MapReduce Based Spatial Hotspots Detection Using Polygon Propagation

Jian Chen [1,2], Satya Katragadda [2], Shaaban Abbady [2]

1: Department of Geography, University of North Alabama

2: NSF Center for Visual & Decision Informatics (CVDI), University of Louisiana at Lafayette

2019 AAG Annual Meeting, Washington, D.C.

Symposium on Frontiers in Geospatial Data Sciences: Big Data Computing for Geospatial Applications

April 4th, 2019

VDI Center for Visual & Decision Informatics

UNIVERSITY OF NORTH ALABAMA 18 30

Outlines

- Background
  - NSF Center fro Visual & Decision Informatics (CVDI)
  - MapReduce
  - Hotspot Analysis
- Our Approach
  - Algorithms
  - Experiments
- Conclusion

Background

NSF Center for Visual & Decision Informatics

Center for Visual & Decision Informatics

Industry Advisory Board (IAB)

mbrain · Schumacher Clinical Partners · Nokia · Johnson & Johnson · GSK · Thomson Reuters · USGS science for a changing world · Capital One · intel · tieto · CGI · Cancer Treatment Centers of America

Clarivate Analytics | Formerly the IP & Science business of Thomson Reuters

Financial Support, Research Direction, Guidance

University Partners

- University of Louisiana Lafayette
- Drexel University
- Tampere University of Technology
- Stony Brook University
- University of Virginia Engineering
- UNC Charlotte
- BYU

Research Expertise & Institutional Support (Infrastructure, Students, etc.)

Collaborative Research Approach

Areas of Focus

Research Projects

- Visual Analytics
- Predictive Analytics
- Data Cleaning
- Interactive Visualization
- Data Summarization
- Social Media Analytics

Research Technology Portfolio

Core and Supplemental Funds, IUCRC Governance

Funding Agencies

- NSF
- Tekes

Value Created

Innovations in Big Data and Analytics

Cooperative Technology Transfer
- Royalty Free Licenses
- Publication Access
- Technology Breakthroughs

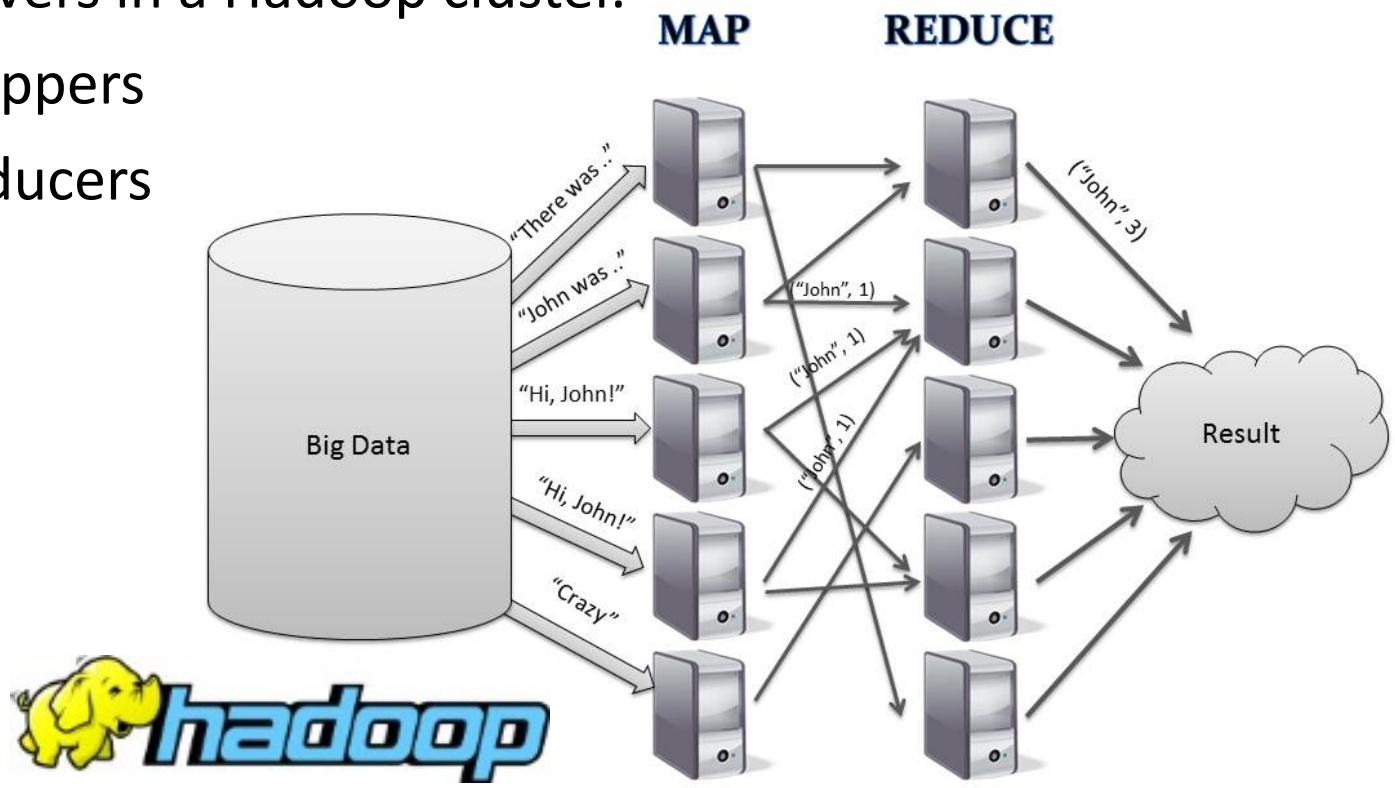
Collaborative Results
- University Alignments
- Faculty/Student Access
- Industry Partnerships

Investment in Future
- Trained Workforce
- Recruitment Opportunities
- Professional Development

Value Return
- Research Cost Savings
- 90% funds dedicated to research
- >40:1 Return on Investment

MapReduce

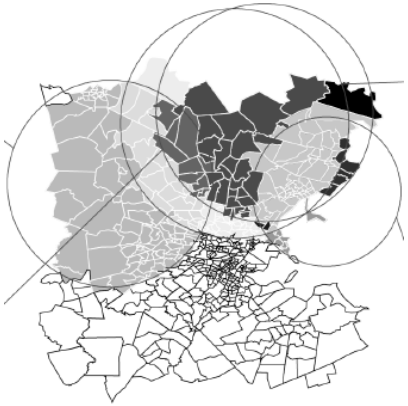
- A processing technique and a program model for distributed computing based on Java.
- Allows massive scalability cross hundreds or thousands of servers in a Hadoop cluster.
- Mappers
- Reducers



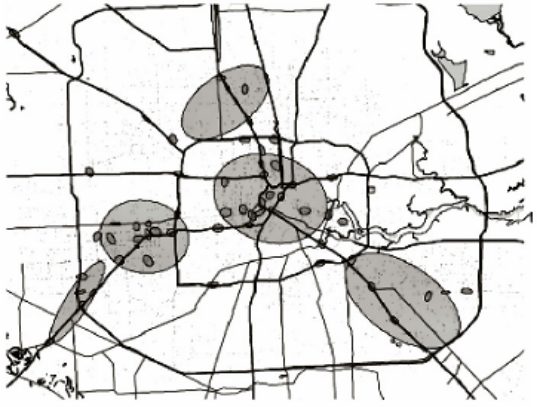
Why hotspots?

- **Spatial Hotspots**
  - Areas of space with unusually high incidence of events

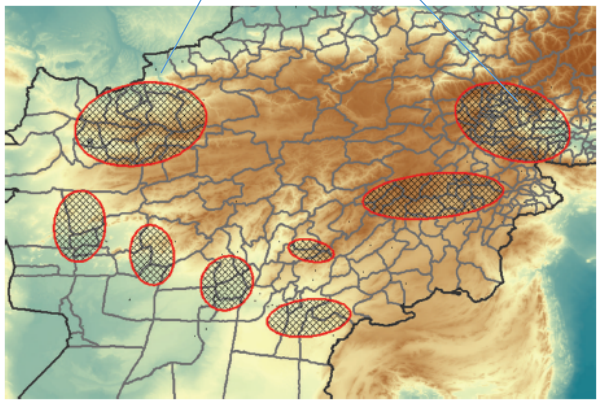
Epidemiology hotspots



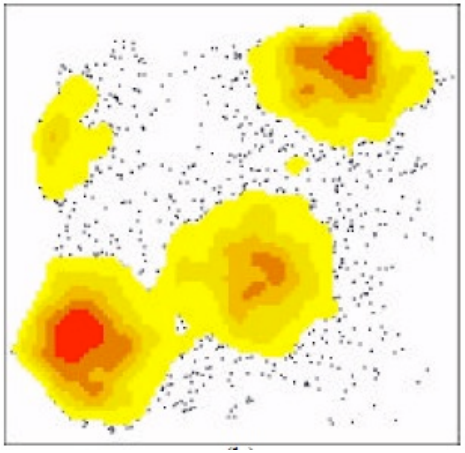
Crash hotspots



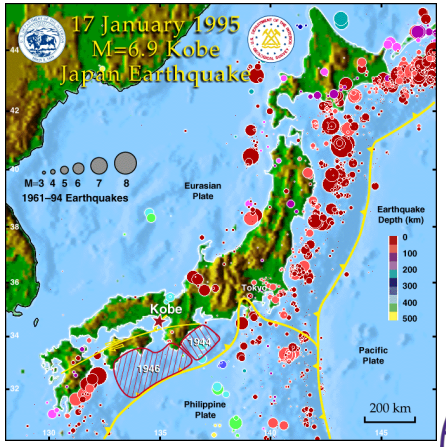
Terrorism hotspots



Crime hotspots

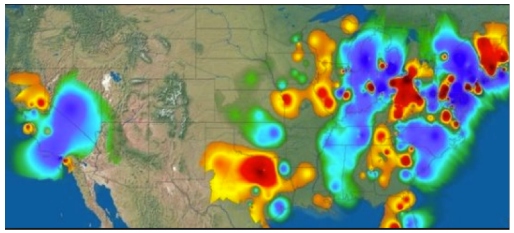


Seismic hotspots



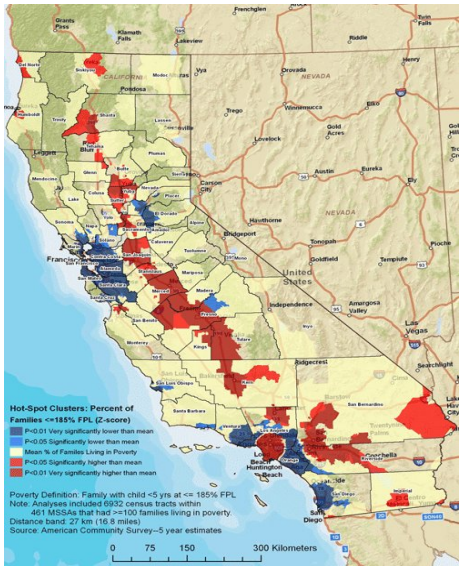
Hotspot Analysis

- Hotspot analysis – spatial clustering/spatial autocorrelation
- Research question - How to know where interesting events (hotspots) occur with overwhelming data compilation?
- Existing approaches
  - Moran's I (Moran 1948), Getis-Ord general G (Getis & Ord 1992)
  - Getis-Ord Gi* (Getis 2007), Anselin's LISA (Anselin 1995)
  - Scan Statistics (Kulldorff 1997, 2006)



