Elements of quantitative research

Francisco Villamil

Research Design for Social Sciences
MA Computational Social Science, UC3M
Fall 2023

Roadmap

Theories and research questions

Concepts and operationalization

Measurement

Description

Prediction

Example: Wartime civilian deaths

Paper discussion

Research and RQs

- Research = answering questions
- Re-cap on research types
  1. **Normative** vs. **positive** research
  2. Positive: **theoretical** vs. **empirical**
  3. Empirical: **Descriptive**, **explanatory**

Research and RQs

- Research = answering questions
- Re-cap on research types
  1. **Normative** vs. **positive** research
  2. Positive: **theoretical** vs. **empirical**
  3. Empirical: **Descriptive**, **explanatory**

     And **predictive**?

Example on types of questions and evidence

- Imagine you are doing the analytics previous to creating a campaign for Mahou
- Think of questions that are:
  - → Quantitative and descriptive
  - → Quantitative and explanatory
  - → Qualitative and descriptive
  - → Qualitative and explanatory

What is a research question?

- Any question we can answer
- Sometimes we say that we derive an RQ from a topic, and a theory from an RQ

  $\rightarrow$ **Topic** $>$ **RQ** $>$ **Theory**

What is a research question?

- Any question we can answer
- Sometimes we say that we derive an RQ from a topic, and a theory from an RQ
  - → **Topic** > **RQ** > **Theory**
- In reality, an RQ can be thought of as an operationalization of an argument
  - → **Previous evidence** > **Argument** > **RQ** > **Hypotheses**
- **Even though** this 'argument' can be something anecdotical that we later develop into a proper, abstract theoretical argument

What is a research question?

- Any question we can answer
- Sometimes we say that we derive an RQ from a topic, and a theory from an RQ
  - → **Topic** > **RQ** > **Theory**
- In reality, an RQ can be thought of as an operationalization of an argument
  - → **Previous evidence** > **Argument** > **RQ** > **Hypotheses**
- **Even though** this 'argument' can be something anecdotical that we later develop into a proper, abstract theoretical argument
  - → And it would actually look into something like this:
  - → Previous > 'Anecdotal argument' > RQ > (Proper) Theory > Hs

Good RQs, in brief

1. Empirically **answerable**
   → i.e. you can answer it with data
2. Theoretically **relevant**
   → i.e. it helps you learn something about your theory/argument

Good RQs, more in detail

1. Consider potential results of the analyses
   - → if you found X, does that answer the question? causally?
   - → example: are kids who play videogames often more aggresive?
   - → does that inform a theory on the aggresiveness effect of VGs?
   - → (you can even try to do a better RQ without causal ID)

Good RQs, more in detail

1. Consider potential results of the analyses
   - → if you found X, does that answer the question? causally?
   - → example: are kids who play videogames often more aggresive?
   - → does that inform a theory on the aggresiveness effect of VGs?
   - → (you can even try to do a better RQ without causal ID)
2. Is it feasible?
   - → do you have the data? is it possible to do it? (e.g. re-offenders)
   - → also: is there any design or strategy to answer it?

Good RQs, more in detail

1. Consider potential results of the analyses
   - → if you found X, does that answer the question? causally?
   - → example: are kids who play videogames often more aggresive?
   - → does that inform a theory on the aggresiveness effect of VGs?
   - → (you can even try to do a better RQ without causal ID)
2. Is it feasible?
   - → do you have the data? is it possible to do it? (e.g. re-offenders)
   - → also: is there any design or strategy to answer it?
3. Keeping it simple and narrow
   - → what are the causes of economic underdevelopment? vs. does exposure to natural disasters hinders economic development?

Example on generating RQs

- Couple things to remember:
  - → RQs are often the link between theory and empirics
  - → So they already suggest which variation to look at

Example on generating RQs

- Couple things to remember:
  - → RQs are often the link between theory and empirics
  - → So they already suggest which variation to look at

- Imagine you have the following argument:

  How good students do at school depends more on the peers they are surrounded by than on the quality of the teaching they receive

- Which RQ could let us test this?

Stories, RQs, and theories

- There're no exclusive definitions of 'stories' and theories
- It's just about getting to a sufficient level of abstraction
- Often, you start with a story or example, and then you move up the ladder for both theory and RQs until you get to a general theory tested with a RQ

Generating theories

- No recipe for this, everyone generates theories *all the time*
- Usually it refers to an analytical argument that explains something
  - → It could also be a descriptive or predictive theory, but even in those cases there's probably an explanation underneath
- Developed inductively, from descriptive data to general explanations
- My advice: if you can't tell a story out of the theory, you're not there yet (i.e. need to be able to travel from/to abstraction)

Generating theories

- No recipe for this, everyone generates theories *all the time*
- Usually it refers to an analytical argument that explains something
  - → It could also be a descriptive or predictive theory, but even in those cases there's probably an explanation underneath
- Developed inductively, from descriptive data to general explanations
- My advice: if you can't tell a story out of the theory, you're not there yet (i.e. need to be able to travel from/to abstraction)

- **Q:** How to identify a **good theory**?

Evaluating theories

1. Simple

Evaluating theories

1. Simple
2. Internally coherent and able to explain variation

Evaluating theories

1. Simple
2. Internally coherent and able to explain variation
3. Testable

Example (of the whole process)

Example

- That's some descriptive evidence that could inspire an anecdote
- The **anecdotal argument** (think of a story)
- The **research question**
- The 'proper' **theory**
- The **hypotheses**?

Example

- That's some descriptive evidence that could inspire an anecdote
- The **anecdotal argument** (think of a story)
  - → My friend John who went on Erasmus has more money than my other friend who couldn't go, and also, John managed to get a job because his father is partner at a local firm
- The **research question**
  - → Is there a causal effect of Erasmus on labor market early success? Is the effect mediated by household income?
- The 'proper' **theory**
  - → Going on Erasmus does not have any causal effect on getting a first job, the relationship is explained by the confounding effect of income
  - → Or: Positive effect among high-income students because they have access to informal networks where this experience is valued
- The **hypotheses**?

Hypotheses

- Hypothesis is just a very formal term for empirical expectations
  - → Which essentially means being able to say what you expect to see given a theory

Hypotheses

- Hypothesis is just a very formal term for empirical expectations
  → Which essentially means being able to say what you expect to see given a theory
  And ideally, knowing what you need to see in order to discredit the theory

Hypotheses

- Hypothesis is just a very formal term for empirical expectations
  → Which essentially means being able to say what you expect to see given a theory
  And ideally, knowing what you need to see in order to discredit the theory

- Imagine that I have the theory that my knee hurts when I do sports on cold days
  → Simplifying it, we have a 2x2 situation:

Hypotheses: what would you expect to observe?

	Cold day	*Hot day*
Run		
Didn't run		

Hypotheses: what would you expect to observe?

	Cold day	*Hot day*
Run	Pain	Not pain
Didn't run	Not pain	Not pain

What if you observe this?

	Cold day	*Hot day*
Run	Pain	Not pain
Didn't run	Pain	Pain

- New theory?
- How would you test it?

Why bother with theory?

- Why start with questions and arguments?

  Why not just exploit data? (Especially if we have big data)

Why bother about theory?

Teen Arrested in Samurai Killing

April 3, 2000

Click to copy

RELATED TOPICS

Archive

MADRID, Spain (AP) _ Police in southeastern Spain arrested a 16-year-old martial arts fan Monday for killing his parents and little sister with a samurai sword.

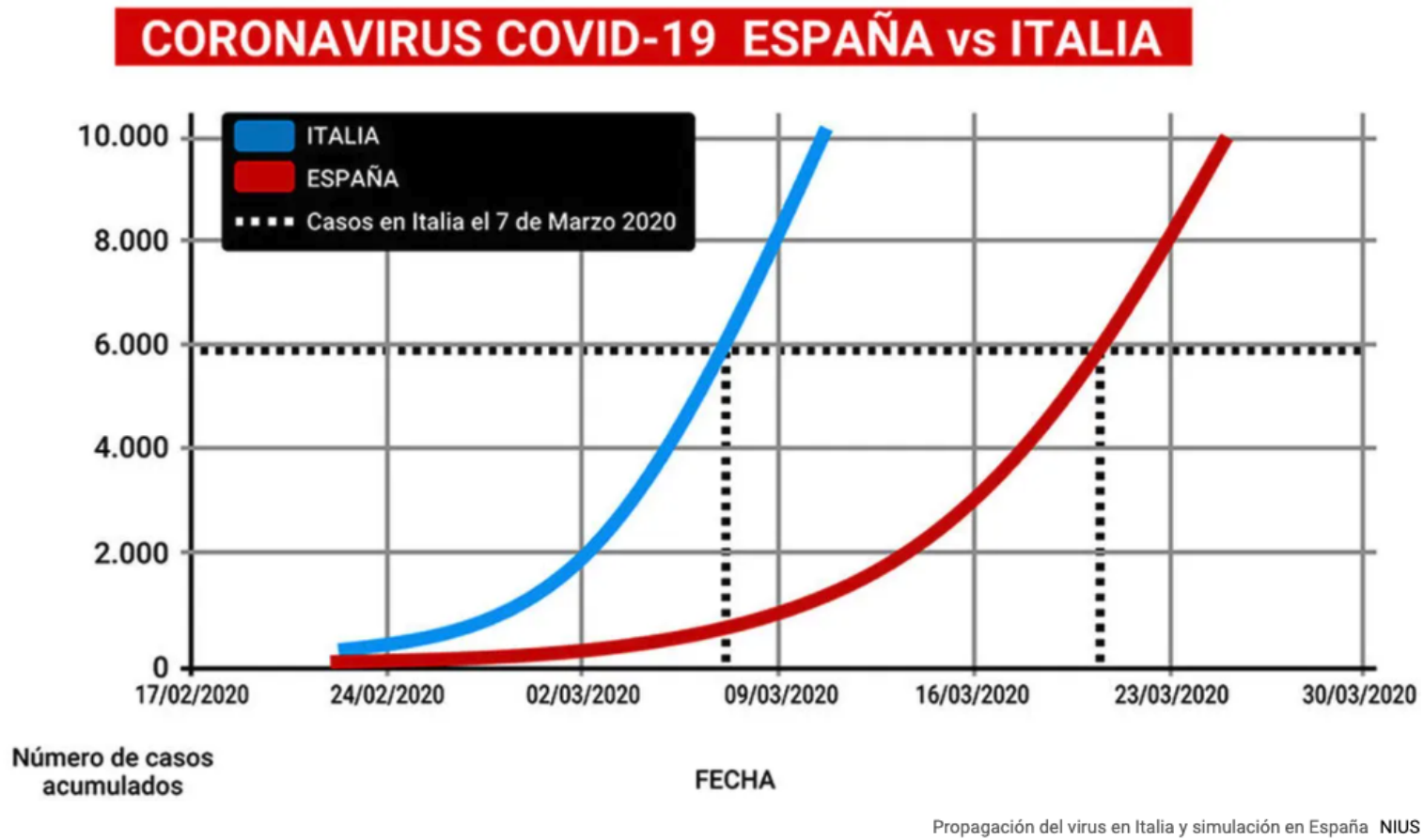
Police arrested the teen in a train station in Alicante as he and a friend were preparing to travel to Barcelona, said Jose Luis Rico, spokesman for the Alicante police.

Why bother about theory?

PORTADA | SOCIEDAD | SANIDAD

La curva del coronavirus: España podría ser la nueva Italia

niusdiario.es • Madrid
10/03/2020 • 05:30h.



Propagación del virus en Italia y simulación en España NIUS

Computational methods and theory

- Limitations of *data mining*
- Focus on the what rather that on the why
- Problems with machine learning
  - → Example of predicting ice cream sales

Why bother about theory?

CIVIL UNREST

Predicting Civil Conflict: What Machine Learning Can Tell Us

Computer programs can be used as early warning systems, allowing the global community to act before violence erupts.

Mechanisms briefly

- A **mechanism** is basically the **how** (or why) of a relationship
- e.g., we know that the flu gives us fever

  $\rightarrow$ flu $>$ fever

Mechanisms briefly

- A **mechanism** is basically the **how** (or why) of a relationship
- e.g., we know that the flu gives us fever

  $\rightarrow$ flu $>$ fever
- What's the mechanism?

  $\rightarrow$ flu $>$ *immune system detects infection* $>$ Δ *body temp* $>$ fever

Mechanisms briefly

- A **mechanism** is basically the **how** (or why) of a relationship
- e.g., we know that the flu gives us fever

  $\rightarrow$ flu $>$ fever
- What's the mechanism?

  $\rightarrow$ flu $>$ *immune system detects infection* $>$ Δ *body temp* $>$ fever
- Good think about mechanisms is that we can try to test them

Testing mechanisms ($\approx$ sub-research questions)

- Let's go back to the Erasmus example
- If our theory is that the effect of going on Eramus is higher for high-income students due to their access to social networks,

  what's the mechanism?
- And how could we try to **test it**?
  - $\rightarrow$ (Think about the sub-RQs)

Recap

- Questions on theory, RQ, or mechanisms?

Roadmap

Theories and research questions

Concepts and operationalization

Measurement

Description

Prediction

Example: Wartime civilian deaths

Paper discussion

Concepts

- What are **concepts**?
- No need to get into epistemological discussions, but basically concepts are the **building blocks of analytical arguments**
- So part of the theoretical framework

Concepts

- What are **concepts**?
- No need to get into epistemological discussions, but basically concepts are the **building blocks of analytical arguments**
- So part of the theoretical framework

- Some times they are a no brainer (income), but in many cases we have to think about them
  - → *Household* income? What's considered a household?
  - → More problematic: Ideology? Democracy?

Concepts

- What are **concepts**?
- No need to get into epistemological discussions, but basically concepts are the **building blocks of analytical arguments**
- So part of the theoretical framework

- Some times they are a no brainer (income), but in many cases we have to think about them
  - → *Household* income? What's considered a household?
  - → More problematic: Ideology? Democracy?
- Also minor point: *concept* $\neq$ *term*
  - → Think about the labels we use to refer to some particular concept, e.g. authoritarian regime (i.e. dictatorship), rationality, etc

Concepts types

- No fixed categories, but some people talk of:

1. **Rule-based** (definition)
   → e.g. what are the rules we could use to define a household?
2. **Ideal types** (or family resemblance?)
   → How do household look like? Can we intuitively identify them?

