ArcGIS Pro: Geoprocessing in Parallel Using Apache Spark

Sarah Ambrose and Bethany Scott

2021 ESRI
DEVELOPER SUMMIT

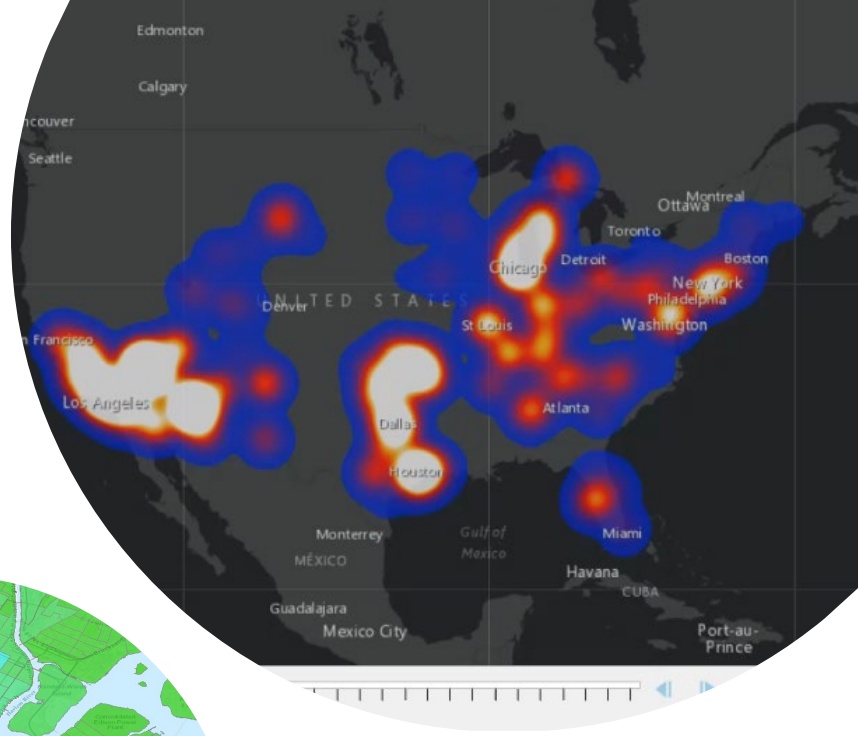
Have you ever had analysis fail because your data was too big?

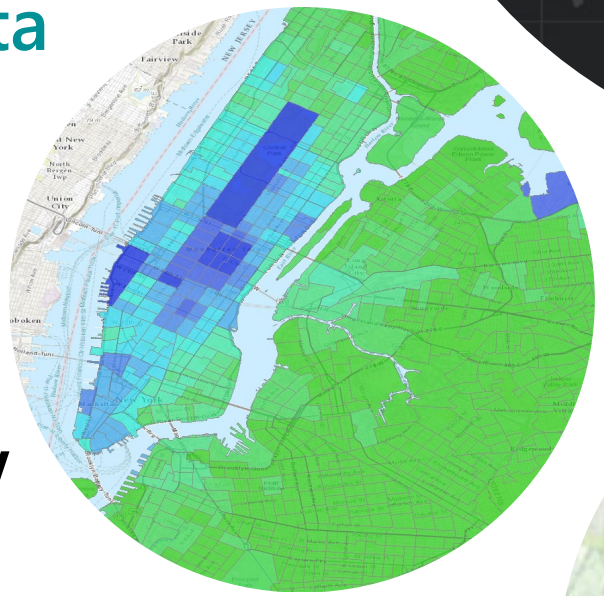
Have you ever waited for a tool to finish, and it took forever?

What is GeoAnalytics?

A toolbox that *parallelizes* computation to quickly analyze large amounts of vector and tabular data

A collection of analysis tools to identify *patterns*, *relationships*, *anomalies* and *incidents* in large amounts of data across *space and time*





How is it faster?

Why use GeoAnalytics?

- Existing tools and workflows aren’t processing data fast enough
- I need to distill my data into something more manageable to use in other analysis
- My data has a ton of noise and I want to explore it to bring out what’s important
- I have datasets that I’m struggling to load into my GIS
- Some of the tools are really cool!

What kind of analysis can I run?

- Which stationary pressure sensors in my pipe network have experienced anomalous events in the past 24 hours? Where are there hot spots of anomalous events?
- Where have my delivery trucks traveled and where is the highest density of unique delivery truck paths? Where do delivery trucks travel the slowest?
- Where and when are events happening close together in space and time?

What kind of analysis can I run?

- Which stationary pressure sensors in my pipe network have experienced anomalous events in the past 24 hours? Where are there hot spots of anomalous events?
- Where have my delivery trucks traveled and where is the highest density of unique delivery truck paths? Where do delivery trucks travel the slowest?
- Where and when are events happening close together in space and time?

What kind of analysis can I run?

- Which stationary pressure sensors in my pipe network have experienced anomalous events in the past 24 hours? Where are there hot spots of anomalous events?
- Where have my delivery trucks traveled and where is the highest density of unique delivery truck paths? Where do delivery trucks travel the slowest?
- Where and when are events happening close together in space and time?

How do I get it?

GeoAnalytics Server

Distributed processing across multiple *server cores and machines* with ArcGIS Enterprise

Requirements: Enterprise + GeoAnalytics Server License

GeoAnalytics Desktop

Parallel processing across cores on your laptop or desktop with ArcGIS Pro

Requirements: Advanced License

When to use Desktop or Server

- Use GeoAnalytics Server when you want to:
  - Bring big data analysis to your entire organization
  - Leverage the power of one or multiple server machines
  - Connect to external big data storage and existing web layers
  - Extend using custom analysis

- Use GeoAnalytics Desktop when you want to:
  - Process local data (from files, databases) faster than before on your own desktop machine
  - Prototyping workflows you want to use with GeoAnalytics Server
  - Connect to and use *big data connections*

When to use Desktop or Server

	GeoAnalytics Server (Enterprise 10.9)	GeoAnalytics Desktop (Pro 2.7)
Input data	- Big data file shares * - Hosted feature layers - Feature services	- File + enterprise geodatabase - Shapefiles - Big data connections (BDC)
Output data	- Hosted feature layers - Big data file shares *	- File and enterprise geodatabase - Shapefiles
Scaling out analysis	- Control the number of machines - Control the percentage of cores and RAM - Scale out data storage with spatiotemporal data store	- One machine only - Control the percentage of RAM (or a value)
Tools	- 28 tools - Run Python Script	- 22 tools - 7 tools for BDC management
Interface	- REST + the ArcGIS API for Python - Pro and Arcpy (and model builder) - Portal Map Viewer	- Pro and Arcpy (and model builder)

Blog post covering this topic

Demo
Running a GeoAnalytics Tool

Bethany Scott

Analysis Overview

Available Tools

Analyze Patterns

Calculate Density
Find Hot Spots
Find Point Clusters
Forest-based Classification and Regression
Generalized Linear Regression

