Big Data Spatiotemporal Analytics
-Trends, Characteristics and Applications

Sangmi Lee Pallickara

Computer Science Department
Colorado State University

September 26, 2019

Geospatial Data and Analytics

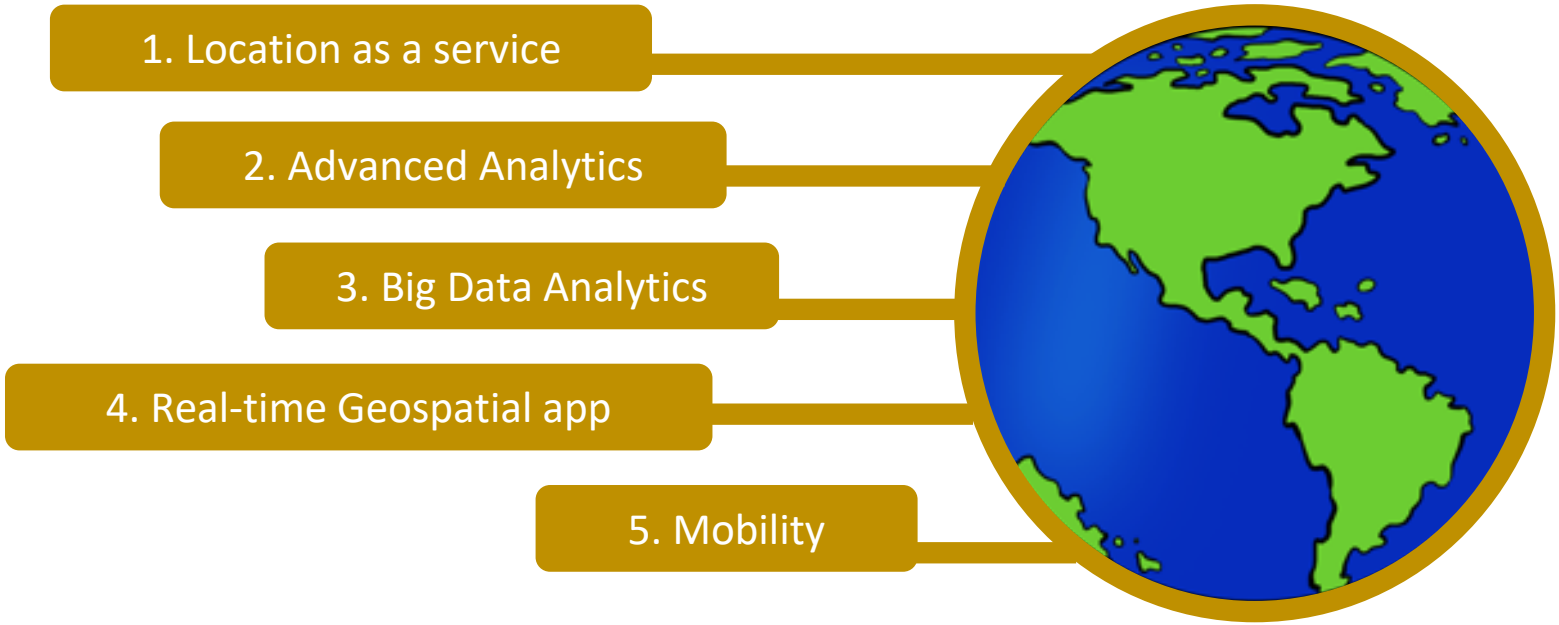
Geospatial Data and Analytics

- Spatial data comprises relative geographic information about the earth and its features
  - A specific location on earth.

- Geospatial analytics uses geographic data to identify relevant information
  - Referenced to geography and time

Five GIS Trends Changing The World

-Jack Dangermond, President of Esri



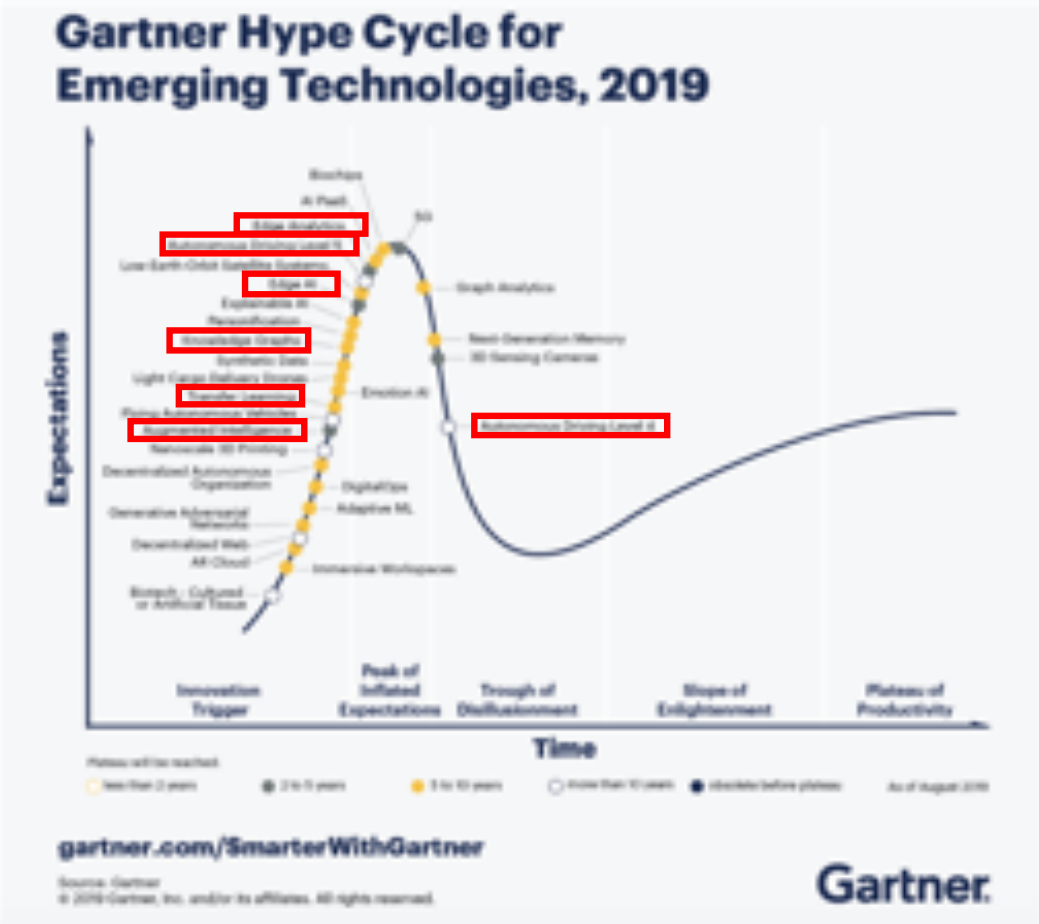
Location as a Service (LaaS) and Mobility

- In 2018, 242 million users were accessing location-based services on their mobile devices
  - More than two-fold increase over 2013

Source: https://www.reuters.com/brandfeatures/venture-capital/article?id=83809

Advanced Analytics and Big Data

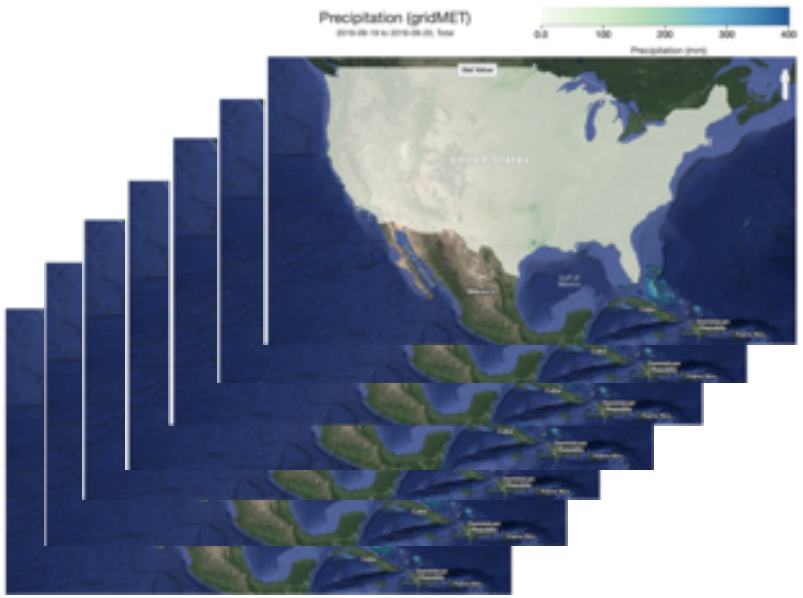
- Edge AI
- Autonomous Driving
- Edge analysis
- Transfer Learning
- Knowledge Graph

