BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Ciències Socials amb dades i computació

Mercè Crosas, Ph.D.

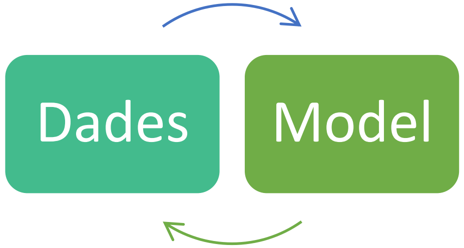
Nou Programa @BSC:
Ciències Socials Computacionals

Domini Científic

- Economia
- Ciències Polítiques
- Psicologia, Ciències Cognitives
- Sociologia, demografia, antropologia
- Humanitats Digitals: Història, arqueologia, lingüística, literatura, patrimoni cultural

Dades

- Enquestes
- Dades estadístiques i administratives
- Dades d´empreses
- Web scraping
- Xarxes socials
- Experimentals
- Satèl·lit, Sensors


Dades

Model

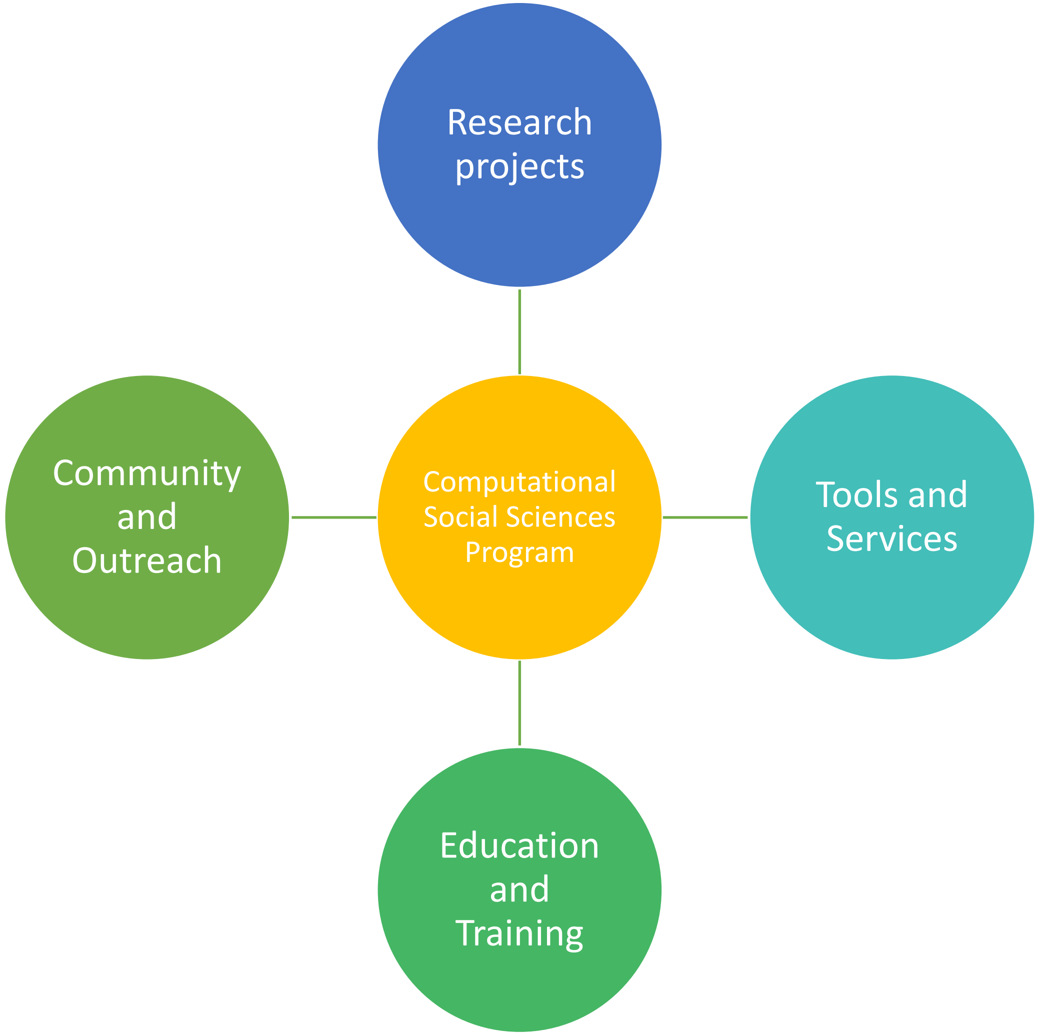
Models

- Simulacions (basades en agents o equacions)
- Models estadístics, regressions
- Aprenentatge automàtic (ML)
- Anàlisi de text, NLP, LLM

