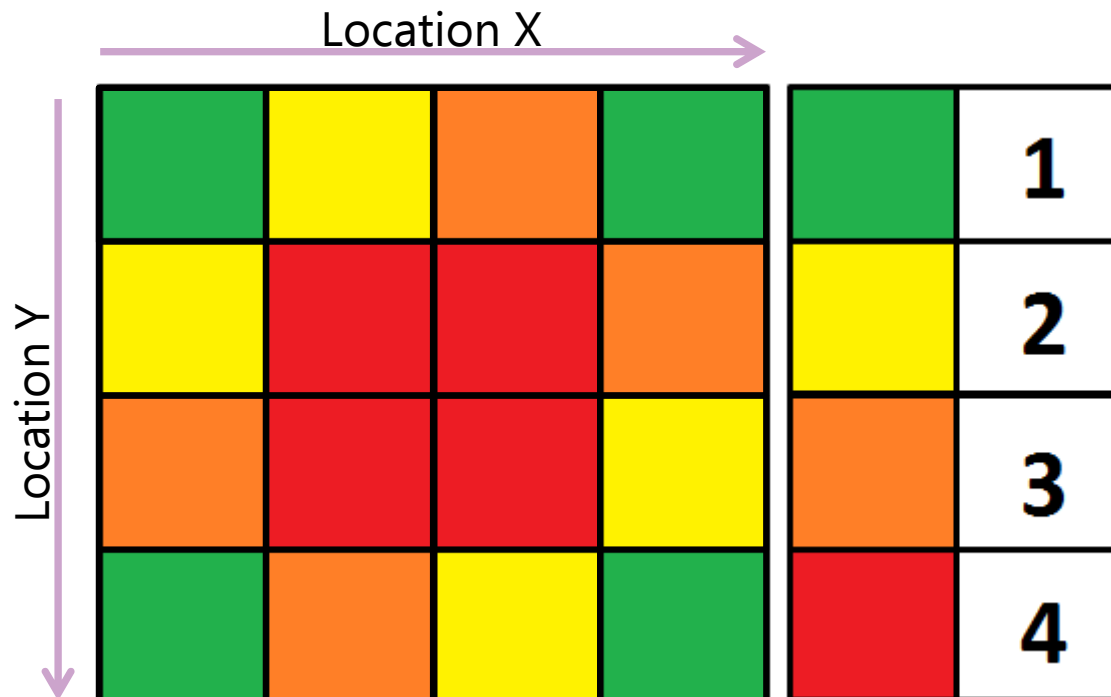
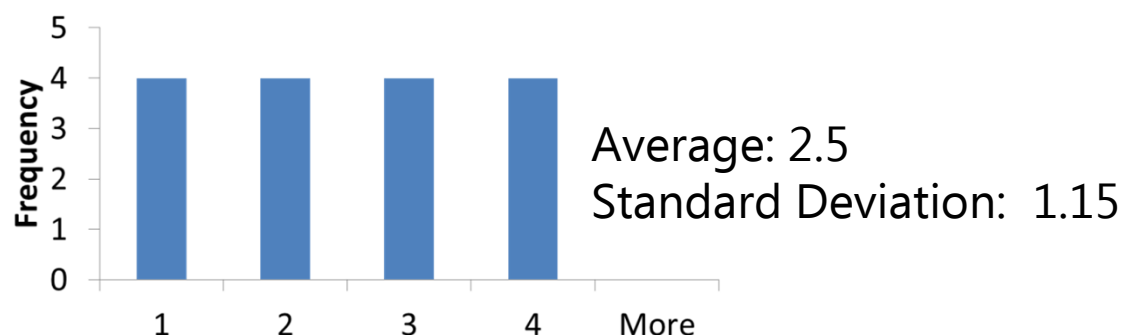
- Analysis of spatial data involves an additional dimension(s) beyond non-spatial data
- Patterns that may not be apparent in non-spatial data can be readily evident when spatial arrangement of the data is considered
- Spatial statistics is the study and implementation of statistically analyzing spatial variability and uncertainty
- Can be useful for data interpolation, analysis and identification of trends and patterns, comparison of attributes in data sets, and identification of co-variation between attributes



# ArcGIS Functions

## Statistical Analysis and Spatial Distributions

Location X	Location Y	Value
1	1	1
2	2	4
4	1	1
3	1	3
3	2	4
4	2	3
1	3	3
1	4	1
2	1	2
2	4	3
1	2	2
4	3	2
4	4	1
2	3	4
3	4	2
3	3	4

