An Approach to Interactive Model Development on Big Data

Benjamin Lewis, Devika Kakkar, Weihe (Wendy) Guan,
Ryan Enos, Jacob R Brown
Harvard University

Center for Geographic Analysis
Harvard University

NSF

Spatiotemporal Thinking, Computing and Applications
STC

omni sci

IQSS
The Institute for Quantitative Social Science

Objectives

Develop methods to:

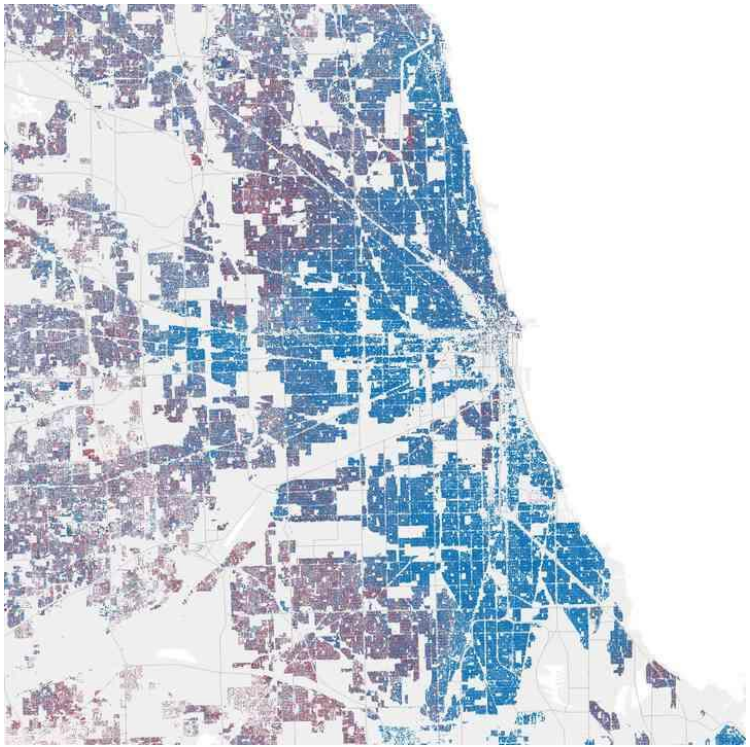
- *analyze* **big geospatial data**
- Interactively *scale* geospatial processes on big data
- Interactively *visualize* results of these geospatial processes on big data

-- which cannot be processed using traditional GIS tools.

By “**big**” we mean datasets too large to be processed using traditional GIS tools.

Use Case - partisan exposure of US voters

- **Objective:** Construct individual levels of partisan exposure for each voter in the US
- **Analysis:**
  - Part 1: Find KNN on a voter dataset of 180 million, with k=1000
  - Part 2: Processing the 180 billion KNN results to perform IDW modelling
- **Goal:** Perform the analysis in a cost and time-efficient manner



