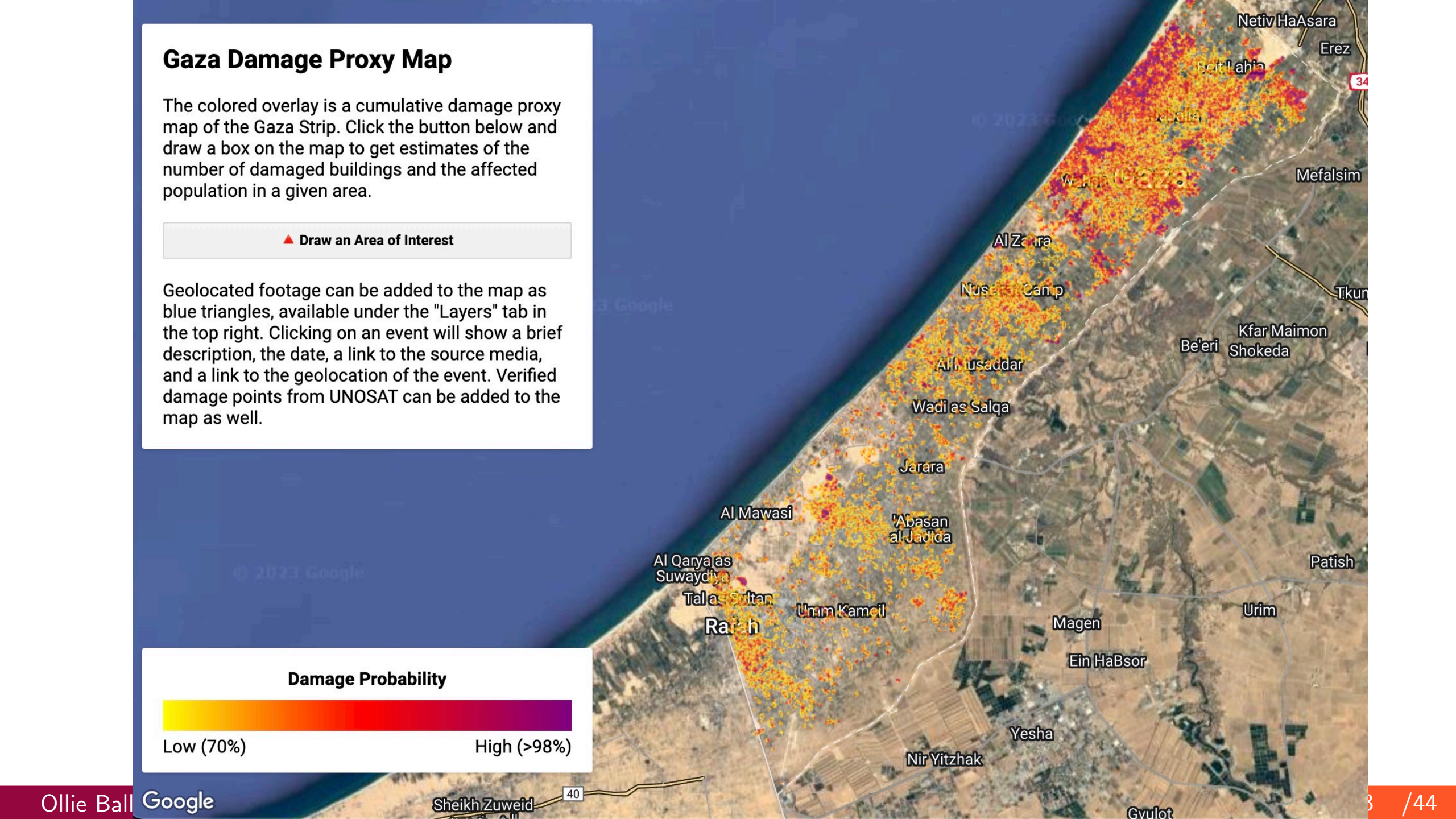
Introduction

CASA0025:
Building Spatial Applications with Big Data

Ollie Ballinger

Outline

1. Spatial Applications
2. Handling big data by type
    1. Vector data
    2. Raster data
3. Module Overview


Gaza Damage Proxy Map
The colored overlay is a cumulative damage proxy map of the Gaza Strip. Click the button below and draw a box on the map to get estimates of the number of damaged buildings and the affected population in a given area.
Draw an Area of Interest
Geolocated footage can be added to the map as blue triangles, available under the "Layers" tab in the top right. Clicking on an event will show a brief description, the date, a link to the source media, and a link to the geolocation of the event. Verified damage points from UNOSAT can be added to the map as well.
Damage Probability
Low (70%)
High (>98%)
Netiv HaAsara
Erez
Mefalsim
Kfar Maimon
Shokeda
Be'eri
Wadi as Salqa
Jarara
Al Mawasi
Al Qarya as Suwaydiya
Tal as Sultan
Rafah
Umm Kameil
Magen
Urim
Patish
Ein HaBsor
Yesha
Nir Yitzhak
Sheikh Zuweid
Google


Ballinger: ‘This is truly sort of irreversible damage that has been caused’ in Gaza
Breaking News ▪ Israel-Hamas War
CBCNN LIVE

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Open access | Published: 03 January 2024

Satellite mapping reveals extensive industrial activity at sea

Abstract

The world's population increasingly relies on the ocean for food, energy production and global trade[1,2,3], yet human activities at sea are not well quantified[4,5]. We combine satellite imagery, vessel GPS data and deep-learning models to map industrial vessel activities and offshore energy infrastructure across the world's coastal waters from 2017 to 2021. We find that 72–76% of the world's industrial fishing vessels are not publicly tracked, with much of that fishing taking place around South Asia, Southeast Asia and Africa. We also find that 21–30% of transport and energy vessel activity is missing from public tracking systems. Globally, fishing decreased by 12 ± 1% at the onset of the COVID-19 pandemic in 2020 and had not recovered to pre-pandemic levels by 2021. By contrast, transport and energy vessel activities were relatively unaffected during the same period. Offshore wind is growing rapidly, with most wind turbines confined to small areas of the ocean but surpassing the number of oil structures in 2021. Our map of ocean industrialization reveals changes in some of the most extensive and economically important human activities at sea.

FINANCIAL TIMES

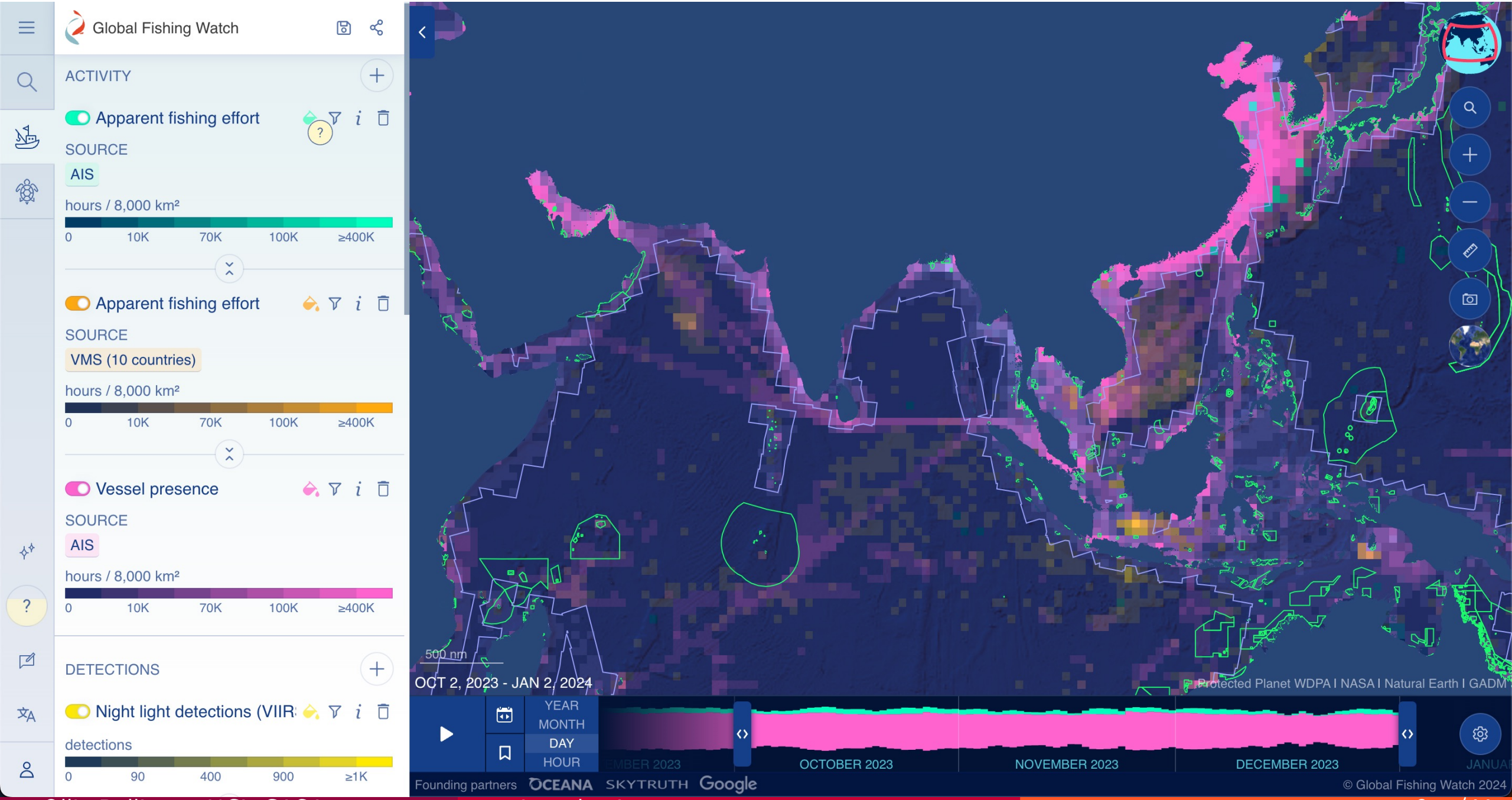
Off-radar fishing threatens efforts to preserve stocks, study warns

Majority of industrial vessels in world's oceans are not publicly tracked, says Global Fishing Watch