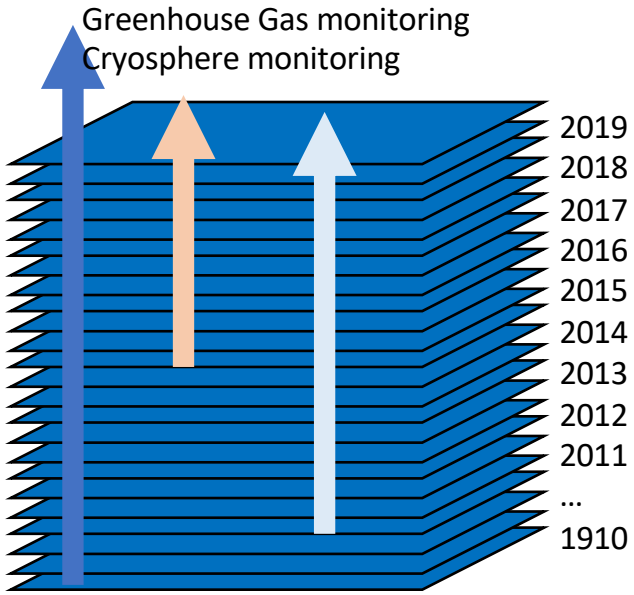


Characteristics and Challenges of Geospatial Data and Analytics

Impedance Mismatch

How to collect/store observations VS. how to access during analytics

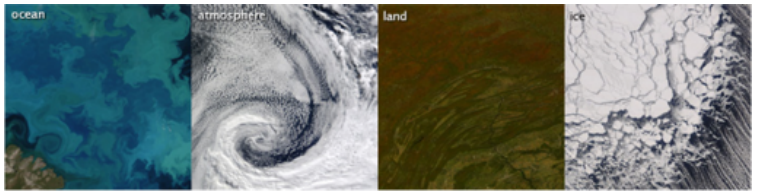




Spatial & Temporal Proximity: Finding what's nearby

- *"Everything is related to everything else, but near things are more related than distant things." —W. R. Tobler*
- Correlated objects such as road system and surrounding area
- Storage and retrieval of data should cope with this specific access style
  - Reduces query throughput
- Finding a needle in a haystack?

Area of Interest
- Finding what's inside



- NASA's MODIS instruments scans the entire surface of the Earth every two days (daily in northern latitude)
- Sentinel-2 earth observation scans the entire Earth every 5 days

- However, interests are not evenly distributed
  - E.g. national disaster, political activity, sports events, etc.
- This is directly related to data access patterns, workload management, and resource organization

Applications and Approach

Colorado State University

Application 1

Big Geo Data on the Street

- **Galileo**: Managing multidimensional time series data
- **Columbus**: Long-running Workflow Engine
- **Confluence**: Realtime Geospatial Data Integration

Colorado State University

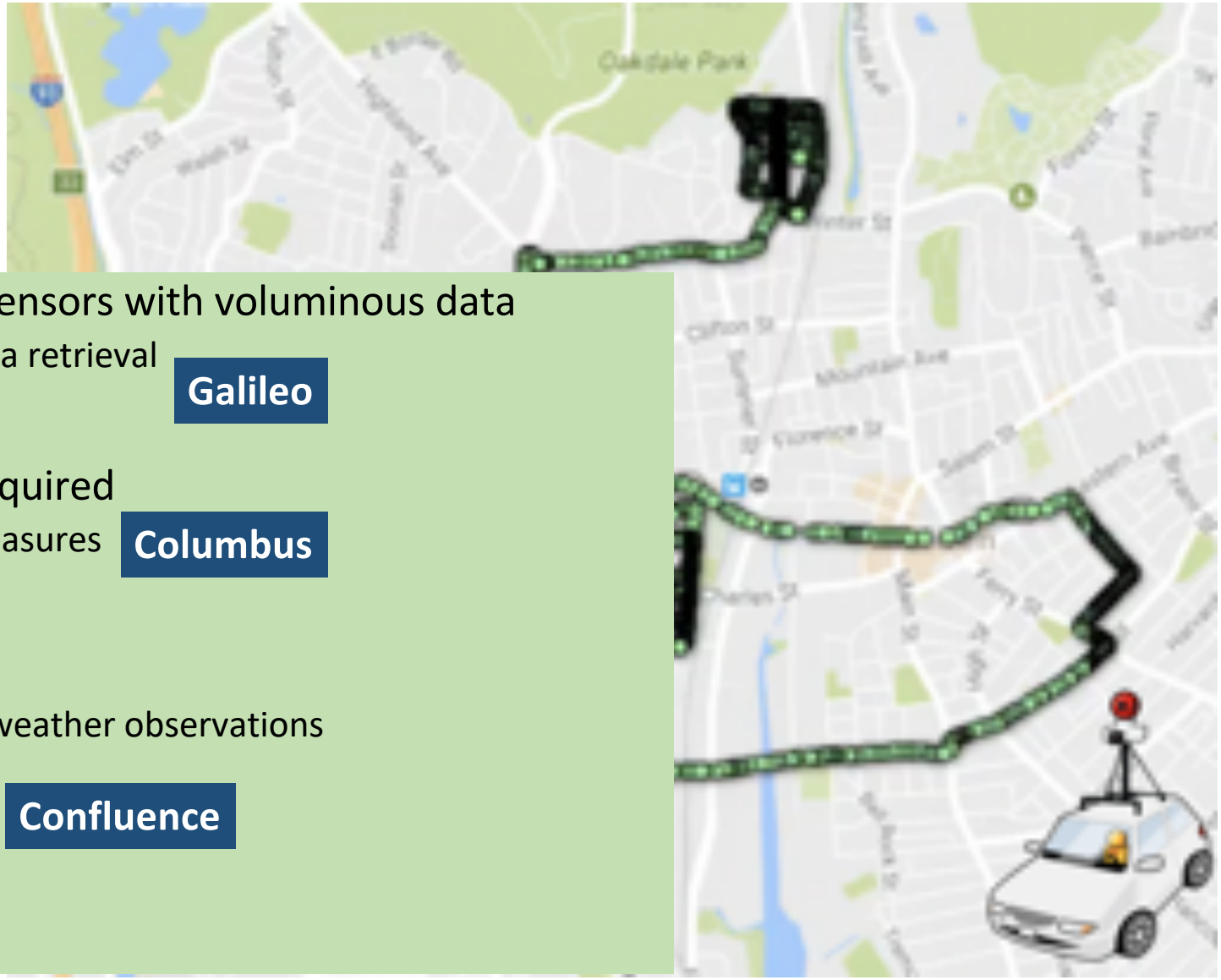
Detecting Natural Gas Leakage

- Environmental Defense Fund, NSF, Google Inc., and Colorado State University

- Google's Street View cars collect the required information
  - car id, car speed, date, time, locality, postal code, cavity pressure, cavity temperature, ch4, gps latitude, gps longitude, wind north, wind east, …

- Methane Gas in urban areas
  - Leakage from the natural gas distribution network

Challenges



- High frequency mobile sensors with voluminous data
  - Scalable storage and data retrieval **Galileo**
- Long-running analysis required
  - 1 month of repeated measures **Columbus**
- Data integration
  - Mobile sensor data VS. weather observations
  - Mismatched locations **Confluence**

Application 1

Big Geo Data on the Street

- **Galileo**: Managing multidimensional time series data
- **Columbus**: Long-running Workflow Engine
- **Confluence**: Realtime Geospatial Data Integration

Colorado State University

Distributed Storage for Multidimensional Geospatial Data

- Data is voluminous
  - Outpaces what is available on a single hard drive
- Storage must be over a collection of machines
  - Avoid central coordinators
  - Cope with failures
  - Preserve data locality without introducing storage imbalances
    - And the accompanying query hotspots
  - Support rich queries and fast ingestion of new data

Galileo: Notable Features

http://galileo.cs.colostate.edu

- High throughput storage and retrieval of observations
  - Support for a large number (~10^{10}) of small files
  - Petascale datasets
- Data: Spatiotemporal and multidimensional with multiple *types*
- Autonomous reconfiguration of data structures
- Query support: Range, geometry and proximity constrained, continuous, approximate & analytical queries

