API-231 / GIS-PubPol
Meeting 00 (GIS 101)

Yuri M. Zhukov
Visiting Associate Professor of Public Policy
Harvard Kennedy School

January 22, 2024

Welcome to API-231!

What is GIS?

What are **Geographic Information Systems**?

1. tools for collection, maintenance, storage, analysis, visualization, distribution of geospatial data
2. a.k.a. "geospatial data science"

Policy applications

1. GIS help us understand
   a) where social, economic, public health problems occur
   b) who is affected by them
   c) how to monitor, manage and mitigate them

Scientific applications

1. GIS help us
   a) acquire data
   b) test hypotheses
   c) make forecasts and predictions

History of geospatial data analysis / Public health

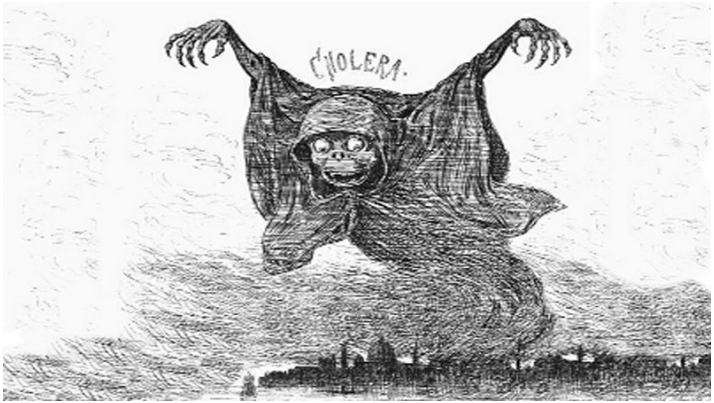


Figure 1: The blue wraith

In 1854, there was an outbreak of Cholera in London's Soho district. It killed 616 people.

Figure 2: The blue death

Cholera is an infectious disease of the small intestine.
It causes severe dehydration in its victims. It is quite deadly.

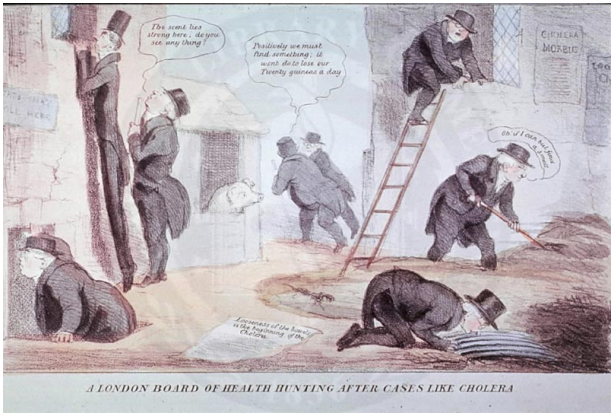


Figure 3: The experts

The Board of Health struggled to find the epidemic's cause.

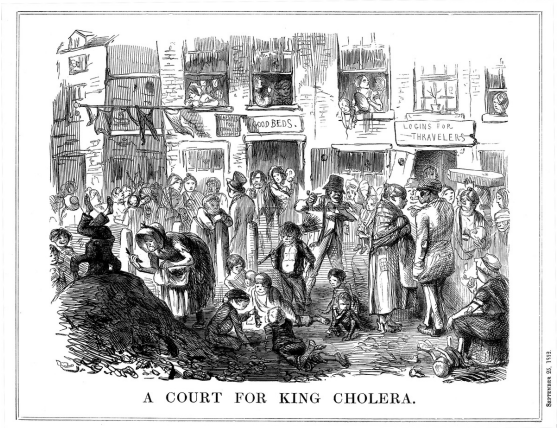


Figure 4: The scapegoats

The leading theory was that Cholera was caused by foul air and poor hygiene.

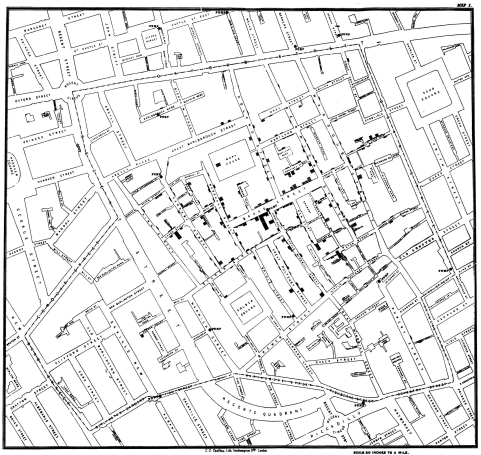


Figure 5: The geospatial data science

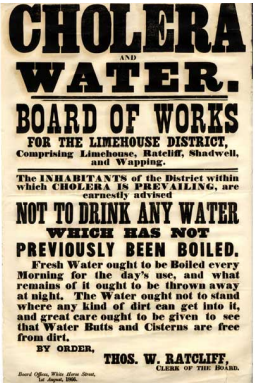
Dr. John Snow believed that Cholera was spread by contaminated water, not foul air. To test his theory, he created a map of Cholera cases at each address in Soho.

Figure 6: The pump

He found a **cluster** of cases around a water pump on Broad St.

Figure 7: The public health campaign

He told authorities to break pump's handle so people couldn't draw water from it.



CHOLERA
AND
WATER.

BOARD OF WORKS
FOR THE LIMEHOUSE DISTRICT,
Comprising Limehouse, Ratcliff, Shadwell, and Wapping.

The INHABITANTS of the District within which CHOLERA IS PREVAILING, are earnestly advised

NOT TO DRINK ANY WATER
WHICH HAS NOT
PREVIOUSLY BEEN BOILED.

Fresh Water ought to be Boiled every Morning for the day's use, and what remains of it ought to be thrown away at night. The Water ought not to stand where any kind of dirt can get into it, and great care ought to be given to see that Water Butts and Cisterns are free from dirt.

BY ORDER,

THOS. W. RATCLIFF,
CLERK OF THE BOARD.

Board Office, White Horse Street, 1st August, 1866.

Figure 8: Data scientist has policy impact

Cholera cases drop in Soho, supporting Snow's theory about contaminated water.

