DATA SCIENCE
JOURNEY

MIDAS Data Science Summer Camp
July 15, 2022

Stephen **Salerno**

Who am I?

I am a PhD student in biostatistics at UM, studying the survival of patients with disease and how healthcare quality is publicly reported

What was my spark?

My first data science research project was to develop methods and analyze how tuberculosis can be detected in endemic areas of the world

How do I get involved?

I volunteer with a student group called Statistics in the Community (STATCOM), which is a community outreach program for data science consulting

A Little About Me

Growing Up

I was born in Cobleskill, a small (~6,000 people) farming town in Upstate NY

My sister, Justine, and I are very close, despite our seven year age difference

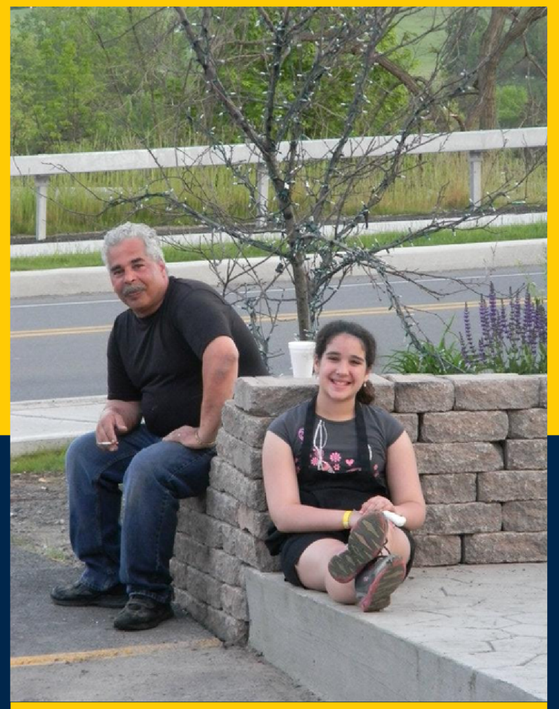
Growing up, I played sports, instruments and was active in various clubs

I loved math and its applications in high school, but did not know what someone could 'do with math' as a profession

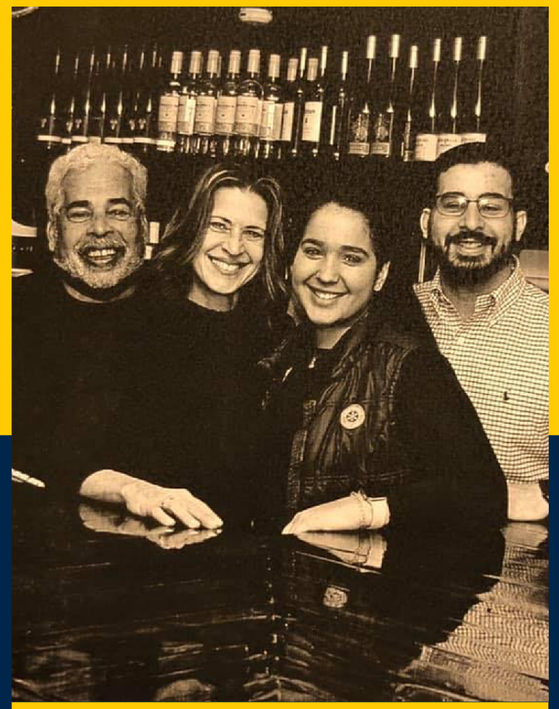
At the encouragement our math teacher, three friends and I competed in a math modeling competition (Moody's Mega Math Challenge)

Namesake

Many Memories



Taking a Break



In the News

THE RESTAURANT

Growing up, my parents owned a restaurant, which had a big part in shaping my identity. In a lot of ways, I 'grew up' working there, and I owe a lot of where I am today (both directly and indirectly) to the restaurant

How I Got Here

And how the restaurant inadvertently helped

An Unexpected Introduction

After being accepted to college as a math major, I learned about data science while waiting on tables

A Pragmatic Switch

I changed my major to biometry and statistics on the advice of a stranger (and the tuition was cheaper)

A Blessing In Disguise

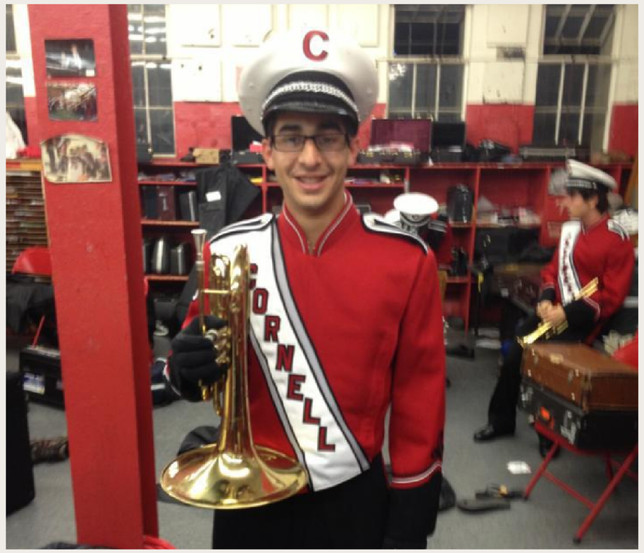
After working late, I missed class sign-ups and needed to take a grad-level class, where I met my mentor

A Perfect Decision

I didn't realize it yet, but I would find something that I was passionate about in biostat and data science

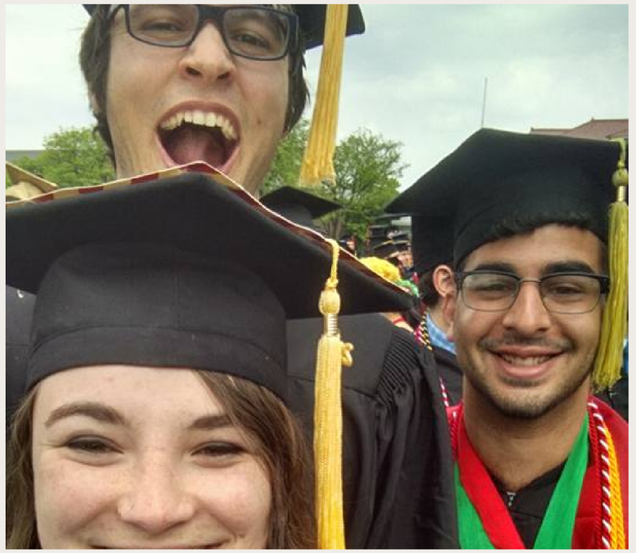
COLLEGE

Too many amazing memories and experiences to include here!



