Lecture 6: Spatial Hotspot Detection

Spatial Data Mining

Instructor: Yiqun Xie

Recap and Technical Roadmap

All the concepts listed here are important for general data mining, data science, machine learning and AI

What we have covered

Probability density function

Statistical distribution

i.i.d.: identical & independent dist.

Maximum likelihood estimation

Optimization

Vectors & matrix

Reinforce →

Keep warm →

Hotspot detection

Probability density function

Statistical distribution

i.i.d.: identical & independent dist.

Maximum likelihood estimation

Optimization

Vectors & matrix

Reinforce →

New

Statistical hypothesis

Spatial point process

Test statistic

Statistical significance

Monte-Carlo simulation

Prediction & Classification

Optimization

Vectors & matrix

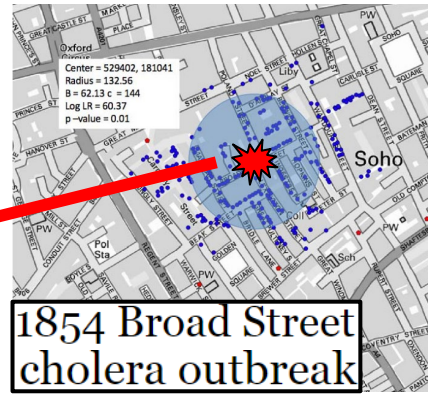
New…

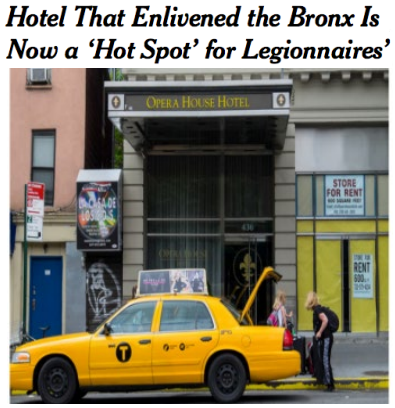
Spatial Hotspot: Unique Characteristics

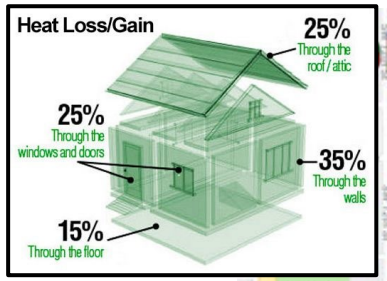
- High cost of spurious results
- Low-dimensional space
- Spatial contiguity
- Event and control

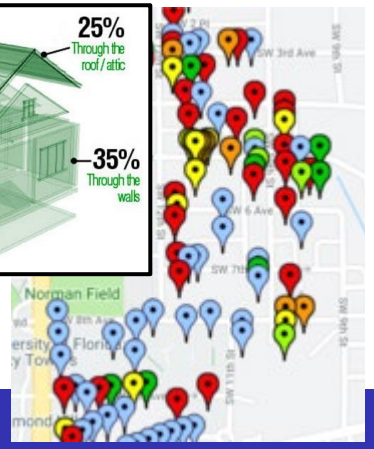
High Cost of Spurious Results

- Public health
  - Disease outbreak, e.g., Legionnaires, cancer...
- Public safety
  - Abnormally high rates of traffic accidents, crimes...
- Sustainability
  - Regions with high energy consumptions...
- Transportation
  - Routes with high emissions (e.g., NOx)...
- Environment, agriculture, forestry, & many more...





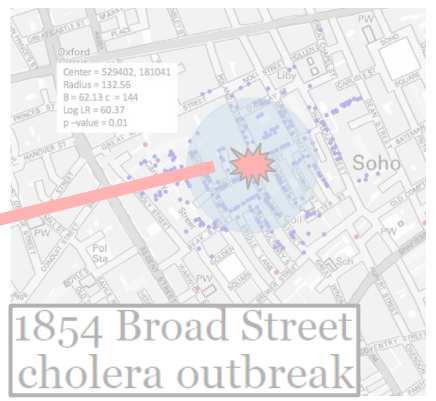


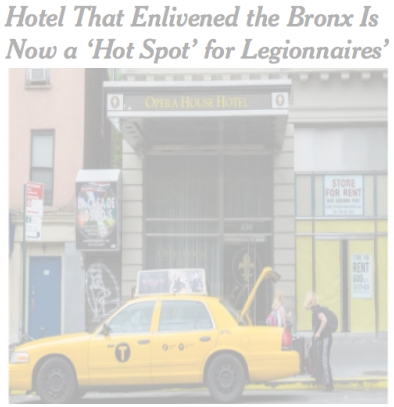


High Cost of Spurious Results

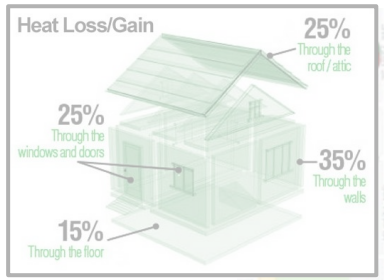
- Public health
  - Disease outbreak, e.g., Legionnaires, cancer...
- Public safety
  - Abnormally high rates of traffic accidents, crimes...
- Sustainability
  - Regions with high energy consumptions...
- Transportation
  - Routes with high emissions (e.g., NOx)...
- Environment, agriculture, forestry, & many more...

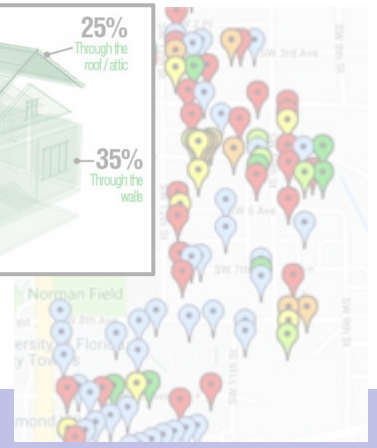
Tight connection with event detection in urban areas or on social networks





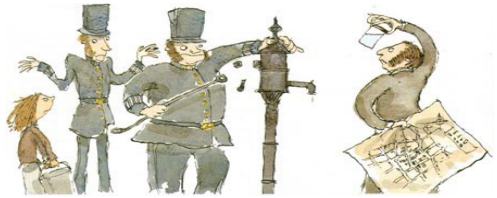
High cost of spurious patterns!

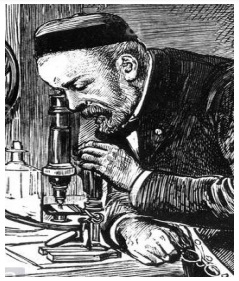




London Cholera Outbreak

1854: What causes Cholera?



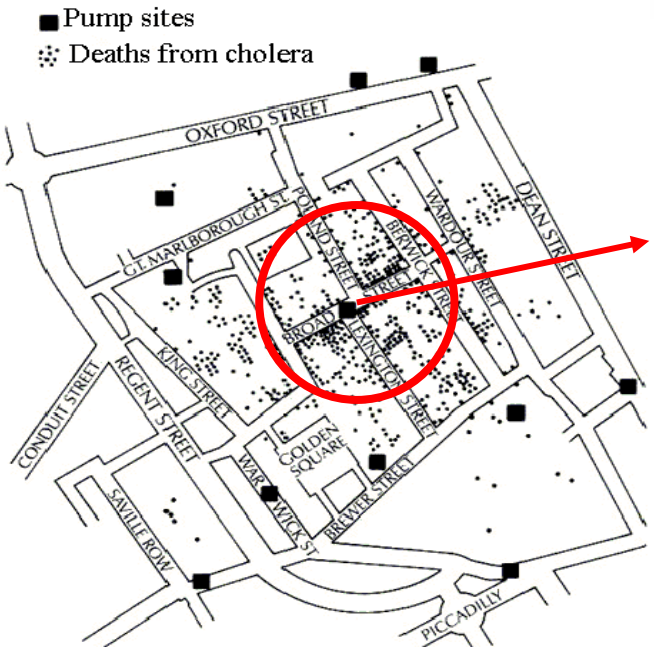


Collect & Curate Data → Discover Patterns, Generate Hypothesis → Test Hypothesis (Experiments) → Develop Theory

? water pump

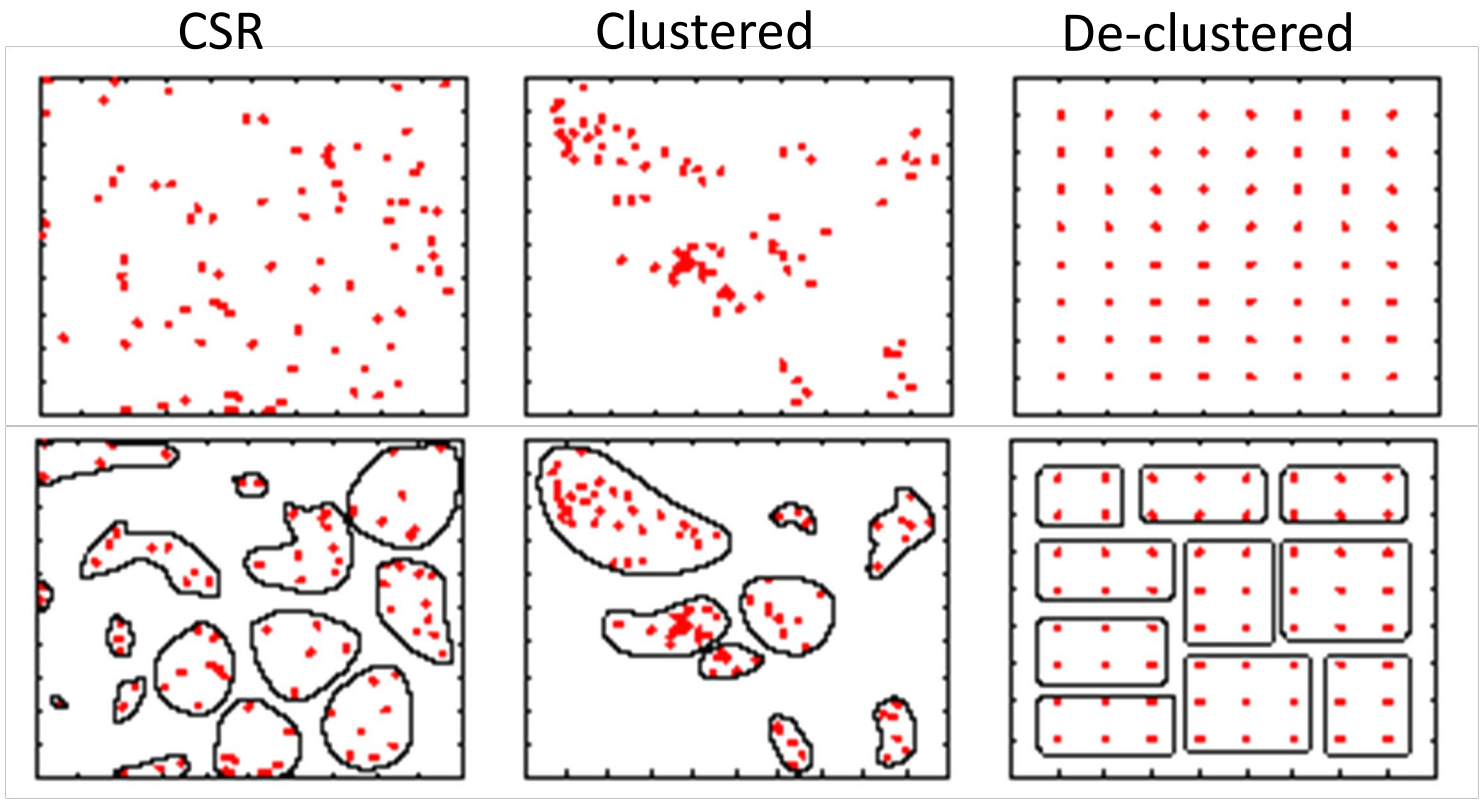
Remove pump handle

TURNING POINTS IN SCIENCE
GERM THEORY



Societal Impact:
Sewage system,
Drinking water supply,
Lower urban density (parks), …

Do Clustering Approaches Give Spurious Results?

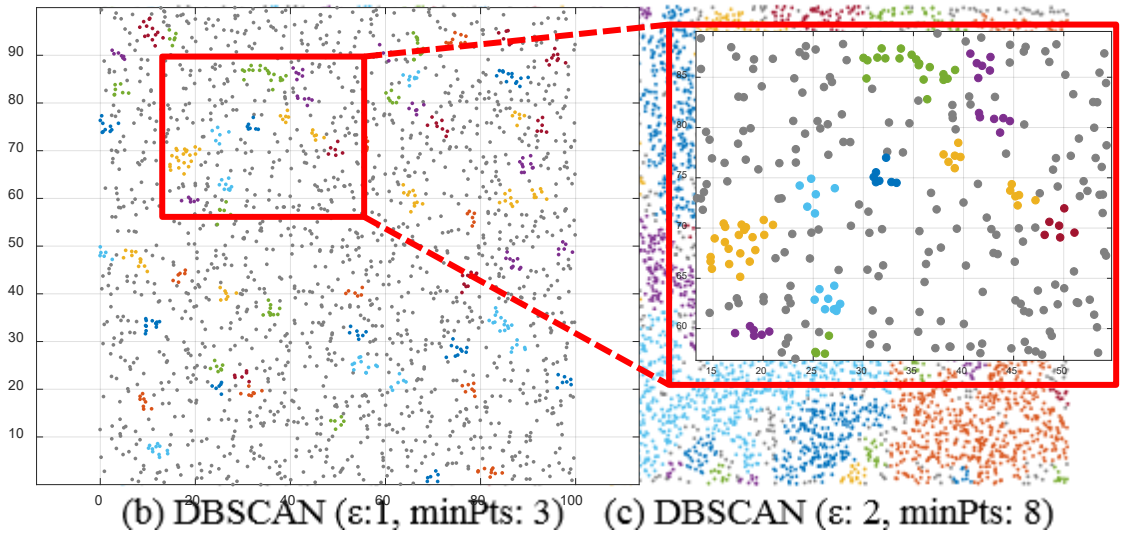


Which approaches we discussed are robust against noise?

- DBSCAN can separate cluster points from noise, but:
  - Points can be close to each other just due to natural randomness

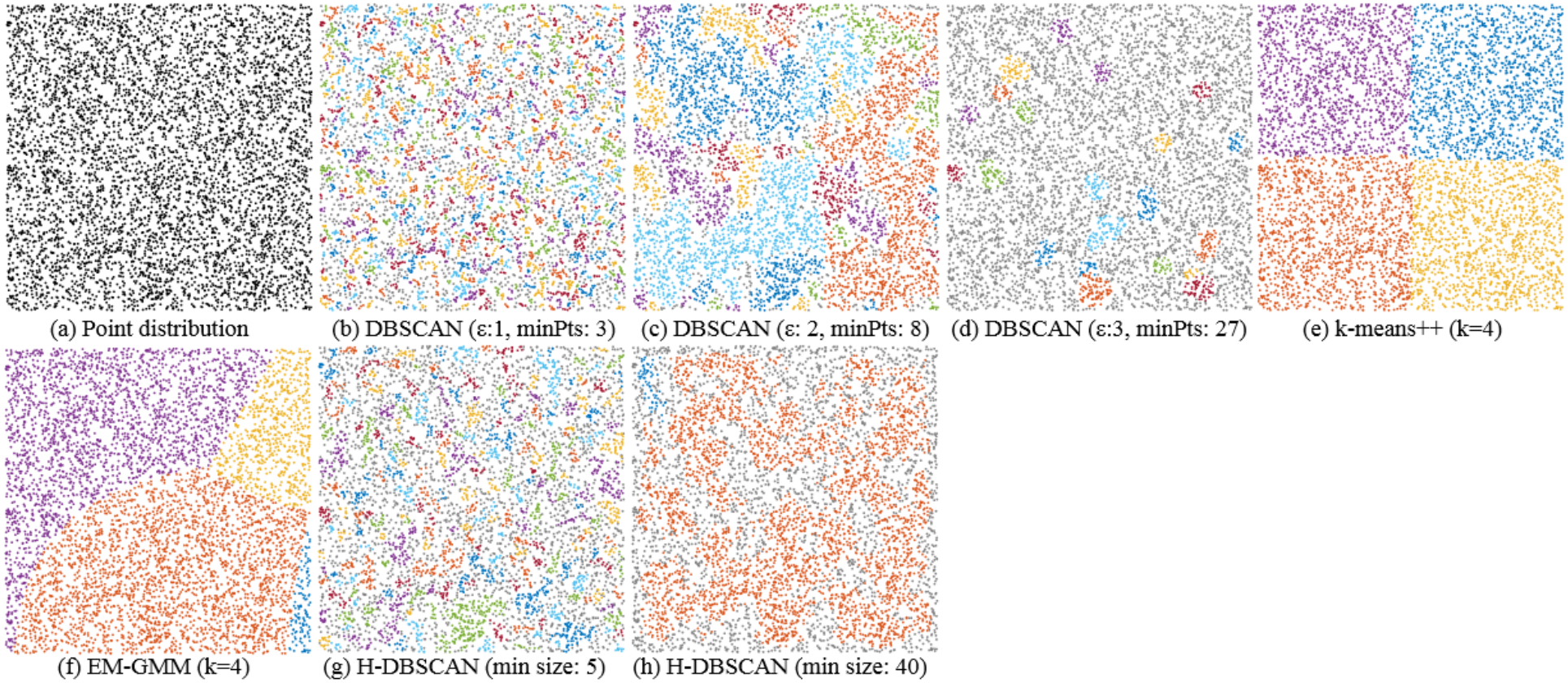
Do Clustering Approaches Give Spurious Results?