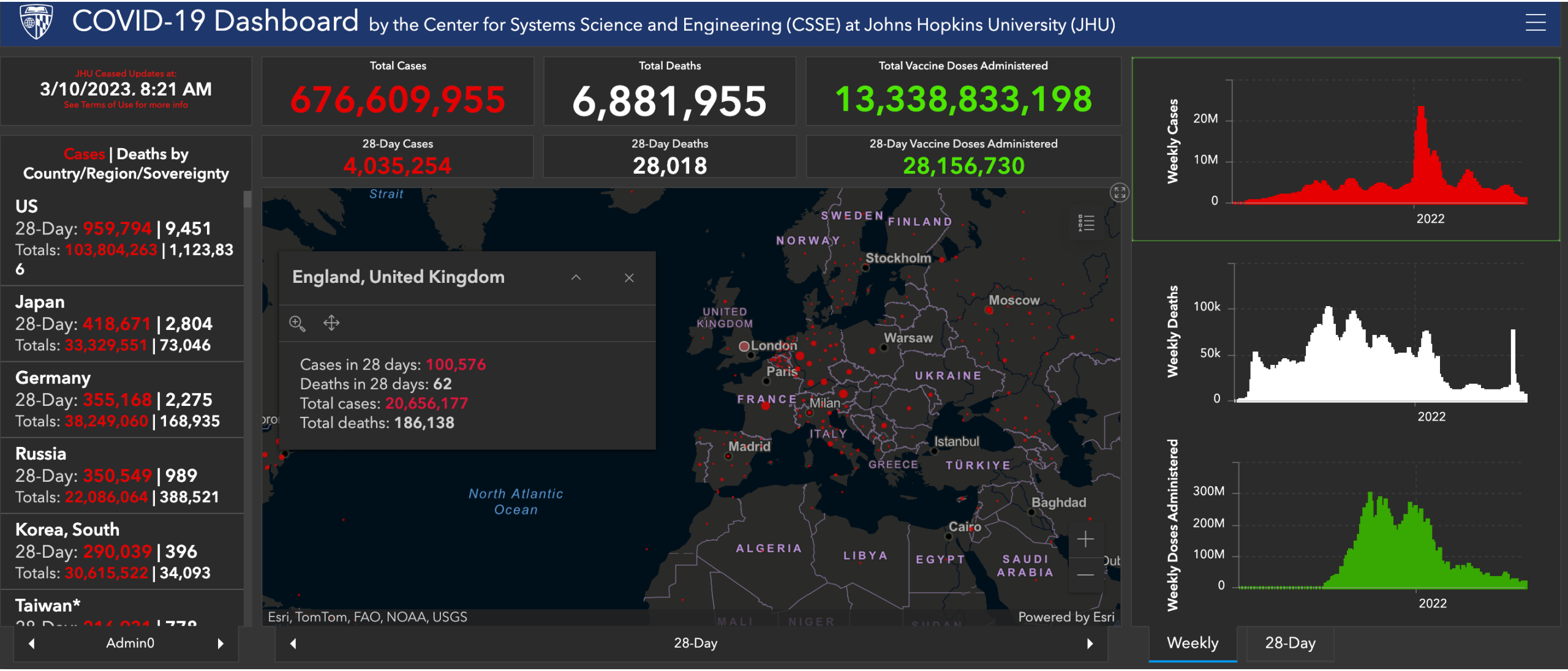
The Guardian
Newspaper of the year

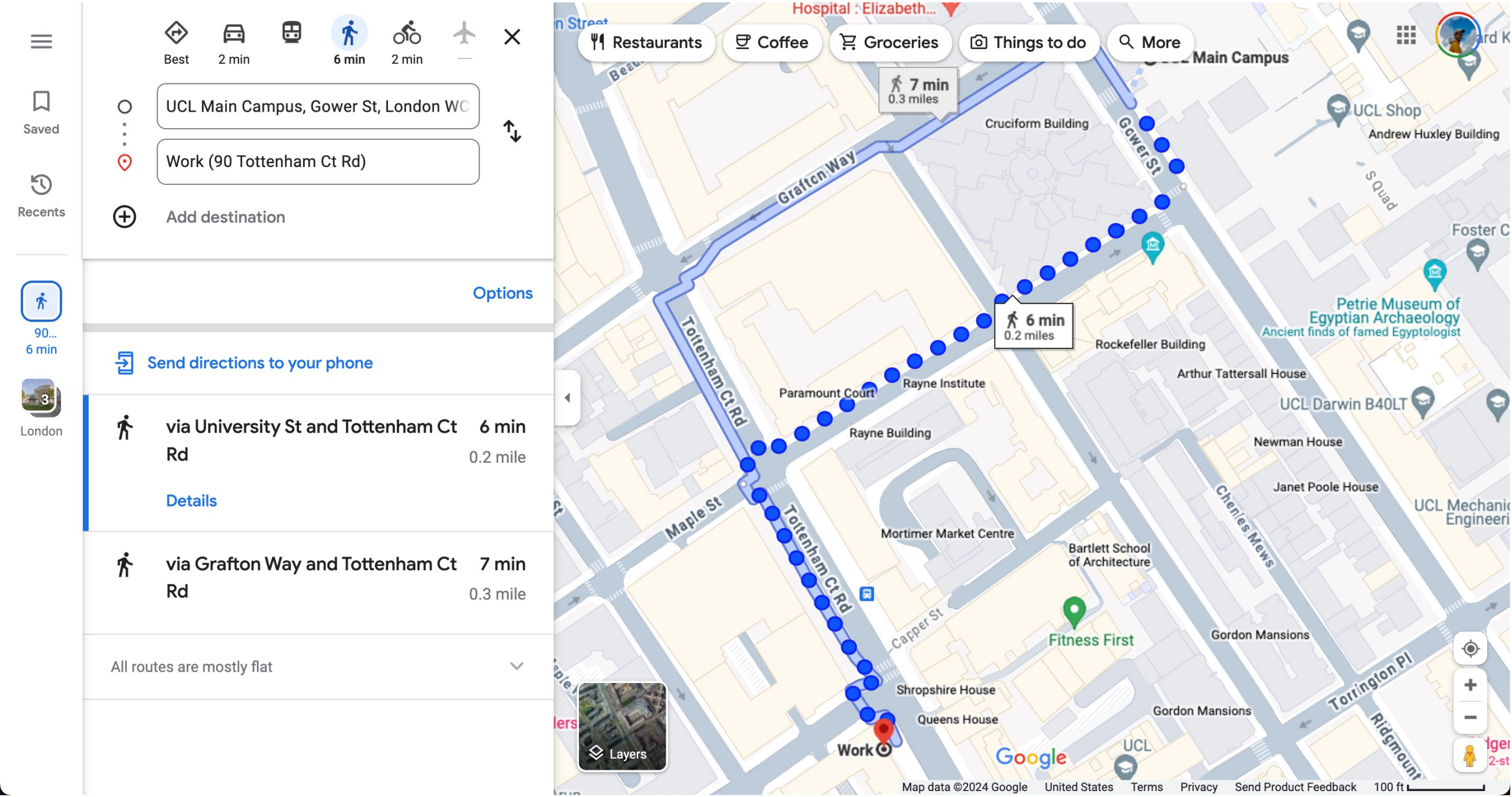
Human activity is powering 'a new industrial revolution' at sea, say experts

Researchers using AI and satellite imagery find 75% of industrial fishing is not being publicly tracked, while wind turbines now outnumber oil platforms

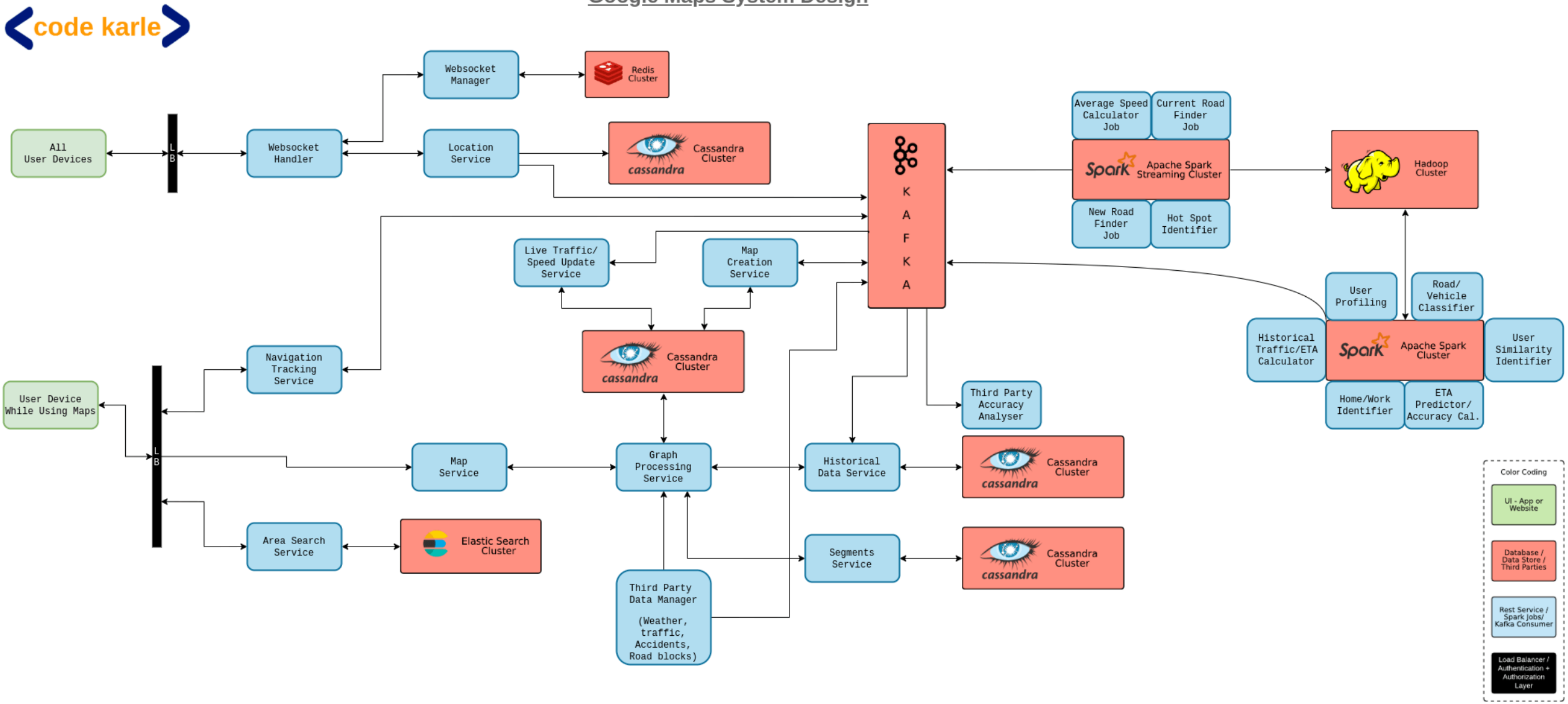

Global Fishing Watch
ACTIVITY
Apparent fishing effort
SOURCE
AIS
hours / 8,000 km²
0
10K
70K
100K
≥400K
Apparent fishing effort
SOURCE
VMS (10 countries)
hours / 8,000 km²
0
10K
70K
100K
≥400K
Vessel presence
SOURCE
AIS
hours / 8,000 km²
0
10K
70K
100K
≥400K
DETECTIONS
Night light detections (VIIR
detections
0
90
400
900
≥1K
500 nm
OCT 2, 2023 - JAN 2, 2024
Protected Planet WDPA I NASA I Natural Earth I GADM
YEAR
MONTH
DAY
HOUR
OCTOBER 2023
NOVEMBER 2023
DECEMBER 2023
Founding partners
OCEANA
SKYTRUTH
Google
© Global Fishing Watch 2024

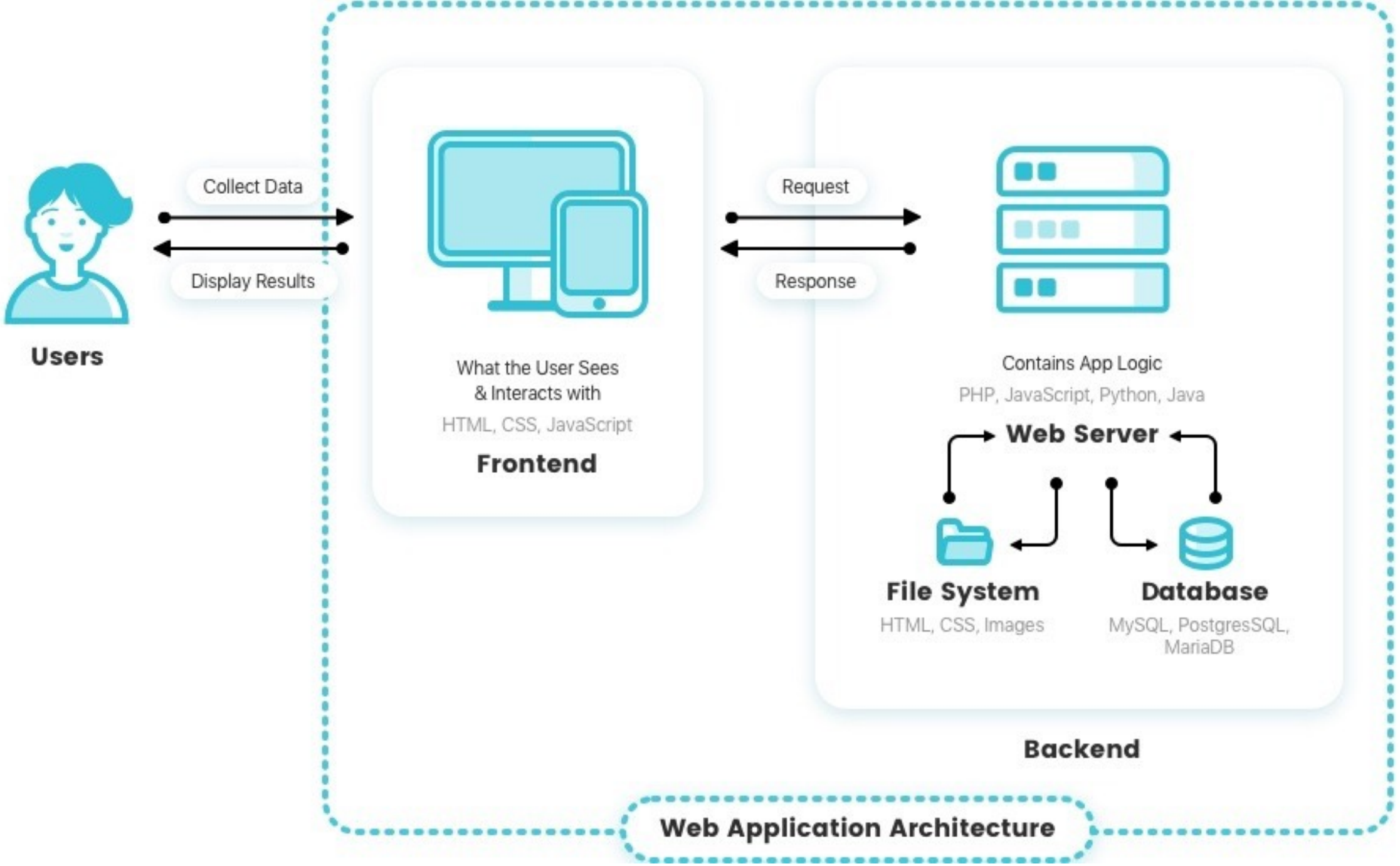

GLOBAL FOREST WATCH
FOREST CHANGE
FOREST COVER
FOREST USE
CONSERVATION
PEOPLE
STORIES
FEEDBACK
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
Google


COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)
JHU Ceased Updates at:
3/10/2023. 8:21 AM
See Terms of Use for more info
Total Cases
676,609,955
Total Deaths
6,881,955
Total Vaccine Doses Administered
13,338,833,198
28-Day Cases
4,035,254
28-Day Deaths
28,018
28-Day Vaccine Doses Administered
28,156,730
Cases | Deaths by Country/Region/Sovereignty
US
28-Day: 959,794 | 9,451
Totals: 103,804,263 | 1,123,836
Japan
28-Day: 418,671 | 2,804
Totals: 33,329,551 | 73,046
Germany
28-Day: 355,168 | 2,275
Totals: 38,249,060 | 168,935
Russia
28-Day: 350,549 | 989
Totals: 22,086,064 | 388,521
Korea, South
28-Day: 290,039 | 396
Totals: 30,615,522 | 34,093
Taiwan*
England, United Kingdom
Cases in 28 days: 100,576
Deaths in 28 days: 62
Total cases: 20,656,177
Total deaths: 186,138
Strait
North Atlantic Ocean
SWEDEN
FINLAND
NORWAY
Stockholm
Moscow
UNITED KINGDOM
London
Warsaw
Paris
UKRAINE
FRANCE
Milan
ITALY
Madrid
Istanbul
GREECE
TÜRKIYE
Baghdad
Cairo
ALGERIA
LIBYA
EGYPT
SAUDI ARABIA
MALI
NIGER
Esri, TomTom, FAO, NOAA, USGS
Powered by Esri
Weekly Cases
20M
10M
0
2022
Weekly Deaths
100k
50k
0
2022
Weekly Doses Administered
300M
200M
100M
0
2022
Admin0
28-Day
Weekly
28-Day