History of geospatial data analysis / Military intelligence

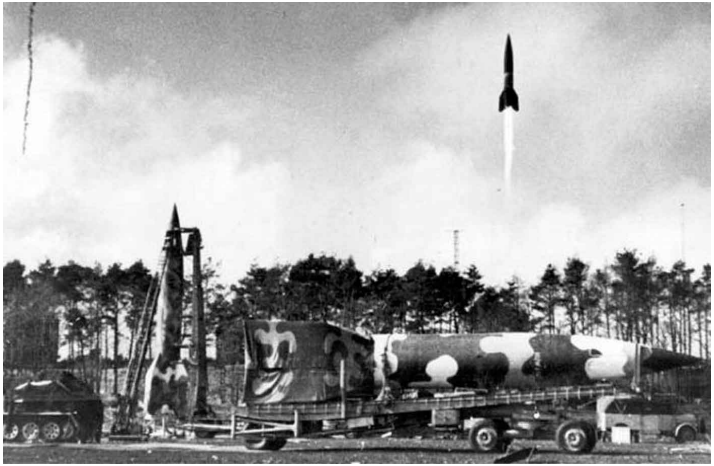


Figure 9: A new weapon

During WWII, Germany launched 1,358 V-2 Rockets at London.

Figure 10: No way to intercept, no way to defend

V-2's speed and trajectory made it invulnerable to anti-aircraft guns and fighter jets.

Figure 11: Terror rains from the sky

V-2 strikes kill 2,724 people in UK + 6,000 civilians and military across Europe.
(+ 15,000 concentration camp prisoners died constructing the V-2)

Figure 12: Locations of V-2 strikes in London

Bomb damage maps were interpreted by some analysts as showing that impact sites were **clustered**. This suggested the V-2's guidance system was more sophisticated than intel estimates thought. Allies tried to jam V-2's guidance system, to no effect.

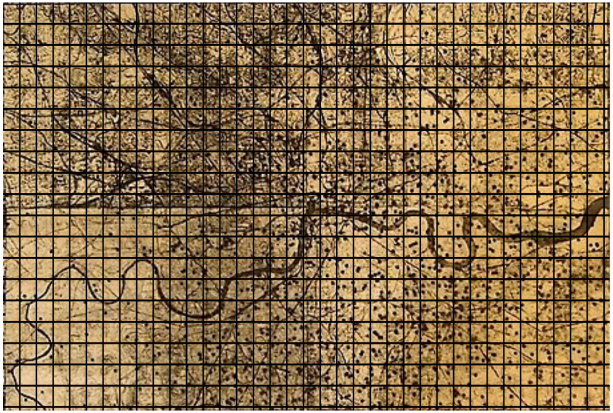


Figure 13: Wartime geostatistics

R.D. Clarke applied a statistical test to assess whether any hard evidence could be found for clustering. For each square, Clark recorded the total number of *observed* bomb hits (537 total in study area), and number of squares with $k = 1, 2, 3, \ldots$ hits.

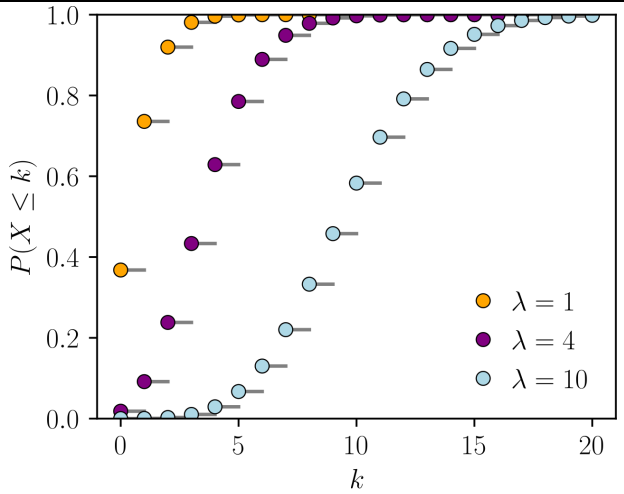


Figure 14: Poisson cumulative density function

Clarke derived the *expected* number of squares with k hits from the cumulative density function of the Poisson distribution $\sum_{k=1}^{n} \frac{e^{-\lambda}\lambda^k}{k!}$, with $\lambda = \frac{537}{576}$ and $n = 576$.

No. of bombs per square	Expected	Observed
1	226.74	229
2	211.39	211
3	98.54	93
4	7.14	7
5+	1.57	1

The distribution of observed V-2 strikes conformed quite closely to the Poisson distribution ($\chi^2 = 1.17, p = 0.88$). If strikes were clustered, we would have seen many more squares with a high number of bombs or none at all.

Conclusion: V-2 impact sites were **random**, not clustered.
Rocket strikes were indiscriminate (within city of London), not targeted.

