# Data in the Wild: A KM Approach to doing a Census Without Asking Anyone and the Issue of Privacy

James L. Kelly
San Diego State University
jameslkellycpa@gmail.com

Murray E. Jennex
San Diego State University
mjennex@sdsu.edu

Kaveh Abhari
San Diego State University
kabhari@sdsu.edu

Alexandra Durcikova
University of Oklahoma
alex@ou.edu

Eric Frost
San Diego State University
efrost@sdsu.edu

## Abstract

*Knowledge Societies strive to better their citizens by maximizing services while minimizing costs. One of the more expensive activities is conducting a census. This paper explores the feasibility of conducting a smart census by using a knowledge management strategy of focusing on actionable intelligence and the use of open source data sources to conduct a national census. Both technical and data privacy feasibility is discussed.*

## 1. Introduction

The federal government collects data on the population of the United States to fulfill its constitutional requirement to count and analyze the population every ten years. As the amount of data collected grows it fuels advances in technology with respect to counting and processing the census data. Unfortunately, new technology for identifying, collecting, and then sharing the census data has yet to be adopted [29], forcing some data to be collected numerous times at varying levels of accuracy. As a solution this study proposes using existing web based sources instead of paper surveys. To show that this is a feasible approach, we analyze current government data collection and analysis efforts, suggest alternate data sources and propose a strategy for an open source system of population statistics. Our study indicates that a strategy that is based on open source data could generate better focused actionable intelligence as well as improve the cost, accuracy, efficiency, timeliness, and collaborative efforts of the census. Specifically, we show how an actionable intelligence strategy is created using current and proposed data sources that can help answer complex questions such what is poverty and a new way to analyze income by using take home pay instead of gross dollar amounts. Lastly, this study proposes a set of data standards for supporting development of an open source census system that also addresses privacy.

## 2. Research Motivation

Why do we need a new approach to census data collection? The census is expensive. The reported cost of the 2010 census was approximately $13 Billion [37]. The 2020 Census if administered the same way will cost approximately $17.5 Billion [11]. Along with the census costs, the American Community Survey (a newer development discussed later) costs as much as $204 million per year to administer [10]. The proposed strategy would drastically cut costs by using existing open source data sets as well as existing government raw data to reduce or eliminate collecting similar or identical information through Census surveys or the American Community Survey. The US Census Bureau would also begin to utilize state information databases to cut down on its data collection efforts. California and Hawaii have robust websites with information on their specific states already in use, making some of the data the census and American Community Survey currently collects redundant for those states.

Accuracy is crucial for decision making. Unfortunately, the census may be inaccurate as most of the questions on the American Community Survey have the potential to produce inaccurate information because of the design of the questions and the way answers are provided. Most questions about annual income and monthly and annual expenses are answered by providing write in totals. This has the potential of producing data that is not accurate. People that fill out the American Community Survey have no incentive to be exact when answering the questions. Not only is the survey voluntary but it is not checked against any other systems to determine the data's accuracy. The survey is long, consisting of a minimum of 11 pages and 48

URI: https://hdl.handle.net/10125/59997
ISBN: 978-0-9981331-2-6

questions per member of a household. Each question generally has multiple parts to answer. Logically if options exist for data collection, self-reporting and long surveys should not be the first choice. Open source or other data sources may be more accurate. Utility companies have a vested interest in keeping accurate information for billing purposes and banks have a vested interest in keeping track of how much money their clients have on account.

The census is inefficient as the same data is collected more than once. The American Community Survey attempts to collect information that has already been obtained by other means and other departments of the federal government. The best example of this is the IRS not sharing data with the Census Bureau, thus forcing the Census Bureau to collect less than accurate information that the IRS already has in its possession [5]. The strategy for data collection calls for collecting data once by allowing government agencies and other entities to share data in a centralized location.

The Census in not timely as it took roughly 9 months for the census bureau to release its first data files for the 2010 census and the census only occurs every ten years. The American Community Survey is timelier than the decennial reports, since it is for one year of data, and is released in total in roughly 9 months. Despite how quickly the results can be tabulated the issue is that the time periods between the surveys are too long. There are important developments that are missed during a 1 or 10-year survey period that could help in creating actionable intelligence. A better strategy is to collect as often as feasible. A perpetually updated, dashboard style reporting system would create timely data for development of actionable intelligence.