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

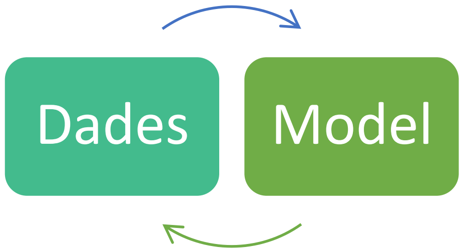
Visió i Estratègia

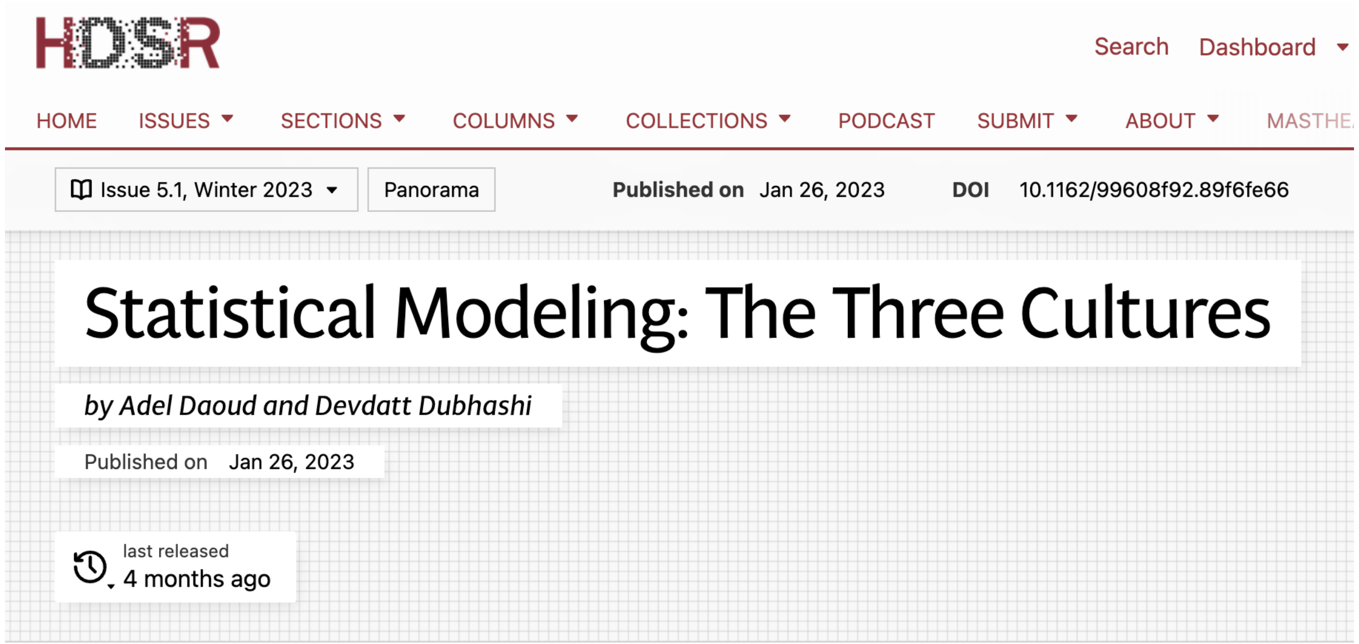
- Preparar les ciències socials i les humanitats per a que es puguin beneficiar de l´era de les dades i la IA.
- Ampliar la col·laboració entre científics socials, humanistes i informàtics.
- Facilitar l´ús del supercomputador a les ciències socials i a les humanitats, posant el BSC a l´abast de *tots* els investigadors.
- Aplicar una recerca eficient i escalable en ciències socials per assistir a les polítiques públiques



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Models i Dades en Ciències Socials: Predicció i Explicació





Issue 5.1, Winter 2023 | Panorama | **Published on** Jan 26, 2023 | **DOI** 10.1162/99608f92.89f6fe66

Statistical Modeling: The Three Cultures

by Adel Daoud and Devdatt Dubhashi

Published on Jan 26, 2023

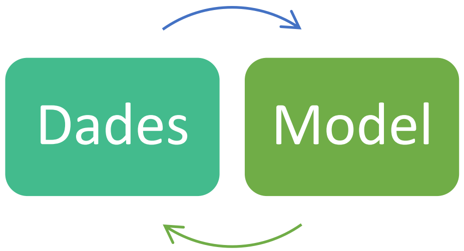
last released
4 months ago

- Les ciències socials quantitatives tradicionalment es basen en:
  - Anàlisis predictiu
  - Anàlisis causal (prova controlada aleatòria, experiments)
- *Data Modeling Culture* (regressions lineals) – prediccions i causalitat
- *Algorithm Modeling Culture (machine learning)* – prediccions
- *Hybrid Modeling Culture:* prediccions i causalitat amb regressions lineals i ML

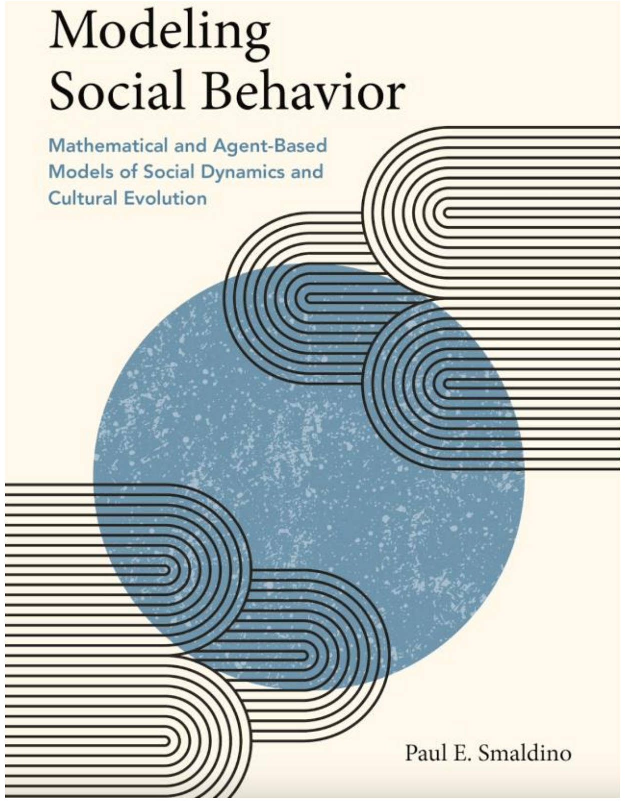
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

https://hdsr.mitpress.mit.edu/pub/uo4hjcx6/release/1

Models i Dades en Ciències Socials: Simulacions


Dades

Model



- Un **Model** en aquest context és “una estructura abstracte o física que potencialment representa un fenomen real” (Weisberg, 2023)
- *Agent-based Modeling:* A on individus estan representats com entitats computacionals (agents) amb un comportament i interacció local.

https://press.princeton.edu/books/paperback/9780691224145/modeling-social-behavior