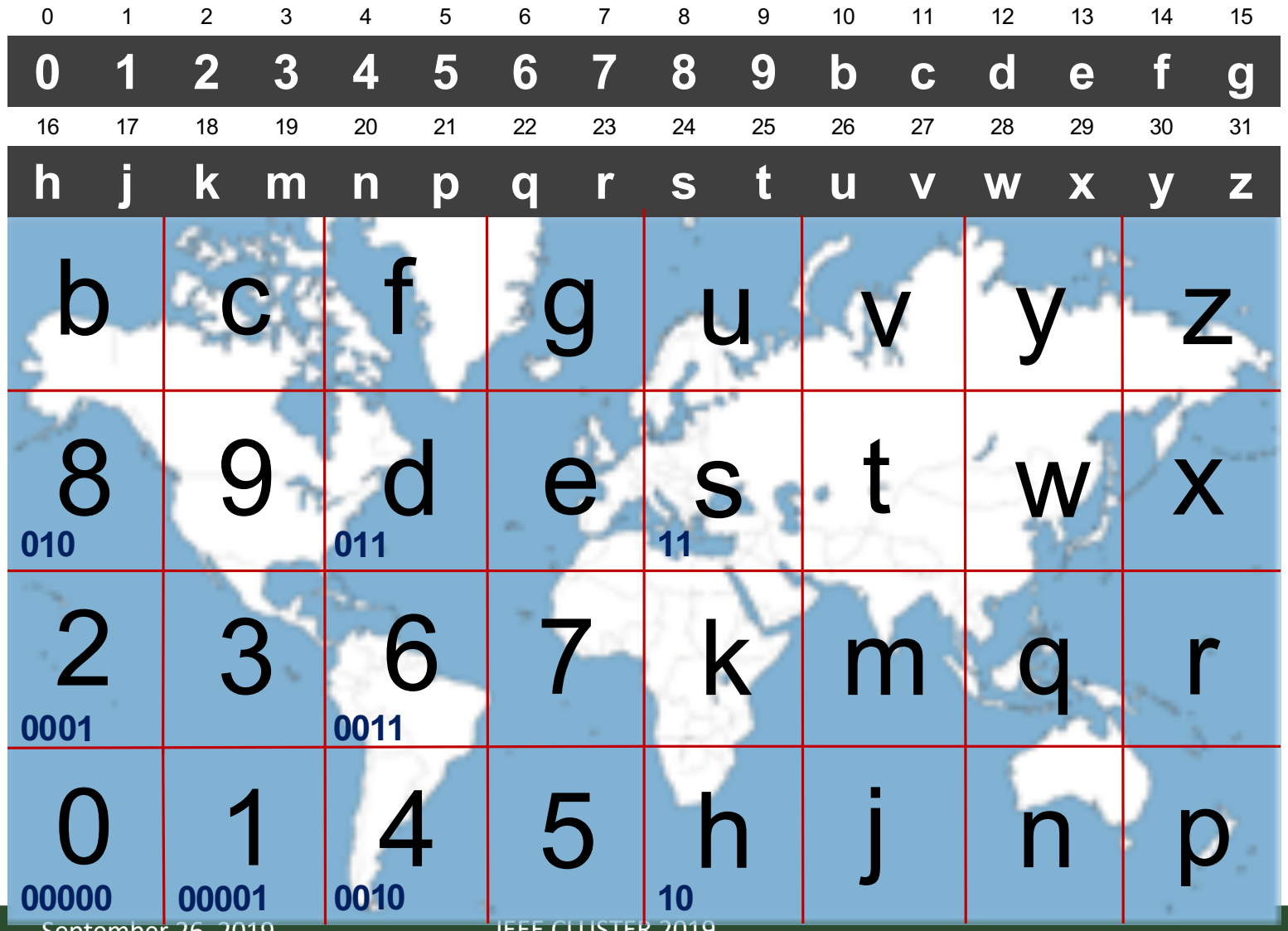
M. Malensek, S. L. Pallickara, and S. Pallickara. Fast, Ad Hoc Query Evaluations over Multidimensional Geospatial Datasets. *IEEE Transactions on Cloud Computing*. Vol. 5(1) pp 28-42. 2017.

M. Malensek, S. L. Pallickara, and S. Pallickara. Analytic Queries over Geospatial Time-Series Data using Distributed Hash Tables. *IEEE Transactions on Knowledge and Data Engineering*. Vol 28(6) pp 1408-1422. 2016.

Data Dispersion Scheme

- Geohash
  - Encodes latitude/longitude as strings representing a bounding box
  - Has found wide use in storing point data in databases (e.g. MongoDB)
- Represents an area
  - Base 32 encoding
    - Subdivides into 32 grids
    - 5 bits per character
      - $10001_2 = 9$
  - Z-order curve

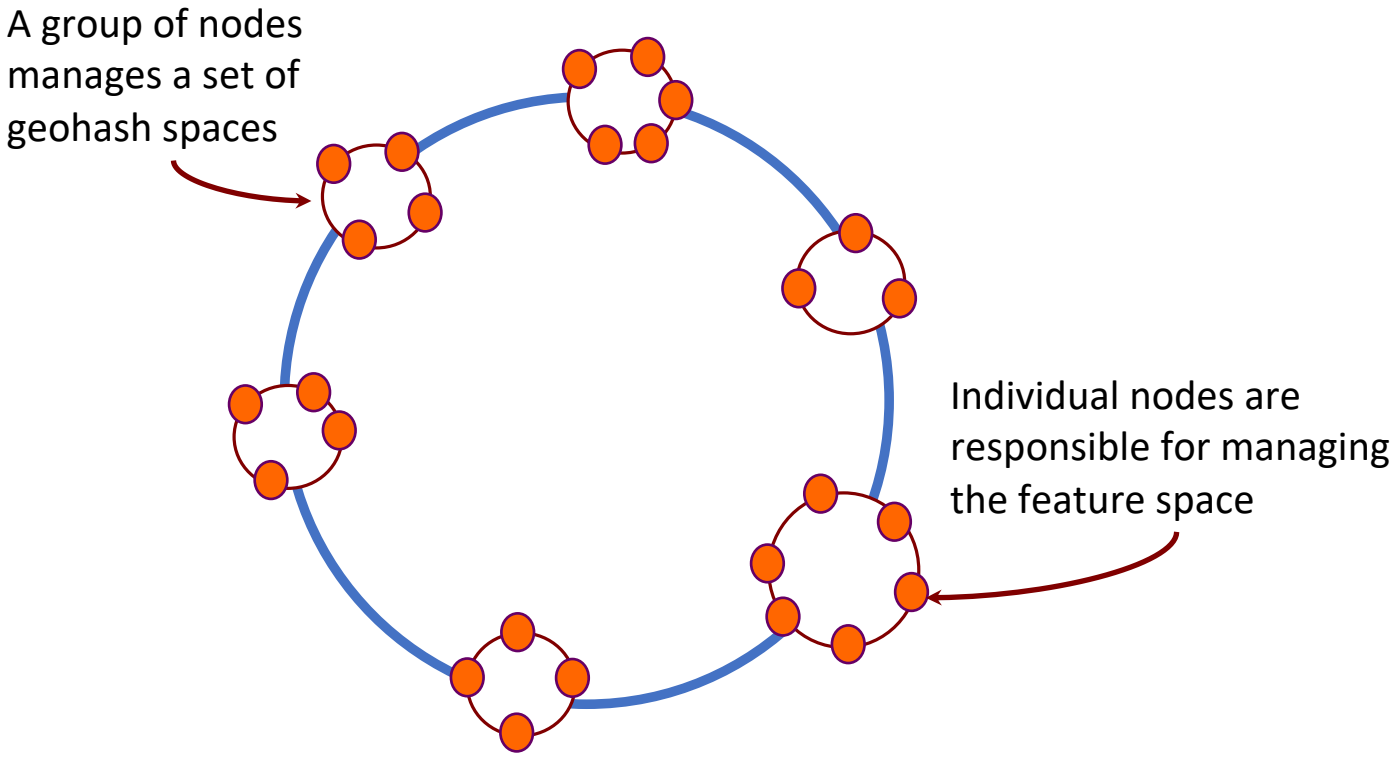
Geohash	Area
9	3110 x 3110 miles^2
9x	777 x 388 miles^2
…	…
9xjqbf2d	38.2 x 19.1 metres^2
…	…
9xjqbf2d7fp	14.9cm x 14.9cm


0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0 1 2 3 4 5 6 7 8 9 b c d e f g
16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
h j k m n p q r s t u v w x y z
b c f g u v y z
8 9 d e s t w x
010 011 11
2 3 6 7 k m q r
0001 0011
0 1 4 5 h j n p
00000 00001 0010 10