Available Computing Resources

FAS RESEARCH COMPUTING
HARVARD UNIVERSITY
FACULTY OF ARTS & SCIENCES

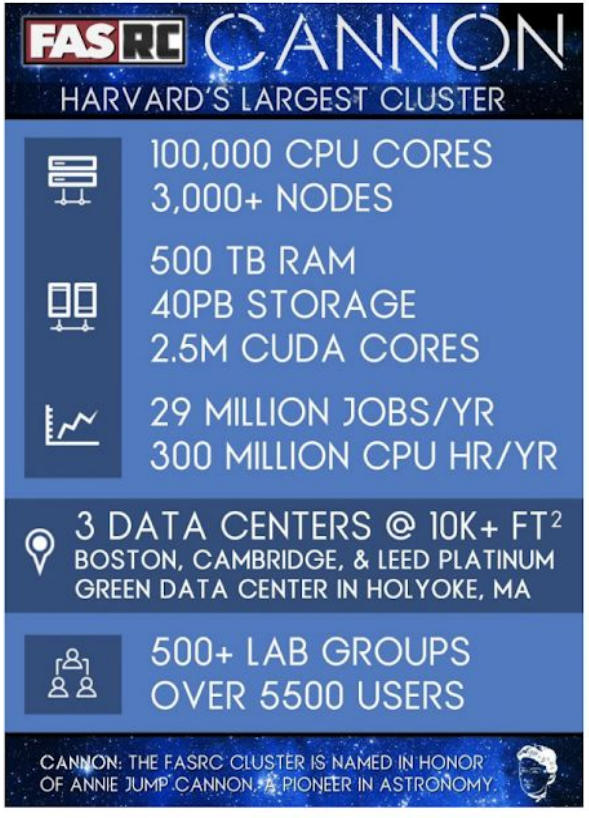
+

omnisci

I/UCRC for Spatiotemporal Thinking, Computing, and Applications

Harvard's FAS Research Computing Cluster

- Compute: 100,000 compute nodes, 8-64 cores/node, 12Gb to 512Gb memory/node, 2,500,000 NVIDIA GPU cores
- Software: CentOS 7 operating system, Slurm job manager, Singularity, 1000+ scientific tools and programs
- Storage: 100 GB (Home dir), 4TB+ (Lab storage), 70Gb/node (Local scratch), 2.4PB (Global scratch), 3PB (Persistent Research data)
- #144 in TOP500 Supercomputers in the world



GIS Databases for Big Data

- **PostGreSQL**
  - Powerful, open source object-relational database system
- **PostGIS**
  - Provides spatial objects for the PostgreSQL database
  - Allows storage and query of information about location and mapping
- **OmniSci**
  - Leverages the massively parallel processing of GPUs alongside traditional CPU compute
  - Super fast queries/analytics (including machine learning) of unindexed data (open source)
  - Super fast interactive rendering of millions or billions of features, on-the-fly on a map
  - Designed to overcome the scalability and performance limitations of legacy analytics tools

Challenges in KNN Computation

Major processing steps

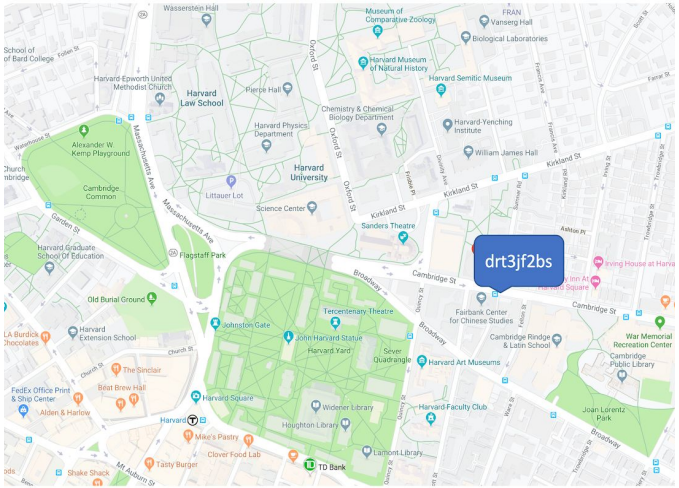
- Create spatial index on the data to speed up the search
- Find the nearest 1000 neighbors per voter
- Compute the Euclidean distances between the voter pairs

Obstacles

- For big spatial datasets, traditional algorithms are slow, resource intensive, costly and inefficient
- For datasets large enough not to fit on RAM, retrieval time from disk is slow without spatial index

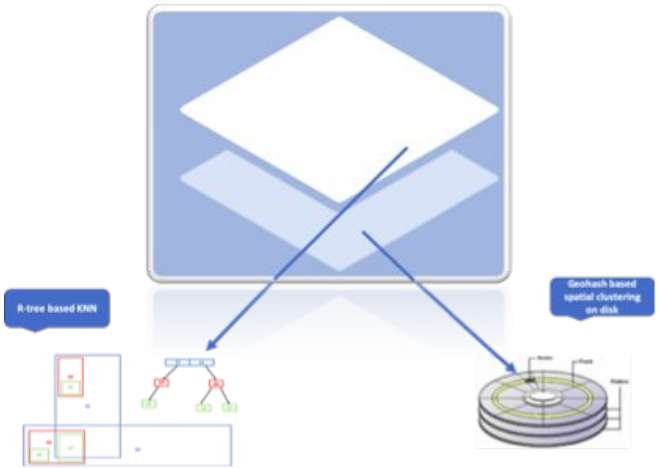
Solution - Geohash and Index based KNN

- Geohash is a method for expressing latitude and longitude using alphanumeric strings or hashes
- The longer a shared prefix between two geohashes is, the spatially closer they are together.
- Index based KNN is faster, cheaper and more efficient



