London School of Hygiene and Tropical Medicine

Improving Health Worldwide

LONDON SCHOOL *of* HYGIENE &TROPICAL MEDICINE

Big data in environmental epidemiology

Arturo de la Cruz Libardi

Environment and Health Modelling (EHM) Lab
London School of Hygiene and Tropical Medicine

2024-03-13

intended learning outcomes

by the end of this session (lecture-demonstration), you will be able to:

1. Critically define big data
2. Describe some implications and applications of big data in public health and epidemiology
3. Evaluate sources of big health and environmental data
4. Think critically about data linkage in the context of exposure assessment

lecture outline

1. motivation

- brief history

2. big data

- definitions
- trends and implications
- epidemiology
- applications

3. health and environmental data

- source examples
- harmonization and modelling
- exposure assessment
- examples

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

big data → epidemiology → public health

LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

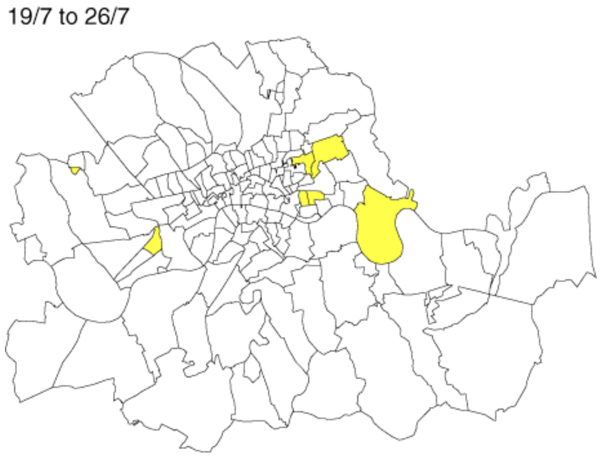
very brief history

integration of subject matter knowledge, (large scale) data, and analysis

from weekly burial counts (1662) to maps (1854) and death certificates to 180k cohort (1952)

