Geospatial Big Data: Software Architectures and the Role of APIs in Standardized Environments

Apache Big Data Europe 2016

Ingo Simonis
Open Geospatial Consortium
isimonis@opengeospatial.org

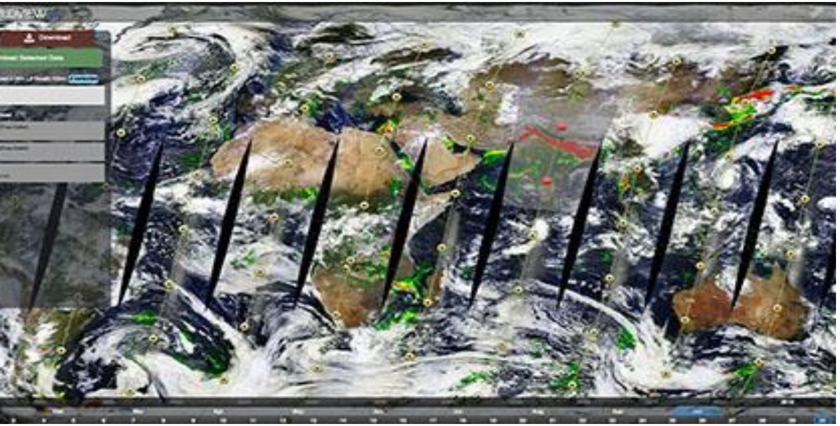
slideset contains material from G. Percivall (OGC), R. Winterton (Pitney Bowes), J. Sanyal (ORNL), A. Asahara (Hitachi), J. Spinney (Pitney Bowes), T. Kolbe(TUM), Rob Emanuele (LocationTech) , ESA

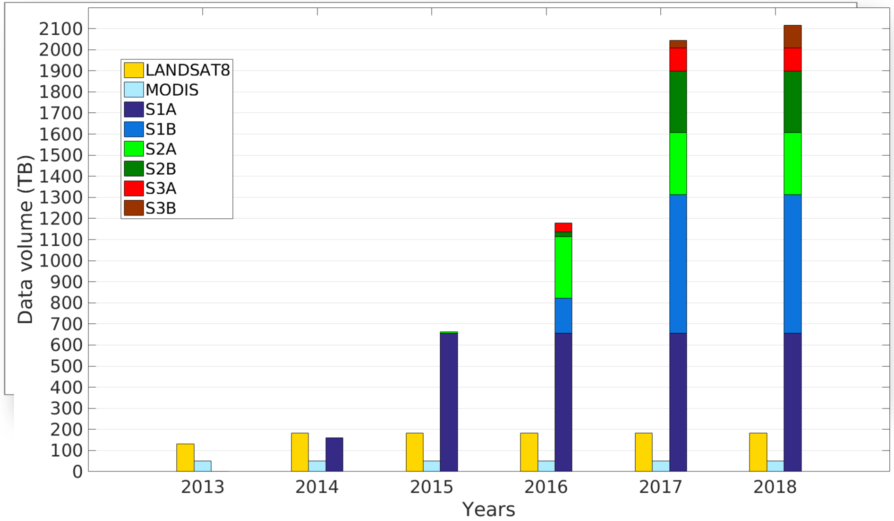
Geospatial Data

- Spatial data is big data
- Apache projects are implementing geospatial functionalities
- Coordination of spatial implementations across Apache projects
- Open standards to increase interoperability and code reuse
- Architectures integrating Big Data Services and Geospatial Services

Earth Observations

- Big Earth Data Initiative (BEDI) - Standardizing and optimizing collection, delivery of U.S. Government's civil Earth observation data.
- Sentinel satellites operated by ESA in the framework of the Copernicus programme funded and managed by the European Commission.





Commercial Cloud Hosting

- DigitalGlobe
  - Entire DG archive in cloud in 2016 - 45 PB - largest EO archive
  - Harris/ENVI processing
- Google Earth Engine
  - 5 PB storage (Landsat and others) - 800+ Library Functions
  - Limiting factor is the ability to pull data from another cloud to support local processing.
- Hexagon Geospatial
  - Cloud hosted dynamic information service
  - AirBus archive in Amazon cloud with Hexagon services

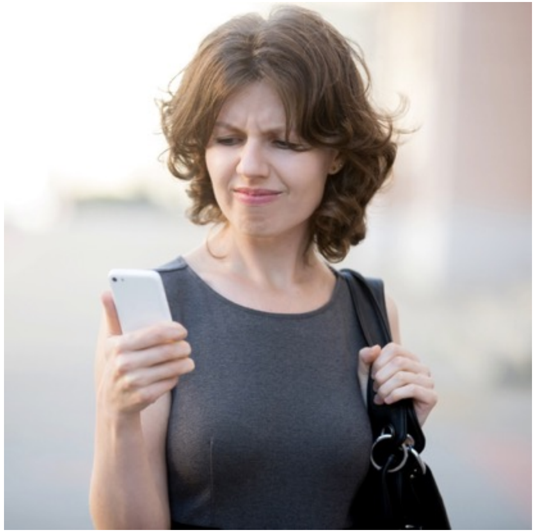
Geo-Enrichment

Allows a wide variety of datasets to be appended to a data record

- What are the property attributes of this insured property?
- What demographic group does this customer belong to?
- What businesses are connected with this area of poor network coverage?

Geo-Analytics

Reduce the complexity of billions of transactional records by assigning data to geographic bins and aggregating results.



- Is the average 4G network coverage in this area better than a competitor?
- Is this data point inside or outside of a geo-fence?

Network Coverage and Performance

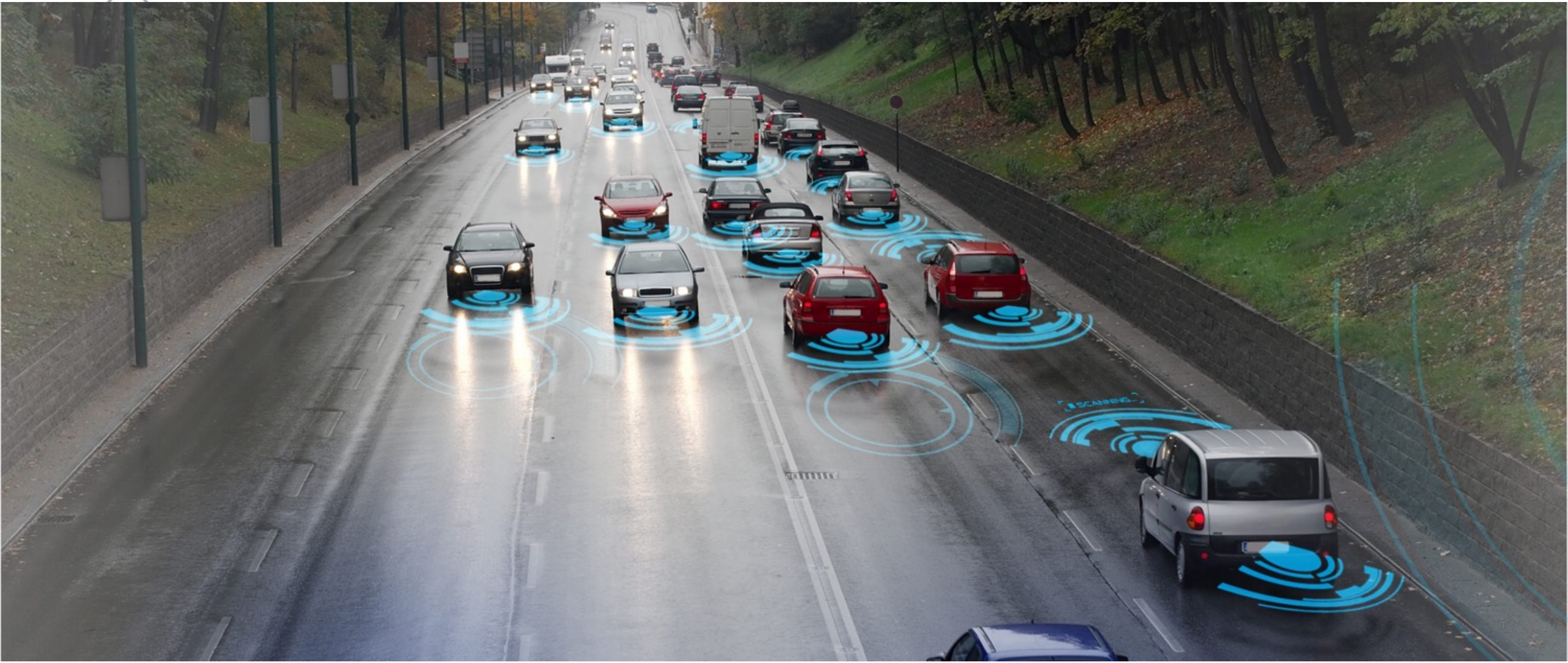
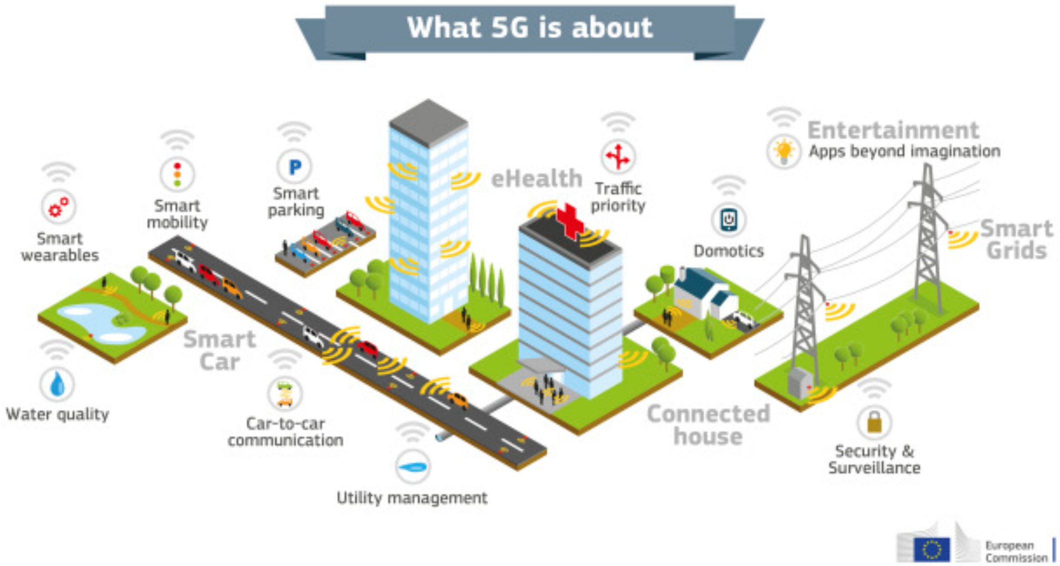


image by: http://barrachd.co.uk

Connected cars will send 25 gigabytes of data to the cloud every hour

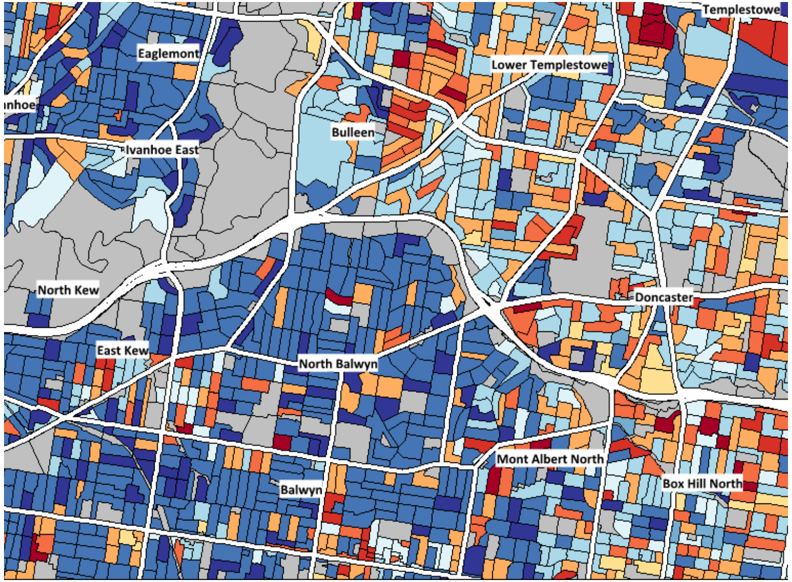
Network Coverage and Performance



- Single view on network:
  - Improve quality of service
  - Increase net promoter scores
  - Enable acquisition
  - Reduce churn

