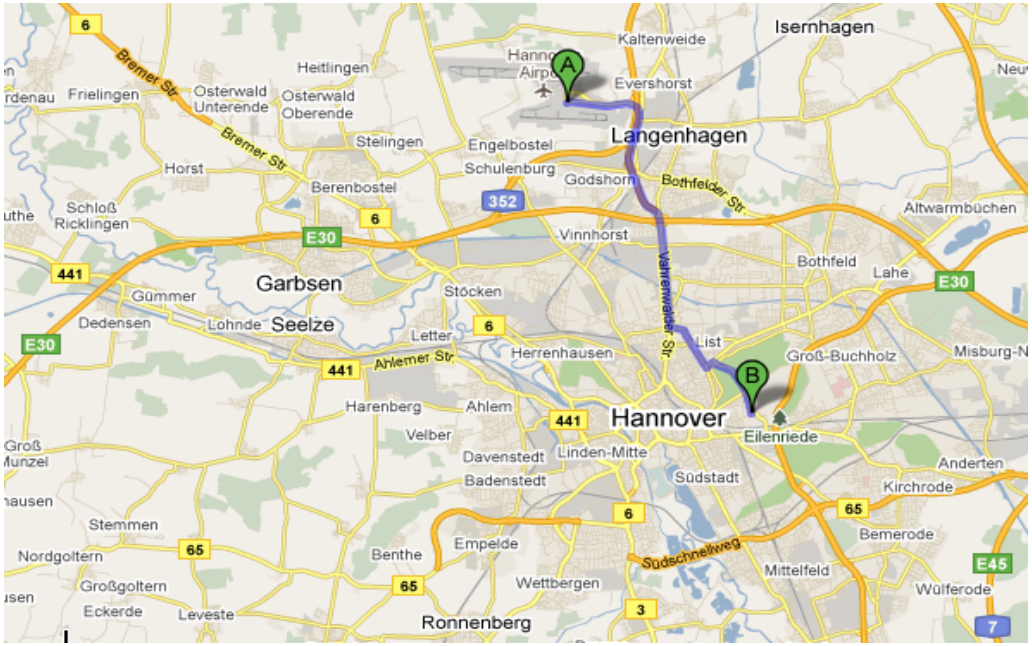
Jackpine: A Benchmark To Evaluate Spatial Database Performance

Suprio Ray, Bogdan Simion, Angela Demke Brown
Computer Science Department
University of Toronto
Toronto, Canada

ICDE 2011, Hannover, Germany

Motivation

- Explosion of spatial data
- Web-mapping and Location-based services are very popular



- How do we evaluate spatial database performance?

TPC-Spatial?

- Where is TPC-Spatial?
- No industry standard spatial database benchmark
- Related work
  - Sequoia 2000 (SIGMOD '93)
- Jackpine - a database benchmark for spatial workloads

Goals for the spatial benchmark

- Comprehensive coverage of spatial features
- Real-world workloads
- Extensible
- Portable

Overview

- Motivation
- Background on spatial database
- Micro benchmark
- Macro benchmark
- Challenges
- Using the benchmark
- Conclusion

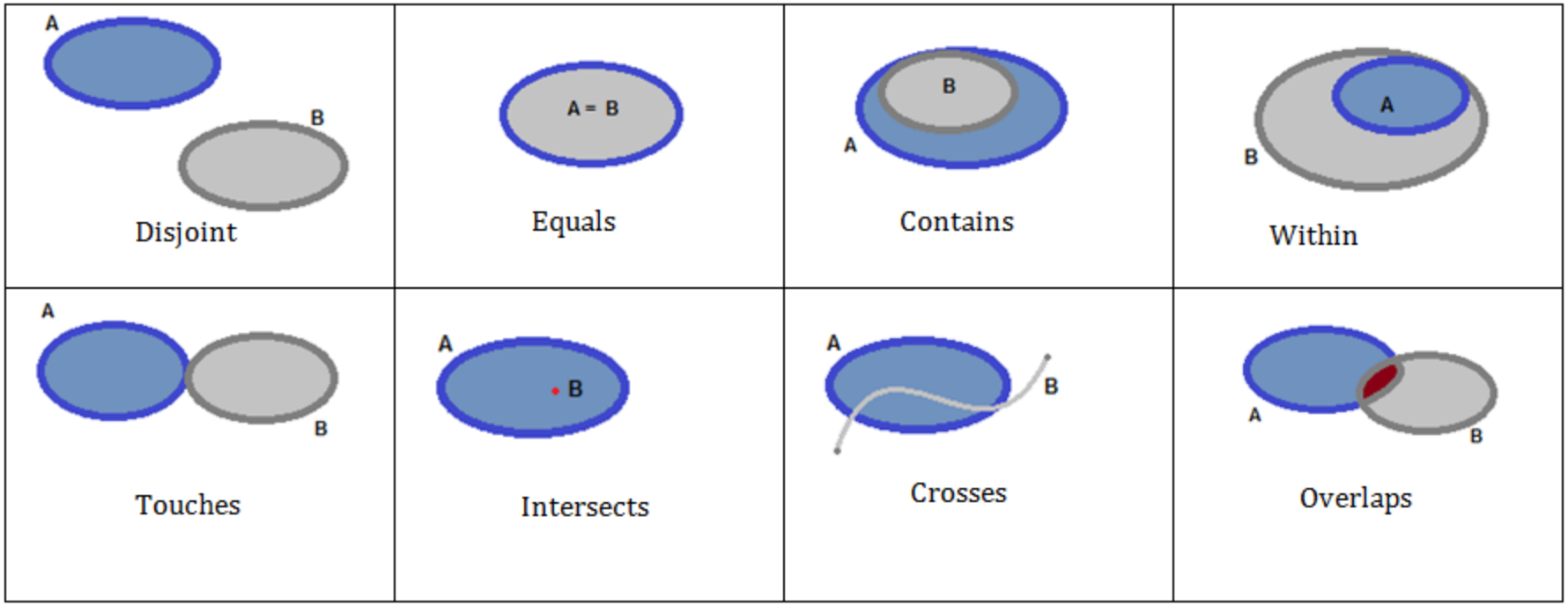
Spatial query execution

- Spatial query
  - allow for the use of geometry data types
    - points, lines and polygons
  - topological relations
    - Intersects, touches, overlaps

- Two step query evaluation process
  - Filter
  - Refine

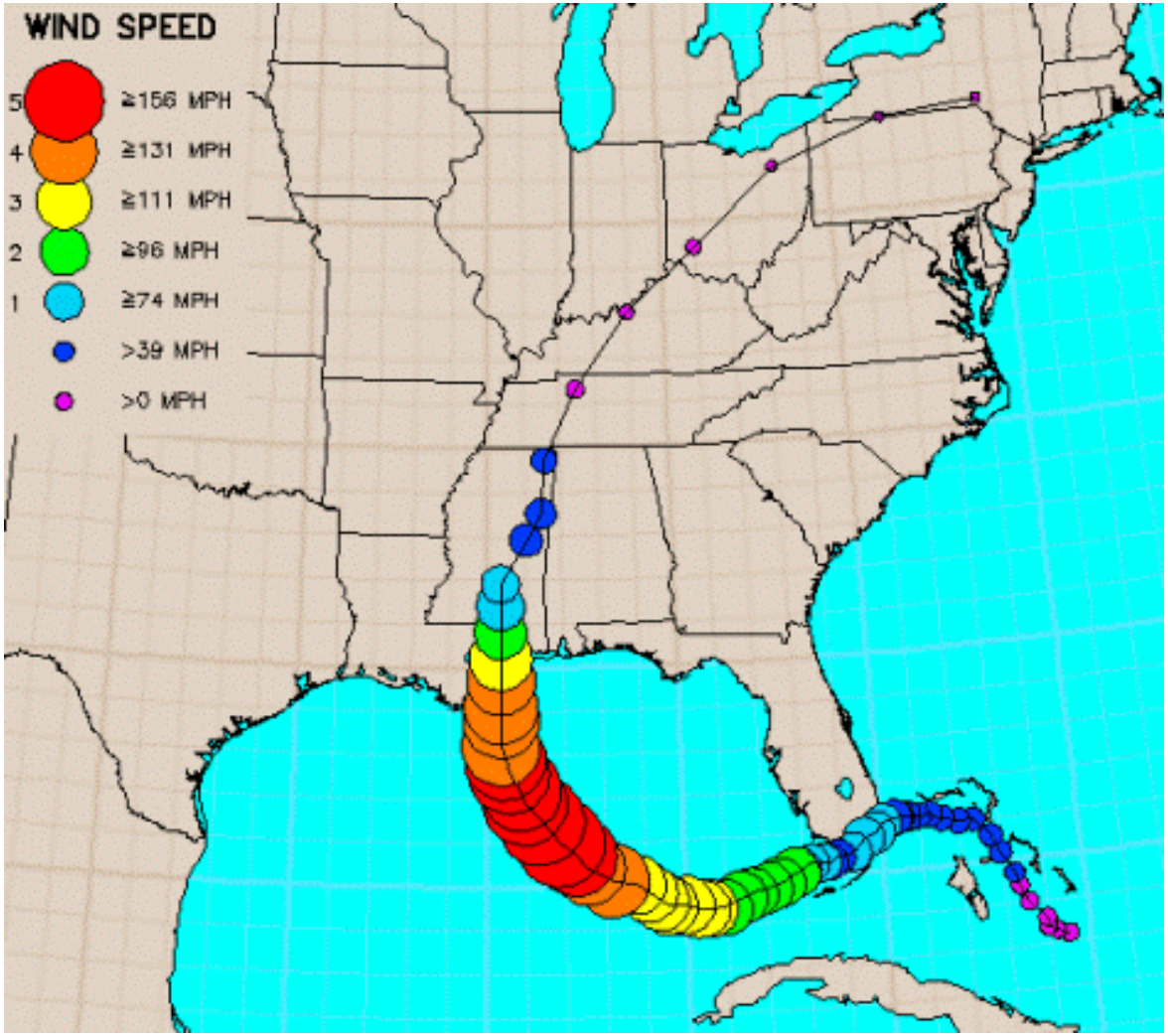
Topological relations

- Dimensionally Extended Nine-intersection model
  - Considers the max dimension of the intersections of two objects
  - Proposes 8 relations:



Filter and Refine

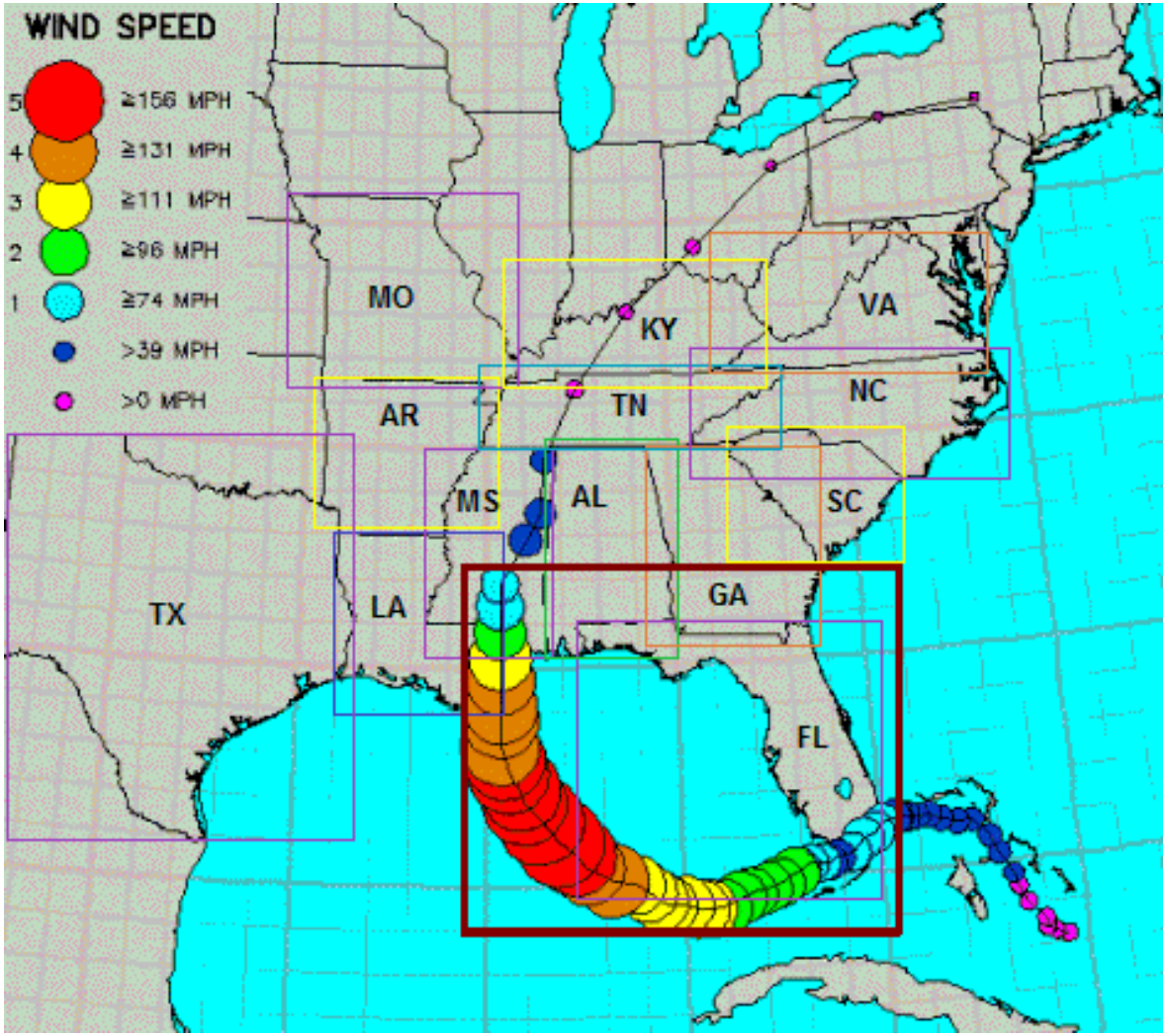
- Spatial query: which US states were affected by Katrina



Filter and Refine

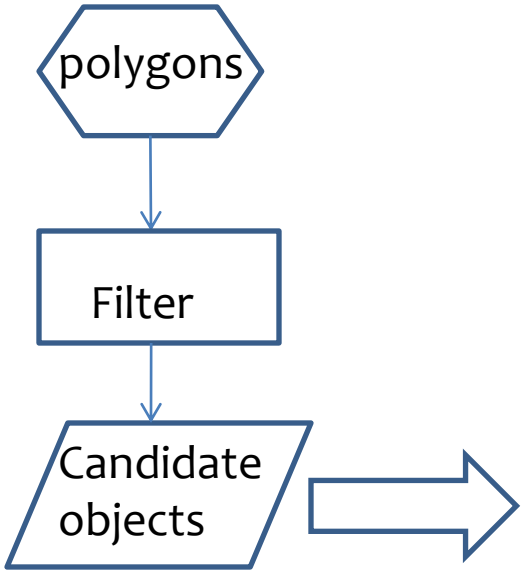
- Spatial query: which US states were affected by Katrina

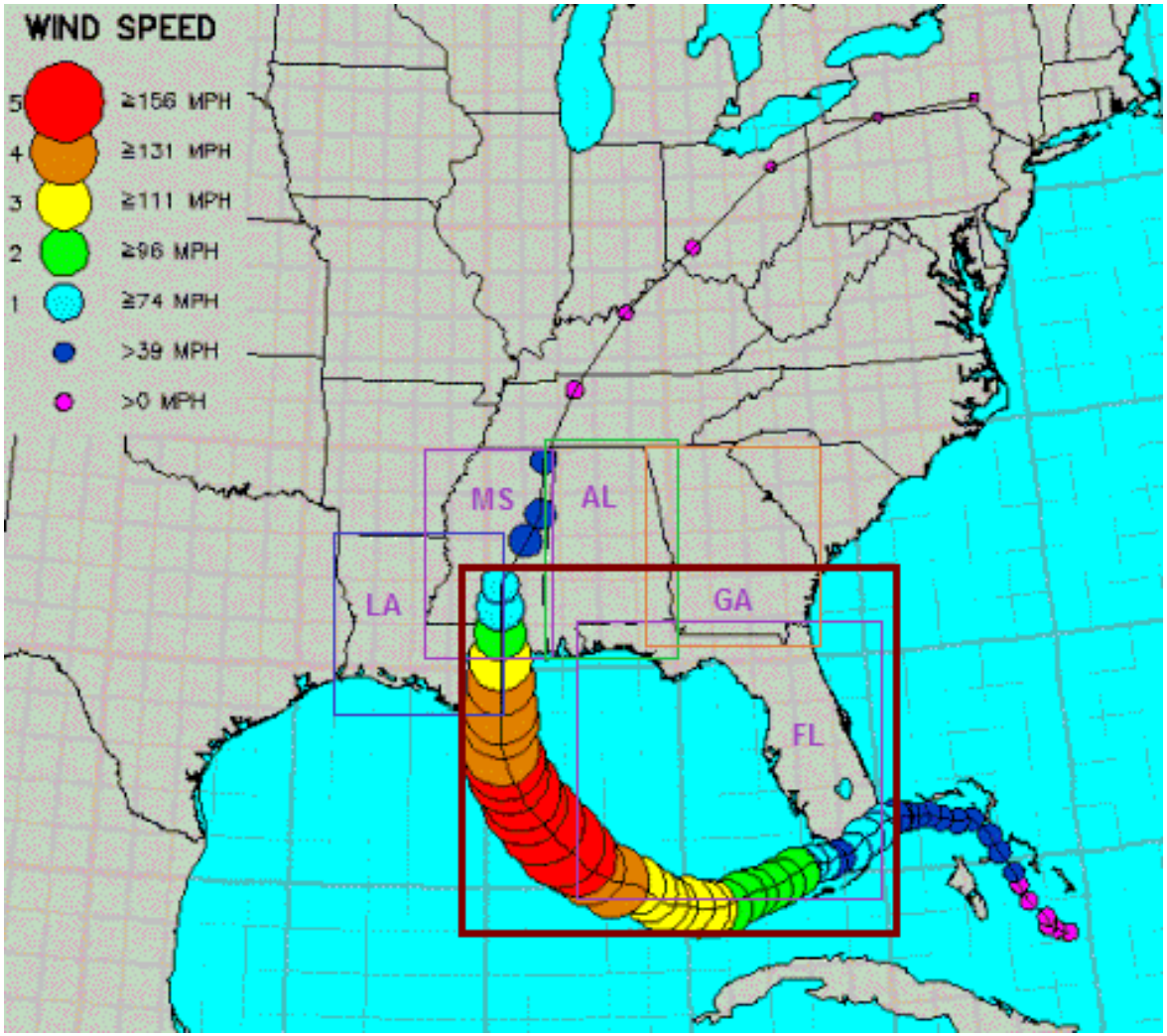
polygons



Filter and Refine

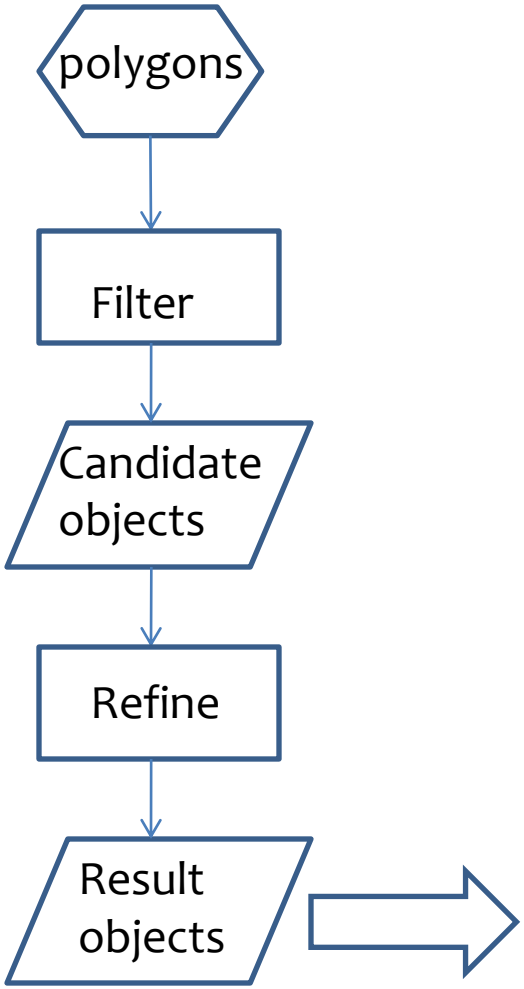
- Spatial query: which US states were affected by Katrina

