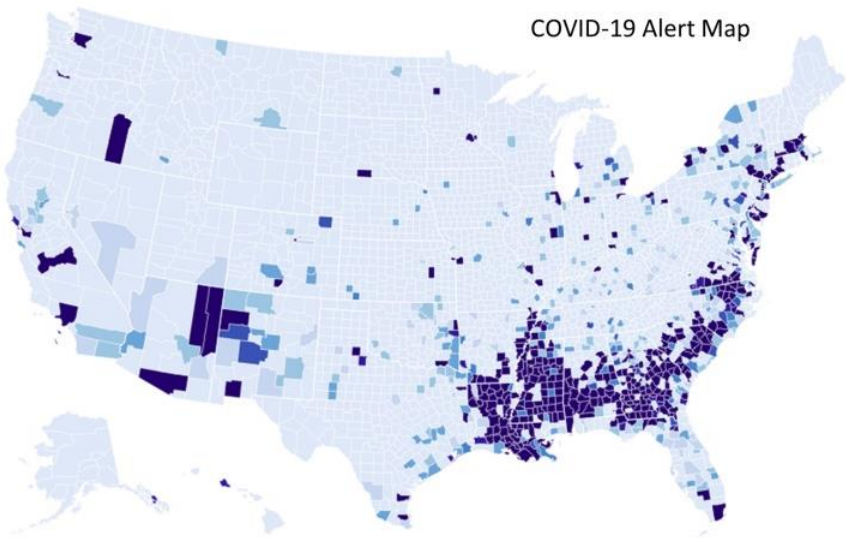
Using Demographic Pattern Analysis to Predict COVID-19 Fatalities on the US County Level

Explainable AI by

Akai Kaeru

http://akaikaeru.com info@akaikaeru.com



Method

COVID-19 Risk Analysis Goals

Understand: what factors expose a community to COVID-19 risk
Inform: health policy on important concurrent risks and correlations
Predict: local COVID-19 mortality, medical resource needs, success of mitigations

More information

Akai Kaeru

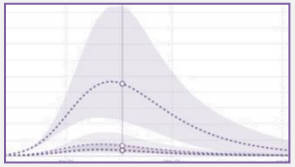
https://akaikaeru.com/covid-19-1

Computational Tools

Methods commonly in use

Simulation (Susceptible - Exposed - Infectious – Recovered (SEIR) model)

- needs almost no data and can look far ahead, but has many unknown parameters
  → observe and keep tuning



Curve fitting (e.g., IHME)

- can learn from other data, but data might not fit perfectly
  → collect data and keep fitting

Machine learning and AI (neural nets, random forests, decision trees)

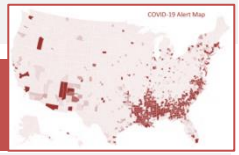
- can provide predictions, but require lots of data, are black boxes, lack explanations
  → this will happen -- but why?

Our new approach, expanding AI to *Explainable* AI

Pattern mining

- learns from associations in the data, learns them explicitly and makes them explainable
  → this will happen – and this is why!

Some of our Many Findings



May: we identified **279 patterns of counties** where COVID-19 death rate > US average
June: **98%** of these sets experienced a death rate growth of **2-3 times the US average**
July: these trends continued – new counties fitting the profiles got infected and deaths ↑

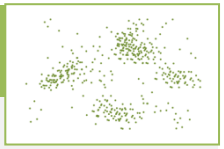
- **at risk:** sparsely populated counties with poor and aging populations
- **at risk:** counties with sleep-deprived, low-educated, uninsured residents
- **at risk:** wealthy counties with high home ownership and housing debt
  counties with more residents in debt have a higher risk of COVID-19 fatalities

Our Data

500 variables for 3,008 US counties

- demographics
- socioeconomic vulnerabilities
- housing composition vulnerabilities
- minority status and language
- housing, transportation, nutrition
- many of these from the CDC
- COVID-19 death rates (evolving)

Our Approach



Objectives

Find patterns (subpopulations) in the high-dimensional feature space where:

1. the data items are **similar** in a set of relevant features (variables)
2. the data items have, on average, **unusually high** (or low) values in some chosen target variable (in our case, COVID-19 death rate)

Benefit: dimension reduction

- typically each patters can be described by just a few features
- it forms a **brief narrative** of the process that caused the target

