Combining GIS and Statistics: Data Visualization to Communicate Findings

Kristen Hocutt

Lu Zhang

March 7, 2019
Washington Statistical Society
WASHINGTON, DC

Lakers Playoff Game
@Staples Center

Cyber attack
@Washington DC

Power Outage
@Eugene area locations

HR Incident
@Scranton, PA

Conference
@Phoenix City Center

Everything happens somewhere

Armed Protester
@Seattle Distribution Center

New Starbucks
@Downtown Phoenix

Amazon Prime

Route Disruption
@Reno & Irvine Locations

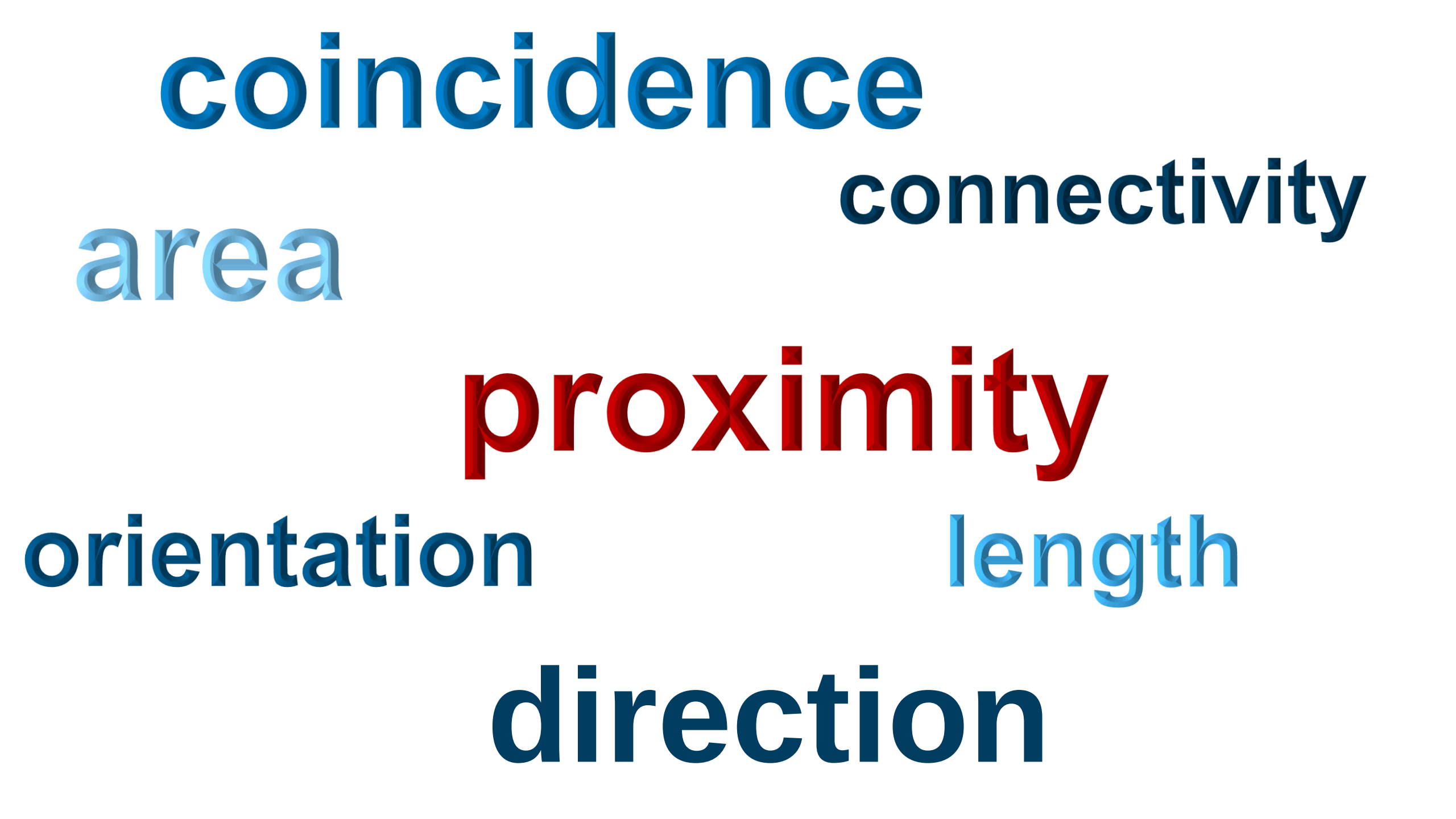
Machine down
@San Francisco P&DC

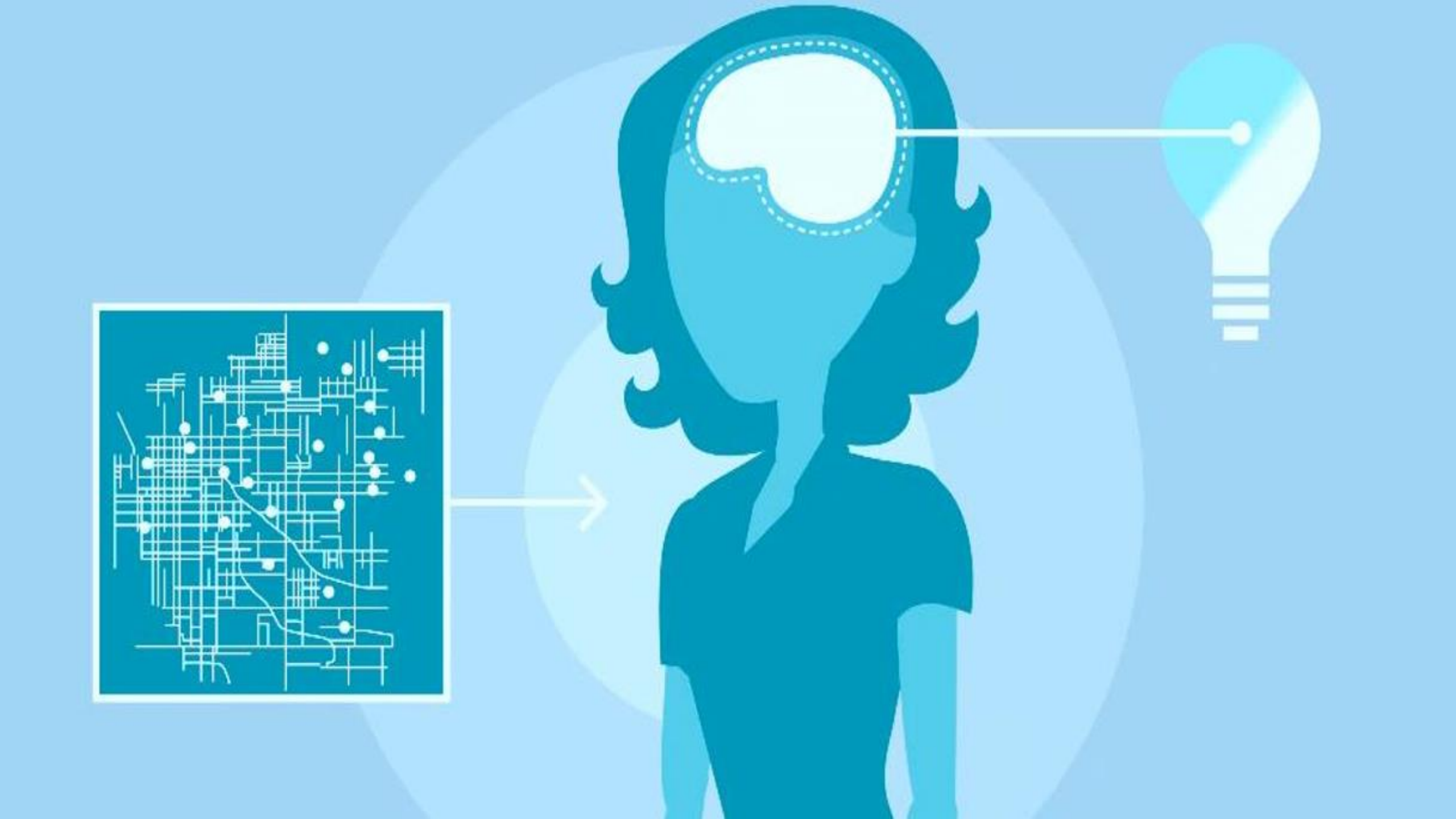
Winter Weather
@Washington DC

Nearest Food truck
Landing @ Salt Lake International Airport

What are Spatial Statistics?

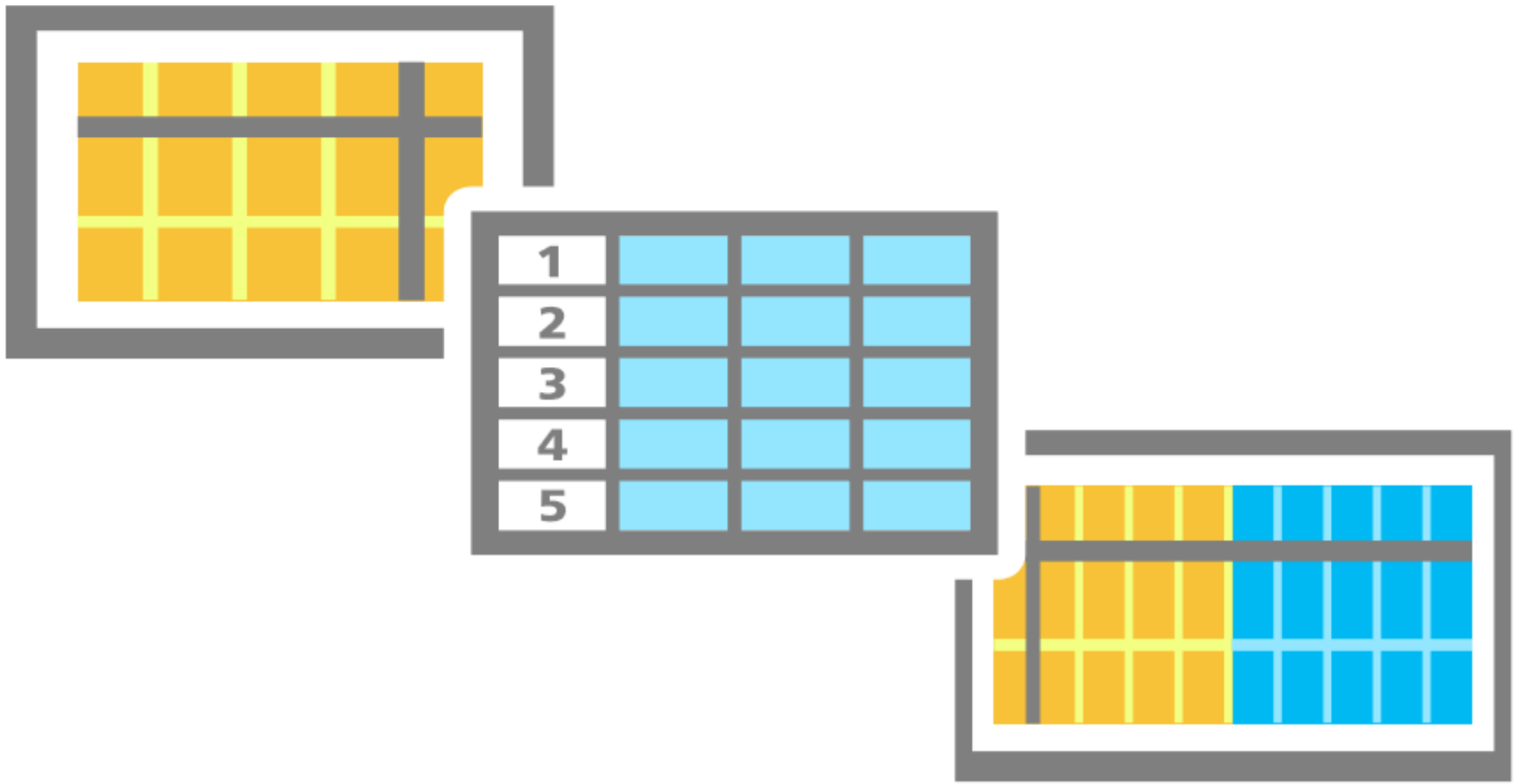
Spatial Statistics are a set of exploratory techniques for describing and modeling spatial distributions, patterns, processes, and relationships.


coincidence
connectivity
area
proximity
orientation
length
direction



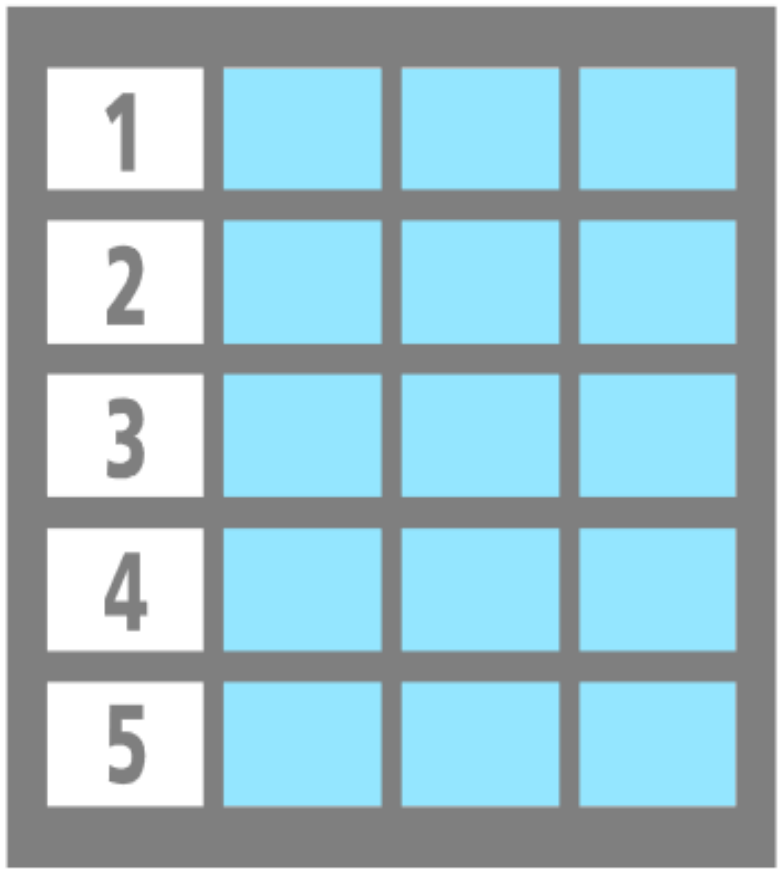
Spreadsheets

Data or Information?

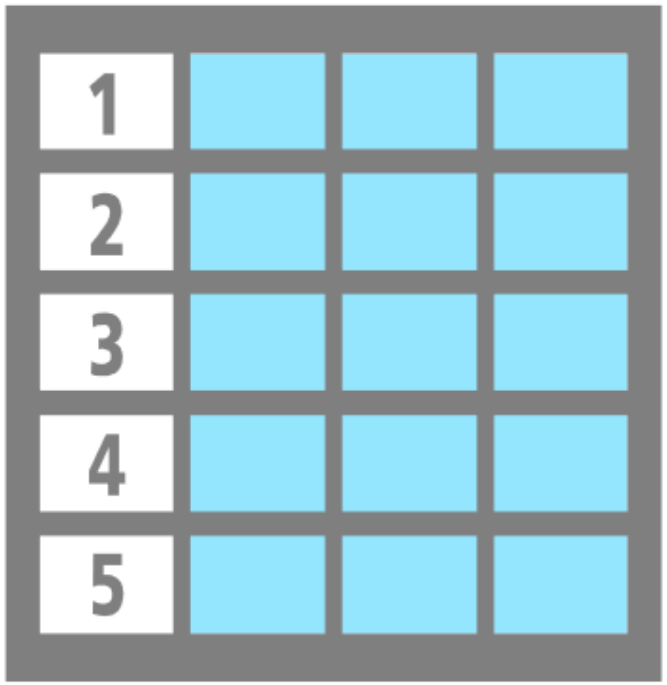


A	B	C	D	E	F	G	H	
4131 021 030916		4131	21	1567 CARS	3/9/2016	Coldstream Homestead Mont	Notheaste	14 (39.3222
4139 026 030916		4139	26	2738 HARF	3/9/2016	Coldstream Homestead Mont	Notheaste	14 (39.3212
0244 034 031418		244	34	1708 W PF	3/14/2018	Union Square	Southern	9 (39.2854
1125 027 031418		1125	27	1526 N CA	3/14/2018	Oliver	Eastern	12 (39.3080
1127 051 031018		1127	51	1516 N BR	3/10/2018	Oliver	Eastern	12 (39.3078
1138 114 031418		1138	114	1424 N BE	3/14/2018	Oliver	Eastern	12 (39.3069
1145 032 031418		1145	32	1336 AISQ	3/14/2018	Oliver	Eastern	12 (39.3058
1149B010 031418	1149B		10	1308 N CA	3/14/2018	Oliver	Eastern	12 (39.3054
7056A018 030816	7056A		18	3438 6TH S	3/8/2016	BROOKLYN	SOUTHERN	10 (39.2395
1488 010 031018		1488	10	2718 E OLI	3/10/2018	Berea	Eastern	13 (39.3082
1499 030 031418		1499	30	2058 E HO	3/14/2018	Broadway East	Eastern	13 (39.3068
1504 014 031518		1504	14	2407 E OLI	3/15/2018	Broadway East	Eastern	13 (39.3077
1505 055 031018		1505	55	1417 N MI	3/10/2018	Berea	Eastern	13 (39.3075
1521 037 031018			37	2525 E HO	3/10/2018	Berea	Eastern	13 (39.3067
1542 076 031218		1542	76	1224 N PO	3/12/2018	Berea	Eastern	13 (39.3056
1544 028 031018		1544	28	3113 E PRI	3/10/2018	Berea	Eastern	13 (39.3059
1661 014 031318		1661	14	3027 MCE	3/13/2018	Ellwood Park/Monument	Southeast	13 (39.2978
0007 053 122915		7	53	1726 APPL	12/29/2015	EASTERWOOD	WESTERN	7 (39.3084
1667 014 031318		1667	14	429 N WAS	3/13/2018	CARE	Eastern	13 (39.2961
1655 083 031318		1655	83	515 N POR	3/13/2018	McElderry Park	Southeast	13 (39.2972
1656 063 031318		1656	63	512 N ROS	3/13/2018	McElderry Park	Southeast	13 (39.2972
1657 062 031418		1657	62	514 N GLO	3/14/2018	McElderry Park	Southeast	13 (39.2973
1658 074 031318		1658	74	531 N BEL	3/13/2018	McElderry Park	Southeast	13 (39.2977
1659 002 031318		1659	2	503 N KEN	3/13/2018	McElderry Park	Southeast	13 (39.2970
4581 010F 030918		4581	010F	3223 SPAU	3/9/2018	Central Park Heights	Northwest	6 (39.3472
4609 055 031418		4609	55	3701 MAN	3/14/2018	Central Park Heights	Northwest	6 (39.3430
4616 005 031318		4616	5	3008 WOC	3/13/2018	Central Park Heights	Northwest	6 (39.3465
5902A004 031518	5902A		4	3307 LAKE	3/15/2018	Belair-Edison	Notheaste	13 (39.3208
5907 031 031518		5907	31	3044 CHES	3/15/2018	Belair-Edison	Notheaste	13 (39.3239
5929 061 031518		5929	61	4340 SHAM	3/15/2018	Belair-Edison	Notheaste	2 (39.3262
0006 032 011509		6	32	2002 PRES	1/15/2009	EASTERWOOD	WESTERN	7 (39.3080
0031 057 032316		31	57	1834 LAUF	3/23/2016	SANDTOWN-WINCHESTER	WESTERN	7 (39.3032
0055C063 032316	0055C		63	1104 N CA	3/23/2016	SANDTOWN-WINCHESTER	WESTERN	9 (39.3013
0073 024 031616		73	24	930 N MO	3/16/2016	SANDTOWN-WINCHESTER	WESTERN	9 (39.2998
0075 016 032416		75	16	1515 MOS	3/24/2016	SANDTOWN-WINCHESTER	WESTERN	9 (39.2997
0082 058 032416		82	58	805 N BRI	3/24/2016	MIDTOWN-EDMONDSON	WESTERN	9 (39.2978
0096 052 031516		96	52	707 N MO	3/15/2016	HARLEM PARK	WESTERN	9 (39.2968
1109 040 031018		1109	40	1704 N BR	3/10/2018	Oliver	Eastern	12 (39.3097
0149 044 031616		149	44	310 N BRU	3/16/2016	FRANKLIN SQUARE	WESTERN	9 (39.2923
0244 074 031516		244	74	1708 LEMI	3/15/2016	UNION SQUARE	SOUTHERN	9 (39.2858
1114 009 031418		1114	9	1216 E FEI	3/14/2018	Oliver	Eastern	12 (39.3084

When you look at a spreadsheet...



