Big Data and Geospatial with HPCC Systems®

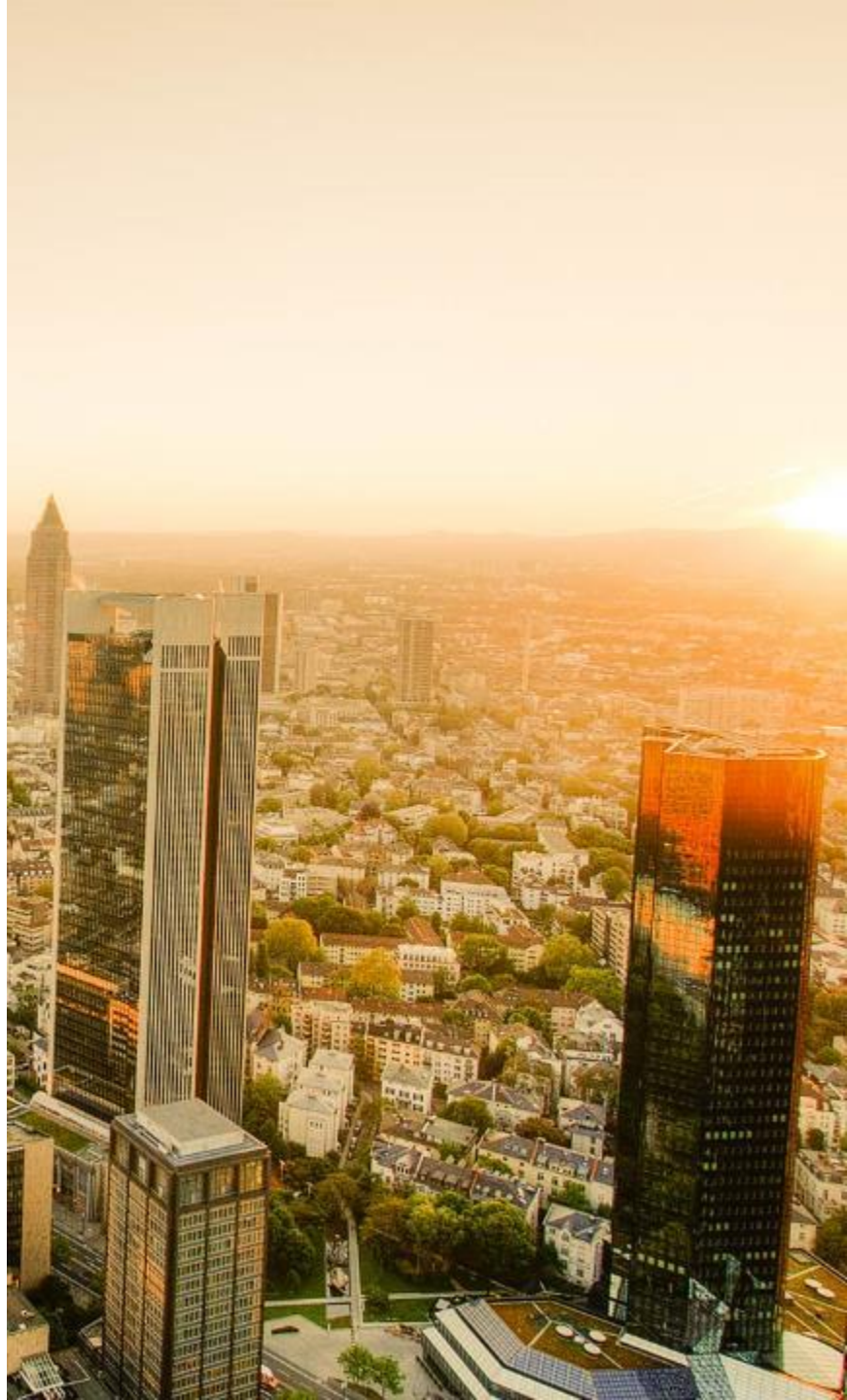
Powered by LexisNexis Risk Solutions

Ignacio Calvo
Greg McRandal

10/05/2016

- Concepts in Geospatial
- How to use them with HPCC
- Use cases
- @HPCCSystems

HPCC SYSTEMS®



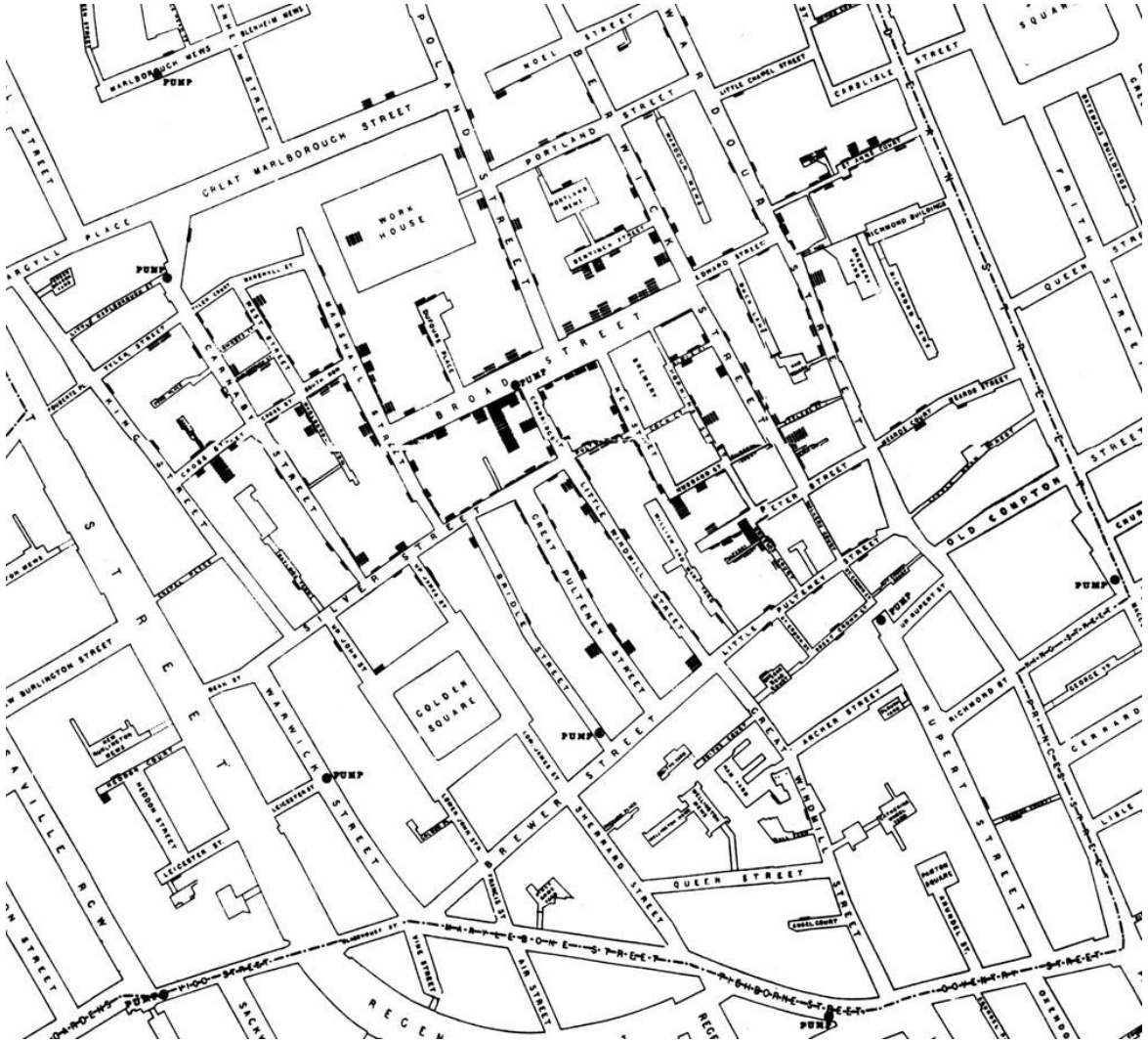
Definition

An approach to applying statistical analysis and other analytic techniques to data which has a geographical or spatial aspect

HPCC SYSTEMS®

HPCC SYSTEMS®

Origin of Geospatial



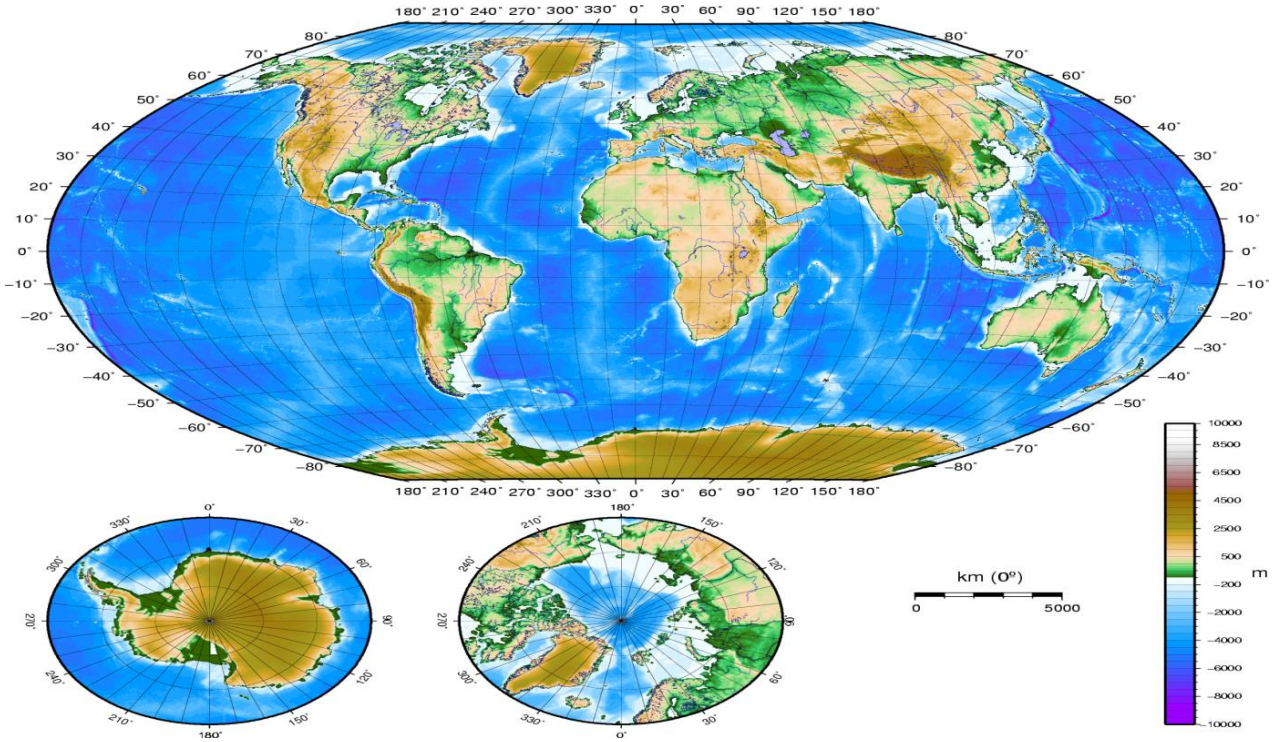
John Snow's original map (1854), using GIS to save lives. This map was used to determine that Cholera was water-borne

