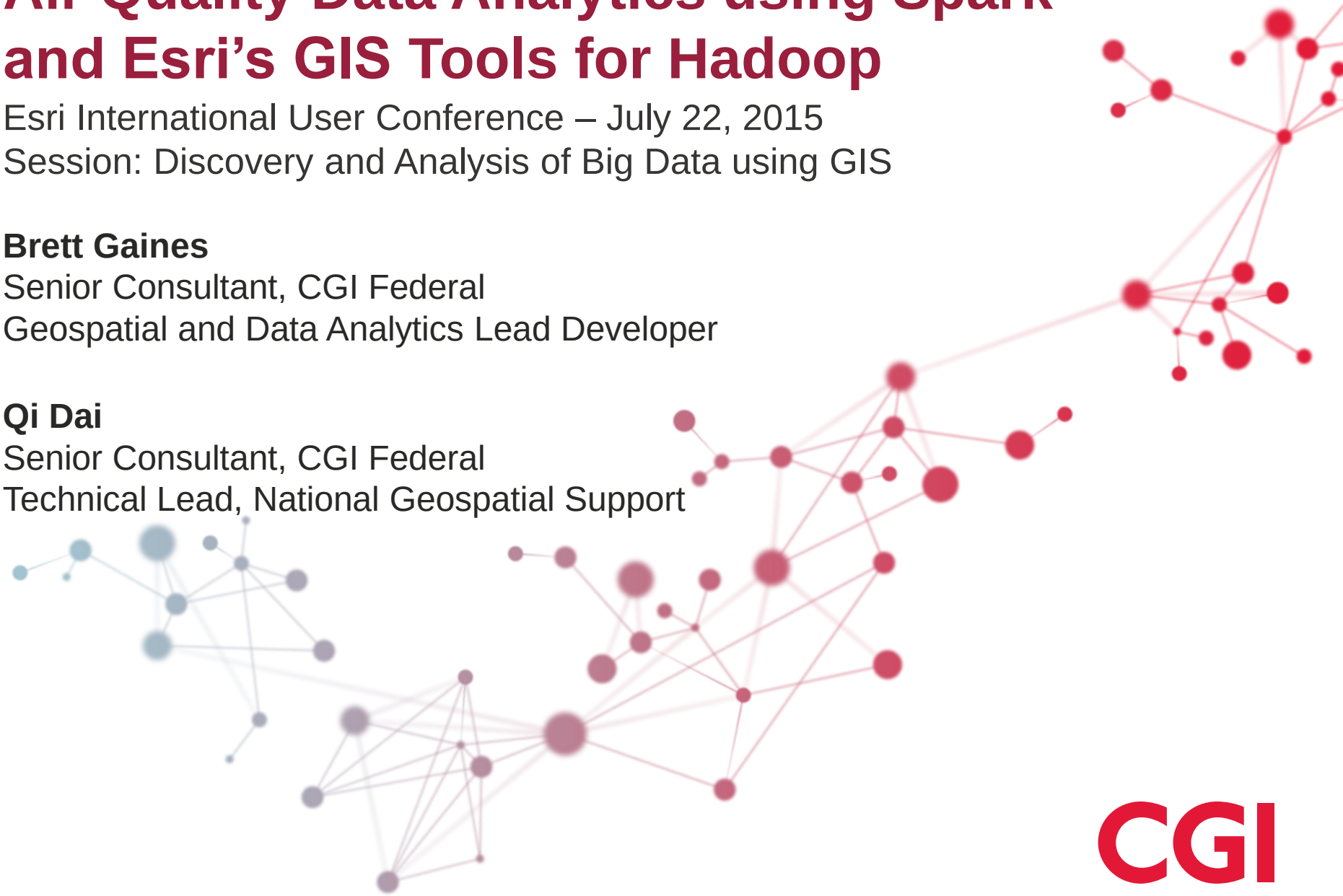
Air Quality Data Analytics using Spark and Esri's GIS Tools for Hadoop

Esri International User Conference – July 22, 2015
Session: Discovery and Analysis of Big Data using GIS

Brett Gaines
Senior Consultant, CGI Federal
Geospatial and Data Analytics Lead Developer

Qi Dai
Senior Consultant, CGI Federal
Technical Lead, National Geospatial Support



CGI

Experience the commitment®

Overview

- Goal of Analysis
- Data Sources
- Hardware Cluster
- Data Processing Steps
  - Anomaly Detection Methods (Statistics)
  - GIS Analysis
- Data Analytics Results and Mapping

Purpose Overview

Data Science

- Apply an anomaly detection algorithm on spatio-temporal static air monitoring pollutant data
- Data is collected hourly by thousands of monitors and contains data for multiple pollutants

