

Challenges and Solutions for Private and Reproducible Environmental Exposure Assessment at Scale

Cole Brokamp

Division of Biostatistics and Epidemiology
Cincinnati Children's Hospital Medical Center

January 12, 2021



Table of Contents

- Background
- DeGAUSS
- Spatiotemporal Geomarkers
- Conclusion

Geomarkers

Geocoding Converting a string of text into spatial coordinates or boundaries

Geomarker Any geospatial measure that influences or predicts health

Geomarkers

Geocoding Converting a string of text into spatial coordinates or boundaries

Geomarker Any geospatial measure that influences or predicts health

place (+ time) → estimating past “exposures”

Geomarkers

- ▶ Geomarkers are the most powerful predictor of disease, disorder, injury, and mortality
- ▶ Data and tools needed for high resolution spatiotemporal geomarker assessment at a population level

Geomarkers

- ▶ Geomarkers are the most powerful predictor of disease, disorder, injury, and mortality
- ▶ Data and tools needed for high resolution spatiotemporal geomarker assessment at a population level

- ▶ Practical usage for exposure assessment is hindered by
 - large data + inefficient manual data curation
 - the need for technical expertise and software skills
 - privacy restrictions

Protected Health Information

- ▶ Confidentiality of research subjects must be safeguarded
- ▶ HIPAA-defined “Safe Harbor” provision prohibits sharing of identifiers and quasi-identifiers, such as:
 - time finer than a calendar year
 - spatial boundary with < 20,000 residents

Protected Health Information

- ▶ Confidentiality of research subjects must be safeguarded
- ▶ HIPAA-defined “Safe Harbor” provision prohibits sharing of identifiers and quasi-identifiers, such as:
 - time finer than a calendar year
 - spatial boundary with < 20,000 residents

- ▶ Sharing PHI
 - consent often not obtained for unforeseen future analyses
 - retrospective consent often not feasible + consent bias
 - IRB and institutional DUA approvals can be lengthy and have different requirements
 - transmission of PHI to a third party often not possible

Protected Health Information

- ▶ Confidentiality of research subjects must be safeguarded
- ▶ HIPAA-defined “Safe Harbor” provision prohibits sharing of identifiers and quasi-identifiers, such as:
 - time finer than a calendar year
 - spatial boundary with < 20,000 residents

- ▶ Sharing PHI
 - consent often not obtained for unforeseen future analyses
 - retrospective consent often not feasible + consent bias
 - IRB and institutional DUA approvals can be lengthy and have different requirements
 - transmission of PHI to a third party often not possible

- ▶ Presents challenges when integrating geomarkers into research studies and clinical applications

Problems with Current Approaches for Multi-Site Studies

- ▶ Anonymization
 - geomasking, date shifting, generalization
- must balance decrease in precision with analysis needs

Problems with Current Approaches for Multi-Site Studies

▶ Anonymization

- geomasking, date shifting, generalization
- must balance decrease in precision with analysis needs

▶ Independent Geomarker Assessment

- specialized expertise and technical skills required at each site
- differences in methods introduce differential error and bias downstream health associations

Problems with Current Approaches for Multi-Site Studies

▶ Anonymization

- geomasking, date shifting, generalization
- must balance decrease in precision with analysis needs

▶ Independent Geomarker Assessment

- specialized expertise and technical skills required at each site
- differences in methods introduce differential error and bias downstream health associations

▶ Existing Software Approaches

- commercial options are cost prohibitive and aren't designed for batch operations
- closed source geocoder prevents transparency and reproducibility

Vision

- Curated and standardized library that researchers can utilize for secure, efficient, automated, and reproducible linkage of geomarkers to their own protected health and geolocation data.

Vision

- Curated and standardized library that researchers can utilize for secure, efficient, automated, and reproducible linkage of geomarkers to their own protected health and geolocation data.
- A generalized framework for geomarker curation and computation to which exposure scientists can contribute.

Vision

- Curated and standardized library that researchers can utilize for secure, efficient, automated, and reproducible linkage of

geomarkers to their own protected health and geolocation data.

