massDOT Transportation Innovation Conference

Using Open Data and Open Tools to Enhance the Transparency and Reproducibility of Geospatial Analysis

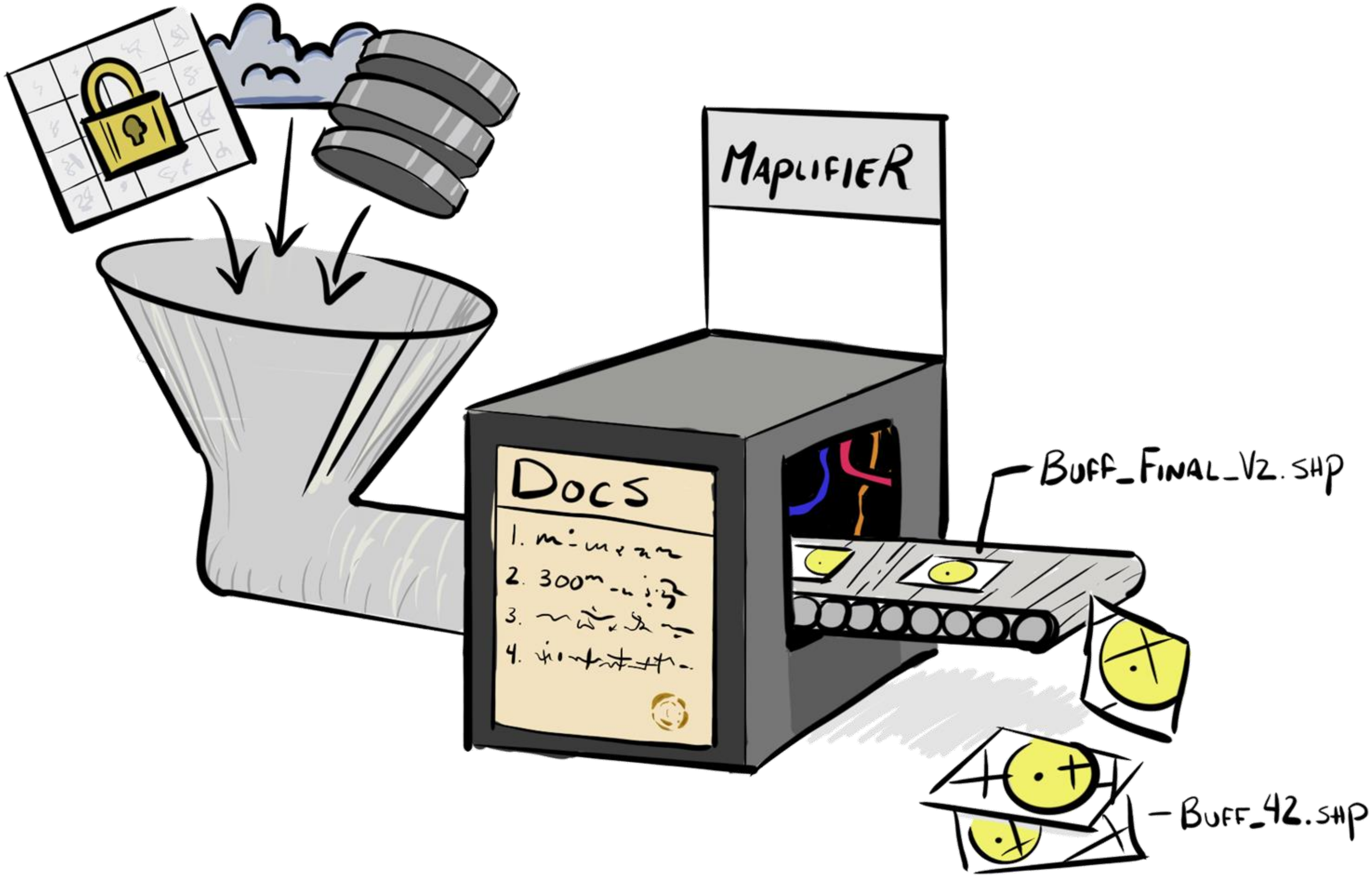
Steven Andrews

Boston Region Metropolitan Planning Organization

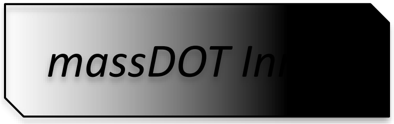
massDOT Massachusetts Department of Transportation