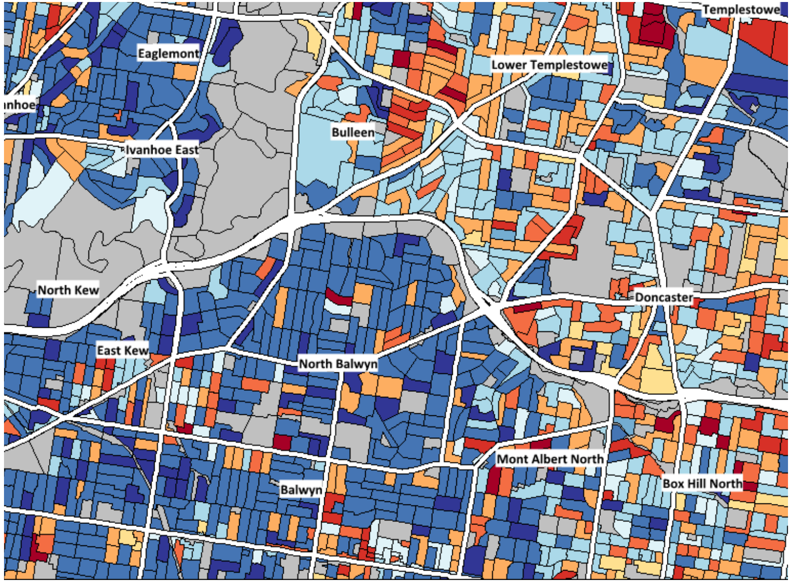
Layered Information

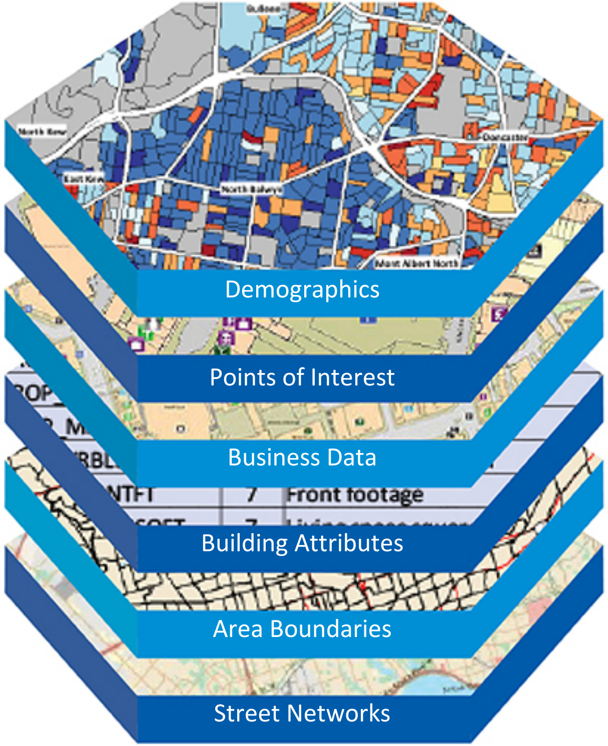
Demographics



Layered Information

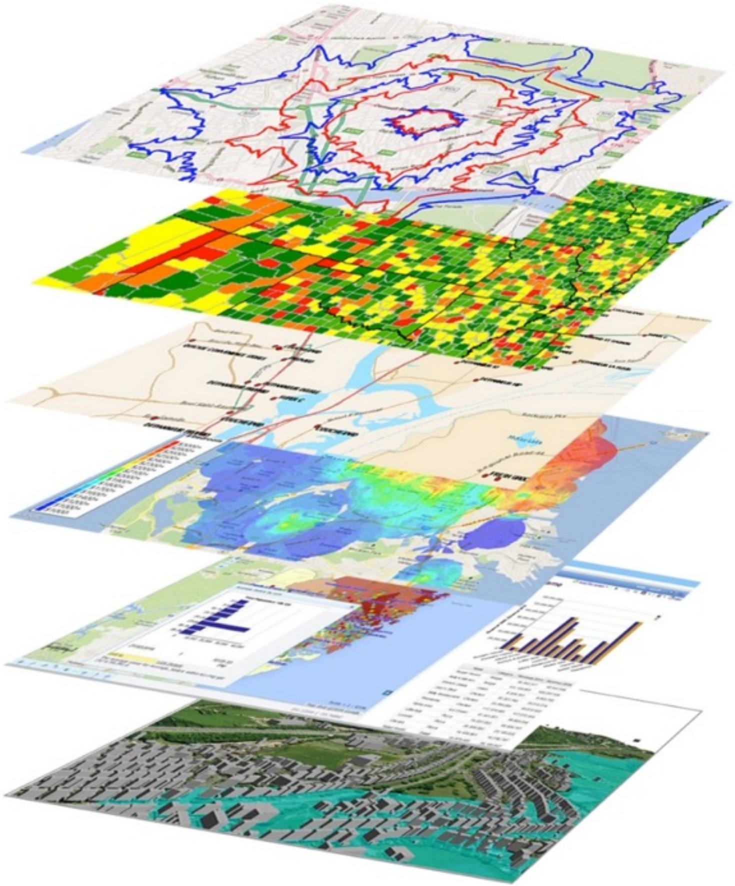
Demographics





Each layer of location data adds new details and additional insight

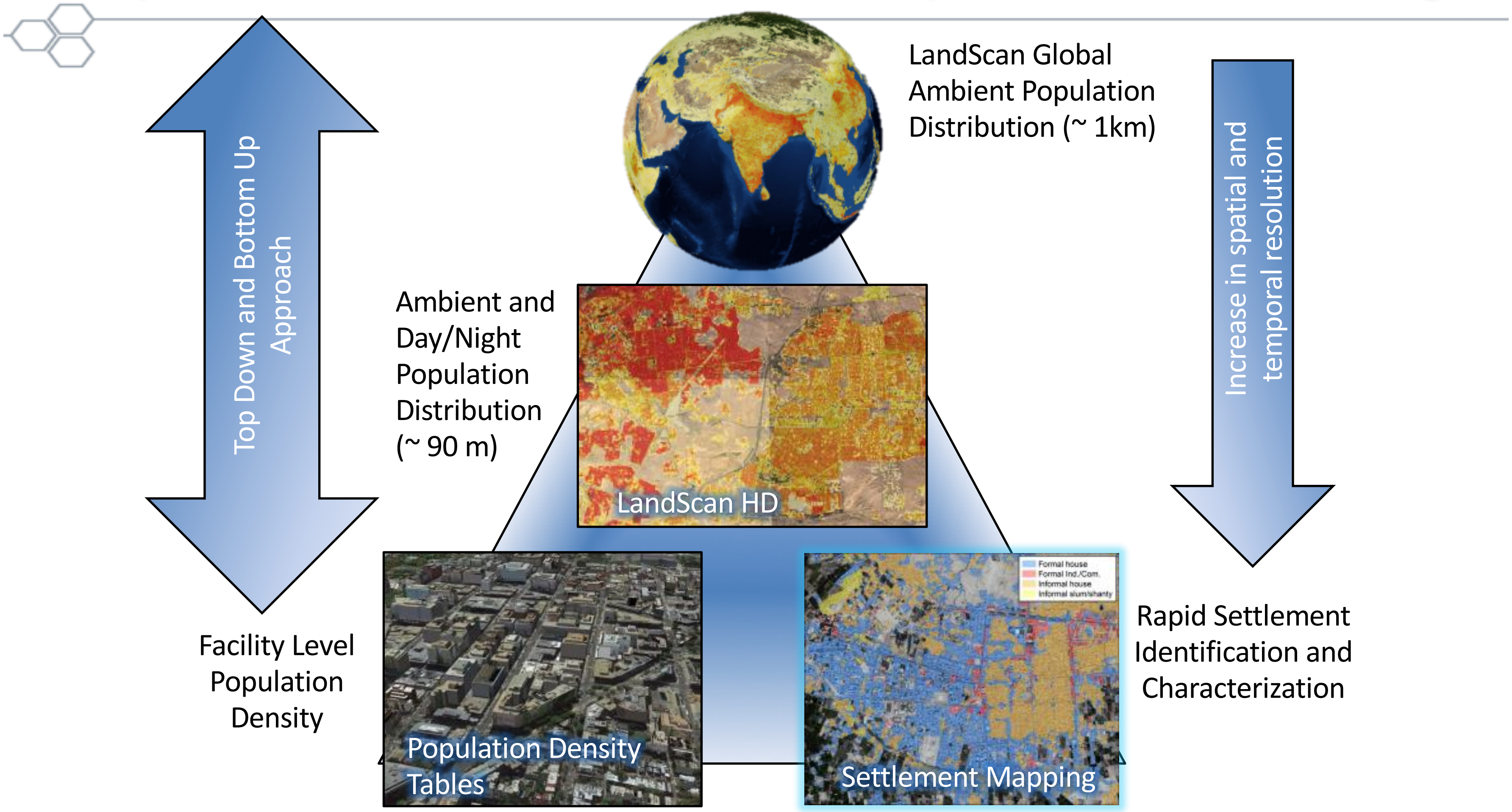
All Domains: Profit from a precise perspective



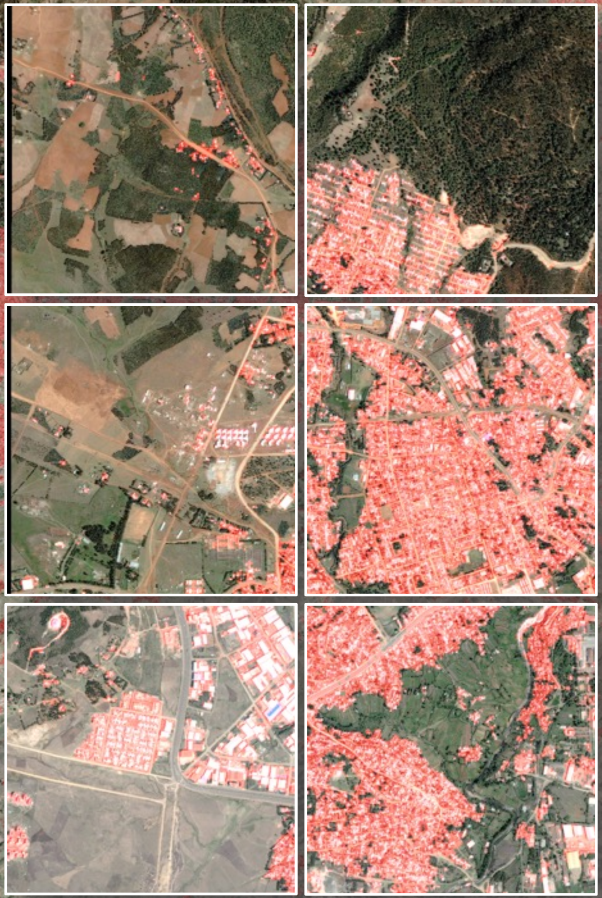
Insurance

- Single view of risk
- Usage based insurance
- Fraud detection

Population Distribution and Dynamics Modeling

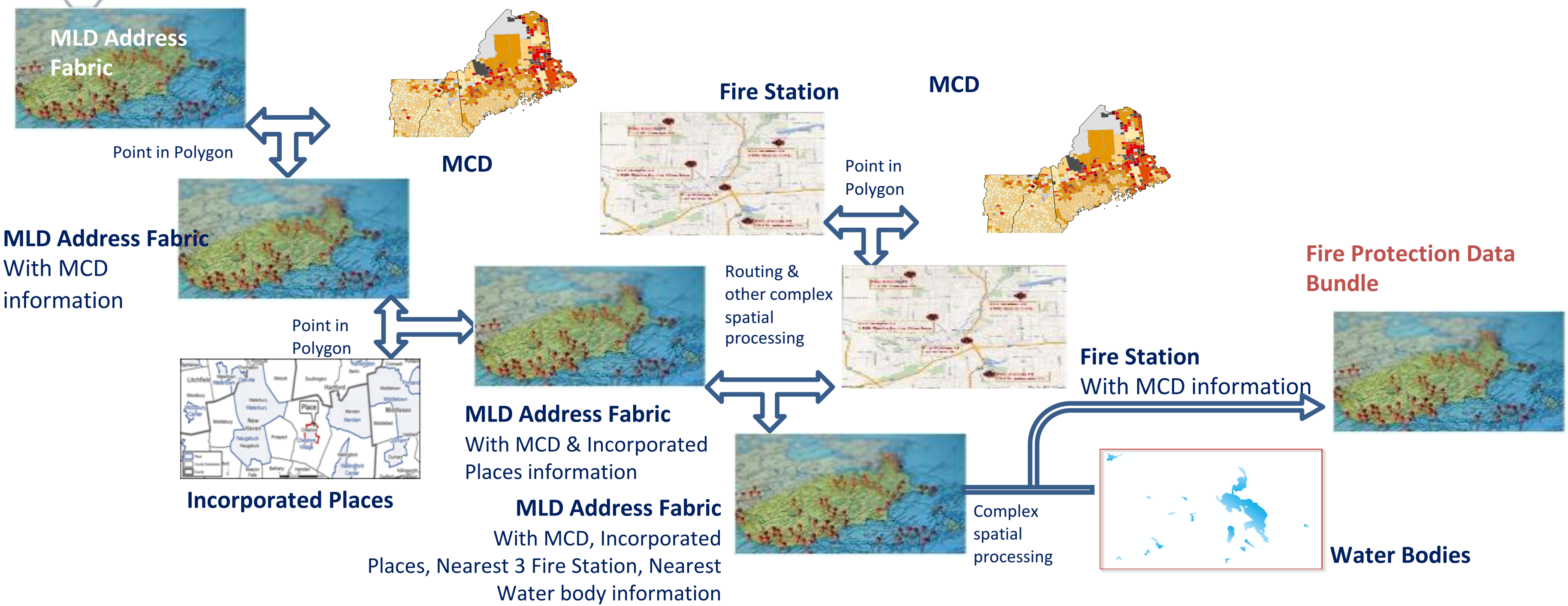


Addis Ababa, Ethiopia



- 2 Xeon Quad core 2.4GHz CPUs + 4 Tesla GPUs + 48GB
- Image analyzed (0.3m)
    - 40,000x40,000 pixels (800 sq. km)
    - RGB bands
- Overall accuracy 93%
    - Settlement class 89%
    - Non-settlement class 94%
- Total processing time
    - 27 seconds

Complex Work Flows



172 Million Addresses to Closest 3 Fire Stations & Nearest Water Boundary- 6 hours Using 10 Node Elastic Map Reduce

Ecology Mapping

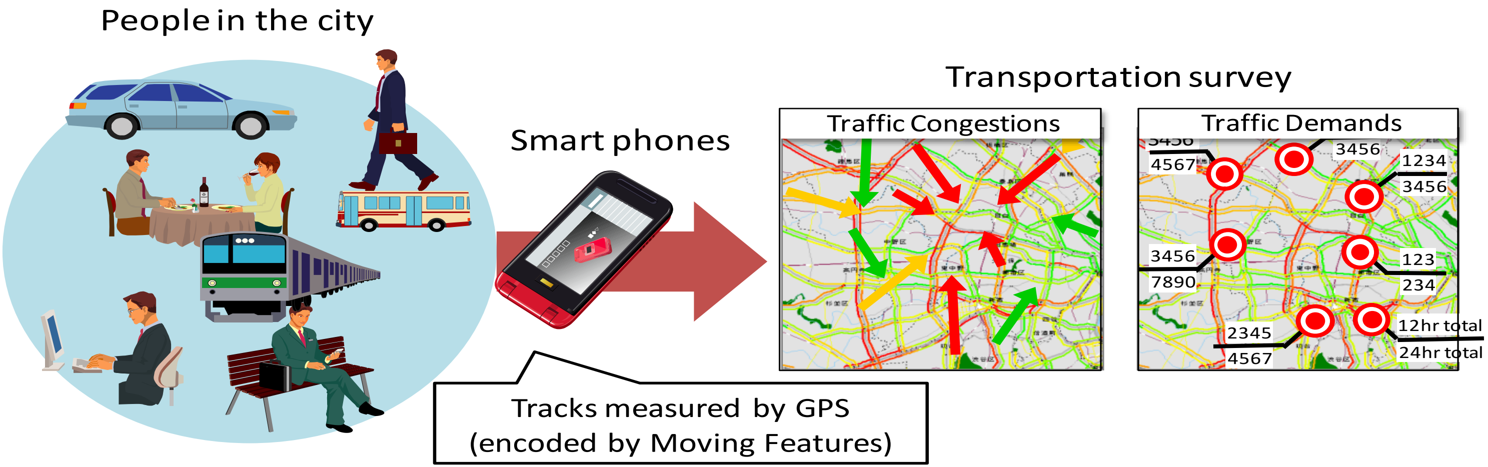
- 1 km sq grid of US each with nine variables, e.g., days below freezing, amount of precipitation in growing season
- Unsupervised statistical multivariate clustering
- Domains: tundra, prairie, alpine, and southeastern forest

NSF NEON Ecological Domains

Science 23 April 2010:
Vol. 328. no. 5977, pp. 418 - 420 DOI:
10.1126/science.328.5977.418

Transportation

- To reduce traffic congestion, trip demand data collected using transportation surveys
- GPS based data collection of trip information is applicable, with the broad availability of location enabled mobile devices
- The GPS tracks are encoded by Moving Features to enable sharing by many stakeholders such as local governments, bus companies, and so on.



Location Based Marketing

PAST	PRESENT	FUTURE
Behaviors & Actuals	Contexts & Possibilities	Predictions & Potentials

City Models for Smart Cities

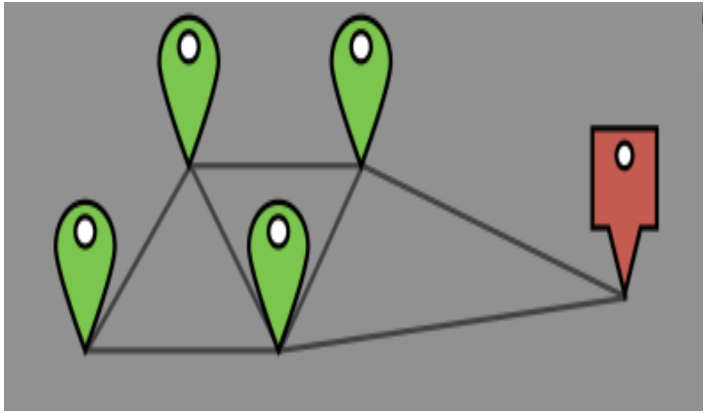
- **Berlin**
  - >500,000 buildings upto Level of Detail 4
  - Modeled according to CityGML
  - Basis for real estate
  - Integration of sensors
- **New York**
  - 1M buildings plus roads at LoD 1
  - NYC Open data
  - Next - Underground critical infrastructure

Geospatial Standards

- Location
- Geometry
- Features
- Coverages
- Sensors and Observations
- Processing, Analytics
- Web Services

Power of Location

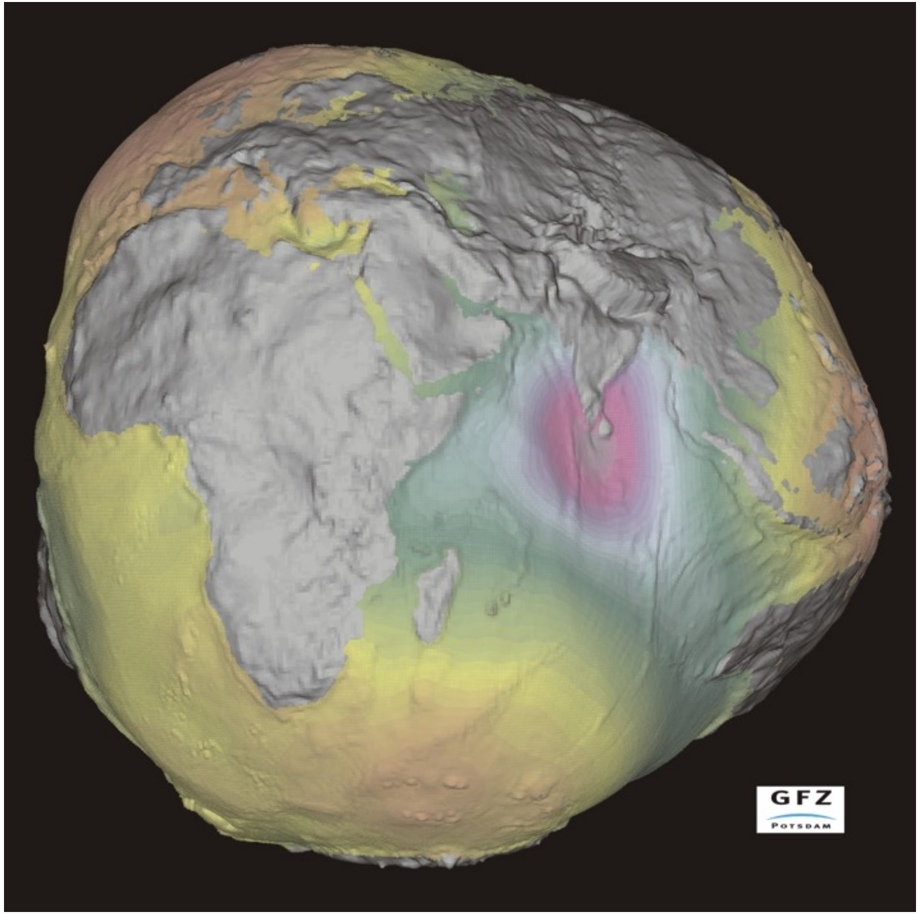
- **1st law of geography**: "Everything is related to everything else, but near things are more related than distant things."
  - Waldo Tobler



