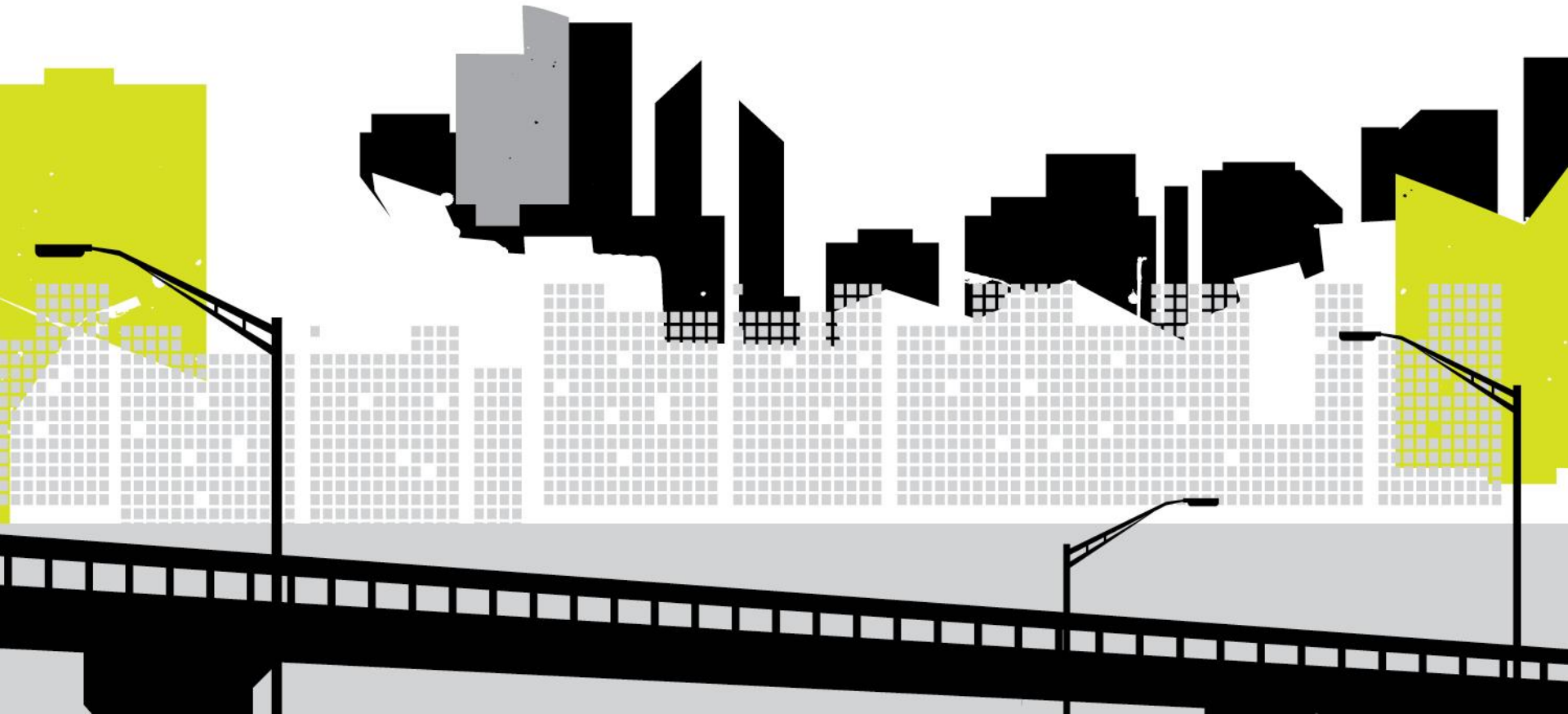**1spatial**
YOUR WORLD SMARTER

Data Governance
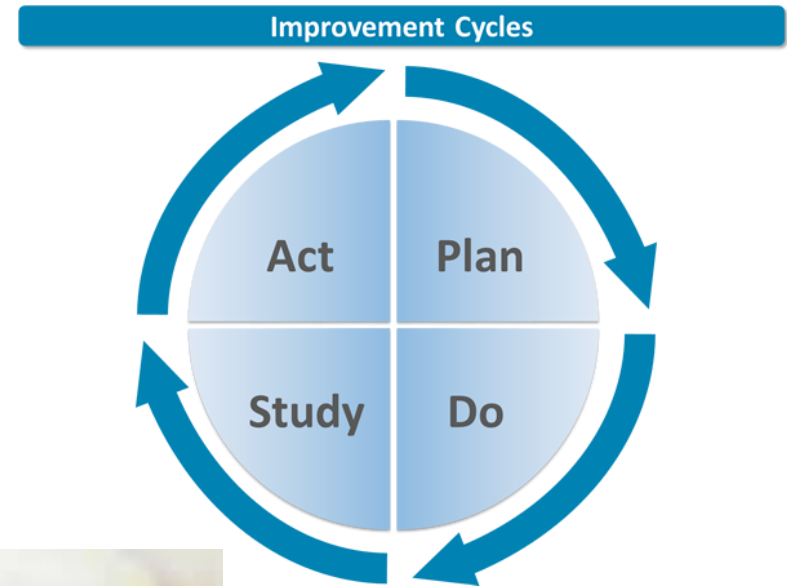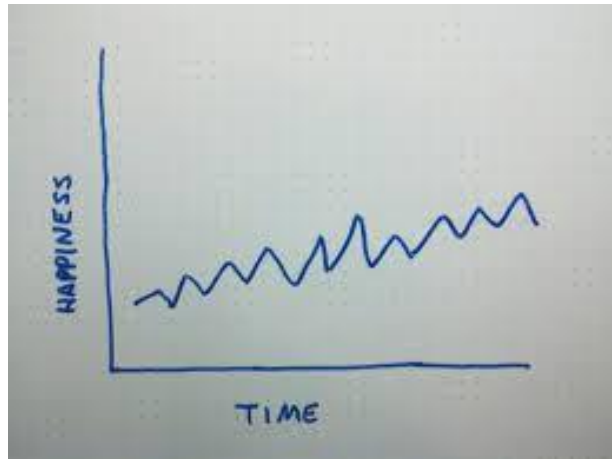# Tracking Data Quality Over Time

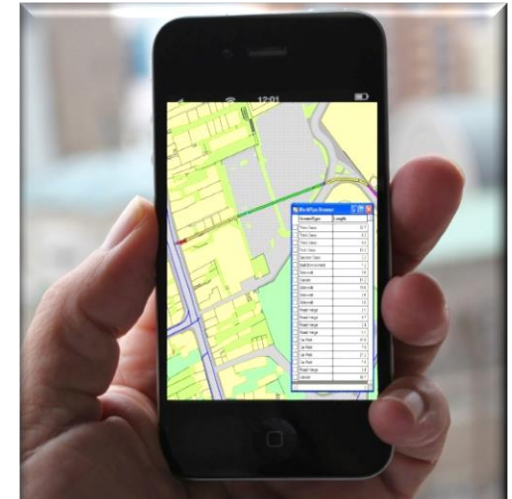# Topics for Discussion

- What is data governance?
- Who does it?
- How is the data governed?
- What tools are available?
  - FME
  - 1Validate
  - The differences
- National Validation
  - What we've done
  - Why we've done it
  - The results
  - A short Video

Data Governance

Data Governance - Why do it?

Data Governance - Why do it?

Data Governance - Why do it?