Marching Band

True to My Roots



Class of 2016

But, most important for this presentation, I learned about **data science** and **research**

What Is Data Science?

"Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. "

Source: https://www.datarobot.com/wiki/data-science/

"All models are wrong, but some are useful"

Source: George Box

Summer Institute for Training in Biostatistics, Boston, MA

Data Science Descriptions from

Current Students

Data science is...

Using statistics, mathematics, and computer science to solve problems

Using data to...

Turn abstract questions into real-world solutions

Finding patterns...

Understand relationships between things and making predictions

REALLY, THOUGH

Data science is such a broad term, so can we make this concrete?

Not one size fits all, and there's lots of opportunity to branch out based on your interests

READ MORE

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

Medicine

In Biostatistics, we use data to predict the health outcomes of patients, test the effectiveness of new treatments, or study what factors make one person more at risk for disease or death than another person

Business

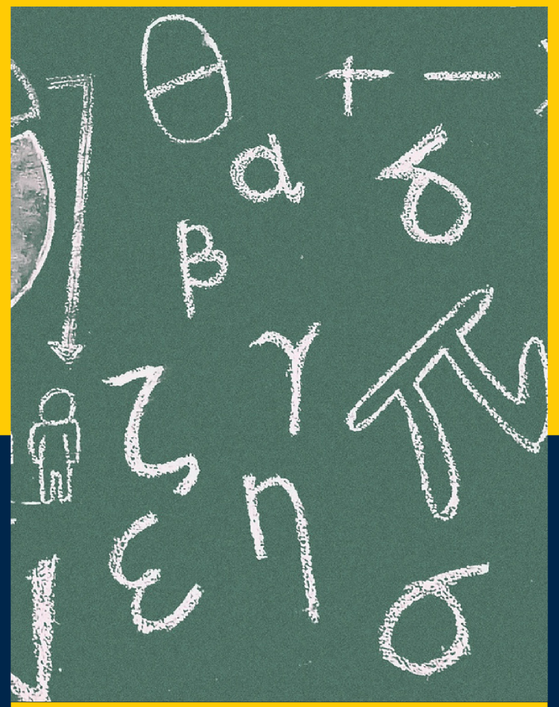
Businesses use data to map out deliveries, connect consumers to products and services that they might be interested in buying, and forecast profits or trends in the stock market over time and in different conditions

Sports

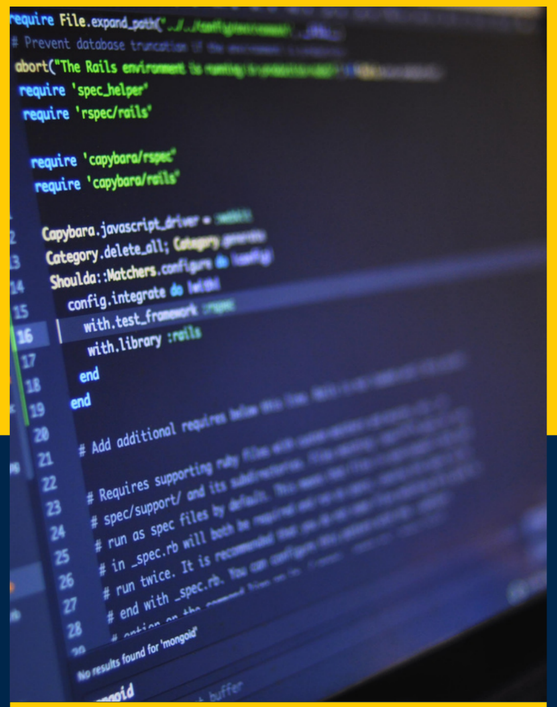
Statistics are used in sports analytics to track the performance of players, to predict who will win a game or tournament, or to evaluate roster picks when trying to build a successful team

Technology

Facial recognition, voice assist, biometric scanners, and many other examples of modern technology that use artificial intelligence are all built on fundamental data science and machine learning models



Mathematics



Computing

Cognates

Writing

WHAT DO YOU LEARN?

As a student studying data science or a related field,
you have the opportunity to take classes in many different areas

Data Science Major

Select Courses at the University of Michigan

Introductory Programming
Programming and Elementary Data Structures
Data Structures and Algorithms

Calculus 1-3
Linear Algebra
Discrete Mathematics

Introduction to Probability and Statistics
Applied Regression Analysis
Machine Learning and Data Mining
Data Management and Applications
Data Science Applications Electives

Capstone Experience

Careers

Many Different Titles

Data Analyst, Programmer, Data Scientist, Data Engineer, Machine Learning Engineer, Research Scientist, Statistician, Professor

Many Different Industries

Academia, Public Health, Medicine, Government, Technology, Law, Sports, Finance, Social Media, Commerce, Insurance

Many Different Opportunities

These are all just some examples! Data science is everywhere and is consistently ranked as one of the best career paths you could choose!