Target Architecture

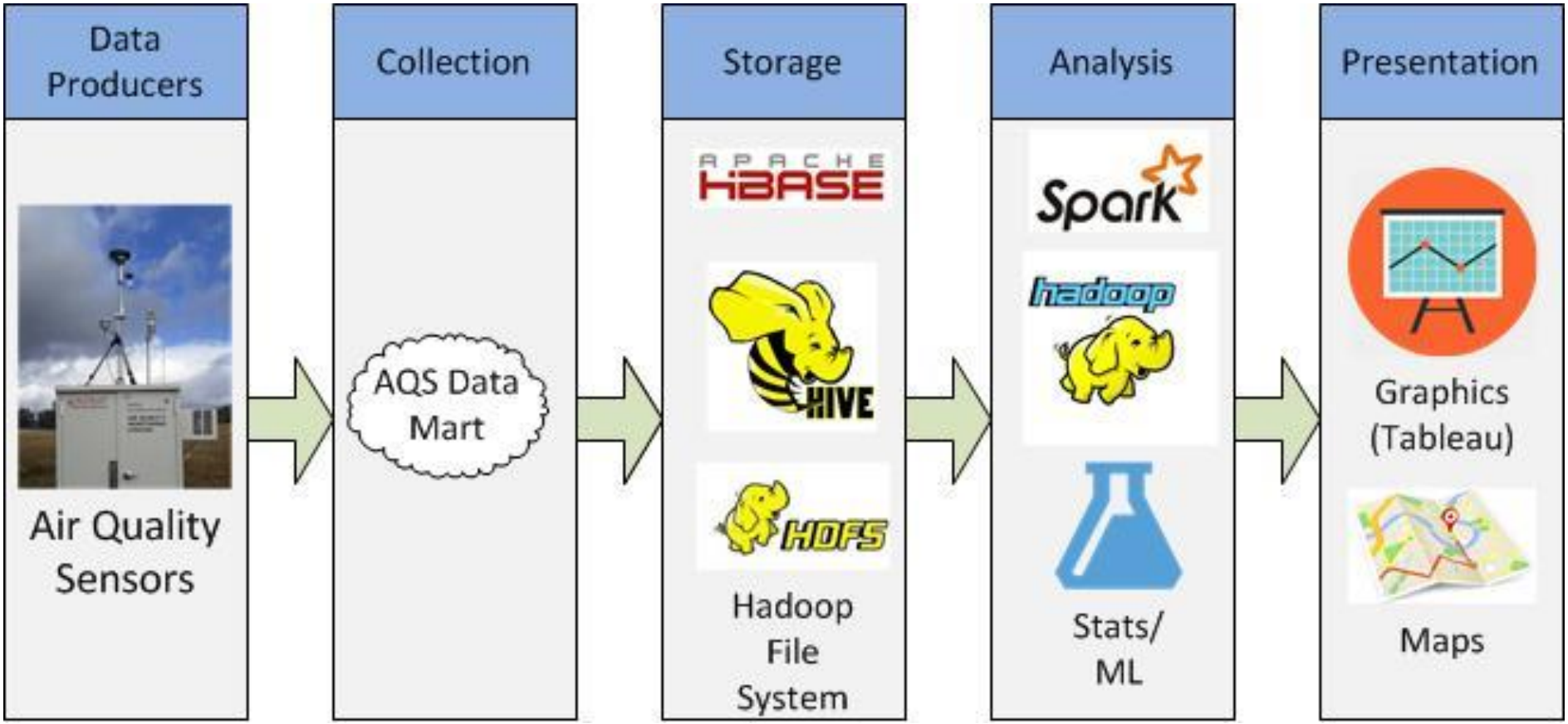
- Hadoop ecosystem & Spark for batch analysis
- Visualization of spatio-temporal results in Tableau and Esri
- Export anomaly datasets to on premise GIS servers & AGOL

Deployment

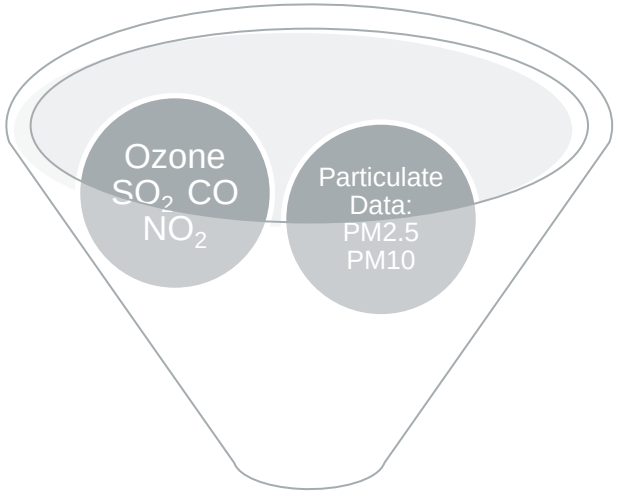
- Hortonworks Data Platform (HDP) cluster
- Esri GIS Tools for Hadoop (extended)

