BASIS TECHNOLOGY

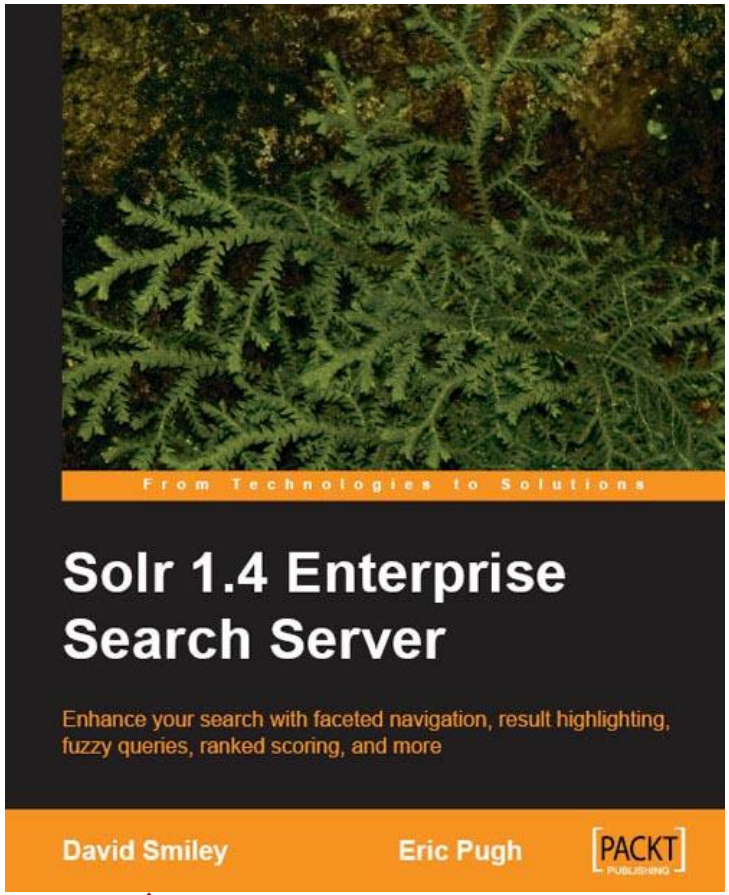
Open Source Search Conference

Lucene and Solr in Government

Geospatial Search Using Geohash Prefixes

David Smiley
Software Systems Engineer, Lead
The MITRE Corporation

My Background



- Solr book author
- Solr instructor
  - (at MITRE)
- At MITRE for 11+ years
  - Supporting internal apps and US DOD sponsors

A new edition is coming soon!

What I Will Cover

- The state and history of geospatial search in Solr
- All about geohashes
- Algorithms for searching geohashes
- The future: a new geo module: “LSP”

Geospatial in Solr 3.1

BASIS TECHNOLOGY

- LatLonType, PointType
  - Internally uses a pair of numeric fields for lat & lon
  - No multi-valued support
- GeoHashField
  - Multi-valued filter, but no sort
  - Doesn't scale well
- Queries:
  - Only point-radius; no bounding box*, polygon

Other Geospatial (1/2)

- LocalLucene, LocalSolr
  - *CartesianTier* concept
  - Abandoned
- Lucene's spatial contrib module
  - Transitioned from LocalLucene
  - Deprecated
- JTeam Spatial Plugin
  - Fork of Lucene's spatial contrib module
  - Supported; questionable future

Other Geospatial (2/2)

- MetaCarta GeoSearch Toolkit for Solr
  - Multi-valued support
  - Relevancy based on search keyword in-text proximity – ***cool!***
  - High memory requirements
  - Supported, not free
- ManTech's Brad Giaccio (attached to SOLR-773) for ICDL project
  - Multi-valued support
  - A good start

Geohashes

- What is a Geohash?
  - A lat/lon geocode system
  - Has a hierarchical spatial structure
  - Gradual precision degradation
  - In the public domain

