

Big Data for Spatial Analysis

D G Rossiter

Cornell University, Soil & Crop Sciences Section
Nanjing Normal University, Geographic Sciences Department

April 7, 2023

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- 1 What is “big data”?
- 2 Examples
- 3 Portals
- 4 Uses of Big Data
- 5 Applications
- 6 How to deal with big (spatio-temporal) data?

What is “big data”?

Examples

Portals

Uses of Big Data

Applications

How to deal with big (spatio-temporal) data?

References

1 What is “big data”?

2 Examples

3 Portals

4 Uses of Big Data

5 Applications

6 How to deal with big (spatio-temporal) data?

Often discussed in terms of **volume** (size of data), **velocity** (frequency of data), and **variety** (diversity of data types):

volume data sets that are **too large** to be handled by common processing methods

- e.g., exceed the storage capacity of main memory or even secondary memory

velocity hyper-temporal data sets, or high-bandwidth streams of data

- e.g., social media activity, real-time locations

variety data sets that have too much **complexity** to be handled by common processing methods

- e.g., observations with 100's to 1000's of attributes of variable or even unknown data quality

What is “big data”?

Examples

Portals

Uses of Big Data

Applications

How to deal with big (spatio-temporal) data?

References

- Too **large** to fully understand summaries or identification of unusual cases (“outliers”)
- Too **complex** to fully understand or control modelling
- Generally need **machine-learning methods** to analyze (e.g., random forests, neural nets)
- Generally too **large** to analyze on individual computers, requires cloud computing.

What is “big data”?

Examples

Portals

Uses of Big Data

Applications

How to deal with big (spatio-temporal) data?

References

“Kilobytes $2^{10} \approx 10^3$ were stored on floppy disks. Megabytes $2^{20} \approx 10^6$ were stored on hard disks. Terabytes $2^{30} \approx 10^9$ were stored in disk arrays. Petabytes $2^{40} \approx 10^{12}$ are stored in the cloud.

“As we moved along that progression, we went from the *folder* analogy to the *file cabinet* analogy to the *library* analogy to – well, at petabytes we ran out of organizational analogies.”

– Anderson, C. (2008, June 23). *The end of theory: the data deluge makes the scientific method obsolete*. **Wired**.
<https://www.wired.com/2008/06/pb-theory/> [1]

What is "big
data"?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- 1 **capturing** massive amounts of obserations
 - e.g., real-time sensors, satellite images, monitoring instruments
- 2 **storage** and data transfer from storage to computation
- 3 **search**, i.e., query to find/subset/summarize
- 4 **processing**, i.e., computation (CPU, memory, parallelization)
- 5 **analysis**: methods and understanding the results
- 6 **sharing**, information policies, e.g., privacy
 - different parts of the database may have different policies
- 7 **visualization**
 - summarizing with appropriate graphic design

What is "big data"?

Examples

Portals

Uses of Big Data

Applications

How to deal with big (spatio-temporal) data?

References

- sensor networks with fine temporal resolution
 - e.g., automated weather stations
- mobile devices
 - e.g, geo-located phones
- remote sensors
 - ever-increasing **spatio-temporal resolution**
- digital laboratory instruments
- point-of-sales or service
 - e.g., pharmacies, retail stores
- user contributions
 - social media, citizen science

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

1 What is “big data”?

2 Examples

3 Portals

4 Uses of Big Data

5 Applications

6 How to deal with big (spatio-temporal) data?

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- NOAA $\approx 50 \cdot 10^9$ multivariate observations of oceans (as of 2018)¹
 - temperature, salinity, oxygen, nitrates, phosphates and silicates at the particular **location** and **depth** collected at a particular **time**, so 4D
- 23andMe² DNA analysis of $5 \cdot 10^6$ individuals
- Twitter has about $500 \cdot 10^6$ tweets per day, or 6k per second ³
- MasterCard processed $74 \cdot 10^9$ transactions per year in 2012⁴

¹<https://www.ncei.noaa.gov/products/world-ocean-atlas>

²<https://www.23andme.com/en-int/dna-ancestry/>, 17-Jan-2021

³<https://www.dsayce.com/social-media/tweets-day/>, as of August 2022

⁴<http://blog.unibulmerchantsservices.com/how-mastercard-processes-74b-transactions-a-year/>

