Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# DATA, COMPUTATION, and SOCIETY: The New Computational Social Science Program at the Barcelona Supercomputing Center

**Mercè Crosas, Ph.D.**
**BSC, CODATA**

Conference on Smart Metadata for Official Statistics

COSMOS

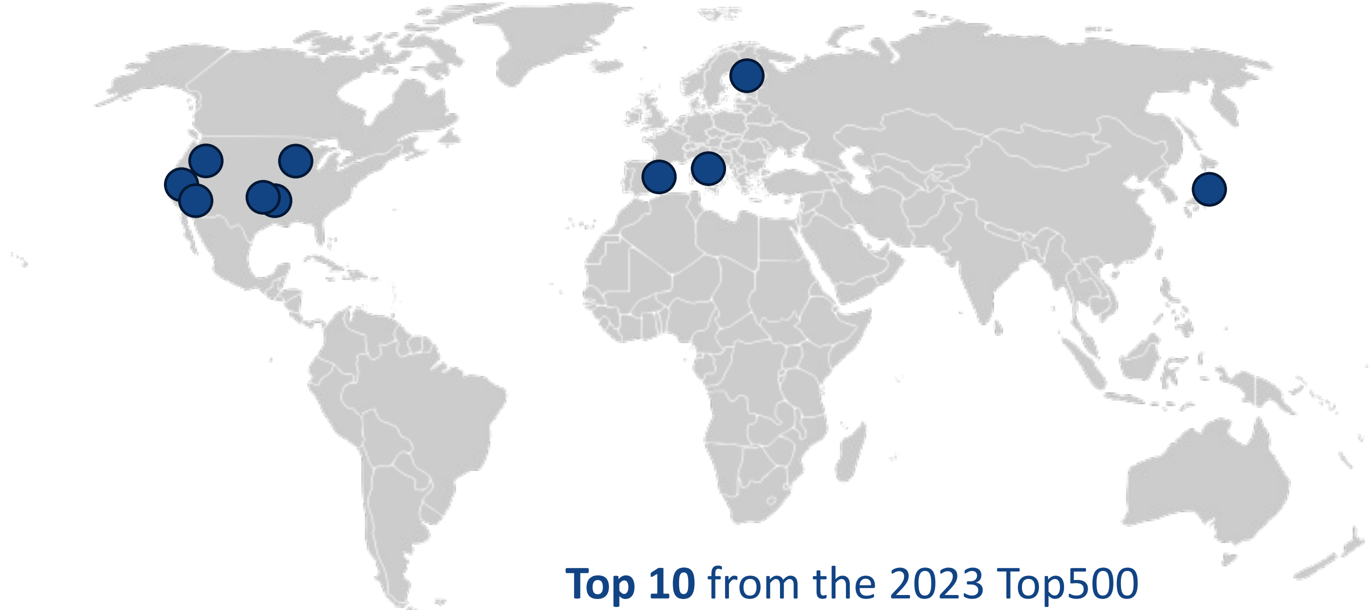
11-12 April 2024
— Paris

1. BSC and the new Program
2. Computational Social Sciences
3. Computational Social Sciences
4. Initial Research Projects
5. Data Collaborations and CODATA
6. Conclusions

# Barcelona Supercomputing Center (BSC)

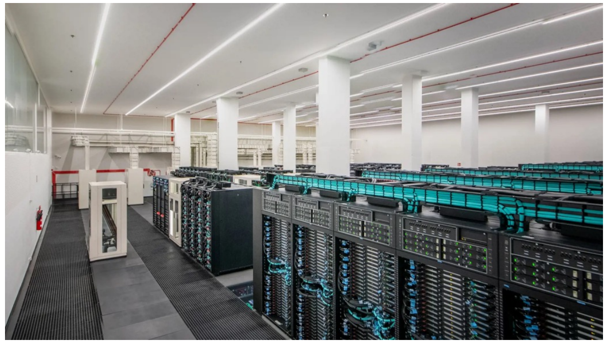
- 2004: Installation of first MareNostrum, ranked 4 in world Top500 list, first in Europe
- 2005: Official creation of the BSC
- A public consortium (Spanish gov, Catalan gov, UPC)
- Provides supercomputing services to Spanish and EU researchers (EuroHPC):
- Now, 1000+ people

BSC/ Centro Nacional de Supercomputación

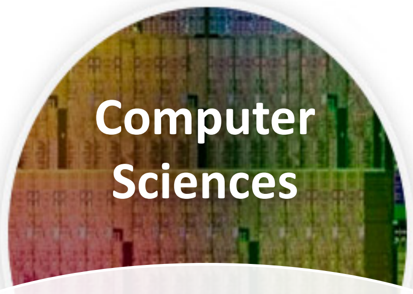


**Top 10** from the 2023 Top500

- **2023: MareNostrum 5, one of three pre-exascale supercomputers in Europe, 8th in the world**
- 300+ Petaflops peak performance
- 600+ Petabytes Data storage



# BSC is also a Research Organization



**Computer Sciences**

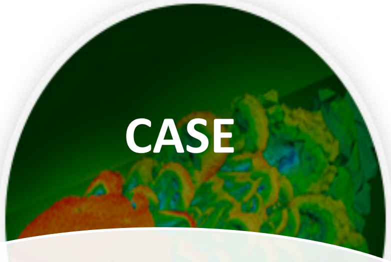
To influence the way machines are built, programmed and used: programming models, performance tools, Big Data, Artificial Intelligence , computer architecture, energy efficiency

**Earth Sciences**

To develop and implement global and regional state-of-the-art models for short-term air quality forecast and long-term climate applications

**Life Sciences**

To understand living organisms by means of theoretical and computational methods (molecular modeling, genomics, proteomics)



**CASE**

To develop scientific and engineering software to efficiently exploit super-computing capabilities (biomedical, geophysics, atmospheric, energy, social and economic simulations)

- Until now, there have been 4 scientific departments, all conducting computational and data-intensive research.

- **In 2023, BSC created a new cross-departmental scientific Program for Computational Social Sciences.**

# Computational Social Sciences Program’s Vision

✓ Prepare the social sciences and humanities to benefit from the age of data and AI

✓ Increase collaboration between social scientists and computer scientists

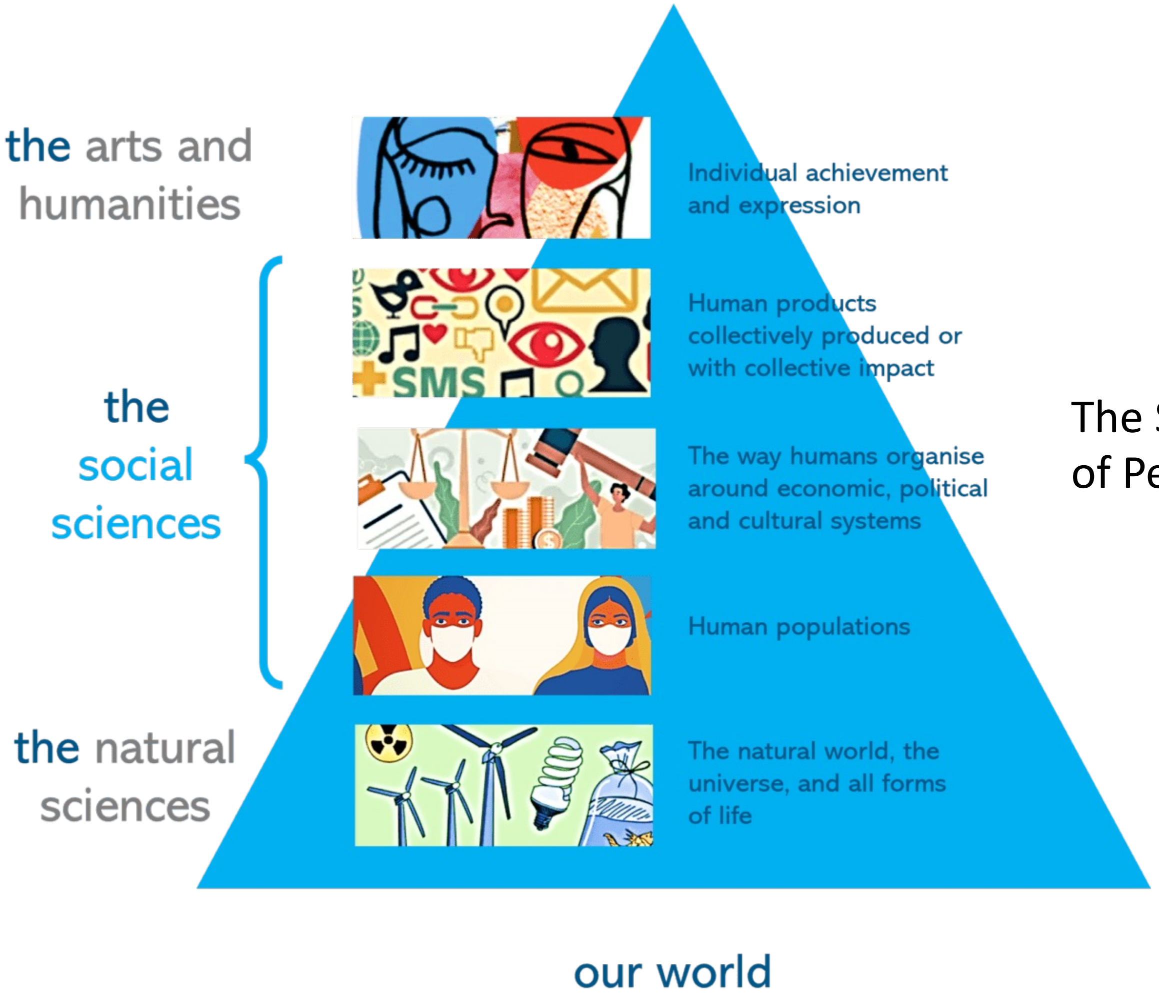
✓ Facilitate the use of High-Performance Computing (HPC) to social science researchers, making HPC more approachable to all scientists

