Beyond Where: Modeling Spatial Relationships and Making Predictions

Lauren Bennett

Jenora D'Acosta

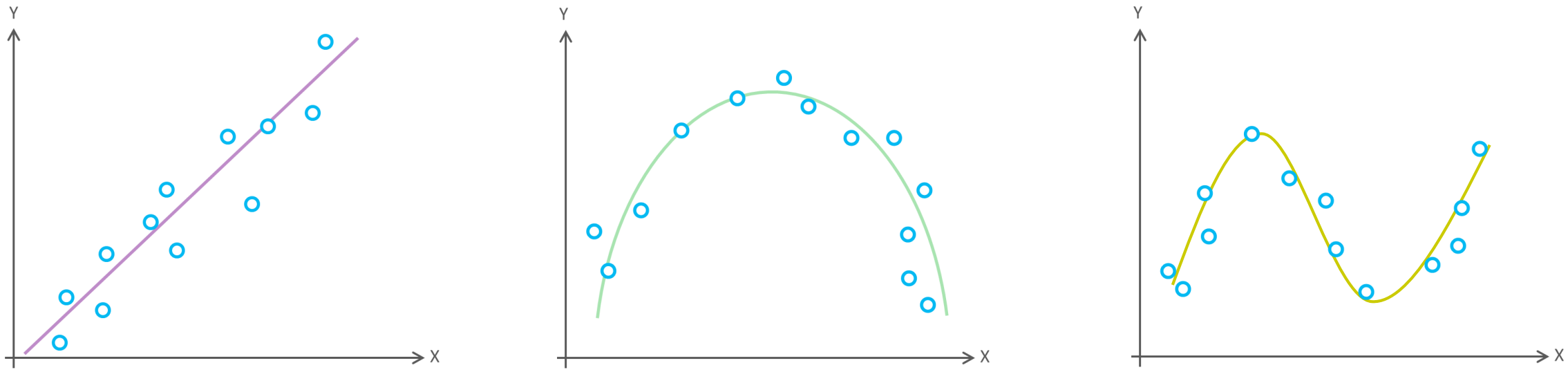
Alberto Nieto

Flora Vale

esriurl.com/spatialstats

SEE WHAT OTHERS CAN'T

Modeling Relationships



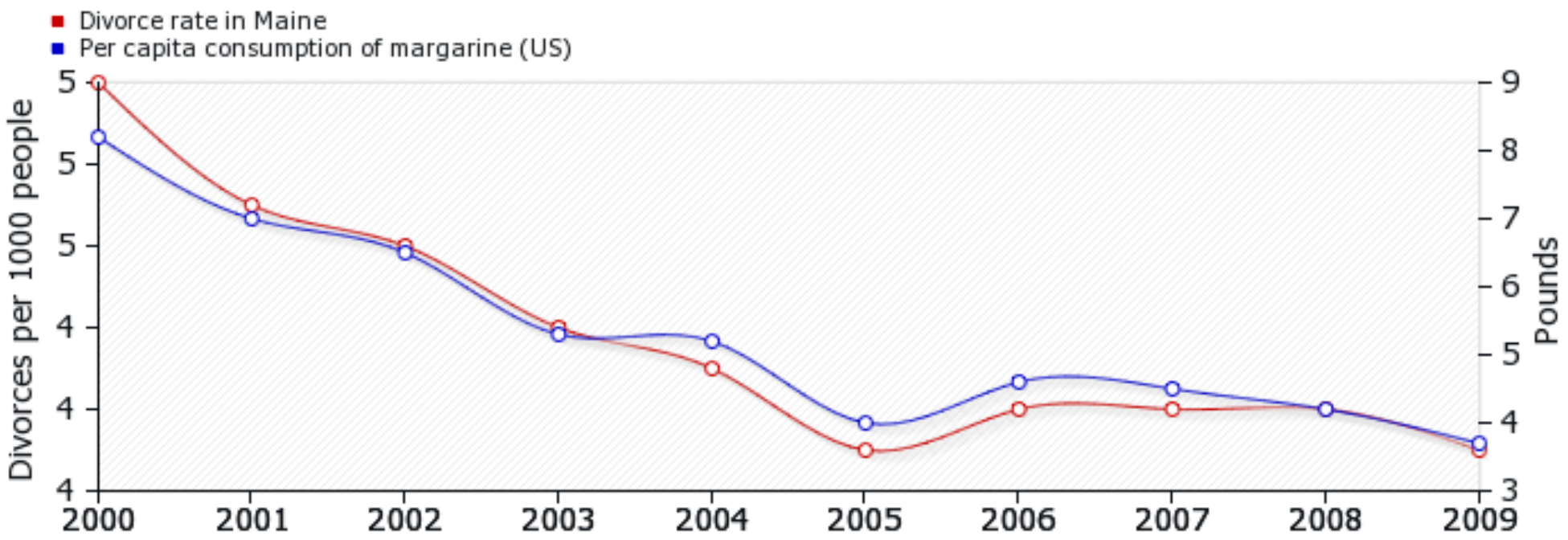
when variables are related, information can be learned about one variable by observing the values of the related variable(s)

Explore correlations

Predict unknown values

Understand key factors

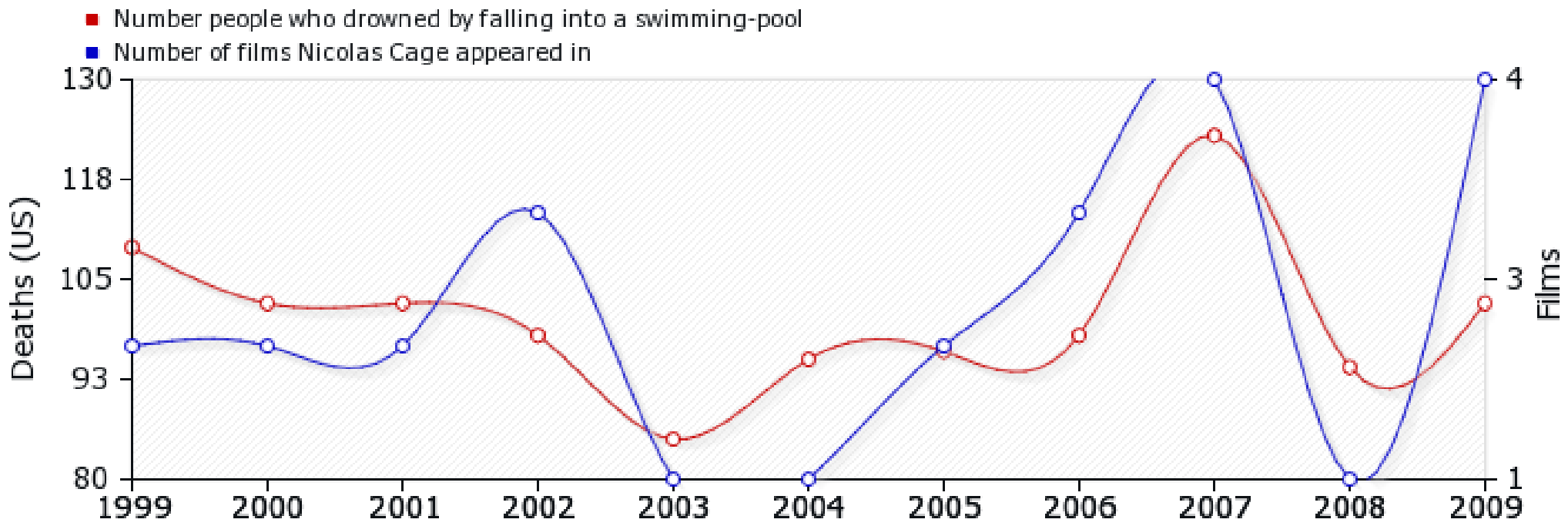
Divorce Rate in Maine vs Per Capita Consumption of Margarine



tylervigen.com

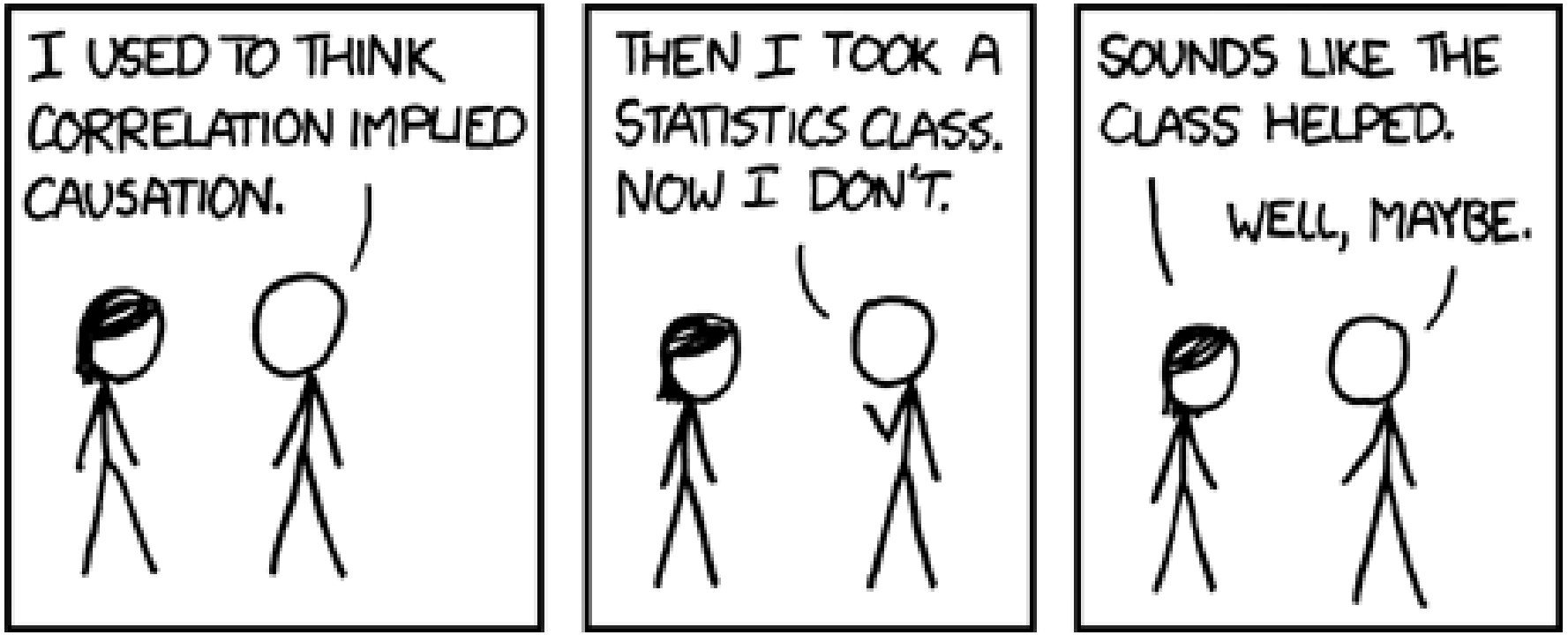
Correlation: 0.992558

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



Correlation: 0.666004

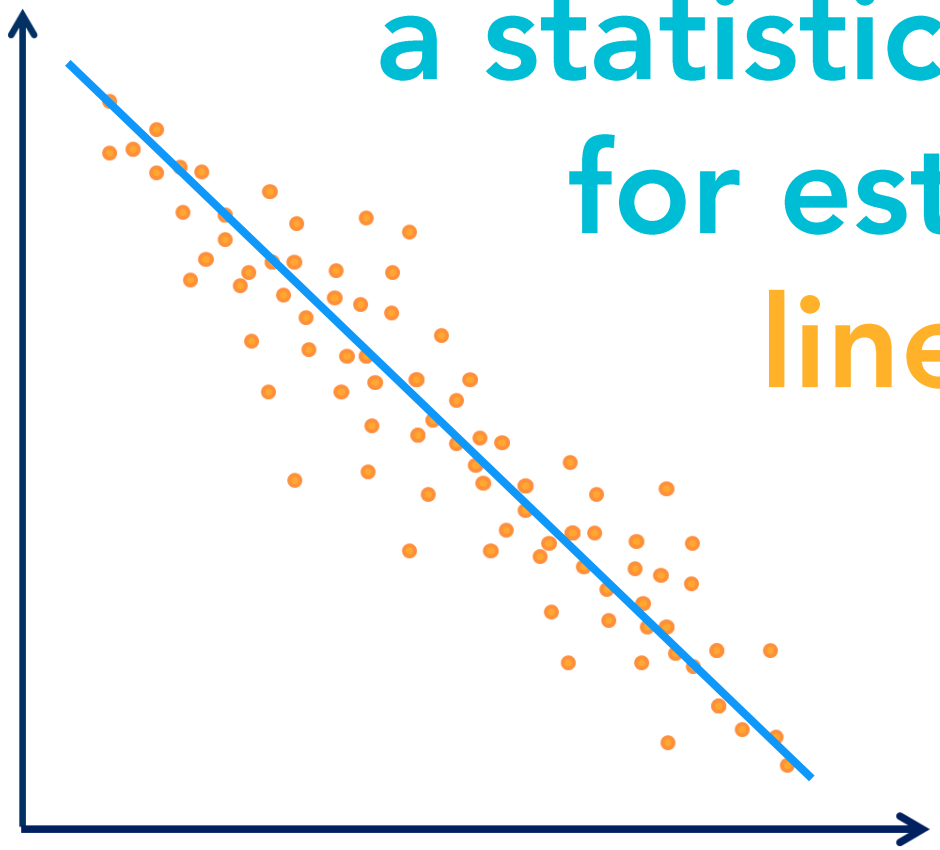
tylervigen.com



http://xkcd.com/552/

Generalized Linear Regression

Modeling linear relationships



a statistical process
for estimating
linear relationships
between variables

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

Dependent Variable

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

What are you trying to predict or understand?

Explanatory Variables

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

Variables you believe to cause or explain the dependent variable

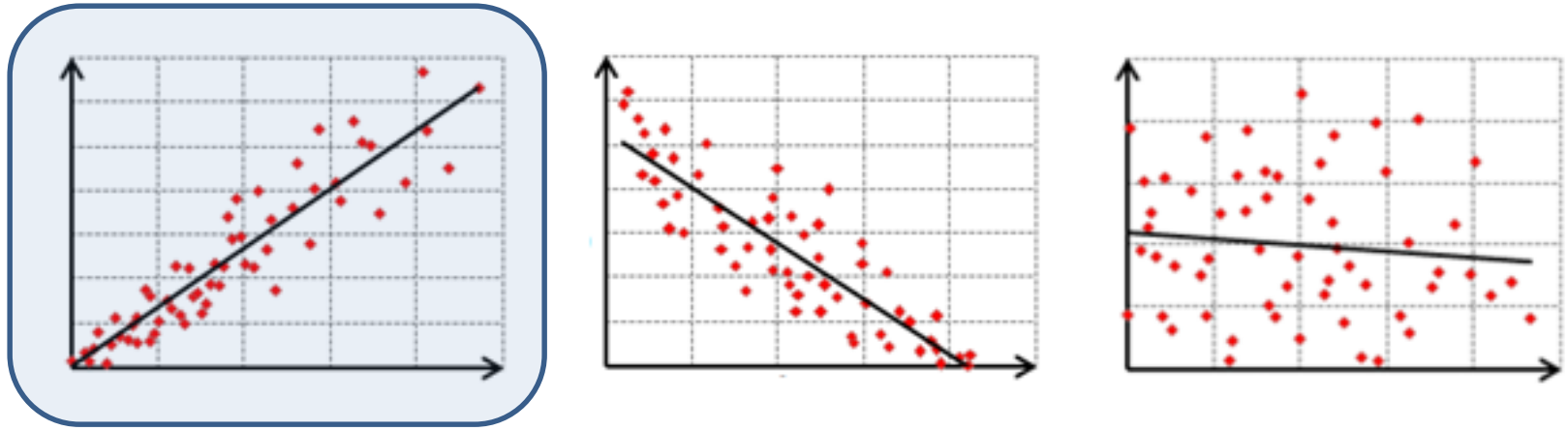
Coefficients

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

Represent the strength and type of relationship that X has to y

Coefficients

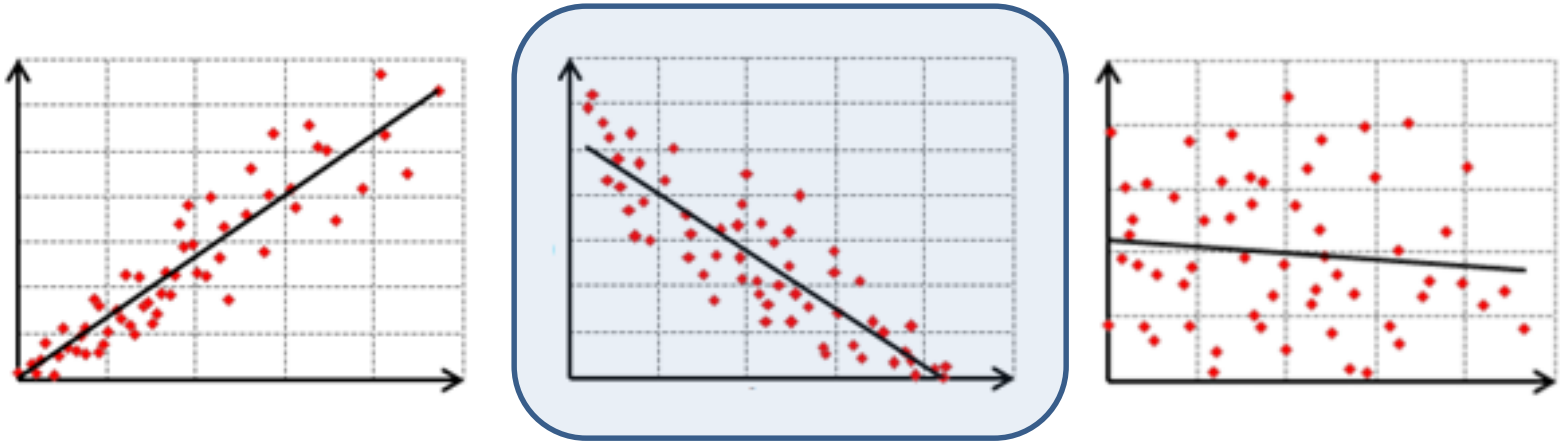
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$



Positive relationship- as obesity rates rise, diabetes rates also rise

Coefficients

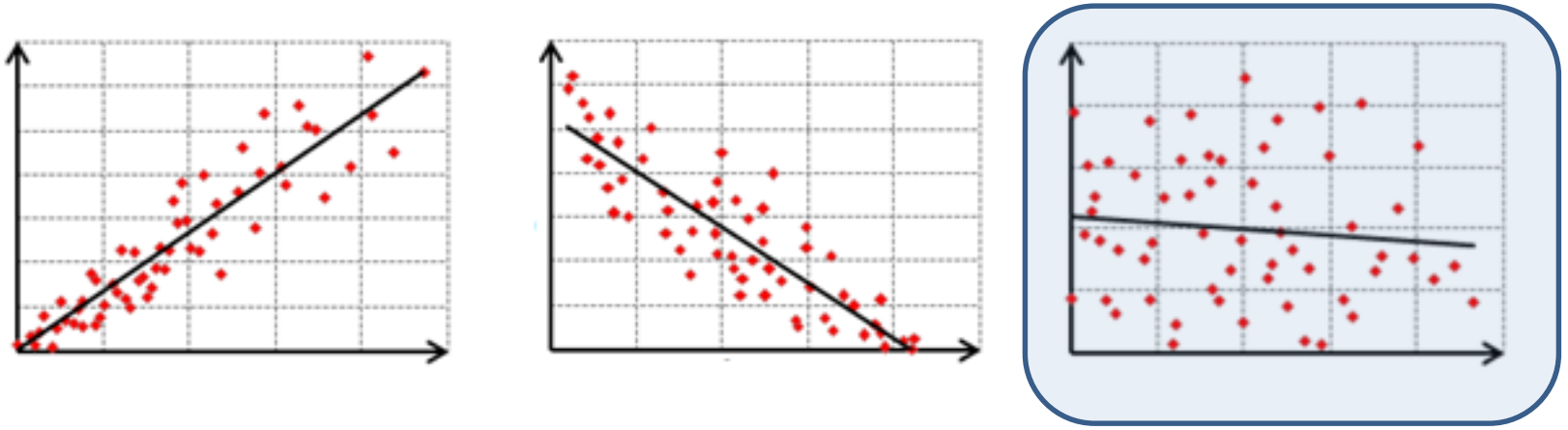
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$



Negative relationship- as foreclosure rates rise, home values drop

Coefficients

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



No relationship- the value for X is not correlated with the value for y

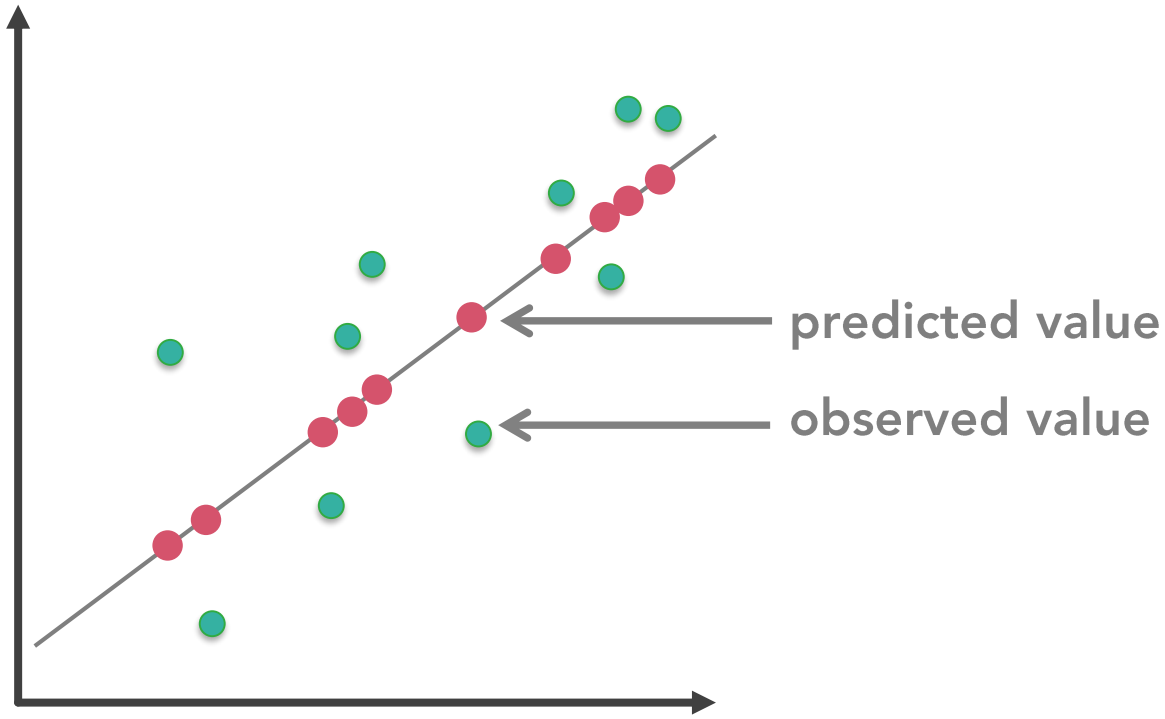
Residual

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

Model over and under predictions

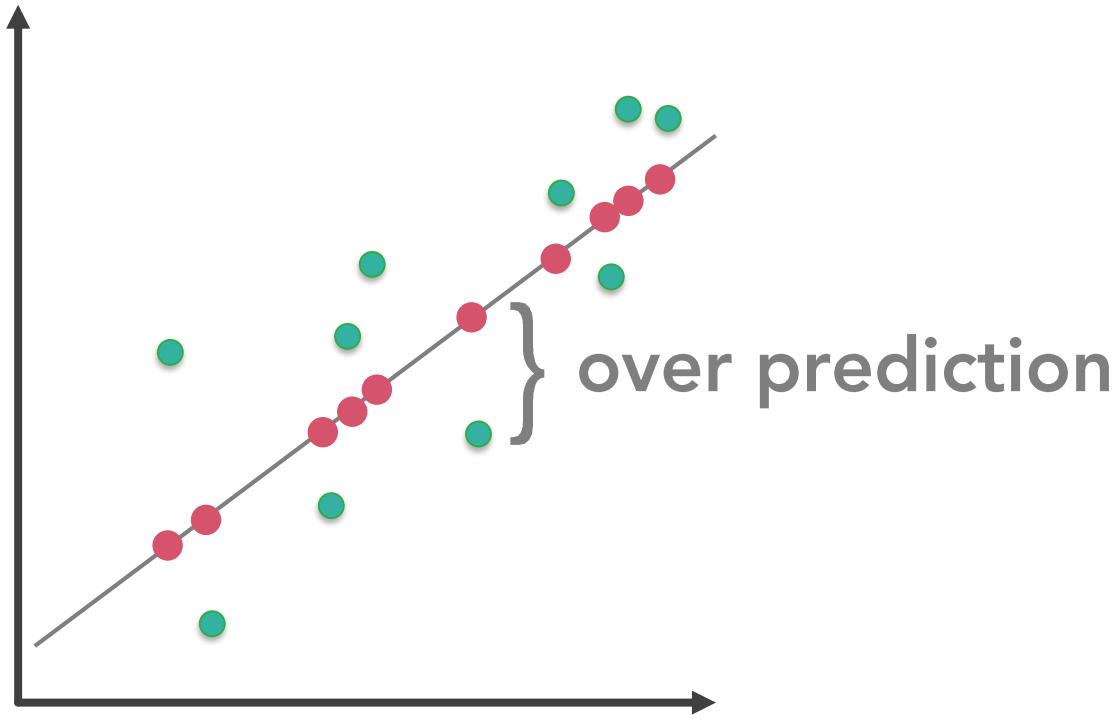
Residual

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



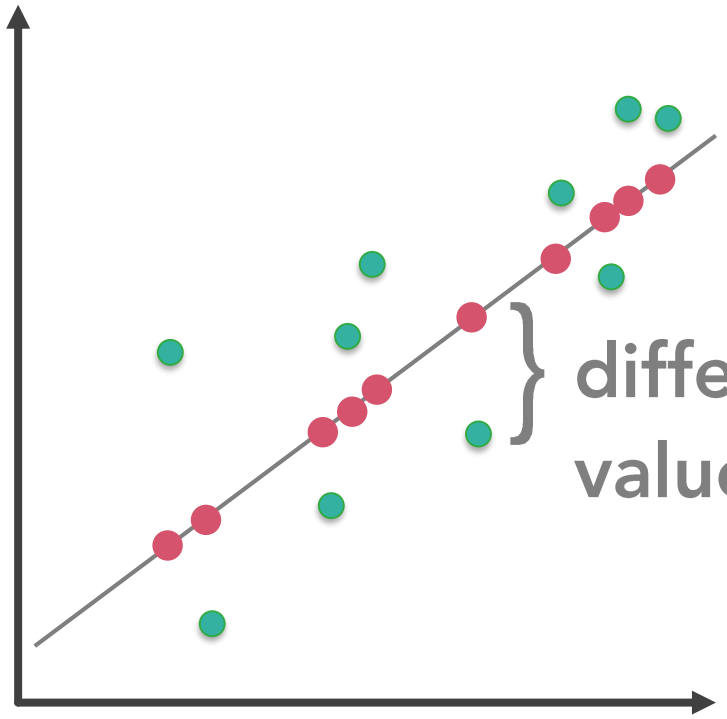
Residual

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



Residual

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



difference between the observed value and the predicted value = ε

Demo

Evaluating our model

Every variable should be statistically significant *

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

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	9947.038596	44.461950	223.720250	0.000000*	43.232352	230.083219	0.000000*	--------
HOSPBEDSD	453.838027	94.410182	4.807088	0.000006*	95.160978	4.769161	0.000007*	1.789789
EVANDMAND	0.399674	0.036395	10.981628	0.000000*	0.041993	9.517526	0.000000*	3.476159
IMAGINGD	0.359023	0.165592	2.168123	0.032424*	0.161761	2.219469	0.028620*	3.044556
HOUSTOND	-0.001145	0.000115	-9.999029	0.000000*	0.000104	-10.976894	0.000000*	1.329077
PQI10D	1.041130	0.220645	4.718573	0.000009*	0.213298	4.881106	0.000005*	1.528101