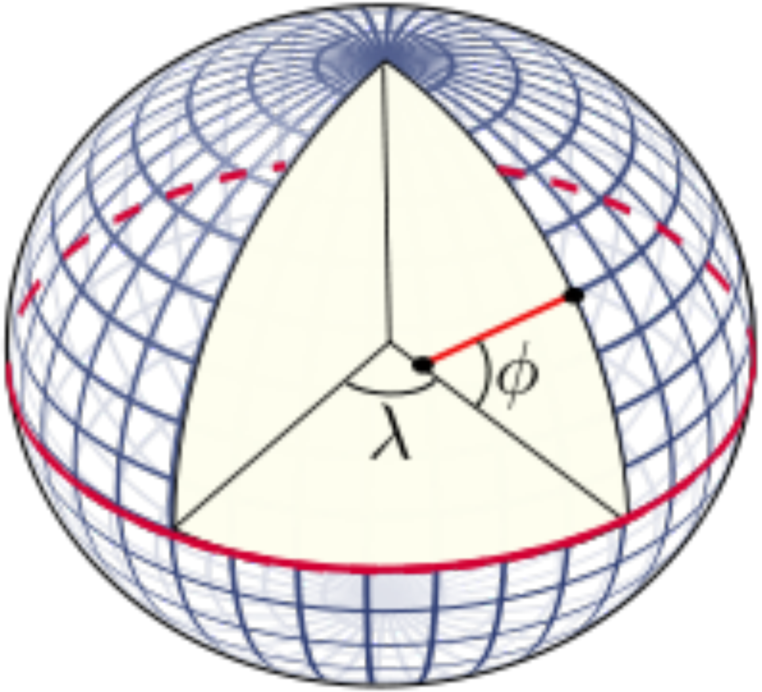
- By measuring entropy of individual's trajectory, we find **93% potential predictability in user mobility**
  - Limits of Predictability in Human Mobility, Science 2010

Some Peculiarities about *Spatial*

Some Peculiarities about *Spatial*

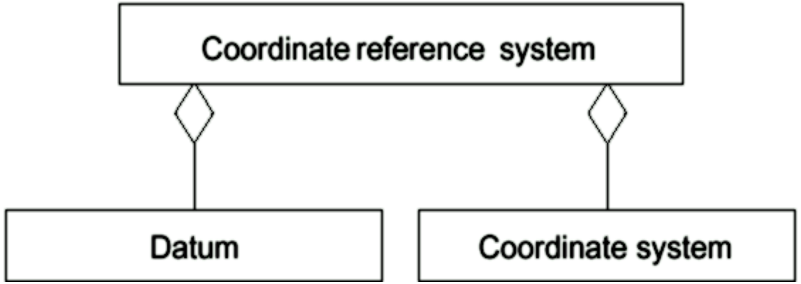


Latitude is not unique !



nor is Longitude!

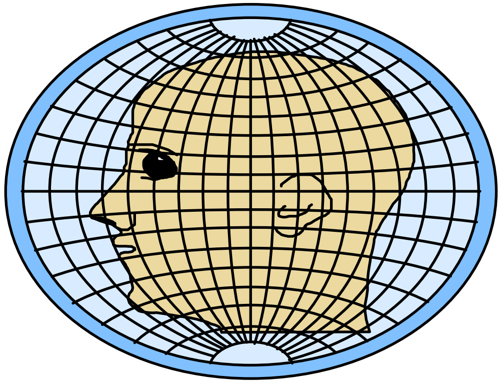
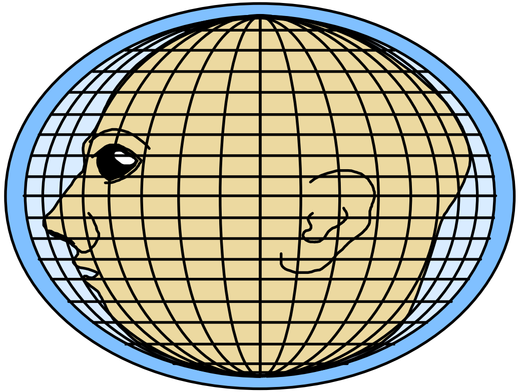
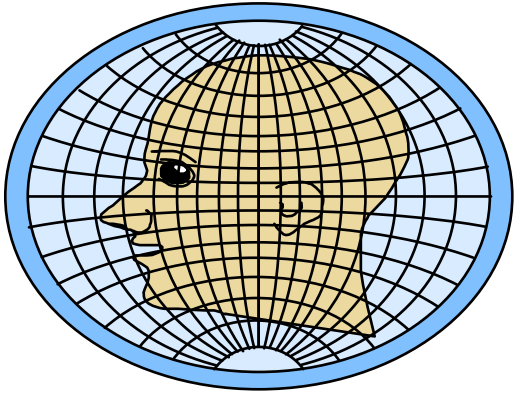
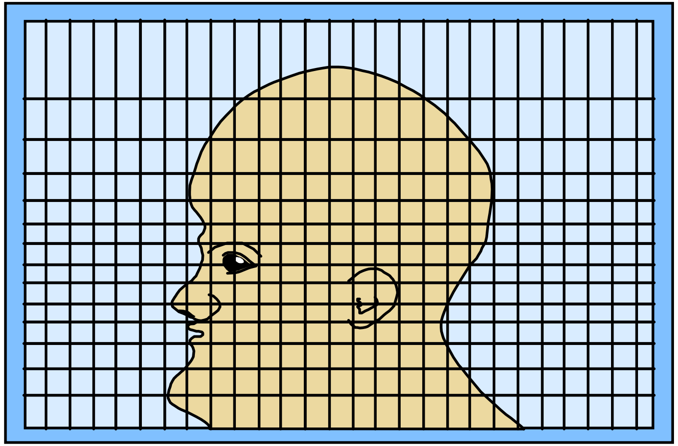
Coordinate Reference Systems



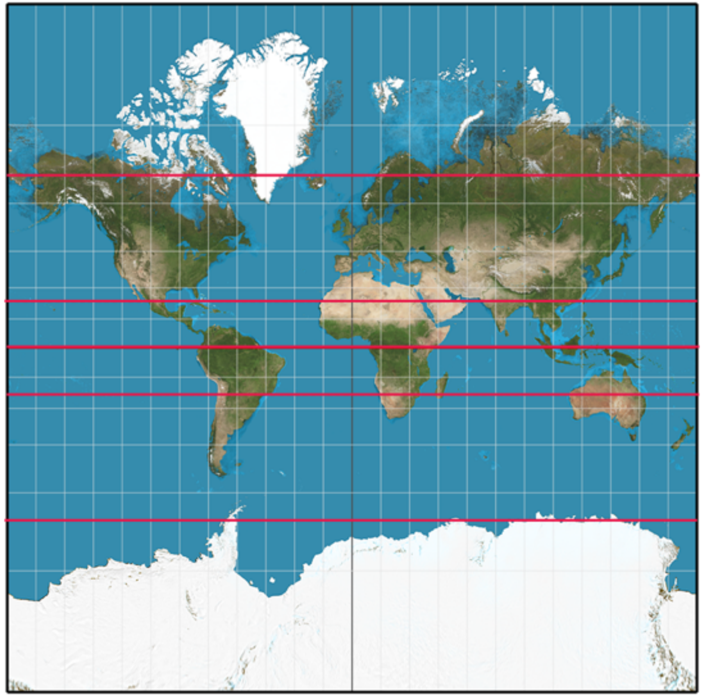
- Coordinate
  - one of a sequence of N numbers designating the position of a point in N-dimensional space
- Coordinate Systems
  - Cartesian 2D and 3D
  - Spherical (3D), Polar (2D)
  - Cylindrical
  - Linear - along a path
  - Ellipsoidal
- Coordinate Reference System
  - coordinate system related to real world by a datum
- Examples
  - Geographic
  - Geocentric
  - Vertical
  - Engineering
  - Image
  - Temporal
  - Derived CRS, e.g., projections

Reference ISO 19111 and OGC Abstract Spec Topic 2

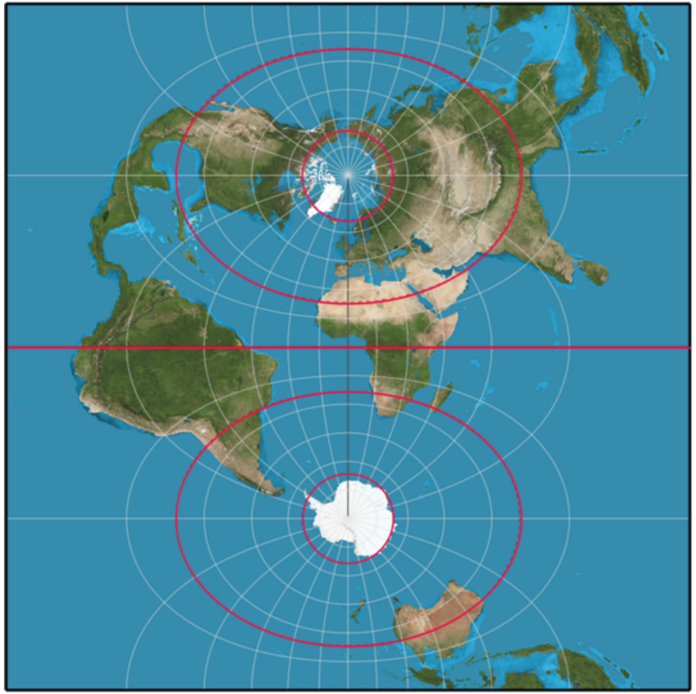
A familiarly shaped ‘continent’ in different map projections

Map Projections

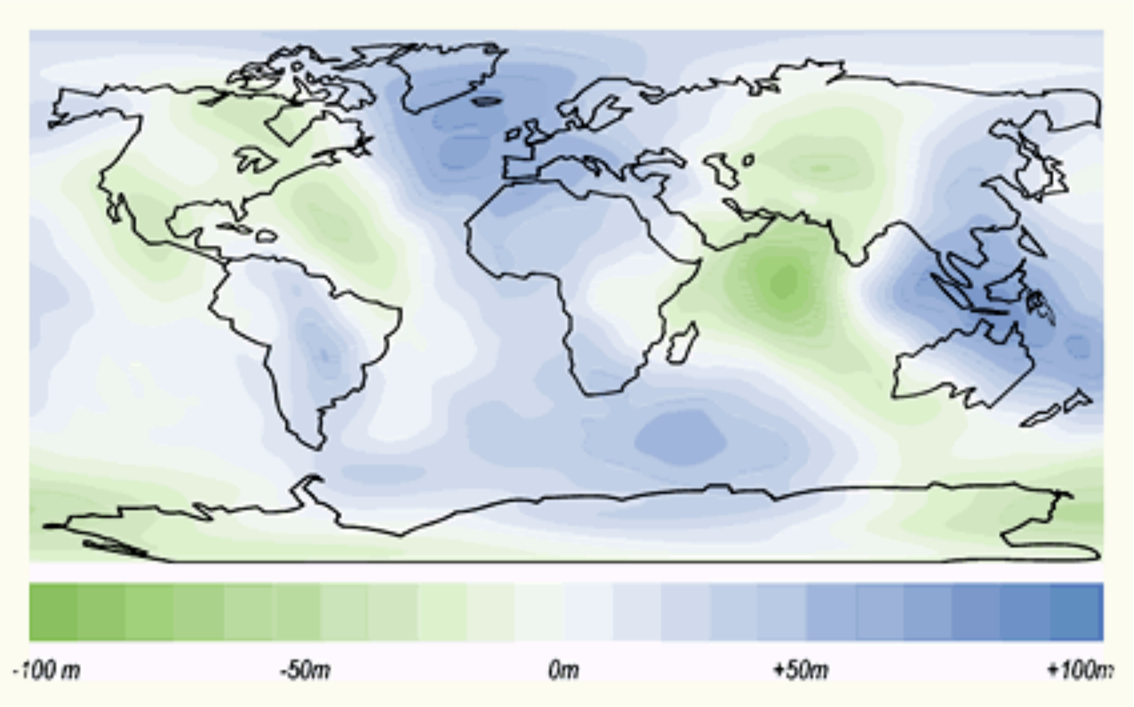


Mercator



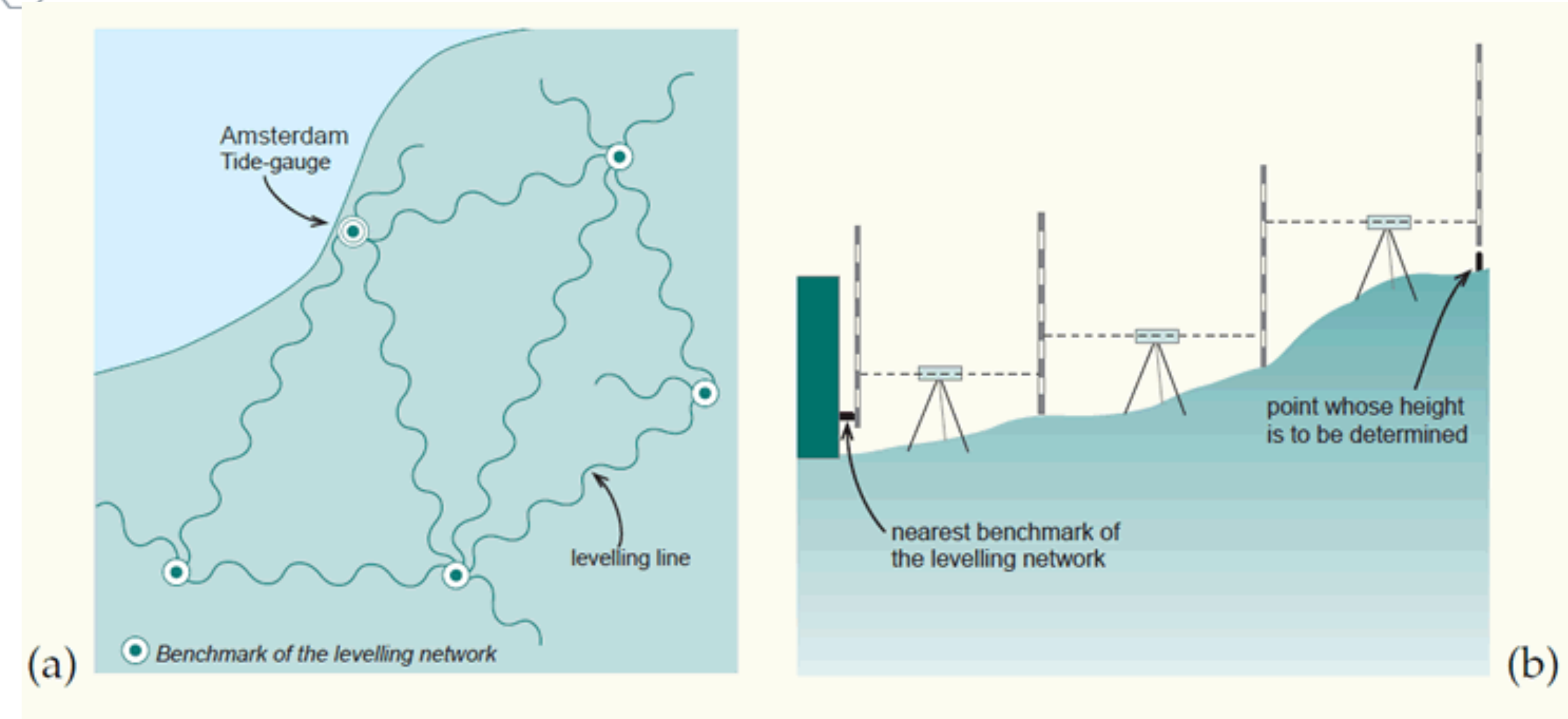
Transverse Mercator

What errors can you expect?



Deviations (undulations) between the Geoid and the WGS84 ellipsoid

Sea Level



Local vertical datum

Sea Level

The Netherlands to Belgium: -2.34m!

No Metadata – No Interpretation

- No geodetic metadata → coordinates cannot be interpreted
  - datum
  - ellipsoid
  - prime meridian
  - map projection

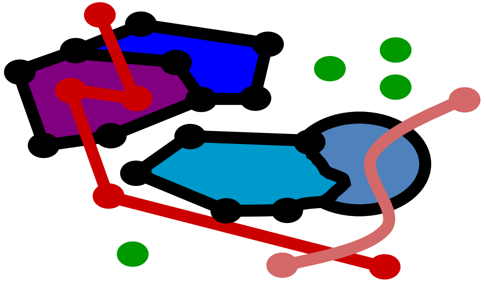
Hiding Geospatial Complexity

Martin Desruisseaux, Geomatys, presentation today about Apache SIS Project

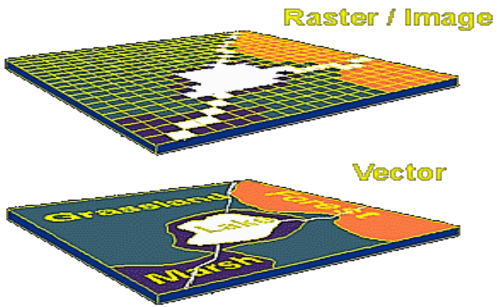
- It is tempting to ignore the complexity of geospatial international standards on the assumption that everyone today uses coordinates given by GPS.
- Apache SIS methods handle a lot of this complexity
- Martin will show example of what happen under the hood during a cube transformation, for demonstrating what the developers gain with SIS.