- Rather that two exclusive types of concepts, they are two ways to think about them which are usually useful in improving concepts

Operationalization

- To translate abstract concepts into concrete stuff we can observe and potentially measure
- Operationalize $\neq$ measure
  - $\rightarrow$ The fact that you can think of a concept in concrete terms does not mean you can always measure it easily
  - $\rightarrow$ Remember the algorithm that models rate of re-offenders
  - $\rightarrow$ Ideology of Twitter users? easy to op, hard to measure **(*)**

Operationalization

- To translate abstract concepts into concrete stuff we can observe and potentially measure
- Operationalize $\neq$ measure
  - → The fact that you can think of a concept in concrete terms does not mean you can always measure it easily
  - → Remember the algorithm that models rate of re-offenders
  - → Ideology of Twitter users? easy to op, hard to measure **(*)**
- More like thinking of real-world attributes that map the conceptual dimensions we think about
  - → *concept*: war intensity; *operationalization*: number of battle deaths
- the **Botswana example** on defining/operationalizing democracies

Importance

- Might seem like something too abstract to care about (especially for computational social science), but it is actually not
- A **huge** part of good quantiative work relies on improving current concepts and their operationalization (which often leads to new ways of measuring them)
- Classic examples with theoretical importance (e.g. Putnam's *social capital*), but also today's paper ('Roads to rule') is a good example of this

Example

- Say we have a question about some x cause of civil war outbreak
- That's two concepts we are actually talking about:

1. **Civil war**
2. **Outbreak**

- How could we define them? And operationalize them?

Another example of a contentious concept

- What is **populism**? How can we operationalize it?
  - → e.g. how could we code a list of *populist* political parties? or leaders?

Roadmap

Theories and research questions

Concepts and operationalization

Measurement

Description

Prediction

Example: Wartime civilian deaths

Paper discussion

Measurement issues

1. Measuring what you really want to measure
   → Careful with the use of proxies
2. Choosing the right unit of analysis
   → Depends on the theory
3. Keep in mind what units you're not observing
   → Missing data, sampling bias

Measuring the right stuff

- Not a lot to say here, other than to **pay attention**
- We normally look at one variable superficially without thinking about how it was created
- How was it exactly measured?
  - → Survey wordings
  - → Coding issues (e.g. level of democracy)
  - → Type of raw data used
- And more importantly, are there **biases related to our question**?

(A few strategies to measure stuff not directly observable)

(Example: we want to measure the ideological or policy positions of political parties)

- **Expert surveys**
  - → You send questionnaires to experts who then reply, aggregate using average or similar
- **Coding written texts**
  - → Manifesto project, but also others based on NLP
- **Observing roll call voting**
  - → Voteview project

(A few strategies to measure stuff not directly observable)

(Example: we want to measure the ideological or policy positions of political parties)

- **Expert surveys**
  - → You send questionnaires to experts who then reply, aggregate using average or similar
- **Coding written texts**
  - → Manifesto project, but also others based on NLP
- **Observing roll call voting**
  - → Voteview project

- All these point to slightly different concepts or operationalizations

(A few strategies to measure stuff not directly observable)

(Example: we want to measure the ideological or policy positions of political parties)

- **Expert surveys**
  - → You send questionnaires to experts who then reply, aggregate using average or similar
- **Coding written texts**
  - → Manifesto project, but also others based on NLP
- **Observing roll call voting**
  - → Voteview project

- All these point to slightly different concepts or operationalizations
- We'll see a different strategy based on *latent variables* in a moment

Measuring the right stuff: Example

- You are doing research on whether discrimination of minorities has a negative effect on overall economic performance of a country
- You find a dataset that lists all minorities in a given country and gives them a yearly score of discrimination from 0 to 10
  - → In the codebook says that discrimination is conceptualized as 'unequal access to state power, which ranges from actual, active discrimination (including mass violence perpetrated against members of the minority group) to lack of access to key political positions in the central government'
- You also learn that the dataset was coded through **expert surveys**, sending a questionnaire to 2–3 researchers from each country
- **What do you think?**

Measuring the right stuff: Example

- Now imagine you use the same dataset to analyze whether more extreme forms of discrimination make violence against minorities more likely
  - → You take the violence data from another dataset that e.g. codes actual violence events from newspapers
- You find a *positive relationship* in the results
- Thoughts?

Measuring the right stuff: Example

- Now imagine you use the same dataset to analyze whether more extreme forms of discrimination make violence against minorities more likely
  - → You take the violence data from another dataset that e.g. codes actual violence events from newspapers
- You find a *positive relationship* in the results
- Thoughts?
- Another issue with expert surveys: within vs between comparisons

Measuring the right stuff: Another example

- Recent debate on **democratic backsliding**
- Problem: how do we measure democracy?
- Available data: many international datasets on democracy rely on subjetive **expert judgement**

EPR example? (https://icr.ethz.ch/data/epr/)

Home People Research Teaching Publications Data GROW^up

ETH Zurich › D-GESS › CIS › ICR › Data › Ethnic Power Relations (EPR) Dataset Family

Ethnic Power Relations (EPR) Dataset Family 2021

The EPR Dataset Family provides data on ethnic groups' access to state power, their settlement patterns, links to rebel organizations, transborder ethnic kin relations, and intraethnic cleavages. The 2014 version has been introduced in Vogt, Bormann, Rüegger, Cederman, Hunziker, Girardin (2015) and has been updated in 2021 in a series of data sets on ethnicity that have stimulated civil war research in the past decade. It features a comprehensive system of tightly integrated data sets:

EPR Core

Politically relevant ethnic groups

The EPR Core dataset identifies all politically relevant ethnic

GeoEPR

Polygons describing ethnic groups

The GeoEPR dataset provides geo-spatial information about

ACD2EPR

Conflicts between ethnic groups

The ACD2EPR docking dataset links conflicts inventoried in

EPR definitions

- **We define ethnicity** as a subjectively experienced sense of commonality based on a belief in common ancestry and shared culture. Different markers may be used to indicate such shared ancestry and culture: common language, similar phenotypical features, adherence to the same faith, and so on. **Our definition of ethnicity thus includes ethnolinguistic, ethnoreligious, and ethnosomatic (or "racial") groups**, but not tribes and clans that conceive of ancestry in genealogical terms, nor regions that do not define commonality on the basis of shared ancestry.
- **An ethnic group is politically relevant if** either at least one significant political actor claims to represent the interests of that group in the national political arena or if group members are systematically and intentionally discriminated against in the domain of public politics.

EPR definitions

- **Monopoly**: Elite members hold monopoly power in the executive to the exclusion of members of all other ethnic groups.
- **Dominance**: ... dominant power in the executive but ... some limited inclusion of "token" members of other groups
- **Senior Partner**: Representatives of the group participate as senior partners in a formal or informal power-sharing arrangement ...(any arrangement that divides executive power among leaders who claim to represent particular ethnic groups and who have real influence on political decision making)
- **Junior Partner**: ... as junior partners in government.
- **Powerless**: Elite representatives hold no political power (or do not have influence on decision making) at the national level of executive power - although without being explicitly discriminated against.
- **Discrimination**: Group members are subjected to active, intentional, and targeted discrimination (formal or informal) by the state, with the intent of excluding them from political power (but not from socio-economic sphere).
- **Self-exclusion**: groups that have excluded themselves from central state power, in the sense that they control a particular territory of the state which they have declared independent

EPR in Spain?

- How would you code Spain? (Or your own country, in case it's multi-ethnic)
- Two main things?

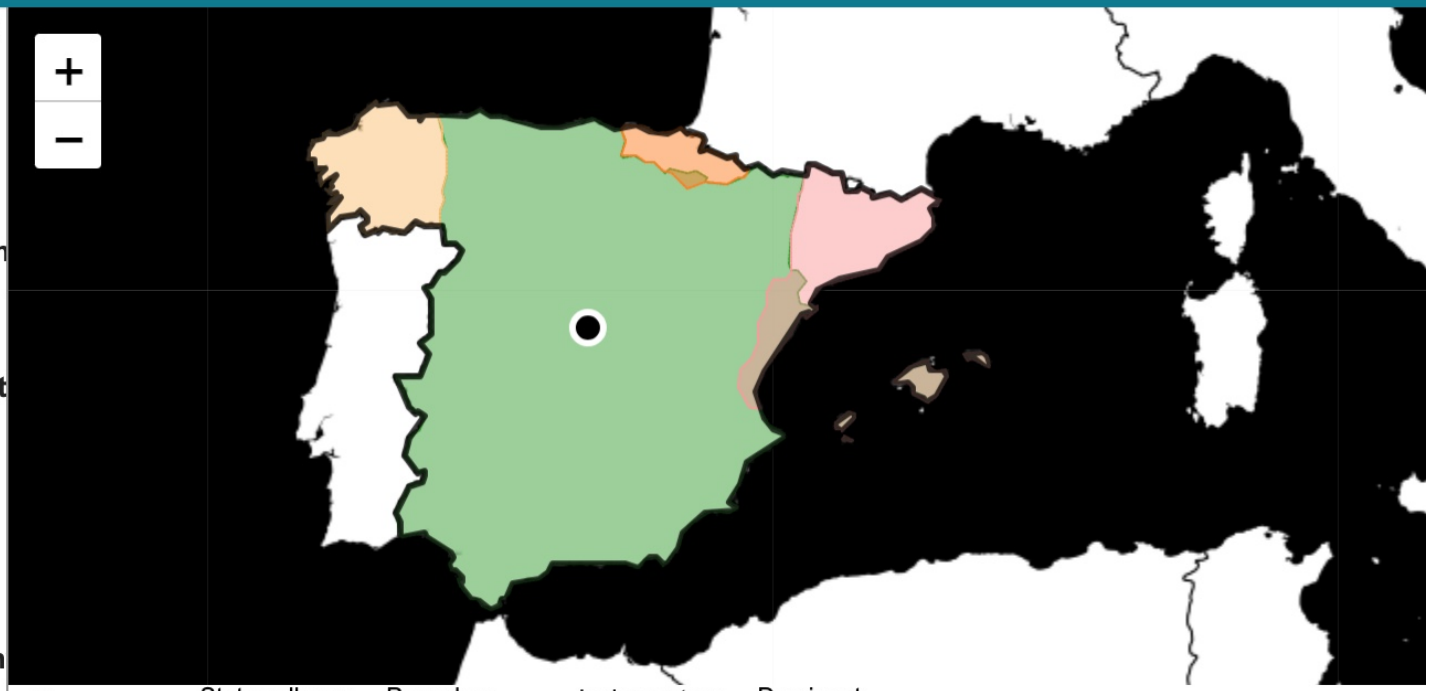
1. How many *politically relevant* ethnic groups?
2. Political status by period?

Department of Humanities, Social and Political Sciences | International Conflict Research

GROWup - Geographical Research On War, Unified Platform

View | Read | Download | Code | About

Solomon Islands
Somalia
South Africa
South Sudan
Spain
Sri Lanka (Ceylon
Sudan
Surinam
Swaziland (Eswat
Sweden
Switzerland
Syria
Taiwan
Tajikistan
Tanzania (Tangan
Thailand
Tibet
Togo
Trinidad and Toba
Tunisia
Turkey (Ottoman
Turkmenistan
Uganda
Ukraine
United Arab Emir



GROUPNAME	COLOR	1970										1980										1990		
		9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
Basques											★	★	★°	★°	★°	°	°	★°	★°	★°	°	°	°	★°
Catalans											°	°	°	°	°	°	°	°	°	°	°	°	°	°
Galicians													°	°	°	°	°	°	°	°	°	°	°	°
Roma																								
Spanish																								

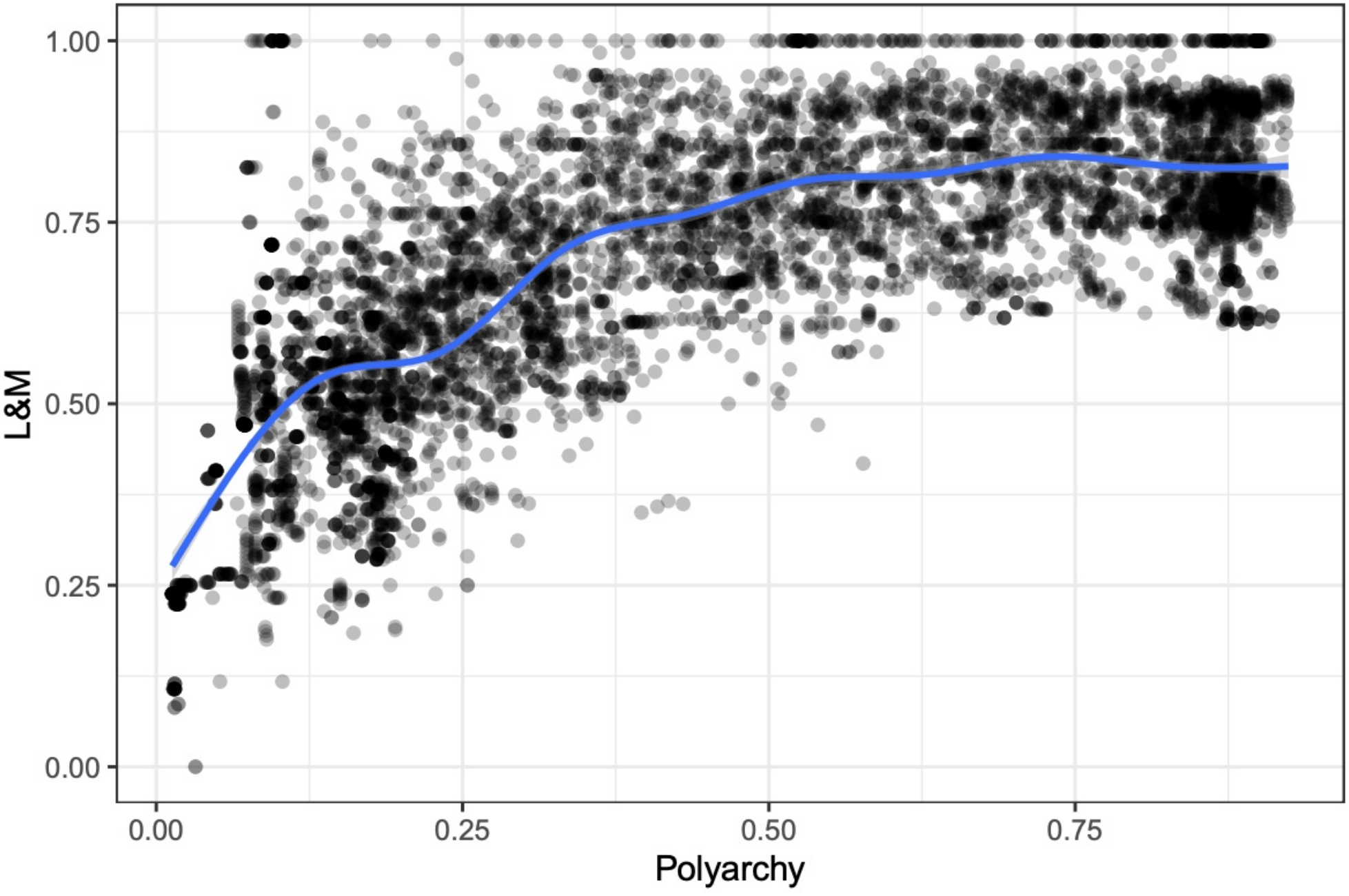
Democratic backsliding

- 'Objective' and 'subjective' operationalization & measurement
- 'Objective' measures usually rely on a *minimalist* conceptualization of democracy
  - → e.g. celebration of contested elections
- 'Subjective' measures tap into *maximalist* definitions of democracy that incorporate more dimensions
  - → regular, contested elections; but also rule of law, participation, media, etc
- Problem: Autocrats are usually pretty skillful

The case for objective measures

- One example from the ACLP (Alvarez, Cheibub, Limongi, Przeworski) Democracy and Dictatorship Dataset
- Coding democracy based on four objective, observable rules:

1. The chief executive must be chosen by popular election or by a body that was itself popularly elected.
2. The legislature must be popularly elected.
3. There must be more than one party competing in the elections.
4. An alternation in power under electoral rules identical to the ones that brought the incumbent to office must have taken place.