- HDBSCAN paper:
  - *"**small clusters** of objects that may be highly similar to each other just by chance, that is, as a **consequence of the natural randomness** associated with the use of a finite data sample"*
- What does small mean?
  - Can we directly define small? (e.g., size = 2, 5 (default in Python lib), 20, 40?)
  - Small cannot be a hard threshold



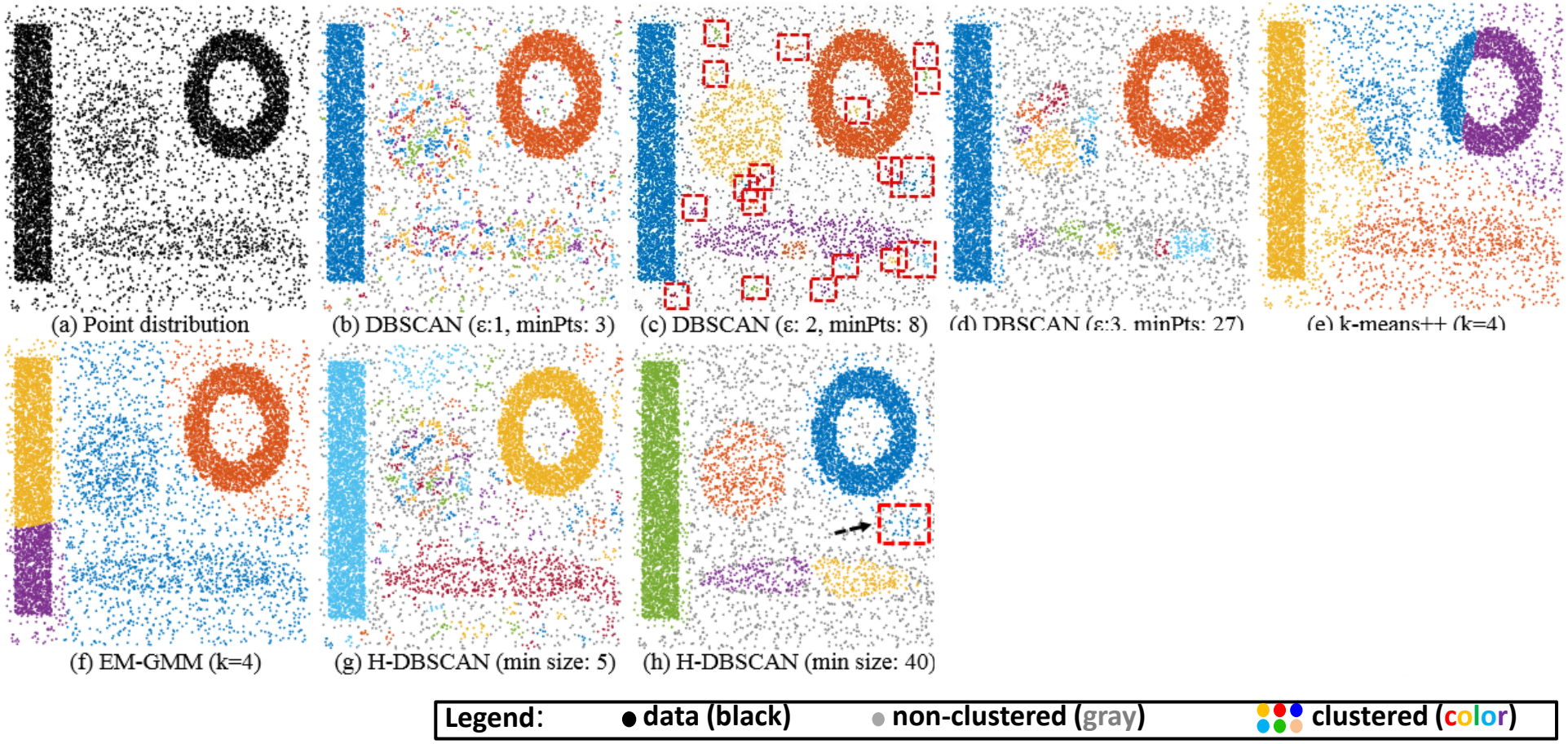
Do Clustering Approaches Give Spurious Results?

Random noise (no cluster)



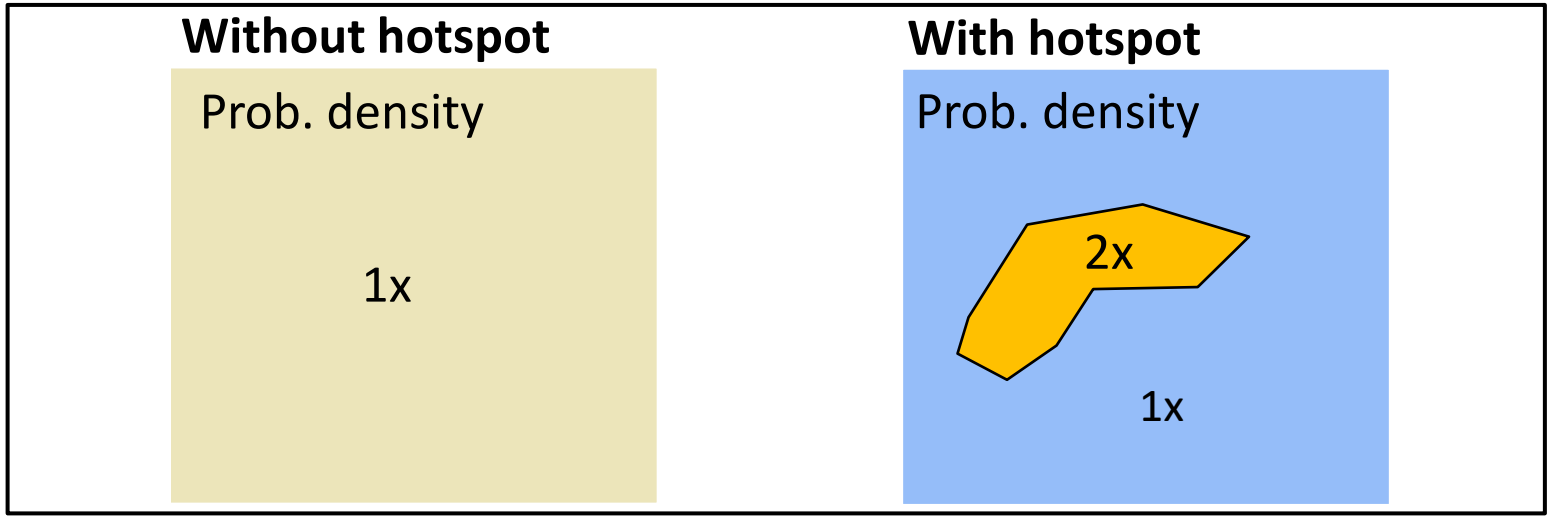
Do Clustering Approaches Give Spurious Results?

Clusters + random noise



What is a Hotspot?

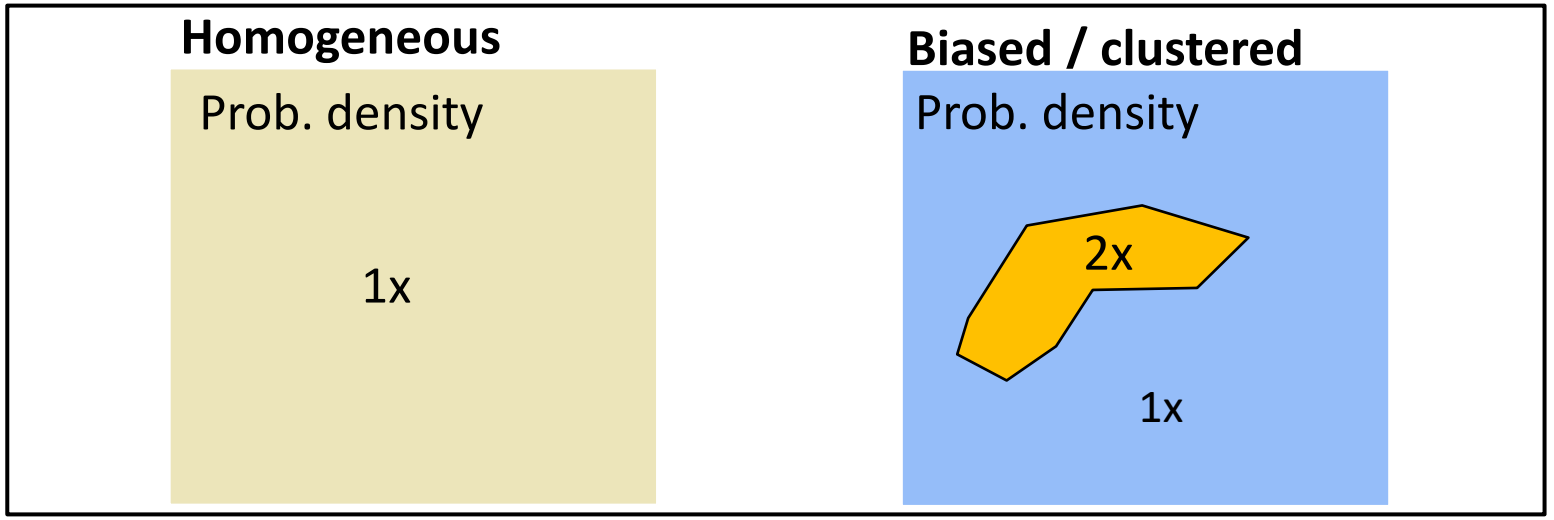
- Example
  - A spatial region that has a higher probability density of generating points (or incidents) of certain events (e.g., disease, crime) compared to the rest.



Spatial point process

Point Processes

- Point process
  - A statistical process that generates the point distribution



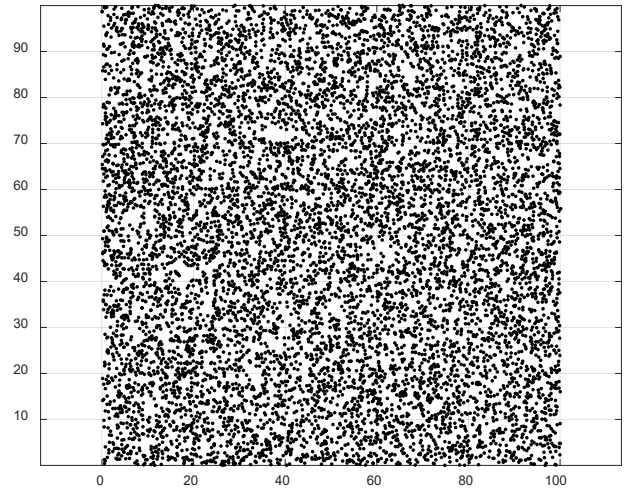
- Homogeneous point process
  - Identical probability density across all locations
- Biased/clustered point process
  - Higher probability density inside clusters and lower outside

For discrete data: Use probability mass at each location (e.g., observations at a fixed number of locations)

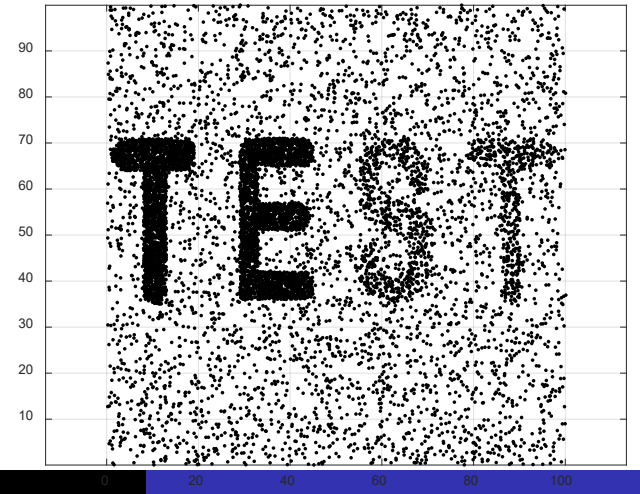
Point Process vs. Data

- In applications, we cannot directly observe the point processes
- We observe the resulting point distribution

Random data (no cluster)



Clustered data (four letters)



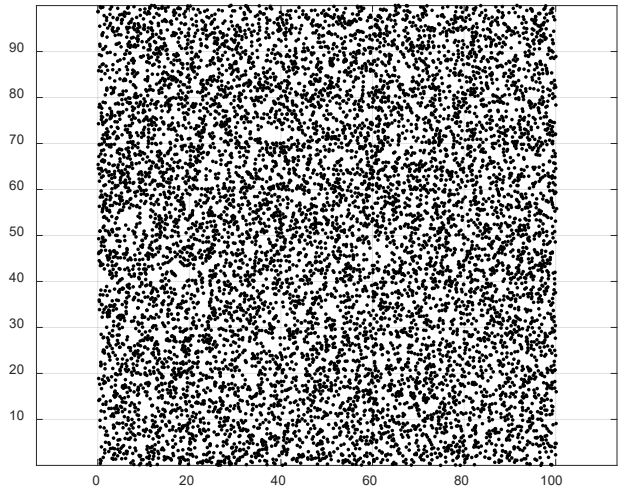
Is there a hotspot?

Will DBSCAN work?

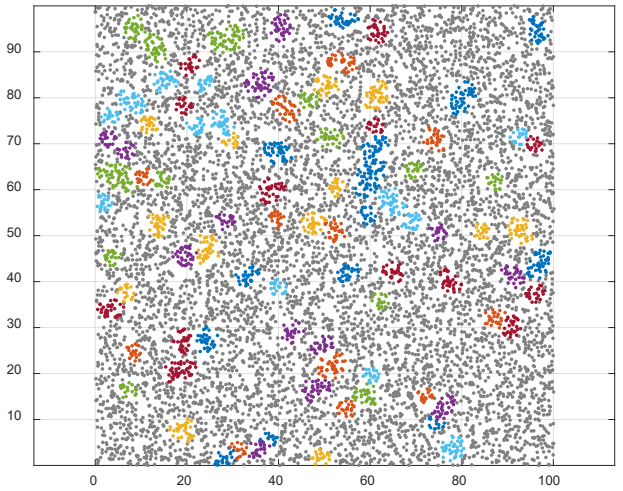
Point Process vs. Data

- In applications, we cannot directly observe the point processes
- We observe the resulting point distributions (realizations)

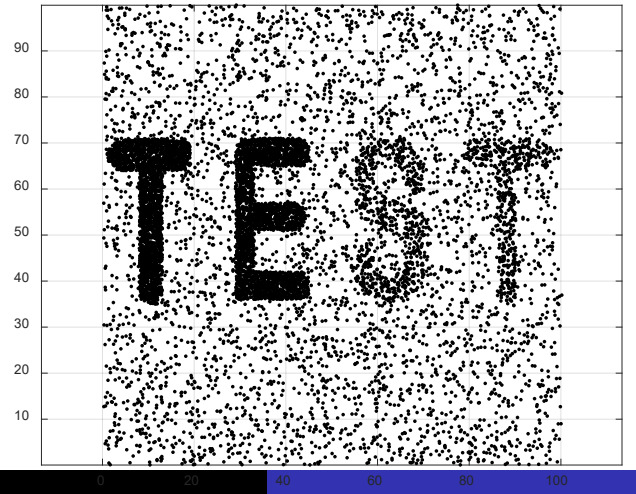
Random data (no cluster)

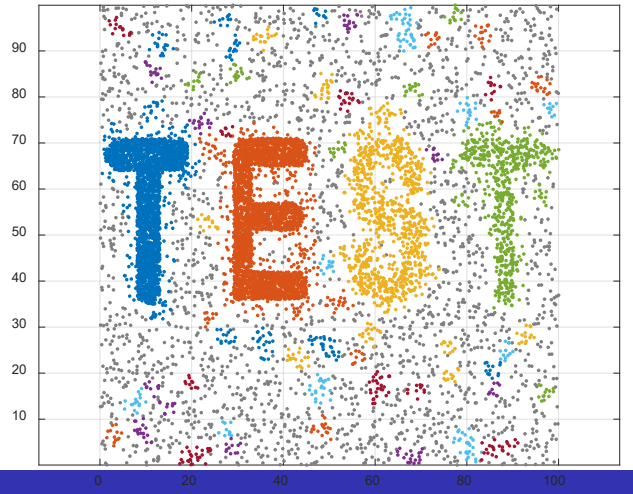


DBSACN results



Clustered data (four letters)





Which Process does Data Come From?