✓ Apply social science research to assist policy making

# Computational Social Sciences

**“Imagine how hard physics would be if particles could think”**

Murray Gell-Mann (physicist, Founder of the Santa Fe Institute for Complexity Science)


the arts and humanities
Individual achievement and expression
the social sciences
Human products collectively produced or with collective impact
The way humans organise around economic, political and cultural systems
Human populations
the natural sciences
The natural world, the universe, and all forms of life
SMS
our world

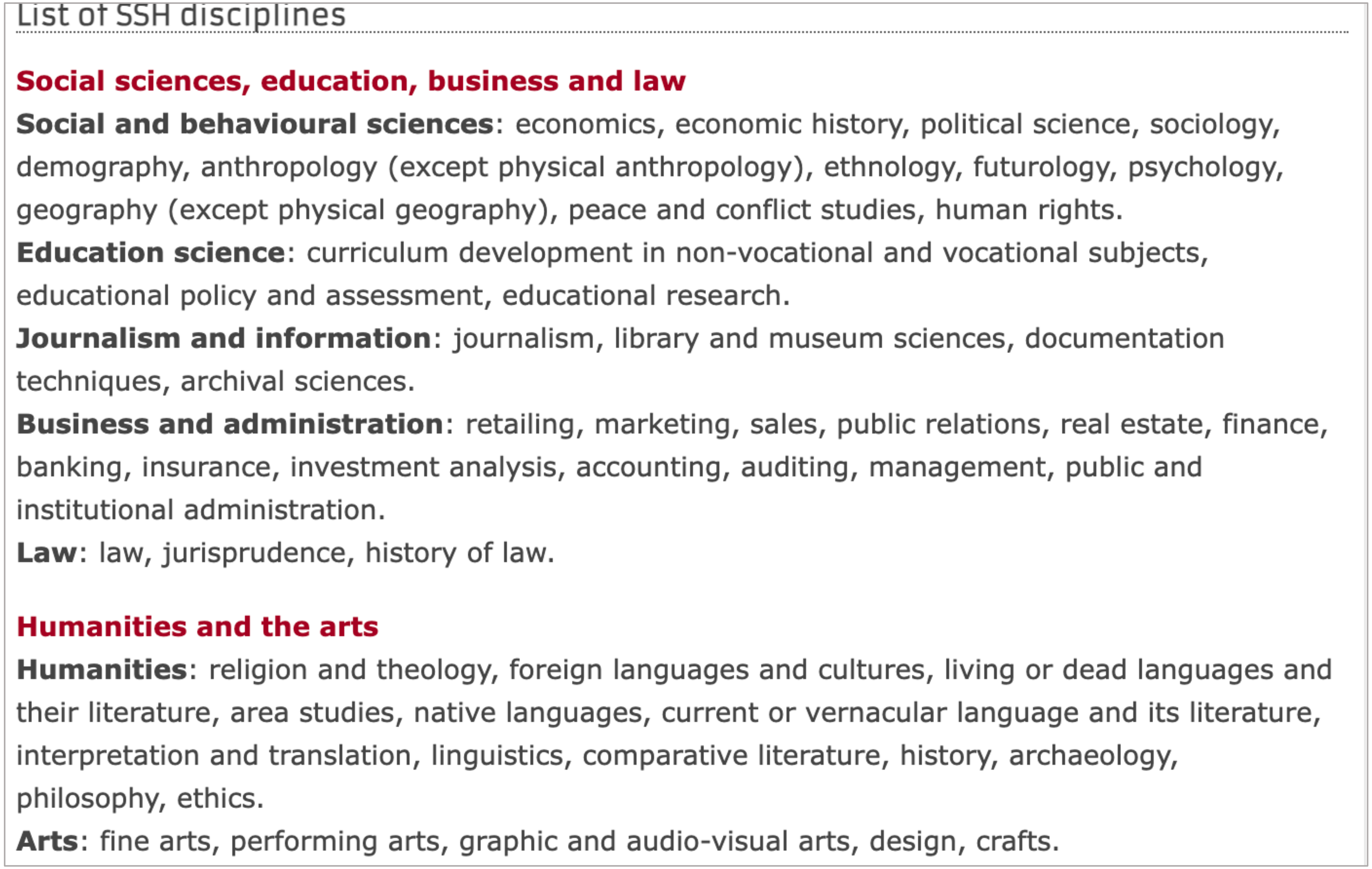
The Study of People

https://stateofthesocialsciences.org.au/about-the-social-sciences/

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# Program´s Scope: Social Science and Humanities (SSH)

- List of SSH disciplines defined by EU
- Based on the UNESCO International Standard Classification of Education



List of SSH disciplines

**Social sciences, education, business and law**

**Social and behavioural sciences**: economics, economic history, political science, sociology, demography, anthropology (except physical anthropology), ethnology, futurology, psychology, geography (except physical geography), peace and conflict studies, human rights.

**Education science**: curriculum development in non-vocational and vocational subjects, educational policy and assessment, educational research.

**Journalism and information**: journalism, library and museum sciences, documentation techniques, archival sciences.

**Business and administration**: retailing, marketing, sales, public relations, real estate, finance, banking, insurance, investment analysis, accounting, auditing, management, public and institutional administration.

**Law**: law, jurisprudence, history of law.

**Humanities and the arts**

**Humanities**: religion and theology, foreign languages and cultures, living or dead languages and their literature, area studies, native languages, current or vernacular language and its literature, interpretation and translation, linguistics, comparative literature, history, archaeology, philosophy, ethics.

**Arts**: fine arts, performing arts, graphic and audio-visual arts, design, crafts.

# Data & Models in Social Sciences

## Census Data, Official Statistics

3800 BC, first census, Babylon

1622, Registration of local citizenship, Wurttemberg, Germany

1665, First modern census, Quebec, Canada

1719, First systematic records in Europe (Prussia)

1790, First Census in the USA

1800s, England, France, Norway,

1828, Australia

1853, Chile passes the first Census Law in South America

*Today, most countries have census data (except for a very few exceptions)*

## Data Archives, Research Repositories

**Surveys, interviews, experimental data**

1924, Odum Institute, first social science research center

1947, Roper Center first social science data archive

1960s, ICPSR social science data archive - **DDI**

UK Data Archive

1970s, IASSIST

CESSDA (UKDA, ZA, DDA, NSD, BASS, ADPSS, Steinmetz)

