SIGGIS Demonstration: Intersection of Social Media Analytics and GeoAI

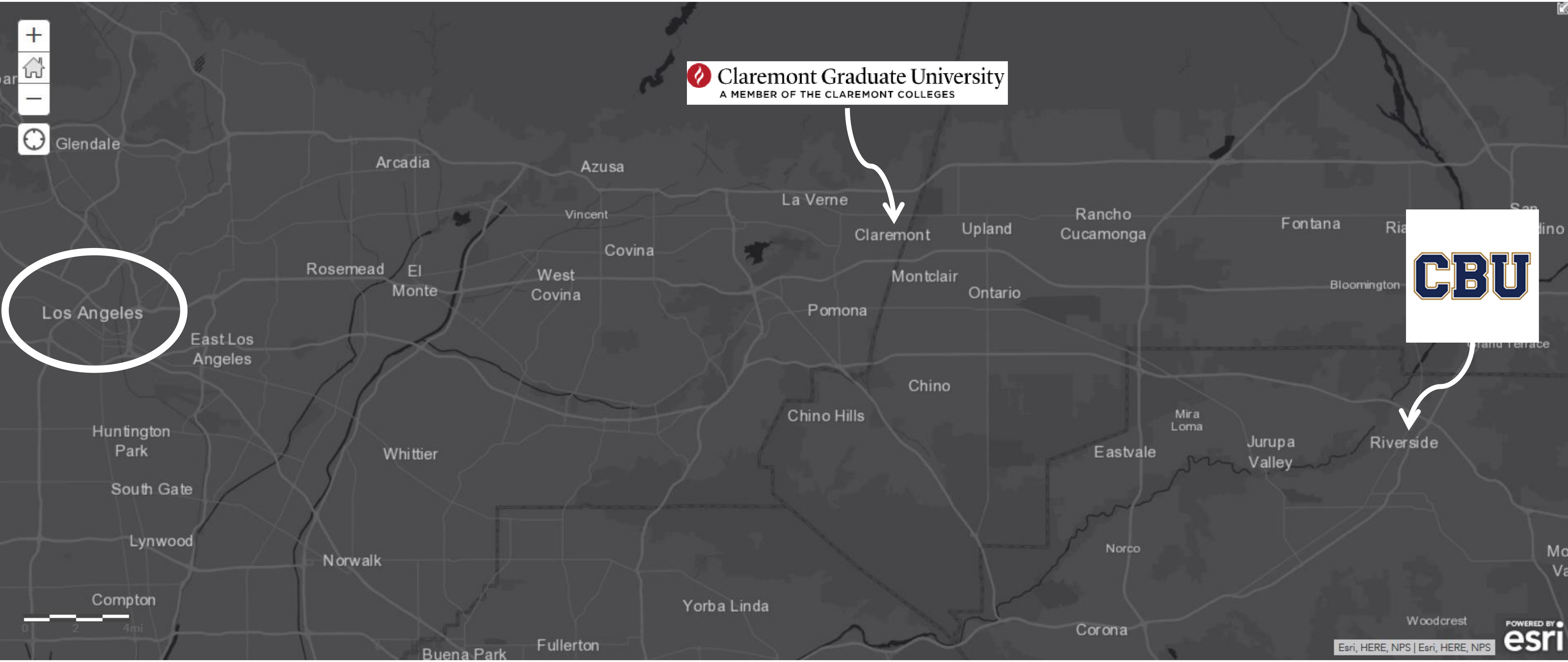
AMCIS 2020

11 August 2020

Social Media Analytics and GeoAI

Anthony Corso, Ph.D.
Associate Professor of Computing, Software and Data Sciences
California Baptist University
Riverside, CA

Brian Hilton, Ph.D.
Clinical Full Professor
Claremont Graduate University
Director, Advanced GIS Lab
Claremont, CA


Claremont Graduate University
A MEMBER OF THE CLAREMONT COLLEGES
CBU
Glendale
Arcadia
Azusa
La Verne
Claremont
Upland
Rancho Cucamonga
Fontana
Vincent
Covina
Rosemead
El Monte
West Covina
Montclair
Ontario
Bloomington
Pomona
Los Angeles
East Los Angeles
Chino
Chino Hills
Mira Loma
Jurupa Valley
Riverside
Eastvale
Huntington Park
Whittier
South Gate
Lynwood
Norwalk
Norco
Compton
Yorba Linda
Corona
Woodcrest
Fullerton
Buena Park
Esri, HERE, NPS | Esri, HERE, NPS
POWERED BY esri

Social Media Analytics and GeoAI

- Today, Social Media such as Twitter, Reddit, and Facebook, have become de facto global communication channels to disseminate news, entertainment, and one’s self-revelations.

- This session will demonstrate Social Media preprocessing techniques, the use of Natural Language Processing to augment the data, and geospatial analysis of this data using GeoAI.

Social Media Analytics

- And now, Anthony…

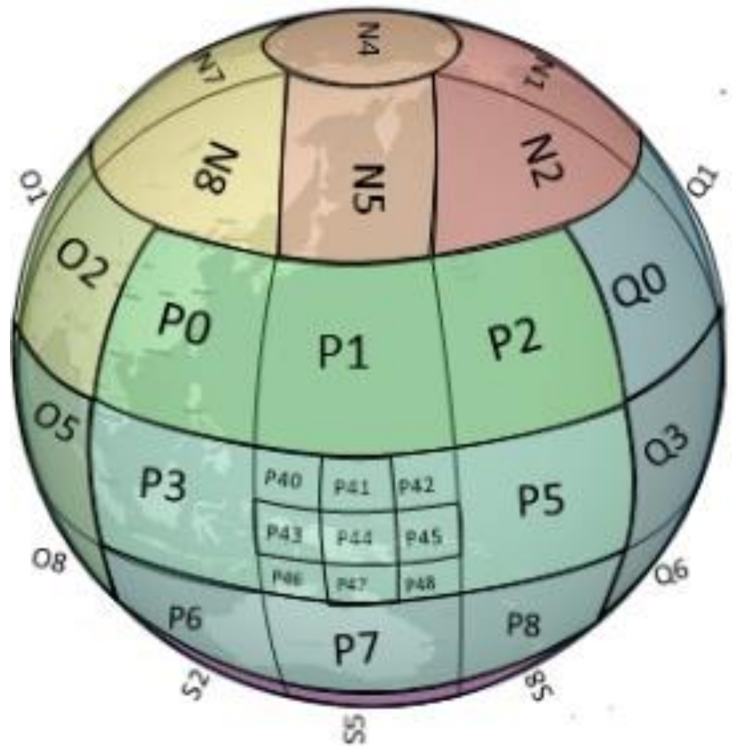
GeoAI

- Brief Discussion
  - Discrete Global Grid Systems
  - Types of Geospatial Data Analytics
  - Types of GeoAI
- Two Examples:
  - “Real-time“, descriptive / diagnostic, spatial-temporal analysis of Tweets
  - Historic, predictive, spatial-temporal analysis of Tweets

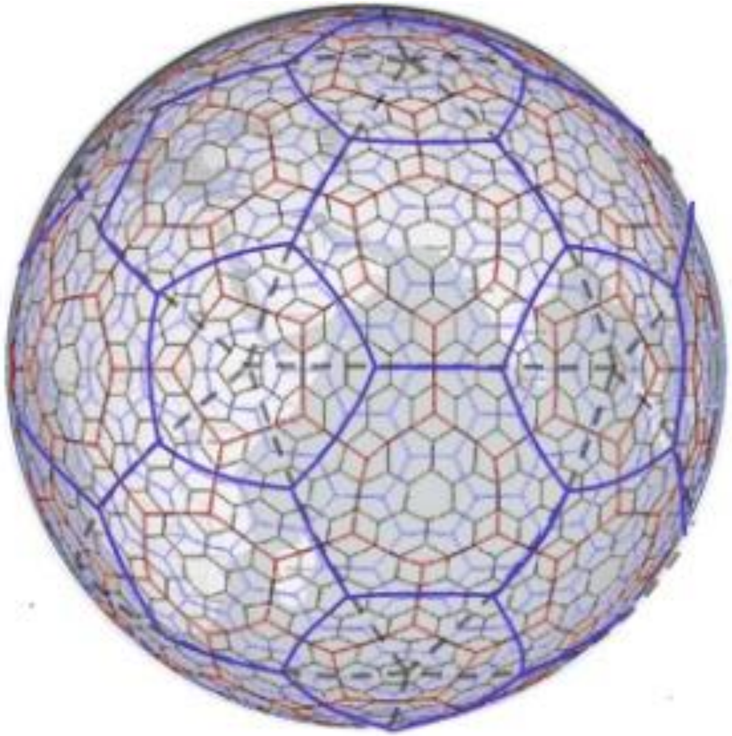
Discrete Global Grid Systems

- **What is a Discrete Global Grid (DGG)?**
- A **Discrete Global Grid** (**DGG**) consists of a set of regions that form a partition of the Earth's surface, where each region has a single point contained in the region associated with it. Each region/point combination is a called a *cell*. Depending on the application, data objects or values may be associated with the regions, points, or cells of a **DGG**. A **Discrete Global Grid System** (**DGGS**) is a series of discrete global grids, usually consisting of increasingly finer resolution grids (though the term **DGG** is often used interchangeably with the term **DGGS**).

Discrete Global Grid Systems



Quadrilateral



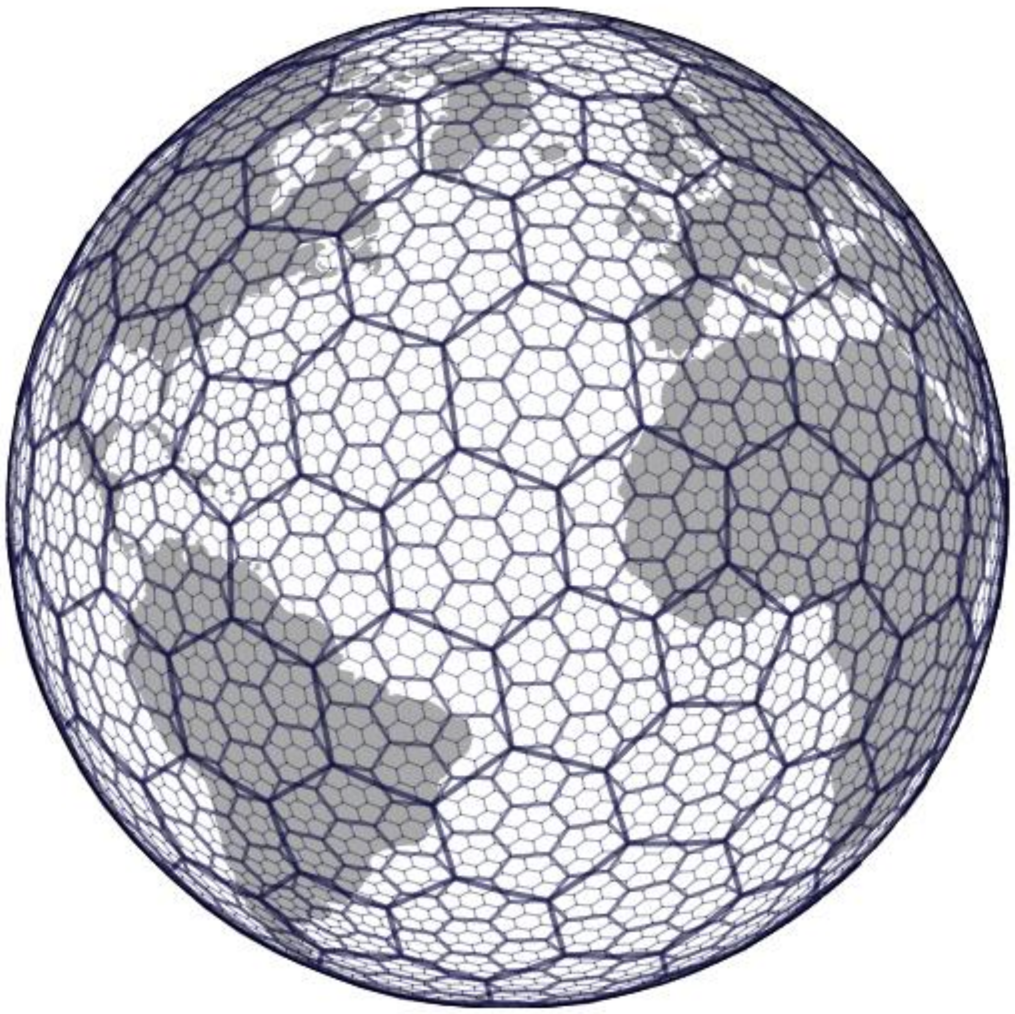
Hexagonal

Triangular

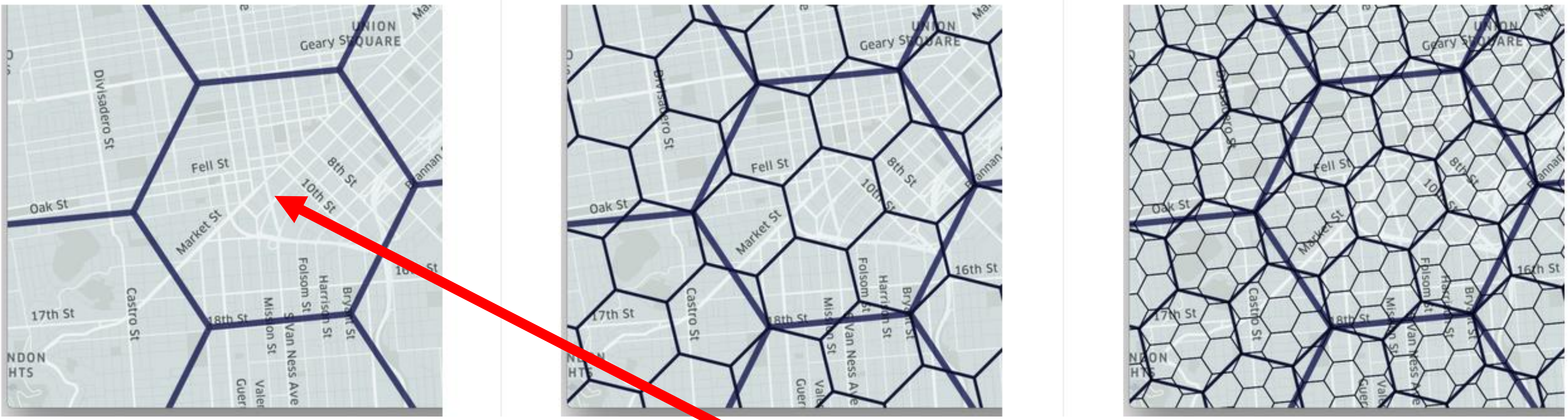
Discrete Global Grid Systems

- DGGS Resources
    - Southern Terra Cognita Laboratory
    - OGC Specification
    - Uber H3

Discrete Global Grid Systems - H3

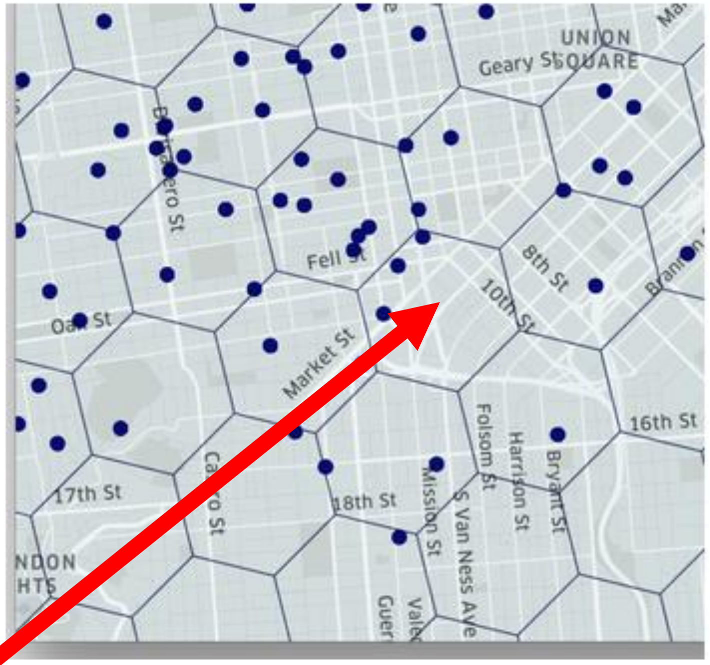


Discrete Global Grid Systems - H3



Each hexagon has a unique index value at a specific resolution
At this location, at resolution 8, the hexID = 8829a1d719fffff

Discrete Global Grid Systems - H3



At this location, at resolution 9, the hexID = 8929a1d7193ffff

The three points here could be "tagged" with this value

Discrete Global Grid Systems - H3

H3 Resolution	Average Hexagon Area (km^2)	Average Hexagon Edge Length (km)	Number of unique indexes
0	4,250,546.8477000	1,107.712591000	122
1	607,220.9782429	418.676005500	842
2	86,745.8540347	158.244655800	5,882
3	12,392.2648621	59.810857940	41,162
4	1,770.3235517	22.606379400	288,122
5	252.9033645	8.544408276	2,016,842
6	36.1290521	3.229482772	14,117,882
7	5.1612932	1.220629759	98,825,162
8	0.7373276	0.461354684	691,776,122
9	0.1053325	0.174375668	4,842,432,842
10	0.0150475	0.065907807	33,897,029,882
11	0.0021496	0.024910561	237,279,209,162
12	0.0003071	0.009415526	1,660,954,464,122
13	0.0000439	0.003559893	11,626,681,248,842
14	0.0000063	0.001348575	81,386,768,741,882
15	0.0000009	0.000509713	569,707,381,193,162

Discrete Global Grid Systems - H3

- What do those resolutions mean?

- For example:
  - Resolution 7: City District
  - Resolution 8: City Neighborhood
  - Resolution 9: 4-8 city blocks
  - Resolution 10: A city block or less

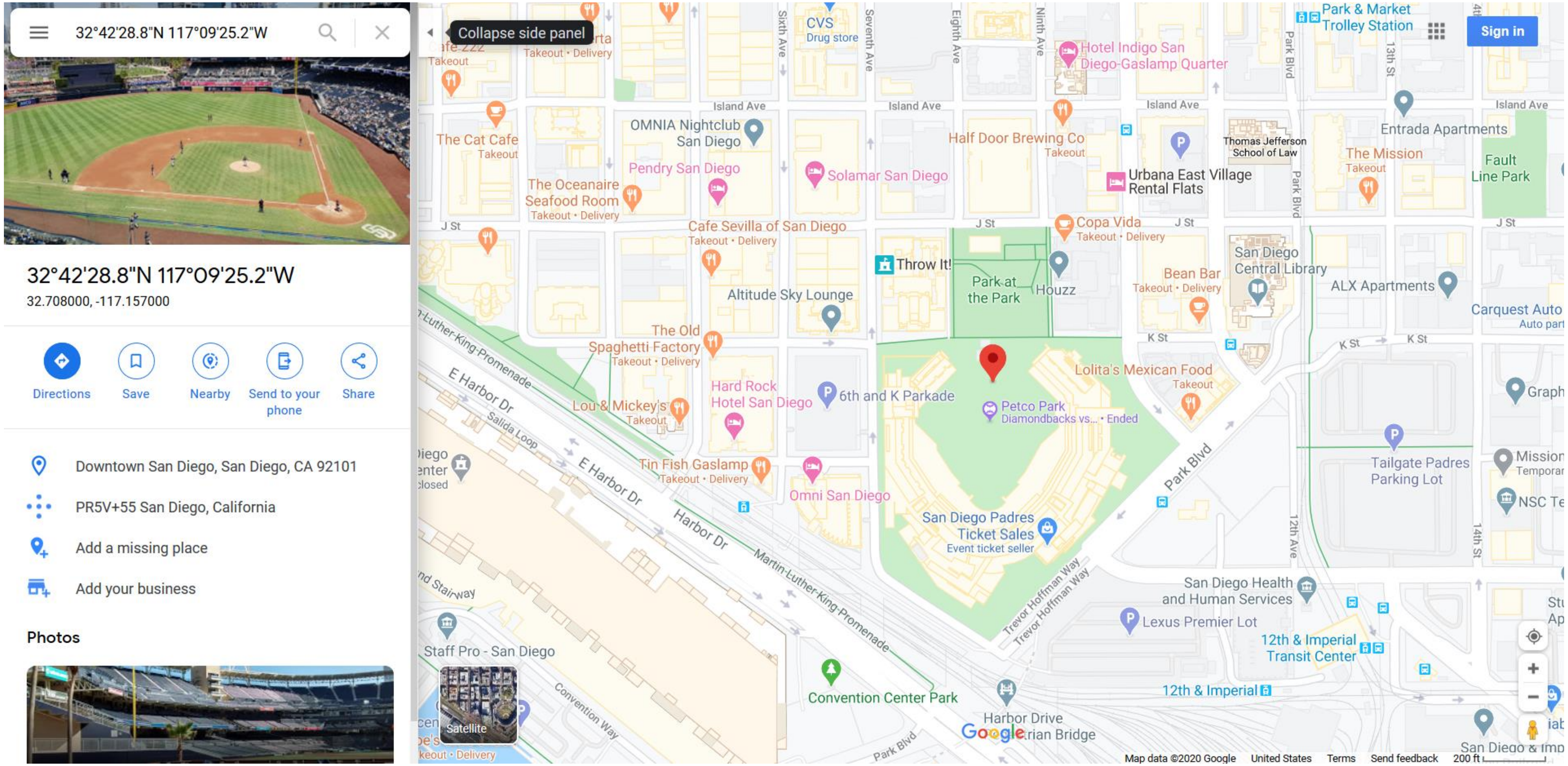
  .