Use Proximity

Create Buffers
Trace Proximity Events

Manage Data

Calculate Field
Clip Layer
Dissolve Boundaries
Overlay Layers

Summarize Data

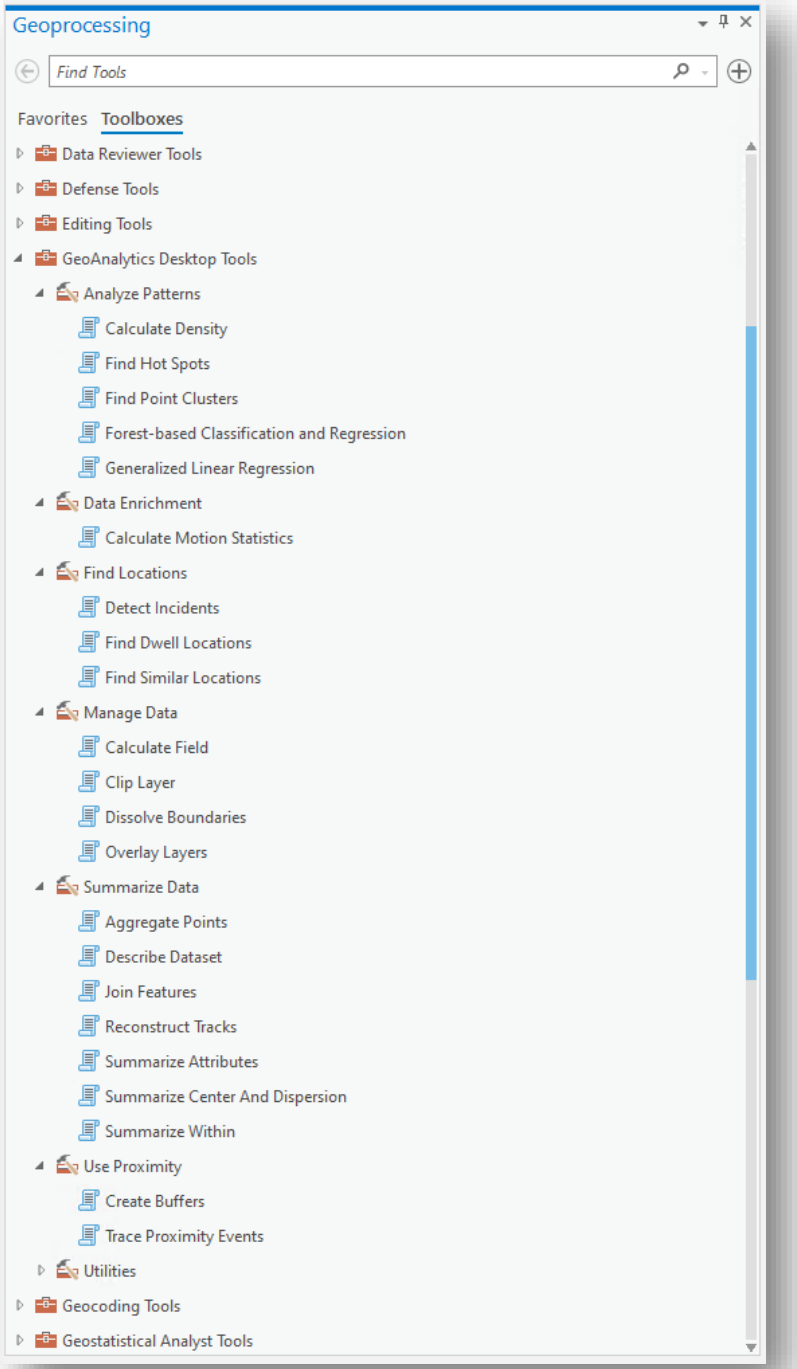
Aggregate Points
Describe Dataset
Join Features
Reconstruct Tracks
Summarize Attributes
Summarize Center and Dispersion
Summarize Within

Find Locations

Detect Incidents
Find Dwell Locations
Find Similar Locations

Data Enrichment

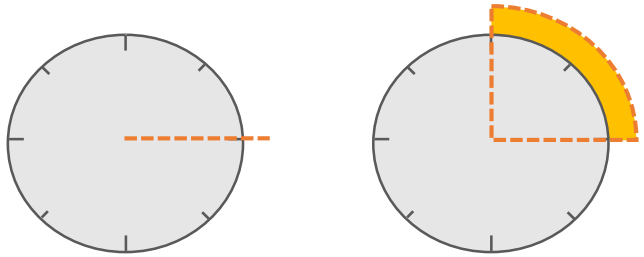
Calculate Motion Statistics



Analysis Capabilities

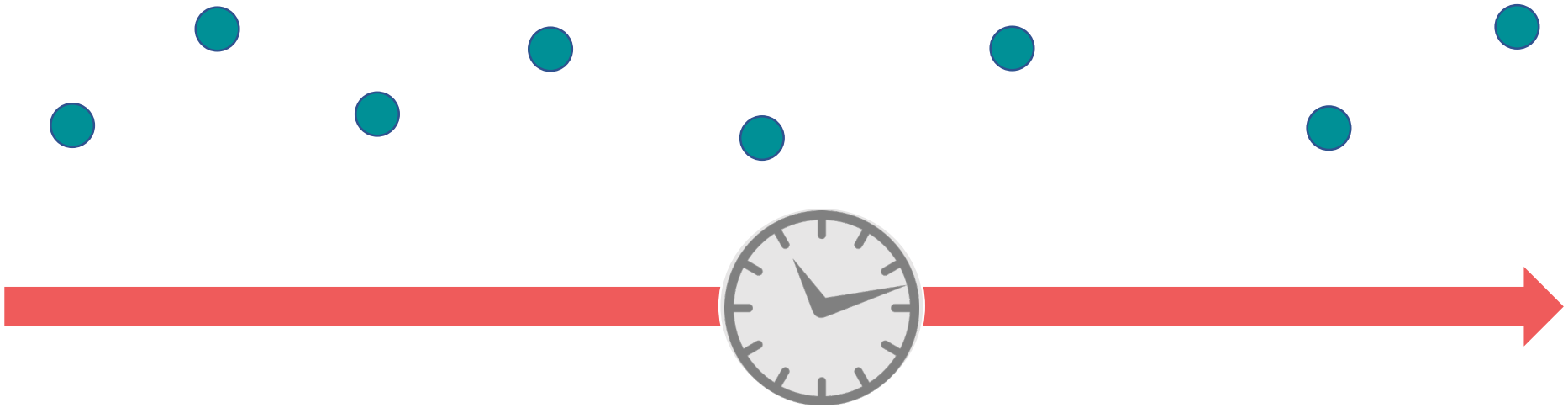
Work with data in both space and time

- Use GeoAnalytics to perform spatiotemporal analysis
- Define your temporal input data:
  - Instants (moment in time)
  - Interval (a duration in time)



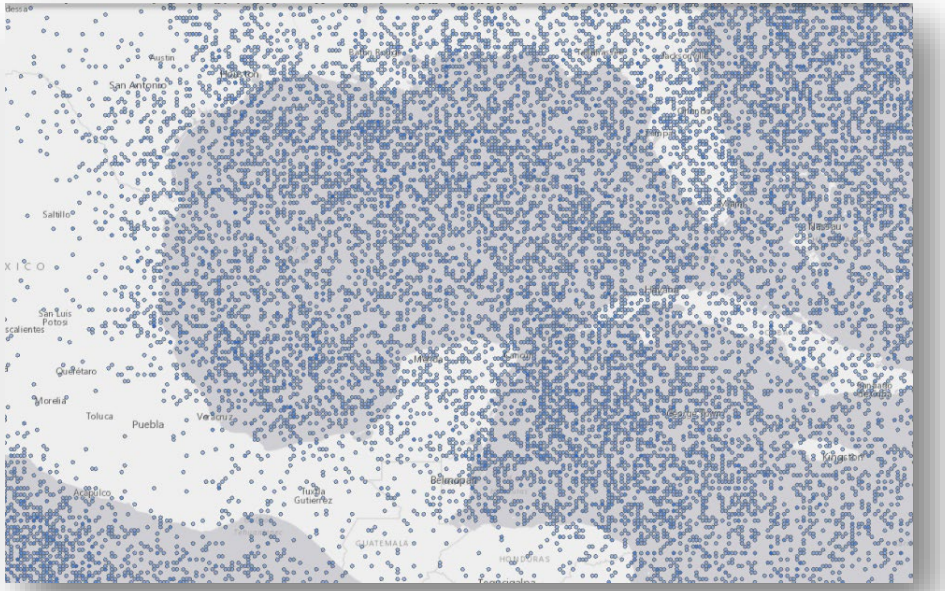
- Visualize results across time using the time slider

