

Reforming Traditional Machine Learning Algorithms with Spatio-Temporal Analytics Capability for Big Data

Jing Xu, Lei Gao
IBM Analytics

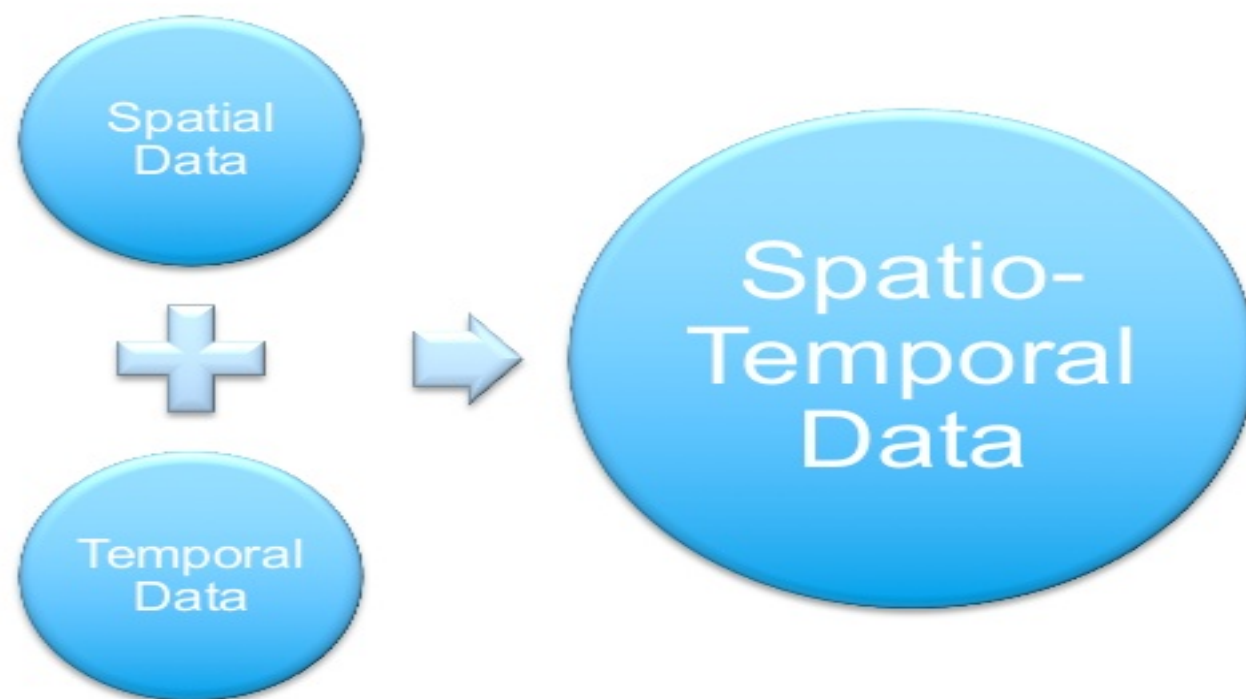
#SAISDS9

Content

- Spatio-Temporal Analysis Background
- Spatio-Temporal Exploratory and Modeling (STEM)
- Data Preparation for STEM
- Use case

