UNIVERSITY OF CALIFORNIA
UC RIVERSIDE

Interactive Exploration of Big Spatial Data

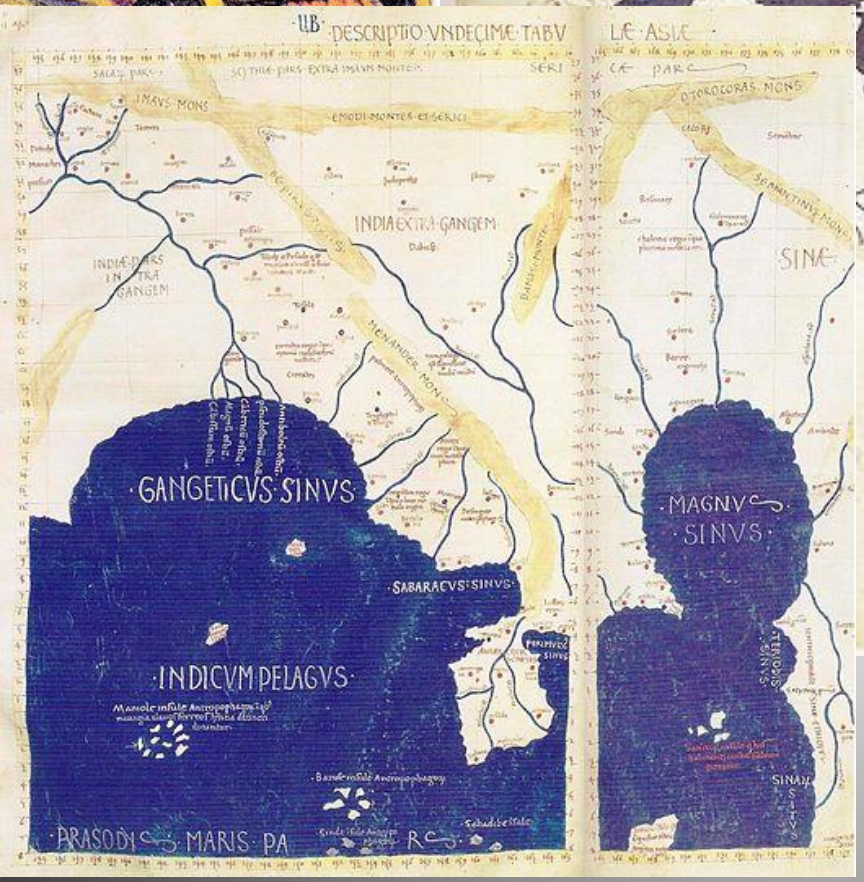
on SpatialHadoop and Beyond

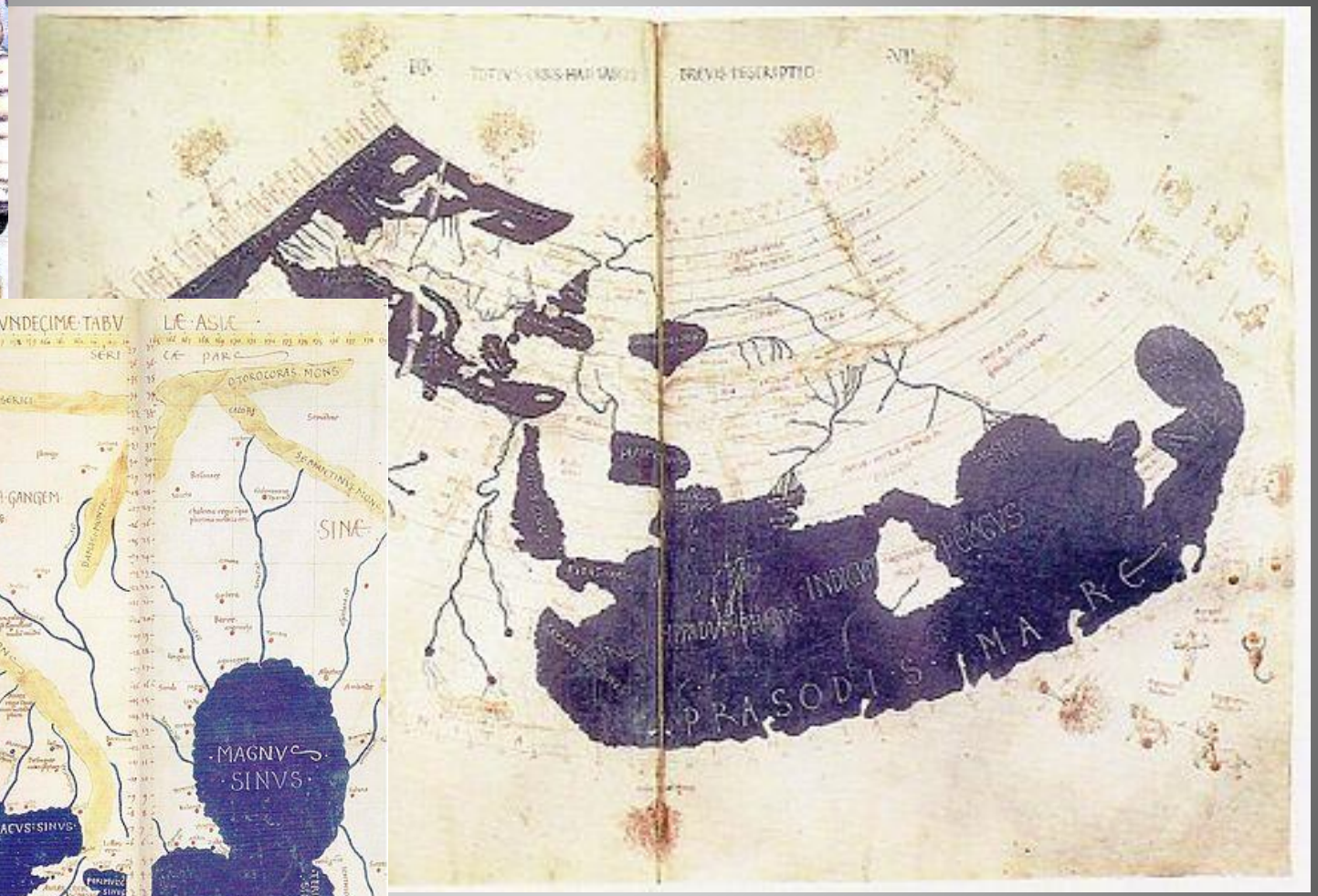
Ahmed Eldawy

Computer Science and Engineering

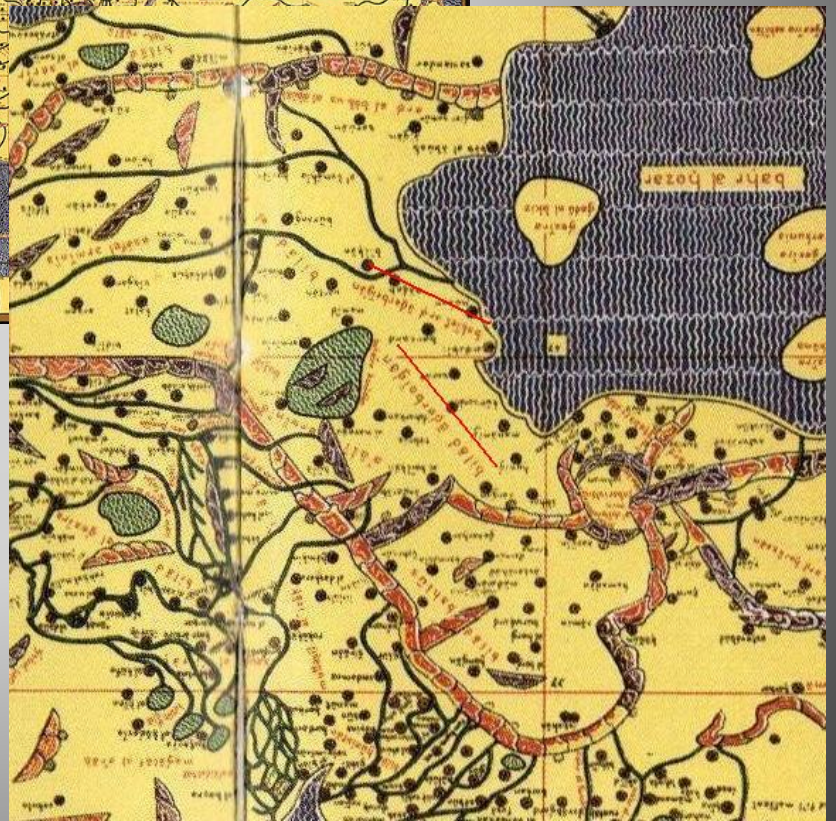

Once upon a
time...

Claudius Ptolemy (AD 90 – AD 168)

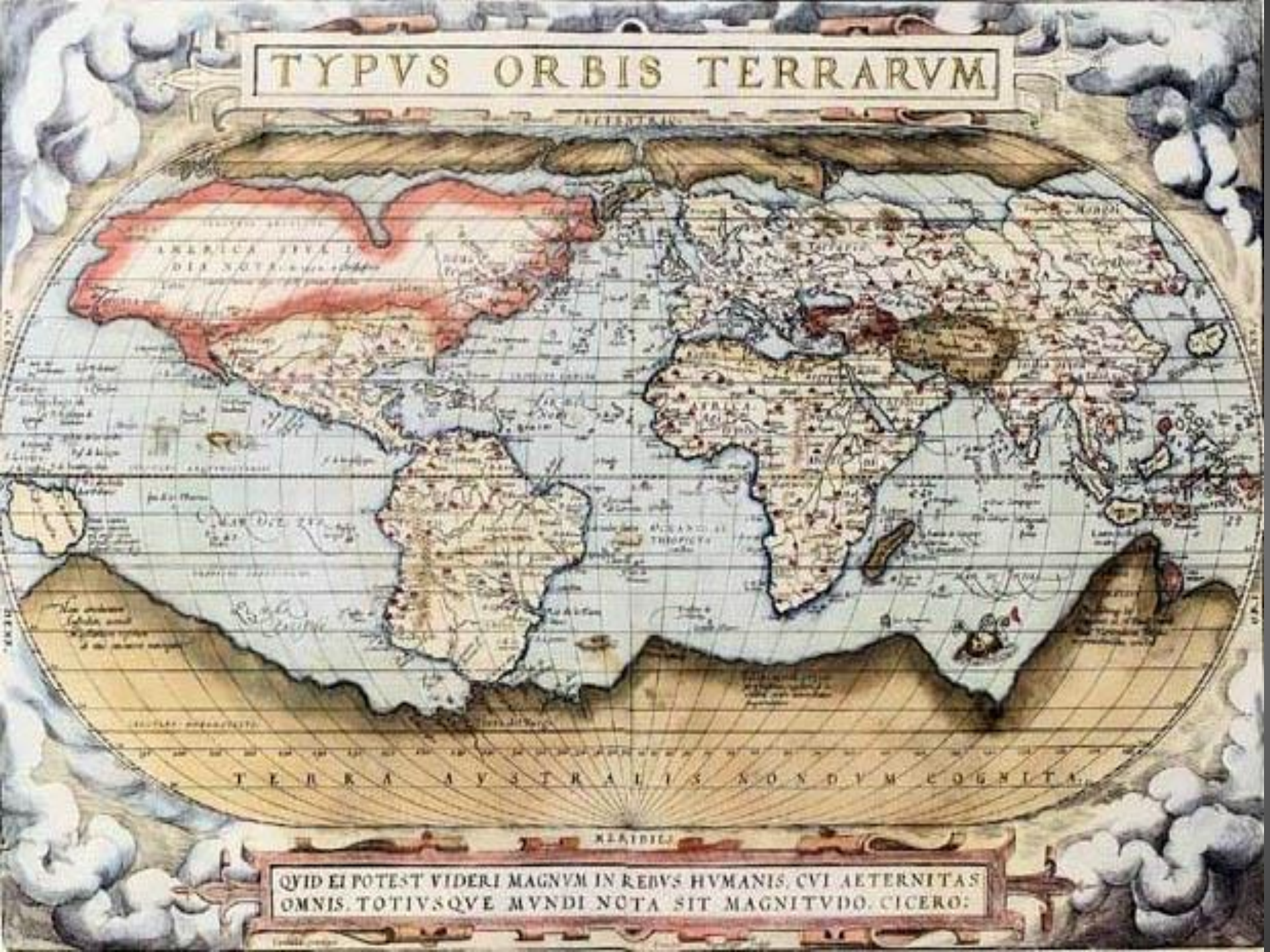




Al Idrisi (1099–1165)

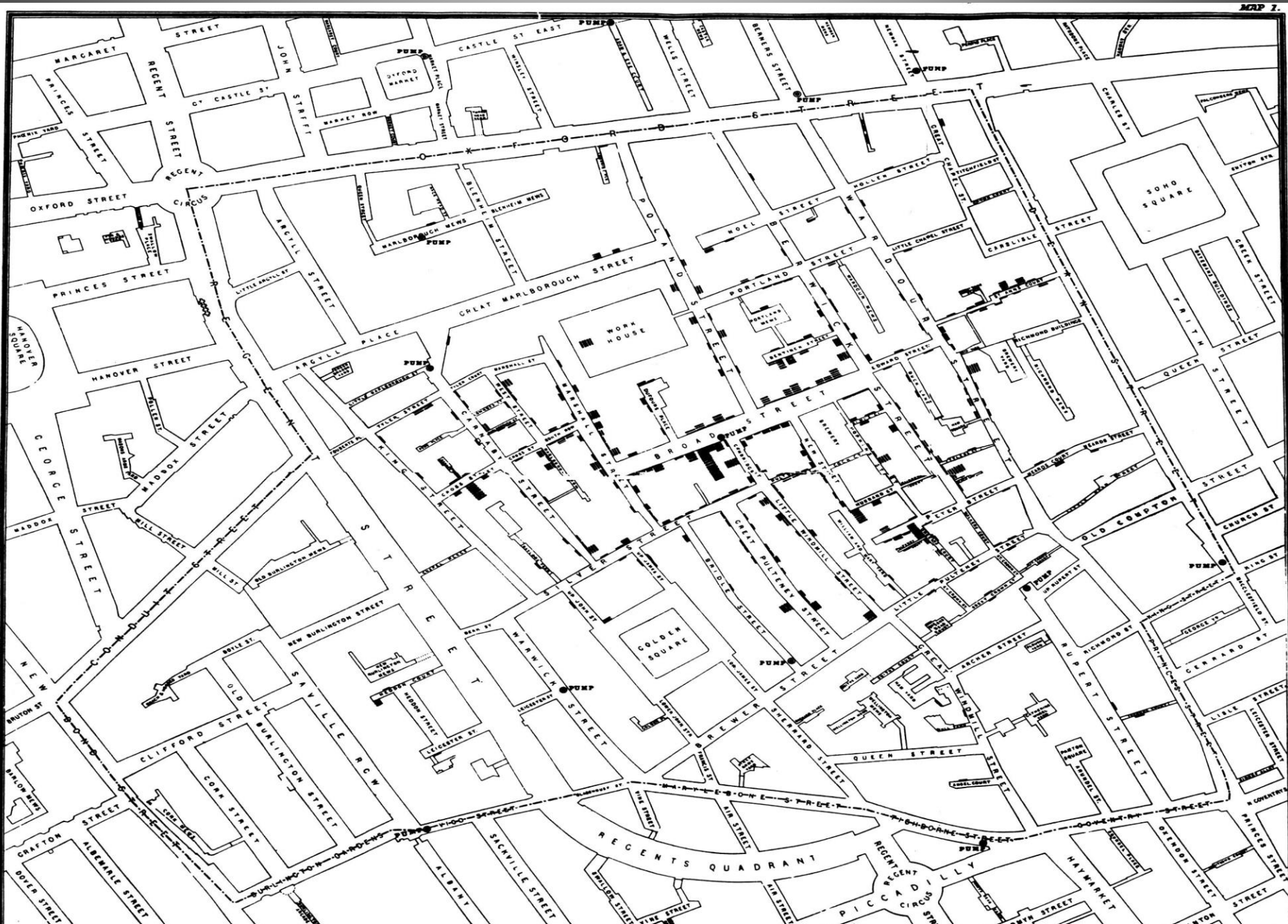



Mar Atlantico
EVROPA
AFRICA
INDIA
Brasil
Tropicus Cancri
Tropicus Capricorni
Mar Indico
Mar Pacifico
Mar Meridional
Mar do Sul
Equinoctialis


TYPVS ORBIS TERRARVM
TERRA AVSTRALIS NONDVM COGNITA
QVID EI POTEST VIDERI MAGNVM IN REBVS HVMANIS, CVI AETERNITAS
OMNIS, TOTIVSQVE MVNDI NOTA SIT MAGNITVDO. CICERO:


ARGONAVTICA.
EVROPAE PARS.
ASIAE PARS.
LIBYAE PARS.
SAVROMATAE.
SCYTHAE.
GETAE
ARIMASPHAE.
MAEOTIS PALVS.
AXENVM Aequor, Iasone pollutum remige primum.
CAVCASEVM MARE.
CAPPADOCIA.
THRACIA.
CELTAE.
SARDOVM PELAGVS.
TYRRHENVM AQVOR.
IONIVM MARE.
LIBYSTICVM MARE.
PELOPIS REGIO.
Maeotim pelagus
THRACIA
PROPONTIS.
BITHYNIA.
ILLVSTRISSIMO PRINCIPI CAROLO COMITI ARENBERGIO, BARONI SEPTIMONTII, DOMINO MIRVARTII, EQVITI AVREI VELLERIS, ETC.

Cholera cases in the London epidemic of 1854



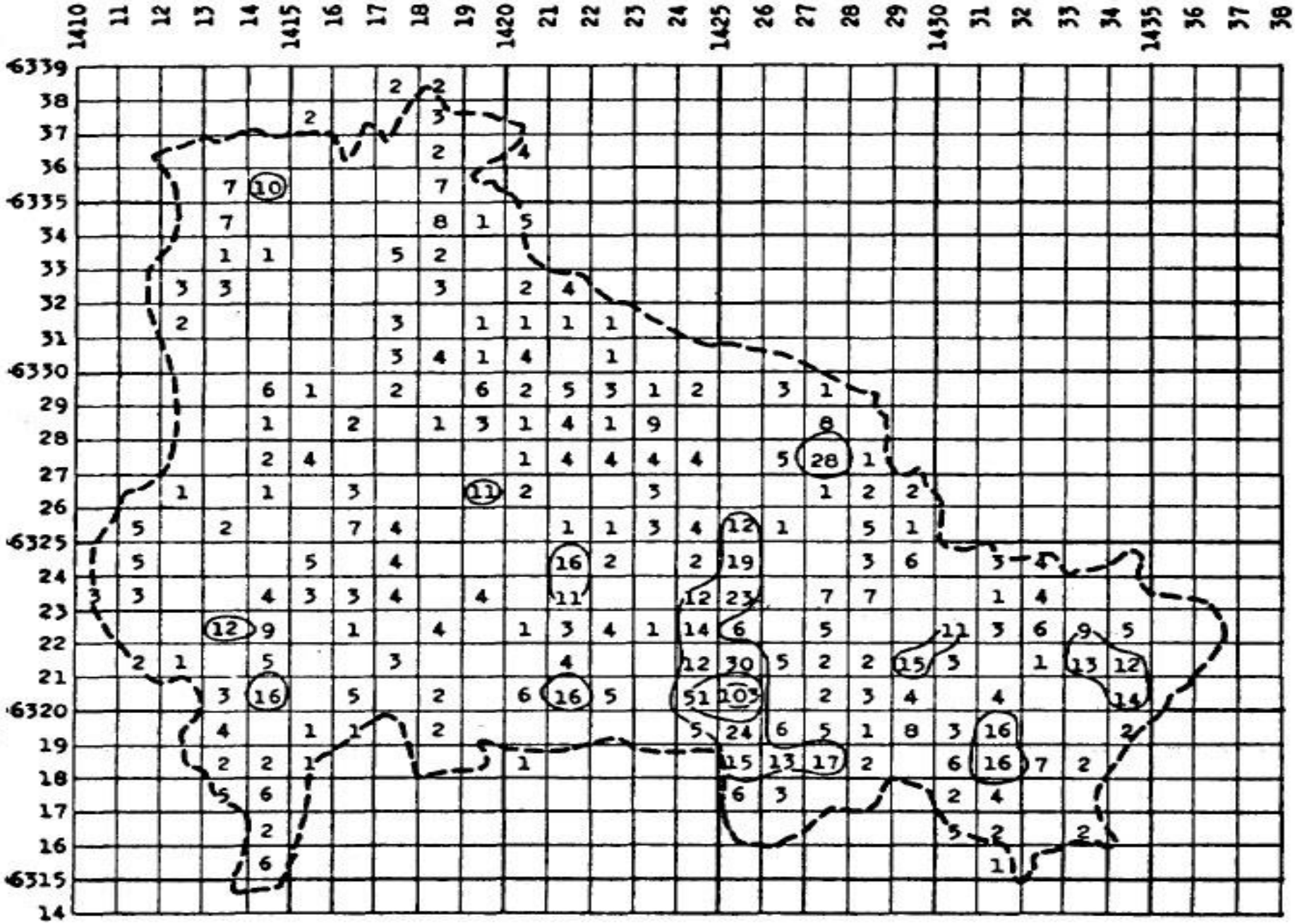


FIGURE 3—Children under 15 years of age in 1940.