Controlled Dispersion

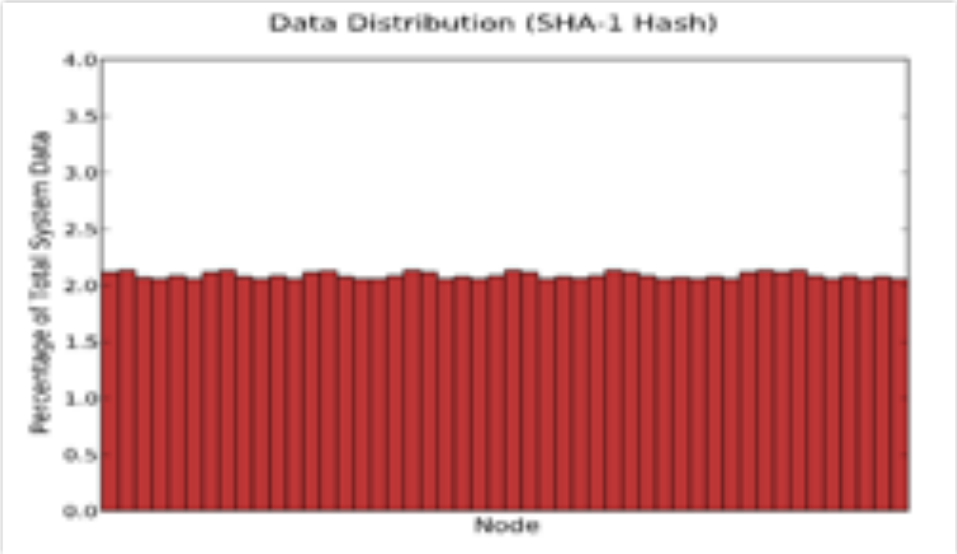
- Makes partitioning a two-stage process:
  1. Place data points into groups based on logical partitions
  2. Generate a standard hash to determine the final location of the data point within its group
- Effectively creates a two-tier DHT
- Can be expanded beyond just two grouping steps
- Balances control and load balancing

Nodes are organized within a ring of rings



Data Partitioning and Dispersion

- Sourced from NOAA NAM Project
- Some Dimensions/Features:
  - Geospatial: Latitude, Longitude
  - Time Series: Start Time, End Time
  - Temperature
  - Relative Humidity
  - Wind Speed
  - Snow Depth
- Composed of 20 billion files, ~1 PB

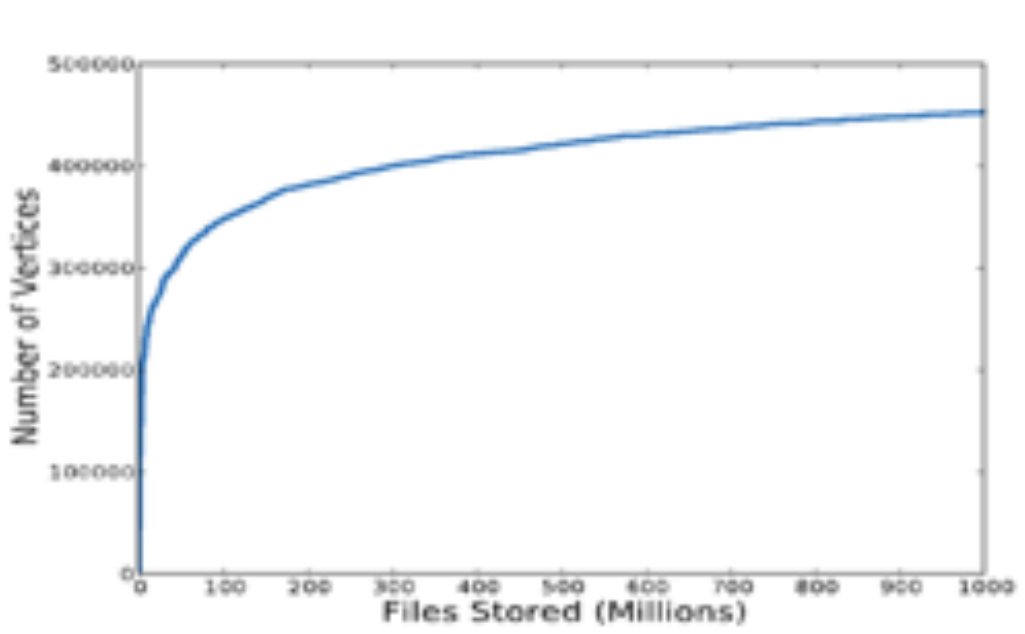


Matthew Malensek, Sangmi Lee Pallickara, and Shrideep Pallickara. Fast, Ad Hoc Query Evaluations over Multidimensional Geospatial Datasets. *IEEE Transactions on Cloud Computing*, Vol. 29(12) pp 1-16. 2017

Matthew Malensek, Sangmi Lee Pallickara, and Shrideep Pallickara. Analytic Queries over Geospatial Time-Series Data using Distributed Hash Tables. *IEEE Transactions on Knowledge and Data Engineering*. Vol. 28(6): pp.1408-1422. 2016

The Feature Graph

- Enables searching for specific feature attributes
  - Compensates for disadvantages of hashing based partitioning
- Each node in the system maintains a copy
- Updates are **gossiped** between nodes at regular intervals
  - Eventually consistent
- When data is inserted:
  - Features become vertices in the graph
  - Each vertex points to a collection of nodes with matching data



Application 1

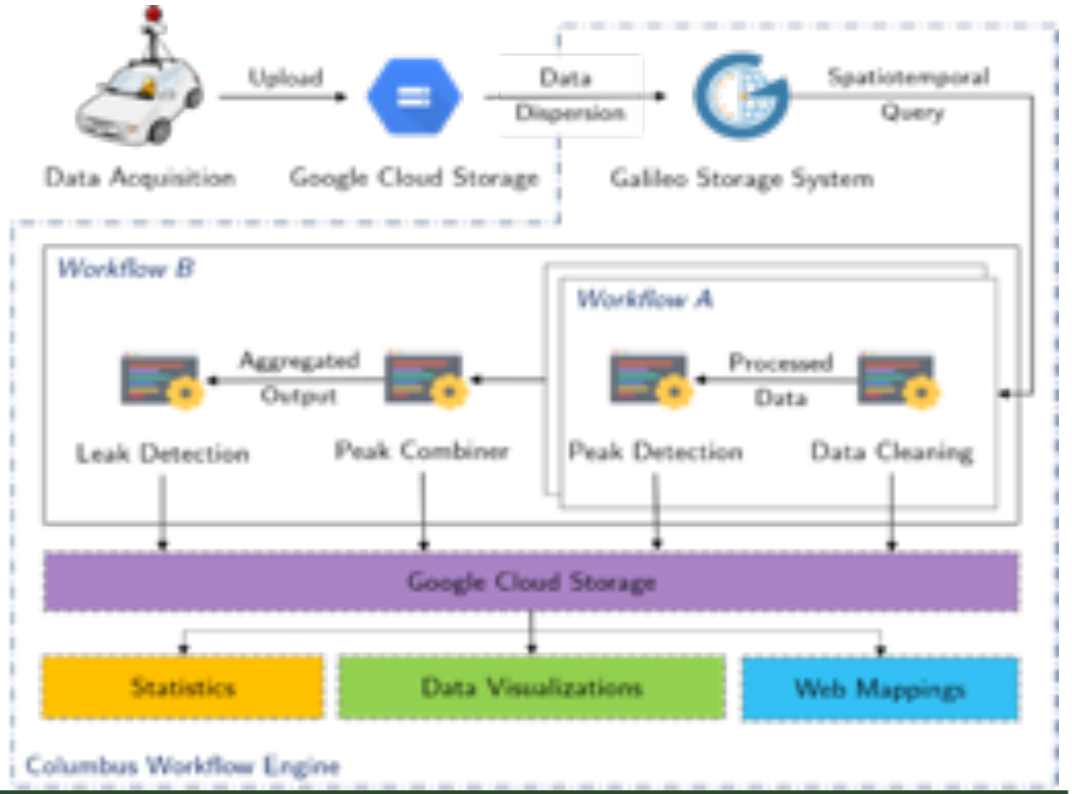
Big Geo Data on the Street

- **Galileo**: Managing multidimensional time series data
- **Columbus**: Long-running Workflow Engine
- **Confluence**: Realtime Geospatial Data Integration

