DATA SCIENCE - PROJECTS

CSC 405/605

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

PROJECT 1

INFORMATION TECHNOLOGY SERVICES :: G SUITE METRICS

Mentor: Nick Young
Enterprise Analytics Architect
nickyoung@uncg.edu

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

ITS: Infrastructure Analytics Overview

- New group in ITS at UNCG (January 2019)
- **“Organizing Information Technology Services’ data to make it useful, accessible, actionable and aligned with data-informed decision making ideals.”**
- 1 Graduate Student for 2019 (already hired)
- Currently interviewing for 2 full time positions
  - Integration Specialist & Data Scientist
- A lot of historical data
  - Homegrown systems, Splunk, ServiceNow, others
- Working on a data pipeline for Google Cloud Platform for analytics (BigQuery, Data Studio etc), but also will be developing/working with existing systems.

Data Science Student Project Overview

- Source Data: Google (G Suite) report metrics
  - Currently pulled automatically into Splunk from Google (daily)
- Data Set for Project
  - Size: 611,914 rows (CSV)
  - Date Range: March 23rd, 2015 to August 17th, 2019
  - Fields: timestamp, metric_name, metric_value
- Project Goals:
  - Analysis, Anomalies, Prediction
  - Visualization / simple dashboard development - Splunk and/or Google Data Studio preferred
- Primary Contact for Project:
  - Nick Young
    Enterprise Analytics Architect
    nickyoung@uncg.edu

Sample Data Screenshot

time	metric_name	metric_value
2015-03-25T00:00:00.000-0400	google.docs:num_7day_active_users	7468
2015-03-25T00:00:00.000-0400	google.docs:num_docs	1899985
2015-03-25T00:00:00.000-0400	google.docs:num_docs_edited	7178
2015-03-25T00:00:00.000-0400	google.docs:num_docs_externally_visible	84246
2015-03-25T00:00:00.000-0400	google.docs:num_docs_internally_visible	1815783
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_edited_for_12months	1031157
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_edited_for_3months	1613176
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_edited_for_6months	1388920
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_viewed_for_12months	810481
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_viewed_for_3months	1484887
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_viewed_for_6months	1203530
2015-03-25T00:00:00.000-0400	google.docs:num_docs_shared_outside_domain	44
2015-03-25T00:00:00.000-0400	google.docs:num_docs_viewed	14113
2015-03-25T00:00:00.000-0400	google.docs:num_docs_with_visibility_anyone_with_link	78631
2015-03-25T00:00:00.000-0400	google.docs:num_docs_with_visibility_people_at_domain	2878
2015-03-25T00:00:00.000-0400	google.docs:num_docs_with_visibility_people_at_domain_with_link	26493
2015-03-25T00:00:00.000-0400	google.docs:num_docs_with_visibility_private	1786412
2015-03-25T00:00:00.000-0400	google.docs:num_docs_with_visibility_public	5571
2015-03-25T00:00:00.000-0400	google.docs:num_drawings	2084
2015-03-25T00:00:00.000-0400	google.docs:num_drawings_edited	6

