DSPG 2017 Project Reviews
July 12, 2017

Virginia Tech
Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Tentative Agenda

- 8:00 - 8:20 coffee & bagels upstairs
- 8:30 AM - NSF Projects:
  - NSF Open Source Software
  - NSF Business Innovation
  - NSF STEM Education
- 10 min Break
- 9:30 AM - Community Projects:
  - Alexandria - Equity
  - Fairfax - Youth Projections
  - Roanoke - Refugees
- 10 min Break
  - New Kent - Teens
  - Arlington - Needs Assessment
  - ARI Base Profiling
- ~12 PM Lunch Break - pizza provided -
- 12:30 PM Projects:
  - WMATA
  - AC 911
  - DHS
  - SCHEV

Measuring the Scope and Impact of Open Source Software

PI & Co-PI: Gizem Korkmaz & Stephanie Shipp
Team: Claire Kelling (Penn State), Sayali Phadke (Penn State)
Ben Swartz (VT), Eirik Iversen (VT), Ronnie Fecso (VT)
Sponsor: National Science Foundation (NSF) –
National Center for Science and Engineering Statistics (NCSES)

VirginiaTech
Biocomplexity Institute

SOCIAL & DECISION ANALYTICS LABORATORY

Introduction

- Can we measure the scope & impact of open source software (OSS) using publicly available data?
- OSS is currently not measured as a university research output (e.g., publications, patents)
- The project aims to test the feasibility and to develop methods to measure the impact (e.g., cost and benefit) of open source software (OSS).
- Goals:
  - Understand the "universe" of open-source software
    - How much is in use (stock), how much is created each year (flow)?
    - Created in which sectors? Who are the creators?
      - **OpenHub**: Commercial, Non-profit, Education, Government
    - Created in what categories? Who uses it?
      - **Sourceforge**: Science and Engineering, Games, Graphics, etc.
  - Case study: R packages in **Depsy**
    - Model-building to understand impact of individual software packages

Universe - OpenHub

- Online community and public directory of free and open source software
- Connected to source repositories and regularly updates information
- Information organized by:
  - Projects (n = ~676,523)
  - Organizations (n = 698)
  - People / Contributors (n = ~291,782)

Data Collection: Using API -- each key allows for 1000 calls per day (using IDs)

- **Project ID** can retrieve six tables:
  - Development activity, Contributors to project, Users of project etc.
- **Organization ID** can retrieve five tables:
  - Contributors, Projects, Organization information
- **User ID** can retrieve five tables:
  - User's geographical and professional details, Achievements, Projects used etc.

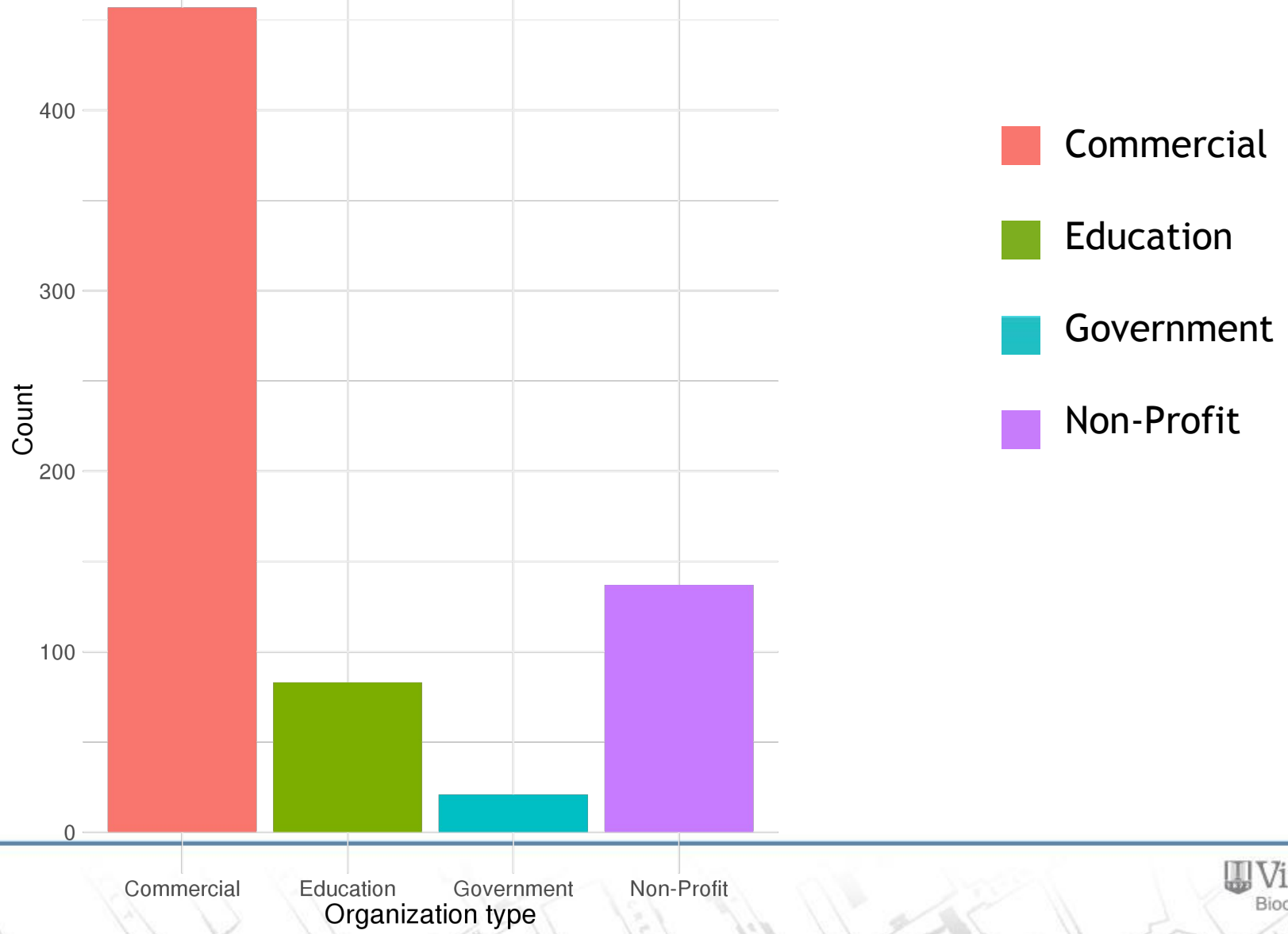
OpenHub: Projects

- Selected 33 variables to collect from API (e.g., description, user_count, tags)
- Two data collection strategies (for ~676,523 projects):
  - Random sample
    - Selected variables queried for random 79,813 projects (12% of all projects)
    - API returned data on 46,005 projects
      - This sample is 6.8% of all OpenHub projects
  - Most relevant projects
    - Selected variables queried for top 73,000 projects
    - API returned data on 55,217 projects
      - This sample is 8.2% of all OpenHub projects
- Most relevant project data tables are better populated than average

Column name	%Missing (Relevant)	%Missing (Random)
Overall data	24.36	42.36
created_at	0	0
updated_at	0	0
description	5.69	6.41
user count	0	0
average rating	75.68	95.48
rating count	0	0
review count	0	0
last analysis update	0	0
last source code access	0.81	0.38
ohloh 1st month of analysis	0	0
ohloh latest month of analysis	0	0

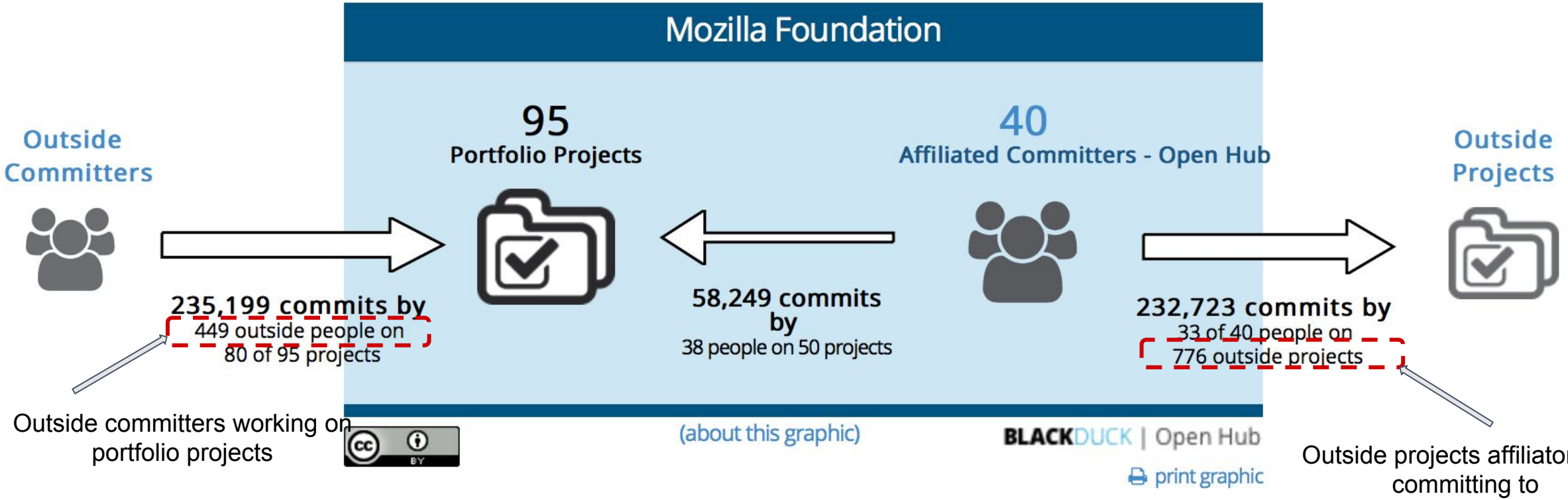
Column name	%Missing (Relevant)	%Missing (Random)
12 month contributor count	0	0
total contributor count	0	0
12 month commit count	0	0
total commit count	0	0
total code lines	0	0
main language	0.87	0.43
factoids	0.02	0.01
tags	51.61	70.18
licenses	36.29	53.06
languages	0.04	0.04
language percentages	0.04	0.04
activity index	0	0

OpenHub: Organizations (n=698)



VirginiaTech.
Biocomplexity Institute

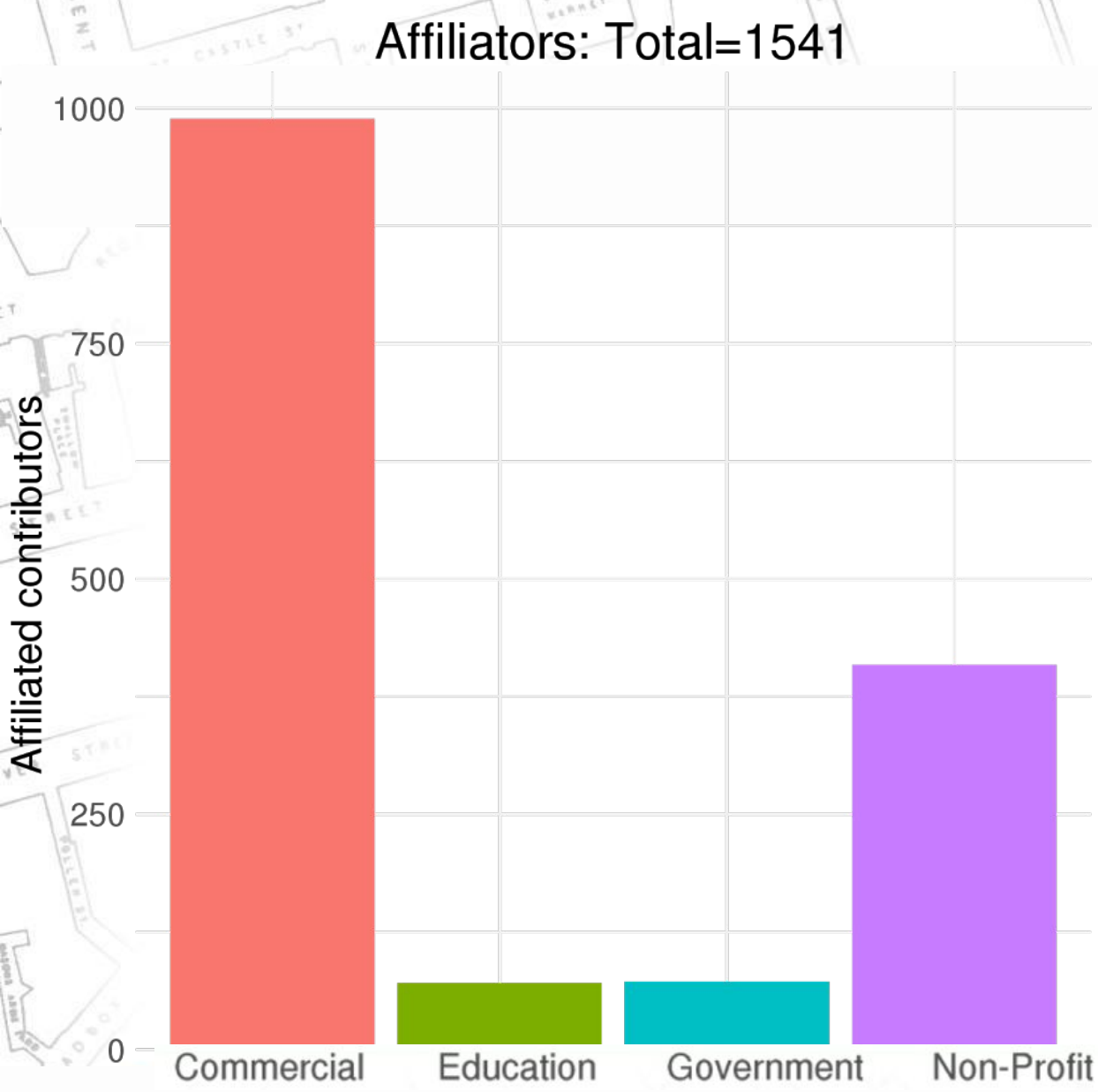
OpenHub: Organizations

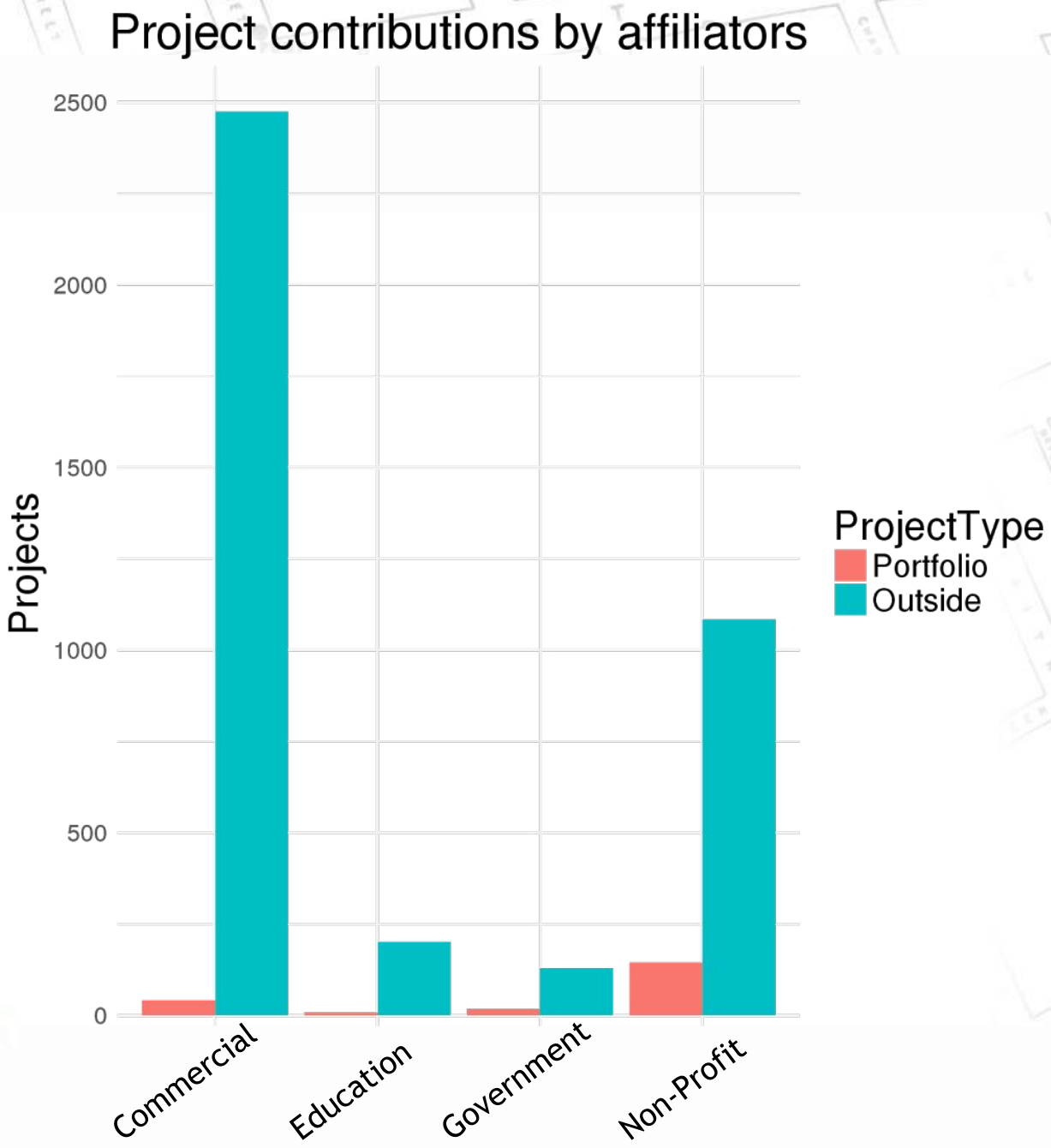


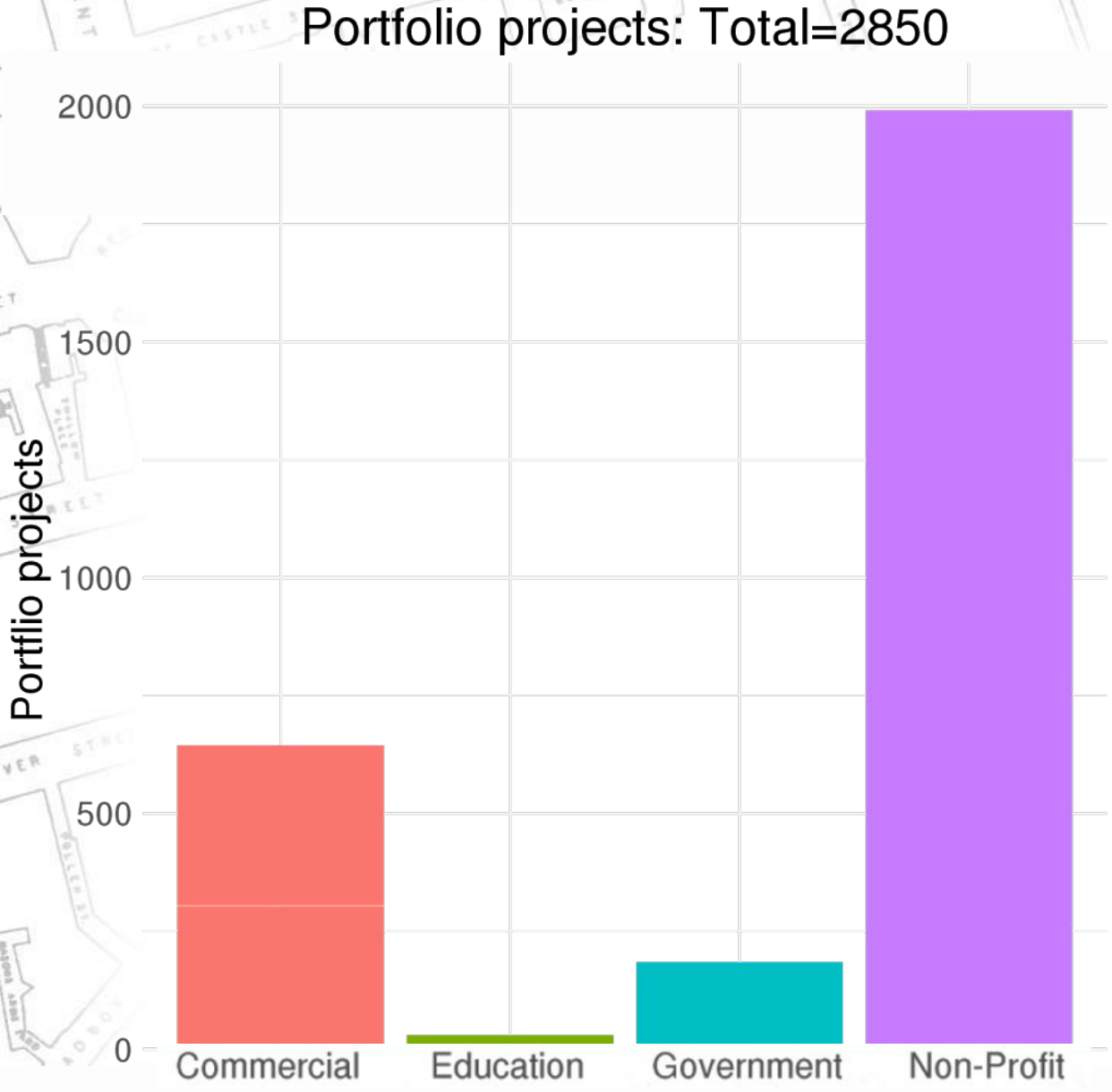
Outside committers working on portfolio projects

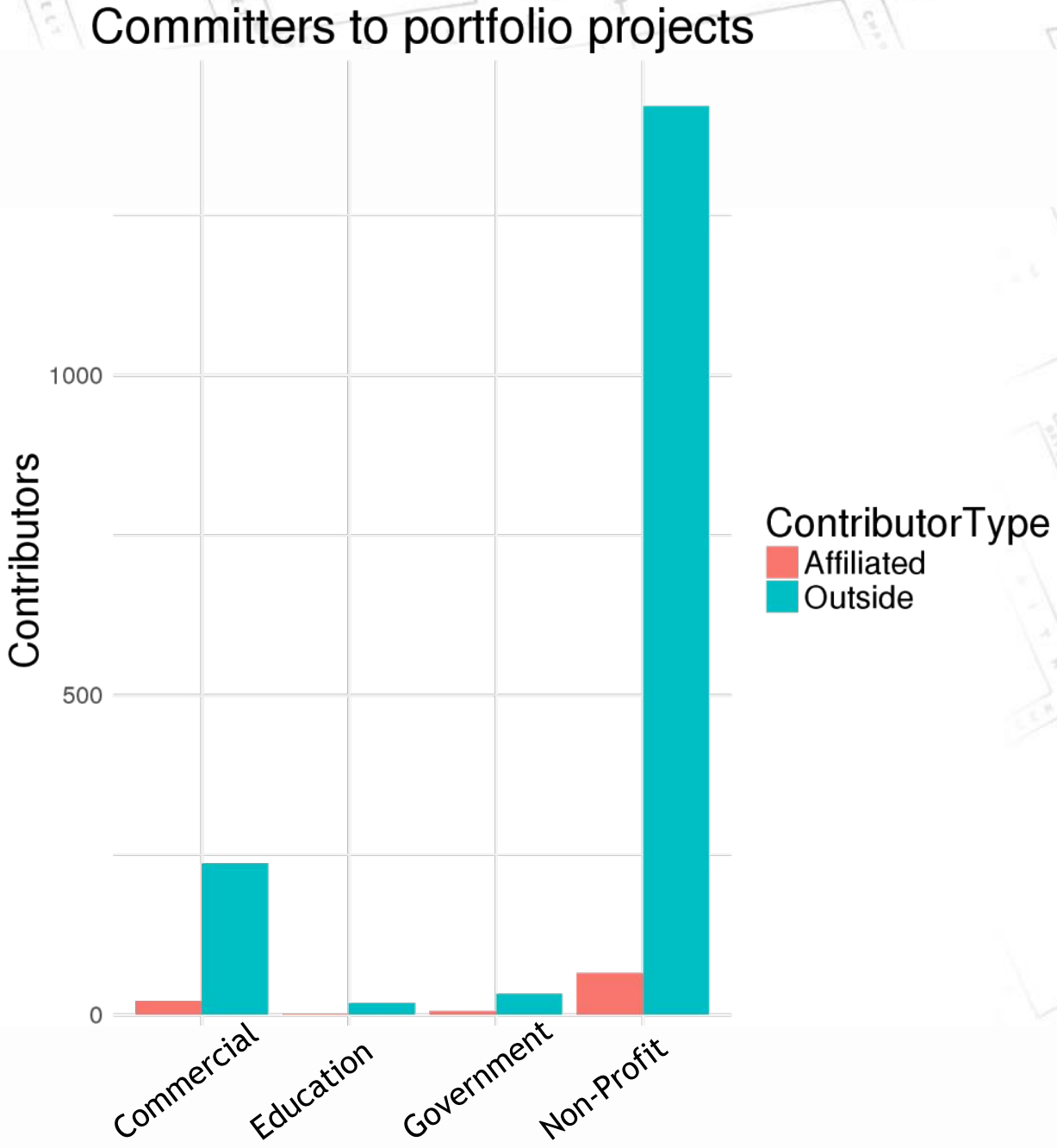
Outside projects affiliators are committing to

Source: https://www.openhub.net/orgs/mozilla


Affiliators: Total=1541
Affiliated contributors
1000
750
500
250
0
Commercial
Education
Government
Non-Profit


Project contributions by affiliators
Projects
2500
2000
1500
1000
500
0
Commercial
Education
Government
Non-Profit
ProjectType
Portfolio
Outside


Portfolio projects: Total=2850
Portflio projects
2000
1500
1000
500
0
Commercial
Education
Government
Non-Profit


Committers to portfolio projects
Contributors
1000
500
0
Commercial
Education
Government
Non-Profit
ContributorType
Affiliated
Outside

OpenHub

- **Challenges**
  - Computational complexity:
    - Lab generated 22 keys => 22,000 API queries each day
    - ~250 days to collect ALL information from OpenHub
    - ~45 days to collect main tables on all projects, users, and organizations
  - Data issues:
    - Website contains sparsely populated data on certain variables
    - API breaks frequently, leading to completely missing rows
- **Next steps**
  - Collecting user information from OpenHub
  - Accessing full OpenHub dataset (through NSF contact?)
  - Analyzing pertinent information such as timeline of OSS development, contributions, types of projects etc.

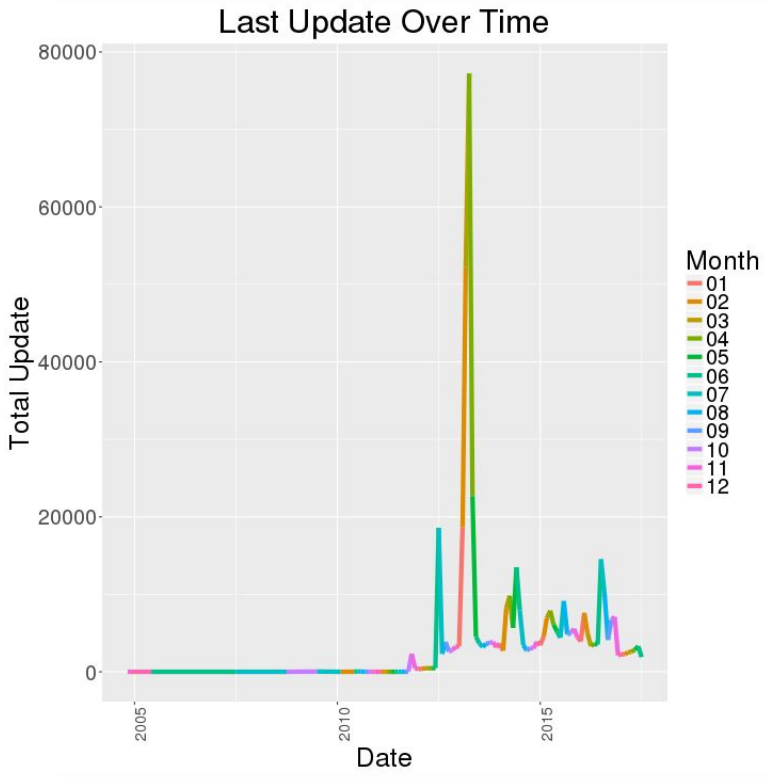
Sourceforge

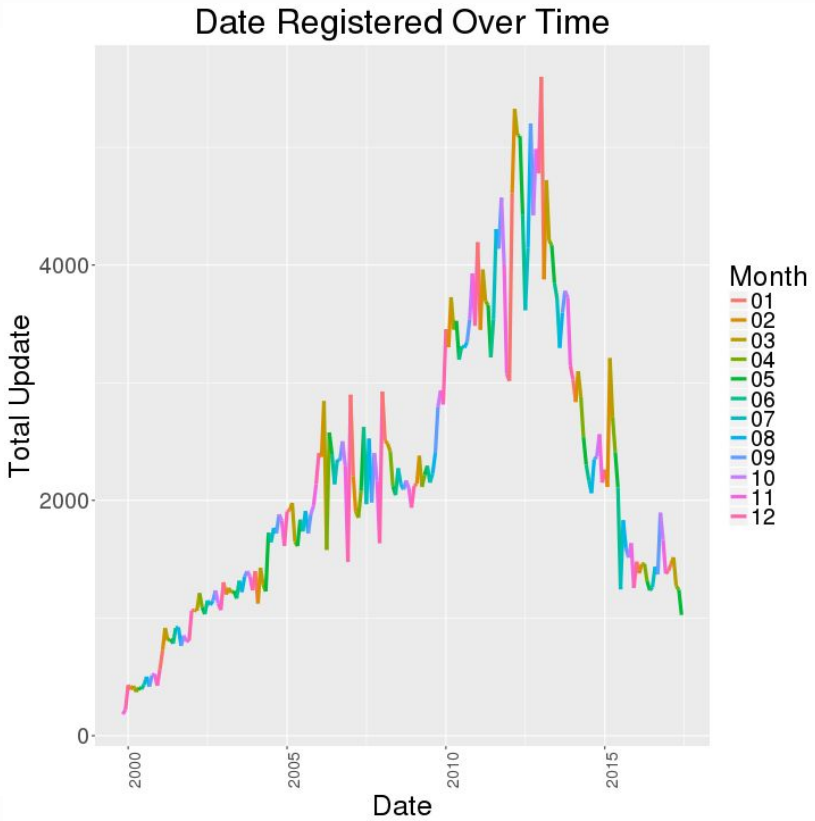
- An “Open Source community resource” dedicated to helping open source projects be as successful as possible
  - provides resources for development and distribution
- Projects are organized in a directory
  - Projects (n = ~488,000)
- Scraping steps
  - Use names provided by Sourceforge through their API
  - Scrape directly from the directories for each package
  - Took 15+ days of near continuous scraping

Sourceforge

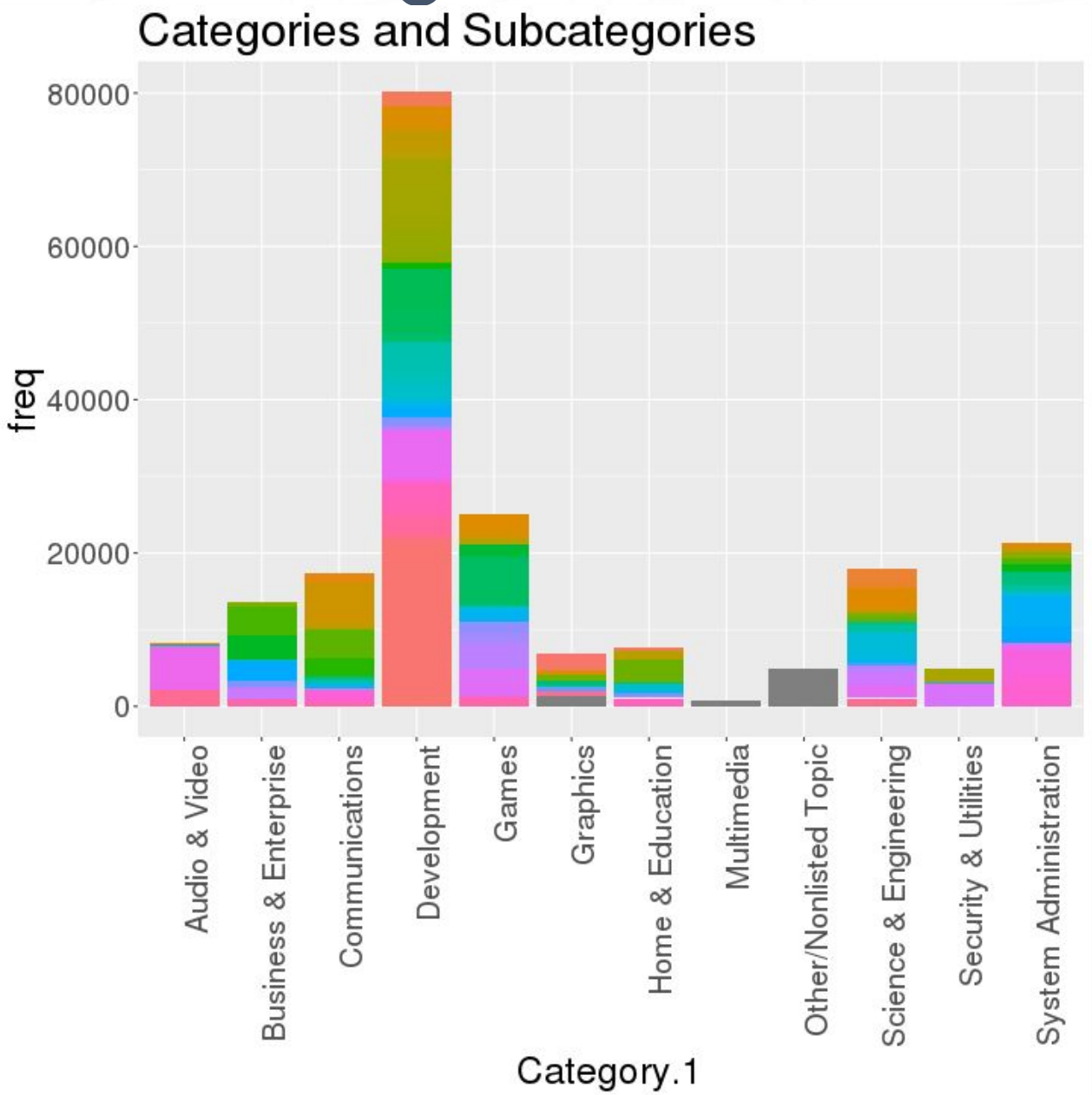
- **Data collection**
  - Pulled 488,000+ project names from the Sourceforge API
  - Number of authors: 352,504
  - Scraped information from each individual webpage

	Column Name	Class	% Missing	Unique Values
1	OSS.Title	character	0.02	449274
2	Average.Rating	list	92.44	132
3	Description	list	22.31	355483
4	Last.Update	list	0	2147
5	Number.of.Ratings	list	0	170
6	Weekly.Downloads	list	0.04	1448
7	Project.Type	character	0	2
8	Category.1	character	55	13
9	Category.2	character	56.2	196
10	Category.3	character	88.38	142
11	Date.registered	list	0	6380
12	Authors	character	3.73	352504
13	Ease	list	92.06	7
14	features	list	92.06	7
15	design	list	92.06	7
16	support	list	92.06	7
17	Total.Downloads	numeric	0.04	27173



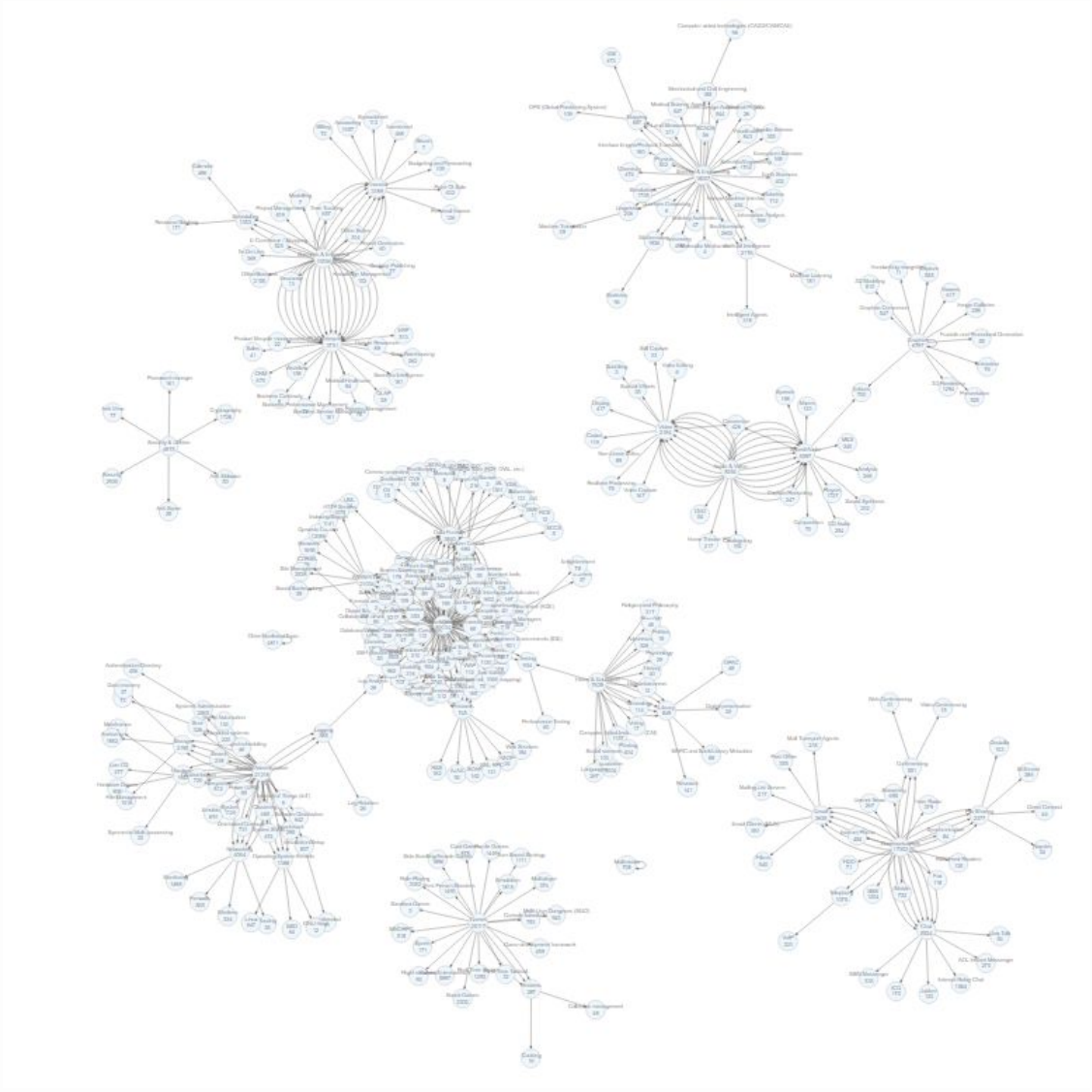


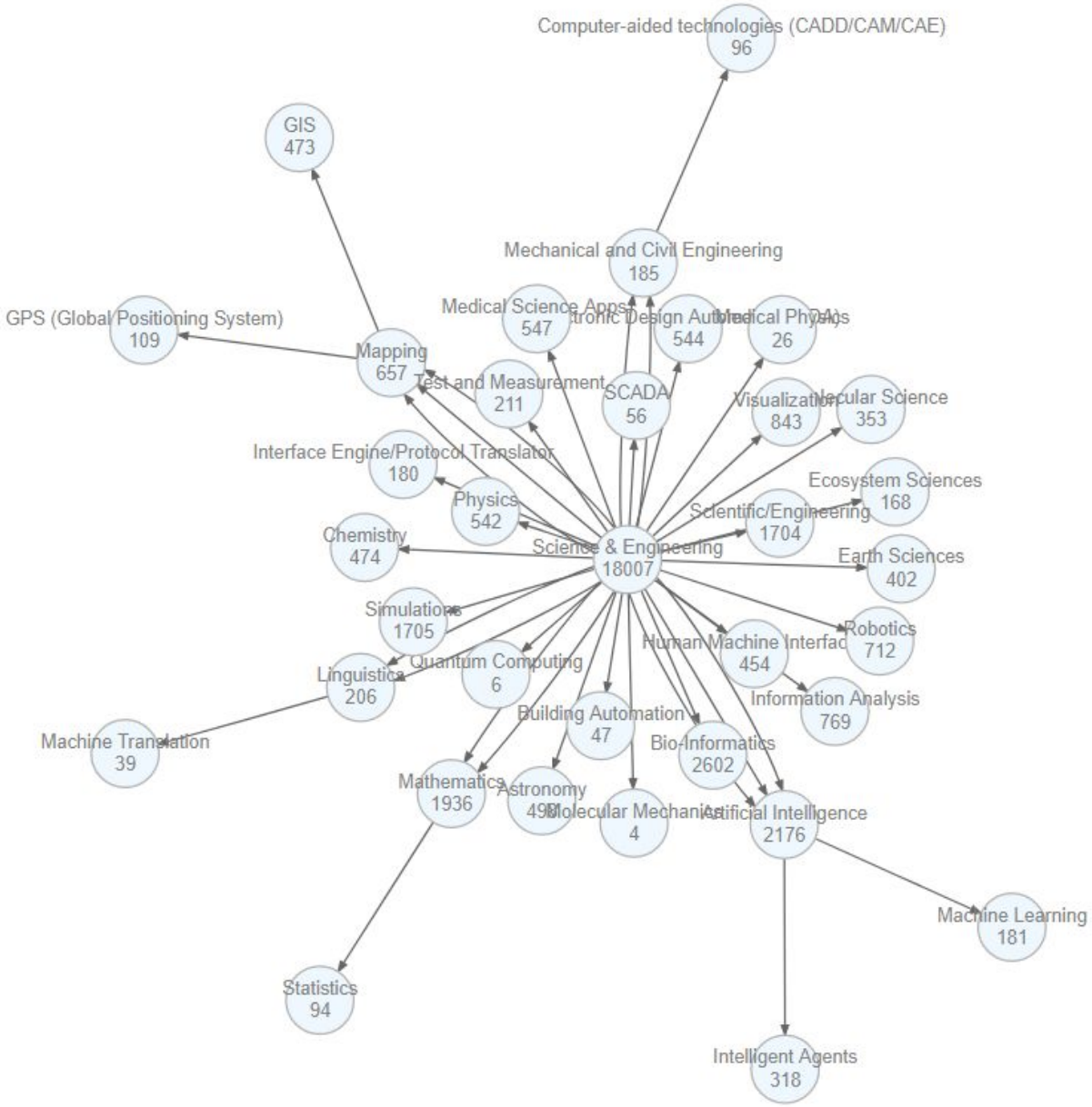
Sourceforge



	Category.1	Category.2	freq
1	Audio & Video	Sound/Audio	5597
2	Audio & Video	Video	2194
3	Business & Enterprise	Enterprise	3701
4	Business & Enterprise	Financial	3189
5	Communications	Chat	5524
6	Communications	Email	3635
7	Development	WWW/HTTP	21034
8	Development	Software Development	6336
9	Games	Games/Entertainment	5897
10	Games	Role-Playing	3352
11	Graphics	NA	1349
12	Graphics	3D Rendering	1294
13	Home & Education	Education	3004
14	Home & Education	Computer Aided Instruction (CAI)	1157
15	Multimedia	NA	708
16	Other/Nonlisted Topic	NA	4811
17	Science & Engineering	Bio-Informatics	2602
18	Science & Engineering	Artificial Intelligence	2176
19	Security & Utilities	Security	2930
20	Security & Utilities	Cryptography	1728
21	System Administration	Networking	4064

Sourceforge



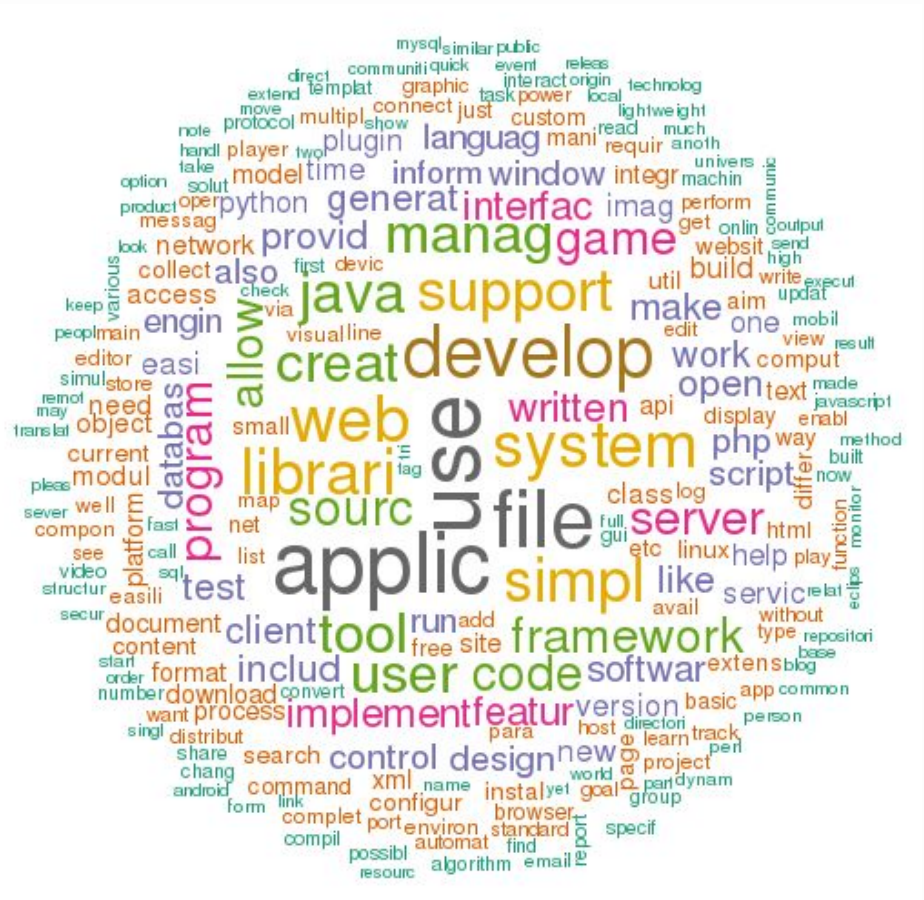


Sourceforge

- **Challenges**
  - Computational complexity:
    - 15+ days to scrape
    - Adding social media information would have drastically increased time
  - Data Complexity
    - Some of the names are in foreign languages
    - Not too many quantitative variables to analyze
- **Next steps**
  - Analyze impact of various factors on Total Downloads
  - Learn more about how these softwares are clustered (machine learning)

Sourceforge vs. OpenHub

- Terms in common and differences:
  - Stemmed words created from the Descriptions
- 11,000+ packages in common



freqSF

freqOH

Case Study - R Packages (source: depsy.org)

A project through ImpactStory, started with NSF funding, to give credit to authors of Open Source projects.

Data Collection

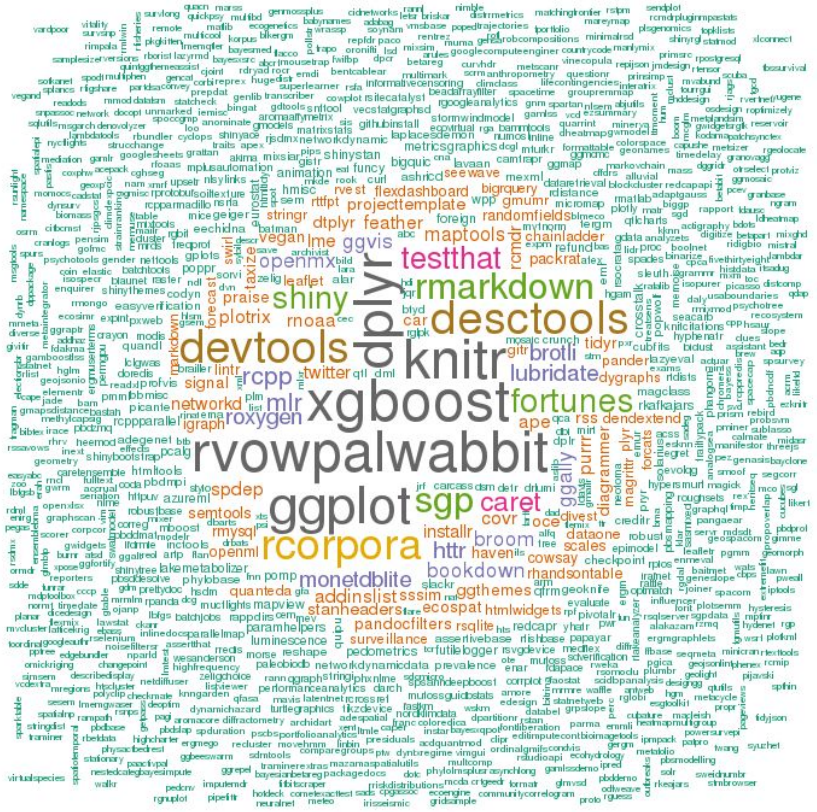
- Gathered all of the names from CRAN
  - 10,926 packages listed
- Scraped the characteristics from the JSON page affiliated with each package
  - 9,810 of these packages were on Depsy
  - 90% of the packages
  - Potentially uploaded after September, 2015 (Depsy's last update)

Case Study - R Packages (source: depsy.org)

- **Six tables through the Depsy API:**
  - Package Matrix:
    - 9,810 entries
    - 24 variables such as git owner, number of committers, number of downloads
  - Contributor Matrix:
    - 24,000 entries
    - 17 variables such as main language, roles, number of citations
  - Tag Lists (such as bayesian, finance, experimental design)
  - Top Neighbors
  - Top Contributors
  - Dependency Matrix

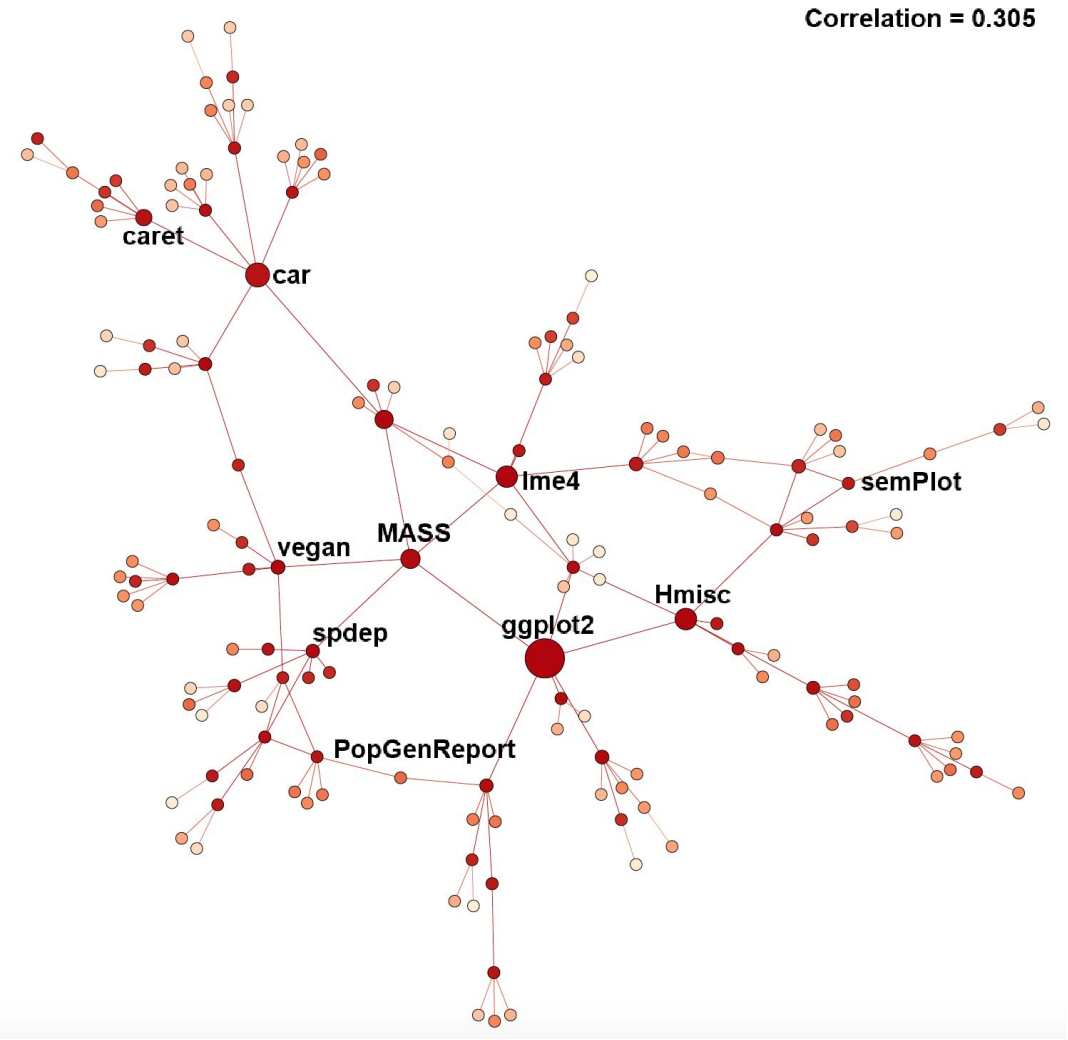
Case Study - R Packages (source: depsy.org)

Packages with the most contributors:


rvowpalwabbit
xgboost
knitr
devtools
dplyr
shiny
rmarkdown
desctools
testthat
ggplot
rcorpora
caret
sgp
fortunes
mlr
roxygen
networkd
lubridate
plotrix
rnoaa
praise
openxlsx
lme
ggvis
maptools
feather
dtplyr
randomfields
chainladder
packrat
leaflet
forecast
twitter
rcpp
signal
spdep
semtools
rmysql
openml
monetdblite
bookdown
handsontable
addinslist
stanheaders
ggthemes
ecospat
pandocfilters
rsqlite
redcapr
broom
installr
haven
cowsay
checkpoint
dataone
scales
epimodel
rmarkdown
shinystan
projecttemplate
flexdashboard
bigquery
seewave
ggmap
ggally
ggraphical
magrittr
plyr
purrr
tidyr
brotli
dygraphs
rdists
magclass
rfalkajars
ape
rss
dendextend
rcpparallel
igraph
pcalg
adegenet
htmltools
shinybootstrap
azureml
semtools
rmysql
mboost
dbarts
openml
lakemetabolizer
phylobase
quanteda
mapview
mucflights
luminescence
paleobiodb
networkdynamicdata
prevalence
surveillance
assertivebase
risthbase
papayar
pedometrics
futile.logger
rsvgdevice
medflex
sdverification
rsomoclu
plumbr
geojsonlint
phenex
rcmip
gedlight
pijavski
sphin
geoshinkleepboost
comptol
glaostat
mmrmr
waffle
anteeb
hgm
metacycle
esgtoolkit
propr
tidyjson
magnitudes
mutossguidebstats
amore
statnetweb
rglob
databel
grpdope
edesign
rstan
parma
eranb
beatmapmultigroup
powersurvepi
patpro
twang
syuzhet
metaklio
pbmodelling
solr
sweidumbr
shinbrowser
rkesjars
odbdemo
odbweave
outreals
rguess
proto
rgnuplot
pipefitr
neuralnet
meteo
irisseismic
gridsample
community
correlogram
hotdeck
comeleacitest
sads
ggpscale
ecomgine
phylolmsplunr
asynchlong
gamlssdemo
riskdistributions
mcda
orgtreed
kernsmooth
grmsd
odbweave
ggpepel
bayesianbetareg
packagedocs
bibliscraper
mazamaspatialutils
multicomp
traminerestras
compareGroups
ptw
dynbenchmark
virtualspecies
waffle
nestedcategbayesimpute
pedcmv
imputemdr
bigmeb

Next Step: Network of R Packages

- Conduct exploratory network analysis
  - Generated dependency network of R Packages
    - 7,547 nodes
    - 20,641 directed edges
- Develop statistical models to characterize factors that affect 'impact' of an R package
  - node characteristics
  - network measures



Conclusion

Next Steps:

- Complete data collection, where appropriate
- Begin model building
- Ongoing questions:
  - How do we define impact/cost/benefit?
  - How do we define the universe?
  - How do we keep up with the speed of OSS innovation?

Questions

Thank you.