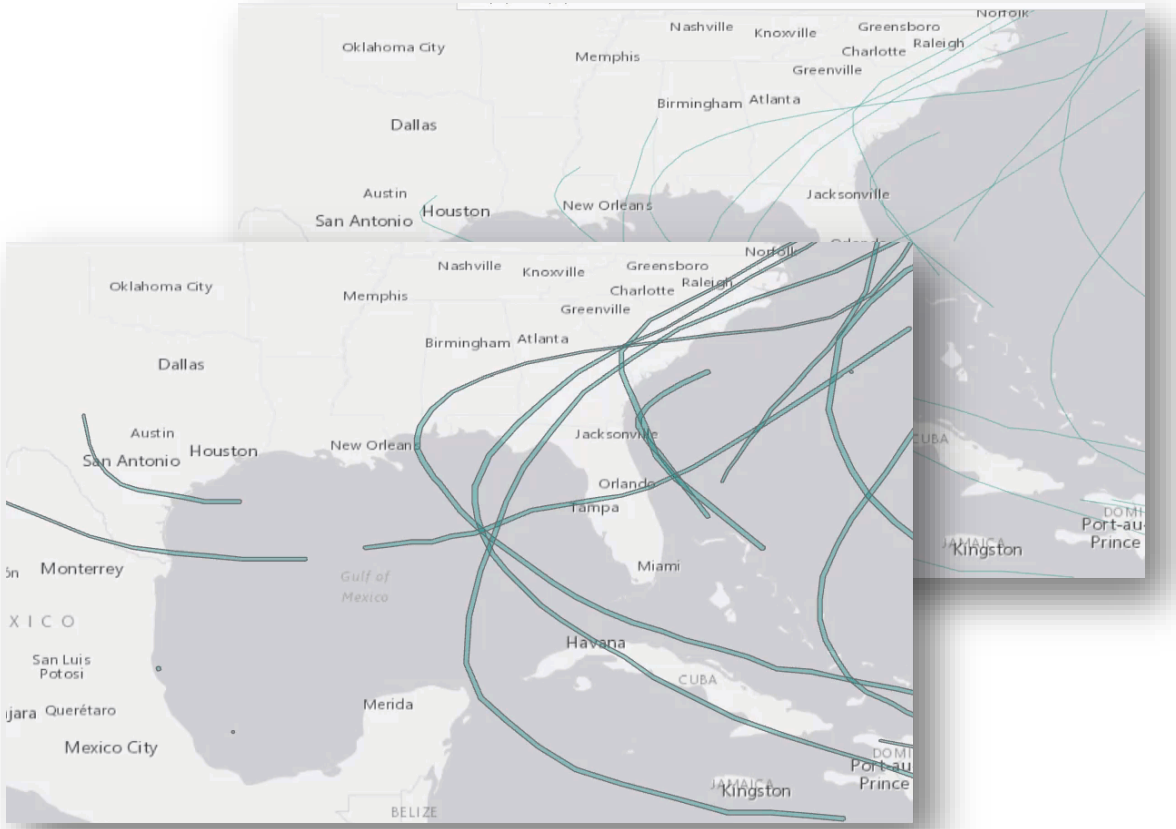
Track Analysis



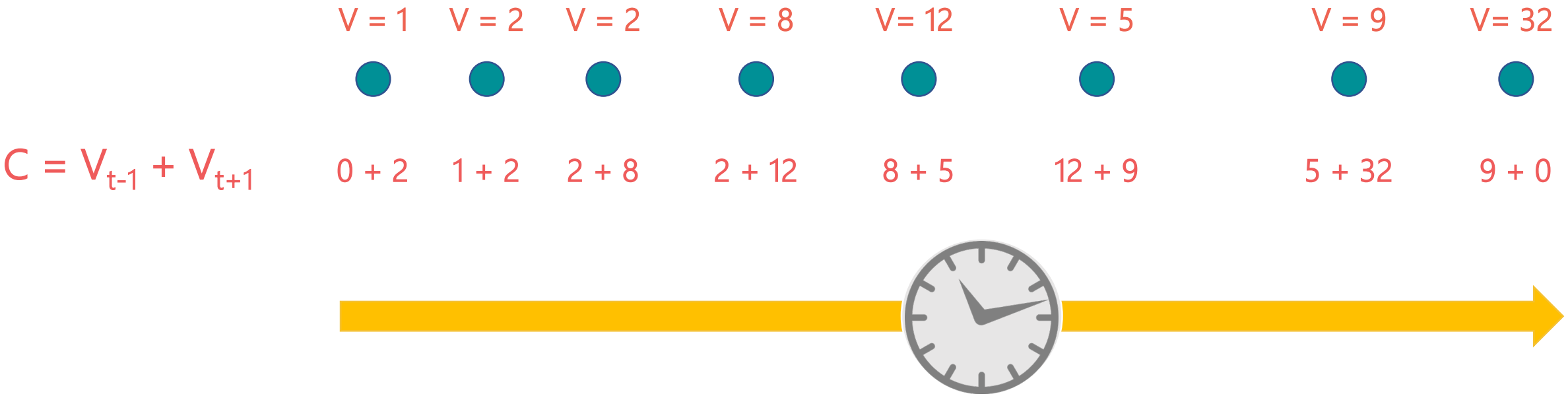
Measurements recorded over time

Visualize where tracks have gone using breadcrumbs as inputs (Reconstruct Tracks)