There is little collaboration between data owners and census takers. Collaboration with groups that specialize in specific types of data collections or with different government agencies that maintain data sets on specific types of data, like the IRS with detailed data on gross income, income sources, tax credits and tax liability, has the potential to create powerful analysis about population statistics that can be turned into actionable intelligence. Currently raw IRS income data is restricted and not easily obtained [22]. This strategy calls for multiple government agencies, local governments, private and public companies to share information in an open source format to share with each other and the public to perform analysis and create actionable data. The Digital Accountability and Transparency Act of 2014 [8] requires the issuance of guidance to federal agencies on such data standards [8]. The federal government should engage with public and private institutions on not just establishment of standards, but also in collecting data and developing more collaborative strategies for collecting data that could help improve the quality and timeliness of data.

Should census be fully automated? There are cases where data was easily obtained from existing datasets for a secondary purpose. Edwin Black in “IBM and the Holocaust” explored the tremendous impact on the ability of the Nazi regime to identify and exterminate persons of Jewish origin or heritage through the application of IBM’s Hollerith Card technology originally developed and applied to the United States Census [4]. Additionally, it has been shown how the Obama campaign used Facebook data to influence voters in the 2012 presidential election. Finally, recent revelations in testimony to the United States Congress by Mark Zuckerberg, Chief Executive Officer of Facebook, showed how Facebook data was obtained by the Trump campaign for potential use in targeted marketing of voters [18]. These examples raise privacy issues as it may be possible to use privately owned data without approval to influence elections, and in the case of Nazi Germany, to actually try to exterminate a targeted group.

This paper proposes a knowledge management, KM, strategy approach to create actionable intelligence for guiding a census using open source data. Additionally, the paper will begin the discussion on the impact on privacy that such an approach may create. Readers are reminded that the technical demands of doing a census have historically driven the development of data analysis technology and that the proposed process is likely to be developed for real world applications. This makes the discussion of privacy issues important as it is better to develop technology with eyes wide open as to the technological impact rather than to develop the technology and then be surprised by the privacy consequences.

## 3. The United States Census

The US census has one constitutional purpose, which is to count the population of the United States every ten years, according to Article 1, Section 2 of the Constitution [35]. This number is used for congressional apportionment to each state based on population. The census Bureau has also taken on the task, though not required by the Constitution, of collecting data other than a straight population count which is apparent by the questions that are asked in historical census and the new American Community Survey. Before the invention of the internet, personal computers, and other more recent advances in computing power and technology, the census survey was one of the only ways of collecting any sort of data on the population of the United States. In today’s hightech world there have been numerous technological advances that would allow more robust data collection at a faster rate with more accuracy, than the methods that have been and are currently being used by the Census bureau. The population of the United States has grown from 3,929,326 in 1790 [30] when the first census was completed to 308,745,538 as of 2010 census, the most recent completed census [31]. The United States of America is simply too large of a population to not be using more advanced data analytics, systems design and KM techniques. Not only does the census bureau have a difficult mission in collecting accurate data on the entire population in its current form, the data it is able to collect is less useful in making decisions, creating policy and helping to manage the huge amounts of resources under the control of the federal government which are estimated at an annual budget of 4.4 trillion and assets of 3.5 Trillion [19].

Additionally, federal data requirements have changed with the addition of an annual survey in 2005. This survey, the American Community Survey, uses statistical modeling to estimate the data that was normally collected on the long form census and goes to an estimated 1 in 6 families each year without repetition for those families in a 5-year period [32]. This was a step in the right direction with respect to timeliness, the problem remains that the American Community Survey still asks questions with little or no analytical value in helping to produce valuable actionable intelligence. The questions analytical value has changed very little over the past 100 years. Moreover, the method of collection, a paper survey, has remained mostly unchanged throughout the history of the census. This survey is mailed to all households in the United States. Follow up is done on people that do not answer the survey and census workers go door to door if necessary to get the survey completed [34]. This is a very inefficient and expensive way to collect data with costs as stated in section 2.

The 2000 census saw a technological advance with the census bureau adding ocular character recognition to help cut down on data entry errors and personnel [33]. The Census Bureau has used many new technologies to help tabulate the results in past census counts. Most of the technological advances are used to count more efficiently, not to get data from a different source which is what this study is proposing [33]. Reducing the time, it takes to count the results as well as the staff necessary to count is a worthwhile endeavor, though it would not be necessary if the census bureau took steps to collect the data in a whole different manner rather than finding solutions to be able to use the current survey method for a longer period of time.

Private companies have embraced advanced data analytics and big data solutions while the government has not yet done so. The Federal government has more opportunity to effect change in this area of technological advances than any other company or group in the United States. Many industries like health care, marketing and finance have taken data analytics to a whole new level in just the last 5 years. Big data is trending, schools are creating programs based on analytics and thousands of books have been written on the subject, big data in its current form started in 2007 [20]. The federal government can collect more “good quality” data than any other organization, and now with technological advances can do it at a lower cost than what is being spent on data with questionable value by following the lead of organizations in for profit and nonprofit industries.