Best
2 min
6 min
2 min
UCL Main Campus, Gower St, London WC
Work (90 Tottenham Ct Rd)
Add destination
Options
Send directions to your phone
via University St and Tottenham Ct Rd
6 min
0.2 mile
Details
via Grafton Way and Tottenham Ct Rd
7 min
0.3 mile
All routes are mostly flat
Saved
Recents
90...
6 min
London
Restaurants
Coffee
Groceries
Things to do
More
7 min
0.3 miles
6 min
0.2 miles
Cruciform Building
Grafton Way
Gower St
Tottenham Ct Rd
Maple St
Paramount Court
Rayne Institute
Rayne Building
Rockefeller Building
Arthur Tattersall House
Mortimer Market Centre
Bartlett School of Architecture
Fitness First
Gordon Mansions
Shropshire House
Queens House
Work
UCL Shop
Andrew Huxley Building
Petrie Museum of Egyptian Archaeology
Ancient finds of famed Egyptologist
UCL Darwin B40LT
Newman House
Janet Poole House
Cheneies Mews
Torrington Pl
Ridgmount
Capper St
Layers
Google
Map data ©2024 Google
United States
Terms
Privacy
Send Product Feedback
100 ft

Google Maps System Design

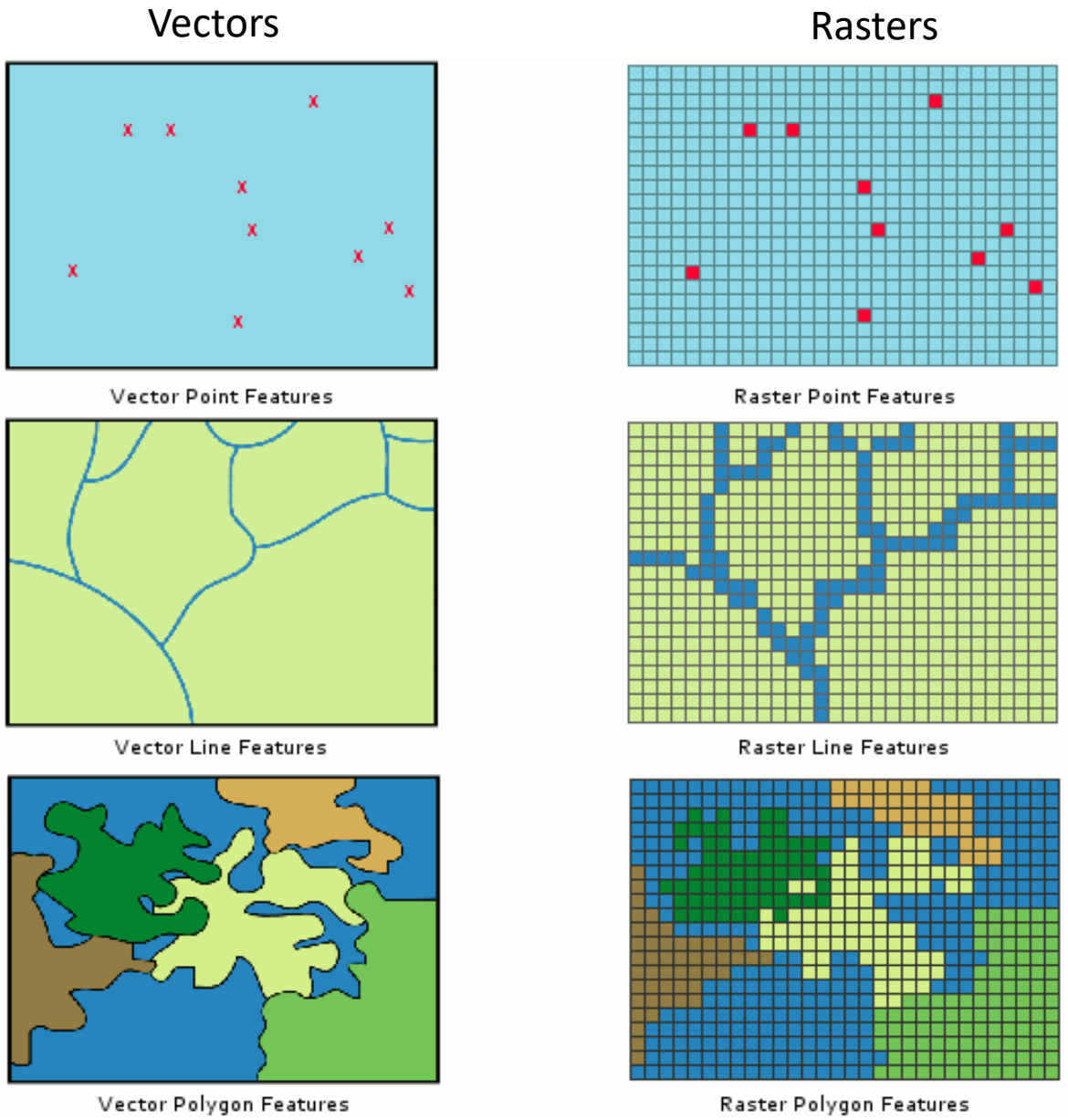



Users
Collect Data
Display Results
What the User Sees
& Interacts with
HTML, CSS, JavaScript
Frontend
Request
Response
Contains App Logic
PHP, JavaScript, Python, Java
Web Server
File System
HTML, CSS, Images
Database
MySQL, PostgresSQL,
MariaDB
Backend
Web Application Architecture

Spatial Data

- There are two main types of spatial data:
  - **Vector**
    - Tabular data with a geometry column
  - **Raster**
    - A digital image, i.e. a matrix of values

- **Handling "big" spatial data depends heavily on its type.**



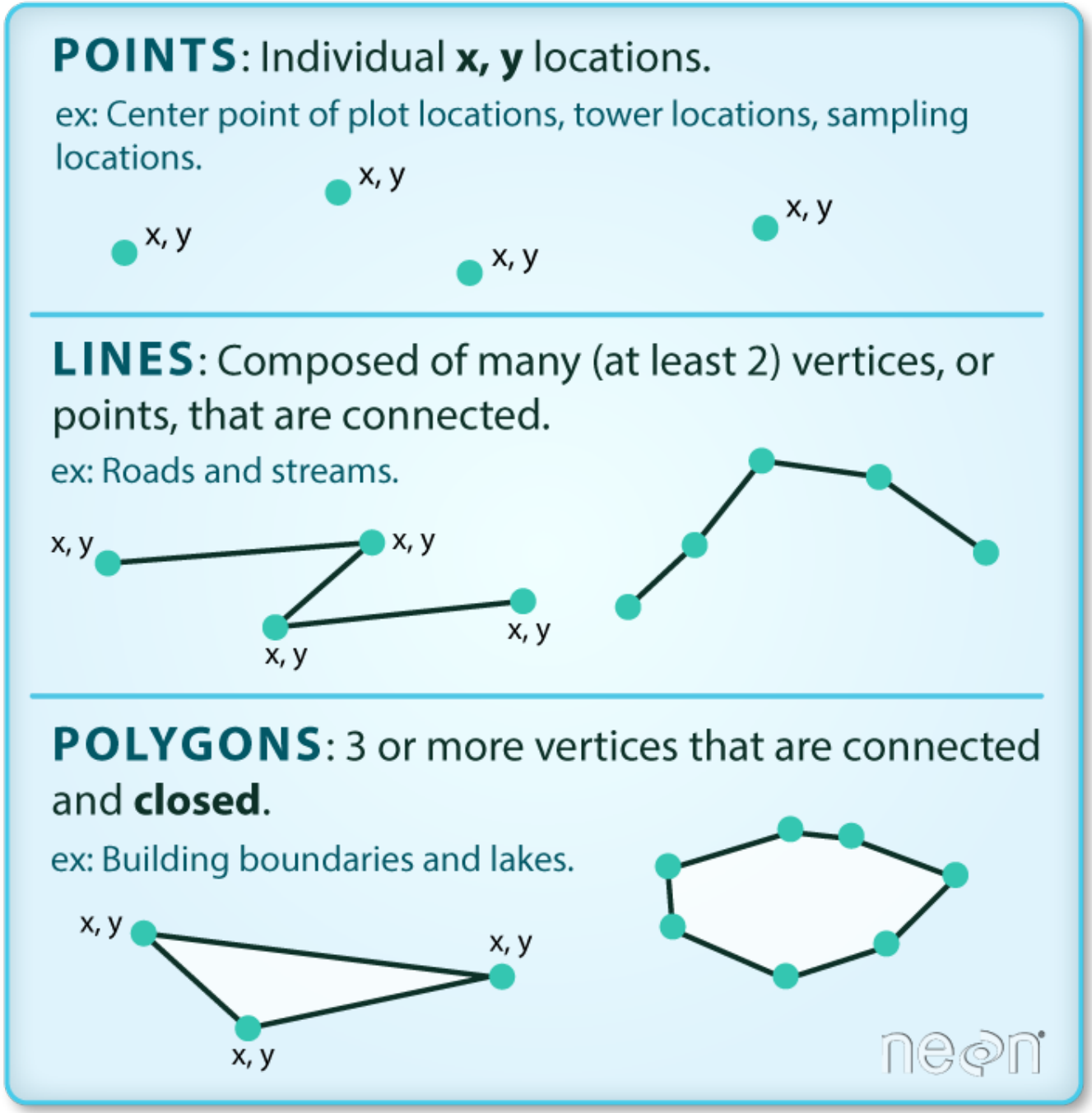
Vector Data

Vector Data: Definition

- Vector data can be thought of as a list of values. The features are recorded one by one, with shape being defined by the numerical values of the pairs of xy coordinates, so that:
  - A point is defined by a single pair of coordinate values.
  - A line is defined by a sequence of coordinate pairs defining the points through which the line is drawn.
  - An area is defined in a similar way, only with the first and last points joined to make a complete enclosure.
- The position and shape of a building is captured as a series of four pairs of numerical coordinates. To reproduce the building in a GIS the computer reads these values and draws a line linking the coordinate positions.
- The vector version can also store additional context information about these features – the **attributes** – a very important aspect.

Vector Data Types

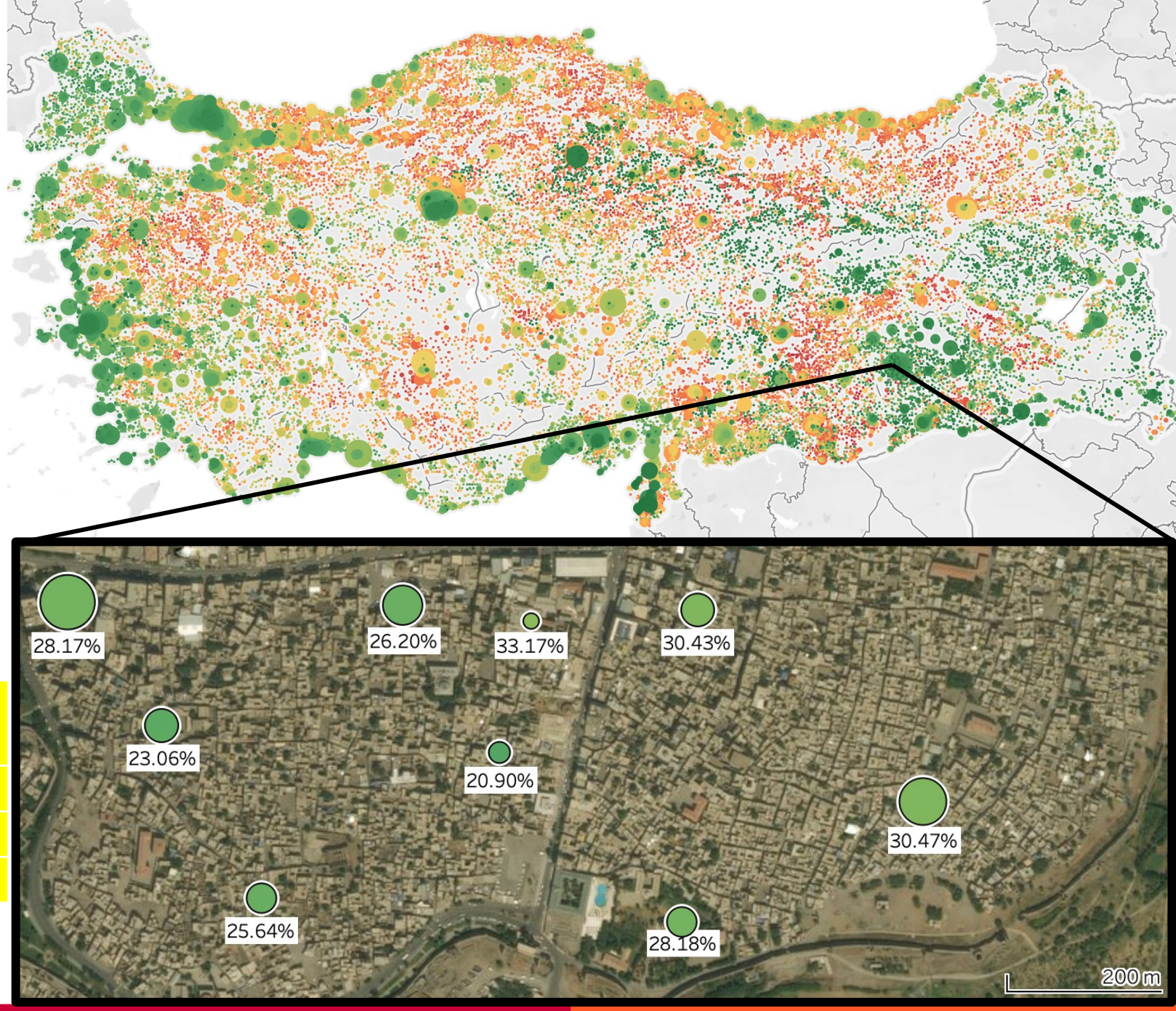
- CSV
  - points
- Shapefile
  - Lines/polygons
- GeoJSON
  - Points/lines/polygons
- KML/GPX
  - Lines/points



