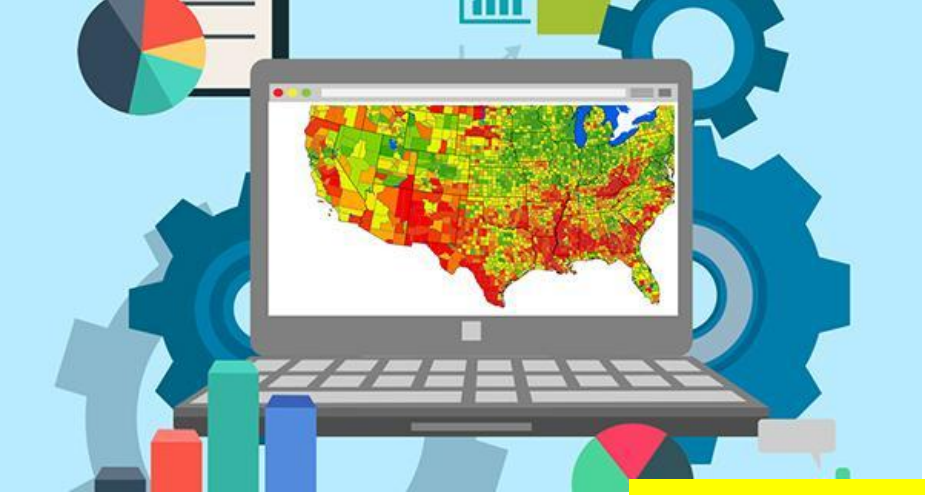
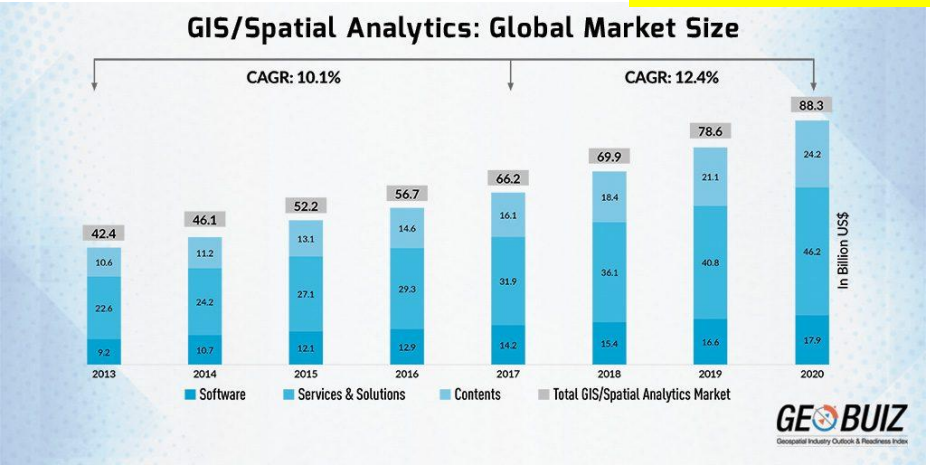


The global GIS and Spatial Analytics mark
to touch US$88.3 Billion by 2020

to US$ 134.23 Billion in 2028!

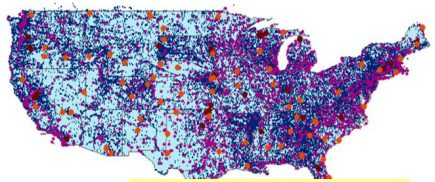


Spatial Data Science

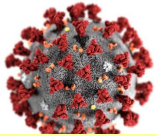
August 11, 2023

SDSC Spatial Information Systems Lab

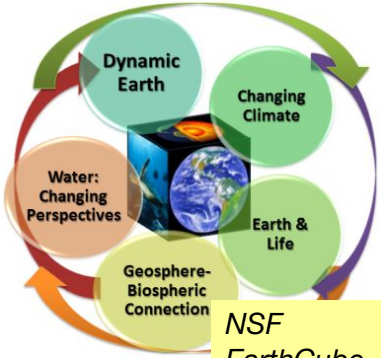
- Spatial data integration
- Interoperability and data standards
- GIS and Geospatial Databases
- Information discovery across distributed sources
- Online analysis systems
- Survey data analysis



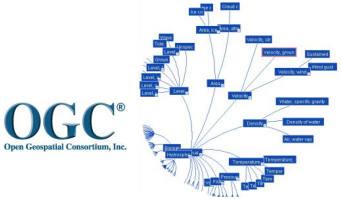
Hydrologic Information System (largest in the world)



COVID-19 modeling

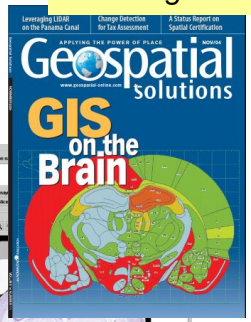


NSF EarthCube

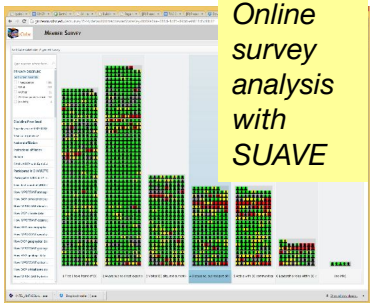


Mexico Health Atlas

Ecosystem Services Dashboard



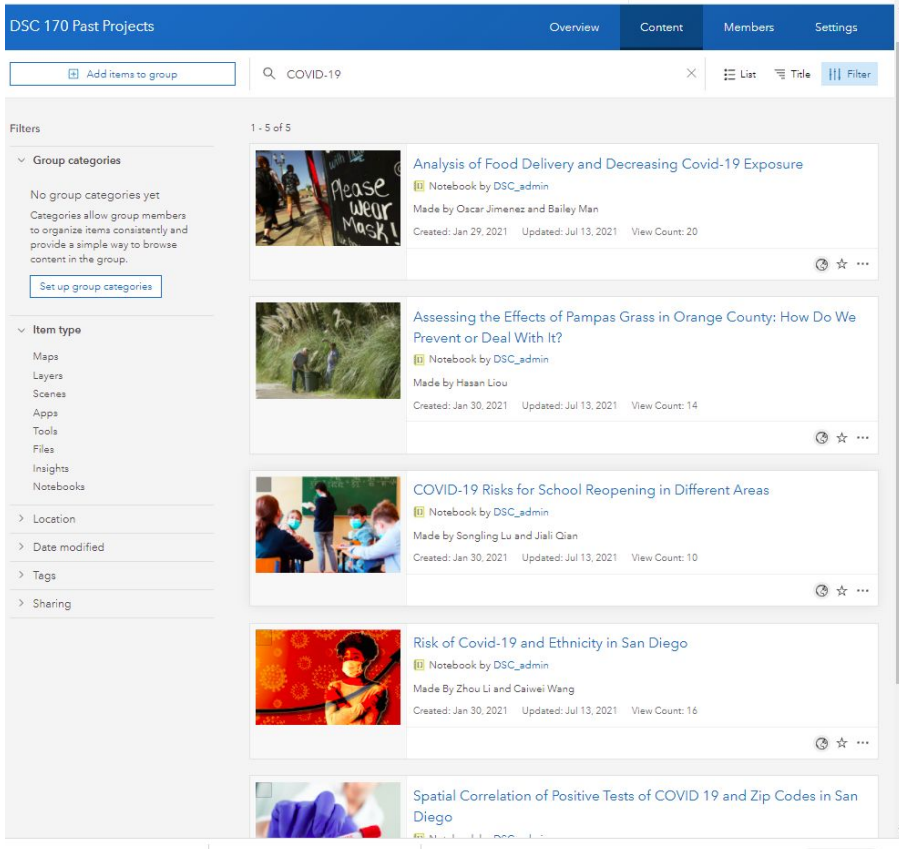
Brain data integration



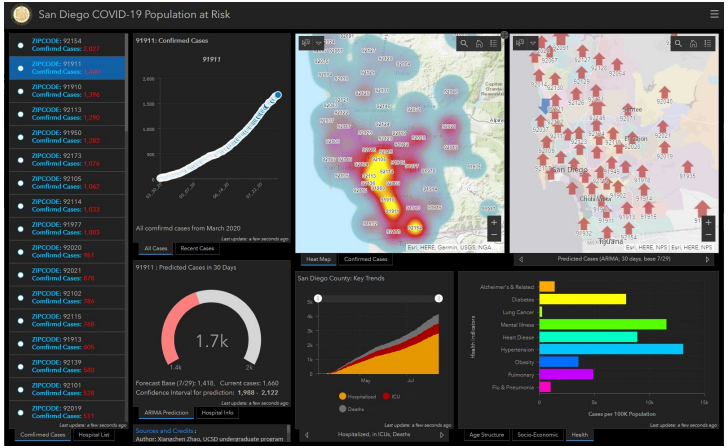
Online survey analysis with SUAVE

Critical Zone Observatories

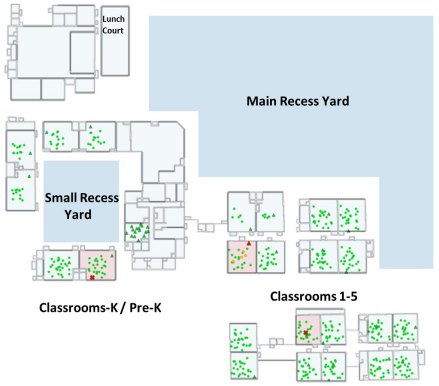
Spatial data science for COVID-19 analysis and modeling



[DSC170 final student projects are publicly available]



COVID-19 Predictive Modeling Dashboard

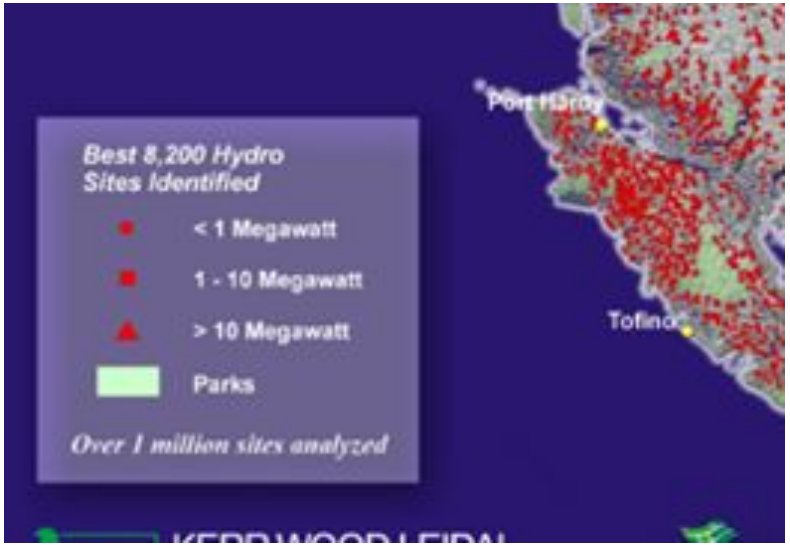


Activities Modeled

Group Activities during class time

[Spatially-explicit agent-based model of COVID-19 in schools]

What is wrong with this map?