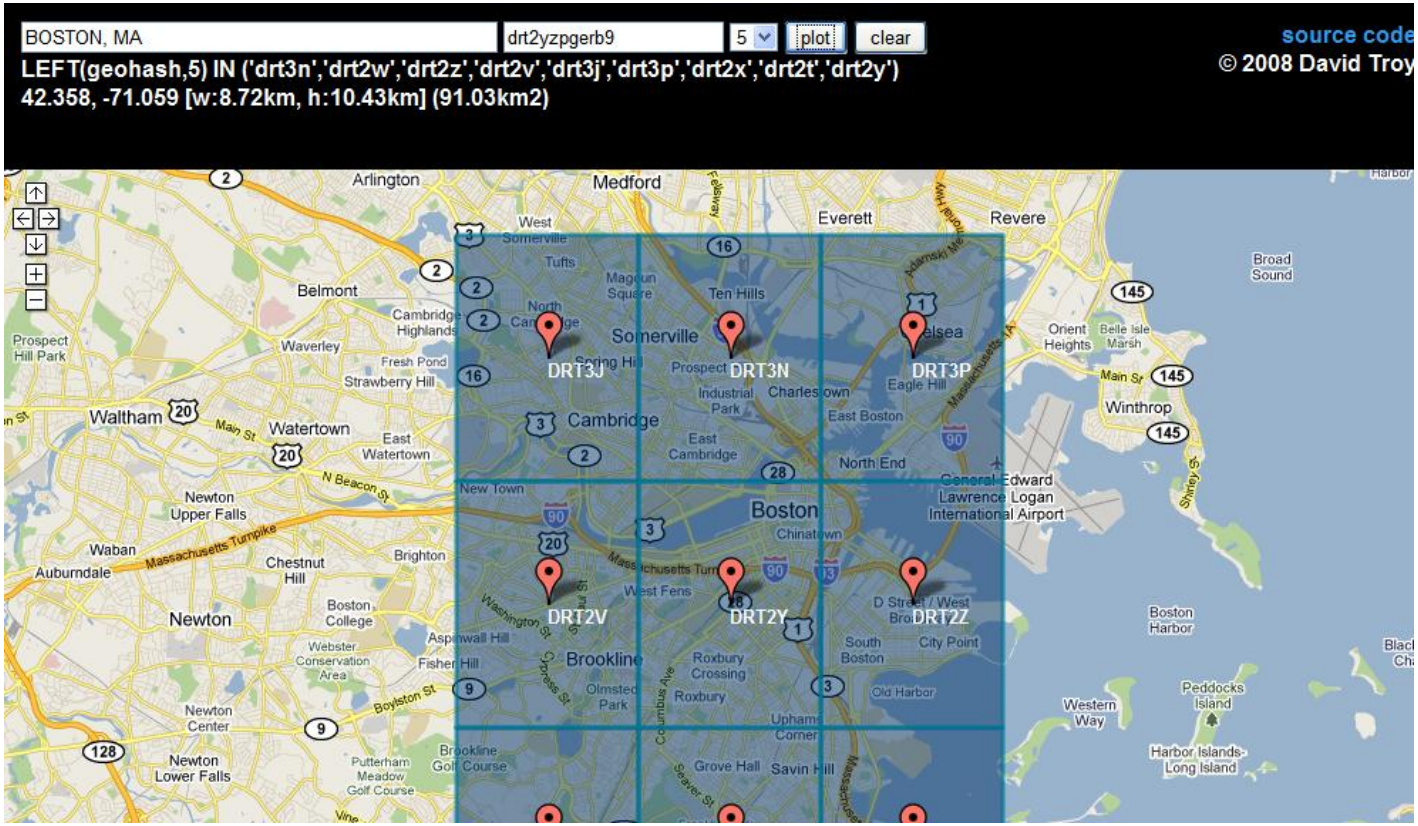
  http://en.wikipedia.org/wiki/Geohash

- Example: (Boston) DRT2Y

Demo

BASIS TECHNOLOGY

http://openlocation.org/geohash/geohash-js/

