A Data Science Framework for the Research Use of Administrative Data

Aaron D. Schroeder

June 28, 2016

1 Introduction

When working with Administraive data, need for a more comprehensive framework than proided by OSEMN. Particularly for identication and access, and data quality (especially record and longitudinal consistency, metadata and provenance).

Plenty written in Data Science literature on modeling and analytic methodologies, the 'Analysi' section. Not enough guidance given on the other 85%, getting, profiling, and fixing data for analysis. That's the focus here. Especially relevant to use of Administrative Data.

2 Administrative Data

These are data sources collected primarily for program administration, and are structured but often have unknown quality and coverage. Despite this, they can provide useful and detailed information that could complement or supplement survey data. The statistical properties of these are elusive since they come with little to no documentation about coverage, representativeness, bias, and longitudinal gaps. In some cases, these statistical properties may be knowable, but simply have not been well-studied. [7]

Operationalization of quality and fitness for use not typical in administrative data.

Administrative data are information collected, used, and stored primarily for administrative (i.e., operational), rather than research, purposes. These data can be an excellent source of information for use in research and impact evaluation. [3]

Administrative records can be defined as data collected for the purpose of carrying out the implementation and functioning of a program or service. They are generally not collected, therefore, for statistical purposes There are many eamples of administrative data such as Social Security Administration records, unemployment records, and traffic violation records. Local school system education records consitute together one of the largest sets of administrative data in the U.S.

As administrative records are generally collected with specific decision-making purposes in mind, and as multiple sets of records must often be combined to provide a more complete decison-making context, the identity of an indidivudal unit corresponding to a given record is crucial. In the case of statistical records, no action on an individual level is ever intended and so identity at the indidividual unit level is not a requirement.

As opposed to Designed or Opportunity Data Designed Data: These are data sources that come from designed data collections such as surveys and experiments. They typically are structured – collected for a specific analytic purpose, representative of a population, and have known data quality. Opportunity Data: These are data generated on an ongoing basis as society moves through its daily paces. The data are captured in a variety of modalities including direct flows, internet searches, web crawling and scraping; volumes of data may be collected opportunistically on topics of one's choosing. These data offer possibilities for understanding human interactions at a societal scale, with rich spatial and temporal dynamics, and for detecting complex interactions and nonlinearities among variables.

There are a number of advantages to using administrative data for research, including cost and ease, reduced participant burden [3], near-universal coverage, accuracy [6], minimized bias [4], and long-term data availability [5].

Given the essential differences between administrative statistics and those based on survey samples, there are inevitably some importance differences in their attributes, their quality and their appropriateness for particular purposes.

Administrative data have greater potential to supplement rather than replace survey data. Agencies currently combine the two data sources in four key ways to cost-effectively increase efficiency and quality. Specifically, agencies use administrative data to: (1) link to survey data to create new data products; (2) supplement surveys' sample frames; (3) compare to survey data to improve accuracy and design of surveys; and (4) combine with survey data to create, or model, estimates. However, expanding the use of administrative data faces key constraints related to the access and quality of the data. While agencies and committees are taking steps to address these constraints and facilitate the process through which agencies work together to share data, individual tools may not be sufficient. A more-comprehensive framework for use by all agencies involved in data-sharing decisions that includes key questions to consider when evaluating potential use of administrative data could make the decision process more consistent and transparent. [2]

Major Difference: Provenance and metadata The provenance of designed data collection (i.e. surveys) is generally, almost by definition, available. The purpose and methods of collection are usually documented as part of the process, as well as any transformations that had to be applied during the creation of the analysis dataset.

The provenance of administrative data is, generally, not available. Personal contact with the personnel directly managing the data is often necessary to discern the methods of collection. Personal contact with policy-level personnel is often necessary to discern the purpose of collection. Also, contact with researchers who have previously used the data in their analyses can aid significantly in gleaning the quality of the administrative data for different purposes. With administrative data, provenance and metadata often must be pieced together.

Other Major Difference Longitudinal = additional focus on consistentcy.

3 A Data Science Framework for Administrative Data: So Much More Than Awesome (OSEMN)

The OSEMN process is quite limited and has been derived from a simple statistical analysis point-of-view where the data under consideration is both designed and readily available.

Mason & Wiggins (2010) gets used everywhere. They define data science according to the following five steps: (1) obtaining data, (2) scrubbing data, (3) exploring data, (4) modeling data, and (5) interpreting data. Together, these steps form the OSEMN model (which is pronounced ”awesome”). 1 ”O” usually discussed as the simple download, in an automated fashion, of an existing reasonably high-quality data set. Usually in about a paragraph. Any discussion of 2 ”S” is generally relegated to some automated processes available in the language under discussion (e.g. R, python). Most content is spent on 3 and 4. 1 and 2 are what take up the most time in the process - deserve a more formal treatment and clearer steps.

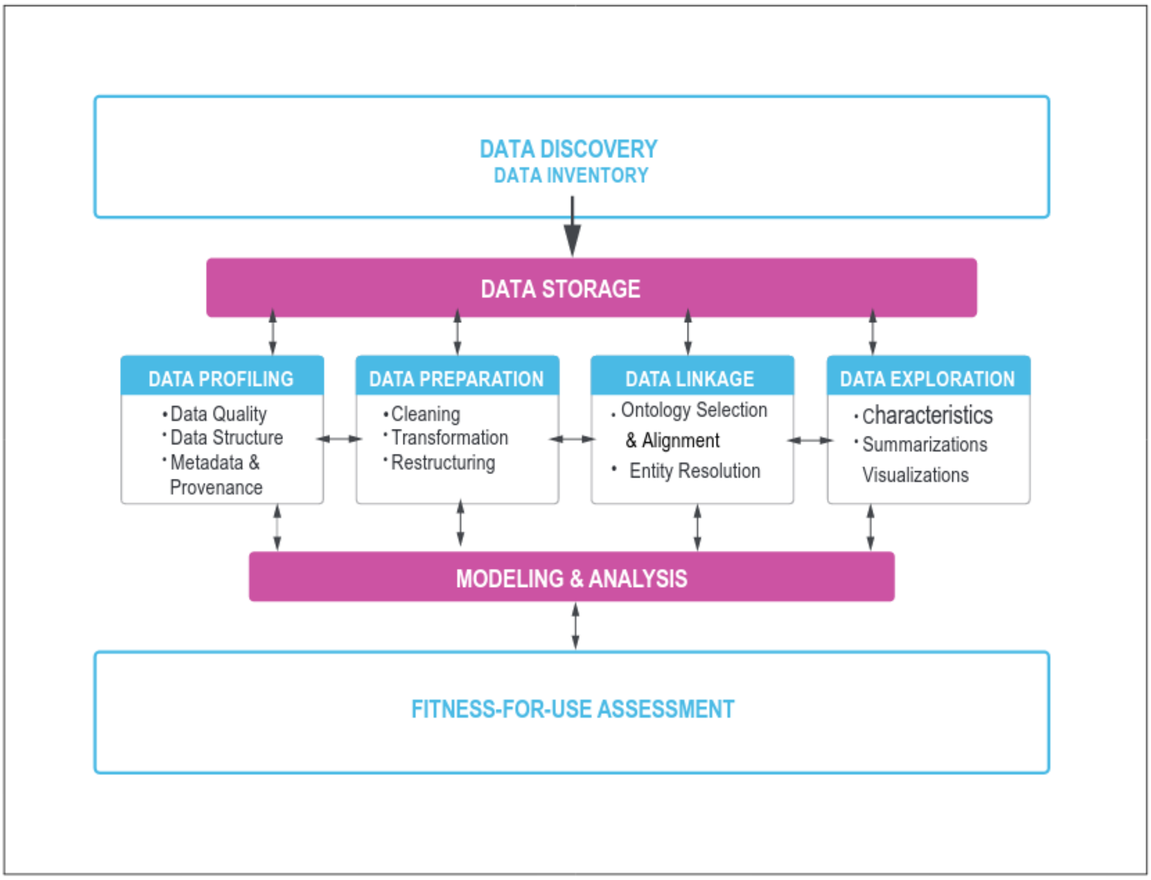


Figure 1: Data Science Framework

Here we will focus most heavily on Data Discovery, Inventory, and Profiling.

3.1 Data Discovery

3.1.1 Identification

The first problem we run into when determining appropraite data targets is who to interview? One approach is to start by simply brainstorming possible data sources with the goal of thinking as broadly and imaginatively as possible in order to assemble a list of potential sources. A more deliberative approach would be to start with a technique like quota sampling. Following this initial identifcation step, we can then procede with another technique called snowballing. The purpose of the initial quota sample is to make sure that you start with a set of interviewees that represent the most obvious set of data stakeholders. This approach is followed by snowballing, however, because it is quite unlikely that you will ‘guess’ all of the appropriate data stakeholders at the beginning, and you will need to solicit the opinion of the first interviewees as to how to proceed.

Although quota sampling is a non-probability method, it is similar to probability sampling. Like a probability sampling method, quota sampling 'addresses the issue of representativeness, although the two methods approach the issue quite differently' [1], p 196. How does quota sampling work? First, the researcher begins with 'a matrix, or table, describing the characteristics of the target population. If the researcher needed a national quota sample, for instance, she "would need to know what proportion of the national population is urban, eastern, male, under 25, white, working class, and the like, and all the other permutations of such a matrix" (196). After constructing this matrix and assigning a relative proportion assigned to each cell in the matrix, the researcher should collect data from people having all the characteristics of a give cell. The researcher could then go farther and assign 'a weight' to each person in each cell 'appropriate to his or her portion of the total population.'

Snowballing is a popular technique used in network studies (Wasserman and Faust, 1994), particularly in situations where stakeholders or other interested representatives are not easily identifiable (Goldenberg, 1992). Hence, this strategy is extremely useful in helping identify all relevant data stakeholders. Snowballing is a simple process of expanding the zone of contacts through initial contacts. The process begins by identifying an initial group of stakeholders, either those already involved in the preliminary stages of the process or those identified via another mechanism (e.g. quota sampling). These actors or participants are then asked to identify those individuals whom they feel should be involved in the process as well. This is the "first-order" zone. The researcher then proceeds to contact those actors (whether individuals or groups) and proceeds to have these "second-order" actors, further identify others who they think would have an interest in the project or process (Wasserman and Faust, 1994: 34; see also Goldenberg, 1992; Babbie, 1998; Doreian and Woodward, 1992).

The process can involve asking respondents to review a "fixed list" from preliminary research or to simply brainstorm to identify those stakeholders that they think are important to add to the list of those whom they think should be included. If being asked to brainstorm, it is advisable to precede this task with another that may stimulate thinking. The key to this process is to be as exhaustive as possible. The purpose of snowballing here is to find, as quickly as possible, the self-limiting reference system of the new network of dataset stakeholders. By self-limiting we mean that after a few iterations of snowballing, the names of suggested new interviewees begin to be repeated. When most of the names suggested have already been interviewed, you have reached the end of the snowball process.

3.1.2 Establishing Access

3.1.2.1 Stakeholder Analysis & Data Access Network Establishment

When a standard access protocol does not exist, necessary to construct network of authority to guarantee data access. People from at least there levels are almost always necessary:

- Policy-level permission

  (Agranoff,1990a, 1990b). need to develop policies or strategies that will support access at the program and operational levels of the data organization. In order for this activity to occur, shared problem definitions and agreed-upon courses of action must be developed at the executive/policy decision level (141). For our purposes it is probably enough to say that the objectives of the goal setting subnetwork is to (1) develop the vision and approach of the data access activity, (2) identify additional stakeholders that need to be added to the goal setting network, and (3) as well as populate the program and operational implementation subnetworks.

- Program-level commitment

Satisfying program-level requirements, MOAs. RUDAs, specification of purpose of access, etc.

- Operational-level willingness & ability

3.1.3 Data Source Screening and Inventory

The first step before conducting a full data inventory is to screen the data sources, identifying which sources are worthy of a deeper look and which are worthy of consideration for profiling. The screening includes five questions and a qualitative evaluation of purpose, data collection method, selectivity, accessibility, and description. Figure 2 shows a sample screening instrument from a recent project.

Following an initial screening inventory, a subset of the sources are selected for a full inventory. Figure 3 shows a sample inventory instrument.