PROJECT 2

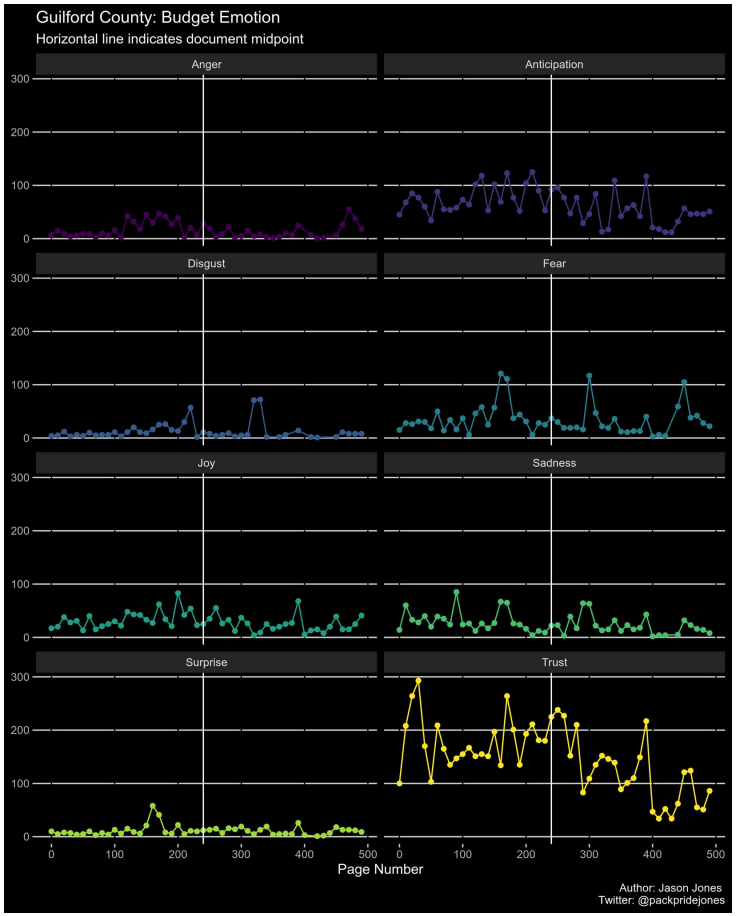
ANALYZING BUDGET TEXT

Mentor: Jason Jones
Analytics and Innovation Manager, Guilford County
jjones6@guilfordcountync.gov

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

Budget Text Processing

- City of Charlotte, Mecklenburg County, Wake County, City of Raleigh, Guilford County, City of Durham, and Durham County
- Emotion and Sentiment Analysis
- Topic Modeling
- Next word(s) recommender



PROJECT 2

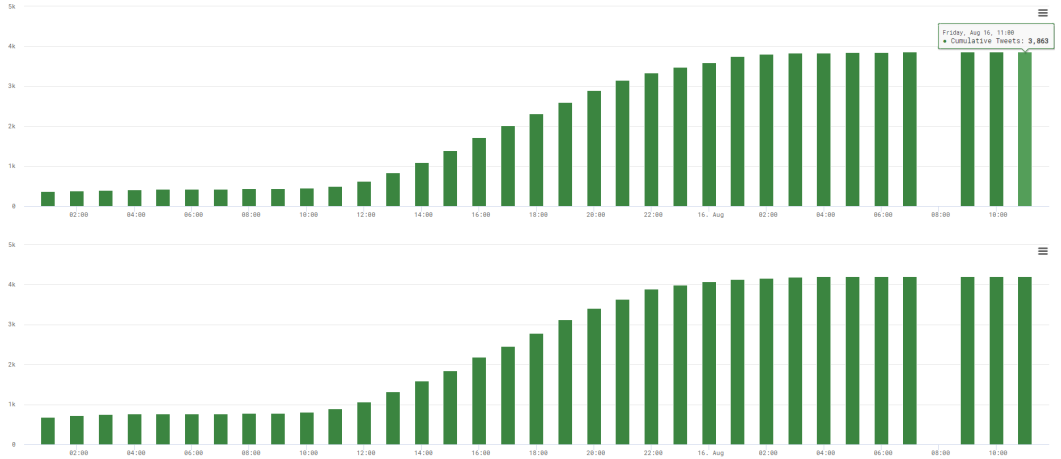
ANALYZING TWITTER GOVERNMENT INTERACTION

Mentor: Jason Jones
Analytics and Innovation Manager, Guilford County
jjones6@guilfordcountync.gov

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

ELGL #CityHallSelfie Day Twitter Data

- 3,863 tweets with 90 different attributes
- Network analysis
- Topic Modeling
- Spatial Analysis?
- Engagement Dashboard



PROJECT 3

ANALYSIS OF POST-STORM RESPONSE IMAGERY

Mentor: Dr. Evan Goldstein
Assistant Professor, Dept. of Geography
ebgoldst@uncg.edu

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

Classification and analysis of post-storm response imagery

Evan Goldstein
@ebgoldstein

Research Scientist
UNCG Geography, Environment, and Sustainability

32 NOAA datasets
37 USGS datasets

each with
~1000 images

Surf City, NC after Hurricane Florence images from Sept.18th

Phase 1
- dashboard to tag images (i.e., to make training data)
- developed vs. undeveloped
- coast vs. river
- washover? erosion?