enabled by technology, creativity, individual and collective effort

ink and paper, punch-cards, telephone…



the data line

big data

Variety

Volume

Velocity

everything else

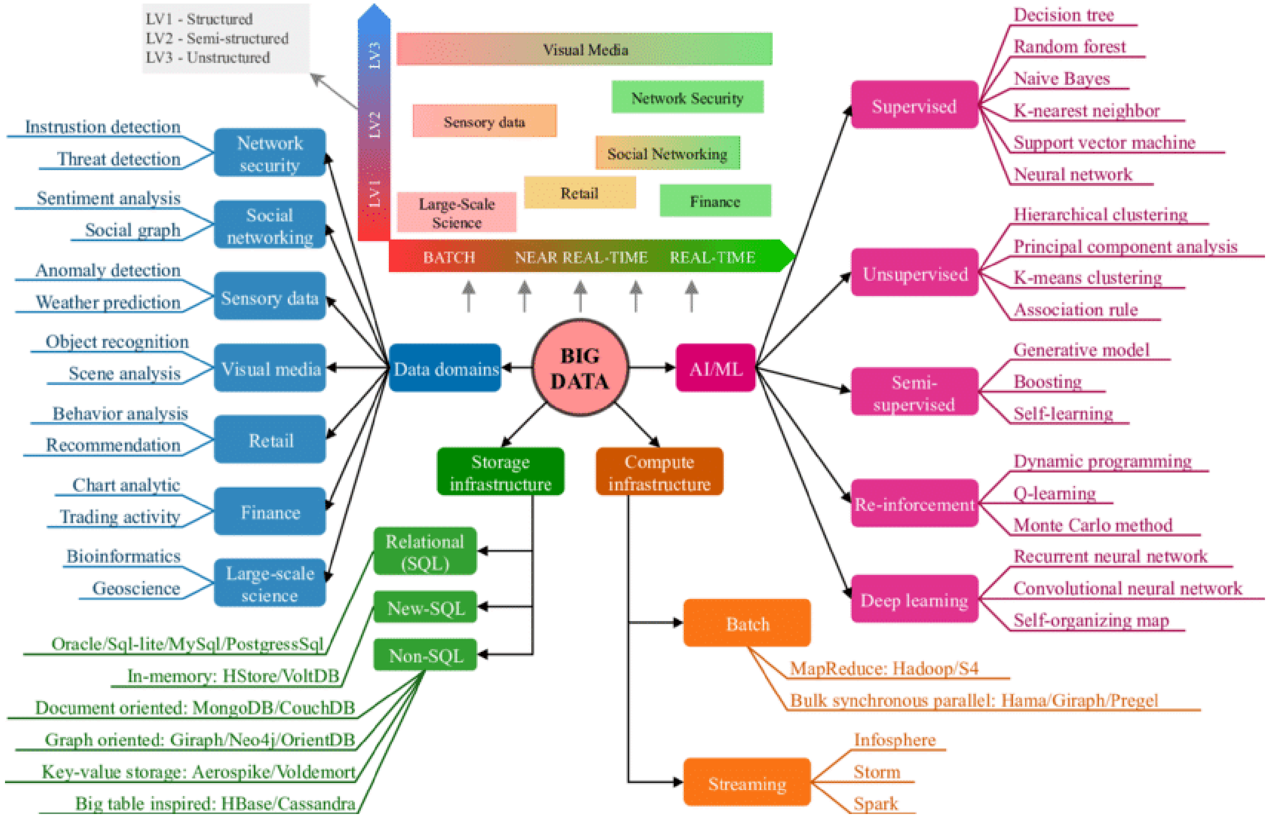
works on your machine

more V's?

is it just about data?

specialized infrastructure, pipelines and jargon

Data - oceans, lakes, warehouses, bases



Gadekallu, Pham, Huynh-The et al. [1]

concurrent global trends

- ageing population - urbanization (demographic change)
- environmentally complex climate change

and technological developments

- powerful unknowable functions (machine-learning)
- smart(ish), cheap(er) and pervasive monitoring

implications

- different training and emphasis
- widened research opportunities

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

big data in epidemiology - challenges and opportunities

Variety - measurement error, confounding etc...

Volume - wide and tall datasets, methods, coverage, power relevant research questions ...

Velocity - highest impact potential, most dependent on infrastructure

anything else?

applications in research and public health

- Research (genomics, EHRs)
- Healthcare administration (logistics)
- COVID-19 (emergency response, tracking, data sharing)
- in references

maternal thyroid function and autism with EHRs

- anatomical brain differences linked to autism are present at birth
- thyroid hormones play key role in brain development

Q 1: is hypothyroidism (ht) associated with inc. risk of autism? 400k births

Q 2: does risk for medicated mothers differ from Q1 ?

Q 3: does risk for lab-tested medicated mothers differ from Q1 and Q2 risks ?

Q 4: are levels of TSH/fT4 associated with inc. risk of autism ? 50k births

- other exposure increases risk of ht and autism: ht → autism association is not causal
- mothers with hypothyroidism have other issues that increase risk of autism

Rotem, Chodick, Shalev et al. [6]

OpenSAFELY

OpenSAFELY: the origin story

> On 7th May 2020, the OpenSAFELY Collaborative pre-printed the world's largest study into factors associated with death from Covid-19, based on an analysis running across the full pseudonymised health records of 40% of the English population. This is an unprecedented scale of data.
>
> … a huge collaboration including the Bennett Institute for Applied Data Science at the University of Oxford, the EHR research group at London School of Hygiene and Tropical Medicine, NHS England, and TPP. Over 42 days during the peak of the first wave of COVID-19 this team worked day and night to produce a fully open-source, privacy-preserving software platform, capable of running open and reproducible analytics across electronic health records, all held securely in situ. Since then the OpenSAFELY platform has expanded to a full scale analytic environment for secure data analysis, reproducible data curation, federated analysis, and code sharing, with every line of code for the platform, for data management, and for data analysis all shared openly by default, in re-usable forms, automatically, and without exception.

