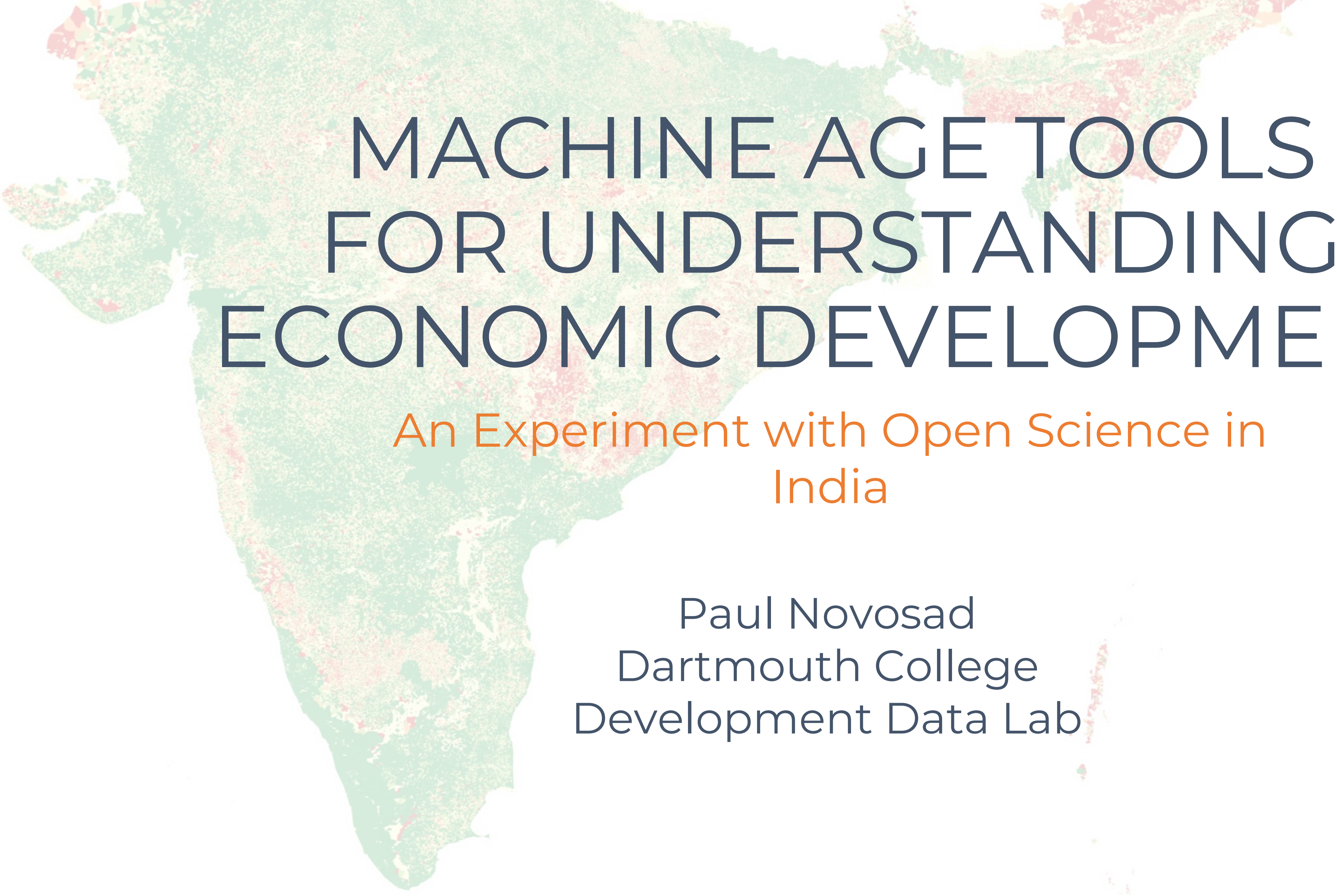


MACHINE AGE TOOLS FOR UNDERSTANDING ECONOMIC DEVELOPMENT

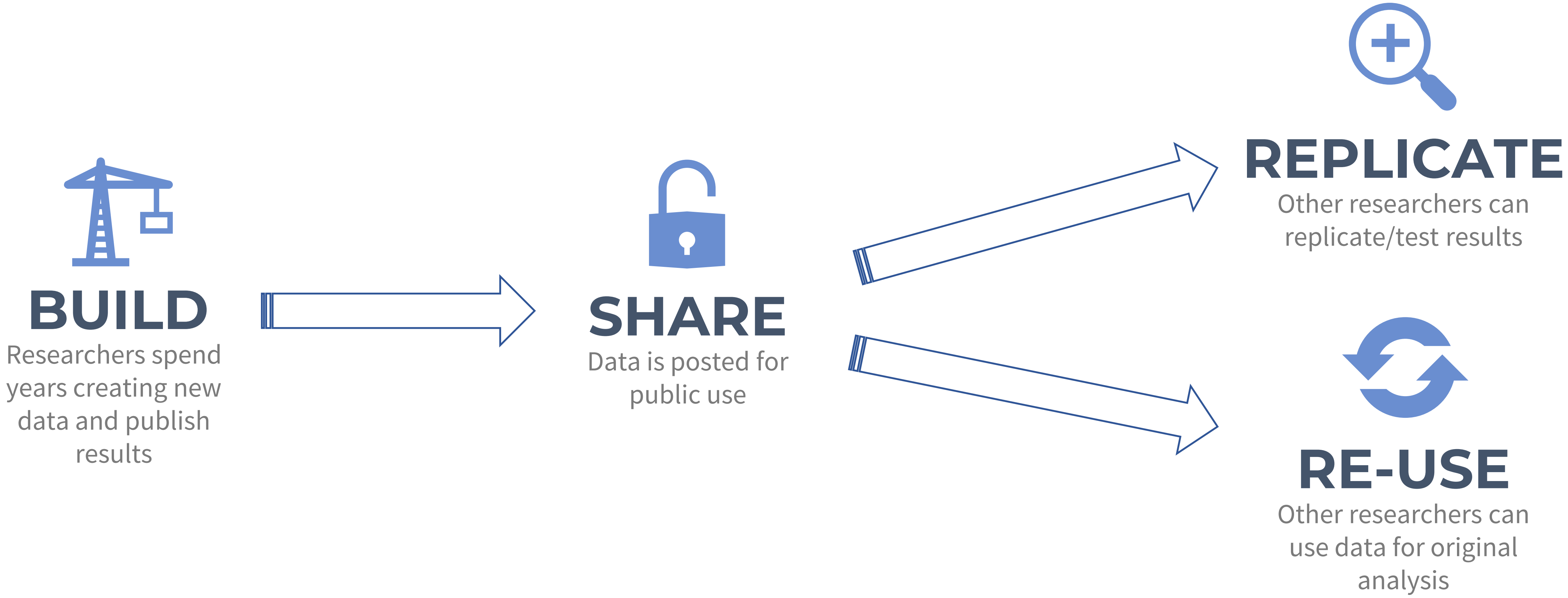
An Experiment with Open Science in India