Phase 1.5
- Evan and Coastal colleagues frantically tag images

Phase 2
- Build and test(Binary) Classifier with tagged images

Phase 3
- Descriptive Statistics on full dataset

Possible add on ideas:
- Deploy for 2019 Hurricane events
- Segment images

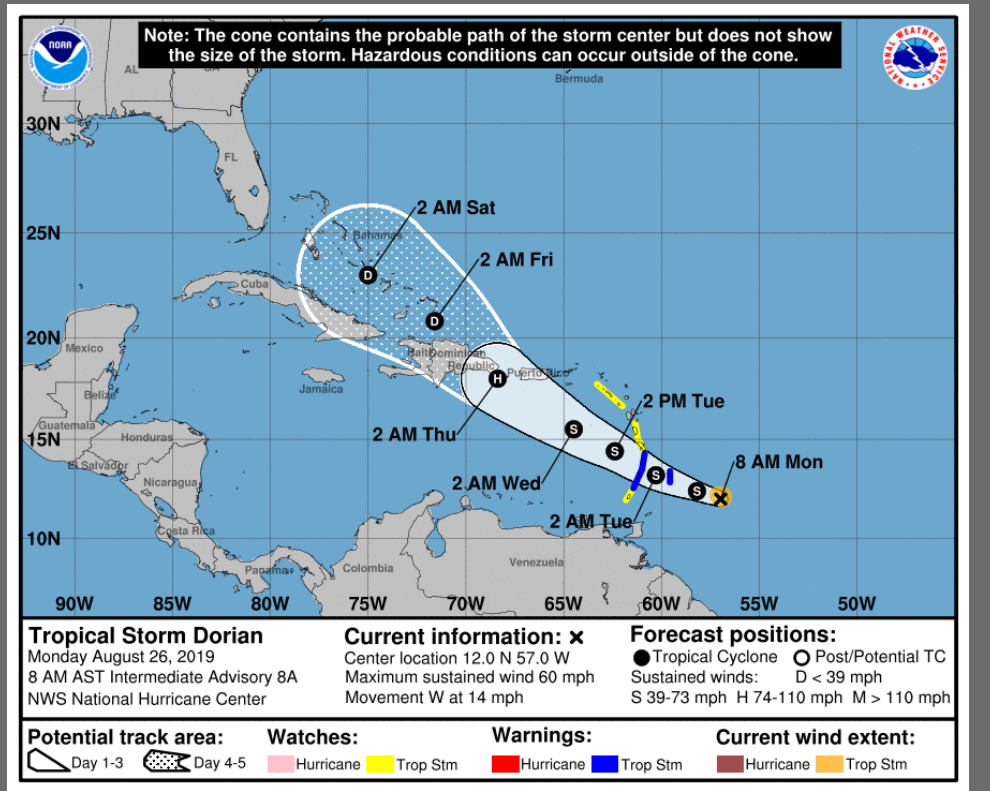
National Oceanic and Atmospheric Administration
U.S. Department of Commerce

August 8, 2019

HOME NEWS & FEATURES

NOAA increases chance for above-normal hurricane season

The end of El Nino could boost Atlantic hurricane activity



PROJECT 4

ANALYZING PEER REVIEW FOR EARTH SCIENCES

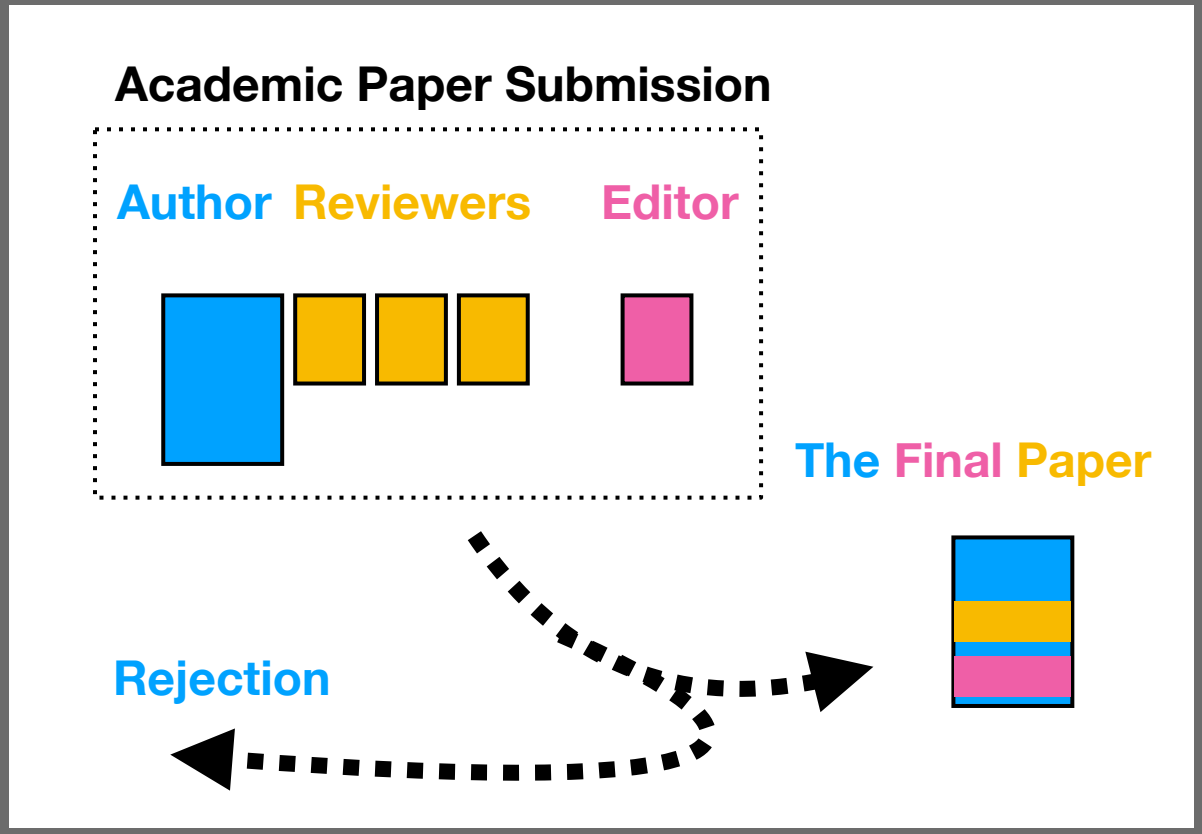
Mentor: Dr. Evan Goldstein
Assistant Professor, Dept. of Geography
ebgoldst@uncg.edu

