Geospatial Big Data

Big Data

Daniel Ciucur

Faculty of Mathematics and Informatics

Wednesday, July 13 2023

Table of Contents

Table of Contents

What is Geospatial data?

- Geospatial data is information that describes objects, events or other features with a location on or near the surface of the earth.

- Typically combines
  - **location information** (usually coordinates on the earth)
  - **attribute information** (the characteristics of the object, event or phenomena concerned)
  - **temporal information** (the time or life span at which the location and attributes exist)

Types and examples of Geospatial data

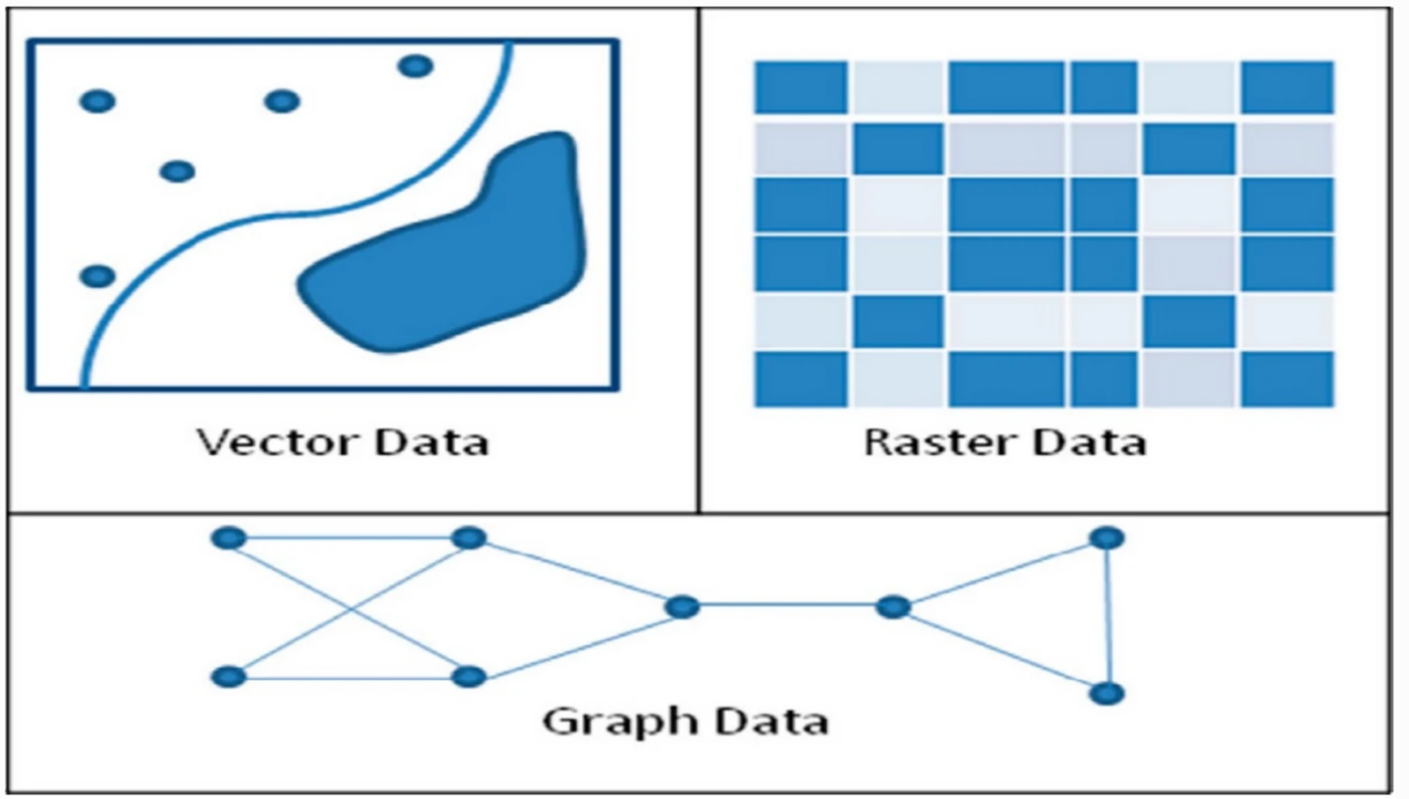


Figure: Types of spatial data

Sources of Geospatial Big Data

- **Earth observations**
  As of 2014, NASA's Earth Observing System Data and Information System (EOSDIS) was managing more than nine petabytes of data, and it is adding about 6.4 terabytes to its archives every day
- **Geoscience model simulations**
- **Internet of Things**
- **Volunteered geographic information**
  VGI efers to the creation and dissemination of geographic information from the public, a process in which citizens are regarded as sensors moving "freely" over the surface of the Earth

Table of Contents

Spatial Hadoop

- full-fledged MapReduce framework with native support for spatial data.
- a comprehensive extension to Hadoop that pushes spatial data inside the core functionality of Hadoop.
- SpatialHadoop runs existing Hadoop programs as is, yet, it achieves order(s) of magnitude better performance than Hadoop when dealing with spatial data.