## 4. Actionable Intelligence

This paper uses a KM strategy approach to determine data sources useful for generating the census. Jennex [15] defined the knowledge content process of KM strategy as the identification of actionable intelligence needed to make a specific decision and then determining what knowledge, information, and data is needed to create that actionable intelligence. For this paper, actionable intelligence is the exact knowledge, information, and data needed to support a specific census question/decision and includes the specific knowledge, information, and data needed to create the actionable intelligence [15] [16]. In general, actionable intelligence is similar to wisdom in the traditional hierarchy of information first presented by Ackoff [1] as shown in the final revised knowledge pyramid, see Figure 1 [16]. This model establishes a top down strategy approach based on the decisions to be made and identifying the technologies and decision support components needed. Creating actionable intelligence starts at the top and emphasizes the question that is being asked before deciding which data to collect and from where [16]. This not only focuses the data collection on the problem to be solved but it also eliminates unnecessary data collection, saving time, money and bolstering data privacy strategies. To apply this to the census we first determined what each census question was trying to answer and then the actionable intelligence needed to generate the answer. Analysis then continued to determine what knowledge, information, and data was needed to create this actionable intelligence.

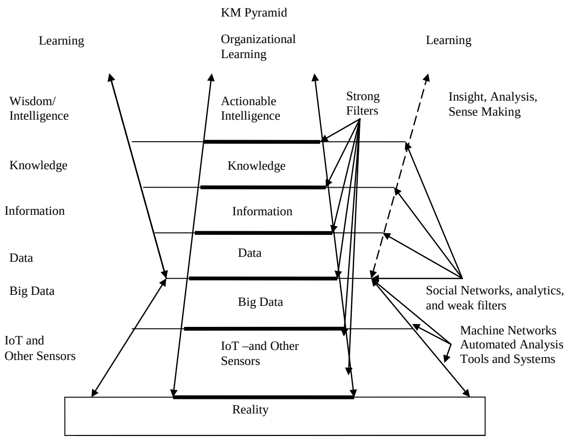

**Figure 1**, The Revised Knowledge Pyramid [16]

KM strategy is being used because what is provided by the American Community Survey, the short form census and the Statistics of Income, SOI, annual reports is best labeled as information or data. This is useful for trend analysis and asking question like what is the average income tax expense on a tax return for reporting income between $40,000.00 and $50,000.00 per year? Or what percentage of the US population has a sink with a faucet in their home in 2010? The answers to both questions are simple and easy to find with the current set of data. What someone can do with those answers is what sets wisdom or actionable intelligence and information/data apart. A question that could only be answered with more advanced data collections and the creation of actionable intelligence would be; what is the solution for poverty in a specific geographical area or demographic? This question cannot be answered with a set of trend data. The answer would require the collection of advanced data and further study and application of knowledge about social programs, government resource limitations, past success in raising families out of poverty as well as privacy policies for data and legal restriction. Another issue is that questions were not specifically linked to the problems needing to be solved when developing the American Community Survey, the short form census or the SOI annual reports. The old approach is inconsistent because the government agencies in charge were collecting data to report to the public as opposed to solving problems.

## 5. Data Privacy Implications

The increased use of connected devices utilizing IoT (Internet of Things), and associated data collection and data usage have generated data privacy concerns with 74% of Americans saying it is very important for them to be in control of who can get information about them [24] and thus have access to Personally Identifiable Information (PII). What is PII? According to the department of labor "Any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means. Further, PII is defined as information: (i) that directly identifies an individual (e.g., name, address, social security number or other identifying number or code, telephone number, email address, etc.) or (ii) by which an agency intends to identify specific individuals in conjunction with other data elements, so-called indirect identification (these data elements may include a combination of gender, race, birth date, geographic indicator, and other descriptors). Additionally, information permitting the physical or online contacting of a specific individual is the same as PII. This information can be maintained in either paper, electronic or other media" [6].

Major data privacy breaches since 2013 including Target, Uber, and Equifax have further increased public awareness of privacy issues [2] with data collected by many organizations in the United States, including but not limited to websites like Facebook and Amazon, government agencies, credit bureaus and telecommunications companies. An open source data system could create major problems with data collections of personal information or information that is able to be traced back to the original subject even after personal data is stripped from the published data. Hackers and criminals have grown sophisticated enough to bypass even some of the best secure systems including government agencies like the IRS in 2015 [21]. Privacy is a big issue that needs to be considered when attempting to create any large data system that includes potentially private data. This open source data model will address the issue of data protection and privacy in the three following ways.