3.2 Data Profiling

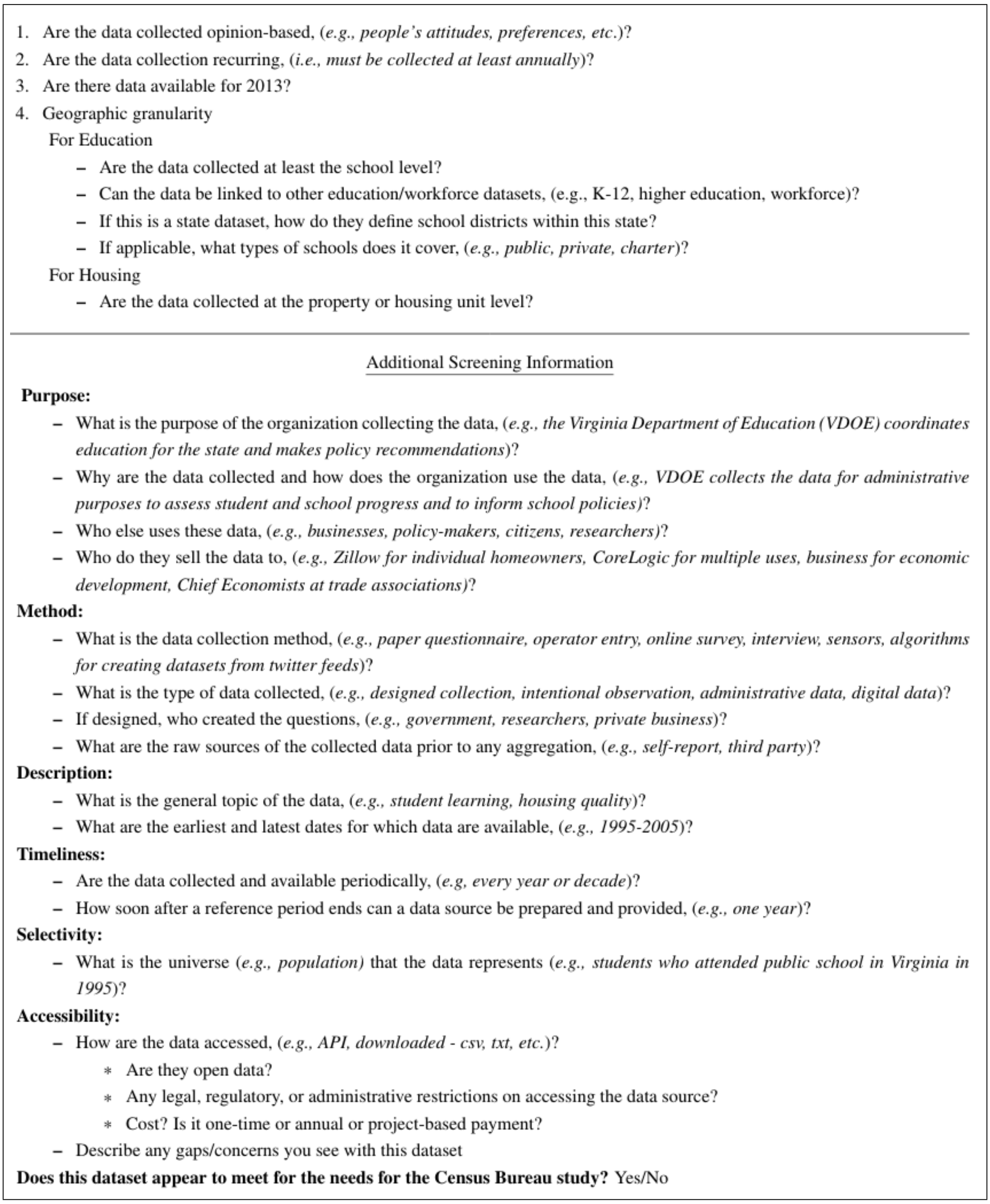
Data profiling starts with a determination of both the quality of the data and its utility to the project at hand. There are numerous rubrics available, but a useful initial assessment of data quality can be achieved through statistical analysis of data field completeness, data field value correctness, and logical consistency between fields and between records (Redman, Wang, DOD, others). In addition, it is useful to ascertain the spread of unique values within a field, as well as the rate of duplication at the record level. For an assessment of dataset utility, the dataset's structure should be analyzed. This is to determine how well the dataset has been structured for the purposes of the intended analyses. The the state of the dataset's metadata should also be analyzed to determine how well observational units and their attributes are defined. An important feature of the data profiling process is that discovered issues are only described and not actually "fixed". The appropriate fix will depend upon the specific needs of the research. If the prescribed "fix" is not appropriate or even possible there would be no need for any action and attempting a fix at this stage could result in wasted time and effort. For example, it may be appropriate to simply normalize city zoning entries into Residential or non-residential versus painfully re-categorizing every missing entry into the 38 zoning classfication.

3.2.1 Data Structure - Is the data appropriately structured for purposes of analysis?

Because datasets are often created for reasons of administration and reporting, and further that there are generally few constraints on their organization, administrative datasets are often constructed in manners not conducive to statistical analyses. During the data profiling step, issues about data structure are identified. During the data transformation step decisions on how to restructure data are made and executed. An example from the housing case study is given in Figure 7. The dataset provided was comprised of single records with 128 fields. All fields in the record were keyed to the variable"List Number." Structured this way, it is not possible to analyze property changes over time, even though the data does in fact exist within the dataset. That is, a restructuring is necessary that pulls out and re-relates the property information to a different key (here, Parcel ID). This situation occurs often as a result of the structural issue 'combined observational unit types' discussed below.

3.2.1.1 Missing Variables

'Missing variables' means that a dataset has values in column headers instead of variable names. It is not uncommon to receive tabular datasets that have been designed for the purpose of presentation,



1. Are the data collected opinion-based, (*e.g., people's attitudes, preferences, etc.*)?
2. Are the data collection recurring, (*i.e., must be collected at least annually*)?
3. Are there data available for 2013?
4. Geographic granularity

   For Education

   - Are the data collected at least the school level?
   - Can the data be linked to other education/workforce datasets, (e.g., K-12, higher education, workforce)?
   - If this is a state dataset, how do they define school districts within this state?
   - If applicable, what types of schools does it cover, (*e.g., public, private, charter*)?

   For Housing

   - Are the data collected at the property or housing unit level?

Additional Screening Information

Purpose:

- What is the purpose of the organization collecting the data, (*e.g., the Virginia Department of Education (VDOE) coordinates education for the state and makes policy recommendations*)?
- Why are the data collected and how does the organization use the data, (*e.g., VDOE collects the data for administrative purposes to assess student and school progress and to inform school policies)*?
- Who else uses these data, (*e.g., businesses, policy-makers, citizens, researchers)*?
- Who do they sell the data to, (*e.g., Zillow for individual homeowners, CoreLogic for multiple uses, business for economic development, Chief Economists at trade associations)*?

Method:

- What is the data collection method, (*e.g., paper questionnaire, operator entry, online survey, interview, sensors, algorithms for creating datasets from twitter feeds*)?
- What is the type of data collected, (*e.g., designed collection, intentional observation, administrative data, digital data*)?
- If designed, who created the questions, (*e.g., government, researchers, private business*)?
- What are the raw sources of the collected data prior to any aggregation, (*e.g., self-report, third party*)?

Description:

- What is the general topic of the data, (*e.g., student learning, housing quality*)?
- What are the earliest and latest dates for which data are available, (*e.g., 1995-2005*)?

Timeliness:

- Are the data collected and available periodically, (*e.g, every year or decade*)?
- How soon after a reference period ends can a data source be prepared and provided, (*e.g., one year*)?

Selectivity:

- What is the universe (*e.g., population)* that the data represents (*e.g., students who attended public school in Virginia in 1995*)?

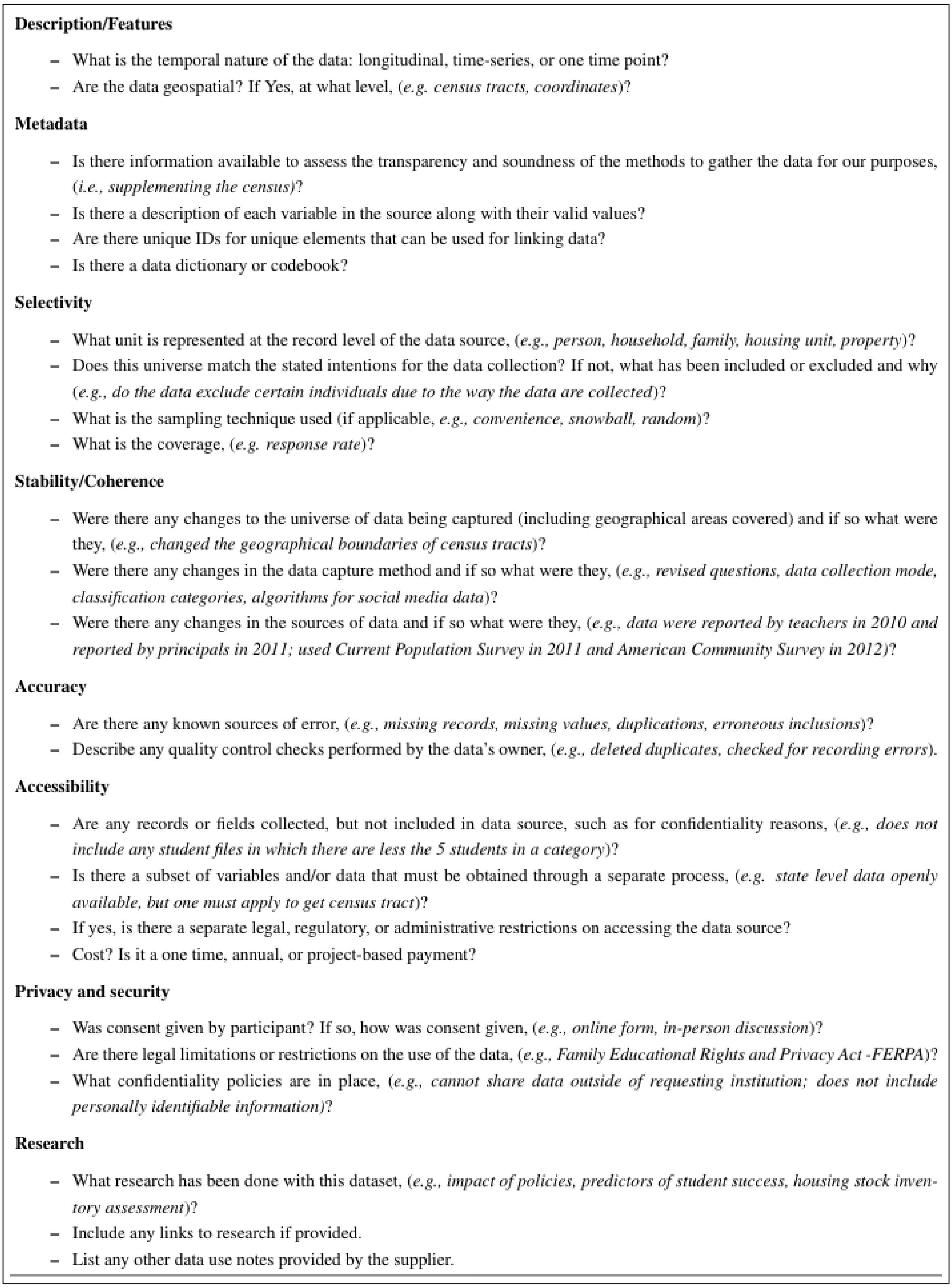
Accessibility:

- How are the data accessed, (*e.g., API, downloaded - csv, txt, etc.*)?
  - Are they open data?
  - Any legal, regulatory, or administrative restrictions on accessing the data source?
  - Cost? Is it one-time or annual or project-based payment?
- Describe any gaps/concerns you see with this dataset

Does this dataset appear to meet for the needs for the Census Bureau study? Yes/No

Figure 2: Sample Data Source Screening Instrument

not analysis, where variables form both the rows and columns, and column headers are values, not variable names. While this type of structural issue was not experienced with the datasets under consideration, it is an issue to be cognizant of, especially, as our experience indicates, when dealing with agency datasets released as tabular data in spreadsheet files. Such files are often initially produced for the purpose of conveying aggregated summary data to decision-makers and are not constructed like a traditional dataset (i.e. one record per row with clear field/column titles).



Description/Features

- What is the temporal nature of the data: longitudinal, time-series, or one time point?
- Are the data geospatial? If Yes, at what level, (*e.g. census tracts, coordinates*)?

Metadata

- Is there information available to assess the transparency and soundness of the methods to gather the data for our purposes, (*i.e., supplementing the census)*?
- Is there a description of each variable in the source along with their valid values?
- Are there unique IDs for unique elements that can be used for linking data?
- Is there a data dictionary or codebook?

Selectivity

- What unit is represented at the record level of the data source, (*e.g., person, household, family, housing unit, property*)?
- Does this universe match the stated intentions for the data collection? If not, what has been included or excluded and why (*e.g., do the data exclude certain individuals due to the way the data are collected*)?
- What is the sampling technique used (if applicable, *e.g., convenience, snowball, random*)?
- What is the coverage, (*e.g. response rate*)?

Stability/Coherence

- Were there any changes to the universe of data being captured (including geographical areas covered) and if so what were they, (*e.g., changed the geographical boundaries of census tracts*)?
- Were there any changes in the data capture method and if so what were they, (*e.g., revised questions, data collection mode, classification categories, algorithms for social media data*)?
- Were there any changes in the sources of data and if so what were they, (*e.g., data were reported by teachers in 2010 and reported by principals in 2011; used Current Population Survey in 2011 and American Community Survey in 2012)*?

Accuracy

- Are there any known sources of error, (*e.g., missing records, missing values, duplications, erroneous inclusions*)?
- Describe any quality control checks performed by the data's owner, (*e.g., deleted duplicates, checked for recording errors*).

Accessibility

- Are any records or fields collected, but not included in data source, such as for confidentiality reasons, (*e.g., does not include any student files in which there are less the 5 students in a category*)?
- Is there a subset of variables and/or data that must be obtained through a separate process, (*e.g. state level data openly available, but one must apply to get census tract*)?
- If yes, is there a separate legal, regulatory, or administrative restrictions on accessing the data source?
- Cost? Is it a one time, annual, or project-based payment?

