

Digital trace data and Harvard Dataverse

Stefano M. Iacus

Senior Research Scientist, Director of Data Science and Product Research,
IQSS, Harvard University

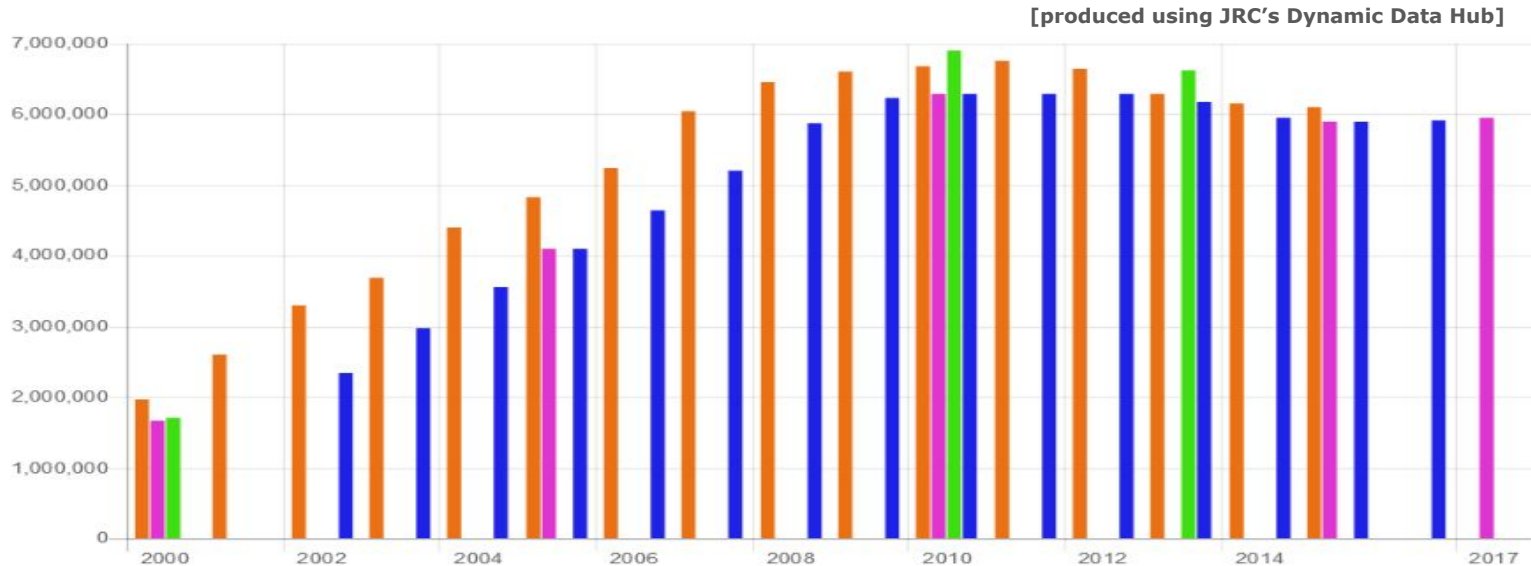
Sep 17th, 2024

79th United Nations General Assembly (UNGA79) Science Summit

Digital traces

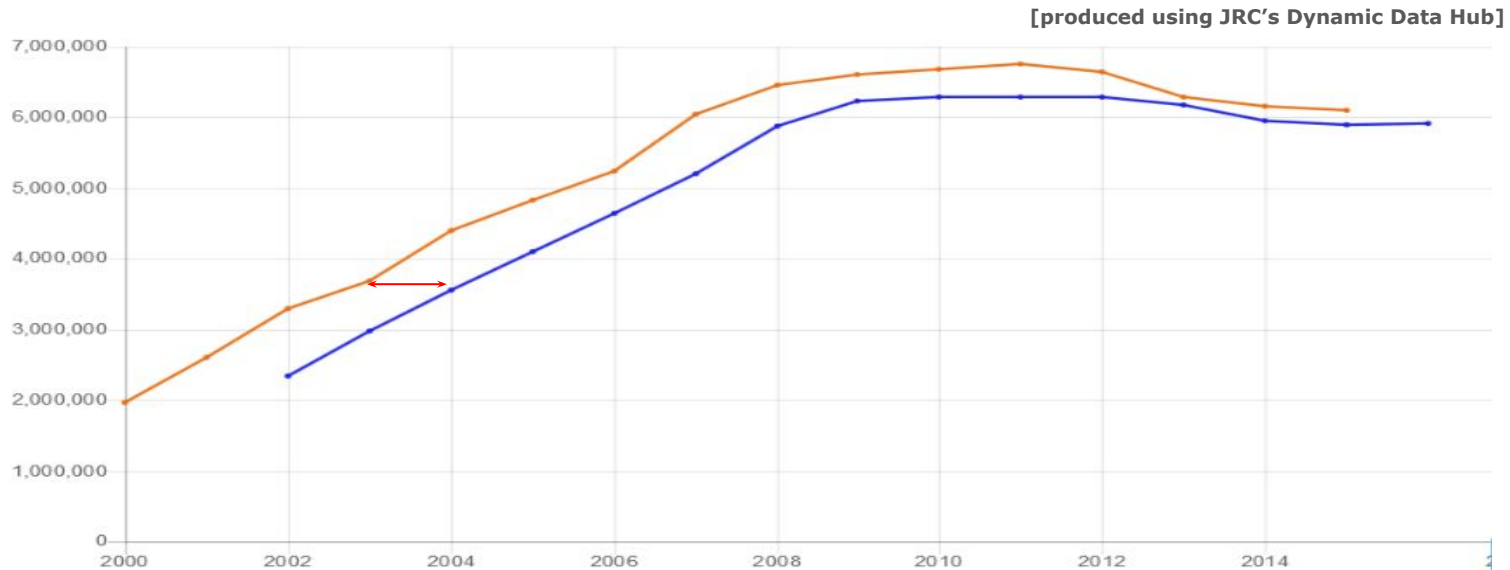
Why (also) digital trace data ?

As an example, stocks of migrants by country of birth in Spain (sources **Eurostat**, **OECD**, **UNDESA**, and **World Bank**) highlight temporal gaps the different datasets








































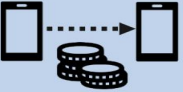









Why (also) digital trace data ?

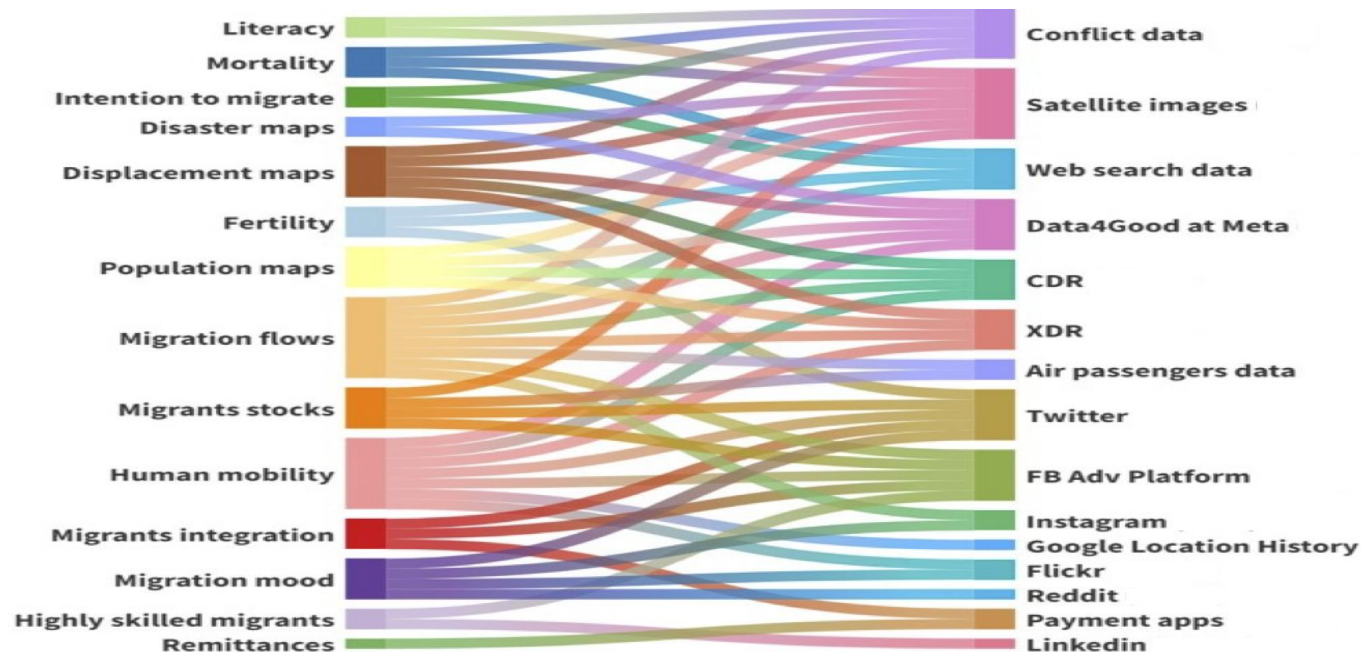
As an example, stocks of migrants by country of birth in Spain (sources **Eurostat**, **OECD**, **UNDESA**, and **World Bank**) highlight temporal gaps the different datasets