- A generalized framework for geomarker curation and computation to which exposure scientists can contribute.
- ▶ FAIR (findable, accessible, interoperable, reusable) data
- ▶ *Reproducible* using computable exposures
- ▶ *Portable* for sharing and mobility of compute

Table of Contents

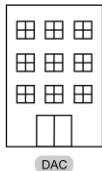
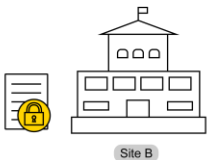
- Background
- DeGAUSS
- Spatiotemporal Geomarkers
- Conclusion

DEcentralized Geomarker Assessment for mULTI Site Studies

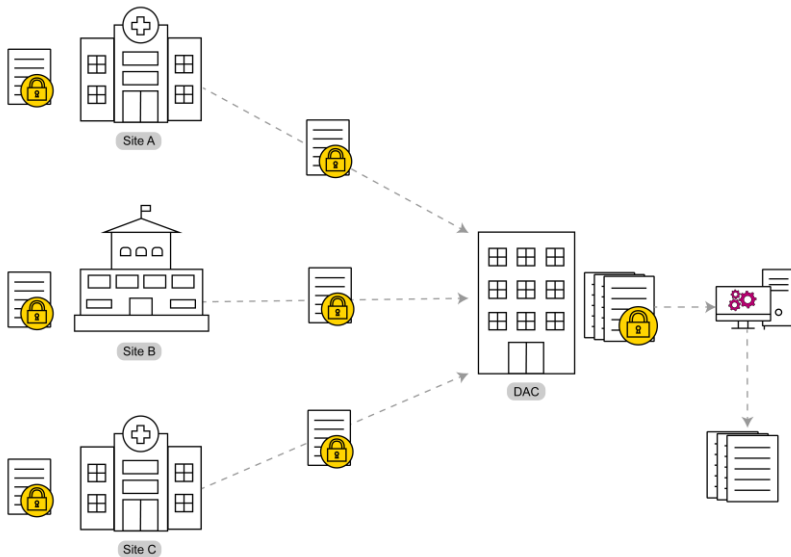


<https://degauss.org>

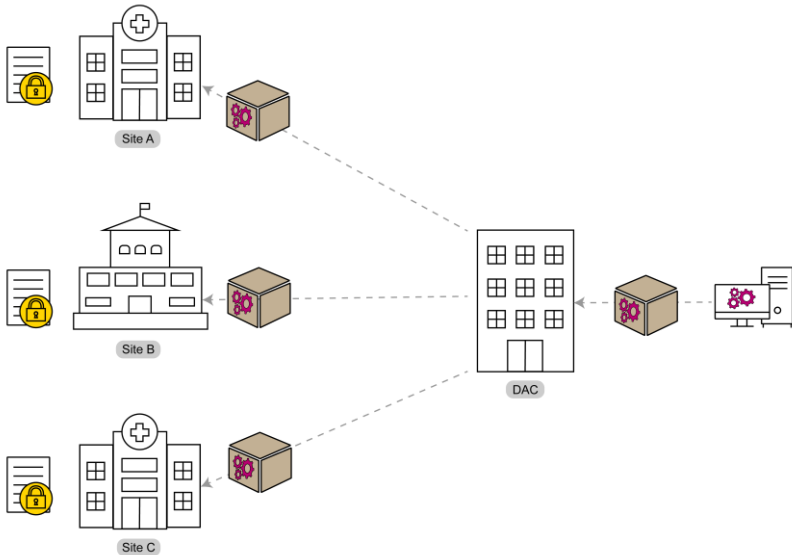
Bringing Computation to Data



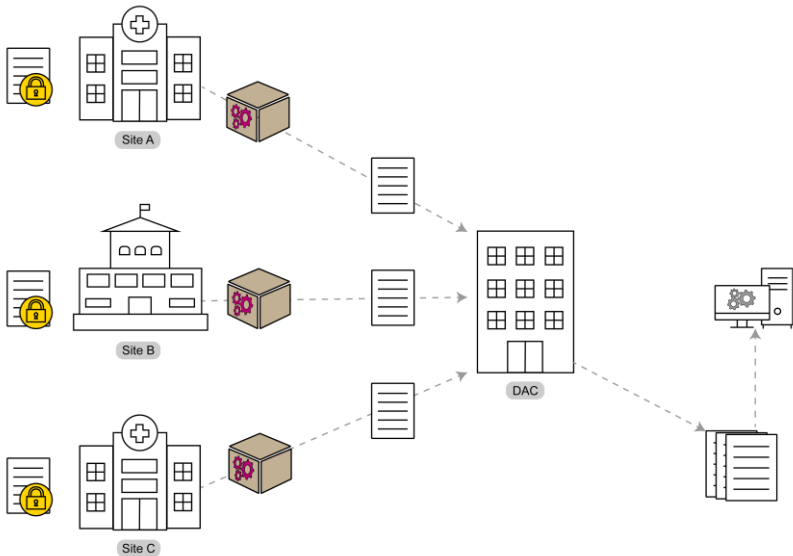
Bringing Computation to Data



Bringing Computation to Data



Bringing Computation to Data



DeGAUSS

- ▶ Decentralized but reproducible and standardized
 - ▶ Container framework that reads and writes CSV files
 - ▶ No extensive computational resources ▶