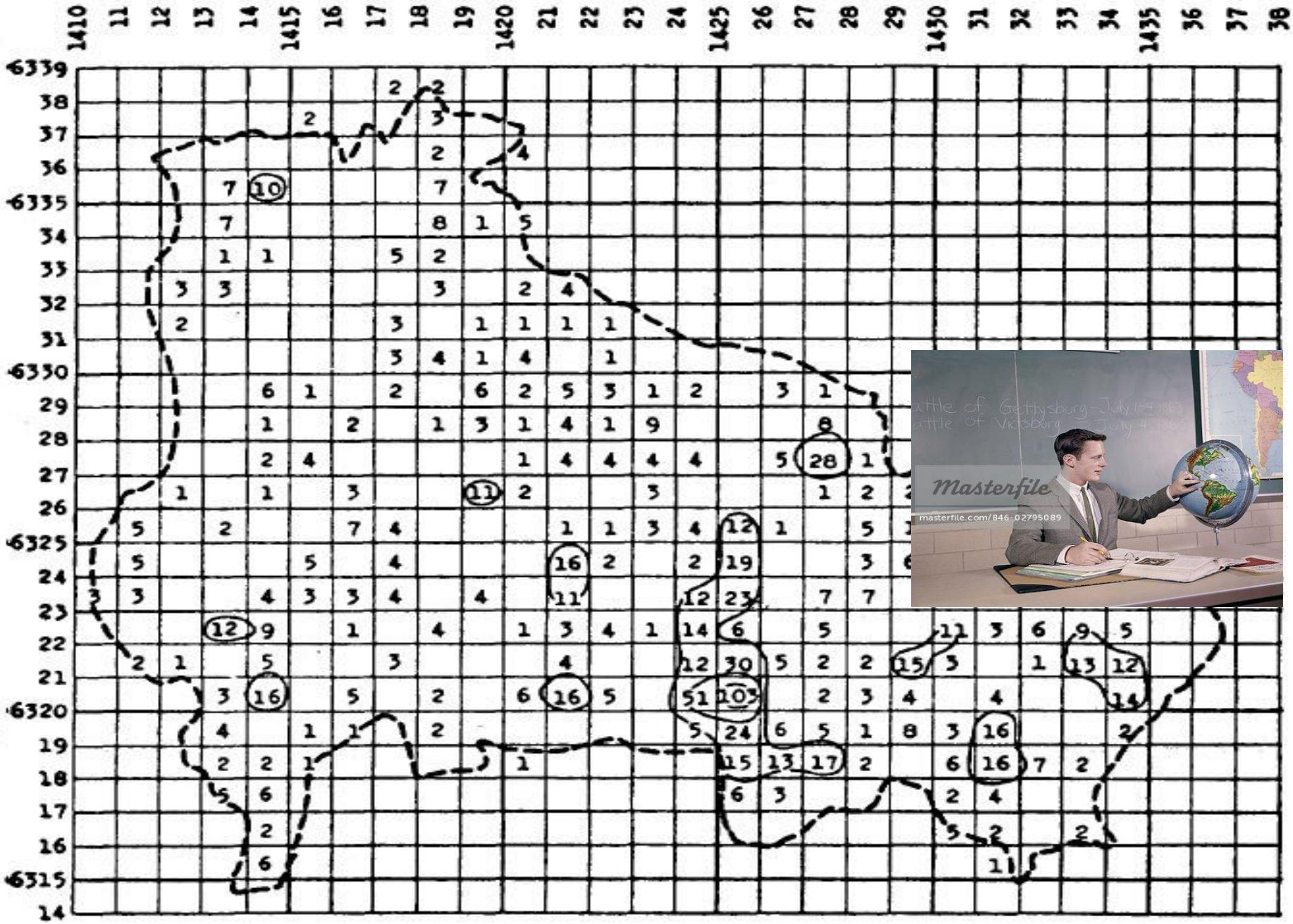
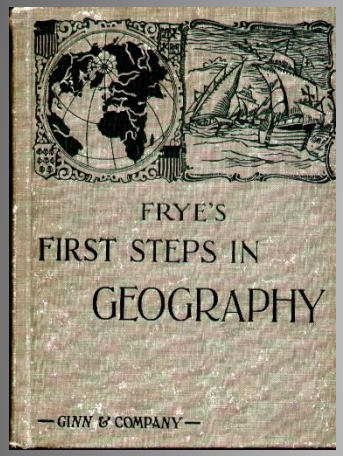


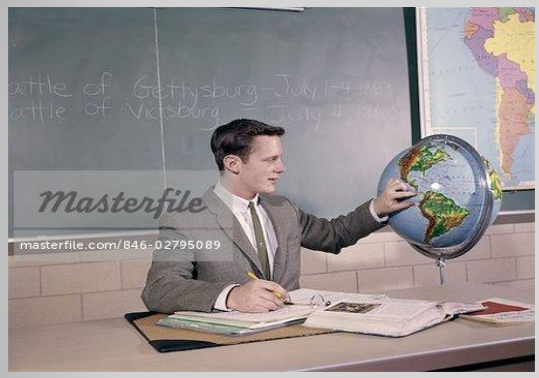
FIGURE 3—Children under 15 years of age in 1940.


UAR
5 M

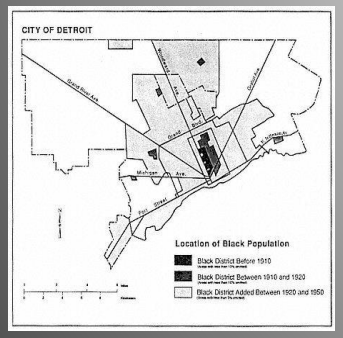

OCÉAN ATLA
Newcastle
Liverpool
Birmingham
Cardiff
Londres
Hambourg
Amsterdam
Rotterdam
Anvers
Le Havre
Paris
Nantes St-Nazaire
Bordeaux
Marseille
Milan
Trieste
Gênes
Madrid
Barcelone
Rome


UNIVAC


FRYE'S
FIRST STEPS IN
GEOGRAPHY
—GINN & COMPANY—


Masterfile
masterfile.com/846-02795089


IBM

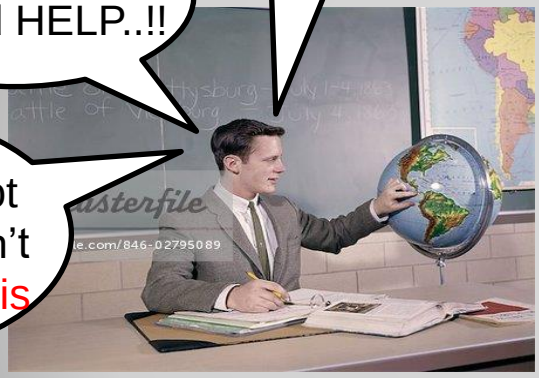

CITY OF DETROIT
Location of Black Population
Black District Before 1910
Black District Between 1910 and 1920
Black District Added Between 1920 and 1950

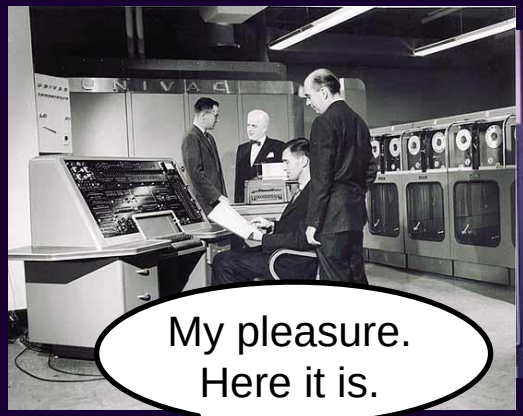

Masterfile
masterfile.com/846-02793618

Cool computer technology..!! Can I use it in my application?

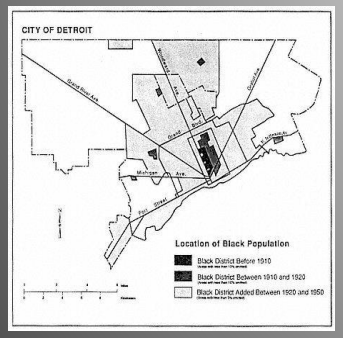
I have **BIG** data. I need HELP..!!

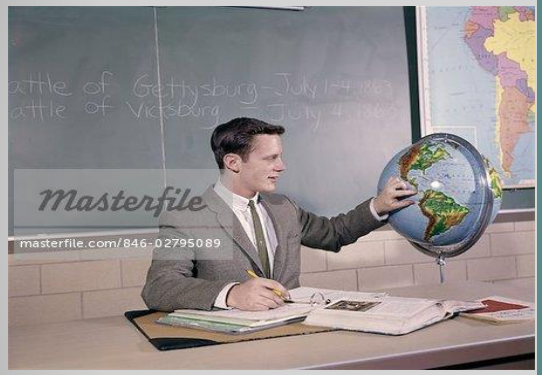
Oh..!! But, it is not made for me. Can't make use of it as is





My pleasure. Here it is.

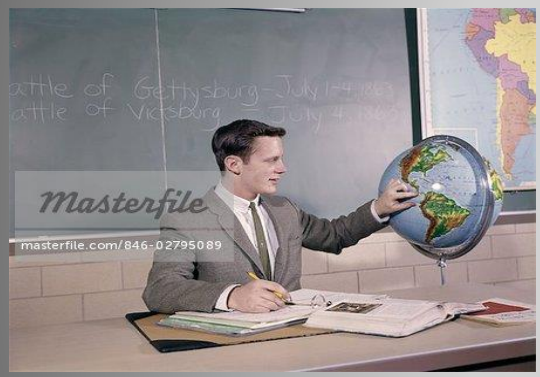



Masterfile
masterfile.com/846-02795089


IBM

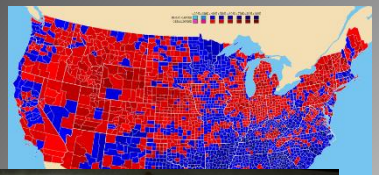
1969

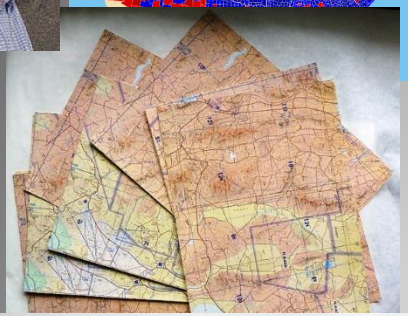
Kindly let me understand your needs

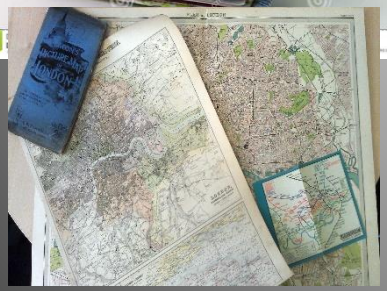


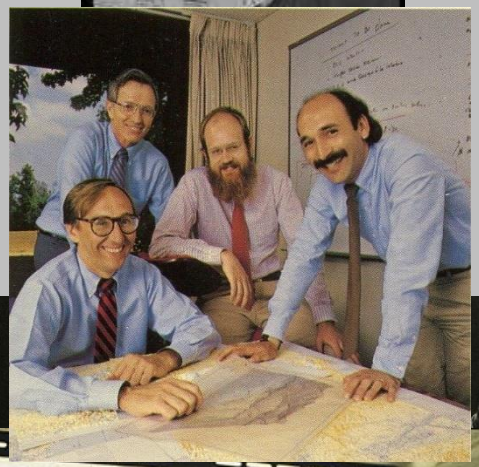
Kindly let me get the technology you have

ESRI

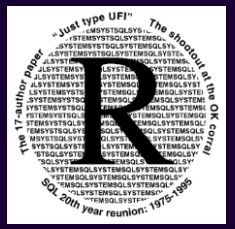









ESRI


"Just type UFI"
The 17-author paper
The shootout at the OK corral
SQL 20th year reunion: 1975-1995


DATABASE
MANAGEMENT
SYSTEMS

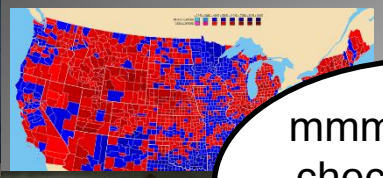

Informix

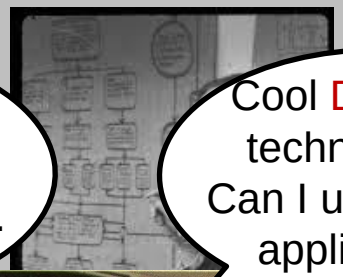

ORACLE
DATABASE
ORACLE

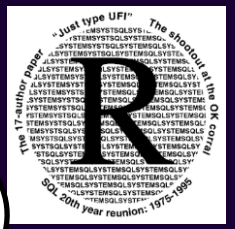
PostgreSQL


AIRLINE BOOKING


Banking







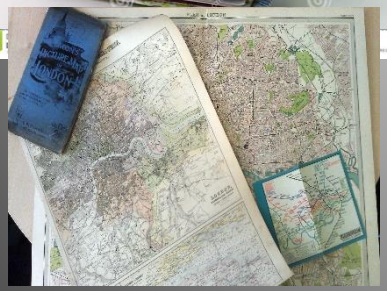
mmm…Let me check with my good friends there.

Cool Database technology..!! Can I use it in my application?

HELP..!! I have **BIG** data. Your technology is not helping me

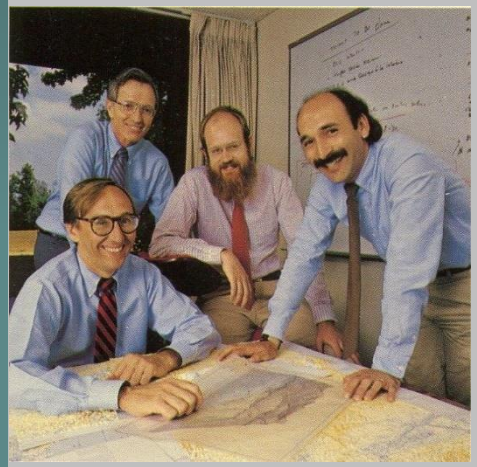
Oh..!! But, it is not made for me. Can't make use of it as is

My pleasure. Here it is.

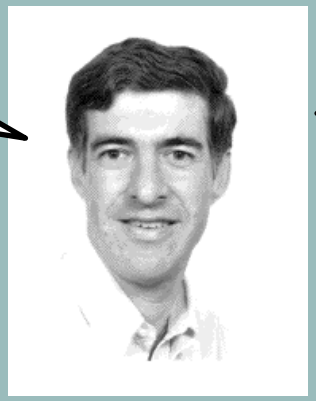


Informix

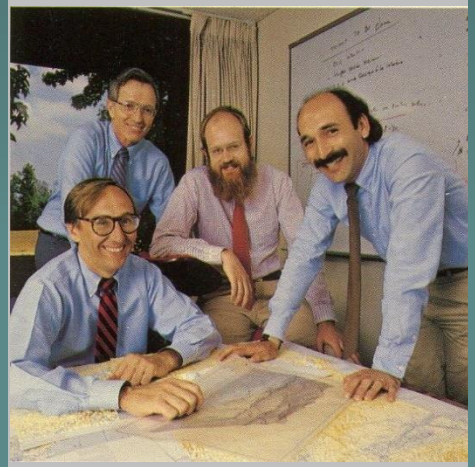
PostgreSQL

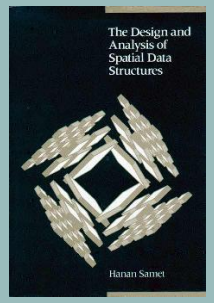


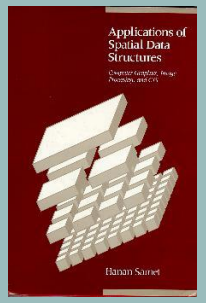
Kindly let me understand your needs

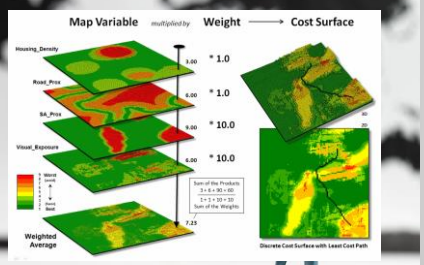


Kindly let me get the technology you have

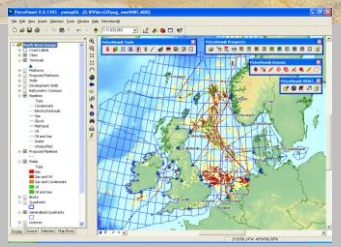




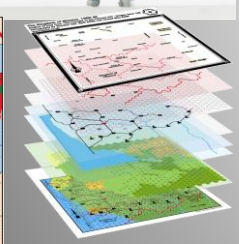


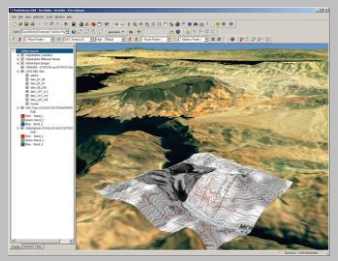

Map Variable
multiplied by
Weight
Cost Surface
Housing_Density
Road_Prox
BA_Prox
Visual_Exposure
* 1.0
* 1.0
* 10.0
* 10.0
Weighted Average
Discrete Cost Surface with Least Cost Path

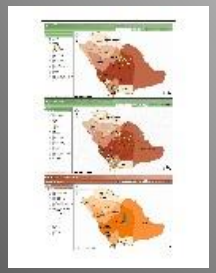

ArcGIS
ESRI




IRAN
Arabian Gulf
BAHRAIN
Manama
SAUDI ARABIA
QATAR
Doha
Riyadh
N
0 km 150

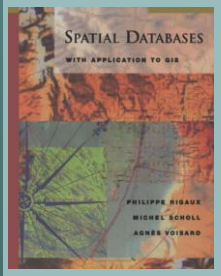


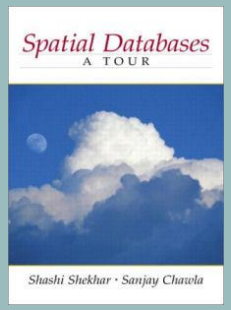





idrisi


The Design and Analysis of Spatial Data Structures
Applications of Spatial Data Structures
PostGIS
Geospatial Objects for PostgreSQL
ORACLE 11g
DATABASE
Oracle Spatial and Graph
ORACLE
Beginning Spatial with SQL Server 2008
Alastair Aitchison


SPATIAL DATABASES
WITH APPLICATION TO GIS
PHILIPPE RIGAUX
MICHEL SCHOLL
AGNÈS VOISARD


Spatial Databases
A TOUR
Shashi Shekhar · Sanjay Chawla

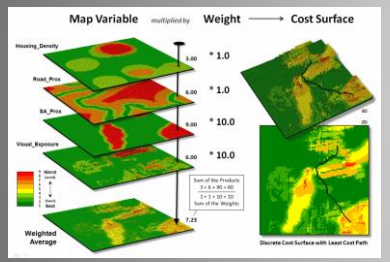

ORACLE
IBM
DB2

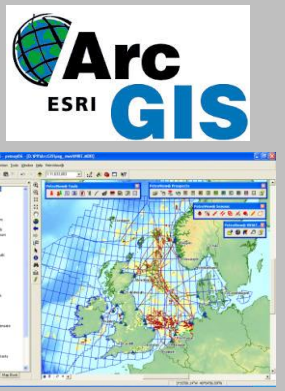

Microsoft
SQL Server


PostgreSQL
the world's most advanced open source database

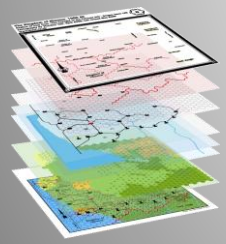

MySQL

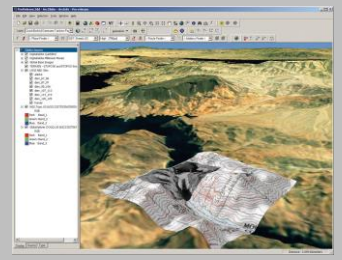

SYBASE
An SAP Company

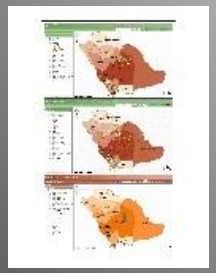

Map Variable
multiplied by
Weight
Cost Surface
Housing_Density
Road_Prox
BA_Prox
Visual_Exposure
* 1.0
* 1.0
* 10.0
* 10.0
Weighted Average
Discrete Cost Surface with Least Cost Path



