CLD3 Processes, Platforms & Examples

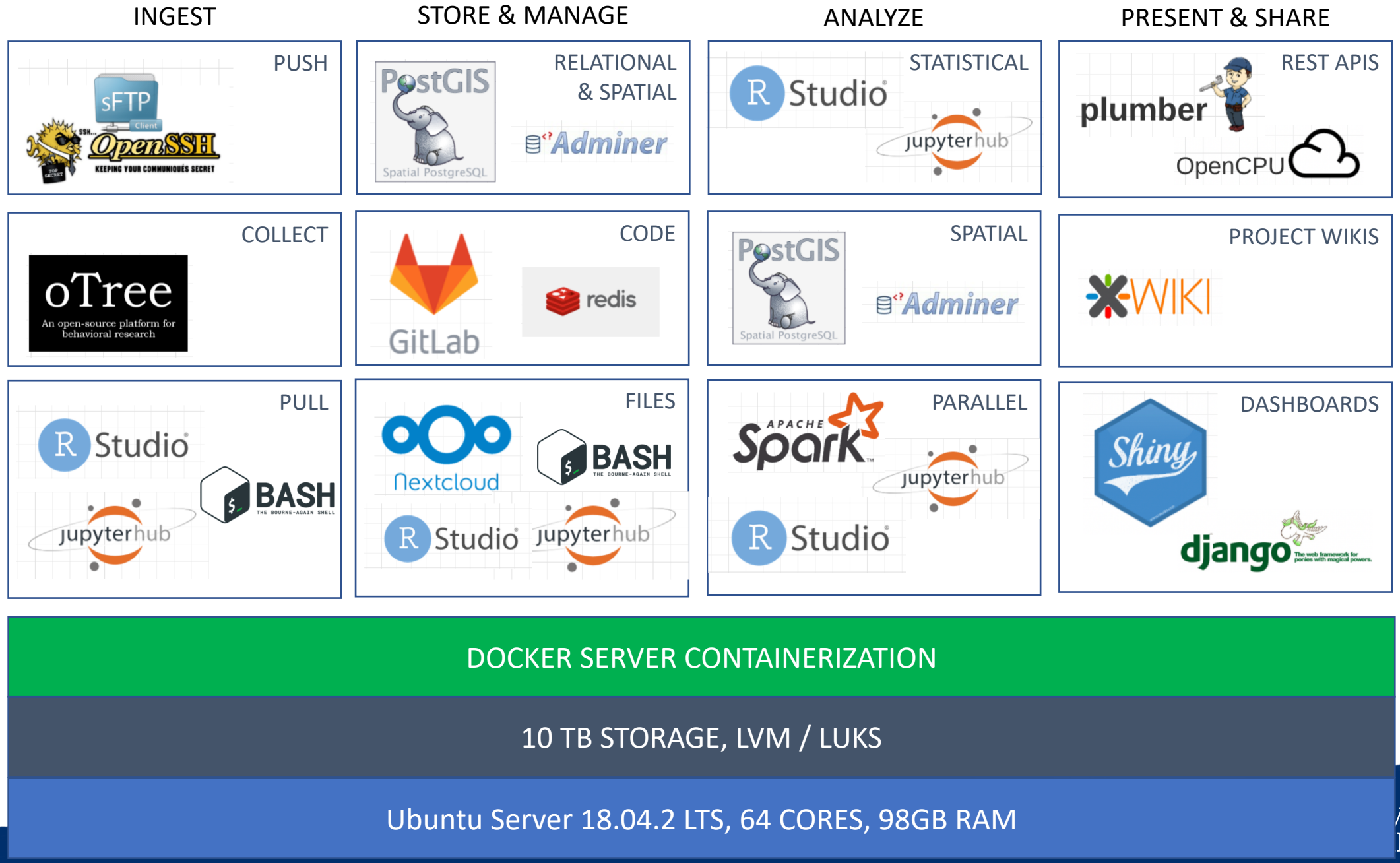
Aaron Schroeder, Research Associate Professor
Social and Decision Analytics Division

CLD3 Workshop
March 25-26, 2019

Biocomplexity Institute & Initiative

CLD3 Processes, Platforms & Examples

- Introduce some of our platforms as they are used in our data science process
- Give some reasons why these platforms were selected
- Show some examples that highlight the use of these platforms at different stages of the process
  - **Including an example of finding data and creating decision making tools for policy makers in Marshalltown**
  - Scenario: Visualizations for Policy Makers to Assess Possible Transit Route Changes
  - Demonstrate
    - Ingestion of various administrative data sources using multiple method
      - job locations
      - vulnerable population locations
      - transit route locations
    - Creation of statistical indexes and maps for sub-county geographies


INGEST
STORE & MANAGE
ANALYZE
PRESENT & SHARE
PUSH
sFTP
Client
SSH...
OpenSSH
KEEPING YOUR COMMUNIQUÉS SECRET
RELATIONAL & SPATIAL
PostGIS
Spatial PostgreSQL
Adminer
STATISTICAL
RStudio
jupyterhub
REST APIS
plumber
OpenCPU
COLLECT
oTree
An open-source platform for behavioral research
CODE
GitLab
redis
SPATIAL
PostGIS
Spatial PostgreSQL
Adminer
PROJECT WIKIS
XWIKI
PULL
RStudio
BASH
THE BOURNE-AGAIN SHELL
jupyterhub
FILES
Nextcloud
BASH
THE BOURNE-AGAIN SHELL
RStudio
jupyterhub
PARALLEL
APACHE
Spark
jupyterhub
RStudio
DASHBOARDS
Shiny
django
The web framework for perfectionists with deadlines.
DOCKER SERVER CONTAINERIZATION
10 TB STORAGE, LVM / LUKS
Ubuntu Server 18.04.2 LTS, 64 CORES, 98GB RAM

VPS + Containerization = Data Ingestion Versatility!

- Simple Parallel
- Simple Additional IP Addresses (for cloud VPS services – use multiple NICs if running your own iron)
- Recent example: NSF project studying the value of Open Source Software
  - Needed many R installations to download just about every R and Python repository on Github!
  - Needed to determine the license being used on each using an online service licenses.io that limits the rate of use by your IP address. So needed multiple IP addresses

Virtual Private Server
(Cloud or Yours)

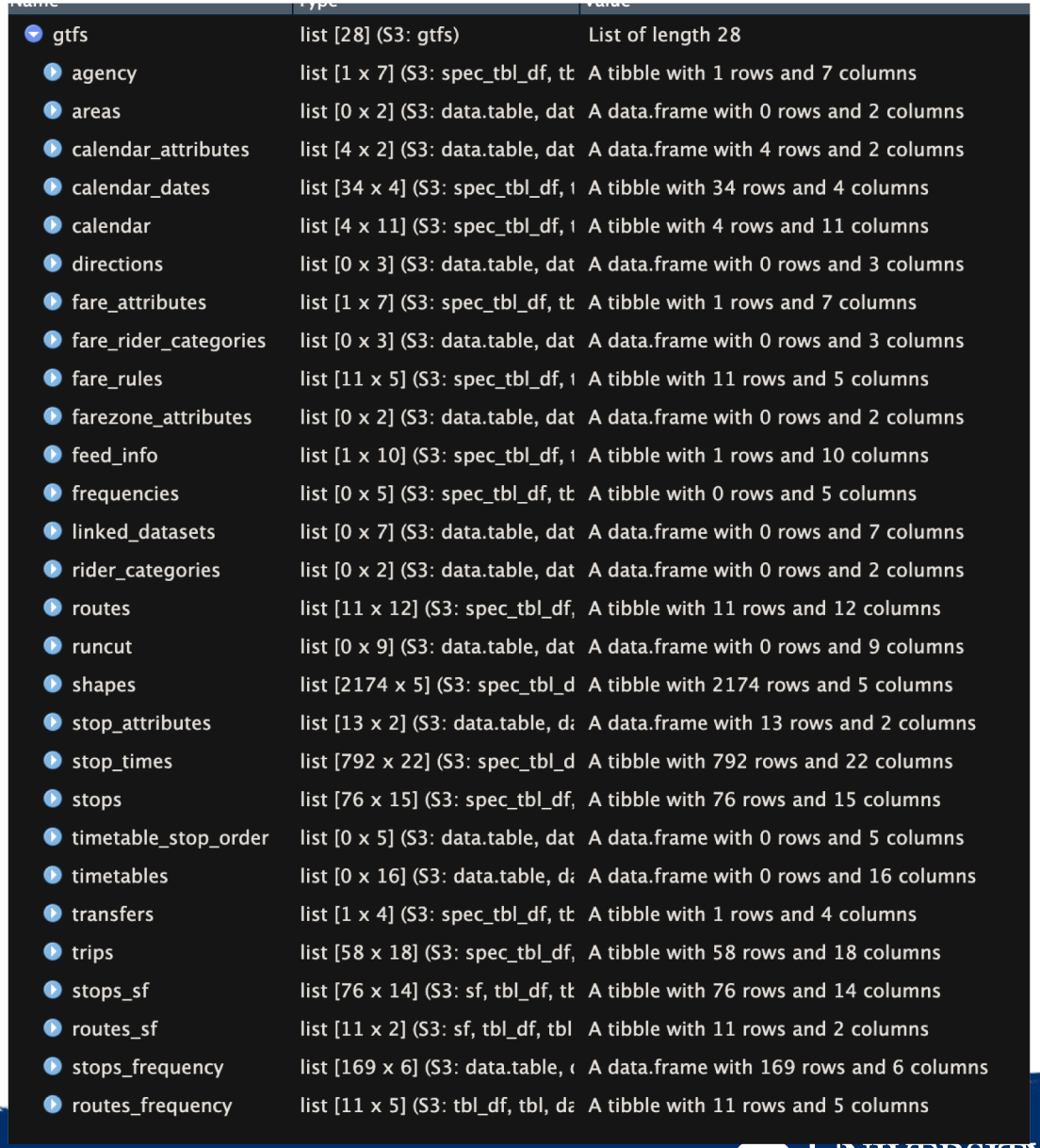
Data Ingestion

- Establish **type and method of data transfer**
  - pushed to or pulled into the cooperative platform?
  - staying where it is and being dynamically queried in a federated manner as needed?
- Establish the **best transfer protocol(s)** to use given the types and method of transfer
  - e.g. SFTP, secure Dropbox, secured REST API, VT SAFR-Data Adapter for secure federated queries
  - Establish designed collection systems (e.g. behavioral experiments)
- Establish **data marshaling processes**
  - system mediation logic, data pipeline and data transformation, transfer schedule, and data provenance maintenance
- Establish **secure data storage procedures**
  - e.g. each project being stored on a new project-dedicated encrypted partition, original data being stored as non-removable and non-editable

Establish Type of Access, Select Method, & Deploy Platform			
Type	Warehouse		Federate
Method	Push	Pull	Leave
Protocol	SFTP, Secure Dropbox	Provider API, WebDAV-HTTPS, Scraping	Provider API, SAFR-Data Adapter
Platform	Apache NiFi		NiFi, SAFR-Data Shaker
Storage	Linux Unified Key Setup (LUKS)		
	Logical Volume Management (LVM) Partition		

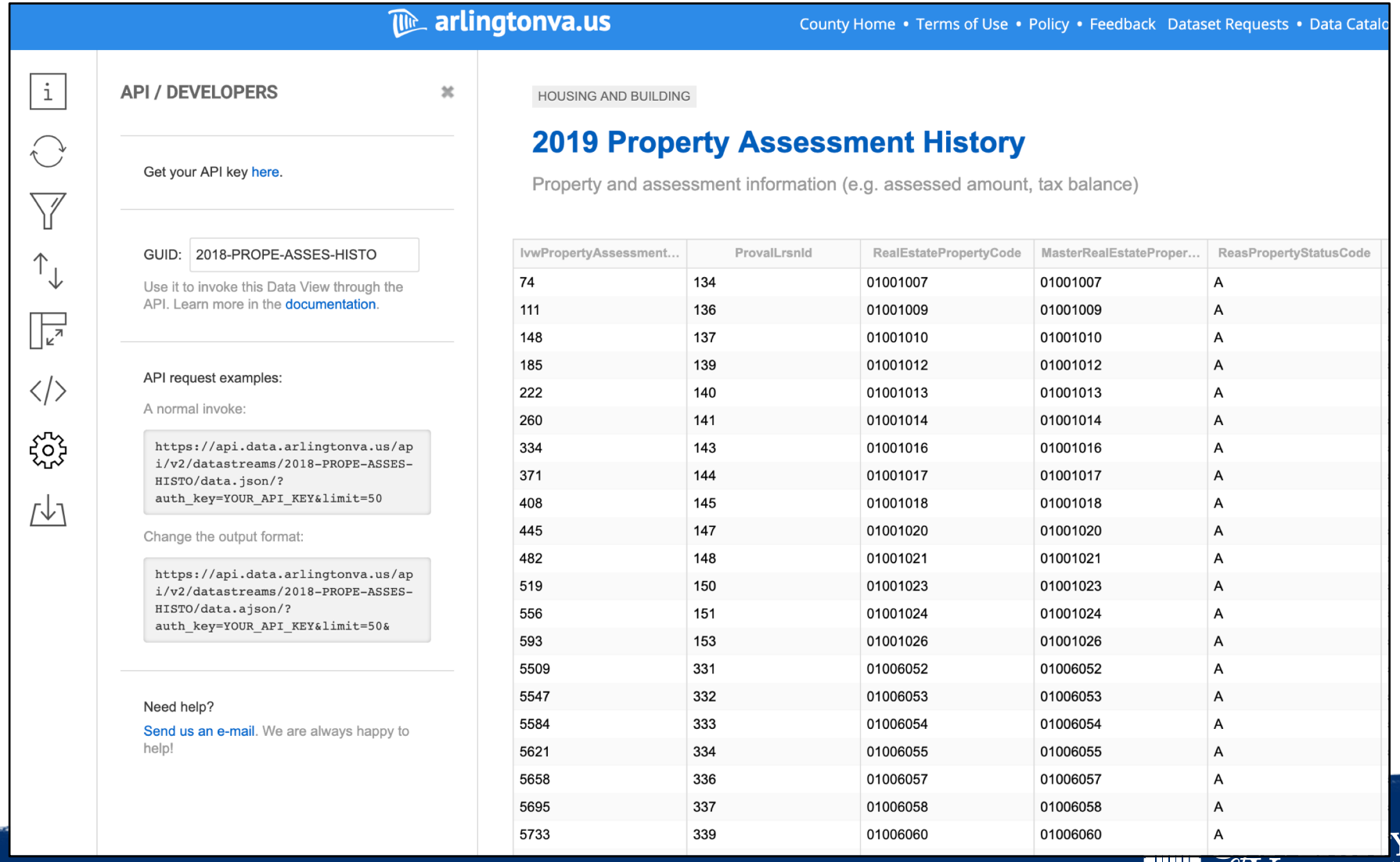
Data Ingestion
Libraries/Packages

```
library(tidytransit)

# get Marshalltown feed URL
feed_url <- feedlist_df %>%
  setDT(.) %>%
  .[loc_t %like% "Marshalltown, IA", url_i]

# read gtfs data from url
gtfs <- read_gtfs(feed_url, geometry = TRUE, frequency =
TRUE)
```