Couldn't load plug-in.

All LSHTM OpenSAFELY projects

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

health data sources

- datasets: ProjectTycho
- cohorts: BioBank and OurFutureHealth
- platforms: CPRD and OpenSAFELY
- raw wearable sensor data

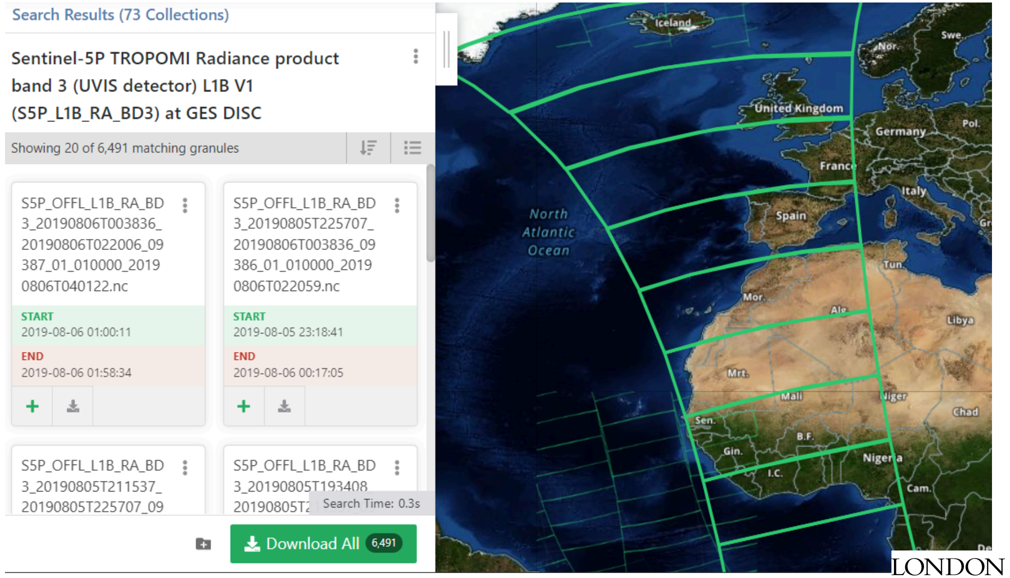
Table 1
Examples of storage needs (per person).

Information	Size
Human genome	1 GB
+ structure determination of the proteins	Several PB[a]
Electronic health record	1 MB-5GB, expected to increase 50-fold from 2012 to 2020[b,c]
Heart rate monitor (per month)	9 GB[d]
Continuous video life-logger (per month)	58GB[e]
Accelerometer (8-h a day, per month)	1 GB[f]
Medical image	MB to GB, up to 1 TB. e.g. 64/128-slice CT scan, 3.0 T MRI and PET often exceeding 100 MB[b].

Tonne, Basagaña, Chaix et al. [2]

environmental data*

- modelled: atmospheric dispersion models, reanalysis, digital twins
- raw: ground monitors, mobile sensors, satellites
- raster vs vector



*some environmental data will already be part of EHRs

** figure from: https://search.earthdata.nasa.gov

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

why use both environment and health data

- a part of disease etiology remains unexplained and is likely due to the environment
- big data processes offer great potential for environmental health research
- public health importance should drive research questions

environment + health data synergy

1. research question
2. **get data**
3. **prepare data**
4. **link health and exposures**
5. analyse

LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

from data to exposure

Data preparation

(complexity)

(none) → continuous modelled output

(simple) → inverse distance weighted surface from point measurement

(complex) → multi-stage spatio-temporal machine-learning ensemble modelling using harmonized features

Linkage

(simple) → matching nearest

(medium) → aggregate over small area

(medium) → from more than one coordinates (with bilinear interpolation) [3]

(complex) → from a trajectory accounting for microenvironments [4]

Vanoli, Mistry, De La Cruz Libardi et al. [3] Smith, Mitsakou, Kitwiroon et al. [4]