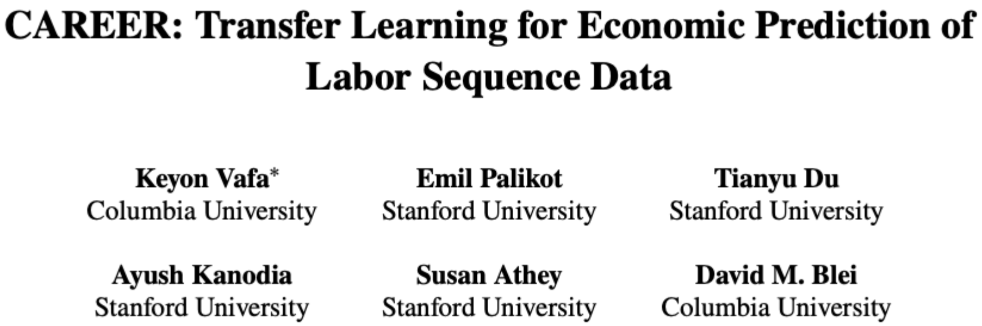
Exemples:

Recerca en ciències socials i humanitats amb l´ús de dades i computació

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Using resumes for economic analysis of labor market

- 24 Million resumes (Zippia)
- Create a transformer-based model that uses transfer learning to learn representations of job sequences: CAREER
- Fine tune model with traditional longitudinal survey data predictive models.