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

Preliminary analysis of a Peer Review Dataset from the Earth Sciences

Evan Goldstein
@ebgoldstein

Research Scientist
UNCG Geography, Environment, and Sustainability



~100s of paper
~5 journals

Phase 1 (wrangling)

- scrape papers (XML) and reviews (XML or PDF) from journal website
- pdf -> easier format -> tokenization

Phase 2 (stats)

- percentage of Signed vs Unsigned reviews
- review corpus as a function of sub-discipline
- Most frequent words and n grams
- Sentiment analysis of words

Possible ML ideas:

- semantic similarity for multiple reviews?
- Text summary with an Attention model?
- sequence to sequence RNN for automatically writing a review
- another NLP task?

PROJECT 5

OPEN SOURCE SOFTWARE VULNERABILITY MITIGATION

Mentor: Dr. Stephen Tate
Professor, CSC
srtate@uncg.edu

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

PROJECT IDEA - 5

- **Datasets:**
  - Software Heritage Graph Dataset (SHGD)
    - "Software Heritage is the largest existing public archive of software source code and accompanying development history: it currently spans more than five billion unique source code files and one billion unique commits, coming from more than 8 million software projects." [approx. 1 TB as a download, or available through Amazon Athena]
  - National Vulnerability Database (NVD)
    - Information on over 120,000 distinct security vulnerabilities (software bugs that have security consequences). Each vulnerability identifies software affected, severity, and other information.
- **Goals:**
  - Cross-reference vulnerabilities with SHGD commits/nodes
    - How long between NVD filing and commit fixing the bug?
    - How long between fix and new software release?
    - Relate project activity to speed of fixes (responsiveness score?)
  - Identify unlabeled commits as corresponding to specific vulnerabilities
- **More context:**
  - SHGD was selected as dataset for the annual "Mining Software Repositories" conference challenge (https://2020.msrconf.org/)
  - Specific "call for challenge papers" with suggested questions due out by the end of this week
  - Successful class project could result in a paper to the conference (paper deadline will probably be early February)

PROJECT 6

VOLVO TRUCKS - LONG HAUL VERSUS REGIONAL

Mentor: Daniel Wingo
Data Analytics Engineer at Volvo
daniel.wingo@volvo.com

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

Volvo Trucks North America

- Volvo Logger Data:
  - There are 2 data sets:
    - Long Haul
    - Regional
- Data Type:
  - Time series data that is collected from two Customer Trucks.
  - These have a mixture of CAN data and sensor data.
- Goals:
  - Compare the two trucks and find unique features that separate the trucks using the channels given.
  - Look at the APU unit on the long haul and give back basic stats about the APU unit.
  - Other goals are TBD

PROJECT 7

PODKNOW - MONITORING PODCAST TRENDS AND PRODUCING PODCAST RATINGS BASED ON THE ACTUAL CONTENT OF PODCASTS.

Mentor: Dr. Aaron Beveridge
Assistant Professor, Department of English
akbeveri@uncg.edu

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

PodKnow Project Goals:

Step 1: Track data on top podcasts with following APIs

- Listen Notes API provides detailed data on podcasts
- iTunes API provides ratings data on podcasts

Step 2: Text mine podcasts

- Option 1: Test scalability of speech-to-text for podcast audio files
- Option 2: Download transcripts from disparate sources

Step 3: Rate podcasts based on the actual content of the podcasts

- Word and N-Gram Frequencies
- Topic models
- Trends

PROJECT 8

ANALYZE COLLISION DATA

Mentor: Sagar Iddyadinesh
DBA and EAM Enterprise Administrator, City of Greensboro
Sagar.Iddyadinesh@greensboro-nc.gov

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

Data

- 5 years of collision data that include pedestrian Collision and fatalities tagged with X,Y co-ordinates and includes Accident Cause and brief description
- Dataset also contains city's infrastructure data such as Pedestrian Signals, Regulatory and Warning street signs and Traffic Signal location.
- Augmenting spatial data such as Streets and Sidewalk can be provided.