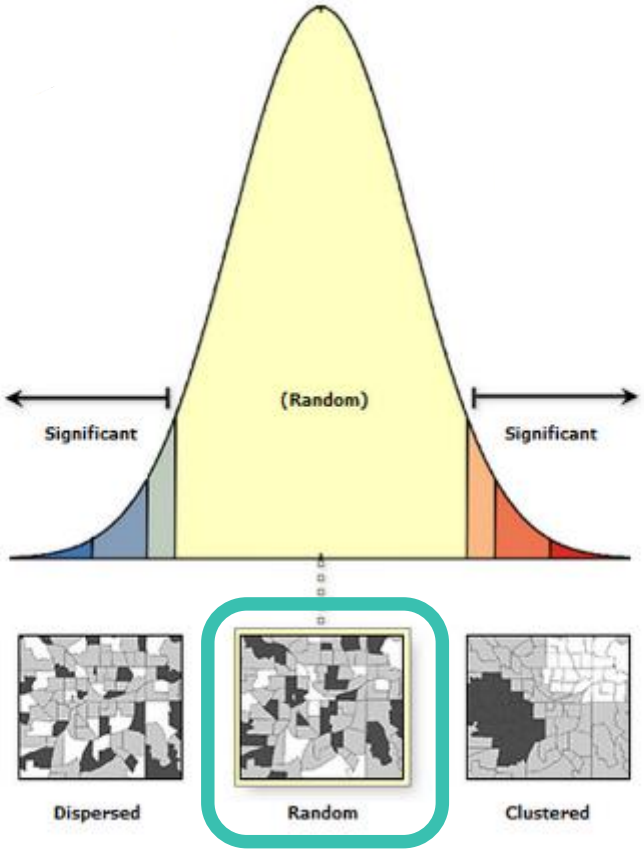
Each variable should tell a different part of the story

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

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	9947.038596	44.461950	223.720250	0.000000*	43.232352	230.083219	0.000000*	--------
HOSPBEDSD	453.838027	94.410182	4.807088	0.000006*	95.160978	4.769161	0.000007*	1.789789
EVANDMAND	0.399674	0.036395	10.981628	0.000000*	0.041993	9.517526	0.000000*	3.476159
IMAGINGD	0.359023	0.165592	2.168123	0.032424*	0.161761	2.219469	0.028620*	3.044556
HOUSTOND	-0.001145	0.000115	-9.999029	0.000000*	0.000104	-10.976894	0.000000*	1.329077
PQI10D	1.041130	0.220645	4.718573	0.000009*	0.213298	4.881106	0.000005*	1.528101

Residuals should not be clustered in location or in value

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



Residuals should not be clustered in location or in value

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

Input Features:	Hot Spot Study Area	Dependent Variable:	TOTAL_COSTS_2010
Number of Observations:	110	Akaike's Information Criterion (AICc) [d]:	1672.966945
Multiple R-Squared [d]:	0.870841	Adjusted R-Squared [d]:	0.864631
Joint F-Statistic [e]:	140.241872	Prob(>F), (5,104) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	556.069919	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	27.470483	Prob(>chi-squared), (5) degrees of freedom:	0.000046*
Jarque-Bera Statistic [g]:	1.591597	Prob(>chi-squared), (2) degrees of freedom:	0.451221

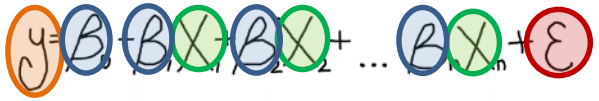
Model should have a strong R-Squared

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

Input Features:	Hot Spot Study Area
Number of Observations:	110
Multiple R-Squared [d]:	0.870841
Joint F-Statistic [e]:	140.241872
Joint Wald Statistic [e]:	556.069919
Koenker (BP) Statistic [f]:	27.470483
Jarque-Bera Statistic [g]:	1.591597

Dependent Variable:	TOTAL_COSTS_2010
Akaike's Information Criterion (AICc) [d]:	1672.966945
Adjusted R-Squared [d]:	0.864631
Prob(>F), (5,104) degrees of freedom:	0.000000*
Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Prob(>chi-squared), (5) degrees of freedom:	0.000046*
Prob(>chi-squared), (2) degrees of freedom:	0.451221

When comparing models, lower AICc is better



Input Features:	Hot Spot Study Area
Number of Observations:	110
Multiple R-Squared [d]:	0.870841
Joint F-Statistic [e]:	140.241872
Joint Wald Statistic [e]:	556.069919
Koenker (BP) Statistic [f]:	27.470483
Jarque-Bera Statistic [g]:	1.591597

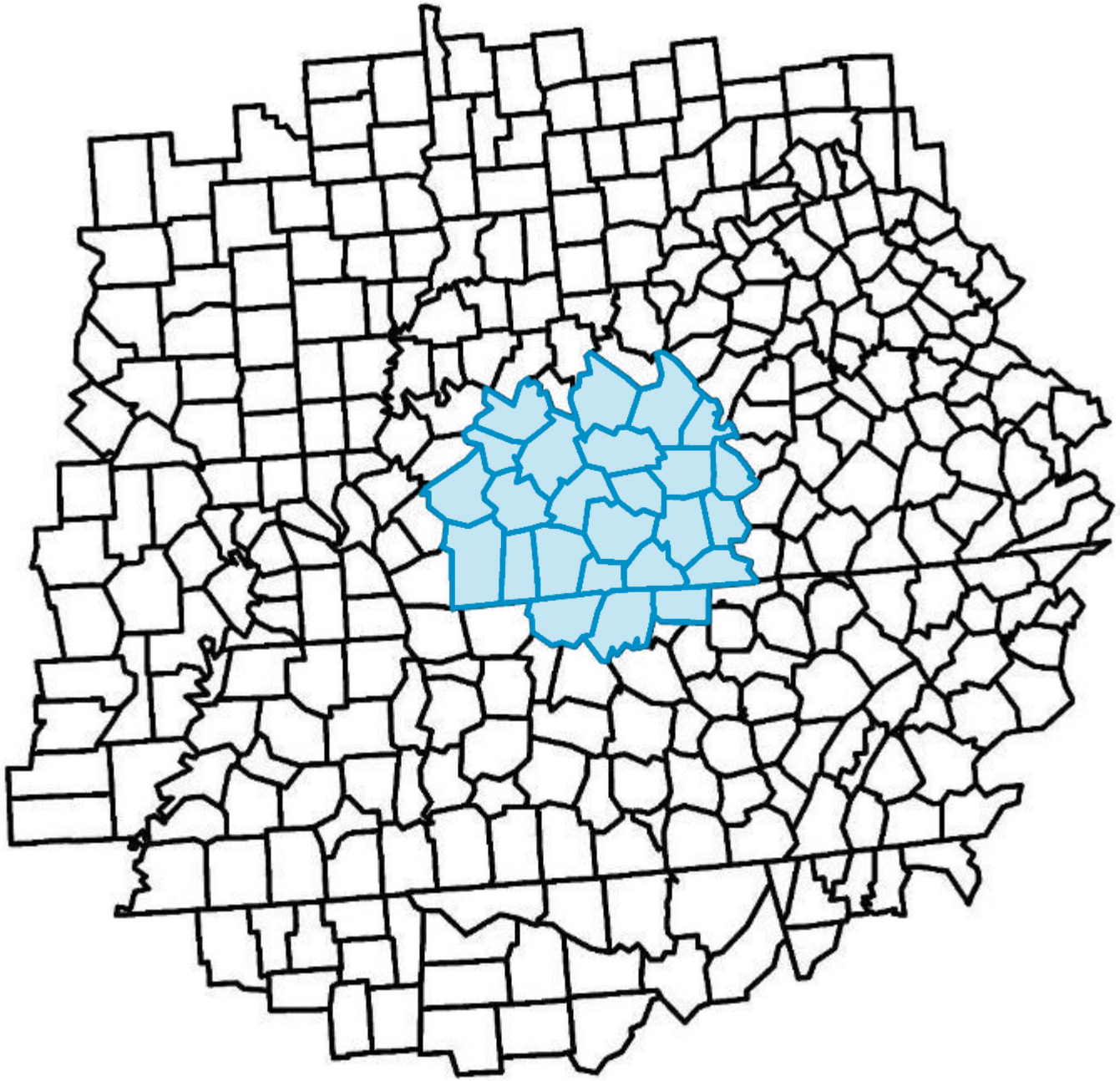
Dependent Variable:	TOTAL_COSTS_2010
Akaike's Information Criterion (AICc) [d]:	1672.966945
Adjusted R-Squared [d]:	0.864631
Prob(>F), (5,104) degrees of freedom:	0.000000*
Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Prob(>chi-squared), (5) degrees of freedom:	0.000046*
Prob(>chi-squared), (2) degrees of freedom:	0.451221