Spatial Hadoop Arhitecture

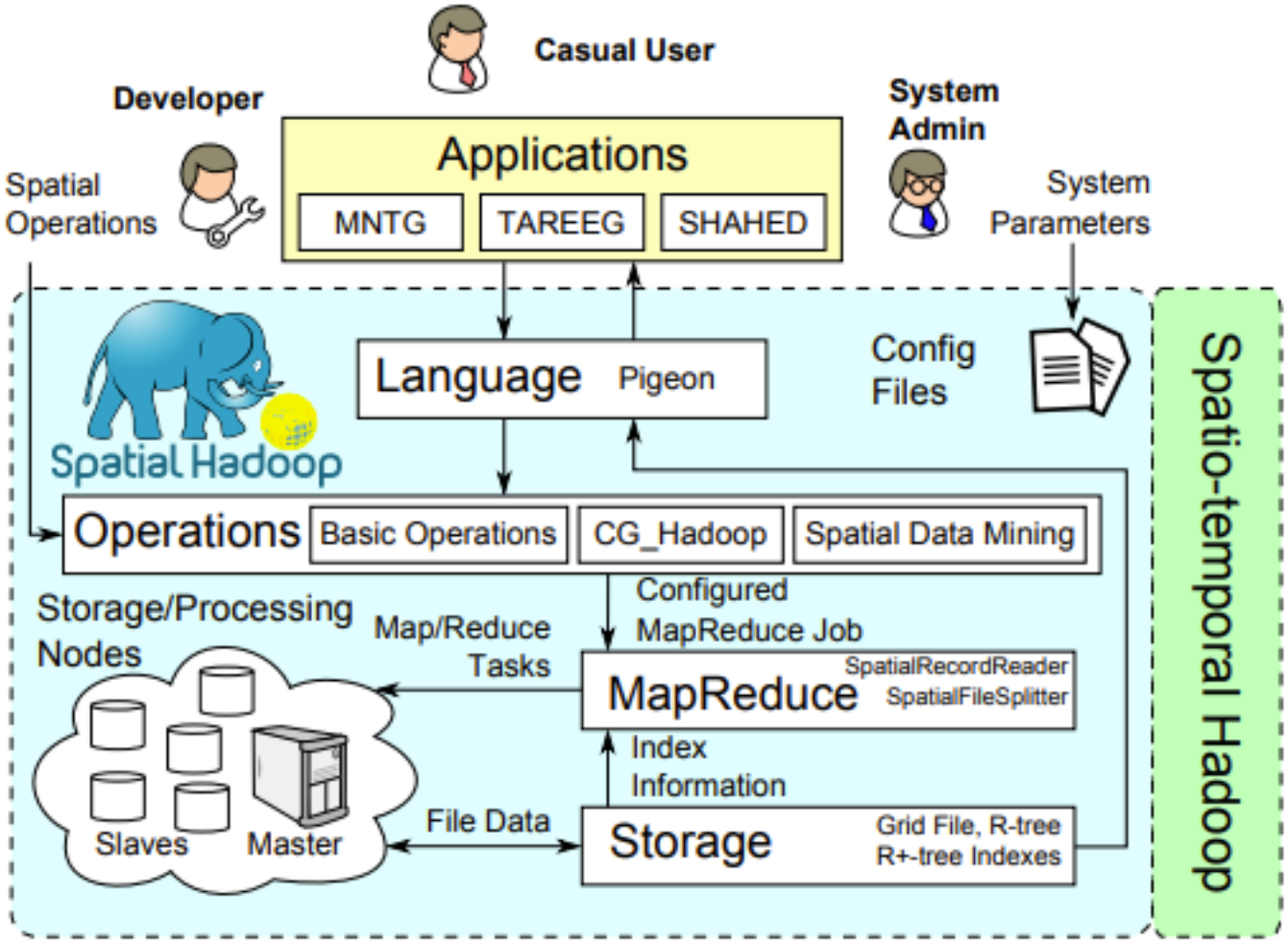


Figure: Spatial Hadoop Arhitecture

Comparrison between Hadoop and SpatialHadoop

```
Objects = LOAD 'points' AS (id:int, x:int, y:int);
Result  = FILTER Objects BY x < x2 AND x > x1
                         AND y < y2 AND y > y1;
```

(a) Range query in Hadoop

```
Objects = LOAD 'points' AS (id:int, Location:POINT);
Result  = FILTER Objects BY
            Overlaps (Location, Rectangle(x1, y1, x2, y2));
```

(b) Range query in SpatialHadoop

Figure: Range query in Hadoop vs Spatial Hadoop

Comparrison between Hadoop and SpatialHadoop

According to one of the pappers I studied for this this report ("A Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial Data") the query in Figure 2.2 was run for 70M spatial objects on a 20 node cluster.

For Hadoop the execution of the query took 200 seconds, while Spatial Hadoop took 2 seconds for the same query.

Table of Contents

Movebank

Movebank is a free, online database of animal tracking data hosted by the Max Planck Institute of Animal Behavior. We help animal tracking researchers to manage, share, protect, analyze and archive their data.