Geospatial Information

Feature Data



Coverage Data



Metadata

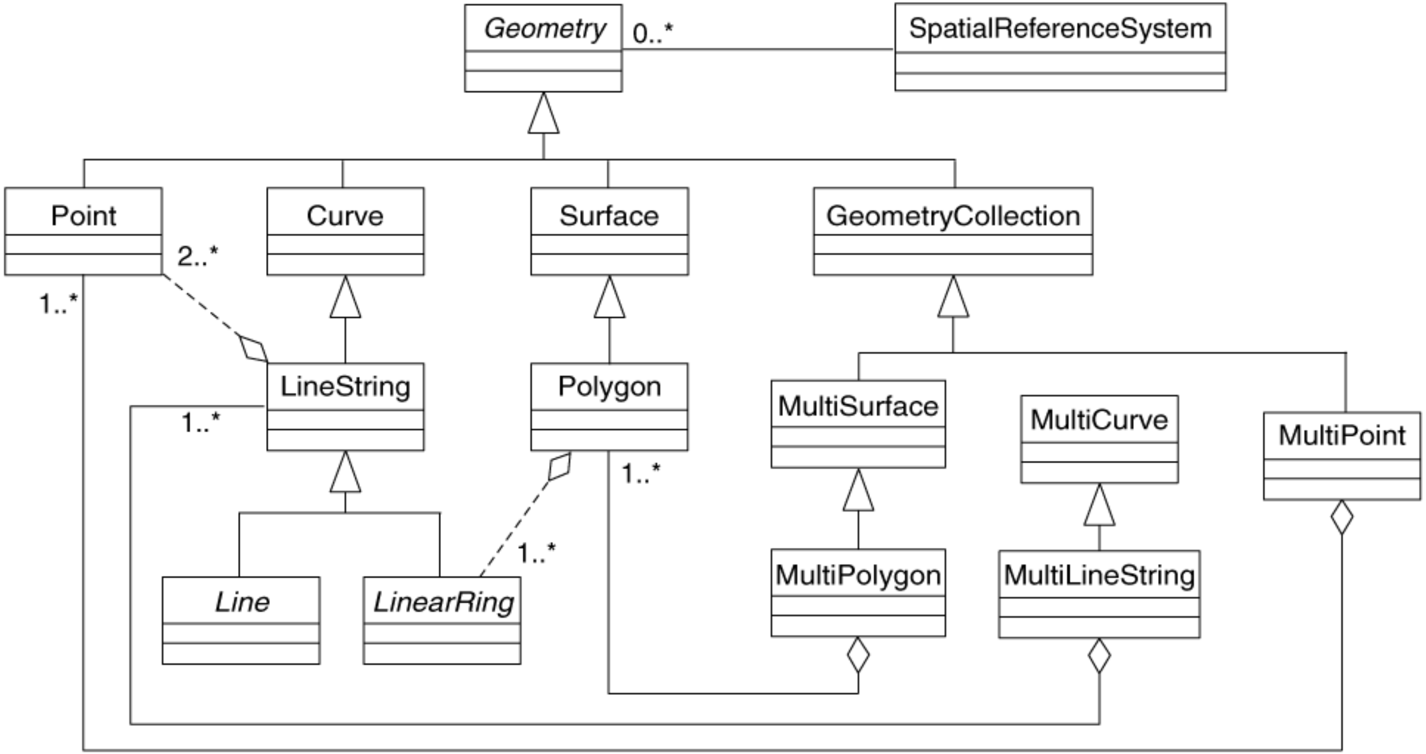
Nutrition Facts
Serv. Size 1/3 cup (56g)
Servings about 3
Calories 80
Fat Cal.10
*Percent Daily Values (DV) are based on a 2,000 calorie diet.

Amount/Serving	% DV*	Amount/Serving	% DV*
Total Fat 1g	2%	**Total Carb.** 0g	0%
Saturated Fat 0g	0%	Fiber 0g	0%
Cholest. 10mg	3%	Sugars 0g	
Sodium 200mg	8%	**Protein** 17g	

Vitamin A 0% • Vitamin C 0% • Calcium 6% • Iron 6%

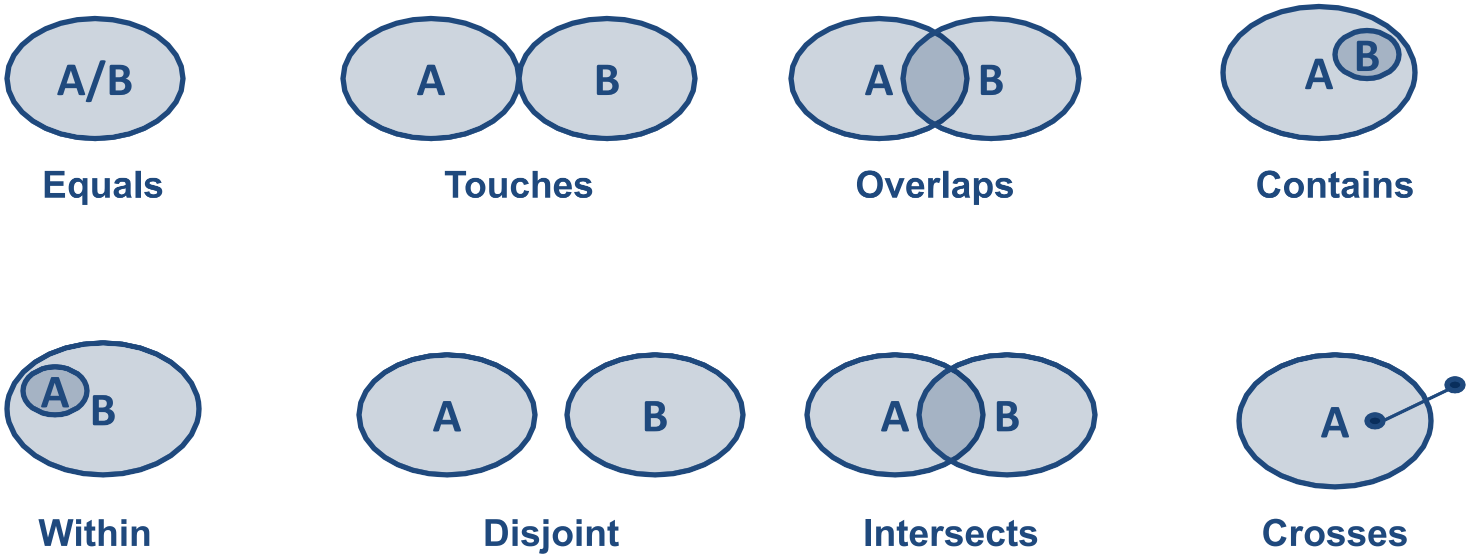
Maps

Simple Geometries for Simple Feature

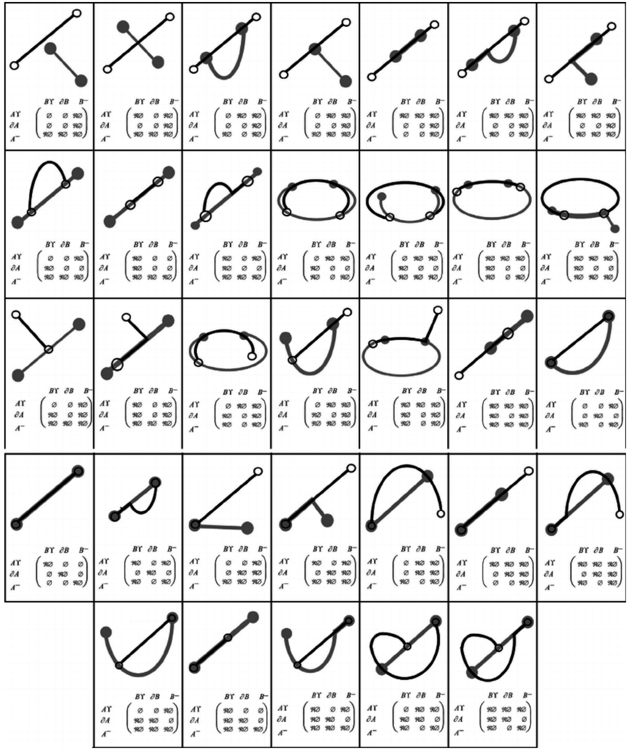


OGC simple features (ISO 1923) geometries are restricted to 0, 1 and 2-dimensional geometric objects that exist in 2-dimensional coordinate space (R2).

Topological Relations between Spatial Objects



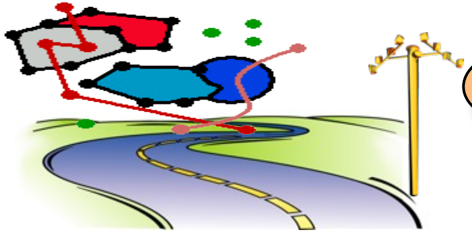
Topological Relations between Spatial Objects



GeoSPARQL for Topological Query Functions

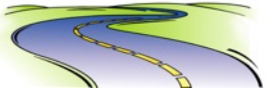
```
– ogcf:relate(geom1: ogc:WKTLiteral, geom2: ogc:WKTLiteral,
              patternMatrix: xsd:string): xsd:boolean
– ogcf:sfEquals(geom1: ogc:WKTLiteral,
                geom2: ogcf:WKTLiteral): xsd:boolean
– ogcf:sfDisjoint(geom1: ogc:WKTLiteral,
                  geom2: ogcf:WKTLiteral): xsd:boolean
– ogcf:sfIntersects(geom1: ogc:WKTLiteral,
                    geom2: ogcf:WKTLiteral): xsd:boolean
– ogcf:sfTouches(geom1: ogc:WKTLiteral,
                 geom2: ogcf:WKTLiteral): xsd:boolean
– ogcf:sfCrosses(geom1: ogc:WKTLiteral,
                 geom2: ogcf:WKTLiteral): xsd:boolean
– ogcf:sfWithin(geom1: ogc:WKTLiteral,
                geom2: ogcf:WKTLiteral): xsd:boolean
– ogcf:sfContains(geom1: ogc:WKTLiteral,
                  geom2: ogcf:WKTLiteral): xsd:boolean
– ogcf:sfOverlaps(geom1: ogc:WKTLiteral,
                  geom2: ogcf:WKTLiteral): xsd:boolean
```

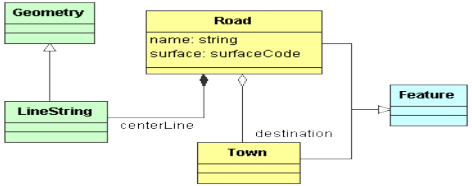
Geographic Features



Concepts *Vocabulary*
Abstract Models
Structure

Encodings
Implementation Specifications
Access





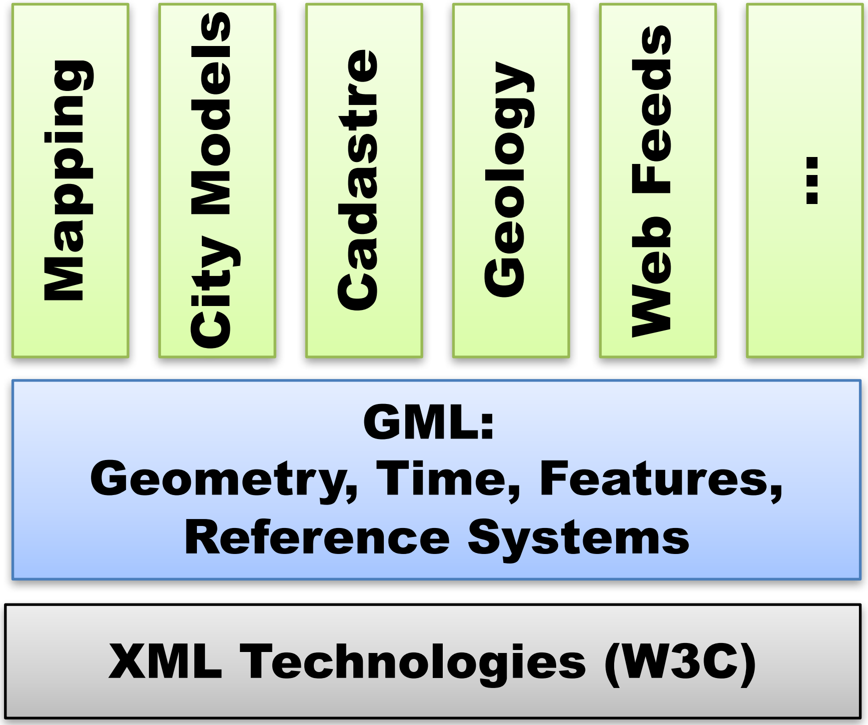
<MultiGeometry gid="c731" srsName="http://www.opengis.net/gml/srs/epsg.xml#4326"> <geometryMember> <Point gid="P6776"> <Coord><x>50.0</x><y>50.0</y></Coord> </Point> </geometryMember> <geometryMember> <LineString gid="L21216"> <Coord><x>0.0</x><y>0.0</y></Coord> <Coord><x>0.0</x><y>50.0</y></Coord> <Coord><x>100.0</x><y>50.0</y></Coord> </LineString> </geometryMember> <geometryMember> </MultiGeometry>

OGC Geography Markup Language

Two Different Usage Patterns

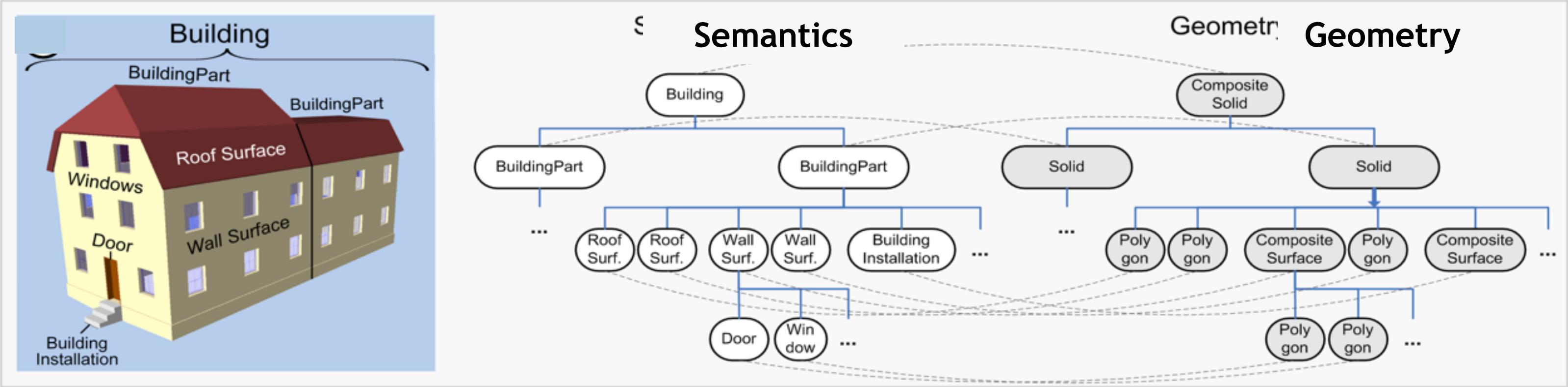
- **Thematic communities describe spatial datasets:** Cadastre, Topography, Geology, Hydrography, Meteorology, Aviation, City Models, etc.

- **Embed location in other XML grammars:** GeoRSS, GeoSPARQL (OGC), Geopriv (IETF), POI (W3C), Sensor Web (OGC), etc.



