Happy or Not: Generating Topic-based Emotional Heatmaps for Culturomics using CyberGIS

Eric Shook †, Kalev Leetaru ‡, Guofeng Cao †,
Anand Padmanabhan †§ and Shaowen Wang †§

{eshook2,leetaru,guofeng,apadmana,shaowen}@illinois.edu

† Department of Geography and Geographic Information Science,
University of Illinois at Urbana-Champaign

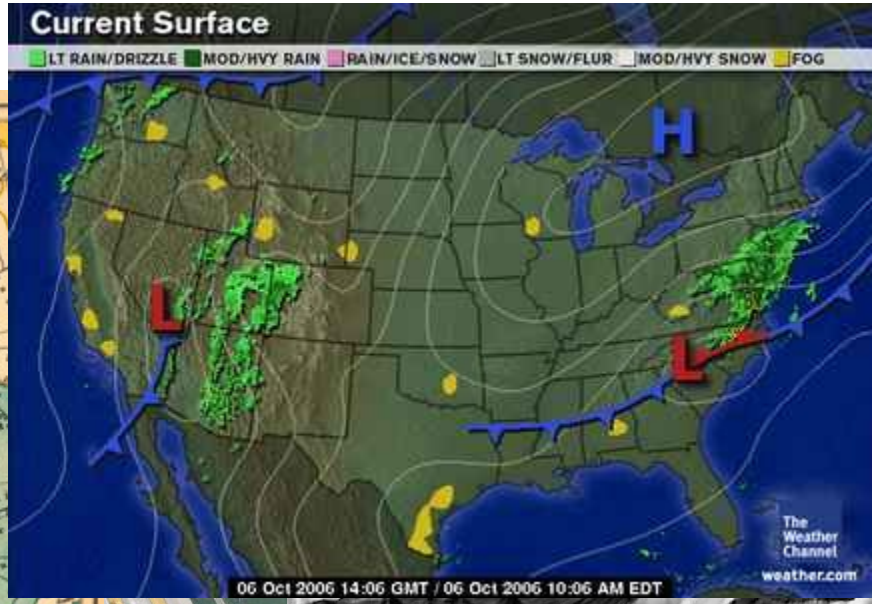
‡ University of Illinois at Urbana-Champaign

§ National Center for Supercomputing Applications,
University of Illinois at Urbana-Champaign

October 8, 2012

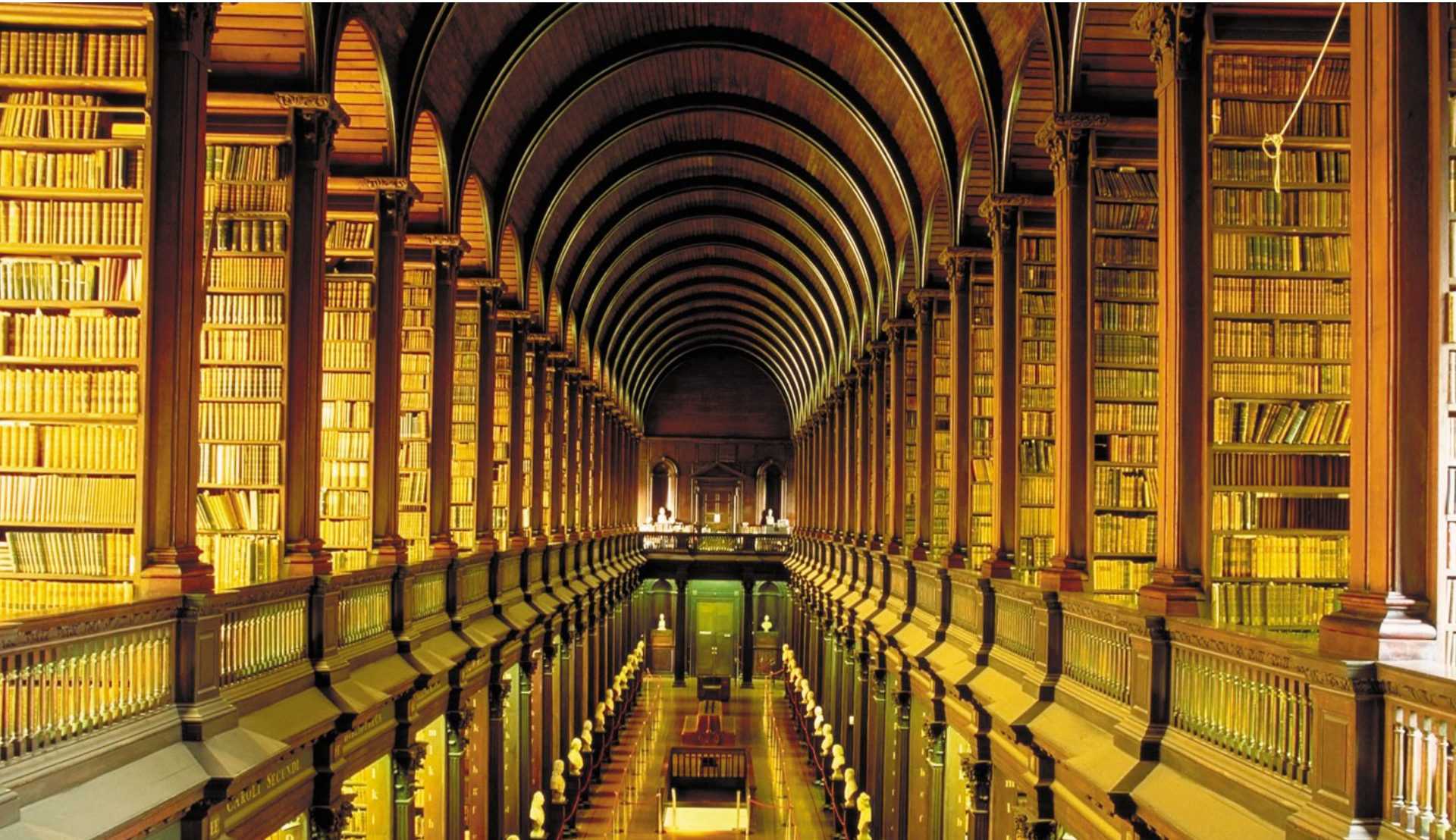
Maps


CENTRAL EUROPE
1789

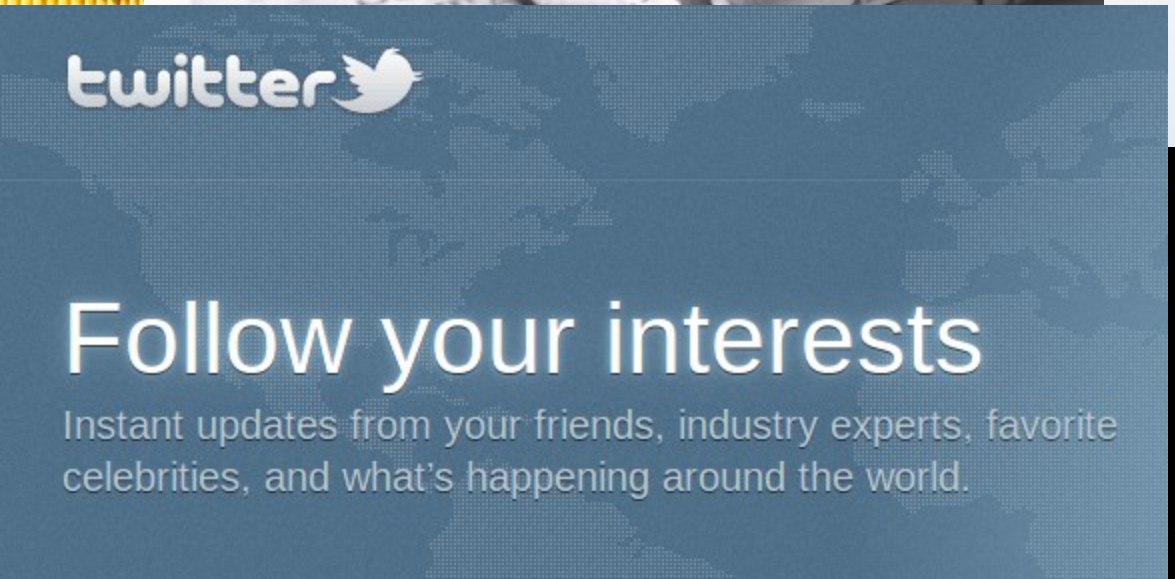

Current Surface
LT RAIN/DRIZZLE
MOD/HVY RAIN
RAIN/ICE/SNOW
LT SNOW/FLUR
MOD/HVY SNOW
FOG
H
L
L
The Weather Channel
weather.com
06 Oct 2006 14:06 GMT / 06 Oct 2006 10:06 AM EDT


Cermak Rd
A
B
McCormick Dr
Donnelly Rd
McCormick Place Metra
McCormick Place Busway
S Lake Shore Dr
Moe Dr
Mc 6th
S Martin L
2315

Untapped Spatial Information



Untapped Spatial Information



Goal

Develop a geospatial visual analytical approach and system architecture that transforms vast amounts of textual data into emotional heatmaps for capturing spatial characteristics of latent tones

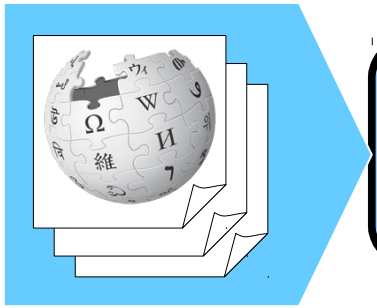
System Architecture

Heatmap Approach

Parallelization of Spatial Methods

Concluding Discussions

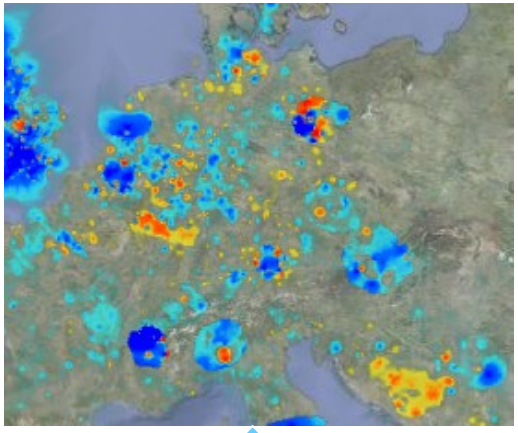
System Architecture



Fulltext Geocoding

Sentiment Mining

CyberGIS



Extracting Information from Text

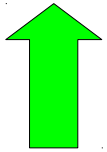
Fulltext
Geocoding

… I love **Chicago, IL** …

… **New York City** is awful ...