Goal

- Broader goal is to improve infrastructure with smart data driven decisions.
  - Historical analysis of accident data.
  - Predictive analysis of pedestrian collision.
- Data can be provided either in CSV format or can be made available on Kaggle.
  - Tools such as QGIS can be used for spatial analysis.

Sample data set

ACCIDENT_CAUSE	COLLISION_DESC	WEATHER	LIGHT_CONDITIONS	ROAD_CLASS	TRAFFIC_CONTROL
IMPROPER BACKING / NO CONTRIBUTING CI	BACKING UP	CLEAR	DARK-LIGHTED ROADWAY	LOCAL	STOP AND GO SIGNAL
IMPROPER BACKING / NO CONTRIBUTING CI	PARKED MOTOR VEHICLE	CLEAR	DARK-LIGHTED ROADWAY	PUBLIC VEHICULAR AREA	NO CONTROL PRESENT
NO CONTRIBUTING CIRCUMSTANCES / ALC	REAR END, SLOW OR STOP	CLEAR	DARK-LIGHTED ROADWAY	LOCAL	STOP AND GO SIGNAL
INATTENTIVENESS /	BACKING UP	CLEAR	DARK-LIGHTED ROADWAY	LOCAL	NO CONTROL PRESENT
SPEED /	RAN OFF ROAD - LEFT	CLEAR	DARK-LIGHTED ROADWAY	LOCAL	NO CONTROL PRESENT
DISREGARDED ROAD MARKINGS /	RAN OFF ROAD - LEFT	CLOUDY	DARK-LIGHTED ROADWAY	LOCAL	NO CONTROL PRESENT
OTHER*	RAN OFF ROAD - RIGHT	CLEAR	DARL-LIGHTED ROADWAY	LOCAL	NO CONTROL PRESENT
DISREGARDED ROAD MARKINGS /	RAN OFF ROAD - LEFT	CLEAR	UNKNOWN	LOCAL	NO CONTROL PRESENT
/	RAN OFF ROAD - RIGHT	CLEAR	DAYLIGHT	LOCAL	NO CONTROL PRESENT
UNABLE TO DETERMINE /	RAN OFF ROAD - LEFT	CLEAR	DAYLIGHT	INTERSTATE	NO CONTROL PRESENT
INATTENTIVENESS / NO CONTRIBUTING CIR	REAR END, SLOW OR STOP	CLOUDY	DAYLIGHT	LOCAL	STOP AND GO SIGNAL
FAIL TO REDUCE SPEED / FOLLOW TO CLO	REAR END, SLOW OR STOP	CLOUDY	DAYLIGHT	LOCAL	STOP AND GO SIGNAL
SPEED /	OVERTURN / ROLLOVER	CLEAR	DAYLIGHT	LOCAL	NO CONTROL PRESENT
OTHER / NO CONTRIBUTING CIRCUMSTANC	RAN OFF ROAD - STRAIGHT	CLOUDY	DAYLIGHT	PUBLIC VEHICULAR AREA	NO CONTROL PRESENT
UNABLE TO DETERMINE / UNABLE TO DETE	ANGLE	CLEAR	DAYLIGHT	LOCAL	STOP AND GO SIGNAL
SWERVE / OVERCORRECTED /	OVERTURN / ROLLOVER	CLEAR	DAYLIGHT	LOCAL	STOP AND GO SIGNAL
FAIL TO YIELD RIGHTAWAY / NO CONTRIB	ANGLE	CLOUDY	DARK-LIGHTED ROADWAY	LOCAL	STOP AND GO SIGNAL
NO CONTRIBUTING CIRCUMSTANCES INDIC	RIGHT TURN, DIFFERENT ROADWAYS	CLEAR	DARK-LIGHTED ROADWAY	LOCAL	FLASHING STOP AND GO SIGNAL