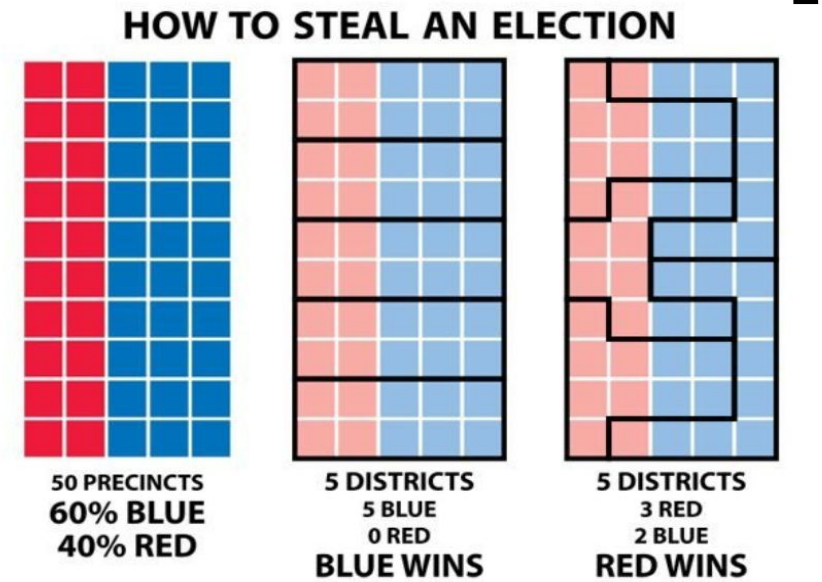
https://cartastrophe.wordpress.com

They improved:
https://www.kwl.ca/wp-content/uploads/2019/07/7-1-1.jpg

What common spatial analysis problem does this analysis illustrate?

Analysis: Thanks to Gerrymandering, N.C. Democrats Wasted 1.3 Million Votes

BY JEFFREY C. BILLMAN NOV. 16, 2018 3:59 P.M.



From:
https://indyweek.com/news/northcarolina/analysis-gerrymandering-efficiency-gap/

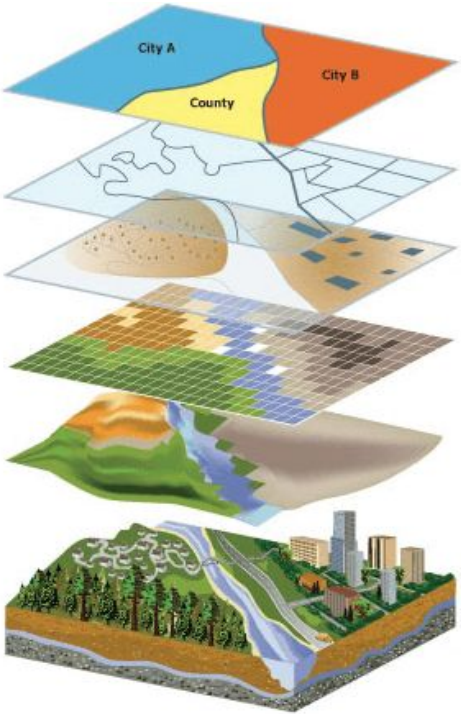
District	A votes	B votes	Winner	A Wasted Votes	B Wasted Votes
1	53	47	A	2	47
2	53	47	A	2	47
3	53	47	A	2	47
4	53	47	A	2	47
5	15	85	B	15	34
total	227	273	4-A, 1-B	23	222

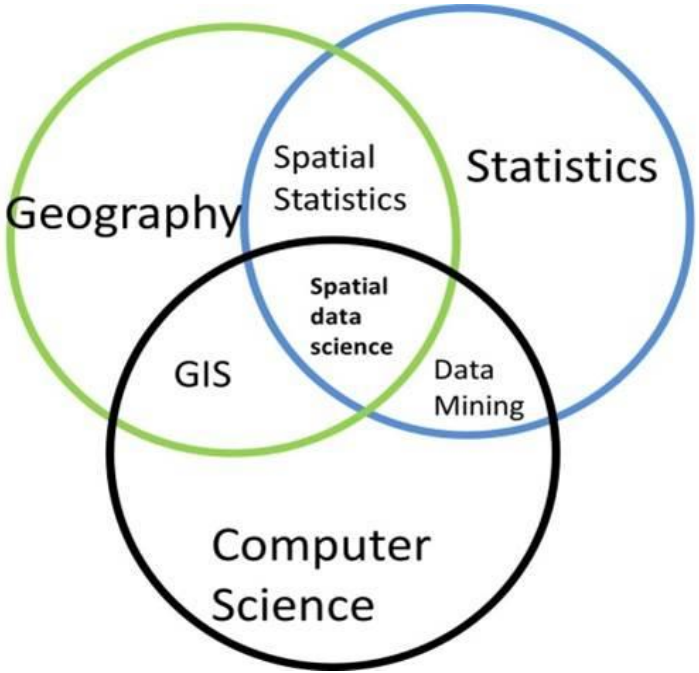
$$\text{Efficiency gap} = \frac{222 - 23}{500} = 39.8\% \text{ in favor of Party A.}$$

https://en.wikipedia.org/wiki/Wasted_vote

Gerrymandering = Gerry (Gov of Massachusetts in 1812 who drew electoral districts) + salamander (how these districts looked like)

From GIS to Spatial Data Science



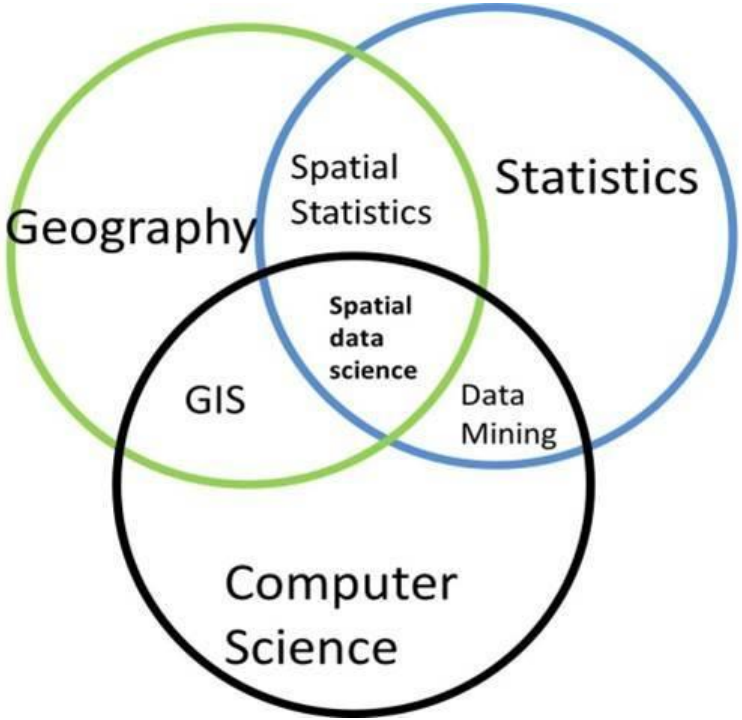


From desktop GIS to online spatial data services and cloud computing

From static maps to interactive spatial databases and advanced analytics

Spatial Data Science

Extracting insights from spatial data to use in practical applications

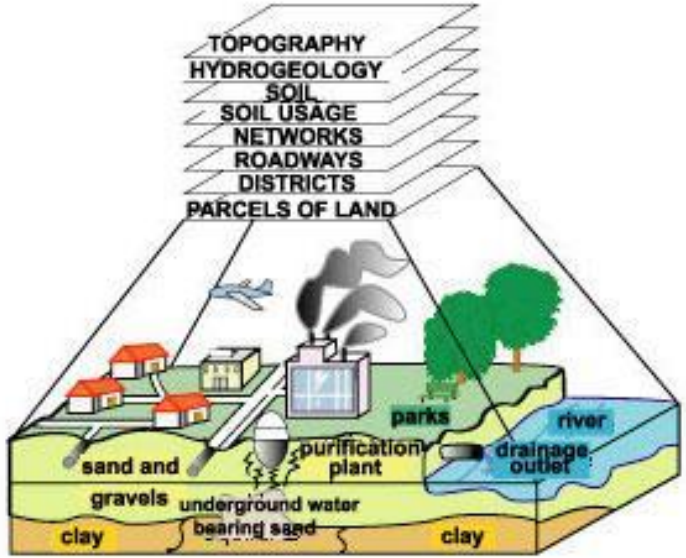


Building blocks:

- Formulating a spatial analysis problem
- Data engineering
- Visualization and exploration
- Spatial analysis and ML
- Big data analytics
- Modeling
- Sharing and collaboration

What is Spatial Analysis?

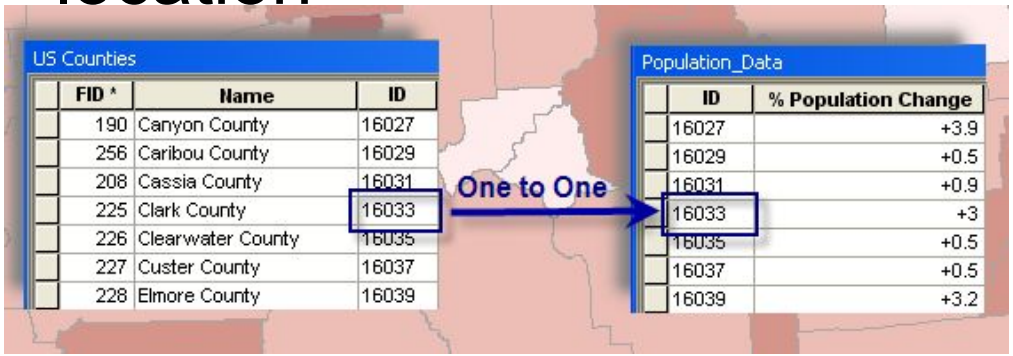
- Generating insights based on analysis of geographic features, spatial relationships between them, and processes that influence their patterns and dynamics
- … the purpose of geographic inquiry is to examine relationships between geographic features collectively and to use the relationships to describe the real-world phenomena that map features represent.” (Clarke 2001)



Source: https://www.planete-tp.com/

Objects or fields?	How accurate?
Same time period?	Spatial reference system?
Update frequency?	Provenance of the data?
Model-generated data?	Cleaned and analysis-ready?