Solution to Part 1: Geohash Clustering with KNN search

- Two layered approach:
  - Bottom layer of Geohash based clustering
  - Top layer of R-tree based KNN search
- Spatial clustering based on Geohash provides fast and efficient access on disk
- It ensures that records which are likely to be retrieved together are located together on disk
- KNN in PostGIS is a pure index based search
- Index based distance operator (<->) is used in the ORDER BY clause to make use of the DB indexes and LIMIT is used to truncate the search
- It works by walking up and down the index made on bounding boxes



Challenges in IDW Modeling

Major processing steps

- Compute Inverse Distance Weighted partisan exposure index of three party affiliations (Democratic, Republican, Independent) for each voter

Obstacles

- Traditional methods like R and Python are slow and resource intensive
- Traditional algorithms require multiple I/O of data which is inefficient for such big data

Solution - OmniSci

- GPU accelerated database, analytics and visualisation platform
- Uses graphics processing units (GPUs) and central processing units (CPUs) to query and visualize big data
- Queries can run across hundreds of CPU cores and tens of thousands of GPU cores per server
- OmniSciDB keeps hot data in GPU memory for the fastest access possible

Solution to Part 2: IDW Modelling Using OmniSci

- Rewrite the original R algorithm to OmniQL, a SQL dialect
- Divide the dataset for 180 billion into smaller chunks to fit OmniSci memory on FASRC
- Combine the results from multiple parallel modelling processes for final results

OmniSci provides the following advantages over traditional methods:

- Eliminates need of multiple I/O (which is a big overhead in these cases)
- Faster GPU based processing (compared to CPU processing)
- Speed improvement from 8 mins to 2.5 sec per 2 billion dataset (compared to traditional programming approaches such as R)

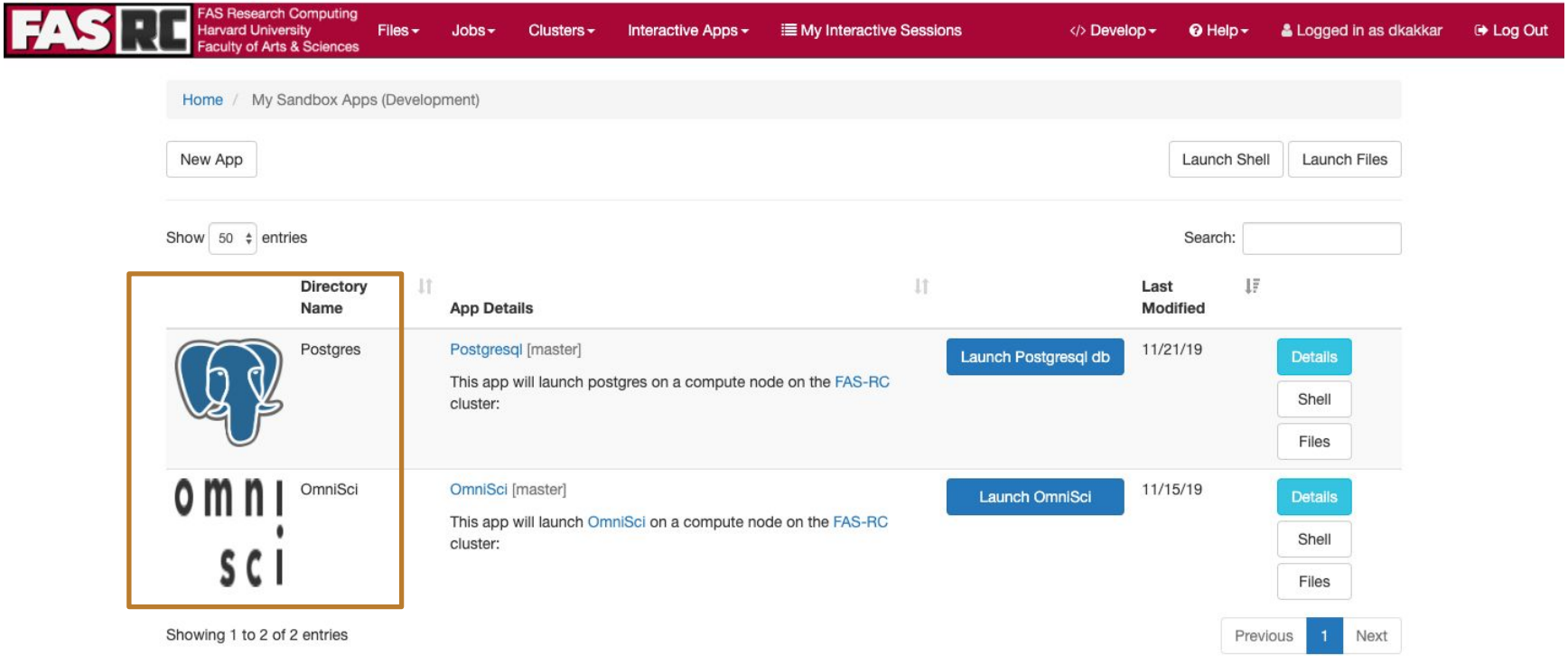
Novelty and Key Features of Our Solution