- Not so simple
  - There can be numerous candidates
- What can we do?
  - Assume a non-interesting distribution (a.k.a., null)
  - Confirm if data can be generated by such a distribution
- Statistical hypothesis testing

Hypothesis Testing

- Widely used in many domains
  - Drug experiments
  - Effects of smoking…

- Two hypotheses
  - **Null hypothesis H_0**: If true, then data is likely not generated by an interesting distribution, or, there is no interesting discovery from the data
  - **Alternative hypothesis H_1**: Often covers anything that differs from the null hypothesis

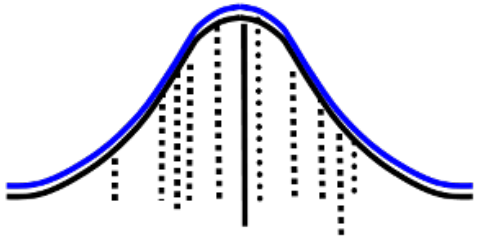
Hypothesis Testing

- Example

Do smokers weigh the same as non-smokers?

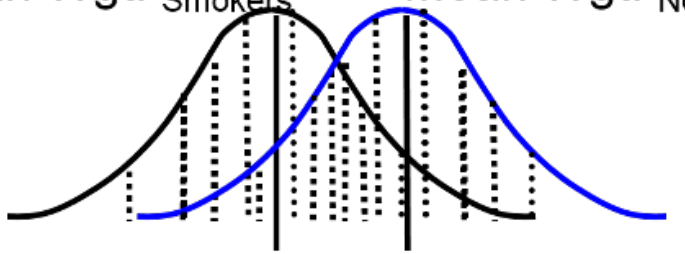
Null Hypothesis (H_0): the average weight does *not* differ

H_0: Mean Wgt. $_{smokers}$ = Mean Wgt. $_{Non\text{-}smokers}$



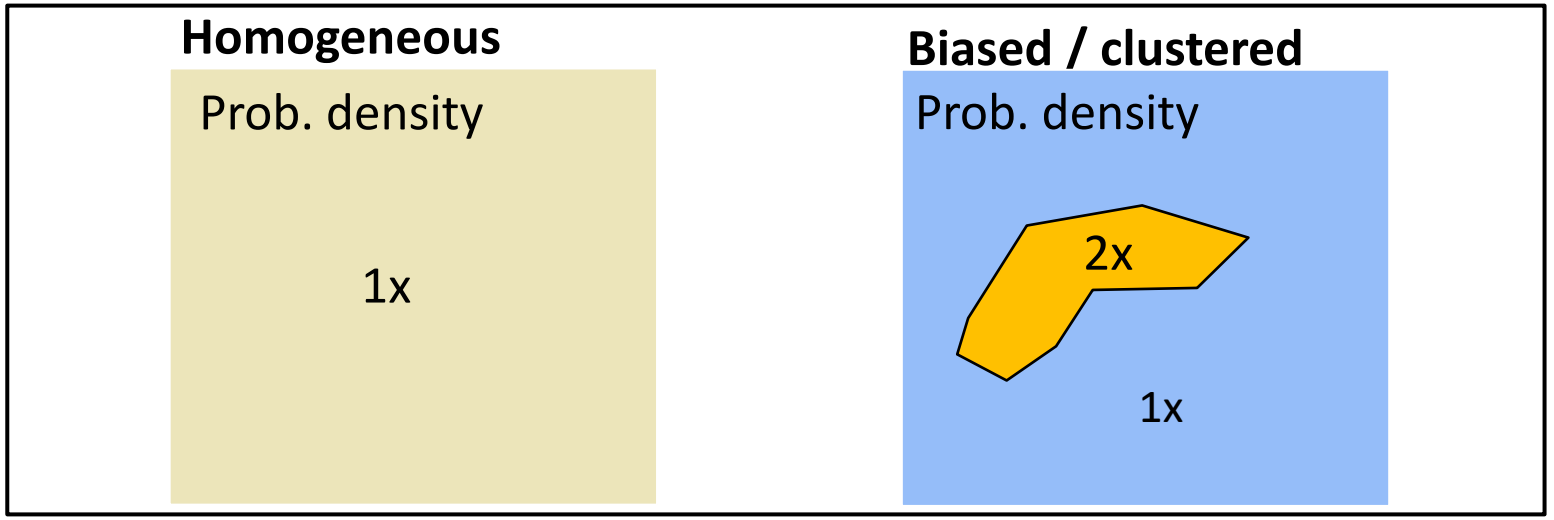
Alternative Hypothesis (H_A): the average weights differ

H_A: Mean Wgt. $_{Smokers}$ $\neq$ Mean Wgt. $_{Non\text{-}smokers}$



Hypotheses for Hotspot Detection

- Point process
  - A statistical process that generates the point distribution



- Homogeneous point process
  - Identical probability density across all locations

→ **Null hypothesis H_0**

Equivalently, p=q for all regions

p: prob. density inside a region

q: prob. density outside

- Biased/clustered point process
  - Higher probability density inside clusters and lower outside

→ **Alternative hypothesis H_1**

p>q for some regions

How to Test the Hypothesis?

- Point process is a random process, so theoretically any point distribution can be generated, with some probability
  - Need a probability threshold to help decide

- Significance level α

How to Test the Hypothesis?

- Significance level α
  - A threshold on the probability of data being generated by H_0
  - If $P(Data \mid H_0) < \alpha$, we reject the null hypothesis; otherwise not
    - By rejecting null, we say the pattern is statistically significant (or significantly different from null)
  - Common values: 0.01, 0.05

- $P(Data \mid H_0)$ is called **p-value**
  - Also called a Type-I error

Quiz

- Which of the following is a threshold?
  A. Significance level
  B. P-value

- Which of the following are true?
  A. Point process is often part of our observation
  B. Point process often cannot be observed
  C. Rejecting the null hypothesis H_0 means data cannot be generated by H_0
  D. P-value = 0 means data cannot be generated by H_0

Key Concepts So Far

- Spatial point process
- Hypothesis testing (especially null hypothesis)
- Significance level
- **P-value**

Next: How to score and enumerate candidates of hotspots and calculate their p-value?

P-value Estimation for Hotspot Detection

- New key concepts
  - Test statistic: Candidate (pattern) scoring
  - Search space: Computation
  - Monte Carlo simulation: Probability via simulation

Test Statistic

- Consider as a score calculated using data samples
- Natural interpretation

Question: What might be a test statistic for the example on the right?

T-test: $$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

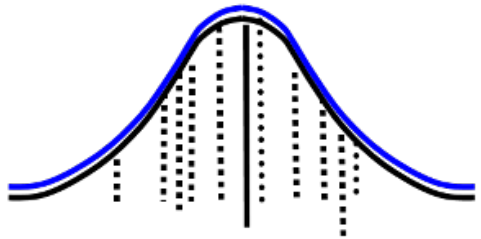
Normalization by standard dev.

Use a standard T-table to find the p-value

Do smokers weigh the same as non-smokers?

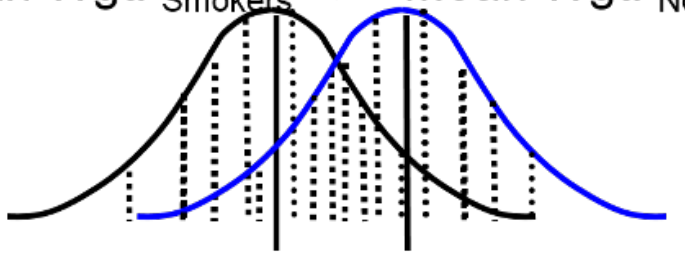
Null Hypothesis (H_0): the average weight does *not* differ

H_0: Mean Wgt. $_{smokers}$ = Mean Wgt. $_{Non\text{-}smokers}$



Alternative Hypothesis (H_A): the average weights differ

H_A: Mean Wgt. $_{Smokers}$ $\neq$ Mean Wgt. $_{Non\text{-}smokers}$



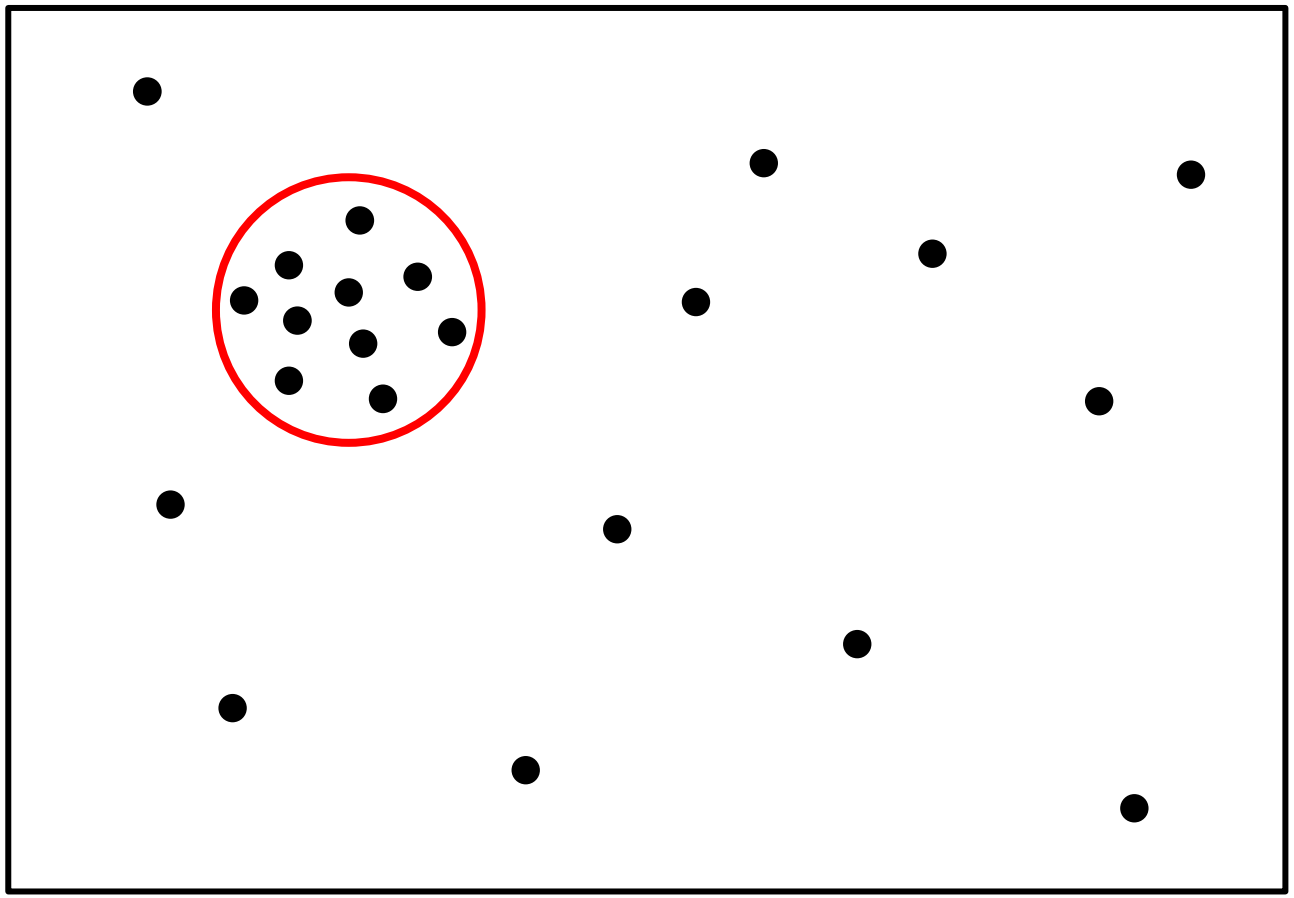
Test Statistics for Hotspot Detection

- H_0: p = q for all regions
- H_1: there exists a region where p > q

- Given a candidate region (hotspot pattern), we can partition the data into two groups
  - What might be a test statistic?
    - What is the characteristic of a hotspot?
    - We need to differentiate data generated by H_0 and H_1
    - How to define a test statistic so that:
      - Data with a hotspot can have a high value
      - Data with no hotspot has a low value

Test Statistics for Hotspot Detection

- An example point distribution



Where is the hotspot?

How did you decide?

density

What is a test statistic?

density

Does the hotspot have the maximum test statistic value?

Why does density not work?

- Monotonicity
- Strong bias towards smaller candidates
  - Draw an infinitely small circle around one point
    - Is this single point pattern with an extremely (or infinitely) high density meaningful? Why?
- Similarly, density ratio does not work either

Inspiration from Density

- Why did you use density?
  - Why not just point count?
- We need to compare candidates with different sizes
- Density provides a way to normalize the score, and make candidates comparable
  - Normalizing by area (density d = n/a)

Spatial Scan Statistic (Software: SaTScan)

- Widely used hotspot detection framework
- Normalize by a probabilistic view
- **Likelihood ratio**
  - Recall definition of likelihood in EM

The methods are commonly referred to as “scan statistics”

Likelihood Ratio (LR)

- Normalize using the likelihood of H_0

$$LR = \frac{Likelihood(H_1)}{Likelihood(H_0)}$$

- Interpretation: How much more likely that the candidate is generated by H_1 than by H_0?