Movebank cyberinfrastructure

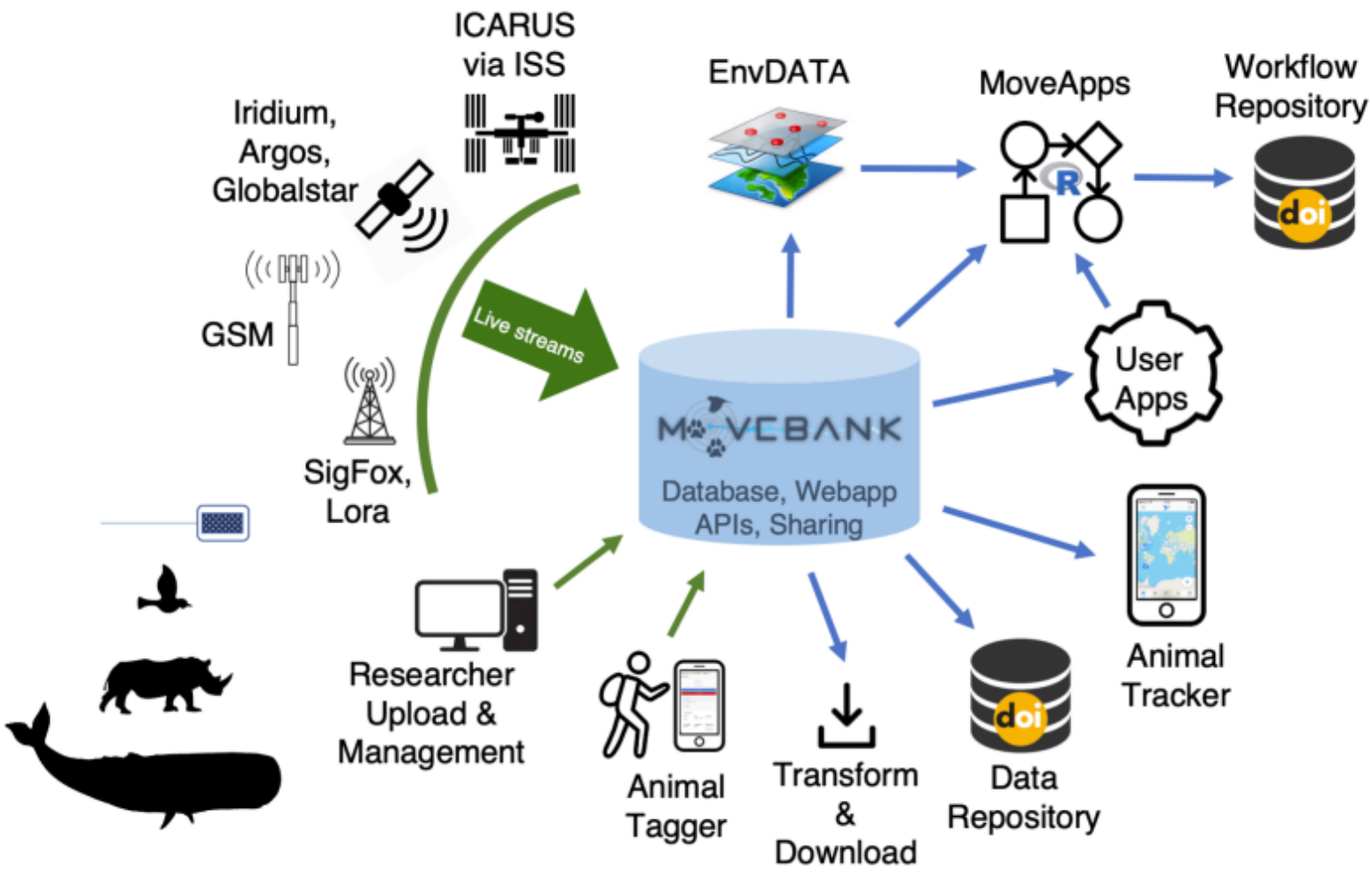