MY START

I was fortunate to have wonderful mentors in colleges, who taught me what it meant to do research. My first project was analyzing how tuberculosis can be detected in areas of the world like Haiti and Vietnam

READ MORE

https://research.cornell.edu/news-features/when-call-statistician

Project

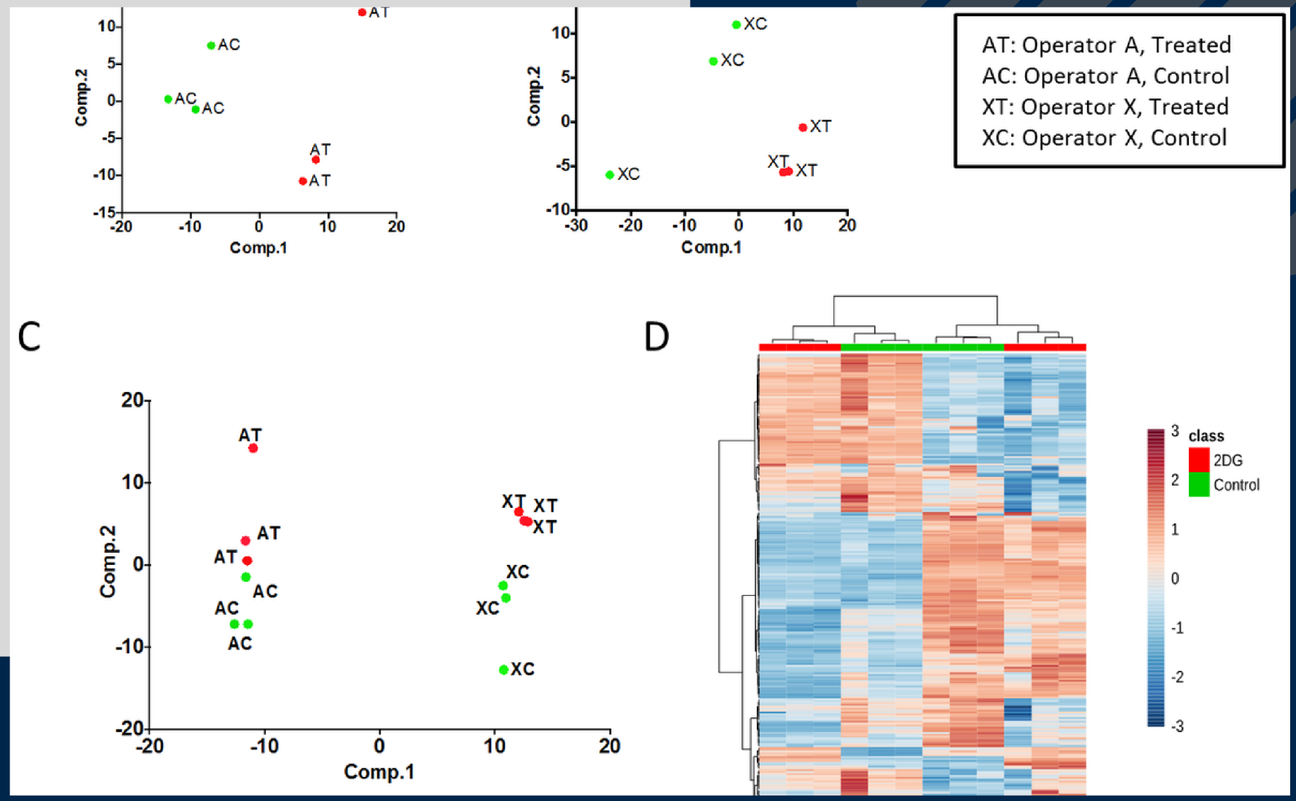
Understanding Batch Effects in Metabolites

There is interest in using metabolism for studying numerous biomedical contexts

However, variation in data that occurs independently of true variation (i.e. batch effects) can be substantial

We investigated a number of algorithms for batch effect correction and differential abundance analysis, and compare their performance

We further introduce an algorithm—RRmix—and illustrate its suitability for differential abundance analysis in the presence of strong batch effects



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179530

MCNAIR SCHOLAR

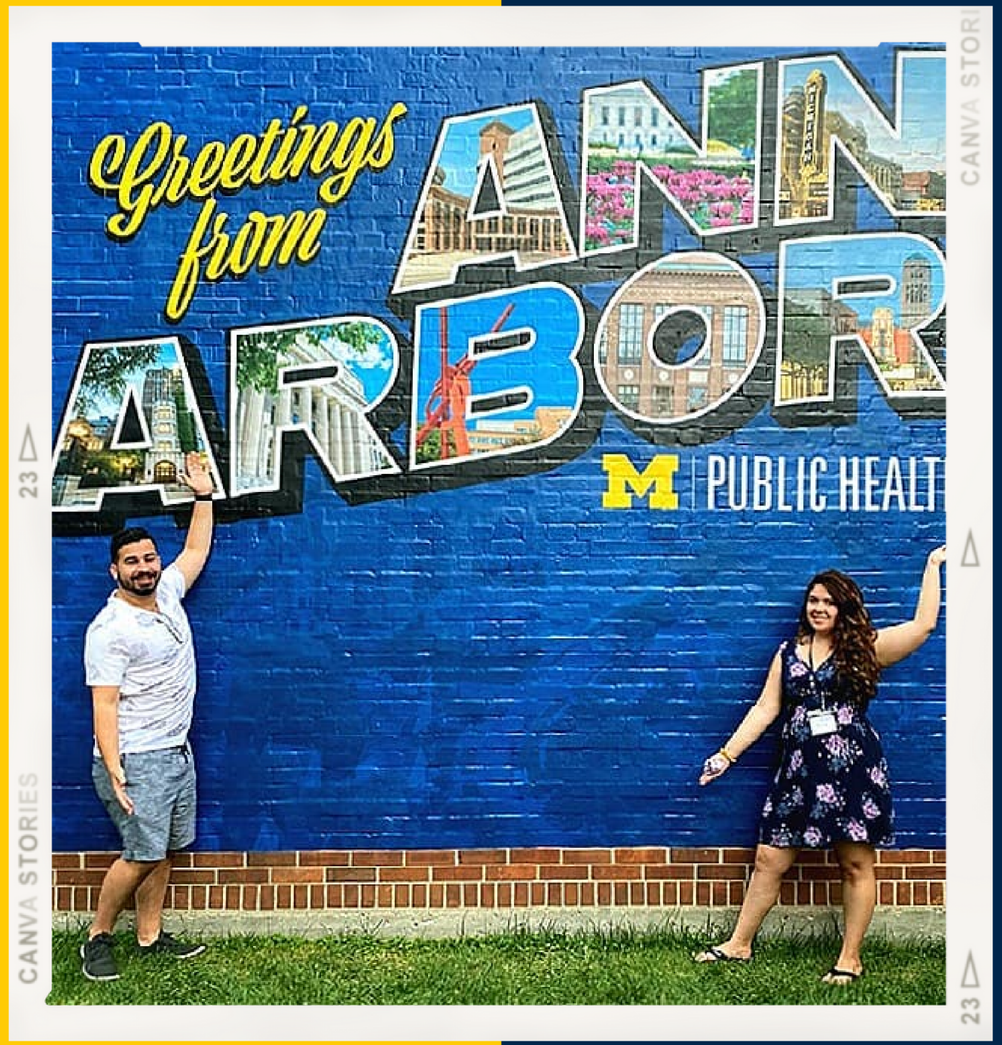
As a first-generation college student, I was encouraged to apply for McNair Fellowship, a post-baccalaureate achievement program for PhD-bound scholars