Extracting Information from Text

Sentiment Mining

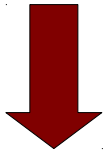


Positive

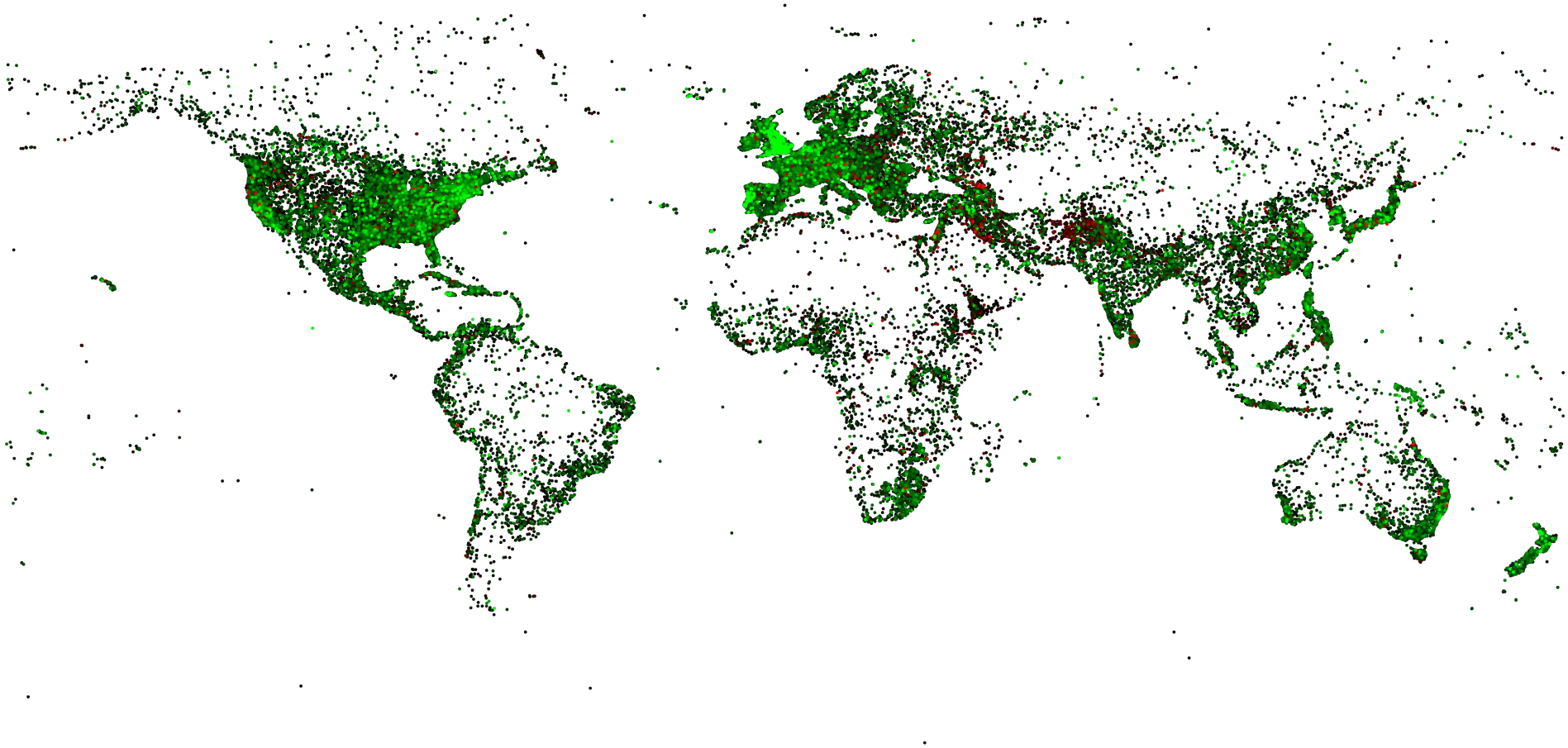
… I **love** Chicago, IL …

… New York City is **awful** ...

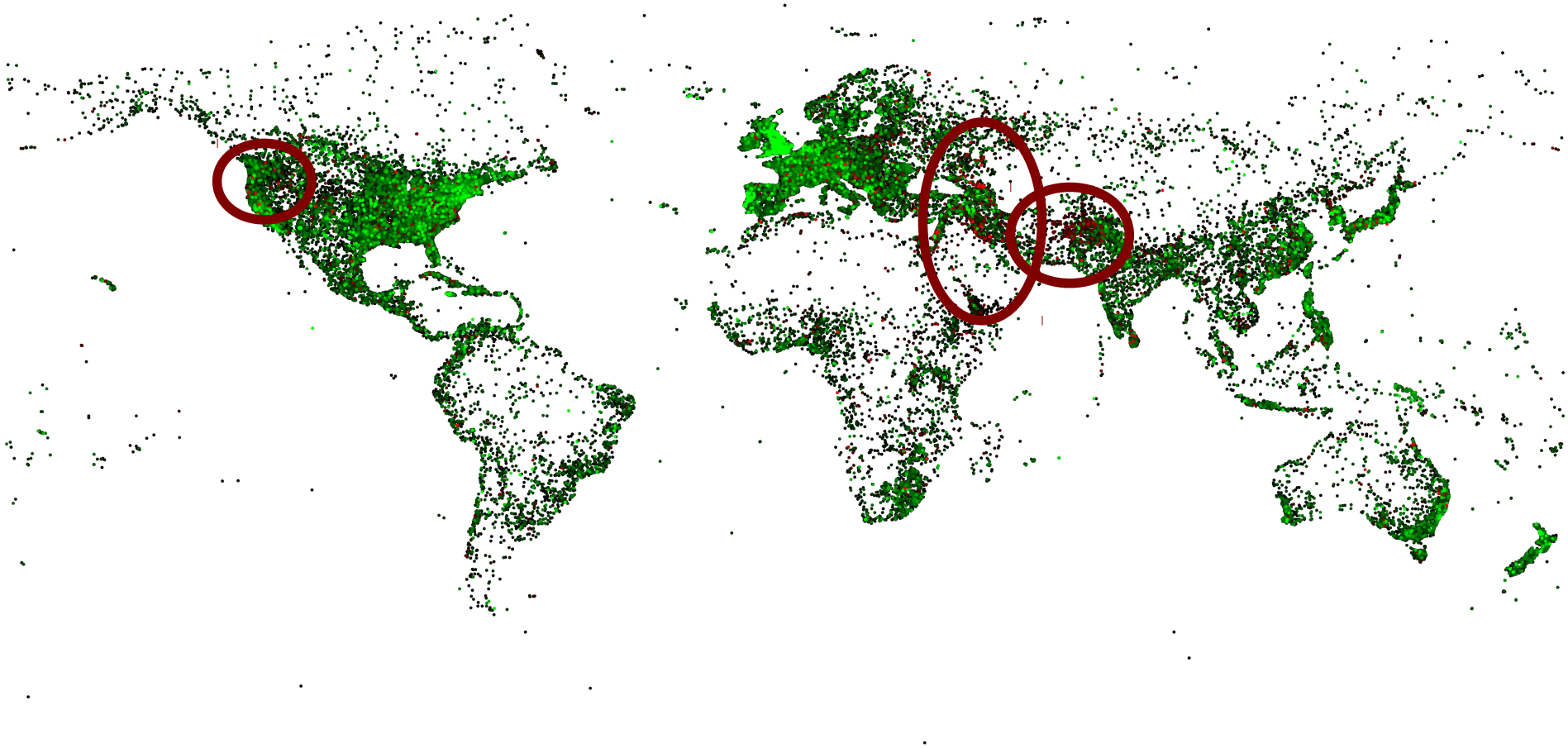
Negative



Point Cloud Visualization



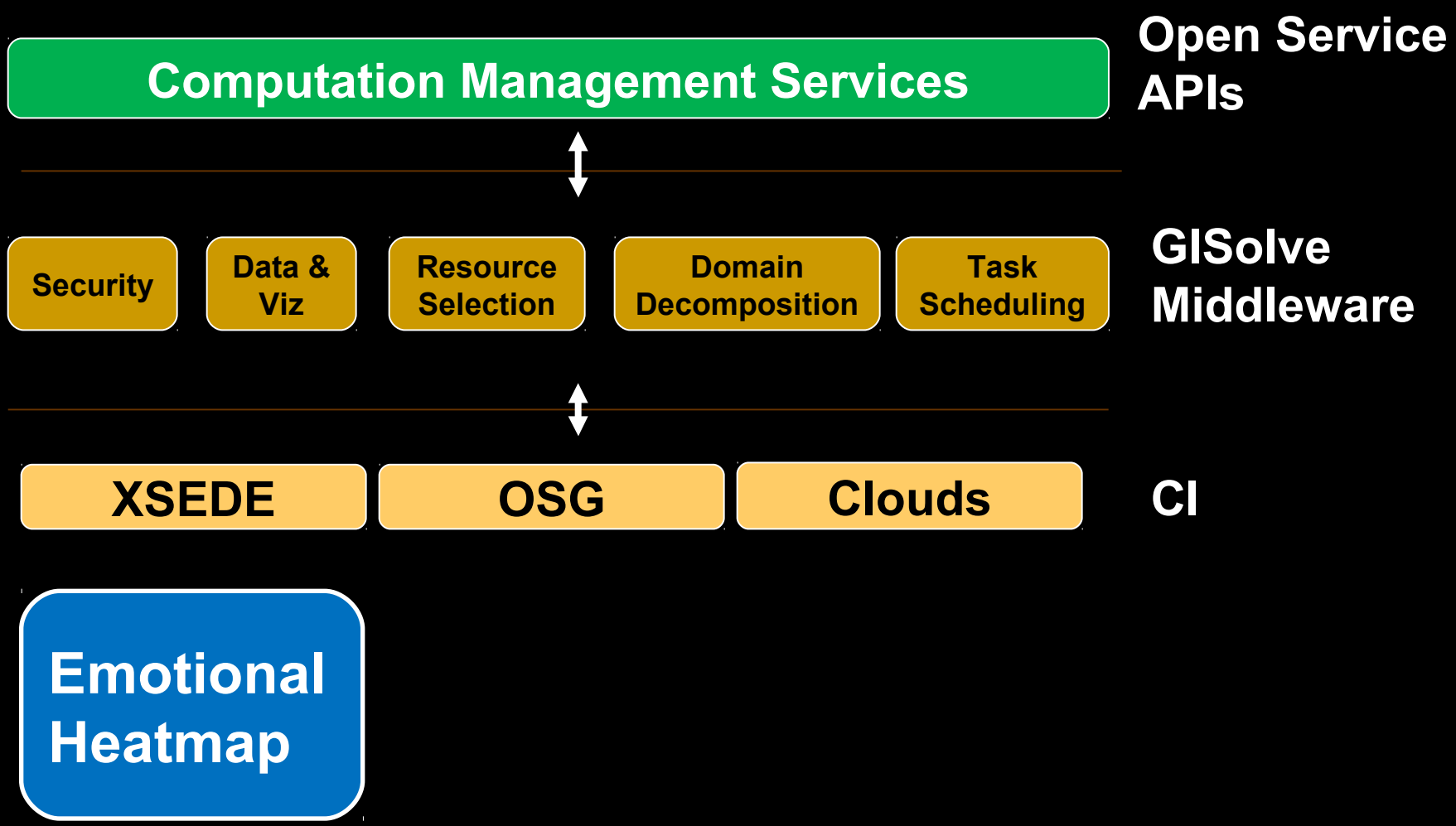
Point Cloud Visualization



Inside the CyberGIS "Blackbox"