1990s, Australian social science data archive

1991, IPUMS, Largest microdata database

2006, IQSS, Harvard Dataverse: Sharing social science research data

**Statistical models for social sciences**

## Big Data

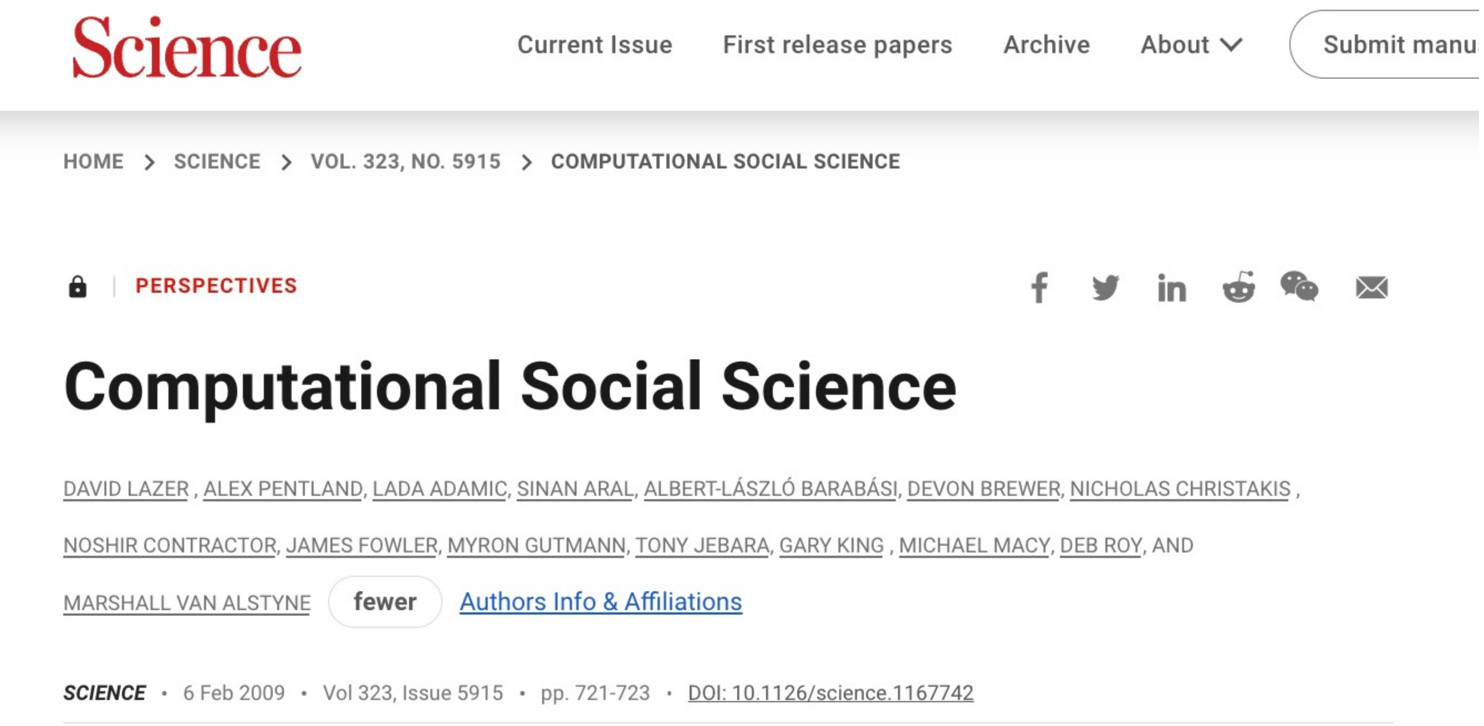
> 2000

social media, web, apps, phones, satellites, sensors, open govs, industrydata, …

**AI models for social sciences**

# "Life in the Network: The Coming Age of Computational Social Science"

**2009**



HOME > SCIENCE > VOL. 323, NO. 5915 > COMPUTATIONAL SOCIAL SCIENCE

PERSPECTIVES

## Computational Social Science

DAVID LAZER , ALEX PENTLAND, LADA ADAMIC, SINAN ARAL, ALBERT-LÁSZLÓ BARABÁSI, DEVON BREWER, NICHOLAS CHRISTAKIS , NOSHIR CONTRACTOR, JAMES FOWLER, MYRON GUTMANN, TONY JEBARA, GARY KING , MICHAEL MACY, DEB ROY, AND MARSHALL VAN ALSTYNE

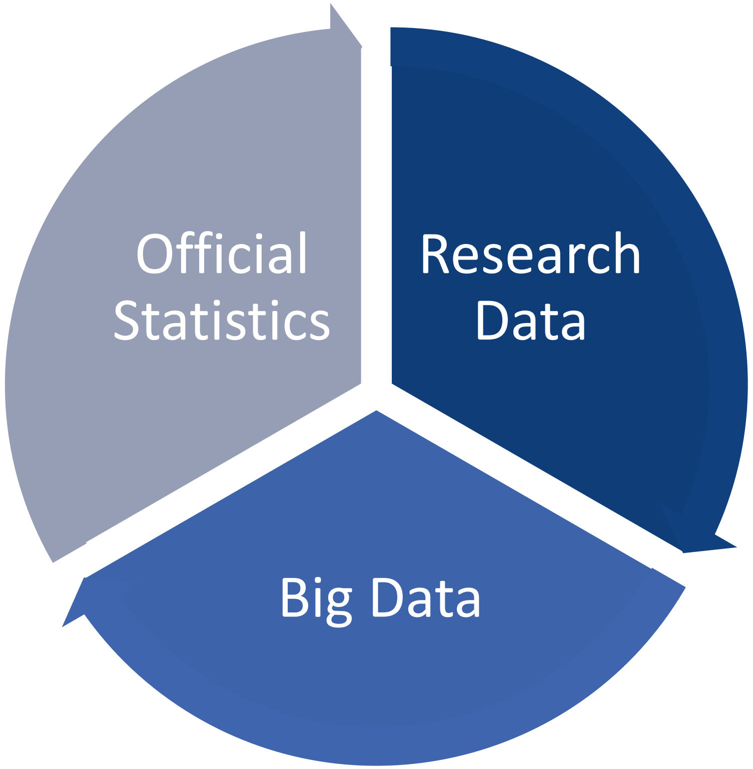
*SCIENCE* • 6 Feb 2009 • Vol 323, Issue 5915 • pp. 721-723 • DOI: 10.1126/science.1167742

"The capacity to **collect and analyze massive amounts of data** has unambiguously transformed such fields as biology and physics.

…

To date the vast majority of existing research on human interactions has relied on **one-shot self-reported data** on relationships. New Technologies (surveillance, email, internet, GPS tracking,….) offer a remarkable **second-by-second picture of interactions over extended periods of time,** providing information about both the structure and content of relationships"

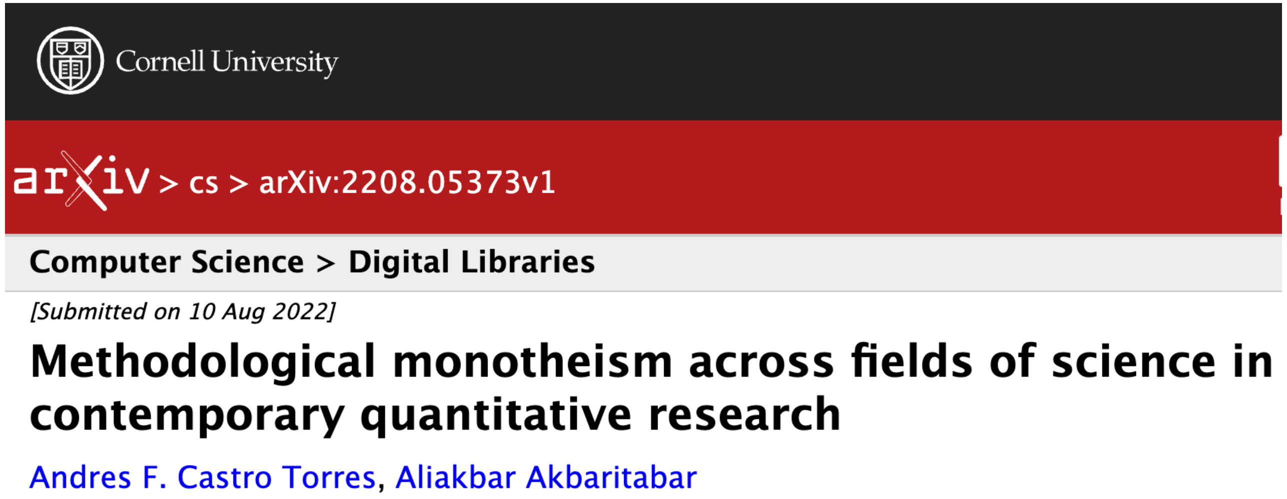
# Program aims to combine data from multiple sectors



**Interoperability across multiple sectors and sources:**

- Harmonize metadata from research datasets, official statistic and governments, and from industry
- Standardize new vocabularies to classify semi-structured data
- Build public-private data partnerships for social science research