  .
  - Resolution 15: Less than one square meter

Discrete Global Grid Systems - H3 - San Diego

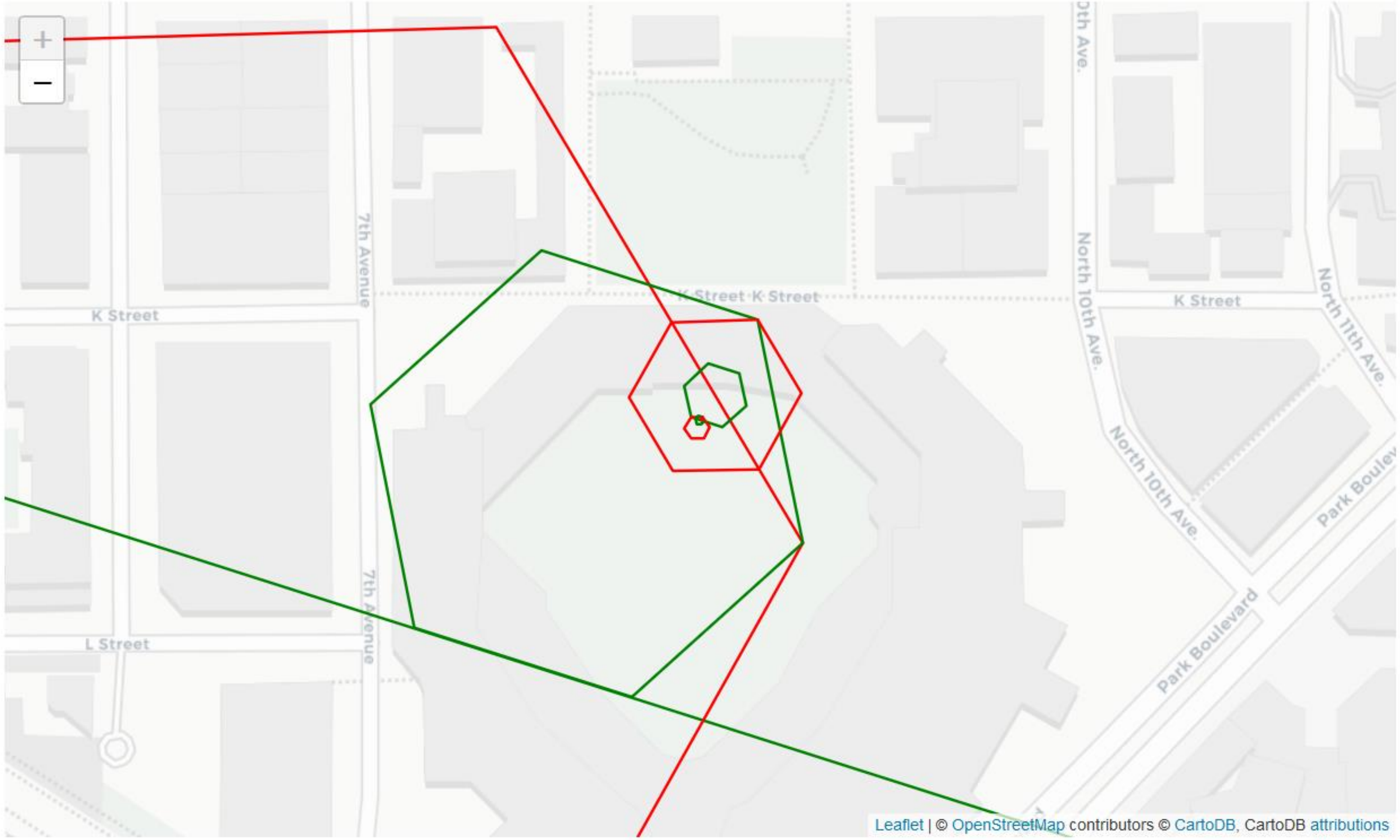
- H3 - San Diego H3 resolution example - Python notebook
- Link

Discrete Global Grid Systems - H3 - San Diego



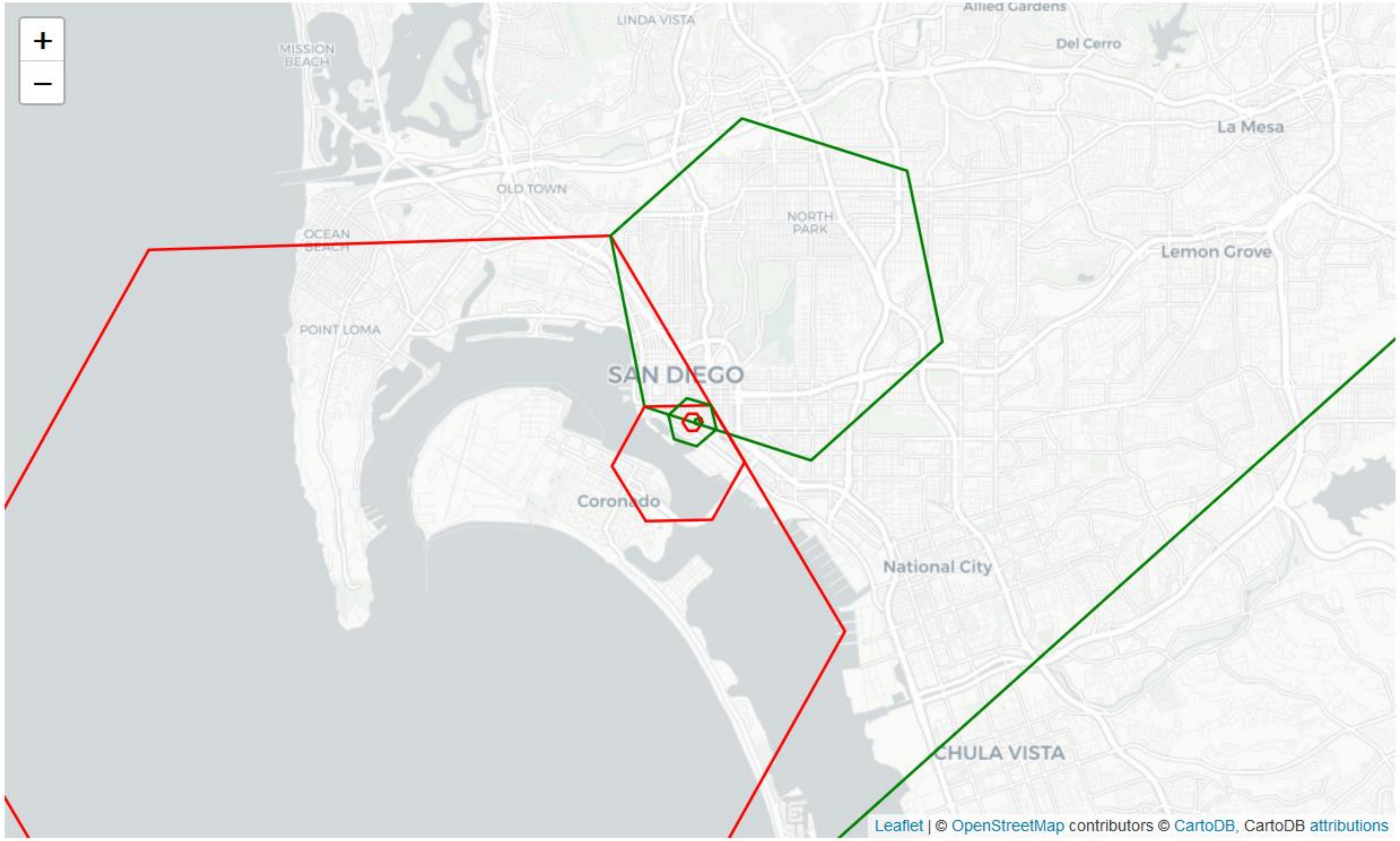
Example location - Petco Park (San Diego, CA) Google Maps (longitude = -117.157, latitude = 32.708)

Discrete Global Grid Systems - H3 - San Diego



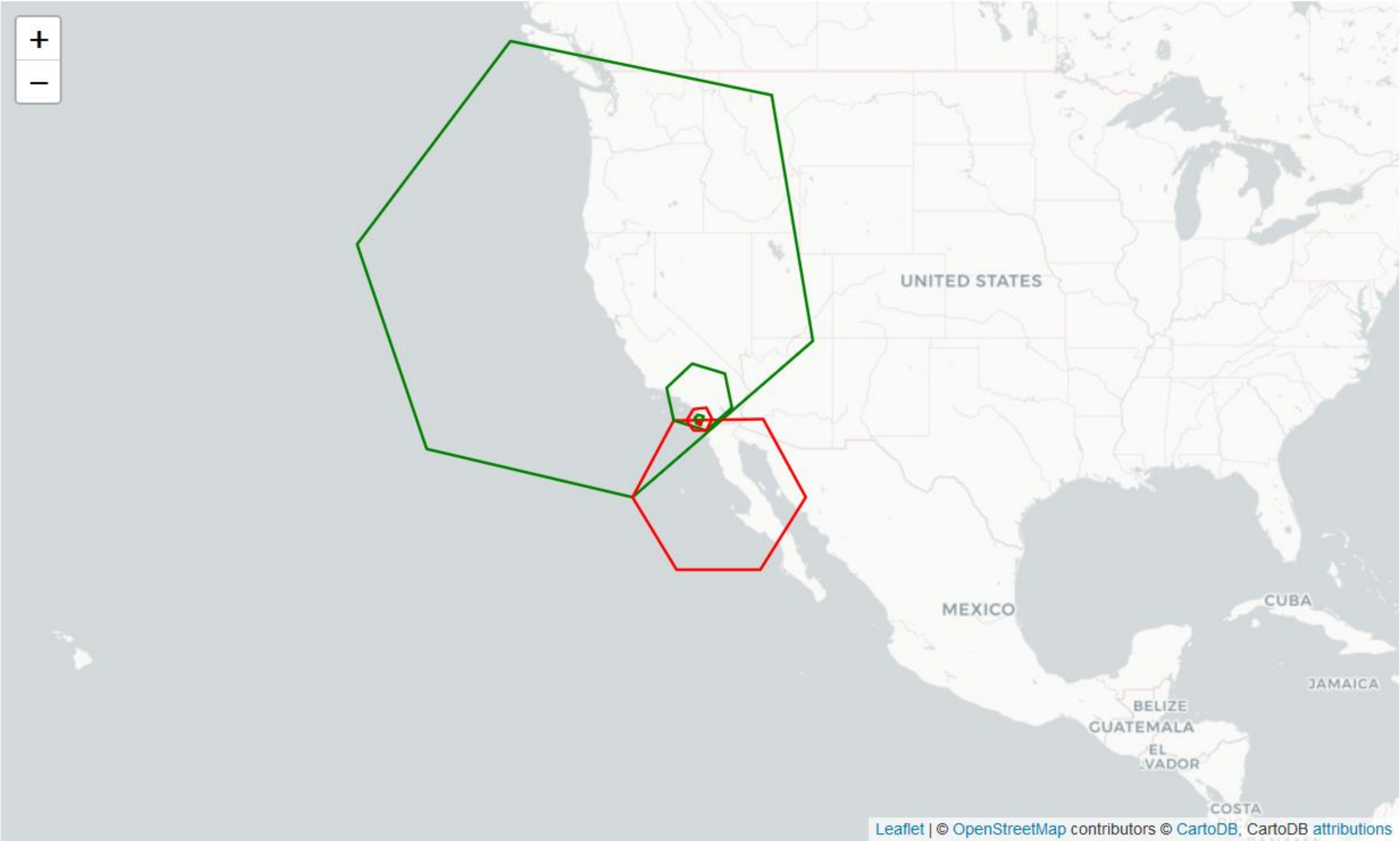
San Diego H3 resolution example

Discrete Global Grid Systems - H3 - San Diego



San Diego H3 resolution example

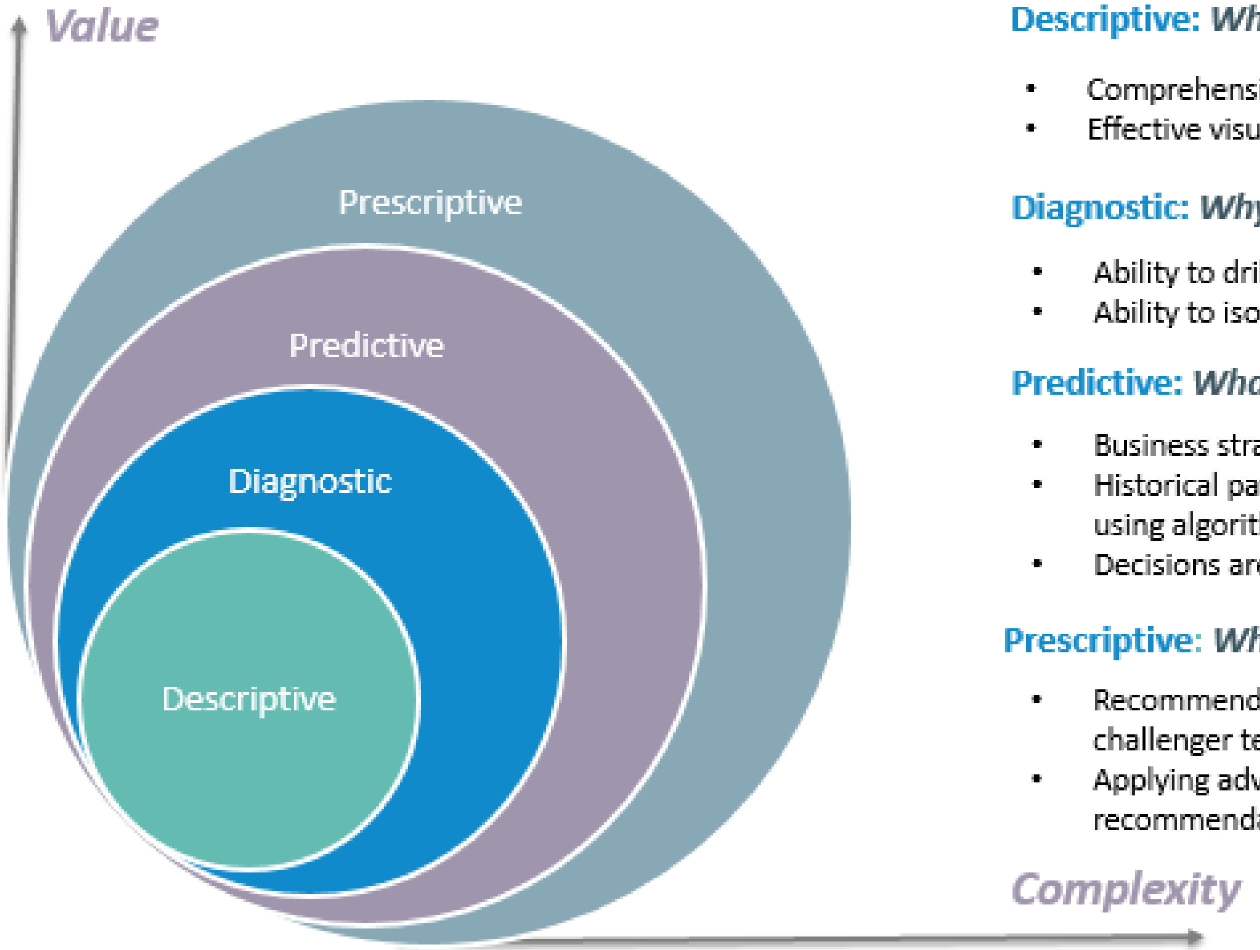
Discrete Global Grid Systems - H3 - San Diego



San Diego H3 resolution example

Types of Geospatial Data Analytics

4 types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

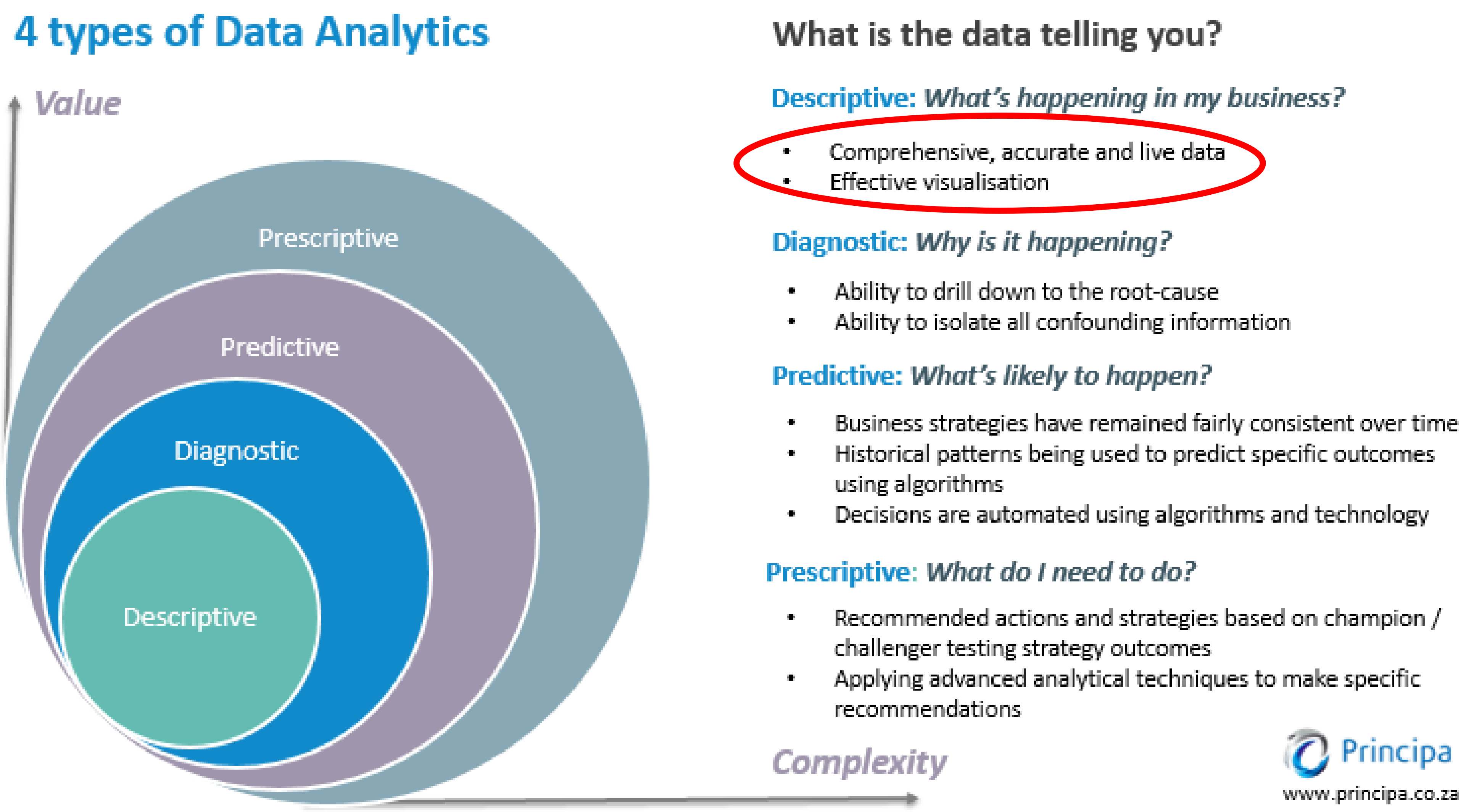
- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