CityGML – Geometry and Semantics

CityGML: (Up to) Complex objects with structured geometry

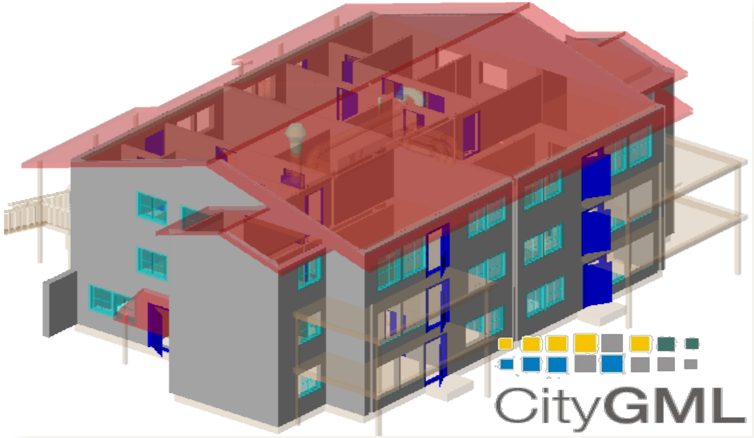


- Geometric entities know **WHAT** they are
- Semantic entities know **WHERE** they are and what their spatial extents are

CityGML and IndoorGML

1st layer: **Topographic space model**

- **building structure**
- geometric-topological model
- network for route planning

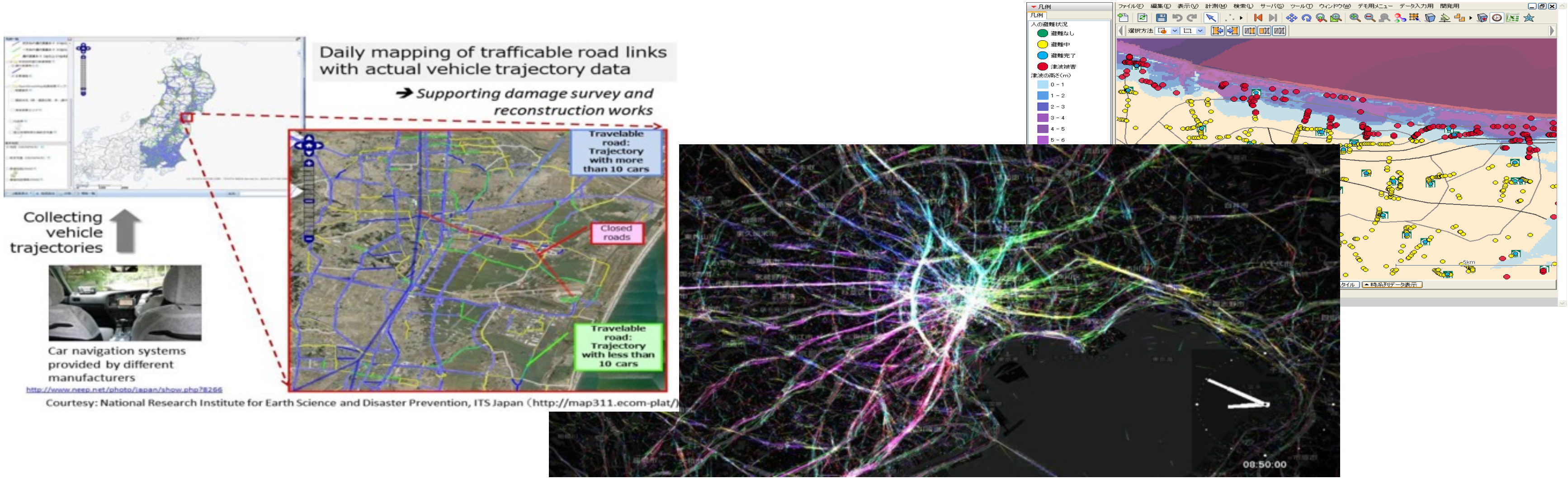


2nd layer: **Sensor space model**

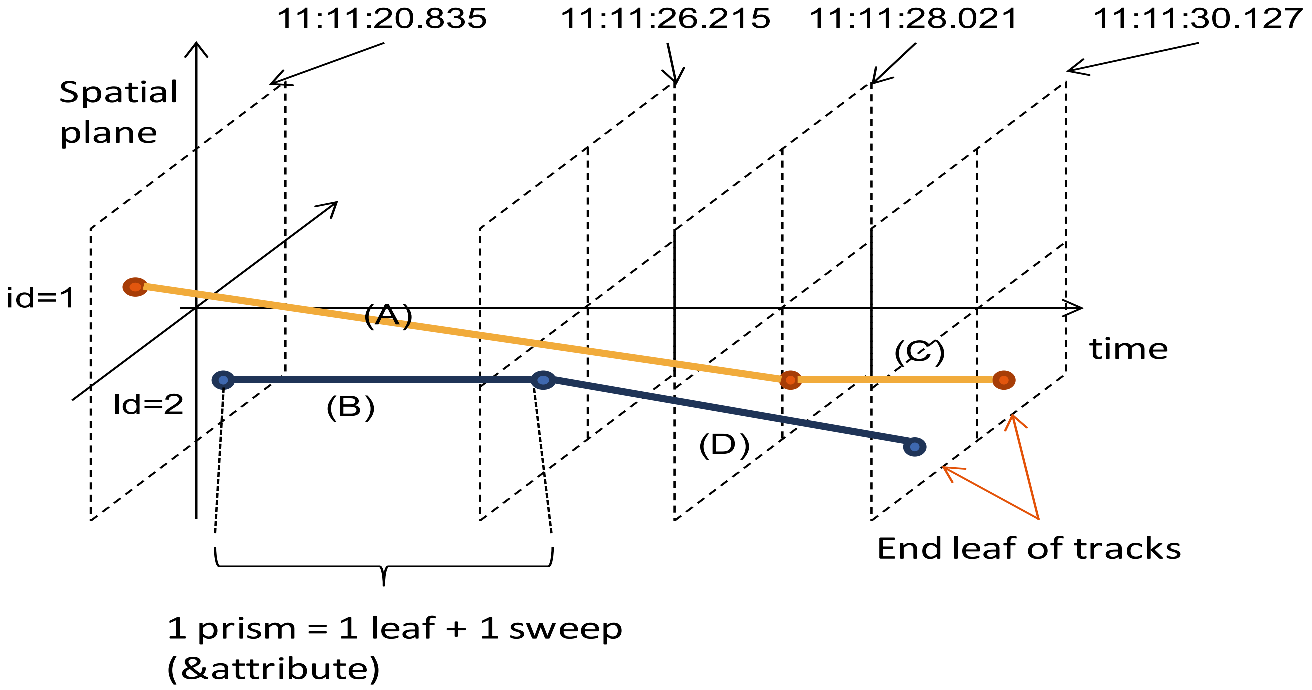
- **Radio/Beacon footprints**
- coverage of sensor areas
- transition between sensor areas

OGC Moving Features

- "Moving features" - vehicles, pedestrians, airplanes, ships.
  - This is Big Data – high volume, high velocity.
- CSV and XML encodings



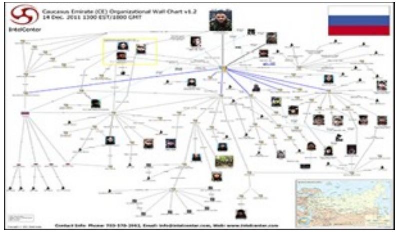
Spatial Temporal Geometry



OGC Moving Features Standard implements ISO 19141

Social Media in Geospatial Analysis

Human-oriented Clients



RDF

JSON-LD

Web Access Layer

OGC Interfaces for Social Media

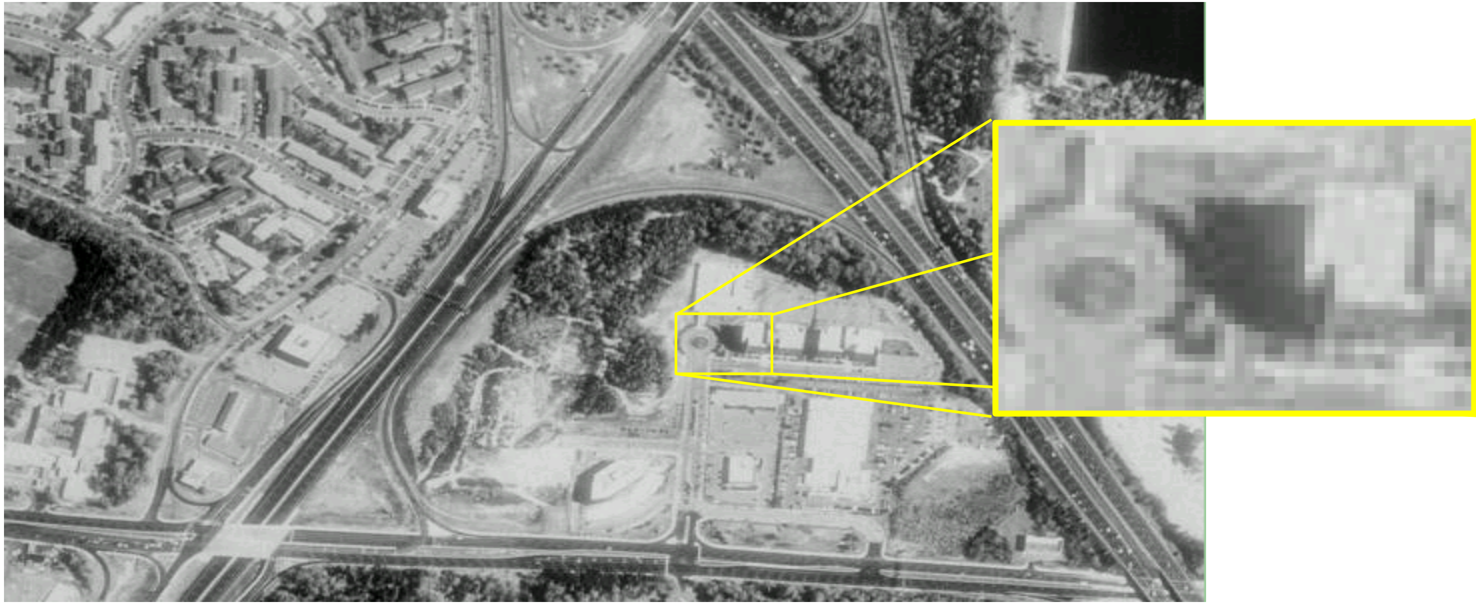
GeoSPARQL	Linked Data REST API	Social Media Analysis WPS

Social Media APIs
Silos

YouTube · flickr · Instagram · twitter · RSS · facebook · …

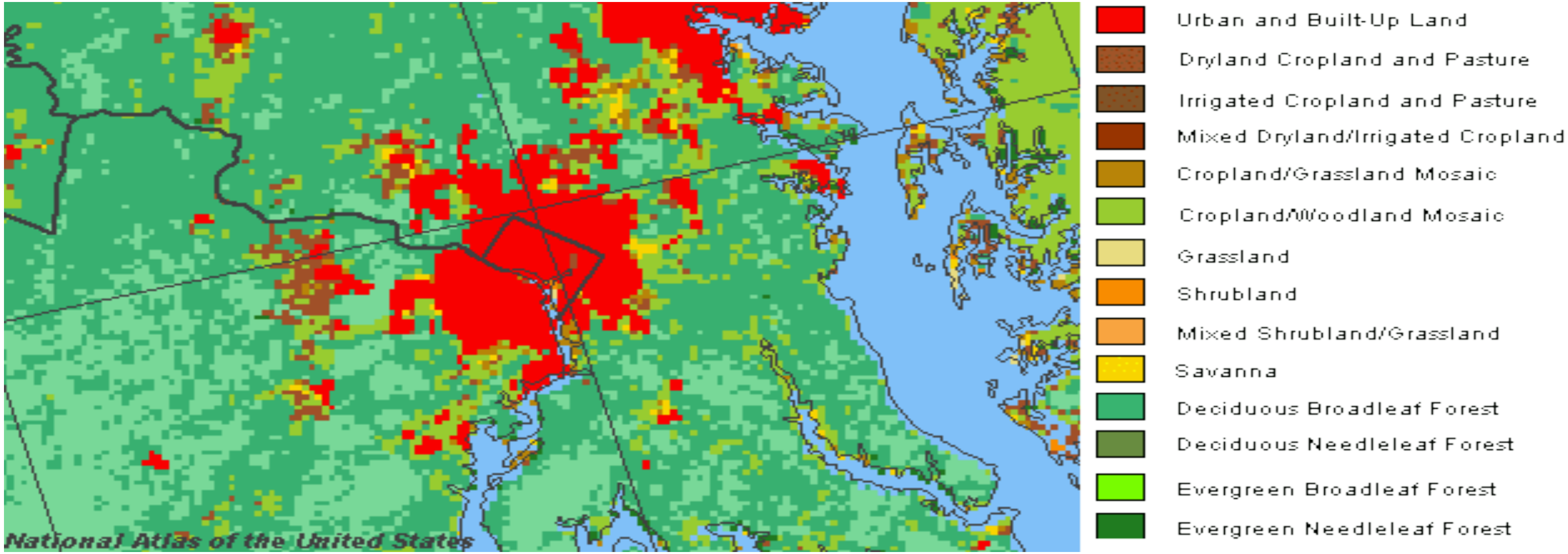
Geospatial Coverages

- Pixel grid (e.g., visible brightness)



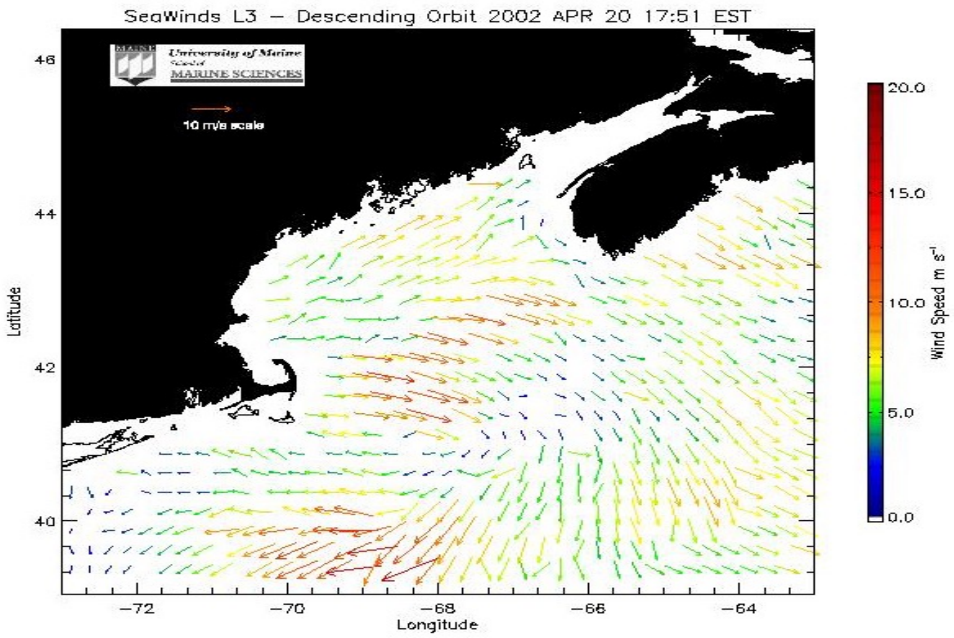
Geospatial Coverages

- Pixel grid (land use / land cover)



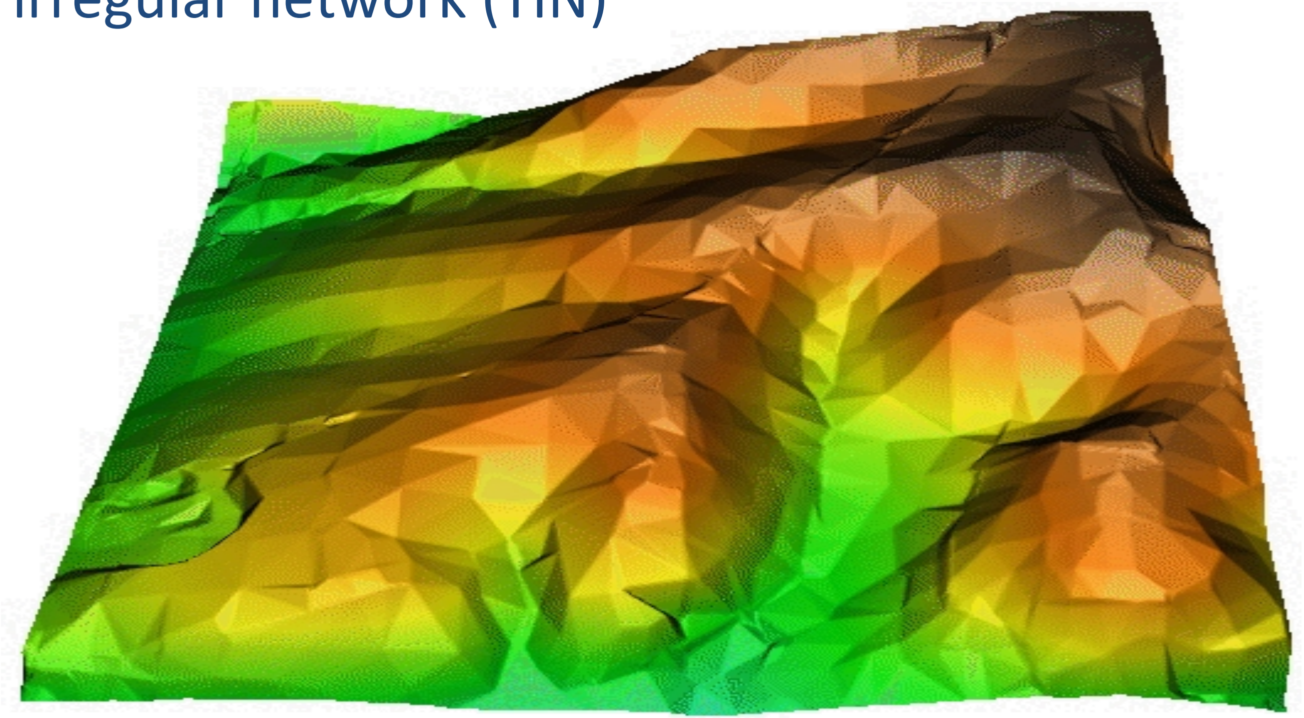
Geospatial Coverages

- Point grid (e.g., wind speed & direction)



