TERRA POPULUS

Integrated Data on Population and Environment

IEEE Big Data Conference, November 1, 2015

Minnesota Population Center

University of Minnesota LIBRARIES

INSTITUTE ON THE ENVIRONMENT
University of Minnesota
Driven to Discover℠

CIESIN
Columbia University

ICPSR
A PARTNER IN SOCIAL SCIENCE RESEARCH
INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

NSF

Background

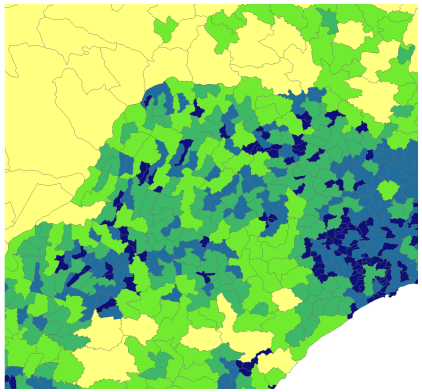
- National Science Foundation: Cyber Infrastructure Grant
- Data Net Initiative

- Enabling research, learning, and policy analysis by providing integrated spatiotemporal data describing people and their environment.

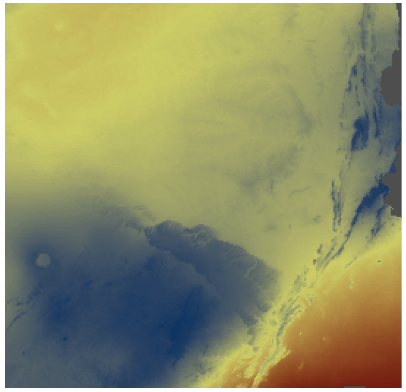
Big Heterogeneous Data

TerraPop Data Formats

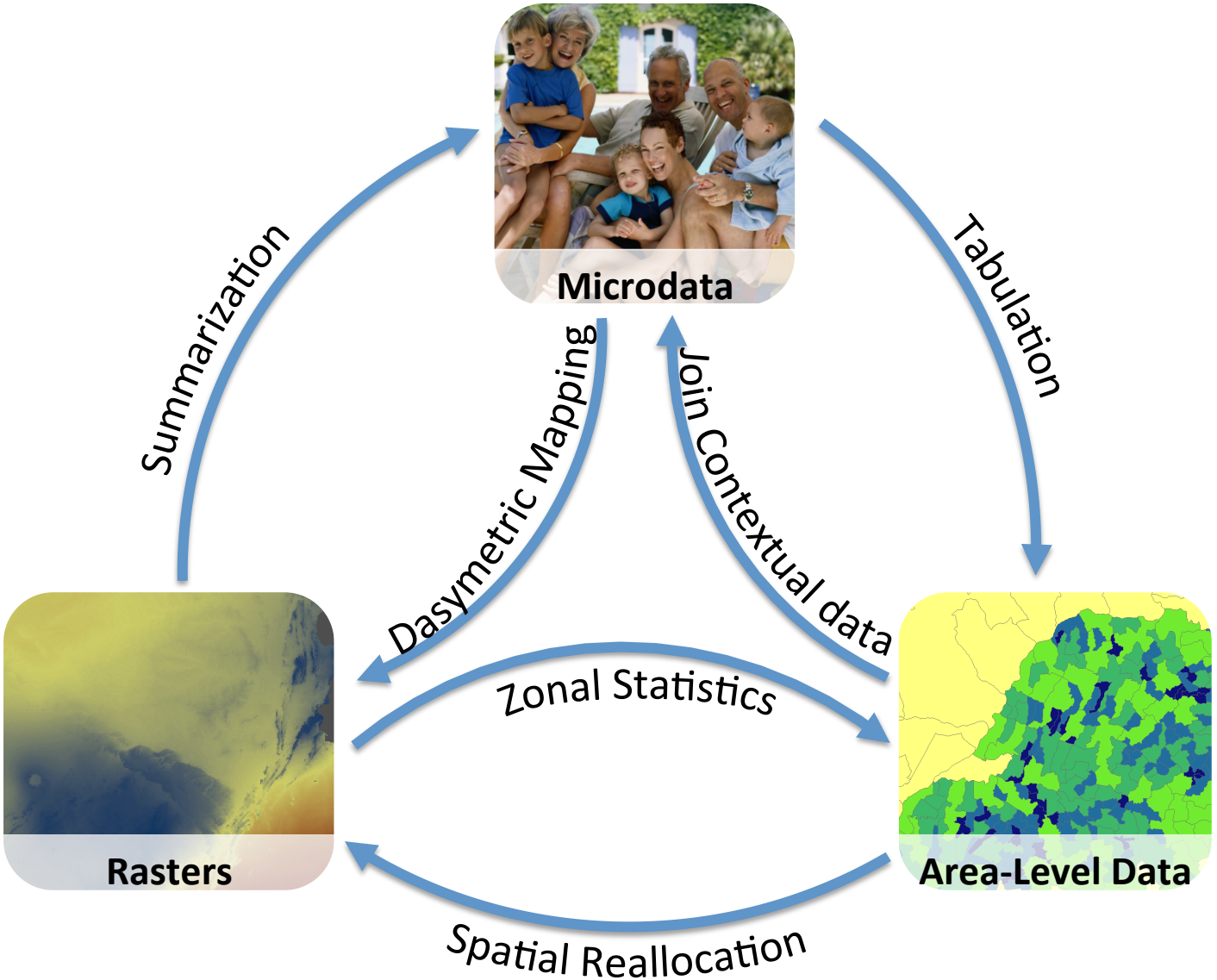
Microdata:
Characteristics of individuals and households



