INTRODUCTION, SPATIAL DATA, AND GEOGRAPHIC INFORMATION

CLASS #1 | GEOG 215

Introduction to Spatial Data Science

Spring 2020

TODAY'S CLASS

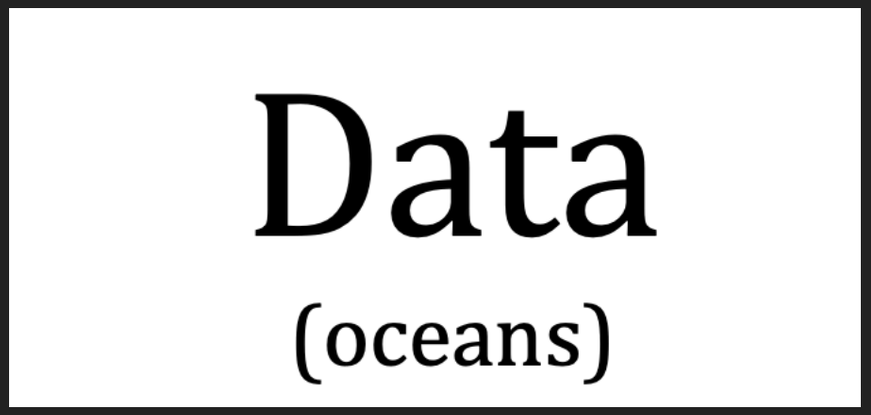
- Introduction
- Spatial Data and Geographic Information
- Course and Syllabus Overview
- Meet and Greet

WHY ARE WE HERE?

- 😩 Just because I had to be here
- Data science is all the buzz(zzz..zzzz)
- I want get a $$$$$$ paying job
- The instructor is supposed to be amazing!
- **Exciting (concerning) time to learn about (geo-) data science**

THE (GEO-) DATA REVOLUTION

- The world is being **"Datafied" -** *taking all aspects of life and turning them into data (Cukier & Mayer-Schoenberg)*
- Advances in Computation, Storage, Geospatial technology
- Data ———> Behavior ————-> Data (Data colonialism)
- Increasing challenges - privacy, misinformation, increasing inequality, issues of technological and ethical responsibilities



- **Data**: Facts and figures, usually raw

Data

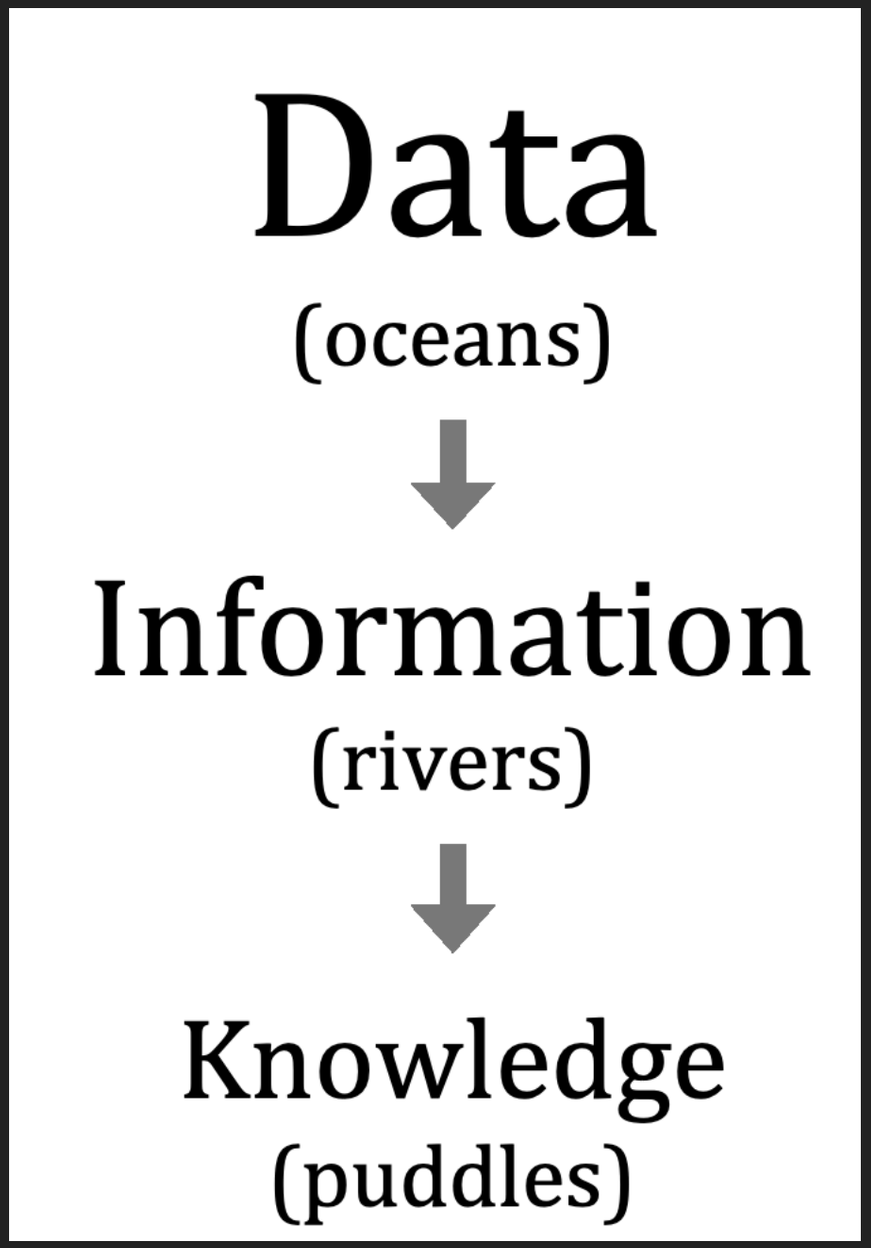
(oceans)

↓

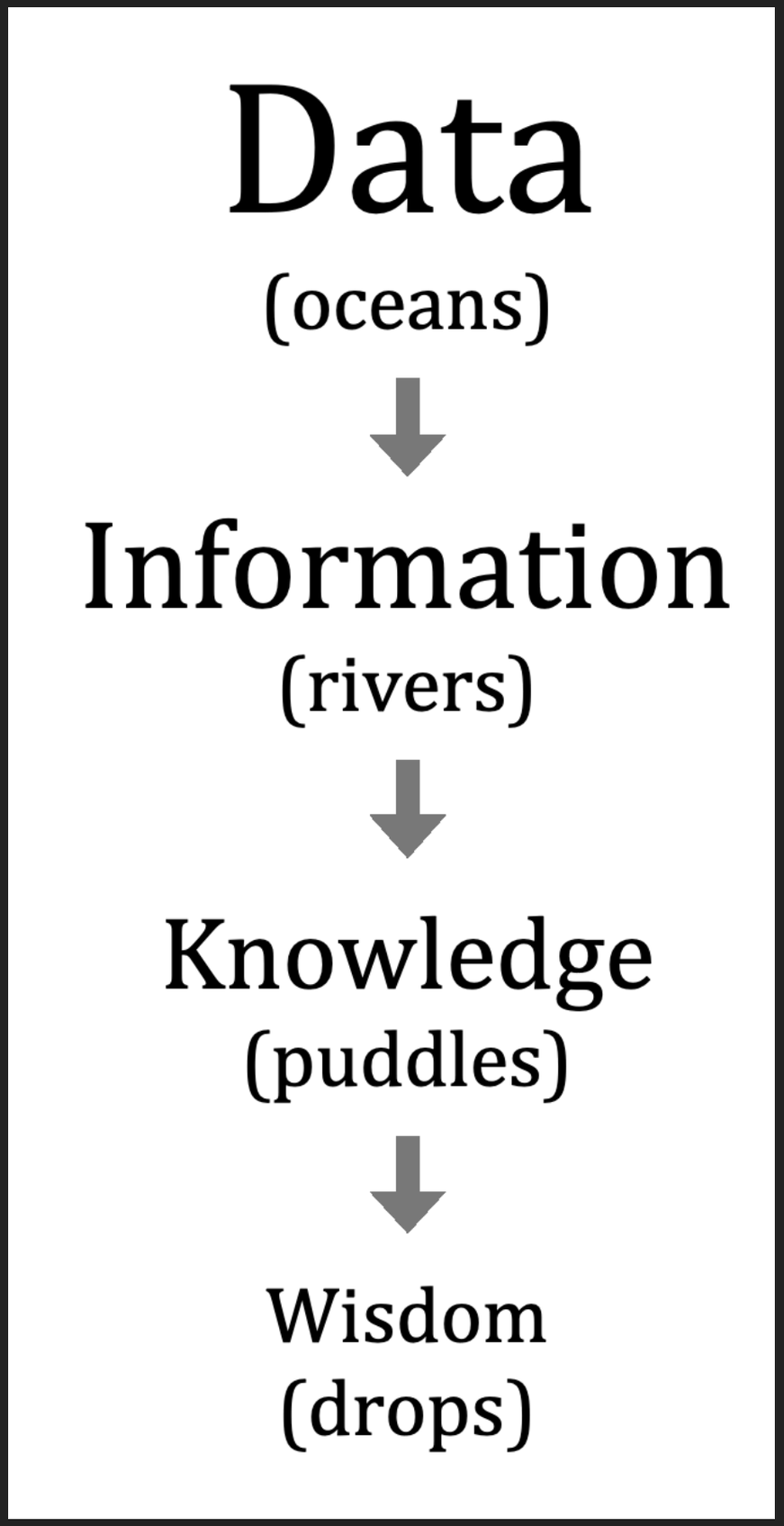
Information

(rivers)

- **Data**: Facts and figures, usually raw
- **Information**: data organized such that it is useful



- **Data**: Facts and figures, usually raw
- **Information**: data organized such that it is useful
- **Knowledge**: accumulated and integrated information on a topic over some period of time and across a broad range of situations



- **Data**: Facts and figures, usually raw
- **Information**: data organized such that it is useful
- **Knowledge**: accumulated and integrated information on a topic over some period of time and across a broad range of situations
- **Wisdom**: application of universal principles, reason, and knowledge to discern what is true and right

WHAT IS DATA SCIENCE ANYWAY?