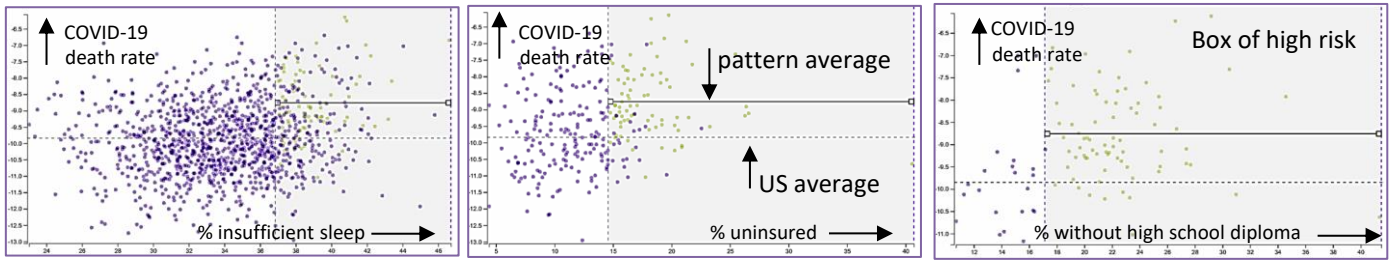
Case Studies

Risk Pattern

This sequence shows how our algorithm automatically identified a subpopulation of counties in the 500-D socio-economic feature space that fits the two search criteria:

- similar in this set's identified three features
- on average **a higher than US-average** COVID-19 death rate

↑ y-axis: May COVID-19 death rate on log scale ● a county ● a county in the pattern's subpopulation



1st feature: % insufficient sleep → 2nd feature: % uninsured → 3rd feature: % w/o high school diploma

Generated explanation:

Counties where a large proportion of residents get insufficient sleep, are uninsured, and did not finish high school are more at risk for these residents to die from COVID-19 when contracting the virus

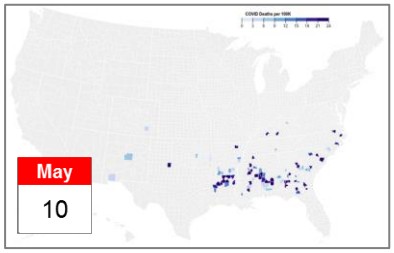
▶ Deeper takeaways:

All of these factors point to a **weakened immune system** which elevates COVID-19 vulnerability risk

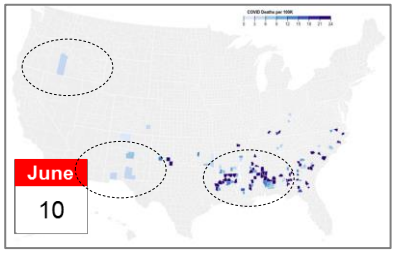
- **lack of sleep:** this weakens the immune system → direct risk
- **low education:** **(1)** residents may work two or more jobs leading to lack of sleep → indirect risk
  (2) they may work outside of the home which leads to social contacts → direct risk
- **uninsured:** residents take less care of their health and are likely physically weak → direct risk

Predictions

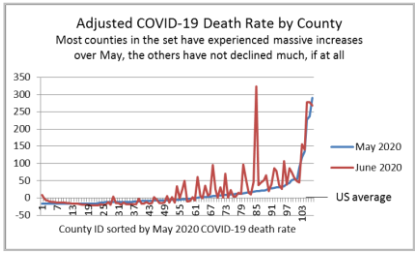
These maps show the pattern's counties colored by death rate



Mostly located in the South



Predicted counties show up

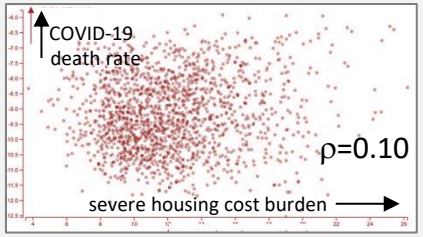


Predicted counties shoot up

Correlation Pattern

Correlations:

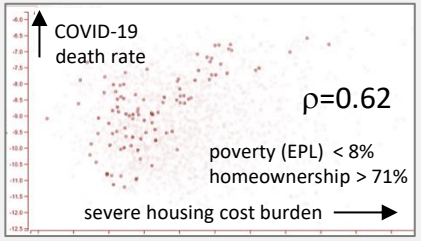
- important correlations are often hidden with conventional correlation analysis that uses all data points indiscriminately
- is there a correlation between **housing debt** and **COVID-19 death rate**? No.