CyberGIS is a new generation of Geographic Information Systems (GIS), which are designed to read, write, analyze, and visualize spatiotemporal data, supported by CyberInfrastructure (CI)

Inside the CyberGIS "Blackbox"



XSEDE

"The most advanced, powerful, and robust collection of integrated advanced digital resources and services in the world" (https://www.xsede.org/overview)

System Architecture

Heatmap Approach

Parallelization of Spatial Methods

Concluding Discussions

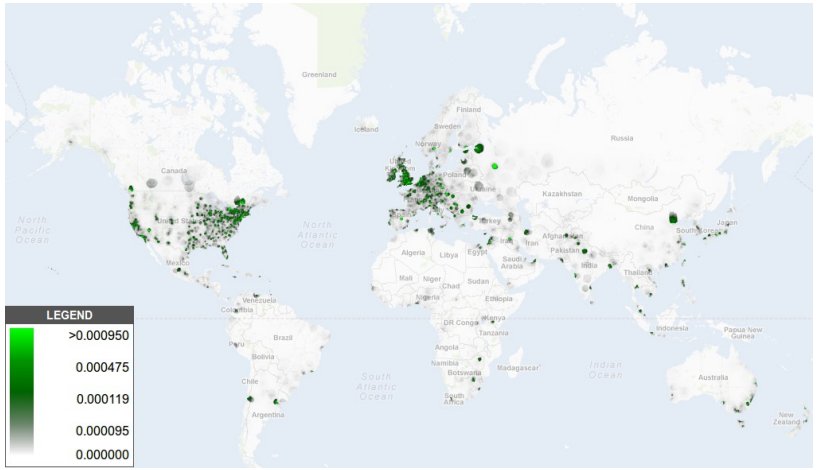
Heatmap Approach

Goal:

Capture 3 important elements extracted from text in a single map

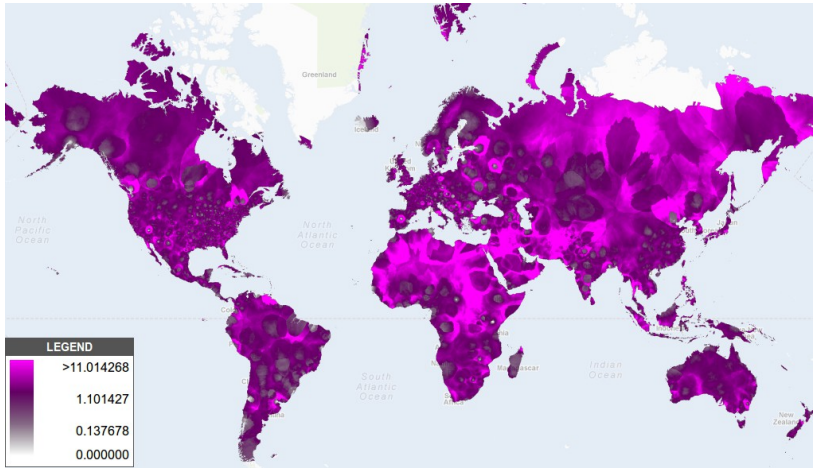
1. Importance of location - Article Density
2. Prevalence of topic - Topic intensity
3. Emotion toward topic - Tone

Emotional Heatmap



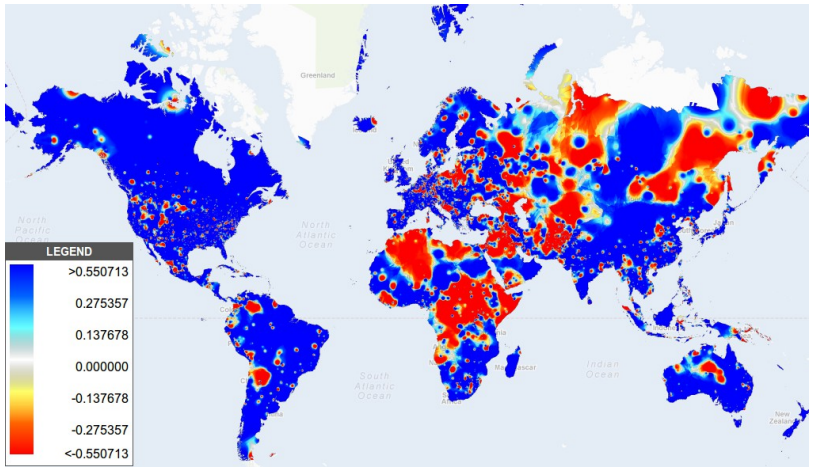
Article Density

*



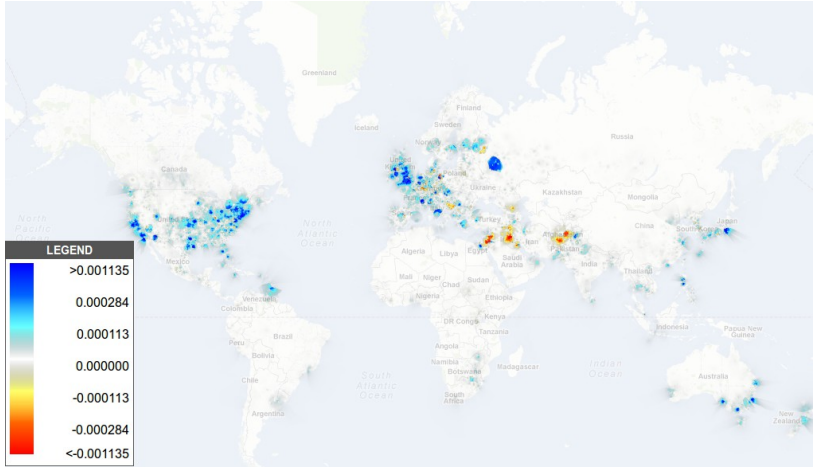
Topic Intensity

*



Tone

=



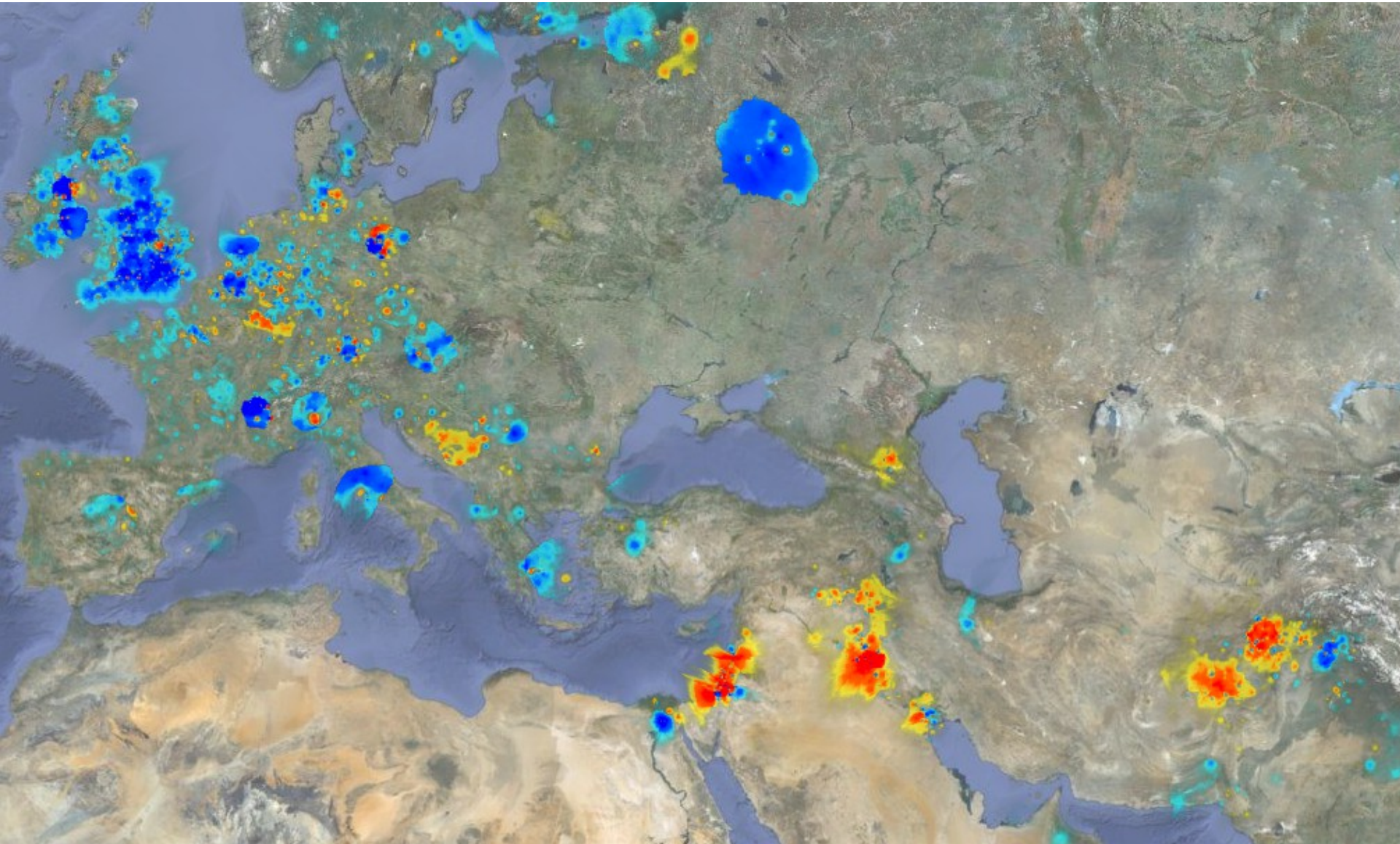
Emotional Heatmap

The SGI Wikipedia Project

Kalev Leetaru used an SGI UV2000 supercomputer to apply fulltext geocoding and sentiment mining to the English edition of Wikipedia extracting over 80 million locations

As a case study we generated emotional heatmaps of "armed conflict" for the year 2003 from this Wikipedia dataset

Emotional Heatmap of Armed Conflict in 2003 (Wikipedia)



System Architecture

Heatmap Approach

Parallelization of Spatial Methods

Concluding Discussions

Spatial Analysis Methods

1) Kernel density estimation (KDE) estimates a probability density function
2) Inverse distance weighted (IDW) interpolation estimates continuous surfaces such as tone

Both methods transform point-based data into 2D lattice data referred to as raster surfaces based on a *K* nearest neighbor search