NO CONTRIBUTING CIRCUMSTANCES /	MOVABLE OBJECT*	CLOUDY	DARK-ROADWAY NOT LIGHTED	INTERSTATE	NO CONTROL PRESENT
EXCEEDED SAFE SPEED FOR CONDITIONS	RAN OFF ROAD - RIGHT	CLEAR	DARK-LIGHTED ROADWAY	LOCAL	NO CONTROL PRESENT
NO CONTRIBUTING CIRCUMSTANCES INDIC	ANIMAL	CLEAR	DARK-LIGHTED ROADWAY	LOCAL	NO CONTROL PRESENT
UNABLE TO DETERMINE /	RAN OFF ROAD - LEFT	CLEAR	DARK-UNKNOWN LIGHTING	PRIVATE PROPERTY	NO CONTROL PRESENT
NO CONTRIBUTING CIRCUMSTANCES / INAT	HEAD ON	CLEAR	DAYLIGHT	LOCAL	NO CONTROL PRESENT
IMPROPER BACKING /	BACKING UP	CLEAR	DAYLIGHT	LOCAL	NO CONTROL PRESENT
FAIL TO YIELD RIGHTAWAY / NO CONTRIB	ANGLE	CLEAR	DAYLIGHT	LOCAL	NO CONTROL PRESENT
NO CONTRIBUTING CIRCUMSTANCES / NO C	BACKING UP	CLEAR	DAYLIGHT	LOCAL	NO CONTROL PRESENT
NO CONTRIBUTING CIRCUMSTANCES / EQUI	REAR END, SLOW OR STOP	CLEAR	DAYLIGHT	LOCAL	STOP AND GO SIGNAL
NO CONTRIBUTING CIRCUMSTANCES / TRA	ANGLE	CLEAR	DAYLIGHT	LOCAL	STOP AND GO SIGNAL
NO CONTRIBUTING CIRCUMSTANCES /	ANIMAL	CLEAR	DAYLIGHT	INTERSTATE	NO CONTROL PRESENT
FAIL TO YIELD RIGHTAWAY / NO CONTRIB	ANGLE	CLEAR	DAYLIGHT	LOCAL	NO CONTROL PRESENT
NO CONTRIBUTING CIRCUMSTANCES / FAIL	REAR END, SLOW OR STOP	CLEAR	DARK-LIGHTED ROADWAY	LOCAL	STOP AND GO SIGNAL
FAIL TO REDUCE SPEED / FOLLOW TO CLO	REAR END, SLOW OR STOP	CLEAR	DARK-LIGHTED ROADWAY	INTERSTATE	NO CONTROL PRESENT
FAIL TO REDUCE SPEED / FOLLOW TO CLO	REAR END, SLOW OR STOP	CLEAR	DARK-LIGHTED ROADWAY	INTERSTATE	NO CONTROL PRESENT
NO CONTRIBUTING CIRCUMSTANCES / FAIL	ANGLE	CLEAR	DUSK	LOCAL	NO CONTROL PRESENT
NO CONTRIBUTING CIRCUMSTANCES / UNA	PARKED MOTOR VEHICLE	CLEAR	DARK-LIGHTED ROADWAY	PUBLIC VEHICULAR AREA	NO CONTROL PRESENT
FAIL TO YIELD RIGHTAWAY / NO CONTRIB	LEFT TURN, SAME ROADWAY	CLEAR	DARK-LIGHTED ROADWAY	LOCAL	STOP AND GO SIGNAL
NO CONTRIBUTING CIRCUMSTANCES / FAIL	HEAD ON	CLEAR	DARK-LIGHTED ROADWAY	PUBLIC VEHICULAR AREA	NO CONTROL PRESENT
INATTENTIVENESS / NO CONTRIBUTING CIR	SIDESWIPE, SAME DIRECTION	CLEAR	DARK-LIGHTED ROADWAY	LOCAL	STOP AND GO SIGNAL
INATTENTIVENESS / NO CONTRIBUTING CIR	REAR END, SLOW OR STOP	CLEAR	DARK-LIGHTED ROADWAY	PUBLIC VEHICULAR AREA	NO CONTROL PRESENT
SWERVE / OVERCORRECTED / NO CONTRIB	ANGLE	CLEAR	DARK-LIGHTED ROADWAY	INTERSTATE	NO CONTROL PRESENT

