Introduction to CSCI-587: Geospatial Information Management

8/24/2024

University of Southern California
USC Viterbi School of Engineering
Fall 2024

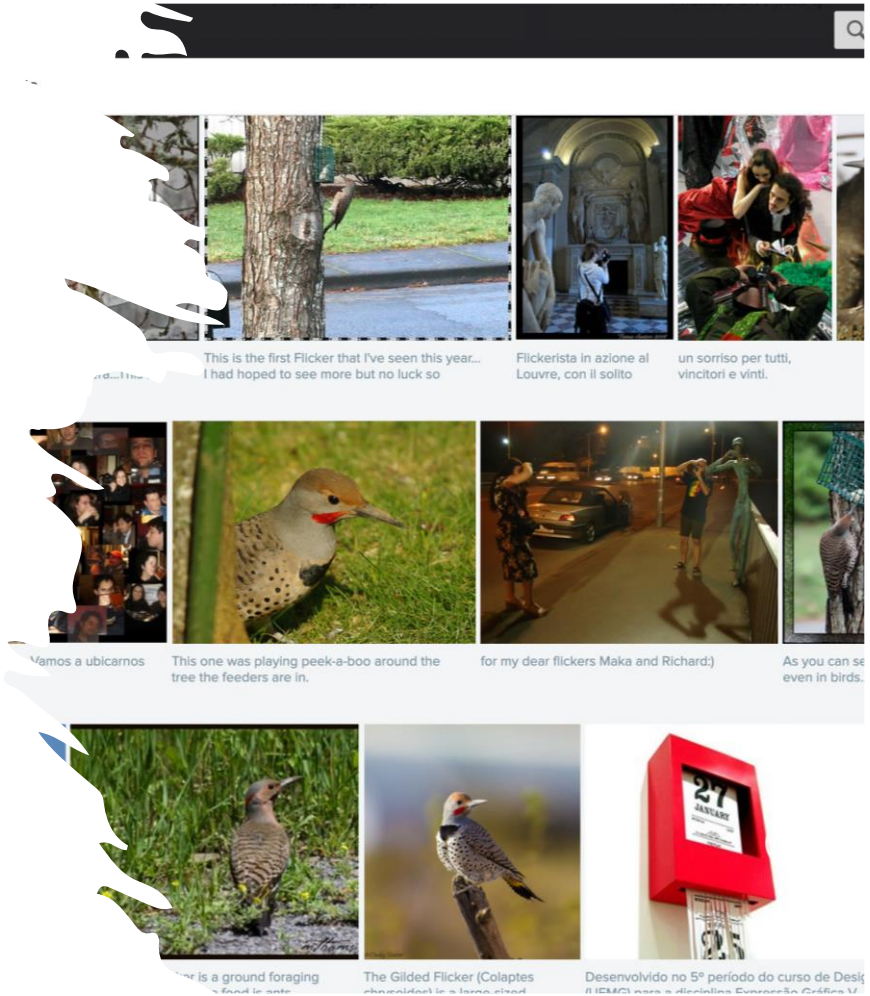
** Some content are adapted from Chiang, Yao-Yi Introduction to Spatial Artificial Intelligence. Available from https://yaoyichi.github.io/spatial-ai.html*

What is Information Management?

What is Data Management?

How do you manage your photos?

- Most cellphones take nice photos
  - Taking 3 photos a day will give you ~1,000 photos a year
  - Taking a 5-day vacation would give you 200 photos
- Ways to managing photos
  - Leave them on the phone?
  - Organize them into folders?
  - Upload them to some cloud services?
- Which method is the best?



Considerations for Managing Photos

Find photos by time

Find photos by subjects

Find photos by locations

Searching must be fast!

Photos need to be secure

Available resources

Data Management (Oracle)

- Data management is the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively.
- help people, organizations, and connected things
  - optimize the use of data within the bounds of policy and regulation
  - (use data to) make decisions and take actions that maximize the benefit to the organization