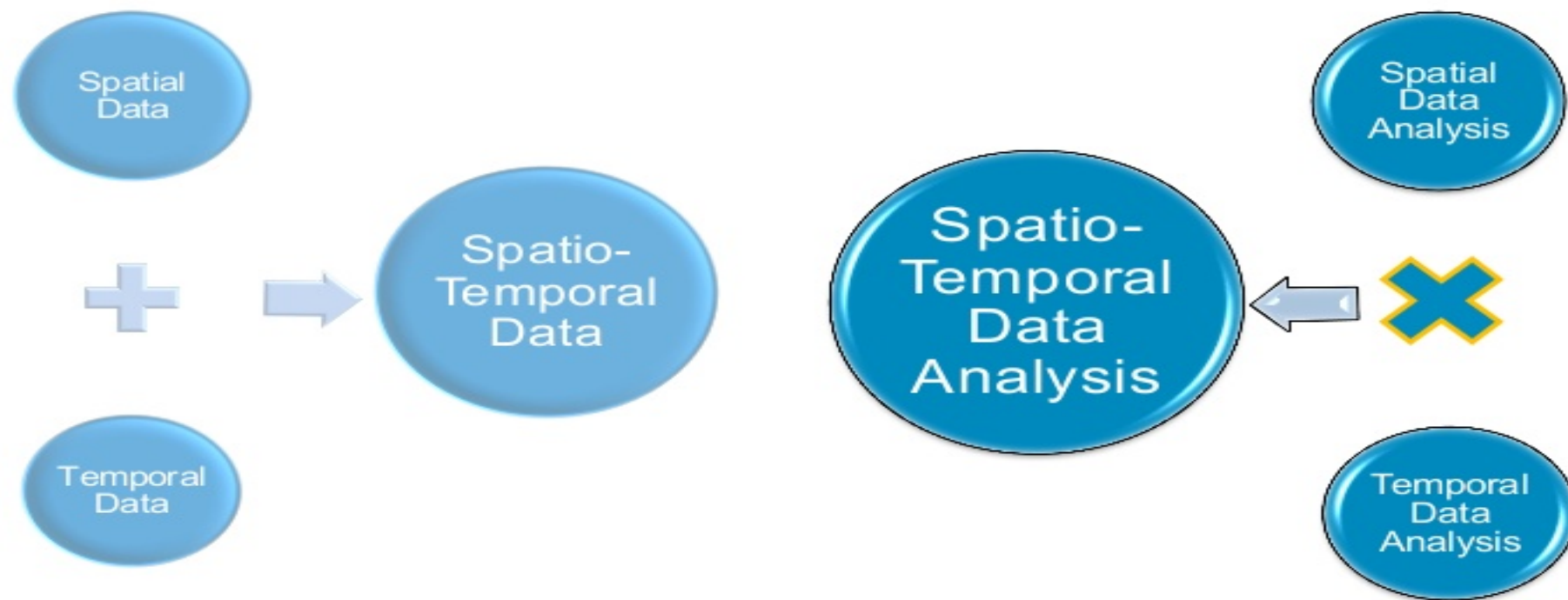
Background

- Spatio-Temporal Data



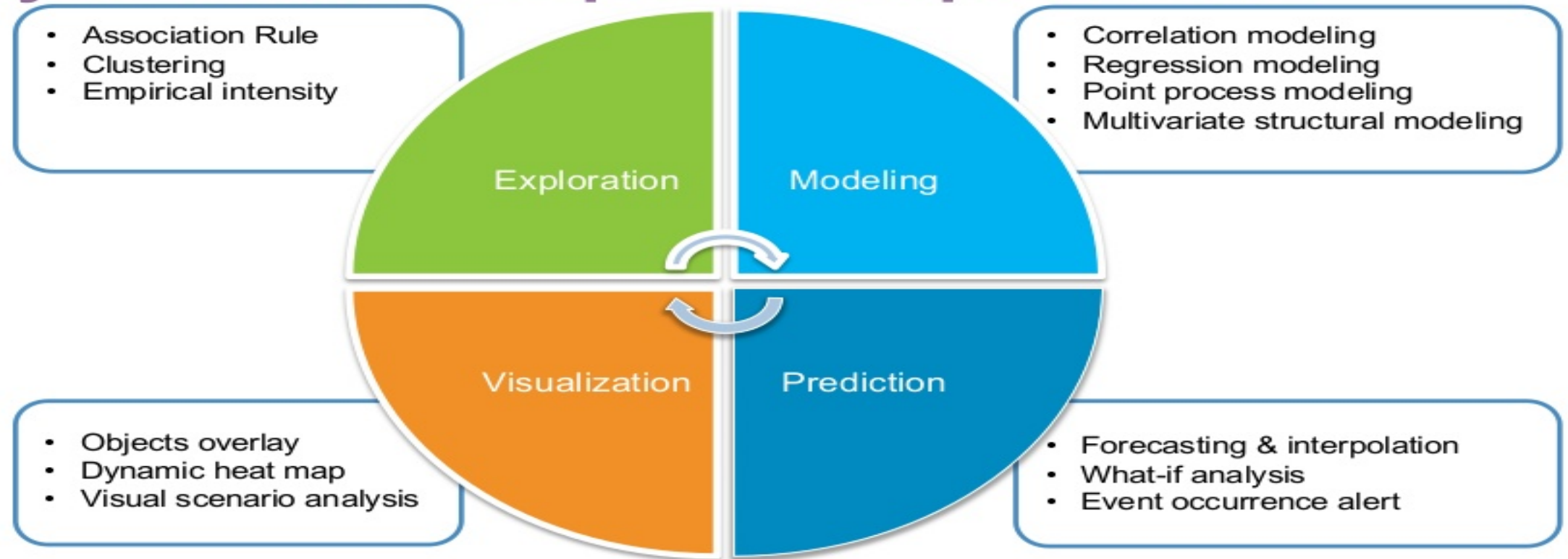
Background

- Spatio-Temporal Data & Analysis

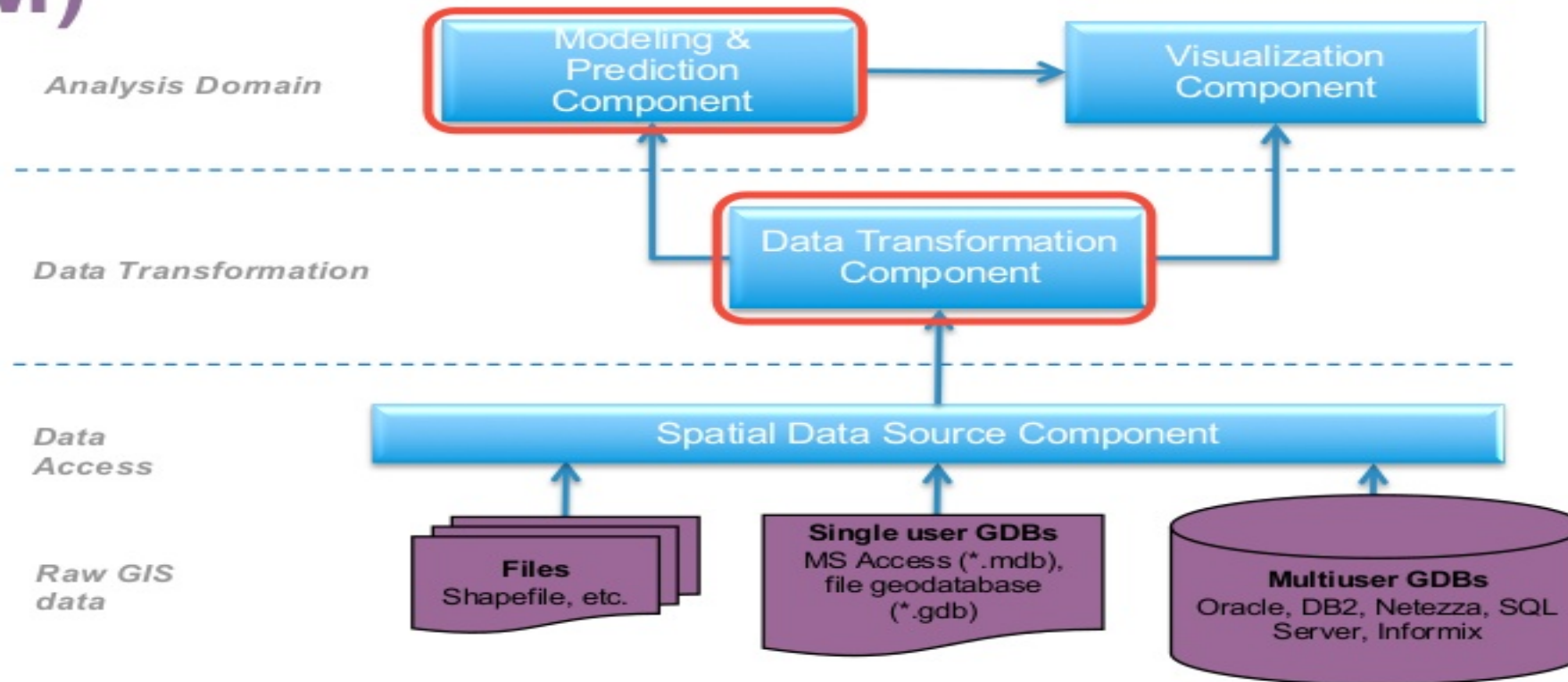