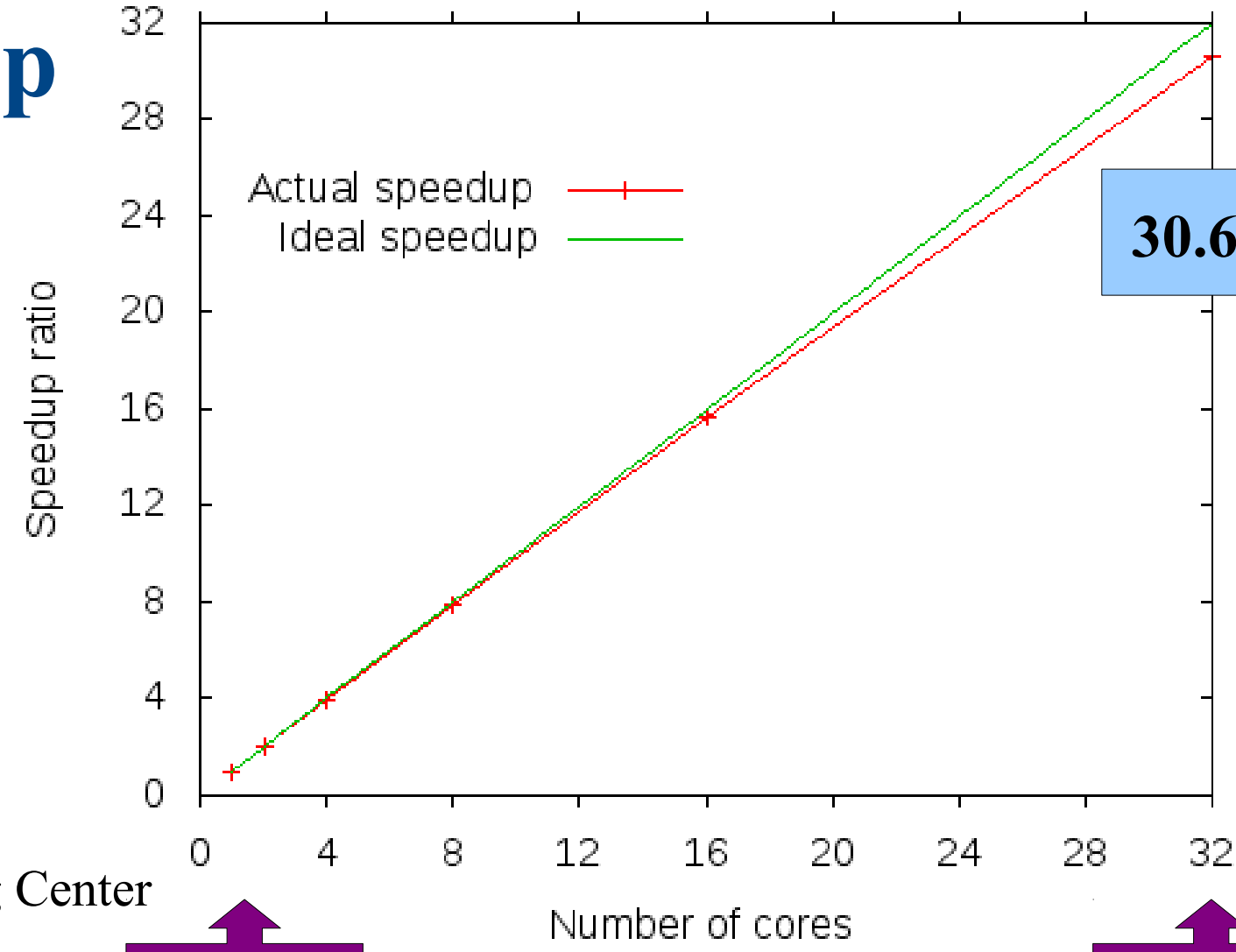
OpenMP

- OpenMP is a shared memory paradigm to facilitate coordination amongst threads executing in parallel
- **Portable** – supported by most supercomputers
- **Scalable** – enables parallel execution of hundreds if not thousands of threads
- **Simple** – The interface is designed for simplicity and flexibility

OpenMP

- Row-based spatial domain decomposition
  - Easy to parallelize in OpenMP
- Dynamic thread scheduling
  - Automatic load-balancing
  - Each thread is assigned a small chunk of rows (e.g. 2)
  - Once a thread finishes their chunk of rows they are assigned a new chunk

Speedup



30.6X

Trestles
San Diego Supercomputing Center (SDSC)

18 hours

35 minutes

Concluding Discussions

CyberGIS provide new opportunities to address computationally intensive problems including the extraction and visualization of spatial information from text

Concluding Discussions

- System Architecture
  - Provides supporting services for computational methods
  - Gain straightforward access to cyberinfrastructure
    - Computational management system
- Parallelization of spatial methods using OpenMP
  - Easy conversion from serial to parallel
  - Automatic dynamic load-balancing
  - Excellent speedup

Acknowledgements

- **National Science Foundation**
  - BCS-0846655
  - OCI-1047916
  - XSEDE SES070004N

Thank you

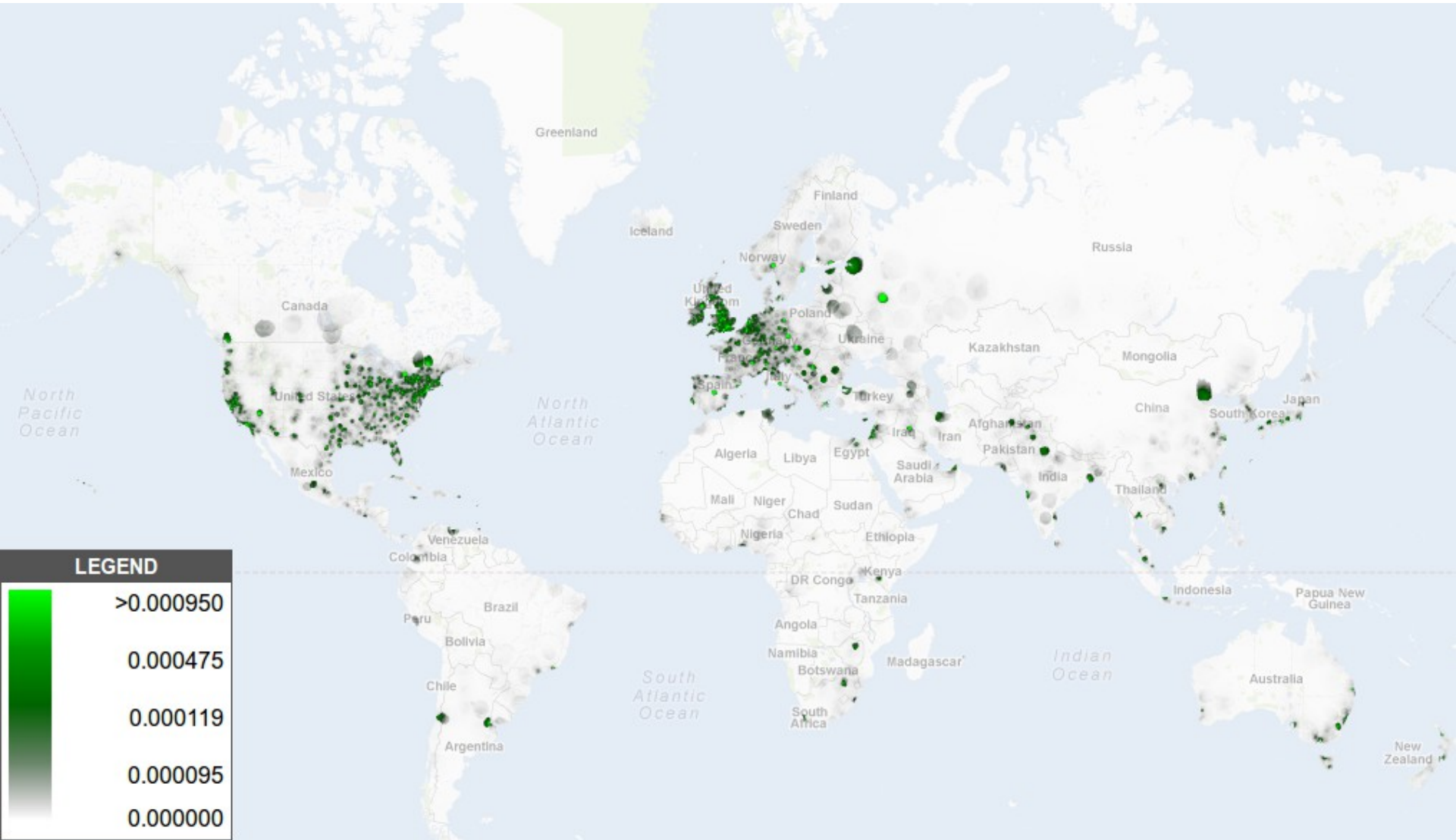
Questions or comments?

1) Importance of Location

- Every mention of a location increases its importance

- Generate a density map of the number of times a location is mentioned in text

1) Importance of Location



2) Prevalence of topic

- We term topic intensity to capture the prevalence of a topic relative to other topics, and adopt a method commonly used in epidemiological studies to estimate it

- Relative risk is a ratio of the densities of disease infection locations and case control locations

