Plan:

1. Introduce approaches to analyzing geospatial data
2. Explain ANN, KNN, and kernel density approaches

Spatial Statistics: Basics

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

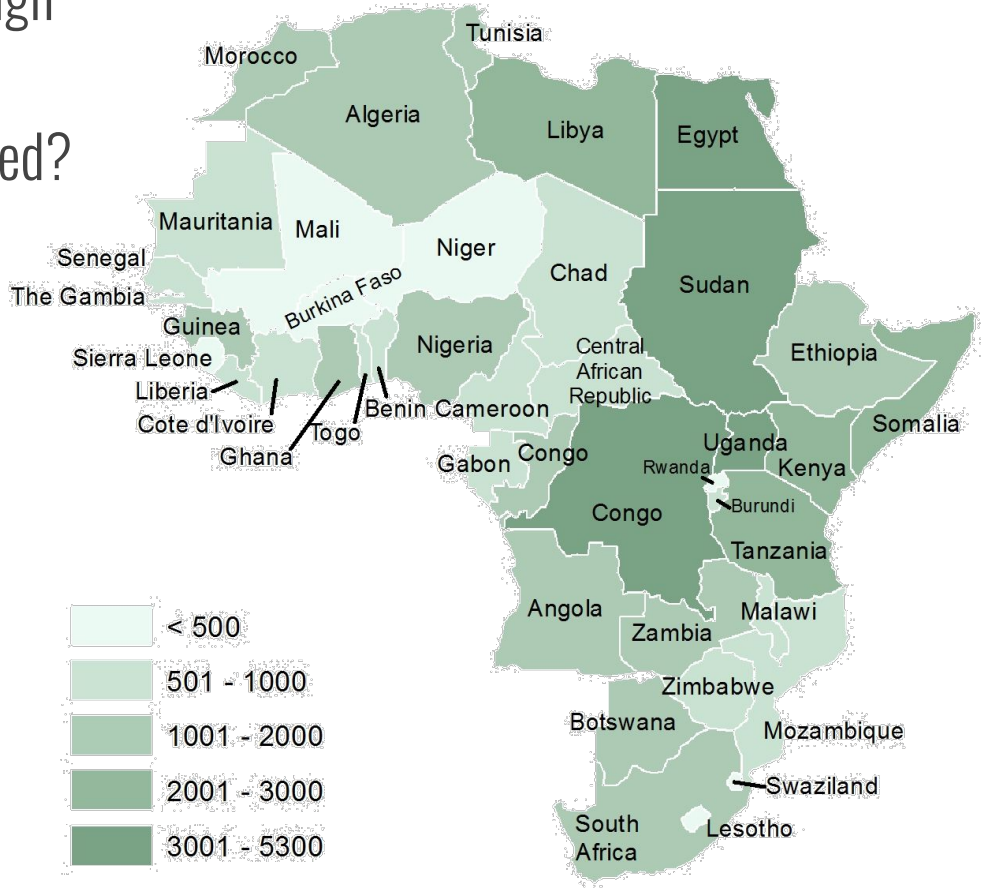
Are countries with a high conflict index score geographically clustered?

Table 1.1: Index of total African conflict for the 1966-78 period (Anselin and O'Loughlin 1992).

Country	Conflicts	Country	Conflicts
EGYPT	5246	LIBERIA	980
SUDAN	4751	SENEGAL	933
UGANDA	3134	CHAD	895
ZAIRE	3087	TOGO	848
TANZANIA	2881	GABON	824
LIBYA	2355	MAURITANIA	811
KENYA	2273	ZIMBABWE	795
SOMALIA	2122	MOZAMBIQUE	792
ETHIOPIA	1878	IVORY COAST	758
SOUTH AFRICA	1875	MALAWI	629
MOROCCO	1861	CENTRAL AFRICAN REPUBLIC	618
ZAMBIA	1554	CAMEROON	604

Data source: Anselin, L. and John O'Loughlin. 1992. Geography of international conflict and cooperation: spatial dependence and regional context in Africa. In The New Geopolitics, ed. M. Ward, pp. 39-75.