Likelihood Ratio (LR)

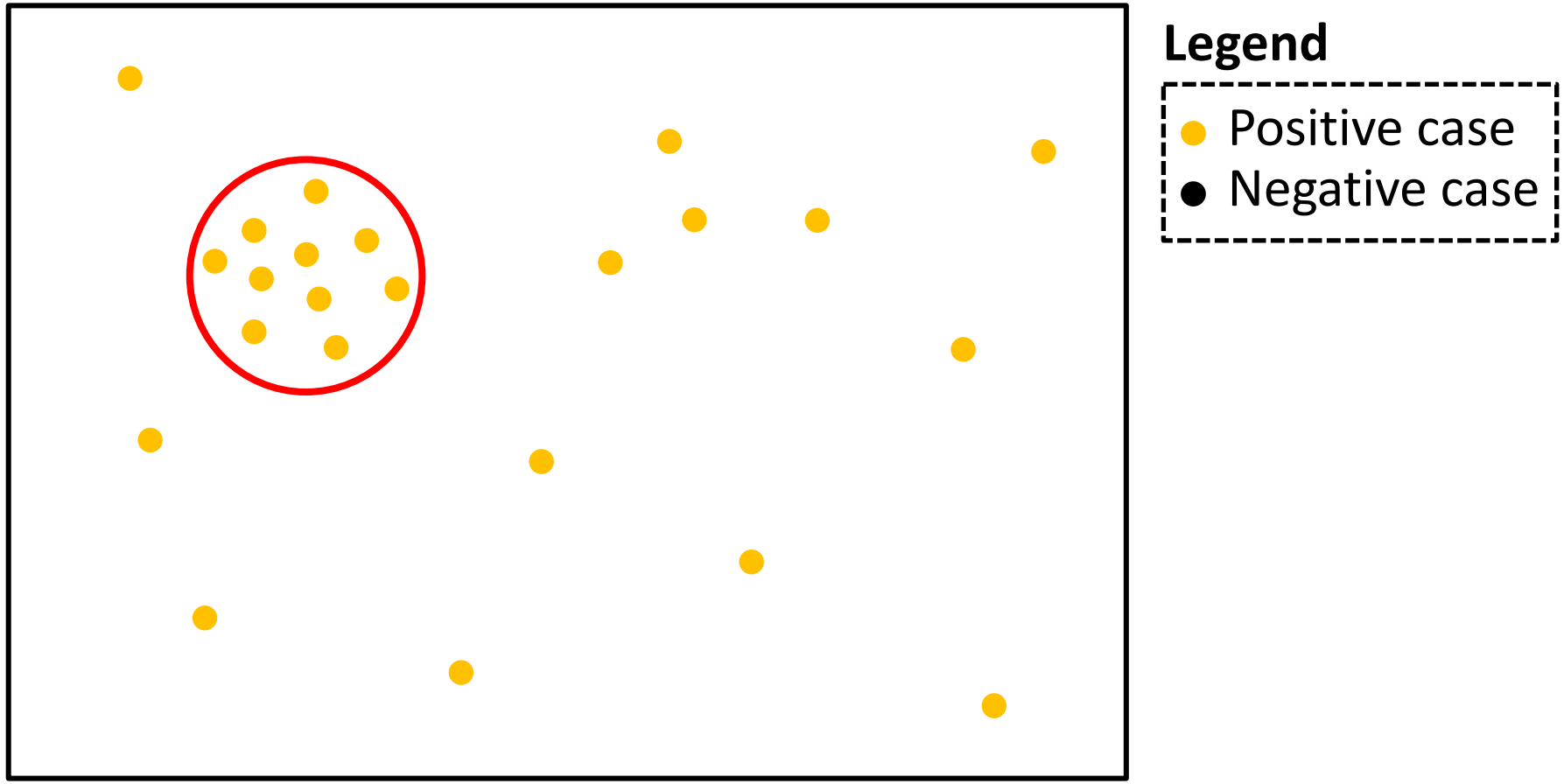
- To calculate likelihoods, we need to first assume data is generated by a statistical distribution (a more specific point process)
  - In EM, we used multivariate gaussian distribution

$$LR = \frac{Likelihood(H_1)}{Likelihood(H_0)}$$

- Examples with Bernoulli-based and Poisson point processes

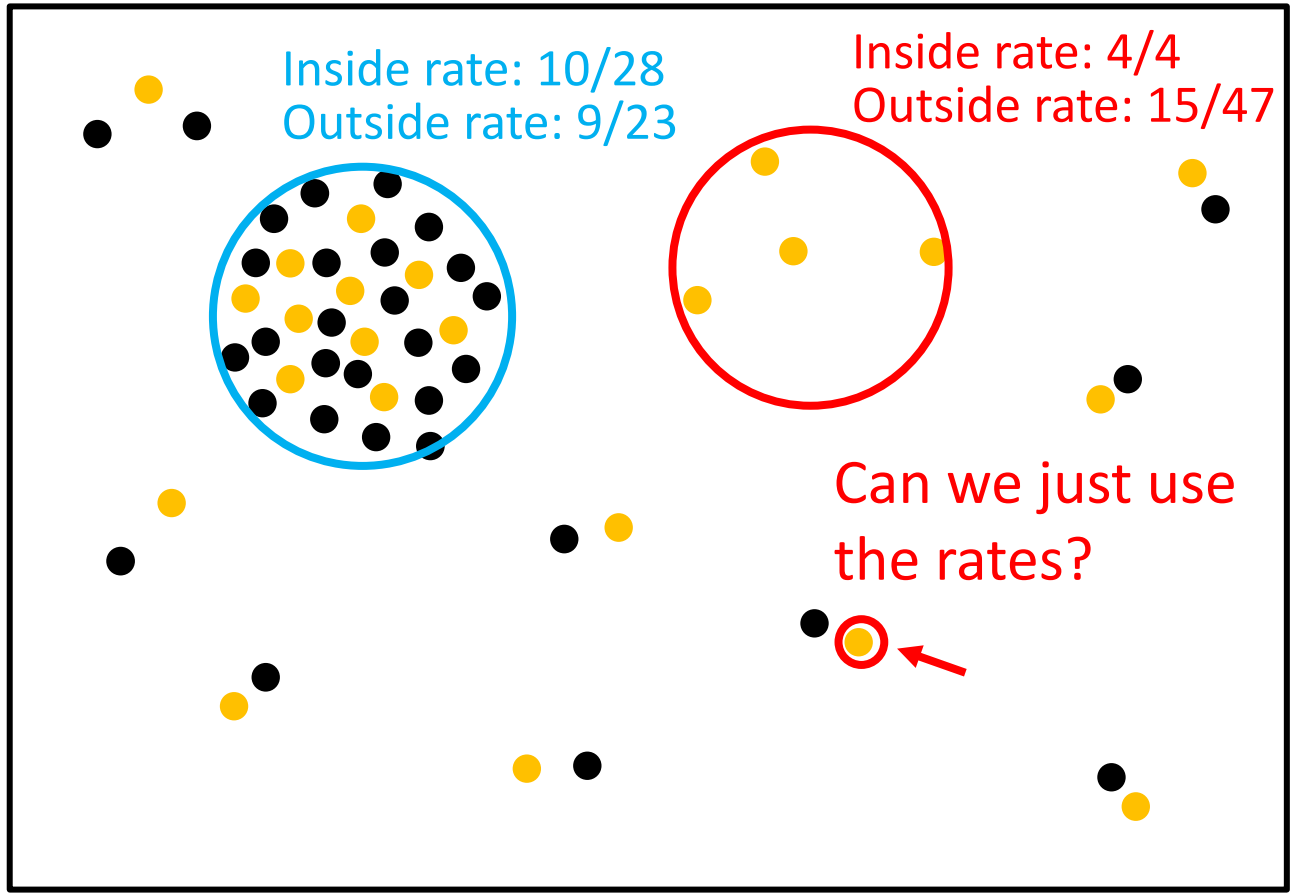
LR for Bernoulli-based Point Process

- New concept: Control points vs. case points
  - Example: Underlying population for a disease
- Why?



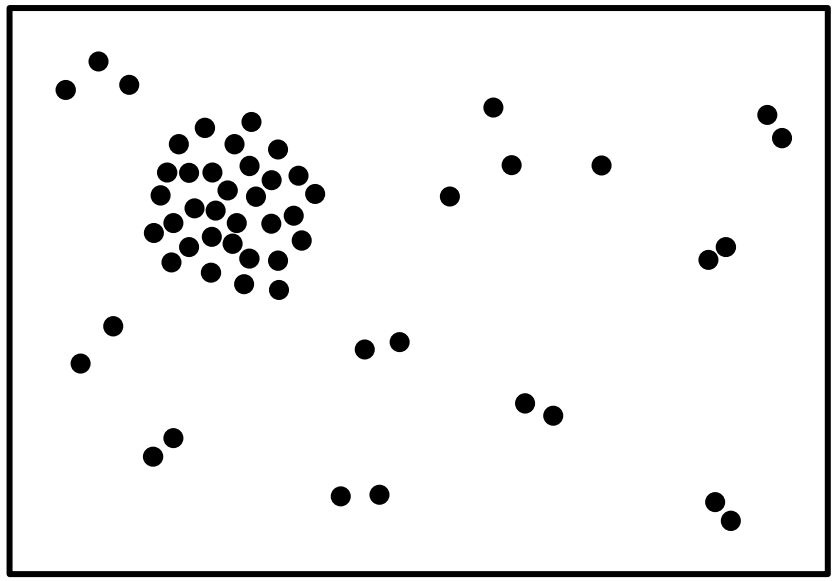
LR for Bernoulli-based Point Process

- New concept: Control points vs. case points
  - Example: Underlying population for a disease
- Why?



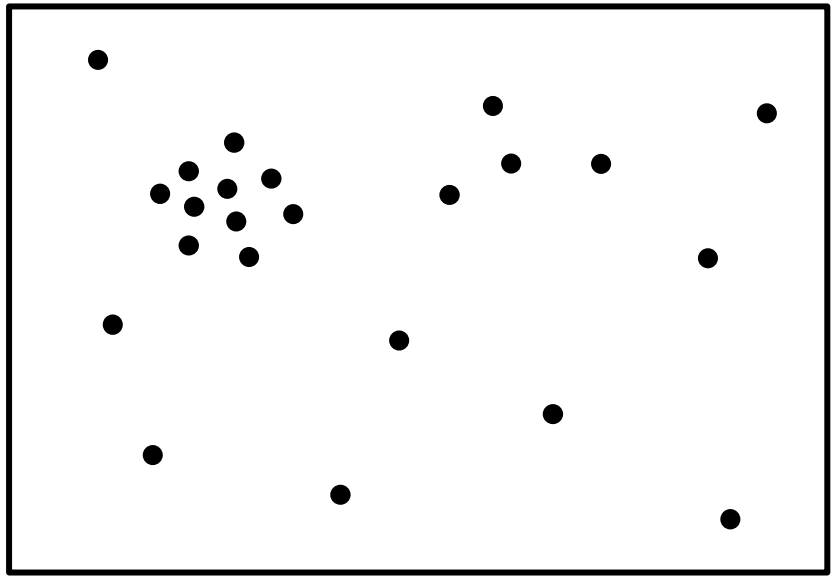
LR for Bernoulli-based Point Process

- We need to consider both case and control points



Control points (positive + negative)

Example: population



Case points (positive only)

Example: disease cases

Bernoulli-based Point Process

- Bernoulli distribution: x~B(p)
  - Example: probability of seeing heads in a coin toss
- What are the two types of points in hotspot detection?
- How does this relate to point process?
  - “heads” as a point being a case point for an event (e.g., disease)
  - “tails”: as a point not being a case point

LR for Bernoulli-based Point Process

- H_0: p = q for all regions
- H_1: there exists a region where p > q
- What is likelihood of H_0 and H_1?
  - How to estimate p?
  - If you do not know how to take derivatives of a function
    - Search online for basics or solutions for similar formats
    - Use Wolfram Alpha
- Derivation on board

Key Notations for Bernoulli version

- Notations (symbols to denote different variables or constants) change from paper to paper
  - Here we list key notations we used during the lecture
  - More important to know what they represent than what symbols are used

- N and n: Total number of control and case points in the entire study area, respectively
- N_z and n_z: Number of control and case points in a candidate region z, respectively
- p and q: Bernoulli probability (probability for a point being a case point); p inside region z, and q outside region z

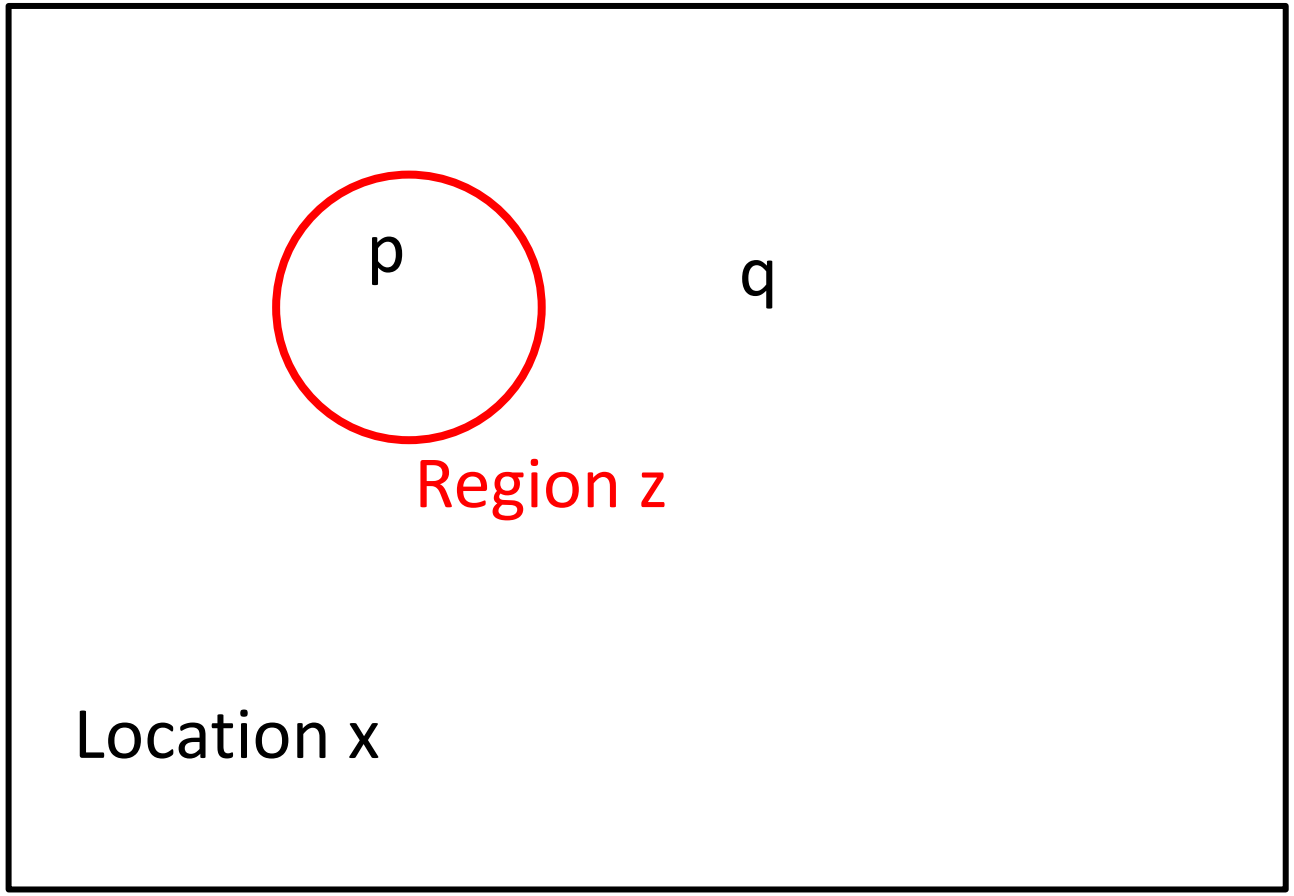
Summary: Bernoulli-based Likelihood

- $L(p,q) = p^{n_z} \cdot (1-p)^{N_z - n_z} \cdot q^{n-n_z} \cdot (1-q)^{(N-N_z)-(n-n_z)}$
- Maximum likelihood estimation
  - Find best p and q to maximize the above likelihood
    - Interpretation: Under the Bernoulli-distribution assumption, find its parameters that have the best chance of generating the observed data
    - For each of the hypotheses
      - H_0: $p = q$ (just replace p with q in the above likelihood function)
      - H_1: $p \neq q$
  - Result (take partial derivatives w.r.t. p and q and set to 0):
    - H_0: $p = q = \frac{n}{N}$
    - H_1: $p = \frac{n_z}{N_z}$, $q = \frac{n-n_z}{N-N_z}$
    - Bring the solutions back to the likelihood function (two versions for two hypotheses) and get the likelihood ratio; this will be the test statistic

LR for Poisson Point Process

- Poisson distribution
  - "… expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event…"
  - **Example use case:** Suppose a person gets 100 calls on average per month, what is the probability of the person receives 10 calls in a month?
  - **Poisson spatial point process:** Assume we know the expected number of points in the whole study area is 100, what is the probability that we observe 90 points?

LR for Poisson Point Process: Two Steps



Step 1. Assuming total number of points follows a Poisson distribution:

Probability of observing k points:

$$f(k;\lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

λ is expected number of points

Step 2: Not enough, also need to estimate probability of observing these N points at their locations

Derivation on board

LR for Poisson Point Process

- Also uses maximum likelihood estimation
- Will see again in machine learning part

From Statistics to Actual Computation

- Key statistical building blocks covered
  - Spatial point processes
  - Hypothesis H_0 and H_1
  - Significance level and p-value
  - Test statistics (likelihood ratios)

- Questions left for computation
  - There are many (infinite) possible regions.
  - How to enumerate the regions and which to use?
  - How to calculate p-value with LR?

Which region to use? Why?