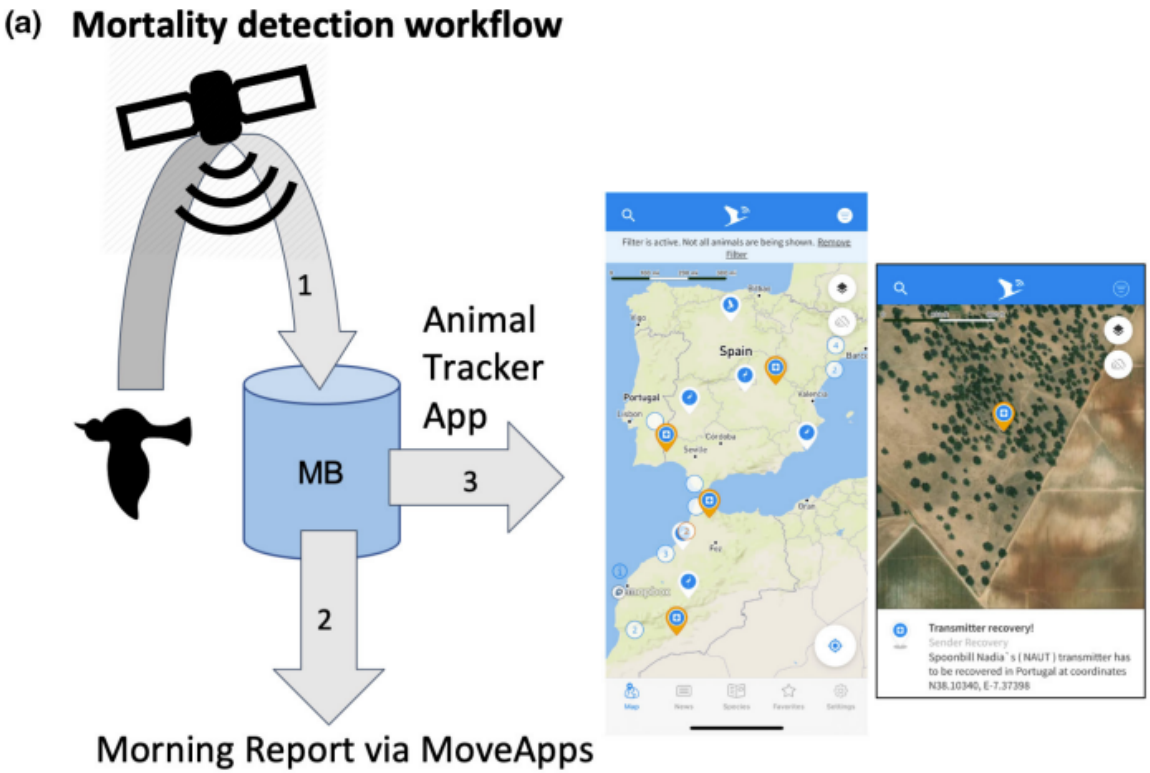
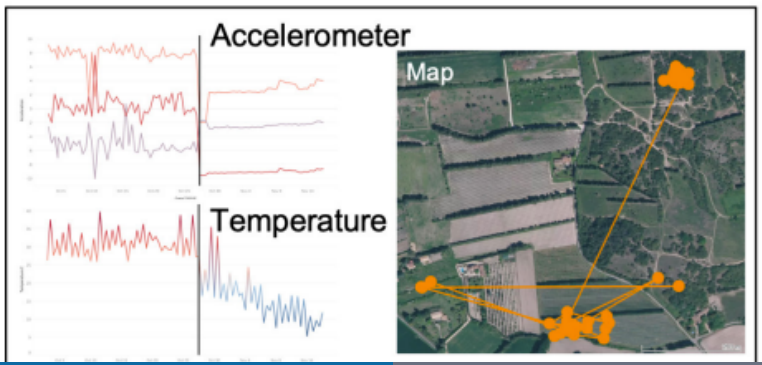