Pros & Cons of Digital trace data

Non-traditional data for demography, migration and human mobility		
Sources	Main strengths	Main challenges
Mobile phones data (CDRs and XDRs) 	Covers large population  Can capture hard-to-reach pop.  Timely info  High resolut. 	Requires anonymisation  Privacy concerns  Cost of data  Lack of standards 
Geo-located social media data and online media contents 	Covers large population  Richness of info  Can capture hard-to-reach pop.  Timely info  High resolut. 	Reliability of self-reported info  Selection & geographic bias  Privacy and ethical issues 
web searches and internet activity 	Can capture people's intentions  Timely information  Free of charge 	Selection bias  Mismatch intentions/actions  Too much info 
IP addresses of websites login and sent e-mails 	Richness of info  Can capture hard-to-reach pop.  Timely info 	Methodological and technical issues  Selection bias  Privacy and ethical issues 
Earth observation data and satellite imagery 	Comparability  Global coverage  High resolut. 	Cost of data  Continuity of service  Method. & technical issues 
Mobile money and mobile payment apps 	Timely info  Reach unbanked recipients  High resolut. 	Cost of data  Selection bias  Privacy and ethical issues  Method. & technical issues 

How the literature on demography, migration and human mobility uses digital trace data



Bosco, C., Grubanov-Boskovic, S., Iacus, S., Minora, U., Sermi, F. and Sphyrtos, S., *Data Innovation in Demography, Migration and Human Mobility*, EUR 30907 EN, Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-46702-1, doi:[10.2760/958409](https://doi.org/10.2760/958409), JRC127369.

Application of innovative data and digital traces

Can be used to:

- **fill in the gaps** in traditional statistics
- obtain estimates at **multiple temporal and spatial scales**
- **overcome measurement errors in survey data**

Current competitive advantages

- **greater geographic and temporal granularity**
- (near-) **real time availability**
- **extensive coverage** (which makes more immediate international comparisons possible).

Facebook Advertising Platform Data



The screenshot displays the Facebook Ads Manager interface for audience targeting. On the left, there are options for 'ience' and 'Use Saved Audience'. Below this is a search bar for 'Add a previously created Custom or Lookalike Audience' and buttons for 'Exclude' and 'Create New'. A dropdown menu shows 'Everyone in this location'. The main area features a map of Uganda with 'Kampala District' selected. Below the map, there are filters for 'Age' (18-65+) and 'Gender' (All, Men, Women). On the right, a summary box indicates 'Your audience selection is fairly broad.' and 'Potential Reach: 1,800,000 people'. Below this, 'Estimated Daily Results' show 'Reach: 21K - 60K'. A disclaimer states: 'The accuracy of estimates is based on factors like past campaign data, the budget you entered and market data. Numbers are provided to give you an idea of performance for your budget, but are only estimates and don't guarantee results.' A link 'Were these estimates helpful?' is also present.

Allows to estimate a target population in terms of age, gender, education, region, interests, etc

Only see Facebook accounts, therefore biased

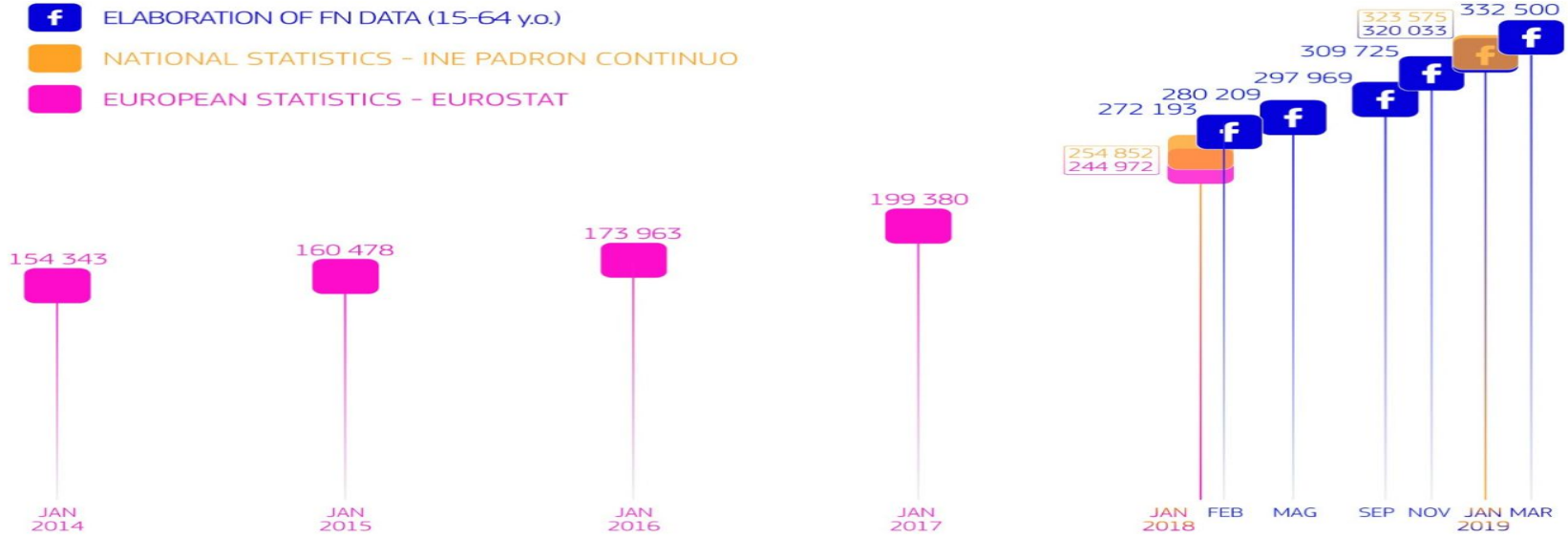
Available over time, not retrospectively

Not all usage of these data are permitted! Check the latest API documentation.

In our toy example for Uganda we show the extraction of these data for one given date only (Nov 9th 2019) but data collection can be done on a weekly basis



Nowcasting human mobility during humanitarian crisis



Venezuelans in Spain: latest **EU official statistics**, **Spanish Statistical Office**, **Facebook Network** data (improved frequency update & timeliness)

Reference: "Quantifying international human mobility patterns using Facebook Network data", S. Spyrtos, M. Vespe, F. Natale, I. Weber, E. Zagheni, M. Rango, *PlosOne*, 2019, <https://doi.org/10.1371/journal.pone.0224134>

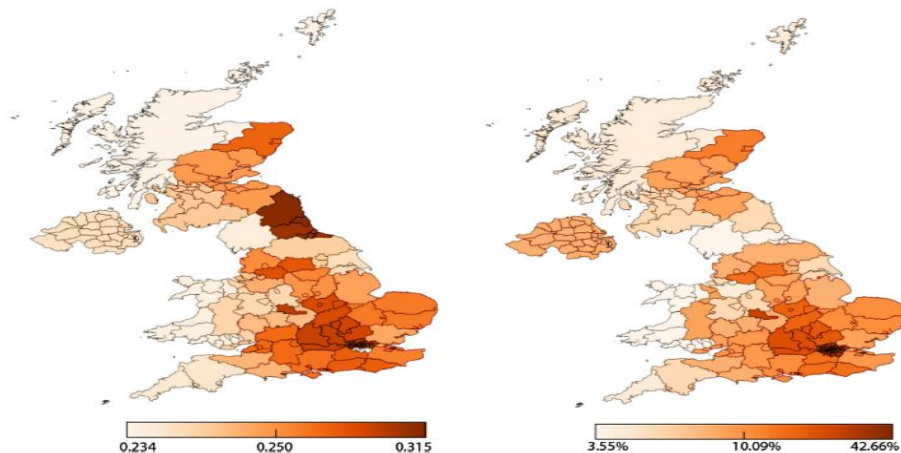
Migrants' integration I



Pollacci (2019), Pollacci et al 2021 and Sirbu et al. (2021) introduced the notion of **Superdiversity** based on Twitter data.