- “Not a Science”
- “An over-hyped phrase”
- “Big Data"
- “Terminator style doomsday”
- “Just a cool name for Statistics”

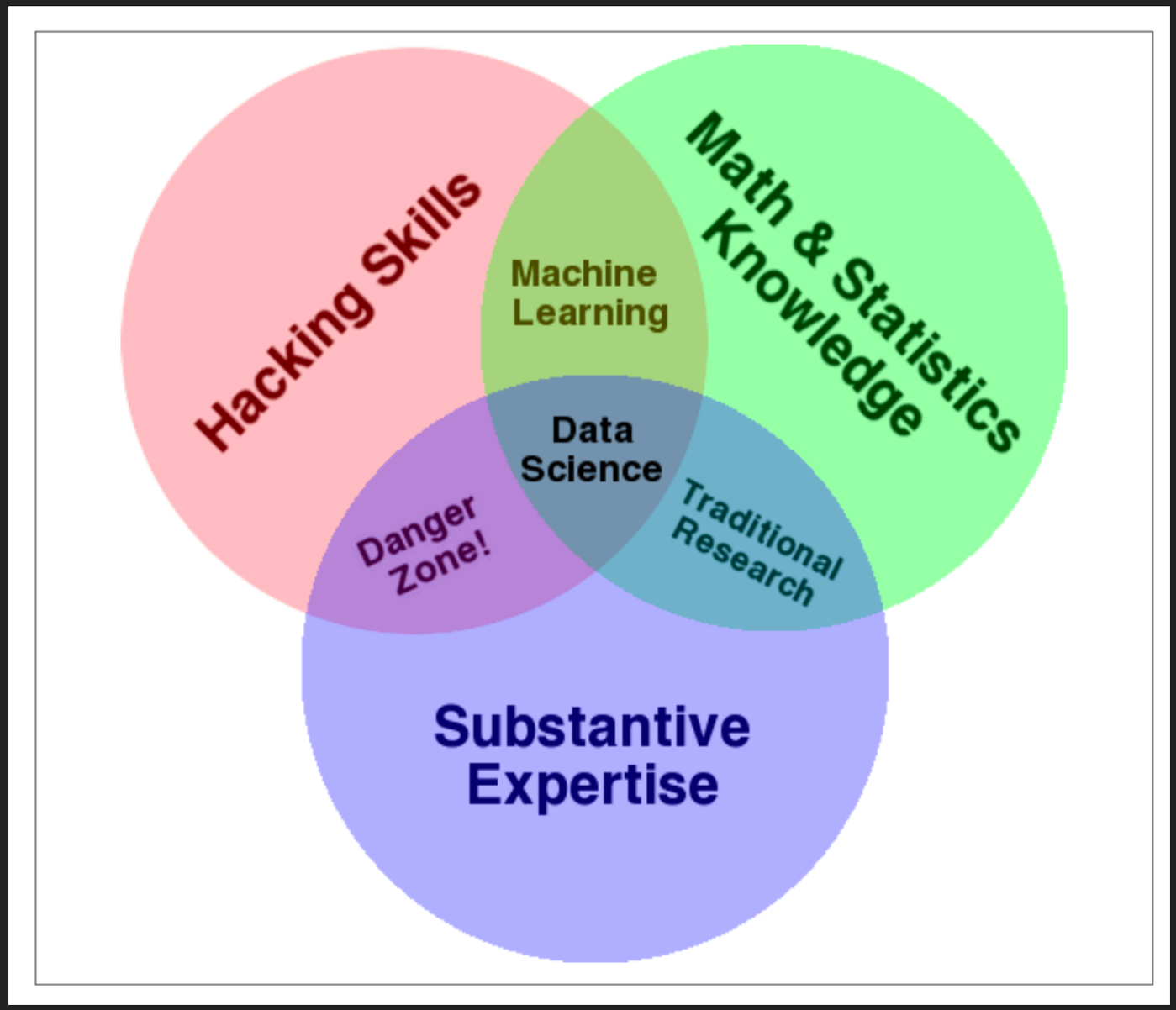
GETTING PAST THE HYPE

"An exciting discipline that allows you to turn raw data into understanding, insight, and knowledge." - *Hadley Wickham*

Data science is the development of methods and techniques to extract information or knowledge from data -

"Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics." - John Driscoll

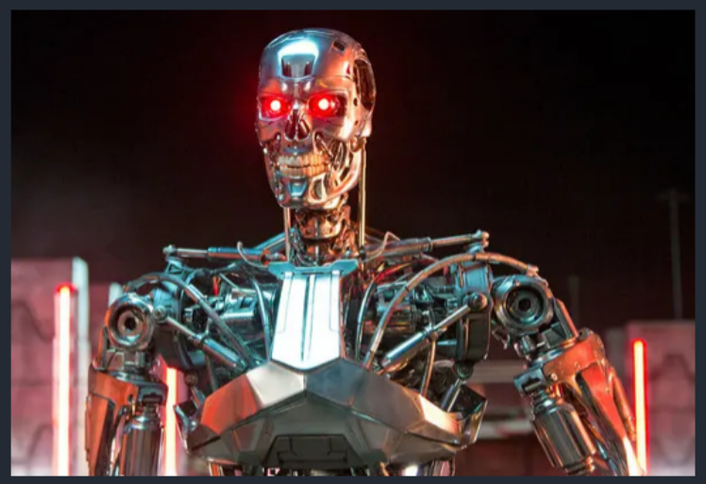
DATA SCIENCE : THINGS TO CONSIDER



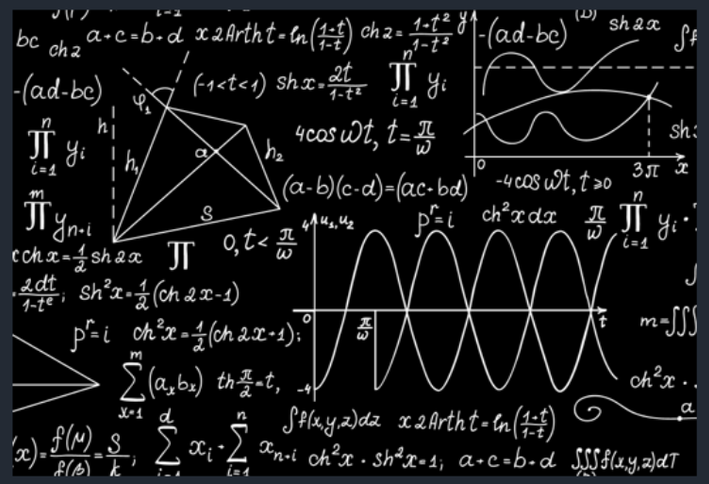
Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

WHAT DO DATA SCIENTISTS DO?