Area-level data:
Characteristics of places defined by boundaries



Raster data:
Values tied to spatial coordinates


Microdata
Tabulation
Area-Level Data
Spatial Reallocation
Rasters
Summarization
Dasymetric Mapping
Join Contextual data
Zonal Statistics

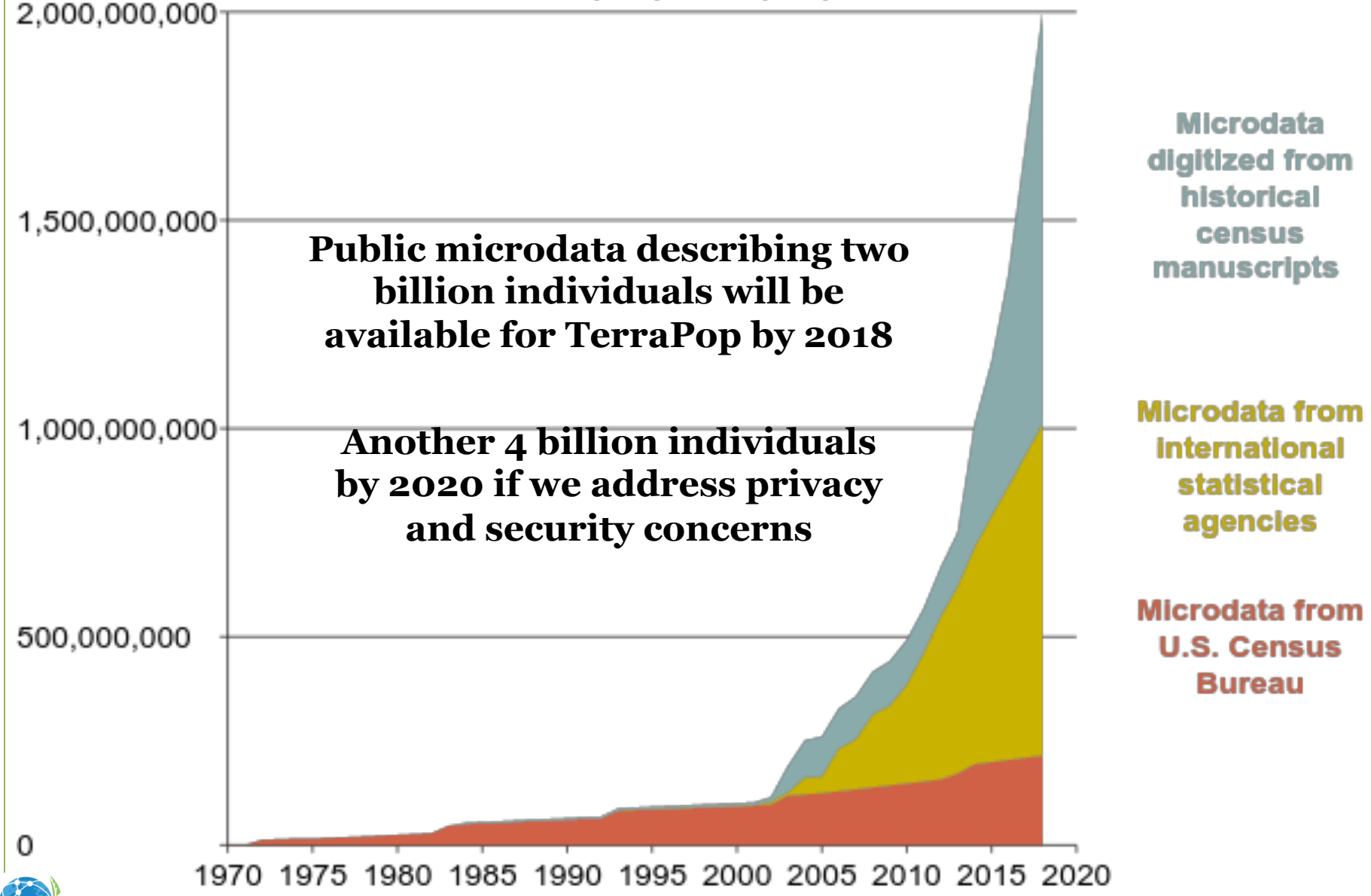
Terra Populus

DATA
PARAGON
TABULATOR
GEOVISUALIZATION

Data

- Microdata
  - 70+ countries
  - 2nd and 3rd level geographies
- Aggregate census data
  - Historical data (48 countries)
  - Variables in addition to population by sex (65 countries)
- Environmental data
  - CRU monthly time series – precipitation & temperature
  - Elevation and derived characteristics
  - Soils

Micro Data


2,000,000,000
1,500,000,000
1,000,000,000
500,000,000
0
1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020
Public microdata describing two billion individuals will be available for TerraPop by 2018
Another 4 billion individuals by 2020 if we address privacy and security concerns
Microdata digitized from historical census manuscripts
Microdata from international statistical agencies
Microdata from U.S. Census Bureau

Areal Data

- US Population
  - 265 billion data points
- 95+ countries whose population < 500,000 people
- 87+ countries with microdata records
- 65 countries have data beyond total population by age and sex
- 54 countries have data at the second geographic level or finer resolution

Raster Data

- MODIS Land Data
  - Yearly land cover data derived from the MODIS Terra and Aqua satellites, available for 2001 – 2013
  - 5 land cover classifications, 240 Gigabytes
- Earth Science
  - Aster 30 Meter DEM resolution - 500 Gigabytes
  - TAUDEM derivatives: slope, solar radiance, wetness index will result in about 6-8 more Terabytes of data
- Climate Datasets
  - NetCDF Format
  - Climate Research Unit – 40 Gigabytes

What Does TerraPop Do?