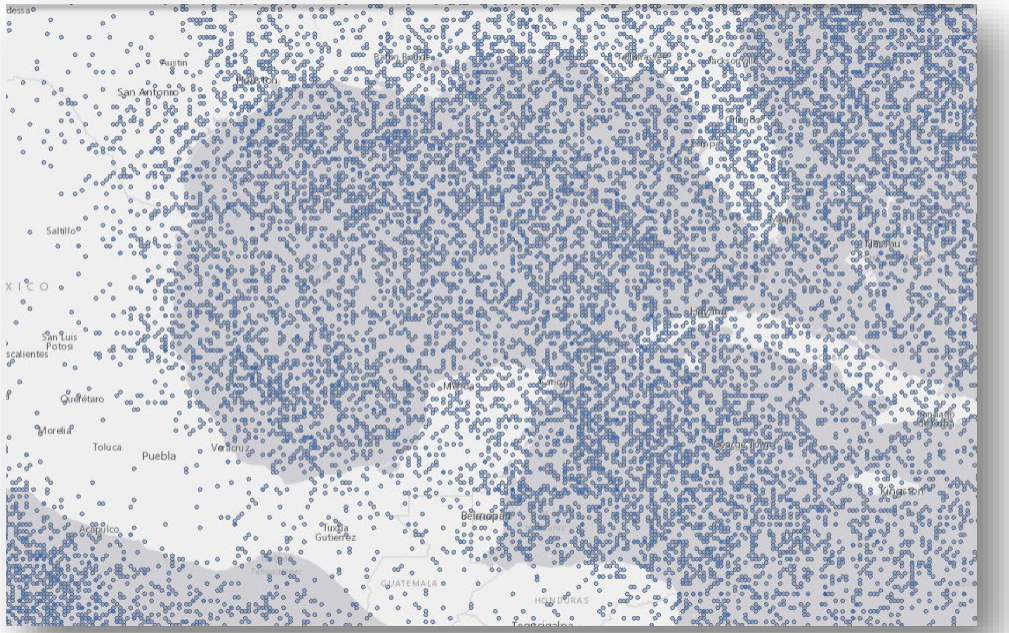


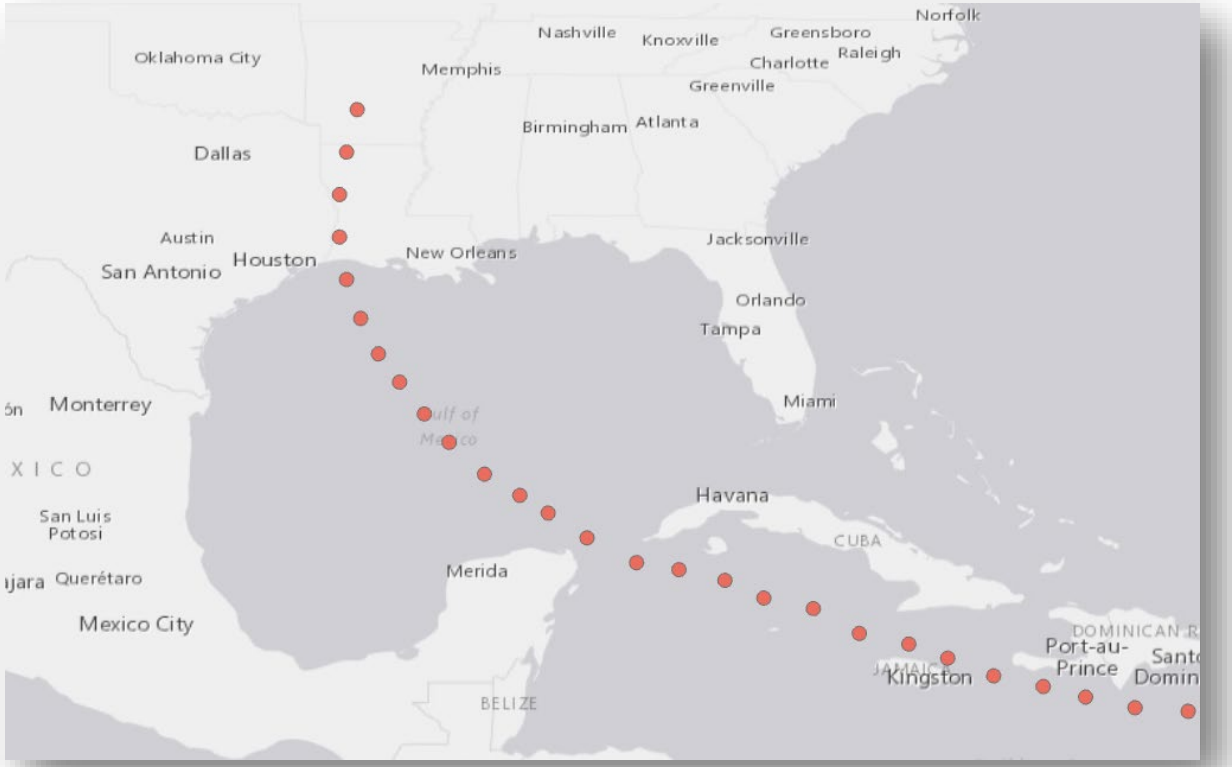


Calculate values and conditions based on previous or subsequent values in a track (Calculate Field)



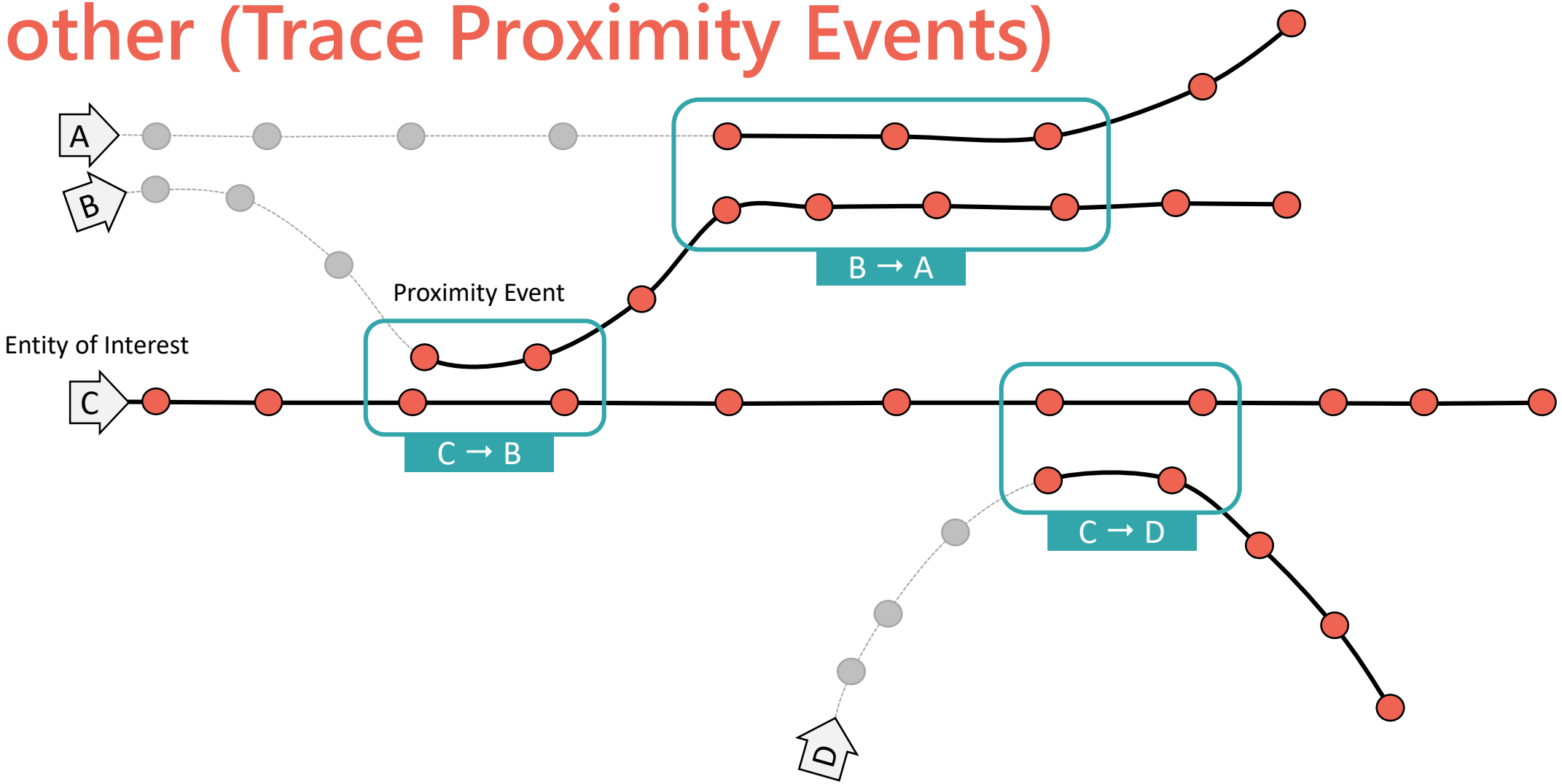
Find features that meet a specified condition (Detect Incidents)