Data Governance – How?
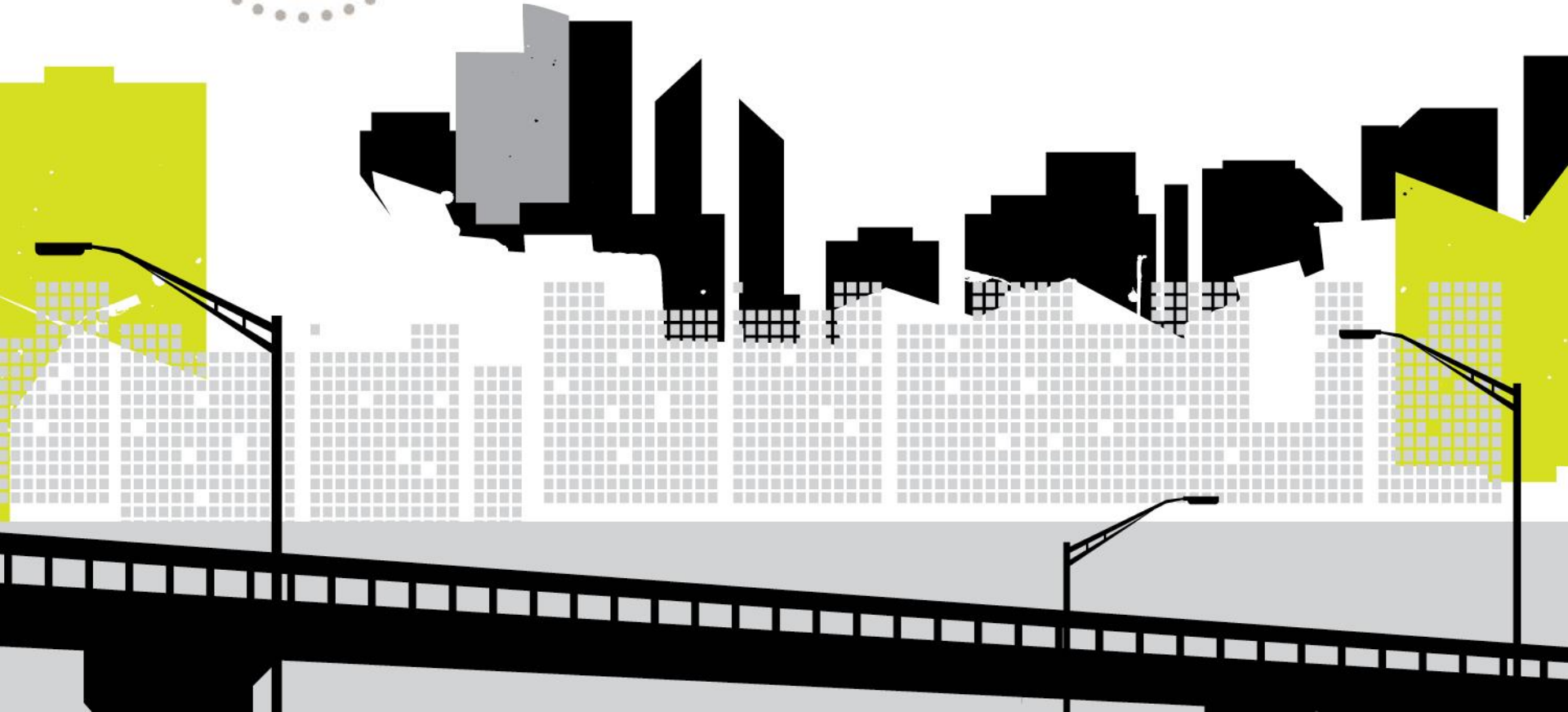
Core Competencies

Convert data formats

E.g. from ESRI Shapefile to MapInfo tab file

Convert Spatial Reference Systems

E.g. from WGS84 to British National Grid

Transform schema

E.g. Change the names of feature types and attributes to a different schema

Report broken geometry or attributes

E.g. Find duplicate vertices or unclosed polygons

E.g. Material must be one of 'Iron, PVC'

E.g. Postal code must be in correct format

Core Competencies

Simple Transformation of attributes or geometries
E.g. Convert polygon to a centroid, buffer points
E.g. Convert 'FC_FOREST_DECIDUOUS' to 'DECIDUOUS'

Basic cleaning
E.g. Remove spikes, kickbacks, duplicate points in geometries
E.g. correct the orientation of rings, break multi-part geometries into single parts
E.g. Correct case, whitespace, empty values in attributes
E.g. Delete lines that are too short, delete polygons that are too small
E.g. Remove small holes from within polygons

Call via REST web service
Invoke processing via REST web service without writing bespoke wrappers

Core Competencies

Advanced Validation

E.g. IGN Belgium: This building is valid if it does not intersect a road, unless there is a bridge or tunnel between us and the road

E.g. UU: This valve is only valid if it connects to two pipes that have different pressures to each other

Building a large rules catalogue

E.g. OSGB, OSi, United Utilities: Run tens or hundreds of contextual spatial rules on tens-of-thousands to millions of features

Run rules and use stats to create drilldown reports or make automatic decisions

E.g. US Census, OSGB, OSi, UU: Run rules and analyse returned XML report, continue if 95% of features pass

E.g. show graphs and stats of rule conformance percentages, listing rules failed for each feature

Complex Scenarios

Basic simplification

E.g. Simplify boundaries of a small or medium sized coverage of polygons by building topology and simplifying edges to maintain connectivity

E.g. Merge small polygons into neighbours of the same type

E.g. delete short dead-end lines in a network

Advanced Geometric Cleaning

E.g. Snapping polygons and lines together to remove gaps, overlaps, undershoots overshoots and misalignments

E.g. Splitting polygons up where they have large overlaps with surrounding polygons, then merging each overlapping part into the best geometrically matching neighbour (the one that creates the 'roundest' result)

Complex Scenarios

Process geometries in relation to other geometries

E.g. Split polygon features into multiple smaller polygon features using the line features in another class

E.g. Split network lines at T-junctions, merge network lines at non-junctions

E.g. Move points away from other points to make them more visible

E.g. Set an attribute on a polygon feature to be the count of the number of a certain type of point within that polygon

Basic inter-feature validation

E.g. Find line or polygon overlaps and output the overlap as a reporting feature

E.g. Find junctions that are not connected to at least two lines

E.g. Check that parcel polygons are completely surrounded by parcel lines

E.g. Check that road lines are connected to at least one other road line at their ends

Complex Scenarios

Simple Conflation

E.g. Conflate two road datasets by using the geometry from one dataset and finding the equivalent feature in the other dataset from which to copy the attributes. Assumes there is a fuzzy but 1-1 match between feature geometries
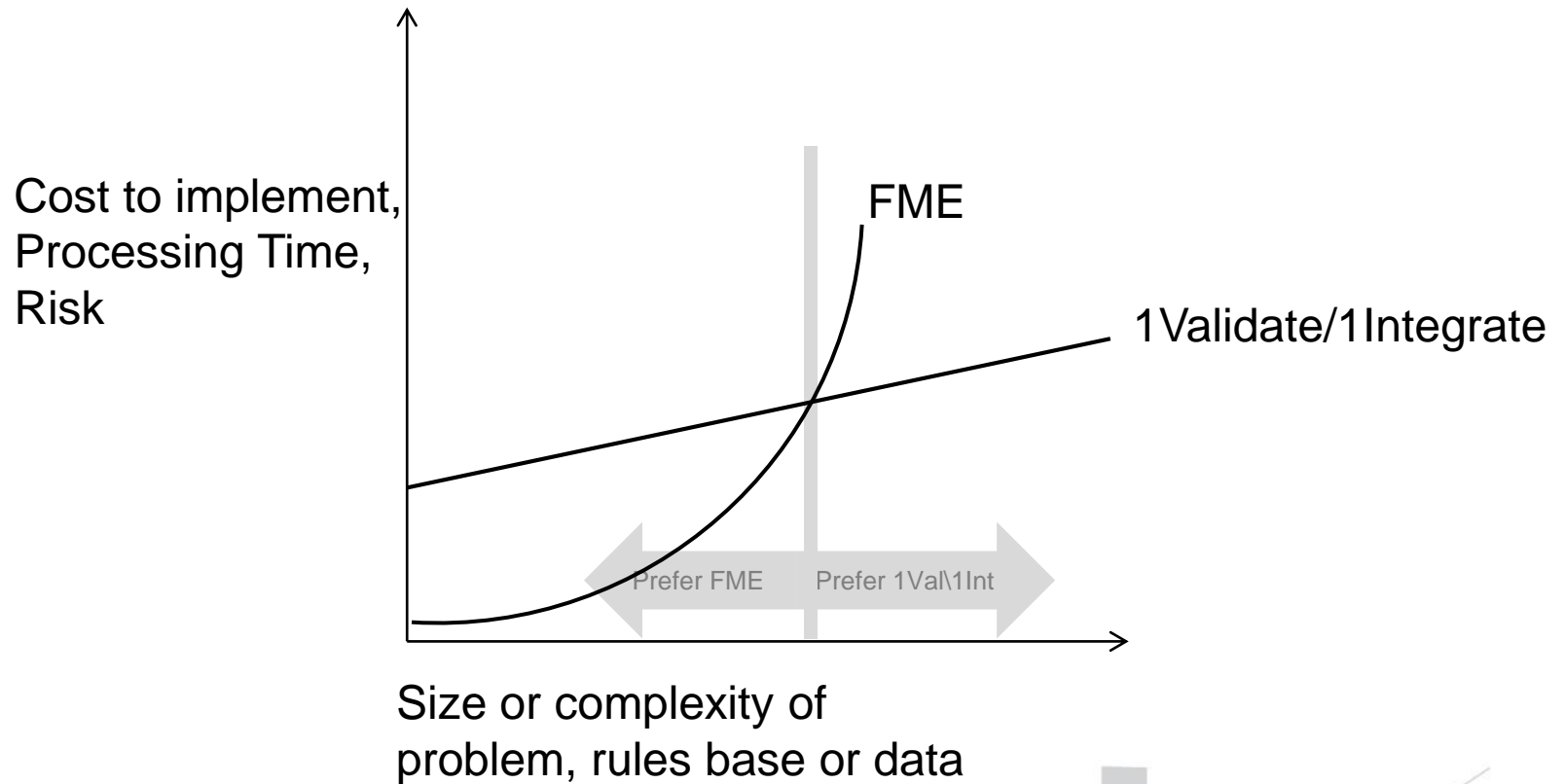
Classification

E.g. some processes for PRIME2 for OSi: Create polygons from lines and classify based on the percentage of boundary shared with various neighbours: e.g. if it is long and thin and has a certain length centreline through it then it's a road surface – unless it is also completely surrounded by a road surface in which case it's a traffic island

Basic Aggregation

Find nearby groups of certain types of features and aggregate their geometries by union convex hull, shrink wrap or other aggregation techniques.