Joining datasets by attribute values vs joining by location



Here, we can use common attributes in two datasets to join them

Using space as a universal index to join disparate data – the key idea behind GIS!

No fields with matching values that we can use to join datasets…

… **but we have geometric information about each object!**

Destination feature class

FID	NAME	POP
158	Adair Village	560
470	Adams	252
58	Adel	0
274	Adrian	141
305	Agate Beach	0
60	Agness	0

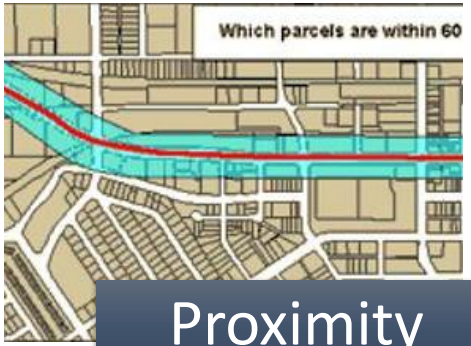
Source feature class

OID	Airport
6	Corvallis Municipal
13	Eastern Oregon Regional At
2	Klamath Falls International
9	La Grande/Union County
10	Newport Municipal
3	Roseburg Regional

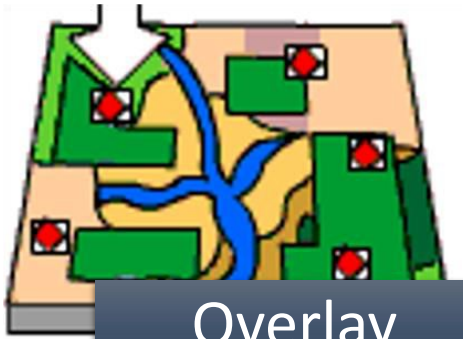
What is the cardinality?

FID	NAME	POP	OID	Airport	Distance
158	Adair Village	560	6	Corvallis Municipal	20,453
470	Adams	252	13	Eastern Oregon Regional At	22,903
58	Adel	0	2	Klamath Falls International	151,547
274	Adrian	141	9	La Grande/Union County	187,488
305	Agate Beach	0	10	Newport Municipal	10,764
60	Agness	0	3	Roseburg Regional	95,752

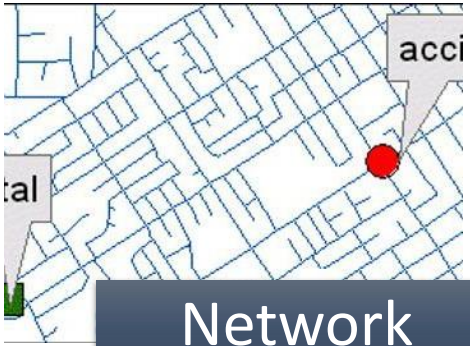
Types of GIS Analysis



Proximity



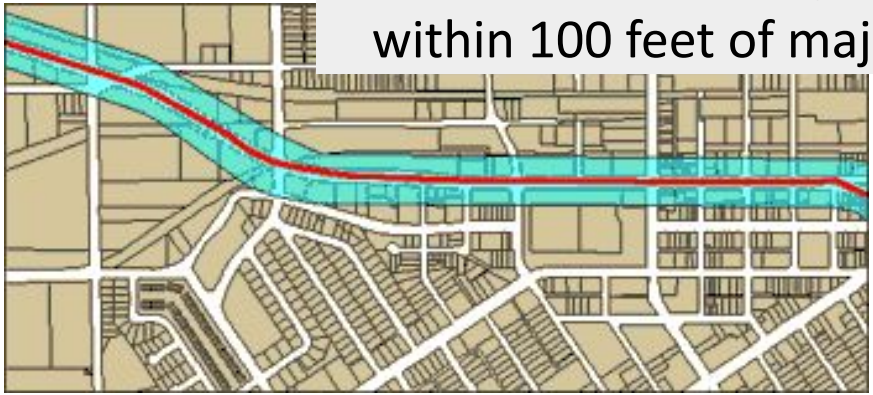
Overlay



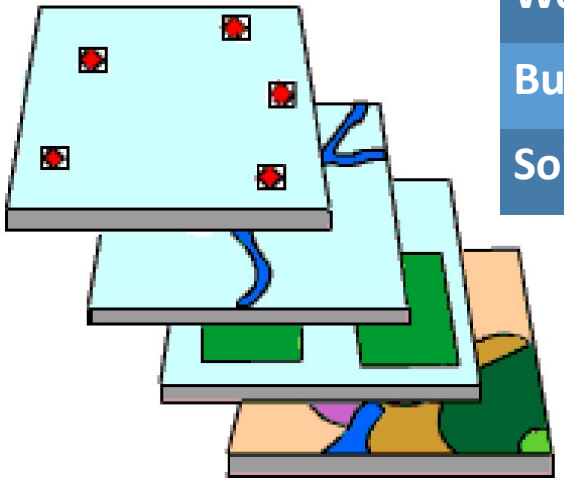
Network

Analysis: Proximity

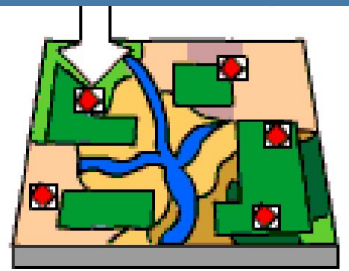
Which asthma patients are within 100 feet of major roads?



Analysis: Overlay



Well type	Drilled
Building Owner	Smith
Soil Type	Sandy



Example: a simple MLP to estimate travel time

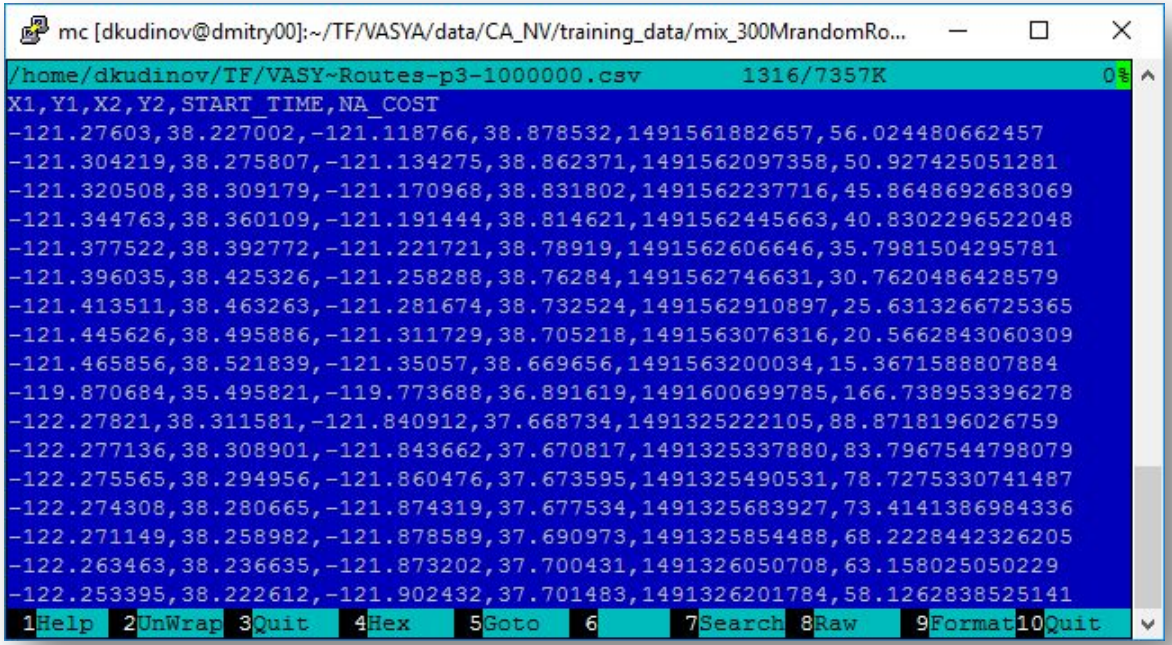
Knowledge about topology of a transportation graph from GPS tracks

Raw input data:

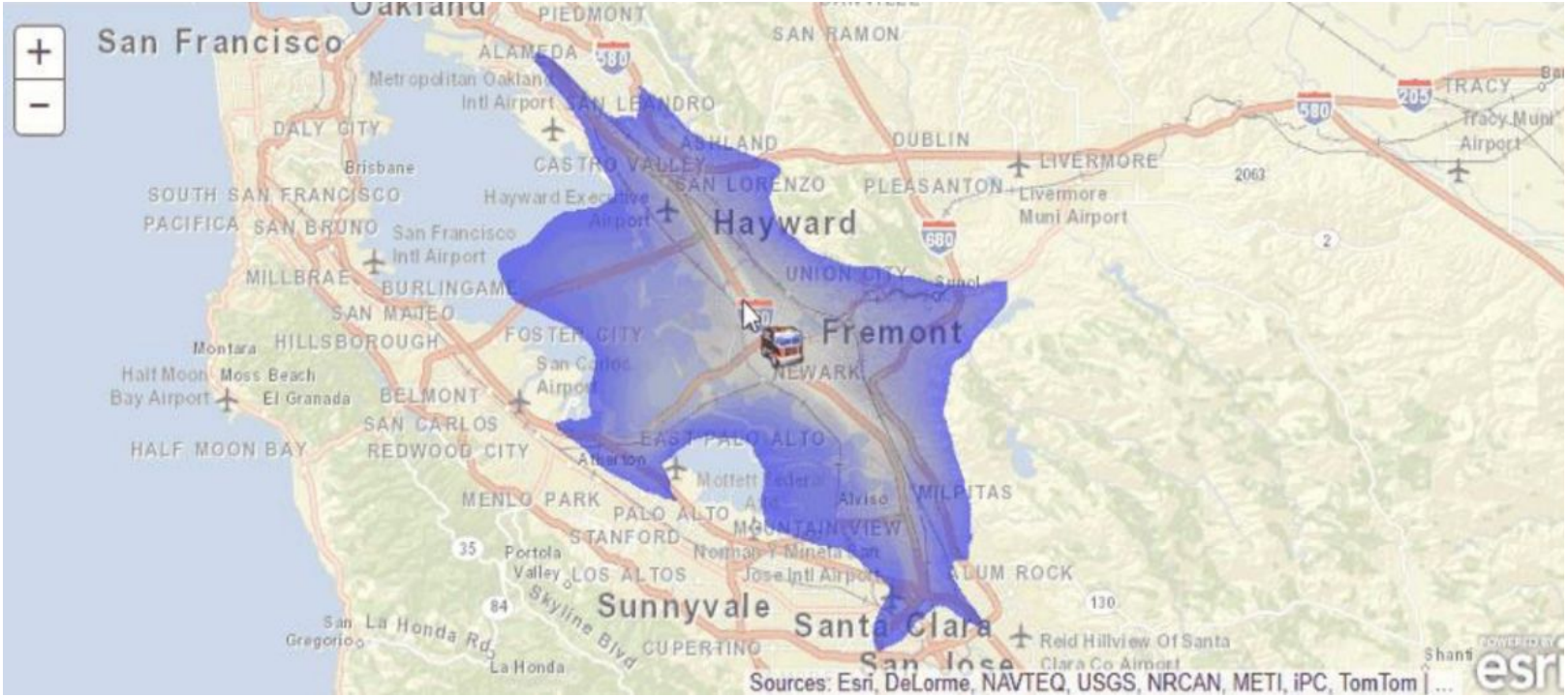
- ~320M From-To GPS pairs n California and Nevada.