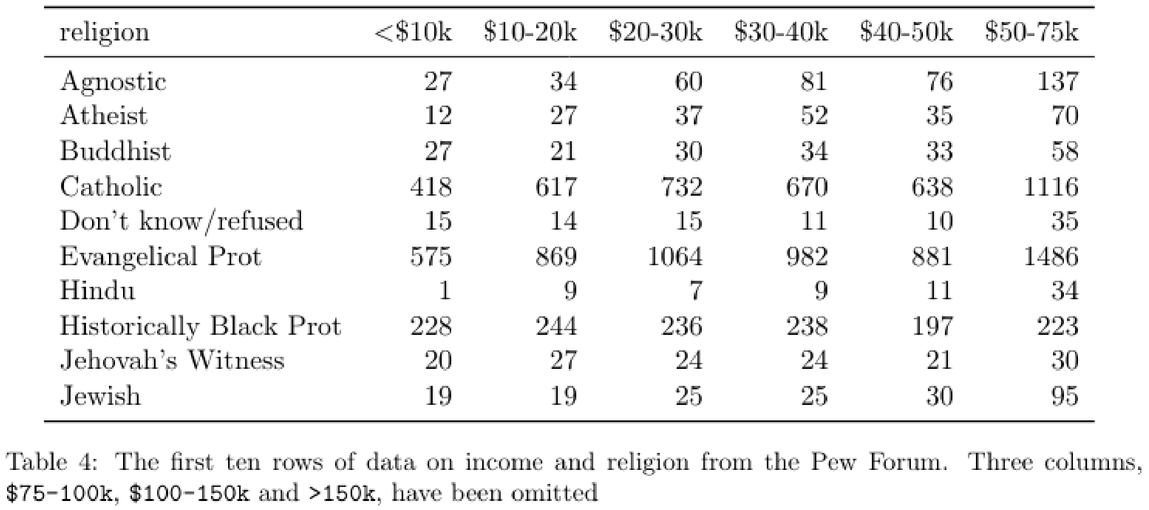
Privacy and security

- Was consent given by participant? If so, how was consent given, (*e.g., online form, in-person discussion*)?
- Are there legal limitations or restrictions on the use of the data, (*e.g., Family Educational Rights and Privacy Act -FERPA*)?
- What confidentiality policies are in place, (*e.g., cannot share data outside of requesting institution; does not include personally identifiable information)*?

Research

- What research has been done with this dataset, (*e.g., impact of policies, predictors of student success, housing stock inventory assessment*)?
- Include any links to research if provided.
- List any other data use notes provided by the supplier.

Figure 3: Sample Data Source Inventory Instrument



religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, `$75-100k`, `$100-150k` and `>150k`, have been omitted

Figure 4: Missing Variables

3.2.1.2 Combined Variables

Combined variables refers to the condition where more than one variable is represented in a record field. Sometimes, especially, after some form of correction to the previous problem of missing variables has been address, we end up with column variable names comprised of a combination of multiple underlying variable names. It should be note that many times, like duplication, this is an issue of problem definition. A common example with administrative record files occurs when we use parts of the date _of _birth field. For example, in the education case study, there was a need at one point to categorize students by birth _year. To achieve this categorization, it was first necessary to divide out from the date _of _birth field the variables birth _month, birth _day, and birth _year. While this is often achieved within the programming of a query (i.e. select datepart('year', date _of _birth), what in fact is occurring is the separation of previously combined variables.

Here is an example that combines gender and age group in the same variable:

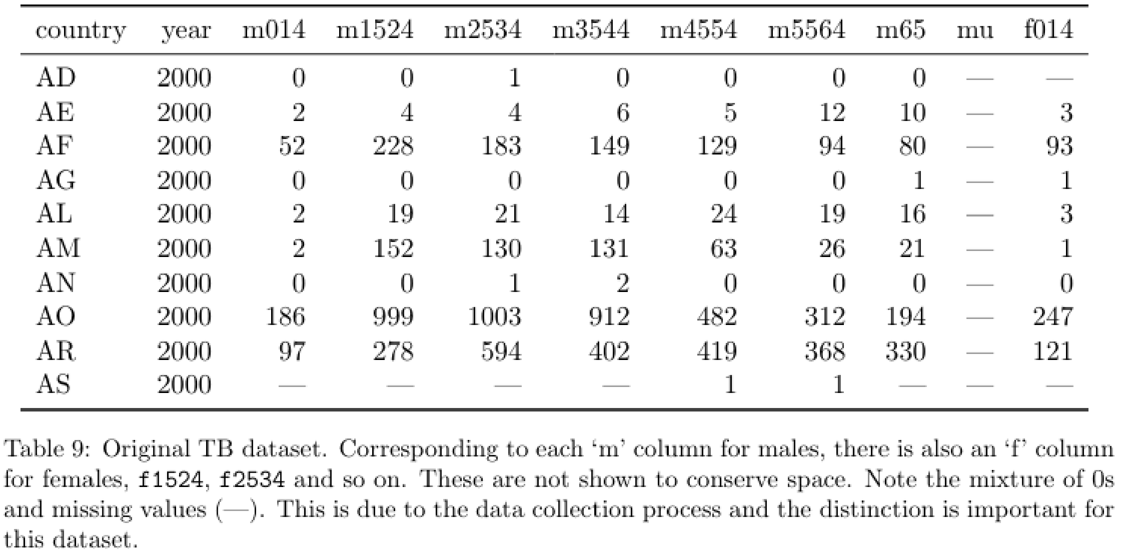
3.2.1.3 Multiple Observation Directions

A particularly messy structural issue occurs when a dataset has variable names in both columns and rows. For example, a dataset with an field/column for each day of the month (on the horizontal) and a row title for 'month' (on the vertical). This situation occurs most often when the data provided comes in the form of and cross-tabulated aggregate data.

Table 11 shows daily weather data from the Global Historical Climatology Networkfor one weather station (MX17004) in Mexico for five months in 2010. It has variables in individual columns (id,year,month), spread across columns (day, d1-d31) and across rows(tmin,tmax) (minimum and maximum temperature).

3.2.1.4 Combined Observational Unit Types

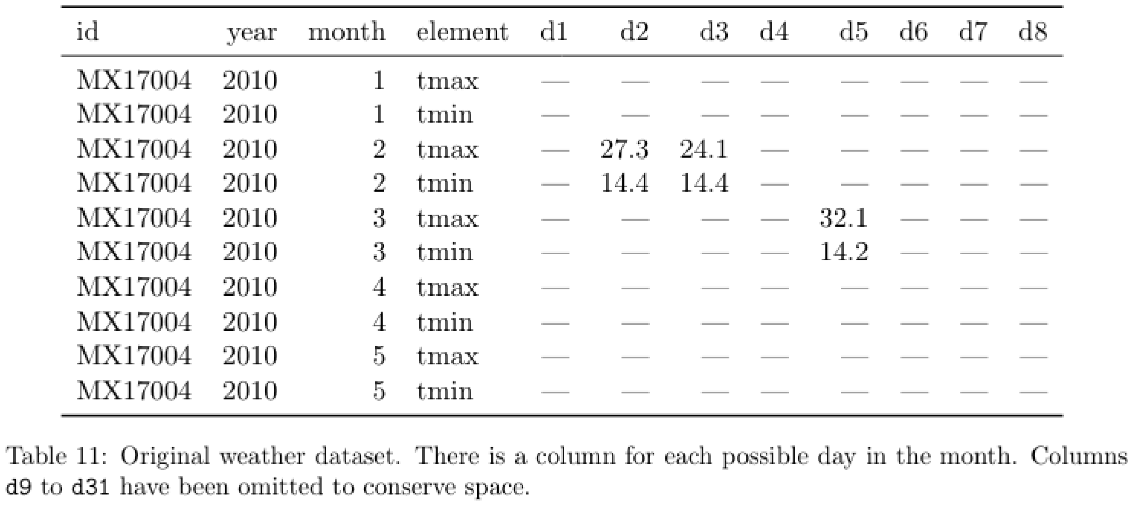
Within administrative records, multiple types of data are often combined for expediency. By multiple types we mean different sets of data fields, each set representing a different type of observational



country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Table 9: Original TB dataset. Corresponding to each 'm' column for males, there is also an 'f' column for females, `f1524`, `f2534` and so on. These are not shown to conserve space. Note the mixture of 0s and missing values (—). This is due to the data collection process and the distinction is important for this dataset.

Figure 5: Combined Variables



id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—

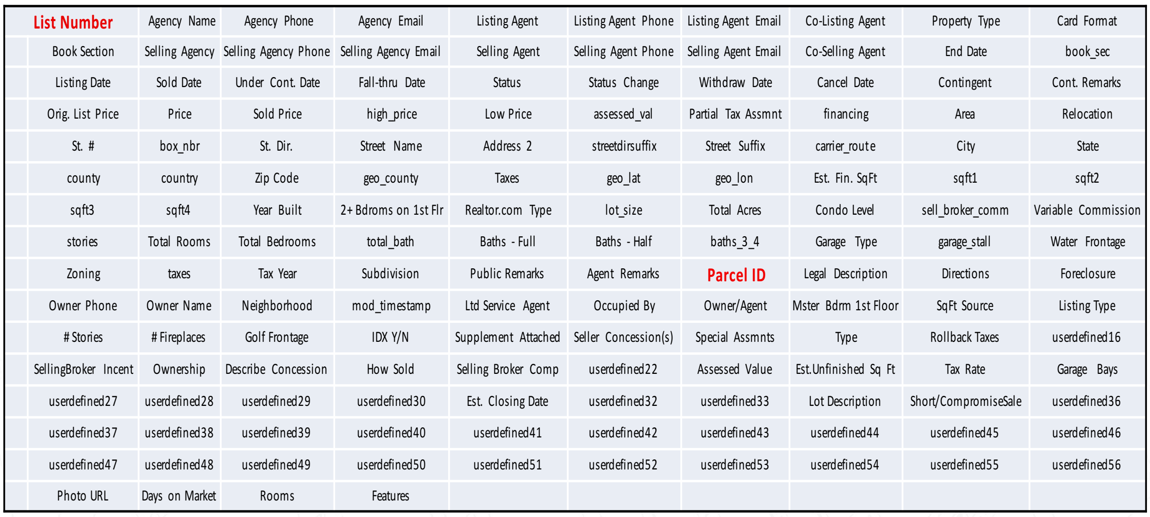
Table 11: Original weather dataset. There is a column for each possible day in the month. Columns `d9` to `d31` have been omitted to conserve space.

Figure 6: Variables in Two Directions.

unit (e.g. property information and listing agent information in the same record). The observational unit types necessary to the project at hand need to be separated out into individual observations or individual datasets in the restructuring phase.

One example encountered in the education case study had to do with a dataset containing both individual demographic data and a periodic measurement of weekly attendance where demographic data and weekly attendance are separate observational units and needed to be in separate datasets for the purposes of the study.

An example from the housing case study is given in Figure 3.3. The dataset provided was-comprised of single records with 128 fields. Each original record was identified by a unique“List Number.” However, if a parcel was listed twice it would have two different “List Num-bers.” As a result, changes in a property or parcel over time could not be tracked from theserecords because the structure only identified the list number not the parcel number. Changing the structure to include the "Parcel ID" allowed the required historical tracking of changes.



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						

Figure 7: Combined data types of *List Number* and **Parcel ID** from MLS data table.

3.2.1.5 Divided Observation Unit Type

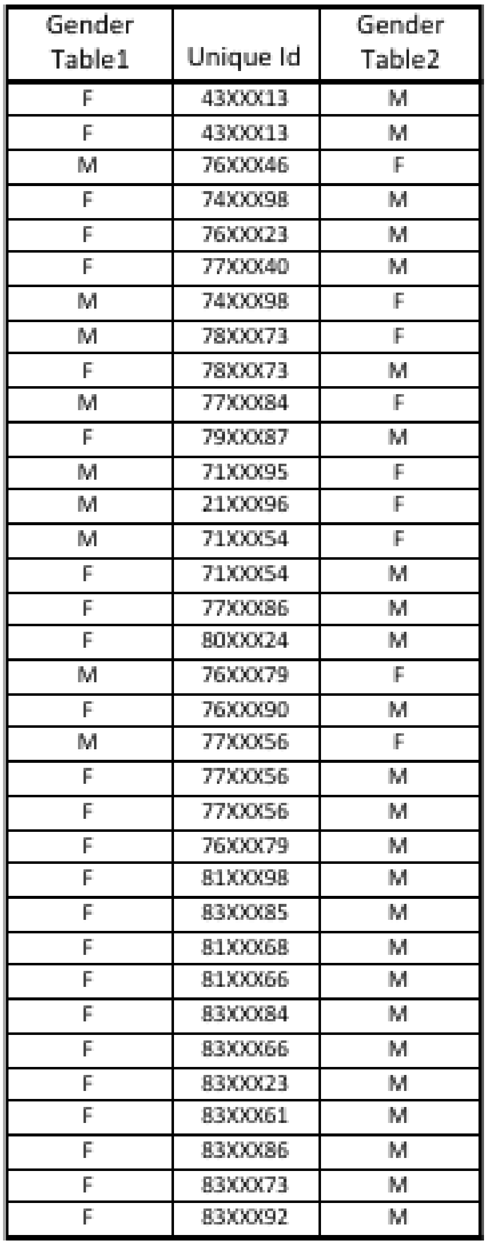
Within administrative data systems, it is also not uncommon to find that a single observation unit type has been split among multiple datasets.This is similar to the consistency discussion of multiple observations with overlapping demographic information. The difference here is that you may have information split among several datasets. For example, there may be, as was the case with one state's educational records, separate tables that duplicate the collection of student demographics. Figure **??** captures gender mismatches across two tables from the same education record information system, linked on the Unique Id of the student. Decisions on whether and how to transform inconsistent data as a result of divided observation unit type need to factor in the magnitude of the issue as well as the ability to accurately correct the data in a timely enough fashion given the project at hand.

