

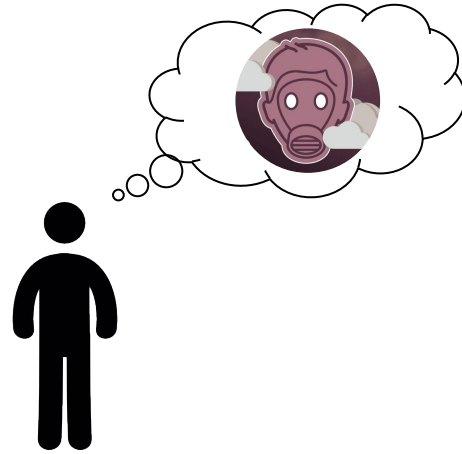
How to conduct a big data analysis on air pollution and health?

Mathematical Institute, Serbian Academy of Arts and Sciences
Mart 29, 2022

Ana Trisovic, Harvard University

How to conduct a big data analysis on air pollution and health?

How to conduct a big data analysis on air pollution and health?



Analysis design

- Secondary data analysis - using data from existing data sources, integrating it and applying study design
- Hypothesis-driven - used to answer presupposed hypothesis or a research question

Exposure - hypothesized cause of the disease

Nitrogen dioxide NO ₂
Ozone O ₃
Particulate Matter PM _{2.5}



Air Quality Data for Health-Related Applications

Collection Overview

Data Sets (2)

*Daily and Annual
PM_{2.5} Concentrations
for the Contiguous
United States, 1-km
Grids, v1 (2000–
2016)*

[+ Show All...](#)

Daily and Annual PM2.5 Concentrations for the Contiguous United States, 1-km Grids, v1 (2000–2016)

Set Overview

Data Download

Documentation

Metadata

Downloads

Data:

View Recommended Citation(s)

This data set consists of daily and annual PM2.5 concentrations for the contiguous United States at 1 km spatial resolution for the years 2000 to 2016 in both GeoTIFF and RDS formats. The daily data are packaged by month in the table below.

The R code and associated inputs used to match latitude/longitude coordinates stored in RDS format (also provided as a Shapefile) to the O3 data in RDS format, and convert to GeoTIFF are [provided here](#) [80 MB zip file].

Annual PM_{2.5} concentration data for the contiguous United States:

GeoTIFF [523 MB zip file]

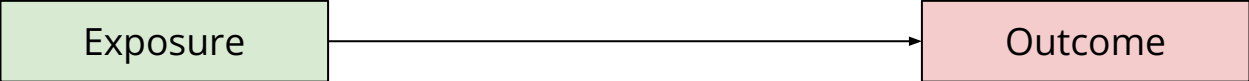
RDS [1.4 GB zip file]

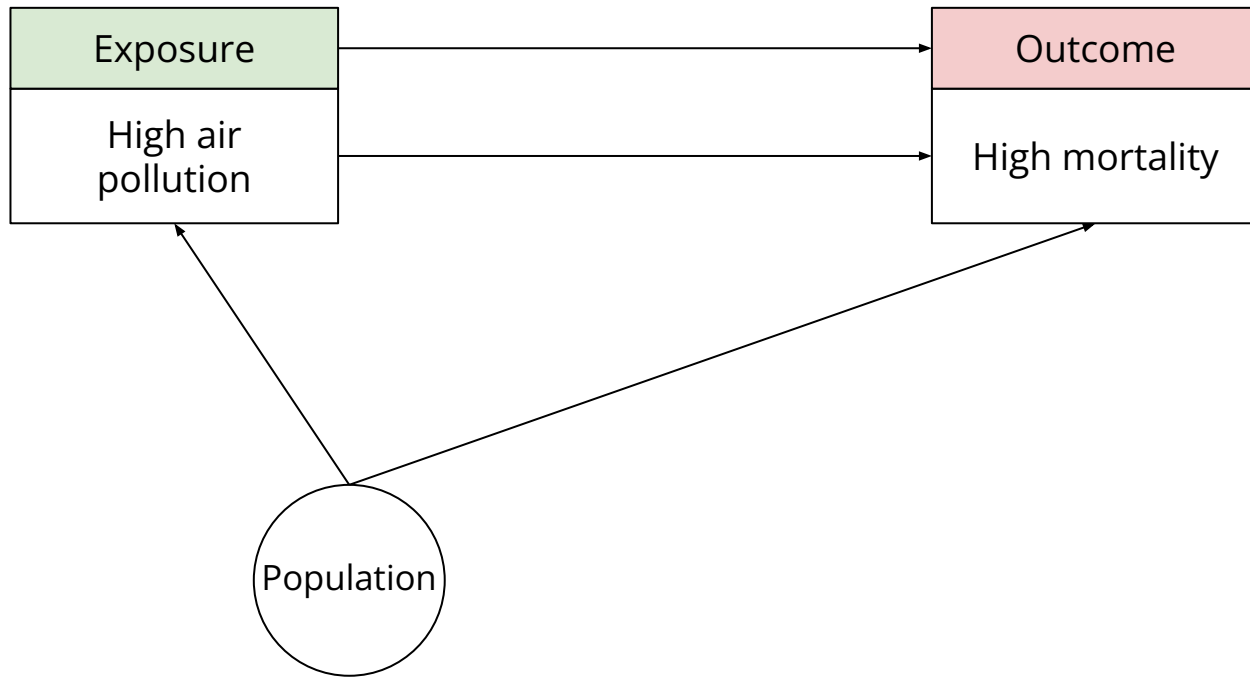
Daily PM_{2.5} concentrations in GeoTIFF (~900 MB/file) and RDS (~2.5 GB/file) packaged by month and year:

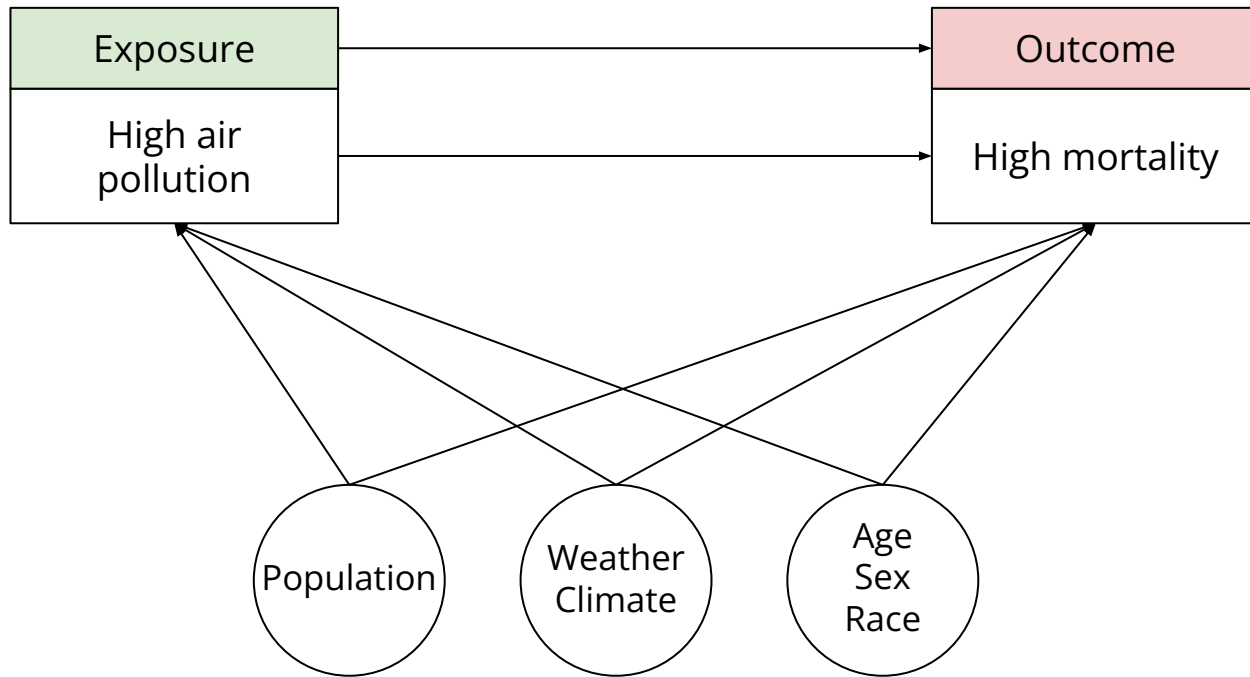
[illegible]