Background

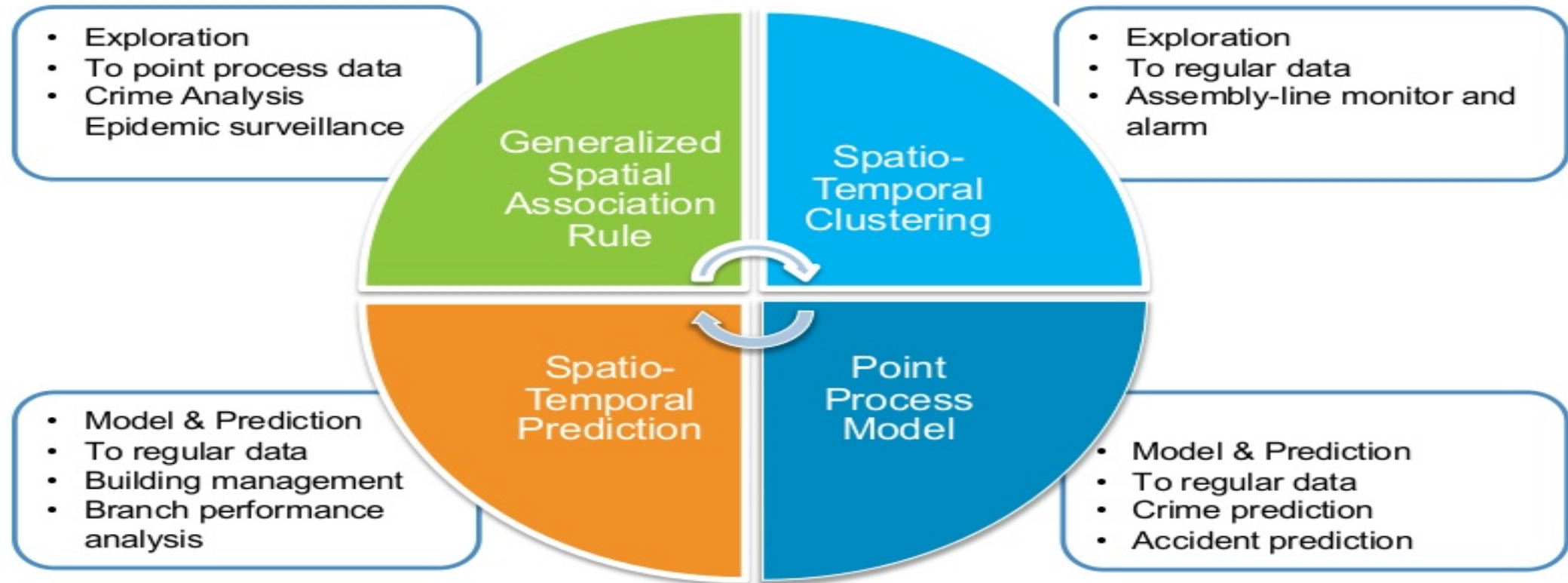
- Analysis Areas for Spatio-Temporal Data



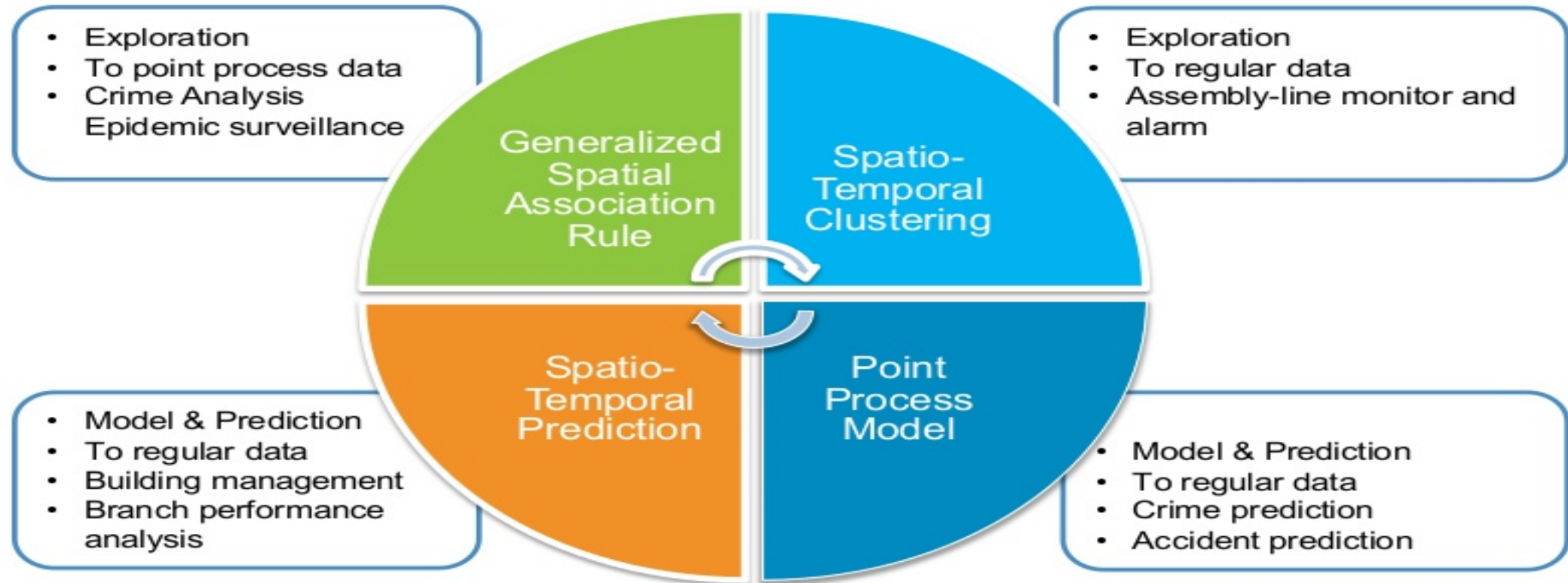
Spatio-Temporal Exploratory and Modeling (STEM)