3.2.1.6 Web-Scraping for Data: A Many Problems Real-Life Example

Scraping data from an emergency services online data system:

3.2.2 Data Quality Profiling & the Dimensions of Quality

A considerable amount of data quality research involves investigating and describing various categories of desirable attributes (or dimensions) of data. These lists commonly include accuracy, correctness, currency, completeness and relevance, as described in Chapter 2. Nearly 200 such terms have been identified and there is little agreement in their nature (are these concepts, goals or criteria?), their definitions or measures (Wang et al., 1993). Here we have let a typology emerge form the project data work. This typology consists of five data quality areas: completeness, value validity, consistency, uniqueness, and duplication. Regardless the typology chosen, the final judgment of data quality is measured by adherence of a dataset to a set of data quality rules. Like data

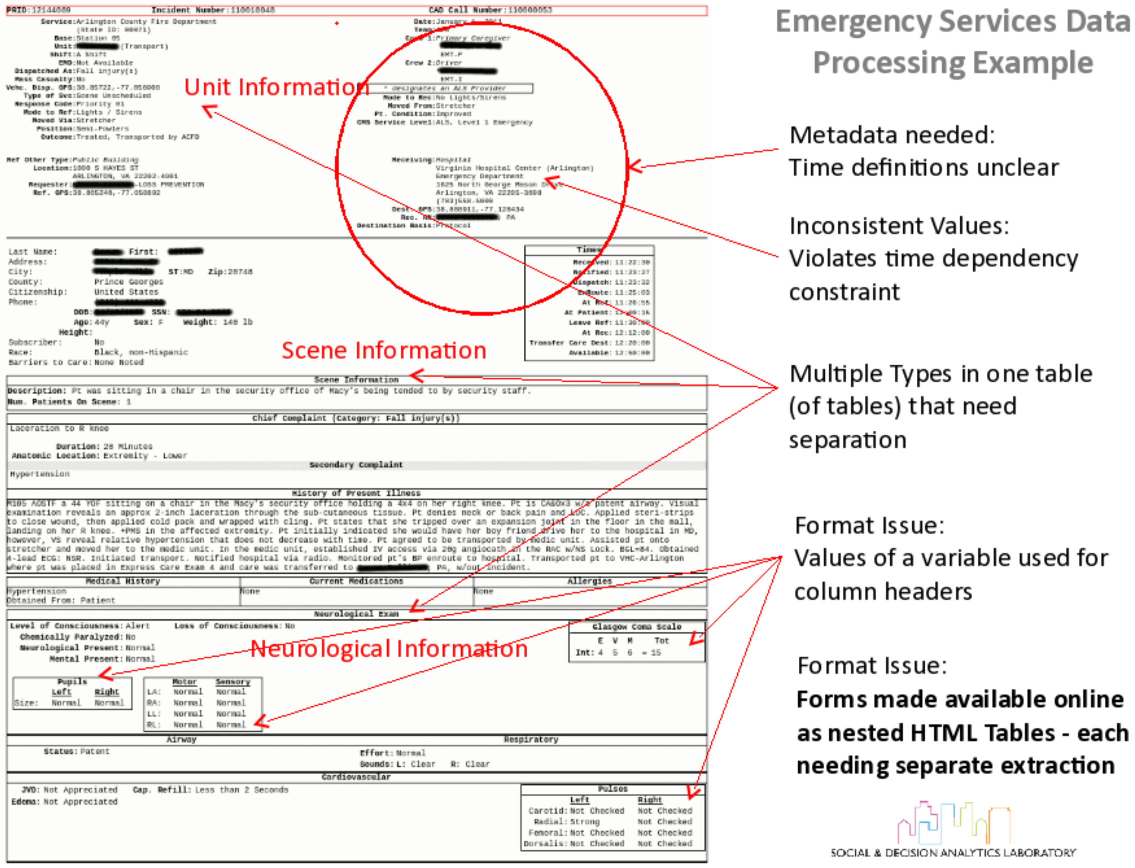


Gender Table1	Unique Id	Gender Table2
F	43XXX13	M
F	43XXX13	M
M	76XXX46	F
F	74XXX98	M
F	76XXX23	M
F	77XXX40	M
M	74XXX98	F
M	78XXX73	F
F	78XXX73	M
M	77XXX84	F
F	79XXX87	M
M	71XXX95	F
M	21XXX96	F
M	71XXX54	F
F	71XXX54	M
F	77XXX86	M
F	80XXX24	M
M	76XXX79	F
F	76XXX90	M
M	77XXX56	F
F	77XXX56	M
F	77XXX56	M
F	76XXX79	M
F	81XXX98	M
F	83XXX85	M
F	81XXX68	M
F	81XXX66	M
F	83XXX84	M
F	83XXX66	M
F	83XXX23	M
F	83XXX61	M
F	83XXX86	M
F	83XXX73	M
F	83XXX92	M

Figure 8: Divided Type

qulibrary(ggplot2)ality attributes, these rules can take one of several forms. Here we choose a typology consisting of three rule-types employed by the DoD in its data quality management efforts. These are: null constraint rules, domain validation rules, and relationship validation rules. Code is developed to enforce these rules. For our study, examples are given in psuedo-code:

Null Constraints select sqft from housing where sqft = 0 or sqft = ” or sqrt = NULL Domain Validation select yearbuilt from housing where yearbuilt is between 1920 and 2015 Relationship Validation select all from housing where type = multifamily and numunits greater than 1



PRID:12144089 Incident Number:110018048 CAD Call Number:110000053

Service:Arlington County Fire Department
(State ID: 00071)
Base:Station 05
Unit:[redacted](Transport)
Shift:A Shift
END:Not Available
Dispatched As:Fall injury(s)
Mass Casualty:No
Vehc. Disp. GPS:38.85722,-77.056908
Type of Svc:Scene Unscheduled
Response Code:Priority 01
Mode to Ref:Lights / Sirens
Moved Via:Stretcher
Position:Semi-Fowlers
Outcome:Treated, Transported by ACFD

Ref Other Type:Public Building
Location:1000 S HAYES ST
ARLINGTON, VA 22202-4001
Requester:[redacted]-LOSS PREVENTION
Ref. GPS:38.865246,-77.058892

Date:January 4, 2011
Team:[illegible]
Crew 1:Primary Caregiver
[redacted]
EMT.P
Crew 2:Driver
[redacted]
EMT.I
* designates an ALS Provider
Mode to Rec:No Lights/Sirens
Moved From:Stretcher
Pt. Condition:Improved
CMS Service Level:ALS, Level 1 Emergency

Receiving:Hospital
Virginia Hospital Center (Arlington)
Emergency Department
1625 North George Mason Drive
Arlington, VA 22205-3698
(703)558.5000
Dest. GPS:38.888911,-77.128434
Rec. RN:[redacted], PA
Destination Basis:Protocol

Last Name: [redacted] First: [redacted]
Address: [redacted]
City: [redacted] ST:MD Zip:20748
County: Prince Georges
Citizenship: United States
Phone: [redacted]
DOB:[redacted] SSN: [redacted]
Age:44y Sex: F Weight: 148 lb
Height:
Subscriber: No
Race: Black, non-Hispanic
Barriers to Care:None Noted

Times	
Received:	11:22:30
Notified:	11:23:27
Dispatch:	11:23:32
EnRoute:	11:25:03
At Ref:	11:28:55
At Patient:	11:09:16
Leave Ref:	11:39:00
At Rec:	12:12:00
Transfer Care Dest:	12:20:00
Available:	12:40:00

Scene Information
Description: Pt was sitting in a chair in the security office of Macy's being tended to by security staff.
Num. Patients On Scene: 1

Chief Complaint (Category: Fall injury(s))
Laceration to R knee
Duration: 28 Minutes
Anatomic Location: Extremity - Lower
Secondary Complaint
Hypertension

History of Present Illness
M105 AOSTF a 44 YOF sitting on a chair in the Macy's security office holding a 4x4 on her right knee. Pt is CA&Ox3 w/a patent airway. Visual examination reveals an approx 2-inch laceration through the sub-cutaneous tissue. Pt denies neck or back pain and LOC. Applied steri-strips to close wound, then applied cold pack and wrapped with cling. Pt states that she tripped over an expansion joint in the floor in the mall, landing on her R knee. +PMS in the affected extremity. Pt initially indicated she would have her boy friend drive her to the hospital in MD, however, VS reveal relative hypertension that does not decrease with time. Pt agreed to be transported by medic unit. Assisted pt onto stretcher and moved her to the medic unit. In the medic unit, established IV access via 20g angiocath in the RAC w/NS Lock. BGL=84. Obtained 4-lead ECG: NSR. Initiated transport. Notified hospital via radio. Monitored pt's BP enroute to hospital. Transported pt to VHC-Arlington where pt was placed in Express Care Exam 4 and care was transferred to [redacted], PA, w/out incident.

Medical History	Current Medications	Allergies
Hypertension Obtained From: Patient	None	None

Neurological Exam
Level of Consciousness: Alert Loss of Consciousness: No
Chemically Paralyzed: No
Neurological Present: Normal
Mental Present: Normal

Glasgow Coma Scale	E	V	M		Tot
Int:	4	5	6	=	15

Pupils	Left	Right
Size:	Normal	Normal

	Motor	Sensory
LA:	Normal	Normal
RA:	Normal	Normal
LL:	Normal	Normal
RL:	Normal	Normal

Airway
Status: Patent
Respiratory
Effort: Normal
Sounds: L: Clear R: Clear

Cardiovascular
JVD: Not Appreciated Cap. Refill: Less than 2 Seconds
Edema: Not Appreciated

Pulses	Left	Right
Carotid:	Not Checked	Not Checked
Radial:	Strong	Not Checked
Femoral:	Not Checked	Not Checked
Dorsalis:	Not Checked	Not Checked

Unit Information

Scene Information

Neurological Information

Emergency Services Data Processing Example

Metadata needed:
Time definitions unclear

Inconsistent Values:
Violates time dependency constraint

Multiple Types in one table (of tables) that need separation

Format Issue:
Values of a variable used for column headers

Format Issue:
Forms made available online as nested HTML Tables - each needing separate extraction

SOCIAL & DECISION ANALYTICS LABORATORY

Figure 9: Web Scraping: Many Problems

3.2.2.1 Uniqueness

The concept of data uniqueness can be generalized as the number of unique valid values that have been entered in a record field, or as a combination of record field values within a dataset. Uniqueness is not generally discussed in terms of data quality, but for the purposes of answering research questions, the variety and richness of the data is of paramount importance. Most notably, if a record field has very little value uniqueness (e.g. entries in the field 'State' for an analysis of housing within a county, which of course would be within a single state), then its utility would be quite low and can be conceptualized as having low quality in terms of the research question at hand.

A basic birth year distribution plot is shown using R and SQL:

```
values_birth_year = dbGetQuery(con, "SELECT birth_year
                                      FROM student_mobility_fields_2005_2015")
```

```
## Warning in postgresqlQuickSQL(conn, statement, ...):  Could not create executeSELECT
birth_year
##                                         FROM student_mobility_fields_2005_2015
```

```
# frequency distribution plot of birth_year values
birth_year_frequencies = table(values_birth_year$birth_year)
barplot(birth_year_frequencies, main="Birth Year Value Distribution", horiz=TRUE)
```

```
## Warning in min(x):  no non-missing arguments to min; returning Inf

## Warning in max(x):  no non-missing arguments to max; returning -Inf

## Warning in min(w.l):  no non-missing arguments to min; returning Inf

## Warning in max(w.r):  no non-missing arguments to max; returning -Inf

## Error in plot.window(xlim, ylim, log = log, ...):  need finite ’xlim’ values
```

3.2.2.2 Completeness

The concept of data compeness can be generalized as the proportion of data provided versus the proportion of data required. Data that is missing may additionally be categorized as record fields not containing data, records not containing necessary fields, or datasets not containing the requisite records. The most common conceptualization of completeness is the first, record field not containing data. This conceptualization of data completeness can be thought of as the proportion of the data that has values to the proportion of data that ’should’ have values. That is, a set of data is complete with respect to a given purpose if the set contains all the relevant data for that purpose. Completeness is application specific. It would be incorrect to simply measure the number of missing field values in a record without first considering which of the fields are actually necessary for completion of the task at hand.

The completeness of the data refers to whether or not there are missing or ”null” values where there should not be any. In the case of entry code, if the active status is N, entry code should be blank. Since there are 1372857 records of active status being N, there should be 1372857 blanks for entry code. However, shown below is a count of the missing values, which is 1355653.

```
# function for truncation used below
column_name = "active_status"
table_name = "student_record_2005_2015"

# get the total count of records
total_record_count <-
  dbGetQuery(con,
             paste0("SELECT count(*) FROM ", table_name))

missing <-
  dbGetQuery(con,
             paste("SELECT count(*)",
                   "FROM", table_name,
                   "WHERE", column_name, "IS NULL"
                   )
             )

print(missing, row.names = FALSE)


##  count
##      0


#give the total count of responses
print(total_record_count, row.names = FALSE)


##     count
##  42134526


#calculate the percent of the variable which is complete
percent_complete <-
  truncate(((total_record_count - missing)/total_record_count)*100, prec = 2)

print(percent_complete, row.names = FALSE)


##  count
##    100
```

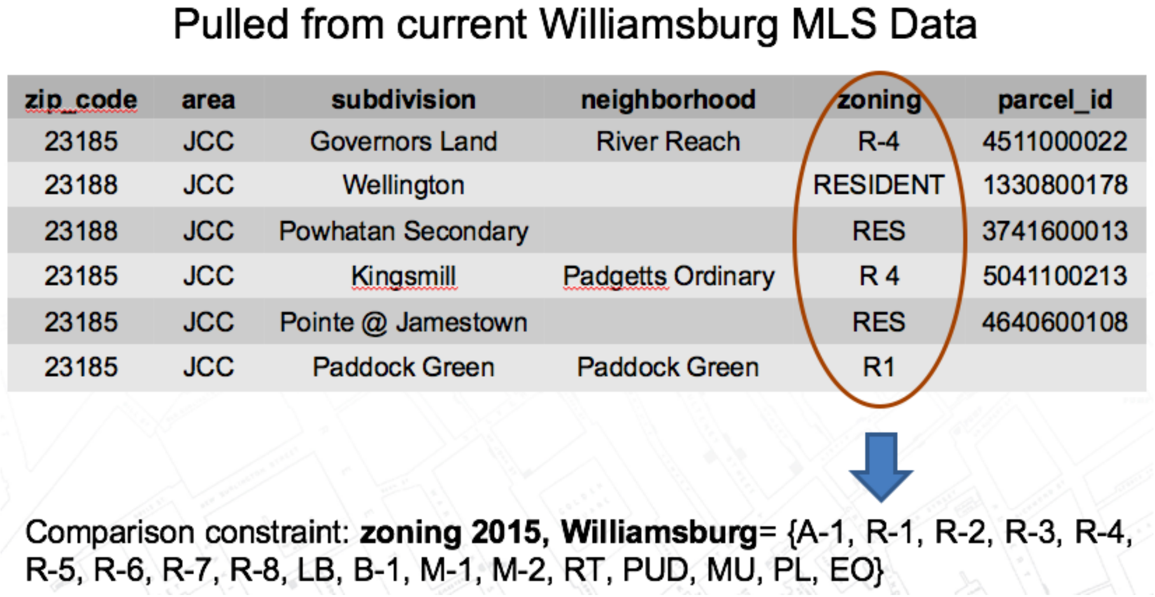
3.2.2.3 Value Validity

The concept of value validity can be conceptualized as the percentage of elements whose attributes possess expected values. The actualization of this concept generally comes in the form of straightforward domain constraint rules. For example, to ascertain how many entries contain non-valid values for a non-empty text field representing gender, an example pseudo-code domain comparison-constraint rule would look something like:

count gender where gender is not in (male, female) Or, to ascertain how many entries contain non-valid values for a non-empty integer field representing age, a pseudo-code domain interval-constraint rule would look something like:

count age where age is not between [0,110] It should be noted that in many discussions of data quality, this concept is simply referred to as "Validity" (Redman, DoD, . . .). However, the term validity has many differing and complex meanings (and attendant sub-definitions) within the social and behavioral sciences. Therefore, here we use our own sub-definition "value validity" to be clear which specific form of validity is being discussed.