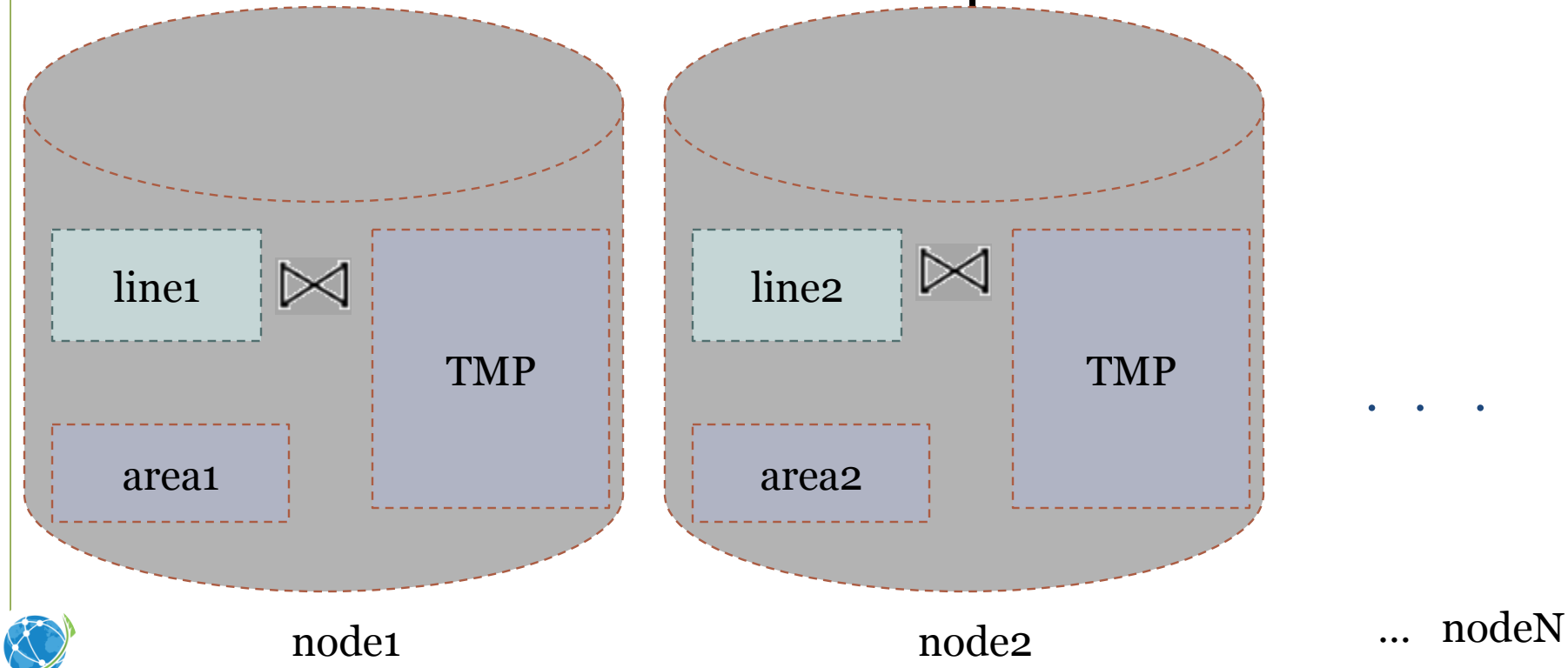
TerraPop

- Access
- Exploration
- Interoperability
- Fusion
- Analysis

Paragon

Joins in Distributed Databases

- Create a temporary table TMP
- Reconstitute area on each node as TMP
- Join TMP with the two local partition of line



Spatial Join Paragon

Query:

```
select a.gid , b.gid from edges_merge_ca_shall
as a, arealm_merge_ca_shall
as b where st_crosses(a.geom, b.geom) ;
```

