SPACESCANS: Semantics Standards and Tools for Spatial and Contextual Exposome Data

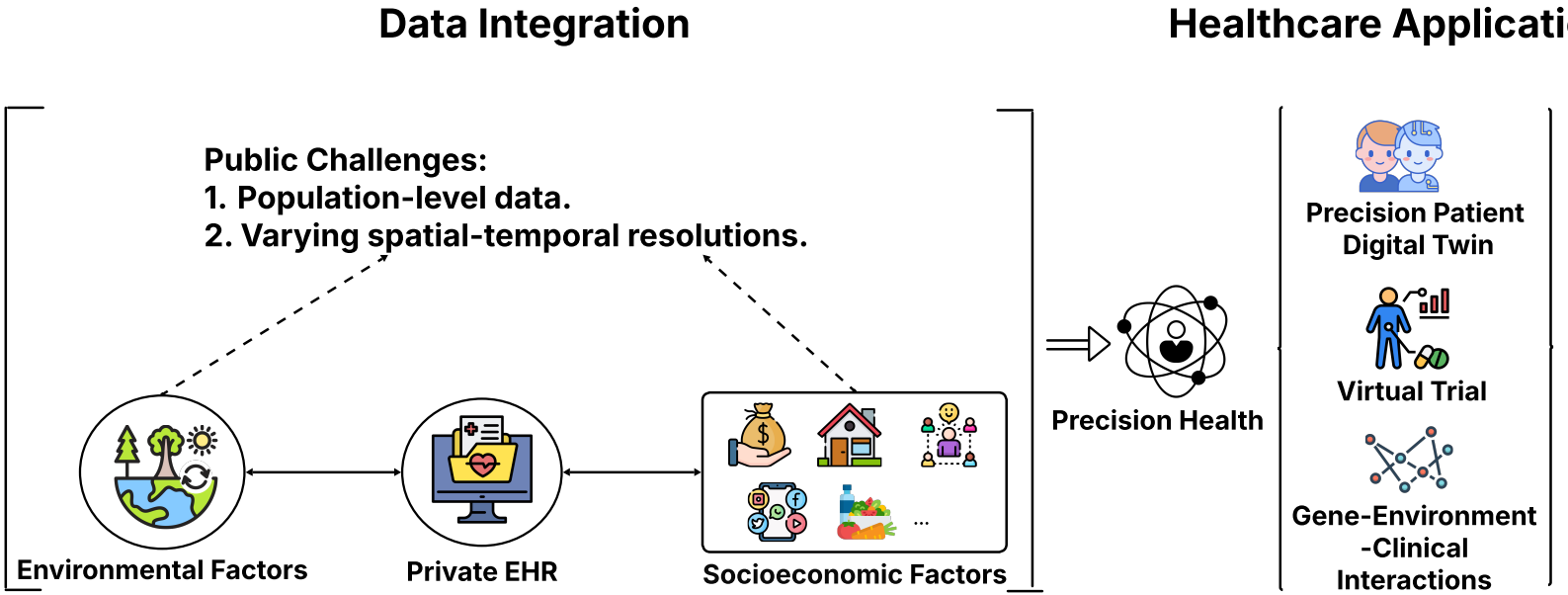
Presented by: Dr. Shuteng Niu, Mayo Clinic and Dr. Xing He, Indiana University

Funded by: NIH NIEHS R24ES036131

PIs: Dr. Jiang Bian, Indiana University | Dr. Cui Tao, Mayo Clinic | Dr. Hui Hu, Harvard University

What Defines Health?

Human health is defined by biological factors, environmental exposures, and social determinants.



SPACESCANS

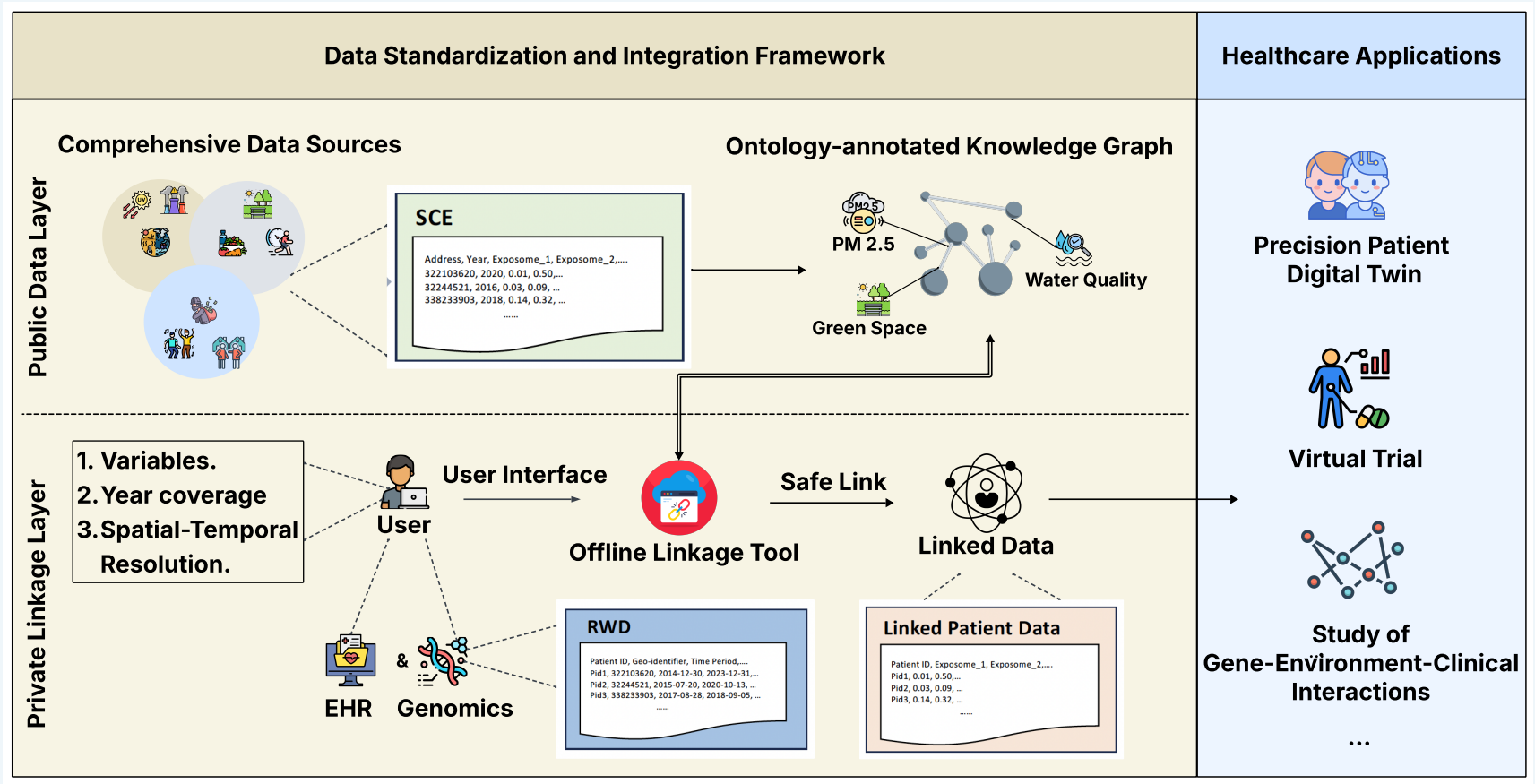
This project fills the gap by:

1. Developing a scalable and ontology-driven data standardization and integration method.

2. Developing a safe and privacy-aware linkage tool.

3. Establishing use cases on real-world applications in healthcare.

An Ontology-based Data Standardization and Integration Framework for Safe Linkage

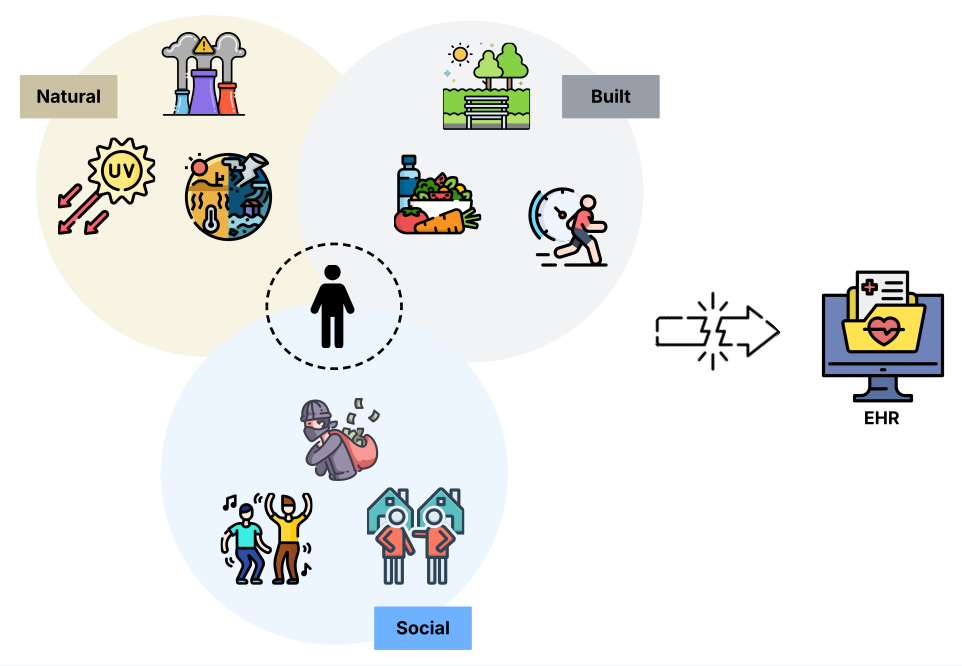


Exposome Coverage

Comprehensive data source collection:

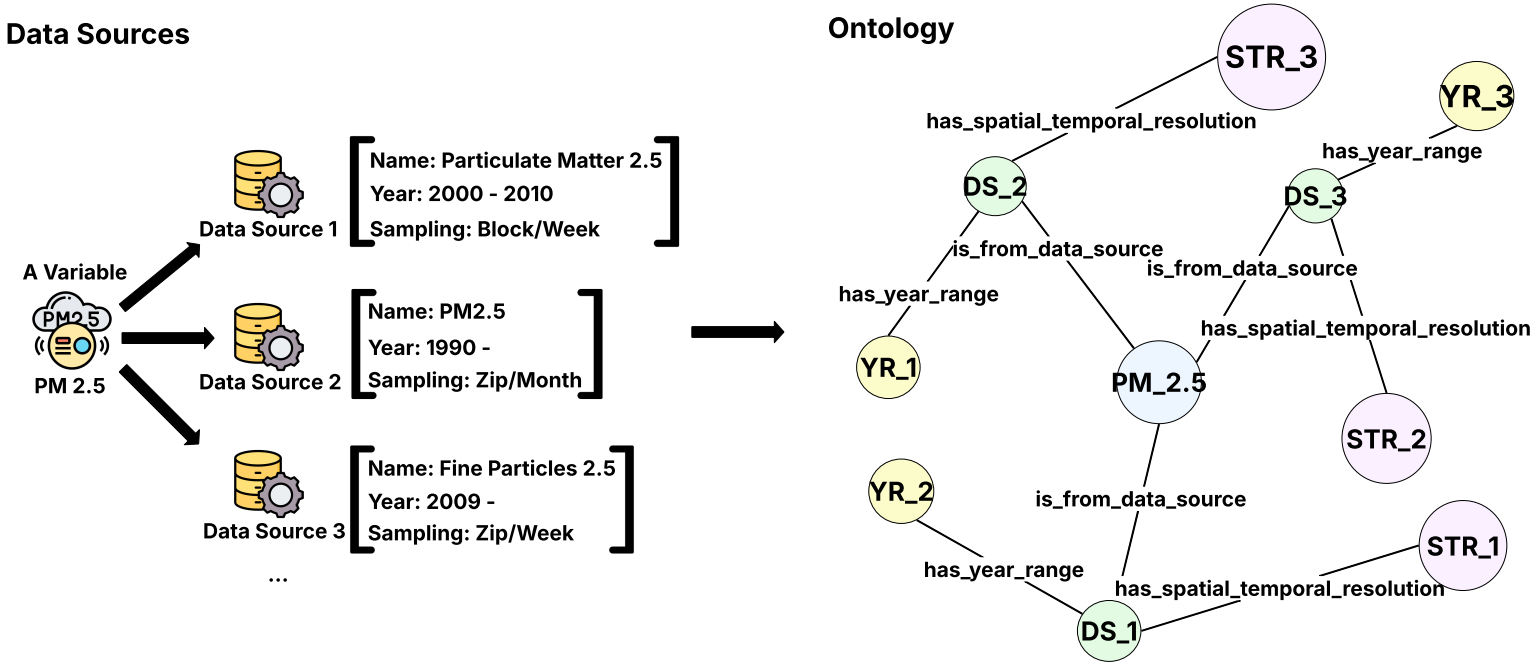
1. **Natural Environmental Exposome**
2. **Built Environment Exposome**
3. **Social Environment Exposome**