An example of a discovered value validity problem can be seen in Figure . While profiling MLS data for a particular locality, it was discovered that the values entered for the field "zoning" were extensively varied. It is clear that the mechanism of input provided for this field was 'free text' where anything can be typed. The domain comparison-constraint for this field is the official list of zoning district names for that locality. It was found that none of the entries for this field in this particular dataset qualify as valid values. However, it should be noted that there may still vary well be usable information contained within the zoning field. The lack of valid values simply points to a potential problem in need of further investigation. For example, if the question at hand simply requires a count of how many properties are "Resdiential", it may very well be possible to transform the existing entries to adequately represent a true or false in this respect. However, the executions of such transformations are left for the Data Transformation stage.



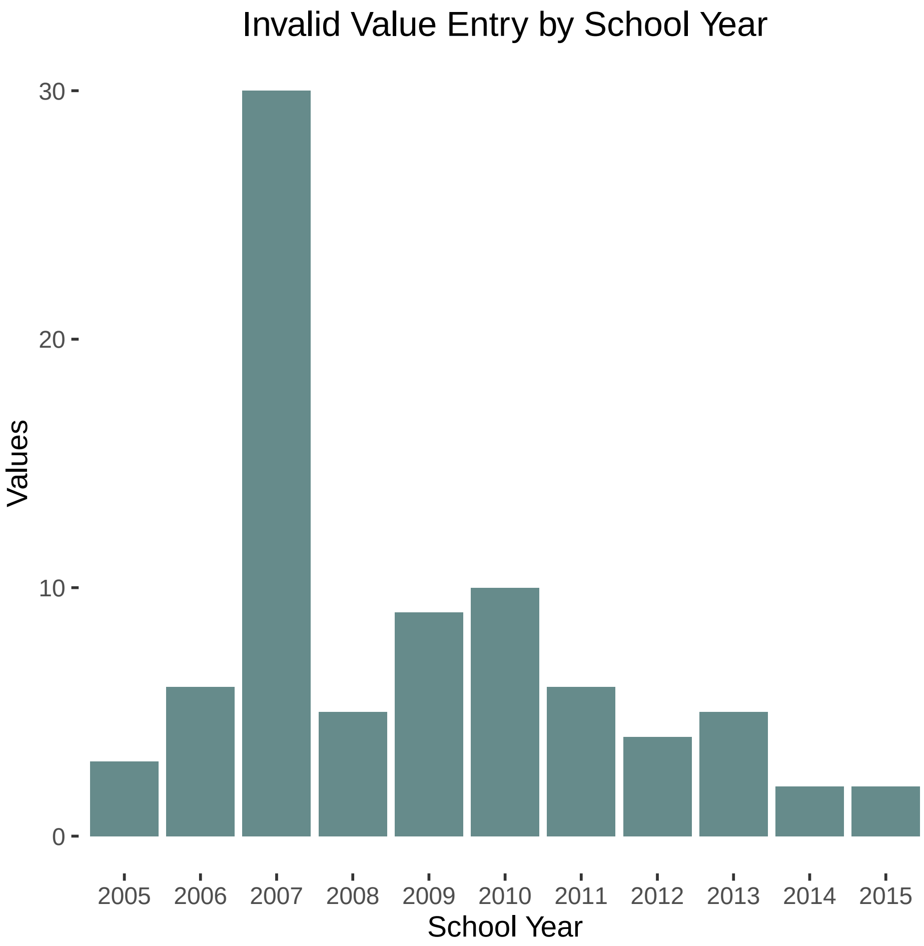
Pulled from current Williamsburg MLS Data

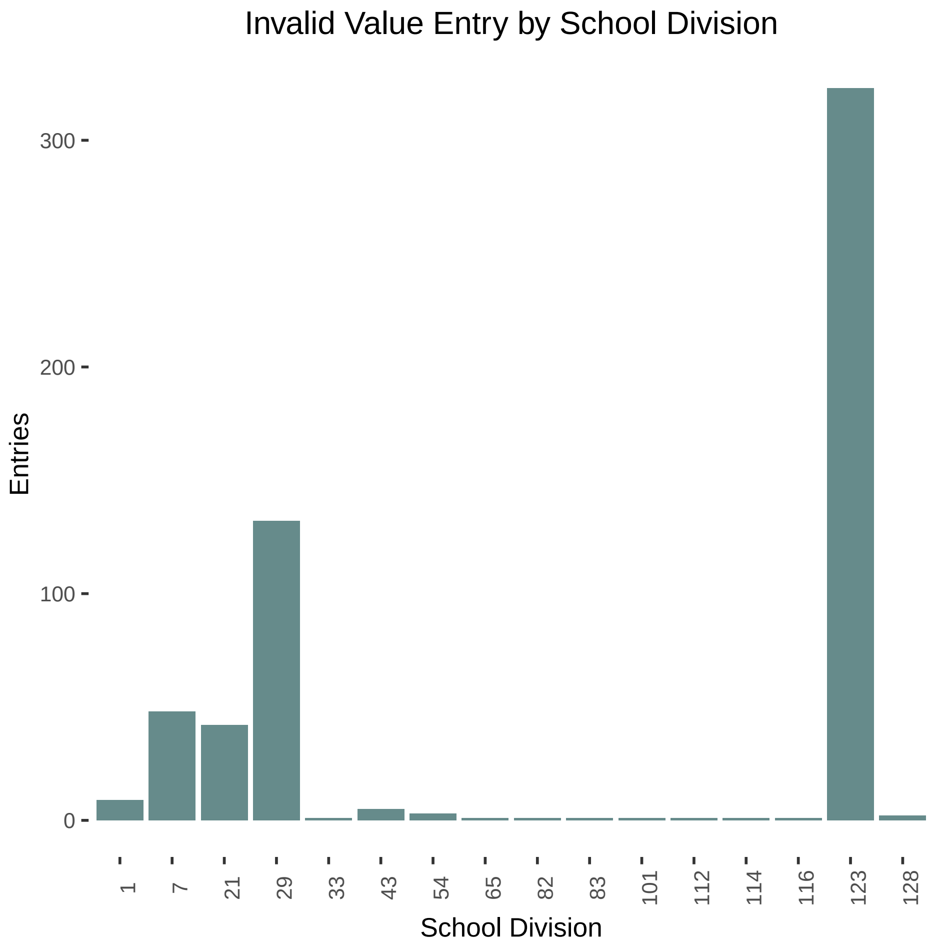
zip_code	area	subdivision	neighborhood	zoning	parcel_id
23185	JCC	Governors Land	River Reach	R-4	4511000022
23188	JCC	Wellington		RESIDENT	1330800178
23188	JCC	Powhatan Secondary		RES	3741600013
23185	JCC	Kingsmill	Padgetts Ordinary	R 4	5041100213
23185	JCC	Pointe @ Jamestown		RES	4640600108
23185	JCC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

Comparison constraint: **zoning 2015, Williamsburg**= {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}

Figure 10: Value Validity

```
column_name = "birth_year"
table_name = "student_record_2005_2015"

invalid = dbGetQuery(con, paste0("SELECT count(*)",
                                 " FROM ", table_name,
                                 " WHERE ", column_name, " IS NOT NULL",
                                 " AND CAST(", column_name, " as character varying)
                                 NOT IN (SELECT CAST(value as character varying)",
                                 " FROM metadata_valid_values",
                                 " WHERE table_name = '", table_name, "'",
                                 " AND column_name = '", column_name,
                                 "')"))

# get number of rows with invlaid values for birth_year
print(invalid, row.names = FALSE)


##  count
##    572


# give the total count of responses
print(total_record_count, row.names = FALSE)


##     count
##  42134526


# calculate the percent of the variable which is valid
percent_valid <-
  truncate(((total_record_count - invalid)/total_record_count)*100, prec = 2)

print(percent_valid, row.names = FALSE)


##  count
##  99.99


if (invalid > 0){
  # get record details
  invalid_details <-
    dbGetQuery(con,
               paste0(" SELECT ", column_name, ", school_year,
                      serving_division__number",
                      " FROM ", table_name,
                      " WHERE ", column_name, " IS NOT NULL",
                      " AND CAST(", column_name, " as character varying) NOT IN
                            (SELECT CAST(value as character varying)",
                            " FROM metadata_valid_values",
                            " WHERE table_name = '", table_name, "'",
                            " AND column_name = '", column_name, "')"))

  # read in data
  year_plot_df <-
    as.data.frame(
      table(invalid_details[[1]], invalid_details[[2]]))

  levels(year_plot_df$Var2) <-
    c("2005","2006","2007","2008","2009","2010","2011","2012","2013","2014","2015")

  # plot inconsistencies by School Year
  school_year <- (ggplot(
    year_plot_df, aes(Var2, Freq)) +
    geom_bar(stat="identity", fill ="paleturquoise4", position="dodge") +
    ggtitle("Invalid Value Entry by School Year") +
    ylab("Values") +
    scale_x_discrete("School Year", drop = "FALSE") +
    theme(panel.background=element_blank()))

  # read in data
  division_plot_df <-
    as.data.frame(
      table(invalid_details[[1]], invalid_details[[3]]))

  # plot inconsistencies by School Division
  division <- (ggplot(division_plot_df, aes(Var2, Freq)) +
    geom_bar(stat="identity", fill ="paleturquoise4") +
    ggtitle("Invalid Value Entry by School Division") +
    xlab("School Division") +
    ylab("Entries") +
    theme(panel.background=element_blank()) +
    theme(text=element_text(size=10),
          axis.text.x=element_text(angle=90, vjust=1)))

  print(school_year)
  print(division)
}
```


Invalid Value Entry by School Year
Values
30
20
10
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
School Year



3.2.2.4 Consistency

The concept of 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. When there is an expected logical relationship between two or more entities, we can refer to the rule specificying this logic as a type of relationship validation called a dependency constraint. Therefore, consistency becomes the degree to which these attributes satisfy said dependency constraint. An example of a logical requirement is that fields A and B must sum to field C. Logical requirements may be quite involved. For example, ‘A has two children’, ‘A is B’s father’, ‘A is C’s father’, and ‘A is D’s father’ are inconsistent.

A simple example of a dependency constraint violation would be a location disagreement like a zip-code that does not agree with a state code. Another might be the identification of a male who is also pregnant.

A more complex example would involve consistency validation of fields associated with student withdrawal in education records. For example, most education records contain some form of an ‘active code’ categorizing the student’s current form of interaction with the educational system. Most systems will have a code for ‘Inactive’ and/or ‘Not enrolled’. Most systems will also have a field for withdrawal date. If a student has been categorized as ‘Inactive’ or ‘Not enrolled’ but does not have an entry for withdrawal date, then there exists a consistency issue.

Causes of inconsistency are varied, but a common source of inconsistency comes from situations where locally derived information is provided with no associated master list or file. An exhaustive ‘master list’ of individuals receiving a public service are, in fact, quite rare. In our study, this occurred with student records such as from the Virginia Department of Education (VDOE). Here the student demographics occur in multiple records about the same student recorded in the same year. Truth must be derived from the multiple observations. For example, the VDOE data we found 16,310 of the 2,346,058 individuals to have more than one value for gender.

A simple consistency check on the field ’gender’ (in pseudo-code) would look like: count students from student_records joined to itself on student_id where gender does not match

3.2.2.5 Record Consistency

A record is inconsistent when active_status is ”I”, but there is no exit code and exit date. The inconsistent records can be plotted by both school district and school year to see if and where there are greater numbers of inconsistencies by location or year.

Get records having an inconsistent relationship between active_status, exit_code and exit_date.

Get the number of inconsistent records as well as the number of unique students associated with those records:

```
column_name = "active_status"
table_name = "student_record_2005_2015"

record_inconsistent_1 =
  dbGetQuery(con,
             paste0(
               "SELECT unique_id",
               " FROM ", table_name,
               " WHERE ", column_name, " = 'I'",
               " AND exit_code IS NULL AND exit_date IS NULL"))

unique_students = as.data.frame(unique(record_inconsistent_1[[1]]))

print(
  paste(
    nrow(unique_students),
    "students with",
    nrow(record_inconsistent_1),
    "inconsistent records"),
    row.names = FALSE)
```

57 unique students with 68 inconsistent records.