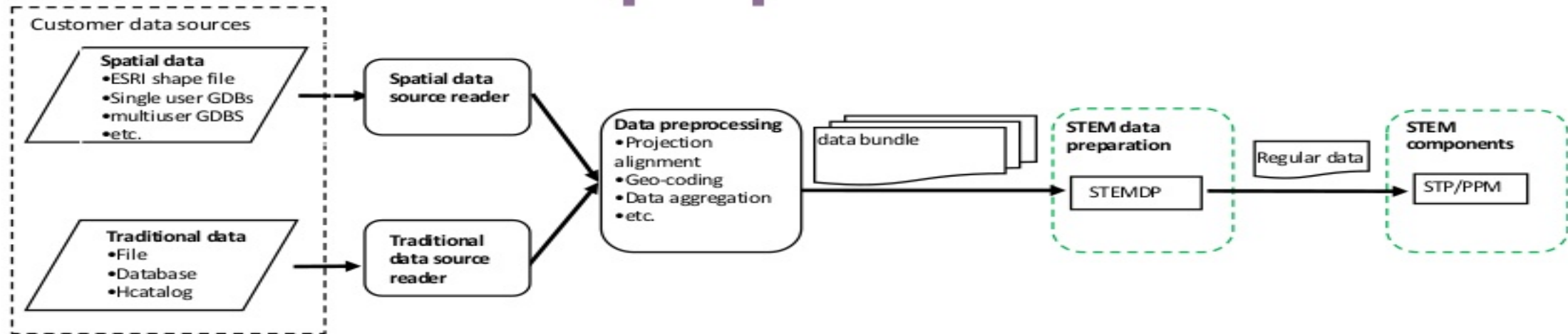
Spatio-Temporal Analysis Suite



Spatio-Temporal Analysis Suite



Work flow of data preparation



- The raw spatial-temporal data from customer involves multiple data sources with different data formats
- STEMDP component
 - Performs the data preparation for STP/PPM
 - Always required before STP/PPM model building
 - Provides the functionality that converts the raw data into regular data that STP/PPM requires
 - Outputs only one regular data source consumable by STP/PPM

Input data of STEM-DP

- **Regular data:** A fixed set of spatial locations and equally spaced time stamps common across locations. There is only one case for each location and time stamp combination

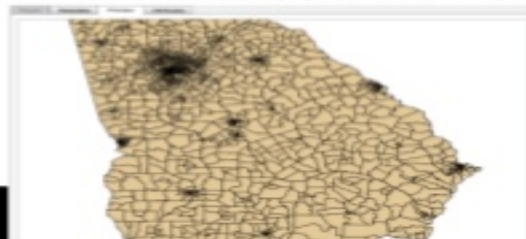
	A	B	C	D	E	F	G	H	I	J
1	longitude	latitude	year	inc	pop	pctw	pctb	pcta	pcth	age
2	-84.431	33.94616	1996	7845.61	181891.1	86.56196	8.80415	3.45804	3.18889	41.3
3	-84.4502	33.92186	1996	9890.284	175523.6	84.40528	10.96279	3.34503	3.24738	34.7
4	-84.4695	33.90059	1996	7375.113	171122.6	82.50816	13.07871	3.0717	3.26355	32.7
5	-84.4522	33.88529	1996	9070.971	160988	83.33412	12.69587	2.67122	3.14153	32.7

- **Point occurrence data:** A list of events labeled by a time stamp and the location of the event

	A	B	C	D
1	year	longitude	latitude	mat_addr
2	1996	-83.23	33.08	301 S WAYNE ST, MILLEDGEVILLE, GA, 31061
3	1996	-83.25	33.11	1900 N COLUMBIA ST, MILLEDGEVILLE, GA, 31061
4	1996	-83.47	33.6	223 N MAIN ST, MADISON, GA, 30650
5	1996	-83.47	33.6	223 N MAIN ST, MADISON, GA, 30650
6	1996	-82.6	31.86	901 S TALLAHASSEE ST, HAZLEHURST, GA, 31539

- **Geospatial layout data:**

- Required if the regular data or point occurrence data requires the region boundary information
- Shape files(*.shp and associated *.dbf, *.prj), JSON format, Geo-enabled databases, etc.
- Convert all of them into unenclosed JSON format that can be used in parallel computing



	ADID	PERMIT ID	FILED DATE	FILED DATE	COUNT	STATUS	STATUS	STATUS	STATUS
1	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
2	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
3	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
4	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
5	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
6	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
7	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
8	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
9	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
10	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
11	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
12	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
13	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
14	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
15	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
16	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
17	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
18	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
19	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
20	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
21	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
22	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
23	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
24	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
25	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
26	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
27	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
28	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
29	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
30	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
31	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
32	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
33	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
34	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
35	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
36	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
37	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
38	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
39	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
40	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
41	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
42	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
43	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
44	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
45	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
46	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
47	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
48	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
49	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
50	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
51	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
52	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
53	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
54	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
55	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
56	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
57	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
58	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
59	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
60	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
61	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
62	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
63	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
64	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
65	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
66	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
67	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000

Data transformation in STEM-DP

- **Data transformation for Regular data**
 - Convert timestamp variable into time-index variable
 - Missing records handling
- **Data transformation for Point occurrence data**
 - **For STP:** Perform kernel-density estimation algorithm by using the point occurrence data
 - **For PPM:** Obtain the centroid of each region and compute the event count in the region by using the point occurrence data. The **count** will be a new field for PPM
 - Missing records handling



** Optional inputs depends on the actual scenario*

Scala API

com.ibm.spss.ml

Score

SmartReports

SmartReportsModel

VectorTest

VectorTransform

com.ibm.spss.ml.attribute

Attribute

AttributeImplicits

AttributeType

BinaryAttribute

DataStorage

NominalAttribute

NumericAttribute

OutputAttribute

ResultFeature

Role

RuleFeature

TypelessAttribute

UnresolvedAttribute

com.ibm.spss.ml.classificationandregression

GeneralizedLinear

GeneralizedLinearModel

LinearRegression

LinearRegressionModel

LinearSupportVectorMachine

LinearSupportVectorMachineModel

com.ibm.spss.ml.classificationandregression

RandomForest

RandomForestModel

com.ibm.spss.ml.classificationandregression

MiscellaneousCosts

MiscellaneousCostsCalculator

root package

package root

This is the documentation for the IBM SPSS Spark Machine Learning standard library.