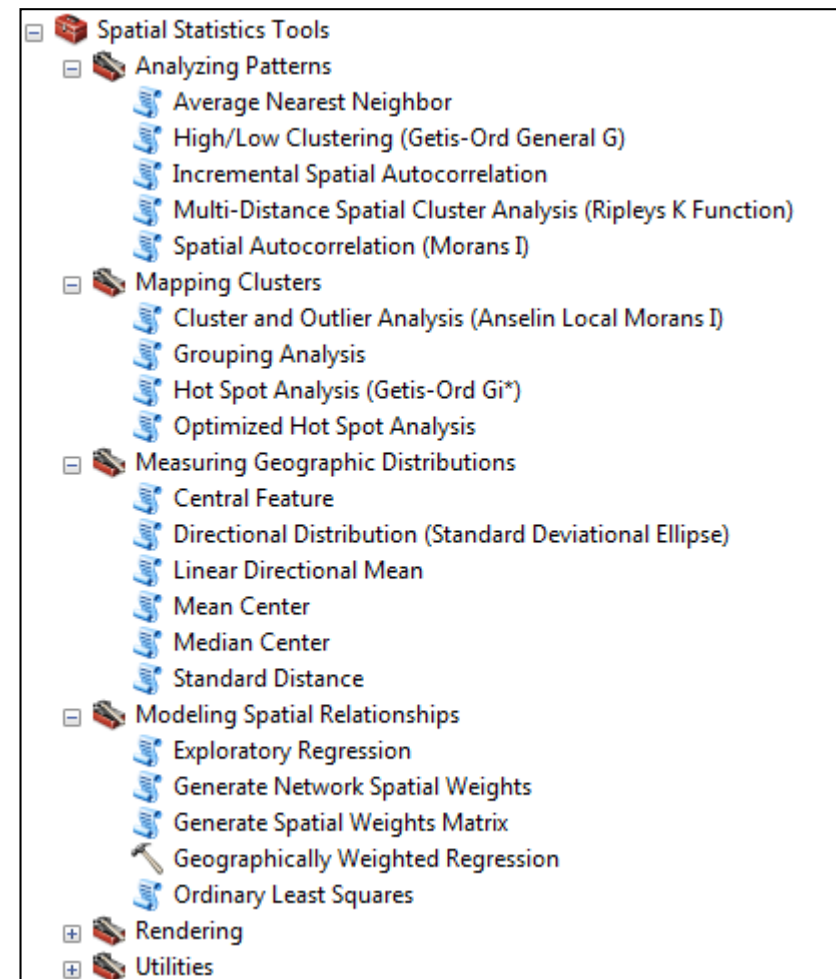
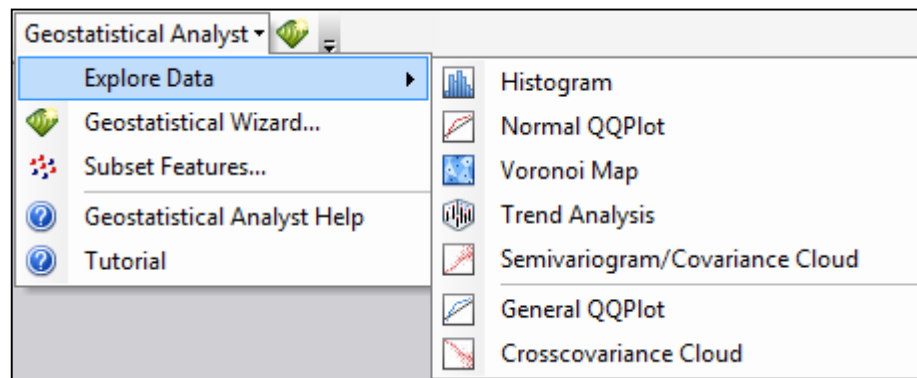


# ArcGIS Functions

## Statistical Analysis and Spatial Distributions

- ArcGIS provides methods to explore and model spatial relationships and patterns in data attributes
- Key concept is Tobler's first law of geography:

*"Everything is related to everything else, but near things are more related than distant things"*