Paul Novosad
Dartmouth College
Development Data Lab

Presentation Preview

- Tremendous unrealized scope for better data collaboration in the social sciences
- The SHRUG: A copyleft dataset and platform for research on India
- Extending the open-source software model to social science research

Open Science: In Theory



BUILD

Researchers spend years creating new data and publish results

SHARE

Data is posted for public use

REPLICATE

Other researchers can replicate/test results

RE-USE

Other researchers can use data for original analysis

Open Science: In Practice

Technical Barriers to Data Sharing

Creating usable public data ≠ just posting code and data

Institutional Barriers to Data Sharing

Keeping data private:

- Takes less work
- Lowers risk of failed replications
- Allows monopoly control of data for future projects

Open Science: In Practice

The result: public data is often of limited use to future researchers

- Posted code is impenetrable, shows final steps but no construction
- Posted datasets are messy and undocumented
- Posted data are often limited to project samples, limiting wider usability

Journal policies focused on replication are not solving the usability problem

Administrative Data
Raises the Returns to Openness

Socioeconomic research in developing countries is usually based on sample surveys

- Useful for aggregate statistics, but not for understanding local variation

Digital exhaust from government programs is barely used

- Universal digital multidimensional paper trail
- But: no documentation, unclear identifiers, survey manuals hard to find
  - (PII)
- Research value scales with the number of datasets
  - Chicken / egg problem: Isolated admin data is of limited use

Researchers in silos cannot mobilize this resource effectively

A New Idea for Data Collaboration

THE DATA BACKBONE

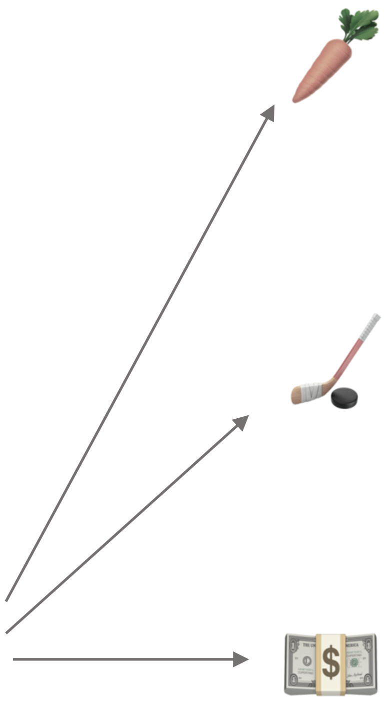
A comprehensive national socioeconomic dataset that is the best starting point for all research (on India)

EASY LINKING

Seamlessly links to all national datasets, so integration is almost costless

OPEN ARCHITECTURE

Lower both the technical and institutional barriers to sharing of data



🥕 **Citation:** Data maintains original reference. Contribution -> citations

🏒 **Copyleft licensing:** if you use, you must share what you link in a principled manner

💵 **Cost reduction:** Standardized data protocols lower cost of making data usable.

The SHRUG

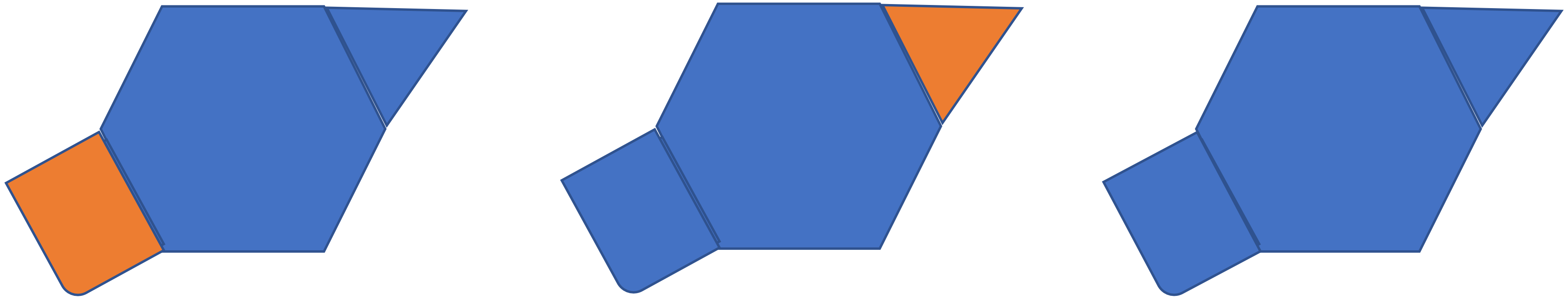
The Socioeconomic High-resolution Rural Urban Geographic data platform for India

SHRUG: The Location Backbone

The backbone is a set of universal locations

- Indian Census locations have new (hard to link) identifiers every 10 years.
- SHRUG has universal identifiers – time series analysis is a cinch.
- We provide simple keys to link SHRUG to any major national Indian dataset.
- Consistent industries, variables definitions, data structure, etc..