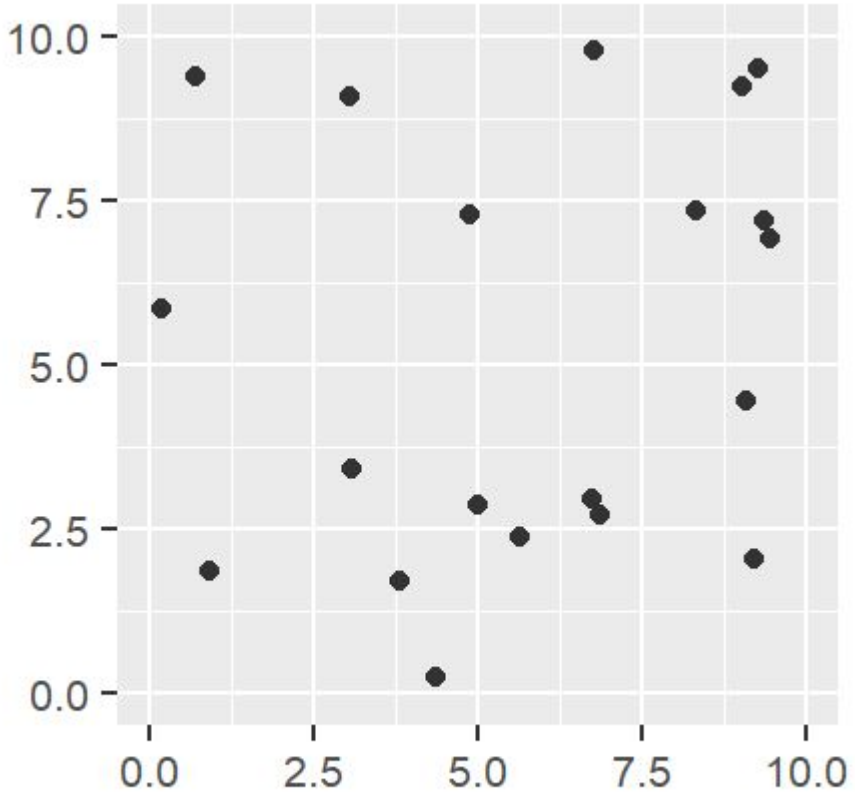
Are countries with a high conflict index score geographically clustered?



https://mgimond.github.io/Spatial/introGIS.html

Global Point Density

the ratio of observed number of points to the study region's surface area

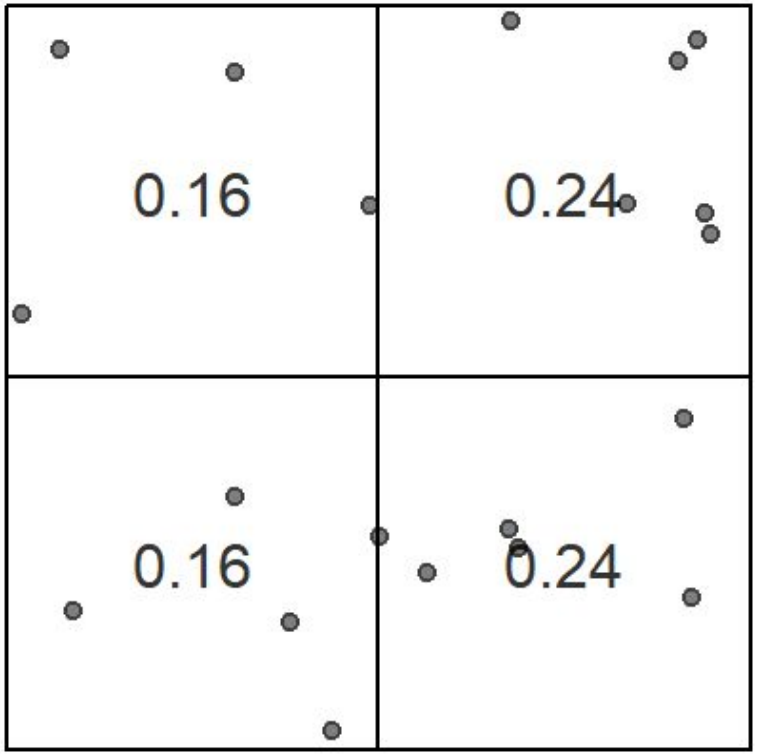


https://mgimond.github.io/Spatial/introGIS.html

Quadrat Density (local)