- What region may best reflect hotspot-ness?
- Choice: use the region that has the highest likelihood ratio
  - $R = \underset{R}{\operatorname{argmax}}\, LR(R)$
- Will the following work?
  - Can we use all?
    - We want to find the actual location of a hotspot
  - Can we use multiple at the same time?
    - Multiple testing, shadowing effects

Multiple Testing

- The probability of falsely rejecting H_0 may no longer satisfy significance level

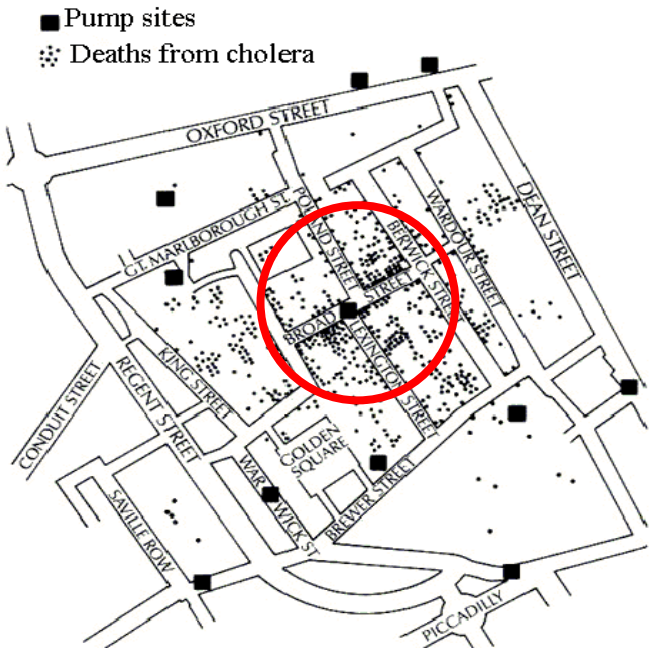
- Example
  - Significance level: 0.01
  - If we perform the test once, the probability of making a false decision is 0.01
  - If we perform the test x times, the probability is?
    - $1 - 0.99^x$
    - $1 - 0.99^{10} \approx 0.096 > 0.01$

How to Enumerate the Regions?

- Search space
  - A.k.a., enumeration space, solution space…
  - Defines the candidates of the final (optimal) solution
    - Candidates for best region in hotspot detection
- Thus far, there is no efficient algorithm that has a search space containing all possible candidates
- We use a reasonable subset as the search space

Different Definitions of Finite Search Spaces

- Circle (most popular, used by SaTScan)
  - Two-point circles (why?)
    - One data point at the center
    - Another data point on the circumference
- Rectangle
  - Four-point rectangles, each on one side
- …

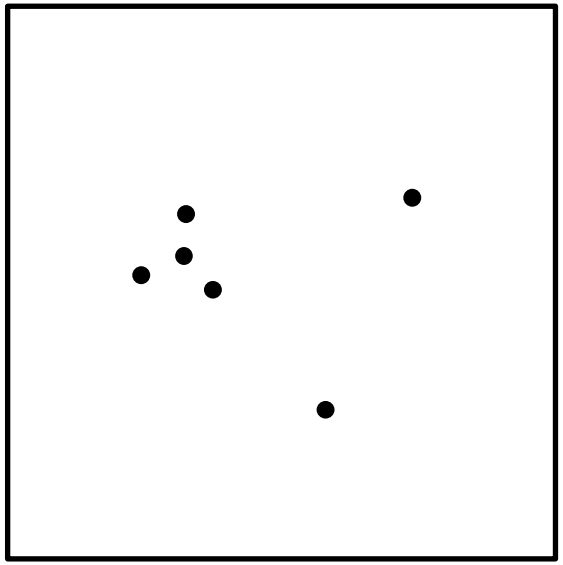


- Search space depends on data size (with these definitions)

Execution Trace on a Small Dataset

Data summary

- Number of points: 6
- Study area: 10 x 10

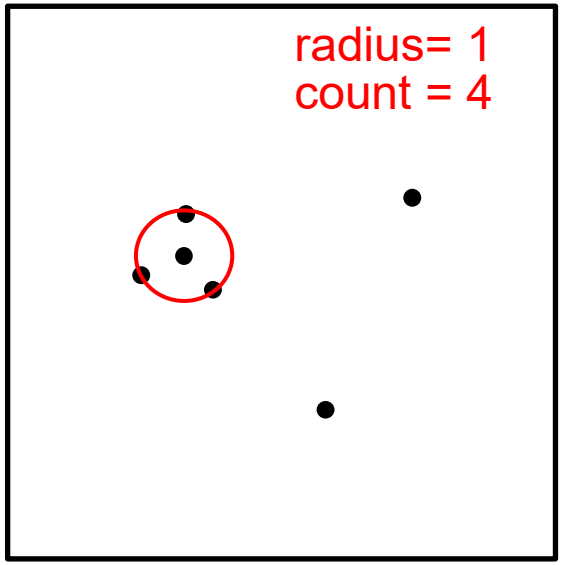


ID	radius	count	Likelihood Ratio (LR)
1			
2			
3			
4			
5			
6			

Execution Trace on a Small Dataset

Data summary

- Number of points: 6
- Study area: 10 x 10

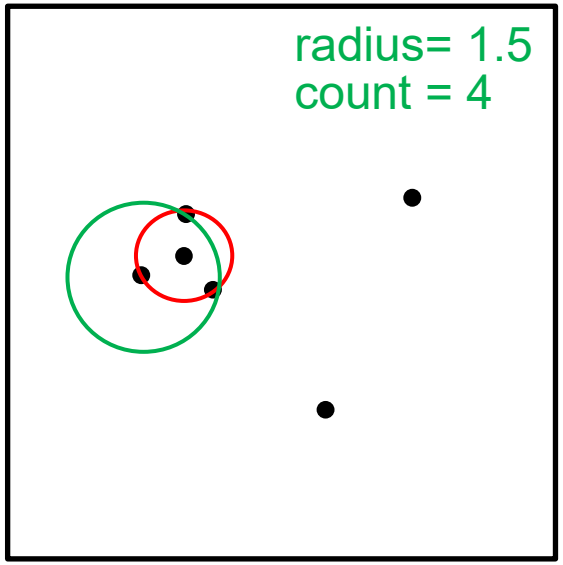


ID	radius	count	Likelihood Ratio (LR)
1	1	4	24064.9
2			
3			
4			
5			
6			

Execution Trace on a Small Dataset

Data summary

- Number of points: 6
- Study area: 10 x 10

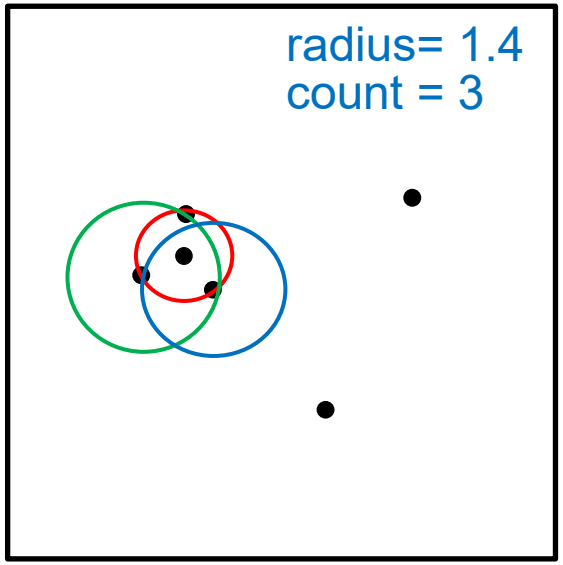


ID	radius	count	Likelihood Ratio (LR)
1	1	4	24064.9
2	1.5	4	1019.96
3			
4			
5			
6			

Execution Trace on a Small Dataset

Data summary

- Number of points: 6
- Study area: 10 x 10

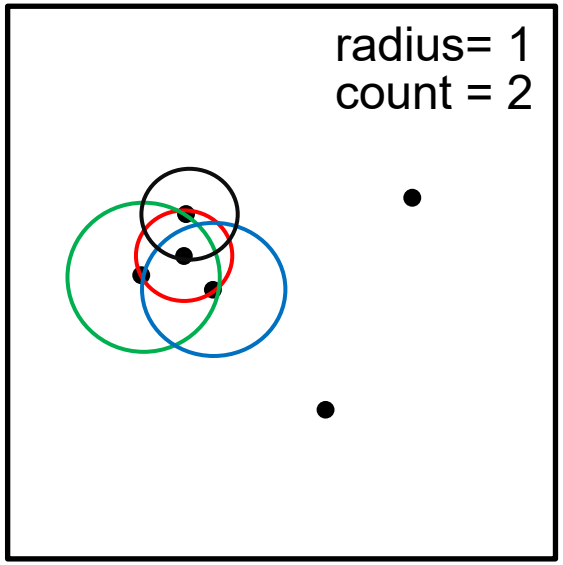


ID	radius	count	Likelihood Ratio (LR)
1	1	4	24064.9
2	1.5	4	1019.96
3	1.4	3	81.10
4			
5			
6			

Execution Trace on a Small Dataset

Data summary

- Number of points: 6
- Study area: 10 x 10

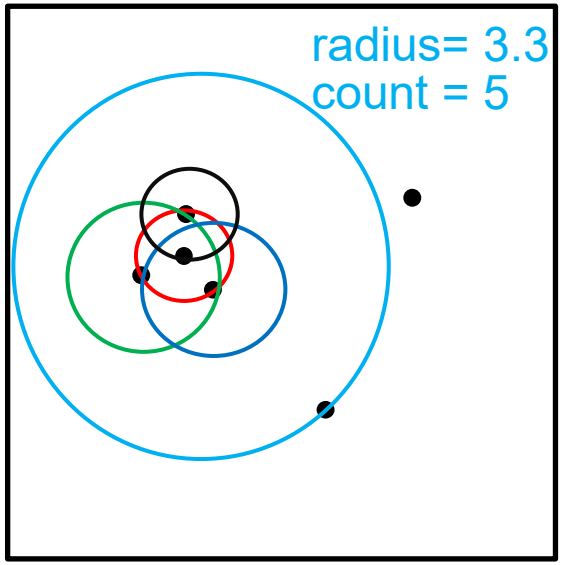


ID	radius	count	Likelihood Ratio (LR)
1	1	4	24064.9
2	1.5	4	1019.96
3	1.4	3	81.10
4	1	2	25.29
5			
6			

Execution Trace on a Small Dataset

Data summary

- Number of points: 6
- Study area: 10 x 10

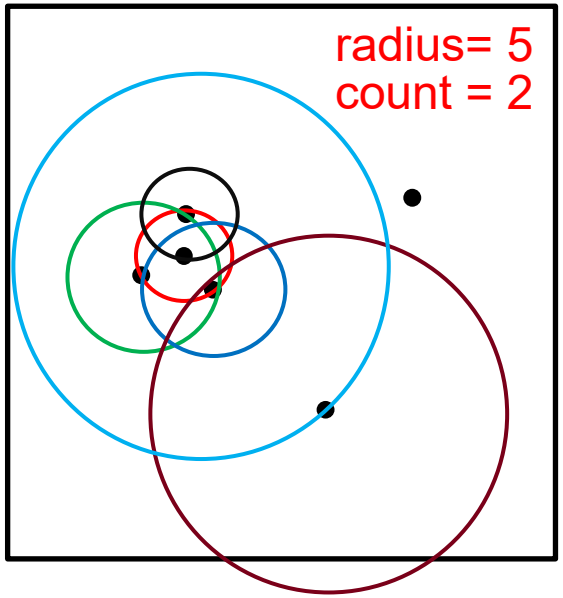


ID	radius	count	Likelihood Ratio (LR)
1	1	4	24064.9
2	1.5	4	1019.96
3	1.4	3	81.10
4	1	2	25.29
5	3.3	5	21.72
6			

Execution Trace on a Small Dataset

Data summary

- Number of points: 6
- Study area: 10 x 10

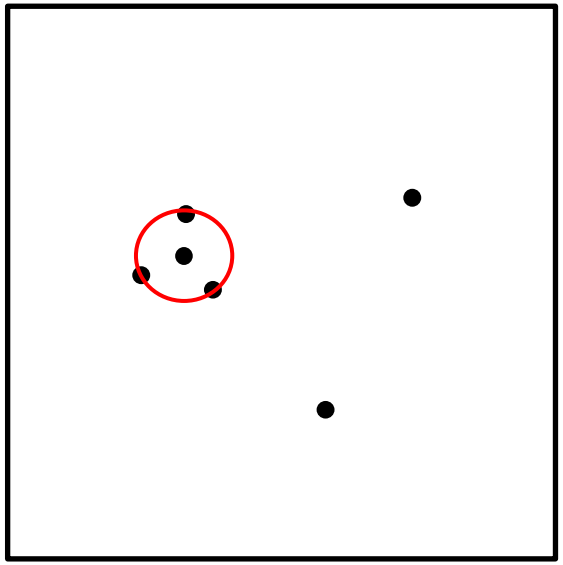


ID	radius	count	Likelihood Ratio (LR)
1	1	4	24064.9
2	1.5	4	1019.96
3	1.4	3	81.10
4	1	2	25.29
5	3.3	5	21.72
6	3	2	1.04

Execution Trace on a Small Dataset

Data summary

- Number of points: 6
- Study area: 10 x 10



ID	radius	count	Likelihood Ratio (LR)
1	1	4	24064.9
2	1.5	4	1019.96
3	1.4	3	81.10
4	1	2	25.29
5	3.3	5	21.72
6	3	2	1.04

What is Needed to Calculate P-Value?

- What does p-value mean?
  - What is the probability that the **best hotspot candidate** in the observed data is generated by the null hypothesis H_0?
- To be more concrete here with our test statistic
  - What is the probability to get an equally-good or better likelihood ratio value in a data generated by H_0?

How to Calculate the Probability (p-value)?

- What can we do if the distribution of the test statistic is standard normal $N(0,1)$?
- Maximum likelihood ratio (LR) is obtained through a complex process involving:
  - Spatial distribution of control points
  - Point process
  - Maximum likelihood estimation (not the max LR!)
  - Enumeration algorithm
  - Data size
  - …

No closed-form solution

Difficult to pre-generate a look-up table

(example: https://www.statisticshowto.com/tables/t-distribution-table/)

Monte Carlo Simulation

- A very powerful and convenient tool to estimate probability
- Quite straightforward in most scenarios
- Simulate the probability (or distribution of a random variable) by generating a large number of random trials
  - Examples
    - Coin-toss
    - π

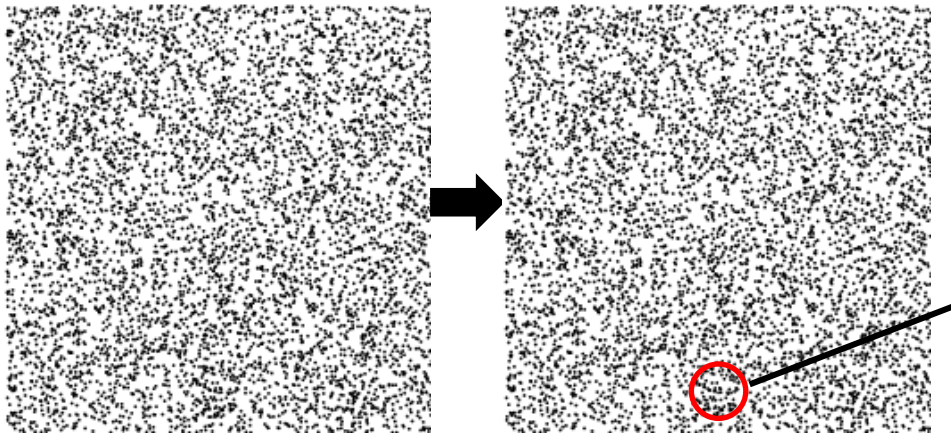
Monte Carlo Simulation for Significance Testing

- Generate M trials (e.g., 1000) of Monte Carlo simulation

For each trial in Monte Carlo simulation

Generate a random point distribution using H_0

Find maximum likelihood ratio by Scan Statistic



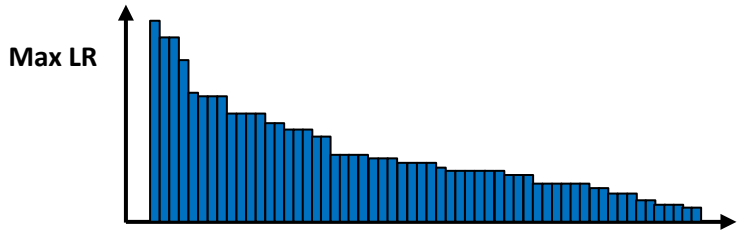
Insert

Monte Carlo table

Trial ID	Max likelihood ratio (log)
1	49
2	70
…	…
126	**61**
…	…
1000	to be computed