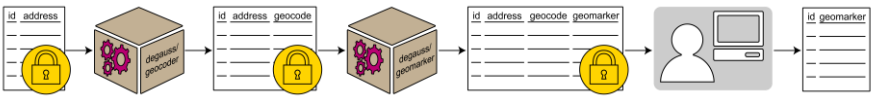
No geospatial or computing expertise required

- ▶ *PHI is never exposed to a third party or the internet*

DeGAUSS

- ▶ Decentralized but reproducible and standardized
 - ▶ Container framework that reads and writes CSV files
 - ▶ No extensive computational resources ▶
- No geospatial or computing expertise required

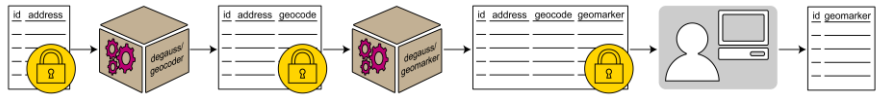
▶ *PHI is never exposed to a third party or the internet*



DeGAUSS

- ▶ Decentralized but reproducible and standardized
- ▶ Container framework that reads and writes CSV files

- ▶ No extensive computational resources
- ▶ No geospatial or computing expertise required
- ▶ *PHI is never exposed to a third party or the internet*



- ▶ Free and open source
- ▶ Automated and continuous documentation and integration
- ▶ Metadata curation and integration
- ▶ Multiple user entry-points (data, geomarker assessment code, Docker/OCI images, GUI, stand-alone application) ▶

Community supports and contributions

Anonymity and Reidentification

- ▶ Anonymity can ensure small, but non-zero, chance of reidentification
 - published examples of reidentification attacks by researchers (Sweeney 2017, Boronow 2020)
 - reidentification tasks are rare and often unsuccessful (Emam 2011, Emam 2015)
- ▶ Don't confate re-identification of identifiers with re-identification of quasi-identifiers
 - quasi-identifiers recovered by merging with extant datasets
 - institutional restrictions on sharing of quasi-identifiers

Table of Contents

- Background
- DeGAUSS
- Spatiotemporal Geomarkers
- Conclusion

High Resolution Spatiotemporal Geomarkers

▶ Pre-computed data “products”

- produced by from interpolation/prediction exposure models
- often uses publicly available spatiotemporal datasets
- ambient air pollution, climate, noise, wildfres, crime

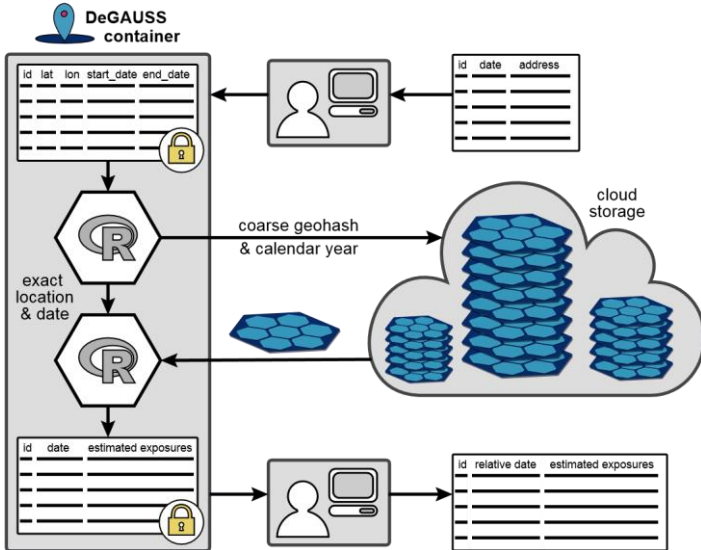
▶ High resolution

- often < 1 km sq. exposures covering entire country
- daily estimates covering 2000 - 2021 ▶ Exposure timing
- used to study acute, short-term, and long-term exposures

- development-based temporal averages during early life

- ▶ Large file sizes require data transmission, when most of data usually not used
- ▶ Most approaches currently require sharing PHI with model developer for addition of estimates