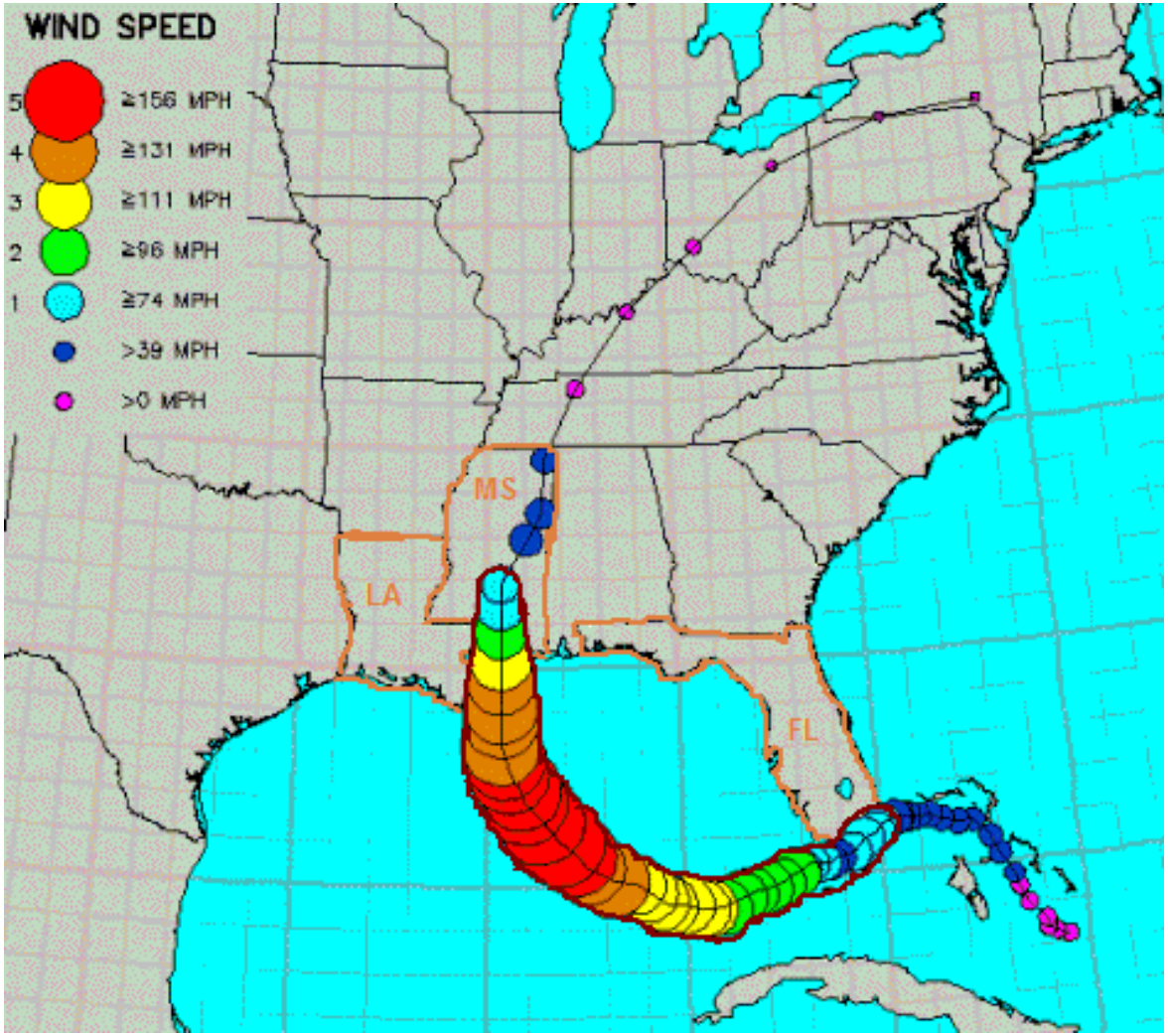




Filter and Refine

- Spatial query: which US states were affected by Katrina





Overview

- Motivation
- Background on spatial database
- Micro benchmark
- Macro benchmark
- Challenges
- Using the benchmark
- Conclusion

Jackpine

- Micro benchmark

  Goal: Comprehensive coverage of spatial features

  - Spatial join queries involving topological relations (based on Dimensionally Extended 9 Intersection model)
  - Queries with spatial analysis and aggregation functions
  - Inserting records with geometric objects

- Macro benchmark

  Goal: Model real-world spatial applications

  - Six spatial applications

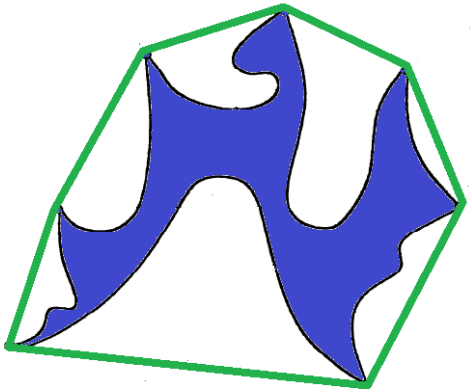
Jackpine - Micro benchmark

- Spatial join queries
  - Each row and each column is included at least once
  - Total number of queries 15

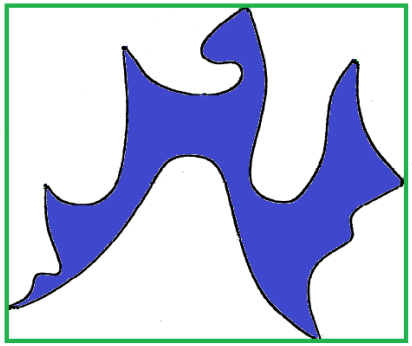
	Polygon and Polygon	Line and Line	Line and Polygon	Point and Polygon	Point and Line	Point and Point
Equals	✓		X	X	X	✓
Disjoint	✓					
Intersects			✓	✓	✓	
Touches	✓		✓			X
Crosses	X	✓	✓	X	X	X
Overlaps	✓		X	X	X	X
Within	✓		✓	✓		X
Contains	✓					X

Micro benchmark

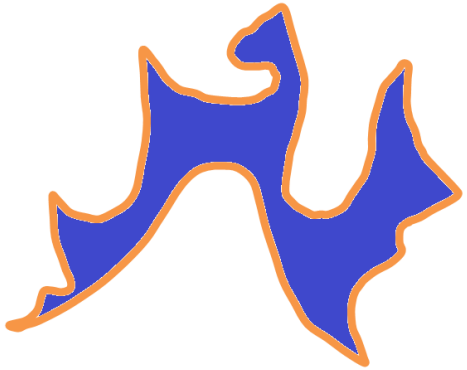
- Spatial analysis - analytic functions
  - distance, dimension, envelope, buffer, convex hull



convex hull



envelope



buffer

- Spatial analysis - aggregation operations
  - longest line, largest area, total length and area

Overview

- Motivation
- Background on spatial database
- Micro benchmark
- Macro benchmark
- Challenges
- Using the benchmark
- Conclusion

Jackpine - Macro benchmark

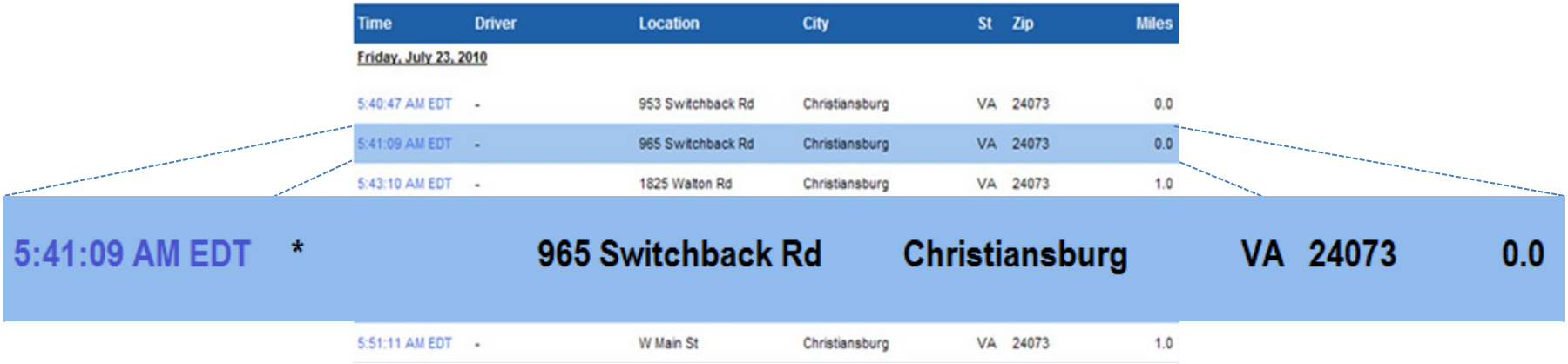
- Attempts to model real-world applications
- Scenarios: consist of a series of queries executed in sequence
  - Geocoding
  - Reverse Geocoding
  - Map search and browsing
  - Land Information Management
  - Flood risk analysis
  - Toxic spill

Macro benchmark - Geocoding

- Determine latitude, longitude from street address
- Common use case is locating addresses of people and organizations in the map
- Scenario queries
  - Finds the matching street segment given street address or zipcode
  - Latitude, longitude of the location is obtained from the street segment

Macro benchmark – Reverse Geocoding

- Obtain textual street address from latitude, longitude
- Common use case is producing trip activity report in location based services