Contemporary uses of geospatial data analysis

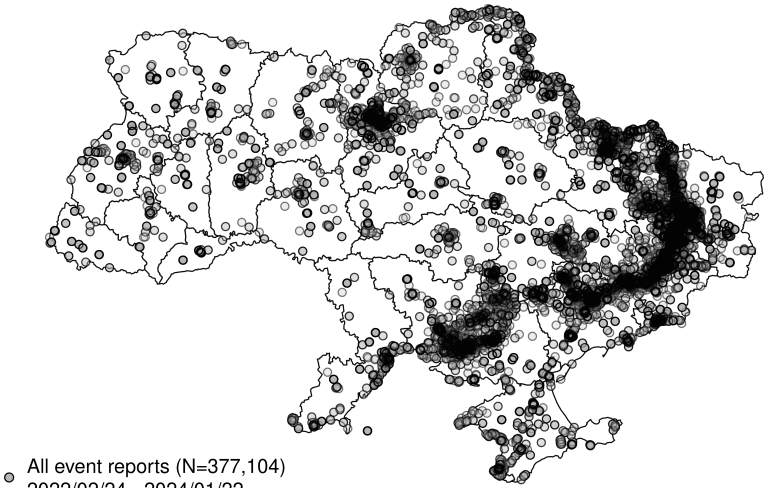


Figure 15: Example: Track violence in the Russia-Ukraine War

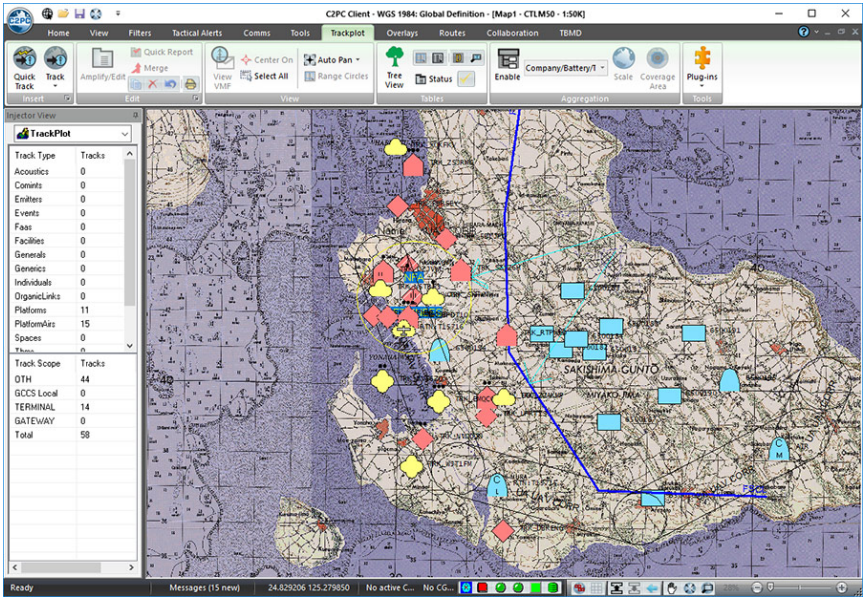


Figure 16: Example: Provide situational awareness during military operations

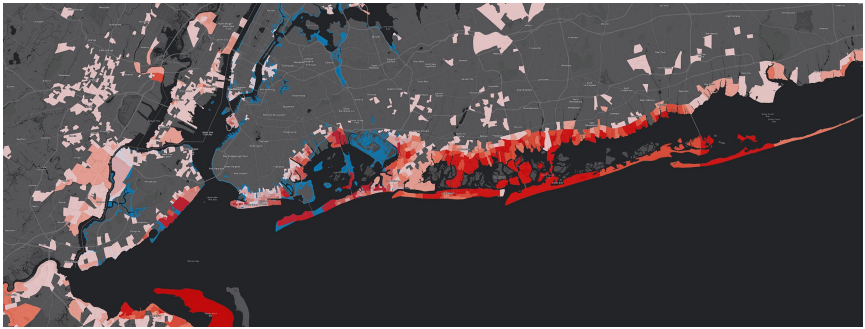


Figure 17: Example: Provide real-time information for emergency management

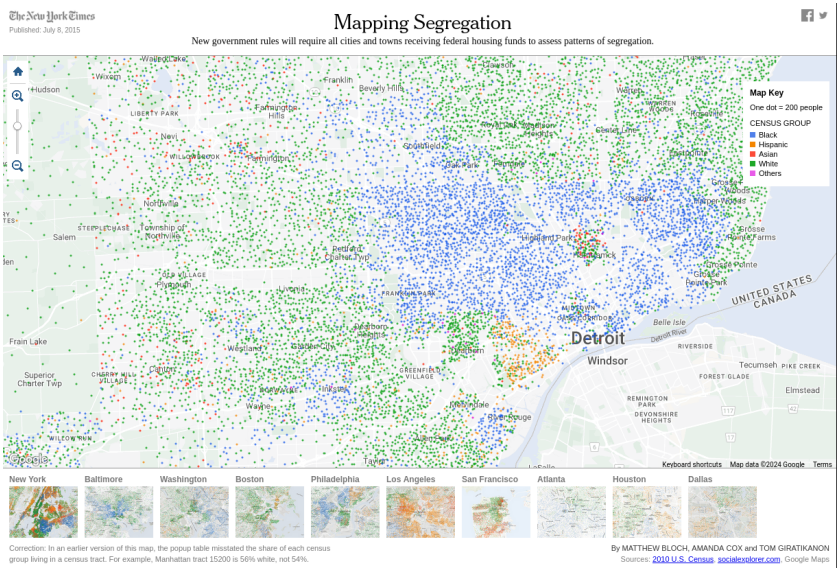


Figure 18: Example: Analyze residential segregation in American cities

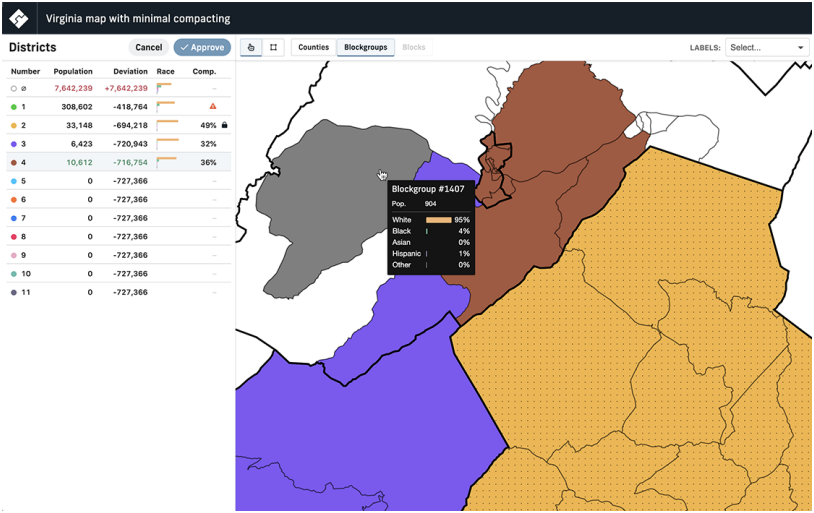


Figure 19: Example: Draw new legislative districts

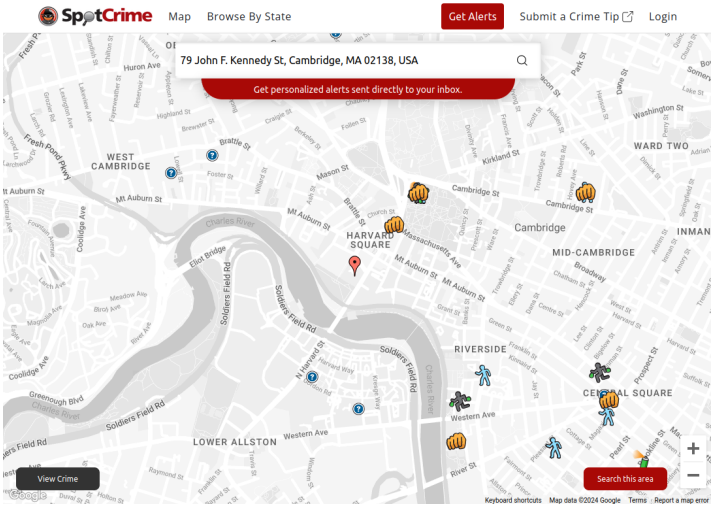


Figure 20: Example: Identify crime hotspots

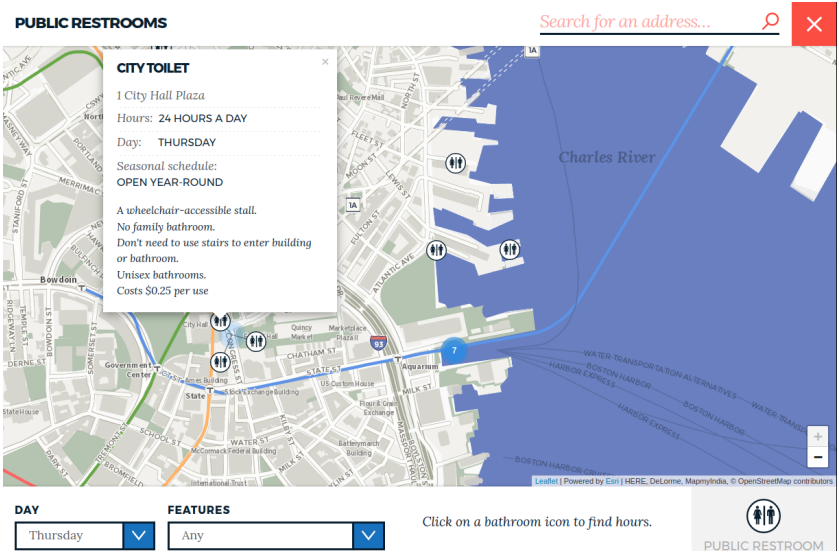


Figure 21: Example: Find a public restroom

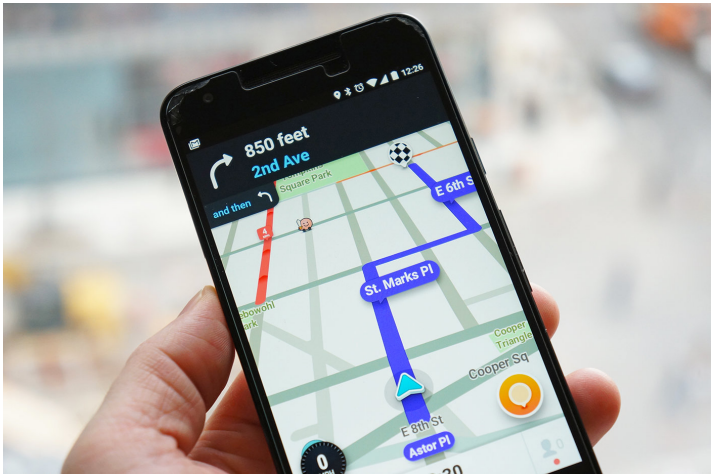


Figure 22: Example: Find your way home

About the Class

Goals of the class

1. Introduce basic GIS concepts
2. Provide hands-on experience in using open-source GIS software
3. Find, open and edit geospatial data
4. Visualize geospatial data (make cool maps)
5. Conduct basic geospatial data analyses
6. Create new geospatial data (georeferencing, geocoding)
7. Apply these skills to an original research project

How will we learn?

1. Methods boot camp
   a) first half of semester
   b) weekly lectures (45-75 min)
   c) weekly computational tutorials
   d) weekly problem sets
2. Research workshop
   a) second half of semester
   b) weekly "walk-throughs" of data collection & analysis on student-selected topics
   c) no problem sets
   d) focus 100% on research project

Figure 23: Learn new methods

Figure 24: Apply them to research

Research "walk-throughs"

1. Step-by-step guides