Surface is divided and then point density is calculated within quadrat

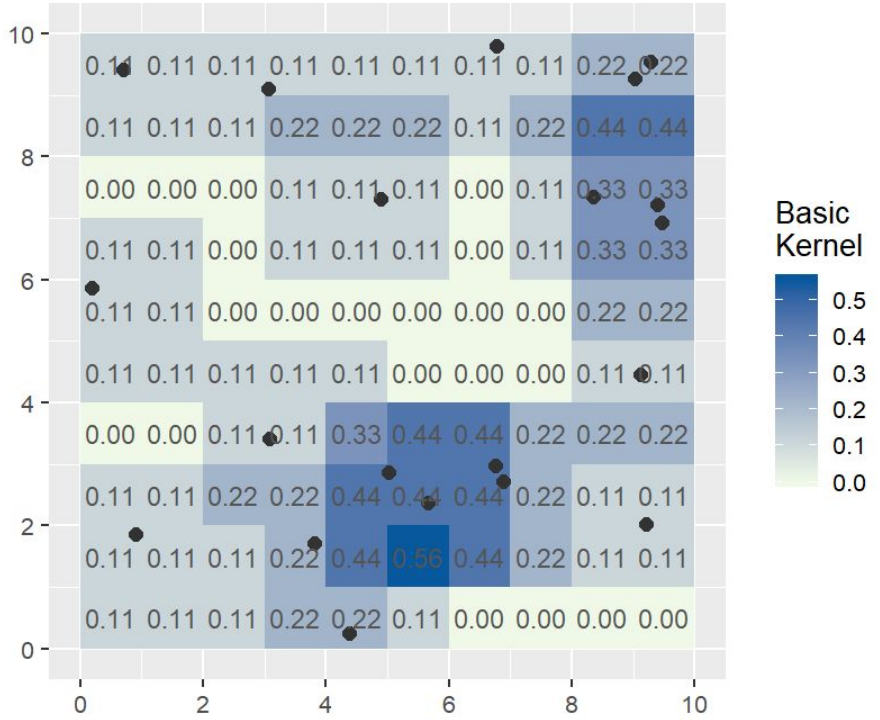
Note: quadrat number and shape will affect measurement estimate.
Suffers from MAUP.



Kernel Density (local)

Point density is calculated within sliding windows (window size = kernel)

Note: kernel will affect measurement estimate, but this is less susceptible to MAUP.



Modeling these data: Poisson Point Process

(Density-based Methods - - how the points are distributed relative to the study space)

$$\lambda(i) = e^{\alpha+\beta Z(i)}$$

λ(i) is the modeled intensity at location *i*
e^{α} is the base intensity when the covariate is *zero*
e^{β} is the multiplier by which the intensity increases (or decreases) for each 1 unit increase in the covariate

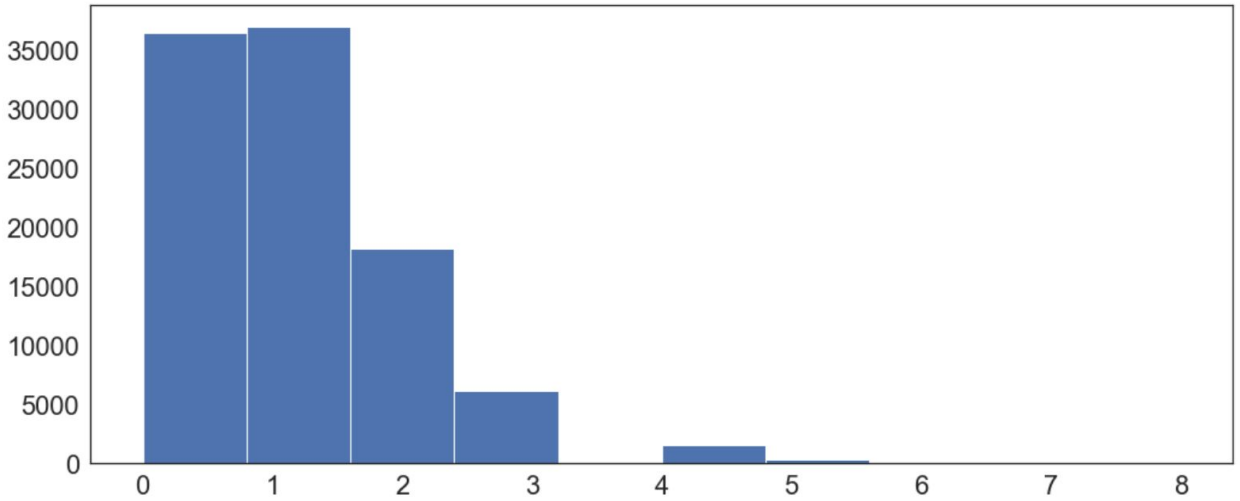
Poisson Distribution

The Poisson Distribution models events in fixed intervals of time, given a known average rate (and independent occurences).