Goal:

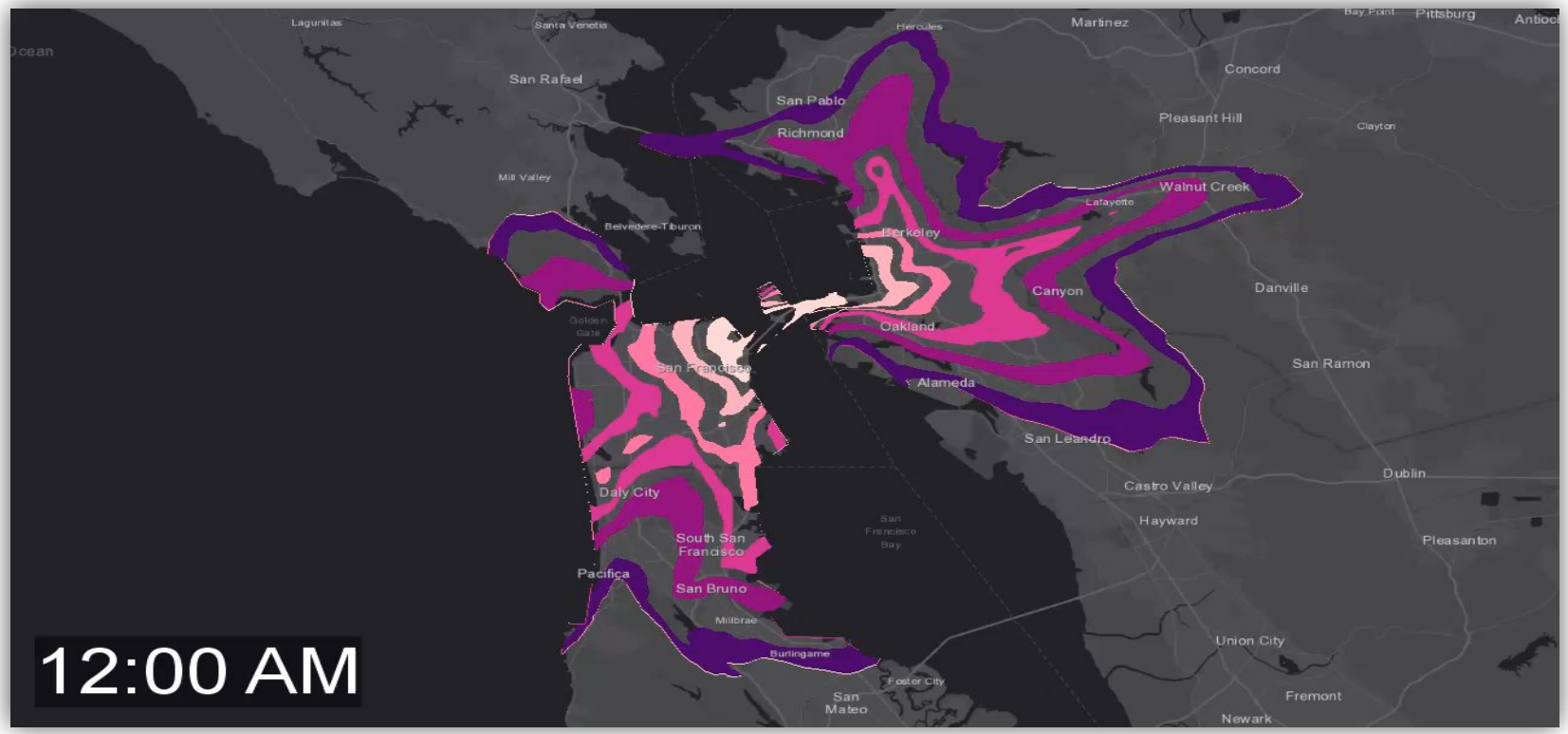
- Predict Travel Time given From, To, Departure Time.



Information products



Information products



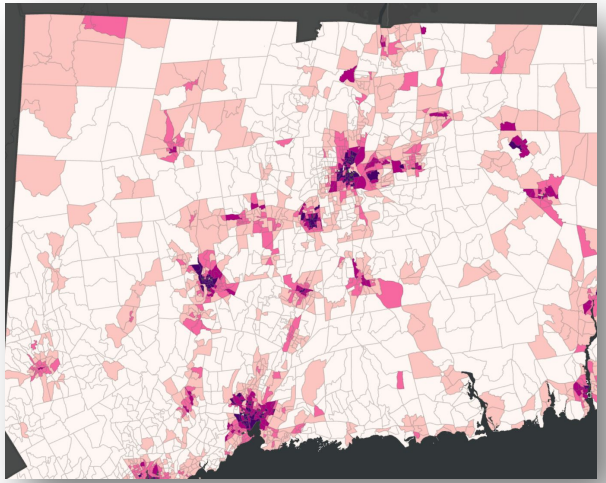
All from ~320M samples -

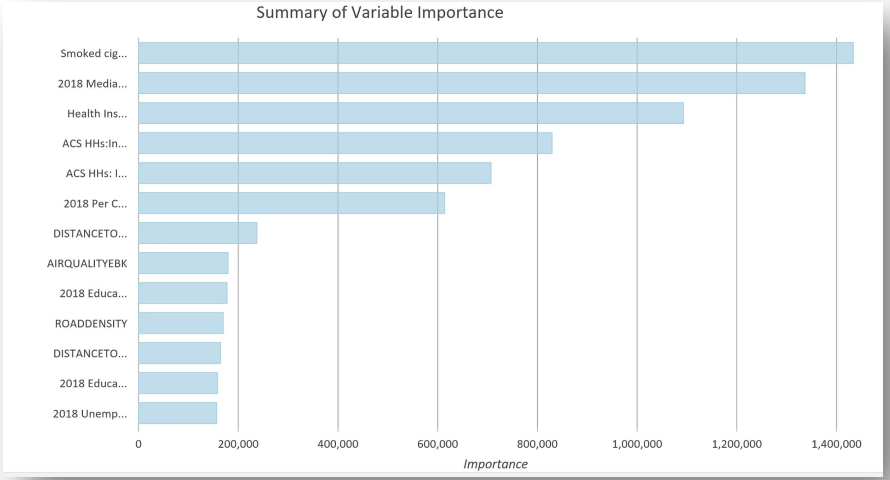
```
/home/dkudinov/TF/VASY~Routes-p3-1000000.csv      1316/7357K                0%
X1,Y1,X2,Y2,START_TIME,NA_COST
-121.27603,38.227002,-121.118766,38.878532,1491561882657,56.024480662457
-121.304219,38.275807,-121.134275,38.862371,1491562097358,50.927425051281
-121.320508,38.309179,-121.170968,38.831802,1491562237716,45.8648692683069
-121.344763,38.360109,-121.191444,38.814621,1491562445663,40.8302296522048
-121.377522,38.392772,-121.221721,38.78919,1491562606646,35.7981504295781
-121.396035,38.425326,-121.258288,38.76284,1491562746631,30.7620486428579
-121.413511,38.463263,-121.281674,38.732524,1491562910897,25.6313266725365
-121.445626,38.495886,-121.311729,38.705218,1491563076316,20.5662843060309
-121.465856,38.521839,-121.35057,38.669656,1491563200034,15.3671588807884
-119.870684,35.495821,-119.773688,36.891619,1491600699785,166.738953396278
-122.27821,38.311581,-121.840912,37.668734,1491325222105,88.8718196026759
-122.277136,38.308901,-121.843662,37.670817,1491325337880,83.7967544798079
-122.275565,38.294956,-121.860476,37.673595,1491325490531,78.7275330741487
-122.274308,38.280665,-121.874319,37.677534,1491325683927,73.4141386984336
-122.271149,38.258982,-121.878589,37.690973,1491325854488,68.2228442326205
-122.263463,38.236635,-121.873202,37.700431,1491326050708,63.158025050229
-122.253395,38.222612,-121.902432,37.701483,1491326201784,58.1262838525141
 1Help  2UnWrap  3Quit   4Hex   5Goto   6        7Search  8Raw   9Format 10Quit
```

ML: Prediction

Predicting Unknowns by Training a Model on a Labelled Dataset

Use Case: Predicting Asthma Hospitalizations per Block Group in Connecticut. Exploring the Top Factors behind the Prediction (Hospitalization) as well