CAREER: Transfer Learning for Economic Prediction of Labor Sequence Data

Keyon Vafa* Columbia University

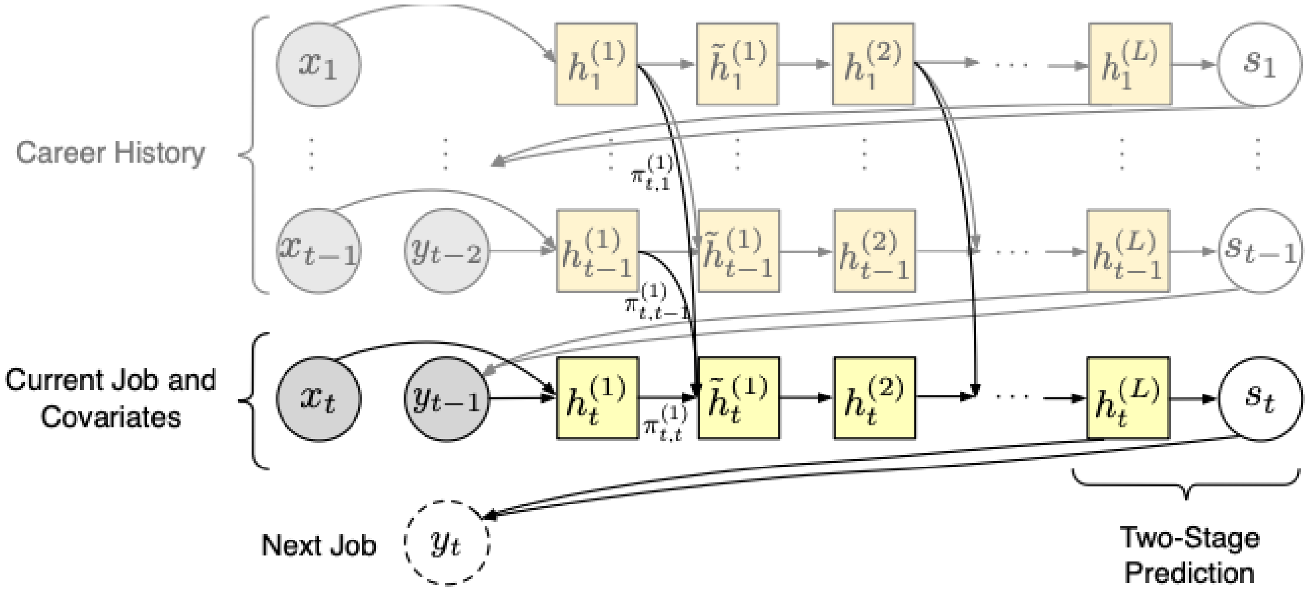
Emil Palikot Stanford University

Tianyu Du Stanford University

Ayush Kanodia Stanford University

Susan Athey Stanford University

David M. Blei Columbia University

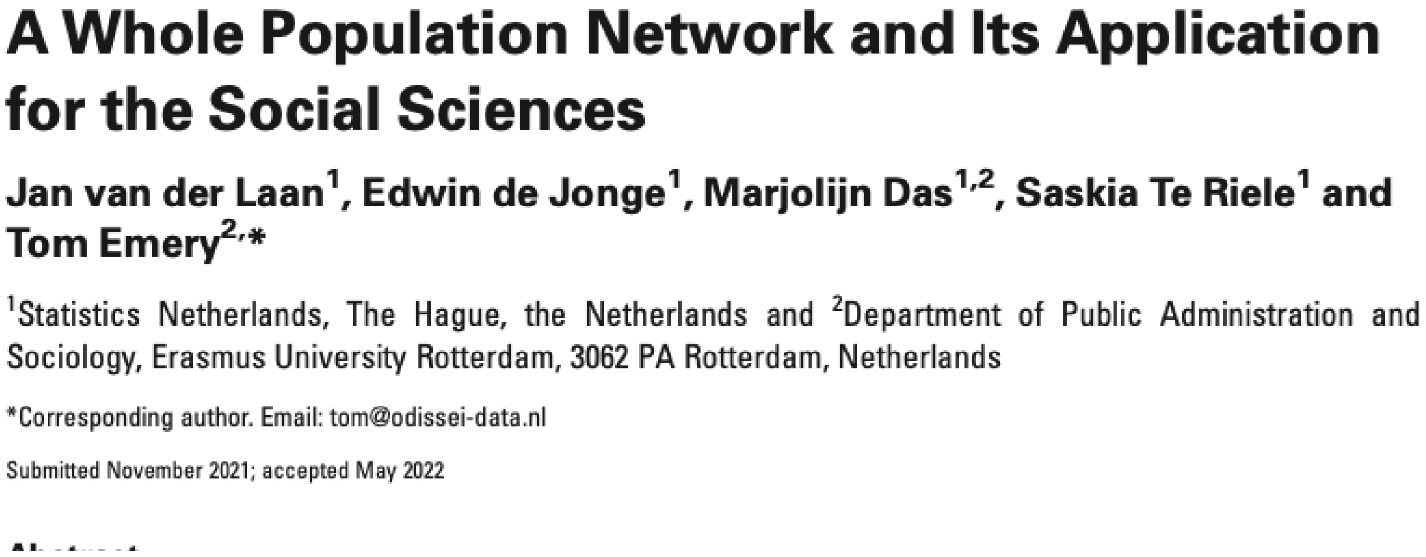


CAREER parameterizes a low-dimensional representation of an individual's career history with a transformer, which it uses to predict the next job.

https://arxiv.org/abs/2202.08370

Population Network from Administrative data in the Netherlands

- 1.4 billion relationships between 17 millions inhabitants of the Netherlands
- Network analysis
- Dataset available for analysis at Statistics Netherlands for research purpose



A Whole Population Network and Its Application for the Social Sciences

Jan van der Laan[1], Edwin de Jonge[1], Marjolijn Das[1,2], Saskia Te Riele[1] and Tom Emery[2,*]

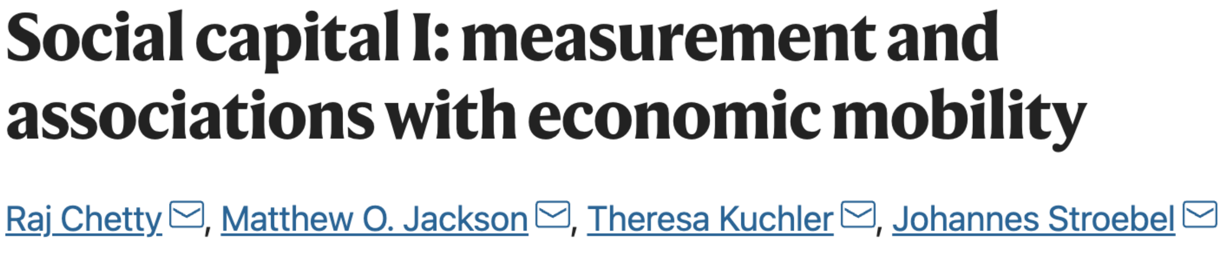
[1]Statistics Netherlands, The Hague, the Netherlands and [2]Department of Public Administration and Sociology, Erasmus University Rotterdam, 3062 PA Rotterdam, Netherlands

*Corresponding author. Email: tom@odissei-data.nl

Submitted November 2021; accepted May 2022

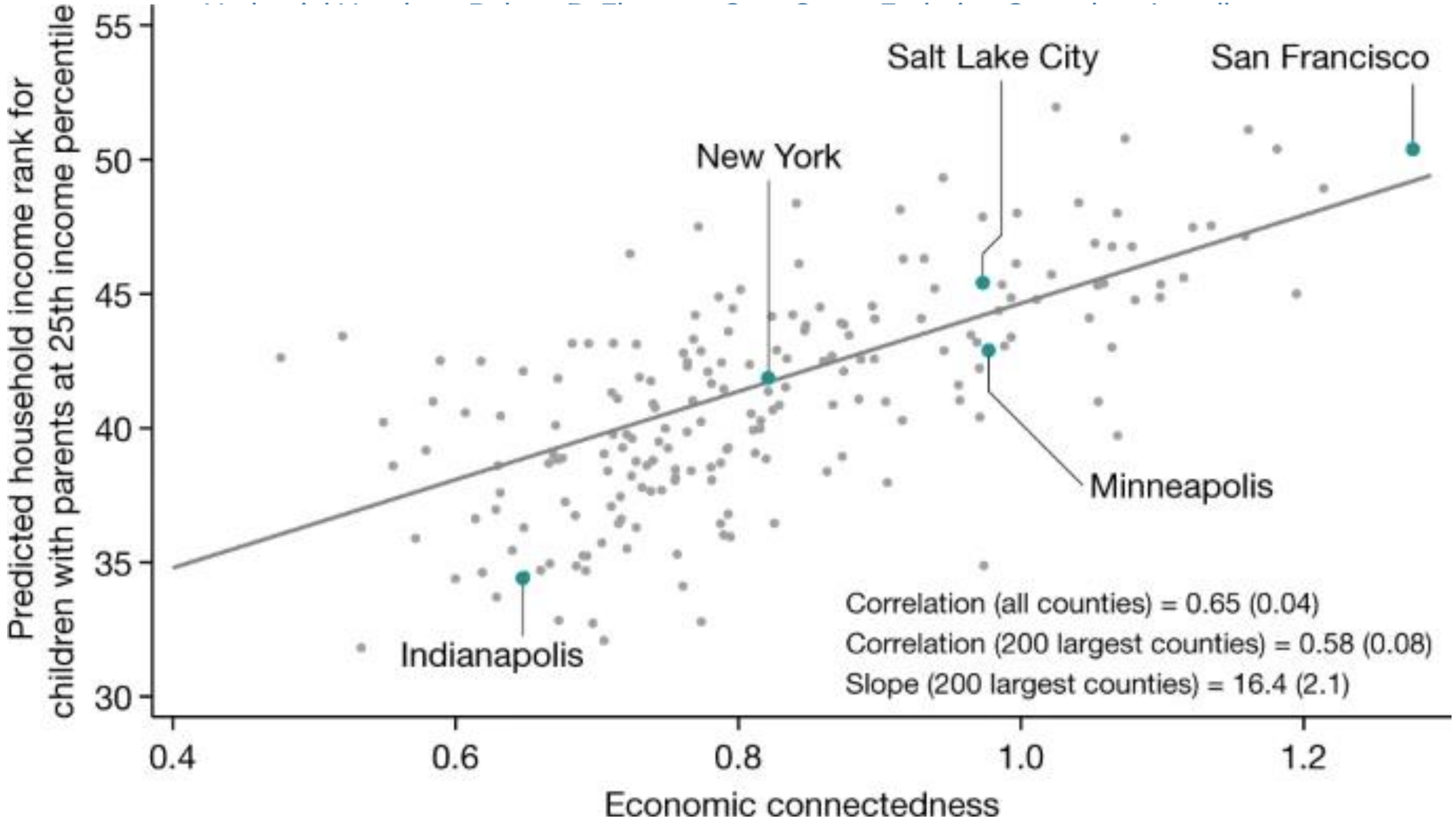
https://academic.oup.com/esr/article/39/1/145/6605763