First, all data submitted and published for public consumption will have to have all PII stripped from the original data before it is accepted and published. Only meta data will be allowed to be included in the system. For example, IRS submits raw tax return information for the tax year 2015. The data including names, address, social security numbers, business names and any other information included on the return will need to be taken off the data set (it is okay to insert a neutral unique identifier for record tracking). The remaining data should be line by line items or aggregated items to make tracing the data back almost impossible.

Second, privacy can be protected by not giving any original data with PII to the agency or organization that controls the meta data. In the example above the agency maintaining the open source system would not have any risk of a breach effecting the personal information of the data since they didn’t have it in the first place. The open source data system is simply a roll up of all the summary data from various sources. This is not as important for data that is already public but would be crucial for data that includes PII in the original data sets like banking records, tax returns etc. It would be crucial when combining several data sets that have a public and private mix. Surprisingly, very few pieces of information are needed to identify individuals in the United States. For example, 5-digit zip, data of birth and gender could be used to identify 87% of the population. Place, gender and date of birth could be used to identify 53% of the population and county, gender and date of birth could be used to identify roughly 18% of the population [28].

Third, only knowledge, data, and information necessary to create actionable intelligence should be included in any meta data to avoid possible reverse engineering of original data from meta data sets. The General Data Protection Regulation (GDPR) can be used as a guide for keeping personal data private and untraceable. GDPR is a European Union regulation that helps to protect private data held by companies that operate in the European Union. GDPR became law on May 25, 2018. It is designed so that data cannot be traced back to the original person that provided the data or the subject of the data.

Debate on data privacy escalated on April 10 and 11, 2018, when Mark Zuckerberg, Chief Executive Officer of Facebook, testified before the United States Congress on data privacy and other issues raised following the disclosure that Cambridge Analytica obtained Facebook member data and used it to aid in election advertising for conservative candidates in the 2016 US presidential elections. [18]. While congress and the public gave the appearance that this was an unethical use of personal data there is evidence to support that the public did not care. There was much celebration in the data analytics community following the 2012 presidential election when data analysts working with the Obama campaign made that campaign the first to use data analytics to drive campaign strategy and marketing [23]. While the resulting Obama victory was wildly hailed as a victory for data analytics which the data scientists celebrated, the announcement that the Trump campaign had done essentially the same thing in 2016 was roundly criticized [9]. Was this a real change in public opinion on data privacy?  Evidence may suggest not, as people increased their Facebook usage following Zuckerberg’s congressional testimony [17]. Our conclusion is that while the privacy debate is raging, citizens of the United States are not so enamored with data privacy to prevent an open source, actionable intelligence based approach to a census as a viable alternative.

## 6. Possible Direction for the Creation of Actionable Intelligence

The United States government does not have a central statistics agency, each department collects and analyzes its own data and does not necessarily share that data with other departments or agencies. This sort of lose knit or decentralized data collections and analysis process lacks an overriding strategy and a set of goals for what they hope to achieve with data collection and analysis. Most of the data collected is centered around trend analysis which does not create the kind of actionable intelligence that can be used to create solutions based on fact. Trend data can be taken out of context easily, or misinterpreted. The proposed strategy advocates collecting data with the intention of providing clear pictures of reality, using multiple sources of differing information and painting a broad picture with context and facts of the overall population of the United States. Below is an example of why trend analysis and the current government data collection process can easily be misleading and misinterpreted. Census data about income is incomplete, the American Community Survey asks about gross income information in the income section of the survey. This does not produce take home pay which can have a wide variance depending on the make-up of a household. Note the following examples of the same size household with different marital status’.

- Example 1: Two adults unmarried each with one child making $50,000 per year each. In this situation let us assume that the adults do not comingle funds, share a bedroom and are not in a relationship and both are able to file head of household.
- Example 2: Two adults married with 2 children making $50,000 per year each.
- Example 3: Two adults unmarried each with one child making $50,000 per year each. In this situation let us assume that the adults are in a relationship comingle funds and only one can file Head of Household filing status.

Table 1 summarizes the federal tax calculations. Note that had the American Community survey gathered the data in Table 1, the incomes would have an average of $100,000.00 per year per household without considering the bottom line number of what the “household” takes home in net pay.

**Table 1**. Federal Income Tax Calculation

| | Example 1 | Example 2 | Example 3 |
|---|---|---|---|
| Income | $100,000.00 | $100,000.00 | $100,000.00 |
| Standard deduction | $(18,700.00) | $(12,700.00) | $(15,700.00) |
| Exemptions | $(16,200.00) | $(16,200.00) | $(16,200.00) |
| Table income* | $65,100.00 | $71,100.00 | $68,100.00 |
| Tax | $(8,342.00) | $(9,734.00) | $(9,036.00) |
| Net Pay ** | $56,758.00 | $61,366.00 | $59,064.00 |

** taxable income for example 1 and 3 is calculated in 2 separate returns using combined totals.*

***for purposes of this example net pay is before state and local tax, SSI and Medicare and ignores tax credits.*

## 7. Recommendations

### 7.1. Proposed Actionable Intelligence Strategy Goals

To create a system that fulfills the needs of multiple independent parties, government and the public the following broad strategic goals are proposed.

