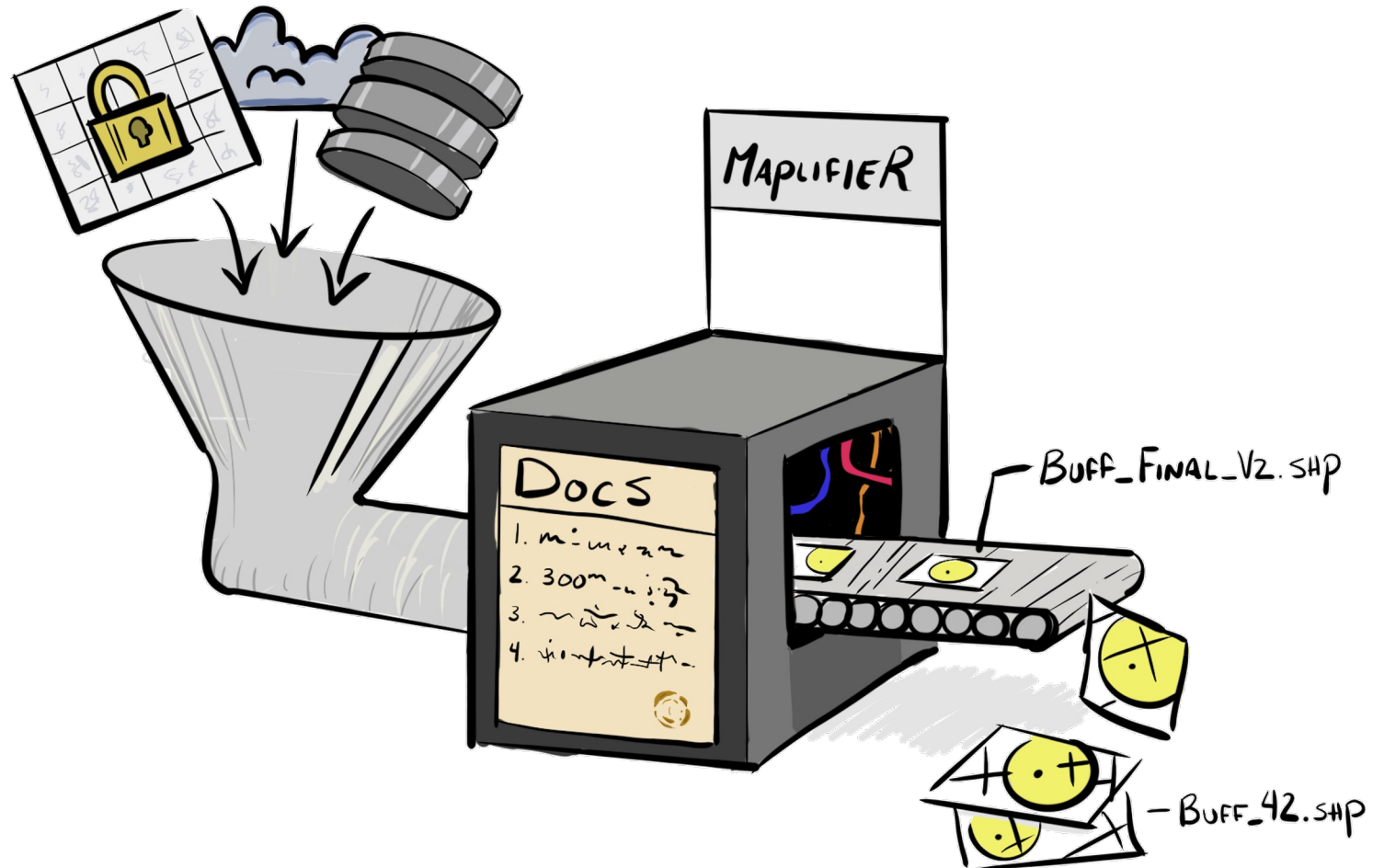


Using Open Data and Open Tools to Enhance the Transparency and Reproducibility of Geospatial Analysis