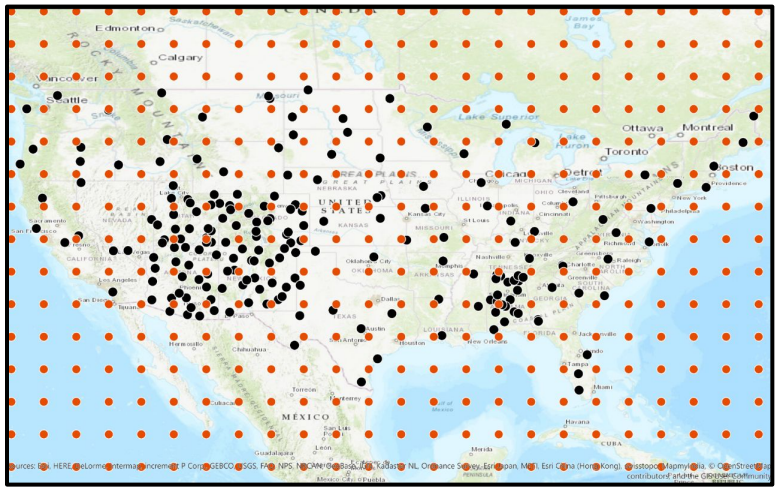


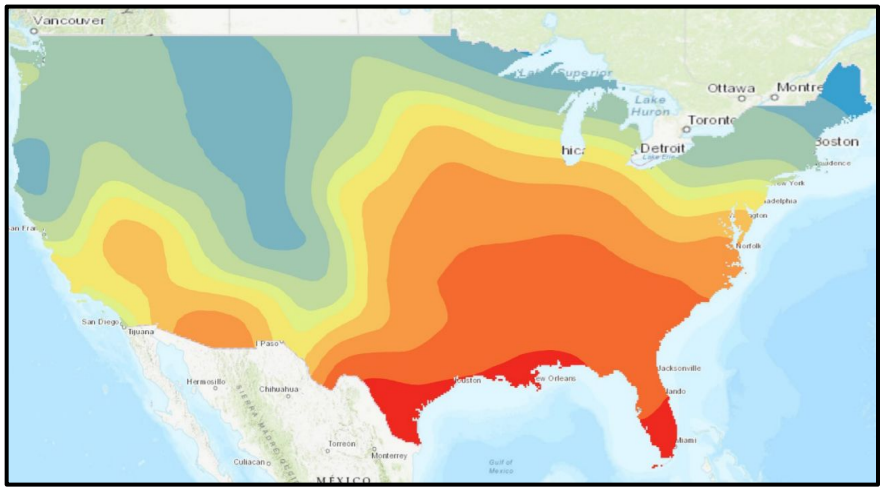
In ArcGIS: Forest Based Classification and Regression, Geographically Weighted Regression, Ordinary Least Squares Regression, Kriging

ML: Prediction / Interpolation

Using the known to estimate the unknown

Use Case: Accurately predict impacts of climate change on local temperature using global climate model data



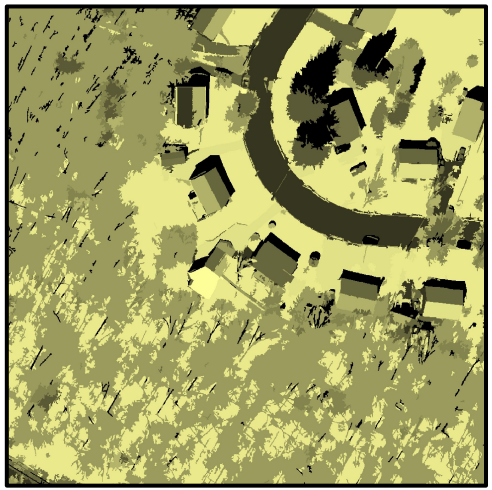


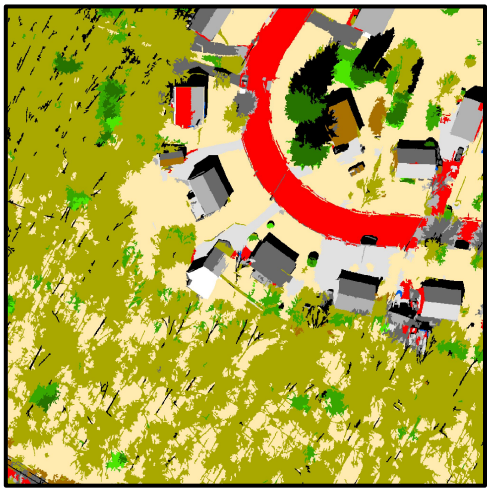
In ArcGIS: Empirical Bayesian Kriging, Areal Interpolation, EBK Regression Prediction, Ordinary Least Squares Regression and Exploratory Regression, Geographically Weighted Regression

ML: Classification

The process of deciding to which category an object should be assigned based on a training dataset

Use Case: Classify impervious surfaces to help effectively prepare for storm and flood events based on the latest high-resolution imagery



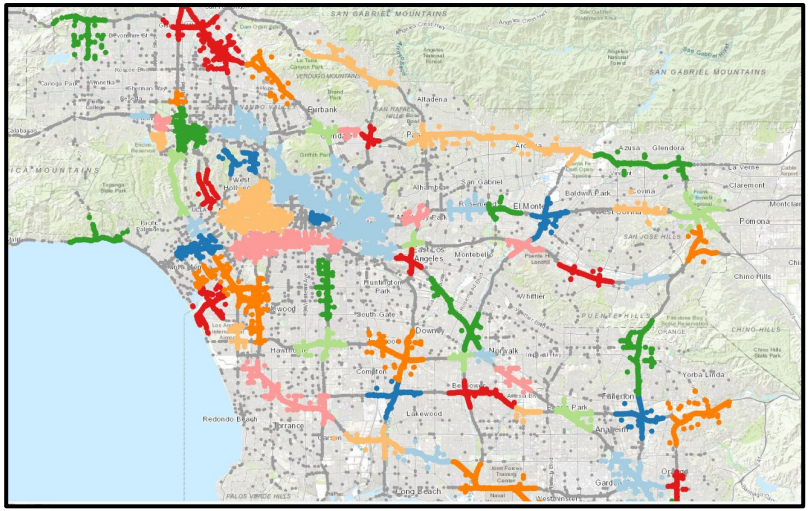


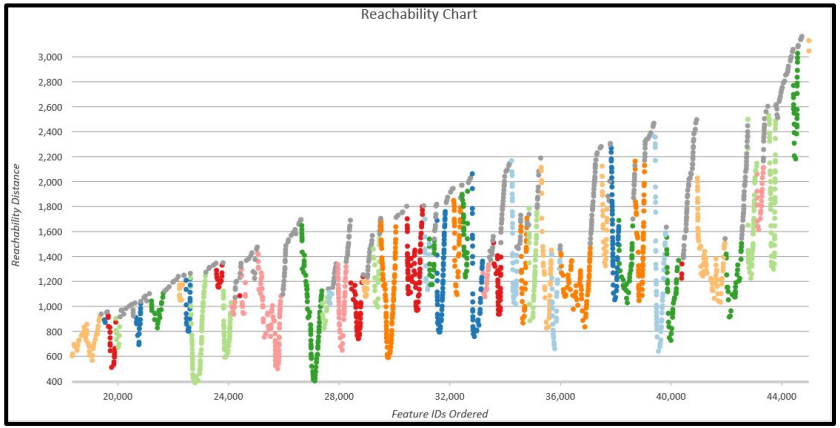
In ArcGIS: Maximum Likelihood Classification, Random Forest, Support Vector Machine

ML: Clustering

The grouping of observations based on similarities of values or locations

Use Case: Given the nearly 50,000 reports of traffic between 5pm and 6pm in Los Angeles (from Traffic Alerts by Waze), where are traffic zones that can be used to elicit feedback from current drivers in the area?



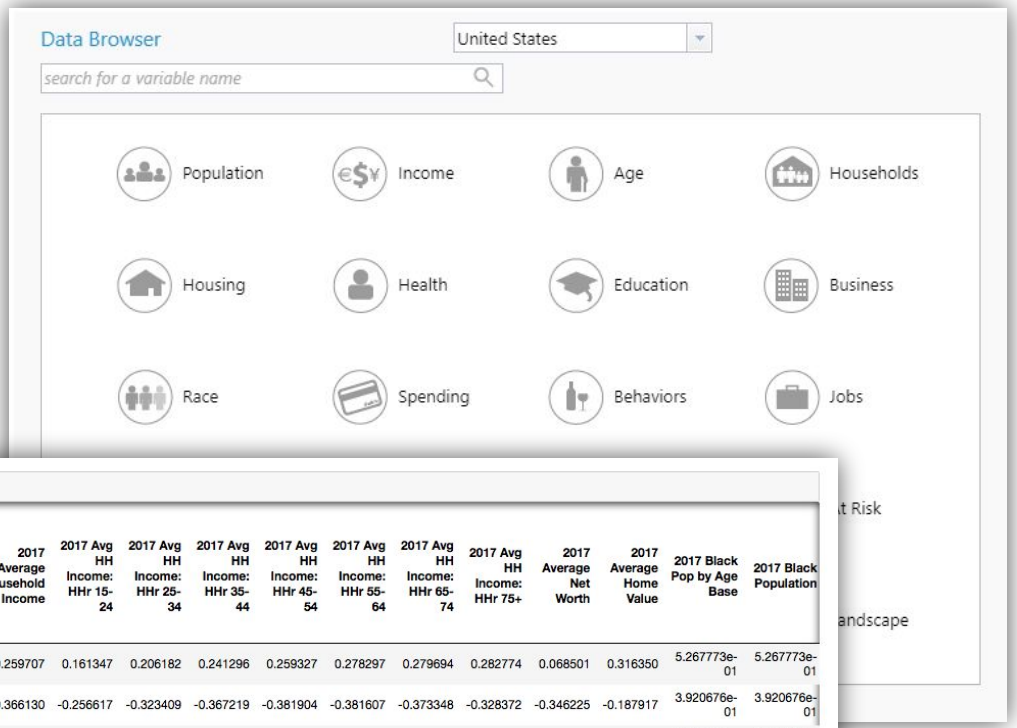


In ArcGIS: Spatially Constrained Multivariate Clustering, Multivariate Clustering, Density-based Clustering, Image Segmentation, Hot Spot Analysis, Cluster and Outlier Analysis, Space Time Pattern Mining

Where to open a **brick-n-mortar store** to maximize the **online sales**?

Open new Offline stores to boost Online sales?