Exploratory Regression

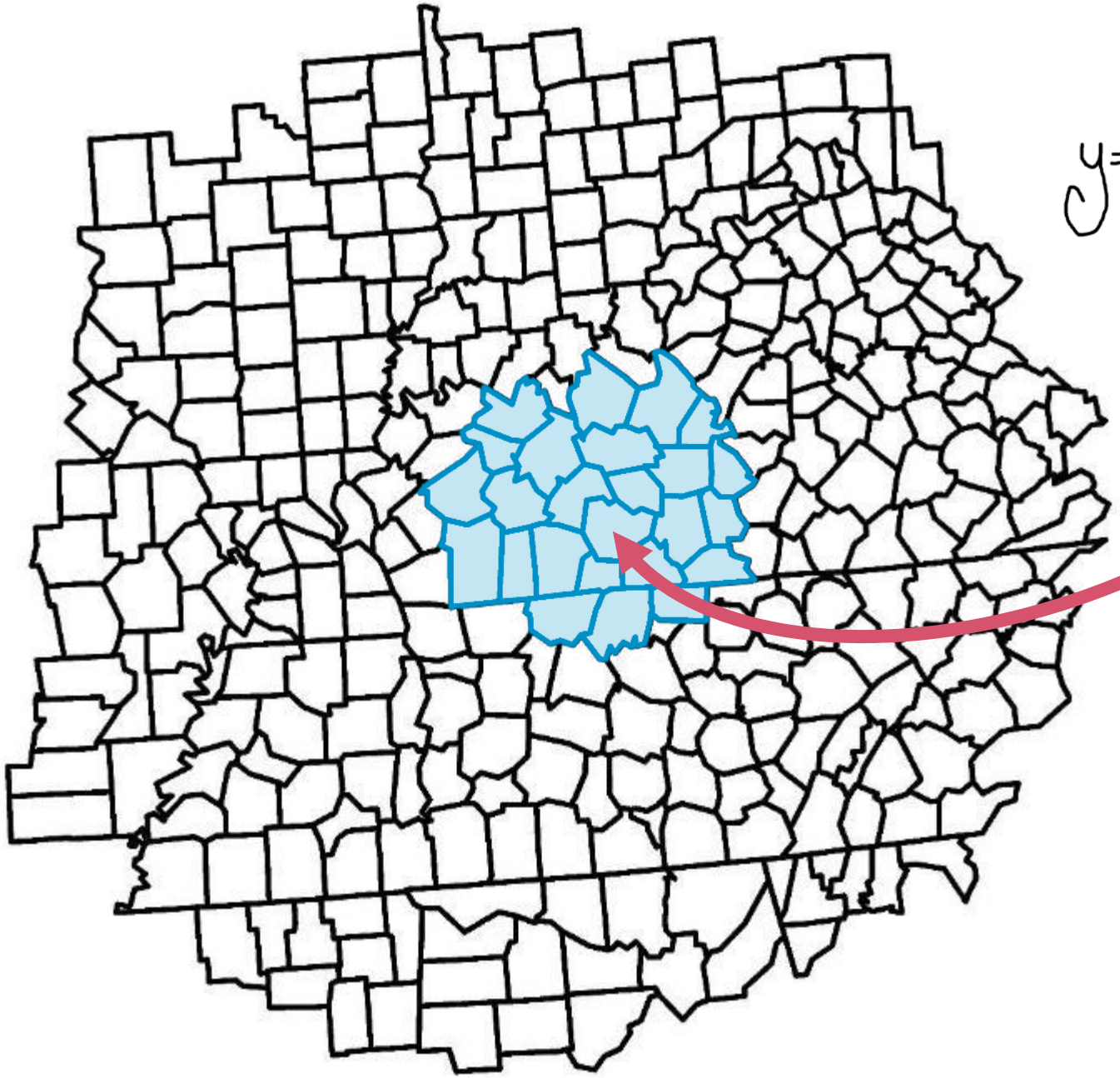
Demo

Geographically Weighted Regression

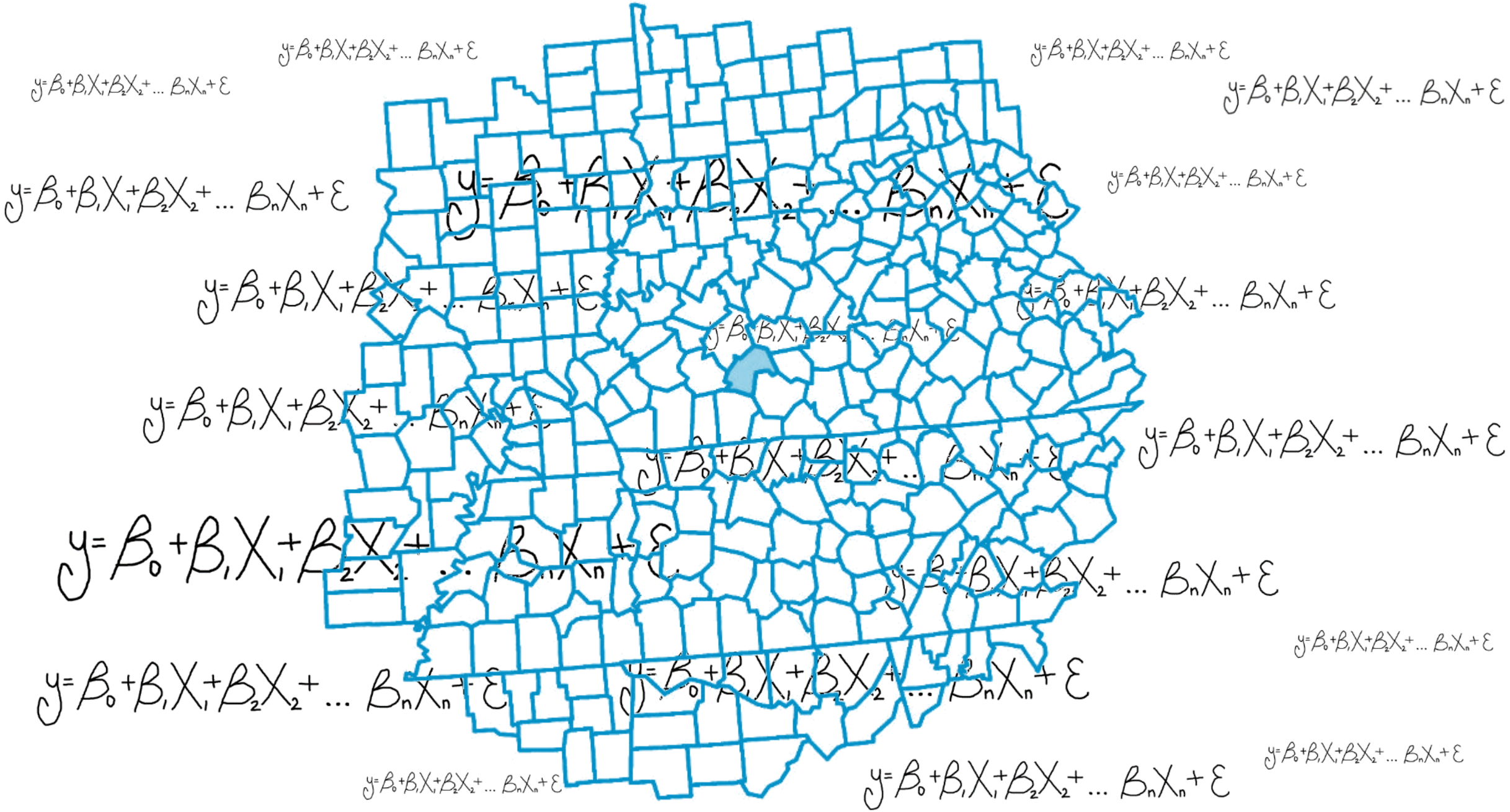
Exploring spatial variation

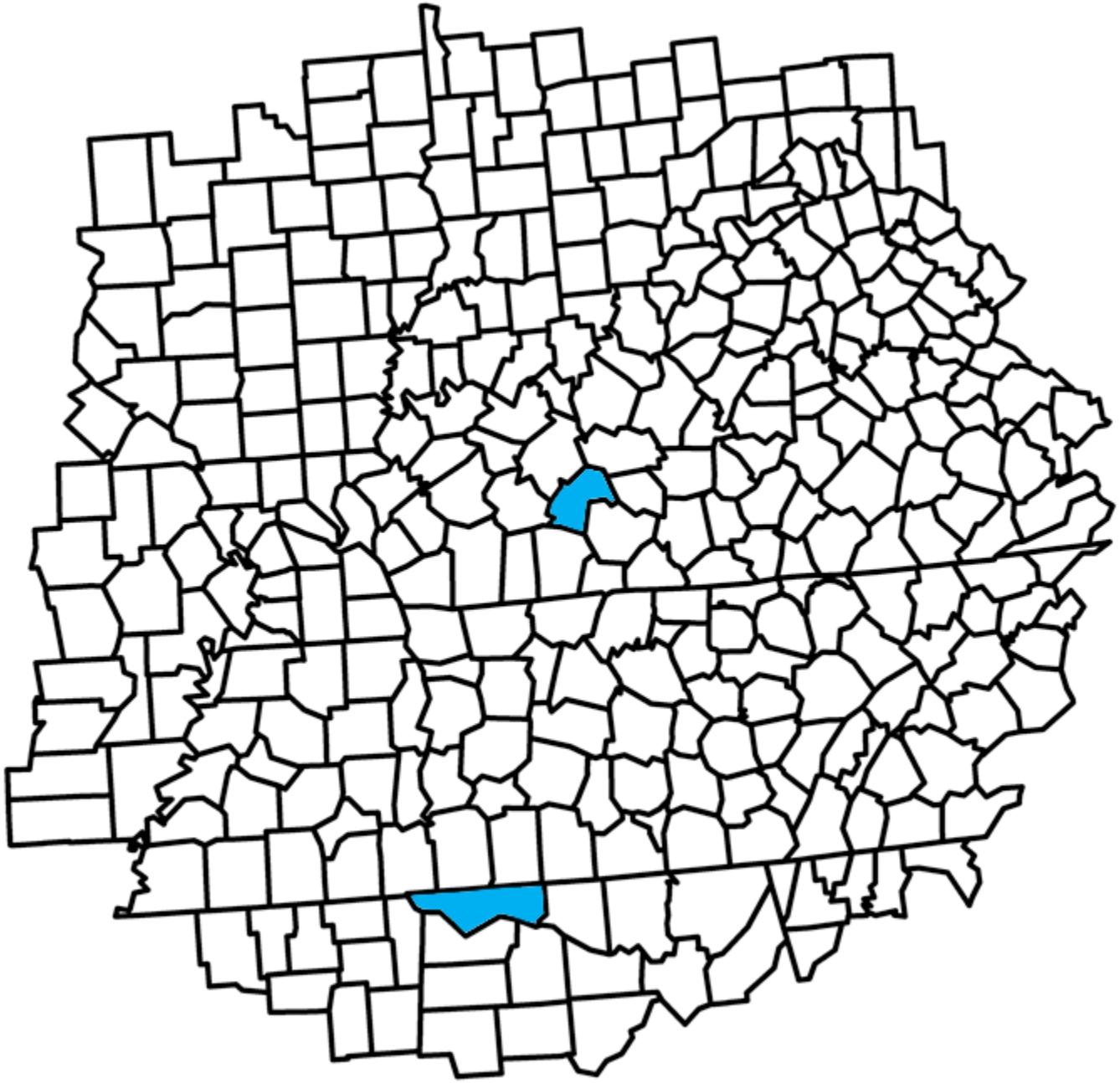


each feature gets a separate equation

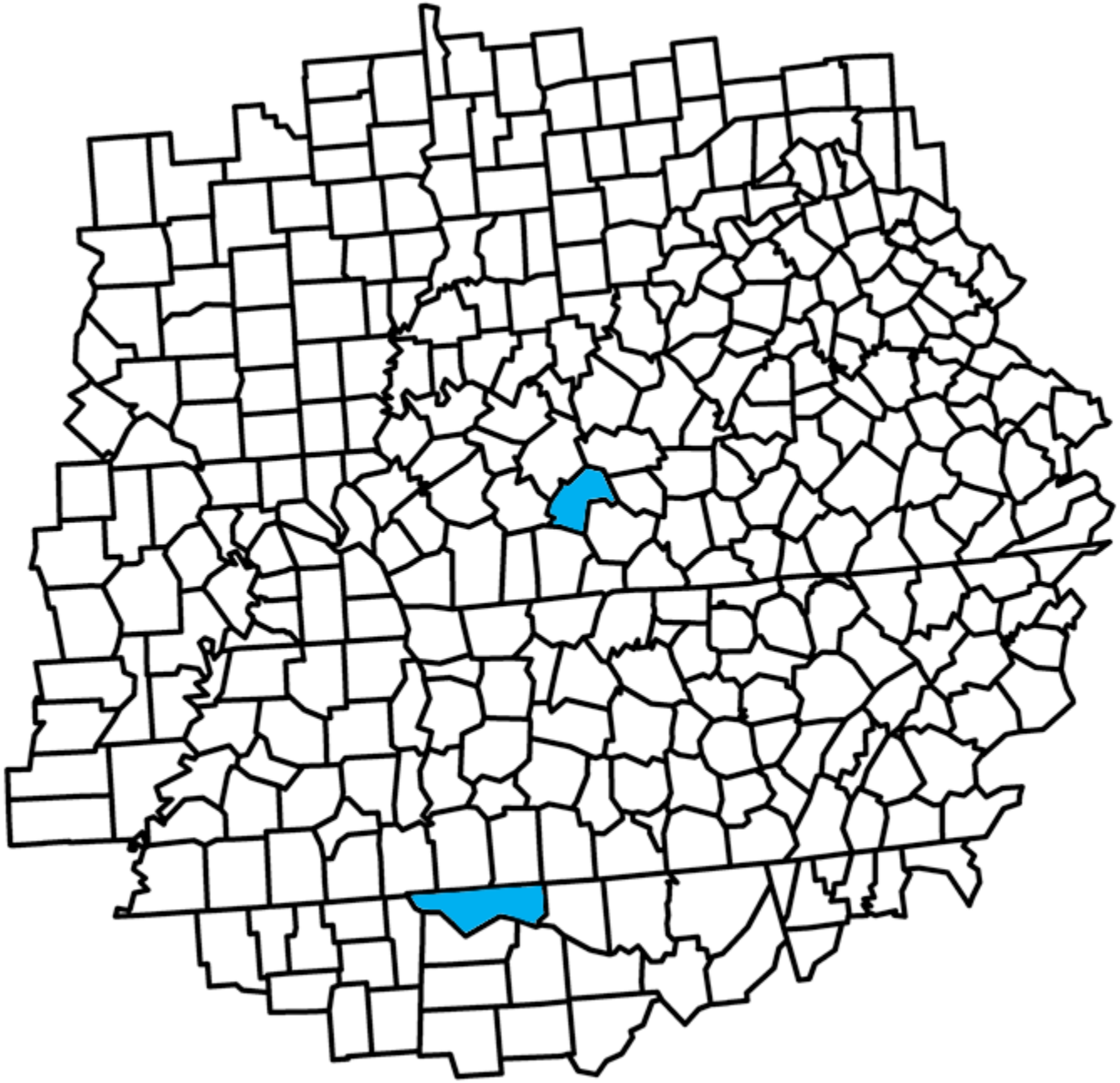


$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$


$y=\beta_0+\beta_1X_1+\beta_2X_2+\dots\beta_nX_n+\varepsilon$



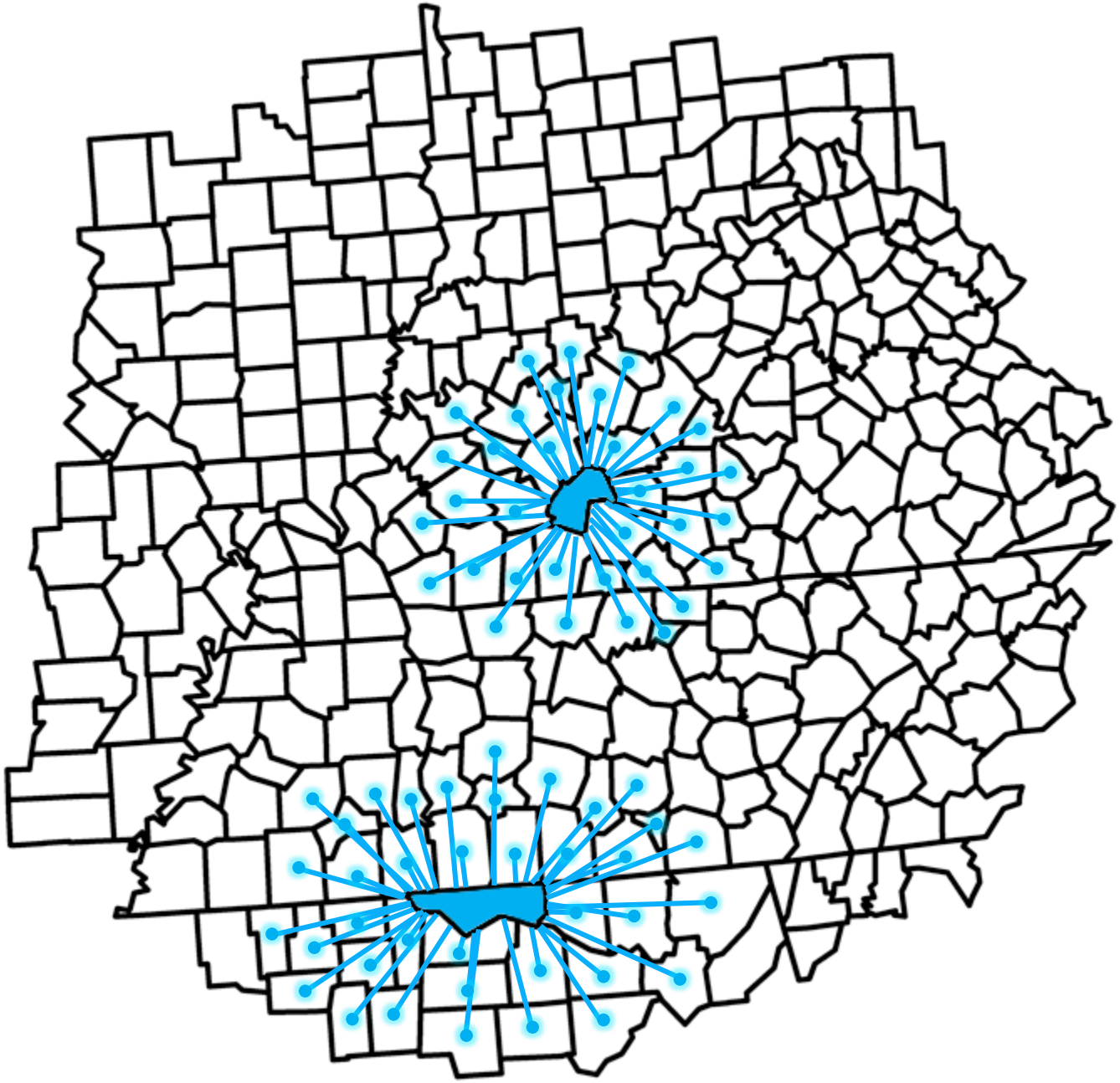
Defining *local*



Defining *local*

Neighborhood Type

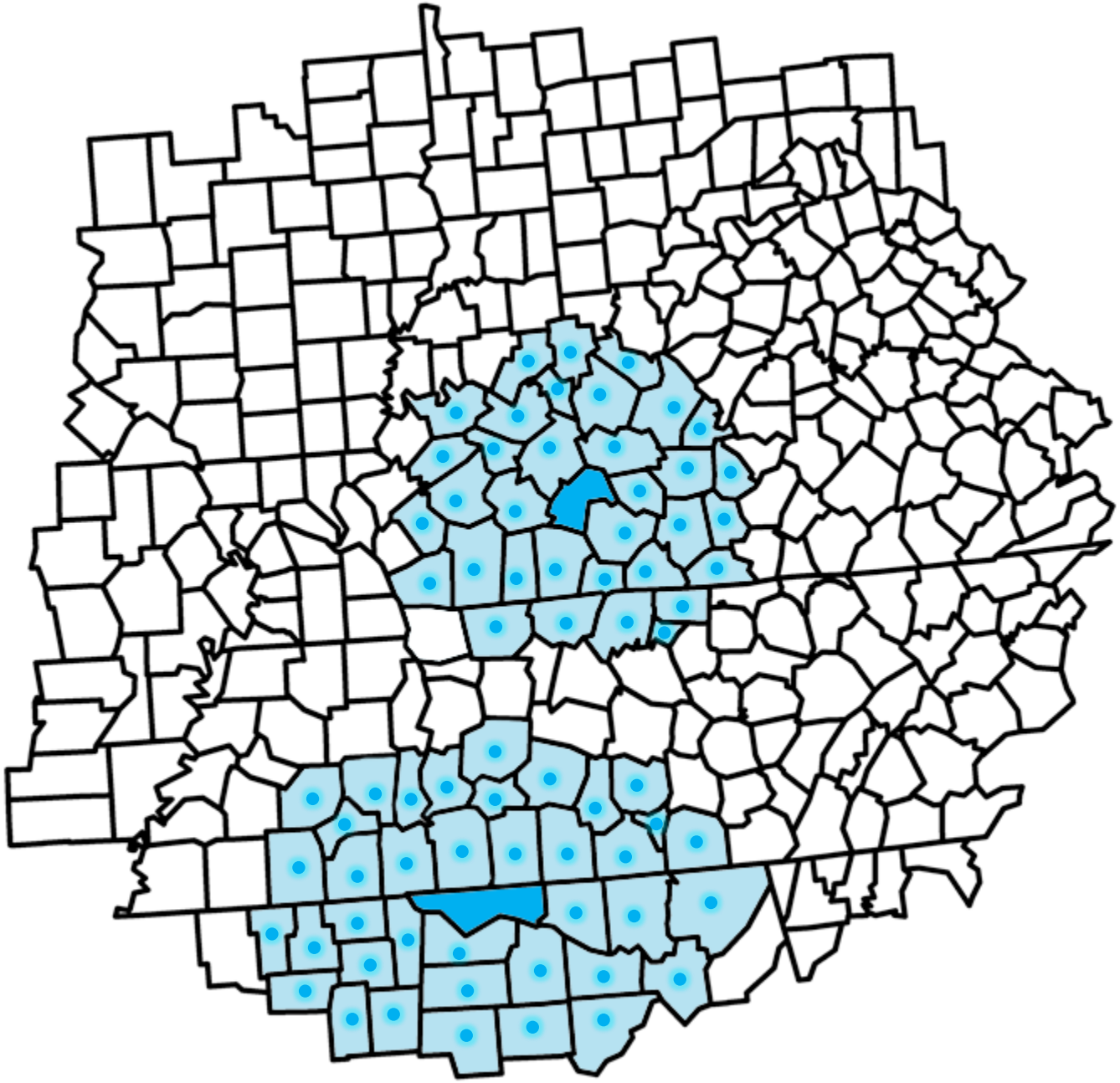
- Number of neighbors



Defining *local*

Neighborhood Type

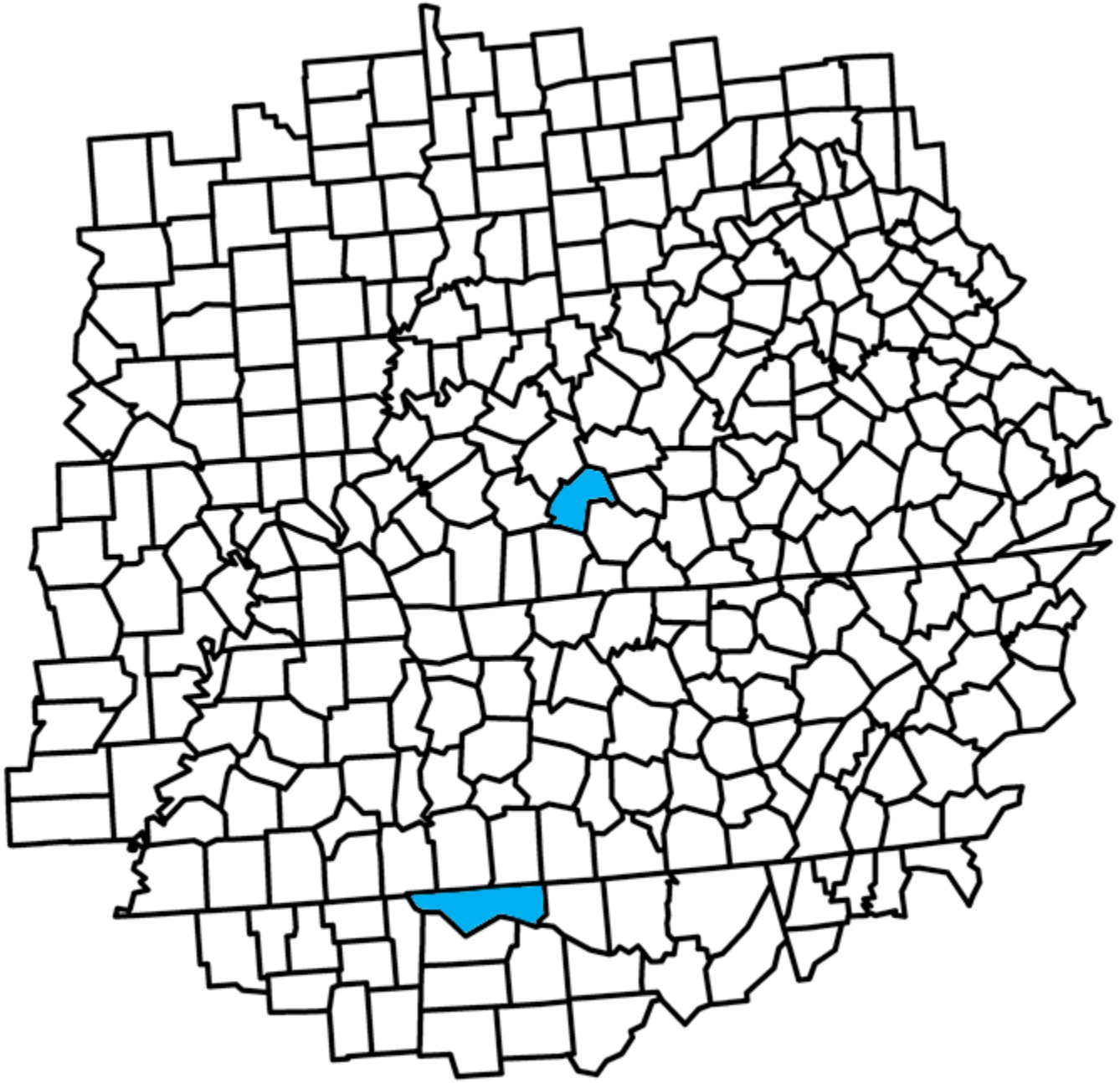
- Number of neighbors



Defining *local*

Neighborhood Type

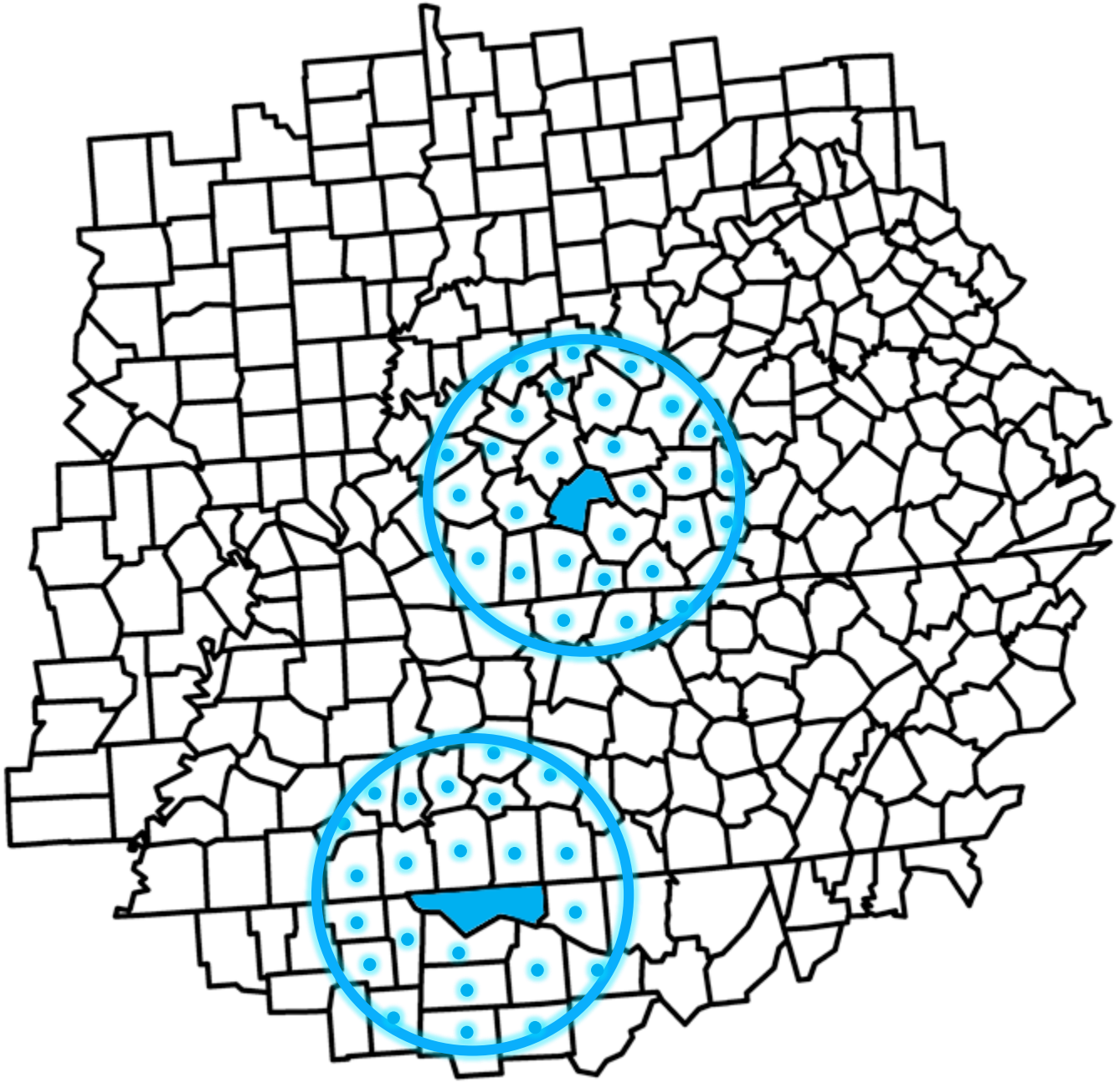
- Number of neighbors



Defining *local*

Neighborhood Type

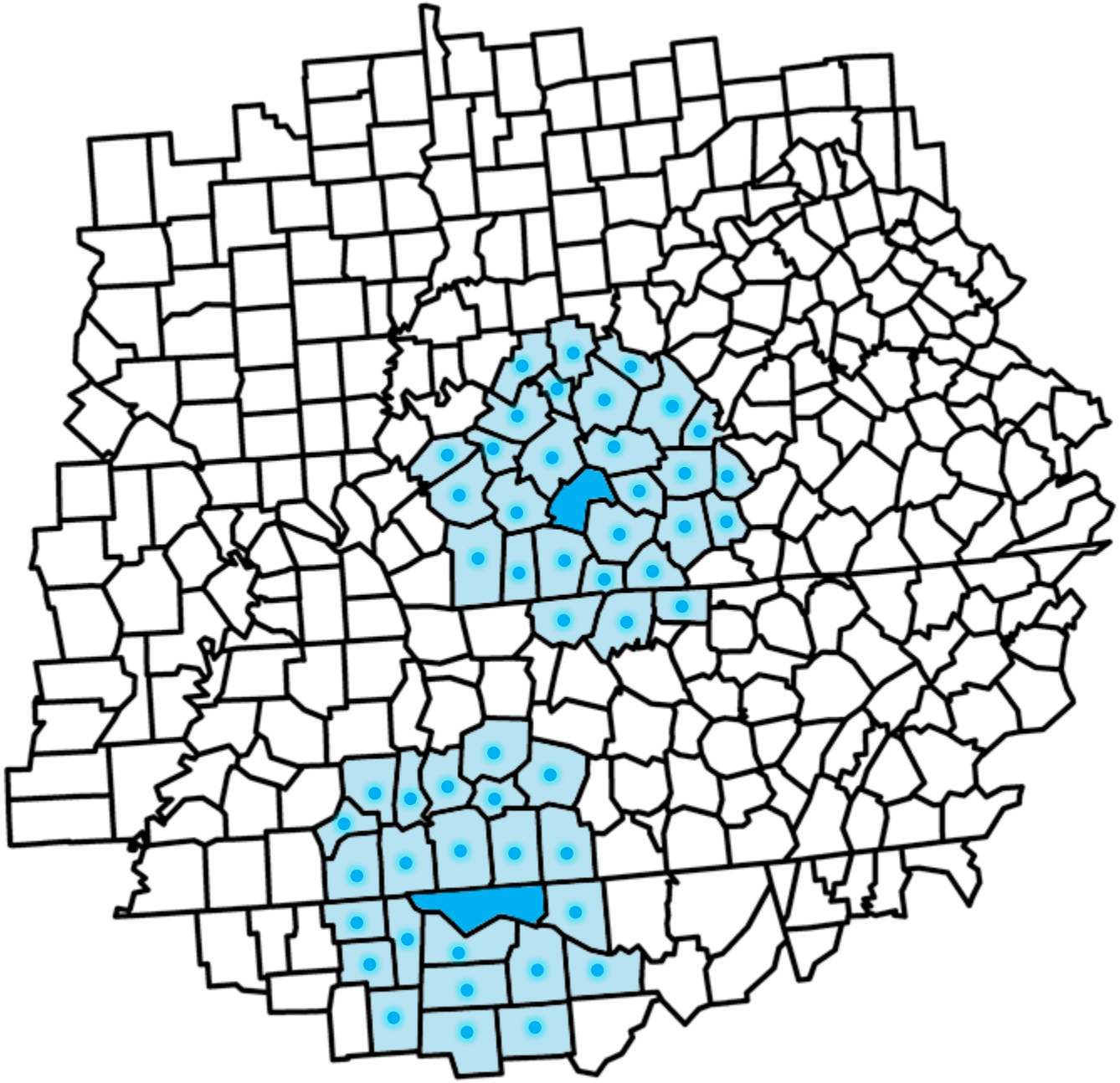
- Number of neighbors
- Distance band



Defining *local*

Neighborhood Type

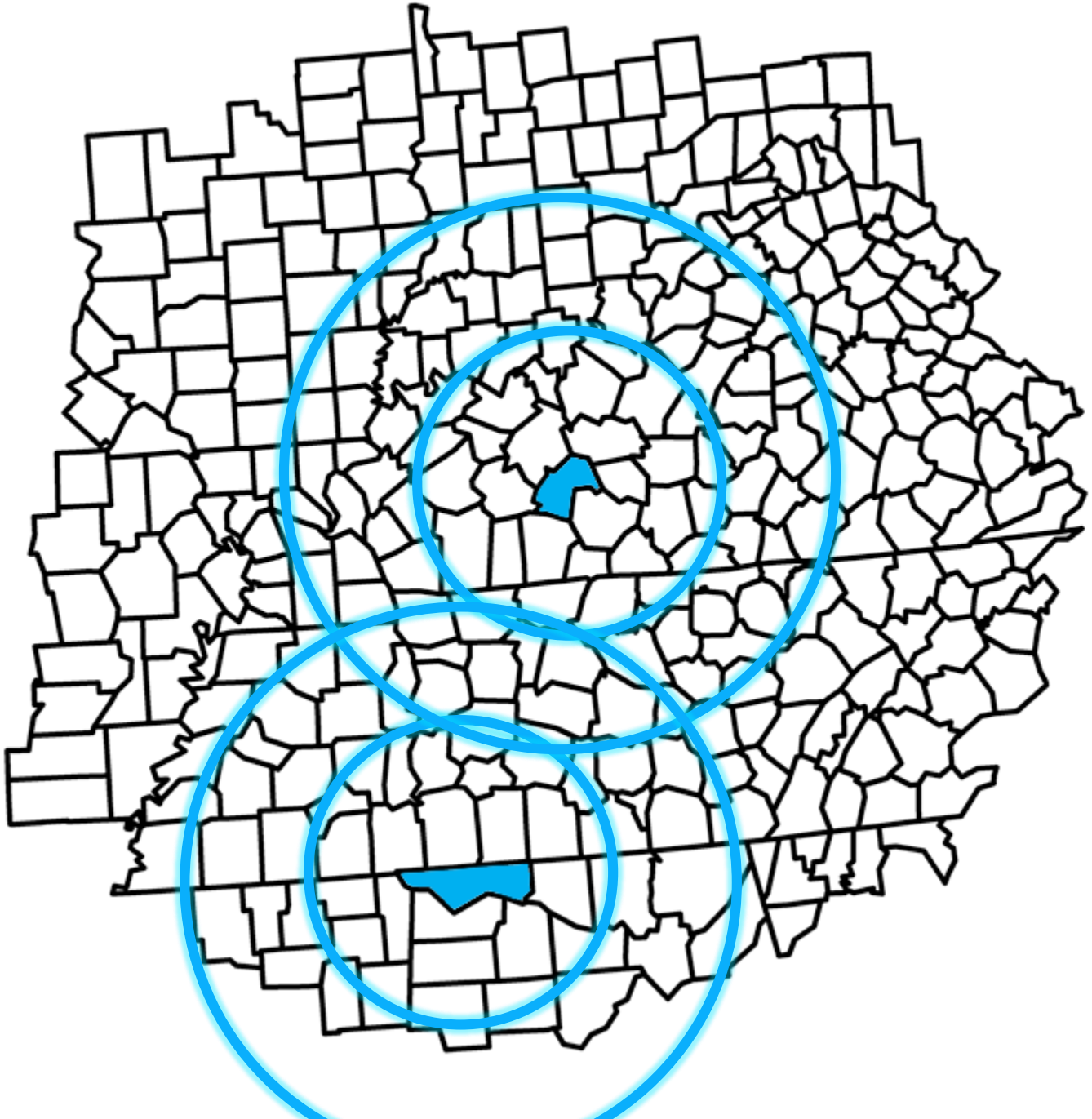
- Number of neighbors
- Distance band



Defining *local*

Neighborhood Type

- Number of neighbors
- Distance band



Defining *local*

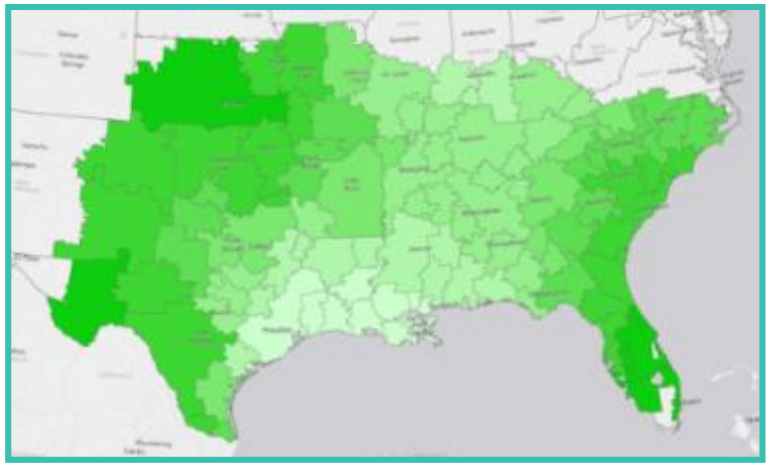
Neighborhood Type

- Number of neighbors
- Distance band

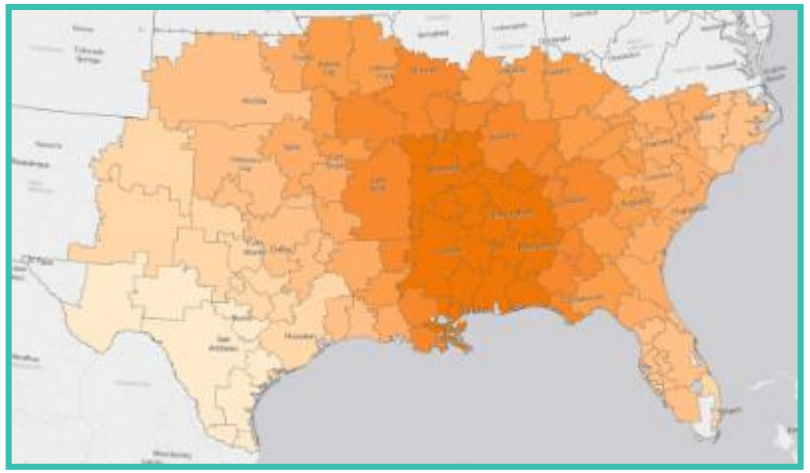
Neighborhood Selection Method

- Golden search
- Manual intervals