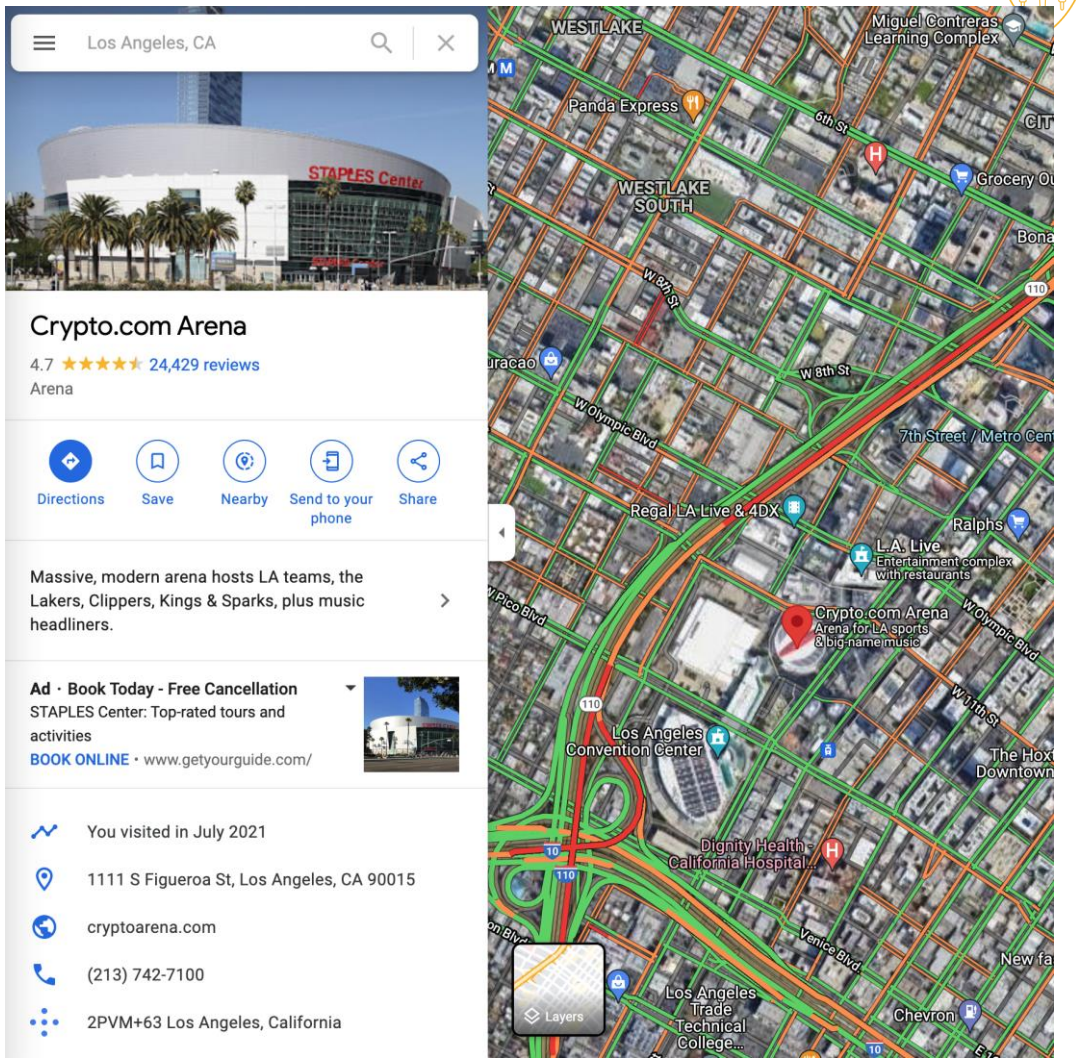
https://www.oracle.com/database/what-is-data-management/

What are Spatial Data?

Spatial Data

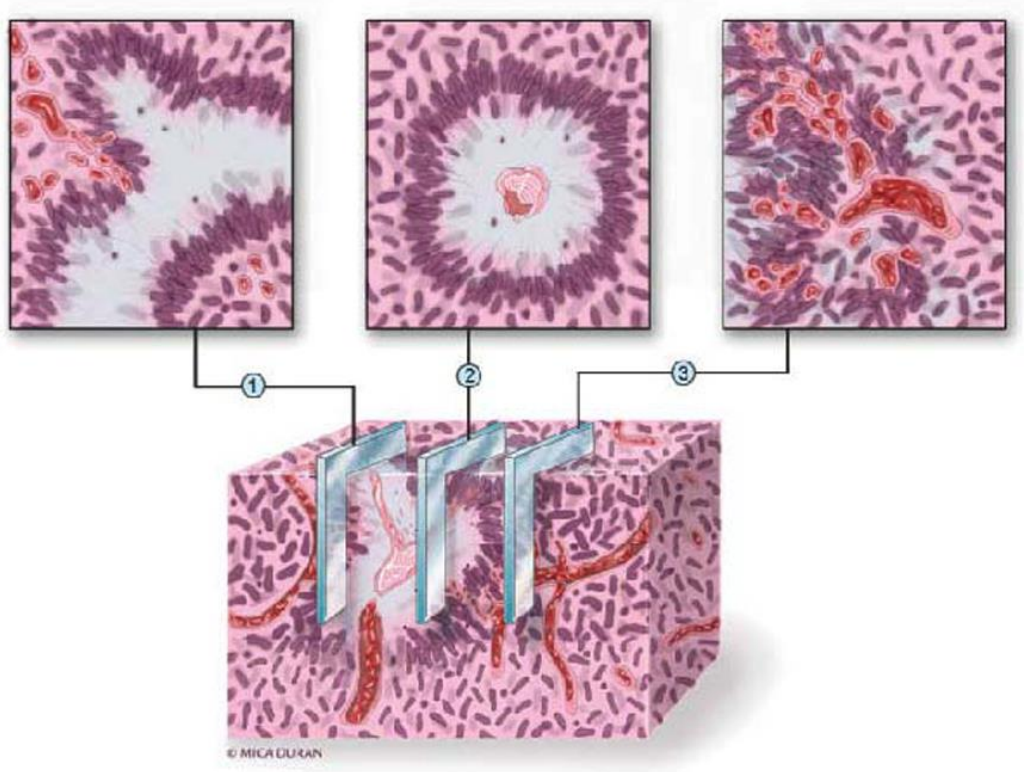
What is Spatial Data?