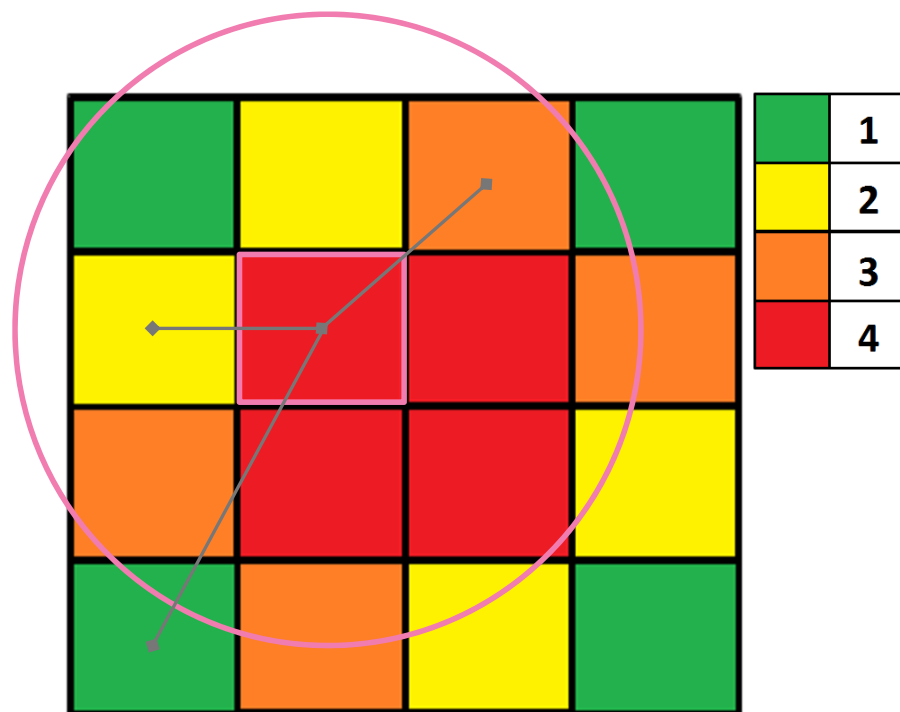
# ArcGIS Functions

## Statistical Analysis and Spatial Distributions

- Tobler's law provides a conceptual framework for finding notable features in the distribution of attributes, like hot spots
- Additionally, it is directly related to predominant methods of interpolation, used to "fill in" data where there are spatial gaps between observed data
- The Inverse Distance Weighted (IDW) interpolation method and its related functions directly assume that interpolated data points are more heavily influenced by nearby data
- Like many statistical modeling methods, care must be taken in defining a number of parameters of the estimation method(s) that may significantly alter the solution

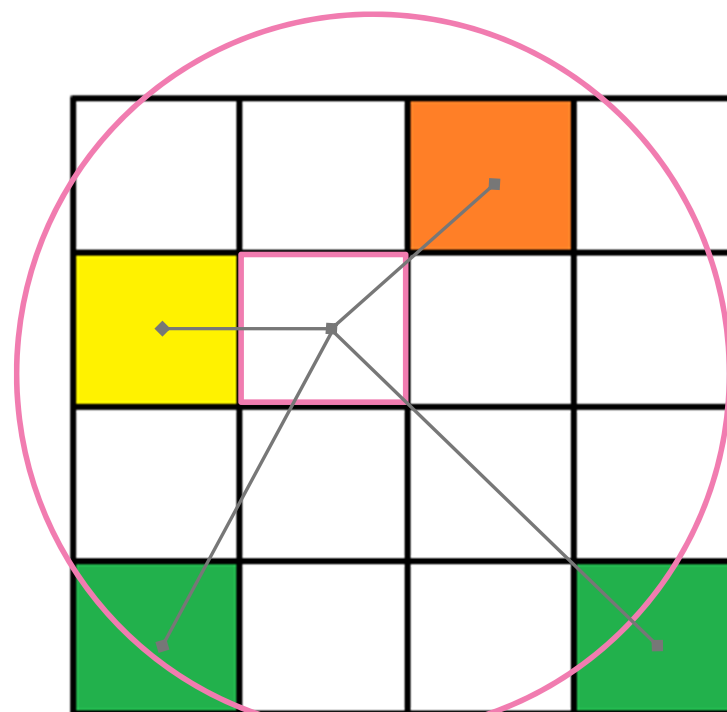
# ArcGIS Functions

## Statistical Analysis and Spatial Distributions



$$\frac{2 * \frac{1}{1} + 3 * \frac{1}{\sqrt{2}} + 1 * \frac{1}{\sqrt{5}}}{2 + 3 + 1}$$

0.76



$$\frac{2 * \frac{1}{1} + 3 * \frac{1}{\sqrt{2}} + 1 * \frac{1}{\sqrt{5}} + 1 * \frac{1}{\sqrt{8}}}{2 + 3 + 1 + 1}$$