CGI

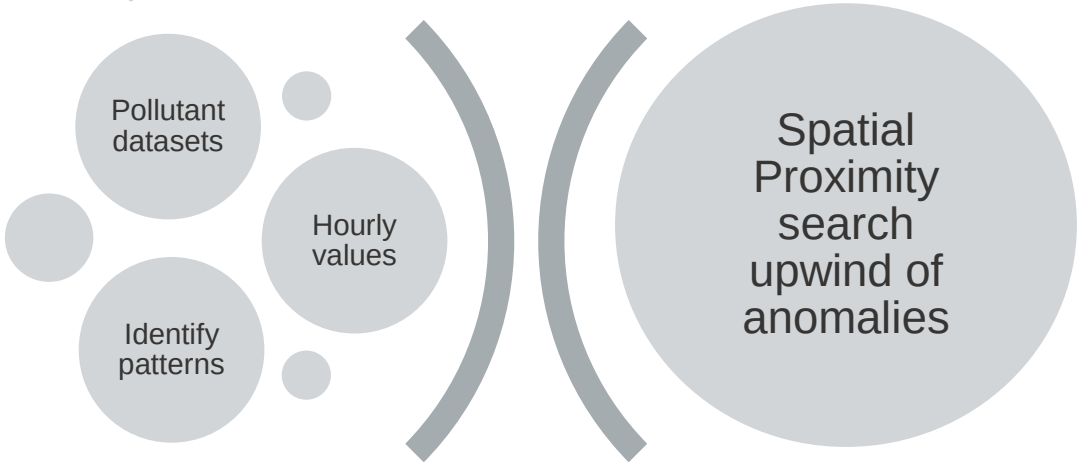
Workflow


Data Producers
Collection
Storage
Analysis
Presentation
Air Quality Sensors
AQS Data Mart
APACHE HBASE
HIVE
HDFS
Hadoop File System
Spark
hadoop
Stats/ ML
Graphics (Tableau)
Maps

Analysis Overview

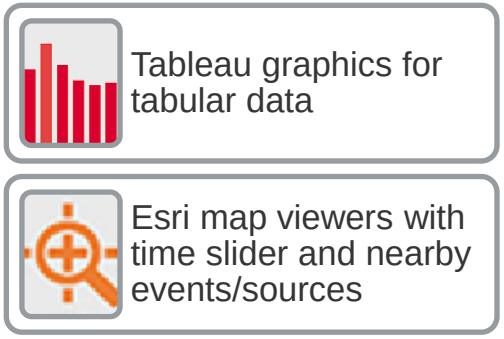


- Distributed spatial search with targeted vectors (wind speed and direction at hour and location of anomaly)
- Extending Esri's Geometry API and integration with Spark



Identify Anomalies

Locate Potential Cause



CGI

Data Sources

- USEPA Air Quality System (AQS)
  - Stores data from >10,000 monitors - Annual, Daily, Hourly, Minute
- 2008-2013 for 6 pollutants
  - Ozone, SO2, CO, NO2, PM25 Non-FRM, PM10 Mass
    - Vary Seasonally and/or Daily
- Meteorological data and possible stationary emission sources

Hourly Data

Criteria Gases

Year	Ozone (44201)	SO2 (42401)	CO (42101)	NO2 (42602)
2014	hourly_44201_2014.zip 7,147,884 Rows 52,900 KB As of 2015-01-02	hourly_42401_2014.zip 2,861,454 Rows 19,046 KB As of 2015-01-02	hourly_42101_2014.zip 1,761,002 Rows 11,998 KB As of 2015-01-02	hourly_42602_2014.zip 2,433,435 Rows 18,323 KB As of 2015-01-02
2013	hourly_44201_2013.zip 9,096,192 Rows 67,040 KB As of 2015-01-02	hourly_42401_2013.zip 3,797,758 Rows 25,191 KB As of 2015-01-02	hourly_42101_2013.zip 2,498,904 Rows 16,893 KB As of 2015-01-02	hourly_42602_2013.zip 3,188,575 Rows 23,803 KB As of 2015-01-02
2012	hourly_44201_2012.zip 9,025,084 Rows 66,896 KB As of 2015-01-02	hourly_42401_2012.zip 3,770,826 Rows 25,073 KB As of 2015-01-02	hourly_42101_2012.zip 2,572,491 Rows 17,258 KB As of 2015-01-02	hourly_42602_2012.zip 3,081,439 Rows 22,857 KB As of 2015-01-02
2011	hourly_44201_2011.zip 8,878,649 Rows 65,644 KB As of 2015-01-02	hourly_42401_2011.zip 3,676,396 Rows 24,565 KB As of 2015-01-02	hourly_42101_2011.zip 2,612,976 Rows 17,500 KB As of 2015-01-02	hourly_42602_2011.zip 3,017,114 Rows 22,321 KB As of 2015-01-02
2010	hourly_44201_2010.zip 8,392,448 Rows 62,172 KB As of 2015-01-02	hourly_42401_2010.zip 3,661,150 Rows 24,041 KB As of 2015-01-02	hourly_42101_2010.zip 2,616,882 Rows 16,937 KB As of 2015-01-02	hourly_42602_2010.zip 3,111,967 Rows 22,388 KB As of 2015-01-02
2009	hourly_44201_2009.zip 8,201,693 Rows 59,443 KB As of 2015-01-02	hourly_42401_2009.zip 3,732,540 Rows 24,115 KB As of 2015-01-02	hourly_42101_2009.zip 2,753,380 Rows 17,597 KB As of 2015-01-02	hourly_42602_2009.zip 3,084,877 Rows 21,618 KB As of 2015-01-02
2008	hourly_44201_2008.zip 8,054,745 Rows 58,634 KB As of 2014-06-13	hourly_42401_2008.zip 3,963,631 Rows 25,656 KB As of 2015-01-02	hourly_42101_2008.zip 2,941,703 Rows 18,759 KB As of 2015-01-02	hourly_42602_2008.zip 3,187,823 Rows 21,976 KB As of 2015-01-02
2007	hourly_44201_2007.zip 8,005,170 Rows 58,479 KB As of 2014-06-13	hourly_42401_2007.zip 4,216,470 Rows 27,239 KB As of 2015-01-02	hourly_42101_2007.zip 3,036,390 Rows 19,309 KB As of 2014-06-13	hourly_42602_2007.zip 3,241,278 Rows 22,207 KB As of 2015-01-02
2006	hourly_44201_2006.zip 7,859,903 Rows 57,396 KB As of 2014-06-13	hourly_42401_2006.zip 4,206,488 Rows 27,119 KB As of 2015-01-02	hourly_42101_2006.zip 3,193,385 Rows 20,324 KB As of 2014-06-13	hourly_42602_2006.zip 3,334,127 Rows 22,912 KB As of 2015-01-02
2005	hourly_44201_2005.zip 7,762,599 Rows 56,660 KB As of 2014-06-13	hourly_42401_2005.zip 4,304,211 Rows 27,873 KB As of 2015-01-02	hourly_42101_2005.zip 3,407,244 Rows 21,699 KB As of 2014-06-13	hourly_42602_2005.zip 3,349,695 Rows 23,073 KB As of 2015-01-02