Point Data

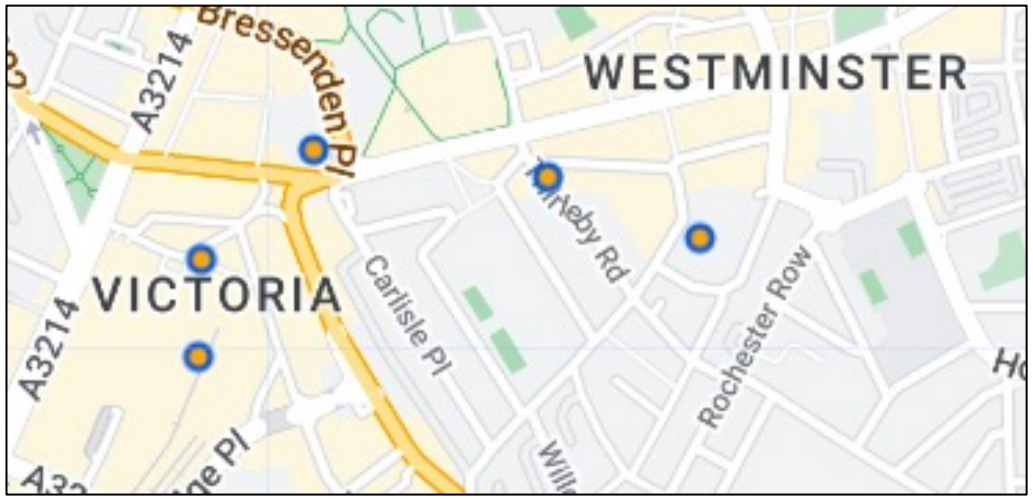
- e.g. ballot box level election results
- Each point can have many attributes including vote share for different parties, etc.

Ballot Box	Party A Vote Share	latitude	longitude
Istanbul 1	0.39	34.349	-84.134
Istanbul 2	0.99	34.079	-84.189
Ankara 1	0.50	34.456	-84.521



Point Data

device_id	timestamp	ip	user_agent	OS	OS_version	manufacturer	model	carrier	latitude	longitude
7b7ad340-630e	1/1/22 9:31	123.456.789	13.3.1	iOS	13	Apple	iPhone	giffgaff	34.349	-84.134
7b7ad340-630e	1/1/22 9:52	123.456.789	13.3.1	iOS	13	Apple	iPhone	giffgaff	34.079	-84.189
7b7ad340-630e	1/1/22 17:13	123.456.789	13.3.1	iOS	13	Apple	iPhone	giffgaff	34.456	-84.521

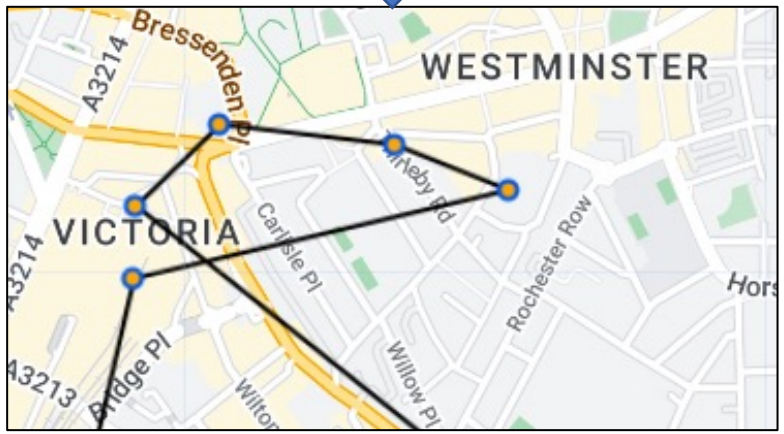


Producer
(app developer)

Line Data

device_id	timestamp	ip	user_agent	OS	OS_version	manufacturer	model	carrier	latitude	longitude
7b7ad340-630e	1/1/22 9:31	123.456.789	13.3.1	iOS	13	Apple	iPhone	giffgaff	34.349	-84.134
7b7ad340-630e	1/1/22 9:52	123.456.789	13.3.1	iOS	13	Apple	iPhone	giffgaff	34.079	-84.189
7b7ad340-630e	1/1/22 17:13	123.456.789	13.3.1	iOS	13	Apple	iPhone	giffgaff	34.456	-84.521

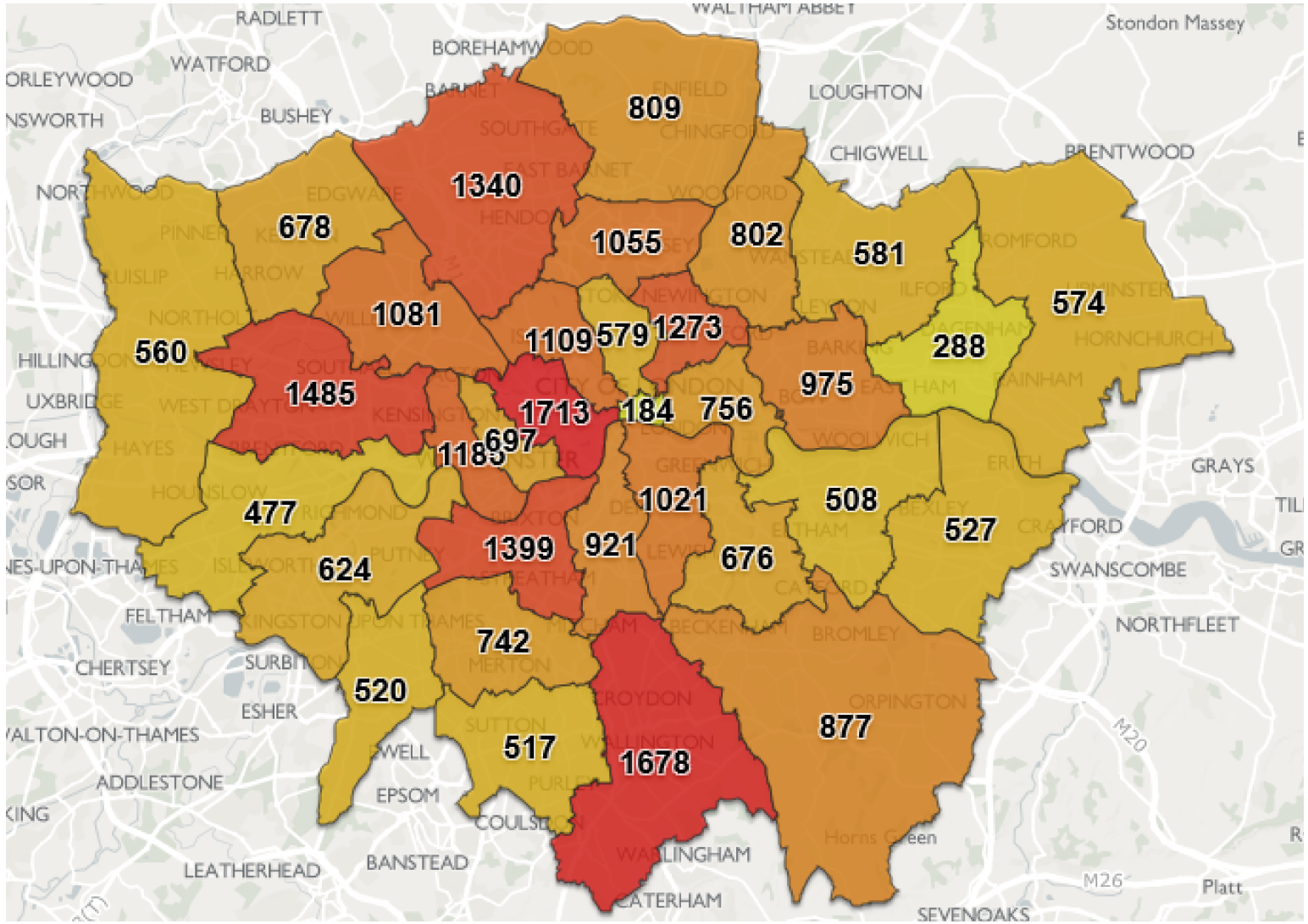
device_id	timestamp	ip	user_agent	OS	OS_version	manufacturer	model	carrier	Geometry
7b7ad340-630e	1/1/22 9:31	123.456.789	13.3.1	iOS	13	Apple	iPhone	giffgaff	LINESTRING(34.349 -84.134, 34.079 -84.189, 34.456 -84.521)



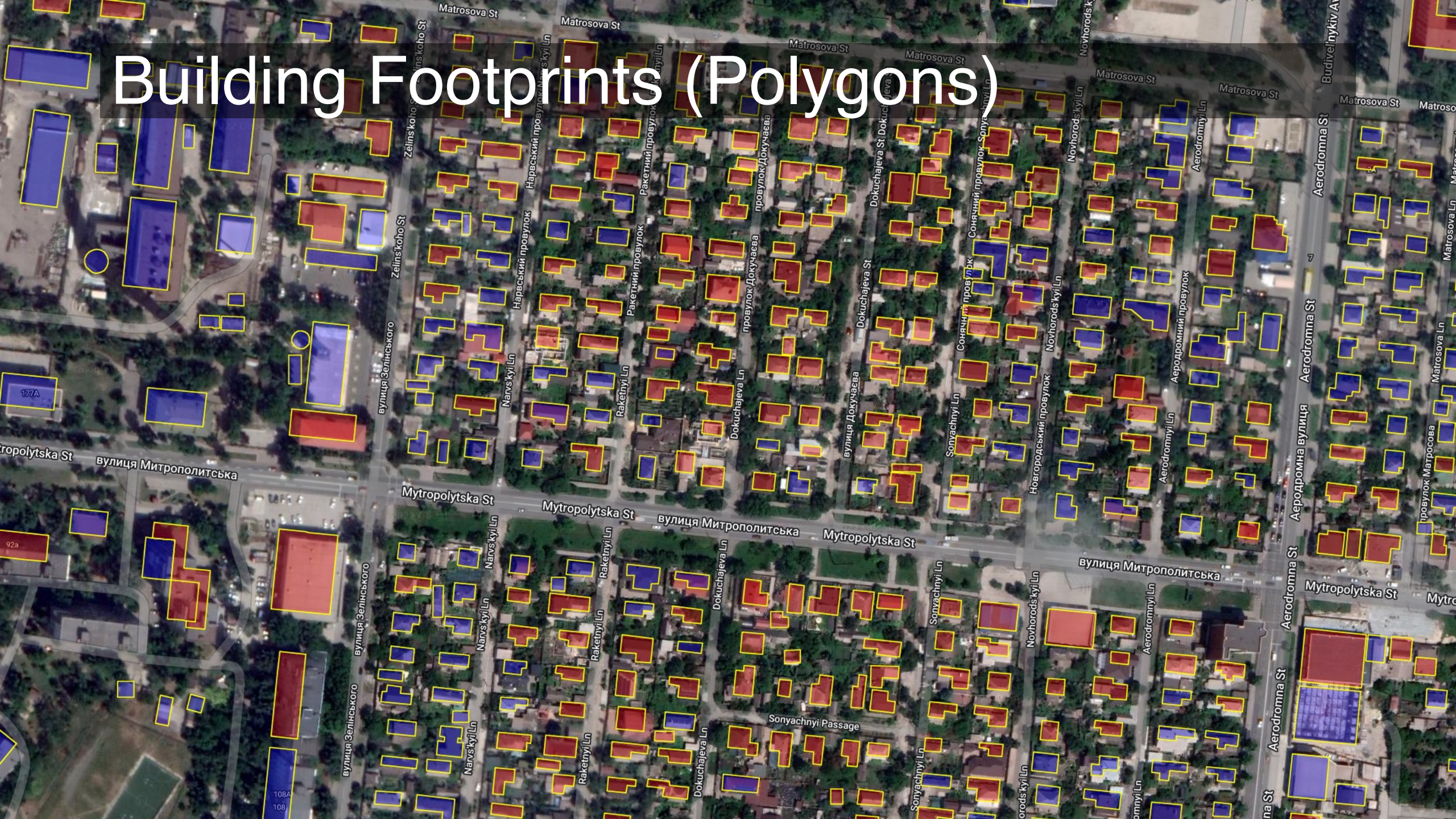
Producer
(app developer)

Polygon Data

Borough	Population	Number of Pubs	geometry
Hackney	280,000	2300	POLYGON(34.079 - 0.381 …
Camden	279,000	1943	POLYGON(35.419 - 0.352 …
Westminster	261,000	532	POLYGON(33.044 0.431 …



Building Footprints (Polygons)

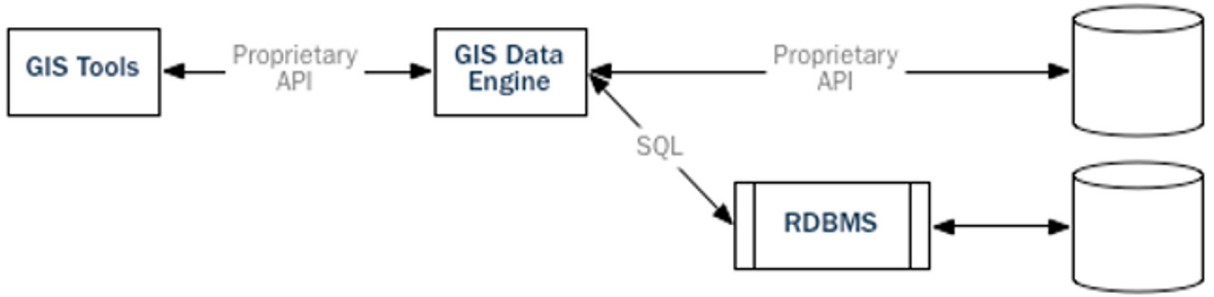


Evolution of GIS Architectures

First-Generation GIS:

Geometry, indexes, and attributes stored in file system

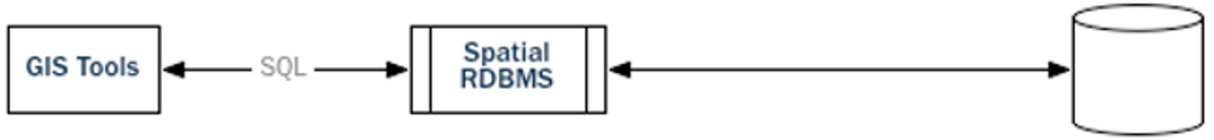
Second-Generation GIS:



Geometry, indexes, and attributes stored in file system (may also be stored in BLOBs in RDBMS)

Attributes stored in RDBMS tables

Third-Generation GIS:



Geometry stored in ADTs in RDBMS tables with related business data

What is a database?