Comments?

Suggestions?

Measuring Business Innovation

PI & Co-PI: Gizem Korkmaz & Stephanie Shipp
Team: Claire Kelling(Penn State), Harsimrat Pandher(VT), Daniel Wilkin(VT), David Park(VT), Joseph Kim(VT)
Sponsor: National Science Foundation (NSF) – National Center for Science and Engineering Statistics (NCSES)

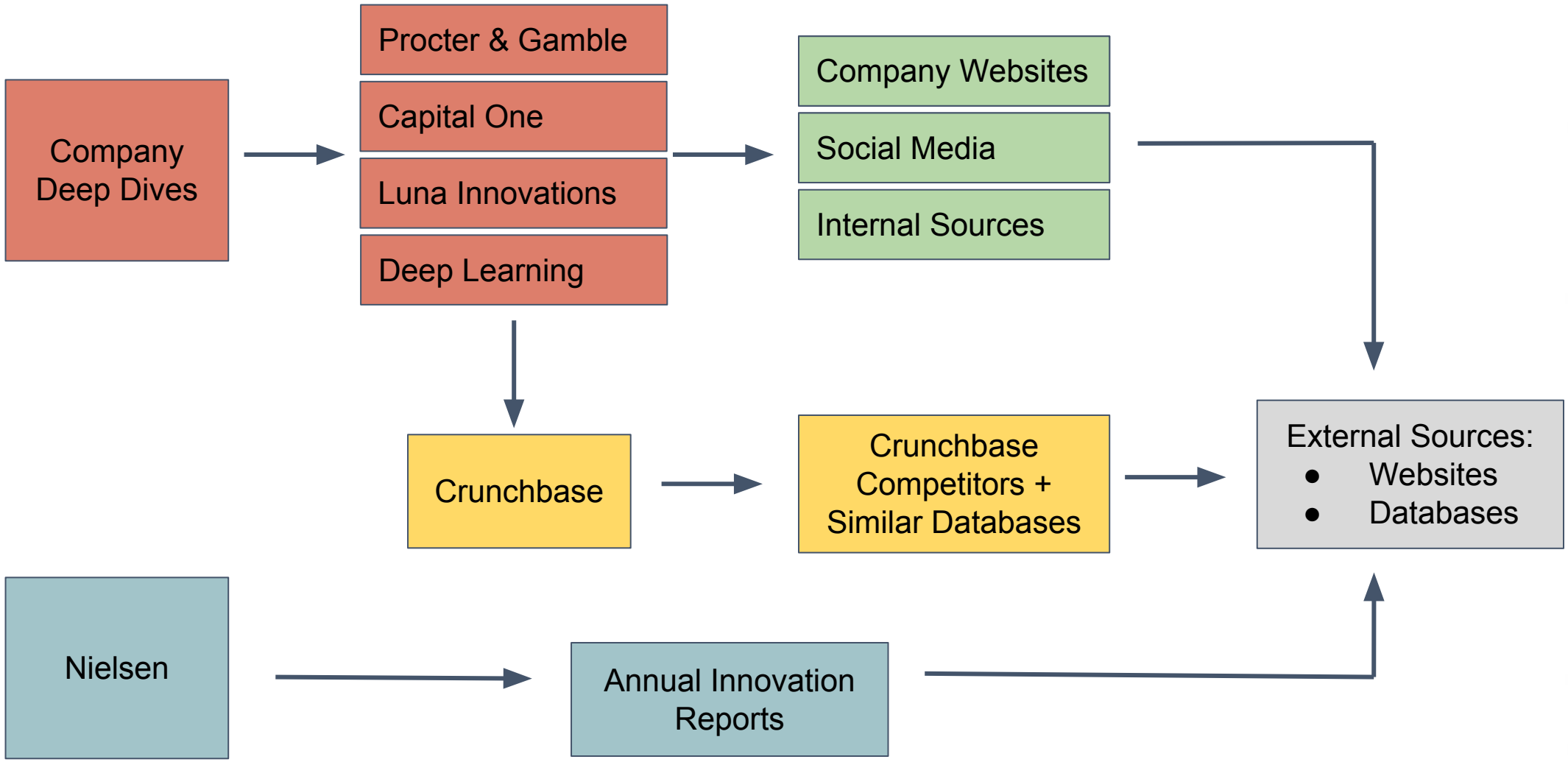
VirginiaTech Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Introduction

- Innovation - “implementation of a new or significantly improved product, process, marketing method, workplace organization, or new organizational method.” (Oslo Manual)
- Current NSF Method - Surveys
  - Business R&D and Innovation (BRDIS) Survey
    - Shortcomings
    - Separating R&D from Innovation Survey
- Research Question - Can we measure business innovation using non-traditional sources (administrative and opportunity data (e.g., websites, product announcements, social media)?
- Goals:
  - Focus on product innovation
  - Identify accessible, external data sources for product launches
  - Company deep dives

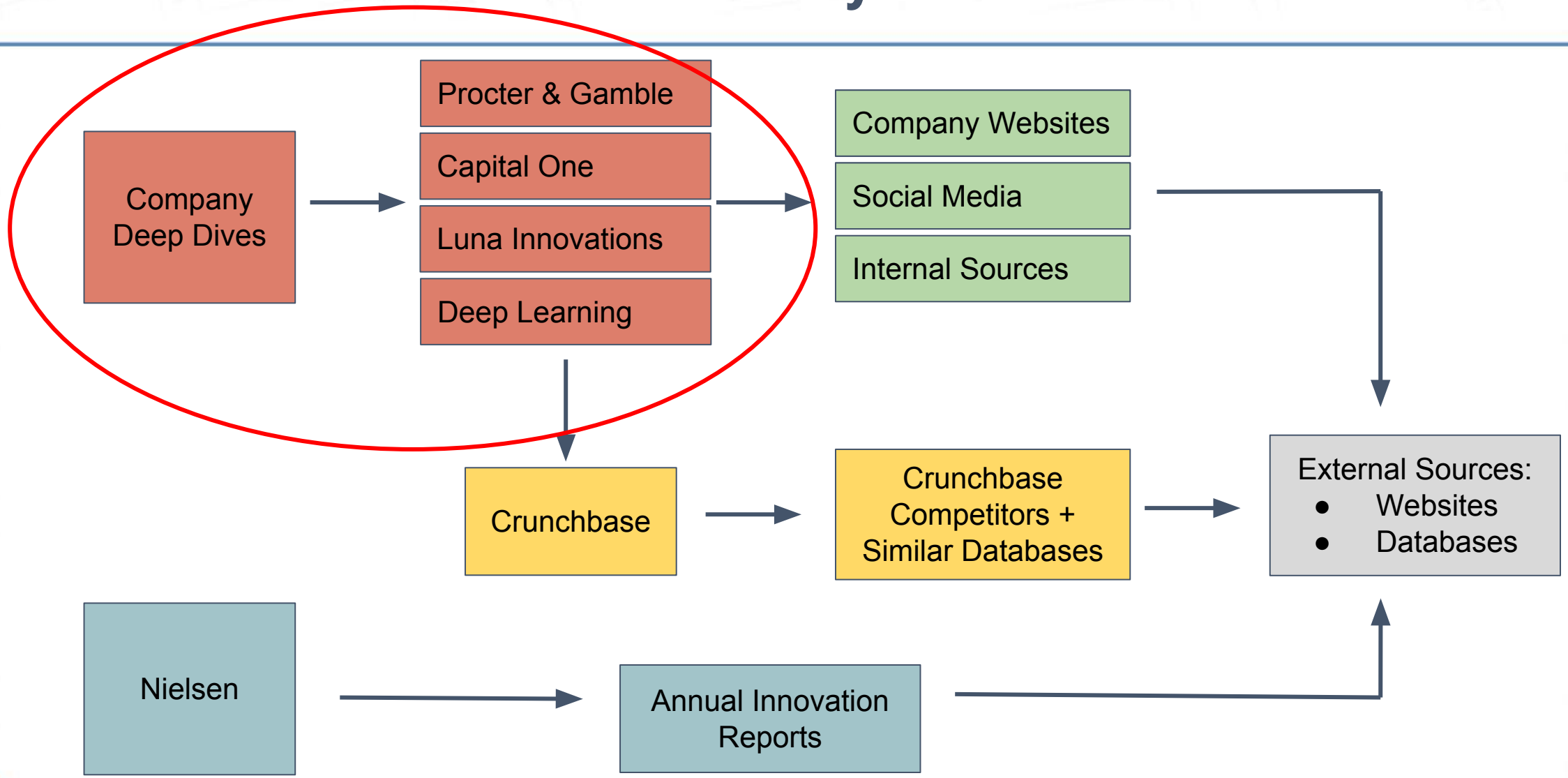
Data Discovery Process



Data Discovery Process

- Identify criteria to classify discovered sites/databases as potentially useful:
  - Does it mention a launch?
  - Does it have a specific launch date?
  - Was the article released at the time of the launch?
  - Can the previous criteria be applied to other products?
  - How accessible is it?
  - Is it scalable?

Data Discovery Process



Company Deep Dive

- Purpose
  - Identify potential data sources to measure product innovation
  - Interview to validate findings (next step)
- Initial set of companies:
  - Procter & Gamble
  - Capital One
  - Luna Innovations
  - Deep Machine Learning
- Selection Criteria
  - Size
  - Industry
  - Connections

Procter & Gamble

General Info:

- Consumer Goods Industry
- 105,000 employees across 6 continents

Innovation:

- "Innovation starts with the customer"
- Open Innovation: Connect + Develop
- Product & Marketing Innovation
  - Gillette Flexball, Tide Pods

Data Sources:

- P&G Website: "Our Approach to Innovation/ Recent Innovations"
- Annual Reports
- Social media

GILLETTE

https://www.pg.com/it_IT/pdf/Relazione_Annuale_2016.pdf

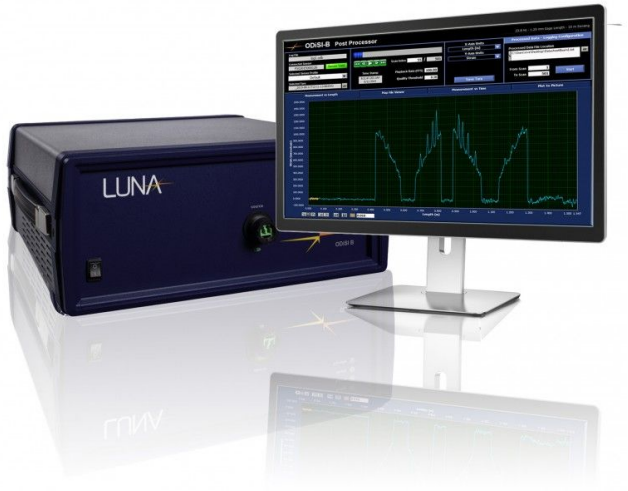
Luna Innovations

General Info:

- Founded in Blacksburg, VA
- About 185 employees
- Aerospace, Telecommunications, Medical, Energy, Defense

Innovation:

- Research + Development
- Product Development
  - Fiber Optic Sensor
- Successful Products

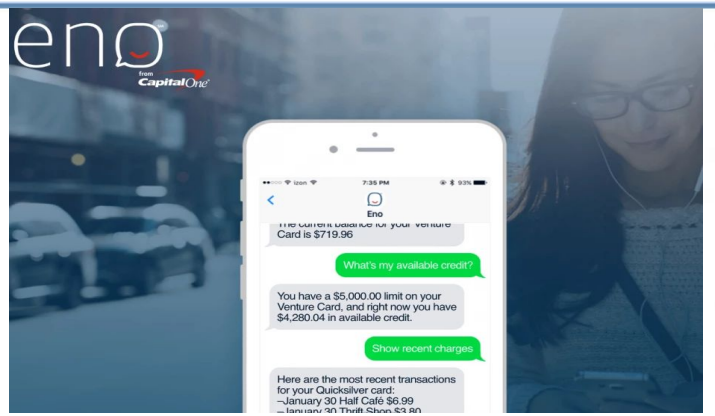


Data sources:

- Local News (Roanoke Times)
- Social Media

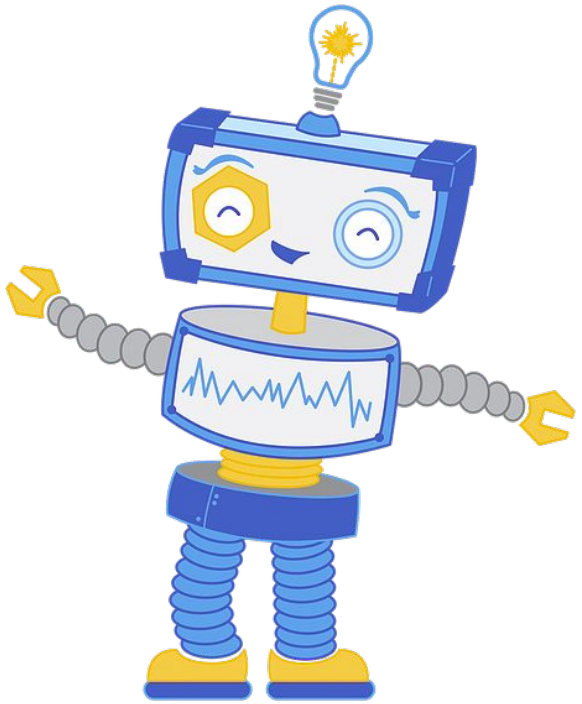
Capital One

- General Info:
  - Commercial banking industry
  - 47,300 employees
  - Headquartered in Mclean VA
- Innovation:
  - “To design new experiences that empower customers to better manage their finances”
  - ENO - a new text message - based assistant that can help customers with a variety of banking and credit card activities
  - Capital One Labs and Capital One Digital - Cross functional team that helps Capital One towards the digital banking
- Data Sources:
  - Capital One Website
  - CrunchBase



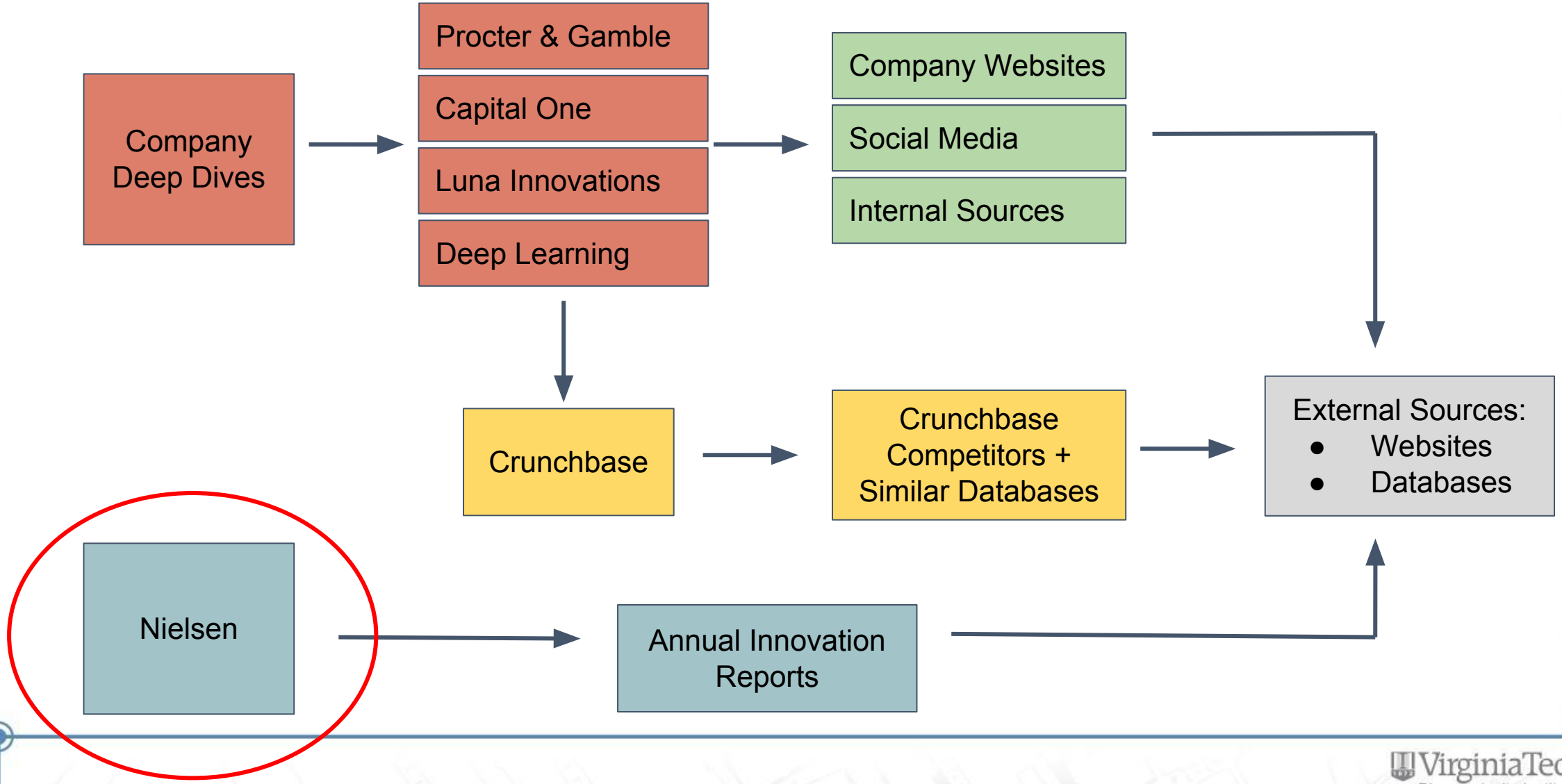
Deep Machine Learning

- General info
  - Startup
  - Founded in Arlington, 2013
  - Data Analytics Service
- Innovation
  - Focus on government service
  - Success due to a government contract
  - There is not sufficient evidence of innovation



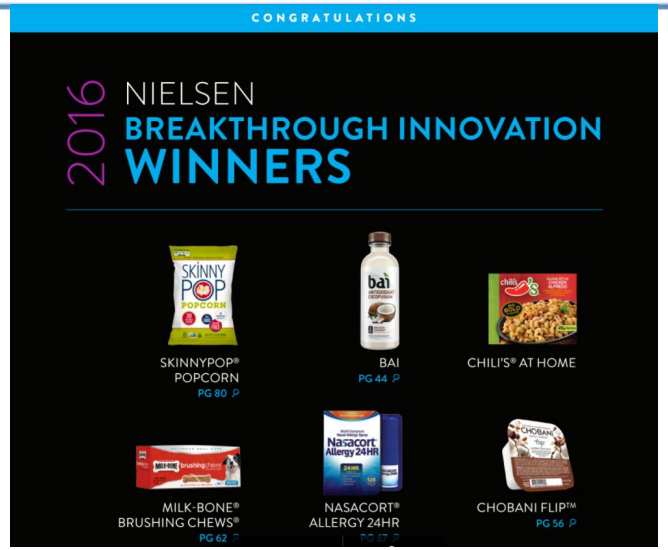
DEEP LEARNING
ANALYTICS

Data Discovery Process



Nielsen

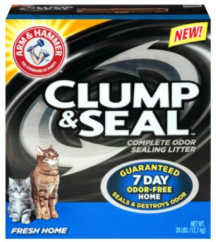
- Report of Top Innovative Products
- Contains product spotlights
- Gives certain criteria behind determining the winning products



SPOTLIGHT #1

ARM & HAMMER™
CLUMP & SEAL™

AN "IMPOSSIBLE" QUESTION UNLEASHES AN UNREAL OUTCOME



THE MAKING OF WINNERS: NIELSEN BREAKTHROUGH INNOVATION CRITERIA

DISTINCTIVENESS

Deliver a new value proposition to the market. Ingredient reformulations, repackaging, size changes, repositioning, and other minor refinements to existing brands are excluded.

RELEVANCE

Generate a minimum of $50 million in U.S. sales during first year of national distribution.

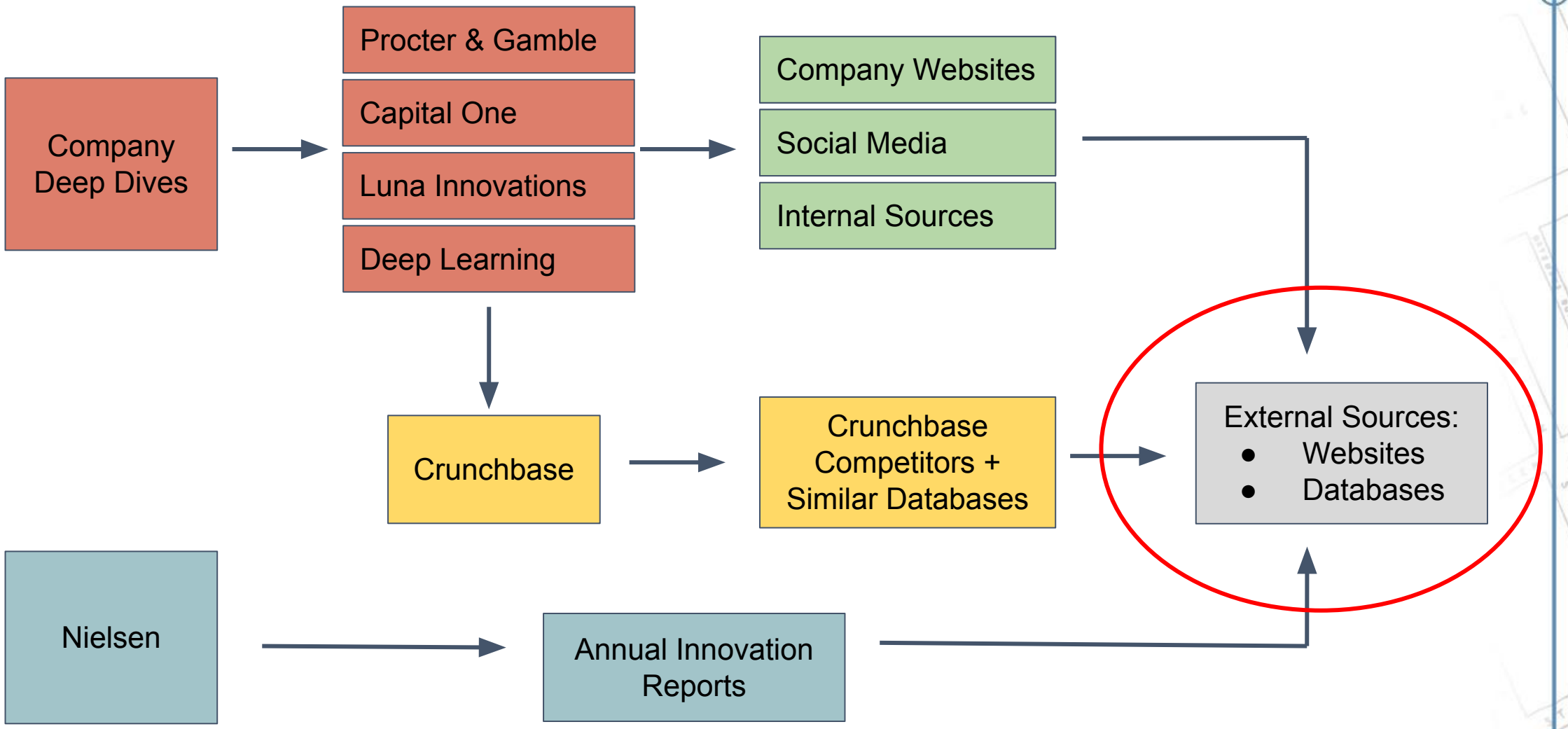
ENDURANCE

Achieve at least 90% of first year sales in year two. This measure confirms a sustained level of consumer demand while allowing for some drop in revenue during the transition from trial to adoption.

Data Discovery

- Documented Information about the company and product
  - company name, sector/industry, product type, product release date, etc.
- Categorized sources based on products in the Nielsen Report
  - social media, news, blogs, reports, etc.
- Evaluated those sources
  - quality
  - reasoning

Data Discovery Process



Websites

Websites	Information Provided
Ars Technica	New Technology Launches, News about Gadgets, Gaming, Science, Cars, Business
The Verge	Product Reviews, News about Technology, Science, Culture, Cars
Digital Trends	Best Products Section, Products Reviews, News about all types of technology
Mashable	New Product Releases in Technology, Science, Business
Inc.	Startups, Innovative Ideas, Technology News
Business Wire	Press Releases, News with Multimedia, Company Information
Business Insider	Tech Insider, New Products, News
PR Newswire	New Presses Releases, Product Launches
Food Business News	Food Business News, Consumer Trends, Supplier Innovations, New Food Product Launches
Food Processing	New Food Products, Food News
Mass Device	Medical Product Releases, Medical News

Company Databases

Databases	Information Provided
CrunchBase	Company Information, News Reports, Social Media Links, Product Releases
AngelList	Company Information, News Reports, Social Media Links, Product Releases
Owler	Company Information, Company Analysis, News
VentureBeat	Technology Innovation, News Reports
F6S	Startup Companies, Conferences, Contests, Hackathons, Investment Information
Mattermark	Company Information, Company Analysis
Dow Jones VentureSource	Company Analysis, Investor Information
CB Insights	Company Analysis, Patents, Startups, Investor Information, News Sentiment, Social Media
Gust	Early-Stage Investments, Startups, Company Information

Future Work

- Continued data discovery of websites and databases
- Data collection by scraping websites and databases for product launches
- Other data sources:
  - Trade Magazines
  - Wall Street Journal Articles
  - SEC Filings
- Company interviews to
  - understand innovation strategies
  - identify data sources
  - validate findings (follow-up interviews)

STEM Education

PI & Co-PI: Bianica Pires & Stephanie Shipp
Team: Madison Arnbarger(VT), Ben Swartz(VT), Chanida Lerdritsomboon(VT), Hannah Brinkley(VT)
Sponsor: National Science Foundation (NSF) –
National Center for Science and Engineering Statistics (NCSES)

VirginiaTech
Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

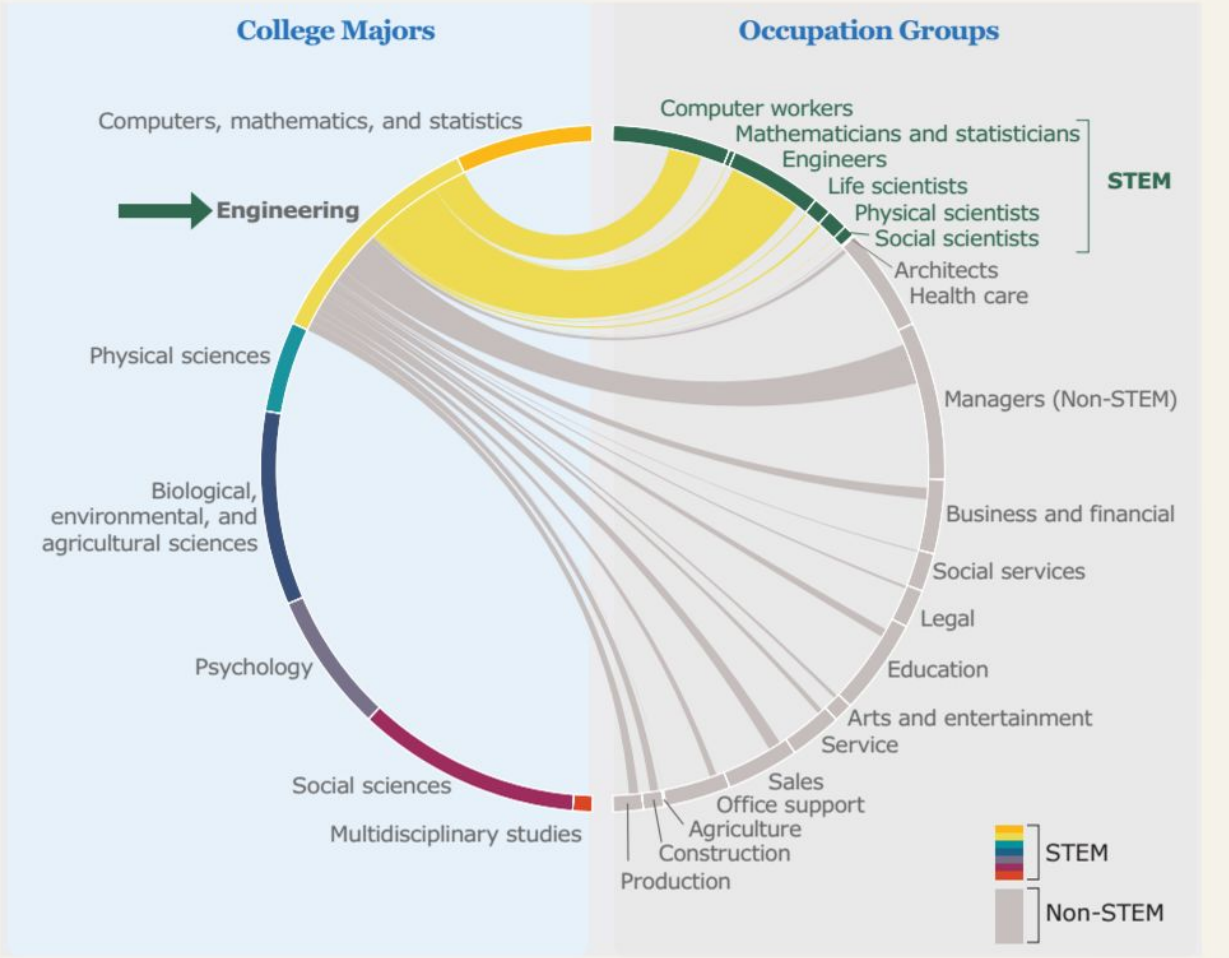
Introduction

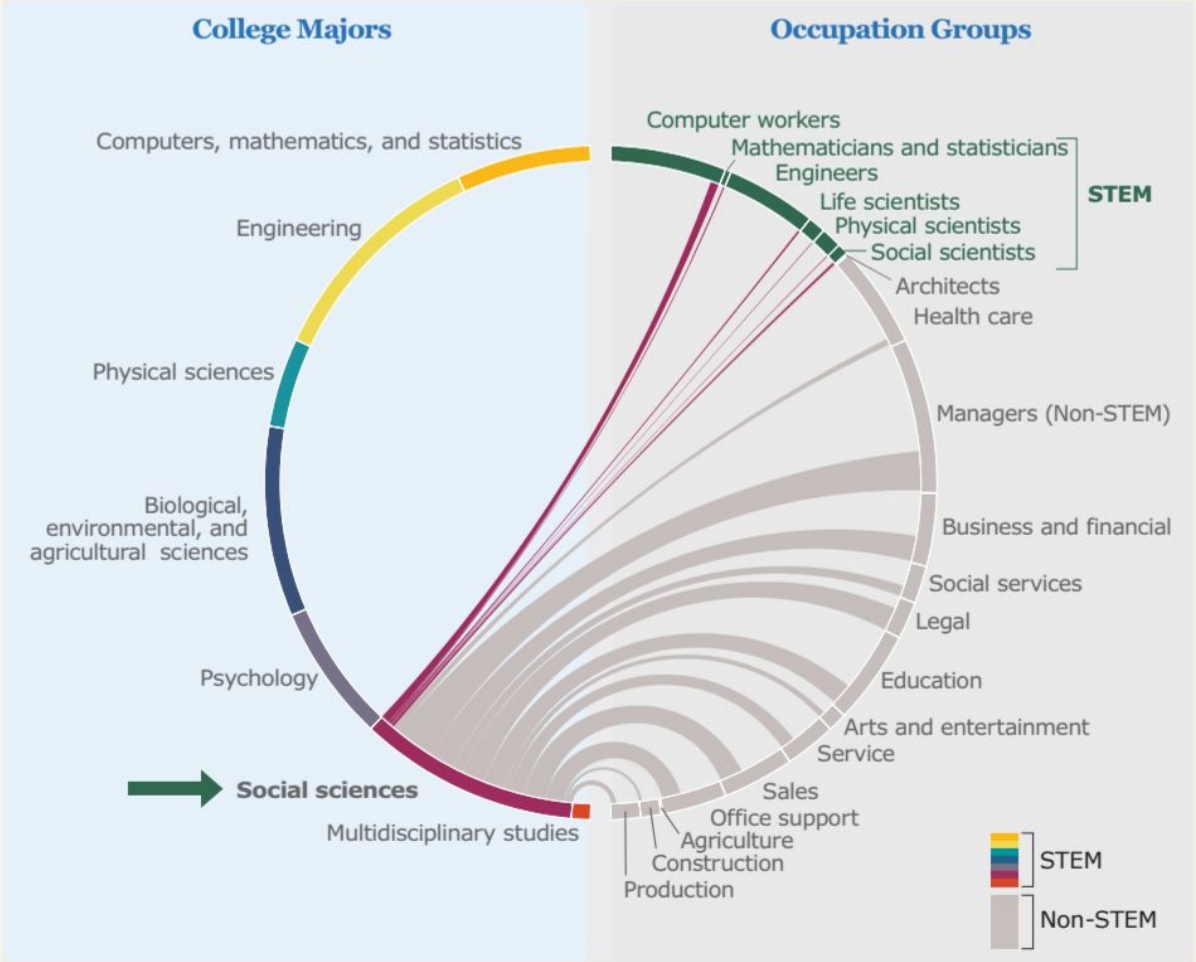
- What are the education and workforce pathways that lead to a STEM-enabled occupation? **(Employee Perspective)**
  - What are STEM pathways for sub-baccalaureate populations? e.g. high school and community college graduates, vocational/ apprenticeship training, other.
  - What are STEM pathways for women and minorities? Where along the pathway do they leave?
- What are the skills and knowledge expected by employers of these occupations? **(Employer Perspective)**

Literature Review

- Definition of STEM is inconsistent throughout the literature
  - STEM versus STEM-related
  - NSF Science & Engineering (S&E) fields; Bureau of Labor Statistics 2012, Standard Occupational Classification Policy Committee (SOCPC); Occupational Information Network (O*NET) STEM classifications
- Focus in the literature is on knowledge, skills, and abilities (KSAs) of STEM occupations versus STEM pathways
- The role of sub-baccalaureate degrees in STEM careers are not as well understood
  - However, half of STEM occupations do not require a bachelor's degree
- Women and minorities do not fare as well in terms of salaries, promotions, and persistence (more tend to shift to non-STEM occupations)
- Existing STEM pathways focus on bachelor's degree to occupation (e.g., Census)

Traditional Pathway Map





SOURCE: U.S. Census Bureau, 2012 American Community Survey.
https://www.census.gov/dataviz/visualizations/stem/stem-html/

Occupation Case Studies

- We selected three STEM occupations with a diversity of traditional and non-traditional pathways
  - Dental Hygienist
  - Environmental Engineer
  - Junior Software Developer

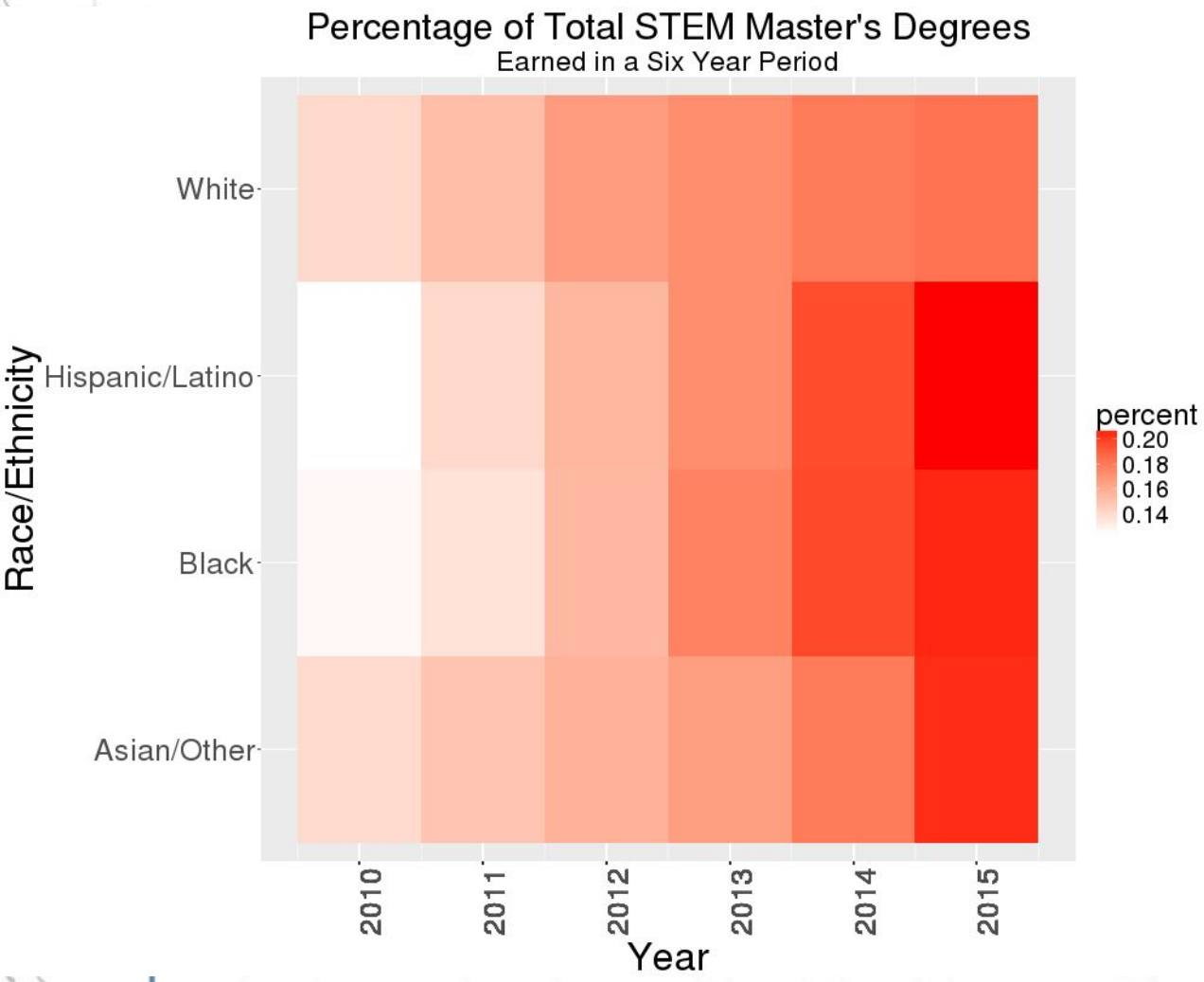
Data Discovery

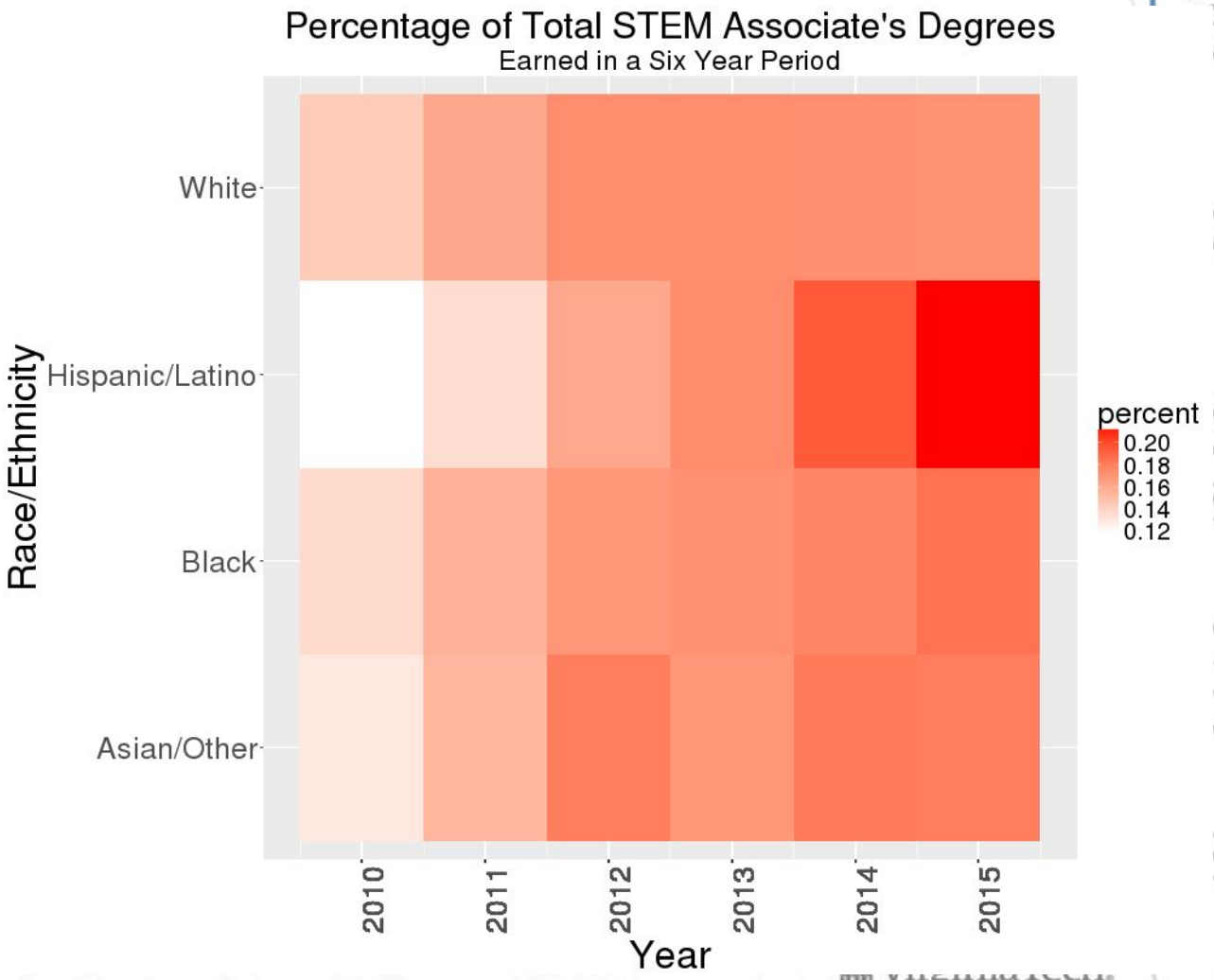
- The data discovery process involved documenting and evaluating approximately 20 different data sources
- Data sources we evaluated include
  - **Surveys**: Integrated Postsecondary Education Data System (IPEDS), Occupational Information Network (O*NET), Occupational Employment Statistics (OES), Scientists and Engineers Statistical Data System (SESTAT), NCES Education Longitudinal Study (ELS)
  - **Resume and Job Websites** for potential web-scraping: linkedin.com, indeed.com, github.io
  - Snapshots of **non-traditional STEM education**, such as online learning: edX, Coursera, other certification programs

Selected Data Sources

Data Source	Description	Unit of Analysis	Potential Use
O*NET	Survey of workers by occupation	Occupation	Gathering knowledge, skills, and abilities of a specific occupation to create a mind map.
IPEDS	Survey of every college, university, technical, and vocational institution that participates in federal student financial aid programs	Institution	Generate trends of certifications and degrees awarded by gender and race/ethnicity in STEM fields from 2010-2014.
IPUMS	Integrated Census microdata and ACS survey data.	Individual	Analyze pathways between education level and occupation type by sex, race, and/or ethnicity.
NCES ELS	Survey of students who were high school sophomores in 2002.	Individual	Describe student transitions between high school, college, and careers.
Indeed.com	Job posting and resume website	Individual	Identify education and skill pathways; Identity employer expected skill and knowledge.
Coursera	Website offering online training courses	Training courses	Identify skills valued by employers for each occupation that can be acquired through taking online courses.
Survey of Allied Dental Association	Surveys that include descriptions on academic programs, admission requirements, and enrollment.	Individual	Provides data on total enrollment by age, gender, and ethnicity and compares different educational settings.