HPCC SYSTEMS®

Understanding the data

Need to know :

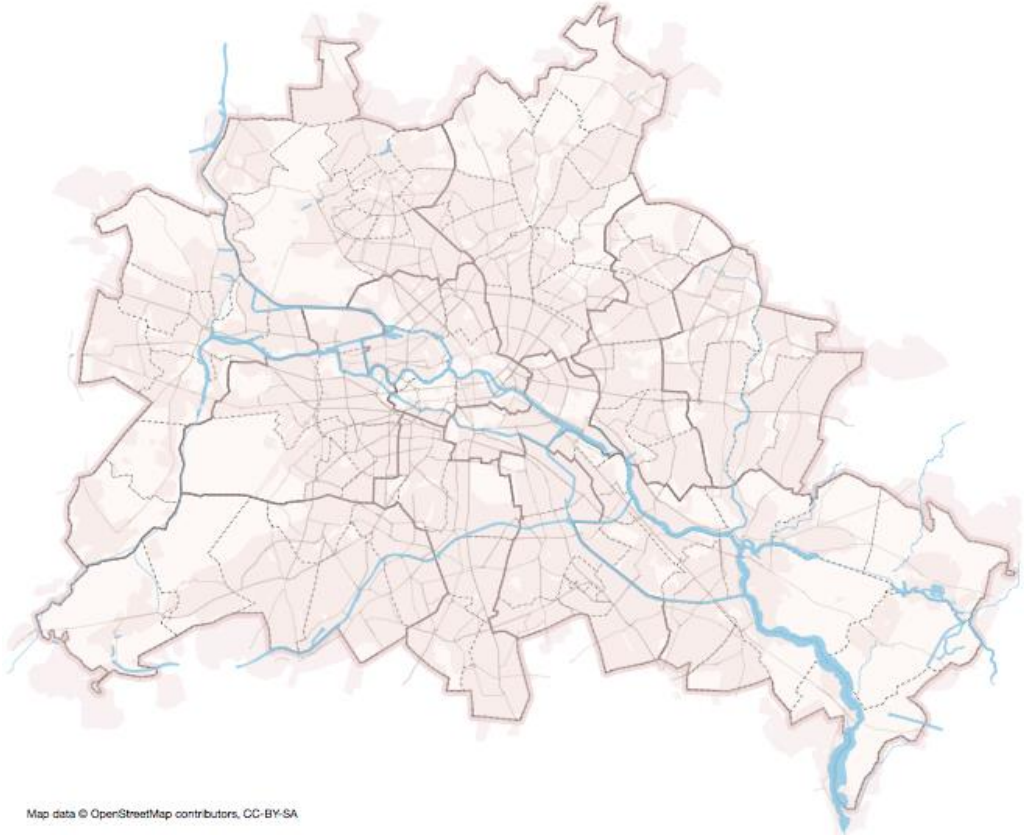
- Format
- Projection / coordinate system



HPCC SYSTEMS®

Formats : Vector vs Raster

Vector



Raster

HPCC SYSTEMS®

What is a projection?

Projections are used to represent the world in ways we can process

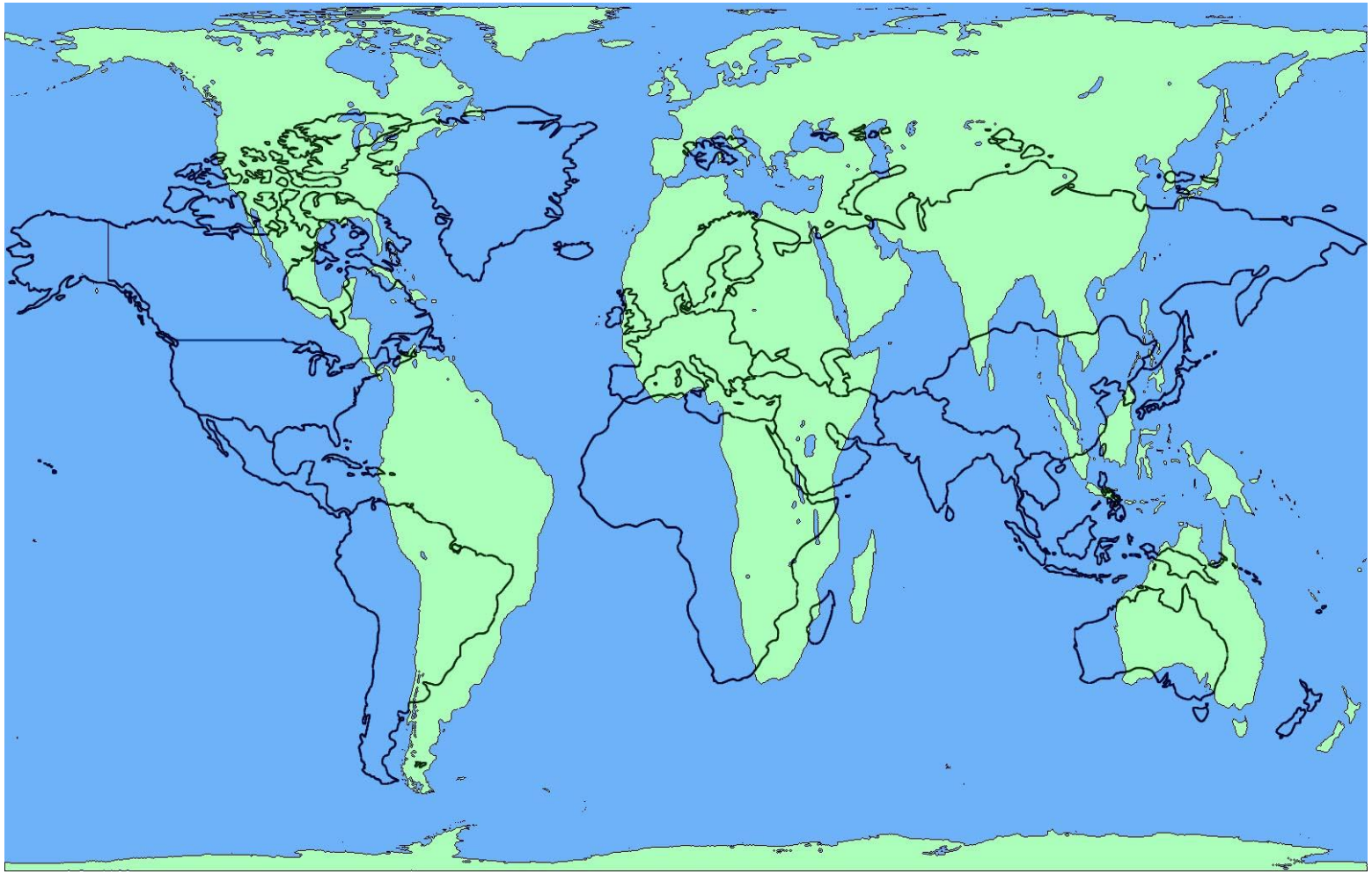
- The Earth is round and maps are flat
- Physical Maps
- Computer Maps

Have I seen projections before?

- Peter vs Mercator vs Winkel tripel
- GPS (latitude/longitude)
- Google Maps

Projections

Two different projections representing the same place.



HPCC SYSTEMS®

Projections to know about

WGS84

- Latitude and longitude
- Our best approximation of the world
- Not always the best for a specific region
- Not technically a projection

Mercator

- Many different ones, choose one based on your location
- Reduces the area it covers to a simple Cartesian plane
- Good near the central axis, bad far away from it :
    - Web Mercator covers the whole world – good near equator, gets worse as you travel north or south
    - Irish National Grid – very good for Ireland, awful anywhere else.

HPCC SYSTEMS®

Lies, damned lies, statistics… and maps!

Tom Phillips
@flashboy

Lovely bit of work on how map projections are weird. Greenland is just taking the piss. thetruesize.com

*https://twitter.com/flashboy/status/641221733509373952

HPCC SYSTEMS®

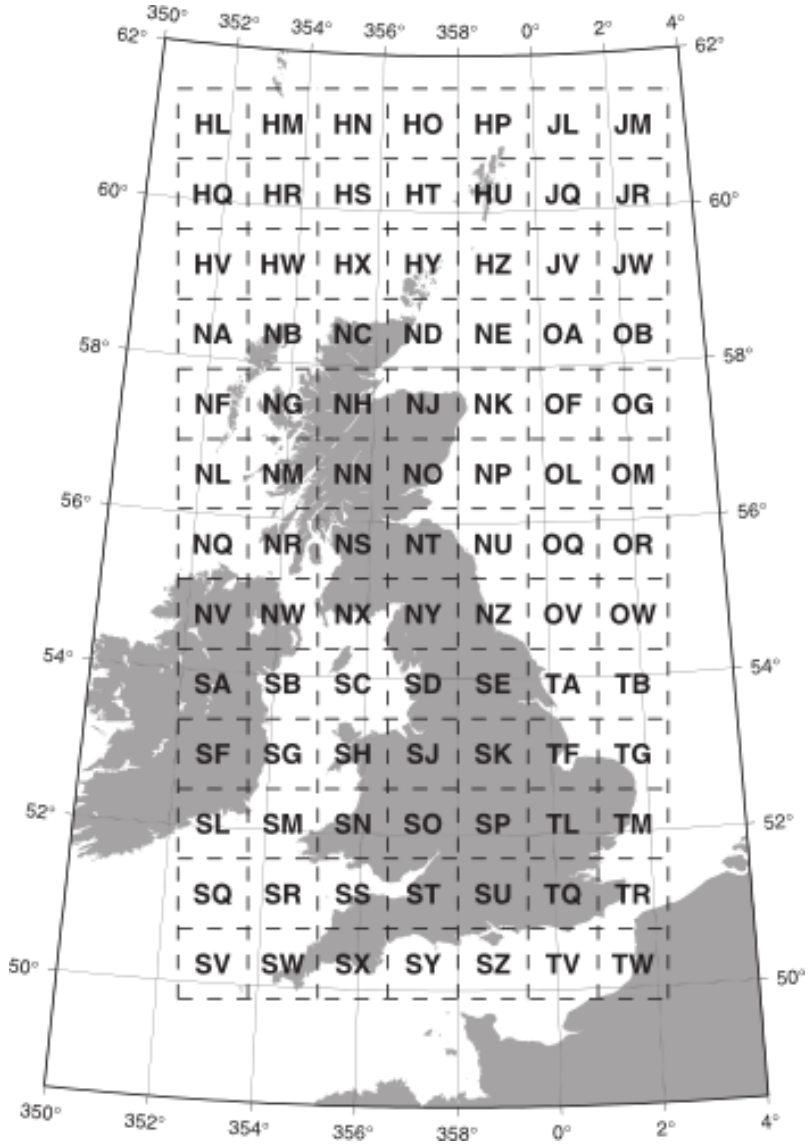
Lies, damned lies, statistics… and maps!