At Scale

Cost to implement,
Processing Time,
Risk

FME

1Validate/1Integrate

Prefer FME   Prefer 1Val\1Int

Size or complexity of
problem, rules base or data

At Scale

Memory

At Scale

Time
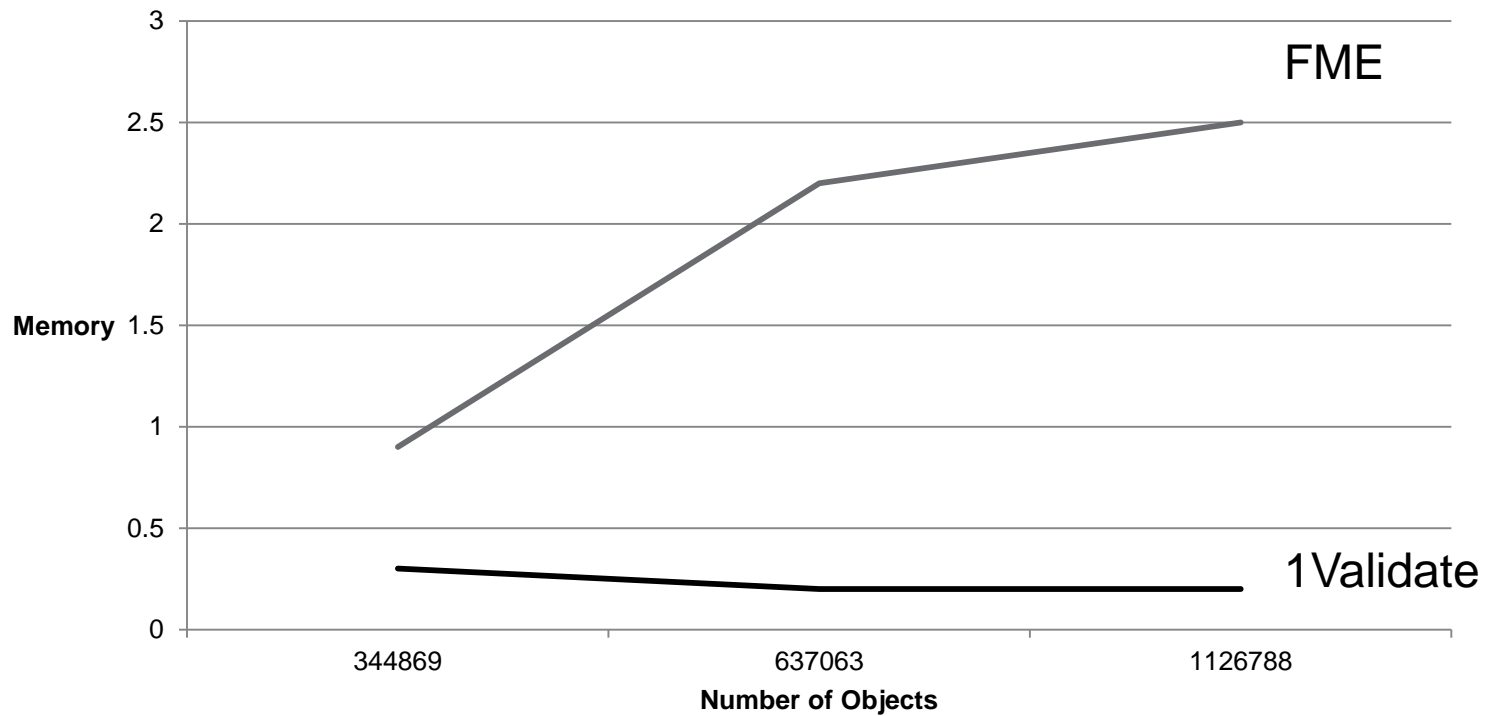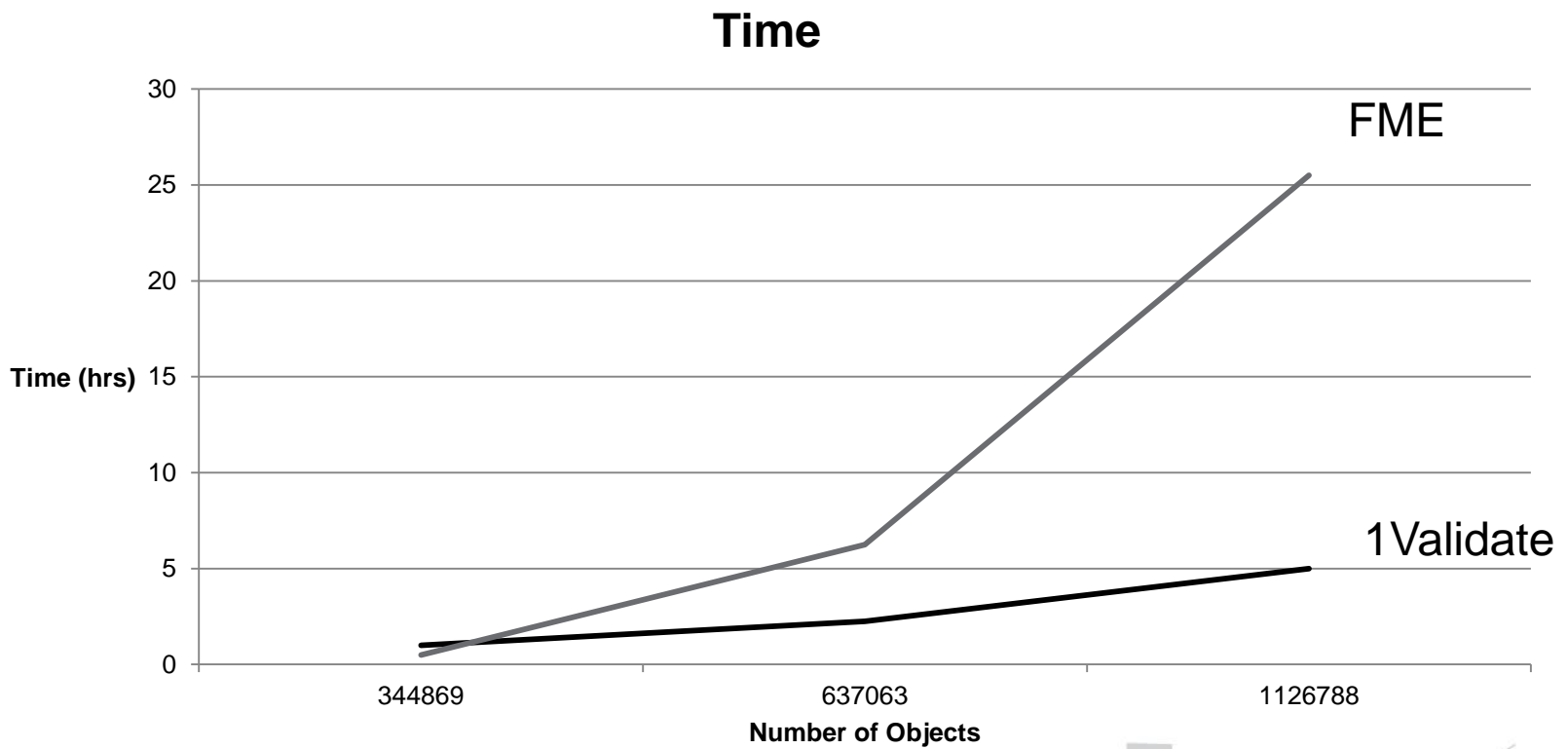
Chart axes: Time (hrs) vertical axis (0 to 30), Number of Objects horizontal axis (344869, 637063, 1126788). Two lines labeled FME and 1Validate.
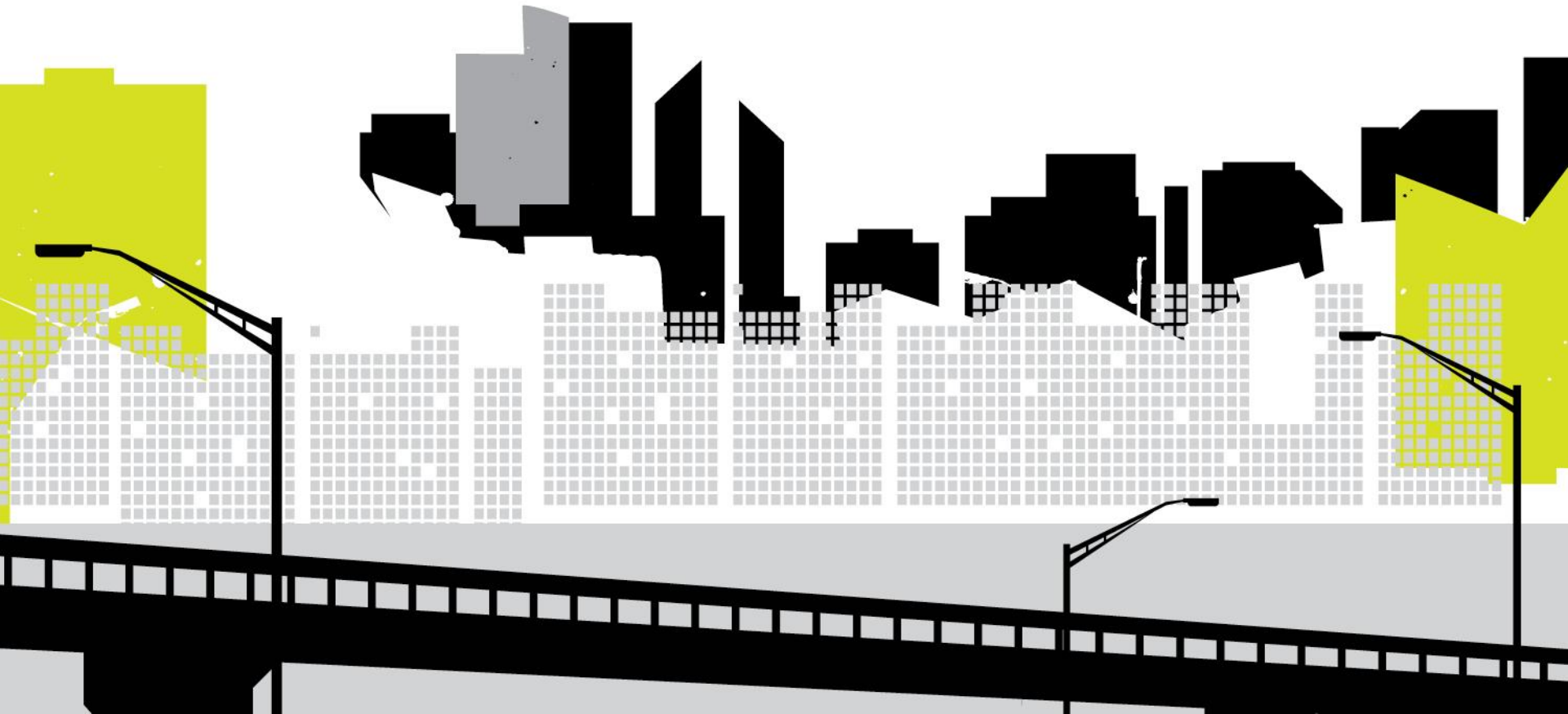
Report broken geometry or attributes **E.g. Find duplicate vertices or unclosed polygons** **E.g. Material must be one of 'Iron, PVC'** **E.g. Postal code must be in correct format**
Advanced simplification **E.g. Simplify boundaries of a possibly large sized coverage of polygons by removing shared vertices and maintaining connectivity. Prevent simplifications that cause other features to intersect with or jump over the boundaries.**
Simple Transformation of attributes or geometries **E.g. Convert polygon to a centroid, buffer points** **E.g. Convert 'FC_FOREST_DECIDUOUS' to 'DECIDUOUS'**

