Open Geospatial Machine Learning

Kevin Stofan

Agenda

- Intro
- Spatial Data Models and Formats
- Spatial Data I/O
- Spatial Autocorrelation
- Spatial Weights Matrix
- Exploratory Spatial Data Analysis (ESDA)
- Rest of the Geospatial ML Workflow
- Related Studies and Competitions
- Discussion

Goals

- Introduce geospatial machine learning workflow
- Exposure to Python spatial tools
- Spatially-explicit modeling
- External resources and further reading

About Me

- Customer Facing Data Scientist at DataRobot
  - Pre- and Post- sales support for customers
  - Assist product and engineering teams with geospatial features
  - Consult customers with geospatial use cases
- Adjunct Professor at Penn State
  - Graduate level Geographic Information Systems (GEOG884)
  - Raster and vector data analysis
  - FOSS4G
- Applied Spatial Analysis
  - Point Pattern Analysis
  - Spatial Econometrics and geostatistics

Repos and Contact

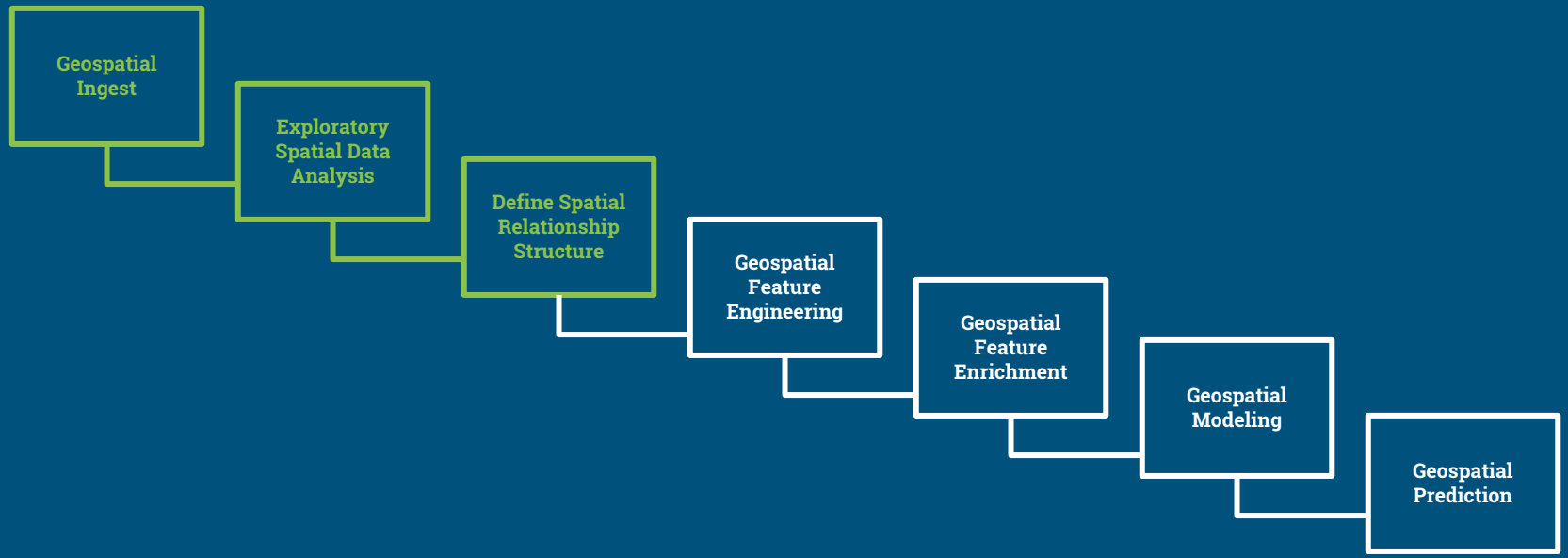
https://github.com/TankofVines/data_intel