Ronald E. McNair

Presenting Research

Graduation

https://oadi.cornell.edu/signature-programs/ronald-e-mcnair-post-baccalaureate-achievement-program



GRAD SCHOOL

Thanks to my mentors in college, I was encouraged to continue on to graduate school, and I applied to Biostatstics here at U of M!

READ MORE

https://sph.umich.edu/biostat/

My Research

And My Academic Family

I work with Dr. Yi Li, who is an expert in an area of statistics called survival analysis

We are coming up with new ways of predicting the survival of patients with cancer and other diseases

We need mathematics and computer programming to develop these methods using "deep learning"

An important part of research and doing a PhD is coming up with new ideas and methods that have an impact on society

Having an incredibly supportive mentor has been an important part of my experience

NEURAL NETWORKS

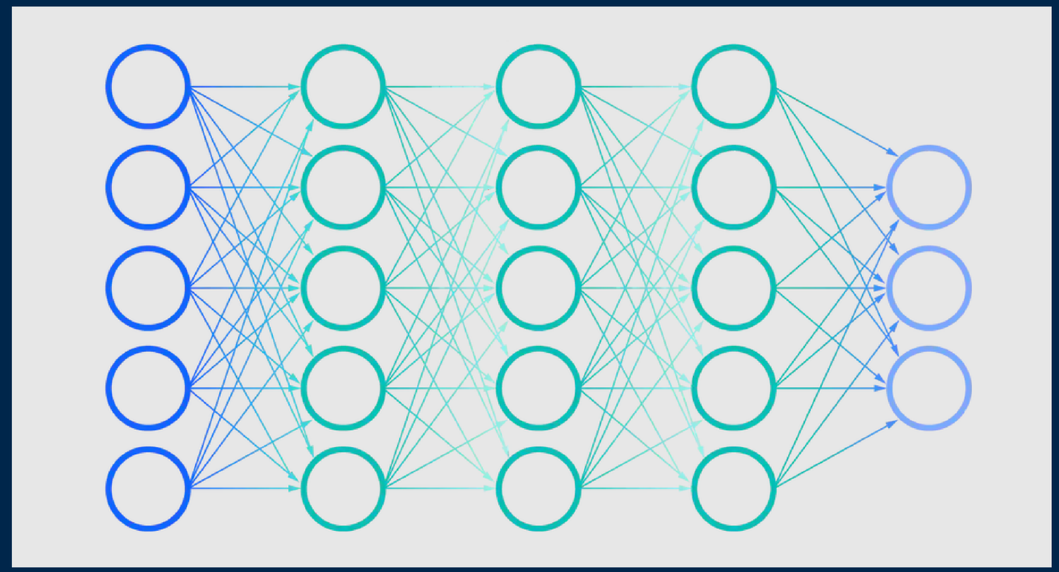
Neural networks reflect the behavior of the human brain, allowing computer programs to recognize patterns and solve common problems in the fields of AI, machine learning, and deep learning.

STRUCTURE

Comprised of an input layer, one or more hidden layers, and an output layer. Each node, or 'neuron', connects another and has an associated weight and threshold.

USE

Use training data to learn and improve predictive accuracy over time and are powerful tools in computer science and artificial intelligence, speech and image recognition, and Google’s search.