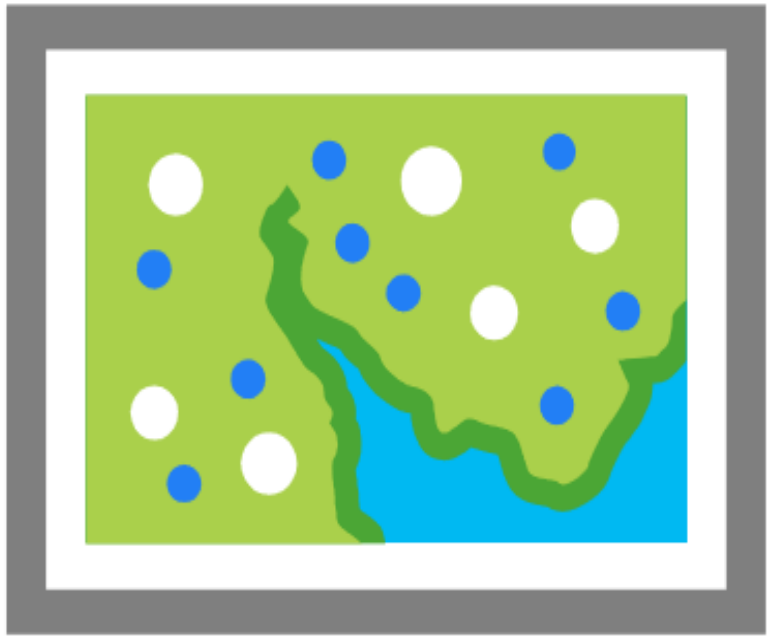
You ask for more



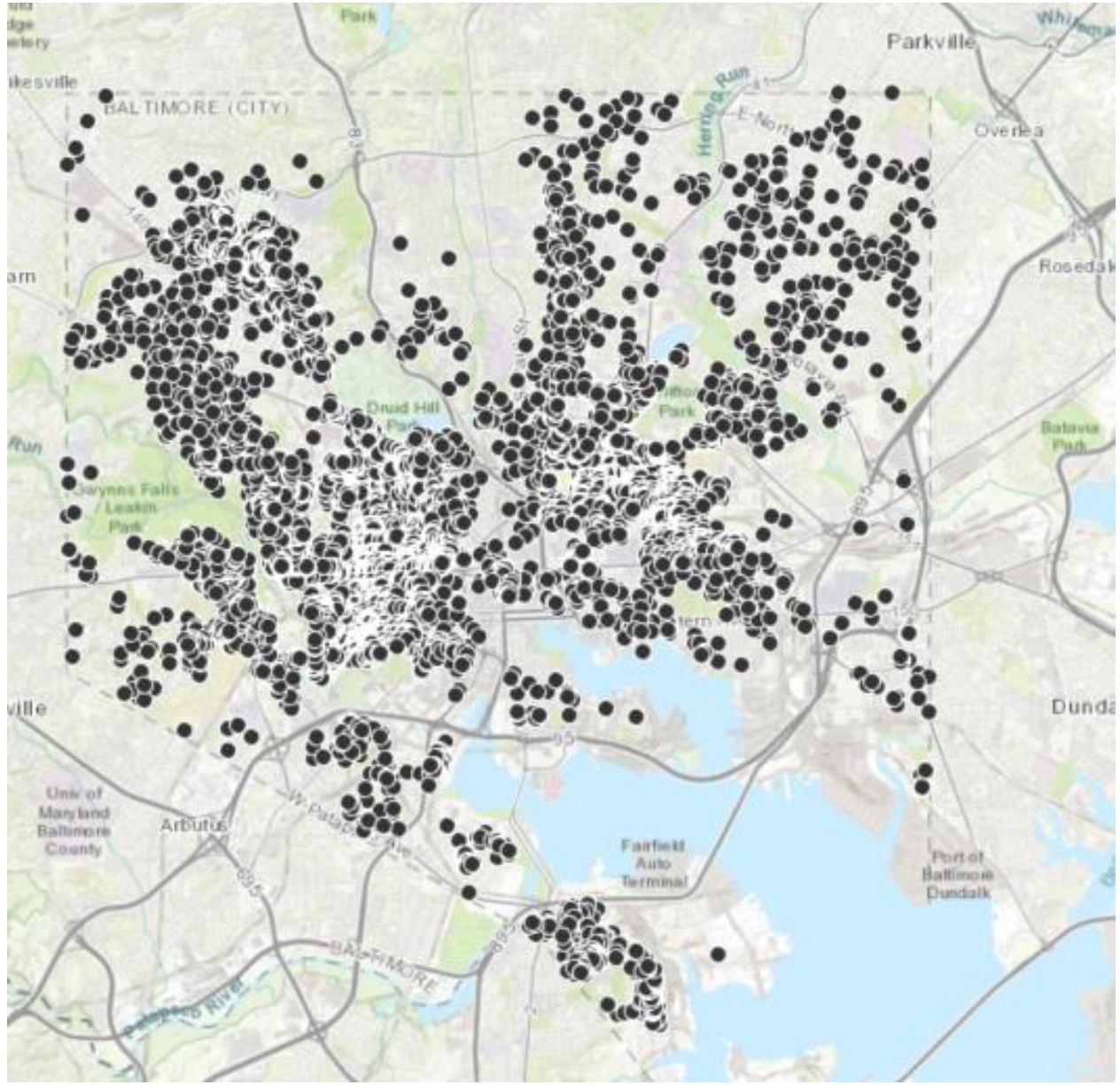
- Mean
- Standard Deviations
- Min and Max
- …

Maps

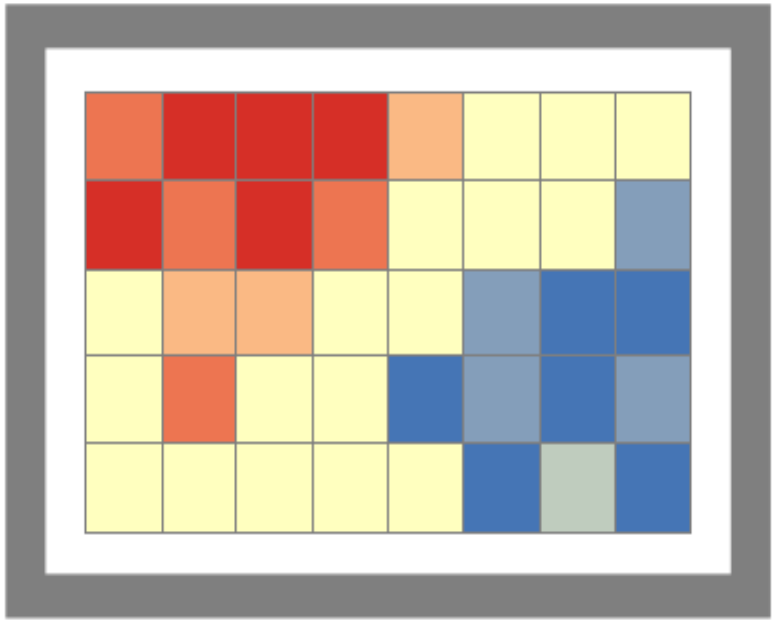
Data or Information?

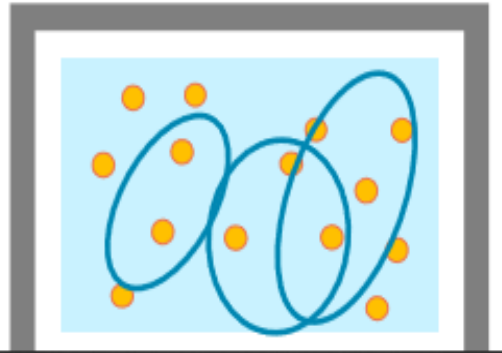


Same goes for maps!


Parkville
BALTIMORE (CITY)
Overlea
Druid Hill Park
Arbutus
Fairfield Auto Terminal
Port of Baltimore Dundalk
Univ of Maryland Baltimore County
Patapsco River

We can do more





Reconnecting Milwaukee

A BikeAble™ Study of Opportunity, Equity and Connectivity

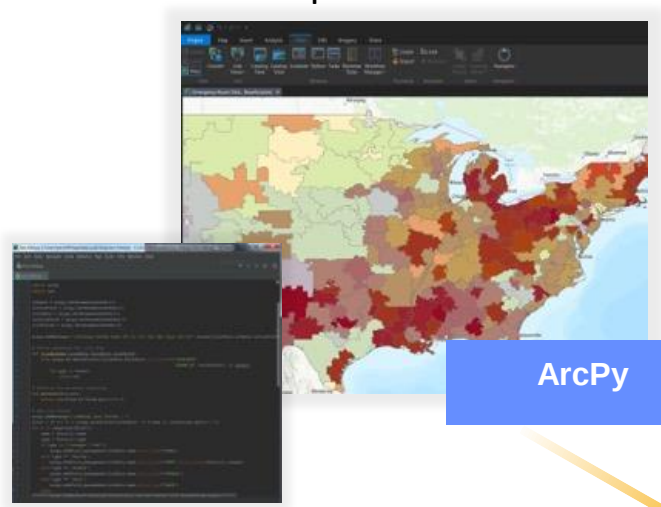
ArcGIS Connection with Open Source Libraries

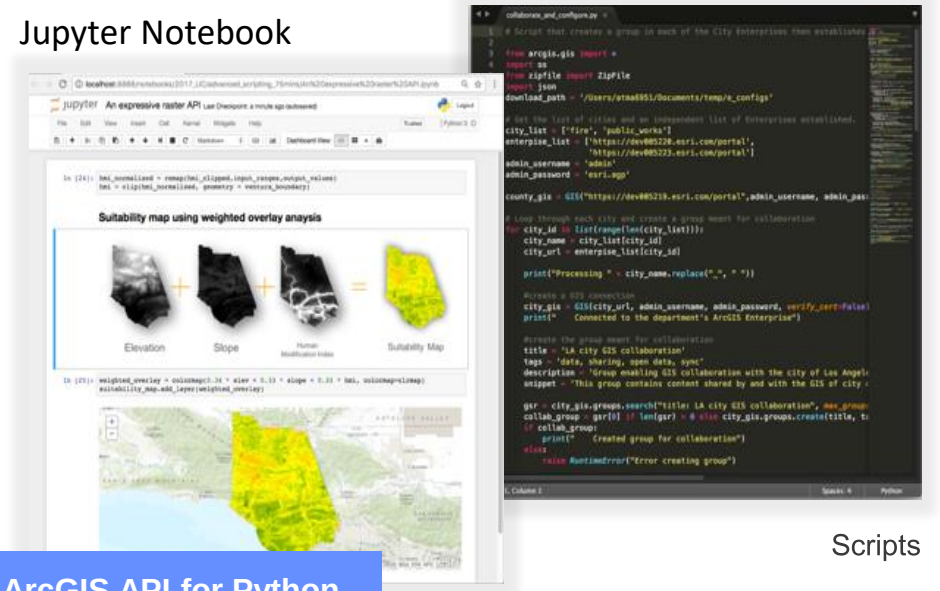
Navigating the Python Ecosystem

Enterprise & Online

Jupyter Notebook

Desktop





ArcPy

ArcGIS API for Python

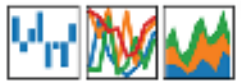
Scripts

PYTORCH

CONDA

SciPy

pandas



TensorFlow

K Keras

matplotlib

dmlc XGBoost

Sample Use Case:

Developing a Weather Repository

- R: A widely used statistical programming language
- More than 10,000 Packages
- Expand Workflows

Sample Use Cases:

Predicting Seagrass Locations

- **Ingesting / Analyzing Multidimensional Scientific Data**
- **Modeling Vacancy Rates in DC -- from Baltimore Vacancy Data**