CGI

Distributed Cluster Environment

- Hortonworks Data Platform (HDP) 2.2
  - RHEL 6.6 OS's
  - 16 CPUs Total
  - 40GB RAM Total
  - 2.5TB Disk Total

- And testing on Azure HDInsight

http://azure.microsoft.com/en-us/services/hdinsight/

Data Analysis – High Level Processing Steps

- Download raw data from Public AQS Data Mart
- Exploratory analysis in R
- Pre-process raw data with Python and ‘sed’
- Import to HDFS
- Create Hive schema-on-read HQL scripts
- Process Hive tables
- Spark jobs with Esri Geometry API (GIS Tools for Hadoop)
- Output from analysis into ArcGIS ecosystem

CGI

Data Analysis – Anomaly Detection and QC

- Detect anomalies at scale and quickly identifiable
- Flag records with identical samples >3 hours in succession
- Check specifically for evening monitor QC samples
- Compare each site only to itself (distinctive “normals”)
- Using anomaly outputs, detect spatial autocorrelation of nearby monitors

Hive – Schema on Read

- Hive can either store a copy of the data or store reference to the data (EXTERNAL command)
- Esri GIS Tools for Hadoop provides Hive UDFs
  - ${env:HOME}/esri-git/gis-tools-for-hadoop/samples/lib/esri-geometry-api.jar
  - ${env:HOME}/esri-git/gis-tools-for-hadoop/samples/lib/spatial-sdk-hadoop.jar;
  - Function ST_Point as ‘com.esri.Hadoop.hive.ST_Point’;

```
1  DROP TABLE IF EXISTS hourly0813co;
2
3  CREATE TABLE IF NOT EXISTS hourly0813co (State_Code STRING, County_Code STRING, Site_Num STRING,
   Parameter_Code string, POC int, Latitude DOUBLE, Longitude DOUBLE, Datum string, Parameter_Name STRING,
   Date_Local STRING, Time_Local STRING, Date_GMT STRING, Time_GMT STRING, Sample_Measurement DOUBLE,
   Units_of_Measure STRING, MDL DOUBLE, Uncertainty STRING, Qualifier STRING, MethodType STRING, Method_Name
   STRING, State_Name STRING, County_Name STRING, Date_of_Last_Change STRING)
4  ROW FORMAT DELIMITED FIELDS TERMINATED BY ",";
5
6  LOAD DATA LOCAL INPATH '/home/bg20/hourly0813no2.csv' OVERWRITE INTO TABLE hourly0813no2;
7
```