In [55]:

Slide Type Fragment

```
dat = poisson.rvs(mu=1, size=100000)
plt.hist(dat);
```

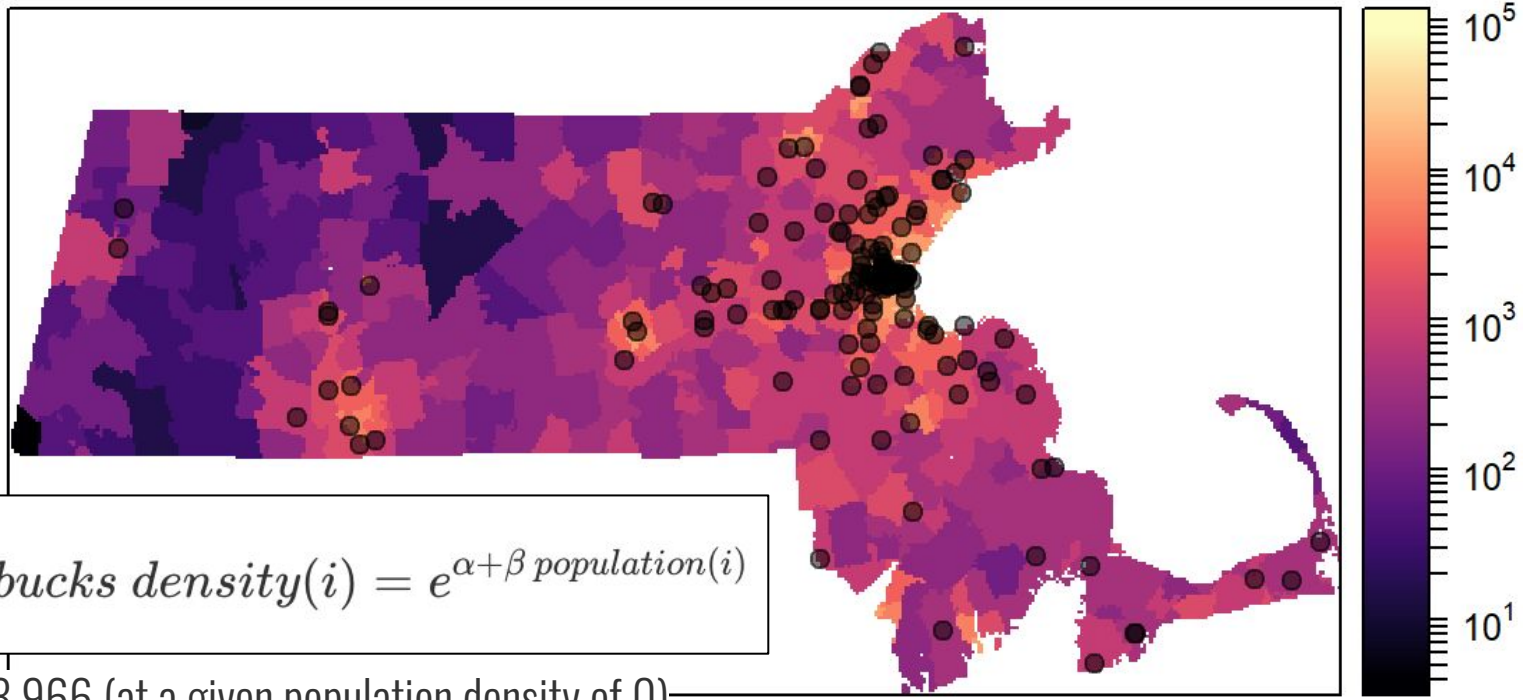


Slide Type Fragment

The **number of visitors a fast food drive-through gets each minute** follows a Poisson distribution. In this case, maybe the average is 3, but there's some variability around that number.

A Poisson distribution can help calculate the probability of various events related to customers going through the drive-through at a restaurant. It will predict lulls (0 customers) and flurry of activity (5+ customers), allowing staff to plan and schedule more precisely.