0.70



# ARCGIS CUSTOM TOOLS

---

# ArcGIS Custom Tools

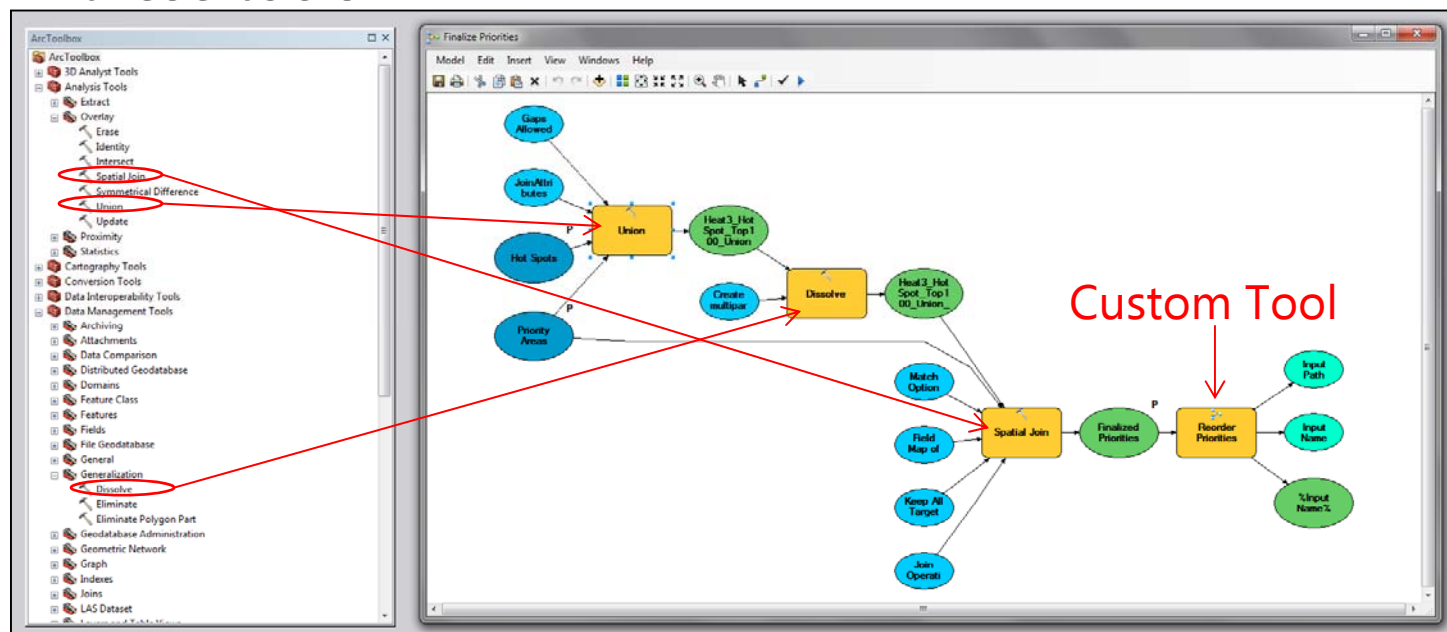
## Models

- The term “model” is a broad term used in many disciplines and can mean many different things depending on discipline, context, and individual interpretation
- For example, a model could be
  - A single equation representing the relationship between a dependent and independent variable
  - A set of equations that must be solved simultaneously
  - A statistical regression of the apparent distribution of attributes in a data set
  - An intricate, branching, possibly iterating set of code meant to represent processes at many scales for a complex (or even simple) system

# ArcGIS Custom Tools

## Custom Tools as a Shareable Analysis Workflow

- Though ArcGIS uses the term “model,” in ARB’s implementation, it may be more appropriate to think of Model Builder as a way to catalog and share analysis in a consistent and repeatable manner
- Model Builder allows an analyst to combine built-in ArcGIS tools (and their own custom tools) to perform tasks that are combinations of these tools