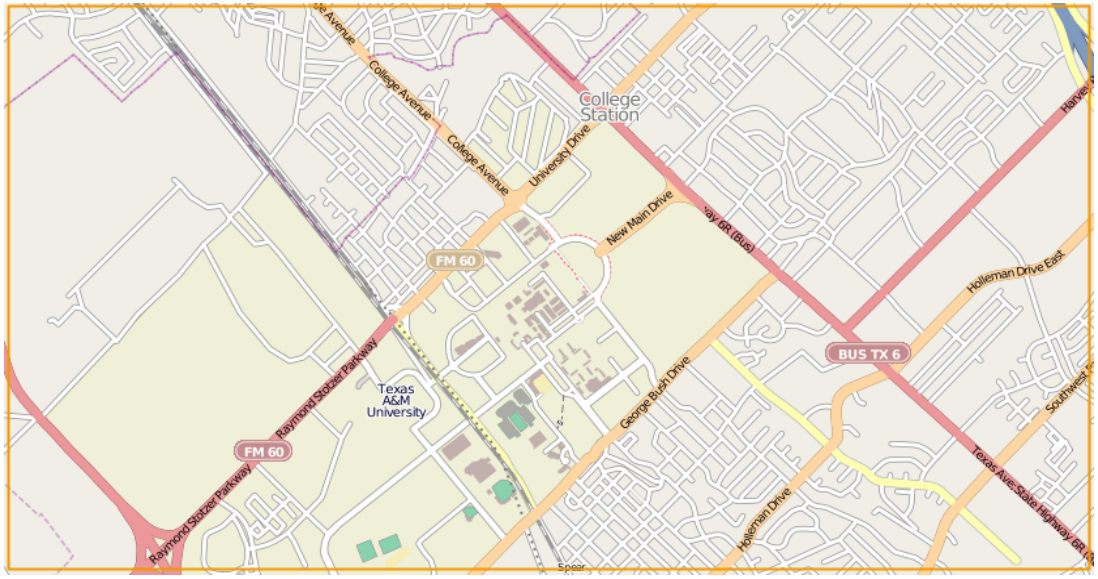
Time	Driver	Location	City	St	Zip	Miles
Friday, July 23, 2010						
5:40:47 AM EDT	-	953 Switchback Rd	Christiansburg	VA	24073	0.0
5:41:09 AM EDT	-	965 Switchback Rd	Christiansburg	VA	24073	0.0
5:43:10 AM EDT	-	1825 Walton Rd	Christiansburg	VA	24073	1.0
5:41:09 AM EDT	*	965 Switchback Rd	Christiansburg	VA	24073	0.0
5:51:11 AM EDT	-	W Main St	Christiansburg	VA	24073	1.0

source: tracknet

- Scenario queries
  - Find the closest city name
  - Find the closest street name

Macro benchmark – Map search & browsing

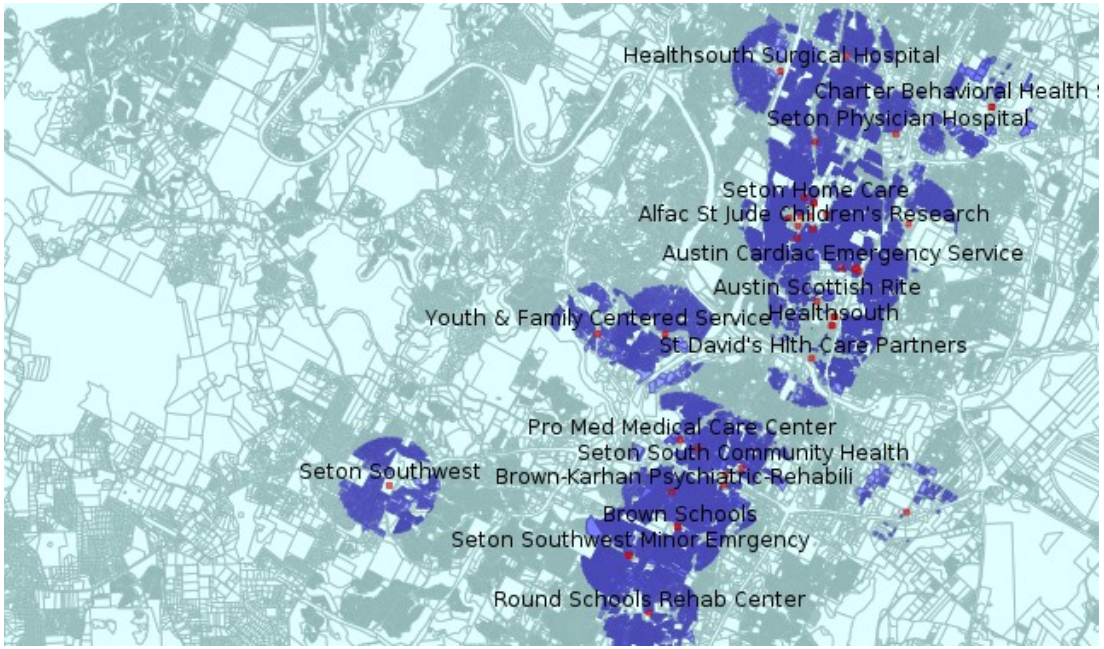
- Simulates searches for points of interest in the map
- Simulates the display of the bounding box around POI in the map
- Two visit cases: Student visit to a campus & Tourist visit to a place



- Scenario queries
  - search by matching keyword
  - series of queries to fetch objects inside the bounding box

Macro benchmark – Land Info Management

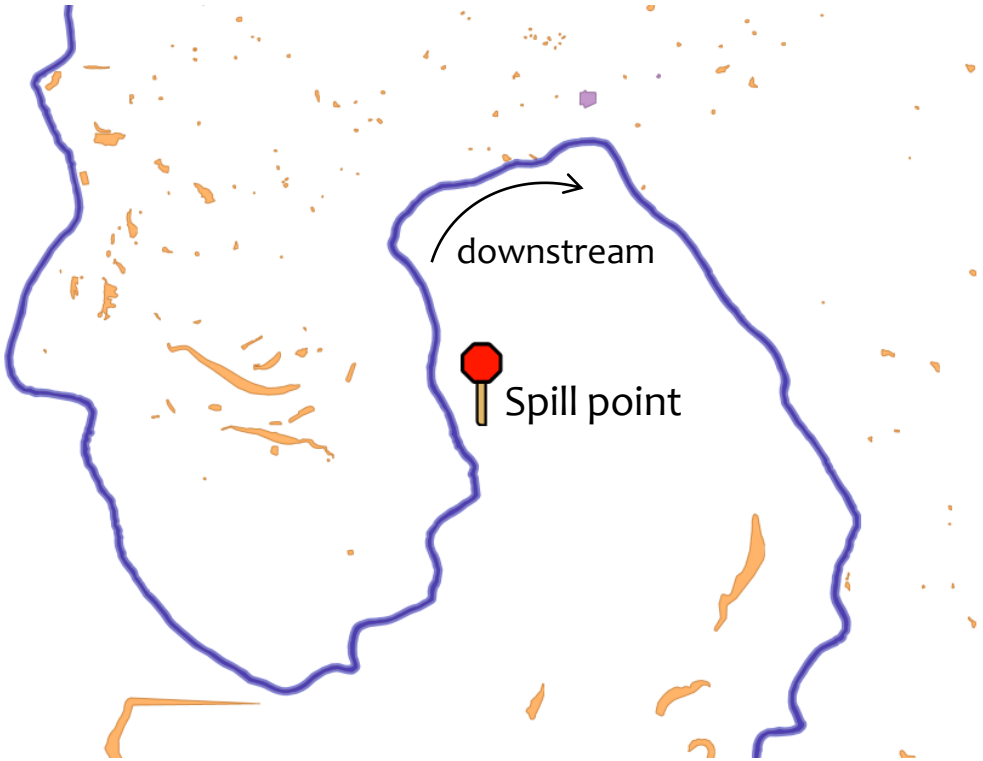
- Maintain precise land parcel location and ownership info
- Used for tax assessment, valuation and mortgage



- Six queries including
  - Avg. value of single family residential properties
  - Commercial properties on unpermitted landfills

Macro benchmark – Toxic spill

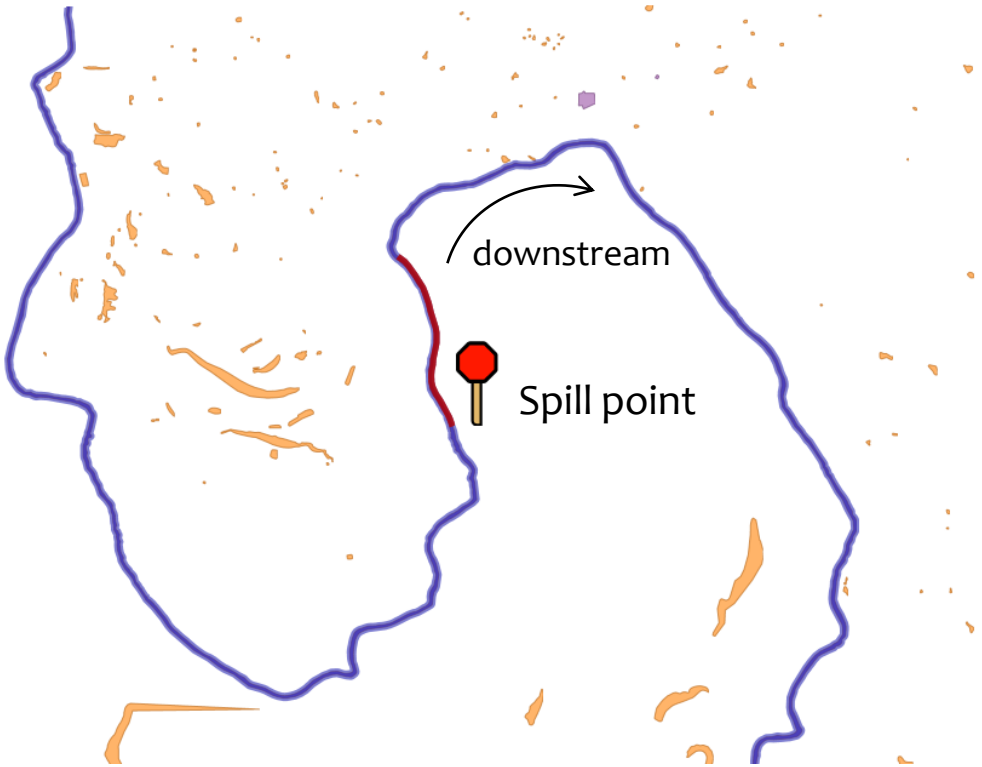
- Toxic chemicals spilled in a waterway may spread miles
- Recursively determines all downstream segments from initial spill point