Package structure

Notable packages include:

- com.ibm.spss.ml
 - com.ibm.spss.ml.classificationandregression
 - com.ibm.spss.ml.classificationandregression.ensemble
 - com.ibm.spss.ml.classificationandregression.tree
 - com.ibm.spss.ml.clustering
 - com.ibm.spss.ml.common
 - com.ibm.spss.ml.common.params
 - com.ibm.spss.ml.datapreparation
 - com.ibm.spss.ml.datapreparation.binding
 - com.ibm.spss.ml.datapreparation.sampling
 - com.ibm.spss.ml.forecasting
 - com.ibm.spss.ml.forecasting.traditional
 - com.ibm.spss.ml.frequencybasedmining
 - com.ibm.spss.ml.survivalanalysis
 - com.ibm.spss.ml.util
 - Score
 - SmartReports

Example Code

Build EventBasedTimeSeriesPatternFinding model and score it in a Jupyter notebook

```
import java.sql.Date
case class SimpleData(customerid: String, times: java.sql.Date)

val sqlContext = new org.apache.spark.sql.SQLContext(sc)
import sqlContext.implicits._

val data = sqlContext.createDataFrame(Seq(
  SimpleData("3", Date.valueOf("2012-6-1")), "Hotel", 131),
  SimpleData("5", Date.valueOf("2012-6-3")), "Flight", 119),
  SimpleData("4", Date.valueOf("2012-6-5")), "Hotel", 130),
  SimpleData("4", Date.valueOf("2012-6-7")), "Flight", 111),
  SimpleData("4", Date.valueOf("2012-6-9")), "Flight", 11),
  SimpleData("5", Date.valueOf("2012-6-11")), "Flight", 119))
```

IBM Watson

Log In

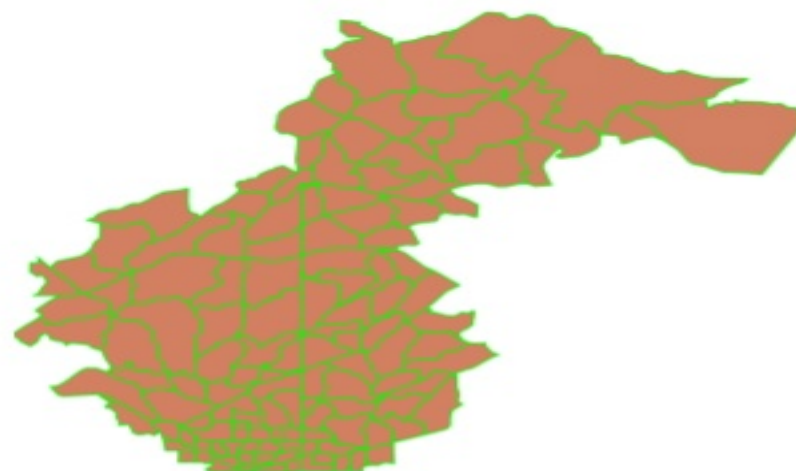
Sign Up

In [20]:
from spss.ml.spatiotemporal.spatiotemporalprediction import SpatioTemporalPrediction
from spss.ml.spatiotemporal.spatiotemporalprediction import SpatioTemporalPredictionModel
from spss.ml.common.wrapper import SPSSJavaWrapper