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- eBird⁵ (Cornell Ornithology)
 - $\approx 1300 \cdot 10^6$ total observations (as of end 2022)⁶
 - $\approx 225 \cdot 10^6$ new observations in 2022 (“volume”, “velocity”)
 - (semi-)automated quality control⁷

⁵<http://www.ebird.org>

⁶<https://ebird.org/wi/news/2022-year-in-review>

⁷<https://support.ebird.org/en/support/solutions/articles/48000795278>

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- Ag-Analytics⁸
 - zoom in, click on individual fields: insurance, revenue, yield forecast
 - combines multiple open layers, with own analytics
- Gro Intelligence⁹
 - “leading the modern agricultural revolution using data and technology, driven by advances in parallel processing, remote sensing, machine learning, and AI”

⁸<https://app.profit.ag/app>

⁹<https://www.gro-intelligence.com>

What is “big
data”?

Examples

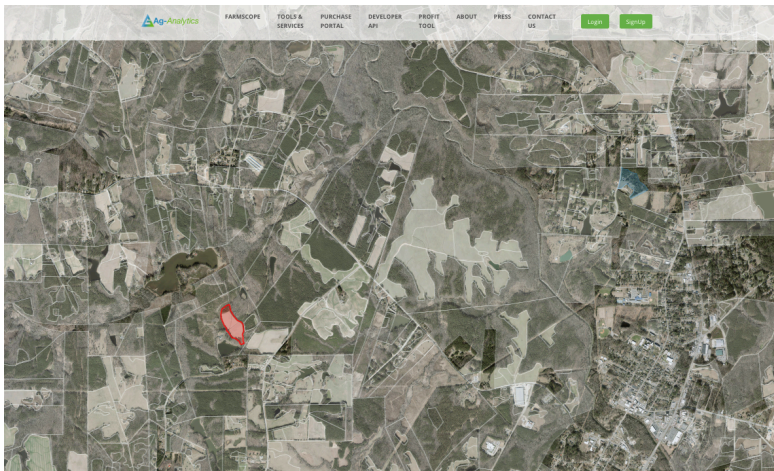
Portals

Uses of Big
Data

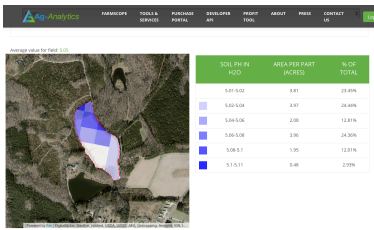
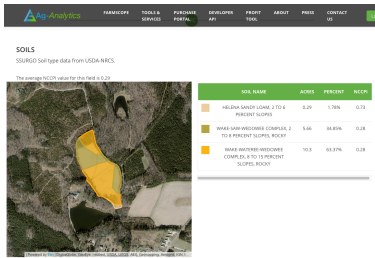
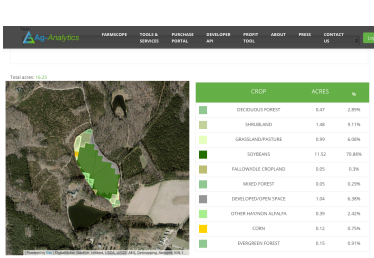
Applications

How to deal
with big
(spatio-
temporal) data?

References



near Franklinton, North Carolina
field at (36.1068N, -78.3530E)



What is “big data”?

Examples

Portals

Uses of Big Data

Applications

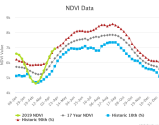
How to deal with big (spatio-temporal) data?

References



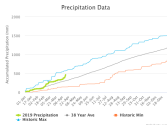
FARMCONDITIONSLIVE™ - NDVI DATA

Normalized Difference Vegetation Index, also known as NDVI, is a measure of the level of greenness inferred by the MODIS Aqua satellite dataset from NASA. It is updated daily.



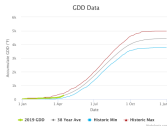
FARMCONDITIONSLIVE™ - PRECIPITATION DATA

Precipitation data. This graph shows accumulated precipitation to date compared to historical norms. These data are updated daily. Source data obtained from PRISM Climate Group, Oregon State University. <http://prism.oregonstate.edu>, tabulation created now.