Goal 1: The system is available to anyone to create reports, aid in research and help to develop actionable intelligence for the development of solutions to the problems associated with population statistics.

Goal 2: The system allows for working with raw data from multiple sources which can be updated on a continuous basis as well as other increments of time depending on the data. Discrete historical data could also be included for context creation and comparisons.

Goal 3: The system allows for maximum flexibility and transparency while maintaining the privacy of PII.

Goal 4: The system will maintain privacy by providing raw data that cannot be traced back to individuals within the data sets using either sources in this system or combining other data sets.

Goal 5: System transparency is on a scale with privacy, the more transparency provided the more likely privacy can be compromised. This scale should be weighted heavier towards privacy than transparency to ensure that privacy is protected.

Goal 6: Data warehouse design should be used to store all data in one place and following industry best practices for data governance, access control, data security and privacy protection.

### 7.2. Proposed General Data Standards

The proposal of collecting census type data, be it socio economic, housing or financial data from multiple sources brings up the question of the consequences of collecting data that is in different formats and asks questions differently. Each data set uploaded, or data source provided should be graded from 1 being the lowest quality in each category and 5 being the best quality in each category. Proposed standards for this strategy are as follows.

Standard 1: All data sets must include age (not date of birth), gender, race and location for segmentation purposes as a minimum. Possible other information to include would be education level or political ideology.

Standard 2: No ranged data. Financial information in census products includes data sets in ranges. Raw data should be provided with no ranges for the end user to be able to segment properly into whatever data ranges they see fit.

Standard 3: Use the most accurate source of information. Any time a collector of data has a vested interest in keeping accurate records the data is more likely to be accurate. For example, utility companies need to keep accurate records of energy use because they must bill for their services. Data sources that are the most likely to provide the most accurate sources of information should always be used. A rating system for data accuracy will be used to score each data set.

Standard 4. Ask questions in a formatted way. Having a no write in option for any answer will not only standardize the answers and data but it will make aggregating the data quicker and more accurate. Keeping the options down to a minimum amount to answer the question in a useful way helps to bring more value to the data collected.

Standard 5: All data must be raw data without PII. Utility records for example would need to be stripped of the specific address of the subject and only include information pertinent to that data set. Data is segmented at the raw level and then stripped of its personal data to avoid PII at the source of the data set as opposed to at the location of the data. Query level data, as well as summary data could also be used. A rating system for quality of data or flexibility of data could be established.

Standard 6: Reports and analysis must create context rich reports. As an example, the current trend analysis style of government data collections and analysis creates opportunities to present data that is out of context. The annual average wage index, produced by the social security administration has increased in all but one year since 1951 [27]. Out of context this could be used as a way to portray that annual wages are on the rise throughout the United States for all but one year in the last 65. To provide context for this dollar amount other data points would need to be collected and used alongside this point. These other data points might include: (a) Inflation index to show how much of the increase is simply due to inflation or a normalized table that takes inflation out of the annual comparison; (b) Spending power which would show how much those dollars are worth in comparison to prior periods; and (c) Average increase in different professions, levels of education, geographic zones, age groups etc. To show how much of the annual increase can be allocated to outlier increases in different segments. Perhaps in this scenario Chief Executive Officer, CEO, pay has increased enough to skew the overall data, or annual salary in a specific profession or location.

### 7.3. Proposed Open Source Process

The process for obtaining data for the below examples would be dependent on the agency that the data would be coming from. Using the precedent set by Chetty and Saez [22] would be how any IRS raw data would be obtained. The Same precedent could be used for obtaining social security information. SSN's would not be kept in the final data tables to avoid additional security risk. Data that is already public would be scraped from the public source and entered into a database, for example, property tax information county by county is public information. Some of the proposed data sources do not currently exist in an open and available environment. Categorized banking and credit card information does exist but is likely unshared, proprietary information. The agency that created the actual system for this strategy would be responsible for developing relationships and possibly influencing legislation that would make a full open source system possible