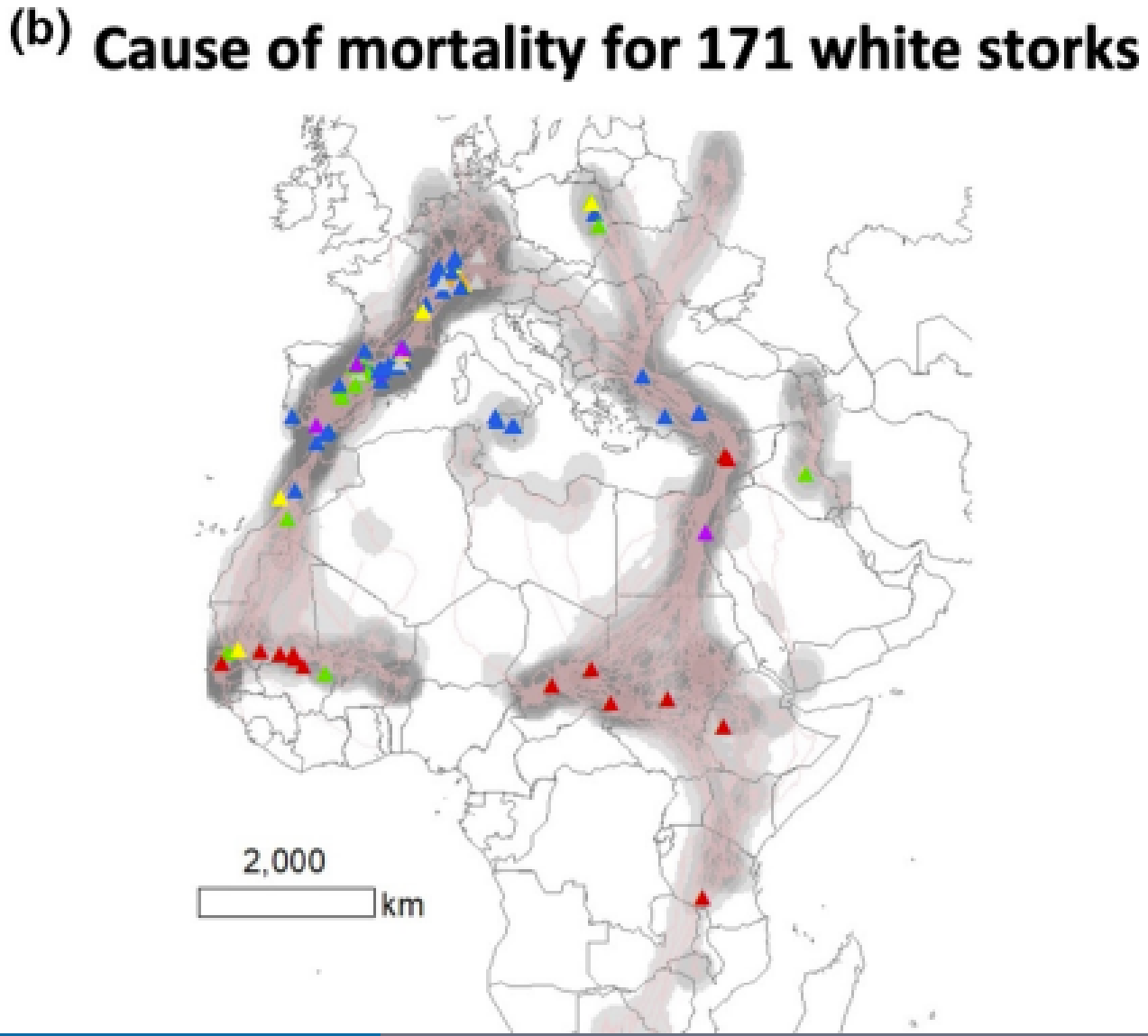
Figure: The Movebank cyberinfrastructure ecosystem of tools to acquire, manage and analyse animal tracking data