- Locations are amalgamated to create the smallest consistent unit:

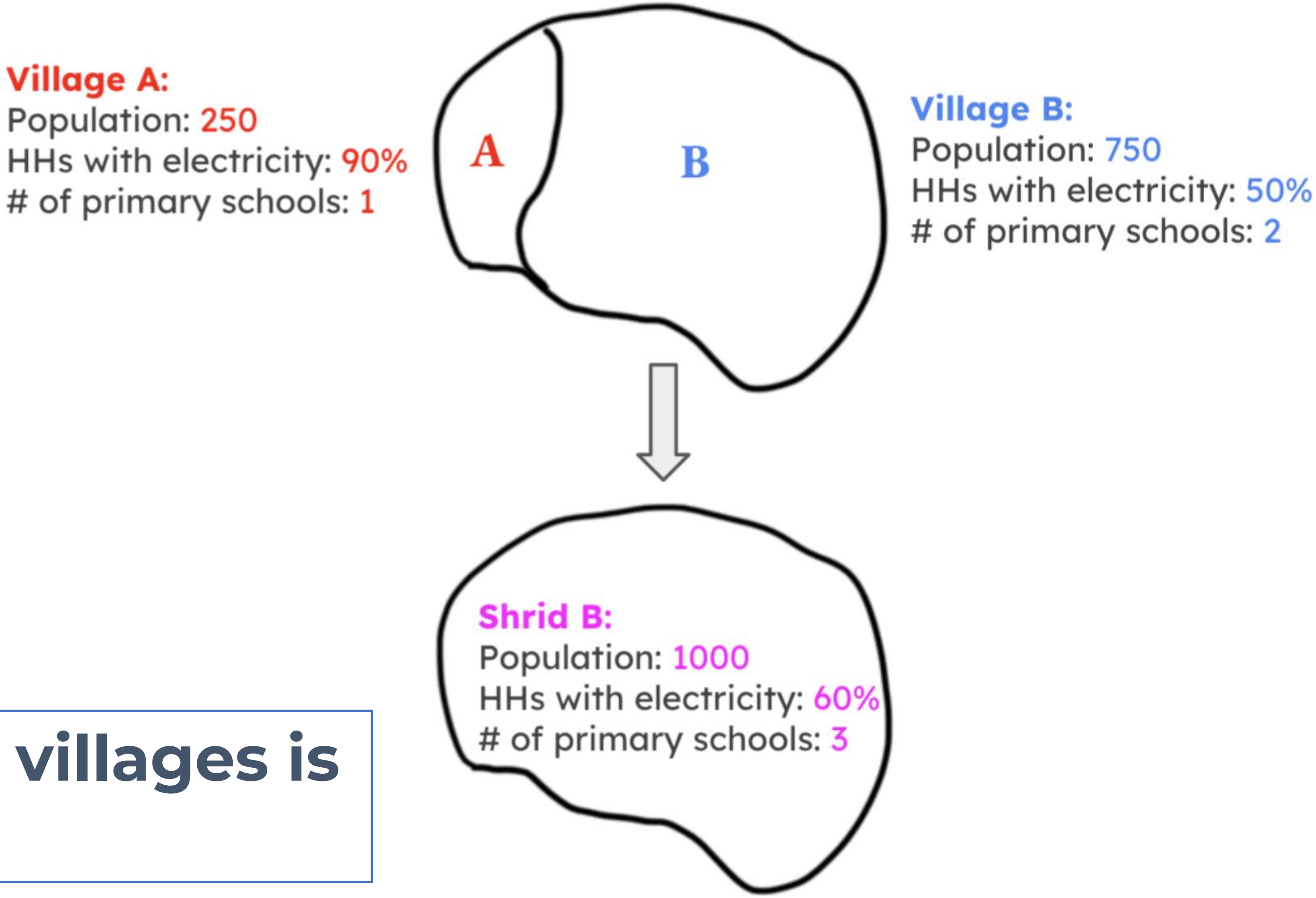


2001 Census Boundaries

2011 Census Boundaries

SHRUG Boundaries (all years)

Building variables in the SHRUG



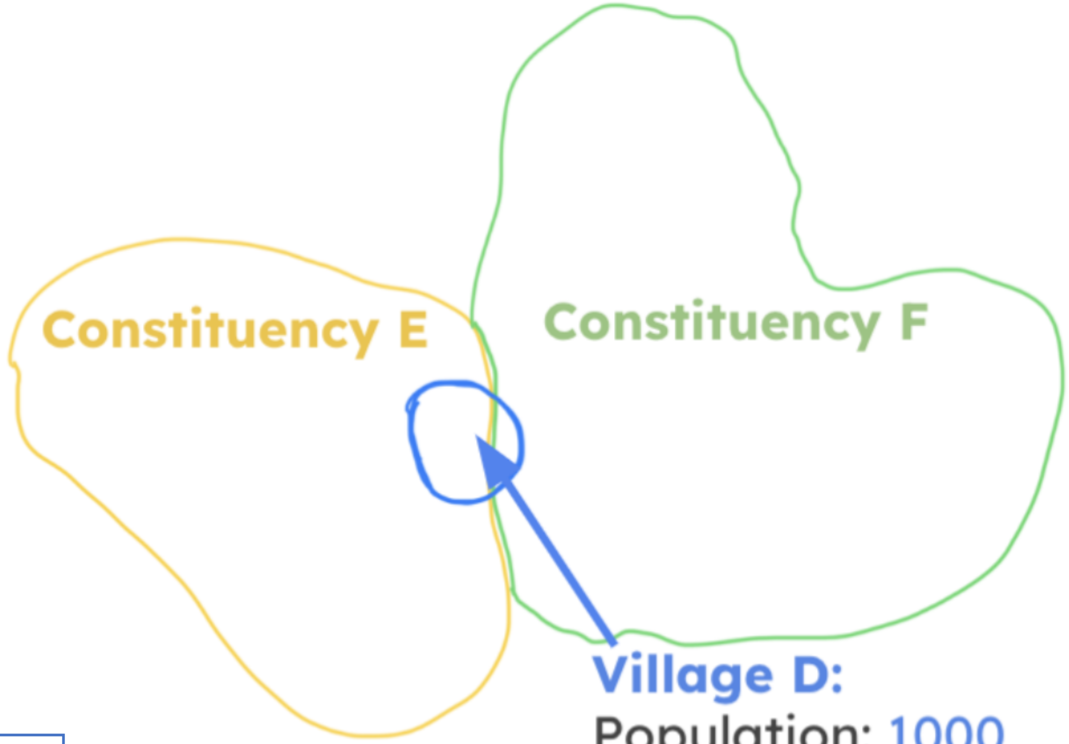
Village A:
Population: 250
HHs with electricity: 90%
# of primary schools: 1

Village B:
Population: 750
HHs with electricity: 50%
# of primary schools: 2

Shrid B:
Population: 1000
HHs with electricity: 60%
# of primary schools: 3

Combining villages is easy.

