Big Data and Analytics: Getting Started with ArcGIS

Mike Park

Erik Hoel

Agenda

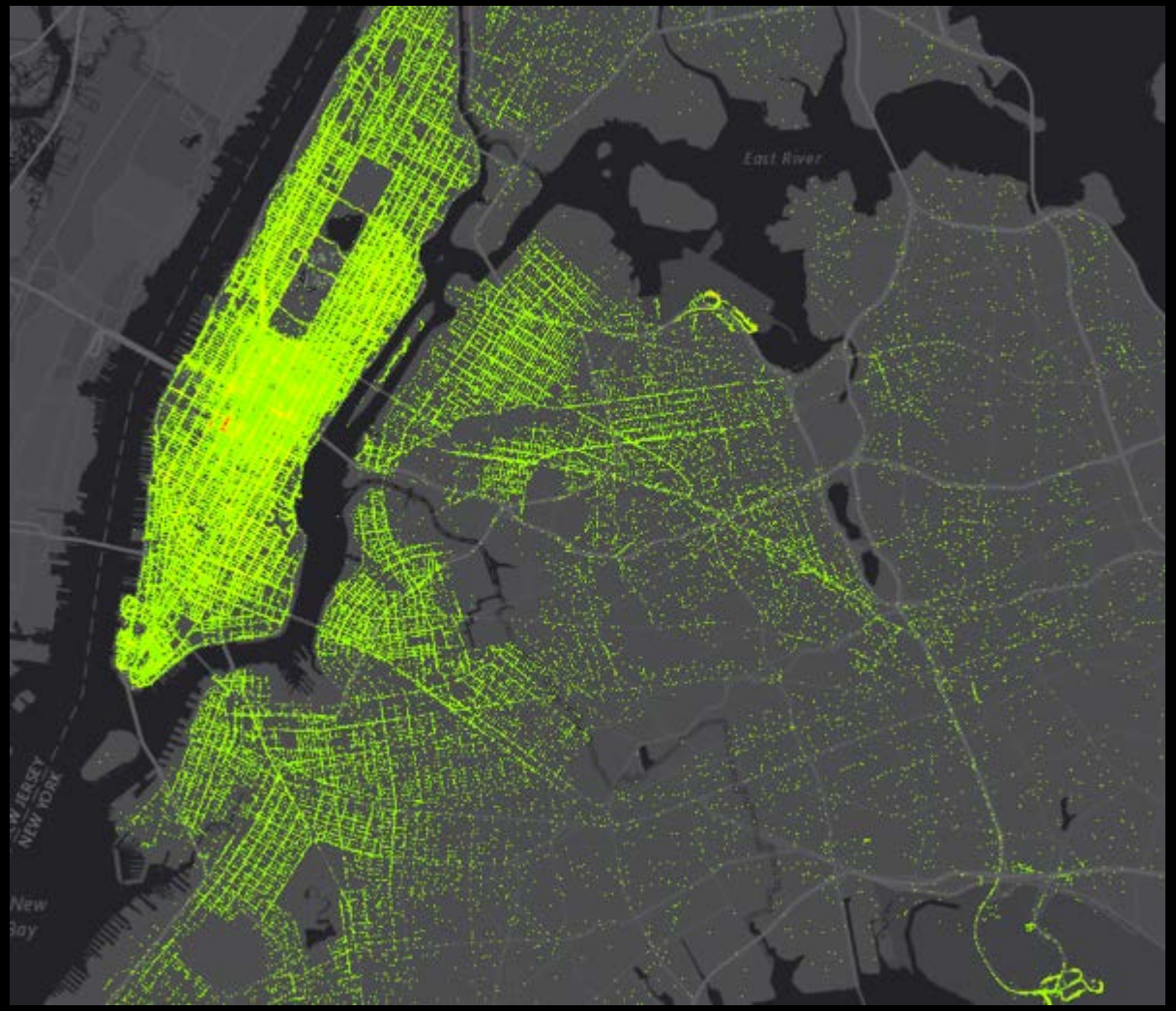
- Overview of big data
- Distributed computation
- User experience
- Data management

Big data

What is it?

- Big Data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard software in a tolerable elapsed time
  - Big data "size" is a constantly moving target, ranging from a few dozen terabytes to many petabytes of data
  - In the past three years, 90% of all recorded data has been generated
- Every 60 seconds:
  - 100,000 tweets
  - 2.4 million Google searches
  - 11 million instant messages
  - 170 million email messages
  - 1,800 TB of data

NYC Taxis by Day



Manhattan Taxis Friday after 8pm

Big data

What techniques are applied to handle it?

- **Data distribu** …oss a collection of
- **Parallel proc** …, combining the partial re
- **Fault toleran** …fails, the dataset can s
- **Commodity** …c architectures
- **Scalability** – …ollections of machines in order to address larger datasets

“Big data is not about the data.”

– Gary King

Harvard University

Director, Inst. For Quantitative Social Science

*(Making the point that while data is plentiful and easy to collect, **the real value is in the analytics**)*

ArcGIS users have big data

- **Smart Sensors**
  - Electrical meters (AMI), SCADA, UAVs
- **GPS Telemetry**
  - Vehicle tracking, smartphone data collectors, workforce tracking, geofencing
- **Internet data**
  - Social media streams, web log files, customer sentiment
- **Sensor data**
  - Weather sensors, stream gauge measurements, heavy equipment monitors, …
- **Imagery**
  - Satellites, frame cameras, drones

GeoAnalytics Examples

- Aggregate vehicle locations into cells for each 10 minute period to reveal traffic patterns
- Aggregate 911 call logs into census blocks by hour to reveal call patterns
- Aggregate web logs of access to map tile servers to determine hotspots of customer interest
- Geocode large address sets in parallel using a geocoding service
- Enrich very large numbers of point locations with contextual data and then select subset of locations meeting certain criterion

Road ahead?

10.4

GeoAnalytics

What is it, and what does it enable me to do?

- GeoAnalytics will be a new *capability* of ArcGIS Server
- It provides me:
  - The ability to do fast batch analysis on large tabular / feature datasets
  - The ability to do fast batch analysis on large raster and image datasets
  - The ability to do fast batch analysis on large geo-event observation archives

10.4

GeoAnalytics

What does ‘batch’ analysis mean

- Batch analysis means the ability to run analysis jobs on large datasets
  - The input is a persisted standard or big dataset
  - The output is a persisted standard or big dataset

- Datasets
  - Standard geospatial data (geodatabases, files, services)
  - Big Data (databases, files, services)

- Key point:

With suitably scaled GeoAnalytics, jobs that would take hours now take minutes

GeoAnalytics Extension for Server

10.4

- Adds out of the box analytics to ArcGIS Server
  - Analysis in ArcGIS Pro and Portal
  - Powered by a new Analysis Service / Toolbox in Server
  - Focused analysis for big data
- Works with:
  - Standard geospatial data (geodatabases, files, services)
  - Big Data (databases, files, services)

10.4

GeoAnalytics Extension for Server

Overview

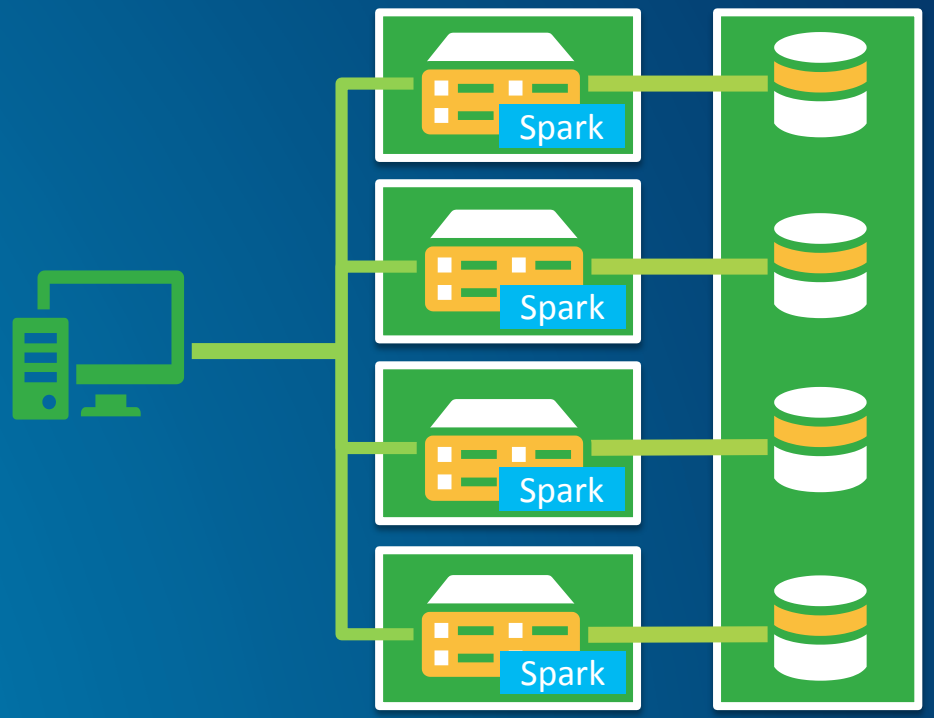
- Users are able to manage, analyze, and visualize big data to derive valuable information
- Previously impossible or slow analytics are made possible by leveraging the power of distributed computation
- Analytics and complicated technologies are made easy by ArcGIS integration
- Ability to perform analysis on vector and raster data