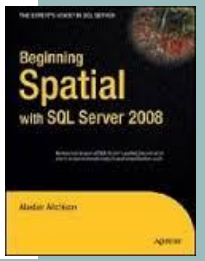

IRAN
Arabian Gulf
BAHRAIN
Manama
SAUDI ARABIA
QATAR
Doha
Riyadh
N
0 km 150

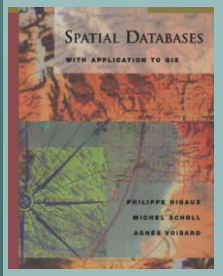


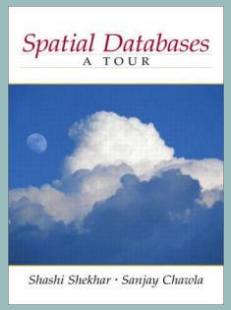





ORACLE 11g
DATABASE
Oracle
Spatial and Graph
ORACLE


Beginning
Spatial
with SQL Server 2008
Alaadin Aljchiaim


Spatial Databases
WITH APPLICATION TO GIS
PHILIPPE RIGAUX
MICHEL SCHOLL
AGNÈS VOISARD


Spatial Databases
A TOUR
Shashi Shekhar · Sanjay Chawla


PostGIS
Geospatial Objects for PostgreSQL


ORACLE


IBM
DB2


Microsoft
SQL Server


PostgreSQL
the world's most advanced open source database

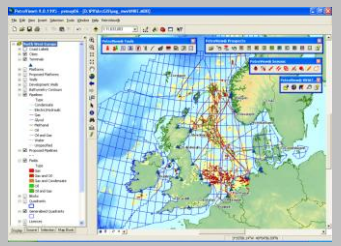

MySQL

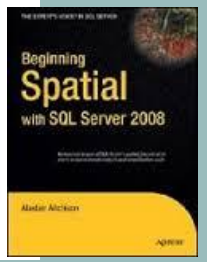

SYBASE
An SAP Company

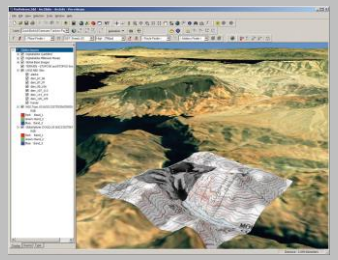
Google
bing
twitter
facebook
Map Reduce
hadoop
HIVE
Spark
amazon web services™

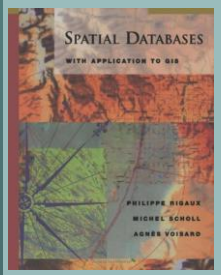
ArcGIS ESRI

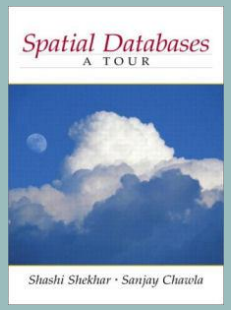
PostGIS
Geospatial Objects for PostgreSQL











ORACLE
IBM DB2
Microsoft SQL Server
PostgreSQL
MySQL
SYBASE | An SAP Company

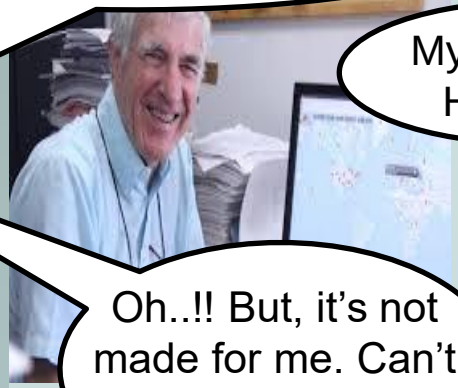
idrisi

HELP..!! Again,
I have **BIG** data.
Your technology is not helping me

Let me check with my other good friends there.

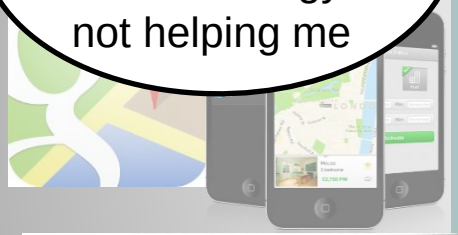
Sorry, seems like the DBMS technology cannot scale more

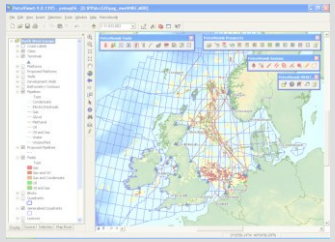
Cool Big Data technology..!!
Can I use it in my application?

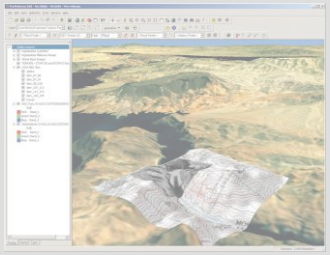


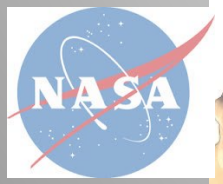
Oh..!! But, it's not made for me. Can't make use of it as is

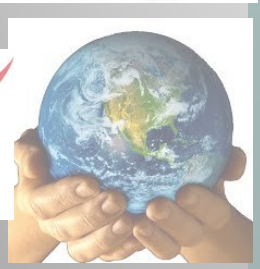
My pleasure.
Here it is.

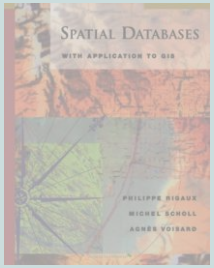


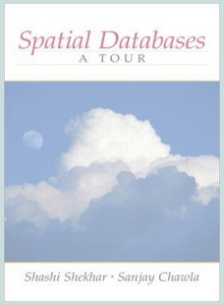












Google

bing

twitter

facebook

Map Reduce

hadoop

amazon web services

HIVE

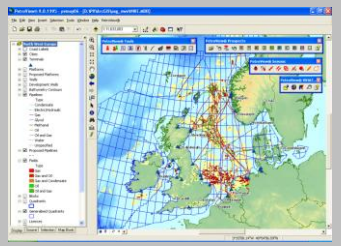
Spark

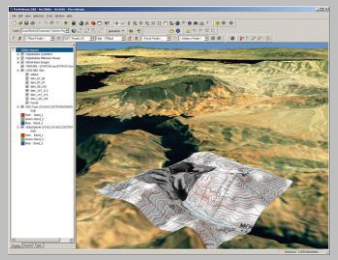
PostGIS

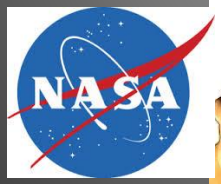
Geospatial Objects for PostgreSQL

ArcGIS

ESRI





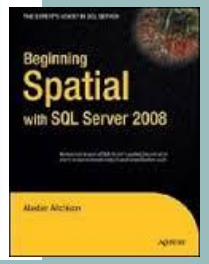


idrisi

ORACLE 11g DATABASE

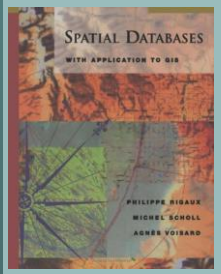
Oracle Spatial and Graph

ORACLE



Beginning Spatial with SQL Server 2008

Alastair Aitchison



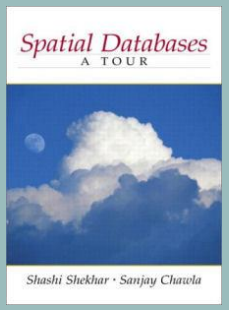
SPATIAL DATABASES

WITH APPLICATION TO GIS

PHILIPPE RIGAUX

MICHEL SCHOLL

AGNÈS VOISARD



Spatial Databases

A TOUR

Shashi Shekhar · Sanjay Chawla

ORACLE

IBM DB2

Microsoft SQL Server

PostgreSQL

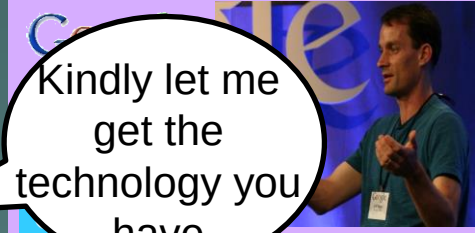
the world's most advanced open source database

MySQL

SYBASE | An SAP Company

Kindly let me understand your needs

Kindly let me get the technology you have



facebook

Map Reduce

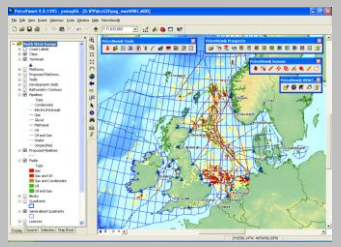
hadoop

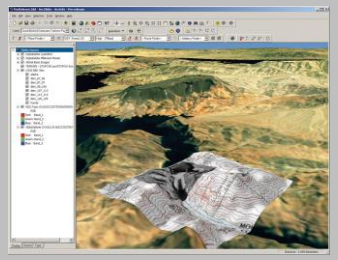
amazon web services

HIVE

Spark

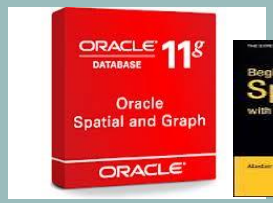
ESRI ArcGIS

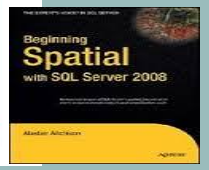




PostGIS

Geospatial Objects for PostgreSQL





ORACLE

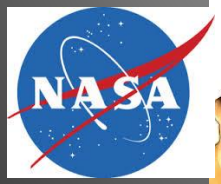
IBM DB2

Microsoft SQL Server

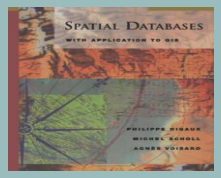
PostgreSQL

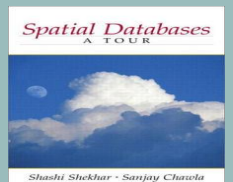
MySQL

SYBASE An SAP Company



idrisi





Big Spatial Data

Google

bing

twitter

facebook

Map Reduce

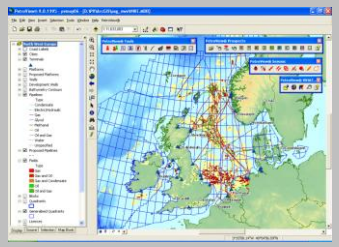
hadoop

amazon web services

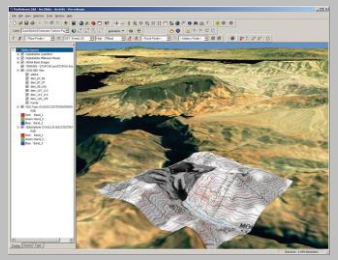
HIVE

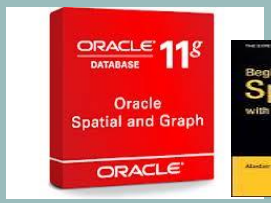
Spark

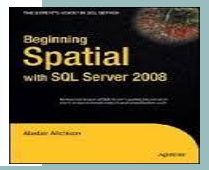
ArcGIS ESRI



PostGIS
Geospatial Objects for PostgreSQL







ORACLE

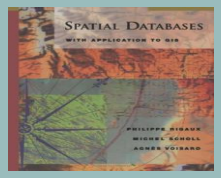
IBM DB2

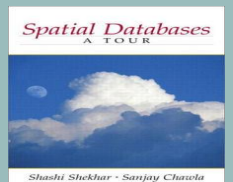
Microsoft SQL Server

PostgreSQL

MySQL

SYBASE | An SAP Company





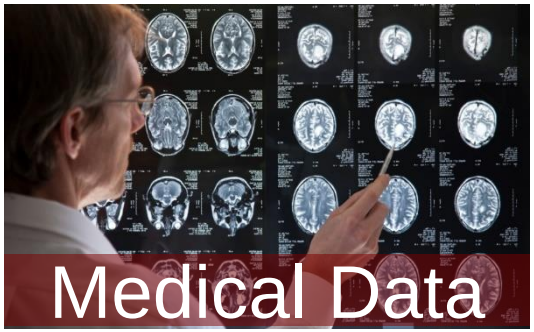
NASA

idrisi

Tons of Spatial data out there…

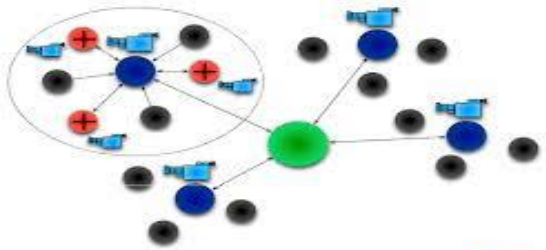
Geotagged Microblogs

Geotagged Pictures

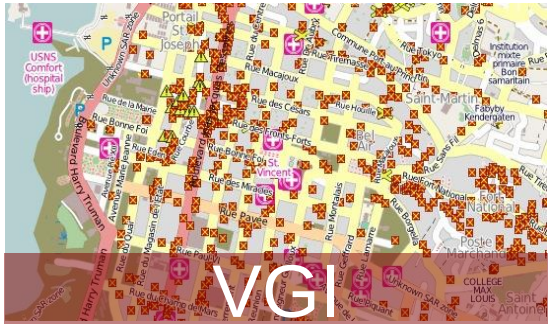


Medical Data

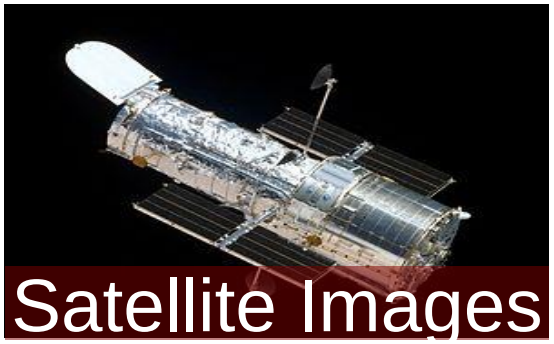
Smart Phones



Sensor Networks



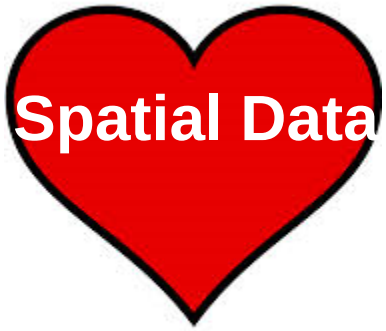
VGI



Satellite Images

Traffic Data

Spatial Data & Hadoop ➔ SpatialHadoop




```
points = LOAD 'points' AS
     (id:int, x:int, y:int);
result = FILTER points BY
  x < xmax AND x >= xmin AND
  y < ymax AND y >= ymin;
```

```
points = LOAD 'points' AS
     (id:int, location:point);
result = FILTER points BY
  Overlap(location, rectangle
    (xmin, ymin, xmax, ymax));
```

Takes 193 seconds

Finishes in 2 seconds

Spatial Hadoop

UCR

Spatial Language


Spatial Indexes


Spatial Operations


Visualization

80,000 downloads in one year

Conducted more than seven keynotes, tutorials, and invited talks

Industry

Microsoft
ORACLE
esri
GEOSYS Growing agriculture
3M
Hortonworks
XEROX PARC
BlackBerry

Academia

Students Projects

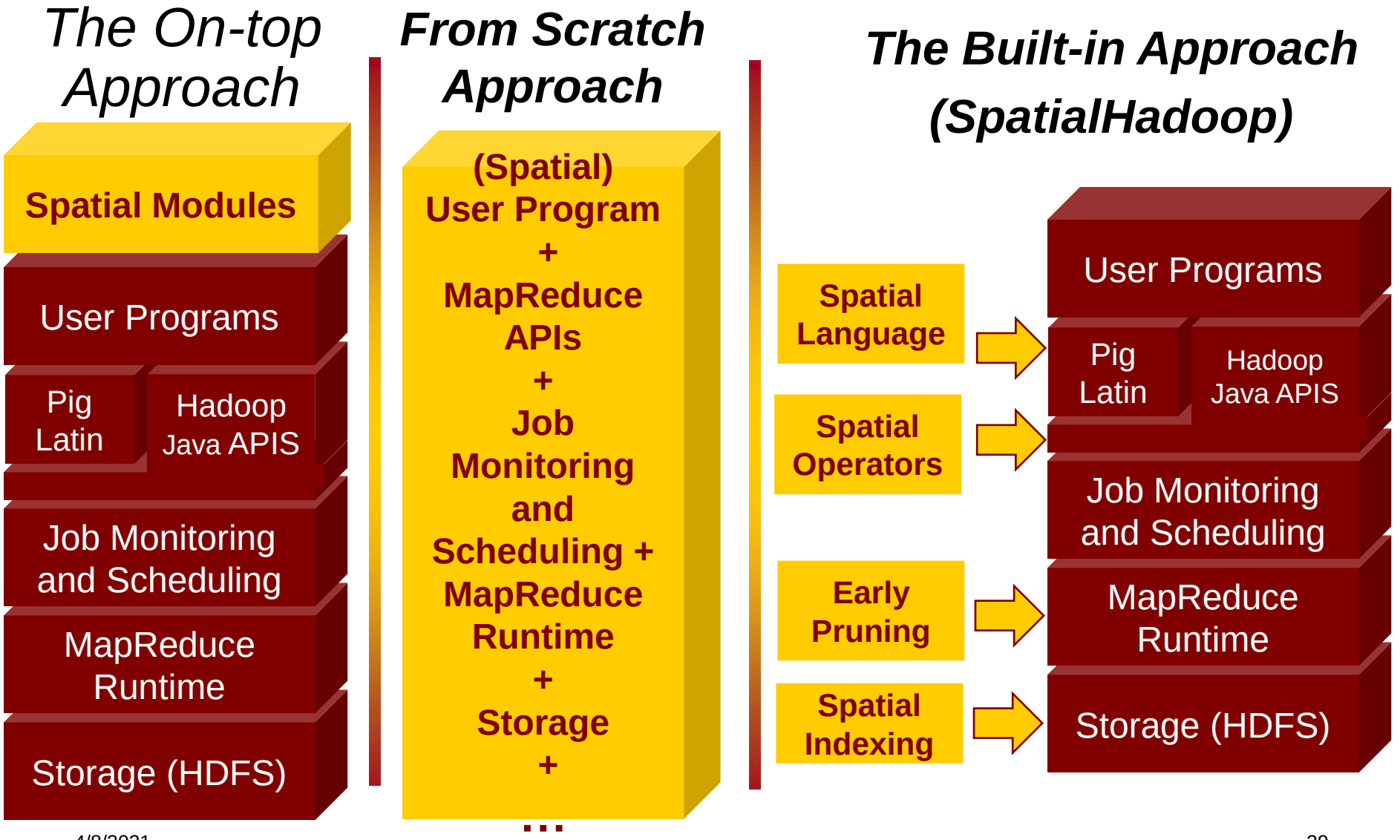
UT DALLAS
Salisbury University
Indian Institute of Technology Bombay
Université de Versailles St-Quentin-en-Yvelines
Sogang University
Alexandria University
Wuhan University
National and Kapodistrian University of Athens

Collaboration

The Cornell Lab of Ornithology
University of Maryland
UC Irvine
GIS Innovation Center
Kunsan National University
University of Genova

>500GB public datasets for benchmarking and testing

The Built-in Approach of SpatialHadoop



Agenda

- The ecosystem of SpatialHadoop
  - ~~Motivation~~
  - Internal system design
  - Applications
  - Related work
  - Performance results
- Interactive data exploration

SpatialHadoop Architecture

Applications: SHAHED [ICDE'15] – MNTG [SSTD'13, ICDE'14]
TAREEG[SIGMOD'14, SIGSPATIAL'14]

VLDB'13
ICDE'15

Language
Pigeon [ICDE'14]

Visualization
[VLDB'15, ICDE'16]

Operations
Basic operations – CG_Hadoop [SIGSPATIAL'13]

MapReduce
Spatial File Splitter
Spatial Record Reader

Indexing
Grid – R-tree – R+-tree – Quad tree
[VLDB'15]

ST-Hadoop

Indexing

Applications: SHAHED [ICDE'15] – MNTG [SSTD'13, ICDE'14]
TAREEG[SIGMOD'14, SIGSPATIAL'14]

VLDB'13
ICDE'15

Language
Pigeon [ICDE'14]

Visualization
[VLDB'15, ICDE'16]

Operations
Basic operations – CG_Hadoop [SIGSPATIAL'13]