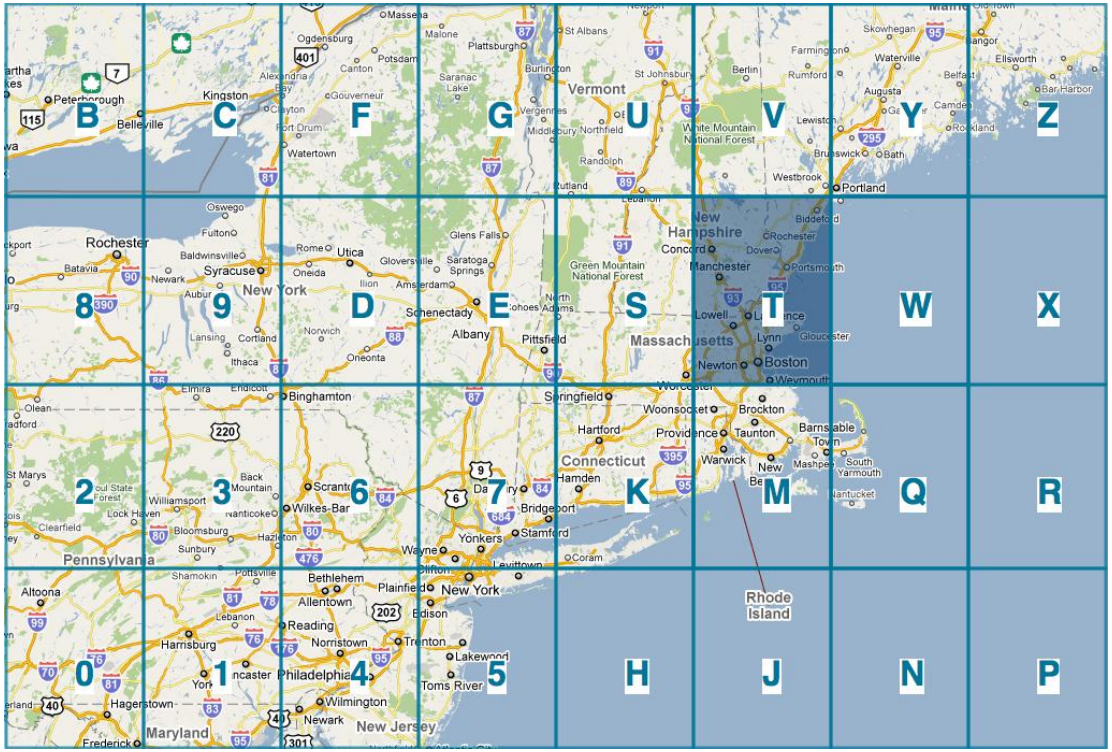


Zooming In: D

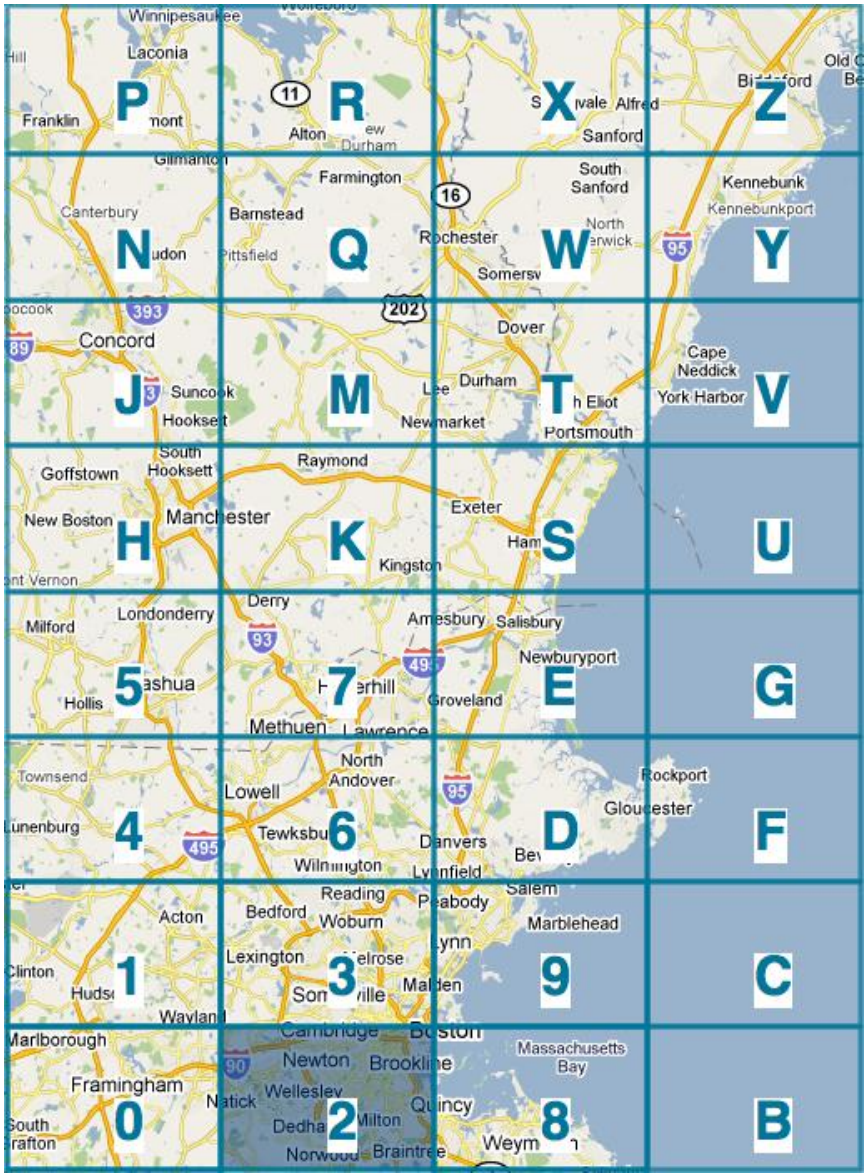

B
C
F
G
U
V
Y
Asia
Europe
8
9
D
E
S
T
W
South America
Indian
2
3
6
7
K
M
Q
0
1
4
5
H
J
N
Antarctica