Steven Andrews

Boston Region Metropolitan Planning Organization

This is how our analysis can look:



Let's start with some definitions:



Transparency

the level of forthcomingness you provide

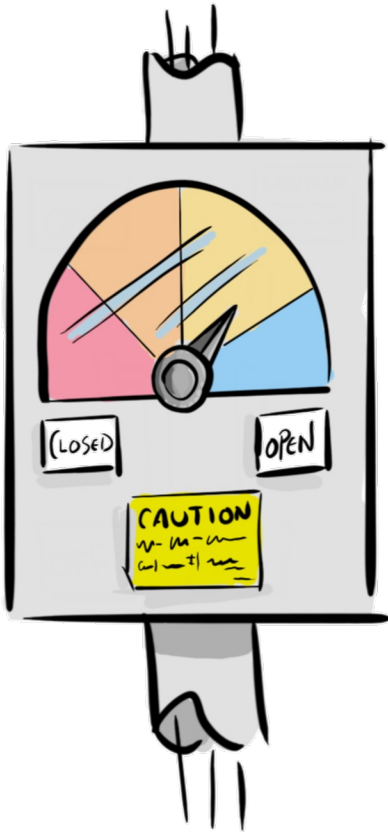


Reproducibility

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:



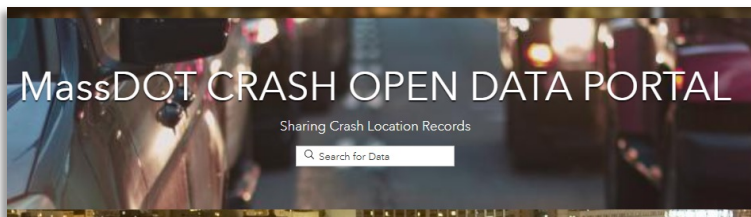
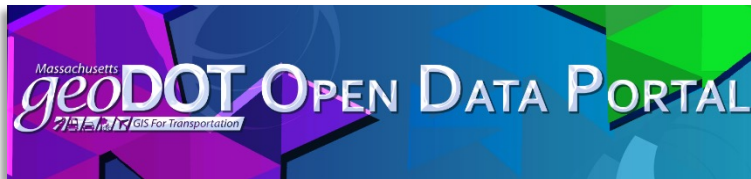
Openness drives **confidence** and **trust** in our analyses.

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.