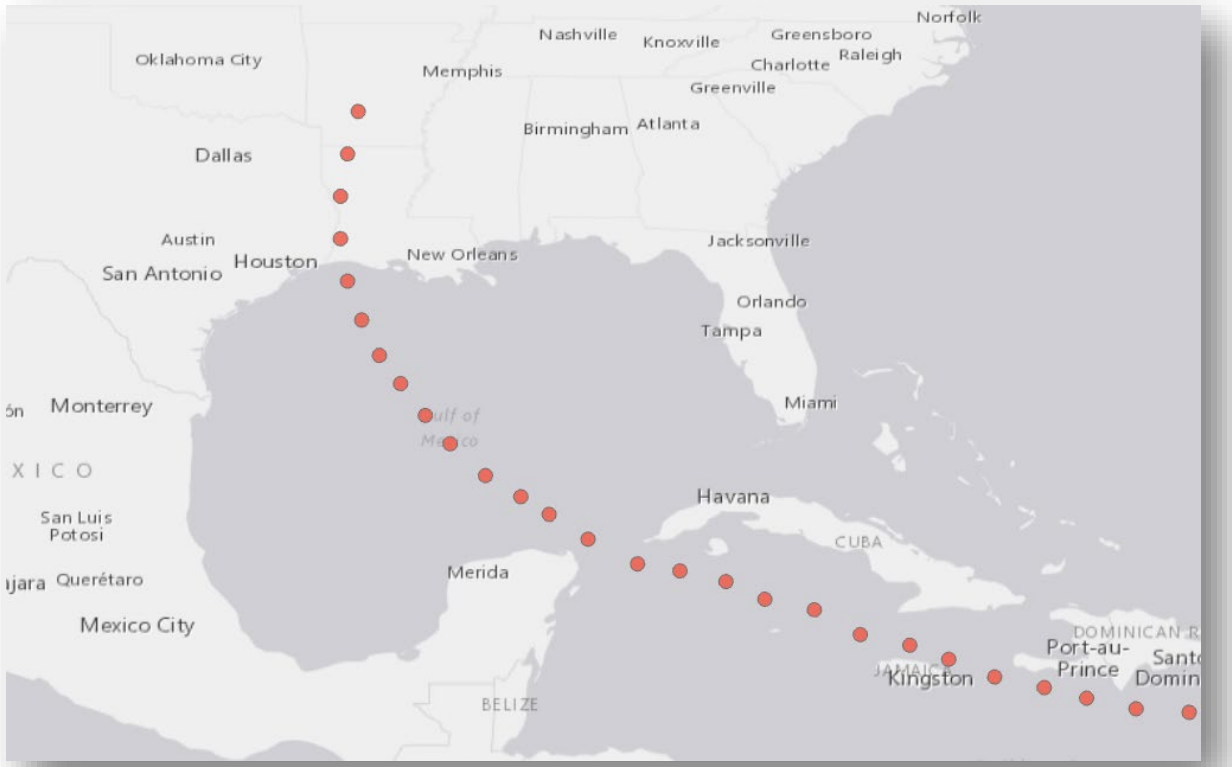


Find where objects are remaining stationary or near-stationary (Find Dwell Locations)

Find downstream events that happened within a spatiotemporal proximity of each other (Trace Proximity Events)

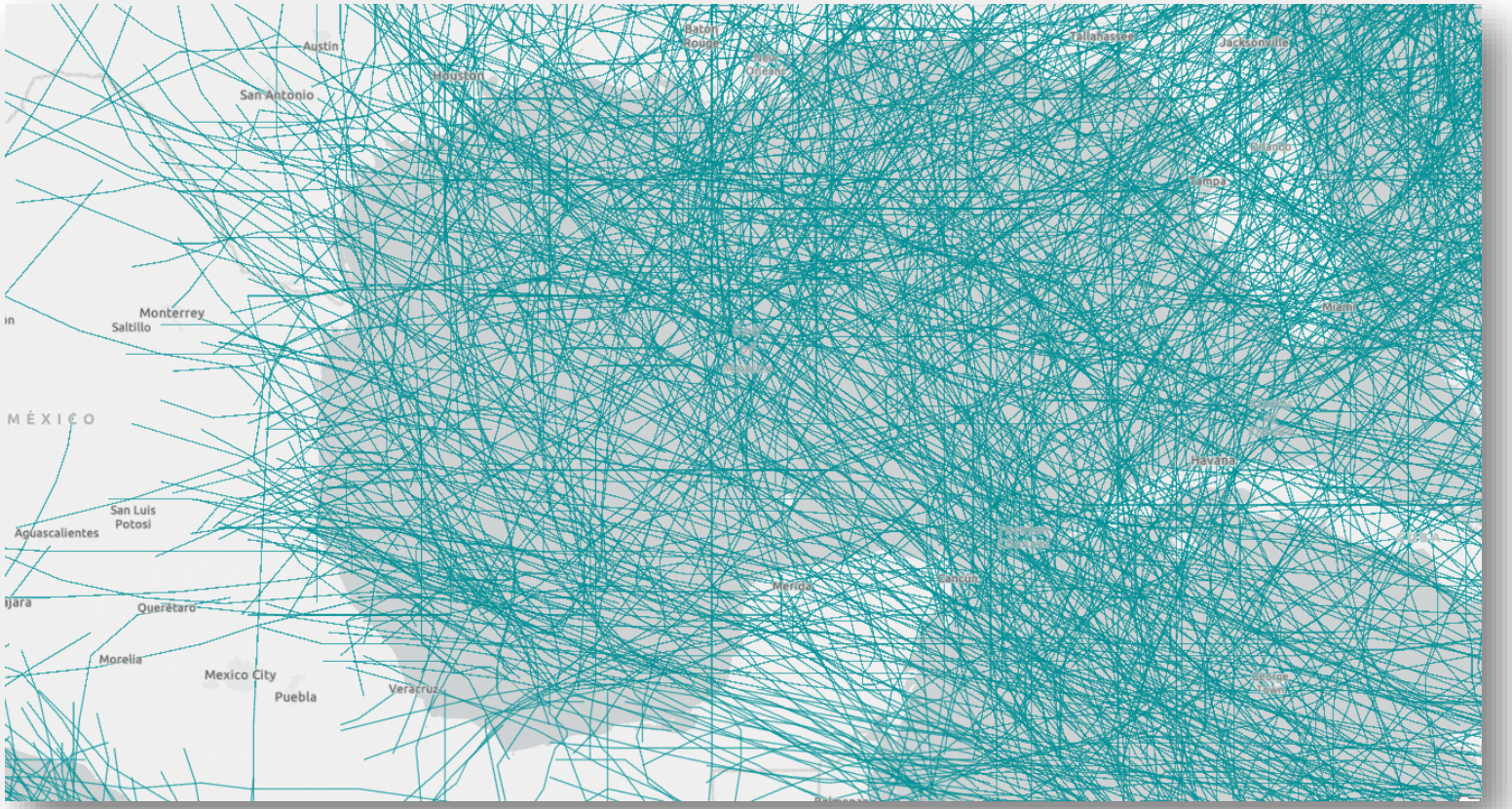


Enrich track data with movement statistics (Calculate Motion Statistics)