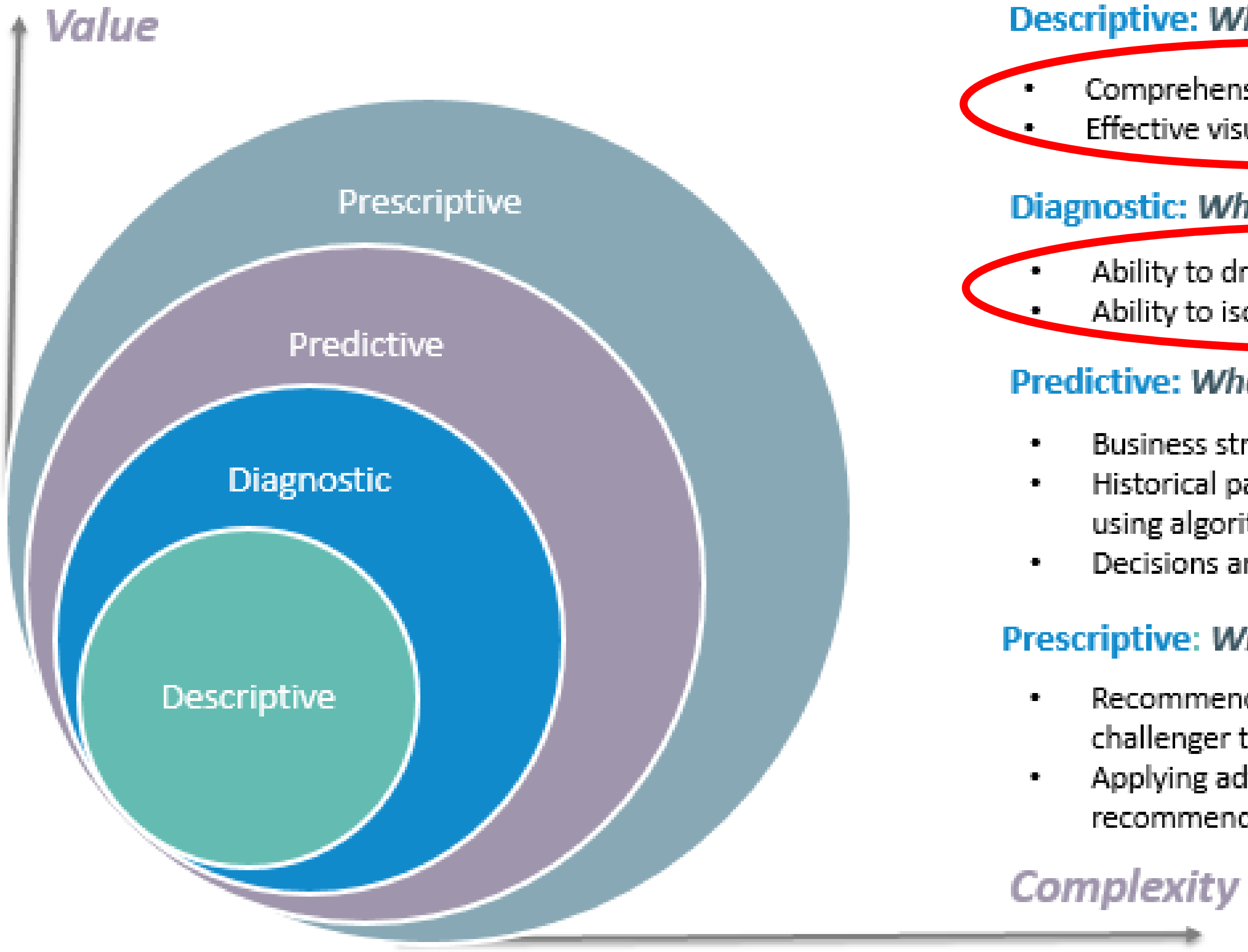
Principa
www.principa.co.za

Types of Geospatial Data Analytics



Types of Geospatial Data Analytics

4 types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

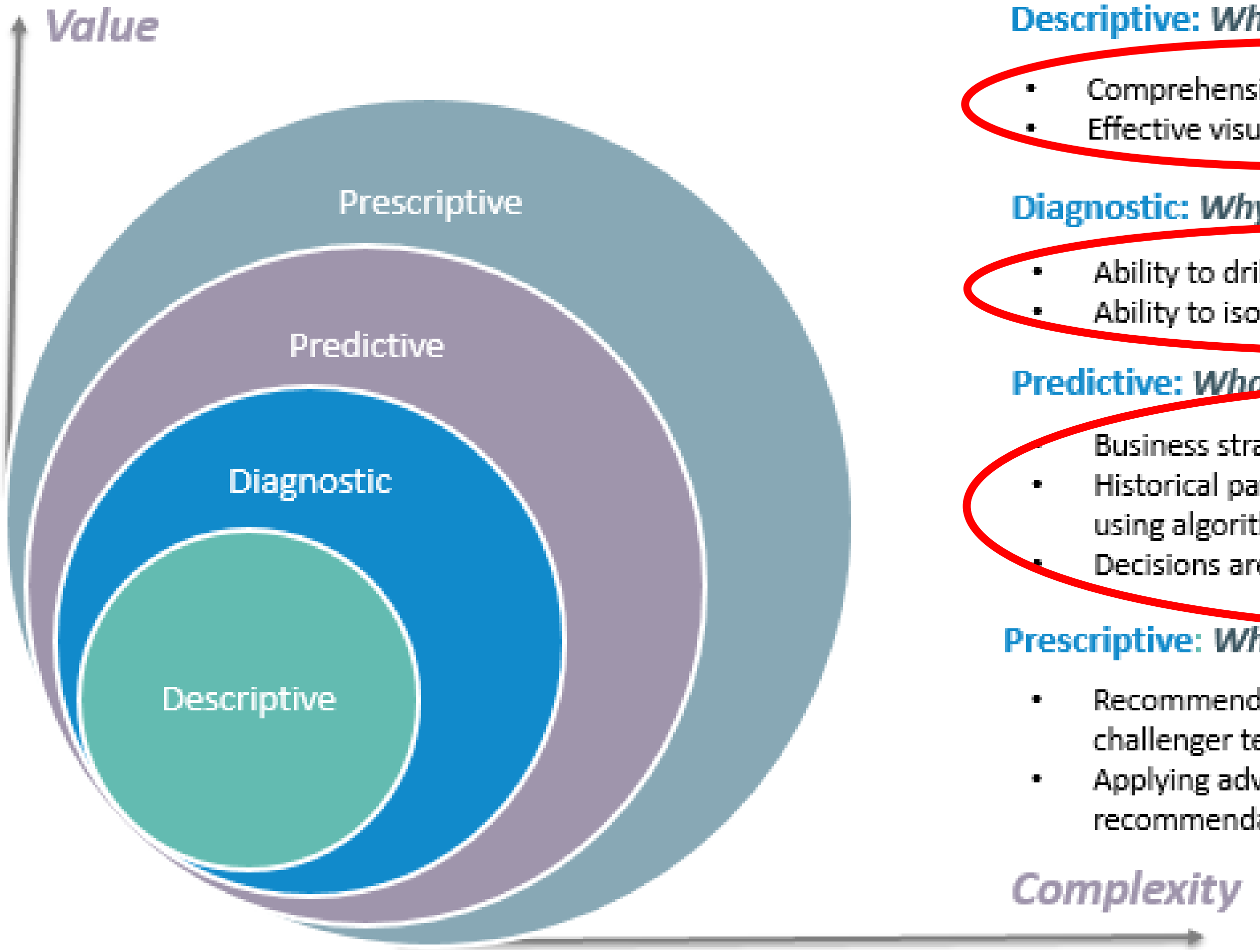
Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

Types of Geospatial Data Analytics

4 types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

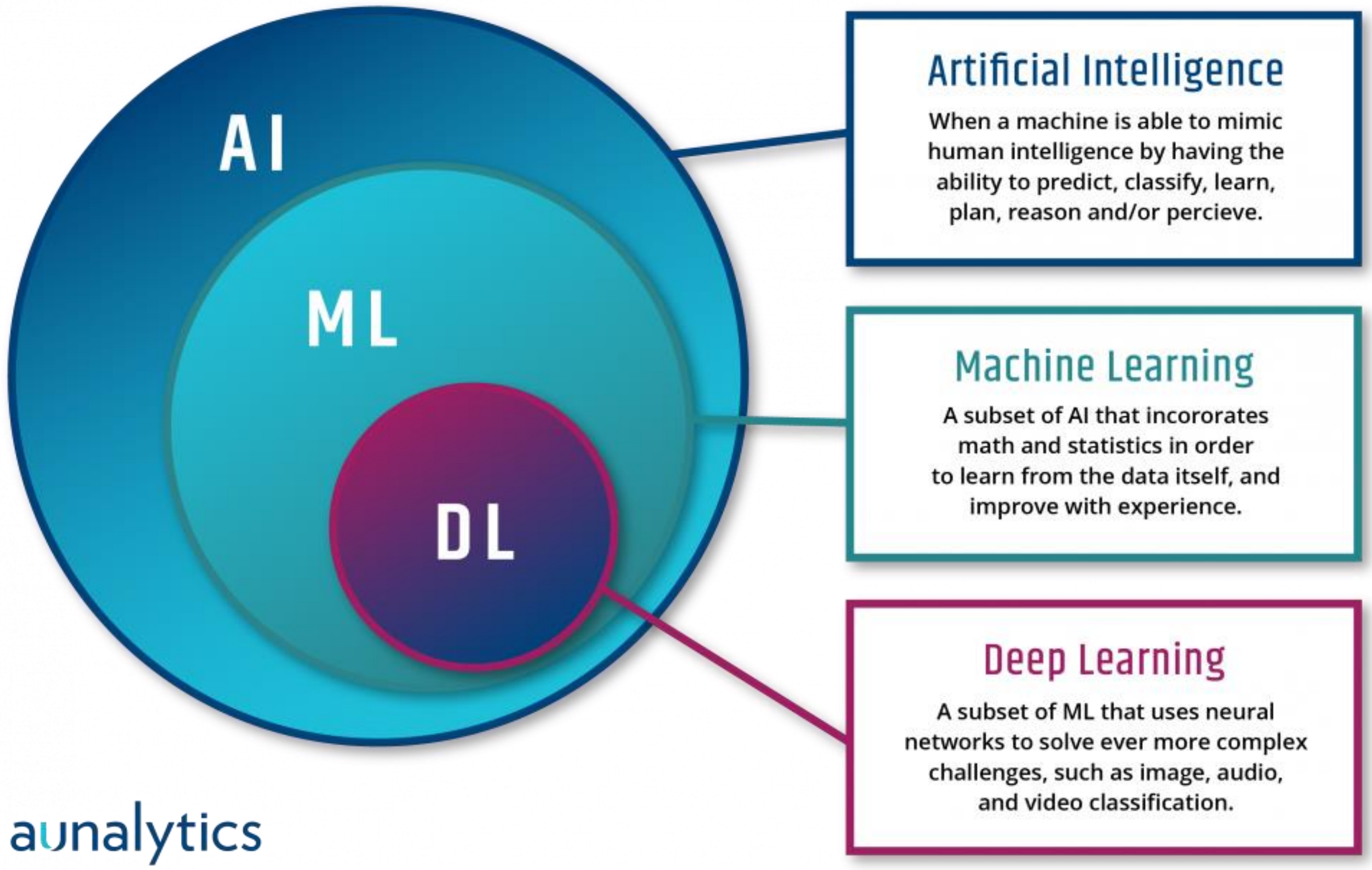
- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

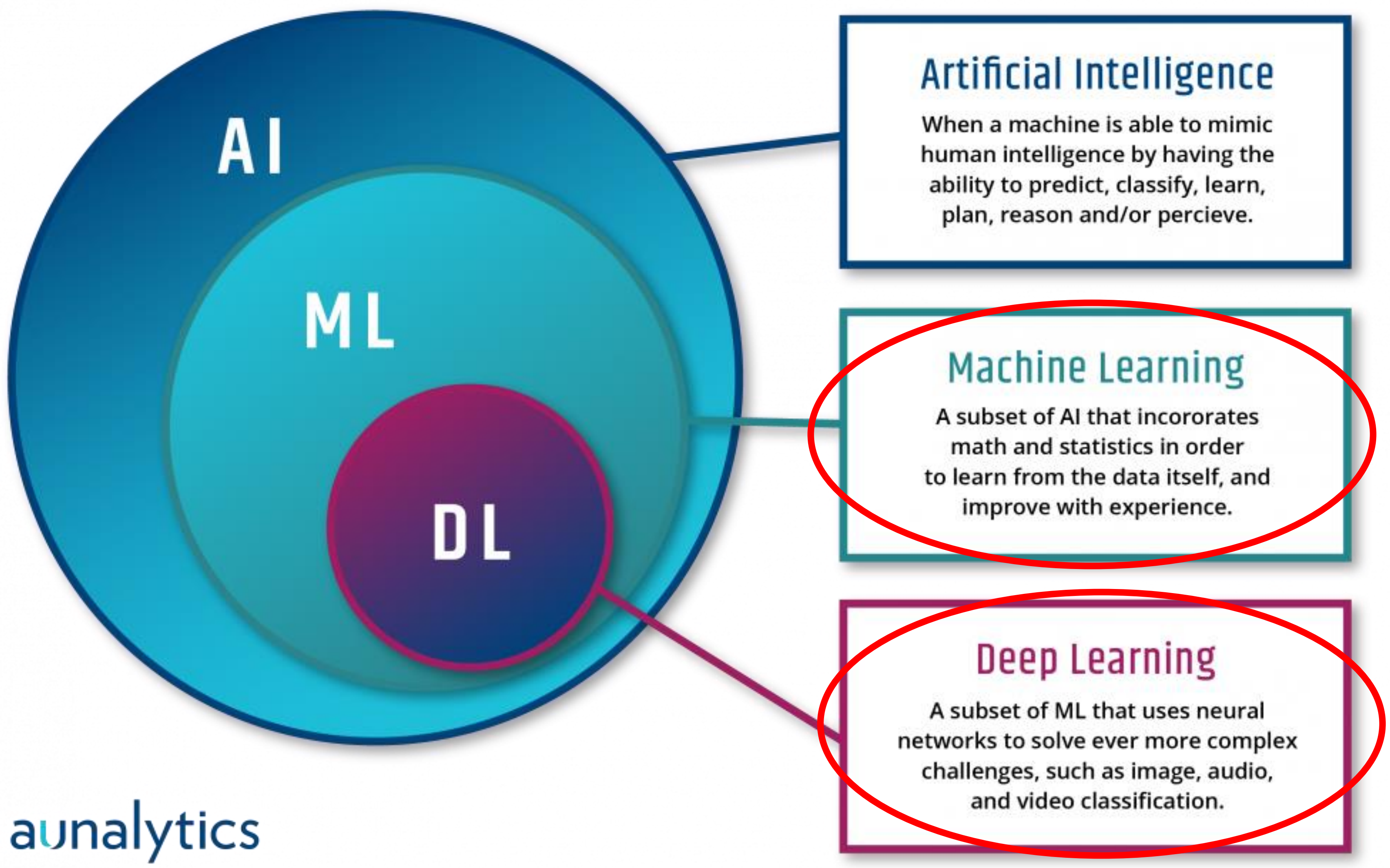
- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

Types of GeoAI



Types of GeoAI



GeoAI - Machine Learning

GeoAI - Machine Learning

Classification

The process of deciding to which category an object should be assigned based on a training dataset

Use Case: Classify impervious surfaces to help effectively prepare for storm and flood events based on the latest high-resolution imagery

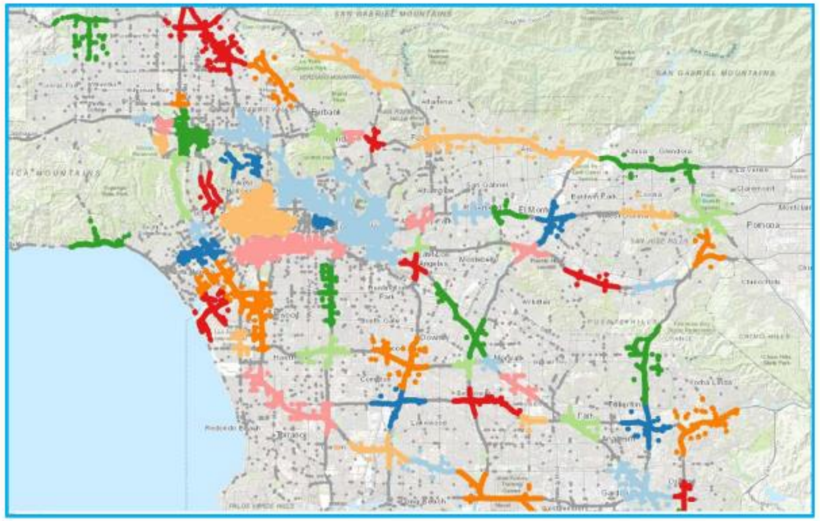
In ArcGIS: Maximum Likelihood Classification, Random Trees, Support Vector Machine , Forest-based Classification and Regression

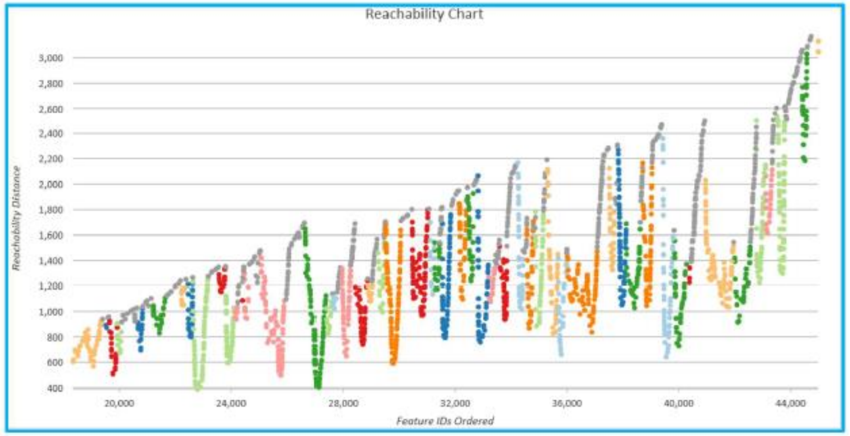
GeoAI - Machine Learning

Clustering

The grouping of observations based on similarities of values or locations

Use Case: Given the nearly 50,000 reports of traffic between 5pm and 6pm in Los Angeles (from Traffic Alerts by Waze), where are traffic zones that can be used to elicit feedback from current drivers in the area?





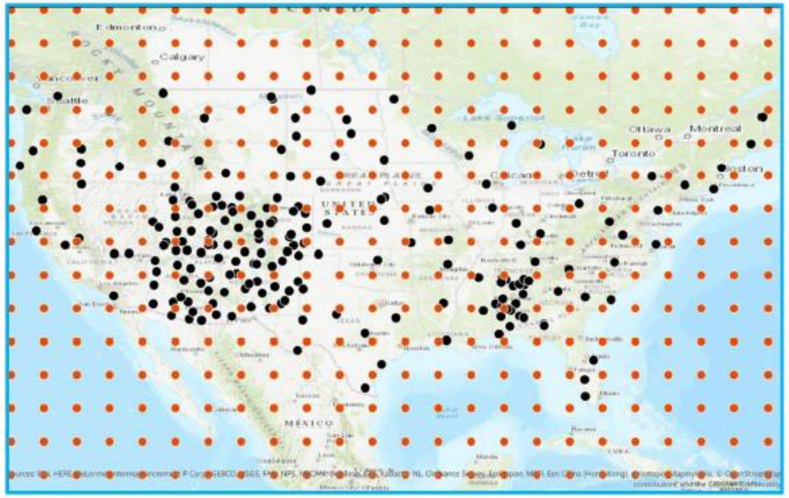
In ArcGIS: Spatially Constrained Multivariate Clustering, Multivariate Clustering, Density-based Clustering, Image Segmentation, Hot Spot Analysis, Cluster and Outlier Analysis, Space Time Pattern Mining

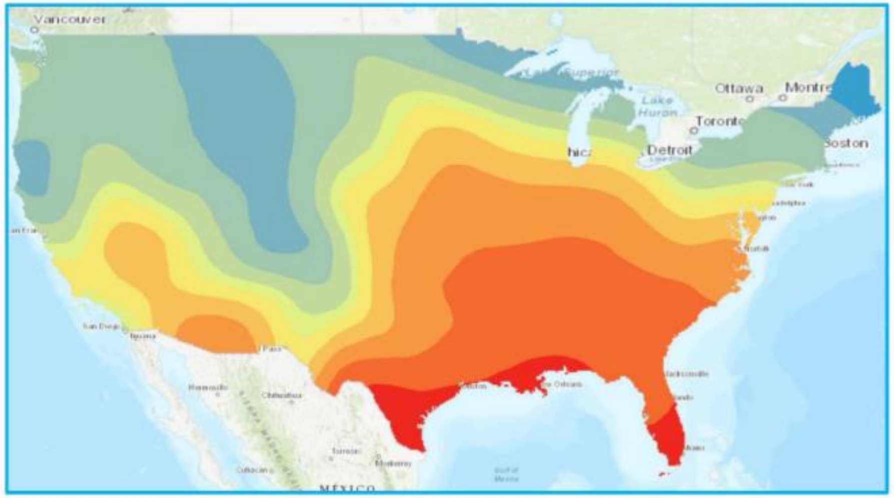
GeoAI - Machine Learning

Prediction

Using the known to estimate the unknown

Use Case: Accurately predict impacts of climate change on local temperature using global climate model data

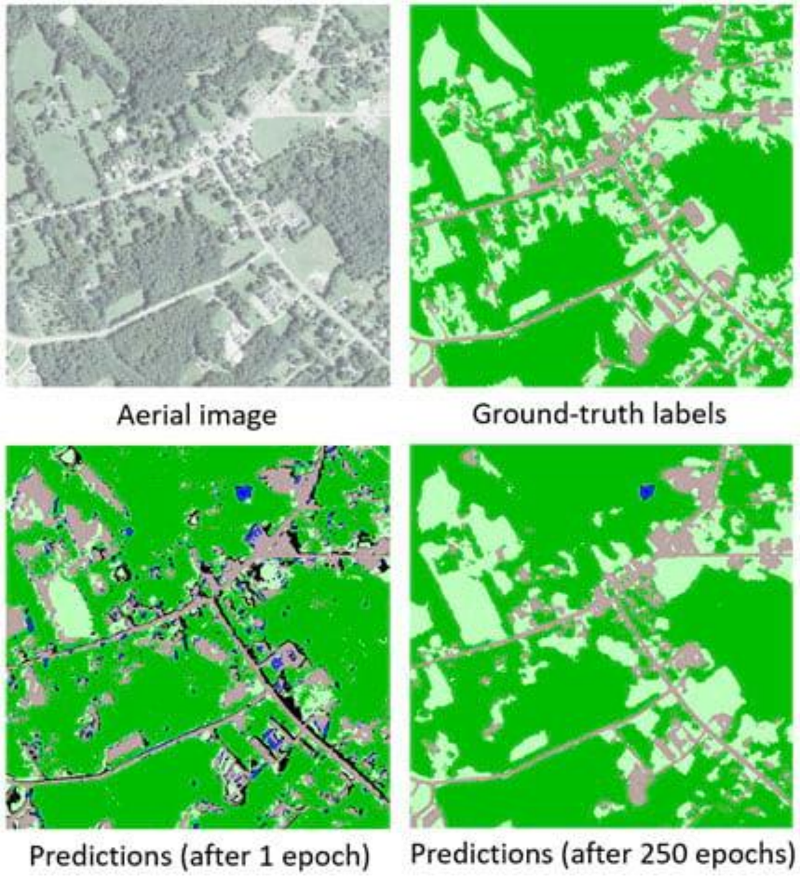




In ArcGIS: Empirical Bayesian Kriging, Areal Interpolation, EBK Regression Prediction, Ordinary Least Squares Regression and Exploratory Regression, Geographically Weighted Regression, Generalized Linear Regression, Forest-based Classification and Regression

GeoAI - Deep Learning

GeoAI - Deep Learning



Object Detection - Swimming Pools

Classification - Land Cover Type

Types of GeoAI

- GeoAI Resources
  - Medium website: GeoAI - thoughts about where AI and GIS intersect
  - Spatial Analysis and Data Science at the 2020 Esri User Conference
  - GeoAI: Vertical Use Cases using AI with ArcGIS
  - Spatial Analysis and Data Science
  - Geographic Data Science Lab
  - Geographic Information Systems and Science
  - Geographic Data Science with PySAL and the PyData Stack
  - Geocomputation with R

“Real-time“, descriptive / diagnostic, spatial-temporal analysis of Tweets

- Study Area - San Diego, CA
- Spatial Resolution - H3 resolution 7, 8, and 9
- Time Period - late December 2019 (hence, “real-time” in quotes)
- Data Sets
  - Twitter
  - San Diego Calls for Service (public safety data)