Colorado State University

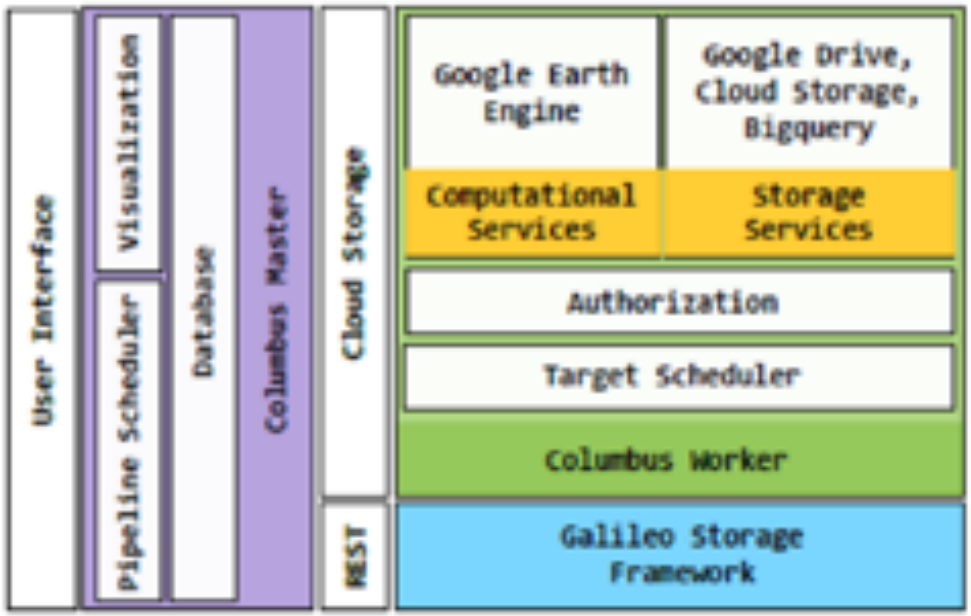
Collecting, organizing, monitoring and scheduling long-lasting analyses

- Local background concentration
  - Moving average of CH4 concentration over a short duration
  - Significantly high concentration?
    - > 10% or > 1 SD above the local average
- Long-running analysis
  - This process is repeated for several weeks
- Continuous monitoring to plan data collection
- Scalable analysis
  - Concurrent data collection and analysis in cities all over the US



System Architecture

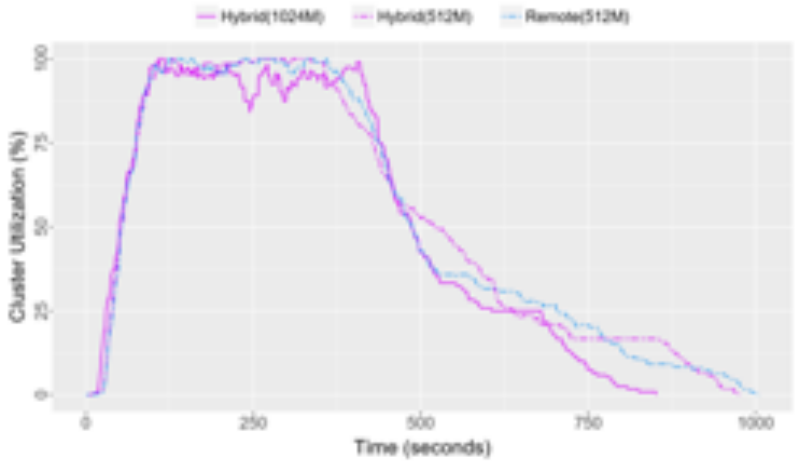
- Workflow
  - Directed Acyclic Graph (DAG) of targets
  - Pull-based data flow
- Weak Association
  - Lazy realization
- Three-tier job queue
  - Waiting queue
  - Ready queue
  - Target queue

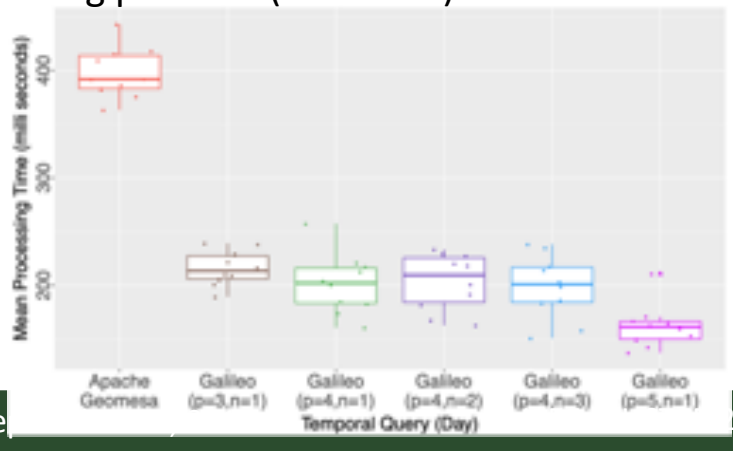


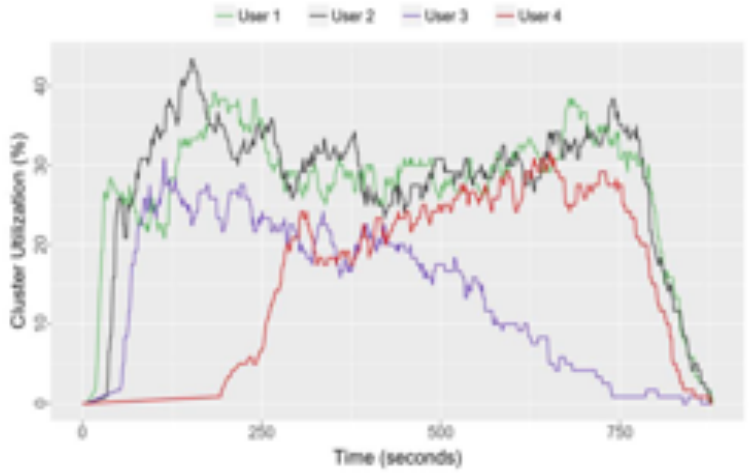
Johnson Charles Kachikaran Arulswamy, and Sangmi Lee Pallickara. Columbus: Enabling Scalable Scientific Workflows for Fast Evolving Spatio-Temporal Sensor Data.Proceedings of the *the 14th IEEE International Conference of Service Computing (IEEE SCC). pp.9-18. Honolulu, Hawaii, USA, 2017*

Locality aware workflow scheduling

- Local
  - Highest data locality by allocating targets to workers housing the data
- Remote
  - Workload-based scheduling
- Hybrid
  - Ratio of the number of workflows waiting to those running per user (WR Ratio)







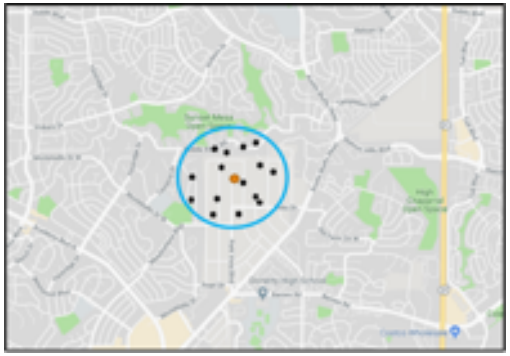
Application 1

Big Geo Data on the Street

- **Galileo**: Managing multidimensional time series data
- **Columbus**: Long-running Workflow Engine
- **Confluence**: Realtime Geospatial Data Integration

Colorado State University

Data Integration

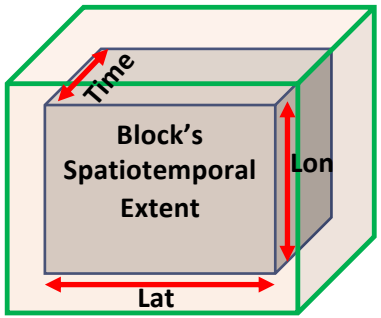


- Our methodology is designed for spatiotemporally distributed block-based storage systems
  - We use Galileo
- Target and source datasets
  - Target records act as the pivot
  - Source dataset data gets moved around, if necessary

Saptashwa Mitra and Sangmi Lee Pallickara, Confluence: Adaptive Spatiotemporal Data Integration Using Distributed Query Relaxation Over Heterogeneous Observational Datasets, *Proceedings of the IEEE/ACM Conference on Utility and Cloud Computing (UCC)*, Zurich, Switzerland 2018