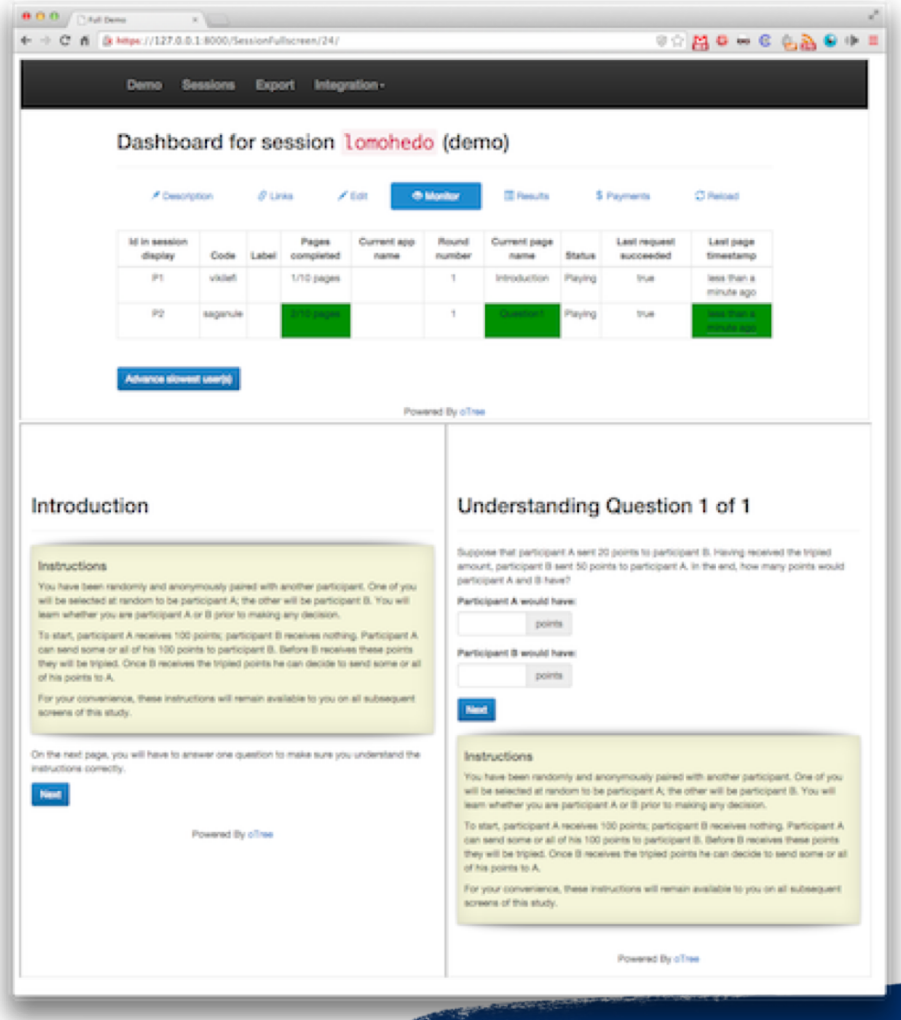
gtfs	list [28] (S3: gtfs)	List of length 28
agency	list [1 x 7] (S3: spec_tbl_df, tb	A tibble with 1 rows and 7 columns
areas	list [0 x 2] (S3: data.table, dat	A data.frame with 0 rows and 2 columns
calendar_attributes	list [4 x 2] (S3: data.table, dat	A data.frame with 4 rows and 2 columns
calendar_dates	list [34 x 4] (S3: spec_tbl_df, t	A tibble with 34 rows and 4 columns
calendar	list [4 x 11] (S3: spec_tbl_df, t	A tibble with 4 rows and 11 columns
directions	list [0 x 3] (S3: data.table, dat	A data.frame with 0 rows and 3 columns
fare_attributes	list [1 x 7] (S3: spec_tbl_df, tb	A tibble with 1 rows and 7 columns
fare_rider_categories	list [0 x 3] (S3: data.table, dat	A data.frame with 0 rows and 3 columns
fare_rules	list [11 x 5] (S3: spec_tbl_df, t	A tibble with 11 rows and 5 columns
farezone_attributes	list [0 x 2] (S3: data.table, dat	A data.frame with 0 rows and 2 columns
feed_info	list [1 x 10] (S3: spec_tbl_df, t	A tibble with 1 rows and 10 columns
frequencies	list [0 x 5] (S3: spec_tbl_df, tb	A tibble with 0 rows and 5 columns
linked_datasets	list [0 x 7] (S3: data.table, dat	A data.frame with 0 rows and 7 columns
rider_categories	list [0 x 2] (S3: data.table, dat	A data.frame with 0 rows and 2 columns
routes	list [11 x 12] (S3: spec_tbl_df,	A tibble with 11 rows and 12 columns
runcut	list [0 x 9] (S3: data.table, dat	A data.frame with 0 rows and 9 columns
shapes	list [2174 x 5] (S3: spec_tbl_d	A tibble with 2174 rows and 5 columns
stop_attributes	list [13 x 2] (S3: data.table, da	A data.frame with 13 rows and 2 columns
stop_times	list [792 x 22] (S3: spec_tbl_d	A tibble with 792 rows and 22 columns
stops	list [76 x 15] (S3: spec_tbl_df,	A tibble with 76 rows and 15 columns
timetable_stop_order	list [0 x 5] (S3: data.table, dat	A data.frame with 0 rows and 5 columns
timetables	list [0 x 16] (S3: data.table, da	A data.frame with 0 rows and 16 columns
transfers	list [1 x 4] (S3: spec_tbl_df, tb	A tibble with 1 rows and 4 columns
trips	list [58 x 18] (S3: spec_tbl_df,	A tibble with 58 rows and 18 columns
stops_sf	list [76 x 14] (S3: sf, tbl_df, tb	A tibble with 76 rows and 14 columns
routes_sf	list [11 x 2] (S3: sf, tbl_df, tbl	A tibble with 11 rows and 2 columns
stops_frequency	list [169 x 6] (S3: data.table,	A data.frame with 169 rows and 6 columns
routes_frequency	list [11 x 5] (S3: tbl_df, tbl, da	A tibble with 11 rows and 5 columns

Data Ingestion APIs



lvwPropertyAssessment...	ProvalLrsnId	RealEstatePropertyCode	MasterRealEstateProper...	ReasPropertyStatusCode
74	134	01001007	01001007	A
111	136	01001009	01001009	A
148	137	01001010	01001010	A
185	139	01001012	01001012	A
222	140	01001013	01001013	A
260	141	01001014	01001014	A
334	143	01001016	01001016	A
371	144	01001017	01001017	A
408	145	01001018	01001018	A
445	147	01001020	01001020	A
482	148	01001021	01001021	A
519	150	01001023	01001023	A
556	151	01001024	01001024	A
593	153	01001026	01001026	A
5509	331	01006052	01006052	A
5547	332	01006053	01006053	A
5584	333	01006054	01006054	A
5621	334	01006055	01006055	A
5658	336	01006057	01006057	A
5695	337	01006058	01006058	A
5733	339	01006060	01006060	A

Ingest Data: Collection, Experiment (oTree)



- oTree lets you create:
  - Controlled behavioral experiments in economics, market research, psychology, and related fields
  - Multiplayer strategy games, like the prisoner's dilemma, public goods game, and auctions.
  - Surveys and quizzes, especially those that require customized or dynamic functionality not available with conventional survey software.

Data Information Process & Platform (Lexicon / PostgreSQL)

- The **Lexicon**: an inventory of and history of changes to:
  - every available data field in every available data source
  - the structure of their storage
  - possible values and meanings of the information
  - possible transformations of each set of field values from one data source to another another data source
  - methods of data source access
  - matching algorithms and how they are to be used in conjunction with possible field value transformations
- Provides fundamental functions for the operation of the framework and is a **requirement** that the data information be collected from all partner communities
- Enables removal of much complexity required for high quality data linkage
  - i.e. No enforcing data standardization schemes on data partners

Data Information Process & Platform

Lexicon ER Diagram

metadata_table		
Key	table_name	
	friendly_name	
	description	
	critical_changes	
	last_updated	

metadata_valid_values		
Key	table_name	
Key	column_name	
	value	
	description	
	valid_use_begin_date	
	valid_use_end_date	
	last_updated	

metadata_transformation		
Key	table_name	
Key	column_name	
	description	
	transformation	SQL code
	ordinal	
	target	value | metadata
	last_updated	

metadata_column		
Key	table_name	
Key	column_name	
	friendly_name	
	description	
	requirements	required | conditional | not required | required before/after date
	demographic_type	race/ethnicity | gender | dob/age | SES | cohort
	domain_type	categorical | numeric | date | ID | open
	numeric_range_min	
	numeric_range_max	
	date_range_min	
	date_range_max	
	id_length_min	
	id_length_max	
	critical_changes	changes in domain (e.g. valid values, evidenced by appearance of new values); changes in collection method (e.g. change to forced-selection from free text, evidenced by frequency of non-entry and error rates)
	valid_use_begin_date	
	valid_use_end_date	
	original_collection_source	where collected originally and why
	original_entry_by	subject | operator | auto-generated
	last_updated	

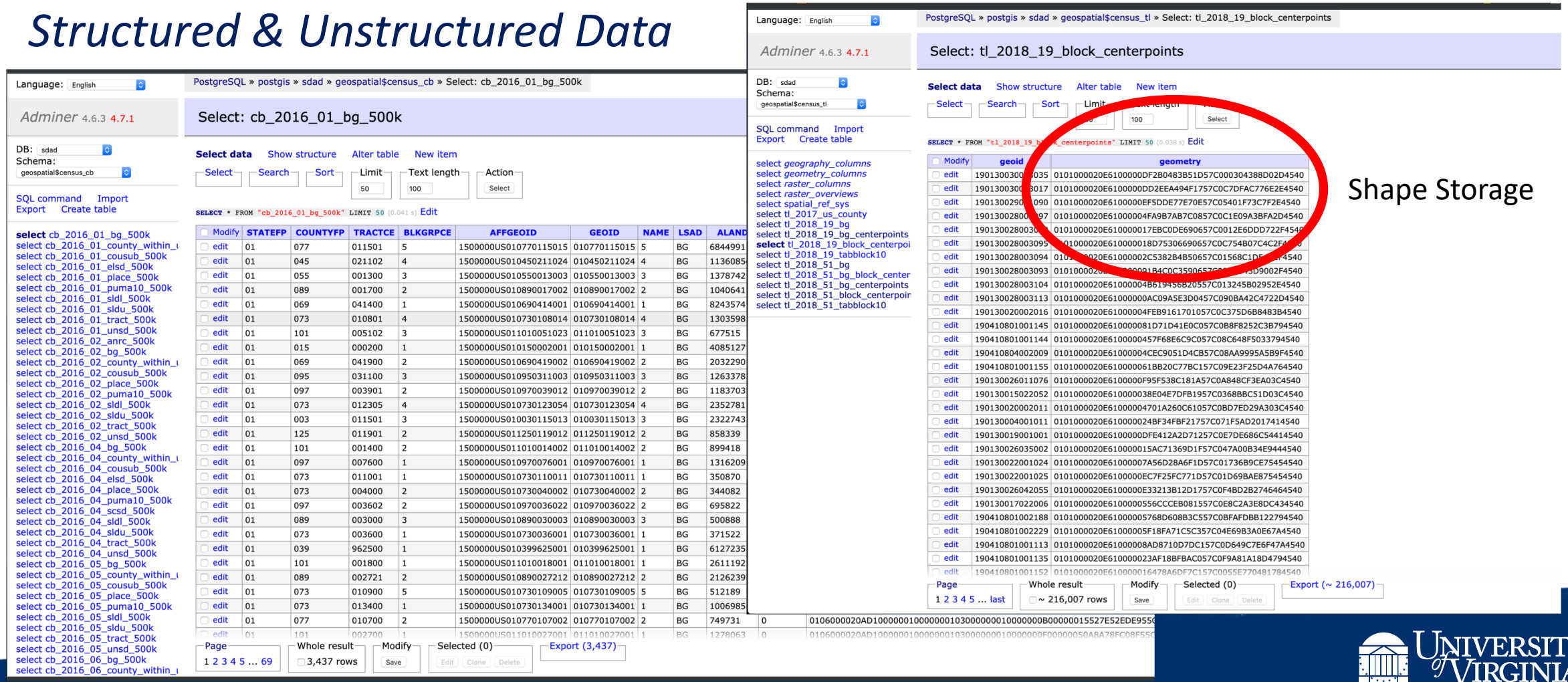
Data Storage and Management

Database Choice: PostgreSQL / PostGIS

Database Interfaces: Adminer, Rstudio, Jupyter, psql

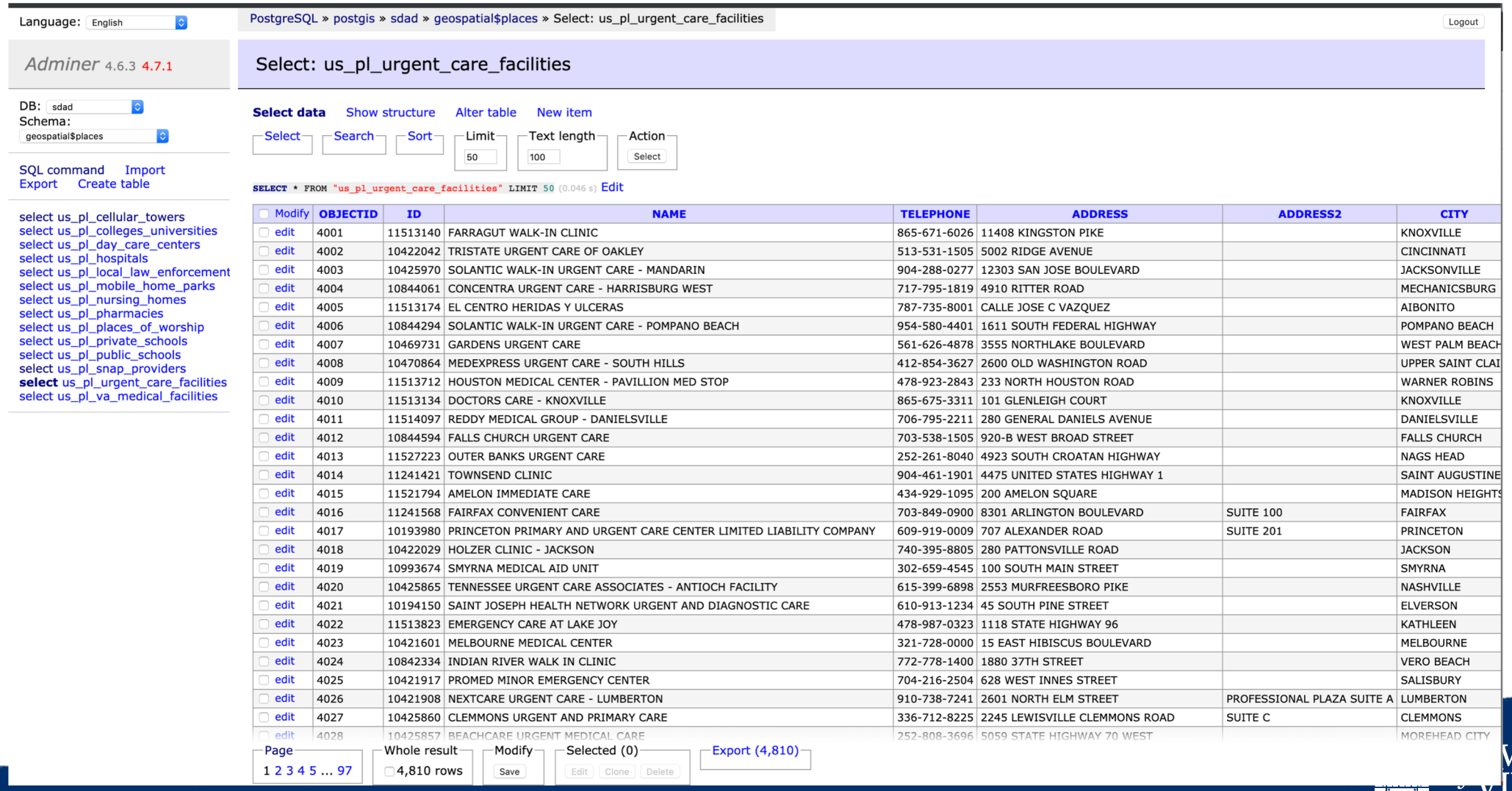
GIN Indexes + trigrams! ☺

Structured & Unstructured Data



Shape Storage

Store most used geographic places



Modify	OBJECTID	ID	NAME	TELEPHONE	ADDRESS	ADDRESS2	CITY
edit	4001	11513140	FARRAGUT WALK-IN CLINIC	865-671-6026	11408 KINGSTON PIKE		KNOXVILLE
edit	4002	10422042	TRISTATE URGENT CARE OF OAKLEY	513-531-1505	5002 RIDGE AVENUE		CINCINNATI
edit	4003	10425970	SOLANTIC WALK-IN URGENT CARE - MANDARIN	904-288-0277	12303 SAN JOSE BOULEVARD		JACKSONVILLE
edit	4004	10844061	CONCENTRA URGENT CARE - HARRISBURG WEST	717-795-1819	4910 RITTER ROAD		MECHANICSBURG
edit	4005	11513174	EL CENTRO HERIDAS Y ULCERAS	787-735-8001	CALLE JOSE C VAZQUEZ		AIBONITO
edit	4006	10844294	SOLANTIC WALK-IN URGENT CARE - POMPANO BEACH	954-580-4401	1611 SOUTH FEDERAL HIGHWAY		POMPANO BEACH
edit	4007	10469731	GARDENS URGENT CARE	561-626-4878	3555 NORTHLAKE BOULEVARD		WEST PALM BEACH
edit	4008	10470864	MEDEXPRESS URGENT CARE - SOUTH HILLS	412-854-3627	2600 OLD WASHINGTON ROAD		UPPER SAINT CLAI
edit	4009	11513712	HOUSTON MEDICAL CENTER - PAVILLION MED STOP	478-923-2843	233 NORTH HOUSTON ROAD		WARNER ROBINS
edit	4010	11513134	DOCTORS CARE - KNOXVILLE	865-675-3311	101 GLENLEIGH COURT		KNOXVILLE
edit	4011	11514097	REDDY MEDICAL GROUP - DANIELSVILLE	706-795-2211	280 GENERAL DANIELS AVENUE		DANIELSVILLE
edit	4012	10844594	FALLS CHURCH URGENT CARE	703-538-1505	920-B WEST BROAD STREET		FALLS CHURCH
edit	4013	11527223	OUTER BANKS URGENT CARE	252-261-8040	4923 SOUTH CROATAN HIGHWAY		NAGS HEAD
edit	4014	11241421	TOWNSEND CLINIC	904-461-1901	4475 UNITED STATES HIGHWAY 1		SAINT AUGUSTINE
edit	4015	11521794	AMELON IMMEDIATE CARE	434-929-1095	200 AMELON SQUARE		MADISON HEIGHTS
edit	4016	11241568	FAIRFAX CONVENIENT CARE	703-849-0900	8301 ARLINGTON BOULEVARD	SUITE 100	FAIRFAX
edit	4017	10193980	PRINCETON PRIMARY AND URGENT CARE CENTER LIMITED LIABILITY COMPANY	609-919-0009	707 ALEXANDER ROAD	SUITE 201	PRINCETON
edit	4018	10422029	HOLZER CLINIC - JACKSON	740-395-8805	280 PATTONSVILLE ROAD		JACKSON
edit	4019	10993674	SMYRNA MEDICAL AID UNIT	302-659-4545	100 SOUTH MAIN STREET		SMYRNA
edit	4020	10425865	TENNESSEE URGENT CARE ASSOCIATES - ANTIOCH FACILITY	615-399-6898	2553 MURFREESBORO PIKE		NASHVILLE
edit	4021	10194150	SAINT JOSEPH HEALTH NETWORK URGENT AND DIAGNOSTIC CARE	610-913-1234	45 SOUTH PINE STREET		ELVERSON
edit	4022	11513823	EMERGENCY CARE AT LAKE JOY	478-987-0323	1118 STATE HIGHWAY 96		KATHLEEN
edit	4023	10421601	MELBOURNE MEDICAL CENTER	321-728-0000	15 EAST HIBISCUS BOULEVARD		MELBOURNE
edit	4024	10842334	INDIAN RIVER WALK IN CLINIC	772-778-1400	1880 37TH STREET		VERO BEACH
edit	4025	10421917	PROMED MINOR EMERGENCY CENTER	704-216-2504	628 WEST INNES STREET		SALISBURY
edit	4026	10421908	NEXTCARE URGENT CARE - LUMBERTON	910-738-7241	2601 NORTH ELM STREET	PROFESSIONAL PLAZA SUITE A	LUMBERTON
edit	4027	10425860	CLEMMONS URGENT AND PRIMARY CARE	336-712-8225	2245 LEWISVILLE CLEMMONS ROAD	SUITE C	CLEMMONS
edit	4028	10425857	BEACHCARE URGENT MEDICAL CARE	252-808-3696	5059 STATE HIGHWAY 70 WEST		MOREHEAD CITY