FARMCONDITIONSLIVE™ - GROWING DEGREE DAYS

Growing degree days are cumulative sum of degrees above 50°F. The more growing degree days, the hotter the growing season. This graph shows accumulation to date compared to historical norms and is updated daily. Source data obtained from PRISM Climate Group, Oregon State University. <http://prism.oregonstate.edu>, tabulation created now. Non-Commercial Use Only.



Browse data

Visualize data from our collection of over 6.8 million data series and growing.



Geography
Land, Water and Land ›



Infrastructure
Production Infrastructure,
Storage,
Telecommunications ›



**Investment, Lending
and Transfers**
Assets, Investment,
Lending ›



**Macroeconomic
Indicators**
Industry Indicators,
Inflation, Labor Statistics ›



Prices
Indices, Market Prices,
Producer Prices ›



Supply
Market Supply, Post-
Harvest Processing,
Production ›



Trade
Trade Flow Matrix, Trade



Weather and Climate
Data Infrastructure



Vast Range of Sources



Meaningful Geospatial Data



Analytical Tools



Visualizations

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- Reference: [10]
- Global predictions for numeric soil properties: OC, bulk density, Cation Exchange Capacity (CEC), pH, soil texture fractions and coarse fragments at seven standard depths (0, 5, 15, 30, 60, 100 and 200 cm; total 280 raster layers)
- based on $\approx$ 150k profiles, 158 gridded covariates
- Maps for on-line view or download¹⁰

¹⁰<https://soilgrids.isric.org>

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

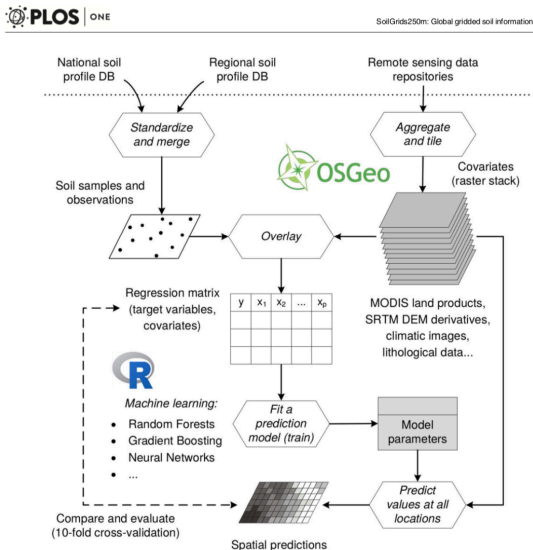


Fig 5. The (data-driven) statistical framework used for generating SoilGrids. SoilGrids are primarily based on publicly released soil profile compilations, NASA's MODIS and SRTM data products and Open Source software compiled with the ATLAS library: > (including contributed packages), and Open Source Geospatial Foundation (OSGeo) supported software tools.

doi:10.1371/journal.pone.0169748.g005

What is “big
data”?

Examples

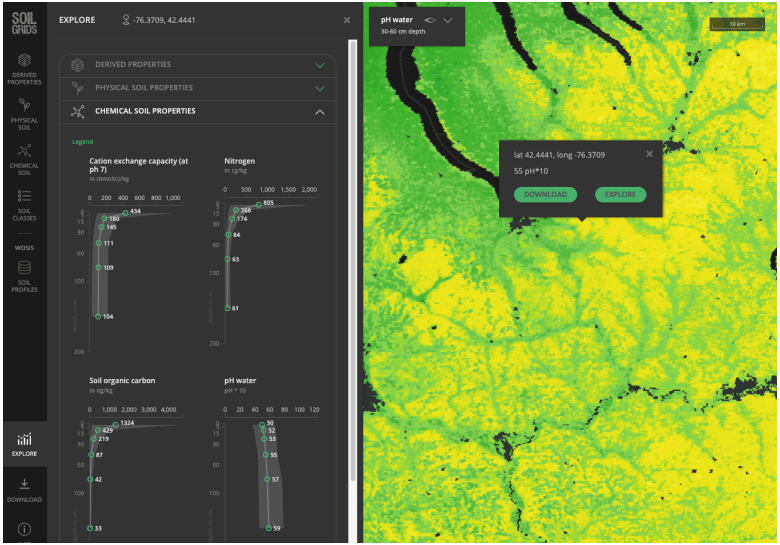
Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References



POLARIS [3]