- User defined

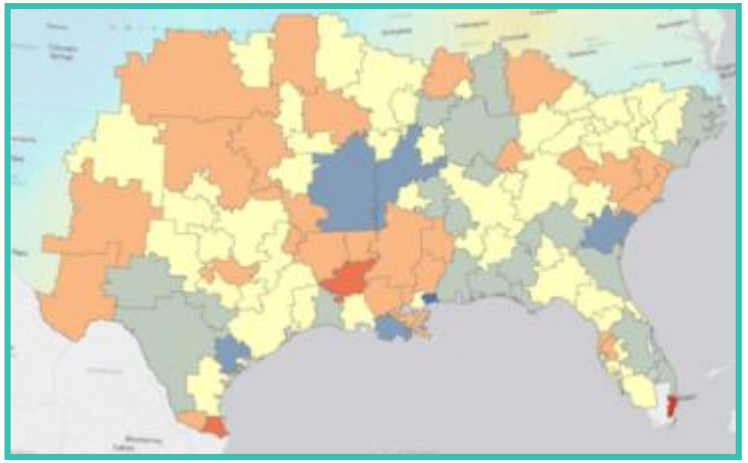
Demo



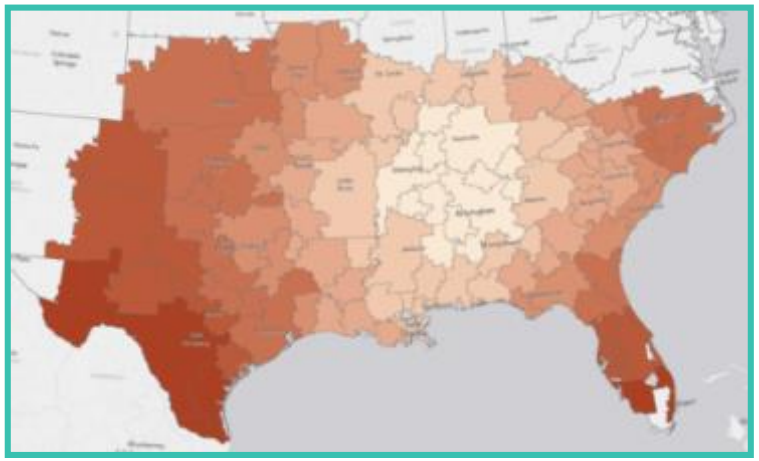
Local R-Squared



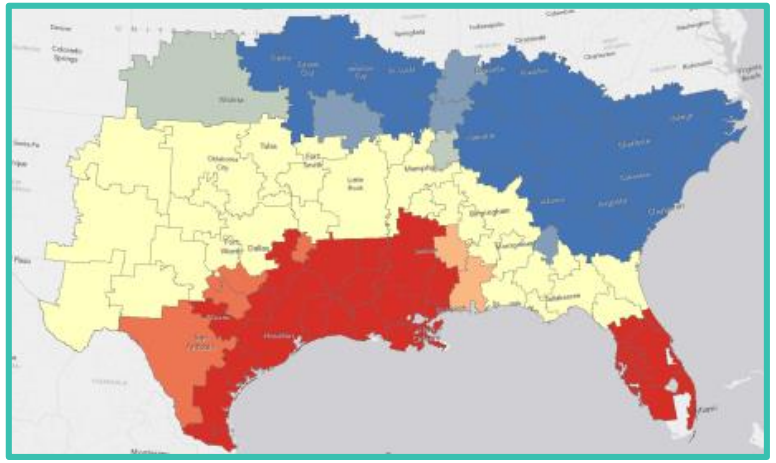
Coefficients



Residuals



Condition Number

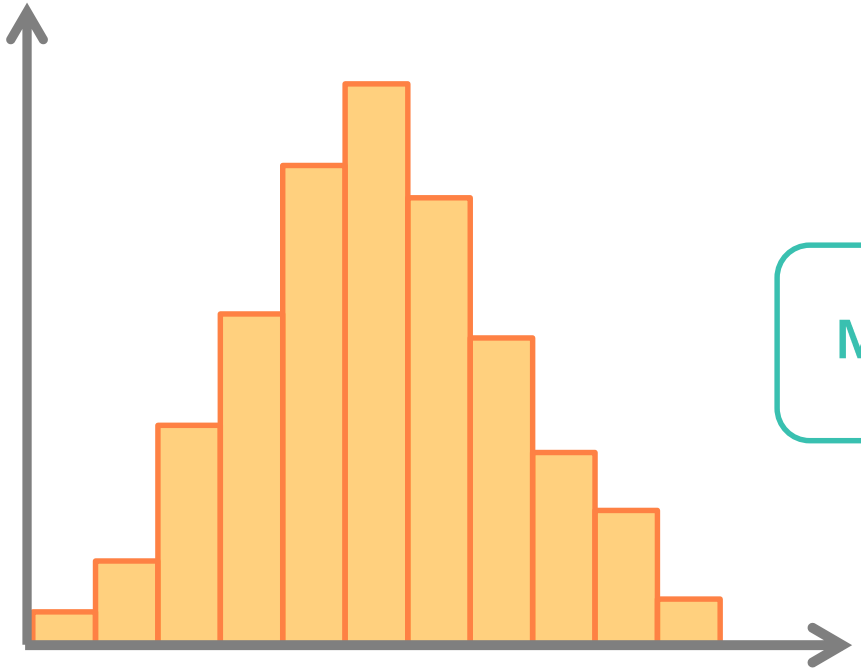


Predictions

Three model types

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

Gaussian – model a continuous variable



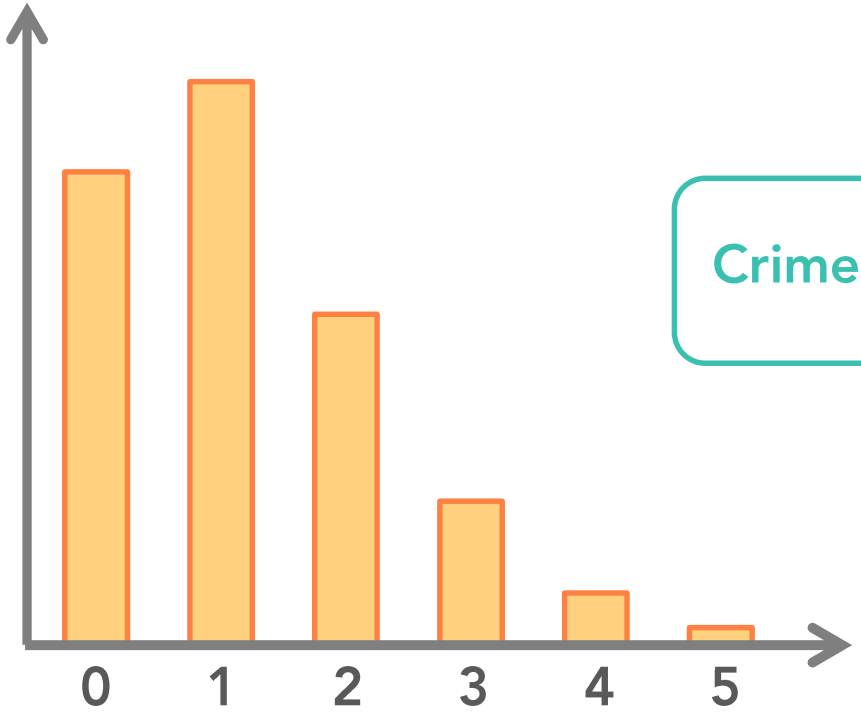
Mortality rate

Sales profits

Temperature

Healthcare spending

Poisson – model a count variable



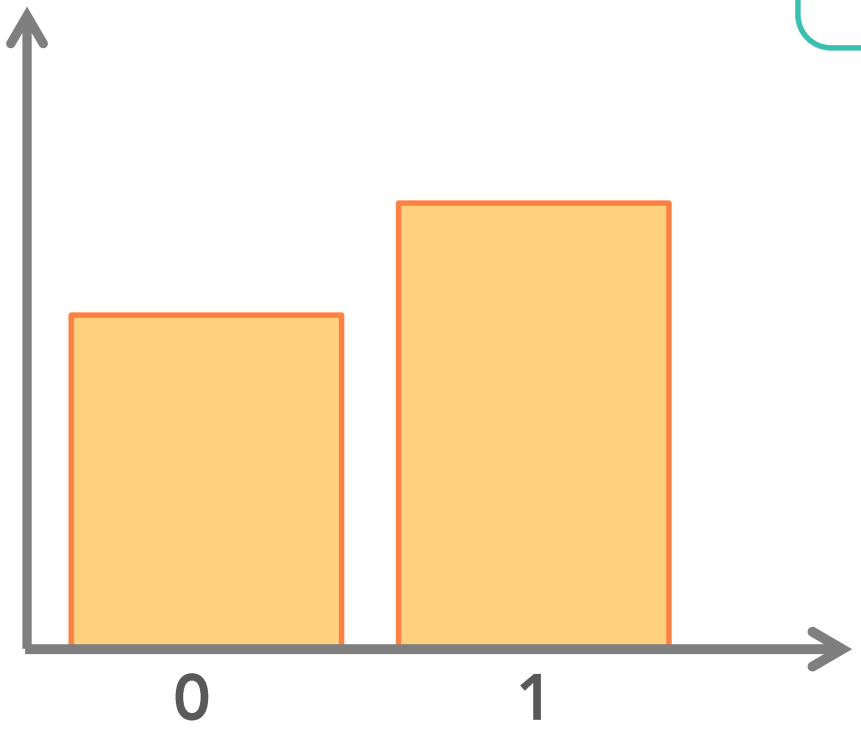
Traffic accidents

Number of people with cancer per 10,000

Crime counts

Sales per month

Logistic – model a binary variable



Disease presence

Fire damage

Insurance fraud

Pass/fail inspection

Demo

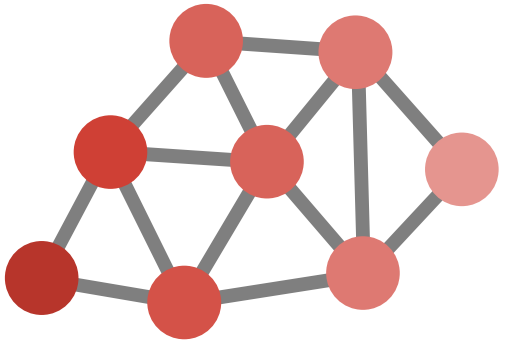
Local Bivariate Relationships

Examining relationships across space

"… the measurement
of a relationship

depends on where

the measurement
is taken."

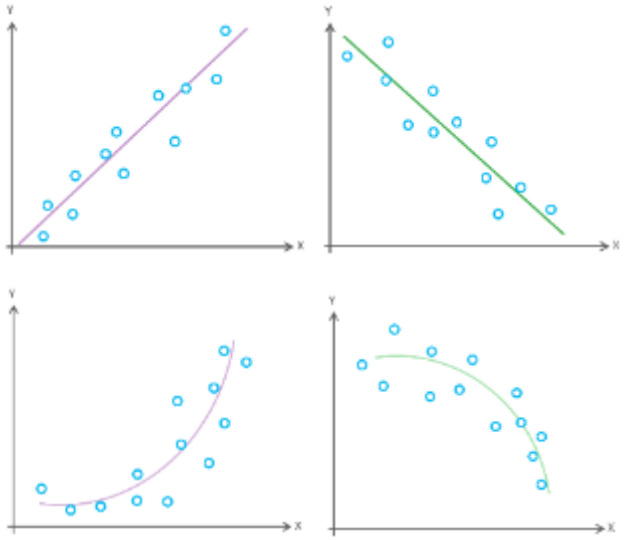


two variables

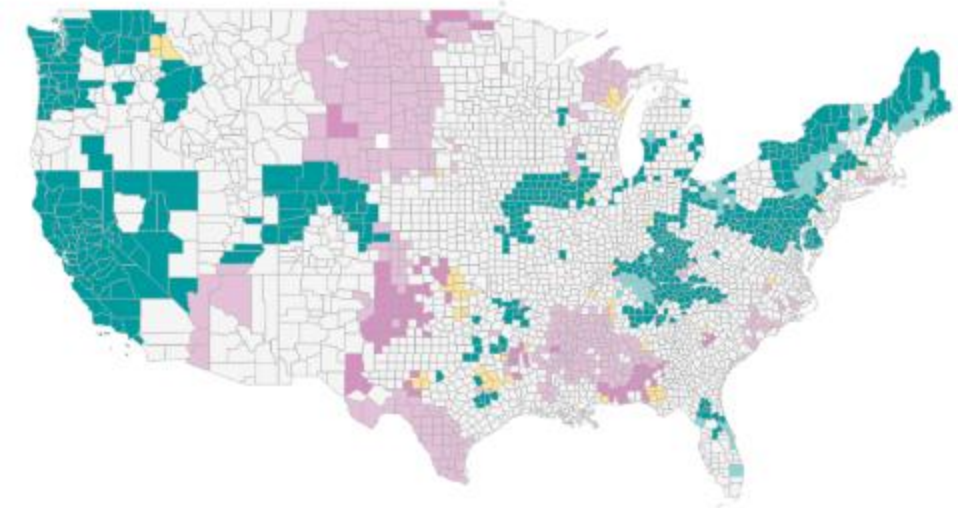
Y

X

determine relationship
significance and type

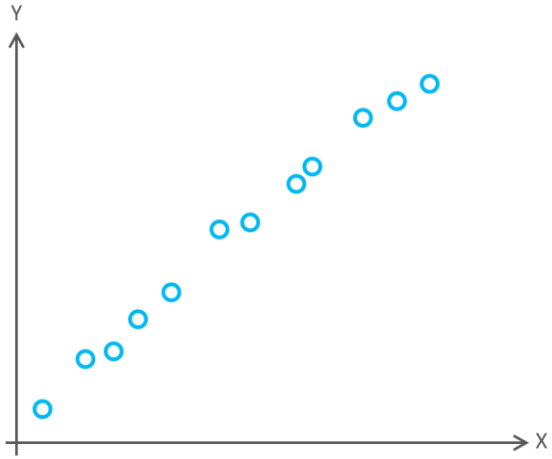


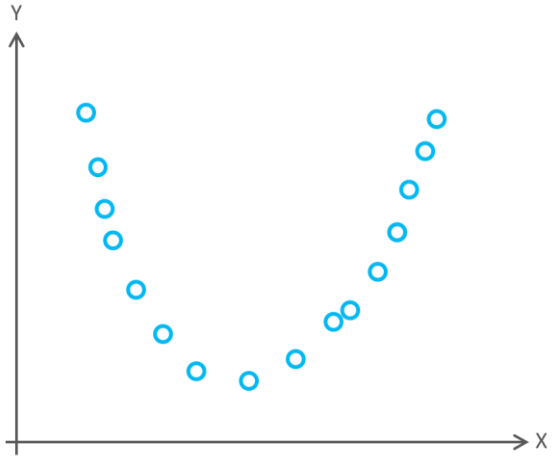
relationships across geography



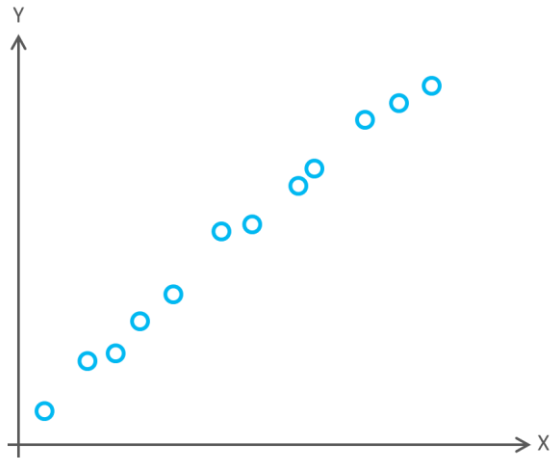
… what does it mean for two variables to be related to each other?