Basic cleaning
E.g. Remove spikes, kickbacks, duplicate points in geometries
E.g. correct the orientation of rings, break multi-part geometries into single parts
E.g. Correct case, whitespace, empty values in attributes
E.g. Delete lines that are too short, delete polygons that are too small
E.g. Remove small holes from within polygons

Advanced Conflation
E.g. US Census: Receive new dataset from submitter, compare with existing data to detect changes and create a record of updates needed to stitch information into existing features. Requires fuzzy geometric matching to determine if feature is already in the data or not, even if the geometry has been changed. This is not simply comparing attributes and checking that the geometry is equivalent: There is not necessarily a 1-1 match between features e.g. one road can be replaced by three, which together have a similar but different geometry. Need to work whether a spatially nearby feature is similar enough geometrically that it's an update to the original or if it's a new feature.

Classification

E.g. some processes for PRIME2 for OSi: Create polygons from lines and classify based on the percentage of boundary shared with various neighbours: e.g. if it is long and thin and has a certain length centreline through it then it's a road surface – unless it is also completely surrounded by a road surface in which case it's a traffic island

Advanced Aggregation

E.g. PRIME2 for OSi: Cluster nearby slope lines together to form slope polygons, only joining correct parallel lines together and classify based on nearby text, e.g. 'Ringfort'

Advanced Inference

E.g. PRIME2, United Utilities inferring missing sewer data: Take existing sewer main network and OS MasterMap topographic polygons and census data on occupants and building type. Apply rules of thumb from engineers to estimate where sewers would have been laid: Calculate whether type of building would have sewers front and back or just at front, calculate path of local sewers along row of houses, connect to each building, estimate location of manhole covers, connect to mains

1validate Core Competencies 1spatial

YOUR WORLD SMARTER

Advanced Validation
E.g. IGN Belgium: This building is valid if it does not intersect a road, unless there is a bridge or tunnel between us and the road
E.g. UU: This valve is only valid if it connects to two pipes that have different pressures to each other

Advanced Network Validation
E.g. TeleAtlas: Check that the whole network is fully accessible considering one way streets and turn restrictions

Building a large rules catalogue
E.g. OSGB, OSi, United Utilities: Run tens or hundreds of contextual spatial rules on tens-of-thousands to millions of features

Run rules and use stats to create drilldown reports or make automatic decisions
E.g. US Census, OSGB, OSi, UU: Run rules and analyse returned XML report, continue if 95% of features pass
E.g. show graphs and stats of rule conformance percentages, listing rules failed for each feature

Basic simplification

E.g. Simplify boundaries of a small or medium sized coverage of polygons by building topology and simplifying edges to maintain connectivity

E.g. Merge small polygons into neighbours of the same type

E.g. delete short dead-end lines in a network

Advanced Geometric Cleaning

E.g. Snapping polygons and lines together to remove gaps, overlaps, undershoots overshoots and misalignments

E.g. Splitting polygons up where they have large overlaps with surrounding polygons, then merging each overlapping part into the best geometrically matching neighbour (the one that creates the 'roundest' result)

Process geometries in relation to other geometries

E.g. Split polygon features into multiple smaller polygon features using the line features in another class

E.g. Split network lines at T-junctions, merge network lines at non-junctions

validate At Scale

Basic inter-feature validation

E.g. Find line or polygon overlaps and output the overlap as a reporting feature

E.g. Find junctions that are not connected to at least two lines

E.g. Check that parcel polygons are completely surrounded by parcel lines

E.g. Check that road lines are connected to at least one other road line at their ends