“Real-time“, descriptive / diagnostic, spatial-temporal analysis of Tweets

- Workflow (in brief)
  - Tag data (Tweets and Calls for Service) with H3 index values
  - Link Tweets and Calls for Service using H3 index

- Purpose
  - Proof-of-concept linking live data
  - Visualize data using various techniques
  - Examine data in an exploratory / drill-down approach

“Real-time“, descriptive / diagnostic, spatial-temporal analysis of Tweets

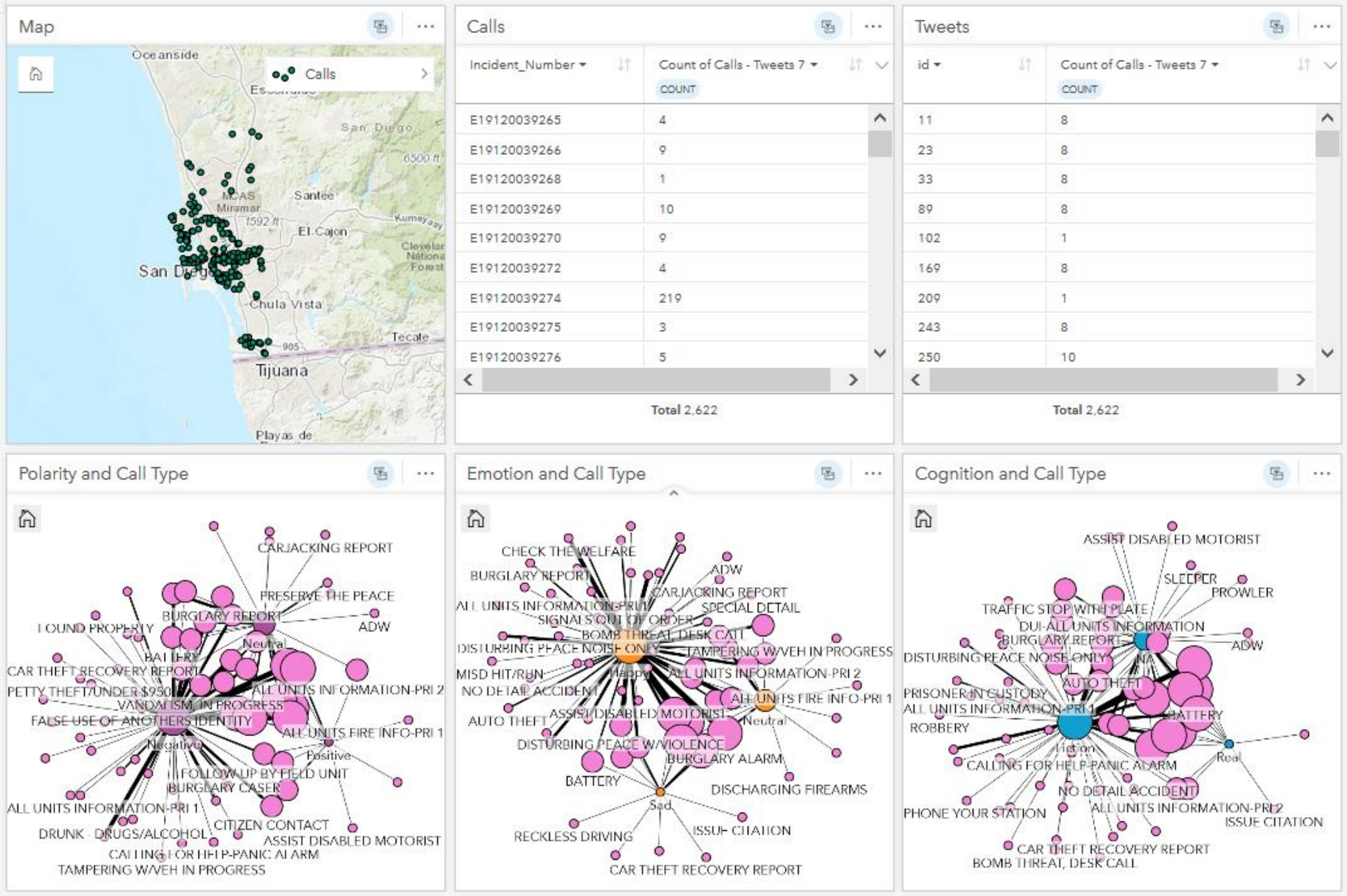
Tweets

created_a	Polarity	Emotion	Cognition	hex_0	hex_1	hex_2	hex_3	hex_4	hex_5	hex_6	hex_7	hex_8	hex_9
Tue Dec 2	Neutral	Neutral	NA	8029ffffff	81487fffff	8229a7fff	8329a4fff	8429a41ff	8529a413	8629a410	8729a411	8829a411	8929a411
Tue Dec 2	Positive	Happy	NA	8029ffffff	81487fffff	8229a7fff	8329a4fff	8429a41ff	8529a413	8629a410	8729a411	8829a411	8929a411
Tue Dec 2	Positive	Happy	Real	8029ffffff	81487fffff	8229a7fff	8329a4fff	8429a41ff	8529a413	8629a402	8729a402	8829a402	8929a402
Tue Dec 2	Positive	Sad	NA	8029ffffff	81487fffff	8229a7fff	8329a4fff	8429a41ff	8529a413	8629a402	8729a402	8829a402	8929a402
Tue Dec 2	Negative	Neutral	Fiction	8029ffffff	81487fffff	8229a7fff	8329a4fff	8429a41ff	8529a403	8629a402	8729a402	8829a402	8929a402
Tue Dec 2	Negative	Sad	NA	8029ffffff	81487fffff	8229a7fff	8329a4fff	8429a41ff	8529a413	8629a411	8729a411	8829a411	8929a411
Tue Dec 2	Neutral	Happy	Fiction	8029ffffff	81487fffff	8229a7fff	8329a4fff	8429a41ff	8529a413	8629a411	8729a411	8829a411	8929a411
Tue Dec 2	Positive	Sad	NA	8029ffffff	81487fffff	8229a7fff	8329a4fff	8429a41ff	8529a413	8629a411	8729a411	8829a411	8929a411

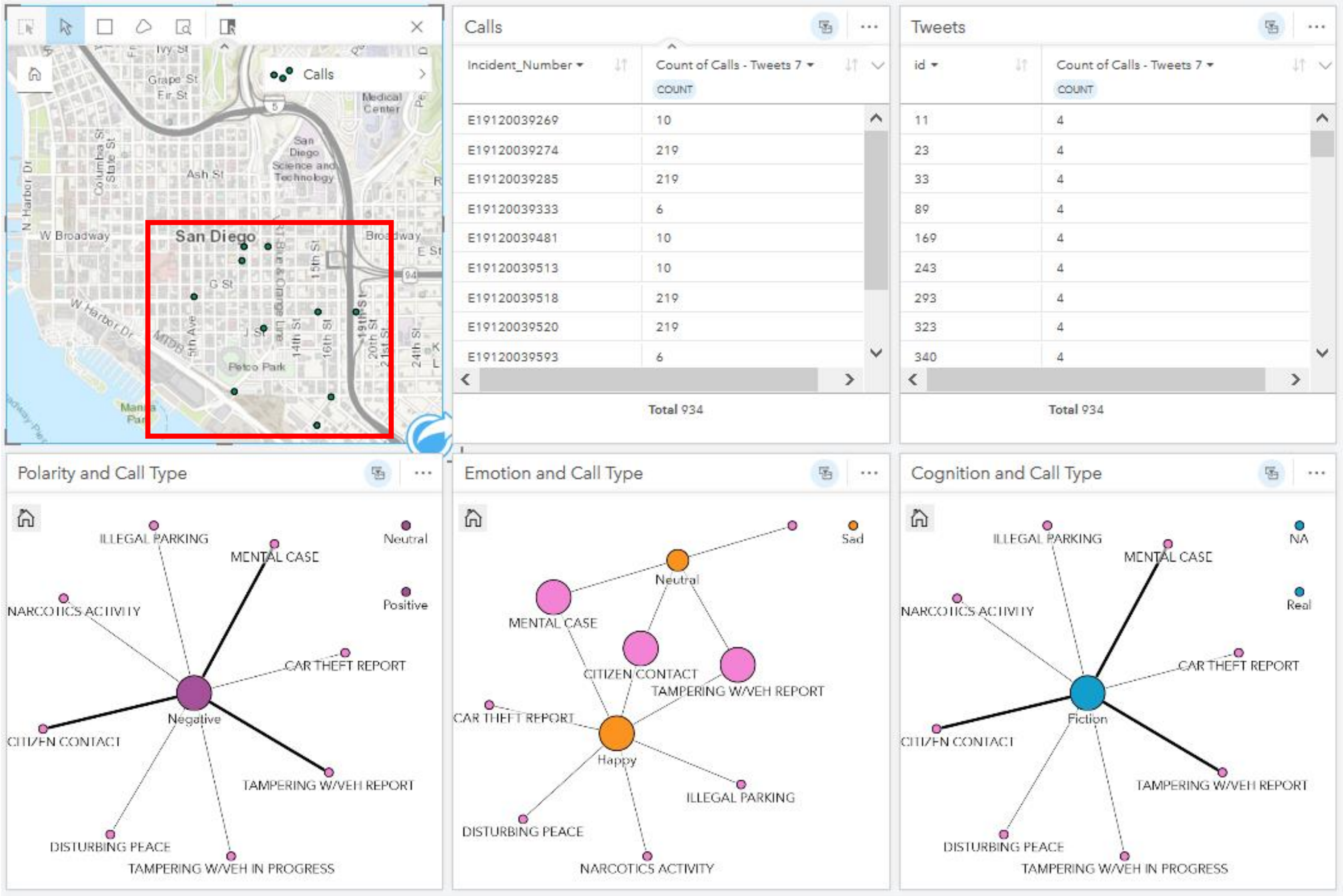
Calls for Service

call_type_	description	priority_1	priority_text	dispo_cod	description_1	date_time_Converted	lat	lon	hex_10	hex_9	hex_8	hex_7
1151	PED STOP/FIELD IN	2	Dispatch as quickly as possibl	CAN	CANCEL	12/23/2019 0:04	32.74872	-117.106	8a29a41a	8929a41a	8829a41a	8729a41aa
1151	PED STOP/FIELD IN	2	Dispatch as quickly as possibl	K	NO REPORT REQUIRED	12/23/2019 0:10	32.73822	-117.11	8a29a41a	8929a41a	8829a41a	8729a41aa
459A	BURGLARY ALARM	2	Dispatch as quickly as possibl	K	NO REPORT REQUIRED	12/23/2019 0:12	32.74855	-117.054	8a29a418	8929a418	8829a418	8729a4180
SLEEPER	SLEEPER	3	Dispatch as quickly as possibl	U	UNFOUNDED	12/23/2019 0:13	32.75441	-117.248	8a29a402	8929a402	8829a402	8729a4026
MPSSTP	TRAFFIC STOP FRO	2	Dispatch as quickly as possibl	O	OTHER	12/23/2019 0:13	32.97904	-117.084	8a29a409	8929a409	8829a409	8729a408a
1151	PED STOP/FIELD IN	2	Dispatch as quickly as possibl	K	NO REPORT REQUIRED	12/23/2019 0:15	32.77491	-117.206	8a29a403	8929a403	8829a403	8729a4035
NARC	NARCOTICS ACTIV	2	Dispatch as quickly as possibl	K	NO REPORT REQUIRED	12/23/2019 0:15	32.75344	-117.248	8a29a402	8929a402	8829a402	8729a4026
1153	SECURITY CHECK	2	Dispatch as quickly as possibl	R	REPORT	12/23/2019 0:17	32.82044	-117.179	8a29a401	8929a401	8829a401	8729a4015
459	BURGLARY IN PRO	1	Dispatch Immediately - seriou	K	NO REPORT REQUIRED	12/23/2019 0:21	32.71397	-117.154	8a29a41a	8929a41a	8829a41a	8729a41ad

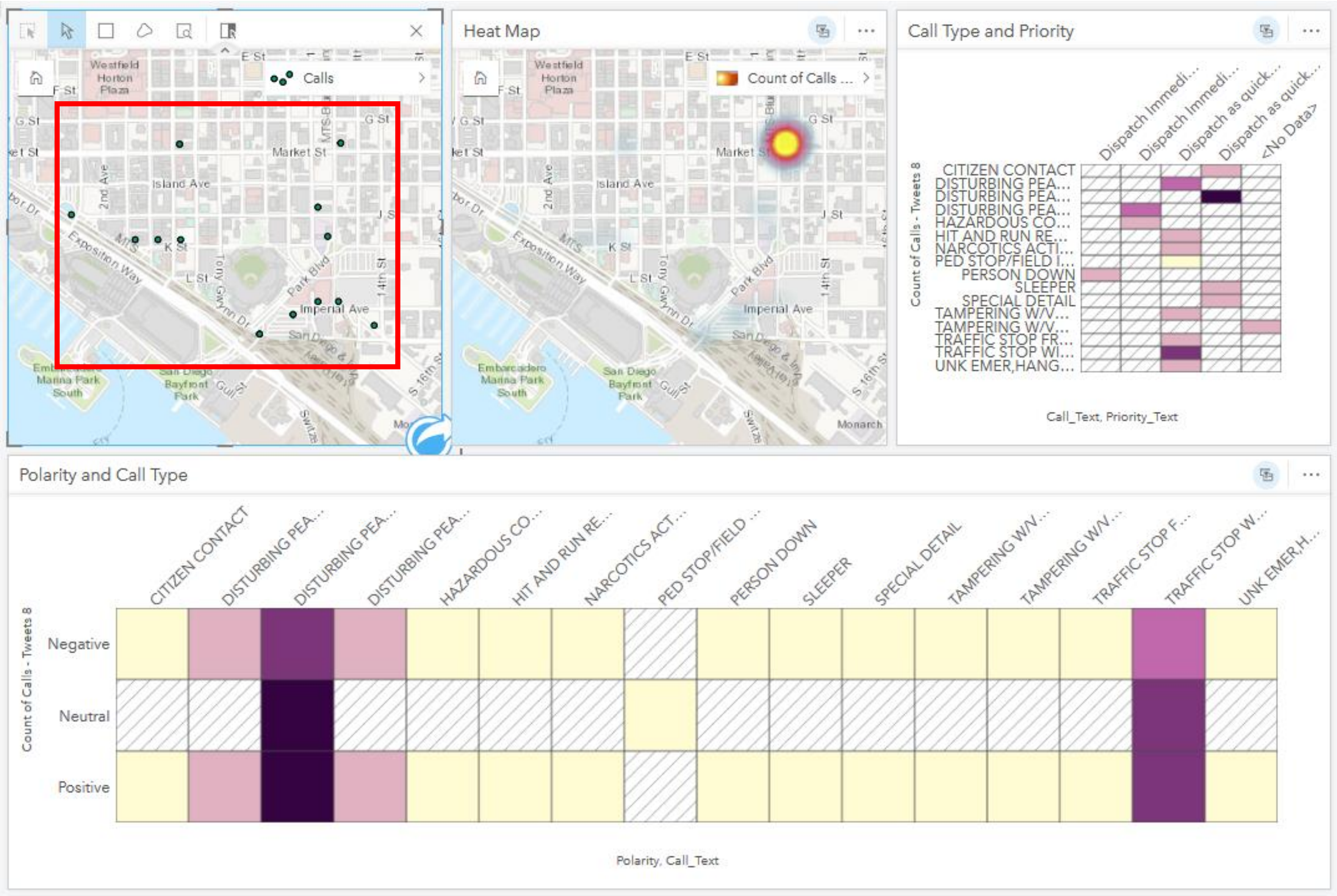
“Real-time“, descriptive / diagnostic, spatial-temporal analysis of Tweets



“Real-time“, descriptive / diagnostic, spatial-temporal analysis of Tweets



“Real-time“, descriptive / diagnostic, spatial-temporal analysis of Tweets



Historic, predictive, spatial-temporal analysis of Tweets

- Study Area - San Diego, CA
- Spatial Resolution - H3 resolution 9
- Time Period - late December 2019
- Data Sets
  - Twitter
  - CalEnviroScreen 3.0 (CES3) Indicators in CalEnviroScreen are measures of either **environmental conditions**, in the case of **pollution burden** indicators, or **health and vulnerability factors** for **population characteristics** indicators.

Historic, predictive, spatial-temporal analysis of Tweets

- Workflow (in brief)
  - Tag data (Tweets) with H3 index values
  - H3 - San Diego H3 hexagons example - Python notebook
  - ArcGIS - San Diego tabulate intersect example - Python notebook
  - Append Tweets data set with CES3 Indicators using H3 index
- Purpose
  - Examine relationships between “NLP-ed” Tweets and CES3 data
  - Predict Emotion (Happy, Neutral, Sad) based on CES3 Population Characteristics

Historic, predictive, spatial-temporal analysis of Tweets

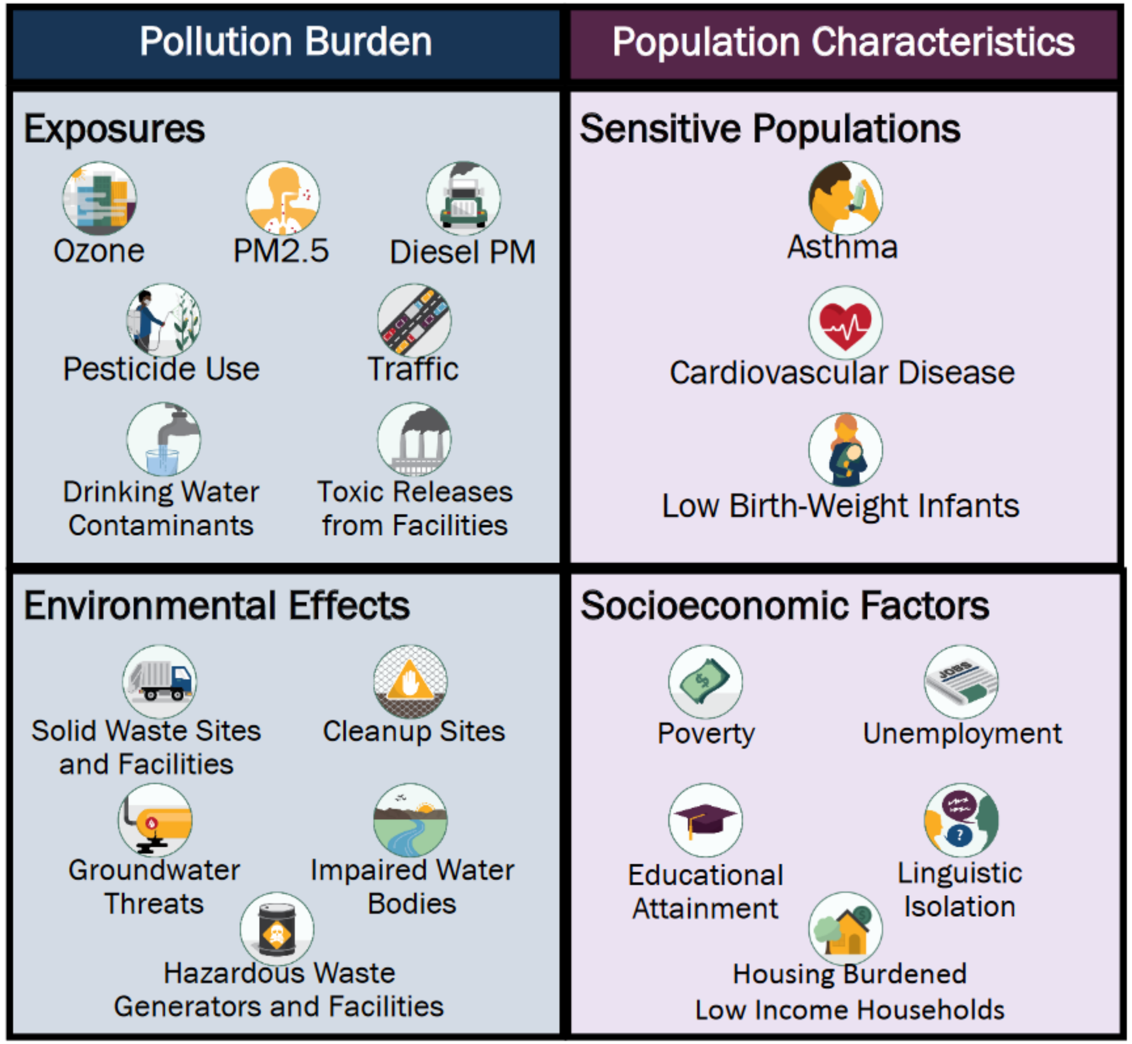
CES3 Indicators:

Pollution

- Exposures
- Environmental Effects

Pollution Characteristics

- Sensitive Populations
- Socioeconomic Factors



Historic, predictive, spatial-temporal analysis of Tweets

- H3 - San Diego H3 hexagons example - Python notebook
- Link

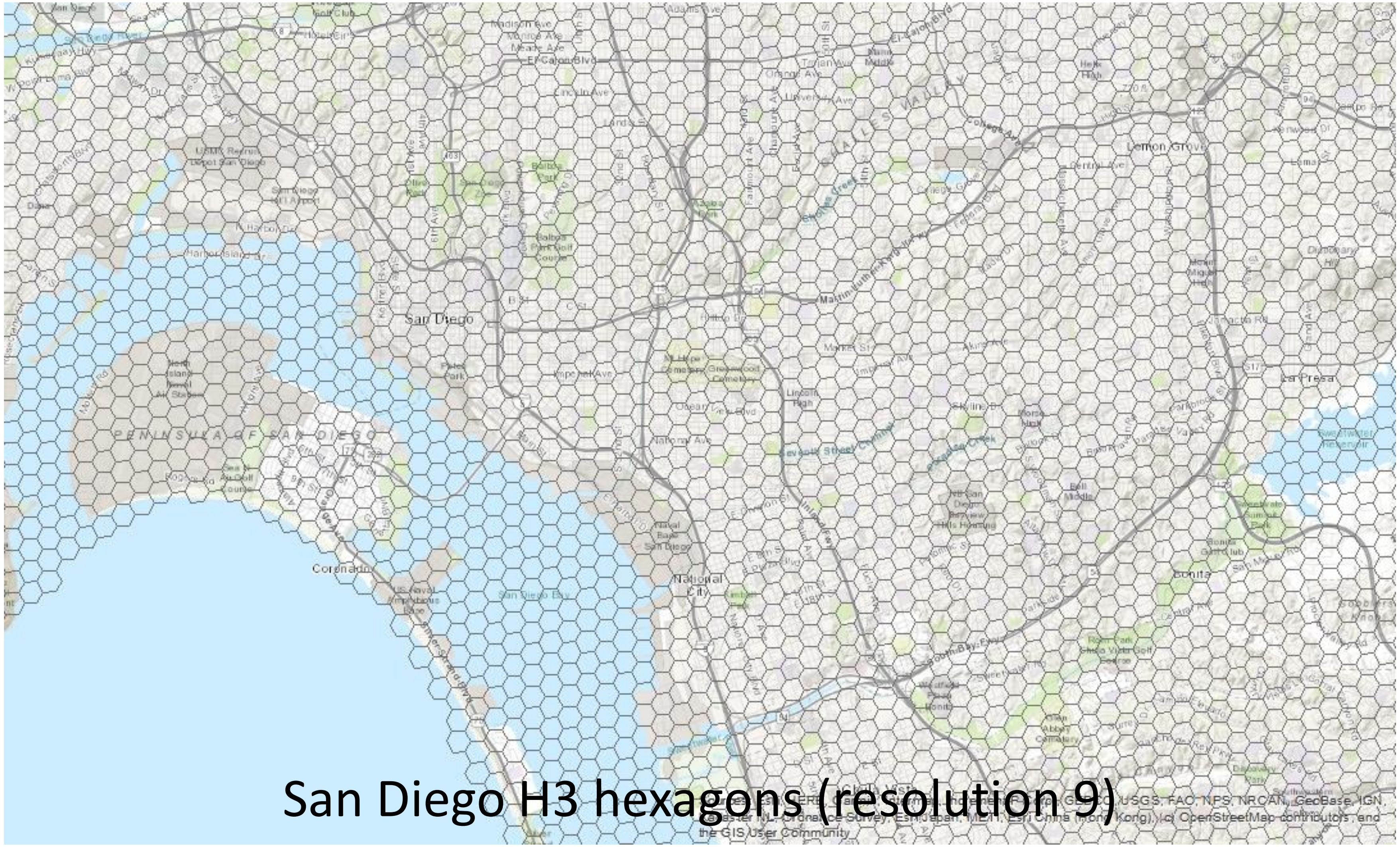
Historic, predictive, spatial-temporal analysis of Tweets

San Diego H3 hexagons (resolution 7)

Historic, predictive, spatial-temporal analysis of Tweets

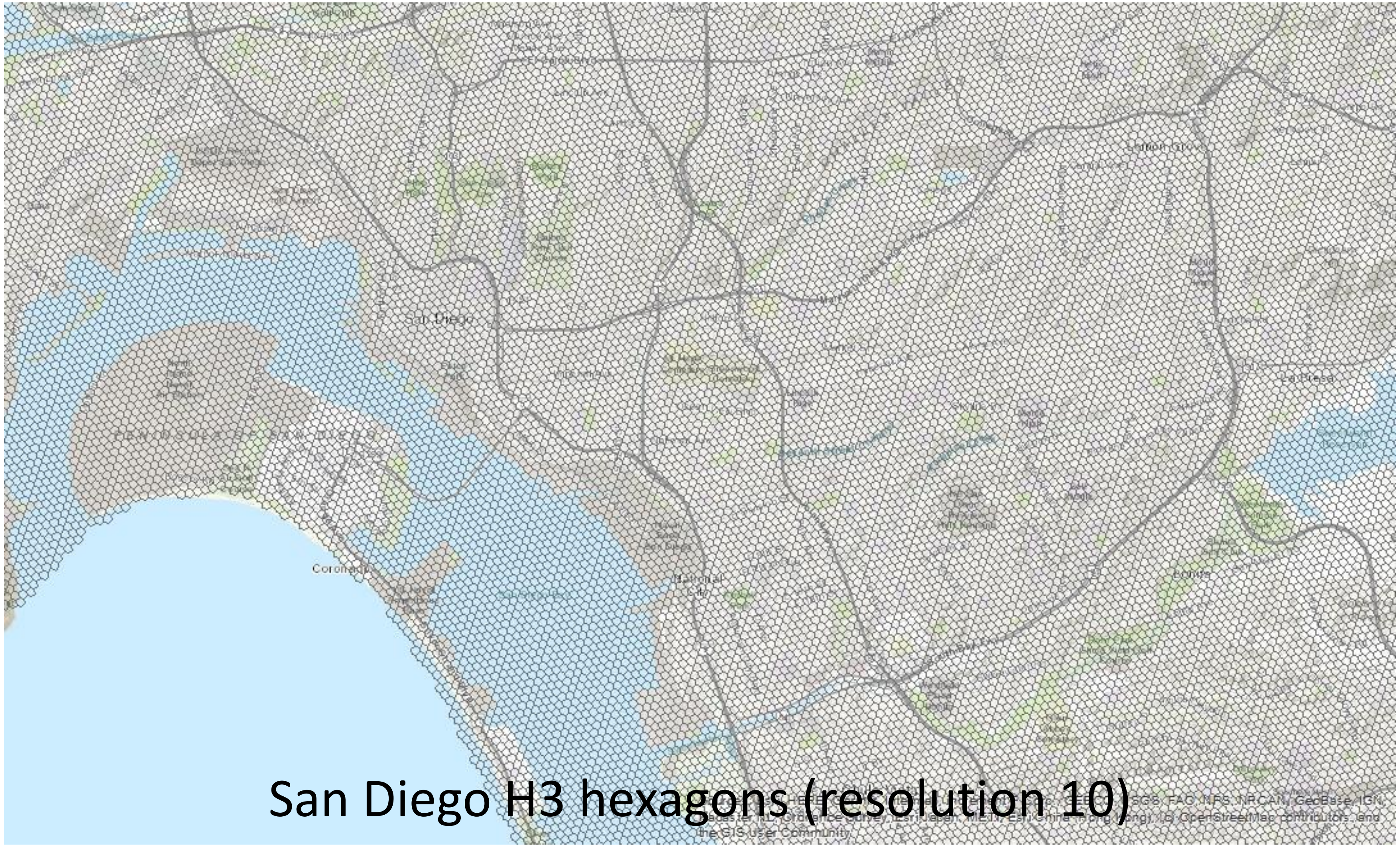
San Diego H3 hexagons (resolution 8)

Historic, predictive, spatial-temporal analysis of Tweets



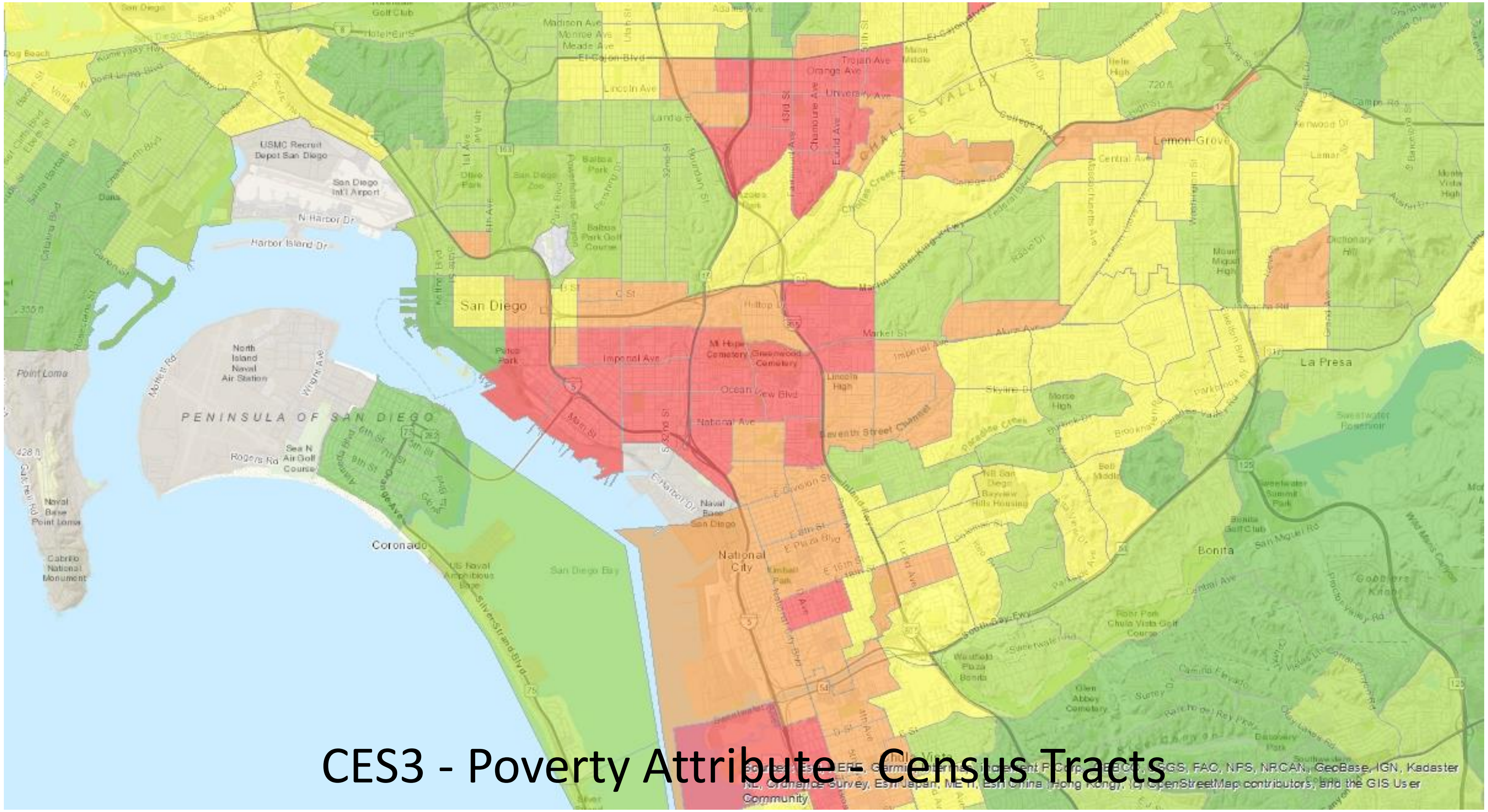
San Diego H3 hexagons (resolution 9)

Historic, predictive, spatial-temporal analysis of Tweets



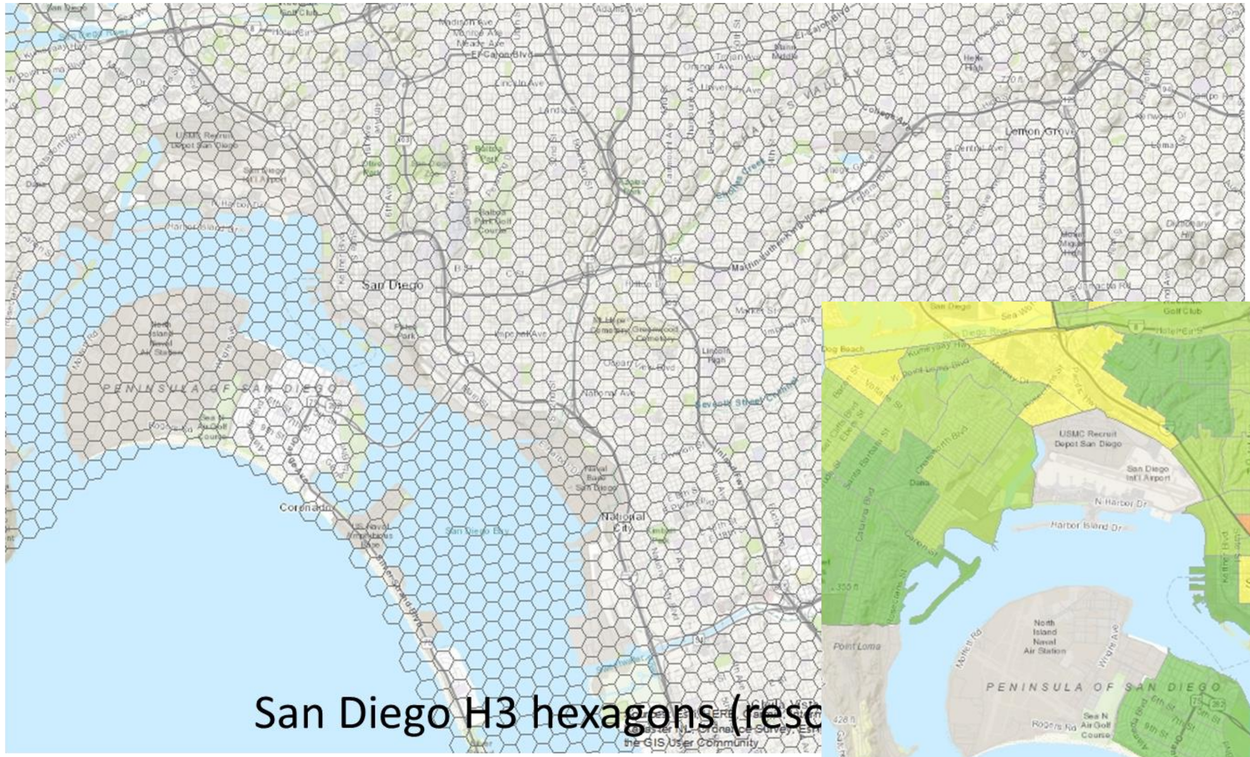
San Diego H3 hexagons (resolution 10)

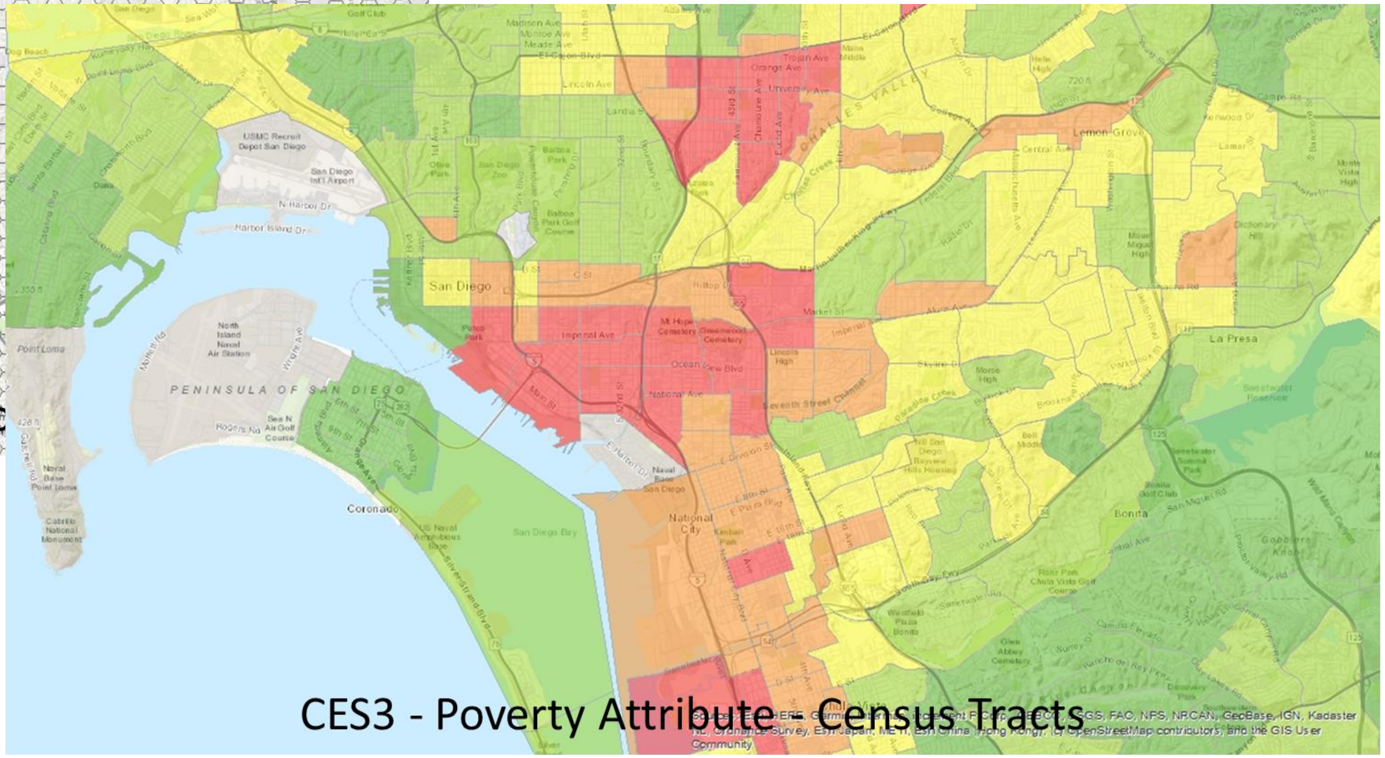
Historic, predictive, spatial-temporal analysis of Tweets



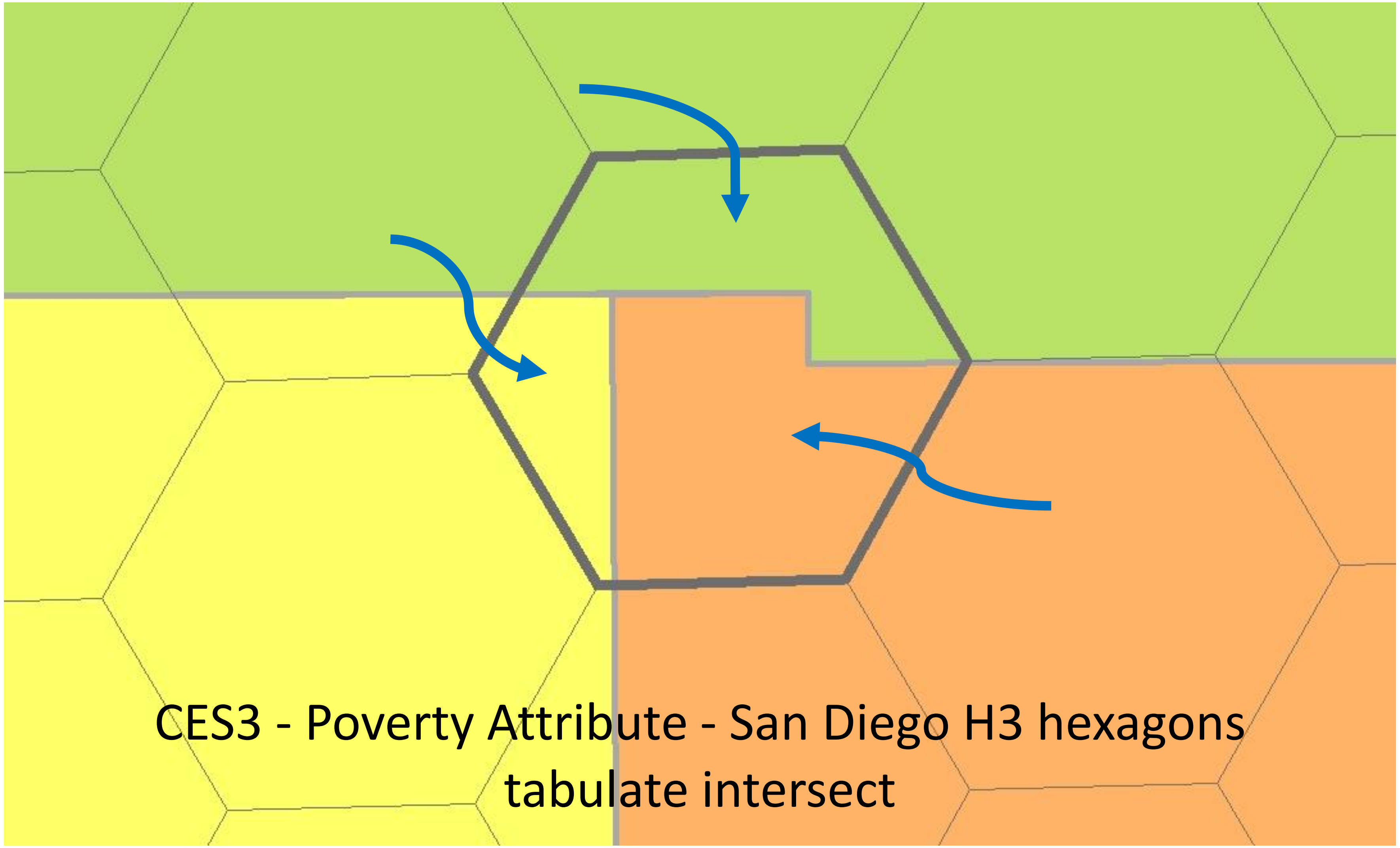
CES3 - Poverty Attribute - Census Tracts

Historic, predictive, spatial-temporal analysis of Tweets





Historic, predictive, spatial-temporal analysis of Tweets



Historic, predictive, spatial-temporal analysis of Tweets

- ArcGIS - San Diego tabulate intersect example - Python notebook
- Link

Historic, predictive, spatial-temporal analysis of Tweets

In [4]:
```
pd.set_option('max_columns', None)
# pd.set_option("max_rows", None)
df = pd.read_csv('gis_analysis\\h3_san_diego_7_areas.csv')
df
```

Out[4]:

	OBJECTID	hex_id	ozone	pm	diesel	drink	pest	RSEIhaz	traffic	cleanups	gwthreats	haz	iwb	swis	asthma	cvd	lbw	edu	housingB
0	1	8729a0902ffffff	0.046	8.28	2.71	554.17	13.408	108.40	2394.83	24.75	199.75	8.10	11	9.5	22.72	4.48	3.97	1.2	36.2
1	2	8729a0906ffffff	0.046	8.28	2.71	554.17	13.408	108.40	2394.83	24.75	199.75	8.10	11	9.5	22.72	4.48	3.97	1.2	36.2
2	3	8729a0910ffffff	0.046	8.28	2.71	554.17	13.408	108.40	2394.83	24.75	199.75	8.10	11	9.5	22.72	4.48	3.97	1.2	36.2
3	4	8729a0910ffffff	0.053	7.44	2.42	907.13	3.031	129.83	682.96	12.00	3.00	0.15	7	0.2	30.39	6.63	4.14	10.6	11.1
4	5	8729a0911ffffff	0.046	8.28	2.71	554.17	13.408	108.40	2394.83	24.75	199.75	8.10	11	9.5	22.72	4.48	3.97	1.2	36.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3037	3038	8729a6b6dffffff	0.048	8.70	0.53	624.70	6.274	11.45	199.96	0.00	0.00	0.00	9	0.0	21.81	5.82	4.96	14.3	14.0
3038	3039	8729a6b6dffffff	0.055	7.38	0.17	1008.75	0.419	6.67	89.93	0.00	8.00	0.00	11	12.4	21.73	4.71	3.16	12.2	11.2
3039	3040	8729a6b6effffff	0.055	7.38	0.17	1008.75	0.419	6.67	89.93	0.00	8.00	0.00	11	12.4	21.73	4.71	3.16	12.2	11.2
3040	3041	8729a6b71ffffff	0.055	7.38	0.17	1008.75	0.419	6.67	89.93	0.00	8.00	0.00	11	12.4	21.73	4.71	3.16	12.2	11.2
3041	3042	8729a6b75ffffff	0.055	7.38	0.17	1008.75	0.419	6.67	89.93	0.00	8.00	0.00	11	12.4	21.73	4.71	3.16	12.2	11.2

3042 rows × 24 columns

Historic, predictive, spatial-temporal analysis of Tweets

In [4]:
```
pd.set_option('max_columns', None)
# pd.set_option("max_rows", None)
df = pd.read_csv('gis_analysis\\h3_san_diego_7_areas.csv')
df
```

Out[4]:

esel	drink	pest	RSEIhaz	traffic	cleanups	gwthreats	haz	iwb	swis	asthma	cvd	lbw	edu	housingB	ling	pov	unemp	AREA	PERCENTAGE
2.71	554.17	13.408	108.40	2394.83	24.75	199.75	8.10	11	9.5	22.72	4.48	3.97	1.2	36.2	0.2	49.4	15.5	8.383652e+06	99.999999
2.71	554.17	13.408	108.40	2394.83	24.75	199.75	8.10	11	9.5	22.72	4.48	3.97	1.2	36.2	0.2	49.4	15.5	8.380721e+06	100.000000
2.71	554.17	13.408	108.40	2394.83	24.75	199.75	8.10	11	9.5	22.72	4.48	3.97	1.2	36.2	0.2	49.4	15.5	7.104579e+06	84.715167
2.42	907.13	3.031	129.83	682.96	12.00	3.00	0.15	7	0.2	30.39	6.63	4.14	10.6	11.1	3.1	31.3	4.5	1.281852e+06	15.284833
2.71	554.17	13.408	108.40	2394.83	24.75	199.75	8.10	11	9.5	22.72	4.48	3.97	1.2	36.2	0.2	49.4	15.5	8.319281e+06	99.198364
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0.53	624.70	6.274	11.45	199.96	0.00	0.00	0.00	9	0.0	21.81	5.82	4.96	14.3	14.0	3.0	38.9	9.0	8.407951e+04	1.002178
0.17	1008.75	0.419	6.67	89.93	0.00	8.00	0.00	11	12.4	21.73	4.71	3.16	12.2	11.2	2.4	47.0	18.5	8.305597e+06	98.997823
0.17	1008.75	0.419	6.67	89.93	0.00	8.00	0.00	11	12.4	21.73	4.71	3.16	12.2	11.2	2.4	47.0	18.5	8.392271e+06	100.000000
0.17	1008.75	0.419	6.67	89.93	0.00	8.00	0.00	11	12.4	21.73	4.71	3.16	12.2	11.2	2.4	47.0	18.5	8.394663e+06	100.000003
0.17	1008.75	0.419	6.67	89.93	0.00	8.00	0.00	11	12.4	21.73	4.71	3.16	12.2	11.2	2.4	47.0	18.5	8.391717e+06	99.999999

Historic, predictive, spatial-temporal analysis of Tweets

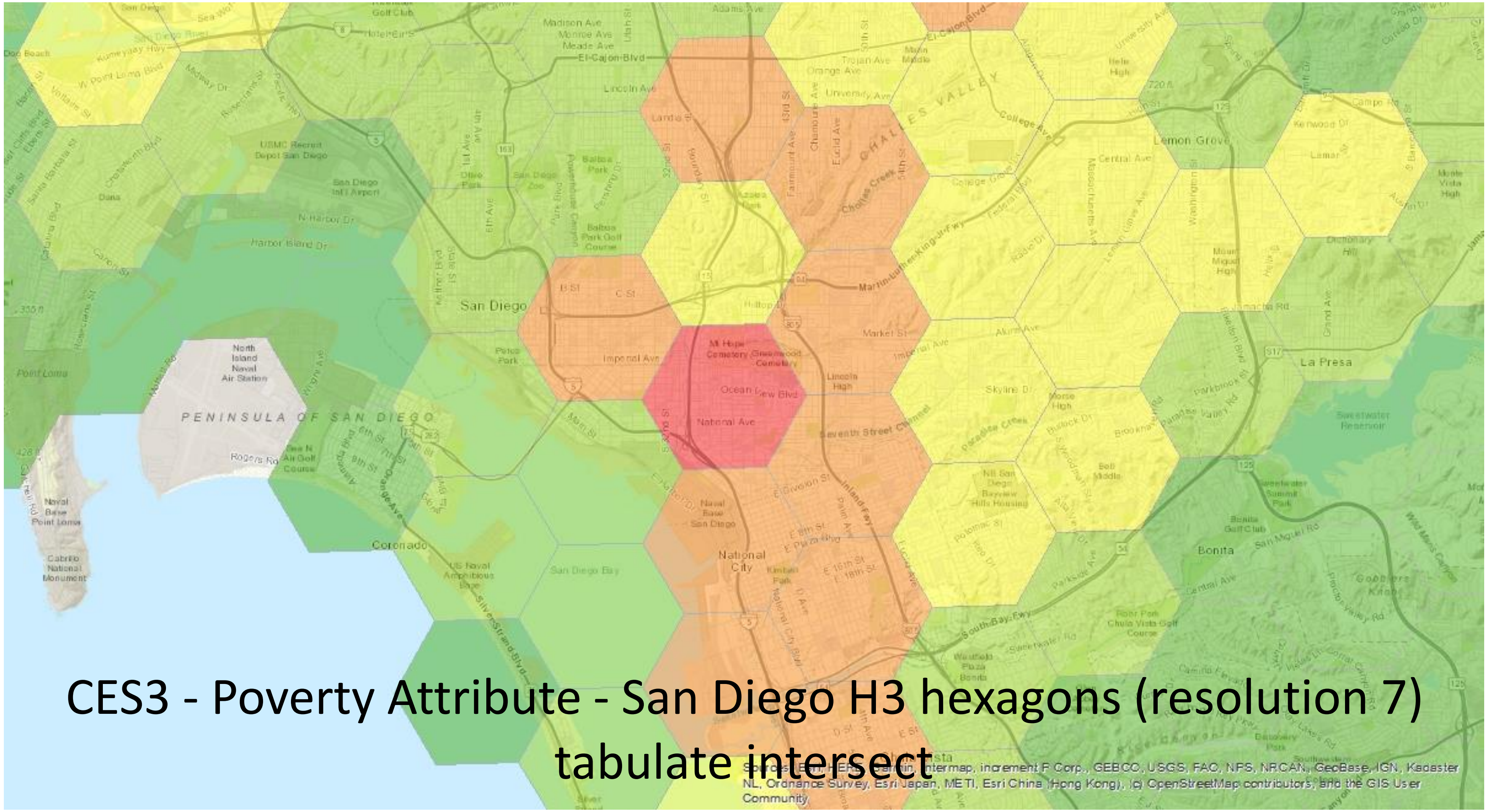
```
y, SUM(PERCENTAGE1*swis) as Solid_Waste_Sites, SUM(PERCENTAGE1*asthma) as Asthma, SUM(PERCENTAGE1*cvd) as Cardiovascula
r_Disease, SUM(PERCENTAGE1*lbw) as Low_Birth_Weight, SUM(PERCENTAGE1*edu) as Educational_Attainment, SUM(PERCENTAGE1*ho
usingB) as Housing_Burden, SUM(PERCENTAGE1*ling) as Linguistic_Isolation, SUM(PERCENTAGE1*pov) as Poverty, SUM(PERCENTA
GE1*unemp) as Unemployment FROM df_sql GROUP BY hex_id;")
df_sql2
```

Out[6]:

	hex_id	Ozone	PM_2_5	Diesel_PM	Drinking_Water	Pesticide_Use	Toxic_Releases	Traffic	Cleanup_Sites	Groundwater_Threats	Hazardous_'
0	8729a0902ffffff	0.046000	8.280000	2.710000	554.169997	13.408000	108.399999	2394.829988	24.750000	199.749999	8.1
1	8729a0906ffffff	0.046000	8.280000	2.710000	554.170001	13.408000	108.400000	2394.830005	24.750000	199.750000	8.1
2	8729a0910ffffff	0.04[illegible]070	8.151607	2.665674	608.119343	11.821893	111.675539	2133.173519	22.801184	169.677090	6.8
3	8729a0911ffffff	0.046056	8.273266	2.707675	556.999438	13.324814	108.571788	2381.107022	24.647791	198.172781	8.0
4	8729a0912ffffff	0.051698	7.596285	2.473956	841.460554	4.961677	125.842873	1001.459408	14.372182	39.606027	1.6
...	...	...	...	...	...	...	...	...	...	...	
963	8729a6b6cffffff	0.055000	7.380000	0.170000	1008.750011	0.419000	6.670000	89.930001	0.000000	8.000000	0.0
964	8729a6b6dffffff	0.054930	7.393229	0.173608	1004.901151	0.477678	6.717904	91.032698	0.000000	7.919826	0.0
965	8729a6b6effffff	0.055000	7.380000	0.170000	1008.749997	0.419000	6.670000	89.930000	0.000000	8.000000	0.0
966	8729a6b71ffffff	0.055000	7.380000	0.170000	1008.750028	0.419000	6.670000	89.930002	0.000000	8.000000	0.0
967	8729a6b75ffffff	0.055000	7.380000	0.170000	1008.749991	0.419000	6.670000	89.929999	0.000000	8.000000	0.0

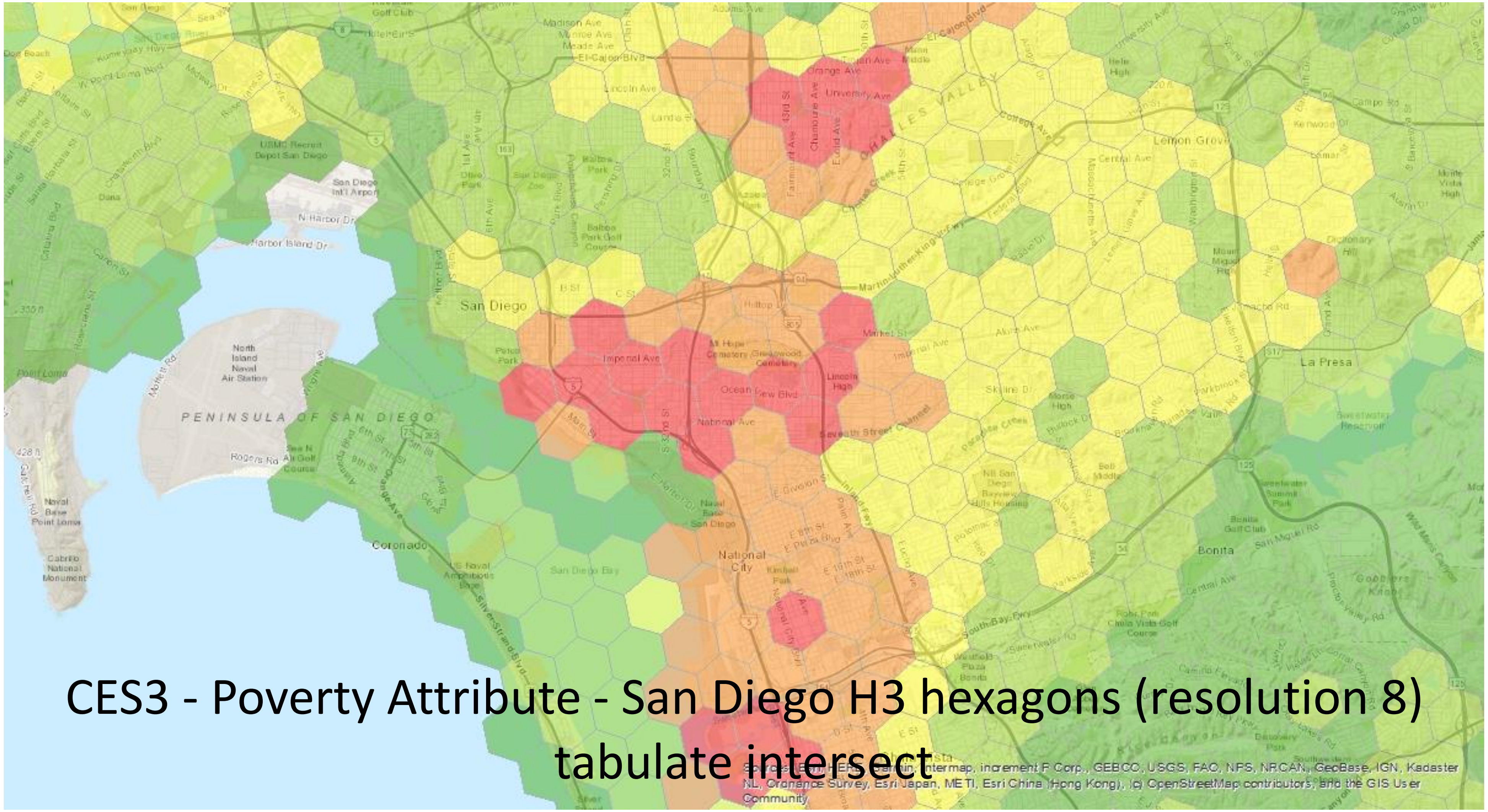
968 rows × 21 columns

Historic, predictive, spatial-temporal analysis of Tweets



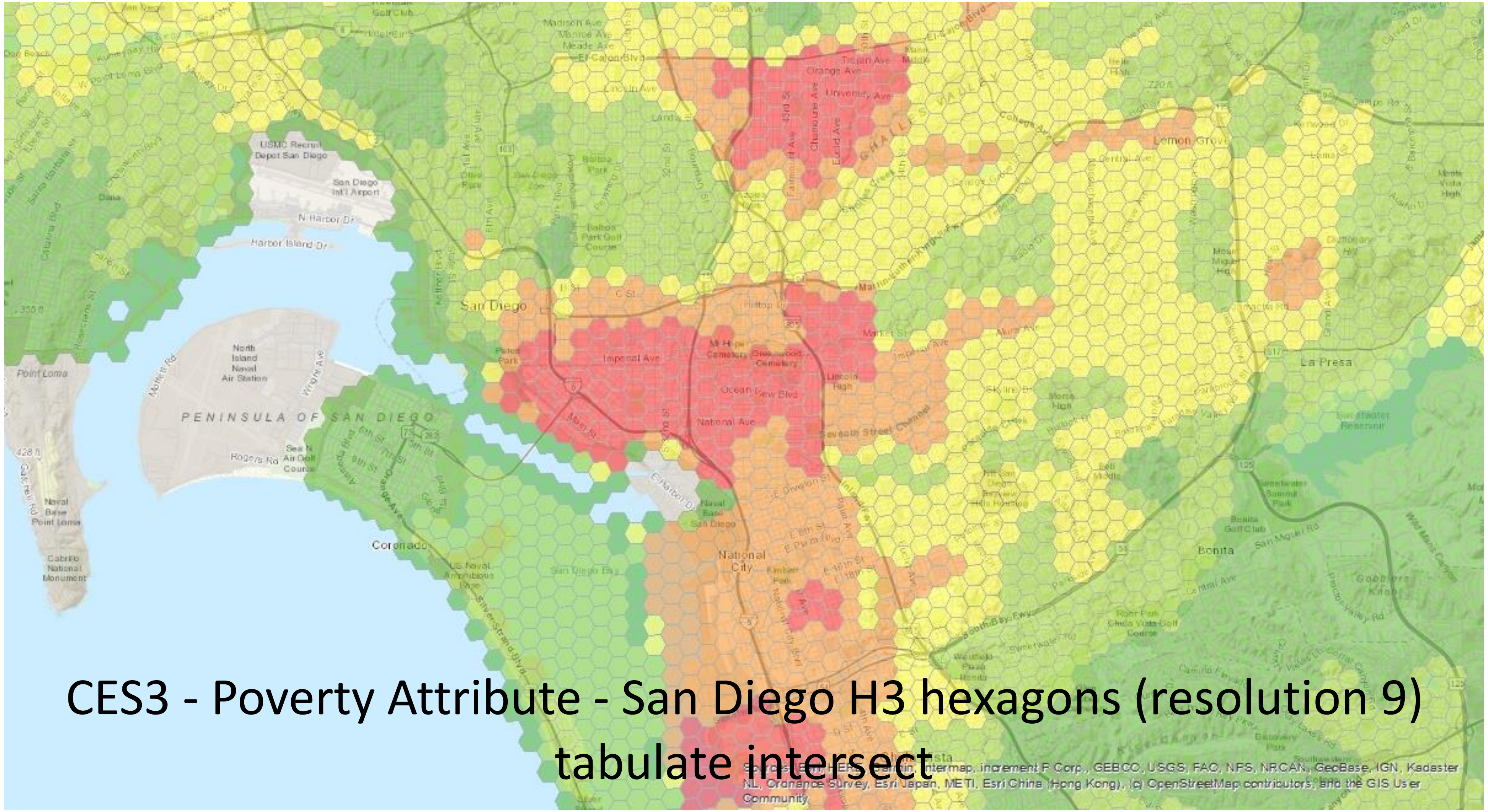
CES3 - Poverty Attribute - San Diego H3 hexagons (resolution 7) tabulate intersect

Historic, predictive, spatial-temporal analysis of Tweets



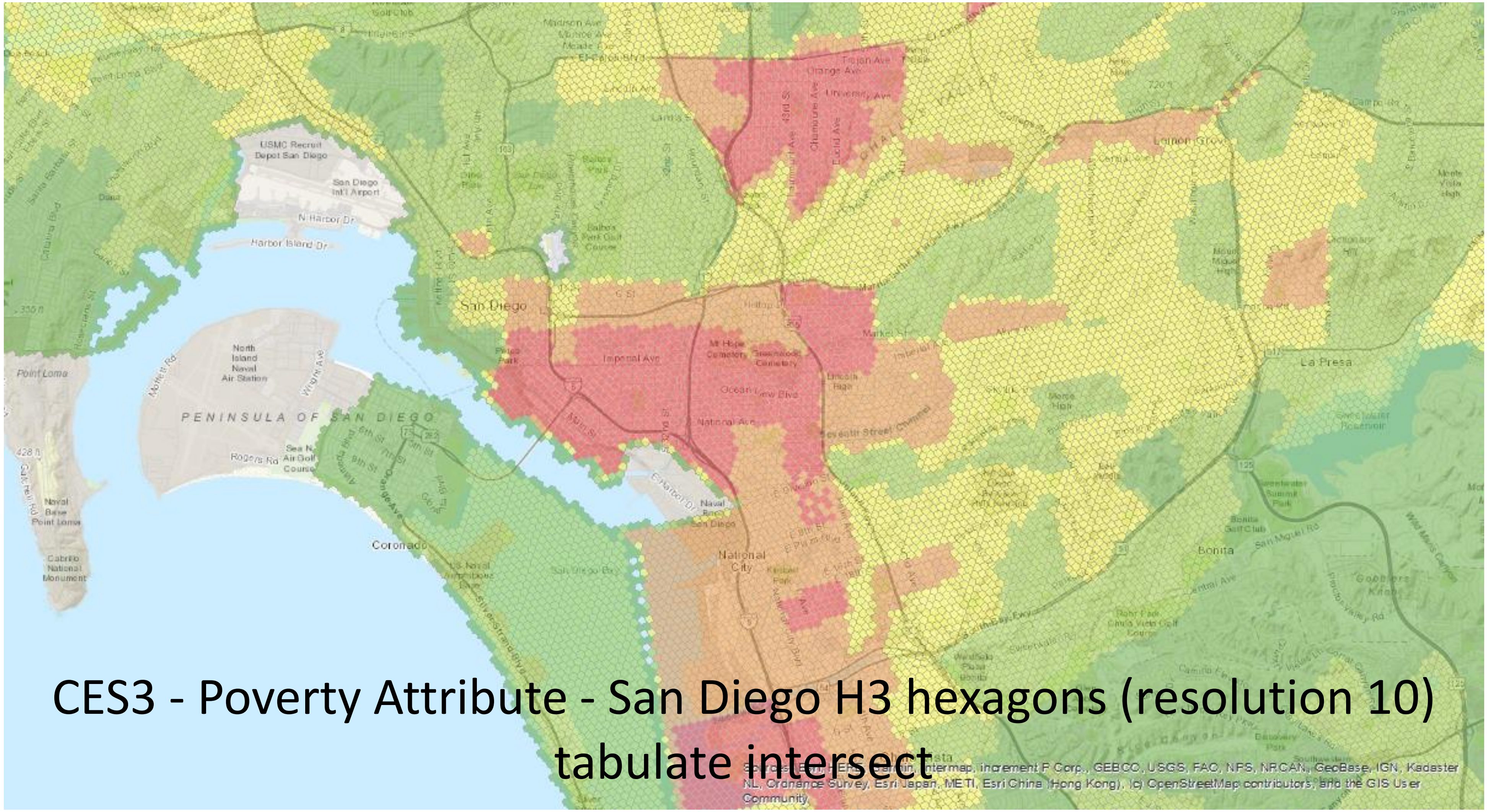
CES3 - Poverty Attribute - San Diego H3 hexagons (resolution 8) tabulate intersect