- 21 billion relationships from Facebook
- Measure Social Capital:
  - (1) cross-type connectedness,
  - (2) network cohesiveness
  - (3) civic engagement
- Multivariable regressions
- Economic connectedness (high economic status friends for individual with low economic status) is the higher predictor of upward income mobility



Social capital I: measurement and associations with economic mobility

Raj Chetty, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel



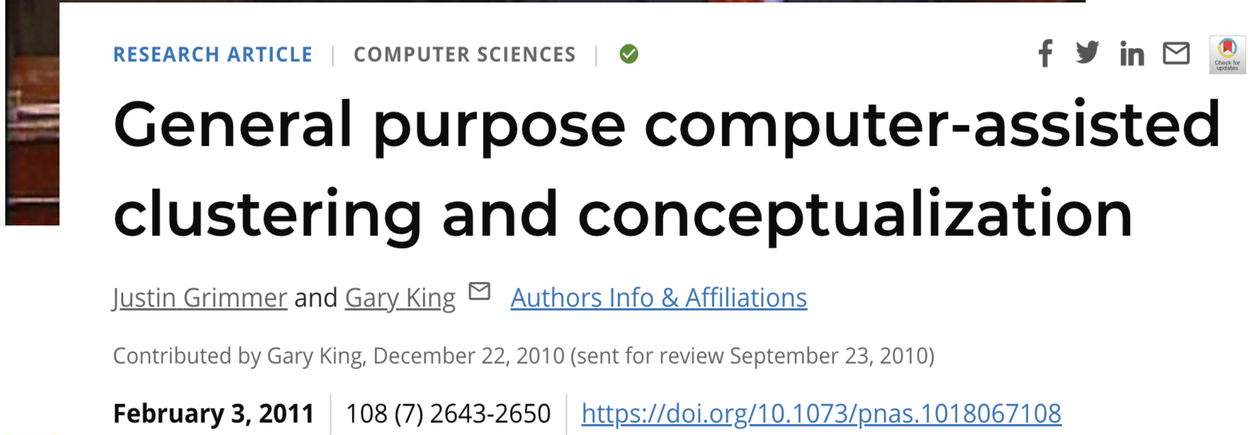
https://www.nature.com/articles/s41586-022-04996-4
https://www.nature.com/articles/s41586-022-04997-3

06:43

Study Finds Congress Spends 27% Of Its Time Taunting

April 21, 2011

This article is more than 12 years old.