System for storage and random access of relationally (tables of rows and columns) structured data, providing the following capabilities for that data.

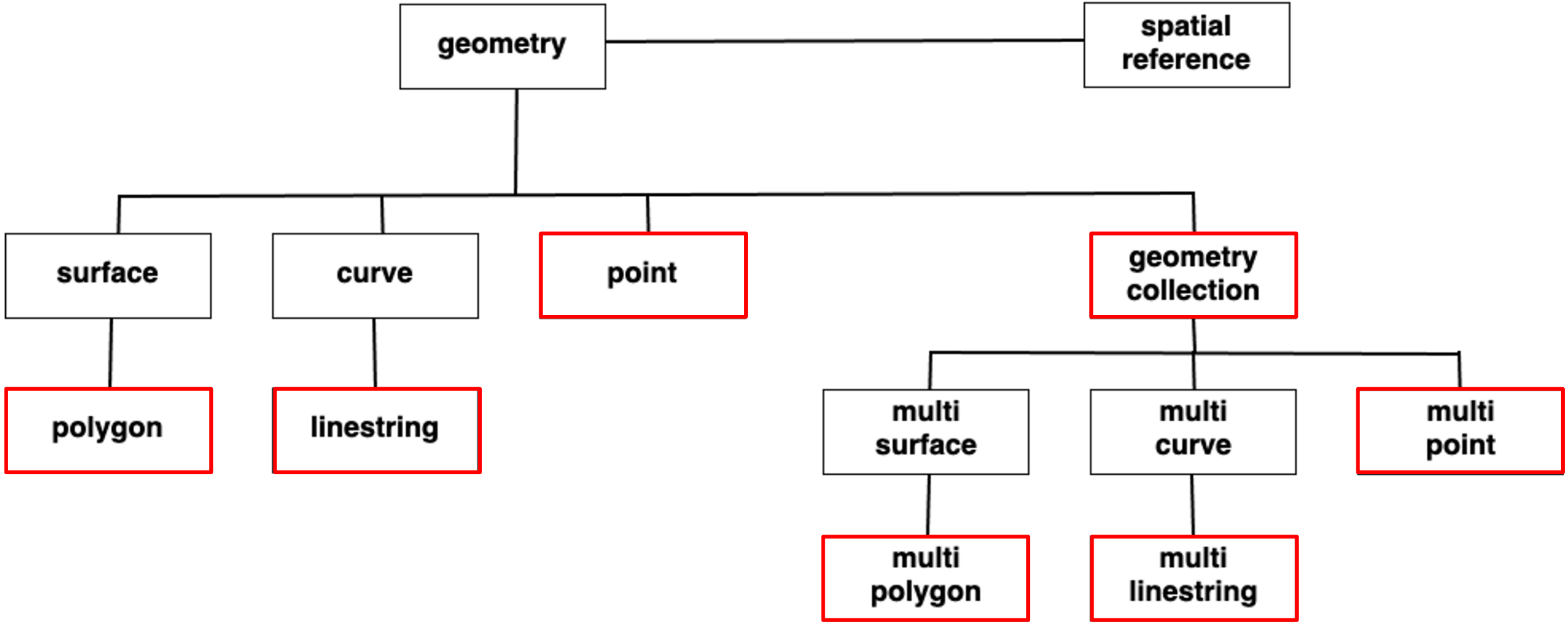
- **Data Types**
  - number, date, and string
- **Indexes**
  - b-tree, hash
- **Functions**
  - strlen(string), pow(float, float), now()

What is a spatial database?

System for storage and random access of relationally (tables of rows and columns) structured data, providing the following capabilities for that data.

- **Data Types** including Spatial Types
  - number, date, string, geometry, geography and raster
- **Indexes** including Spatial Indexes
  - b-tree, hash, rtree, quadtree
- **Functions** including Spatial Functions
  - strlen(string), pow(float, float), now(), ST_Area(), ST_Distance()

Spatial Types



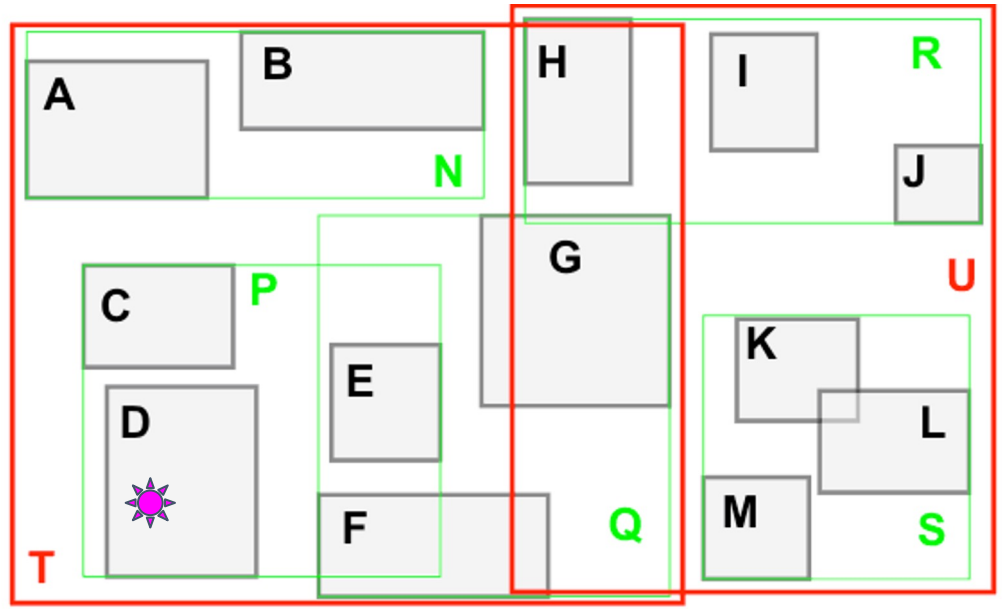
Spatial Indexes

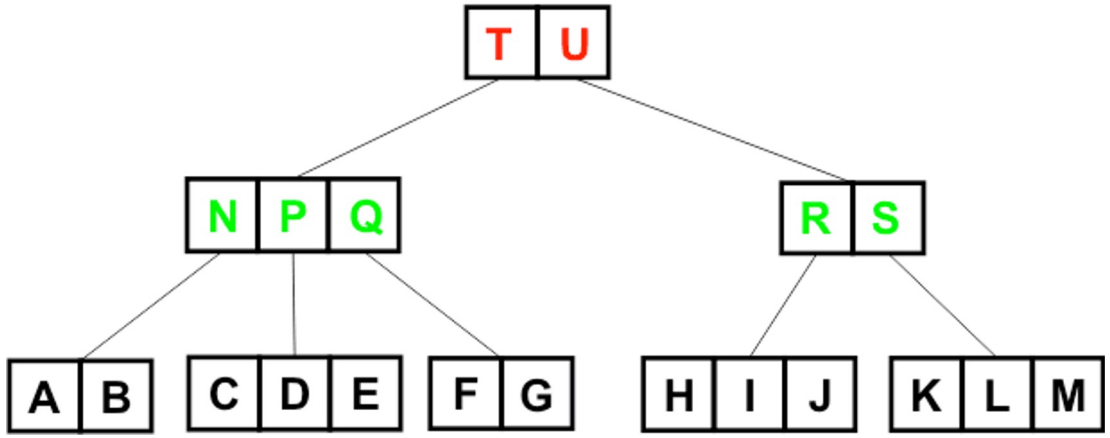
This R-Tree organizes the spatial objects so that a spatial search is a quick walk through the tree.

To find what object contains ✹ ?

- The system first checks if it is in **T** or **U** (**T**)
- Then it checks if it is in **N**, **P** or **Q** (**P**)
- Then it checks if it is in **C**, **D** or **E** (**D**)

Only 8 boxes have to be tested. A full table scan would require *all 13 boxes* to be tested. The larger the table, the *more powerful* the index is.



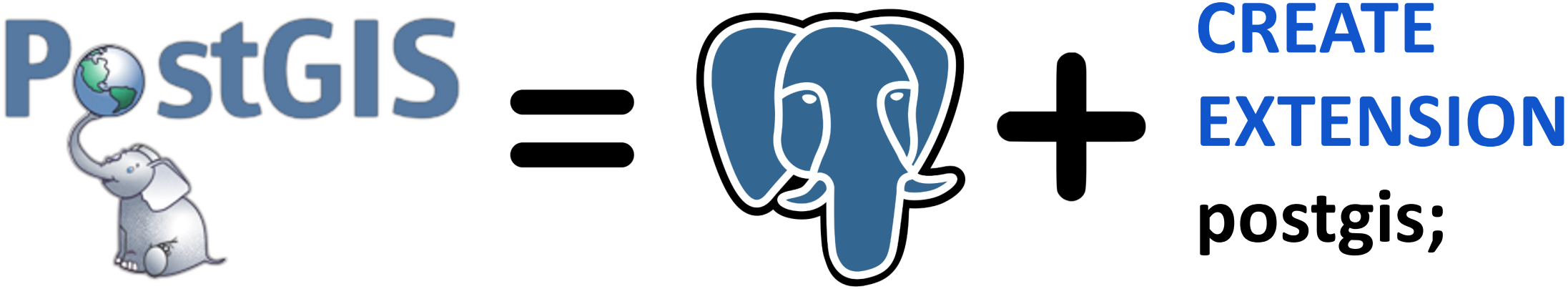


Spatial Functions

For example:

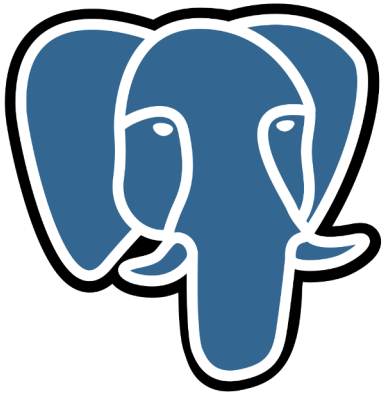
- ST_GeometryType(geometry) → text
- ST_Area(geometry) → float
- ST_Distance(geometry, geometry) → float
- ST_Buffer(geometry, radius) → geometry
- ST_Intersection(geometry, geometry) → geometry
- ST_Union([geometry]) → geometry

What is PostGIS?



What is PostgreSQL?

- Enterprise RDBMS
- Functionally equivalent to Oracle / MSSQL
- Multi-vendor open source community
- Multi-platform support, and available on all clouds
- Highly extensible by design
  - Types, functions, indexes, replication slots, foreign data, core hooks
  - What makes PostGIS possible



Why not files?

- Shape, FGDB, GeoPackage?
- No lingua franca for file access, every format has its own library
  - Database has SQL
  - Databases have JDBC, ODBC, others
- Multi-user access to files results in either
  - Global file locking and performance issues (best case)
  - File corruption and data loss (worst case)
- All database-style queries have to be implemented on client
  - Joins, aggregations, set-based logic

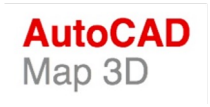
PostGIS reference users - private sector

Ball Aerospace
& Technologies Corp.

PostGIS reference users - government

PostGIS 3rd party integration

Desktop

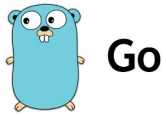



and more...

Middle

and more...

