

The Map Is Not the Territory: Lessons from Mapping the Modern Agora

Jae Yeon Kim (SNF Agora Institute, JHU)

Prepared for NLP for CSS Guest Lecture (April 7, 2025)

Quick Intro & Goals

- Who am I?
 - Public policy scholar
 - Computational *social* scientist
 - Public interest technologist
- Goals for today:
 1. Why domain knowledge matters
 2. Lessons from Mapping the Modern Agora
 3. Practical takeaways

The map is not the territory



Know the Map-Territory gap

- **Data science** helps us turn data into *actionable insights*
- But... **is your data a *good enough* map of the territory?**

To answer that, we need to:

1. **Know the territory**

Understand how the data came from the real world.

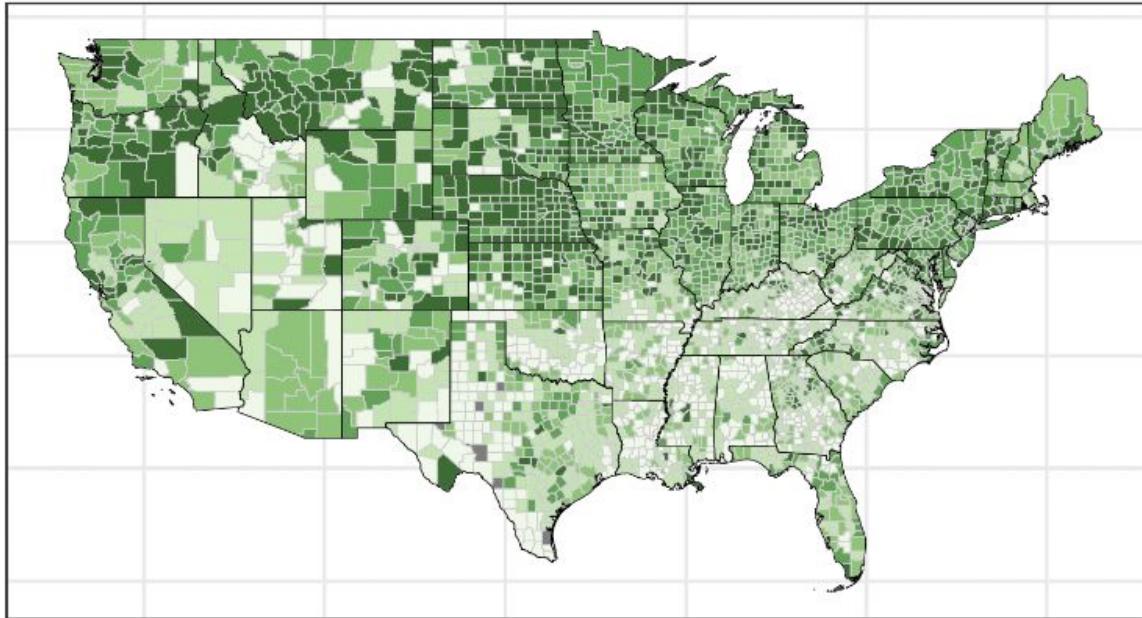
2. **Compare the map with the territory**

Validate data and models with the real world.

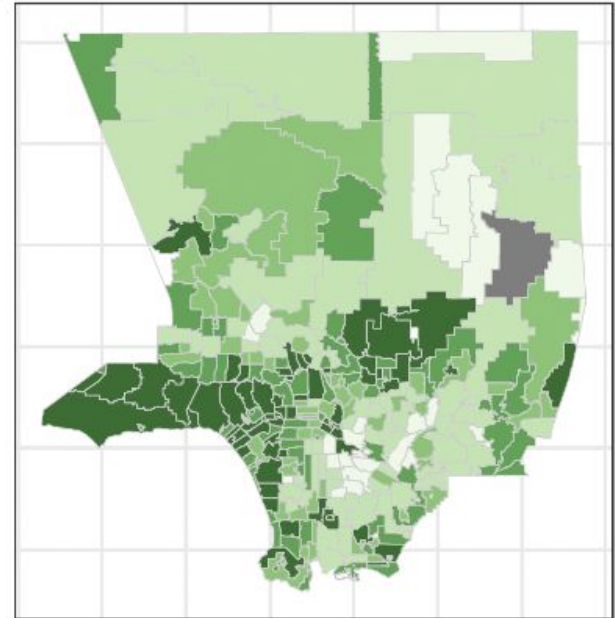
→ For both steps, **domain knowledge is essential**