- Historical georeferenced data of online and offline sales.
- History of brick-n-mortar store openings and closings.
- GeoEnrichment to add more variables:
  - Income,
  - Family size,
  - Age,
  - Real Estate value,
  - Education,
  - Health, …



In [15]: factor_table

2017 Asian Pop Age 85+	2017 Average Disposable Income	2017 Average Family Size	2017 Average Household Size	2017 Average Household Income	2017 Avg HH Income: HHr 15-24	2017 Avg HH Income: HHr 25-34	2017 Avg HH Income: HHr 35-44	2017 Avg HH Income: HHr 45-54	2017 Avg HH Income: HHr 55-64	2017 Avg HH Income: HHr 65-74	2017 Avg HH Income: HHr 75+	2017 Average Net Worth	2017 Average Home Value	2017 Black Pop by Age Base	2017 Black Population
0.337762	0.267120	-0.000403	-0.002960	0.259707	0.161347	0.206182	0.241296	0.259327	0.278297	0.279694	0.282774	0.068501	0.316350	5.267773e-01	5.267773e-01
0.072360	-0.369341	0.005300	0.006875	-0.366130	-0.256617	-0.323409	-0.367219	-0.381904	-0.381607	-0.373348	-0.328372	-0.346225	-0.187917	3.920676e-01	3.920676e-01
0.586181	0.222665	0.004702	0.006868	0.233728	0.180754	0.232275	0.225480	0.224534	0.227649	0.214383	0.211953	0.122326	0.355778	-3.603345e-01	-3.603345e-01
0.058719	-0.066830	0.000532	0.002885	-0.097569	-0.052679	-0.104736	-0.104749	-0.101179	-0.095976	-0.081660	-0.078063	-0.034022	-0.153067	-4.764484e-01	-4.764484e-01
0.242423	-0.122336	-0.002263	-0.003279	-0.132532	-0.094038	-0.125207	-0.132322	-0.134337	-0.133141	-0.121797	-0.109517	-0.089141	-0.137234	3.286951e-01	3.286951e-01

Open new Offline stores to boost Online sales?

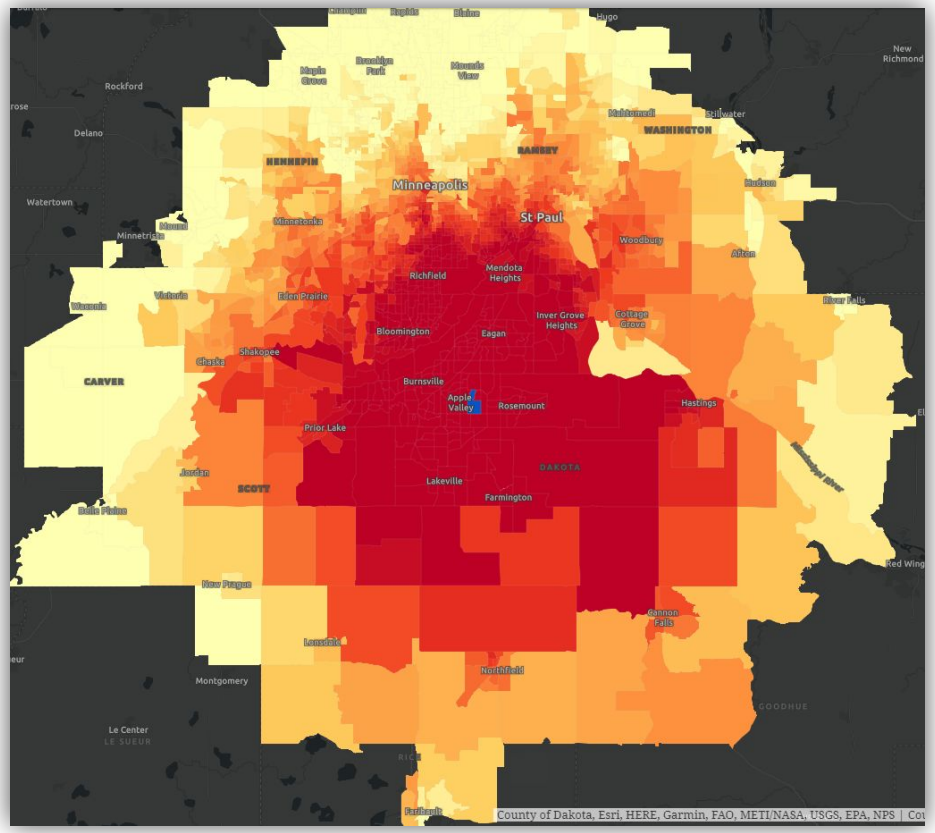
Input

- For each offline store location the socio-demographic values were aggregated within a polygon equal to an hour of drive time.

- Online Sales within the same polygon before and after the offline store was opened / closed.

Goal:

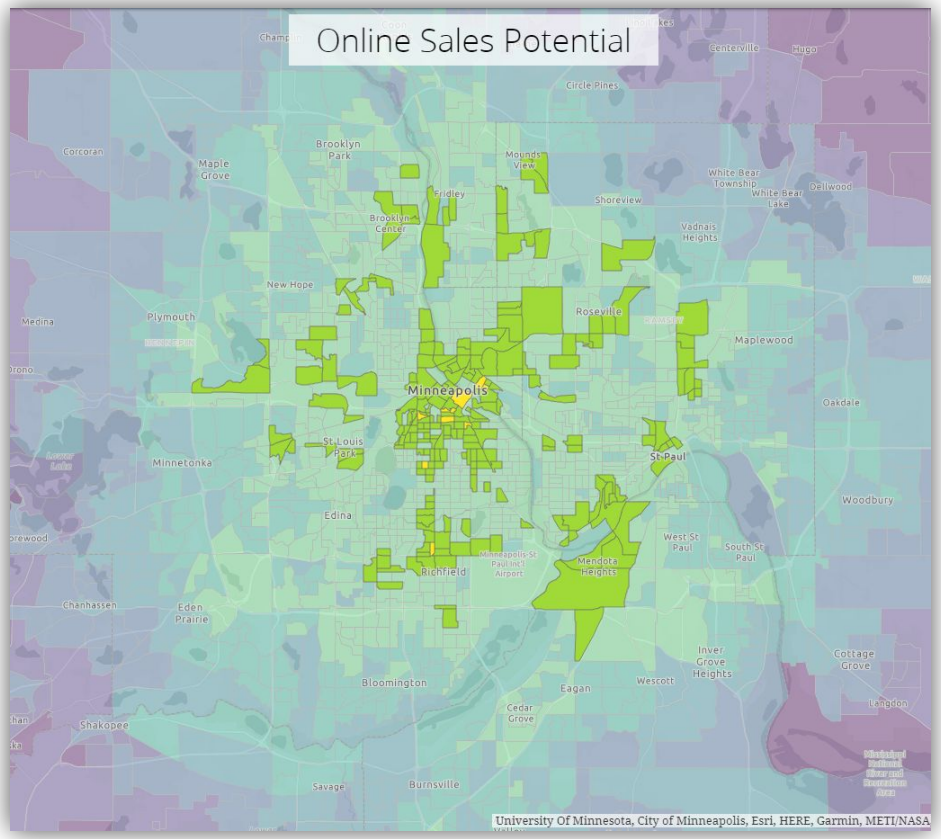
- Find the optimal location for a new offline store to maximize the online sales boost.



Open new Offline stores to boost Online sales?

- After the model was trained, we selected regions with a projected online sales growth above $1.5M per year ->

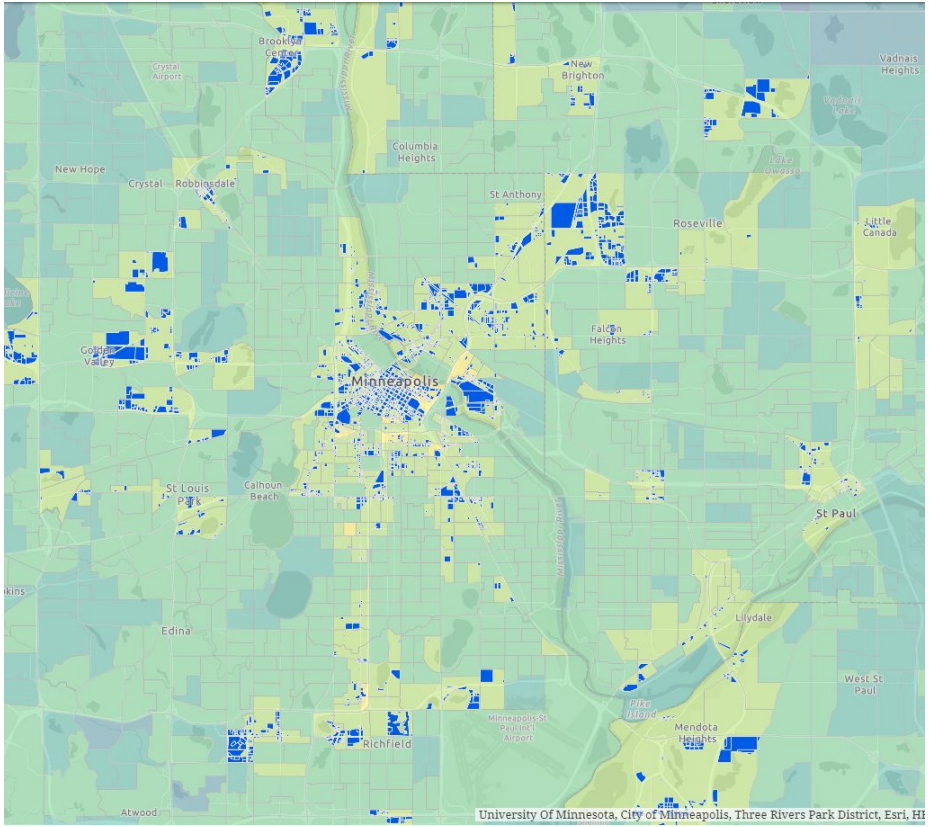
- R^2 on Test set ~0.94



Open new Offline stores to boost Online sales?

- Using traditional GIS tools, we filtered out non-commercial zones, places with a little foot traffic.

- The remaining candidates were further filtered with Network Analyst Location-Allocation Solver to maximize the location accessibility based on transportation network and historical road traffic.



HDSI UC SAN DIEGO

SDSC SAN DIEGO SUPERCOMPUTER CENTER

GeoACT

Geographically-assisted Agent-based model of COVID-19 Transmission

Bailey Man, Kaushik Ganapathy, Jiaxi Lei, Ilya Zaslavsky

Halıcıoğlu Data Science Institute, San Diego Supercomputer Center, UCSD

Presented at Gateways 2021

https://zenodo.org/record/5569451

Introduction

What's the Problem?