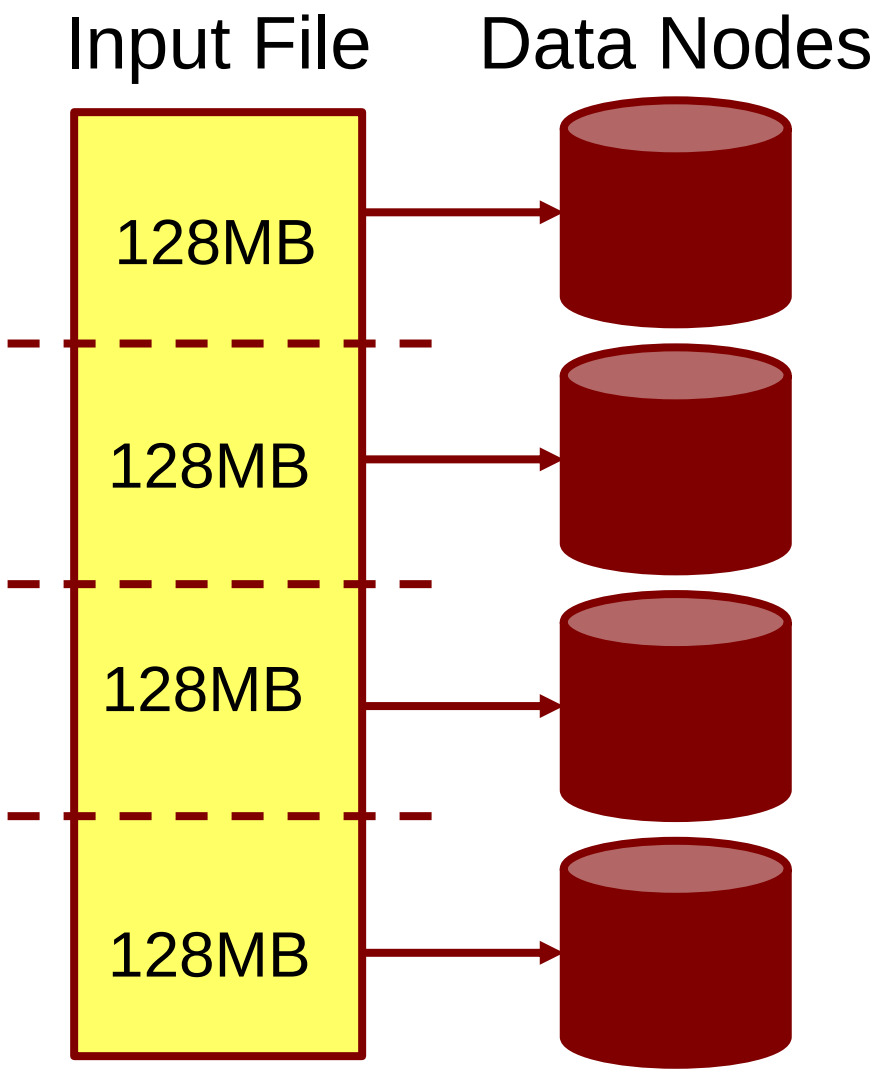
MapReduce
Spatial File Splitter
Spatial Record Reader

Indexing
Grid – R-tree – R+-tree – Quad tree
[VLDB'15]

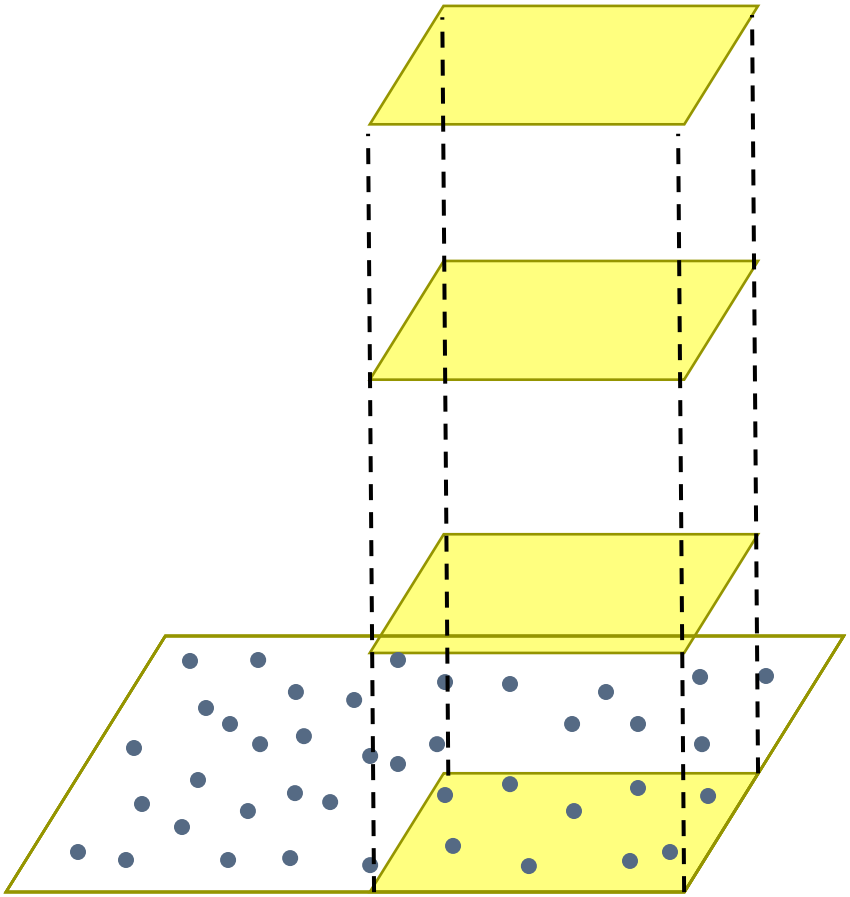
ST-Hadoop

Data Loading in Hadoop

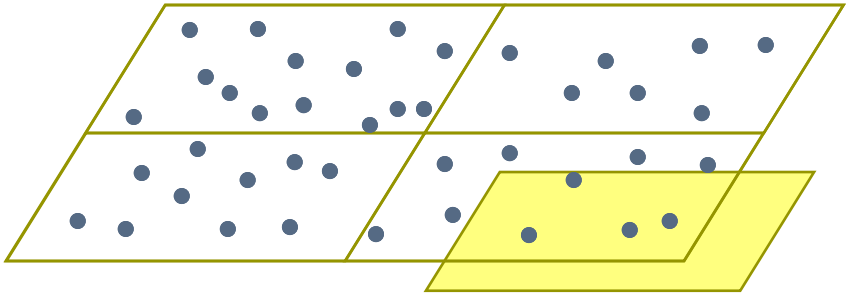
- Blindly chops down a big file into 128MB chunks
- Values of records are not considered
- Relevant records are typically assigned to two different blocks
- HDFS is too restrictive where files cannot be modified



Spatial Distributed File System

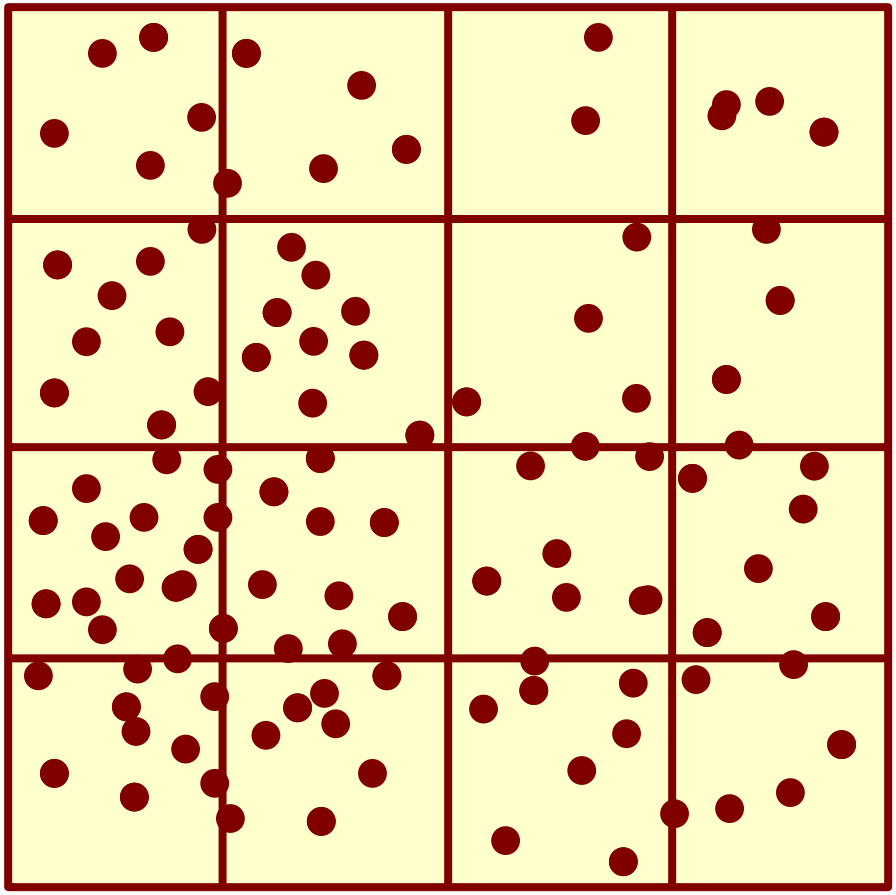


Default Partitioning



Spatial Partitioning

Uniform Grid



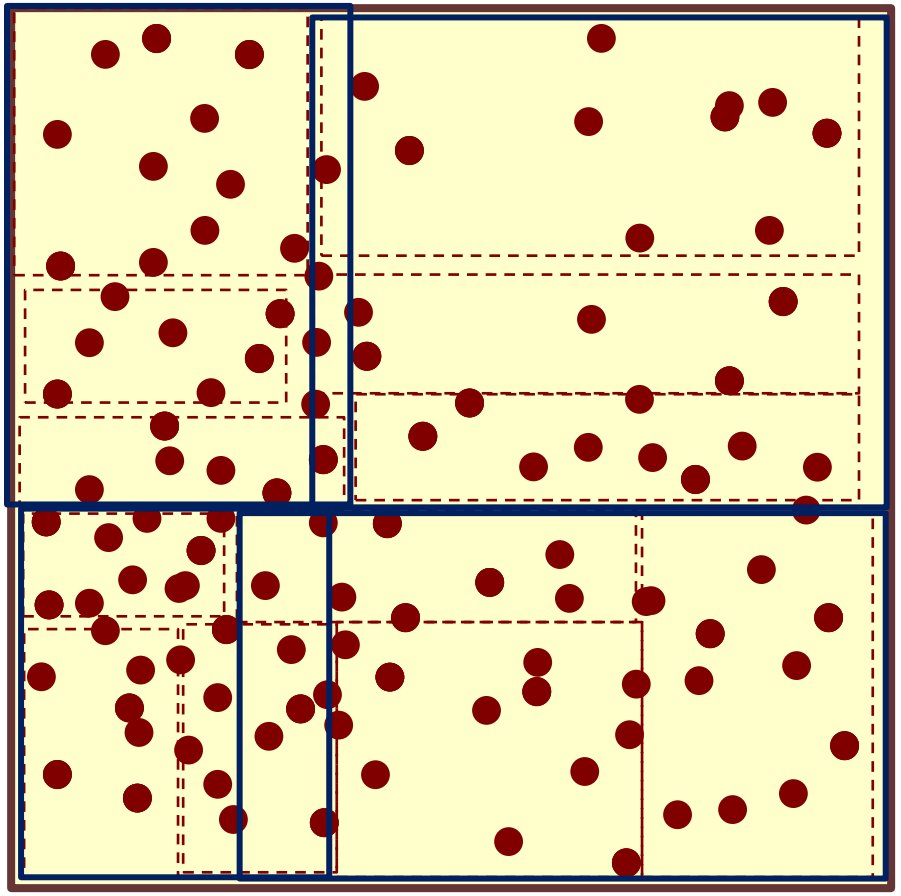
Works only for uniformly distributed data

R-tree

- Read a sample
- Bulk load the sample into an R-tree
  - Leaf node capacity C

$$C = \frac{k.B}{|R|(1+\alpha)}$$

  - k: Sample size
  - B: HDFS Block capacity
  - |R|: Input size
  - α: Index overhead
- Use MBR of leaf nodes as partition boundaries

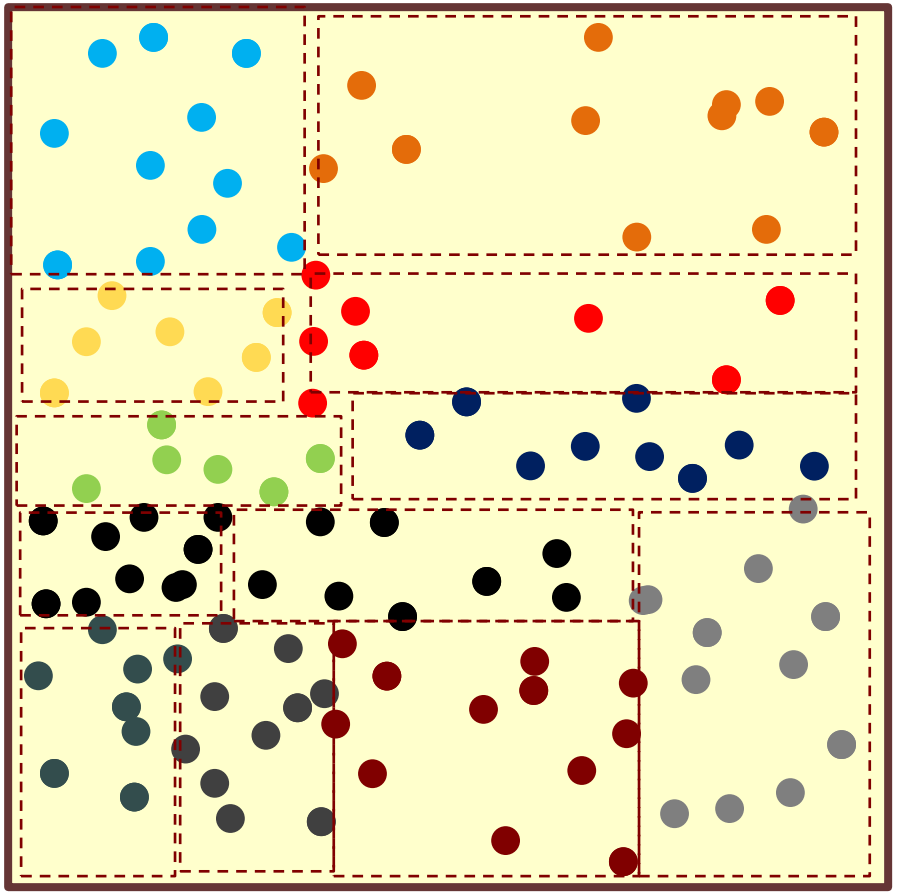


R-tree

- Read a sample
- Bulk load the sample into an R-tree
  - Leaf node capacity C

$$C = \frac{k.B}{|R|(1+\alpha)}$$

  - k: Sample size
  - B: HDFS Block capacity
  - |R|: Input size
  - α: Index overhead
- Use MBR of leaf nodes as partition boundaries
- Partition the data

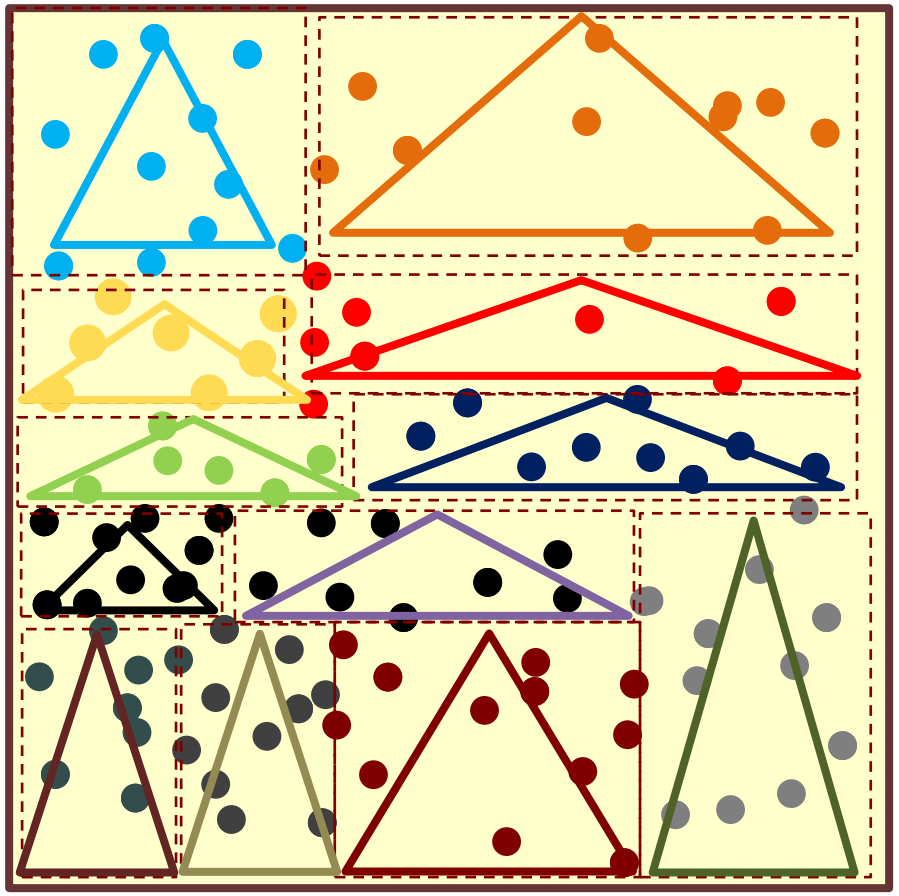


R-tree

- Read a sample
- Bulk load the sample into an R-tree
  - Leaf node capacity C

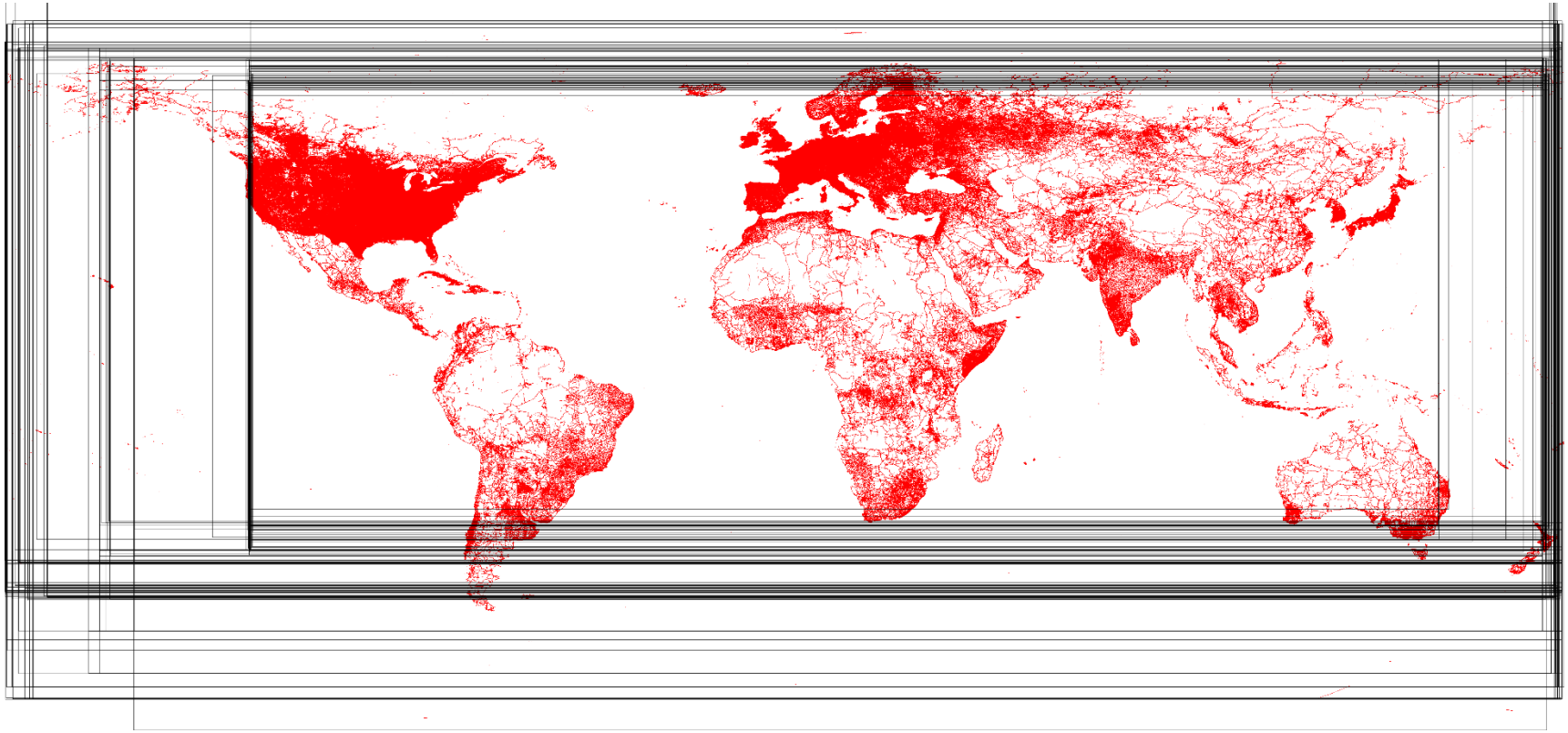
$$C = \frac{k.B}{|R|(1+\alpha)}$$

  - k: Sample size
  - B: HDFS Block capacity
  - |R|: Input size
  - α: Index overhead
- Use MBR of leaf nodes as partition boundaries
- Partition the data
- Optional: Build R-tree Local indexes