Adds critical perspectives in health on top of EHR data.



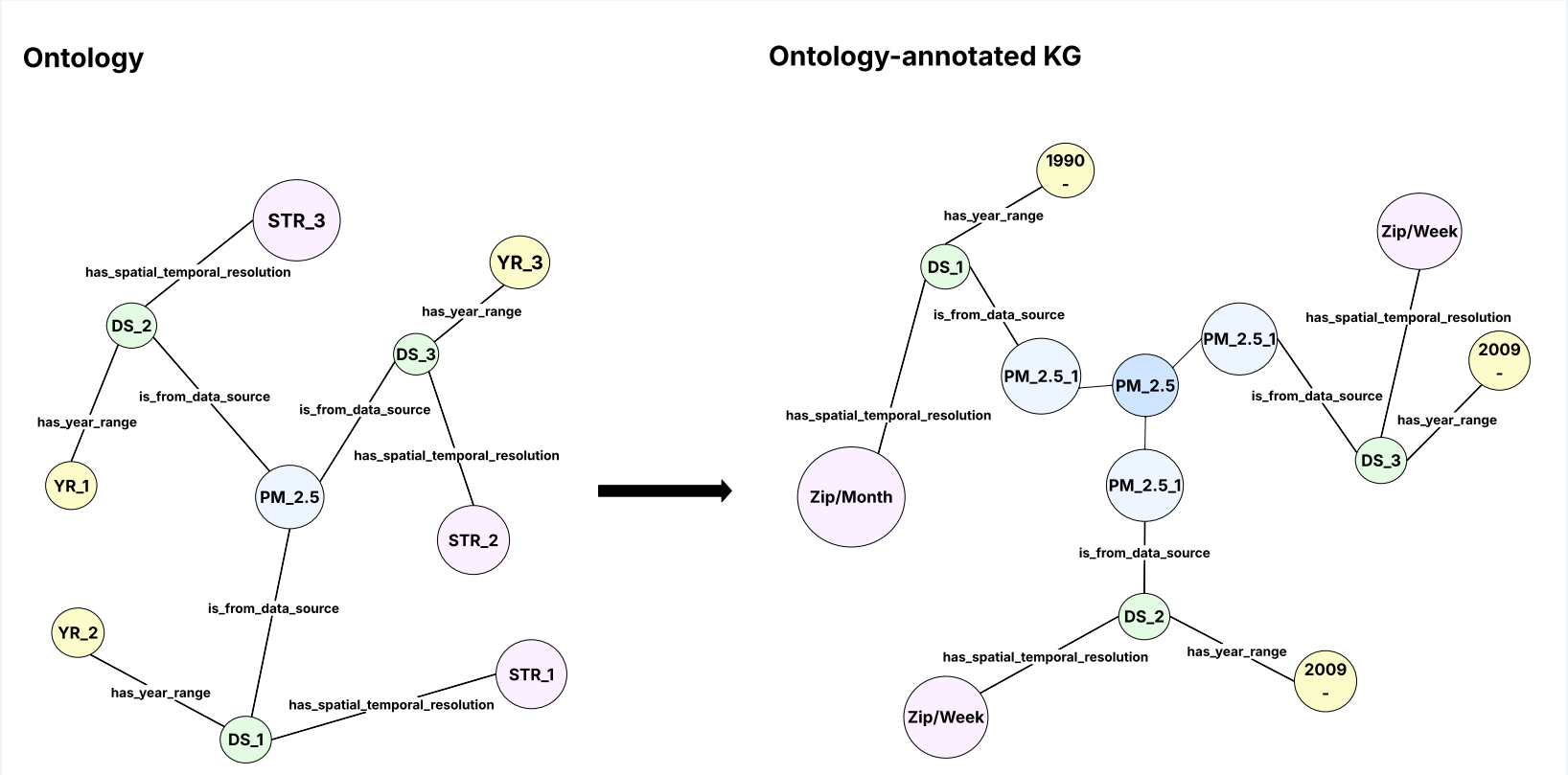
What Ontology Is and How it Supports the Data Layer

An ontology uses formal concepts and relationships that specifies the meaning and structure of data within a domain.



From Ontology to Knowledge Graph

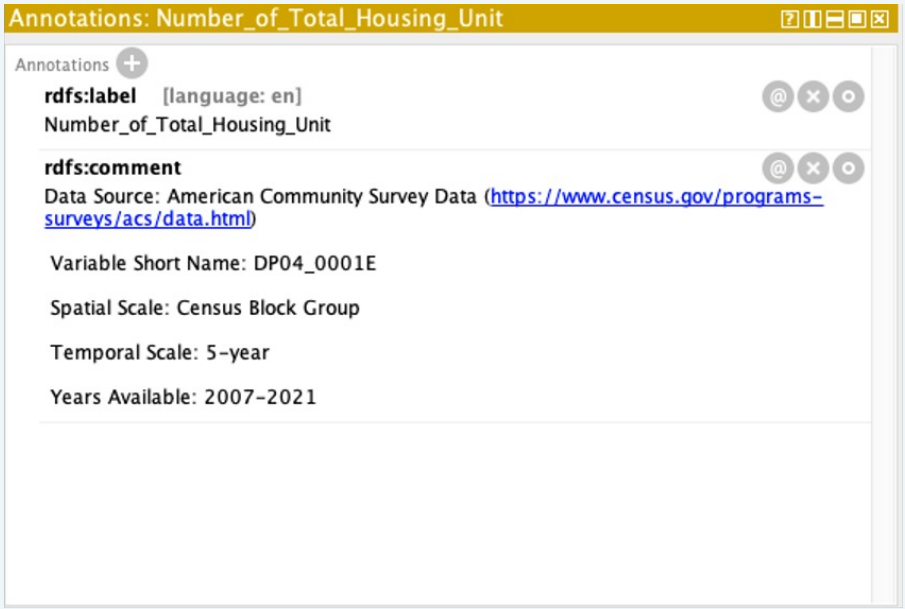
We store actual values of data using the ontology as a semantic structure