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

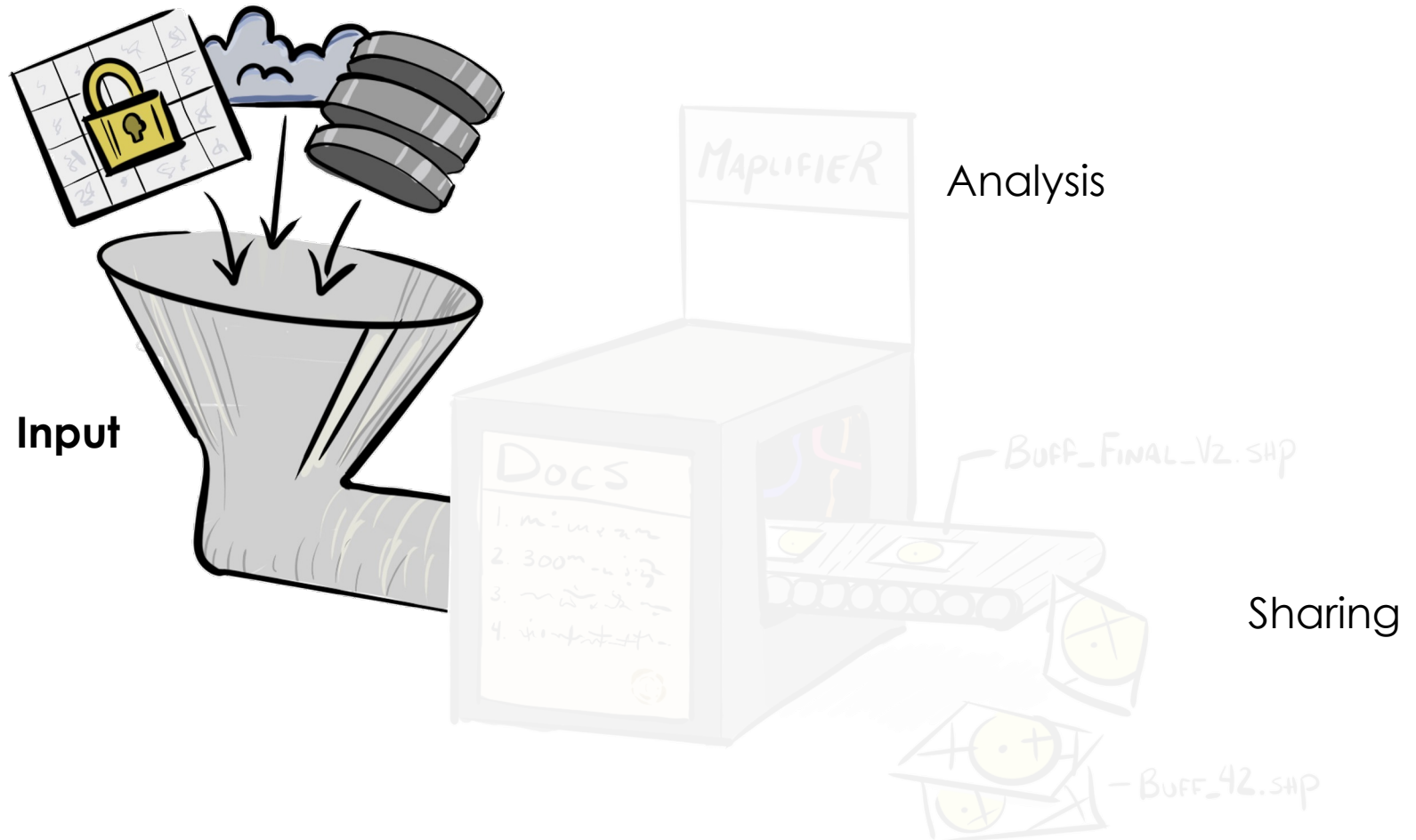
Station Locations