State-of-the-art: Scan Statistics

- Identifies clusters of high activity using spatiotemporal data
- Good for discrete localized outbreaks
- Can be combined with others for multivariate analysis
- Requires pre-defined boundaries
- Resultant hotspots w/ artificial shape may include non-hotspots (high false positive errors)



Hotspot analysis in public health, source: ESRI

Terrorism activities hotspots in Philippines Mar – Jun 2002, sources (Gao, et.al, 2013)

VDI Center for Visual & Decision Informatics

UNIVERSITY OF NORTH ALABAMA 18 30

Our Approach

Our Hotspots Detection Algorithm

- Aim:
  - Unsupervised
  - Absence of geographical boundary
  - Identifies compact hotspots
- Approach:
  - Density-based
  - Polygon propagation
- Details: Katragadda et. al, 2018

Our Hotspots Detection Algorithm

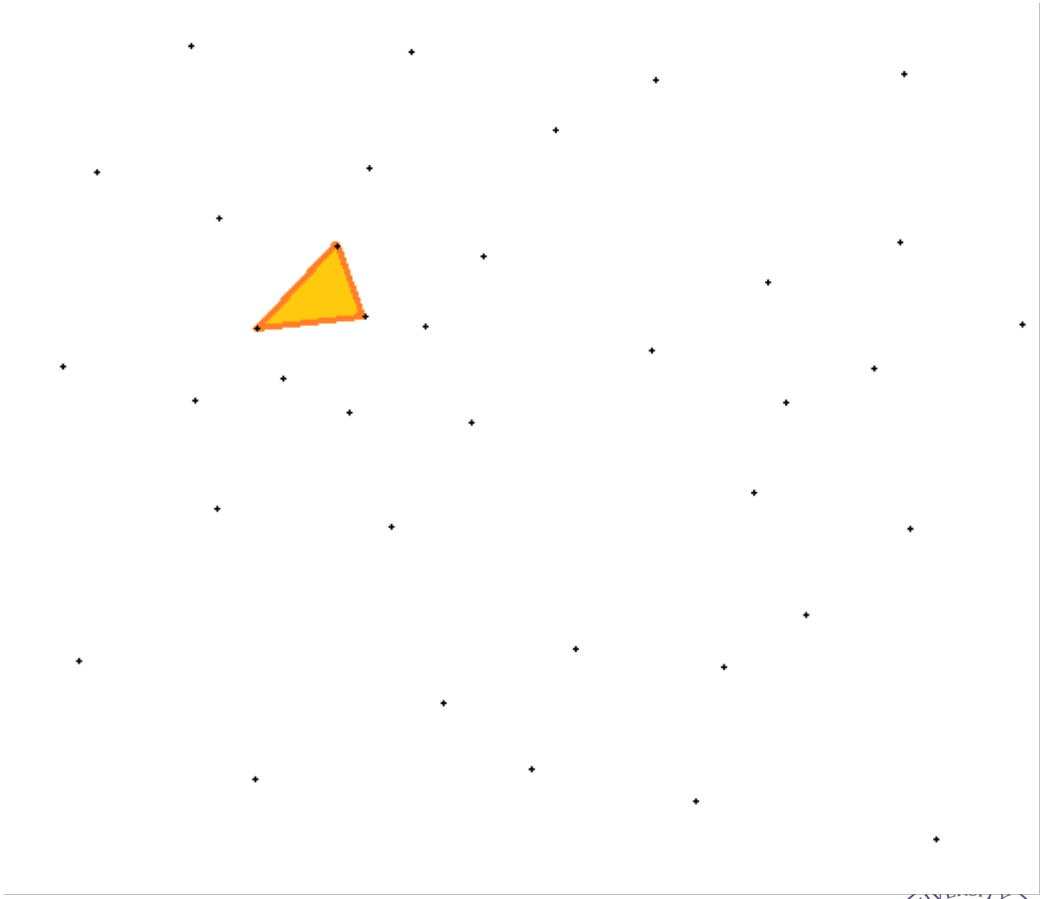
Given a set of points

- Identify a triangle: density > *t*
  - *Add nearest point to polygon:* density > *t*
  - *Until no more points can be added*
- Until no more triangles can be formed
- For each Polygon
  - *Calculate log-Likelihood value*
  - *If log-likelihood value > t2, it is categorized as hotspot*

Our Hotspots Detection Algorithm

Given a set of points

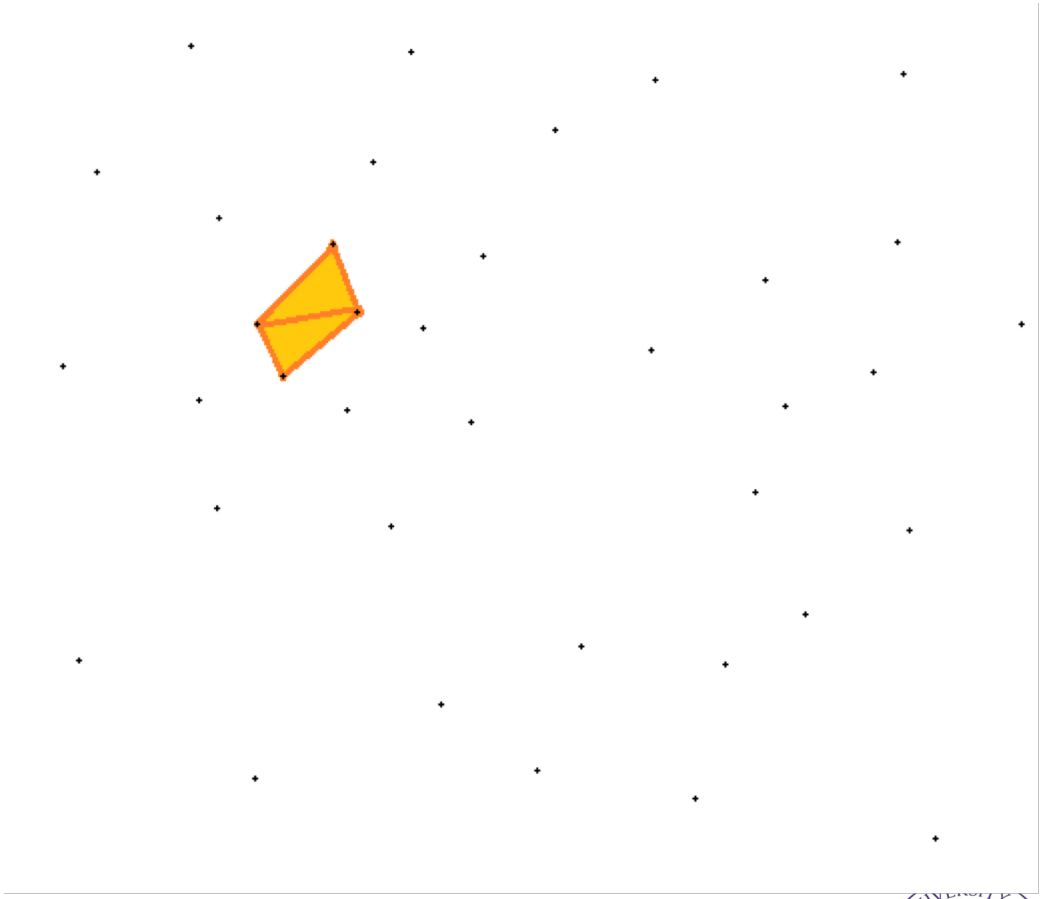
- Identify a triangle: density > *t*
  - *Add nearest point to polygon:* density > *t*
  - *Until no more points can be added*
- Until no more triangles can be formed
- For each Polygon
  - *Calculate log-Likelihood value*
  - *If log-likelihood value > t2, it is categorized as hotspot*