But we found a correlation for counties in a pattern where

- home ownership is high
- poverty is low

More debt → more deaths



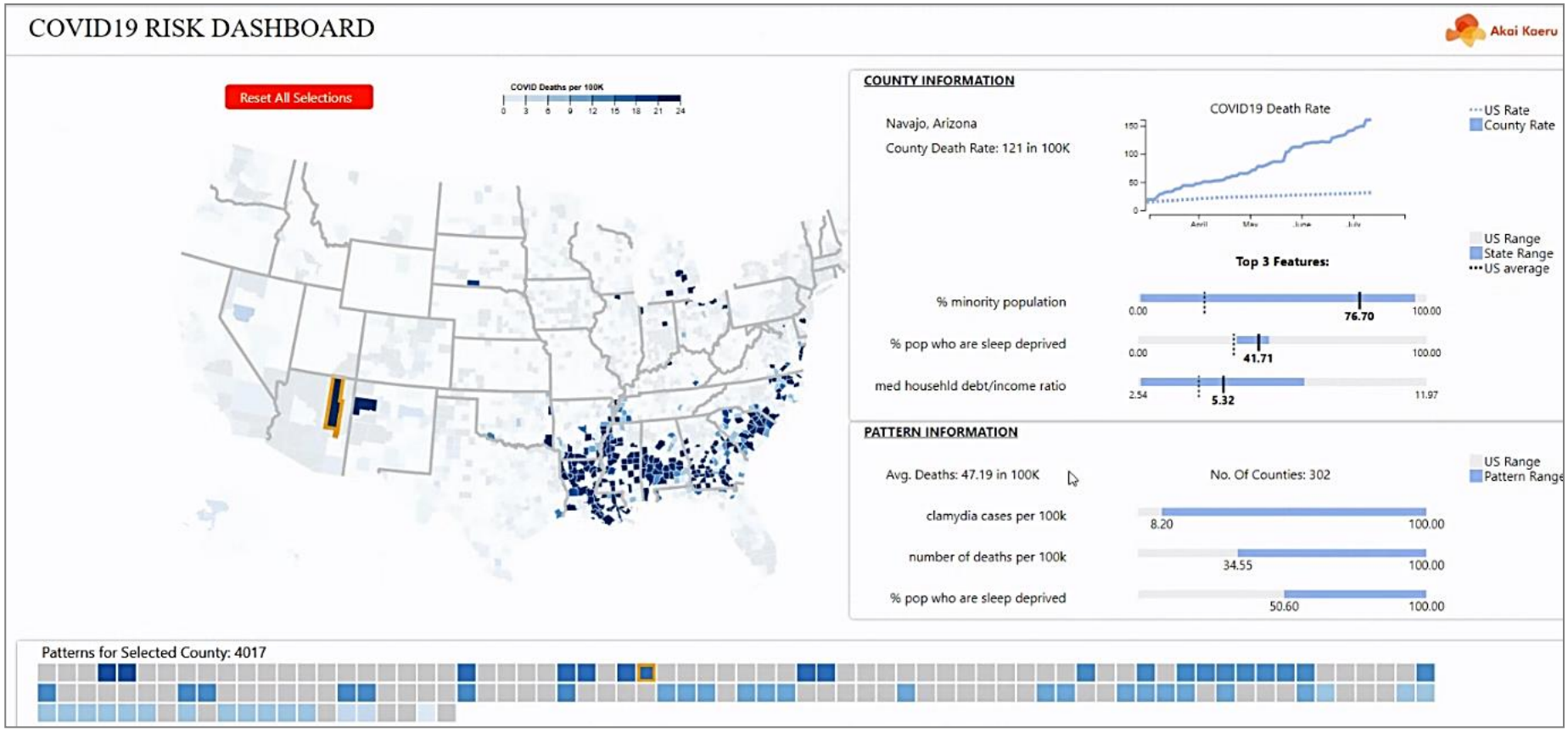
The affected counties are in the North East and at the big lakes.

▶ Weakened immune response

- stress & worries about debt
- low money → poor nutrition

Soon to Come

Interactive Web Browser-Based Dashboard



The dashboard supports the following assessments

- **Evaluate**: click on a county and see its risk profiles
- **Compare** : see what other counties have these risk profiles
- **Predict:** project what death rate might be on the horizon
- **Review:** see the risk profiles in context of the overall US

Virus Mitigation Recommendations

Learn from other counties what to do next

- complete lockdown or just close bars, restaurants?
- how much routine cleaning and disinfection?
- how much protective gear and what?
- how strongly to enforce social distancing?

Again, we can learn from data

- find patterns of counties where a certain strategy worked (or not)
- look which of these patterns your county fits to
- predict what will work and what will not