Sort in desc. order

Trial ID	Max likelihood ratio (log)
17	101
22	97
…	…
3	90
…	…



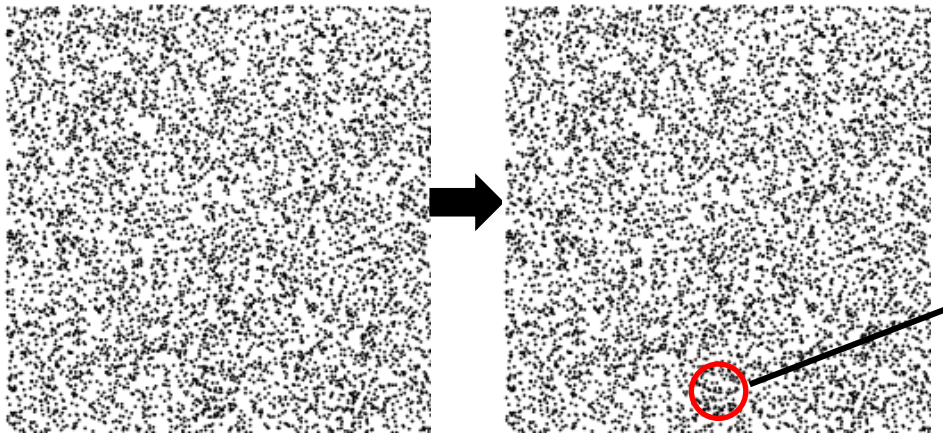
Monte Carlo Simulation for Significance Testing

- Generate M trials (e.g., 1000) of Monte Carlo simulation

For each trial in Monte Carlo simulation

Generate a random point distribution using H_0

Find maximum likelihood ratio by Scan Statistic



Insert

Monte Carlo table

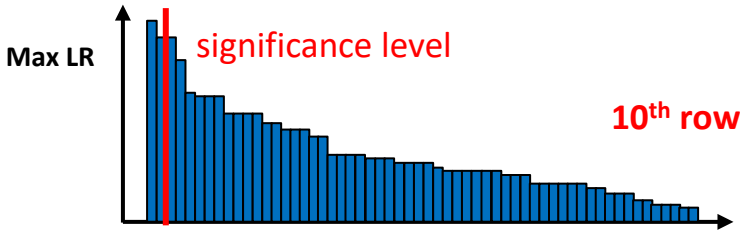
Trial ID	Max likelihood ratio (log)
1	49
2	70
…	…
126	**61**
…	…
1000	to be computed

Sort in desc. order

Trial ID	Max likelihood ratio (log)
17	101
22	97
…	…
3	**90**
…	…

Determine significance (sig. level: 0.01)

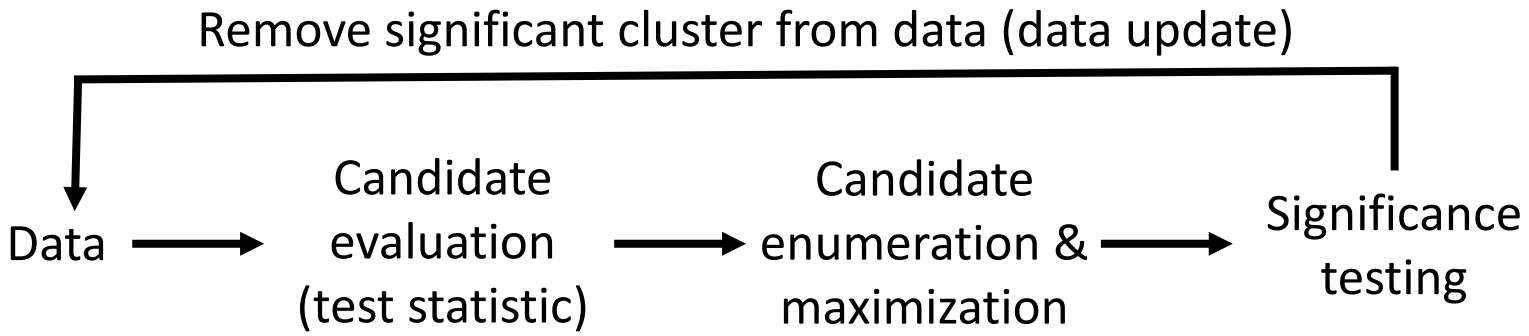
A cluster in real-data is { significant, if LR > 90; chance pattern, otherwise



How to Detect Multiple Clusters?

- After a significant hotspot is detected
  - Remove the corresponding points
  - Remove the corresponding space (no points should be put in there in Monte-Carlo simulation)
- Re-run the same algorithm on the remaining data
- Terminate if no significant hotspot is detected

Summary of Key Steps

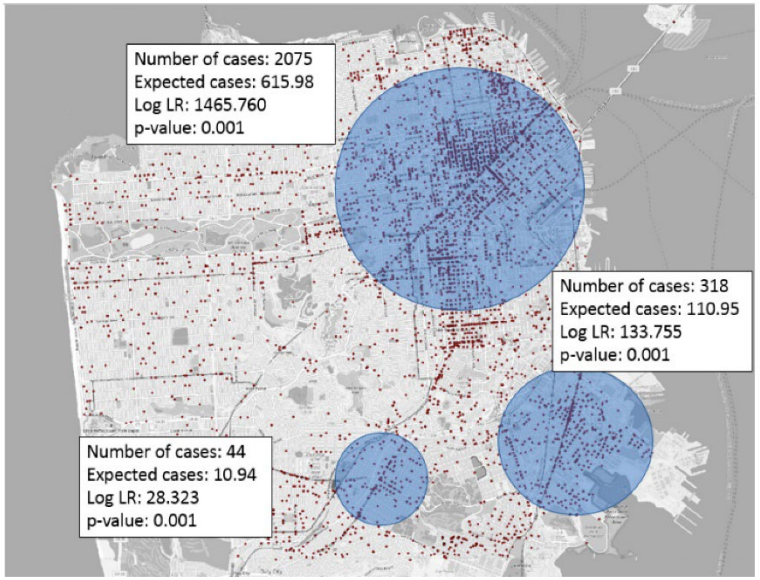


Example Results

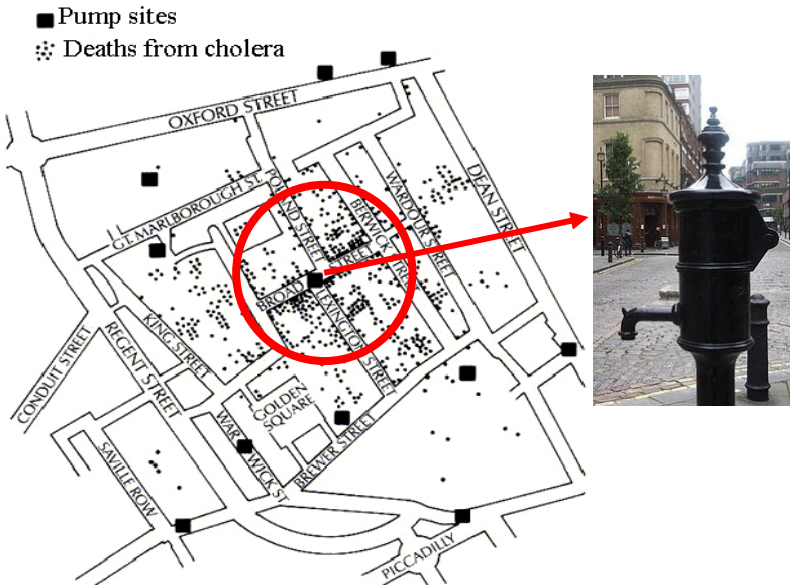
SaTScan™
Software for the spatial, temporal, and space-time scan statistics

NIH NATIONAL CANCER INSTITUTE
Division of Cancer Control & Population Sciences

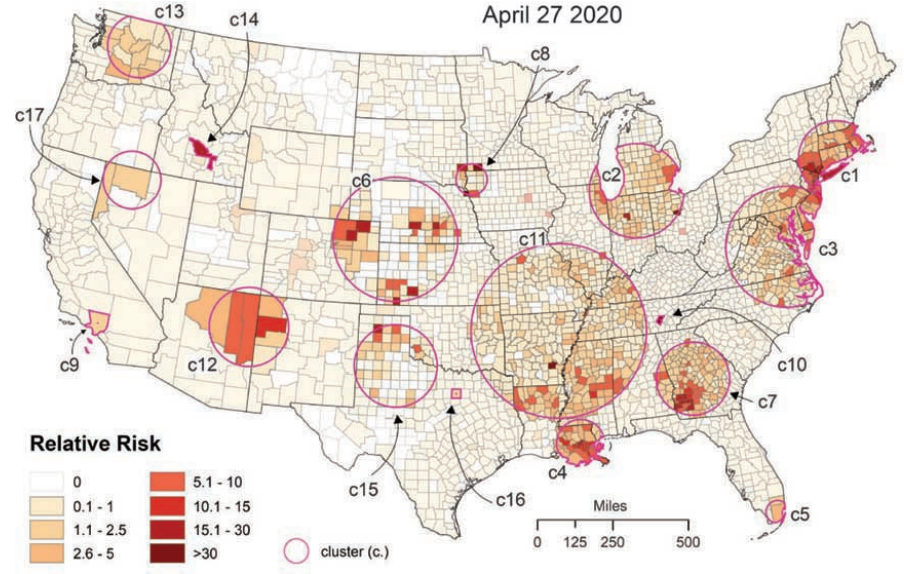
Surveillance Research Program



Crime hotspot Example



London Cholera outbreak



COVID-19 Example

Topics

- Extensions and reinforcements

- Software: SaTScan
  - Documentation
  - Run examples
  - Installed on lab machines

Extensions (and Reinforcements)

- Point processes for different problems
  - You might be able to carry out many of those
  - Real-valued: Life-expansion, Airbnb price
  - Categorical: Diversity
  - Space-time...
- Hotspots of different shapes
  - Ring (e.g., serial criminals), irregular, linear (e.g., on a road network)
- Data
  - Social network, trajectories, polygon, raster...

Point Processes (Reinforcement)

- Real valued, or continuous, point processes
  - Bernoulli and Poisson can only handle count data
    - Number of disease cases, crime cases
  - Examples of real-valued hotspot patterns
    - Regions with significant longer or shorter life expansions
    - Survival time
    - Abnormally higher or lower Airbnb prices
    - Air quality indices
    - Crop yield
    - House insulation
    - …

Real-Valued/Continuous Point Process

- Which statistical distribution naturally models distribution of real values?
  - Normal distribution
  - Similar: Family of exponential distributions

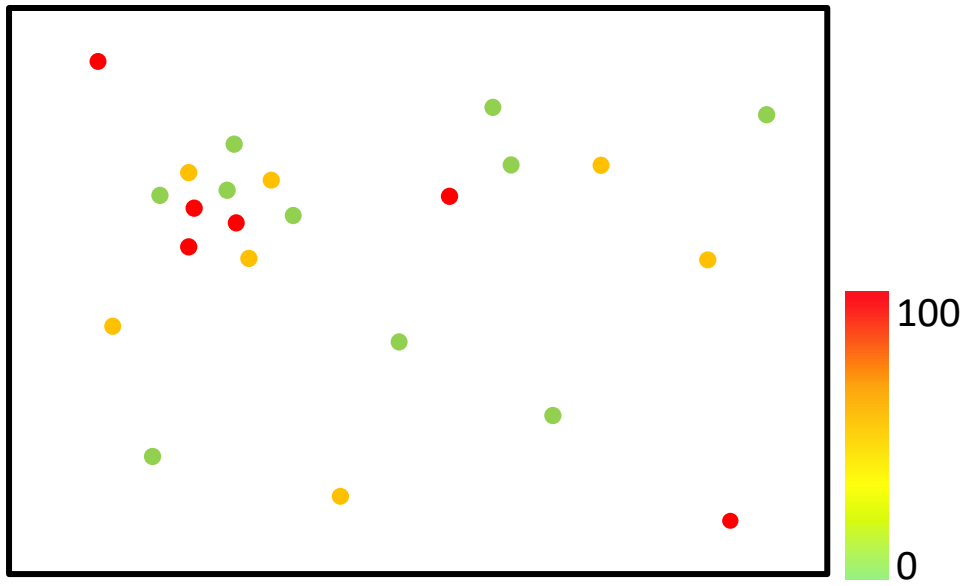
- Normal point process
  - Assuming the value on each point follows a normal distribution
    - Mean μ
    - Variance σ^2
    - Simpler than EM (only scalars, no vectors/matrices)

Real-Valued/Continuous Point Process

- Normal point process $\mathrm{N}(\mu, \sigma^2)$: assuming the value on each point follows a normal distribution
  - Mean μ
  - Variance σ^2
- Given a region, denote:
  - Inside values are generated by $\mathrm{N}(\mu_1, \sigma^2)$
  - Outside values are generated by $\mathrm{N}(\mu_2, \sigma^2)$
- Hypotheses
  - H_0: $\mu_1 = \mu_2$ for all regions
  - H_1: There is a region with $\mu_1 > \mu_2$

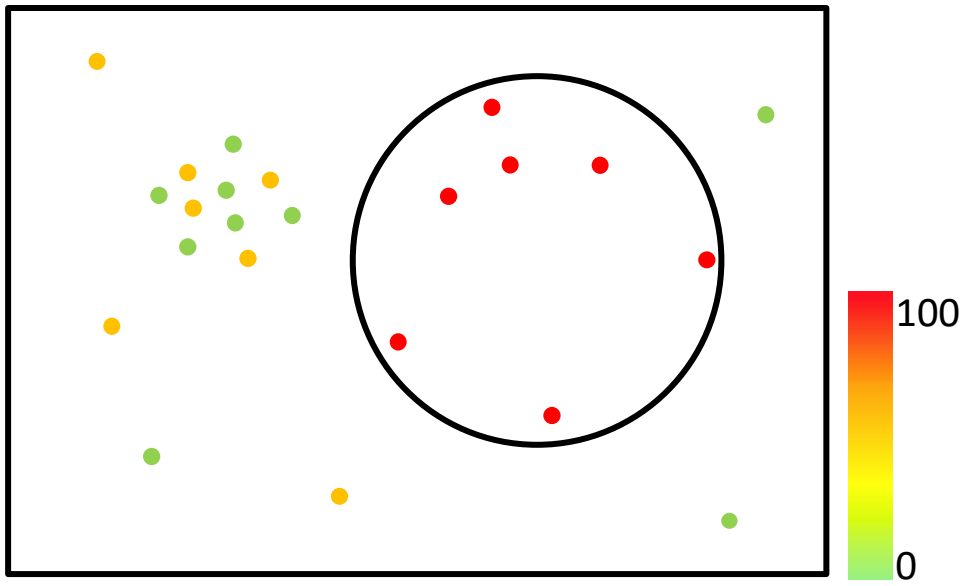
Real-Valued/Continuous Point Process

- Note: No longer about point density (only values)
- Example:



H_0

$\mu_1 = \mu_2$ for all regions



H_1

There exists a region with $\mu_1 > \mu_2$

Multinomial Point Process

- Distribution of categorical values

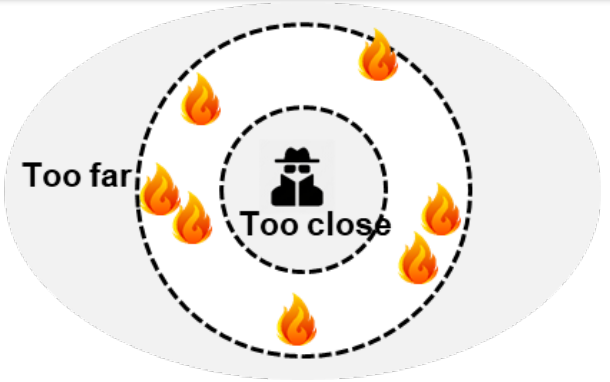
- Examples
  - Regions with a distinct composition of diseases
  - Regions with a distinct composition of disease symptoms
  - Regions with abnormally higher/lower biodiversity
  - Crop types
  - Tree types
  - POI (business) types…

