# Putting Our Industry's Data to Work: A Case Study of Large-Scale Data Aggregation

## Preprint

Elaina Present,[1] Chris CaraDonna,[1] Eric Wilson,[1] Natalie Frick,[2] Janghyun Kim,[1] Rajendra Adhikari,[1] Anna C. McCreery,[3] and Elizabeth Titus[4]

*1 National Renewable Energy Laboratory*
*2 Lawrence Berkeley National Laboratory*
*3 Elevate Energy*
*4 Northeast Energy Efficiency Partnerships*

*Presented at the 2020 ACEEE Summer Study on Energy Efficiency in Buildings*
*August 17-21, 2020*

**NREL is a national laboratory of the U.S. Department of Energy**
**Office of Energy Efficiency & Renewable Energy**
**Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

**Conference Paper**
NREL/CP-5500-77102
September 2020

# Putting Our Industry's Data to Work: A Case Study of Large-Scale Data Aggregation

## Preprint

Elaina Present,[1] Chris CaraDonna,[1] Eric Wilson,[1] Natalie Frick,[2] Janghyun Kim,[1] Rajendra Adhikari,[1] Anna C. McCreery,[3] and Elizabeth Titus[4]

*1 National Renewable Energy Laboratory*
*2 Lawrence Berkeley National Laboratory*
*3 Elevate Energy*
*4 Northeast Energy Efficiency Partnerships*

**Suggested Citation**
Present, Elaina, Chris CaraDonna, Eric Wilson, Natalie Frick, Janghyun Kim, Rajendra Adhikari, Anna C. McCreery, and Elizabeth Titus. 2020. *Putting Our Industry's Data to Work: A Case Study of Large-Scale Data Aggregation: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-5500-77102. https://www.nrel.gov/docs/fy20osti/77102.pdf.

**NREL is a national laboratory of the U.S. Department of Energy**
**Office of Energy Efficiency & Renewable Energy**
**Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

**Conference Paper**
NREL/CP-5500-77102
September 2020

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

**NOTICE**

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via www.OSTI.gov.

*Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.*

NREL prints on paper that contains recycled content.

# Putting Our Industry's Data to Work: A Case Study of Large-Scale Data Aggregation

*Elaina Present, National Renewable Energy Laboratory*
*Chris CaraDonna, National Renewable Energy Laboratory*
*Eric Wilson, National Renewable Energy Laboratory*
*Natalie Frick, Lawrence Berkeley National Laboratory*
*Janghyun Kim, National Renewable Energy Laboratory*
*Rajendra Adhikari, National Renewable Energy Laboratory*
*Anna C. McCreery, Elevate Energy*
*Elizabeth Titus, Northeast Energy Efficiency Partnerships*

**ABSTRACT**

With increasing deployment of advanced metering infrastructure (AMI), building automation system (BAS) controls, internet of things (IoT) network devices, and data-driven evaluation, measurement, and verification studies, the building sector is currently generating a staggering amount of energy-related data. In the right hands, these data sets can contribute to increased comfort and energy savings for building occupants and a more reliable electrical grid; however, due to a combination of factors, including significant privacy concerns, much of the data that are presently generated and stored are not used outside basic operational applications.

In the past year, our team has dedicated over 2,000 person-hours to accessing building energy data for a project funded by the U.S. Department of Energy (DOE) Building Technologies Office. We sought whole-building or end-use (e.g., lighting) timeseries data at the individual-building or equipment level where possible or aggregated information, such as timeseries averages and quartiles by building type (e.g., office, retail, hospital), where sharing individual building information was not an option. We are additionally working with IoT and BAS data sets to derive information important to the project.

We have assembled an extensive data set that will enable the development of publicly available end-use load profiles to benefit the U.S. building and electricity industries. Here we present an overview of the data set that we have assembled to date, the motivators and approaches that got us here, and the lessons we learned through our efforts. We also discuss work underway that presents additional options for future data access.

## Background and Motivation

### Building Energy Data

The building sector is generating a staggering amount of data. Advanced metering infrastructure (AMI) deployment is widespread across the country. In 2018, 57% of residential meters and 54% of commercial meters were currently AMI meters (EIA 2019). This deployment is uneven—20% of electric utilities with at least 500 meters have over 90% of both their residential and commercial customers on AMI meters (EIA 2019), (NEEP 2017). Nonutility players are also gathering large quantities of relevant building energy and operation data. Smart



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

thermostats, for example, record indoor temperatures and setpoints; these are currently installed in 13% of U.S. homes, a figure that is rapidly increasing (Parks Associates 2019).

Collectively, these data contain a wealth of information about how and when energy is used in buildings, yet data owners are often reluctant to share data externally. In some cases, this is for reasons of competitive advantage provided by having the data, but in most cases it is due to customer privacy concerns and the optics of sharing data. Anonymous data sharing is sufficient for many potential data uses and could allow these data to contribute to a more efficient, reliable, and cleaner energy industry—but data are rarely shared even in anonymous form.