https://www.ibm.com/cloud/learn/neural-networks

Lung Cancer

Prevalence

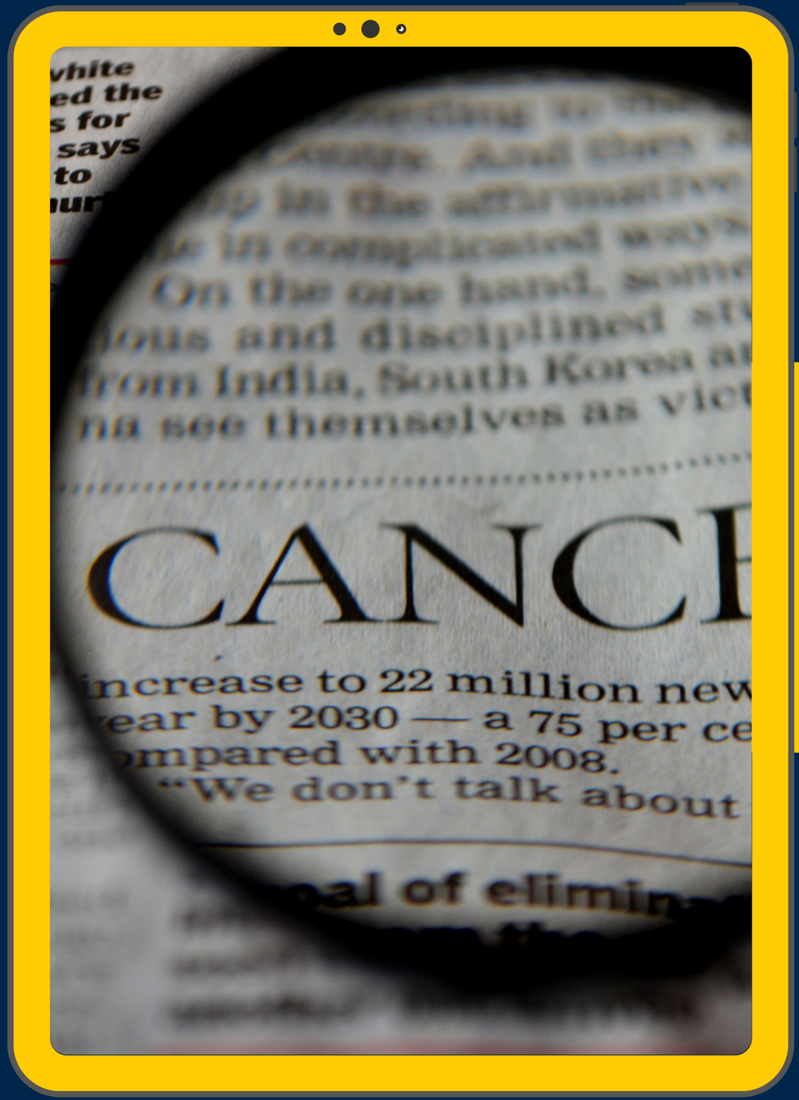
Lung cancer remains one of the leading causes of cancer-related deaths to date, with a 5-year survival rate of approximately 1 in 5

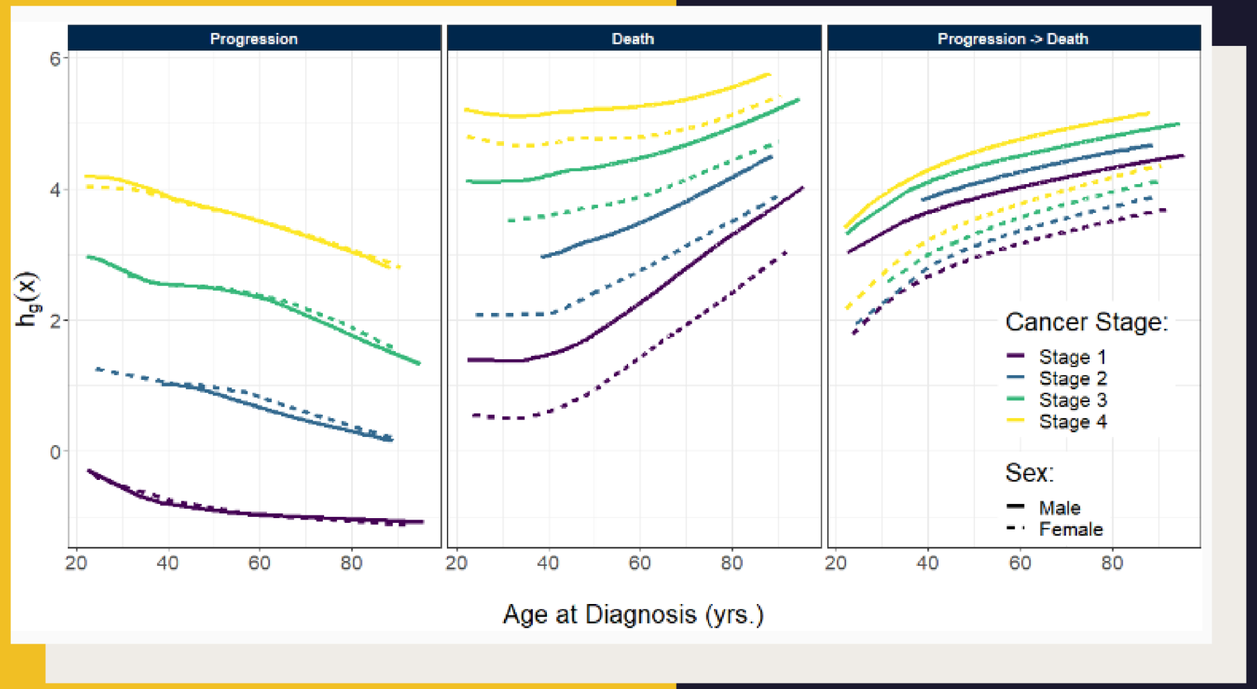
Prognosis

Prognosis varies greatly and depends on several individualized risk factors including smoking status, genetic variants, and other comorbid conditions

Disease Trajectory

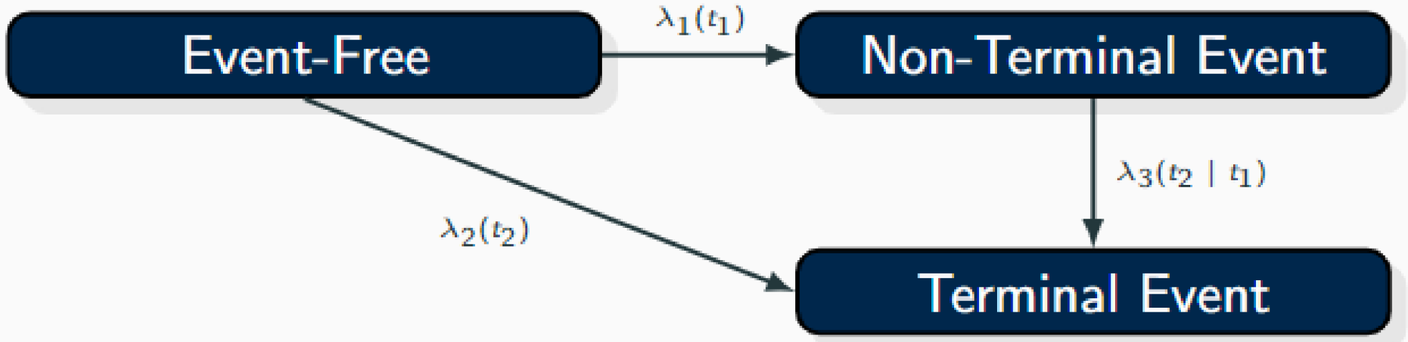
Patients diagnosed with lung cancer may experience a disease progression, go into remission, or have a recurrence prior to death





MODEL

Our approach is based on the illness-death model, and we predict cancer progression and death based on factors such as smoking status and age



1.