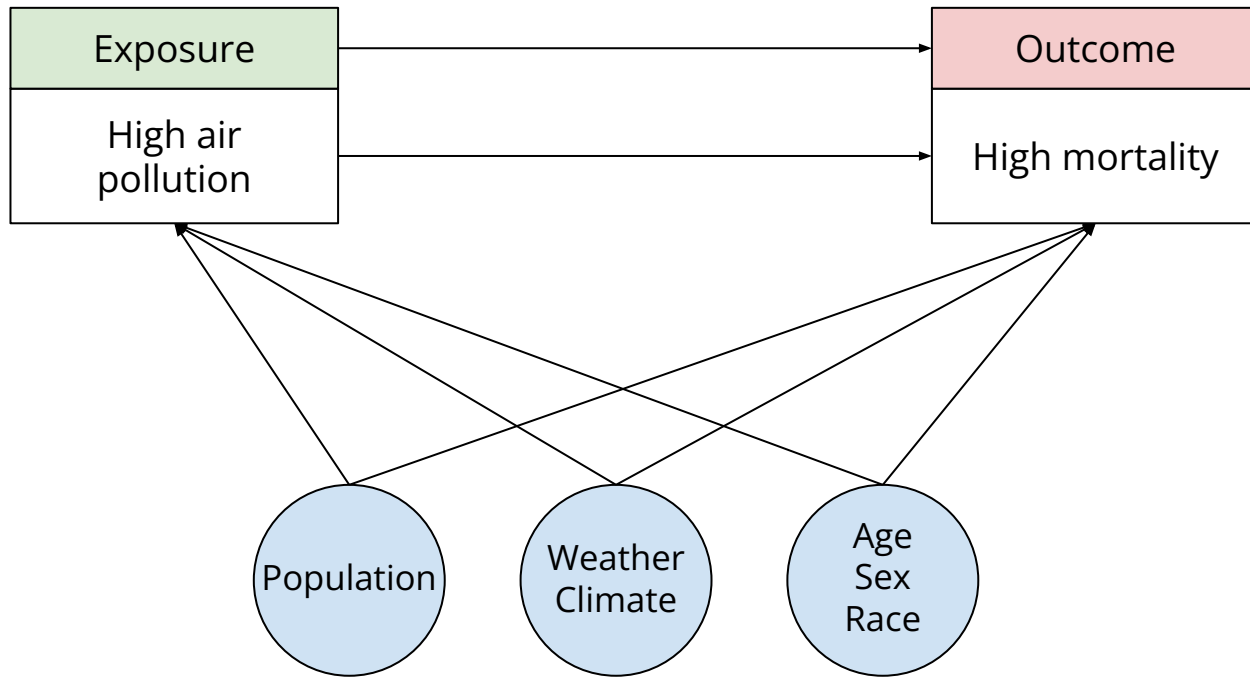
Outcome - hypothesized to have a causal relationship with exposure

- Medicare administrative data, also known as health services utilization data, are collected by the Centers for Medicare and Medicaid Services (CMS) and derived from reimbursement information or the payment of bills.
 - Medical diagnosis
 - Mortality









Confounder

Census



Centers for Disease
Control and Prevention



Climatology Lab



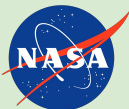
Outcome

Centers for Medicare
and Medicaid Services



Exposure

Air Quality



Confounder

Census



Centers for Disease
Control and Prevention



Climatology Lab



Independent
Variable

Outcome

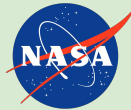
Centers for Medicare
and Medicaid Services



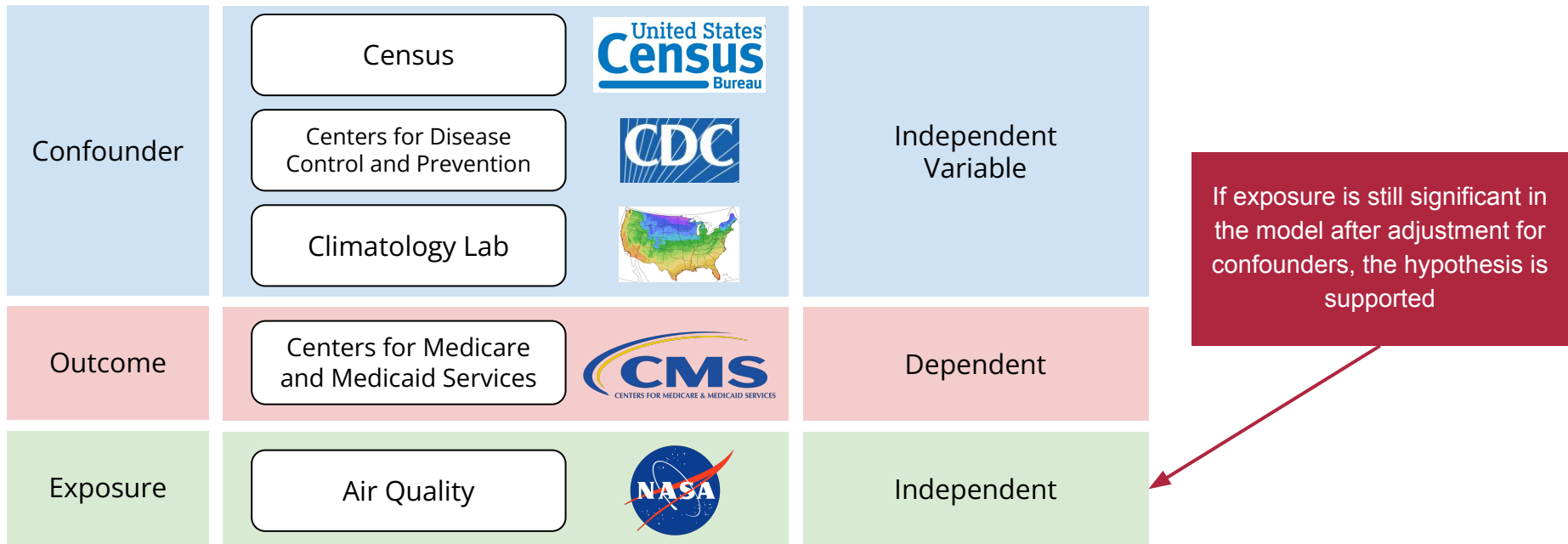
Dependent

Exposure

Air Quality



Independent



Confounder

Census



Centers for Disease
Control and Prevention



Climatology Lab



Independent
Variable

If exposure is still significant in
the model after adjustment for
confounders, the hypothesis is
supported

Outcome

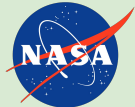
Centers for Medicare
and Medicaid Services



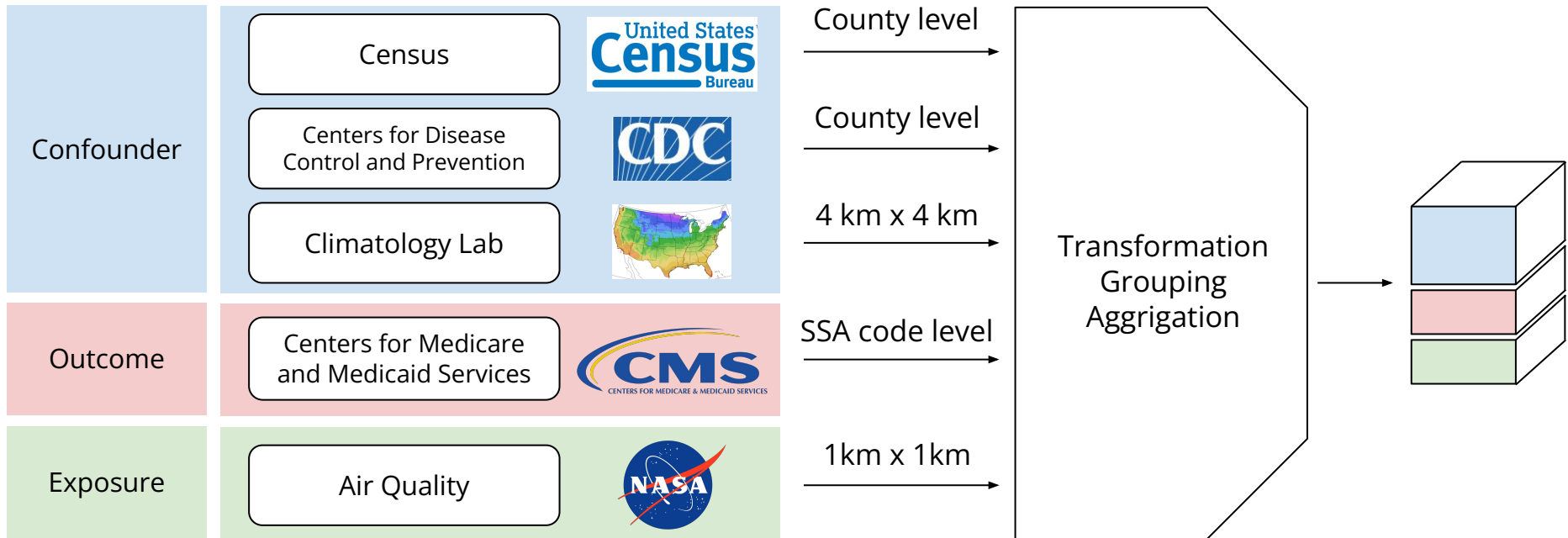
Dependent

Exposure

Air Quality



Independent



Troubles with data

- Temporal and spatial resolution
 - Different in different datasets
- Missing data
 - More-or-less every dataset
- Inconsistent data
 - I.e., single person having inconsistent information on age/sex/race
- Badly formatted data
 - I.e., medical, satellite