Our Hotspots Detection Algorithm

Given a set of points

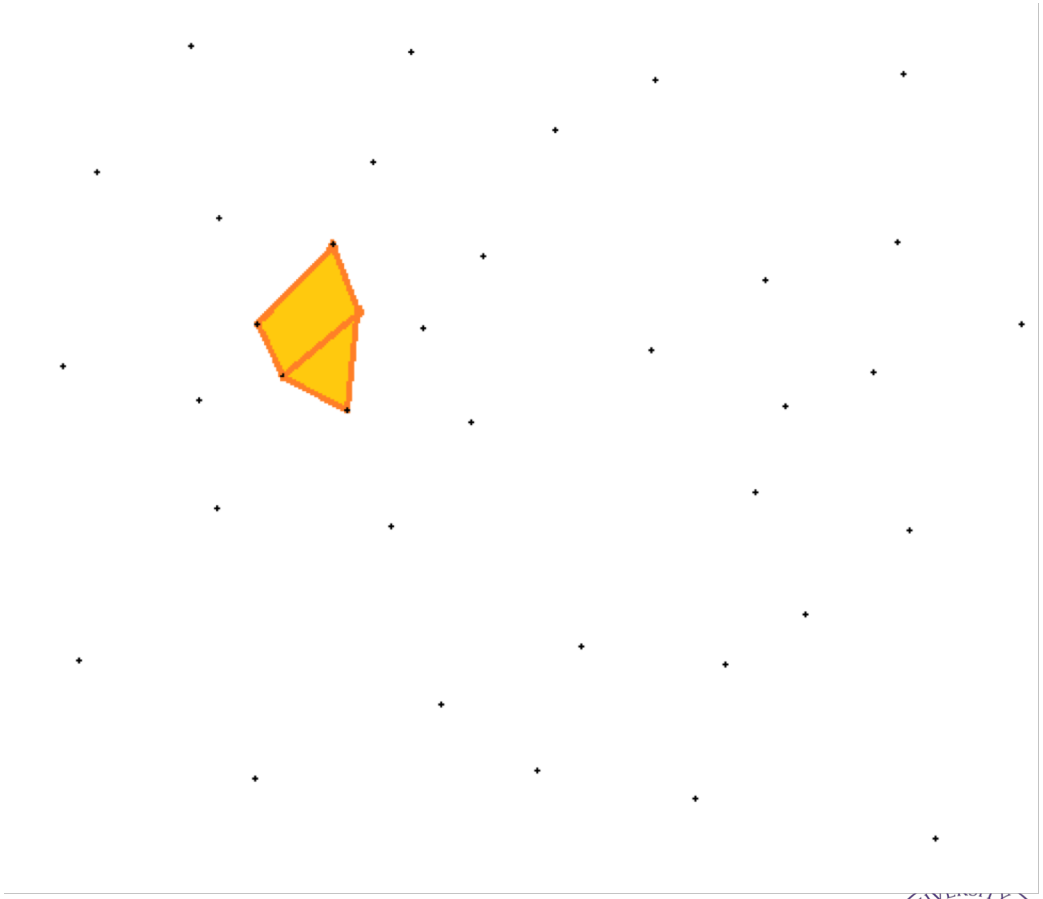
- Identify a triangle: density > *t*
  - *Add nearest point to polygon:* density > *t*
  - *Until no more points can be added*
- Until no more triangles can be formed
- For each Polygon
  - *Calculate log-Likelihood value*
  - *If log-likelihood value > t2, it is categorized as hotspot*



Our Hotspots Detection Algorithm

Given a set of points

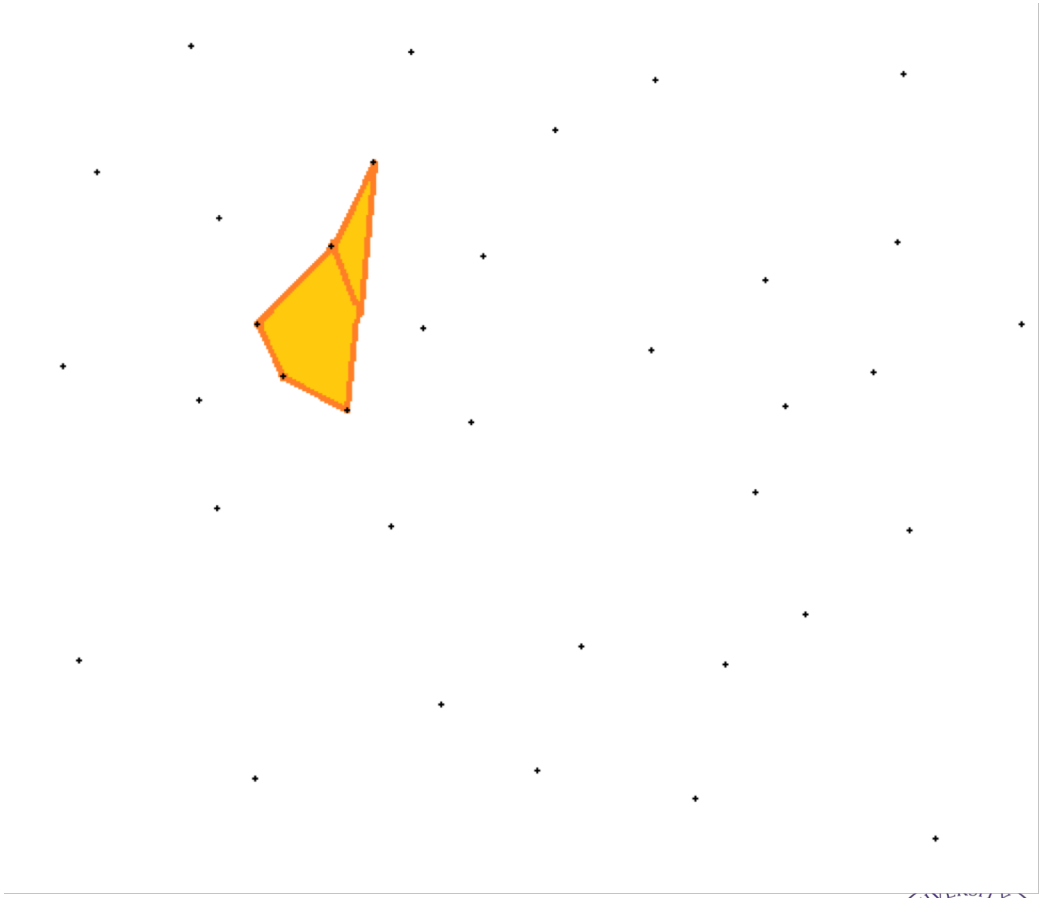
- Identify a triangle: density > *t*
  - *Add nearest point to polygon:* density > *t*
  - *Until no more points can be added*
- Until no more triangles can be formed
- For each Polygon
  - *Calculate log-Likelihood value*
  - *If log-likelihood value > t2, it is categorized as hotspot*



Our Hotspots Detection Algorithm

Given a set of points

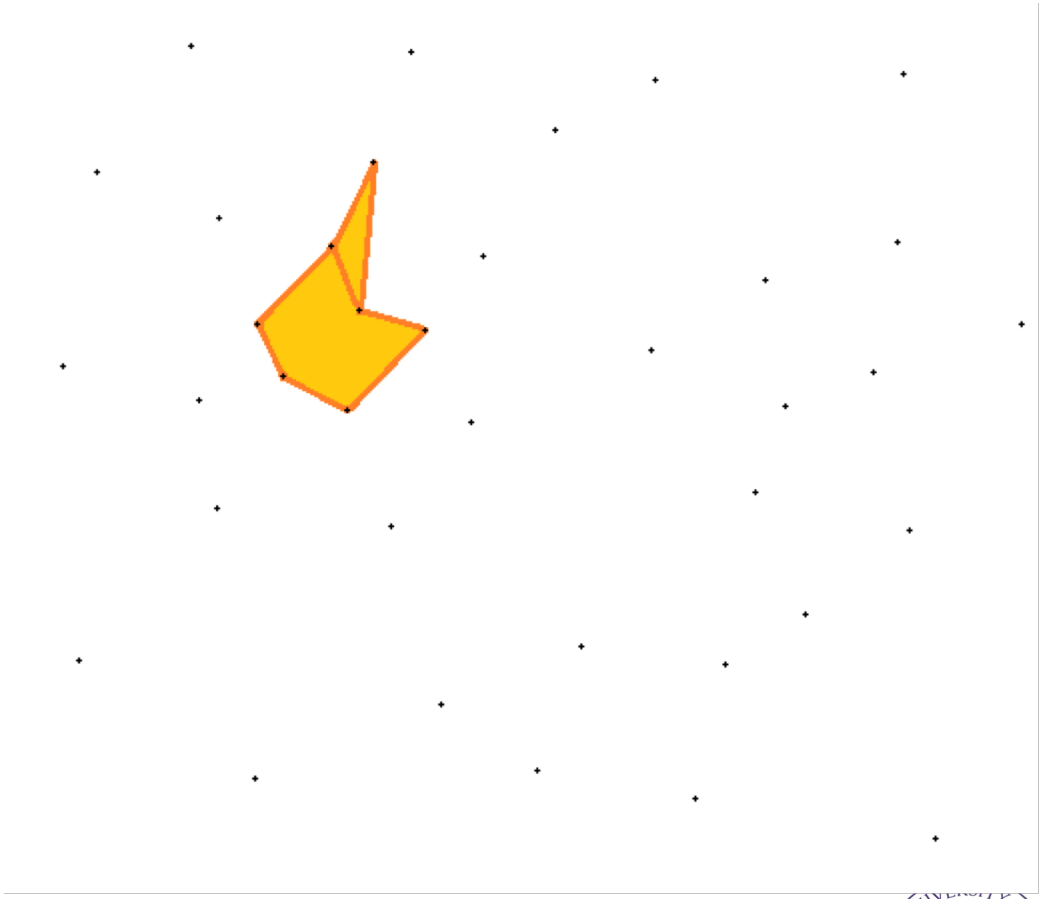
- Identify a triangle: density > *t*
  - *Add nearest point to polygon:* density > *t*
  - *Until no more points can be added*
- Until no more triangles can be formed
- For each Polygon
  - *Calculate log-Likelihood value*
  - *If log-likelihood value > t2, it is categorized as hotspot*



Our Hotspots Detection Algorithm

Given a set of points

- Identify a triangle: density > *t*
  - *Add nearest point to polygon:* density > *t*
  - *Until no more points can be added*
- Until no more triangles can be formed
- For each Polygon
  - *Calculate log-Likelihood value*
  - *If log-likelihood value > t2, it is categorized as hotspot*