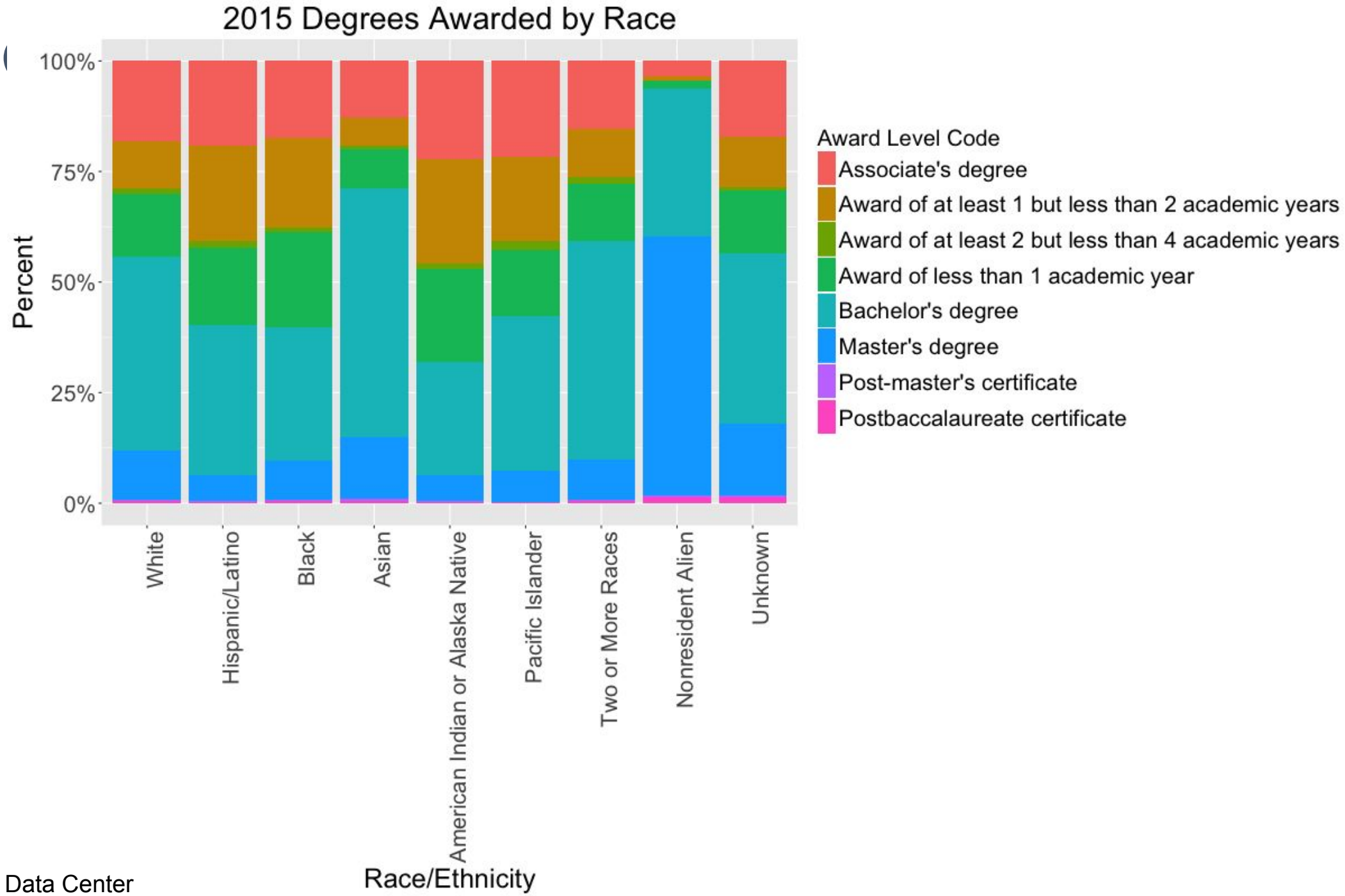
Trends in STEM Degrees by Race





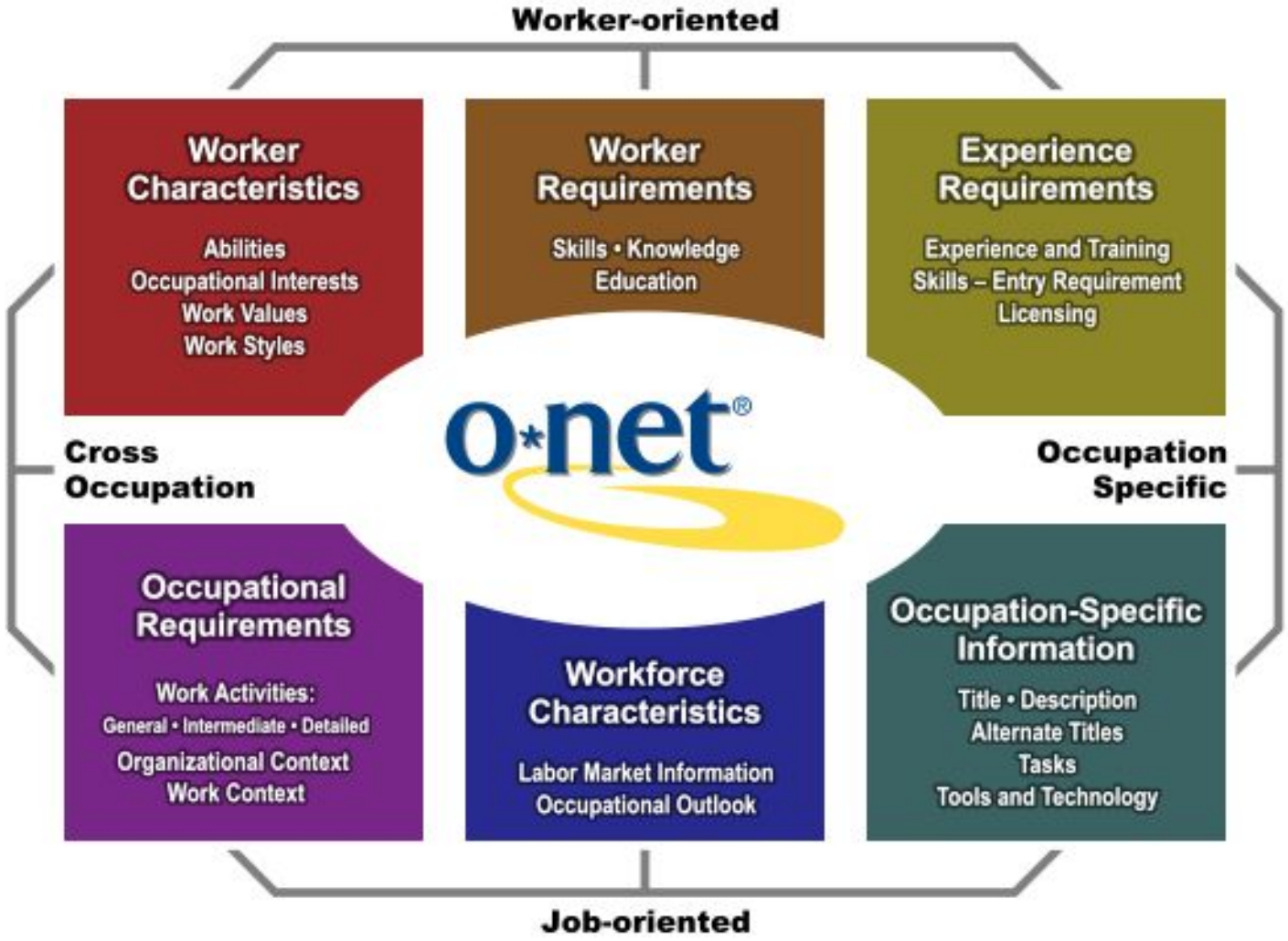
NCES IPEDS Data Center

Biocomplexity Institute

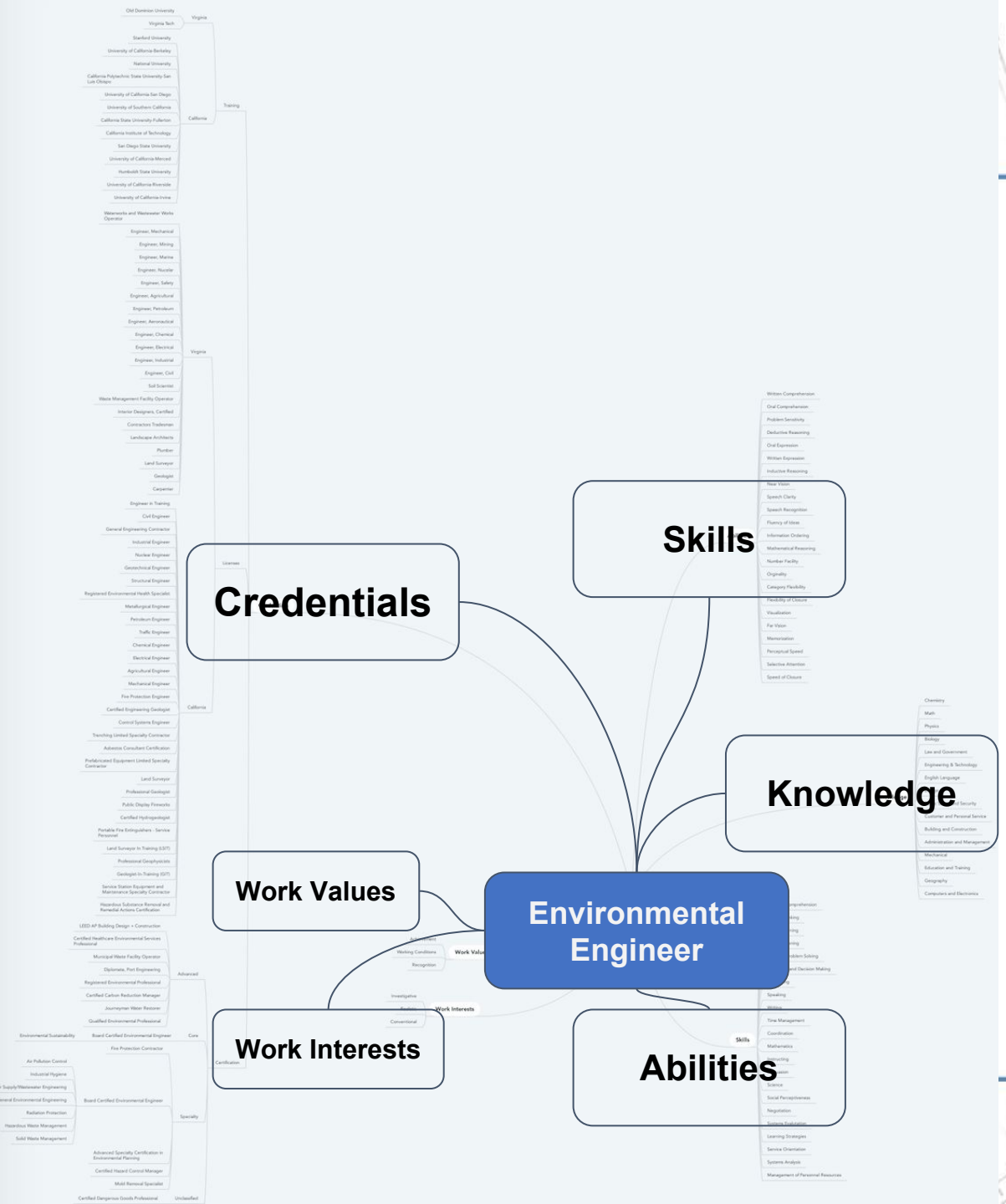


NCES IPEDS Data Center

O*NET Content Model



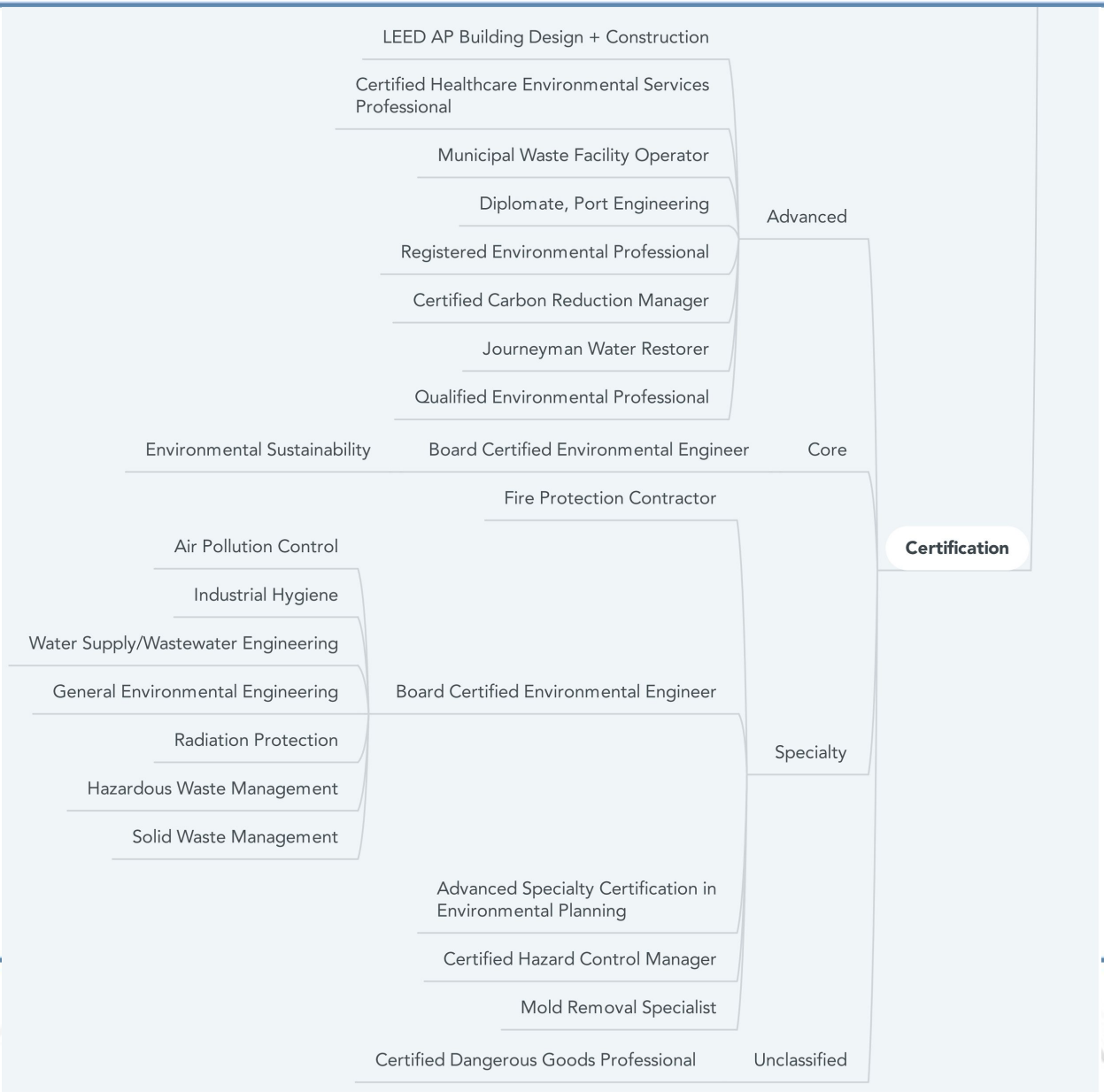
https://www.onetcenter.org/content.html

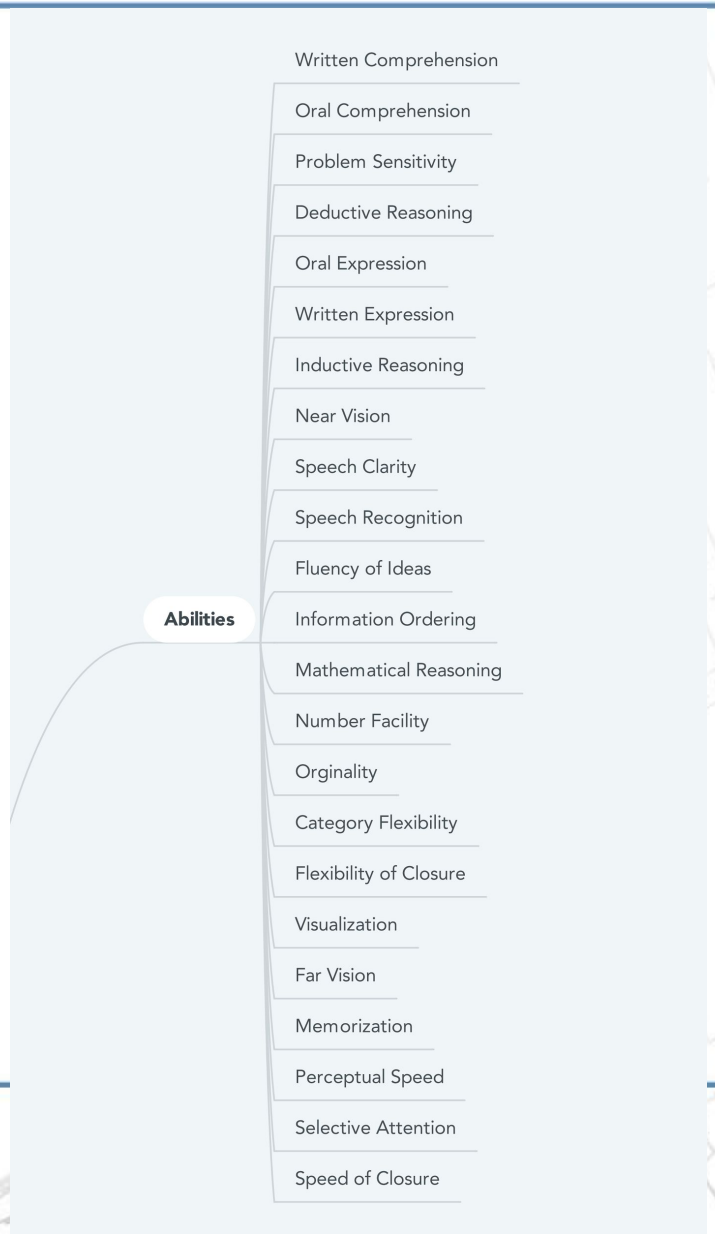


Mind Map for Environmental Engineer

VirginiaTech.
Biocomplexity Institute

Mind Map - Snapshot

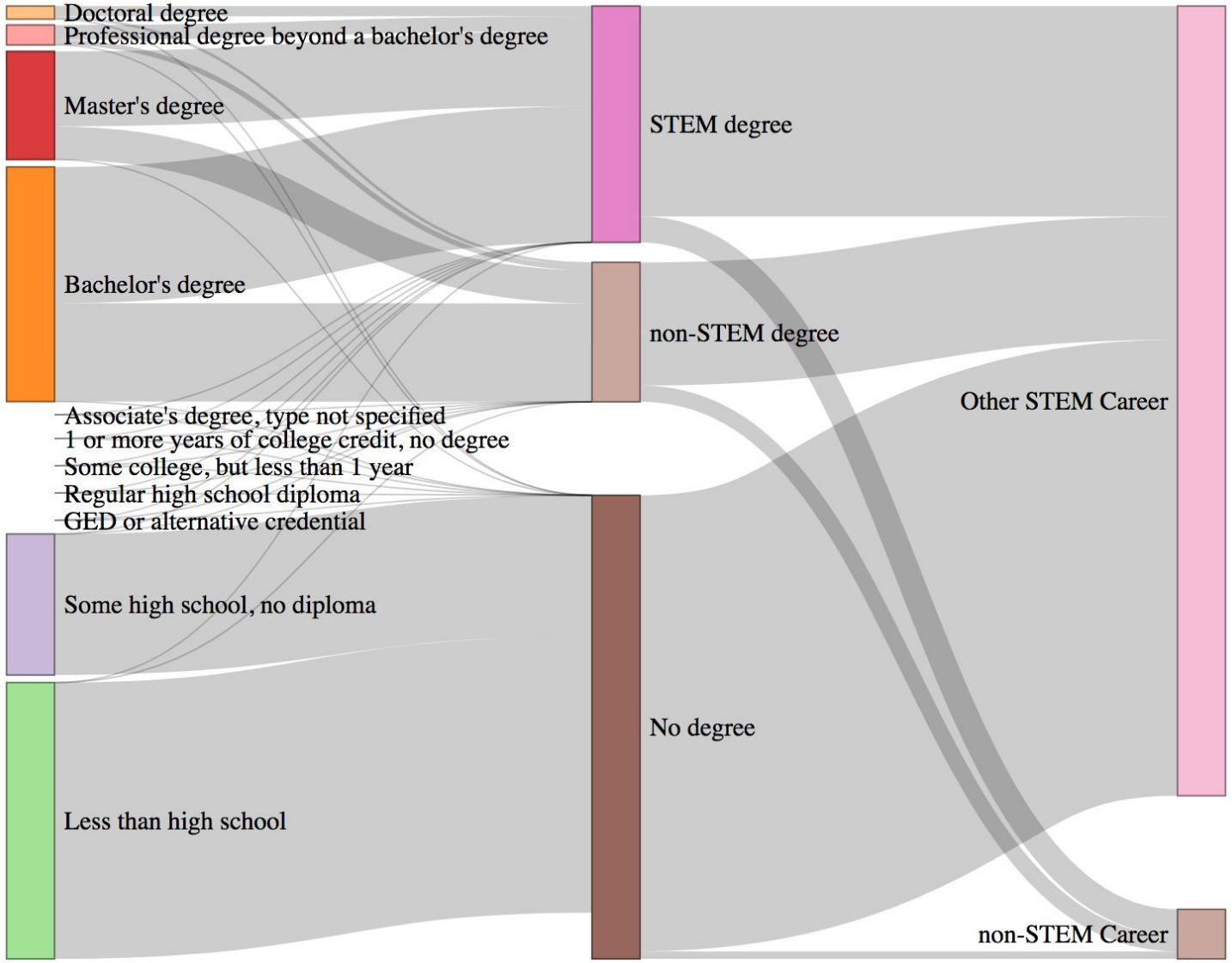




Employee Perspective: Generating Pathways from Data

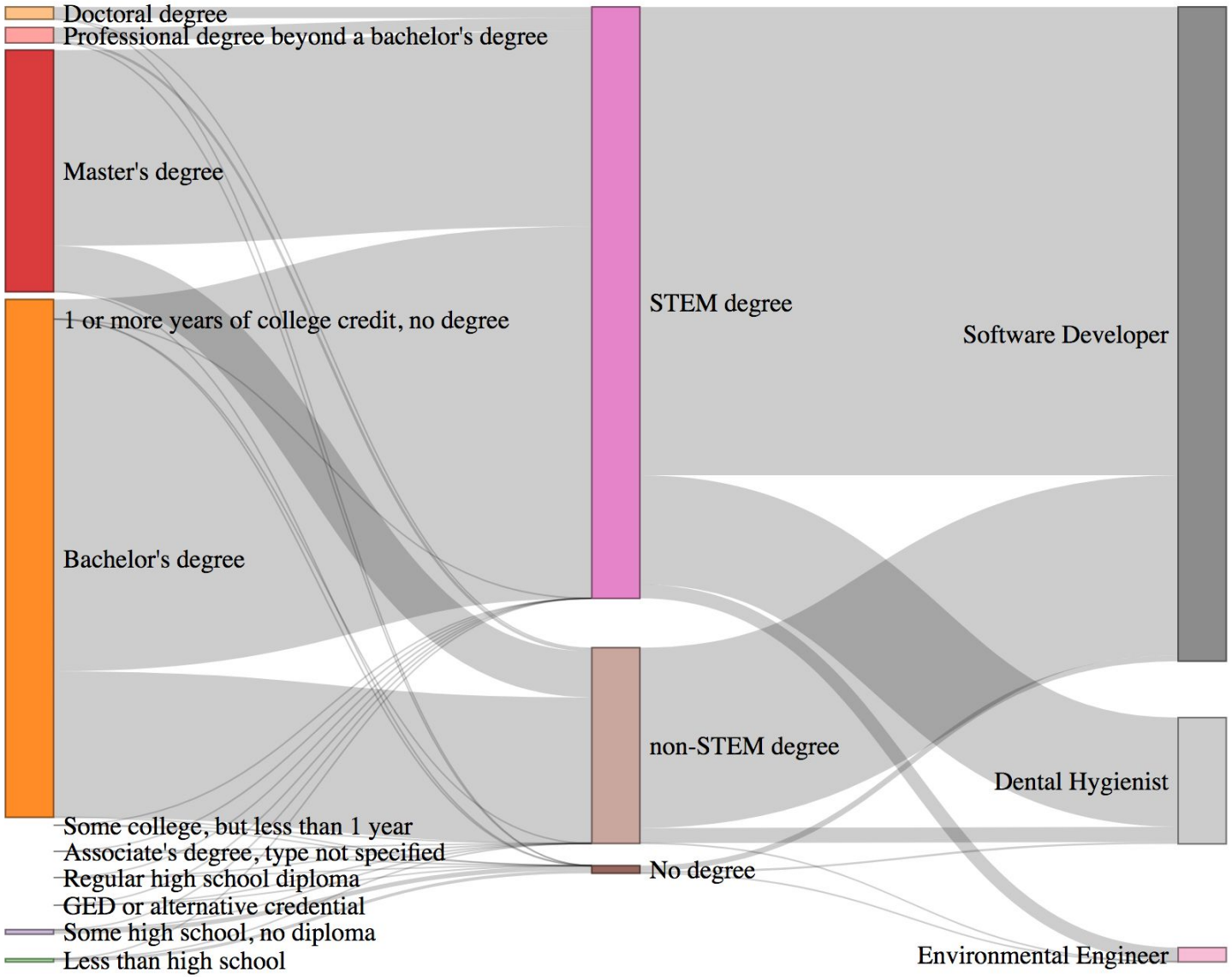
Integrated Public Use Microdata Service (IPUMS) is a publicly available source of harmonized Census, ACS, and CPS data.

At the individual level, this data is conducive to pathway mapping by education level, occupation type, and minority identity.



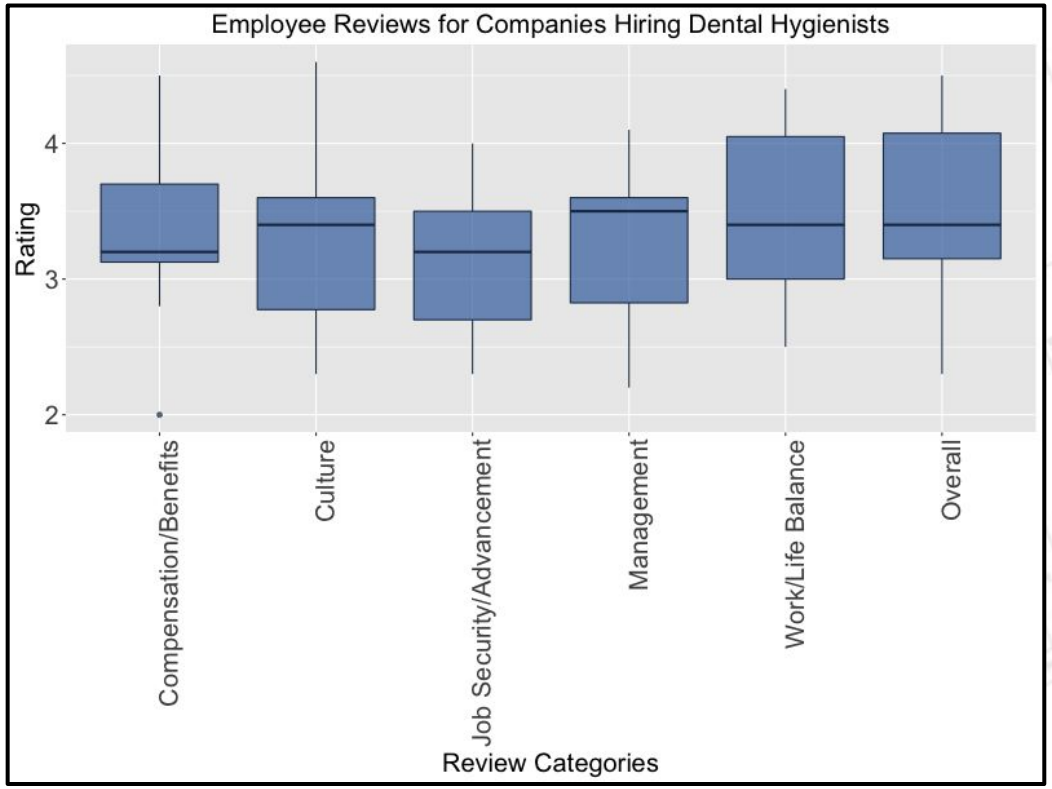
Employee Perspective: Generating Pathways from Data

- Next steps:
  - Focus pathway analysis on three selected occupations.
  - Identify which non-STEM fields attract STEM degree-holders (attrition).
  - Understand demographic patterns of STEM leavers and persisters.



Employee Perspective: Understanding the Workforce

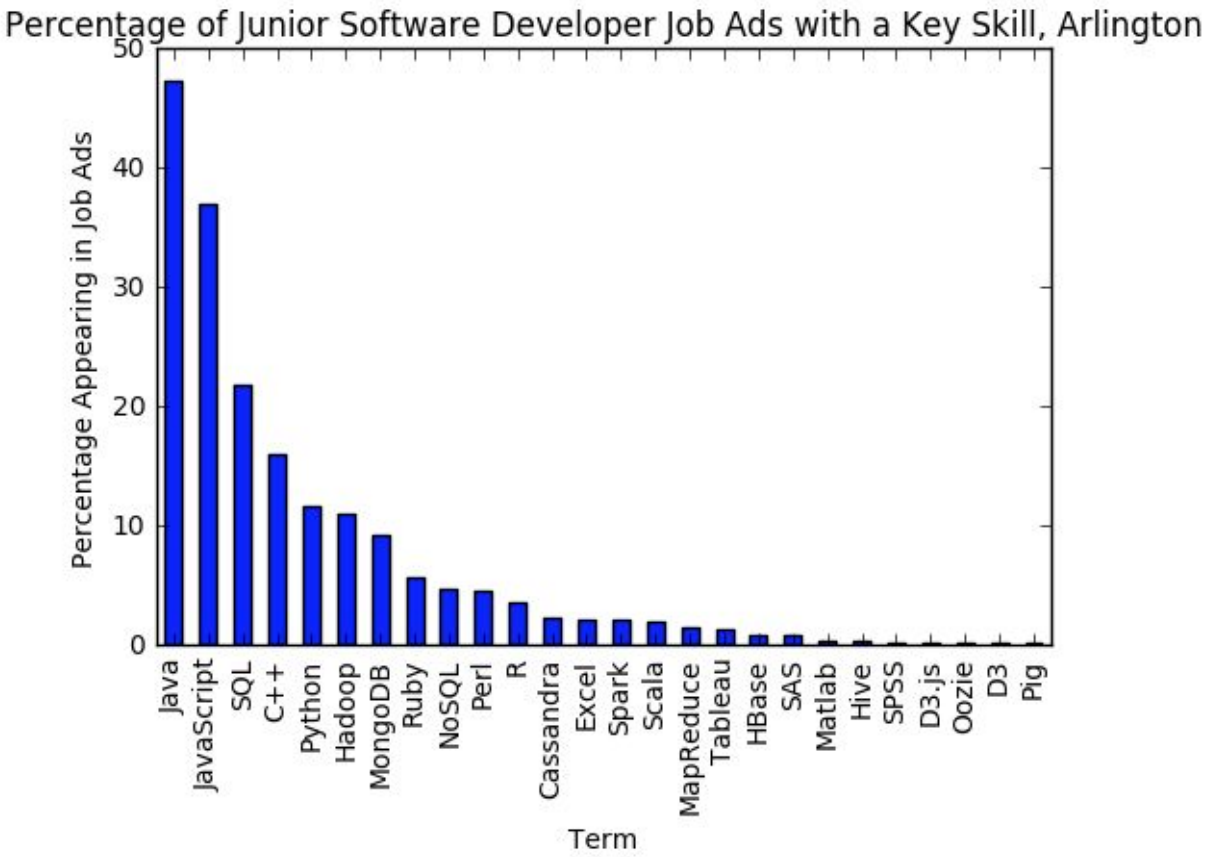
- **Indeed** is a job site that allows employers to post jobs and job seekers to post their resume, search for jobs, and provide company reviews
- **Coursera** is an online training site developed in 2012 by two Stanford Computer Science professors who wanted to share their knowledge and skills



Title	Description	Skills
Python for Everybody Specialization	This Specialization will introduce fundamental programming concepts including data structures, networked application program interfaces, and databases, using the Python programming language….	This course aims to teach everyone the basics of programming computers using Python. We cover the basics ….

Employer Perspective: Expected Knowledge and Skills

- Key skills required/desired from companies currently hiring junior software developers (or similar)
- We can compare these skills to those offered by Coursera and other institutions



Steps Ahead

- Use IPUMS to better understand pathways of STEM persisters and leavers
  - Pathways to STEM from non-STEM backgrounds
  - Patterns along race and gender lines
- Use IPEDS to finalize general trends in the STEM degrees awarded
  - include PhDs
  - subset by gender
- Continue analysis of web-scraped content (Indeed, oursera, etc.)
  - Map the knowledge/skills obtained by the workforce of our three occupations to the knowledge/skills desired by employers
- Identify and document additional longitudinal surveys that may have data to support our STEM pathways or reveal new pathways (e.g., ECLS-K, B&B:08)

Community Projects: Alexandria - Equity

Team: Sayali Phadke, Daniel Wilkin, Ronnie Fresco
PI & Co-PI: Emily Molfino, Mark Orr
Sponsor: Alexandria City Government, Office of Performance and Accountability

VirginiaTech
Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

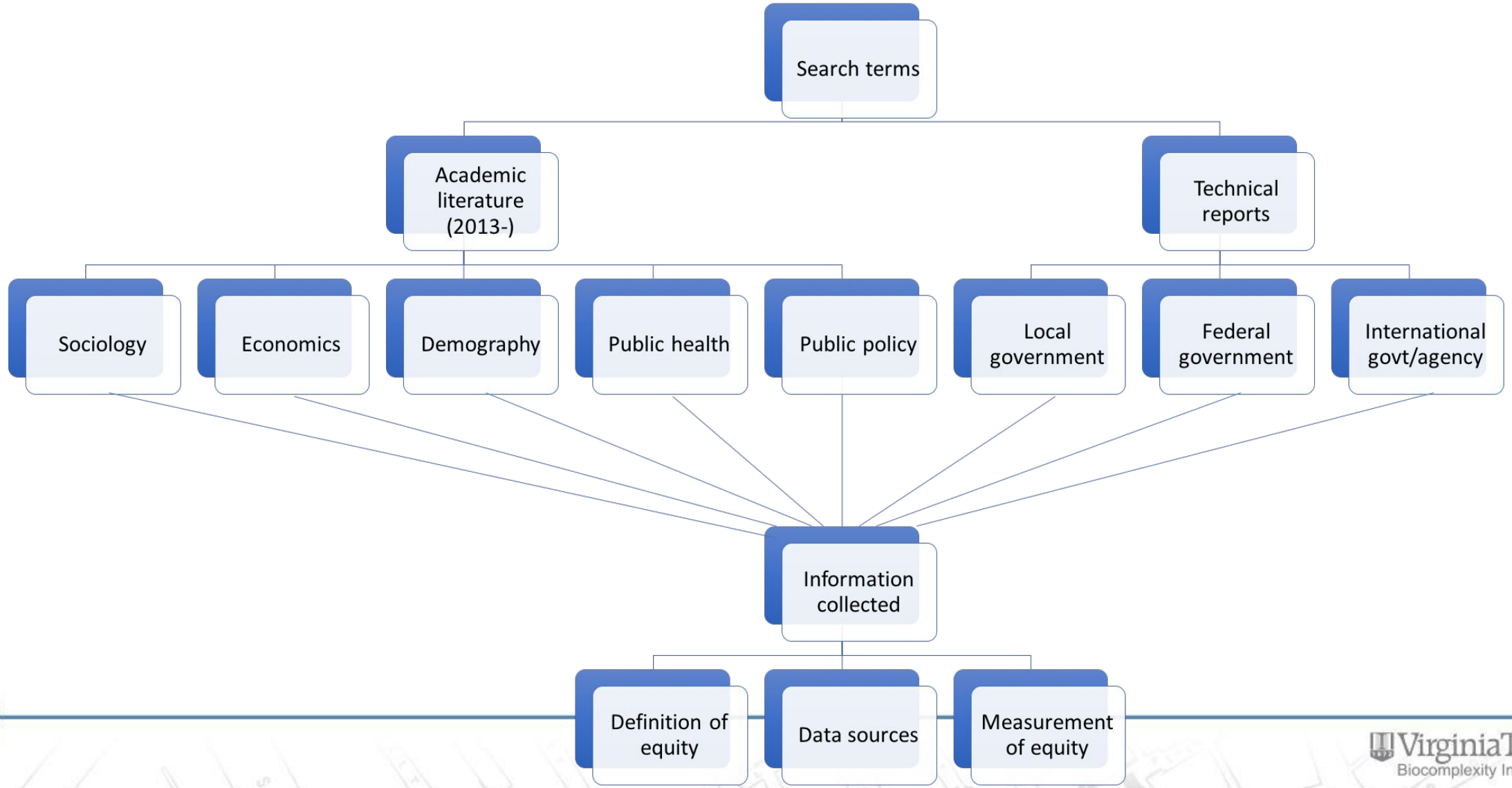
Introduction

- Equity is important for the well-being of citizens, and hence an important policy consideration
- City of Alexandria striving towards combating issues of inequity in the city
- ‘Equity in Alexandria’ team would like to understand:
  - How is equity defined?
  - How can equity be measured?

Objectives of the project

- Literature review on equity
  - Conceptualizing equity
  - How is it defined?
  - How is it measured?
- Discuss the link between definitions of equity and data sources w.r.t. the City of Alexandria
- Provide recommendation to the city government about data sources and a data portal

Methodology: literature review



Methodology: lit review (contd.)

Lit review: Output

- Equity can be understood in its many forms:
  - Equity of access
  - Equity of distribution
  - Equity as achieved by inequality of treatment
  - Process equity
- Equity can be measured in different ways:
  - Simple comparison of variables-of-interest across groups (education by race, driving privileges by age etc.)
  - Composite indices (e.g. Gini coefficient, Theil statistic etc.)
  - Other methods such as Principle of Proportional Shortfall, Expected Inequality-Averse model

Lit review: Output (contd.)

- Equity is studied with respect to one aspect or a combination of various aspects:
  - Who you are: Age, race, gender etc.
  - What is provided: Transportation services, housing facilities etc.
  - What is available: Health, education etc.
  - What you can achieve: Employment etc.
- Several reports focusing on city or county level equity:
  - King’s county, WA
  - Multnomah county, OR
  - Dane county, WI
  - Boston, MA
  - San Francisco, CA

Lit review: Output (contd.)

- Several research groups aim to understand and measure equity, presented via a web-based interface:
  - http://nationalequityatlas.org
  - http://thrivingcities.com

- Data sources
  - Administrative data from local, state, and federal level
  - Survey conducted by local government
  - Policy and planning documents

Equity in Health & Generalization

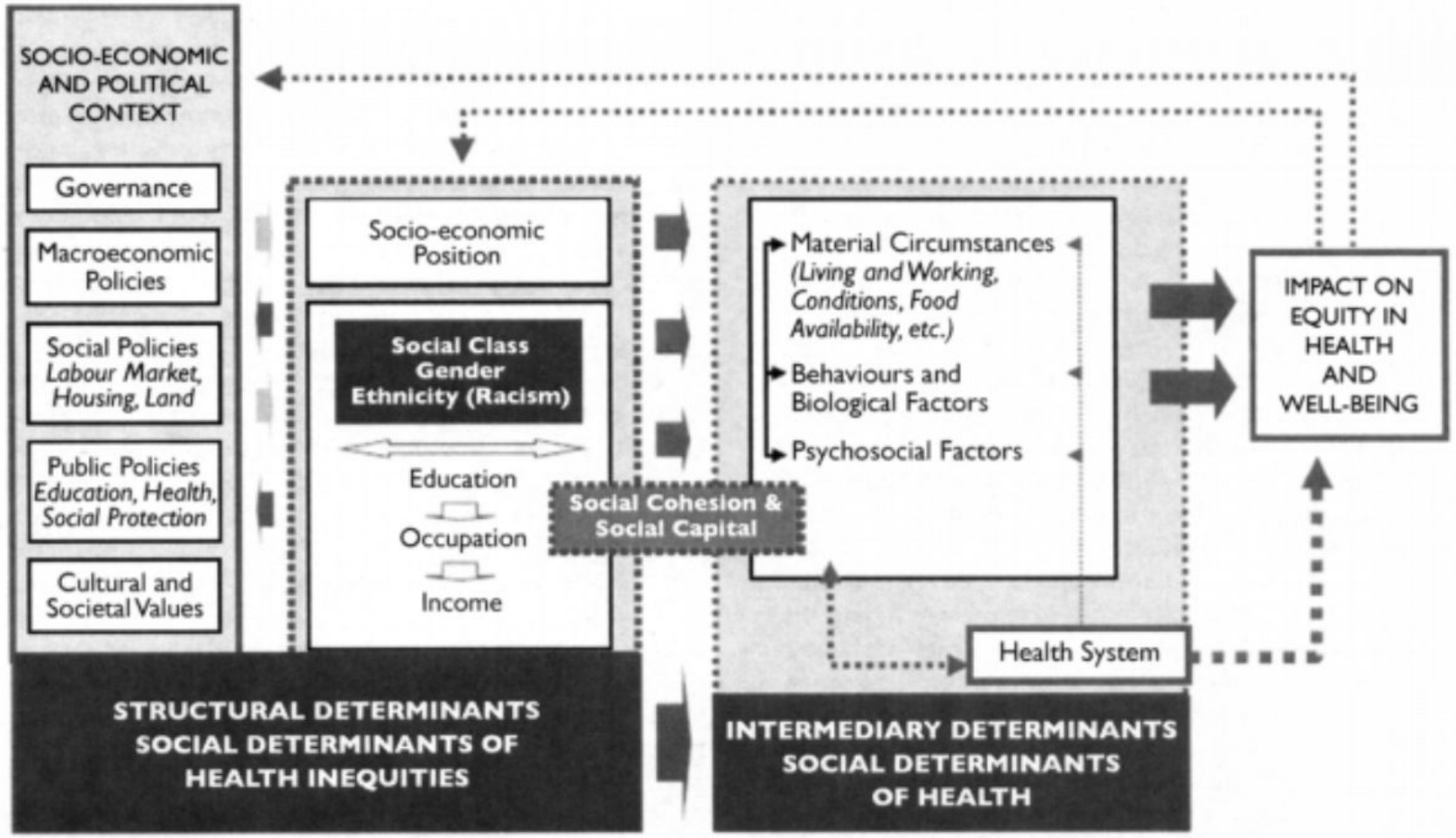


Figure 1: WHO Conceptual Framework on the Social Determinants of Health

Source: WHO (2010, 6).

ALIMEZELLI, H., LEIS, A., DENIS, W., & KARUNANAYAKE, C. (2015). Lost in Policy Translation: Canadian Minority Francophones and Health Disparities. *Canadian Public Policy / Analyse De Politiques, 41*, S44-S52.

VirginiaTech. Biocomplexity Institute

Baseline measure

- Calculate baseline measure of racial equity in Alexandria
- Data source:
  - ACS does not provide cross-tables with race variable
  - Use data from NHGIS (National Historical Geographic Information System)
- Look at racial breakdown of key variables such as education, income, household details, health status etc.
- Geographical levels:
  - Zip codes
  - Block groups
- Visualize for the most recent time period for which data is available
- Compare with residents' sentiment as understood by the Alexandria Resident Survey

Deliverables

- Literature review resulting in annotated bibliography
- Extensive data bibliography
- Baseline on racial equity in Alexandria, referenced with results from the Alexandria Residents' Survey

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Source	**Type (Academic, administrative (Local, Fed, International))**	**Who prepared it?**	**Citation and/or URL**	**Category of variable (education, healthcare etc.)**	**Variable**	**Variable definition (if available)**	**Data source**	**Administrative or survey based?**	**Level/unit of measurement (individual, block group, census tract etc.)**	**Frequency of collection (if known)**	**Additional notes**	**Publicly available?**	**Quality (bias, sampling frame, sampling quality etc.)**	**Comments on how they measure equity**	**Who might be excluded?**
National Equity Atlas	**Administrative (From local to national)**	National Equity Atlas (and varying US agencies)	http://nationalequityatlas.org/indicators							Varies based upon indicator					
			http://nationalequityatlas.org/indicators/Detailed_race-ethnicity	Demographics	Race/Ethnicity	The composition of the population by race/ethnicity and nativity. Latinos include people of Hispanic origin of any race and all other groups are non-Hispanic.	IPUMS	Administrative	Individual	2010-2014		Yes			
			http://nationalequityatlas.org/indicators/Population_growth_rates	Demographics	Population Growth Rates	The net percentage change in population by race/ethnicity. Latinos include people of Hispanic origin of any race and all other groups are non-Hispanic.	U.S. Census Bureau	Administrative	Individual	2000-2010		Yes			
			http://nationalequityatlas.org/indicators/Contribution_to_growth%3A_Immigrants	Demographics	Contribution to growth: immigrants	The net change in population by nativity and the share of the net change in population attributable to immigrants.	IPUMS	Administrative	Individual	1980-2014		Yes			
			http://nationalequityatlas.org/indicators/Contribution_to_growth%3A_People_of_color	Demographics	Contribution to growth: people of Color	The net change in population by race/ethnicity and the share of the net change in population attributable to people of color (all people who do not identify as non-Hispanic white).	U.S. Census Bureau	Administrative	Individual	1980-2010		Yes			
			http://nationalequityatlas.org/indicators/Racial_generation_gap	Demographics	Racial Generation Gap	The percentage of youth (persons under 18) and seniors (persons 65 or older) who are people of color (all people who do not identify as non-Hispanic white). The racial generation gap is the difference in the percent of youth and seniors of color.	U.S. Census Bureau	Administrative	Individual	1980-2014		Yes			
			http://nationalequityatlas.org/indicators/Diversity_index	Demographics	Diversity Index	The Diversity score is a measure of racial/ethnic diversity in a given region.	U.S. Census Bureau	Administrative	Individual	1980-2015		Yes			
			http://nationalequityatlas.org/indicators/Median_age	Demographics	Median Age	The median age for the population by race/ethnicity.	IPUMS	Administrative	Individual	2014		Yes			

Community Projects: Fairfax - Youth Projections

Team: Zarni Htet, Chanida Lerdritsomboon, Eiriki Iversen, Ben J. Swartz
SDAL: Josh Goldstein and Stephanie Shipp
Sponsor: Fairfax Health and Human Services (HHS) -
Michelle Gregory, Sophia Dutton, Linda Hoffman

VirginiaTech Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Project Description

- Fairfax Health and Human Services (HHS) is interested in developing long-term population and economic forecasting capabilities

- **Research Questions**
  1. Which HHS programs and services may be affected by changes in population size, behaviors, and demographic makeup?
  2. What current and future trends in demographics, housing, and other factors would inform the models to predict changes in their programs?
  3. What are trends in the data that warrant more investigation?

Specific Research Questions

- Are mental health services accessible to address depression, anxiety, and eating disorders?
- Are mental health services available?

Literature Review

What is Mental Health Disorder?

- According to the World Health Organization (WHO), a mental health disorder is:
  - behavioral/psychological,
  - of clinical significance,
  - accompanied by a concomitant distress and/or a raised risk of death, or an important loss of freedom
  - involves an unexpected cultural response to any situation

Source:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC556080/

Literature Review

Mental health problems and disorders in adolescents

Problem or disorder	Prevalence (%)
Common	
Depression	3-5
Anxiety	4-6
Attention-deficit/hyperactivity disorder	2-4
Eating disorders	1-2
Conduct disorder	4-6
Substance misuse disorder	2-3

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC556080/

Literature Review continued..

Service availability barriers

- Physical Location
  - –Low visibility for anonymity
  - –High visibility for increased awareness
  - –Comfortable environment for patient
    - •Potential Solutions: Outreach programs, Recreational activities
- Cultural Barriers
  - –The idea of what mental health is will vary from culture to culture
  - –Stigma towards receiving care
- Administrative Barriers
  - –Operating hours conflict with schedule of the clientele
  - –Inside school hours
    - •what is the appropriate time to request mental health care?

Source: https://www.ncbi.nlm.nih.gov/pubmed/24898523

Literature Review continued..

Los Angeles Unified School District and the San Francisco School District Mental Health Services

- Similar to the mental health services provided by Fairfax County
  - Wellness/prevention services, community based organizations, intervention programs, Medication support services, etc.

Barriers to mental health in Fairfax county

- Inconsistent/incomplete information with parents, mostly regarding insurance coverage, available providers, etc.
  - Lack of a coordinated, cross system training curriculum. Also, low income families lack transportation, health insurance and other resources that inhibits them from accessing these mental health services.

Data Discovery

- ★ American Community Survey (ACS)
  - Characterize and identify areas of low mental health accessibility
- ★ Fairfax Community Health Dashboard
  - Identifying possible relation between community indicators to predict mental health disorder trends
- ★ Web Scrape of Psychology Today
- ★ Fairfax Youth Survey

Star - selected sources

Data Discovery

- Substance Abuse and Mental Health Services (SAMSHA) -
  - Mental Health Facilities Data
- Virginia Department of Education (VDOE)
  - Fall enrollment - characteristics by school
  - On time graduation rates by school
  - Dropout rates by school

Topics from Survey

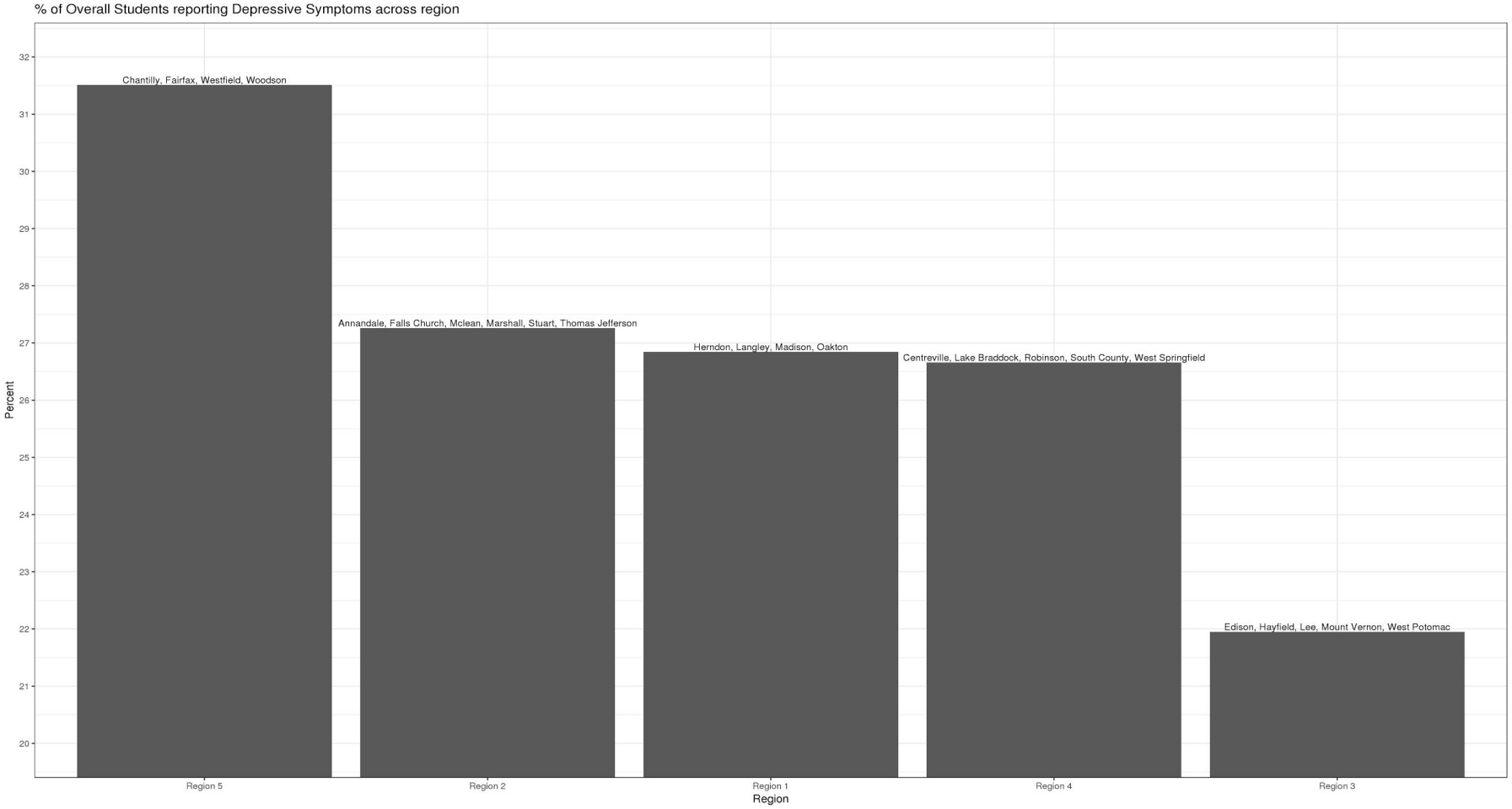
- Indicators examined from Fairfax County Youth Survey:
  - depression
  - suicide consider
  - suicide attempt
  - high stress
  - adults talking:
    - "There are lots of adults in my neighborhood I could talk to about something important?"
  - binge drinking
  - bullied victims in school
  - bullied victims not in school
  - cigarette usage
  - marijuana usage
  - cyberbullying victim
  - cyberbullying aggressor
  - extracurricular activities available
  - extracurricular activities regularly
  - food insecurity
  - fruit and vegetables consumption
  - gratitude
    - "how much do you agree with the statement: 'I have much in life to be thankful for' ?"
  - parent help available
  - physical activity daily
  - physical activity none
  - sleep 4 hours or less
  - sleep exactly 6 hours

Source: http://www.fairfaxcounty.gov/demogrph/youthpdf.htm

Survey Question Definitions

- Depressive Symptoms:
  - Those who felt so sad or hopeless almost every day for two weeks or more in a row during the past 12 months
- Food Insecurity:
  - Those who go hungry because there was not enough food at home during the past 30 days
- Physical Activity None:
  - Those physically active for less than 60 minutes during the past 7 days
- Parent Help Available:
  - Those who had a personal problem that could ask their parents for help
- Extracurricular Regularly:
  - Those who participated in school or non-school-based activities after the regular school day ended

% Students Reporting Depression by Fairfax School Regions



5 2 1 4 3

REGION

List of High Schools by Region

Region 5 - Chantilly, Fairfax, Westfield, Woodson
Region 2 - Annandale, Falls Church, McLean, Marshall, and Stuart Pyramids + Thomas Jefferson High School for Science
Region 1 - Herndon, Langley, Madison, Oakton, and South Lakes
Region 4- Centreville, Robinson, South County, West Springfield, Lake Braddock
Region 3 - Edison, Hayfield, Lee, Mount Vernon, and West Potomac

Fairfax Youth Survey - Selected Results from 2014-2015

- For those with high % of depressive symptoms (relative to other high school pyramids), and other factors that may be related

High School pyramid	Demographic	Depressive Symptoms (%)	Food Insecurity (%)	No Physical Activity (%)	Parent Help Available (%)	Extra-curricular Regularly (%)
FAIRFAX	Overall	32	25	12	76	72
LEE	Overall	29	28	17	76	58
MOUNT VERNON	Overall	28	31	19	77	59

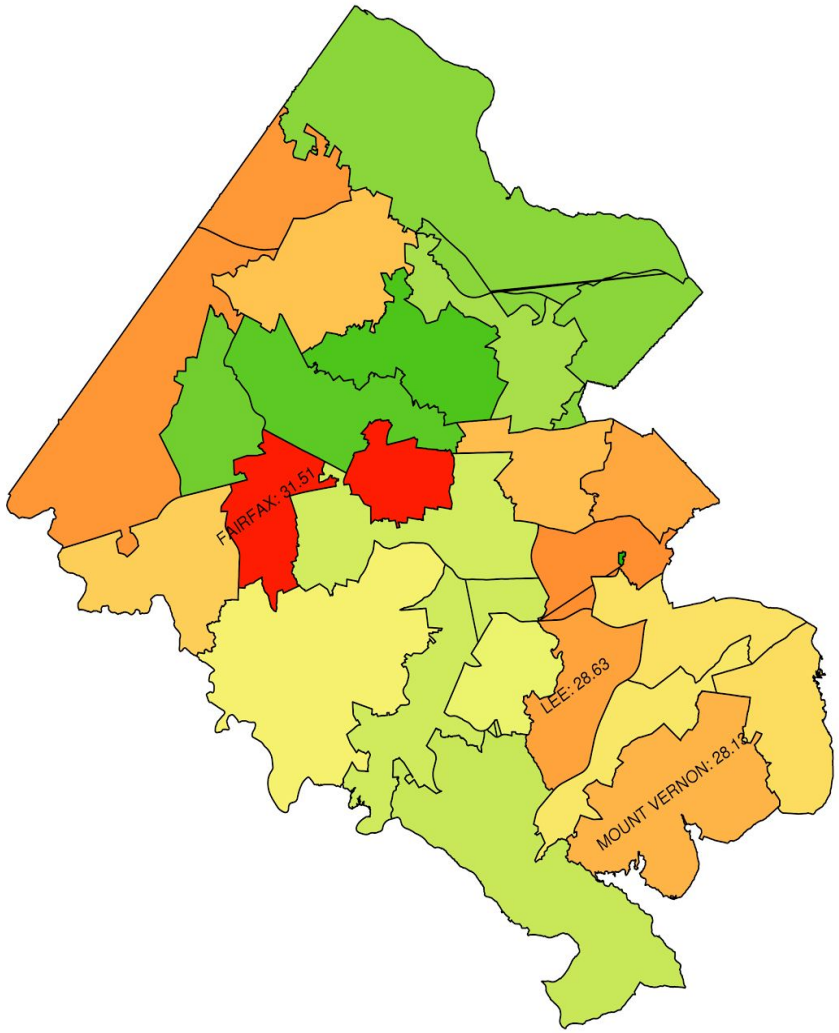
Source: http://www.fairfaxcounty.gov/demogrph/youthpdf.htm

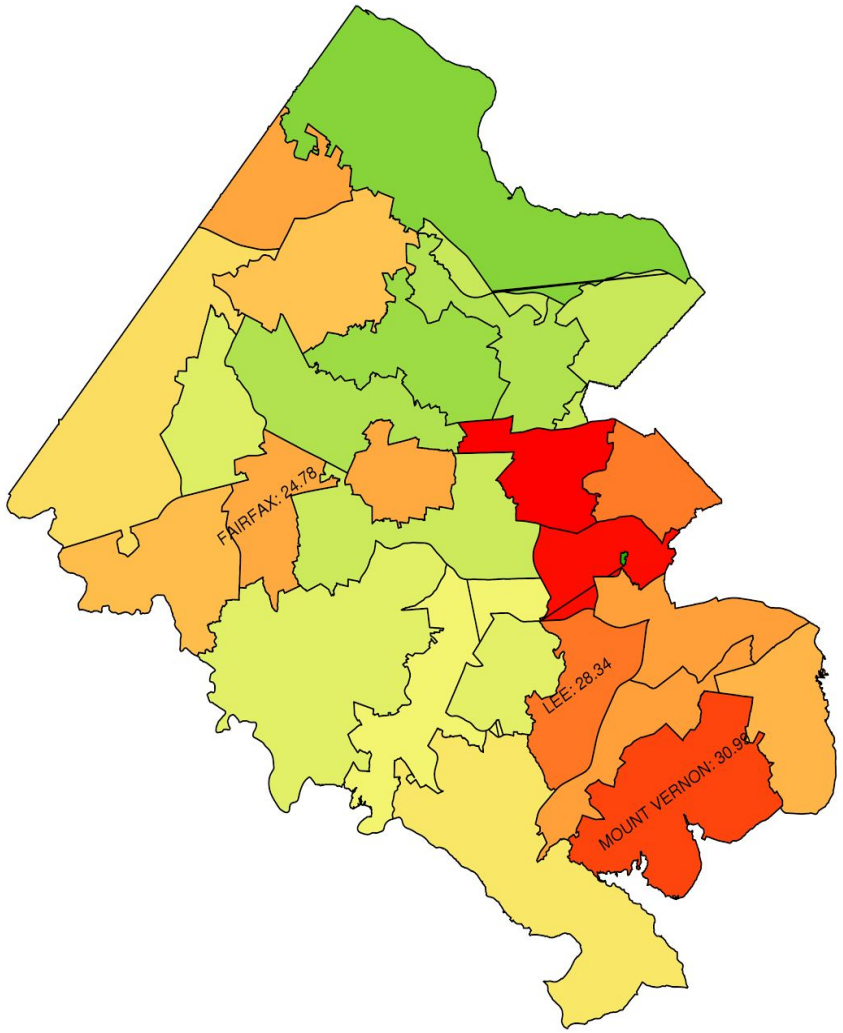
% Students Reporting Depressive Symptoms (left) and % Students Reporting Food Insecurity (right)

Shades of green = relatively good (light) to good dark)

Red = bad; e.g. high % of students with depressive symptoms

Shades of gold = in the middle from light gold (top middle) to dark gold (bottom middle)





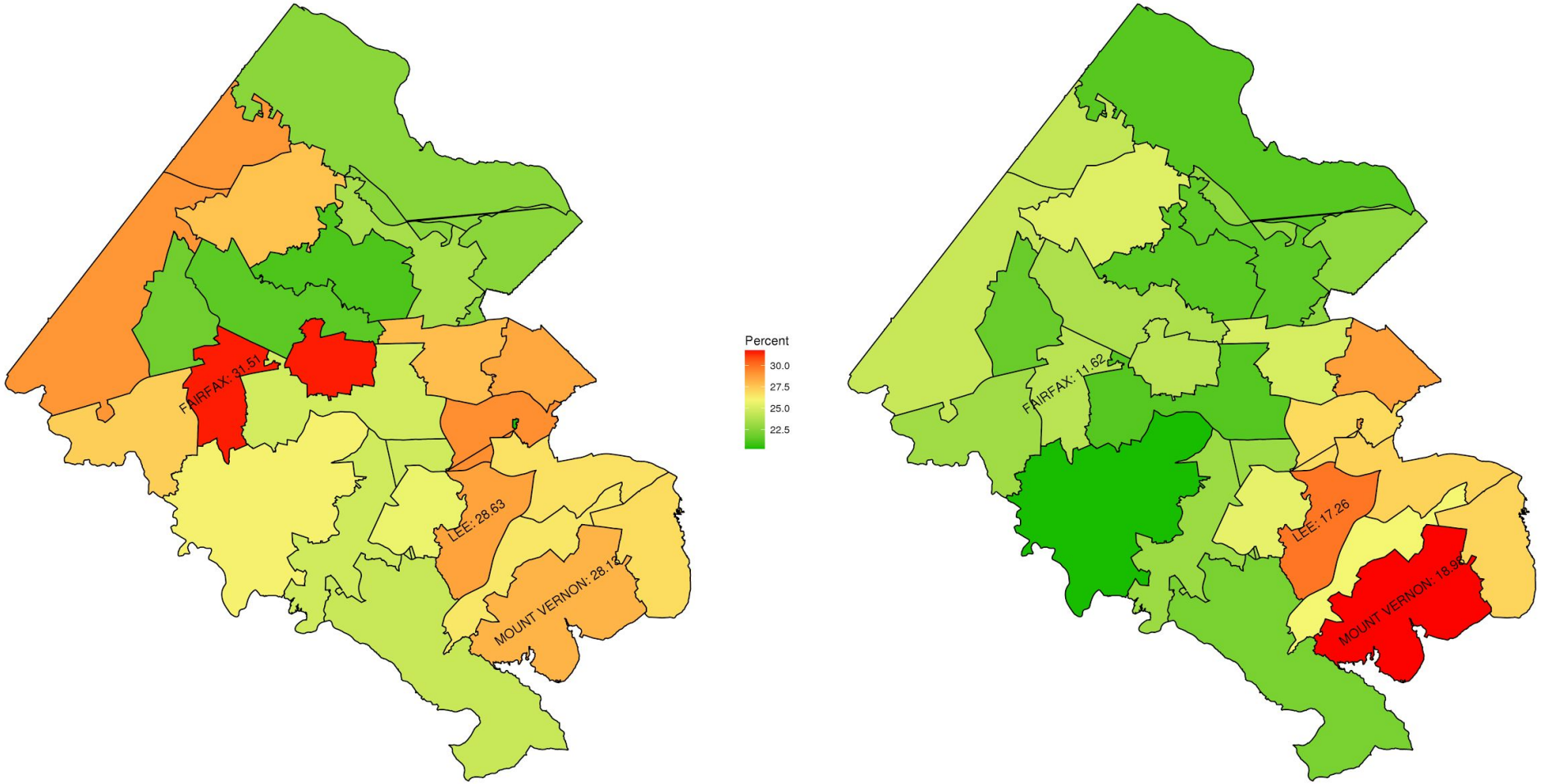
Each area on the map is a high school pyramid boundary that includes the high school, middle schools, and elementary schools for that boundary.

% Students Reporting Depressive Symptoms (left) and % Students Reporting Physical Activity (right)

Shades of green = relatively good (light) to good dark)

Red = bad; e.g. high % of students with depressive symptoms

Shades of gold = in the middle from light gold (top middle) to dark gold (bottom middle)



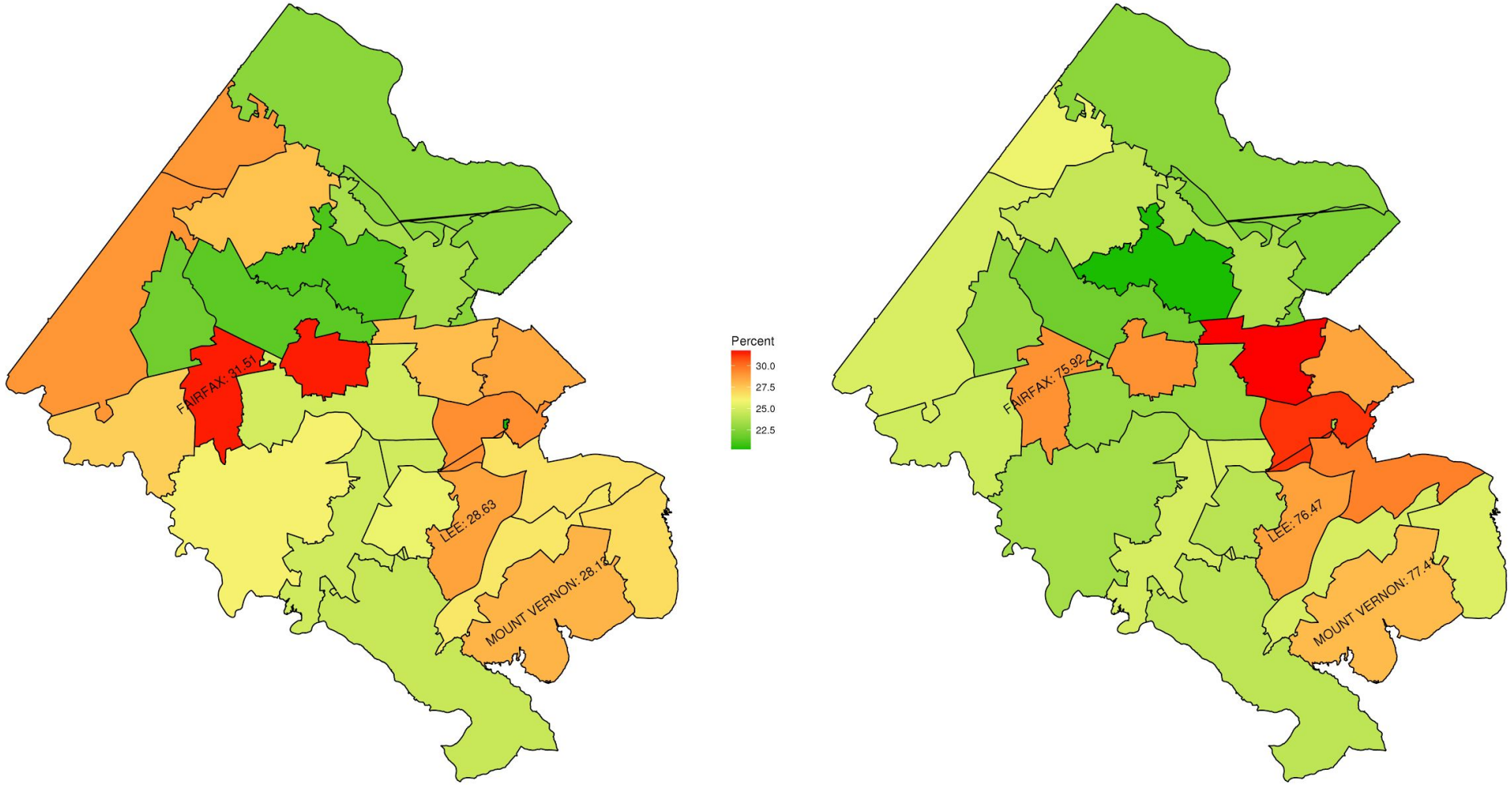
Each area on the map is a high school pyramid boundary that includes the high school, middle schools, and elementary schools for that boundary.