- Purpose: map **probability** of **soil series** at **every grid cell** in the lower 48 (USA) States; total $\approx 1.25 \cdot 10^9$ grid cells
- algorithm: DSMART (Disaggregation and Harmonization of Soil Map Units Through Resampled Classification Trees)
- supercomputer “Blue Waters”; 12 474 nodes 30 x 30 km, with 60 km buffer to ensure continuity
- 30 m horizontal spatial resolution, so $1000 \times 1000 = 10^6$ per 1° tile¹¹
- required 450 000 core-hours = 5 wall-clock hours
- “This is negligible computer time at current HPC facilities that can handle 10 million (≈ 1100 years) core-hour tasks.”

¹¹ <http://hydrology.cee.duke.edu/POLARIS/>

What is "big
data"?

Examples

Portals

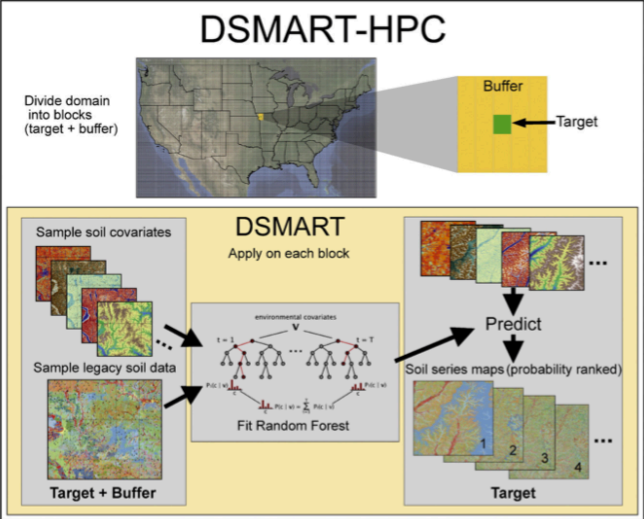
Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

N.W. Chaney et al. / Geoderma 274 (2016) 54–67



What is "big
data"?

Examples

Portals

Uses of Big
Data

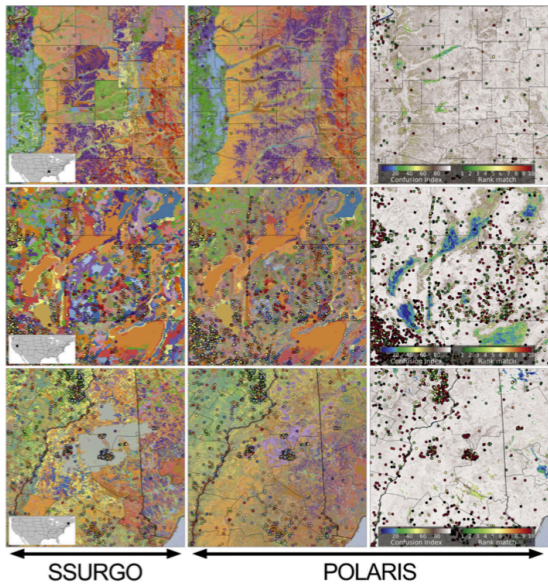
Applications

How to deal
with big
(spatio-
temporal) data?

References

N.W. Chaney et al. / *Geoderma* 274 (2016) 54–67

59



POLARIS covariate importance

What is “big data”?

Examples

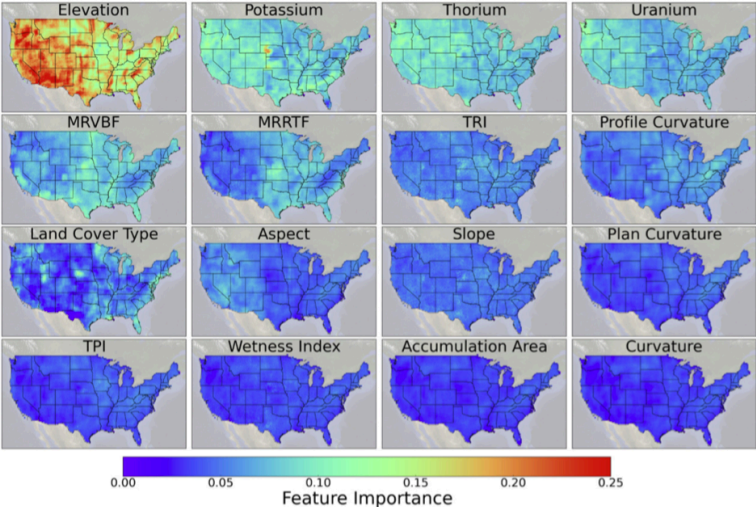
Portals

Uses of Big Data

Applications

How to deal with big (spatio-temporal) data?