- KNN calculations using PostGIS on AWS:
  - Computing speed: 200,000 distances/sec
  - Hardware: m4.xlarge EC2 server
  - Software: PostGIS based KNN
  - Spatial index: Geohash based spatial indexing
  - Storage of results: Amazon S3, Compressed from 18 TB reduced to 1.5 TB
  - Customized AMI: Optimized PostGIS installed; ready to use for parallel processing
  - Cost Effective: ~ $175/month on AWS
- IDW modeling results visualization on FASRC:
  - Extremely fast GPU-based processing of results for interactive visualization
  - NVIDIA GPUs, 256GB RAM, 2 CPU cores, 1 GPU core per 2 million features

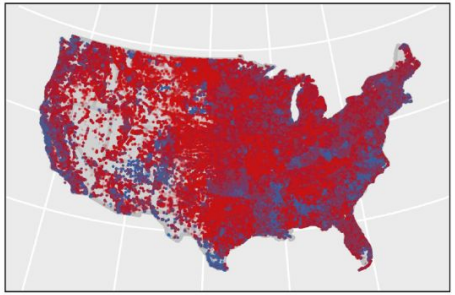
Deployment on FASRC

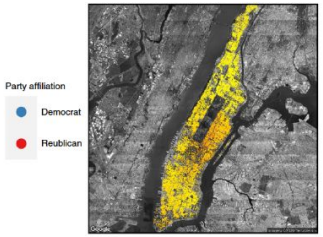
- FASRC provides a more sustainable and cost-efficient solution compared to AWS, especially for expanding the research to even bigger datasets
- PostGIS and OmniSci are installed as public apps on the FASRC to implement the solution
- I/O is a big overhead in such problems and FASRC makes it more efficient by loading the data once
- Several asynchronous jobs on FASRC run independent calculations and then combine the results in the end
- Non-Harvard users may implement the solution on AWS or their own slurm cluster
- Other slurm cluster users may install PostGIS and Omnisci on their cluster following our installation guide here: https://github.com/cga-harvard/GIS_Apps_on_HPC/tree/master/dev