Data Integration Query

- `STJoin (target, source, coverage_spatial, coverage_temporal, attributes, relaxation_spatial, relaxation_temporal, model)`
- Data Integration – finding matching pairs.
- Feature Interpolation for a one-to-one relationship between points
  - Creating a synthetic value from neighboring values
    - Using well-known mathematical interpolation techniques
      - Dynamic Interpolation Parameter Optimization (In our case, β for IDW).
    - Weighted Mean
  - Uncertainty Estimation
    - Using machine learning model to predict interpolation error
    - Weighted Standard Deviation



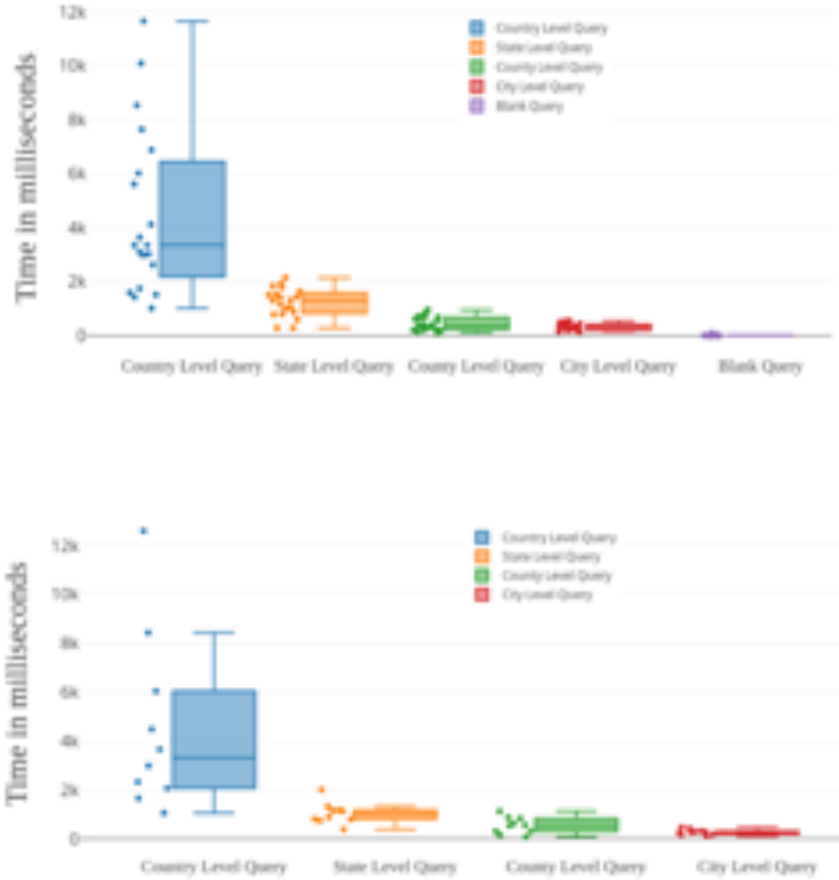
Interpolation with Uncertainty

- *Inverse Distance Weighting*
  - Neighbors influence interpolated value
  - Degree of influence dependent on closeness
  - Estimate: $$Z_j = \frac{\sum_{i=1}^{n} \frac{Z_i}{(h_{ij})^{\beta}}}{\sum_{i=1}^{n} \frac{1}{(h_{ij})^{\beta}}}$$
  - Uncertainty : $\boldsymbol{\beta}$ predicted by model
- *Attribute-based Uncertainty Estimation for Data Integration (AUEDIN)*
  - If degree of influence is not dependent on closeness
  - Weight of a neighbor observation based on some other feature in the record
  - Estimate: Weighted mean
  - Uncertainty : Weighted Standard Deviation

Saptashwa Mitra, Yu Qiu, Haley Moss, Kaigang Li, and Sangmi Lee Pallickara, Effective Integration of Geotagged, Ancillary Longitudinal Survey Datasets to Improve Adulthood Obesity Predictive Models, IEEE Big Data Science and Engineering 2018

Data integration query - latency

- HP Z420 with the configuration of 8-core Xeon E5-2560V2, 32 GB RAM and 1 TB disk.
- Cluster of 90 nodes
- 3 Groups of equal size
- Vector-to-Vector : NAM dataset and NOAA ISD dataset (~ 3.3 TB & ~50GB)
- Vector-to-Raster: Methane emission dataset and NOAA NOMADS Climate Data



U.S. DEPARTMENT OF ENERGY

arpa·e
CHANGING WHAT'S POSSIBLE

Application 2

Big Geo Data in the Field

Root Genetics in the Field to Understand Drought Adaptation and Carbon Sequestration

Radix: High-throughput In-Situ Georeferencing for Sensor data from Test Fields

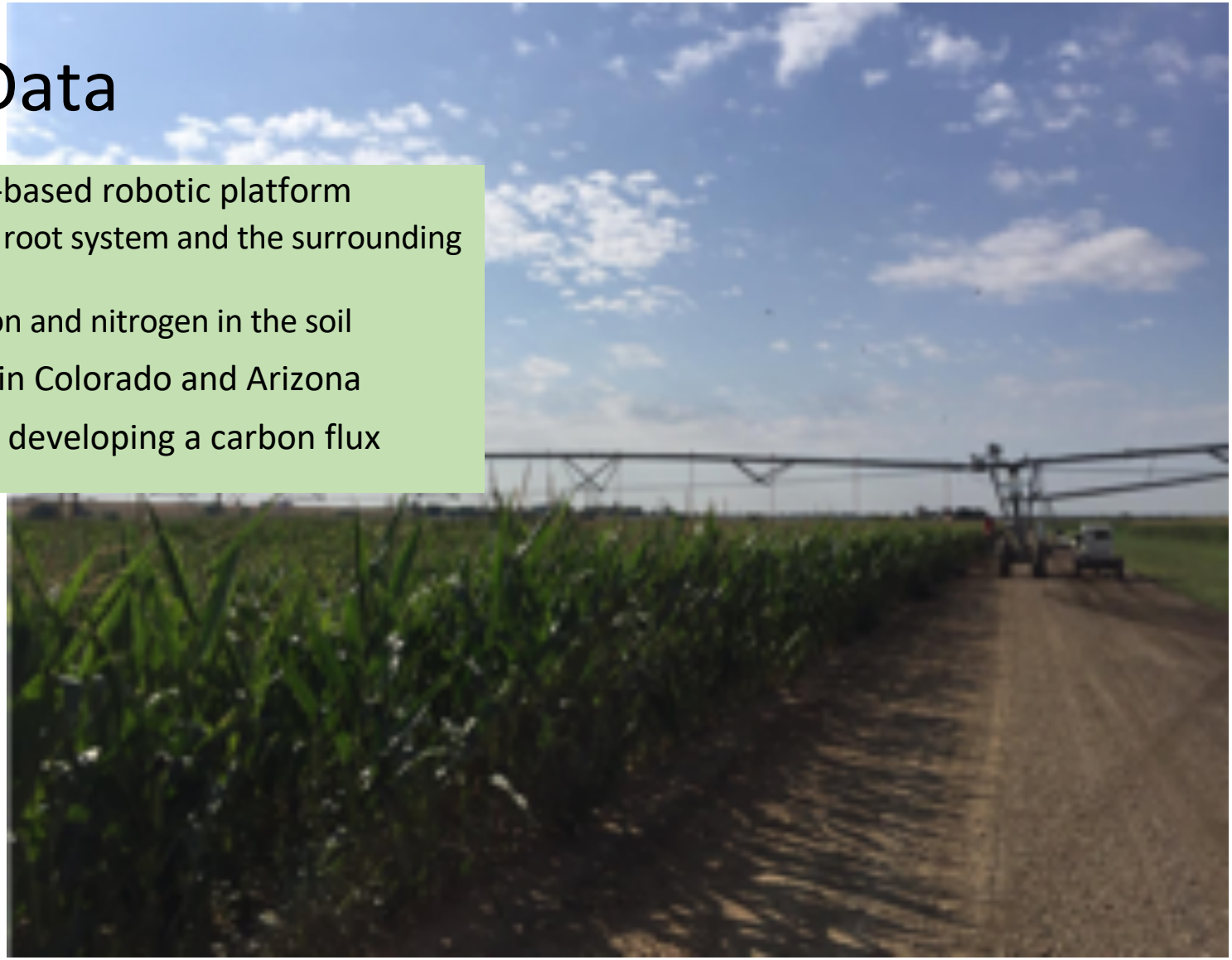
Stash: In-memory Distributed Cache for Visual Analytics

IEEE CLUSTER 10:00-10:30 AM September 26 (Today!)