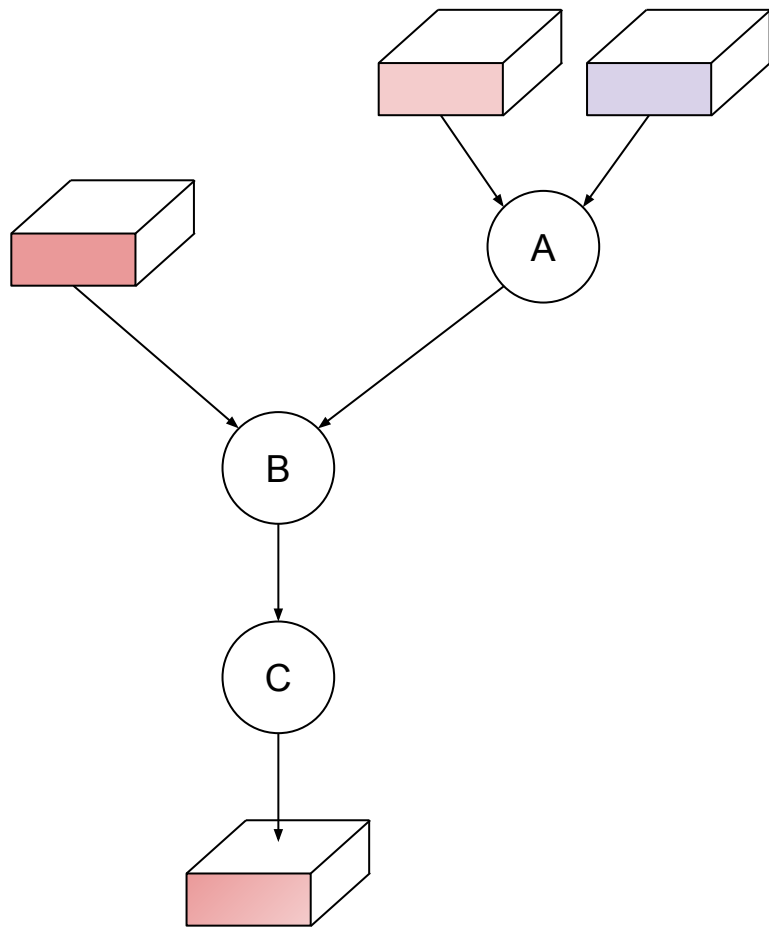
Missing data

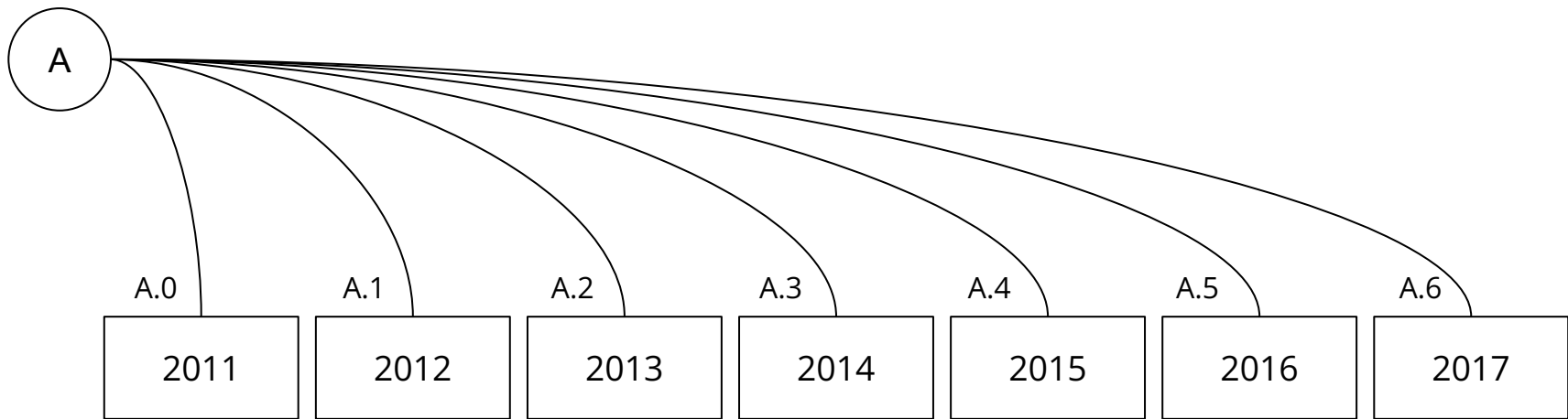
- Exclude records with missing values if they measure subpopulation, exposure or outcome
 - If there are few records with missing values (<5% of records)
- Don't exclude if they measure a confounder
- If there are over 5% of the records with missing values on subpopulation, exposure, outcome or important confounder - rethink study design