10.4

Distributed computation

Integrated into ArcGIS Server

- Distributed analytics against distributed data
- Many frameworks/technologies exist for distributing computation
  - E.g., Hadoop, MapReduce, Spark
  - Spark: processes distributed data in memory
- ArcGIS Server integrates these technologies on a cluster to solve analytic problems

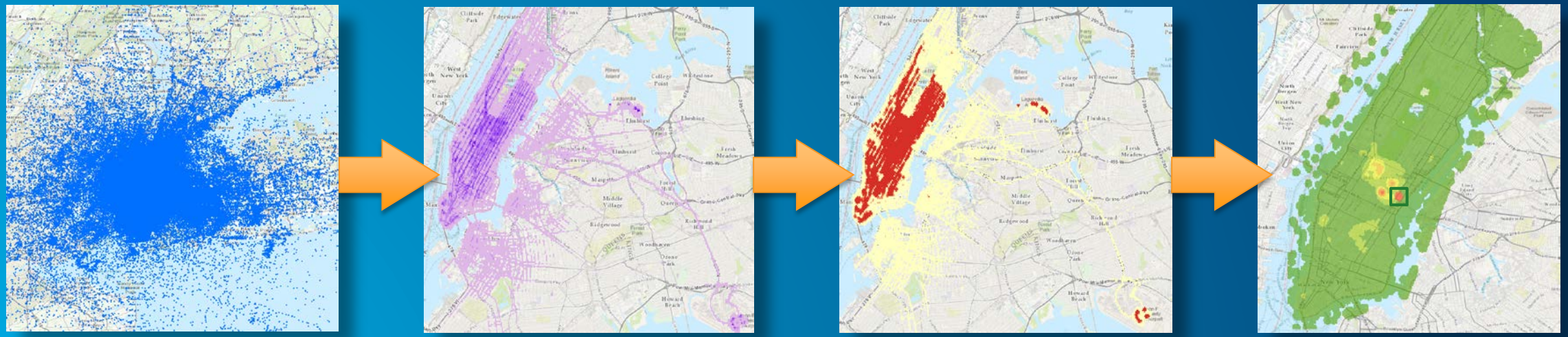


10.4

GeoAnalytics

Distributed analysis on distributed data

- Parallelized batch analytics on tabular, vector, raster, and imagery datasets (big and standard data)



Raw Data | Aggregated Data | Hotspots | Analysis Results

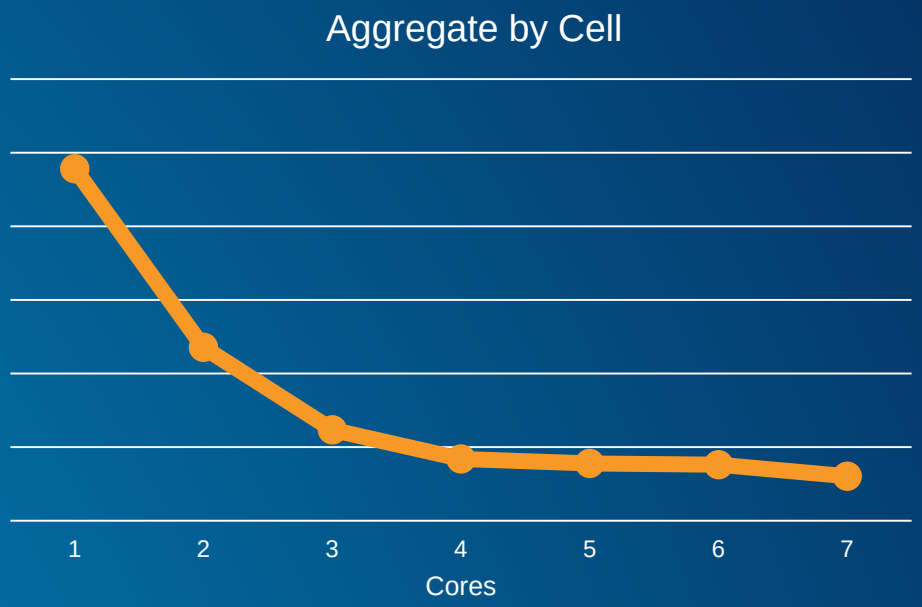
- Supports data exploration via feature, map, and image layers