U.S. Department of Transportation Federal Highway Administration

This is how our analysis can look:



Let’s start with some definitions:

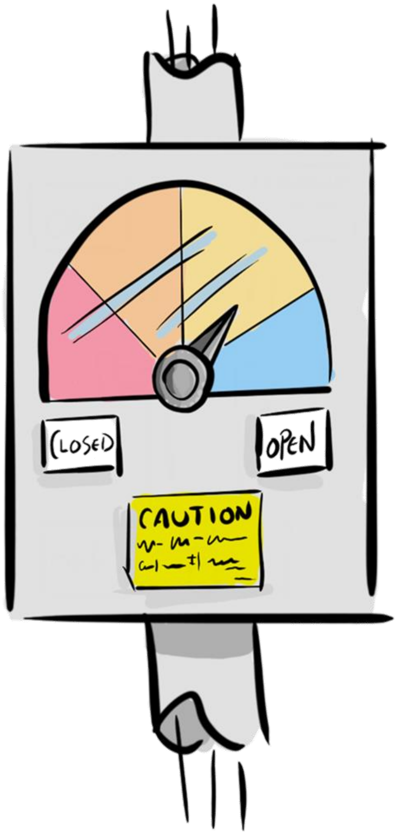


Transparency
the level of forthcomingness you provide

Reproducibility
the ability for someone to use the same inputs and methods to get the same results as you.

Replicability (new inputs, same or similar analysis, same outcome) is also neat,
but perhaps outside this talk.

and we'll follow that up with some proclamations:



Openness drives **confidence** and **trust** in our analyses.

Openness lets us **share and grow** with our colleagues, researchers, and the public

Openness is a continuum, and incremental progress is better than no progress.

Sometimes we won't be as open as we'd like. We should do what we can such that we're not limited by design decisions or technology.

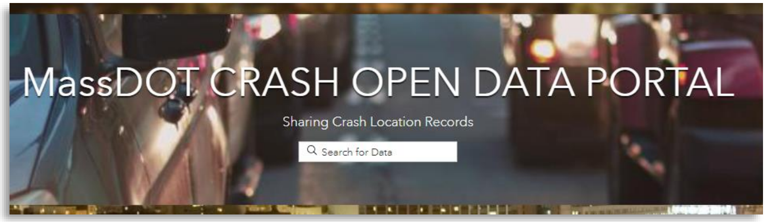
Thanks to hard work across the state, we’ve come a long way.

Massachusetts geoDOT Open Data Portal

GIS For Transportation

MBTA Open Data Portal

Transparency in Transit


MassDOT CRASH OPEN DATA PORTAL

Sharing Crash Location Records

Search for Data

United States® Census Bureau

The proliferation in open data portals has given rise to a highly available place for us to grab data for our analyses.