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

A Satellite-Based Spatio-Temporal Machine Learning Model to Reconstruct Daily PM2.5 Concentrations across Great Britain [5]

- Ground observations of $PM_{2.5}$
- A lot of environmental data
- Random forest algorithms

Schneider, Vicedo-Cabrera, Sera et al. [5]

A Satellite-Based Spatio-Temporal Machine Learning Model to Reconstruct Daily PM2.5 Concentrations across Great Britain [5]

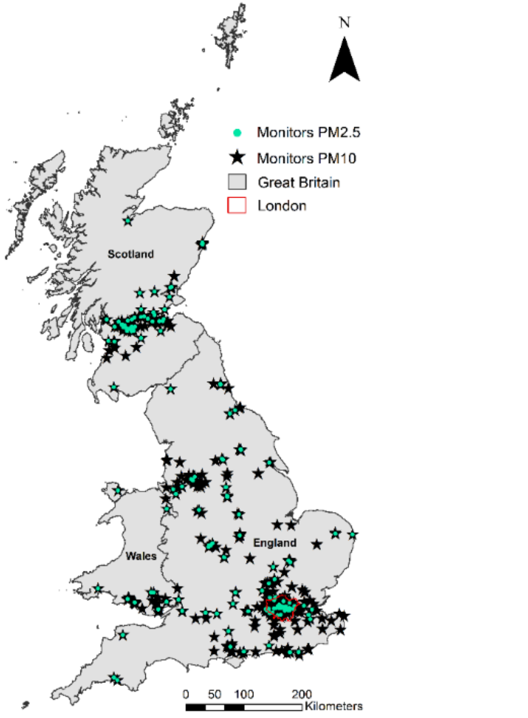


Figure 1. Spatial distribution of 581 PM_{10} (black star) and 183 $PM_{2.5}$ (turquoise dots) monitors across Great Britain during the study period.

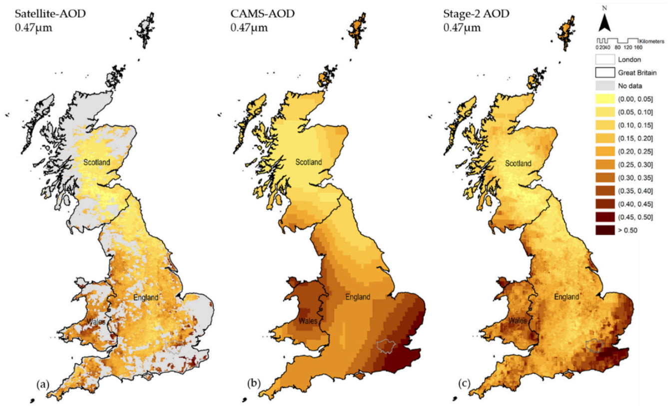


Figure 2. Satellite-AOD 0.47 µm values are represented in Figure 2a (mean of all Terra- and

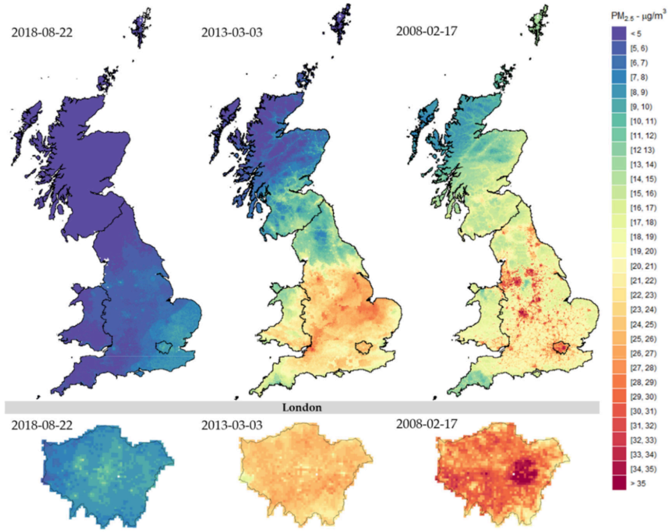


Figure 5. Stage-4 day-specific $PM_{2.5}$ estimations across Great Britain (Top) and London (...