Zooming In: DR

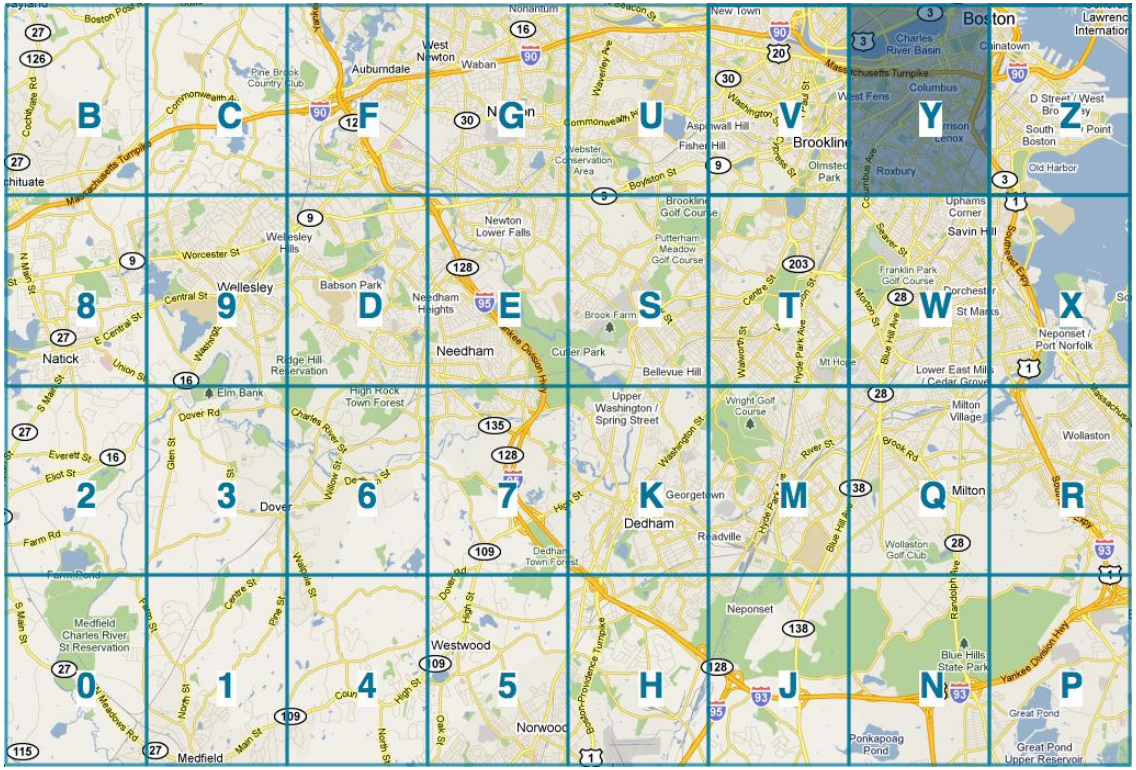
Zooming In: DRT



Zooming In: DRT2



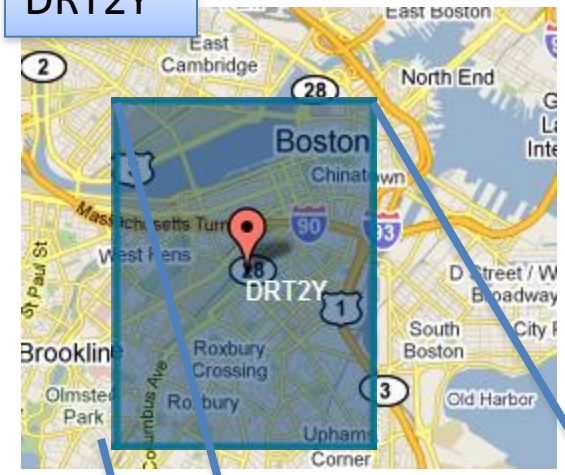
Zooming In: DRT2Y



Geohash Grids

Internal coordinates of an **odd** length geohash…

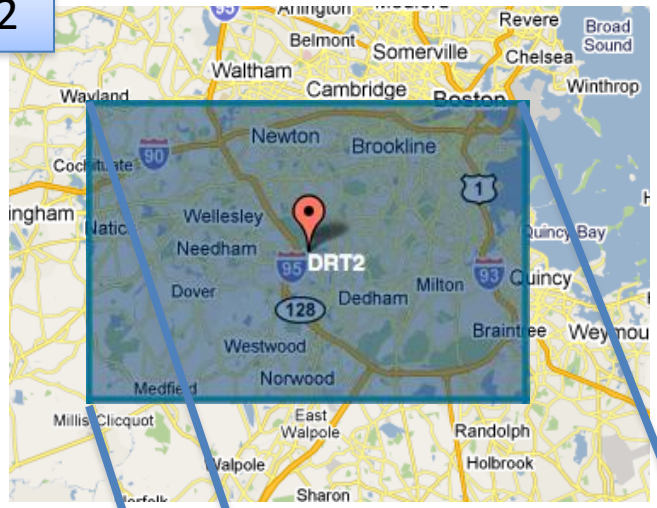
DRT2Y



P	R	X	Z
N	Q	W	Y
J	M	T	V
H	K	S	U
5	7	E	G
4	6	D	F
1	3	9	C
0	2	8	B

…and an **even** length geohash