% Students Reporting Depressive Symptoms (left) and % Students Reporting Parents' Help Available (right)

Shades of green = relatively good (light) to good dark)

Red = bad; e.g. high % of students with depressive symptoms

Shades of gold = in the middle from light gold (top middle) to dark gold (bottom middle)



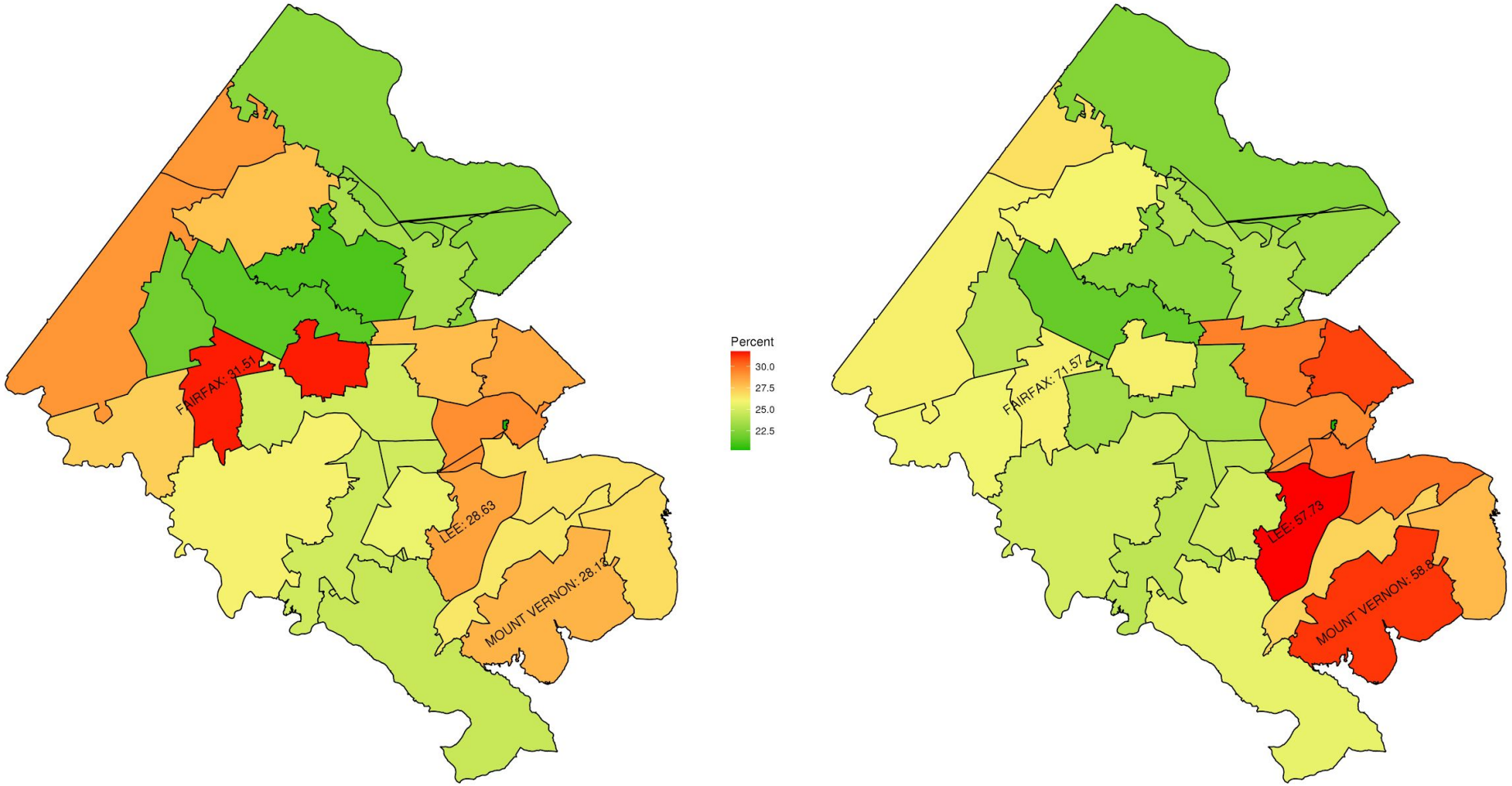
Each area on the map is a high school pyramid boundary that includes the high school, middle schools, and elementary schools for that boundary.

% Students Reporting Depressive Symptoms (left) and % Students Reporting Extracurricular Regularly (right)

Shades of green = relatively good (light) to good dark)

Red = bad; e.g. high % of students with depressive symptoms

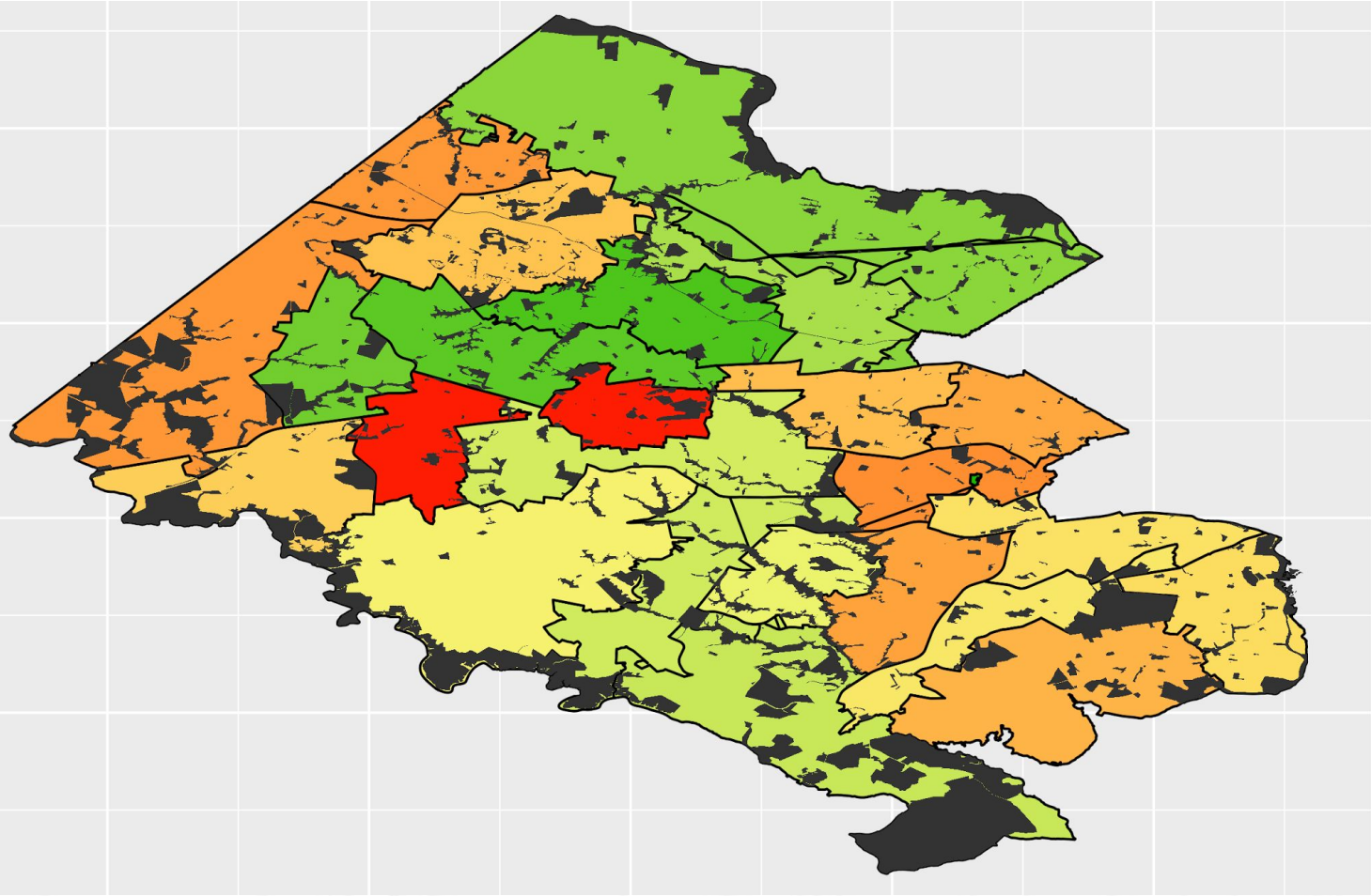
Shades of gold = in the middle from light gold (top middle) to dark gold (bottom middle)



Each area on the map is a high school pyramid boundary that includes the high school, middle schools, and elementary schools for that boundary.

VirginiaTech
Biocomplexity Institute

% Students Reporting Depressive Symptoms and Availability of Parks in Fairfax County (black shaded areas)



Shades of green = relatively good (light) to good dark)

Red = bad; e.g. high % of students with depressive symptoms

Shades of gold = in the middle from light gold (top middle) to dark gold (bottom middle)

Psychology Today Web Scrape

- SAMHSA lists mental health service providers that are funded by a state agency -- incomplete
- psychologytoday.com has a directory of mental health service providers
- Scraped addresses of every therapist, psychologist, and psychiatrist in Fairfax County listed on their website
- Converted address strings to lat-longs in Google API
- Overlaid locations on a heat map of students reporting depressive symptoms by school boundary

Steps Ahead

- Creating a synthetic data set of all residents in Fairfax county using PUMS_PERSONAL and ACS to capture ACS variables of interest by high school pyramid boundary
  - Reason for imputation: ACS variables are by ZCTAs and not School Boundaries
- Run a simple logistic model with the Fairfax Youth Survey microdata

ZCTA are Zip Code Tabulation Areas used by Census Bureau

Community Projects: Roanoke - Refugees

Team: Claire Kelling, Adrienne Rogers, Hannah Brinkley
Kyle Morgan, Craig Mortan
SDAL: Gizem Korkmaz, M. Orr
Sponsor: Virginia Tech Global Forum for Urban and Regional Resiliency

Virginia Tech
Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

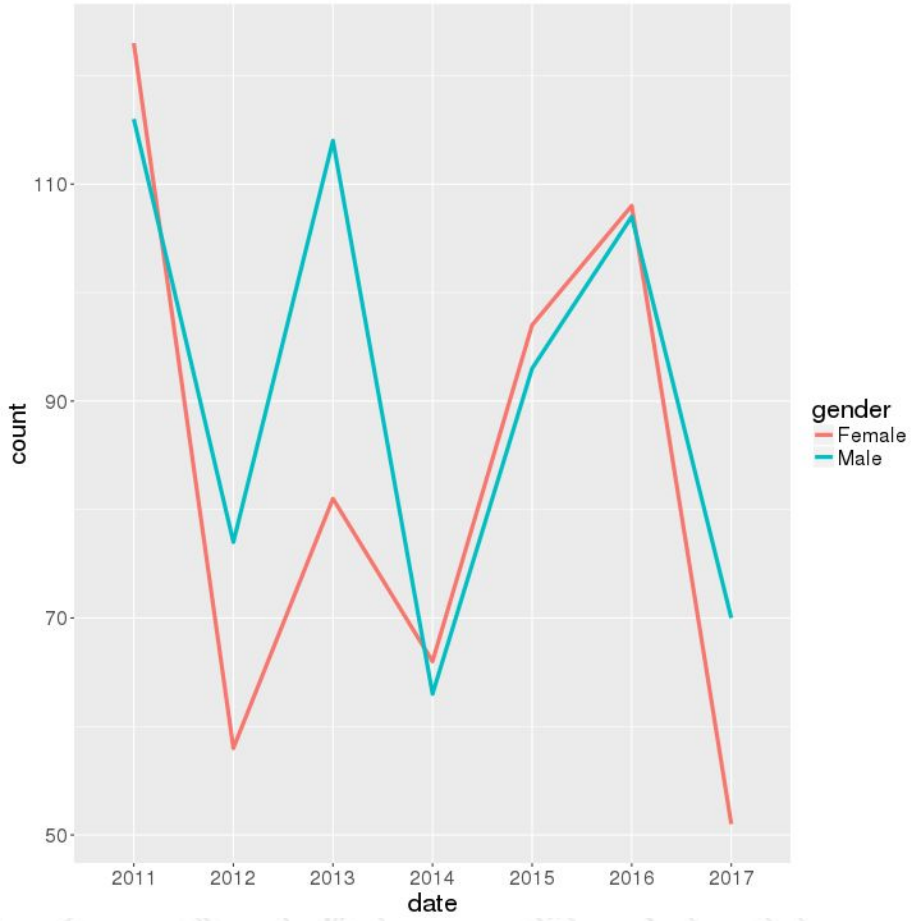
Project Description

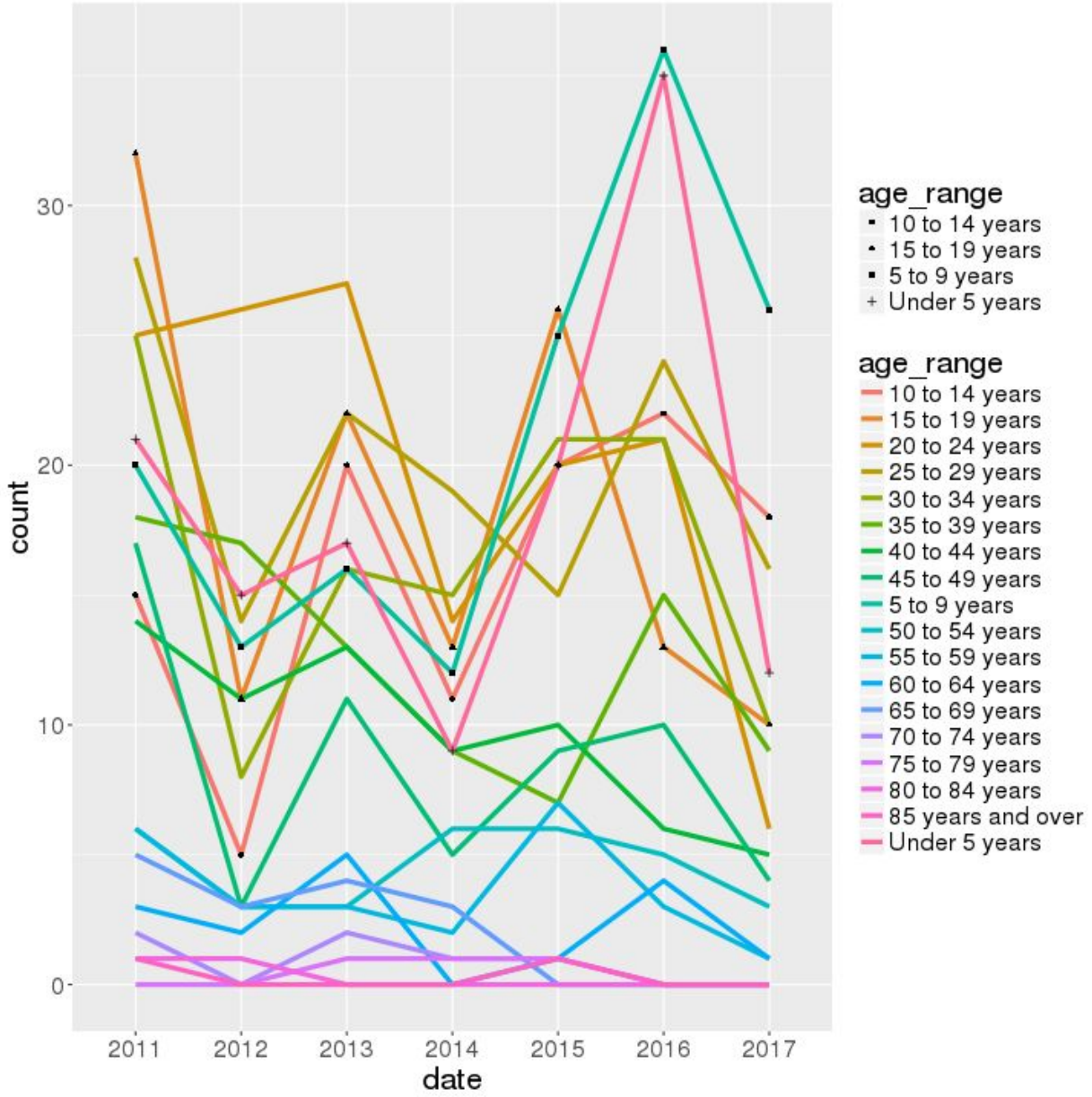
- Part of the Community Indicators Project
- SPIA, Global Forum and other groups are interested in expanding capacity at Virginia Tech for understanding refugees
- Research Questions:
  - 1) How can we measure the **integration** of refugees into host communities and countries?
  - 2) How can we measure the **economic impact** of refugees on host communities and countries?
- Literature Review:
  - Cleveland Case Study
  - UN provided guidance on research questions to the Virginia Tech teams

Local Data Analysis

Virginia Department of Health:

- Most refugees in Roanoke are young (0-19)
- Gender distribution is fairly even over time

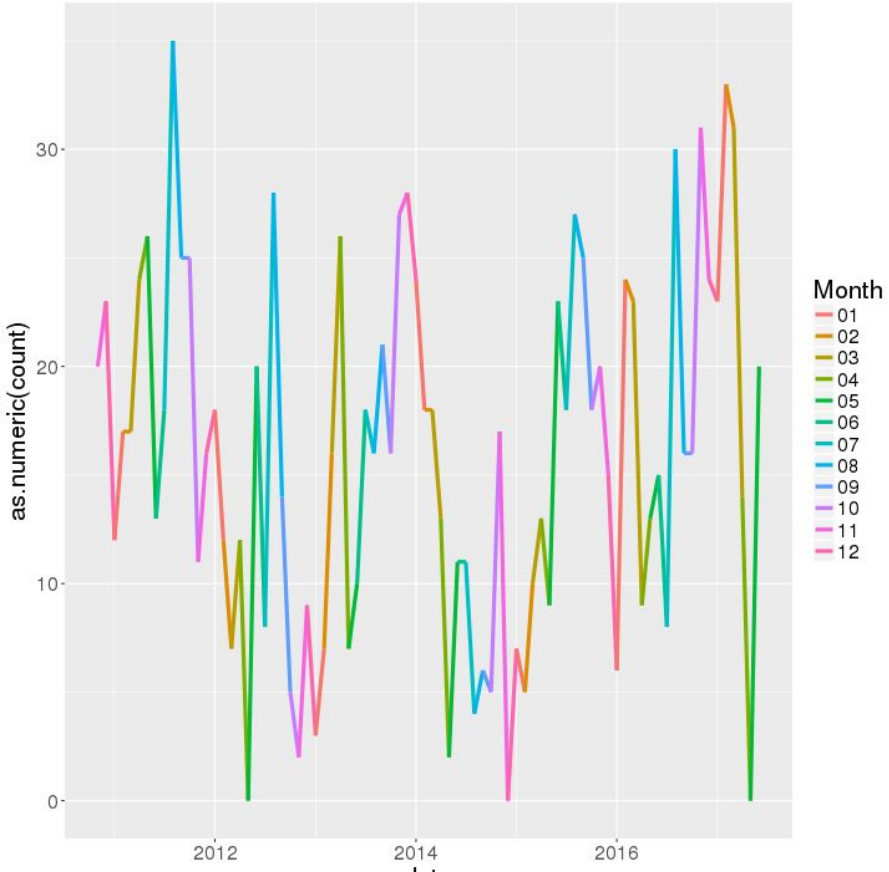


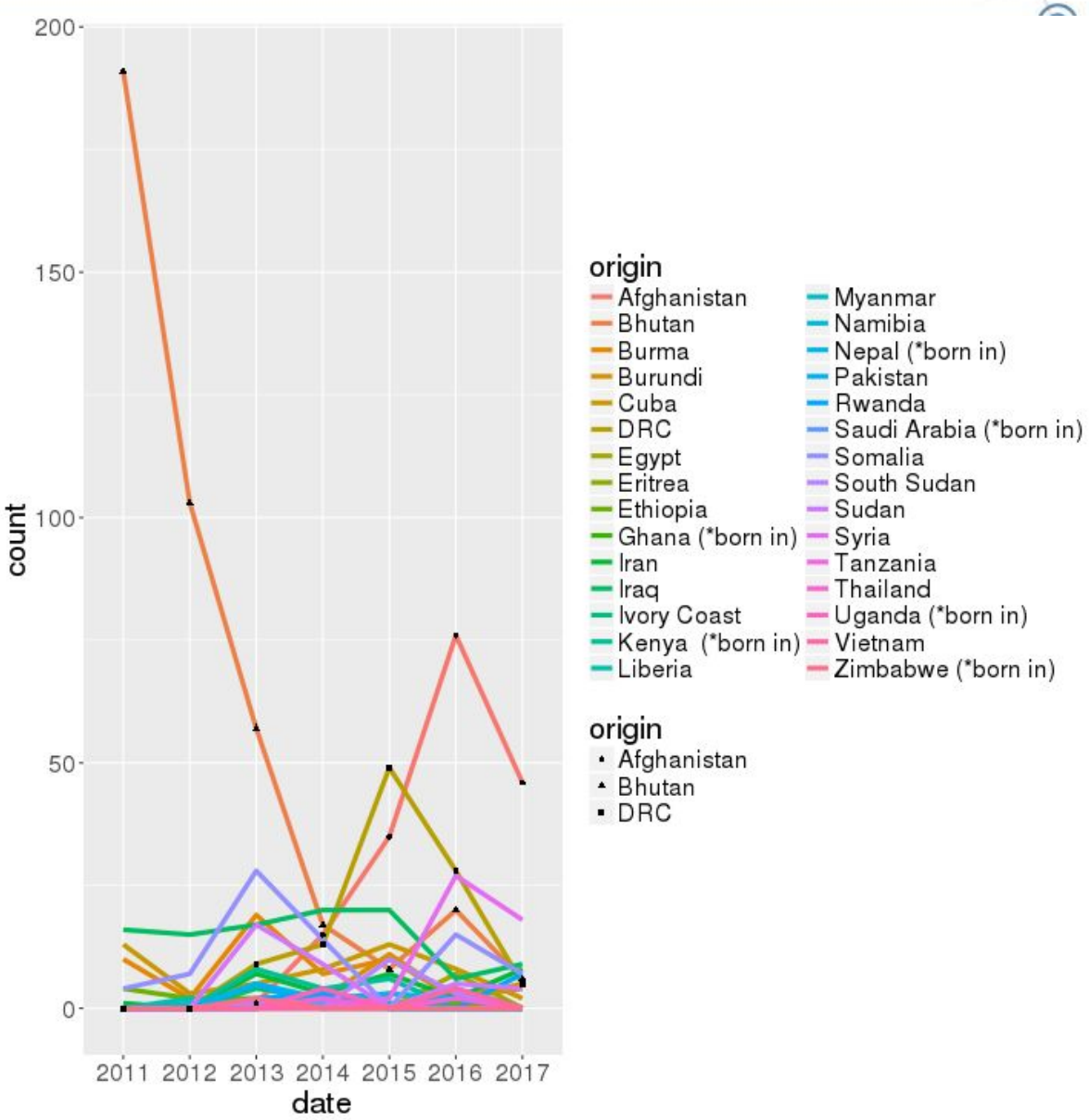


Local Data Analysis

Virginia Department of Health:

- Not much of a seasonal trend (such as spikes in the beginning of the year)
- Most refugees in Roanoke are from Afghanistan, Democratic Republic of the Congo, and Bhutan (with recent increases from Syria)





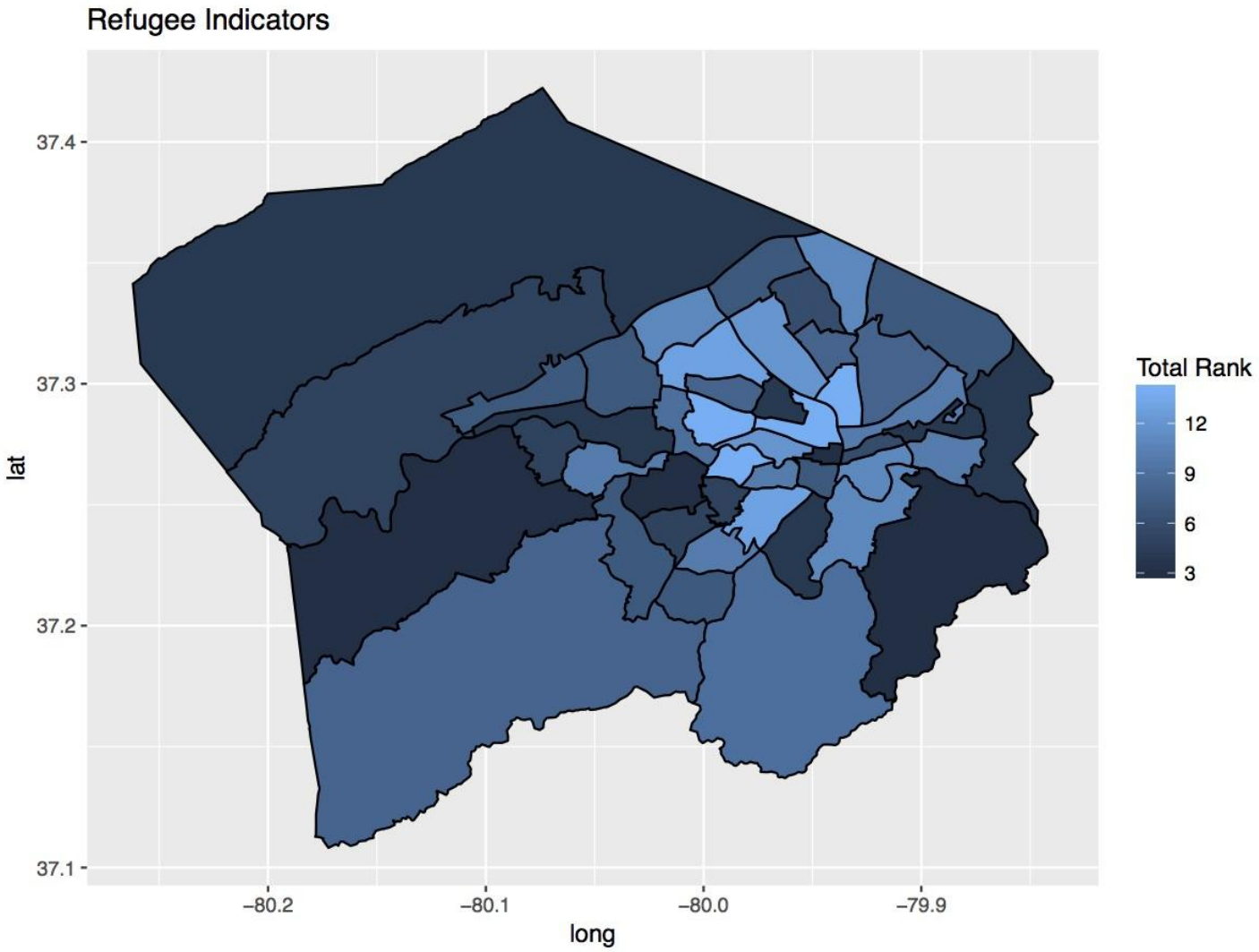
Local Data Analysis

- Virginia Newcomer Information System:
  - Clients: Asylee, Refugees, SIV virus, Other (Trafficking Victims and Cuban Haitian Entrants)
  - Clients entering employment within 90 days of intake:
    - Roanoke: 51%
    - State Average: 48%
  - Percent of employed Clients working Full Time:
    - Roanoke: 73%
    - State Average: 68%

Refugee Indicators Index

Used American Community Survey (ACS) data tables from Roanoke County, Roanoke City, and Salem City at the Census Tract level

- Variables used from the ACS:
  - Number of foreign-born people who are not citizens
  - Number of people from countries that the U.S. takes refugees from
  - Number of people who either speak English poorly or do not speak English at all

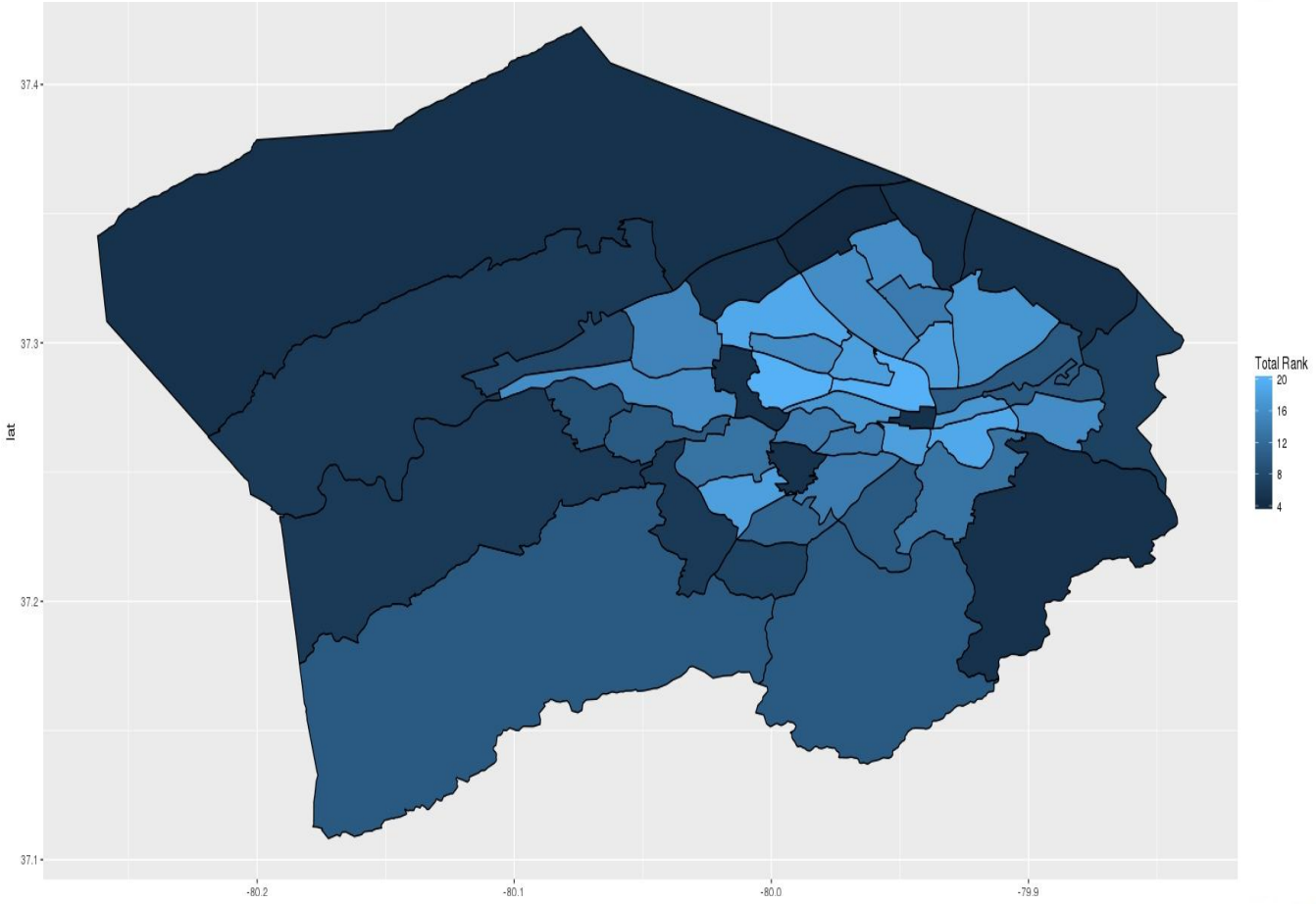


Economic Indicators Index

Similar to the Refugee Indicator's approach, we put together a map of some variables that have to do with economic vulnerability

- Variables we looked at were all on the household level:
  - households that received food stamps
  - households that have no vehicle
  - households that have income below the poverty line
  - households that spend more than 50% of their income on rent / mortgage
- Variables were found for the census block level to keep it homogenous
- Similar to the map before, a lot of the variables have high composite scores closer to the city of Roanoke

Vulnerable populations based on Socio-Economic Factors



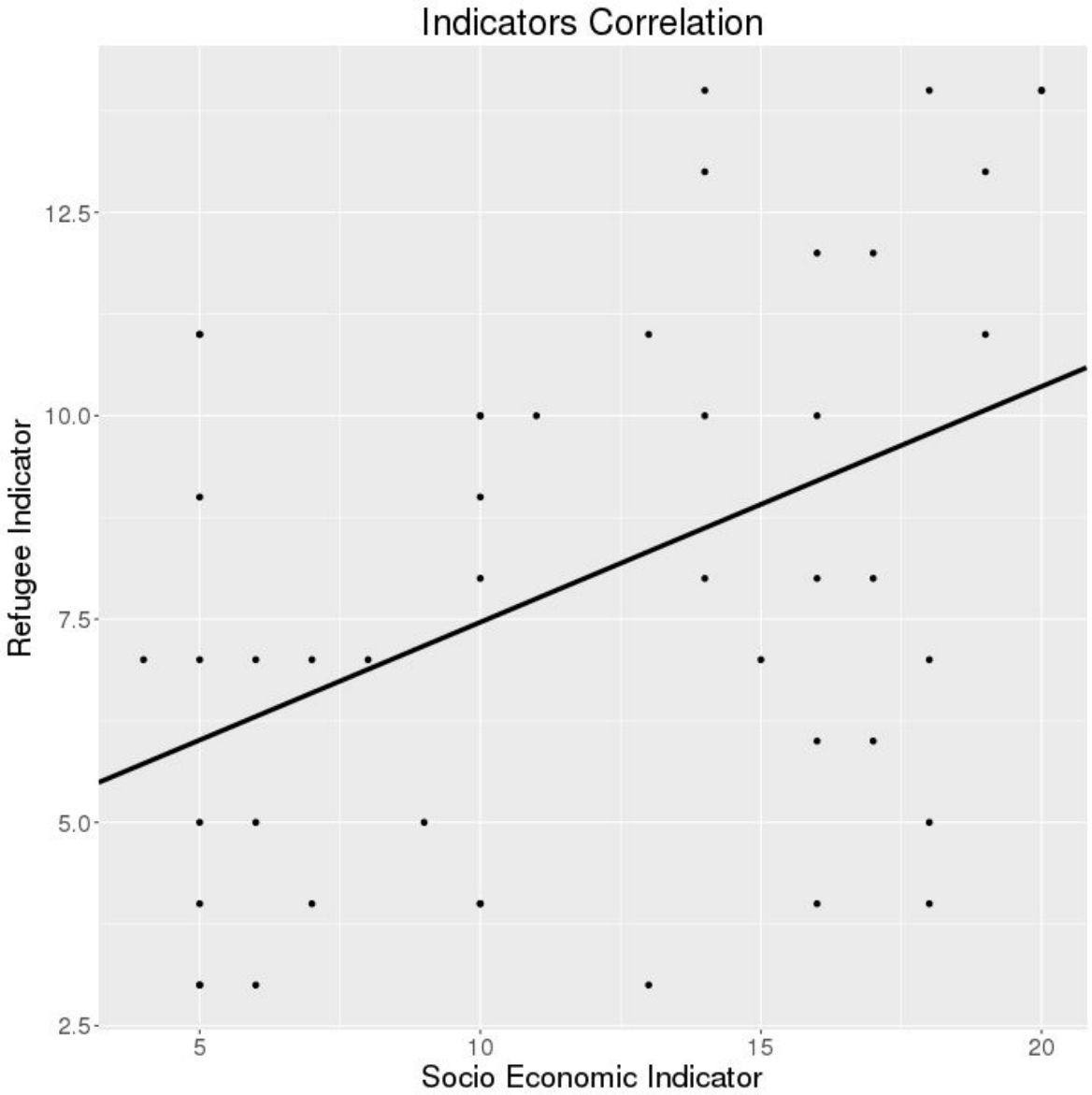
VirginiaTech
Biocomplexity Institute

Refugees and Socio Economic Indicators

Refugee Indicator & Socio-Economic Indicator

- These two indices do not seem to be well correlated (46 census tracts)
  - Correlation: 0.4382
  - Regression r^2: 0.192

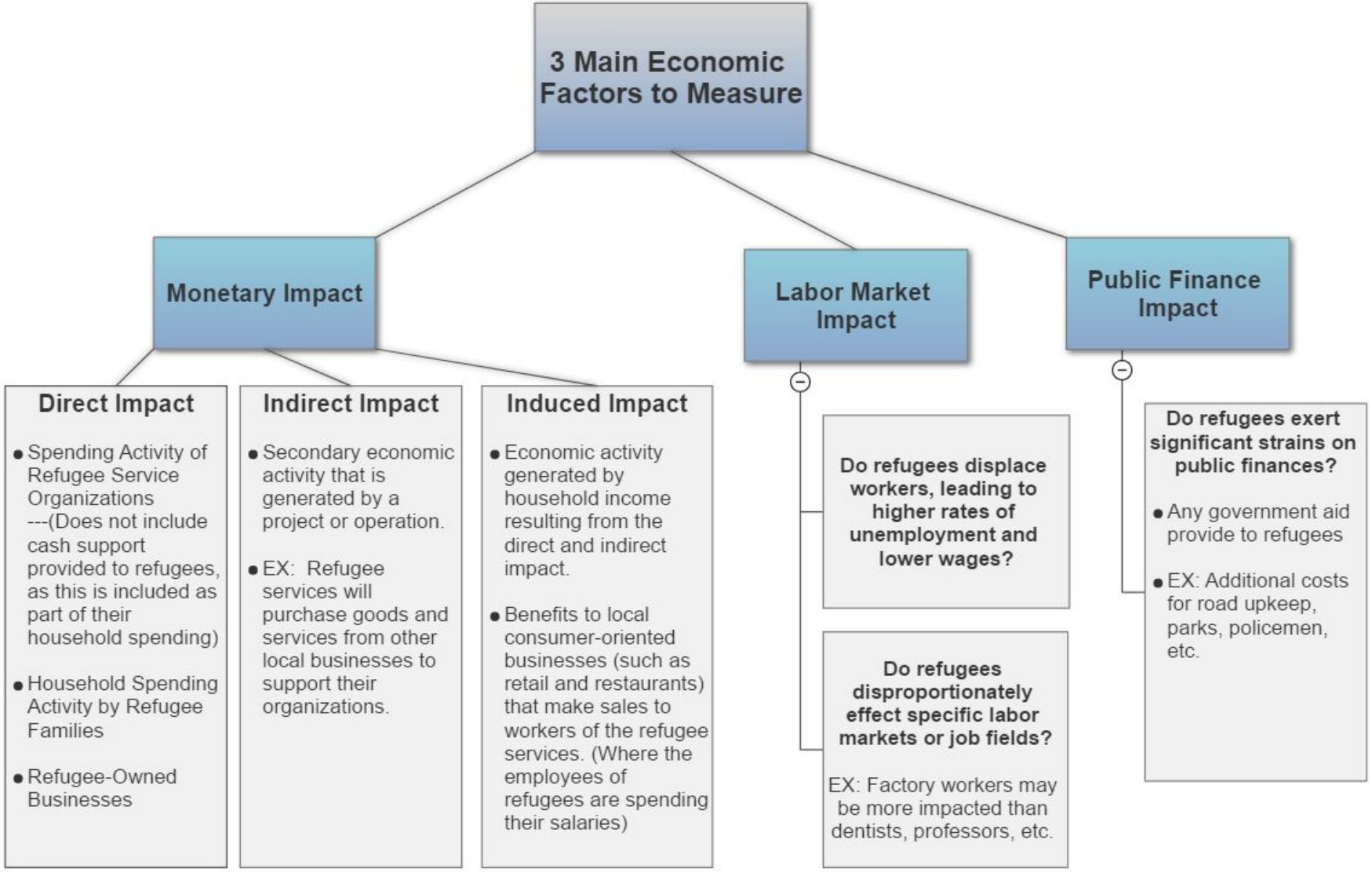
- Considering additional/removing variables



Economic Indicators

Cleveland Case Study:

- Used a survey of refugee organizations and refugees to obtain data on refugee spending, housing, and employment



Challenges

Challenges of measuring economic impact:

- Lack of access to refugees / local data
- Staggered arrival dates
- Getting input from local businesses
- Infeasibility of determining whether refugees are “replacing” other workers or simply filling open jobs

Cleveland Case Study:

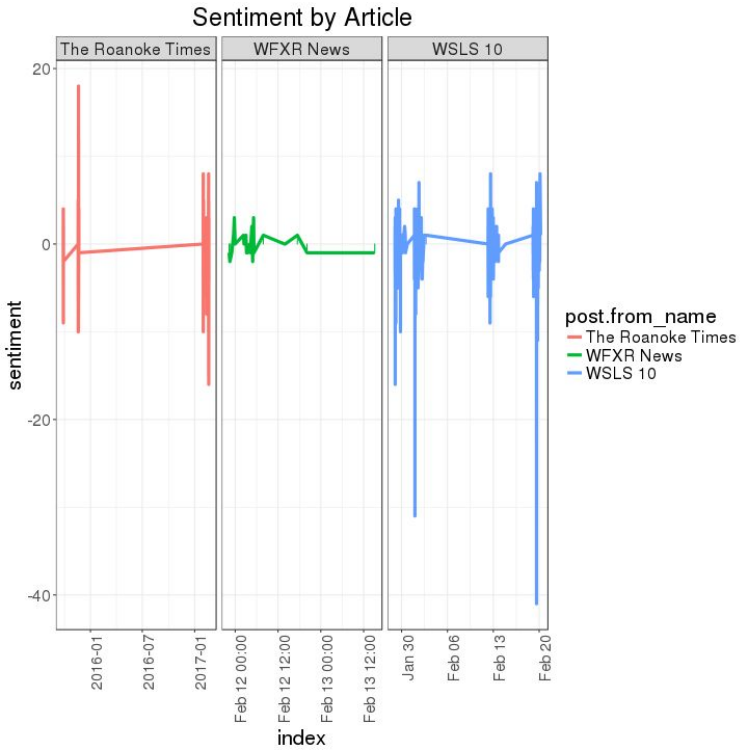
- Not a viable option due to our lack of access to refugees that would have been necessary for a survey

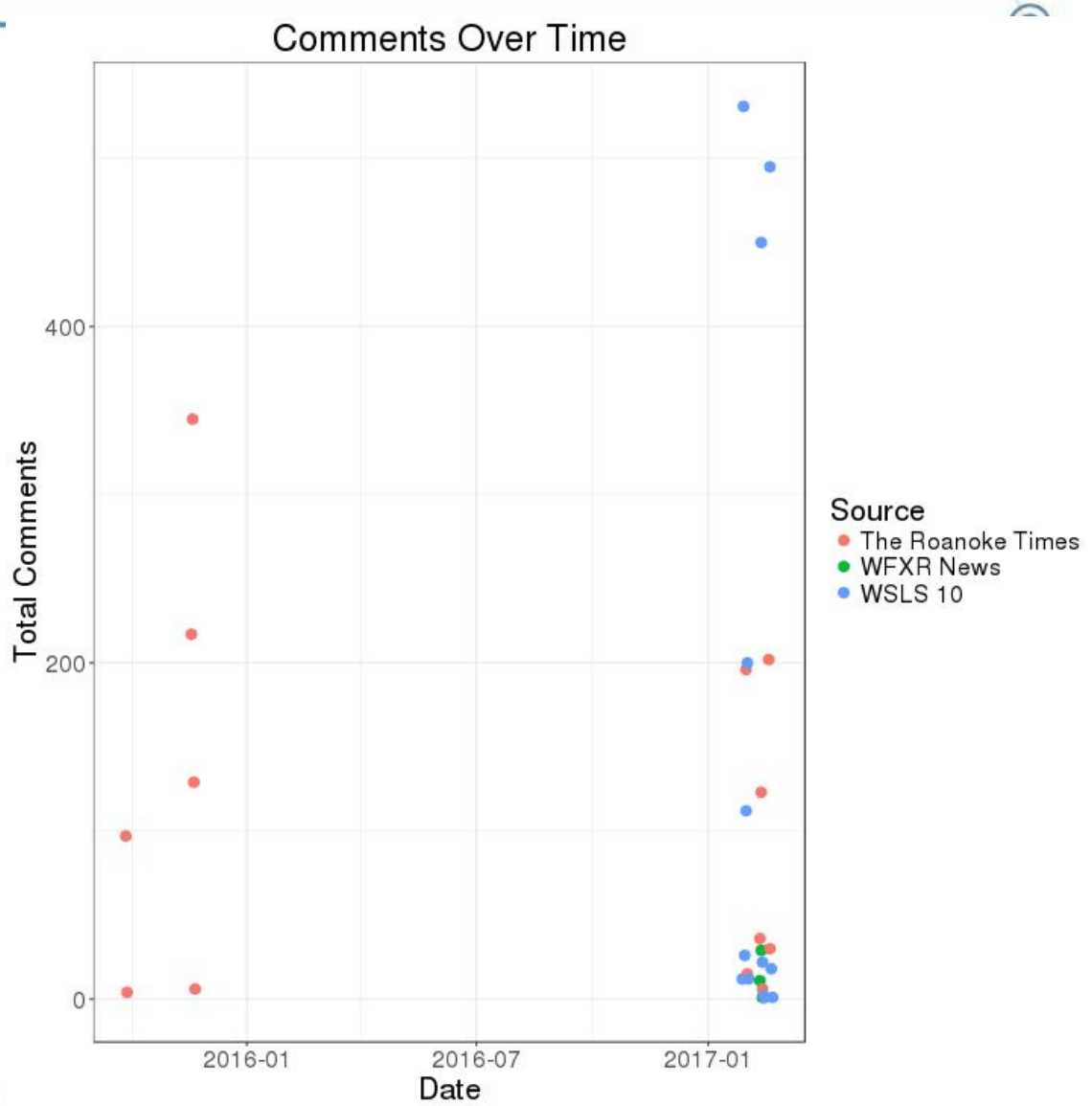
Analyzing Economic Characteristics

- Assess Economic Characteristics of communities likely to have larger Refugee populations compared to Roanoke County averages
- Create Index of Economic Status compared to Roanoke County Averages
  - *Based on research by Appalachian Regional Commission*
- Variables to look into for assessment:
  - ACS: Food Stamps/SNAP, Public Assistance, Employment, Renter Housing Costs
  - Others: Tax Brackets, Charitable Income
- Issues:
  - This analysis will only measure economic characteristics of refugee-inhabited areas, not the impact on refugees themselves
- Future Questions:
  - How do these areas compare to known areas with more and less refugees?

Sentiment Analysis

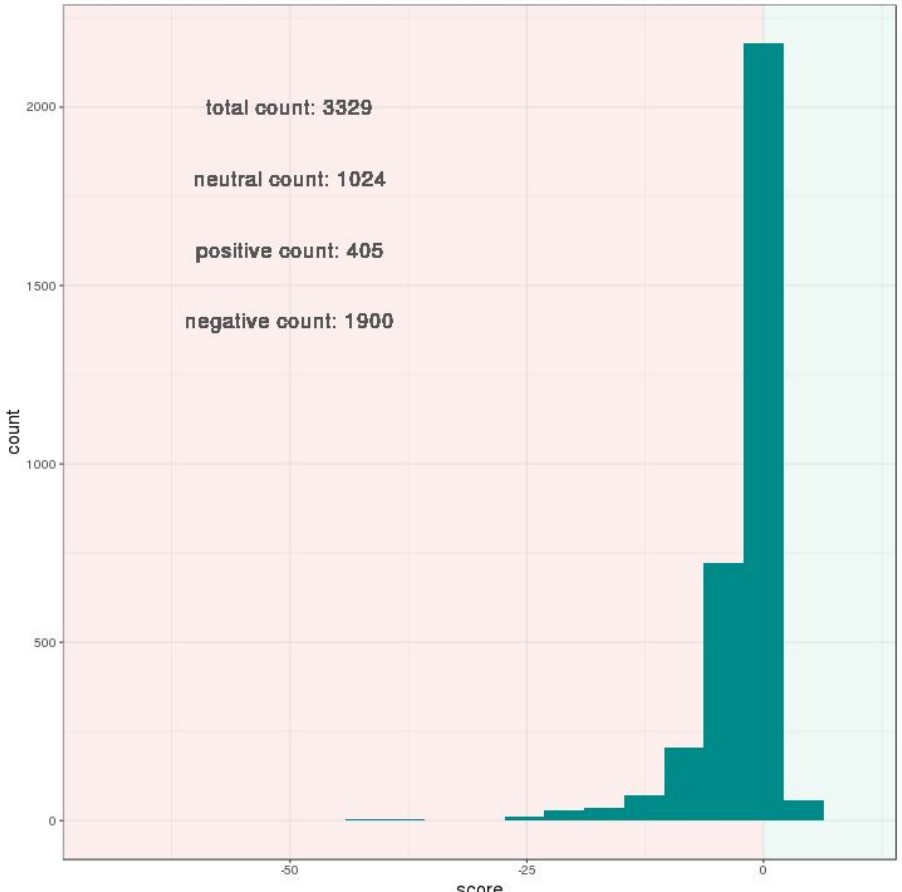
- Data Source: Local Roanoke News Facebook pages
  - The Roanoke Times, WSLS 10, WFXR News
  - Scraped 3,329 comments from 12 articles
- Focused around two events:
  - Roanoke Mayor David Bowers requested suspended assistance to Syrian refugees
  - Recent political events surrounding presidential policies

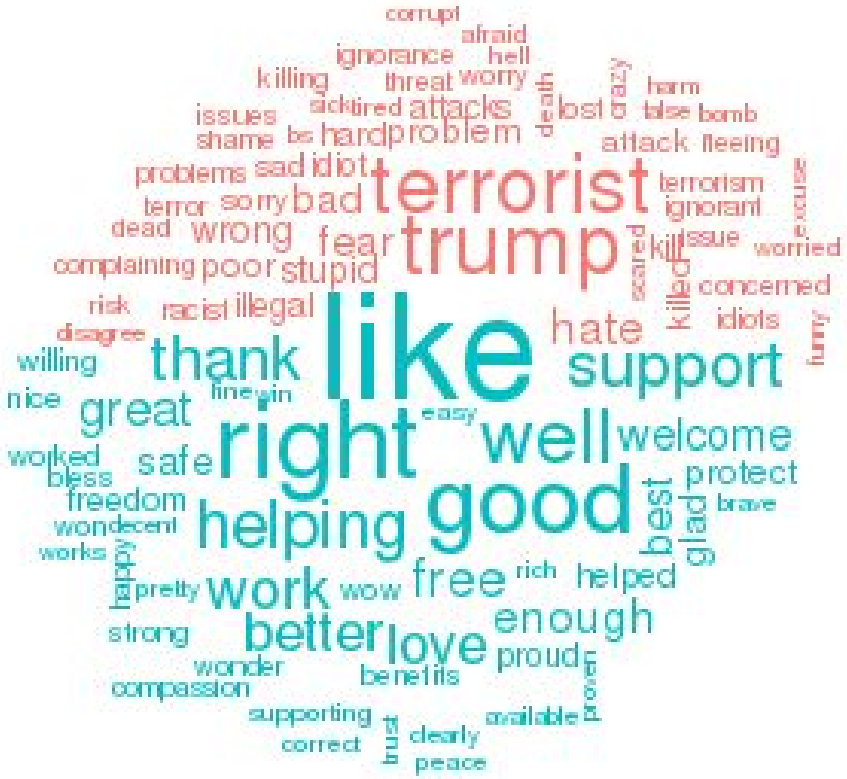




Sentiment Analysis

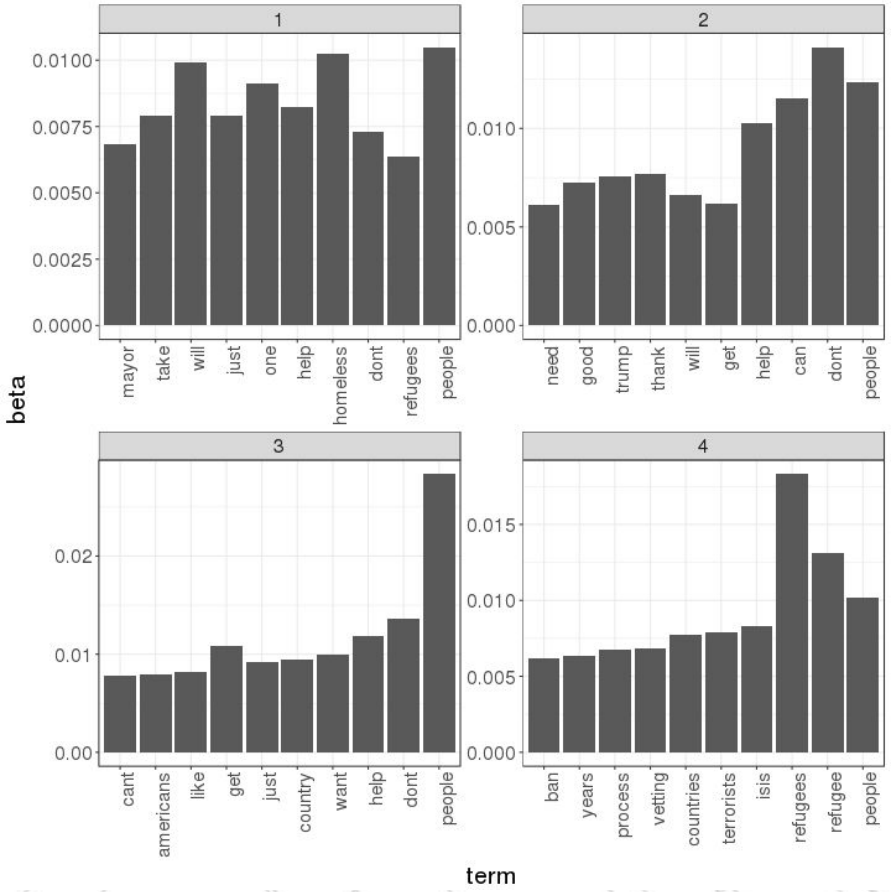
- Sample: "Mayor Bowers was thinking of the welfare and security of all Roanoker's. Also, it appears that while taking the position he is also concerned with what would happen if the group of refugees were possibly infiltrated by a terrorist that, like the Paris attacks, would only be hours away from the Nation's Capitol and seemingly in a nice quiet place where havoc and widespread destruction could be planned. Like the Paris attacks. I, for one, stand and give Mayor Bowers my full support. The current Administration has no idea who these people are and the possibility exists that just one of however many refugees that would settle in Roanoke could be a terrorist. And that is one too many. Mayor Bowers is thinking clearly and should stand his ground. Regardless of what the city council or any celebrities have to say. Mayor Bowers represents the people of Roanoke."