Projection Woes:

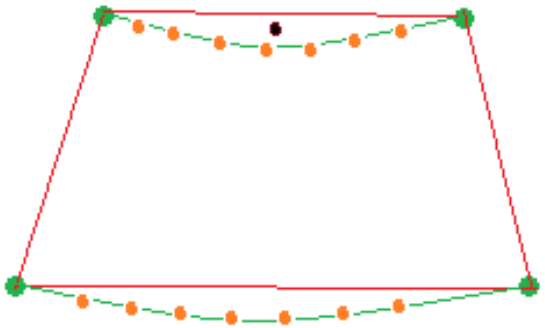
A straight line in Mercator is not a straight line in WGS84

Don’t re-project polygons!

This “solution” is only good enough for visuals, not for maths.



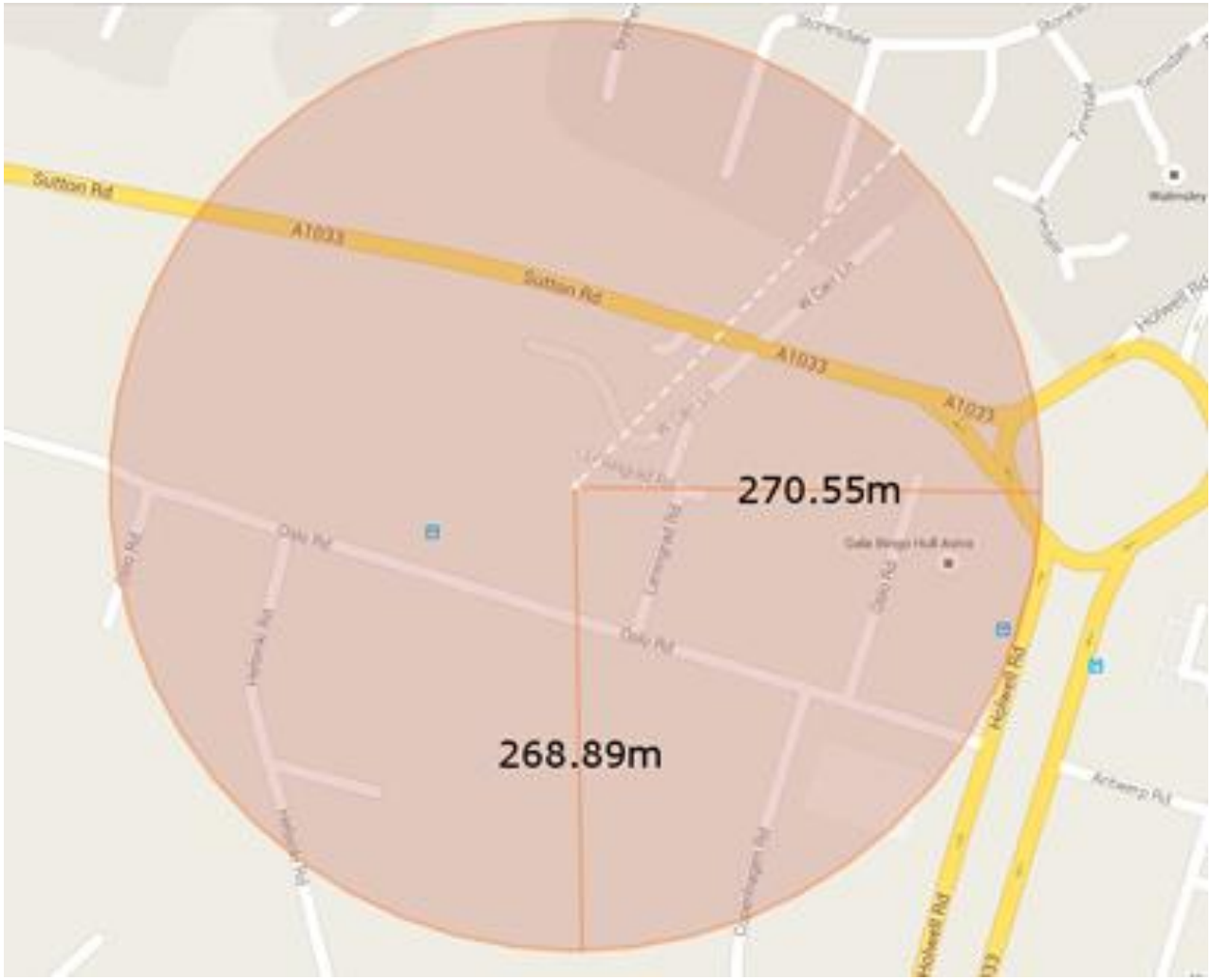
Four points converted to WGS84



Where the lines should be

HPCC SYSTEMS®

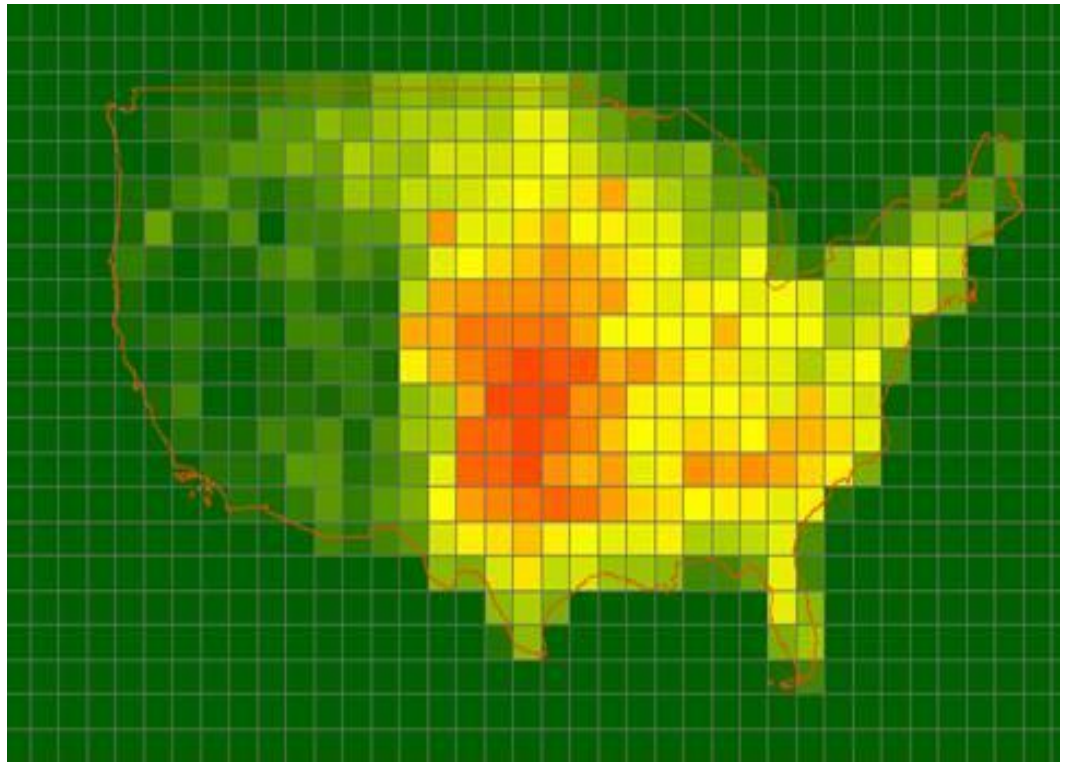
Lies, damned lies, statistics… and maps!



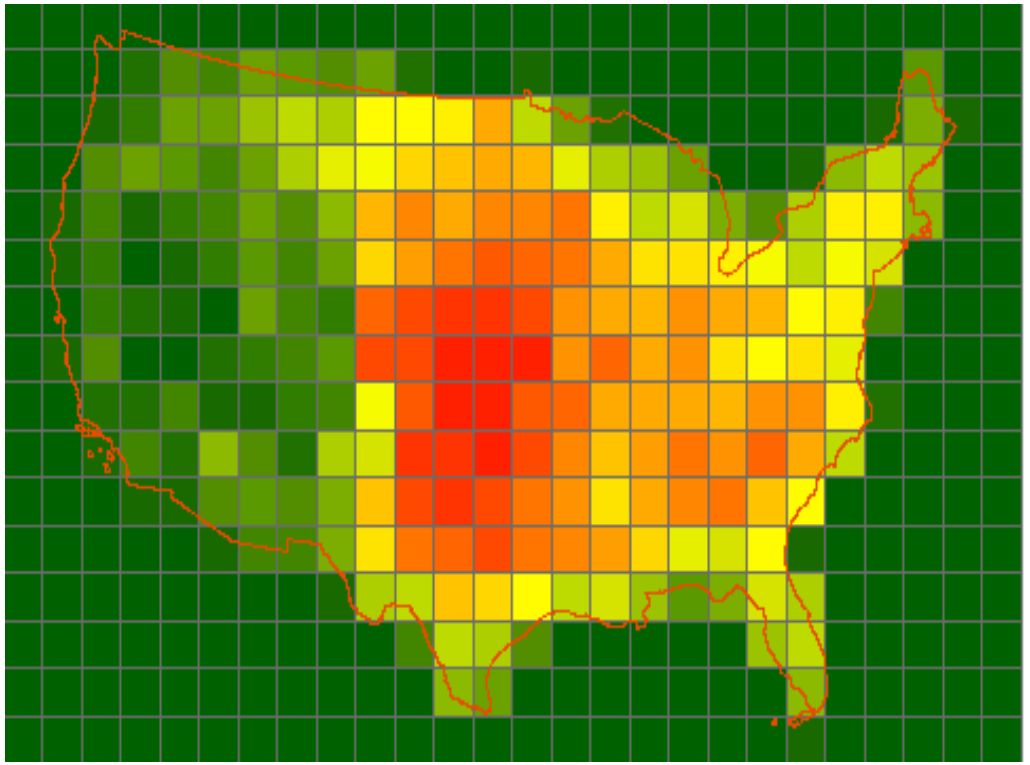
Lies, damned lies, statistics… and maps!