- Data that can be spatially referenced, e.g.,
  - Time series from fixed-site sensors (e.g., traffic, air quality)
  - Remotely sensed data (e.g., satellite imagery)
  - Geotagged photos and tweets
  - Documents mentioning location entities



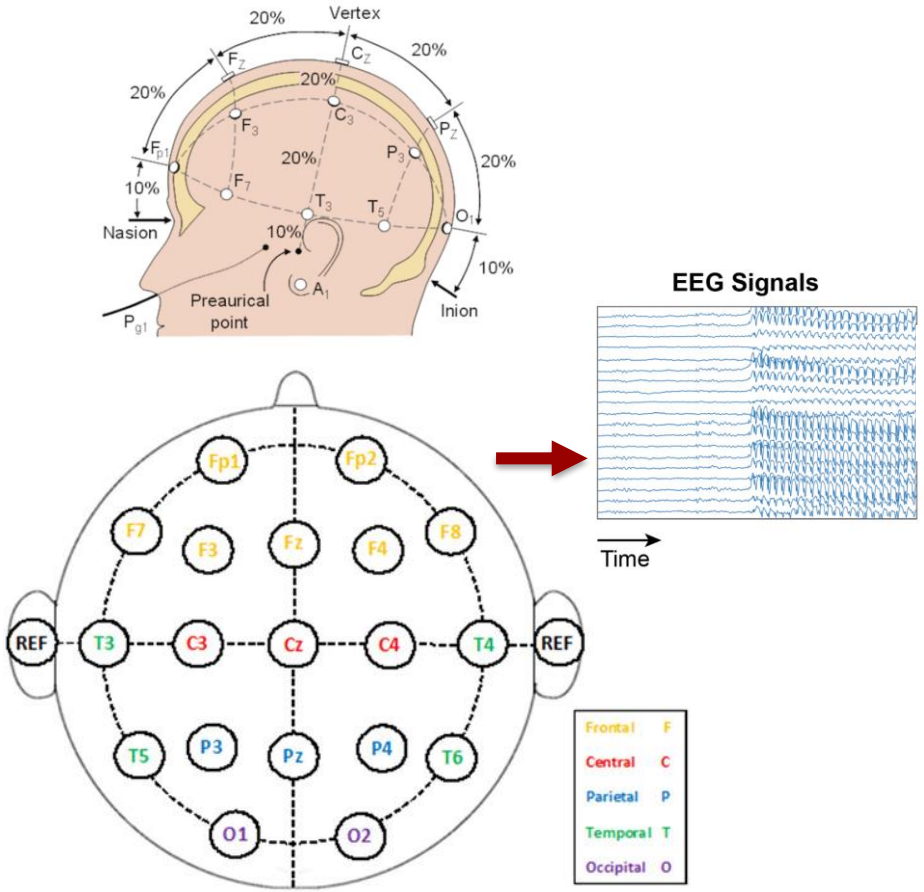
Spatial Data Do Not Have to be Geo Data

- **Digital Pathology**



Rong, Y., Durden, D. L., Van Meir, E. G., & Brat, D. J. (2006). 'Pseudopalisading'necrosis in glioblastoma: a familiar morphologic feature that links vascular pathology, hypoxia, and angiogenesis. *Journal of Neuropathology & Experimental Neurology*, *65*(6), 529-539.

- **EEG Data**



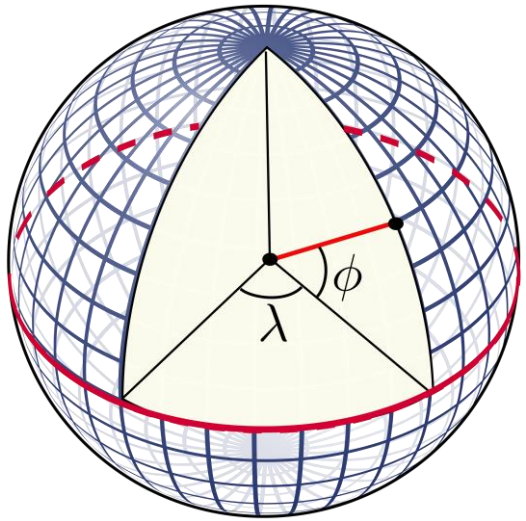
What does “Spatially Referenced” Mean?