Simple Conflation

E.g. Conflate two road datasets by using the geometry from one dataset and finding the equivalent feature in the other dataset from which to copy the attributes. Assumes there is a fuzzy but 1-1 match between feature geometries
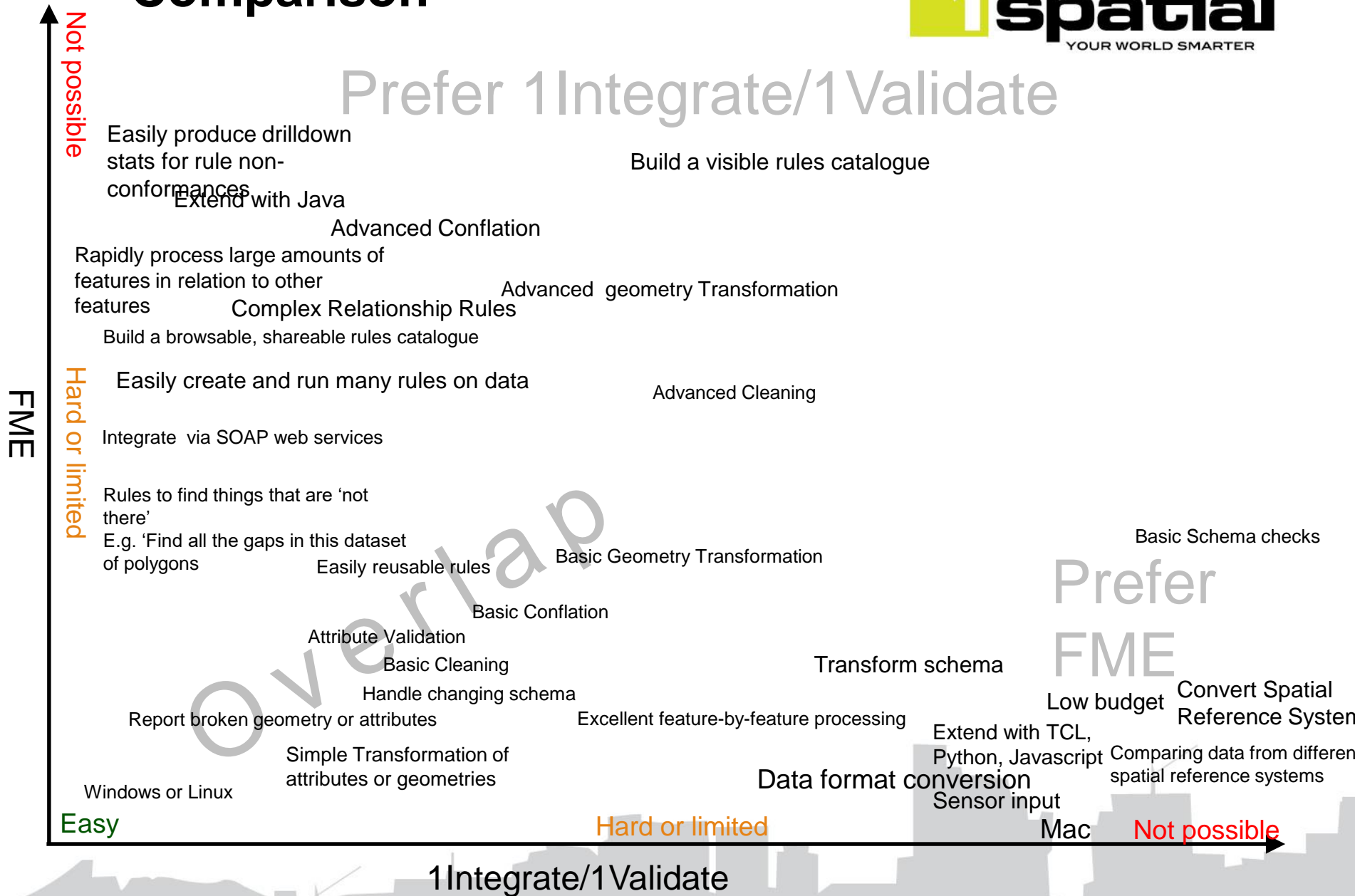
Basic Aggregation

Find nearby groups of certain types of features and aggregate their geometries by union convex hull, shrink wrap or other aggregation techniques.
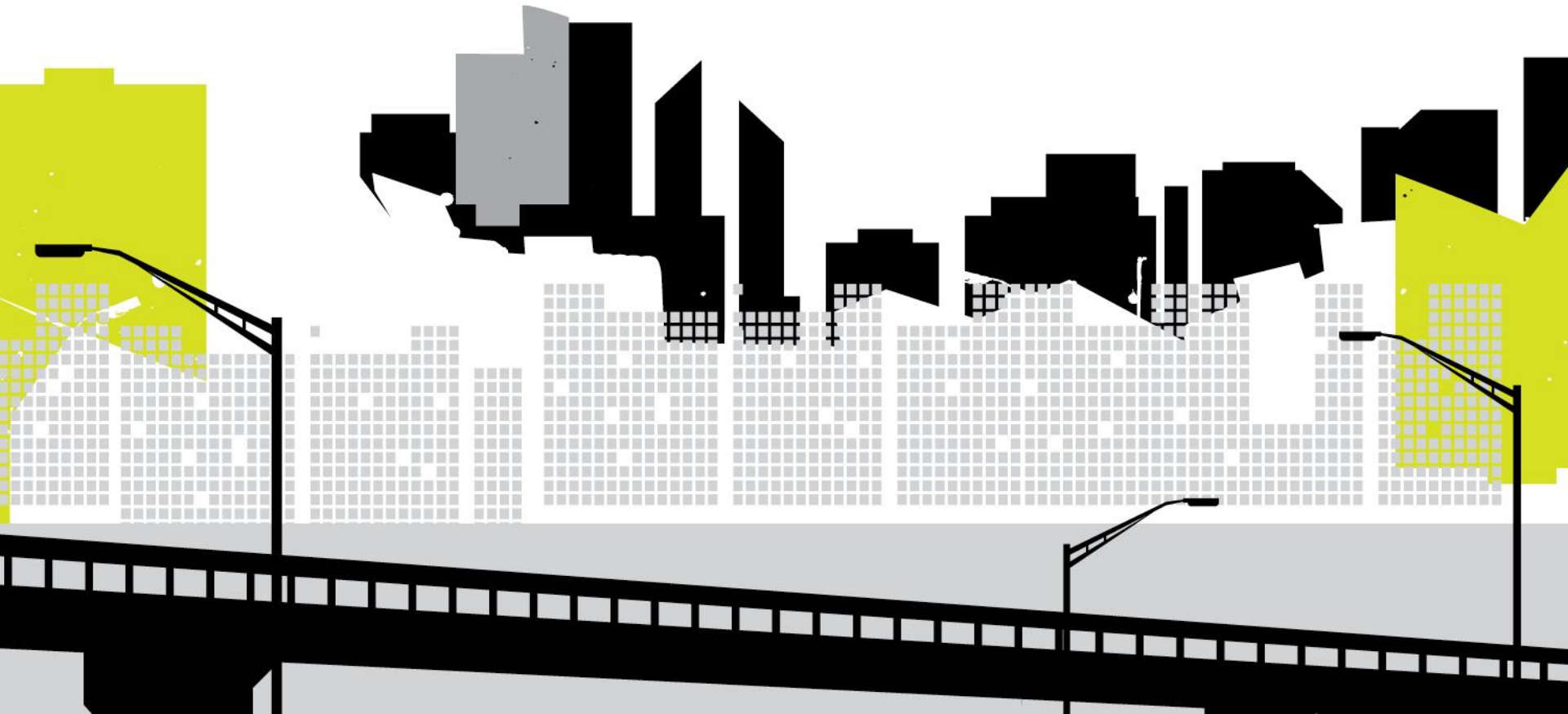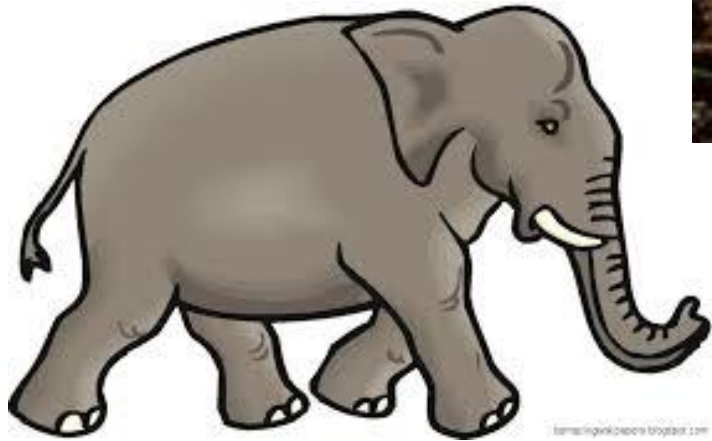
Comparison

1spatial
YOUR WORLD SMARTER

Prefer 1Integrate/1Validate

FME

Not possible

Hard or limited

Easily produce drilldown stats for rule non-conformances

Build a visible rules catalogue

Extend with Java

Advanced Conflation

Rapidly process large amounts of features in relation to other features

Advanced geometry Transformation

Complex Relationship Rules

Build a browsable, shareable rules catalogue

Easily create and run many rules on data

Advanced Cleaning

Integrate via SOAP web services

Rules to find things that are 'not there'
E.g. 'Find all the gaps in this dataset of polygons