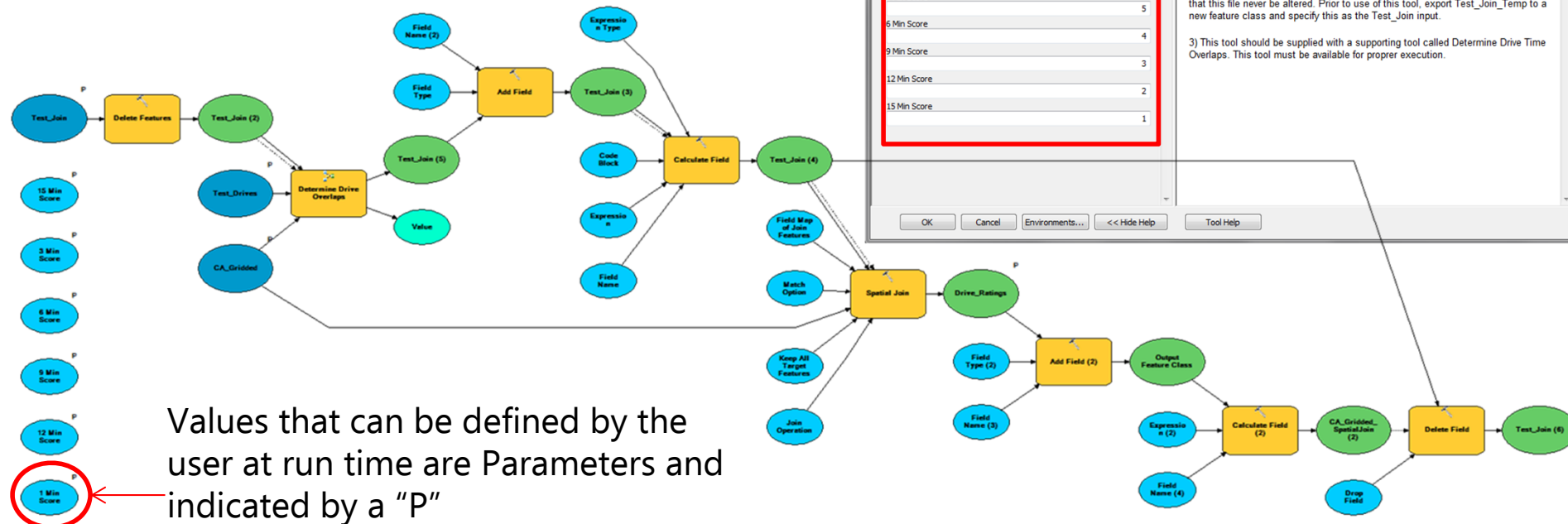
- CHIT\_Tools
- Determine Drive Overlaps
  - Determine Station Counts 6-Factor Match\_Norm
  - Determine Station Counts 6-Factor Match\_Norm\_Alt
  - Filter Priorities by Area
  - Finalize Priorities
  - Generate Coverage Factor
  - Generate Heat Map
  - Prioritize Heat Map Areas
  - Reorder Priorities
  - Statistical Hot Spots

# ArcGIS Custom Tools

## Building Custom Tools from Built-in and Custom Tools

- Creating custom tools also allows for documentation of the process and the opportunity to consistently apply key input values that can affect analyses

Default values for parameters can be stored with custom tools

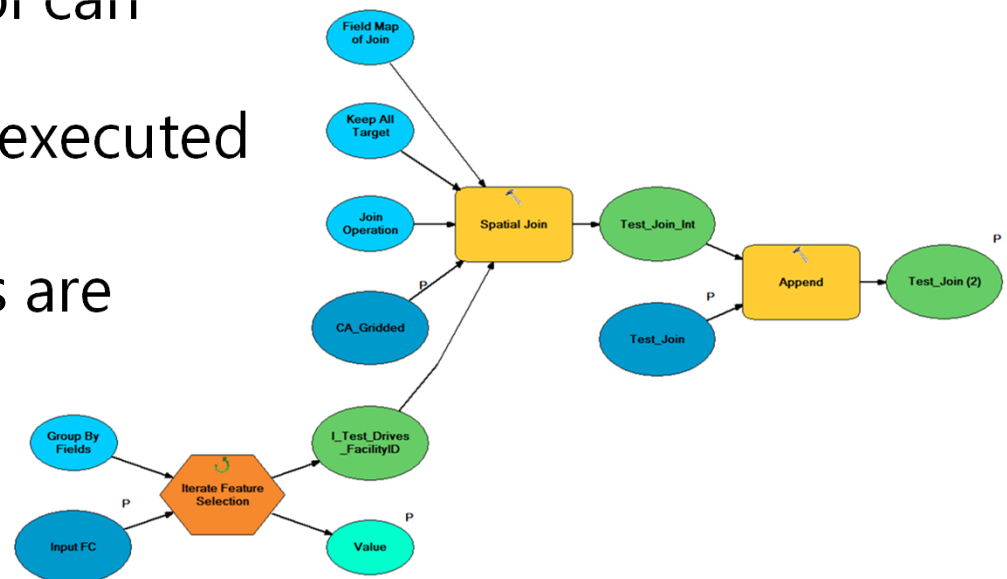




# ArcGIS Custom Tools

## Iteration

- At times, an analysis may require the same steps to be performed on every feature in a dataset separately (or every dataset in a database, etc...)
- In these cases, iteration is required; however, ArcGIS does have strict limits on the types of iteration available
- Additionally, each custom tool can only have one iterator and all operations in the tool will be executed in the iteration
- For this reason, iterative steps are often kept in separate tools and included as a step in other custom tools