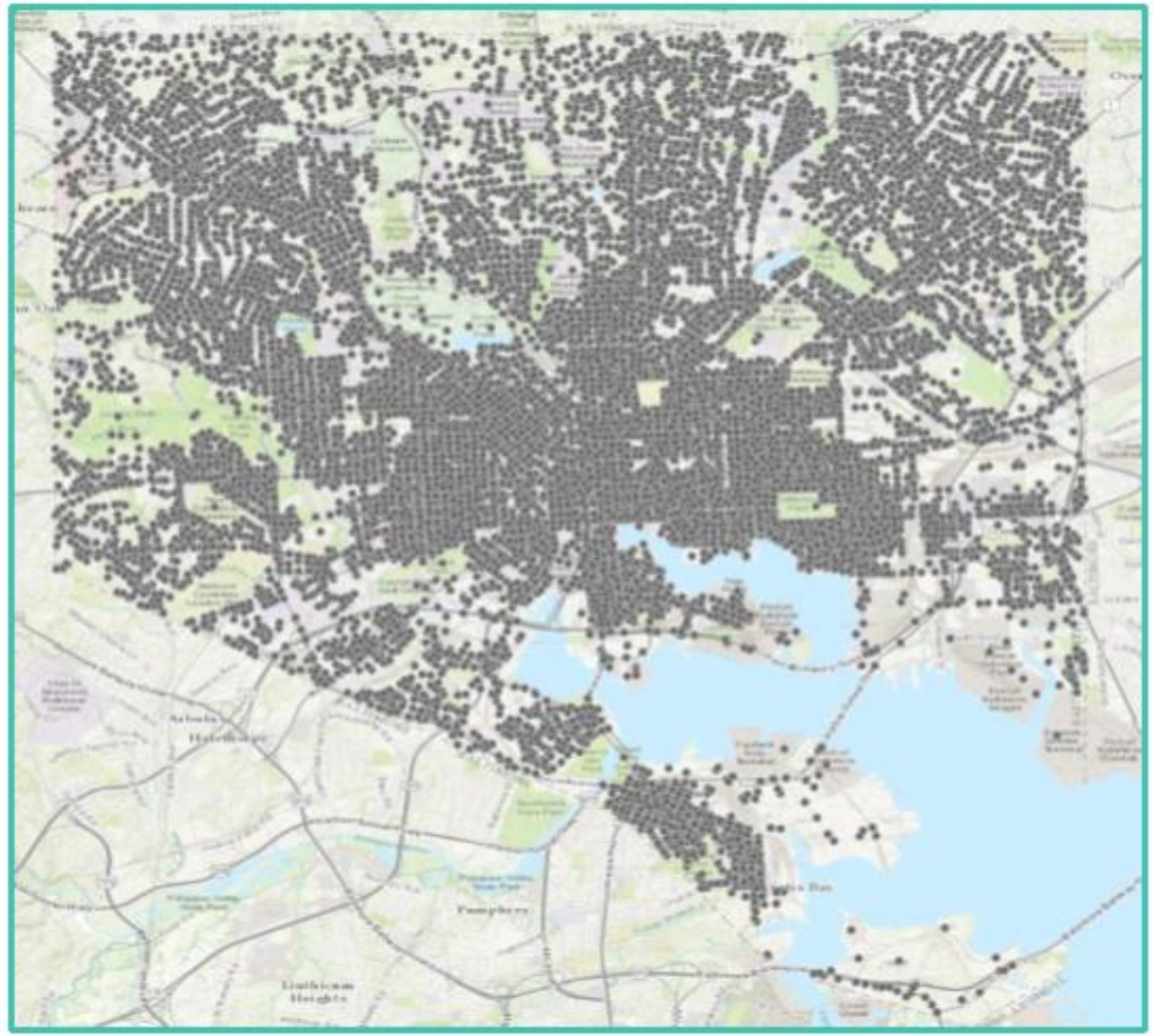
Clustering, Prediction & Classification

Kristen Hocutt

Subjectivity of Maps

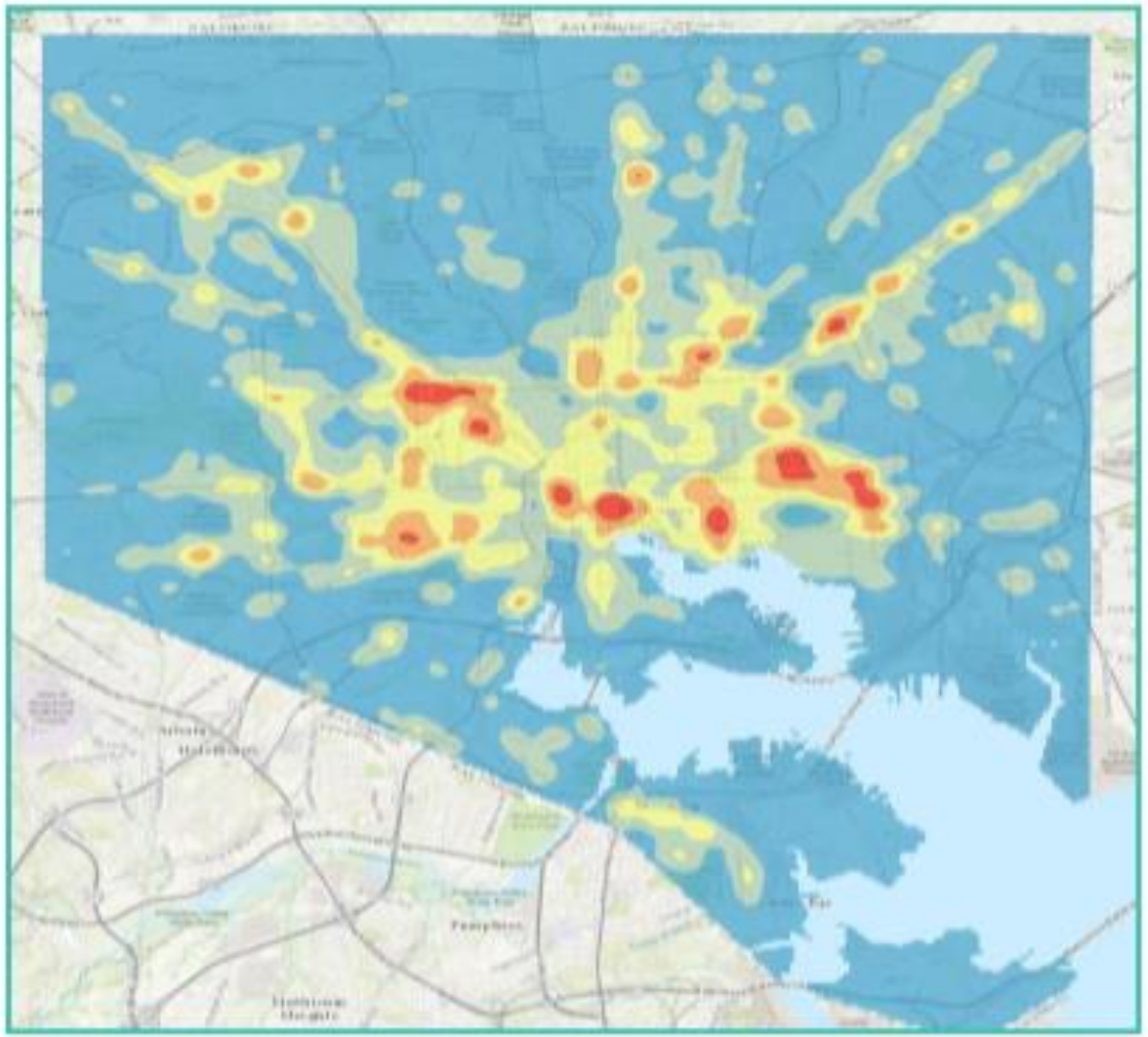
The map as data

High Priority 911 Calls in Baltimore



The map as data

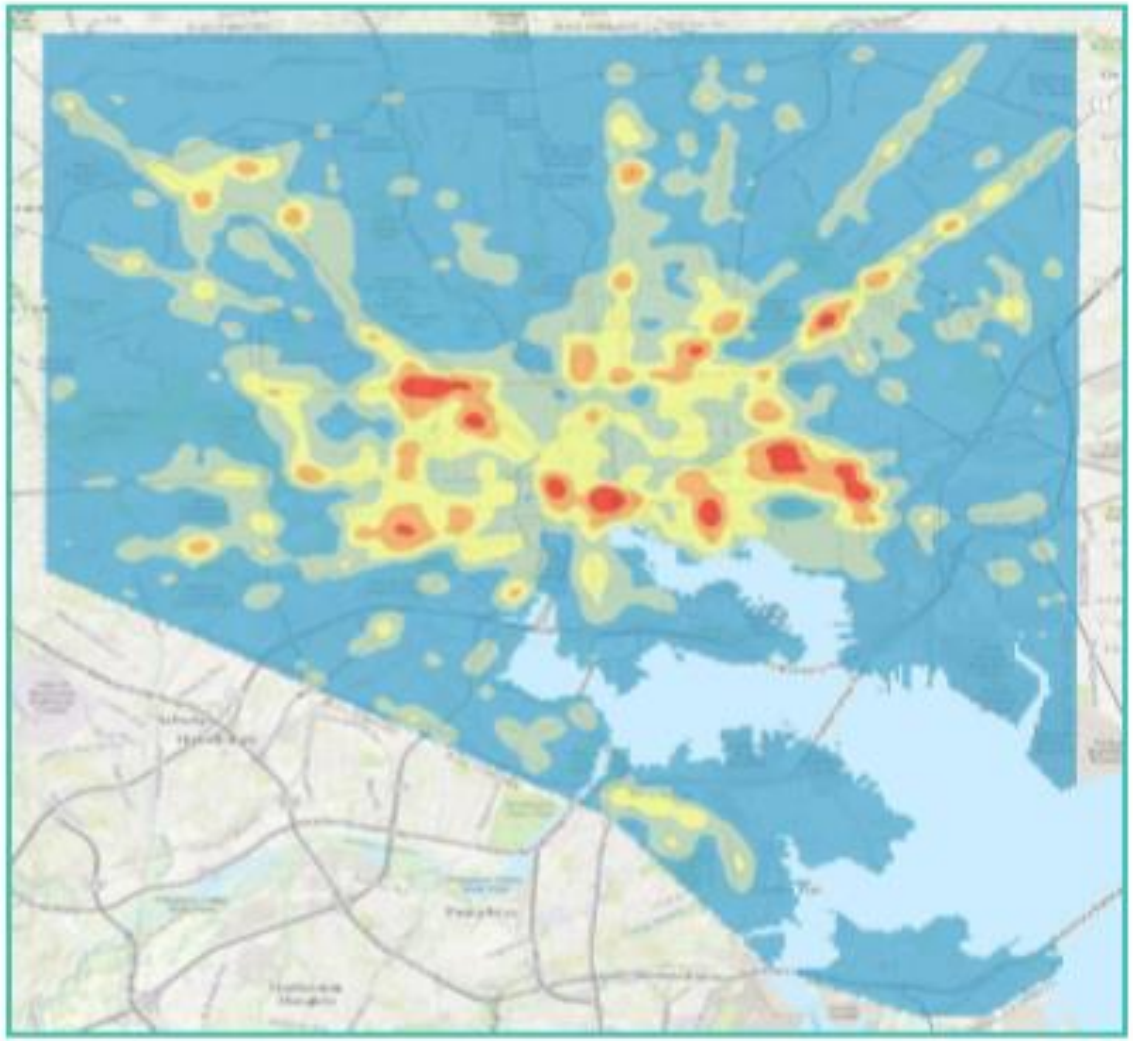
High Priority 911 Calls in Baltimore

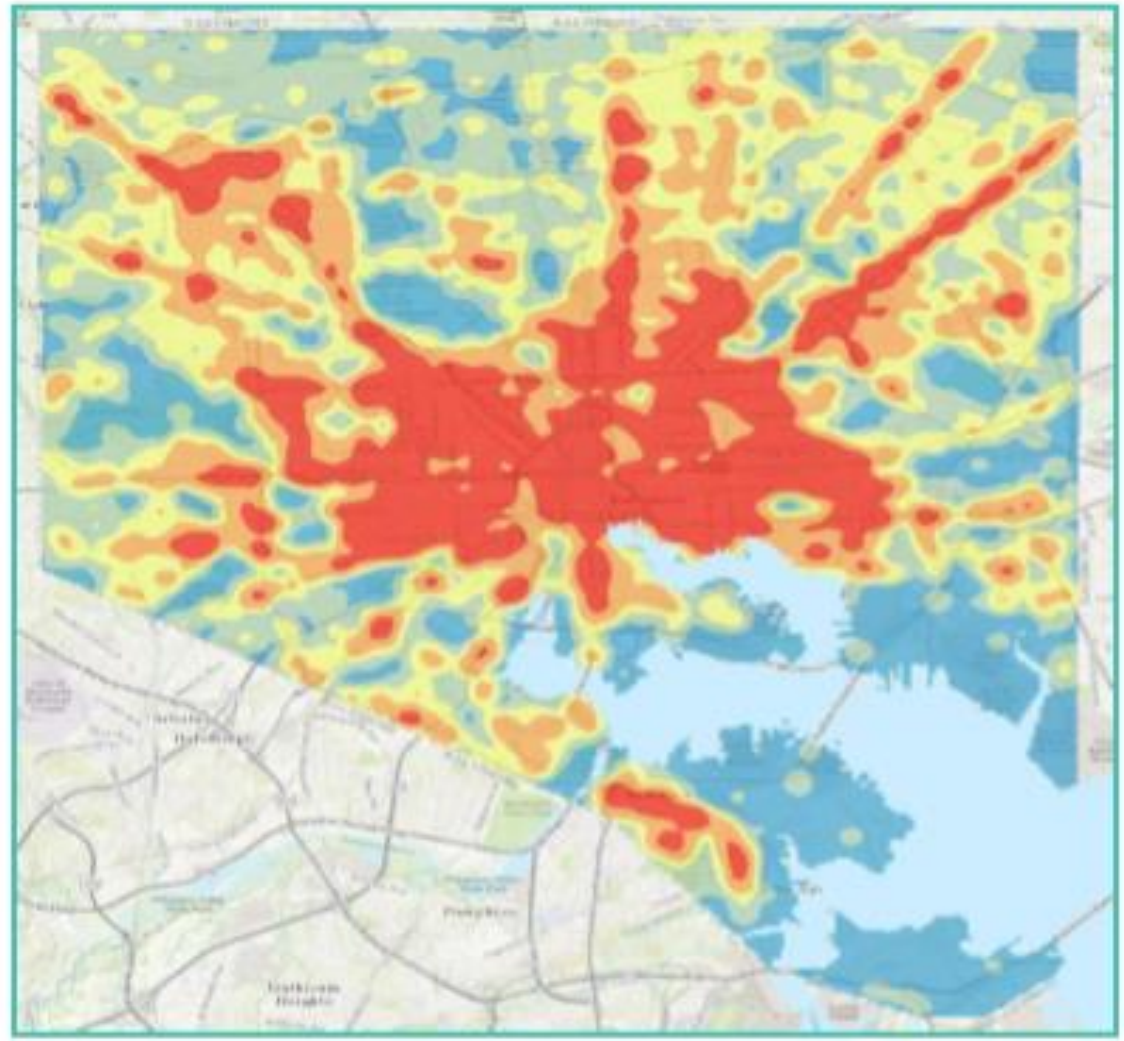


Where are the hot spots? Where is the variation greater?

The map as data

High Priority 911 Calls in Baltimore

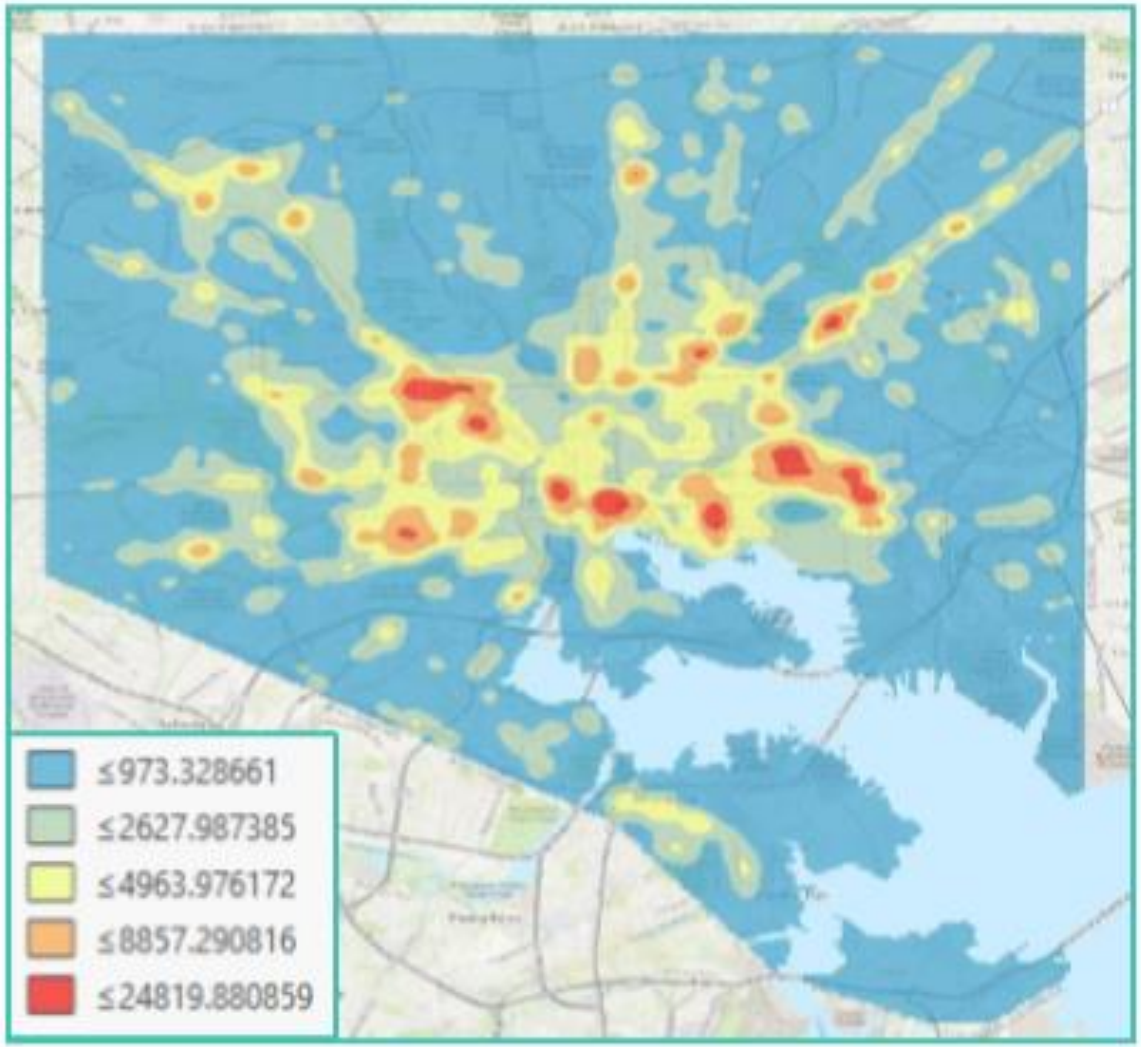




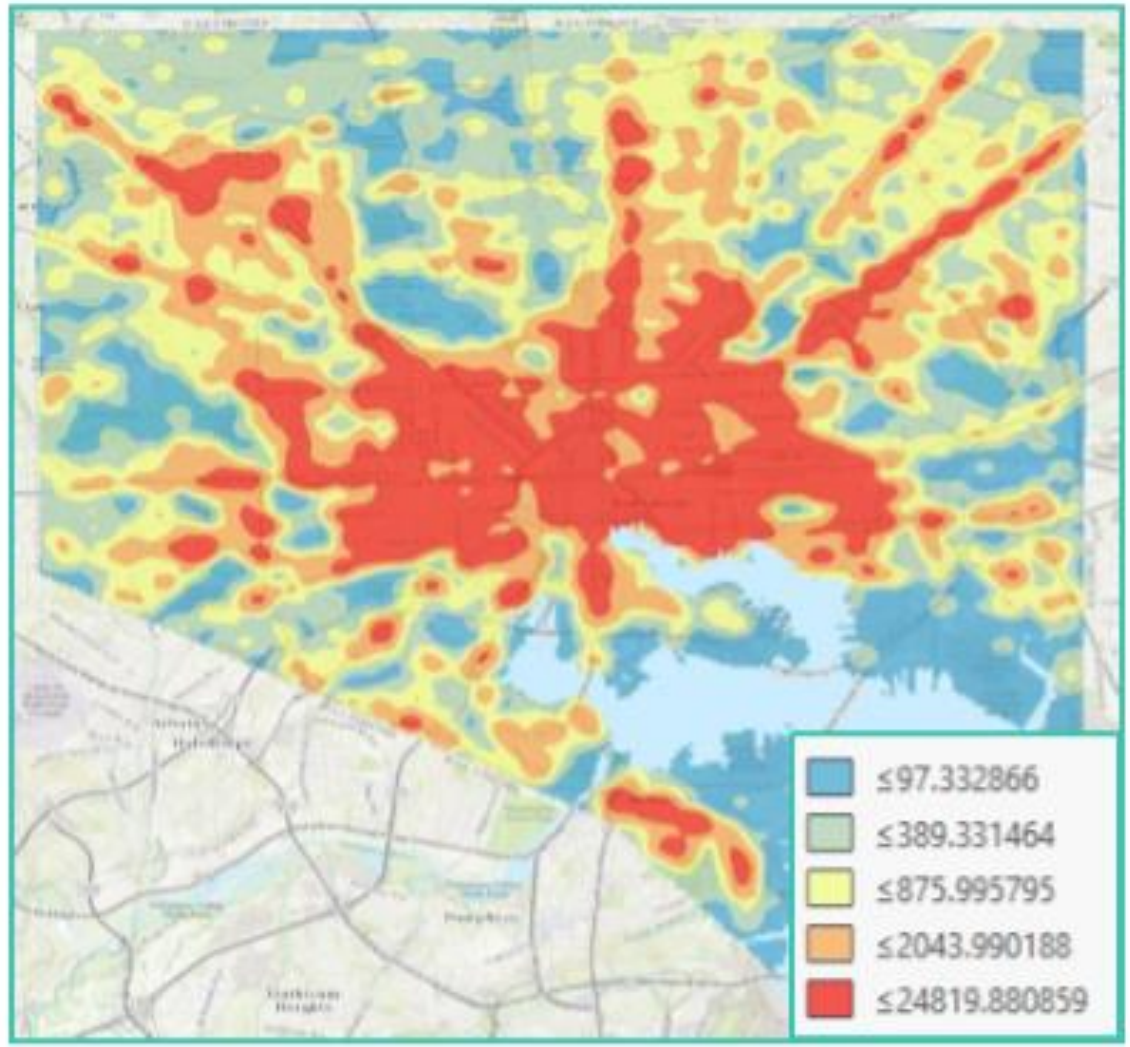
Where are the hot spots? Where is the variation greater?

The subjectivity of visual pattern analysis

Natural Breaks



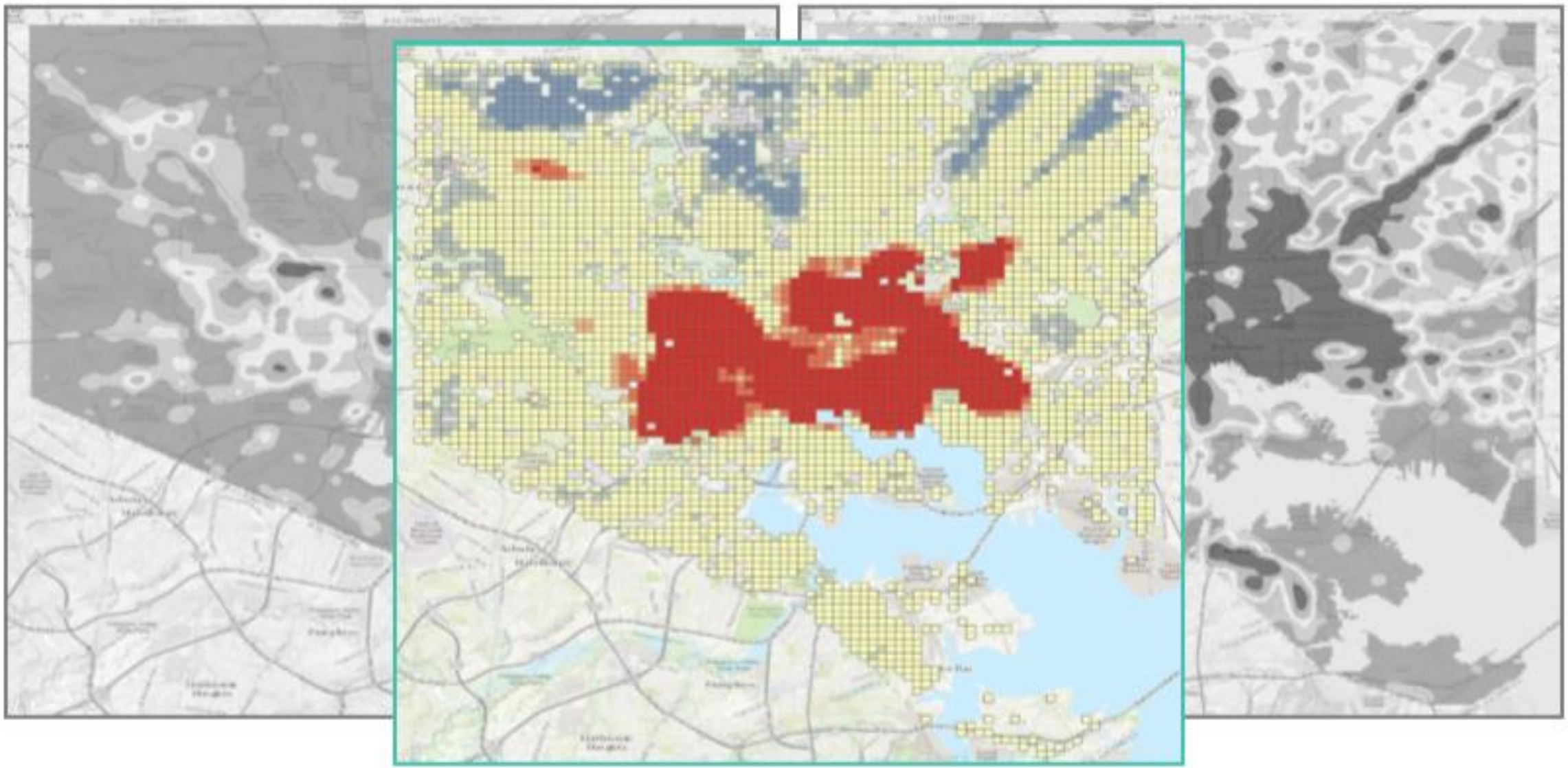
Quantile



Where are the hot spots? Where is the variation greater?