A Sample Variable in Ontology

Each variable has:

1. **Label with semantic meaning.**
2. **Short variable name for linakge.**
3. **Available spatiotemporal scales.**
4. **Available years.**



Current Data Coverage

9 data sources, 543 variables, and 6 spatial-temporal resolutions.

Summary of processed and additional spatial and contextual exposome data sources.

	Data Source	Available	Geometry	Spatiotemporal Scale	# of Vars	Example Measures
Natural Environment						
Air pollution	CACES	1979-2020	Polygon	BG/1-year	6	$PM_{2.5}$
	ACAG	2000-2023	Raster	0.01deg/2-week	21	Black carbon
	FAQSD	2002-2021	Polygon	CT/1-day	2	$PM_{2.5}$
	SEDAC	2005-2016	Raster	1km/1-day	3	O_3
	NATA and AirToxScreen	2011-2020	Raster	CT/1-year	175	Benzene
Ultraviolet	TEMIS	2002-2025	Raster	0.25deg/1-day	4	Clear-sky DNA-damage UV dose
Blue space	NHDPlusV2	CS	Polygon	Vector/CS	5	Proximity to NHD coastline
Meteorology	PRISM	1981-2025	Raster	800m/1-day	9	Mean temperature
Built Environment						
Vacant land	US HUD	2006-2025	Polygon	CT/3-month	19	Percent of vacant addresses
Walkability	National Walkability Index	CS	Polygon	BG/CS	1	Walkability index
Food access	USDA FARA	2010, 2015, 2019	Polygon	CT/1-year	44	Percent of low-access population at 1 mile
Green space	MODIS	2000-2025	Raster	250m/16-day	1	NDVI
	HLS	2013-2025	Raster	30m/2-day or 3-day	1	NDVI
Noise	NPS	CS	Raster	270m/CS	3	Existing A-weighted L50 sound pressure level
Road proximity	TIGERLine Roads	2000-2024	Line	Line/1-year	4	Proximity to A1 road (primary highways with limited access)
Light at night	VIIRS	2012-2024	Raster	15sec/1-year	1	Light at night
Social Environment						
Sociodemographic factors	ACS	2005-2023	Polygon	BG/5-year	1,041	Neighborhood deprivation index
Social Capital	CBP	1986-2024	Polygon	ZCTA/1-year	10	Religious, civic, and social organizations
Crime and Safety	UCR	1974-2024	Polygon	County/1-year	7	Burglary rate, aggravated assault rate

BG: Census Block Group; CT: Census Tract; ZCTA: ZIP Code Tabulation Area; CS: Cross-sectional

Use Case: Green Space to Major Depression and Asthma

- **Objective:** Investigate the association between major depression and green space coverage using linked HER data

- **Data Sources:**
- Mayo Clinic EHR: demographics, diagnoses, and SDoH.
- SPACESCAN: Green space percentage from 2008 – 2021, including 5751 unique zip codes covering 13 states around three Mayo campuses.

- **Approach:** Temporal data representation and ML model development.

Partial data credits to Dr. Huaqing Wang.

Preliminary Results of Causal Inference

```
Optimization terminated successfully.
         Current function value: 0.692969
         Iterations 3
                           Logit Regression Results
==============================================================================
Dep. Variable:                  label   No. Observations:               239832
Model:                          Logit   Df Residuals:                   239830
Method:                           MLE   Df Model:                            1
Date:                Thu, 29 May 2025   Pseudo R-squ.:               0.0002570
Time:                        18:48:29   Log-Likelihood:            -1.6620e+05
converged:                       True   LL-Null:                   -1.6624e+05
Covariance Type:            nonrobust   LLR p-value:                  2.374e-20
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0247      0.005      5.061      0.000       0.015       0.034
treatment     -0.0824      0.009     -9.243      0.000      -0.100      -0.065
==============================================================================
```

$$Odd\_Ratio = e^{-0.0824} \approx 0.921$$

Risk_Reduction= $(1 - Odd\_Ratio) \times 100 = 7.9\%$

Patients exposed to more green space have ***7.9%*** lower odds of depression.

Preliminary Results of Causal Inference

```
Optimization terminated successfully.
         Current function value: 0.693045
         Iterations 3
                           Logit Regression Results
==============================================================================
Dep. Variable:                  label   No. Observations:               101742
Model:                          Logit   Df Residuals:                   101740
Method:                           MLE   Df Model:                            1
Date:                Fri, 30 May 2025   Pseudo R-squ.:               0.0001479
Time:                        09:10:18   Log-Likelihood:                -70512.
converged:                       True   LL-Null:                       -70522.
Covariance Type:            nonrobust   LLR p-value:                 4.937e-06
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0186      0.007     -2.491      0.013      -0.033      -0.004
treatment      0.0627      0.014      4.567      0.000       0.036       0.090
==============================================================================
```

$$Odd\_Ratio = e^{-0.0824} \approx 1.065$$

$$\text{Risk\_Reduction} = (1 - Odd\_Ratio) \times 100 = -6.5\%$$

Patients exposed to more green space have ***6.5%*** higher odds of asthma.

Use Case: Air Quality to PBC prediction

- **Objective:** Predict PBC using linked data from EHR, SPACESCAN environmental data, and water quality data.