Topic Intensity

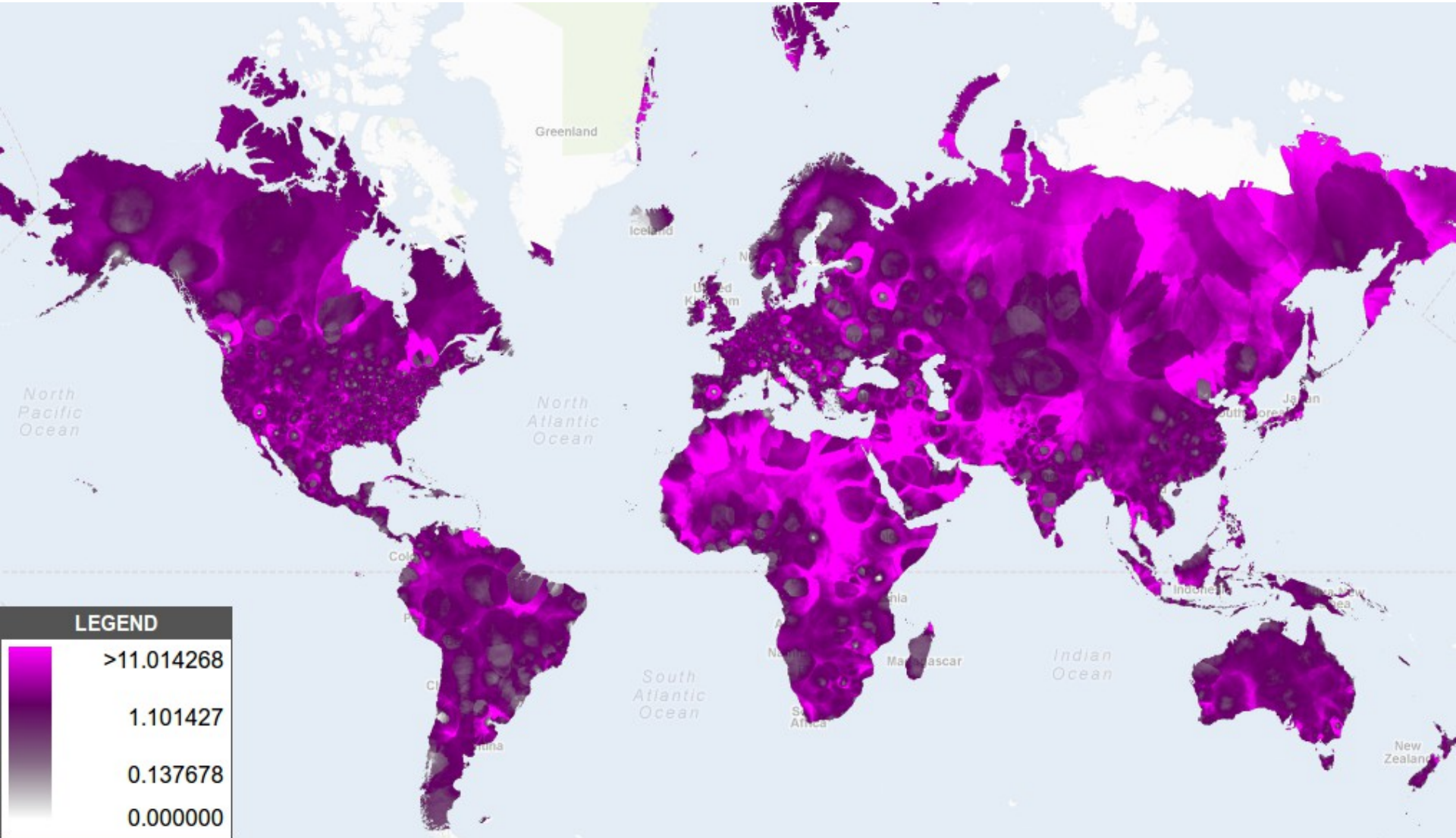
Topic Intensity

$$\frac{\text{Density of articles that mention a topic}}{\text{Density of articles that do not mention a topic}}$$

Relative Risk

$$\frac{\text{Density of locations with disease}}{\text{Density of locations without disease}}$$

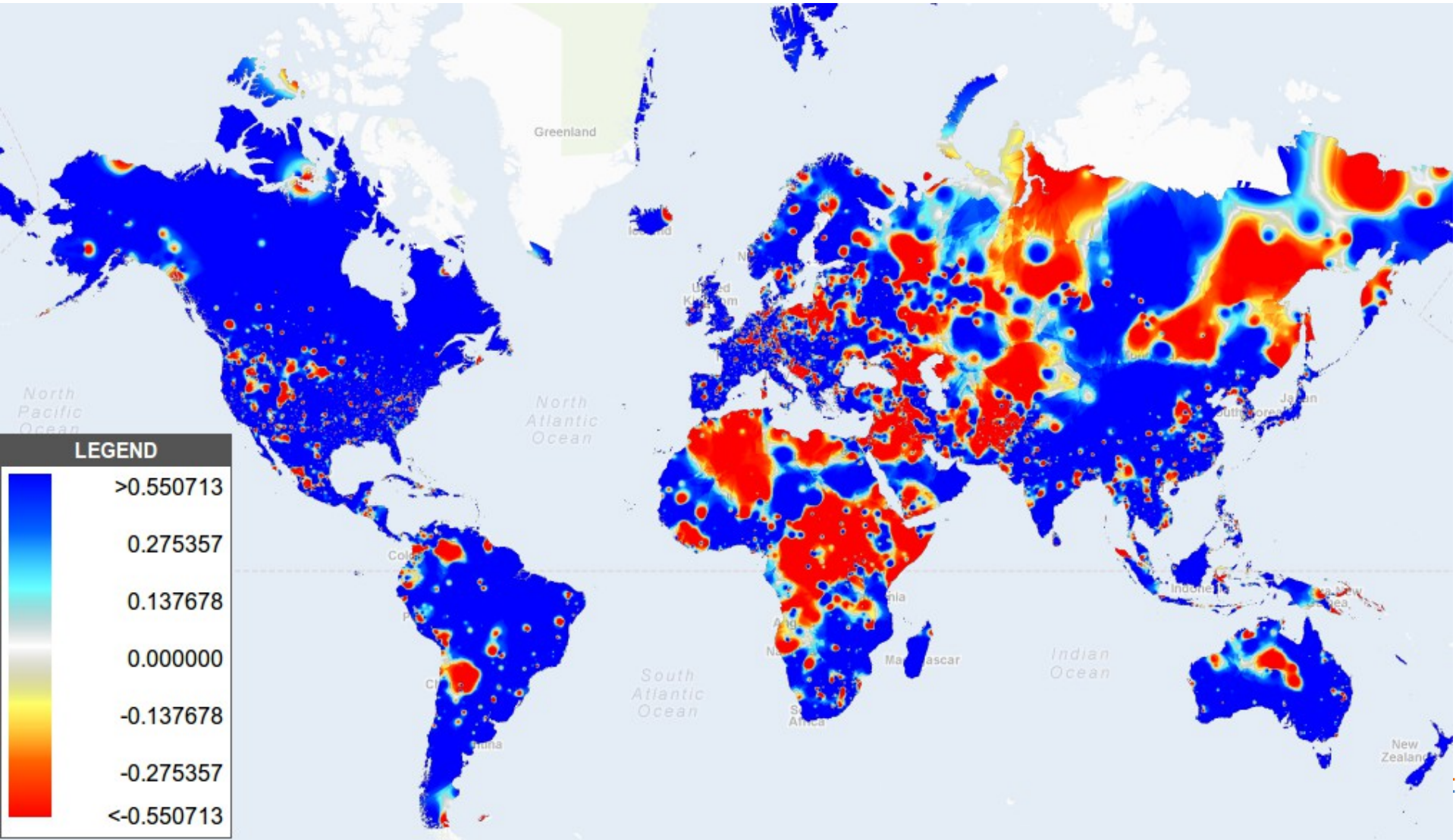
Topic Intensity



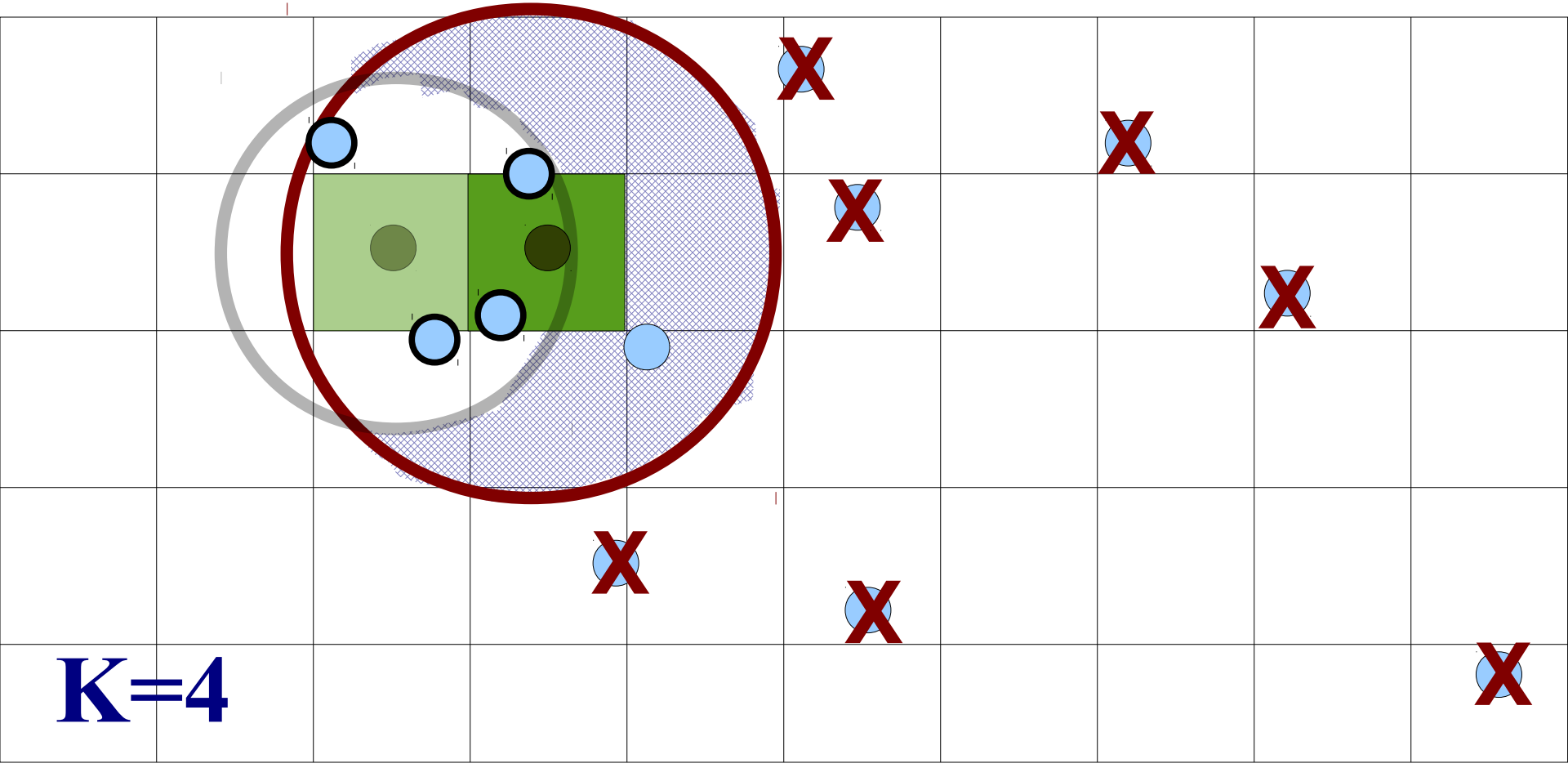
3) Emotion toward topic

- Tone map captures positive and negative tone toward a particular topic across space