Our Hotspots Detection Algorithm

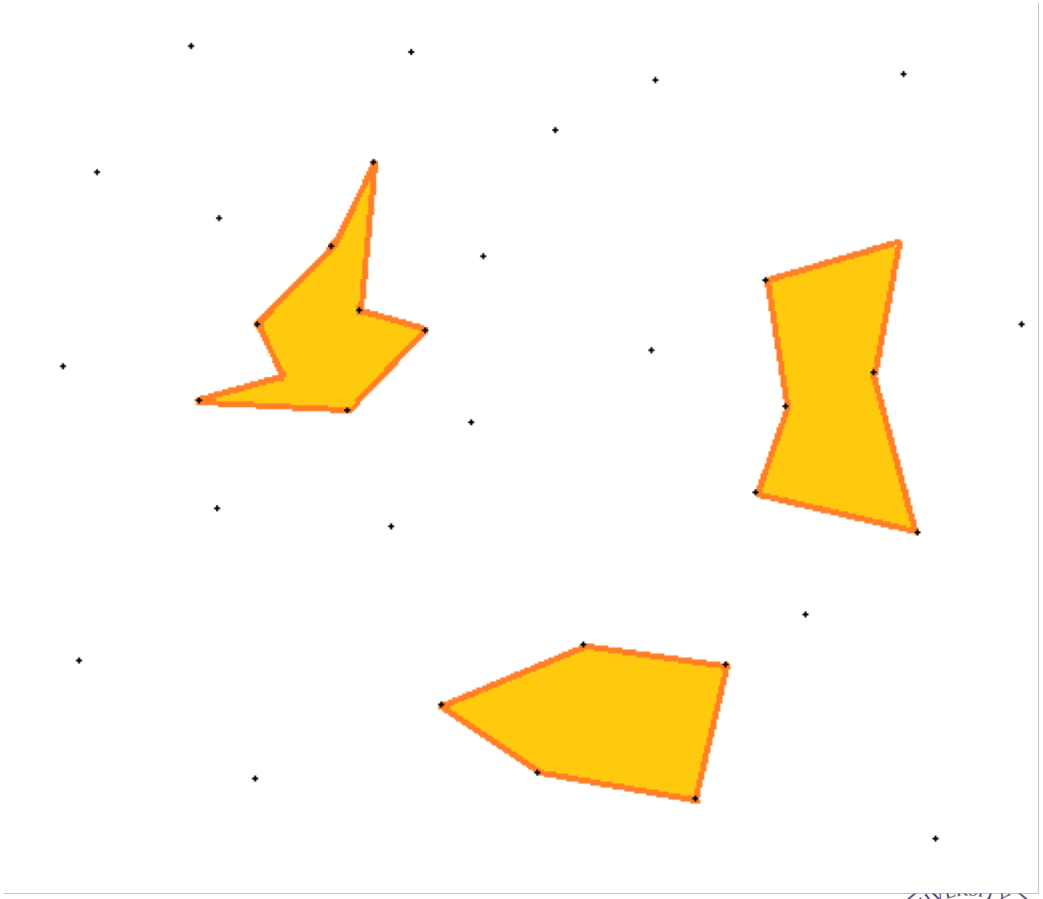
Given a set of points

- Identify a triangle: density > *t*
  - *Add nearest point to polygon:* density > *t*
  - *Until no more points can be added*
- Until no more triangles can be formed
- For each Polygon
  - *Calculate log-Likelihood value*
  - *If log-likelihood value > t2, it is categorized as hotspot*

Our Hotspots Detection Algorithm

Given a set of points

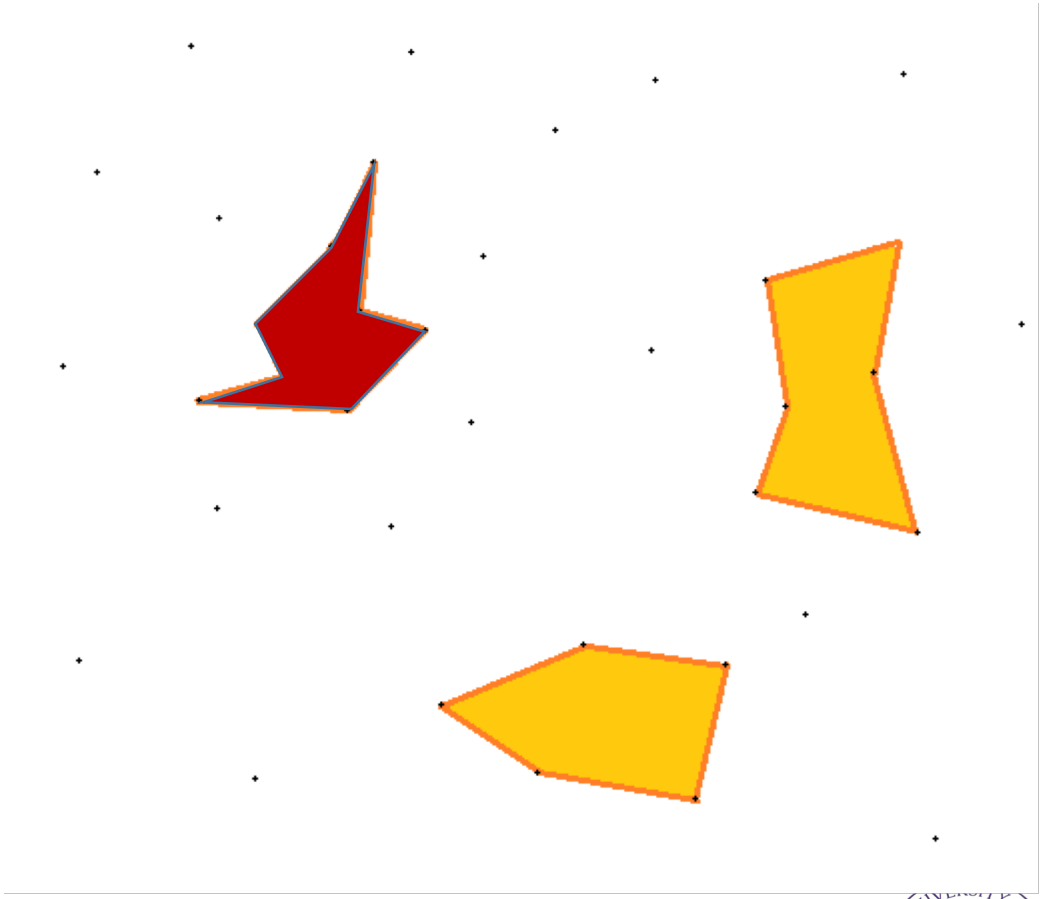
- Identify a triangle: density > *t*
  - *Add nearest point to polygon:* density > *t*
  - *Until no more points can be added*
- Until no more triangles can be formed
- For each Polygon
  - *Calculate log-Likelihood value*
  - *If log-likelihood value > t2, it is categorized as hotspot*



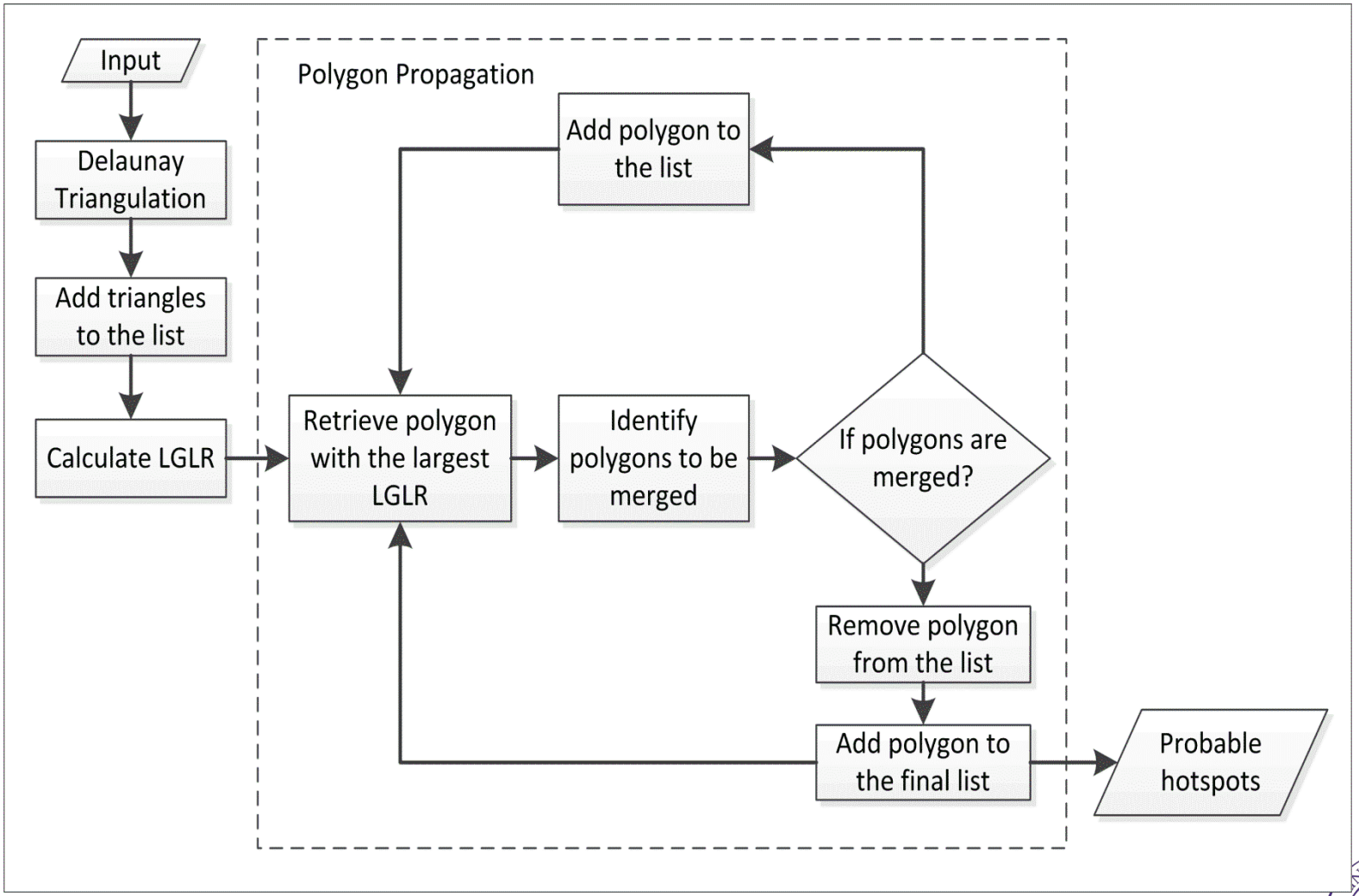
Our Hotspots Detection Algorithm

Given a set of points

- Identify a triangle: density > *t*
  - *Add nearest point to polygon:* density > *t*
  - *Until no more points can be added*
- Until no more triangles can be formed
- For each Polygon
  - *Calculate log-Likelihood value*
  - *If log-likelihood value > t2, it is categorized as hotspot*