CGI

Anomaly Detection Methods

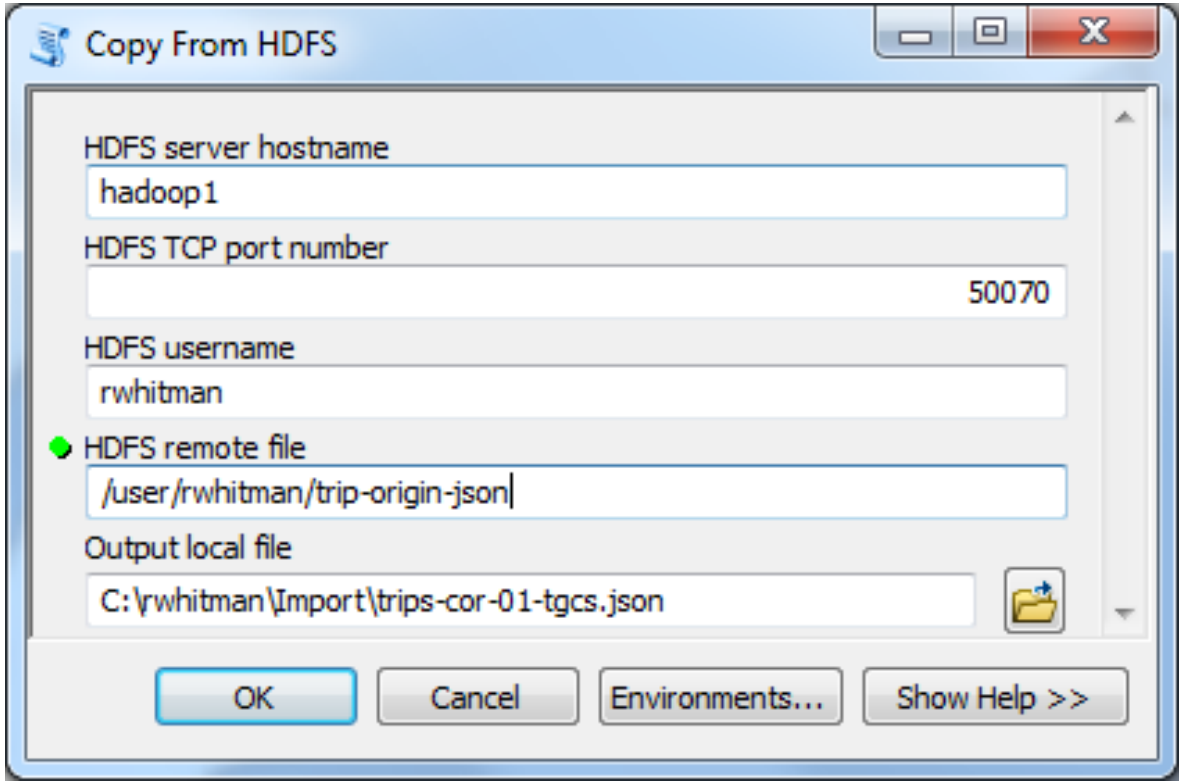
- Global and local maximas
- Mean, variance, standard deviation
- Support Vector Machines, Density-based (KNN), Neural Nets, Fuzzy Logic

- Median Absolute Deviation (ModZScore)
  - Implemented in PySpark
    - By site, by month, by hour
    - Threshold determined by ModZScore

CGI

Anomaly Results in HDFS – Now what?

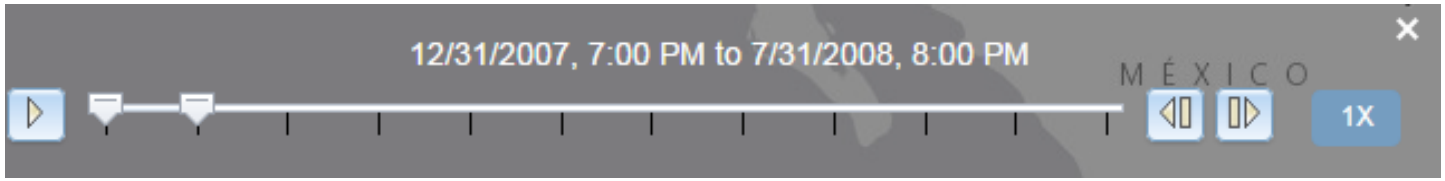
- Transfer output and perform spatial analysis on anomaly data



http://blogs.esri.com/esri/arcgis/2013/08/09/vehicle-trip-discovery-with-gis-tools-for-hadoop/

Time-series spatial correlation

- Checking if anomalies occur at same exact time for nearby monitors over 6 years
  - These monitors can be possible candidates for ‘buddy sites’
  - Time-series filtered buffer/intersect
    - Esri Geometry API in Spark Job:
    - Proximity2DResult.**getCoordinate**()
      - Returns the closest coordinate