Data Scientist

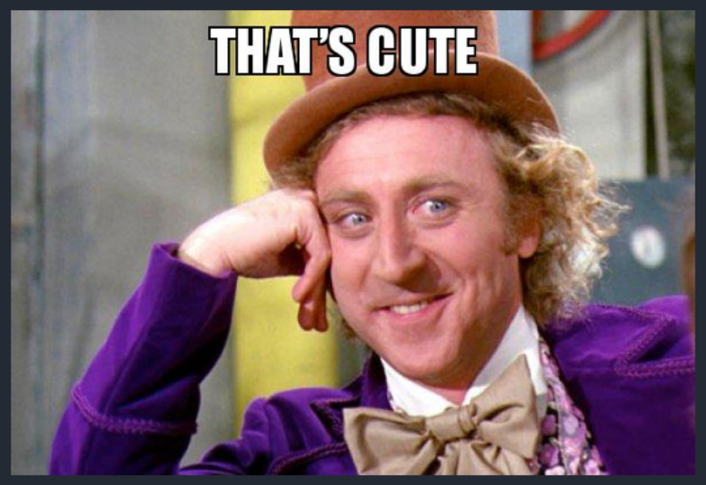


What society thinks I do



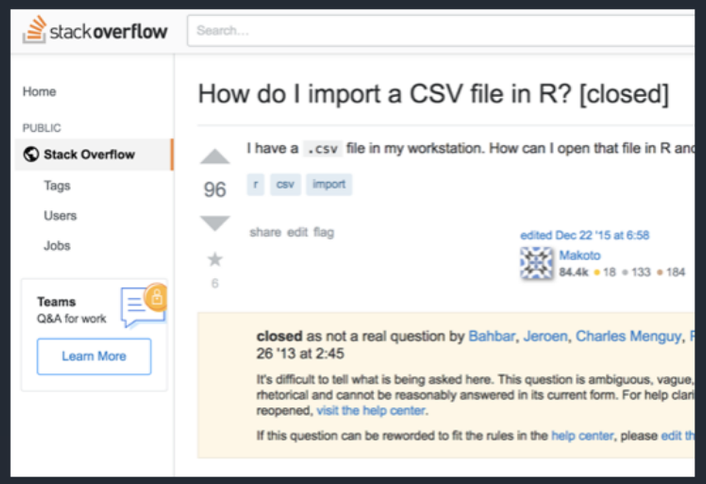
What my friends think I do

What my boss thinks I do



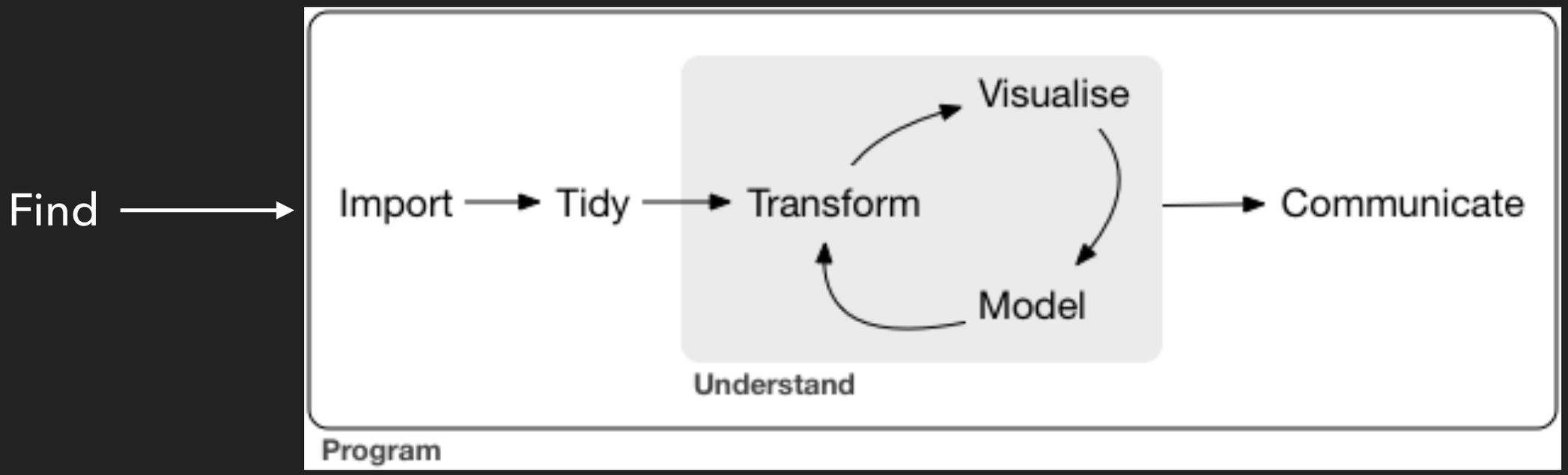
What mathematicians think I do

What I think I do



What I really do

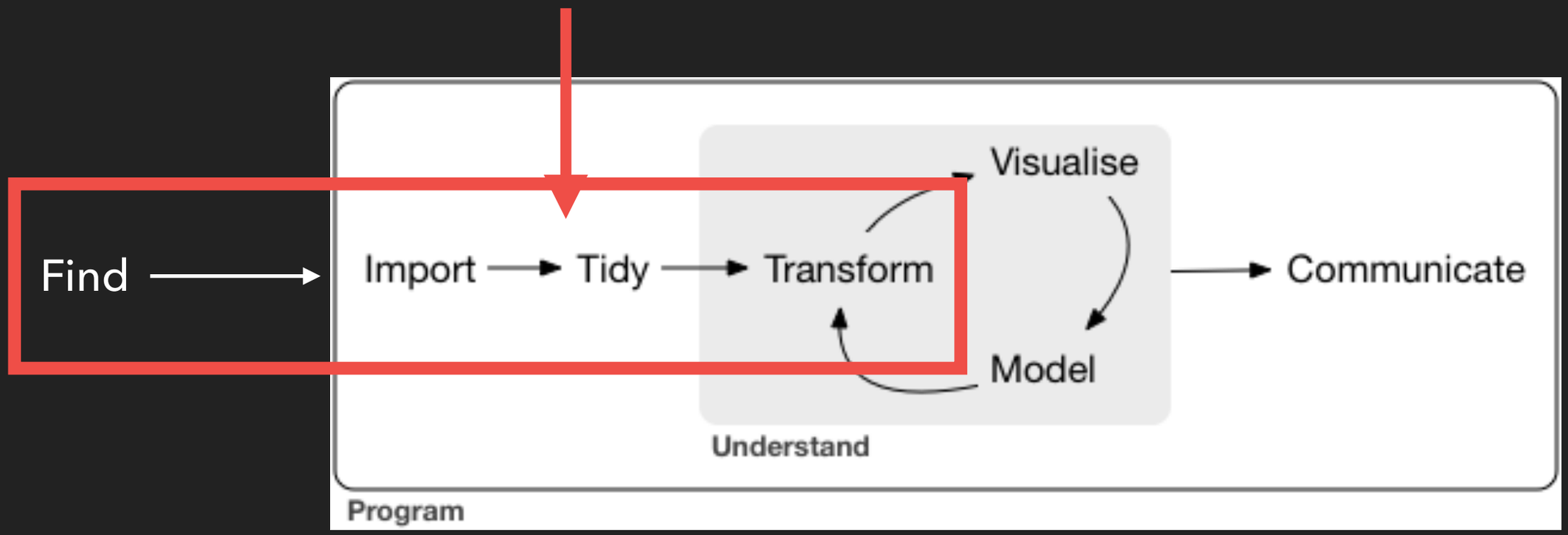
DATA SCIENCE PROCESS



Source: R for Data Science

DATA SCIENCE PROCESS

"The most time consuming and least fun"



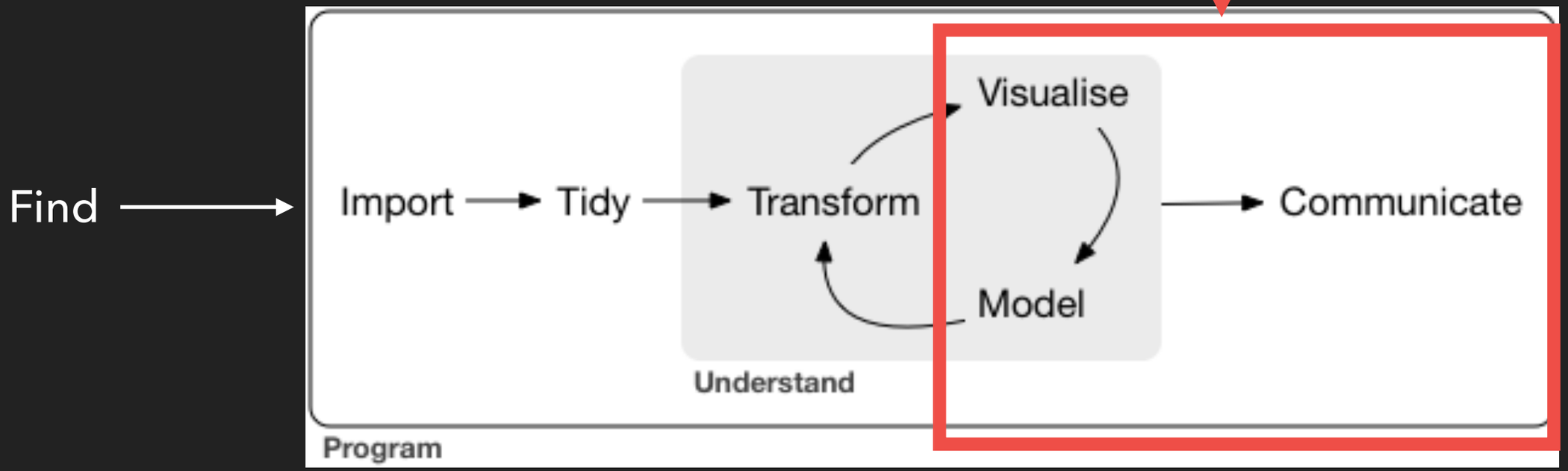
Source: R for Data Science

DATA SCIENCE PROCESS

"Data science is 80% preparing data, 20% complaining about preparing data"- *Almost every data scientist*