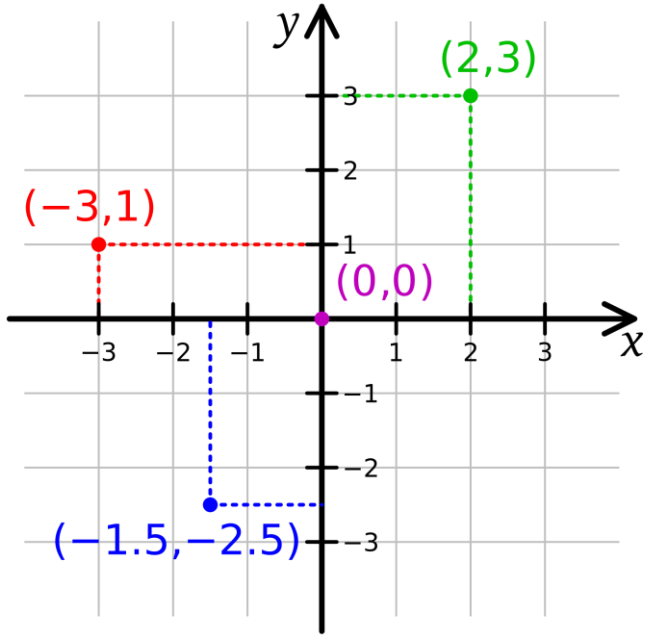
Spatial Coordinates

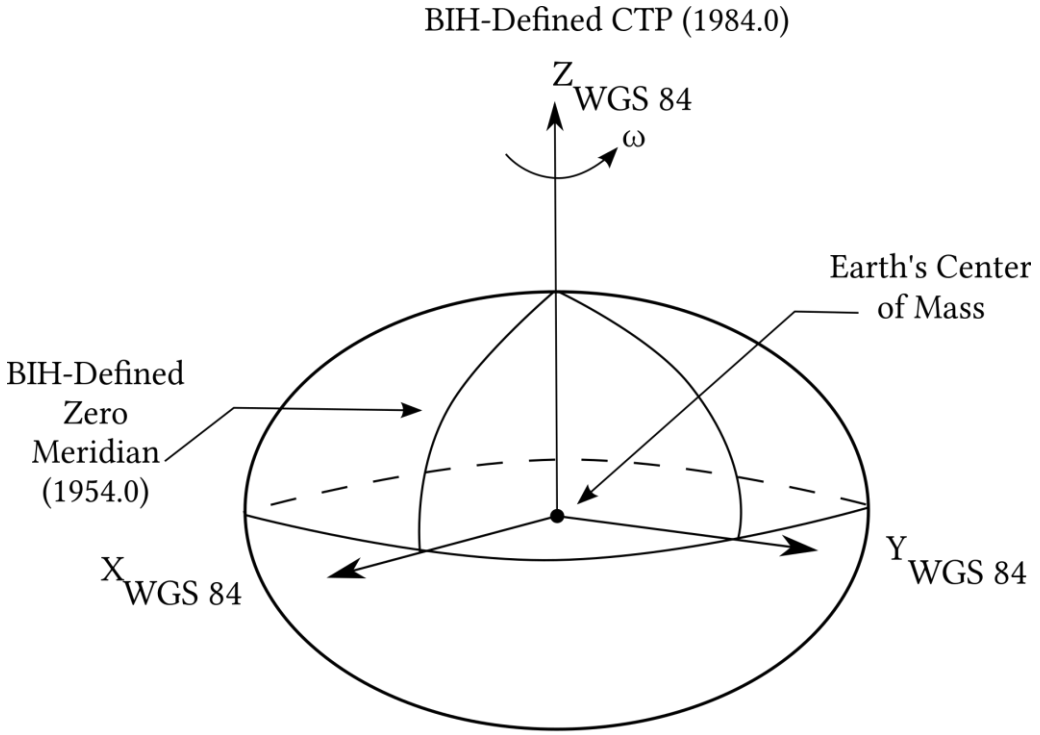
e.g., latitude and longitude, X and Y

Spatial Reference System