By superdiversity is intended a new level of cultural diversity due to immigration and cultural differences among immigrants themselves (Vertovec, 2007).

The authors create the superdiversity index (SI) from Twitter data based on the idea that "*different cultures assign different emotional valence to different words*"



Superdiversity index (left) and immigration levels (right) across UK regions at NUTS2 level.

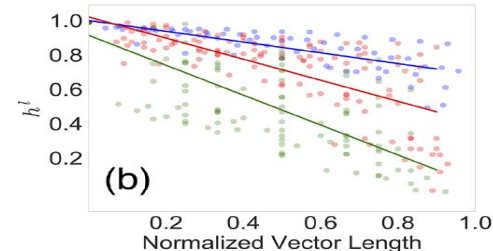
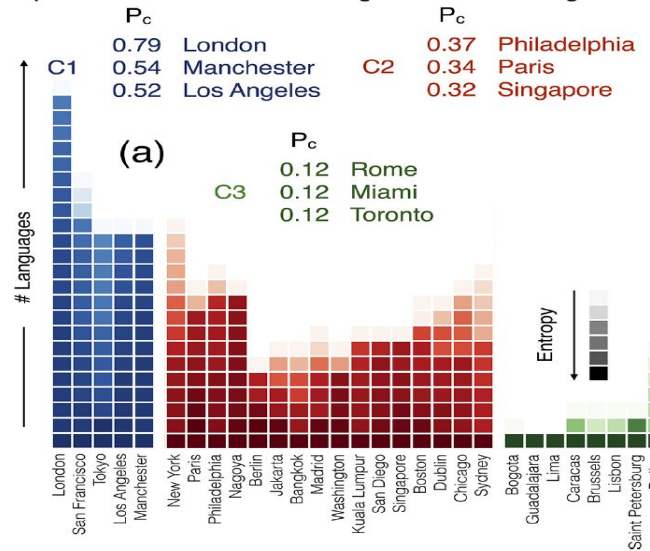
Migrants' integration II



Lamanna et al (2018) proposed a metric to assess the **spatial segregation of immigrant communities** in a subset of 53 most populated cities in the world.

The spatial integration measure (**Power of Integration**) is based on the concept of **Shannon entropy on languages spoken on Twitter**. The immigrant community is considered to be well-integrated into the city if also large portion of native residents live in the same neighborhoods.

Top 3 Cities / Cluster according to *Power of Integration*



Power of Integration. In (a), three groups of cities show similar behavior in the number of communities detected and in their levels of integration. The height of the bars represents the number of languages (communities) detected in each city.

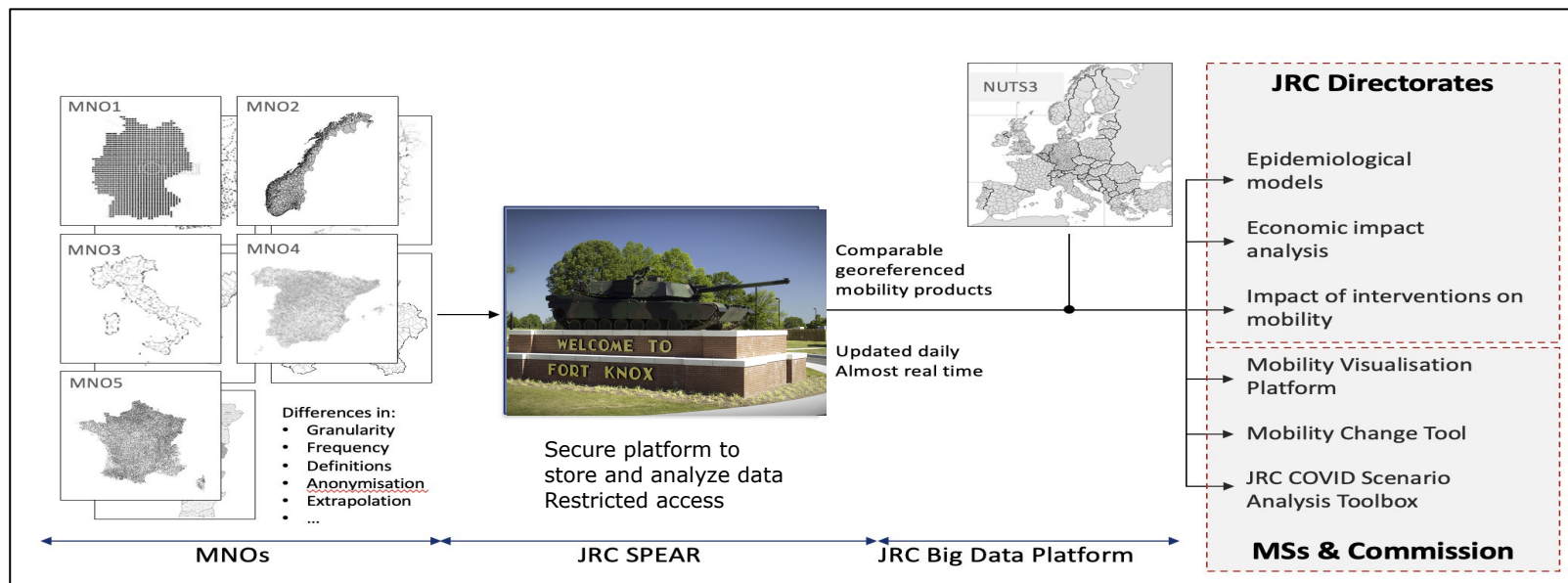
Mobile Phone Data for crisis management

17 Mobile Network Operators, covering **22 EU Member States** (AT, BE, BG, CZ, DE, DK, EE, ES, FI, FR, GR, HR, HU, IE, IT, LT, LV, PT, RO, SE, SI, SK) + NO

→ except CY, MT, PL, LU and NL

Challenges: privacy, fundamental rights, security, commercial sensitivity, communication challenges

Harmonisation: common denominator to de-sensitise and reduce heterogeneity at the price of further aggregation and normalisation



Mobility Functional Areas (MFA) - Mobile Phone Data

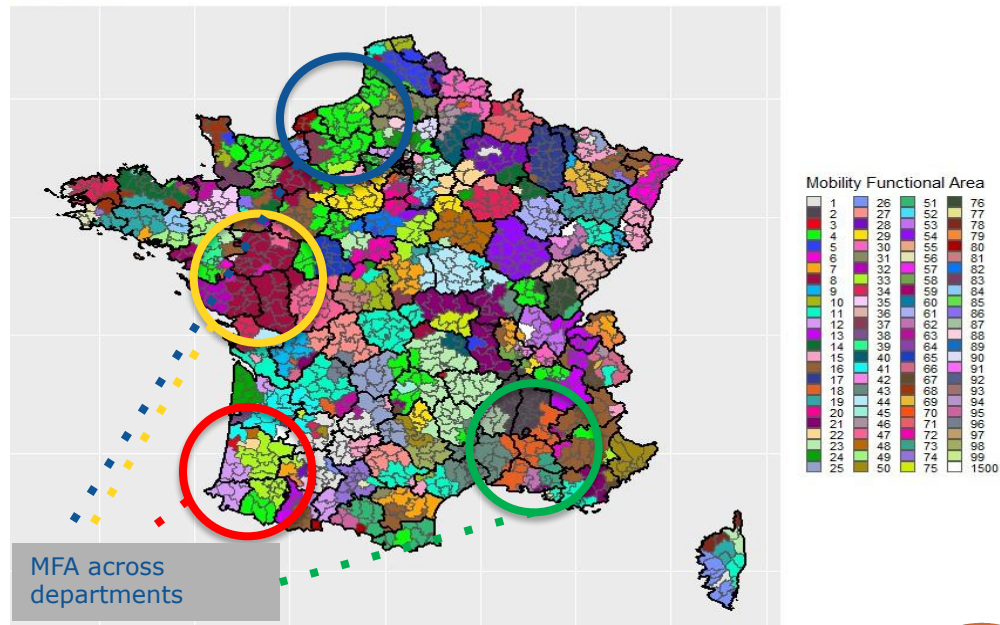
Mobility Functional Areas (MFA) are data-driven concepts of territorial units forming clusters identified through the inbound and outbound movements between territorial units looking at **mobile phone** traces.