- **Data Sources:**
- Mayo Clinic EHR: demographics, diagnoses, CCI, autoimmune flags, and addresses.
- SPACESCAN: environmental exposures including air and neighborhood-level data.
- Water Data: ZIP-code level municipal water source types and quality indicators.

- **Approach:** Multimodal data integration and ML model development.

Preliminary Results and Next Steps

- **Preliminary Model Performance and Future Work**
- **Model:** Random Forest using
  - EHR only EHR + water
  - EHR + air pollution
  - EHR + water + air pollution data
- **Initial AUC: 0.74 (baseline model without lab data).**

- **Planned Enhancements:**
  - Add lab data (e.g., LFTs, AMA).
  - Explore time-series and other ML or DL models.
  - Validate using external cohort.

- **Goal: Develop a clinically actionable prediction model for early detection of PBC.**

Exposome Linkage Pipeline

& SPACESCANS Web Application

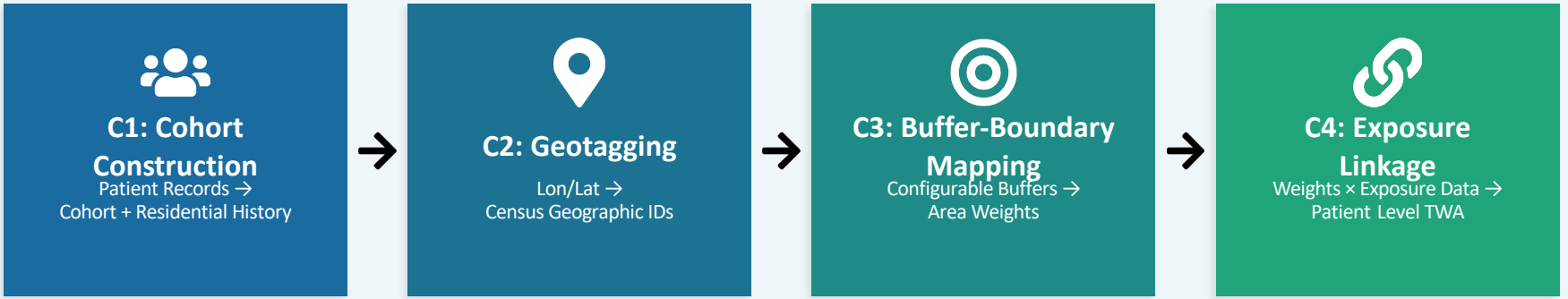
Building a Multi-Domain Environmental Exposure Linkage Pipeline
with an Ontology-Guided Web Interface for Reproducible Linkage

Input: patient_id + residential address history (lon/lat, start_date, end_date)
Output: patient-level time-weighted average exposures across multiple domains

NIH NIEHS R24ES036131 | Indiana University | Mayo Clinic | Harvard University

PIPELINE

Four-Stage Pipeline: From Patient Records to Linked Exposures



Pipeline Implementation Status

Initially developed as R scripts with proven production use (e.g., 802K pregnancies × 14 data sources in the Exposome-HDP study). **Currently extending to a Python package** with multi-processing support for significantly faster execution, designed as a general-purpose library and the computational backend of the SPACESCANS Web Application.

STAGE C1

Cohort Construction & Residential History

From Patient Records to Geocoded Residential Histories

Required Input: Residential History Table

patient_id	lon	lat	start_date	end_date

Each row = one address period for one patient. Multiple rows per patient capture residential mobility.

1. Define study cohort: identify patients and exposure time windows from clinical data or user-supplied records
2. Reconstruct residential history: timestamped addresses from EHR encounter records or user-provided address logs
3. Geocode addresses: resolve to lon/lat coordinates (e.g., ZIP+4 centroid, address-level geocoder)
4. Validate: filter invalid points, deduplicate overlapping address intervals, ensure temporal coverage

OUTPUT

N

patients

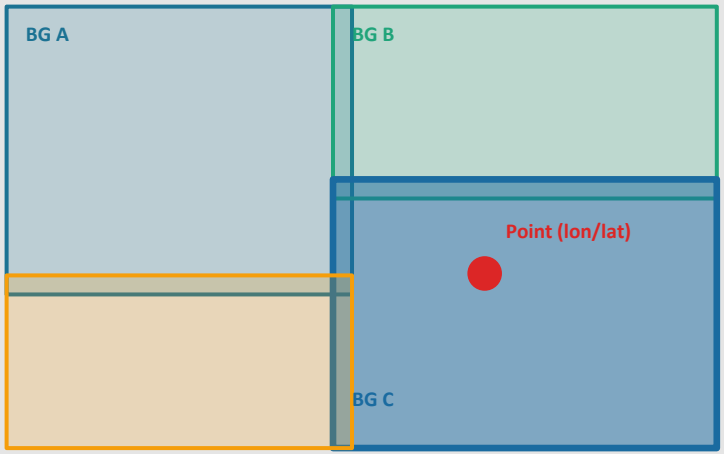
Geocoded Residential History

`patient_id, geoid, lon, lat, start_date, end_date`

e.g., 802K pregnancies in HDP study

STAGE C2

Geotagging: Point-in-Polygon Assignment



Residence assigned to Block Group C

Spatial overlay

Each geocoded residence → Block Group (BG) and ZCTA5

Derive geography

Tract and County from BG GEOID (numeric truncation)

Primary purpose

Filter out invalid points that fall outside any census polygon

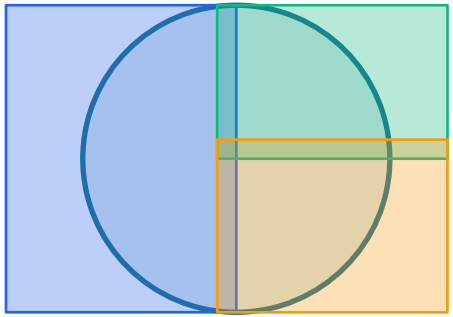
Secondary purpose

Consolidated geocoded point file for all downstream scripts

STAGE C3

Buffer-Boundary Weight Generation

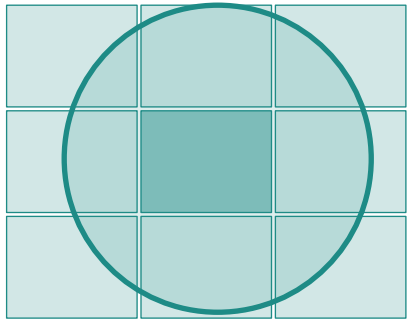
Polygon Approach (BG, Tract, County, ZCTA5)



`BG_A: 0.60` `BG_B: 0.30` `BG_C: 0.10`

1. Rasterize polygon → 25m binary grid
2. exact_extract(mean) on binary raster
3. Result = fraction of buffer inside polygon

Raster Approach (ACAG, PRISM, TEMIS, MODIS, etc.)