e.g., WGS84, Cartesian System

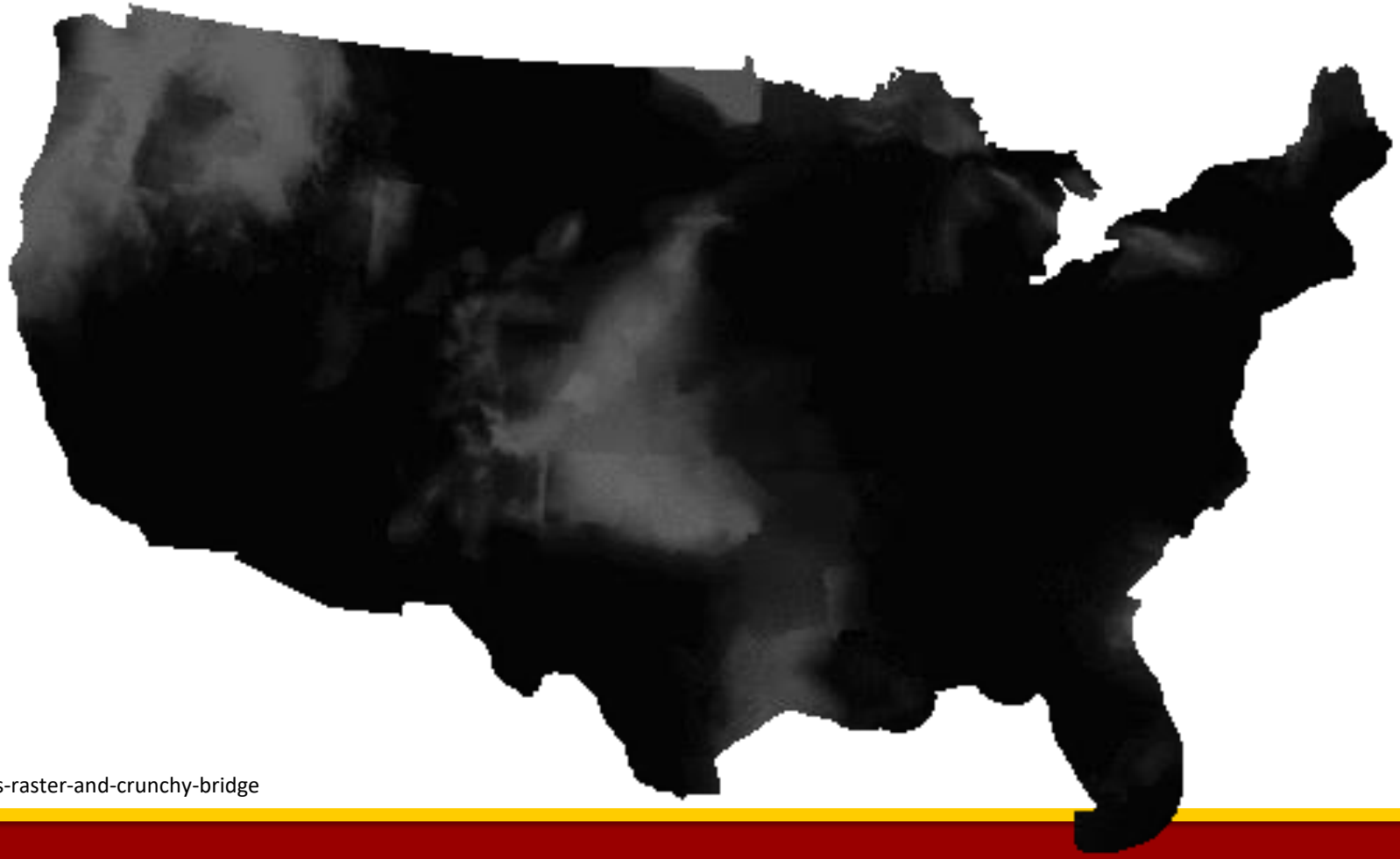






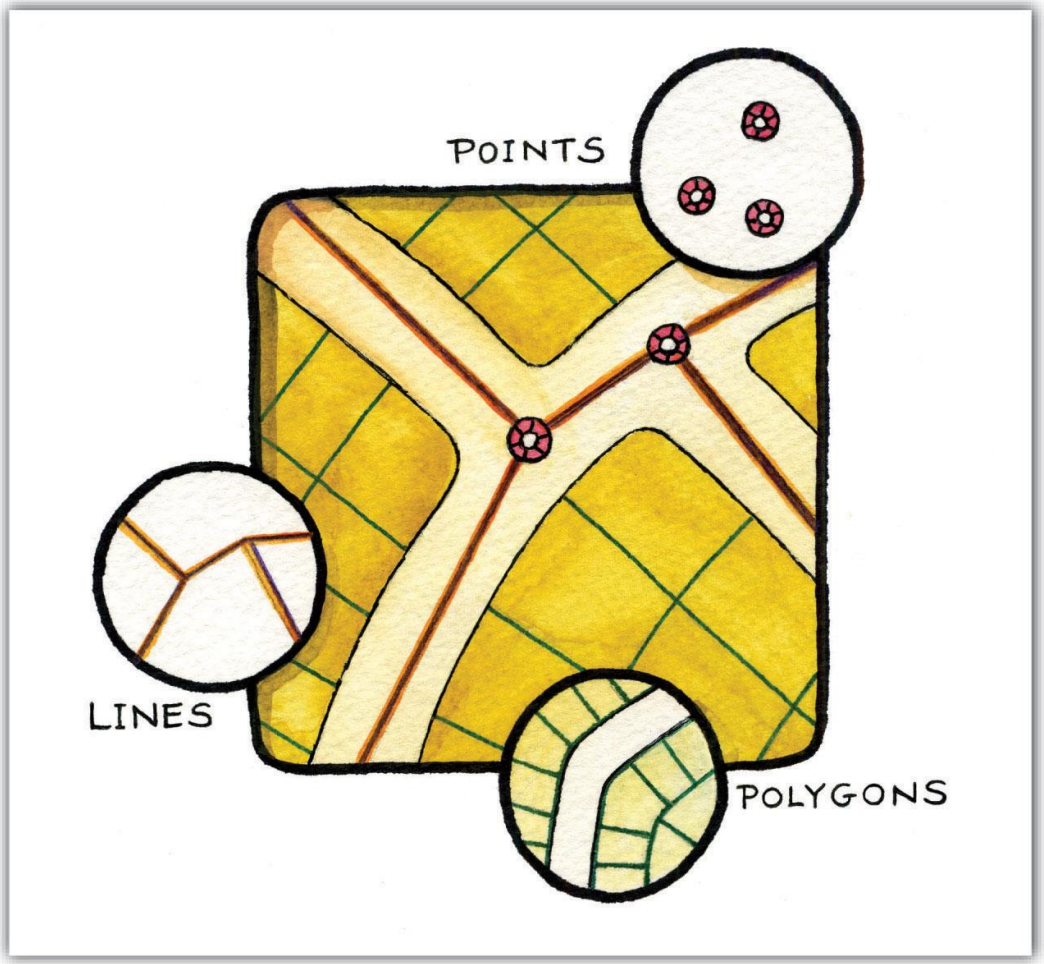
Spatial Data Representations – Raster Data