Mapping the Modern Agora

A



B



Civic opportunity index

| | | |
|---|---|---------|
| 1 | 3 | 5 |
| 2 | 4 | No Data |

- **Goal:** Understand where and how people can participate in civic life today
- **Built a national dataset** of organizations providing civic opportunities, including:
 - Using IRS tax records (~1.8M)
 - Scraped ~1.1M organizational websites
- **Classified civic opportunity types and their providers**
- **Mapped civic infrastructure** at the county and ZIP-code level
- Linked civic data to **real-world behaviors** (e.g., vaccination rates, mutual aid group formation)
- **Compiled contact info** to survey organizational staff and leaders

Why study civic infrastructure?

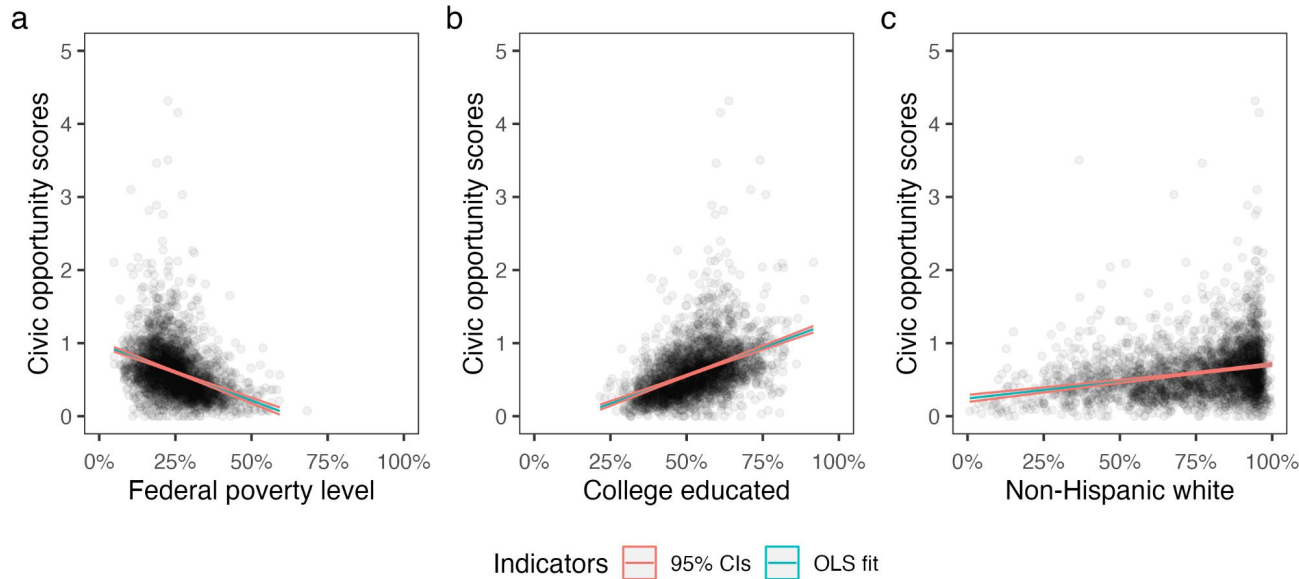
People learn to engage civically **by practicing it**
→ Civic infrastructure enables that practice

But studying it is hard:

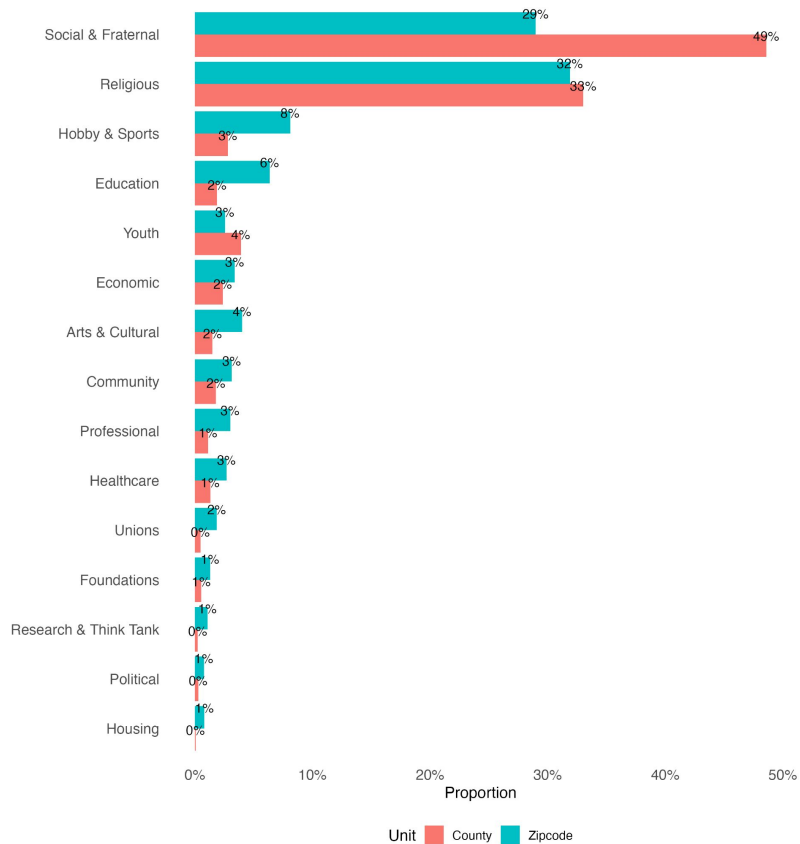
- Over **1 million decentralized, organic organizations**
- No ready-made dataset

This project helps **fill that gap**

→ Opening up new research opportunities and unanswered questions



Civic opportunities are more common in educated, white, and affluent counties



Religious and social-fraternal organizations make up 37% of civic opportunities—and are the top providers in 85% of counties.

Practice quiz (sorry, no bonus point)

1. What is the *map* in the MMA project?
2. What is the *territory* in the MMA project?



The Map

- Administrative data (IRS records)
- Web-scraped organizational websites
- Machine-classified civic opportunities and their providers
- Aggregated civic opportunity scores by county and ZIP code



The Territory

- Actual civic activity on the ground (*some captured*)
- Informal, under-the-radar, or unregistered groups (*mostly not captured*)

How to build a useful civic opportunity map



Step 1: Analyzed tax-exempt organization types

→ Compared organization categories and filing requirements (who files, who doesn't)



Step 2: Examined IRS Form 990 structures

→ Identified content differences across 990, 990-EZ, 990-PF, 990-N, etc.
→ Parsed and standardized XML versions for structured analysis



Step 3: Built a custom data pipeline

→ Linked IRS records to geolocations and organizational websites
→ Scraped website content at scale
→ Classified civic opportunities and provider types using NLP + machine learning

Step 3 isn't possible without Steps 1 and 2.

Understanding the structure and limitations of IRS data is key to building a useful civic map.

Step 1 : Understand tax-exempt orgs

| Type | IRS Code | Example Activities | Filing Required? |
|---------------------------|-----------|-----------------------------|------------------|
| Public Charities | 501(c)(3) | Education, health, arts | Yes (mostly) |
| Private Foundations | 501(c)(3) | Grantmaking | Yes (990-PF) |
| Social Welfare Orgs | 501(c)(4) | Advocacy, civic leagues | Yes |
| Business Leagues | 501(c)(6) | Trade associations | Yes |
| Religious Orgs (Churches) | 501(c)(3) | Worship, spiritual services | No (voluntary) |

Some examples of tax-exempt organizations and their filing requirements.