Minimizing the subjectivity

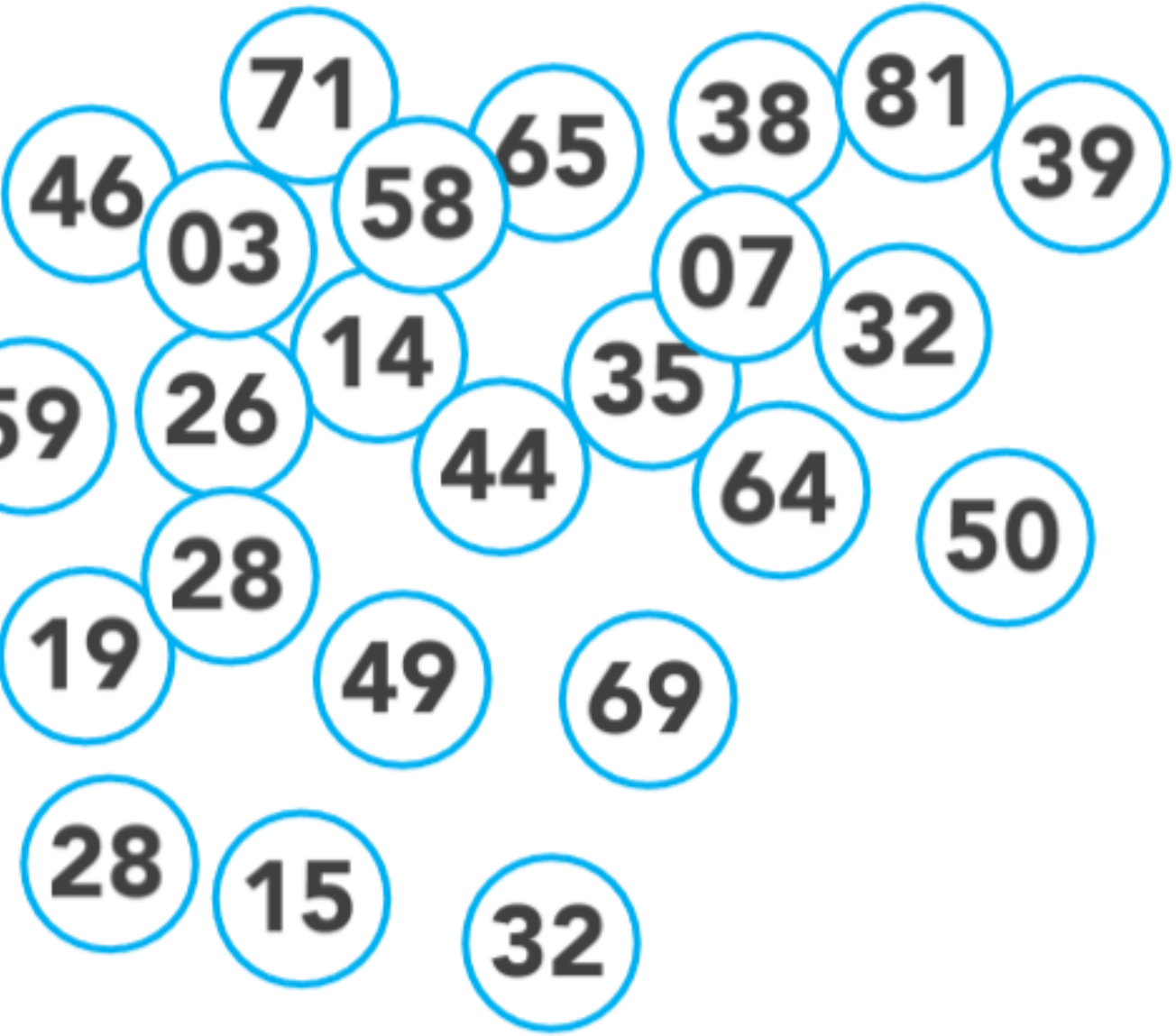
Turning the map into information

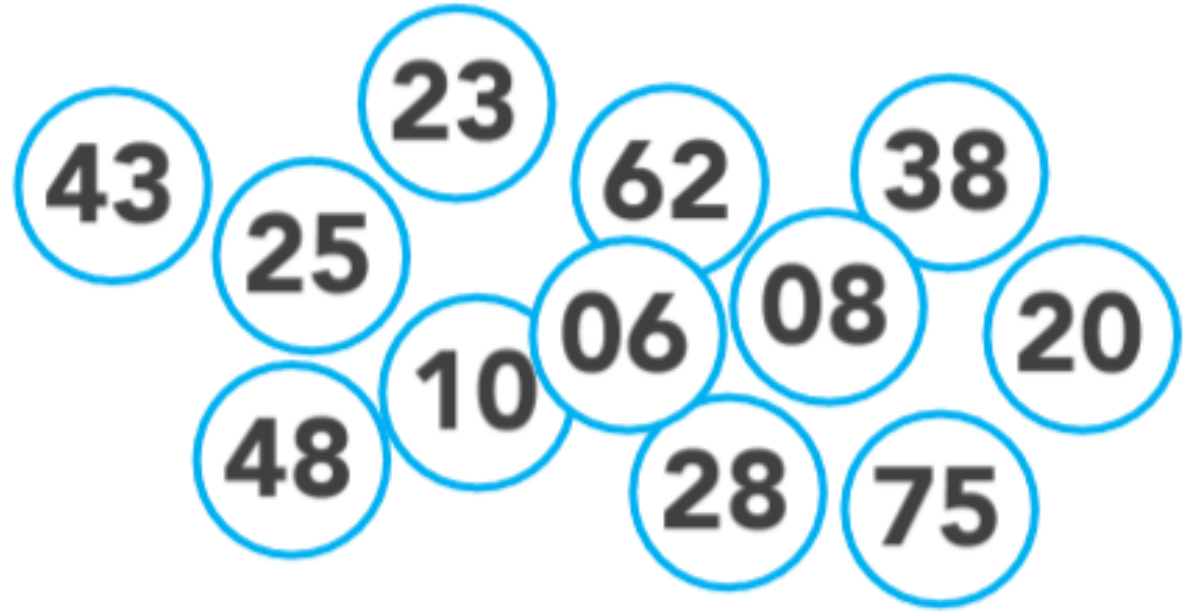



Complete Spatial
RANDOMNESS

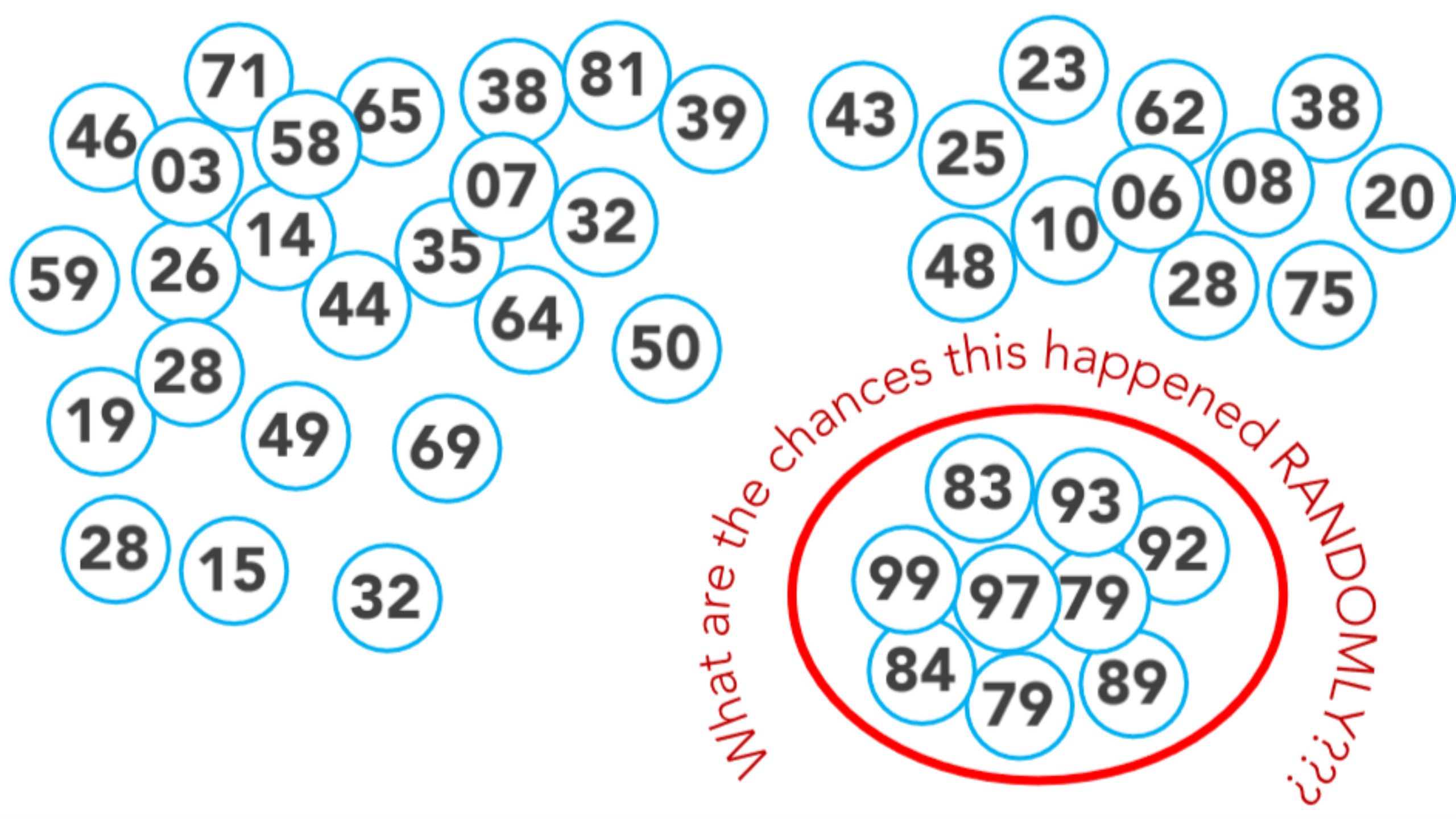
Is there a PATTERN?


14 15 92 65 35 89 79 32 38
46 26 43 38 32 79 50 28 84
19 71 69 39 93 75 10 58 20
97 49 44 59 23 07 81 64 06
28 62 08 99 86 28 03 48 25


71
65
38
81
46
58
39
03
07
14
32
35
59
26
44
64
50
28
19
49
69
28
15
32

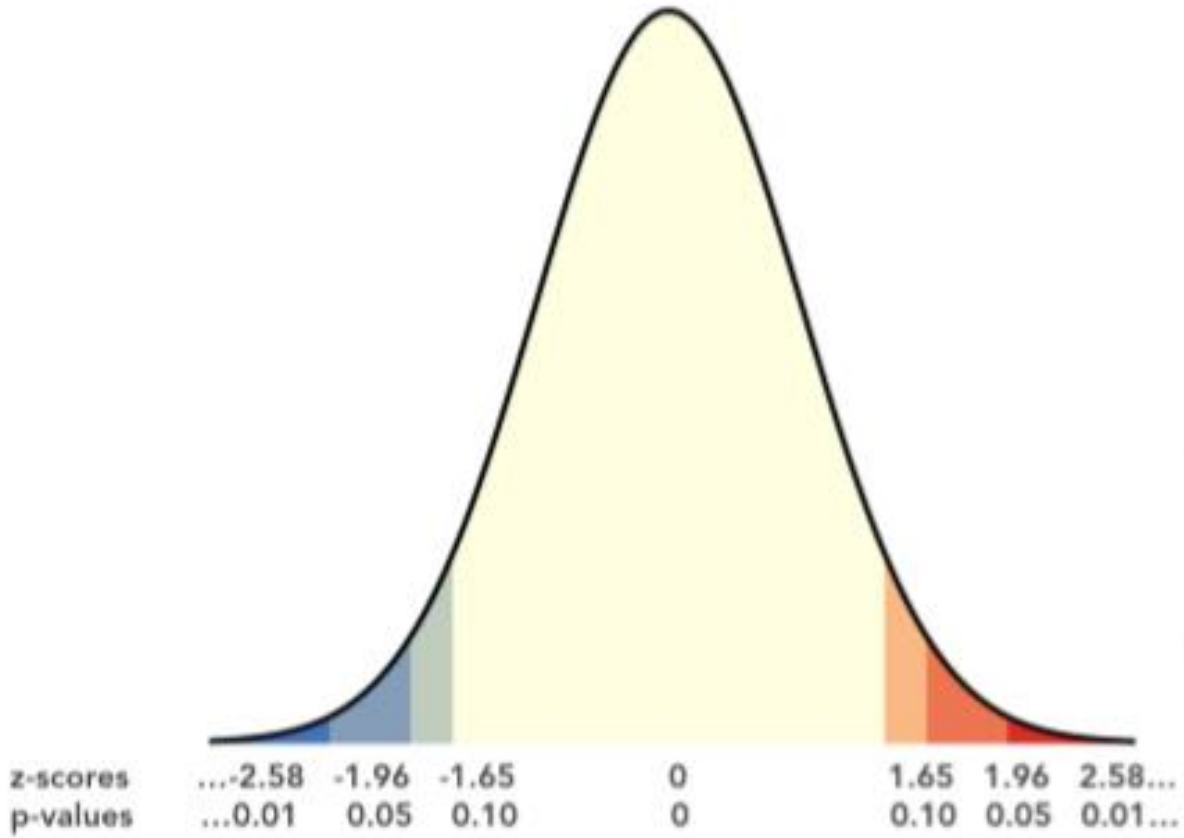

23
43
62
38
25
06
08
20
10
48
28
75

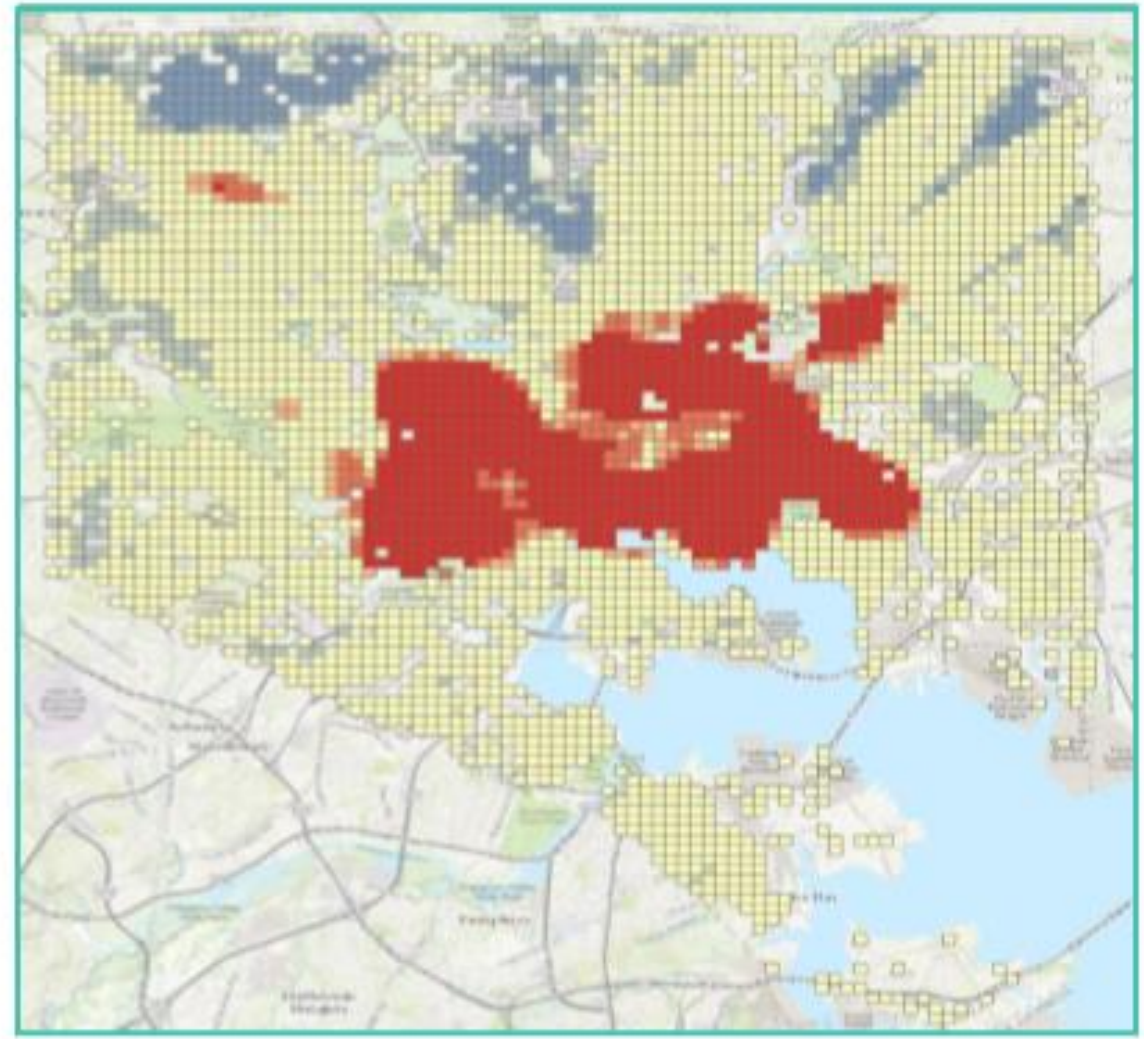

83
93
92
99
97
79
84
79
89


71
46
03
58
65
38
81
39
07
32
14
35
59
26
44
64
50
28
19
49
69
28
15
32
43
23
25
62
38
06
08
20
10
48
28
75
What are the chances this happened RANDOMLY???
83
93
92
99
97
79
84
79
89

z-scores

p-values


z-scores ...-2.58 -1.96 -1.65 0 1.65 1.96 2.58...
p-values ...0.01 0.05 0.10 0 0.10 0.05 0.01...