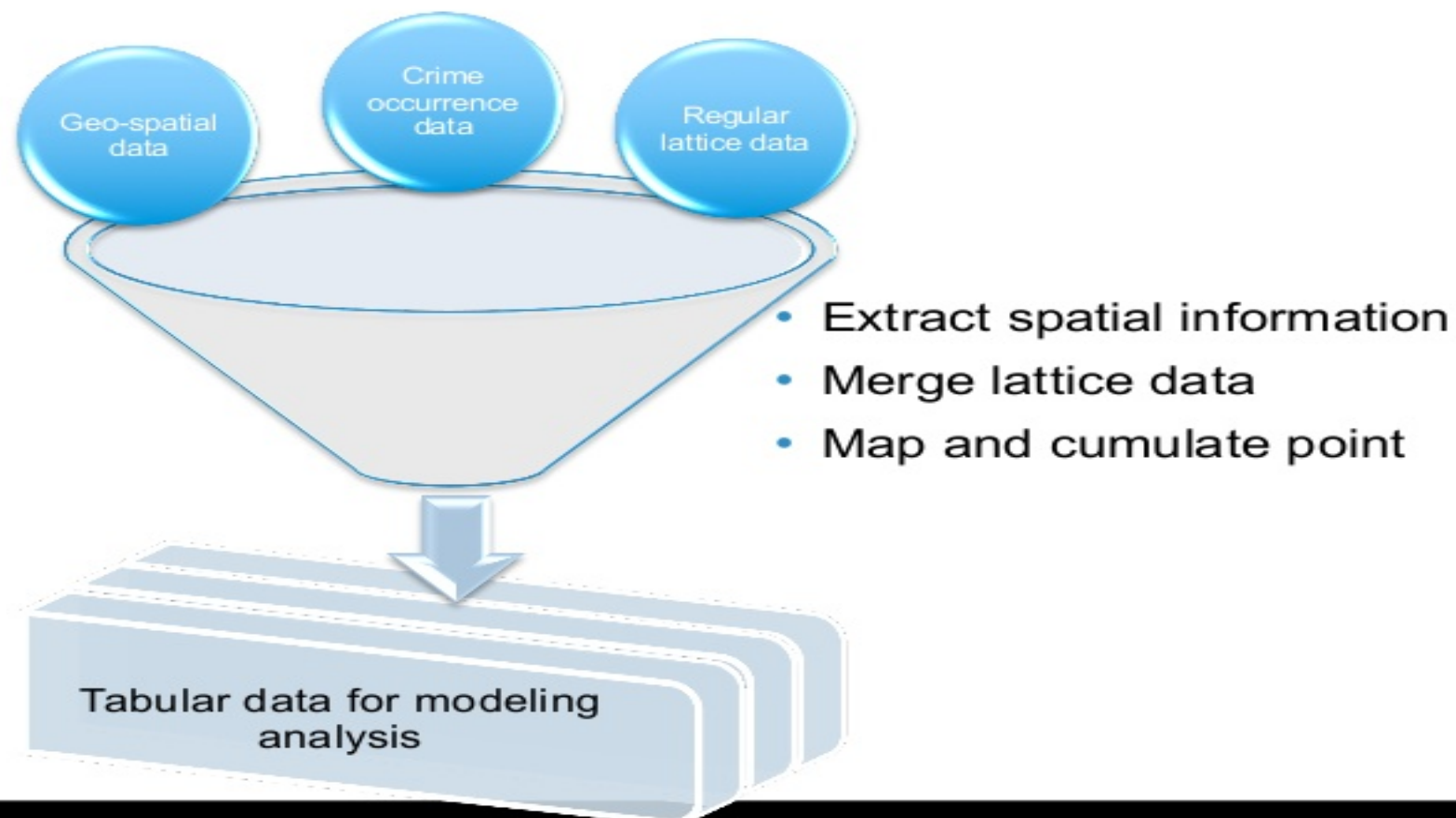
In [21]:
#stp = SpatioTemporalPrediction().setTargetField("aqi").setIntercept(True).setInputFieldList(["temp_min", "temp_max", "temp_ave", "humidity_min", "humidity_max", "humidity_ave", "windid", "winpid", "westid"])
stp = SpatioTemporalPrediction().setTargetField("aqi").setIntercept(True).setInputFieldList(["temp_min", "temp_max", "temp_ave", "humidity_min", "humidity_max", "humidity_ave", "windid", "winpid", "westid"])
.setLocationFieldList(["Locx", "Locy"]).setTimeIndexField("Time")
.setLag(1)
model = stp.fit(edf_model)

Use Case

- Crime Occurrences Modeling and Prediction
 - Police department in a city want to predict crime occurrences of next several months in order to better plan and allocate resources
- Available information for analysis
 - Geo-spatial data defined census tracts of the city
 - Past crime occurrences
 - The time and coordinates of crime event
 - Local demographic profiles
 - population density
 - per capita income
 - ethnic diversity
 - median age
 - male-to-female ratio



Data Preparation



Prepared Data for Modeling

- Align crime occurrences points with layout, and merge with demographic profiles to get the model ready data for PPM

	A	B	C	D	E	F	G	H	I	J	K
1	Time.Index	Tract.ID	Longitude	Latitude	Area	log_Popula	log_Incom	Diversity	Age	MFRatio	Counts
2	1	130670303.2	-84.43097	33.94616	0.001159	18.86818	10.41436	-413.322	42	97.189	628
3	1	130670303.2	-84.45015	33.92186	0.000906	19.05783	10.27308	-404.59	35	91.049	161
4	1	130670303.4	-84.46951	33.90059	0.000645	19.39024	10.48969	-406.498	32	104.713	285
5	1	130670303.4	-84.45218	33.88529	0.000418	19.77336	10.38474	-408.388	32	105.296	151
6	1	130670311.1	-84.48765	33.89422	0.000391	19.91062	10.12567	-398.619	32	96.5612	107
7	1	130670312	-84.47668	33.87345	0.000638	19.31286	10.36606	-398.038	34	85.4478	99
8	1	130670312	-84.47881	33.85592	0.000902	18.92945	10.1344	-389.845	34	92.8024	270
9	1	131210001	-84.35463	33.79865	0.000316	20.65016	10.20267	-380.088	40	100.368	94
10	1	131210002	-84.36368	33.79139	0.000285	20.79336	10.17474	-376.83	38	126.696	131

Time dimension:
'time.index' is encoded time point when the aggregated counts are occurs in one time interval

Space dimension:
'Longitude', 'Latitude' are coordinates where the demographic data are belong to

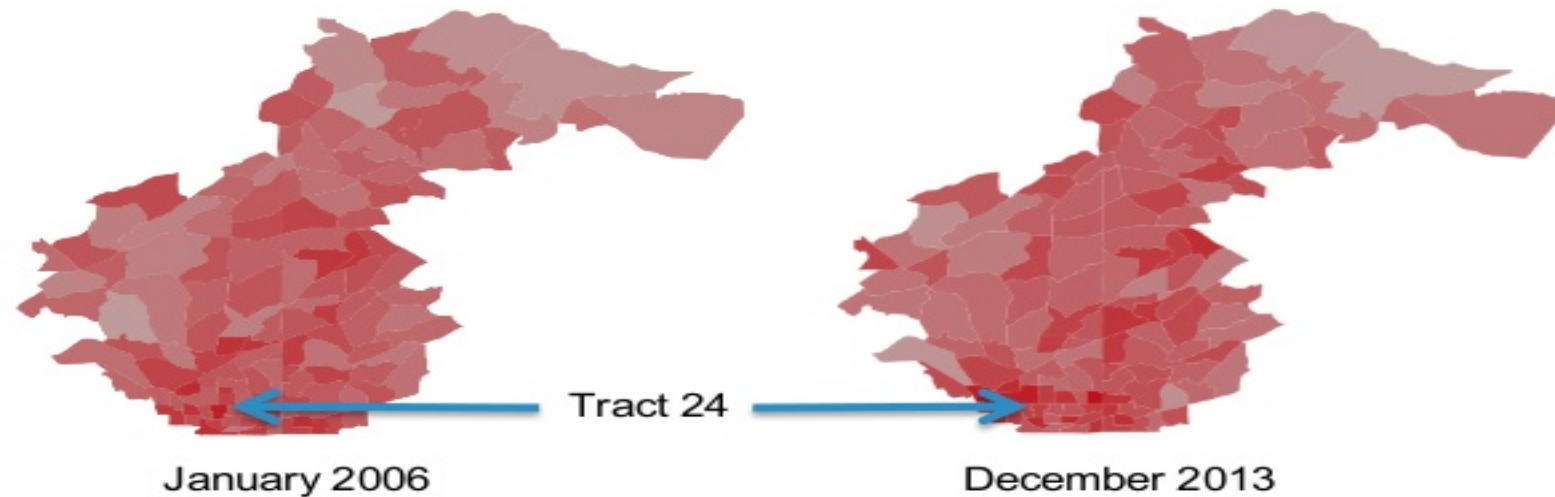
Inputs:
There demographic input are the external predictors used in PPM