Building variables in the SHRUG



Village D:

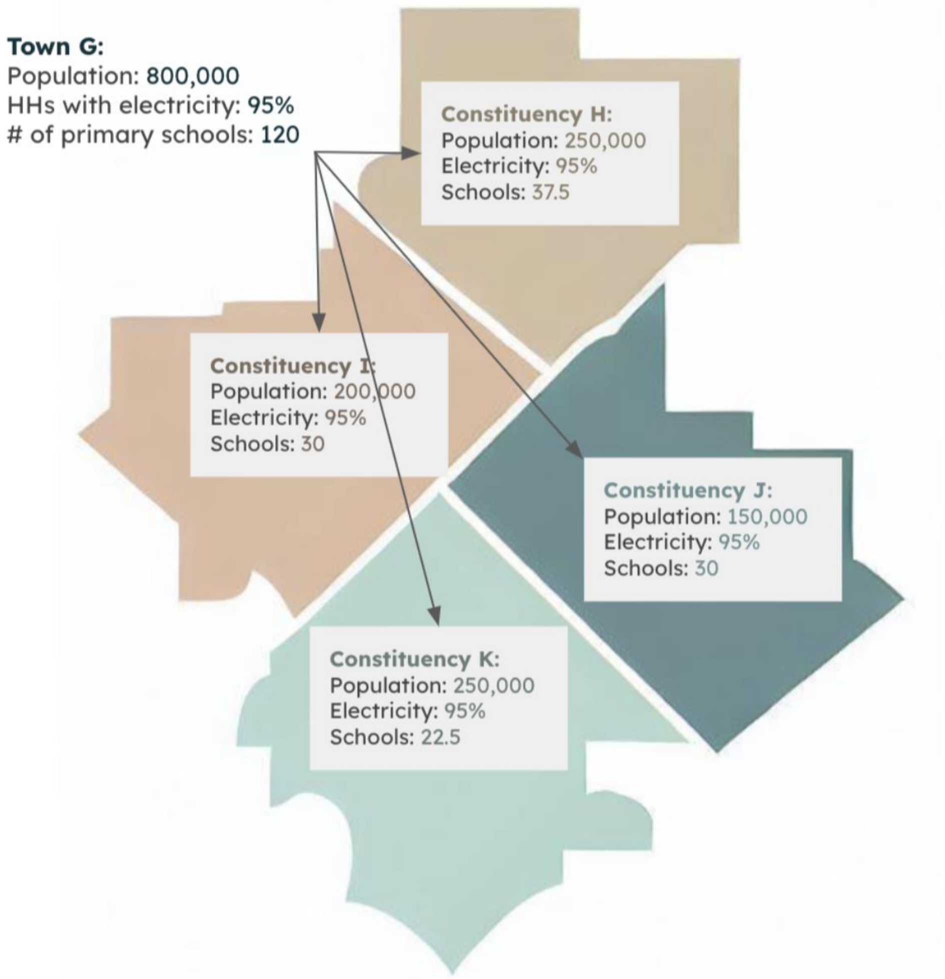
Population: 1000

HHs with electricity: 60%

# of primary schools: 3

Constituencies require some imputation.

Building variables in the SHRUG



Town G:
Population: 800,000
HHs with electricity: 95%
# of primary schools: 120

Constituency H:
Population: 250,000
Electricity: 95%
Schools: 37.5

Constituency I:
Population: 200,000
Electricity: 95%
Schools: 30

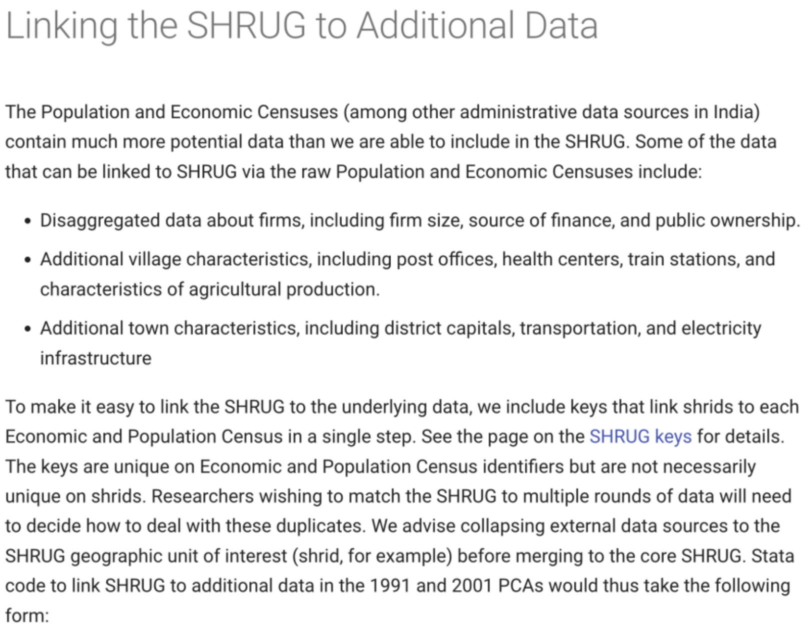
Constituency J:
Population: 150,000
Electricity: 95%
Schools: 30

Constituency K:
Population: 250,000
Electricity: 95%
Schools: 22.5

Urban constituencies are especially hard.

Usage advice: RTFM

docs.devdatalab.org

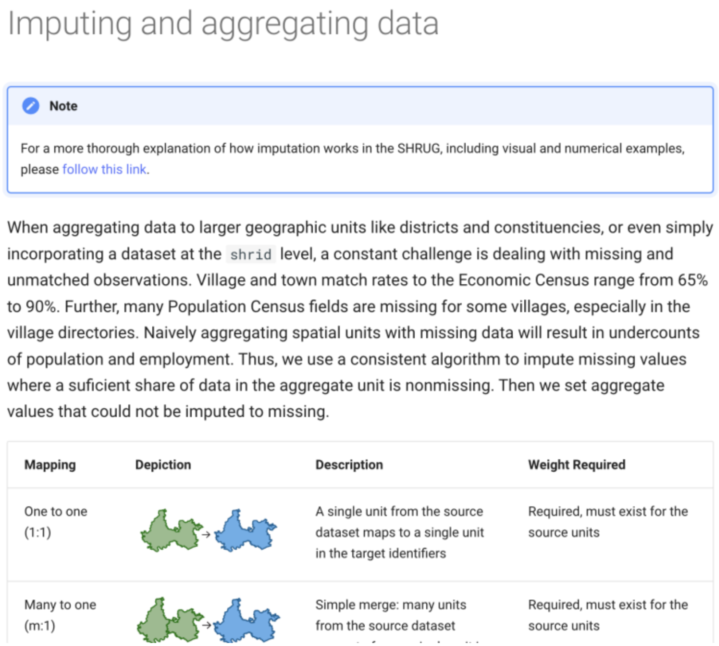


Linking the SHRUG to Additional Data

The Population and Economic Censuses (among other administrative data sources in India) contain much more potential data than we are able to include in the SHRUG. Some of the data that can be linked to SHRUG via the raw Population and Economic Censuses include:

- Disaggregated data about firms, including firm size, source of finance, and public ownership.
- Additional village characteristics, including post offices, health centers, train stations, and characteristics of agricultural production.
- Additional town characteristics, including district capitals, transportation, and electricity infrastructure

To make it easy to link the SHRUG to the underlying data, we include keys that link shrids to each Economic and Population Census in a single step. See the page on the SHRUG keys for details. The keys are unique on Economic and Population Census identifiers but are not necessarily unique on shrids. Researchers wishing to match the SHRUG to multiple rounds of data will need to decide how to deal with these duplicates. We advise collapsing external data sources to the SHRUG geographic unit of interest (shrid, for example) before merging to the core SHRUG. Stata code to link SHRUG to additional data in the 1991 and 2001 PCAs would thus take the following form:



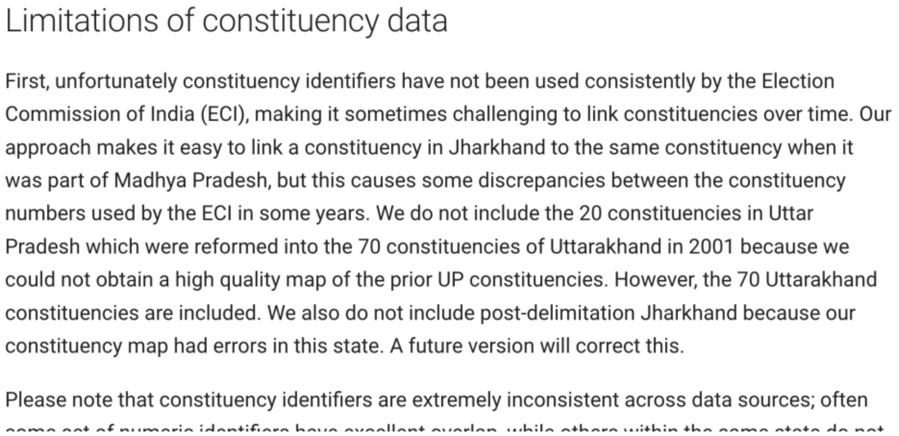
Imputing and aggregating data

Note

For a more thorough explanation of how imputation works in the SHRUG, including visual and numerical examples, please follow this link.

When aggregating data to larger geographic units like districts and constituencies, or even simply incorporating a dataset at the `shrid` level, a constant challenge is dealing with missing and unmatched observations. Village and town match rates to the Economic Census range from 65% to 90%. Further, many Population Census fields are missing for some villages, especially in the village directories. Naively aggregating spatial units with missing data will result in undercounts of population and employment. Thus, we use a consistent algorithm to impute missing values where a suficient share of data in the aggregate unit is nonmissing. Then we set aggregate values that could not be imputed to missing.

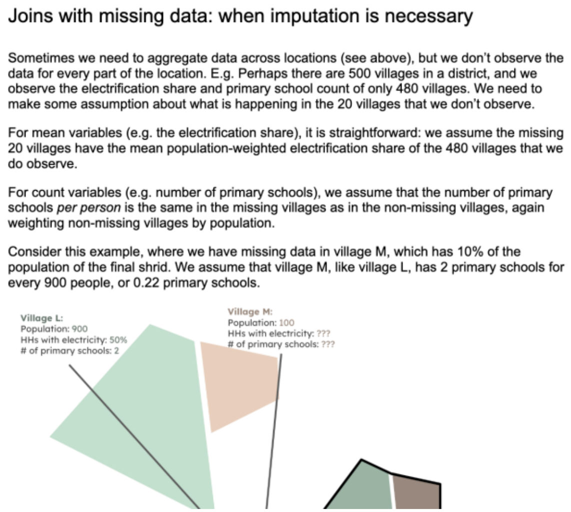
Mapping	Depiction	Description	Weight Required
One to one (1:1)	→	A single unit from the source dataset maps to a single unit in the target identifiers	Required, must exist for the source units
Many to one (m:1)	→	Simple merge: many units from the source dataset	Required, must exist for the source units



Limitations of constituency data

First, unfortunately constituency identifiers have not been used consistently by the Election Commission of India (ECI), making it sometimes challenging to link constituencies over time. Our approach makes it easy to link a constituency in Jharkhand to the same constituency when it was part of Madhya Pradesh, but this causes some discrepancies between the constituency numbers used by the ECI in some years. We do not include the 20 constituencies in Uttar Pradesh which were reformed into the 70 constituencies of Uttarakhand in 2001 because we could not obtain a high quality map of the prior UP constituencies. However, the 70 Uttarakhand constituencies are included. We also do not include post-delimitation Jharkhand because our constituency map had errors in this state. A future version will correct this.

Please note that constituency identifiers are extremely inconsistent across data sources; often some set of numeric identifiers have excellent overlap, while others within the same state do not



Joins with missing data: when imputation is necessary

Sometimes we need to aggregate data across locations (see above), but we don't observe the data for every part of the location. E.g. Perhaps there are 500 villages in a district, and we observe the electrification share and primary school count of only 480 villages. We need to make some assumption about what is happening in the 20 villages that we don't observe.

For mean variables (e.g. the electrification share), it is straightforward: we assume the missing 20 villages have the mean population-weighted electrification share of the 480 villages that we do observe.

For count variables (e.g. number of primary schools), we assume that the number of primary schools *per person* is the same in the missing villages as in the non-missing villages, again weighting non-missing villages by population.

Consider this example, where we have missing data in village M, which has 10% of the population of the final shrid. We assume that village M, like village L, has 2 primary schools for every 900 people, or 0.22 primary schools.

Village L:
Population: 900
HHs with electricity: 50%
# of primary schools: 2
Village M:
Population: 100
HHs with electricity: ???
# of primary schools: ???