Hot Spot Analysis

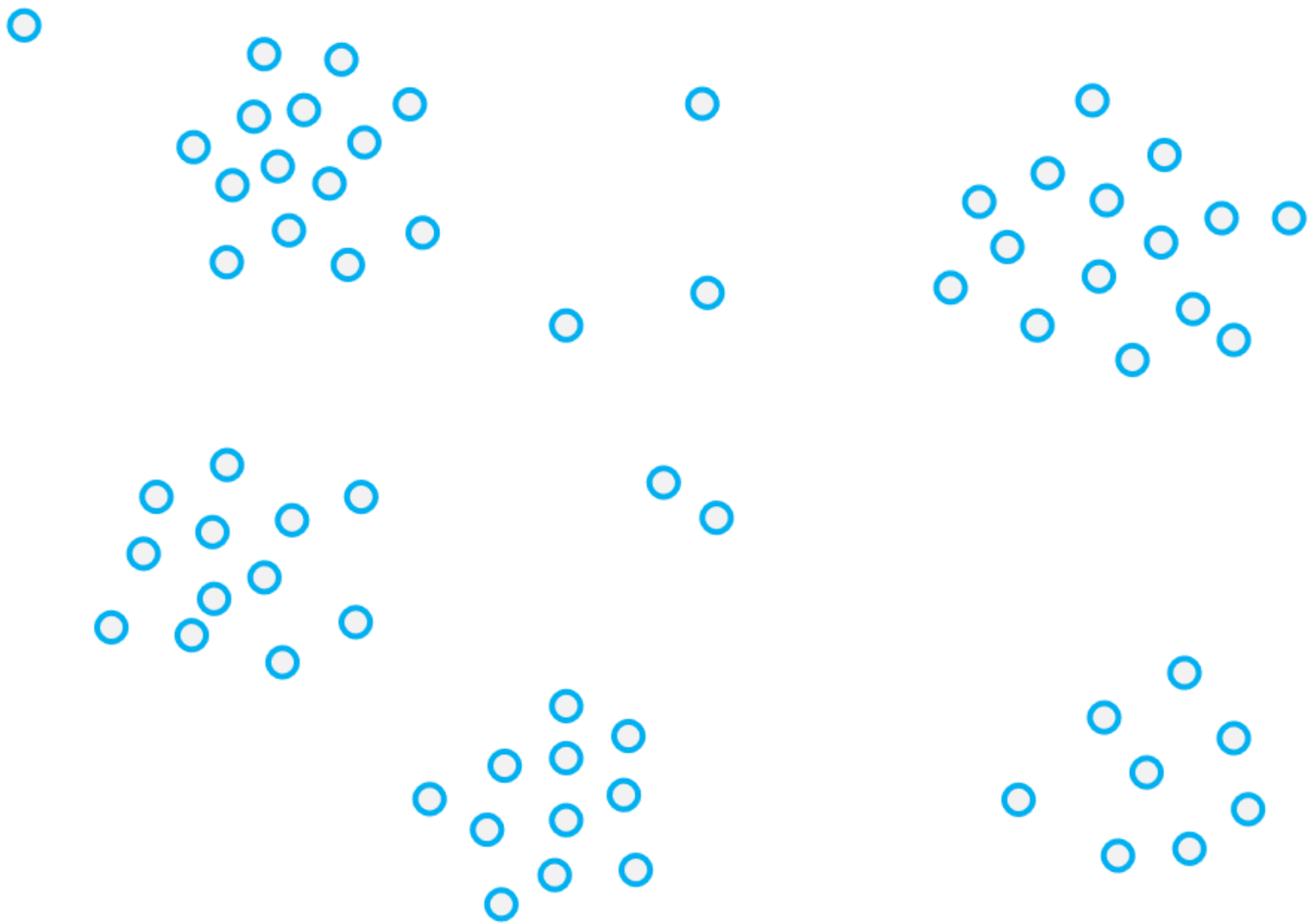
given a set of weighted features, identifies statistically significant hot spots and cold spots using the Getis-Ord Gi* statistic

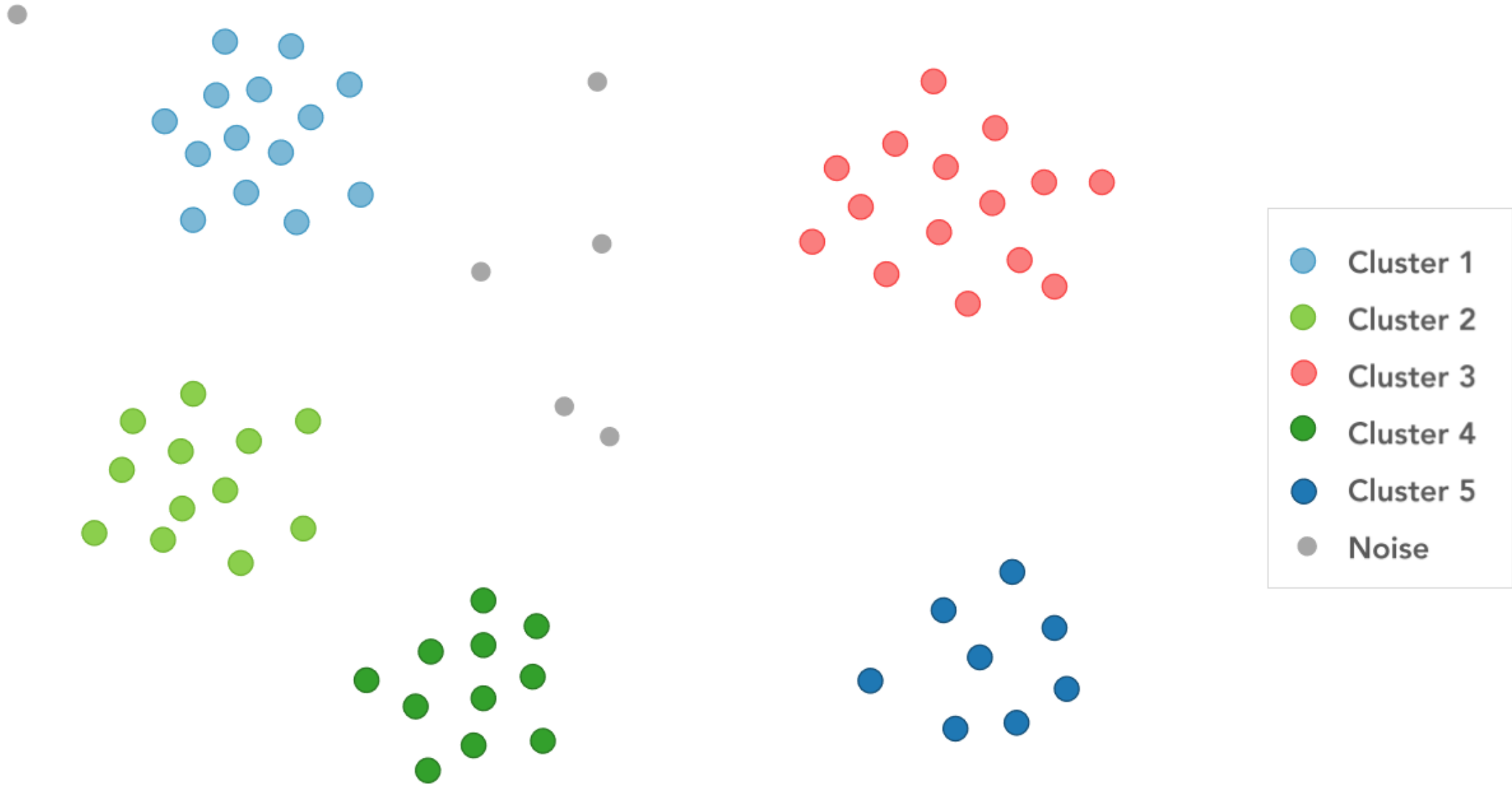
…how do we know if it's SIGNIFICANTLY different???

MATH!

Density-based Clustering

finds clusters based on feature locations




Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Noise

DBSCAN – defined distance

HDBSCAN – self adjusting

OPTICS – multi-scale

Spatial Analysis of Traffic Accidents

A Density Based Clustering Approach

Models

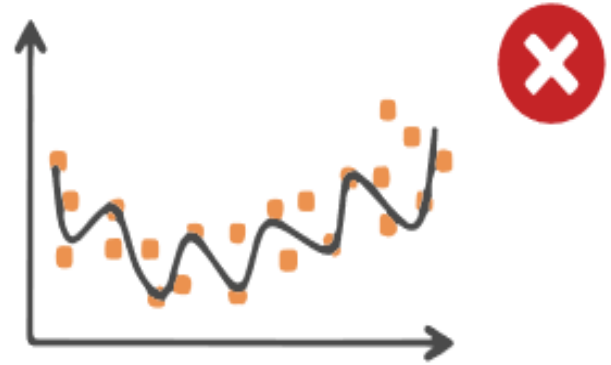
Representative generalizations used for prediction

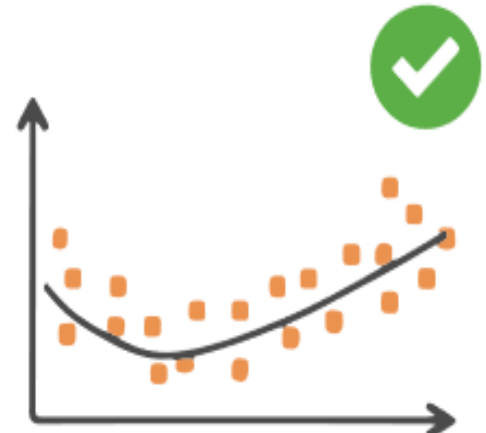
Why model

Use information we have to predict information we don't have

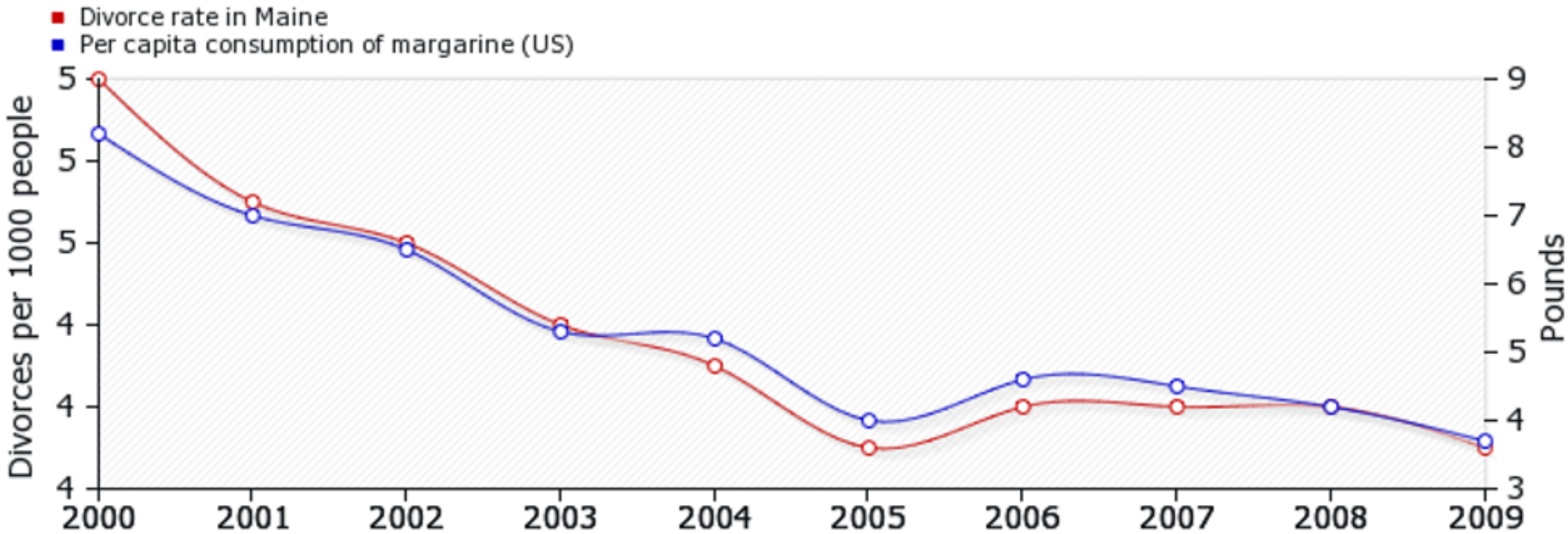
When we can't trust a model

Mimics training dataset and models noise instead of generalizing a trend





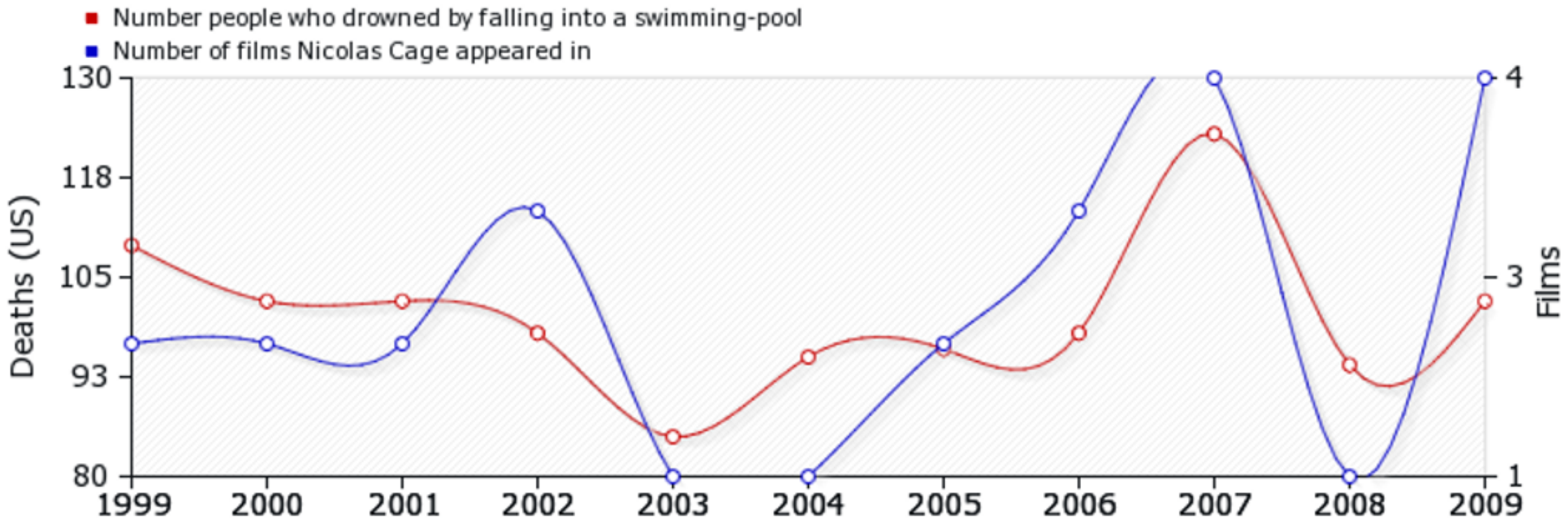
Divorce Rate in Maine vs Per Capita Consumption of Margarine



tylervigen.com

Correlation: 0.992558

Number People Who Drowned by Falling into a Swimming-Pool vs Number of Nicolas Cage Films



Correlation: 0.666004

tylervigen.com

Many ways to model

Generalized Linear Regression
Geographically Weighted Regression

Forest-based Classification and Regression

Generalized Linear Regression

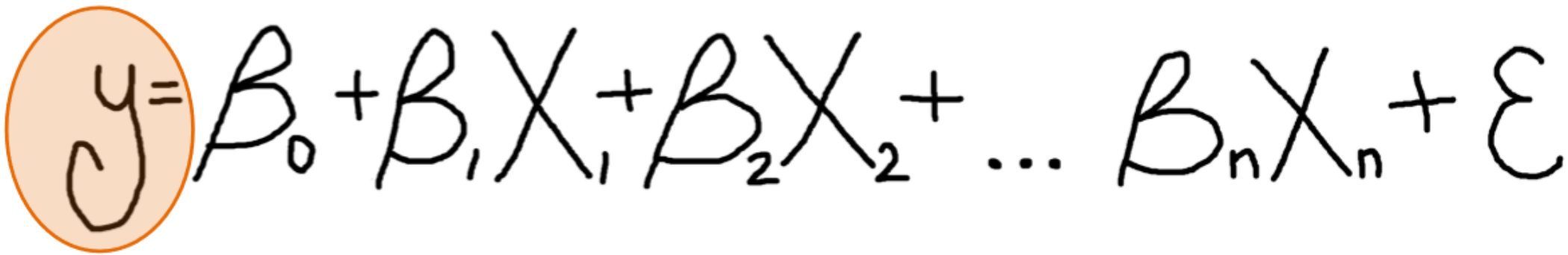
Modeling linear relationships

Three model types

- Gaussian – continuous
- Logistic – binary
- Poisson – count

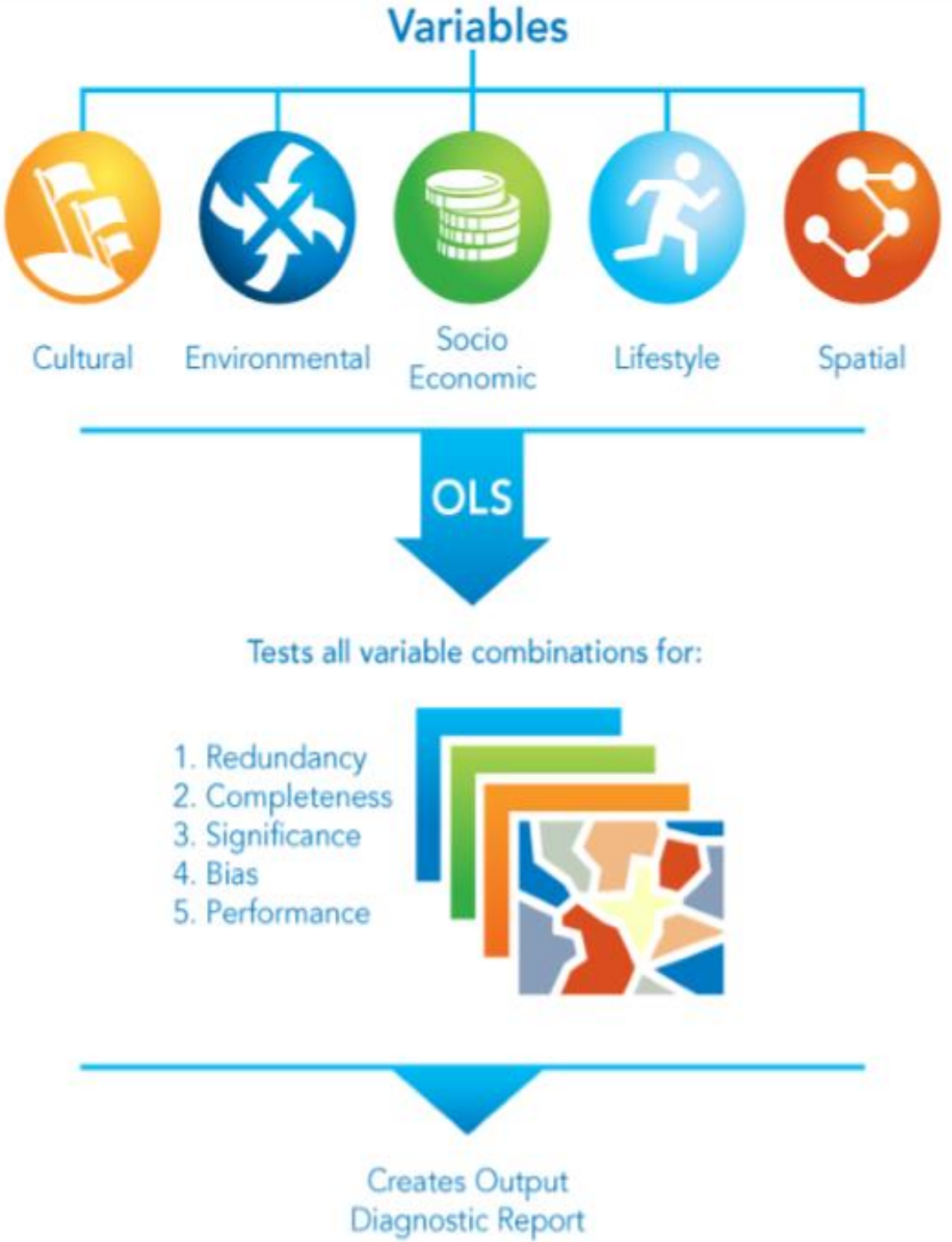
Gaussian
Ordinary Least Squares

Dependent Variable



What are you trying to predict or understand?