CGI

Directed Proximity Search on Anomalies

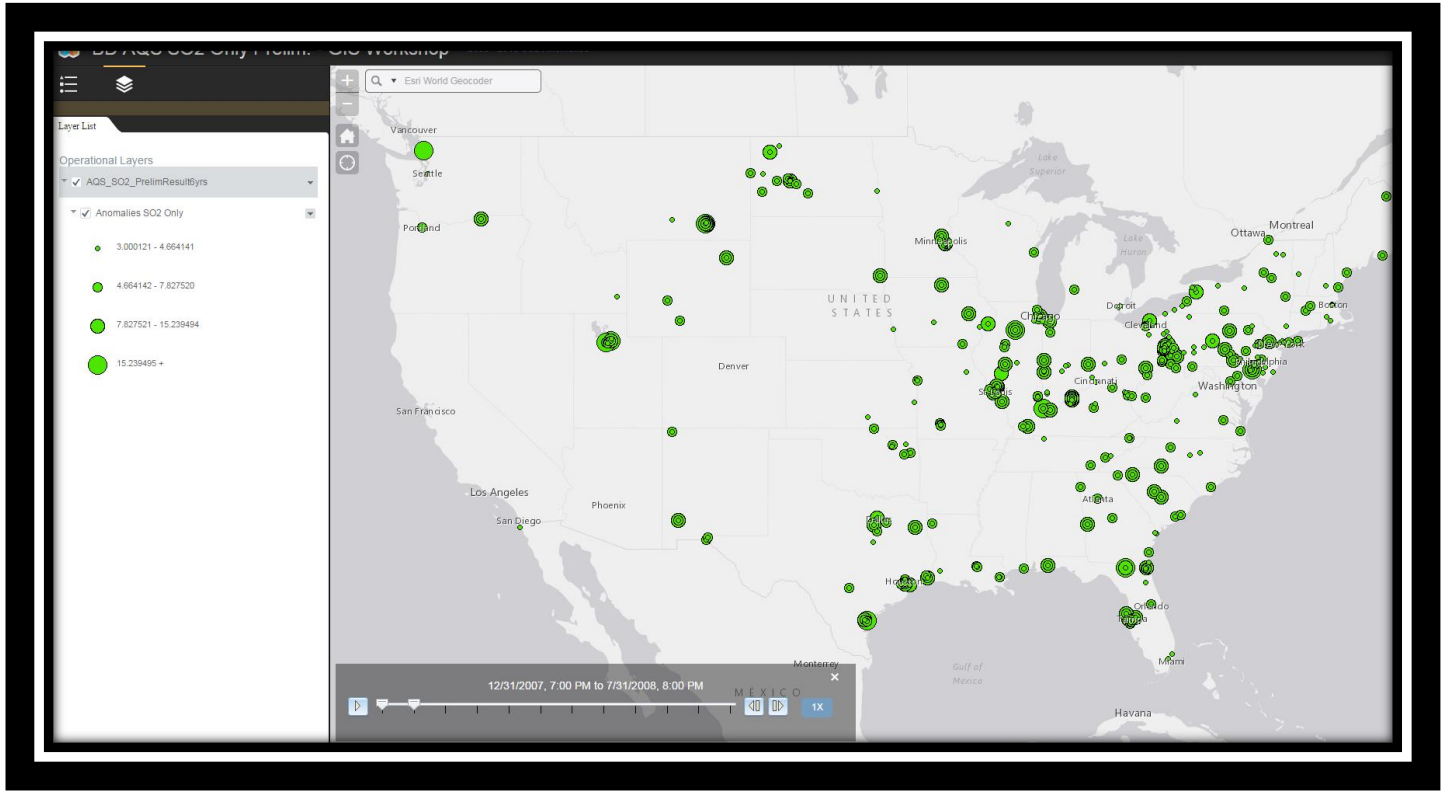
- Using wind speed/direction – search upwind of anomaly via targeted vector
  - Wildfires, oil spills, airports, industrial, road network, dust events, agriculture, etc. (mix of temporal and non-temporal datasets)
    - USCG NRC, USGS Fed. fires, EPA FRS, Esri Streets, EPA AQS (Wind), US Census Bureau, FAA NFDC
- One-to-Many with possible sources per anomaly time/location

CGI

Visualizing the Results

- ArcGIS Server Map Services
- ArcGIS Online Webmaps and Web App Builder
- Tableau Desktop and Server