- Two queries including
  - If spill point is on any waterway segment

Macro benchmark – Toxic spill

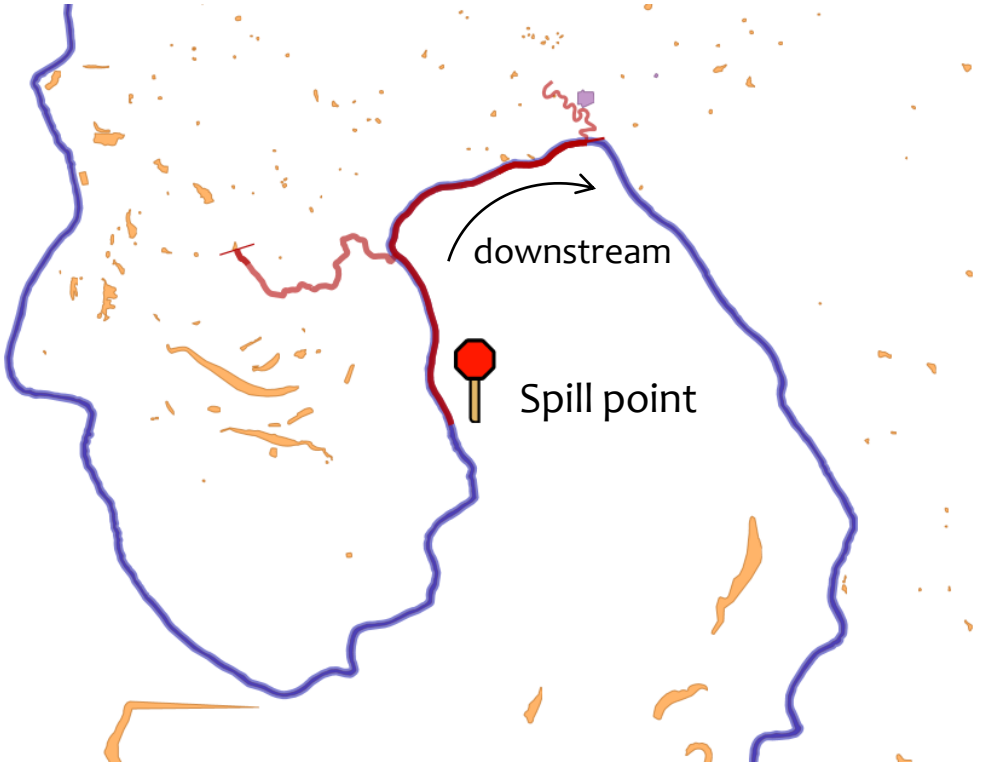
- Toxic chemicals spilled in a waterway may spread miles
- Recursively determines all downstream segments from initial spill point



- Two queries
  - If spill point is on any waterway segment
  - Waterway segments within 20 mile downstream of spill point

Macro benchmark – Toxic spill

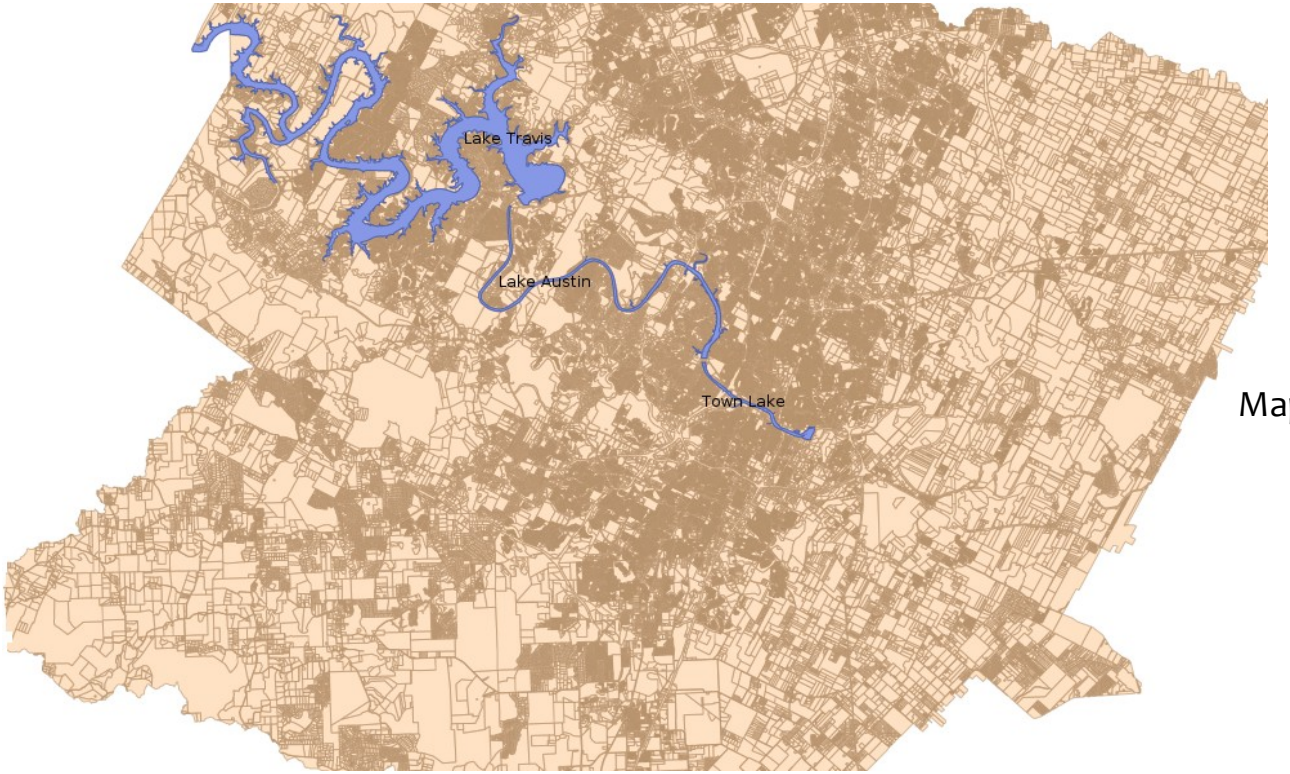
- Toxic chemicals spilled in a waterway may spread miles
- Recursively determines all downstream segments from initial spill point



- Two queries
  - If spill point is on any waterway segment
  - Waterway segments within 20 mile downstream of spill point

Macro benchmark – Flood Risk Analysis

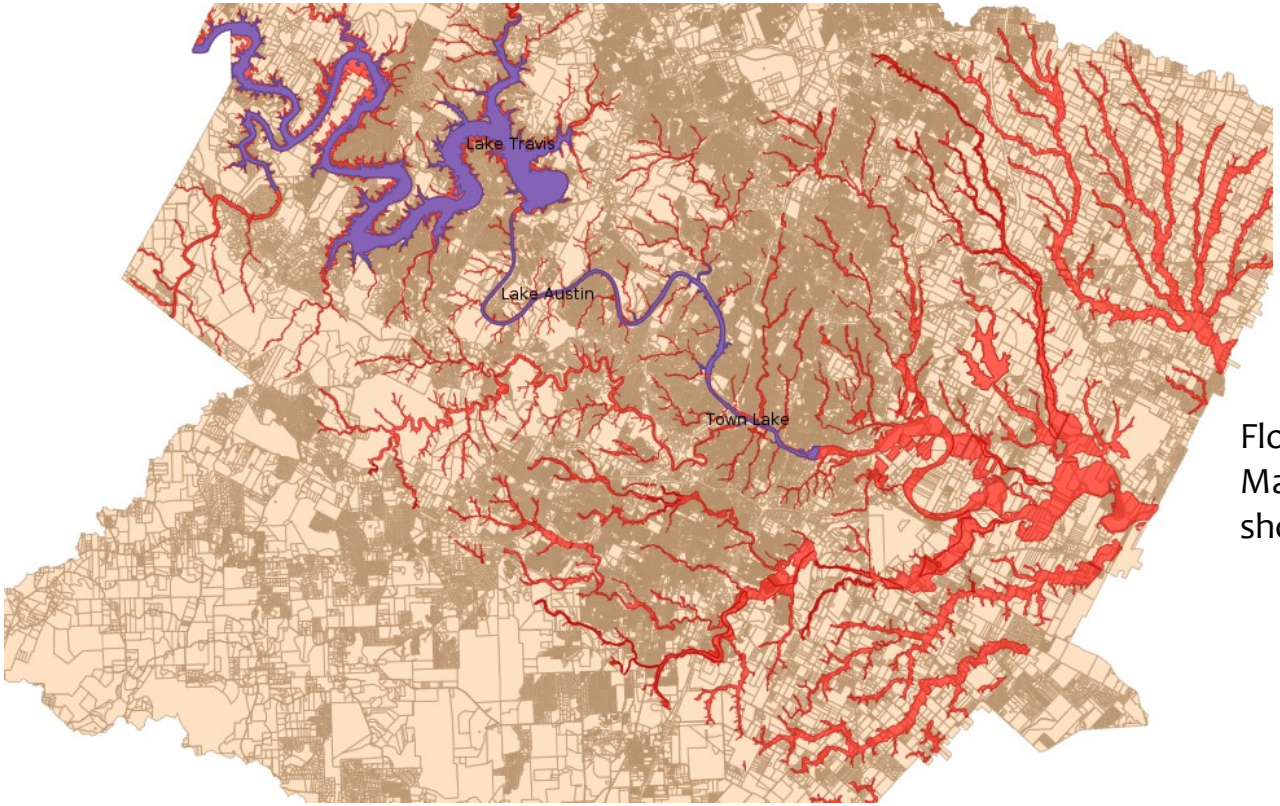
- Flood Insurance Rate Map depicts flood hazard areas
- DFIRM database is used to determine flood insurance rate



Map of Travis County, TX

Macro benchmark – Flood Risk Analysis

- Flood Insurance Rate Map depicts flood hazard areas
- DFIRM database is used to determine flood insurance rate