3) Emotion toward topic



K Nearest Neighbor Search – Learned Search



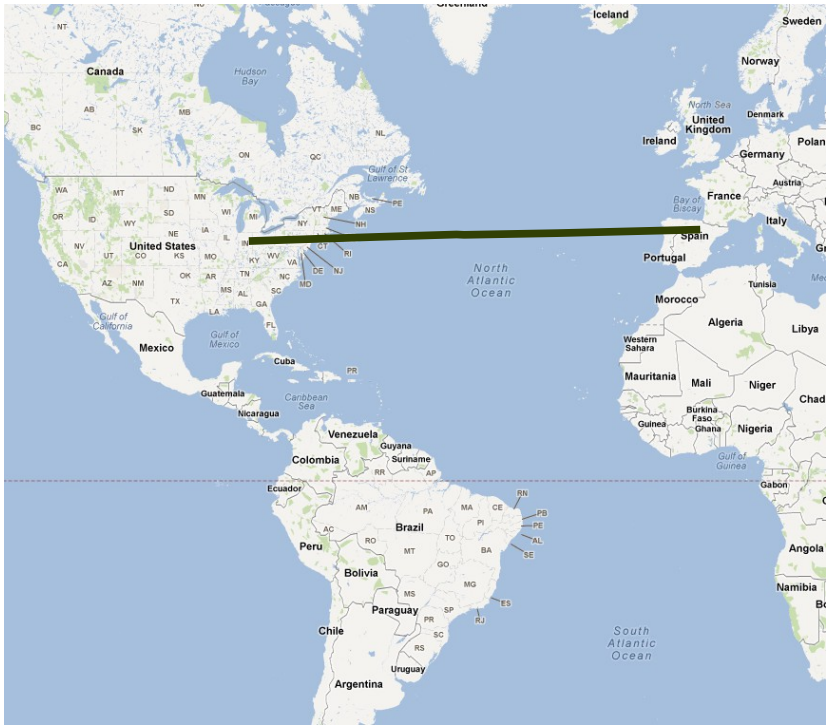
(Hodgson, 1989)

Algorithmic Tradeoffs

Great circle distance

$$= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Euclidean distance



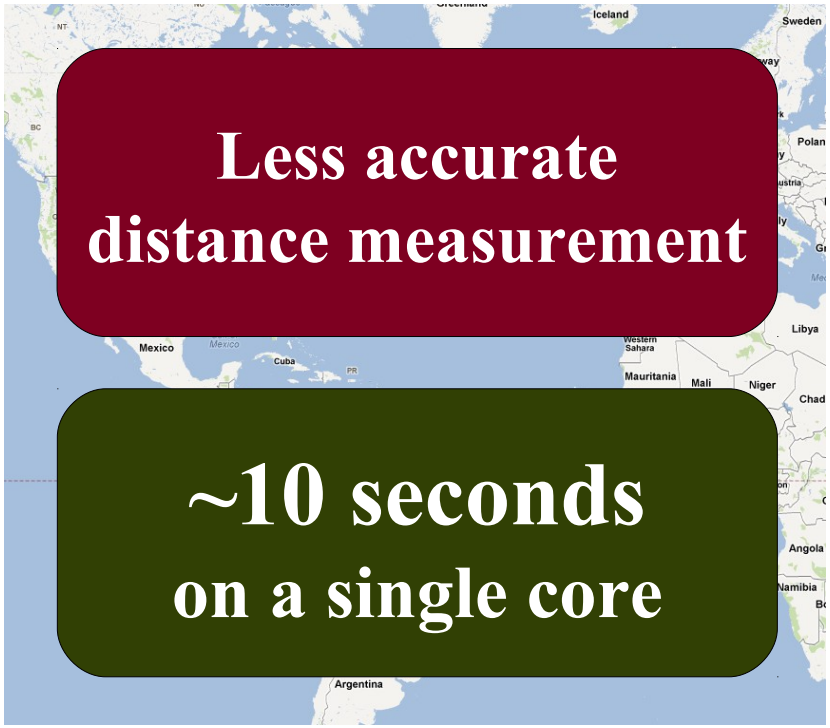
$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

Algorithmic Tradeoffs

Great circle distance

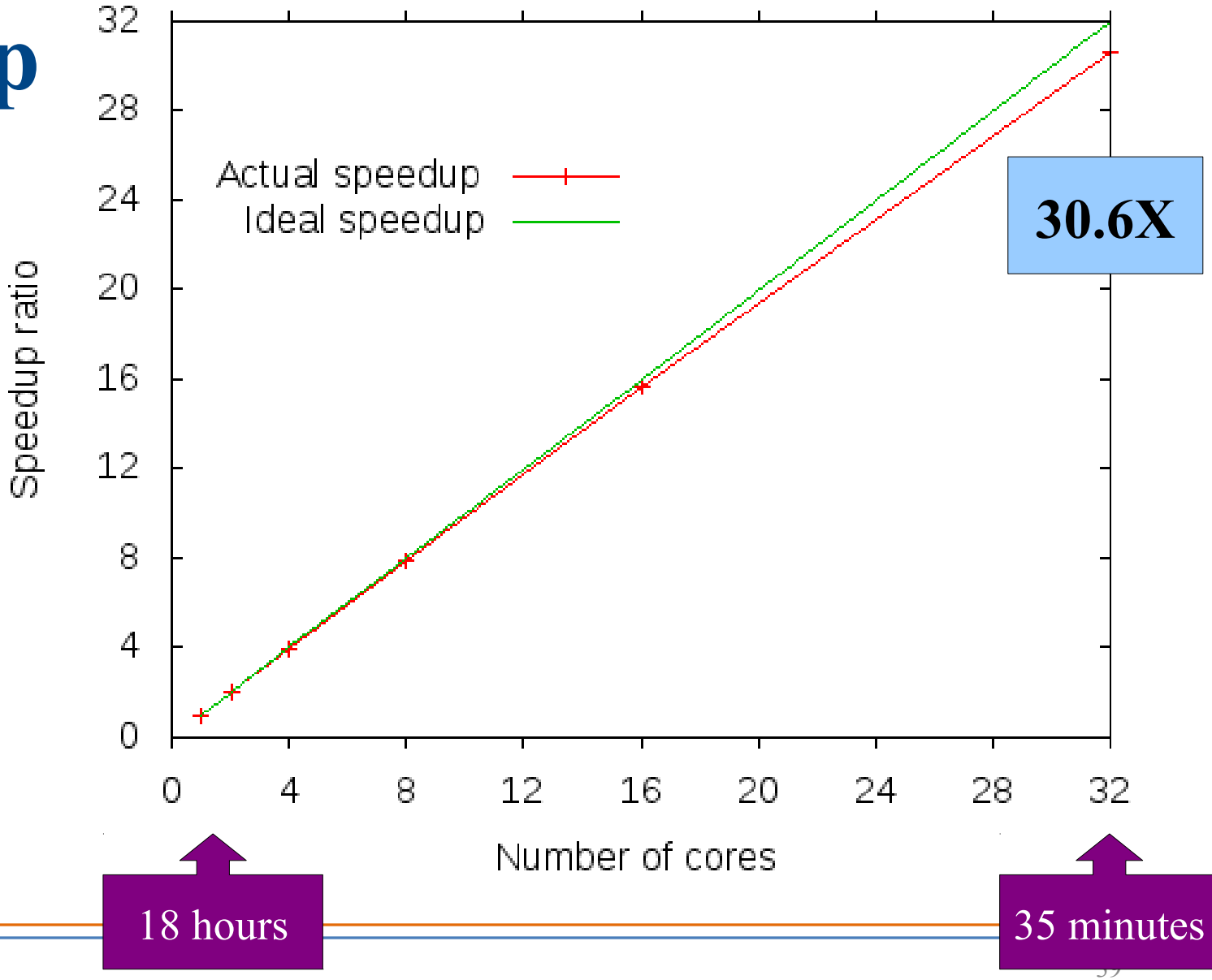
$$= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Euclidean distance



$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

Speedup


Actual speedup
Ideal speedup
Speedup ratio
Number of cores
0
4
8
12
16
20
24
28
32
30.6X
18 hours
35 minutes

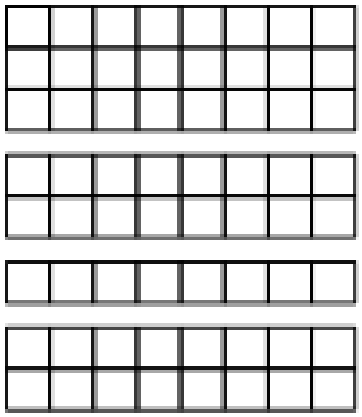
Open Service APIs

- Goal: Provide an easy-to-use API for cyberinfrastructure-based geospatial computation
- API design
  - Not the synchronous request-response model
  - Methods:
    - Submission (as REST POST)
    - Monitoring (as REST GET)
    - Fetching results (as REST GET)

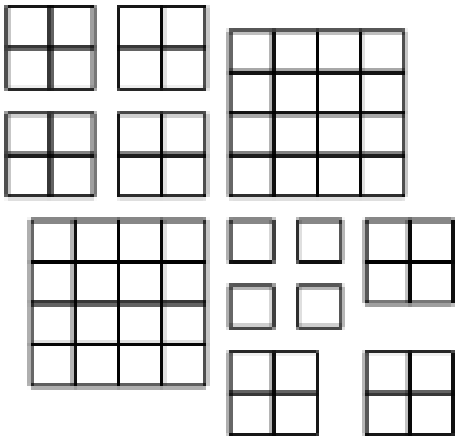
GISolve Middleware

- Resource Selection
- Task Scheduling
  - Tasks (i.e. jobs) are scheduled based on a series of dependencies (input data copy, job submission, job status, job completion, output data copy, visualization)
- Data and Visualization
  - Data and visualization management systems were developed to handle large-scale geospatial data

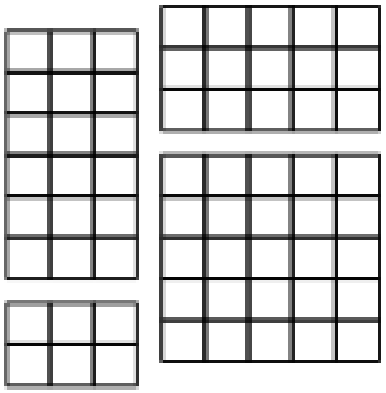
Spatial Domain Decomposition



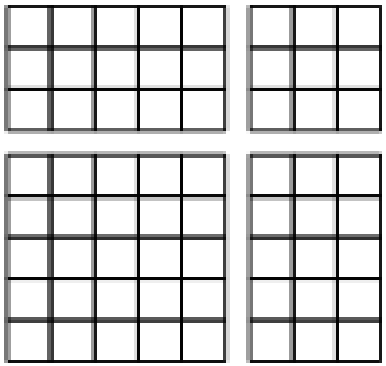
Row or Column



Quadtree



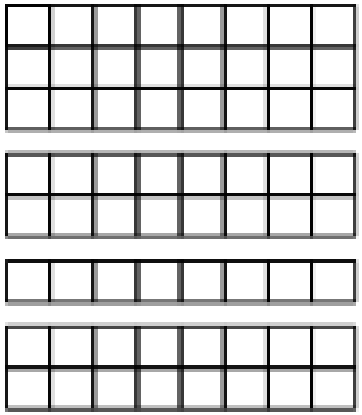
Recursive Bisection



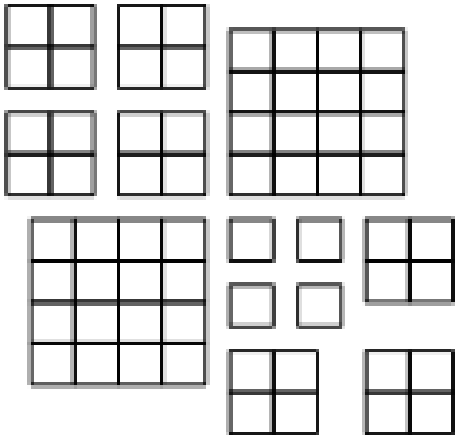
Grid

(Ding and Densham, 1996)