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London Hybrid Exposure Model: Improving Human Exposure Estimates to NO2 and PM2.5 in an Urban Setting [4]

> …the London Hybrid Exposure Model (LHEM), (…) calculates exposure of the Greater London population to outdoor air pollution sources, in-buildings, in-vehicles, and outdoors, using survey data of when and where people spend their time.

- London Travel Demand Survey, trip route simulation

> Exposure to outdoor air pollution was provided by CMAQ-urban, which couples the Weather Research and Forecasting (WRF) meteorological model, the Community Multiscale Air Quality (CMAQ) regional scale model, and the Atmospheric Dispersion Modeling System (ADMS) roads model

- I/O ratio for indoor air levels
- for in-vehicle levels: $\frac{dC_{in}}{dt} = \lambda_{in}(C_{out} - C_{in}) - n\lambda_{HVAV} \cdot C_{in} - V_g(\frac{A^*}{V}) \cdot C_{in} + \frac{Q}{V}$
- constant value for the underground
- **"microenvironments"**

Smith, Mitsakou, Kitwiroon et al. [4]

London Hybrid Exposure Model: Improving Human Exposure Estimates to NO2 and PM2.5 in an Urban Setting [4]

residential vs modelled exposure

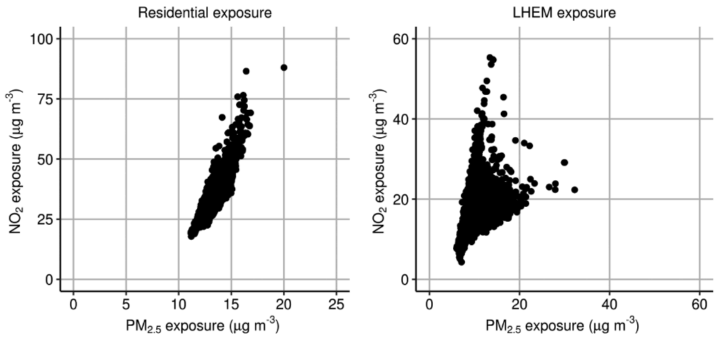


Figure 4. Scatter plots of NO_2 and $PM_{2.5}$ exposure outdoors at the residential address (left) ($R = 0.90$) and using the LHEM (right).