Flood Insurance Rate Map of Travis County, TX showing high risk areas

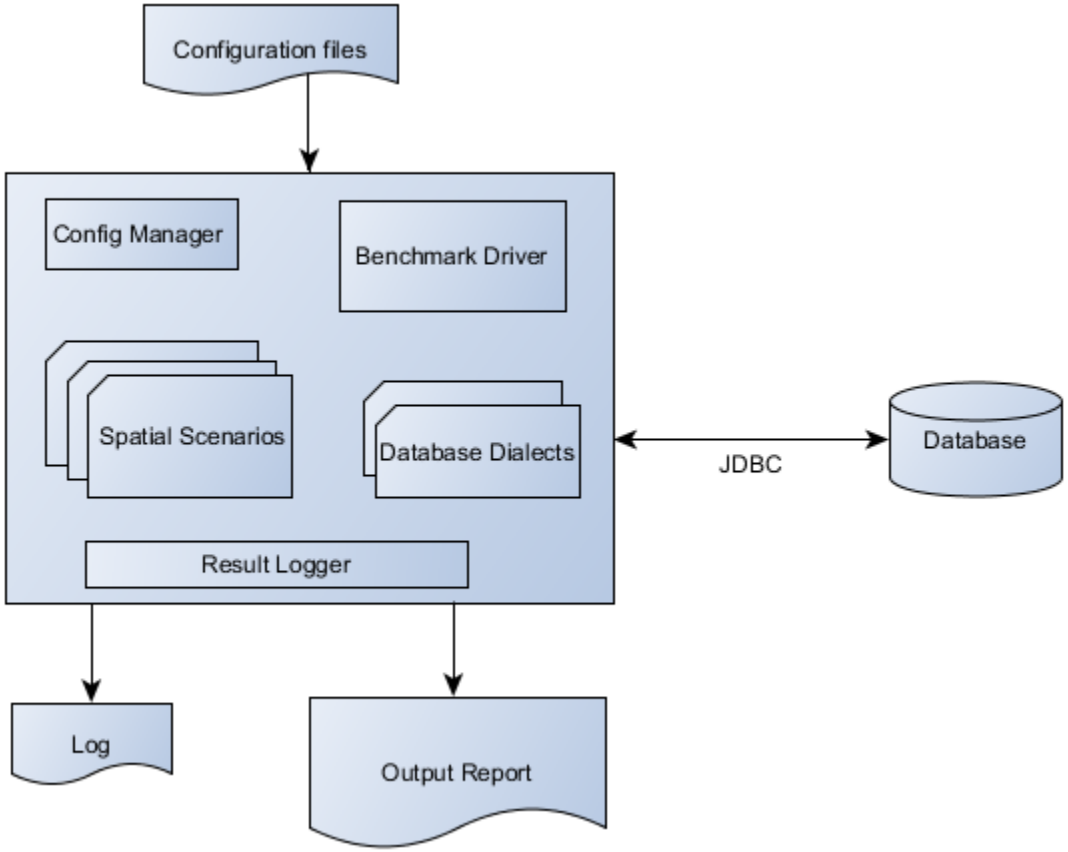
- Four queries including
  - Residential property owners required to carry flood insurance
  - Industrial complexes in high risk areas

Goals for the spatial benchmark

- Comprehensive coverage of spatial features 
- Real-world workloads 
- Extensible 
- Portable 

Implementation

- Spatial Scenario: initialization, invocation of the queries and cleanup
- Database Dialect: contain actual SQL queries



Implementation

- Supported runtime parameters
  - number of warm-up runs
  - iterations
  - number of threads
  - database to run against

- Output report

Avg Duration	Avg Ops Sec	Count Of Resultset	Duration	Iterations	Other Exceptions	SQL Exceptions	Warmup Count	Warmup Duration	DBType
0.236	4.237288	310.0	0.708	3.0	0.0	0.0	1.0	0.972	MySQL 5.0.91
0.945333	1.057827	80.0	2.836	3.0	0.0	0.0	1.0	2.415	PostgreSQL 8.4.2
1.954666	0.511596	80.0	5.864	3.0	0.0	0.0	1.0	2.759	Informix

Experimental setup

- Database
  - MySQL 5.0.91, PostgreSQL 8.4.2 and Informix 11.50
- Dataset – TIGER® for Texas and Travis County dataset

	MySQL
Total data set size (GB)	4.6
Number of tables	15
Size of the largest data table (MB)	1651.6
Size of the largest table index (MB)	416.9
Cardinality of the largest table	5,895,060

- Machine – Pentium 4 CPU, 512 MB RAM, 240 GB disk

Challenges and some observations

- Database idiosyncrasies
  - MySQL: no refine step, only filter step
  - MySQL: table order matters
- OGC standards compliance
  - Some spatial functions unsupported
- Configuration issues
- Tuning

Challenges and some observations

- Query execution phase
  - MySQL performs MBR-based filter, but does not execute the refinement phase

Area and Area queries: # of records returned

	Intersects	Contains	Equals	Overlaps	Touches	Within
MySQL	7141	6246	5973	573	37	6246
PostgreSQL	6621	6019	5971	22	532	6019
Informix	6621	6018	5971	23	532	6018

Challenges and some observations

- Query execution phase
  - MySQL performs MBR-based filer, but does not execute the refinement phase

	Intersects	Contains	Equals	Overlaps	Touches	Within
MySQL	7141	6246	5973	573	37	6246
PostgreSQL	6621	6019	5971	22	532	6019
Informix	6621	6018	5971	23	532	6018

Area and Area queries: # of records returned

Actual geometries touch

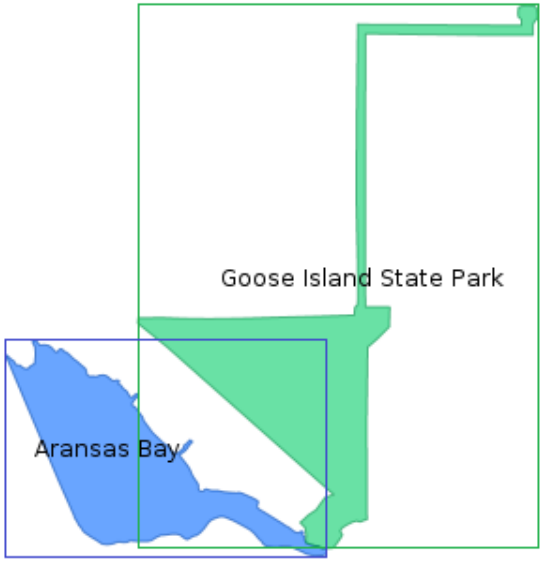
Challenges and some observations

- Query execution phase
  - MySQL performs MBR-based filer, but does not execute the refinement phase

	Intersects	Contains	Equals	Overlaps	Touches	Within
MySQL	7141	6246	5973	573	37	6246
PostgreSQL	6621	6019	5971	22	532	6019
Informix	6621	6018	5971	23	532	6018

Area and Area queries: # of records returned

But, MBRs overlap



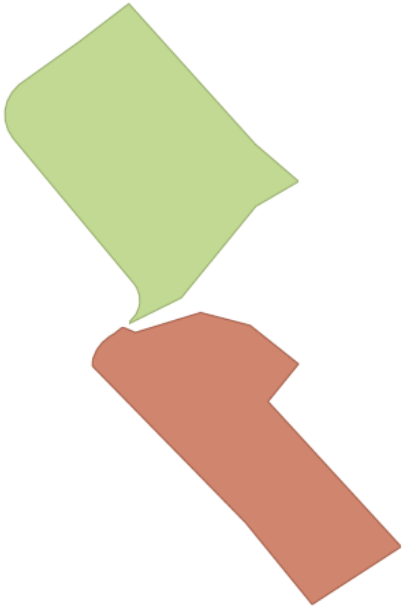
Challenges and some observations

- Query execution phase
  - MySQL performs MBR-based filer, but does not execute the refinement phase

	Intersects	Contains	Equals	Overlaps	Touches	Within
MySQL	7141	6246	5973	573	37	6246
PostgreSQL	6621	6019	5971	22	532	6019
Informix	6621	6018	5971	23	532	6018

Area and Area queries: # of records returned

Actual geometries are disjoint



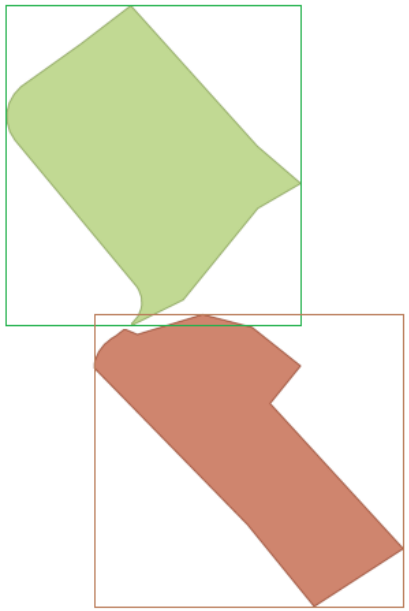
Challenges and some observations

- Query execution phase
  - MySQL performs MBR-based filer, but does not execute the refinement phase

	Intersects	Contains	Equals	Overlaps	Touches	Within
MySQL	7141	6246	5973	573	37	6246
PostgreSQL	6621	6019	5971	22	532	6019
Informix	6621	6018	5971	23	532	6018

Area and Area queries: # of records returned

But, MBRs overlap



Challenges and some observations

- Support for spatial functions not complete
  - MySQL does not support Distance, Dwithin, Buffer, ConvexHull
  - Informix does not support Dwithin
- Simulated functions
  - MySQL: Distance, Dwithin
  - Informix: Dwithin
- Runtime issues
  - Informix had runtime issues with Buffer, StartPoint, EndPoint

Challenges and some observations

- Table order in the spatial predicate is important for MySQL

- select count(*) from arealm a, edges e where

	intersects(e.shape, a.shape)	intersects(a.shape, e.shape)
Use spatial index	X	✓
Exec time	23 hours	3 minutes

- select count(*) from arealm a, edges e where

	intersects(e.shape, a.shape) and a.ogr_fid= 3332	intersects(a.shape, e.shape) and a.ogr_fid= 3332
Use spatial index	✓	X
Exec time	0.08 seconds	16 minutes

- select count(*) from arealm a1, arealm a2 where

	intersects(a1.shape, a2.shape)	intersects(a2.shape, a1.shape)
Use spatial index	X	✓
Exec time	2 minutes	0.35 seconds

Challenges and some observations

- Table order in the spatial predicate doesn't matter for

- PostgreSQL

  select count(*) from arealm_merge a1, arealm_merge a2 where

	ST_Intersects(a1.the_geom, a2. the_geom)	ST_Intersects (a2. the_geom, a1. the_geom)
Use spatial index	✓	✓
Exec time (sec)	15.09	19.22

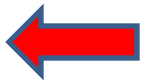
- Informix

  select count(*) from arealm_merge a1, arealm_merge a2 where

	ST_Intersects(a1.shape,a2.shape)	ST_Intersects (a2.shape,a1.shape)
Use spatial index	✓	✓
Exec time (sec)	3.2	3.1