MFA's often **spread across multiple administrative areas** and/or cover only part of them.

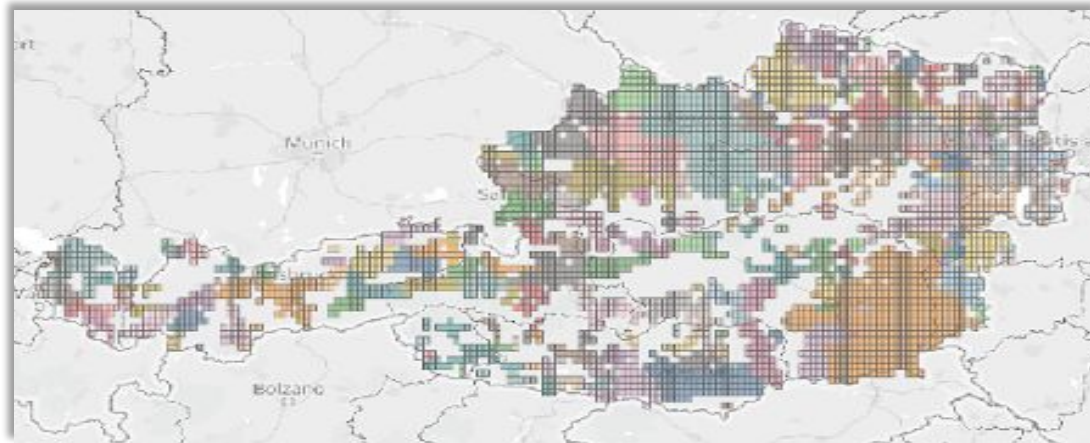
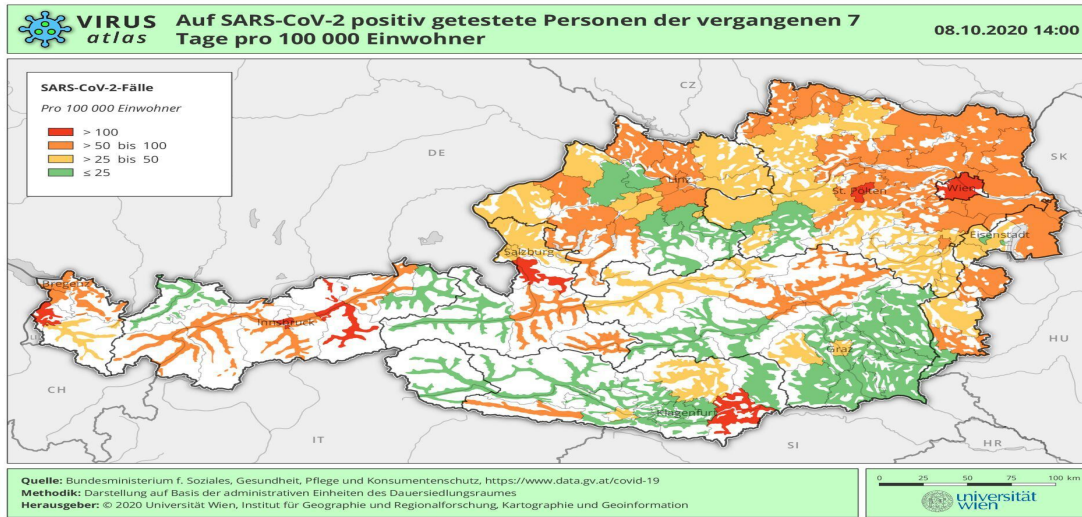
The **MFA's change with time** and are cyclical, like mobility

There is **not necessarily a central node** for the **MFA's** (compared, e.g., to FUA – Functional Urban Areas).

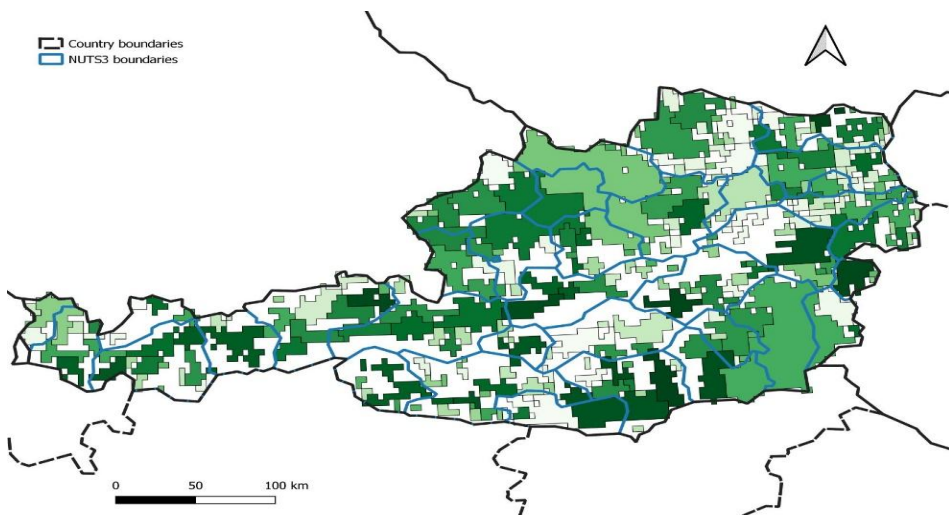
2020-01-13



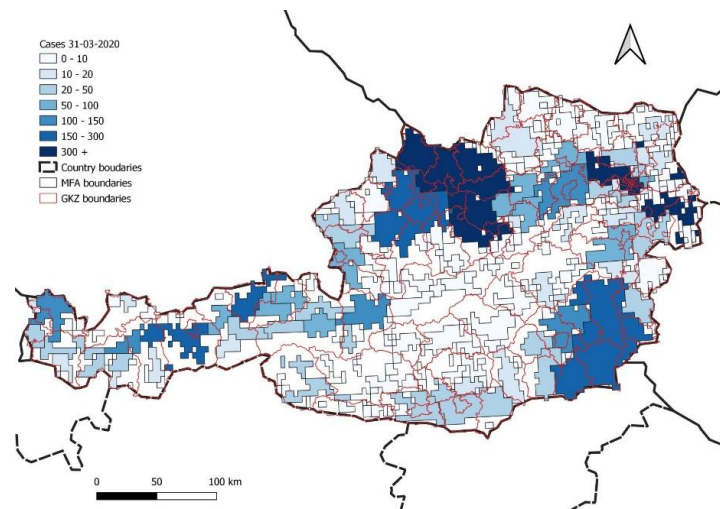
Mobility Functional Areas and Epidemiology



Mobility Functional Areas and Epidemiology



MFAs versus administrative borders

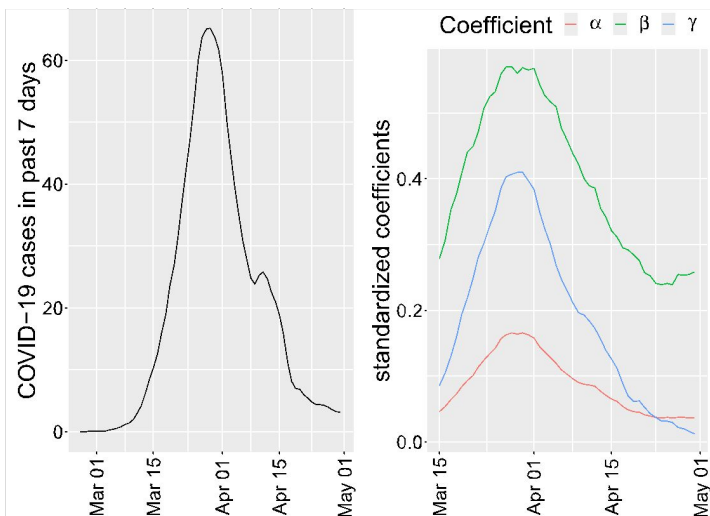


MFAs versus COVID-19 cases (31-03-2020)