Colorado State University

Phenotyping Data

- High-throughput ground-based robotic platform
  - Characterizes a plant's root system and the surrounding soil chemistry
  - How plants cycle carbon and nitrogen in the soil
- Studies at two field sites in Colorado and Arizona
- Public access to data and developing a carbon flux model



U.S. DEPARTMENT OF ENERGY

arpa·e

CHANGING WHAT'S POSSIBLE

Challenges

- Large volume of data with variety
  - 1,000s of genotypes with several combination of treatments
  - High-frequency data (250Hz)
  - Mapping to the plot and corresponding treatment
  - Multiple data sources
    - Mobile sensor array
    - UAV images
    - LiDAR observations
    - Robotic Platform sensors and images
  - **Galileo, Radix**
- Visual analytics over the daily observations
  - **Stash** (Talk: 10:30AM, Thursday)

U.S. DEPARTMENT OF ENERGY

arpa·e
CHANGING WHAT'S POSSIBLE

Application 2

Big Geo Data in the Field

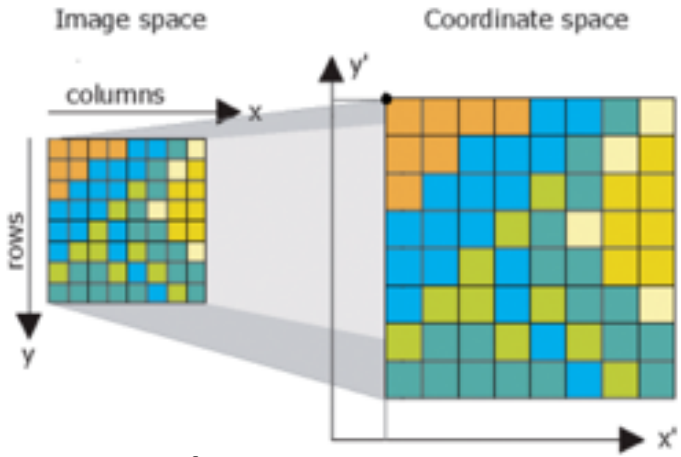
Root Genetics in the Field to Understand Drought Adaptation and Carbon Sequestration

Radix: High-throughput In-Situ Georeferencing for Sensor data from Test Fields

Stash: In-memory Distributed Cache for Visual Analytics

Colorado State University

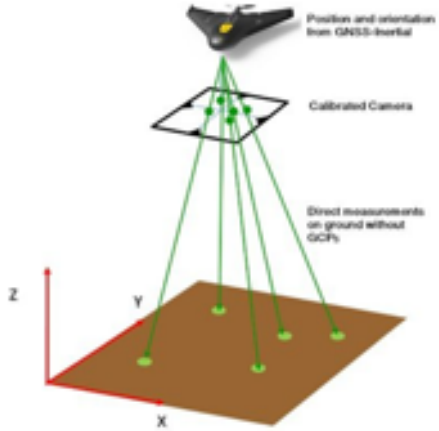
What is Georeferencing?



Raster to geolocation

- An unreferenced raster image consists of pixels
- Requires use of GCPs (Ground Control Points)

Image source: http://desktop.arcgis.com/en/arcmap/latest/manage-data/geodatabases/raster-basics.htm



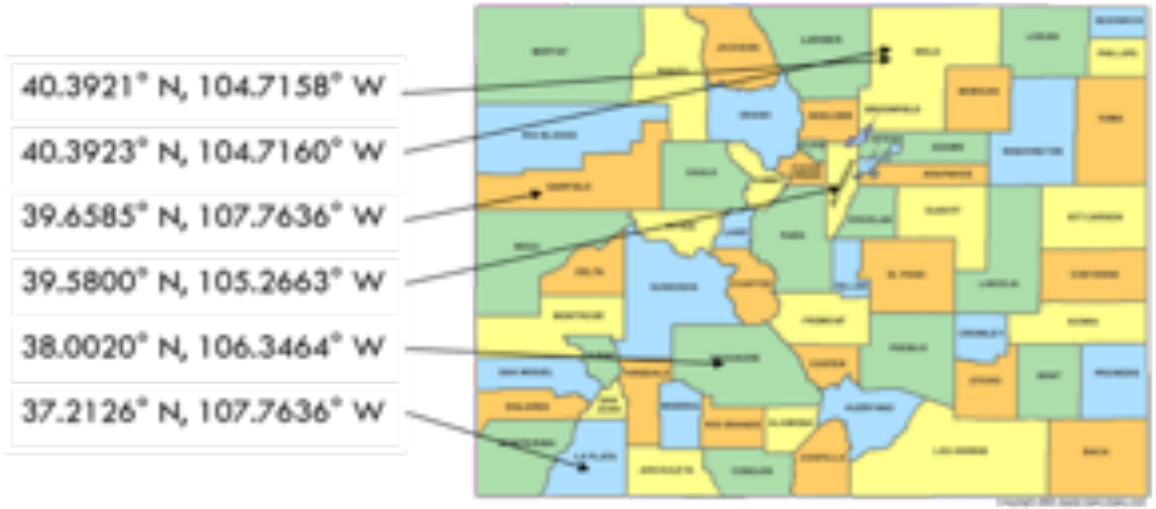
Direct Georeferencing

- GPS coordinates and a calibrated camera
- No ground control points during flight needed

Image source: http://www.uavexpertnews.com/2017/09/direct-georeferencing-on-unmanned-aerial-vehicles/

Reverse geocoding

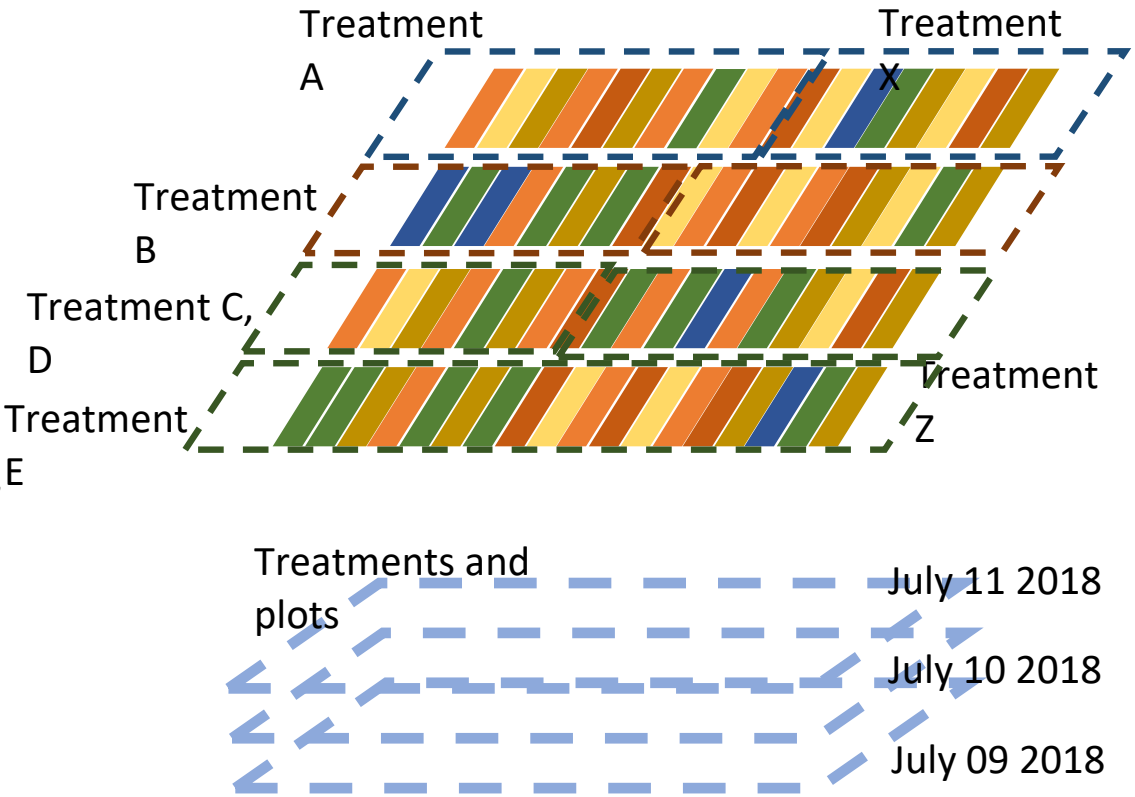
- Assigning geolocations to existing shapes or boundaries
- Requires pre-defined shape files or polygon definitions
- Very expensive for large volumes of data