Language

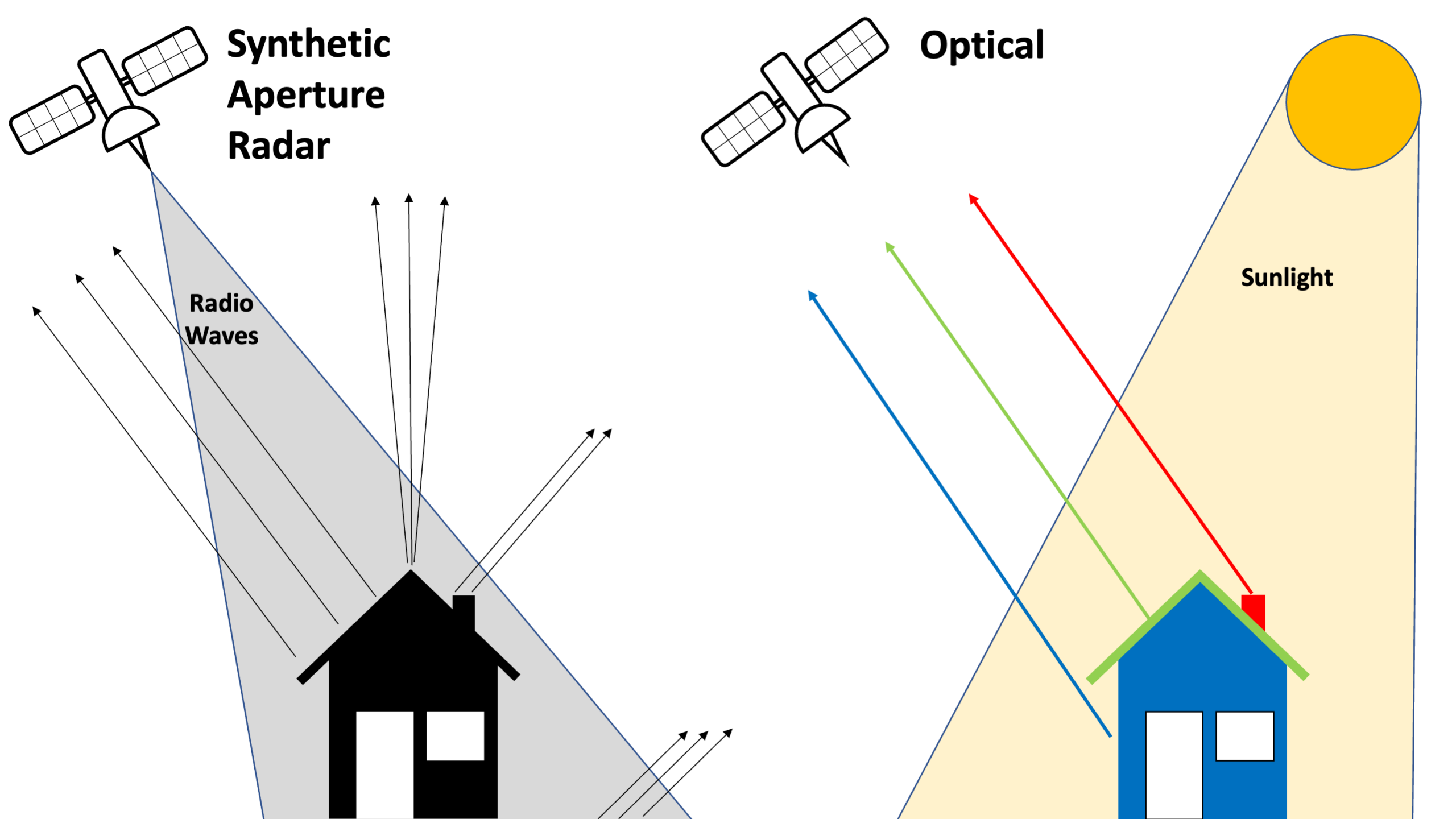


and more...

Raster Data

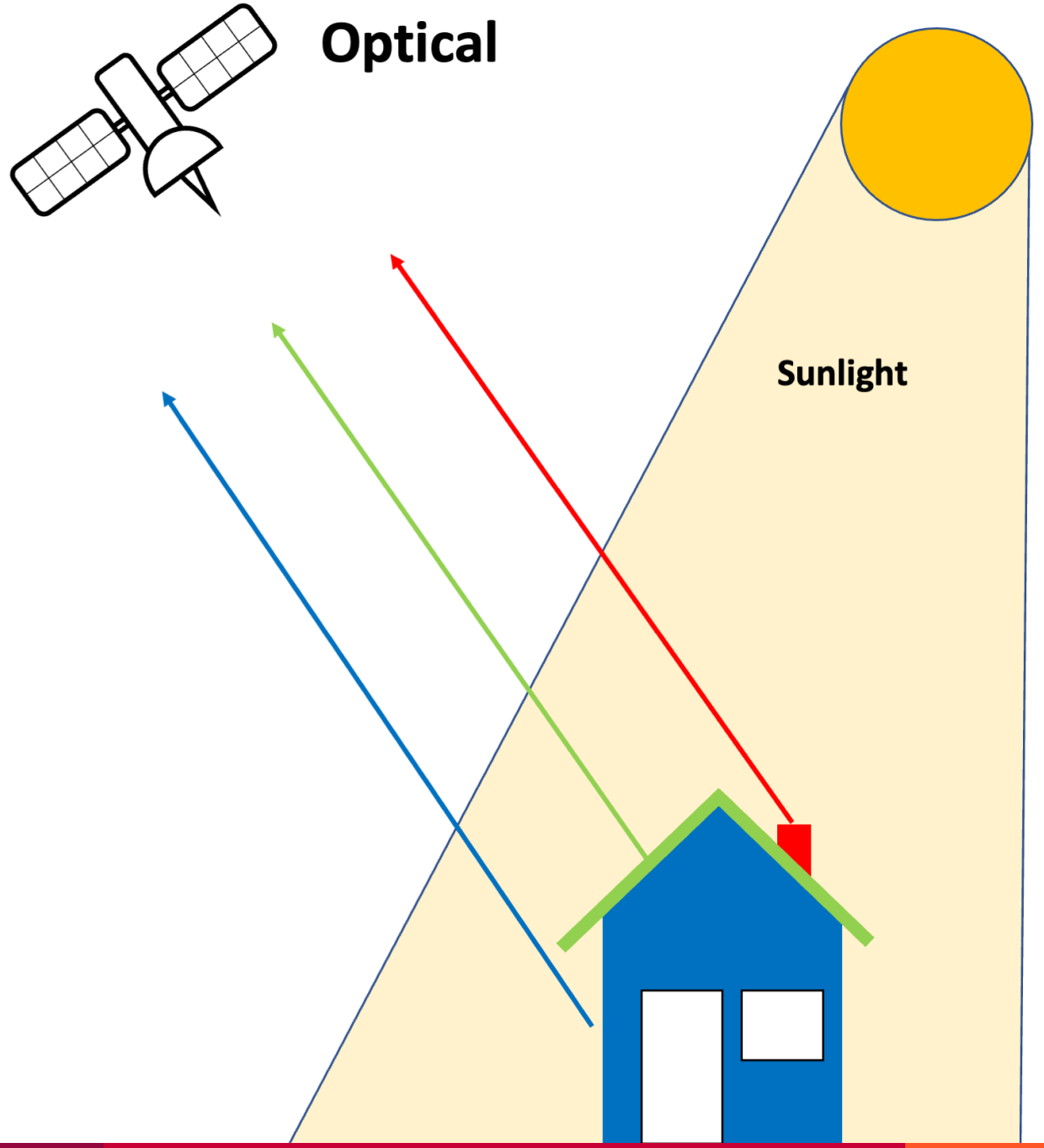
Raster Data: Definition

- Similar to a digital photograph. The entire area of the map is subdivided into a grid of tiny cells, or pixels. A value is stored in each of these cells to represent the nature of whatever is present at the corresponding location on the ground.
- The major use of raster data involves storing map information as digital images, in which the cell values relate to the pixel colors of the image. To reproduce the image the computer reads each of these cell values one by one and applies them to the pixels on the screen.

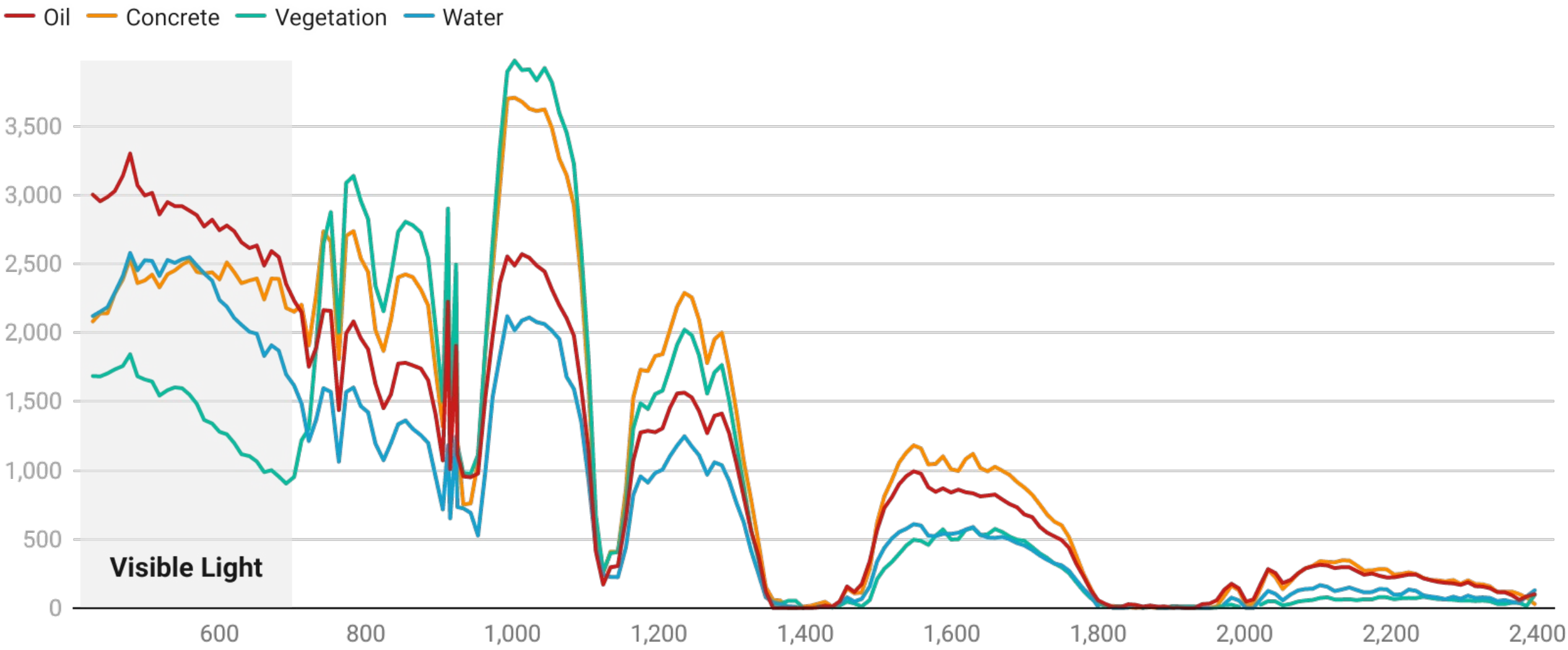

Synthetic Aperture Radar
Radio Waves
Optical
Sunlight

Raster Data: Examples

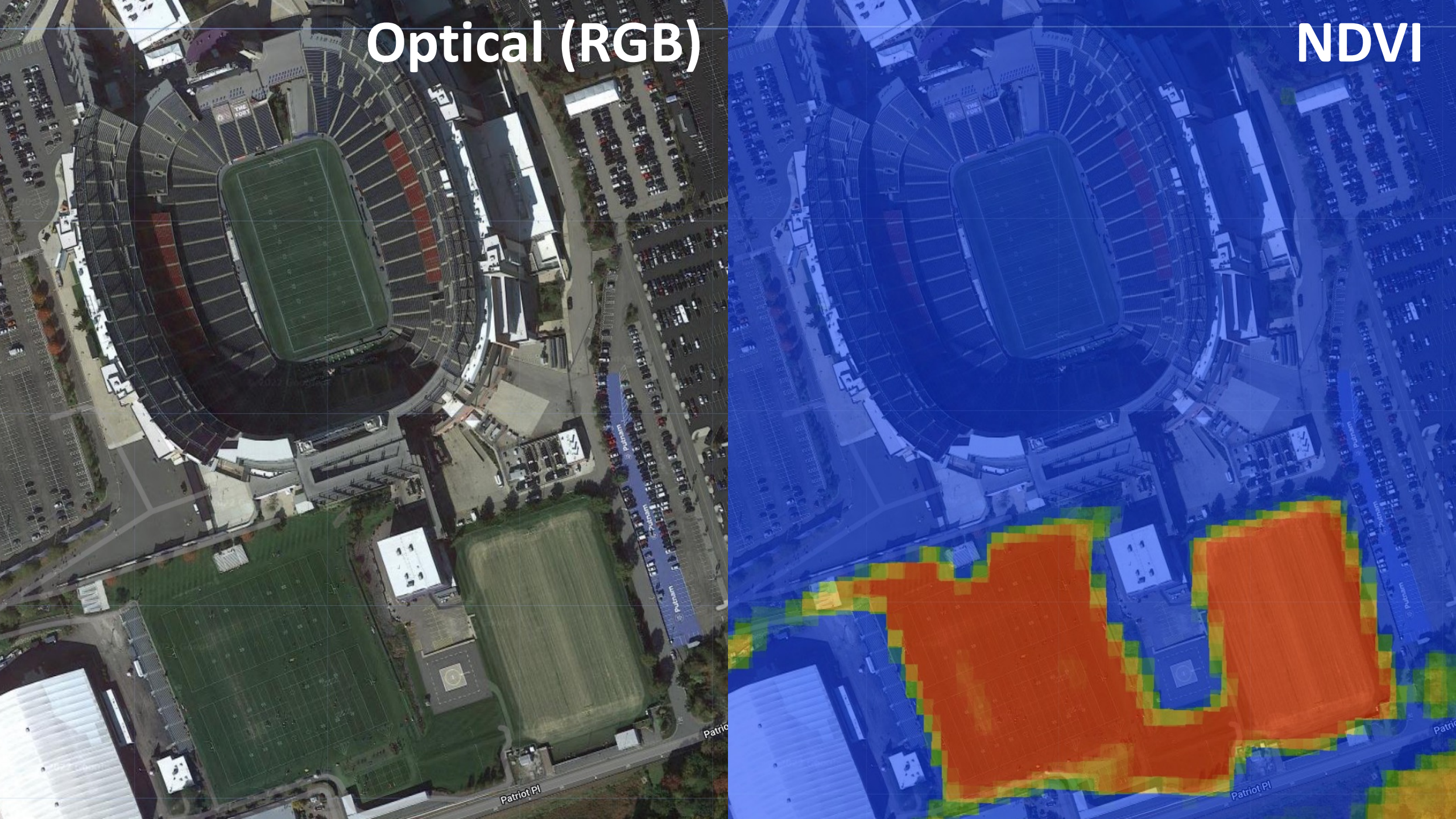
- Satellite imagery
  - Optical
    - High resolution (<1m/pixel)
      - MAXAR, PLEIADES
    - Medium resolution (10-30m per pixel)
      - ESA Sentinel-2, NASA Landsat
    - Low resolution (>30m/pixel)
      - MODIS, VIIRS, GOES
  - Radar
    - High resolution (<1m/pixel)
      - Capella Space, ICEYE
    - Medium resolution (10-30m per pixel)
      - ESA Sentinel-1
    - Low resolution (>30m/pixel)