# Computational Social Sciences



Cornell University

arXiv > cs > arXiv:2208.05373v1

Computer Science > Digital Libraries

*[Submitted on 10 Aug 2022]*

**Methodological monotheism across fields of science in contemporary quantitative research**

Andres F. Castro Torres, Aliakbar Akbaritabar

Linear regression has been dominant in quantitative social science

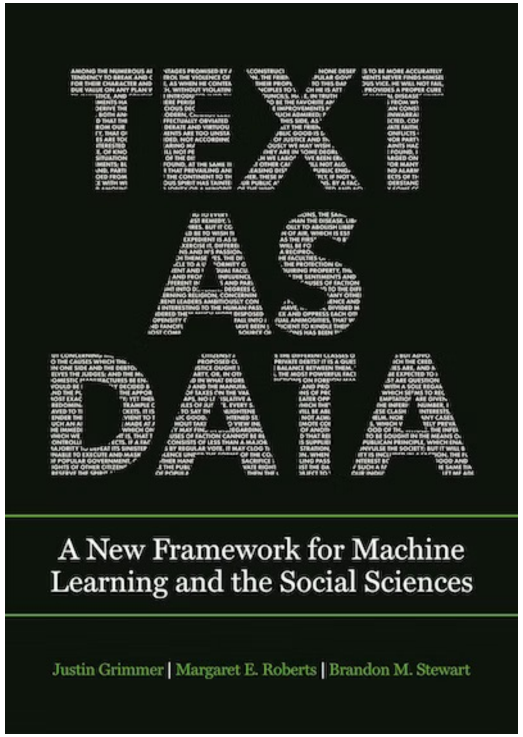
- Using bibliometric data from the Web of Science, conduct a large-scale and cross-disciplinary assessment of the prevalence of linear-model-based research from 1990 to 2022.
- Corpus 1: 7,164,784 articles - articles reporting methods or quantitative data in abstract
- Corpus 2: 13,720,556 articles – any kind of empirical evidence
- **Results:**
  - **High prevalence of linear-model-based research in the social sciences (60%)**

# Recent Methods in Computational Social Science
## (beyond Simple Linear Regressions)

- **Text, image/video as Data (Clustering, Topic Modeling, Positioning)**
  - Artificial Intelligence (AI)/Machine Learning(ML), Natural Language Processing (NLP), Transformers/Large Language Models (LLM)
- **Regression Analysis beyond SLR:**
  - Multilevel models (nested data models)
  - Multivariate models
- **Complex Systems:**
  - Social Network Analysis (SNA)
  - Agent-Based Modeling (ABM)
  - (Scaling theory)

A wider variety of methods are now used as more computing power and software become available

# Text as Data in Social Sciences


TEXT AS DATA
A New Framework for Machine Learning and the Social Sciences
Justin Grimmer | Margaret E. Roberts | Brandon M. Stewart

**From text to data:**

- N dimensional space, distance among documents

**Used in Social Science for:**

- Discovery
- Measurement
- Inference: predictive and causal

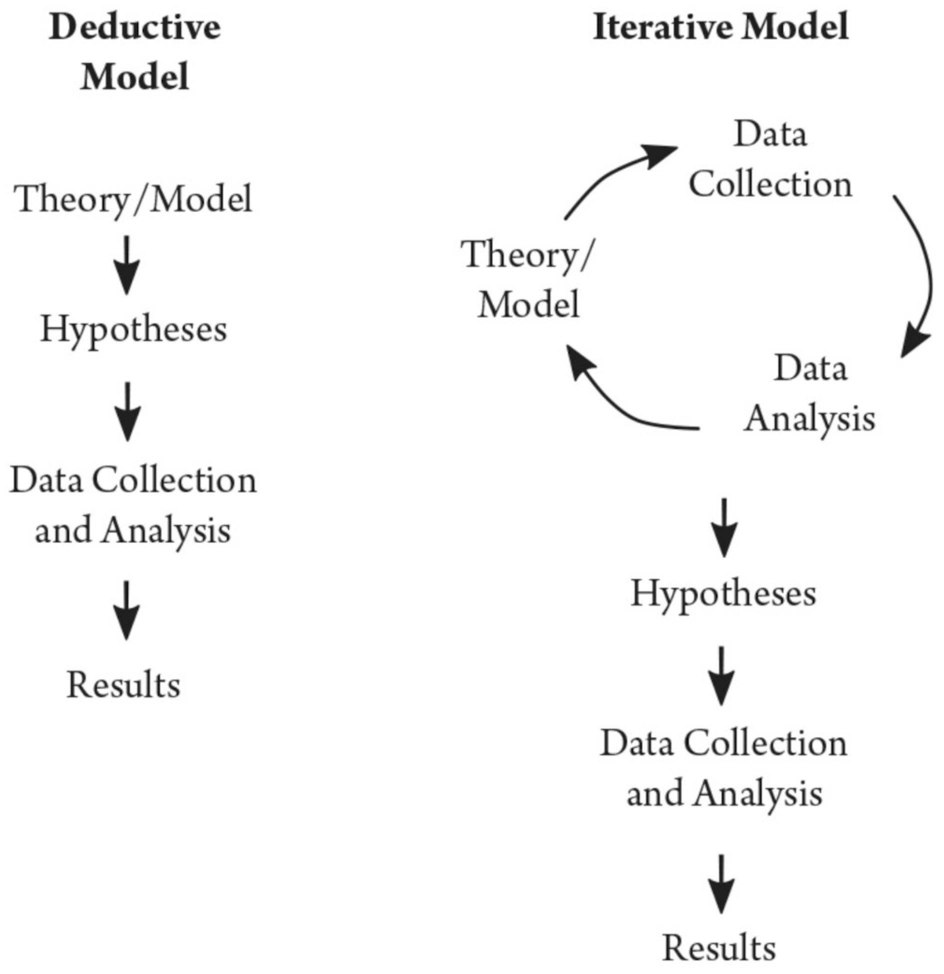
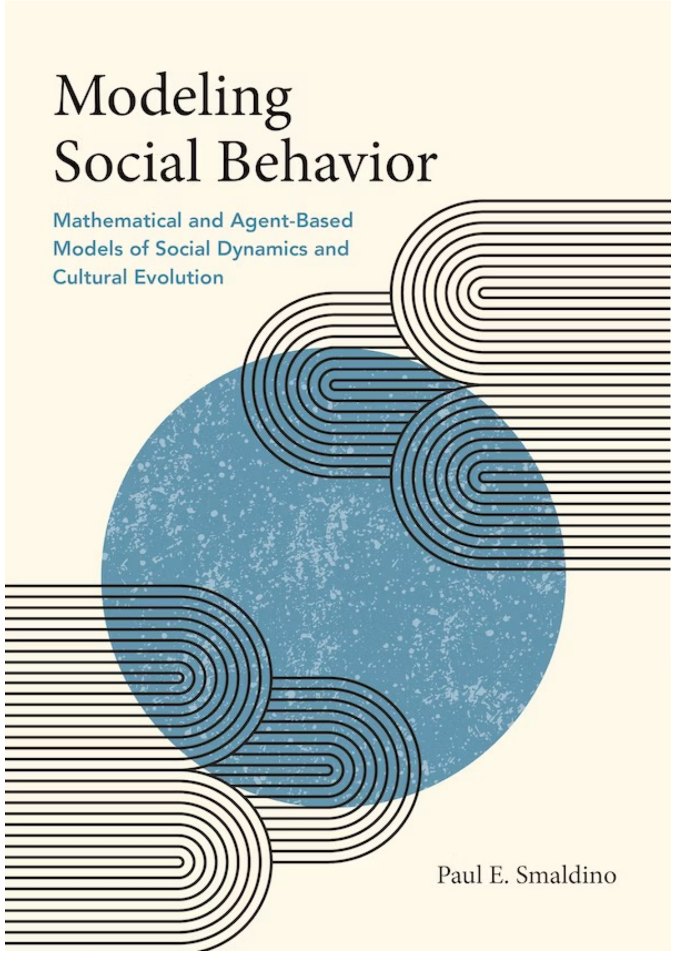
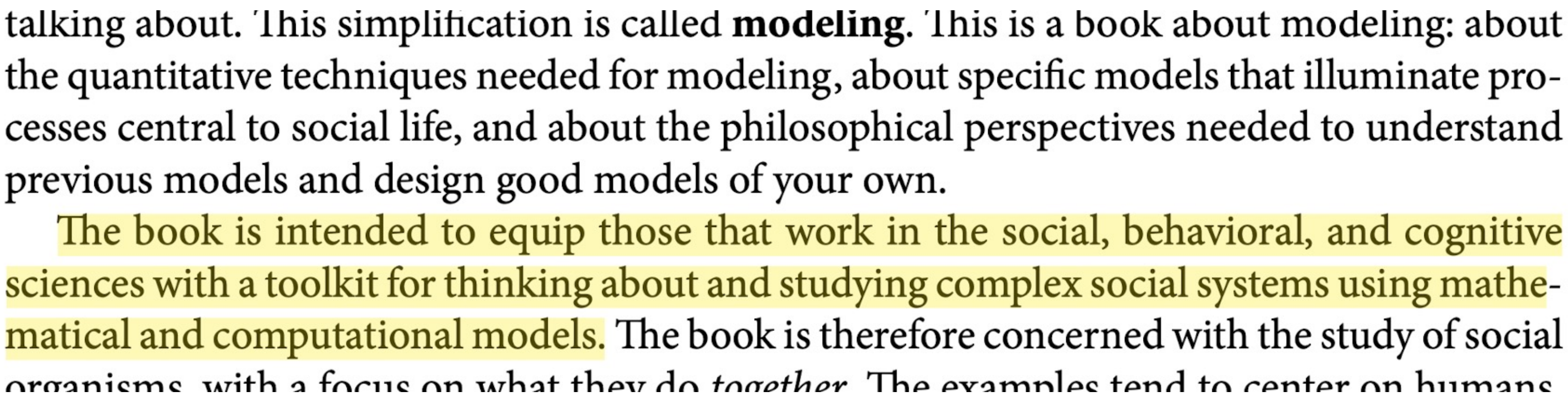