Get the total count of responses and calculate the percent of the columns values that are consistent.

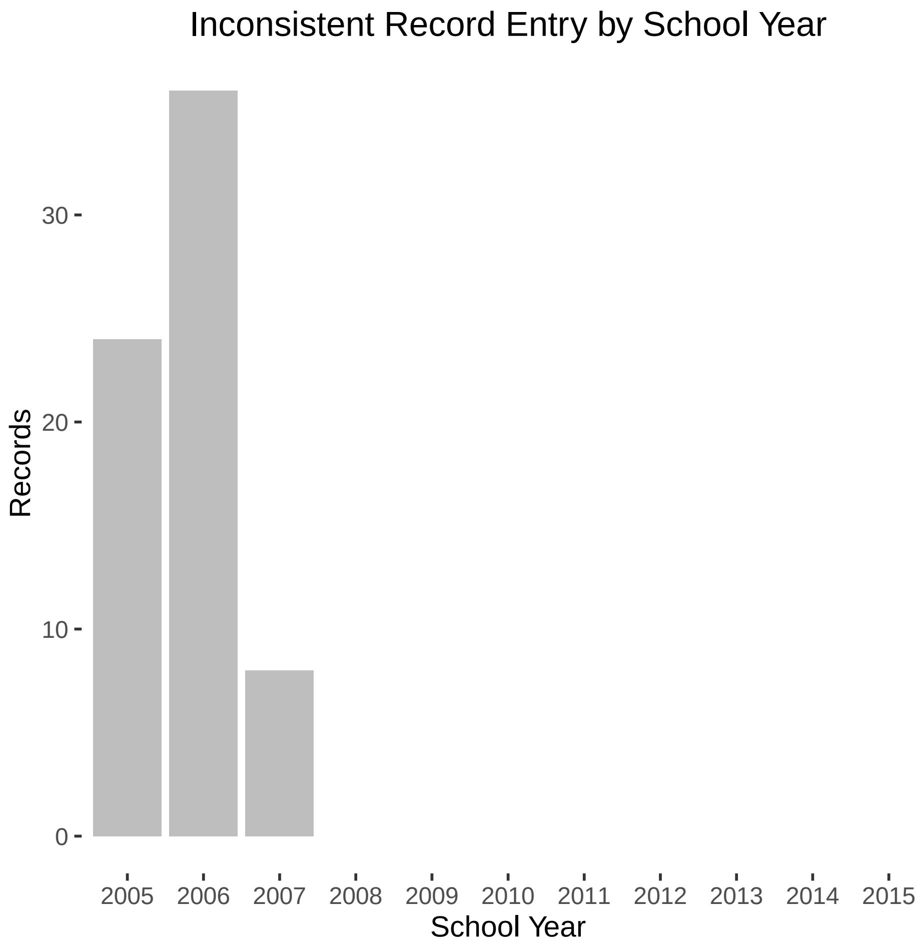
```
total_count = dbGetQuery(con, paste0("SELECT count(*)",
                                     " FROM ", table_name))

print(total_count, row.names = FALSE)
```

```r
# function for truncation used below
truncate = function(x, ..., prec = 0) base::trunc(x * 10^prec, ...) / 10^prec

# calculate the percent of the variable which is consistent

record_percent_consistent_1 =
  truncate(((total_count - nrow(record_inconsistent_1))/total_count)*100, prec = 2)

print(record_percent_consistent_1, row.names = FALSE)
```

99.99 percent consistent.

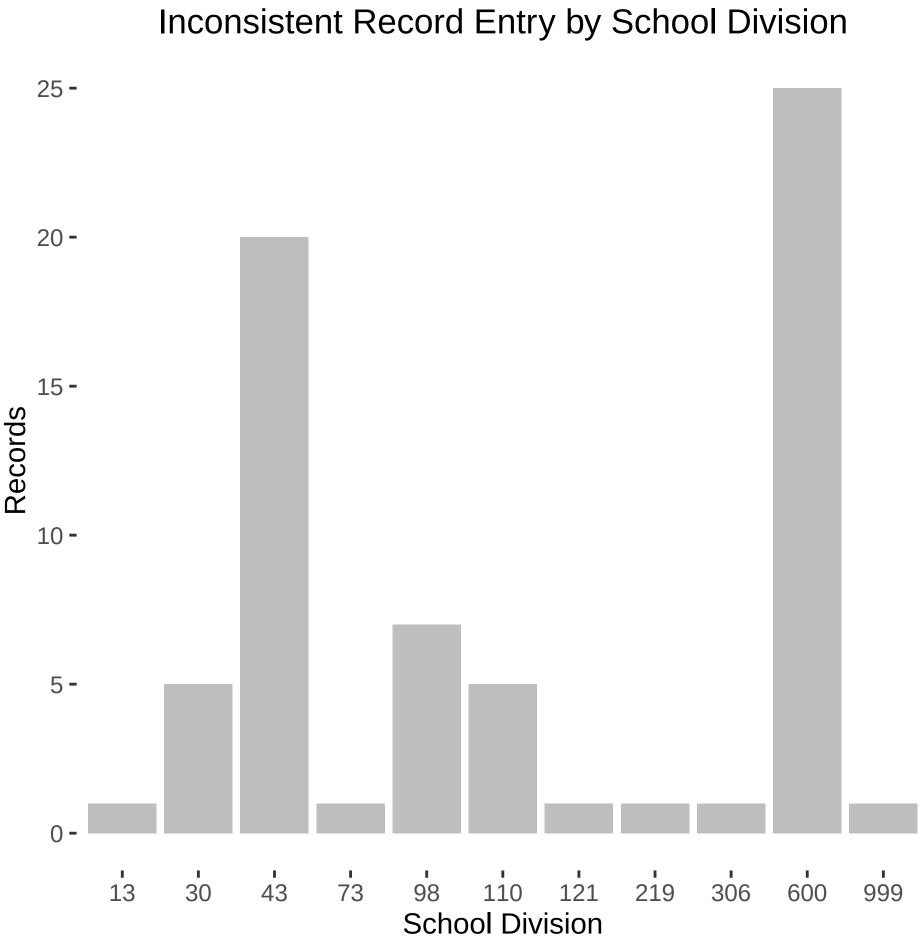
Create a plot of inconsistent records vs school year.

```r
record_inconsistent_1_details =
  dbGetQuery(con,
             paste0("SELECT distinct unique_id, ", column_name,
                    ", school_year ",
                    ", serving_division__number ",
                    " FROM ",
                    table_name,
                    " WHERE ",
                    column_name,
                    " = 'I' AND exit_code IS NULL AND exit_date IS NULL"))

year_plot_df =
  as.data.frame(
    table(record_inconsistent_1_details[[3]], record_inconsistent_1_details[[2]]))

levels(year_plot_df$Var1) =
  c("2005","2006","2007","2008","2009","2010","2011","2012","2013","2014","2015")

school_year_plot <- ggplot(
  year_plot_df, aes(Var1, Freq)) +
  geom_bar(stat="identity", fill ="grey") +
  ggtitle("Inconsistent Record Entry by School Year") +
  ylab("Records") +
  scale_x_discrete("School Year", drop = "FALSE") +
  theme(panel.background=element_blank())

print(school_year_plot)
```



Create a plot of inconsistent records vs school division.

```
division_plot_df =
  as.data.frame(
    table(record_inconsistent_1_details[[4]], record_inconsistent_1_details[[2]]))

school_division_plot <- ggplot(
  head(division_plot_df[order(-division_plot_df$Freq),], n=20), aes(Var1, Freq)) +
  geom_bar(stat="identity", fill ="grey") +
  ggtitle("Inconsistent Record Entry by School Division") +
  xlab("School Division") +
  ylab("Records") +
  theme(panel.background=element_blank())

print(school_division_plot)
```



3.2.2.6 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).

Causes of longitudinal inconsistency are varied, but a common source of inconsistency comes from situations where locally derived information is provided with no associated master list or file. An exhaustive 'master list' of individuals receiving a public service are, in fact, quite rare. In our study, this occurred with student records such as from the Virginia Department of Education (VDOE). Here the student demographics occur in multiple records about the same student recorded in the same year. Truth must be derived from the multiple observations. For example, the VDOE data we found 16,310 of the 2,346,058 individuals to have more than one value for gender.

A simple consistency check on the field 'gender' (in pseudo-code) would look like:

```
count students from student_records joined to itself on student_id where gender does not match
```

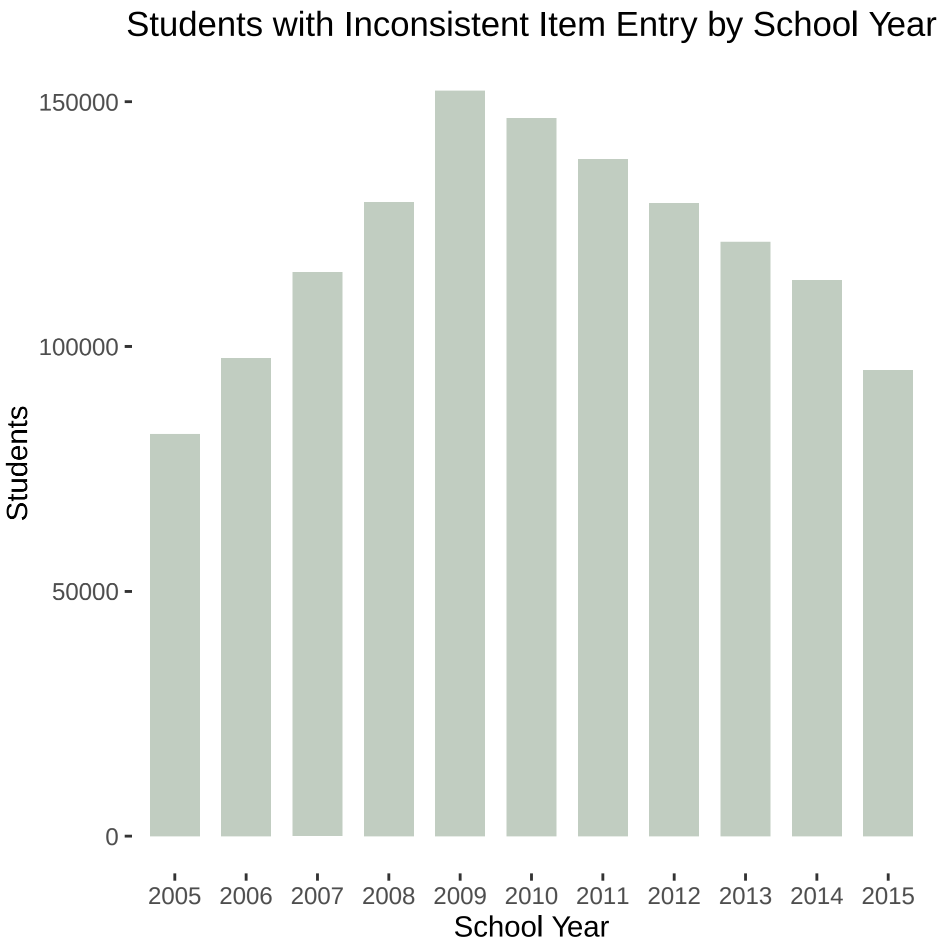
A consistently troublesome demographic variable, from a longitudinal consistency viewpoint, is race. Race categorization schemes change fairy frequently (in comparison to other demographic categories). In addition, people will periodically elect to change the racial category with which the identify.

A lonitudinal consistency check of race in the Virginia Student Record Collection:

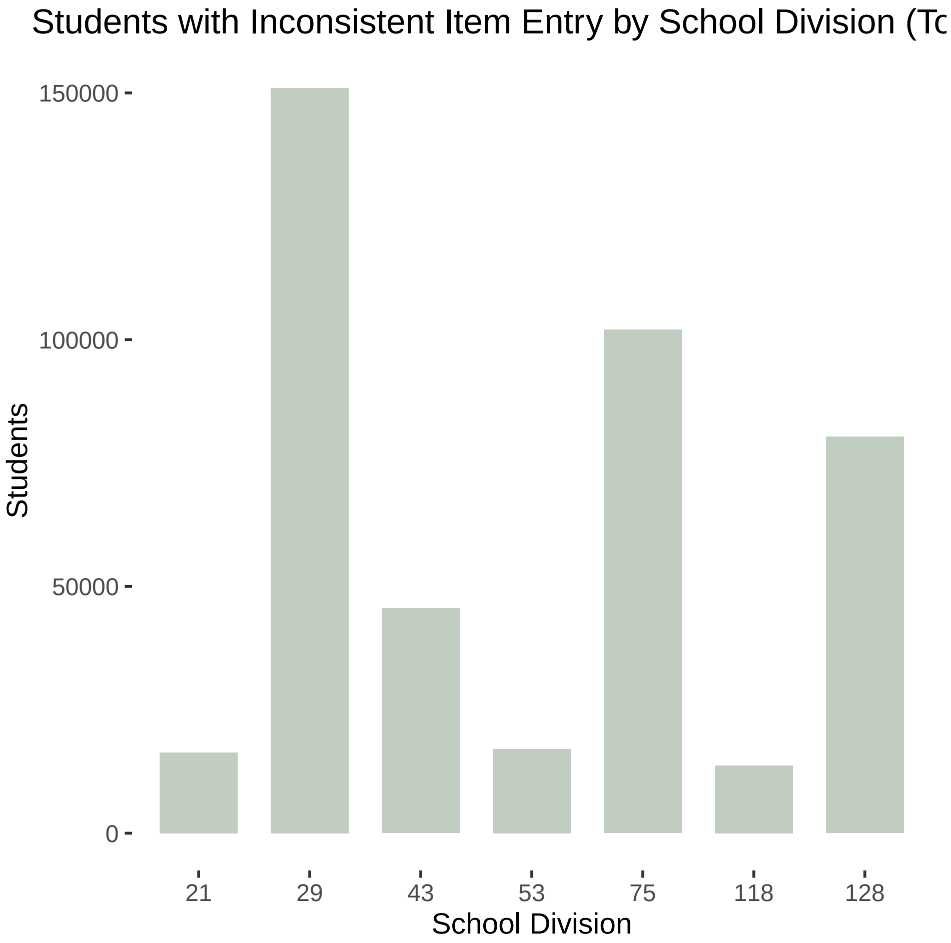
In the Virginia Student Record Collection there are 160055 such inconsistencies. This means the dataset is about .37% inconsistent (which actually means it is still a pretty good variable for many uses).