SHRUG 2.0

Some highlights:

- Economic Census: every firm, every village, 1990–2013
- Antyodaya: Some of the only post-2015 village data
- SECC: Village- and town-level consumption
- Air pollution
- Elections and Affidavits
- Town/Village polygons

Coming soon:

- Rainfall & Temperature
- Intergenerational Mobility
- Cities: segregation, inequality, neighborhood services

Use Cases for the SHRUG

Studying Local Development

Most variation in socioeconomic status and in policy is local.

Baseline Data for RCTs

Plug a village list into the SHRUG and get 30 years of multidimensional data.

Cities

SHRUG is the first broad dataset that identifies the full set of towns and cities.

Media / Civil Society

Journalists / citizens are hungry for data but lack resources to build it themselves.

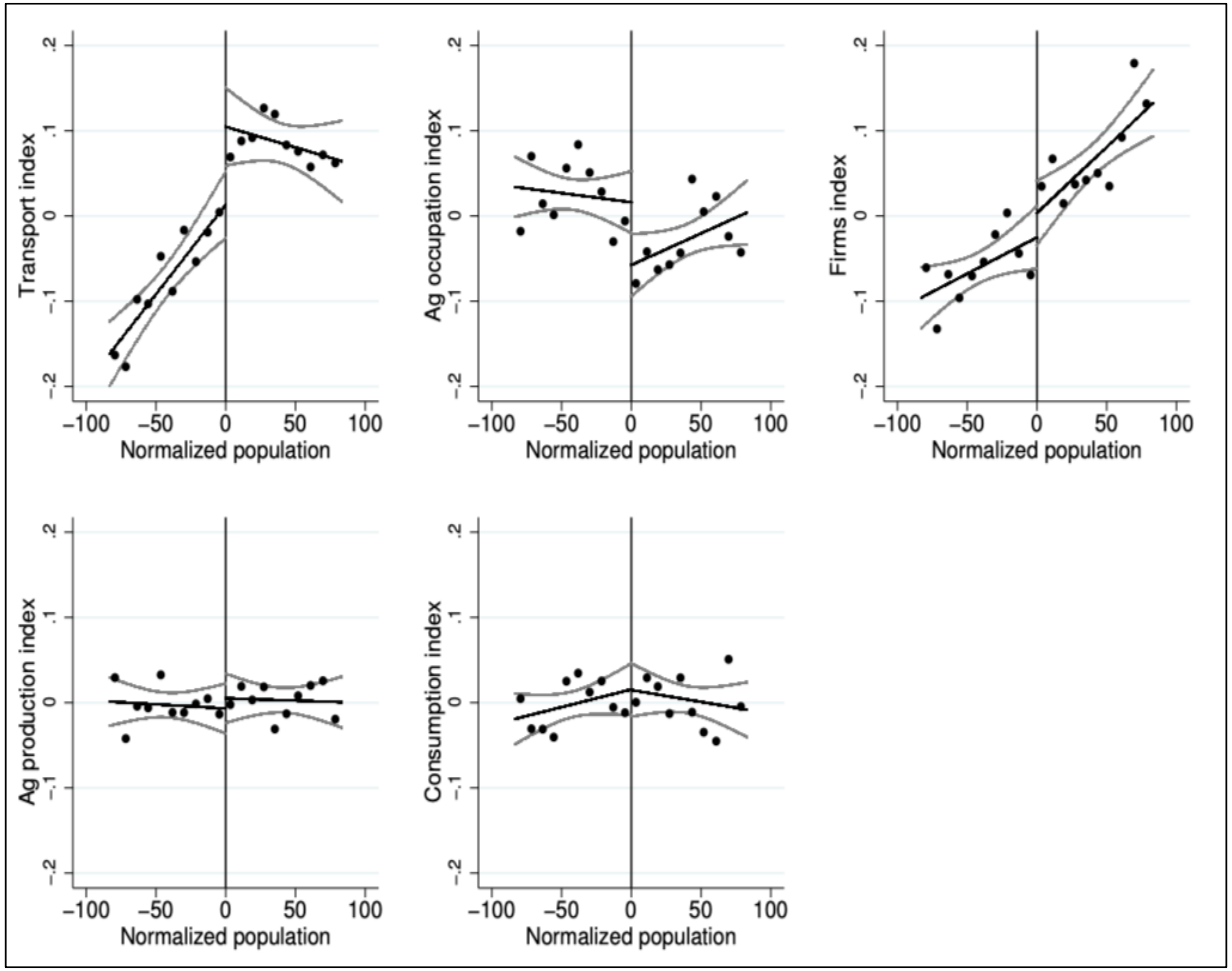
Example 1: Rural Roads

What are the impacts of India's large-scale rural roads program?

- 100,000 new village roads were built from 2000-15
- District-level (old data) approach:
  - Districts building more roads were way better off
  - (But correlation ≠ causation, and effective districts built more roads)

- We use village variation and RD to measure causal impacts. Findings:
  - New roads did not affect consumption, entrepreneurship, investment, or agriculture
  - They did help people get jobs outside of villages

- Required national data on a broad set of village outcomes
  - Very hard to do without administrative data



Asher and Novosad, "Rural Roads and Local Economic Development," AER (forthcoming)

Example 2: Impacts of Mines

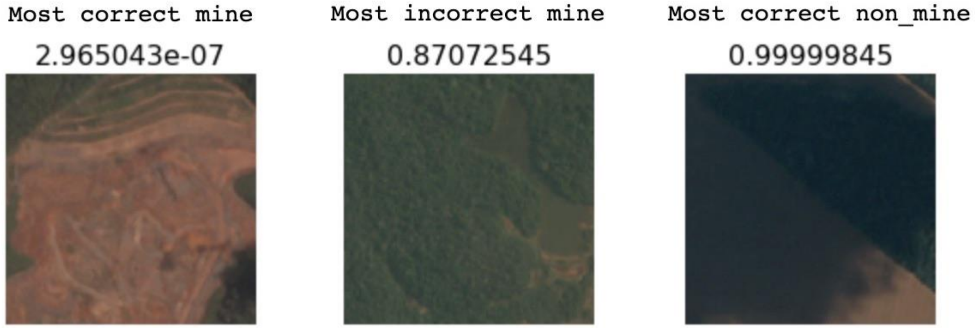
How does mineral extraction affect local opportunity?