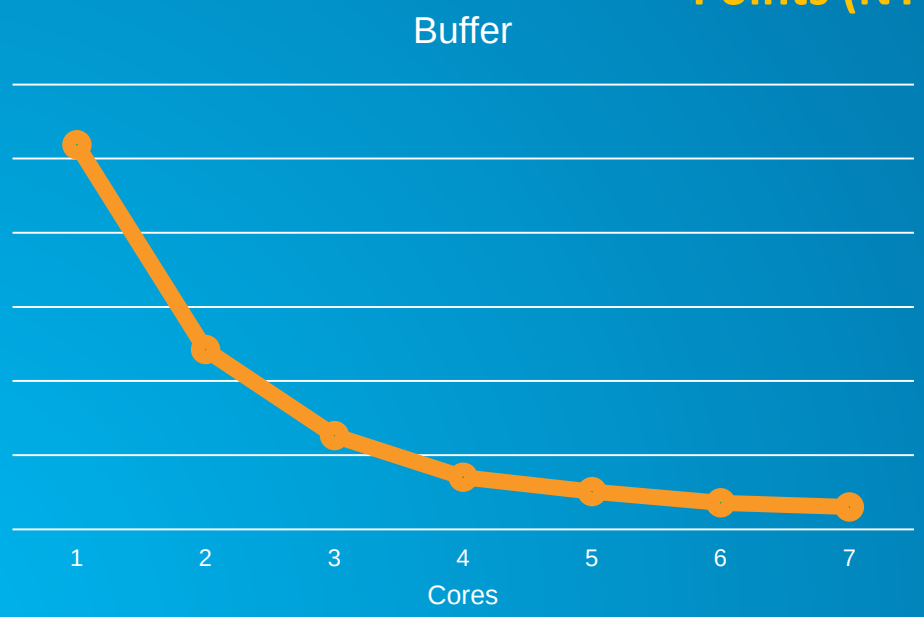
GeoAnalytics

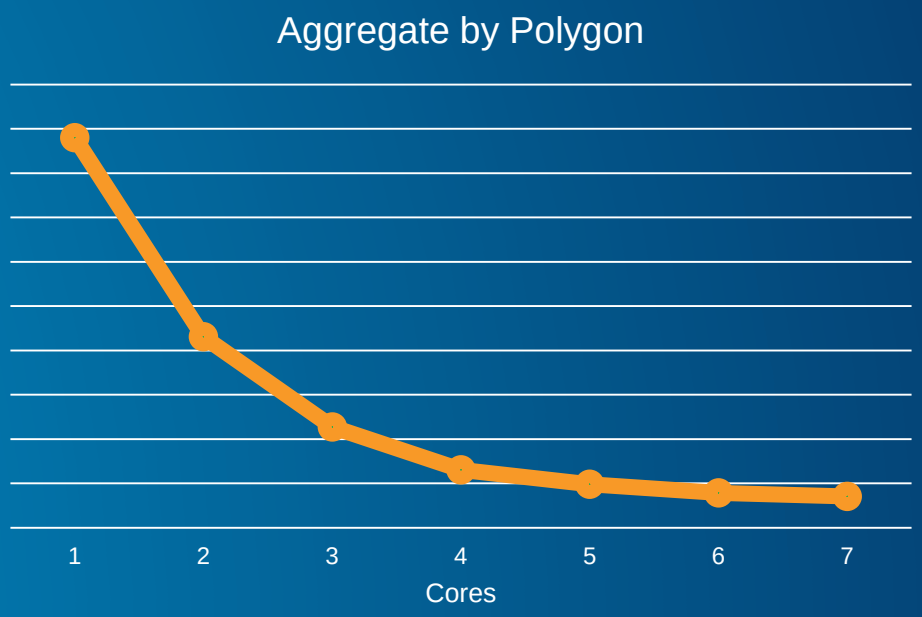
Performance: minutes, not hours

- 16 nodes in the cluster
  - 4 cores per node
  - 8 – 16GB RAM per node

Polygons (NYC Blocks) 40K
Points (NYC Taxi) 170M





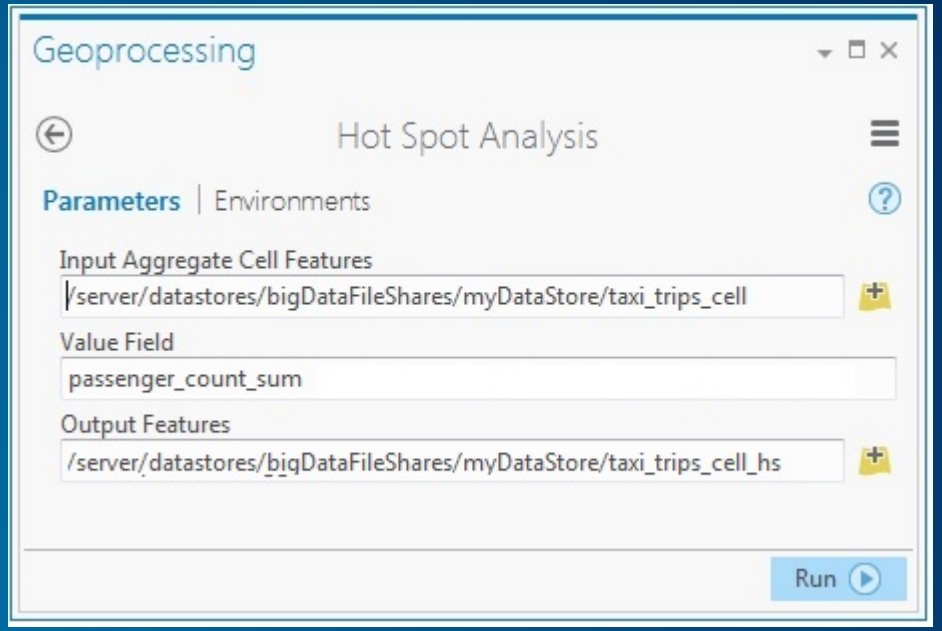


GeoAnalytics

User Experience - Analysis

- **ArcGIS Pro:**
  - **Out of the box tools that run in Server and process services and registered data using a GP tool interface**

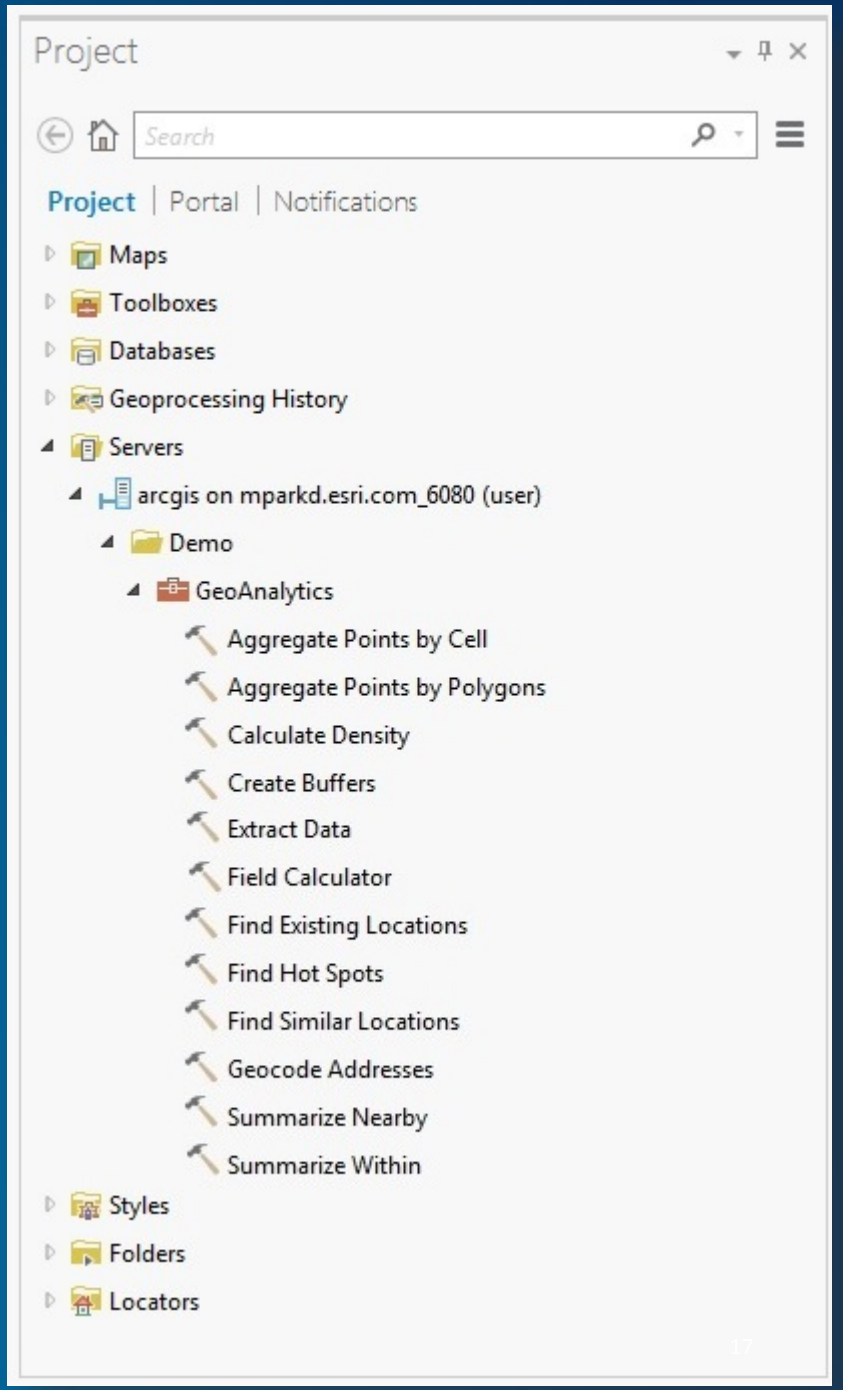
- **Tools are exposed through a REST-based interface that can be used by ArcGIS Pro or web clients**



Initial release

ArcGIS 10.4 - Analysis

- Analysis capabilities patterned after the ArcGIS Online Spatial Analysis service
  - Contains a useful subset of the current tasks
- GeoAnalytics includes additional tools useful for a big data workflows
  - Move data to and from the client
  - Register and manage data resident in the Big Data Server's directories
  - Addition of temporal capabilities
  - Ability to write to NetCDF



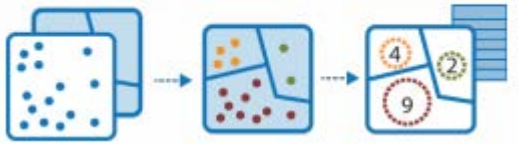
Analytic capabilities
ArcGIS 10.4 release

- **Summarize Data**
  - Aggregate Points by Polygon + time
  - Aggregate by Cell + time
  - Summarize Nearby + time
  - Summarize Within + time

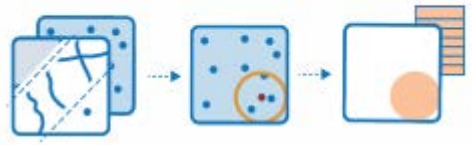
- **Find Locations**
  - Find Existing Locations
  - Find Similar Locations