   a) where to find and download data
   b) how to pre-process, integrate the data
   c) how to conduct a very rudimentary analysis of the data
2. Options (students select 3 of 10)
   a) agriculture and crop productivity
   b) Congressional redistricting
   c) climate-conflict nexus
   d) crime and policing
   e) international migration
   f) nighttime luminosity
   g) piracy and transnational shipping
   h) political repression
   i) racial and ethnic segregation
   j) Russian-Ukrainian War

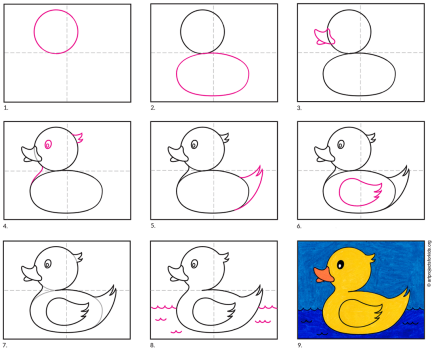


Figure 25: Like this, but for GIS

Grading

1. Problem sets (40%)
   a) $8 \times 5\%$ each
   b) due no later than 11:59 PM each Sunday
   c) collaboration encouraged
2. Final project (40%)
   a) 1-paragraph project abstract
      - due 11:59 PM, 3/8
   b) 5-minute class presentation
      - 4/23 or 4/25
   c) 5-7 page report
      - due 11:59 PM, 5/3
3. Attendance & participation (20%)
   a) show up, ask questions, help others

Figure 26: Don't worry

Final Project

1. Overview
   a) goal: use GIS to answer a *political/social/economic question*
   b) *descriptive question*: answer through mapping & visualization (e.g. "Which neighborhoods are the most violent?")
   c) *explanatory question*: answer through analysis of geospatial data (e.g. "Why are some neighborhoods more violent than others?")
   d) collaboration/co-authorship permitted
2. Project abstract (1 paragraph)
   a) summarize research idea, needed spatial & non-spatial data
3. In-class presentation (5 min, 2 slides)
   a) slide 1: Research question
   b) slide 2: Map(s)
4. Written report (5-7 pages)
   a) section 1: Research question
   b) section 2: Data & methods
   c) section 3: Preliminary results