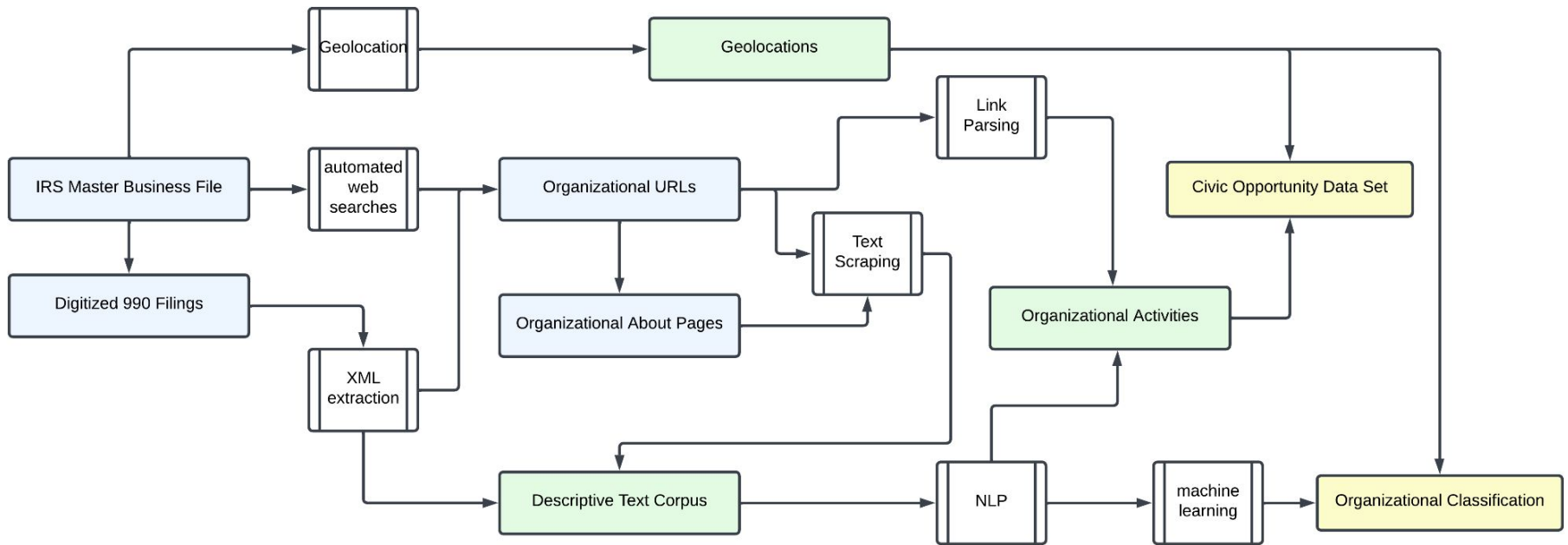
Step 2 : Examine IRS forms

| | | |
|--|--|---|
| Form 990 | Return of Organization Exempt From Income Tax | OMB No. 1545-0047 |
| Department of the Treasury Internal Revenue Service | Under section 501(c), 527, or 4947(a)(1) of the Internal Revenue Code (except private foundations) Do not enter social security numbers on this form as it may be made public. Go to www.irs.gov/Form990 for instructions and the latest information. | 2021 Open to Public Inspection |
| A For the 2021 calendar year, or tax year beginning , 2021, and ending , 20 | | |
| B Check if applicable: <input type="checkbox"/> Address change <input type="checkbox"/> Name change <input type="checkbox"/> Initial return <input type="checkbox"/> Final return/terminated <input type="checkbox"/> Amended return <input type="checkbox"/> Application pending | C Name of organization Doing business as Number and street (or P.O. box if mail is not delivered to street address) Room/suite City or town, state or province, country, and ZIP or foreign postal code | D Employer identification number E Telephone number G Gross receipts \$ |
| F Name and address of principal officer: | | H(a) Is this a group return for subordinates? <input type="checkbox"/> Yes <input type="checkbox"/> No H(b) Are all subordinates included? <input type="checkbox"/> Yes <input type="checkbox"/> No If "No," attach a list. See instructions. |
| I Tax-exempt status: <input type="checkbox"/> 501(c)(3) <input type="checkbox"/> 501(c) () (insert no.) <input type="checkbox"/> 4947(a)(1) or <input type="checkbox"/> 527 | H(e) Group exemption number ▶ | |
| J Website: ▶ | K Form of organization: <input type="checkbox"/> Corporation <input type="checkbox"/> Trust <input type="checkbox"/> Association <input type="checkbox"/> Other ▶ | |
| L Year of formation: | | M State of legal domicile: |
| Part I Summary | | |
| 1 Briefly describe the organization's mission or most significant activities: | | |