COVID-19 OUTCOMES

In a first project with the hospital here at U of M, we looked at the different types of outcomes for patients over the course of their COVID-19 infection

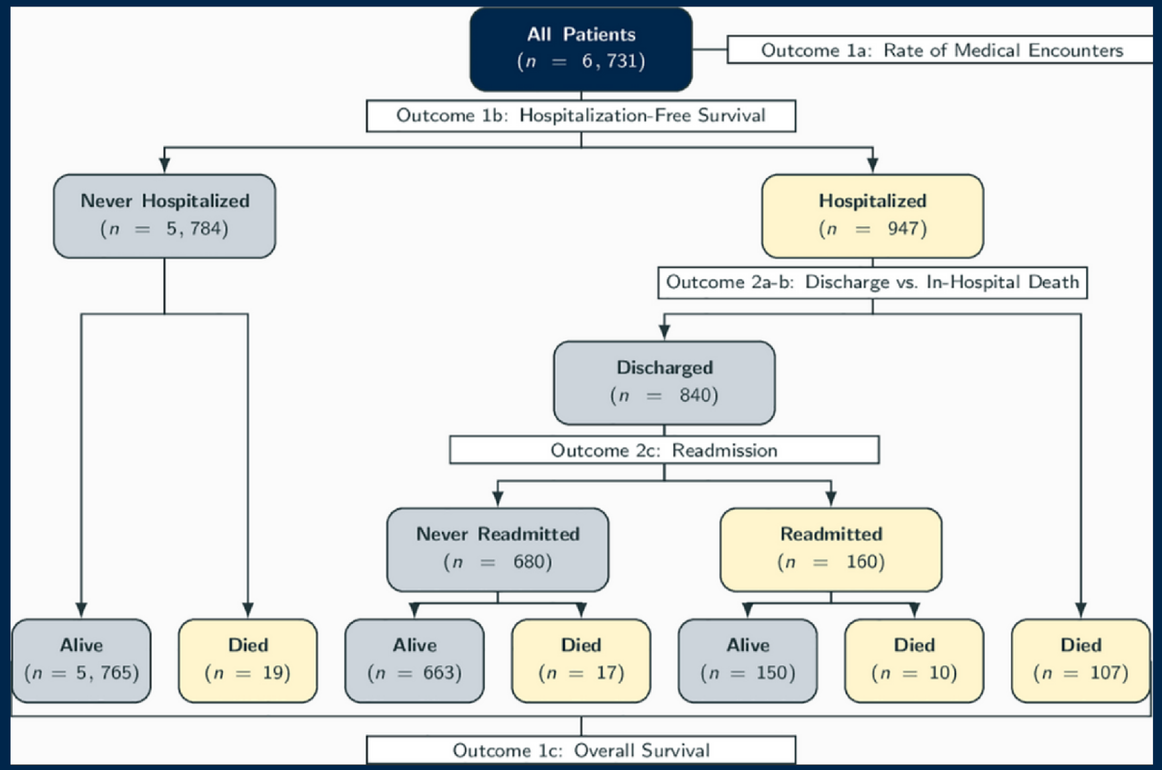
2.

PREDICTING SURVIVAL

In a second project, we looked at how we can use information about patients when they are admitted to the hospital, as well as their actual chest x-ray images to predict their survival

COVID-19

Our group has also been dedicated to understanding the impact of COVID-19 on certain populations in Southeast Michigan throughout the pandemic