GIS Basics

geospatial = situated in *geographic space*

space is about more than geography

- "space" refers to any dimension for which a notion of distance between objects can be defined (e.g. social networks, trade, culture, ideology)
- "geographic space" refers to Earth's surface and near-surface

geospatial data = information on "where" + "what"

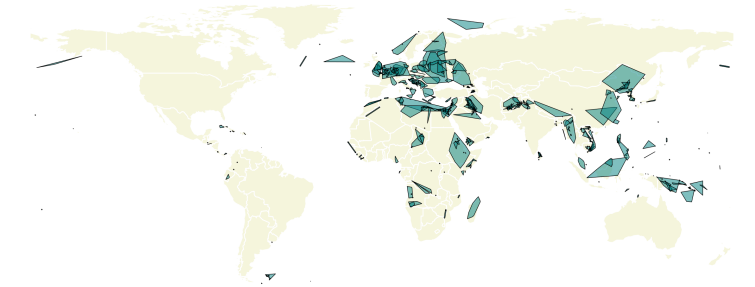
where: absolute and relative locations of features
(e.g. coordinates, distance, clustering, dispersion)
- dimension 1: x, horizontal position, longitude, easting
- dimension 2: y, vertical position, latitude, northing
- dimension 3: z, elevation, altitude, depth

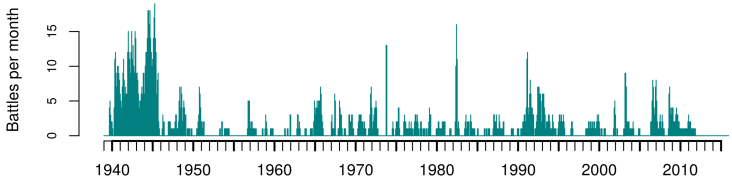
what: properties and attributes of those features
(e.g. vote share, number of fatalities, temperature)

spatio-temporal data = info on "where" + "when" + "what"

when: absolute and relative timing of observation
(e.g. year, day, electoral cycle, round)
- dimension 4: t, time