- How do we keep students and teachers healthy?
  - Teachers and parents need to know infection risks for **THEIR children** and **THEIR schools**

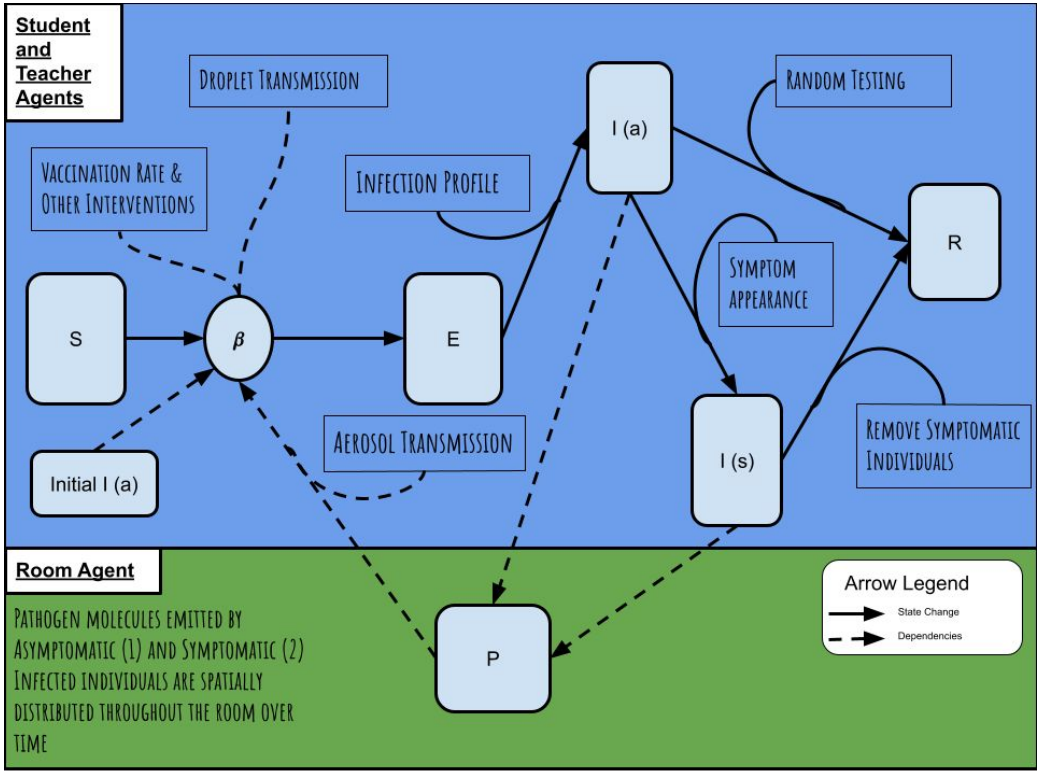
- How can we help schools prepare?
  - Simulate their plans in advance
  - Determine which activities are conduits of transmission
  - Evaluate which interventions will work best for a given school

Our Solution:

We will provide a set of simulations to assess transmission and NPIs in relevant school settings

Methodology

SEIR-P Diagram



Susceptible

Exposed

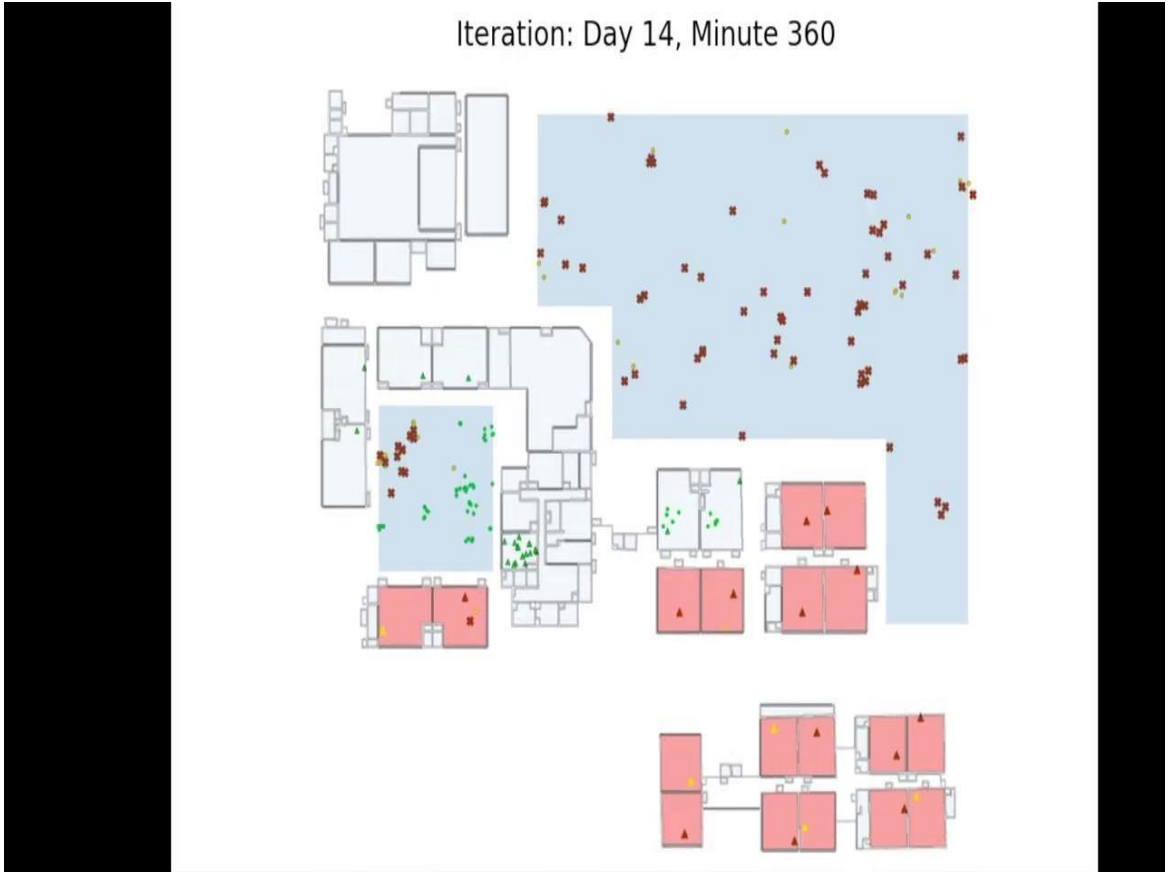
Infected (**A**symptomatic)

Infected (**S**ymptomatic)

Removed

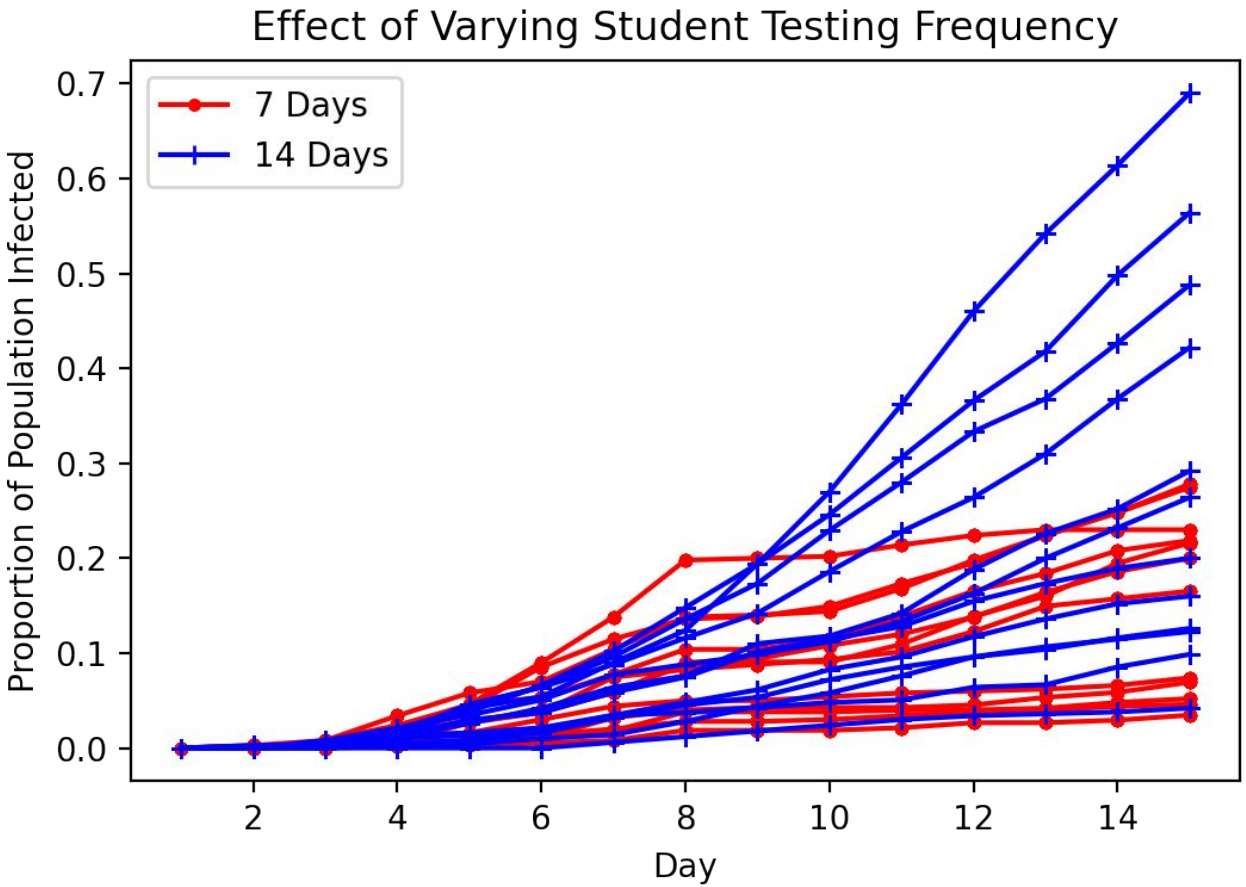
Pathogen

Simulation Video



Results

School Simulation Results: Example



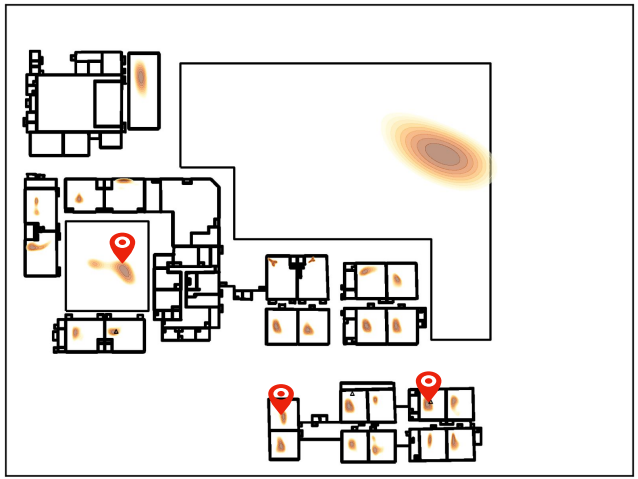
More frequent testing allows schools and parents to identify COVID-19 infected asymptomatic students and move them from in-person to remote learning. Under no other NPIs, this cuts infection rates after 2 weeks by roughly 2.5 times.