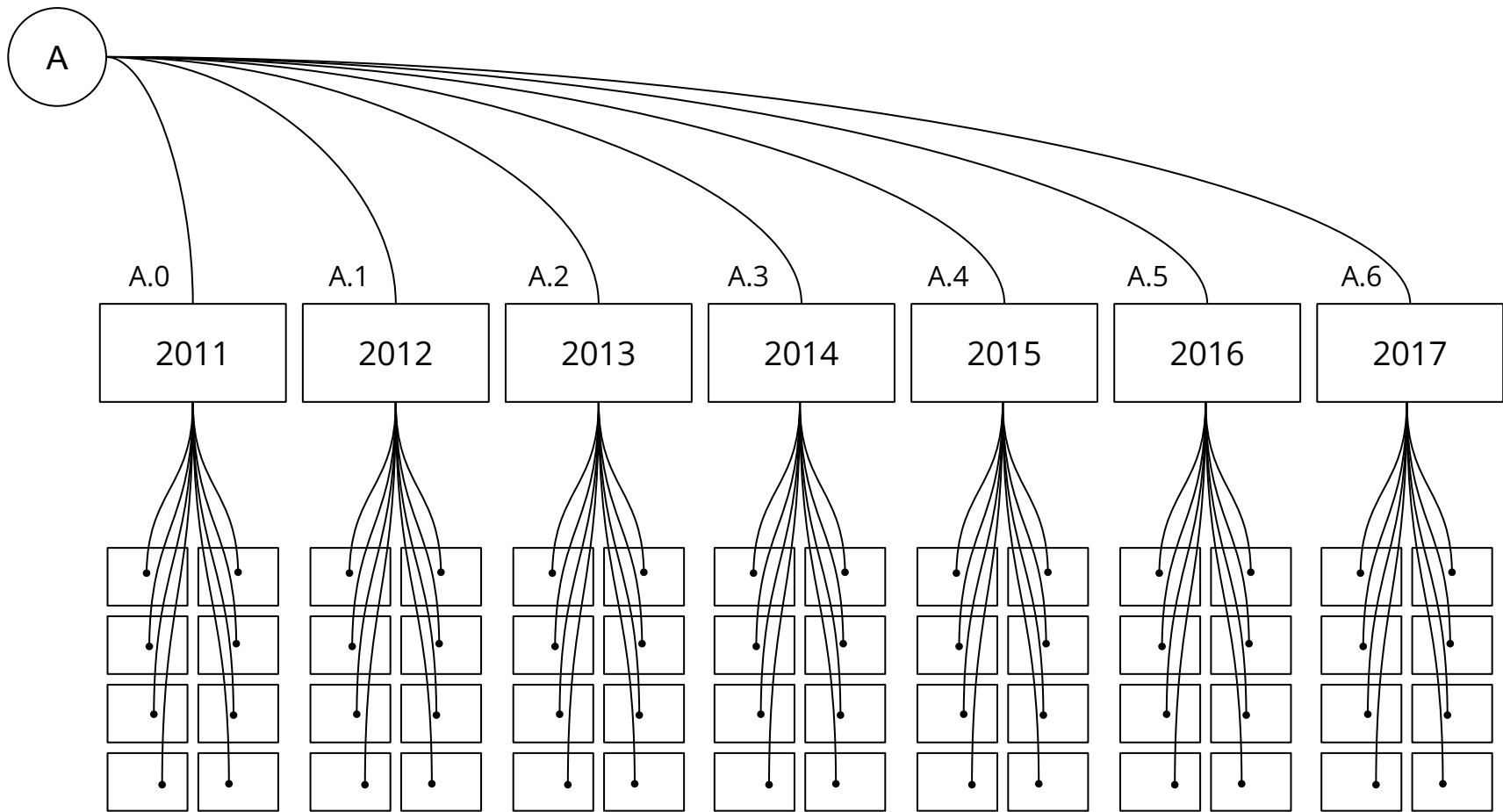
Data sources

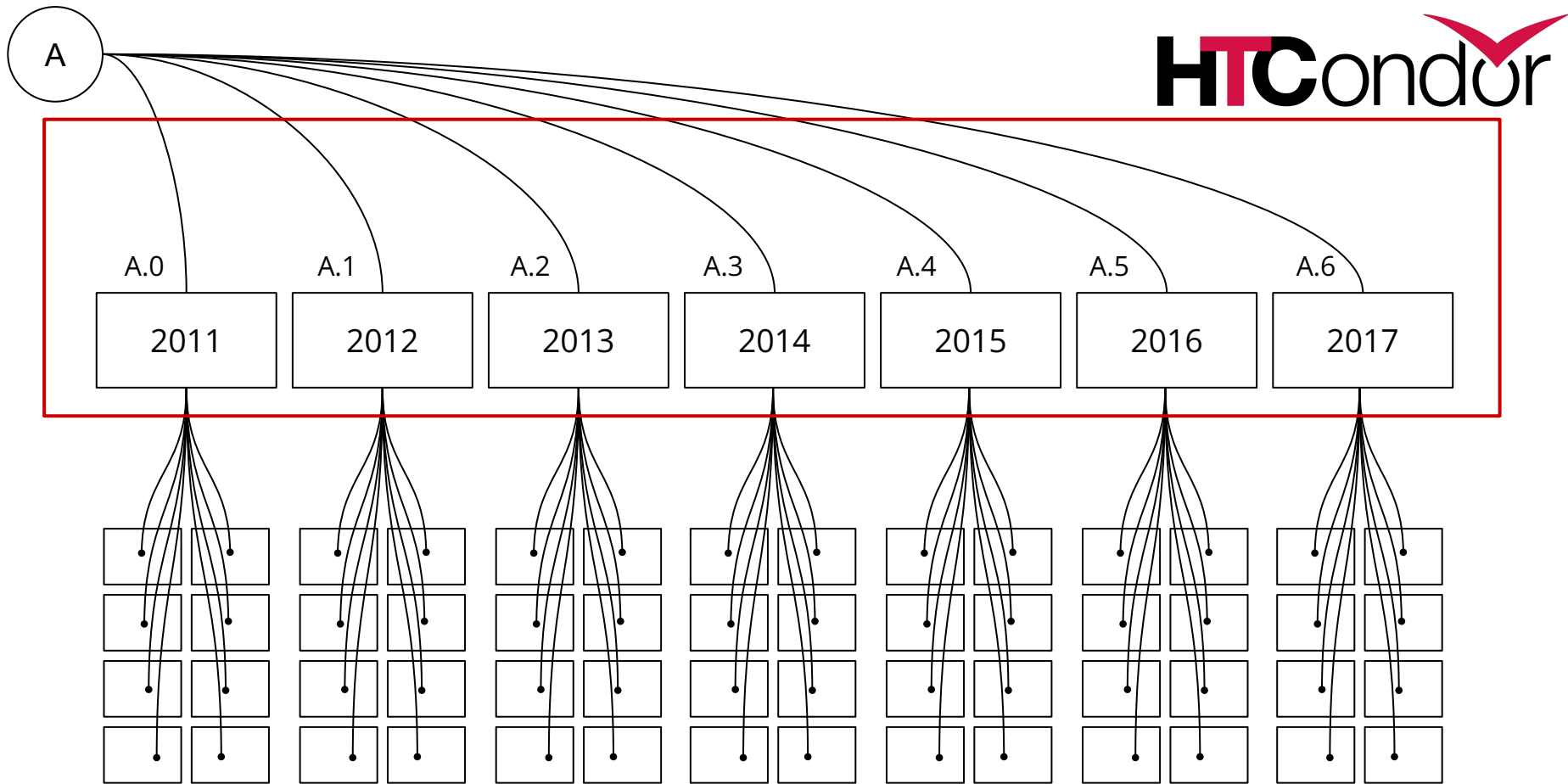
Summary

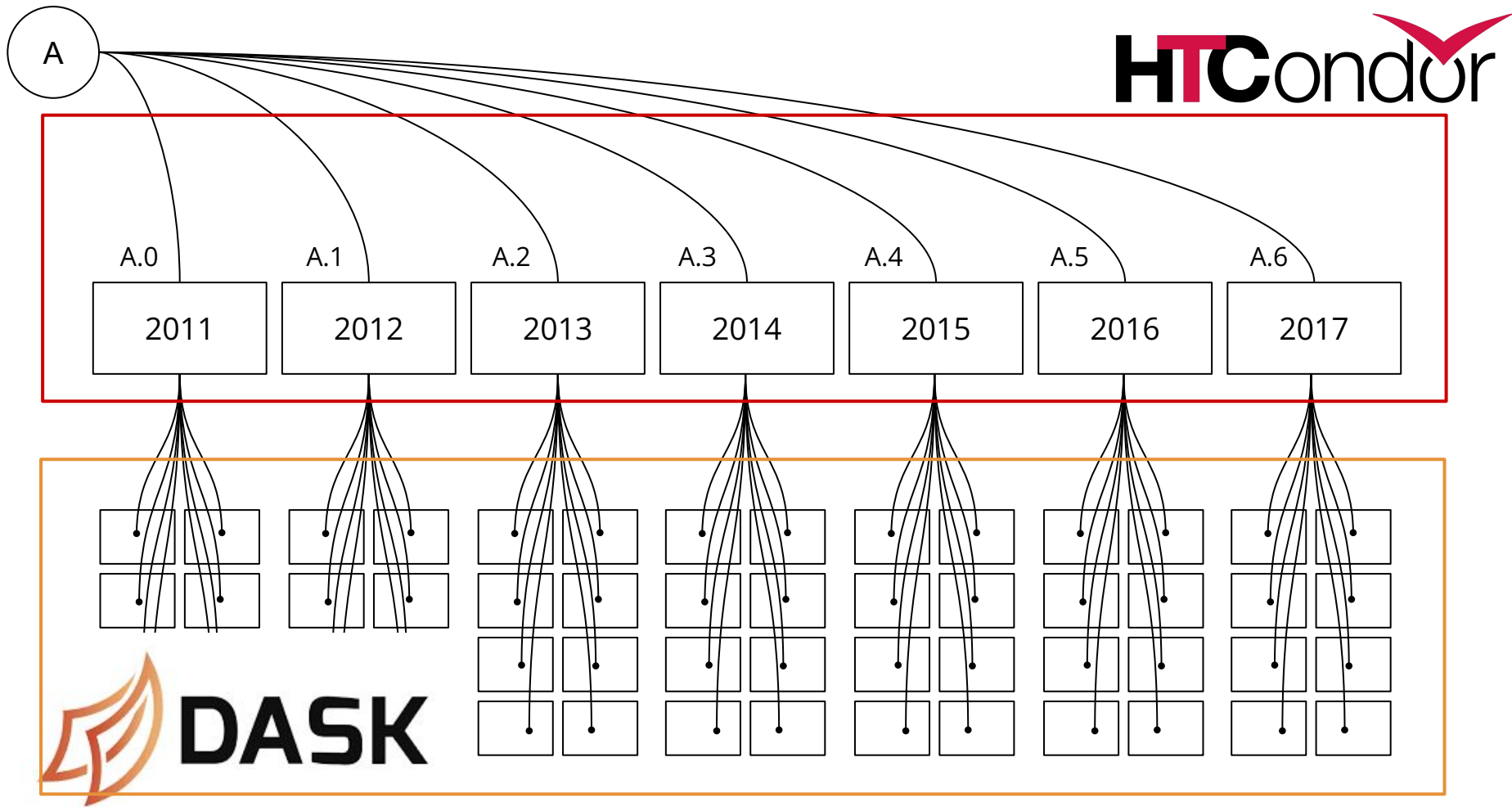
- Hypothesis-driven epidemiological data analyses require data for exposures, outcomes and confounders
 - Secondary data analysis often require data cleaning, transformation and aggregation
-