https://github.com/TankofVines/odsc

kevin.stofan@gmail.com

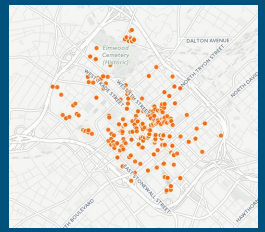
@tankofvines

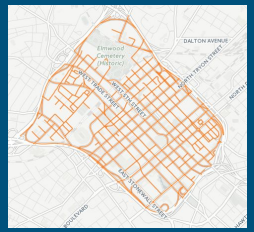
Geospatial ML Workflow

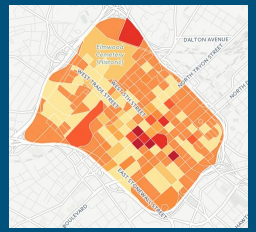


Geospatial ML Techniques

Geospatial Data Ingestion



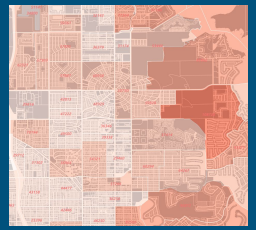




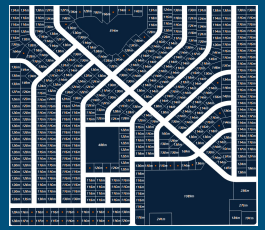
KML SHP JSON GPX WKT WBT GML

Spatial Feature Engineering/Enrichment

Spatial Imputation

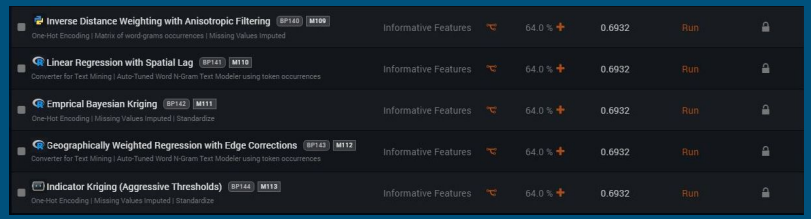


Feature Enrichment and Dasymetric Mapping



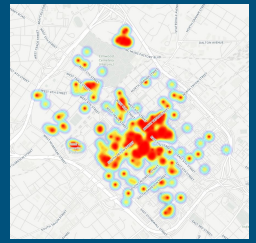
Derived features and transformations (e.g. area, centroid, contiguity, etc.)

Spatially-explicit Models

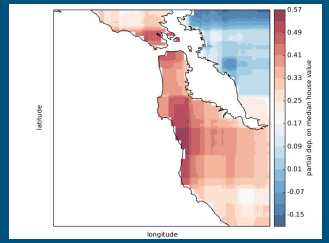


Spatial econometric, geostatistical, geographically weighted regression models

Geospatial Visualization

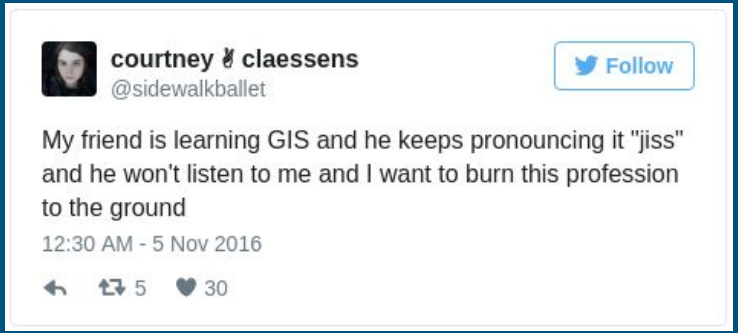


Heatmaps, kernel density estimates, and hexagonal binning



Two-way partial dependence plots using coordinates

A Note on Terminology



Tool: QuantumGIS

- Geographic Information Systems (GIS)
  - Collection - Maintenance - Storage - Analysis - Output - Distribution
  - Handles vector and raster data models
  - Historically dominated by ESRI and ArcGIS