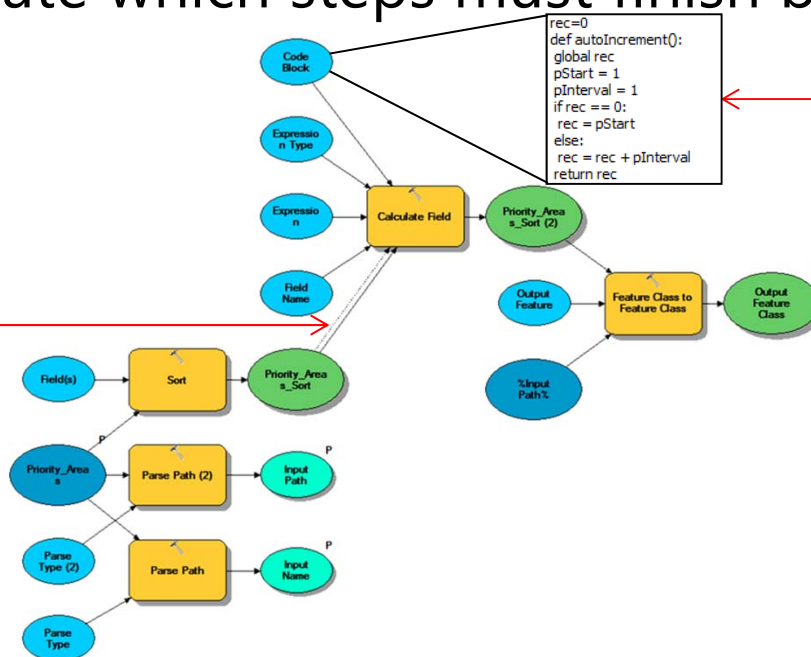


# ArcGIS Custom Tools

## Controlling Execution Flow

- Unlike other development languages or environments, ArcGIS does not initially assume the visual structure of the model indicates the intended order of execution
- For models that require sequential steps, preconditions must be put in place to indicate which steps must finish before another can begin

Preconditions are indicated with a dashed line in the Model Builder interface



In this example, the Sort step must occur before the field calculation because the expression depends on the index order of features in the data set



# CENSUS AND DMV DATA GEOGRAPHIES

---

# Data Source Geographies

## Census Data

- Census Data are available from American Fact Finder:  
<https://factfinder.census.gov>
- American Fact Finder provides data from multiple datasets
- Selection of dataset determined by geographic scale, attributes desired, and statistical certainty desired
- CHIT utilizes the American Community Survey (ACS), chosen for being a program with continual, annual updates and the richness of the attribute data available

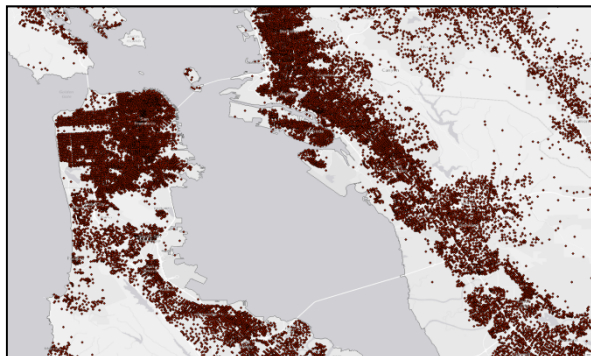
# Data Source Geographies

## Census Data

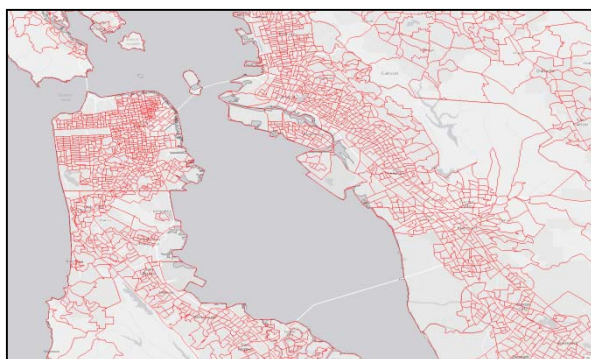
- ACS data are available on a variety of geographic scales. Choosing the right scale depends on the intent.
  - Block: The fundamental unit in census data sources and the only unit for which 100% data are published and collected (not just representative samples). Represented as points in the census data.
  - Blockgroup: The smallest geography for which representative sample data are published. Represented as polygons.
  - Tract: Roughly equivalent to a neighborhood; largely constant geography through time but may change. Represented as polygons.
  - County, State, etc...
  - Others : County Subdivision, Region, Urbanized Area, Metropolitan Statistical Area, Legislative Districts, etc...