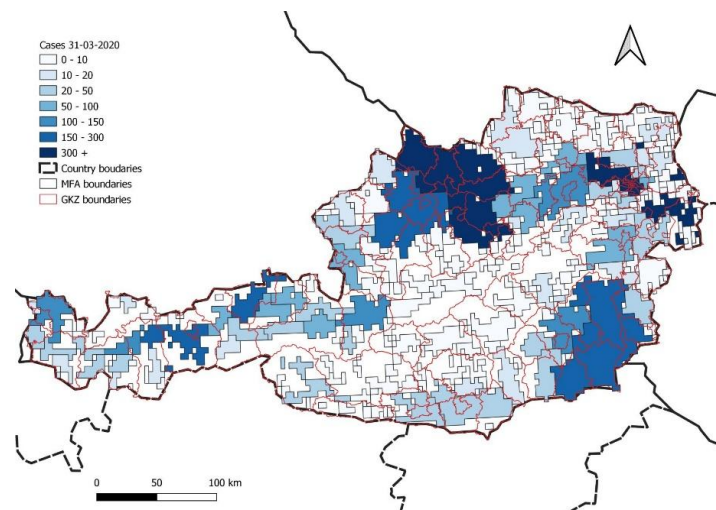
Mobility Functional Areas and Epidemiology

$$y_{i,t} = \alpha + \beta \cdot \text{mfaInd}_i + \gamma \cdot \text{population}_i,$$

(simple regression model)



Relative impact of MFA wrt population



MFAs versus COVID-19 cases (31-03-2020)

Pros & Cons of Digital Traces

	innovative	traditional
Purpose of data collection	commercial and other purposes	administrative/statistical purposes
Collected by	mobile devices, sensors, users of internet applications	trained professionals
Transparent methodology	rarely	always
Data heterogeneity	high	low
Demographic representativeness	can be low	high
Spatial and temporal completeness	varies	harmonized
Reference area	user/device location	administrative areas
Reference time period	undefined, continuous, can be real-time	defined
Timeliness	high	low
QC/QA mechanism	unknown	present
Data cost	varies but can be low	high
Stability/Sustainability of the data source	low	high

(Why) Harvard Dataverse

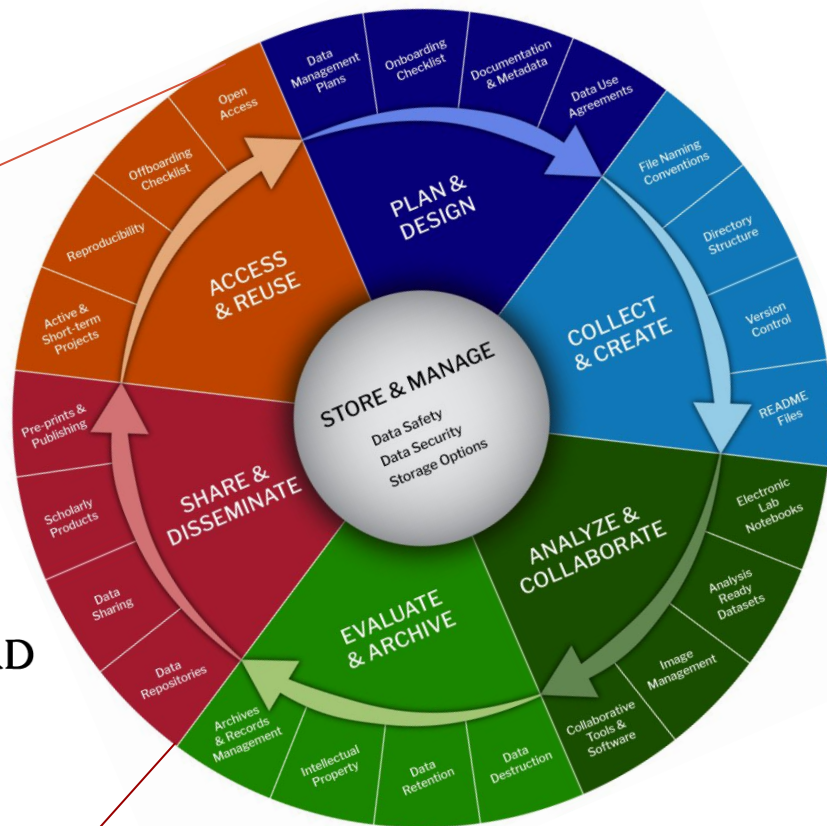
Research Data Management

“The **active and ongoing** management of data **through its lifecycle** of interest and usefulness to scholarship, science, and education.”



Data Sharing

- Depositing data in a **repository**
- Choosing data **licenses**
- Applying **metadata** to **make** published **data** more **findable (FAIR)**
- Write good documentation so shared **data** is actually **reusable**
- Steward shared data over its **useful term**



What is Dataverse?

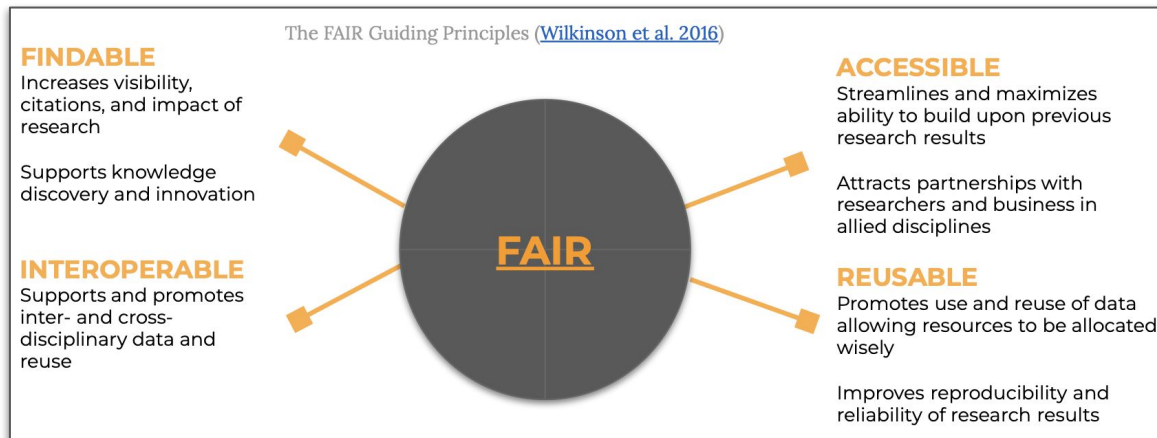
An **open-source** platform that provides a **generalist** repository to **publish, cite, and archive research data**

Built to support **multiple types of data, users, and workflows**

Supports **FAIR** principles and **Signposting**.

Developed mainly at Harvard's Institute for Quantitative Social Science (IQSS) since 2006 + key contributors from our large community

Started as a data sharing platform for the social science now **covers a wide range of disciplines**.



Agricultural Sciences 4,904

Arts and Humanities 36,716

Astronomy and Astrophysics 1,350

Business and Management 2,341

Chemistry 955

Computer and Information Science 3,798

Earth and Environmental Sciences 9,554

Engineering 2,292

Law 5,849

Mathematical Sciences 722

Medicine, Health and Life Sciences 10,548

Physics 1,760

Social Sciences 64,287

What is Dataverse?

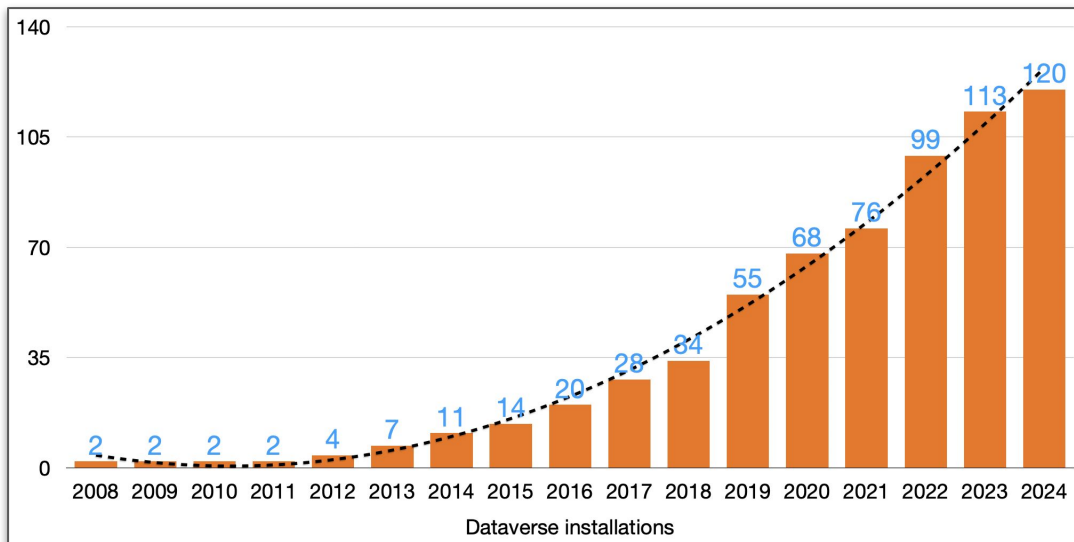
An **open-source** platform that provides a **generalist** repository to **publish, cite, and archive research data**

Built to support **multiple types of data, users, and workflows**

Supports **FAIR** principles and **Signposting**.