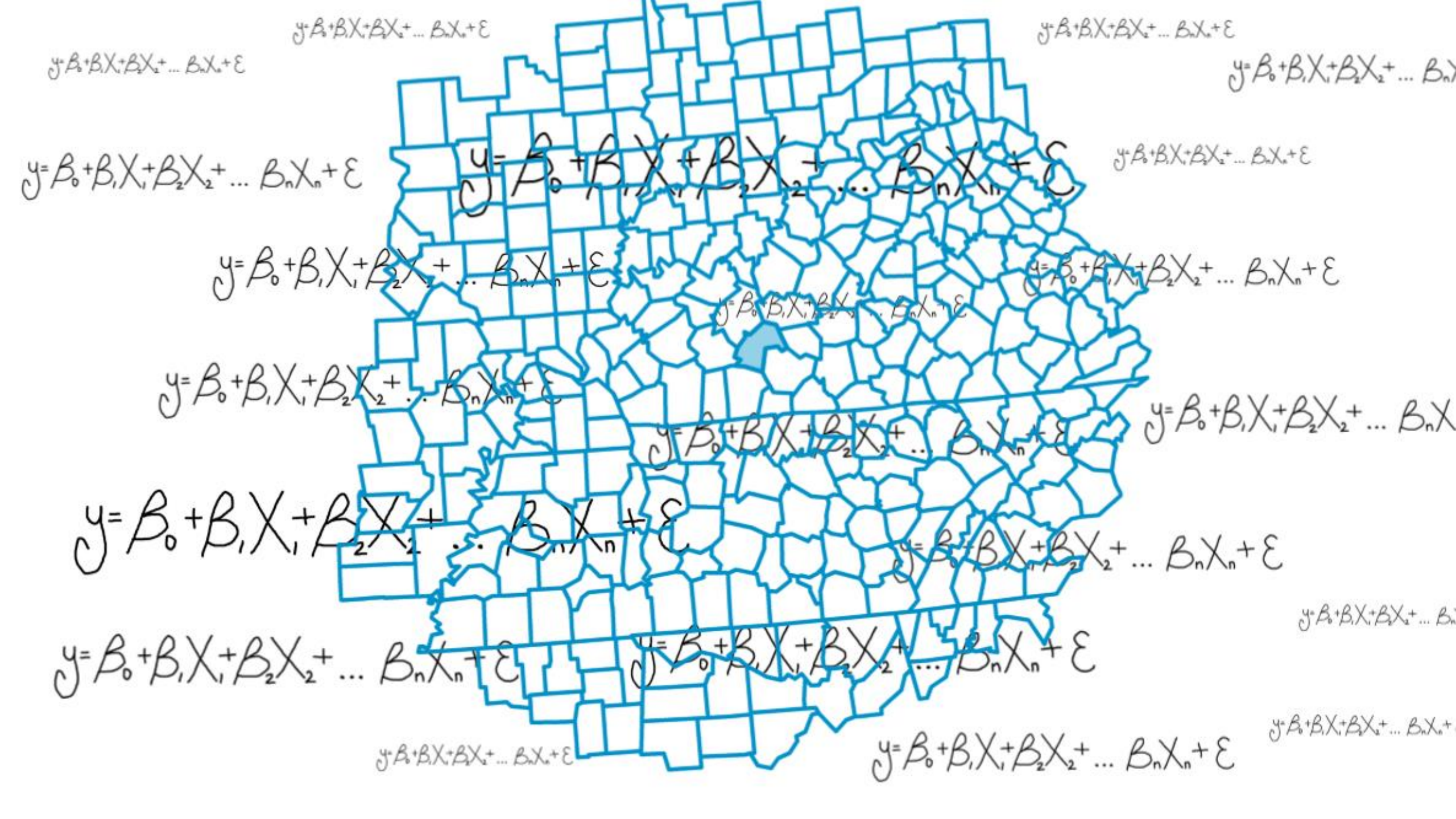
Exploratory Regression

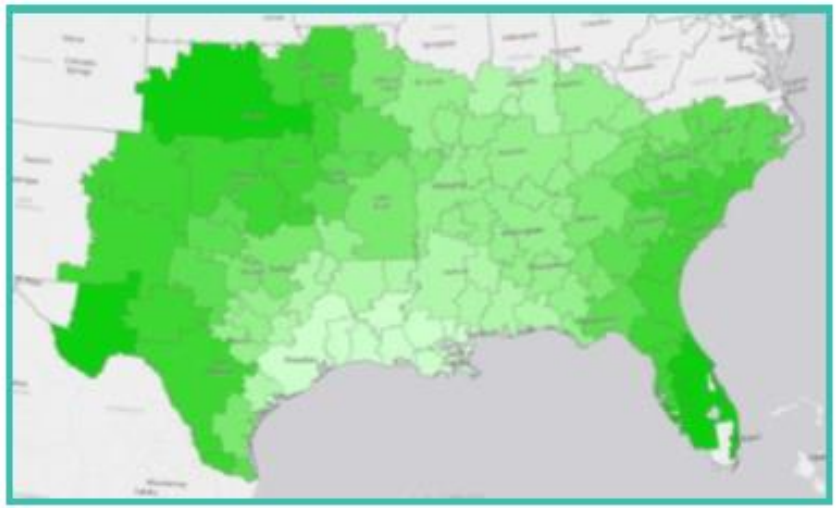


Geographically Weighted Regression

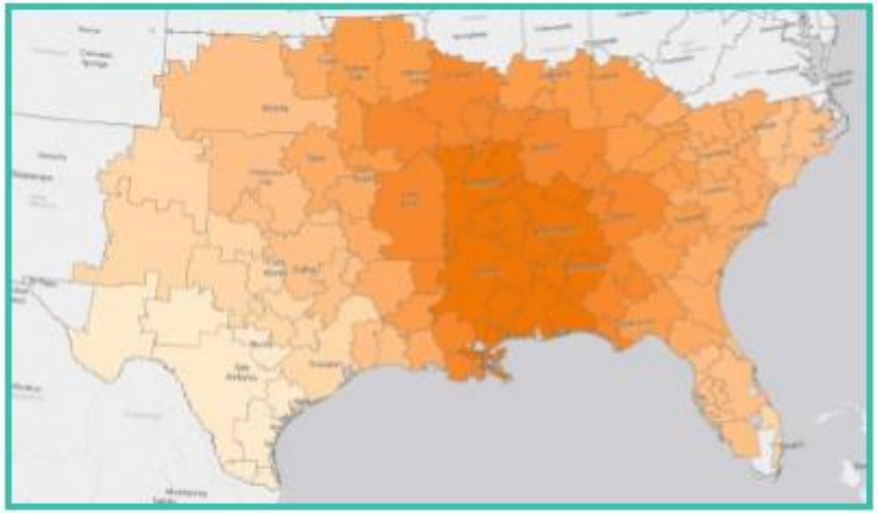
Exploring spatial variation

each feature gets a separate equation

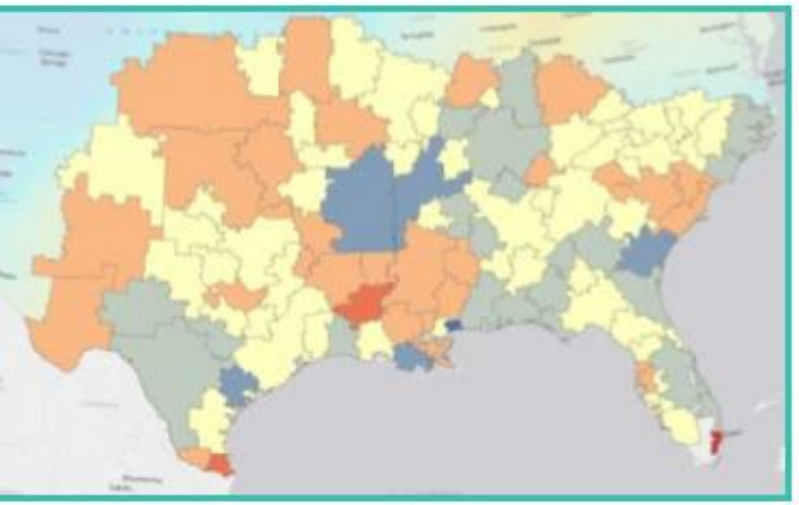

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$



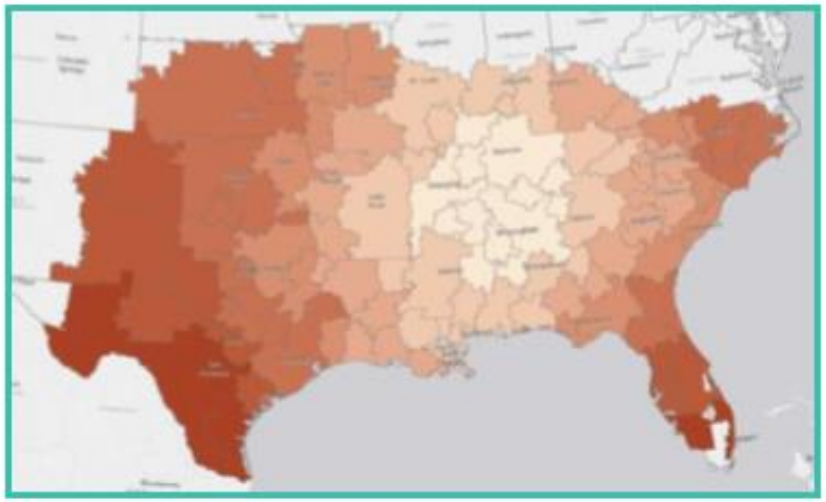
Local R-Squared



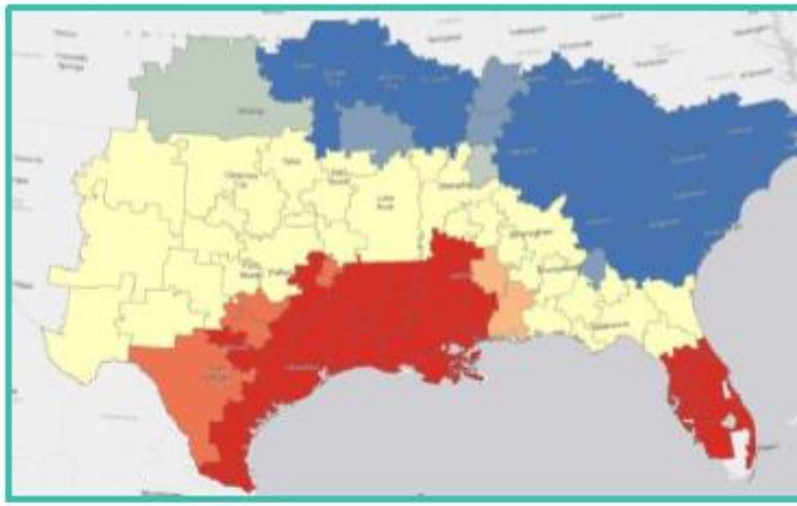
Coefficients



Residuals



Condition Number



Predictions

Forest-based Classification & Regression

Predicting using machine learning

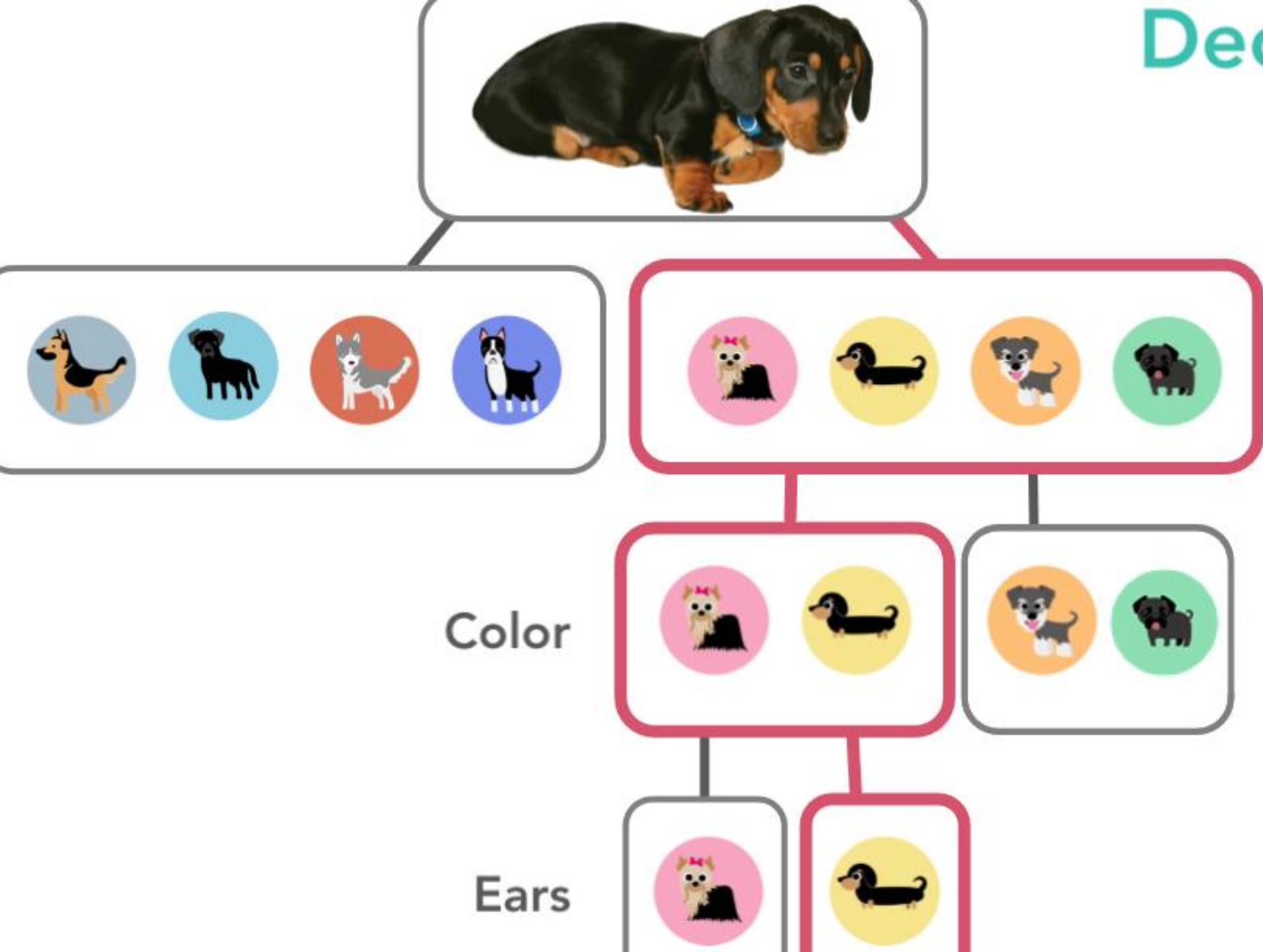
Training

variable to predict

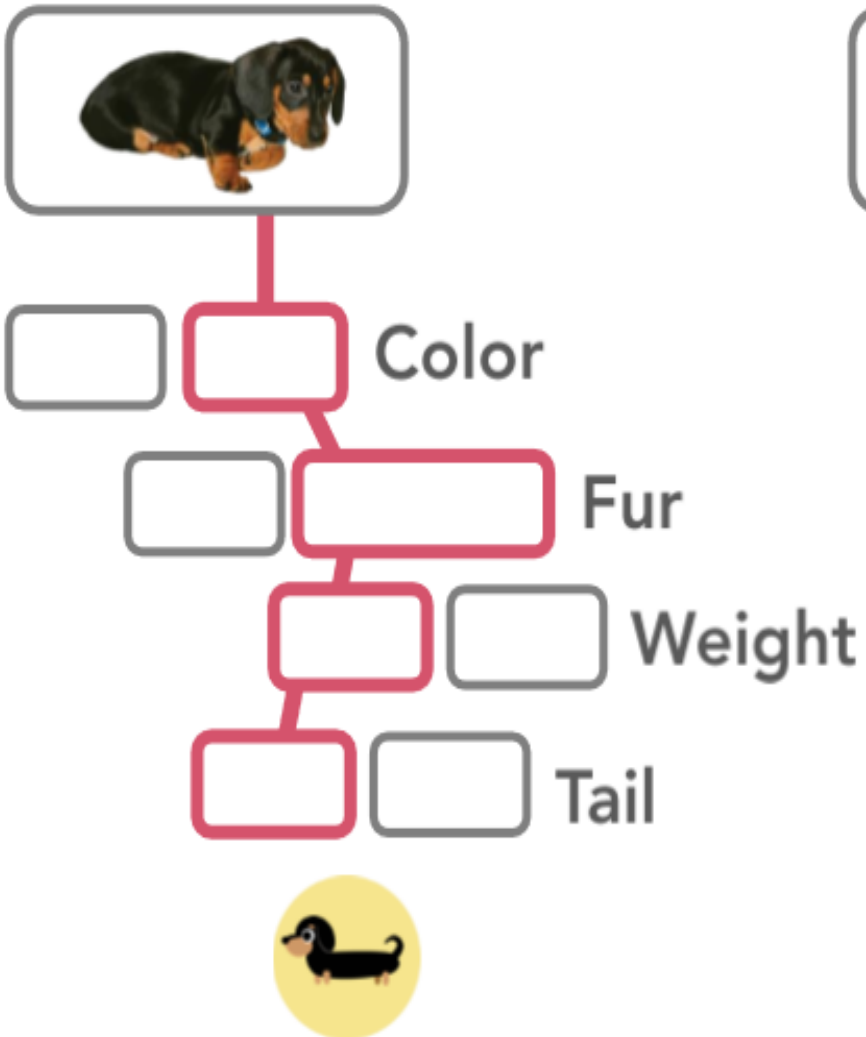
Breed