References



What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

1 What is “big data”?

2 Examples

3 Portals

4 Uses of Big Data

5 Applications

6 How to deal with big (spatio-temporal) data?

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- Collect and catalog many sources of geographic data
- example: World Food Programme (WFP) Geonode¹²
- example: Soil Geographic Databases¹³
- example: OpenGovernment¹⁴
- problem: **searching** for relevant (to the user) information
 - WFP Geonode: search by region, type of information, keyword, date, extent, file type

¹²<https://geonode.wfp.org>

¹³<https://www.isric.org/explore/soil-geographic-databases>

¹⁴<https://www.data.gov/open-gov/>

What is "big
data"?

Examples

Portals

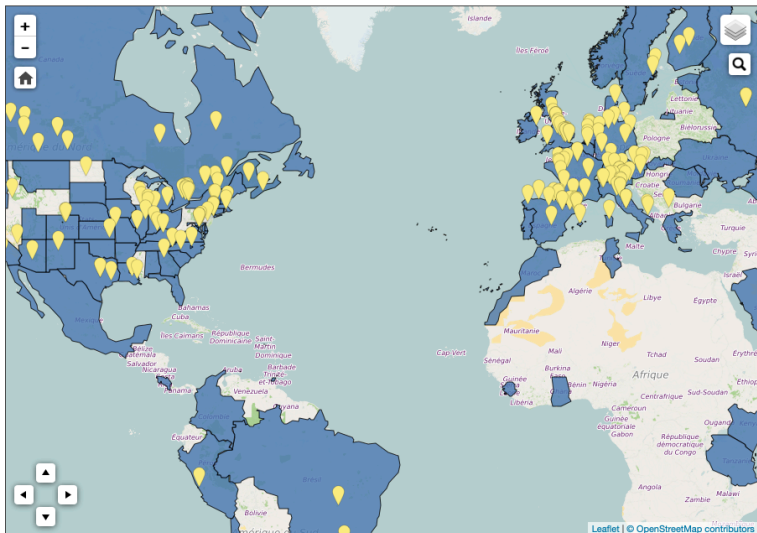
Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

Map representation of Open Data Sites



What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

1 What is “big data”?

2 Examples

3 Portals

4 Uses of Big Data

5 Applications

6 How to deal with big (spatio-temporal) data?

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- Scientific research into processes
 - e.g., eBird Science applications¹⁵:
 - “The eBird Status and Trends project now provides modeled abundance and range maps for 2068 species including 868 species added this month [Dec 2022] ... ”¹⁶
- Mapping
 - e.g. SoilGrids, POLARIS
- Prediction, decision support
 - e.g., Ag-Analytics, Gro Intelligence
- Visualization, hypothesis formation

¹⁵<https://ebird.org/science>

¹⁶<https://ebird.org/wi/news/2022-year-in-review>

What is "big data"?

Examples

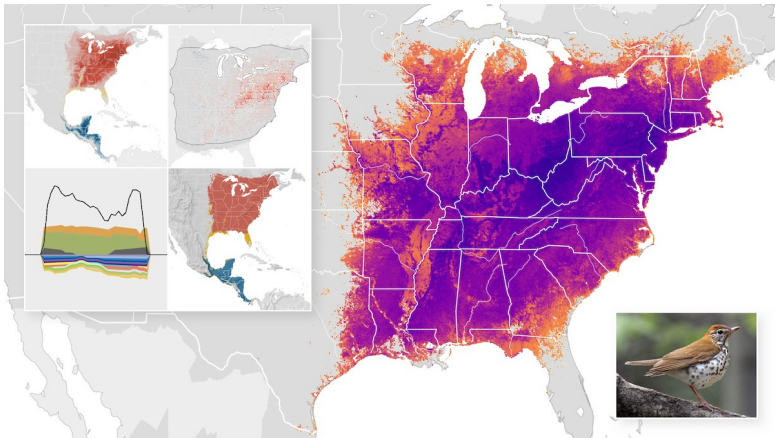
Portals

Uses of Big Data

Applications

How to deal with big (spatio-temporal) data?