Location of Starbucks relative to population density in MA



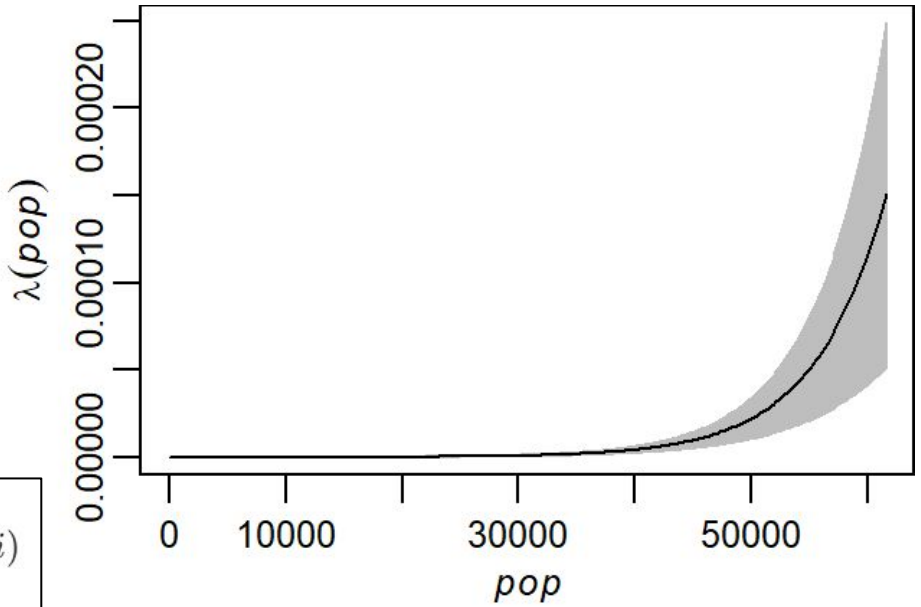
$$Starbucks\ density(i) = e^{\alpha + \beta\ population(i)}$$

α = -18.966 (at a given population density of 0)

$e^{-18.966}$ = 5.80 x 10^{-9} cafes per square meter

β = 0.00017; $e^{0.00017}$ or 1.00017

Location of Starbucks relative to population density in MA

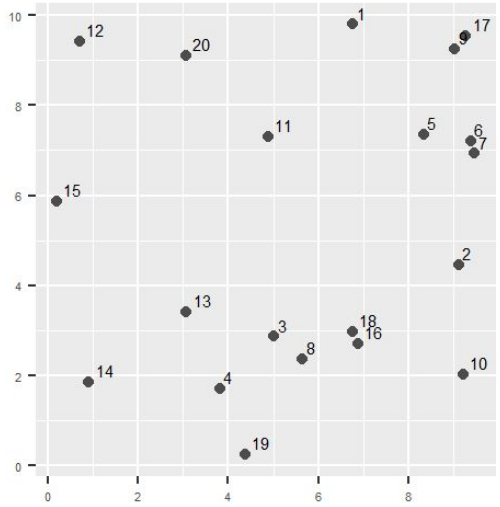


$$Starbucks\ density(i) = e^{\alpha + \beta\ population(i)}$$

α = -18.966 (at a given population density of 0)
$e^{-18.966}$ = 5.80 x 10^{-09} cafes per square meter
β = 0.00017; $e^{0.00017}$ or 1.00017

Modeling these data: Average Nearest Neighbor

(Distance-based Methods - how the points are distributed relative to one another)



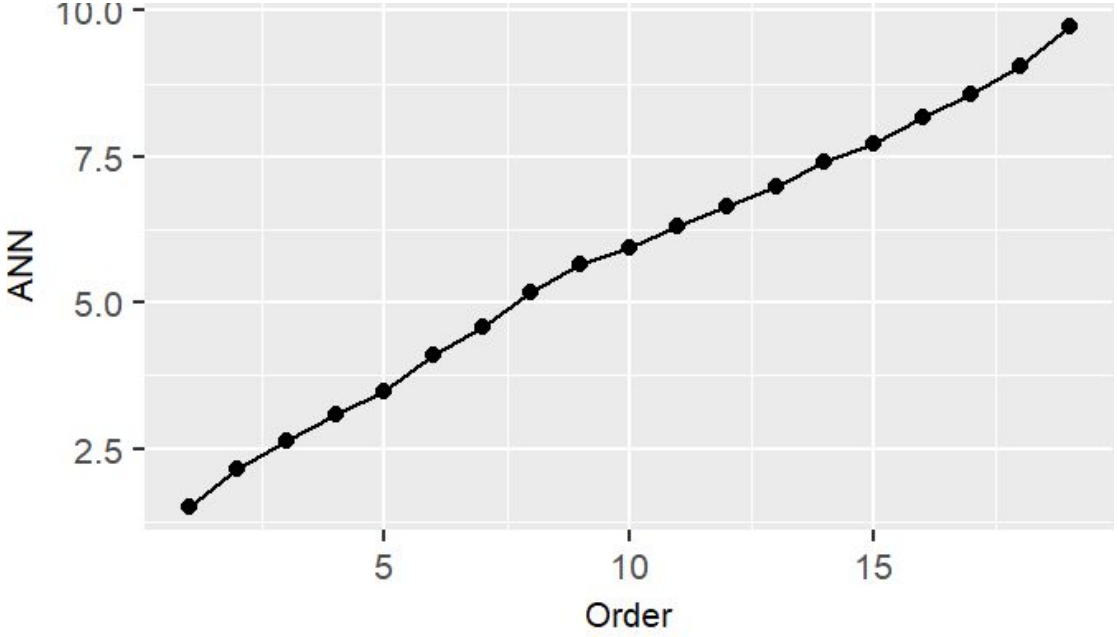
From	To	Distance	From	To	Distance
1	9	2.32	11	20	2.55
2	10	2.43	12	20	2.39
3	8	0.81	13	4	1.85
4	19	1.56	14	13	2.67
5	6	1.05	15	12	3.58
6	7	0.3	16	18	0.29
7	6	0.3	17	9	0.37
8	3	0.81	18	16	0.29
9	17	0.37	19	4	1.56
10	2	2.43	20	12	2.39