1.00
0.75
0.50
0.25
0.00
L&M
0.00
0.25
0.50
0.75
Polyarchy

One view on this (by V-Dem people)

- Even if using expert surveys, you can take some measures
  - → Aim for *replicability*
  - → Incorporate measures of uncertainty
  - → Build it differently: e.g. incorporate different dimensions, use an ordinal scale, aggregate differently, etc
- 'Objective measures' are not that objective
  - → Botswana example, and systematic downward bias against young democracies with economic growth
  - → e.g. how do you detect fraud? election forensics methods (based on distribution) can be incorporated by autocrats in later elections

- To know more: Knutsen *et al.* 'Conceptual and Measurement Issues in Assessing Democratic Backsliding.' V-Dem Working Paper, May 2023.
  - → v-dem.net/media/publications/wp_140.pdf

Proxies

- A **proxy variable** is a variable that we use to substitute another variable we cannot observe or measure
- This is a matter of creativity, but the important thing is to think about **potential biases**
- A real example:
  - → Trying to know if leftist/Basque nationalist priests during Francoist Spain had an effect on later terrorism
  - → First problem (among many): how do you measure the ideology of these people?
  - → Using the 1963 letter to the Vatican

Latent variables

- Some concepts are just not directly observable
  - $\rightarrow$ (or very expensive / unfeasible to do so)
- Another option is to create the variable out of other observables
- This is sometimes called **latent variables**

Latent variables

- Let's look at one example: imagine you want to do research on whether left-wing or right-wing people tweet differently (or some other outcome, e.g. echo chambers idea)
- It's easy to get data on the outcome variable (Tweet content, frequency, ...)
  - → if you don't know how now, you'll learn in the spring
- But **how do you code ideology?**

Latent variables

- Let's look at one example: imagine you want to do research on whether left-wing or right-wing people tweet differently (or some other outcome, e.g. echo chambers idea)
- It's easy to get data on the outcome variable (Tweet content, frequency, ...)
  - → if you don't know how now, you'll learn in the spring
- But **how do you code ideology?**
  - → Some people have done it focusing only on a subset, e.g. politicians, for which you have information (problem of selection)

Latent variables

- Let's look at one example: imagine you want to do research on whether left-wing or right-wing people tweet differently (or some other outcome, e.g. echo chambers idea)
- It's easy to get data on the outcome variable (Tweet content, frequency, ...)
  - → if you don't know how now, you'll learn in the spring
- But **how do you code ideology?**
  - → Some people have done it focusing only on a subset, e.g. politicians, for which you have information (problem of selection)
  - → Of even some others have linked survey data to Twitter activity, asking for consent (problem of cost, non-response)

Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data

Pablo Barberá

Wilf Family Department of Politics, New York University, 19 W 4th Street, 2nd Floor, New York, NY 10012.
e-mail: pablo.barbera@nyu.edu

Edited by R. Michael Alvarez

Politicians and citizens increasingly engage in political conversations on social media outlets such as Twitter. In this article, I show that the structure of the social networks in which they are embedded can be a source of information about their ideological positions. Under the assumption that social networks are homophilic, I develop a Bayesian Spatial Following model that considers ideology as a latent variable, whose value can be inferred by examining which politics actors each user is following. This method allows us to estimate ideology for more actors than any existing alternative, at any point in time and across many polities. I apply this method to estimate ideal points for a large sample of both elite and mass public Twitter users in the United States and five European countries. The estimated positions of legislators and political parties replicate conventional measures of ideology. The method is also able to successfully classify individuals who state their political preferences publicly and a sample of users matched with their party registration records. To illustrate the potential contribution of these estimates, I examine the extent to which online behavior during the 2012 US presidential election campaign is clustered along ideological lines.

2 Ideal Point Estimation Using Twitter Data

2.1 *Assumptions*

In this article, I demonstrate that valid ideal point estimates of individual Twitter users and political actors with a Twitter account can be derived from the structure of the "following" links between these two sets of users. In order to do so, I develop a Bayesian spatial model of Twitter users' following behavior.

The key assumption of this model is that Twitter users prefer to follow politicians whose positions on the latent ideological dimension are similar to theirs. This assumption is equivalent to that of spatial voting models (see, e.g., Enelow and Hinich 1984). I consider following decisions to be costly signals about users' perceptions of both their ideological location and that of political accounts. Such cost can take two forms. If the content of the messages users are exposed to as a

Validating measure

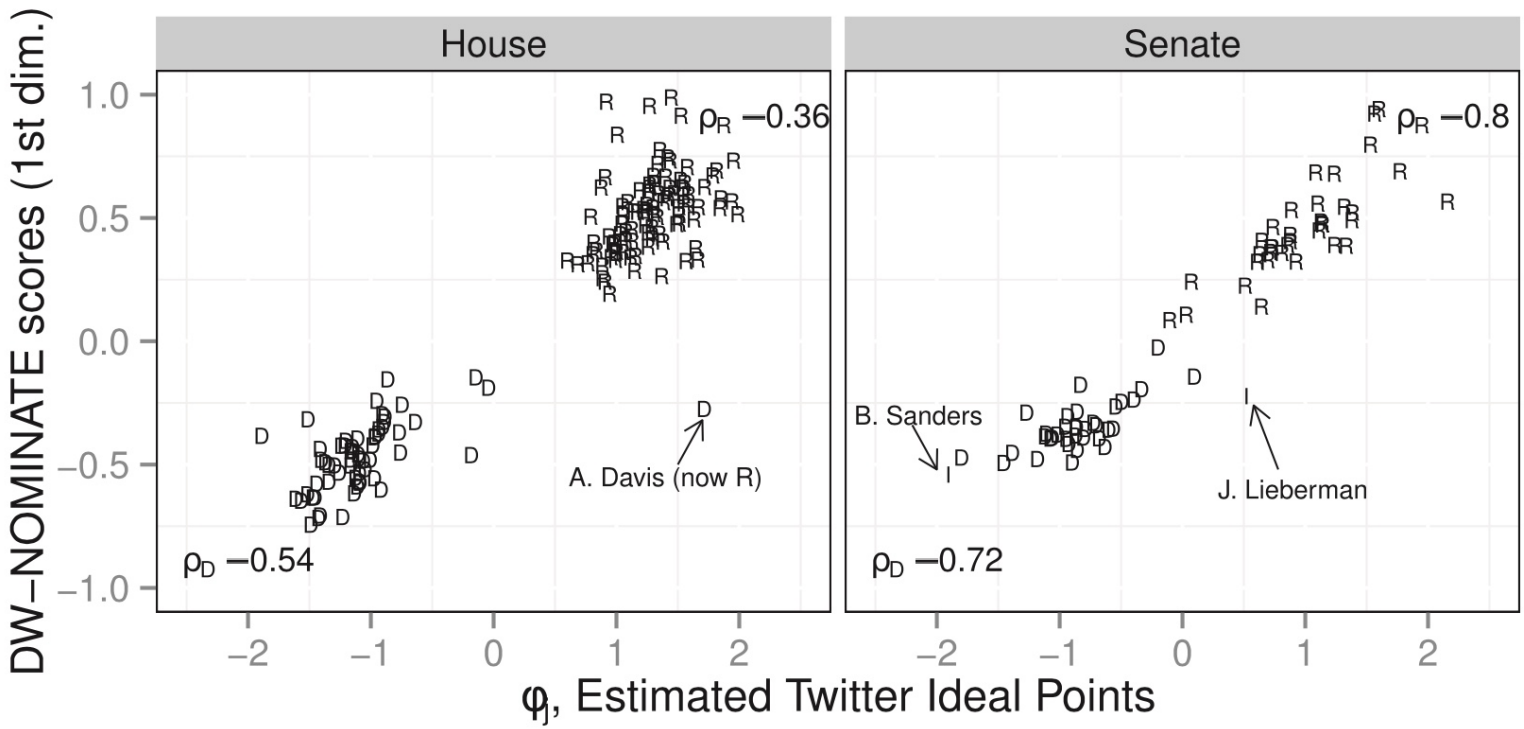


Fig. 1 Ideal point estimates for members of US Congress.

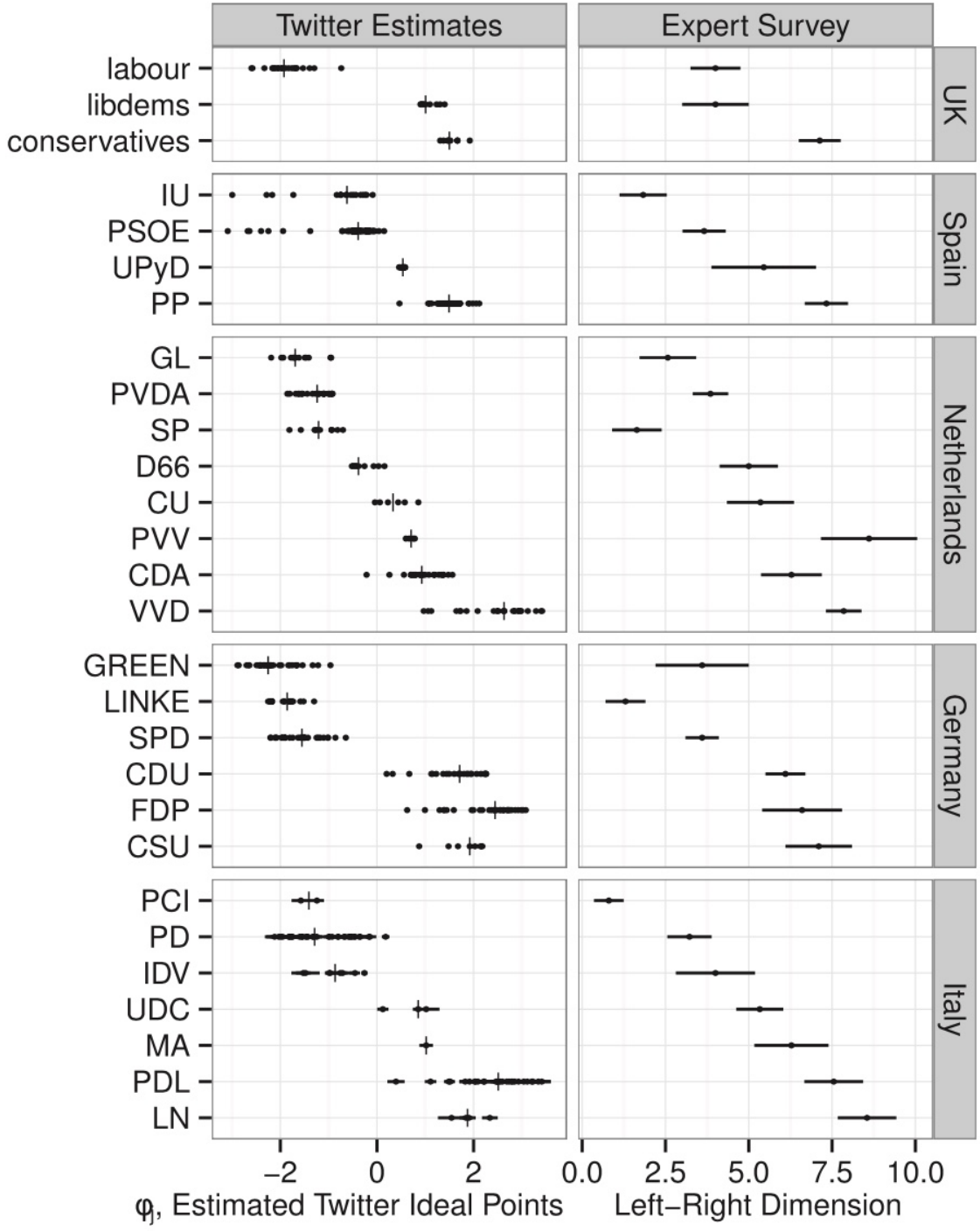


Fig. 3 Ideological location of parties in five European countries.

Psychological Science
1–12

DOI: 10.1177/0956797615594620
pss.sagepub.com
SAGE

Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?

Pablo Barberá[1], John T. Jost[1,2,3], Jonathan Nagler[3], Joshua A. Tucker[3], and Richard Bonneau[4]

[1]Center for Data Science, [2]Department of Psychology, [3]Department of Politics, and [4]Center for Genomics and Systems Biology, New York University

Abstract

We estimated ideological preferences of 3.8 million Twitter users and, using a data set of nearly 150 million tweets concerning 12 political and nonpolitical issues, explored whether online communication resembles an "echo chamber" (as a result of selective exposure and ideological segregation) or a "national conversation." We observed that information was exchanged primarily among individuals with similar ideological preferences in the case of political issues (e.g., 2012 presidential election, 2013 government shutdown) but not many other current events (e.g., 2013 Boston Marathon bombing, 2014 Super Bowl). Discussion of the Newtown shootings in 2012 reflected a dynamic process, beginning as a national conversation before transforming into a polarized exchange. With respect to both political and nonpolitical issues, liberals were more likely than conservatives to engage in cross-ideological dissemination; this is an important asymmetry with respect to the structure of communication that is consistent with psychological theory and research bearing on ideological differences in epistemic, existential, and relational motivation. Overall, we conclude that previous work may have overestimated the degree of ideological segregation in social-media usage.