Web App Builder w/Time Slider Widget and Quantitative Z Scores



Several other GIS products produced to visualize the analysis results as well

Tableau Workbook and Dashboard Anomalies Per Month, Per Pollutant

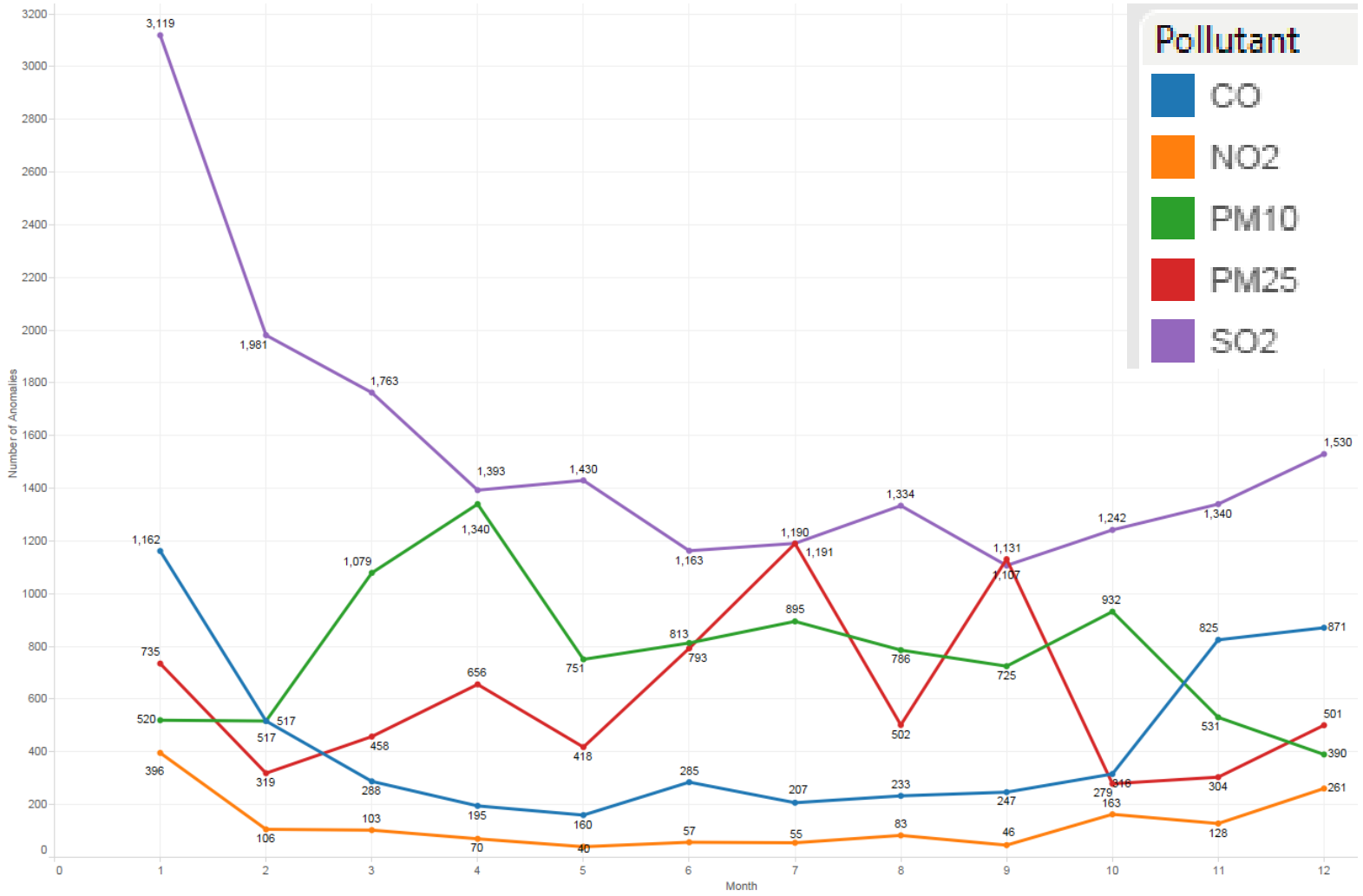
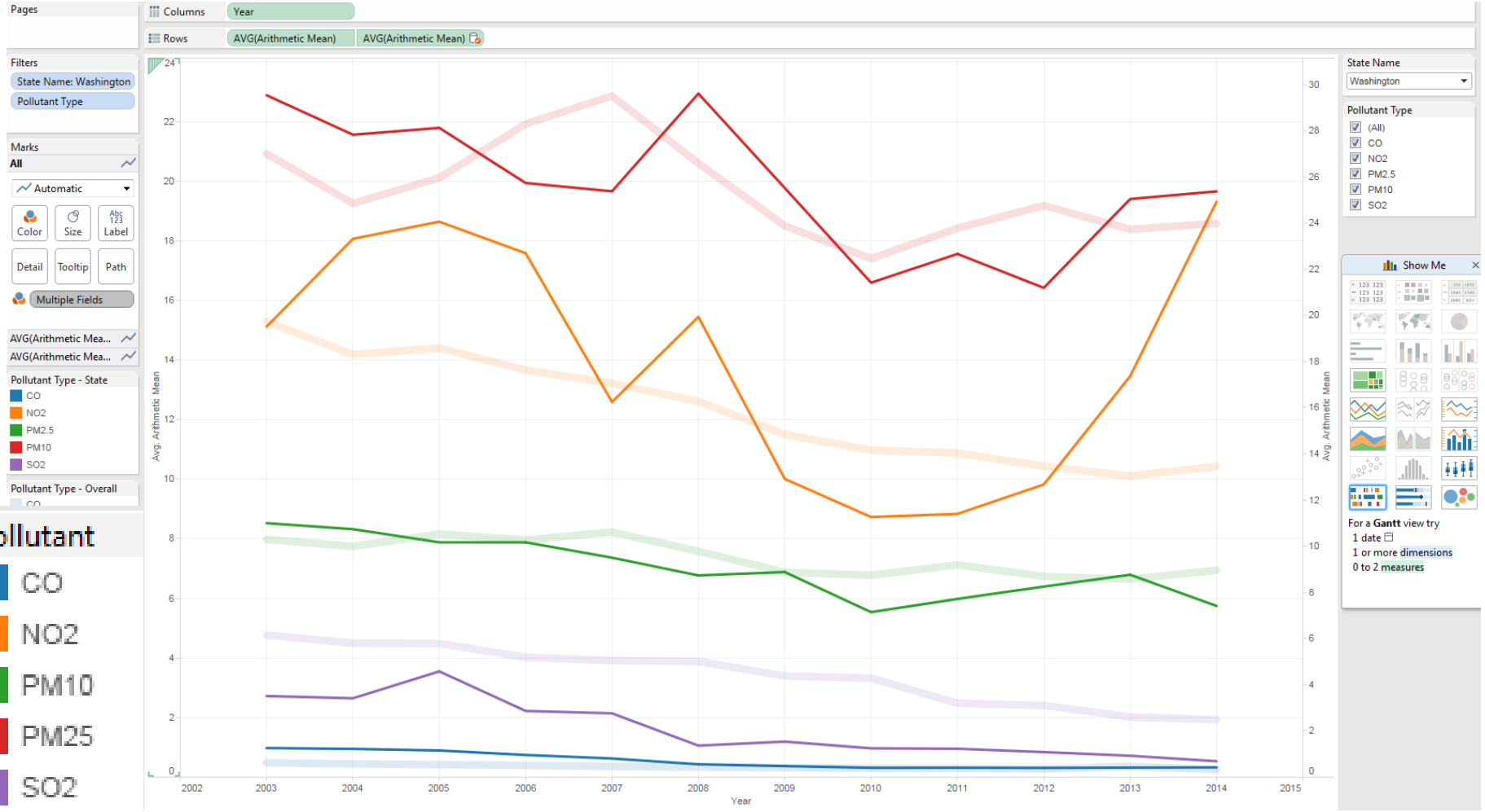