Optical
Sunlight

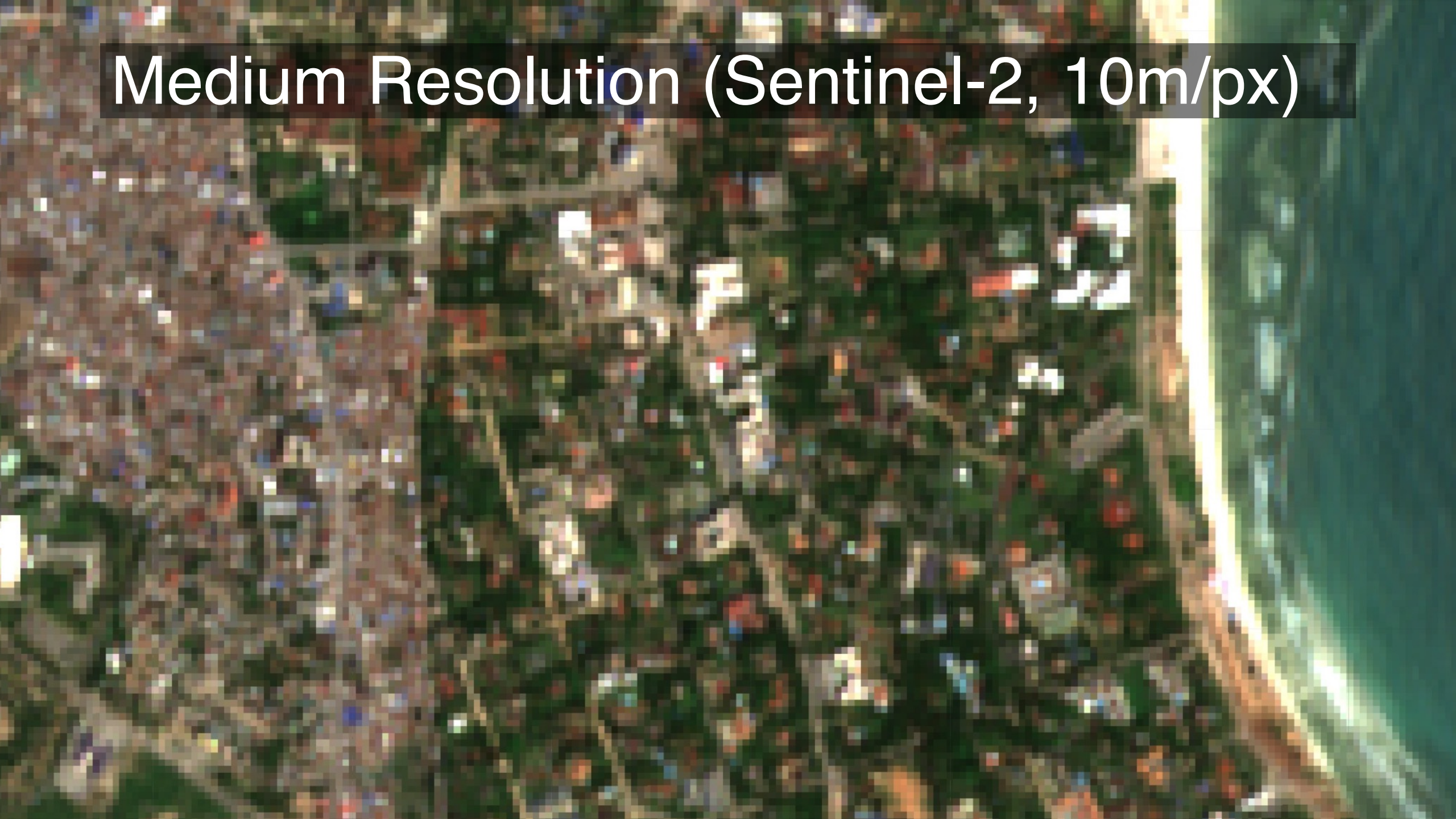
Spectral Profiles of Different Materials



Created with Datawrapper


Optical (RGB)
NDVI
Patriot Pl
Patriot Pl

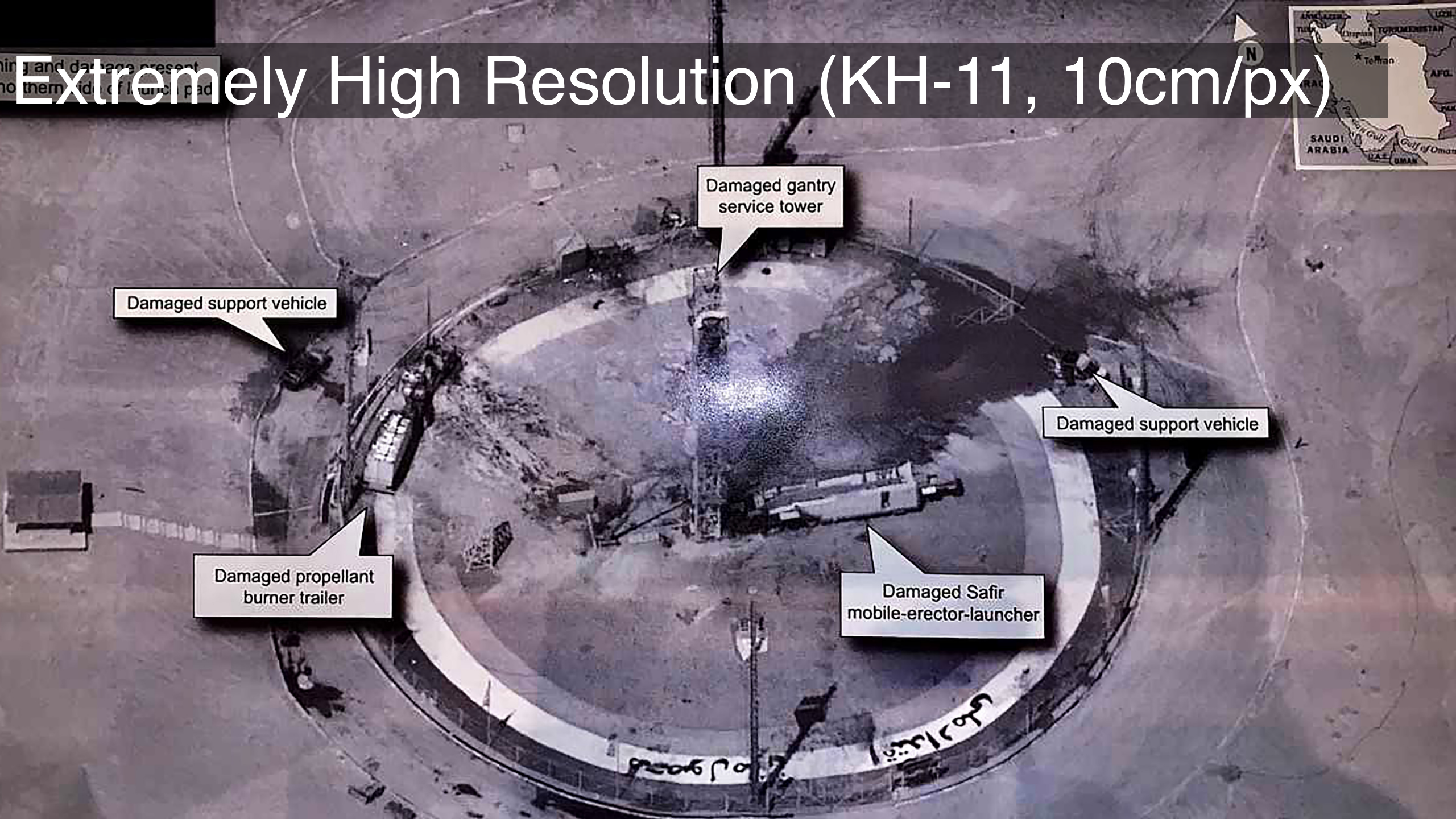

Medium Resolution (Landsat 8, 30m/px)

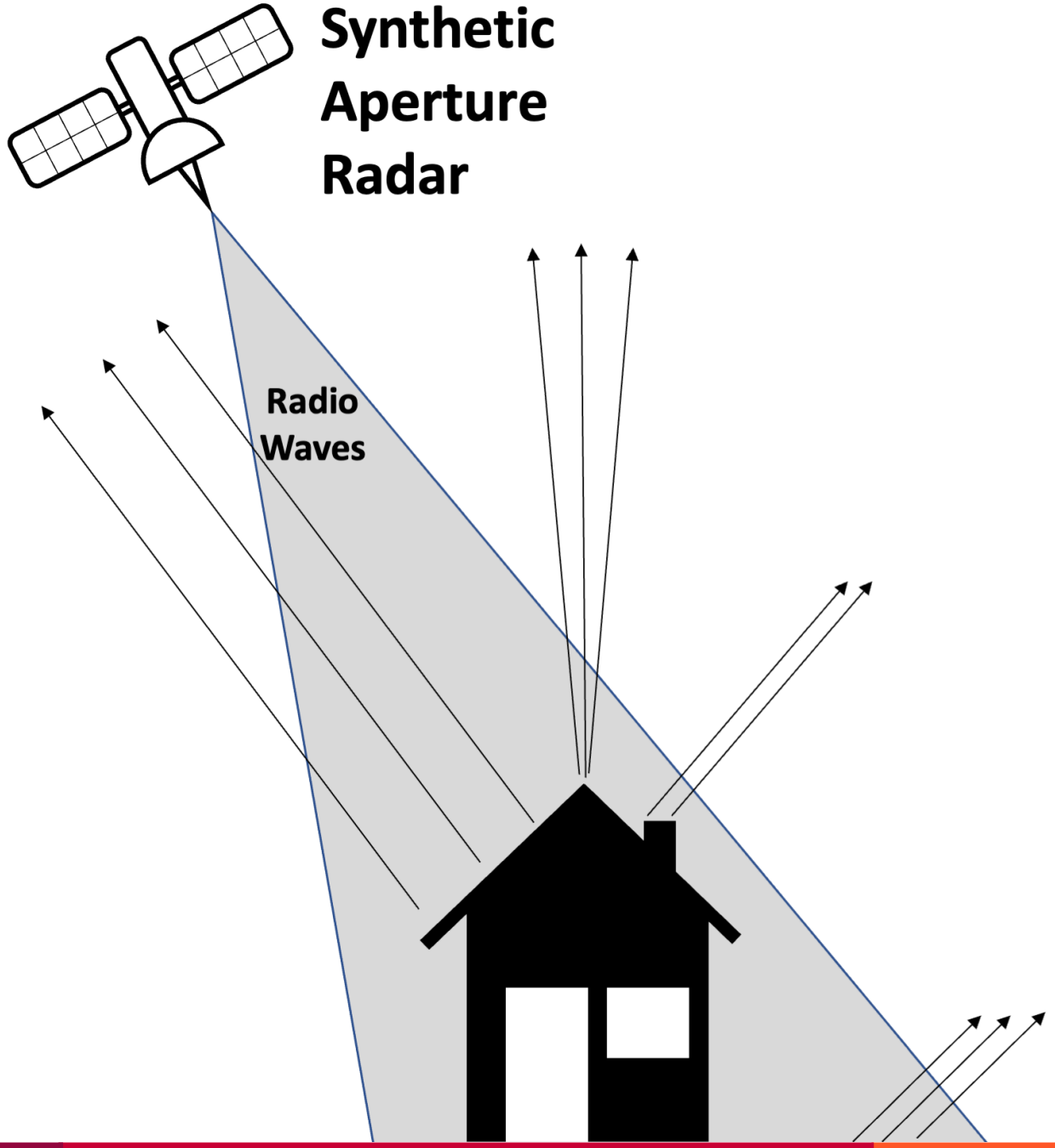

Medium Resolution (Sentinel-2, 10m/px)


High Resolution (Maxar, 50cm/px)


High Resolution (Maxar, 50cm/px)