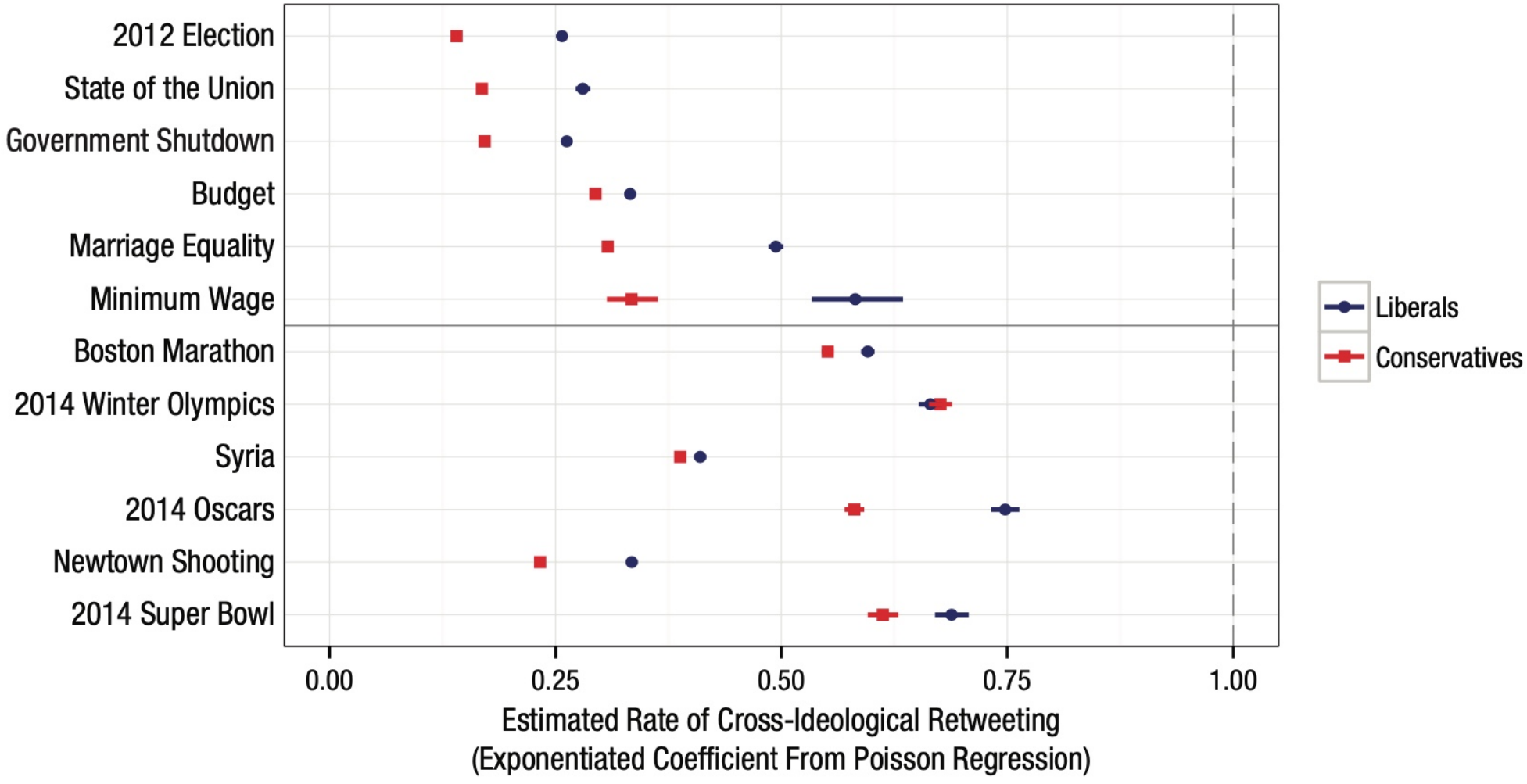


Fig. 4. Liberal-conservative asymmetries in cross-ideological retweeting. The graph shows the estimated rate of cross-ideological retweeting for each tweet collection and for each ideological group after adjusting for each group's propensity to retweet and be retweeted; each point corresponds to an exponentiated coefficient of a Poisson regression for the indicated topic and ideological group. The error bars indicate 99.9% confidence intervals (not visible in some cases because of their small size). An exponentiated coefficient of 1 (highlighted by the dashed vertical line) would indicate identical retweeting rates for individuals of the same and different ideological orientations—that is, a rate of cross-ideological retweeting that is equal to the rate of within-group retweeting.

Unit of analyses

- Level at which we have our observations
- Deeply related to the variables we have
  - $\rightarrow$ Even though not all variables have to/can be measured at the same level

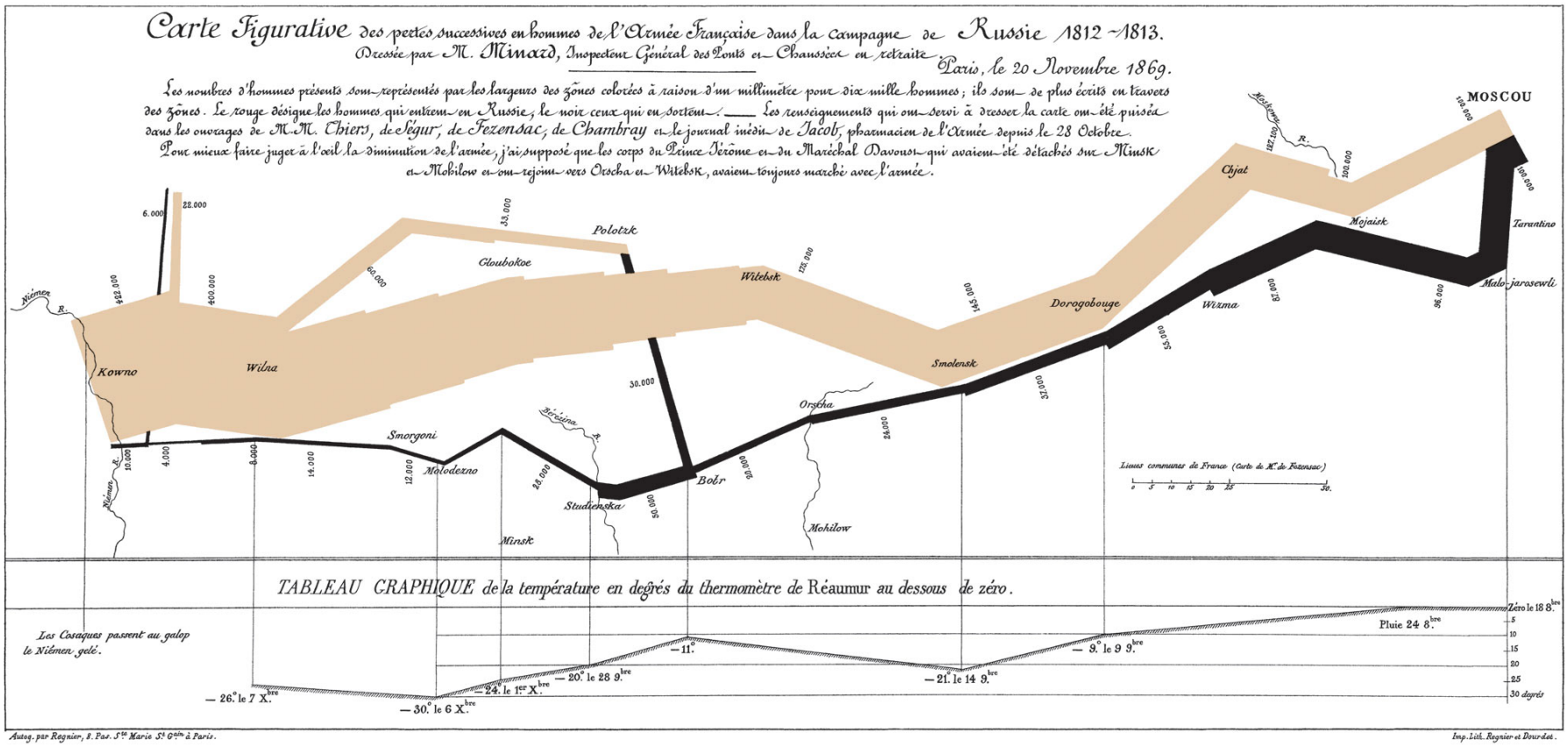
Unit of analyses

- Level at which we have our observations
- Deeply related to the variables we have
  - → Even though not all variables have to/can be measured at the same level
  - → e.g. individual-level data and household income

Unit of analyses

- Level at which we have our observations
- Deeply related to the variables we have
  - → Even though not all variables have to/can be measured at the same level
  - → e.g. individual-level data and household income
- Most important thing: we need to **choose the right unit of analyses** depending on the theory (and mechanism) we are testing

A more difficult example



- How many variables and what's the unit of observation?

Extra: What's the causal argument being told here?

Could we test it with the data we have?

Theories, hypotheses, and measurement

- Let's say I want to explain the effect of school choice on future salaries
- My argument is: going to private schools leads to higher salaries in the future because increased resources lead to better educational attainment through lower teacher/pupil ratio, which signals individuals as more skillful in the labour market, explaining higher salaries

Theories, hypotheses, and measurement

- Let's say I want to explain the effect of school choice on future salaries
- My argument is: going to private schools leads to higher salaries in the future because increased resources lead to better educational attainment through lower teacher/pupil ratio, which signals individuals as more skillful in the labour market, explaining higher salaries

- Hypotheses?

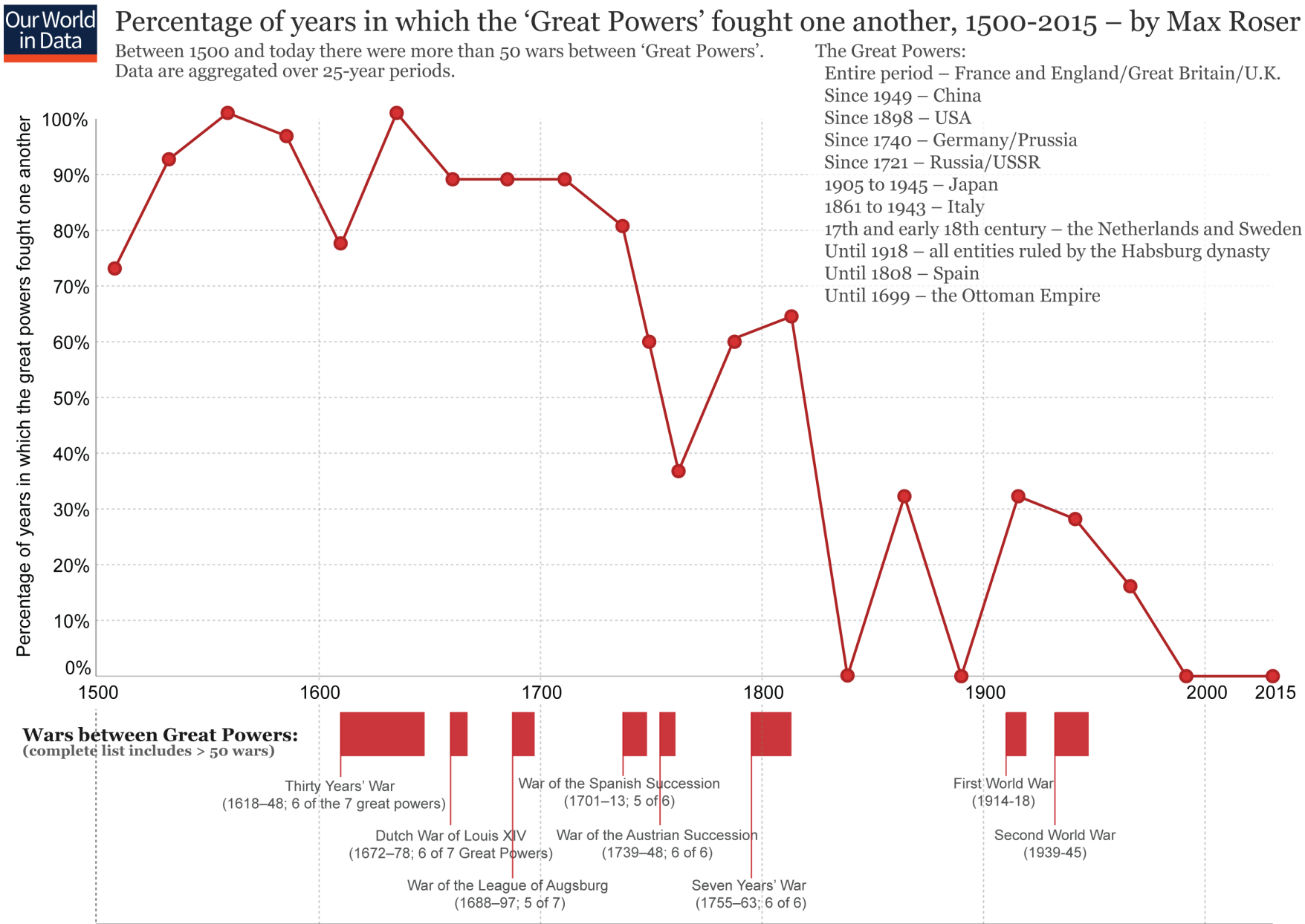
Theories, hypotheses, and measurement

- Let's say I want to explain the effect of school choice on future salaries
- My argument is: going to private schools leads to higher salaries in the future because increased resources lead to better educational attainment through lower teacher/pupil ratio, which signals individuals as more skillful in the labour market, explaining higher salaries