Example of multi-dimensional data / Battles in space and time





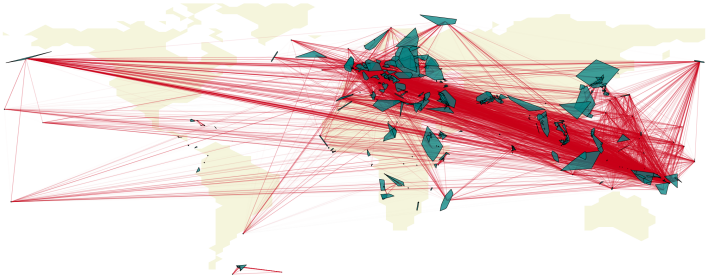


Figure 27: Let's add another dimension

Key: red lines denote pairs of battles with common participant

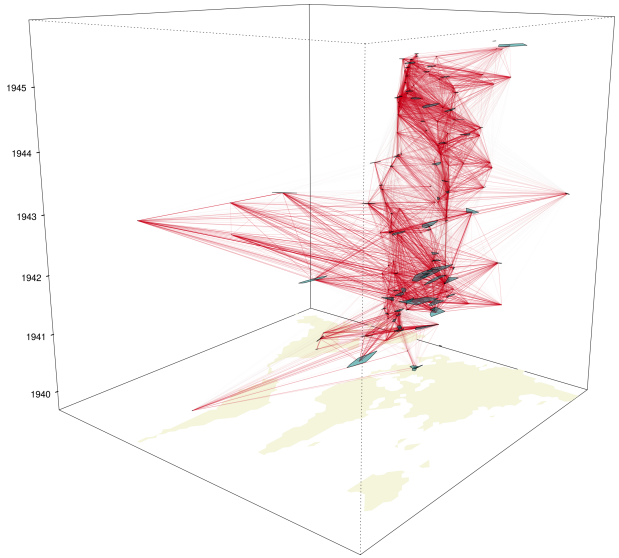


Figure 28: WWII battles in multidimensional space

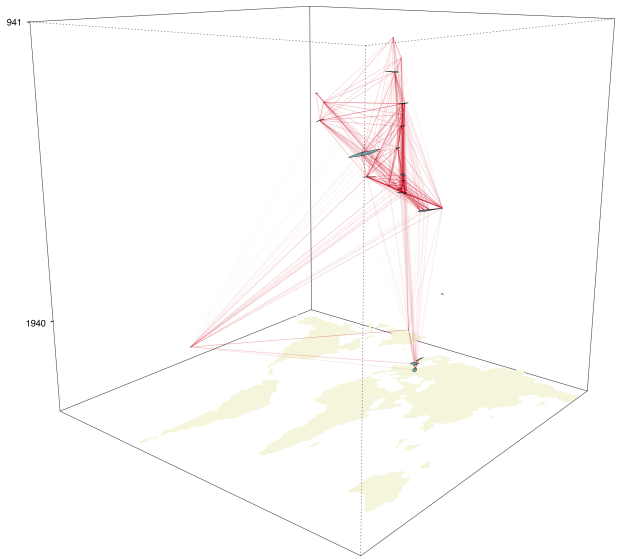


Figure 29: Battles (1939-1941), linked by combatant

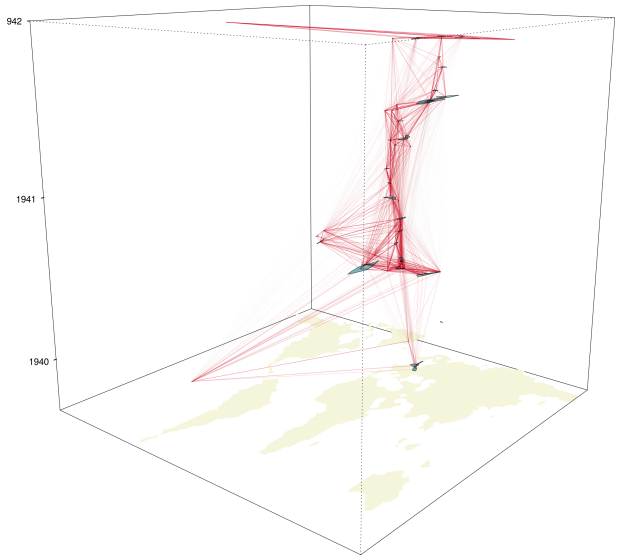


Figure 30: Battles (1939-1942), linked by combatant

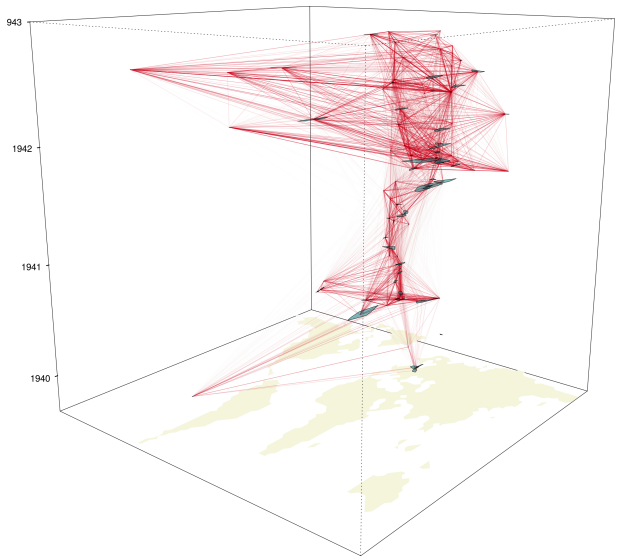


Figure 31: Battles (1939-1943), linked by combatant

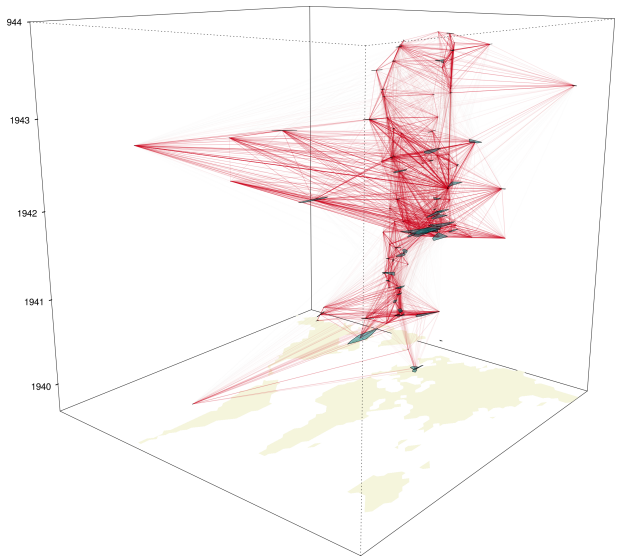


Figure 32: Battles (1939-1944), linked by combatant

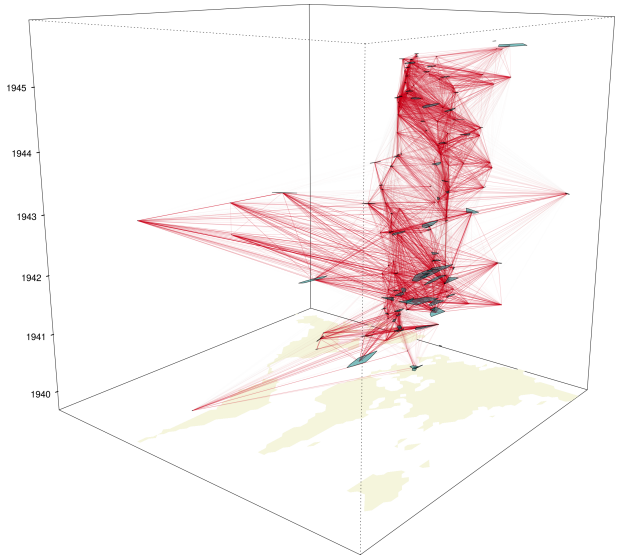


Figure 33: Battles (1939-1945), linked by combatant

Types of spatial data

Vector data

discrete objects in space

- *point*: pair of coordinates
  (e.g. small objects, events)
- *polyline*: open, connected set of points
  (e.g. roads, rivers)
- *polygon*: closed, connected set of points
  (e.g. countries, administrative units)

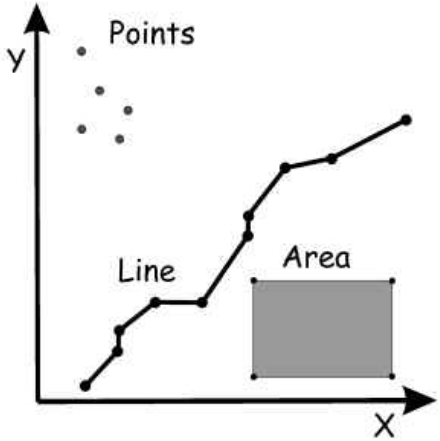


Figure 34: Vector data objects

Figure 35: Points

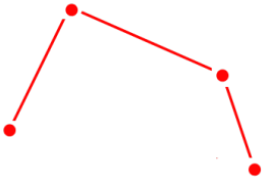


Figure 36: Polyline

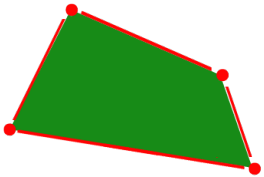


Figure 37: Polygon

Raster data

space as continuous field

- *image*: regular, equally-spaced grid
- *pixel*: individual grid cell
- each pixel represents value or presence/absence of some quantity of interest (e.g. temperature, rainfall, elevation, land cover)