There are various reasons for this. One is concern about theoretical re-identification of anonymized data by cross-correlation with other public data sets (Narayanan and Shmatikov 2008), (Barth-Jones 2012). Another is that data are often stored in a way that creates significant effort to share (e.g., ad hoc data formatting, nonstandard storage medium, stored among multiple parties). A third concern is that the level of metadata (e.g., primary use type of a commercial building) that may be needed to make data useful is not available independent of customer-identifiable data fields (e.g., address). Therefore, currently, the data generated by the buildings sector is overwhelmingly used only by its owners (the utilities, property managers, data analytics companies, and so forth), their contractors, and their regulators. It is not used in bulk by research organizations, such as national laboratories, or consumer advocacy organizations operating independently of the data-owning organization.

Data aggregation across sources is always difficult, but the fragmented nature of these valuable data sets makes this an even more challenging situation. Some of the questions that these data sets could help answer cover large geographic areas—grid balancing in the Western Electricity Coordinating Council or federal energy policy, for example—and require compiling insights from across myriad data sets. Each disparate data set will include its own specified fields, units of measurement, time interval, and other norms, as well as its own data security requirements. The holders of each data set will all have different data-sharing guidelines and infrastructures, many of which have not been exhaustively tested in practice. These obstacles take time, resources, and a motivating project to unravel.

## The End-Use Load Profiles Project and Its Data Needs

The National Renewable Energy Laboratory (NREL) is leading a 3-year project funded by the Department of Energy (DOE) to create end-use load profiles for the U.S. building stock, in collaboration with Lawrence Berkeley National Laboratory (LBNL) and Argonne National Laboratory and under the guidance of a Technical Advisory Group (TAG). We are creating these profiles using our best-in-class building stock energy modeling tools, ComStock and ResStock (NREL n.d.), informed by and validated against the best available ground-truth building energy and operation data. These end-use load profiles will be publicly available at the end of the project, along with quantified uncertainty characterizations of key metrics. We expect the resulting load profiles will be used for time-sensitive valuation of energy efficiency—one of the original motivators for the project—along with demand side management planning, electrification impact studies, program design, nonwires alternatives analysis, economic analyses of solar and storage, policy analysis, and many other purposes. We will be creating end-use load profiles for all regions of the contiguous United States, regardless of data availability, but expect that there will be lower uncertainty in regions where we have the most robust ground-truth data.



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

We have generally seen two approaches used by other experts to create end-use load profiles. The first is a top-down approach, typically using statistical methods to disaggregate AMI data into constituent end uses. One version of this is *conditional demand analysis (CDA),* which can be a powerful approach in regions with high AMI saturation; however, CDA requires detailed knowledge about the characteristics of each of the metered buildings. For example, to derive the load profile for residential dishwashers, it is necessary to know which customers have dishwashers and which do not. The second is a bottom-up approach—submetering a statistically selected subset of buildings and extrapolating the results to represent an entire region. This is also a strong approach that can yield a high degree of accuracy with sufficient sample size and extensive submetering; however, sample sizes and submetering are often constrained due to high cost, limiting the diversity in end-use consumption patterns that can be captured.

We are leveraging what has been learned from each of these approaches. We are using white-box, physics-based models with inputs based on real data that describes the existing U.S. residential and commercial building stock. We run hundreds of thousands of statistically generated building energy models—approximately one model for every 200 buildings that we are representing—to appropriately represent the diversity seen in the stock. Physics-based modeling is computationally expensive but greatly reduces the amount of measured data necessary to complete the project, allowing us to adjust for the impacts of weather, daylight availability, and so on, on different end uses, and to extend data from one area to areas that have minimal data available. Physics-based modeling also enables running different energy efficiency or demand flexibility upgrade scenarios, which would be very hard to do with purely data-driven black-box modeling.

This project requires extensive data to inform, calibrate, and validate its models, including four basic types: weather and location, building characteristics, end-use energy consumption, and population-level whole-building data. The weather data we generated was based on the National Oceanic and Atmospheric Integrated Surface Database (Smith, Lott, and Vose 2011) and the National Weather Service stations (Bianchi and Smith 2019). For building characteristics data, we rely primarily on saturation and survey data from a variety of sources, including the U.S. Energy Information Administration (EIA), U.S. Census, and real estate databases (Frick et al. 2019). Further discussion of these data types is beyond this paper's scope.

No amount of building characteristics data, from window type to heating fuel, can tell us how people actually use buildings and thus provide the occupancy, setpoints, schedules, and other behavior-driven inputs to our models. Therefore, the third data type needed is timeseries end-use energy consumption data. This type of data is limited in quantity and quality and is generally never available for entire geographic populations of buildings, which would allow us to treat it as unbiased and use it not just to inform but to calibrate and validate our models. This leads us to rely also on population-level *whole-building* data to calibrate and validate our models. This includes widely available load research data and AMI data. Load research data—typically developed by metering a sample of customers in each customer class—is a valuable check but is not typically categorized by commercial building type, and it tells us nothing about the diversity of energy consumption across buildings. We value geographically complete AMI data sets for an entire city, county, utility territory, or other region, because they provide an inherently unbiased perspective on that location and give us insight into energy use diversity both among buildings in a region and across regions throughout the country.