# Data Source Geographies

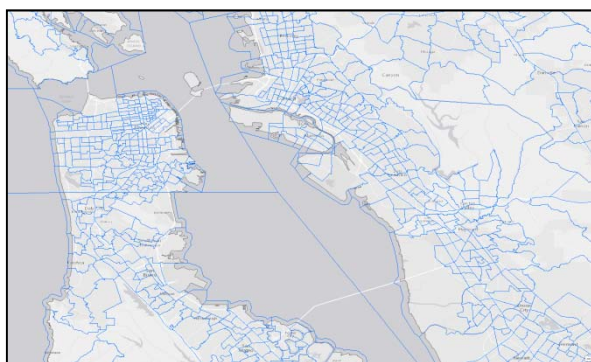
## Census Data



- Blocks: 410,559 in California
  - Most limited in attributes available; ARB considered best used for populations only



- Block Groups: 23,212 in California
  - Typically 600 to 3,000 people
  - Income data aggregates limited to:
    - Aggregate (Sum), Median
  - Average area: ~7 sqmi



- Tracts: 8,057 in California
  - Typically 1,200 to 8,000 people; optimum 4,000
  - Income data available in a variety of forms
    - Aggregate (Sum), Median, Mean, Quintile Mean, Quintile Bounds, Quintile Share
    - Quintile data includes top 5%
  - Average area: ~20 sqmi

# Data Source Geographies

## Census Data

- Counts in the form of aggregate sums or means are useful as directly-usable data points when other estimations do not need to be made
  - Cannot use aggregates alone to develop an understanding of the shape of the population's distribution of an attribute
    - Ex: The mean of income alone cannot help determine the cutoff for the top 20% without significant additional assumptions
- Descriptive statistic attributes beyond counts allow for statistical modeling, interpolation, and extrapolation
  - Will see in afternoon that this was important for implementing income data in particular

# Data Source Geographies

## Department of Motor Vehicles Data

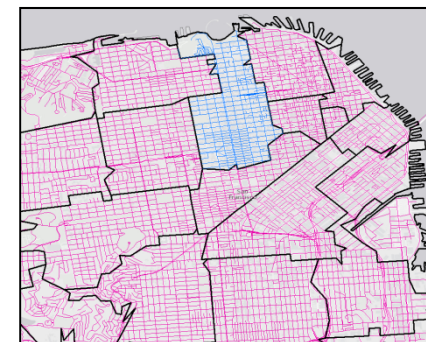
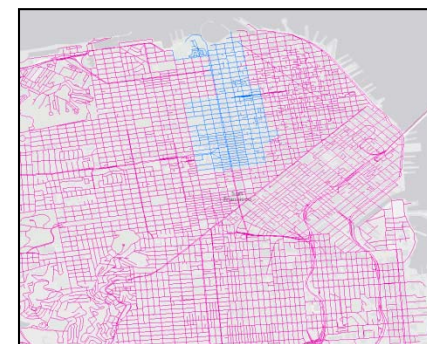
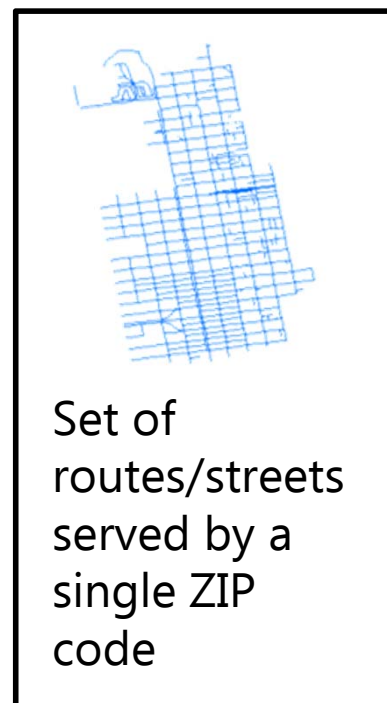
- DMV data are based on registrations of vehicles
- For tools or analyses that can remain confidential, data could be analyzed at resolutions down to the street address (point resolution even finer than Census blocks)
- CHIT is intended to be public and transparent; non-confidential data is available from DMV only at the ZIP code level