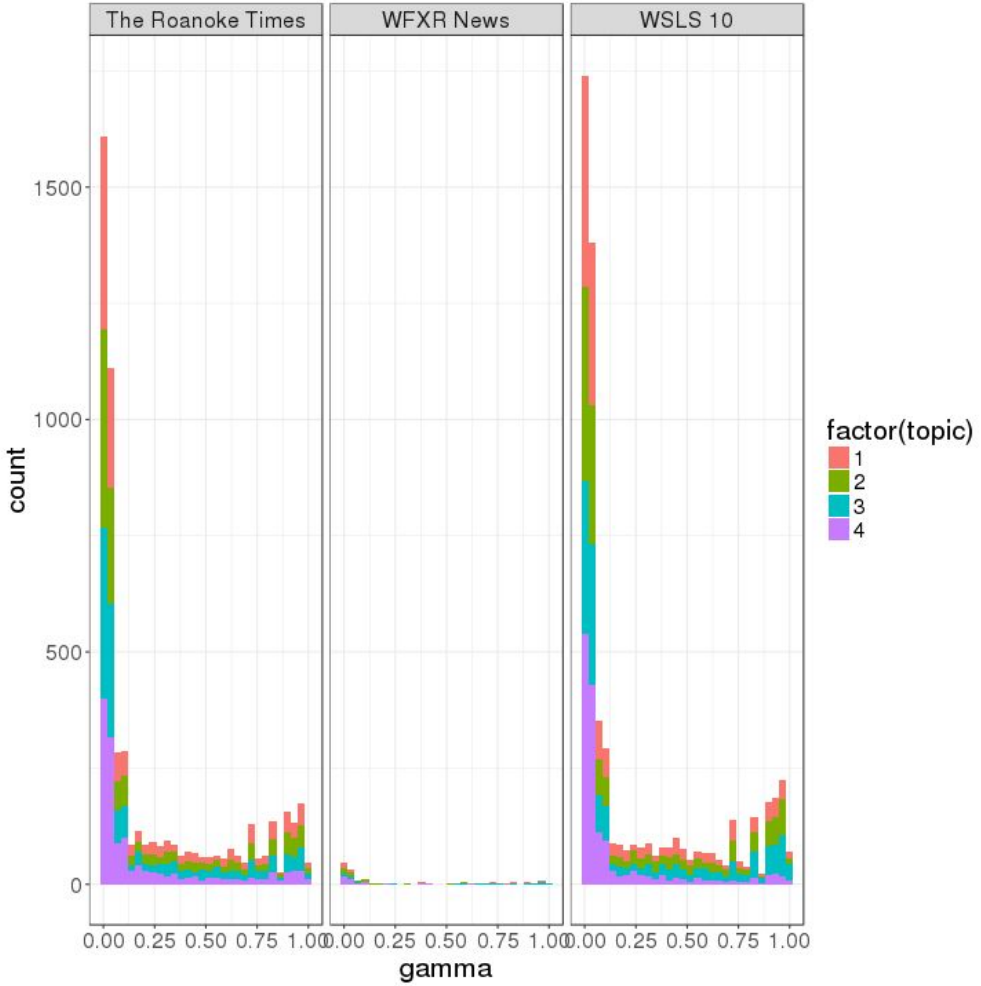




Latent Dirichlet Allocation (LDA)

- LDA: a generative, unsupervised statistical model that allows a set of observations to be explained by groups
- We performed this analysis for the Facebook comments, from our three sources
  - Chose the number of categories to be 4





Next Steps

- We have analyzed sentiment in two ways
  - How can humans validate our findings?
  - Amazon Mechanical Turks to validate our sentiments/categorization
- Formulate a manuscript

New Kent County

DSPG - David Park, Joseph Kim, David Hinkle, Lata Kodali

SDAL - Dave Higdon, Aaron Schroeder, Stephanie Shipp

Sponsor - Virginia Cooperative Extension (VCE)

VirginiaTech Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Project Description

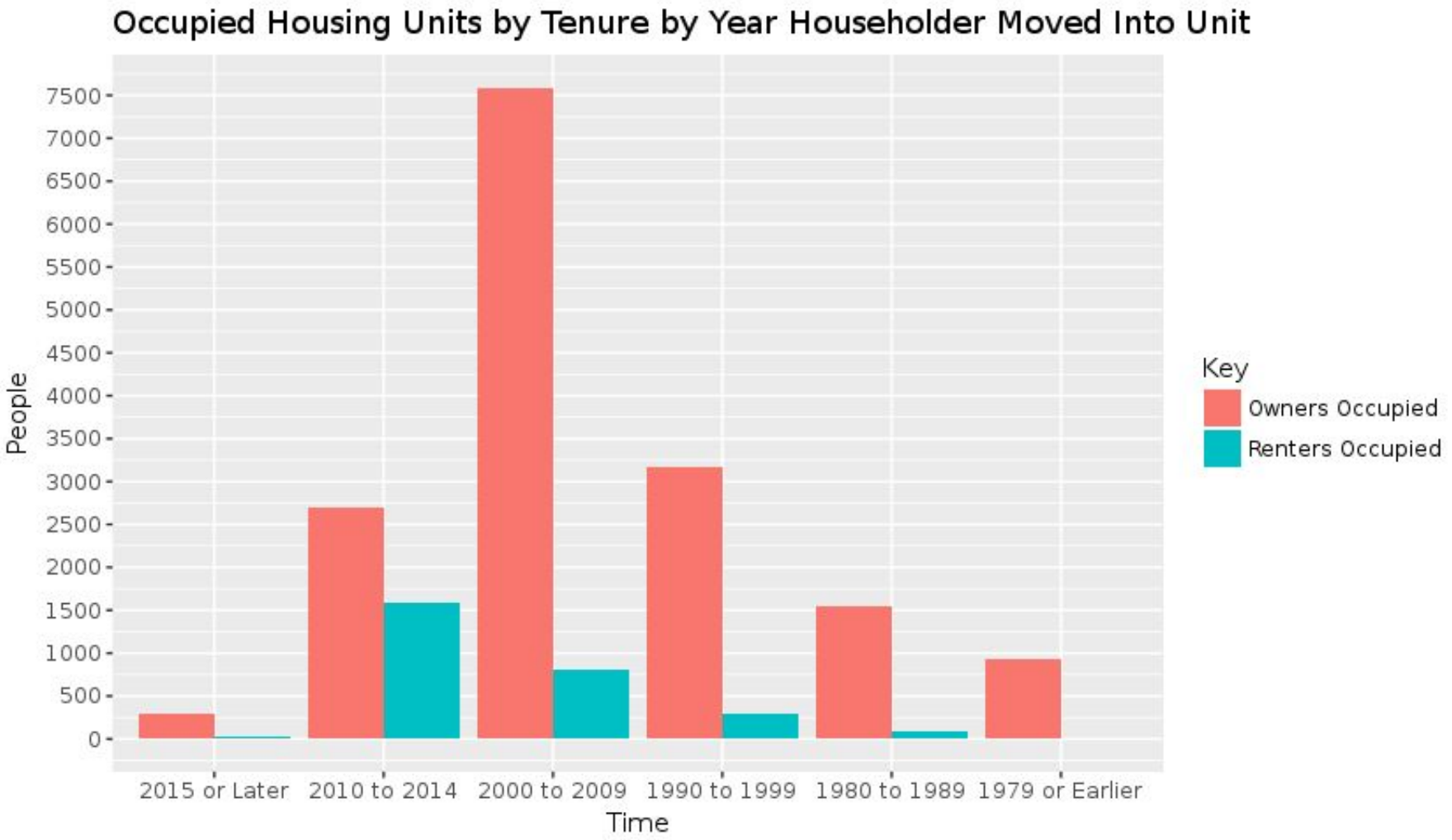
- **VCE Representative:** Carl Frick
- **Issues:**
  - New Kent County is growing, and there is a perceived increase in the number of youth who have nowhere to go after school. These teenagers might be engaging in risky behaviors.
  - Lack of transportation for programs such as 4-H and barriers to enrollment in 4-H.
- **Goal 1:** Find data related to drug use amongst the youth in New Kent. Discover what is causing the change in drug use.
- **Goal 2**: Provide a way to compare New Kent County to other VA counties.

Project Overview

- Basic profiling of New Kent County using:
  - ACS data
  - Housing data
  - Virginia Department of Education (VDOE) data
  - Social media data

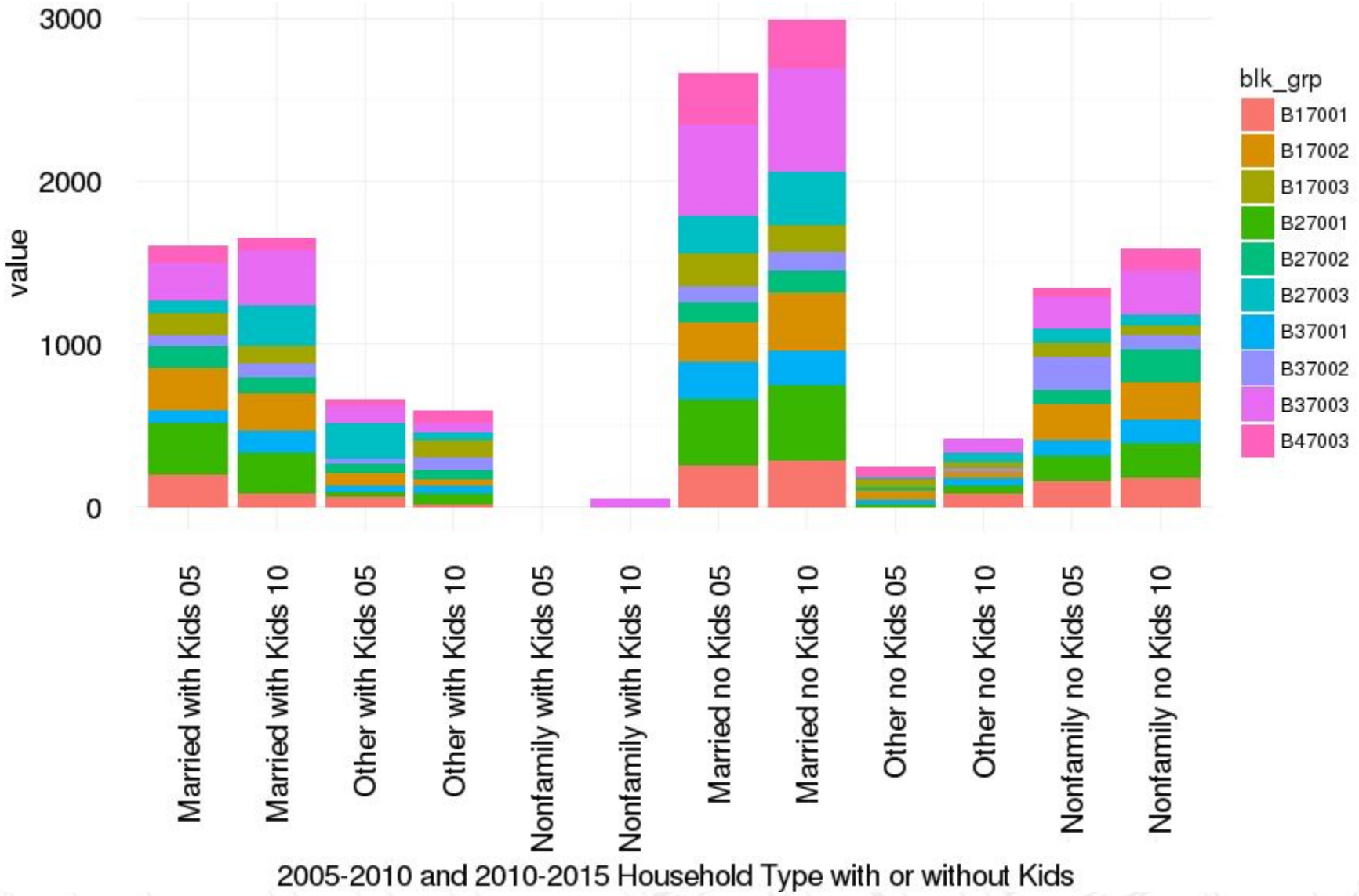
- Plotted the data collected to see trends in New Kent
- Conducted a field study by visiting New Kent and met with Carl (Co-op Extension agent)

ACS New Kent Housing Data



VirginiaTech.
Biocomplexity Institute

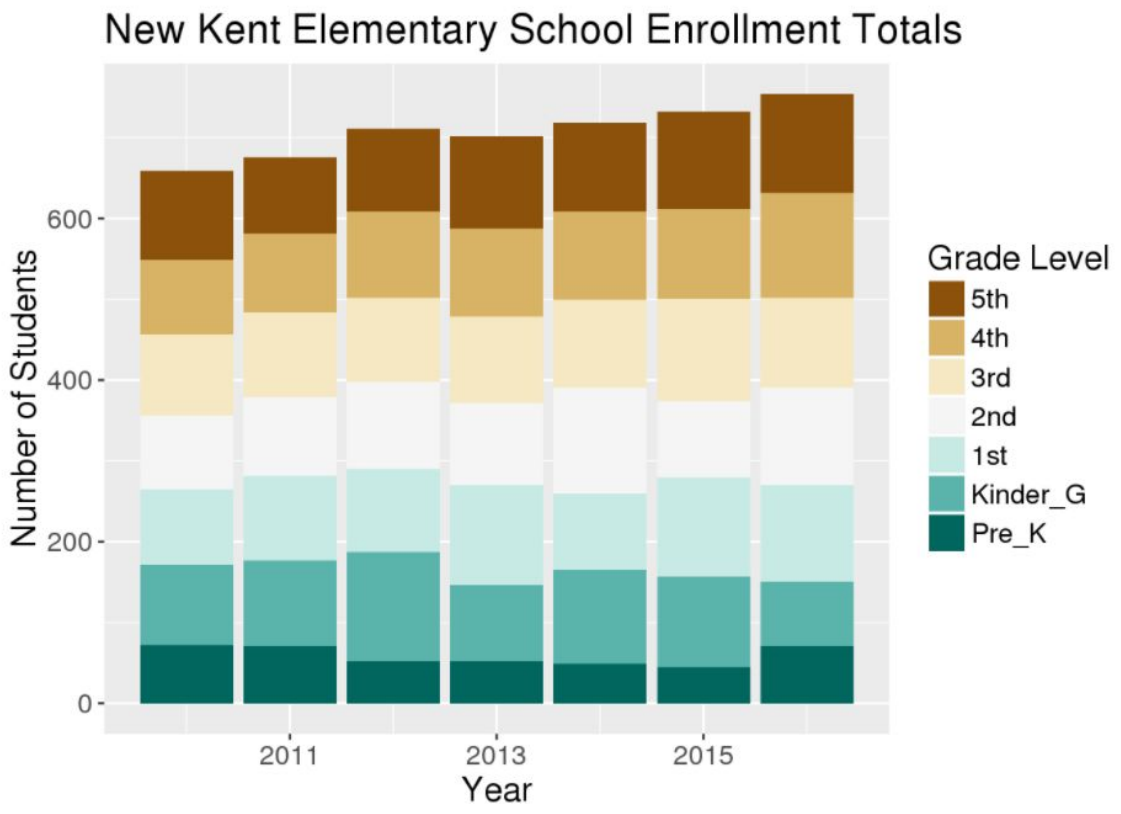
ACS Household Type Profile



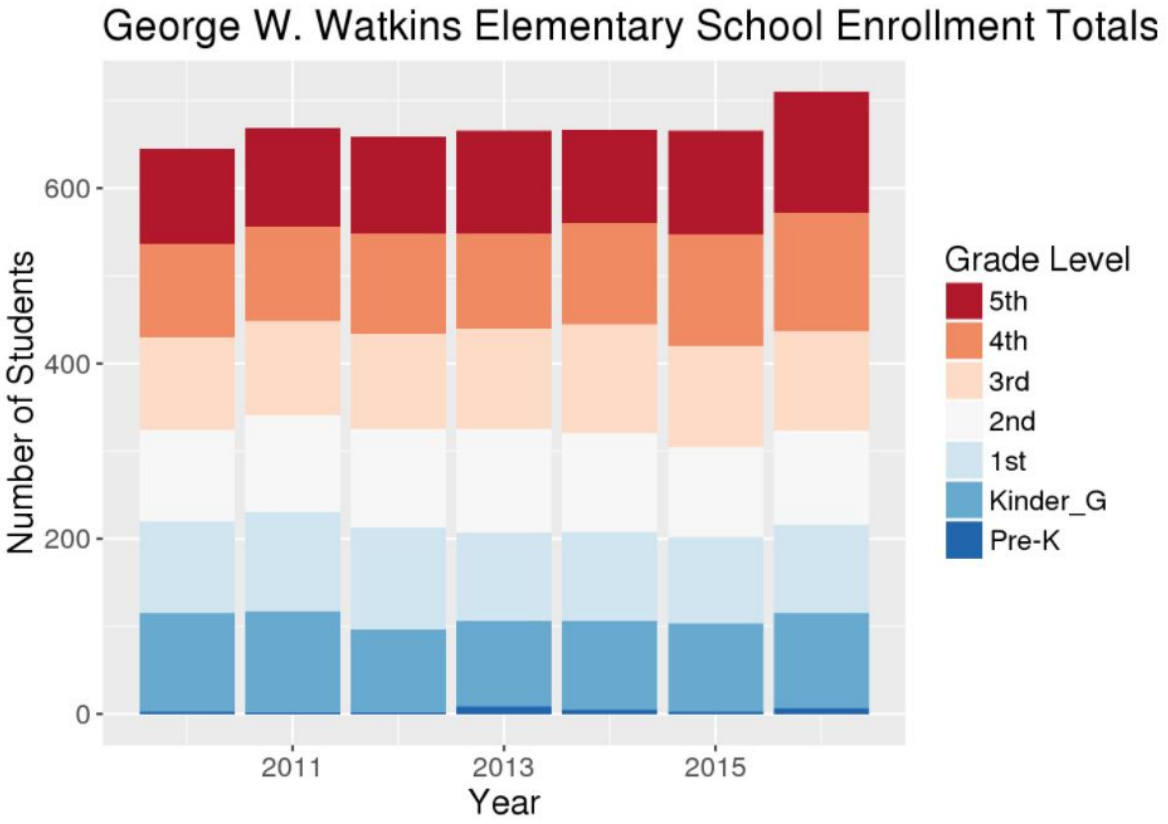
VirginiaTech
Biocomplexity Institute

New Kent School Profiles

- New Kent Elementary

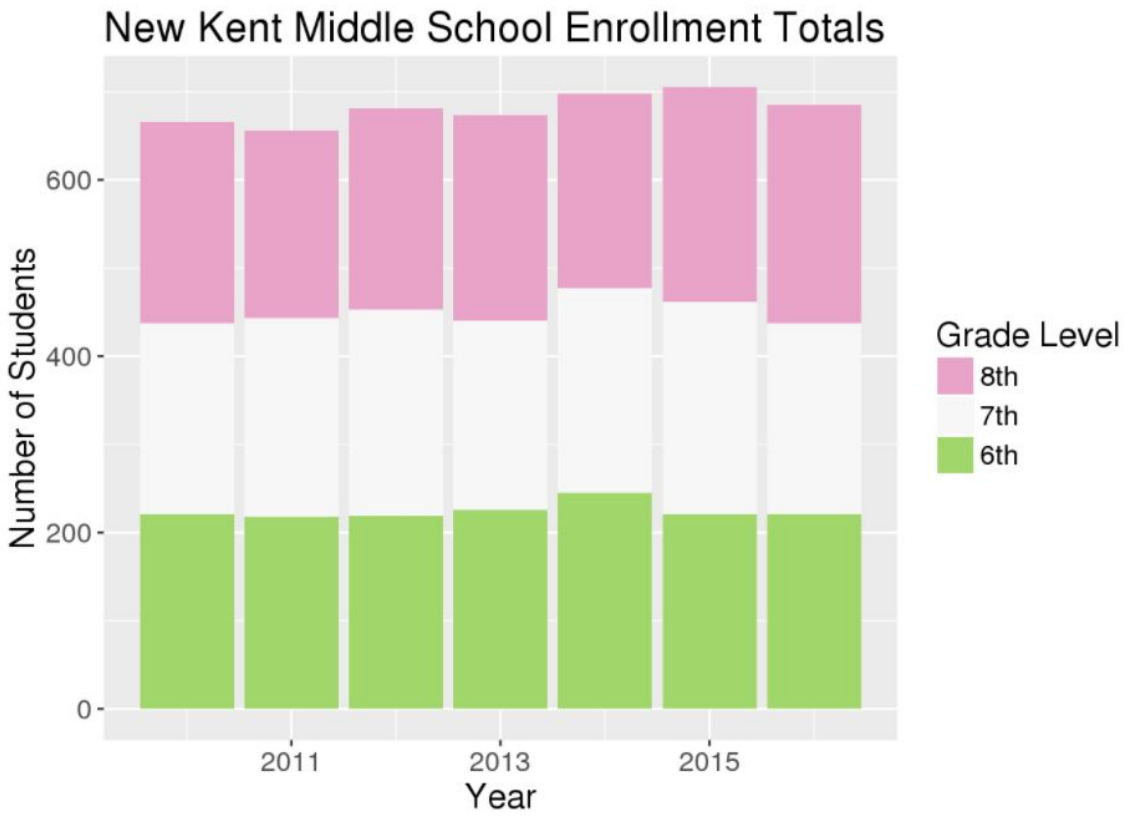


- George W. Watkins

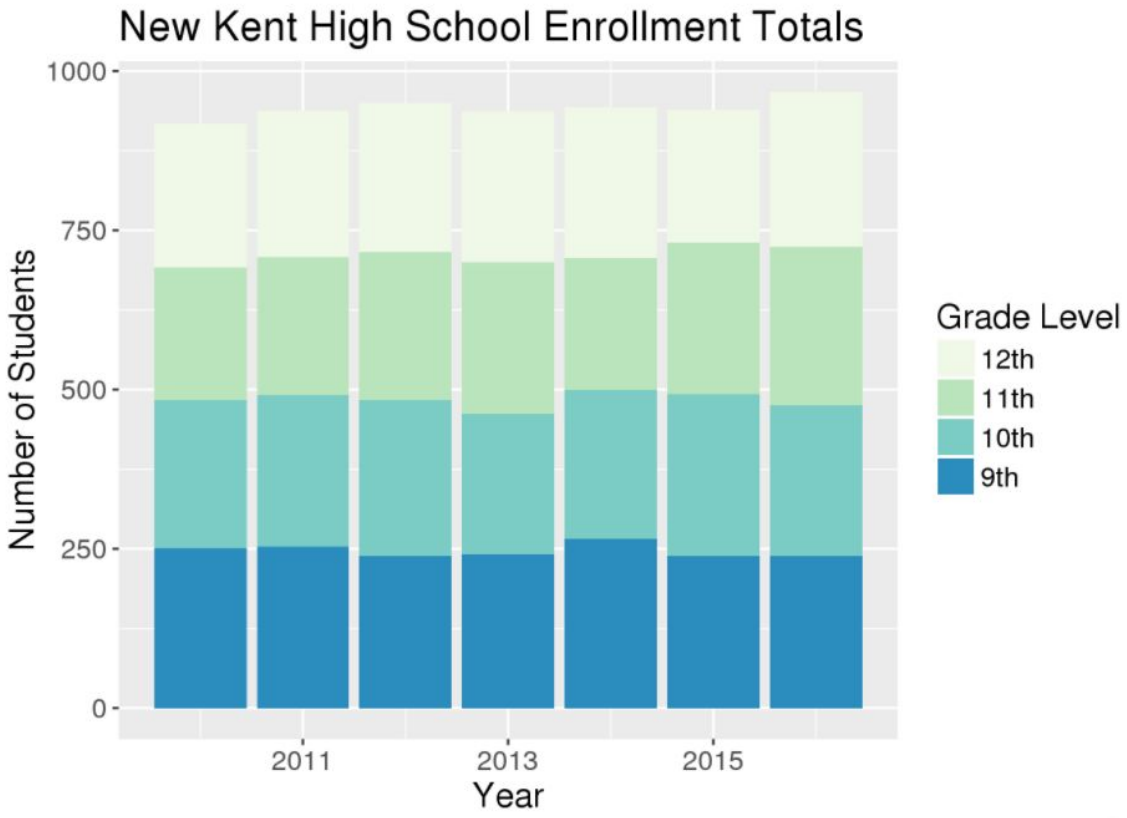


New Kent School Profiles

- New Kent Middle



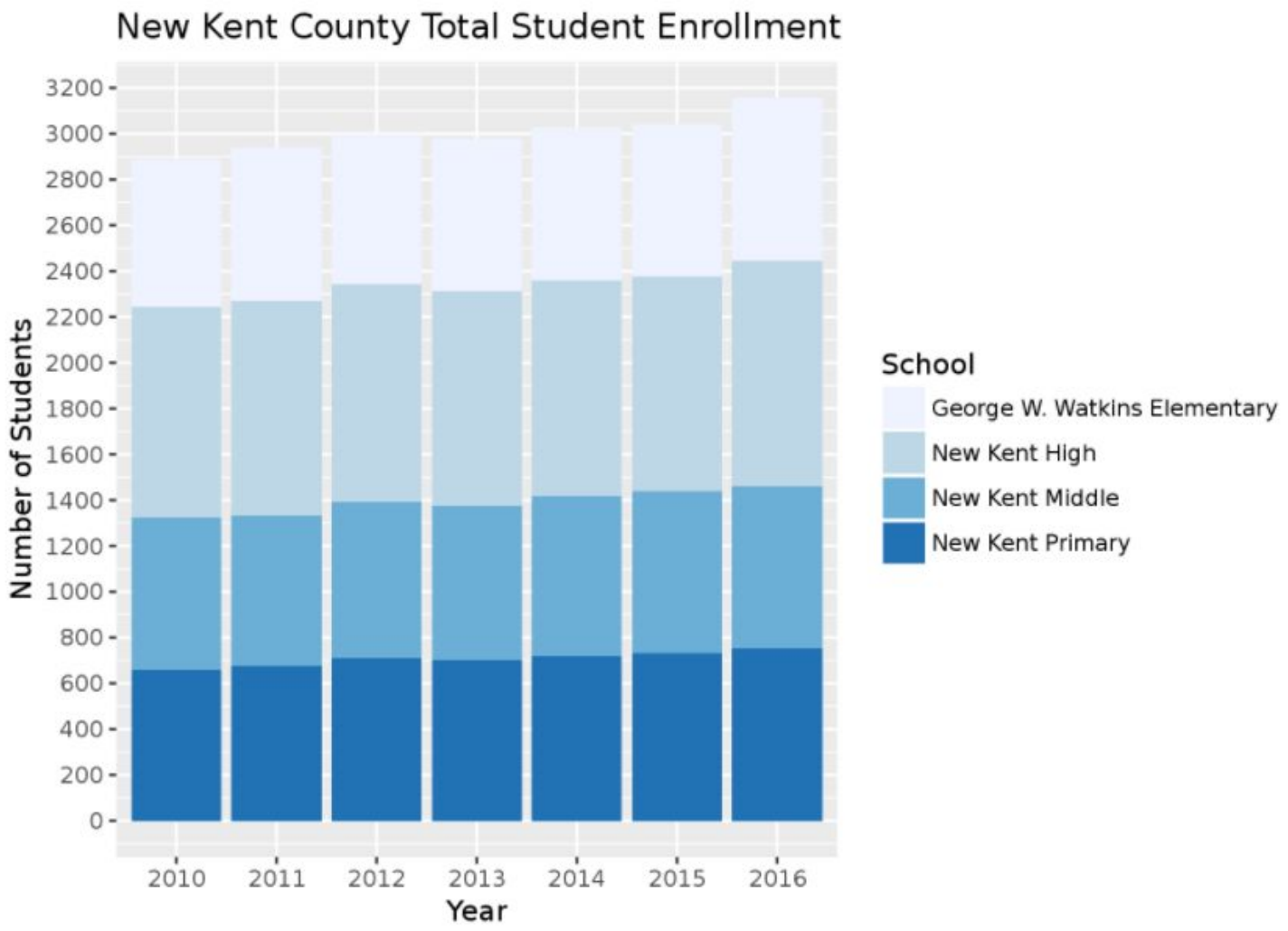
- New Kent High



VirginiaTech.
Biocomplexity Institute

New Kent School Profiles

- Total School Population



New Kent School Profiles

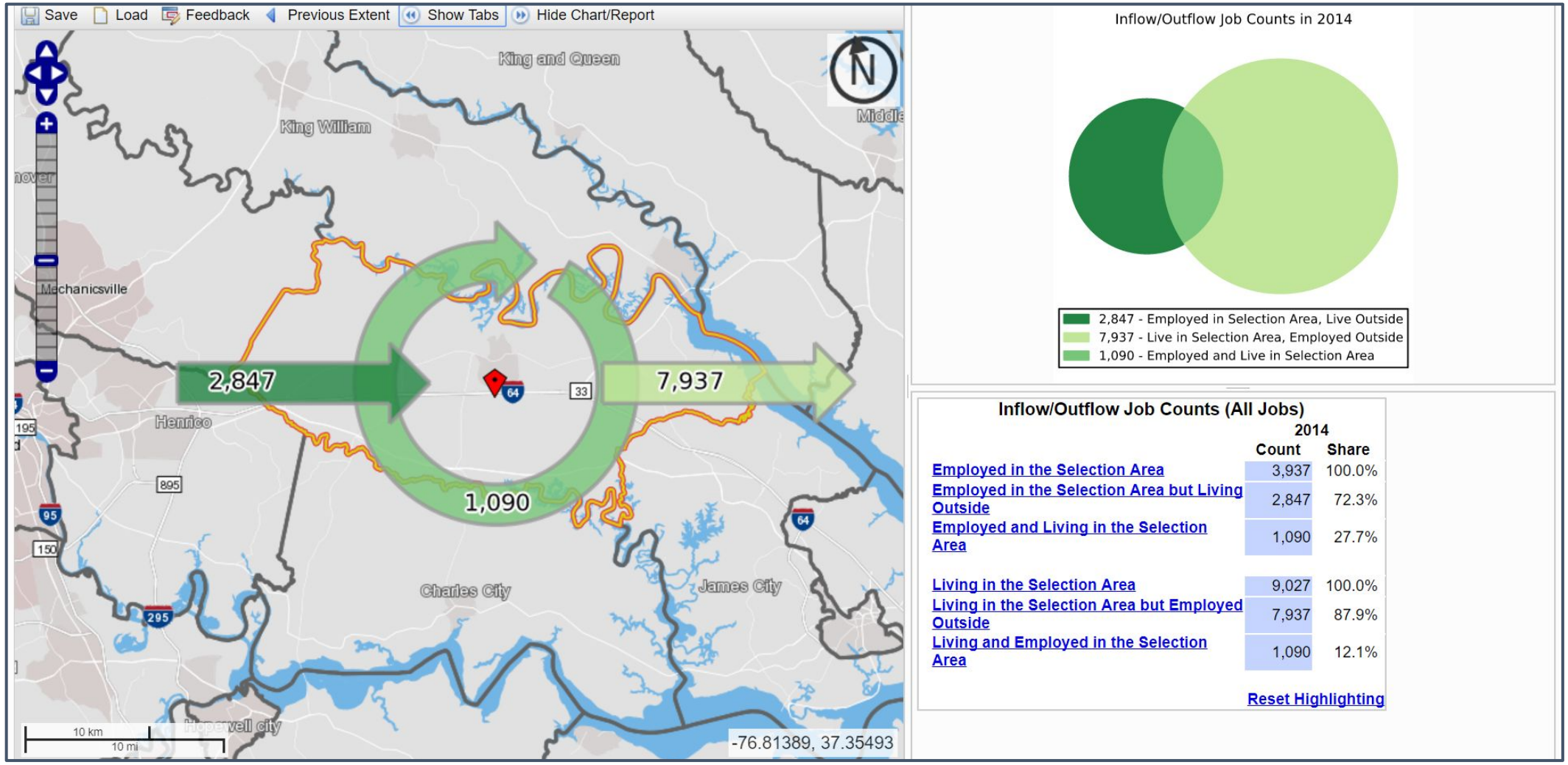
school name	year	total number of students	Hispanic Male	Hispanic Female	Asian Male	Asian Female	Black Male	Black Female	White Male	White Female
New Kent Elementary	2012	711	3.1	2.7	0.7	0.8	8.6	5.2	39.5	34.2
	2013	702	3.4	2.1	0.7	0.7	7.1	6.3	39.6	34.5
	2014	719	3.1	3.1	0.6	0.6	6.8	4.6	39.2	34.4
	2015	732	3.4	3.7	0.7	0.4	5.7	3.6	36.3	36.9
	2016	754	4.2	3.4	0.1	0.4	5.2	3.8	40.3	34.5
George W. Watkins	2012	659	1.1	1.2	0.3	0.5	5.2	3.9	40.7	42
	2013	666	2	1.8	0.8	0.6	4.4	4.1	39.8	42.6
	2014	667	1.8	1.5	0.9	0.4	3.9	3.3	41.8	42.1
	2015	666	2	1.8	0.6	0.3	3.8	4.1	42	42
	2016	710	1.8	1.8	0.4	0.4	3.4	4.8	41	41.5
New Kent Middle	2012	681	2.5	1.2	0.1	0.1	5.7	6.9	43.5	35.2
	2013	673	2.4	1.5	0.3	0.6	6.4	5.8	41.9	35.5
	2014	698	2.7	1.6	0.3	0.7	5.9	4.9	41.8	37.1
	2015	705	1.8	1.6	0.6	0.9	6.2	4	40.9	37.9
	2016	707	2.3	2.3	1	0.6	5.5	4.1	36.6	41.3
New Kent High	2012	950	1.5	1.5	0.4	0.5	6.9	4.6	40	38.1
	2013	936	1.5	1.6	0.4	0.4	6.2	5.3	40.8	36.9
	2014	943	2.1	1.5	0.5	0.4	5.8	5.4	40.5	37.4
	2015	939	2.8	1.7	0.3	0.7	6	5.9	42.2	34.5
	2016	985	2.7	2.1	0.3	0.7	6.6	5.9	42.9	33

4-H Survey Data

- Risk Factors survey answered by 49 Cooperative Extension Agents
- Example Questions
  - What are the biggest issues facing youth in your unit?
    - Poverty/Lack of Money
    - Drugs
  - What are the most significant barriers to 4-H participation in your unit?
    - Transportation
    - Competition with other activities

Source: 4-H Survey provided by Carl Frick.

OnTheMap: Employment In and Outflow



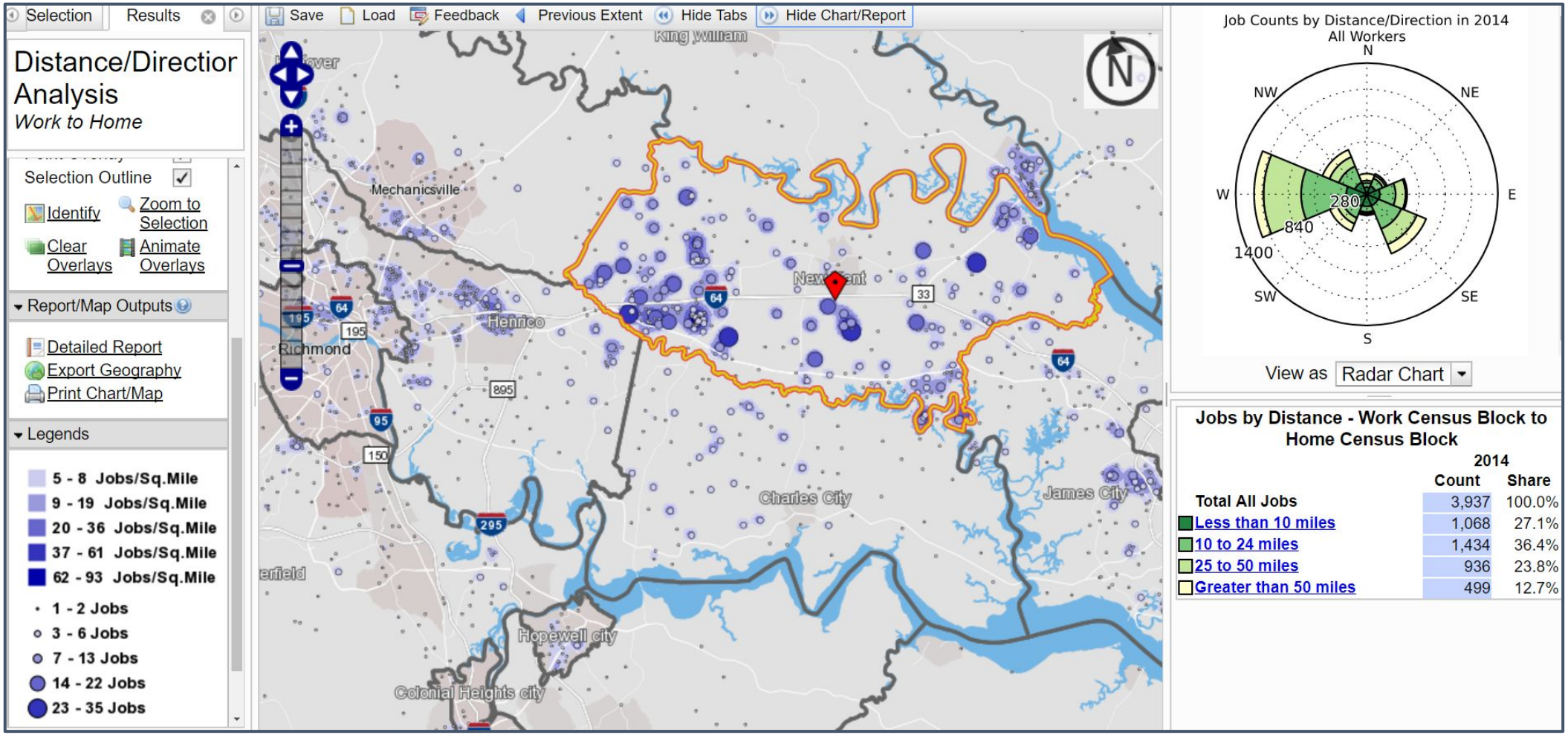
Inflow/Outflow Job Counts (All Jobs)

	2014 Count	2014 Share
Employed in the Selection Area	3,937	100.0%
Employed in the Selection Area but Living Outside	2,847	72.3%
Employed and Living in the Selection Area	1,090	27.7%
Living in the Selection Area	9,027	100.0%
Living in the Selection Area but Employed Outside	7,937	87.9%
Living and Employed in the Selection Area	1,090	12.1%

Reset Highlighting

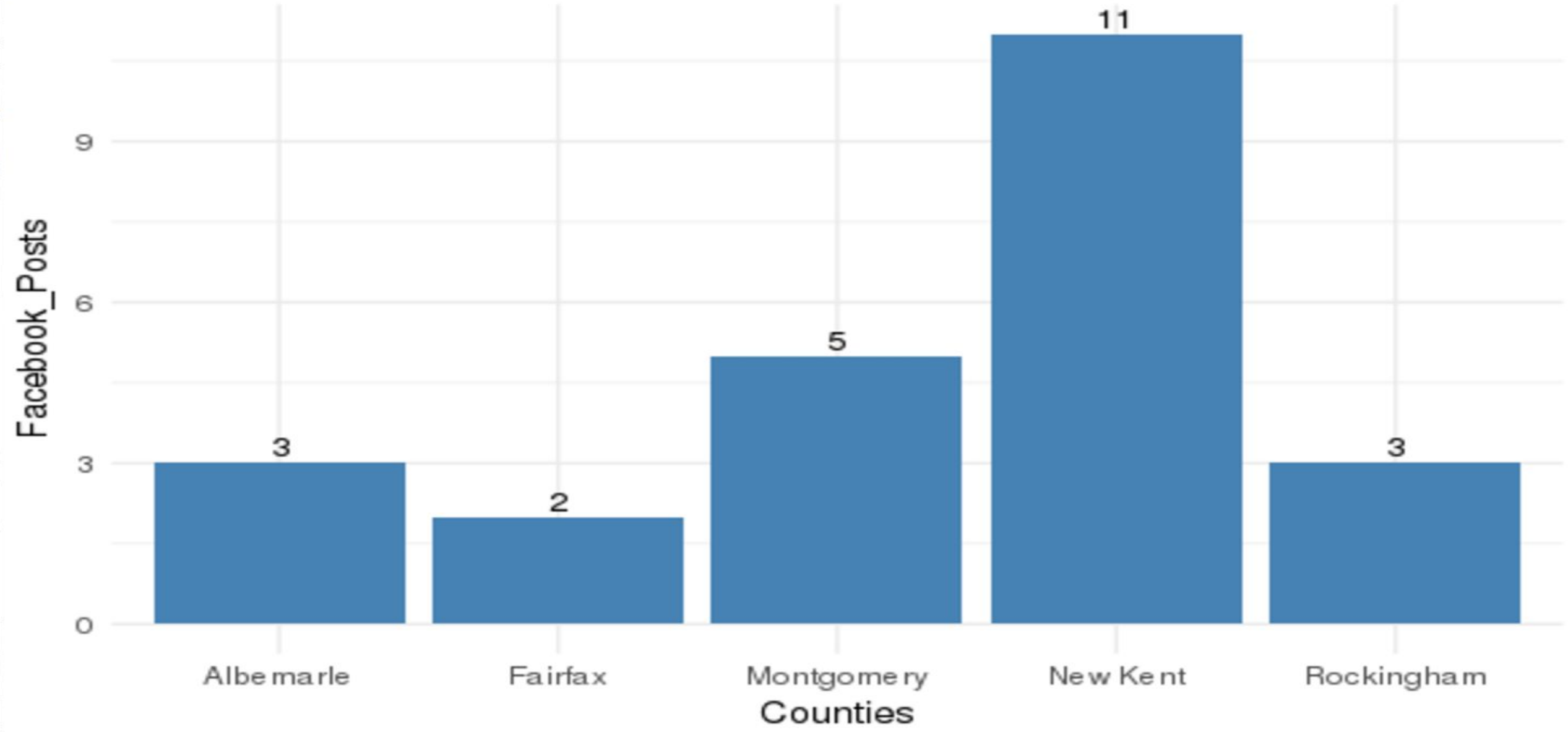
Source: https://onthemap.ces.census.gov/

OnTheMap: Distance to Work



Source: https://onthemap.ces.census.gov/

New Kent Social Media Data



Future Work

- Dashboard:
  - Compare variables across VA counties
  - Provide visuals including map and basic plots
- Model:
  - Variables of Interest vs. Problems in Respective Counties (e.g. drug use, poverty, transportation).
- Challenges:
  - Finding data sources that show trends for New Kent County.
  - Lack of data to analyze for New Kent, e.g. crime data

Questions

Thank you.

Comments?

Suggestions?

Community Projects: Arlington - Needs Assessment

PI : Vicki Lancaster
Team: Madison Arnsbarger, Chanida Lerdritsomboon, Harsimrat Pandher (Simmu), Sean Pill
Sponsor: APAH (Arlington Partnership for Affordable Housing)

VirginiaTech
Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Project Overview

Our Sponsor- APAH (Arlington Partnership for Affordable Housing)

- Identify service gaps and mobilize citizens to close these gaps in zip code 22204.
  - Preschool and Childcare
  - Workforce
- They are funded by the Citibank Community Progress Makers Fund and must report to their donors.
- Plan to host a workshop for APAH and the Community Progress Makers roster where we show them how to find the information they need to conduct a needs-based assessment of zip-code 22204 independently, since most of it is readily available.

Background: Snapshot of Zipcode 22204:

Report Area Data Taken from Community Commons	Percent Free/Reduced Price Lunch Eligible	Percent Families with Income Over $75,000	Percent Population Age 25+ with No High School Diploma	Percent Population Under Age 18 in Poverty
22204 Estimates	52.38%	54.59%	13.55%	***22.3%***
Arlington County, VA	30.91%	73.82%	6.52%	***11.37%***

Population Below the Poverty Level, Children (Age 0-17), Percent by Tract, ACS 2011-15

- Over 30.0%
- 22.6 - 30.0%
- 15.1 - 22.5%
- Under 15.1%
- No Population Age 0-17 Reported
- No Data or Data Suppressed

- Report Area

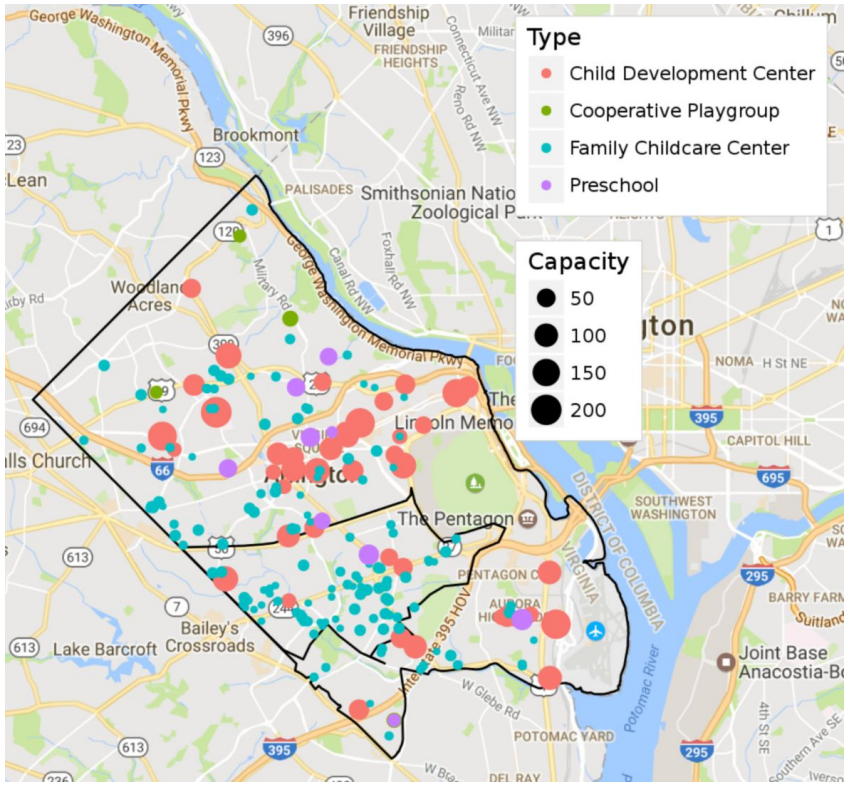
Background: The story of Zipcode 22204

- **Socioeconomic/Demographic Factors**
  - Community Commons: Community Health Needs Assessment Report:
    - Compares 22204 data to the rest of Arlington.
- **Quality of Public Transit**
  - AllTransit Score: Combines information regarding "can I get transit" with "Jobs accessible in a 30 Minute Transit ride and combines them in a way that is reflective of what fraction of people use transit for their journey to work. Ranked 0-10 and ***available at block-group level***
- **Social Vulnerability**
  - CDC: SVI Index, a 0-100 index, ***ranked by Census Tract***, uses 15 ACS variables to measure a Tract's relative Vulnerability to a disaster

Child Care Availability in 22204

Types of Child Care Services:

- Child Development Centers
- Family Child care Centers
- Preschools
- Church-Based Child care



Data Sources:

- Arlington Country Open Data
- Childcarecenter.us
- Carelulu.com
- Daycare.com
- American Community Survey 2010 - 2015

Childcare for children ages 0-5 in 22204:

- **Assumption: Children who are not in child care services utilize family and friends for care.**

Total Number of Children	3,418
Children in Child Care Services	1,059 (31%)
NOT in Child Care Services	2,359 (69%)

Childcare Affordability in 22204

"The US Department of Health and Human Services has established that in order to be considered affordable, families would spend no more than 10% of their gross income on child care."

- Arlington County 2018 Proposed Budget

	Median Income of Families with Own Children in Home ($, 2015)	Annual Cost of Childcare for an infant and 4-year-old ($)	Cost of Childcare as Percentage of Income
22204	65,037	28,350 - 29,484	43.6 - 45.5%
Arlington	148,629	40,521*	27.3%
Virginia	79,400	21,476	27.0%

* annual cost for an infant and 4-year-old in Washington, D.C. (DC Office of the State Superintendent of Education, May 2014)

Barriers for using existing services

- Physical space needs and constraints
  - Childcare providers prefer to be located on the first floor of a building
  - Must provide outdoor element for children's play and activity
- Costs
  - Rent is expensive
  - Cost of services is overall expensive
- Regulatory requirements
  - Reduction in capacity of programs
  - Parking
- Perceptions
  - Negative perception of child care uses within office buildings

Average weekly rate for 12 child care providers:

Age	Cost
Infant	$300.58
Toddlers	$300.58
Age 2-3	$291.26
Age 3-4	$285.30
Age 4-5	$267.45

Source: Arlington County 2018 Proposed Budget

Home Care Providers Hours and Summer Child Care Services Available

Home Care Providers Hours:

- Ranges from 6:00am- 5:30pm

Arlington 2017 Summer Camps Catalog:

- 23 camp locations within zip code 22204
- Ages range between: 3-13 year olds
- Fees are different depending on the type of activities provided by the camp
    - Ranges from $120-$160 weekly

Data Sources:

- Arlington Country Open Data
- childcarecenter.us
- carelulu.com
- Daycare.com
- County Manager's Message from FY 2018 Proposed budget page 41

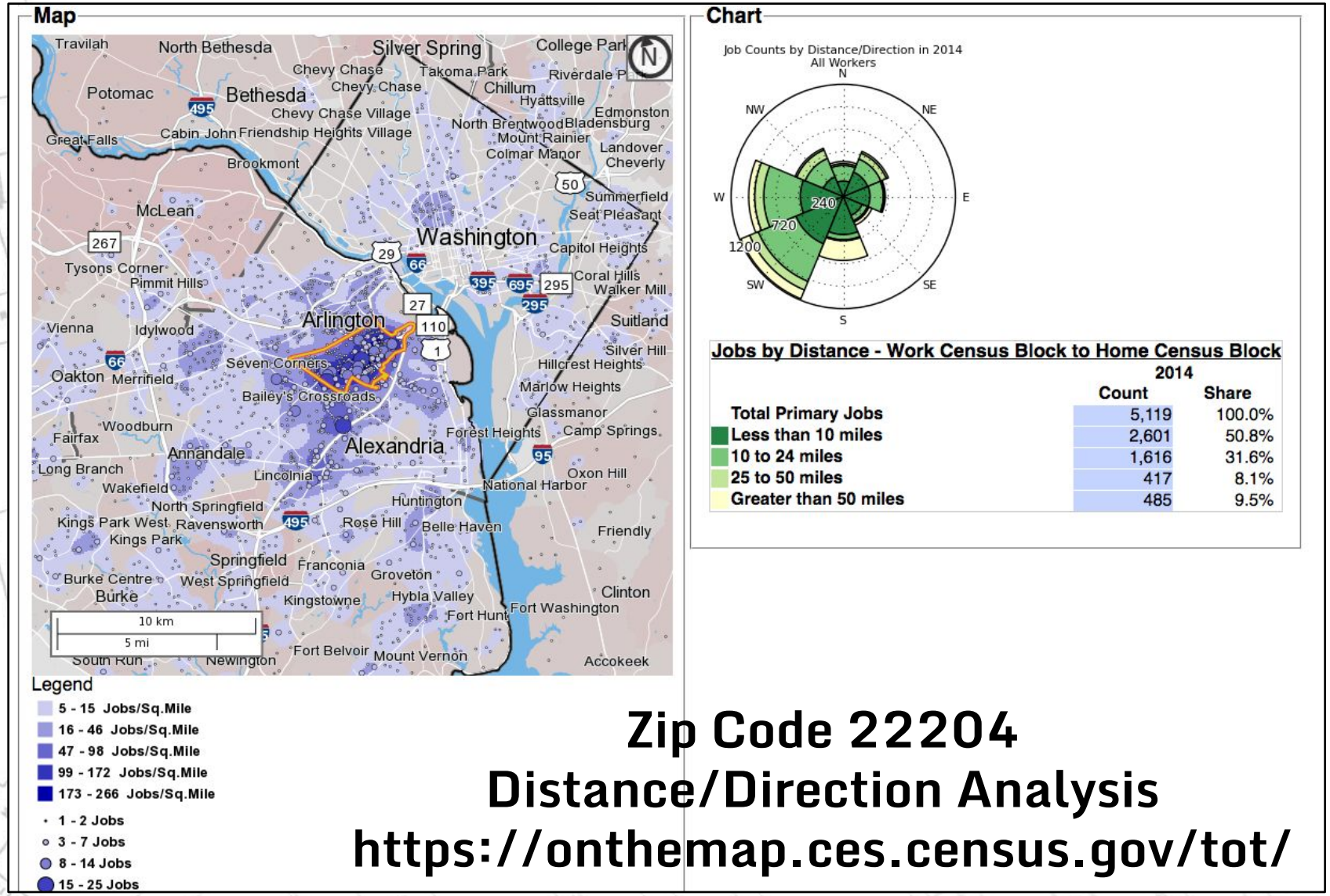
Potential additional services:

- **Look to other cities** (Richmond, VA)
  - **Office of Community Wealth Building** "serves as the collective impact hub for an informed, energized, and aligned community of non-profits, ministries, government agencies, funders, businesses and people - Community Wealth Builders."
  - **Send-A-Kid-To-Camp Campaign**
  - **Transitional Daycare Subsidy:** provides childcare assistance for 12 months after TANF benefits end (for income-eligible families).
- **Incorporate affordable childcare facilities in first floor of APAH buildings**
  - Convenient locations for low income families
- **Create a survey for residents**

Workforce

1. Where do people work?
2. What is the self-employment rate?
3. What benefits and packages are available to individuals?
4. What type of training, if any, would they need to make a career advancement? What are the barriers to attending the training?
5. What are the commutes like?
6. Do people use public or private transit?

What are the commutes like?



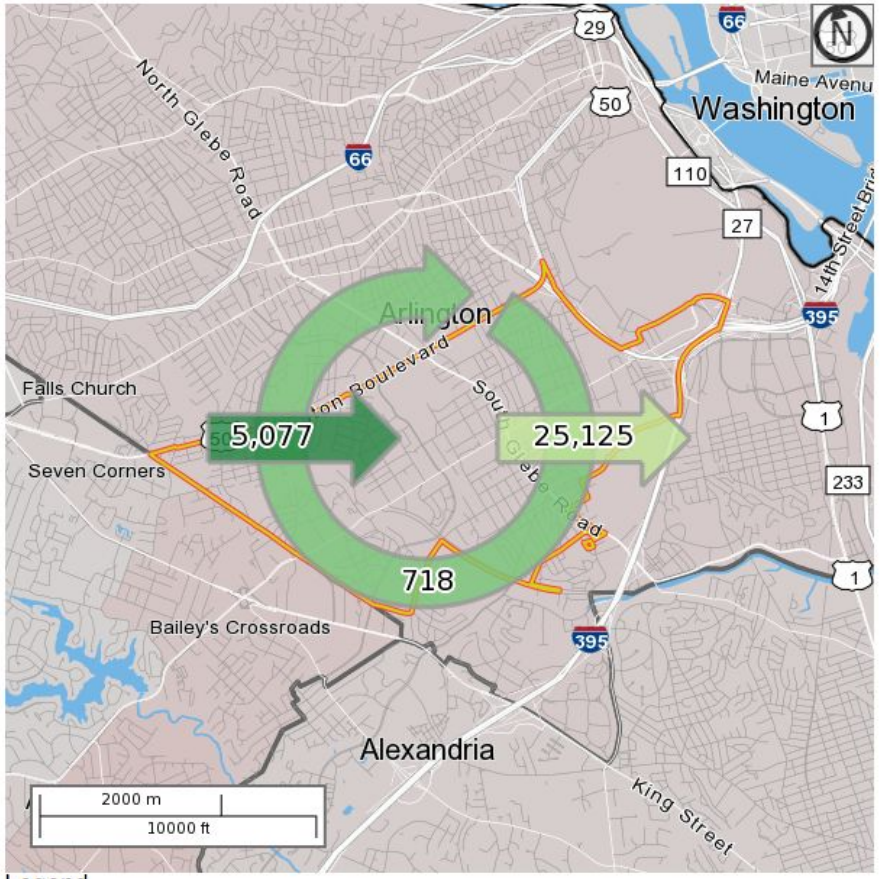
Jobs by Distance - Work Census Block to Home Census Block

	2014 Count	2014 Share
Total Primary Jobs	5,119	100.0%
Less than 10 miles	2,601	50.8%
10 to 24 miles	1,616	31.6%
25 to 50 miles	417	8.1%
Greater than 50 miles	485	9.5%

Zip Code 22204
Distance/Direction Analysis
https://onthemap.ces.census.gov/tot/

OnTheMap (V6.5)
Web-based mapping and reporting application that shows where workers are employed and where they live. Based on the a partnership with the U.S. Census Bureau and Local Employment Dynamics (LED) partner states.

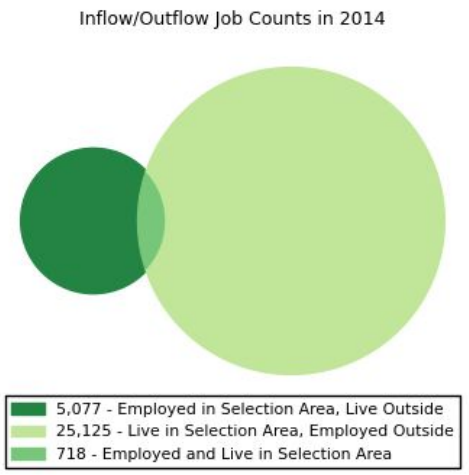
Where do people work?



Legend

Note: Overlay arrows do not indicate directionality of worker flow between home and employment locations.