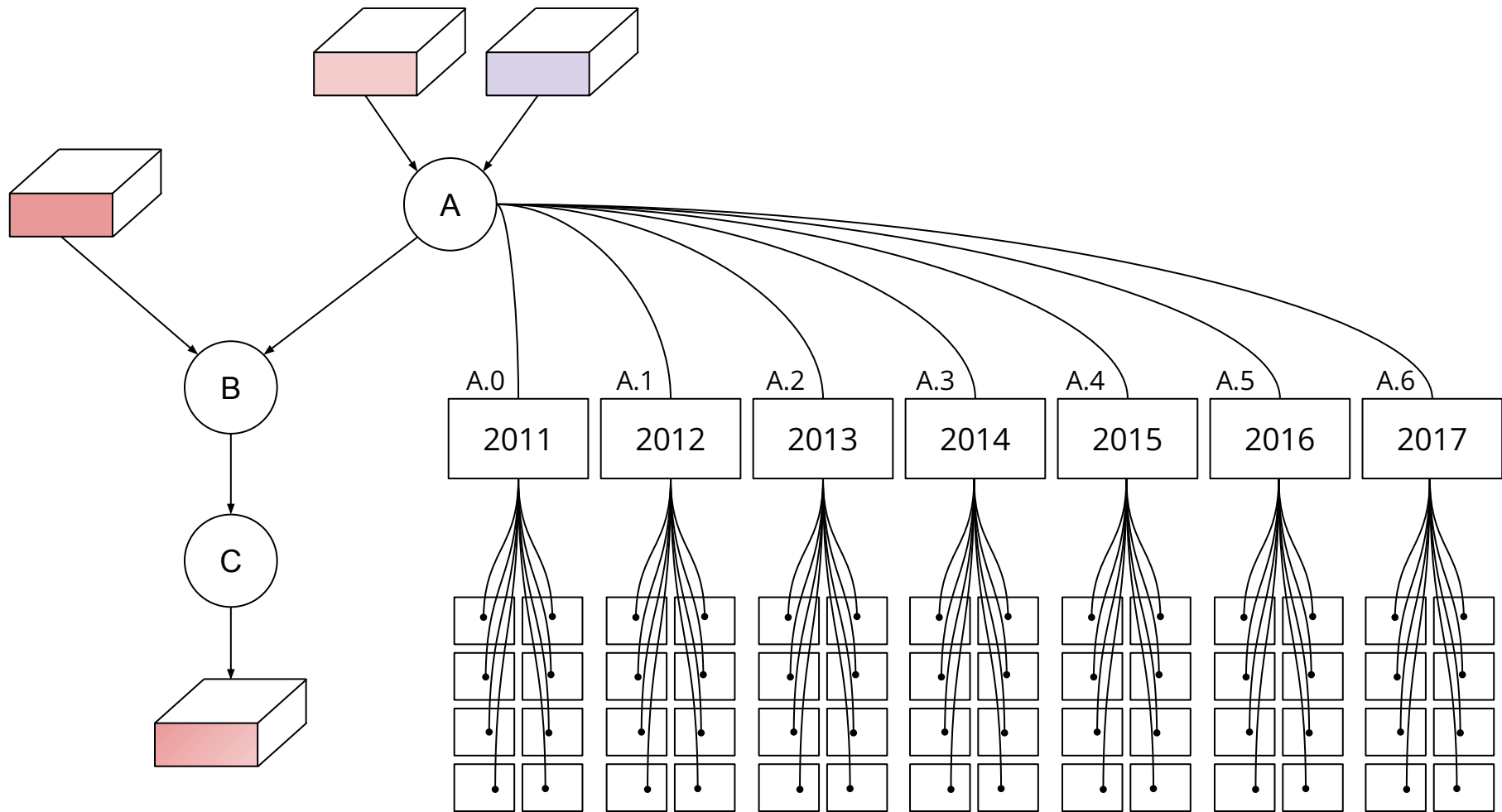


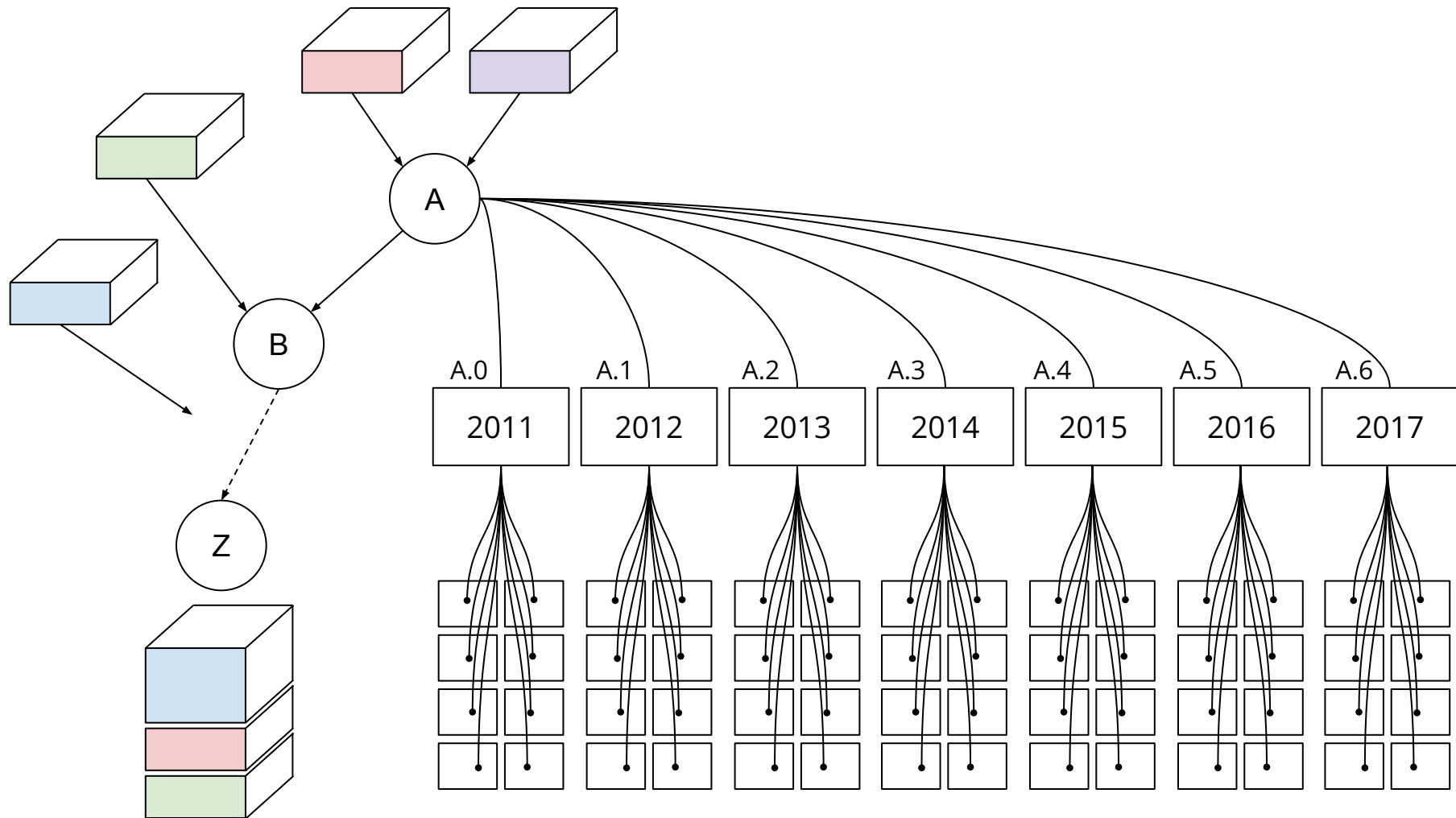


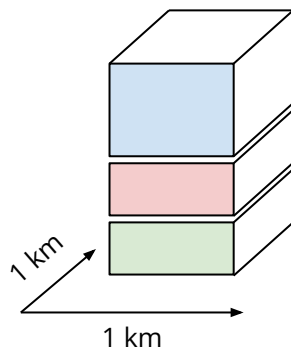


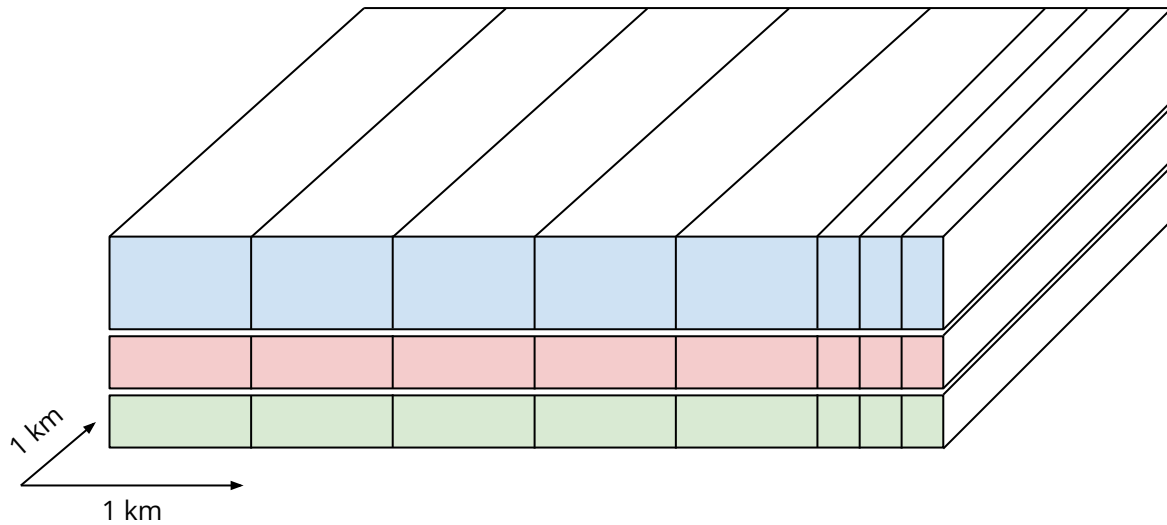


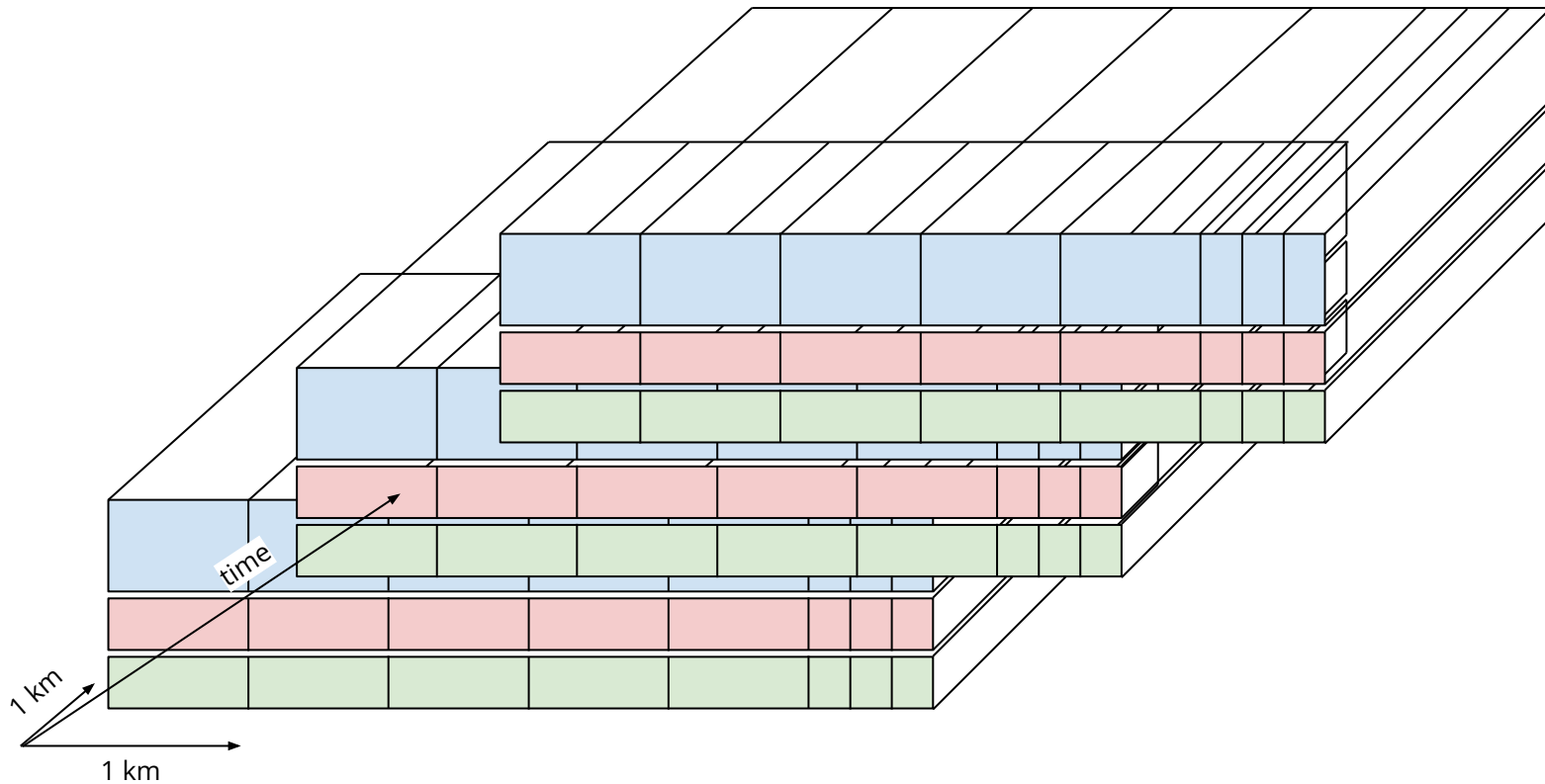






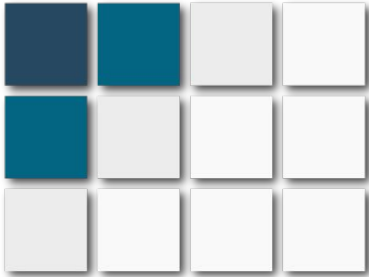






NetCDF format

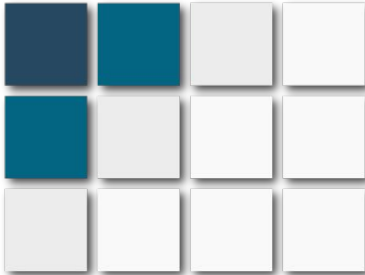
- Dimensions
- Variables
- Data
- Metadata



netCDF

NetCDF format

- Dimensions
- Variables
- Data
- Metadata

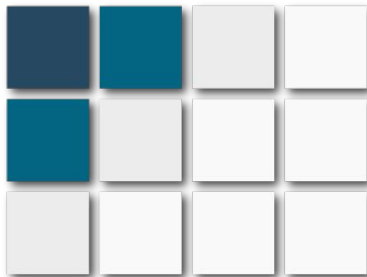


netCDF

```
VI_terra
<xarray.Dataset>
Dimensions:                                (lat: 134, lon: 182, time: 345)
Coordinates:
  * time                                   (time) object 2004-12-18 00:00:00 ... 2019-12-19 00:00:00
  * lat                                   (lat) float64 21.75 21.75 ... 21.2
  * lon                                   (lon) float64 -158.3 ... -157.6
Data variables:
  crs                                     int8 ...
  _500m_16_days_EVI                      (time, lat, lon) float32 ...
  _500m_16_days_MIR_reflectance           (time, lat, lon) float32 ...
  _500m_16_days_NDVI                     (time, lat, lon) float32 ...
  _500m_16_days_NIR_reflectance           (time, lat, lon) float32 ...
  _500m_16_days_VI_Quality                (time, lat, lon) float64 ...
  _500m_16_days_blue_reflectance          (time, lat, lon) float32 ...
  _500m_16_days_composite_day_of_the_year (time, lat, lon) float32 ...
  _500m_16_days_pixel_reliability         (time, lat, lon) float64 ...
  _500m_16_days_red_reflectance           (time, lat, lon) float32 ...
  _500m_16_days_relative_azimuth_angle    (time, lat, lon) float32 ...
  _500m_16_days_sun_zenith_angle          (time, lat, lon) float32 ...
  _500m_16_days_view_zenith_angle         (time, lat, lon) float32 ...
Attributes:
  title:      MOD13A1.006 for aid0001
  Conventions: CF-1.6
  institution: Land Processes Distributed Active Archive Center (LP DAAC)
  source:     AppEEARS v2.40
  references: See README.txt
```

NetCDF format

- Dimensions
- Variables
- Data
- Metadata



netCDF

```
VI_terra

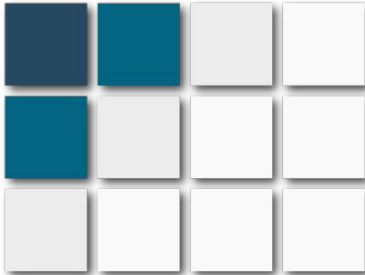
<xarray.Dataset>
Dimensions:                                (lat: 134, lon: 182, time: 345)
Coordinates:
  * time                                   (time) object 2004-12-18 00:00:00 ... 2019-12-19 00:00:00
  * lat                                    (lat) float64 21.75 21.75 ... 21.2
  * lon                                    (lon) float64 -158.3 ... -157.6

Data variables:
  crs                                     int8 ...
  _500m_16_days_EVI                      (time, lat, lon) float32 ...
  _500m_16_days_MIR_reflectance           (time, lat, lon) float32 ...
  _500m_16_days_NDVI                     (time, lat, lon) float32 ...
  _500m_16_days_NIR_reflectance           (time, lat, lon) float32 ...
  _500m_16_days_VI_Quality                (time, lat, lon) float64 ...
  _500m_16_days_blue_reflectance          (time, lat, lon) float32 ...
  _500m_16_days_composite_day_of_the_year (time, lat, lon) float32 ...
  _500m_16_days_pixel_reliability         (time, lat, lon) float64 ...
  _500m_16_days_red_reflectance           (time, lat, lon) float32 ...
  _500m_16_days_relative_azimuth_angle    (time, lat, lon) float32 ...
  _500m_16_days_sun_zenith_angle          (time, lat, lon) float32 ...
  _500m_16_days_view_zenith_angle         (time, lat, lon) float32 ...

Attributes:
  title:      MOD13A1.006 for aid0001
  Conventions: CF-1.6
  institution: Land Processes Distributed Active Archive Center (LP DAAC)
  source:     AppEEARS v2.40
  references: See README.txt
```