# Data Source Geographies

## Department of Motor Vehicles Data

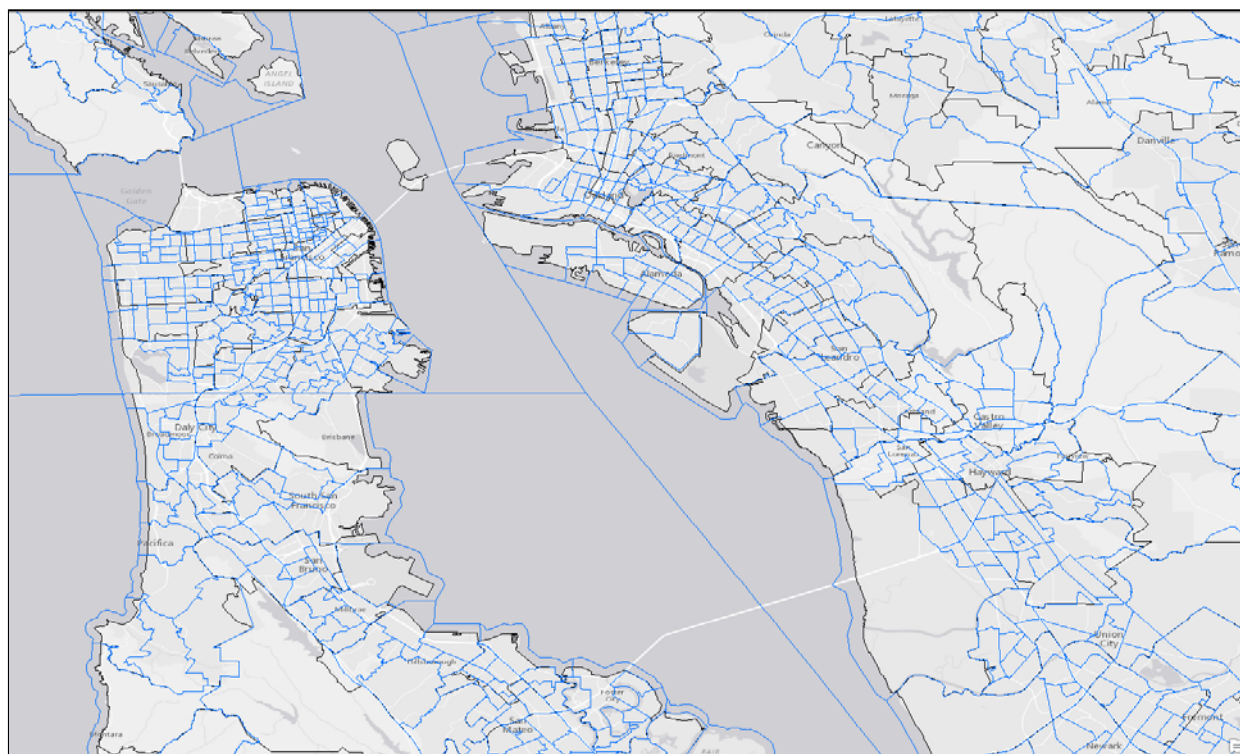
- ZIP codes are often conceptualized as polygons
- However, ZIP codes are actually defined by a set of routes
- Defining a bounding polygon is not a trivial matter and can have multiple equivalently valid solutions
- US Census has developed a standardized set of polygons (ZCTA: Zip Code Tabulation Area) that ARB has adopted for use in CHIT to define ZIP codes



# Data Source Geographies

## Department of Motor Vehicles Data

- 1,808 ZIP Codes represented in DMV data
- ZCTAs are typically larger than tracts, though may be similar in size in urban areas



# Data Source Geographies

## Summary of Spatial Bases Used

- Population:
  - Block (Highest Resolution Available)
- Income:
  - Tract used for modeling of distribution of income and education
- DMV Data:
  - ZIP codes (as represented by ZCTAs)
- OEM Surveys:
  - Counties or Statewide



Increasing  
Size

leads to...

Greater  
Statistical  
Certainty,  
Lower  
Resolution

# Data Source Geographies

ACS Multi-Year Estimates

- ACS data are provided in 1-, 3-, and 5-year aggregates
- Aggregating years allows for inclusion of more samples, which increases statistical certainty
- For similar reasons, the Census provides aggregated data on increasingly large geometries as fewer years included
- CHIT prioritizes certainty over currency



Including  
more years

leads to...

Greater  
Statistical  
Certainty,  
Less  
Current

# Wrap-Up

- Covered a variety of topics that will be instrumental in the main technical formulation discussion and further exploration of CHIT
  - Brief Introduction and review of CHIT and AB 8 process
  - Introduction to GIS fundamentals
  - ArcGIS tools utilized in formulation of CHIT
  - Creating custom tools in ArcGIS
  - Working with geographies in Census and DMV data sources
- What are some of the major built-in ArcGIS tools used in CHIT and what do they accomplish?
- What are the differences in the spatial structure of the data sources and how are they reconciled?
- How does ArcGIS assess spatial distribution of data? What statistical analyses are applied?
- How is a tool like CHIT made in ArcGIS?



# DISCUSSION

---

For questions or comments, contact:

Andrew Martinez

(916) 322-8449

[andrew.martinez@arb.ca.gov](mailto:andrew.martinez@arb.ca.gov)