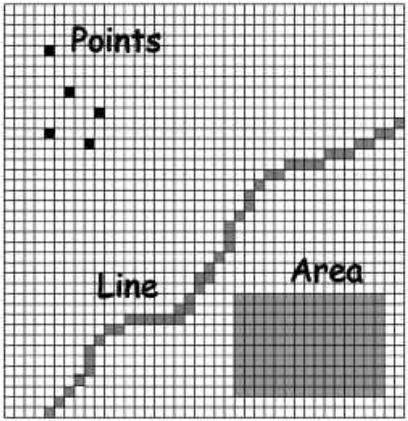


Figure 38: Raster data

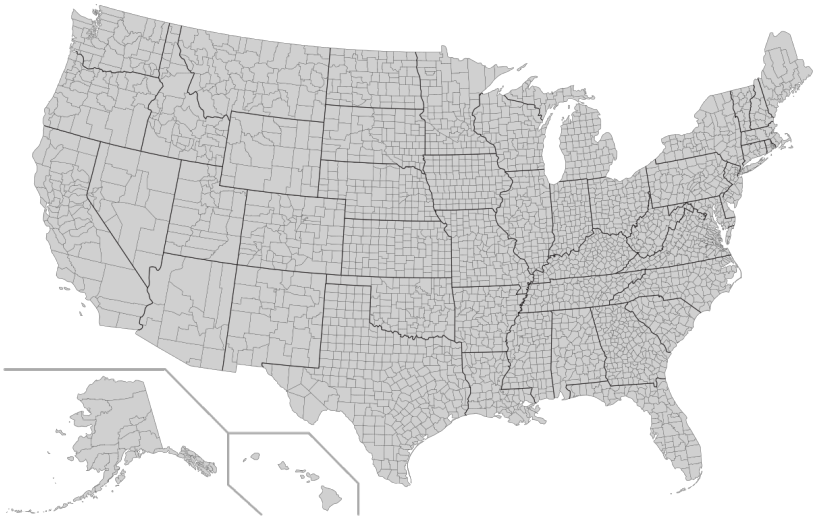


Figure 39: Vector or raster?

Figure 40: Vector or raster?

Figure 41: Vector or raster?

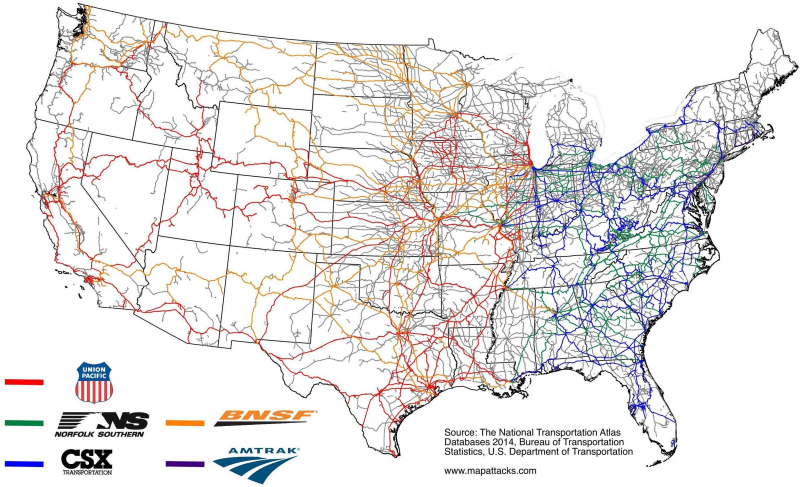


Figure 42: Vector or raster?

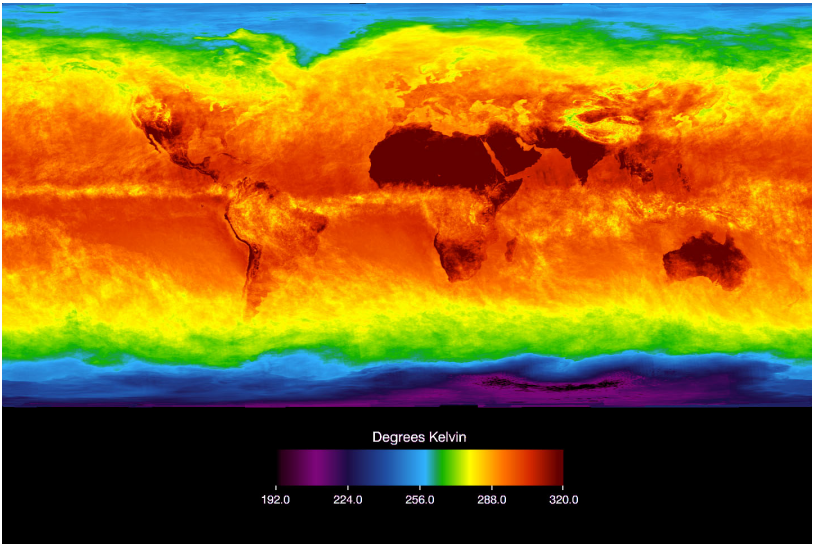


Figure 43: Vector or raster?

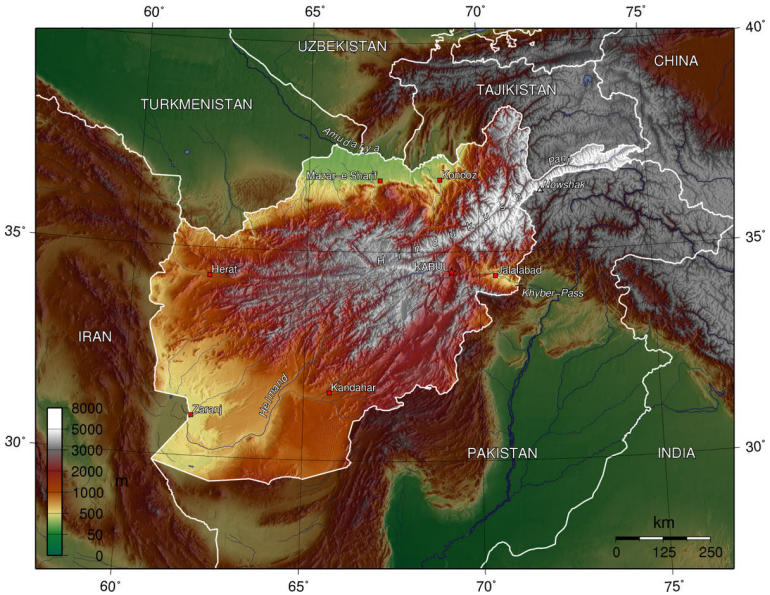


Figure 44: Vector or raster?