## 8. Examples

**Example 1: What is the Average Take Home Pay for The Population of the United States?** Most measures of income including average salaries and average household income concentrate on gross pay, for example the American Community Survey asks about income in its income section and refers to gross income not net. The calculation for this number is much easier to obtain that any other income trend figure. In section 6, we provided an example of three different tax scenarios that started with the same number of household members and the same income and with a simplified tax calculation produced three different take home pay numbers. After adding in credits, other deductions and state and local tax items these take-home pay numbers could be drastically different. That example shows, even with a simplified calculation method, that take-home pay will have a much different outcome than looking at trends in gross pay which is simply the amount of money that a person earns before any taxes, including income tax, social security and local and state tax, are deducted. Take home pay is really the only number that matters. Wages could go up every year by 10% for everyone on average and if it is being outpaced by an increase in deductions from pay or a decrease in credits that a person received the previous year the inference from the data would be very different when looking at take home versus gross pay. This could paint an unrealistic picture of the economy and the populations spending power in any given year or other period. Using a simplified method can hide things that are happening between gross pay and take home pay.

It is important to create content rich actionable intelligence with data projects. Example 1 is an input to Example 2 about poverty. Data segmentation is very important in this data set. The population of the data set is large enough that multiple data segments would be possible without creating groups of data that do not have a large enough sample size to make good data inferences [7]. According to the IRS, approximately 145,329,000 tax returns where filed during the 2017 filing season for tax year ending 12/31/16 [14]. The following data would need to be collected to provide enough information to create actionable intelligence in accordance with this proposed strategy in section 7.1 for population statistical analysis. We propose that, to provide trend data as well as context that at least 50 years of data is taken. Adjustments would need to be made for inflation on an annual basis. Data needed would be as follows; SSN, zip, age, gender race, filing status, gross income, deductions tax due and credits

Changes in take home pay can be assigned to actual changes in gross pay versus take home pay. For example, from one year to the next changes in the tax code could increase or decrease take home pay and produce results that were hidden when only looking at trend data for gross pay. The effectiveness of certain tax credits could be measured as well by looking at the before and after trends of take home pay. The EITC (Earned Income Tax Credit) and child tax credit – created in 1975 and 1997 respectively– would create good breaking points for detailed take home pay analysis [12]. A measure of effectiveness for a tax credit would be valuable in creating future tax credits or other tax policy.

**Example 2: What is Poverty?** Poverty is defined by two widely used formulas, the official poverty measure and the supplemental poverty measure [26]. The formula does not go into what poverty means beyond the amount a household must make in annual gross income to be “under the poverty line” nor does it adjust fully for pricing differences across the United States in either formula, though the supplemental formula does adjust for some pricing differences in housing. Using this system, the following data would be needed to develop a comprehensive definition of what amount the poverty thresholds should be set at as well as what poverty is, beyond the simple dollar amount. The main features of the actionable intelligence that this data would create is an ability to subset data and rapidly determine poverty levels using multiple different criteria. As well as to make decisions about which anti-poverty programs should have high priority and which ones should not or to design new programs and solutions based on the results. The following chart provides the data that would be needed and a prime source for it along with an alternative.

All data would need to include common population descriptions including, age, gender, race, marital status, education level and geographic tags like county, city of zip code. The preference for geographic tags should be what most closely matches the reality of price differences from one geographic code to the next. For example, there are 43,000 zip codes in the united states [38], it is unlikely that zip codes that are right next to each other produce a statistically significant price or wage difference. Conversely, there are 3,141 county or county equivalents in the United States [36] counties are more likely to have a statistically significant difference in prices and wages than zip codes. We suggest that future research determine significant geographic boundaries to base population statistical analysis on, for now the tables include zip code as the geographic tag.

Non-data set related information to create an overall context would need to include proposed budget levels for each item. Definitions for a needed item versus a non-need item for each geographic region must be included. For example, transportation could be considered a need though public versus private would have to be taken into consideration depending on the geographic attributes, New York City where owning a car is not a need versus San Diego, where public transportation is not as robust and a car is a need.

The results could be used to aid in the creation of anti-poverty programs targeted by geographic regions, household sizes, and other population segments. It could also be used to develop personal benchmarks to help the public develop their own personal healthy spending habits as well as answer questions about systematic poverty versus families that are simply living beyond their means. The information could also help determine clearly stated goals for anti-poverty programs and tracking their effectiveness. (a) What are needs per person segmented by geographic location? (b) What is a needs budget for a household of 4 (2 children,2 adults)? (c) What is the delta between the needs budget and the current average budget? and (d) Do spending patterns show that part of the problem with poverty is overspending?