Visuals don’t agree with maths: Wind and Hail.

Web Mercator

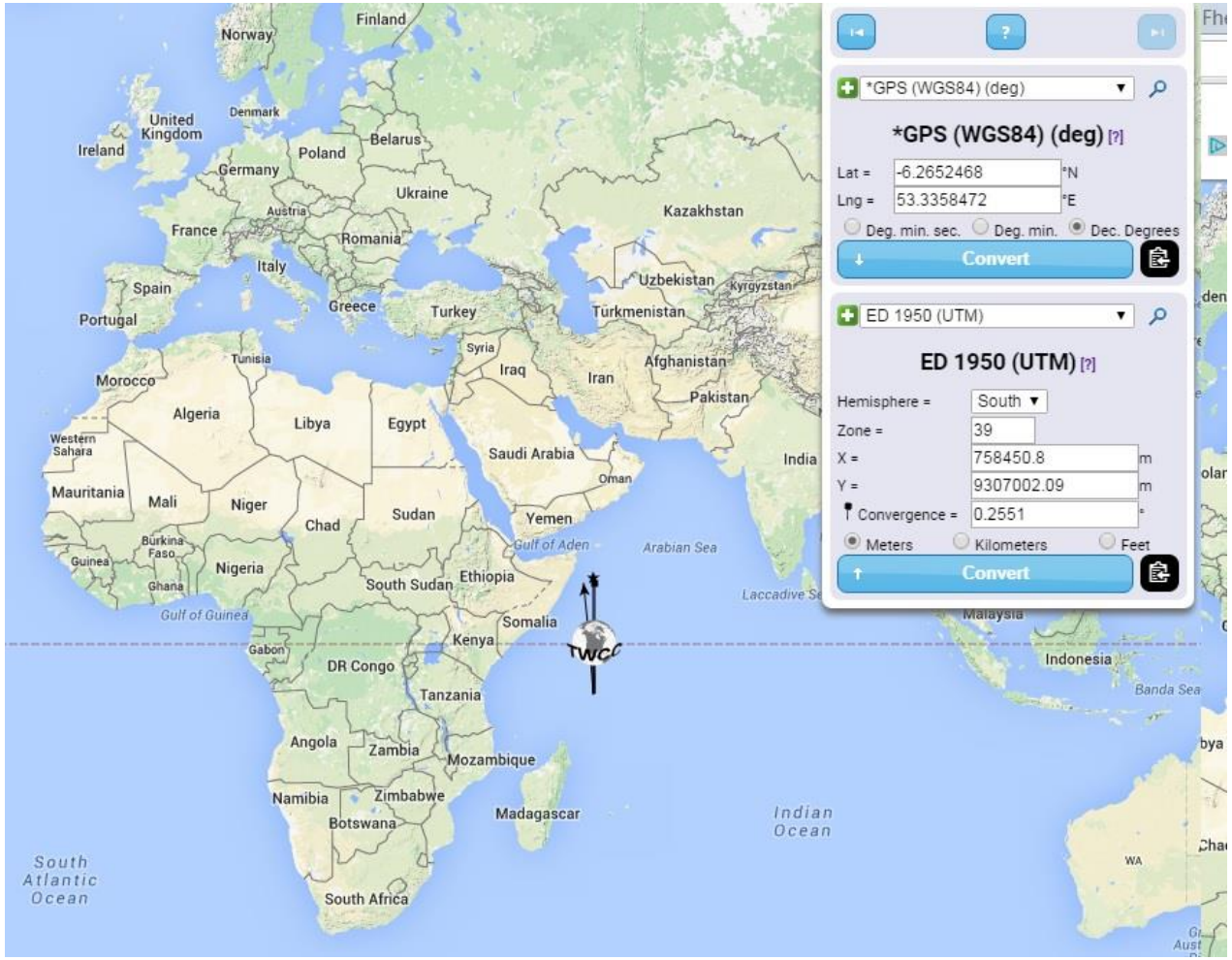


WGS84



HPCC SYSTEMS®

Number one bug in Geospatial



HPCC SYSTEMS®

*http://twcc.fr

Number one bug in Geospatial

LatY LonX

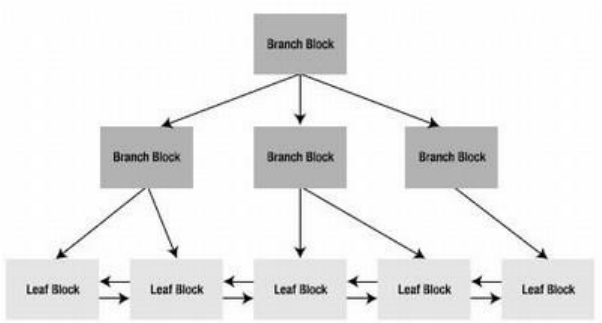
HPCC SYSTEMS®

Now I understand my data, what's next?

Data

Ingest

Index



Query

HPCC SYSTEMS®

Bringing Geospatial into HPCC

GOAL

Bring our geospatial processes into the realm of Big Data

HPCC SYSTEMS®

Big Data

STEPS

Extend HPCC and ECL to support the following main capabilities :

- Spatial filtering of vector geometries
- Spatial operations using vector geometries
- Spatial reference projection and transformation
- Reading of compressed geo-raster files

HPCC SYSTEMS®

Big Data

STEPS

Integration of open source libraries

GDAL

GEOS Geometry Engine Open Source

PROJ.4

HPCC SYSTEMS®

Ingesting Vector Data

It's a CSV file.

Id	Name	Geometry	Projection	Value
1	Alice's place	POINT (53.78925462 -6.08354321)	4326*	€5,973,000
2	Bob's place	POINT (-34.78925462 7.08354321)	4326	€872,000
3	Celine's place	POINT (102.78925462 -6.08354321)	4326	€9,324,000

* WGS84 (Lat/Lon)

HPCC SYSTEMS®

Ingesting Vector Data

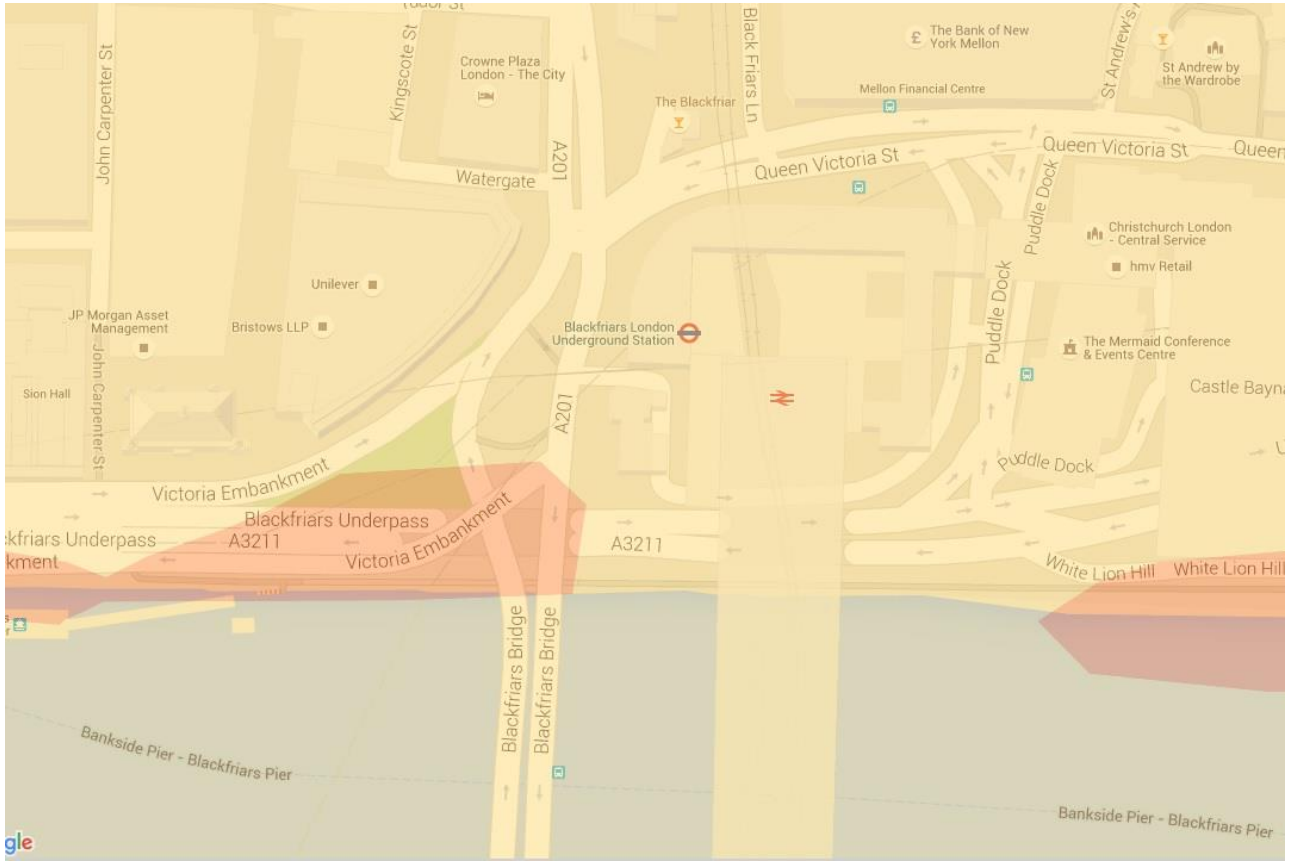
1. Shape data

2. Parse XPATH

3. Process and index

Data ready to query

It's a GML / XML file.