Approach



○○○○○

Background
Geomarkers

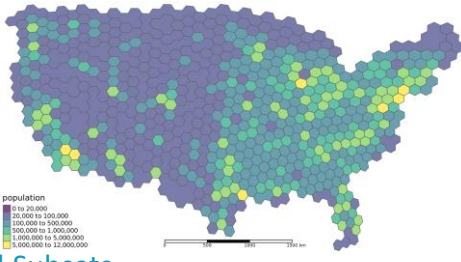
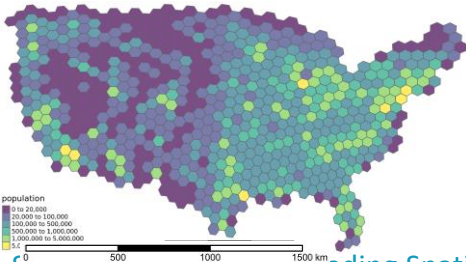
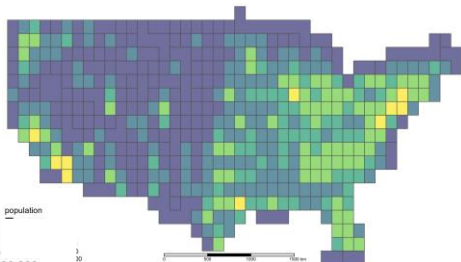
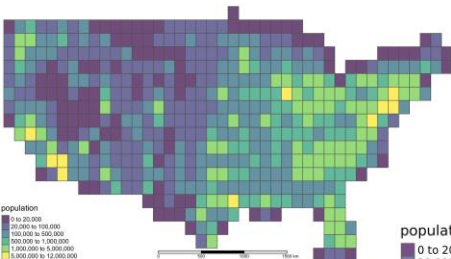
○○○○○○○

○○○○○

DeGAUSS
Conclusion

Spatiotemporal

○○○○○



Sampling Spatial Subsets

Applications

- ▶ ECHO, eMERGE, government organizations, electronic health data warehouses
 - different levels of consent, data management and coordination centers
- ▶ Applied within DeGAUSS containers for several different daily, high resolution ambient pollution estimates
 - <https://degauss.org/pm>
 - <https://degauss.org/schwartz>

Advantages

- ▶ Prevents download of unnecessary spatial and/or temporal “slices” of data
- ▶ Decreases time and resources needed by end user to run software without sharing PHI
- ▶ Automated downloading, parsing, and spatiotemporal joining

DeGAUSS

image*	description	version**
ghcr.io/degauss-org/geocoder	batch geocoding	version v3.0.2
ghcr.io/degauss-org/census_block_group	census block group and tract FIPS	version v0.4.1
ghcr.io/degauss-org/st_census_tract	spatiotemporal census tract FIPS 1970 - 2020	version v0.1.2
ghcr.io/degauss-org/dep_index	census tract-level deprivation index	version v0.1
ghcr.io/degauss-org/roads	proximity and length of major roads	version v0.1
ghcr.io/degauss-org/aadt	average annual daily traffic	version v0.1.1
ghcr.io/degauss-org/greenspace	enhanced vegetation index	version v0.2
ghcr.io/degauss-org/nlcd	land cover (imperviousness, land use, greenness)	version v0.1
ghcr.io/degauss-org/pm	daily PM2.5	version v0.1.3
ghcr.io/degauss-org/narr	daily weather data (air temperature, humidity, etc)	version v0.1
ghcr.io/degauss-org/drivetime	distance and drive time to various care sites	version v1.0
degauss/schwartz_grid_lookup	schwartz grid for spatiotemporal pollutant models	version v0.4.1
degauss/schwartz	daily PM2.5, NO2, and O3 concentrations	version v0.5.5

https://degauss.org/available_images

Table of Contents


- Background
- DeGAUSS
- Spatiotemporal Geomarkers
- Conclusion

Future Directions

- ▶ GUI interfaces for researchers and scientists
- ▶ Metadata curation for data science workflows and clinical informatics pipelines
- ▶ Cloud Optimized Geotifs (COG)
- ▶ Integrating methods for “deidentifying” area-level data
- ▶ Homomorphic encryption
- ▶ Facilitating community contributions

Discussion

G x E x Time

- geomarkers and epigenome change over time
- report back for spatiotemporal exposures
- less focus on privacy/precision tradeoffs for time 

Geospatial data collection and sharing

- empower people to donate their own spatiotemporal data collected via cloud-hosted location trackers
- think about consent in the future: limited sharing of pseudo-identifiers only?

Discussion

- ▶ HIPAA Safe Harbor not sufficient to guarantee anonymity, but should this be our goal in research studies?
- ▶ Updated guidance & policies needed
 - zip code. . .
 - details on spatial and temporal generalization strategies
 - update examples to use census-defined boundaries
 - reidentification of pseudo-identifiers versus identifiers
 - how to deal with datasets that may be considered de-identified now, but will change to identified after unforeseen datasets and methods arise?
- ▶ Must maintain reproducibility *and* privacy

Thank You

<https://degauss.org> @degauss-
org

cole.brokamp@cchmc.org
<https://colebrokamp.com>
@cole_brokamp

DeGAUSS is supported by NIH R01LM013222 & U2COD0233754



Department of Health