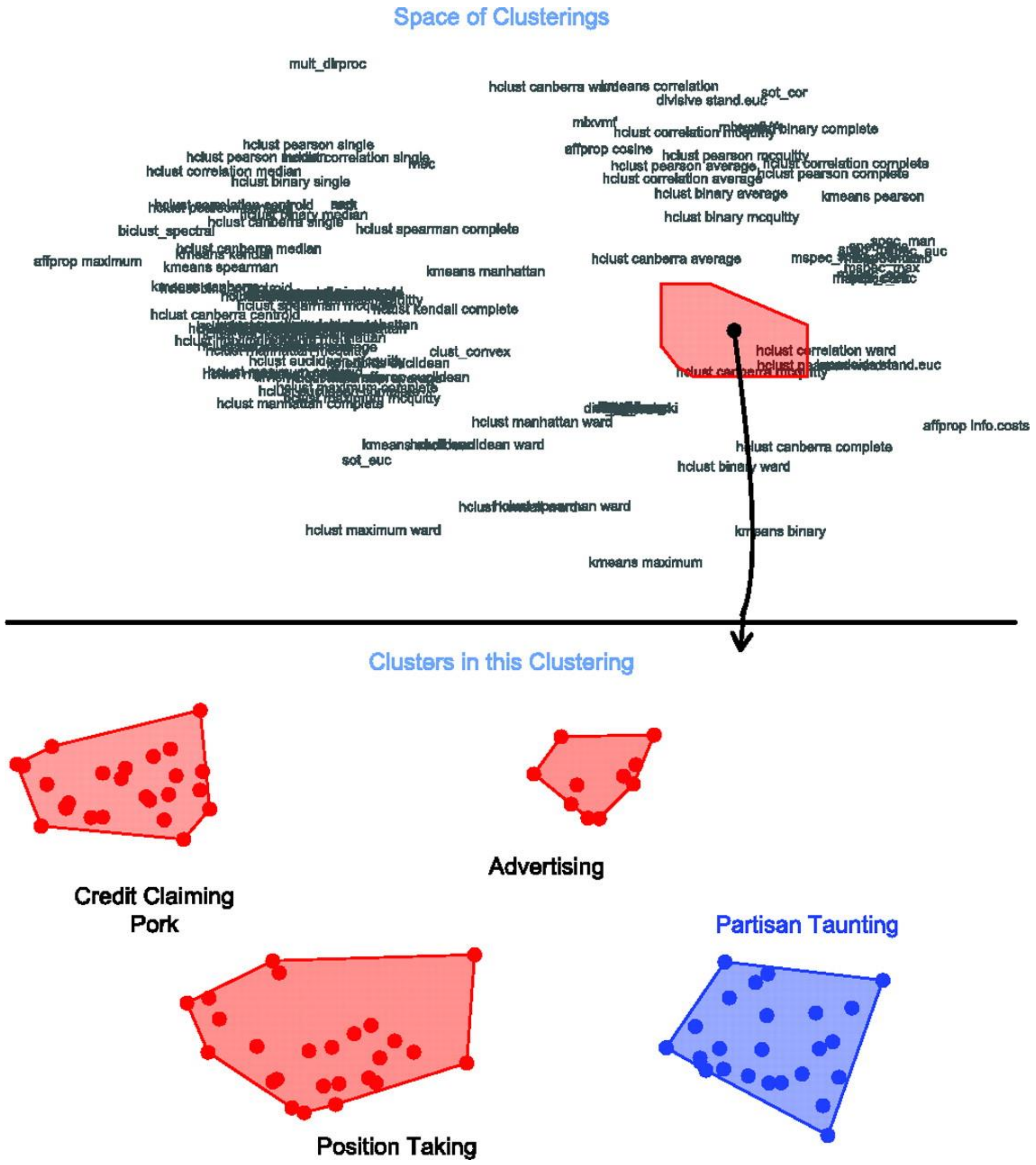
RESEARCH ARTICLE | COMPUTER SCIENCES |

General purpose computer-assisted clustering and conceptualization

Justin Grimmer and Gary King Authors Info & Affiliations

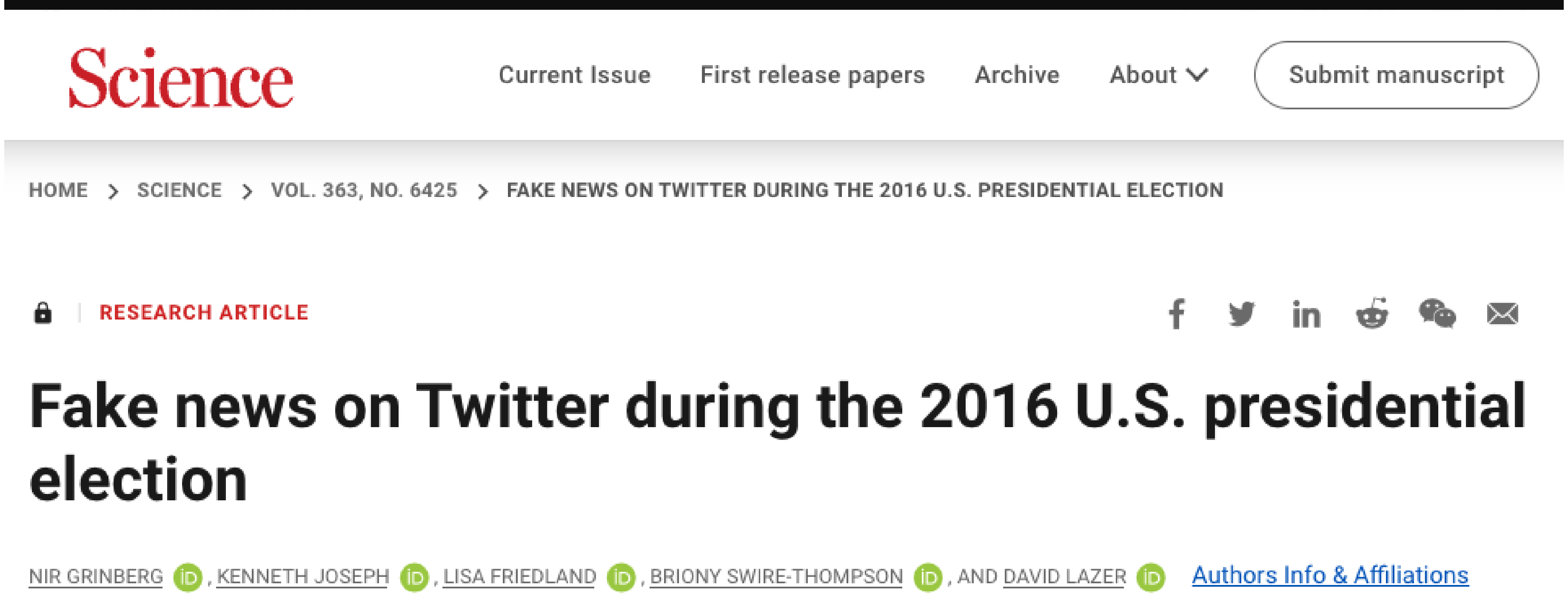
Contributed by Gary King, December 22, 2010 (sent for review September 23, 2010)

February 3, 2011 | 108 (7) 2643-2650 | https://doi.org/10.1073/pnas.1018067108



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

https://www.pnas.org/doi/10.1073/pnas.1018067108



Science

Current Issue First release papers Archive About Submit manuscript

HOME > SCIENCE > VOL. 363, NO. 6425 > FAKE NEWS ON TWITTER DURING THE 2016 U.S. PRESIDENTIAL ELECTION

RESEARCH ARTICLE

Fake news on Twitter during the 2016 U.S. presidential election

NIR GRINBERG, KENNETH JOSEPH, LISA FRIEDLAND, BRIONY SWIRE-THOMPSON, AND DAVID LAZER Authors Info & Affiliations