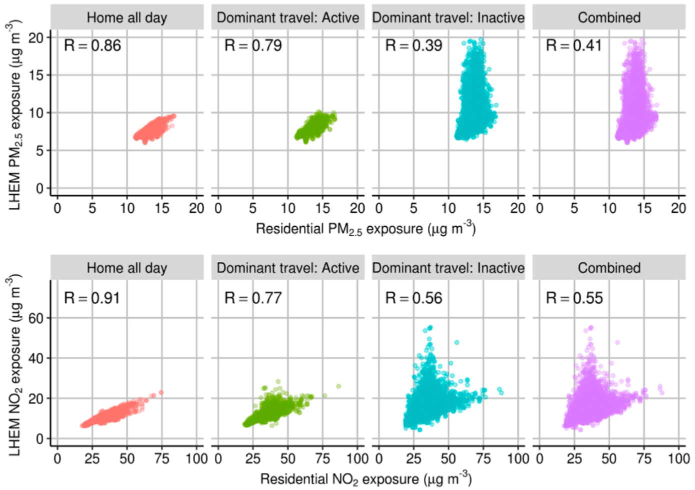
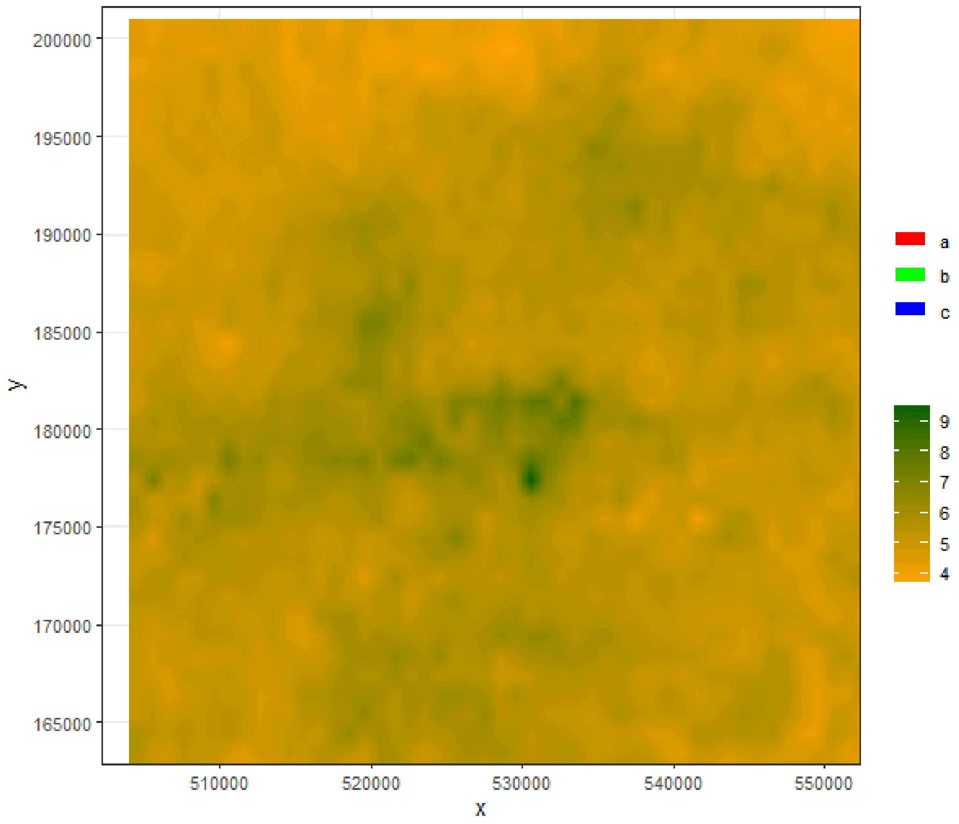
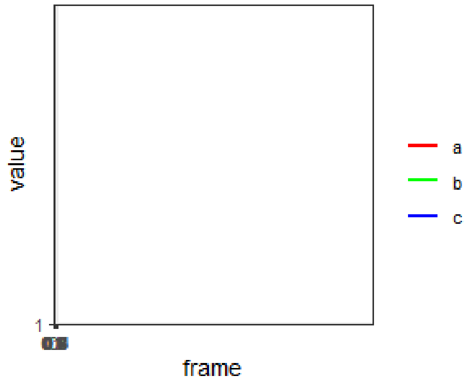
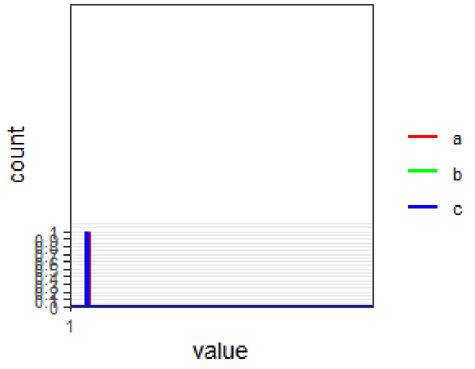


Figure 3. Scatter plots of NO_2 (bottom) and $PM_{2.5}$ (top) LHEM exposure versus exposure at the residential address - demonstrating the relative strength of the relationship between those who undertake active travel (cycle and walk), those that stay at home, and those who undertake inactive travel (car, motorcycle, bus, train, and tube).