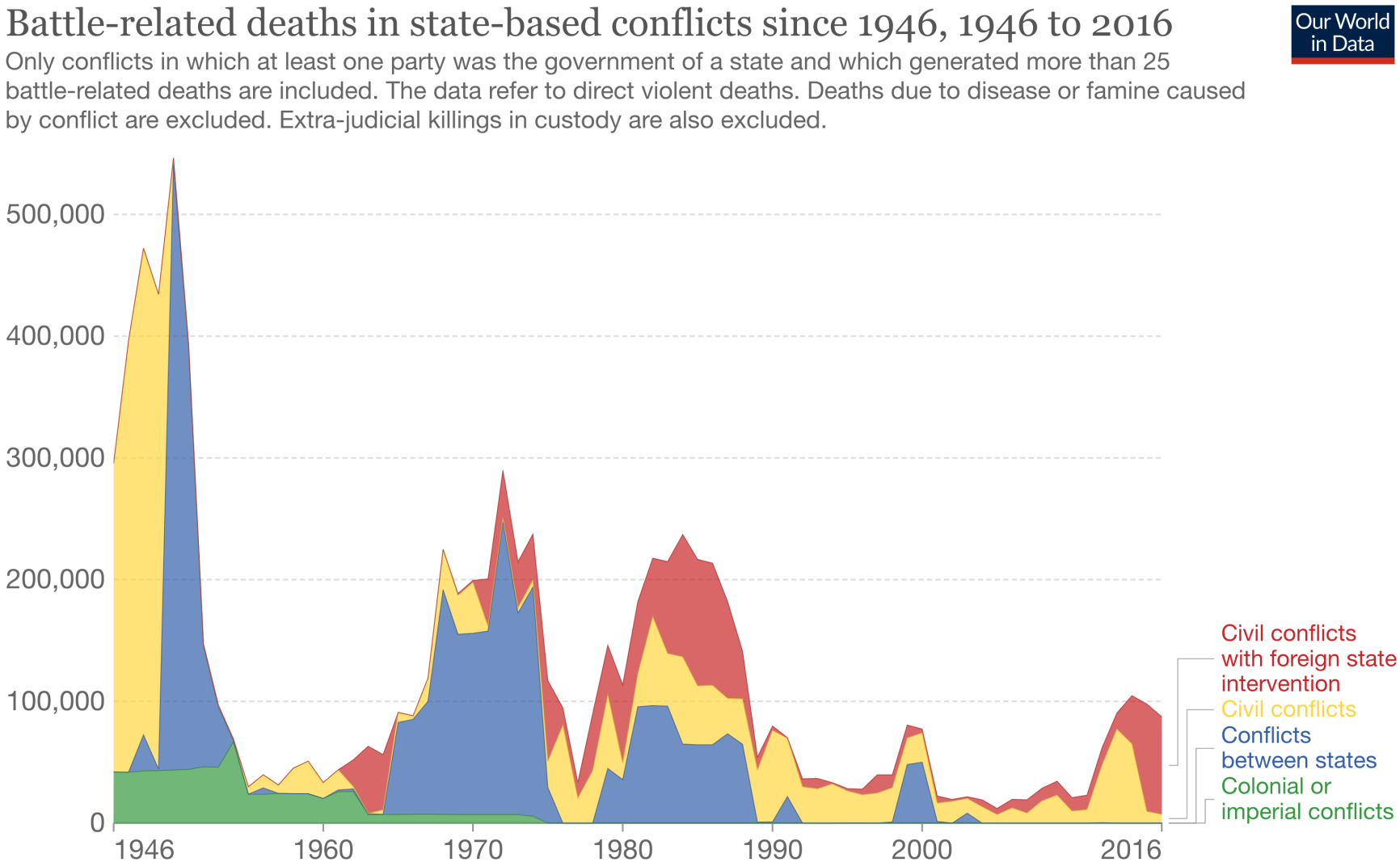
- Hypotheses?
- Testing the relationship and the mechanism?

  And alternative explanations?

Another example



Another example



Another example

- Now, let's say my theory is that inter-state war has declined because democratic countries are less likely to go to war because they face higher domestic costs for waging wars

- Hypotheses? Testing the mechanism? Unit of analyses? Measurement? Alternative explanations?

Another example

- What if I say that it is because democracies do not fight *each other*, as they have shared interests in the international system and shared conflict resolution mechanisms?

- And if I say that democratic countries face higher costs when fighting another democracy, but not otherwise?

- What should I observe in each case? At different levels? How to measure?

Beyond our observations: missing data

- **Missing data** is often more important that it looks
- Important to understand if it's biasing our analyses
- Three types:

1. Missing **completely at random**
   → No problem, random observations are missing (Probably not very often)
2. Missing **at random**
   → One variable explains whether obs are missing or not, but it's not related to our question
3. Missing **not at random**
   → The variable that explains 'missingness' is key to our question

- Example from patients, smoking, gender, emergency

Beyond our observations: sampling bias

- **Sampling bias** could be thought of as missing data or, rather, as a controlling variable we'll indirectly including
- Easy case: we're dealing with a pre-designed sample that might have some biases
  - → Online survey and +65
- More easy to miss: there is an 'invisible' determining which observations we have or not
  - → e.g. when using Twitter data,
- We'll talk more about how this affects inference
  - → Collider bias example?

Complexity of the social world and micro/macro

- Why do mass protests emerge?

Levels of explanation

Can you think of...?

- Macro-level mechanisms
- Micro-level mechanisms

Levels of explanation

Can you think of...?

- Macro-level mechanisms
- Micro-level mechanisms

- What's the point of macro-level explanations, actually?

Extra: From qualitative to quantitative

- Imagine you have this question:

  What are the differences between neoclassicism and romanticism in music?

- Could you try to make a quantitative question out of it?

Roadmap

Theories and research questions

Concepts and operationalization

Measurement

Description

Prediction

Example: Wartime civilian deaths

Paper discussion

Describing variables

- What is a **variable**?
- Types?

Describing variables

- What is a **variable**?
- Types?
  - → Continuous

Describing variables

- What is a **variable**?
- Types?
  - → Continuous
  - → Count

Describing variables

- What is a **variable**?
- Types?
  - → Continuous
  - → Count
  - → Ordinal

Describing variables

- What is a **variable**?
- Types?
  - → Continuous
  - → Count
  - → Ordinal
  - → Categorical (binary)

Describing variables

- What is a **variable**?
- Types?
  - → Continuous
  - → Count
  - → Ordinal
  - → Categorical (binary)
  - → Qualitative (* are really a variable?)

Describing variables

- What is a **variable**?
- Types?
  - → Continuous
  - → Count
  - → Ordinal
  - → Categorical (binary)
  - → Qualitative (* are really a variable?)
- Why does it matter?
- Conceptual meaning vs statistical meaning

Describing variables

- Main idea: you are describing the variable distribution (i.e. how the frequency of values looks like)
- You probably know this from basic statistics
  - → In practice, the measures of distribution do not matter so much
- But one important thing: we are talking about **real-world observations**, so before you do anything (analyses, etc), do look at them
  - → At least, **plot the main variables**
  - → Is it coherent with the **theoretical** or **expected distribution**?

Describing variables

- Also, sometimes the distribution is important to think about actual effect sizes, so it's good to summarize variables (mean, SD, IQR...)
  - → Maybe this makes sense if you've learn logistic regression?
  - → We'll talk more tomorrow about the concept of average effect in causality
- In a normal distribution, there's probably not much to say
- But what if a key independent variable has a bimodal distribution? What does this say about the **causal mechanism?**
  - → e.g. think about the effect of income on X in two societies: one is extremely unequal and the other is normally distributed

Describing relationships

- What is a **relationship**?
- Essentially that as you know about the values of one variable, you learn about the values of the other variables

  e.g. a *negative* relationship means that you know that higher values in x imply lower values in y

Describing relationships

- What is a **relationship**?
- Essentially that as you know about the values of one variable, you learn about the values of the other variables

  e.g. a *negative* relationship means that you know that higher values in x imply lower values in y

  Imagine you have a small car, and a friend of yours is coming and is bringing along his two kids. Concerned about space, you ask '*how old are they?*' And the answer is: '*They're 6 and 2.*'

  $\rightarrow$ What do you imagine about size?

Describing relationships

- What is a **relationship**?
- Essentially that as you know about the values of one variable, you learn about the values of the other variables

  e.g. a *negative* relationship means that you know that higher values in x imply lower values in y

  Imagine you have a small car, and a friend of yours is coming and is bringing along his two kids. Concerned about space, you ask '*how old are they?*' And the answer is: '*They're 6 and 2.*'

  → What do you imagine about size?

  → Now imagine you ask '*are they blonde, red-haired, or brown-haired?*'

Statistical relationship $\neq$ causal relationships

- Last example: Is there a causal relationship *age* $\rightarrow$ *size*?

Statistical relationship $\neq$ causal relationships

- Last example: Is there a causal relationship *age* $\rightarrow$ *size*?
- What if the variable you want to guess is **the time of the day**, and someone tells you that she just heard the rooster crow? Causal?

Statistical relationship $\neq$ causal relationships

- Last example: Is there a causal relationship *age* $\rightarrow$ *size*?
- What if the variable you want to guess is **the time of the day**, and someone tells you that she just heard the rooster crow? Causal?

- Why are non-causal descriptive relationships **useful**?

Univariate description

Facebook says there are only 3.57 degrees of separation

By James Titcomb

4 February 2016 • 10:06am

Univariate description

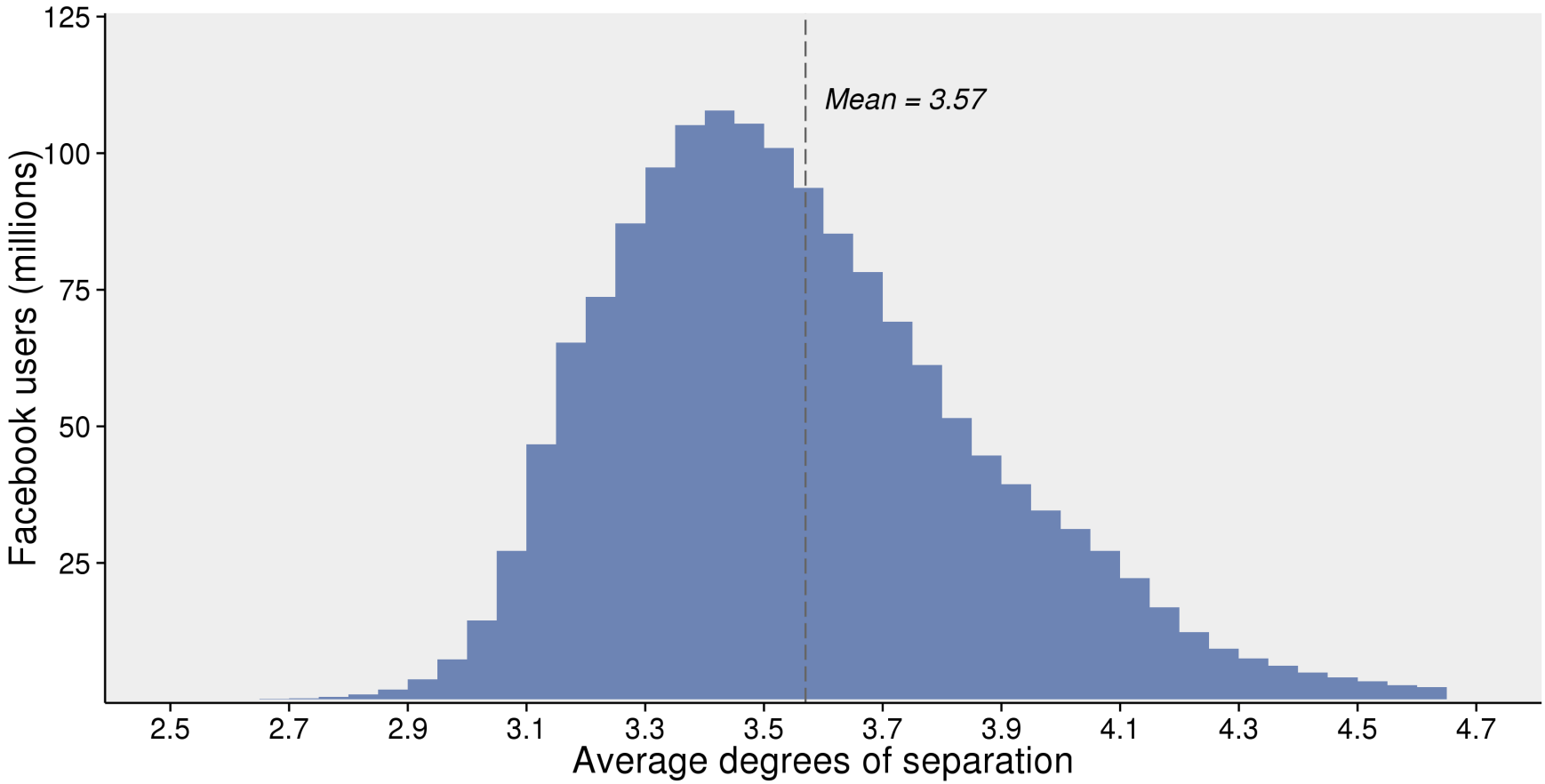


Fig 1. Estimated average degrees of separation between all people on FB.

(https://research.facebook.com/blog/2016/2/three-and-a-half-degrees-of-separation/)

The original 'theory'

I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice. fill in the names.

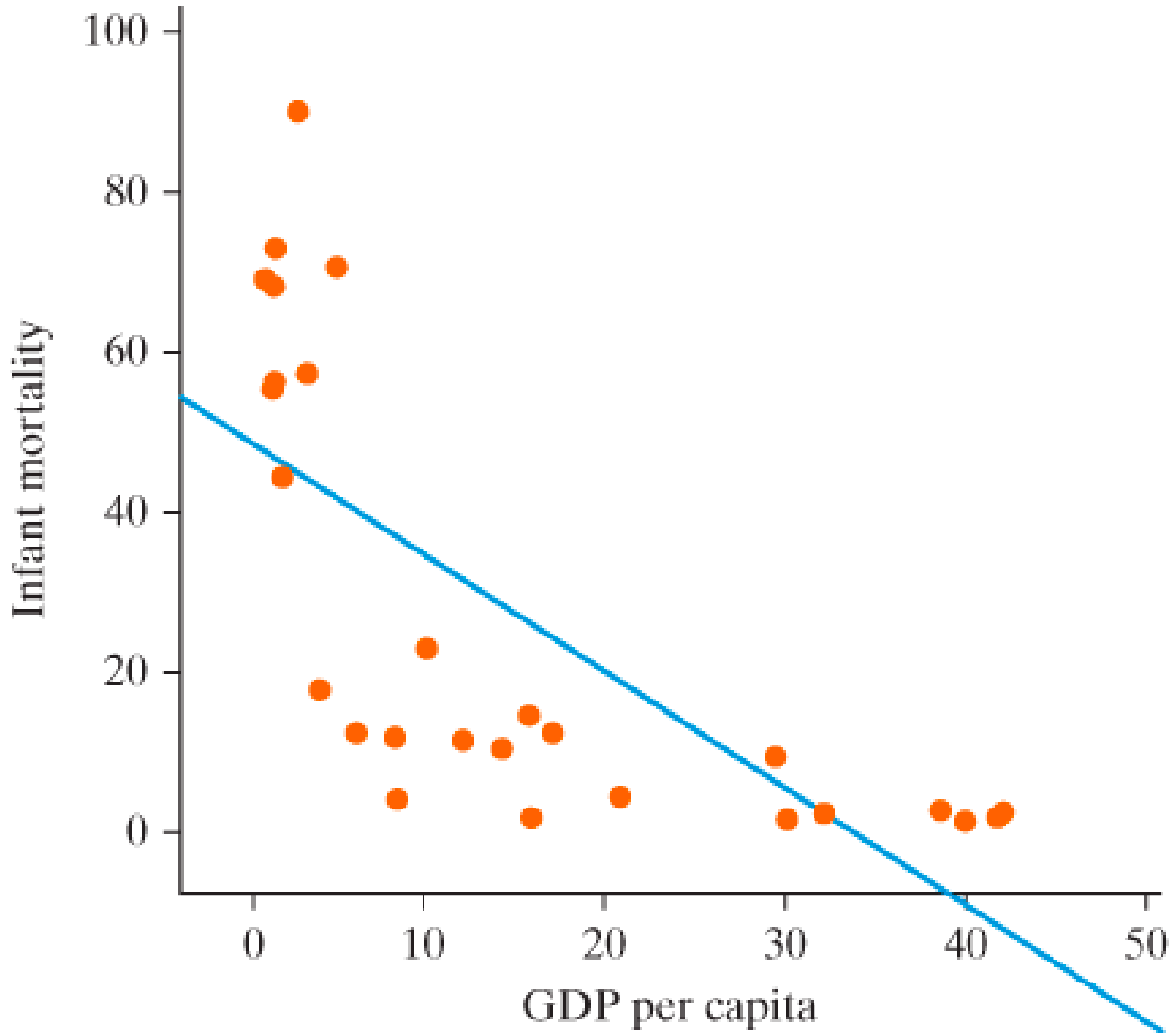
Six Degrees of Separation, John Guare

- Do we have an answer?