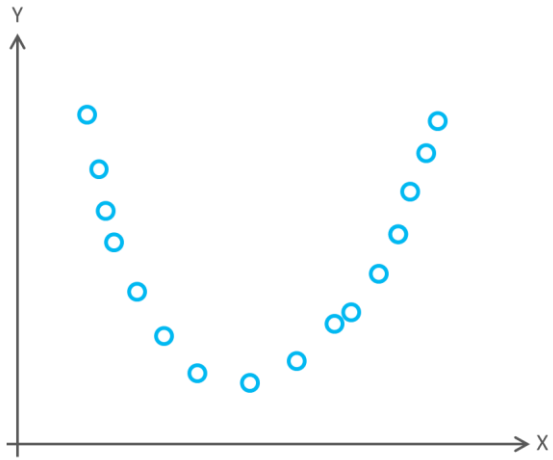
"good" relationships



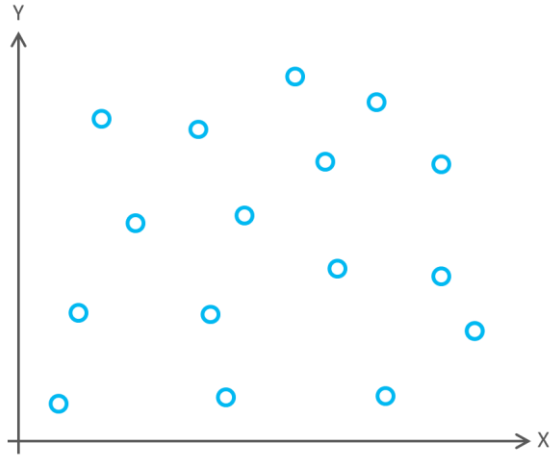


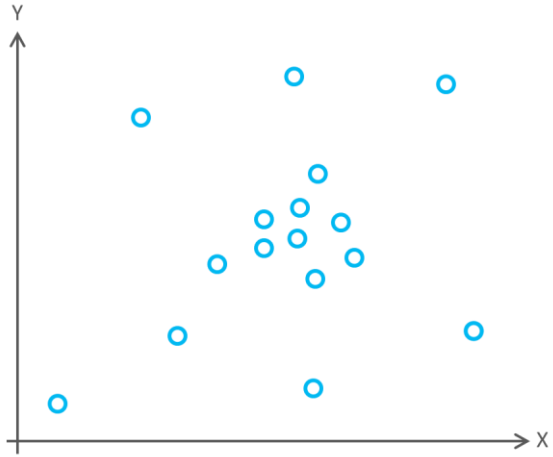
"good" relationships



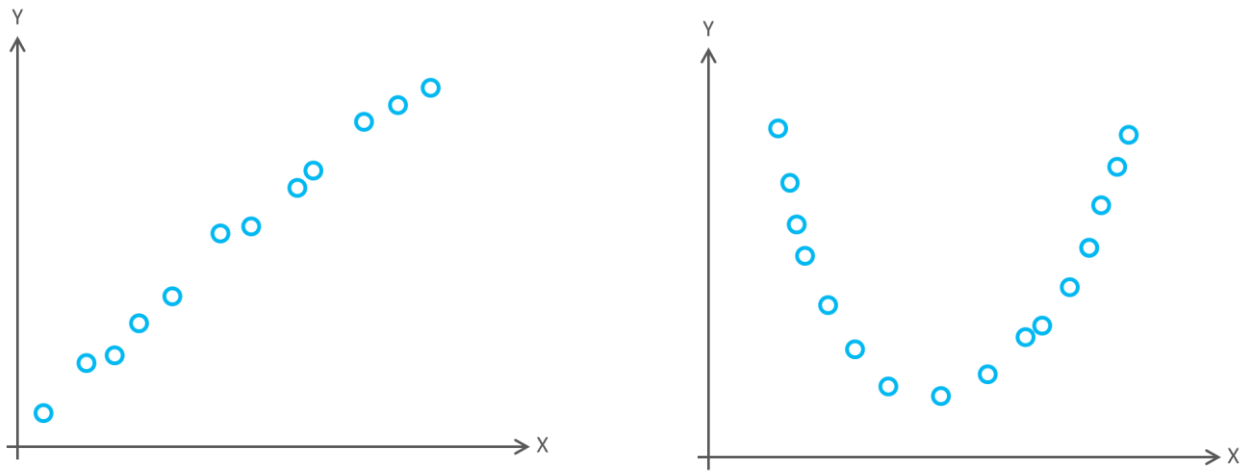


null relationships



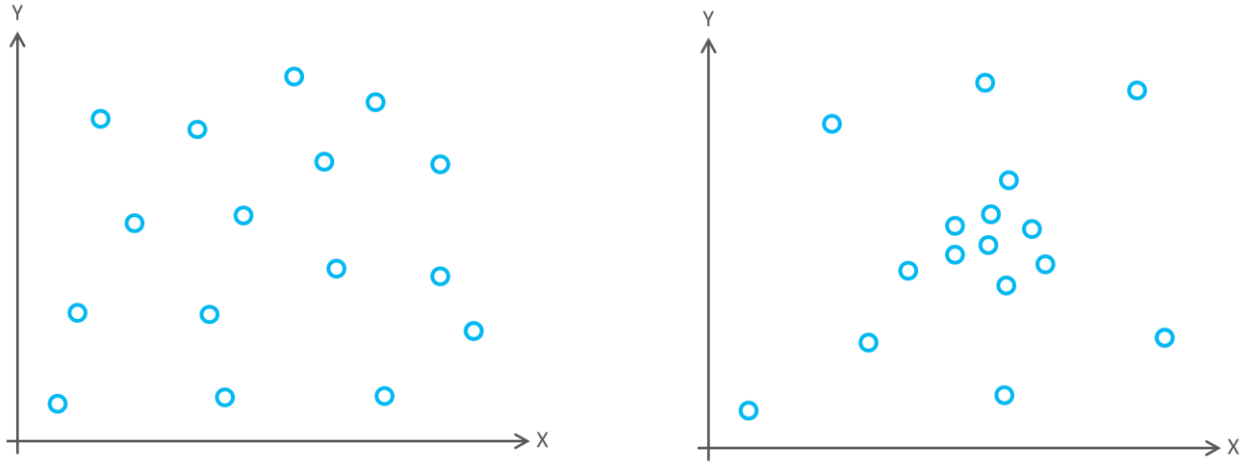


"good" relationships



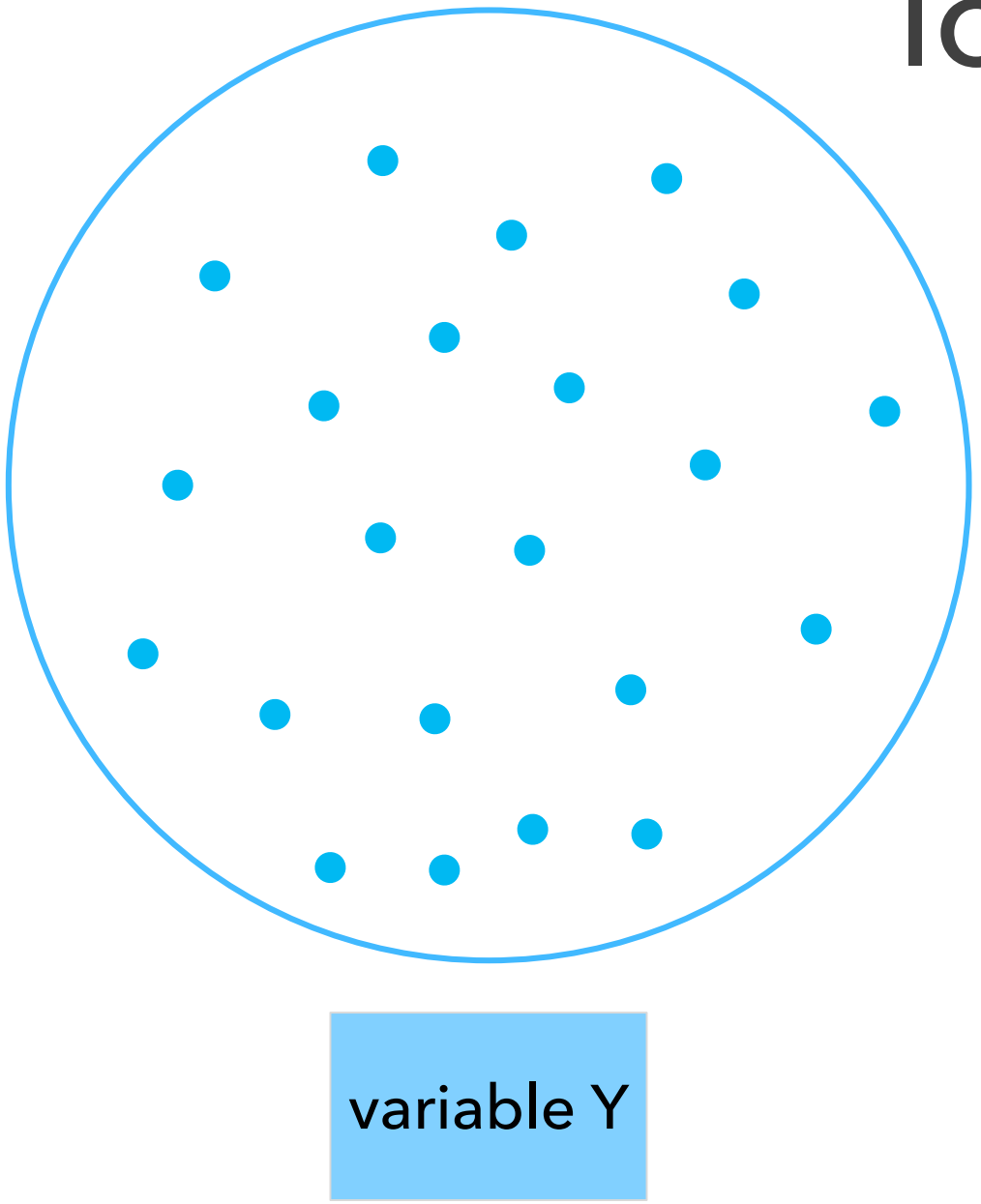
→ low entropy

null relationships

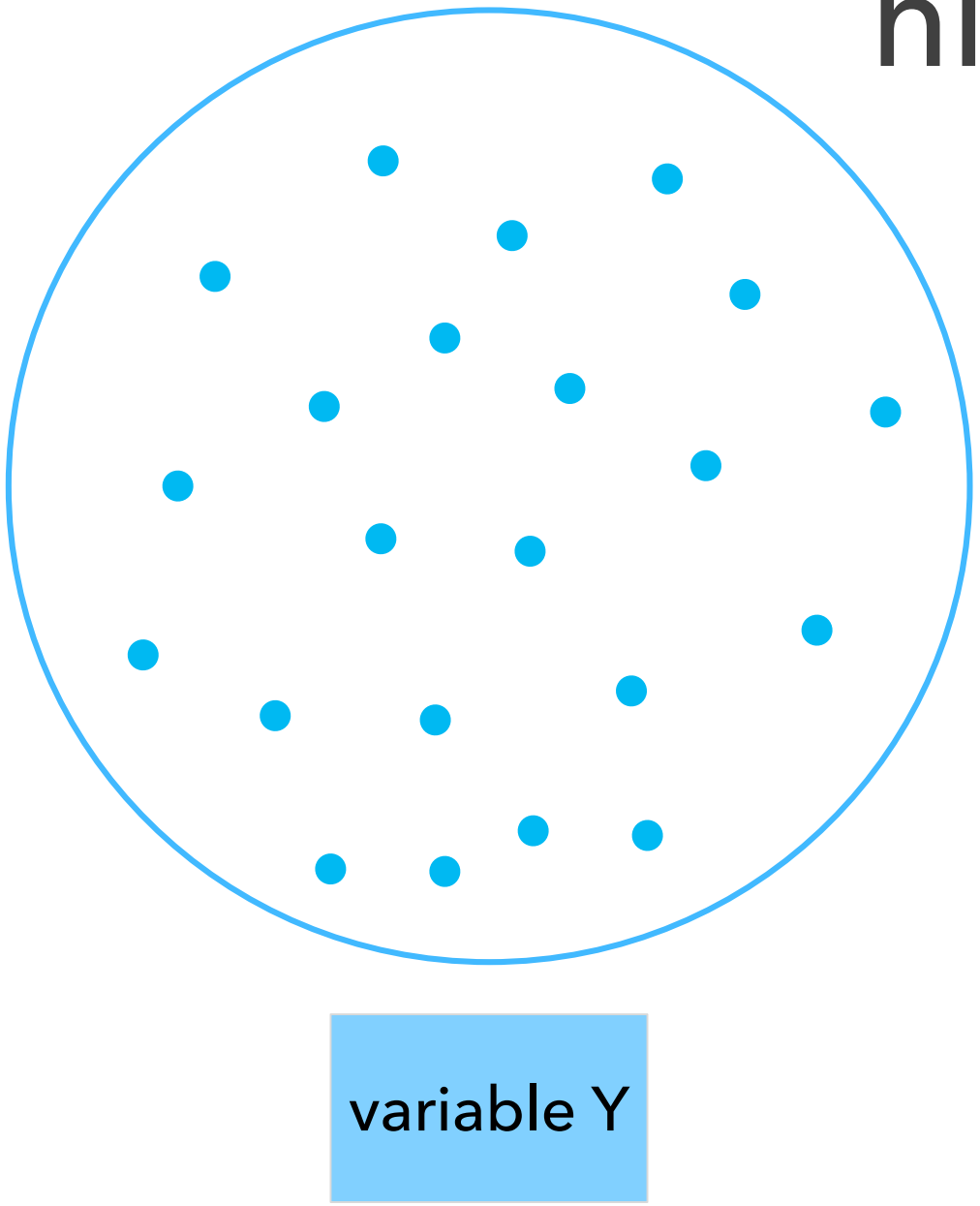


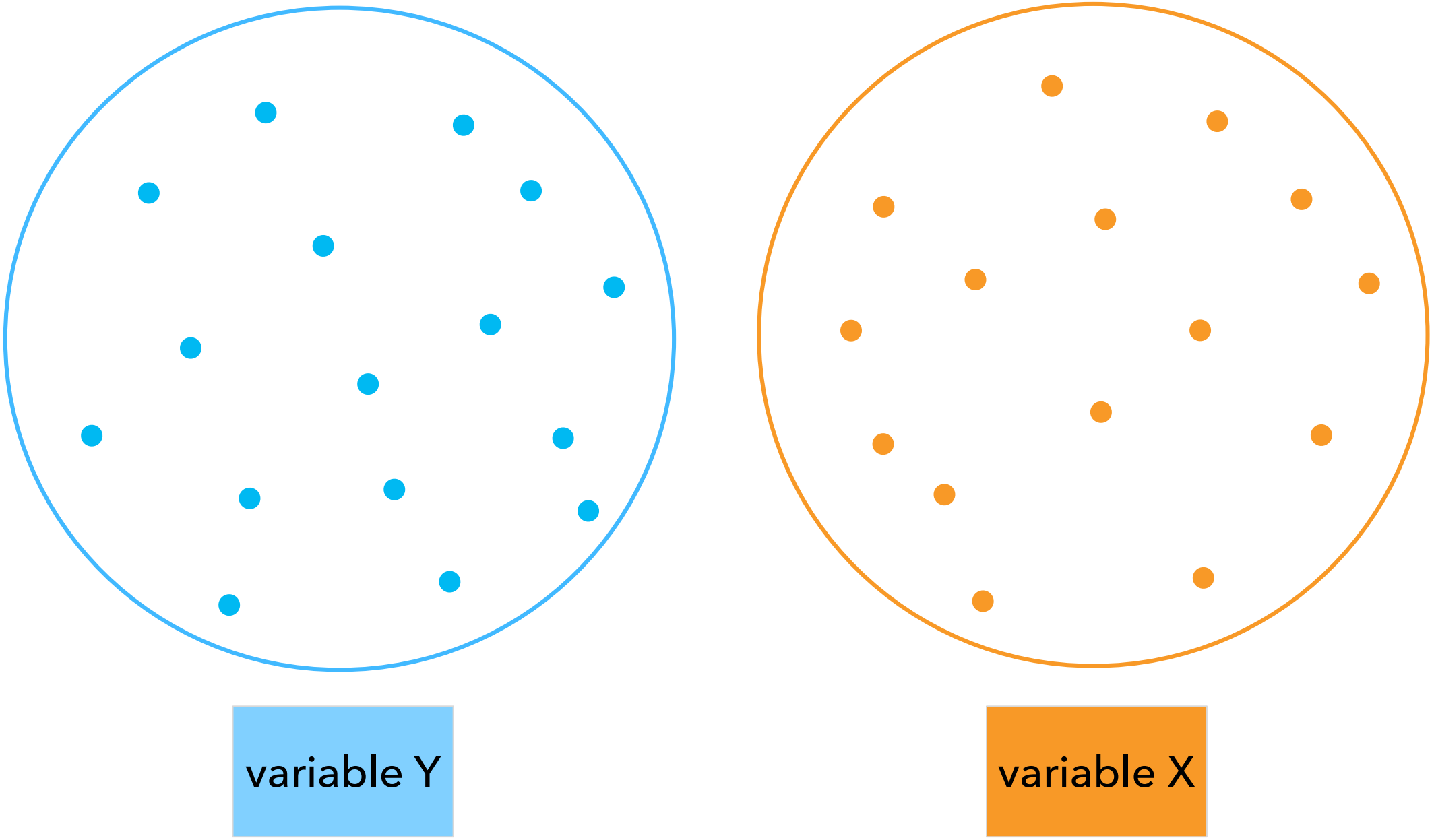
→ higher entropy

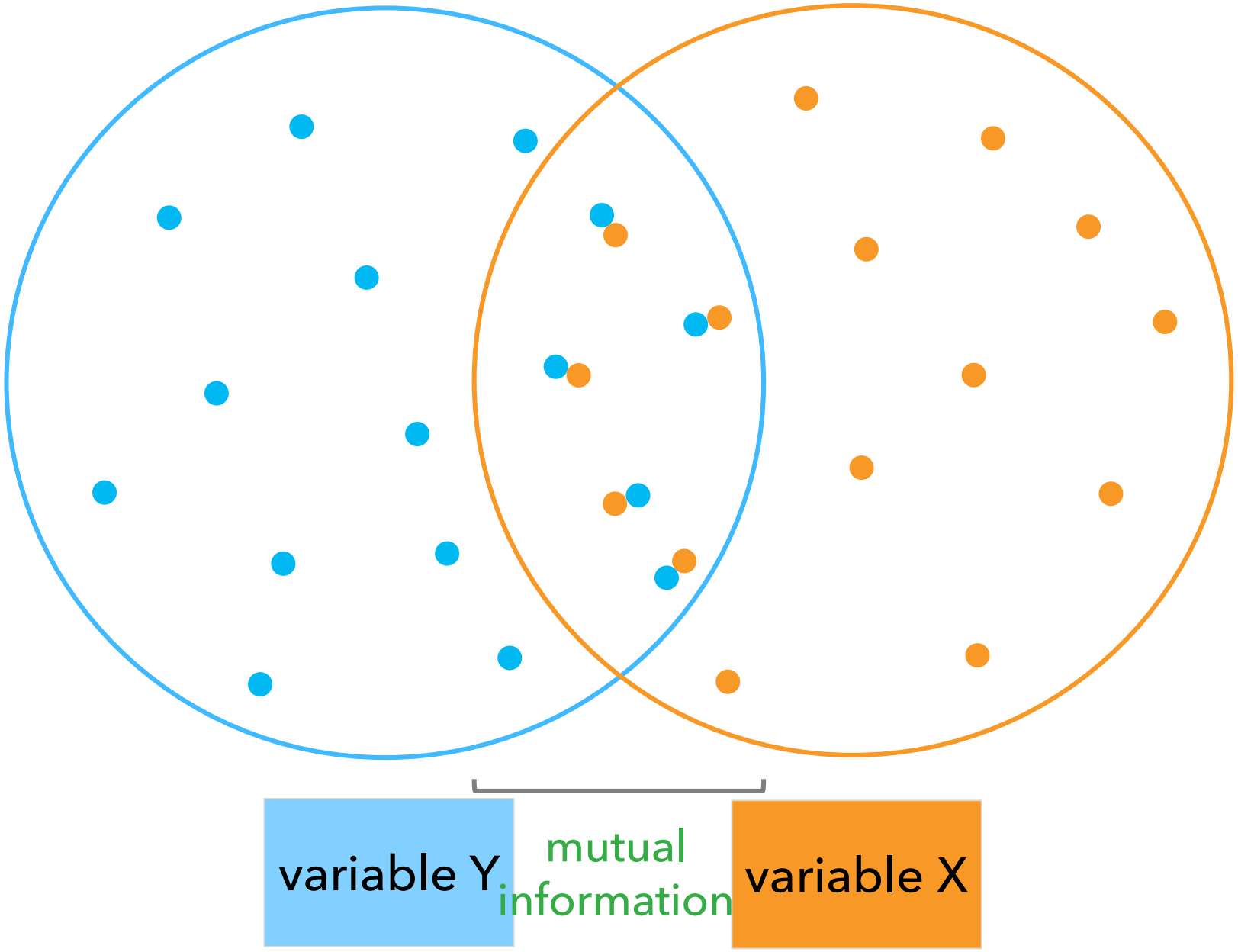
low entropy



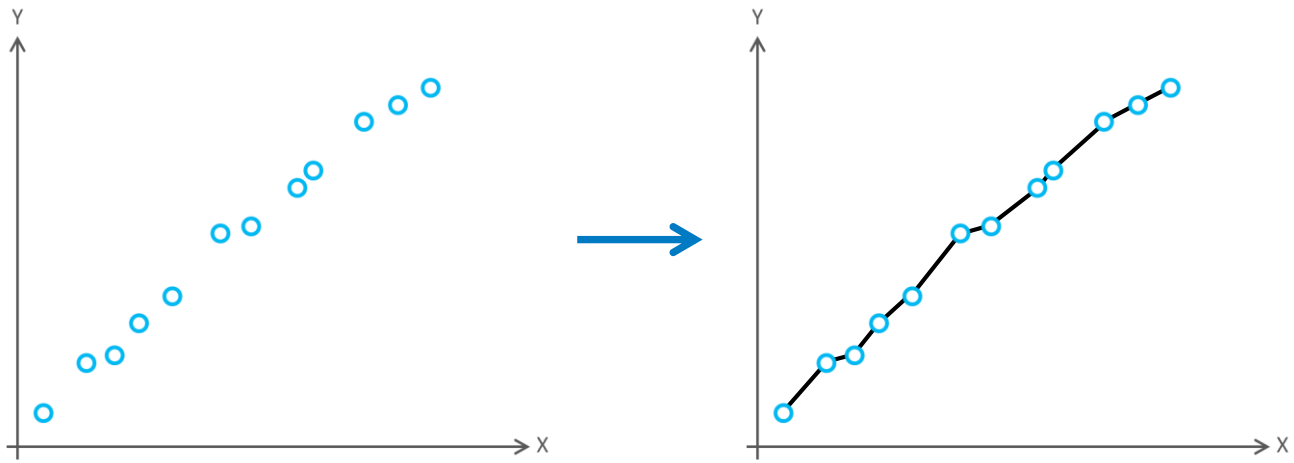
high entropy



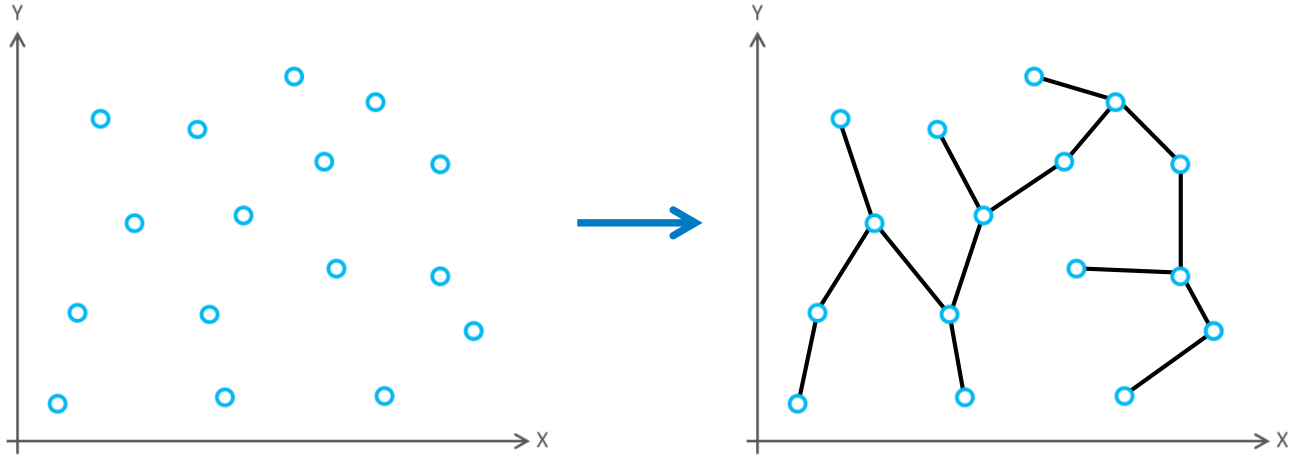

variable Y
variable X


variable Y
mutual information
variable X

minimum spanning trees



low entropy

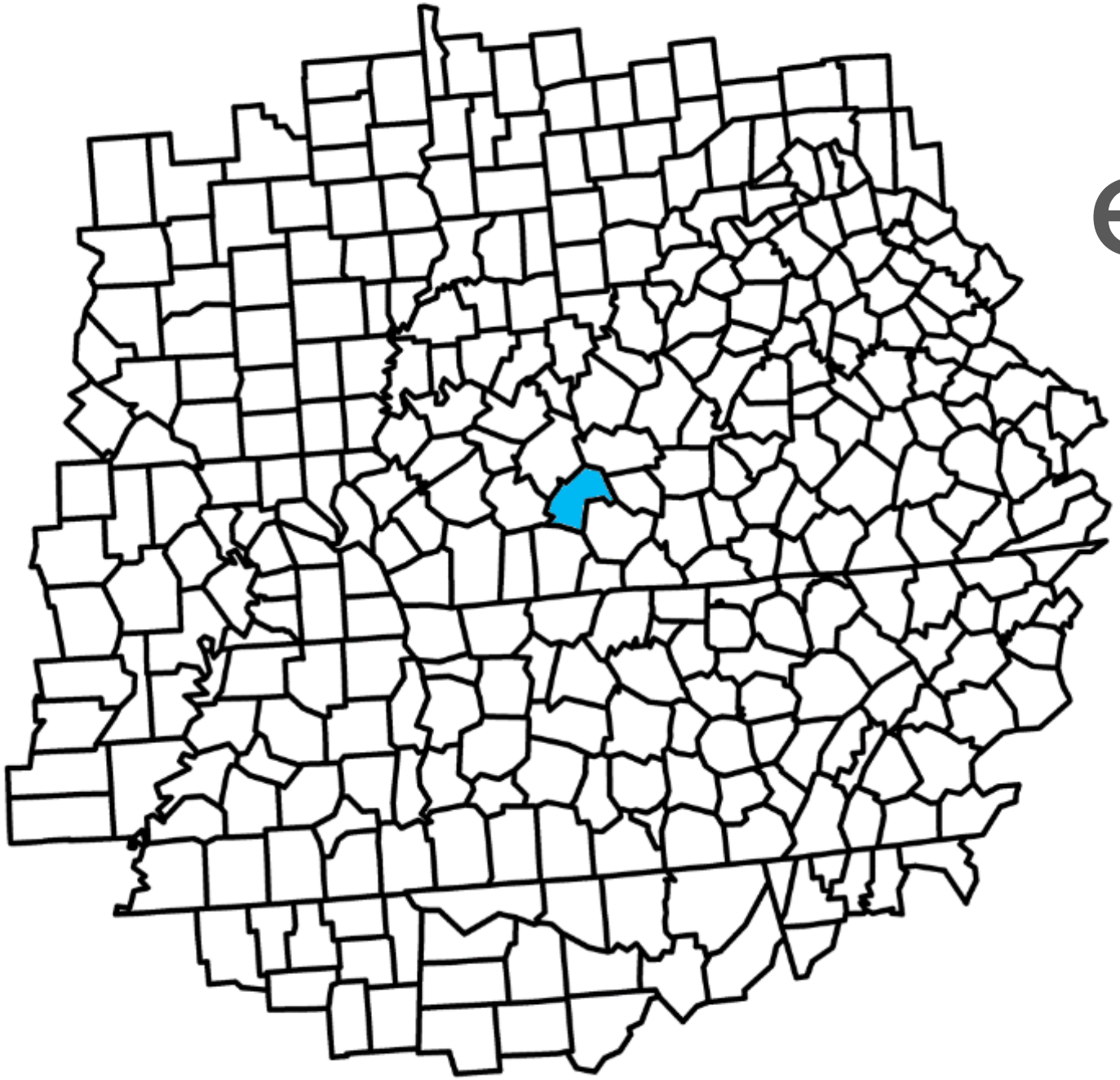


higher entropy

…how do we know if the relationship is SIGNIFICANT???

Permutation-based distribution estimation

feature



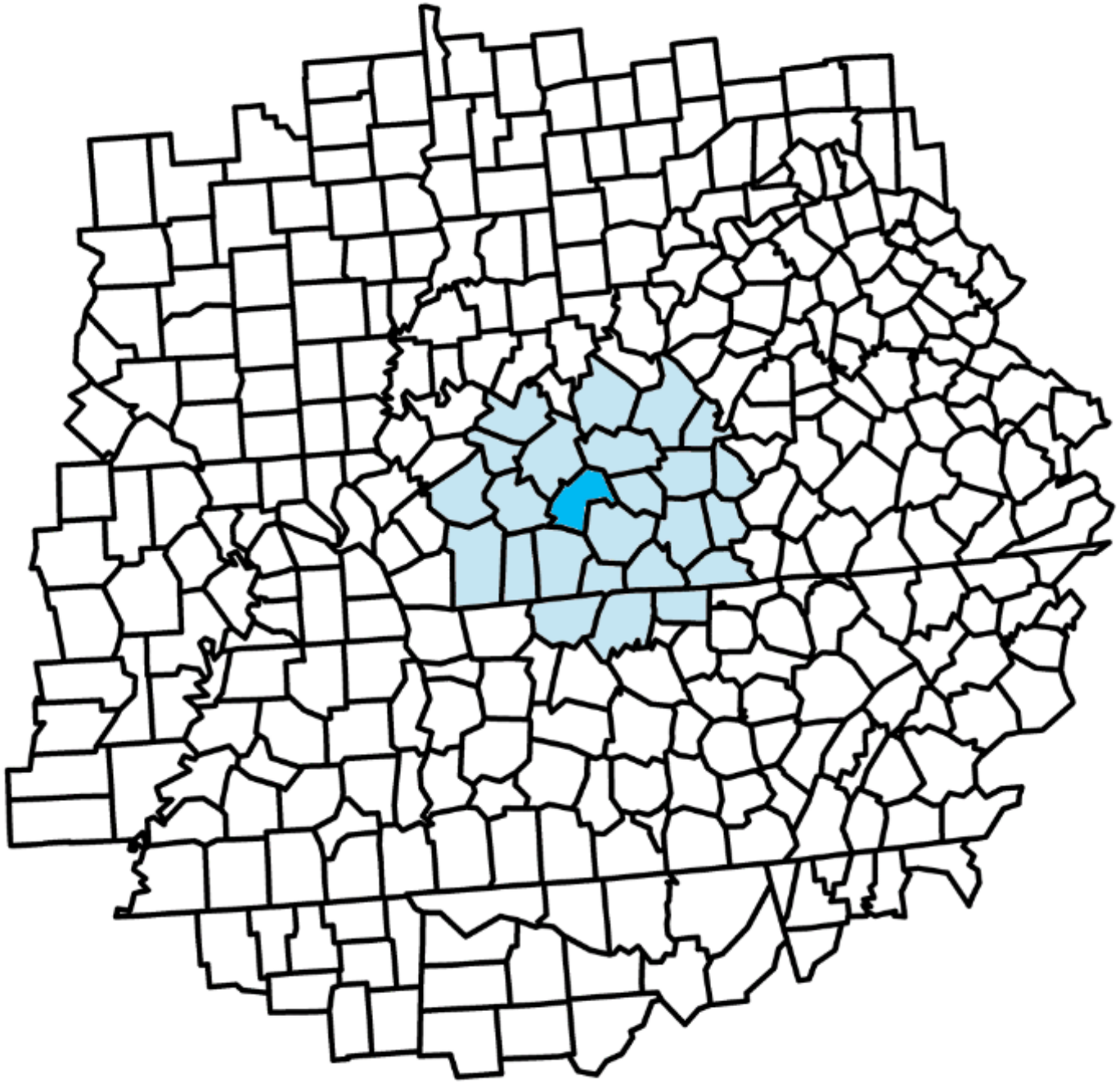
each feature has two values

Y

dependent

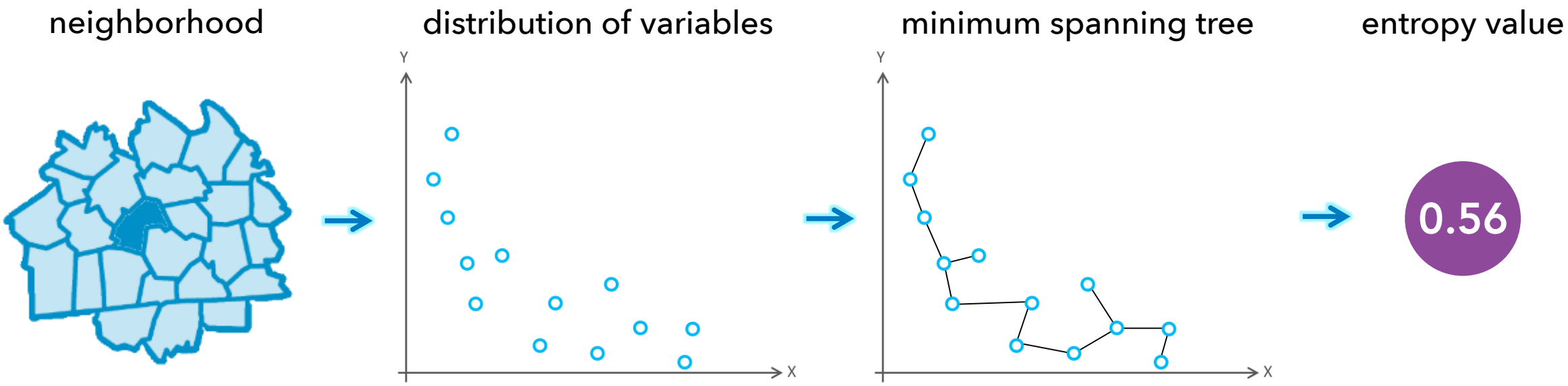
X

explanatory



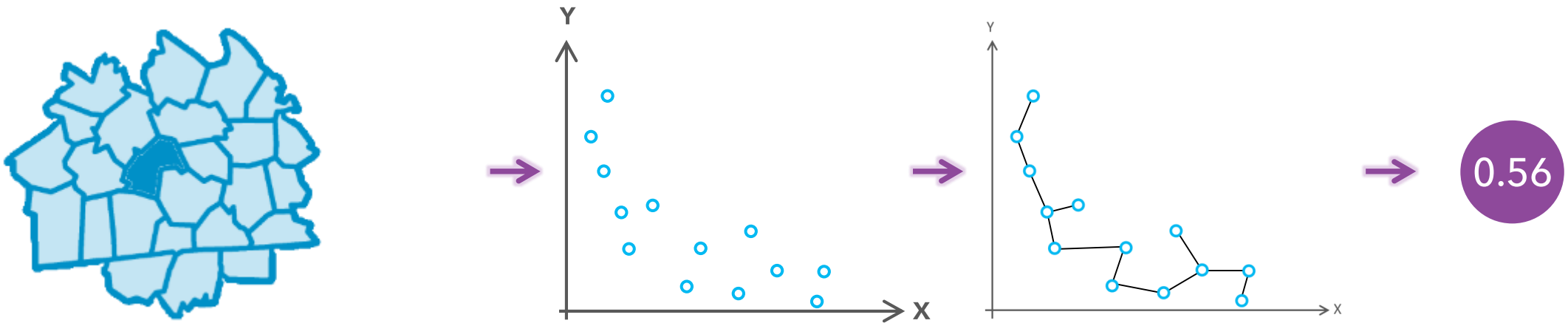
relationship between Y and X is evaluated for a neighborhood

relationship
between Y and X
is evaluated for a
neighborhood

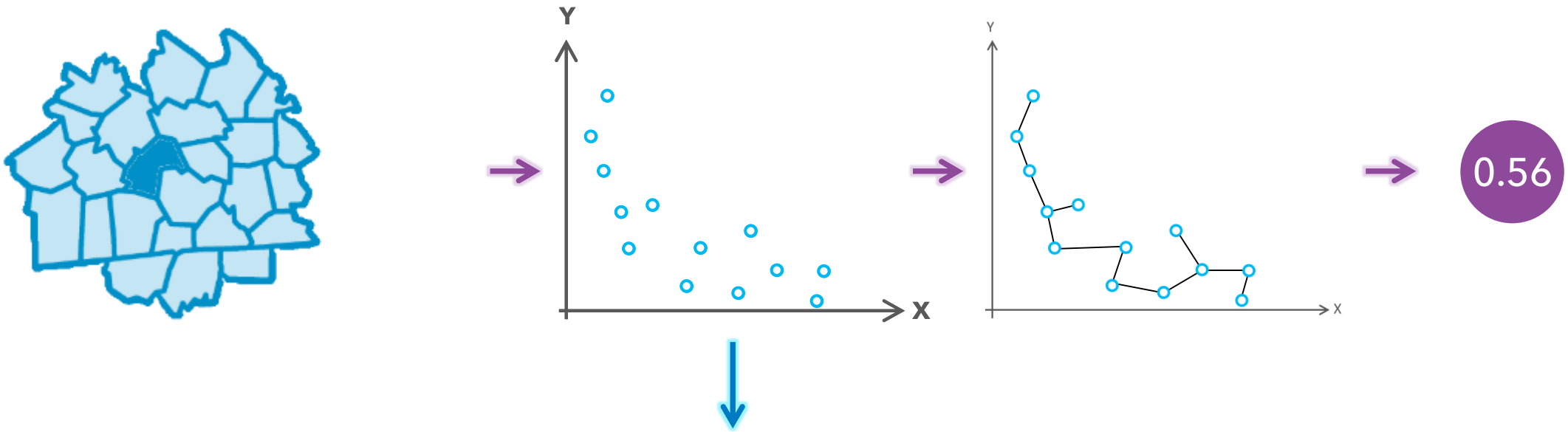

neighborhood
distribution of variables
minimum spanning tree
entropy value
Y
X
Y
X
0.56

what is the probability that the observed entropy exists while Y and X are actually independent from each other?