Star Ratings

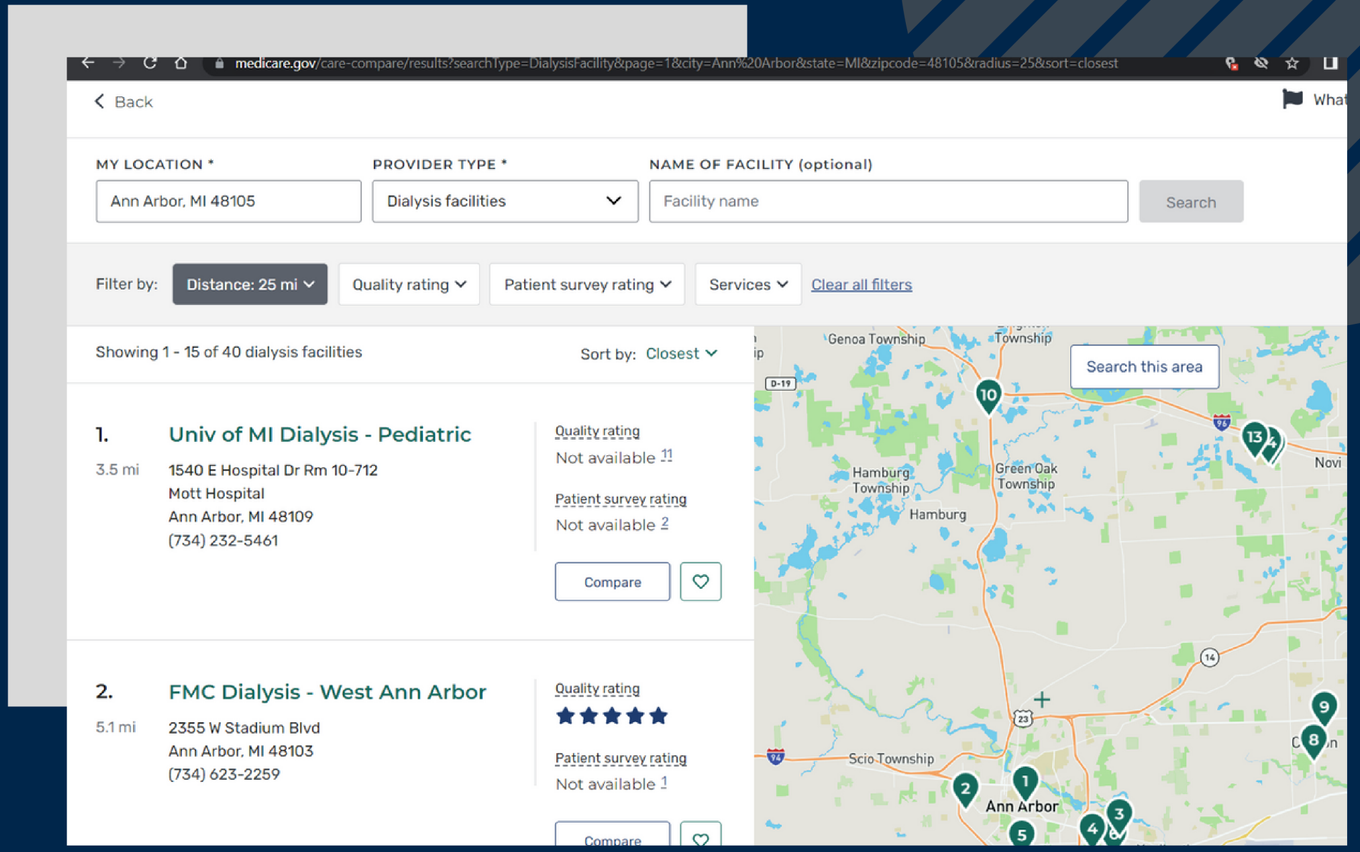
And the Kidney Epidemiology and Cost Center

Another important project that I work on uses statistics to develop a rating system for patients on dialysis to use

The group that I work with collects data and develops measures of healthcare quality that relate to dialysis, hospitalizations, and death

We turn these measures into a five-star rating that is easier for patients and their families to interpret

Patients can go online and use this rating, like Yelp, to find a dialysis provider in their area that meets their needs



THE BEST THING ABOUT BEING A DATA SCIENTIST IS THAT YOU GET TO PLAY IN EVERYONE'S BACKYARD

- JOHN TUKEY

Data for Public Good Symposium, 2019

SERVICE

One thing I am very passionate about is using my time and skills to help others through 'data for good' efforts and projects

READ MORE

https://sph.umich.edu/pursuit/2018posts/working-in-everyones-backyard.html

Why Care About DATA FOR GOOD?

https://sph.umich.edu/biostat/statcom/

The Need

Non-profit organizations and humanitarian causes, for example, can benefit from answering data-driven questions

The Avenue

Just like Teach for America or Doctors without Borders, data scientists donate their time and skills in meaningful ways

The Solution

I work with a group, Statistics in the Community (STATCOM) here at U of M

About STATCOM

STATCOM (Statistics in the Community) at the University of Michigan is a community outreach program provided by graduate students in the Departments of Biostatistics, Statistics, and others at University of Michigan. The program offers the expertise of graduate students, free of charge, to non-profit and community organizations in the areas of data organization, analysis, and interpretation.

READ MORE

https://sph.umich.edu/biostat/statcom/

Bloomberg Data for Good Exchange, 2018

Services We Provide

Advice and assistance are offered on a wide variety of statistical issues including:

Decision Making

Use of data to improve decision making processes

Surveys

Survey/sample design and analysis

Design

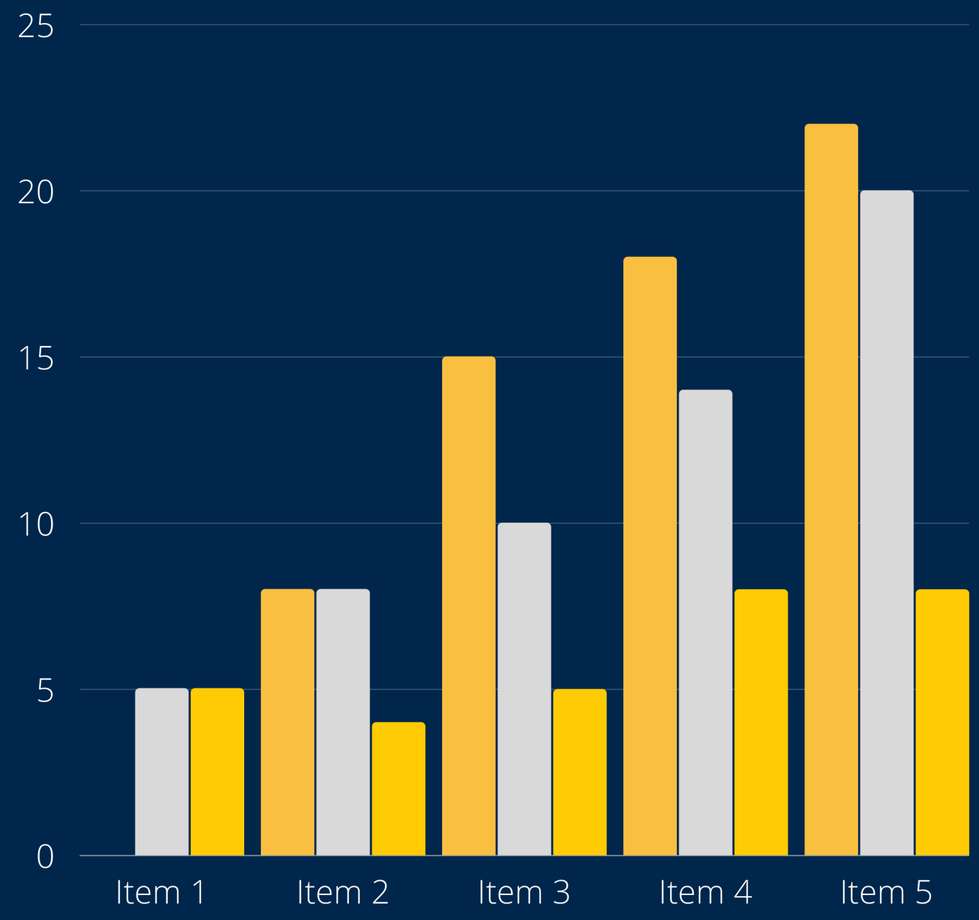
Design and analysis of studies and experiments