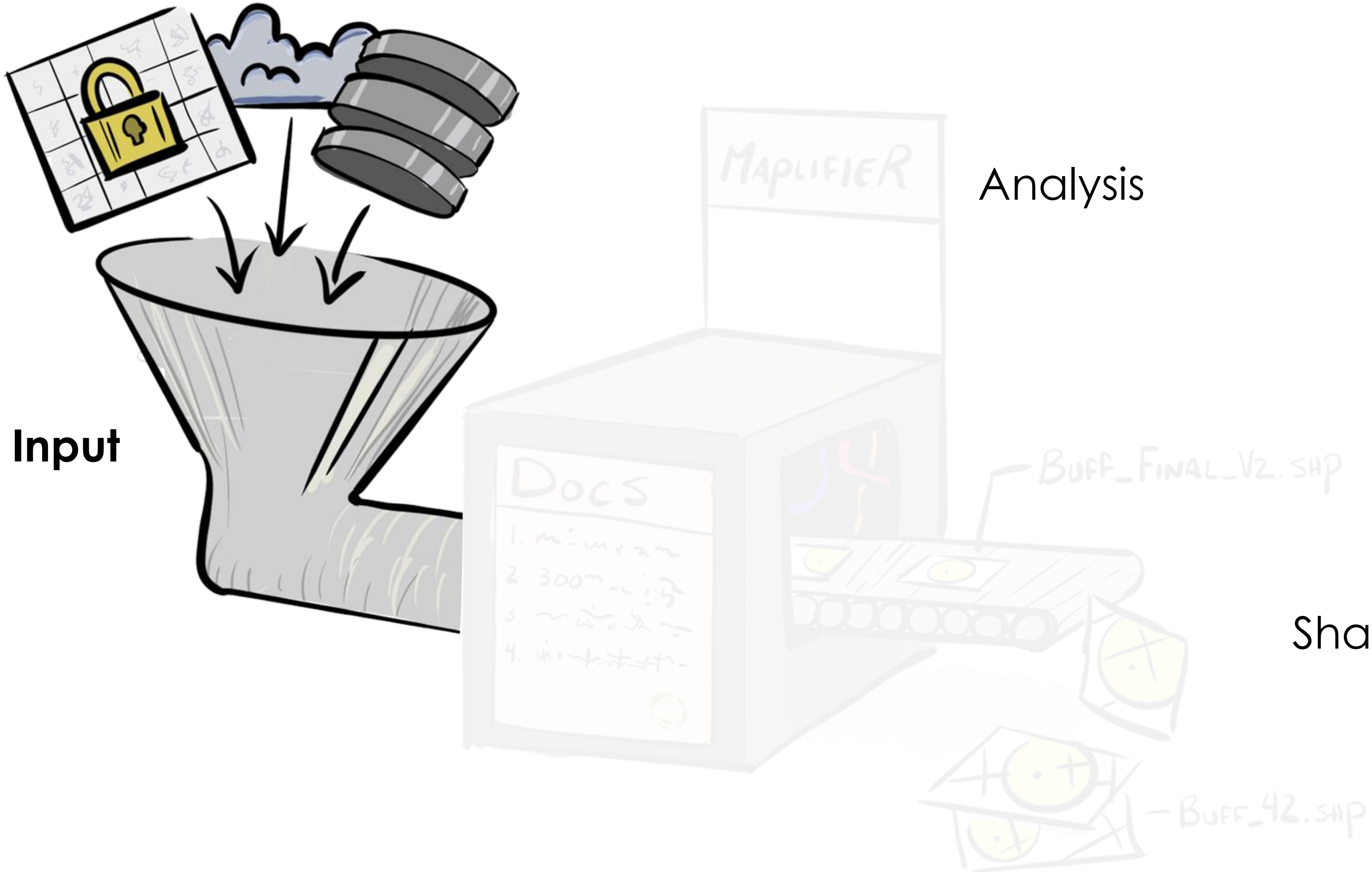
Roadways

Ridership
Schedules
Station Locations
Vehicle locations

Crash information

Census Data *(thanks Census Bureau!)*

When we have a ‘private’ and a public version of a dataset, we can use the public version.