Historic, predictive, spatial-temporal analysis of Tweets



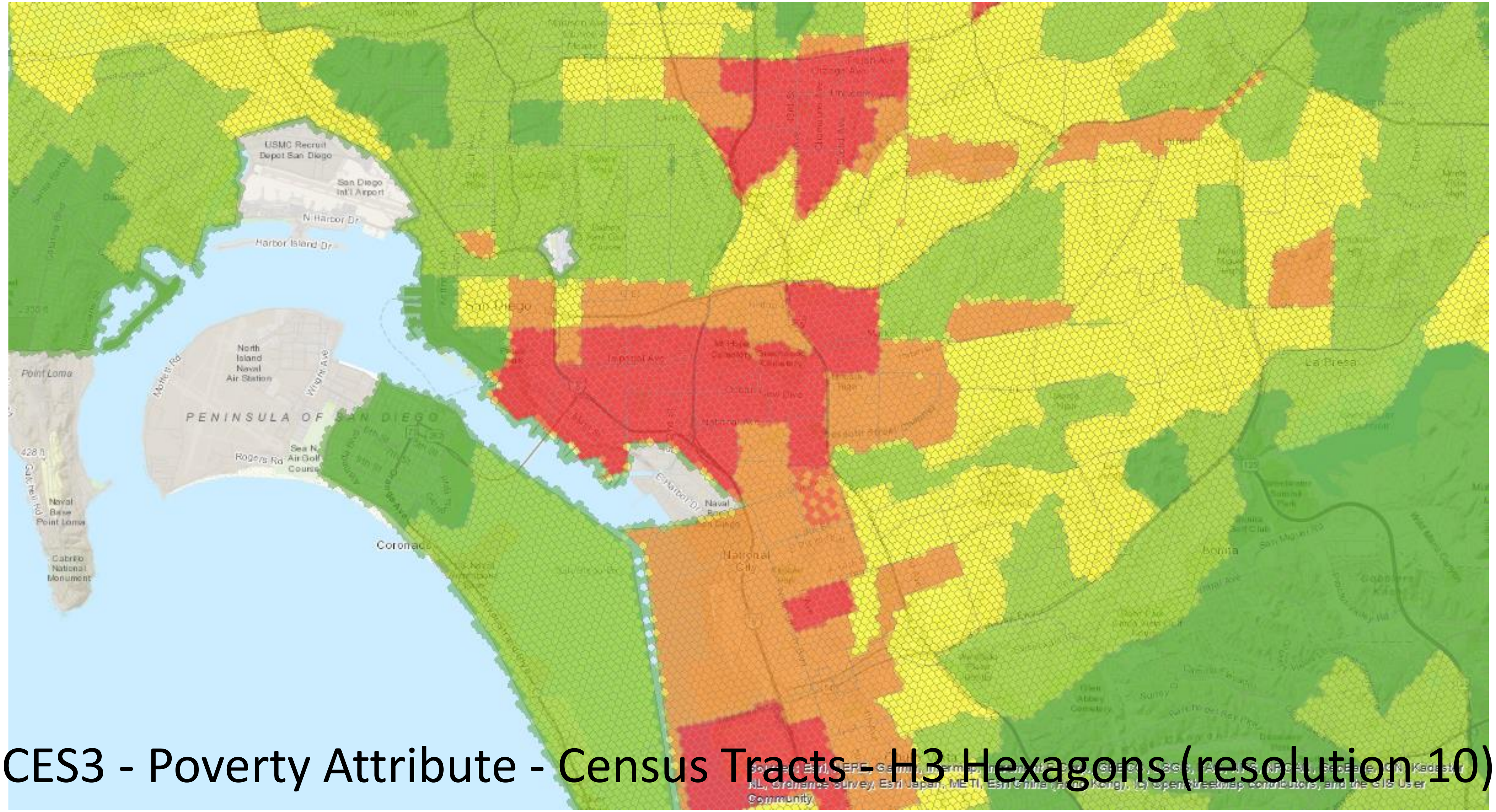
CES3 - Poverty Attribute - San Diego H3 hexagons (resolution 9) tabulate intersect

Historic, predictive, spatial-temporal analysis of Tweets



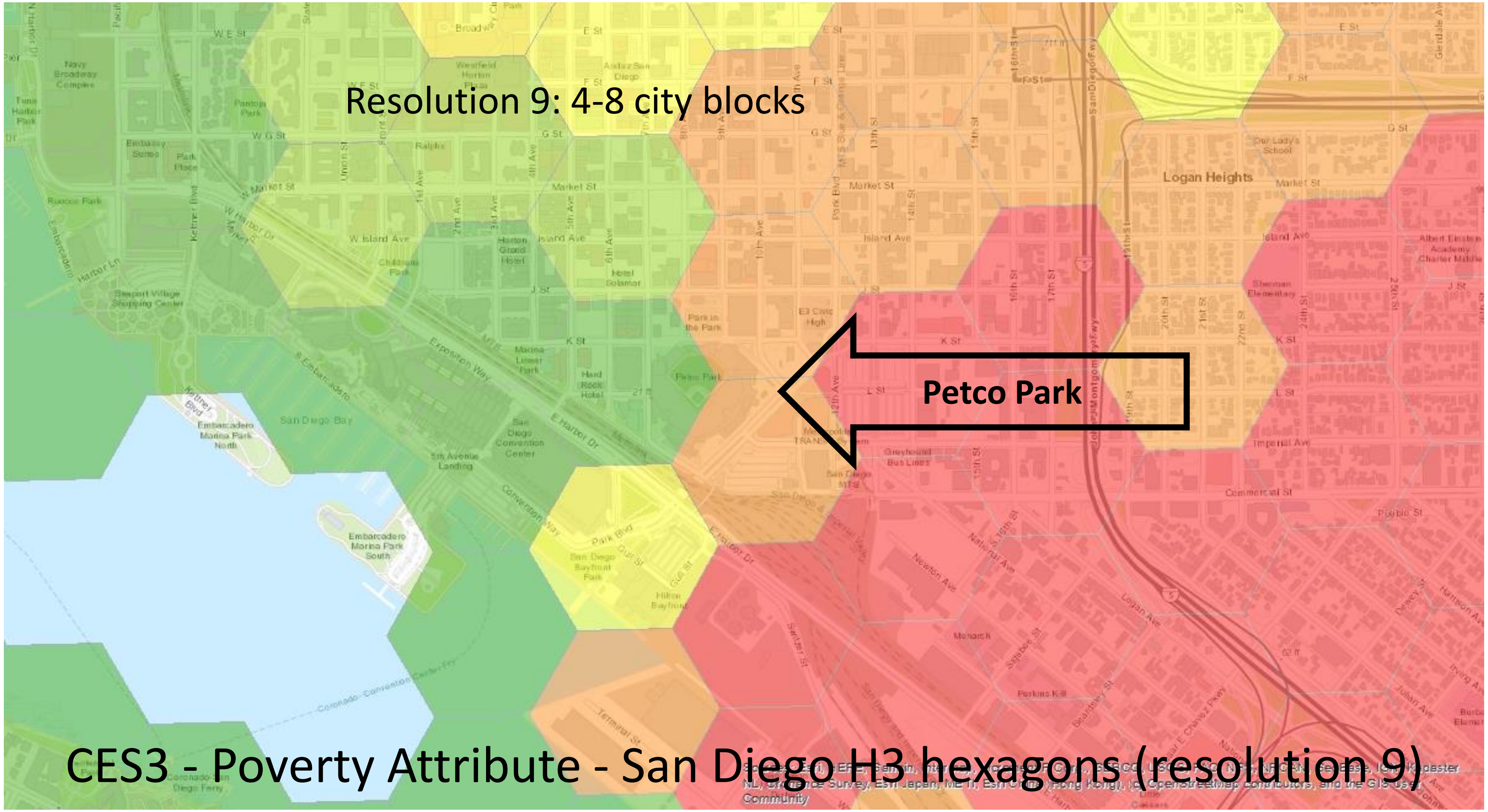
CES3 - Poverty Attribute - San Diego H3 hexagons (resolution 10) tabulate intersect

Historic, predictive, spatial-temporal analysis of Tweets



CES3 - Poverty Attribute - Census Tracts - H3 Hexagons (resolution 10)

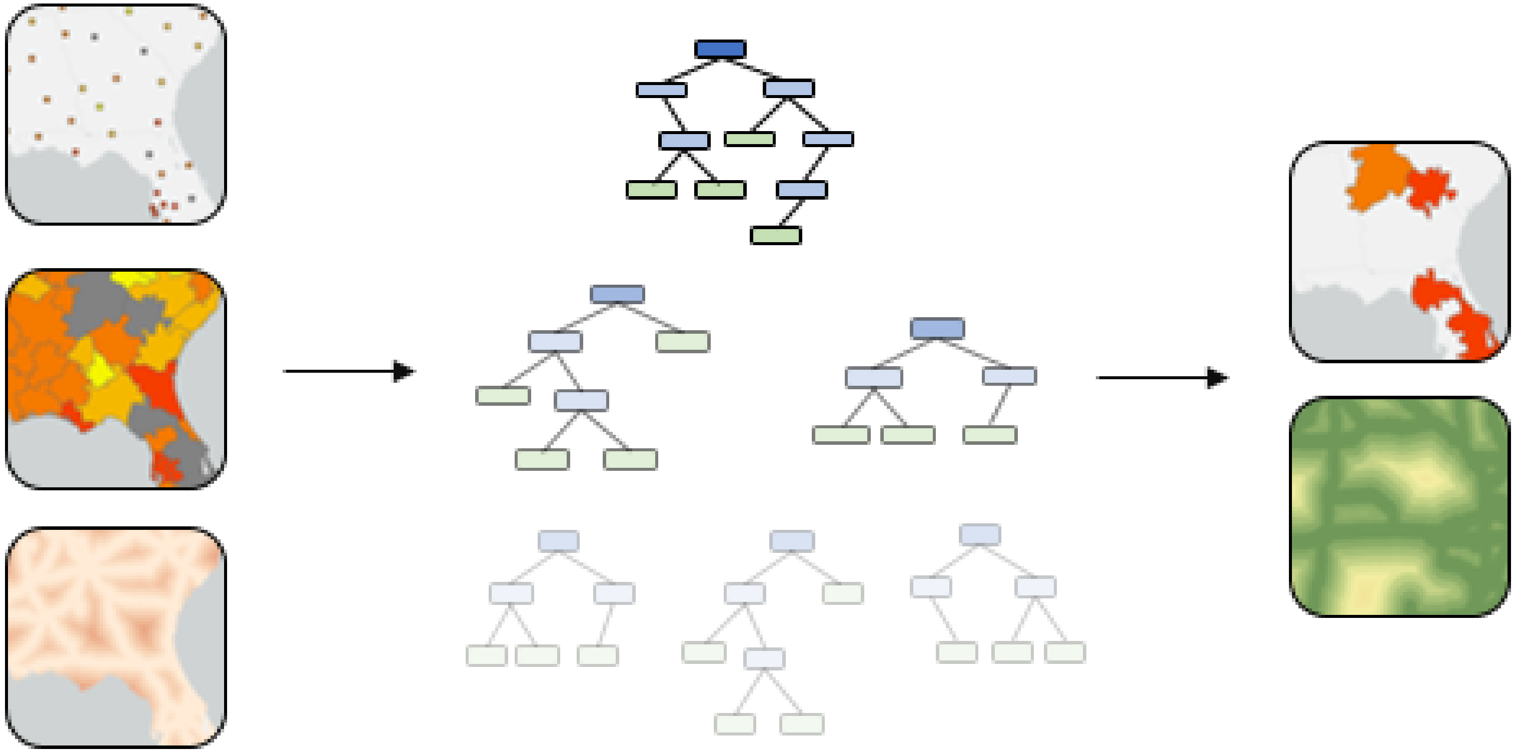
Historic, predictive, spatial-temporal analysis of Tweets



CES3 - Poverty Attribute - San Diego H3 hexagons (resolution 9)

Historic, predictive, spatial-temporal analysis of Tweets

- Forest-based Classification and Regression



Historic, predictive, spatial-temporal analysis of Tweets

- Forest-based Classification and Regression

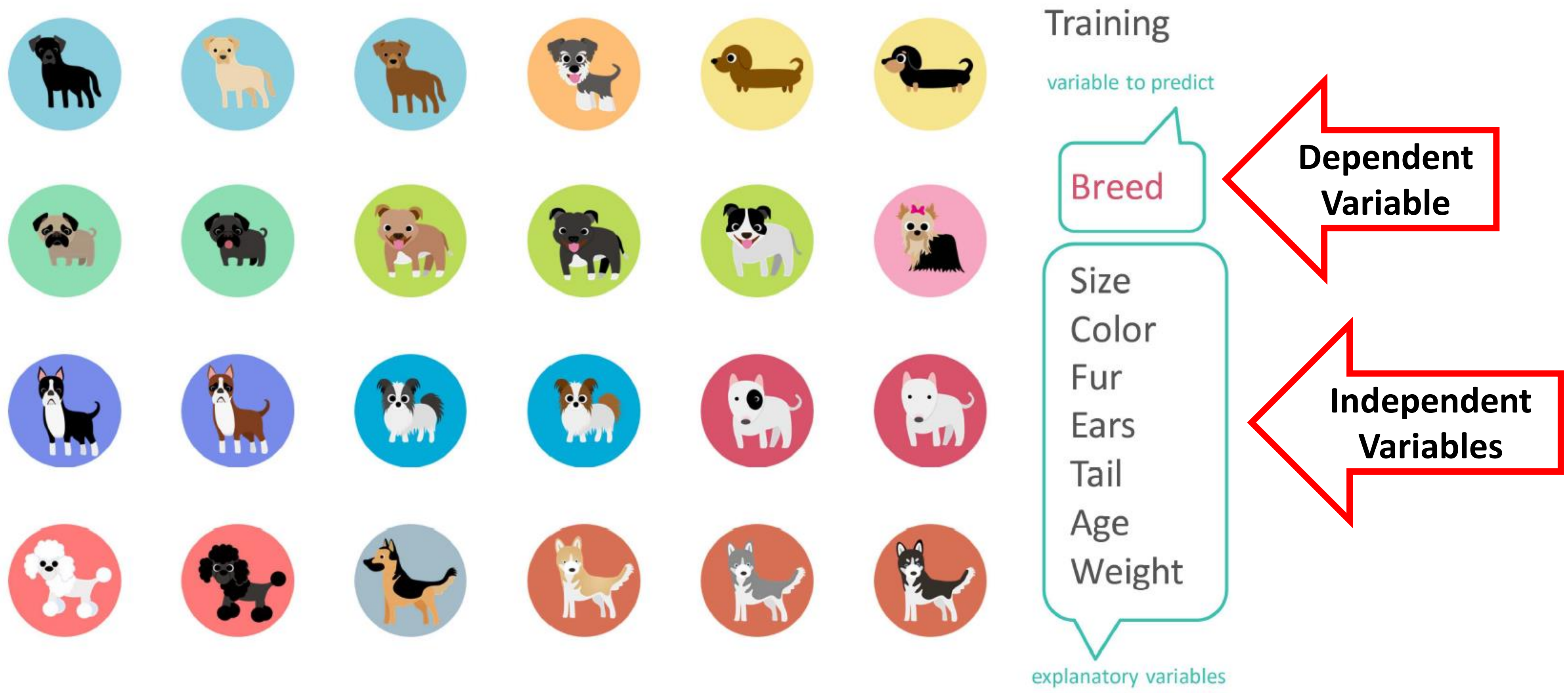
- Many **decision trees** are created, called an ensemble or a forest, that are used for **prediction**.
- Each **tree** generates its own prediction and is used as part of a **voting scheme** to make **final predictions**.
- Final predictions are not based on **any single tree** but rather on the **entire forest**.

Historic, predictive, spatial-temporal analysis of Tweets

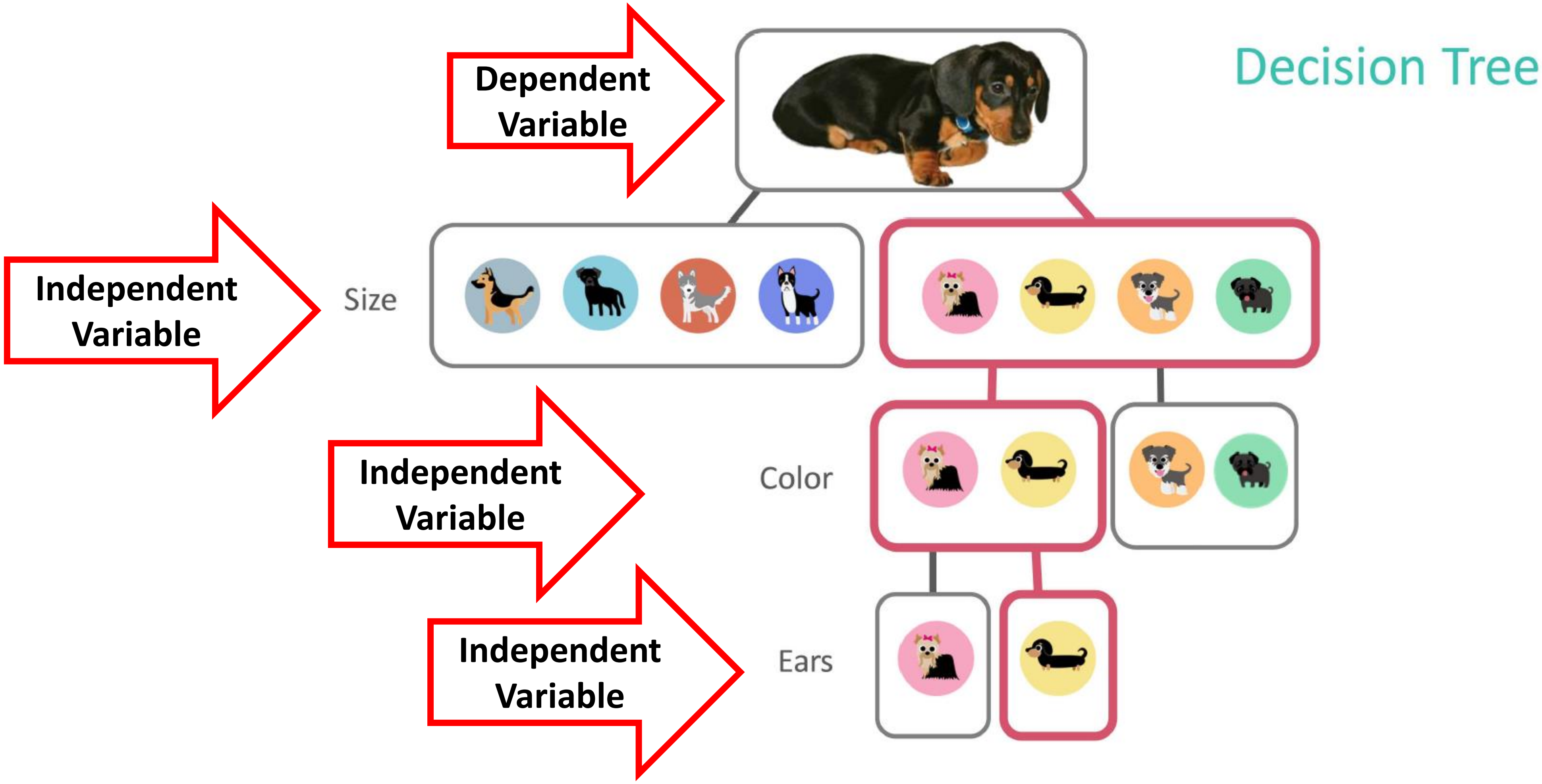
- Forest-based Classification and Regression

- The use of the entire forest helps **avoid overfitting the model** to the training dataset,
- as does the use of both a **random subset** of the training data and a **random subset of explanatory variables** in each tree that constitutes the forest.