Where to find (free/open-source) **spatial data?**

Coordinates and basemaps:

- Geographical place names: `geonames.org`
- Global administrative units: `geoboundaries.org`, `gadm.org`
- Land cover and elevation: `www.usgs.gov/centers/eros`

Geo-referenced data:

- Geocoded U.S. Census data: `nhgis.org`
- City of Cambridge GIS data: `cambridgema.gov/GIS`
- Environmental data: `sedac.ciesin.columbia.edu`
- Armed conflict data: `x-sub.org`, `ucdp.uu.se`
- Nighttime lights and fires: `payneinstitute.mines.edu/eog/`
- Electoral districts: `electiondataarchive.org`, `cdmaps.polisci.ucla.edu`

A large number of links is also available at

- `freegisdata.rtwilson.com`
- `hgl.harvard.edu`
- `guides.library.upenn.edu/globalgis`

This is *not* a comprehensive list

Data file formats

Vector data:

- `GeoJSON` (JavaScript Object Notation) is the new standard for vector data
- but points, polylines, polygons are often stored in older `Shapefile` format
  - each `Shapefile` includes: shapes/geometries (`.shp`), positional index (`.shx`), attribute table (`.dbf`)
  - sometimes also includes: projection (`.prj`), spatial index (`.sbx`), metadata (`.shp.xml`), other elements
- other common formats include
  - `GBD/MBD` (File/Personal Geodatabase)
  - `KML/KMZ` (Keyhole Markup Language, used for Google Earth)
  - `OSM` (OpenStreetMap's XML-based file format)

Raster data:

- common formats include
  - `ASC` (ASCII delimited text file)
  - `GeoTIFF` (georeferenced TIFF image file)
  - `IMG` (ERDAS Imagine file)
  - `DEM` (Digital Elevation Model)
  - `DTED` (Digital Terrain Elevation Data)

Software options

Popular software for the analysis of spatial data

Application	Availability	Learning Curve	Key Functionality
ArcGIS	License	Medium	Geoprocessing, visualization, georeferencing
QGIS	Free	Medium	Geoprocessing, visualization, georeferencing
GRASS	Free	High	Image processing, spatial modeling
Matlab	License	High	Spatial econometrics, basic visualization
Stata	License	Medium	Spatial econometrics, basic visualization
Python	Free	High	Geoprocessing, visualization, geostatistics, spatial econometrics, point processes
R	Free	High	Geoprocessing, visualization, geostatistics, spatial econometrics, point processes

We will be using **QGIS** and **R**

Software & programming

1. QGIS (option 1)
   a) free, open-source alternative to ESRI ArcGIS
   b) visualize, manage, edit, analyze spatial data, create maps
   c) intuitive graphical user interface (GUI)
   d) multiplatform (runs on Linux, Mac, Windows, Android)
   e) download it here: `qgis.org`

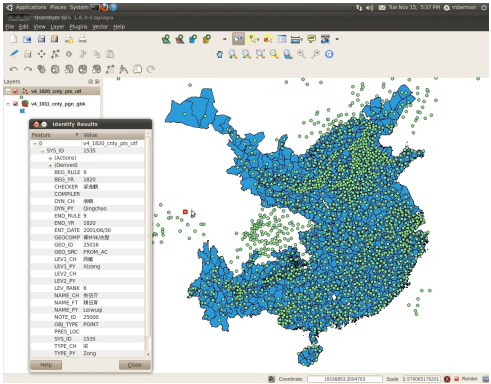


Figure 45: QGIS

Software & programming

2. R (option 2)
   a) open-source statistical programming language
   b) can do (most) of what you can do in QGIS, and lots more
   c) can run R from the command line
      ... or using source code editor (e.g. Sublime Text, XEmacs)
      ... or using integrated development environment (e.g. RStudio Cloud)
   d) also multiplatform (runs on Linux, Mac, Windows, Android)
   e) download R here: `r-project.org`
      ... or RStudio here: `posit.co`

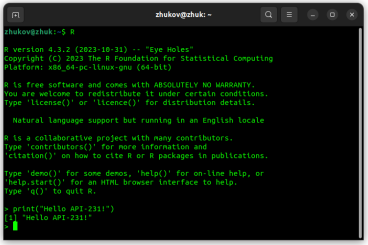


Figure 46: R

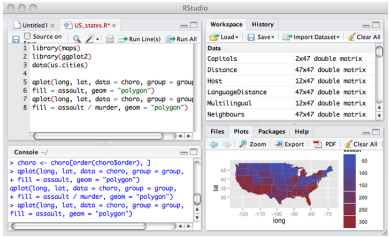


Figure 47: RStudio

3. RStudio Cloud (option 2.5)
   a) same as RStudio, but accessible through web browser
   b) advantages:
      - packages/dependencies already installed
      - no software to download
   c) all R lab exercises will be made available through RStudio Cloud
   d) you can access it through link posted on Canvas
   e) set up RStudio Cloud account w/ your `harvard.edu` credentials
   f) link to sign-up page: `posit.cloud`

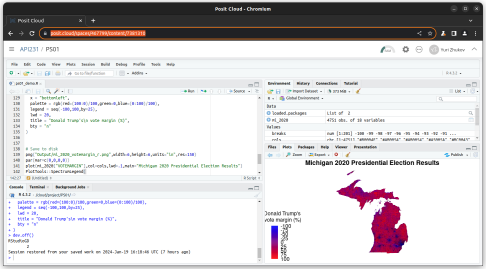


Figure 48: RStudio Cloud

Geospatial analysis in R

Task	R Packages
Data management	`sf`, `terra`, `rgdal`, `rgeos`, `rmapshaper`
Integration with other GIS	`rgdal`, `RArcInfo`, `SQLiteMap`, `spgrass6`, `rpostgis`, `RPyGeo`, `RQGIS`, `R2WinBUGS`
Access spatial data	`RgoogleMaps`, `rnaturalearth`, `geonames`, `OpenStreetMap`
Point pattern analysis	`spatstat`, `splancs`, `spatialkernel`
Geostatistics	`gstat`, `geoR`, `geoRglm`, `spBayes`
Disease mapping	`DCluster`, `spgwr`, `glmmBUGS`, `diseasemapping`
Spatial regression	`spdep`, `spatcounts`, `McSpatial`, `splm`, `spatialprobit`, `mgcv`, `spatialreg`

Full(-ish) list: `cran.r-project.org/web/views/Spatial.html`

QGIS and R help at HKS

1. *GIS + Mapping Office Hours*
   (Th 1300-1500, HKS Library Office G-16)
   POC: `@belle_lipton`
2. *R, Python, + Programming Office Hours*
   (W 1330-1430, Library Commons)
   POC: `@james_adams`, `@james_capobianco`
3. *Introduction to GIS Workshop*
   (F, 2/9, 1330-1530, Rubenstein G-21)
   Registration: `link`
4. *Advanced Data Cleaning for GIS Workshop*
   (F, 2/16, 1330-1530, Rubenstein G-21)
   Registration: `link`