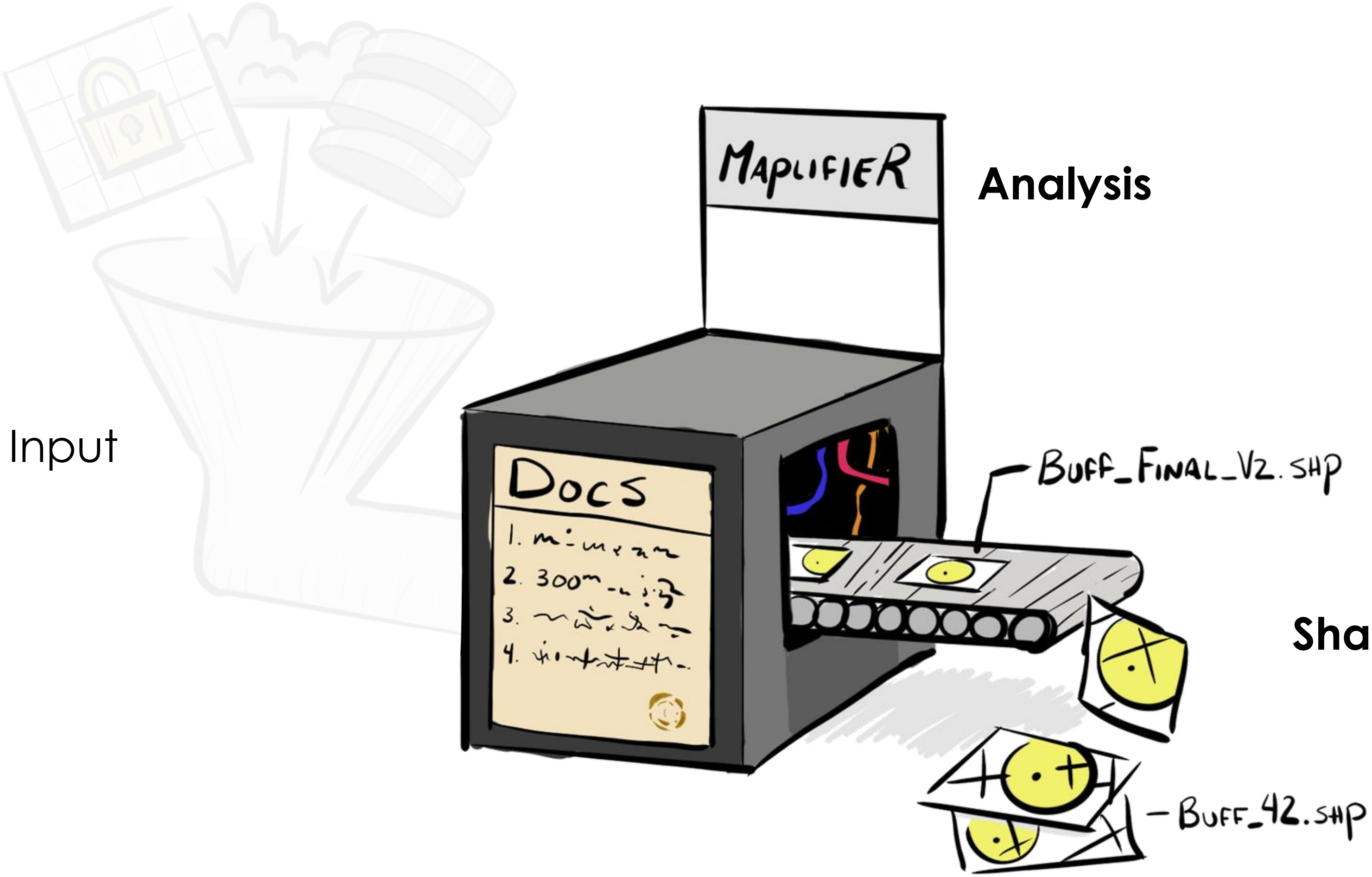


Input

Analysis

Sharing

We still have a ways to go on the rest of this.



Input

Analysis

Sharing

We have adopted open tools into *(some)* of our analysis pipelines.

Input

United States® Census Bureau

MBTA Open Data Portal
Transparency in Transit

OSM

Output

GEOJSON

geopackage

(geospatial) Analysis

R

sf

sfnetworks

tidycensus

Stars

Python

GeoPandas

igraph

rasterio

arcpy*

Platforms

QGIS

Sharing

Code

GitHub

Notebooks

quarto®

jupyter

Display

Leaflet

deck.gl

Here's an example of what a (basic) geospatial analysis looks like in this workflow:

Interactive Quarto Doc

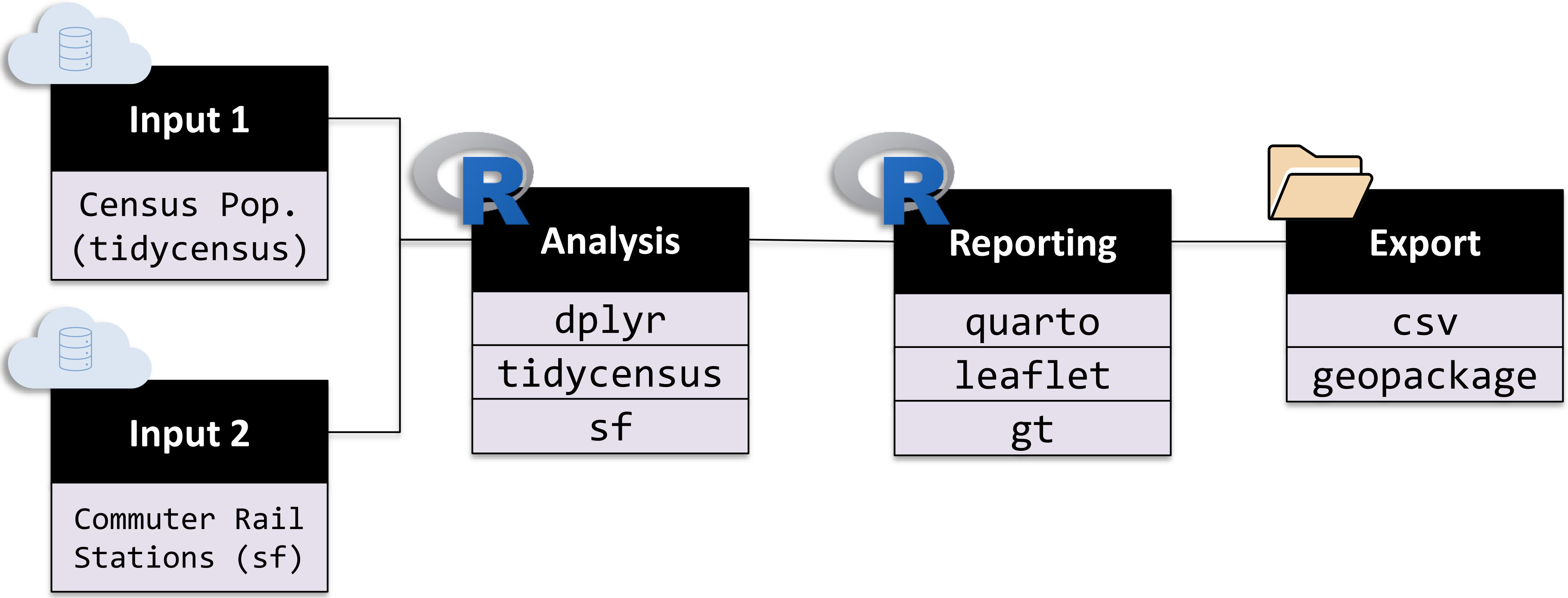
ctpsstaff.github.io/MBTA_CR_muni_population/

Public GitHub repo

github.com/CTPSSTAFF/MBTA_CR_muni_population

Question: What is the population of MA municipalities that have a commuter rail station within their borders?

Here’s an example of what a (basic) geospatial analysis looks like in this workflow:



Input 1

Census Pop. (tidycensus)

Input 2

Commuter Rail Stations (sf)

Analysis

dplyr

tidycensus

sf

Reporting

quarto

leaflet

gt

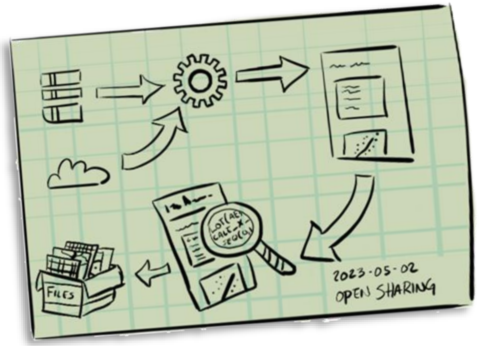
Export

csv

geopackage

All of the analysis, geospatial operations, stats, and summaries are in one place, and people can follow it through top to bottom. No intermediate files are saved if I don’t want them.