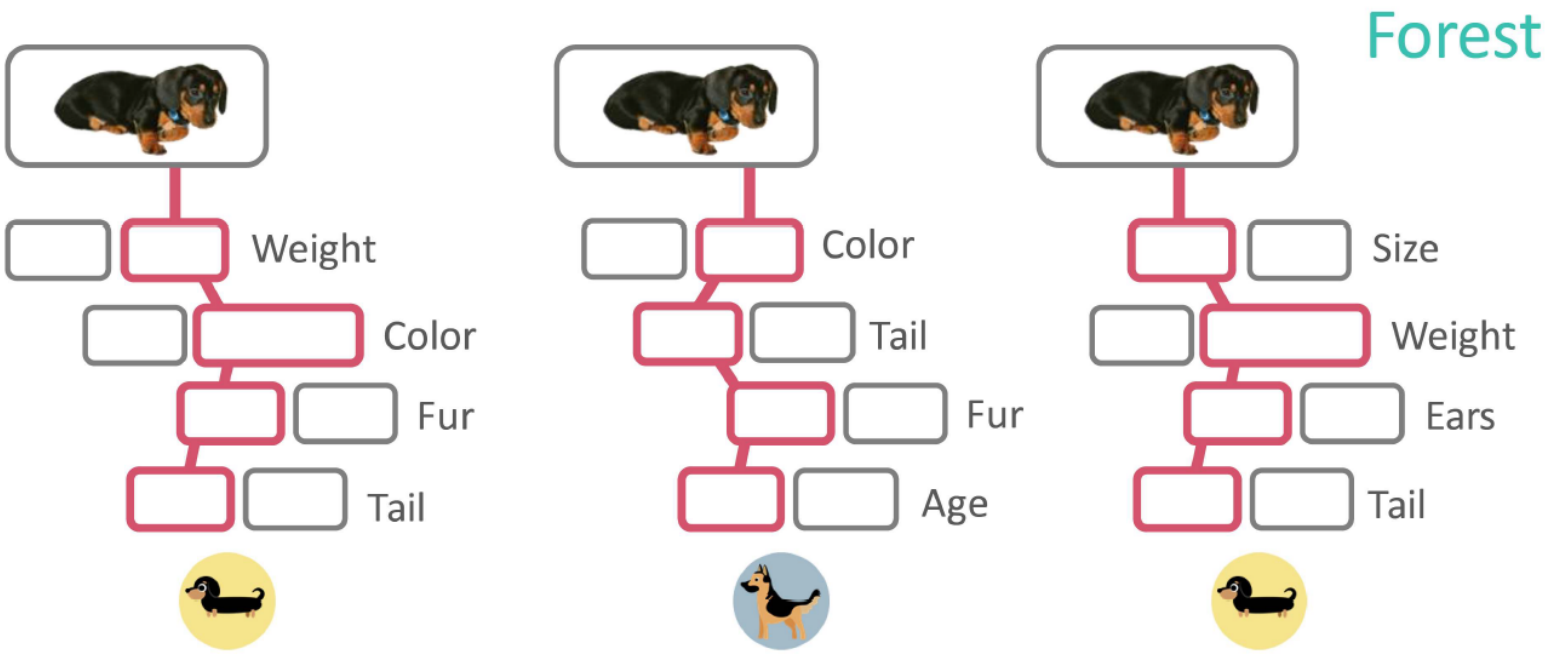
Historic, predictive, spatial-temporal analysis of Tweets



Historic, predictive, spatial-temporal analysis of Tweets

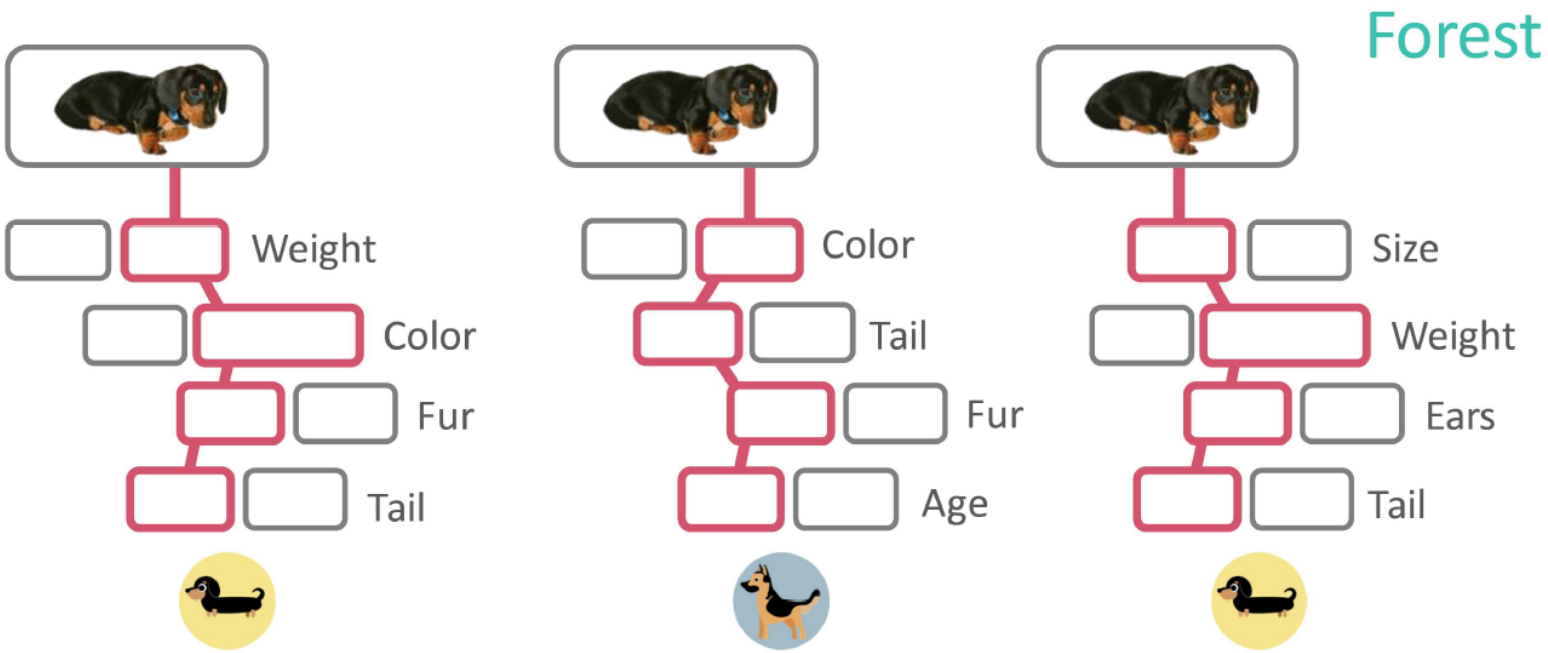


Historic, predictive, spatial-temporal analysis of Tweets



Random subset of data and variables used in each tree

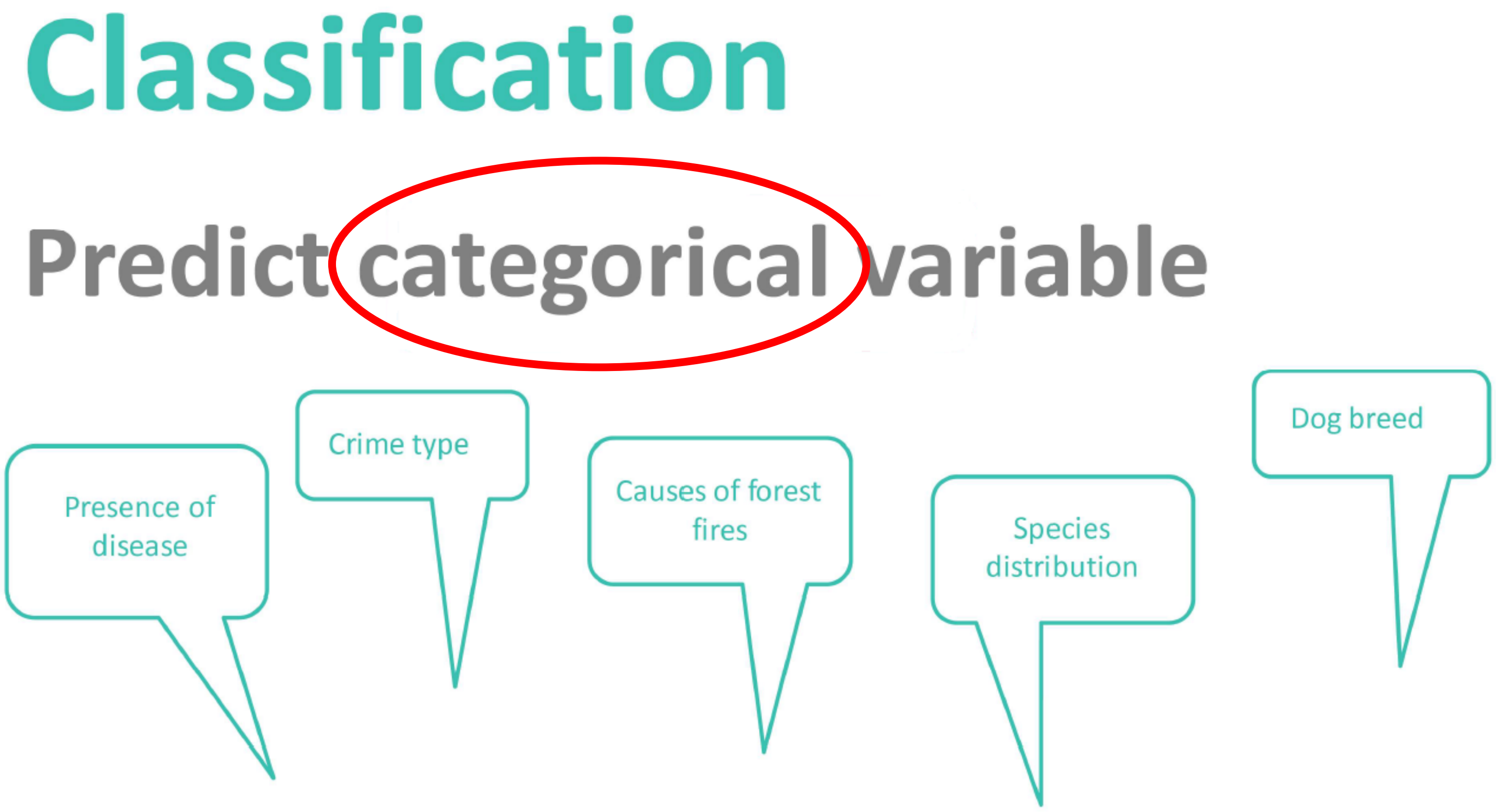
Historic, predictive, spatial-temporal analysis of Tweets



Majority vote wins

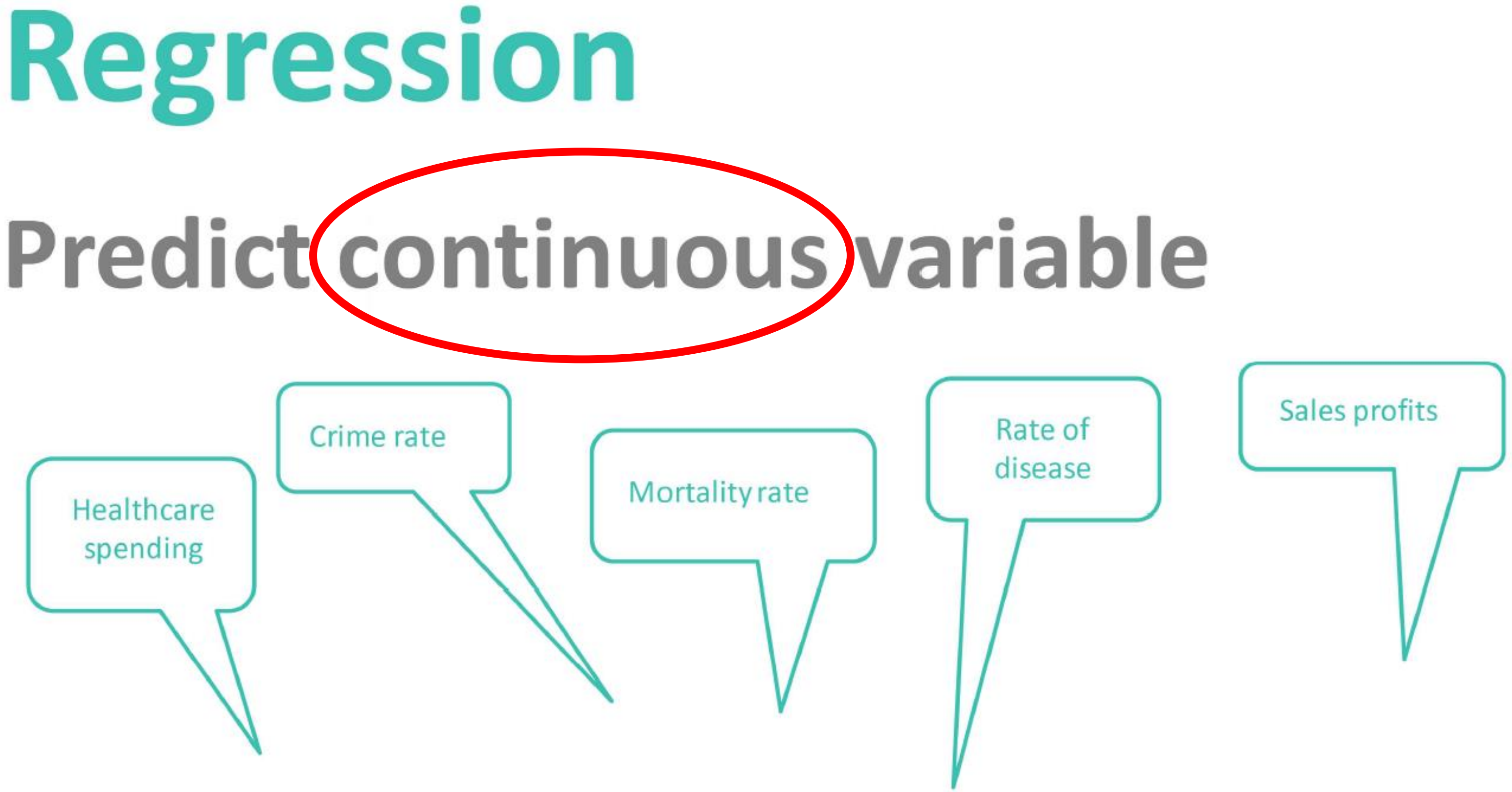
Historic, predictive, spatial-temporal analysis of Tweets

- Forest-based Classification and Regression



Historic, predictive, spatial-temporal analysis of Tweets

- Forest-based Classification and Regression



Historic, predictive, spatial-temporal analysis of Tweets

- Forest-based Classification and Regression - Model Parameters
  - Predict Emotion (Happy, Neutral, Sad) based on CES3 Population Characteristics
  - 90 training / 10 validation split, 100 trees, 100 iterations

Model 1 Variables	**Model 2 Variables**	**Model 3 Variables**
Unemployment	Unemployment	Poverty
Poverty	Poverty	Asthma
Linguistic Isolation	Housing Burden	Cardiovascular Disease
Housing Burden		
Educational Attainment		

Historic, predictive, spatial-temporal analysis of Tweets

- Forest-based Classification and Regression - Results

	Model 1			Model 2			Model 3			
Emotion	Actual	Predicted	%_Correct	Actual	Predicted	%_Correct	Actual	Predicted	%_Correct	
Happy	266	176	66.16541353	269	204	75.83643123	269	195	72.49070632	**Under**
Neutral	280	421	150.3571429	280	433	154.6428571	280	400	142.8571429	**Over**
Sad	261	210	80.45977011	263	175	66.53992395	263	217	82.5095057	**Under**

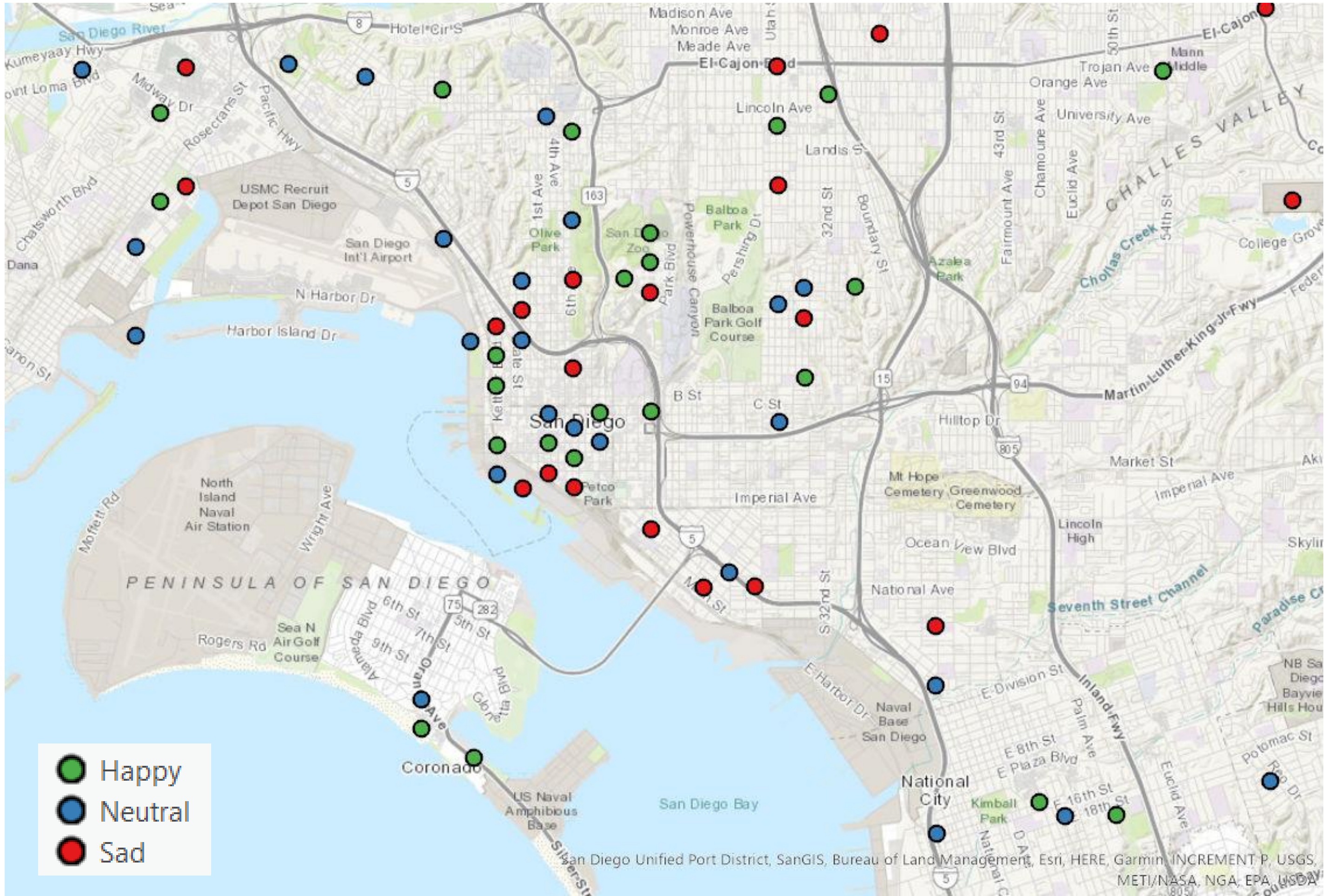
Historic, predictive, spatial-temporal analysis of Tweets

- Forest-based Classification and Regression - Results

Model 1 Variables	Importance	Model 2 Variables	Importance	Model 3 Variables	Importance
Unemployment	21%	Unemployment	36%	Poverty	34%
Poverty	18%	Poverty	29%	Asthma	35%
Linguistic Isolation	17%	Housing Burden	35%	Cardiovascular Disease	31%
Housing Burden	22%				
Educational Attainment	21%				

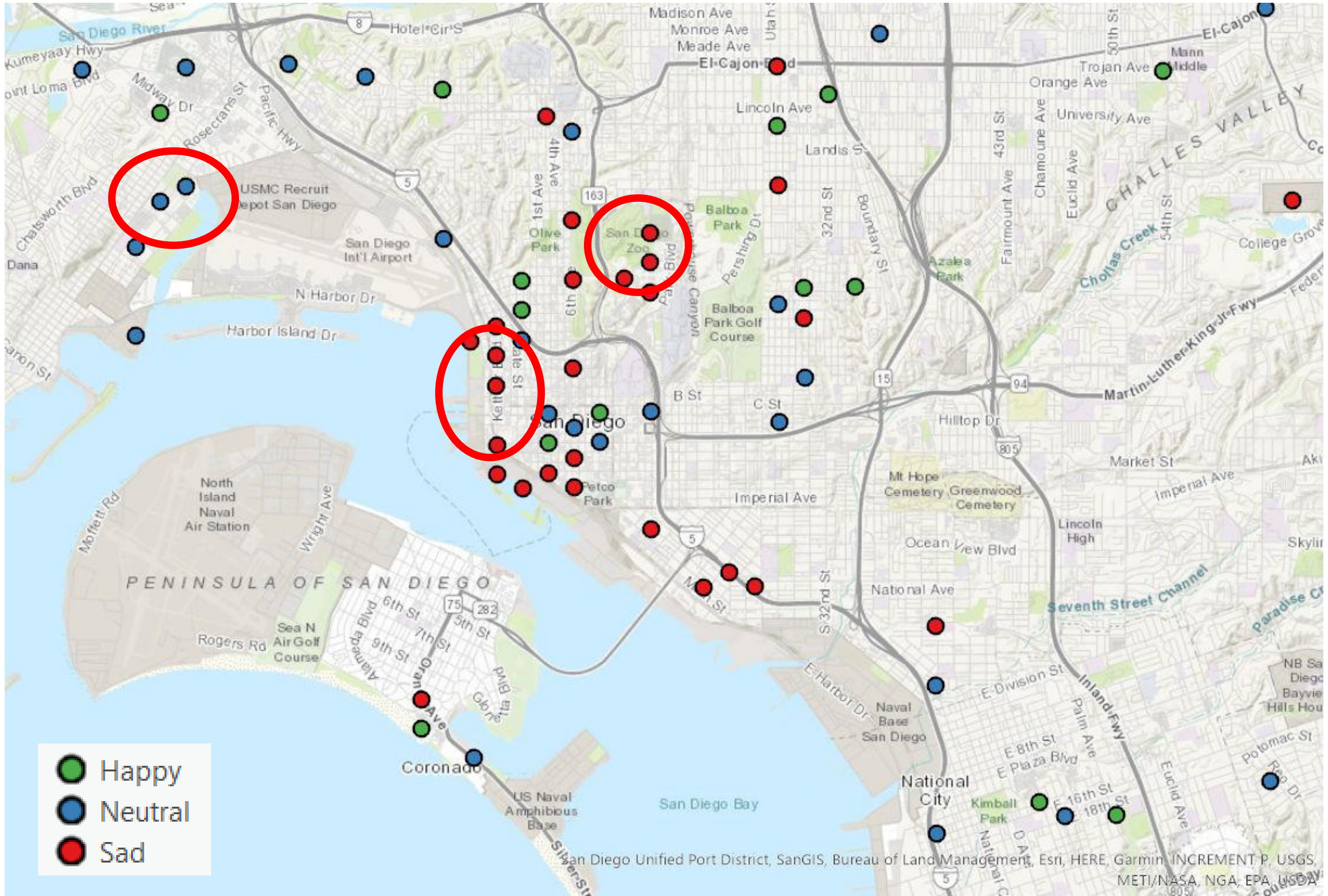
Historic, predictive, spatial-temporal analysis of Tweets

- Forest-based Classification and Regression - Model 1 Actual



Historic, predictive, spatial-temporal analysis of Tweets

- Forest-based Classification and Regression - Model 1 Predicted



Live Demo

- Demonstration using ArcGIS Insights

- Demonstration using ArcGIS Pro