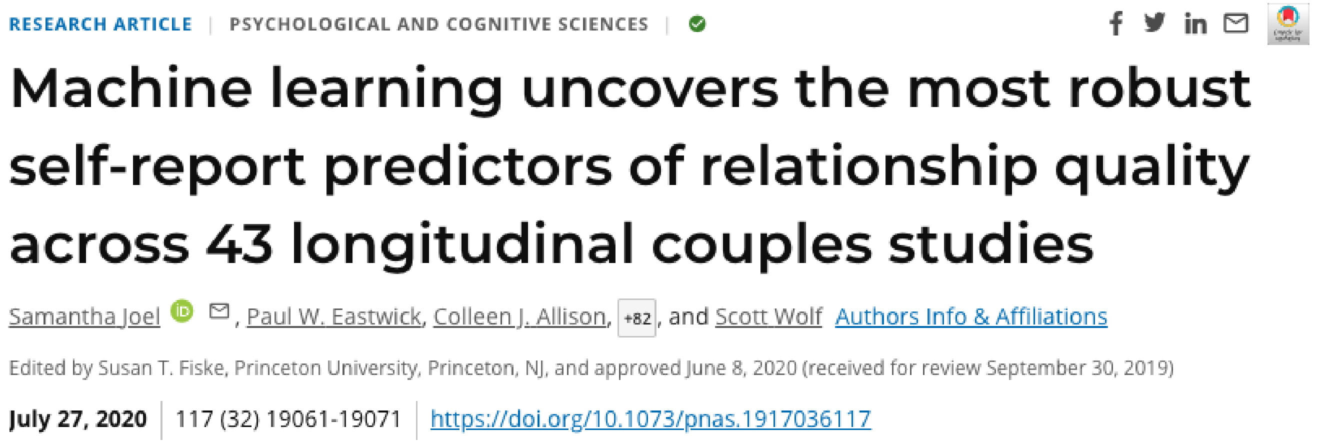
“Fake news accounted for nearly 6% of all news consumption, but it was heavily concentrated—only 1% of users were exposed to 80% of fake news, and 0.1% of users were responsible for sharing 80% of fake news.”

- Twitter data linked to public voter registration records, studying the tweets sent by more than 16,000 accounts from August to December 2016.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Meta-analysis of relationship quality

- 43 longitudinal datasets from 29 labs
- Use Random Forests to quantify predictability of relationship quality
- Quality is predicted from a variety of constructs, but higher predictor is a person´s perception of the relationship itself



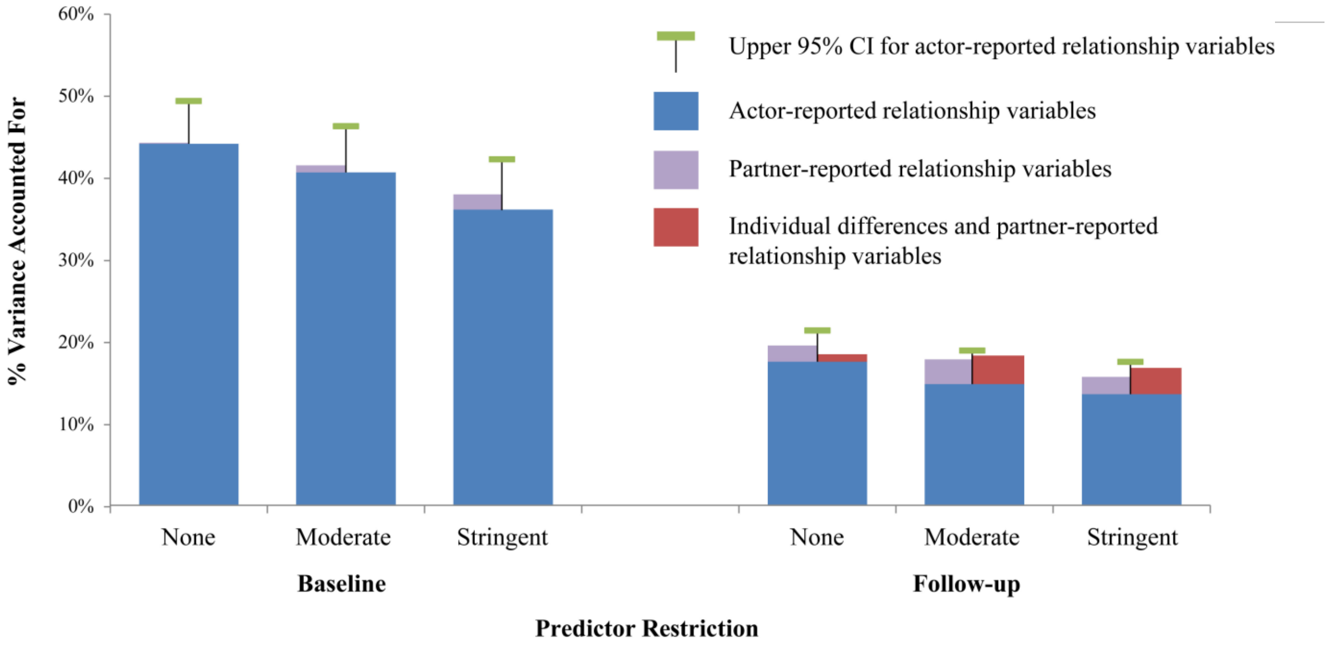
RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES |

Machine learning uncovers the most robust self-report predictors of relationship quality across 43 longitudinal couples studies

Samantha Joel, Paul W. Eastwick, Colleen J. Allison, +82, and Scott Wolf Authors Info & Affiliations

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved June 8, 2020 (received for review September 30, 2019)