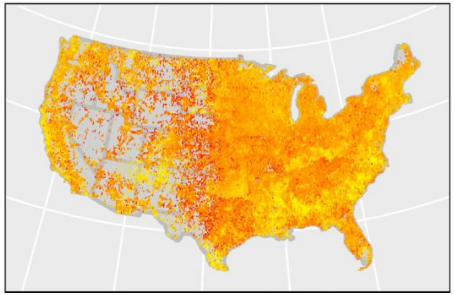
GIS Databases Deployed on FASRC

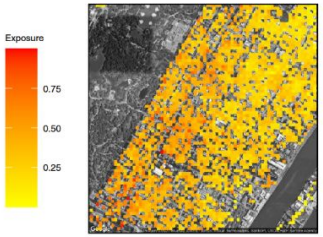


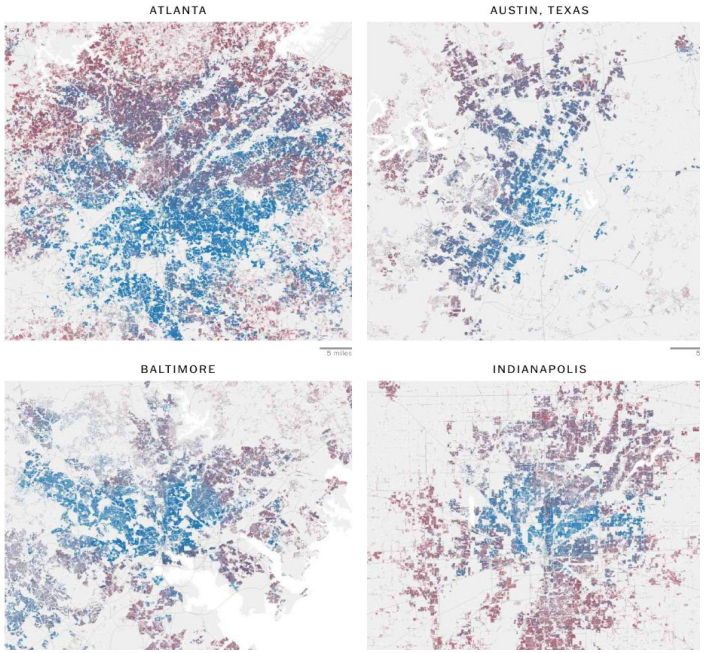
Results - partisan exposure of US voters visualizable at any scale











Recent Publications Supported by This Work

Article on Nature (3/8/2021)

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-021-01066-z

Check for updates

The measurement of partisan sorting for 180 million voters

Jacob R. Brown[1,2] and Ryan D. Enos[1,2]

Segregation across social groups is an enduring feature of nearly all human societies and is associated with numerous social maladies. In many countries, reports of growing geographic political polarization raise concerns about the stability of democratic governance. Here, using advances in spatial data computation, we measure individual partisan segregation by calculating the local residential segregation of every registered voter in the United States, creating a spatially weighted measure for more than 180 million individuals. With these data, we present evidence of extensive partisan segregation in the country. A large proportion of voters live with virtually no exposure to voters from the other party in their residential environment. Such high levels of partisan isolation can be found across a range of places and densities and are distinct from racial and ethnic segregation. Moreover, Democrats and Republicans living in the same city, or even the same neighbourhood, are segregated by party.

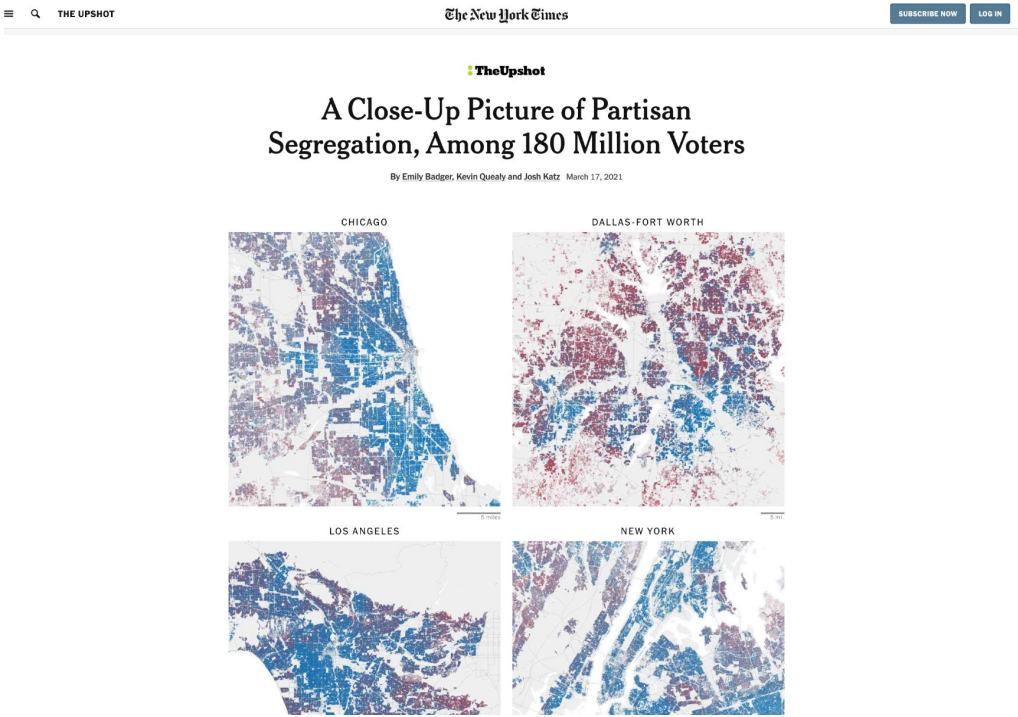
Segregation between human social groups is associated with a range of profoundly negative outcomes, including intergroup conflict, prejudice, inefficient resource allocation, poor democratic governance and other socially deleterious effects[1–4]. Segregation is also implicated in topics of intense interest across the social sciences, including interpersonal contact and intergroup relations[5,6], the bridging nature of social networks[7–11], poverty[11,12] and political representation[13,14]. Drawing on these associations and using aggregate data, popular and scholarly accounts of politics in the United States—and, increasingly, other Western democracies—describe stark partisan segregation, with members of different political parties living separate lives, resulting in partisan rancour and threatening the functions of the democracy[14–18]. Yet, despite the association between segregation and important outcomes, and the claims of increasing partisan segregation, the measurement of segregation among partisans, as with the measurement of segregation for most social groups, is severely limited: researchers must usually rely on data aggregations that do not include the actual locations of individuals, and thus measurements are limited to summaries across large geographical areas, and the experience of individual exposure across groups is masked.