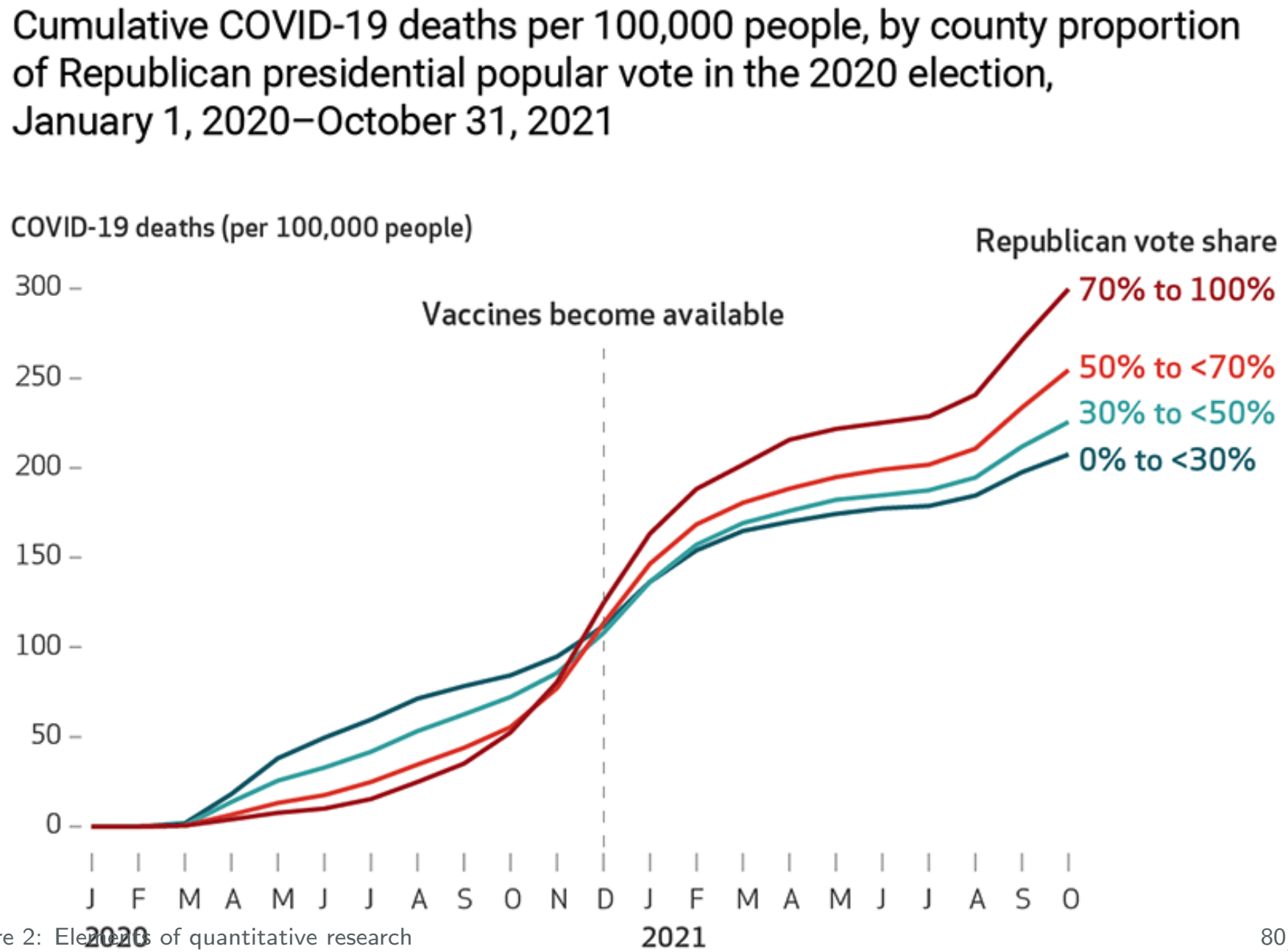
Bivariate relationships

- Examples?

Bivariate relationship

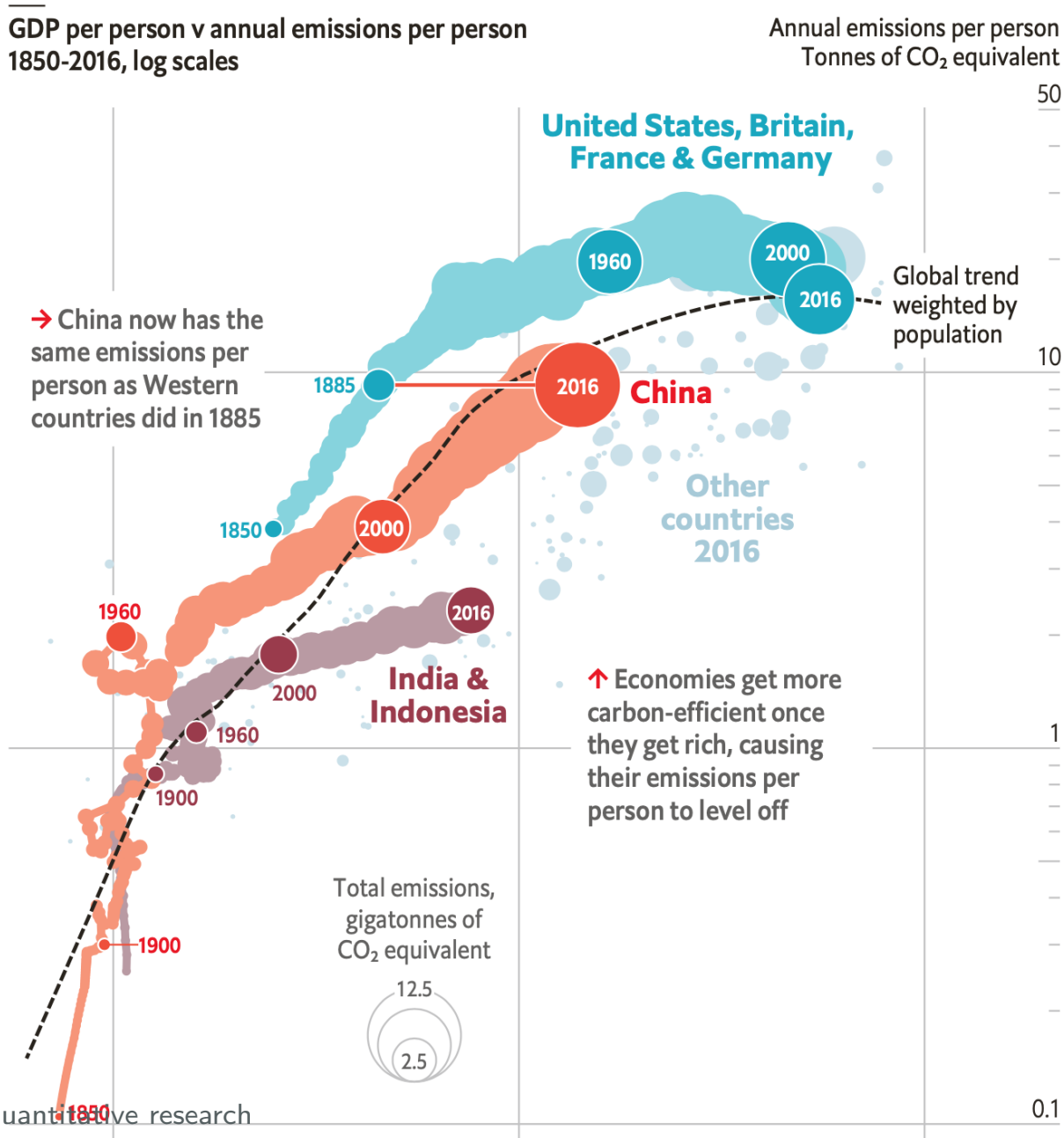


How many variables? Unit?



How many variables? Unit?

How many variables? Unit?



Bivariate relationships

- What do we use this for?

Bivariate relationships

- What do we use this for?
- Essentially, we are trying to detect whether two variables are dependent
  - → In other words, it's about conditional values: $E(Y|X)$

Bivariate relationships

- What do we use this for?
- Essentially, we are trying to detect whether two variables are dependent
  - → In other words, it's about conditional values: $E(Y|X)$
- Example graph about infant mortality:

  $E(IM|GDPpc = 1000)$?

  $E(IM|GDPpc = 30000)$?

Bivariate relationships

- This is what statistics is about, and only this

Bivariate relationships

- This is what statistics is about, and only this
- Even if it can get complicated: non-linear relationships, multivariate dependencies, etc

Bivariate relationships

nature > nature communications > articles > article

Article | Open Access | Published: 20 August 2018

Sequences of purchases in credit card data reveal lifestyles in urban populations

Riccardo Di Clemente, Miguel Luengo-Oroz, Matias Travizano, Sharon Xu, Bapu Vaitla & Marta C. González

Nature Communications **9**, Article number: 3330 (2018) | Cite this article

14k Accesses | **31** Citations | **268** Altmetric | Metrics

This article has been updated

Abstract

Zipf-like distributions characterize a wide set of phenomena in physics, biology, economics, and social sciences. In human activities, Zipf's law describes, for example, the frequency of appearance of words in a text or the purchase types in shopping patterns. In the latter, the uneven distribution of transaction types is bound with the temporal sequences of purchases of individual choices. In this work, we define a framework using a text compression technique

Bivariate relationships

RESEARCH ARTICLE

Faces in the crowd: Twitter as alternative to protest surveys

Christopher Barrie[1☯*], **Arun Frey**[2,3☯]

1 School of Social & Political Science, University of Edinburgh, Scotland, United Kingdom, **2** Department of Sociology, University of Oxford, England, United Kingdom, **3** Leverhulme Centre for Demographic Science, Oxford, England, United Kingdom

☯ All these authors contributed equally to this work and are listed in alphabetical order.
* christopher.barrie@ed.ac.uk

Abstract

Who goes to protests? To answer this question, existing research has relied either on retrospective surveys of populations or in-protest surveys of participants. Both techniques are prohibitively costly and face logistical and methodological constraints. In this article, we investigate the possibility of surveying protests using Twitter. We propose two techniques for sampling protestors on the ground from digital traces and estimate the demographic and ideological composition of ten protestor crowds using multidimensional scaling and machine-learning techniques. We test the accuracy of our estimates by comparing to two in-protest surveys from the 2017 Women's March in Washington, D.C. Results show that our Twitter sampling techniques are superior to hashtag sampling alone. They also approximate the ideology and gender distributions derived from on-the-ground surveys, albeit with some bias, but fail to retrieve accurate age group estimates. We conclude that online samples are

Describing relationships

- When we find a conditional relationship, we often say that X *explains* Y
- But these statistical relationships do *not* tell us anything about cause and effect, only about conditional means (or $E(Y|X)$, or conditional conditional means if we also control for Z)
- We need another strategy to understand *why*

Explanatory questions and data

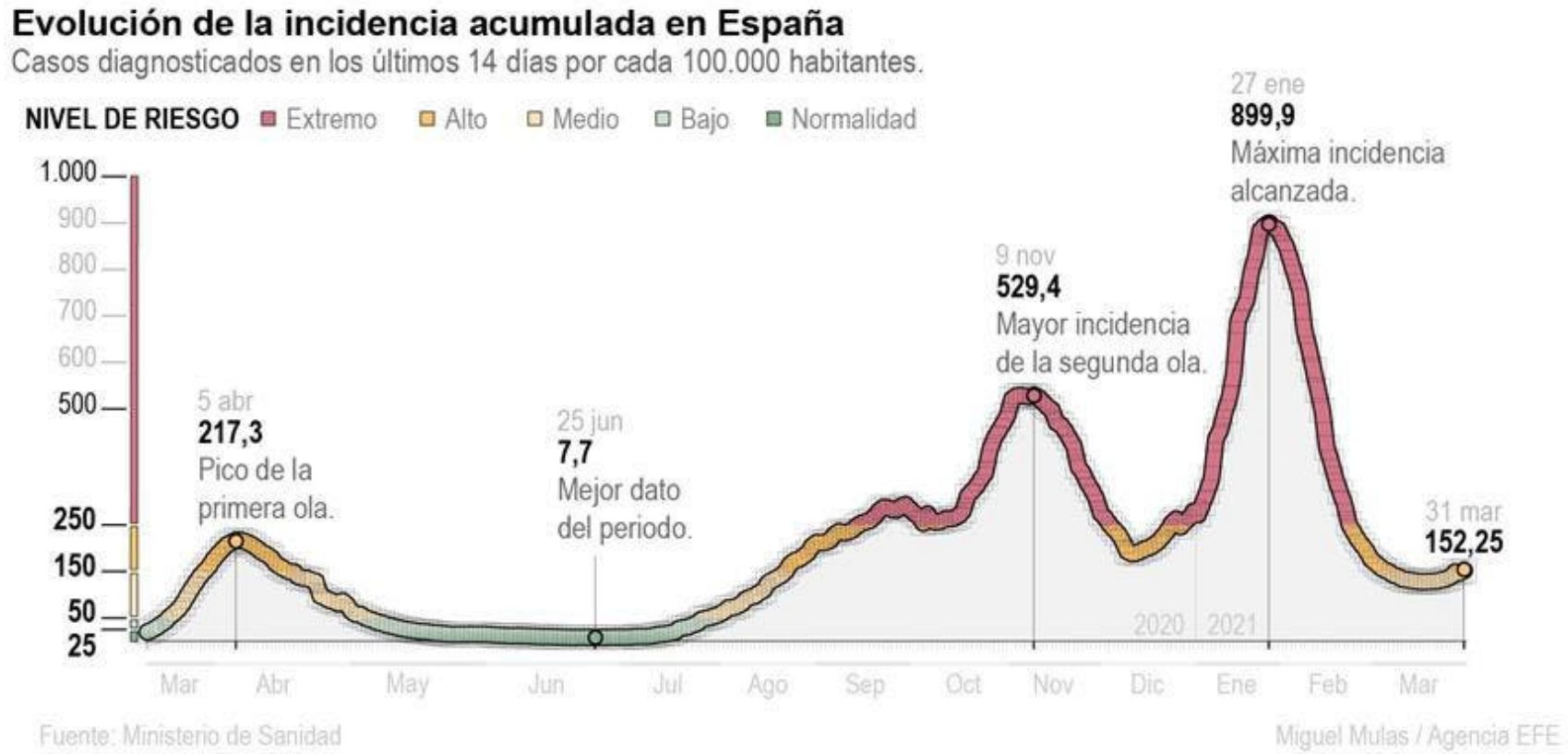
- Let's we want to know what's going on between two things (the process or mechanism between two variables)
- Data always comes from somewhere
- So when we look at or analyze data, we try to uncover the *data generating process*
  - → There could be several mechanisms generating the same data (variable), or they might vary over time
- (A model is actually our simplified guess of that process)

Data generating process

- How does the data look like, and what's the data generating process in...