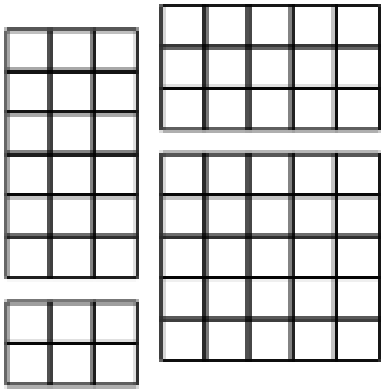
Spatial Domain Decomposition



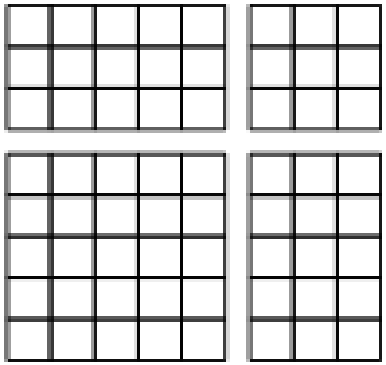
Row or Column



Quadtree



Recursive Bisection



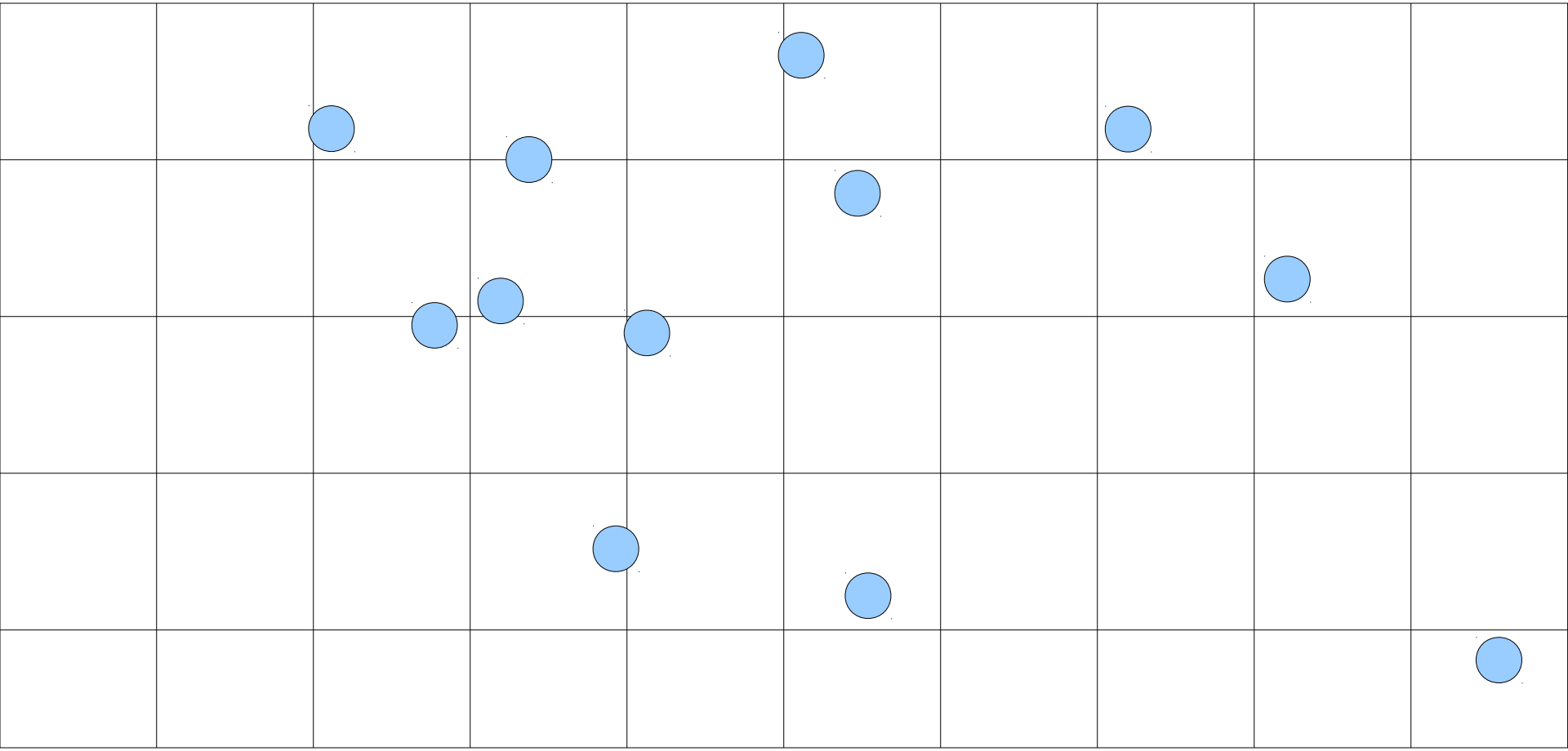
Grid

(Ding and Densham, 1996)

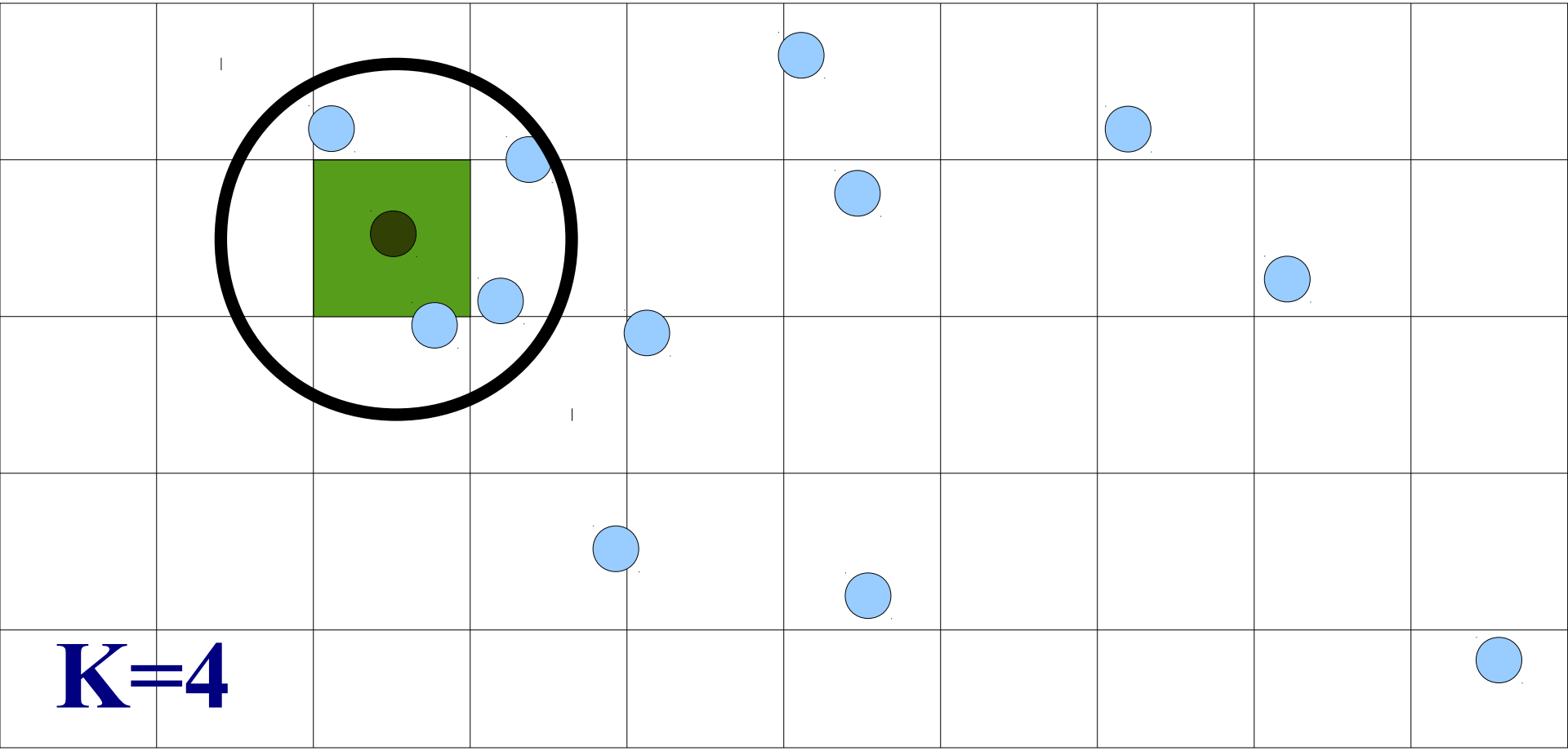
3 Important Elements

1) Importance of Location (Article Density)
   - Every mention of a location increases its importance
2) Prevalence of topic (Topic Intensity)
   - We term topic intensity to capture the prevalence of a topic relative to other topics and map it across space
3) Emotion toward a topic (Tone)
   - Tone map captures positive and negative tone toward a particular topic across space