**Table 2**. Needed Data Set’s and Sources for Poverty Data Example

| Data set | Current source of data | Proposed source of data |
|---|---|---|
| Take home pay output example 1 | N/A | IRS Raw data |
| Household size | ACS | IRS Raw data |
| Rent/Mortgage | ACS | Mortgage holder raw data |
| Monthly debt | Census Data | Debt holder raw data |
| Vehicle expenses | ACS | Categorized bank/Credit card raw data |
| Gas | ACS | Categorized bank/Credit card raw data |
| Groceries | BLS (1) | Categorized bank/Credit card raw data |
| Telecommunications | BLS (1) | Categorized bank/Credit card raw data |
| Health Insurance | ACS | Categorized bank/Credit card raw data |
| Medical Expenses | BLS (1) | Insurance company raw data |
| Property Taxes | BLS (1) | Hospital and bank Credit card raw data |
| Other expenses | BLS (1) | Local public property tax raw data |

## 9. Conclusions

This study has explored the feasibility and need for radically changing the way the United States census is performed to that of using an actionable intelligence approach to generating a data strategy and then using that strategy to identify open sources of data. We have explored the issue of data privacy and while much was made of data privacy in 2018 due to Facebook and the GDPR, we conclude that the public is willing to utilize this approach as long as data privacy is addressed and there is a large cost reduction in the census process.

Our recommendations to ensure data privacy include stripping out personally identified information from the data and metadata and replace with nontraceable identifiers. Control access to the source data to only those with a need to know while allowing general access to the stripped data. Store data in secure data storage at the source. Not storing multiple copies of the data. Transmit data using secure connections. Use the most accurate data available. Do not use paper surveys.

This study is limited to the role of the federal government in data collection activities and designing a strategy that can be used for other groups as well. An open source system for government data collections means that it will be more accessible and not fall under the direct management of the federal government though it will be able to be accessed by agencies within the government. This study does not go into the absolute effectiveness of possible data that will be collected, it is simply a proposal for a strategy that can offer more agility, transparency and actionable intelligence for decision making to be used on by many different groups. The study is also limited to two specific government agencies the IRS and the US Census bureau in analyzing what information is currently collected, their methods and purpose of the data that is collected. Many other departments of the federal government collect data on the US population and publishes multiple reports ranging from weekly reports to reports published every ten years, like the census short form results. Other major agencies that participate in data collections include the Department of labor, agriculture, education and energy.

## 10. References

[1] Ackoff, R. L. (1989). From data to wisdom. Journal of Applied Systems Analysis, 16(1), 3-9.

[2] Armerding, T. (2017). The 16 biggest data breaches of the 21st century. CSO. Retrieved December 1, 2017 from https://www.csoonline.com/article/2130877/data-breach/the-16biggest-data-breaches-of-the-21st-century.html

[3] Beine, J. (n.d.). The cost of the U.S. census. Retrieved June 13, 2018 from http://www.genealogybranches.com/censuscosts.html

[4] Black, E. (2001). IBM and Holocaust. Crown Publishers, New York, New York, United States of America.

[5] Card, D., Chetty, R., Feldstein, M., & Saez, E. (2011). Expanding access to administrative data for research in the United States (National Science Foundation 10-069). Alexandria, VA: National Science Foundation.

[6] Department of Labor. (n.d.). Guidance on the protection of personal identifiable information. Retrieved June 13, 2018 from https://www.dol.gov/general/ppii.

[7] Deziel, C. (2017). The effects if a small sample size limitation. Retrieved June 10, 2018 from https://sciencing.com/effects-small-sample-size-limitation-8545371.html.

[8] Digital Accountability and Transparency Act of 2014, (2014). S. 994, 113d Congress. Retrieved June 11, 2018 from https://www.congress.gov/bill/113th-congress/senate-bill/994.

[9] Editorial Staff, (2018). Funny, When Obama Harvested Facebook Data On Millions Of Users To Win In 2012, Everyone Cheered. Investor's Business Daily, March 19, 2018.

[10] Griffin, D. H. (2011). Cost and workload implications of a voluntary American community survey (Report No. 20233-0001). Washington D. C.: U. S. Census Bureau.

[11] House Oversight Committee. (2017). Hearing on the 2020 Census. Retrieved June 9, 2018 from https://oversight.house.gov/hearing/hearing-2020-census/.

[12] Hungerford, T. L., & Thiess, R. (2013). The earned income tax credit and the child tax credit (Report No. 370). Washington, DC: Economic Policy Institute.

[13] Internal Revenue Service. (2016). 1040 Tax Tables 2016 (IRS Publication No. 24327A). Retrieved May 30, 2018 from https://www.irs.gov/pub/irs-prior/i1040tt--2016.pdf..

[14] Internal Revenue Service. (2017). Individual income tax returns, 2015 (IRS Publication No. 1304). Retrieved May 30, 2018 from https://www.irs.gov/statistics/soi-tax-stats-individual-incometax-returns-publication-1304-complete-report.