PROJECT 9

DETERMINATION OF THE PREVALENCE, INCIDENCE, AND IMPACT OF ALCOHOL AND SUBSTANCE MISUSE IN GUILFORD COUNTY

Mentor: Dr. Stephen Sills
Director, Center for Housing and Community Studies, UNCG
sjsills@uncg.edu

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

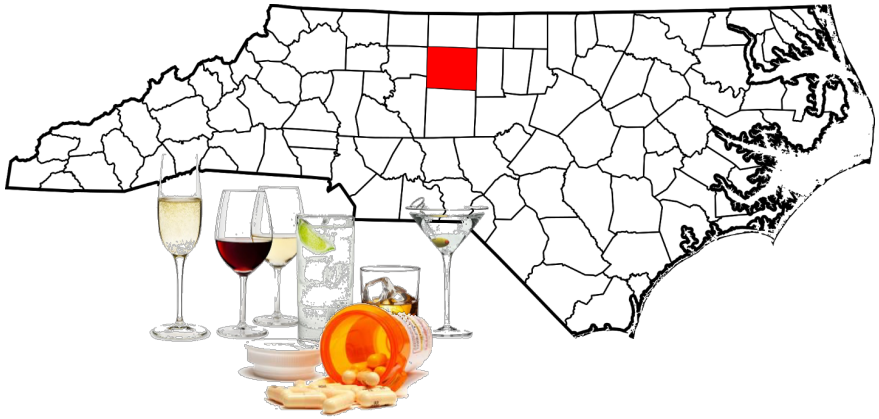
UNC GREENSBORO

Center *for* Housing & Community Studies

Committed to investigating the social, economic, environmental, and geospatial aspects of home and neighborhood and how they impact people's health, well-being, and life course opportunities.

Determination of the Prevalence, Incidence, and Impact of Alcohol and Substance Misuse in Guilford County

Life expectancy in the United States has been declining in recent years due in large part to alcohol, drugs, and suicide. Alcohol use has economic, health, and social costs to individuals, communities, and societies. Guilford County now averages 30 visits per week to the ED for heroin and opiate related overdoses. Locally we know that there is substantial impact on our local health care system, higher demand for law enforcement and emergency services, lost employment and lower tax receipts, increased use of health and human services, etc. Communities like ours are struggling to ascertain the full scope of the impact of these substances and to develop an adequate response to the crisis.



PI Dr. Stephen Sills, UNCG Center for Housing and Community Studies
Co-PI Dr. Ken Gruber, UNCG Center for Youth Family and Community Partnership
Co-PI Dr. Jeremy Bray, UNCG Department of Economics

Data soup…

CDC, ABC, ACS, NCDOT, HPD, GPD, GCS, EMS, GFD, HPFD, MCO, YRBS, etc…

3.80 GB 277 Files, 76 Folders

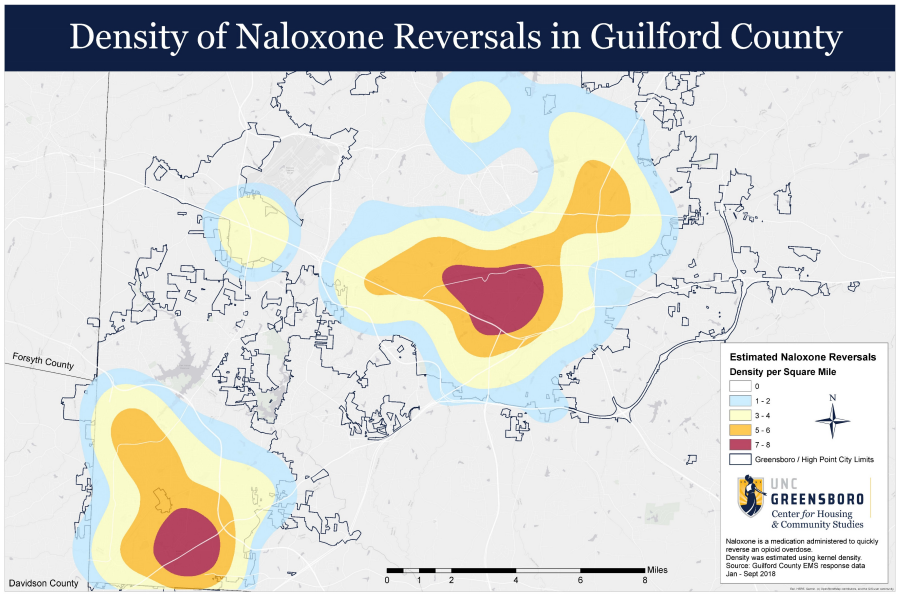
XLS, CSV, GIS, SHP, KLM, PDF

Analysis to date…

Compiled Data, geocoded all at block-group level, constructed GIS, conducting analysis report.