Target:
Crime counts for each area as model building target

Prepared Data Visualization

- Event occurring intensity

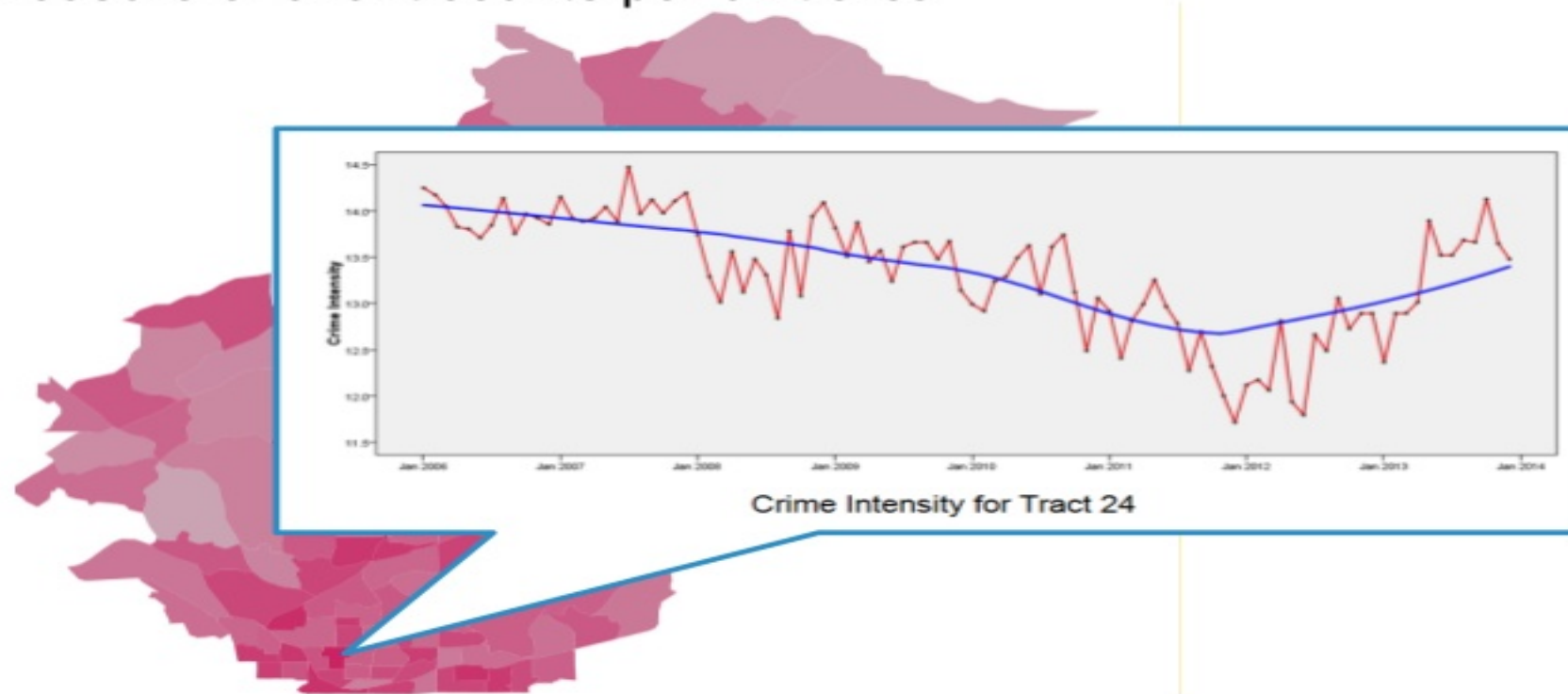
- Intensity is a measure of event counts per unit area