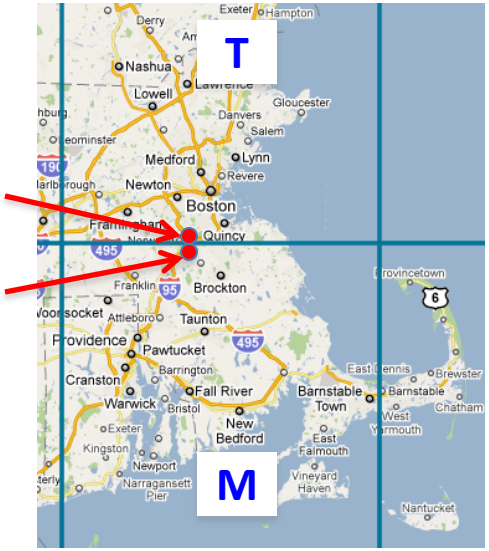
DRT2



B	C	F	G	U	V	Y	Z
8	9	D	E	S	T	W	X
2	3	6	7	K	M	Q	R
0	1	4	5	H	J	N	P

Proximity with Geohashes

- Points with a long common prefix are near each other
  - But the converse is not always true!
- ~~Points near each other share a long common prefix~~
  - Edges cases exist at every level:
    - D R **T**….
    - D R **M**…
    - (only share 2 letters)