R-tree-based Index of a 400 GB road network

Non-indexed Heap File



Operations

Applications: SHAHED [ICDE'15] – MNTG [SSTD'13, ICDE'14]
TAREEG[SIGMOD'14, SIGSPATIAL'14]

VLDB'13
ICDE'15

Language
Pigeon [ICDE'14]

Visualization
[VLDB'15, ICDE'16]

Operations
Basic operations – CG_Hadoop [SIGSPATIAL'13]

MapReduce
Spatial File Splitter
Spatial Record Reader

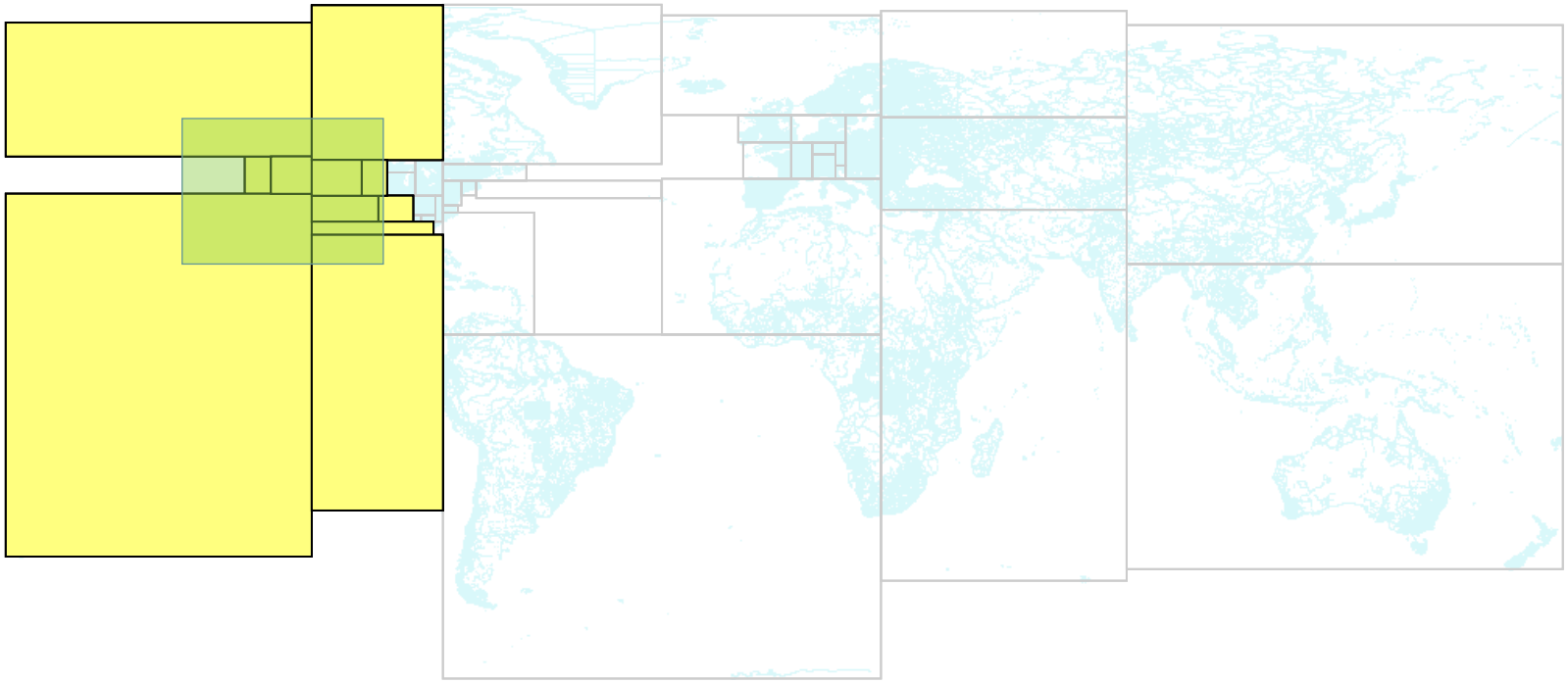
Indexing
Grid – R-tree – R+-tree – Quad tree
[VLDB'15]

ST-Hadoop

Operations Layer

- Basic Operations: e.g, Range query and KNN
- Spatial Join Operations
- Computational geometry operations: e.g., Polygon Union, Voronoi diagram, Delaunay Triangulation, and Convex Hull
- User-defined operations: e.g., kNN join