Data Storage and Management

Database Choice: PostgreSQL / PostGIS

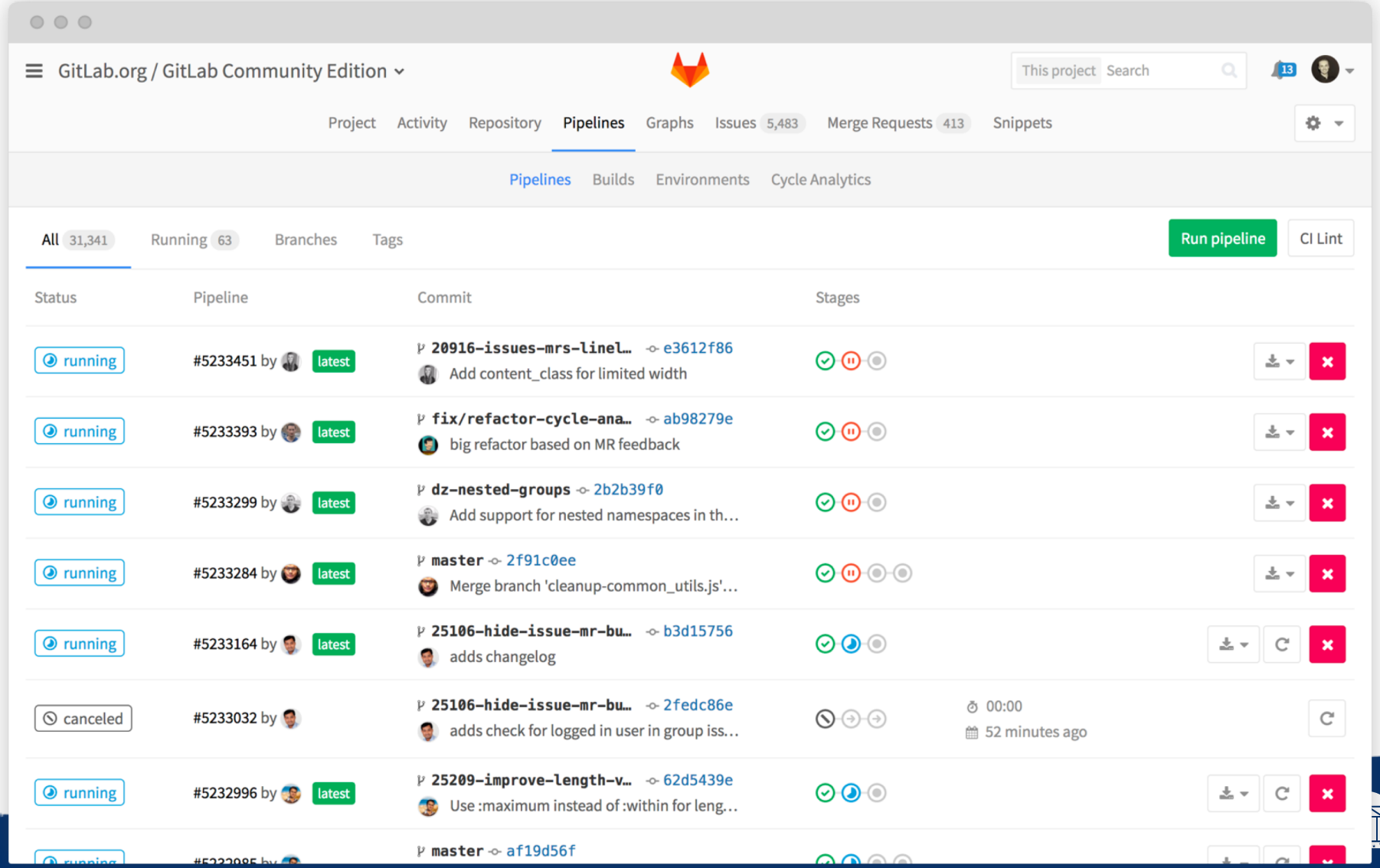
Database Interfaces: Adminer, Rstudio, Jupyter, psql

Very fast access to geo files and database GIS functions

```
# create db connection
con <- sdalr::con_db(dbname = "sdad", host = "127.0.0.1", port = 5433, user = "anonymous", pass =
"anonymous")

# create SQL query
sql <- "SELECT distinct \"GEOID10\" geoid, geometry
        FROM tl_2018_19_tabblock10 where left(\"GEOID10\", 5) = '19127’”

# get census blocks
marshall_county_blocks <- sf::st_read(con, query = sql) %>%
  st_transform(crs = 4269)
```

Code Management: GitLab

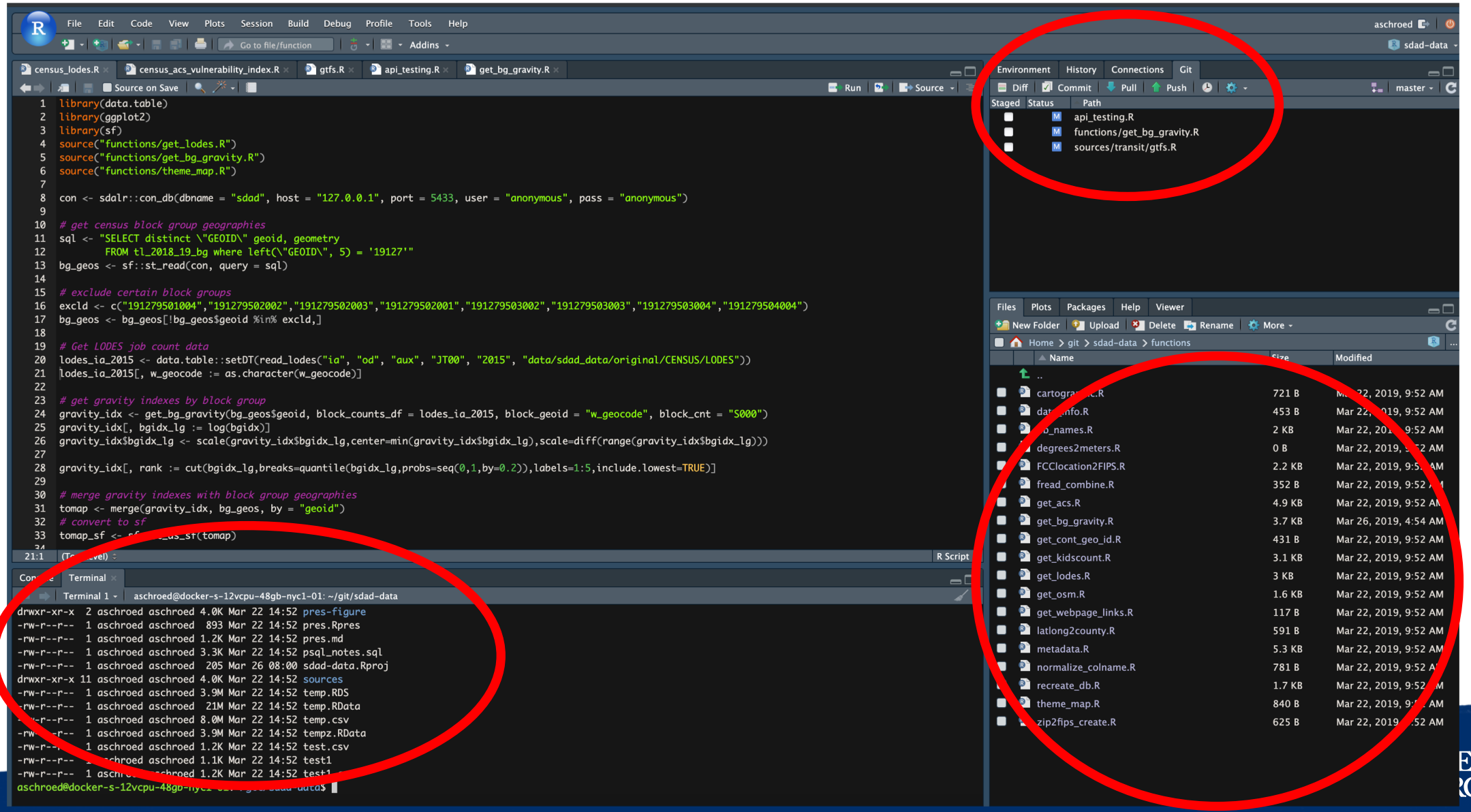
Multiple Branch Management Critical in Team Science



Data Profiling & Preparation

Preferred Platform: RStudio Server

Git Integration



Command Line (Bash) Access

Server Files

Data Profiling: Quality

Completeness

percentage of elements properly populated

e.g. Testing for NULLs and empty strings where not appropriate

Value Validity

percentage of elements whose attributes possess meaningful values

e.g. A comparison constraint like {male; female} or an interval constraint like age = [0,110]

Consistency

a measure of the degree to which two or more data attributes satisfy a well-defined dependency constraint – relationship validation

e.g. Zip-code – state consistency or gender – pregnancy consistency

Uniqueness

the number of unique values taken by an attribute, or a combination of attributes in a dataset

e.g. Frequency distribution of an element

note. The more homogeneous the data values of an element, the less useful the element is for analysis

Duplication

a measure of the degree of replication of distinct observations per observation unit type

e.g. Greater than 1 registration per student per official reporting period

note. Duplication can occur as a result of choice of level of aggregation; for example, aggregating to a single student registration per academic year when registration information is actually collected multiple times per academic year

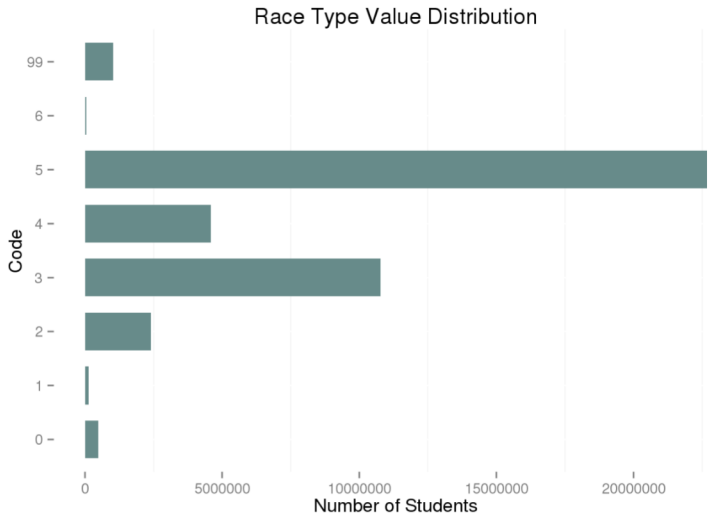
Race Type Report

SUMMARY OF DATA ELEMENT

Race Type is recorded for each student as the one or more races they identify with. This set was collected from students in the school system during the school years of 2005-2015. During our collection period, starting with the 2010-2011 school year, the race type definition expanded to allow students to choose multiple races and separately specify Hispanic ethnicity. In this process categories were eliminated. Eliminated categories included the 'Unspecified or Unknown' category and the 'Hispanic' category. A student's Hispanic ethnicity is now captured under an ethnic flag.

Number of Unique Values: 8

value	value_description
99	Multiple race types reported
1	American Indian or Alaskan Native
2	Asian
3	Black or African American
4	Hispanic (valid before 2010 school year)
5	White
6	Native Hawaiian or Other Pacific Islander
0	Unspecified or Unknown (valid before 2010 school year)



DATA ANALYTICS SUMMARY OF RESULTS

Test	Measurementvalues	Value
Completeness	Number of missing values	0
	Percent of complete values	100 %
Validity	Number of invalid values	0
	Percent of valid values	100 %
Uniqueness	Number of unique values	8
Record Consistency I	Number of inconsistent records	0
	Percent consistent records	100 %
Record Consistency II	Number of inconsistent records	2830444
	Percent consistent records	93.28 %
Longitudinal Inconsistency I	Individuals with inconsistent records	160055

Data Profiling: Valid Values

- Data elements with proper values have ***value validity***
- The percentage of data elements whose attributes possess values within the range expected for a legitimate entry is a measure of value validity
- Checking for value validity generally comes in the form of straight-forward domain constraint rules
  - How many entries contain non-valid values for a non-empty text field representing gender?
    - *< count gender where gender is not* (*male, female*) *>*
  - How many entries contain non-valid values for a non-empty integer field representing age?
    - *< count age where age is not between* [0, 110] *>*

Pulled from current James City County MLS Data

zip_code	area	subdivision	neighborhood	zoning	parcel_id
23185	JCC	Governors Land	River Reach	R-4	4511000022
23188	JCC	Wellington		RESIDENT	1330800178
23188	JCC	Powhatan Secondary		RES	3741600013
23185	JCC	Kingsmill	Padgetts Ordinary	R 4	5041100213
23185	JCC	Pointe @ Jamestown		RES	4640600108
23185	JCC	Paddock Green	Paddock Green	R1	

Comparison constraint: **zoning 2015, James City County**= {A-1, R-1, R-2, R-3, R-4, R-5, R-6, R-7, R-8, LB, B-1, M-1, M-2, RT, PUD, MU, PL, EO}

- During Data Profiling issues are described, not “fixed”
- The appropriate fix depends upon the needs of the research
- It may be appropriate to simply normalize all zoning entries to the five major categories of zoning: Residential, Mixed Residential-Commercial, Commercial, Industrial, and Special

Data Profiling: Consistency

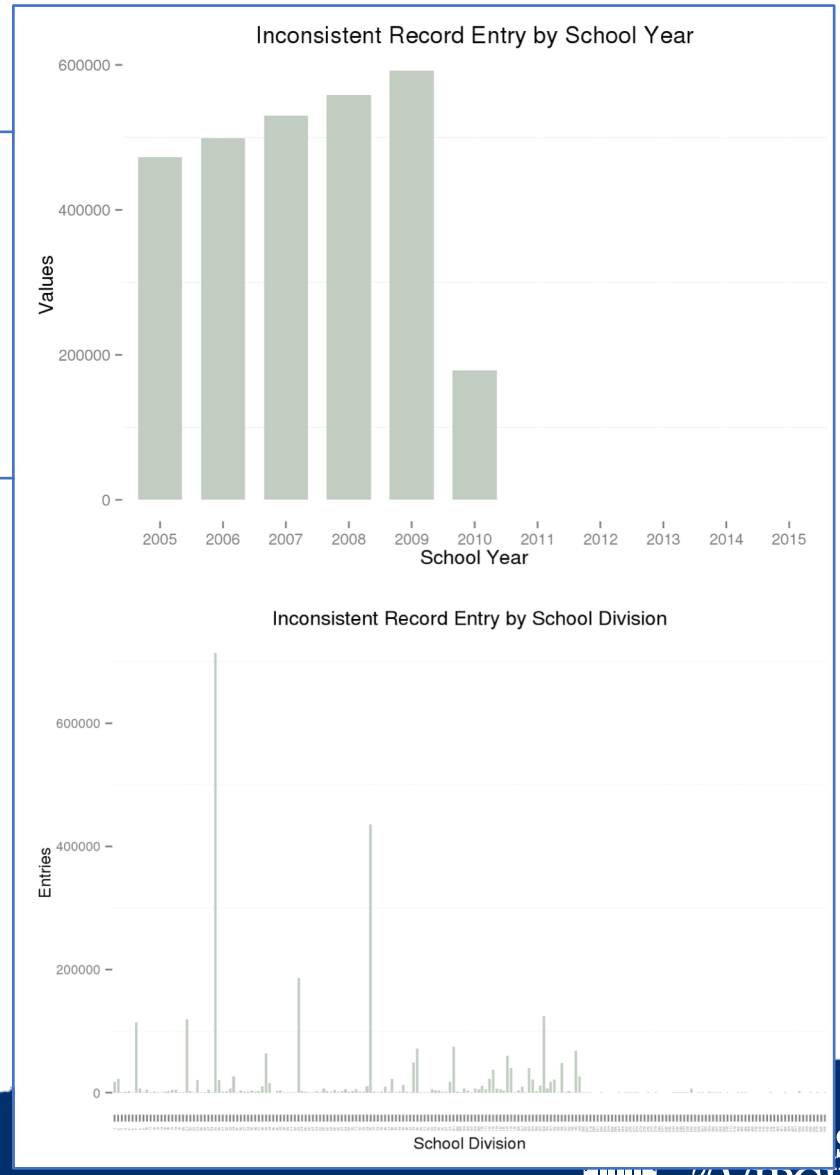
RECORD CONSISTENCY II

Find Records with an inconsistent relationship between Race Type and School Year .