Overview

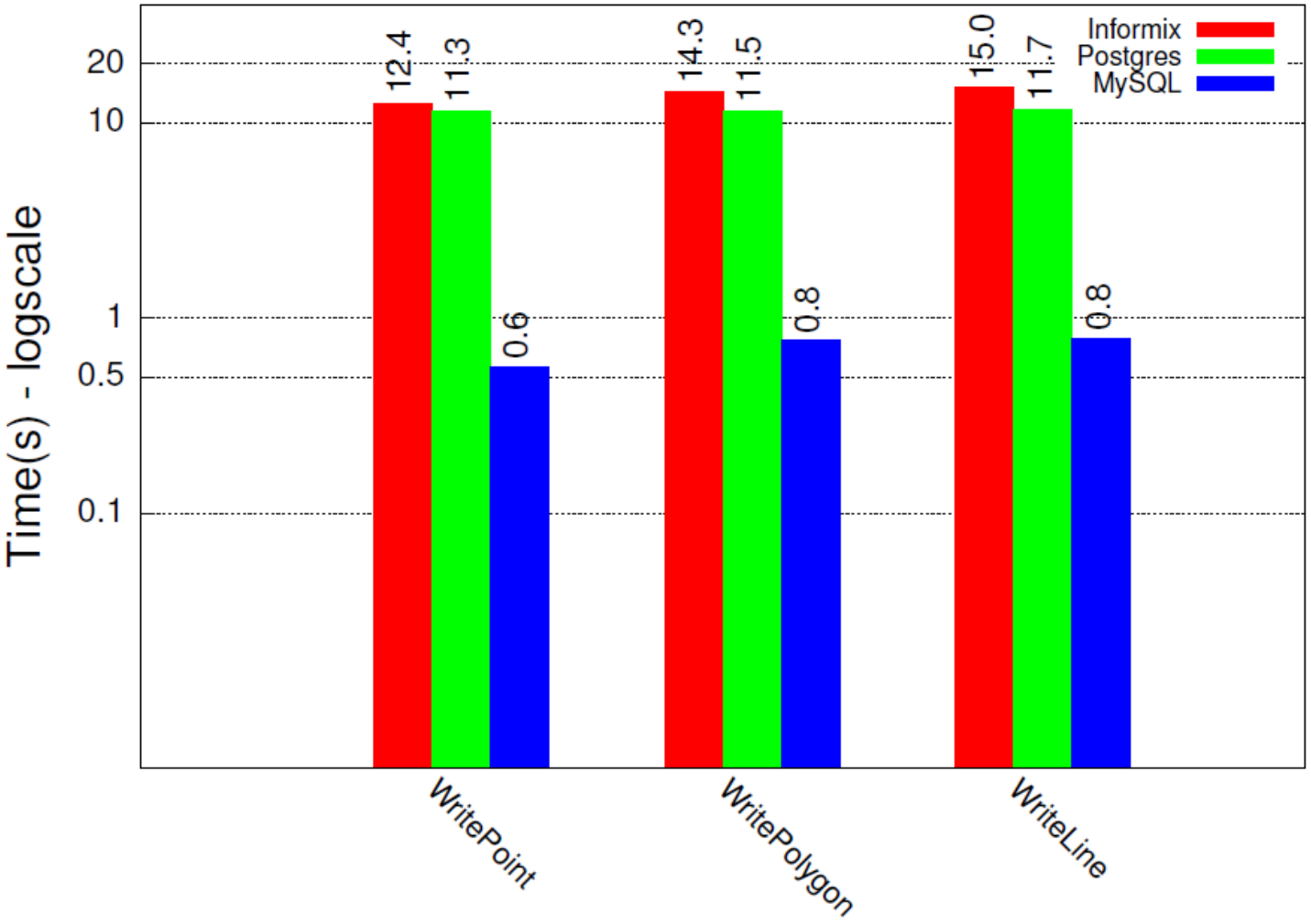
- Motivation
- Background on spatial database
- Micro benchmark
- Macro benchmark
- Challenges
- Using the benchmark
- Conclusion



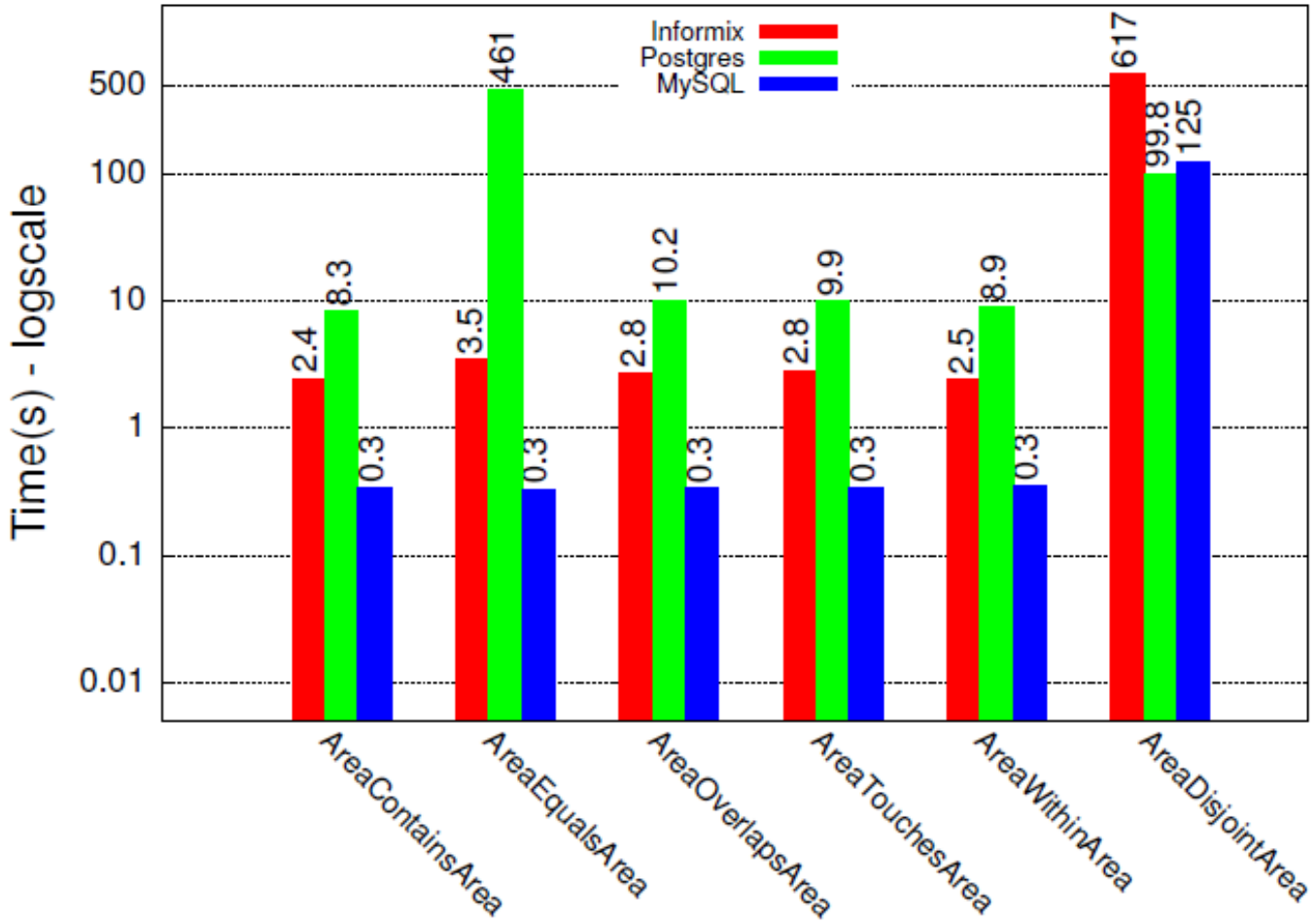
Using the benchmark

- Upload the dataset from shape files
- Configure
- Run the benchmark
- View results
- Example:
  - Compare with PostgreSQL, MySQL and Informix
    - Not a fair comparison - no refine step in MySQL