200000
195000
190000
185000
180000
175000
170000
165000
y
510000
520000
530000
540000
550000
x
a
b
c
9
8
7
6
5
4


value
frame
1
a
b
c

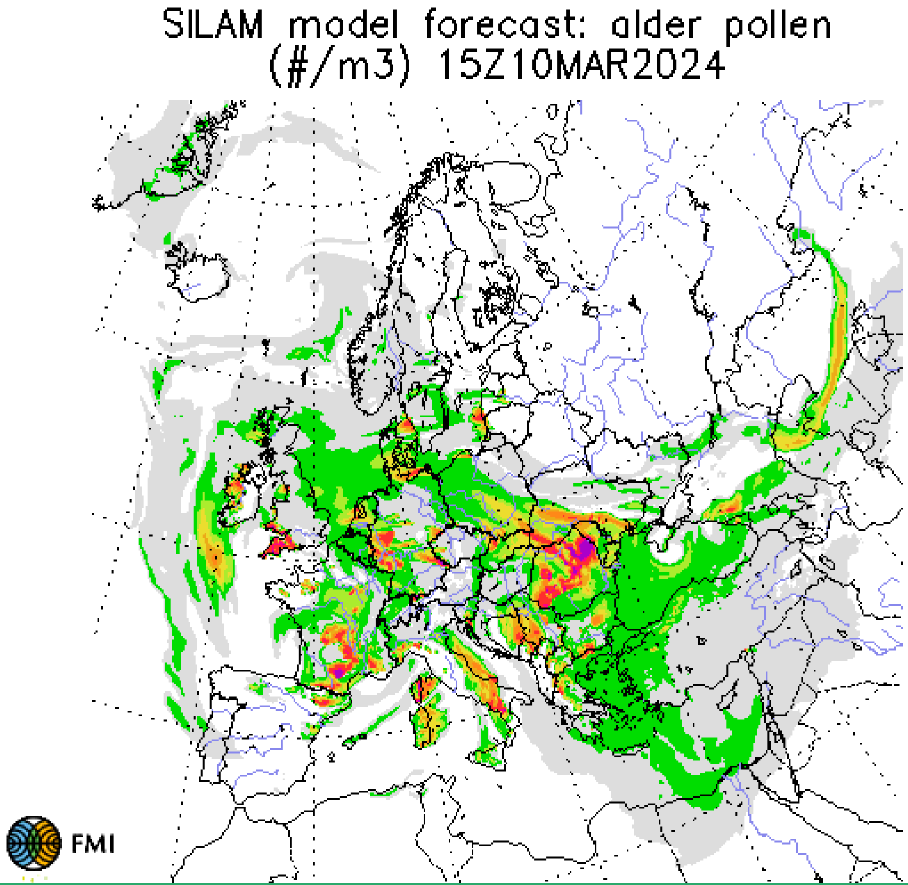

count
value
1
0
a
b
c

SILAM v.5.7

FINNISH METEOROLOGICAL INSTITUTE (http://en.ilmatieteenlaitos.fi/)

System for Integrated modeLling of Atmospheric coMposition

Region: europe



Control panel

Pollen taxon:

Alder | Birch | Grass

Olive | Ragweed | Mugwort

Pollen Index | Allergy Risk

Region:

Europe

Hour: **-57**

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

OpenSAFELY query (ehrQL) reliability testing using generative artificial intelligence!

we have learned to

1. Critically define big data as big data processes

2. Describe some implications and applications of big data in public health and epidemiology

- **classical (measurement error, confounding) challenges**
- **new (comprehensive health data, real-time action) opportunities**

3. Evaluate sources of big health and environmental data

- **health genetic data, EHRs, wearable sensors**
- **environment reanalyses, satellites, ground sensors**

4. Think critically about data linkage in the context of exposure assessment

- **depends on individual/areal nature of health data and resolution of environmental data**

references

[1] T. R. Gadekallu, Q. Pham, T. Huynh-The, et al. *Federated Learning for Big Data: A Survey on Opportunities, Applications, and Future Directions*. En. 2021.