* New GeoAnalytics capabilities in orange

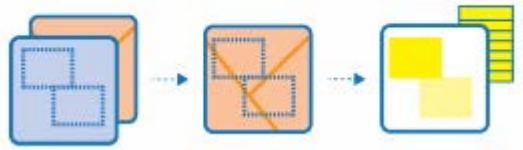
Aggregate Points



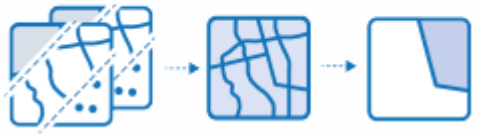
Summarize Nearby



Summarize Within



Find Existing Locations



Find Similar Locations

Analytic capabilities

ArcGIS 10.4 release

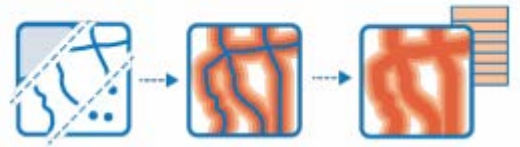
- **Analyze Patterns**
  - Calculate Density
  - Find Hot Spots + time
- **Use Proximity**
  - Create Buffers + time
- **Manage Data**
  - Extract Data
  - Field Calculator
  - Geocode Addresses

* New GeoAnalytics capabilities in orange

Calculate Density

Find Hot Spots

Create Buffers



Extract Data

Field Calculator

Geoprocessing

Aggregate by Cells

Parameters | Environments

Input Features

http://sarahdesk.esri.com:6080/arcgis/rest/services/taxi_trips/FeatureServer/0

Fields to Summarize

passenger_count

Output Features

/server/datastores/bigDataFileShares/myDataStore/taxi_trips_cell

Cell Size

330 Feet

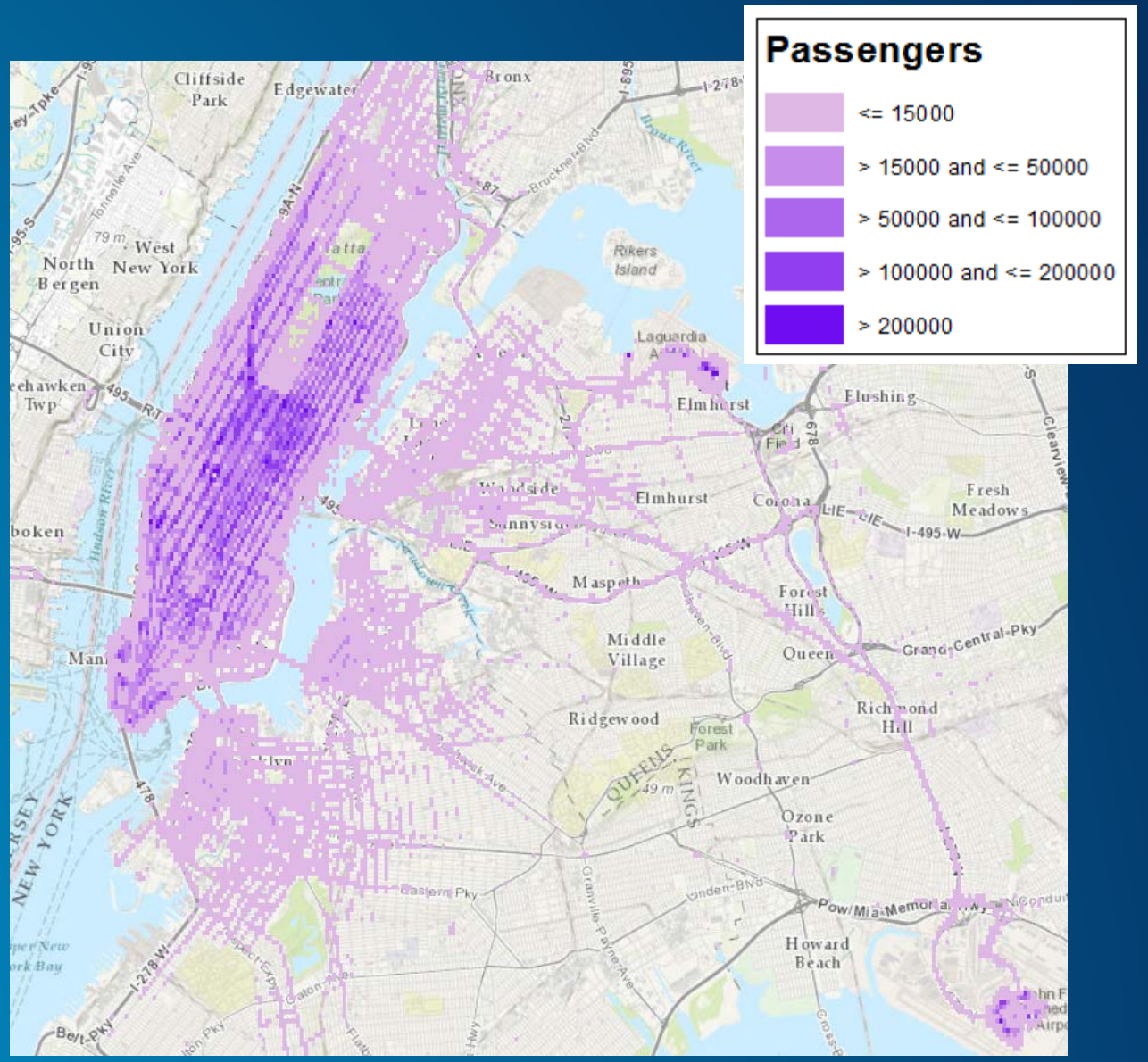
SQL Expression

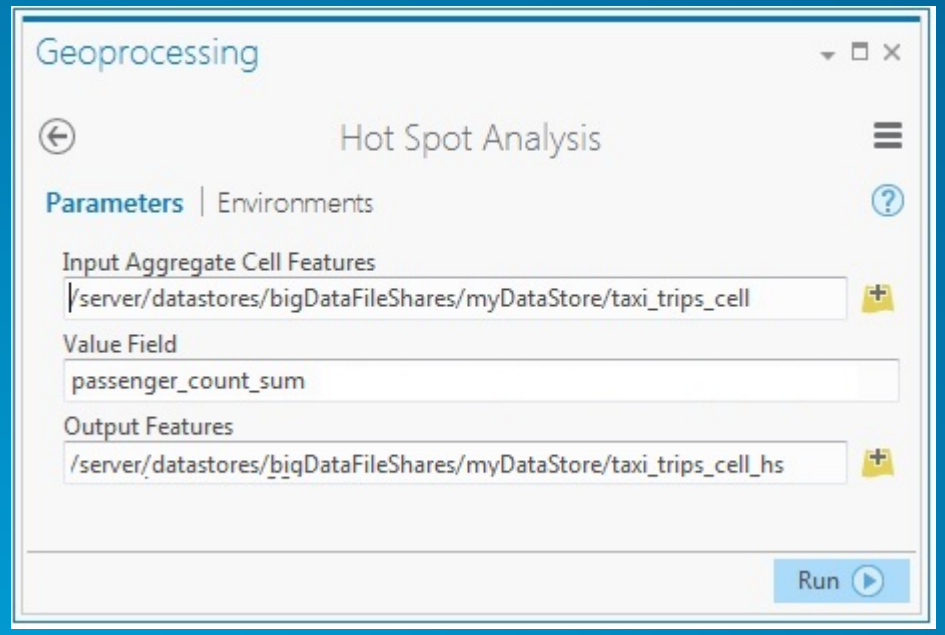
Extent: As Specified Below

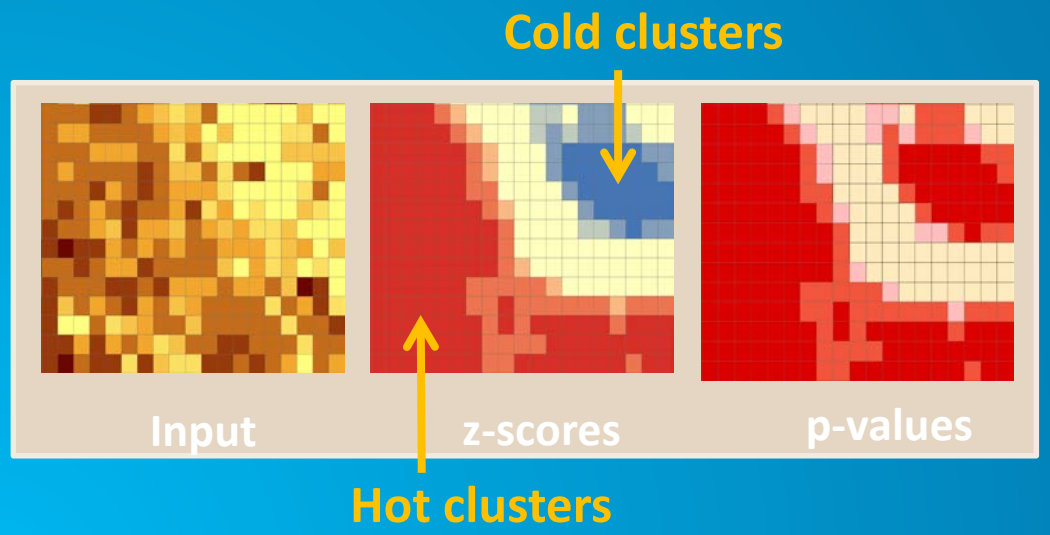
-74.0256386116715 -73.9601940094528