DATA SCIENCE PIPELINE

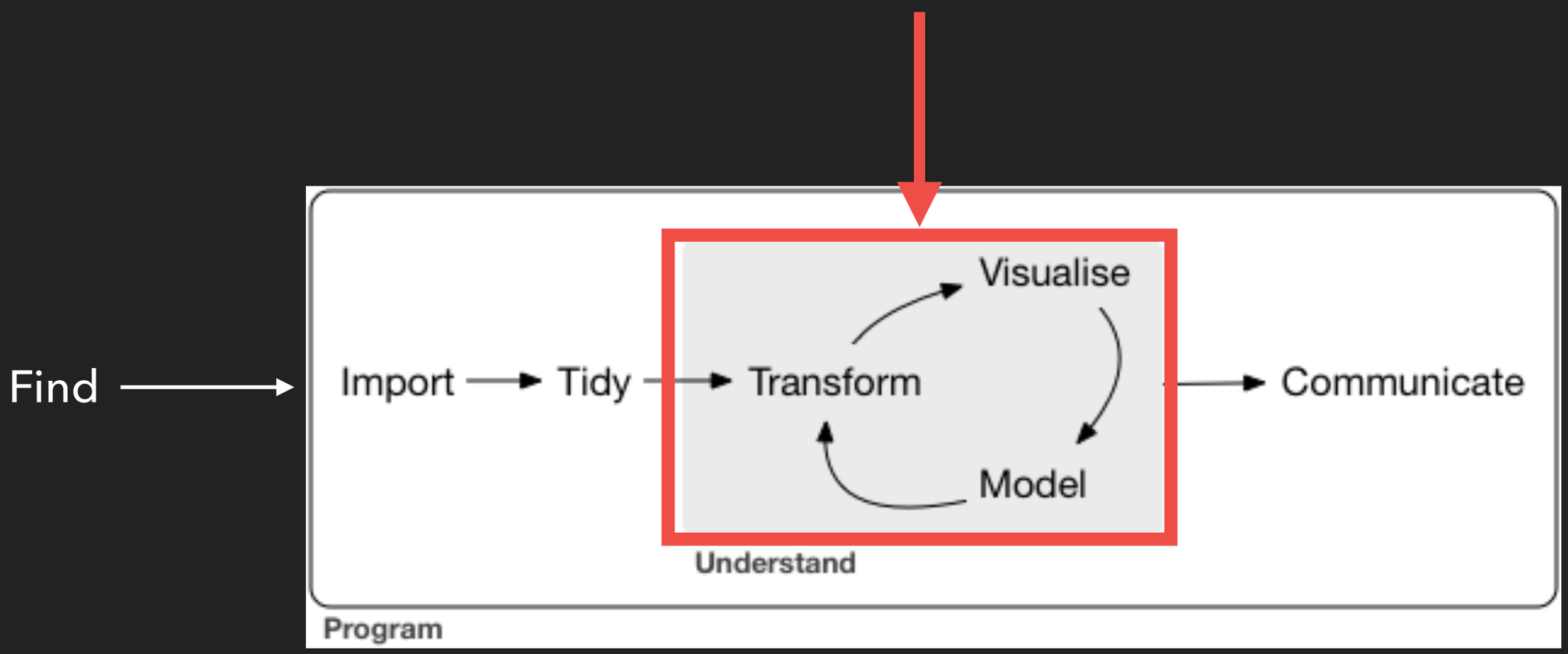
"The least time consuming and most fun"



Source: R for Data Science

DATA SCIENCE PIPELINE

"Where the MAGIC happens"



Source: R for Data Science

WHAT IS (GEO–) SPATIAL DATA SCIENCE ?

- Similar to (subset of) Data Science
  - Focus on observations/data that are geo-referenced to a location on Earth's surface
    - Have a defined location
- Focus on answers and understanding spatial/geographic questions, e.g.,
  - Where?
  - Why there?

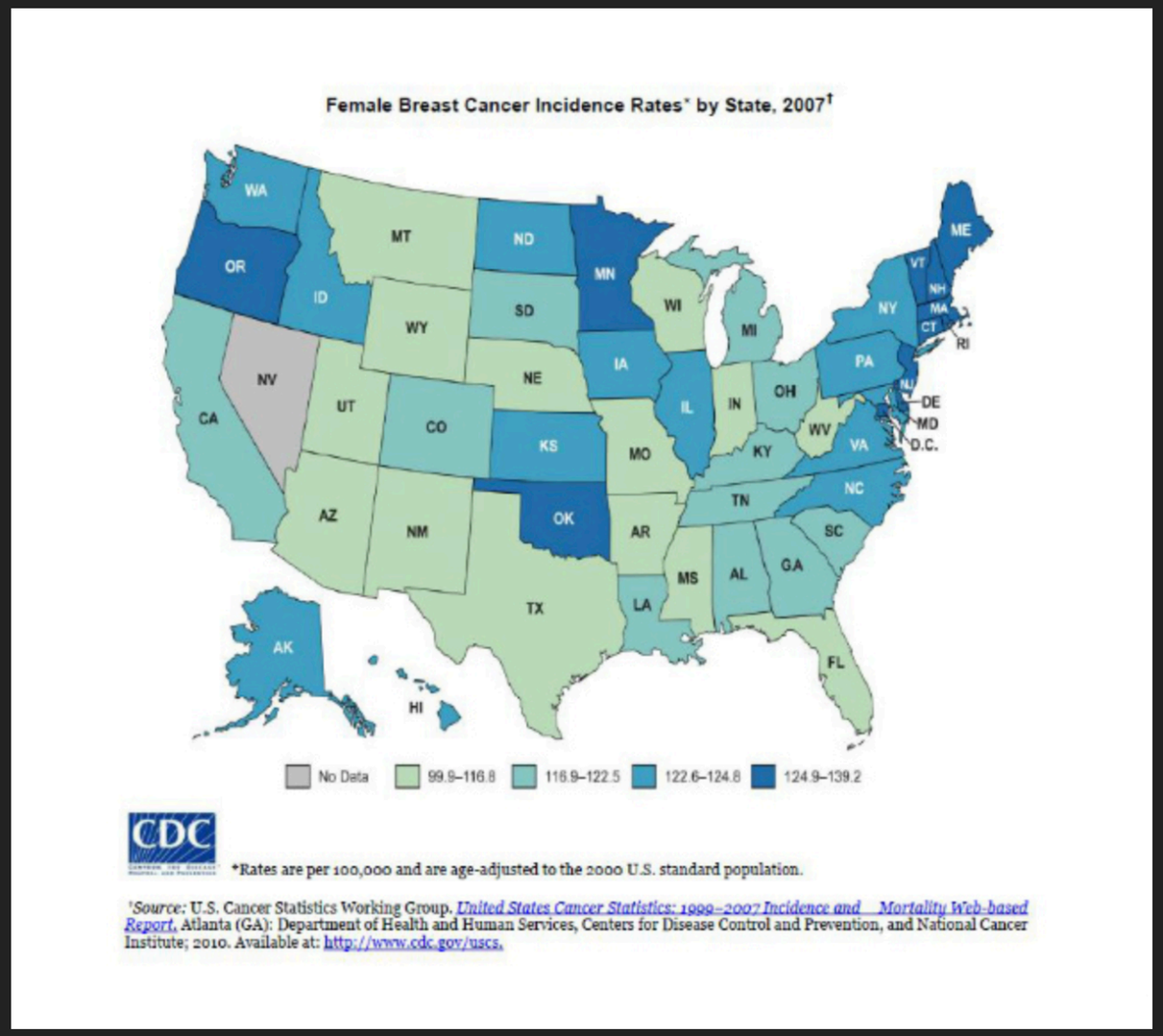
SPATIAL DATA IS SPECIAL

- Linked geometric and tabular data
  - Position and attributes
    - Geometric data is NOT the same as adding a “location” name or coordinate values
- Enables simultaneous spatial and statistical analysis
  - May provide causal insights that otherwise won’t be visible