Tableau – Comparing states to national mean



Conclusion

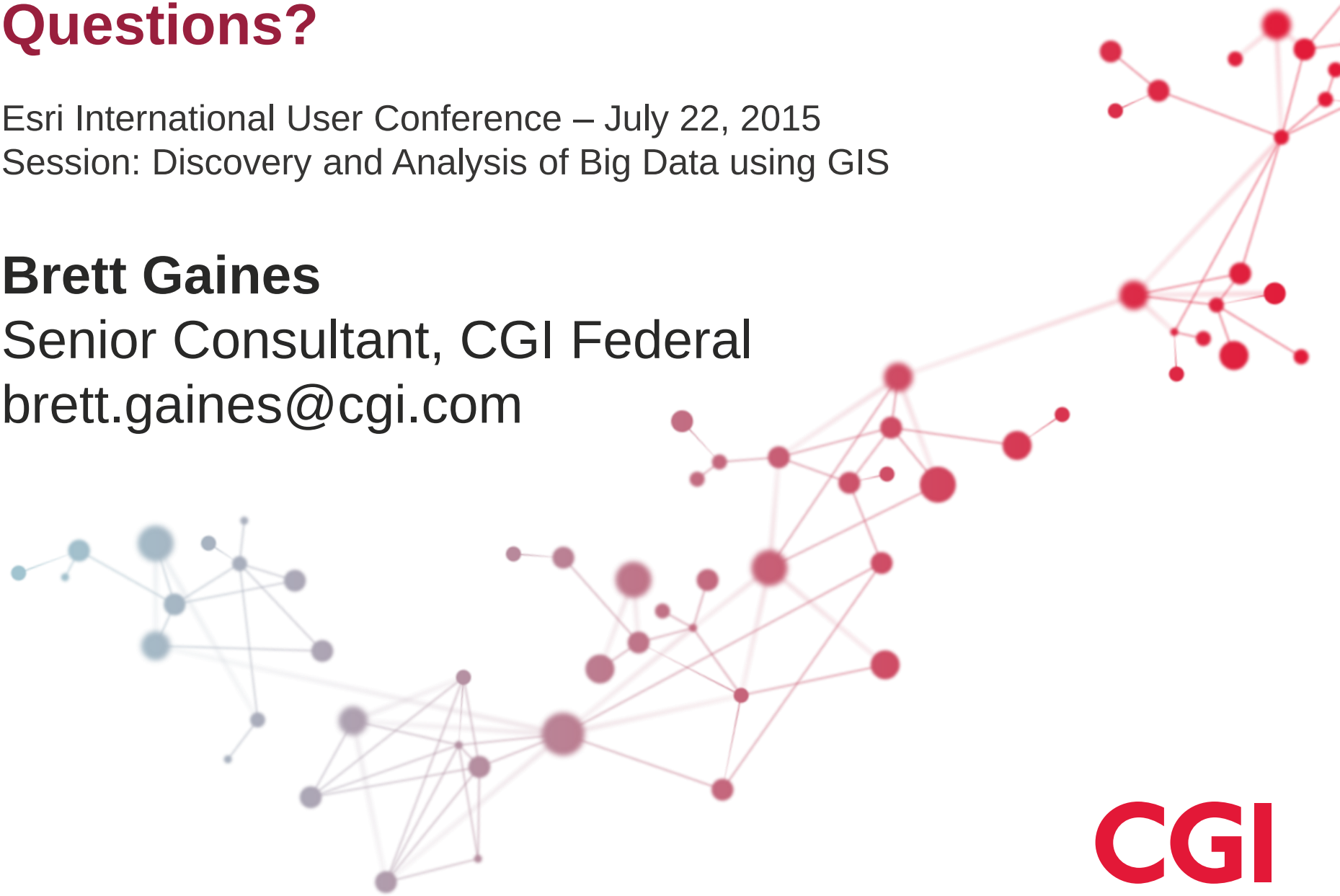
- Esri's open source GIS Tools for Hadoop allows large-scale distributed computing on spatial data
  - Utilize via Spark, Hive, Traditional MR
  - Esri Geometry API for Java is easily customized and extensible for particular use-cases
- https://github.com/Esri/geometry-api-java

CGI

Questions?

Esri International User Conference – July 22, 2015
Session: Discovery and Analysis of Big Data using GIS

Brett Gaines
Senior Consultant, CGI Federal
brett.gaines@cgi.com



CGI

Experience the commitment®