Data Model for Phenotyping Data

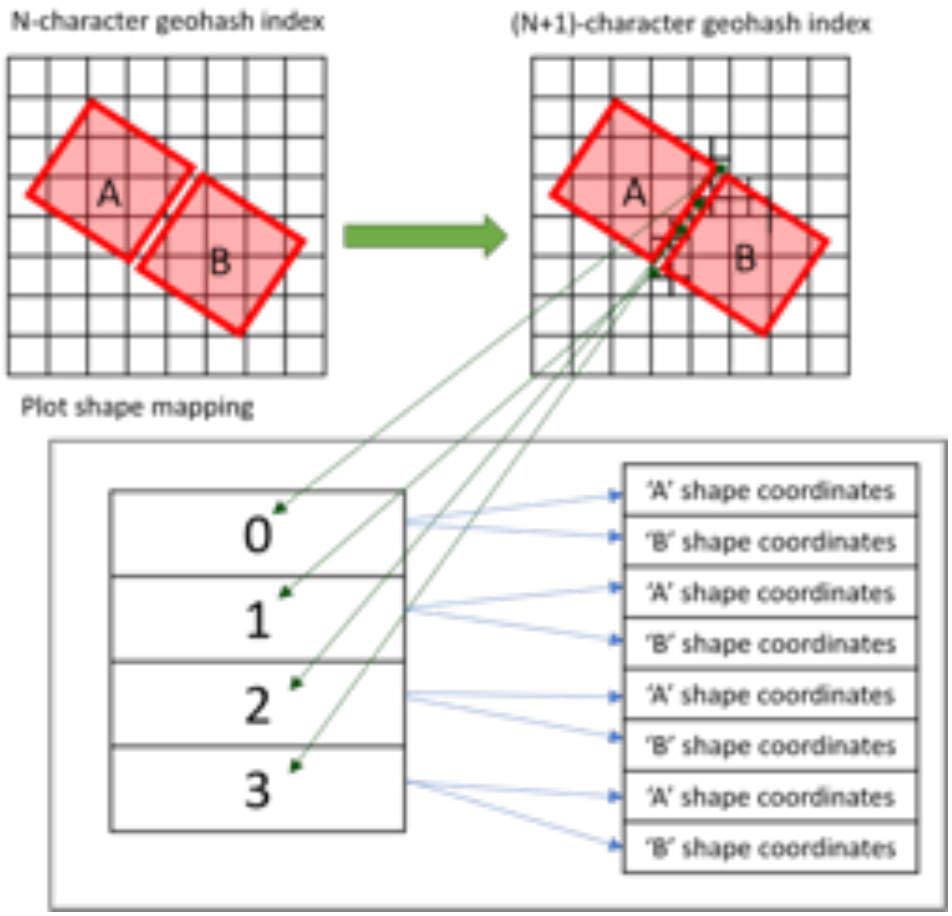
- "Plot" based data model
  - Treatment(s) and genotype(s)
  - Temporal and phenotyping attributes

- Existing solution
  - Geospatial queries in databases
  - Creates a point object
  - Creates a polygon object
  - Examine whether the point is "inside" the polygon object



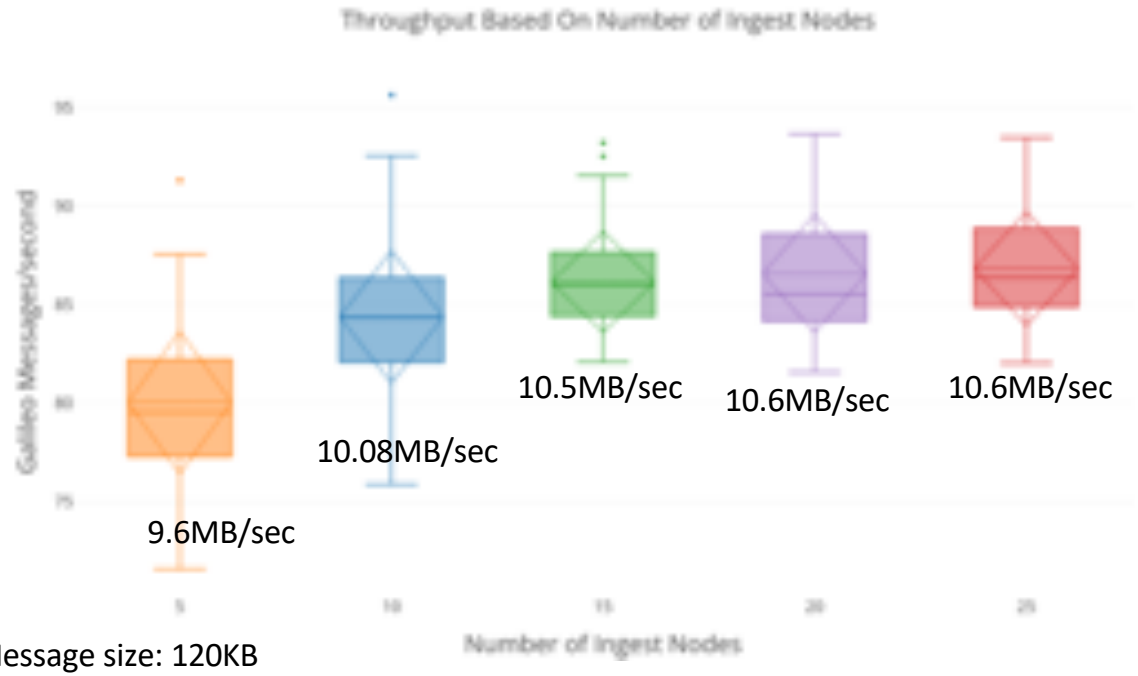
Nested Hashgrid

- First tier bitmap represents the coarser-grained geohash precision N
- For each index that intersects multiple shapes, break down that index into geohash precision of N+1
- If indices in the N+1 precision bitmap index still contain multiple collisions, store the raw shape for highest possible precision

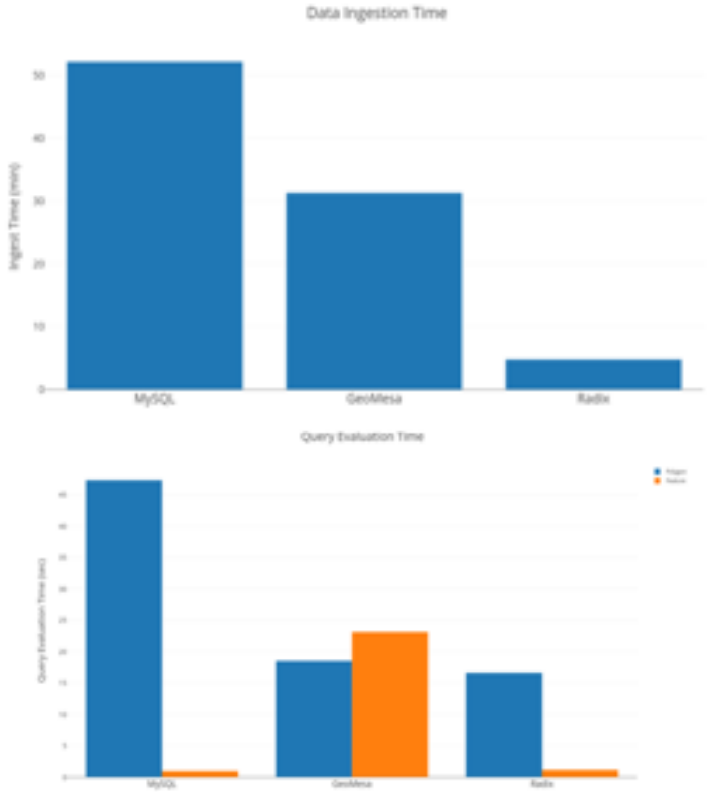


Max Roselius and Sangmi Lee Pallickara, Enabling High-throughput Georeferencing for Phenotype Monitoring over Voluminous Observational Data, *Proceedings of the IEEE International Conference on Big Data and Cloud Computing (BDCloud2018)*, Melbourne, Australia, 2018.

Data ingestion and Query evaluation



Message size: 120KB



Conclusions

- Geospatial attributes of sensor data provides fundamental capabilities of monitoring and reasoning for geosciences
- To achieve high-throughput interactive data retrieval over voluminous datasets, spatiotemporal proximity must be preserved for data dispersion
- Distributed data integration scheme along with data uncertainty provides real-time access to the fused data and improves model accuracy
- Passive workflow management handles long-running analytics without overloading computing resources

Thank you

Sangmi Lee Pallickara
http://www.cs.colostate.edu/~sangmi
sangmi@colostate.edu