- India has many mines but their impacts are highly local

- Few districts depend on mining: aggregate approach misses local effects

- We want to know:
  - How are the villages directly in the path of mining development affected?
  - Care about a wide range of outcomes: education, consumption, work, health, pollution

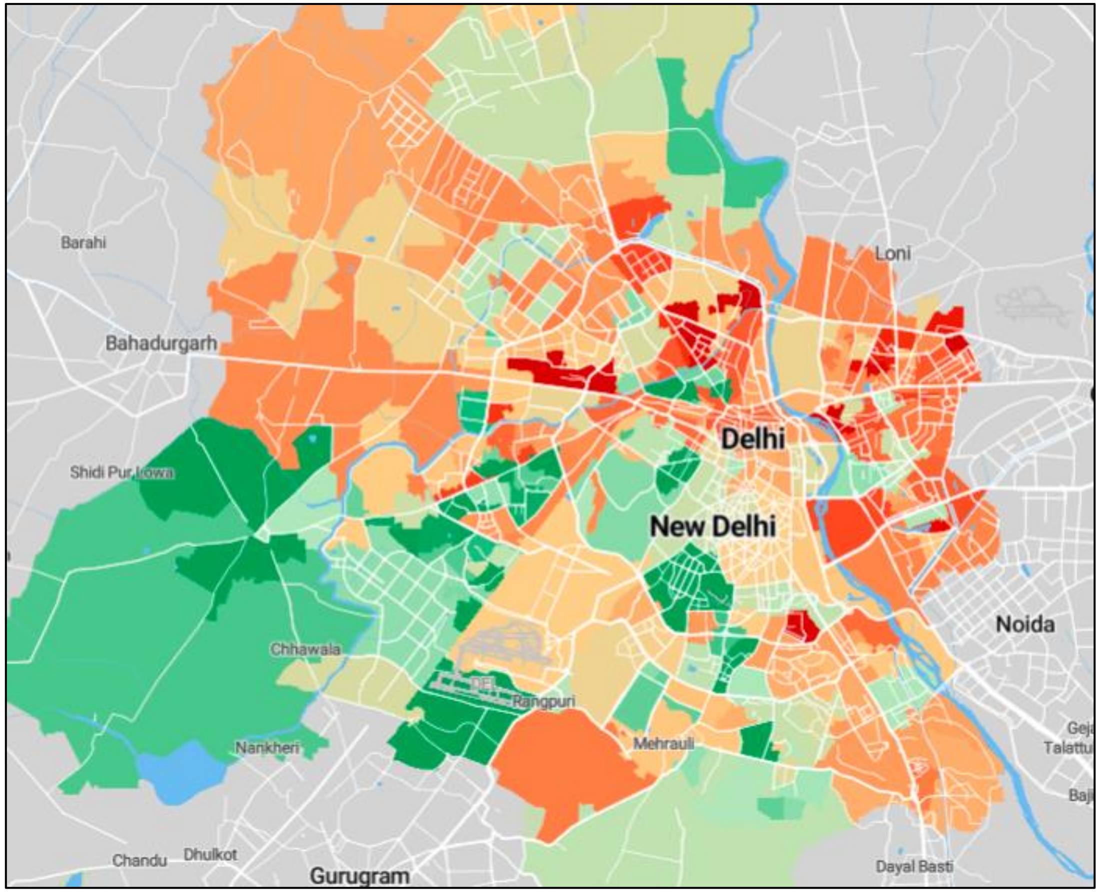
- Approach:
  - Computer vision and satellite data detect mine location and expansion
  - International prices generate exogenous variation in mine growth -> causality

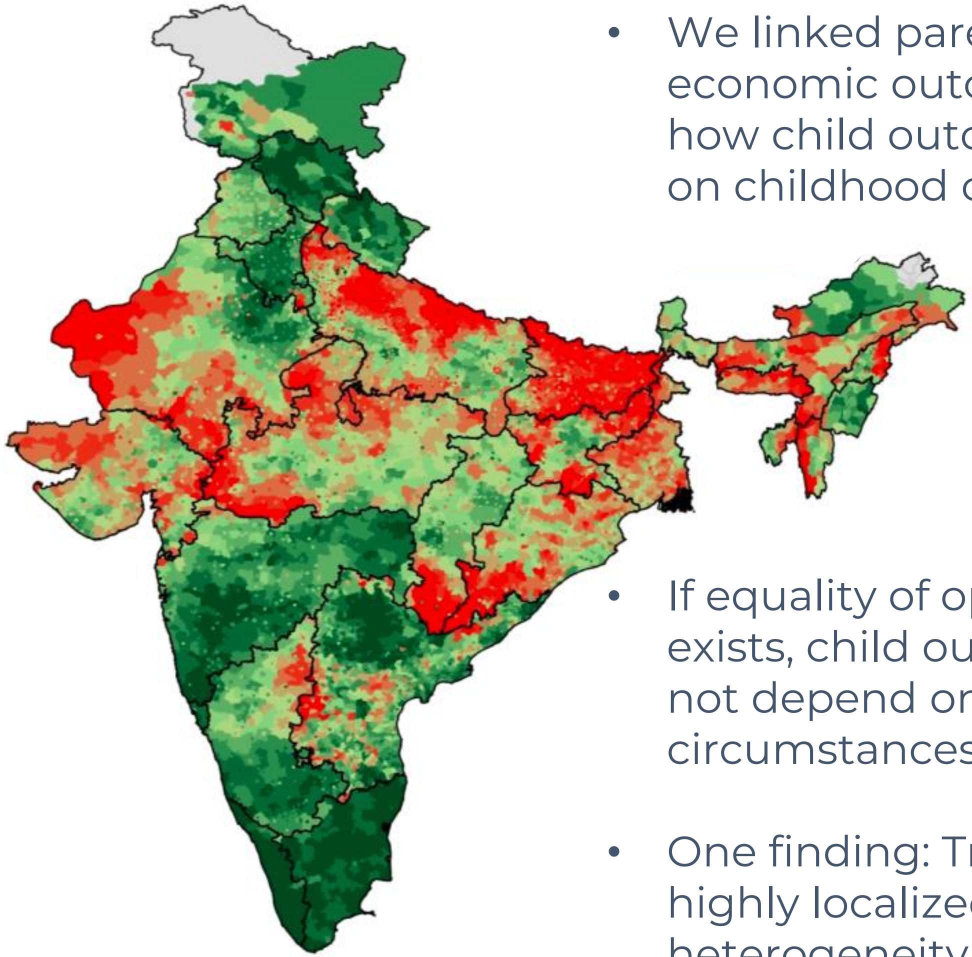


Asher, Lunt, and Novosad, “Digging for Development: Economic Impacts of Mining Booms”

Example 3: Equality of Opportunity

What is the geography of upward mobility in India?