July 27, 2020 | 117 (32) 19061-19071 | https://doi.org/10.1073/pnas.1917036117

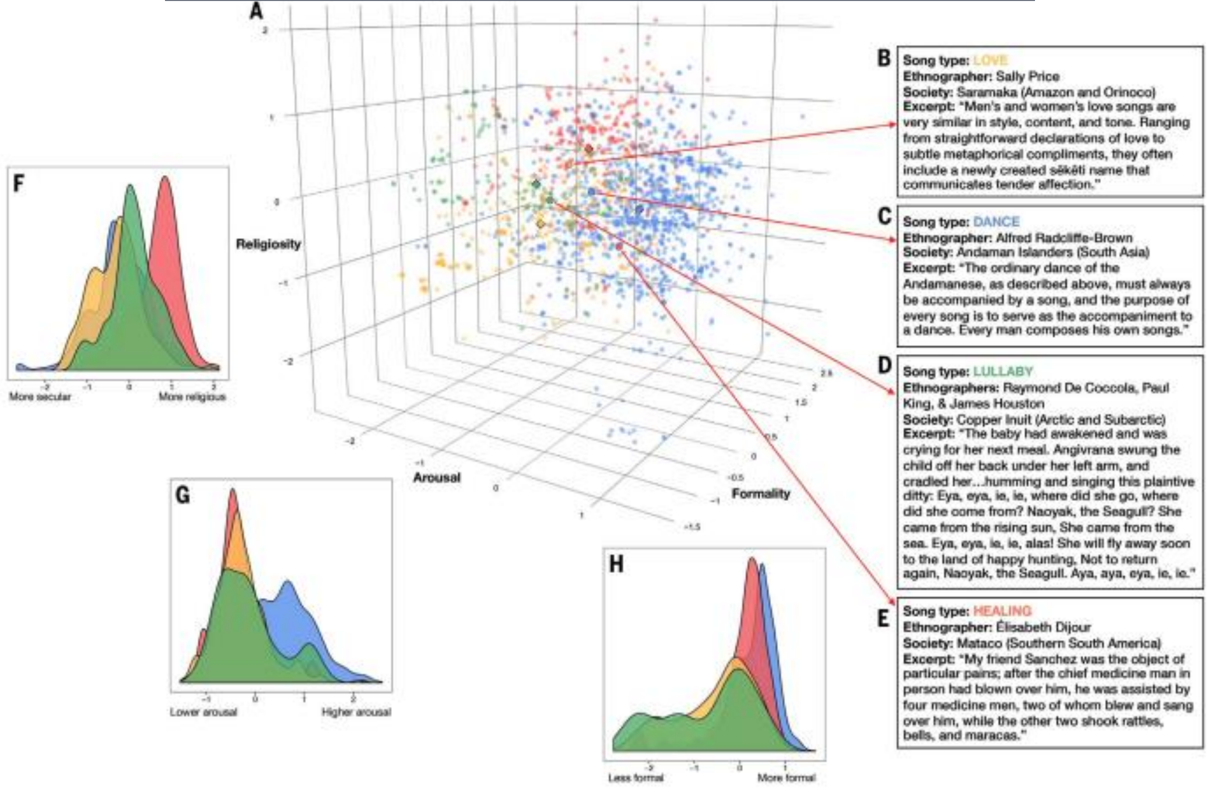


https://www.pnas.org/doi/full/10.1073/pnas.1917036117

Universality and Diversity in Human Song

- Data:
  - A corpus of ethnographic text on musical behavior
  - A discography of audio recordings of the music itself
- Datasets annotated by humans and automated algorithms (matching algorithm, Markov chain Monte Carlo, Bayesian principal component analysis)
- Computational social science applied to rich humanistic data reveals universal features and patterns of variability.

THE MUSIC LAB



B

Song type: LOVE
Ethnographer: Sally Price
Society: Saramaka (Amazon and Orinoco)
Excerpt: "Men's and women's love songs are very similar in style, content, and tone. Ranging from straightforward declarations of love to subtle metaphorical compliments, they often include a newly created sëkëti name that communicates tender affection."

C

Song type: DANCE
Ethnographer: Alfred Radcliffe-Brown
Society: Andaman Islanders (South Asia)
Excerpt: "The ordinary dance of the Andamanese, as described above, must always be accompanied by a song, and the purpose of every song is to serve as the accompaniment to a dance. Every man composes his own songs."

D

Song type: LULLABY
Ethnographers: Raymond De Coccola, Paul King, & James Houston
Society: Copper Inuit (Arctic and Subarctic)
Excerpt: "The baby had awakened and was crying for her next meal. Angivrana swung the child off her back under her left arm, and cradled her...humming and singing this plaintive ditty: Eya, eya, ie, ie, where did she go, where did she come from? Naoyak, the Seagull? She came from the rising sun, She came from the sea. Eya, eya, ie, ie, alas! She will fly away soon to the land of happy hunting, Not to return again, Naoyak, the Seagull. Aya, aya, eya, ie, ie."

E

Song type: HEALING
Ethnographer: Élisabeth Dijour
Society: Mataco (Southern South America)
Excerpt: "My friend Sanchez was the object of particular pains; after the chief medicine man in person had blown over him, he was assisted by four medicine men, two of whom blew and sang over him, while the other two shook rattles, bells, and maracas."

Fig. 2. Patterns of variation in the NHS Ethnography. (**A** to **E**) Projection of a subset of the NHS Ethnography onto three principal components. Each point represents the posterior mean location of an excerpt, with points colored by which of four types (identified by a broad search for matching keywords and annotations) it falls into: dance (blue), lullaby (green), healing (red), or love (yellow). The geometric centroids of each song type are represented by the diamonds. Excerpts that do not match any single search are not plotted but can be viewed in the interactive version of this figure at http://themusiclab.org/nhsplots, along with all text and metadata. Selected examples of each song type are presented here [highlighted circles and (B) to (E)]. (**F** to **H**) Density plots show the differences between song types on each dimension. Criteria for classifying song types from the raw text and annotations are shown in table S17.

Mehr et al.
https://mehr.nz/pdf/2019_MehrEtAl_Science.pdf

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Gràcies

merce.crosas@bsc.es