Extensions

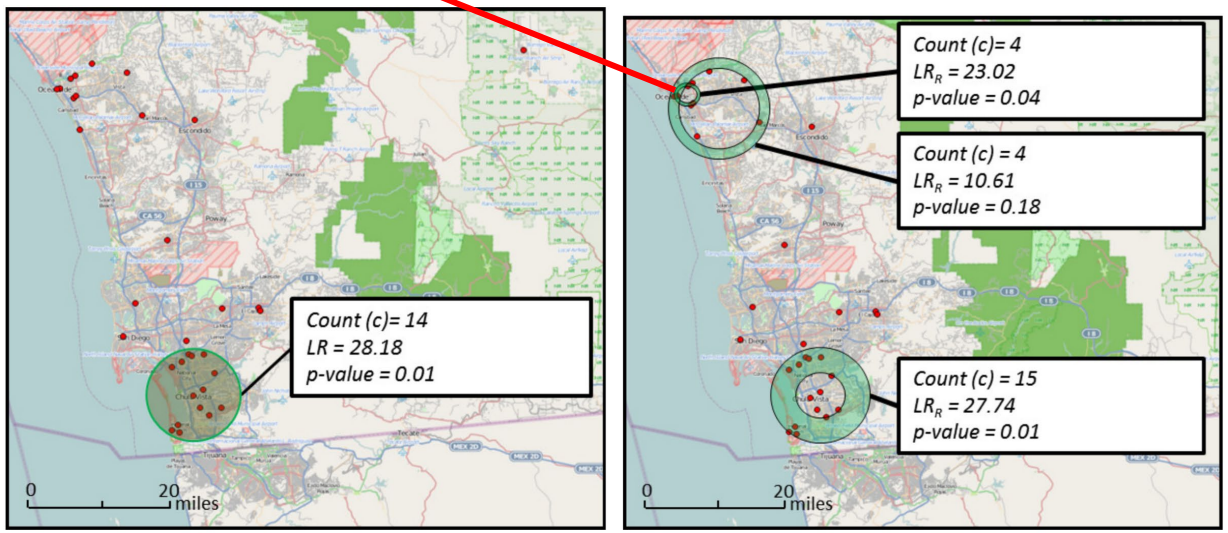
- Point processes for different problems
  - Real-valued: Life-expansion, Airbnb price
  - Categorical: Diversity
  - Space-time...
- Hotspots of different shapes
  - Ring (e.g., serial criminals), irregular, linear (e.g., on a road network)
- Data
  - Social network, trajectories, polygon, raster...

Shape Extensions

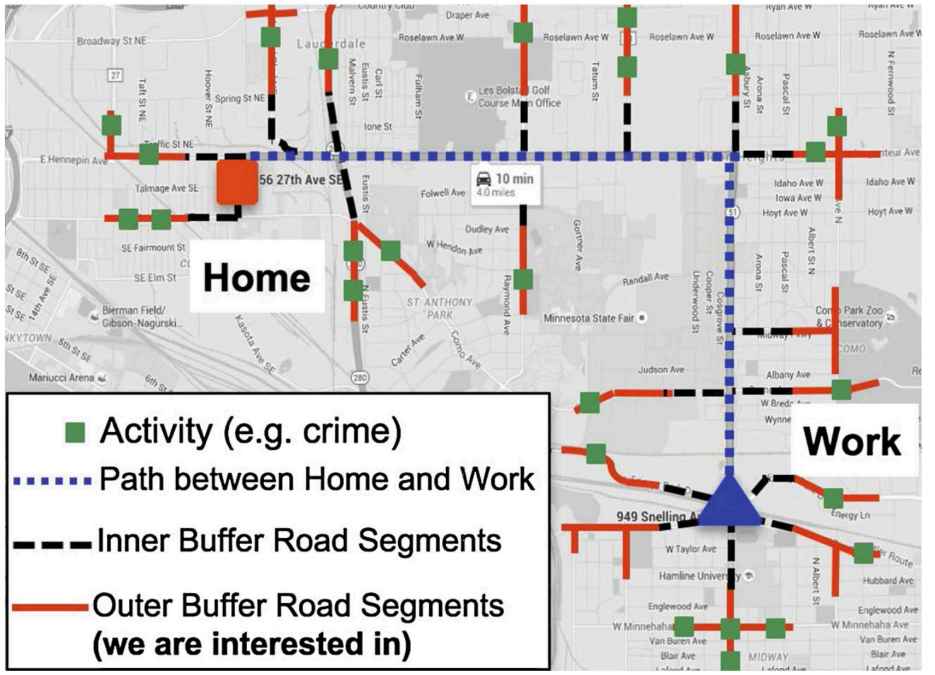
- Shapes: ring, rectangle, linear…
- Spatial data models: road network…



An arsonist was caught



SaTScan Ring-shaped



Extensions

Which shape best describes this pattern?

N.Y. / REGION

Hotel That Enlivened the Bronx Is Now a 'Hot Spot' for Legionnaires'

By WINNIE HU and NOAH REMNICK AUG. 10, 2015

Contaminated Cooling Towers

Five buildings have been identified as the potential source of the Legionnaires' disease outbreak in the South Bronx.

- Possible sources of Legionnaires' outbreak
- Additional sites found with legionella bacteria
- Locations of people with Legionnaires'

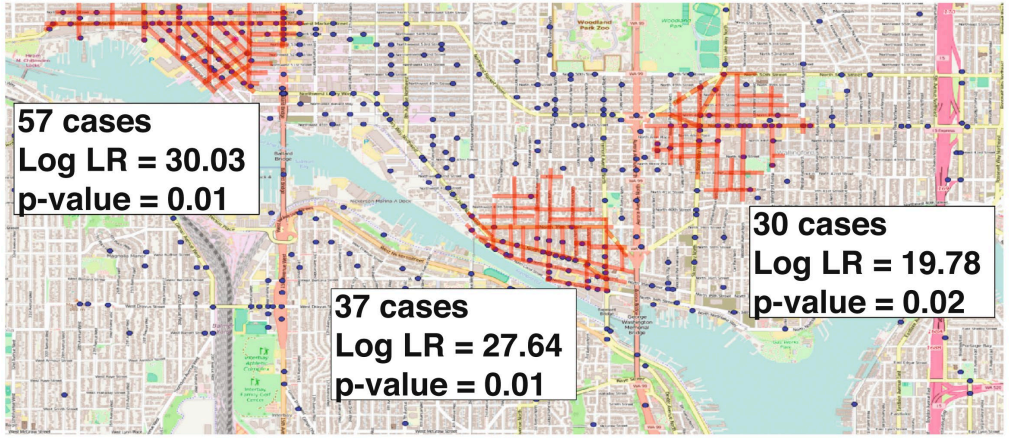
Source: New York Mayor's Office

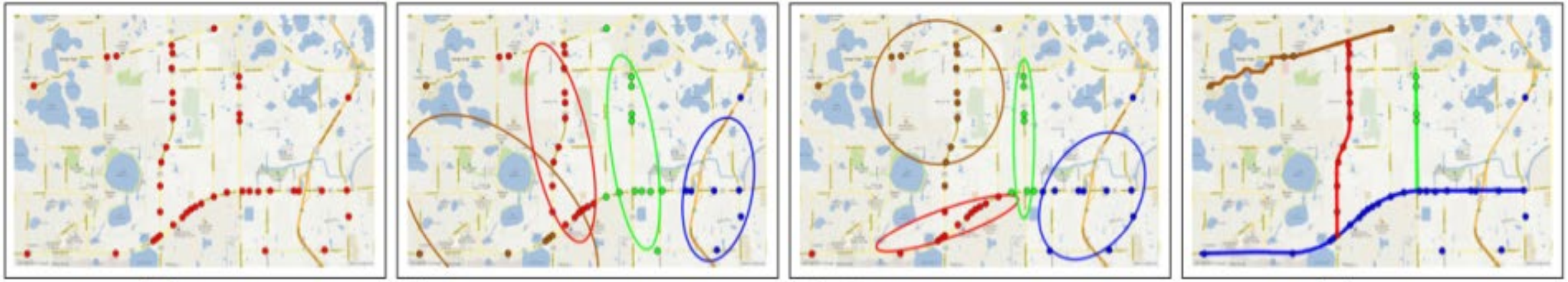
By The New York Times

The Opera House Hotel is at the center of the outbreak. Edwin J. Torres for The New York Times

Shape Extensions

- Shapes: ring, rectangle, linear…
- Spatial data models: road network…

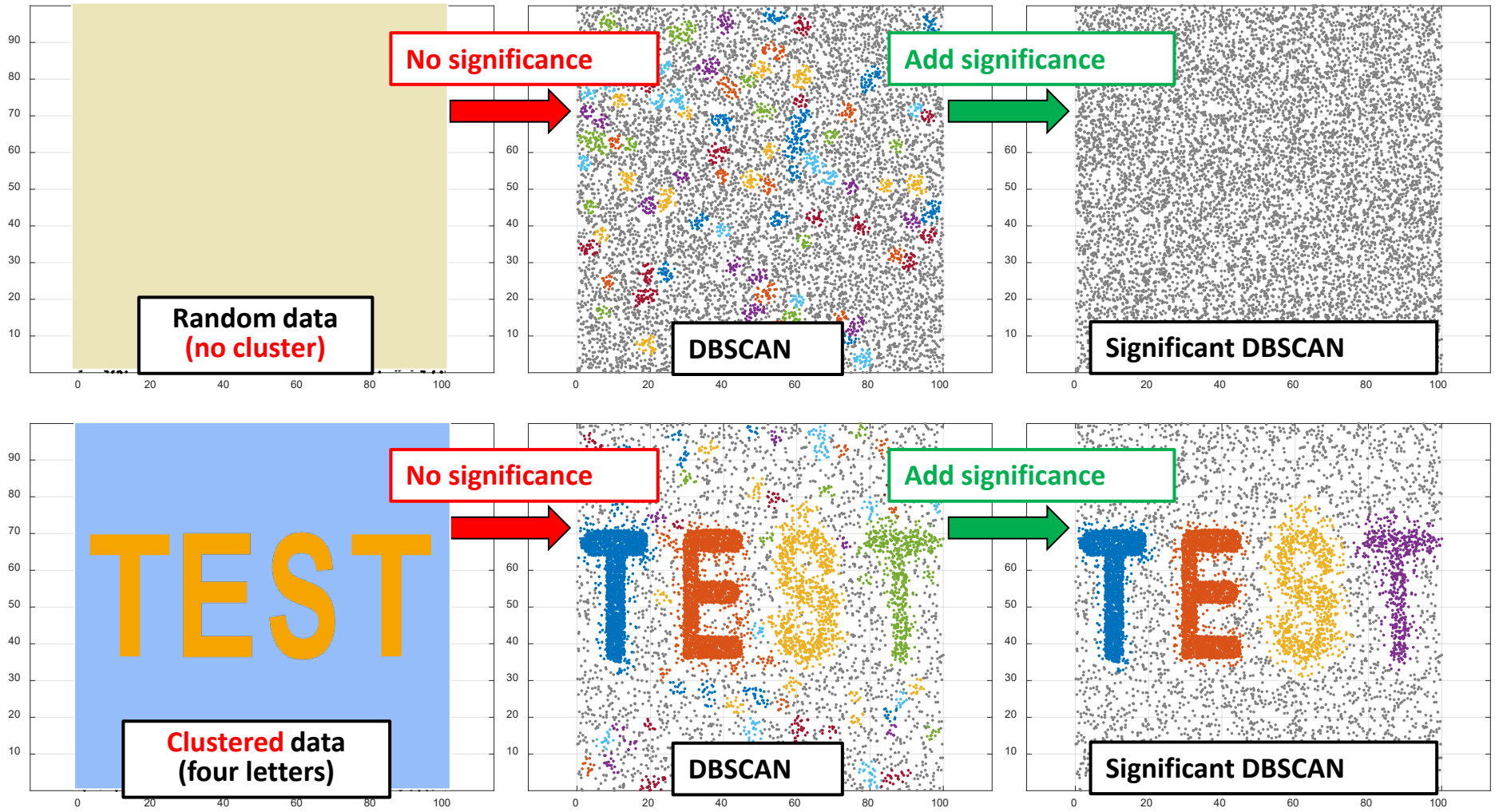




(a) Input (b) Crimestat K-means with Euclidean Distance (c) Crimestat K-means with Network Distance (d) KMR

Source: A K-Main Routes Approach to Spatial Network Activity Summarizationin IEEE Transactions on Knowledge and Data Eng. (www.computer.org/csdl/trans/tk/preprint/06574853-abs.html)

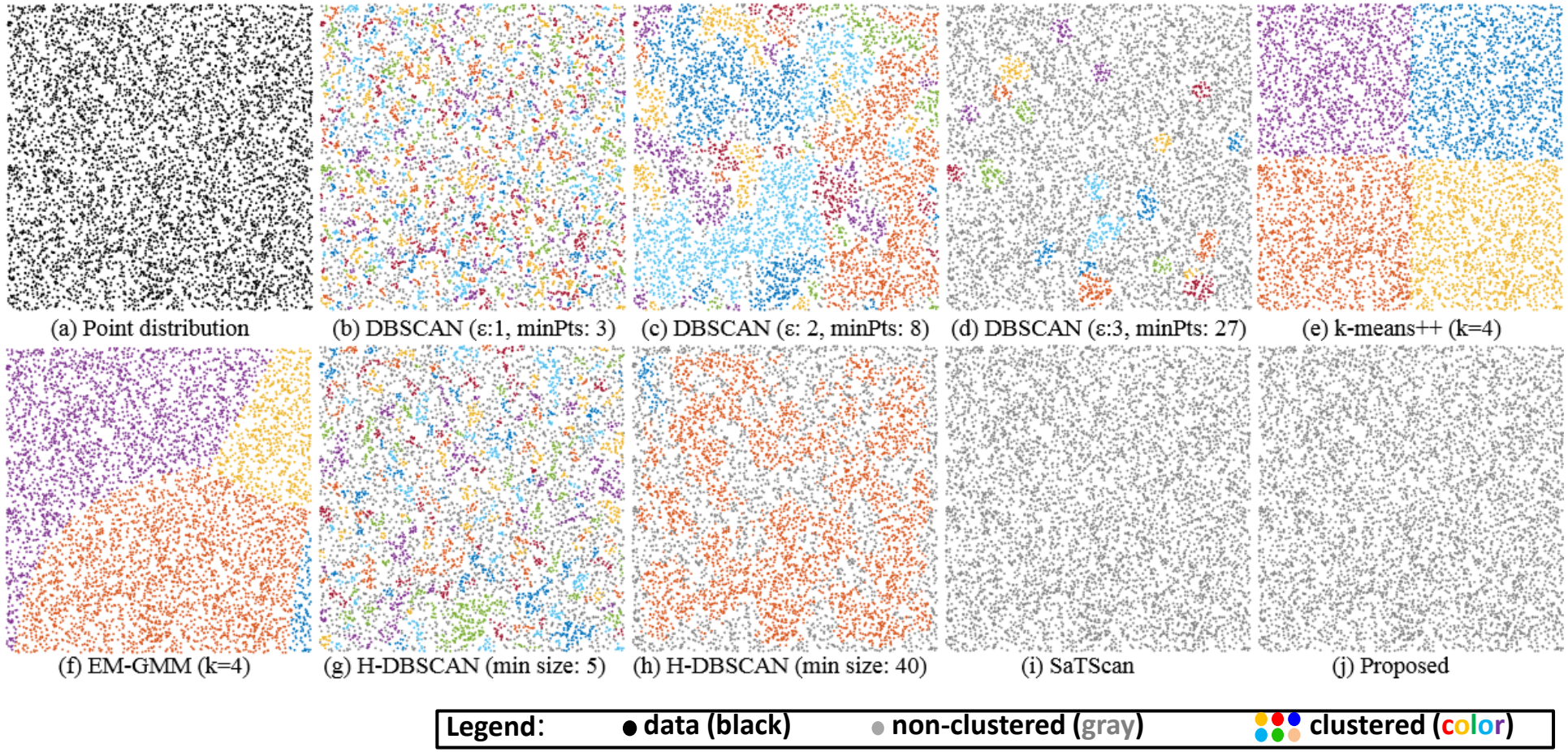
Irregular Shapes



Xie, Y. and Shekhar, S., 2019, August. Significant DBSCAN towards Statistically Robust Clustering. In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases* (pp. 31-40).