- Employed and Live in Selection Area
- Employed in Selection Area, Live Outside
- Live in Selection Area, Employed Outside
- Analysis Selection



Inflow/Outflow Job Counts (All Jobs)

	2014 Count	2014 Share
Employed in the Selection Area	5,795	100.0%
Employed in the Selection Area but Living Outside	5,077	87.6%
Employed and Living in the Selection Area	718	12.4%
Living in the Selection Area	25,843	100.0%
Living in the Selection Area but Employed Outside	25,125	97.2%
Living and Employed in the Selection Area	718	2.8%

Reset Highlighting

Zip Code 22204
Inflow/Outflow Analysis
https://onthemap.ces.census.gov/tot/

Data Sources
Longitudinal Employer-Household Dynamics (LEHD) which has several data products that characterize workforce dynamics, such as the LEHD Origin-Destination Employment Statistics (LODES).

What is the self-employment rate?

SOI Tax Stats - Individual Income Tax Statistics - 2014 ZIP Code Data (SOI)

Return to Tax Stats home page

ZIP Code Data - 2014

Data files are available for download in Comma Separated Values files (.csv file extension) where data items are separated by commas and each new line represents a new row, and Excel files by State (.xls file extension).

Documentation

ZIP Code Data Users Guide and Record Layouts (Word Document)

ZIP Code Data

.csv Files

2014 (all States, includes AGI) 2014 (all States, does not include AGI)

Zip File (contains the Documentation, the .csv files, and all State files in Excel format)

2014

ZIP Code Data by Individual State

VIRGINIA

Individual Income Tax Returns: Selected Income and Tax Items by State, ZIP Code, and Size of Adjusted Gross Income, Tax Year 2014

[Money amounts are in thousands of dollars]

Release Date of 2015

Returns

08/31/2017	Individuals, Zip Code Data	This study provides detailed tabulations of individual income tax return data at the State and ZIP code level. Data for the Tax Year 2013 study are from returns filed through December 2014.	2015

ZIP code [1]	Size of adjusted gross income	Self-employment tax	
		Number of returns	Amount
		(92)	(93)
22204	$1 under $25,000	2,000	2,735
22204	$25,000 under $50,000	780	1,723
22204	$50,000 under $75,000	430	1,222
22204	$75,000 under $100,000	300	1,004
22204	$100,000 under $200,000	540	2,278
22204	$200,000 or more	280	2,996

Self-employment Rate =
(Total Self-Employed 22204)/(Total Employed 22204) =
4330/27500 = **0.16**

https://www.irs.gov/uac/soi-tax-stats-individual-income-tax-statistics-2014-zip-code-data-soi

What benefits and packages are available?

SOI Tax Stats - Individual Income Tax Statistics - 2014 ZIP Code Data (SOI)

Return to Tax Stats home page

ZIP Code Data - 2014

Data files are available for download in Comma Separated Values files (.csv file extension) where data items are separated by commas and each new line represents a new row, and Excel files by State (.xls file extension).

Documentation

ZIP Code Data Users Guide and Record Layouts (Word Document)

ZIP Code Data

.csv Files

2014 (all States, includes AGI) 2014 (all States, does not include AGI)

Zip File (contains the Documentation, the .csv files, and all State files in Excel format)

2014

ZIP Code Data by Individual State

VIRGINIA

Individual Income Tax Returns: Selected Income and Tax Items by State, ZIP Code, and Size of Adjusted Gross Income, Tax Year 2014

[Money amounts are in thousands of dollars]

Individual Retirement Accounts

ZIP code [1]	Size of adjusted gross income	IRA payments	
		Number of returns	Amount
		(47)	(48)
22204		430	1,580
22204	$1 under $25,000	50	152
22204	$25,000 under $50,000	100	330
22204	$50,000 under $75,000	130	333
22204	$75,000 under $100,000	40	200
22204	$100,000 under $200,000	110	565
22204	$200,000 or more	**	**

Retirement Savings Contributions Credits,
Pensions and Annuities in AGI

https://www.irs.gov/uac/soi-tax-stats-individual-income-tax-statistics-2014-zip-code-data-soi

Next Steps

- Prepare data workshop for APAH and their partners
  - Share data sources
  - Explain readily available search and analysis tools
  - Walk through less-intuitive data collection techniques
- Make recommendations to APAH to address service gaps in 22204

ARI - Army Base Profiling

Team: Zarni Htet, Adrienne Rogers, Joseph Kim, David Hinkle, David Park
SDAL: Josh Goldstein, Stephanie Shipp
Sponsor: The US Army Research Institute for Behavioral and Social Science Research

VirginiaTech
Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Project Description

Sponsors: The US Army Research Institute for Behavioral and Social Science Research

ARI representatives: Greg Ruark and Andrew Slaughter

Project Title: Towards an Integrated Data Framework for Understanding the Context of Army Environments: A Case Study of Attrition of Army Soldiers

DSPG Project Goals: Do Soldiers leave the Army because of better opportunities in civilian life?

- Identify publically available data sources (e.g., Census and BLS data) to create social, demographic, economic, and other quantitative profiles of Army bases and their surrounding areas

Process Overview

Data Exploration:

- ACS data
  - Ex: Educational Attainment and Marital Status
- BLS data
  - Ex: JOLTS*, OES*, CES*, and CPS*

Exploratory Analysis: Uses visual analysis to understand demographics

1. Location of the fort
2. Bar Graph of side by side comparison of the fort and the surrounding areas
3. An aggregated form of the Bar Graph
4. ACS data plotted into a geographic map

National Analysis: Creating a user friendly dashboard for visuals, comparisons, and to allow for automated approach to produce each profile

*Acronyms - JOLTS- Job Opportunity and Labor Turnover, OES- Occupational Employment Statistics, CES- Current Employment Statistics, CPS- Current Population Survey

Data Discovery

U.S Census Bureau, American Community Survey (ACS): Gathered all tables of interest from ACS. Pulled military and demographic variables for future use

Bureau of Labor Statistics (BLS): Gathered relevant BLS data and also wrote a brief description of each survey

Base Names / ZIP Codes: Manually compiled a list to match zip codes and their bases.

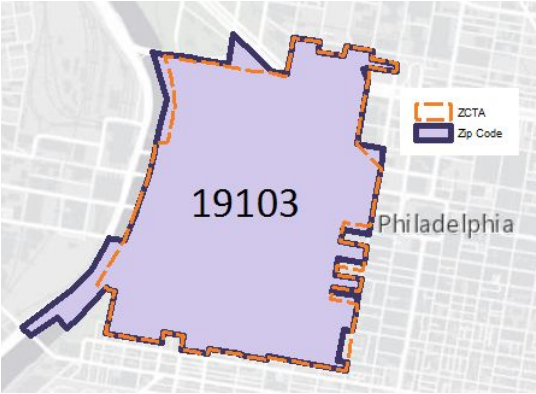
Zip Code v. ZIP Code Tabulated Area

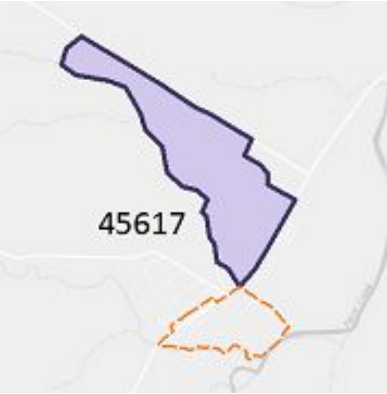
Zip Code: a five-digit number representing a USPS delivery area

- Can lead to over counting

*ZIP Code Tabulated Area (ZCTA): a five digit number representing a generalized area of the USPS Zip Code service area. **NOT** the same as Zip Codes*

- A ZCTA does not cross county lines
- A ZCTA is based on population





https://www.policymap.com/2016/03/what-are-zip-code-tabulation-areas/

Exploratory Analysis

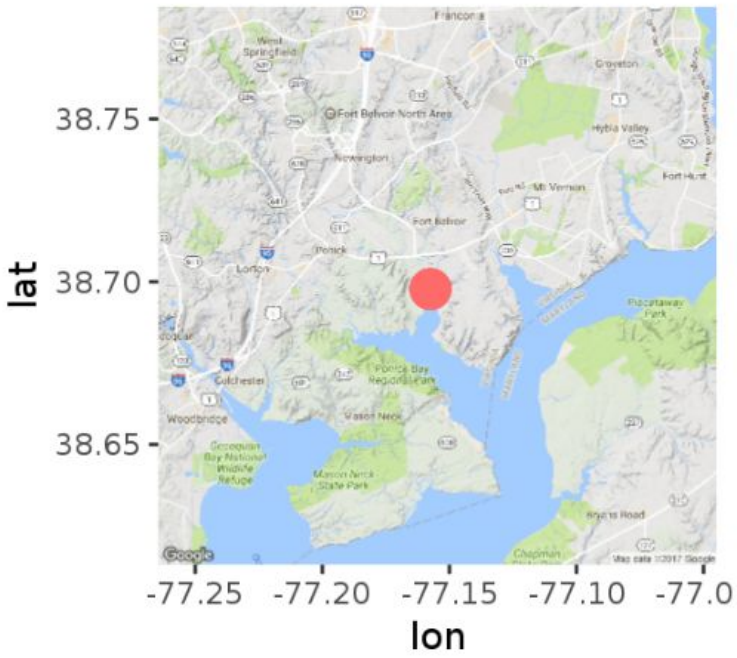
- Goal: Use ACS data to understand the social and economic makeup of Fort compared to surrounding area
- Targeted ACS Variables: Educational Attainment, Children Under 18, Poverty Line Status, Transportation to Work, Marital Status, Sex, Race, Ethnicity, Government Assistance, Veteran Status
- Process:
  1. Collect Zip Codes and convert to ZCTAs
  2. Gather data from ACS: Zip Matcher function
  3. Create and automate plots: Function making training
  4. Run RMD file to create HTML documents

Ft. Belvoir Overview

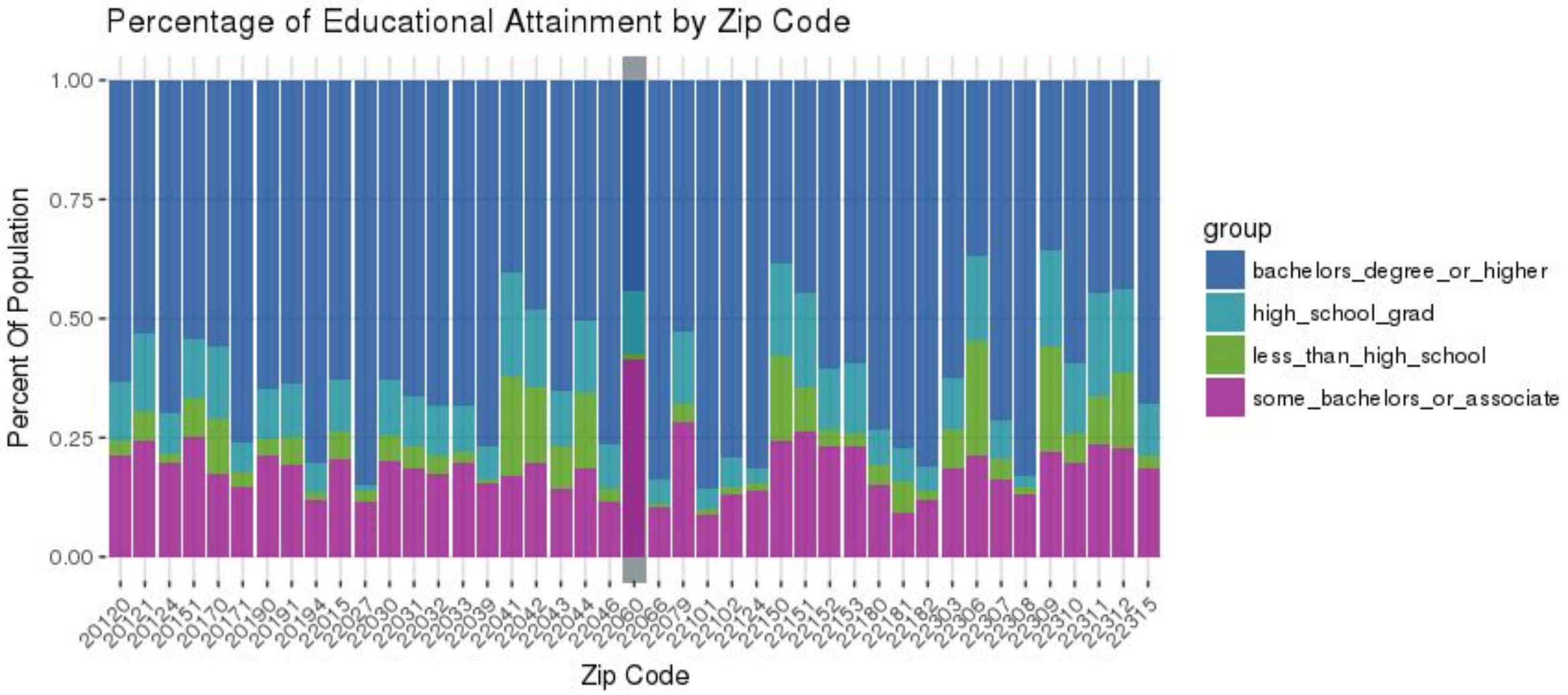
Compared to Fairfax County, Va., households that live on Fort Belvoir

- **more** people have a High School degree or higher but **less** people have a Bachelor's degree or higher
- **more** people have children under the age of 18 in their household
- percentage of population above poverty line is **equivalent**
- **more** workers drive alone to work
- population is **more** predominantly male
- Percentage of population that is Black/African American is **twice** as large

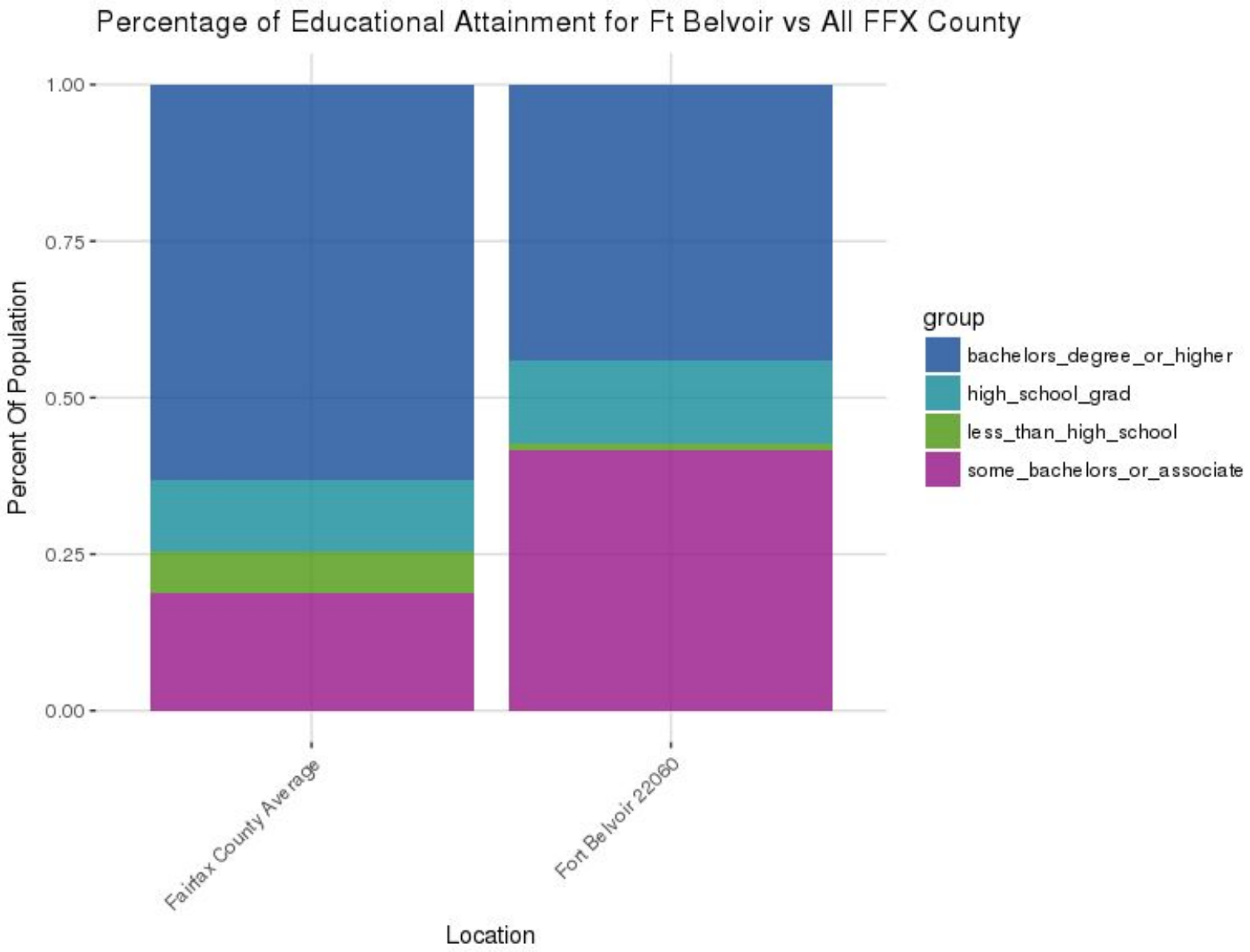
Location(s): Fort Belvoir; Fairfax County, VA



Exploratory Analysis

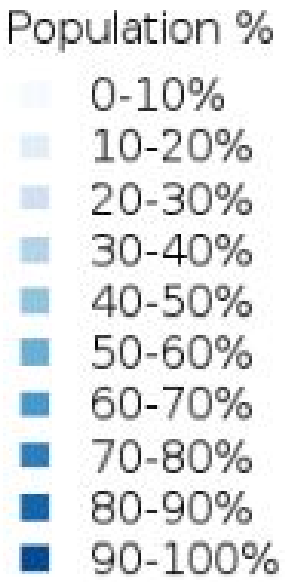


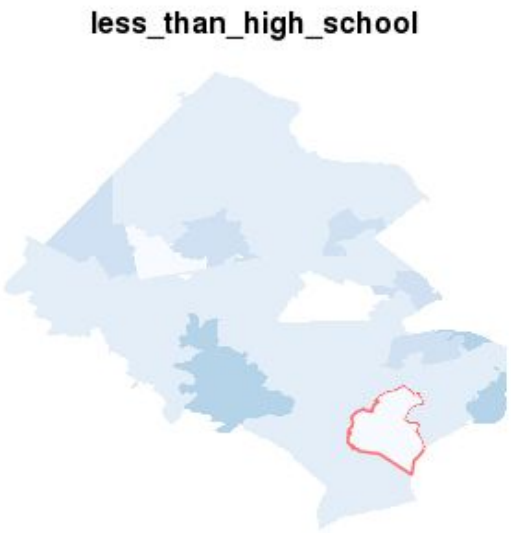
Exploratory Analysis

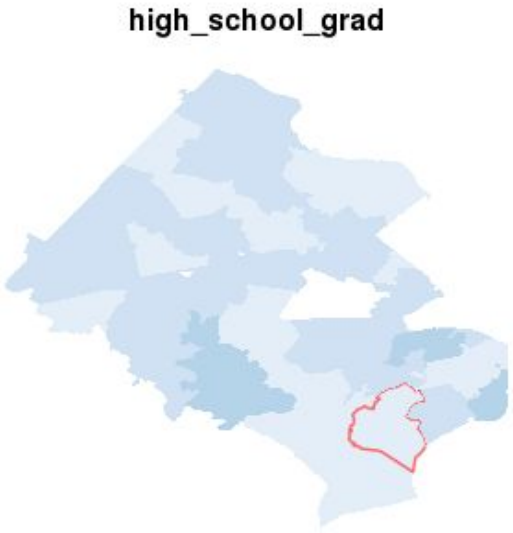


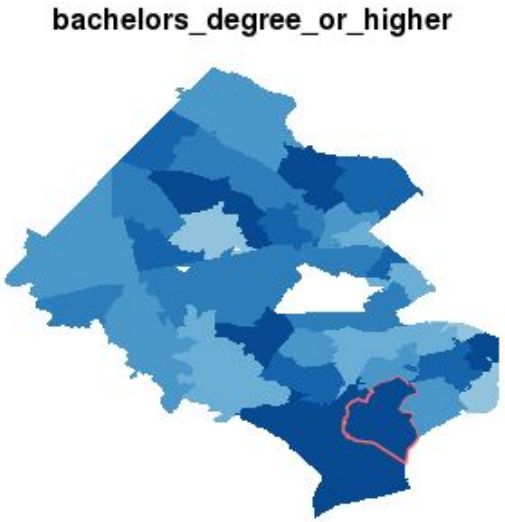
VirginiaTech.
Biocomplexity Institute

Exploratory Analysis









VirginiaTech.
Biocomplexity Institute

National Analysis

- Purpose
  - Compare across multiple bases and surrounding areas
- Dashboard
  - User Friendly Platform for Analysis
    - Visuals
    - Specific Variable Comparison
    - Saves Time
- Progress
  - Prepared Data
  - Mapped Bases
  - Dropdowns (size & color)

Future Work

- Add Plots
  - Base vs. County Zips
  - Base vs. County Average
  - Base vs. All Other Bases
- Incorporate Data
  - PDE with ACS
- Model
  - Combined Data vs. Attrition Rates

Questions

WMATA Bus Fare Evasion

DSPG - Craig Morton, Eirik Iversen, Harsimrat Pandher (Simmu), David Hinkle, Zarni Htet, Lata Kodali

SDAL - Vicki Lancaster, Gizem Korkmaz, Aaron Schroeder, Joshua Goldstein

Virginia Tech Biocomplexity Institute

SDAL Social & Decision Analytics Laboratory

Project Description

- Sponsor: Washington Metropolitan Area Transit Authority (WMATA)
- WMATA Representative: Catherine Vanderwaart
- Problem: WMATA loses approximately 10-20 million dollars a year due to fare evasion on its 1300-1500 daily bus trips.

Research Goal: Define fare evasion and analyze several different data sources from WMATA to find determinants of fare evasion.

Project Outline

We will provide:

- Literature Review
  - Will cover history of fare evasion in context of WMATA
  - Tried methods of dealing with evasion
- Data Discovery
  - Involves working with WMATA given data and ACS data to understand fare evasion.
- Data Integration & Statistical Analyses
  - Possibility quantify bus fare evasion
  - Demographic profiles of neighborhoods where evasion occurs

Plan

- Pinpoint fare evasion at the census block group level by observing morning commutes between 5-10 AM on weekdays.
- Understand economic vulnerability of those census block groups which contain bus stops.
- Define fare evasion and determine highest fare evading census block groups.
- Determine sources of variability within those highest groups.
- Combine economic vulnerability along with feasible tactics to reduce fare evasion.

WMATA Datasets

Bus stops

- Contains bus stop ID, description, latitude, and longitude.
- 10,988 observations of 5 variables.

Approximate Person Counter (APC)

- Contains front and back door entries and exits for a given bus, route, and bus stop.
- 3,779,527 observations of 20 variables.

Farebox

- Contains cash and smartrip transactions for given bus, trip number, and bus stop.
- 2,729,668 observations of 17 variables.

Data Processing

- Attached census block group information
- Concentrated on weekdays from 5-10 AM
- Calculated number of riders on the bus for APC
- Cleaned APC and Farebox datasets
- Created master file aggregated at the census block group (CBG) level with following variables:

cbg	rte	trip	dir	stop_IDs	desc
510594302021	2	20	1	6	BRADDOCK RD + FORESTDALE DR,BRADDOCK RD + T...
510594301011	2	5	1	1	BRADDOCK RD + ROANOKE LA,BRADDOCK RD + ROA...
510594302031	4	18	1	3	BRADDOCK RD + SIDEBURN RD,BRADDOCK RD + CAR...

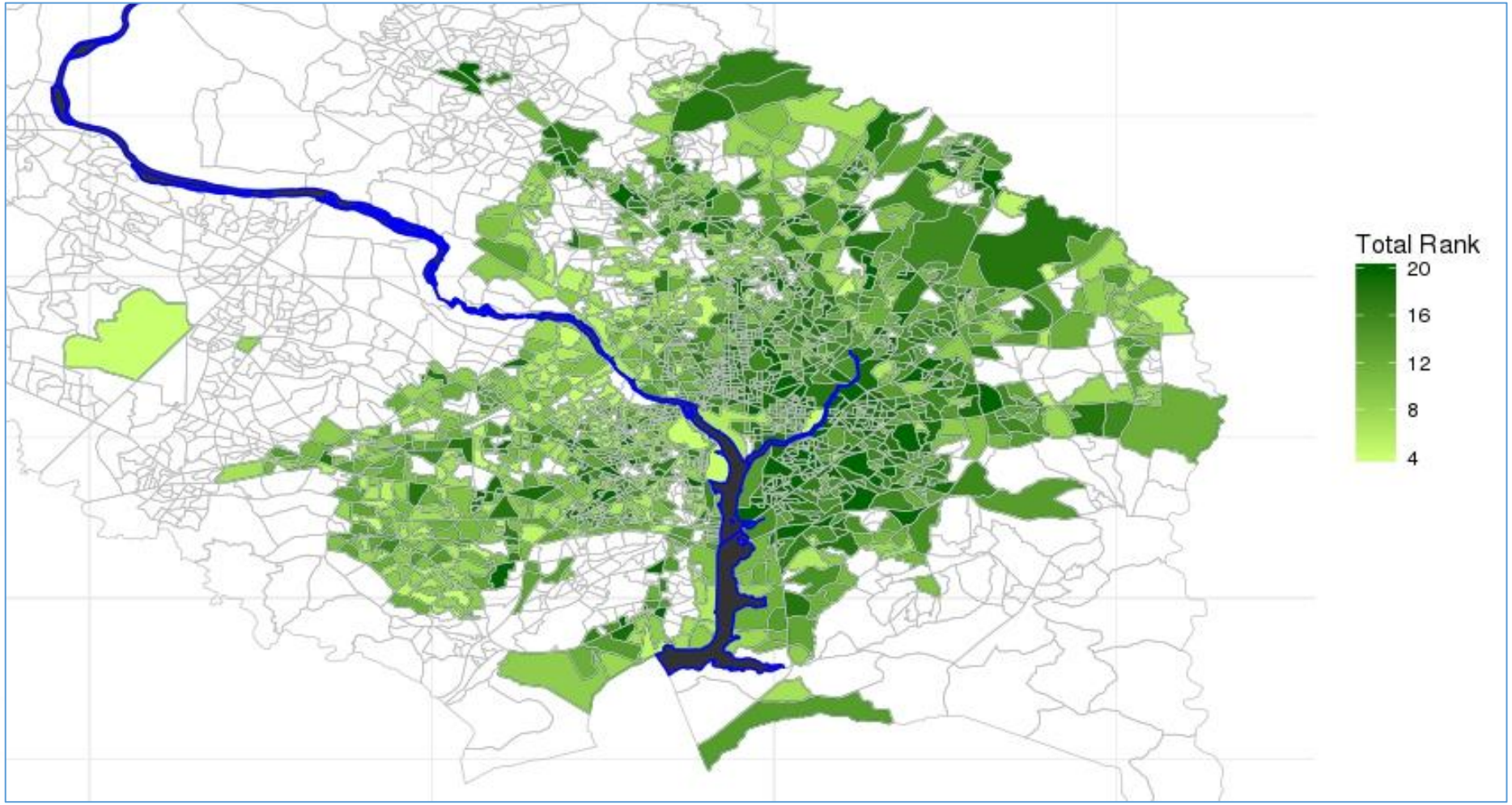
bus_IDs	in_front	in_back	out_front	out_back	on_bus	key8	cash	smartrip
28	73	0	1	0	155	0	0	65
27	9	0	6	0	107	0	0	7
39	39	0	1	0	323	0	0	30

Economic Vulnerability

In order to understand populations that might be more at risk to not be able to pay their fares, we:

- Created an economic vulnerability index
- Looked at four variables on the household level for block groups
  - Block groups that have a high number of houses receiving food stamps
  - High number of households that have no vehicle
  - High number of households that have an income below the poverty line
  - High number of households that spend more than 50% of their income on their rent / mortgage
- Take these variables and assign them a one to five quantile based on the range of the number of households
- Add up scores across the variables to get a composite rank out of 20

Economic Vulnerability Map



- 7 counties WMATA services
- Colored BGS contain bus stops
- High rank means high economic vulnerability

Fare Evasion

We estimate fare evasion by the following measures:

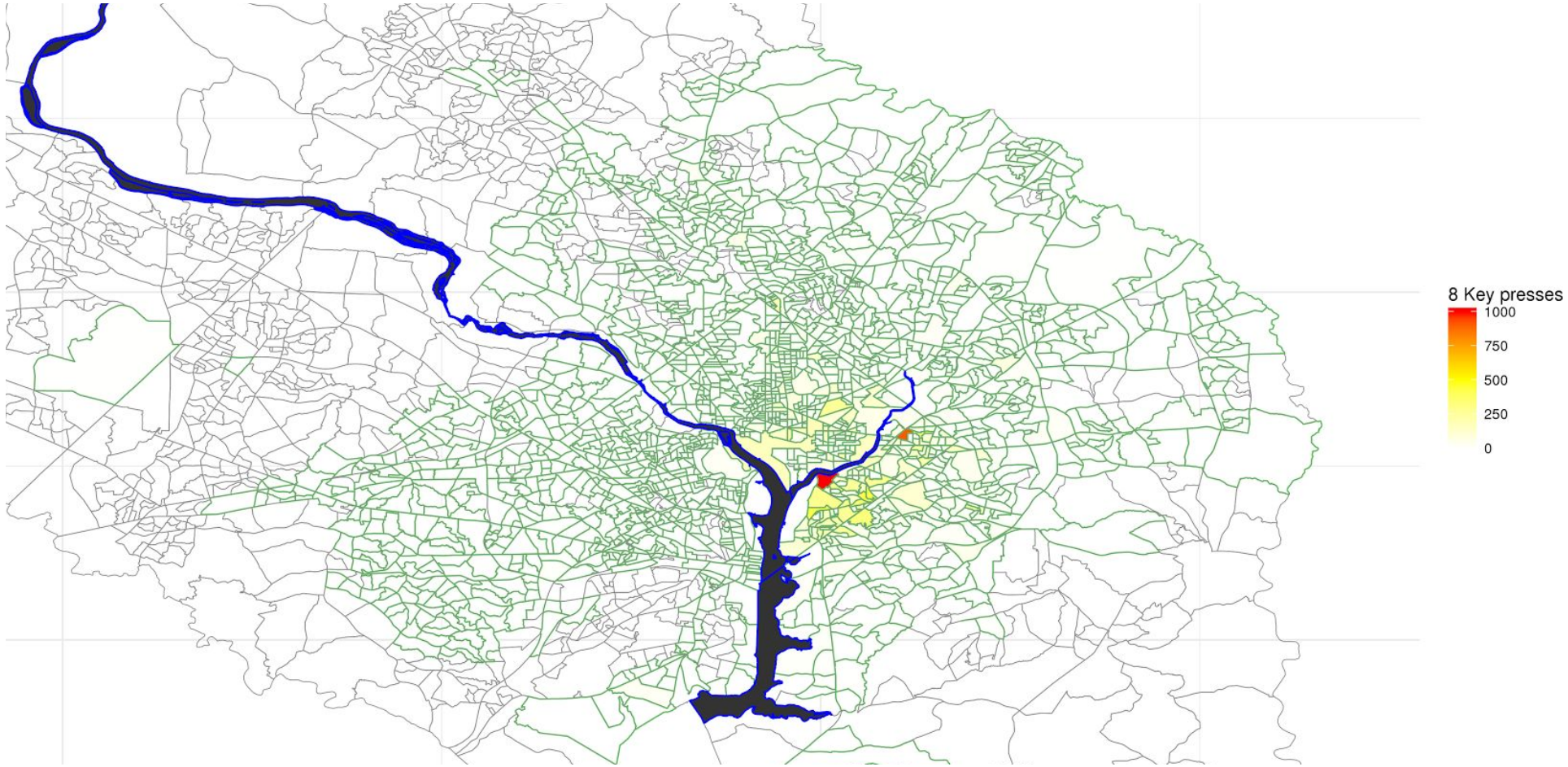
- 8 Key presses
  - A button the bus operator presses when they observe an instance of fare evasion
- Count of APC back door entries
  - People entering the bus from the back door

Back Door Entries

Green outline indicates a block group containing WMATA serviced bus stops.

VirginiaTech.
Biocomplexity Institute

8 Key Presses



Green outline indicates a block group containing WMATA serviced bus stops.

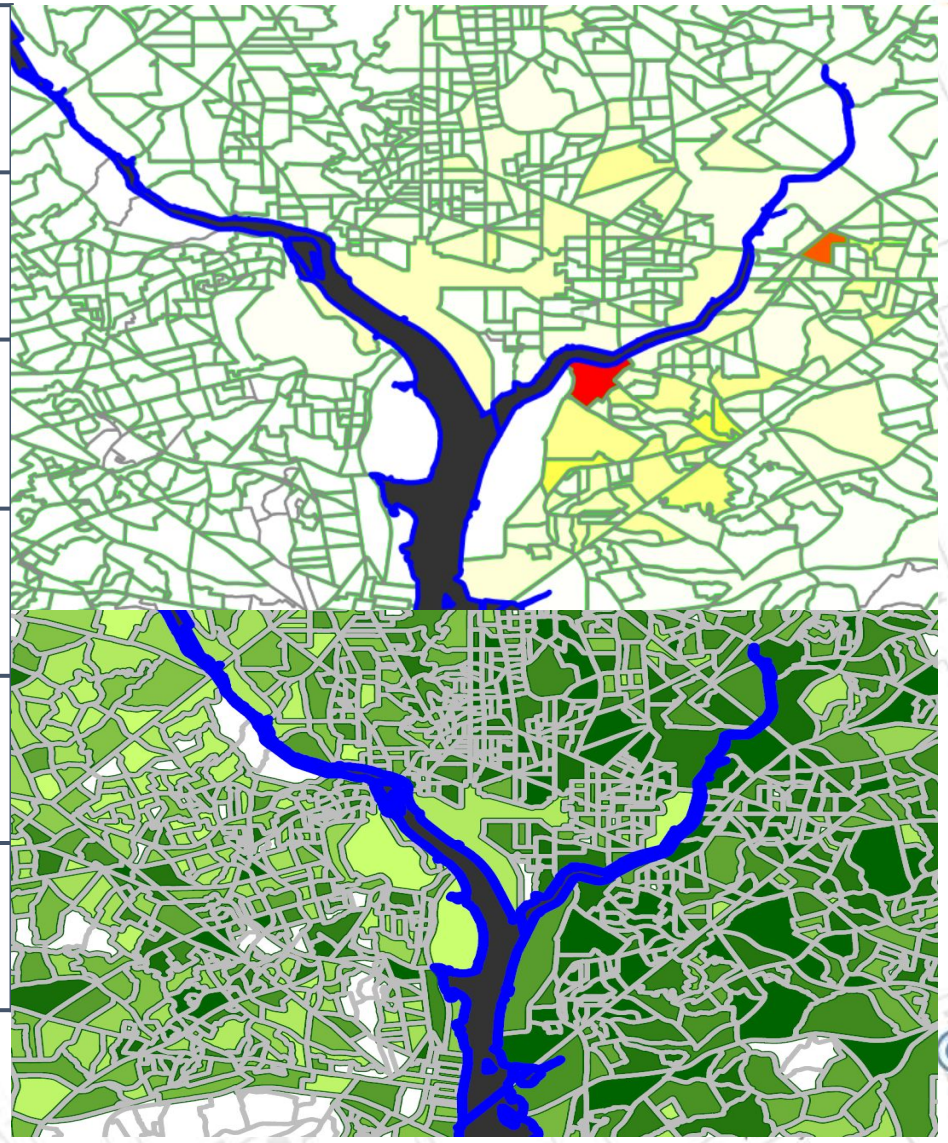
Top 5 in Back Door Entries

CBG	No.	Top Routes (Start-End Stations)
110010107001 (Central DC)	113	16Y (Columbia Pike-McPherson), S2 (Silver Spring-Federal Tri)
110010108002 (Central DC)	88	16Y (Columbia Pike-McPherson), S1 (16h St. Heights-Potomac Park)
110010074011 (Southeast DC)	76	A2 (Southern Ave-Anacostia), A4 (D.C Village - St.Elizabeths Gate 4)
110010044002 (Downtown DC)	51	64 (Fort Totten Station- Federal Tri), 90 (Congress Heights - Duke Ellington)
110010062021 (TJ Memorial)	51	64 (Fort Totten Station- Federal Tri), S2 (Silver Spring-Federal Tri)

VirginiaTech
Biocomplexity Institute

Top 5 in 8 Key Presses

CBG	No.	Top Routes (Start-End Stations)
110010074011 (Southeast DC)	1000	W4 (Deanwood-Anacostia), A2 (Anacostia-Southern)
110010078031 (Lower Central NE DC)	911	W4 (Deanwood-Anacostia), X2 (Minnesota - Lafayette)
110010075022 (Southeast DC)	442	W4 (Deanwood-Anacostia), 92 (U Street- Garfield Line)
110010074081 (Southeast DC)	428	W4(Deanwood-Anacostia), 92 (U Street- Garfield Line)
110010104001 (Southeast DC)	428	W4(Deanwood-Anacostia), A2 (Southern Ave-Anacostia)



Top 5 with Economic Vulnerability Index

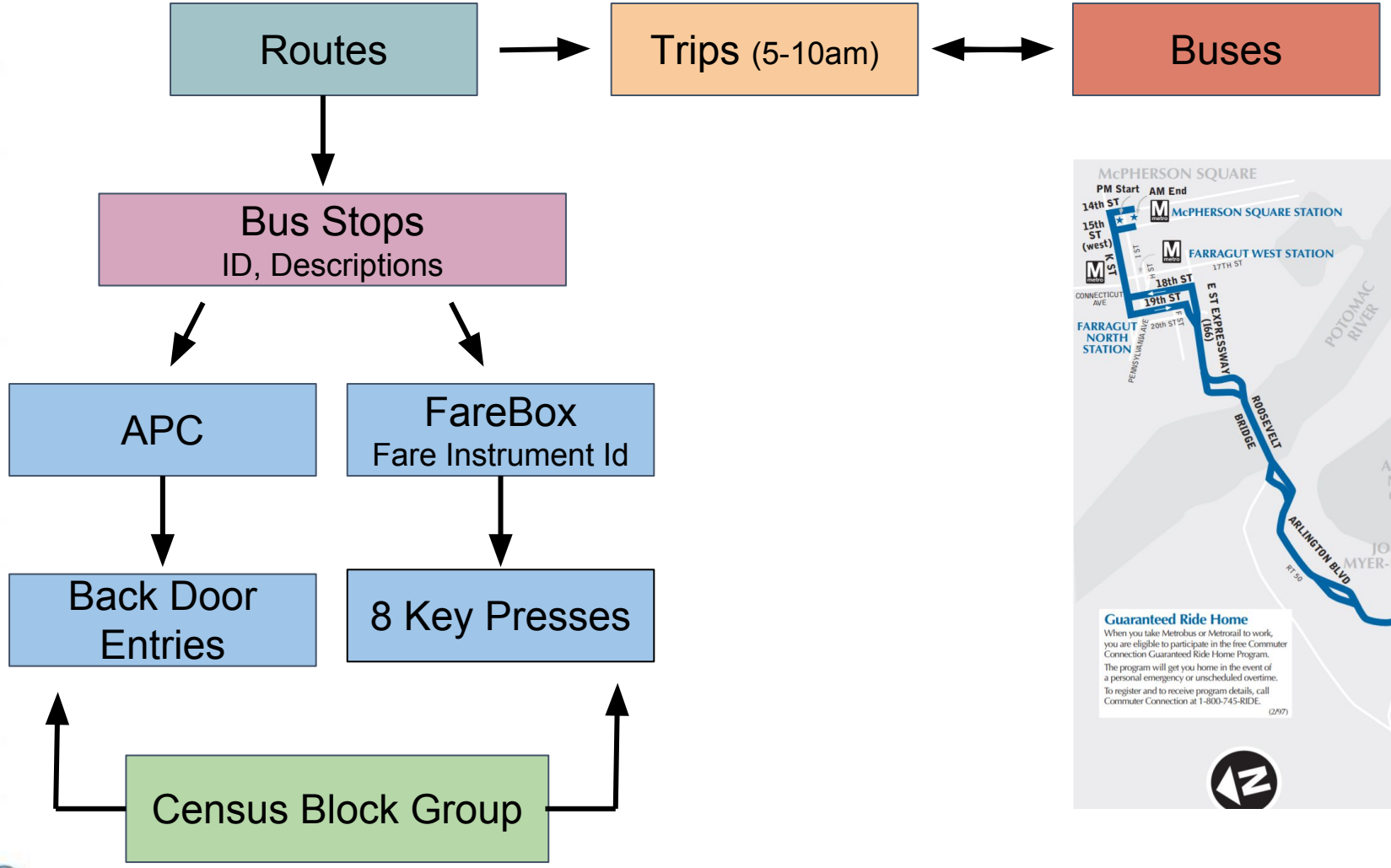
Back Door Entries:

CBG	Rank
110010107001 (Central DC)	6
110010108002 (Central DC)	17
110010074011 (Southeast DC)	19
110010044002 (Downtown DC)	18
110010062021 (TJ Memorial)	16

8 Key Presses:

CBG	Rank
110010074011 (Southeast DC)	20
110010078031 (Lower Central NE DC)	19
110010075022 (Southeast DC)	20
110010074081 (Southeast DC)	15
110010104001 (Southeast DC)	19

Next Steps


Routes
Trips (5-10am)
Buses
Bus Stops
ID, Descriptions
APC
FareBox
Fare Instrument Id
Back Door Entries
8 Key Presses
Census Block Group

MetroExtra
McPHERSON SQUARE
PM Start
AM End
14th ST
McPHERSON SQUARE STATION
15th ST (west)
FARRAGUT WEST STATION
17TH ST
K ST
CONNECTICUT AVE
18th ST
19th ST
20th ST
FARRAGUT NORTH STATION
PENNSYLVANIA AVE
E ST EXPRESSWAY (166)
POTOMAC RIVER
ROOSEVELT BRIDGE
ARLINGTON NATIONAL CEMETERY
ARLINGTON BLVD
RT 50
JOINT BASE MYER-HENDERSON HALL
PERSHING DR
WASHINGTON BLVD
WALTER REED DR
2ND ST
COURTHOUSE RD
BARTON ST
WALTER REED DR
COLUMBIA PIKE
GLEBE RD
OAKLAND ST
GEORGE MASON DR
BARCROFT
FOUR MILE RUN DR
RT 244
RT 27
Legend
Metrorail Station
Terminal Stands
Designated stops in Virginia
Guaranteed Ride Home
When you take Metrobus or Metrorail to work, you are eligible to participate in the free Commuter Connection Guaranteed Ride Home Program.
The program will get you home in the event of a personal emergency or unscheduled overtime.
To register and to receive program details, call Commuter Connection at 1-800-745-RIDE.
(2/97)

Comments

Thank you!

Questions?

Suggestions?

Thoughts?

Repurposing 911 Data toward Evidence-Based Policy

PI & Co-PI: Gizem Korkmaz & Josh Goldstein
Team: Claire Kelling (Penn State), Sayali Phadke (Penn State), Madison Arnsbarger(VT)
Kyle Morgan (VT), Sean Pili (VT), Craig Morton(VT), Hannah Brinkley (VT)
Sponsor: Arlington County Police and Fire Departments

VirginiaTech Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Project Overview

- Overall goal is to improve **"Situational Awareness":** to improve quality of life and services while accelerating the county's efficiency and resiliency

- **Research Questions**
  - Featured Events: What to expect from specific events so that the Police Department can be better prepared?
  - Characterizing Communities: What are characteristics of communities that encounter specific types of crimes?
  - Response Time: Which factors affect emergency response time?

Impact of Featured Events on Crime Patterns

PI & Co-PI: Gizem Korkmaz & Josh Goldstein
Team: Sayali Phadke (Penn State), Craig Morton (VT), Hannah Brinkley (VT)
Sponsor: Arlington County Police Department

Virginia Tech Biocomplexity Institute

SDAL Social & Decision Analytics Laboratory

Main idea

- Specific events may affect crime
- Effect will depend on:
  - Type of event
  - Number of people involved/affected
  - Location of the event
  - Type of crime
  - Other time series confounders?
- Police department stipulates that such an effect exists. We are testing the hypothesis

Types of events

- Political events
- Holidays
- Weather-related events
- Sports
- Social events
- Full-moon

Event-date collection

- Political events
  - All elections during 2005-2015 for Arlington, VA manually recorded*
- Sports
  - Events that would have a high number of people attending / draw to bars:
    - Marine Corps Marathon
    - 2013 Super Bowl
    - Playoff games for the Washington sports teams
- Social events
  - Arlington events since 2007 retrieved from Yelp.com
    - Able to get counts of number of people interested
    - Looking into the events with highest number of people interested
- Weather-related events
  - 2010 Snowmageddon, hottest days of the year, hurricanes/tropical storms
- Holidays, and Full-moon
  - Manually retrieved dates for Thanksgiving, Halloween, Christmas etc. and full-moons

*https://arlingtonva.s3.dualstack.us-east-1.amazonaws.com/wp-content/uploads/sites/3/2013/09/Candidate_History.pdf

Data preparation

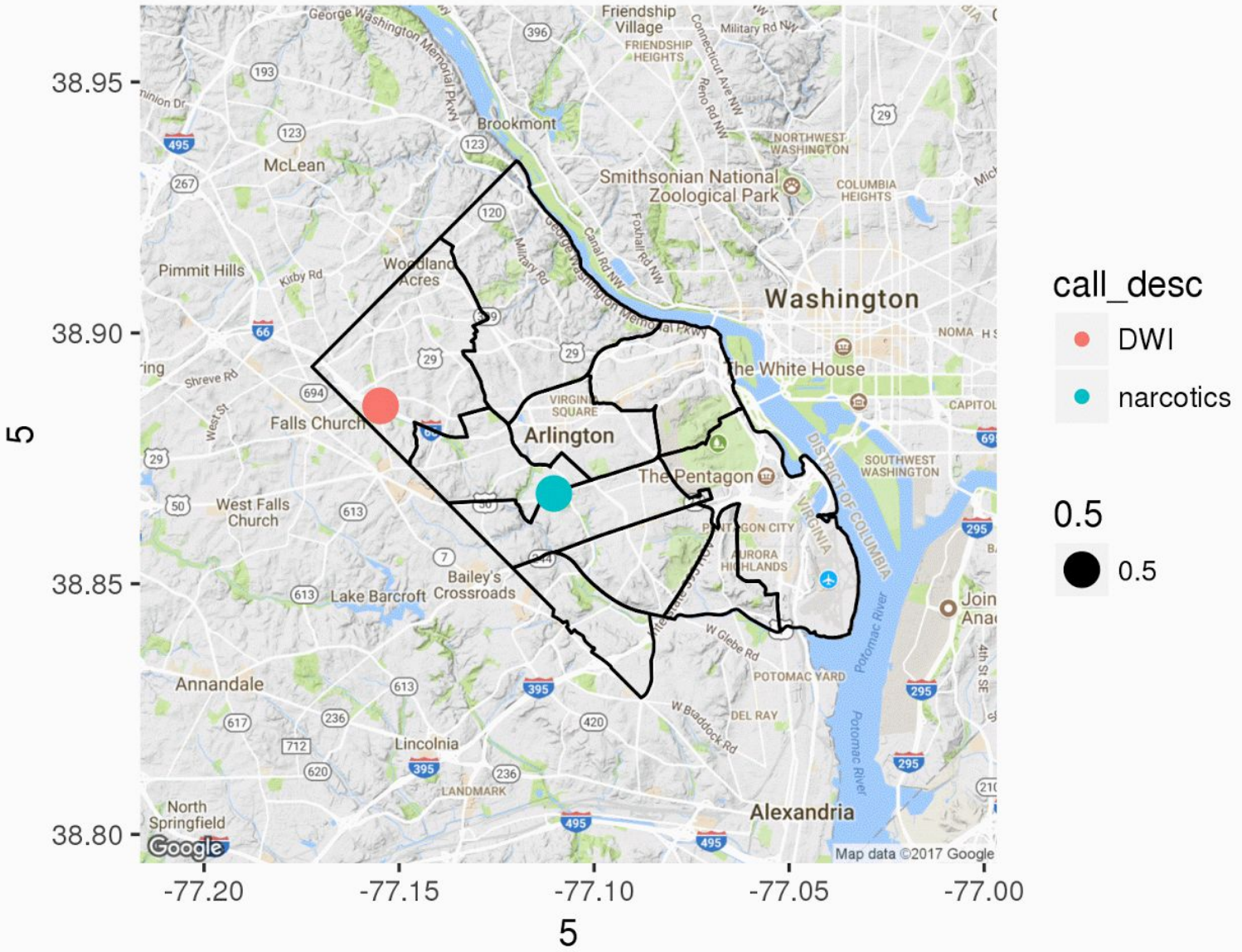
- Police have provided two separate datasets
  - Incident table (inmast):
    - Information on crime incidents
      - Location
      - Type of incident
        - This is a four digit code; too many values
        - First two digits indicative of overall category of crime
    - Doesn’t contain information on time at which call was received
  - Call for Service table (cfs):
    - Exact date and time of 911 call
- Merged inmast and cfs to get both the exact date and time of the call, and the location and type of incident

2009 Inauguration

The 2009 Presidential Inauguration was President Obama's first inauguration

- Washington D.C. on January 20th, 2009 around 11am
- Interest
  - Possible increase in traffic-related crimes
  - Possible increase in protest-related crimes

2009 Presidential Inaguration: 2009-01-19

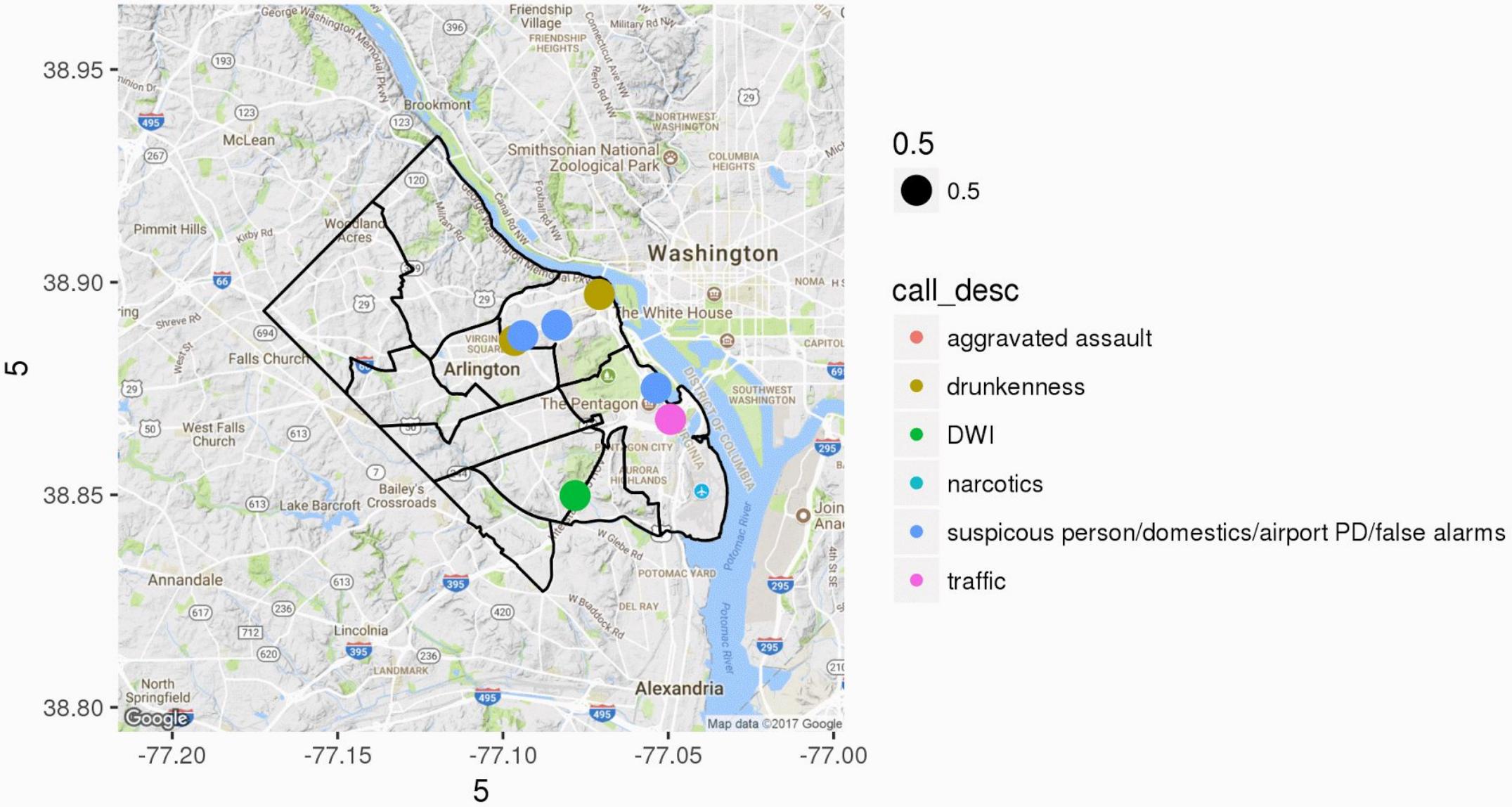


Super Bowl 2013

This Super Bowl in 2013 was the Baltimore Ravens vs the San Francisco 49ers

- Took place in New Orleans
- Started at 6:30 PM EST
- A local team played, so it seemed interesting
  - Increase in drinking-related offenses:
    - Disturbing the peace
    - Drunkenness
    - DWI

2013 Super Bowl: 2013-02-02

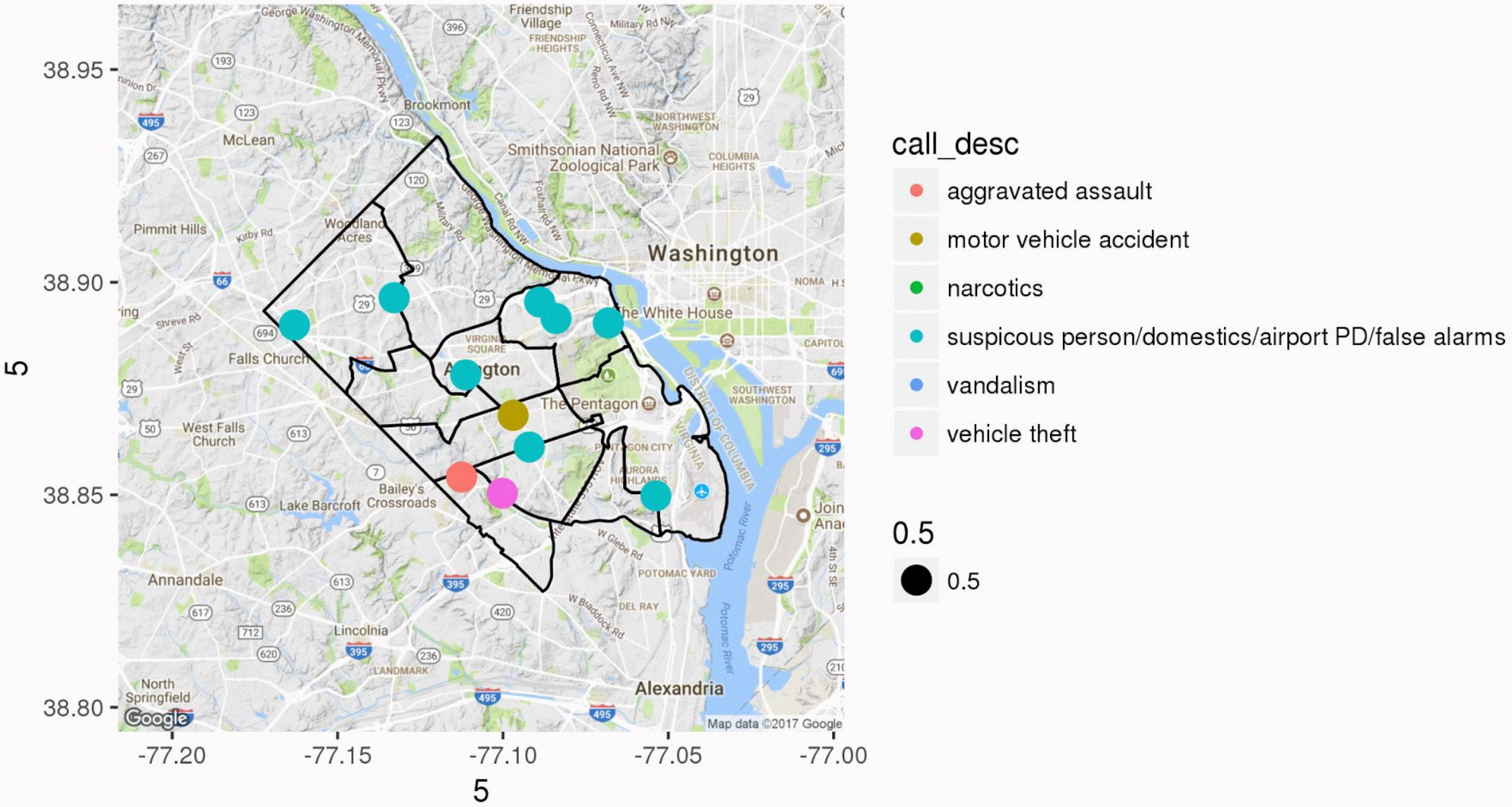


Snowmageddon 2010

Snowstorm across North America resulting in ~18.7” of snow

- Occurred between February 5th and 6th in 2010
- Interest
  - Possible increase in domestic violence/ domestic abuse crimes

Snowmageddon: 2010-02-01



Statistical Model

- A stats model lets us test hypotheses about events.
- To account for dependence in rate over time as well as across space, fit a spatiotemporal log-Cox point process to Arlington crimes.
- The rate at location s_i and time t is modeled as

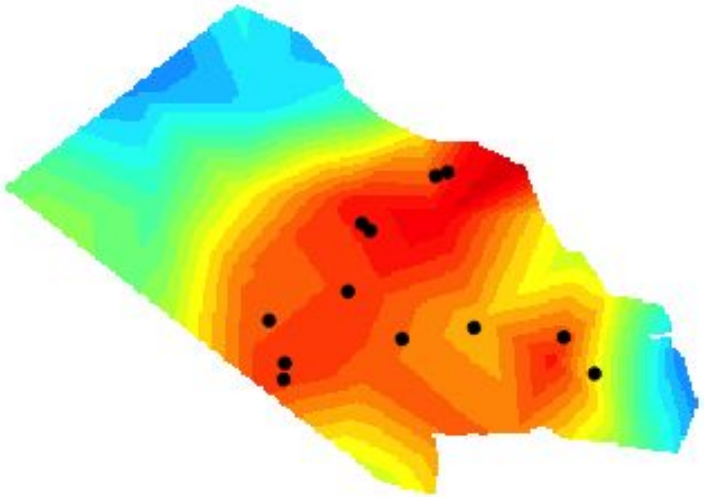
$$\log(\lambda(s_i, t)) = z(s_i, t)\boldsymbol{\beta} + \xi(s_i, t) + \epsilon(s_i, t)$$
$$\xi(s_i, t) = a\xi(s_i, t-1) + \omega(s_i, t)$$

Where $\epsilon(s_i, t) \sim N(0, \sigma_\epsilon^2)$, and $\omega(s_i, t)$ is a time-independent spatial process with Matern covariance. Inferece relies on the SPDE approximation[1] implemented in the INLA R package.