HPCC SYSTEMS®

Ingesting Vector Data

1. Shape data → **2.** Parse XPATH → **3.** Process and index → **Data ready to query**

It's a GML / XML file.

```
<?xml version="1.0" encoding="utf-8" ?>
<ogr:FeatureCollection
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="http://ogr.maptools.org/SolubleRocks_v6_SP.xsd"
     xmlns:ogr="http://ogr.maptools.org/"
     xmlns:gml="http://www.opengis.net/gml">
  <gml:boundedBy>
    <gml:Box>
      <gml:coord><gml:X>126787.0000000005</gml:X><gml:Y>50221.99999999939</gml:Y></gml:coord>
      <gml:coord><gml:X>640306.0000000012</gml:X><gml:Y>973634.0000000012</gml:Y></gml:coord>
    </gml:Box>
  </gml:boundedBy>
  <gml:featureMember>
    <ogr:SolubleRocks_v6_SP fid="F0">
      <ogr:geometryProperty><gml:Polygon><gml:outerBoundaryIs><gml:LinearRing><gml:coordinates>245859.99999999942,52505.999999999272 245851.999
      <ogr:CLASS>A</ogr:CLASS>
      <ogr:LEGEND>Soluble rocks are present within the ground. Very few dissolution features are likely to be present. Potential for difficult
      <ogr:VERSION>SolubleRocks_v6</ogr:VERSION>
      <ogr:ORIG_FID>0</ogr:ORIG_FID>
    </ogr:SolubleRocks_v6_SP>
  </gml:featureMember>
<ogr:FeatureCollection>
```

HPCC SYSTEMS®

Ingesting Vector Data

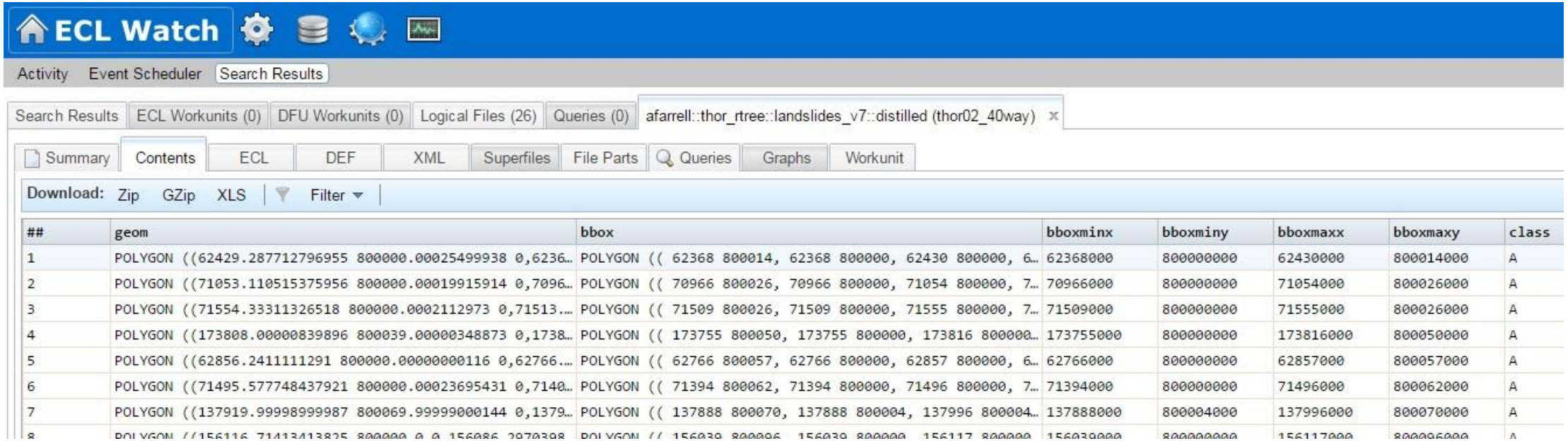
1. Shape data

2. Parse XPATH

3. Process and index

Data ready to query

It's a GML / XML file.

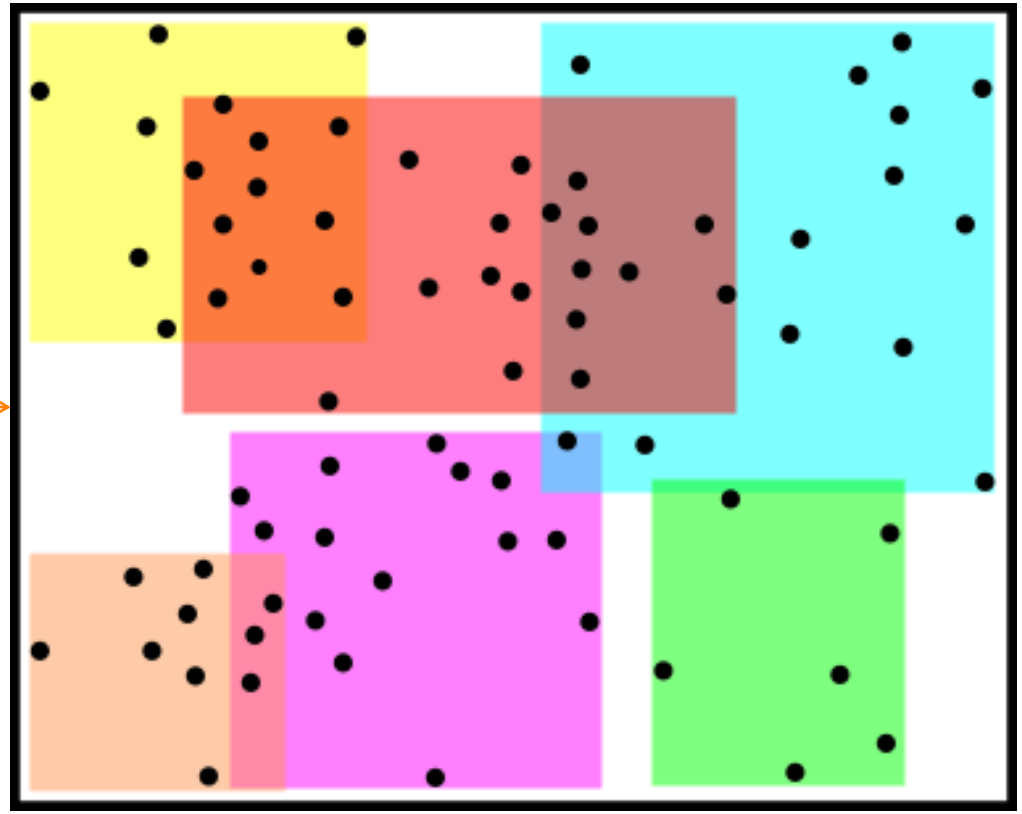


##	geom	bbox	bboxminx	bboxminy	bboxmaxx	bboxmaxy	class
1	POLYGON ((62429.287712796955 800000.00025499938 0,6236…	POLYGON ((62368 800014, 62368 800000, 62430 800000, 6…	62368000	800000000	62430000	800014000	A
2	POLYGON ((71053.110515375956 800000.00019915914 0,7096…	POLYGON ((70966 800026, 70966 800000, 71054 800000, 7…	70966000	800000000	71054000	800026000	A
3	POLYGON ((71554.33311326518 800000.0002112973 0,71513.…	POLYGON ((71509 800026, 71509 800000, 71555 800000, 7…	71509000	800000000	71555000	800026000	A
4	POLYGON ((173808.00000839896 800039.00000348873 0,1738…	POLYGON ((173755 800050, 173755 800000, 173816 800000…	173755000	800000000	173816000	800050000	A
5	POLYGON ((62856.2411111291 800000.00000000116 0,62766.…	POLYGON ((62766 800057, 62766 800000, 62857 800000, 6…	62766000	800000000	62857000	800057000	A
6	POLYGON ((71495.577748437921 800000.00023695431 0,7140…	POLYGON ((71394 800062, 71394 800000, 71496 800000, 7…	71394000	800000000	71496000	800062000	A
7	POLYGON ((137919.99998999987 800069.99999000144 0,1379…	POLYGON ((137888 800070, 137888 800004, 137996 800004…	137888000	800004000	137996000	800070000	A
8	POLYGON ((156116.71413413825 800000.0 0.156086.2970398	POLYGON ((156039 800096, 156039 800000, 156117 800000	156039000	800000000	156117000	800096000	A

HPCC SYSTEMS®

Indexing vector data

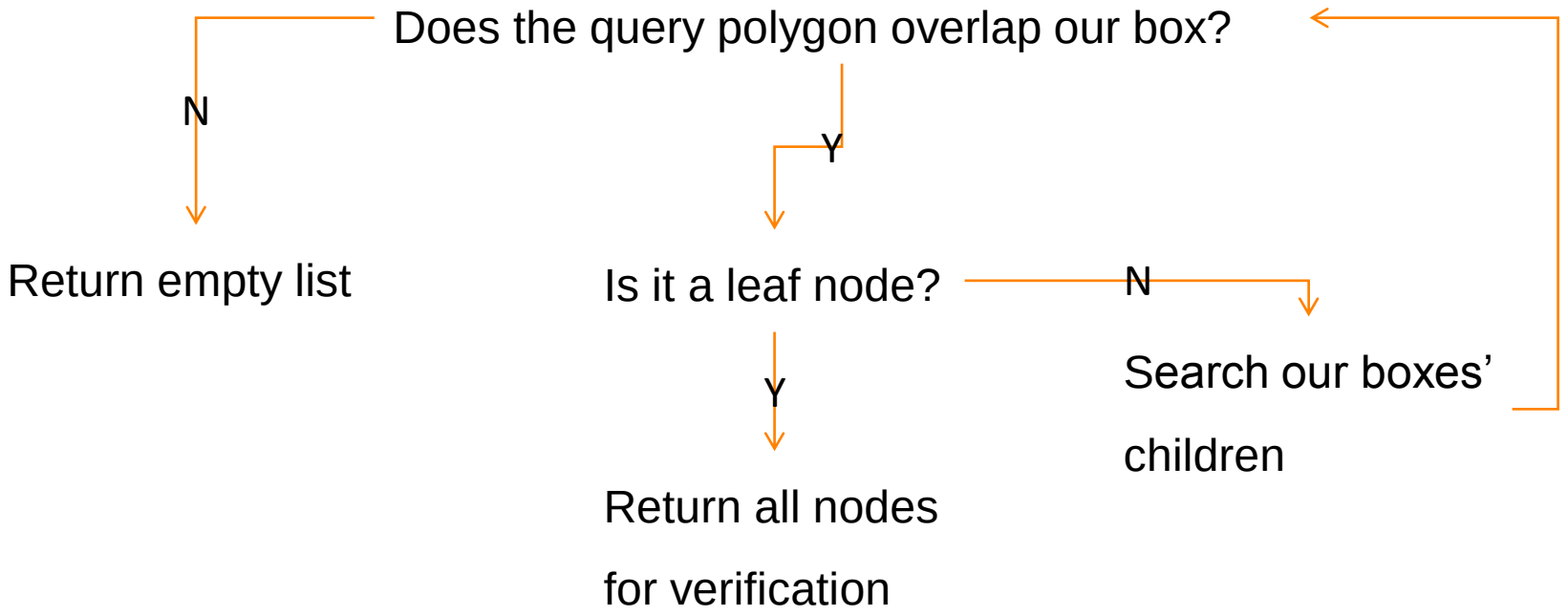
- Outline Box: Biggest rectangle
- Boxes contain boxes
- Bottom box in the tree contains actual geometries
- Here, 3 levels pictured
- Boxes can overlap (entries are only in one)