http://www.qgis.org/en/site/index.html

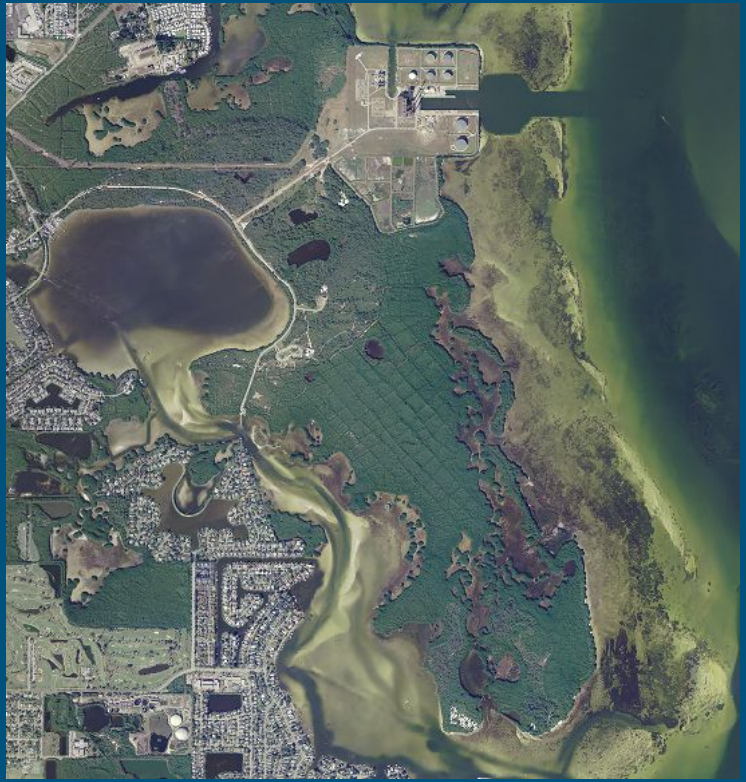
Spatial Data Models

Real World Entities

Vector Objects

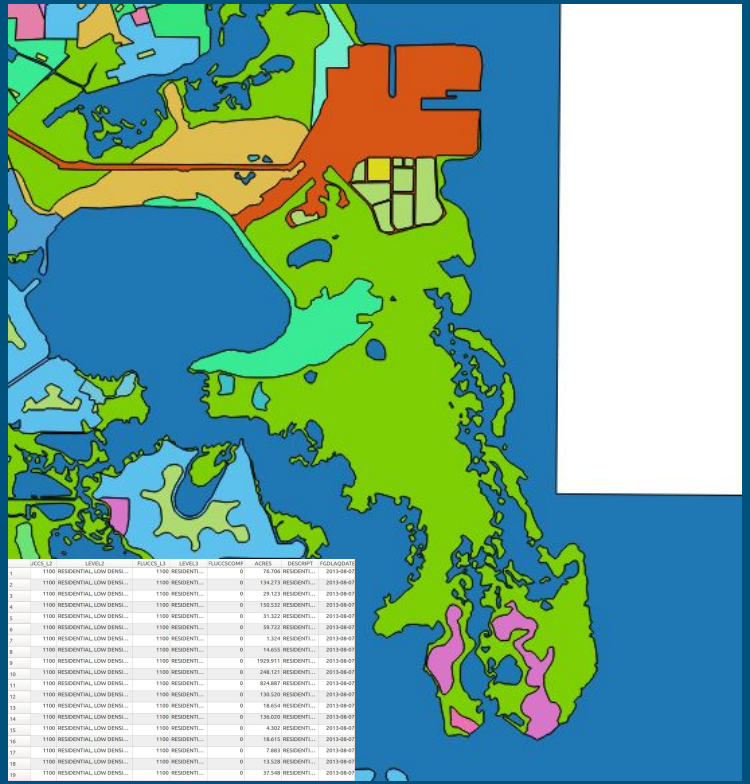
Pixel-based Raster

Spatial Data Models



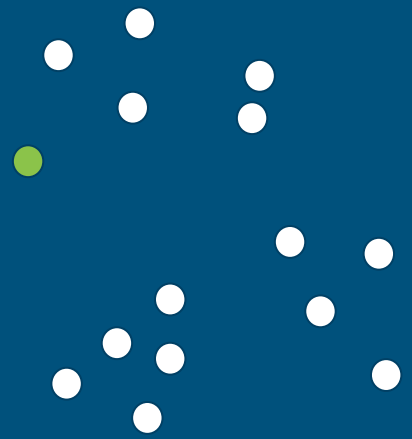
Spatial Data Models

Spatial Data Models



	JCCS_L2	LEVEL2	FLUCCS_L3	LEVEL3	FLUCCSCOMP	ACRES	DESCRIPT	FGDLAQDATE
1	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	76.706	RESIDENTI...	2013-08-07
2	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	134.273	RESIDENTI...	2013-08-07
3	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	29.123	RESIDENTI...	2013-08-07
4	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	150.532	RESIDENTI...	2013-08-07
5	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	31.322	RESIDENTI...	2013-08-07
6	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	59.722	RESIDENTI...	2013-08-07
7	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	1.324	RESIDENTI...	2013-08-07
8	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	14.655	RESIDENTI...	2013-08-07
9	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	1929.911	RESIDENTI...	2013-08-07