ANN = 1.52 units

Modeling these data: Average Nearest Neighbor

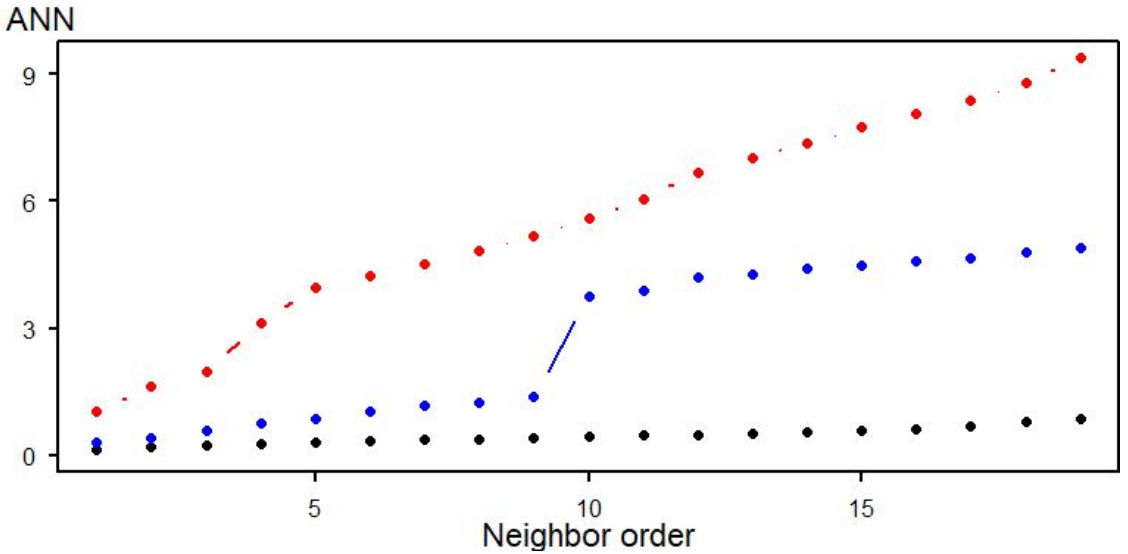
(Distance-based Methods - how the points are distributed relative to one another)

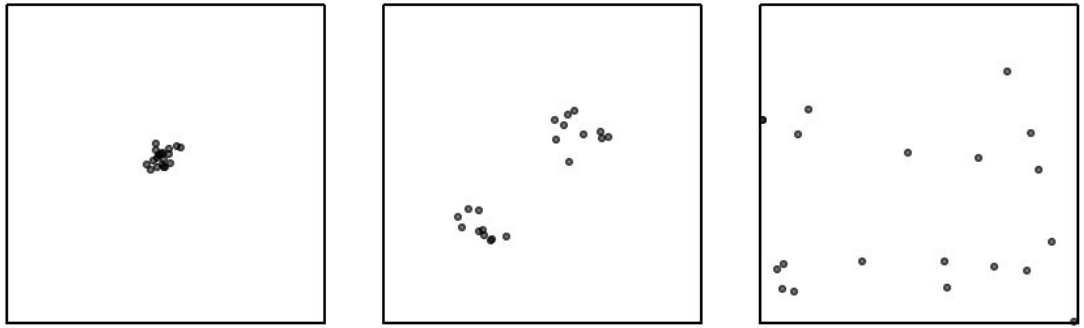
plot the ANN values for different order neighbors, that is for the first closest point, then the second closest point, and so forth.