[2] C. Tonne, X. Basagaña, B. Chaix, et al. "New frontiers for environmental epidemiology in a changing world". In: *Environment International* 104 (Jul. 2017), pp. 155-162. ISSN: 0160-4120. DOI: 10.1016/j.envint.2017.04.003. URL: https://www.sciencedirect.com/science/article/pii/S0160412017301459 (visited on 02/15/2024).

[3] J. Vanoli, M. N. Mistry, A. De La Cruz Libardi, et al. "Reconstructing individual-level exposures in cohort analyses of environmental risks: an example with the UK Biobank". En. In: *Journal of Exposure Science & Environmental Epidemiology* (Jan. 2024). ISSN: 1559-0631, 1559-064X. DOI: 10.1038/s41370-023-00635-w. URL: https://www.nature.com/articles/s41370-023-00635-w (visited on 03/10/2024).

[4] J. D. Smith, C. Mitsakou, N. Kitwiroon, et al. "London Hybrid Exposure Model: Improving Human Exposure Estimates to NO2 and PM2.5 in an Urban Setting". En. In: *Environmental Science & Technology* 50.21 (Nov. 2016), pp. 11760-11768. ISSN: 0013-936X, 1520-5851. DOI: 10.1021/acs.est.6b01817. URL: https://pubs.acs.org/doi/10.1021/acs.est.6b01817 (visited on 02/02/2023).

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

references

[5] R. Schneider, A. Vicedo-Cabrera, F. Sera, et al. "A Satellite-Based Spatio-Temporal Machine Learning Model to Reconstruct Daily PM2.5 Concentrations across Great Britain". En. In: *Remote Sensing* 12.22 (Nov. 2020), p. 3803. ISSN: 2072-4292. DOI: 10.3390/rs12223803. URL: https://www.mdpi.com/2072-4292/12/22/3803 (visited on 02/03/2022).

[6] R. S. Rotem, G. Chodick, V. Shalev, et al. "Maternal Thyroid Disorders and Risk of Autism Spectrum Disorder in Progeny". En-US. In: *Epidemiology* 31.3 (May. 2020), p. 409. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000001174. URL: https://journals.lww.com/epidem/fulltext/2020/05000/maternal_thyroid_disorders_and_risk_of_autism.15.aspx (visited on 03/13/2024).

[7] D. Cox, C. Kartsonaki, and R. H. Keogh. "Big data: Some statistical issues". In: *Statistics & Probability Letters* 136 (May. 2018), pp. 111-115. ISSN: 0167-7152. DOI: 10.1016/j.spl.2018.02.015. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5992743/ (visited on 03/11/2024).

[8] E. J. Williamson, A. J. Walker, K. Bhaskaran, et al. "Factors associated with COVID-19-related death using OpenSAFELY". En. In: *Nature* 584.7821 (Aug. 2020). Publisher: Nature Publishing Group, pp. 430-436. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2521-4. URL: https://www.nature.com/articles/s41586-020-2521-4 (visited on 03/12/2024).

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

references

[9] M. J. Khoury and J. P. A. Ioannidis. "Big data meets public health". En. In: *Science* 346.6213 (Nov. 2014), pp. 1054-1055. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaa2709. URL: https://www.science.org/doi/10.1126/science.aaa2709 (visited on 03/08/2024).

[10] S. J. Mooney and V. Pejaver. "Big Data in Public Health: Terminology, Machine Learning, and Privacy". In: *Annual Review of Public Health* 39.1 (2018). _ eprint: https://doi.org/10.1146/annurev-publhealth-040617-014208, pp. 95-112. DOI: 10.1146/annurev-publhealth-040617-014208. URL: https://doi.org/10.1146/annurev-publhealth-040617-014208 (visited on 03/05/2024).

other info

Presentation made with xaringan in RStudio.

Contact: Arturo.de-la-Cruz-Libardi@lshtm.ac.uk Slides:

LONDON SCHOOL *of* HYGIENE &TROPICAL MEDICINE

suggestions?

- DASH 26th March opening event
- hundreds of hours of free and open resources
- a lot of local and global circumstances to improve