Check if there any records that have a race_type of '4' after the 2009 school year.

Number Inconsistent: 2830444

Percent Consistent: 93.28 %



Record Consistency The concept of record consistency is best understood as the degree of logical agreement "between" record field values in either a single dataset or between two or more datasets. Simple Example: location disagreement between zip code and state FIPS code.

Longitudinal Consistency An inconsistency in the data when checked over time (longitudinally), to see if the same value is recorded for every new record when it should be (i.e. birthdate and other demographics).

Data Profiling & Preparation

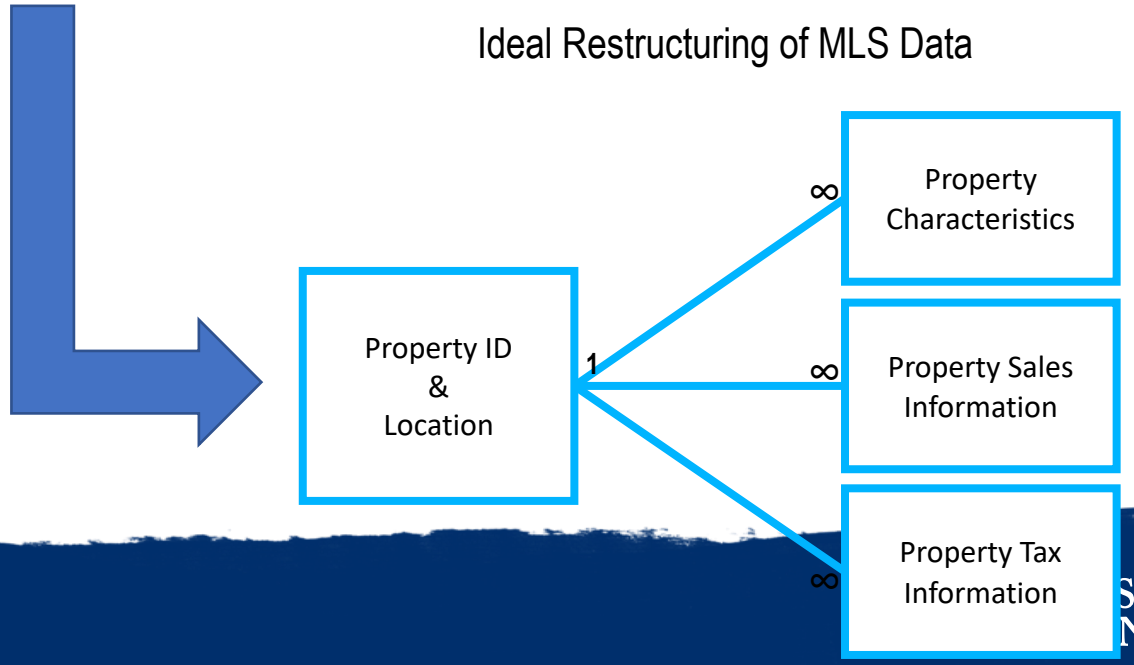
The **Data Preparation** Phase includes the activities necessary to **"fix"** the issues of Quality, Structure, and Metadata discovered during Data Profiling – activities can include:

- Cleansing
  - Missing Values
  - Date Formats
  - Nominal => numeric
  - Outliers
  - Inconsistent Data
  - De-duplication
- Transformation
  - Aggregation
  - Normalization
  - Smoothing/Winsorization
  - Imputation
  - Feature Construction
- Restructuring

Current Structure of Williamsburg MLS Data

List Number	Agency Name	Agency Phone	Agency Email	Listing Agent	Listing Agent Phone	Listing Agent Email	Co-Listing Agent	Property Type	Card Format
Book Section	Selling Agency	Selling Agency Phone	Selling Agency Email	Selling Agent	Selling Agent Phone	Selling Agent Email	Co-Selling Agent	End Date	book_sec
Listing Date	Sold Date	Under Cont. Date	Fall-thru Date	Status	Status Change	Withdraw Date	Cancel Date	Contingent	Cont. Remarks
Orig. List Price	Price	Sold Price	high_price	Low Price	assessed_val	Partial Tax Assmnt	financing	Area	Relocation
St. #	box_nbr	St. Dir.	Street Name	Address 2	streetdirsuffix	Street Suffix	carrier_route	City	State
county	country	Zip Code	geo_county	Taxes	geo_lat	geo_lon	Est. Fin. SqFt	sqft1	sqft2
sqft3	sqft4	Year Built	2+ Bdroms on 1st Flr	Realtor.com Type	lot_size	Total Acres	Condo Level	sell_broker_comm	Variable Commission
stories	Total Rooms	Total Bedrooms	total_bath	Baths - Full	Baths - Half	baths_3_4	Garage Type	garage_stall	Water Frontage
Zoning	taxes	Tax Year	Subdivision	Public Remarks	Agent Remarks	**Parcel ID**	Legal Description	Directions	Foreclosure
Owner Phone	Owner Name	Neighborhood	mod_timestamp	Ltd Service Agent	Occupied By	Owner/Agent	Mster Bdrm 1st Floor	SqFt Source	Listing Type
# Stories	# Fireplaces	Golf Frontage	IDX Y/N	Supplement Attached	Seller Concession(s)	Special Assmnts	Type	Rollback Taxes	userdefined16
SellingBroker Incent	Ownership	Describe Concession	How Sold	Selling Broker Comp	userdefined22	Assessed Value	Est.Unfinished Sq Ft	Tax Rate	Garage Bays
userdefined27	userdefined28	userdefined29	userdefined30	Est. Closing Date	userdefined32	userdefined33	Lot Description	Short/CompromiseSale	userdefined36
userdefined37	userdefined38	userdefined39	userdefined40	userdefined41	userdefined42	userdefined43	userdefined44	userdefined45	userdefined46
userdefined47	userdefined48	userdefined49	userdefined50	userdefined51	userdefined52	userdefined53	userdefined54	userdefined55	userdefined56
Photo URL	Days on Market	Rooms	Features						

Ideal Restructuring of MLS Data



Recent Example: Distinct Counts of Children 0-5

Datasets Under Consideration

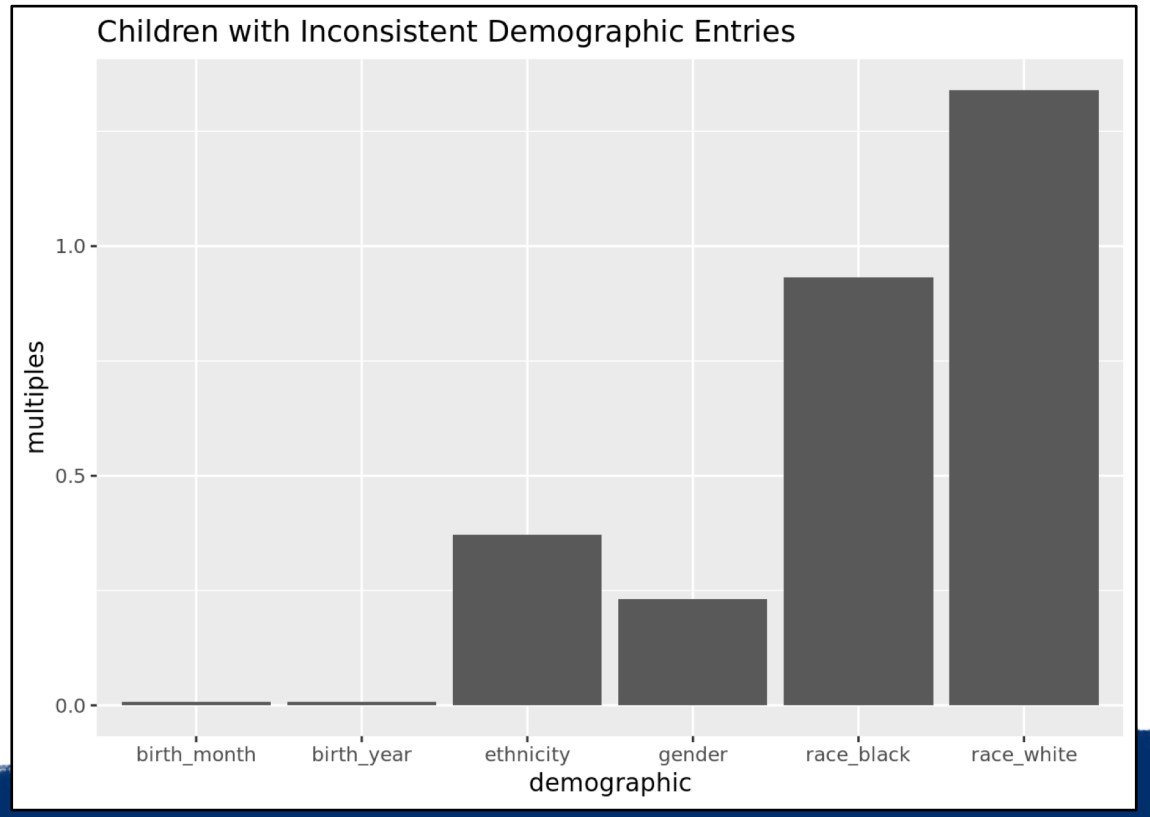
Finding the best representations of
Who, What, When & Where Across Agencies, Across Datasets

Source		Demographics				Location		Time
Agency	Dataset	Race	Ethnicity	Gender	Age	Person	Service	Service
VDSS	Customers By Year	cust_race_is_amer_indian_alaska_native_ind customer_race_is_asian_indicator customer_race_is_black_indicator customer_race_is_hawaiian_pacific_islander_ind customer_race_is_other_indicator customer_race_is_white_indicator	ethnicity_code	gender_code	age_class_code age_group_code age_type_code month_of_birth year_of_birth	-	-	calendar_year_number service_year
VDSS	SNAP Customers By Year	-	-	-	-	zip_code	county_fips_code	calendar_year_number
VDSS	TANF Customers By Year	-	-	-	-	zip_code	county_fips_code	calendar_year_number
VDSS	Foster Customers By Year	-	-	-	-	-	county_fips_code	calendar_year_number
VOCS	OCS Services By Year	-	-	-	-	-	-	service_begin_date service_end_date service_duration program_year
VDOE	Unique Student Listing	race_type		gender	birth_month birth_year [grade_code]		[division_number_reporting_school_numbe]	school_year entry_date
VDOE	VPI+	students_race_ethnicity	students_race_ethnicity	gender	birth_month birth_year [grade]		[school]	assessment_date school_year

Recent Example: Distinct Counts of Children 0-5

Longitudinal Consistency

- Each dataset profiled to discover the most consistent recording of demographic information over time



Recent Example: Distinct Counts of Children 0-5

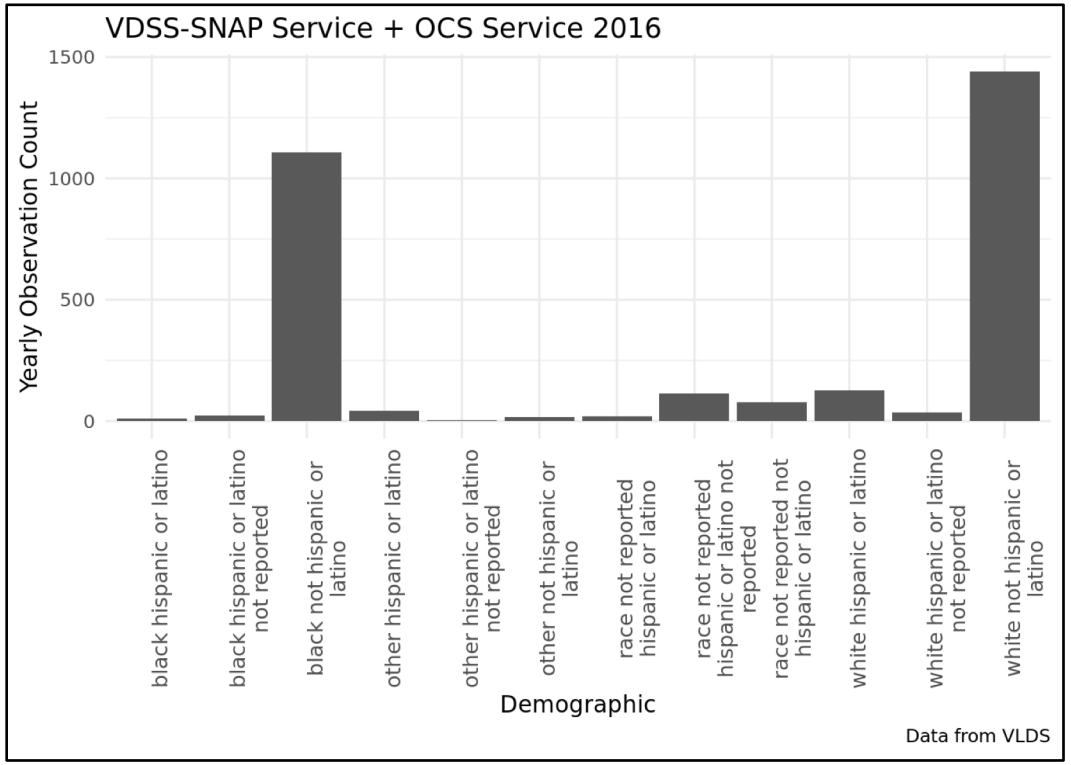
Transformation & Linkage Enables Distinct Count Cross-Tabulation

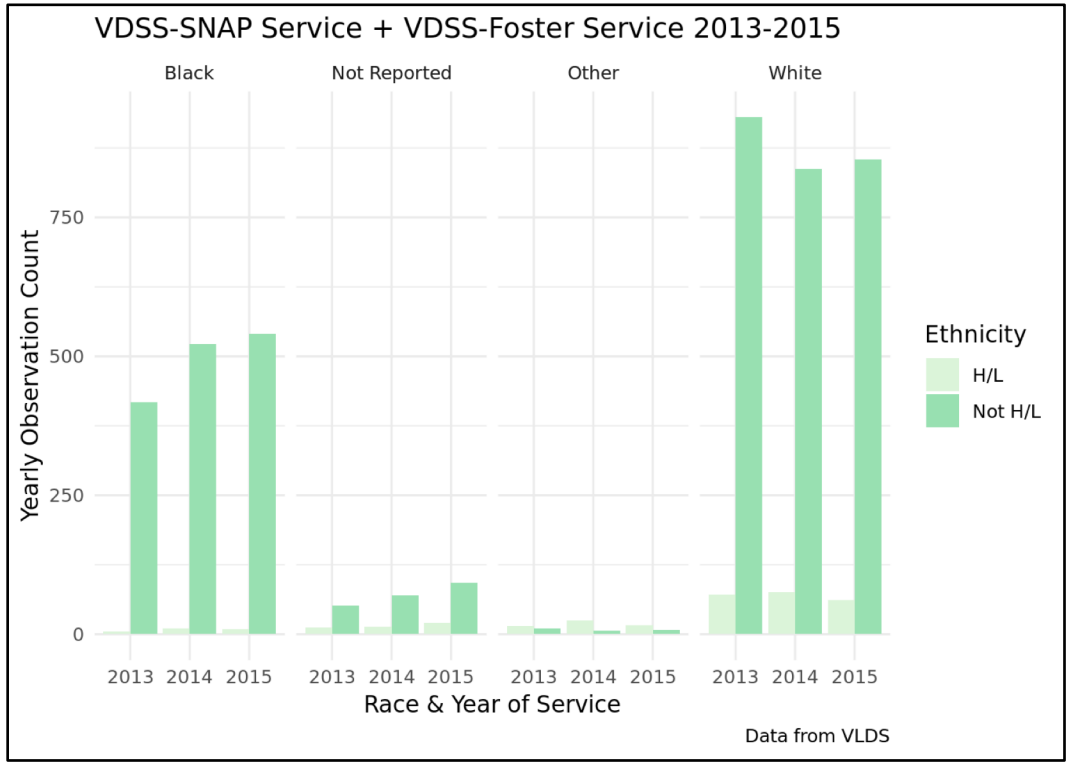
Distinct Count Cross-Tabulation Service By Race/Ethnicity By Year