Result Comparison

Random noise (no cluster)

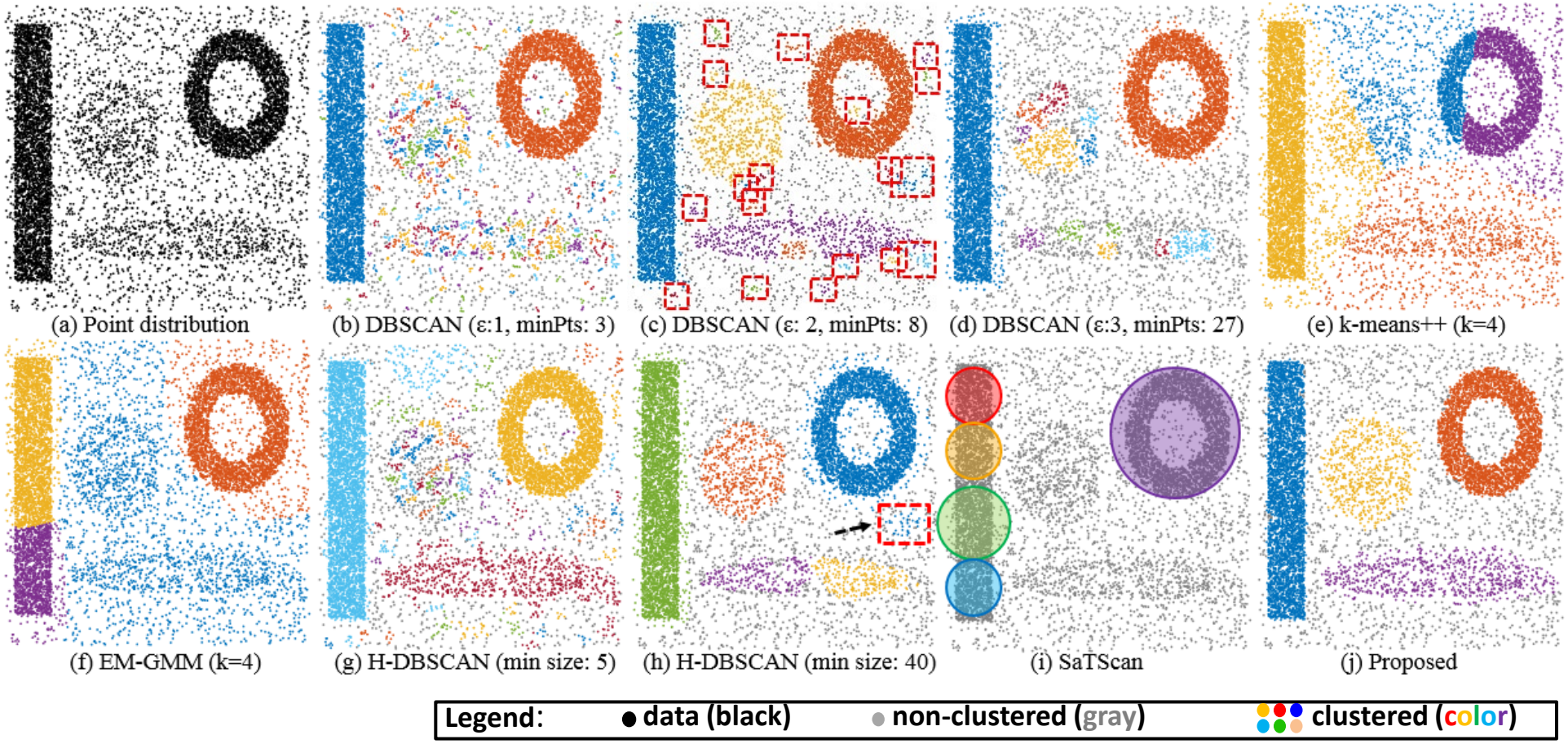


Significant DBSCAN

Xie, Y. and Shekhar, S., 2019, August. Significant DBSCAN towards Statistically Robust Clustering. In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases* (pp. 31-40).

Result Comparison

Clusters + random noise



Significant DBSCAN

Xie, Y. and Shekhar, S., 2019, August. Significant DBSCAN towards Statistically Robust Clustering. In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases* (pp. 31-40).

Extensions (and Reinforcements)

- Point processes for different problems
  - Real-valued: Life-expansion, Airbnb price
  - Categorical: Diversity
  - Space-time...
- Hotspots of different shapes
  - Ring (e.g., serial criminals), irregular, linear (e.g., on a road network)
- Data
  - Social network, trajectories, polygon, raster...

Data Extensions

- Social network data
- Trajectories
- Polygons and rasters

Social Network

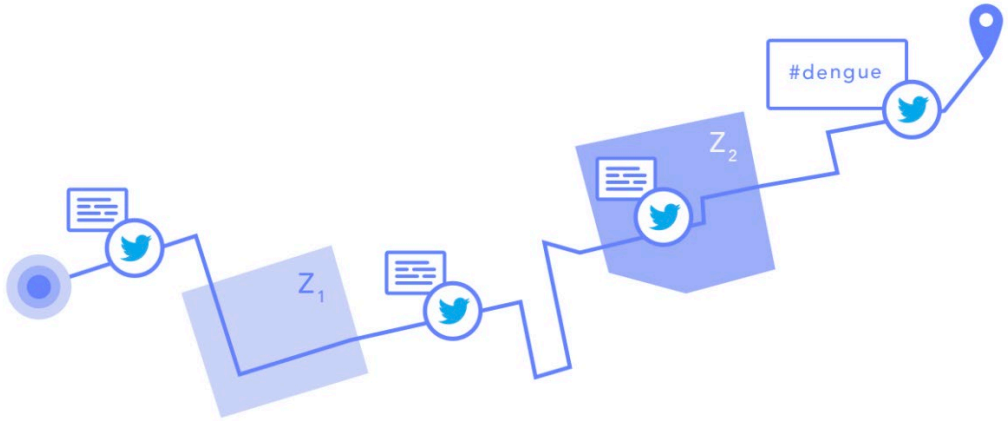
- Why?
  - Example: Disease data may be delayed and not real-time
    - Spread of disease can be very fast so early-warnings are important
    - Social network (with geotags) may compliment traditional approach with more real-time data
  - Other events
    - Flooding and natural disasters, crimes and terrorisms, …

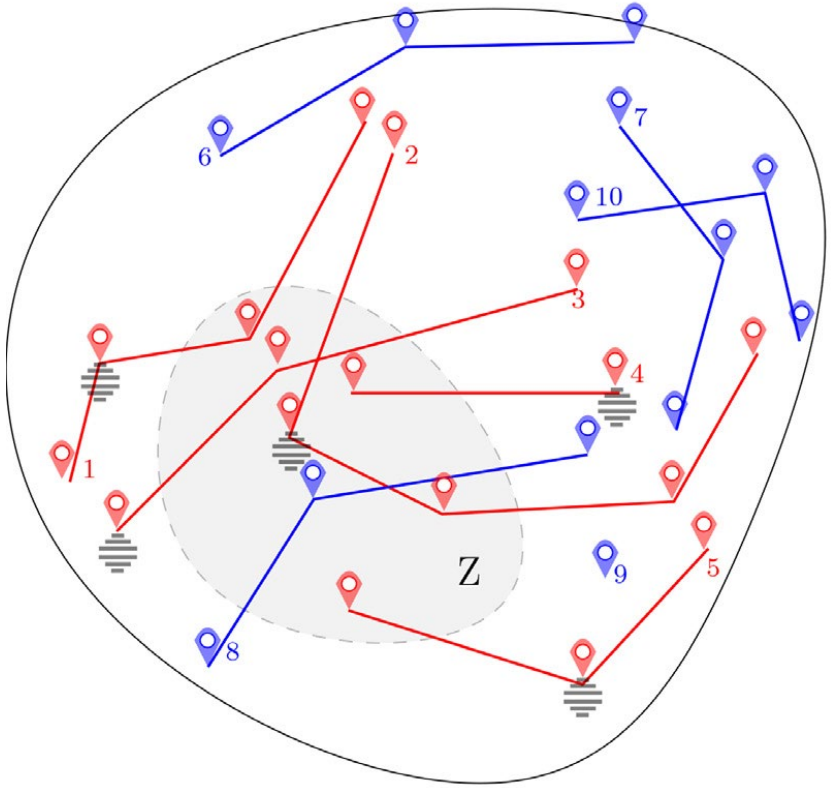
Social Network

- Why?
  - Example: Disease data may be delayed and not real-time
    - Spread of disease can be very fast so early-warnings are important
    - Social network (with geotags) may compliment traditional approach with more real-time data
  - Other events
    - Flooding and natural disasters, crimes and terrorisms, …

Social Network

- Twitter data to send early warnings of Dengue fever outbreaks



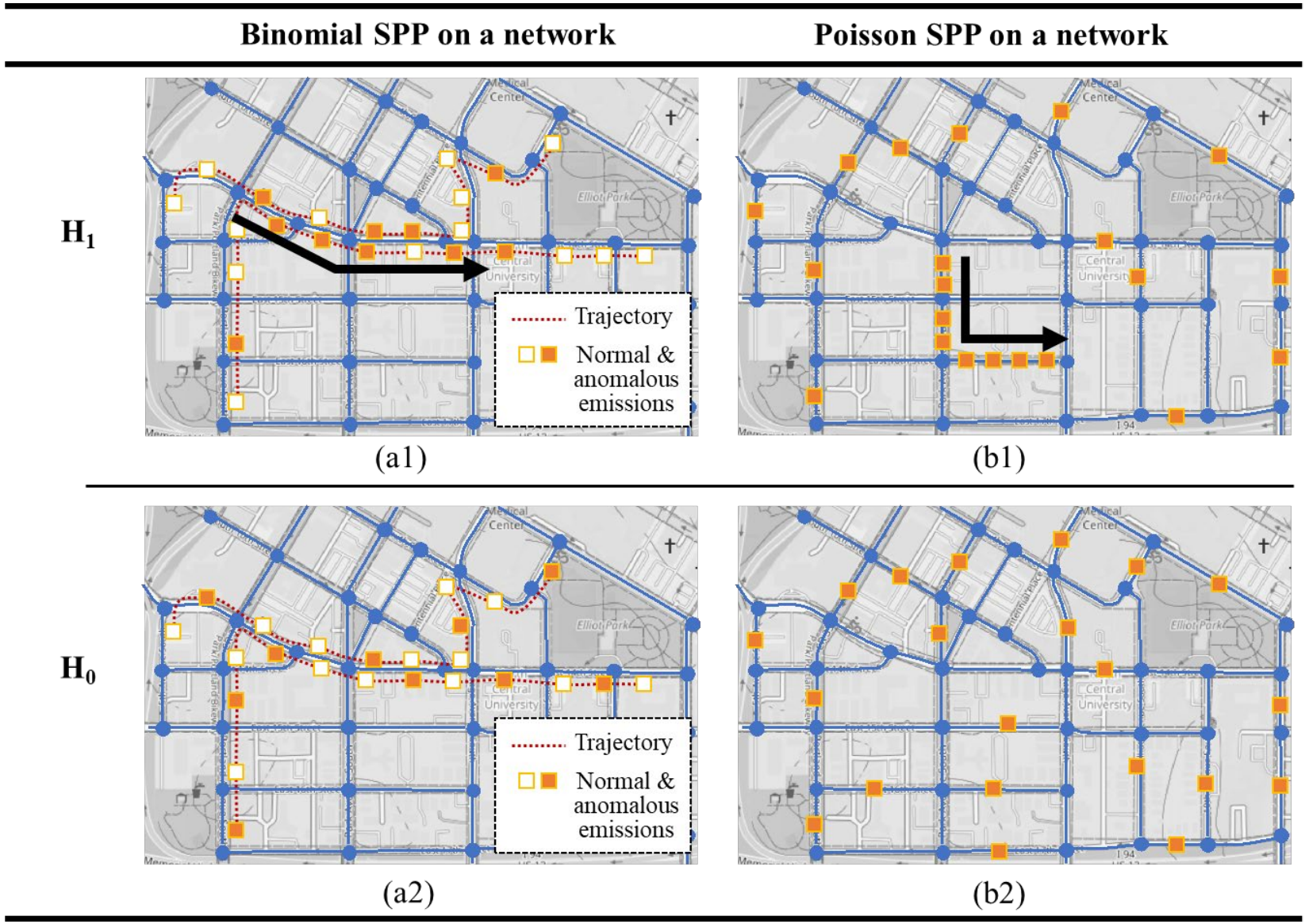


Trajectories

- Regional shapes or linear shapes

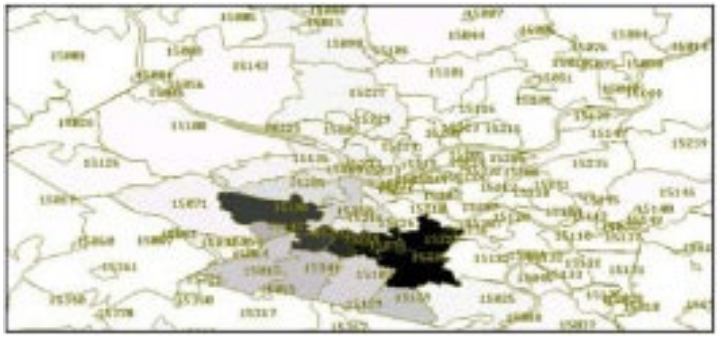
Trajectories

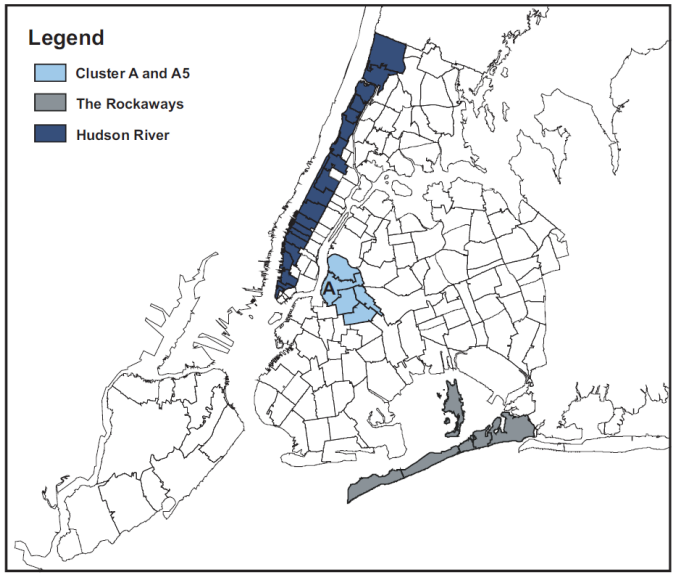
- High energy consumptions
- High emissions

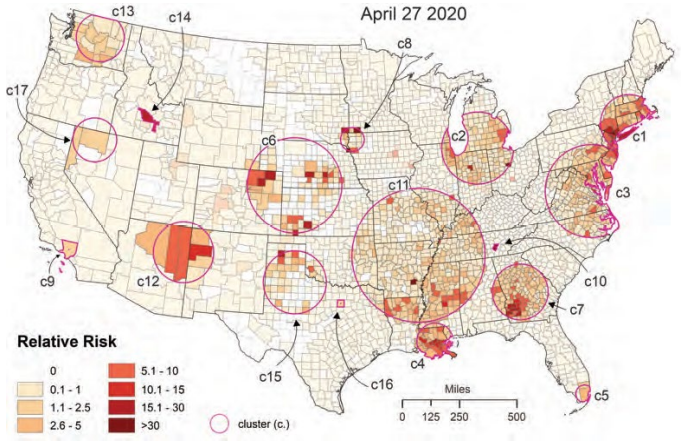


Other Data Types

- Polygons, rasters, time-series
  - Polygon: use centroids as points
  - Raster: use cells as points
  - Time-series
    - Add time-periods into enumeration
    - Expectation-based scan statistics
    - Bayesian scan statistics
    - …







SaTScan (Demo in the Week after Midterm)

- A widely-used software by M. Kulldorff (Harvard)
  - Detailed user guide (we will go through together)
- Popularizes scan statistics
- Includes implementations for a variety of point processes
- See lecture notes for step-by-step guidance