Visualization

Graphical methods of summarizing and gaining meaning from data

Prediction

Use of data to detect trends and make predictions and projections.



PROJECT

Find the optimal pantry locations and order they should be visited to best allocate resources to households across the city within a given month

RESULT

Food for Thought is now scheduling routes for their mobile pantries based on our recommendations and is reaching more people than ever before

FOOD FOR THOUGHT

Food for Thought is a non-profit that serves 400+ families experiencing food insecurity in Toledo, Ohio each month through a mobile pantry service

The Data

Demographics

Families were asked to report an address and other demographic data such as how many people lived in their household

Health Outcomes

Several health characteristics and food preferences were also surveyed to better gauge the needs of those being served

Socioeconomic Status

Neighborhood socioeconomic data were also collected using community surveys and governmental/census resources

1. UNDERSTANDING THE DATA

Clustering analysis used to group/rank areas with similar characteristics into need categories; based on regional poverty, unemployment, and health characteristics. Distances traveled to a pantry used to minimize total distance traveled, constrained to higher priority neighborhoods and other limiting factors

2. OPTIMIZING PANTRY LOCATIONS

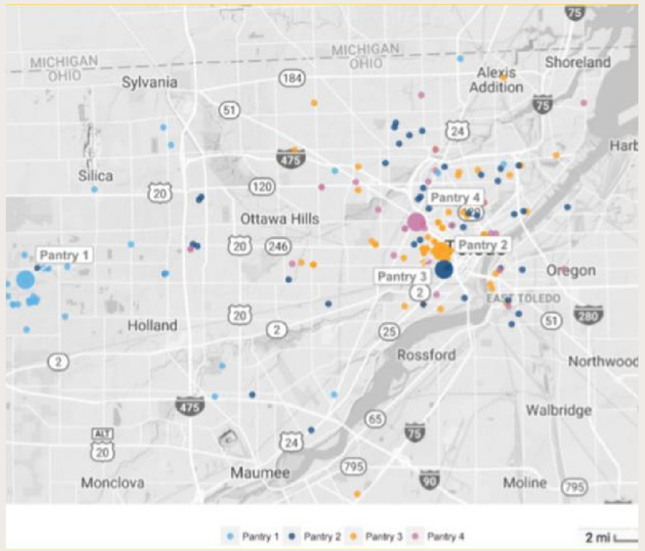
Location modeling used to create network of demand nodes, optimized locations of accruing unmet demand with constraints related to the number of visits per month and the needs of the households covered by a candidate location

PROJECT

Having collected data on the performance of 35 different locations for several years, they sought to identify optimal pantry locations

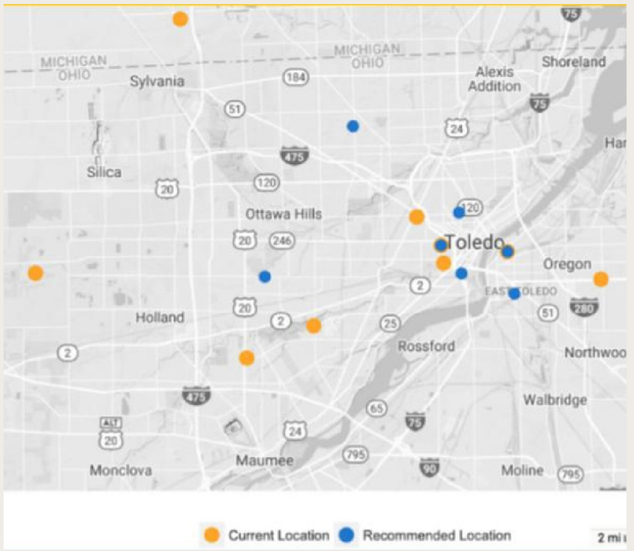
HOW WE MAPPED IT

We used a statistical model to find the most 'optimal' pantry locations



Previous Pantry Locations

Statistical Model



Recommended Locations

https://data.bloomberglp.com/company/sites/2/2018/09/Data-for-Good-In-Your-Neighborhood.pdf

Cassidy

Molly + Lola

Running

Family

OUTSIDE OF SCHOOL

Finding time in life and having balance are important for your success and overall health and wellbeing!

Four Things

Some Advice to Leave With Today

01 Do What You Love

You don't have to be perfect at something to love it, and doing what you love will bring you more success and happiness than doing something you think is safe or smart.

02 You Are Not 'Bad' At Anyting

Like playing a sport or an instrument, math takes practice. If you think you are not good at something, in reality, you probably just haven't done it enough yet.

03 Practice Balance

Set aside time for you. Excercising, reading, cooking, sleeping, or anything else that is important to you should have priority in your day, just as studying and productivity does.

04 Build Strong Communities

I am very fortunate for the people in my life and the people who left impacts on my journey along the way. Surround yourselves with good people who uplift you.

THANK YOU!
ANY QUESTIONS?

Feel free to reach out to me at any time if you want to talk more!

salernos@umich.edu