Probability of Precipitation



https://blog.crunchydata.com/blog/postgis-raster-and-crunchy-bridge

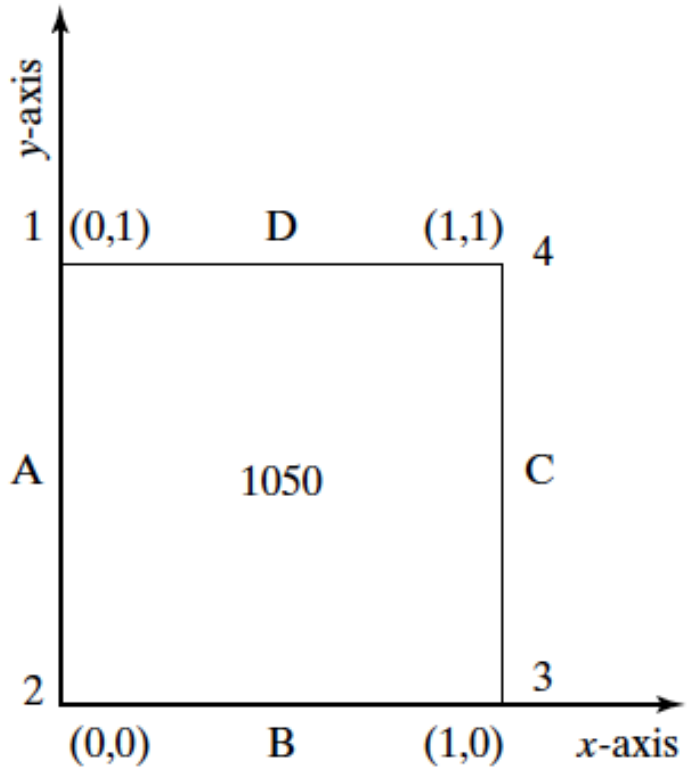
Spatial Data Representations – Vector Data



https://saylordotorg.github.io/text_essentials-of-geographic-information-systems/s08-02-vector-data-models.html

Spatial Data in Traditional Database Management Systems

- Use common Abstract Data Types, e.g., integer, string, floating points



Census_blocks

Name	Area	Population	boundary-ID
340	1	1839	1050

Polygon

boundary-ID	edge-name
1050	A
1050	B
1050	C
1050	D

Edge

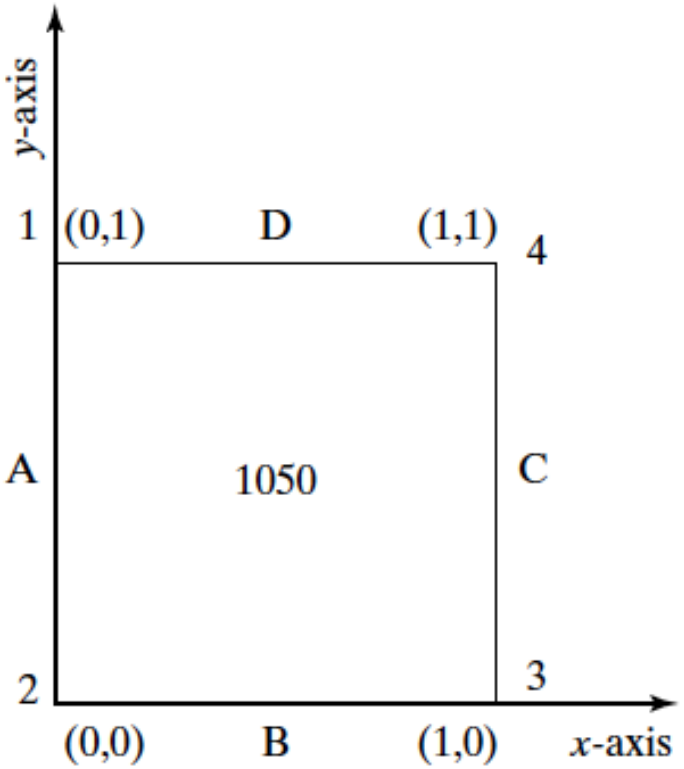
edge-name	endpoint
A	1
A	2
B	2
B	3
C	3
C	4
D	4
D	1

Point

endpoint	*x*-coor	*y*-coor
1	0	1
2	0	0
3	1	0
4	1	1

Spatial Data in Traditional Database Management Systems

- How about this? What is the data type of the boundary column? String?