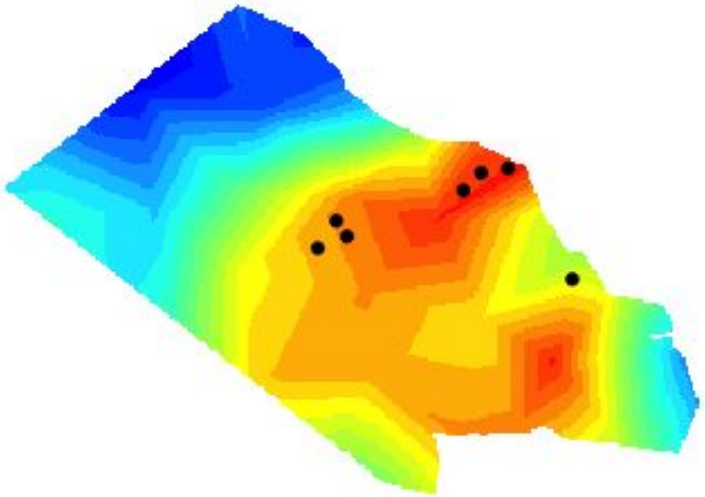
[1] Lindgren et al., 2011

Model fit (Superbowl)

No significant effect on crime rate February 3rd, 6pm-12pm...



01/27/13 06AM



02/03/13 06PM

Challenges and next steps

- Challenges
  - Lost data folder!
  - Lack of clearer hypothesis
- Next steps:
  - Whatever visualizations/modeling we don't achieve by Wednesday
  - Show the results to police department, see if it makes sense to them, extend analysis depending on feedback
  - Drill down on crime types depending on initial findings

Characterizing Communities for Crime Rate Inference

PI & Co-PI: Gizem Korkmaz & Josh Goldstein
Team: Claire Kelling (Penn State), Kyle Morgan (VT), Sean Pili (VT), Craig Morton(VT)
Sponsor: Arlington County Police Department

Virginia Tech Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Research Questions

- We aim to “Characterize Communities” that are vulnerable to certain crime types
- Modeling specific crime types
  - Mental Health
  - Drug-Related
  - Domestic Violence

- Research Questions:
  - Which communities are specifically vulnerable to each of these crime types?
  - Do Points of Interest play a role in crime rates?
  - Does spatial proximity and/or travel between groups affect crime rates?

Literature

- Most recent work is spatially focused (hotspots)
  - Some work in time and population-centric paradigms (community and individual profiling)

Wang et al (2016). Crime Rate Inference with Big Data: *Association for Computing Machinery*.

- Data Sources:
  - (Nodal) Demographic Data: 2000 Census
  - (Nodal) Point-of-Interest (POI) Data: Foursquare
  - (Edge) Geographic Influence: boundary shapefiles
  - (Edge) Hyperlink by Taxi Flow: Illinois Freedom of Information Act

- Our contributions:
  - Spatial modeling, specific crime types, new data sources/types

Data Sources (Descriptions)

- Crime Data
  - Provided by ACPD
- ACS
  - 2015 ACS 5-year sample
  - Characteristics such as income, poverty rate, education, etc.
- Facebook: 1 page per neighborhood (on avg)
- Points of Interest
  - Public and private mental health facilities
  - Social: restaurants, bars, etc
  - Other facilities: metro stops, community centers, churches, schools, etc.

ACPD Data - Data Exploration

- Created 3 subsets of the data based on incident and Original_Call Types
  - Mental Health: (6190)
  - Domestic Violence ((5619)
  - Drugs (7945)
- Original call codes (DRUG) - Call codes (XXXX)
- Kept observations if they had correct incident OR Original_Call Types
- Wanted to find a way to asses the quality of this data before modeling it.
- Studied the mismatch between original call and call incident

Data Exploration

First, found the number of Original_Call was correct but the incident wasn't.
Next slide: Number of times the incident type was correct but Original_Call wasn't.

Subset	Orignal_Calls	Frequency	Percent Mismatched	Popular_Incident_Types
Domestic Violence	DOME	5282	.36	Assault (576), Drunk/Liquor, (230) Miscellaneous Officer(182)
Mental Health	MO, MOP	2100	.15	Miscellaneous Officer (75), Assault(66), Drunk/Liquor (56)
Drugs	DRUG, PID	3337	.07	Lost/Found Persons (73), Miscellaneous Officer(50), Citizen/Miscellaneous Public (16)

ACPD Data Exploration

Subset	Incidents	Frequency	Percent Mismatched	Popular Original Call_types
Domestic Violence	ASSAULT FAMILY , DOMESTICFAMILY DISTURBANCE etc.	3715	.09	Assault With Injuries (32) 911 Suspicious Circumstances (30) ,Assault, Just Occurred (20)
Mental Health	MENTAL CASES, ATTSUICDSUICIDE ATTEMPTS DRUGS	5867	.70	Administrative Services (1528) Overdose (426)
Drugs	NARCOTICMARAJUANA-POS SESING, NARCOTICNARCOTICS COCAINE - SELLING	7714	.60	Traffic Stop (2456), Subject Stop (369), Suspicious Persons (209)

Data Exploration - Social Media

- Facebook pages of 53 neighborhoods in AC - community exploration.
- Assessed the usefulness of each page
- Currently have information on the number of likes per page and frequency of posts. In future will collect avg number of comments/likes per post etc. and will scrape the useful pages to use for modeling.

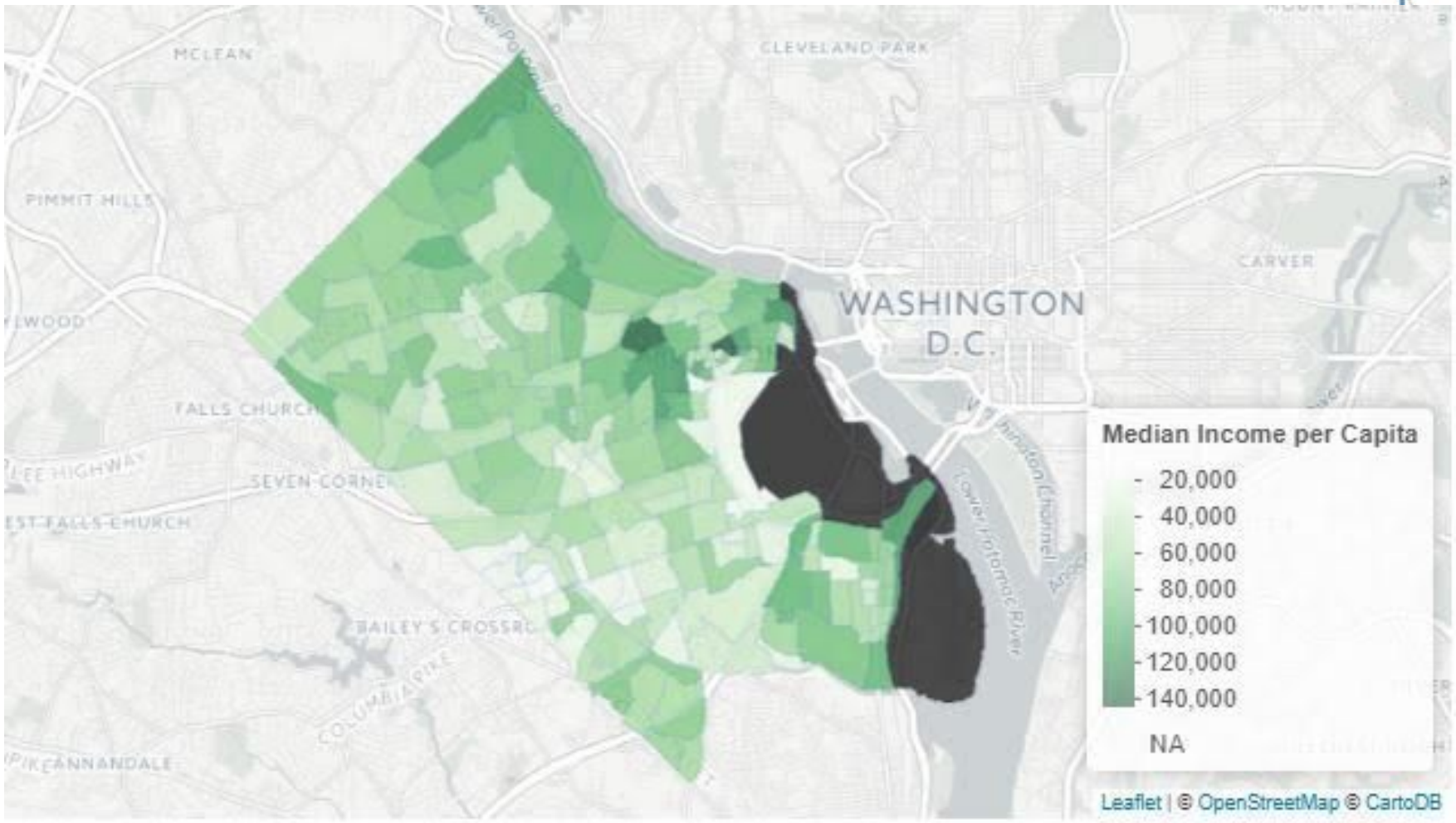
FB page Link	Useful/Not Useful	Likes/Membe	Post Frequency (average per month)
https://www.facebook.com/LFLAlcovaH	not useful	59	6.66
https://www.facebook.com/pages/Arling	not useful	0	0.42
https://www.facebook.com/pg/Arlington	not useful	42	1.16
https://www.facebook.com/pages/Arling	useful	88	14
https://www.facebook.com/Ashton-Heig	useful	51	2.5
https://www.facebook.com/aurorahighla	not useful	19	1.16
https://www.facebook.com/BallstonBID/	useful	4189	35
https://www.facebook.com/pages/Barcr	useful	59	43
NA	NA	NA	NA
https://www.facebook.com/BluemontFo	not useful	70	0.41
https://www.facebook.com/bluemontcivi	useful	231	1
https://www.facebook.com/BYB1998/	not Useful	19	1.16
NA	NA	NA	NA
https://www.facebook.com/pages/Cherr	Useful	60	15.41
https://www.facebook.com/claremontciv	Useful	54	1.72
https://www.facebook.com/Clarendon-N	Useful	303	9.08
https://www.facebook.com/pages/Colon	Not Useful	40	1
https://www.facebook.com/ColumbiaFo	Useful	103	78
NA	NA	NA	NA

Data Exploration - Social Media

FB page Link	Useful/Not Useful	Likes/Membe	Post Frequency (average per month)
https://www.facebook.com/LFLAlcovaH	not useful	59	6.66
https://www.facebook.com/pages/Arling	not useful	0	0.42
https://www.facebook.com/pg/Arlingtoni	not useful	42	1.16
https://www.facebook.com/pages/Arling	useful	88	14
https://www.facebook.com/Ashton-Heig	useful	51	2.5
https://www.facebook.com/aurorahighla	not useful	19	1.16
https://www.facebook.com/BallstonBID/	useful	4189	35
https://www.facebook.com/pages/Barcr	useful	59	43
NA	NA	NA	NA
https://www.facebook.com/BluemontFo	not useful	70	0.41
https://www.facebook.com/bluemontcivi	useful	231	1
https://www.facebook.com/BYB1998/	not Useful	19	1.16
NA	NA	NA	NA
https://www.facebook.com/pages/Cherr	Useful	60	15.41
https://www.facebook.com/claremontciv	Useful	54	1.72
https://www.facebook.com/Clarendon-N	Useful	303	9.08
https://www.facebook.com/pages/Colon	Not Useful	40	1
https://www.facebook.com/ColumbiaFo	Useful	103	78
NA	NA	NA	NA

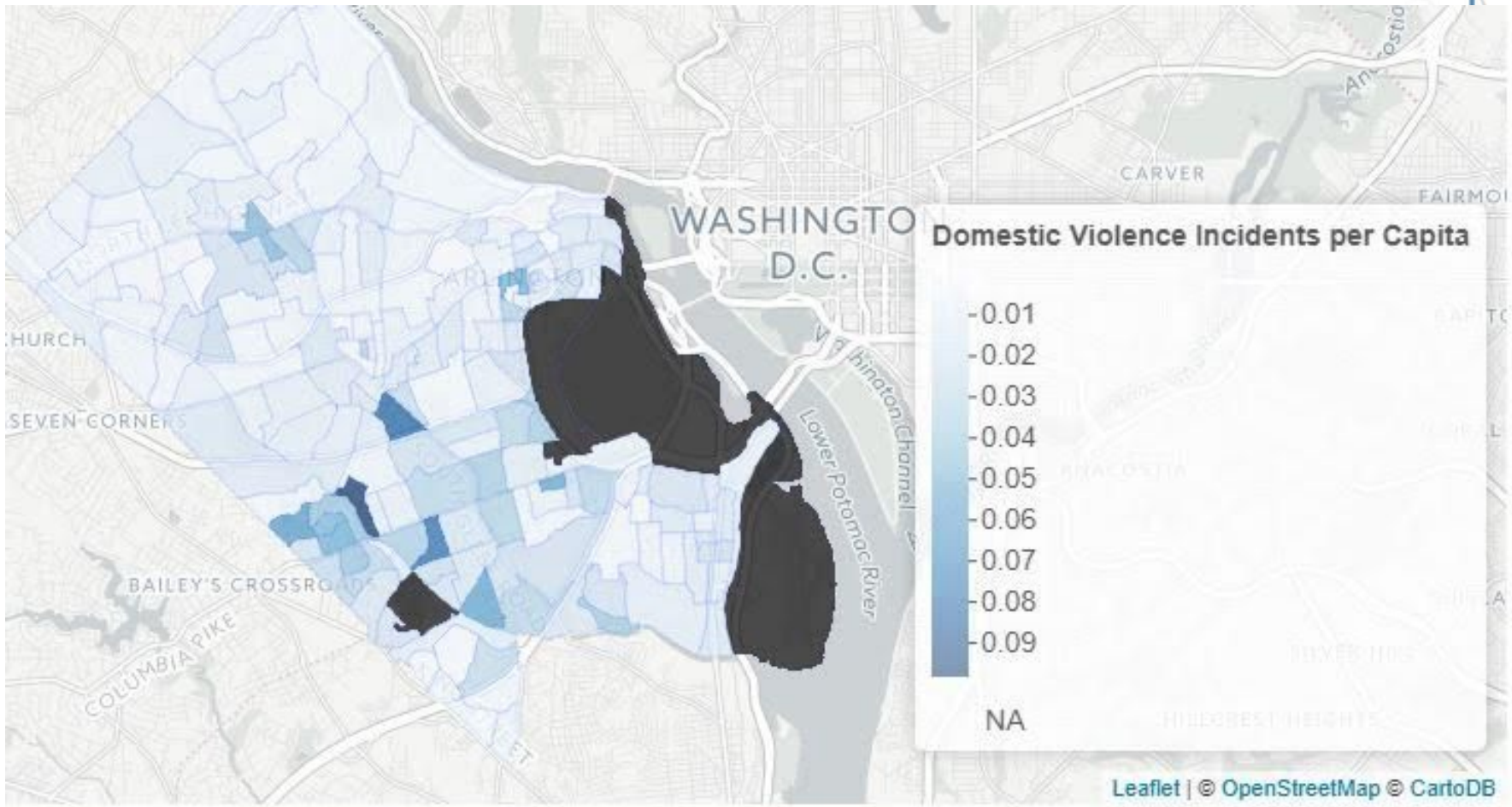
Characterization of Arlington- ACS

Plotting ACS variables by block group to better understand the demographics of Arlington



Characterization of Arlington- Crime

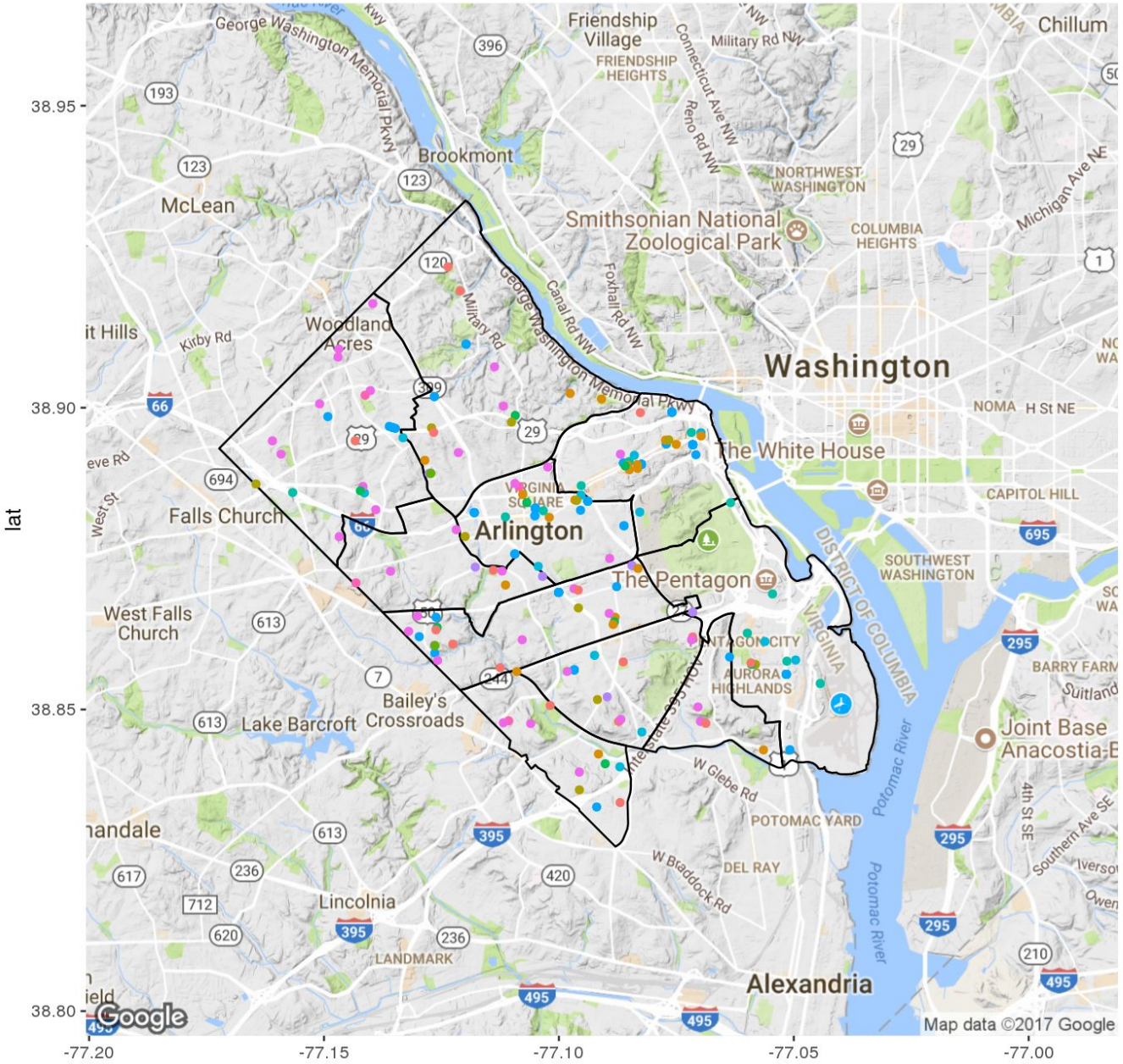
Plotting crime totals and crimes per capita to understand which areas are most susceptible to certain types of crime



Points of Interest/ Radar Search

- Public Arlington Data: (Open GIS Data Portal)
  - Hospitals/Metro/Schools
- Mental Health Providers (government and private)
  - Government: Substance Abuse and Mental Health Administration (SAMHSA)
  - Private: PsychologyToday.com
- Google Maps API Radar Search:
  - Bars
  - Churches

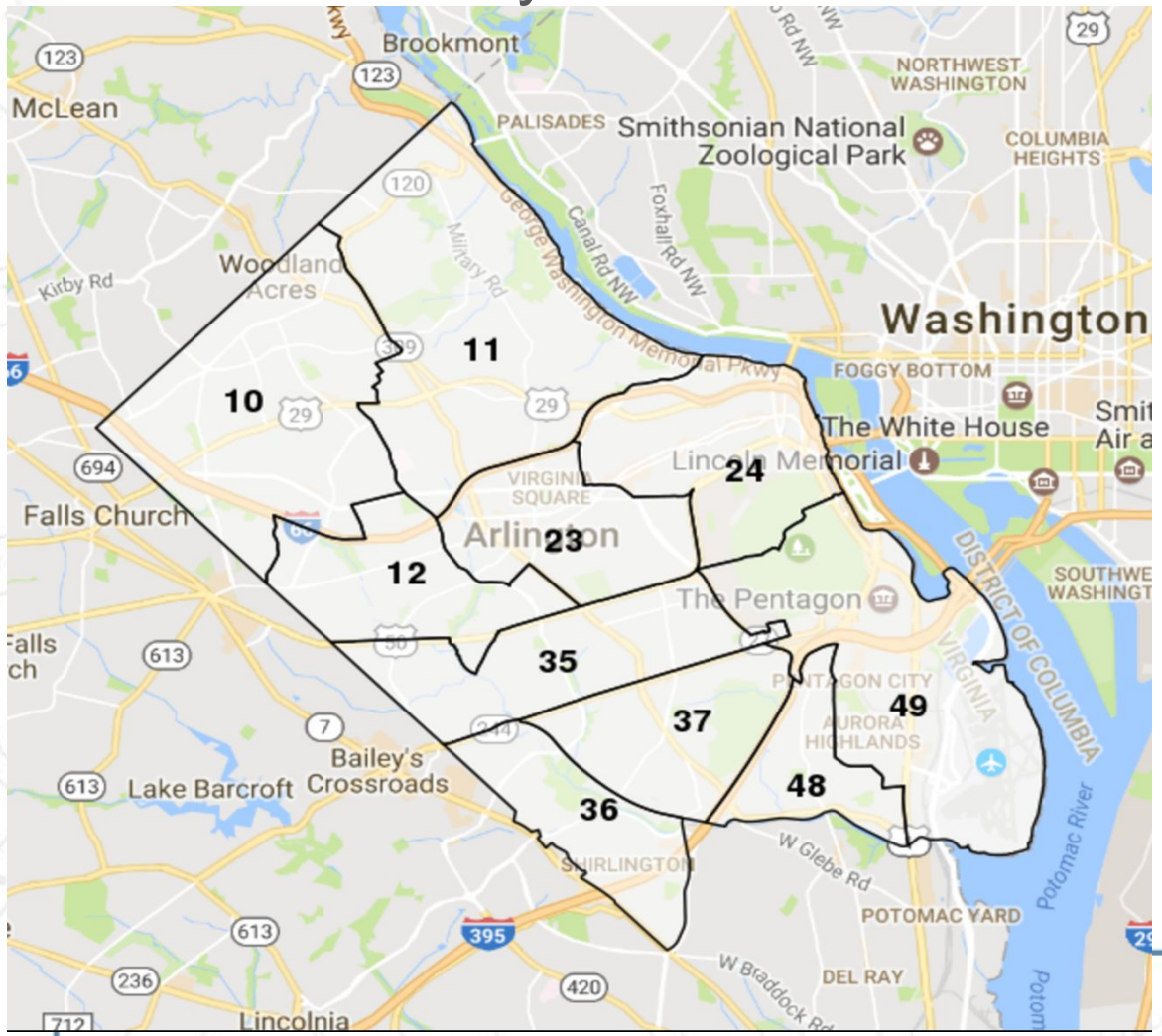
Points of Interest/ Radar Search

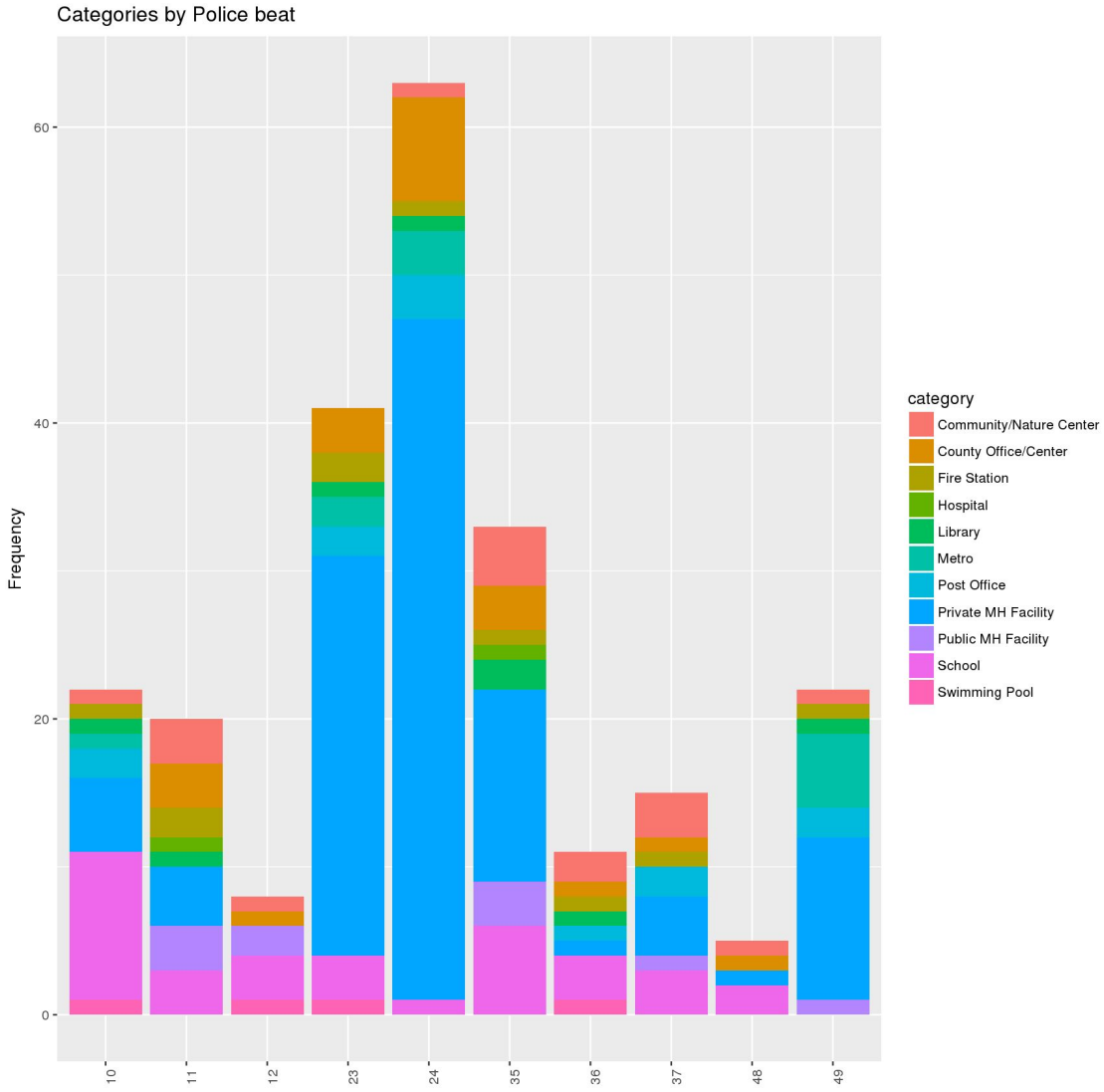


VirginiaTech.
Biocomplexity Institute

Points of Interest/ Radar Search

- Similar analysis will be done on the block and tract level





Transportation Data

- The goal is to identify flows between communities as they influence each other due to interactions (in addition to geographical influence)
- We use WMATA bus routes to compute the active routes (passenger volume) in Arlington County and travel volume between census block groups

	GEOSTOPID_dimstop	GEOID	route	Route_Direction	STOP_SEQUENCE	total_entry	total_exit
1	1786	510131030003	22C	WEST	9	2	15
2	1786	510131030003	22A	WEST	16	1	1
3	1786	510131030003	22F	WEST	9	0	7
4	1786	510131030003	22C	WEST	9	2	15
5	1786	510131030003	22A	WEST	16	1	1
6	1786	510131030003	22F	WEST	9	0	7
7	1786	510131030003	22C	WEST	9	2	15
8	1786	510131030003	22A	WEST	16	1	1
9	1786	510131030003	22F	WEST	9	0	7
10	1786	510131030003	22C	WEST	9	2	15
11	1786	510131030003	22A	WEST	16	1	1
12	1786	510131030003	22F	WEST	9	0	7
13	1786	510131030003	22C	WEST	9	2	15

The Model

- Spatial Modeling based off of literature review paper
- Unit of analysis: census block groups

- Modeling ideas:
  - Markov Random Fields
  - Poisson CAR Model
  - Network approach to geographic data
  - Spatial Generalized Linear Mixed Model (SGLMM)
  - Spatial Point Process

Emergency Response Time

PI & Co-PI: Gizem Korkmaz & Josh Goldstein
Team: Claire Kelling (Penn State), Madison Arnsbarger(VT)
Sponsor: Arlington County Fire Department

Virginia Tech Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Research Question

- Goals:
  - To understand the impacts of various factors on fire department response time.
  - To ultimately improve allocation of emergency response resources and safety of Arlington residents.
- Definitions:
  - Processing time = Alarm time → Dispatch time
  - Turnout time = Dispatch time → Enroute time
  - Travel time = Enroute time → Arrival time

Data:

- 399,158 rows, 68 columns
- Years 2010 - 2015
- Subset data:
  - exclude non-AC stations
  - SF cad call type
  - exclude turnout times longer than 10 minutes

Message Incident:

- 381,832 r ows

incidentkey	100
station	95.4
locationtype	100
directions	5.9
address1	53.1
latitude	90.4
longitude	90.4
alarms	100
cadcallnumber	100
cadcalltype	100

Incident:

- 164,900 rows

incidentkey	100
userkey	100
incidentnumber	100
exposure	100
incidenttype	91.1
mutualaid	91.1
alarmdatetime	100
shift	100
district	97
completeddatetime	84.2

Apparatus:

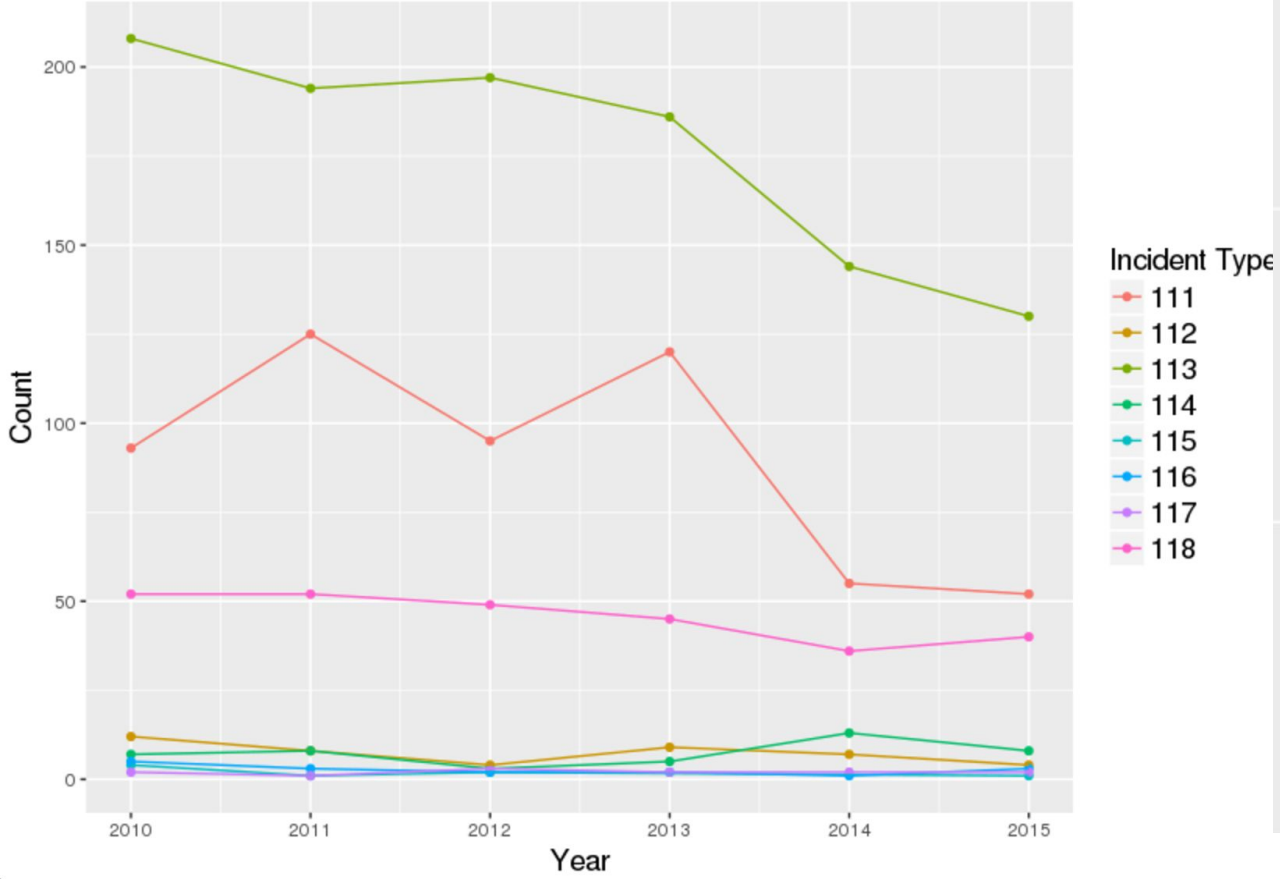
- 344,409 rows

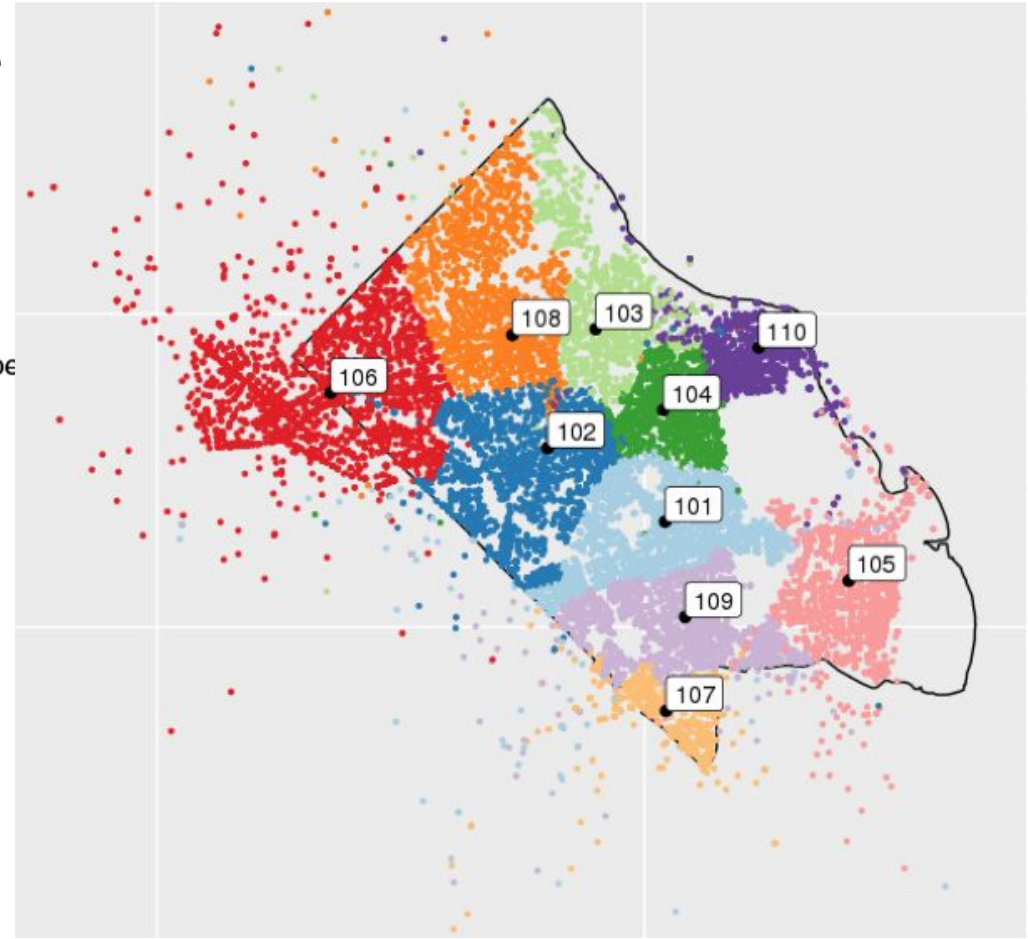
incidentapparatuskey	100
incidentkey	100
sequence	100
apparatusid	99.99
apparatustype	99.99
dispatchdatetime	100
enroutescenedatetime	92.82
arrivaldatetime	80.59
cleardatetime	99.98
apparatususe	88.95

Exploratory Analysis:

SF CAD Calls by Incident Type

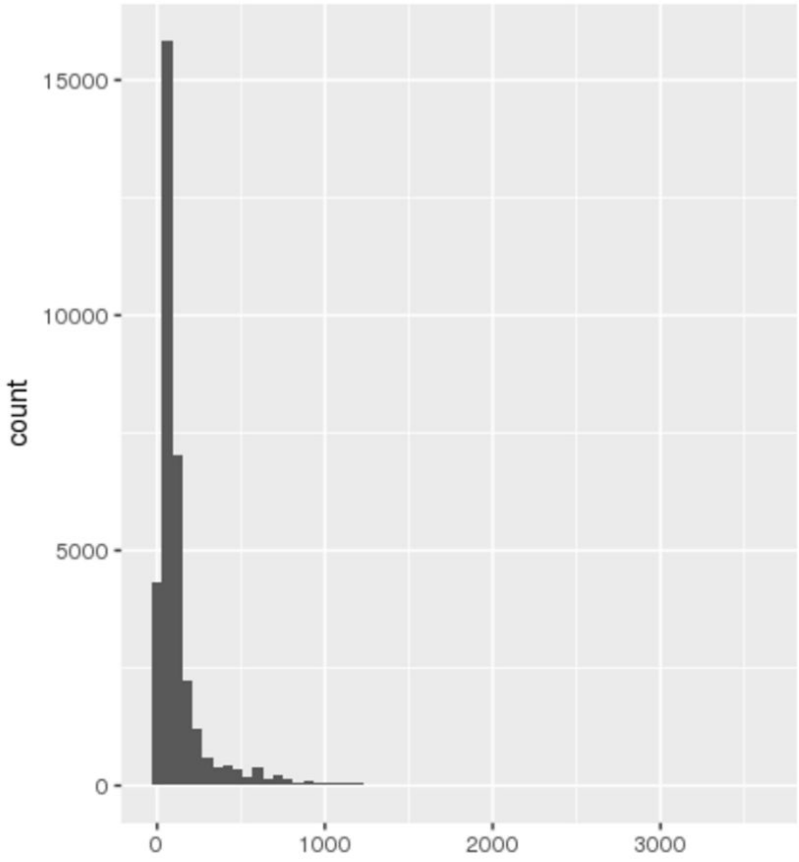
111 = building fire, 113 = kitchen fire, 118 = trash or rubbish fire





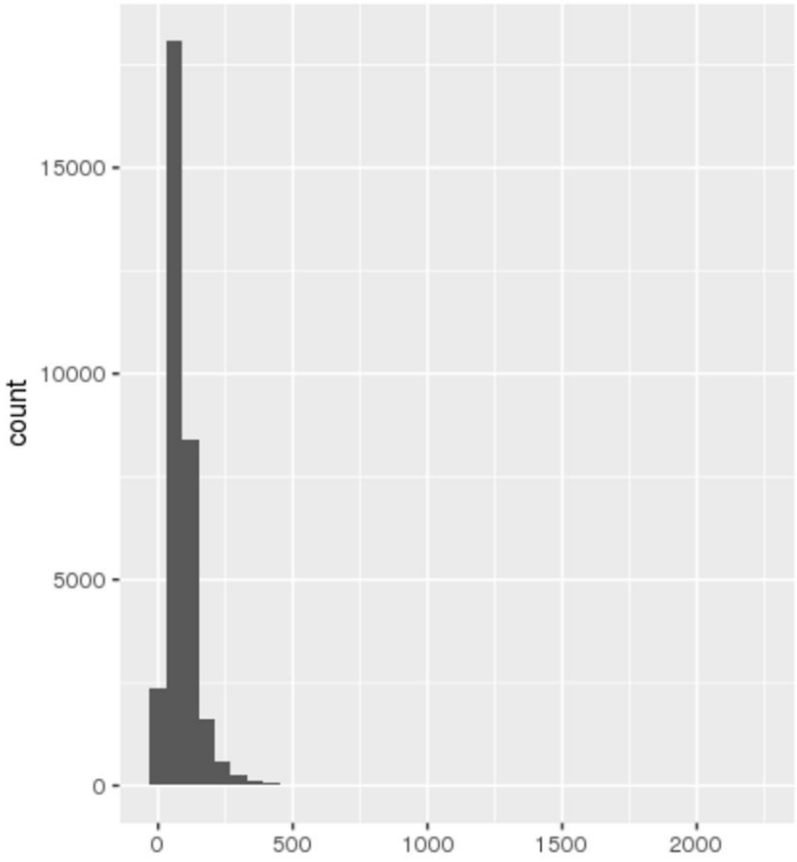
Exploratory Analysis: Response Time Intervals

Processing Times



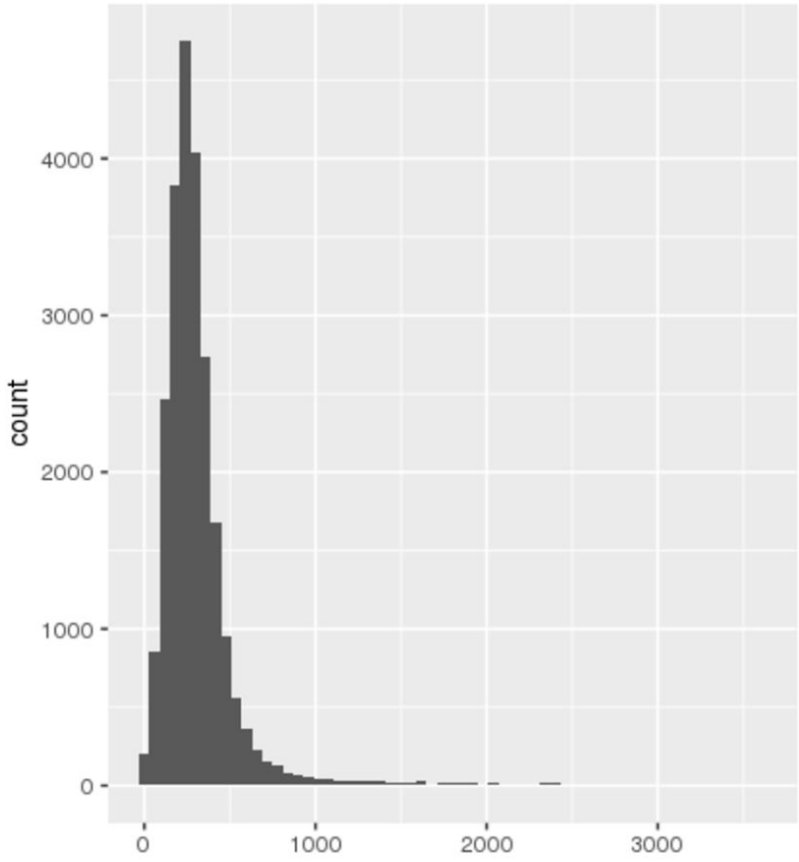
Processing Time (sec)

Turnout Times



Turnout Times (sec)

Travel Times



Travel Times (sec)

Turnout Time Model

Predictors:

- Year (spline fit)
- Month (spline fit)
- Hour (spline fit)
- Apparatus type
- Station
- CAD Call type*
- Number of floors in station*

Model:

$$f(\text{turnout time}) = \beta_0 + \beta_1 \text{apparatus} + \beta_2 \text{station} + \beta_3 \text{month} + \beta_4 \text{year} + g(\text{hour}) + \epsilon$$

$$\text{where we assume } \epsilon_i \sim N(0, \sigma^2)$$

$$f() \text{ is a transformation of the response}$$

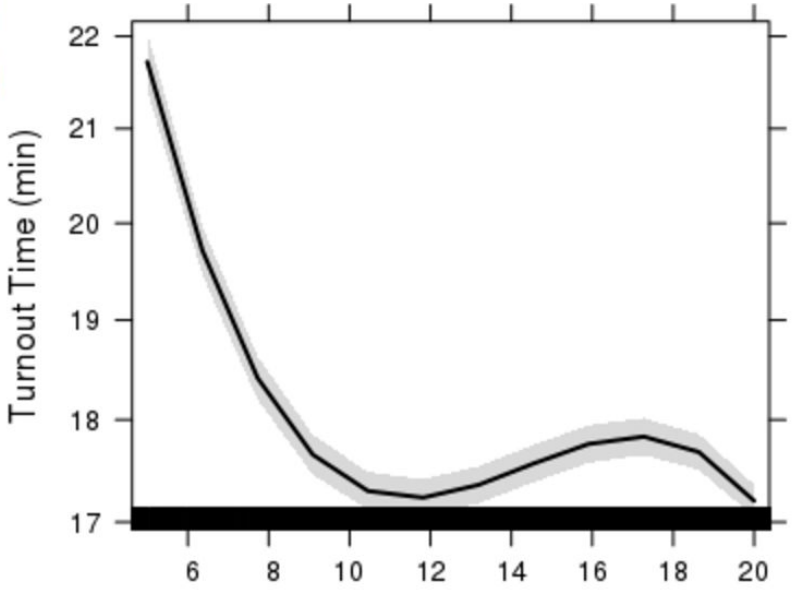
$$g() \text{ is a spline effect}$$

VirginiaTech
Biocomplexity Institute

* these predictors will be added to the model at a later time.

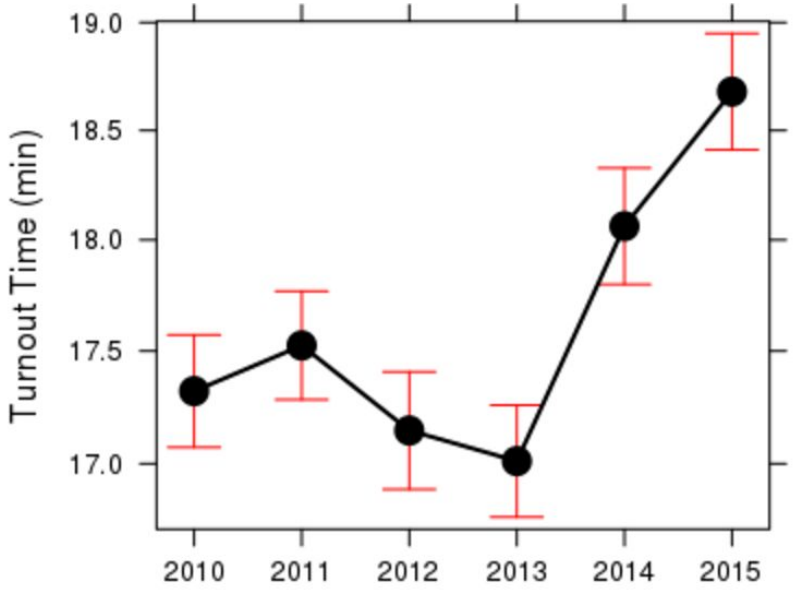
Effects of Time Predictors:

Hour effect plot (with spline)



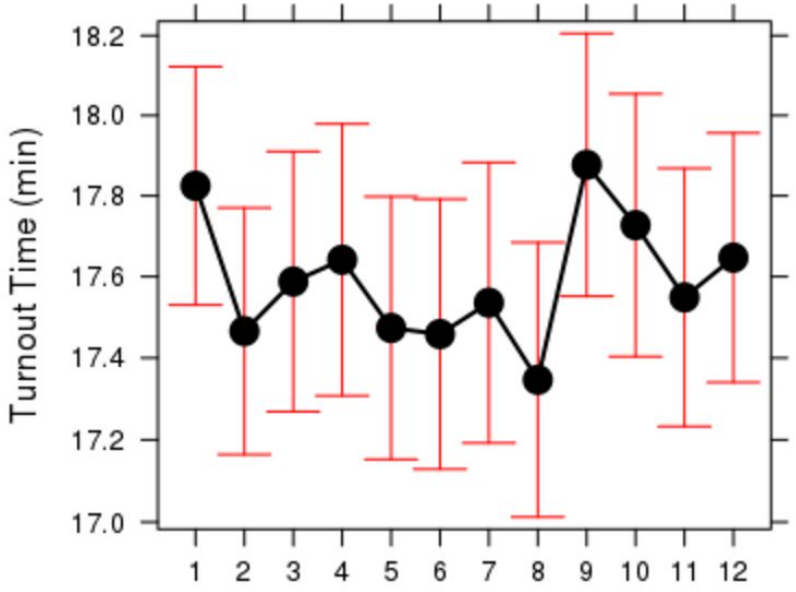
Hour

Year effect plot



Year

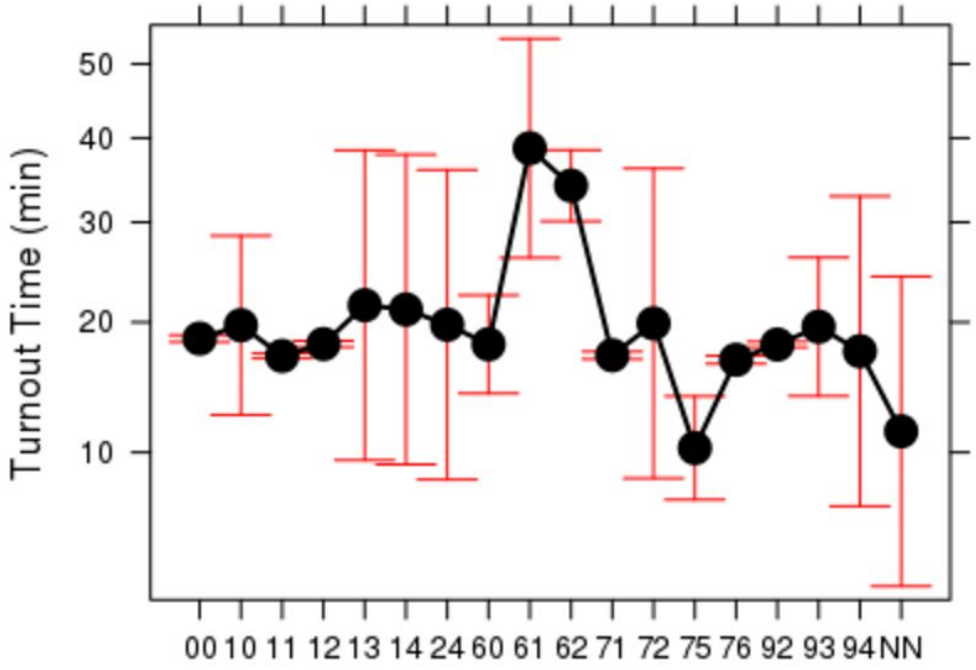
Month effect plot



Month

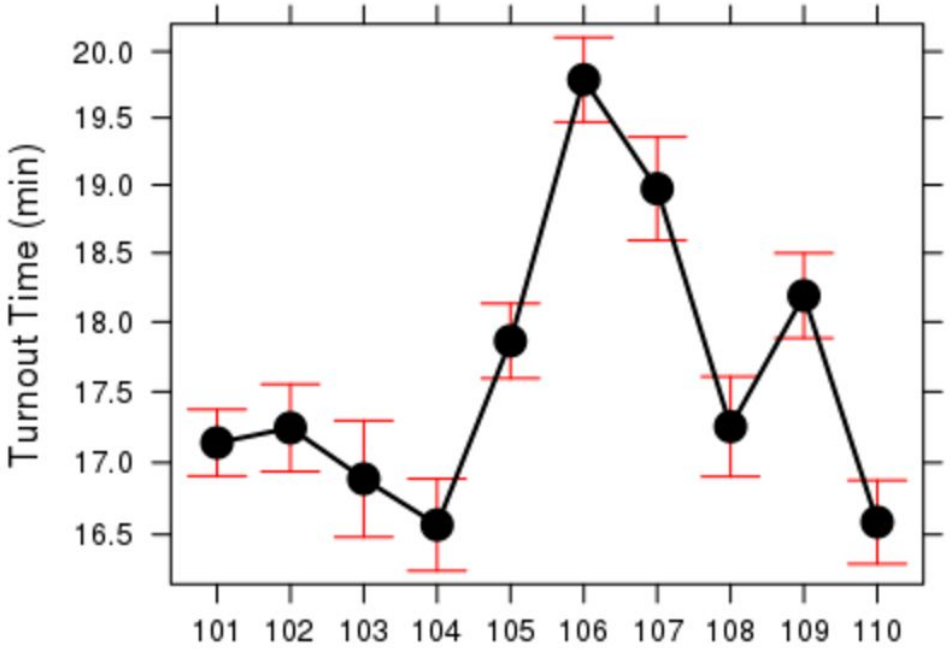
Effects of Station and Apparatus Type:

Apparatus Type effect plot



Apparatus Type

Station effect plot



Station

Future work:

- Incorporate other variables into Turnout Time model.
  - CAD Call Type
  - Number of stories per station
  - Weather
- Model other aspects of response time.
  - Processing time
  - Travel time
- Investigate CAD Call types other than SF (structure fire).