Basic Schema checks

Easily reusable rules

Basic Geometry Transformation

Overlap

Prefer

FME

Basic Conflation

Attribute Validation

Basic Cleaning

Transform schema

Handle changing schema

Low budget

Convert Spatial Reference System

Report broken geometry or attributes

Excellent feature-by-feature processing

Extend with TCL, Python, Javascript

Comparing data from different spatial reference systems

Simple Transformation of attributes or geometries

Data format conversion

Windows or Linux

Sensor input

Easy

Hard or limited

Mac

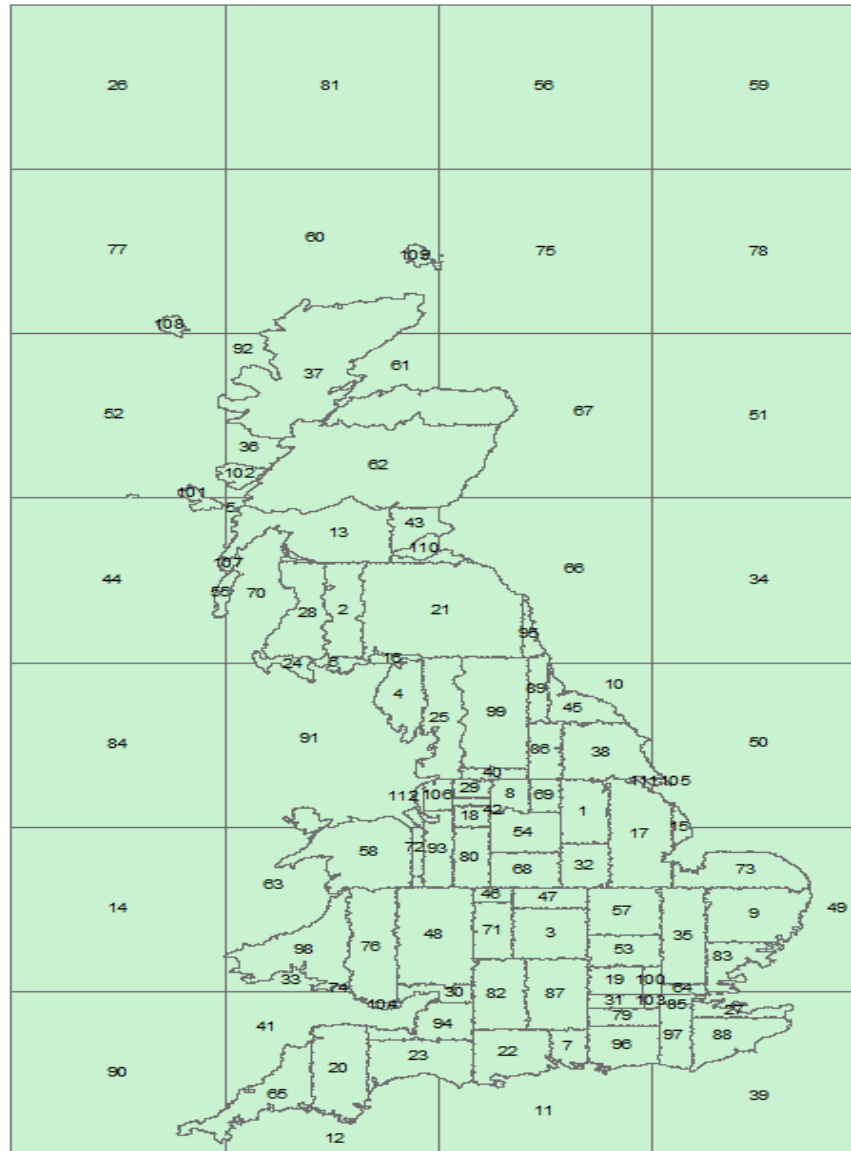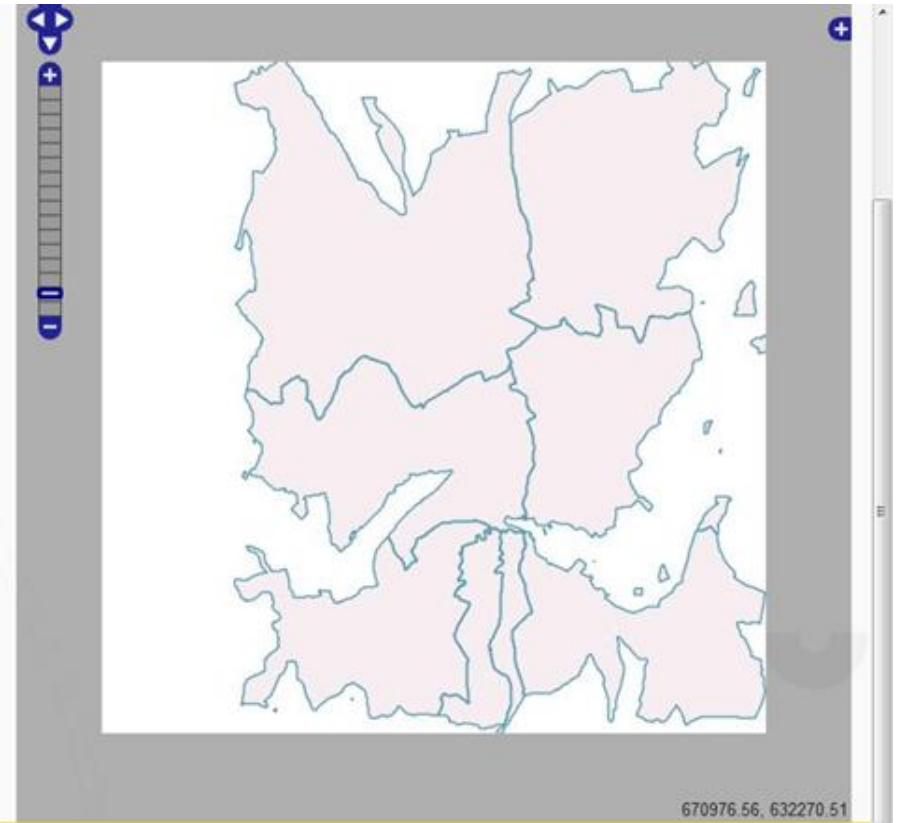Not possible
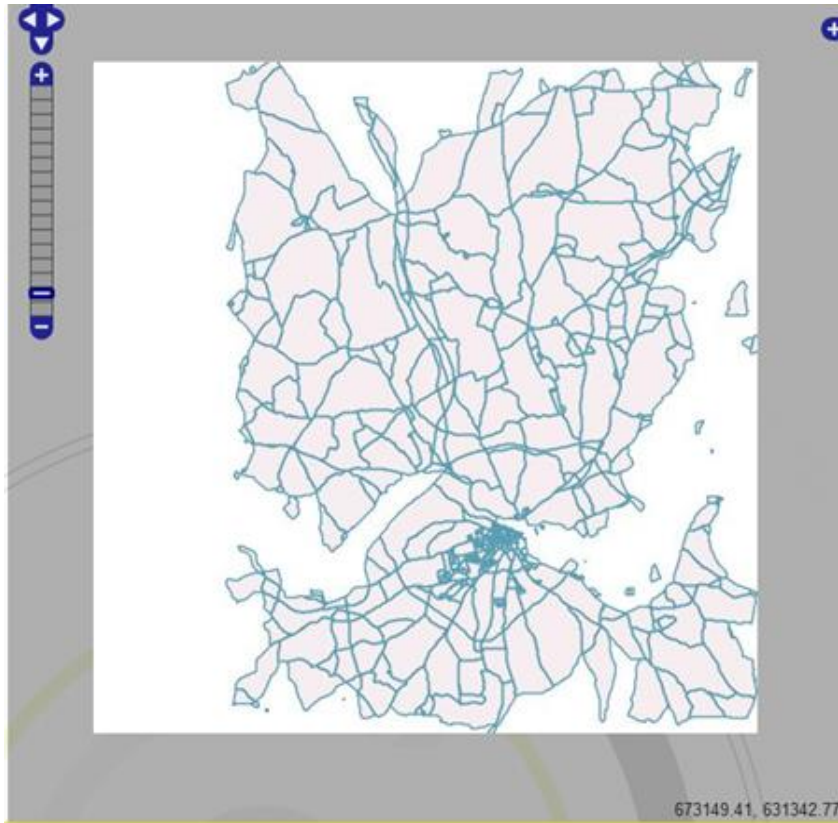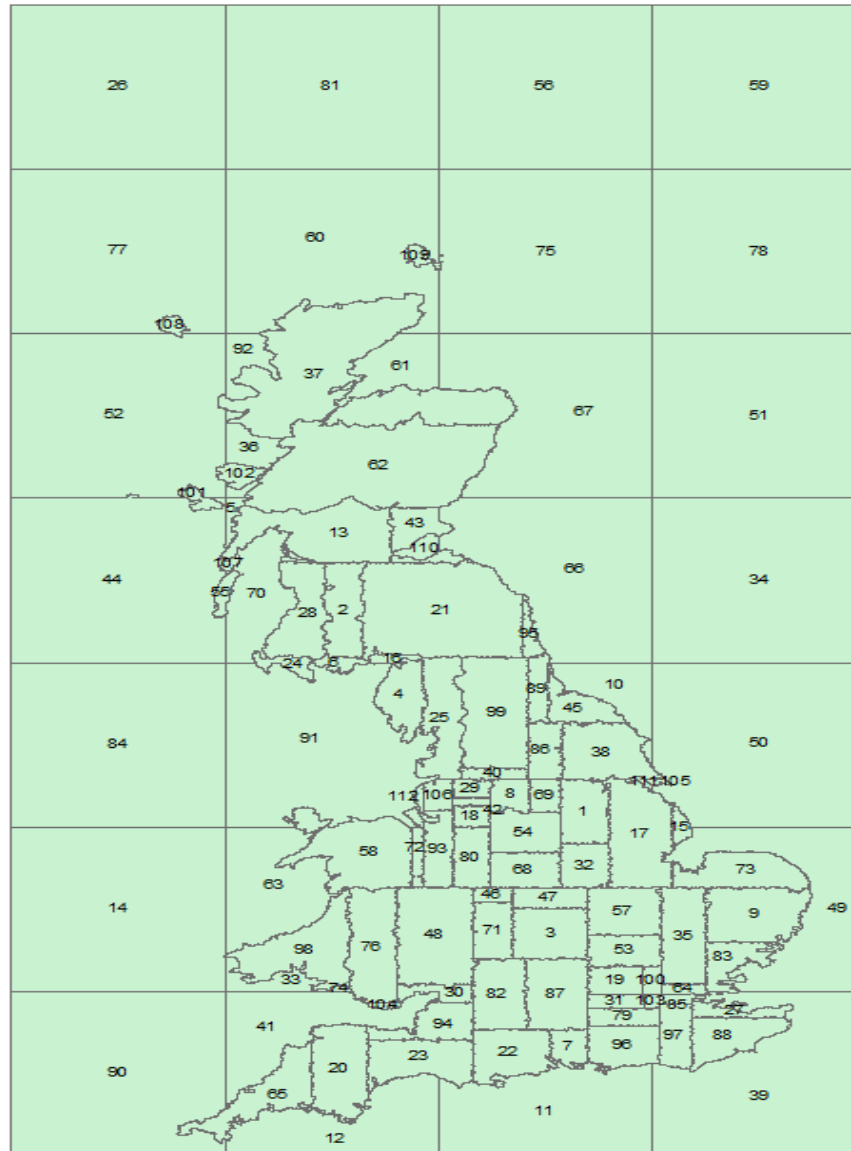
1Integrate/1Validate