Services	Demographic	2013	2014	2015	2016
snap & tanf	black hispanic or latino not reported	1159	1710	1766	1333
snap & tanf	black not hispanic or latino	72710	62084	56655	40908
snap & tanf	other hispanic or latino	2112	1752	1529	1025
snap & tanf	other hispanic or latino not reported	143	162	129	66
snap & tanf	other not hispanic or latino	1870	1616	1287	877
snap & tanf	race not reported hispanic or latino	1438	1123	1064	749
snap & tanf	race not reported hispanic or latino not reported	4461	5727	8795	6941
snap & tanf	race not reported not hispanic or latino	5409	4375	4190	3213
snap & tanf	white hispanic or latino	5357	4412	4325	3424
snap & tanf	white hispanic or latino not reported	902	1889	1958	1264
snap & tanf	white not hispanic or latino	41489	34649	30686	21683

Recent Example: Distinct Counts of Children 0-5

Distinct Count Cross-Tabulation Enables Distinct Count Plots

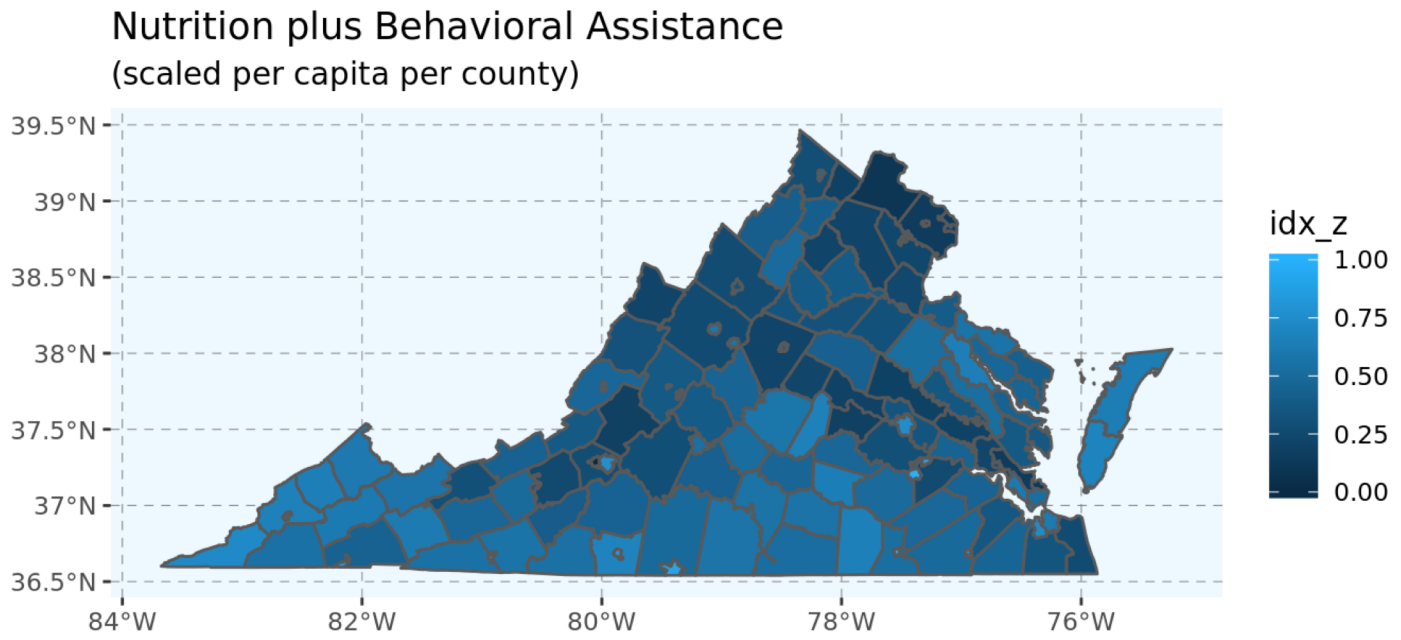




Recent Example: Distinct Counts of Children 0-5

Distinct Count Cross-Tabulation Enables Creation of New Indices

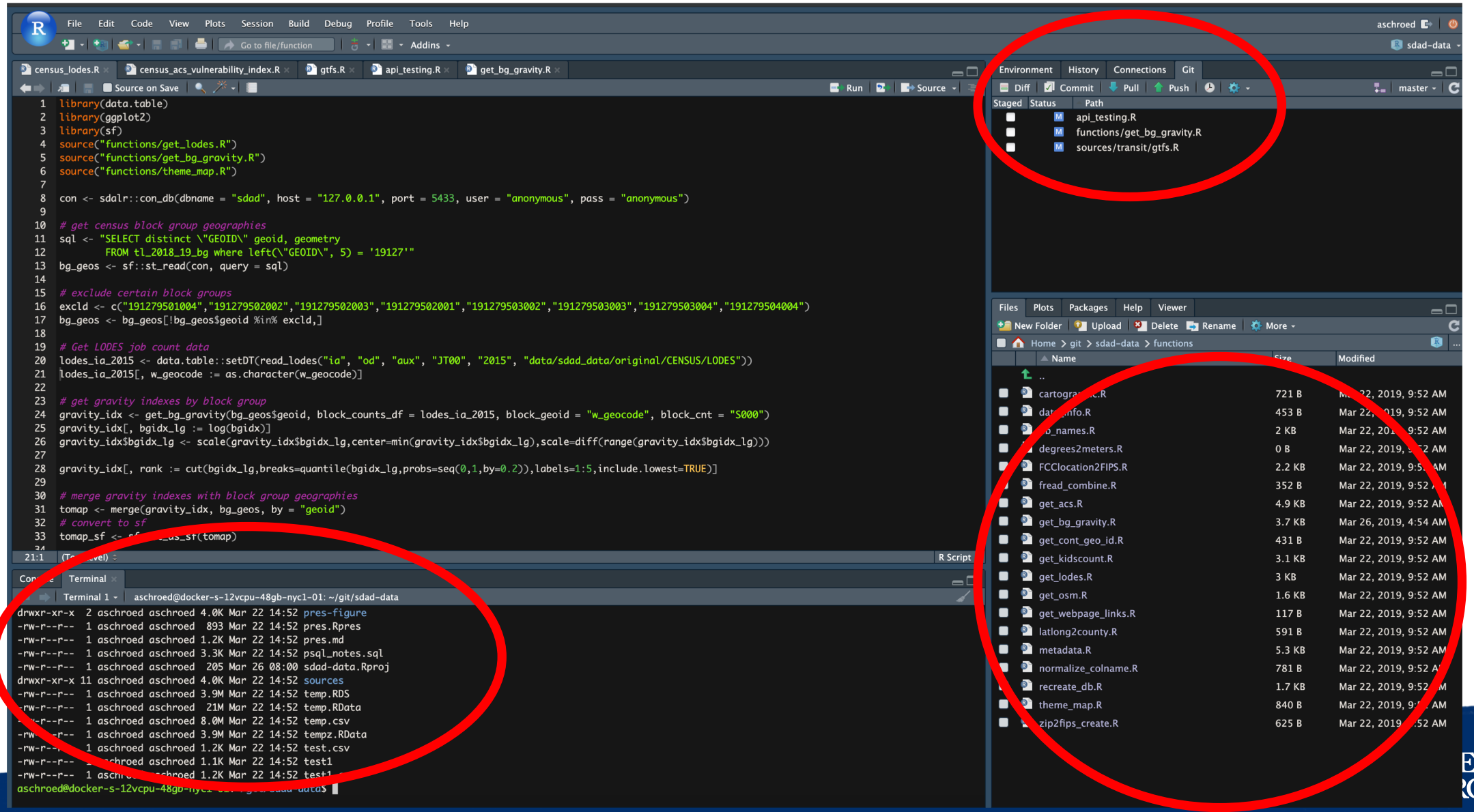
- Example Composite Index Combining SNAP (Nutrition) and OCS (Behavioral Assistance) Data



Data Analysis & Data Product Creation

Preferred Platform: RStudio Server

Git Integration



Command Line (Bash) Access

Server Files

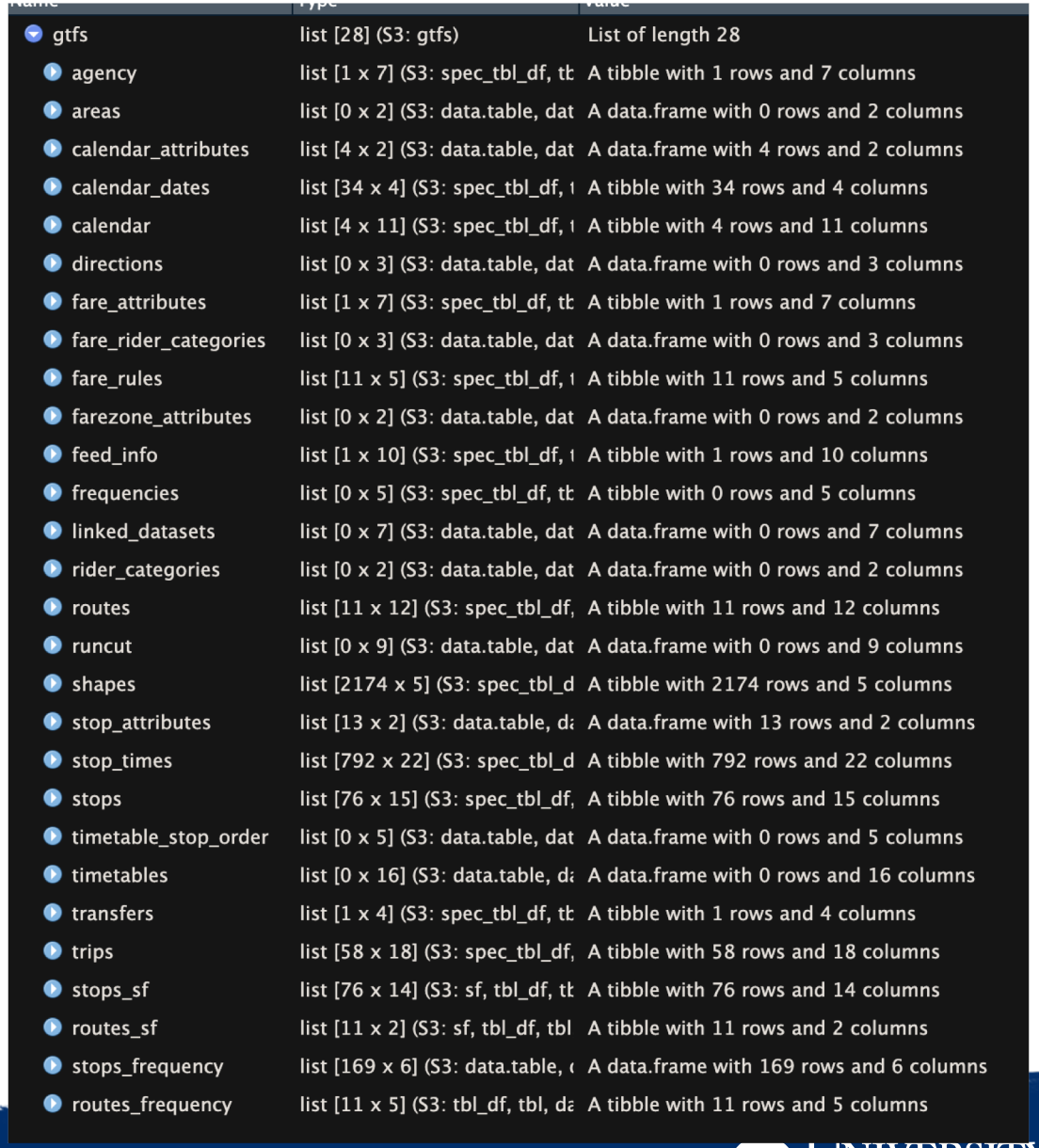
Data Ingestion
Marshalltown

Shapefiles from our GIS database (PostGIS)

```
# create db connection
con <- sdalr::con_db(dbname = "sdad", host = "127.0.0.1", port = 5433, user = "anonymous", pass =
"anonymous")

# create SQL query
sql <- "SELECT distinct \"GEOID10\" geoid, geometry
        FROM tl_2018_19_tabblock10 where left(\"GEOID10\", 5) = '19127’”

# get census blocks
marshall_county_blocks <- sf::st_read(con, query = sql) %>%
  st_transform(crs = 4269)
```

Data Ingestion
Marshalltown

```
library(tidytransit)

# get Marshalltown feed URL
feed_url <- feedlist_df %>%
  setDT(.) %>%
  .[loc_t %like% "Marshalltown, IA", url_i]

# read gtfs data from url
gtfs <- read_gtfs(feed_url, geometry = TRUE, frequency =
TRUE)
```



gtfs	list [28] (S3: gtfs)	List of length 28
agency	list [1 x 7] (S3: spec_tbl_df, tb	A tibble with 1 rows and 7 columns
areas	list [0 x 2] (S3: data.table, dat	A data.frame with 0 rows and 2 columns
calendar_attributes	list [4 x 2] (S3: data.table, dat	A data.frame with 4 rows and 2 columns
calendar_dates	list [34 x 4] (S3: spec_tbl_df, t	A tibble with 34 rows and 4 columns
calendar	list [4 x 11] (S3: spec_tbl_df, t	A tibble with 4 rows and 11 columns
directions	list [0 x 3] (S3: data.table, dat	A data.frame with 0 rows and 3 columns
fare_attributes	list [1 x 7] (S3: spec_tbl_df, tb	A tibble with 1 rows and 7 columns
fare_rider_categories	list [0 x 3] (S3: data.table, dat	A data.frame with 0 rows and 3 columns
fare_rules	list [11 x 5] (S3: spec_tbl_df, t	A tibble with 11 rows and 5 columns
farezone_attributes	list [0 x 2] (S3: data.table, dat	A data.frame with 0 rows and 2 columns
feed_info	list [1 x 10] (S3: spec_tbl_df, t	A tibble with 1 rows and 10 columns
frequencies	list [0 x 5] (S3: spec_tbl_df, tb	A tibble with 0 rows and 5 columns
linked_datasets	list [0 x 7] (S3: data.table, dat	A data.frame with 0 rows and 7 columns
rider_categories	list [0 x 2] (S3: data.table, dat	A data.frame with 0 rows and 2 columns
routes	list [11 x 12] (S3: spec_tbl_df,	A tibble with 11 rows and 12 columns
runcut	list [0 x 9] (S3: data.table, dat	A data.frame with 0 rows and 9 columns
shapes	list [2174 x 5] (S3: spec_tbl_d	A tibble with 2174 rows and 5 columns
stop_attributes	list [13 x 2] (S3: data.table, da	A data.frame with 13 rows and 2 columns
stop_times	list [792 x 22] (S3: spec_tbl_d	A tibble with 792 rows and 22 columns
stops	list [76 x 15] (S3: spec_tbl_df,	A tibble with 76 rows and 15 columns
timetable_stop_order	list [0 x 5] (S3: data.table, dat	A data.frame with 0 rows and 5 columns
timetables	list [0 x 16] (S3: data.table, da	A data.frame with 0 rows and 16 columns
transfers	list [1 x 4] (S3: spec_tbl_df, tb	A tibble with 1 rows and 4 columns
trips	list [58 x 18] (S3: spec_tbl_df,	A tibble with 58 rows and 18 columns
stops_sf	list [76 x 14] (S3: sf, tbl_df, tb	A tibble with 76 rows and 14 columns
routes_sf	list [11 x 2] (S3: sf, tbl_df, tbl	A tibble with 11 rows and 2 columns
stops_frequency	list [169 x 6] (S3: data.table,	A data.frame with 169 rows and 6 columns
routes_frequency	list [11 x 5] (S3: tbl_df, tbl, da	A tibble with 11 rows and 5 columns

Transit route data from transitfeeds.com

UNIVERSITY of VIRGINIA

Data Ingestion
Marshalltown

```
acs_vars <- c(
  "B25070_001","B25070_010",
  "B25091_001","B25091_011","B25091_022",
  "B25044_001","B25044_003","B25044_010",
  "B22010_001","B22010_002",
  "B17021_001","B17021_002"
)
acs_est <- get_acs(geography="block
group",state=state_names,county=county_names,

variables=acs_vars,year=year,cache_table=TRUE,out
put="wide", geometry = TRUE)
```