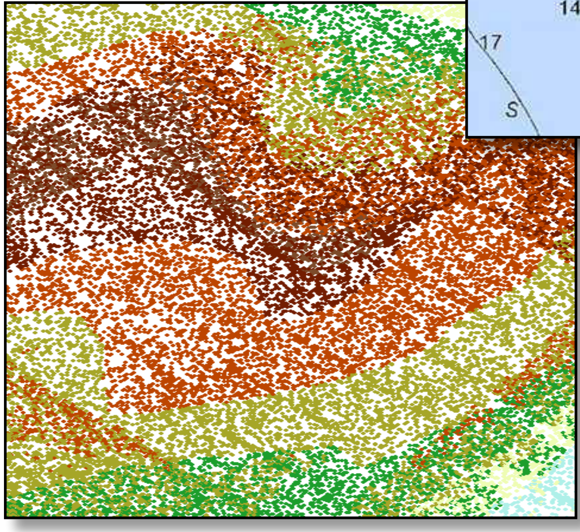
Geospatial Coverages

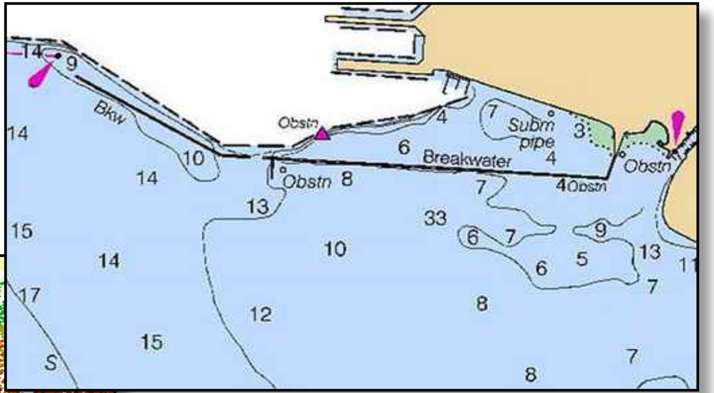
- Triangulated irregular network (TIN)

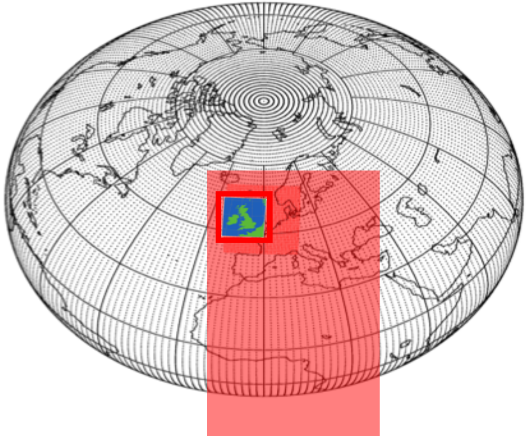


OGC Point Clouds

- WG established in 2015
- Focus on all types of point clouds: LiDAR/laser, bathymetric, meteorologic, photogrammetric…





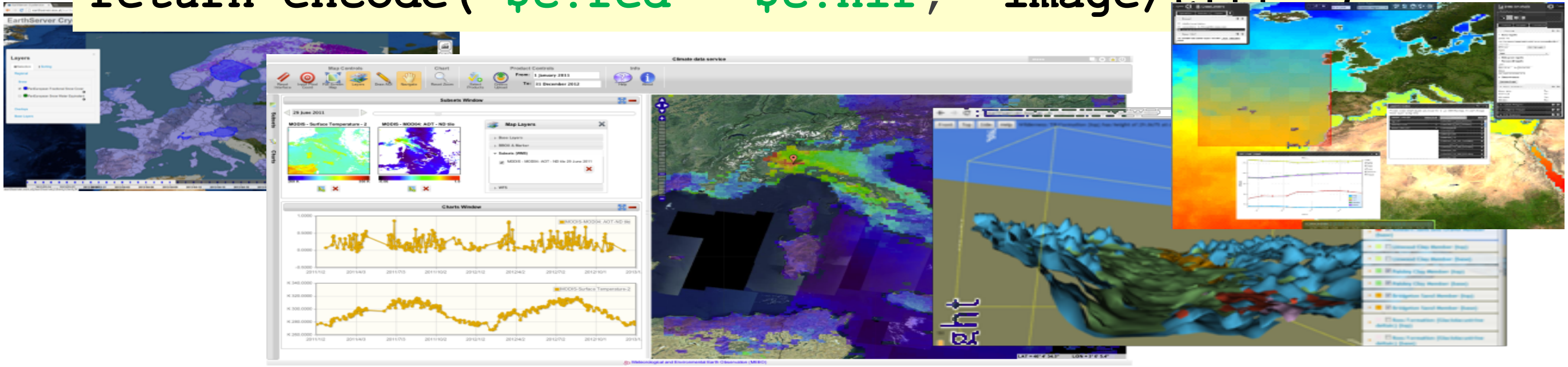


Web Coverage Processing Service

- Query Language for nD sensor, image, simulation, statistics data
  - Syntax close to XQuery (WCPS 2.0: integration)
- Ex: "*From MODIS scenes M1, M2, and M3, the difference between red and nir, as TIFF where nir exceeds 127 somewhere*"

```
for $c in ( M1, M2, M3 )
where some( $c.nir > 127 )
return encode( $c.red - $c.nir, “image/tiff” )
```

→ ($tiff_1$, $tiff_2$)

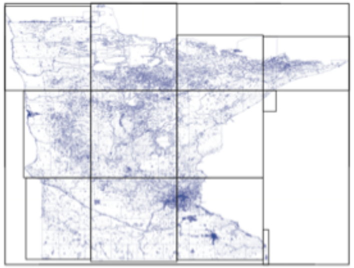


Geospatial Analytics

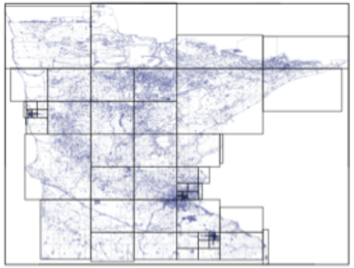
- Analytic exploitation of the space-time features will usher in advances in high-quality prediction systems.
  - Space time features: the highest order bits - Jonas, Tucker

- Using algorithmic extraction and big data graphs to create and relate entities on the Web, organising them through a semantic taxonomy and enabling natural access
  - The future is ‘Where’" - S. Lawler, Bing

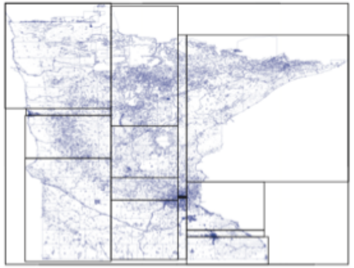
Spatial Partitioning Techniques



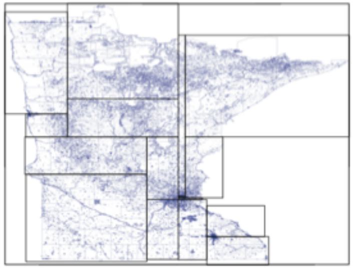
Grid



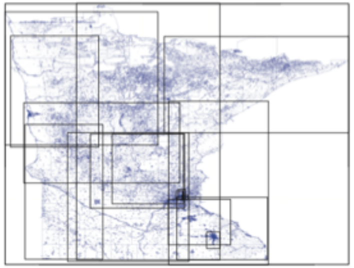
Quad-tree



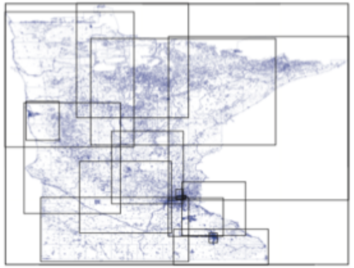
STR/STR+



K-d Tree



Z-Curve



Hilbert Curve

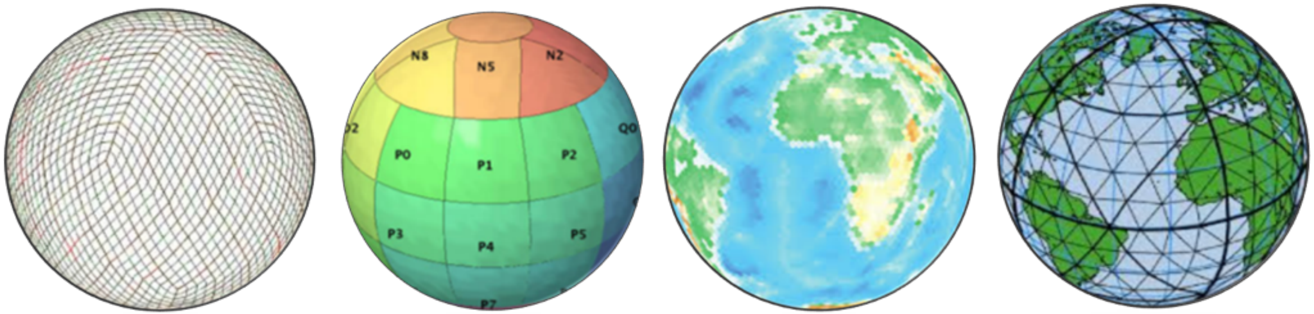
Spatial Indexing

GEOJINNI (formerly SPATIALHADOOP)	Spatial index stored per file on HDFS
GeoTrellis	Z order (2D and 3D), Hilbert (N-Dimensional)
geomesa	Z order (2D and 3D) Binned per week for spatiotemporal
GeoWave	N-Dimensional Hilbert with arbitrary binning and tiered indexing

Discrete Global Grid Systems

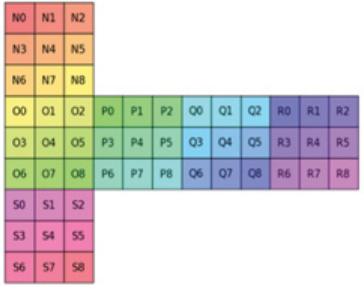
- *"…a spatial reference system that uses a hierarchical tessellation of cells to partition and address the globe. DGGS are characterized by the properties of their cell structure, geo-encoding, quantization strategy and associated mathematical functions."*

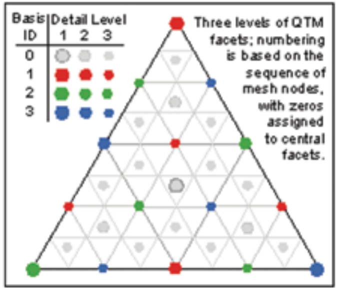
  – *OGC DGGS Candidate Standard*

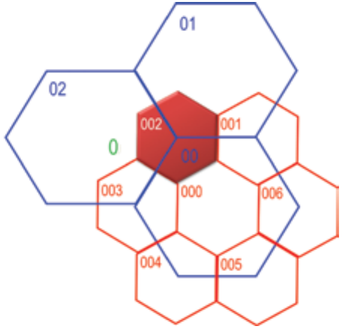


Standardizing Discrete Global Grid Systems

Different Cell Shapes







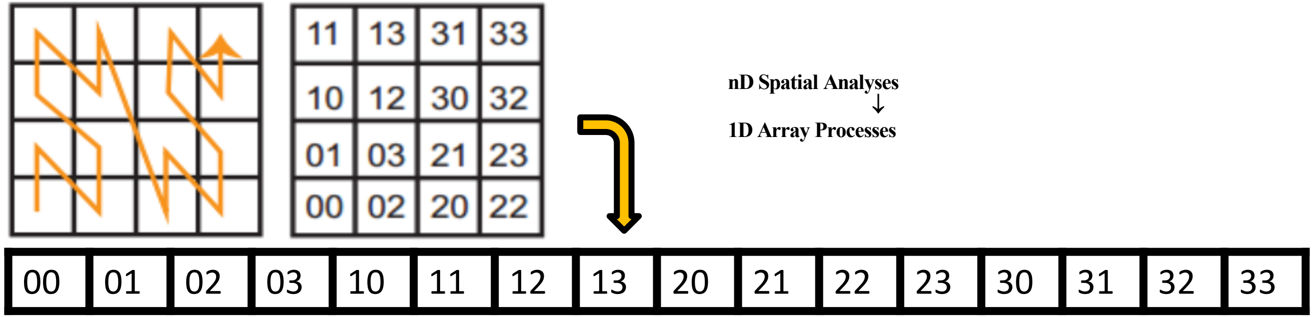
Square = Familiar

Triangular = Fast

Hexagonal = Fineness of Fit

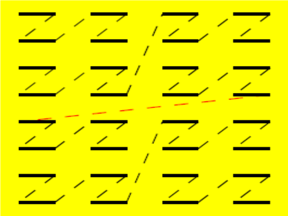
Unique Cell Indices

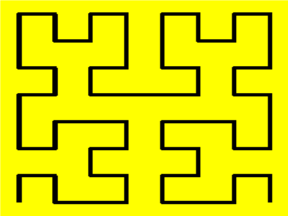
- *Hierarchy-based, Space-filling Curve, Axes-based or Encoded Address*

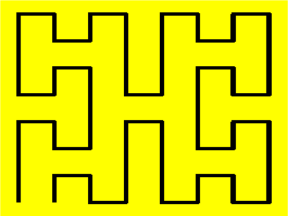


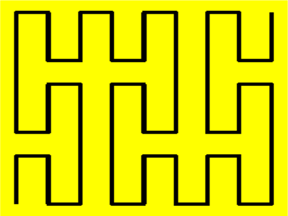
Space Filling Curves

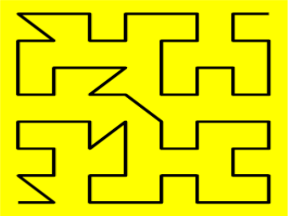
A few different choices…

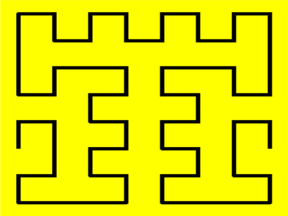

Z-order

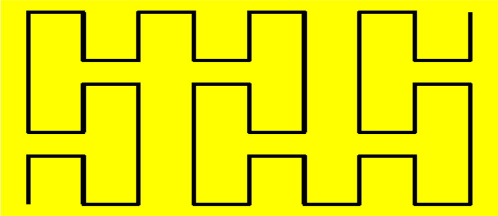

Hilbert curve


H-order


Peano's curve


AR^2W^2-curve

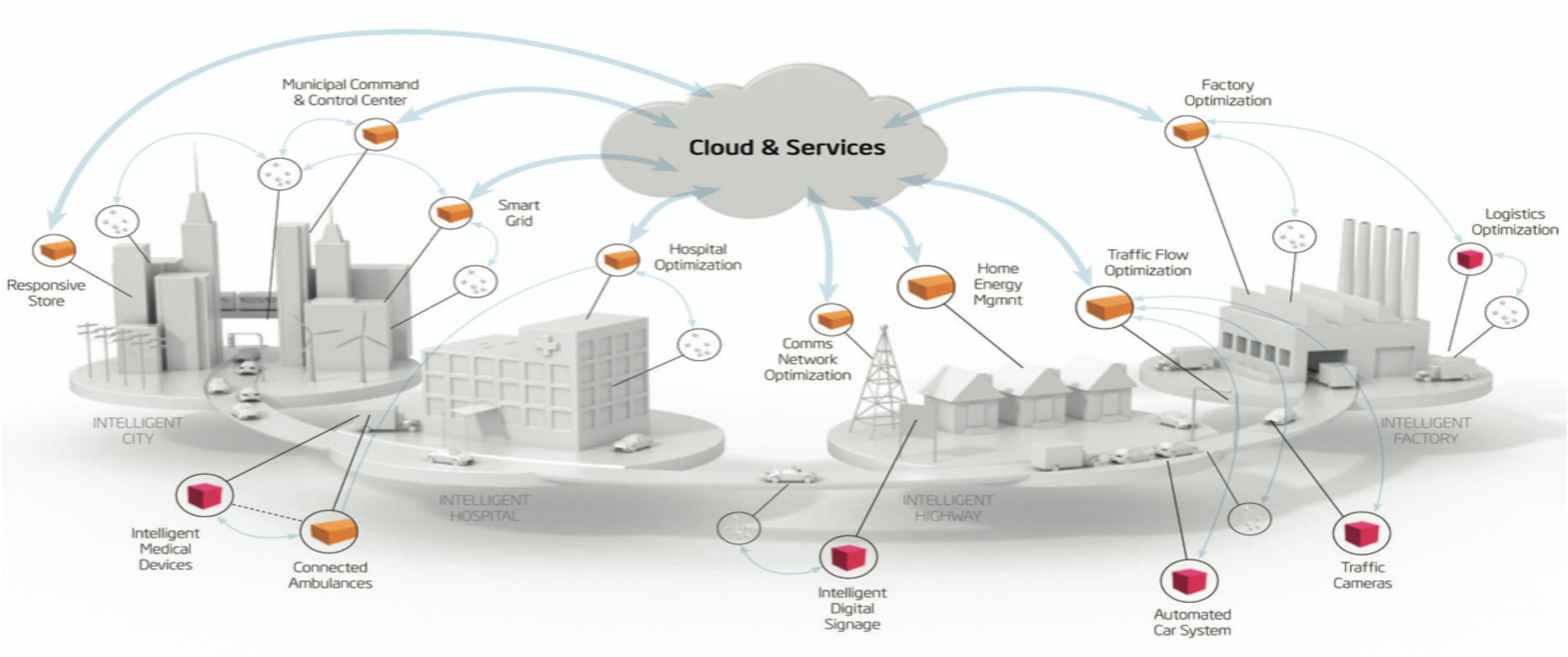

$\beta\Omega$-curve


Balanced Peano

50 billions Internet-connected things by 2020

Sensors Everywhere

(Things or Devices)