References



Status & trends for wood thrush (*Hylocichla mustelina*)

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- European Radioactivity Environmental Monitoring¹⁷
 - “Gamma dose rate averages and maxima for the last 24 hours in almost real time”
 - European Radiological Data Exchange Platform (EURDEP): a **network** for the **exchange** of radiological monitoring data between most European countries
 - Large network of sensors, automatic reporting and summarizing
- Methods: see [5] “**Real-time automatic interpolation** of ambient gamma dose rates from the Dutch radioactivity monitoring **network**”

¹⁷<https://remap.jrc.ec.europa.eu>

What is “big data”?

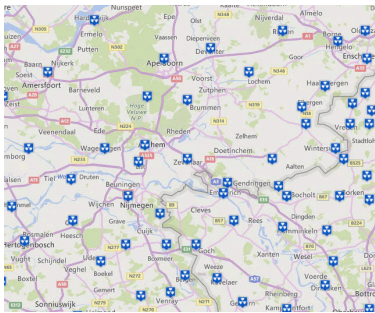
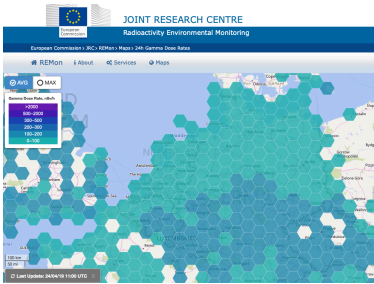
Examples
Portals

Uses of Big
Data

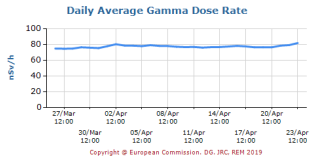
Applications

How to deal
with big
(spatio-
temporal) data?

References



 ARNHEM



What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

1 What is “big data”?

2 Examples

3 Portals

4 Uses of Big Data

5 Applications

6 How to deal with big (spatio-temporal) data?

What is “big data”?

Examples

Portals

Uses of Big Data

Applications

How to deal with big (spatio-temporal) data?

References

- General discussions: [2, 4]
- Ecology: [6]
- Geographic sociology: [7]
- Epidemiology: [9, 12]
- Agroecosystems: [8]
- Agricultural entomology: [11]
- Radiation monitoring: [5]

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

1 What is “big data”?

2 Examples

3 Portals

4 Uses of Big Data

5 Applications

6 How to deal with big (spatio-temporal) data?

How to deal with big (spatio-temporal) data?

What is "big
data"?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- **automated** methods for data capture, data screening (quality control)
- **robust** statistical methods, not sensitive to outliers, not dependent on manual selection of model form
- **massive** computing resources
- **collaborative** science – combine different disciplinary knowledge; also requires experts in inter-, multi-disciplinary collaboration

Massive datasets can not be handled on many personal/departmental computers.
So **cloud computing** must be used.

- Example: Google Earth Engine¹⁸
 - “A **planetary-scale** platform for Earth science data & **analysis** – Powered by Google’s **cloud infrastructure**”
 - “hosts satellite imagery and stores it in a **public data archive** that includes **historical earth images** going back more than **forty years** ... made available for global-scale **data mining**.”
 - Aimed at consistent Earth-wide analyses, but can be used regionally or locally
 - Must register an account¹⁹ (free for research, education, and nonprofit use), and log in to use.

¹⁸<https://earthengine.google.com/>

¹⁹<https://signup.earthengine.google.com/>

What is "big
data"?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- Computation is all done **remotely** (parallel processing)
 - local computer only for coding/viewing
- Accessible via an Application Programming Interface (**API**)
 - Javascript (built into GEE)²⁰, Python, **R** with rgee package²¹
- Built-in code editor
- Direct access to the datasets
- **Image processing, Geometry** algorithms
- **Machine-learning** algorithms: un/supervised classification

²⁰[https:](https://developers.google.com/earth-engine/guides/getstarted)

[//developers.google.com/earth-engine/guides/getstarted](https://developers.google.com/earth-engine/guides/getstarted)