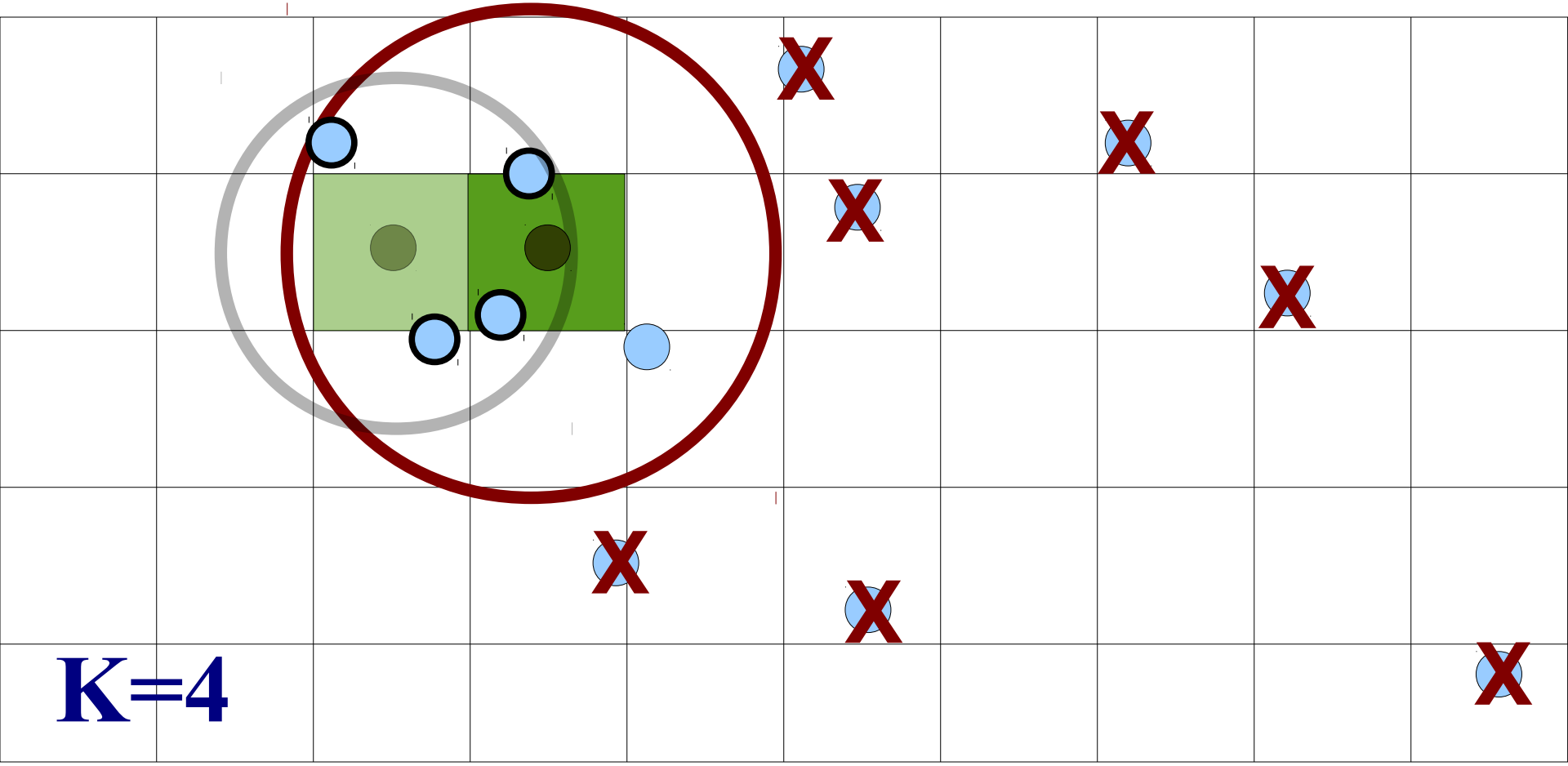
Raster Surface and Locations



K Nearest Neighbor Search

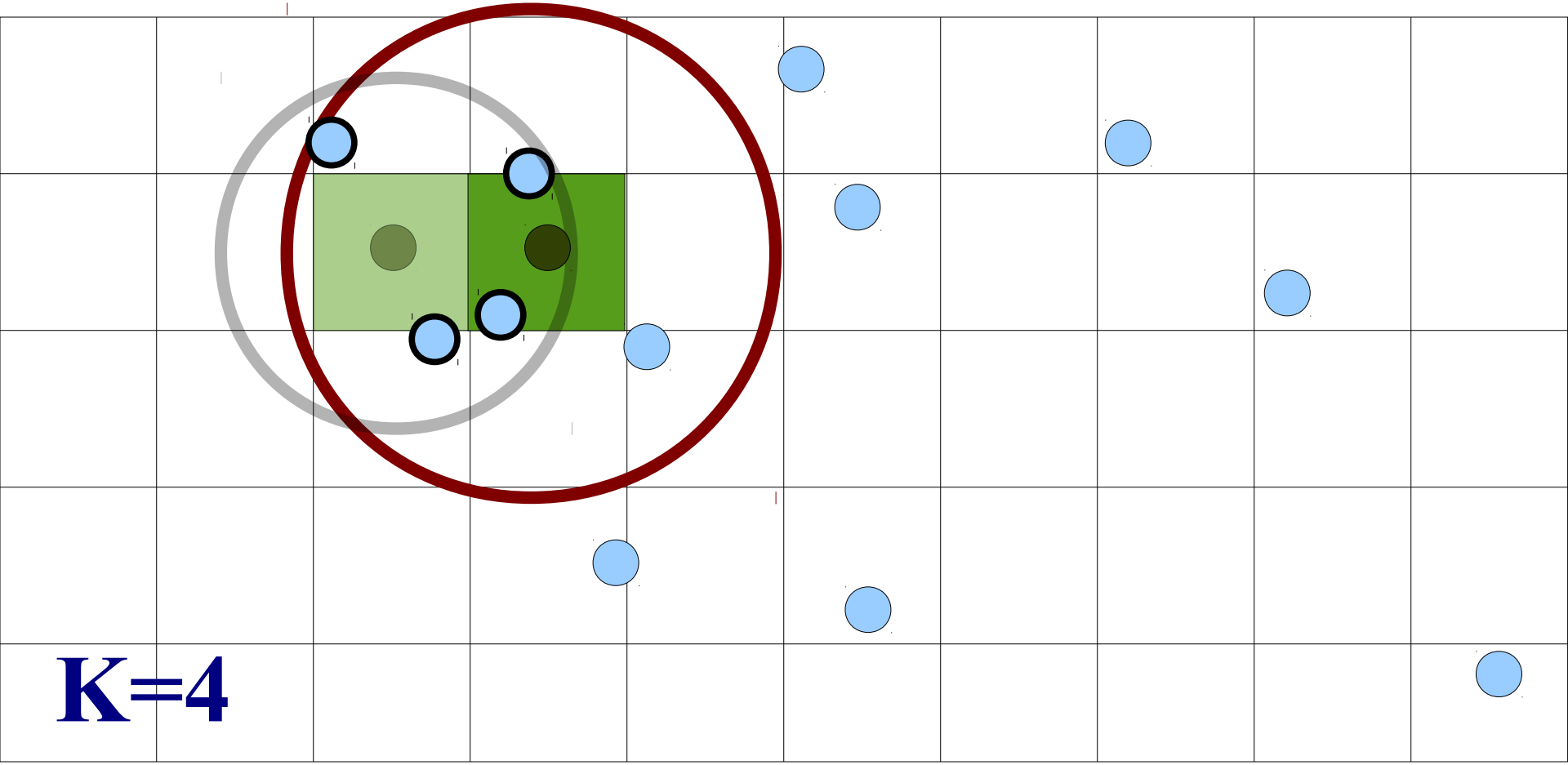


K Nearest Neighbor Search – Learned Search



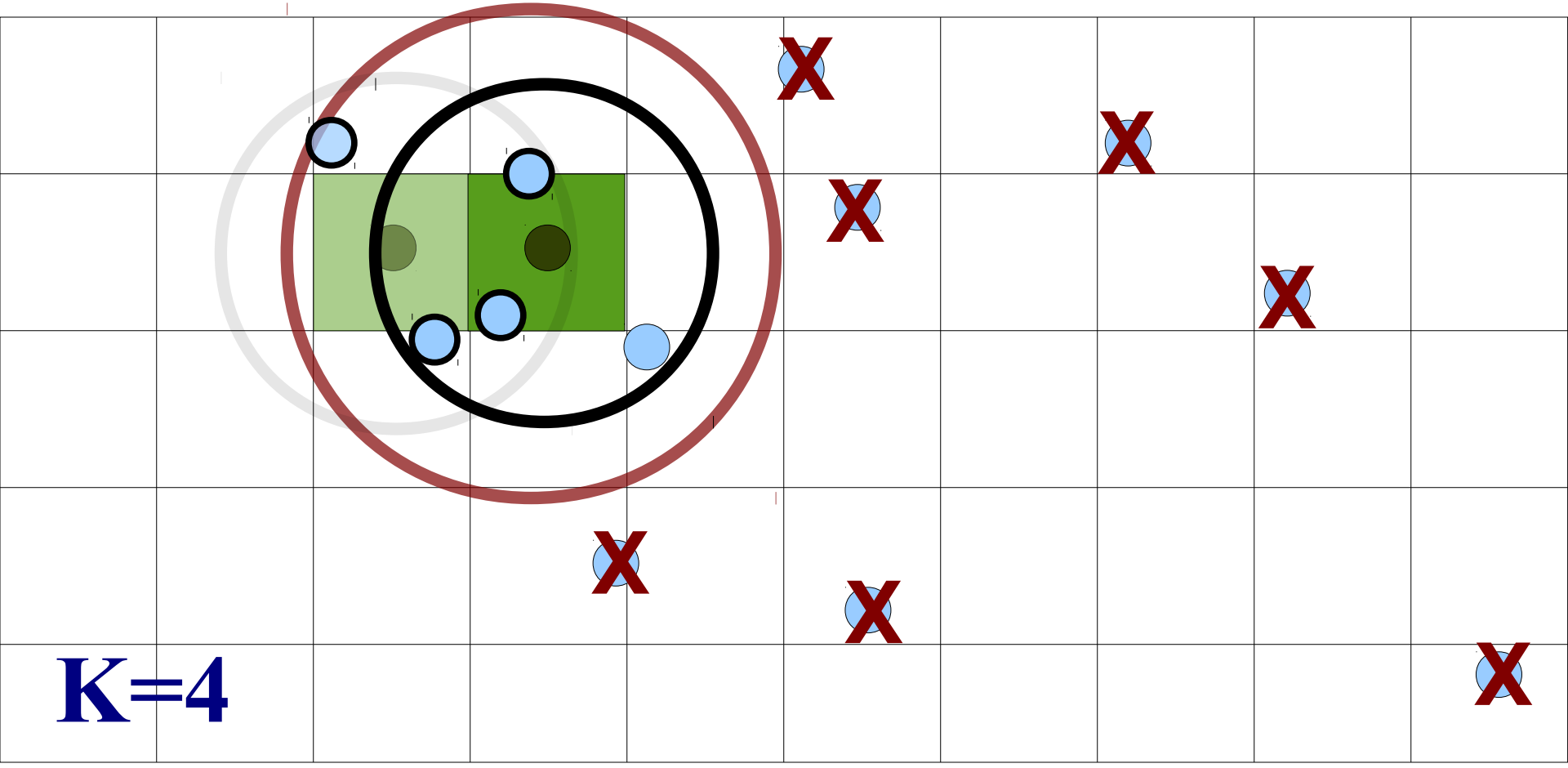
(Hodgson, 1989)

K Nearest Neighbor Search – Learned Search



(Hodgson, 1989)

K Nearest Neighbor Search – Learned Search



(Hodgson, 1989)

Parallelization of Spatial Methods

- No data dependencies among cells
  - Embarrassingly parallel problem
- However, spatial characteristics among neighboring cells can be used to improve *k*NN search speed through spatial pruning
  - Learned search techniques
- These "soft" data dependencies should be taken into account while parallelizing methods