PostgreSQL (standalone): 463 seconds
Stado-Spatial (2 nodes): 96 seconds

Tabulator

Tabulator

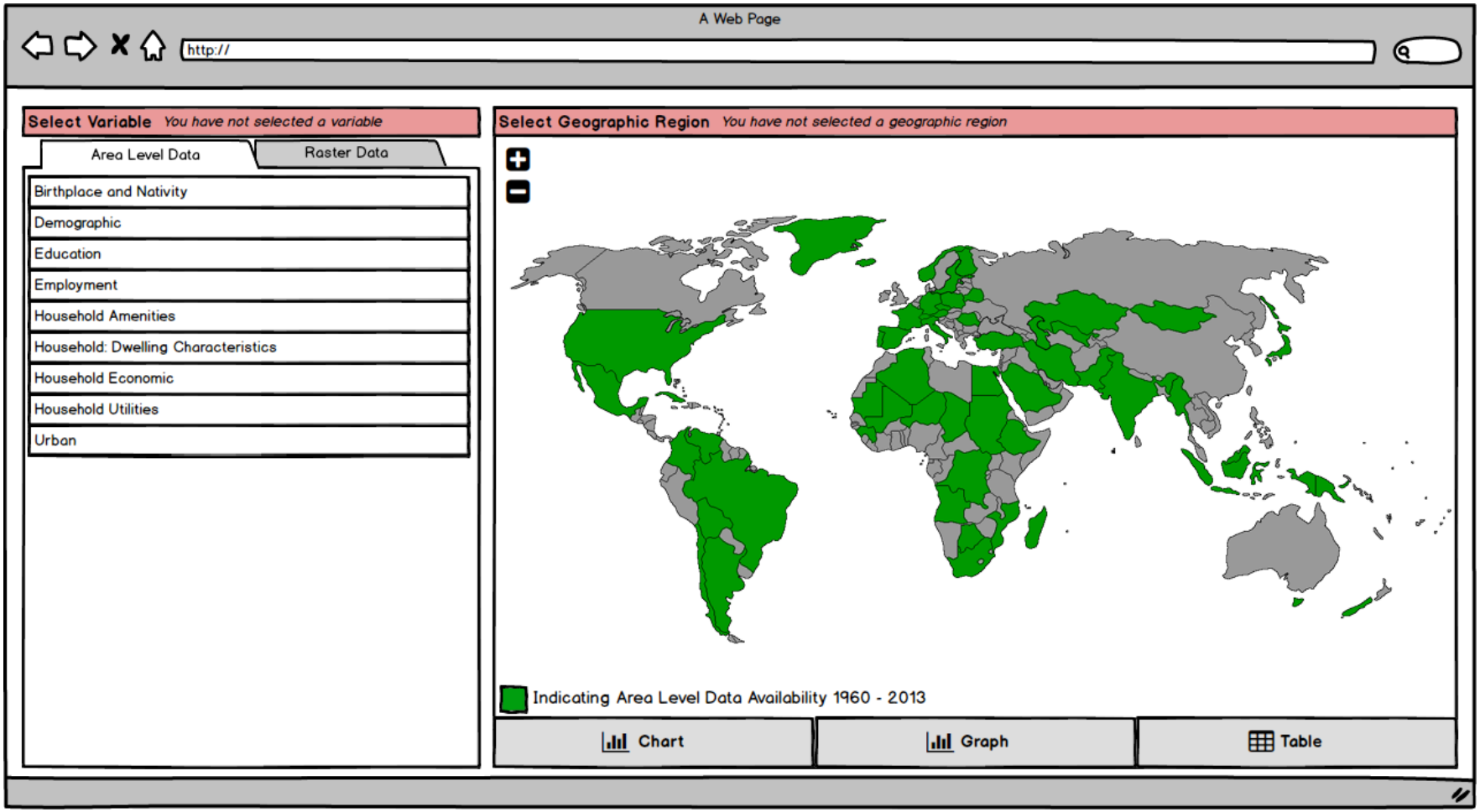
- Generates area-level data from microdata using geographic level codes
  - National, First Level (e.g. State), Second Level (e.g. County)
- Parquet on Apache Spark
  - High Compression Ratio
    - 8 Gigabytes gzip compressed
    - 3 Gigabytes parquet compressed
  - Columnar Storage (3,000+ columns)
  - 3 Nodes with each allocated 8 gigs of memory

Query Performance

Number of datasets	Number of records	Time to aggregate by 1 column	Time to aggregate by 2 columns	Time to aggregate by 3 columns
1	12 million	5 seconds	9 seconds	10 seconds
10	25 million	7 seconds	10 seconds	18 seconds
20	82 million	10 seconds	12 seconds	27 seconds
56	128 million	7.5 seconds	20 seconds	30 seconds

Visualization

Landing Page



Terra Populus Software Stack

Web Application

Data Processing

Geospatial Data Repository

Websites

EXTRACT BUILDER
HTTPS://DATA.TERRAPOP.ORG/

TERRACLIP
HTTPS://DATA.TERRAPOP.ORG/TERRACLIP

Questions

?