[15] Jennex, M., (2017a). "Re-Examining the Jennex Olfman Knowledge Management Success Model," 50th Hawaii International Conference on System Sciences, HICSS50, IEEE Computer Society, January 2017.

[16] Jennex, M.E., (2017b). Big Data, the Internet of Things and the Revised Knowledge Pyramid, Data Base for Advances in Information Systems, 48(4), pp. 69-79.

[17] Kanter, J. (2018). The backlash that never happened: New data shows people actually increased their Facebook usage after the Cambridge Analytica scandal. Business Insider, May 20, 2018. Retrieved on June 9, 2018 from http://www.businessinsider.com/people-increased-facebook-usage-after-cambridge-analytica-scandal-2018-5

[18] Kozlowska, H. and Timmons, H. (2018). What we learned from Mark Zuckerberg's Congressional testimony. Quartz, April 13, 2018. Retrieved on June 3, 2018 from https://qz.com/1251646/what-we-learned-from-mark-zuckerbergs-congressional-testimony/.

[19] Malenich, J. L., & Dacey, R. F. (2017). Fiscal years 2016 and 2015 consolidated financial statements of the U.S.

government (Report No. GAO-17-283R). Washington, DC: GAO U.S. Government Accountability Office.

[20] Marr, B. (2015). A brief history of big data everyone should read. World Economic Forum. Retrieved May 24, 2018 from https://www.weforum.org/agenda/2015/02/a-briefhistory-of-big-data-everyone-should-read/.

[21] McCoy, K. (2016). Cyber hack got access to over 700,000 IRS accounts. USA Today. Retrieved June 1, 2018 from https://www.usatoday.com/story/money/2016/02/26/cyberhack-gained-access-more-than-700000-irs-accounts/80992822/.

[22] Mervis, J. (2014). How two economists got direct access to IRS tax records. Retrieved May 25, 2018 from http://www.sciencemag.org/news/2014/05/how-two-economists-got-direct-access-irstax-records.

[23] Pilkington, E. and Michel, A., (2012). Obama, Facebook and the power of friendship: the 2012 data election. The Guardian, Feb 17, 2012.

[24] Rainie, L. (2016). The state of privacy in post-Snowden America. Pew. Retrieved June 13, 2018 from http://www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy-in-america/.

[25] Scherer, M., & Bahrampour, T. (2017). 2020 census needs cash infusion, commerce secretary will tell congress. Chicago Tribune October 20, 2017,. Retrieved May 15, 2018 from http://www.chicagotribune.com/news/nationworld/politics/ct-2020-census-money-20171010-story.html.

[26] Short, K. (2011). The Research Supplemental Poverty Measure: 2010 (Report No. P60-241). Suitland, MD: United States Census Bureau.

[27] Social Security Administration. (n.d.). National average wage index. Retrieved from https://www.ssa.gov/oact/cola/AWI.html

[28] Sweeney, L. (2000). Simple demographics often identify people uniquely (Data Privacy Working Paper 3). Pittsburgh, PA: Carnegie Mellon University. .

[29] United States Census Bureau, (n.d.a). History. Retrieved on June 5, 2018 from: https://www.census.gov/history/

[30] United States Census Bureau. (n.d.b). Pop culture: 1790. Retrieved from https://www.census.gov/history/www/through_the_decades/fast_facts/1790_fast_fact s.html

[31] United States Census Bureau. (n.d.c). Pop culture: 2010. Retrieved from https://www.census.gov/history/www/through_the_decades/fast_facts/2010_fast_fact s.html

[32] United States Census Bureau. (n.d.d). American community survey. Retrieved from https://www.census.gov/history/www/programs/demographic/american_community_ survey.html.

[33] United States Census Bureau. (n.d.e). Tabulation and processing. Retrieved from https://www.census.gov/history/www/innovations/technology/tabulation_and_process ing.html.

[34] United States Census Bureau. (2010). Door-to-door visits begin for 2010 census: Census takers to follow up with about 48 million households nationwide. Retrieved from https://www.census.gov/newsroom/releases/archives/2010_census/cb10-cn59.html.

[35] United States Census Bureau. (2011). What is the census? Retrieved from https://www.census.gov/2010census/about/.

[36] United States Geological Survey. (n.d.). How many counties are there in the United States? Retrieved June 3, 2018 from https://www.usgs.gov/faqs/how-many-counties-are-there-united-states

[37] United States Government Accountability Office. (2011). 2010 Preliminary Lessons Learned Highlight the Need for Fundamental Reforms (Report No. GAO-11-496T).

[38] ZIP Boundary. (n.d.). ZIP code FAQs. Retrieved from http://www.zipboundary.com/zipcode_faqs.html