Figure 2. from Text as Data, Grimmer, Roberts, Stewart, 2022

# Simulations of Social Behavior


Modeling Social Behavior
Mathematical and Agent-Based Models of Social Dynamics and Cultural Evolution
Paul E. Smaldino

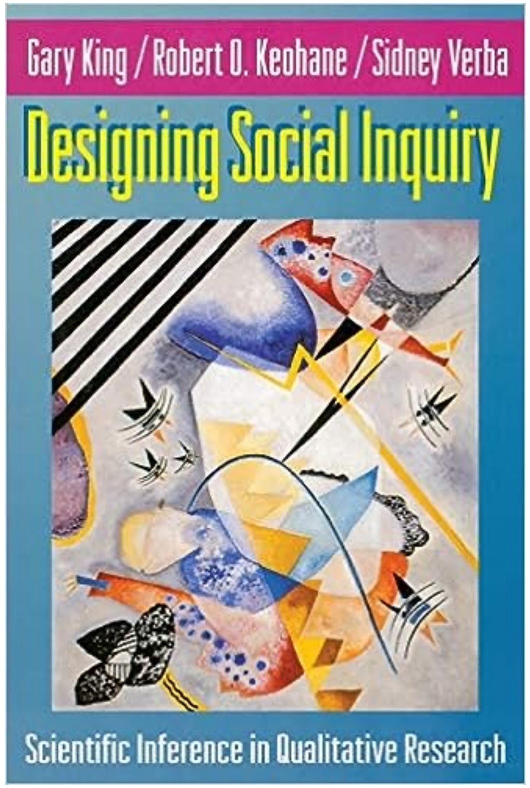
**“The sciences of social behavior are more important than ever.”**


talking about. This simplification is called **modeling**. This is a book about modeling: about the quantitative techniques needed for modeling, about specific models that illuminate processes central to social life, and about the philosophical perspectives needed to understand previous models and design good models of your own.

The book is intended to equip those that work in the social, behavioral, and cognitive sciences with a toolkit for thinking about and studying complex social systems using mathematical and computational models. The book is therefore concerned with the study of social organisms, with a focus on what they do *together*. The examples tend to center on humans

Smaldino, 2023

# “The Science in Social Science”


Gary King / Robert O. Keohane / Sidney Verba
Designing Social Inquiry
Scientific Inference in Qualitative Research

“ … social scientific research can be quantitative or qualitative in style. “ But ought to be scientific in approach:

- **The goal is inference:** Scientific research is designed to make **descriptive or explanatory inferences** on the basis of empirical information about the world.
- **The procedures are public:** Scientific research uses explicit, codified, and ***public* methods** to generate and analyze data whose reliability can therefore be assessed.
- **The conclusions are uncertain**: … Inference without **uncertainty estimates** is not science as we define it.
- **The content is the method:** (Karl Pearson, 1892) `**The unity of all science consists alone in its method**, not in its material´”

# Initial Areas of Research and Accompanied Data

- Households, populations – census, population microdata
- (Social) Media diets, political communication - (social) media data and volunteered data
- Jurisdiction environment – laws, strategic plans
- History and cultural heritage - historical text archives
- Sustainable reporting, management science – reporting businesses data
- Social ecology and urbanism – social media, public administration data, industry data
- Social Innovation in Public Policy, Social Services, Education - administrative & industry data

# The L.A. Project: Worldwide Evolution of Living Arrangements

Aims to:

- Understand the evolution of household compositions for 156  nations (and 3500 subnational regions) since 1970
- With 150 million individual records representing 98% of world´s population
- **Conduct the first world-wide multilevel analysis**: individual, household, sub-national, national
- World-wide dataset + multilevel regression = **computational intensive**

Collaboration with:

Albert Esteve, CED

Juan Galeano, CED

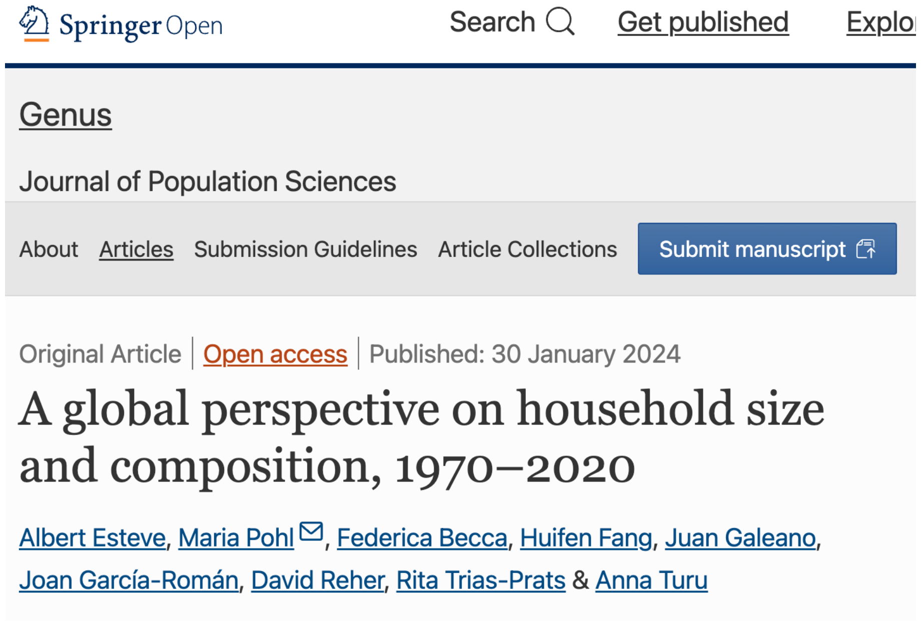
CED
Centre d'Estudis
Demogràfics

Fundación "la Caixa"

HARMONIZED INTERNATIONAL CENSUS DATA FOR SOCIAL SCIENCE AND HEALTH RESEARCH

# Prior study: CORESIDENCE

- **CORESIDENCE database:**
  - Household levels indicators at the national and sub-national level (use only national level in study)
  - 156 countries
  - 793 data points over time
- **Combines data from 4 main repositories:**
  - Integrated Public Use of Microdata Series-internationa (IPUMS-i) (Minnesota Population Center, 2020)
  - Demographic Health Surveys (DHS)
  - Multiple Indicator Cluster Surveys (MICS)
  - European Labor Force Surveys (EU-LFS)
- **Harmonization of indicators:**
  - Harmonize the construction of indicators (e.g., household size typology)
  - Definitions across samples cannot be harmonized



Journal of Population Sciences

Original Article | Open access | Published: 30 January 2024

A global perspective on household size and composition, 1970–2020

Albert Esteve, Maria Pohl, Federica Becca, Huifen Fang, Juan Galeano, Joan García-Román, David Reher, Rita Trias-Prats & Anna Turu