NetCDF format

- Dimensions
- Variables
- Data
- Metadata



netCDF

VI_terra

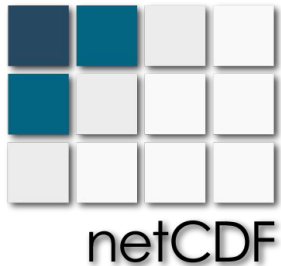
```
<xarray.Dataset>
Dimensions:                                (lat: 134, lon: 182, time: 345)
Coordinates:
  * time                                   (time) object 2004-12-18 00:00:00 ... 2019-12-19 00:00:00
  * lat                                   (lat) float64 21.75 21.75 ... 21.2
  * lon                                   (lon) float64 -158.3 ... -157.6
Data variables:
  crs                                     int8 ...
  _500m_16_days_EVI                     (time, lat, lon) float32 ...
  _500m_16_days_MIR_reflectance          (time, lat, lon) float32 ...
  _500m_16_days_NDVI                     (time, lat, lon) float32 ...
  _500m_16_days_NIR_reflectance          (time, lat, lon) float32 ...
  _500m_16_days_VI_Quality                (time, lat, lon) float64 ...
  _500m_16_days_blue_reflectance         (time, lat, lon) float32 ...
  _500m_16_days_composite_day_of_the_year (time, lat, lon) float32 ...
  _500m_16_days_pixel_reliability         (time, lat, lon) float64 ...
  _500m_16_days_red_reflectance          (time, lat, lon) float32 ...
  _500m_16_days_relative_azimuth_angle   (time, lat, lon) float32 ...
  _500m_16_days_sun_zenith_angle         (time, lat, lon) float32 ...
  _500m_16_days_view_zenith_angle        (time, lat, lon) float32 ...
```

```
Attributes:
  title:      MOD13A1.006 for aid0001
  Conventions: CF-1.6
  institution: Land Processes Distributed Active Archive Center (LP DAAC)
  source:     AppEEARS v2.40
  references: See README.txt
```

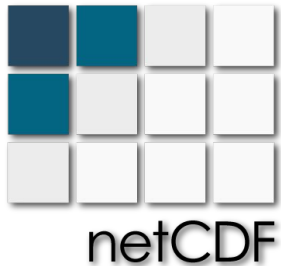
Data analysis toolbox



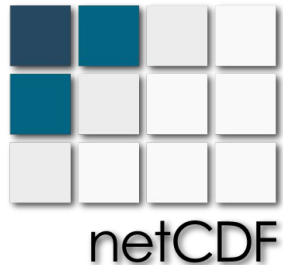
Data analysis toolbox



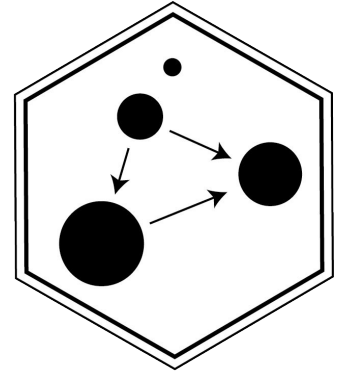
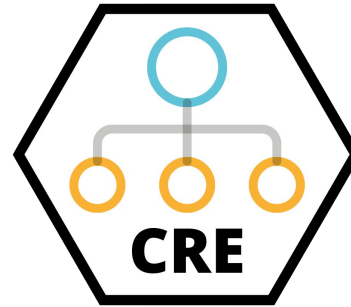
Data analysis toolbox



Data analysis toolbox



Google Earth Engine



<https://fasrc.github.io/CRE/>
<https://fasrc.github.io/CausalGPS/>

Data processing and analysis

Summary

- Big data analysis can be conducted using solely free and open source software!
 - NetCDF file format is great when working with high-dimensional geospatial datasets
-

How to conduct a big data analysis on air pollution and health?

[UKRAINE](#)[CORONAVIRUS](#)[RECODE](#)[THE GOODS](#)[FUTURE PERFECT](#)[THE HIGHLIGHT](#)[CROSSWORD](#)[MORE ▾](#)

Science has been in a “replication crisis” for a decade. Have we learned anything?

Bad papers are still published. But some other things might be getting better.

By Kelsey Piper | Oct 14, 2020, 12:20pm EDT

TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

[NEWS & OPINION](#)[PUBLICATIONS](#)[CATEGORIES](#)

[Home](#) / [News & Opinion](#)

Potential Causes of Irreproducibility Revealed

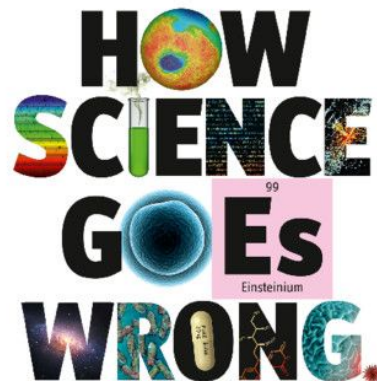
Five independent groups got different results in a drug-response experiment, despite sharing protocols, reagents, and cell lines. The researchers identify technical variables could be to blame.

The
Economist

OCTOBER 19TH-20TH 2013

economist.com

Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar



nature

[Explore content ▾](#)[About the journal ▾](#)[Publish with us ▾](#)

[nature](#) > [news feature](#) > [article](#)

[Published: 25 May 2016](#)

1,500 scientists lift the lid on reproducibility

[Monya Baker](#)

[Nature](#) **533**, 452–454 (2016) | [Cite this article](#)

34k Accesses | **1489** Citations | **3920** Altmetric | [Metrics](#)

Capturing the data pipeline

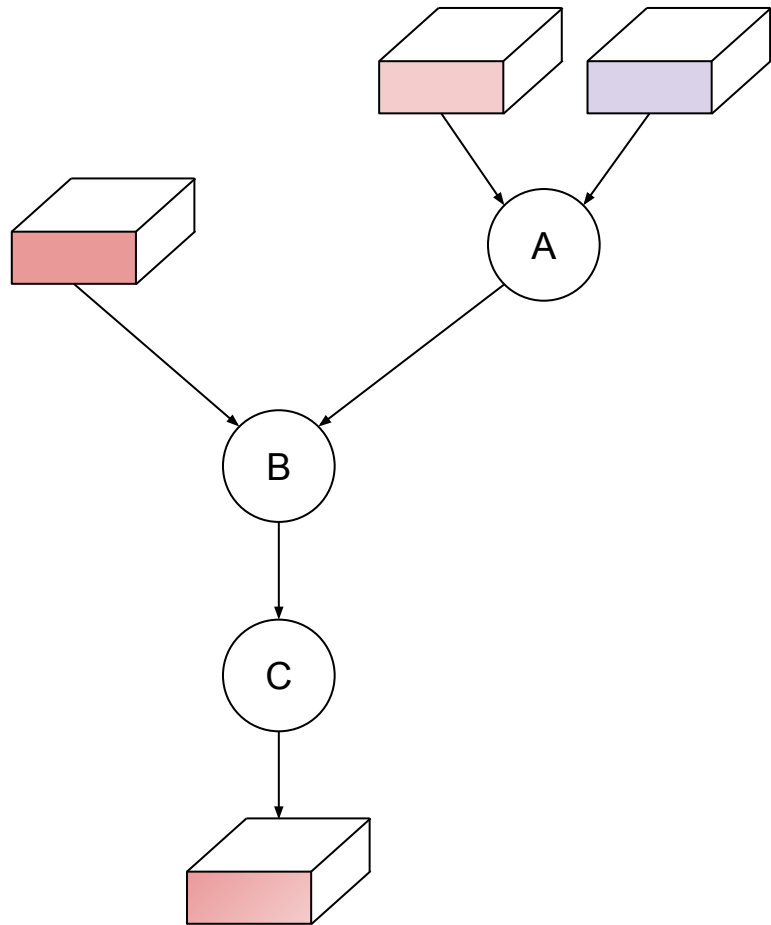
- Review
- Verification
- Collaboration

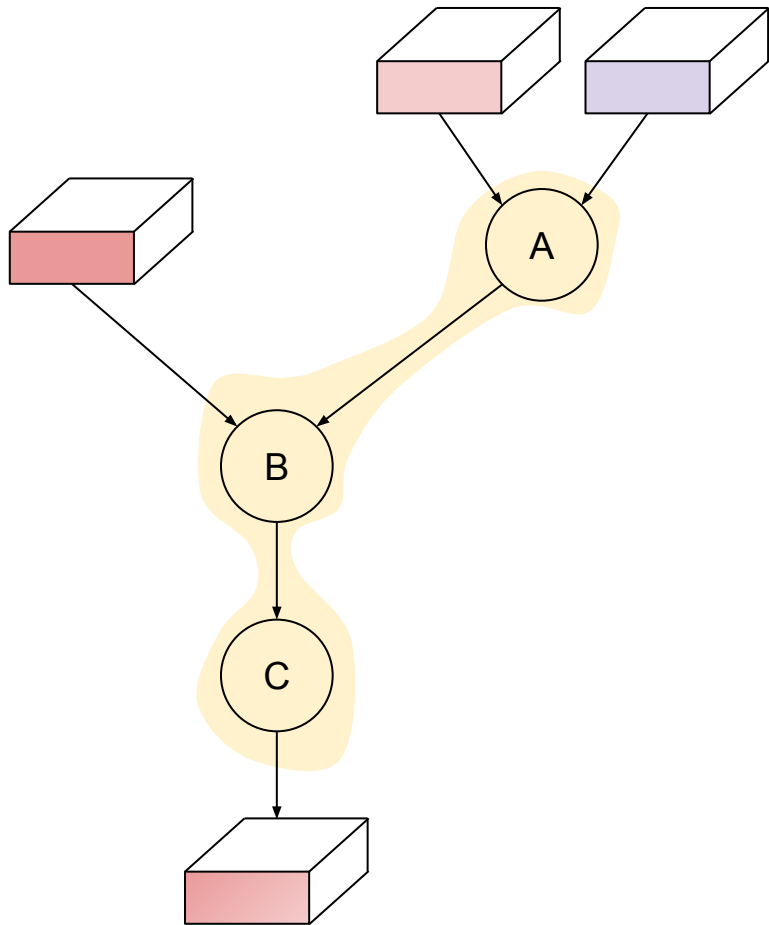
“An article about computational science in a scientific publication is **not** the scholarship itself, it is merely **advertising** of the scholarship. The actual scholarship is the software, [data] ... and set of instructions which generated the figures. ”