IS THIS GEOMETRIC DATA

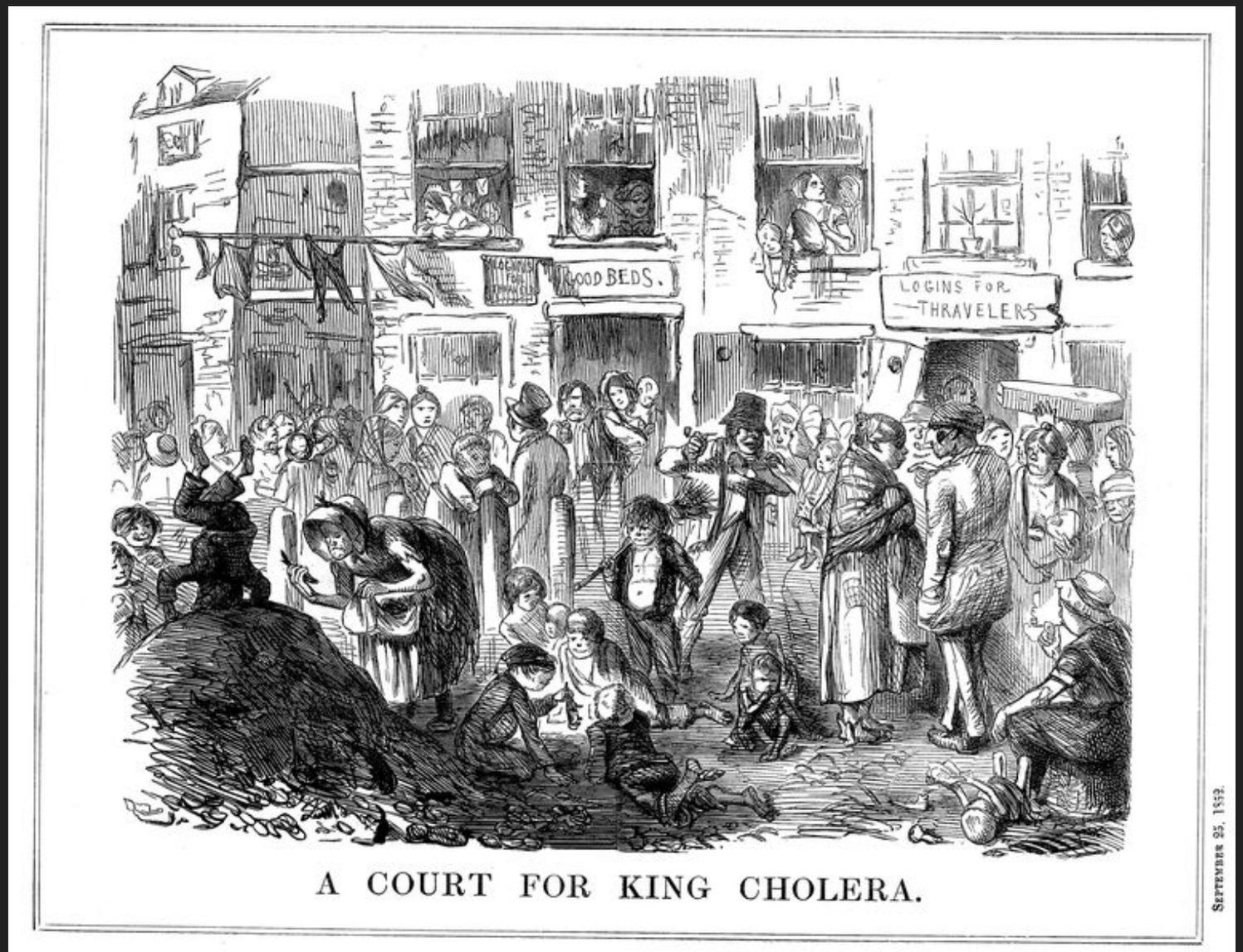
Table 3
Number of Infant Deaths, Live Births and Infant Death Rates by Michigan and Michigan County of Residence, 2017 and 2013 - 2017 Average

County of Residence	2017			2013 - 2017		
	Infant Deaths	Live Births	Infant Death Rate	Average Infant Deaths	Average Live Births	Average Infant Death Rate
MICHIGAN	762	111,507	6.8 ±0.5	770.2	113,256.8	6.8 ±0.2
Alcona	1	58	*	0.4	59.0	*
Alger	-	63	-	-	62.0	-
Allegan	4	1,348	*	7.4	1,374.0	5.4 ±1.7
Alpena	2	290	*	1.2	271.4	4.4 ±3.5
Antrim	-	161	-	0.6	192.6	*
Arenac	-	113	-	1.2	123.8	9.7 ±7.7
Baraga	1	69	*	1.0	72.6	*
Barry	5	635	*	2.6	633.8	4.1 ±2.2
Bay	7	1,011	6.9 ±5.1	5.4	1,035.4	5.2 ±2.0
Benzie	1	159	*	1.2	156.4	7.7 ±6.1
Berrien	8	1,728	4.6 ±3.2	11.2	1,782.0	6.3 ±1.6
Branch	5	530	*	4.2	539.6	7.8 ±3.3
Calhoun	15	1,536	9.8 ±4.9	11.0	1,641.6	6.7 ±1.8
Cass	3	507	*	3.4	495.4	6.9 ±3.3
Charlevoix	2	217	*	1.2	226.6	5.3 ±4.2
Cheboygan	2	229	*	0.8	211.8	*
Chippewa	1	324	*	1.2	344.8	3.5 ±2.8
Clare	3	331	*	1.6	326.4	4.9 ±3.4
Clinton	2	824	*	4.0	814.8	4.9 ±2.1
Crawford	1	138	*	1.4	122.4	11.4 ±8.4