Vehicle locations

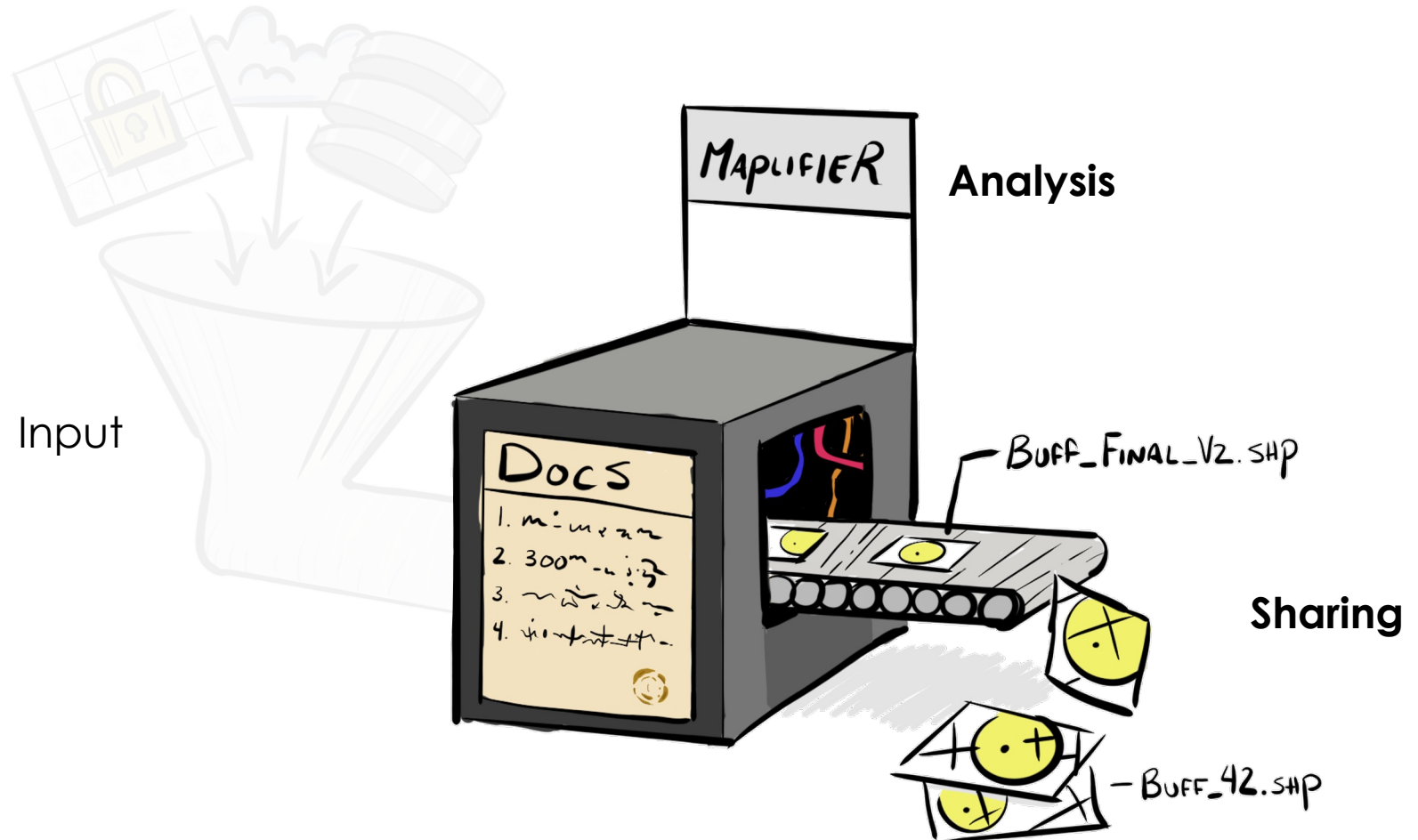
Crash information

Census Data (*thanks Census Bureau!*)