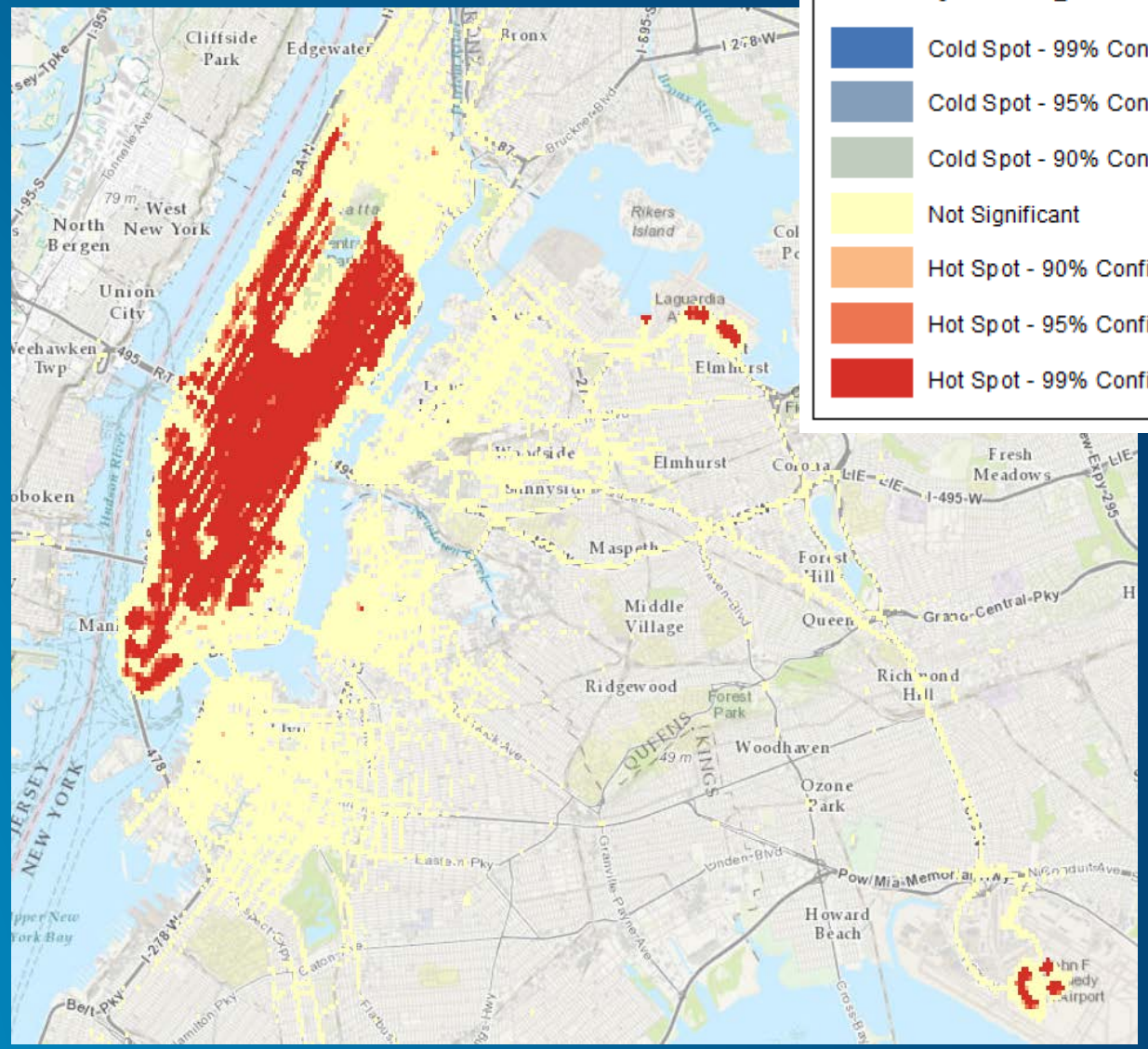
40.6970326732393 40.7944178151074

Run



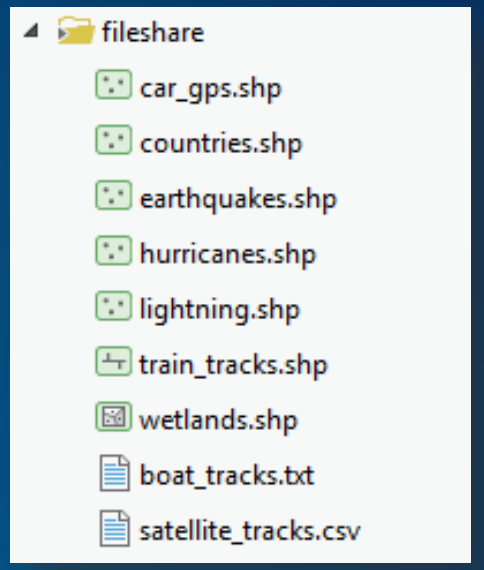

Geoprocessing
Hot Spot Analysis
Parameters | Environments
Input Aggregate Cell Features
/server/datastores/bigDataFileShares/myDataStore/taxi_trips_cell
Value Field
passenger_count_sum
Output Features
/server/datastores/bigDataFileShares/myDataStore/taxi_trips_cell_hs
Run

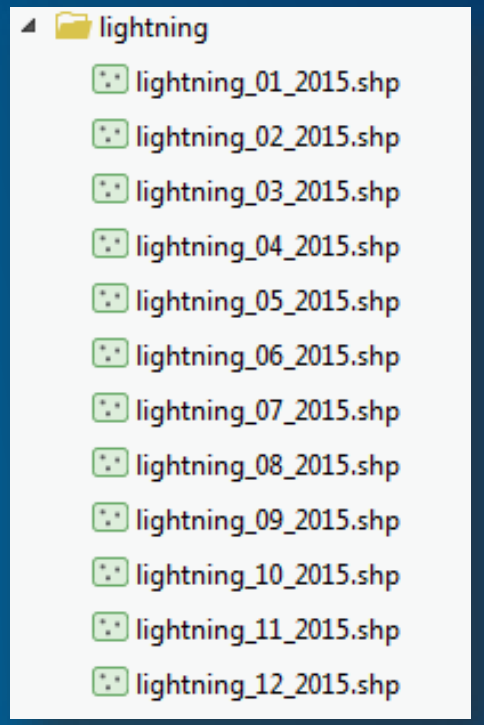

Cold clusters
Input
z-scores
p-values
Hot clusters


Hot Spot Significance
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence
Cliffside Park
Edgewater
Bronx
North Bergen
West New York
Union City
Rikers Island
Laguardia
Elmhurst
Middle Village
Ridgewood
Forest Park
Woodhaven
Ozone Park
Howard Beach
Fresh Meadows
Richmond Hill
Queens
Kings
Jersey
New York

Data Stores
Management

- Both GIS data stores and big data stores are supported
  - Map and Feature services
  - ArcGIS SQL Data Store
- Directories of files (shapefiles, CSVs, etc.) serve as data stores
  - GIS file shares
    - Each file represents a single dataset
  - Big data file shares
    - Folder of sharded shapefiles or other file formats
- ArcGIS Big Data Store