Search Filter Algorithms

Filter Algorithm: Linear scan

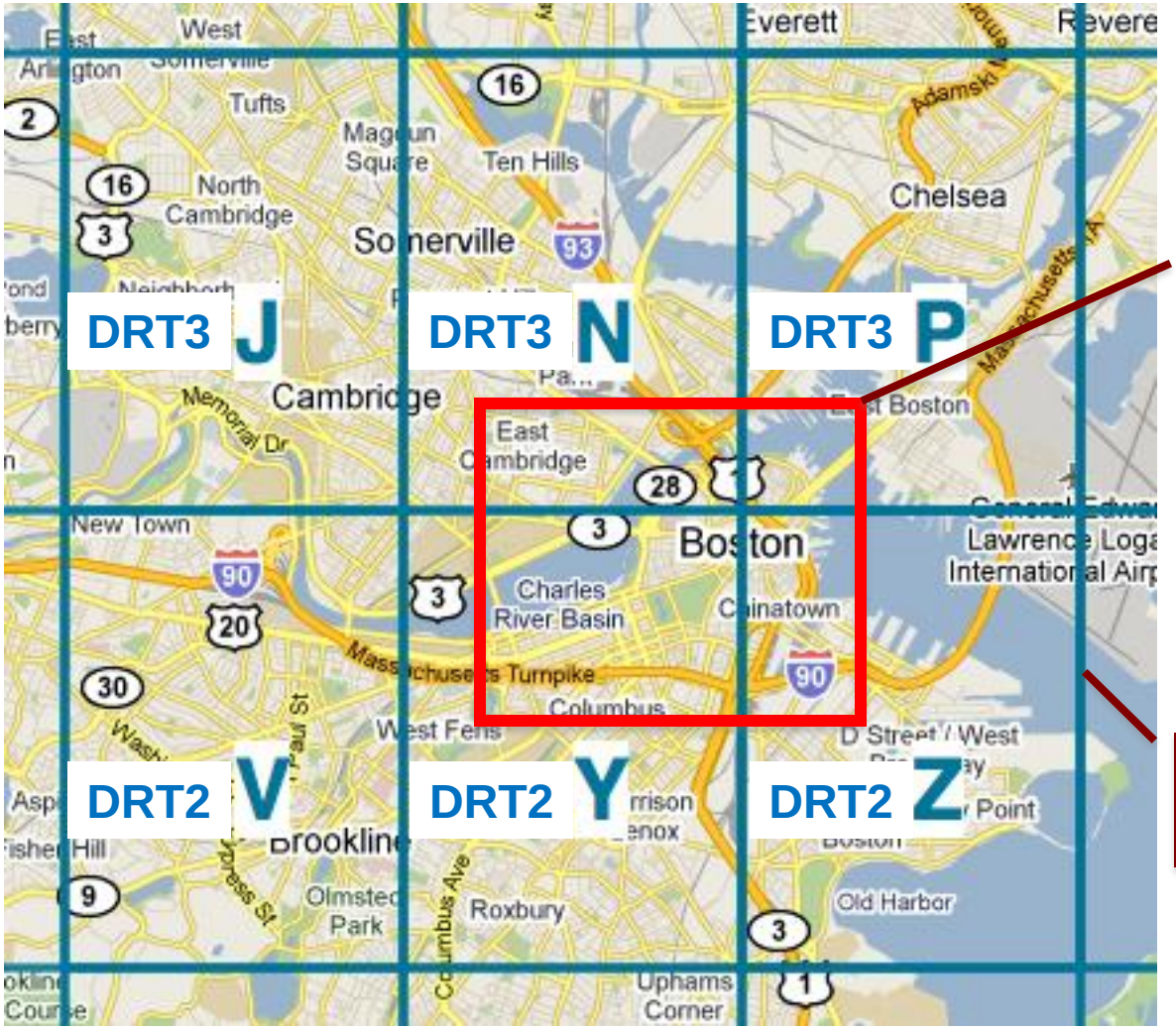
BASIS TECHNOLOGY

- The road not taken: dumb brute force
  - (linear scan)
- Algorithm:
  - Iterate over every indexed term a (geohash string), decode it, intersect with query shape
- Solr's GeoHashField does this
- Doesn't leverage unique spatial geohash properties

Index:

…
6A299
DKF30
DRT2Y
DRT2Z
DRT26
DRT3H
DRT3N
DRT3V
DRK54
DY8B2
F7FBZ
…

Filtering by Lat-Lon Box



Filter Algorithm: Fixed Grid Depth

BASIS TECHNOLOGY

- Determine a "good" grid cell size with minimal overlap of grid to query shape
- Algorithm:
  - Get list of "good" overlapping grid cells
  - For each grid cell *g*:
  - TermsEnum.seek(g)
  - Loop: TermsEnum.next() while *g* is a prefix of the current term:
  - Decode & intersect with query shape

Index:

…
6A299
DKF30
DRT2Y
DRT2Z
DRT26
DRT3H
DRT3N
DRT3V
DRK54
DY8B2
F7FBZ
…

Filter Algorithm: NGram Tree Traversal

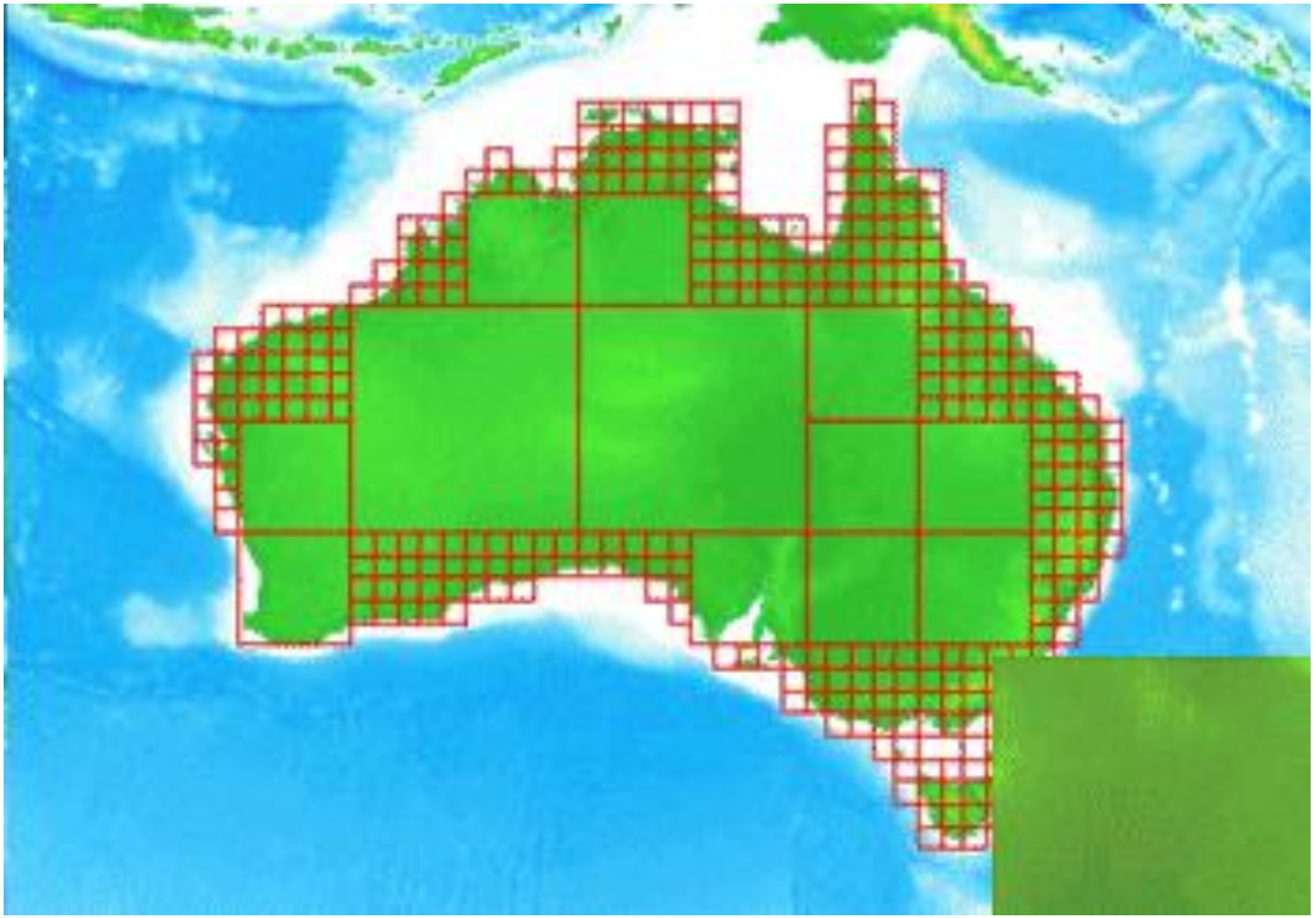
BASIS TECHNOLOGY

- Index each point at every grid level
  - D, DR, DRT, DRT2, DRT2Y
- Algorithm:
  - Recursive loop across top grid cells:
  - If cell is *within* query shape, simply add all assigned documents to result
  - If cell intersects query shape, recursive(cell.subcells)
- Actual details have more to it!

JIRA SOLR-2155

Tree Traversal Illustrated

BASIS TECHNOLOGY



Benchmarking

BASIS TECHNOLOGY

- Geonames.org, US data set – 2M points
- Point-radius query shape
- Geohash length 9 (~2 meters accuracy)

JIRA LUCENE-2844

km	Place/query	ms/query (LatLonType)	ms/query (SOLR-2155)
11	587	10.0	4.8
44	3404	11.5	4.3
230	45,536	21.8	24.0
1,800	1,319,692	288.5	142.3

Distance Sorting, Polygon Search

BASIS TECHNOLOGY

- Distance sorting
  - All points decoded into memory
  - Use Solr's geodist()
- Polygon query shape
  - Uses 3rd party JTS library (LGPL licensed)
  - Outstanding pole & prime meridian bugs

JIRA SOLR-2155

The Future: a new Lucene & Solr geospatial framework “LSP”

The Future: LSP

- A new geospatial framework for Lucene and Solr
  - http://code.google.com/p/lucene-spatial-playground/
- *Possible* successor to Lucene's module
- Committers:
  - Ryan McKinley (Voyager GIS)
  - Chris Male (JTeam)
  - David Smiley (MITRE)

What's in LSP

- Core shape interfaces (point, box, circle)
  - JTS extensions, adds polygon
- PrefixGrid / Tree abstraction
  - Geohash & quad implementations
- Multiple index / search strategies
- Query parser for Solr
- Benchmark
- Testing (of course)
- Demonstration web app

Future work

BASIS TECHNOLOGY

Features

- Indexing polygons
- Projections
- 1-dimensional grid?

Misc

- Benchmarking
- Testing

Performance

- Hilbert-curve optimized grids
- Faster distance sorting
  - Via projected data
  - Via sharing info w/ filtering

Thank You!

For more information

https://issues.apache.org/jira/browse/SOLR-2155

http://code.google.com/p/lucene-spatial-playground/