Developed mainly at Harvard's Institute for Quantitative Social Science (IQSS) since 2006 + key contributors from our large community

Started as a data sharing platform for the social science now **covers a wide range of disciplines**.



Agricultural Sciences 4,904

Arts and Humanities 36,716

Astronomy and Astrophysics 1,350

Business and Management 2,341

Chemistry 955

Computer and Information Science 3,798

Earth and Environmental Sciences 9,554

Engineering 2,292

Law 5,849

Mathematical Sciences 722

Medicine, Health and Life Sciences 10,548

Physics 1,760

Social Sciences 64,287

What is Dataverse?

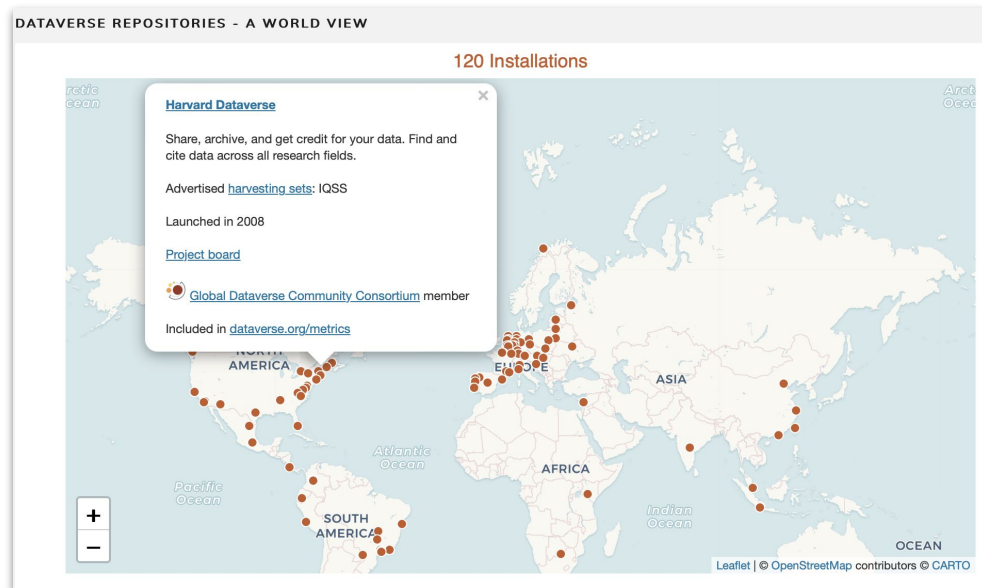
An **open-source** platform that provides a **generalist** repository to **publish, cite, and archive research data**

Built to support **multiple types of data, users, and workflows**

Supports **FAIR** principles and **Signposting**.

Developed mainly at Harvard's Institute for Quantitative Social Science (IQSS) since 2006 + key contributors from our large community

Started as a data sharing platform for the social science now **covers a wide range of disciplines**.



Agricultural Sciences 4,904

Arts and Humanities 36,716

Astronomy and Astrophysics 1,350

Business and Management 2,341

Chemistry 955

Computer and Information Science 3,798

Earth and Environmental Sciences 9,554

Engineering 2,292

Law 5,849

Mathematical Sciences 722

Medicine, Health and Life Sciences 10,548

Physics 1,760

Social Sciences 64,287

What is Dataverse?

An **open-source** platform that provides a **generalist** repository to **publish, cite, and archive research data**

Built to support **multiple types of data, users, and workflows**

Supports **FAIR** principles and **Signposting**.

Developed mainly at Harvard's Institute for Quantitative Social Science (IQSS) since 2006 + key contributors from our large community

Started as a data sharing platform for the social science now **covers a wide range of disciplines**.

- **datasets**
 - **177 K** Harvard DV
 - **[426 K]** whole DV network
- **files**
 - **1.8 M** Harvard DV
 - **[6.12 M]** whole DV network
- **downloads**
 - **63.1 M** (1.3 M / month) Harvard DV
 - **[86.5 M]** whole DV network
- **storage**
 - **< 70 TB** Harvard DV [as of **end 2023**]
- Contributed by **70K users** overall [Harvard DV]
 - 5.3 K unique users depositing data in 2023
 - 4.0 K unique users downloading data in 2023
 - **8.3 K unique active users** in 2023

Agricultural Sciences 4,904

Arts and Humanities 36,716

Astronomy and Astrophysics 1,350

Business and Management 2,341

Chemistry 955

Computer and Information Science 3,798

Earth and Environmental Sciences 9,554

Engineering 2,292

Law 5,849

Mathematical Sciences 722

Medicine, Health and Life Sciences 10,548

Physics 1,760


Social Sciences 64,287

Deluge of Large Data

2004 - 2023: < 70TB

Gen - Apr '24: + 35 TB (+50%)

June '24: +420 TB (a single project)

Sustainability model: 
data owner contributes to storage costs

Moving to **MOC** (Mass Open Cloud):

- no egress costs
- 30% (storage/computing) cost wrt external cloud solutions



What is special about Dataverse?

The concept of “**collection**”

Datasets with standardized **metadata**

Upload data files/documentation, with metadata

Publish and share capabilities

Link your datasets to coauthors and other data

Persistent identifiers: (**DOIs, ORCID, etc**)

Dataset and file level **citations**

Set “**terms of access**” and **licenses**

Export metadata

Export data citation

Private URL to share your dataset in draft format

Data Analysis, File Previewer, **AI** integration

File folder hierarchy preservation

Restrict/Open files for **access/request access**

Workflows for data deposit and publishing

Custom metadata blocks

Metadata Types

1. **Citation Metadata:** any metadata that would be needed for generating a data citation and other general metadata that could be applied to any dataset;
2. **Domain Specific Metadata:** with specific support currently for Social Science, Life Science, Geospatial, and Astronomy datasets;
3. **File-level Metadata:** varies depending on the type of data file and include options like file tags, descriptions, variable names, and hierarchy preservation.

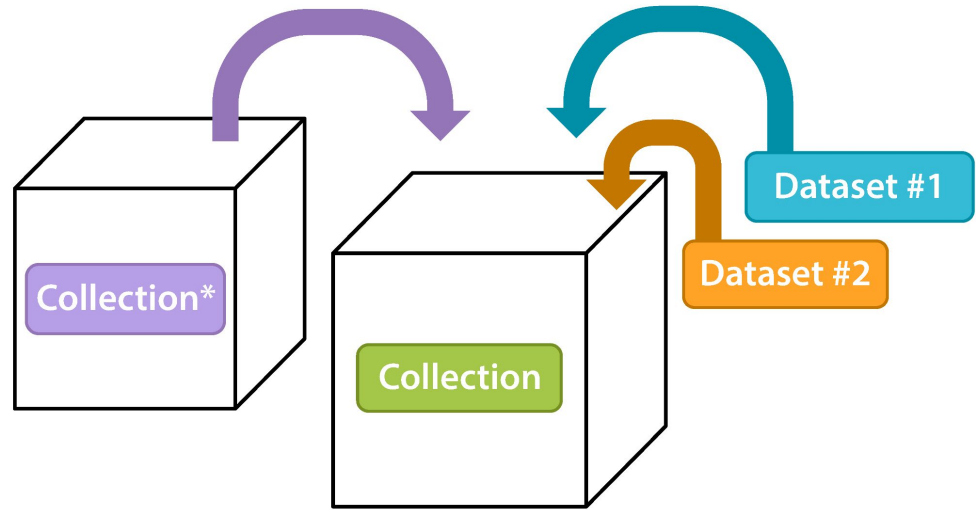
Provenance,
authorship, etc

Higher
granularity of
information

Dataverse Collections

- Ability to create Dataverse **collections** to organize datasets according to your needs
- Dataverses collections can also contain other collections, enabling any **hierarchical** structure
- **Different rules** can be applied for different Dataverse collections, e.g. for Metadata, Permissions, etc.

Schematic Diagram of a **Collection** in Dataverse Software 5.0



Container for your **Datasets** and/or **Collections***

* Collections can contain other Collections

Dynamic Metadata

- **Metadata is defined dynamically** at the database level, allowing for modularly adding new Metadata blocks
- Supports:
 - single or **multiple** values
 - simple or **compound** values
 - controlled vocabularies
 - **external** vocabularies

Choose the metadata fields to use in dataset templates and when adding a dataset to this dataverse.