Study case: Cause of mortality for 171 white storks





Study case: Cause of mortality for 171 white storks



Study case: Cause of mortality for 171 white storks

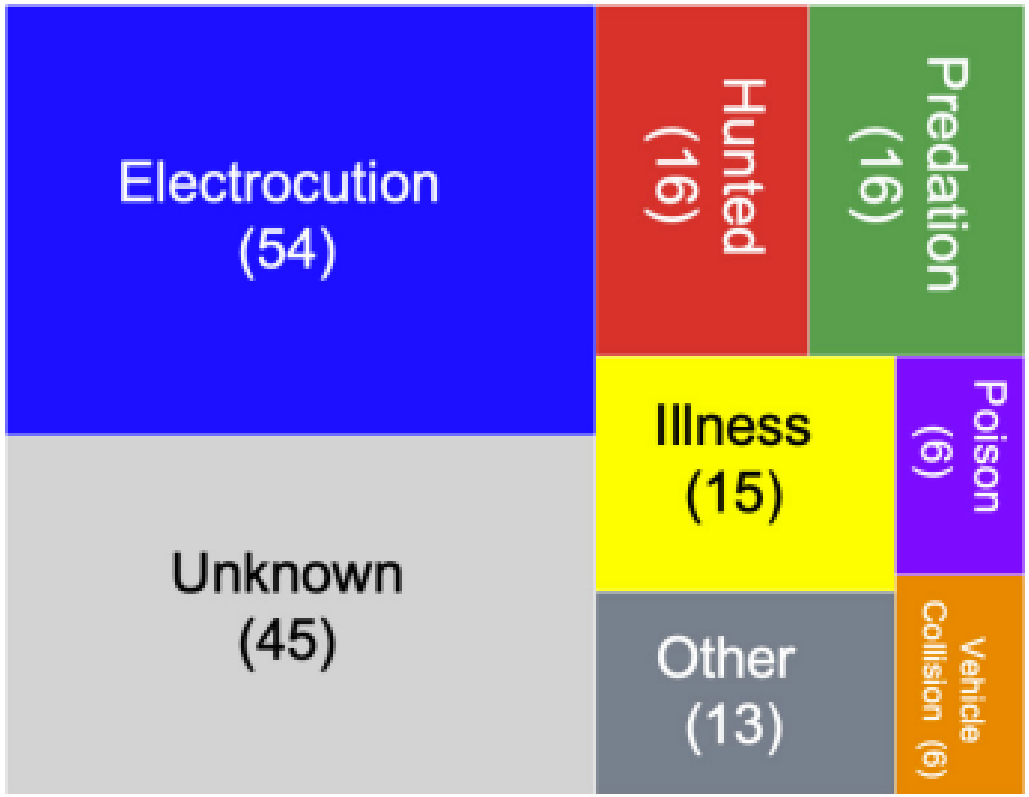


Figure: Storks mortality causes

PostGIS

PostGIS extends the capabilities of the PostgreSQL relational database by adding support storing, indexing and querying geographic data.