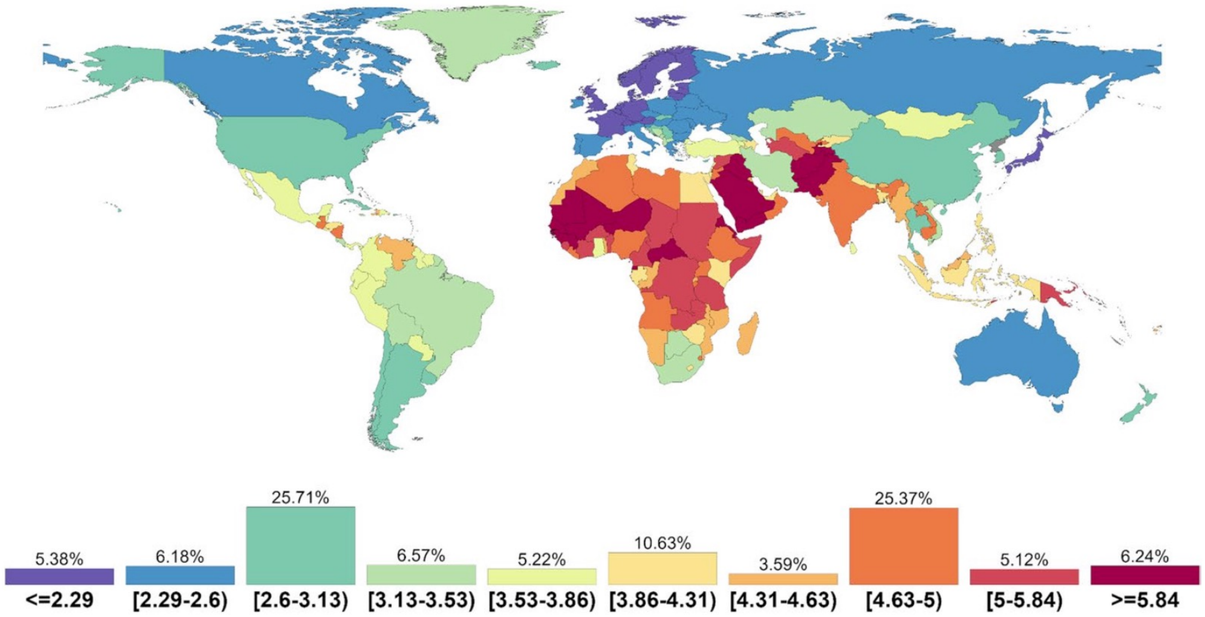


**Fig. 1** Average household size by country, most recent year available since 2000. Histogram legend shows the percentage of the world's population in each category. Each category represents 10 percent of the 156 countries represented in the map. Sources: CORESIDENCE database and UN Household database

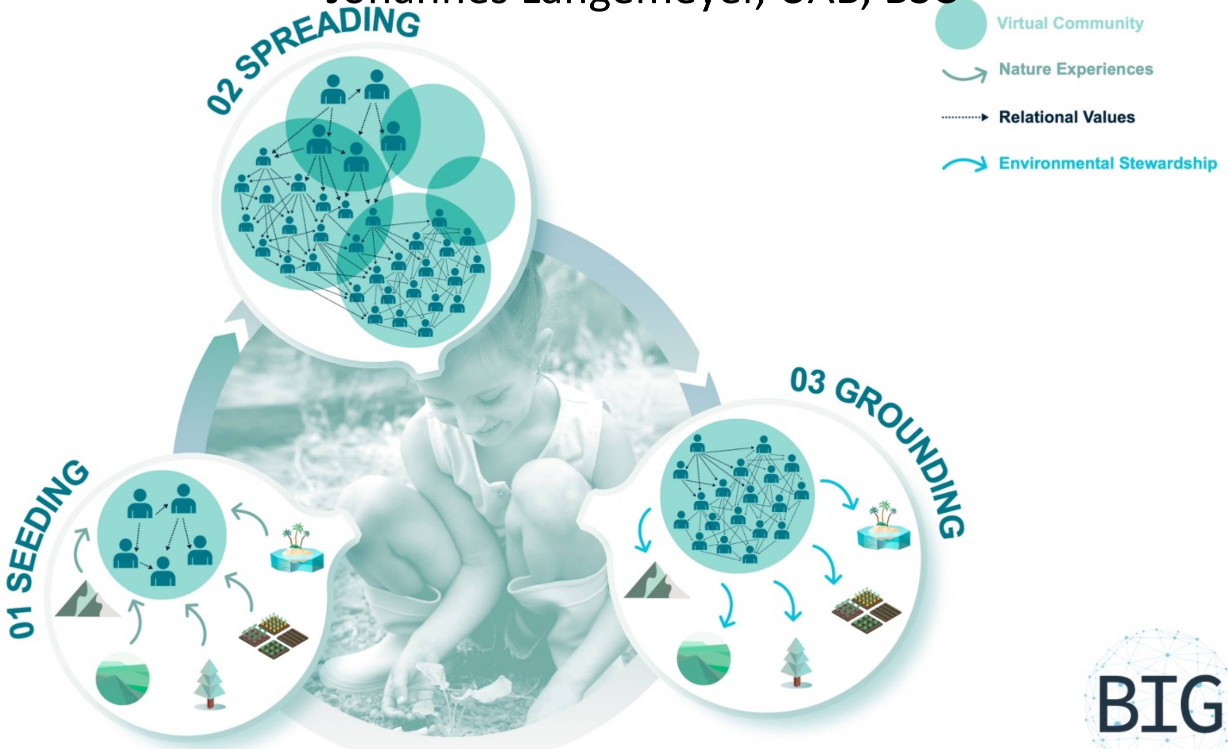
# BIG 5: Do Digital Experiences shape societal nature values and foster environmental stewardships?

- ERC Starting Grant 2023, **Johannes Langemeyer**, UAB (with Crosas, PI-advisor)
- Analyze data from 5 platforms: TikTok, YouTube, Instagram, Twitter, Facebook
- Development of methodology to identify DRVs: **Digital Relational Values**, triggered by indirect experience of nature
- Digital social network analysis/topic modeling/LLMs/experiments to assess and understand multiplying effect of social media to environmental stewardship through DRVs

Collaboration with:

ICTA
Institut de Ciència i Tecnologia Ambientals
ICTA-UAB

Johannes Langemeyer, UAB, BSC



BIG 5

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC Computational Social Sciences

# AI_Text: Catalan Historical Archives

- Catalunya holds one of the largest document archives in the world from XIII-XV
- *L´Arxiu General de Protocols de Barcelona* alone contains 111.115 volumes (17km)
- 750.000 images from 3.336 volumes digitalized
- AI-based transcription and analysis of texts
- Apply also to *l´Arxiu dels Marquesos de Barberà* and *Registres de Cancelleria de l'Arxiu de la Corona d'Aragó*

- Collaboration with:
  - Coral Cuadrada: AMSMB, URV. Doctora Història Medieval
  - Daniel Piñol: UB. Paleògraf. Doctor Història Medieval
  - Roser Salicrú: CSIC. Investigadora. Doctora Història Medieval
  - Antoni Albacete: AHPB. Director de Recerca i Difusió. Doctor en Història
  - Marc Estapé, coordinador, IBM
- From BSC, with:
  - Dario Garcia (AI), Nadia Tonello (Data Management)

# Proposed Workflow for Social Innovation of Complex Problems

In collaboration with:

AGIRRE LEHENDAKARIA CENTER
for Social and Political Studies

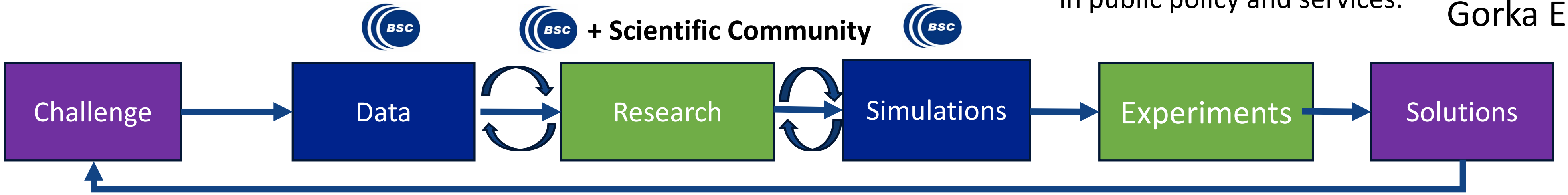
**1. Administration**
identifies a complex socio-techno problem, for which has no solution

**3. Research Calls**
Further identification of research questions, analysis, and learning of the given challenge by the scientific community.