- Citation Metadata (Required) [\[+\] View fields + set as hidden, required, or optional](#)
- Geospatial Metadata [\[+\] View fields](#)
- Social Science and Humanities Metadata [\[+\] View fields](#)
- Astronomy and Astrophysics Metadata [\[+\] View fields](#)
- Life Sciences Metadata [\[+\] View fields](#)
- Journal Metadata [\[+\] View fields](#)

Citation Metadata [^](#)

Title * [?](#)

Author * [?](#)

Name * ?	<input type="text" value="Admin, Dataverse"/>	Affiliation * ?	<input type="text" value="Dataverse.org"/>	<input type="button" value="+"/>
Identifier Scheme * ?	<input type="text" value="Select..."/>	Identifier * ?	<input type="text"/>	

Contact * [?](#)

Name * ?	<input type="text" value="Admin, Dataverse"/>	Affiliation * ?	<input type="text" value="Dataverse.org"/>	<input type="button" value="+"/>
E-mail * ?	<input type="text" value="dataverseadmin@iq.harvard.edu"/>			

Description * [?](#)

This field supports only certain [HTML tags](#).

Text * [?](#)

Date * [?](#)

CroissantML: AI/ML-Ready Datasets Simplified

arXiv > cs > arXiv:2403.19546v1 Search... Help | Ad

Computer Science > Machine Learning

[Submitted on 28 Mar 2024 (this version), latest version 30 May 2024 (v2)]

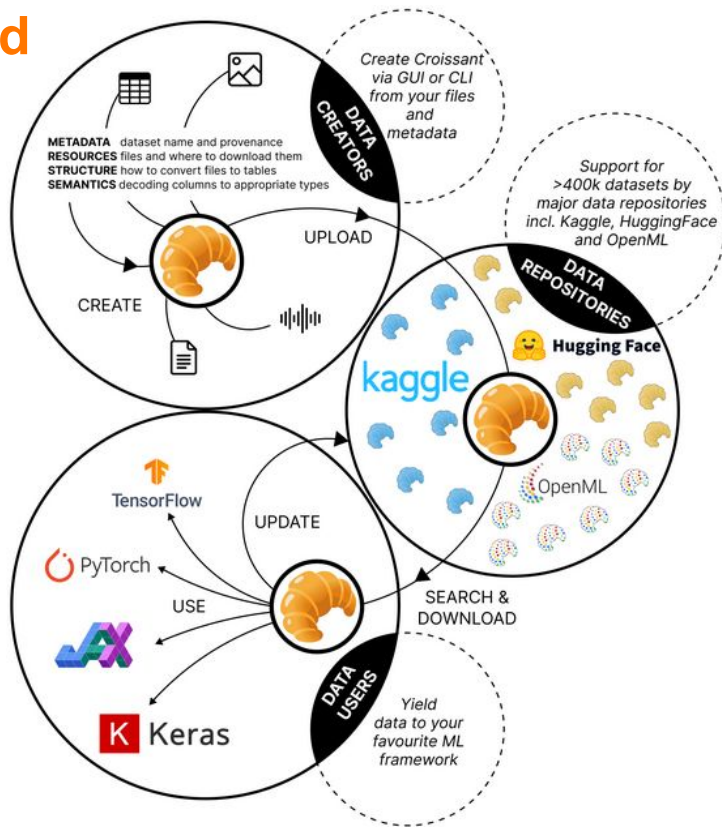
Croissant: A Metadata Format for ML-Ready Datasets

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Joan Giner-Miguel, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruysen, Rajat Shinde, Elena Simperl, Geoffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Steffen Vogler, Carole-Jean Wu

Data is a critical resource for Machine Learning (ML), yet working with data remains a key friction point. This paper introduces Croissant, a metadata format for datasets that simplifies how data is used by ML tools and frameworks. Croissant makes datasets more discoverable, portable and interoperable, thereby addressing significant challenges in ML data management and responsible AI. Croissant is already supported by several popular dataset repositories, spanning hundreds of thousands of datasets, ready to be loaded into the most popular ML frameworks.

<https://arxiv.org/abs/2403.19546v1>

IQSS/DV contributed to definition of this standard for ML/AI workflows. HDV exposes CroissantML metadata to crawlers to increase discoverability of this type of data and has a built-in exporter.



Flexible Permission System

- Supports multiple workflows by controlling who can add to your Dataverse collection, what they can do, and what role they have on and created Datasets
- Roles are defined as a set of permissions to grant to users or to groups
- **Groups** can be defined **statically** or **dynamically** (e.g. users logging in from the same institution)

Edit Access

Who can add to this dataverse?

- Anyone adding to this dataverse needs to be given access
- Anyone with a Dataverse account can add sub dataverses
- Anyone with a Dataverse account can add datasets
- Anyone with a Dataverse account can add sub dataverses and datasets

When a user adds a new dataset to this dataverse, which role should be automatically assigned to them on that dataset?

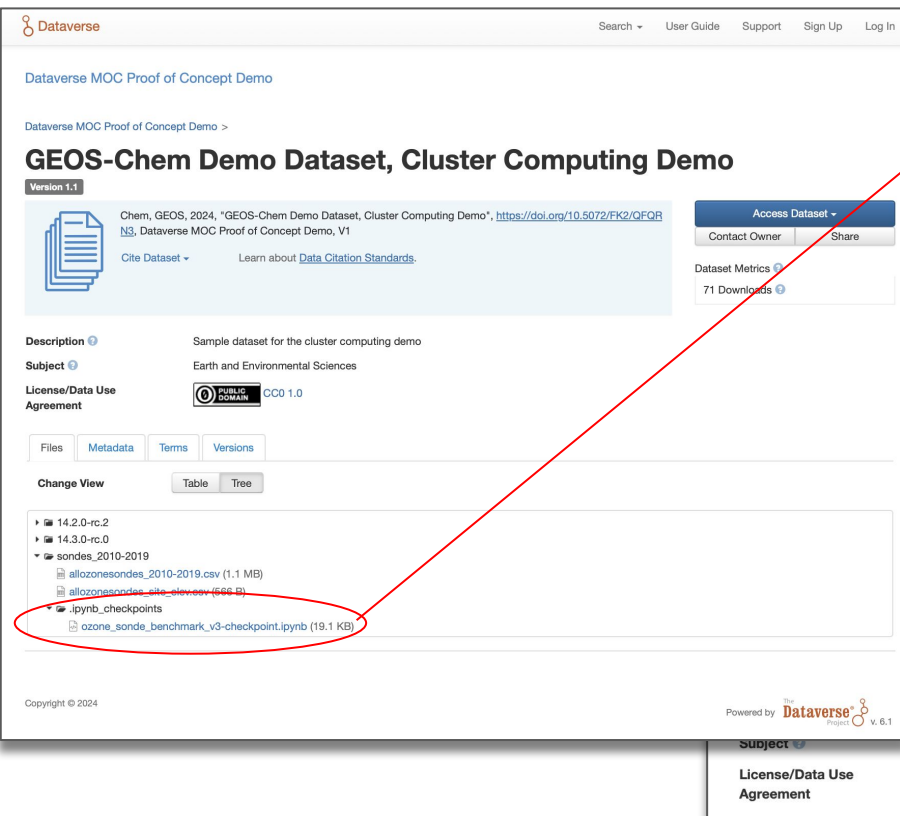
- Contributor - Edit metadata, upload files, and edit files, edit Terms, Guestbook, Submit datasets for review
- Curator - Edit metadata, upload files, and edit files, edit Terms, Guestbook, File Restrictions (Files Access + Use), Edit Permissions/Assign Roles + Publish