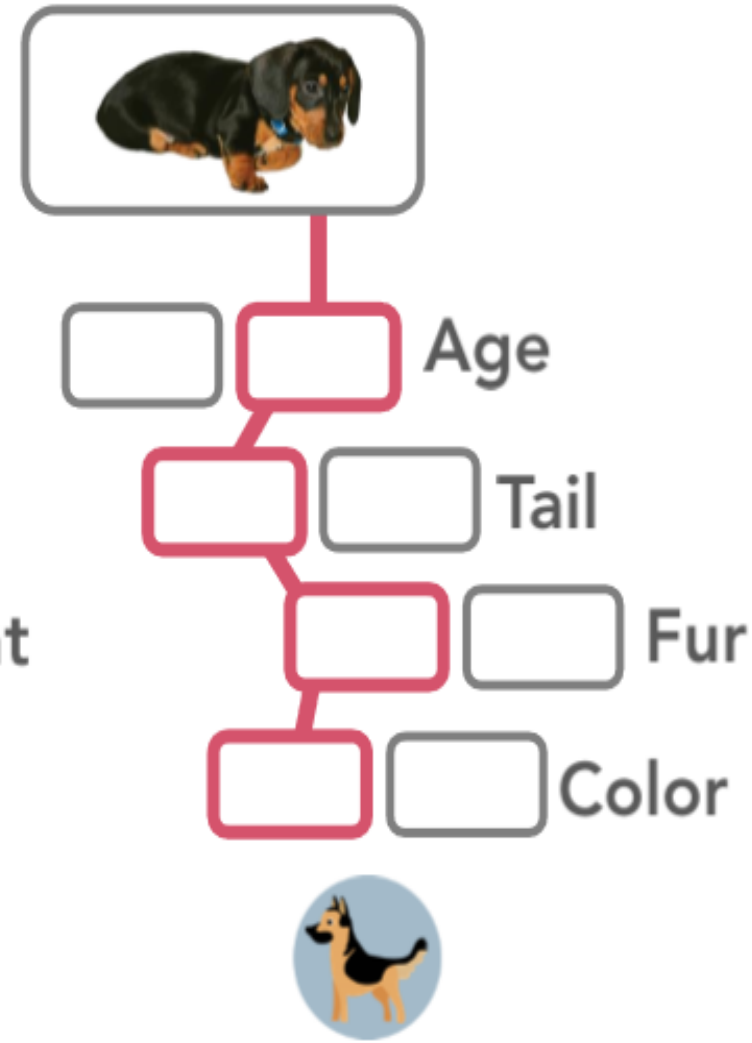
Size
Color
Fur
Ears
Tail
Age
Weight

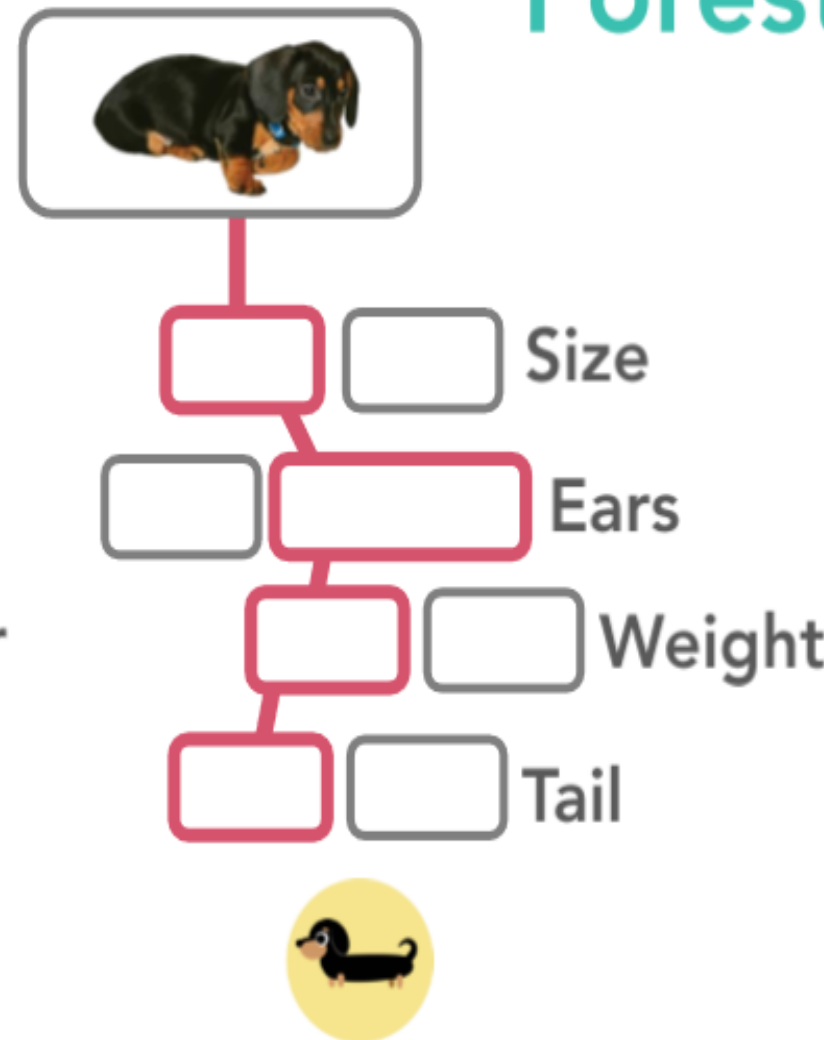
Decision Tree



Forest

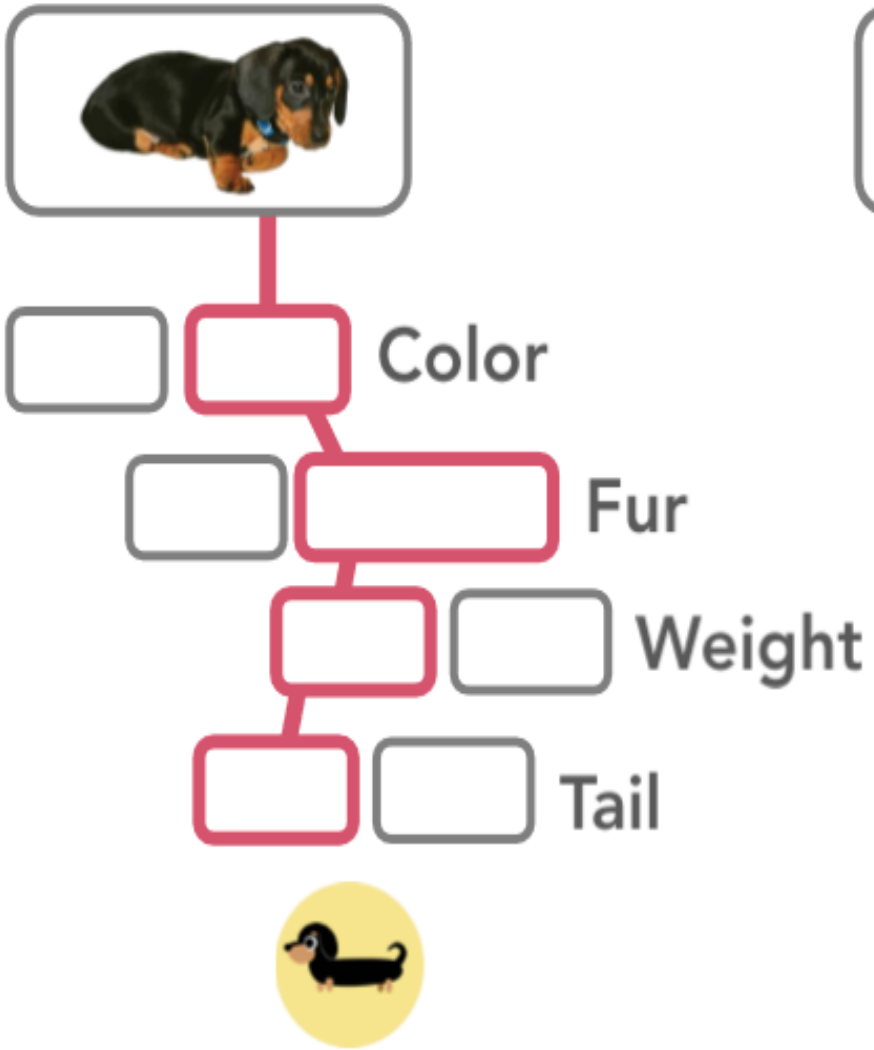


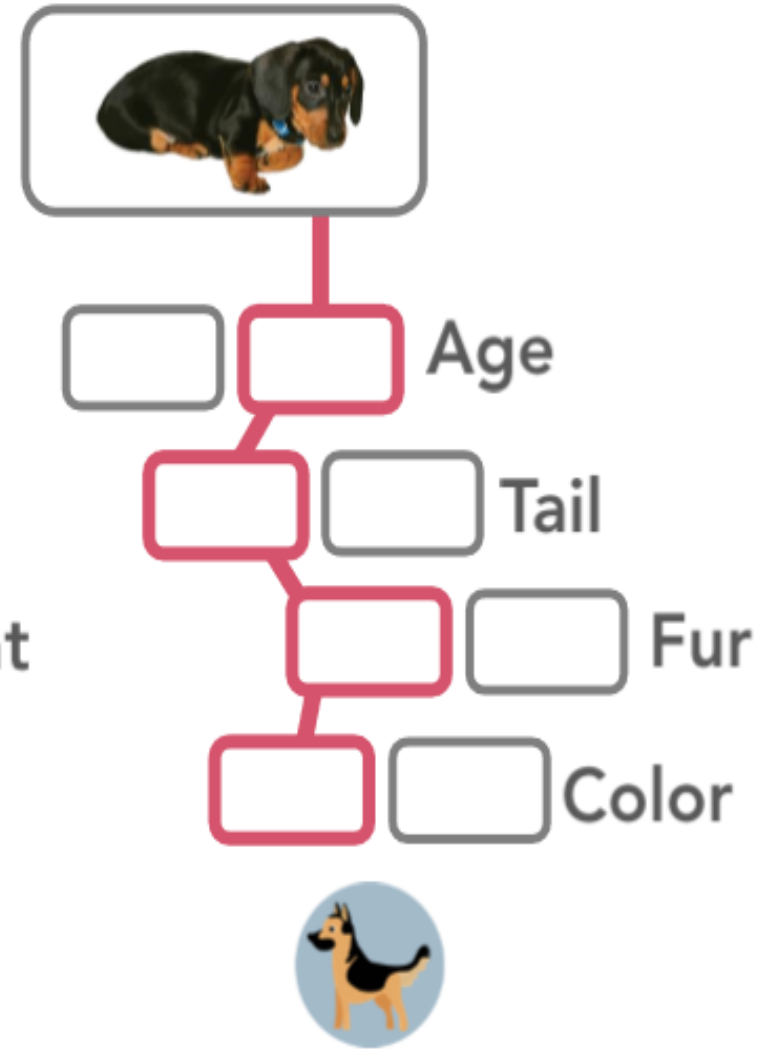


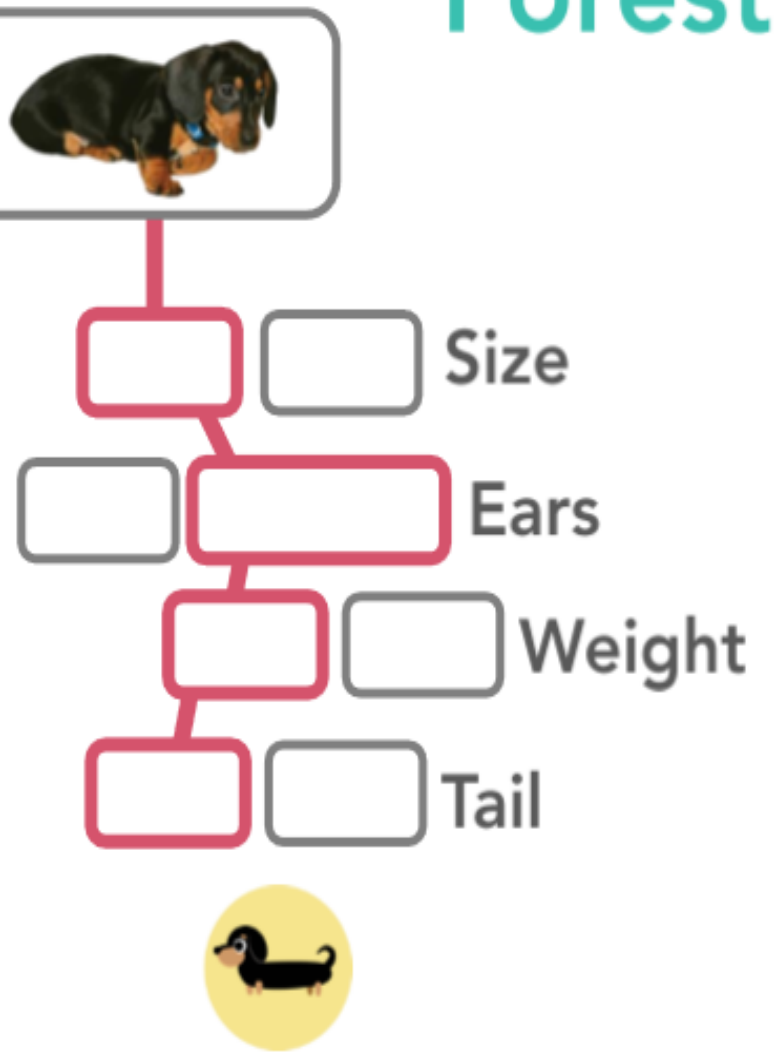


Random subset of data and variables used in each tree

Forest



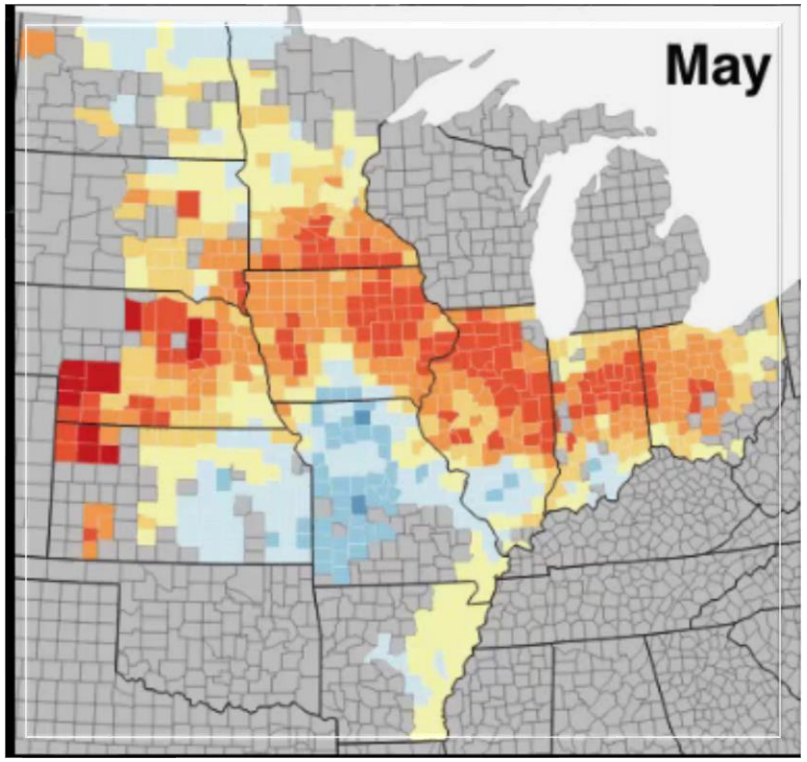




Majority vote wins

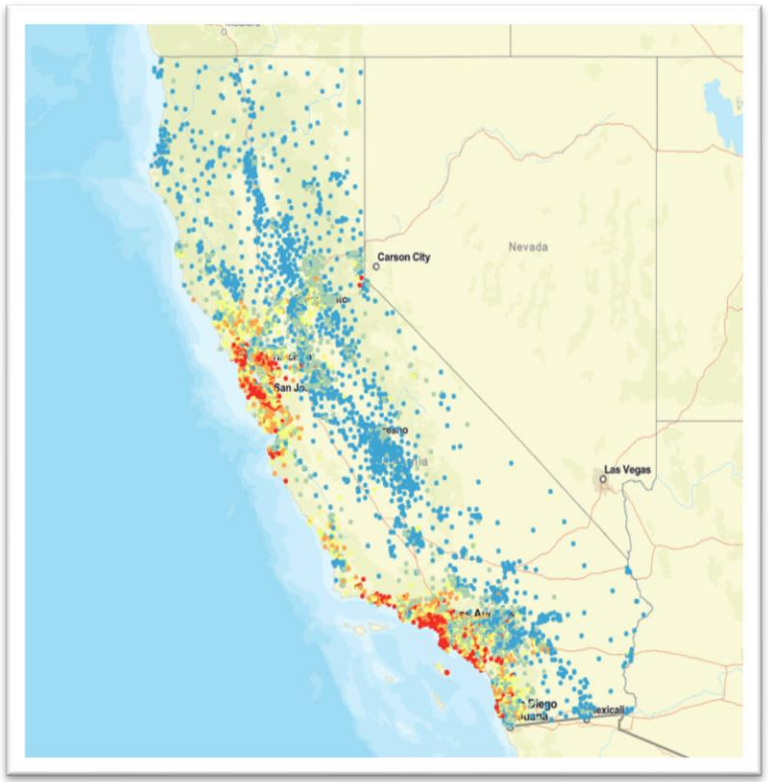
=

Potential Applications/Where it's used Today

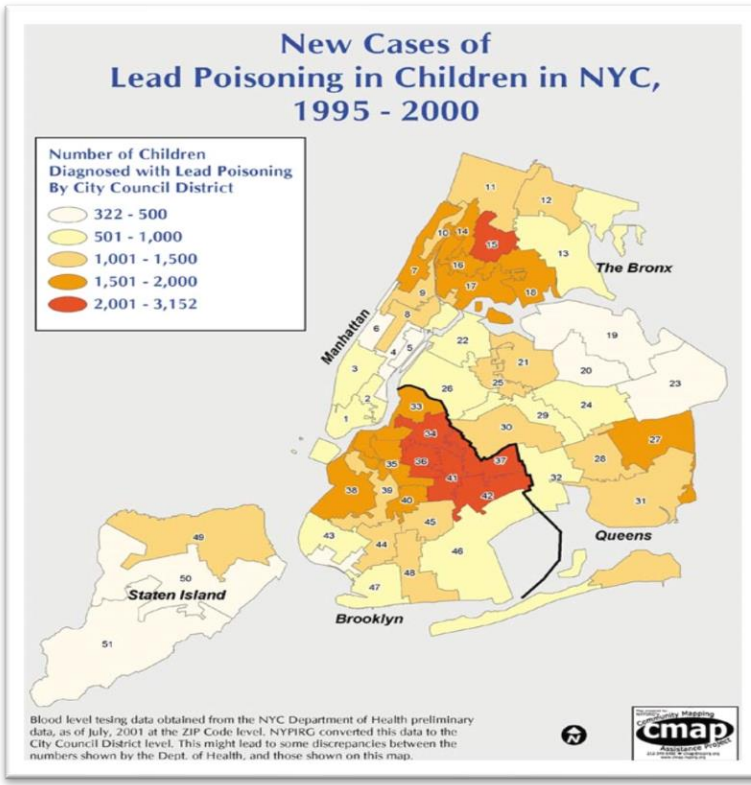


Crop Yield

JIAXUAN YOU, XIAOCHENG LI, MELVIN LOW, DAVID B. LOBELL, STEFANO ERMON-
STANFORD UNIVERSITY



Housing Values



Lead Poisoning

Data Visualization

Kristen Hocutt

Why visualize data?