10	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	248.121	RESIDENTI...	2013-08-07
11	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	824.887	RESIDENTI...	2013-08-07
12	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	130.520	RESIDENTI...	2013-08-07
13	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	18.654	RESIDENTI...	2013-08-07
14	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	136.020	RESIDENTI...	2013-08-07
15	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	4.302	RESIDENTI...	2013-08-07
16	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	18.615	RESIDENTI...	2013-08-07
17	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	7.883	RESIDENTI...	2013-08-07
18	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	13.528	RESIDENTI...	2013-08-07
19	1100	RESIDENTIAL, LOW DENSI...	1100	RESIDENTI...	0	37.548	RESIDENTI...	2013-08-07

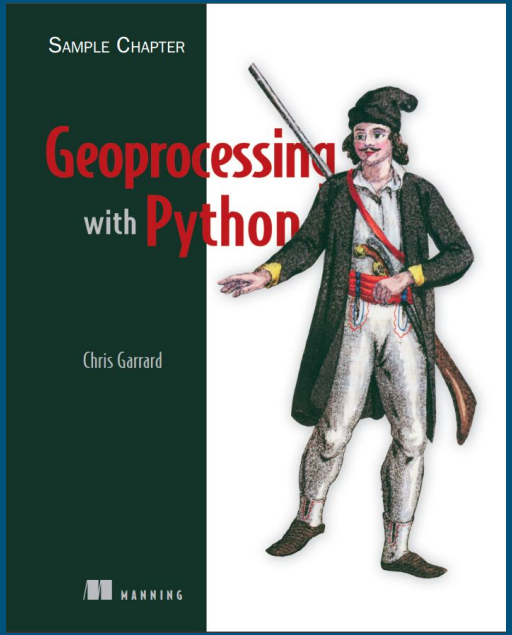
Vector Data Types



Vector Data Formats

- File-based
  - ESRI Shapefiles
  - GeoJSON
  - File Geodatabase
- Database
  - PostGIS Table
  - SpatialLite
  - Various proprietary DBs (Oracle, Mongo, MS SWL Server)
- Binary/Text
  - WKT/WKB