Anatomy of a Feature

Not just spatial

Attributes

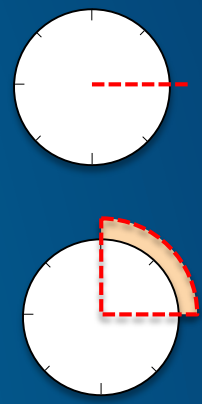
- Text
- Numbers
- Dates
- Binary
- …

Geometry

- Point
- Polyline
- Polygon

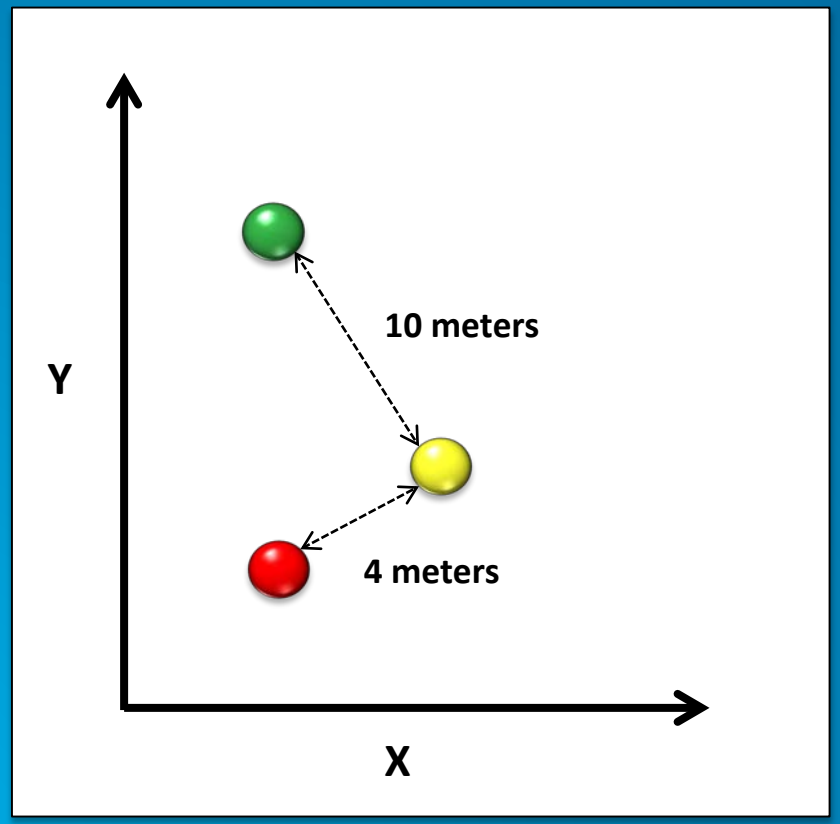
Time

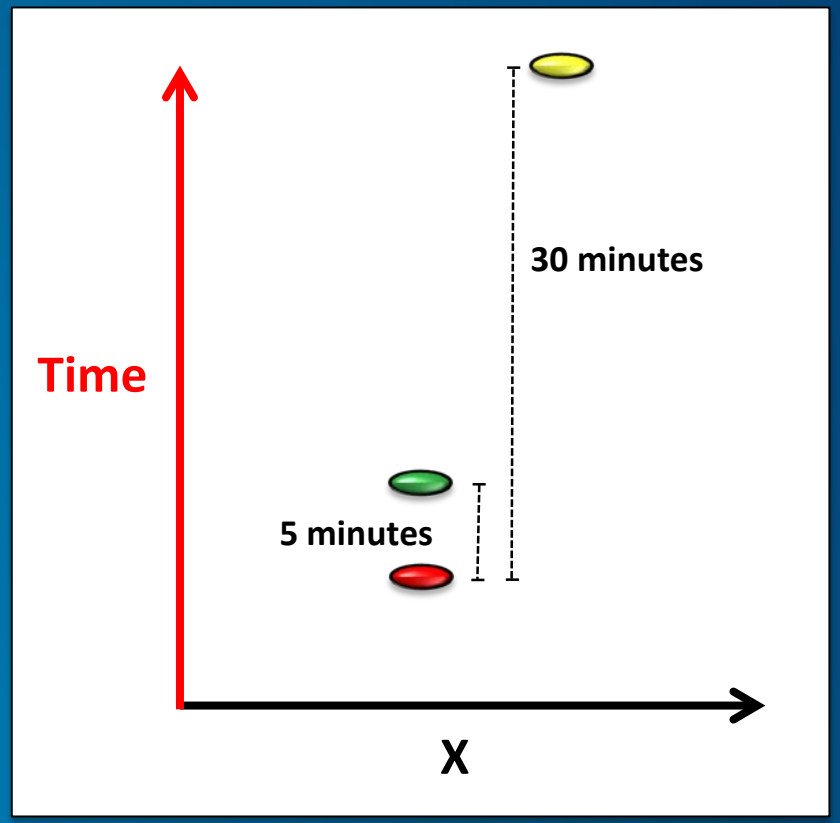
- **Instant**
- **Interval**



Why Time Is Important

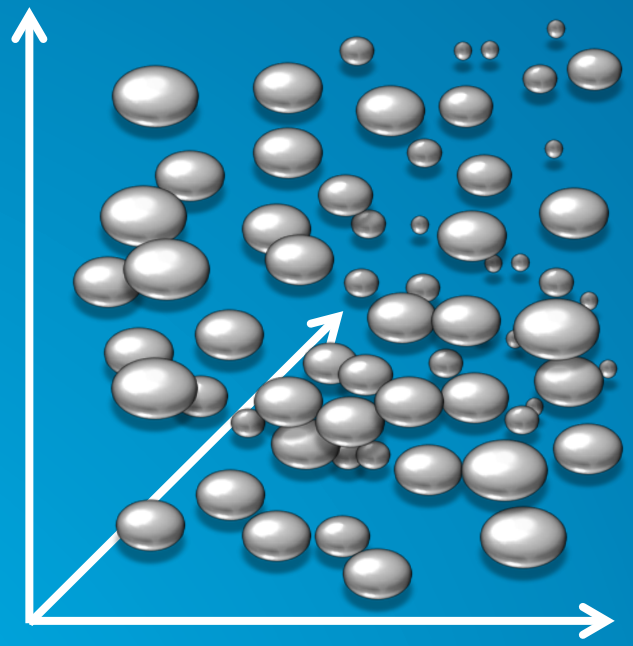
Space/time relationship

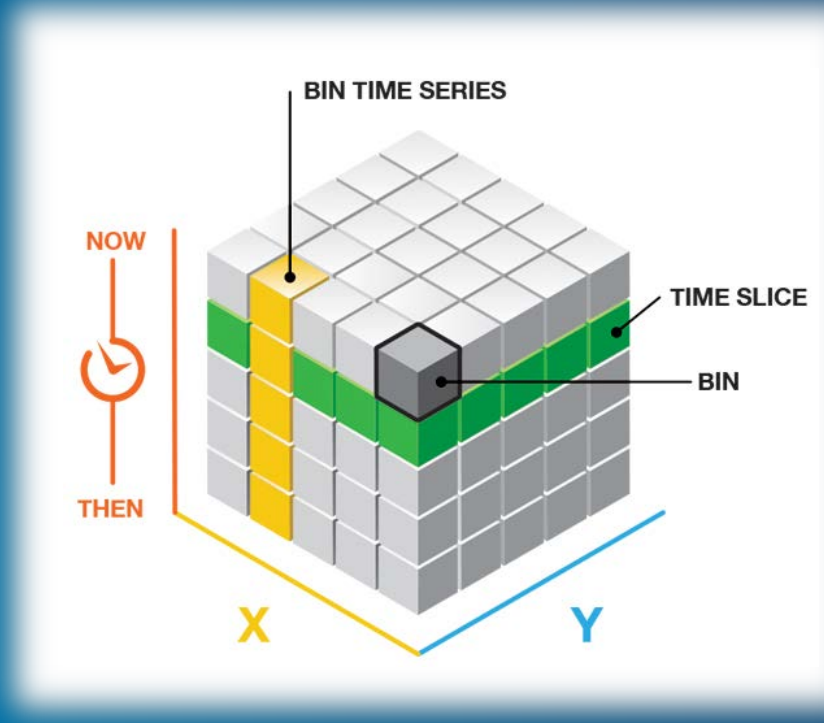




Why Time Is Important

Summarization

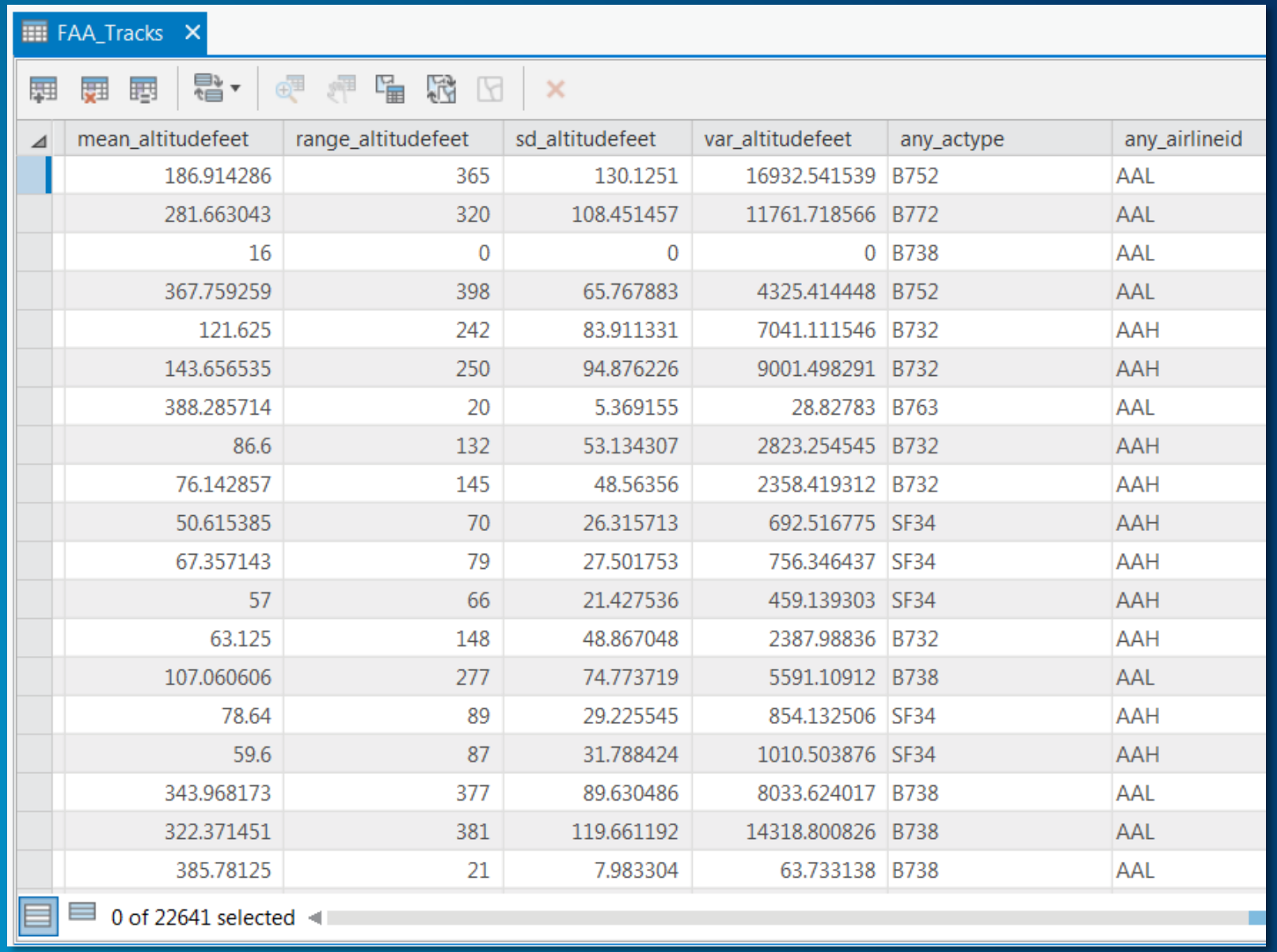




Aggregation

Summary statistics

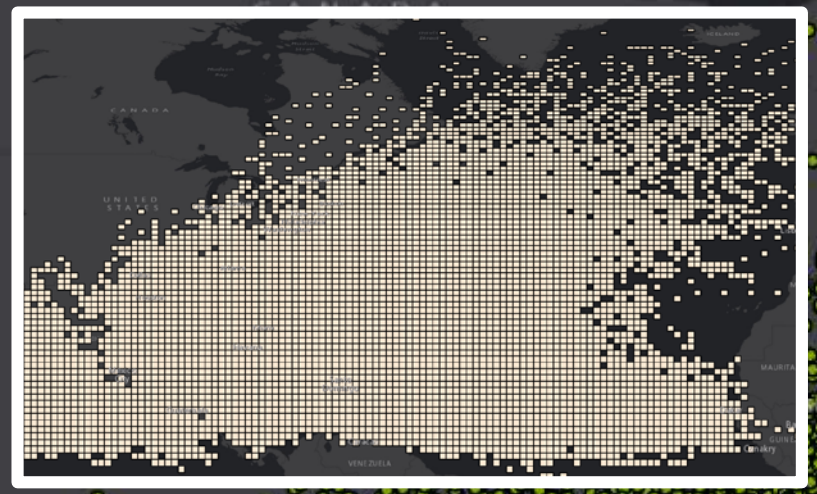
- Numeric Statistics
  - Count
  - Min
  - Max
  - Sum
  - Mean
  - Standard Deviation
  - Variance
- Text Statistics
  - Min (alphabetical ordering)
  - Max (alphabetical ordering)
  - Any

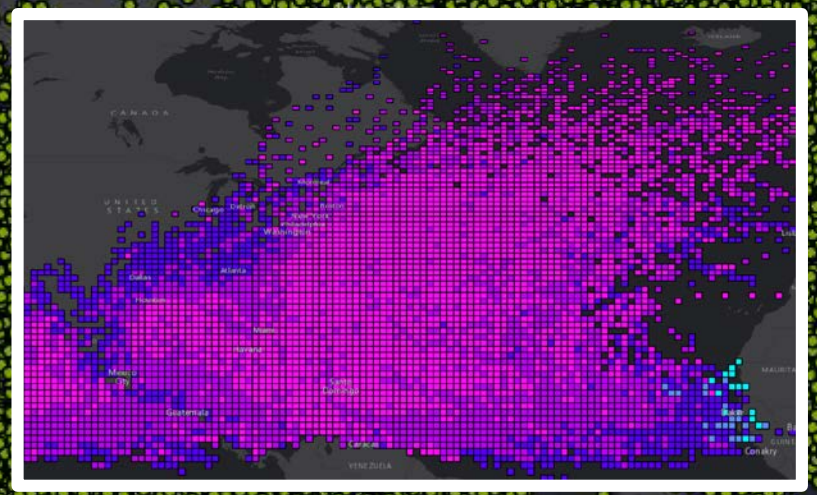


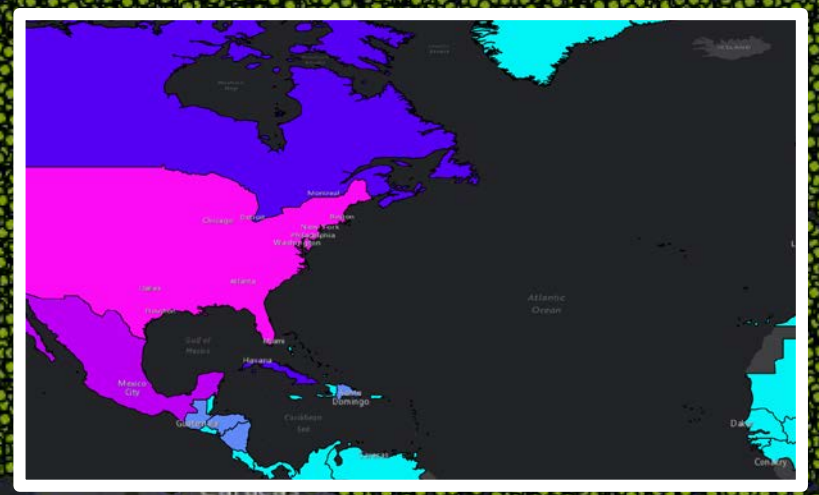
mean_altitudefeet	range_altitudefeet	sd_altitudefeet	var_altitudefeet	any_actype	any_airlineid
186.914286	365	130.1251	16932.541539	B752	AAL
281.663043	320	108.451457	11761.718566	B772	AAL
16	0	0	0	B738	AAL
367.759259	398	65.767883	4325.414448	B752	AAL
121.625	242	83.911331	7041.111546	B732	AAH
143.656535	250	94.876226	9001.498291	B732	AAH
388.285714	20	5.369155	28.82783	B763	AAL
86.6	132	53.134307	2823.254545	B732	AAH
76.142857	145	48.56356	2358.419312	B732	AAH
50.615385	70	26.315713	692.516775	SF34	AAH
67.357143	79	27.501753	756.346437	SF34	AAH
57	66	21.427536	459.139303	SF34	AAH
63.125	148	48.867048	2387.98836	B732	AAH
107.060606	277	74.773719	5591.10912	B738	AAL
78.64	89	29.225545	854.132506	SF34	AAH
59.6	87	31.788424	1010.503876	SF34	AAH
343.968173	377	89.630486	8033.624017	B738	AAL
322.371451	381	119.661192	14318.800826	B738	AAL
385.78125	21	7.983304	63.733138	B738	AAL