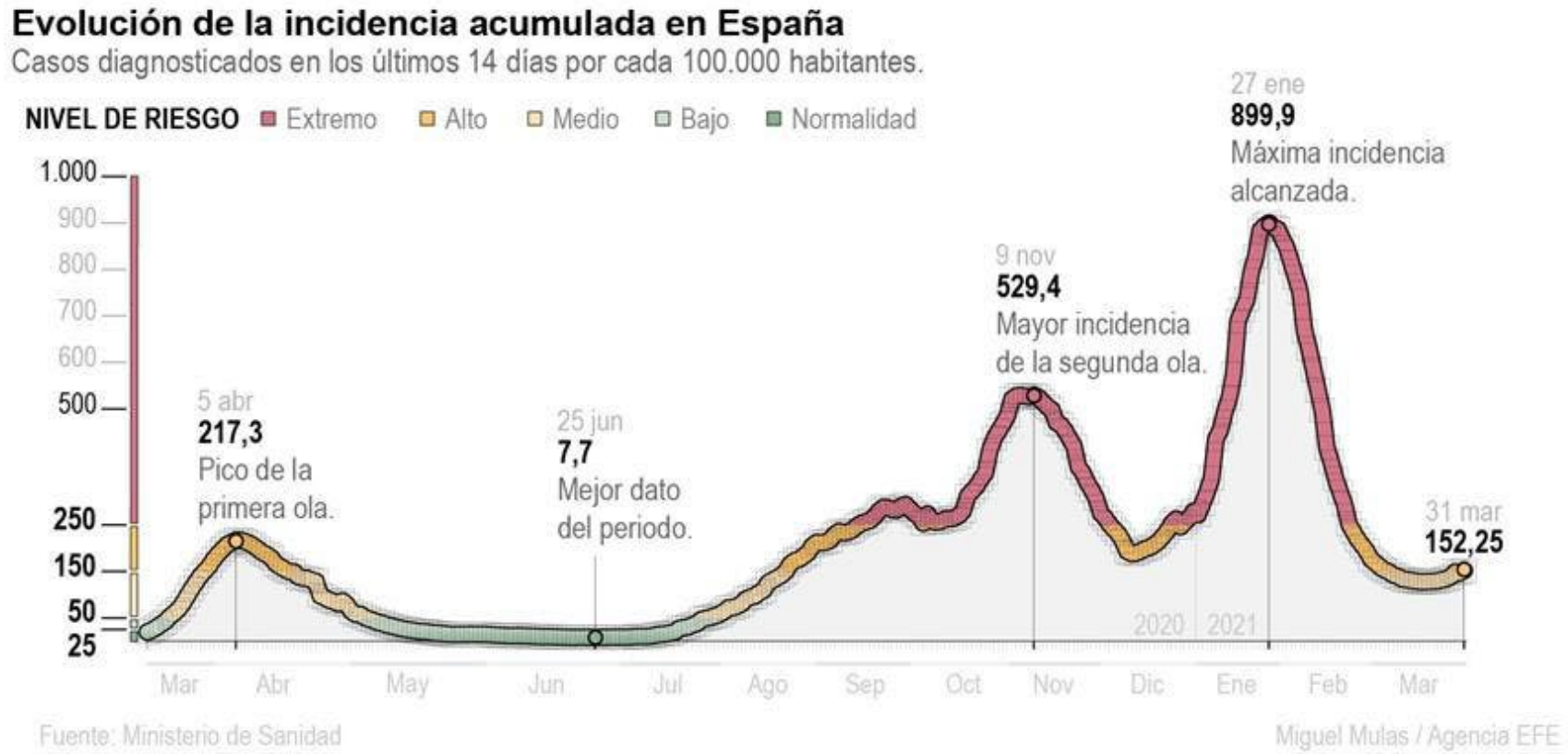
- Flipping a coin?
- Choosing Mahou instead of another beer brand?
- Going on Erasmus and getting a job?

Data generating process



- What the generating process?

Data generating process

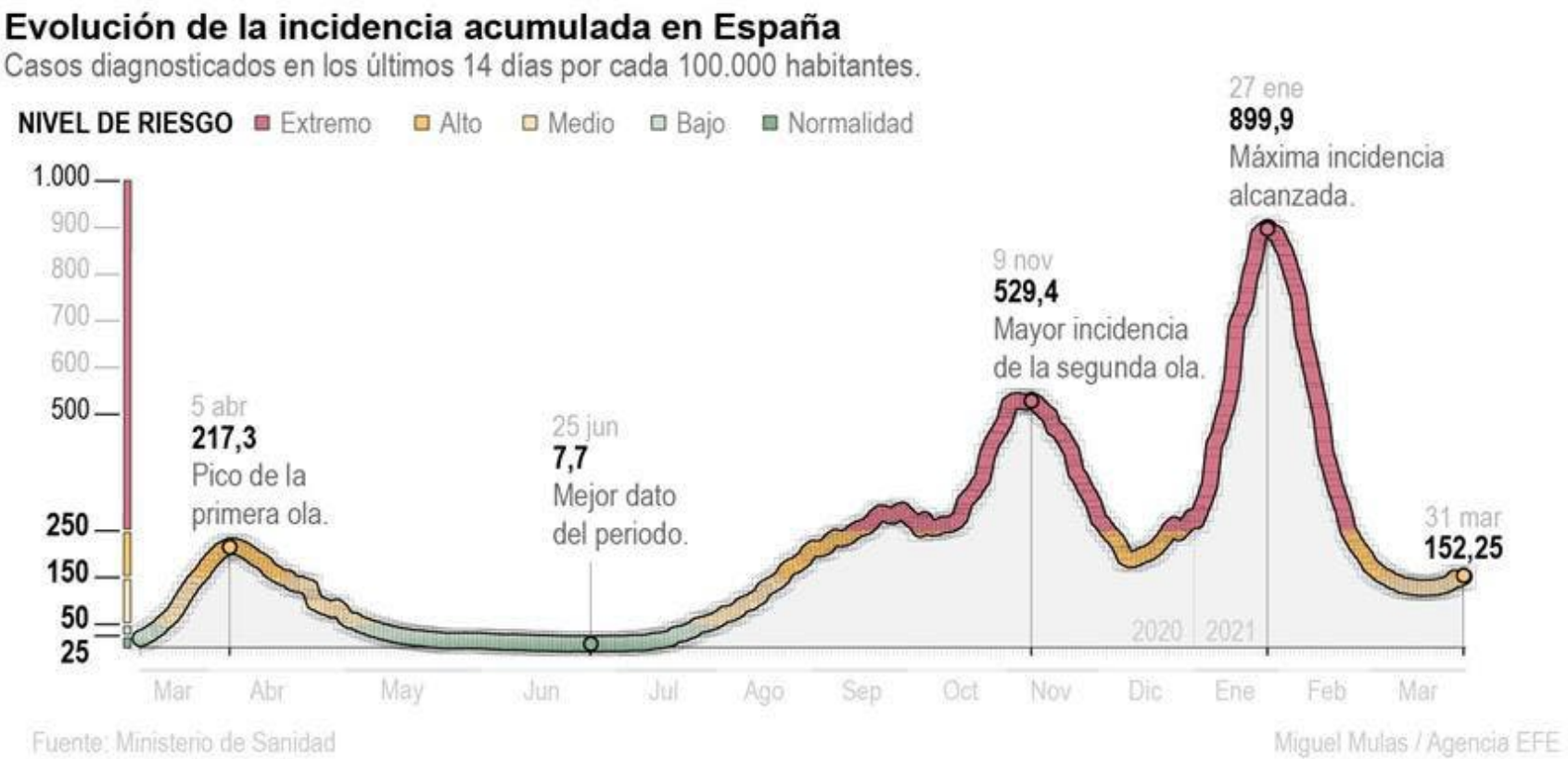


- What the generating process?
- How many variable are involved in generating this outcome?

Thinking about variables and processes/mechanisms

- Think we explain or explain with: variables
- Relationship between them: process
- The main idea is that if we compare different combinations of their values we are going to discover something about the process

Data generating process

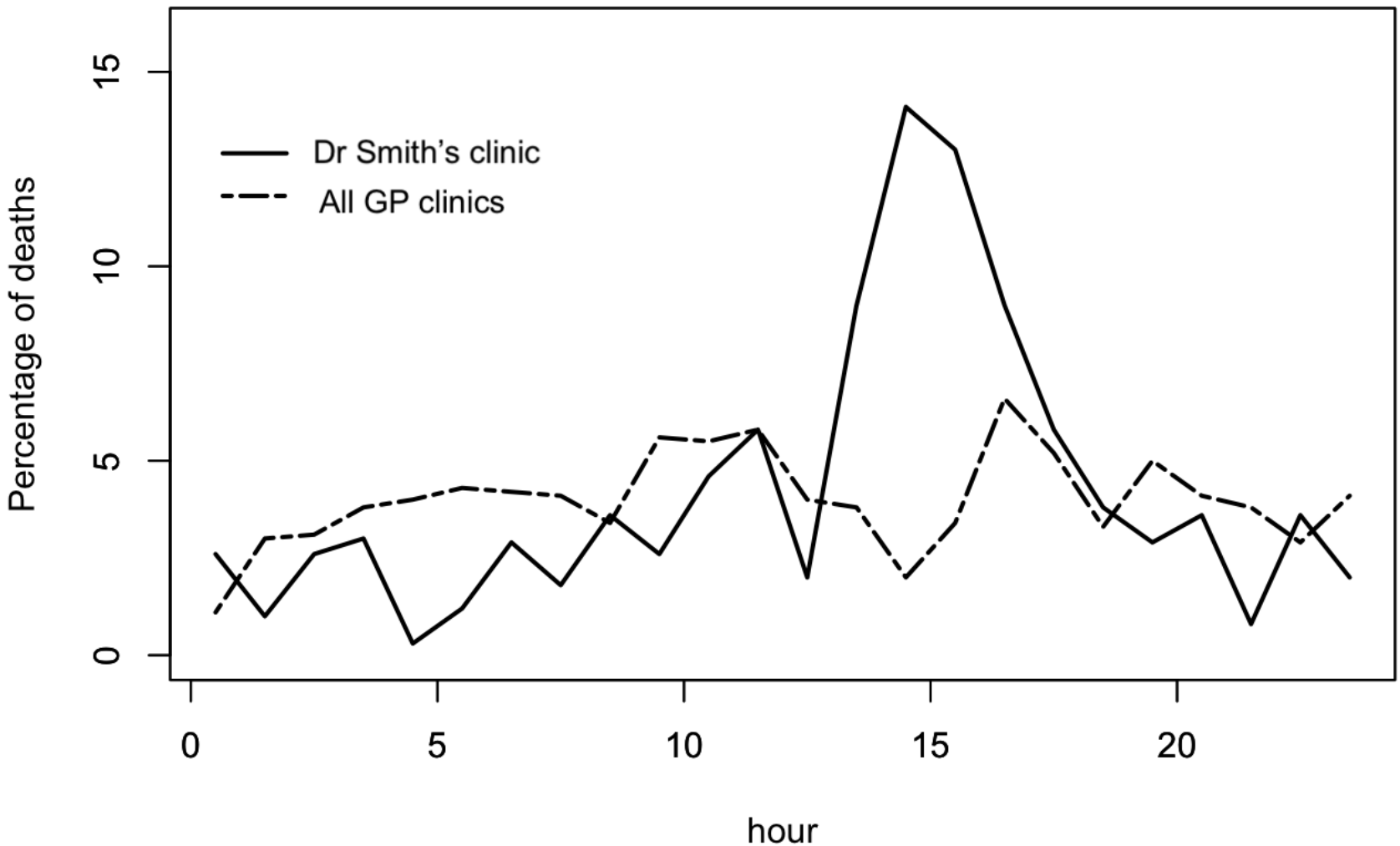


- How could we test different hypotheses? (schools, bars, weather, etc)

DGP and RQs

- Again, we want to state the research question as close as possible to that generating process
  - → What explains Covid incidence in Madrid?
  - → What explains Covid incidence in Madrid over time between June 2020 and June 2021
  - → Do school openings increase Covid incidence in Madrid?
  - → School openings and Covid incidence evolution in Madrid

Example


Dr Smith's clinic
All GP clinics
Percentage of deaths
0
5
10
15
0
5
10
15
20
hour

Roadmap

Theories and research questions

Concepts and operationalization

Measurement

Description

Prediction

Example: Wartime civilian deaths

Paper discussion

About prediction

- What is prediction?