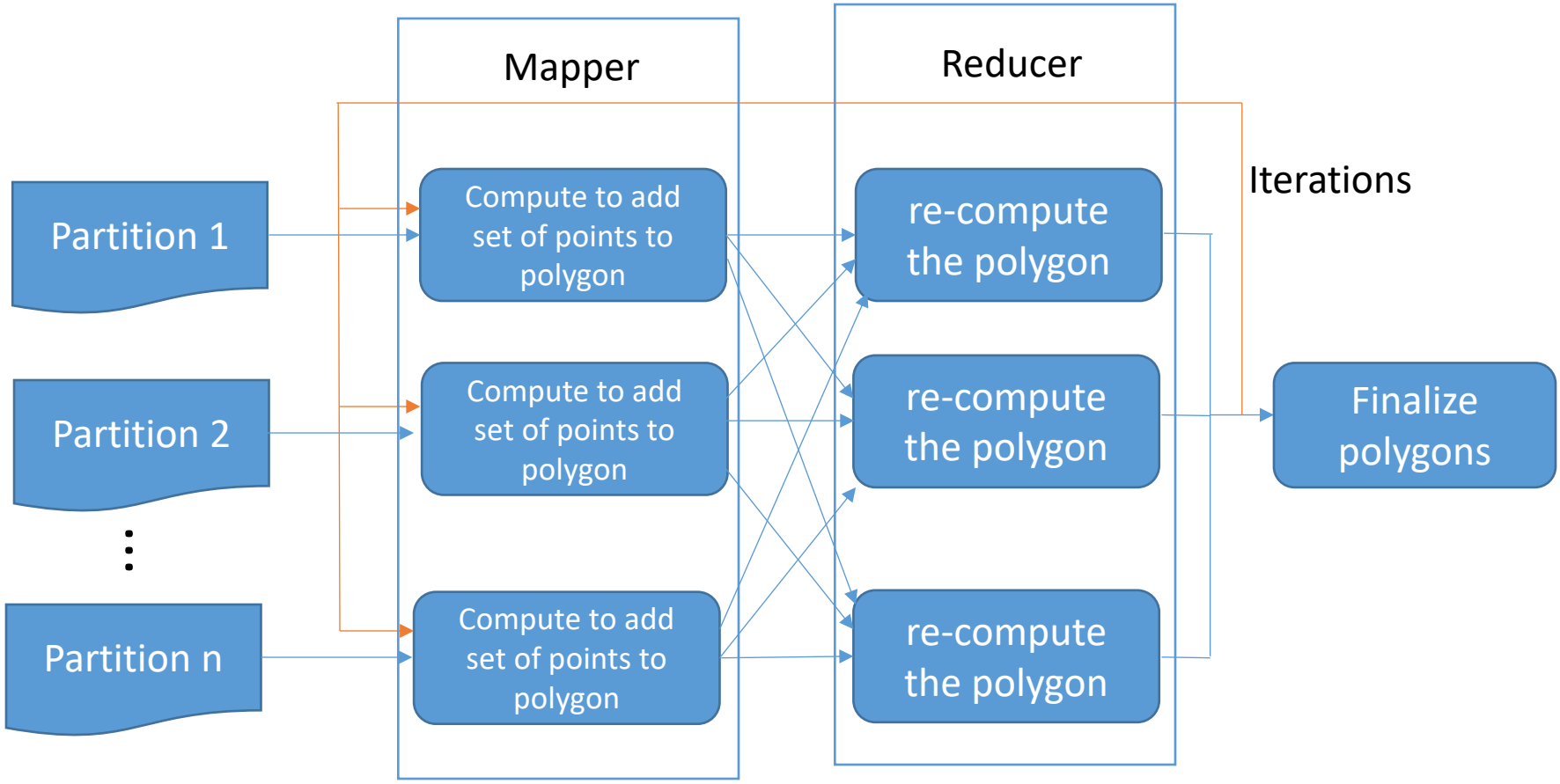
Our Hotspots Detection Algorithm



Challenge of Polygon Propagation based Methods

- Creating a patch of areas covering the whole map
  - ✓ Using scan statistics to identify most likely cluster
- Polygons may spread to a large area W/O constraints
  - ✓ A compactness parameter introduced to limit the size of the cluster
  - ✓ $\alpha = \frac{Area(P)}{Area(E)}$ P – the polygon, E- the smallest enclosing eclipse compasses all the points in polygon P
  - ✓ Prioritize polygon propagate within the encompassing eclipse
- Polygon propagation is expensive, complexity – $O(N^3 logN)$
  - ✓ All possible polygons are pre-computed by Delaunay Triangulation ($O(NlogN)$)
  - ✓ Distributed polygon propagation

Our Hotspots Detection Algorithm – MapReduce



MapReduce Version of Algorithm

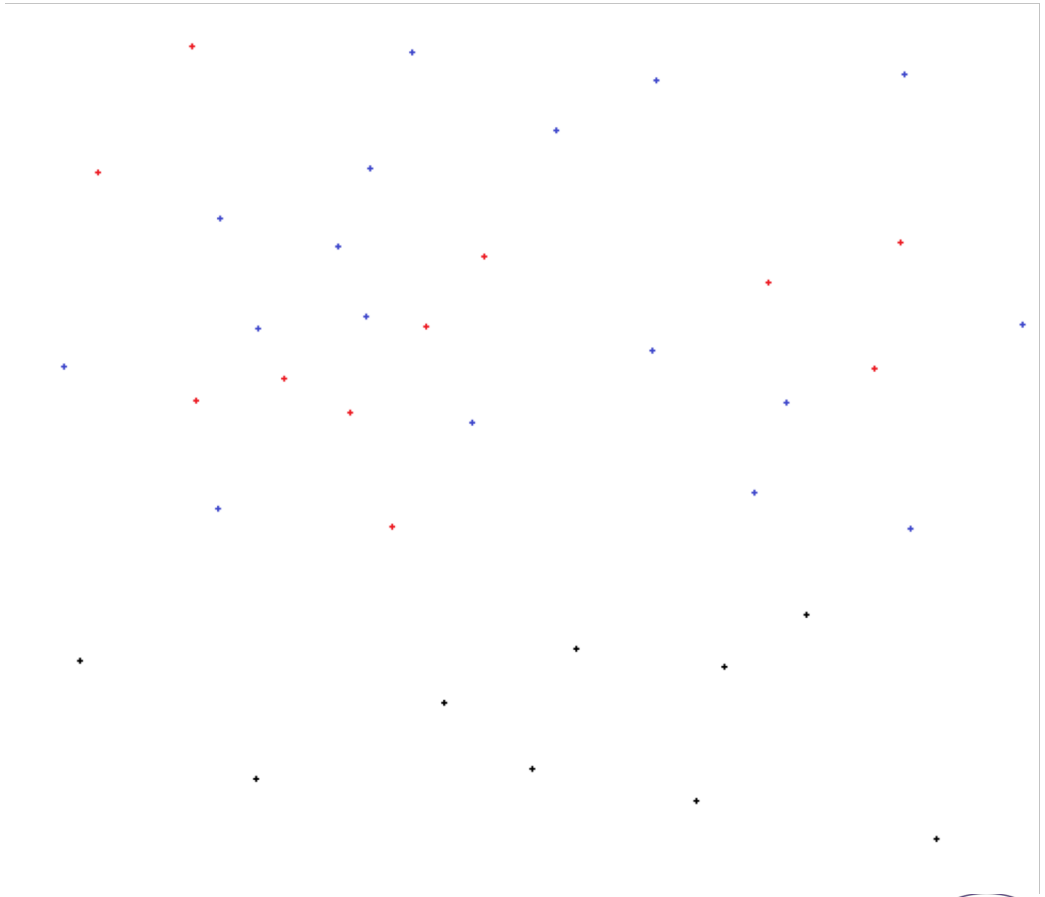
Given a set of input points

Mapper
- Create/extend polygons:
  - density of polygon>t1

Reducer
- re-compute the polygons

Repeat until stabilization

For each Polygon
- *Calculate log-Likelihood value*
- *If log-likelihood value > t2, it is a hotspot*