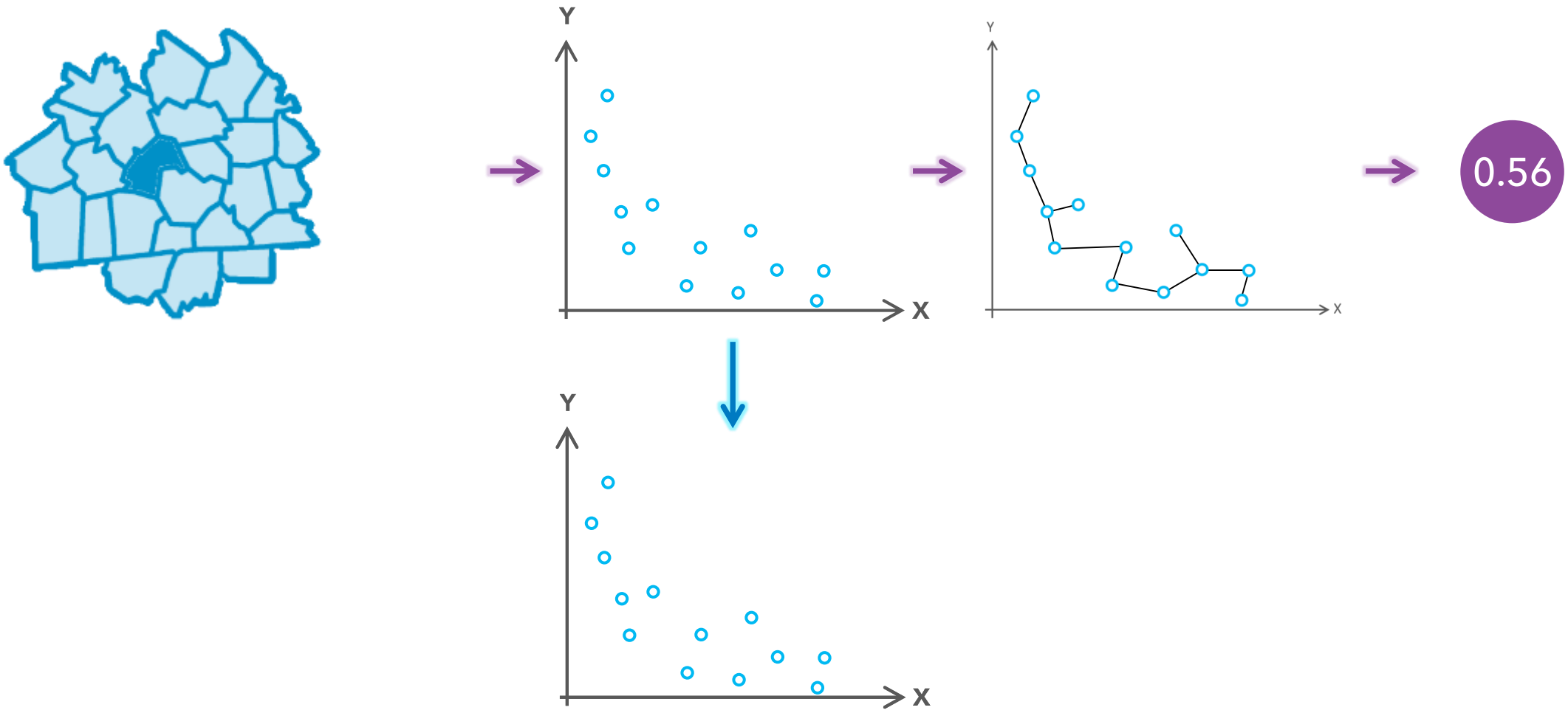
observed local entropy



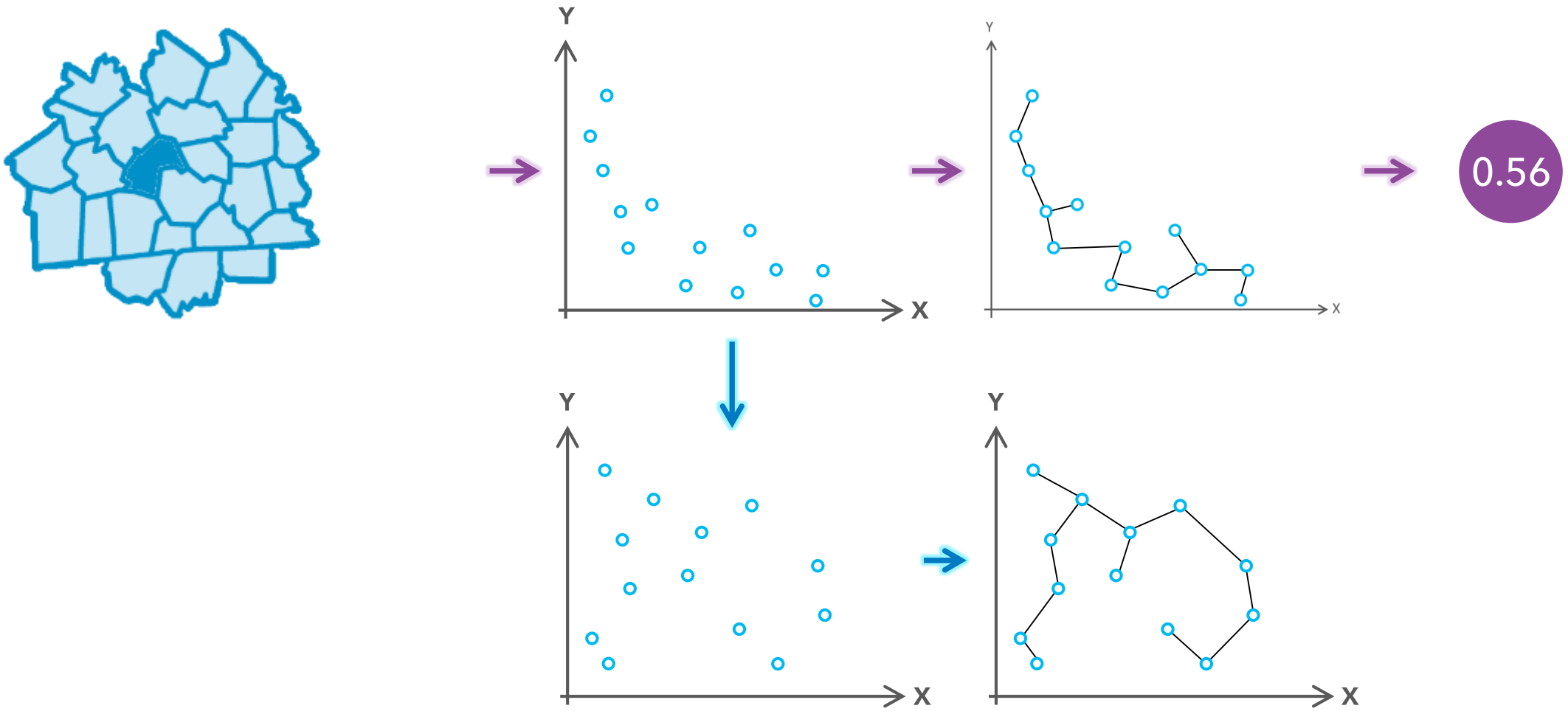
start a permutation



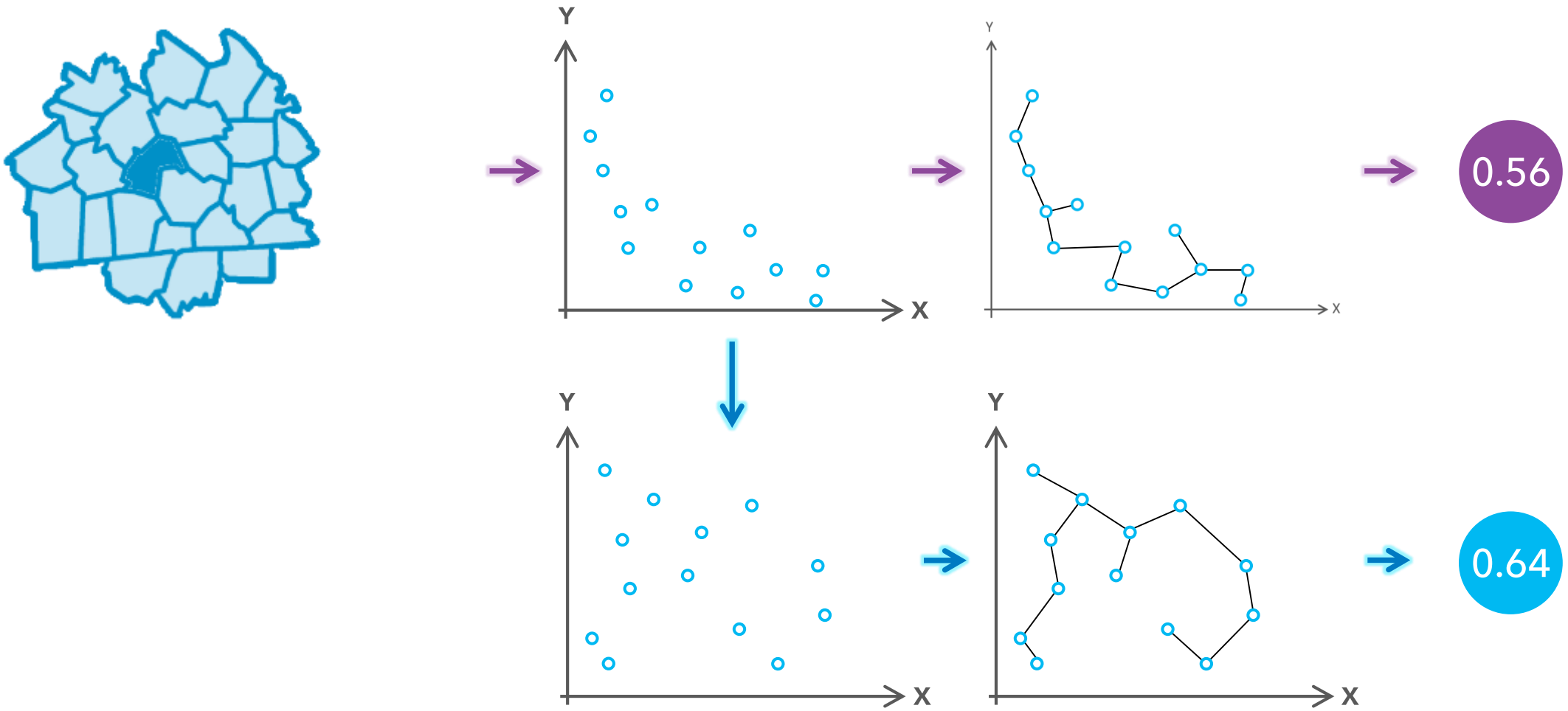
the dependent values Y are shuffled, while explanatory values X are kept the same

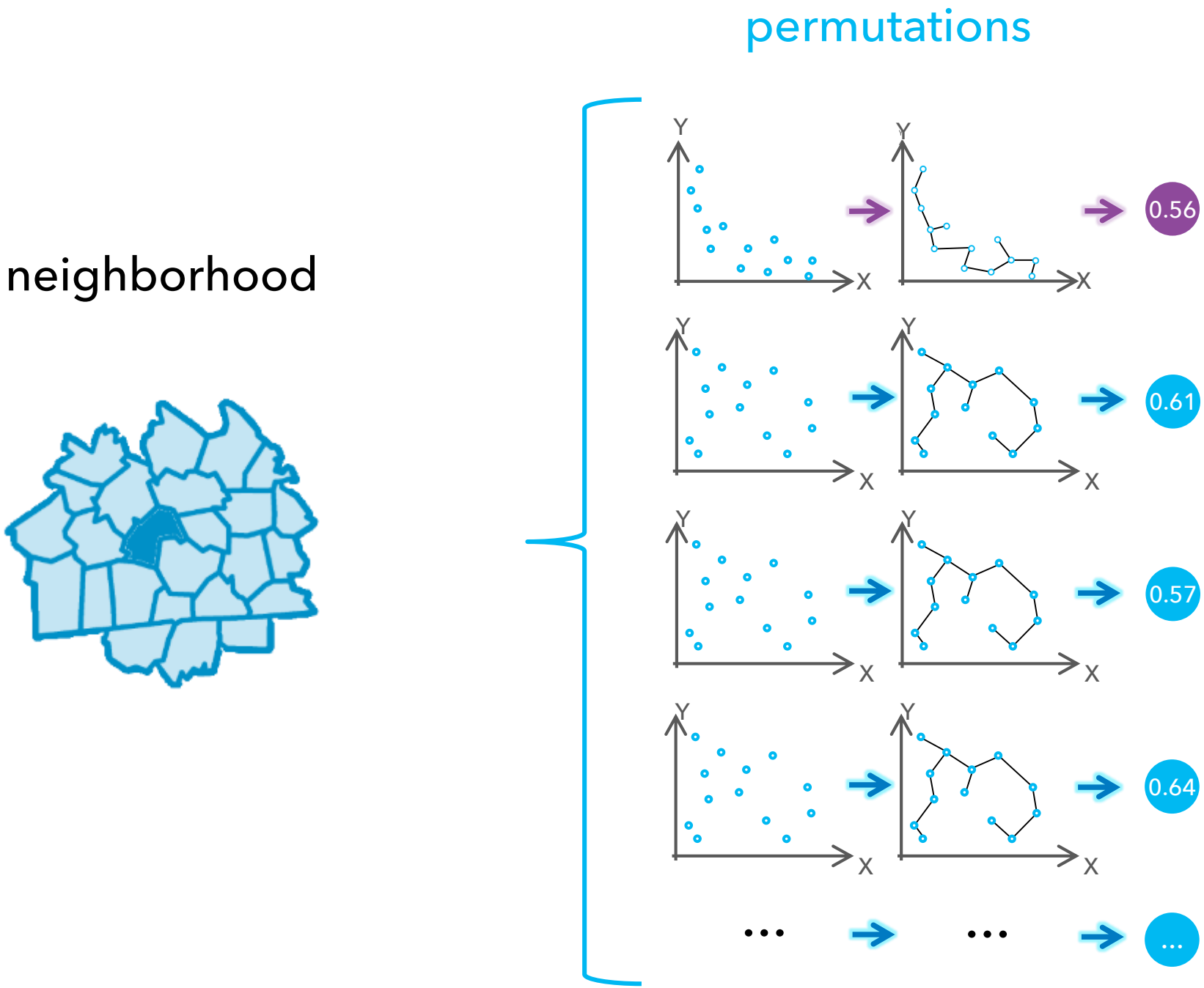


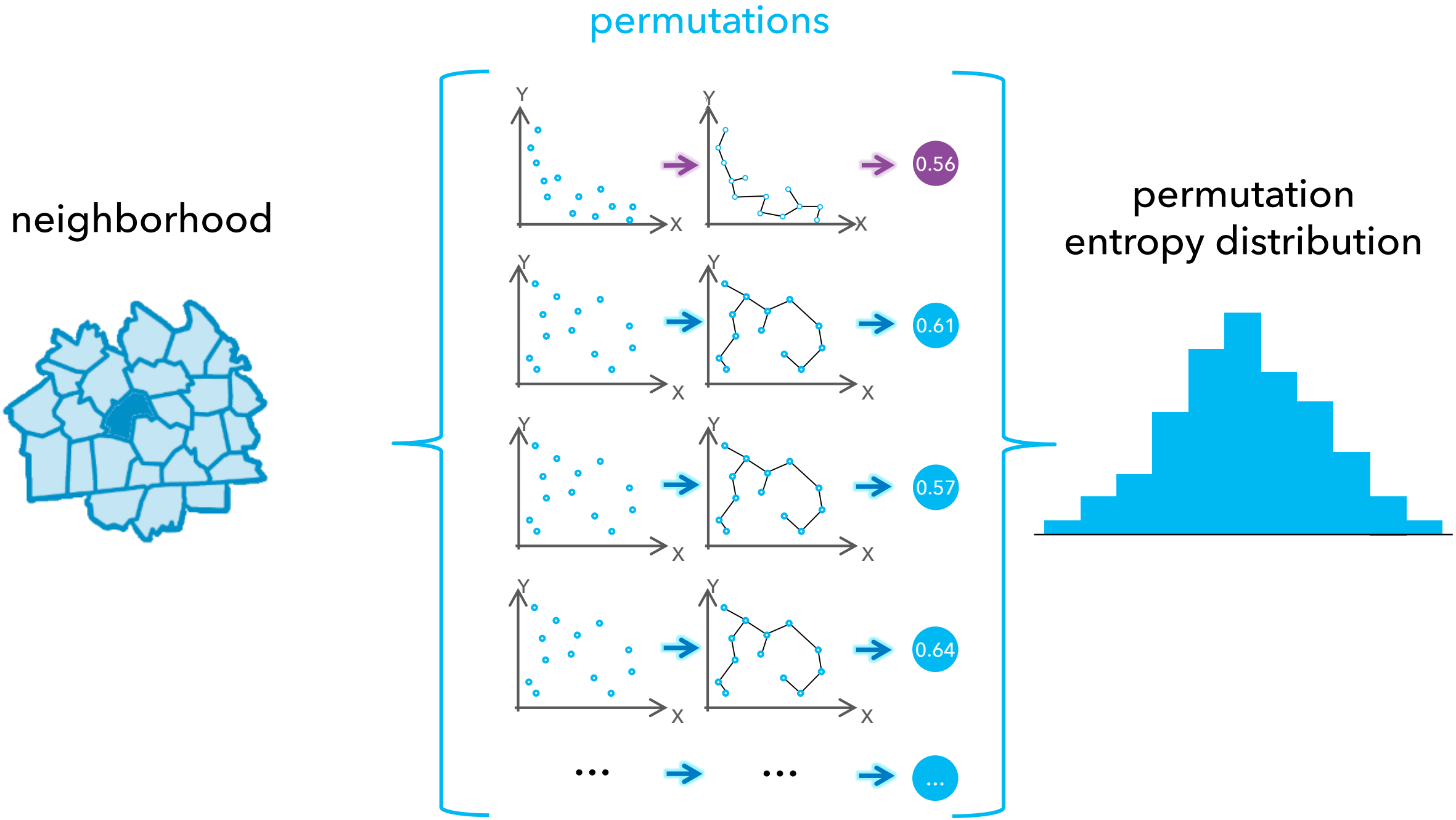
minimum spanning trees and entropy are calculated for the permutation



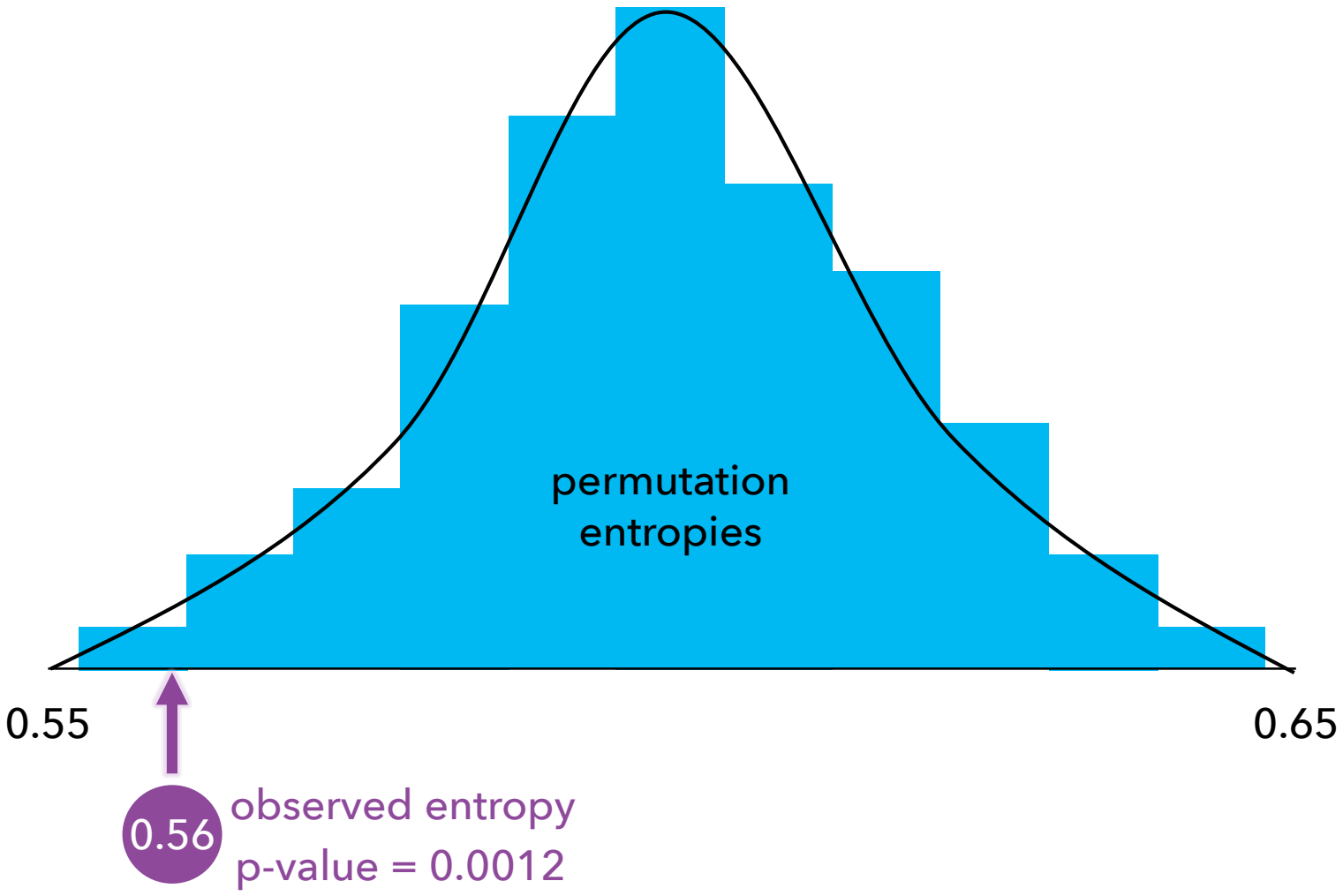
minimum spanning trees and entropy are calculated for the permutation




permutations
neighborhood
Y
X
0.56
0.61
0.57
0.64
...

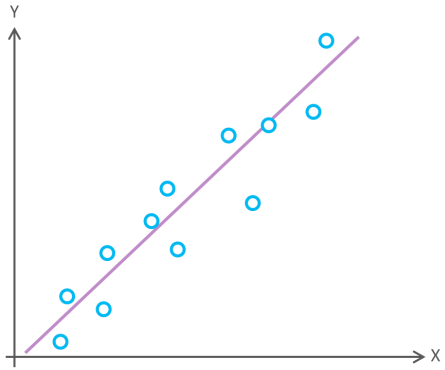

permutations
neighborhood
Y
X
0.56
0.61
0.57
0.64
...
permutation
entropy distribution

observed entropy is converted to a p-value

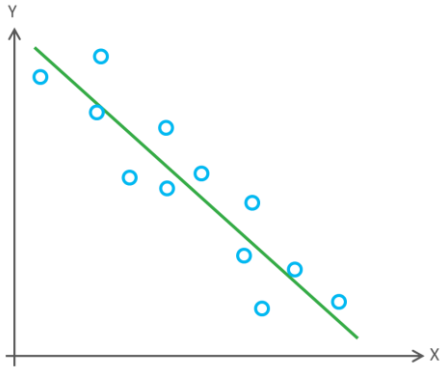


entropy helps determine
if relationships are significant,
but it does not tell us
what type of relationship exists