Tool: OGR/GDAL



https://www.manning.com/books/geoprocessing-with-python#downloads

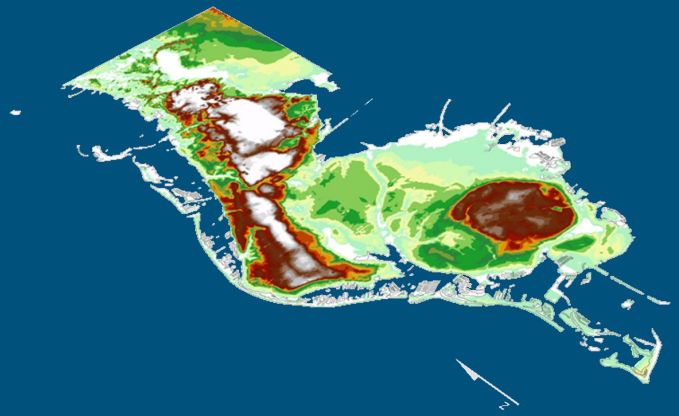
Tool: Geopandas

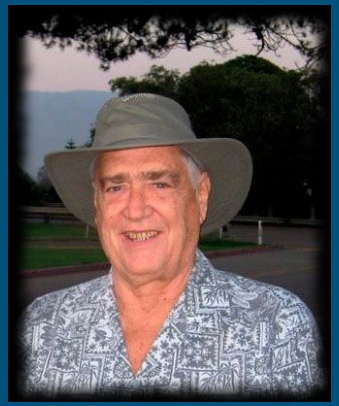
- Python Geospatial Library
  - I/O
  - Feature Engineering
  - Visualization

Spatial Autocorrelation

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

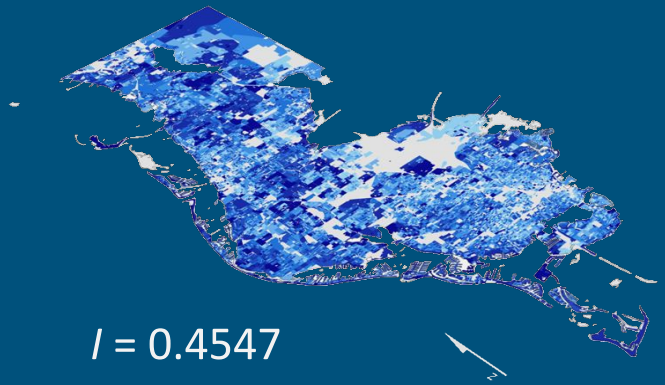
- Waldo Tobler





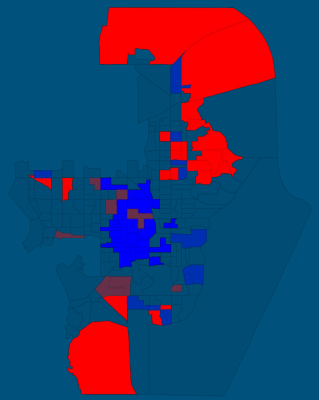
Global Spatial Autocorrelation

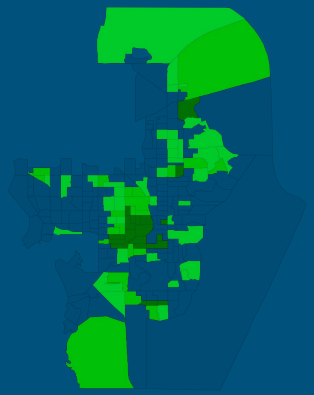
$$I = \left[\frac{\text{observations}}{\text{normalize for overall dataset variance}} \right] \times \left[\frac{\text{weighted covariance}}{\text{total map spatial neighbor weight}} \right]$$