²¹<https://r-spatial.github.io/rgee/>

What is "big
data"?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

Many datasets are also too large for local storage, so they are also in the cloud and used as needed; Google Earth Engine includes many datasets²², including:

- Imagery (Landsat, Sentinel, MODIS . . .)
- Atmospheric conditions (can help correct other products)
- Weather
- Geophysical: terrain (e.g., SRTM), elevation
- Highlights
- Administrative
- Interpreted: land cover, land use, cropland (e.g., USDA NASS; Global Food Security)

²²<https://developers.google.com/earth-engine/datasets/catalog>

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

Ever-increasing amounts of data, ever-increasing computer power, allow integrating many “big” data sources to deal with “big”, complex, inter-disciplinary problems.

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- [1] Chris Anderson. The end of theory: the data deluge makes the scientific method obsolete. *Wired*, Jun 2008. ISSN 1059-1028. URL <https://www.wired.com/2008/06/pb-theory/>.
- [2] Roger Bivand and Konstantin Krivoruchko. Big data sampling and spatial analysis: “which of the two ladles, of fig-wood or gold, is appropriate to the soup and the pot?”. *Statistics & Probability Letters*, 136:87–91, May 2018. doi: 10.1016/j.spl.2018.02.012.
- [3] Nathaniel W. Chaney, Eric F. Wood, Alexander B. McBratney, Jonathan W. Hempel, Travis W. Nauman, Colby W. Brungard, and Nathan P. Odgers. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. *Geoderma*, 274:54–67, Jul 2016. doi: 10.1016/j.geoderma.2016.03.025.
- [4] Hamid Ekbia, Michael Mattioli, Inna Kouper, G. Arave, Ali Ghazinejad, Timothy Bowman, Venkata Ratandeep Suri, Andrew Tsou, Scott Weingart, and Cassidy R. Sugimoto. Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, 66(8):1523–1545, 2015. doi: 10.1002/asi.23294.
- [5] Paul H. Hiemstra, Edzer J. Pebesma, Chris J.W. Twenhöfel, and Gerard B.M. Heuvelink. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. *Computers & Geosciences*, 35(8):1711–1721, Aug 2009. doi: 10.1016/j.cageo.2008.10.011.

What is "big
data"?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- [6] S. L. LaDeau, B. A. Han, E. J. Rosi-Marshall, and K. C. Weathers. The next decade of big data in ecosystem science. *Ecosystems*, 20(2):274–283, Mar 2017. doi: 10.1007/s10021-016-0075-y.
- [7] Jiwei Li, Qingqing Ye, Xuankai Deng, Yaolin Liu, and Yanfang Liu. Spatial-temporal analysis on Spring Festival travel rush in China based on multisource big data. *Sustainability*, 8(11):UNSP 1184, Nov 2016. doi: 10.3390/su8111184.
- [8] M. Susan Moran, Philip Heilman, Debra P. C. Peters, and Chandra Holifield Collins. Agroecosystem research with big data and a modified scientific method using machine learning concepts. *Ecosphere*, 7(10):e01493, 2016. doi: 10.1002/ecs2.1493.
- [9] Dirk U. Pfeiffer and Kim B. Stevens. Spatial and temporal epidemiological analysis in the big data era. *Preventive Veterinary Medicine*, 122(1–2):213–220, Nov 2015. doi: 10.1016/j.prevetmed.2015.05.012.
- [10] Laura Poggio, Luis M. de Sousa, Niels H. Batjes, Gerard B. M. Heuvelink, Bas Kempen, Eloi Ribeiro, and David Rossiter. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7(1):217–240, June 2021. doi: 10.5194/soil-7-217-2021.

What is “big
data”?

Examples

Portals

Uses of Big
Data

Applications

How to deal
with big
(spatio-
temporal) data?

References

- [11] Jay A. Rosenheim and Claudio Gratton. Ecoinformatics (big data) for agricultural entomology: pitfalls, progress, and promise. *Annual Review of Entomology*, 62(1):399–417, Jan 2017. doi: 10.1146/annurev-ento-031616-035444.
- [12] Kim B. Stevens and Dirk U. Pfeiffer. Spatial modelling of disease using data- and knowledge-driven approaches. *Spatial and Spatio-temporal Epidemiology*, 2(3):125–133, Sep 2011. doi: 10.1016/j.sste.2011.07.007.