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



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



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

Input



OSM

Output

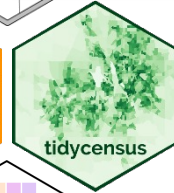
GEOJSON



geopackage

(geospatial) Analysis

R



Python



GeoPandas



igraph

rasterio

arcpy*

Platforms



Sharing

Code



Notebooks



Display



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/

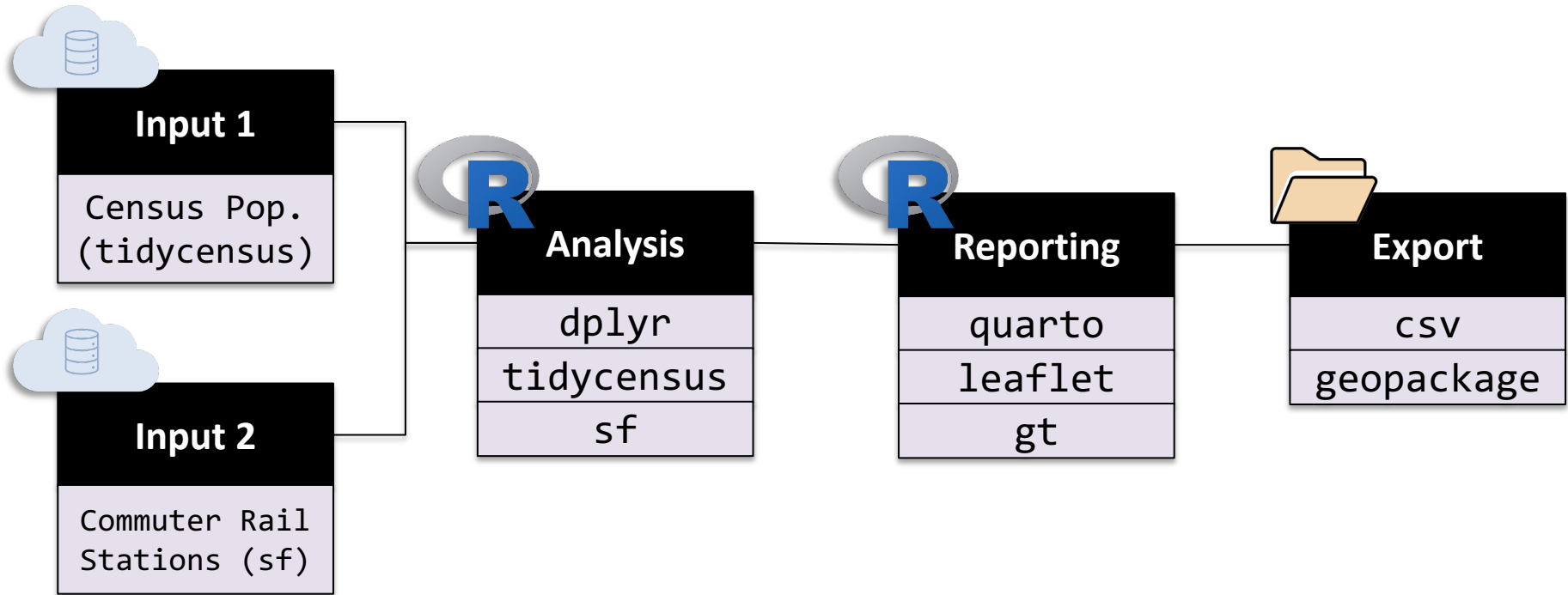


Public GitHub repo

github.com/CTPSSTAFF/MBTA_CR_muni_population

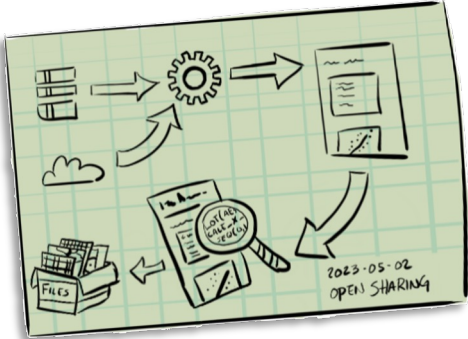
Question: What is the population of MA municipalities that have a commuter rail station within their borders?

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



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.



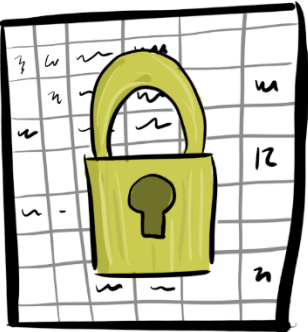
Still have to plan



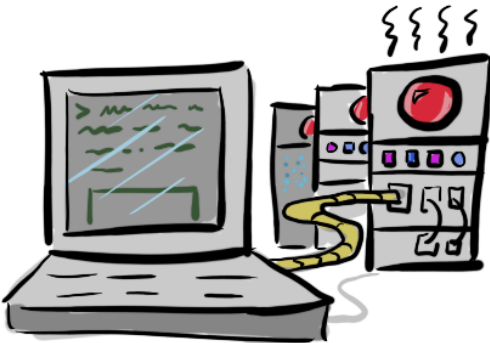
Scrubbing private data



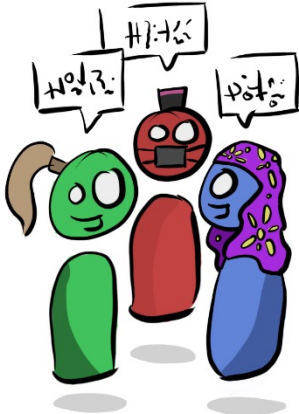
Package it all up



(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://hdr.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/>