`cell_1: 0.45` `cell_2: 0.35` `cell_3: 0.15` `cell_4: 0.05`

1. Use template raster from data source
2. exact_extract with include_cell + include_area
3. Normalize: weight = coverage / total_coverage

STAGE C4

Exposure Linkage: Three-Step Aggregation

Step A: Spatial Aggregation

For each geoid × time period, area-weighted mean across overlapping spatial units

$$\text{exposure\_geoid} = \Sigma(\text{weight}_i \times \text{value}_i) / \Sigma(\text{weight}_i)$$

Step B: Temporal Aggregation

For each address, day-weighted mean across overlapping time periods

$$\text{exposure\_addr} = \Sigma(\text{exp\_geoid}_t \times \text{overlap\_days}_t) / \Sigma(\text{overlap\_days}_t)$$

Step C: Patient Aggregation

Aggregate across all residential addresses (folded into Step B)

$$\text{Patient TWA} = \Sigma(\text{area-weighted exp} \times \text{overlap\_days}) / \Sigma(\text{overlap\_days})$$

Core: Patient TWA = Σ(area weighted exposure × overlap days) / Σ(overlap days)

DATA SOURCES

Environmental & Climate Data Sources

Source	Resolution	Temporal	Key Variables
ACAG V5	~1 km	Biweekly	PM2.5, BC, dust, NH4, NO3, OM, SO4, SS
FAQSD	Tract	Daily	Ozone (8hr max), PM2.5 (daily avg)
PRISM	800m	Daily	Precip, Tmax/min/mean, VPD, RH, Heat Index
TEMIS	Global	Daily	UV doses (DNA-damage, erythemal, vitamin-D)
MODIS NDVI	250m	16-day	NDVI (Terra+Aqua blended, gap-filled)
VIIRS NTL	~500m	Annual	Nighttime radiance
DOT Noise	~270m	Static	L50 dBA: existing, impact, natural
NHDPlus HR	Vector	Static	Dist. to flowlines, waterbodies, coastline

DATA SOURCES

Social & Built Environment + Processing Notes

Source	Geography	Temporal	Key Variables
NDI (ACS)	Block Group	Annual	Neighborhood Deprivation Index
EPA Walkability	Block Group	Static (2016)	National Walkability Index
USDA FARA	Tract	Annual	64 food access variables + derived flags
FBI UCR	County	Annual	8 crime rates (murder, robbery, assault, etc.)
Census CBP/ZBP	County/ZCTA5	Annual	Business establishment counts by sector
TIGER Roads	Vector	Annual	Distance to primary (S1100) and secondary (S1200) roads

OUTPUT

Final Deliverable: Patient-Level Linked Exposures

PATID	Air Quality	Climate	UV	Greenness	Noise	Social	Proximity
PAT_1							
PAT_2							
PAT_3							

~100+ exposure variables per patient

N patients × ~100+ exposure variables across multiple domains

Spatial heterogeneity: buffer area weights blend across neighboring spatial units

Residential mobility: time-weighted averaging across address changes

R scripts (production) → **Python package (in progress)** with multi-processing → SPACESCANS Web backend

SPACESCANS Web Application

An intuitive web interface for the linkage pipeline

Ontology-Driven

The developed SPACEO ontology powers both the data catalog browsing and the variable selection in the task wizard — ensuring FAIR, machine-computable metadata throughout.

Pipeline Backend

The R-based pipeline is being rewritten as a Python package with multi-processing. This package serves as a general-purpose library and the computational engine behind the web interface.