An inconsistency in the data is checked over time (longitudinal), to see if the same value is recorded for every new record. For birth_month, this would be if a student recorded as one birth month one year another birth month next year. For this data set there are 11379 inconsistencies. This means the dataset is about 2.7e-2% inconsistent.

```
#LONGITUDINAL CONSISTENCY
column_name = "race_type"
table_name = "student_record_2005_2015"
demographic_table_name = "student_record_self_join"

longitudinal_inconsistent_1 <-
  dbGetQuery(con, paste0(" SELECT distinct unique_id_a",
                         " FROM ", demographic_table_name,
                         " WHERE ", column_name, "_a <> ", column_name, "_b"))

# get the total count of individuals
total_person_count <-
  dbGetQuery(con,
             paste0("SELECT count(distinct unique_id) FROM ", table_name))

longitudinal_inconsistent_count_1 <-
  nrow(longitudinal_inconsistent_1)

longitudinal_percent_consistent_1 <-
  truncate(((total_person_count - longitudinal_inconsistent_count_1)/total_person_count)*100, prec

print(longitudinal_inconsistent_count_1)


## [1] 160055


print(longitudinal_percent_consistent_1)


##   count
## 1 93.94


# convert dataframe column to a single-quoted, comma-delimited string for use in next SQL query
unique_id_list <-
  paste0("'", paste0(longitudinal_inconsistent_1$unique_id, collapse = "','"), "'")

# get record details
longitudinal_inconsistent_1_details <-
  dbGetQuery(con,
             paste0("SELECT distinct unique_id, ",
                    column_name, ", school_year,",
                    " serving_division__number division_number",
                    " FROM ", table_name,
                    " WHERE unique_id IN (", unique_id_list,")"
                    )
             )
# read in data
year_plot_df <-
  as.data.frame(
    table(longitudinal_inconsistent_1_details[[3]], longitudinal_inconsistent_1_details[[2]]))

levels(year_plot_df$Var1) <-
  c("2005","2006","2007","2008","2009","2010","2011","2012","2013","2014","2015")

# plot inconsistencies by School Year

  school_year <- (ggplot(
  year_plot_df, aes(Var1, Freq)) +
  geom_bar(stat="identity", fill ="honeydew3", width=.7) +
  ggtitle("Students with Inconsistent Item Entry by School Year") +
  scale_x_discrete("School Year", drop = FALSE) +
  ylab("Students") +
  theme(panel.background=element_blank()))


# read in data
division_plot_df <-
  as.data.frame(
    table(longitudinal_inconsistent_1_details[[4]], longitudinal_inconsistent_1_details[[2]]))

# plot inconsistencies by School Division
  division <- (ggplot(
  head(division_plot_df[order(-division_plot_df$Freq),], n=20), aes(Var1, Freq)) +
  geom_bar(stat="identity",  fill ="honeydew3", width=.7) +
  ggtitle("Students with Inconsistent Item Entry by School Division (Top 20)") +
  xlab("School Division") +
  ylab("Students") +
  theme(panel.background=element_blank()))
  print(school_year)
```



```
print(division)
```



3.2.2.7 Duplication

Duplication refers to the degree of replication of distinct observations per observation unit type. For example, in state-level secondary-education registration records, greater than 1 registration per student per official reporting period would represent duplication.

However, it is important to remember that while duplication can occur as a result of the accidental entering of the same information multiple times, duplication can occur many times as a direct result of the 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.

3.2.3 Metadata: Are datasets, observation units, and their attributes consistently named, sufficiently described, and appropriately formatted for combination with other datasets?

Metadata is generally defined "data that provides information about other data". [http:www.merriam-webster.comdictionarymetadata]. The main purpose of metadata is the facilitation of the discovery relevant information pertaining to a particular object/resource. It does this by "allowing resources to be found by relevant criteria, identifying resources, bringing similar resources together, distinguishing dissimilar resources, and giving location information."[National Information Standards Organization; Rebecca Guenther; Jaqueline Radebaugh (2004). Understanding Metadata. Bethesda, MD: NISO Press. ISBN 1-880124-62-9. Retrieved 2 April 2014.] Therefore, generally speaking, a lack of metadata for a dataset can present significant impediments to the use of said dataset. Being more specific, when dealing with data that is to be used for research purposes, it is of vital importance to discover if the datasets (tables), their observational observation units (records/rows), and their attributes (fields/columns) are consistently named, sufficiently described, and appropriately formatted for analysis and for combination with other project datasets. Additionally, does information exist regrading any transformations that have occurred to original data sources in the creation of said dataset, as well as who did the transforming?

Observation Unit Definition When a dataset is provided without definition of the purpose of that dataset, we have an issue with the Observation Unit Definition. Why was this data collected? When dealing with datasets not orignally collected for research purposes (e.g administrative data), there is often no easy answer to this question. To correct issues of observational unit definition, it is often necessary to first generate separate new datasets from the dataset provided, each representing only a single observational unit type. At this point a new observational unit definition can be created.

This type of metadata is issue is quite common, and we experienced said issue when dealing with certain 3rd-party produced housing datasets. A single dataset would include, within each record(row), multiple potential observational units (e.g housing unit data, listing service data, owner data, neighborhood data). Only after defining the observational units needed and extracting the necessary fields could a new observational unit definition be generated (e.g. Housing Unit Specifications for Arlington County VA from CY 20XX to CY 20XX).

Observation Unit Attributes Definition Attributes (columns) without definition and/or non-meaningful/confusing naming

note. It is sometimes necessary to separate multiple variables represented in a single attribute (column) before creating definitions

Semantic Confusion The concept of semantic interoperability here refers to the ability of data systems to exchange data with other data systems unambiguously. Semantic interoperability is concerned not just with the data syntax, but also with the transmission of the meaning with the data (semantics). This is generally accomplished by adding metadata to a dataset, thereby defining a controlled, shared vocabulary. Without this shared vocabulary, **Semantic Confusion** can occur, where names and syntax may agree, but definitions don't. For example, while combining two data sets, it may be found that two fields(attributes) have the same name (say "Grade") but their definitions are completely different (because "Grade" can refer to both a 'score' for a test/class or the 'level/year' of schooling).

Multiple Attribute Names Attributes with different names but the same definition

e.g. attributes name "Grade" and "Year" both referring to 'level/year' of schooling

Inconsistent Attribute Formats Attributes of the same type that are formatted differently

e.g. most commonly an issue when dealing with dates and times

3.2.4 Provenance

The concept of Provenance is very broad and has different meanings within different fields of inquiry. For the purposes of Data Profiling, we find it useful to apply the definition provided by the World Wide Web Consortium (W3C)

"Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance." https://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance

An example of why it is important to consider data provenance was seen in the housing case study. Some of the datasets used were provided by 3rd party vendors. As part of the value-added of these data products, 3rd party vendors often perform some set of transformations on the original data to enhance data consistency and quality. Sometimes the transformation processes used are readily available to the client, and the client can validate their application by repeating the transformations and producing identical results. Other this information is not made available, thus necessitating further investigation and experimentation on the part of the client to ensure that the data provided is, in fact, a true representation of the original source data.

Provenance MUCH more important with admin data, especially private data with proprietary algorithms (USE PROPERTY CRIME EXAMPLE).

The provenance of designed data collections (i.e. surveys) is generally, almost by definition, available. The purpose and methods of collection are usually documented as part of a study design, as well as any transformations that had to be appied during the creation of the analysis dataset from the collection instrument.

For administrative data, provenence is generally NOT readily available. Personal contact with data management personnel is critically necessary to understand the methods of collection and any transformations that occur. Also, personal contact with policy-level personnel is equally needed to discern the original purpose of collection. Also, other researchers to help glean quality. More and more metadata being generated.

Property Crime Example Another example from our studies was from a commercial data provider which provided indicators of neighborhood quality based on patented algorithms. We were unable to reconcile differences found in their crime indexes and data from Arlington County, Virginia Police Incident Tracking system. Figure 3.5 presents the misalignment in these data sources by census tract. The figure shows property crime counts as calculated by the commercial provider and as directly pulled from the Arlington County Police Incident Tracking system. The county data had five census tracts with counts greater than 300. These were not shown in the boxplot to allow a better scaled comparison to the commercial provider. The commericial provider did not describe, nor makes available, their methods to adjust the counts.

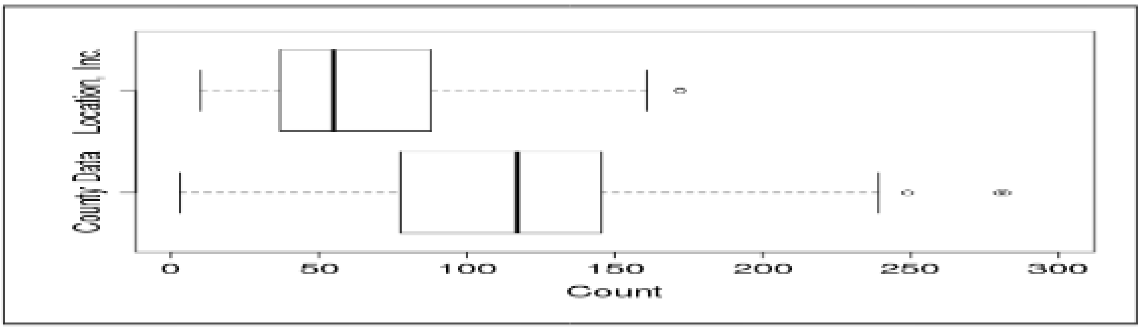
3.3 Data Preparation

3.3.1 Cleaning & Transformation

3.3.2 Restructuring

To address issues of structure discovered during Data Profiling, it often necessary to restructure the data source (dataset) into multiple new datasets that are more easily analyzed. This activity can be thought of as being akin to the process of database normalization, the process of organizing the columns (attributes) and tables (relations) of a relational database to minimize data redundancy.

An example of restructuring in the education case study was the subsetting of a database into three tables to account for student's race, gender, and disadvantaged status. Each of the three tables were aggregated by "School Year," "Division Number," and "Grade Code" according to a set



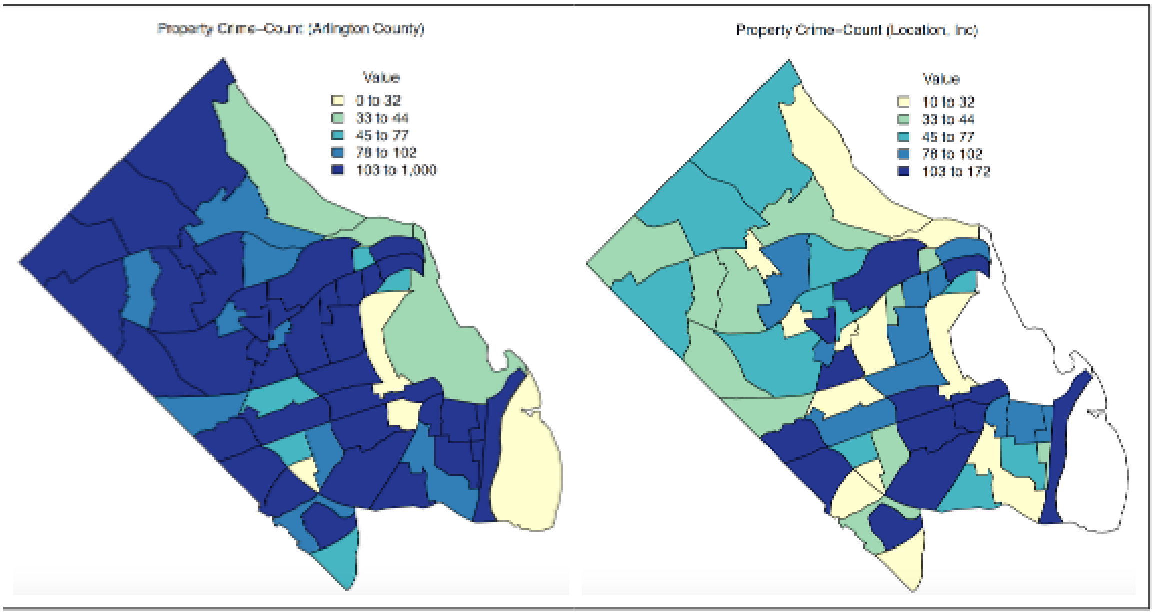


Figure 11: Geographic distribution of 2013 property crime counts by census tract using crime data provided by Arlington County, VA and by a commercial data provider.

of rules determining inclusion/exclusion of each variable in the Fall Membership table.

Another example of restructuring the data occurred when dealing with third-party MLS data. It was necessary to divide the dataset into multiple separate datasets,"Property ID & Location," "Property Characteristics," "Property Sales Information," and "Property Tax Information." Each of these new datasets represented a distinct unit of analysis. All of the new datasets were then associated via a new identifier, in this case, "Parcel ID" (see Figure 3.7). However, "Parcel ID" was left blank in over 7% of the entries. Therefore, extra work was required employing the use of a geocoding API (application program interface) to locate a property within county parcel maps that already included a "Parcel ID." Further, an additional complication is the fact that no standardized address format was used in the creation of the MLS record. Therefore, direct interaction and decision making by an analyst was also necessary to finalize the geographic matching.