**5. Co-creation**
Identifies network of innovation agents. Provides portfolio of possible **real experiments** for innovation in public policy and services.

Gorka Espiau

Itziar Moreno

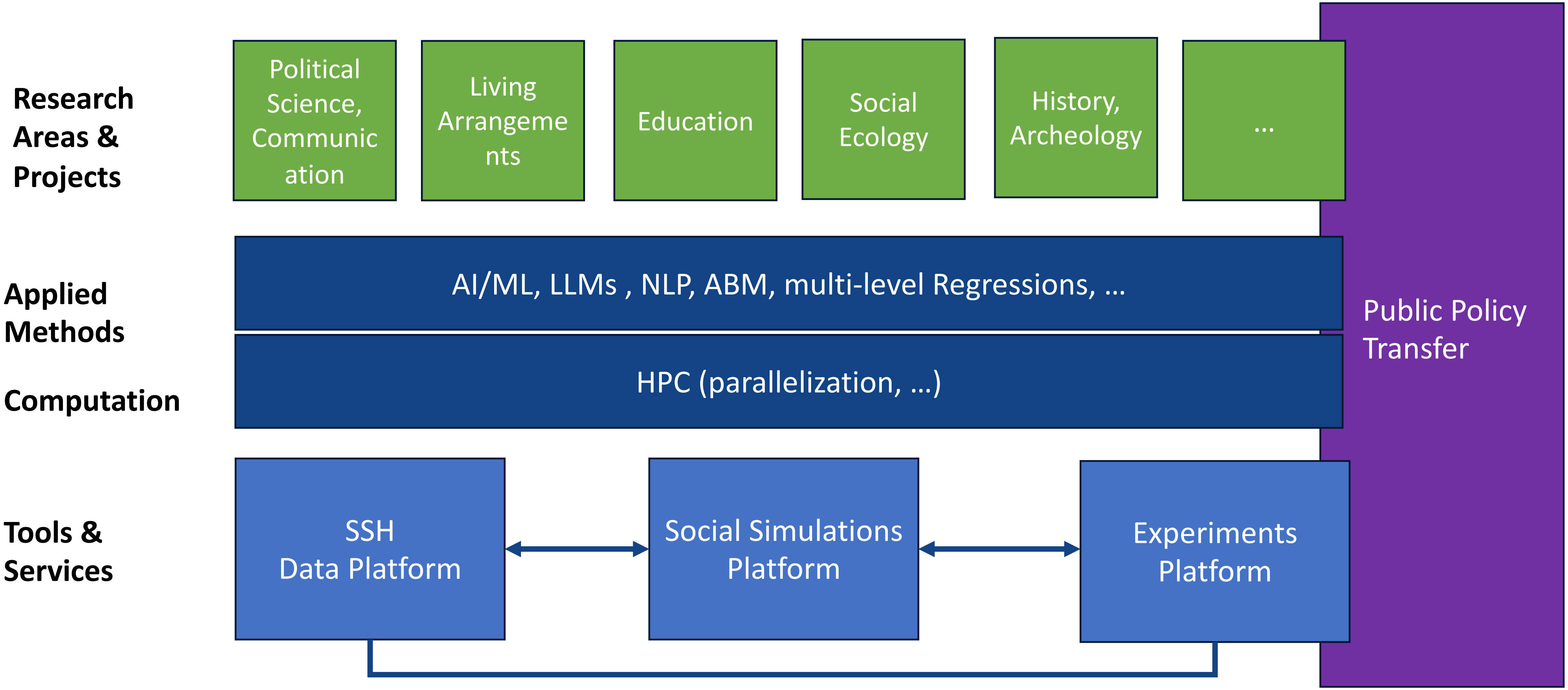


**2. Scientific Data Group**
Identifies existent data and data needed.  Helps define research questions that could be answered. Provides FAIR data catalog, management and harmonization, Follows protocols for data access to research.

**4. Virtual experiments of public policy**
Simulations of possible scenarios based on data. It models the implementation of solutions.

**6. Real Experiments**
Administration identifies the experiments that are applied the real world. Evaluation of the solution follows, iteratively.

# From Research and Computing to Public Policy



Research Areas & Projects

Political Science, Communication

Living Arrangements

Education

Social Ecology

History, Archeology

...

Applied Methods

AI/ML, LLMs , NLP, ABM, multi-level Regressions, …

Computation

HPC (parallelization, …)

Tools & Services

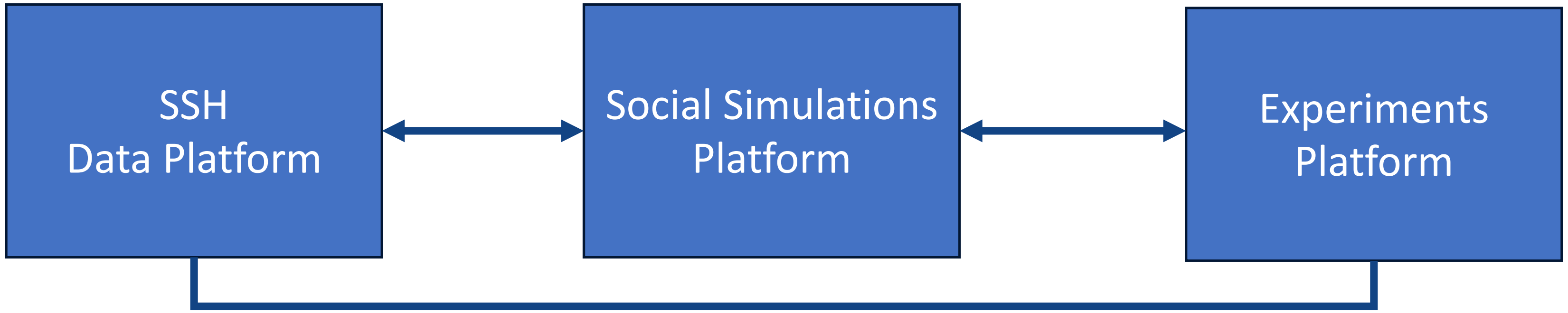
SSH Data Platform

Social Simulations Platform

Experiments Platform

Public Policy Transfer

# Program´s Tools & Services will support Research and Public Policy Transfer



**Data Science:**

- Data collection & acquisition
- FAIR Data and Metadata for Harmonization
- Data integration with HPC workflows
- Data readiness for applied methods
- Sensitive data

**Complexity Science:**

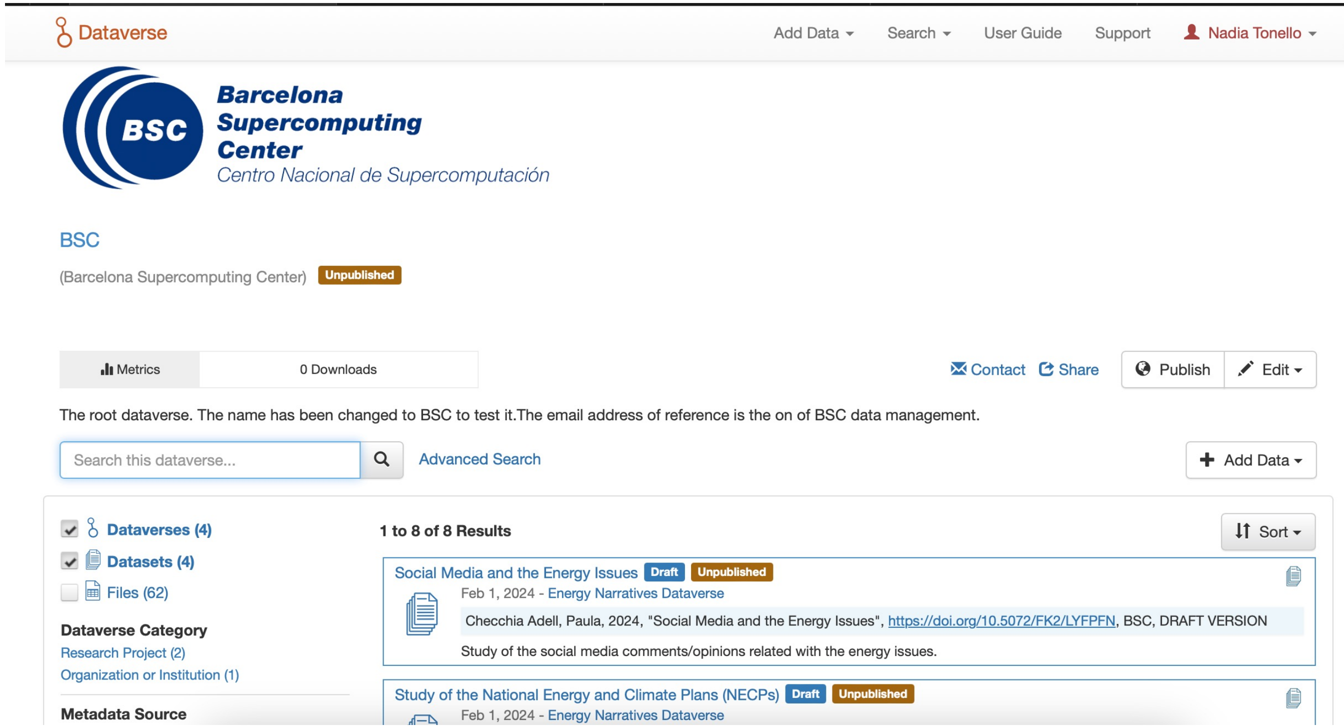
- Combine Agent-Based Models with fine-tuned LLMs
- Network analysis
- Explore scenarios and virtual experiments
- Reduce number of real-world experiments