Arlington Department of Human Services (DHS)

PI & Co-PI: Aaron Schroedner
Team: Zarni Htet, Sayali Phadke
Sponsor: Arlington Department of Human Services

Virginia Tech
Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

String matching

- Distance functions
  - Compute how two strings are similar or different based on prespecified algorithm involving operations on each character

    e.g. we have a = “s phadke” and b = “sayali p”

Distance	dist(a,b)
JW	0.5
LV	7

- Types of distance functions
  - Jaro-Winkler (JW)
    - Calculates how far a character is from one string to a character in another
      - Gives more weight to the prefix of a letter
    - Good for names
  - Levenshtein Distance (LV)
    - Minimum number of operations needed to transform string ‘a’ into string ‘b’
  - Soundex
  - Hamming Distance

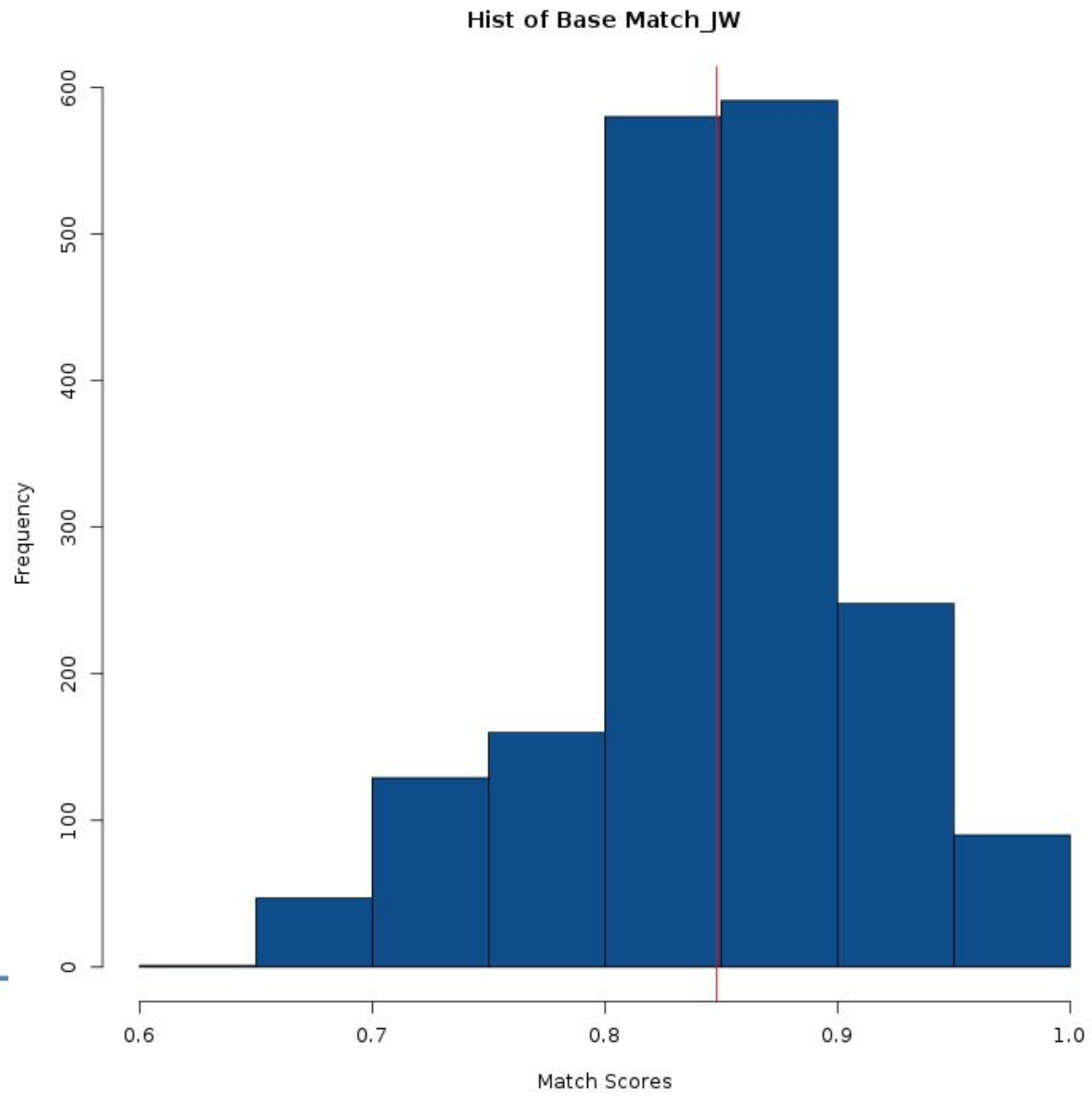
Albuquerque Police Department (APD)

- **APD** maintains arrest records/**crime description**
- Manually entered by the officer-on-duty => lack of consistency in records
- Department of Public Safety (**DPS**) maintains a **crime description** and category
- Nine unique categories of crimes
- Crime Categorization Table (**CCT**) contains **crime description**, whether it is a felony or misdemeanor, and the degree of crime
- Neither of the three crime descriptions have a sufficient overlap!
- **Overall goal:** merge the three datasets
- **Purpose:** to allow a new arrest record to be easily matched to an existing crime description, and its corresponding category, whether it is a felony or misdemeanor, and the degree of crime

APD: Method

- Compare each word from string1 (word length A) to each word in string2 (word length D)
- Threshold Jaro Winkler score (at 0.95), and create binary word-to-word matching matrix
- Find the maximum (0 or 1) for each word from string1, and sum over those to calculate N
- Apply a penalty (P) of 0.01 to the each of the unmatched words to calculate N-P
- For a given string1, pick the match from string2 with highest value of (N-P); giving a score S
- Also document a standardized version of this score calculated as S/A

APD: Histogram of Matched Scores



VirginiaTech.
Biocomplexity Institute

DHS Arlington Motivation

Name (Sys1)	SSN	DOB	Gender
Sayali Phadke	123-45-6789	10/19/1990	F
Zarni Htet	**456-78-8901**	**05/25/1991**	**F**
Aaron Schroedner	789-01-2345	03/07/1975	M
Zarni	**456-78-8901**	**05/25/1991**	**M**

Name (Sys2)	SSN	DOB	Gender	Zipcode
Sayli S Phadke	NA	10/19/1990	M	22203
Htet Zarni	NA	05/27/1991	M	22202
A Schrodner	789013445	--/07/1975	F	22203

Problem statement

- Various DHS services collect information about clients
- Data stored across nine systems
- Single system problem
  - Clerical errors (Names & Other variables being misspelt) resulting in duplicates
- Multiple system problem
  - Each service requires different information, leading to missing data in other fields
- The two problems compounded

On a given day, how many citizens did the DHS serve?

Three systems

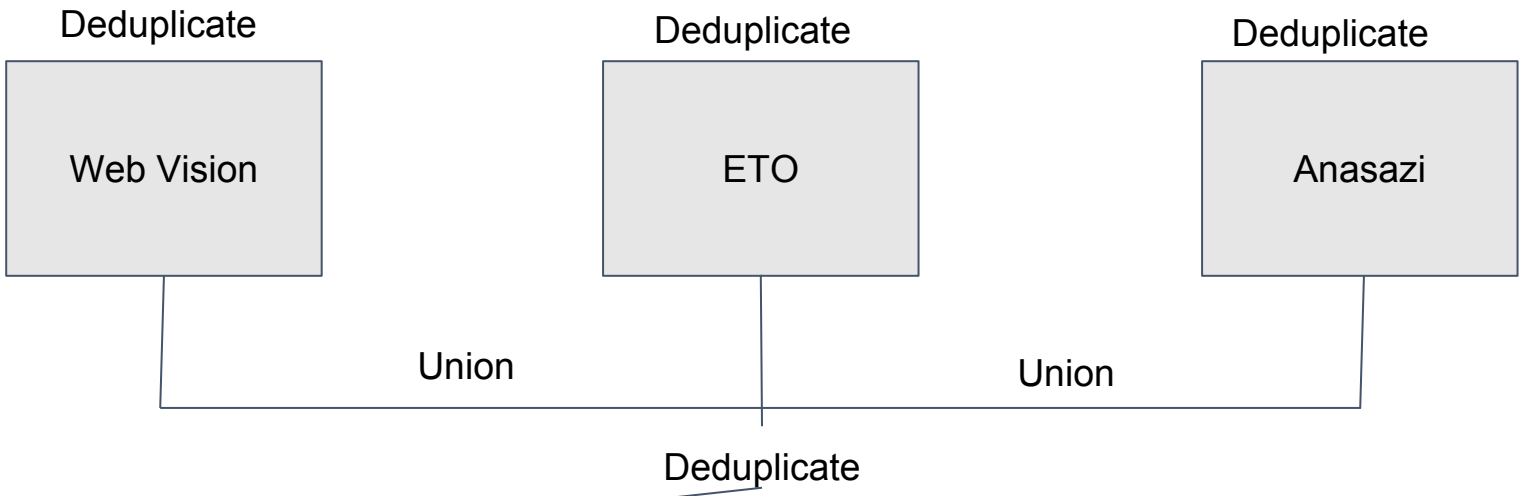
- What they are?
  - Web Vision : Behavioral Health for Children and Aging Population
  - ETO : Economic Independence Data
  - Anasazi : Community Service
- What is the problem with each of them
  - Different data requirements
    - Web Vision: No SSN requirement, Babies need not have names
  - ETO
    - Running exact deduplication on some variables
  - Lots of legacy code with not much documentation
    - Our understanding: **AND** operations on columns like first name and last name with exact matching!!!
  - Data entry error

Probabilistic linkage

- What does it mean?
  - A method to determine whether two items are the same, even if they do not match exactly due to different specification, or spelling error etc.
- M probability (quality/reliability)
  - P(two field values agree | It's a match)
    - Determines how good a record has been documented (SSN, for instance)
- U probability (commonness)
  - P(two field values agree | It's not a match)
    - Imagine a column of City. To distinguish between 2 people who live in D.C, not very helpful.
- Composite Weight Scores
  - Agreement Weights and Disagreement Weights for each fields depending on agreement or disagreement
  - Sum those weights across each row of comparison to get Composite Score
  - Positive Composite: A match, Negative Composite: A mis-match

Ref : http://www.heinz.cmu.edu/~roehrig/Record%20Linkage/Fellegi-Sunter.pdf

Approach



Intermediate Table

New ID	Old ID	Table Code
1	66	WebVision
1	77	ETO
1	88	Anasazi

Master Table(User Access Portal)

New ID	Name	DOB	SSN	……
1	Zarni	1300	333	
2	Sayali	1400	444	
3	Aaron	1500	555	

Method (contd.)

- Preprocessing: DOB split, SSN etc
- Two functions for deduplication by manually specifying M weights
  - RecordLinkage package
    - compare.dedup, compare.delink ⇒ Generate match pairs (expand.grid)
      - Single data set, Pair of data set
      - Parameters
        - strcomp: Jaro Winkler (default)
        - block field: eg. (first name, last name)
    - pairs$pairs (From Webvision)

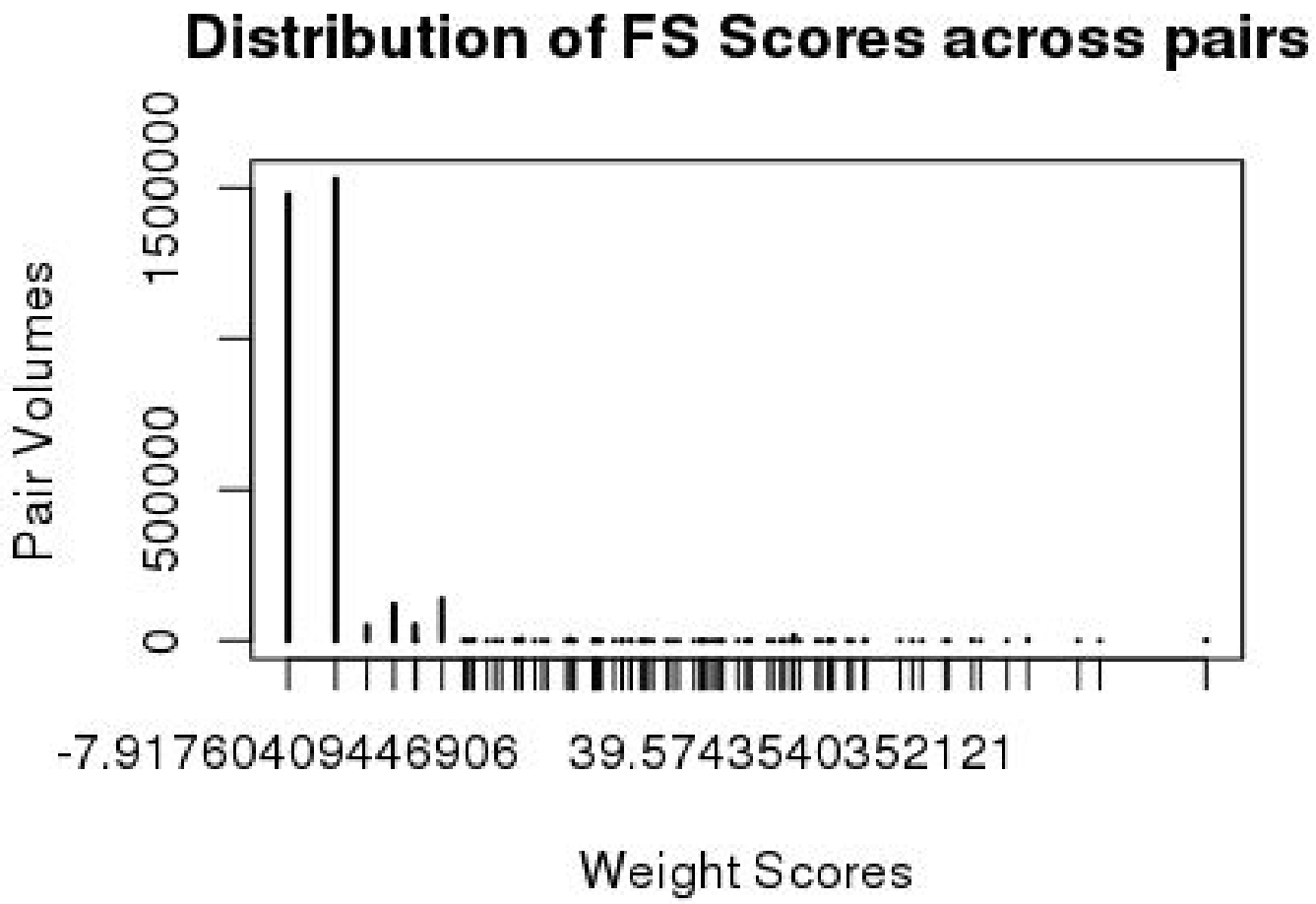
	id1	id2	last_name	first_name	gender	ssn	street	zip	birth_yr	birth_mo
1	1	2	1.0000000	1.0000000	1	NA	1.0000000	1.0000000	1	1
2	1	1237	0.5555556	0.5000000	0	NA	0.4810458	0.6400000	1	1
3	1	2008	0.6357143	0.0000000	0	NA	0.0000000	0.9200000	1	1
4	1	2359	0.4537037	0.5119048	1	NA	0.0000000	0.9200000	1	1

Method (contd.)

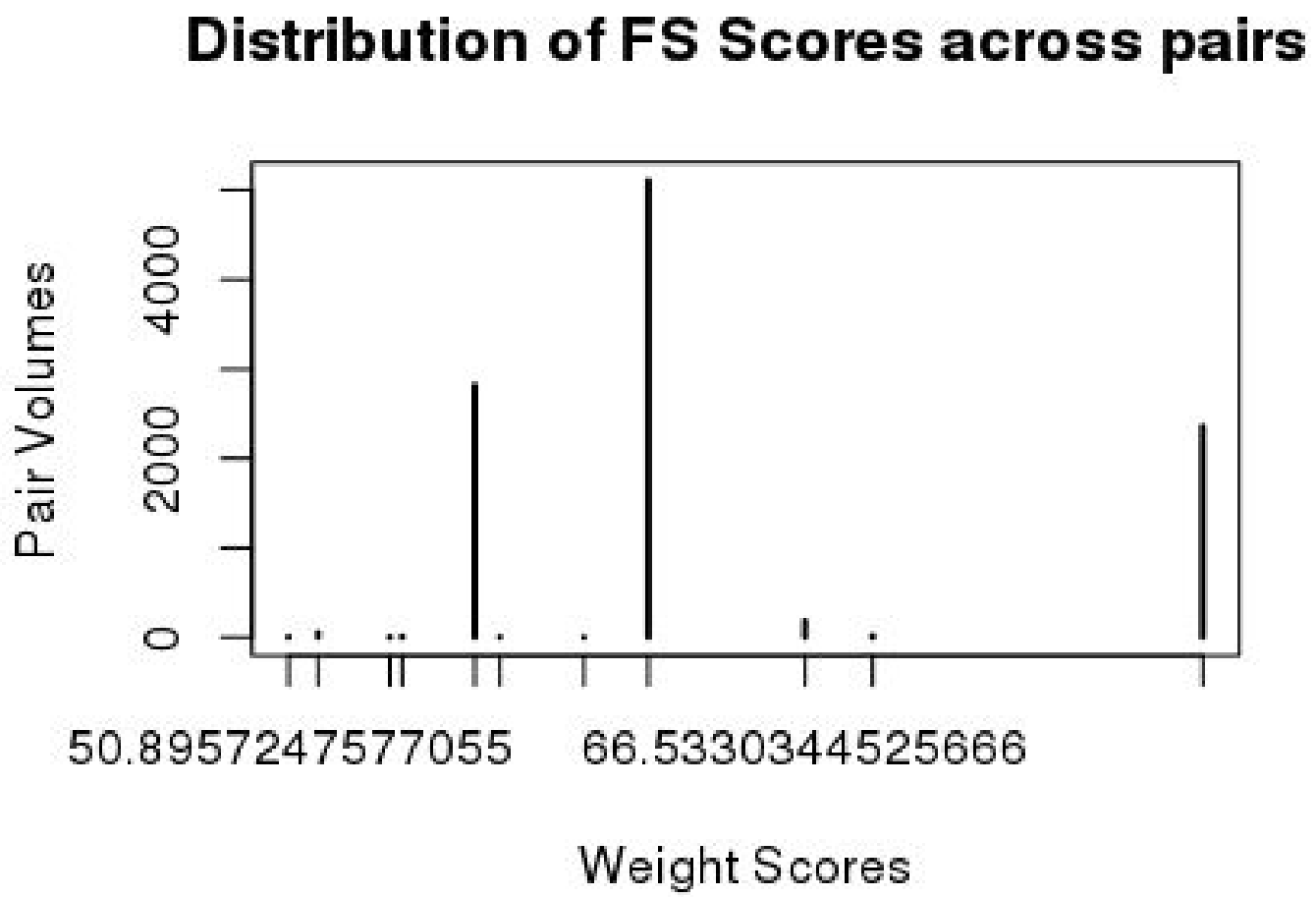
- Jaro Winkler threshold for pairs$pairs
  - Columnwise - 0, Rowwise: Avg JW score - 0.25
- Weights applied
  - m probability weights from Aaron's experience
- Supply into Fellegi & Sunter function ⇒ get composite scores
- Clerical Review (basically eyeball for clusters)

Weight Score	Volume	Weight Score	Volume
78.57	2363	56.53	2819
61.78	5099	51.79	54

Method (contd.)

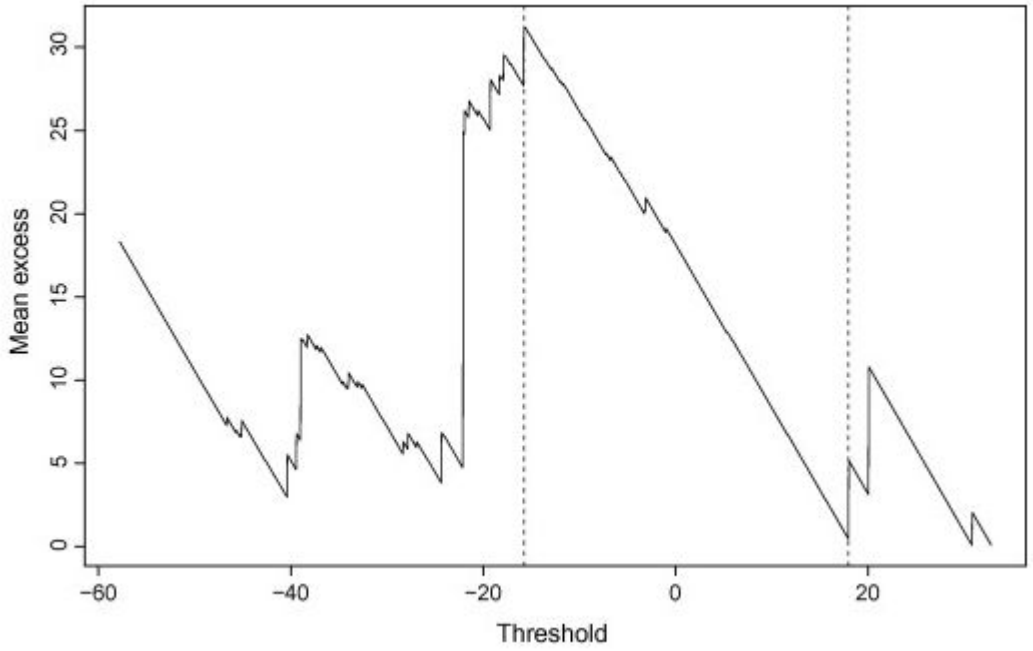


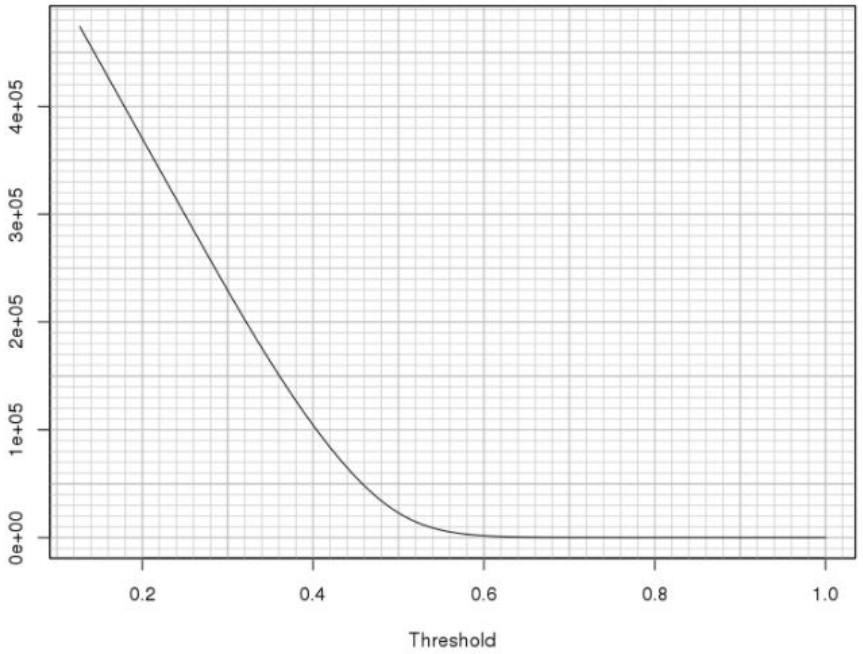
Method (contd.)



Automatic Weights

- Automatically figuring out weights
  - Epi weights Current problem
    - a weight calculation for each pair built from Epilink calculation
    - run pareto threshold: Will give us an interactive way to grab thresholds!





Ref: http://www.sciencedirect.com/science/article/pii/S1532046411000372?via%3Dihub#s0030

VirginiaTech.
Biocomplexity Institute

Automatic Weights(contd.)

- EM algorithm
  - Used to estimate the values required for weight calculation
  - E-step estimates whether the pair of records is actually the same
  - M-step updates conditional probabilities for computing weights of fields
  - Further reading underway!

Next steps

- Investigate whether epiWeights or EMWeights are understandable and computationally feasibly option
- Apply the method to ETO, Anasazi, and WebVision to check performance
- Fine-tune parameters to create best possible matches
- Package final method code in a way that the DHS team can implement it on all systems

SCHEV

Midterm Presentation

DSPG: Sean Pilli, Kyle Morgan, Ronnie Fesco, Lata Kodali

PI: Aaron Schroeder

VirginiaTech Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Project Description

- **Sponsor**: State Council of Higher Education for Virginia
- **SCHEV Representative:** Todd Masa, Director, Policy Analytics & VLDS Lead
- **Problem:** Virginia Longitudinal Data System (VLDS) contains sensitive information that cannot be released.
  - Researchers are interested in exploring these data, but they are typically denied access.
  - These data contain many variables, and machine learning methods can be used to determine important ones.
- **Solution:** Create synthetic VLDS data that is privatized and usable for public.

Project Goals

Research Goal: What qualifies as useable and what is privatized enough for generated synthetic data?

- **Usability:** Are relationships preserved between original and synthetic data?
  - Compare using regression models and machine learning.
- **Privacy:** How vulnerable are certain individuals to re-identification?
  - Use techniques such as smoothing or top and bottom coding.
  - Note exact replicates and determine vulnerable demographics/ unique demographics combinations.

Synthpop Overview

- Synthpop package in R utilizes several techniques to create synthetic data and disclosure.
- Features of the package include:
  - Creation of synthetic data using `syn()`.
  - Compare original to synthetic using `compare()`.
  - Compare regression models using `compare.fit.synds()`.
  - Evaluate or ensure privacy using `sdc()`.

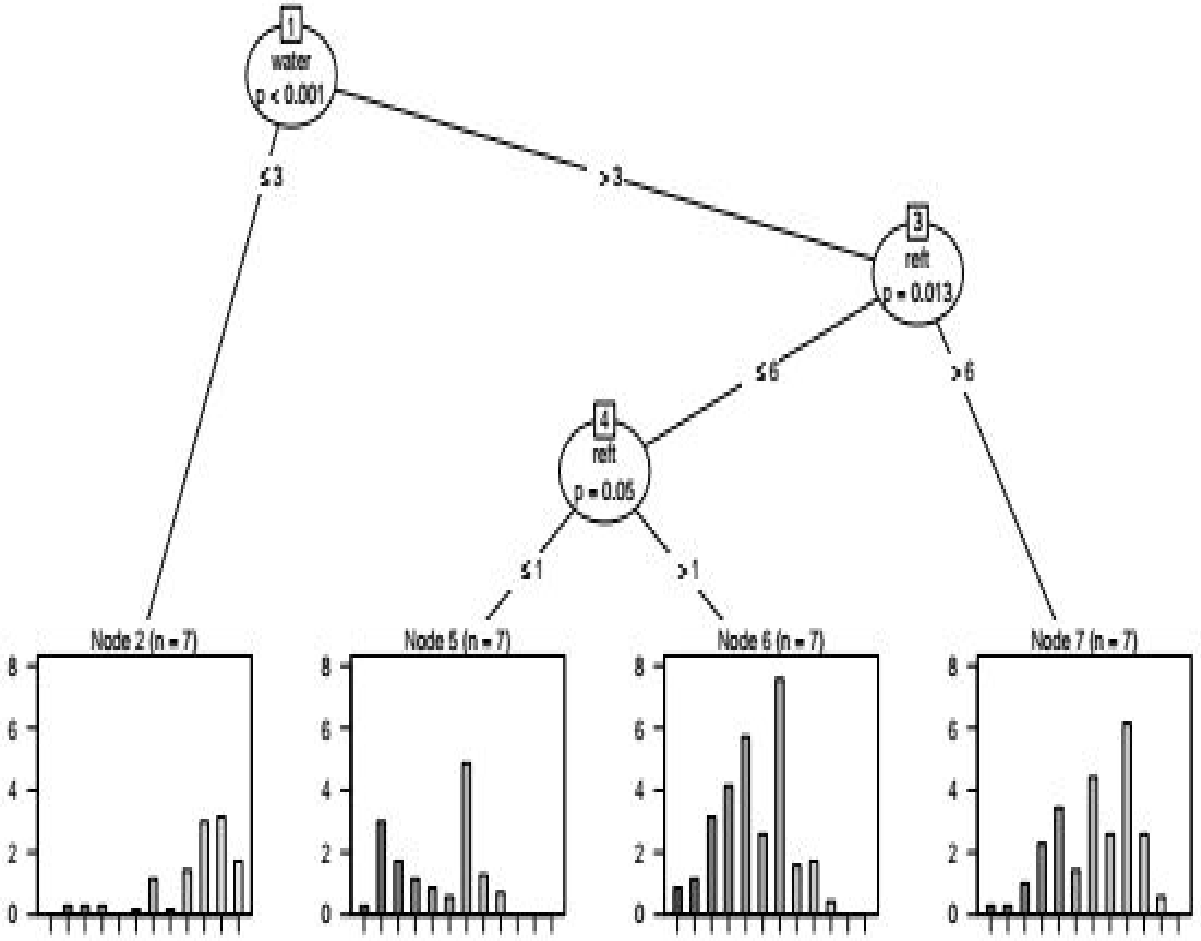
Methods for Synthetic Generation

Synthpop includes many methods:

- Sample
- CART
  - Classification and Regression Trees
- Parametric
  - Norm
  - Logreg
- Ctree
  - Conditional Inference Tree
  - Reduces selection bias

After exhaustive comparisons, decided on Ctree method.

Ctree Diagram:



VirginiaTech
Biocomplexity Institute

VLDS Sample Dataset

	unique_id	birth_year	gender	race_type	aggregate_days_present
1	HKKUQQQ	2002	1	3	0
2	VXY8AQQ	1993	1	3	9
3	MX6JQQQ	2004	1	5	0
4	SRYZAQQ	2005	1	3	169
5	5KRFZQQ	1990	1	5	102
6	FX3KQQQ	2000	1	3	NA
7	8RRSQQQ	2001	0	3	131
8	U8Y7QQQ	1999	0	5	170
9	F2B5QQQ	2000	1	3	165
10	PH2WQQQ	2002	1	3	91
11	6XVCQQQ	2001	1	3	NA
12	N75XQQQ	2001	0	3	0
13	4ZHVAQQ	1996	1	3	177
14	L6LHAQQ	2000	1	3	128
15	ZD7EAQQ	1997	0	3	180
16	S8HVAQQ	1996	0	3	NA
17	OXBYAQQ	1993	1	3	0

Showing 1 to 17 of 1,000 entries

Dataset with 1000 observations.

Birth Year	1980s	1990s	2000s
Count	25	507	468

Gender	Male	Female
Count	517	483

Race Type	0	1	2	3	4	5	6	99
Count	1	2	8	719	23	228	2	17

ADP: ranges from 0-181 with 137 NA's

VirginiaTech
Biocomplexity Institute

Usability

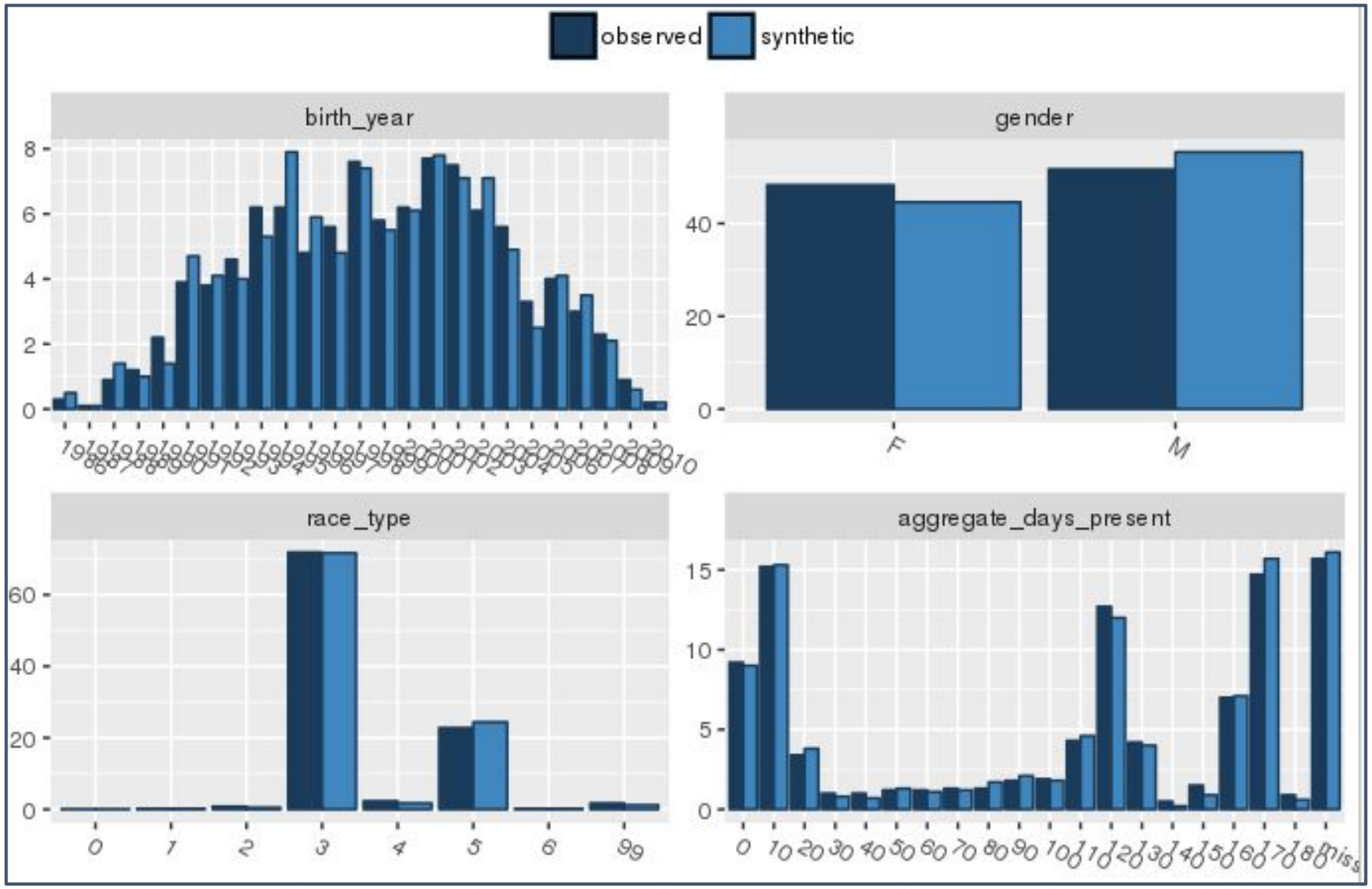
Preserving **relationships** are between original and synthetic data

- Regression models: statistical properties such as model estimates and inferences
- Machine learning: unique demographic combinations preservations

Original (left) vs. Synthetic via Ctree (right)

	birth_year	gender	race_type	aggregate_days_present	birth_year	gender	race_type	aggregate_days_present
1	2002	M	3	0	1997	F	3	0
2	1993	M	3	9	2002	M	5	NA
3	2004	M	5	0	1997	M	5	112
4	2005	M	3	169	2001	F	3	0
5	1990	M	5	102	1993	M	3	102

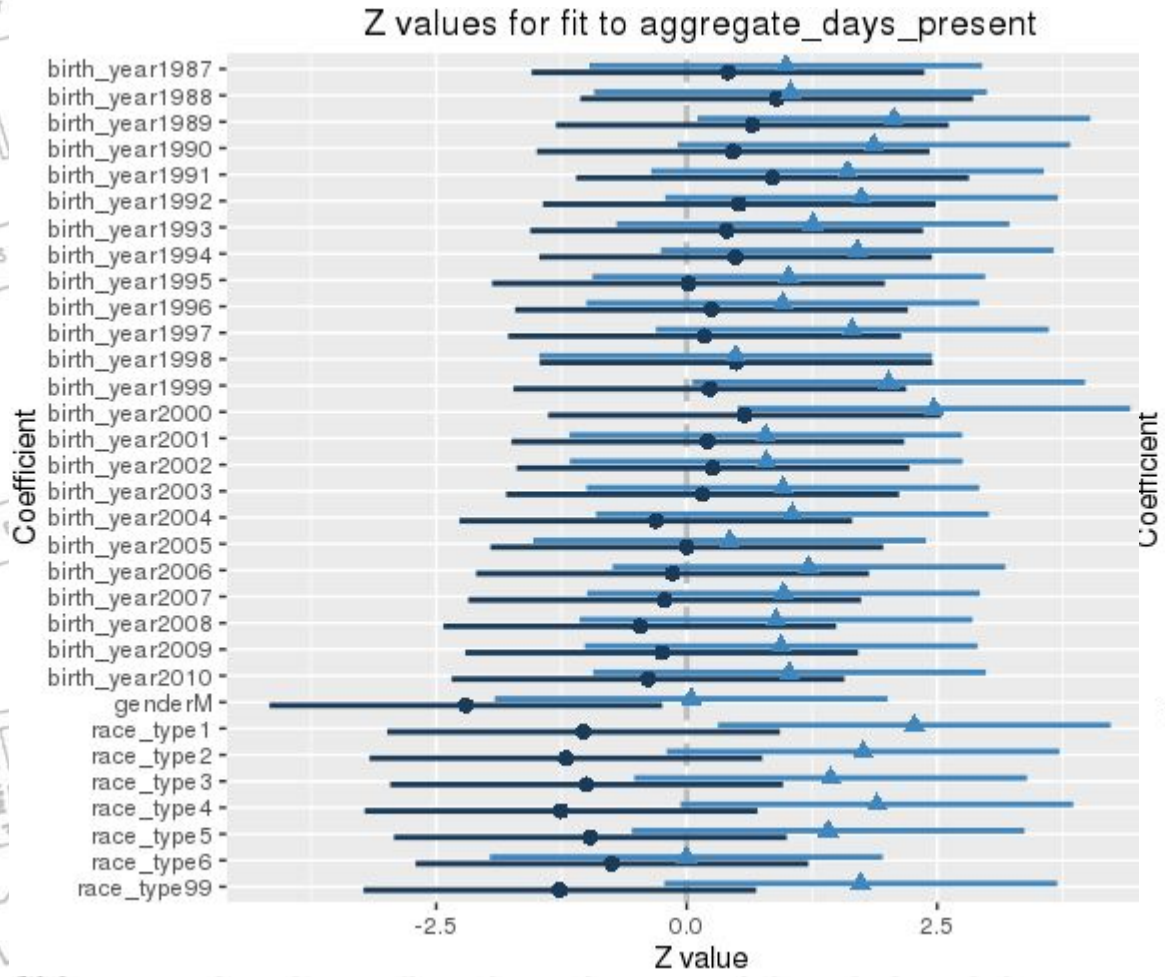
Comparison of Original vs. Synthetic



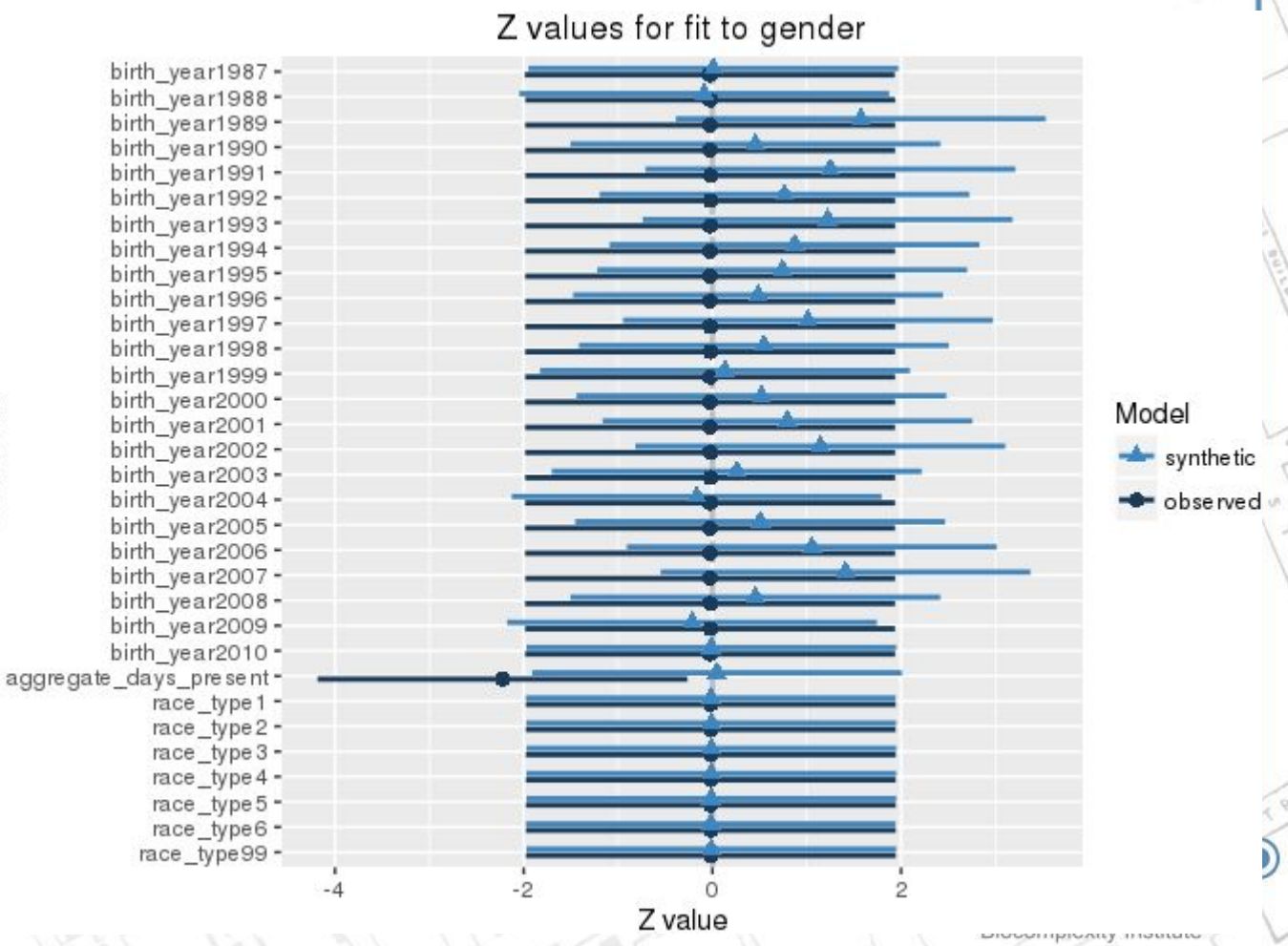
VirginiaTech
Biocomplexity Institute

Regression Models

Regressing on aggregrate_days_present:



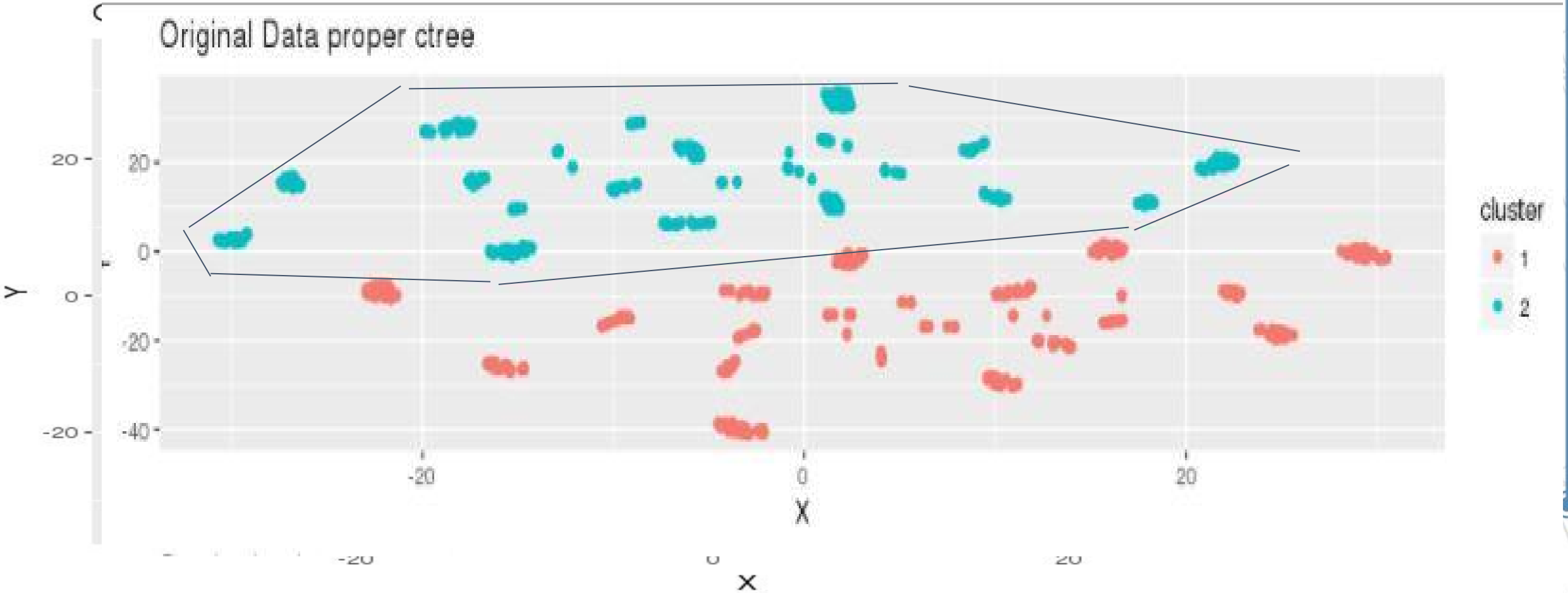
Regressing on gender:



Machine Learning

Partitioning Around Medoids (PAM) Clustering with Gower Distances (incorporates categorical variables) .

Essentially K-Means but the centers are physical data-points.



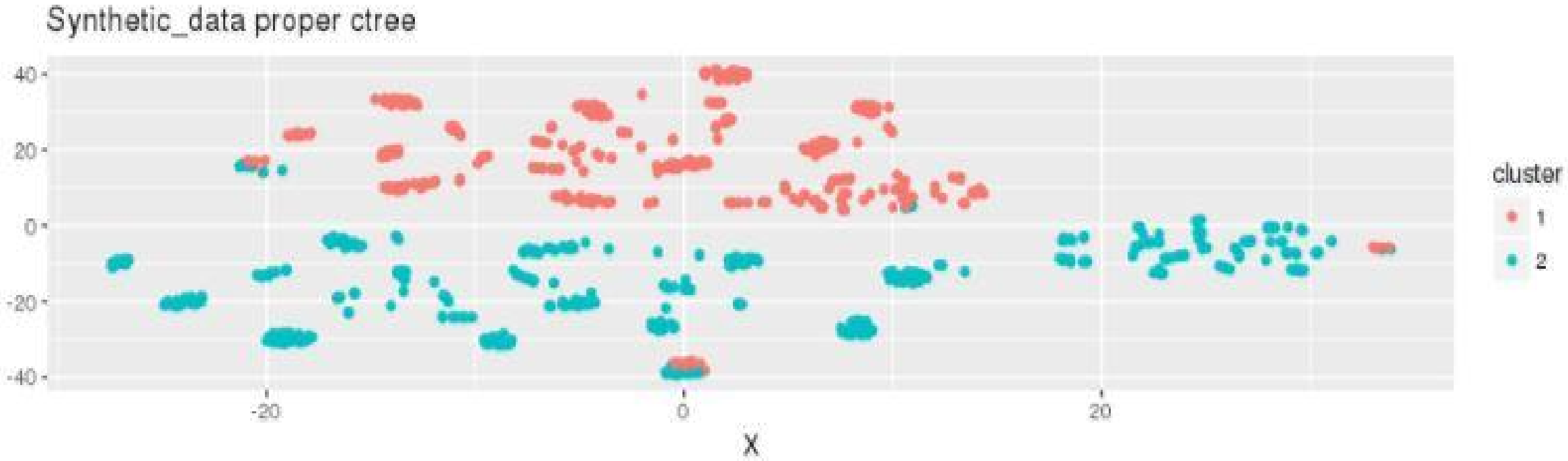
Machine Learning

How do I interpret axes in higher than 2D?

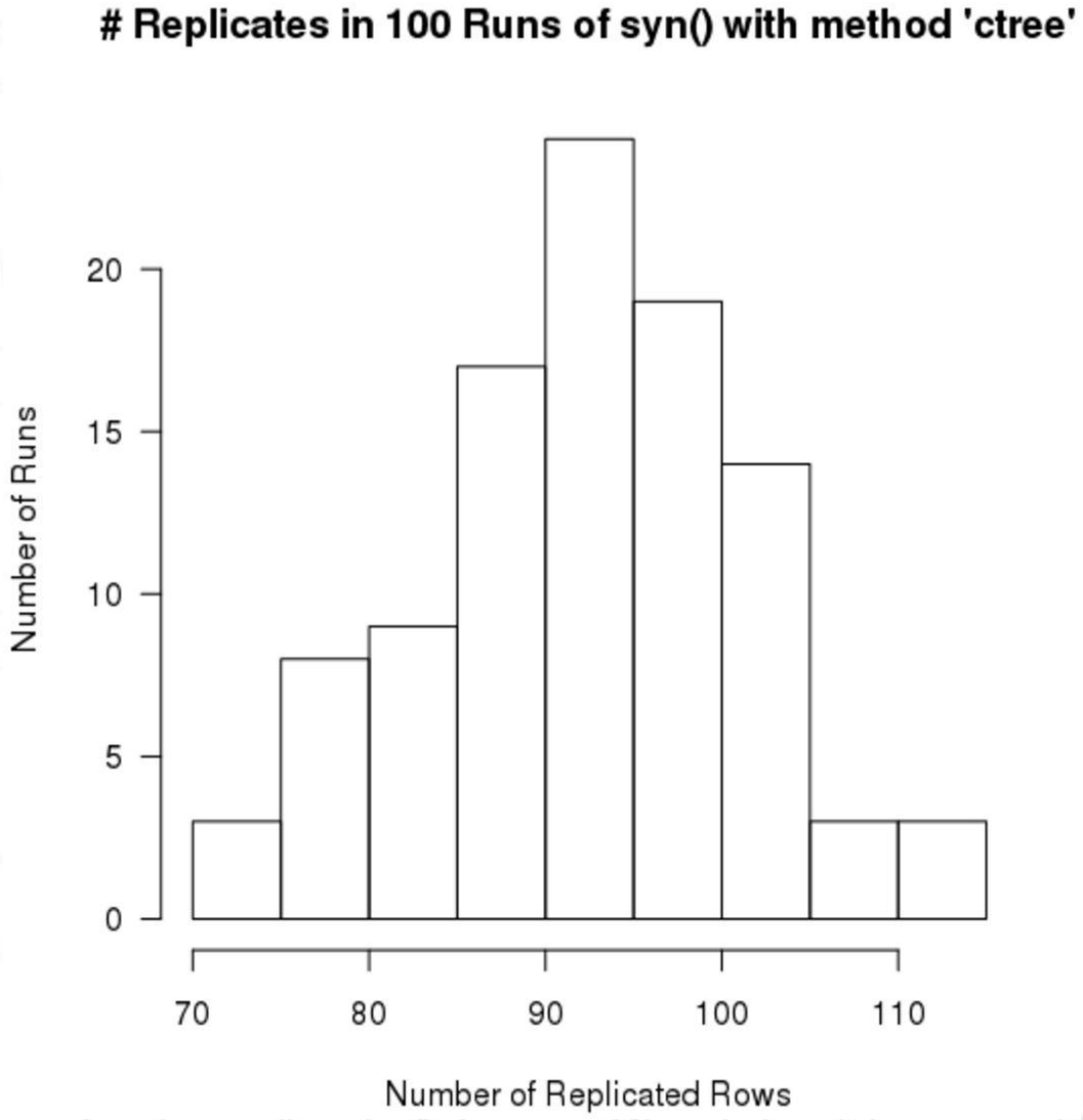
- Planning on Graphing 2 numeric variables at a time via a scatterplot and allowing their colors to be defined by their cluster identity.

Determining Usability:

- PAM algorithm in R allows us to save the clusters as data frames so we intend to find clusters that have similar frequencies of unique demographic combinations in the dataset.



Data Privacy



Method:

- Run syn() to create synthetic data
- Use sdc() to detect and remove any rows of data that are exact replicates
- Repeat 100 times

Out of 1,000 rows of data, an exact copy of a row showed up an average of 94 times in the 1,000-row synthetic data

Data Privacy

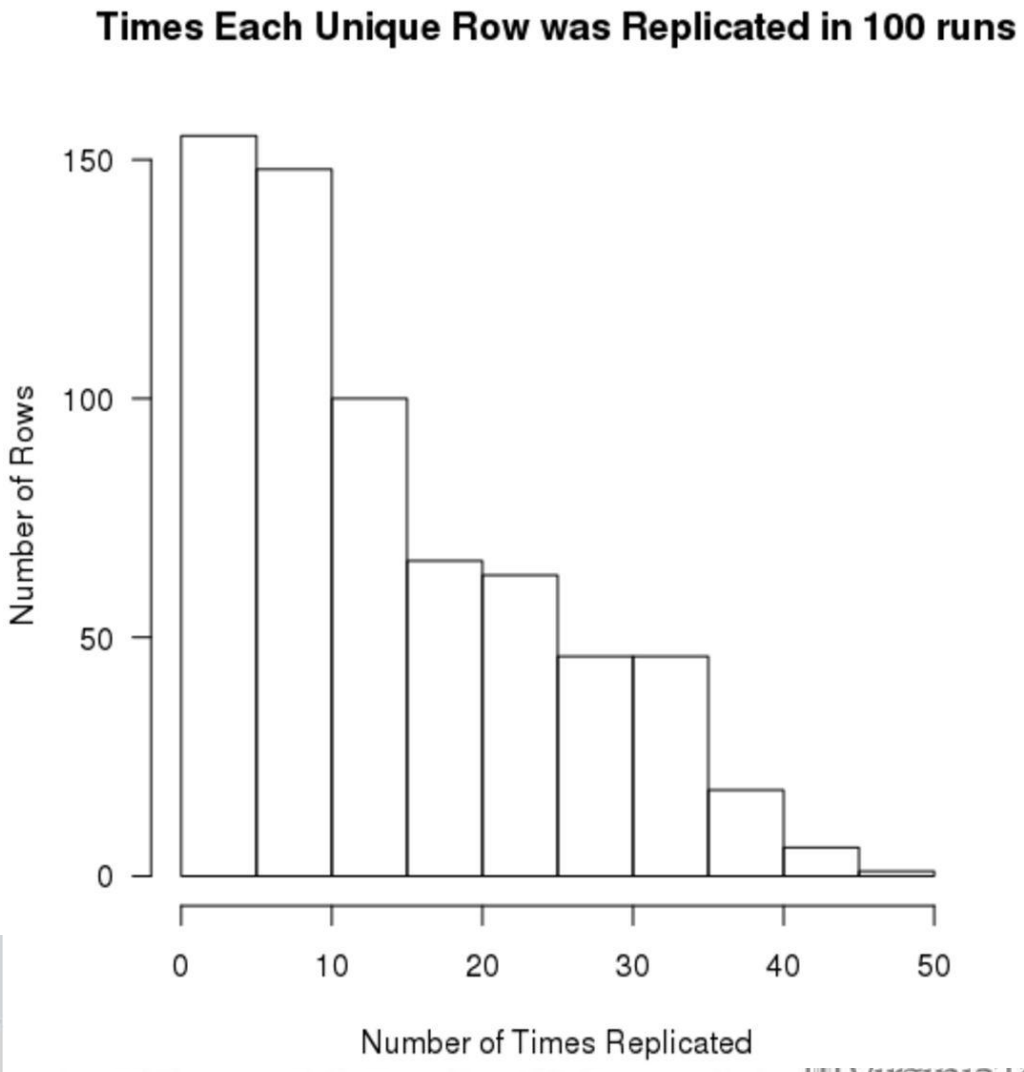
Method:

- Run syn() to create synthetic data
- Use sdc() to detect and remove any rows of data that are exact replicates
- Repeat 100 times

There were a total of 649 unique rows of data that showed up unchanged in the synthetic data at least once.

Of these, the maximum was 46 of 100 runs.

birth_year	gender	race_type	aggregate_days_present	Number_Reps
1993	F	5	NA	46



Virginia Tech
Biocomplexity Institute

Next Steps

- Repeat everything using a larger VLDS dataset.
- Machine Learning:
    - Compare clusters from original and synthetic to see if demographics are preserved.
- Data Privacy:
    - Use top and bottom coding to remove outliers that have high risk of disclosure
    - Evaluate whether smoothing numerical variables in the syn() function is better

Comments

Thank you!

Questions?

Suggestions?

Thoughts?