Aggregation

Summary methods

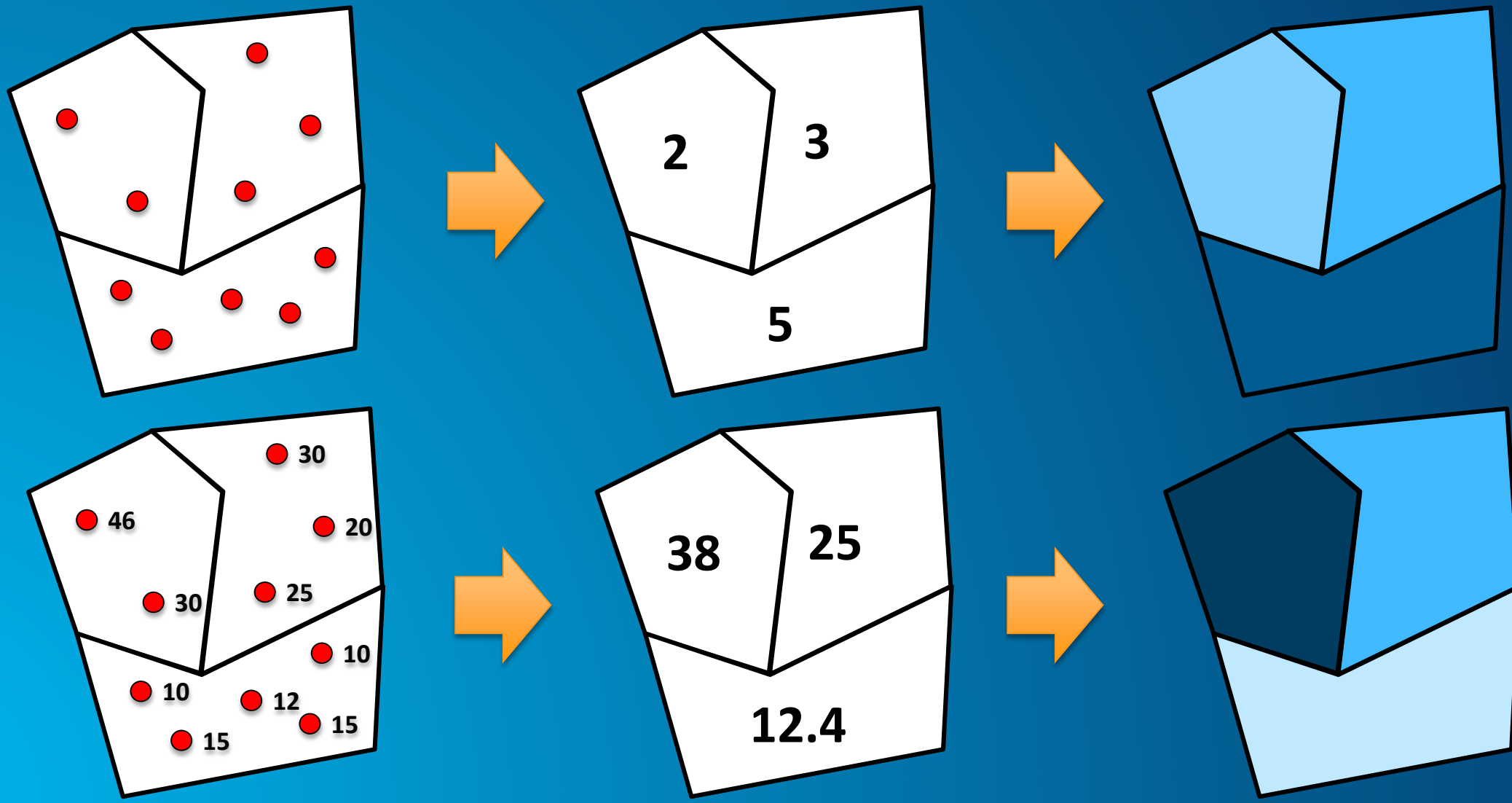






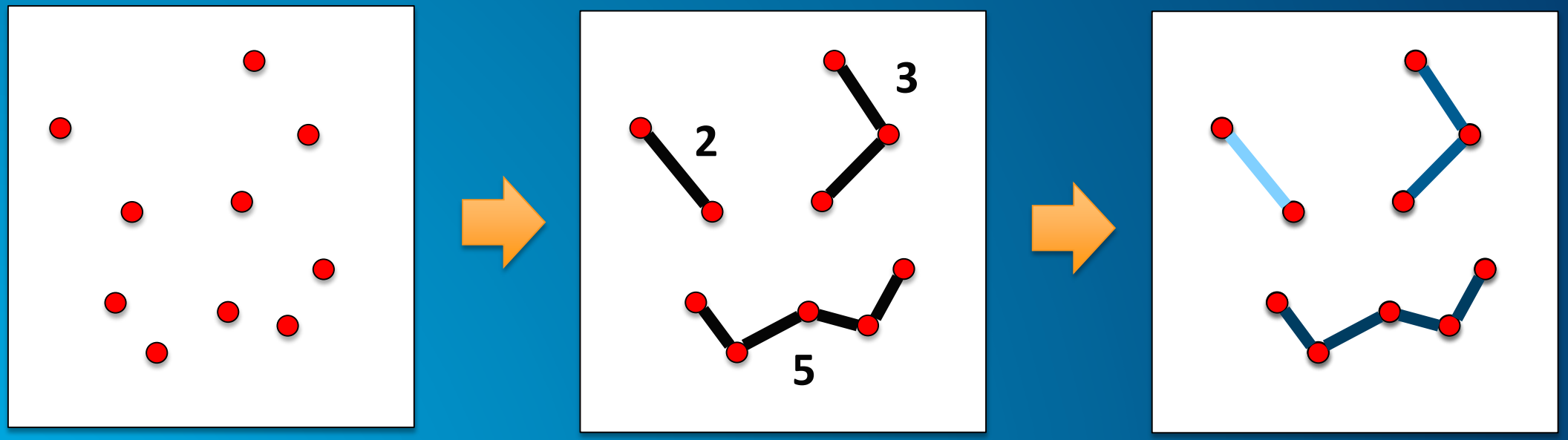
Aggregation

Point counts and attribute means



Path generation

Vertex count aggregation



10.4

Spatio-temporal big data store

Management

- Distributed data store for high velocity, high volume data
- Available to GeoAnalytics and GeoEvent
  - Supports high velocity continuous analytics with GeoEvent services
  - Supports high volume batch analytics with GeoAnalytics services
- Accessible through feature services
- Based upon Elasticsearch for storage and indexing
  - Open-source real-time distributed search engine and data store built on top of Apache Lucene

10.4

Integration with GeoEvent

ArcGIS 10.4 Release

- Enhanced GeoEvent service integration
- Partnership to better support persisting high velocity, high volume streaming data into the Big Data cluster
  - Spatio-temporal Big Data Store
- Shared platform service for distributed computation

10.4

GeoAnalytics capabilities for server

Summary

- Allows you to run GeoAnalytics on dedicated server nodes
- Uses services and data stores to expose the results of analyses
- Supports management and analytics against massive spatio-temporal datasets

Why would I want to use it?

Summary

- Functionality available out of the box in Portal; no need to publish
- Runs on big data collections (observational data)
  - Data collections whose size was previously problematic
- Runs fast, and is scalable
- I don’t need to learn anything new; I use it just like existing GP tools

Understanding our world.