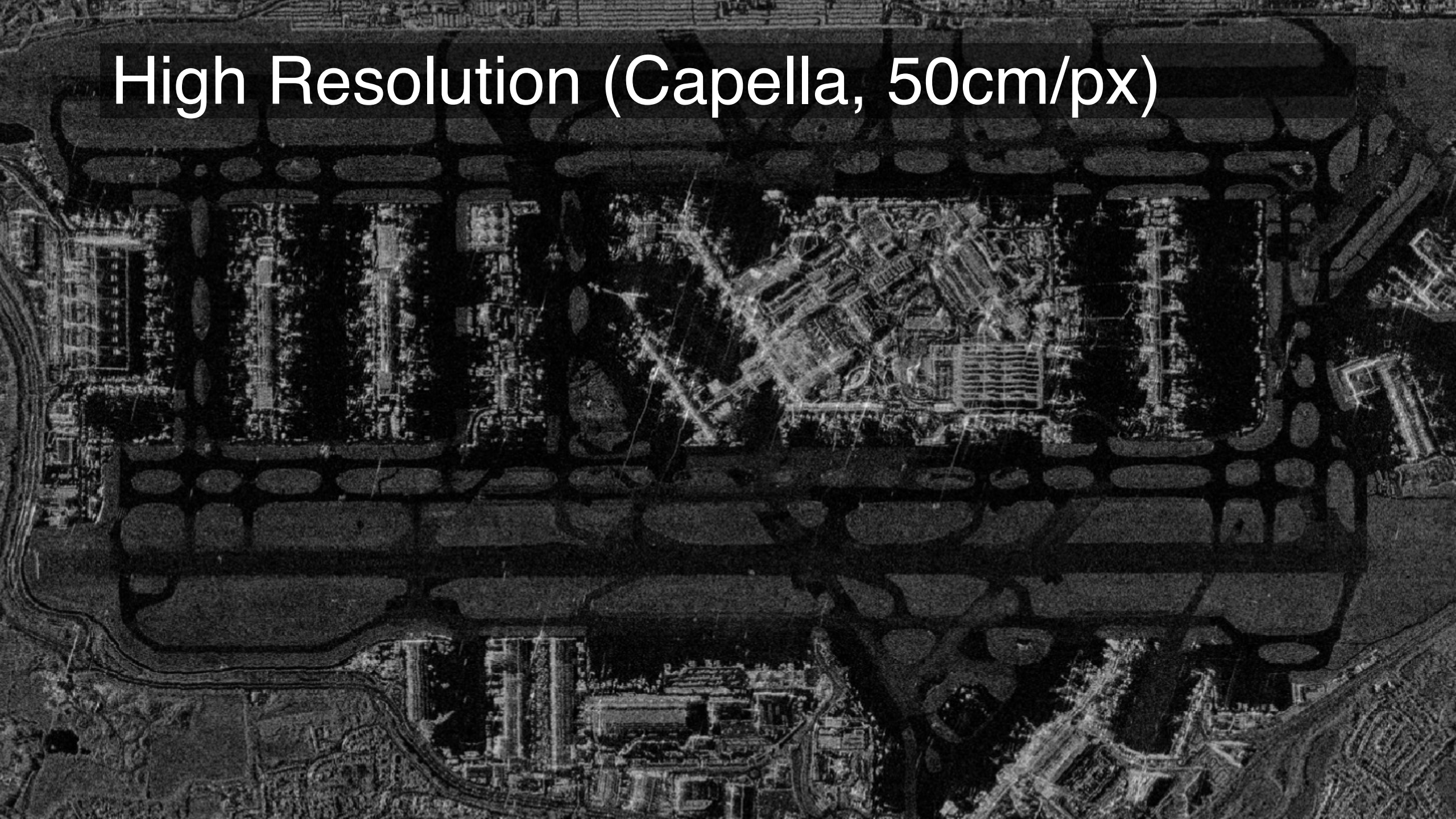
Extremely High Resolution (KH-11, 10cm/px)




Synthetic
Aperture
Radar
Radio
Waves

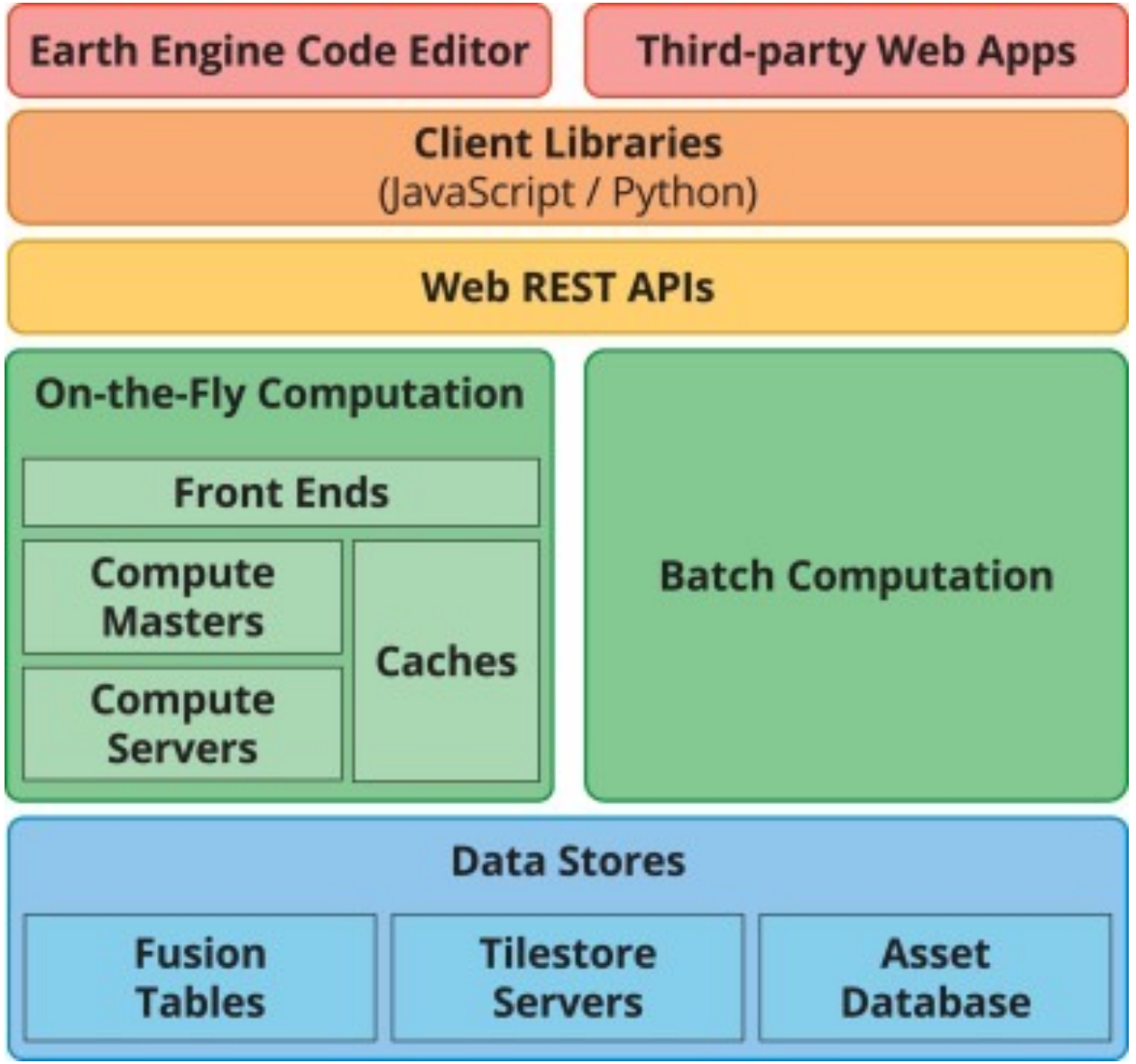

Medium Resolution (Sentinel-1, 10m/px)

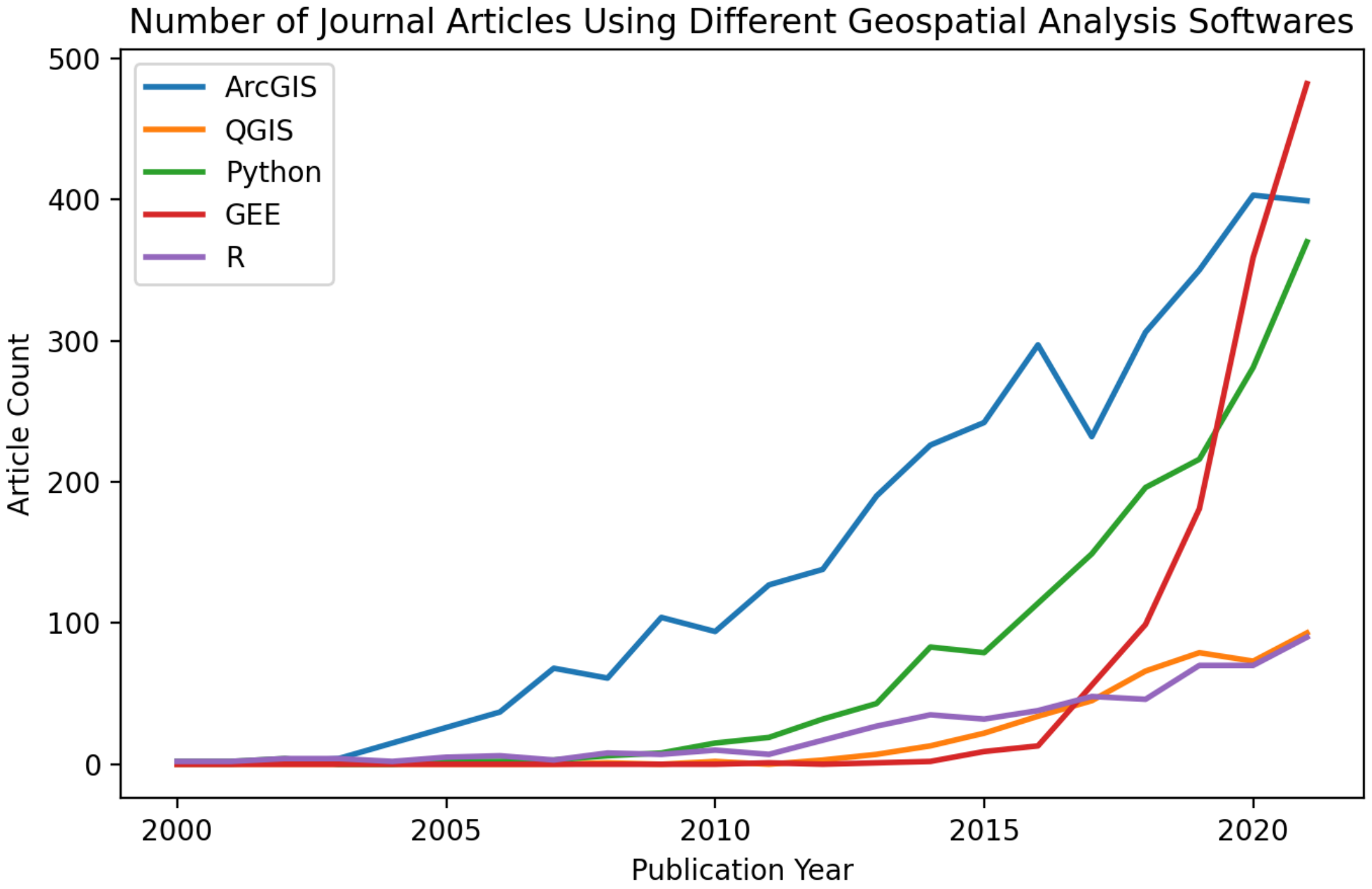
High Resolution (Capella, 50cm/px)



Google Earth Engine

- Asset catalog of over 30 petabytes of geophysical data, mostly sat imagery
  - If you stored this on macbook pros and laid them end-to-end in a straight line, they would stretch from here to Heathrow airport
- Runs large spatial computations on google servers
- Allows for the building and deployment of apps, all in one place




Number of Journal Articles Using Different Geospatial Analysis Softwares
ArcGIS
QGIS
Python
GEE
R
Article Count
0
100
200
300
400
500
2000
2005
2010
2015
2020
Publication Year

Module Overview

Module Outline

Week	Lecture	Tools
1	Introduction	PostgreSQL
2	Databases and SQL	DuckDB
3	Spatial Databases I	PostGIS
4	Spatial Databases II	PostGIS
5	Guest Lecture: GFW	PostGIS
6	Intro to Earth Engine	Google Earth Engine
7	Classification I	Google Earth Engine
8	Classification II	Google Earth Engine
9	Synthetic Aperture Radar	Google Earth Engine
10	User Interface Design	Google Earth Engine

Lectures

- Lectures will provide the conceptual foundations for different aspects of application design.
- Weeks 1-5 will focus on handling large vector datasets using Databases and SQL.
  - Week 5 will feature a guest lecture from Global Fishing Watch
- Weeks 6-10 will introduce analytical concepts in Google Earth Engine
  - Lectures 6-9 will be co-taught with CASA0023 Remote Sensing

Workshops

- Practical sessions involving the application of concepts learned during the lecture
- Aim is to hone our coding skills
- Weeks 1-5
  - We'll be using SQL in the pgAdmin database client.
  - We will also learn how to use Python + VScode to integrate SQL directly into data science workflows
- Weeks 6-10
  - We'll be working in Google Earth Engine's browser-based IDE
  - Weeks 6 and 7 will be joint with CASA0023 Remote Sensing, location TBD

Assessments

- **Database Quiz - 30%**
  - **7th February 2024**
  - Administered during the Workshop on Week 5. It will last one hour. Students will be provided with a dataset and will have to answer 10 questions.
- **Group Application - 50%**
  - **22nd April 2024**
  - The application will import data and apply analysis to one of the datasets from the Earth Engine catalogue, with optional integration of third-party data. The application must allow users to interactively query results.
- **Presentations - 20%**
  - **22nd April 2024**
  - Each group is expected to produce a presentation showing off the application, how the group carried out the analysis of the dataset, the limitations of the analysis and how the interactive tool works under the hood

Workshop