In this article, using data on the exact residential address of every registered voter in the United States and harnessing advances in spatial data computation, we measure the local partisan segregation for each of these voters, creating a spatially weighted measure of cross-partisan exposure for more than 180 million individuals. Using this measure, we present evidence of extensive partisan segregation in the country. A large proportion of voters live with virtually no exposure to voters from the other party in their residential environment. Such high levels of partisan isolation can be found across a range of places and densities and are distinct from racial and ethnic segregation. Moreover, Democrats and Republicans living in the same city, or even the same neighbourhood, are segregated by party.

These high levels of partisan isolation have several important implications. In the United States, political party affiliation is considered a social identity, analogous to race or religion[19], and is a powerful predictor of a range of attitudes and behaviours[20], including behaviours outside of the explicitly political realm[21,22]. Because partisanship is correlated with political ideology and other attitudes and behaviours, the extent of a voter's partisan isolation is likely to affect their exposure to individuals different from themselves and to competing sociopolitical viewpoints, thus affecting a range of important outcomes. Cross-group exposure can be consequential for the shaping of intergroup attitudes and behaviours[6], including the prejudicial attitudes that are levelled across parties in the United States[23].

Isolated partisan environments may also affect behaviour through channels other than (a lack of) interpersonal contact: indeed, human behaviour can be shaped by low-level environmental cues[24,25], such as the norms displayed by neighbours, and randomized controlled trials have shown that political messaging from neighbours, such as the posting of yard signs, has a persuasive effect

Article on New York Times (3/17/2021)



Future Plan

- Expanding the partisan analysis to 20 years of US voter dataset to study the changes in voting behavior over time
- Combining Partisan Analysis with Twitter data to analyse the effect of social media on voting behavior
- Use distributed Omnisci infrastructure which allows single queries to span more than one physical host for big data
- Installing ArcGIS Enterprise on FASRC to make more GIS tools readily available to Harvard researchers
- Exploring other possible big data use cases, such as:
    - Disease Surveillance
    - Global Internet access mapping
    - National Water Model
    - EPA Air Quality Modeling

References

[1] Brown J. & Enos R., The measurement of partisan sorting for 180 million voters, Nature Human Behavior, 2021 https://www.nature.com/articles/s41562-021-01066-z.epdf

[2] Badger B., Quealy K. & Katz. J, A Close-Up Picture of Partisan Segregation, Among 180 Million Voters, 2021 https://www.nytimes.com/interactive/2021/03/17/upshot/partisan-segregation-maps.html

[3] Kakkar D., Lewis B., Singh R., OmniSci Virtual Summit, 2020 https://www.youtube.com/watch?v=3DlOeWqDMSs

[4] Kakkar D., Lewis B., Scaling geospatial processes on Harvard's high-performance cluster, Harvard DataFest, 2020 https://drive.google.com/file/d/1FEnh-okCNLuthtyQtoBldyid7D6Sb-F_/view?usp=sharing

[5] Introduction to Cluster Computing on FASRC: https://www.rc.fas.harvard.edu/wp-content/uploads/2019/12/Intro-to-Cannon.pdf

[6] About Postgis: https://postgis.net/

[7] OmniSci Overview: https://docs.omnisci.com/latest/4_distributed.html

Acknowledgement

This work is partially sponsored by NSF Awards #1841403 and OmniSci.

Co-authors of this presentation:

Ben Lewis

Devika Kakkar

Ryan Enos

Jacob R. Brown