NAME	B25070_001E	B25070_001M	B25070_010E	B25070_010M	B25
Block Group 1, Census Tract 9501, Marshall County, I...	24	19	3	4	
Block Group 2, Census Tract 9501, Marshall County, I...	16	12	0	9	
Block Group 3, Census Tract 9501, Marshall County, I...	56	38	3	5	
Block Group 4, Census Tract 9501, Marshall County, I...	15	8	4	4	
Block Group 1, Census Tract 9502, Marshall County, I...	76	32	3	5	
Block Group 2, Census Tract 9502, Marshall County, I...	83	55	2	5	
Block Group 3, Census Tract 9502, Marshall County, I...	91	28	5	7	
Block Group 1, Census Tract 9503, Marshall County, I...	45	26	0	9	
Block Group 2, Census Tract 9503, Marshall County, I...	76	27	10	12	
Block Group 3, Census Tract 9503, Marshall County, I...	28	17	3	4	
Block Group 4, Census Tract 9503, Marshall County, I...	87	43	6	6	
Block Group 1, Census Tract 9504, Marshall County, I...	90	29	2	3	
Block Group 2, Census Tract 9504, Marshall County, I...	18	17	0	9	
Block Group 3, Census Tract 9504, Marshall County, I...	29	20	3	6	
Block Group 4, Census Tract 9504, Marshall County, I...	47	27	2	3	
Block Group 1, Census Tract 9505, Marshall County, I...	288	99	51	45	
Block Group 2, Census Tract 9505, Marshall County, I...	175	77	38	45	
Block Group 3, Census Tract 9505, Marshall County, I...	160	81	0	9	
Block Group 4, Census Tract 9505, Marshall County, I...	140	71	14	17	
Block Group 1, Census Tract 9506, Marshall County, I...	161	79	5	9	
Block Group 2, Census Tract 9506, Marshall County, I...	135	63	68	53	
Block Group 3, Census Tract 9506, Marshall County, I...	315	95	35	25	
Block Group 4, Census Tract 9506, Marshall County, I...	48	40	11	18	

Vulnerability data composited from Census ACS housing, transportation, and nutrition subsidy variables

Gravity Models for Job Access and Transit Access

census_lodes.R | census_acs_vulnerability_index.R | gtfs.R | api_testing.R | get_bg_gravity.R

Source on Save

```r
#* @apiTitle Block Group Gravity Index

#* Create a Gravity Model Index using a list of block group geoids (12 characters) and
#* a data.frame of block geoids and a count of something for each one
#* @param bg_geoid_list list of Census block group geoids (12 characters)
#* @param block_counts_df data.frame with a geoid column (15 characters) and a count of somethin
#* @param block_cnt the name of the column holding the counts in block_counts_df
#* dist_mi the distance from each block group centerpoint from which block will be retrieved
#* @post /get_bg_gravity
get_bg_gravity <- function(bg_geoid_list, block_counts_df, block_geoid = "geoid", block_cnt = "t
  library(data.table)
  for (bgid in bg_geoid_list) {
    # Get Blocks
    bg_blocks <- get_blocks_within_distance_of_bg(bg = bgid, dist_mi = dist_miles)

    # Merge Blocks and LODES
    bg1_count <- merge(bg_blocks, block_counts_df, by.x = "geoid_block", by.y = block_geoid)
    setDT(bg1_count)

    # Aggregate Jobs per Block
    bg1_count_2 <- bg1_count[,.(block_cnt=sum(as.numeric(get(block_cnt)))), .(geoid_bg, geoid_block, dist_mi)]

    # Calculate Index
    idx <- bg1_count_2[, .(block_cnt, d_sqr=dist_mi^2, e=block_cnt/(dist_mi^2))][d_sqr==0, e := 0][, sum(e)]
    idx_dt <- data.table(geoid = bgid, bgidx = idx)

    # Combine
    if (exists("idxes") == TRUE) idxes <- rbindlist(list(idxes, idx_dt))
    else idxes <- idx_dt
  }
  idxes
}
```

$$E \equiv \sum_{i=1}^{n} \frac{p_i}{r_i^2}$$

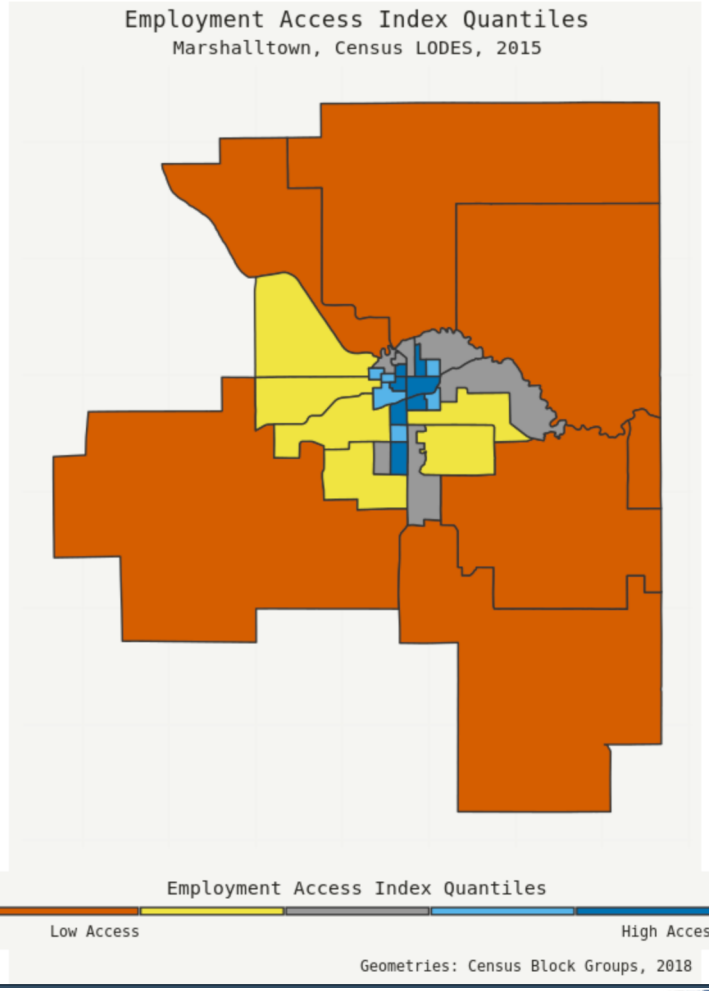
Equation 3: Employment Access Index Definition

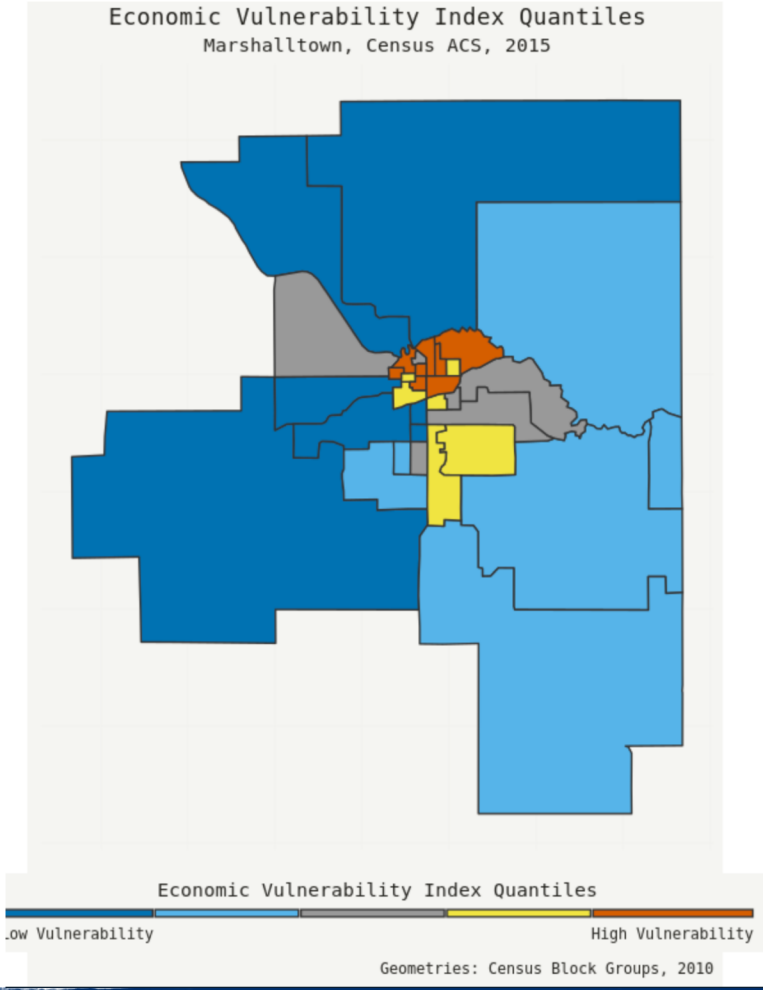
Where:

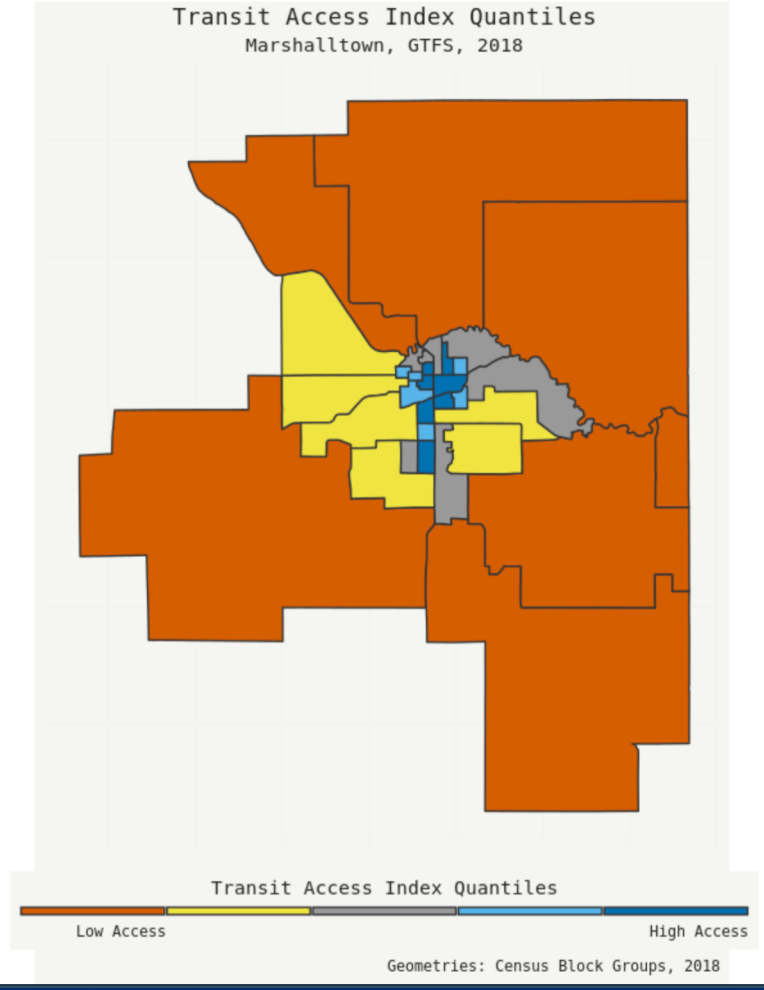
- E is the Employment Access for a given Census block group
- n is the total number of Census blocks
- p_i is the number of jobs in the i^{th} Census block
- r_i is the distance (in miles) from the center of the given Census block group to the center of the i^{th} Census block

Data Visualization & Sharing: Maps (ggplot)

Combining Administrative and Survey Data to Relate Job Availability, Economic Vulnerability and Access to Transit





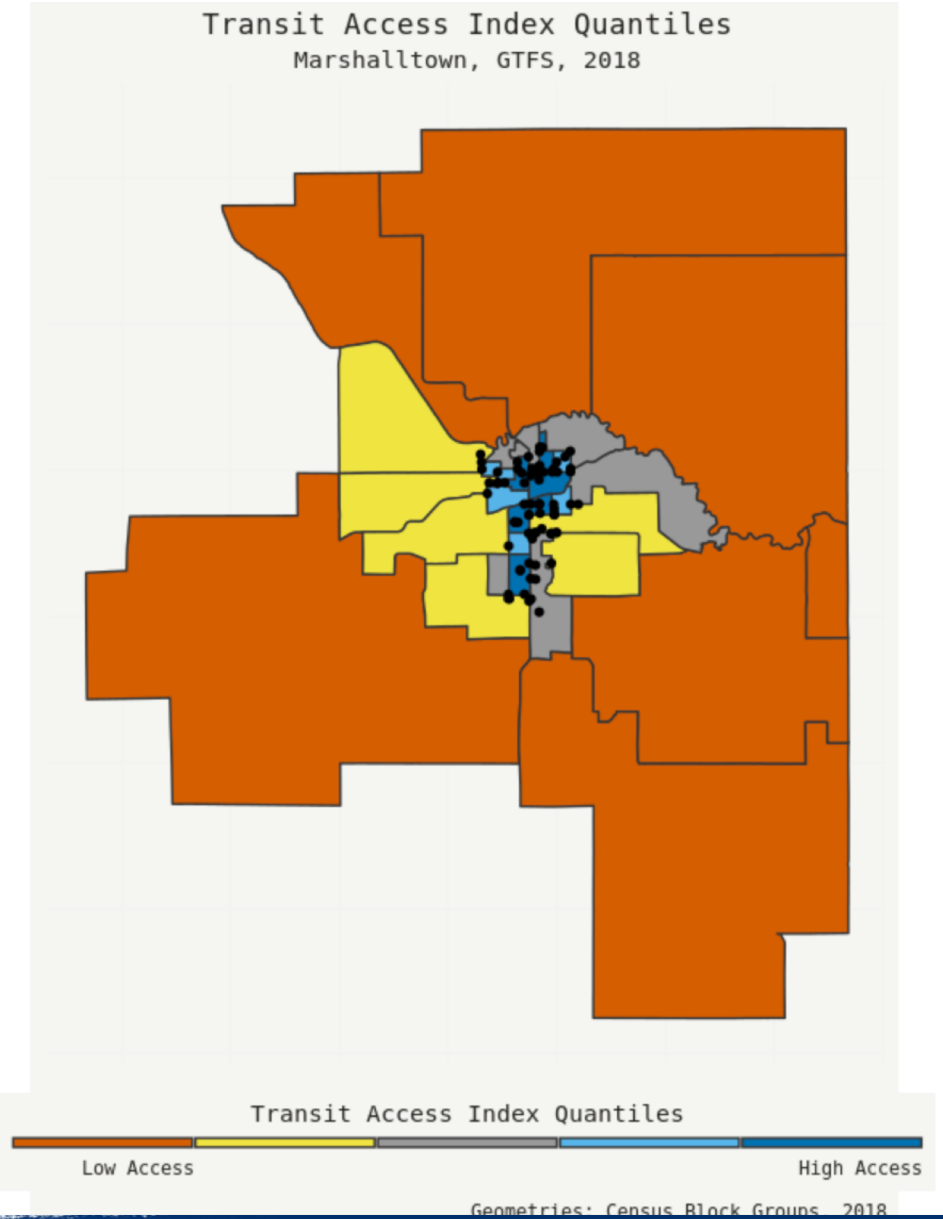


LEHD Origin-Destination Employment Statistics (LODES),

American Community Survey (ACS)