Save Changes

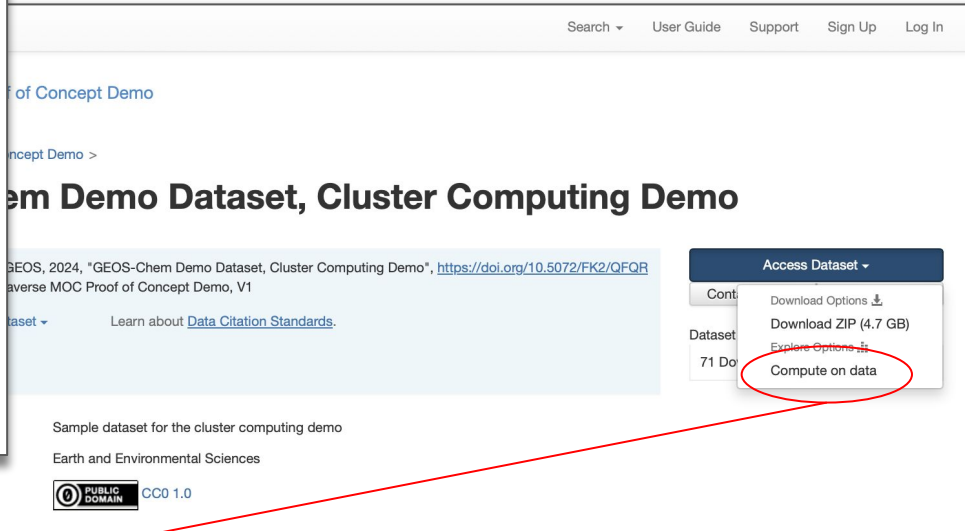
Cancel

2 Users/Groups

User/Group Name (Affiliation) ⚙	ID ⚙	Role ⚙
Dataverse Admin (Dataverse.org)	@dataverseAdmin	Admin
Anyone with a Dataverse account	:authenticated-users	Dataverse + Dataset Creator



If Dataverse sees a (python) **notebook**, the new menu item “**Compute on data**” appears in the *Access Dataset* drop down menu



This **menu will launch** the JupyterLab VM with the pre-loaded notebook taken from the dataset. All files in this collection are seen as **local** to the Jupyter instance. Python will simply load them into memory for computing purposes.

“Ask the data” DV chatbot

HARVARD
Dataverse

Add Data - Search - About - User Guide - Support - Sign Up - Log In

Harvard Dataverse >

Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends

Version 1.0

[Access Dataset -](#)
Contact Owner | Share

[Cite Dataset -](#) | [Learn about Data Citation Standards.](#)

Description

Research into the causes of violence against civilians has increased significantly in recent years, yet the mechanisms governing spatial patterns of victimization remain poorly understood. My investigation explores if and why one specific locality, capital cities, experiences a higher frequency of violence against civilians perpetrated by armed insurgent organizations. I argue that the political value associated with capitals allows these groups to asymmetrically impose higher costs on the regime by targeting civilians in these localities. I lay out and validate three specific mechanisms to explain this pattern: elite coercion, popular intimidation, and international persuasion. In the first scenario insurgents aim to influence domestic elites directly. In the second, they aim to affect domestic civilians' resolve. In the third, they seek to influence international audiences. Using new geolocated global atrocities data for the years 1996-2009, I evaluate this linkage by employing different methodological approaches and accounting for potential reporting biases. Finally, I show that ethnic and secessionist wars are more likely to experience atrocities in the capital compared with other conflicts. The findings illustrate potential benefits from explaining the temporal and spatial variation in violence by insurgents, with a focus on strategic conditions and power asymmetries. (2017-09-21)

Subject Social Sciences

Keyword Political violence; Civilian victimization; Conflict; Spatial analysis

License/Data Use Agreement CC0 1.0

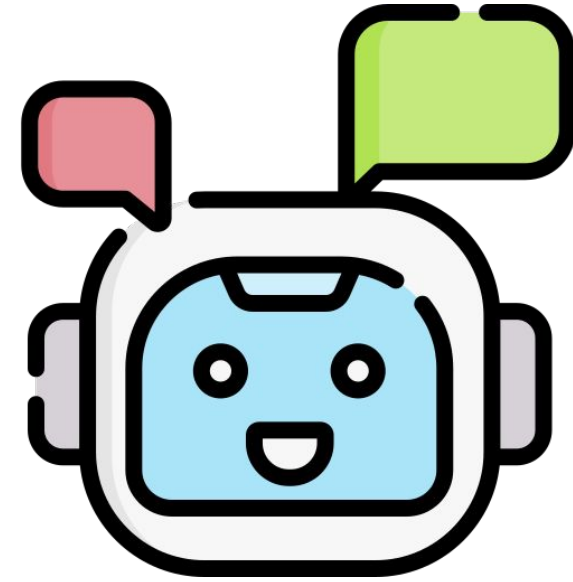
Files | Metadata | Terms | Versions

Change View Table | Tree

Search this dataset...

Filter by
File Type: All - Access: All - [Sort -](#)

21 to 30 of 43 Files		Download -
	A.8. Forecasting and Cross Validation Exercises.R Replication files 09_20_17/ R Syntax - 4.9 KB Published Sep 21, 2017 39 Downloads MDS: 9f8...7a5	
	A.9. Capital Atrocities and Regime Survival.R Replication files 09_20_17/ R Syntax - 11.1 KB Published Sep 21, 2017 39 Downloads MDS: 7c1...990	
	eth_sec_war_dat.tab Replication files 09_20_17/Data/Conflict Type Analysis/ Tabular Data - 50.5 MB Published Sep 21, 2017 46 Downloads 48 Variables, 214840 Observations UNF:6:URZ8...MBw===	



This chatbot only sees the tabular data but is clueless about the metadata

Dataset Citation

Paolucci, Matteo, 2020, "Stroke and COVID-19 cases in March 2020 compared to March 2019 in the Italian provinces of Forlì-Cesena and Rimini", <https://doi.org/10.7910/DVN/EVHWSB>, Harvard Dataverse, V2, UNF:6:Cs2UBASEs+8UI/398OA+GA== [fileUNF]

Cite Dataset ▾

Learn about [Data Citation Standards](#).

File Tools

Metadata

Versions

File Tools ▾

 Open in New Window

Tell me what you want to know

create a regression model that explains the number of strokes as a function of covid cases (show me the exact variable names you used). Show me the coefficients and predict the number of strokes when covid cases = 50

Simple statistical reasoning

Answer please

The regression model used the variable `COVID-19FCcuminc` to explain the number of strokes (`FCstrokecuminc20`). The coefficients of the model are as follows: - Constant: 1.205324 - COVID-19FCcuminc: 5.159673 When the number of COVID cases is 50, the predicted number of strokes is approximately 259.189.

March2020	COVID-19FCcuminc	COVID-19RNcuminc	COVID-19FC+RNcuminc	E	FCstrokecuminc20	FCcumnumb20	RNstrokecuminc20	RNstrokecum
1	0.25	4.7	2.31		1.2670770307	5	0.5880191106	2
2	0.25	5.59	2.72		2.2807386552	9	1.7640573319	6
3	0.25	7.06	3.4		3.2944002798	13	3.2341051084	11
4	0.51	9.7	4.76		4.8148927166	19	4.4101433297	15
5	1.01	19.99	9.8		6.0819697473	24	5.2921719956	18
6	1.01	27.34	13.2		6.8422159657	27	6.4682102168	22
7	1.77	30.58	15.11		7.0956313718	28	7.9382579934	27
8	3.8	33.22	17.42		8.3627084025	33	8.8202866593	30
9	4.05	48.77	24.5		9.376370027	37	9.9963248806	34

Tell me what you want to know

¿Cuántas personas fuman en este conjunto de datos?

Answer please

60 personas fuman en este conjunto de datos.

Tell me what you want to know

combien y a-t-il de fumeurs dans cet ensemble de données ?

Answer please

Il y a 60 fumeurs dans cet ensemble de données.

Tell me what you want to know

hvor mange røykere er det i dette datasettet?

Answer please

There are 60 smokers in this dataset.

Tell me what you want to know

このデータセットには喫煙者が何人いますか?

Answer please

このデータセットには60人の喫煙者がいます。

for some reasons
German does not
work for this data

```

scutor chain...
les

/data" seems to be the only table
a person smokes or not.

CREATE TABLE user (
  "r"+u" VARCHAR

Smoke
No
Yes
No

e has a "Smoke" column which indic
his table.

Action: sql_db_query_checker
Action Input: "SELECT COUNT(*) FROM data WHERE Smoke = 'Y
Observation: The original query seems to be correct as it
s'. There are no joins, unions, between statements, or f
correct as well, assuming 'Smoke' is a text column. The

Here is the reproduced query:

SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'
Thought:The query seems to be correct. Now I can execute it to get the number of people
Action: sql_db_query
Action Input: "SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'"
Observation: [(60,)]
Thought:I now know the final answer
Final Answer: 60 personas fuman en este conjunto de datos.

> Finished chain.

```

このデータセットには60人の喫煙者がいます。

Kono datasetto ni wa 60-ri no kitsuen-sha ga imasu.

↓ 🔊

There are 60 smokers in this dataset.

Thank you!



Contact: siacus@iq.harvard.edu