MapReduce Version of Algorithm

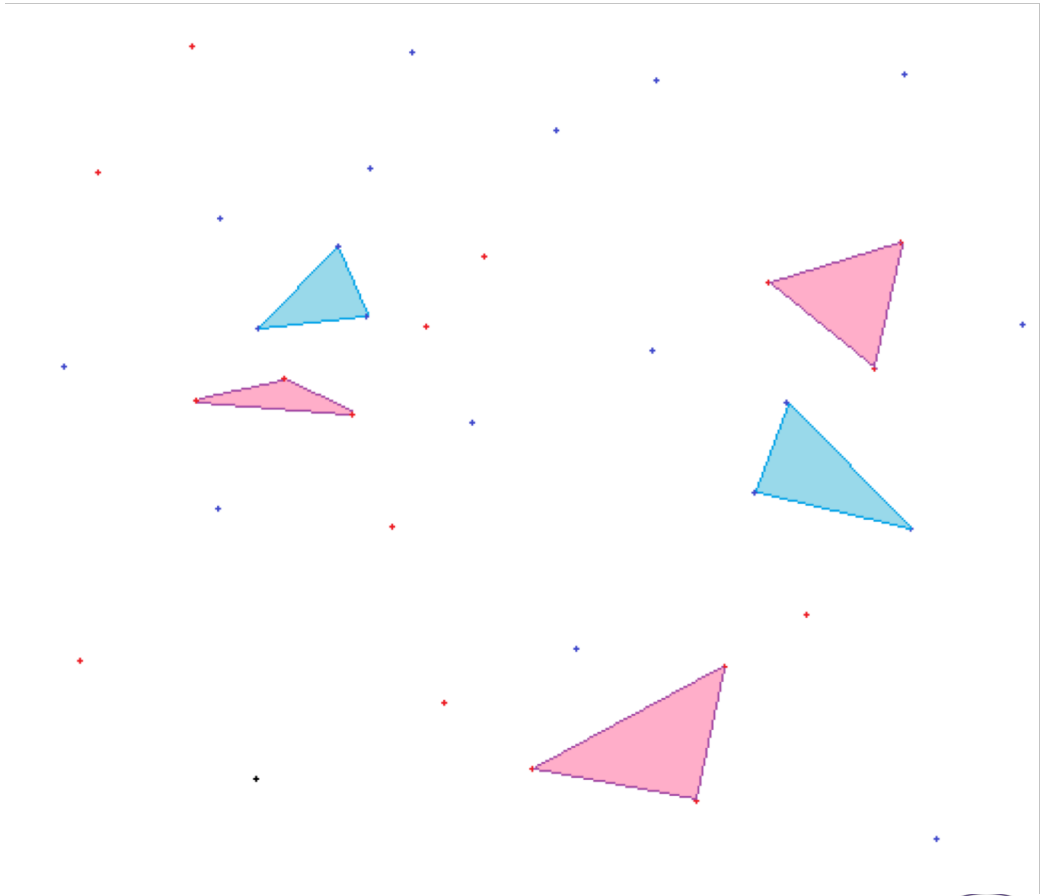
Given a set of input points

Mapper
- Create/extend polygons:
  - density of polygon>t1

Reducer
- re-compute the polygons

Repeat until stabilization

For each Polygon
- *Calculate log-Likelihood value*
- *If log-likelihood value > t2, it is a hotspot*



MapReduce Version of Algorithm

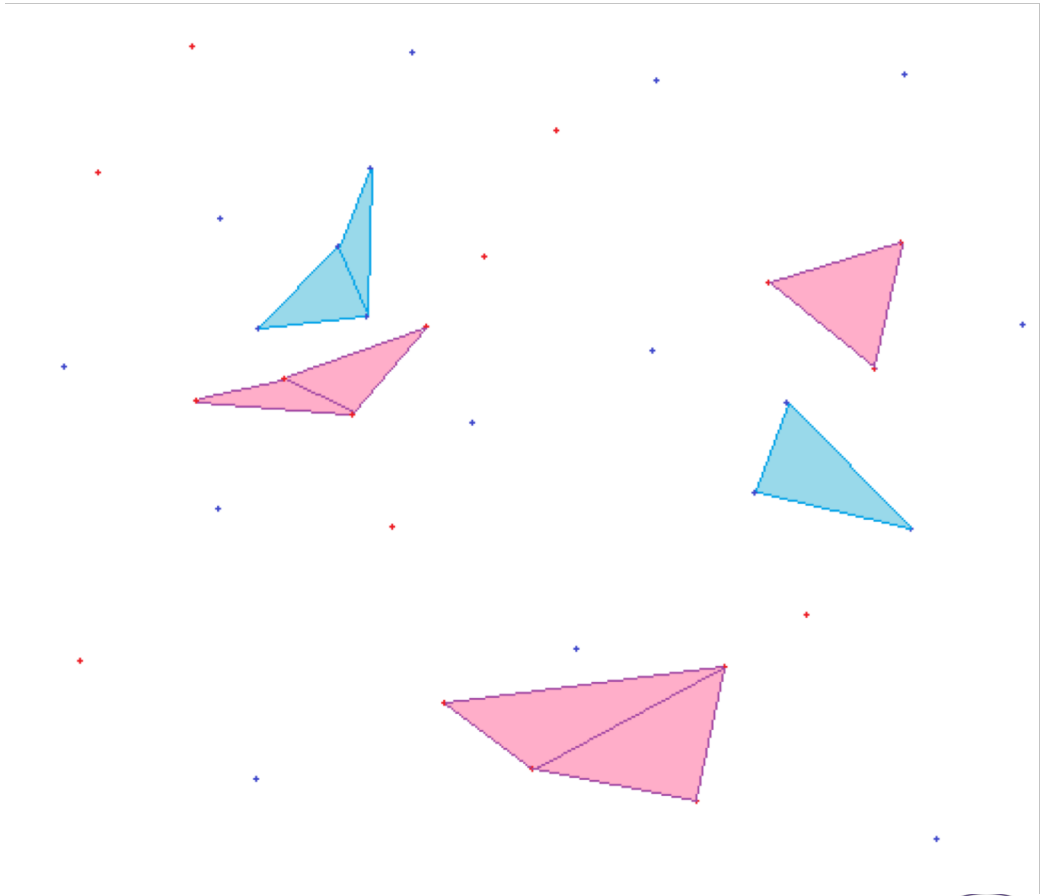
Given a set of input points

Mapper
- Create/extend polygons:
  - density of polygon>t1

Reducer
- re-compute the polygons

Repeat until stabilization

For each Polygon
- *Calculate log-Likelihood value*
- *If log-likelihood value > t2, it is a hotspot*



MapReduce Version of Algorithm

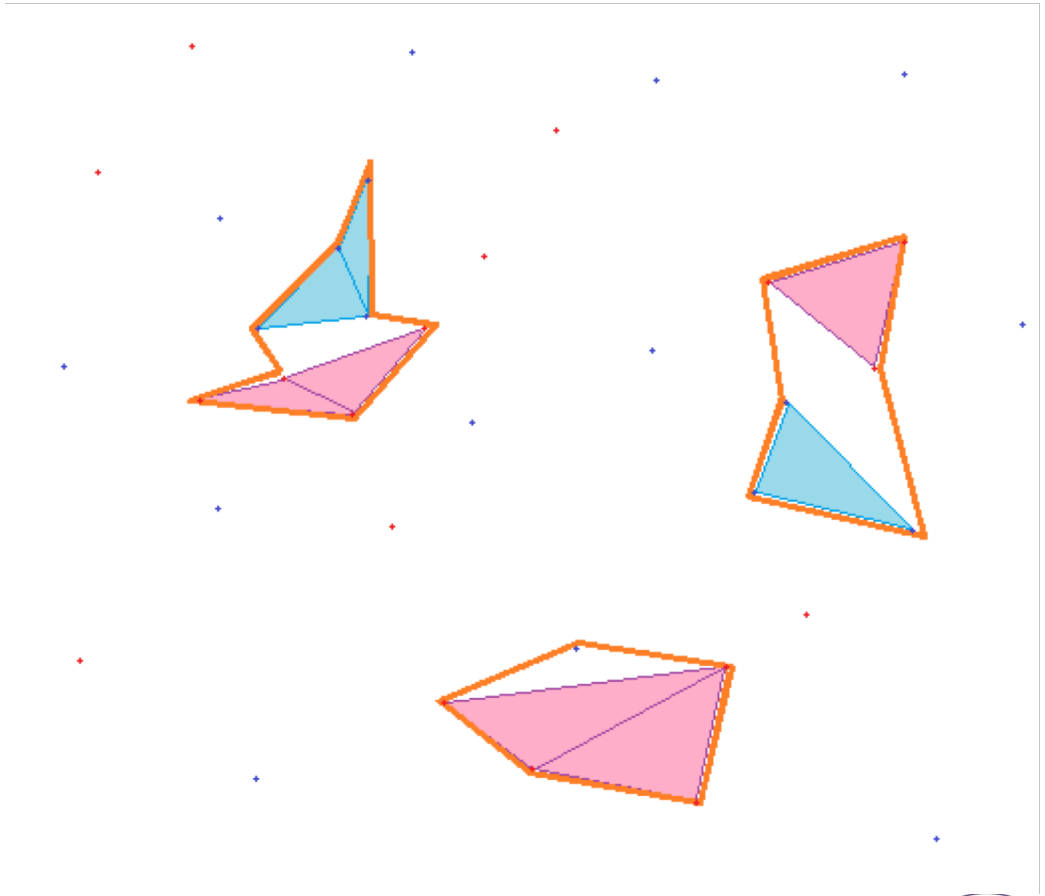
Given a set of input points

Mapper
- Create/extend polygons:
  - density of polygon>t1

Reducer
- re-compute the polygons

Repeat until stabilization

For each Polygon
- *Calculate log-Likelihood value*
- *If log-likelihood value > t2, it is a hotspot*