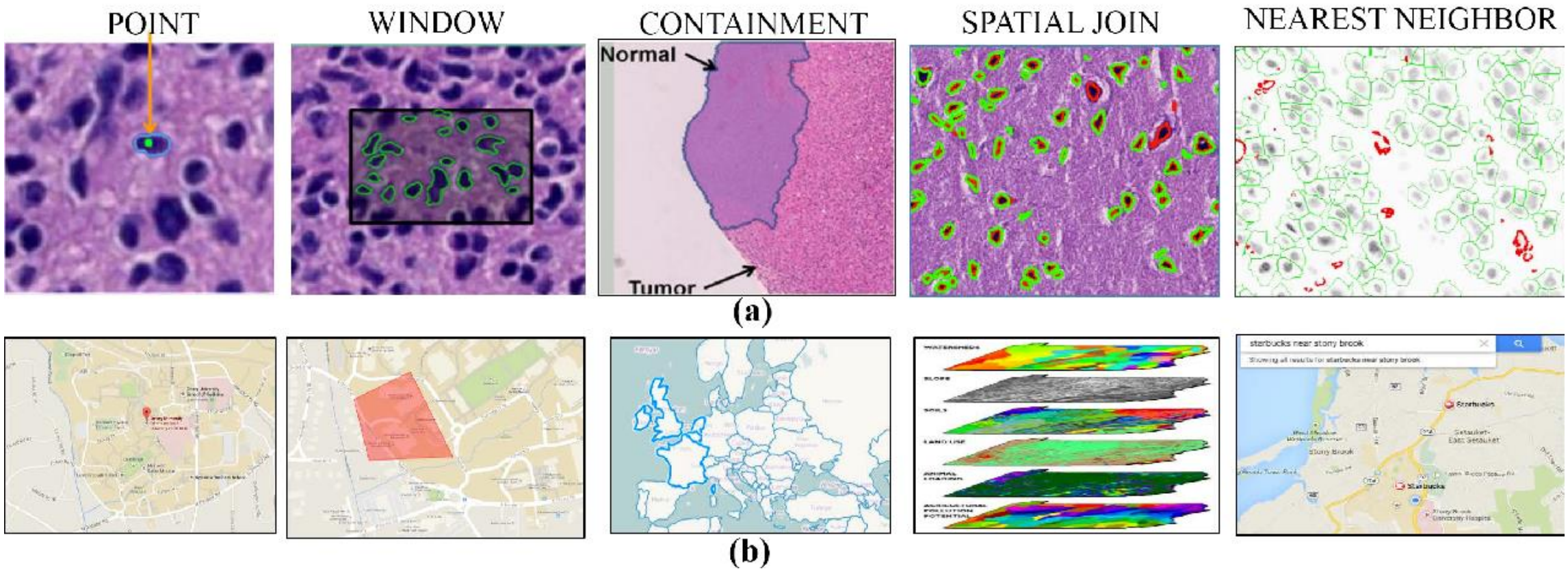


Census_blocks

Name	Area	Population	Boundary
1050	1	1839	Polyline((0,0),(0,1),(1,1),(1,0))

Why do We Need Spatial DBMS?

- Example of spatial query use cases in (a) pathology imaging; (b) GIS applications



Wang, Fusheng, Ablimit Aji, and Hoang Vo. "High performance spatial queries for spatial big data: from medical imaging to GIS." Sigspatial Special 6, no. 3 (2015): 11-18.

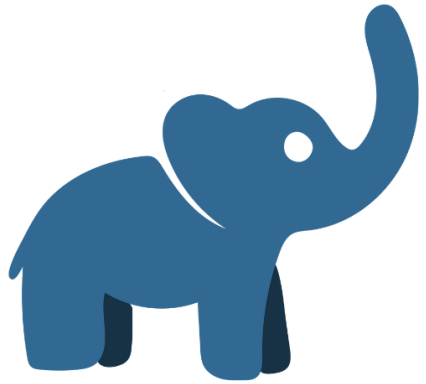
Spatial Data Management

Spatial Databases

- Support spatial data manipulations using SQL like languages
- Require a relational database engine
- e.g., PostGIS

Spatial Big Data Platforms

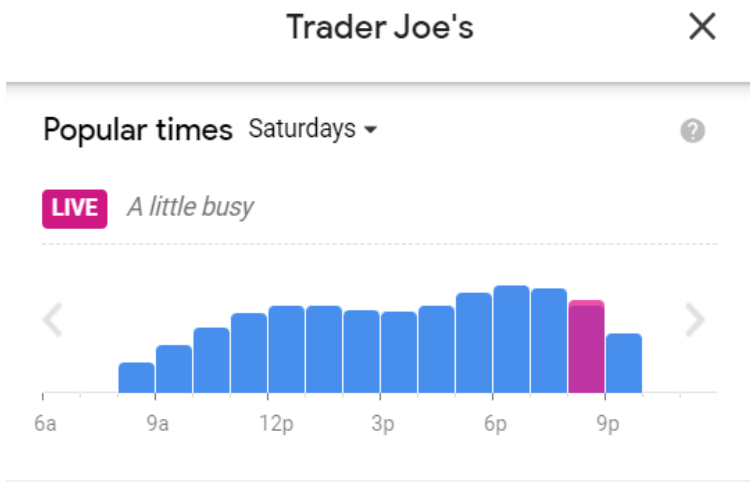
- Support highly parallelized spatial data manipulations
- Require a Big Data processing platform
- e.g., GeoMESA (MapReduce), Apache Sedona (based on Spark)

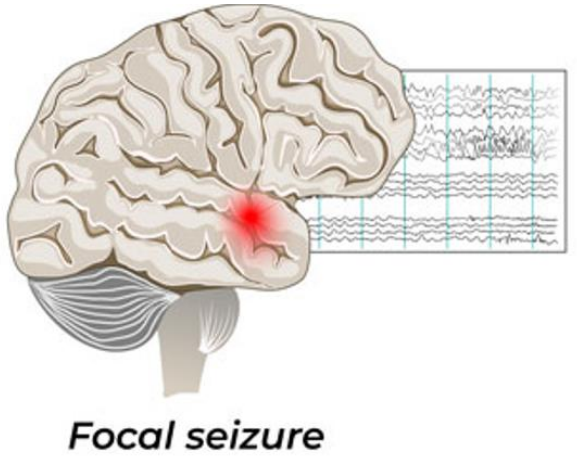


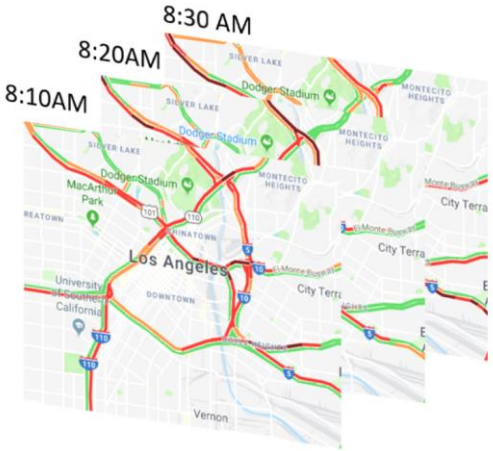
ML-based Applications using Spatial Data?