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0" encoding="UTF-8" ?>
<Return xmlns="http://www.irs.gov/efile" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.irs.gov/efile returnVersion="2014v5.0">
  <ReturnHeader binaryAttachmentCnt="0">
    <ReturnTa>2015-05-14T18:01:56-05:00</ReturnTa>
    <TaxPeriodEndDt>2014-12-31</TaxPeriodEndDt>
  </ReturnHeader>
  <PreparerFirmGrp>
    <PreparerFirmID>330885895</PreparerFirmID>
    <PreparerFirmName>
      <BusinessNameLine1Txt>LINDSAY & BROWNELL LLP</BusinessNameLine1Txt>
    </PreparerFirmName>
    <PreparerUSAddress>
      <AddressLine1Txt>4225 EXECUTIVE SQUARE SUITE 1150</AddressLine1Txt>
      <City>LA VOILLAK</City>
      <StateAbbreviationCd>CA</StateAbbreviationCd>
      <ZIPCd>92037</ZIPCd>
    </PreparerUSAddress>
    <PreparerFirmGrp>
      <ReturnTypeCd>990</ReturnTypeCd>
      <TaxPeriodBeginDt>2014-01-01</TaxPeriodBeginDt>
    </PreparerFirmGrp>
  </PreparerFirmGrp>
  <Preparer>
    <EIN>201585919</EIN>
    <BusinessName>
      <BusinessNameLine1Txt>VOICE OF SAN DIEGO</BusinessNameLine1Txt>
    </BusinessName>
    <BusinessNameControlTxt>VOIC</BusinessNameControlTxt>
    <PhoneNum>6193250525</PhoneNum>
    <USAddress>
      <AddressLine1Txt>2508 HISTORIC DECATUR SUITE 120</AddressLine1Txt>
      <City>SAN DIEGO</City>
      <StateAbbreviationCd>CA</StateAbbreviationCd>
      <ZIPCd>92106</ZIPCd>
    </USAddress>
    <Preparer>
      <PersonName>ANN ALPERT</PersonName>
      <PersonTitle>CFO</PersonTitle>
      <PhoneNum>6193250525</PhoneNum>
      <SignatureDt>2015-05-13</SignatureDt>
      <DiscussWithPaidPreparerInd>1</DiscussWithPaidPreparerInd>
    </Preparer>
  </Preparer>
  <PreparerPersonGrp>
    <PreparerPersonName>MARY H MCGROARTY</PreparerPersonName>
    <PTIN>P00735101</PTIN>
    <PhoneNum>6193250525</PhoneNum>
  </PreparerPersonGrp>
  <TaxYear>2014</TaxYear>
  <BuildTS>2016-02-25 16:41:14Z</BuildTS>
</Return>
```


Step 3 : Build a pipeline



Why is this map more useful than just using the IRS tax records?



Coverage Error

- Many orgs aren't required to file detailed forms (i.e., 990).
- Our response: combined **IRS data** with **web scraping** to increase coverage beyond official tax records



Processing Error

- IRS forms are inconsistent across types (990, 990-EZ, PF, N)
We parsed and standardized **XML filings** to ensure consistent extraction of key variables.



Classification Error

- Organization types and their activities aren't neatly labeled
- We used **NLP + machine learning** to classify civic opportunity and their providers based on text from websites and filings

Making data useful =
knowing what it misses




Data isn't useful if your audience doesn't find it meaningful

Data is just a **map**—not the **territory**

To build a useful map, we must:

1. **Understand the territory**
2. **Examine what's missing** from existing maps
3. **Improve the map** to better reflect what matters

Your Turn: Reflect on Your Own Work

1.  **What's the map and the territory in your project?**
(What data are you using, and what real-world phenomena are you trying to understand?)
2.  **How would you evaluate the map–territory gap?**
(What's missing, misrepresented, or overly simplified?)
3.  **How might you improve the map?**
(What data, domain knowledge, or methods could help?)