A MOTIVATING EXAMPLE

Year 1854, Location: London

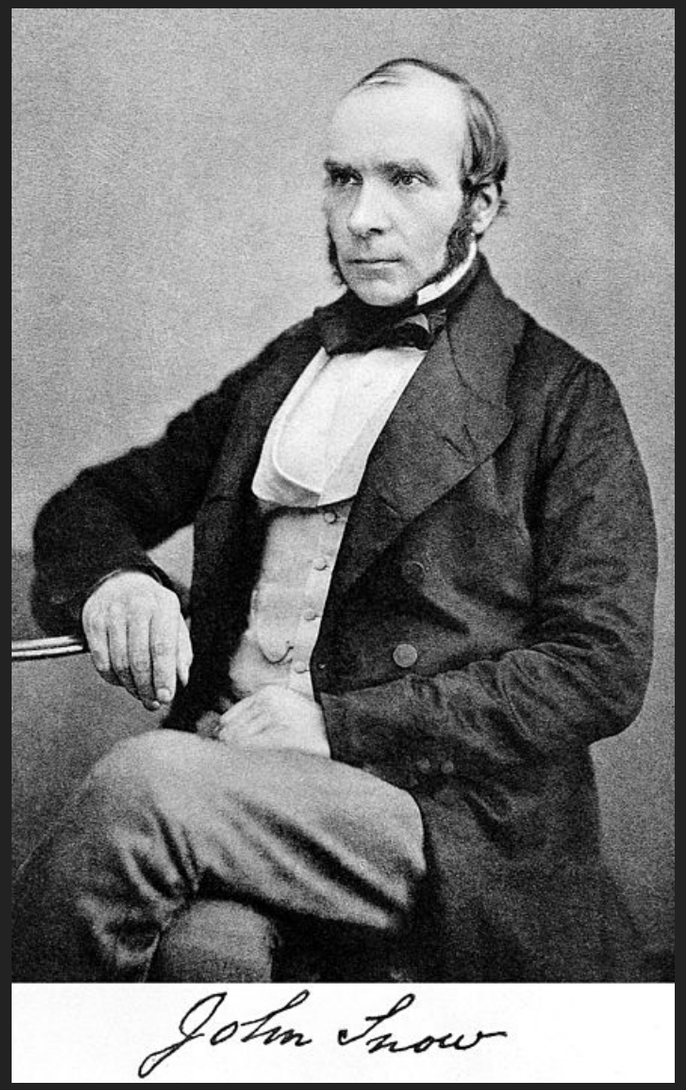


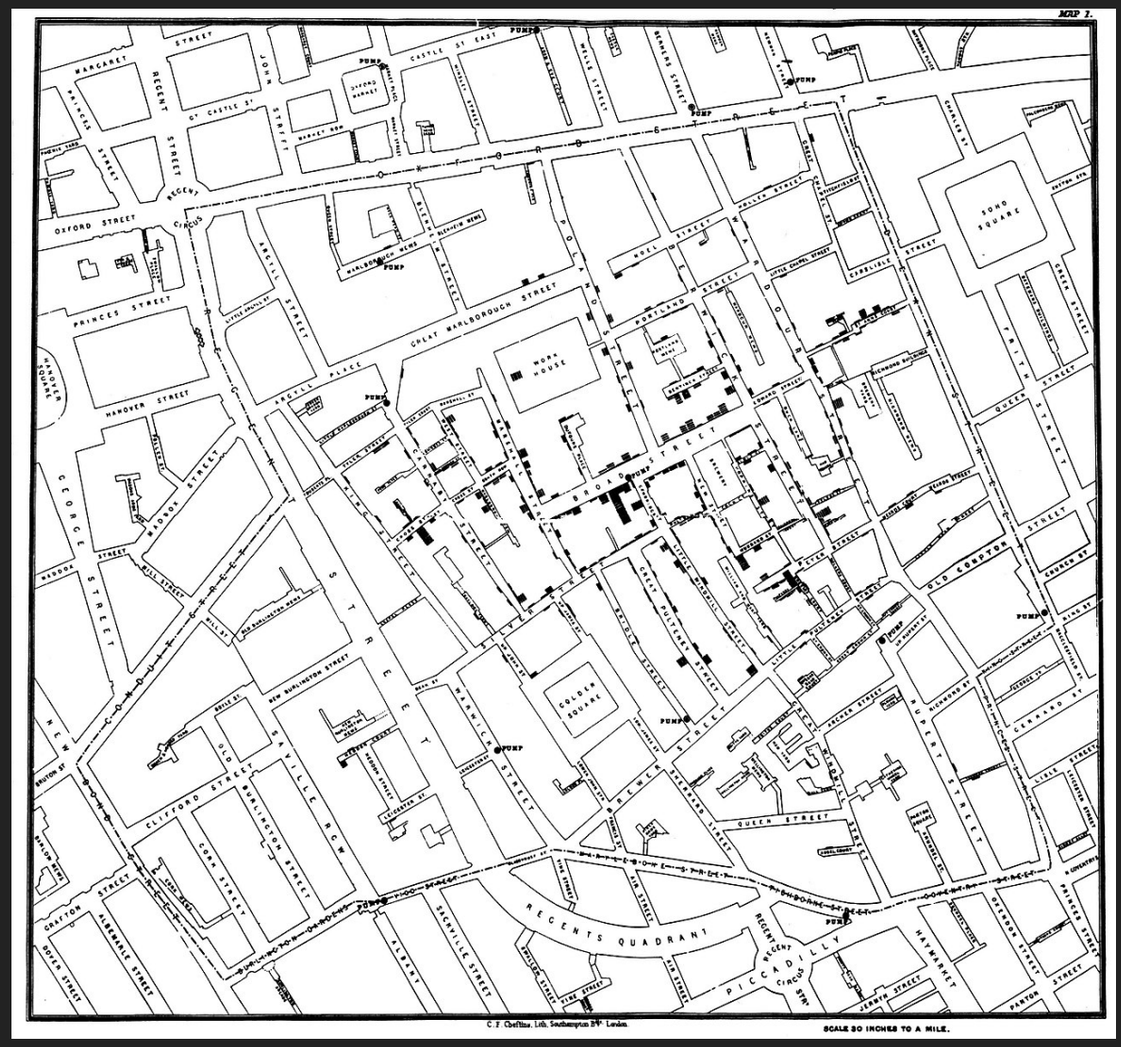
A COURT FOR KING CHOLERA.

SEPTEMBER 23, 1832.