3.4 Data Linkage

Once of the newer challenges fueled by the all data revolution has to do with combining data across data sources, particularly in light of multiple owners of the sources. The statistics literature has

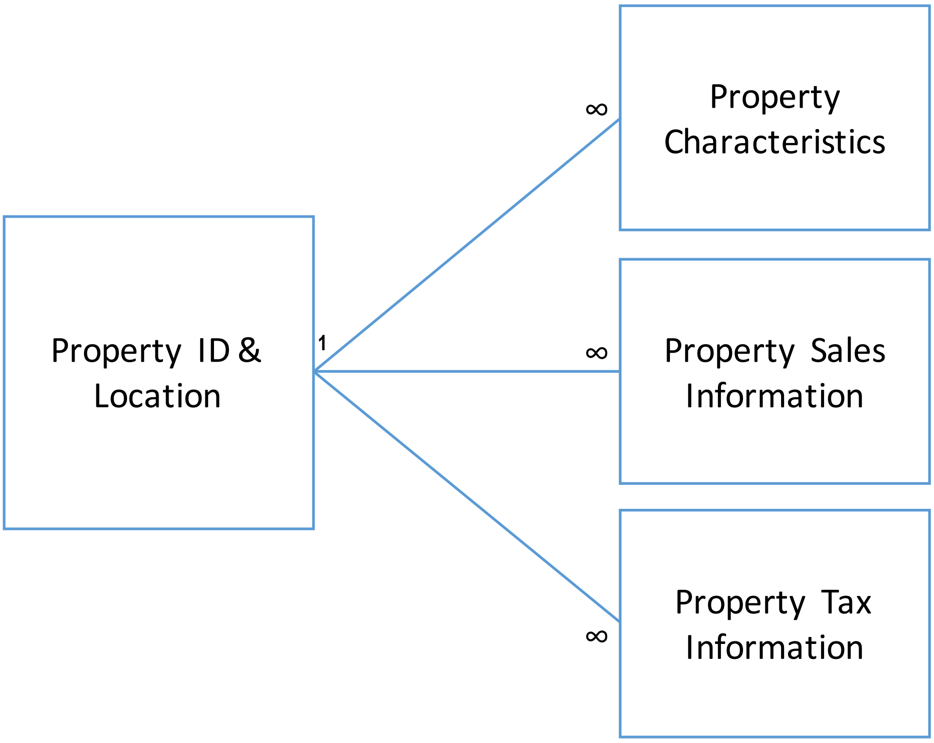


Figure 12: Restructured

a long history of methodology development for linking data (Fellegi & Sunter, 1969; NRC, 1999). This research was initially developed to aid data combination across multiple datasets for official statistics reporting. Simply stated, the goal of a record linkage is to determine whether a record from one data set has a corresponding record in another data set. As identification disclosure risk became magnified due to technology enhancements and network connectivity (Sweeney, 2002), the statistical disclosure limitation framework began to expand to record linkage methodology (Skinner, 2008).

Today record linkage technology must take on a new dimension, namely data governance and the desire to combine data across organizations operating under different, sometimes contradictory, polices and regulations. The integration of record-level data from the data systems of multiple agencies (owners) has the potential for generating high-quality evidence to be used in research and in the assessment of public policy outcomes. However, when attempting to combine data records from these systems, a number of complex legal issues must be considered, not the least of which is the privacy of persons represented by the data in the systems. Because there exist many overlapping and often inconsistent privacy-related restrictions at multiple levels of government, the linkage of data, particularly administrative data records, across public agencies can prove exceedingly difficult.

In recent years, one of the most pronounced areas of development in the technological control of data access has been in the area of privacy protecting record linkage (also referred to variably as data matching, data sharing, or entity resolution). This development is most especially pronounced in the literatures of computer science, medical informatics, as well as statistics, with each field producing multiple literature syntheses and taxonomies of the research area (Churches & Christen, 2004; Hall & Fienberg, 2010; Kum, Krishnamurthy, Machanavajjhala, Reiter, & Ahalt, 2014; Vatsalan, Christen, & Verykios, 2013; Verykios, Karakasidis, & Mitrogiannis, 2009). When speaking specifically to methods of record linkage that support or enhance personal privacy, the current acronym gaining widespread usage is PPRL (denoting Privacy-Protecting Record Linkage).

The requirement of 'privacy' adds a third major challenge to the two traditional major challenges of record linkage: quality and scalability. Data quality varies. Real-world data contain errors or 'dirty data'. Therefore, approximate (e.g., probabilistic) matching and classification techniques are required to achieve accurate linkage (Christen, 2006; Hernández & Stolfo, 1998). The size of a potential match between record sets is the product of the two data sets. Therefore, the computational complexity of a single match between record sets grows quadratically. With large data sets (not even big data), such growth quickly becomes a performance bottleneck, especially when each comparison utilizes approximate matching algorithms made necessary by the existence of dirty data. Accordingly, significant effort has been put forth in the development of techniques to make record matching more scalable (Baxter, Christen, & Churches, 2003; Christen, 2012; Christen & Goiser, 2007; Herzog, Scheuren, & Winkler, 2007). Now, in addition to these two challenges, we are adding the need to consider privacy protection (or more precisely, data access restrictions) at every step of the linkage process. It is in this third area of privacy protection that much development has recently occurred.

Recent and extensive overviews of the techniques are being developed specifically to address the privacy protection requirements of the linkage process (Vatsalan, Christen, O'Keefe, & Verykios, 2014). In their taxonomy, they produce an analysis that begins with the aspects of the privacy situation and details how each techniques attempts to address that specific situation. The portion of their taxonomy specific to privacy includes consideration of three dimensions: how many parties are involved, the adversary model assumed, and the actual techniques used in the PPRL approach.

Proposed solutions in the record linkage privacy context can be categorized as belonging to either a three-party protocol or a two-party protocol. In three-party protocols, a third party (in addition to the two with the data) is employed to conduct the linkage. This third party is evaluated at a certain level of 'trust'. Two-party protocols involve only the data sharers in an attempt to achieve higher-level security by reducing the possibility of any third-party collusion. Such approaches necessarily involve more complex computational approaches than the typical three-party approach. In addition, a proposed solution's privacy context can be characterized as containing two types of adversary. 'Honest-but-curious' adversaries will attempt to find out as much as they can without breaking protocol, while 'malicious' adversaries are willing to break protocol and attempt different attacks to access as much data as possible. An analogous framework for high-dimensional data analysis is based on cautious or generous sharing of data (Fienberg & Jin, 2012).

Lastly, a given technique can be classified as either exact or approximate. Given that in most real-world scenarios the data being matched is often dirty or noisy, if a single trusted shared-identifier (e.g., verified Social Security Number) is not available, or is not permitted under the data governance constraints, then the scenario will typically involve some form of probabilistic linkage. Accordingly, most contemporary techniques seek to address privacy in probabilistic record linkage, reaching back to the earliest approaches of (Fellegi & Sunter, 1969), but for different reasons.

The most difficult problems involve privacy-preserving probabilistic record-linkage in an environment of potentially malicious adversaries. Protecting data in a potentially malicious environment invariably involves the application of some form(s) of one-way hashing (where a new unique value is generated, but information is lost, so recovering the original value is impossible) and/or encryption (where no information is lost and the original value may be recovered). The issue that arises for most approaches to privacy protecting probabilistic linkage is that the most common methods of demographic comparison (e.g., string similarity functions) are quickly rendered inadequate by the usage of hashing/encryption techniques. For example, a slight misspelling in "first name" will result in completely different hashes and, therefore, a 100% non-match. A number of approaches to rectify this issue have been proposed. However, most proposed methods necessarily involve some level of reduction in matching performance as compared to string-similarity functions used on unobscured data. Since this literature may be less familiar to the statistical science community and other users of statistics, some additional detail is given below.

A protocol employing a common table of reference strings to which the actual strings in two data sets can be compared is proposed so that edit-distances from the actual strings to the reference strings can be computed. (Pang & Hansen, 2006). The reference strings closest to the actual strings can be encrypted and sent along with the edit distances to be used for match determination (Pang & Hansen, 2006). Unfortunately, testing of the process results in a fairly sizable reduction in both recall and precision in comparison to a reference string similarity measure used on unprotected strings (Bachteler, Schnell, & Reiher, 2010).

A protocol that employs mathematical stenography is proposed. The method is to embed a given string using the SparseMap method (Hjaltason, 2003) into a Euclidean space already populated with random strings (Scannapieco, Figotin, Bertino, & Elmagarmid, 2007). The coordinates for a given string are then given as the approximate distances between the string and the random strings. A third party compares the Euclidean distances between the string to determine a match. Testing shows a markedly better result than the table of reference strings approach, but still a not insignificant difference from a reference string similarity measure used on unprotected strings (Bachteler et al., 2010).

A novel method that takes advantage of the properties of Bloom filters is proposed by (Schnell, Bachteler, & Reiher, 2009). A Bloom filter is a space-efficient probabilistic data structure represented by a bit array that is used to test whether an element is a member of a set. Strings are stored as bits that represent the key-hashed message authentication code of the string's constituent n-grams. Bloom filters with similar strings will have a high proportion of the same bits set to one. Using this knowledge, a string similarity measure, the Dice coefficient, can be used to calculate the ratio of similarity. This method works quite well, in fact testing shows that the precision-recall curves using 2-grams (a contiguous sequence of n items, in this case, 2 items) is nearly identical to the benchmark metric (Bachteler et al., 2010). However, the Bloom filter approach is only applicable to using similarity measures that determine the overall similarity of one set of characters to another. That is, when order does not matter. Because of the nature of the encoding scheme and Bloom filters themselves, there is not a way to use this method with standard distance-metrics (Winkler, 2006). Demonstrating an approach that has been successfully deployed in both research and live agency environments (Schroeder, 2012), the approach described here uses a newly invented form of "ordered hashing" where both n-gram based similarity functions and distance-based similarity functions can be employed without loss of recall or precision. That is, the algorithm allows for the comparison of strings using standard string-similarity functions when the strings must be obfuscated in some secure manner (e.g. hash, encrypt). The approach employs two separate processes. The first, using a one-time-pad to create cipher for each string, provides for a type of obfuscation that is both theoretically unbreakable and not vulnerable to frequency analysis (Denning & Elizabeth, 1982). The second process enhances the first by employing the method of chaffing and winnowing which involves the addition to the cipher of fake characters ("chaff") to the valid characters ("wheat") so as to result in all encoded strings being the same length (Rivest, 1998). The proposed methodology is currently being deployed in two multi-agency data-integration projects in the State of Virginia (Schneider, Massa, & Vivari; Spears, Bradburn, Schroeder, Tester, & Forry, 2012).

This research on privacy-preserving data systems, able to integrate data across multiple sources governed by different polices open more opportunities for gaining interoperability between data assets that could be used productively for research. These systems also offer the opportunity for monitoring use through audits or other mechanisms. The potential benefits of moving to a new system that controls use of data and not the collection of the data is expected to yield large societal benefits (Kalapesi, 2013; PCAST, 2014). Challenges to implementing this approach are creating the political will and public understanding of these approaches (Acquisti et al., 2015; Kagal & Abelson, 2010), the creation of significant criminal penalties for privacy violations, and figuring out practical ways for individuals to express their preferences about personal data. One suggestion is to allow individuals to delegate their choices to organizations they trust (Mundie, 2014). Education and technology are also important for changing perceptions and increasing understanding about the proposed new approach, especially for older generations that may perceive a trust-based based approach as threatening (PCAST, 2014).

3.5 Data Exploration

Data Exploration refers to the analysis of the datasets by summarizing main characteristics, often with visual methods. Data exploration is used throughout the data framework. Descriptive statistics play a principal role in data profiling, from identifying valid attribute values to checking for semantic consistency. The use of visual techniques like boxplots support iterations between data cleaning and transformation during the data preparation.

3.6 Data Analysis

3.7 Data Fitness for Use

The purpose of developing the data framework in the context of specific problems is to find synergies across application domains with respect to the data framework development and use. The ultimate goal is to develop a disciplined process of identifying data sources, preparing them for use, and then assessing the value of these sources for the intended use(s).

Understanding how to approach fitness for use starts with considering the modeling and analyses that will use the data. Modeling depends on the research questions and the intended use of the data to support the research hypotheses. Fitness assessment should be about the fitness of the data for the modeling, from straight forward tabulations to complex analyses. Therefore, fitness is a function of the modeling, data quality needs of the models, and data coverage (representativeness) needs of the models. Finally, fitness should characterize the information content in the results.

An Example:

4 Discussion

References

[1] Earl Babie. *The Practice of Social Research.* Wadsworth Publishing Company, 1998.

[2] Ronald S. Fecso and Robert Goldenkoff. *FEDERAL STATISTICAL SYSTEM Agencies Can Make Greater Use of Existing Data, but Continued Progress Is Needed on Access and Quality Issues.* Tech. rep. GAO-12-54. Government Accountability Office, 2012.

[3] Laura Feeney et al. *USING ADMINISTRATIVE DATA FOR RANDOMIZED EVALUATIONS.* Tech. rep. J-PAL North America, 2015.

[4] Helen Lee, H Warren, and Lakhpreet Gill. "Cheaper, Faster, Better: Are State Administrative Data the Answer?" In: *The Mother and Infant Home Visiting Program Evaluation-Strong Start Second Annual Report. OPRE Report 2015-09.* Office of Planning, Research et al., 2015.

[5] Jens Ludwig et al. *Long-term neighborhood effects on low-income families: Evidence from Moving to Opportunity.* Tech. rep. National Bureau of Economic Research, 2013.

[6] Bruce D Meyer and Nikolas Mittag. *Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness and holes in the safety net.* Tech. rep. National Bureau of Economic Research, 2015.

[7] National Academies Press. *Frontiers in massive data analysis.* Tech. rep. National Research Council, 2013.