ML-based Applications using Spatial Data

- **Examples**
  - **POI Visit Forecasting:** Predicting the # of visits to specific POIs at different times
  - **Seizure Detection:** Predicting the occurrence of seizures in specific brain regions during particular time intervals.
  - **Traffic Forecasting:** Predicting traffic flow on roads during various times of day







This Course

Course Overview

- **Course Focus:**
  - Efficient storage, manipulation, and analysis of geospatial data
  - Introduction to real-world location-based applications and datasets
  - Materials will cover both traditional methods and modern ML-based approaches
- **Topics**
  - **Spatial Index Structures:** Quadtrees, K-d Trees, R-Trees, Voronoi indexes
  - **Spatial Queries:** k-NN, Reverse-NN, Skyline, Spatial Skyline
  - **Non-Euclidean Spaces:** Road networks, land surfaces analysis
  - **Geo-Spatial Applications:** Spatial crowdsourcing, geo-social networks, ride sharing
  - **Data Privacy:** Geospatial data privacy and security
  - **Machine Learning:** Clustering, classification, anomaly detection
  - **AI for Spatiotemporal Analysis:** Predictive analytics, traffic forecasting, dynamic routing

Course Prerequisites

- **Some Level of Familiarity with:**
  - **Data Modeling**: ER model, relational model, SQL, normalization
  - **Physical Data Design:** Storage devices, B+-tree, hash indexes
  - **Database Protocols:** Concurrency control, crash recovery
  - **Machine Learning:** Supervised, unsupervised, reinforcement learning
    - **CSCI-567 (ML):** Recommended, but not mandatory
  - **Programming Languages:** Proficiency in C/C++ or Python
    - Some understanding of deep learning frameworks like PyTorch is recommended

Grading and Course Work

Grading is as follows

- **Two midterms** (60%)
  - Midterm 1: 10/14 (30%)
    - Covers material up to Lecture 14
  - Midterm 2: 12/04 (30%)
    - Covers material from Lecture 15 onwards
- **Assignments:** (30%)
  - Deliverable 1: Due 9/29 (10%)
  - Deliverable 2: Due 10/28 (10%)
  - Deliverable 3: Due 12/3 (10%)
- **Participation:** (10%)

Assignments

- **Assignment 1:** Will be released on 09/04
  - **Topic:** Spatial Indexing and Simple Queries
  - **Main Tasks:**
    - Implement a spatial index using KD-Tree
    - Implement and analyze nearest neighbor and range queries.
    - Compare simple partitioning techniques with spatial indexing.
- **Assignment 2:** Will be released on 09/30
  - **Topic:** Advanced Spatial Queries and Non-Euclidean Spaces
  - **Main Tasks:**
    - Develop Reverse K-Nearest Neighbor queries.
    - Process queries within road networks and other complex spaces.

Assignments (cont'd)

- **Assignment 3:** Will be released on 10/30
  - **Topic:** Traffic Forecasting with Graph Neural Network Models
  - **Main Tasks:**
    - Implement a GNN forecasting model.
    - Experiment with different graph constructions and compare outcomes.

Logistics

- **Platform:** Released and submitted via Brightspace and course website.
- **Late Policy:** 5% deduction per day late, with a maximum of 3 days, no exception
- **Collaboration:** Individual work unless stated otherwise.
- **Grading Criteria:** Evaluated based on the correctness of results, clarity of presentation, and quality of code (details in assignment descriptions).

Course Staff and Office Hours

- **Instructor:** Professor Cyrus Shahabi
  - **Email:** shahabi@usc.edu
  - **Office Hours:** Mon-Wed, 4-5 PM, PHE-306a.

Teaching Assistants:

- **Maria Despoina Siampou**
  - **Email:** siampou@usc.edu
  - **Office Hours:** TBD
- **Arash Hajisafi**
  - **Email: hajisafi@usc.edu**
  - **Office Hours: TBD**

Communications

- **Course Website:** infolab.usc.edu/teaching/csci587/
  - **Material distributions:** e.g., lecture slides, reading materials
  - **Lectures schedule**
- **Assignments:** Through Brightspace + course website
- **Questions/Requests:** Email

Readings

- **Textbook:** Foundations of Multidimensional and Metric Data Structures by Hanan Samet.
  - Not required, but recommended.
  - A 30% coupon is available at https://www.cs.umd.edu/~hjs/.
- **Additional Readings:** Published Papers **(strongly recommended)**
  - A list of reading material is available at the course schedule through the website: infolab.usc.edu/teaching/csci587/
  - Readings are available through infolab.usc.edu/teaching/csci587_syllabus/
  - All USC students have automatic access to these digital archives.
  - Strongly recommended for the exams.

Q&A