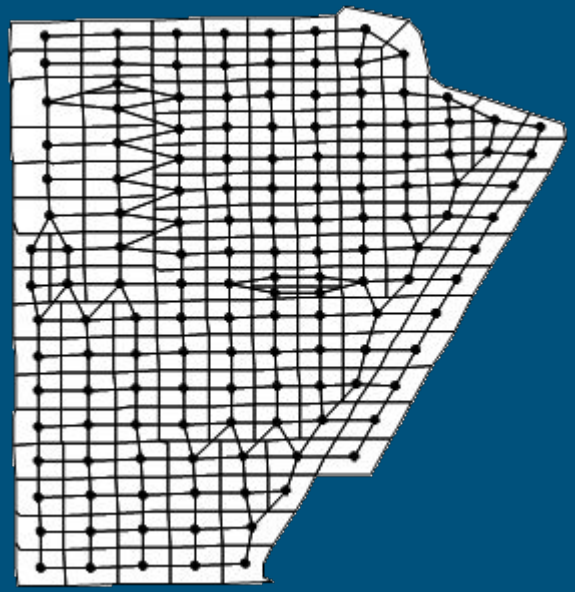
$I = 0.4547$

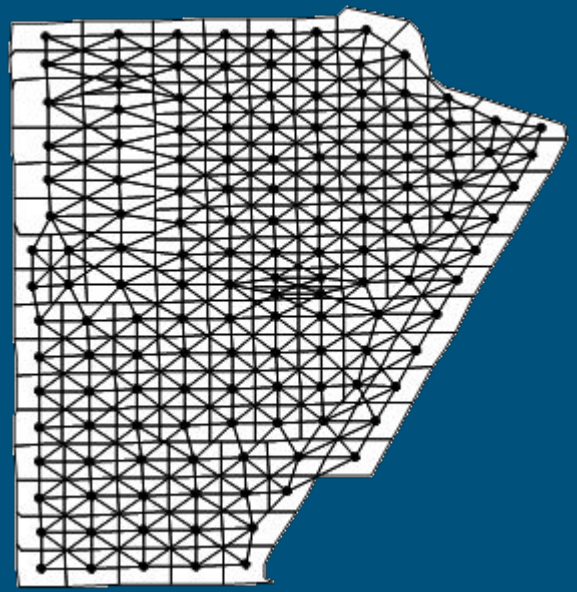
Local Spatial Autocorrelation





Spatial Weights Matrix



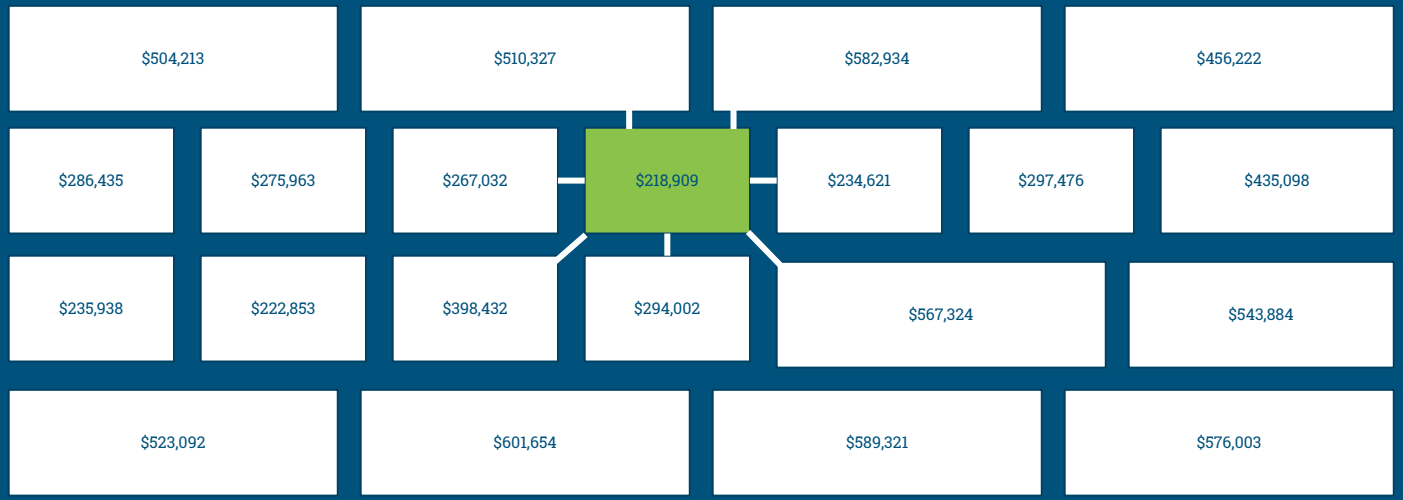


Tool: GeoDa

- Exploratory Spatial Data Analysis (ESDA)
  - Desktop
  - Spatial Autocorrelation
  - Spatial Regression