Validation for the Nation

Validation for the Nation
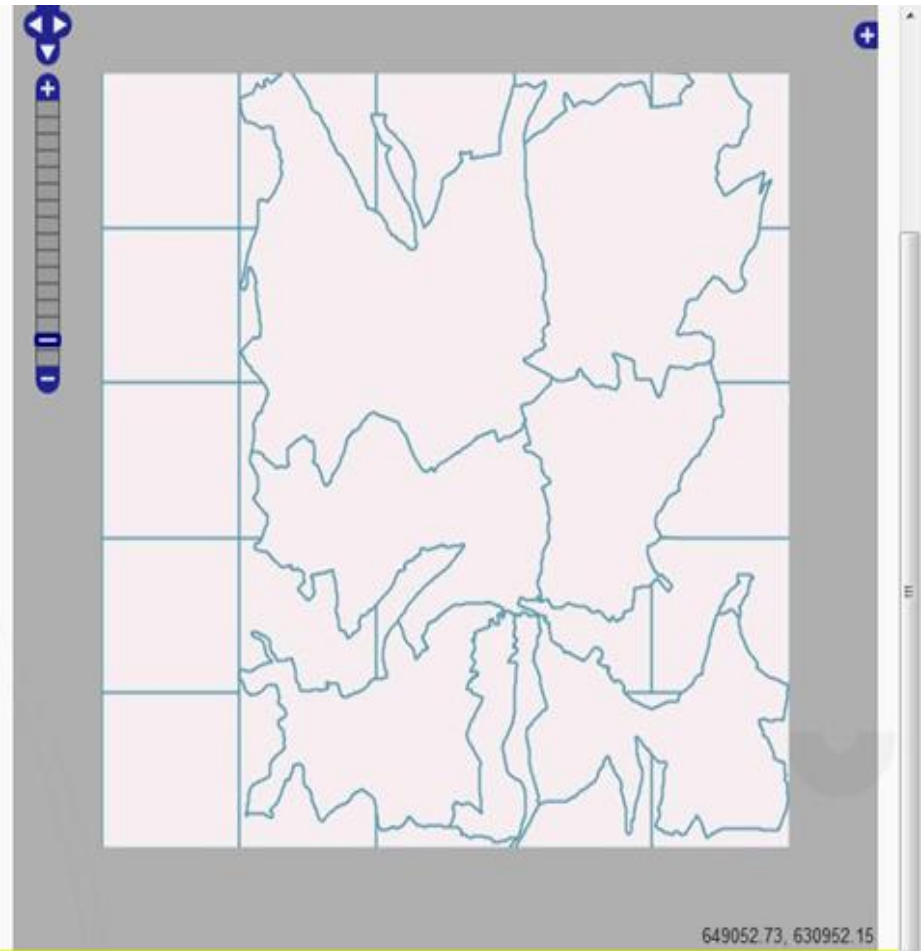
Validation for the Nation

Validation for the Nation

673149.41, 631342.77

670976.56, 632270.51

Validation for the Nation

Validation for the Nation

1spatial
YOUR WORLD SMARTER

Chart | Session

62 WAITING

9 RUNNING   117 FINISHED
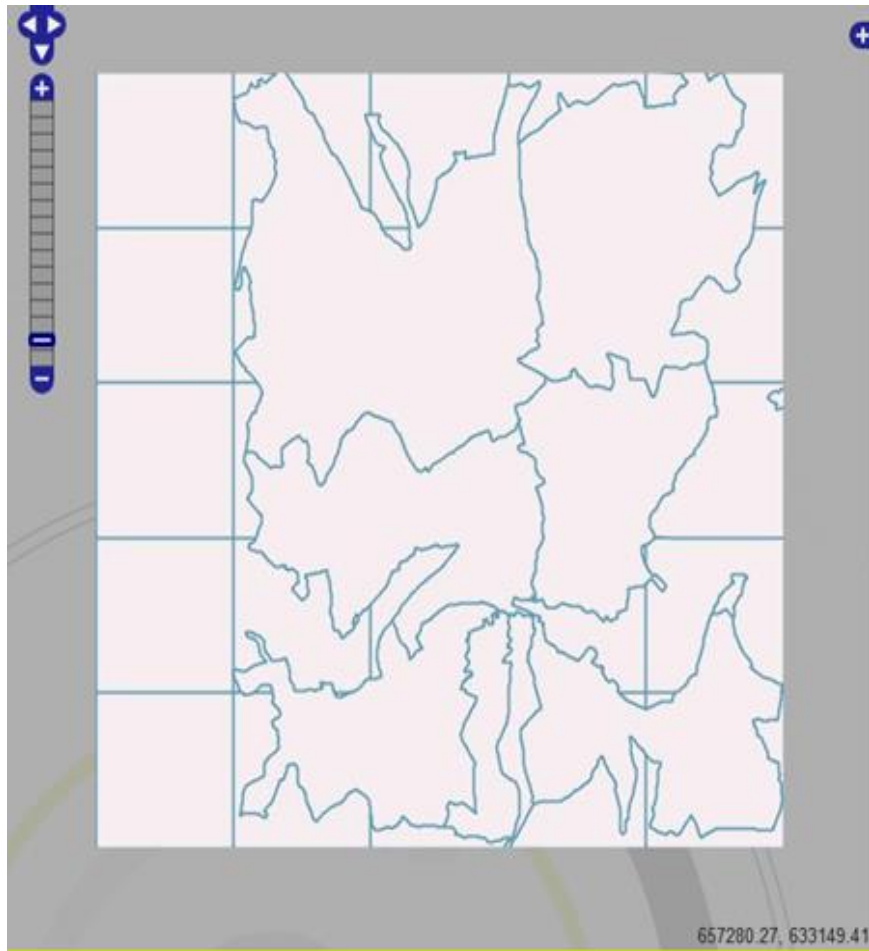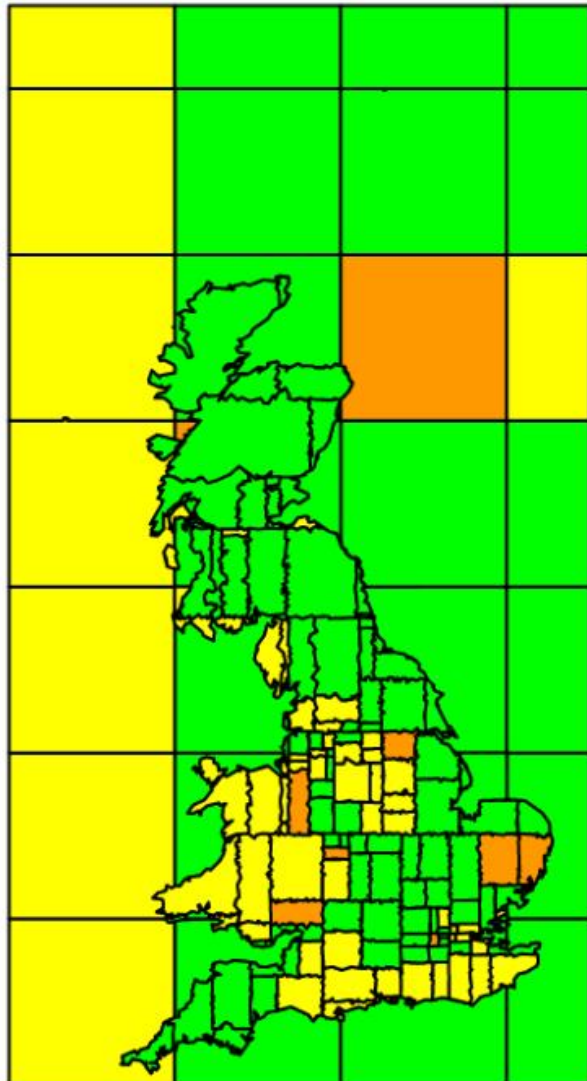
Chart | Session

📈 Timing ⚠ Errors

1: OpenData ✔
2: CheckRules !