MapReduce Version of Algorithm

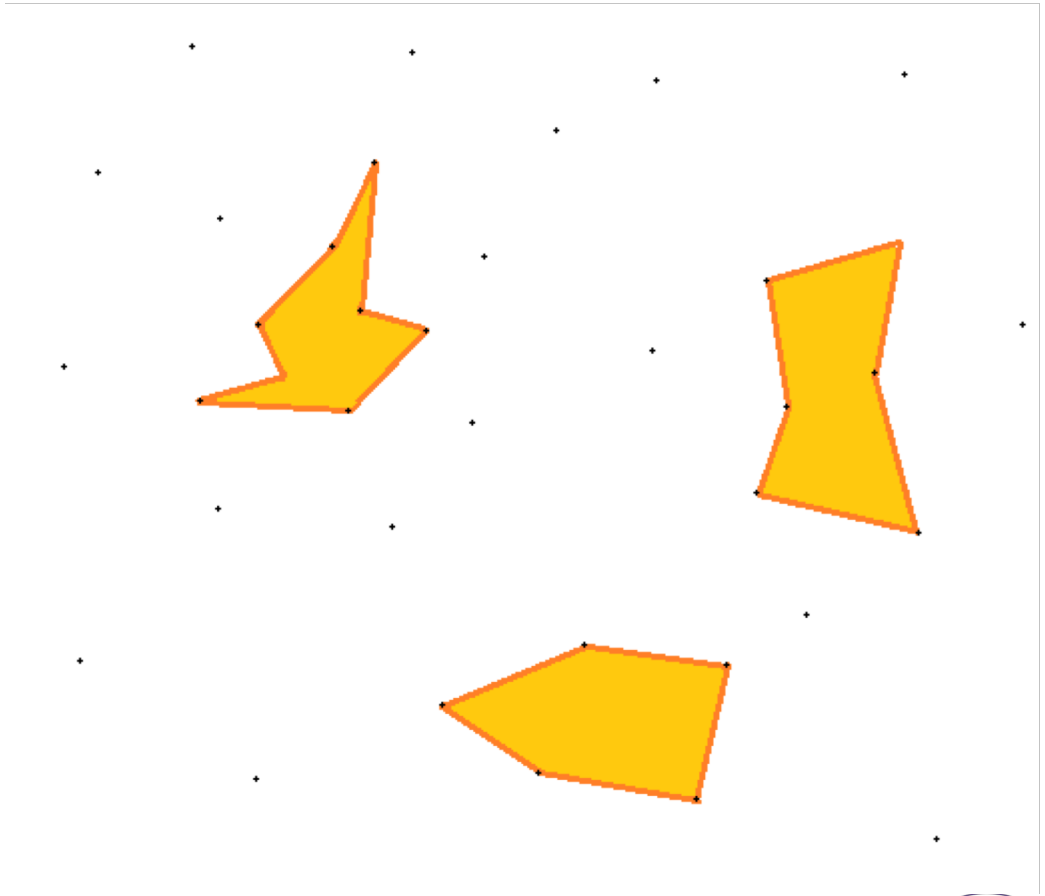
Given a set of input points

Mapper
- Create/extend polygons:
  - density of polygon>t1

Reducer
- re-compute the polygons

Repeat until stabilization

For each Polygon
- *Calculate log-Likelihood value*
- *If log-likelihood value > t2, it is a hotspot*



MapReduce Version of Algorithm

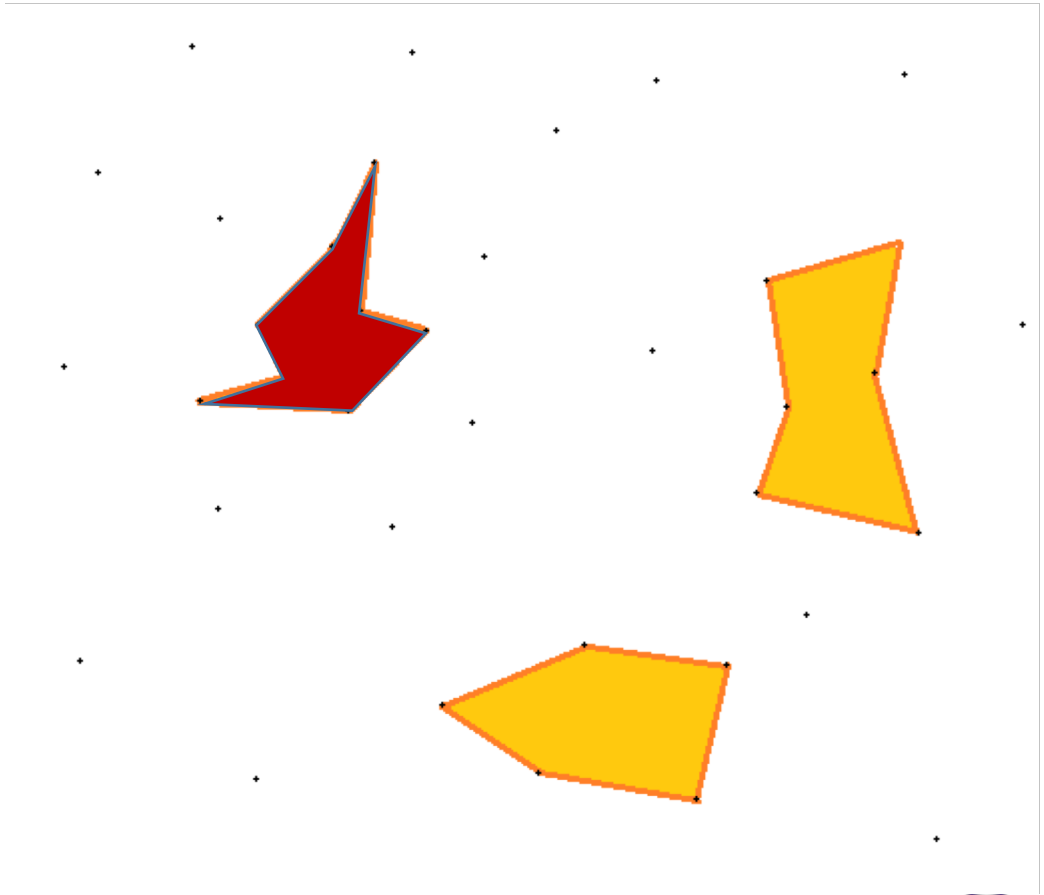
Given a set of input points

Mapper
- Create/extend polygons:
  - density of polygon>t1

Reducer
- re-compute the polygons

Repeat until stabilization

For each Polygon
- *Calculate log-Likelihood value*
- *If log-likelihood value > t2, it is a hotspot*

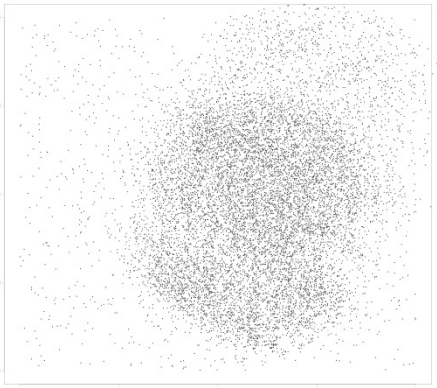


Run Time Comparison

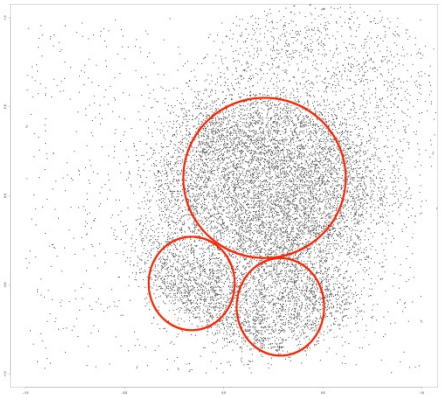
- Serial Implementation
  - CPU: Quad-Core 2.5 GHz
  - 8 GB RAM
  - 9 hours 32 minutes
- Hadoop Implementation (3-node cluster of)
  - CPU: Dual Six-Core 2.0 GHz
  - 16 GB RAM
  - 48 minutes

Run time reduced about 90%!

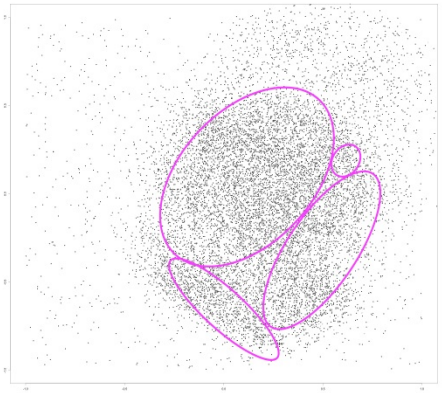
Hotspots Detection on a Simulated Dataset