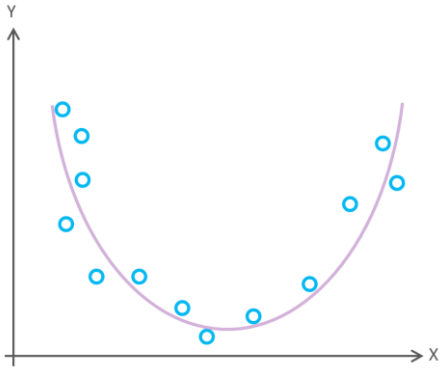
classifying the local relationships



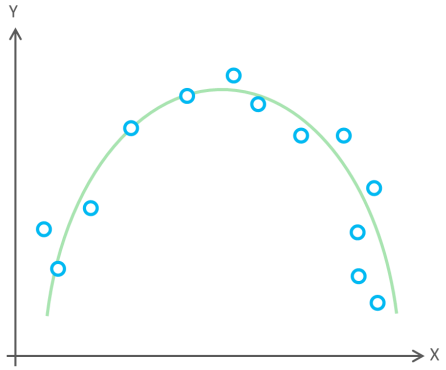
Positive Linear



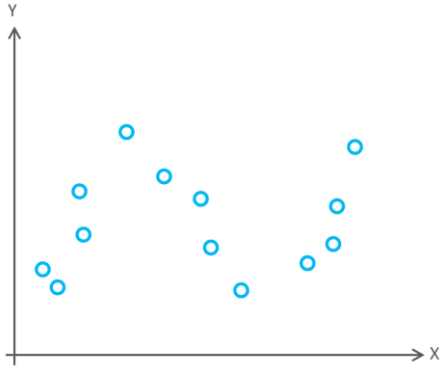
Negative Linear



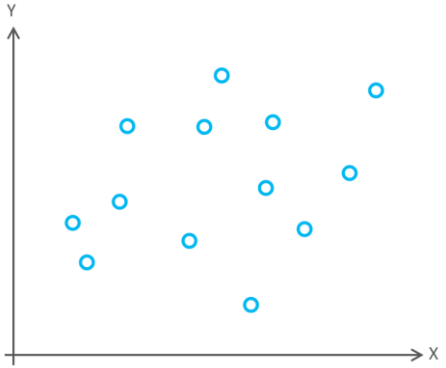
Convex



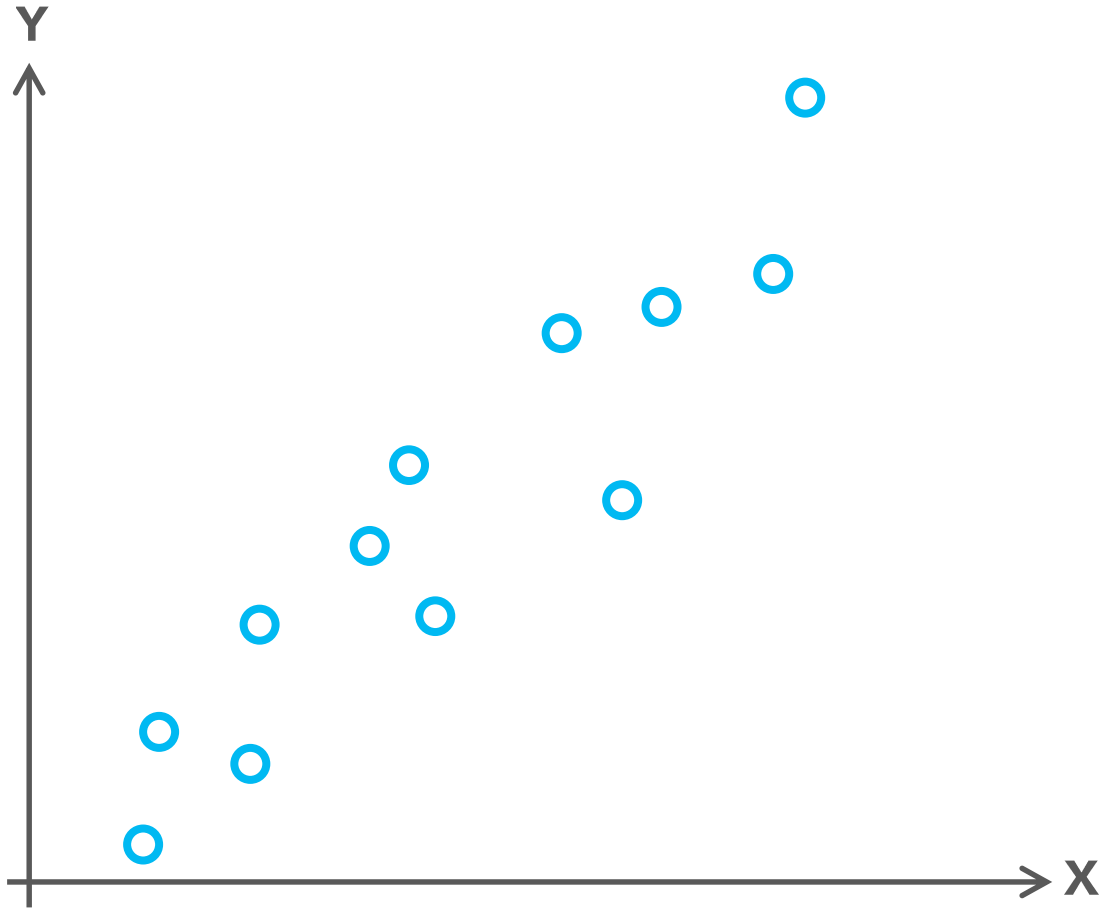
Concave



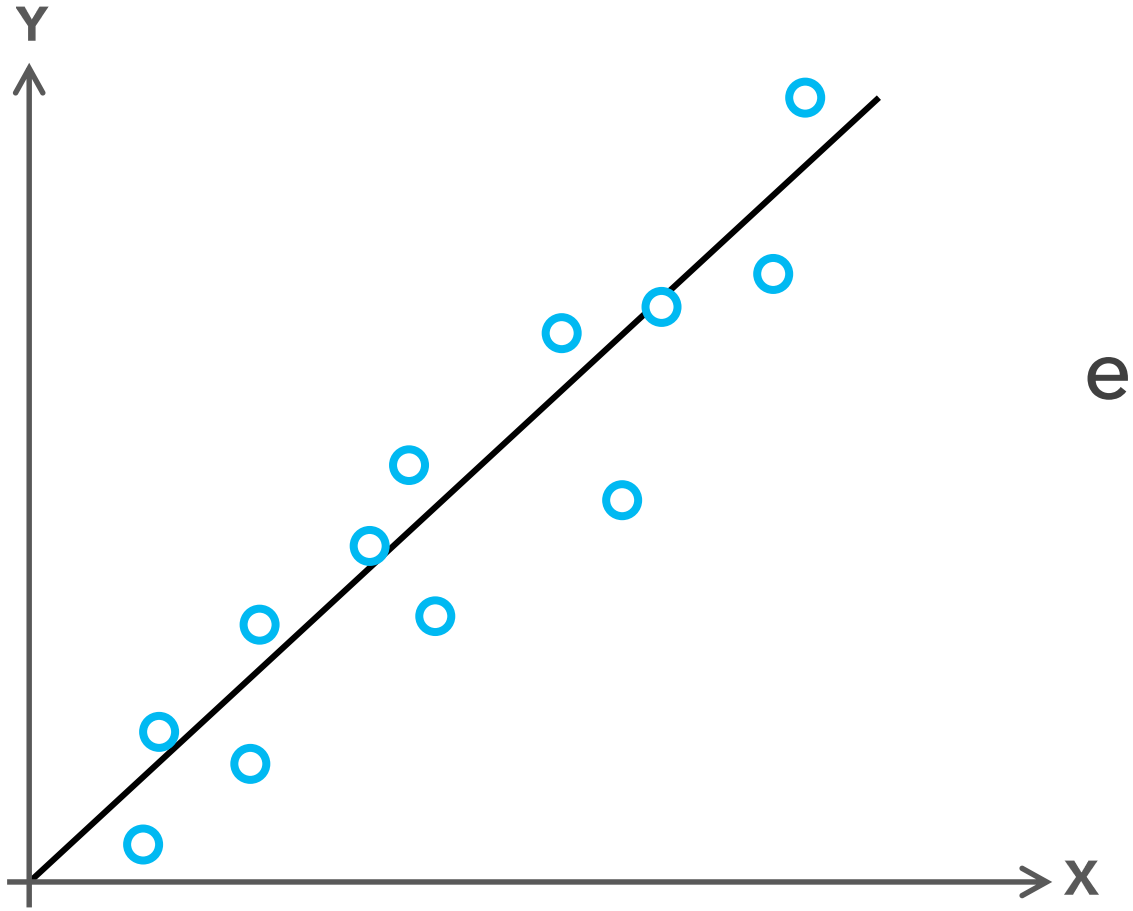
Undefined Complex



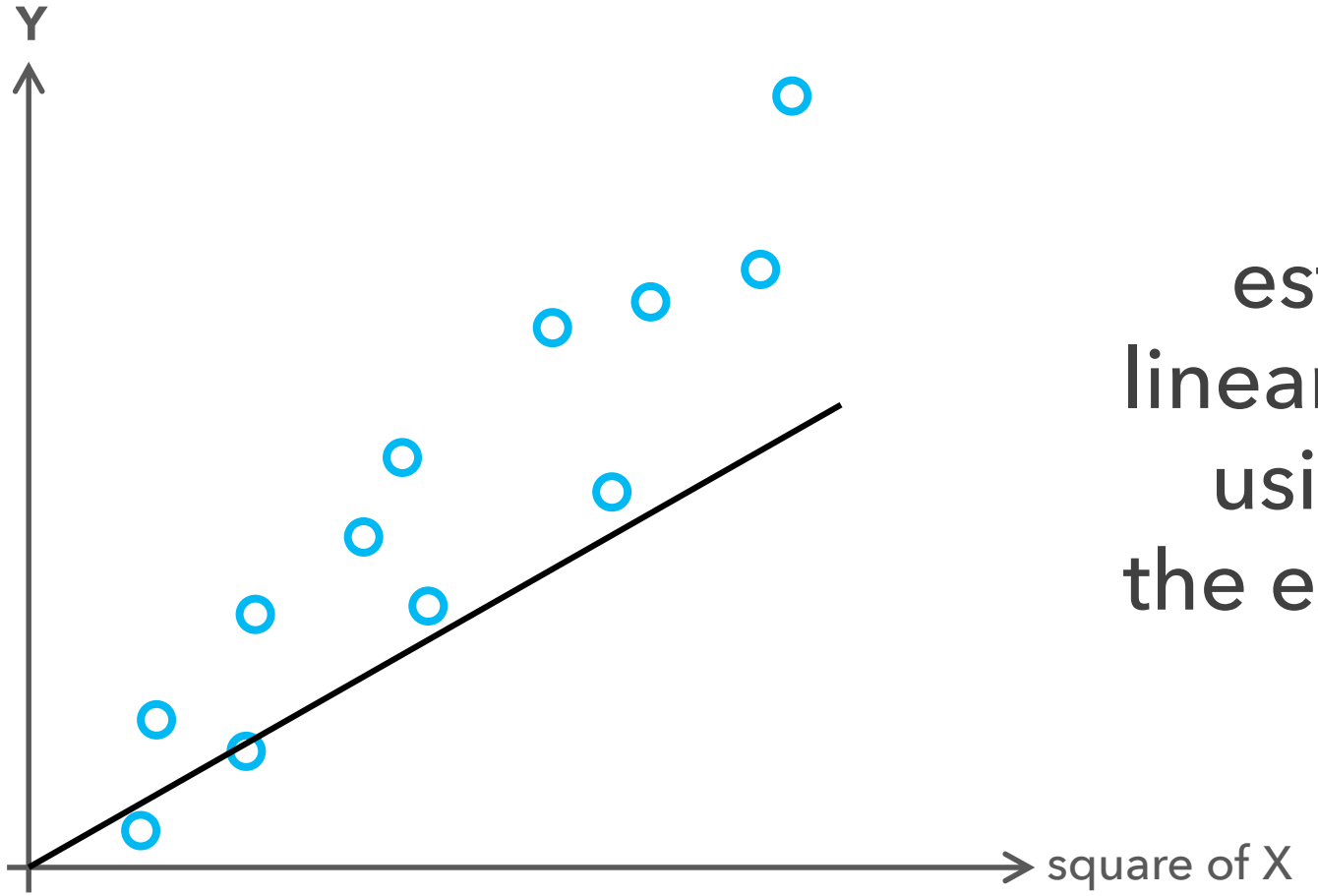
Not Significant



start with the distribution of the significant feature's neighborhood



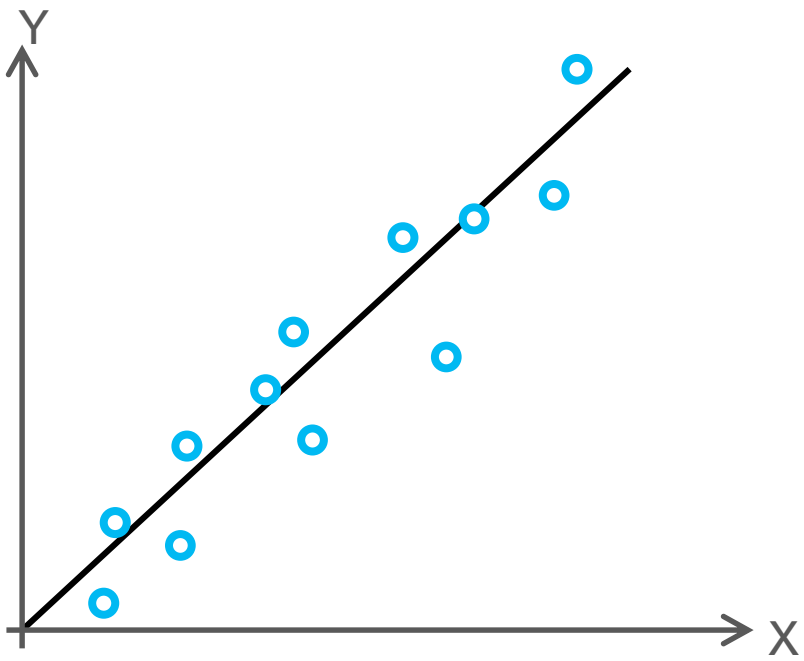
estimate an ordinary linear regression model and calculate AICc



estimate a second linear regression model using the square of the explanatory variable

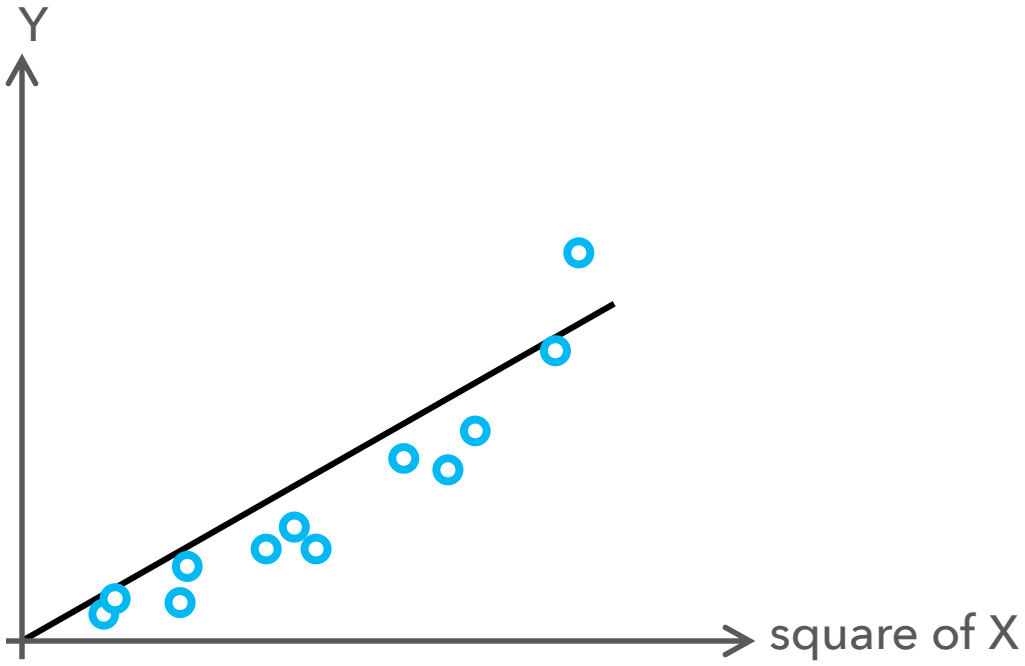
AICc values are compared, and a model is selected

linear model

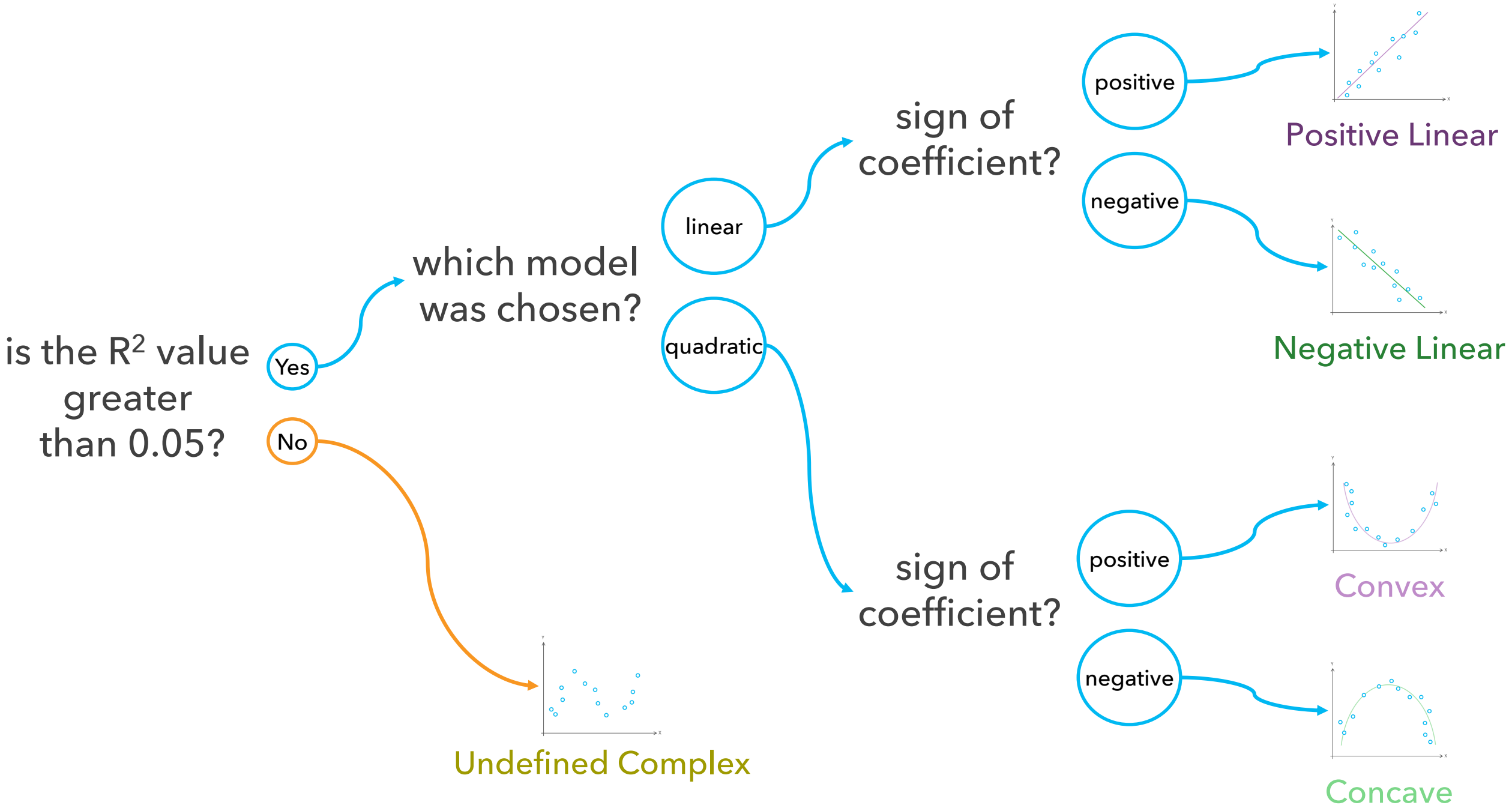


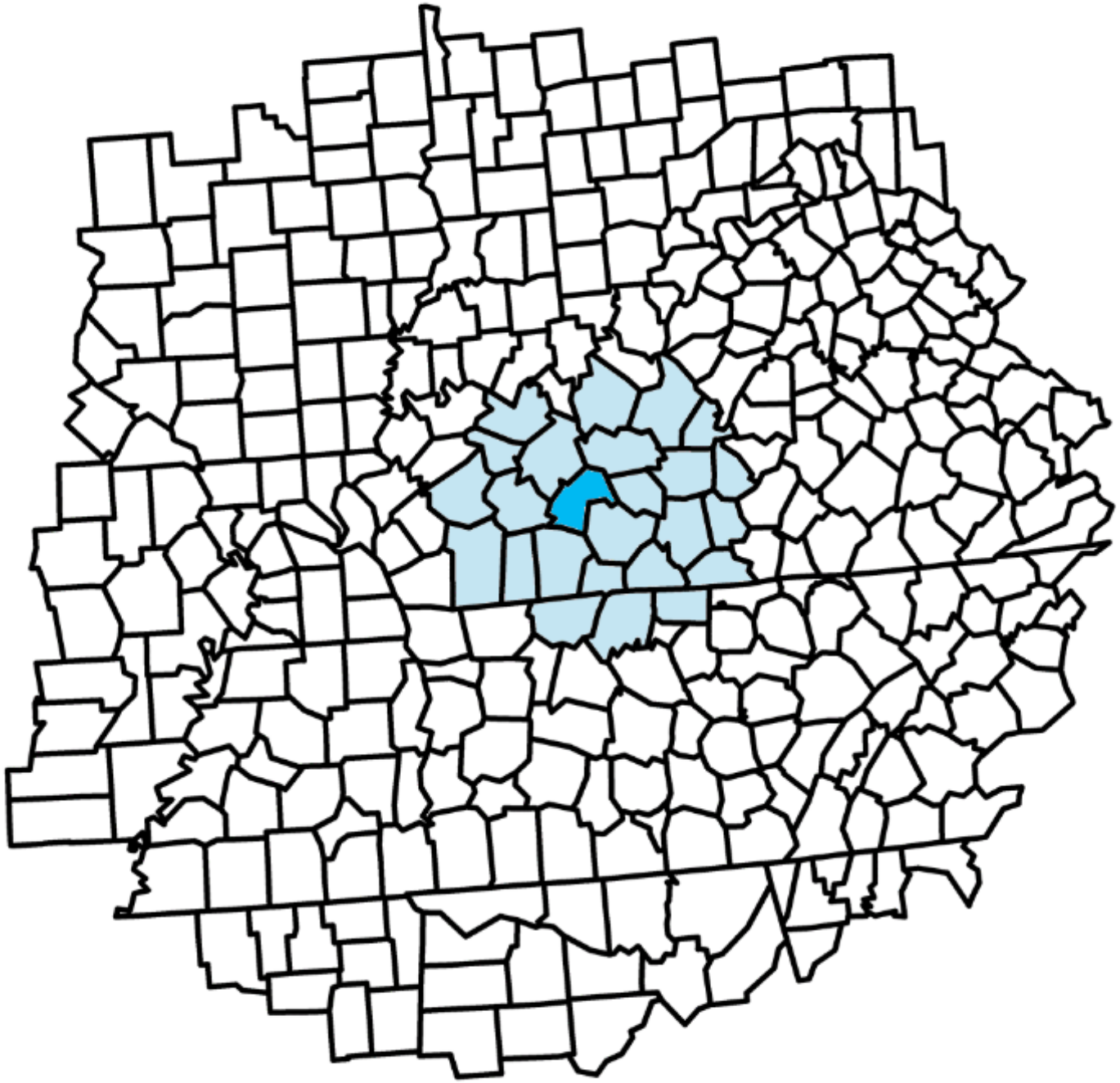
AICc: -42.3

quadratic model

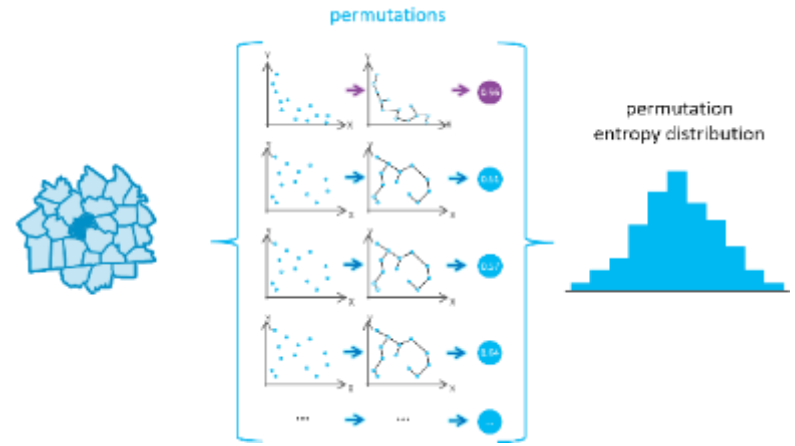


AICc: -51.4

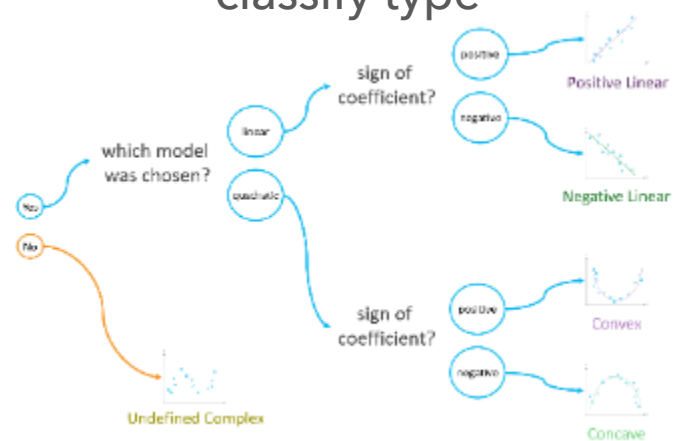

is the R^2 value greater than 0.05?
Yes
No
which model was chosen?
linear
quadratic
sign of coefficient?
positive
negative
Positive Linear
Negative Linear
sign of coefficient?
positive
negative
Convex
Concave
Undefined Complex