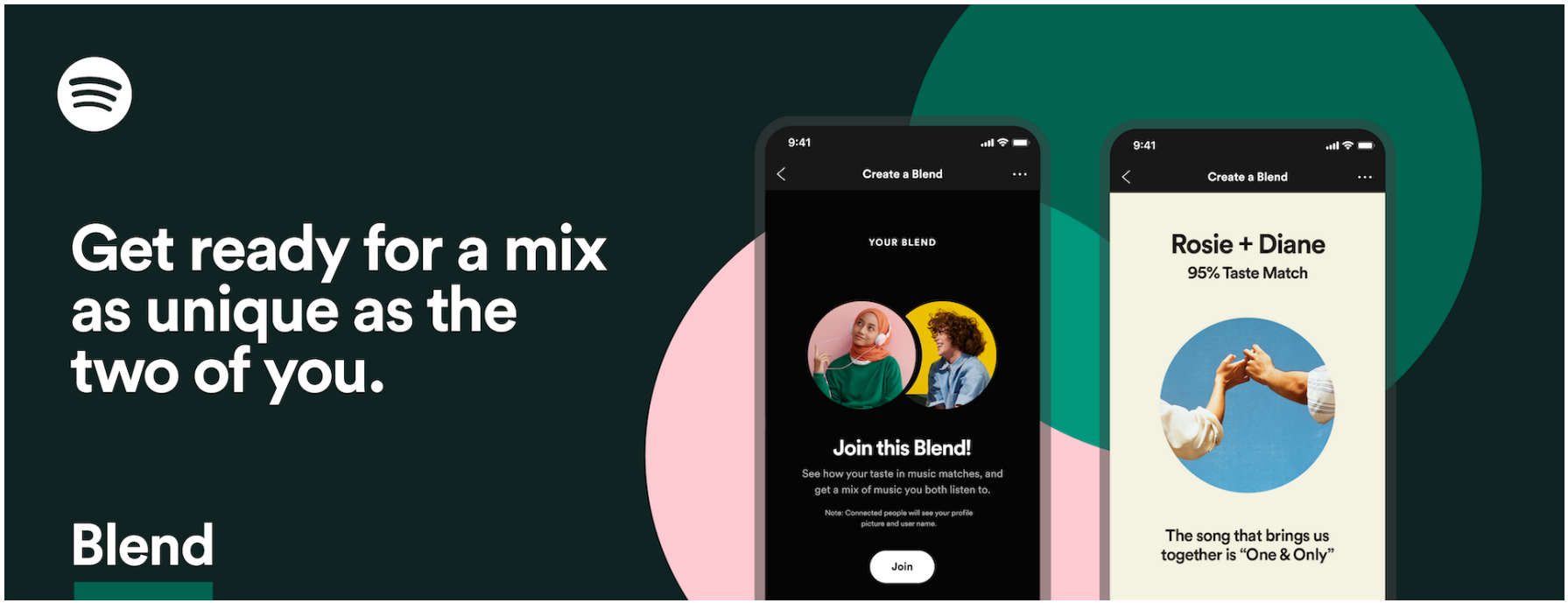
About prediction

- What is prediction?

The two concepts of prediction:

- Predicting another variable
- Predicting the future (or out of sample prediction)

About prediction



About prediction

TECH

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff

Welcome to The Not-So Private Parts where technology & privacy collide

Follow

Feb 16, 2012, 11:02am EST

About prediction

ECONOMICS

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1*] Gabriel Cadamuro,[2] Robert On[3]

Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer sources of big data exist. We show that an individual's past history of mobile phone use can be used to infer his or her socioeconomic status. Furthermore, we demonstrate that the predicted attributes of millions of individuals can, in turn, accurately reconstruct the distribution of wealth of an entire nation or to infer the asset distribution of microregions composed of just a few households. In resource-constrained environments where censuses and household surveys are rare, this approach creates an option for gathering localized and timely information at a fraction of the cost of traditional methods.

About prediction

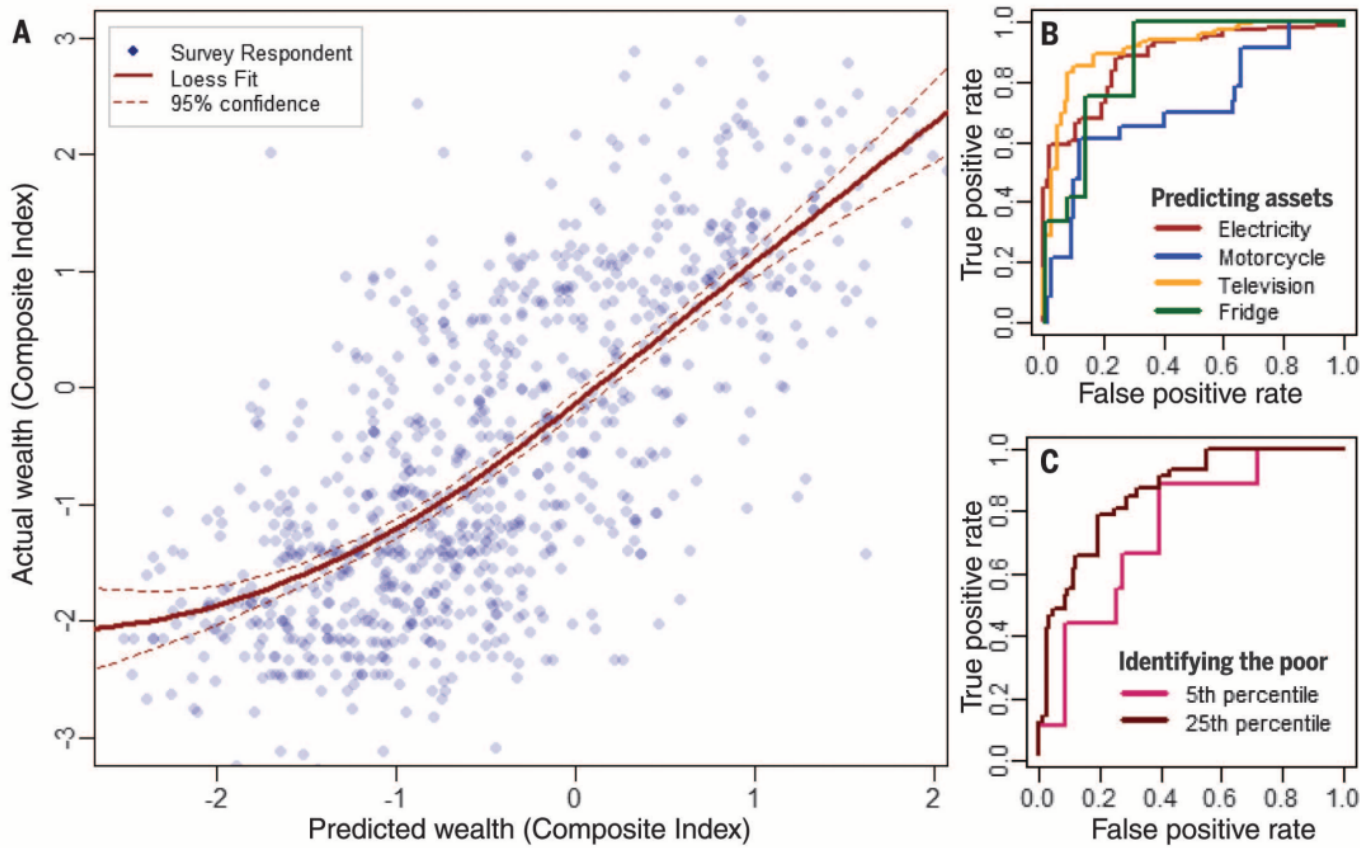


Fig. 1. Predicting survey responses with phone data. (**A**) Relation between actual wealth (as reported in a phone survey) and predicted wealth (as inferred from mobile phone data) for each of the 856 survey respondents. (**B**) Receiver operating characteristic (ROC) curve showing the model's ability to predict whether the respondent owns several different assets. AUC values for electricity, motorcycle, television, and fridge, respectively, are as follows: 0.85, 0.67, 0.84, and 0.88. (**C**) ROC curve illustrates the model's ability to correctly identify the poorest individuals. The poor are defined as those in the 5th percentile (AUC = 0.72) and the 25th percentile (AUC = 0.81) of the composite wealth index distribution.

About prediction

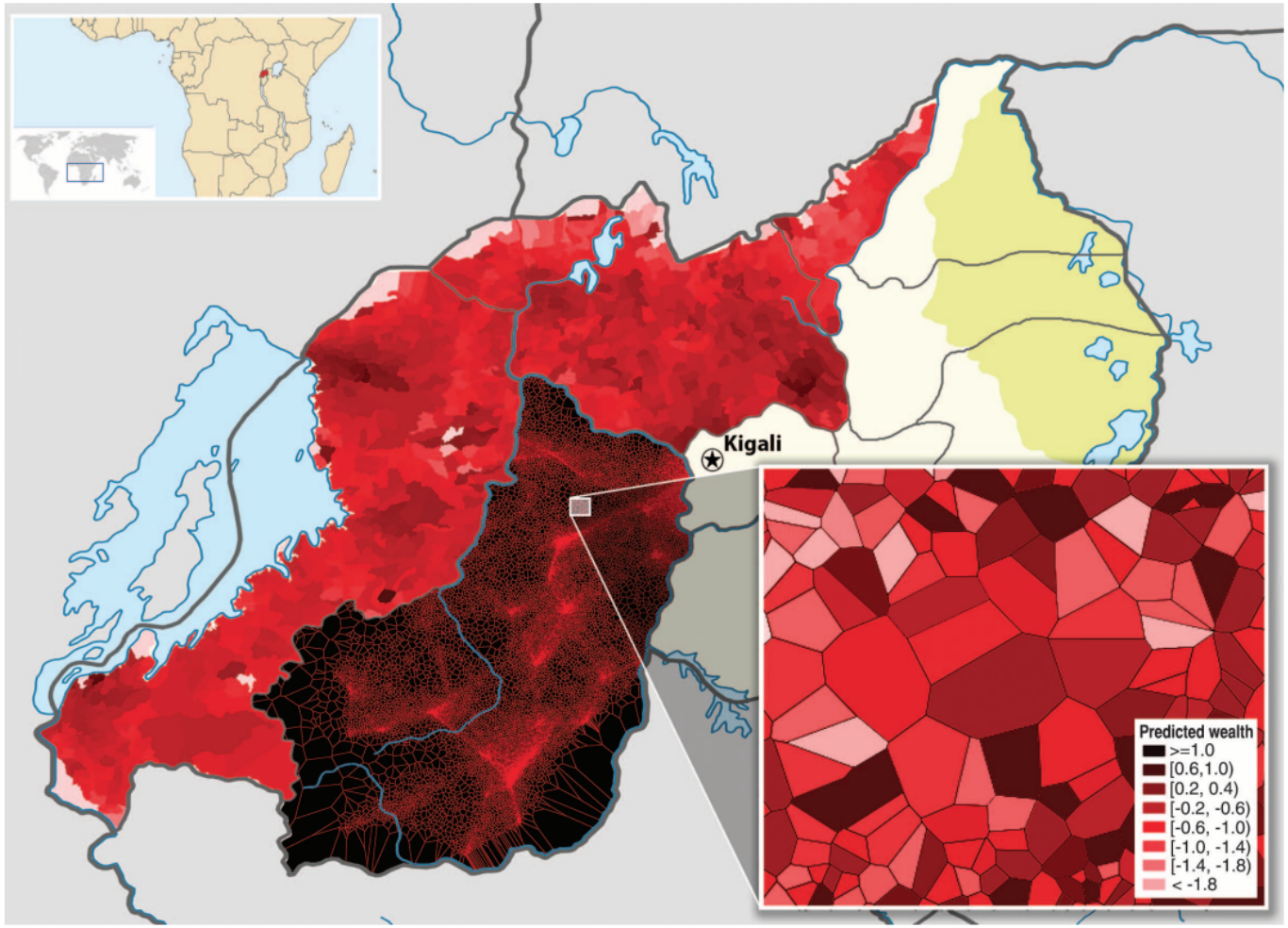


Fig. 2. Construction of high-resolution maps of poverty and wealth from call records. Information derived from the call records of 1.5 million subscribers is overlaid on a map of Rwanda. The northern and western provinces are divided into cells (the smallest administrative unit of the country), and the cell is shaded according to the average (predicted) wealth of all mobile subscribers in that cell. The southern province is overlaid with a Voronoi division that uses geographic identifiers in the call data to segment the region into several hundred thousand small partitions. (**Bottom right inset**) Enlargement of a 1-km^2 region near Kiyonza, with Voronoi cells shaded by the predicted wealth of small groups (5 to 15 subscribers) who live in each region.

About prediction

- Causality and prediction, is it the same?

About prediction

- Causality and prediction, is it the same?
- The importance of **counterfactuals** (more on this later)

Roadmap

Theories and research questions

Concepts and operationalization

Measurement

Description

Prediction

Example: Wartime civilian deaths

Paper discussion

Practical example

- You want to test an argument about **wartime civilian deaths:**
  - → The intuition you have is that civilians will be more likely to be treated well (and not killed) by rebel groups during civil wars when they need their resources (e.g. labor) to survive

- Clean up the theory, decide on the main concepts
- Develop different RQ at different levels
- How can we measure the main concepts? Variables?
- What answers could we get from the data?
  - → Are we learning something about our theory?

Roadmap

Theories and research questions

Concepts and operationalization

Measurement

Description

Prediction

Example: Wartime civilian deaths

Paper discussion

Journal of Conflict Resolution
2021, Vol. 65(2-3) 563-590

DOI: 10.1177/0022002720963674
journals.sagepub.com/home/jcr


Roads to Rule, Roads to Rebel: Relational State Capacity and Conflict in Africa

Carl Müller-Crepon[1], Philipp Hunziker[2], and Lars-Erik Cederman[3]

Abstract

Weak state capacity is one of the most important explanations of civil conflict. Yet, current conceptualizations of state capacity typically focus only on the state while ignoring the relational nature of armed conflict. We argue that opportunities for conflict arise where relational state capacity is low, that is, where the state has less control over its subjects than its potential challengers. This occurs in ethnic groups that are poorly accessible from the state capital, but are internally highly interconnected. To test this argument, we digitize detailed African road maps and convert them into a road atlas akin to Google Maps. We measure the accessibility and internal connectedness of groups via travel times obtained from this atlas and simulate road networks for an instrumental variable design. Our findings suggest that low relational state capacity increases the risk of armed conflict in Africa.