PostGIS Features

- **Spatial Data Storage**: Store different types of spatial data such as points, lines, polygons, and multi-geometries, in both 2D and 3D data.
- **Spatial Indexing**: Quickly search and retrieve spatial data based on its location.
- **Spatial Functions**: A wide range of spatial functions that allow you to filter and analyze spatial data, measuring distances and areas, intersecting geometries, buffering, and more.
- **Raster Data Support**: Storage and processing of raster data, such as elevation data and weather data.
- **Integration**: Access and work with PostGIS using third party tools such as QGIS, GeoServer, ArcGIS, Tableau, and MapServer.

PostGIS Example

- nyc_census_blocks
  - blkid, popn_total, boroname, geom
- nyc_streets
  - name, type, geom
- nyc_subway_stations
  - name, geom
- nyc_neighborhoods
  - name, boroname, geom

What subway station is in 'Little Italy'? What subway route is it on?

```
SELECT s.name, s.routes
FROM nyc_subway_stations AS s
JOIN nyc_neighborhoods AS n
ON ST_Contains(n.geom, s.geom)
WHERE n.name = 'Little Italy';
```

```
   name     | routes
------------+--------
 Spring St  | 6
```

What is the GeoJSON representation of the 'Broad St' subway station?

```
SELECT
 ST_AsGeoJSON(geom)
FROM nyc_subway_stations
WHERE name = 'Broad St';
```

```
{"type":"Point",
 "crs":{"type":"name","properties":{"name":"EPSG:26918"}},
 "coordinates":[583571.905921312,4506714.341192182]}
```

Table of Contents

Conclusion

In my report I managed study the topic of Geospatial Big Data.

I was interested to know what types of geospatial data we have, how we collect this data and what challenges we have. After that I studied tools/technologies which were created to help us deal with geospatial data, for this I analyzed SpatialHadoop and Beast. Lastly, I researched an open-source project which consists of an ecosystem of tools used by thousands of researchers to collect, manage, share, visualize, analyse and archive their animal tracking and other animal-borne sensor data.

References I

[1] Jae-Gil Lee and Kang Minseo. "Geospatial Big Data: Challenges and Opportunities". In: *Big Data Research* 2 (Feb. 2015). DOI: 10.1016/j.bdr.2015.01.003.

[2] Ahmed Eldawy and Mohamed F Mokbel. "A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data". In: *Proceedings of the VLDB Endowment* 6.12 (2013).

[3] Harshawardhan S Bhosale and Devendra P Gadekar. "A review paper on big data and hadoop". In: *International Journal of Scientific and Research Publications* 4.10 (2014).

[4] Ahmed Eldawy et al. "Beast: Scalable exploratory analytics on spatio-temporal data". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021.

References II

[5] Antonin Guttman. "R-trees: A dynamic index structure for spatial searching". In: *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*. 1984.

[6] Ahmed Eldawy and Mohamed F Mokbel. "The era of big spatial data". In: *2015 31st IEEE International Conference on Data Engineering Workshops*. IEEE. 2015.

[7] Yassine Loukili, Younes Lakhrissi, and Safae Elhaj Ben Ali. "Geospatial Big Data Platforms: A Comprehensive Review". In: *KN-Journal of Cartography and Geographic Information* 72.4 (2022).

[8] Zhenlong Li. "Geospatial big data handling with high performance computing: Current approaches and future directions". In: *High Performance Computing for Geospatial Applications* (2020).

References III

[9] Roland Kays et al. “The Movebank system for studying global animal movement and demography”. In: *Methods in Ecology and Evolution* 13.2 (2022).

Questions?