- Distance
- Duration
- Speed
- Acceleration
- Elevation
- Slope
- Bearing
- Idle

Summarize where your assets have traveled into bins to get a track “heat map” (Reconstruct Tracks + Summarize Within)

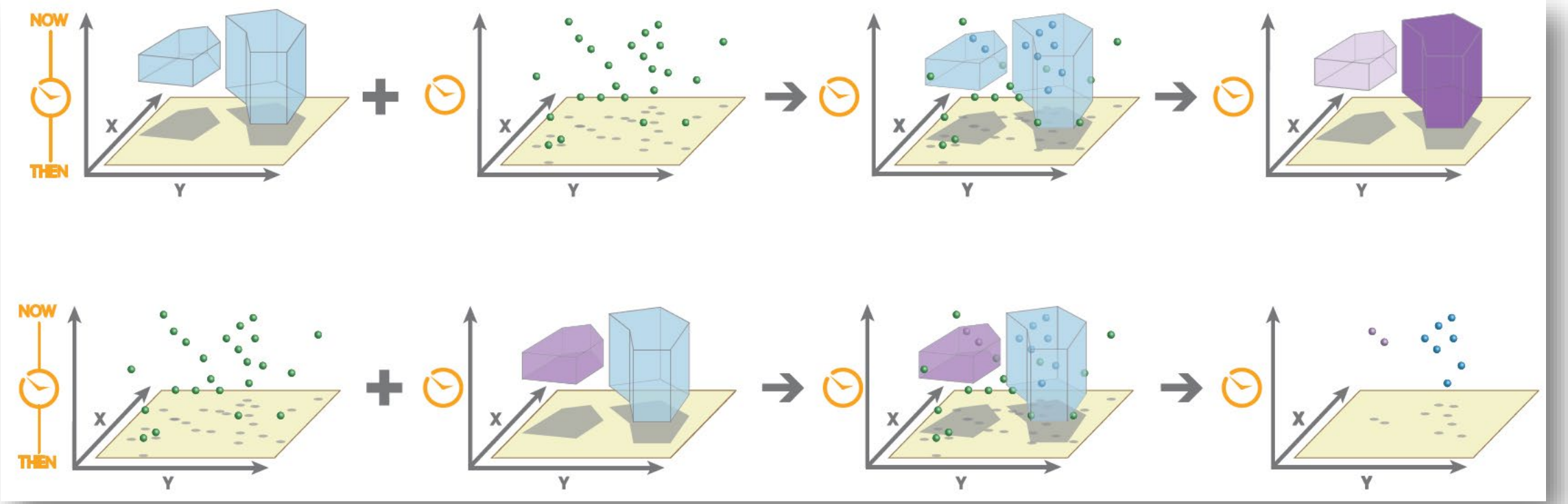


Additional Space-Time Analysis

- Spatiotemporal Joins
  - Join Features
- Time Stepping
  - Aggregate Points
  - Calculate Density
  - Find Hot Spots
  - Create Space Time Cube
- Spatiotemporal Clustering
  - Find Point Clusters

Spatiotemporal Joins

Join features based on their relationships in time

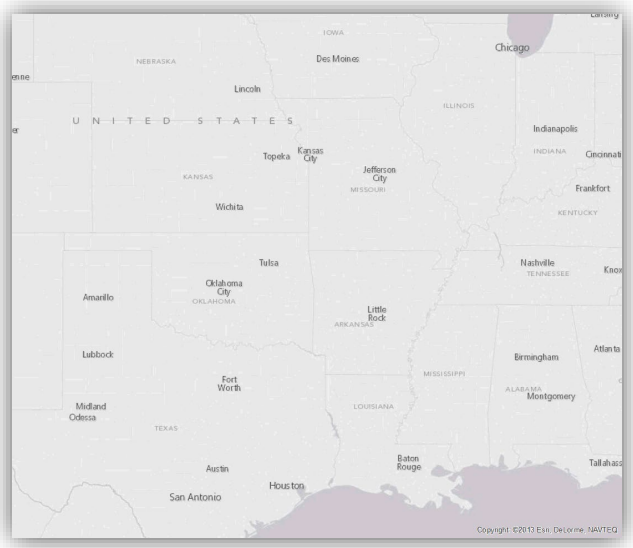


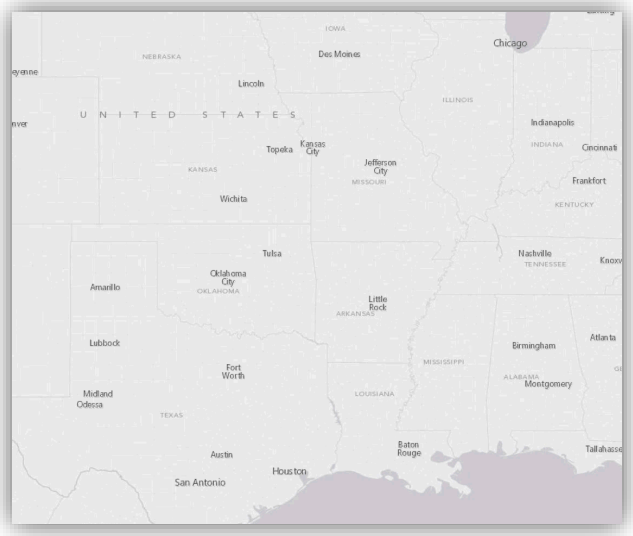
Space-time join of (a) polygon to point features and (b) point to polygon features

Time Stepping

Analyze your data at specified intervals:

- Interval (the duration)
- Repeat (how often you do it)
- Reference (how you align it)





Spatiotemporal Clustering

Find clusters of points based on:

- A minimum number of features in each cluster
- A distance
- A time distance

Demo
Track Analysis

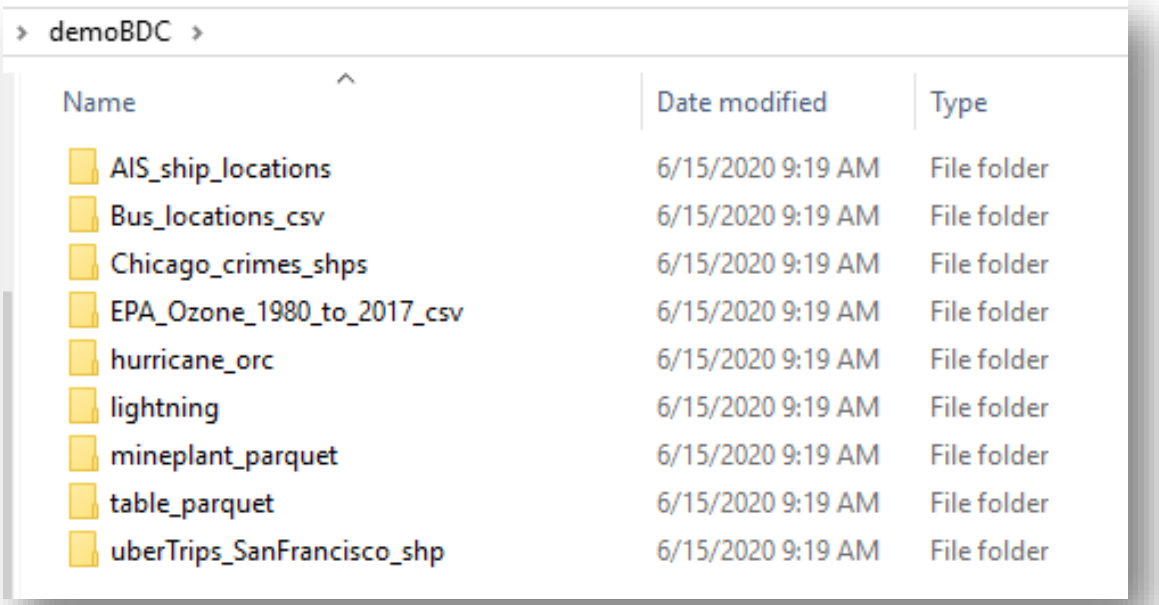
Bethany Scott

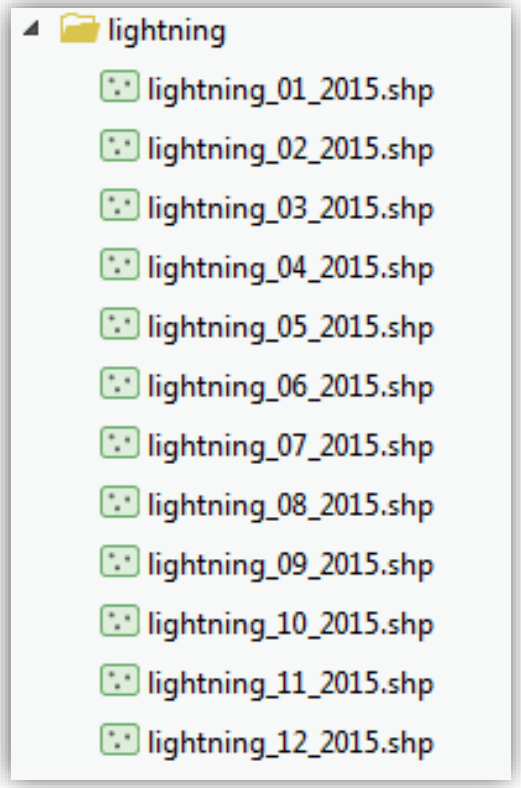
Big Data Connections (BDCs)

BDC details

Read directly from collections of datasets in a directory:

- Delimited files
- Shapefiles
- Parquet
- ORC





Examples of when to use BDCs

- You can represent multiple datasets of the same schema and file type as a single dataset.
- A BDC accesses the data when the analysis is run, so you can continue to add data to an existing dataset in your BDC without reregistering or publishing your data.
- You can modify the BDC to remove, add, or update which datasets are visible.
- BDCs are flexible in how time and geometry can be defined and allow for multiple time formats on a single dataset.

Examples of when to use BDCs

- You can represent multiple datasets of the same schema and file type as a single dataset.
- A BDC accesses the data when the analysis is run, so you can continue to add data to an existing dataset in your BDC without reregistering or publishing your data.
- You can modify the BDC to remove, add, or update which datasets are visible.
- BDCs are flexible in how time and geometry can be defined and allow for multiple time formats on a single dataset.

Examples of when to use BDCs

- You can represent multiple datasets of the same schema and file type as a single dataset.
- A BDC accesses the data when the analysis is run, so you can continue to add data to an existing dataset in your BDC without reregistering or publishing your data.
- You can modify the BDC to remove, add, or update which datasets are visible.
- BDCs are flexible in how time and geometry can be defined and allow for multiple time formats on a single dataset.

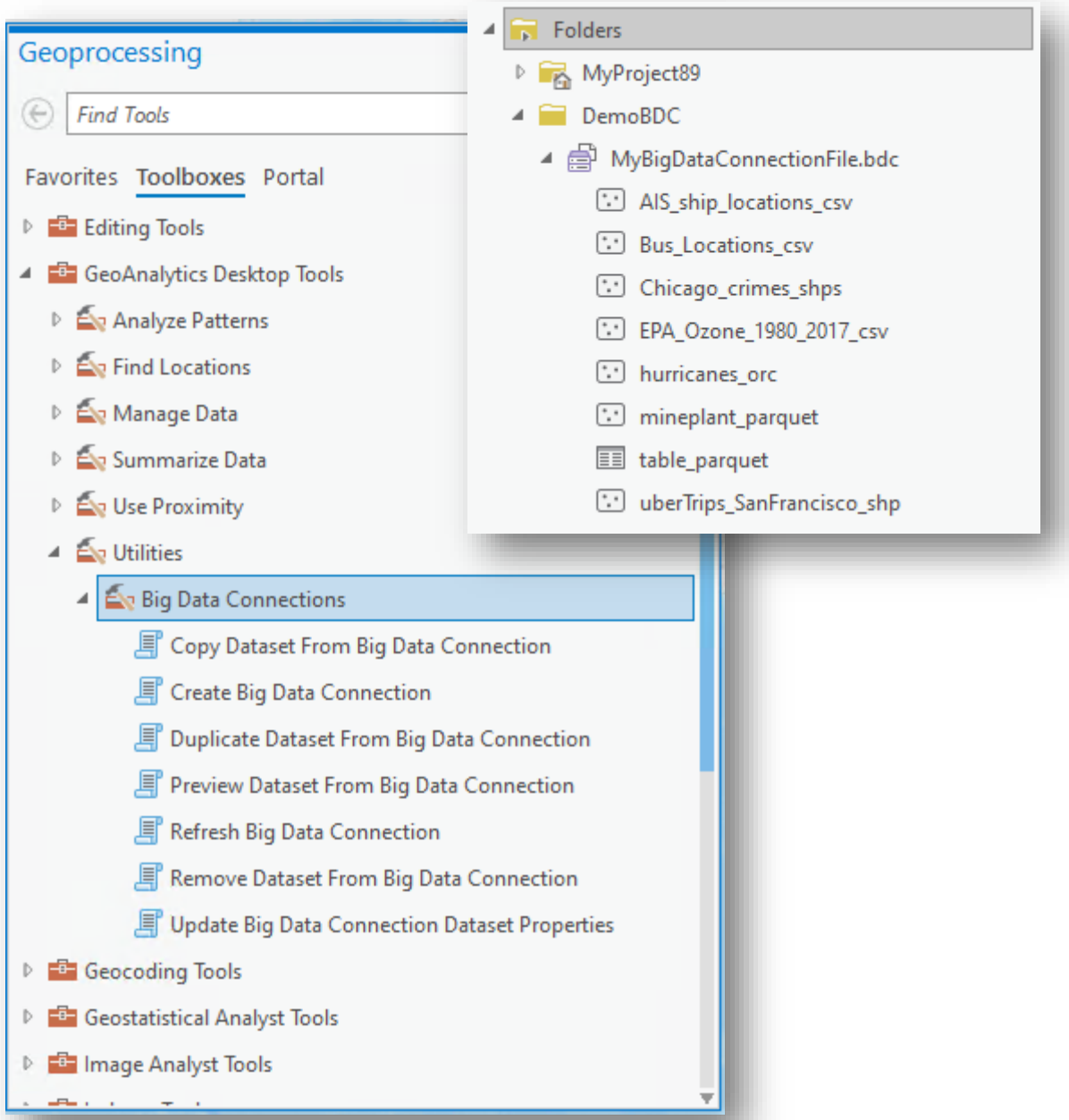
Examples of when to use BDCs

- You can represent multiple datasets of the same schema and file type as a single dataset.
- A BDC accesses the data when the analysis is run, so you can continue to add data to an existing dataset in your BDC without reregistering or publishing your data.
- You can modify the BDC to remove, add, or update which datasets are visible.
- **BDCs are flexible in how time and geometry can be defined and allow for multiple time formats on a single dataset.**