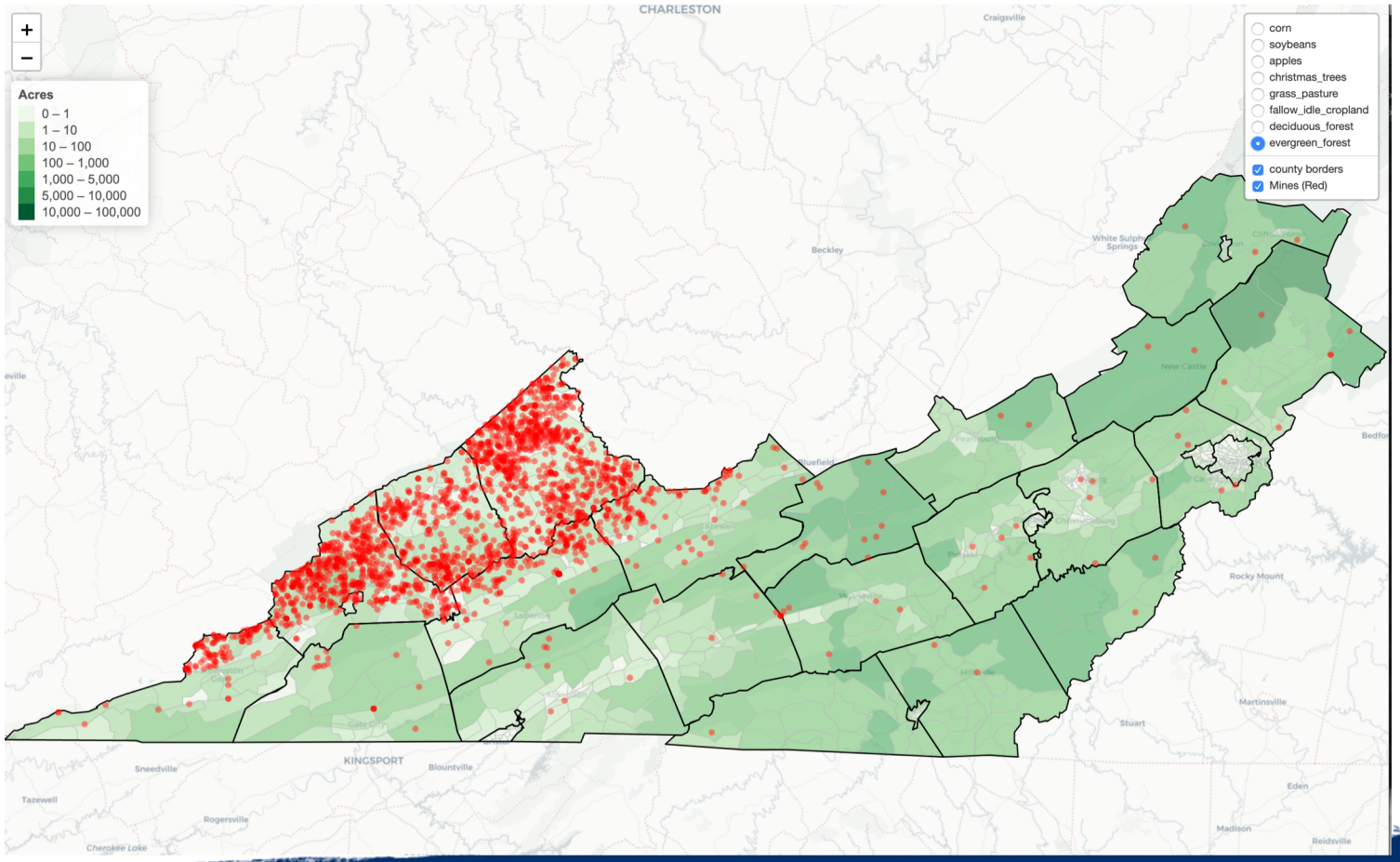
General Transit Feed Specification (GTFS)

UNIVERSITY of VIRGINIA


Transit Access Index Quantiles
Marshalltown, GTFS, 2018
Transit Access Index Quantiles
Low Access
High Access
Geometries: Census Block Groups, 2018

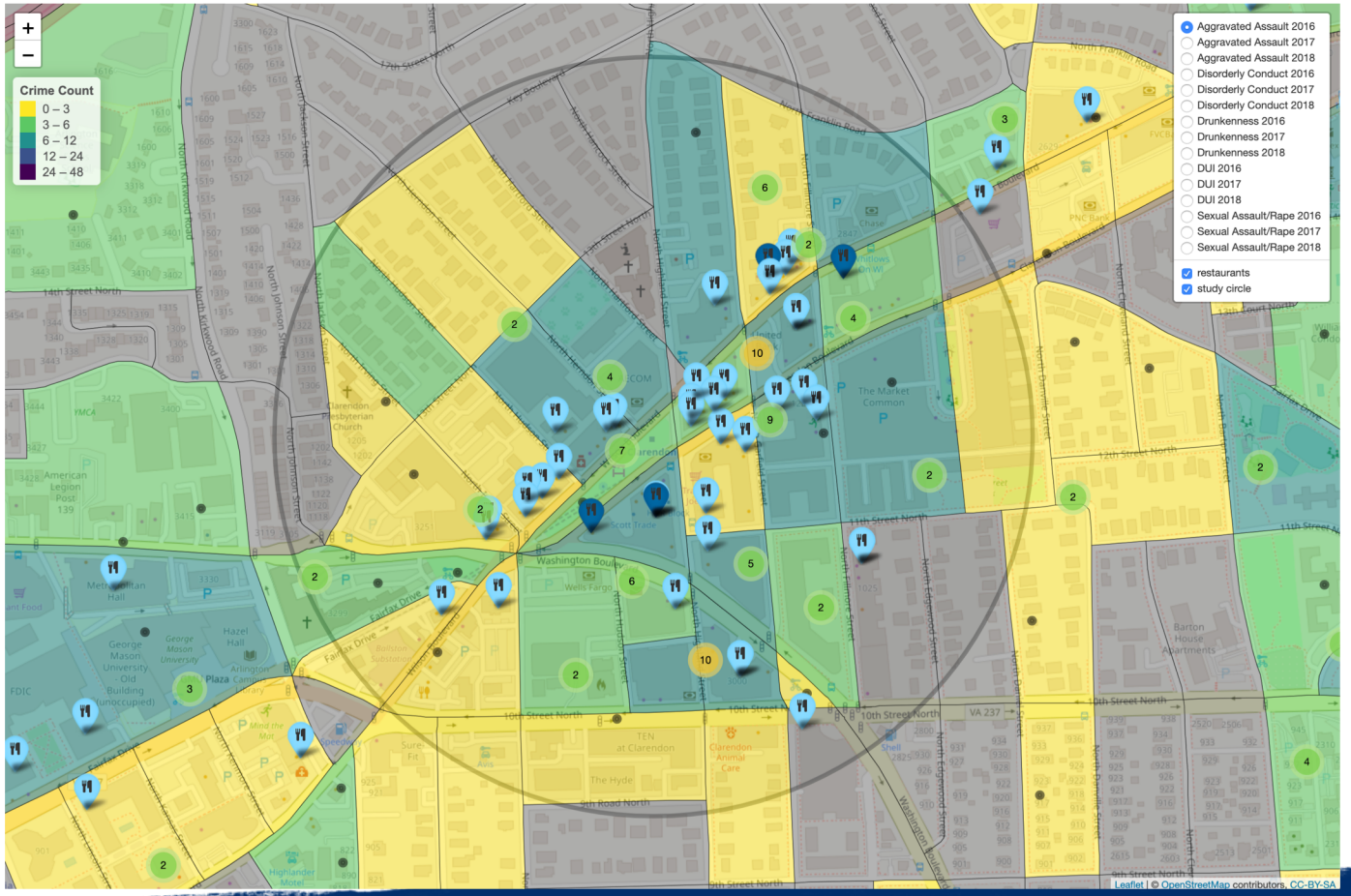
Data Visualization & Sharing

Interactive Maps (leaflet)



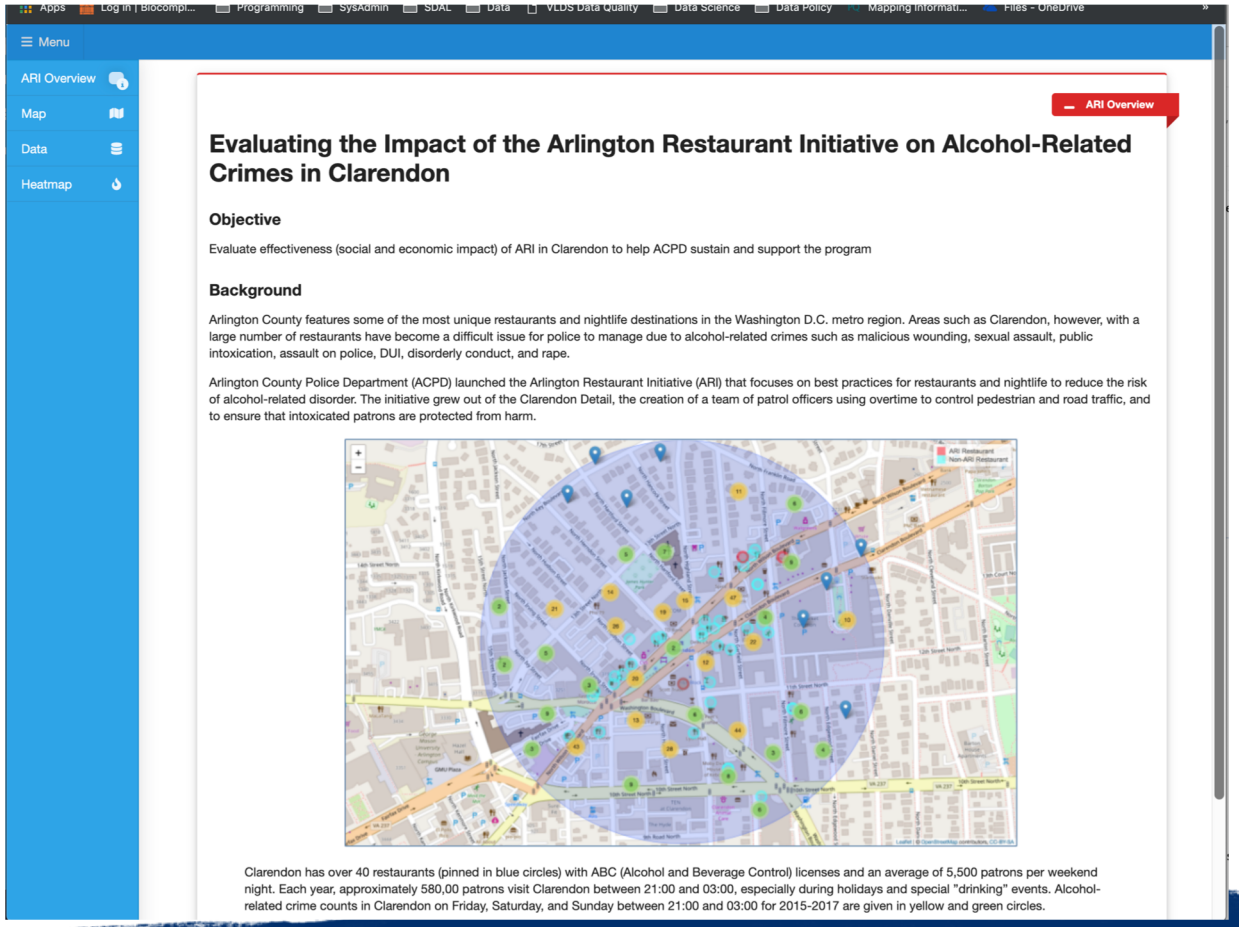
Data Visualization & Sharing

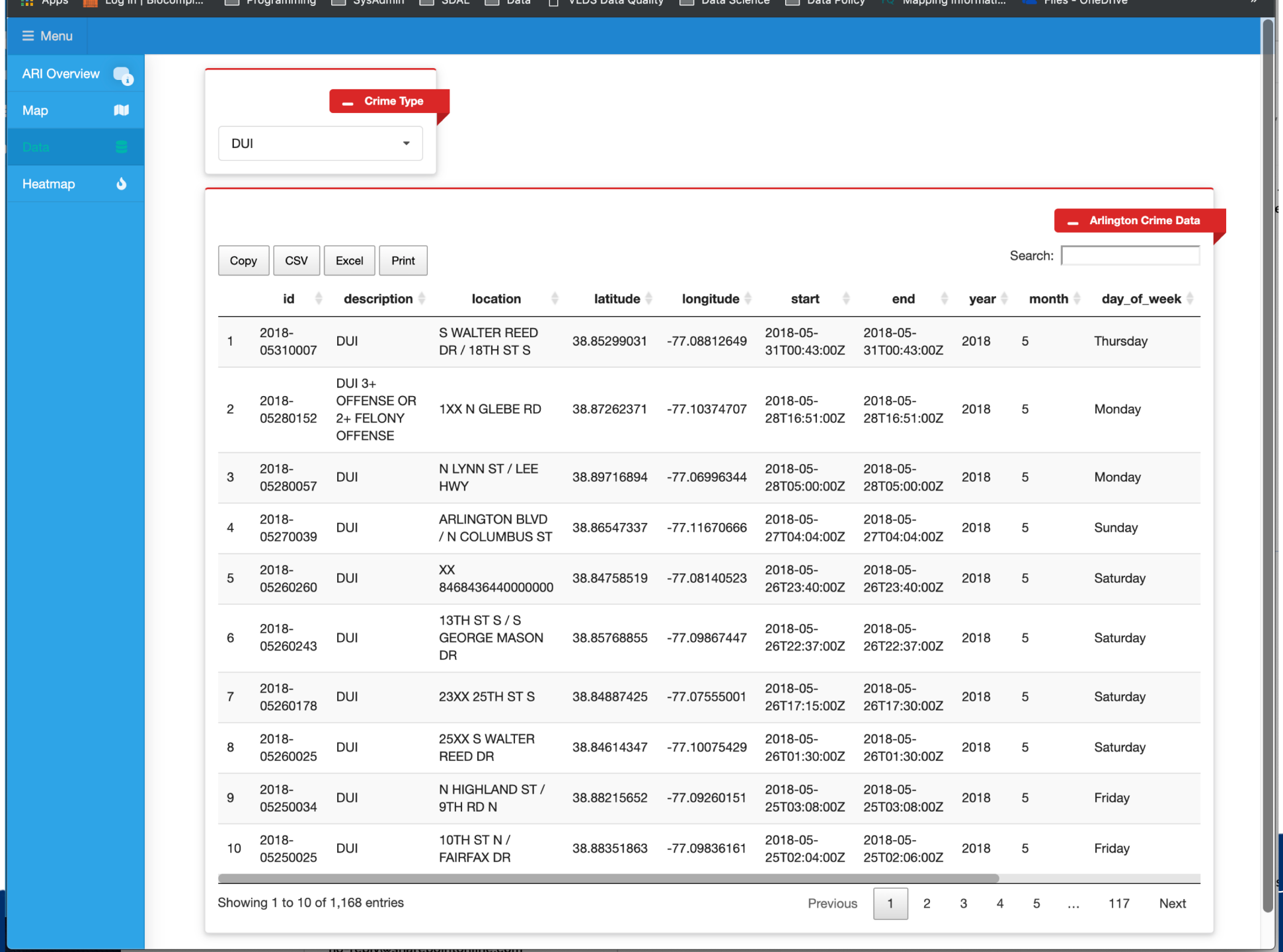
Interactive Maps (leaflet)



Data Visualization & Sharing

Dashboards (R Shiny, Django)





	id	description	location	latitude	longitude	start	end	year	month	day_of_week
1	2018-05310007	DUI	S WALTER REED DR / 18TH ST S	38.85299031	-77.08812649	2018-05-31T00:43:00Z	2018-05-31T00:43:00Z	2018	5	Thursday
2	2018-05280152	DUI 3+ OFFENSE OR 2+ FELONY OFFENSE	1XX N GLEBE RD	38.87262371	-77.10374707	2018-05-28T16:51:00Z	2018-05-28T16:51:00Z	2018	5	Monday
3	2018-05280057	DUI	N LYNN ST / LEE HWY	38.89716894	-77.06996344	2018-05-28T05:00:00Z	2018-05-28T05:00:00Z	2018	5	Monday
4	2018-05270039	DUI	ARLINGTON BLVD / N COLUMBUS ST	38.86547337	-77.11670666	2018-05-27T04:04:00Z	2018-05-27T04:04:00Z	2018	5	Sunday
5	2018-05260260	DUI	XX 8468436440000000	38.84758519	-77.08140523	2018-05-26T23:40:00Z	2018-05-26T23:40:00Z	2018	5	Saturday
6	2018-05260243	DUI	13TH ST S / S GEORGE MASON DR	38.85768855	-77.09867447	2018-05-26T22:37:00Z	2018-05-26T22:37:00Z	2018	5	Saturday
7	2018-05260178	DUI	23XX 25TH ST S	38.84887425	-77.07555001	2018-05-26T17:15:00Z	2018-05-26T17:30:00Z	2018	5	Saturday
8	2018-05260025	DUI	25XX S WALTER REED DR	38.84614347	-77.10075429	2018-05-26T01:30:00Z	2018-05-26T01:30:00Z	2018	5	Saturday
9	2018-05250034	DUI	N HIGHLAND ST / 9TH RD N	38.88215652	-77.09260151	2018-05-25T03:08:00Z	2018-05-25T03:08:00Z	2018	5	Friday
10	2018-05250025	DUI	10TH ST N / FAIRFAX DR	38.88351863	-77.09836161	2018-05-25T02:04:00Z	2018-05-25T02:06:00Z	2018	5	Friday