Range Query



Use **local indexes** to find matching records

Use the **global index** to prune disjoint partitions

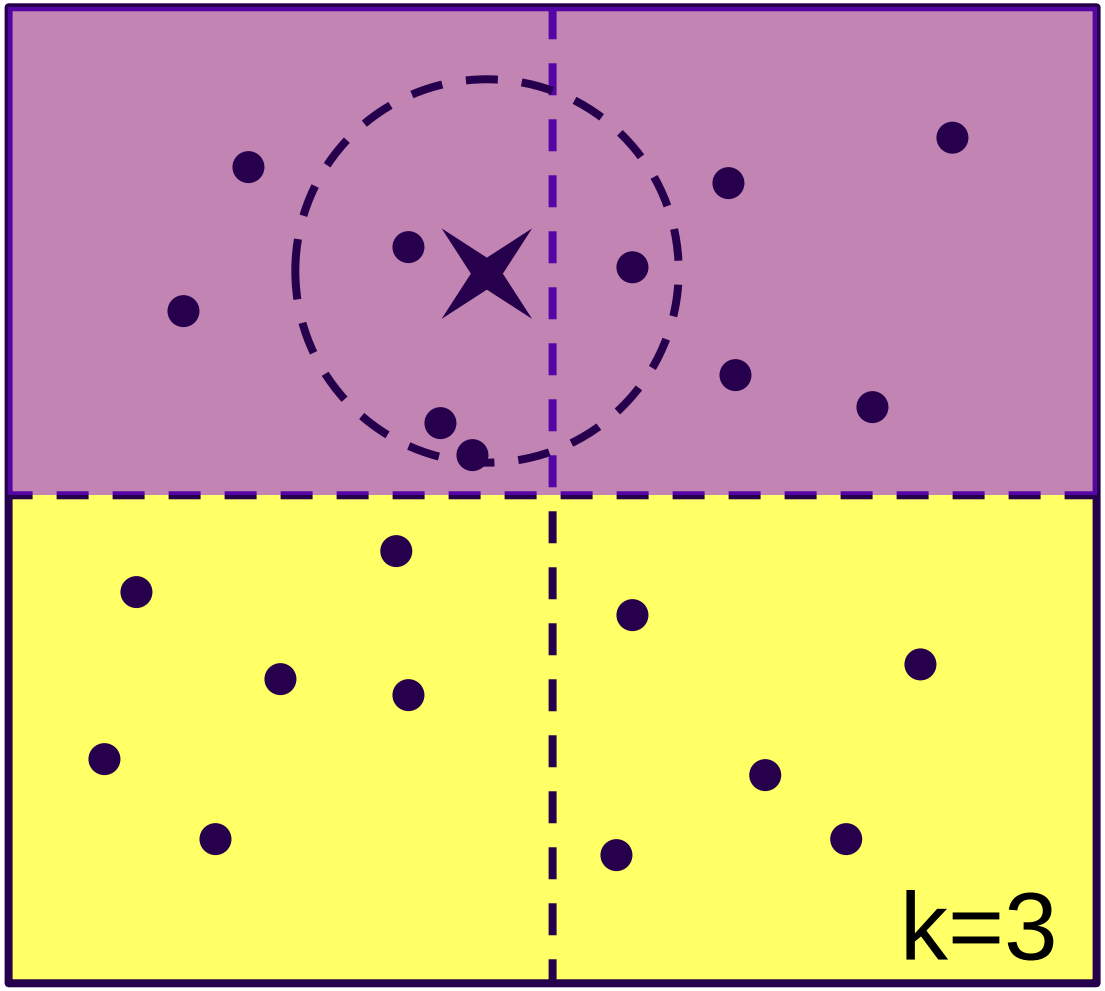
KNN over Indexed Data

First iteration runs as before and result is tested for correctness

✘ Answer is incorrect

Second iteration processes other blocks that might contain an answer

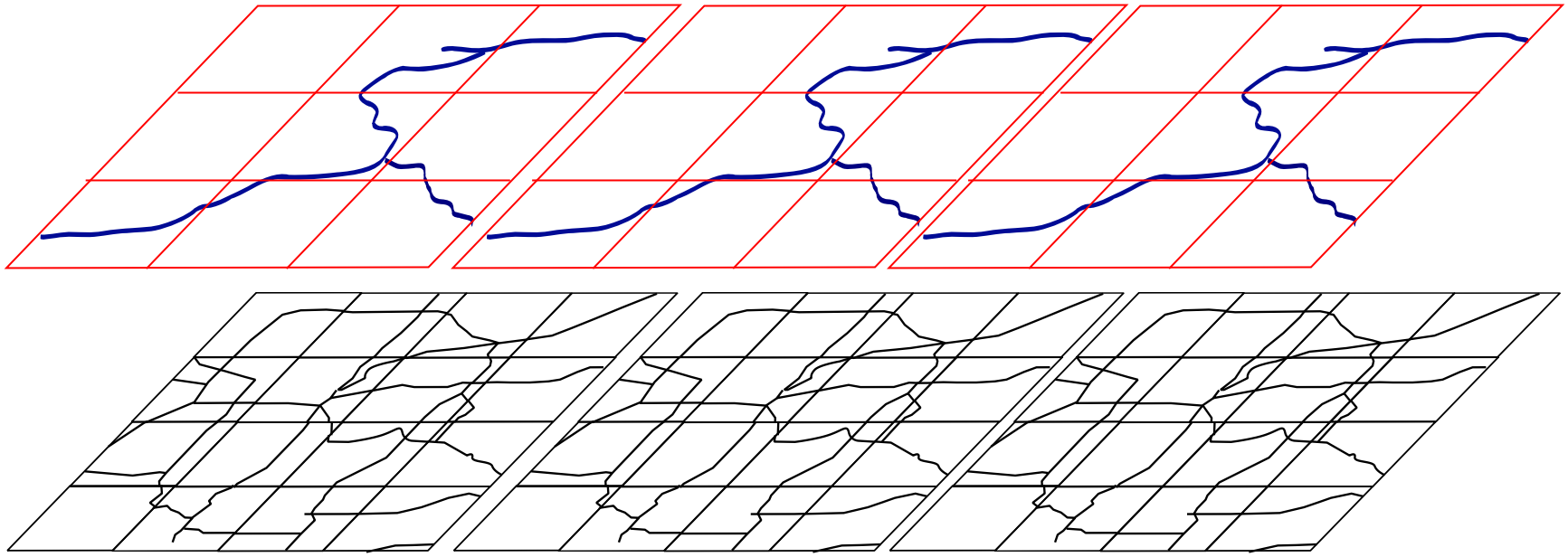
✓ Answer is correct



Spatial Join

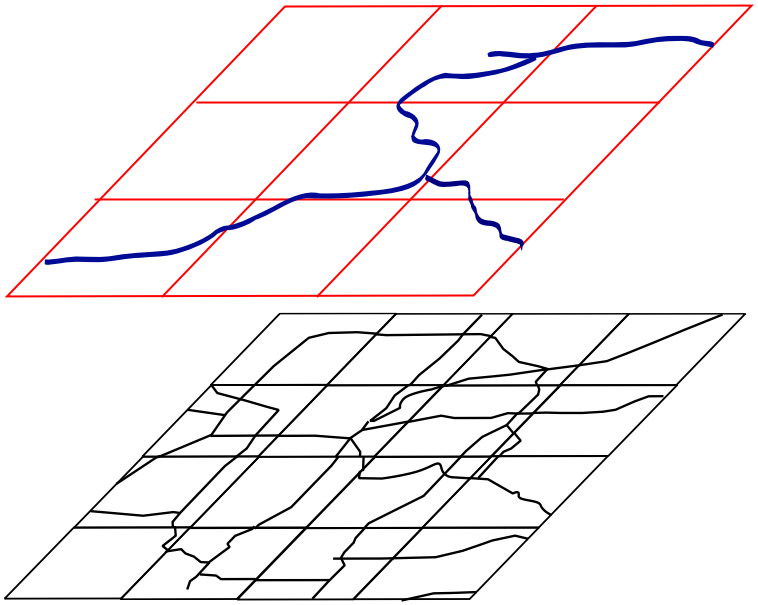
Join Directly

Partition – Join



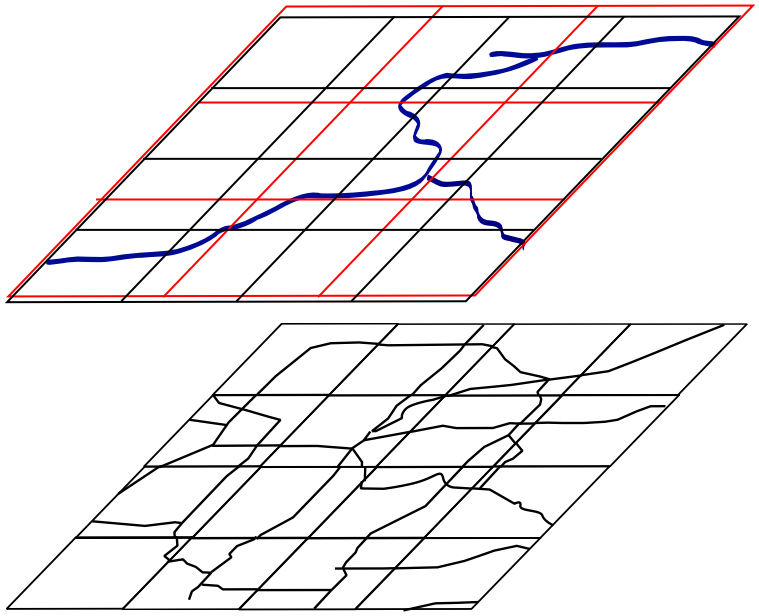
Spatial Join

Join Directly



Total of 36 overlapping pairs

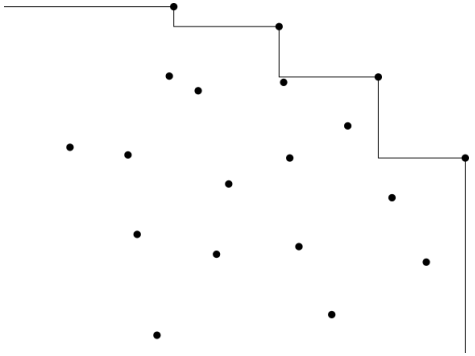
Partition – Join



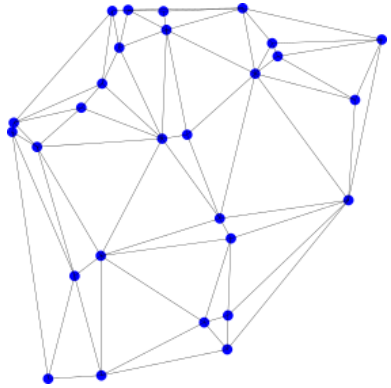
Only 16 overlapping pairs

CG_Hadoop

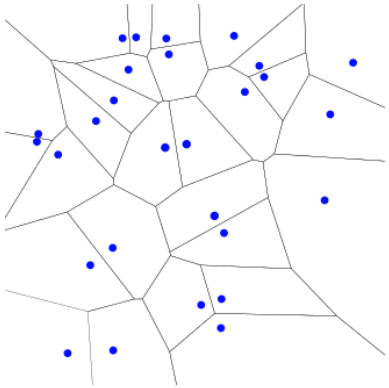
Polygon Union



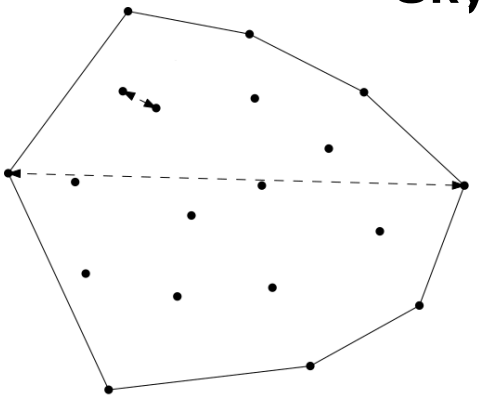
Skyline



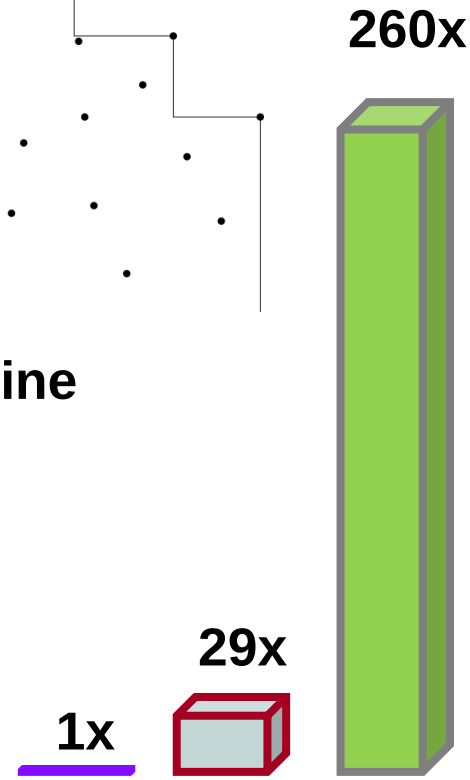
Delaunay Triangulation



Voronoi Diagram



Convex Hull
Farthest/closest pair

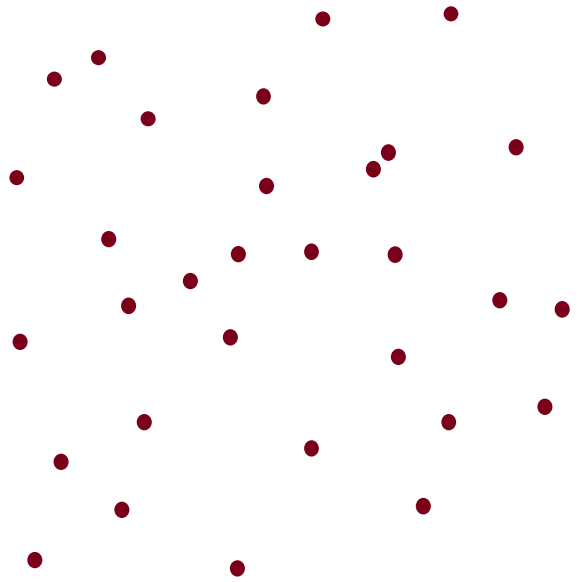


A. Eldawy, Y. Li, M. F. Mokbel, R. Janardan. “CG_Hadoop: Computational Geometry in MapReduce”, **ACM SIGSPATIAL’13**

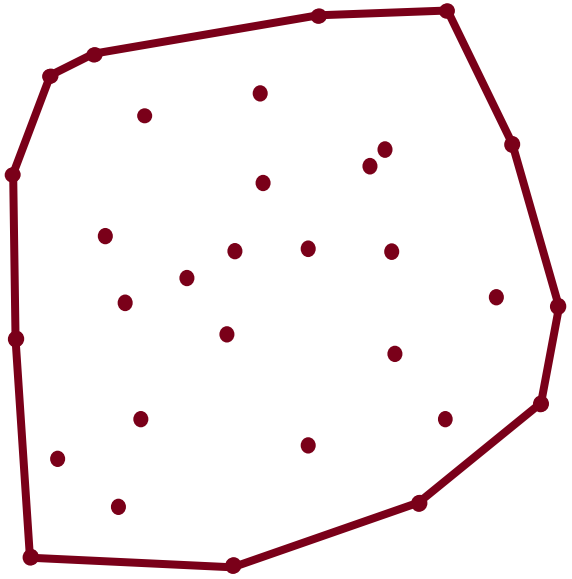
Convex Hull

Find the minimal convex polygon that contains all points

Input



Output



Convex Hull in CG_Hadoop

Hadoop

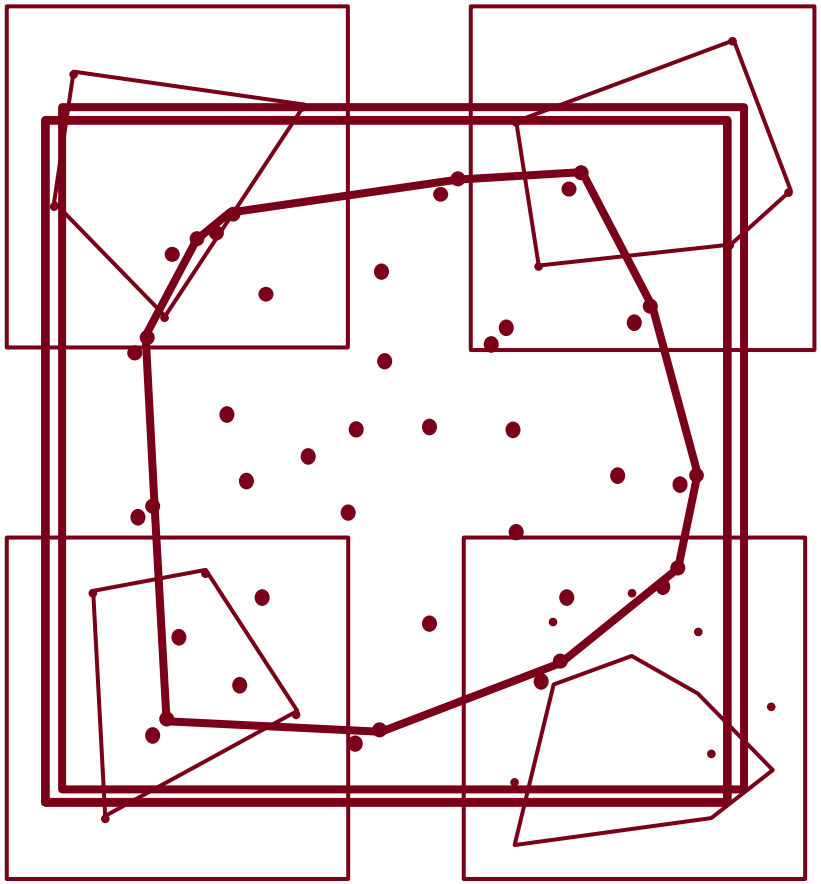
SpatialHadoop

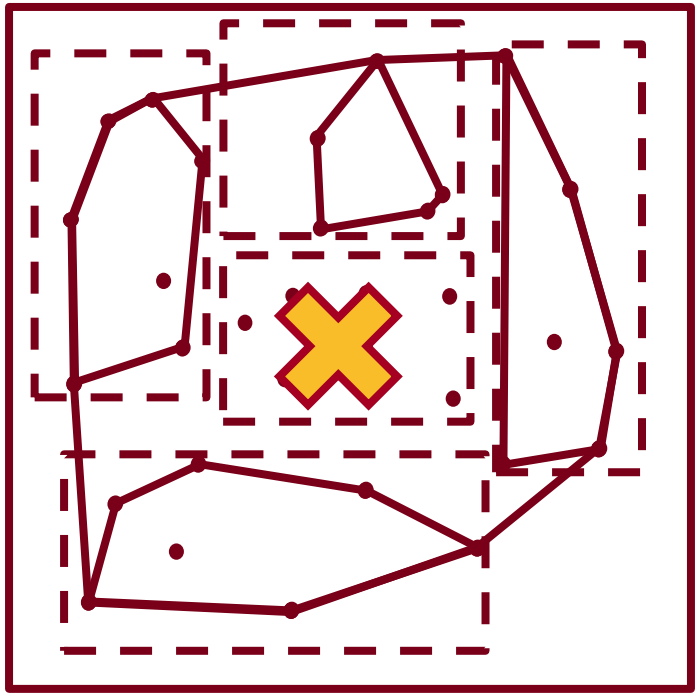
①Partition

②Pruning

③Local hull

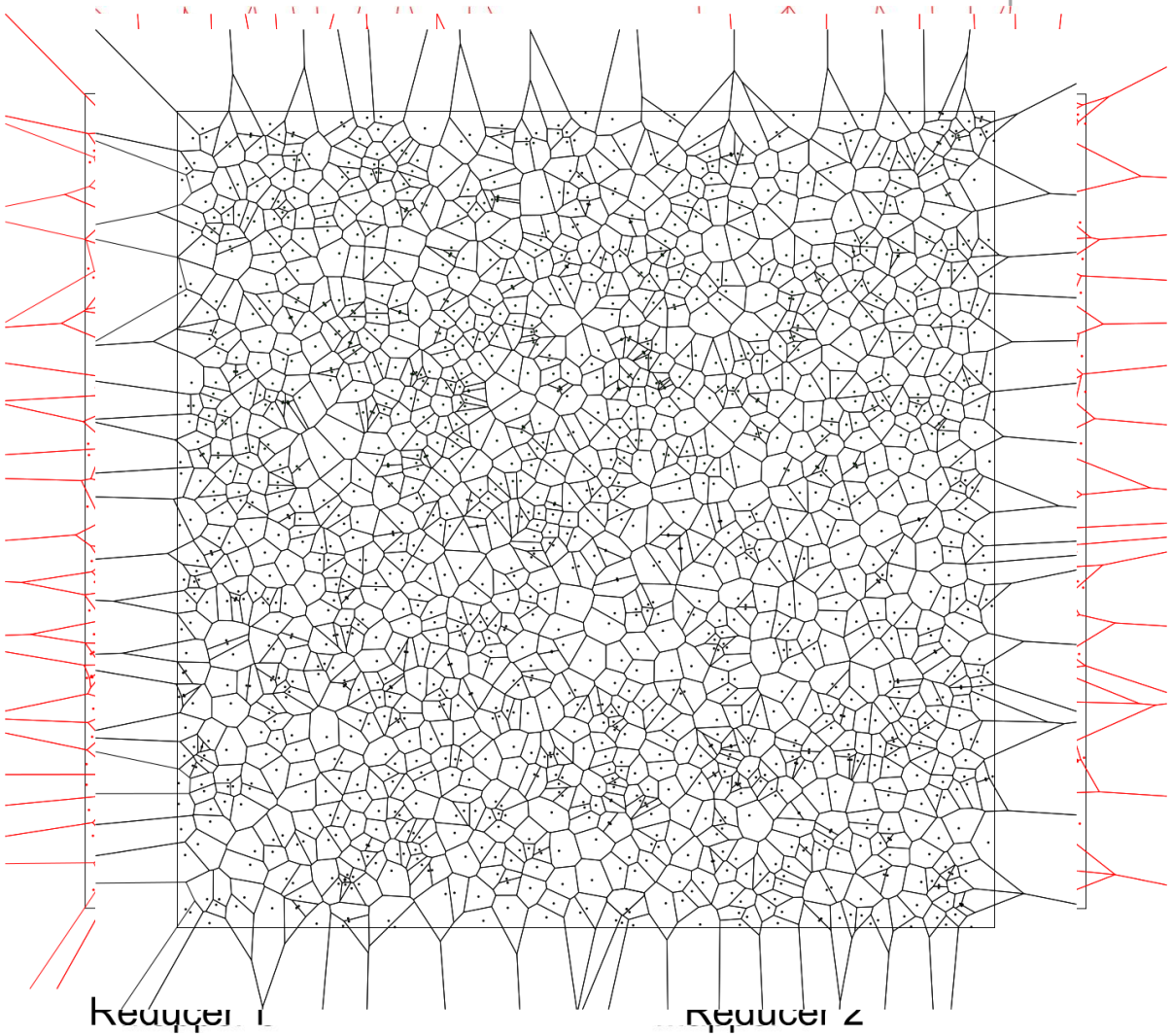
④Global hull





Advanced Analytics

- **Partitioning**
- Local VD
- Pruning
- Vertical Merge
- Pruning
- Horizontal Merge
- Final output



Y. Li, A. et al“Scalable computational geometry in MapReduce”, In VLDBJ 2019

Agenda

- The ecosystem of SpatialHadoop
  - ~~Motivation~~
  - ~~Internal system design~~
  - ~~Applications~~
  - Related work
  - Performance Results
- Interactive data exploration

Other Big Spatial Data Systems

Parallel SECONDO

MD-HBase

GeoSpark

geomesa

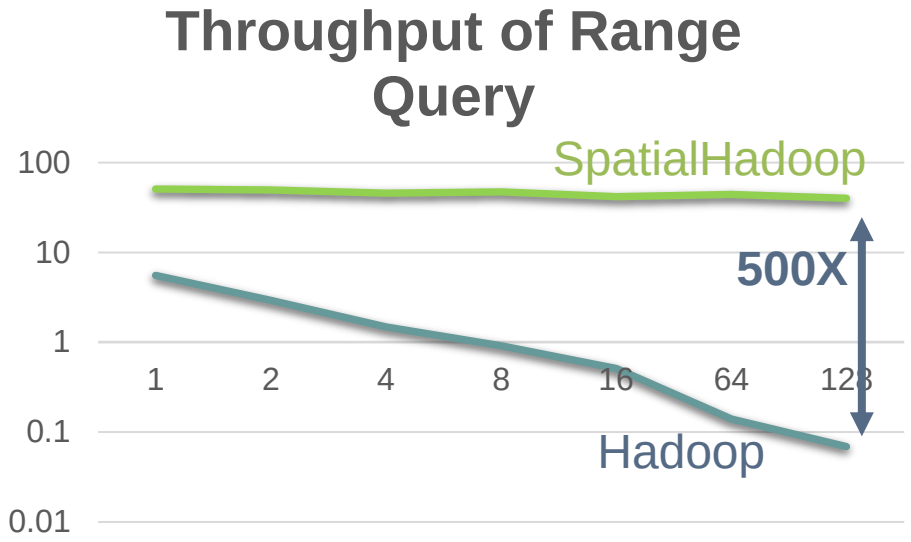
Hadoop-GIS
Spatial Big Data Solutions

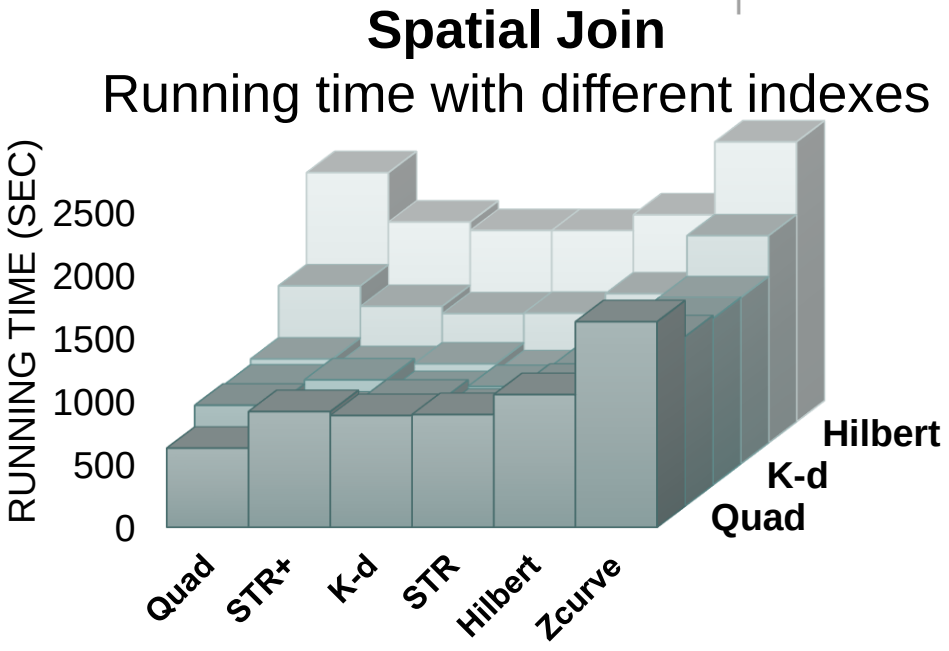
ESRI Tools for Hadoop

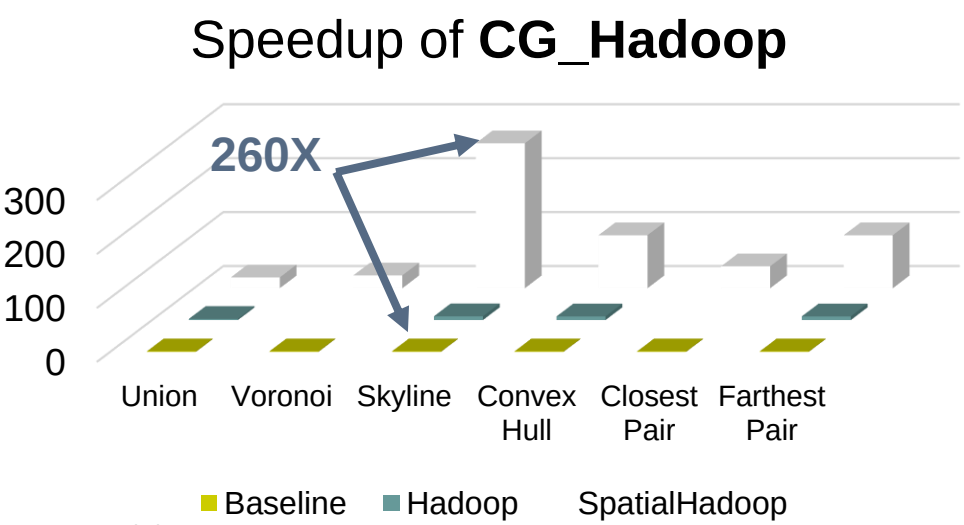
SpatialHadoop is the only extensible system that can be easily expanded by researchers and developers

A. Eldawy and M. Mokbel. “The Era of Big Spatial Data: A Survey”, **Foundations and Trends in Databases 2016**

Performance Results



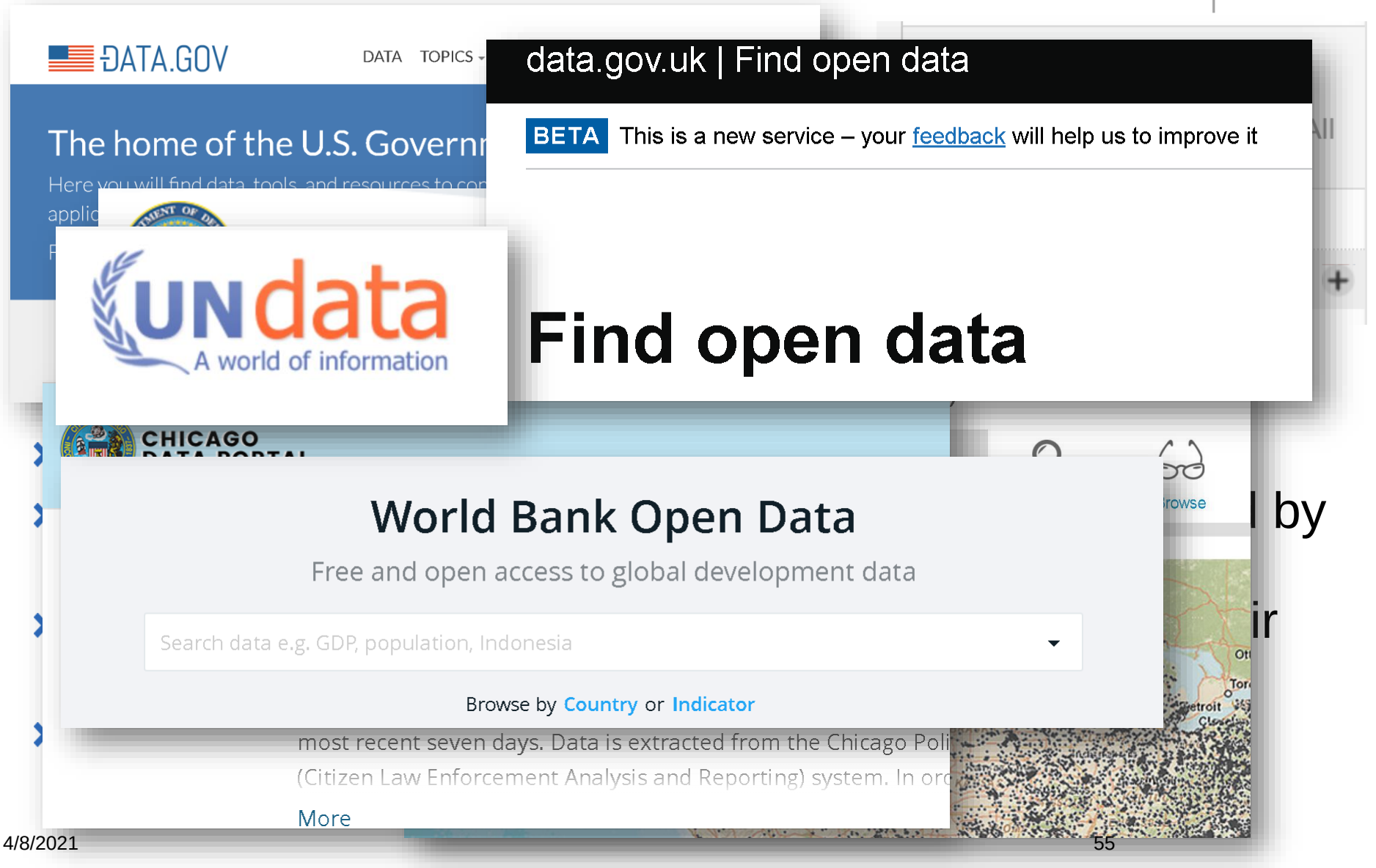




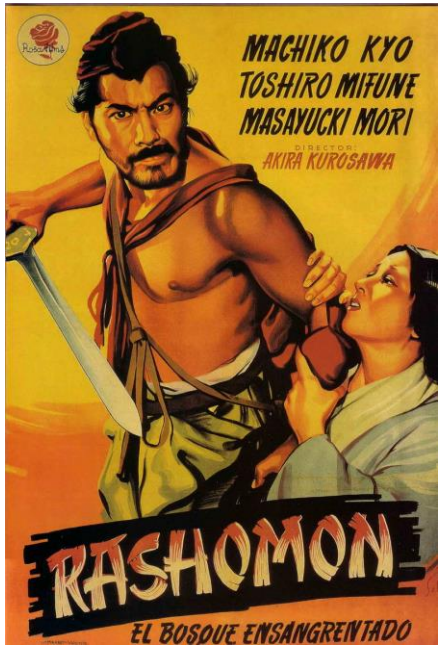
Agenda

- ~~The ecosystem of SpatialHadoop~~
  - ~~Motivation~~
  - ~~System design~~
  - ~~Applications~~
  - ~~Related work~~
  - ~~Performance results~~
- Interactive data exploration

Rise of Big Open Data



Did these data repositories work as expected?

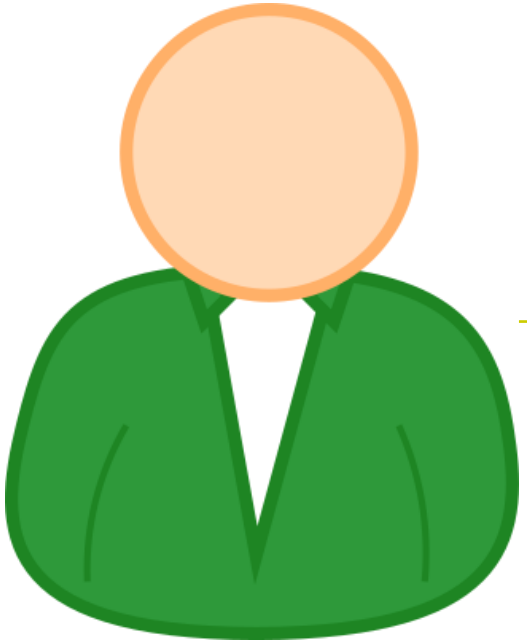


DATA.GOV

DATA TOPICS RESOURCES STRATEGY DEVELOPERS CONT

The home of the U.S. Government's open data

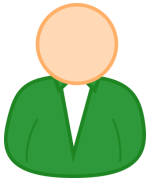
Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.