HPCC SYSTEMS®

Querying vector data

Searching an R-Tree: e.g. Finding all buildings (points) inside a flood zone (polygon)



HPCC SYSTEMS®

Ingesting Raster Data

1. Raster data → **2.** Tile and spray → **3.** Process and index → **Data ready to query**

It's a raster / TIFF file. Bitmap image

HPCC SYSTEMS®

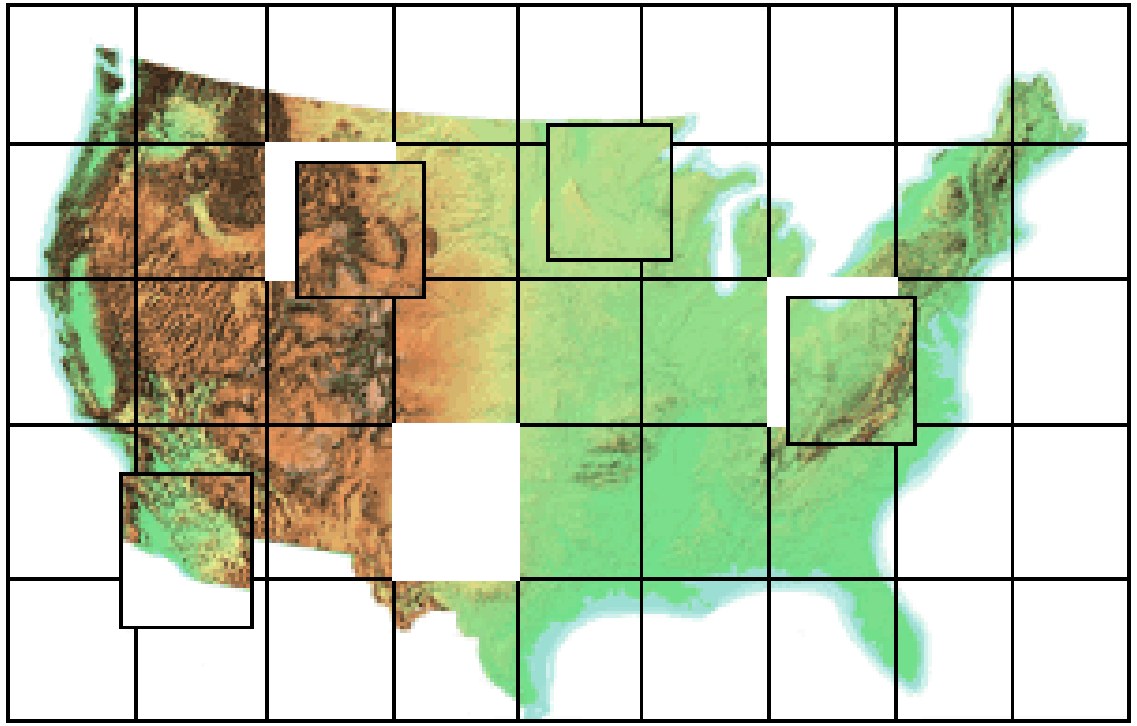
Ingesting Raster Data

1. Raster data → **2.** Tile and spray → **3.** Process and index → **Data ready to query**

Tiling divides raster images into small manageable areas of known dimensions.

These tiles have their own metadata:

- Bounding box
- Grid position



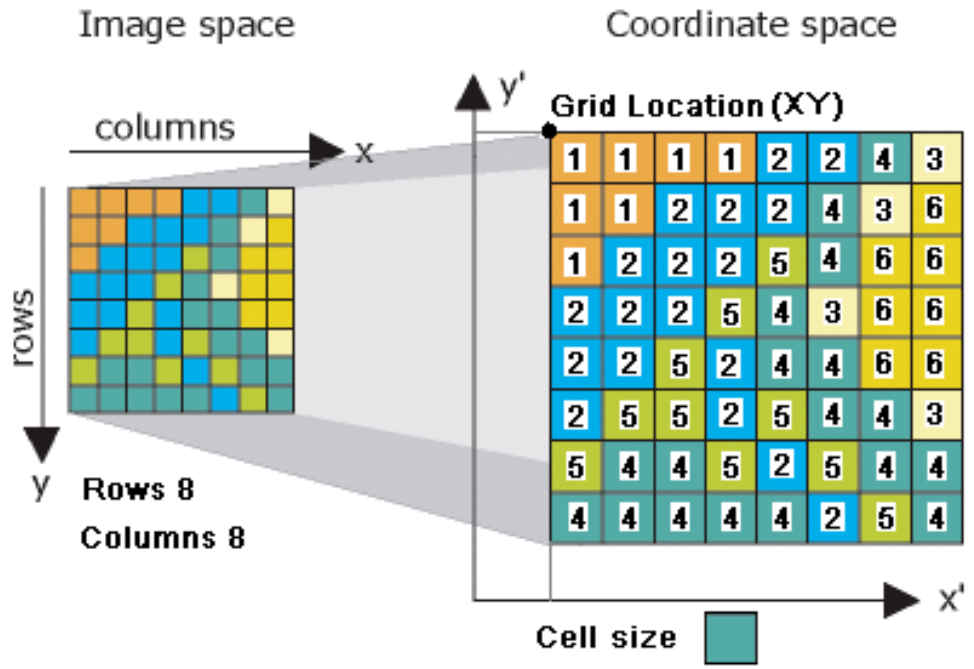
Tiles in a Raster

HPCC SYSTEMS®

Ingesting Raster Data

1. Raster data → **2.** Tile and spray → **3.** Process and index → **Data ready to query**

1. Figure out which grid position the geometry needs
2. Extract the required pixel
3. Interrogate the pixel for its value
4. Interpret its value
5. Return to user