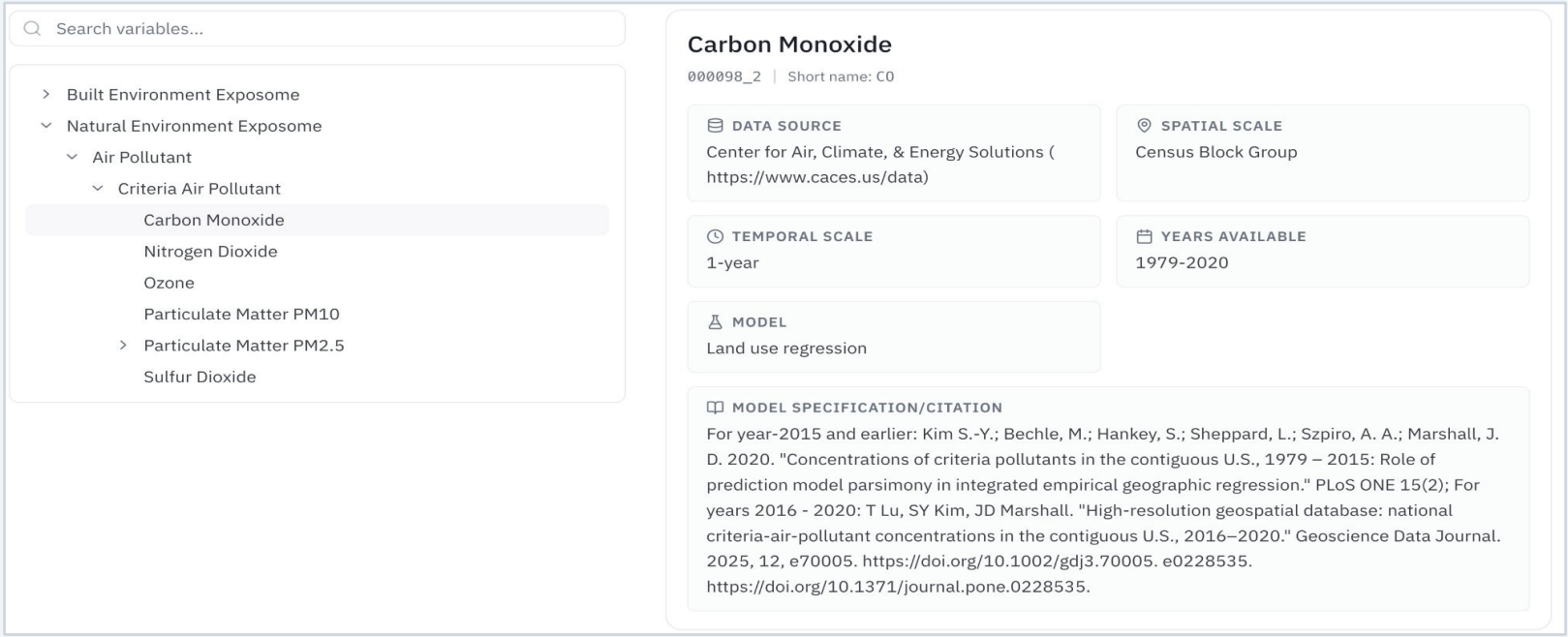
User-Friendly Interface

The web application provides researchers a point-and-click interface to upload data, configure buffers, select variables, and execute linkage — no coding required.

DATA CATALOG

Ontology-Guided Data Catalog

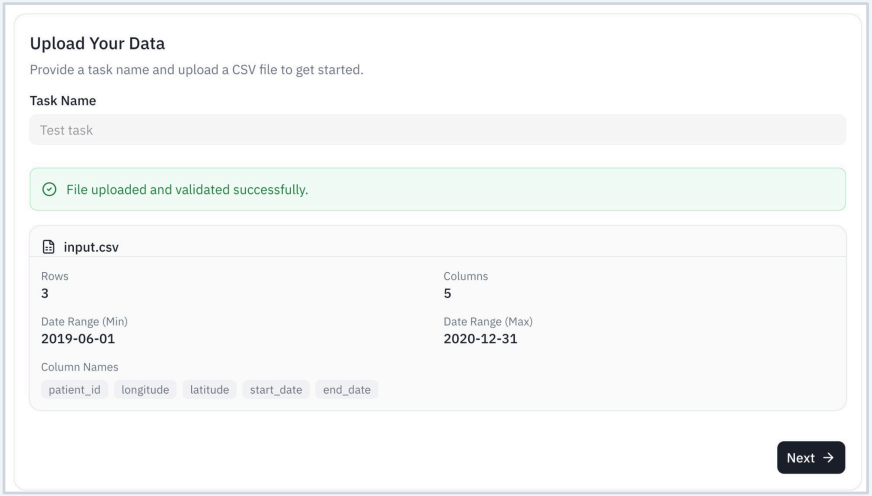
The SPACEO ontology structures both the data catalog browsing (left panel: ontology tree) and the variable detail metadata (right panel), ensuring consistent, FAIR-compliant variable definitions across the entire platform.



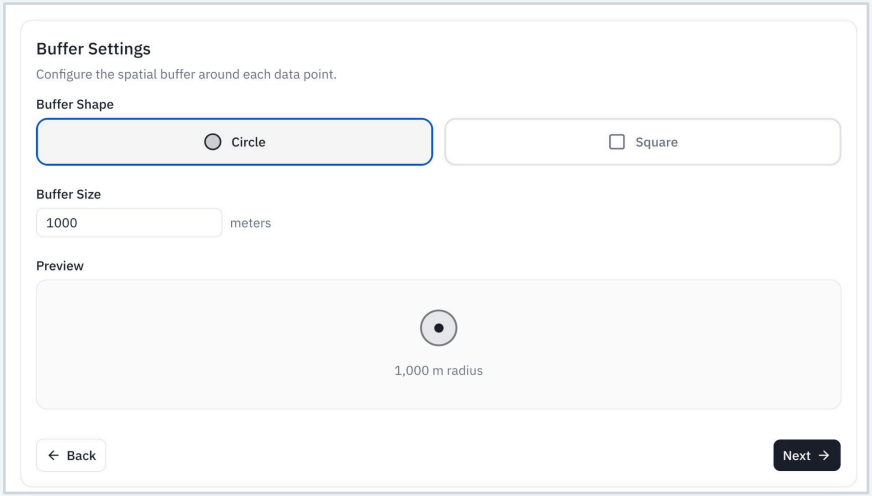
WORKFLOW

Task Wizard: Upload Data & Configure Buffer

1 Upload Data



2 Buffer Settings

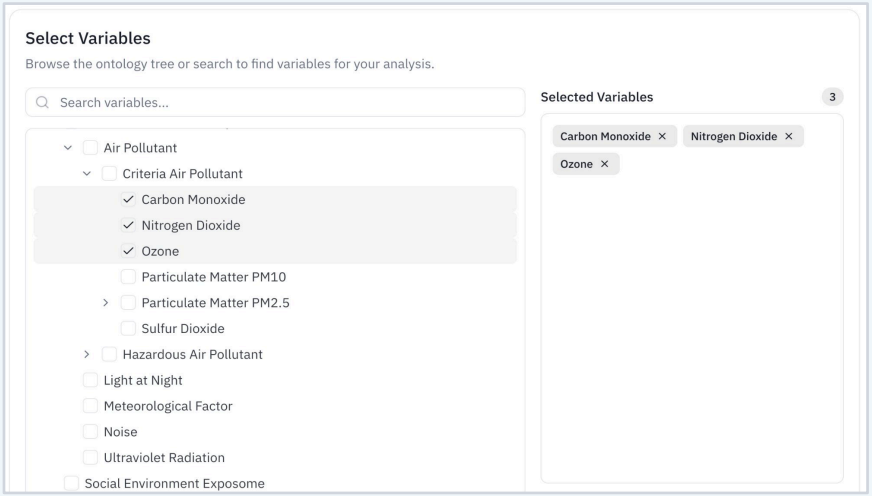


CSV upload with automatic validation and schema summary • Circle/square buffer with configurable radius in meters