~ Prof Claerbout

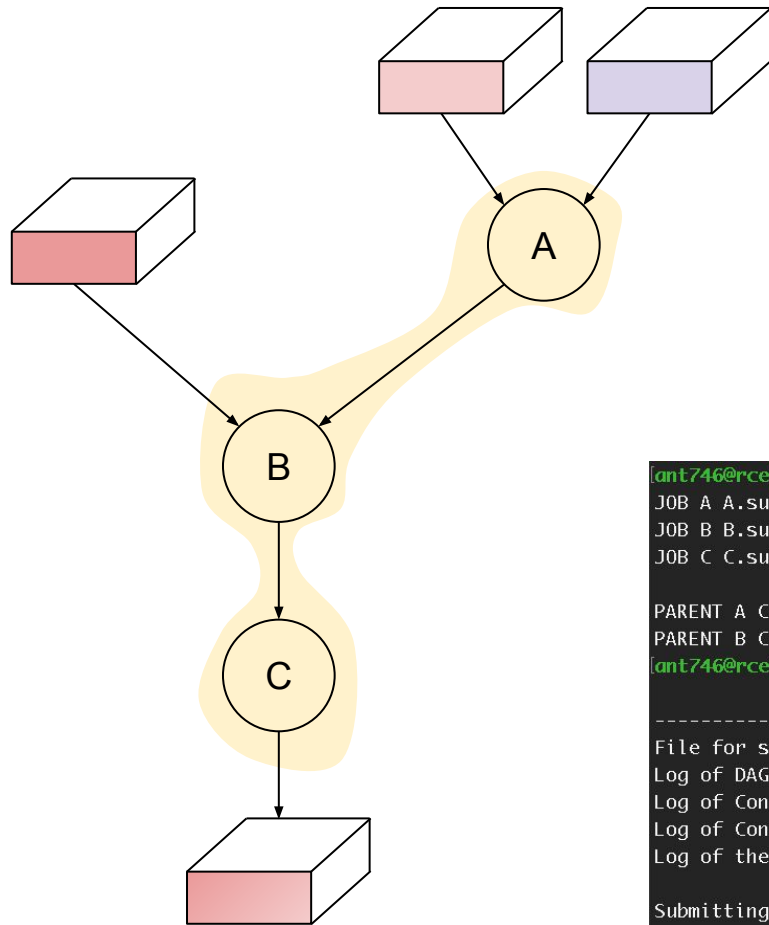
Automation

- Fast(er) analysis
- Troubleshooting
- Reuse
- Education and training





JOB	A	A.submit
JOB	B	B.submit
JOB	C	C.submit
PARENT	A	CHILD B
PARENT	B	CHILD C



```
JOB  A  A.submit
JOB  B  B.submit
JOB  C  C.submit
PARENT A CHILD B
PARENT B CHILD C
```

```
[ant746@rce6-5:~/shared_space/ci3_ant746/mcbs-medpar-mbsf (master)] $ cat workflow.dag
```

```
JOB A A.submit
JOB B B.submit
JOB C C.submit
```

```
PARENT A CHILD B
PARENT B CHILD C
```

```
[ant746@rce6-5:~/shared_space/ci3_ant746/mcbs-medpar-mbsf (master)] $ condor_submit_dag workflow.dag
```

```
-----
File for submitting this DAG to Condor           : workflow.dag.condor.sub
Log of DAGMan debugging messages                 : workflow.dag.dagman.out
Log of Condor library output                    : workflow.dag.lib.out
Log of Condor library error messages             : workflow.dag.lib.err
Log of the life of condor_dagman itself          : workflow.dag.dagman.log
-----
```

```
Submitting job(s).
1 job(s) submitted to cluster 84471.
-----
```

Workflow engines



Snakemake
nextflow



COMMON
WORKFLOW
LANGUAGE






- A free and open-source software platform to archive, share, and cite research data
 - Focus on data sharing and making data available


78 institutions around the globe run Dataverse installations as their official data repository



Data sharing

 **HARVARD**
Dataverse

[Add Data](#) ▾ [Search](#) ▾ [About](#) [User Guide](#) [Support](#) [Sign Up](#) [Log In](#)


 **HARVARD T.H. CHAN**
SCHOOL OF PUBLIC HEALTH


NSAPH Dataverse (Harvard University)


[Harvard Dataverse](#) >


[✉ Contact](#) [↗ Share](#)

Welcome to the dataverse collection by the National Studies on Air Pollution and Health group at the Harvard T.H. Chan School of Public Health. The group releases analysis and open data relating to air quality, demographics, and health.

 [Advanced Search](#)

☒  **Dataverses (2)**

☒  **Datasets (2)**

☐  **Files (18)**


Dataverse Category
[Research Group \(2\)](#)


Publication Year
[2022 \(4\)](#)

Subject
[Earth and Environmental Sciences \(1\)](#)
[Mathematical Sciences \(1\)](#)
[Medicine, Health and Life Sciences \(1\)](#)

Author Name
[Braun, Danielle \(1\)](#)
[Khoshnevis, Naeem \(1\)](#)
[Woodward, Sophie \(1\)](#)
[Wu, Xiao \(1\)](#)

1 to 4 of 4 Results [↑↓ Sort ▾](#)

**Replication Data for: Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mortality in the United States**
Feb 14, 2022 - NSAPH Analysis Data

Woodward, Sophie, 2022, "Replication Data for: Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mortality in the United States", <https://doi.org/10.7910/DVN/3ZU0AS>, Harvard Dataverse, V1, UNF:6:D8W7/RcF4rlkKWjYBggFeQ== [fileUNF]

This is the data repository for publicly available data to reproduce analyses in Woodward, S., Wu, X., Hou, Z., Mork, D., Braun, D., Dominici, F., 2022. Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mort...

NSAPH Open Data (Harvard University)
Feb 7, 2022

NSAPH Analysis Data (Harvard University)
Feb 7, 2022

Data sharing

Data collections

The screenshot displays the Harvard Dataverse interface. At the top, the navigation bar includes links for 'Add Data', 'Search', 'About', 'User Guide', 'Support', 'Sign Up', and 'Log In'. The main header features the Harvard logo and the text 'HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH'. Below this, the page is titled 'NSAPH Dataverse (Harvard University)' with a breadcrumb link 'Harvard Dataverse >'. A welcome message states: 'Welcome to the dataverse collection by the National Studies on Air Pollution and Health group at the Harvard T.H. Chan School of Public Health. The group releases analysis and open data relating to air quality, demographics, and health.' A search bar with the placeholder 'Search this dataverse...' and a magnifying glass icon is present, along with a link to 'Advanced Search'. On the left sidebar, filters are shown for 'Dataverses (2)', 'Datasets (2)', and 'Files (18)'. Under 'Dataverse Category', 'Research Group (2)' is selected. Under 'Publication Year', '2022 (4)' is selected. Under 'Subject', 'Earth and Environmental Sciences (1)', 'Mathematical Sciences (1)', and 'Medicine, Health and Life Sciences (1)' are listed. Under 'Author Name', 'Braun, Danielle (1)', 'Khoshnevis, Naeem (1)', 'Woodward, Sophie (1)', and 'Wu, Xiao (1)' are listed. The main content area shows '1 to 4 of 4 Results'. The first result is 'Replication Data for: Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mortality in the United States', dated Feb 14, 2022, by Woodward, Sophie. A description follows: 'Woodward, Sophie, 2022, "Replication Data for: Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mortality in the United States", https://doi.org/10.7910/DVN/3ZU0AS, Harvard Dataverse, V1, UNF:6:D8W7/RcF4rlkKWjYBggFeQ== [fileUNF]'. Below this, a paragraph states: 'This is the data repository for publicly available data to reproduce analyses in Woodward, S., Wu, X., Hou, Z., Mork, D., Braun, D., Dominici, F., 2022. Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mort...'. Below the search results, two more entries are visible: 'NSAPH Open Data (Harvard University)' dated Feb 7, 2022, and 'NSAPH Analysis Data (Harvard University)' dated Feb 7, 2022. Red arrows from the 'Data collections' box point to the 'Subject' and 'Author Name' filters.

Data sharing

- Data should be licensed
- Metadata
- It should be complete
- It should be shared in a (free, open) machine-readable format

Dissemination

Summary

- Sharing of data, code and computational processes is necessary due to the requirements of policy makers, journals, funding agencies.

Email: anatrisovic@g.harvard.edu
GitHub & Twitter: [atrisovic](#)