A MOTIVATING EXAMPLE

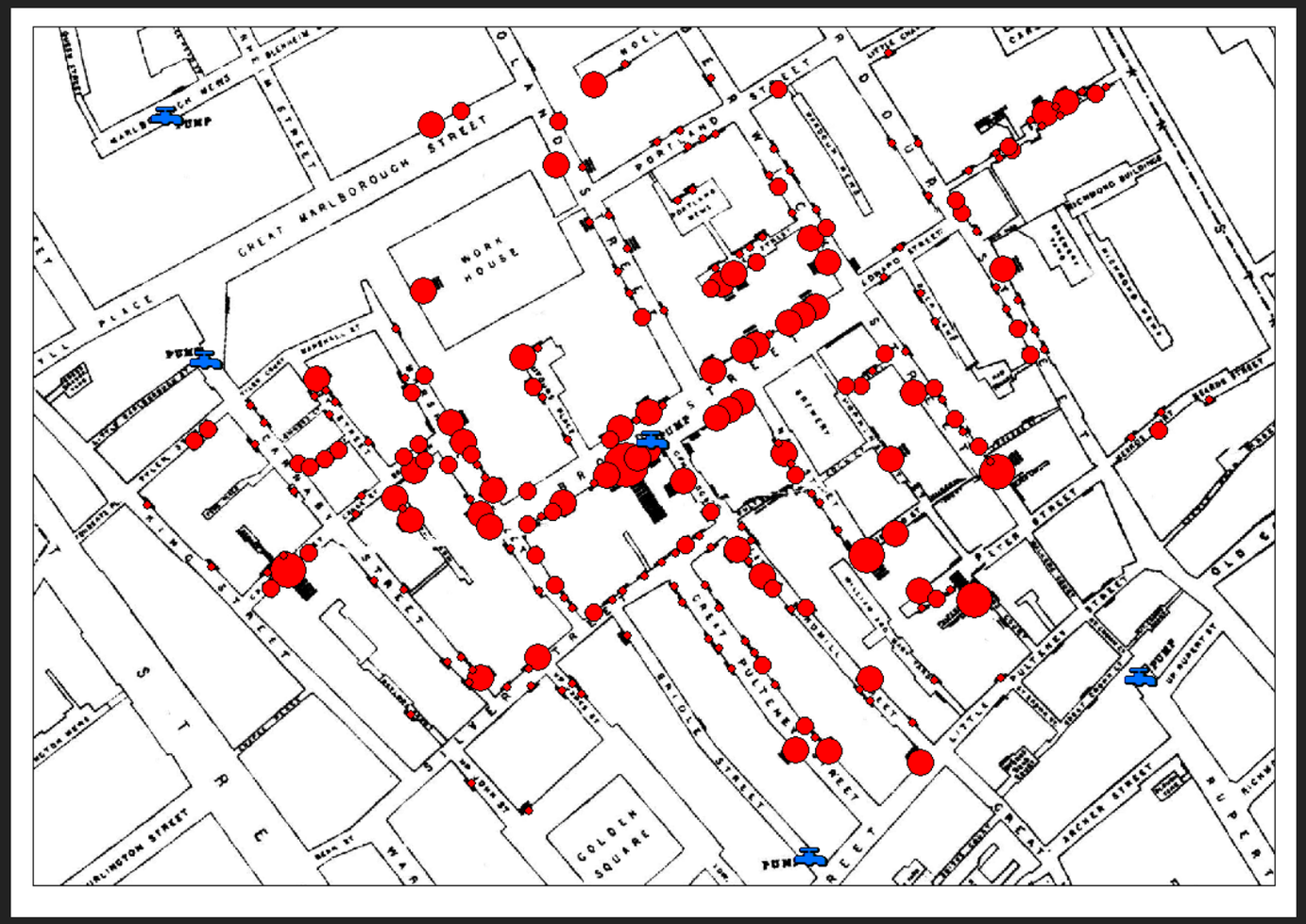
Enter John Snow



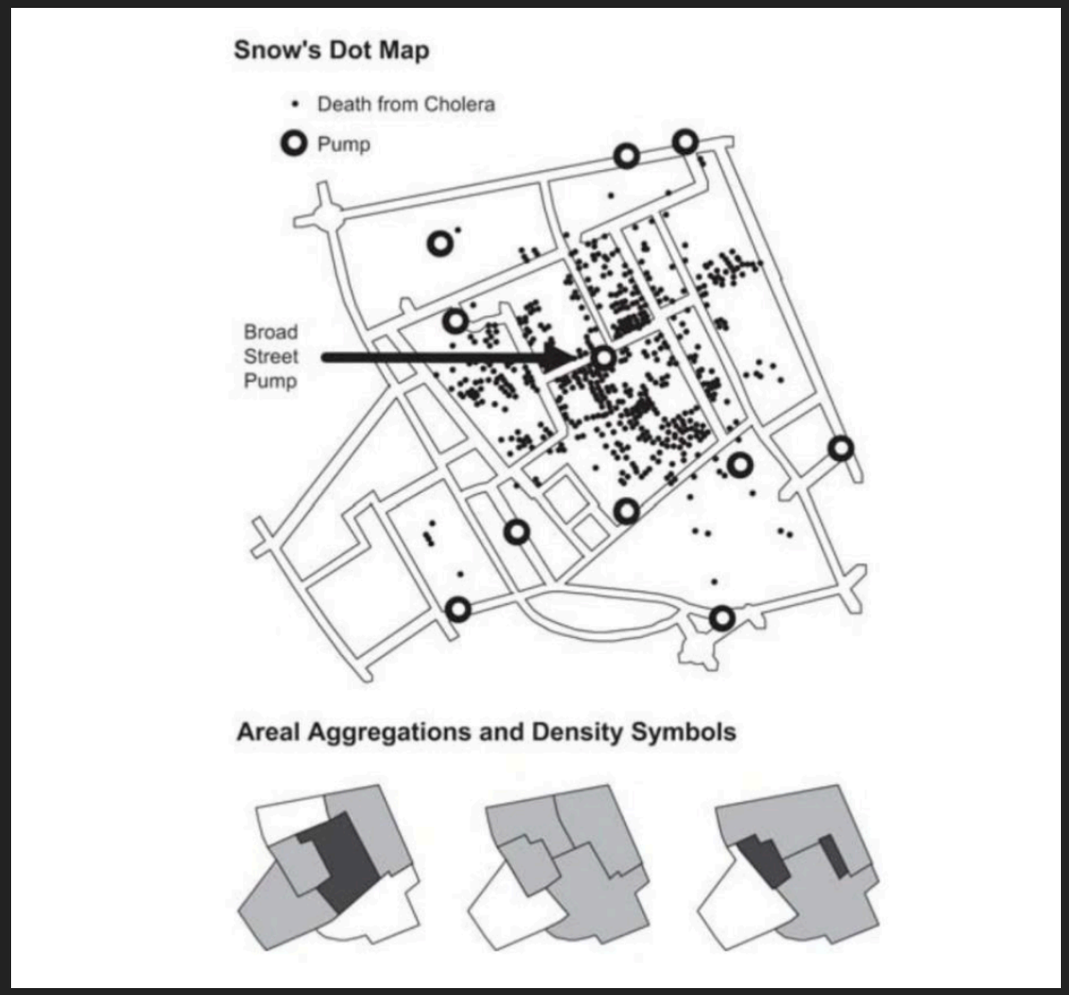


A MOTIVATING EXAMPLE

Year 1854, Location: London



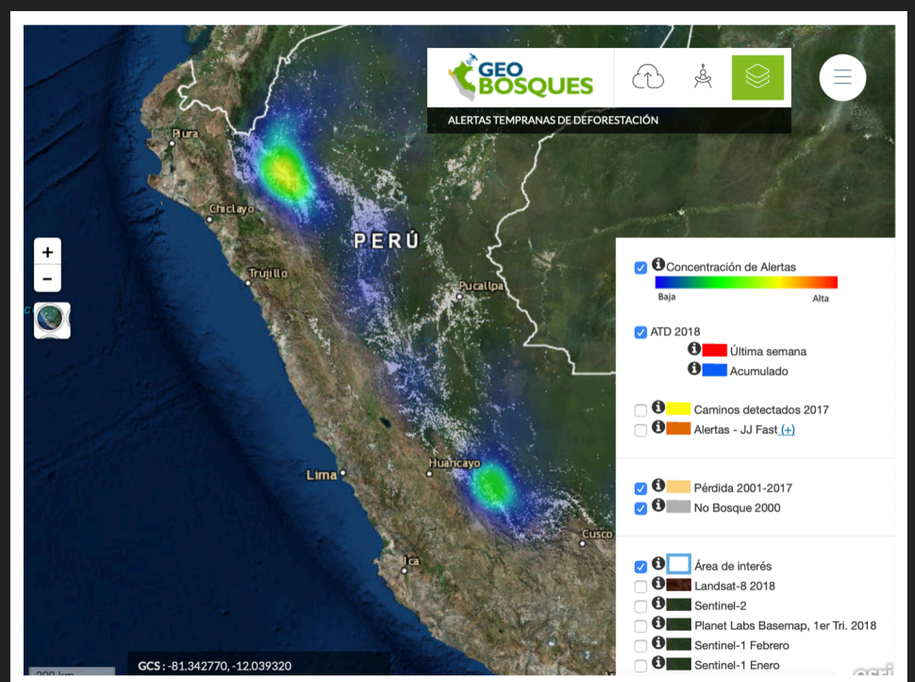
SPATIAL DATA IS CHALLENGING



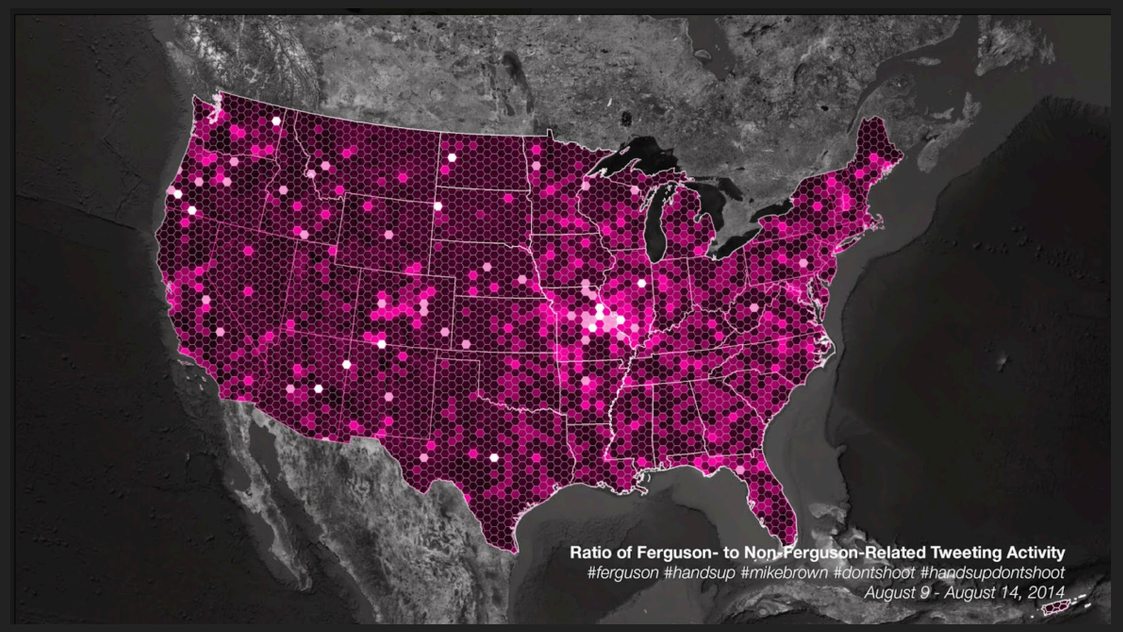
Example of the ***Modifiable Areal Unit Problem (MAUP)***

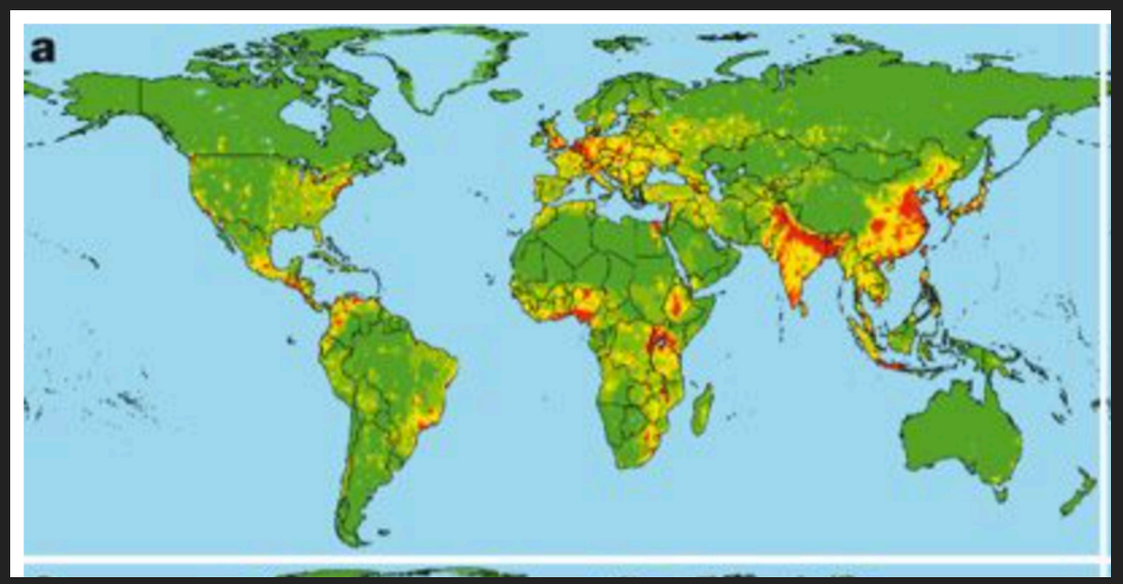
Source: Cromley & Mclafferty, GIS & Public Health, 2015

APPLICATIONS



GeoBosques, provides an alert system in areas that could be deforested.





COMPONENTS OF SPATIAL DATA

- Location
  - Often in 2-D space, but can be 3-D
- Attribute
  - Some measurable or observable property
- Time
  - Maps can be a snapshot, or be variable over time
- Metadata
  - Describes the data and assumption

QUESTIONS SO FAR?

COURSE OVERVIEW

- Mainly divided into 2 parts
  - Part 1 - “Spatial Data and Data Science techniques”
    - Course material focuses on the tenets of spatial data
    - Labs focus on core data science programming skills
  - Part 2 - “Exploratory Spatial Data Analysis”
    - Focus on spatial analysis theory
    - Labs focus on the spatial aspect of programming

PHILOSOPHY

- (Lots of) methods and techniques (class, readings and labs)
  - General overview
  - Intuition
  - Very little math (we will introduce statistical principles where needed)
  - Lots of ways to continue on your own
- Emphasis on the application and use (labs and homework)
- Close connection to “real world” applications (homework and projects)

PROGRAMMING (IN R AND R MARKDOWN)

- Do not be afraid
- Is a highly valuable skill
- Helps you to think logically
- Increases reproducibility
- Will make you more efficient
- Will allow you to ask and answer more challenging and interesting questions!