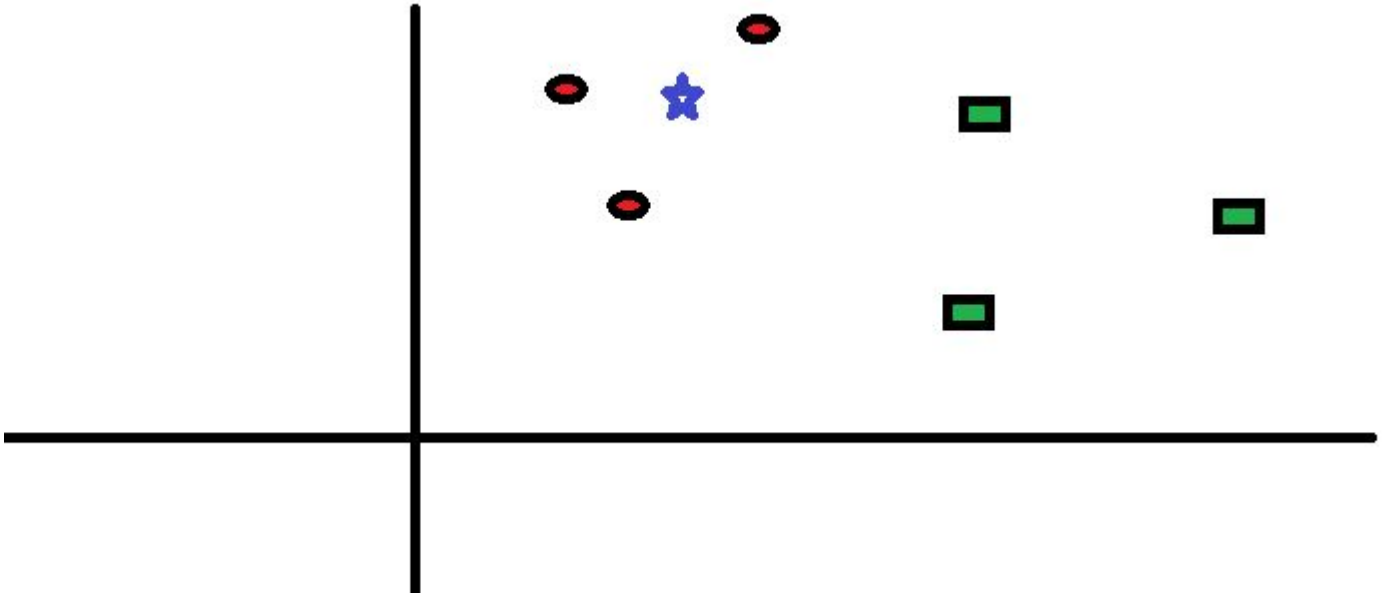
ANN vs neighbor order offers insight into underlying spatial relationship