http://geodacenter.github.io/

Spatial Lag


$504,213
$510,327
$582,934
$456,222
$286,435
$275,963
$267,032
$218,909
$234,621
$297,476
$435,098
$235,938
$222,853
$398,432
$294,002
$567,324
$543,884
$523,092
$601,654
$589,321
$576,003

Higher Order Weights

$504,213

$510,327

$582,934

$456,222

$286,435

$275,963

$267,032

$218,909

$234,621

$297,476

$435,098

$235,938

$222,853

$398,432

$294,002

$567,324

$543,884

$523,092

$601,654

$589,321

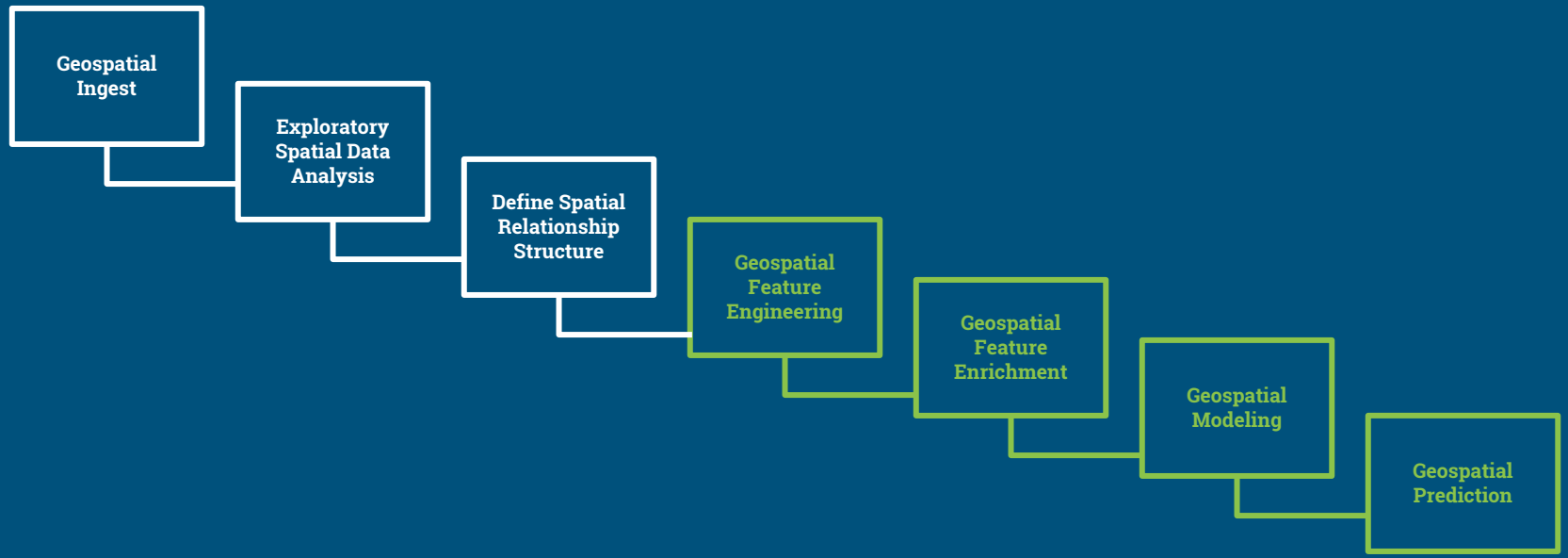
$576,003