- We linked parent to child economic outcomes to study how child outcomes depend on childhood circumstances.

- If equality of opportunity exists, child outcomes should not depend on birth circumstances.

- One finding: Tremendous highly localized geographic heterogeneity

SHRUG: Returns to Scale

- The SHRUG is fully open
- We are continuing to build this data platform
- But to achieve maximal scale, we need to mobilize the crowd

1. Rewards for Contributors

Contributing to SHRUG helps your research get found

Posting your own data is great, but very slow for others to find, evaluate, and link.

SHRUG-connected data has a high-quality standard and is immediately linked to dozens of other data sources.

SHRUG is structured to maintain attribution

- Use three components → cite three papers
- Downloads automatically generate citation files
- Repeated nudges to eliminate accidental omissions

2. Copyleft Licensing

🚨 **If you use SHRUG, you publish your data with SHRUG standards**

ODbL-based license requires derivative products to be released with same license at time of publication.

🔓 **Modeled on the Gnu Public License, a copyleft license for software that undergirds the open-source software movement.**

Like a time-limited patent, the license trades off the scientist’s interest in not getting scooped and the public interest of having open data.

3. We Will Help You

Releasing highly usable data takes a lot of additional work.

Research teams aren't rewarded for this work and may not have the capacity to release highly usable products.

We work with teams generating high value national data to help them normalize and integrate it with SHRUG.

Committing funds to this phase ex ante could further improve quantity and quality of data sharing in equilibrium.

A Vision for Data Collaboration

A health researcher is working with state-wide medical claims data.

By linking the data to SHRUG, she can study the highly local social determinants of health.

When she publishes, she also publishes village-level aggregates describing health outcomes with SHRUG identifiers.

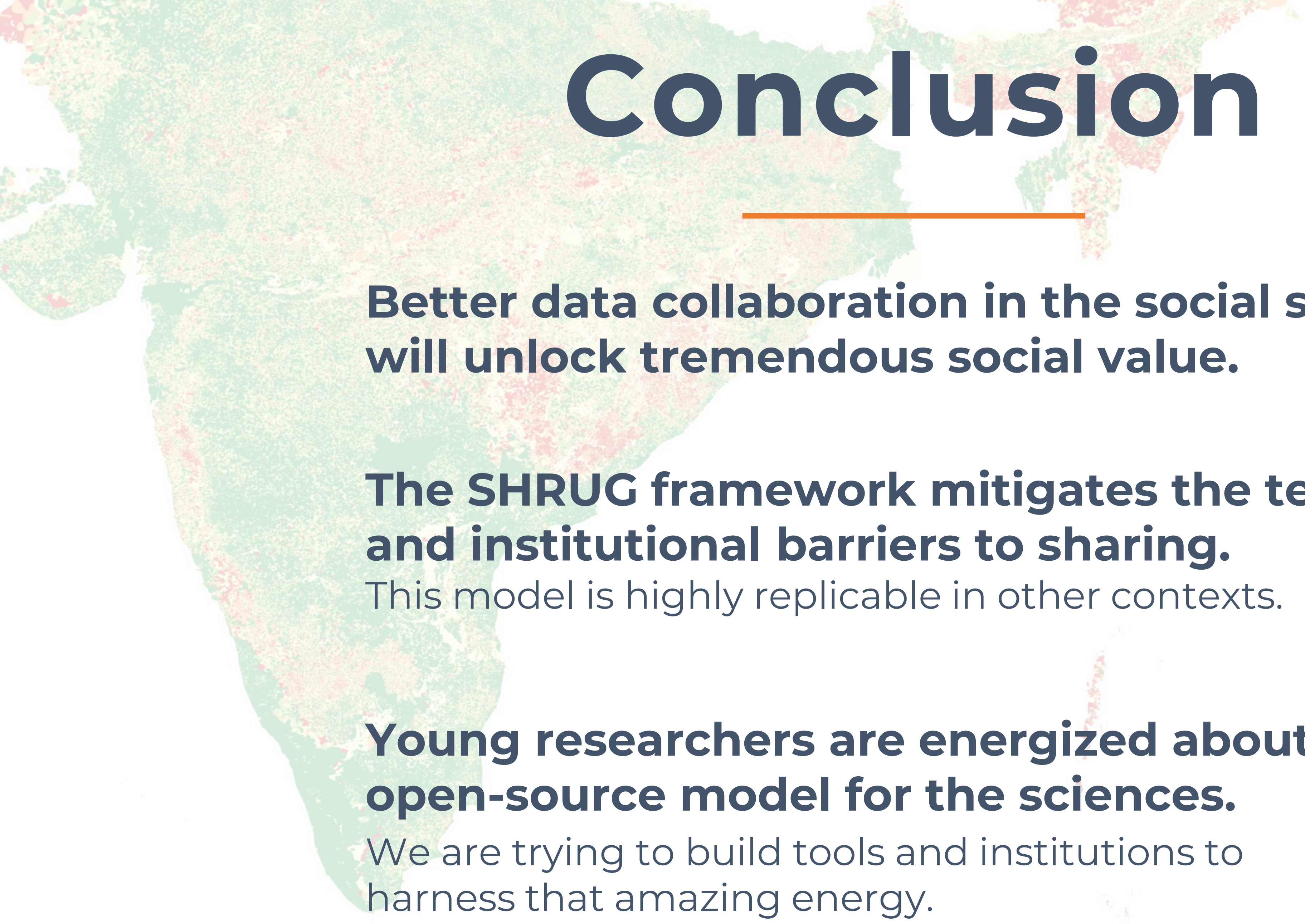
Health module is now available to future users of the SHRUG, enabling dozens of additional studies.

Scale this process by all the researchers working with administrative data in India

Next Steps for Shrug

- API: Direct access in R/Stata/Python
- Working with governments / firms to expand data availability
- Making contribution seamless

Conclusion



Better data collaboration in the social sciences will unlock tremendous social value.

The SHRUG framework mitigates the technical and institutional barriers to sharing.
This model is highly replicable in other contexts.

Young researchers are energized about an open-source model for the sciences.
We are trying to build tools and institutions to harness that amazing energy.