Note: study space definition affects this measure



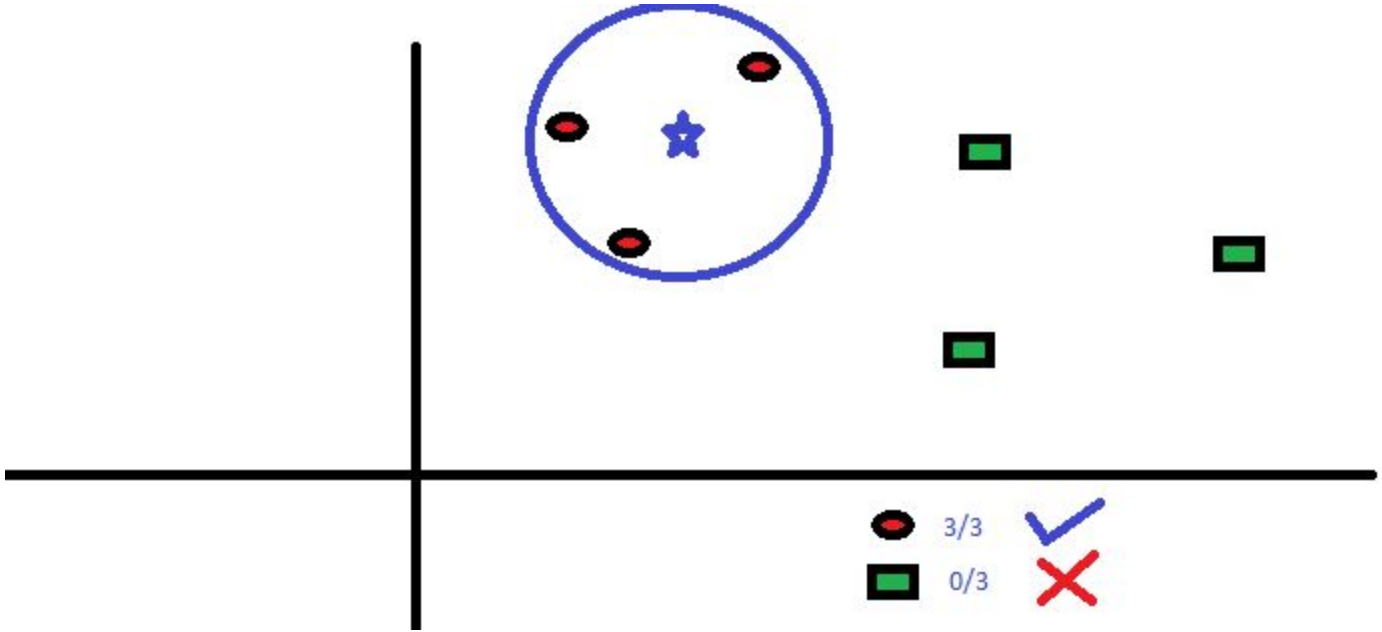


KNN: K Nearest Neighbor for Classification



https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

KNN: To which class does the blue star belong?

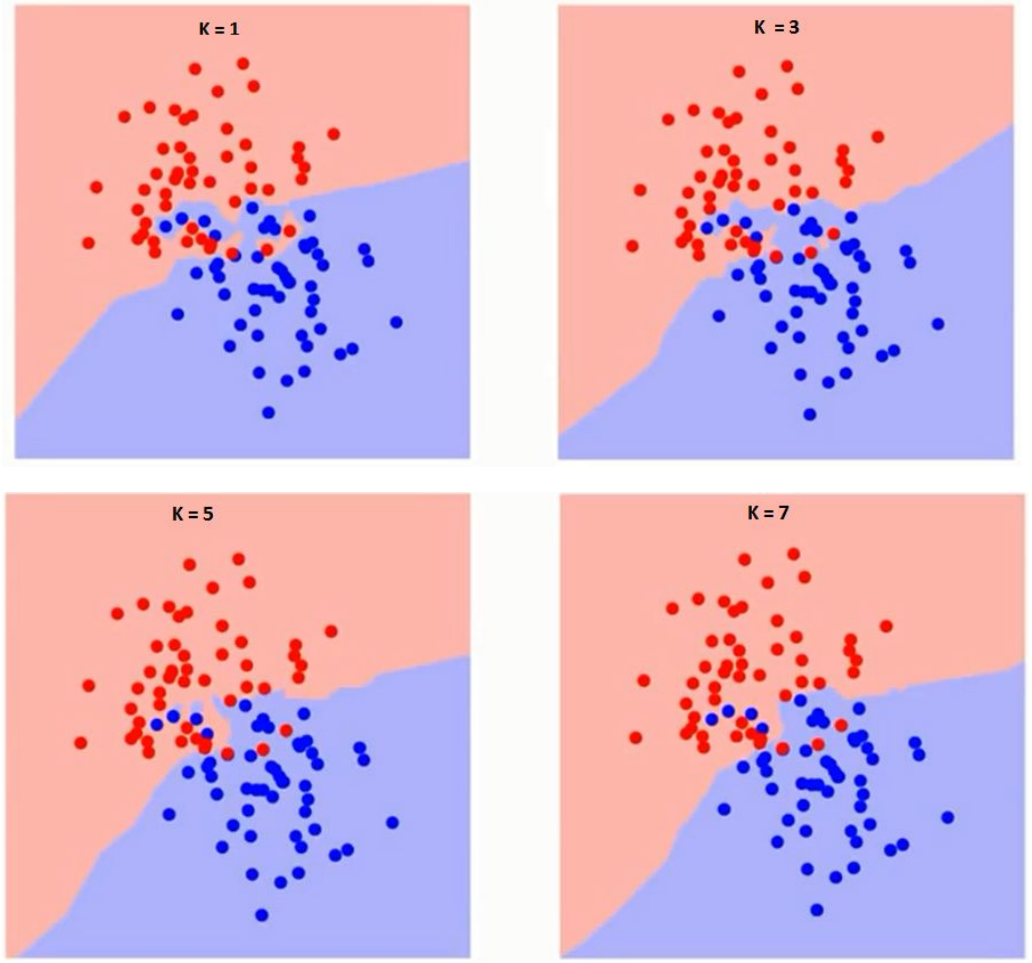


https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

KNN: Choosing K

K specifies how many neighbors to consider.

Note that as more neighbors are considered, the boundary smooths out.



https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

KNN: Pros & Cons

Pros:

- No assumptions about data (good for nonlinear)
- Simple and interpretable
- Relatively high accuracy
- Versatile (classification & regression)

Cons:

- Computationally intensive
- High Memory requirements
- Stores all (or most) of training data
- Prediction slow with large N
- Sensitive to outliers/irrelevant features

https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7