Tool: PySAL

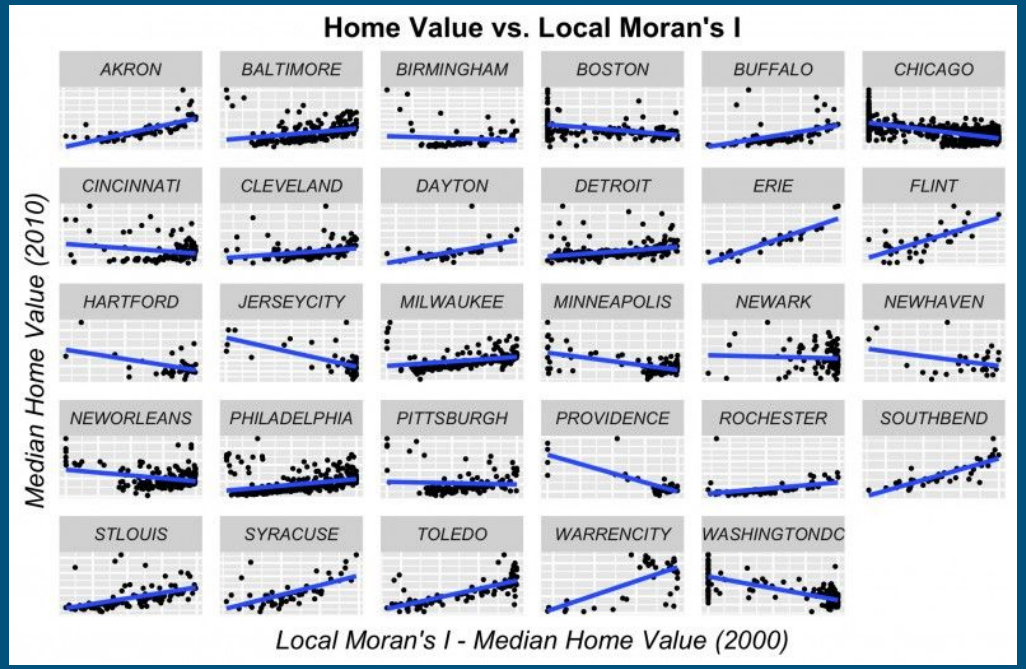
- Python Library for ESDA
  - GeoDa Equivalent
  - Spatial weights
  - Spatial lag, smoothing, regionalization, and more

http://pysal.readthedocs.io/en/v1.11.0/#

Geospatial ML Workflow

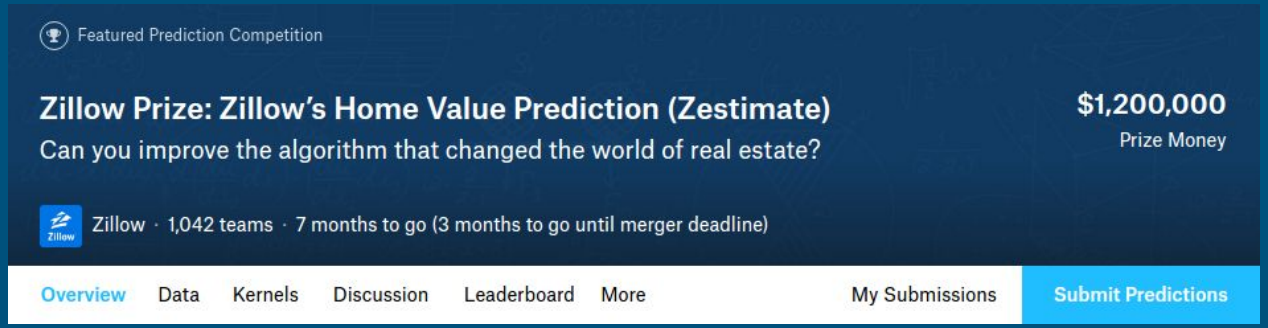


Further Reading: Urban Spatial



http://urbanspatialanalysis.com/portfolio/predicting-gentrification-using-longitudinal-census-data/

Further Reading: Kaggle Zillow Zestimate



https://www.kaggle.com/c/zillow-prize-1

Discussion