CheckRulesTask

Started   5/16/16 6:46 PM
Duration a minute

Cattle Grid	9
Closing Primitive	
Construction Site	4
Construction Site Limit	7
Edge Of Dataset	
Edge Of Supported Structure	
Gate	
Gate + Leading Step Tread	
Inland Water	
Inter Tidal Road	
Inter Tidal Water	
Land	
Leading Step Tread	
Made Surface	

Task 1 Details - DebugTask

▼ Summary

	Minimum	Average Per Job	Maximum	Total	Per Feature
Duration	a few seconds	a few seconds	a few seconds	a few seconds	ms
Processed	15	15	15	2,820.00	1
Total	15	15	15	2,820	1
Errors	0	4.351	10	818	290.071 / 1000

▼ Task Results for All Sessions

Filter Results...

Session	Processed	Total	Errors	Started	Duration
Partition 0	15	15	2		a few seconds
Partition 1	15	15	4		a few seconds
Partition 10	15	15	3		a few seconds
Partition 100	15	15	4		a few seconds
Partition 101	15	15	6		a few seconds
Partition 102	15	15	3		a few seconds
Partition 103	15	15	5		a few seconds
Partition 104	15	15	7		a few seconds
Partition 105	15	15	5		a few seconds
Partition 106	15	15	3		a few seconds
Partition 107	15	15	7		a few seconds
Partition 108	15	15	4		a few seconds
Partition 109	15	15	5		a few seconds
Partition 11	15	15	3		a few seconds
Partition 110	15	15	7		a few seconds
Partition 111	15	15	3		a few seconds
Partition 112	15	15	5		a few seconds
Partition 113	15	15	7		a few seconds

Chart | Session

📈 Timing ⚠ Errors

- 1: OpenData ✔ 36.783
- 2: CheckRules ! 30.936
- 3: Pause ⏸

CheckRulesTask

Started 5/16/16 6:02 PM
Duration a minute

Cattle Grid	9
Closing Primitive	488
Construction Site	4
Construction Site Limit	7
Edge Of Dataset	106
Edge Of Supported Structure	94
Gate	37
Gate + Leading Step Tread	
Inland Water	272
Inter Tidal Road	
Inter Tidal Water	577
Land	5.415
Leading Step Tread	15
Made Surface	580

Session

Partition 0: RUNNING

Timing ⚠ Errors
1: OpenData ✔ 36.783
2: CheckRules ⟳ 17075

OpenDataTask

01-release-mapped-datastore

Started 5/16/16 6:45 PM
Duration a minute

Feature	Count
A Road Dual Carriageway	190
A Road Extent	1
A Road Link	
A Road Roundabout	110
A Road Single Carriageway	409
A Road Sliproad	19
Access	
Access Link	
Admin Boundary Link	1
Admin Boundary Node	2
Admin Exception	
Admin Unit	
Administrative	
Anomaly	
Anomaly Area	
Anomaly Point	
Area	
Artificial Cartographic Landform	7
Artificial Cartographic Slope	56
B Road Dual Carriageway	
B Road Extent	2
B Road Link	
B Road Roudabout	
B Road Single Carriageway	60
B Road Sliproad	
Base Or Side Of Cliff	507
Bottom Or Side Of Slope	60
Cartographic Cliff	516
Cartographic Landform Area	
Cartographic Landform Label	
Cartographic Landform Line	
Cartographic Landform Point	
Cartography	
Cattle Grid	9
Change of Mereing Boundary Node	
Classified Junction	
Cliff Edge	635
Closing Primitive	488
Coastal Slope	81
Compound Anomaly	
Construction Site	4
Construction Site Limit	7
Control	
Culvert Entrance	523
Decision Node	
Dual Carriageway	
Edge Of Dataset	106
Edge Of Supported Structure	94
Electricity Pylon or Lattice Tower Area	
Feature	

Rule	Value		
GM8.2	8.200	Processed	8,200
R11.1	22.676	Error	0
R81.2	60	Count	239
		Total	8,200

Video
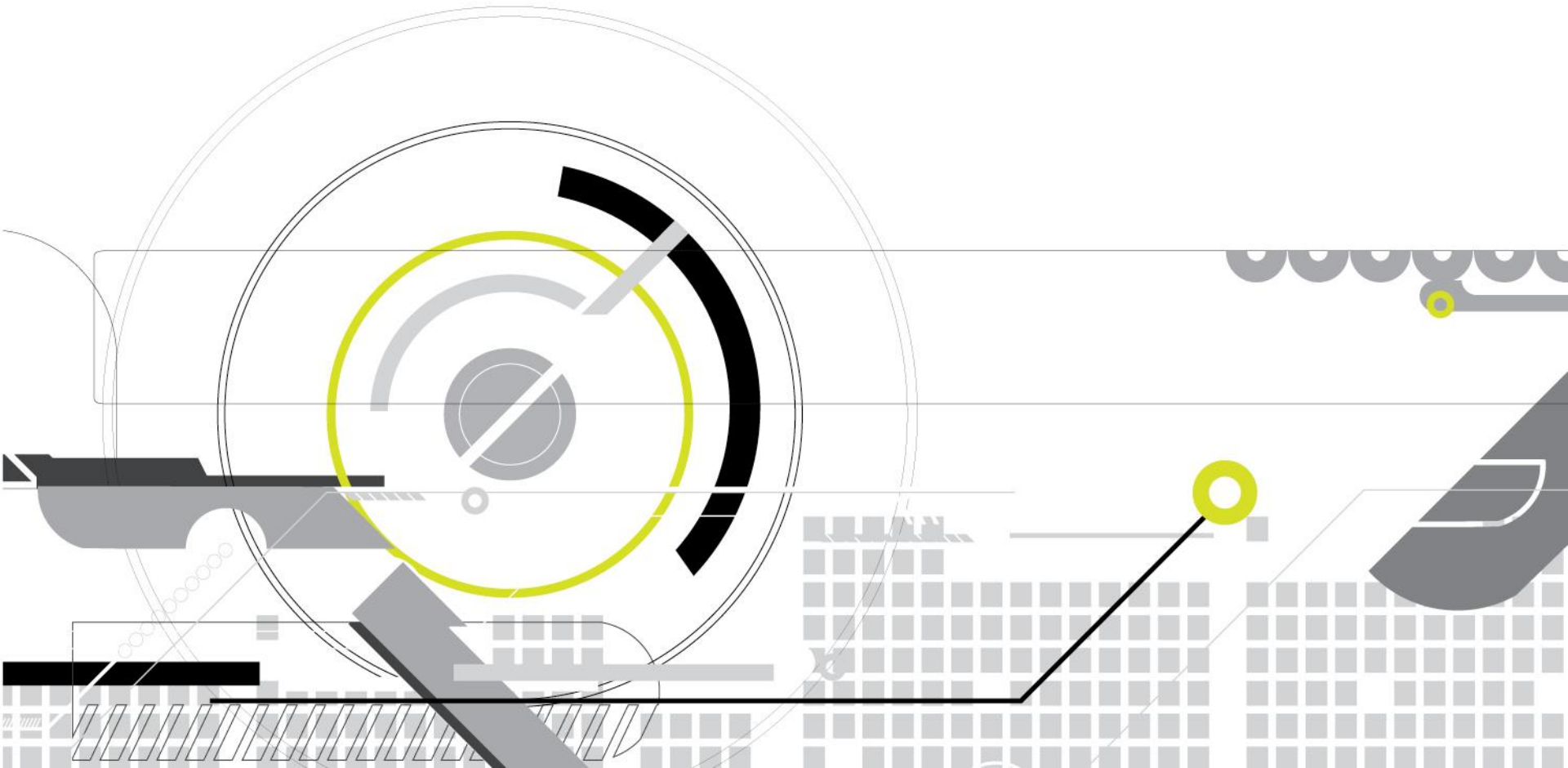
Questions

Follow us @1spatial