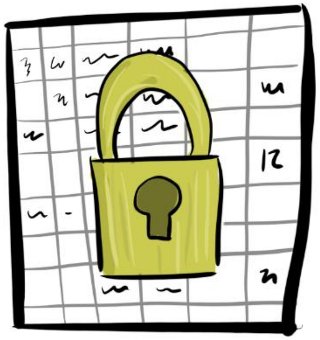
Merely using these tools is not a cure-all.



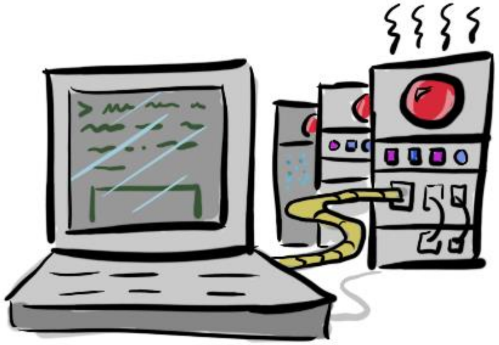
Still have to plan

Scrubbing private data

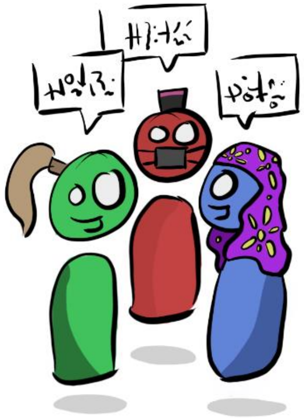
Package it all up



(Useful) proprietary data



Large scale computing



Openness culture

Being more open is something we should make active progress towards, and the tools to open our analyses exist today.

Being more open is something we should make active progress towards, and the tools to open our analyses exist today.

I challenge us as an industry to reach beyond merely sharing our results and our narratives and do the work to share our analysis processes and our work more fully.

Resources I came upon while creating this presentation:

Reproducibility, Replicability, and Reliability (2020)
https://hdsr.mitpress.mit.edu/pub/hn51kn68/release/4

Reproducibility and Replicability in Science (2019)
https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science

A Beginner's Guide to Conducting Reproducible Research (2021)
https://esajournals.onlinelibrary.wiley.com/doi/full/10.1002/bes2.1801

Reasons, challenges, and some tools for doing reproducible transportation research (2021)
https://www.sciencedirect.com/science/article/pii/S2772424721000044

And if you're interested in an entire conference series dedicated to this, check out foss4g https://foss4g.org/

FOSS4G
Prizren, 2023

Using Open Data and Open Tools to Enhance the Transparency and Reproducibility of Geospatial Analysis

Much of the geospatial analysis performed in the transportation industry today is completed using proprietary tools, typically those in the ESRI ecosystem, using data that lives behind seamless public access on private servers. Analyses using these technologies, while powerful, make it difficult for the public–and even other analysts–to discover or recreate how data was transformed into results.

The work of many data and GIS professionals, including the GIS-giant ESRI and workers across MassGIS, MassDOT, and the MBTA, has led to the proliferation of open data initiatives. At the same time, talented and committed open source programming teams have created exceptionally powerful and programmer-friendly suites of geospatial analysis tools–perhaps most notably, though not exclusively–the `geopandas` library in Python and the `sf` library in R.

Using these open resources, we at the Boston Region MPO have created analyses that are auditable and reproducible from ingesting data, through manipulation and transformation, to generating output. We have used open tools to support equity analyses, to evaluate transit coverage areas, to aggregate US Census data, and to calculate federally-mandated roadway performance metrics.

While there are certainly start-up costs–one needs to invest in learning new ways of working and developing a culture where using alternative techniques is encouraged–the tools themselves, which are supported by enthusiastic developer communities, are relatively straightforward to use. By choosing an open solution when it is an option and leaving the proprietary software for those times when it is truly needed, we can continue pushing towards our shared openness and transparency goals.