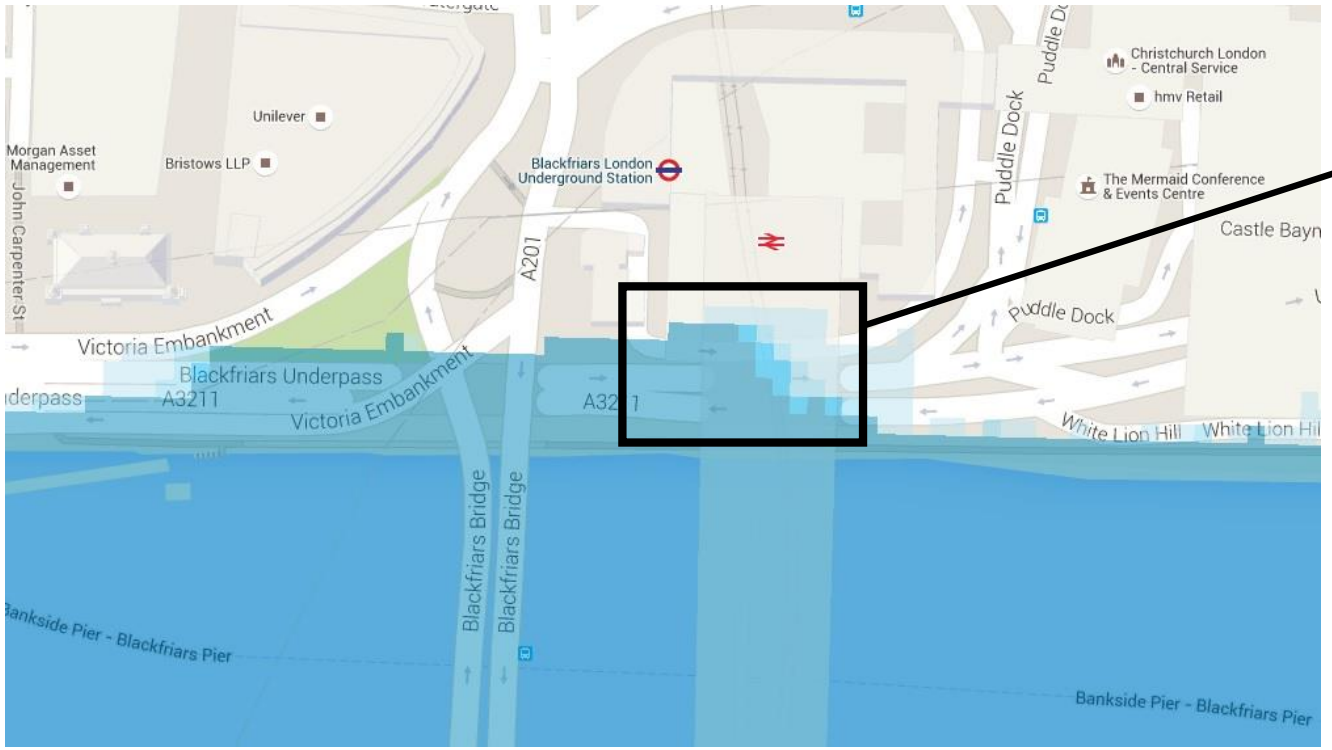


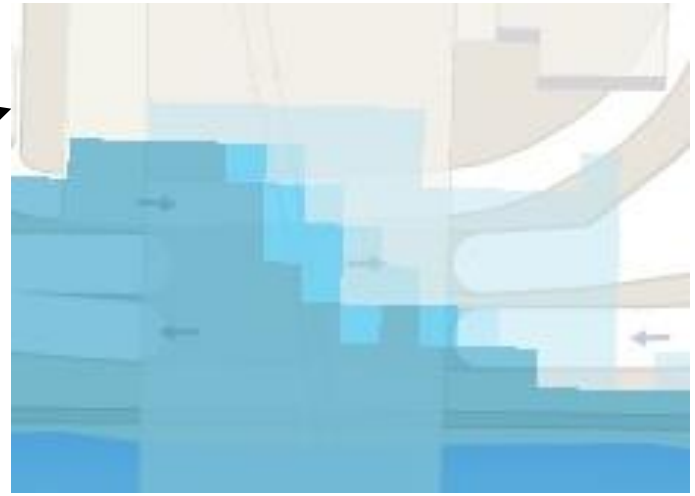
HPCC SYSTEMS®

Ingesting Raster Data

1. Raster data → **2.** Tile and spray → **3.** Process and index → **Data ready to query**

It's a raster / TIFF file. Bitmap image





HPCC SYSTEMS®

Ingesting Raster Data

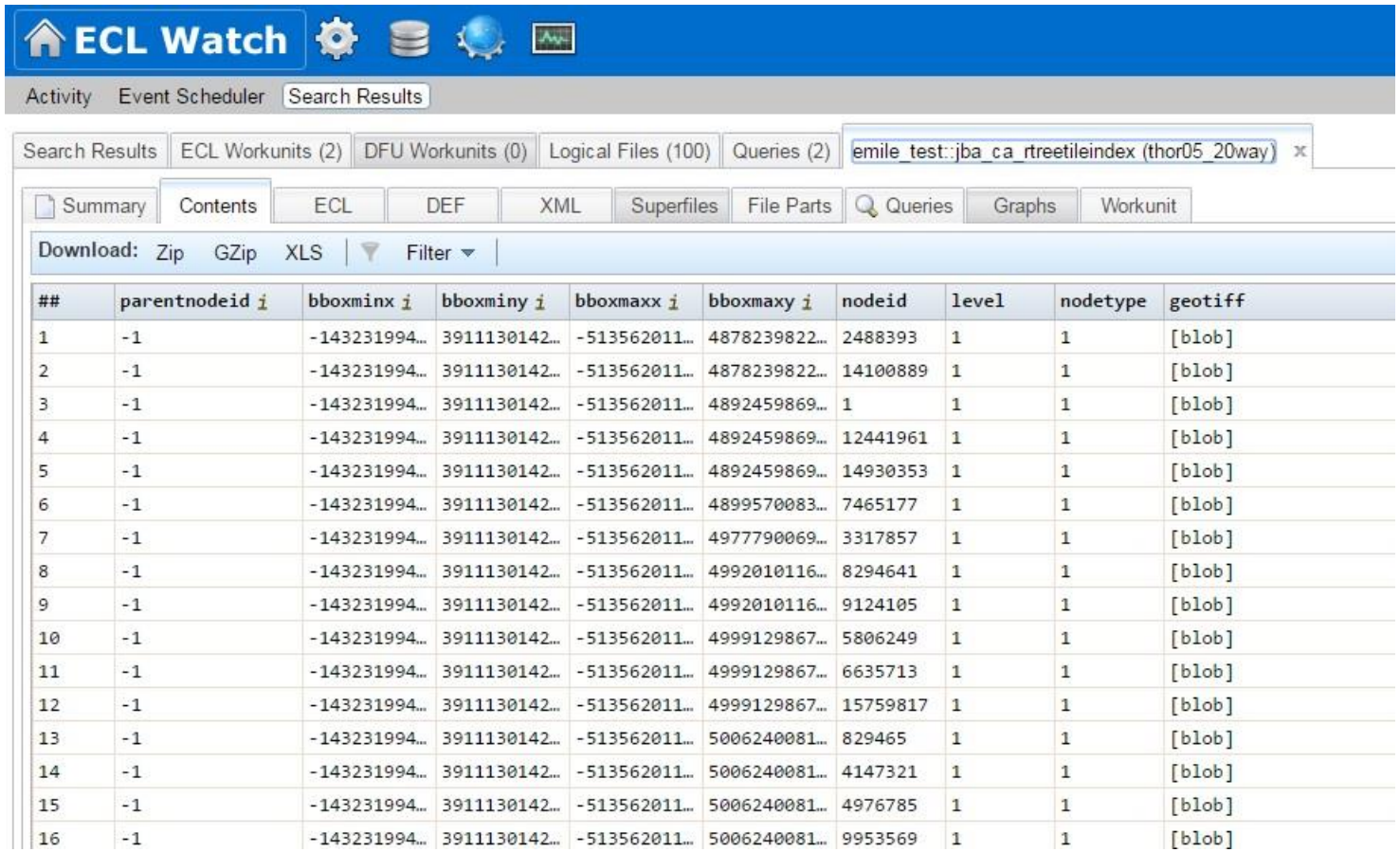
1. Raster data

2. Tile and spray

3. Process and index

Data ready to query

It's a raster / TIFF file.



##	parentnodeid	bboxminx	bboxminy	bboxmaxx	bboxmaxy	nodeid	level	nodetype	geotiff
1	-1	-143231994…	3911130142…	-513562011…	4878239822…	2488393	1	1	[blob]
2	-1	-143231994…	3911130142…	-513562011…	4878239822…	14100889	1	1	[blob]
3	-1	-143231994…	3911130142…	-513562011…	4892459869…	1	1	1	[blob]
4	-1	-143231994…	3911130142…	-513562011…	4892459869…	12441961	1	1	[blob]
5	-1	-143231994…	3911130142…	-513562011…	4892459869…	14930353	1	1	[blob]
6	-1	-143231994…	3911130142…	-513562011…	4899570083…	7465177	1	1	[blob]
7	-1	-143231994…	3911130142…	-513562011…	4977790069…	3317857	1	1	[blob]
8	-1	-143231994…	3911130142…	-513562011…	4992010116…	8294641	1	1	[blob]
9	-1	-143231994…	3911130142…	-513562011…	4992010116…	9124105	1	1	[blob]
10	-1	-143231994…	3911130142…	-513562011…	4999129867…	5806249	1	1	[blob]
11	-1	-143231994…	3911130142…	-513562011…	4999129867…	6635713	1	1	[blob]
12	-1	-143231994…	3911130142…	-513562011…	4999129867…	15759817	1	1	[blob]
13	-1	-143231994…	3911130142…	-513562011…	5006240081…	829465	1	1	[blob]
14	-1	-143231994…	3911130142…	-513562011…	5006240081…	4147321	1	1	[blob]
15	-1	-143231994…	3911130142…	-513562011…	5006240081…	4976785	1	1	[blob]
16	-1	-143231994…	3911130142…	-513562011…	5006240081…	9953569	1	1	[blob]

HPCC SYSTEMS®

Bringing it all together

##	geom	id
7	POINT (-84.25332677705793571 34.07173235399688593)	1
6	POINT (-84.2606472418293464 34.06904953470633046)	2
5	POINT (-84.26401909226950693 34.07132809899221115)	3
4	POINT (-84.27511070555951278 34.07632602535307598)	4
3	POINT (-84.28034594703238724 34.06441871524530995)	5
2	POINT (-84.28096707737662996 34.07386384848369687)	6
1	POINT (-84.26180076961152565 34.07911885643056848)	7

Execute ECL

PROJECT : For Each Building

TRANSFORM :

- <<Calculate distance to Fire station & assign risk category
- <<Calculate distance to Police station & assign risk category
- << Determine if building is in flood zone

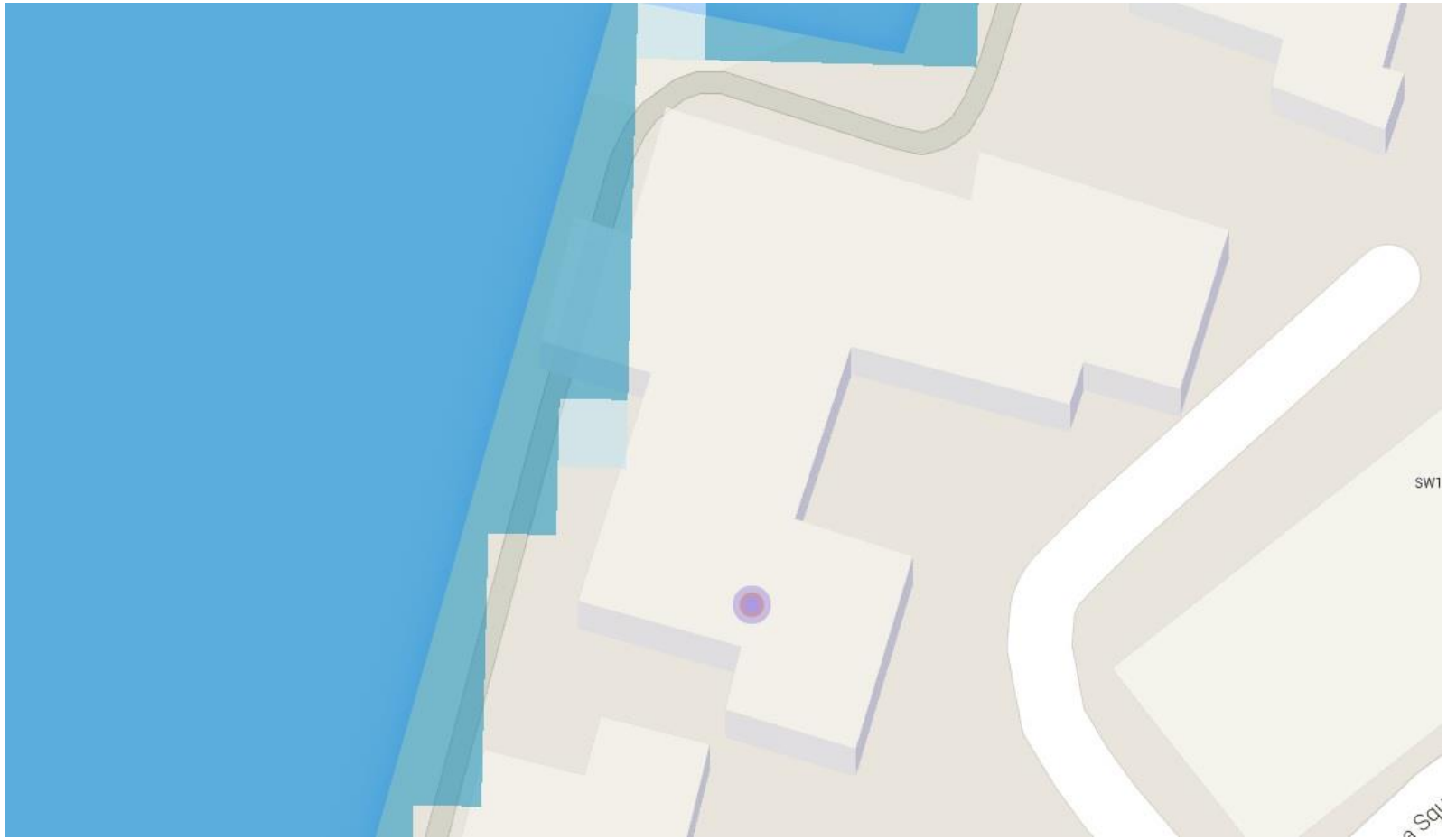
OUTPUT : Each building augmented with peril data:

Display Output

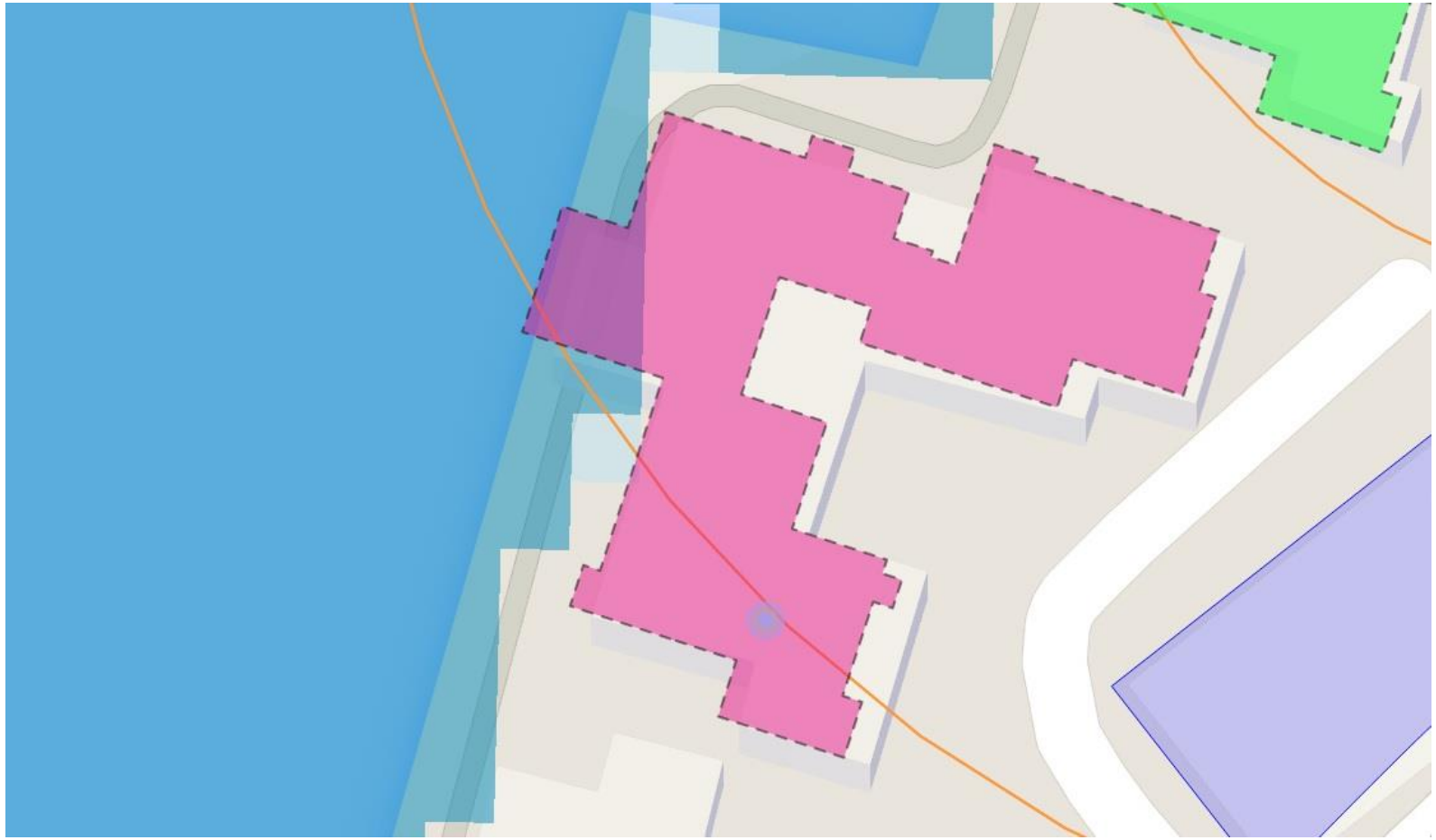
##	geom	id	distancetofirestation	fireriskcategory	distancetopolicestation	crimeriskcategory	isatriskofflood
7	POINT (-84.253326...	1	1930	A	2846	A	true
6	POINT (-84.260647...	2	1420	B	2147	A	true
5	POINT (-84.264019...	3	1022	B	1863	B	false
4	POINT (-84.275110...	4	144	D	1194	C	false
3	POINT (-84.280345...	5	1420	B	551	D	false
2	POINT (-84.280967...	6	716	C	662	C	false
1	POINT (-84.261800...	7	1147	B	2360	A	false

*Andrew Farrell
In pursuit of perils : Geo-spatial risk analysis through HPCC Systems
https://hpccsystems.com/resources/blog/afarrell/pursuit-perils-geo-spatial-risk-analysis-through-hpcc-systems

Add even more value



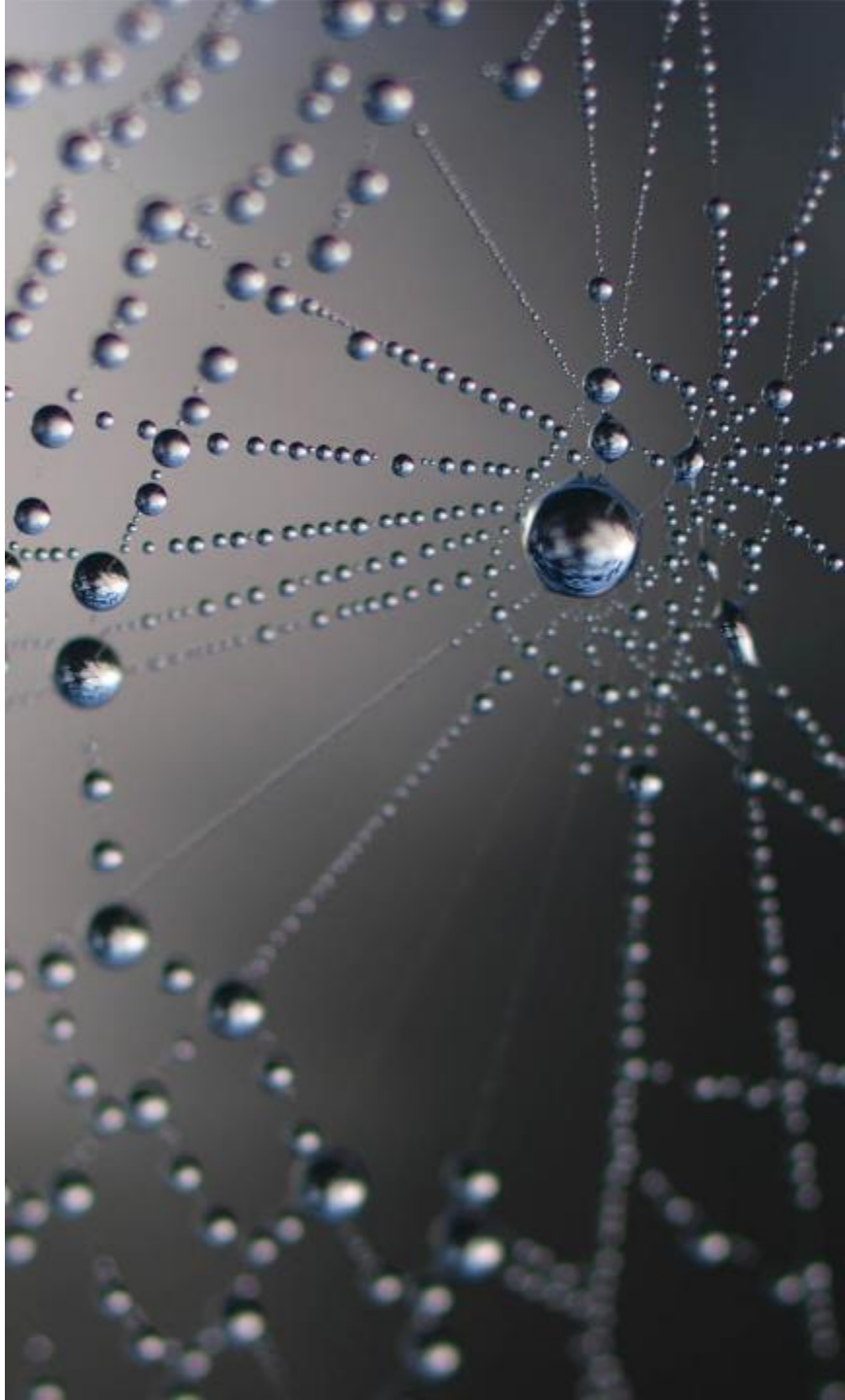
Add even more value



Why Geospatial with HPCC?

- Efficient parallel processing
- Ability to import libraries from different languages
- Good coverage of functions and spatial predicates
- Fast ingestion
- Support for different formats
- Sub-second queries

HPCC SYSTEMS®



Questions?

HPCC SYSTEMS®

hpccsystems.com