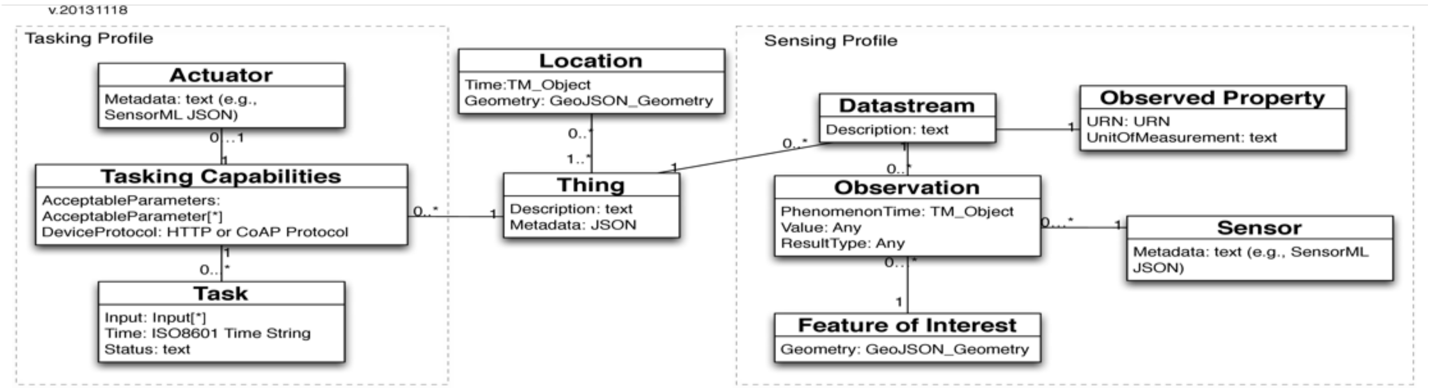


OGC Sensor Web Enablement

- Quickly discover sensors and sensor data (secure or public) that can meet my needs – location, observables, quality, ability to task
- Obtain sensor information in a standard encoding that is understandable by me and my software
- Readily access sensor observations in a common manner, and in a form specific to my needs
- Task sensors, when possible, to meet my specific needs
- Subscribe to and receive alerts when a sensor measures a particular phenomenon

OGC SensorThings for IoT

- Builds on OGC Sensor Web Enablement (SWE) standards that are operational around the world
- Builds on Web protocols; easy-to-use RESTful style
- OGC candidate standard for open access to IoT devices

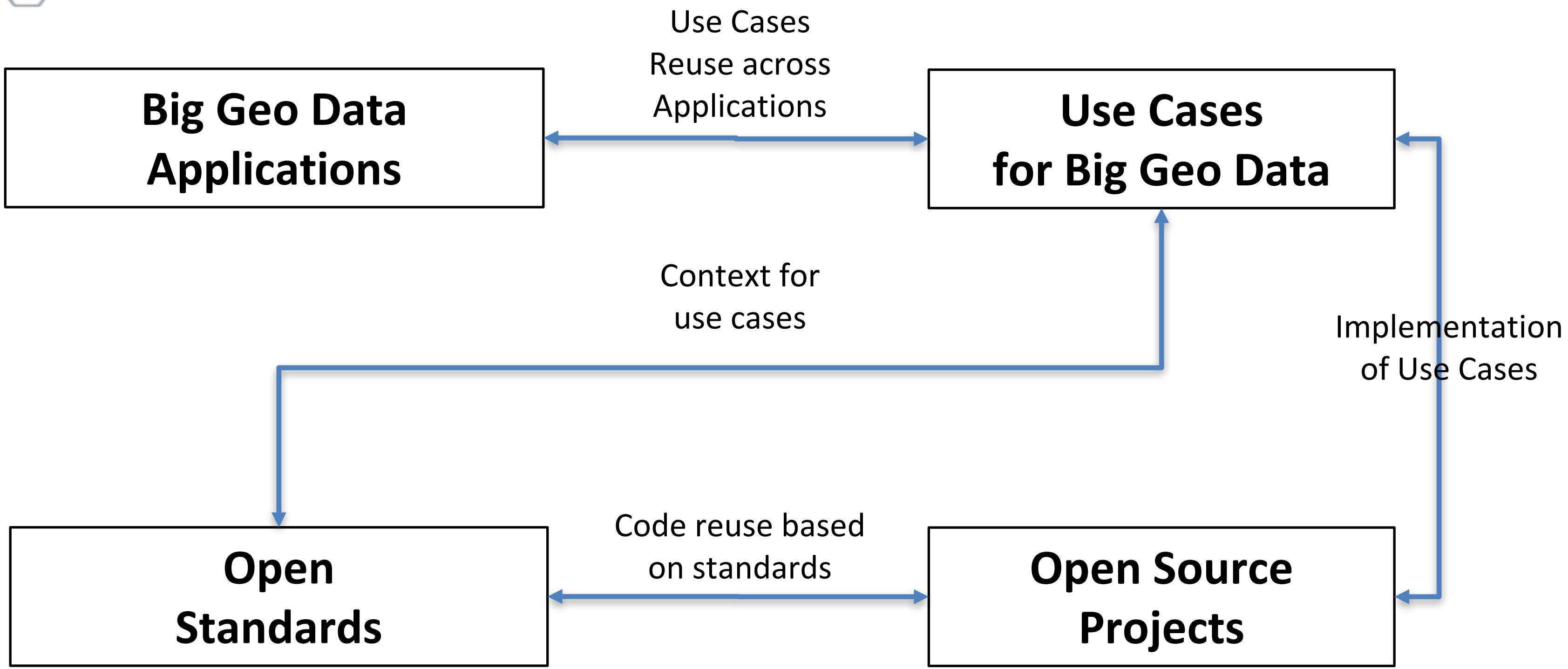


http://ogc-iot.github.io/ogc-iot-api/datamodel.html

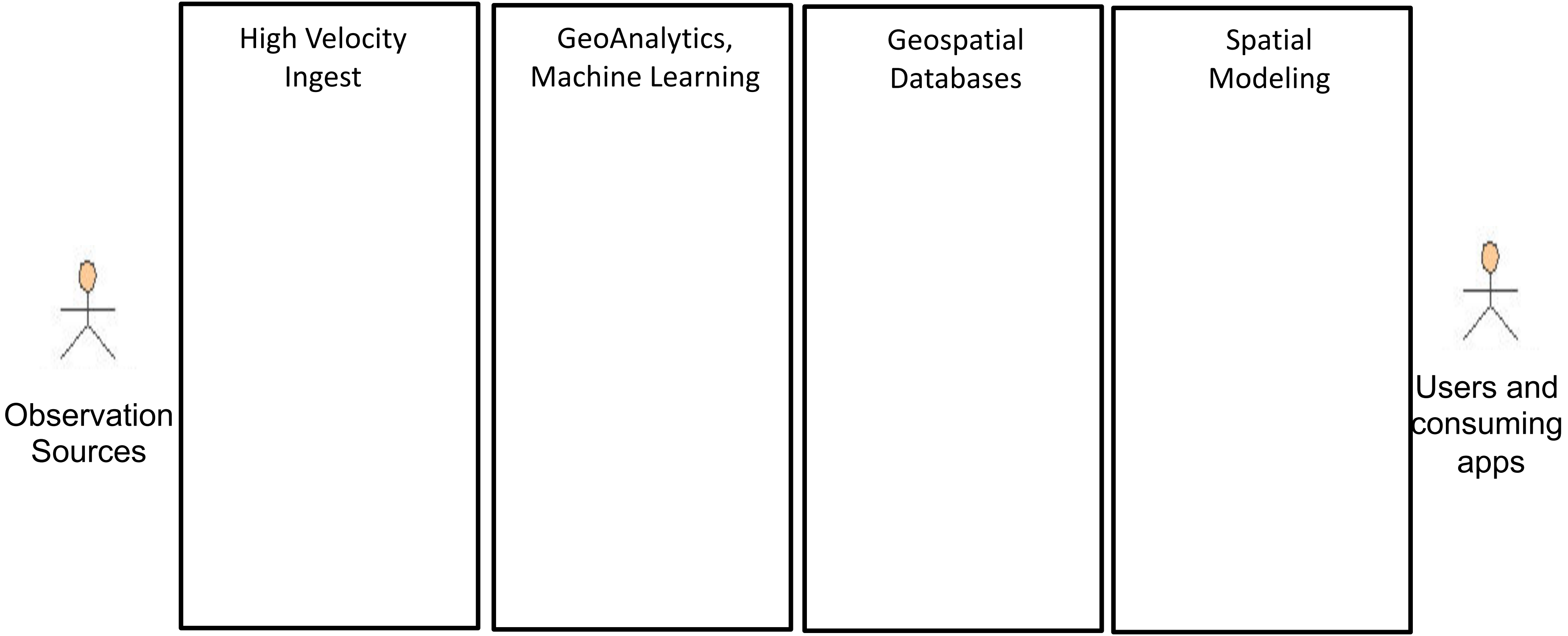
OGC Essentials

- Simple Features for SQL: Fundamental geometries and operations which underlie all OGC standards.
- Well Known Text: Text encoding of Simple Features geometries
- Well Known Binary: binary encoding of Well Known Text.
- CQL/Filter: Common Query Language and Filter language
- GeoPackage: SQLlite for geospatial
- WMTS Simple Tile Matrix

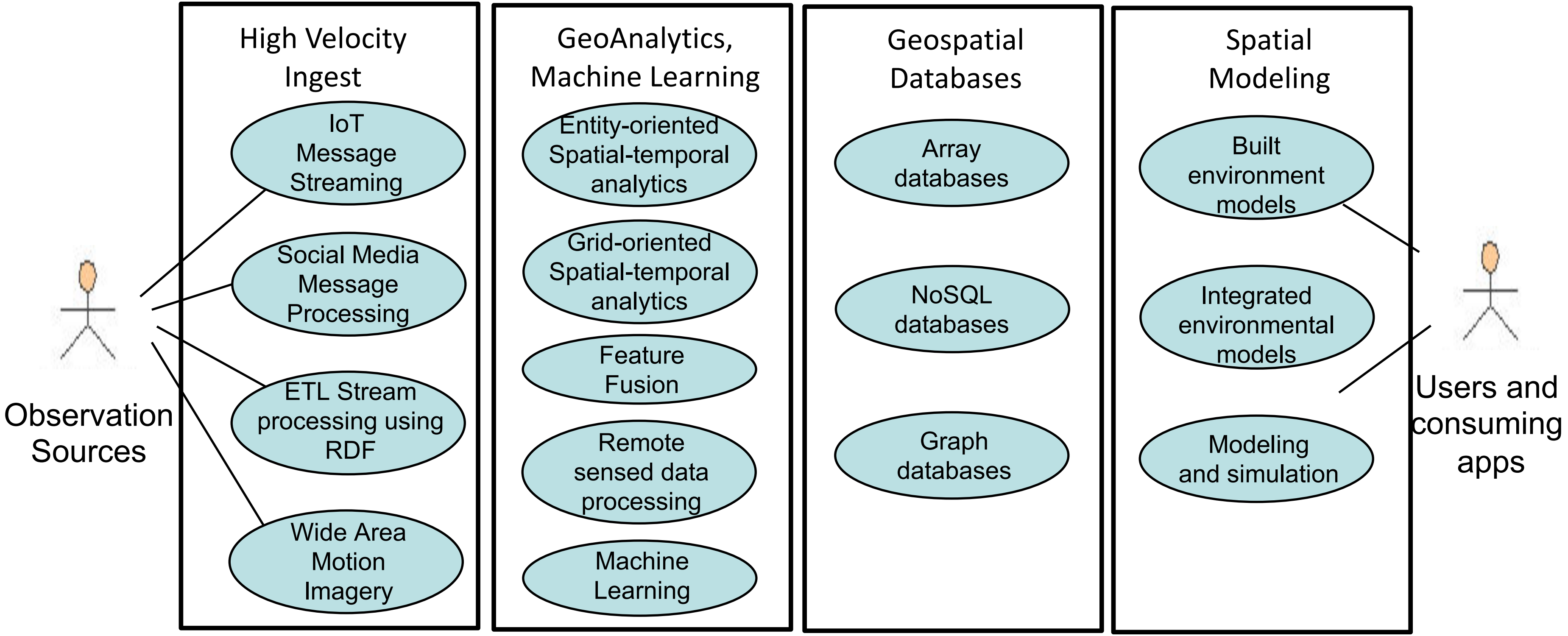
OGC Big Geo Data White Paper



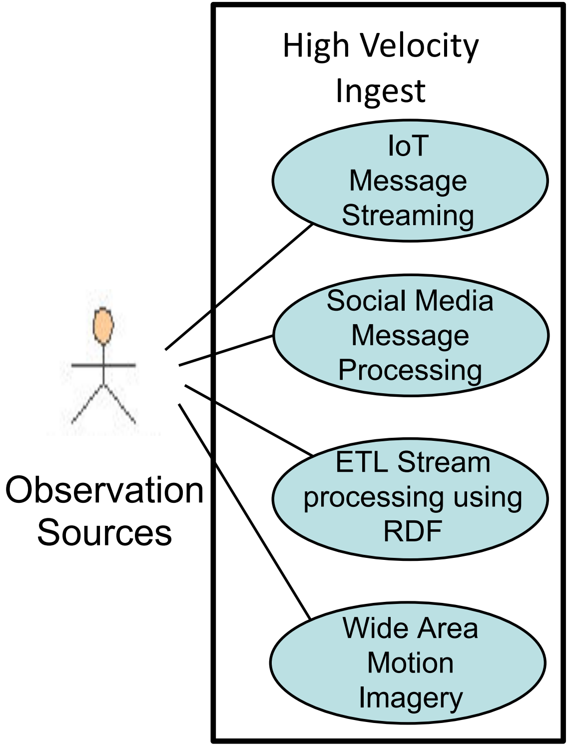
Use Cases for Big Geo Data



Use Cases for Big Geo Data



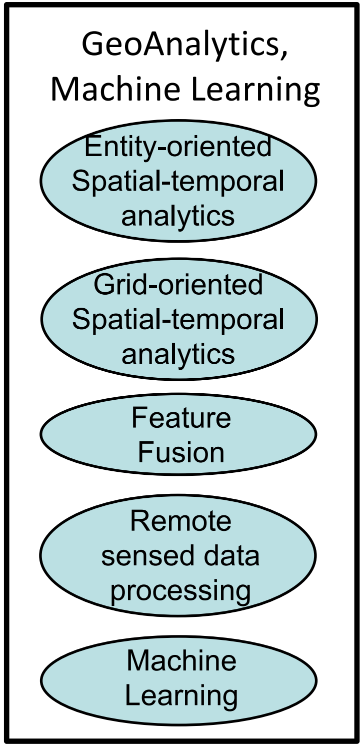
High Velocity Ingest - Use Cases



- Open Source Projects
  - Apache Kafka, Apache NiFi, Apache Jena,
  - SensorHub, SensorUp
- Open Standards
  - IoT: MQTT, COAP, IPSO,
  - OGC Sensor Web Enablement (SWE), SensorThings
  - RDF, OWL, GeoSPARQL,
  - Web Processing Service (WPS)
  - Wide Area Motion Imagery (WAMI)

DRAFT

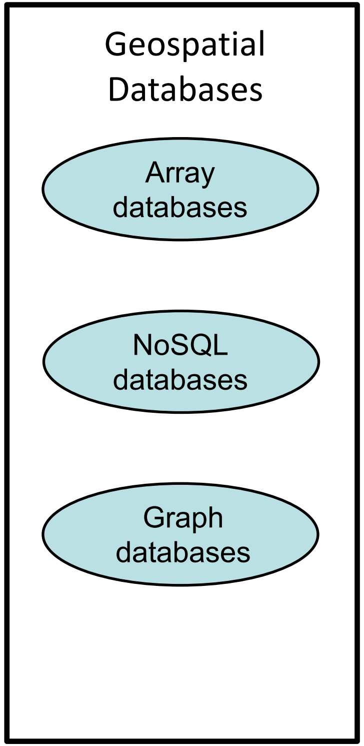
GeoAnalytics, Machine Learning Use Cases



- Open Source Projects
  - Apache: Accumulo, Storm, Lucene, Hadoop, SIS, Magellan, Marmotta, Mahout, Spark
  - LocationTech: GeoWave, GeoTrellis, GeoMesa, GeoJinni, JTS Topology Suite
  - OSGeo: GDAL/OGR, OSSIM, pycsw
  - Others: MrGeo, MonetDB
- Open Standards
  - OGC Simple Features, DGGS
  - GeoTIFF, NetCDF, HDF encodings
  - Web Processing Service (WPS)