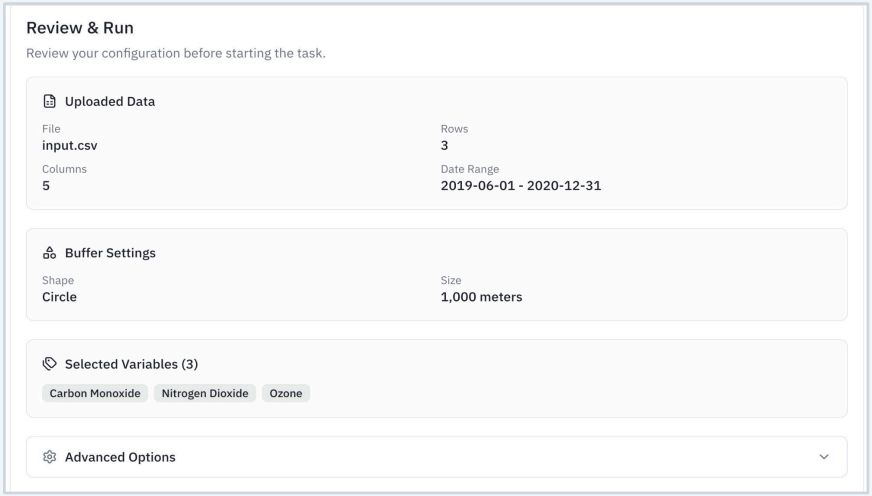
WORKFLOW

Task Wizard: Select Variables & Review

3 Select Variables



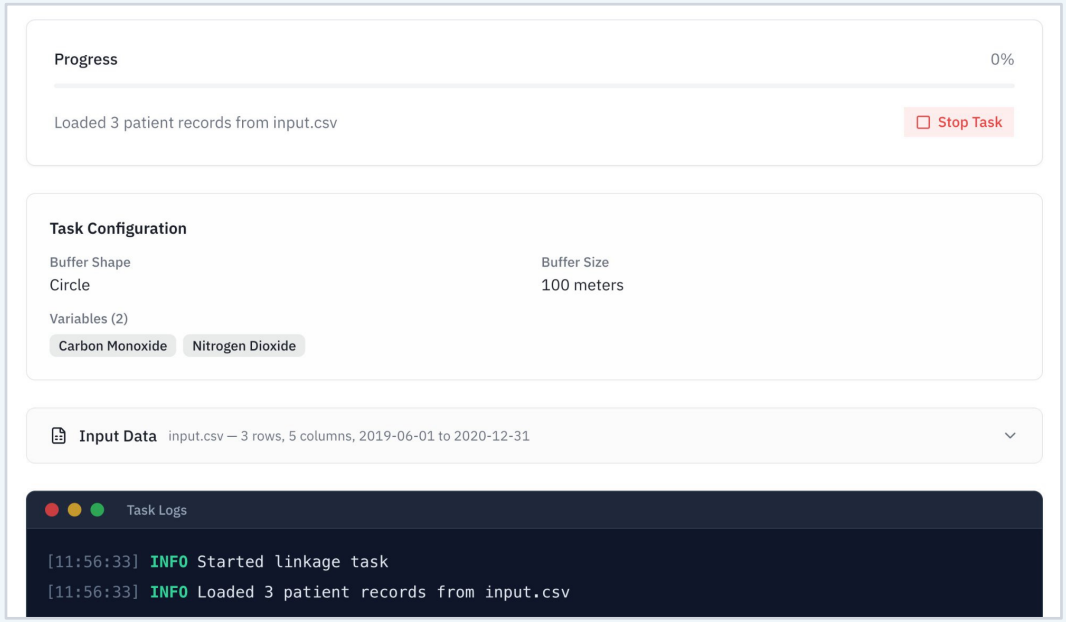
4 Review & Run



Ontology-driven variable tree with search and chip-based selection • Full configuration summary before pipeline launch

EXECUTION

Task Execution & Monitoring



Real-time progress

Live progress bar and status updates

Task configuration

View buffer shape, size, and selected variables

Live log streaming

Terminal-style task logs with timestamps

Download results

Analysis-ready linked dataset on completion

Stop / Retry

Controls for running or failed tasks

Summary

Pipeline

- 4-stage pipeline: Cohort → Geotagging → Buffer Weights → Exposure Linkage
- Patient-level TWA exposures across 100+ variables from 14+ data sources
- R scripts (production) → Python package with multi-processing (in progress)

Web Application

Browse data catalog via ontology-guided navigation (SPACEO)

Upload patient records with geospatial and temporal identifiers

Configure spatial/temporal buffers and select exposome variables

Execute reproducible linkage pipelines with real-time monitoring

Download analysis-ready linked datasets

Open-source | **MIT License** | NIH NIEHS R24ES036131

Team

Indiana University / Regenstrief Institute

- Jiang Bian, PhD - Principal Investigator
- Xing He, PhD - Faculty Lead, Strategic Data Science Infrastructure
- Yu Huang, PhD - Investigator
- Laura Ruppert, MPH - Director of Center Operations
- Sarah Zappone, PM - Project Manager
- Sneha Manoharan, MS - Project Manager

Mayo Clinic

- Cui Tao, PhD - Principal Investigator
- Lu Kang, MD - Project Manager
- Shuteng Niu, PhD - Project Lead Investigator – Biomedical Ontologies/Knowledge Graph Engineer
- Haifang Li, PhD - Investigator – Computational Ontologies development
- Huo Nan, PhD - Investigator – Ontologies development, data analysis

MGB / Harvard

- Hui Hu, PhD - Principal Investigator
- Jaime Hart, ScD - Investigator – environmental health, GIS
- Francine Laden, PhD - Investigator – environmental epidemiology
- Claire Leiser, PhD - Investigator – environmental epidemiology

HANDS ON WORKSHOP

Join Us at ISEE 2026

Pre-Conference Workshop on Spatial Exposome Linkage

Date and Location: Saturday, August 30, 2026; Munich, Germany

Time: 8:00 AM – 11:30 AM

Conference: ISEE 2026 — International Society for Environmental Epidemiology

Register & Details: https://www.isee26.org/pre-conference-workshops