Convert slow reasoning tasks into fast perception tasks

1399683

6962578

1019462

1345609

724027

844322

587106

5731663

1628760

635200

5401748

3922722

1336265

3111844

6656872

1877879

19631599

12817894

3601157

1050228

8875318

6581982

2815039

2917750

38120066

11575704

12914651

629120

927030

1885932

2908933

8275961

6074504

6636256

3880520

9913774

6539407

26538203

5903286

5235100

4419036

2106392

4885854

3018484

9978939

4796559

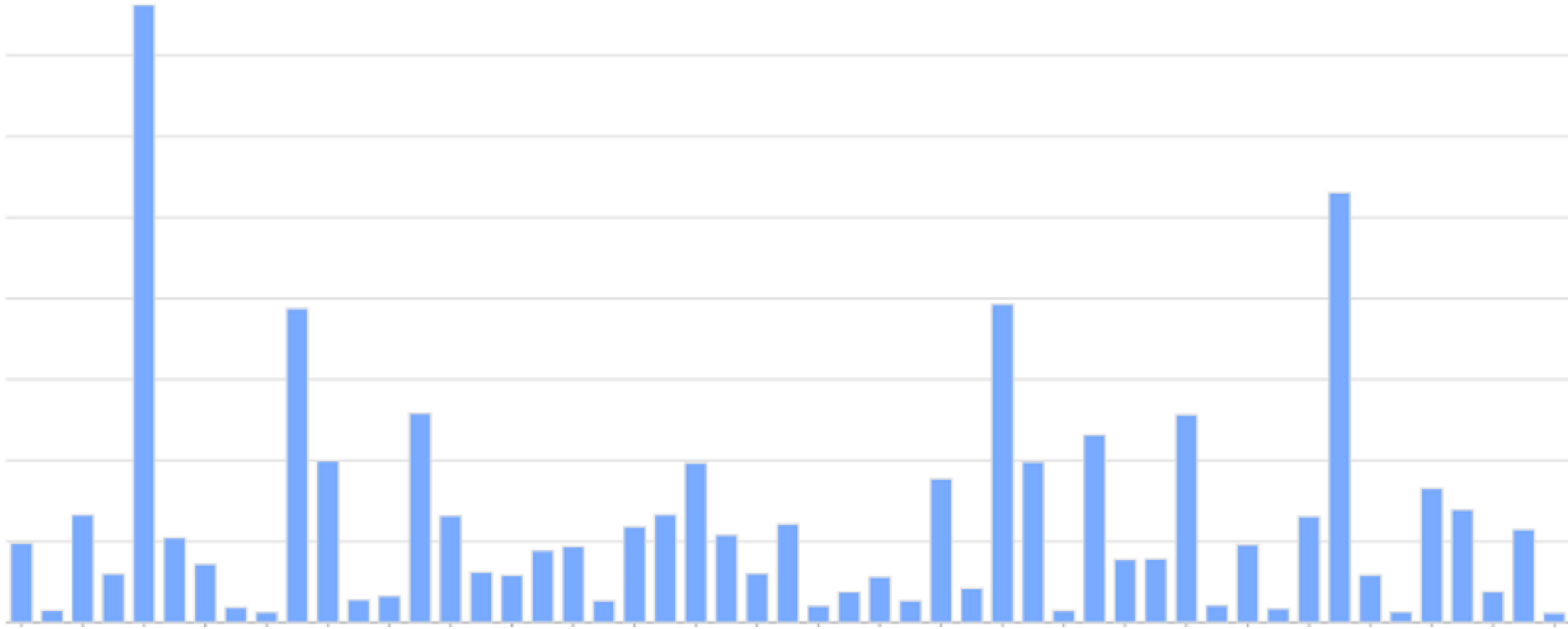
2995330

4681639

19383475

9853722

742404



1399683

6962578

1019462

1345609

724027

844322

587106

5731663

1628760

635200

5401748

3922722

1336265

3111844

6656872

1877879

19631599

12817894

3601157

1050228

8875318

6581982

2815039

2917750

38120066

11575704

12914651

629120

927030

1885932

2908933

8275961

6074504

6636256

3880520

9913774

6539407

26538203

5903286

5235100

4419036

2106392

4885854

3018484

9978939

4796559

2995330

4681639

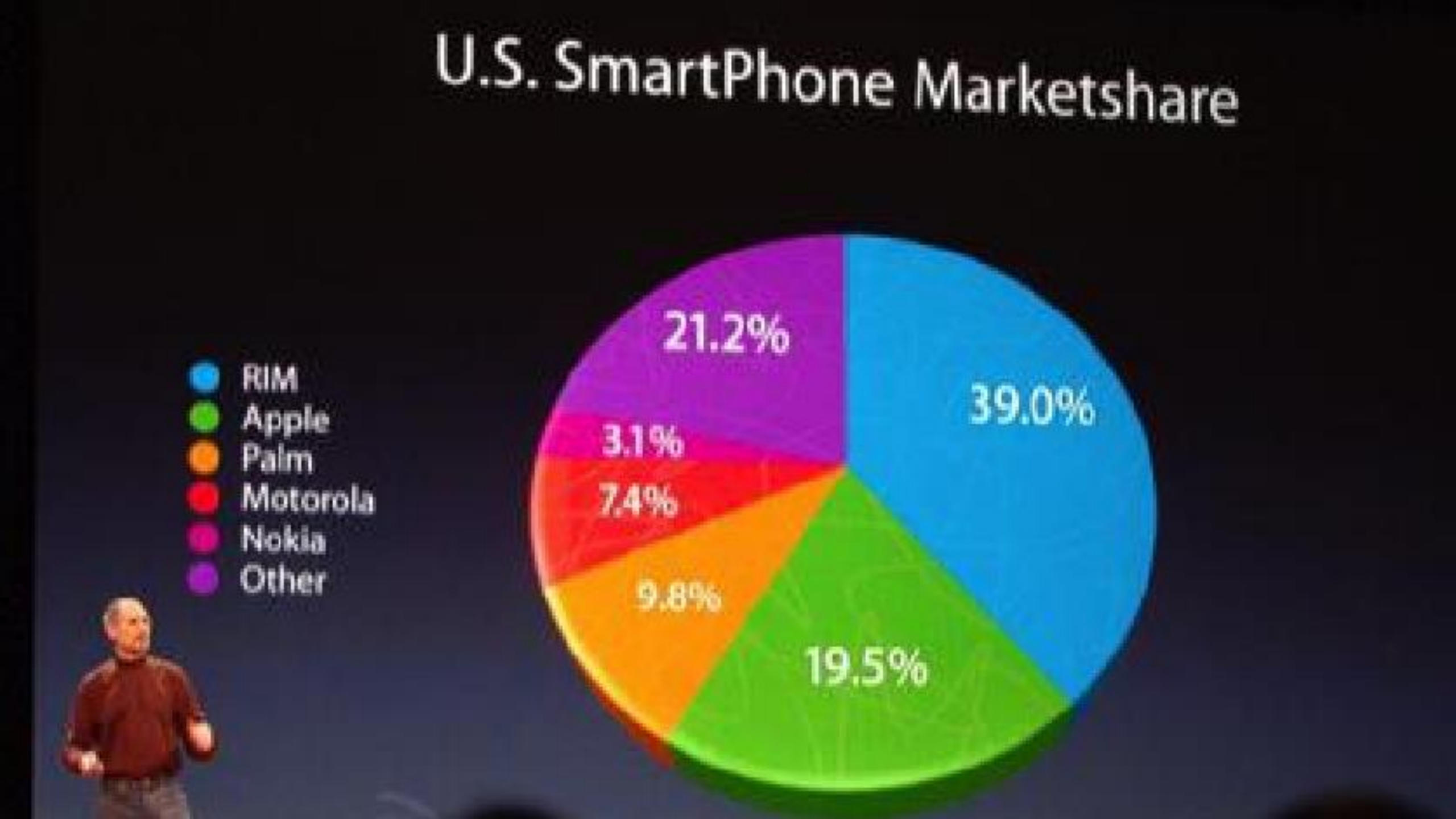
19383475

9853722

742404

Good Viz vs Bad Viz

U.S. SmartPhone Marketshare



Visualizations to support spatial analysis

Distributions and frequency

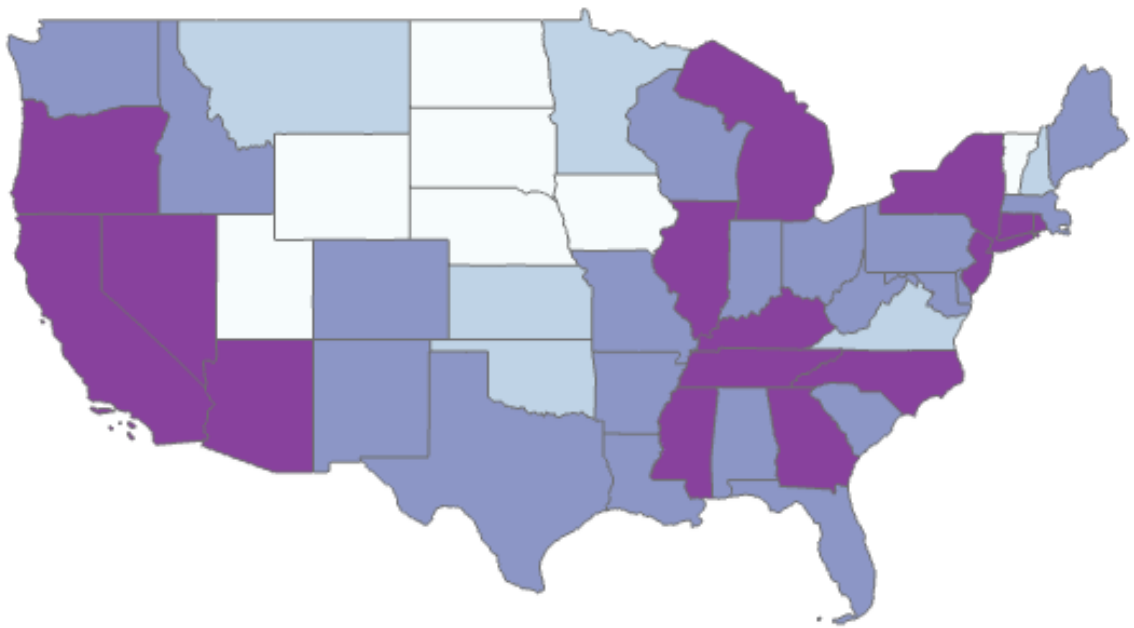
Category comparisons

Relationships and correlations

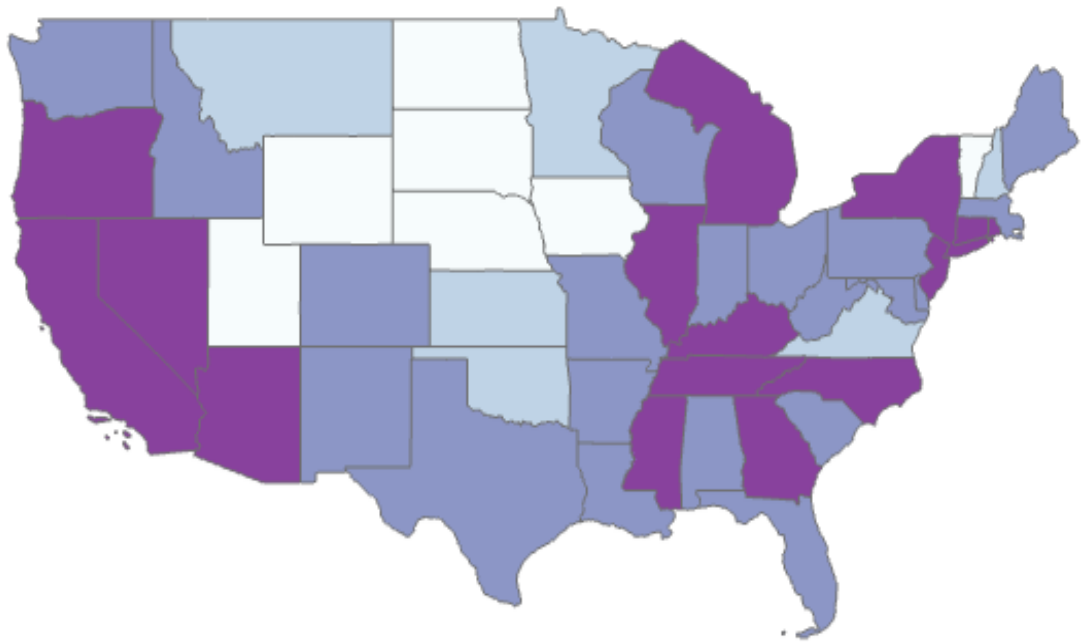
Change over time or distance

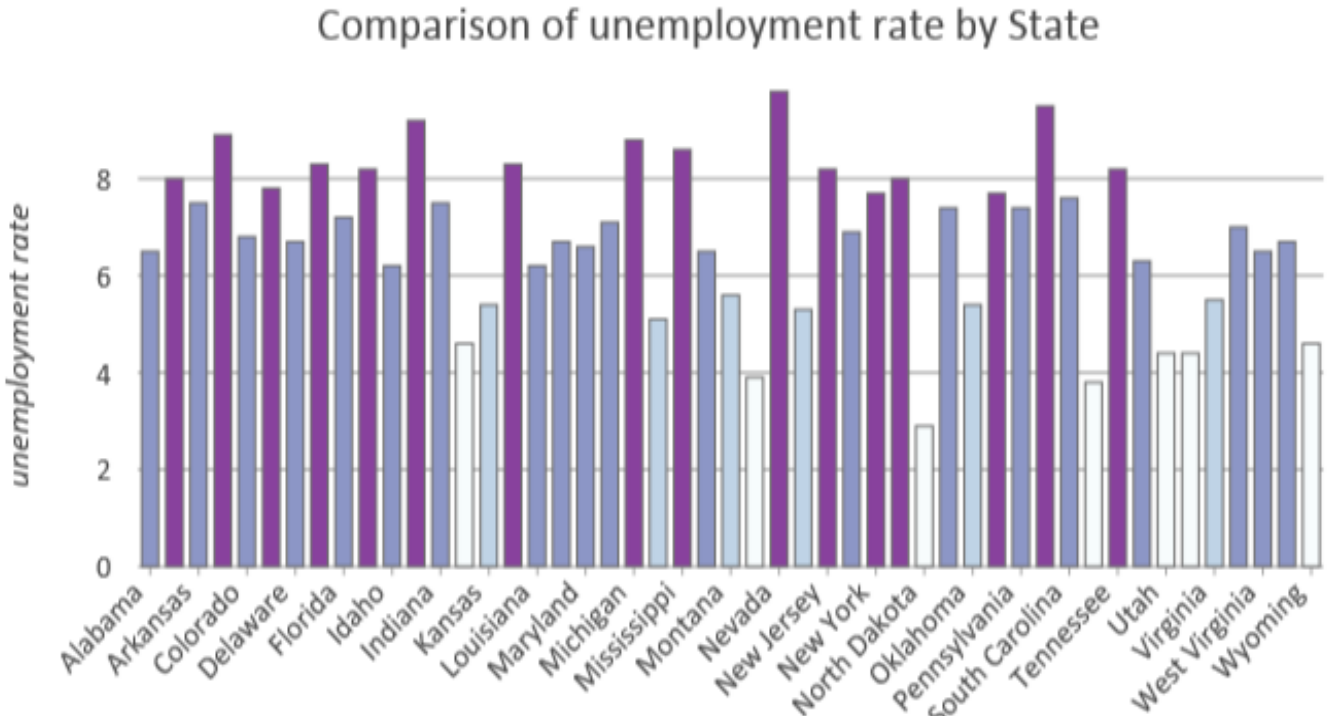
When a map (alone) isn't the best option…

When a map (alone) isn't the best option…

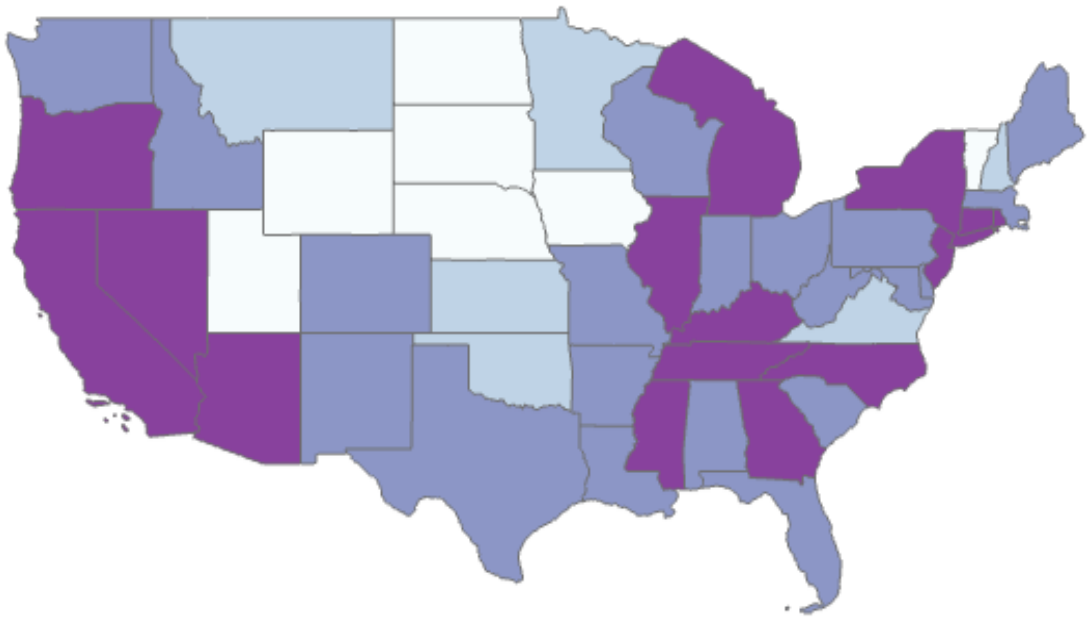


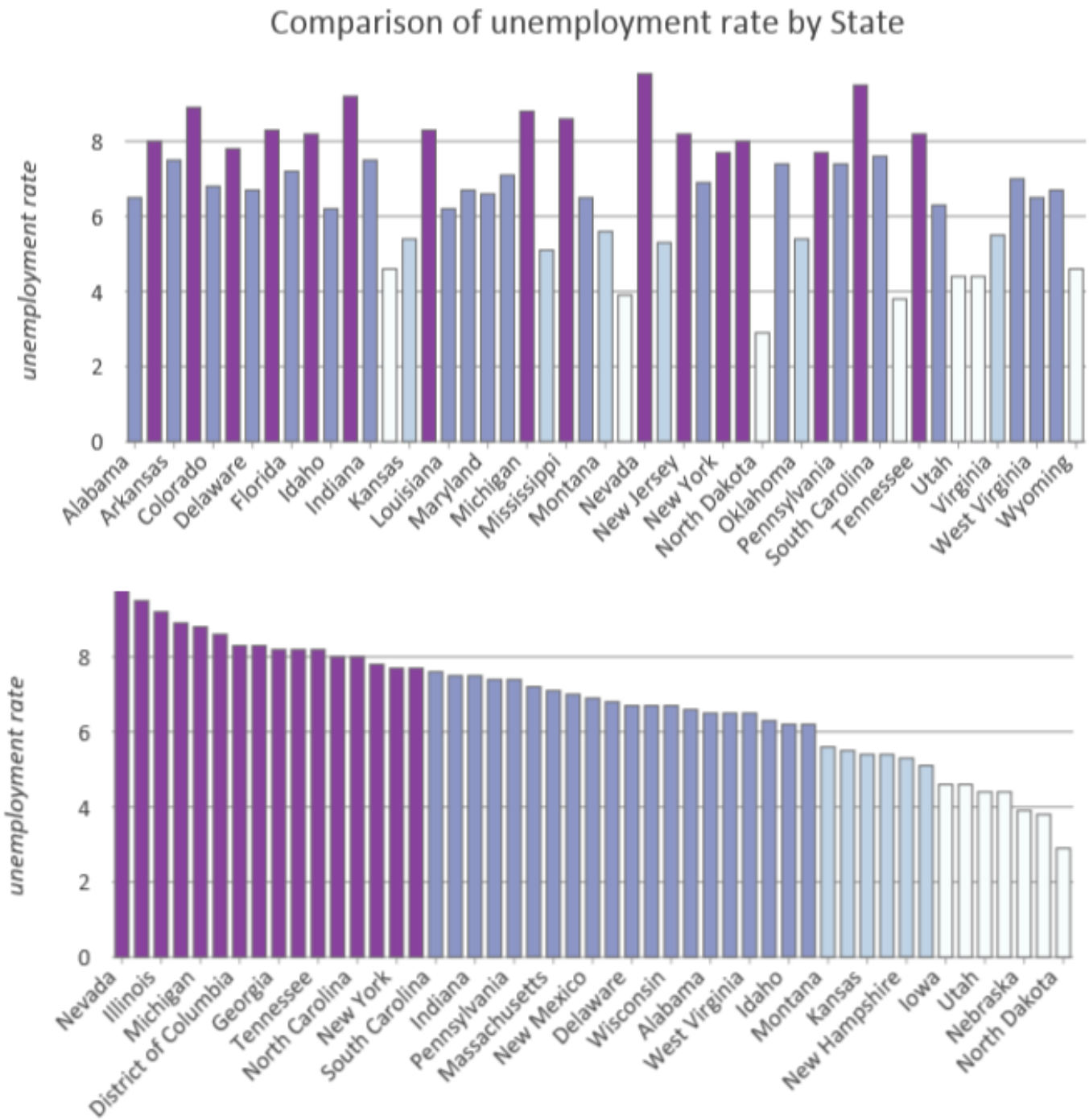
When a map (alone) isn't the best option...





When a map (alone) isn't the best option...





Key Takeaways

Something happens Everywhere

Open Source paired with COTS can enhance workflows

Data Visualization & Messaging drive quick decisions

Resources

- Learn ArcGIS
- R Bridge
- ArcGIS Pro
- Spatial Statistics Page

Questions?

Kristen Hocutt- khocutt@esri.com
Lu Zhang- l.zhang@esri.com