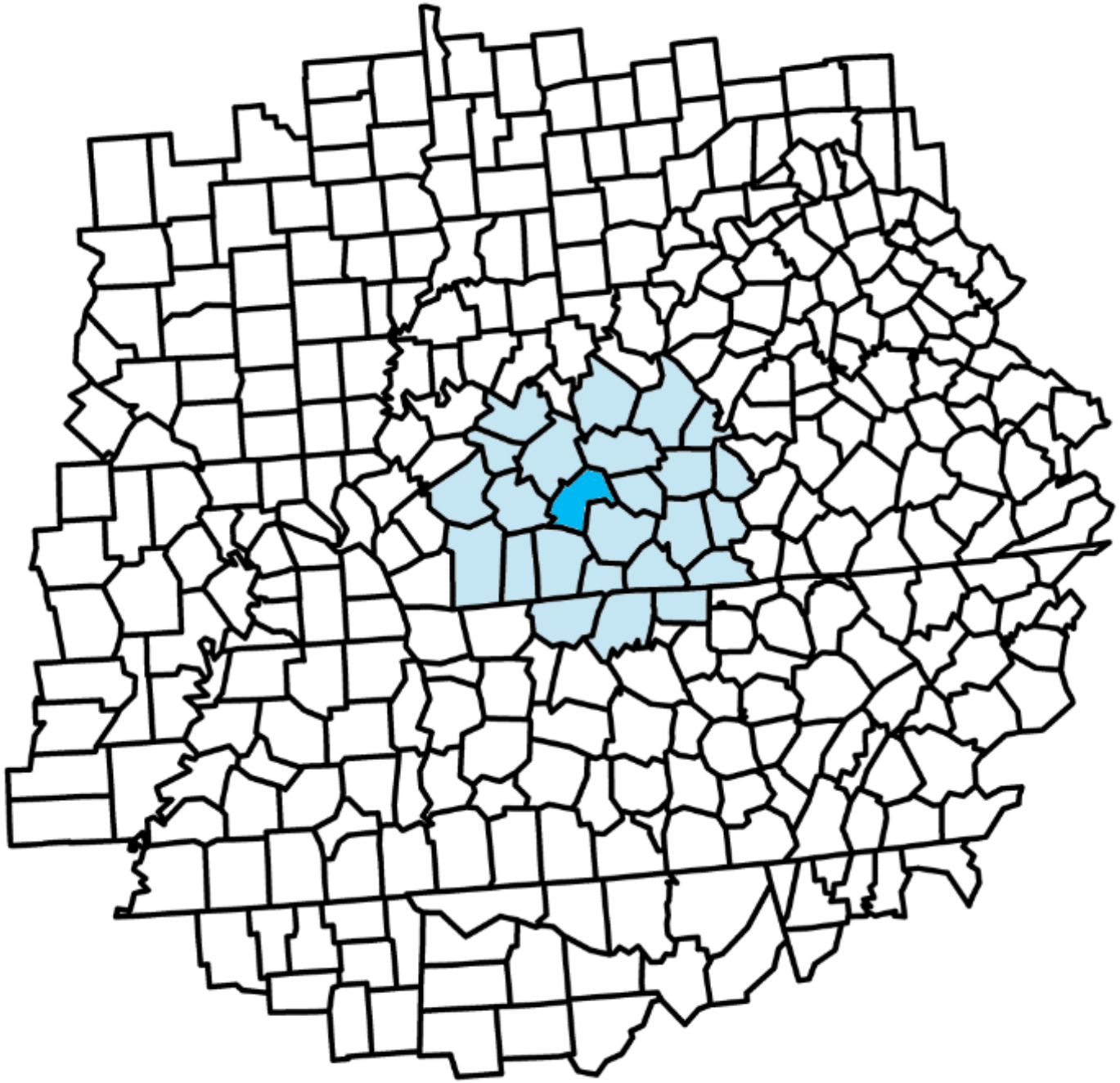


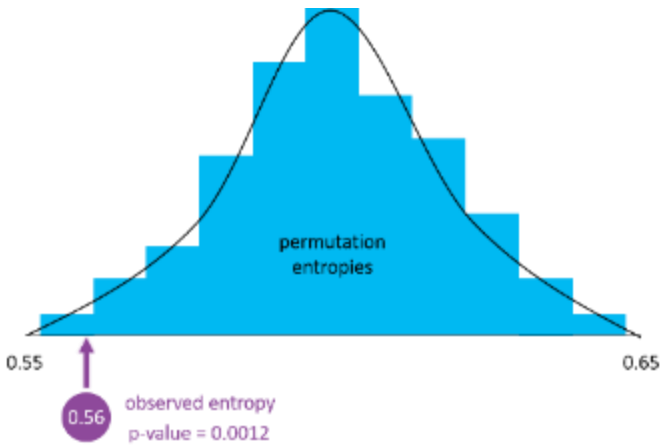
entropy

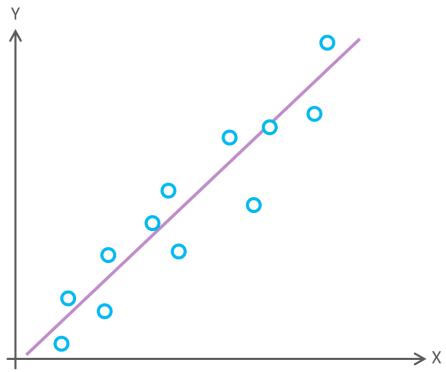


classify type

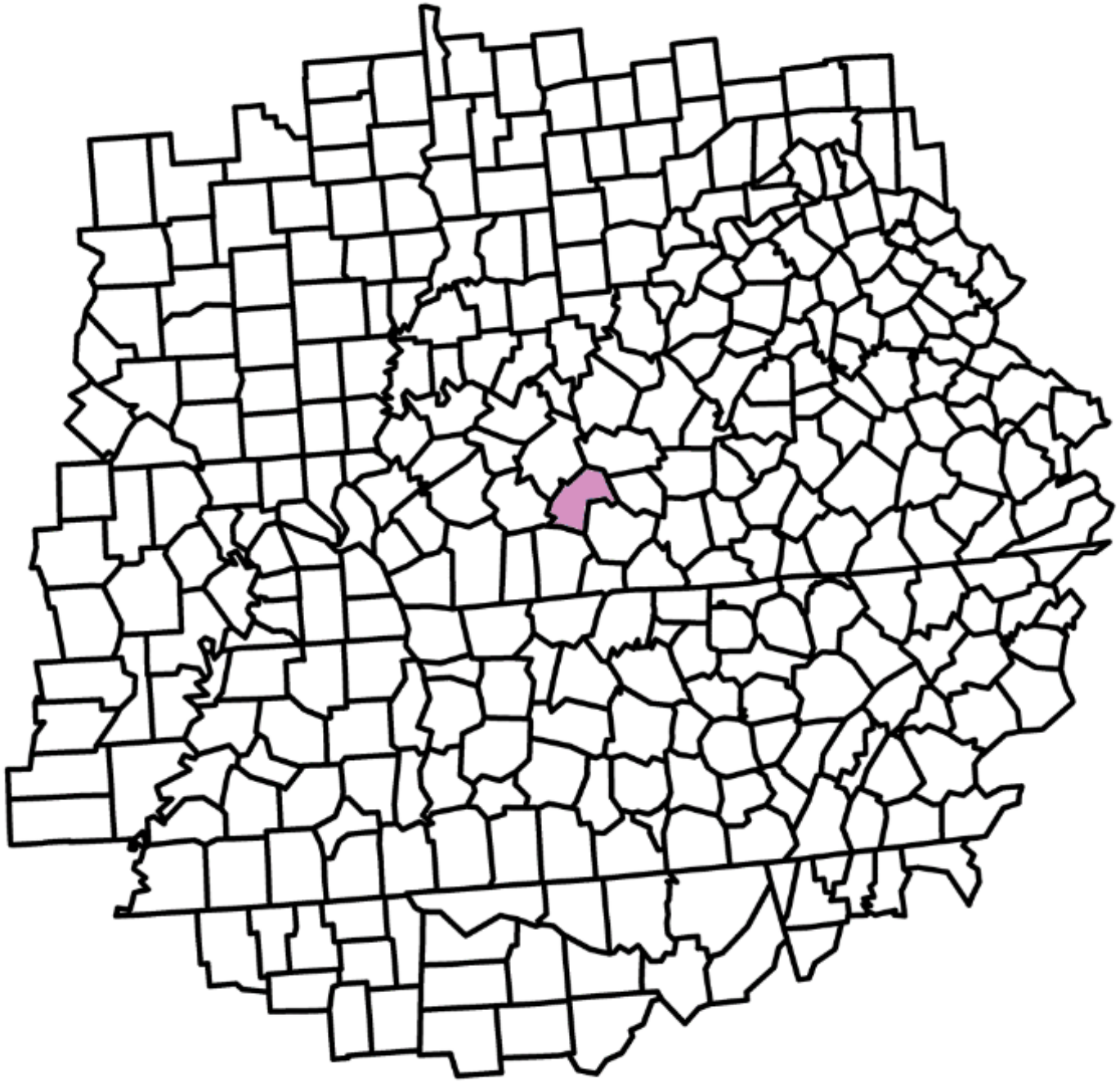


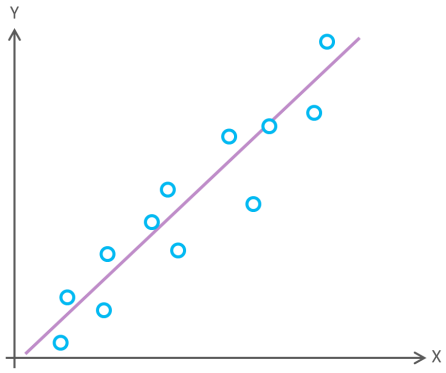




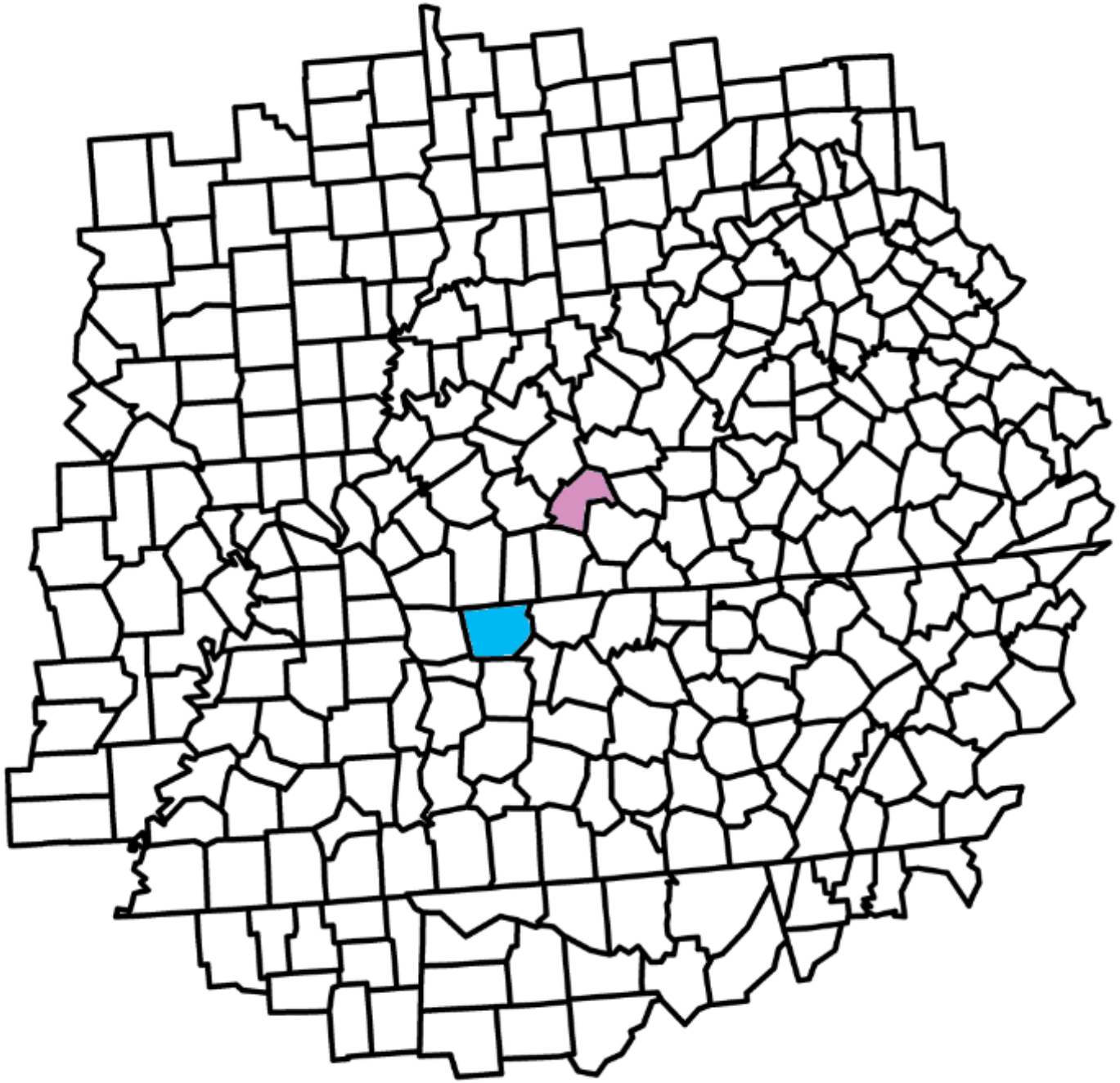


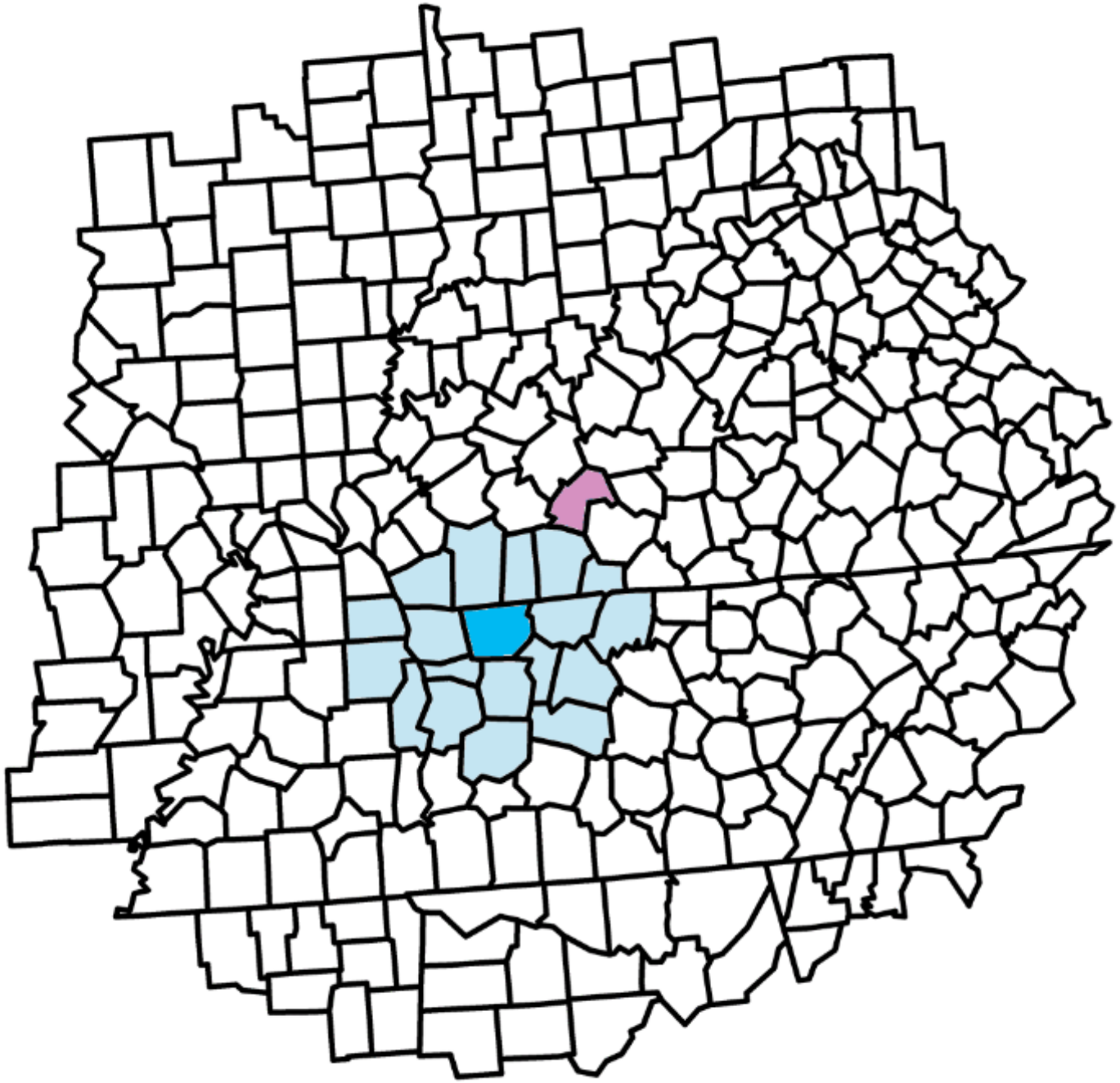
Positive Linear

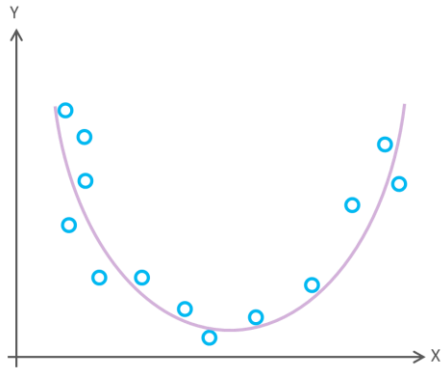




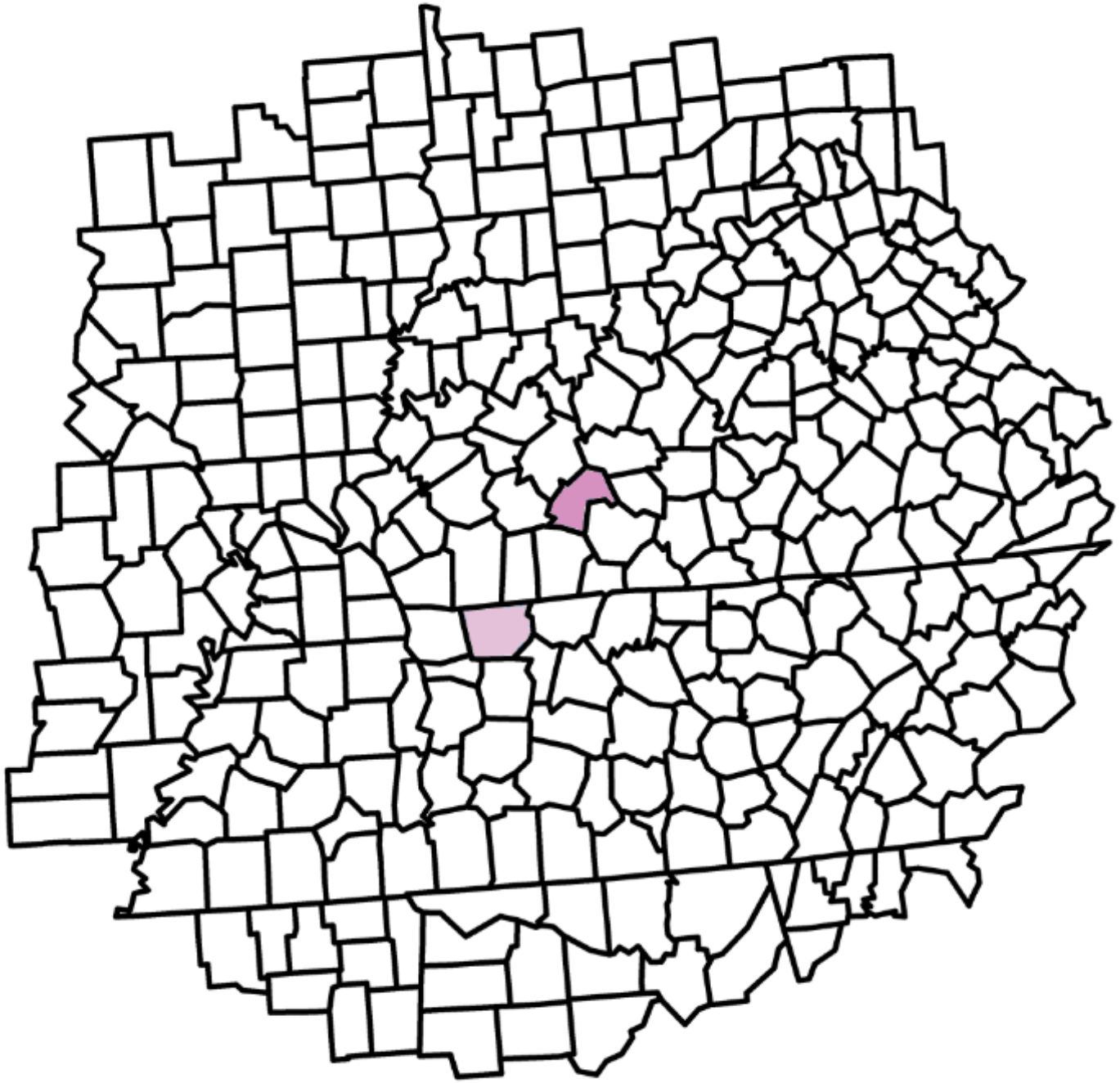
Positive Linear

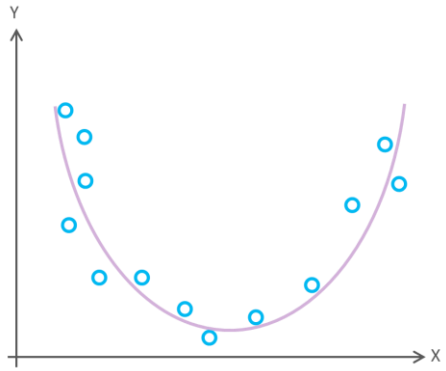




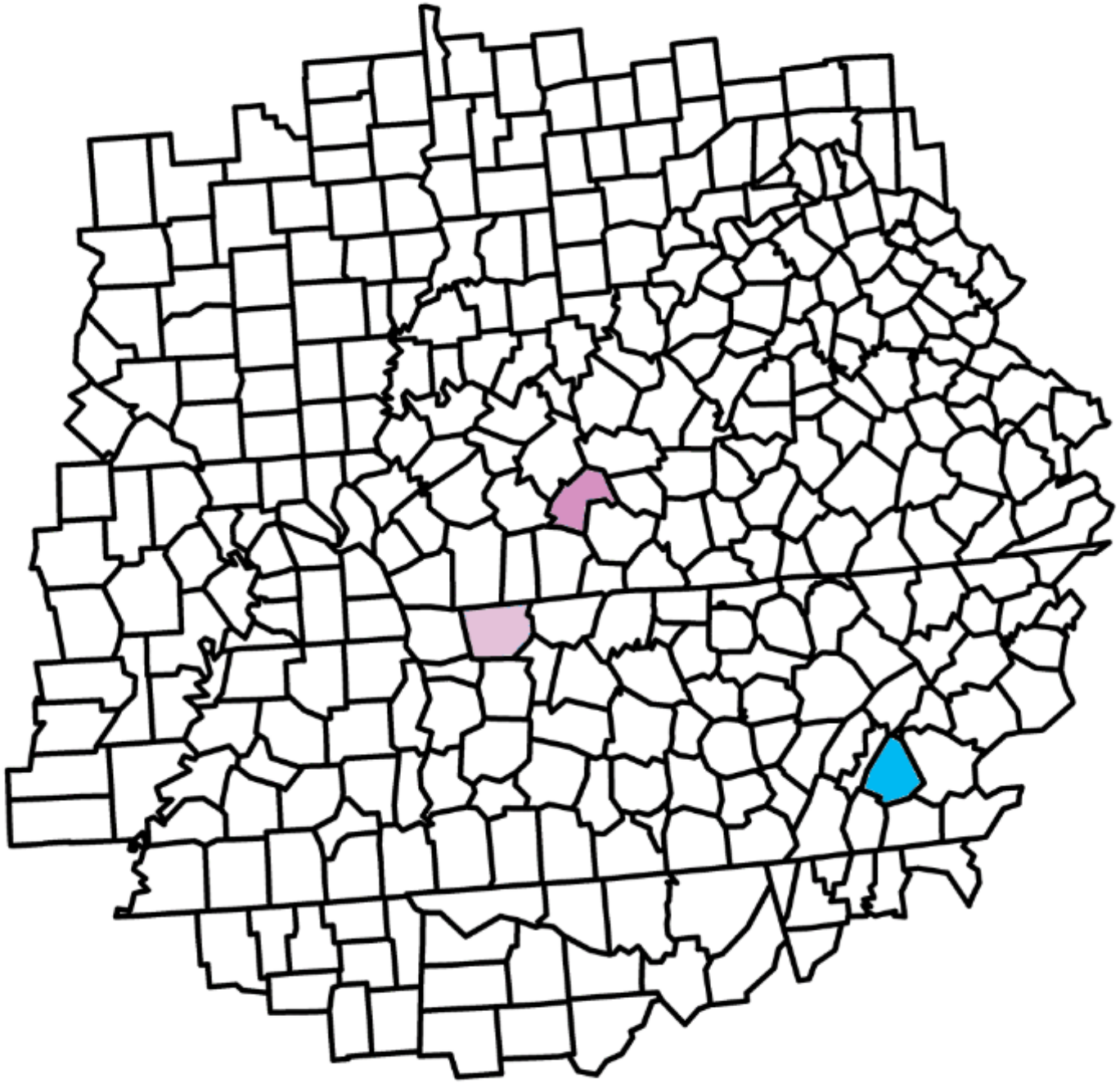


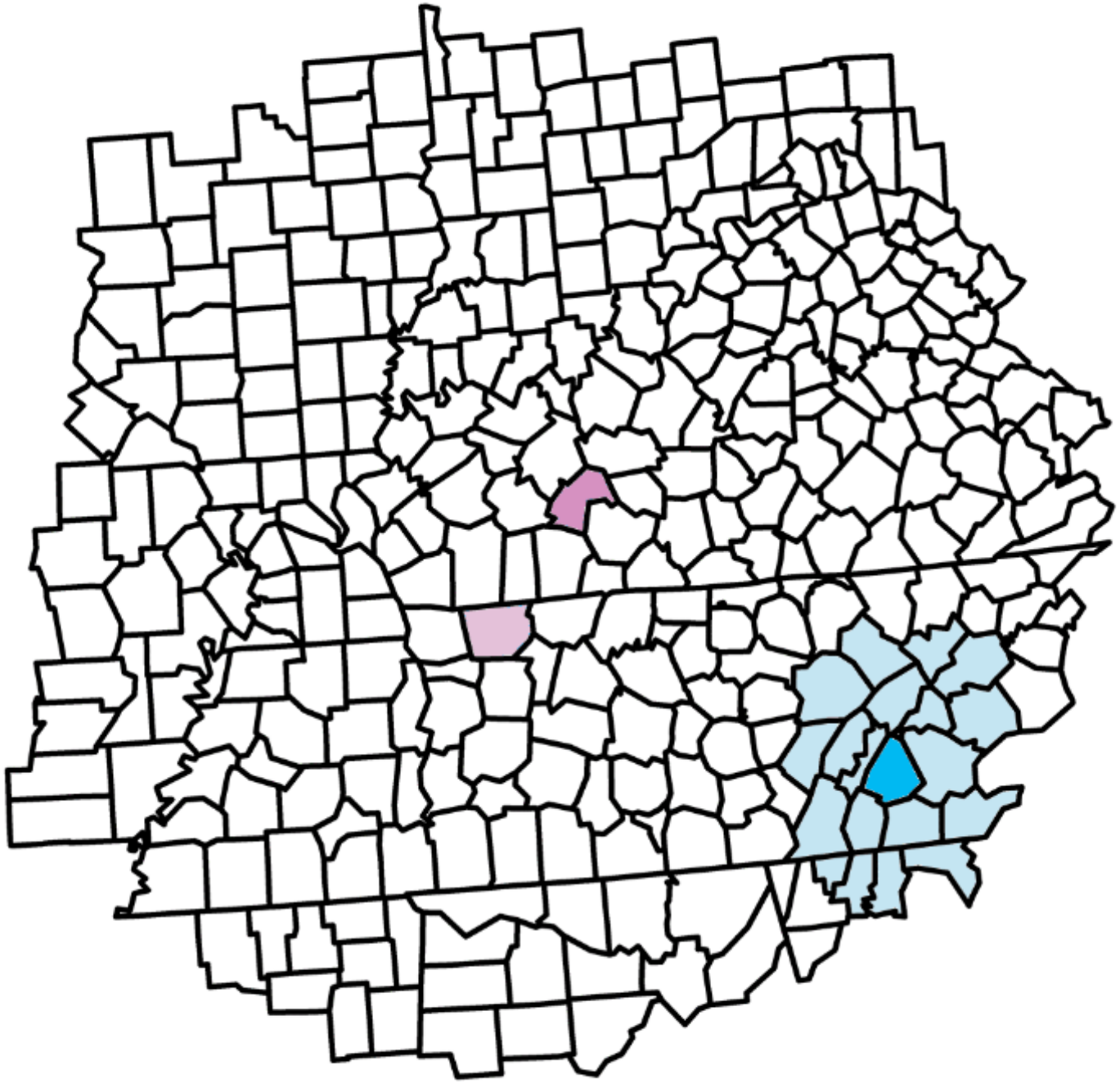
Convex

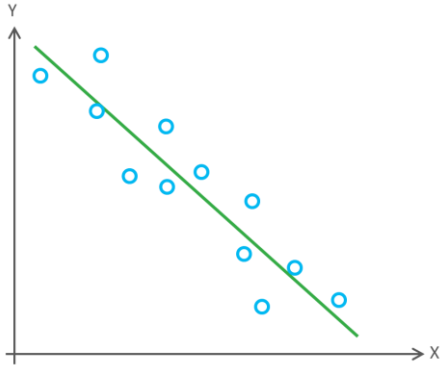




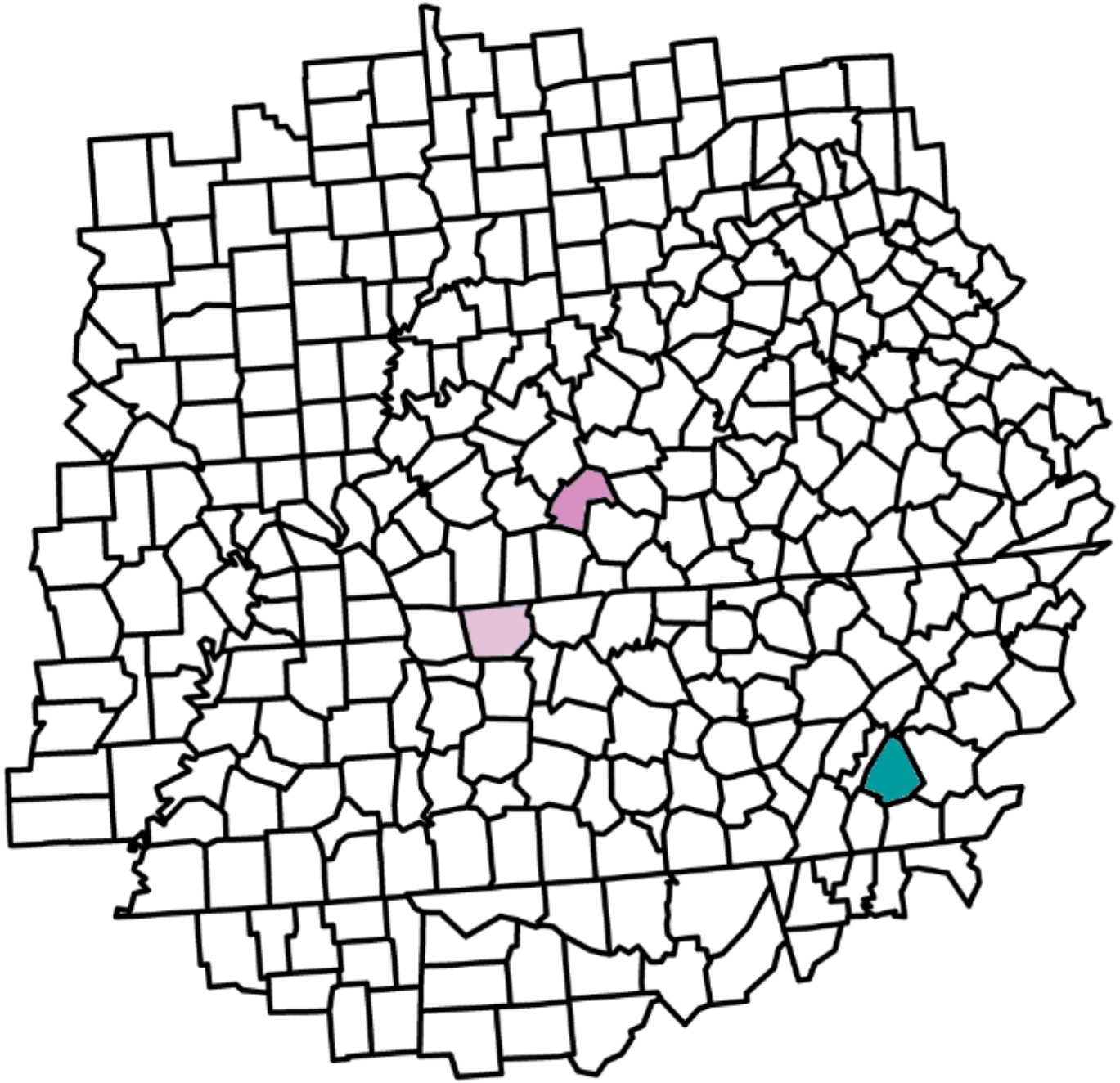
Convex

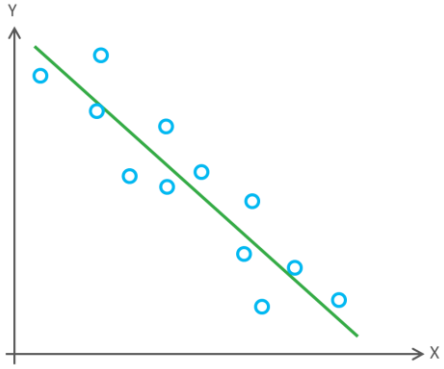




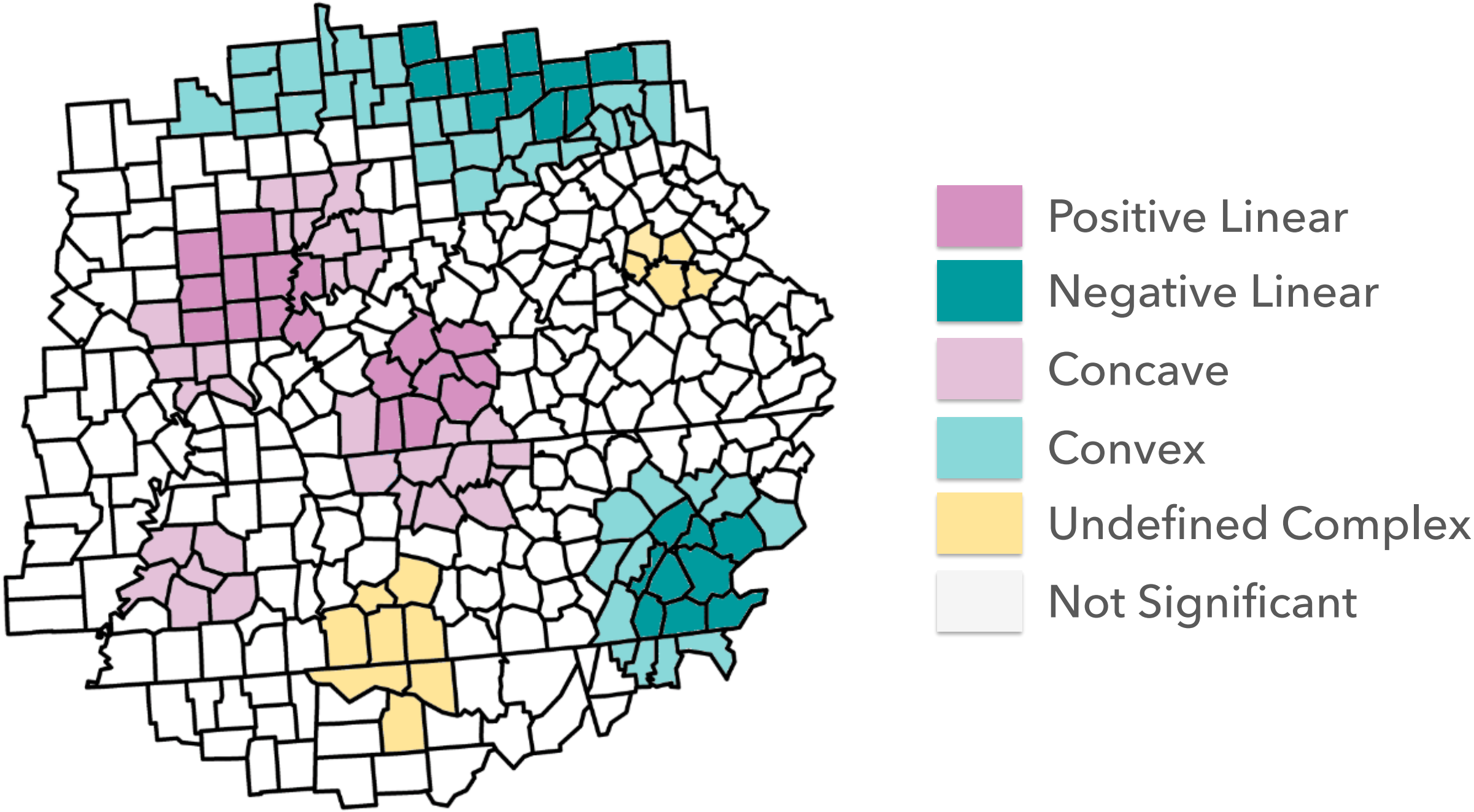


Negative Linear





Negative Linear


Positive Linear
Negative Linear
Concave
Convex
Undefined Complex
Not Significant

Demo

"Essentially, all models are wrong, but some are useful."

- George E. P. Box

esri | THE SCIENCE OF WHERE

esriurl.com/spatialstats