361 variables, 292 rows in the following categories:

- GEOGRAPHY (BLOCK GROUP)
- DEMOGRAPHICS
- EDUCATION
- SOCIO-ECONOMIC STATUS
- HEALTH & MEDICAL INFO
- MENTAL & BEHAVIORAL HEALTH UTILIZATION
- ENVIRONMENT
- 911 CALLS
- SUBSTANCE USE
- SUBSTANCE USE DISORDERS
- SUBSTANCE USE RELATED CRIME
- OTHER CRIME NOT SUBSTANCE USE RELATED
- COST ESTIMATES (ONGOING)



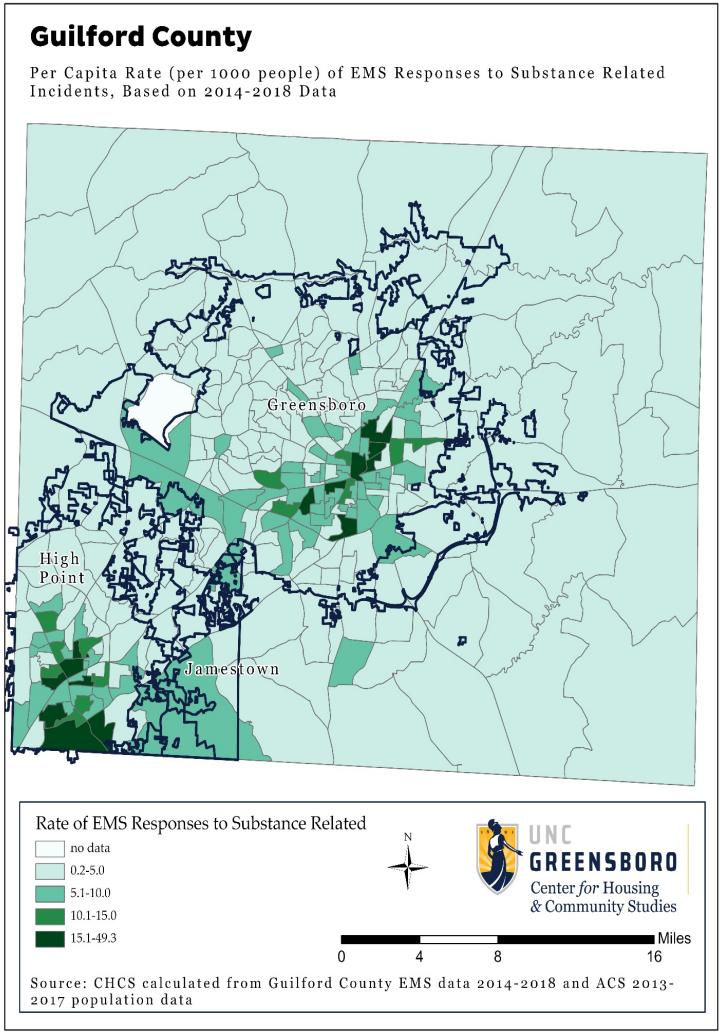
Guilford County Emergency Services Substance Use Response Rates

Data from Guilford County Emergency Services was compiled for the five-year period of 2014-2018. All responses were mapped and aggregated to the Census Block Group level. After annualizing the data, incidence rates per 1,000 were computed using total population per block group (ACS 2013-2017).

There were over 10,124 responses from Guilford County Emergency Services related to substance use during the five-year period; or about 2025 per year on average. The annualized rate for the entire county was 3.9 substance related responses per 1,000 population.

The lowest rate (.2 responses per 1,000 persons) was found in north central Greensboro in the vicinity of Kirkwood and Irving Park neighborhoods (Census Block Group 370810104041). The highest rate (49.3 responses per 1,000 persons) was found in South High Point near I-85 and Surrett Dr in an industrial area with a relatively low population of 710 persons and a notable transient population at several motels near the highway. Other areas of high response included:

- The central business district of Greensboro (Census Block Group 370810108002) where response rates were 35 per 1,000. This is likely related to the concentration of bars and clubs in the area, as well as a concentration of homelessness in vicinity of Center City Park; and
- The neighborhood south of Oak Grove and Southmont just north of I-40 and centered on Randleman Rd and W. Meadoview. This area, with a rate of 30.7 responses per 1,000, has a lower population (801 persons), but is also a commercial corridor along a major highway with many low-cost, off-brand motels, at least two Adult bookstore/theaters, and a history of street-level prostitution and drug dealing.



Project…

How can we make this data available to other researchers at UNCG?

Convert raw data sources to consistent format

Data cleaning, find duplications, unifying sources of data

Create a catalog or index of data

Consistent metadata format

Report on Descriptive data for each variable in each file (Code book)

Cloud based lookup

PROJECT SELECTION AND TEAMS

- **Teams**
  - Teams should consist of 5 members
  - Should include both graduate and undergraduate students

- **Project Assignment – Next Steps**
  - Establish a team
  - Select top 3 projects
  - Send me an email with
    - Subject: Data Science Final Project - 2019
    - Team Name (Choose a cool one)
    - Team Member Names – Github ID
    - Top 3 choices of projects
      - <Project Number – Project Title>

- **You will be given the project based on FCFS**