Menu

ARI Overview

Map

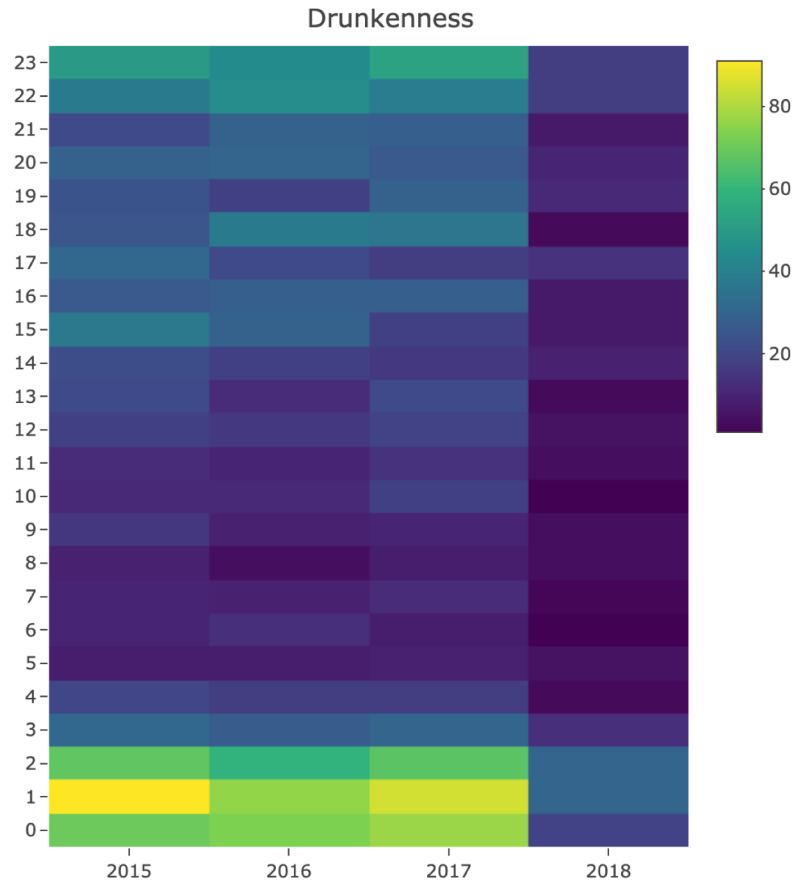
Data

Heatmap

Heatmap Control

Drunkenness

Arlington Crime Heatmap



VERSITY
IRGINIA

Menu

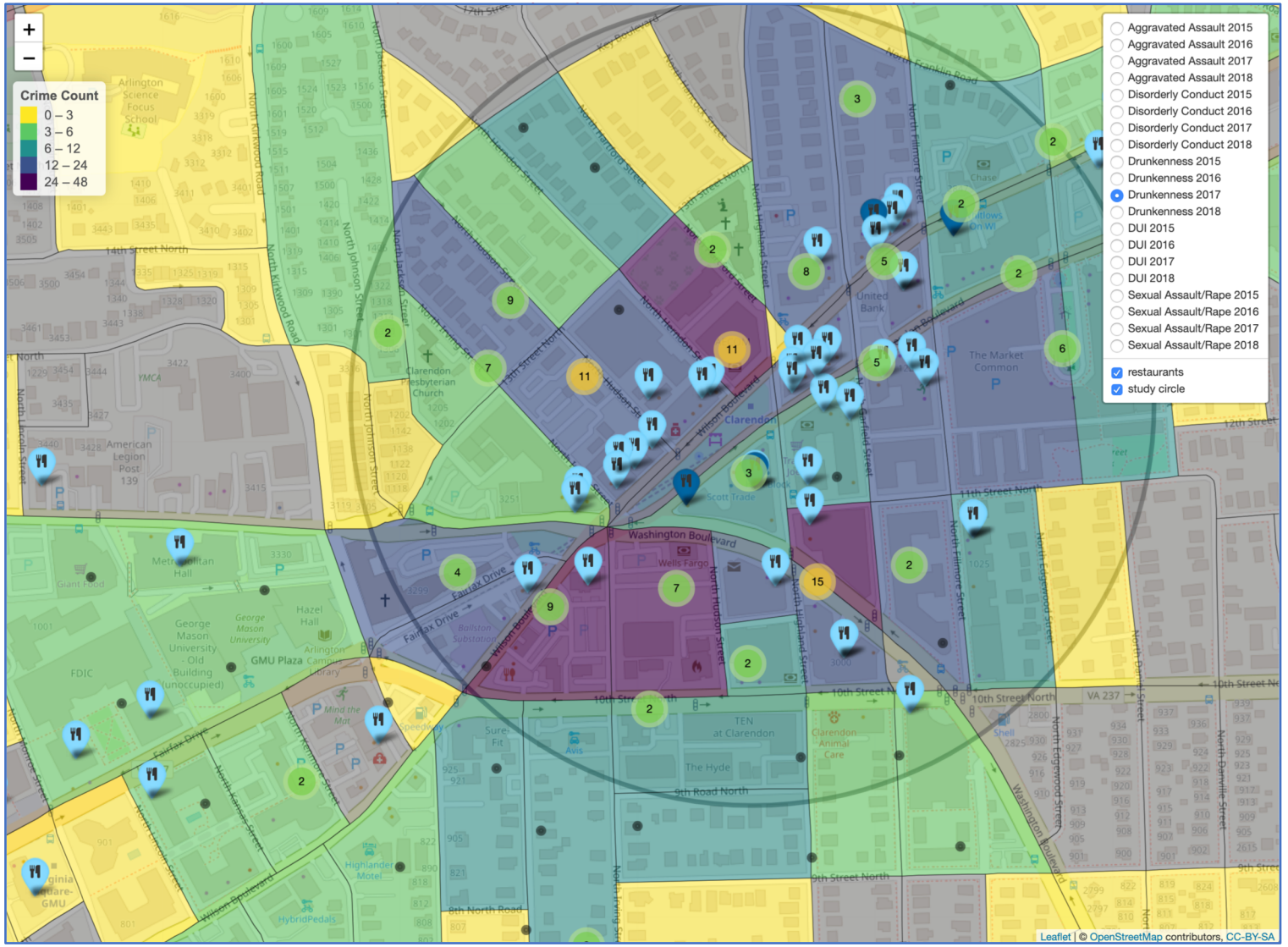
ARI Overview

Map

Data

Heatmap

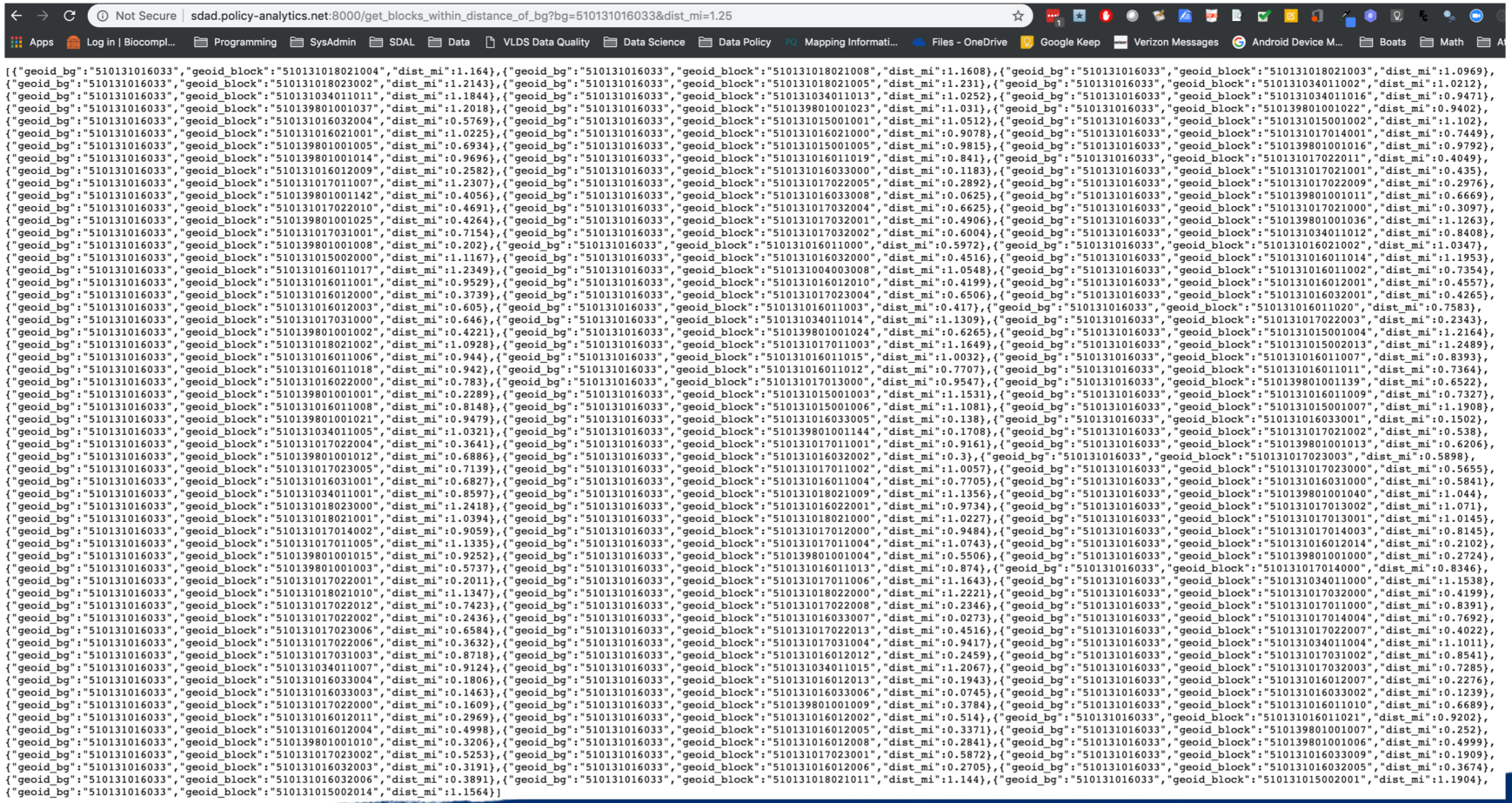
Arlington Crime Map



UNIVERSITY of VIRGINIA

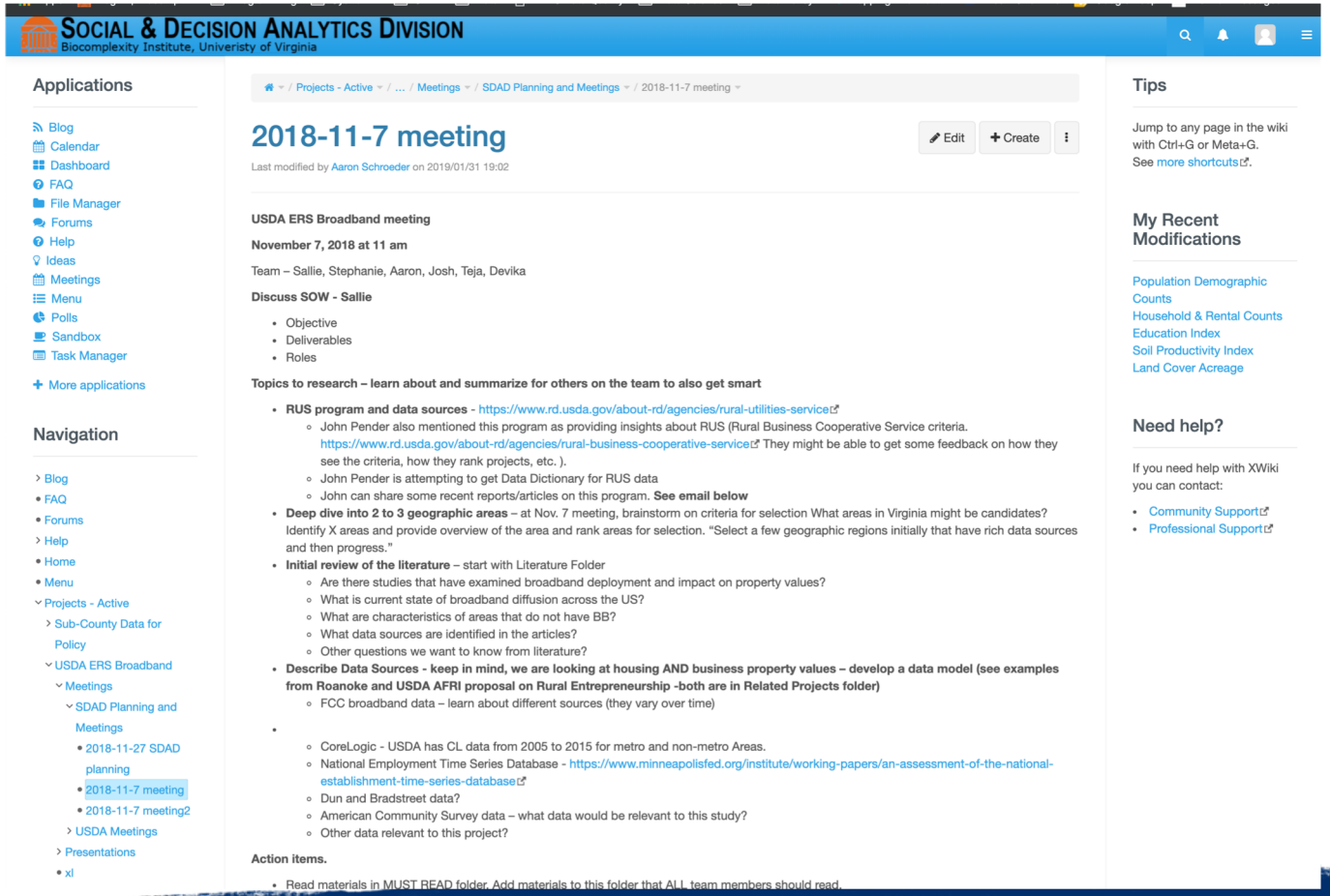
APIs (Plumber, OpenCPU)

1. Distance from BG center-points to Block center-points
2. Gravity model API



Data Visualization & Sharing

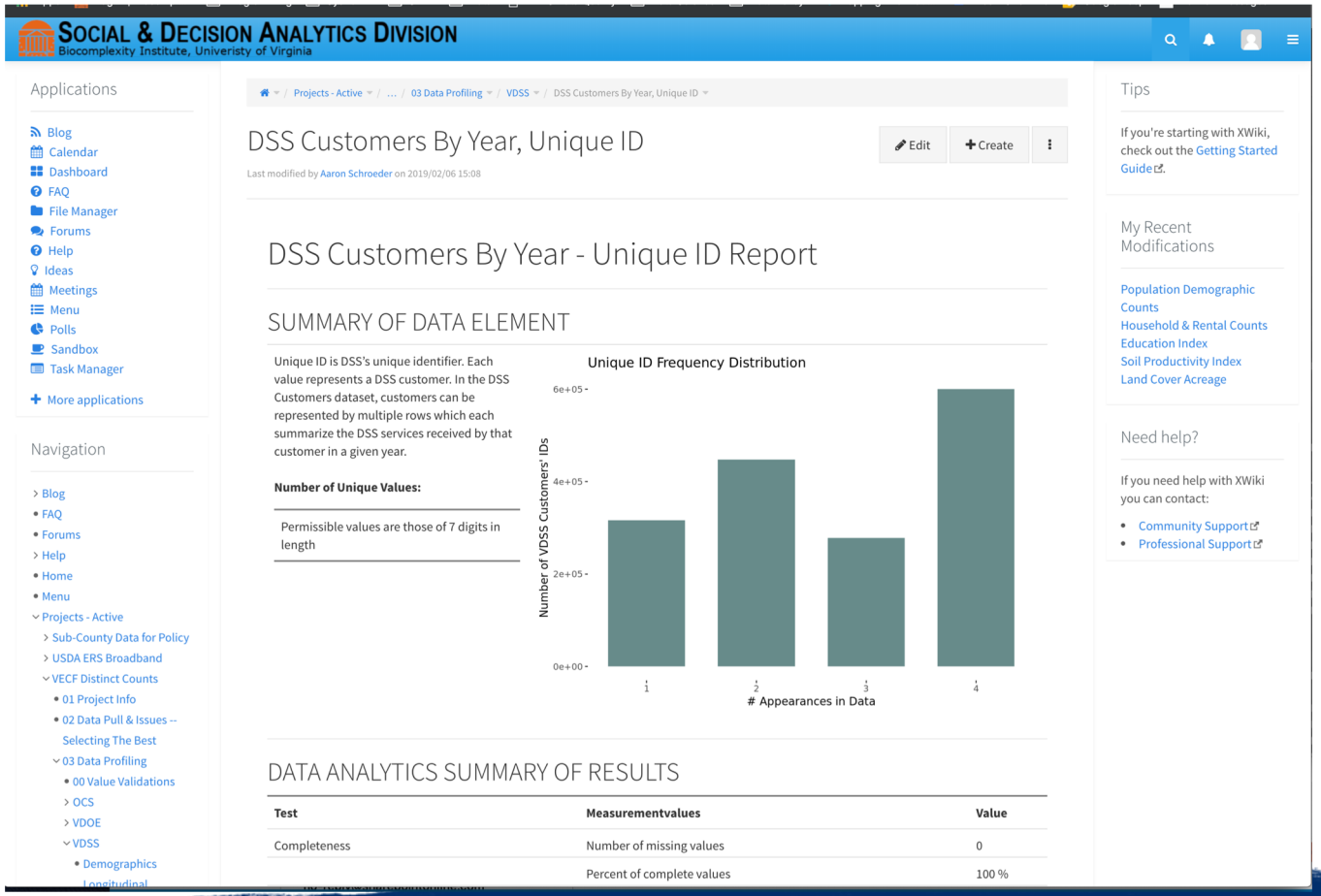
Project Wikis (XWiki)



UNIVERSITY of VIRGINIA

Data Visualization & Sharing

Project Wikis (XWiki)



UNIVERSITY of VIRGINIA

Thank You!