Prepared Data Visualization

- Event occurring trend

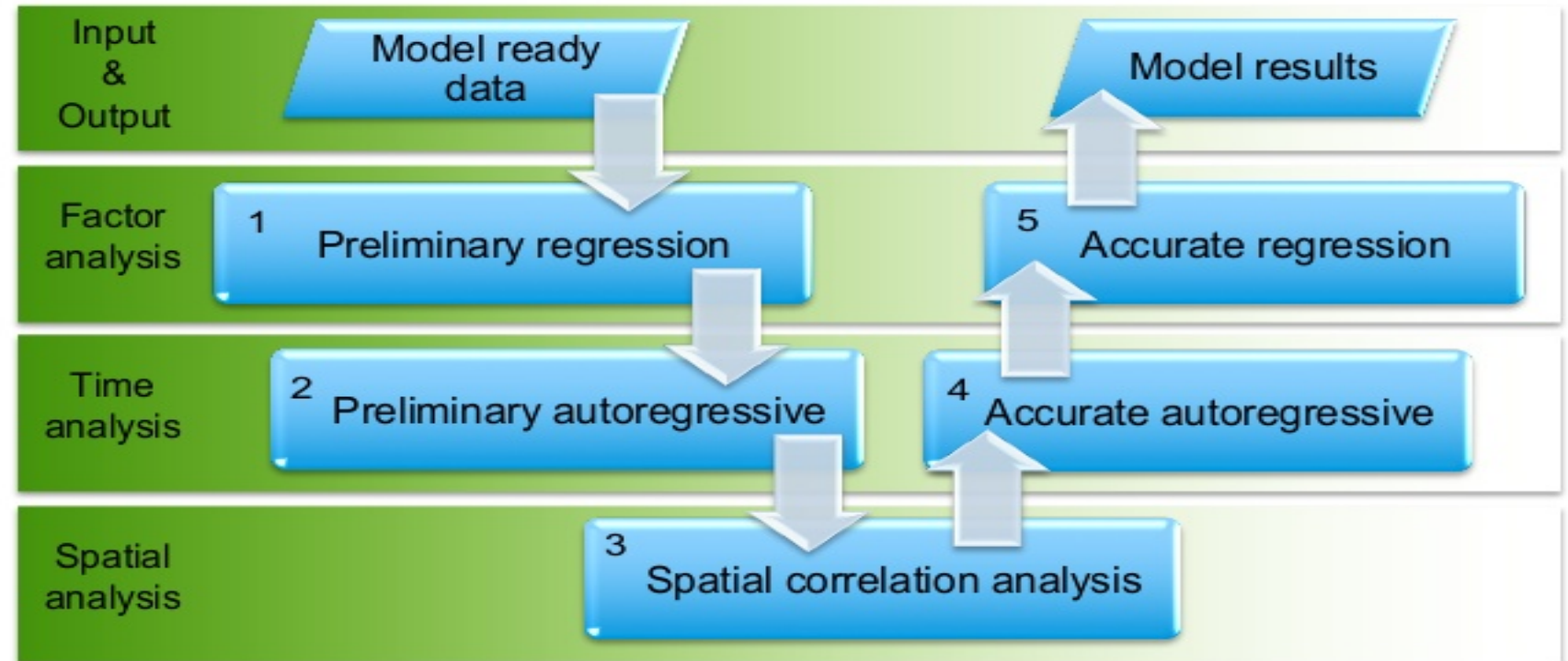
- Intensity is a measure of event counts per unit area



Spatio-Temporal Modeling

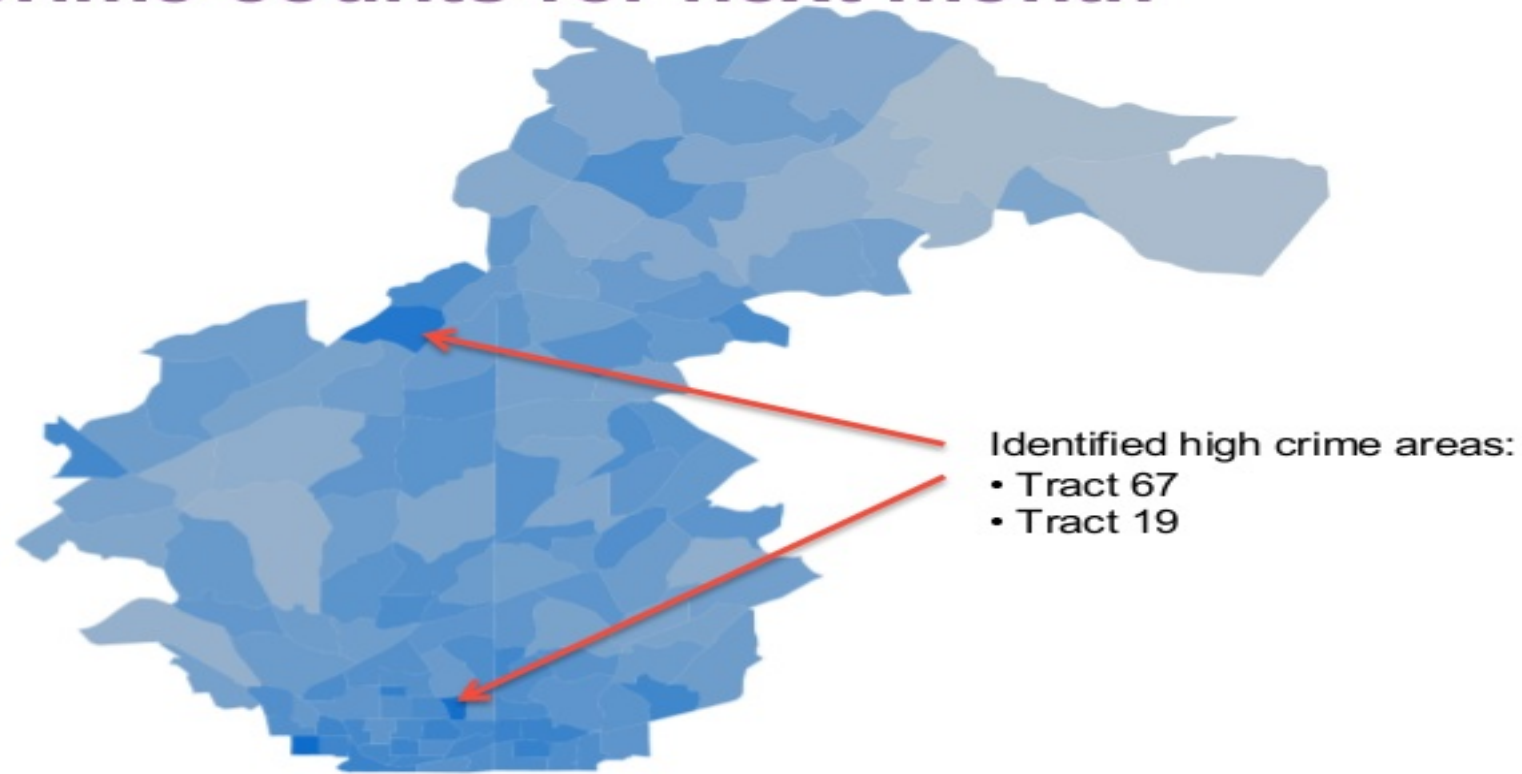
Modeling process including 3 layers to handle different types of information:

- The influence of external factors
- Time-series autocorrelation
- Spatial correlation among all the cities



Prepared Result Visualization

- Predicted crime counts for next month



What-if Scenario Analysis

- Scenario

- A large sports festival will be held in the city
- Police department wants to estimate how the crime will increase due to
 - Large influx of people to some areas
 - Fans are mainly adult males



Thanks!