DRAFT

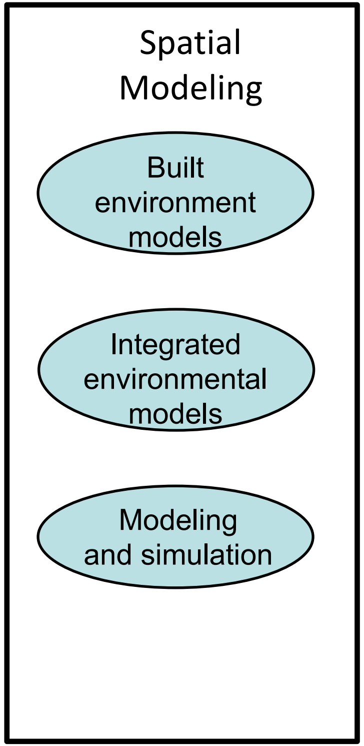
Geospatial Databases Use Cases



- Open Source Projects
  - Apache: Accumulo, Lucene/Solr, Cassandra, SIS, Marmotta
  - OSGeo: degree, GeoServer, OpenLayers, QGIS
  - EarthServer, THREDDS, Raster Storage Archive
  - MonetDB
- Open Standards
  - Web Feature Service (WFS)
  - Web Coverage Service (WCS)
  - Web Map Service (WMS)
  - Geography Markup Language (GML)

DRAFT

Spatial Modeling Use Case



- Open Source Projects
  - Apache SIS
  - CityDB
  - Cesium
- Open Standards
  - CityGML
  - OpenMI
  - OGC CDB

DRAFT

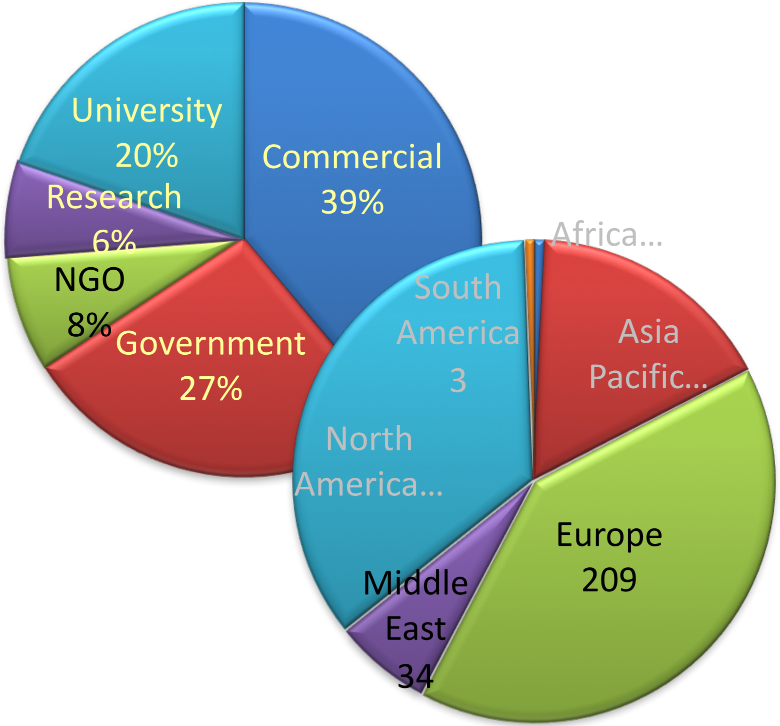
Open Source and Open Standards

- Importance of coordination
  - "Having just one implementation of something is risky" - Tom Hardie, IETF
  - Need to define stable interfaces with stable standard reference
  - Protocols, Interfaces and encodings documented in open standards
- Open Standards use of Open Source
  - Reference Implementations of Open Standards
  - Code snippets in Open Standards.

The Open Geospatial Consortium

Not-for-profit, international voluntary consensus standards organization; leading development of geospatial standards

- Founded in 1994
- 515+ member organizations
- 48 standards
- Thousands of implementations
- Broad user community implementation worldwide
- Alliances and collaborative activities with ISO and many other SDO’s



Apache BD USA May 2016 - Geospatial Track

- Open Geospatial Standards and Open Source
  - George Percivall, Open Geospatial Consortium (OGC)
- Magellan: Spark as a Geospatial Analytics Engine
  - Ram Sriharsha
- Applying Geospatial Analytics Using Apache Spark Running on Apache Mesos
  - Adam Mollenkopf, Esri
- SciSpark: MapReduce in Atmospheric Sciences
  - Kim Whitehall, NASA Jet Propulsion Laboratory
- Geospatially Enable Your Hadoop, Accumulo, and Spark Applications with LocationTech Projects
  - Robert Emanuele, Azavea

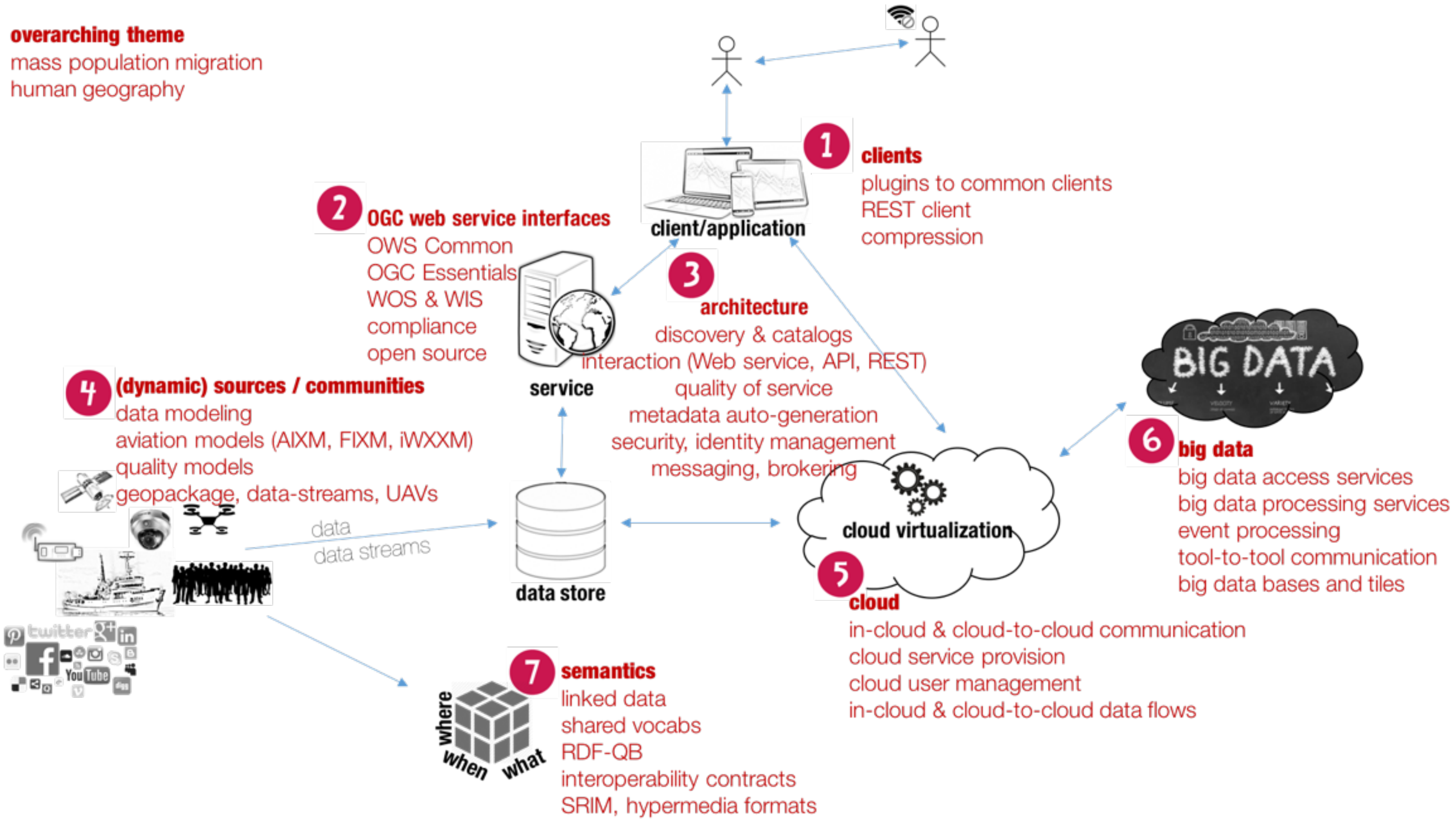
© 2016 Open Geospatial Consortium

Apache BD USA May 2016 - Geospatial Track II

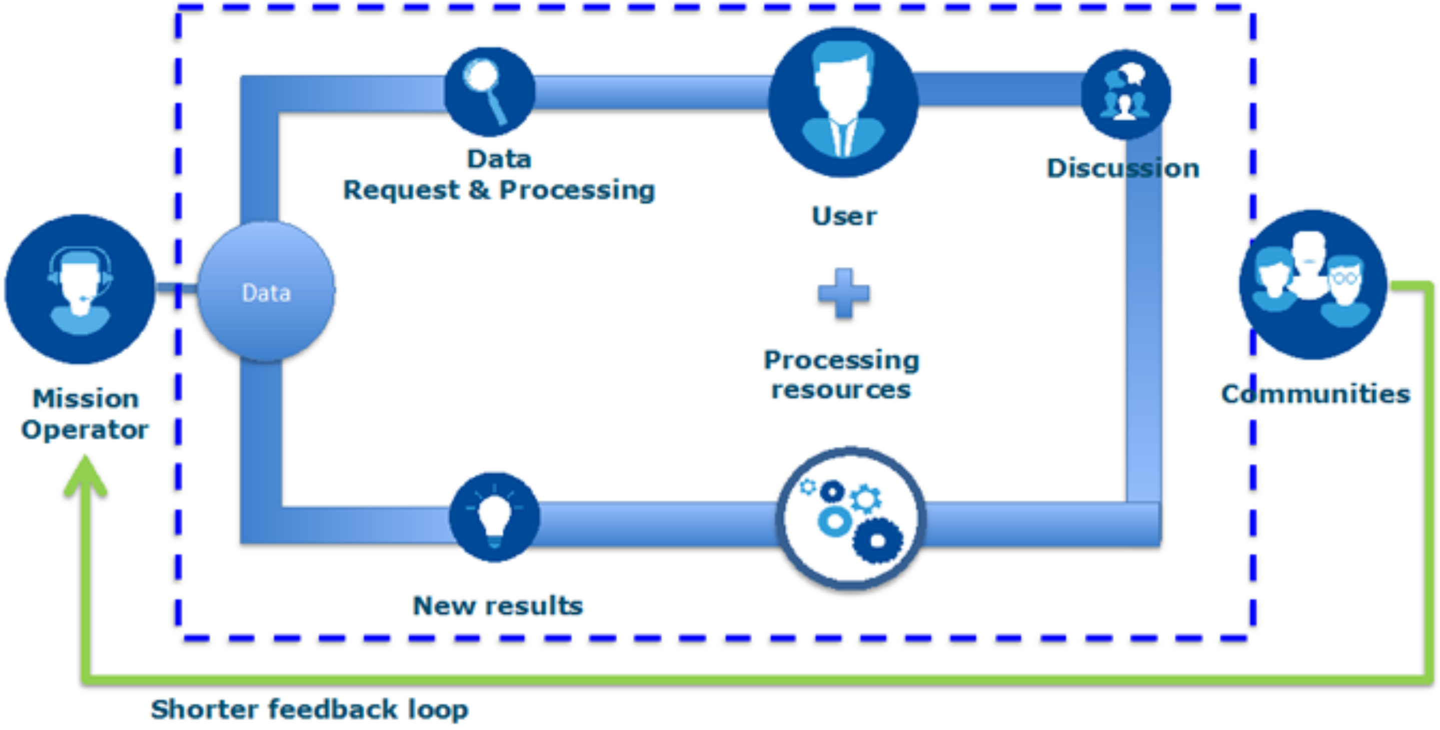
- Hiding Some of Geospatial Complexity
  – Martin Desruisseaux, Geomatys
- Geospatial Querying in Apache Marmotta
  – Sergio Fernandez, Redlink GmbH
- Spatial Data Based People/Vehicles Trails Analysis to Support Precision Urban Planning
  – Yonghua (Henry) Zeng, IBM
- Crowd Learning for Indoor Positioning
  – Thomas Burgess, indoo.rs GmbH

© 2016 Open Geospatial Consortium

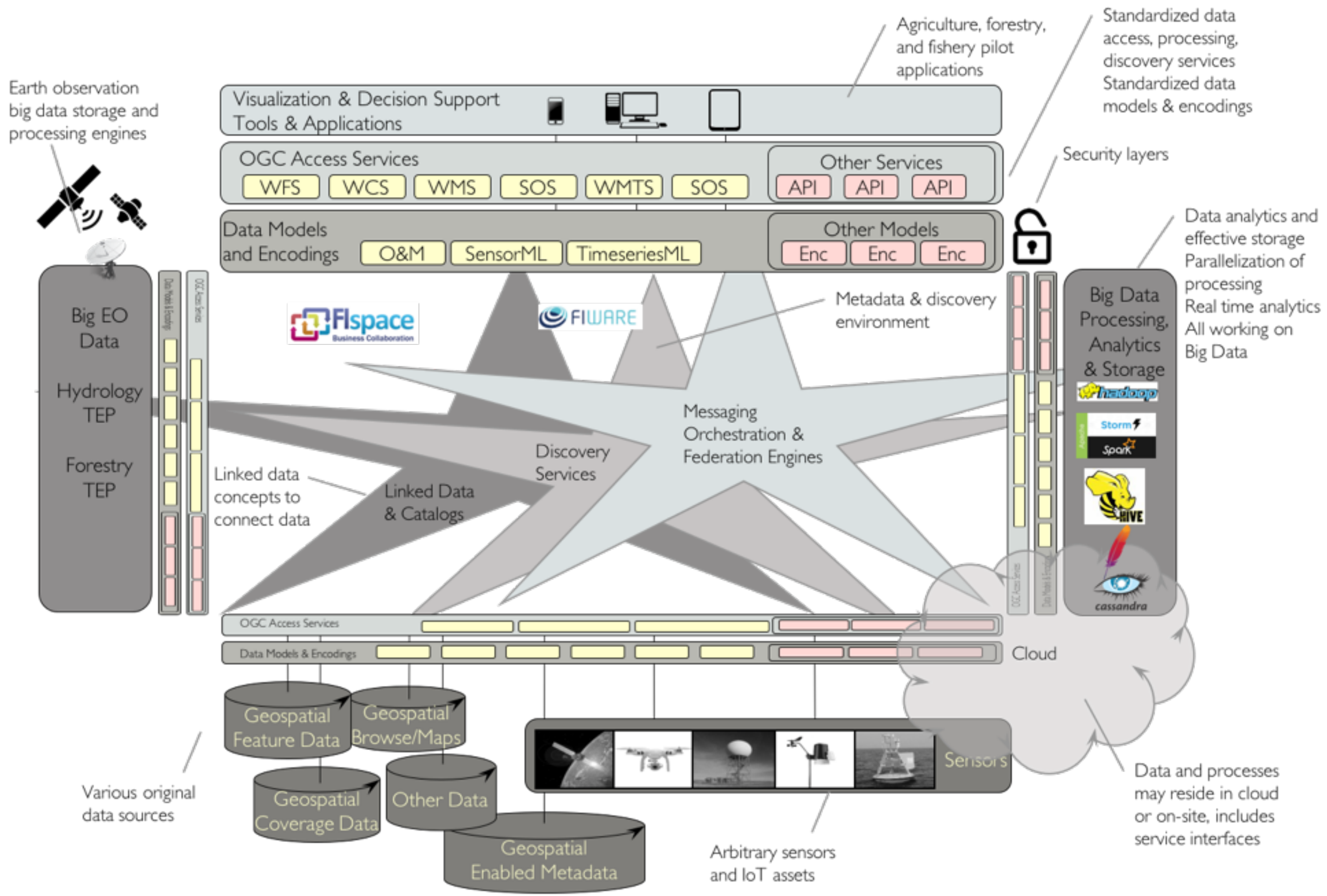
OGC Testbed-13



ESA EO (Exploitation) Platform



Big Data Integration Architectures



The Open Geospatial Consortium

Open Geospatial Consortium

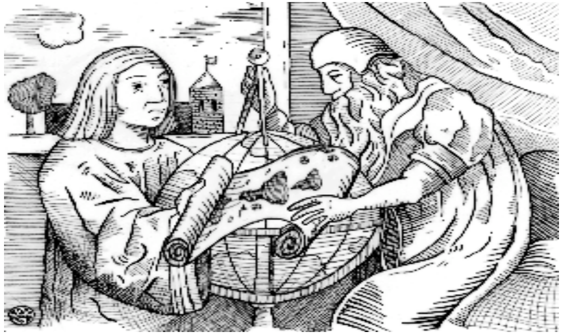
www.opengeospatial.org

OGC Standards - freely available

www.opengeospatial.org/standards

OGC on YouTube

http://www.youtube.com/user/ogcvideo



Dr. Ingo Simonis

isimonis@opengeospatial.org