Available Tools

Big Data Connections

Copy Dataset from Big Data Connection
Create Big Data Connection
Duplicate Dataset From Big Data Connection
Preview Dataset From Big Data connection
Refresh Big Data Connection
Remove Dataset From Big Data Connection
Update Big Data Connection Dataset Properties



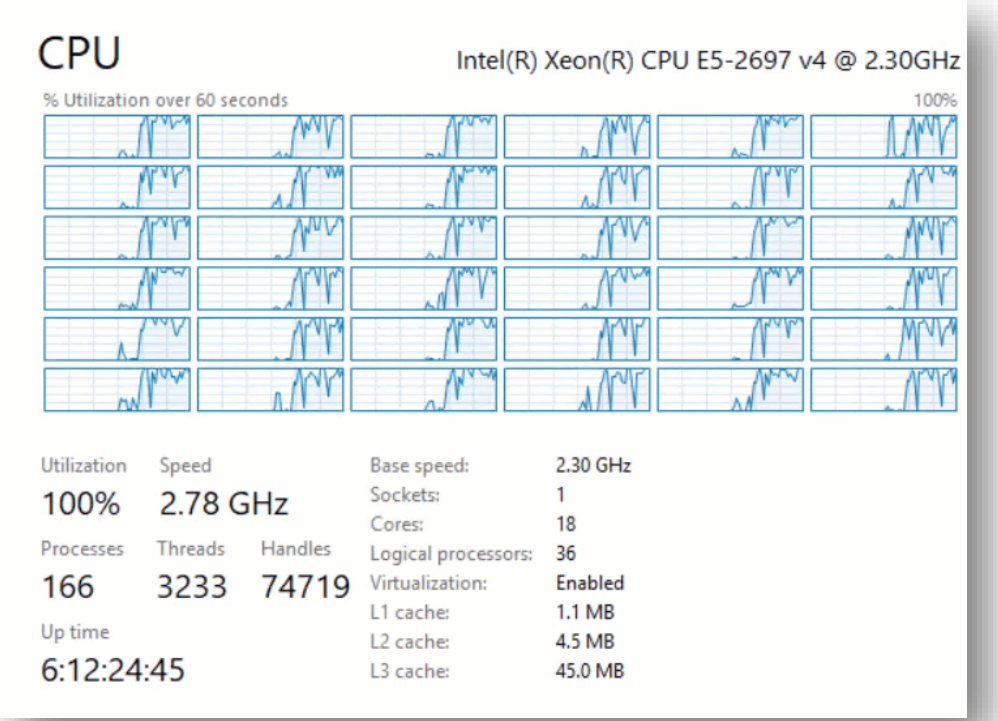
Demo
Connect to your big data

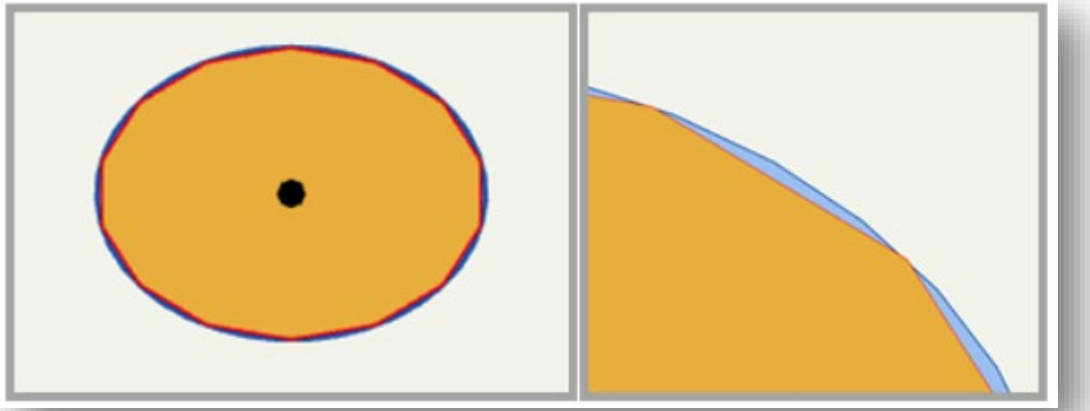
Bethany Scott

Considerations when running GeoAnalytics Desktop Tools

Need to knows

- GeoAnalytics Tools use Spark
  - Uses 95% of the memory on your machine and all cores
- Processing factor
  - Control the core usage
- Start up time
  - GeoAnalytics takes about 5-15 seconds to start Spark
- Different implementation of tools
  - Results won't be exactly the same as similar tools in Pro





Performance

Duration of the tool depends on multiple factors:

Input Data	Data Sources	Tool + Parameters	Hardware
Number of features	Colocation	Operation being used	RAM
Geometry Type	File type	Many neighbors?	Drive space and speed
Number of Vertices		Distribute the data	Cores

Performance

Duration of the tool depends on multiple factors:

Input Data	Data Sources	Tool + Parameters	Hardware
Number of features	Colocation	Operation being used	RAM
Geometry Type	File type	Many neighbors?	Drive space and speed
Number of Vertices		Distribute the data	Cores

Performance

Duration of the tool depends on multiple factors:

Input Data	Data Sources	Tool + Parameters	Hardware
Number of features	Colocation	Operation being used	RAM
Geometry Type	File type	Many neighbors?	Drive space and speed
Number of Vertices		Distribute the data	Cores

Performance

Duration of the tool depends on multiple factors:

Input Data	Data Sources	Tool + Parameters	Hardware
Number of features	Colocation	Operation being used	RAM
Geometry Type	File type	Many neighbors?	Drive space and speed
Number of Vertices		Distribute the data	Cores

RAM Details

Spark uses memory (RAM) to run analysis

- Increase the memory if possible (64 > 32 > 16 GB RAM)
- If the job can’t be completed in memory, it will spill to disk
  - Uses your temp drive
  - If you run out of space on the temp drive, you can modify it using your Windows settings

What if I need to run on even bigger data?

Bigger Data?

GeoAnalytics Server

Distributed processing across multiple *server cores and machines* with ArcGIS Enterprise

GeoAnalytics Desktop

Parallel processing across cores on your laptop or desktop with ArcGIS Pro

Summary

- Accelerated: Speeds up analytical processing time using built-in parallel compute
- Spatiotemporal: Many tools are designed to analyze data in space and time
- Integrated: Works as is in ArcGIS Pro with your data in ArcGIS Pro!
- Actionable: Able to crunch through large volumes of data to generate actionable insights and intelligence. Enabling organizations to visualize & react to large amount of data in a clearer and more meaningful way.

Helpful Links

Blog post introducing GeoAnalytics in Pro

GeoAnalytics Desktop Documentation

Dev Summit 2020 Plenary Blog – Spatiotemporal clustering

Blog: Heat mapping with GeoAnalytics

Email us at geoanalytics-pes@esri.com

esri®

THE SCIENCE OF WHERE®

Copyright © 2021 Esri. All rights reserved.

Please provide your feedback for this session by clicking on the session survey link directly below the video.