**Real-world Experiments**

- Design/manage Randomized Control Trials
- Connect innovation experiments network
- Causal inference analysis
- Feed data and results back to simulations

# SSH + HPC Data Platform: BSC Dataverse

- Publish datasets used for **Computational Social Science research projects** for transparency, reproducibility and reuse.
- House the metadata for the **Public-Private Data Partnership program for SSH** research, and data user agreements to request access to the data
- Integrate with High-Performance Computing (HPC) with MareMostrum (workflows)
- **DDI support (future DDI-CDI)**
- Harvest metadata from other SSH data repositories
- Not only a data repository, also **publish code** for SSH **computational methods** using **HPC**



BSC

(Barcelona Supercomputing Center) Unpublished

The root dataverse. The name has been changed to BSC to test it.The email address of reference is the on of BSC data management.

☑ Dataverses (4)
☑ Datasets (4)
□ Files (62)

**Dataverse Category**
Research Project (2)
Organization or Institution (1)

**Metadata Source**

**1 to 8 of 8 Results**

Social Media and the Energy Issues Draft Unpublished
Feb 1, 2024 - Energy Narratives Dataverse
Checchia Adell, Paula, 2024, "Social Media and the Energy Issues", https://doi.org/10.5072/FK2/LYFPFN, BSC, DRAFT VERSION
Study of the social media comments/opinions related with the energy issues.

Study of the National Energy and Climate Plans (NECPs) Draft Unpublished
Feb 1, 2024 - Energy Narratives Dataverse

Additional benefits:

- Enables **responsible open science** by making data as open as possible and as restricted as needed
- Aligns with FAIR principles to support **Findable, Accessible, Interoperable, and Reusable** data/metadata
- Builds on top of the Dataverse **open-source platform** (dataverse.org)

# BSC-Harvard Collaboration

The Dataverse® Project

OpenDP

IQSS

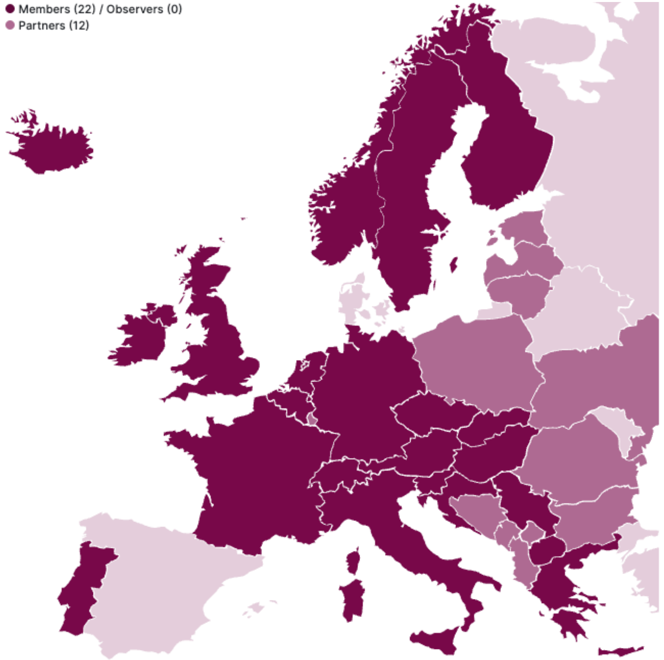
The Institute for Quantitative Social Science

HARVARD UNIVERSITY

- BSC Dataverse will be part of the collaboration with Harvard´s IQSS
- Collaboration will help to extend the functionality to integrate with large storage and HPC
- Future integration with OpenDP (differential privacy tools) to support sensitive data
- Fine-tuning LLMs with social science data and metadata

# Collaboration with European Infrastructures

- Coordination of  CLARIAH-CAT, and CLARIAH-ES
- Coordination with Social Science and Humanities Open Cloud
- Aim to work with:
  - CESSDA – steps towards Spain being part of CESSDA ERIC
  - MEDem – liaising ESFRI Roadmap



CESSDA membership: Consortium of European Social Science Data Archives

# International Data Collaborations – with CODATA

COMMITTEE ON DATA

CODATA

INTERNATIONAL SCIENCE COUNCIL

- To empower science to address universal challenges through the transparent, trustworthy and equitable use of data and information.
- To work with and provide recommendations to UN, UNESCO, OECD on the use and sharing of data for science

**Four priority goals:**

- **Making Data Work**
- **Improving Data Policies**
- **Advancing Data Science**
- **Enhancing Data Skills**

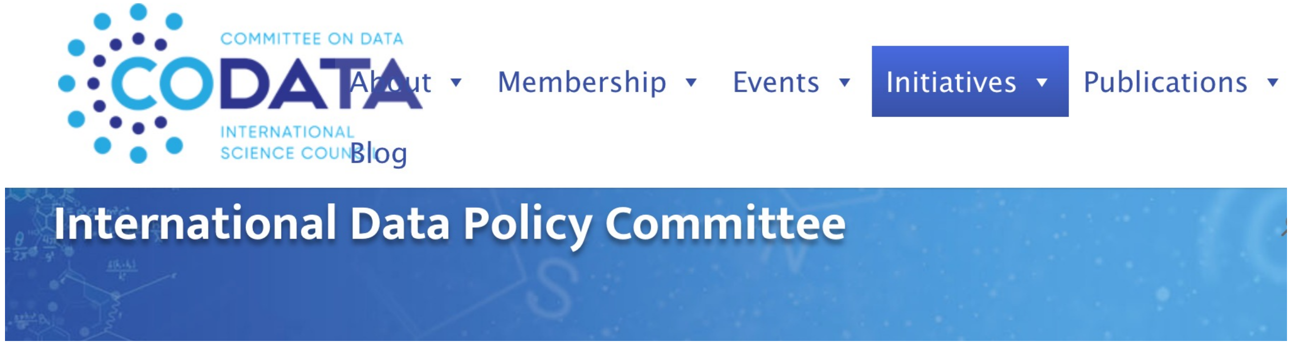
Decadal Programme: Making Data Work for Cross Domain Grand Challenges

# CODATA key initiatives

WorldFAIR

Cross-Domain Interoperability Framework (CDIF)

- CDIF has been developed based on input from the 11 case studies within WorldFAIR
- Use existing standards
- Recommendations for specific standards (Schema.org, DCAT, ODRL, DDI-CDI, SKOS/XKOS, SSSOM, etc.)
- Draft in May 2024



International Data Policy Committee

**IDPC Mission:**

- The mission of the CODATA IDPC is to serve as a global platform for **collaboration, research, and advocacy in shaping effective data policies** that address the complex challenges of today’s increasingly digital societies
- UNESCO Toolkit by the end of 2024

# Conclusions

- **Social Sciences** are **more needed than ever** to address today´s challenges.
- With larger amounts of data, better software for applied methods, and more computational power, **computational social sciences** emerged and is **growing.**
- Computational social science benefits from **combining official statistics data with big data** (mostly from industry) and data generated by **research.**
- The **combined data can be used for social simulations** and help reduce the number of experiments conducted to test **social innovation and public policy**
- For this, we need to **harmonize everchanging variables/indicators/vocabularies** from multiple sources, sectors, and types: structured, semi-structured, and unstructured data.
- **FAIR metadata becomes the critical step to Interoperable data** – *work presented at COSMOS (DDI-CDI, STDF, i-ADOPT, harmonize SDMX – DDI-CDI) can help*

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

# THANK YOU!

merce.crosas@bsc.es

www.bsc.es