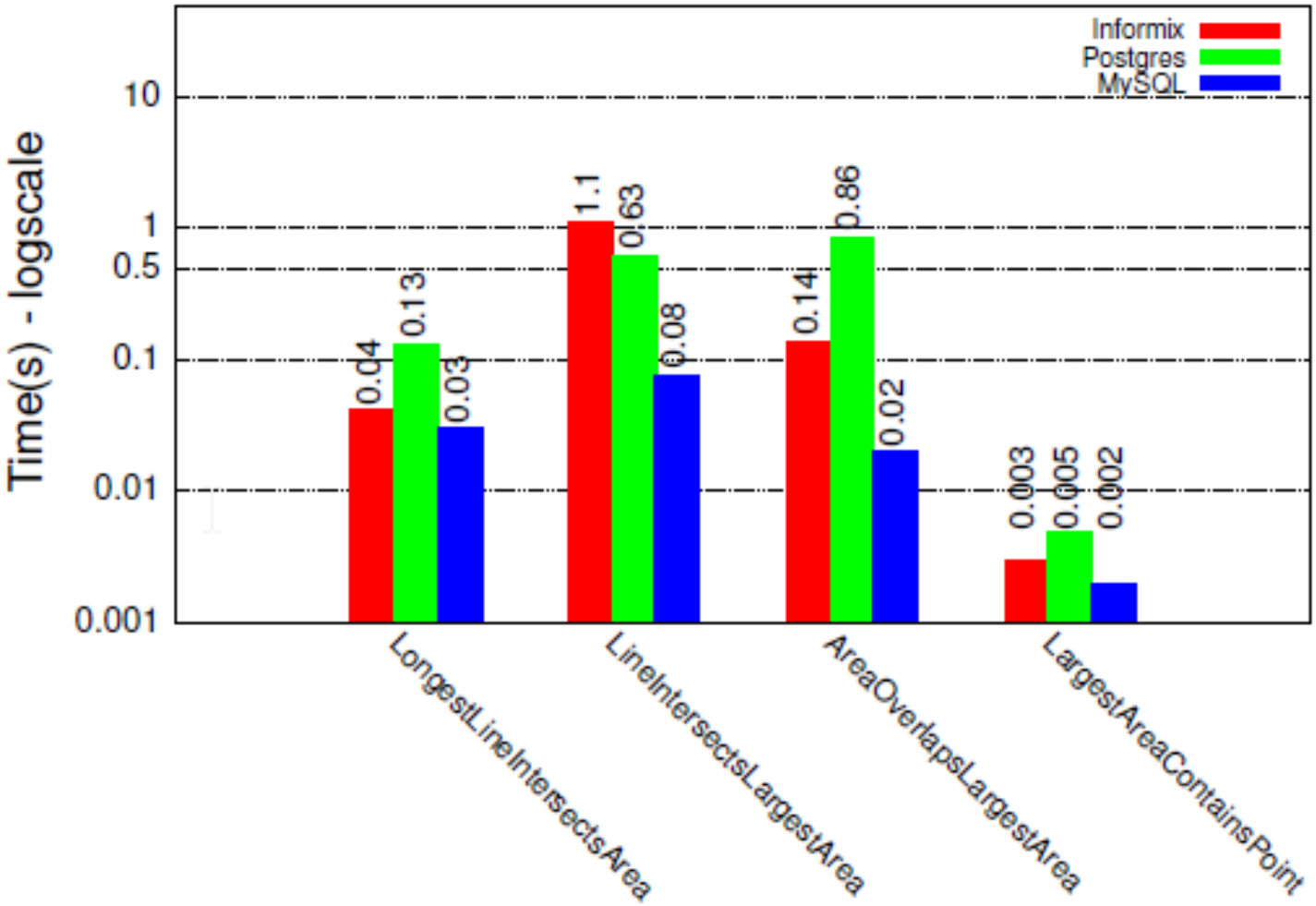
Results - Record insertion



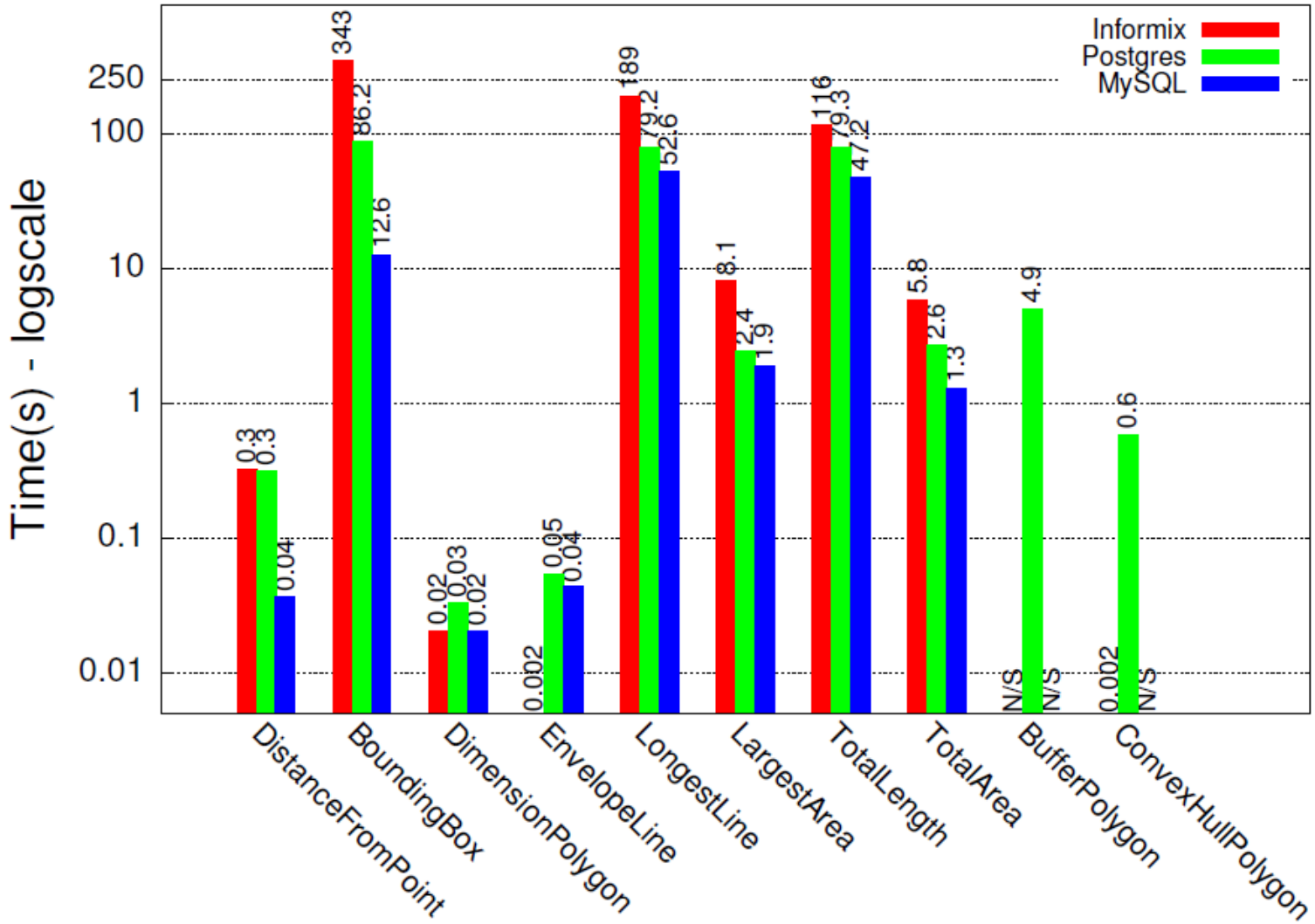
Results - Spatial join



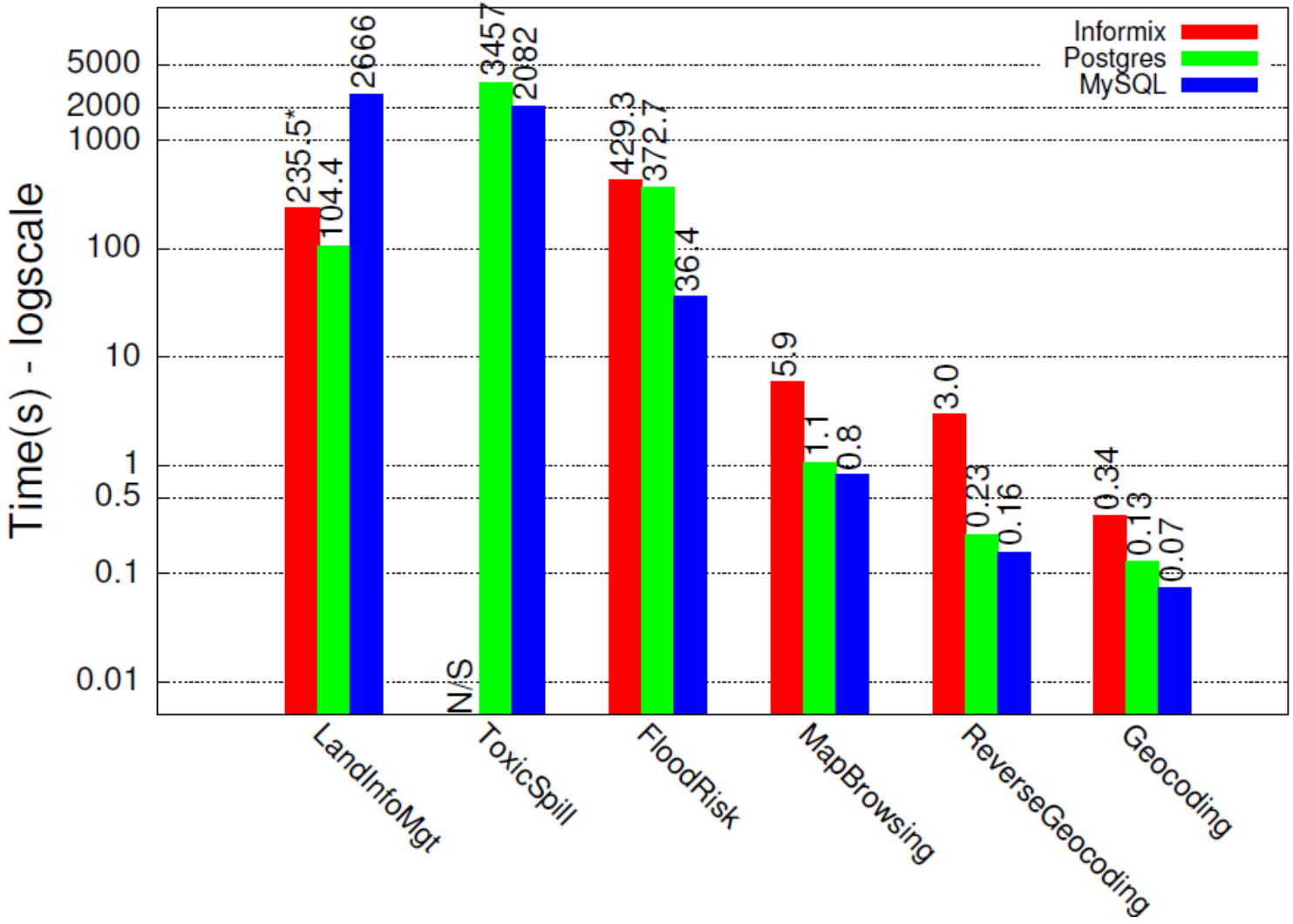
Results – Spatial join with an object



Results - Spatial analysis



Results - Macro benchmark



Score Metric

- Methodology
  - Geometric mean over all queries, each exec time is normalized over a reference DBMS
  - Metric

$$Score_{DBMS_X} = \sqrt[N]{\prod_{q=1}^{N} \frac{Time_q^{DBMS_{ref}}}{Time_q^{DBMS_X}}}$$

Score Metric

- Which queries to be included in the score?
  - Ideally all of them
  - Not all are supported by all databases

	Micro benchmark	Macro benchmark
MySQL	Incomplete	Incomplete
PostgreSQL	1.00	1.00
Informix	Incomplete	Incomplete

Score Metric

- Scores
  - Included only those queries supported by all databases

	Micro benchmark	Macro benchmark
MySQL	5.17	1.06
PostgreSQL	1.00	1.00
Informix	1.48	0.29

Conclusion

- Comprehensive coverage in the micro benchmark workloads ✓
- Macro benchmarks includes representative real-world applications ✓
- Portable ✓
- Extensible ✓
- Flexible ✓

Future work

- Dataset scalability
- Spatio-temporal workloads
- Invite suggestions from the community

Thank you!