Helped the police department publish their data on Data.gov

The Politician

The Politician



Published a paper

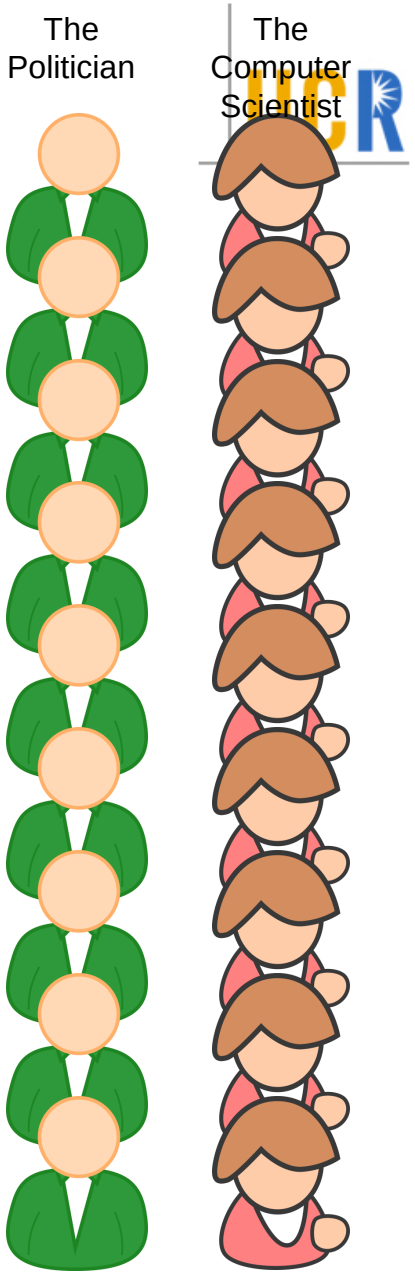
Developed new spatial analytics algorithms to study event data

Released the code

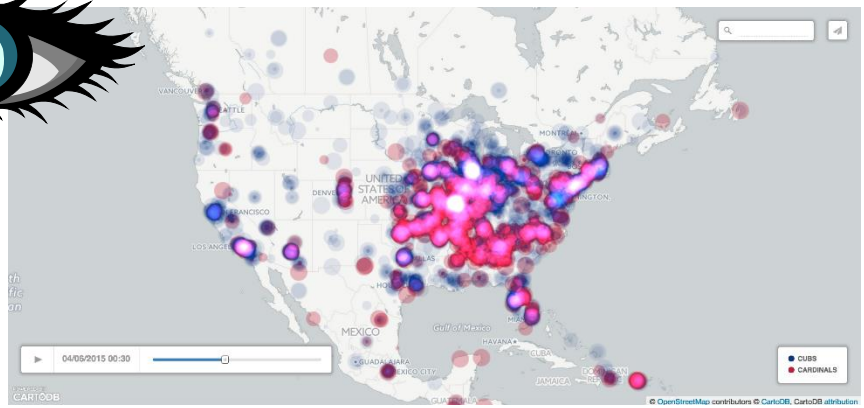
The Computer Scientist

Want to study the relationship between crimes, demographics, and homelessness using data science

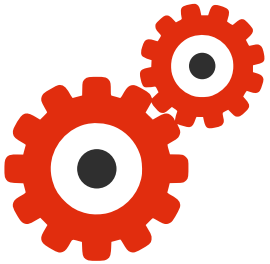
The Domain Scientist



Exploratory Analytics

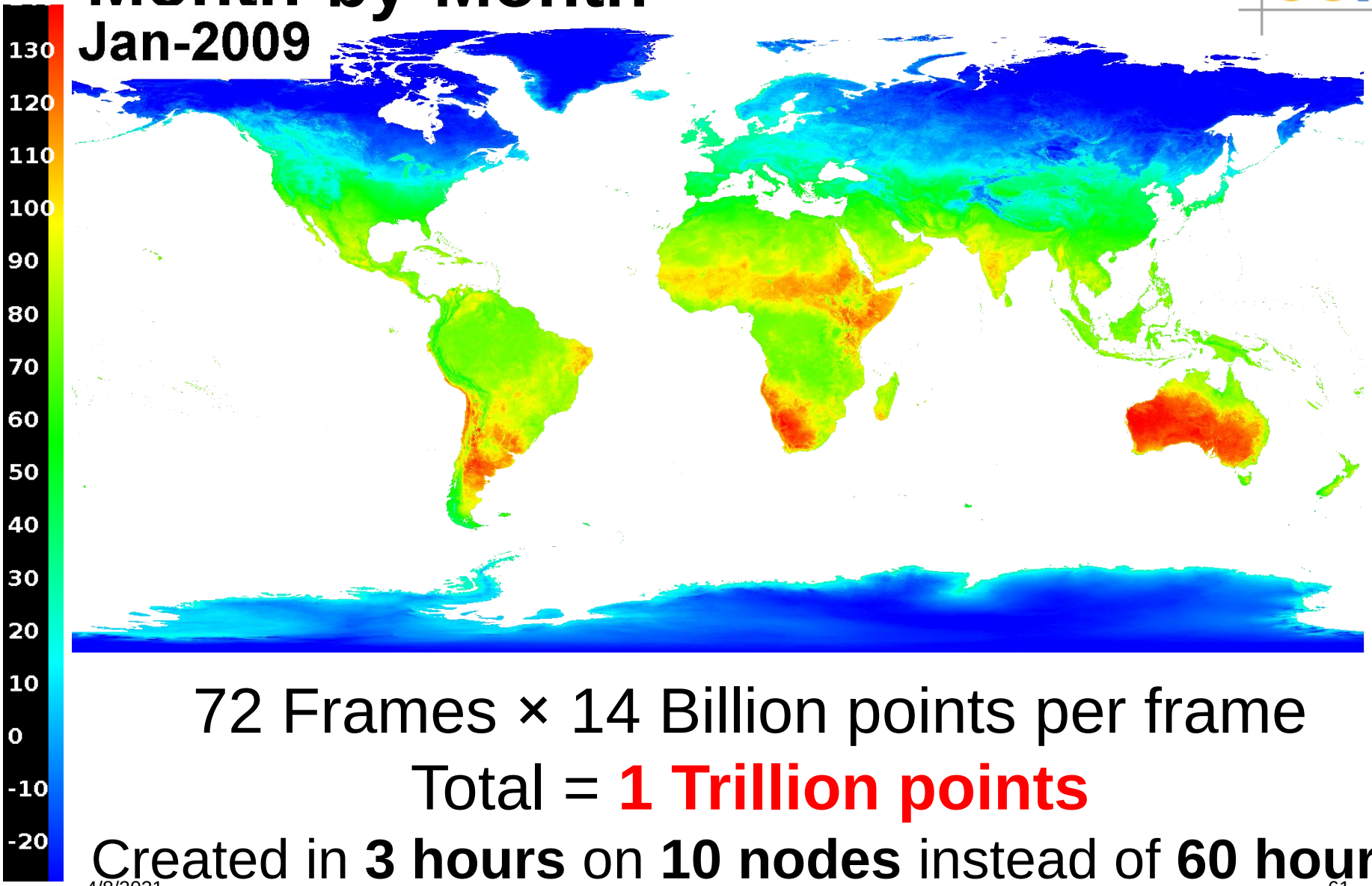


0.5 – 5.0 seconds



Big Spatial Data

Heat Map From 2009 to 2014 Month-by-Month



72 Frames × 14 Billion points per frame

Total = **1 Trillion points**

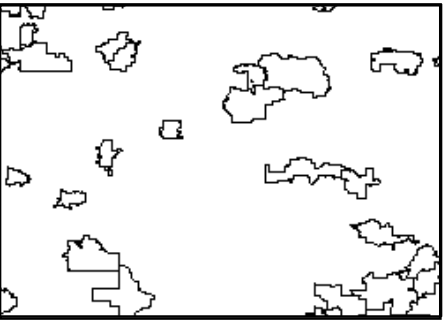
Created in **3 hours** on **10 nodes** instead of **60 hours**

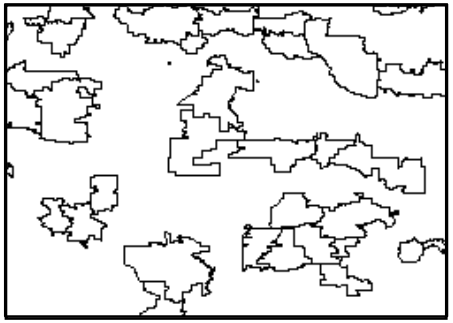
Single Level Image

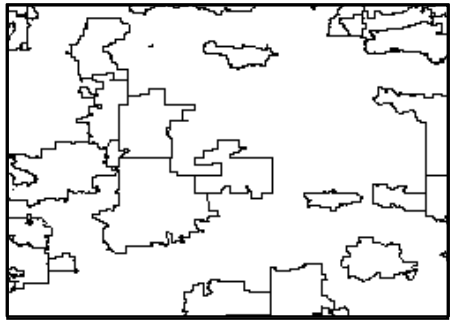
Input

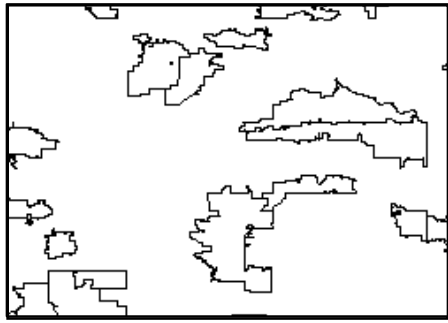
Split | Split | Split | Split

createRaster | createRaster | createRaster | createRaster



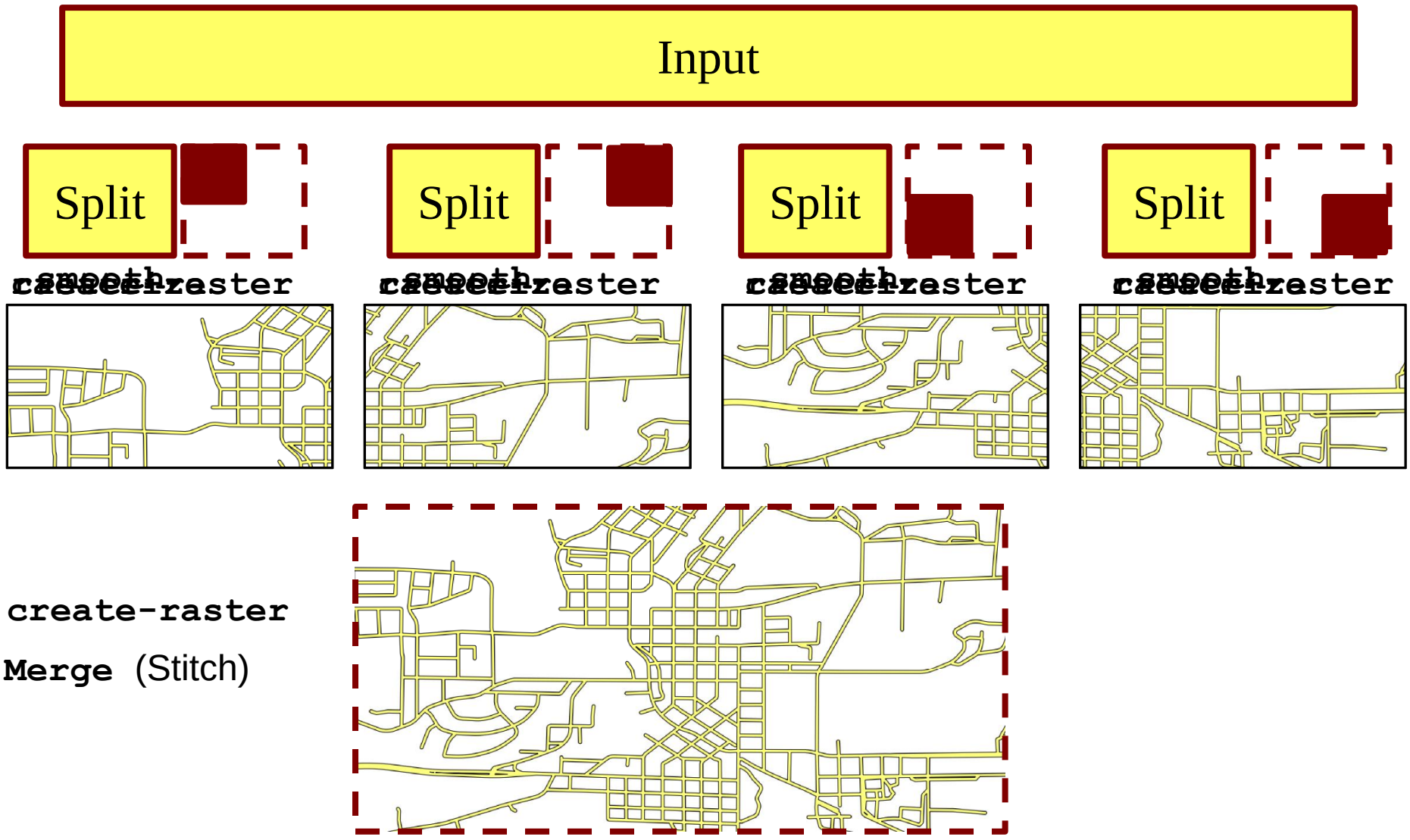




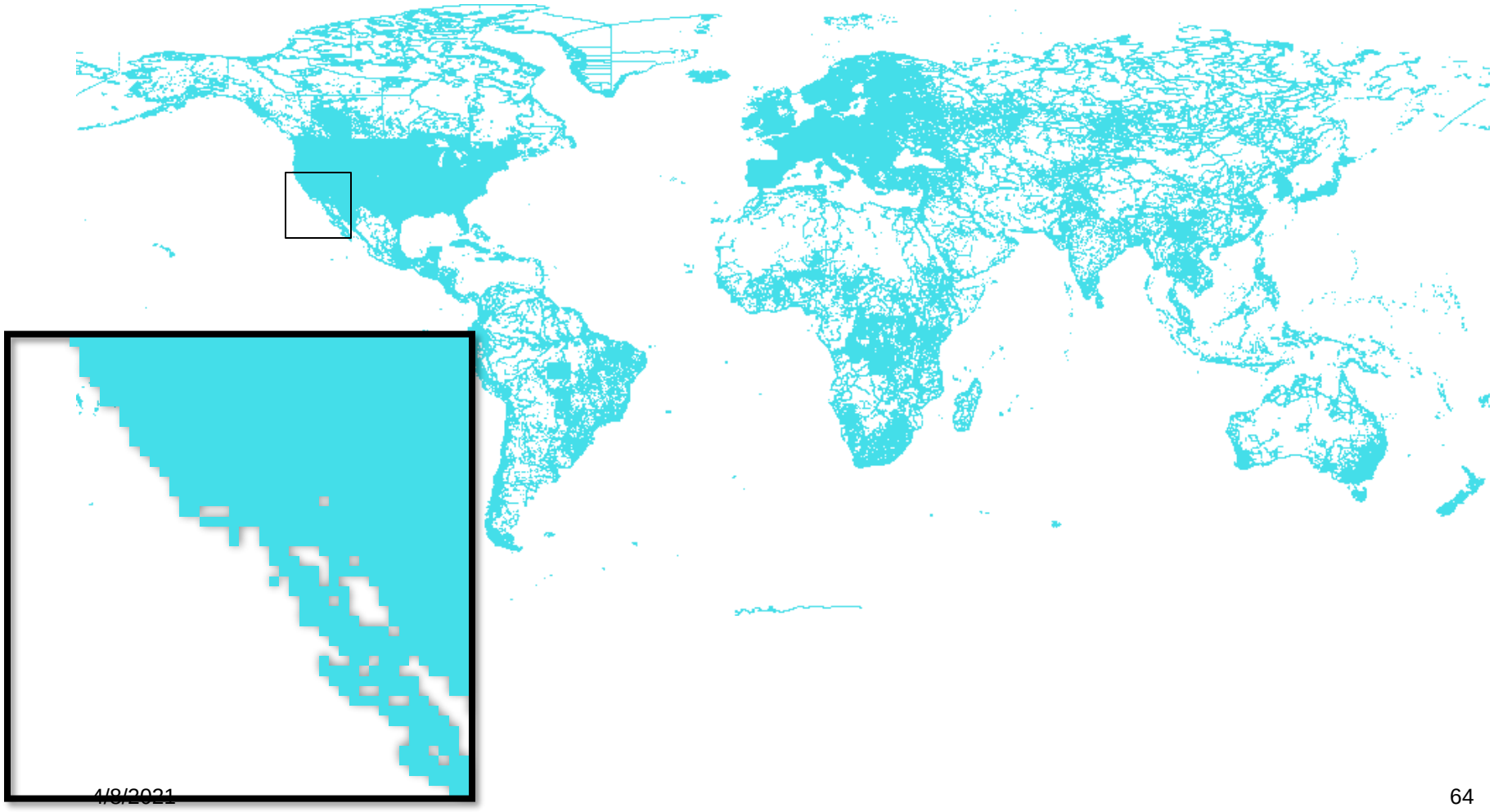


Merge
(Overlay)

Space Partitioning



Level of Details



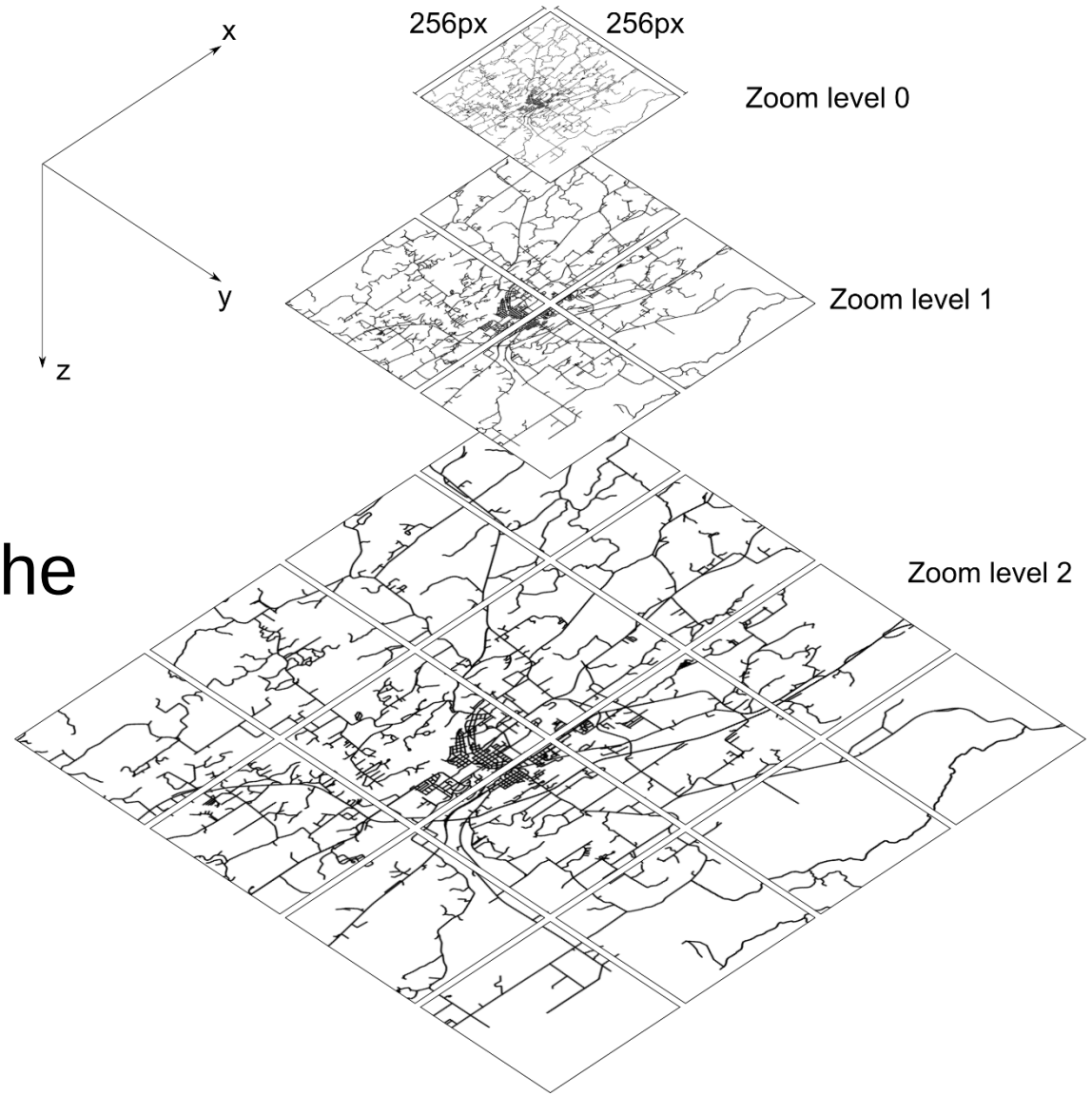
Multilevel Visualization

Map of California – 2GB

Generated in 2 minutes on 10-node cluster using SpatialHadoop instead of one hour

Multi-level Image

- Many images at different zoom levels
  - Pan
  - Zoom in/out
  - Fly to
- More details as the zoom level increases
- Number of tiles increases exponentially