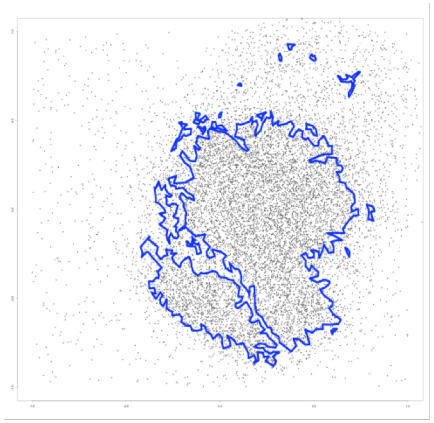
Original data



Circular scan statistics

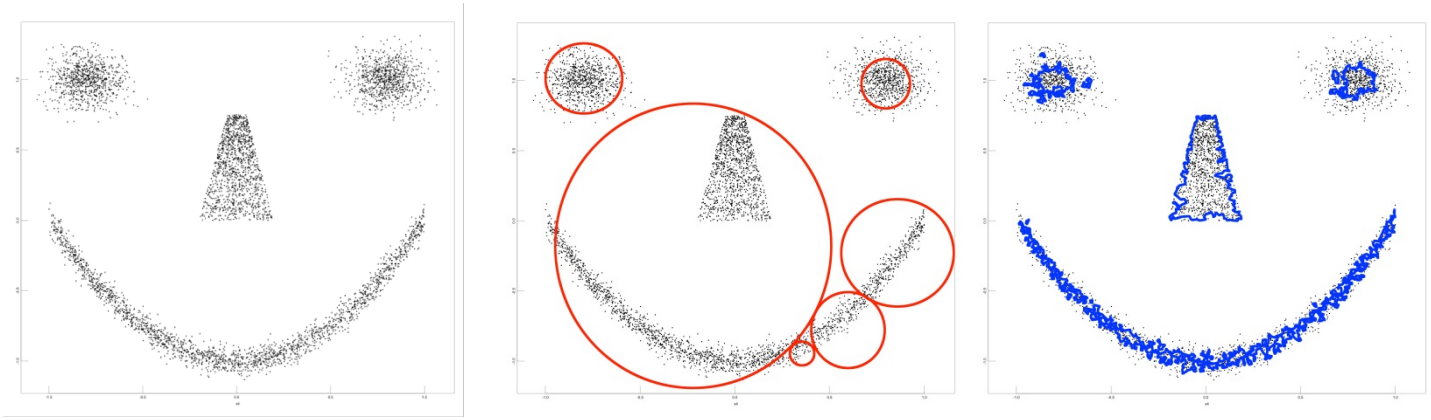


Elliptical scan statistics



Our method

Hotspots Detection on Smiley Dataset



Original data

Circular scan statistics

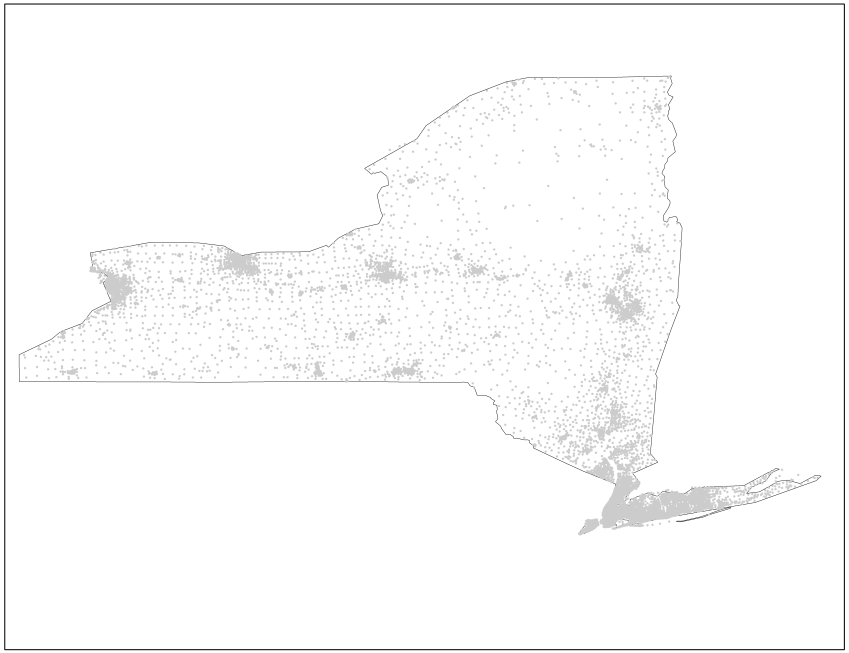
Our method

Significant reduction in false positives – a major advantage

Test on Public Health Dataset

New York Breast Cancer Data

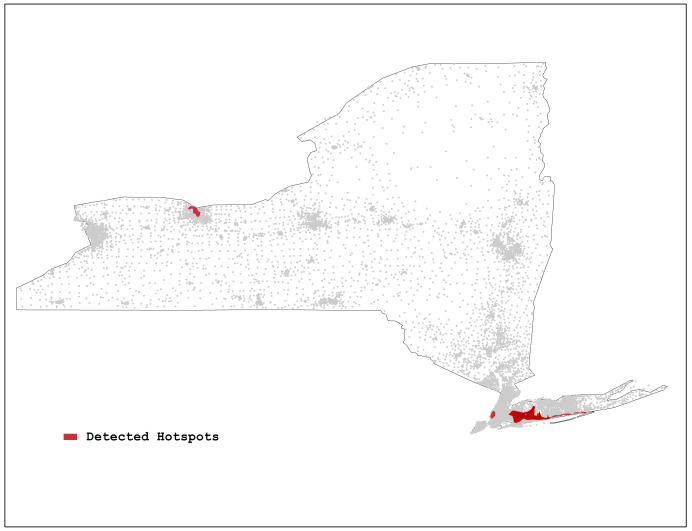
- Sources: NYSDH 2015, Bosceo et. al, 2016
- Breast cancer cases aggregated to census block
- 13,848 points (centroid of a census block), representing 72, 926 patients and a population of 27, 820, 632 in 2009 in the State of New York



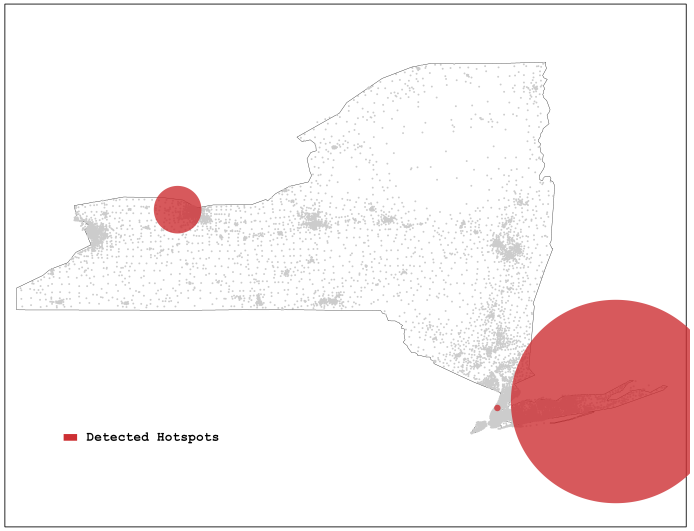
VDI Center for Visual & Decision Informatics

UNIVERSITY OF NORTH ALABAMA 18 30

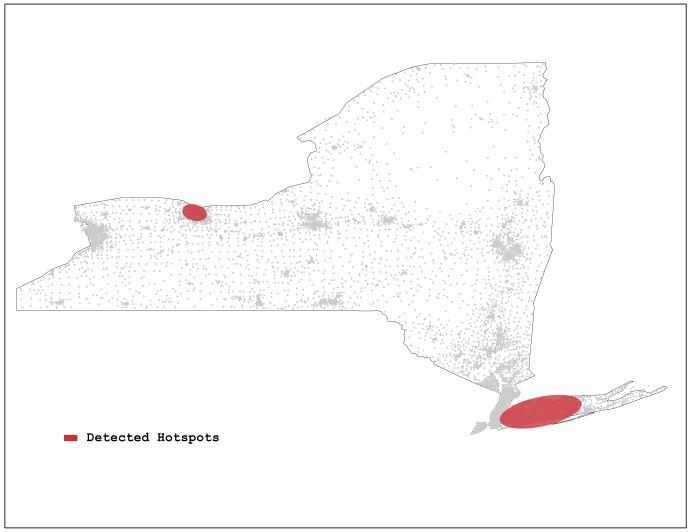
Test on Public Health Dataset



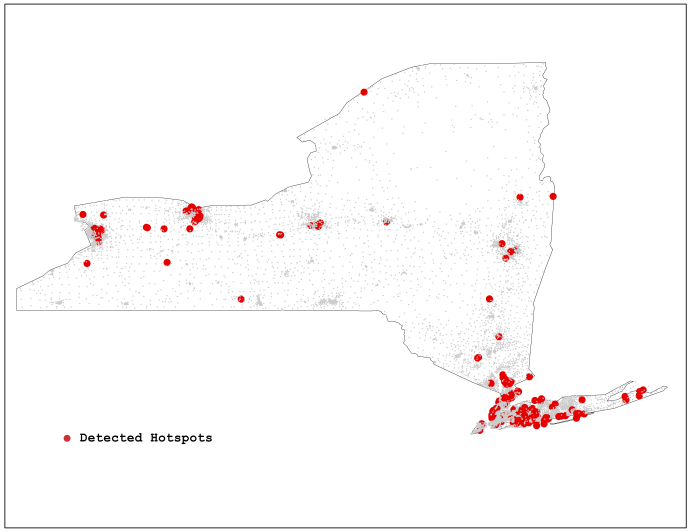
Our Method



Circular Scan Statistics



Elliptic Scan Statistics



Getis-Ord Gi*

Conclusion

- Polygon propagation (PP) can detect better heterogeneous cluster that are irregularly shaped
- PP detects more compact hotspots area (has less false positive errors) than the state-of-the-art Scan Statistics
- MapReduce based PP significantly outperforms the traditional Scan Statistics and PP algorithms, make it suitable for large-scale spatial hotspots detection

- Future work
  - Identify better ways to calculate the stability of the detected hotspot
  - Reduce the computational time of the greedy search

Acknowledgement

- NSF Grants CNS-16505511 and IIP-1160958

- NSF I/UCRC CVDI IAB grant CS-14-4

- University Research Grant, University of North Alabama

References

- Anselin, L. 1995. “Local Indicators of Spatial Association-LISA.” Geographical Analysis 27 (2): 93–115.
- Boscoe, F., T. Talbo, and M. Kulldorff. 2016. “Public Domain Small-area Cancer Incidence Data for New York State, 2005–2009.” Geospatial Health 11 (1): 304.
- Getis, A., and J. K. Ord. 1992. “The Analysis of Spatial Association by Use of Distance Statistics.” Geographical Analysis 24 (3): 189–206.
- Getis, A. 2007. “Reflections on Spatial Autocorrelation.” Regional Science and Urban Economics 37: 491–496.
- Gao, P, P. Guo, K. Liao, J. J. Webb, and S. L. Cutter. 2013. "Early detection of terrorism outbreaks using prospective space–time scan statistics." *The Professional Geographer* 65(4): 676-691.
- Katragadda, S., J. Chen, and S. Abbady. 2018. “Spatial Hotspot Detection using Polygon Propagation”, International Jornal of Digital Earth, DOI: 10.1080/17538947.2018.1485754
- Kulldorff, M. 1997. “A Spatial Scan Statistic.” Communications in Statistics - Theory and Methods 26 (6): 1481–1496.
- Kulldorff, M., L. Huang, L. Pickle, and L. Duczmal. 2006. “An Elliptic Spatial Scan Statistic.” Statistics in Medicine 25 (22): 3929–3943.
- Moran, P. A. P. 1950. “Notes on Continuous Stochastic Phenomena.” Biometrika 37 (1/2): 17–23.
- NYSDH (New York State Department of Health). 2015. “Cancer Mapping Data: 2005-2009”, https://health.data.ny.gov/Health/Cancer-Mapping-Data-2005-2009/cw3n-fkji?category=Health&view_name=Cancer-Mapping-Data-2005–2009
- Ord, J. K., and A. Getis. 2001. “Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation.” Journal of Regional Science 41 (3): 411–432.

Questions & Comments