If infected students are identified, this would also trigger quarantine regime for others in the classroom.

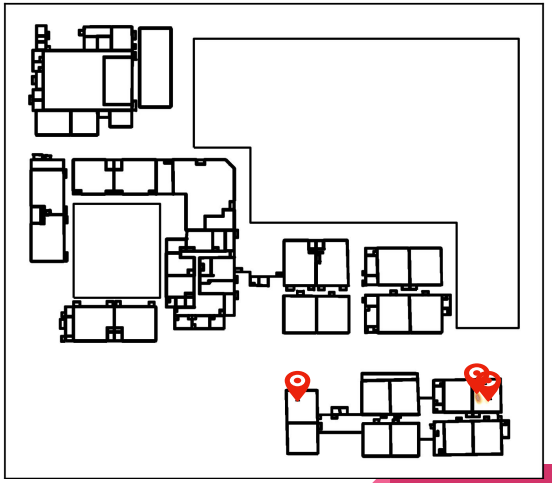
Mask wearing compliance and Indoor/Outdoor Lunch

[testing frequency: 14 days, attendance: 100%]

Cafeteria Lunch & 0% Masks



Class Lunch & 100% Masks



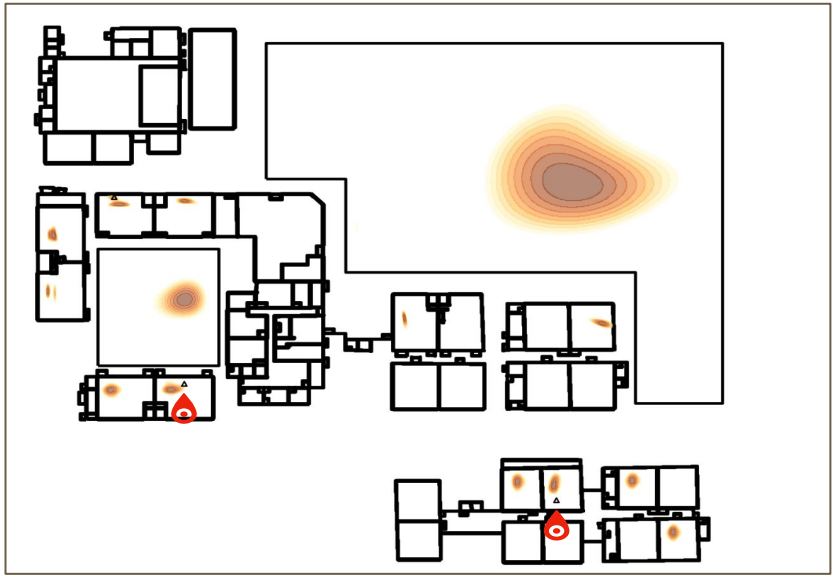
Mask wearing and avoiding indoor cafeteria lunches help in reducing transmission

Transmission in recess yards assumes no sanitization of common surfaces

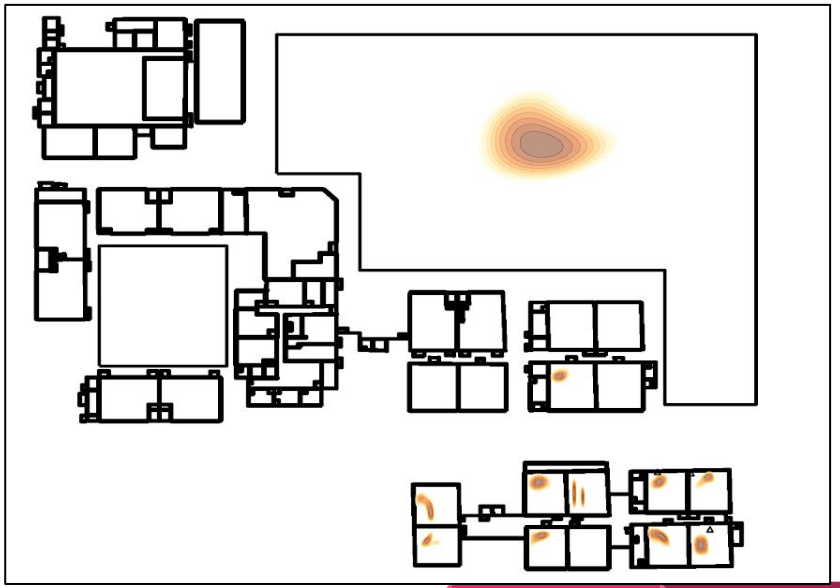
Effects of Class Attendance

(under 14-day test frequency, 50% mask compliance, lunch in classrooms)

50% Masks, 100% Attendance & Class Lunch



50% Masks, 75% Attendance & Class Lunch



Mask wearing and reducing cohort sizes can reduce extents of transmission

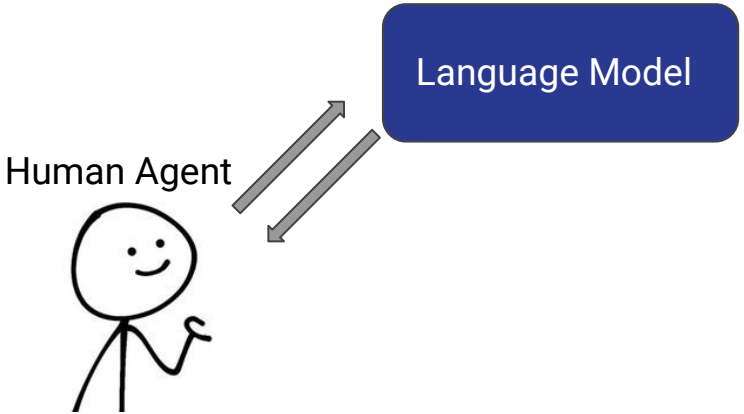
Transmission in recess yards assumes no sanitization of common surfaces

Recent Addition

Smart Agents

What if the human agents can make more intelligent/independent decisions?

With the recent advancements in Generative Language Models such as GPT4, we can!



Personas: Synthetic Background

	age	gender	home
0	17	male	Diego
1	11	female	Julian
2	11	female	Valley
3	13	male	Oceanside
4	14	female	Lemon
5	18	female	Rey
6	18	female	Campo
7	7	male	Palomar
8	8	male	El
9	26	male	Jacumba

Personas: Synthetic Background

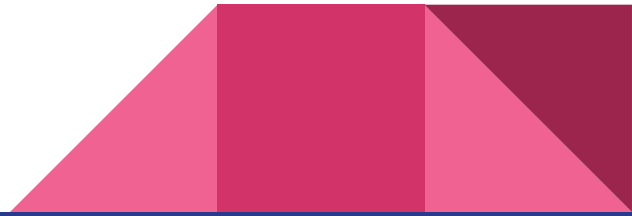
Sample input: “Your name is p1, you are a 26-year-old male from Luis, CA. Tell me something more about yourself. Use at most 30 words to explain.”

Sample output: “p1 is a 26-year-old male from Luis, CA who has an interest in the occult, videogames, literature, nature, and the world.”

Smart Agents: action phase

Problem: LLM output at agent level does not consider other agents' actions at the current step. Each agent is "play their own game". Hard to extract spatial information.

Example output:
P1: Talk to P2 about a future travel plan.
P2: Play the guitar and sing a song for everyone.



Smart Agents: action phase

Problem: LLM output at agent level does not consider other agents' actions at the current step. Each agent is "play their own game". Hard to extract spatial information.

Example output:
P1: Talk to P2 about a future travel plan.
P2: Play the guitar and sing a song for everyone.

Solution: phrase the input text at room-level and ask as a series of questions to determine the spatial relationship between each acting agents.

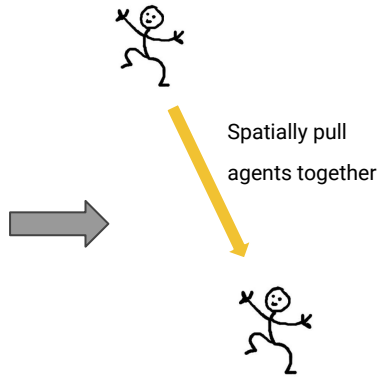
Smart Agents: action phase

Phrase the input text at room-level and ask as a series of questions to determine the spatial relationship between each acting agents.

With information of room and agents, what are the actions of agents?

Is there social interaction in each action?

Who will the agent interact with?



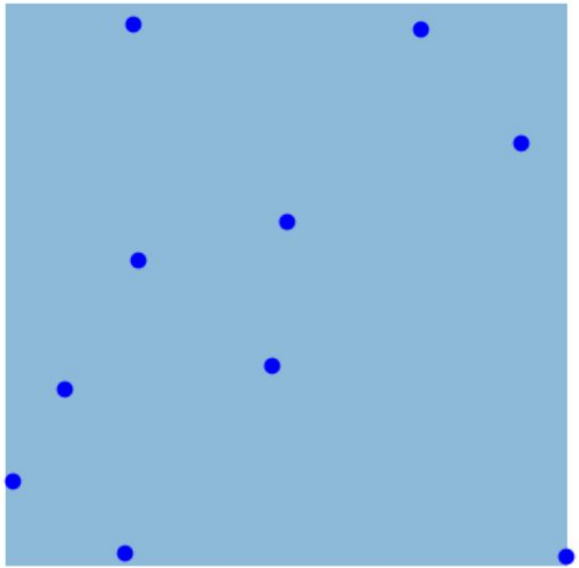
Smart Agents: action phase

Phrase the input text at room-level and ask as a series of questions to determine the spatial relationship between each acting agents.

Example action output: “Read a book and discuss it with a friend.”
Actors decided by LLM: ['p5', 'p6', 'p7']

Smart Agents: action phase

Step 0



Step 1