Adaptive Multilevel Visualization

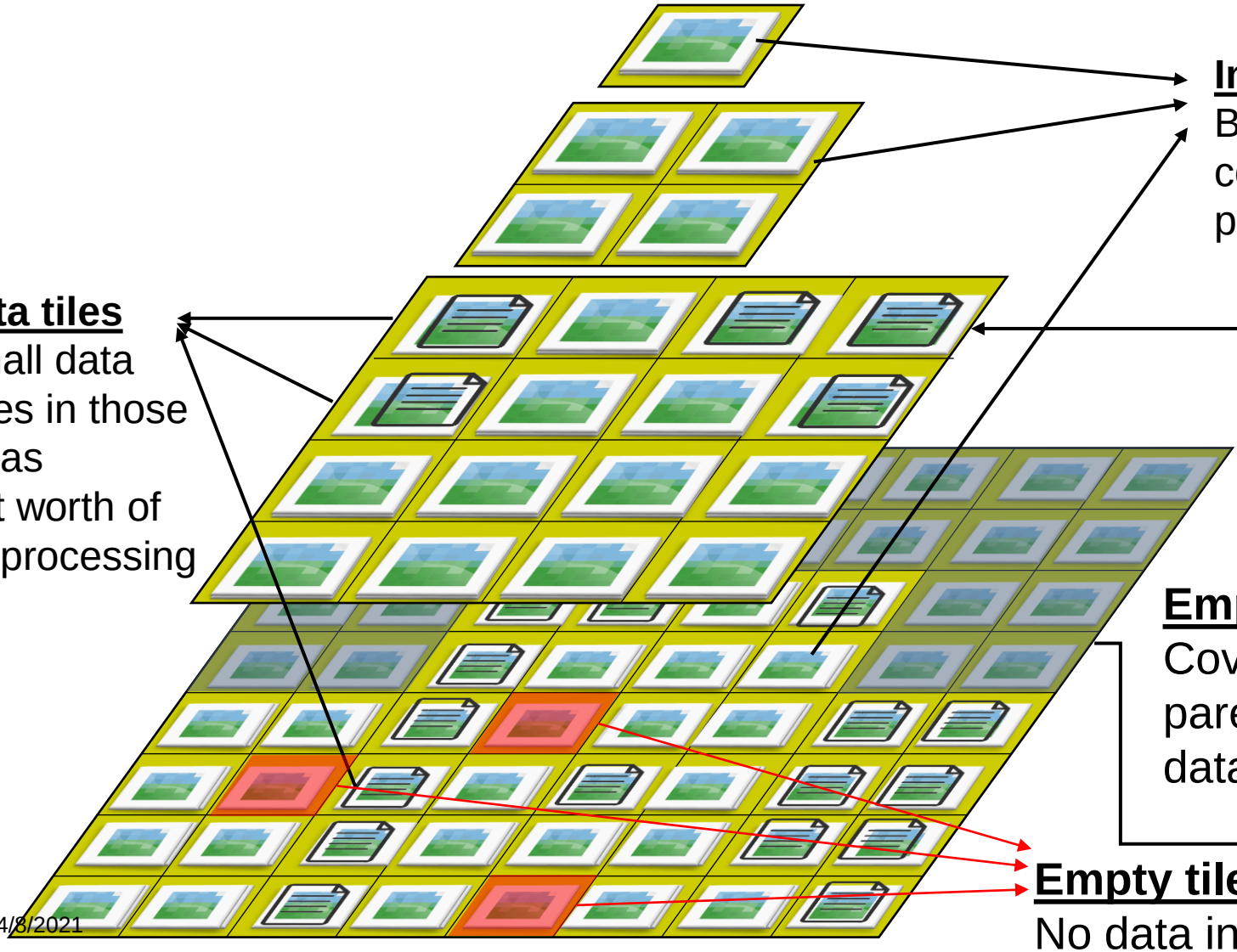


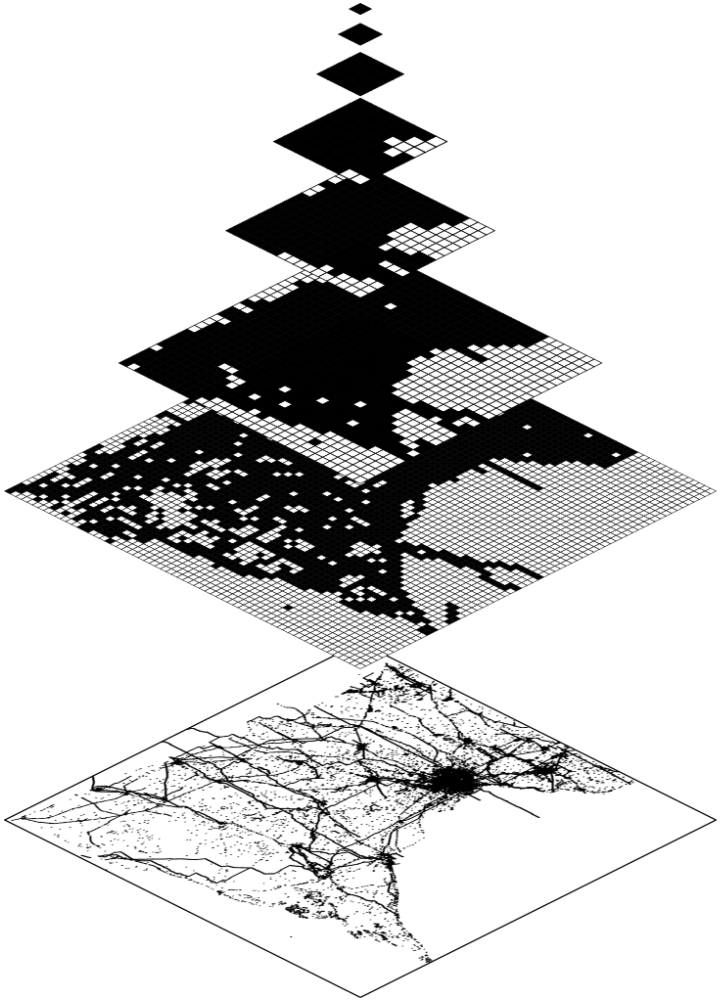
Image tiles
Big enough to consider preprocessing

Data tiles
Small data sizes in those areas
Not worth of preprocessing

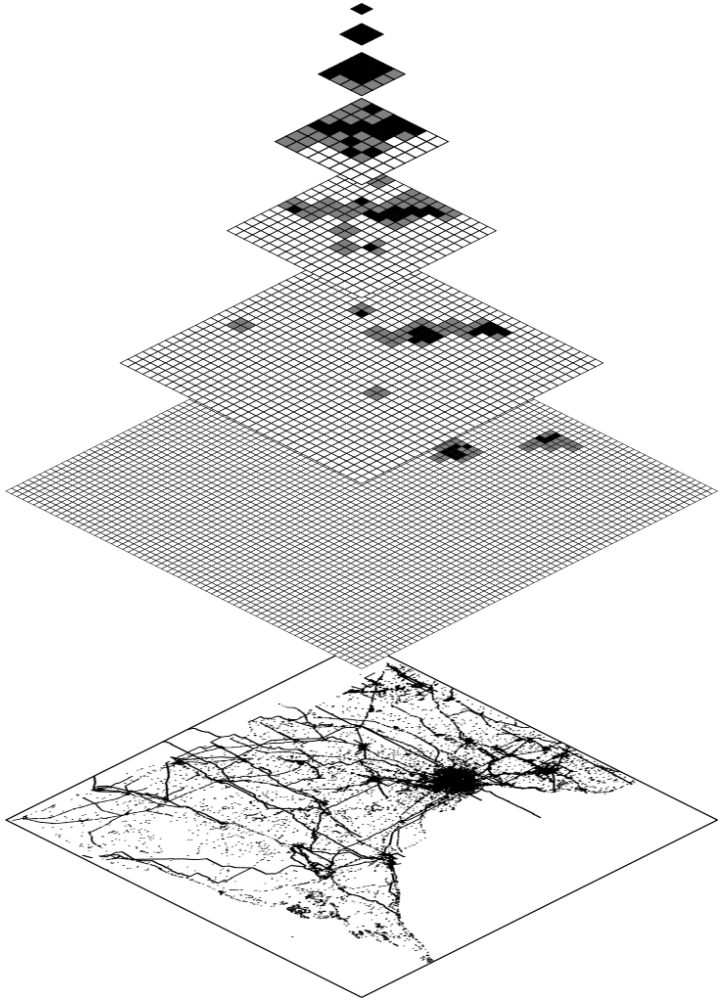
Empty tiles
Covered by parent data tiles

Empty tiles
No data in there

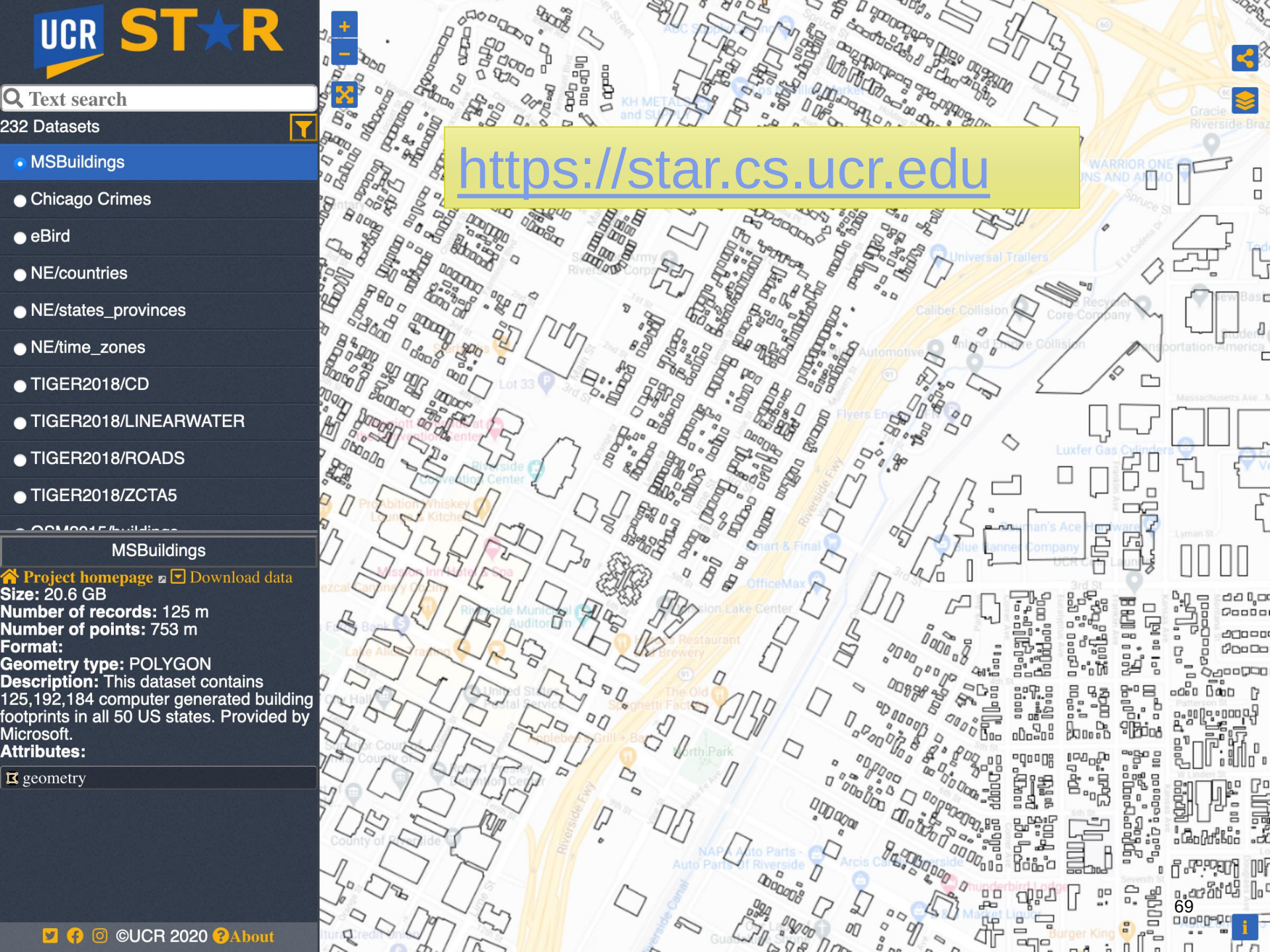
Adaptive Multilevel Images



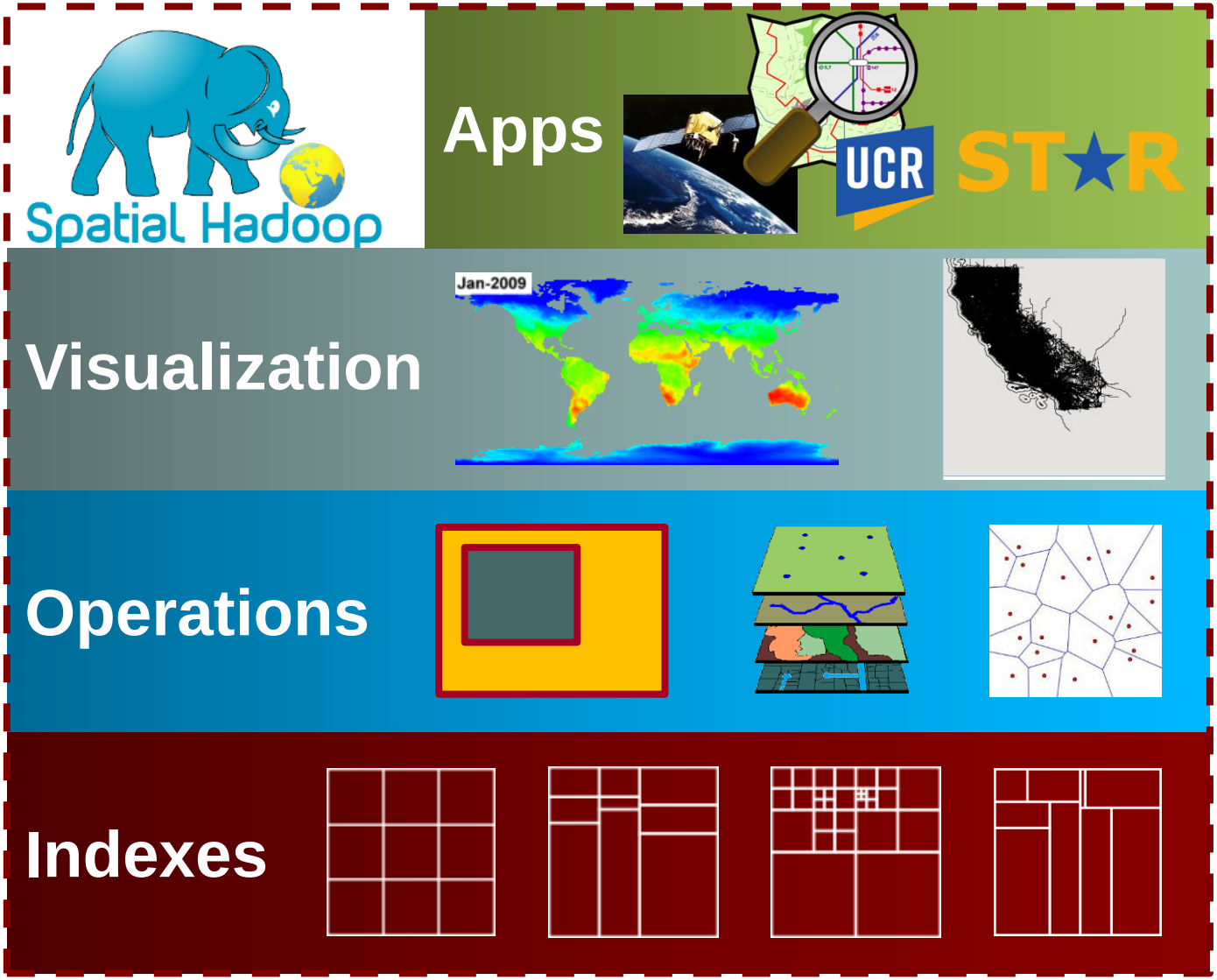
Full Image (3,160 tiles)



Adaptive Image (231 files)


UCR STAR
Text search
232 Datasets
MSBuildings
Chicago Crimes
eBird
NE/countries
NE/states_provinces
NE/time_zones
TIGER2018/CD
TIGER2018/LINEARWATER
TIGER2018/ROADS
TIGER2018/ZCTA5
MSBuildings
Project homepage
Download data
Size: 20.6 GB
Number of records: 125 m
Number of points: 753 m
Format:
Geometry type: POLYGON
Description: This dataset contains 125,192,184 computer generated building footprints in all 50 US states. Provided by Microsoft.
Attributes:
geometry
©UCR 2020
About
https://star.cs.ucr.edu

Summary

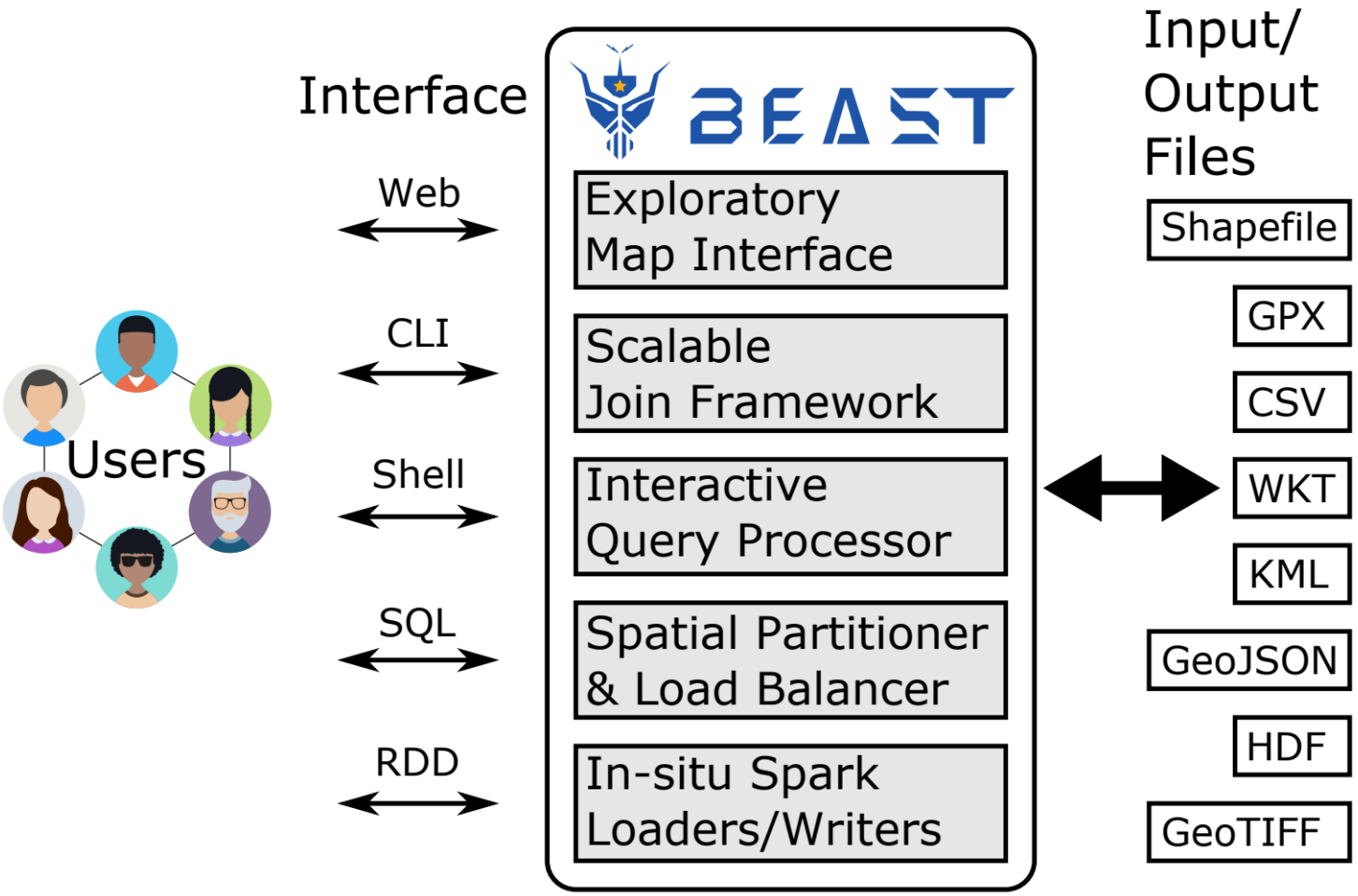

Spatial Hadoop
Apps
UCR
ST★R
Visualization
Jan-2009
Operations
Indexes

Future Work

- More interaction in visualization
  - Non-spatial filters, e.g., temporal
  - Select individual objects on the map (on big data of course)
- Approximate query processing on big spatial data
- Utilize machine learning for query optimization and better user experience
- Support incremental updates to datasets

Beast

- Big Exploratory Analytics on Spatio-Temporal Data



Thank You

Questions?