Intro to Spatial Data Science

GEOG 215

Welcome to the official website of *Introduction to Spatial Data Science* for Spring 2020 at the University of North Carolina Chapel Hill

- **Instructor:** Varun Goel |
- **Teaching Assistant:** Prabisha Shreshta
- **Graduate Research consultant:** Kate Brandt
- **Lectures:** Monday 3:35 pm - 4:50 pm
- **Labs:** Wednesday 3:35 pm - 4:50 pm,
- **Location:** Carolina Hall 204
- **Class Discussion Page:** **Click here**

HOW DOES THIS COURSE FIT IN GEOGRAPHY

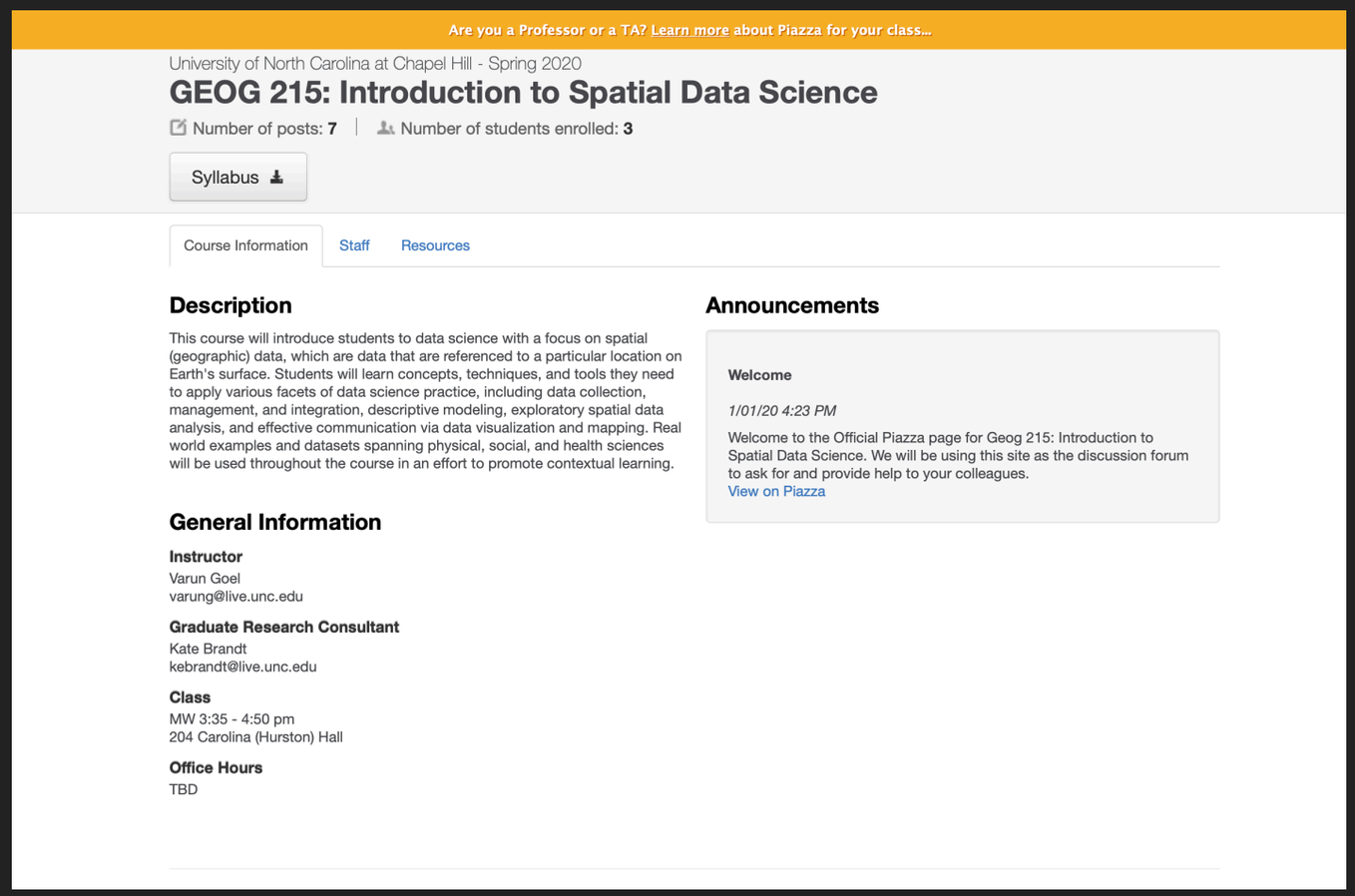
- Not quite a Quantitative Methods course
  - Not quite a Statistics Course
  - Not quite a GIS Course
  - A little bit of all them
  - Main focus on programming and pipeline required to acquire, handle, and process spatial data using Exploratory Spatial Analysis)

SAKAI

- NOT my favorite platform for this course
- Primarily for submission of assignments and grades
- I will initially send class emails through Sakai (then transition to Piazza)

PIAZZA

- Our go to online class discussion platform (FERPA compliant)
- Engagement on Piazza WILL be considered for participation grade
- All announcements will be posted there



CLASSES

- Lectures
  - Slides posted on course website after class
  - Majority of class time spent on discussion, in class activities and exercises
    - Prepare by reading assigned material
- Software demonstrations

EVALUATION

- Labs (roughly weekly) [10] **(20%)**
  - Posted (on Sakai and website) and turned in on Sakai
  - Generally, completed outside of class (
- Homeworks [2] **(15%)**
  - Cover multiple labs to help solidify understanding
- Exams [2] **(30%)** -written and applied, (Take Home !!)
- Final project and oral presentation **(20%)**
- Participation **(15%)**
  - Come to class, ask questions, make use of discussion board (benefits everyone in class)

COURSE MATERIAL AND SCHEDULE

- 2nd cohort of spatial data science students
  - Highly likely to change throughout the semester
    - I will update the syllabus on the website and send announcements
- Want to learn something in particular not on the syllabus?
  - One class time reserved for advanced topics
  - Let me know! I cannot promise, but will see what I can do

COMPUTERS/ TECHNOLOGY REQUIREMENTS

- Bring computers to class everyday
  - In-class exercises/polls
- Install software (I will upload a detailed document on the website)
  - R and RStudio (open-source and cross-platform)

CLASSROOM EXPECTATIONS

- Please show up on time
- Please be respectful of your classmates' learning environment
  - Silence your devices
    - Please, no texting, emailing, social media, etc
      - While you may be an expert at multi-tasking, research shows your peers are not!
- Be respectful of others' views or opinions
- Humor my attempts at humor

OFFICE HOURS

- 223 Carolina (Hurston) Hall
- Monday 11am-12pm
- Thursday 2-4pm
- Friday 2-3 pm
- By appointment
- If my door is open, feel free to knock and I will be happy to talk for a few minutes
- TA office hours (to be updated soon)

TEXT

QUESTIONS SO FAR?

ABOUT ME

- 4th year PhD candidate in Geography
- Health and Medical Geographer
  - Also, GiScience and spatial analysis
- Please call me Varun (pronounced like Maroon but with a V).
- I am not technically a doctor or a professor (hopefully soon someday!)

MAJOR RESEARCH THEMES

- Infectious Disease ecology - Malaria in the Democratic Republic of Congo, Diarrheal Disease in Bangladesh
- Applying spatial methods to evaluating impacts of health policies and public health programs
- Applying spatial thinking to public health studies -
  - Impacts of herd immunity (or vaccination exemptions) on surrounding populations?
  - How does an individual's neighborhood affect their health?

OTHER "NON RESEARCH" INTERESTS

- Developing tools and interacting web applications for non-profit organizations (in R)
- Contributing to voluntary humanitarian mapping efforts
- Contributing to open-source projects -Data**R**fying aspects of my workflow in R (including making the course website, writing papers, writing presentations)
- Cooking everything from appetizers, mains to dessert with chickpea -
  - Secretly hoping to opening a chickpea themed restaurant one day

YOUR TA AND GRADUATE RESEARCH CONSULTANT

- Prabisha Shrestha (Course TA)
- Kate Brandt (Course GRC)

YOU!

- Preferred Name
- Major
- Your favorite show at the moment, or something you have enjoyed binge watching lately

BEFORE NEXT CLASS

- Install R, RStudio and Rmarkdown (document will be posted tonight)
- Thoroughly read assigned readings (will be uploaded tonight: *required* and *supplemental* readings
- Get access to class piazza page, and complete survey (will be sent tonight)

QUESTIONS ?