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

A few relevant data sets are available for public purchase. These include the ComEd Anonymous Data Service (ComEd n.d.) and the Pecan Street Dataport (Pecan Street n.d.). Though we are using each of these to their fullest extent, these two data sets fall short of the scope of data we need to feel confident we were capturing the full variety of buildings on a national scale.

For residential end-use energy consumption data, we are using the results of the Massachusetts Residential Baseline Study (Navigant 2018), the underlying data from Florida Solar Energy Center's Phased Deep Energy Retrofit Project (Fenaughty, Parker, and Martin 2017), and the Northwest Energy Efficiency Alliance (NEEA) 2011 Residential Building Stock Assessment (RBSA)—Metering Study (NEEA 2019). These three sources are freely available. The RBSA Metering data and the Massachusetts data are based on metering of homes selected with an intent to be statistically representative. The Pecan Street and Florida Solar Energy Center data sets were not designed to have representative sampling, but have sufficient sample sizes to be of interest nonetheless. These four end-use data sets from very different parts of the country provide a solid foundation on the residential side. We found no parallel data sets from the past 10 years on the commercial side.

It is, as previously mentioned, possible for a project such as ours to meter the energy consumption of a sample of buildings at the end-use level. This is the approach used by NEEA in the Home Energy Metering Study and the Commercial Energy Metering Study; however, within our budget, it was simply not possible to do this well or comprehensively for the whole United States. Given the amount of existing data in the building sector, we also did not consider this the most efficient path. Therefore, we spent the time and resources to leverage existing data first, only resorting to metering to fill gaps.

We decided to see how well we could cover our project's remaining data needs—namely, commercial building end-use data and whole-building, whole-population data—using energy data that already exists. Others have written on the complex situation of permissions and storage of energy data (McKibbin 2014), so we understood this was a significant undertaking. The American Council for an Energy-Efficient Economy (ACEEE) maintains a state-by-state listing of regulations on third-party data access (ACEEE 2019).

## Approach

### Commercial Building End-Use Data

The limited availability of commercial building end-use data was identified as an area of focus early in the project. The team determined minimum data sample sizes that were considered sufficient for calibrating stock energy models for each end use and building type and presented them to subject matter experts on a webinar. This was done separately for weather-driven end uses, for which key metrics were presumed to be transferable across building types, and schedule-driven end uses, for which key metrics were presumed to be transferable across geographic regions but not across building types.

Through discussions with the TAG, subject matter experts, and within the project team, we realized that many of the holders of commercial end-use data are not conventional players in the space, such as utilities and the consulting firms, state energy offices, and regional energy efficiency organizations (REEOs) that support them. Commercial end-use data, where it exists, is



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

often in the hands of large commercial real estate owners, energy metering companies, energy service companies, dashboard companies, and data analytics firms, as well as research organizations such as national laboratories. And it is not necessarily in the form of energy consumption timeseries data. So, we considered the full range of potential sources, including internet of things (IoT) and building automation system (BAS) data sources that do not necessarily provide energy consumption timeseries data, but nonetheless can still provide useful insight into the operation and scheduling of building systems. And we formed a budget—because, in most cases, these potential data-sharing organizations would not expect to benefit directly from the work we would be doing with their data, we offered funding for data access, and in particular for any time spent processing or labeling data for us.

Our requirements for spending include a cost justification, and since there are not price sheets publicly available for commercial building end-use data, we put a market research process in place. We surveyed all NREL staff in the Buildings and Thermal Sciences Center to create a list of organization types and specific parties that might have commercial end-use energy consumption data, BAS data, or IoT data at scale, and worked through our networks to find contacts at each. We then reached out using a template that included a brief project overview and a list of data types we were seeking. This initial outreach was generally by email and occasionally through LinkedIn. We scheduled calls with as many of the identified organizations as were interested. On these calls, we explained our project goals and why we needed data, and asked what data might be available, what funding level or other steps would be necessary for us to access that, and if they knew of others we should contact. We followed each call with a data request tailored to the organization. This market research outreach effort lasted over two months and consumed roughly 700 NREL person-hours. Follow-up conversations and emails were often necessary.

Once we had a substantial body of funding and effort estimates from organizations, we were able to start evaluating the total cost and data coverage that would result from pursuing different combinations of data sets. Then, with DOE approval, we started the process of reaching back out to everyone we had spoken to, asking those with whom we wanted to move forward to make final estimates of their available data, and beginning the process of working through any necessary nondisclosure agreements (NDAs), as well as contracts. Once NDAs (if necessary) and contracts were in place, the data sets were transferred. Some small data sets were transferred by email. Other approaches we used included direct upload to secure cloud services,[1] a dedicated secure file transfer protocol server, and, in at least one case, encrypted hard drive sent through the mail.

Upon receipt of each data set, we did a first round of quality checks within the first 2 weeks, checking labels and units and verifying that our technical team understood what was contained in what we had received. We conducted follow-up discussions to ask any questions.

### Whole-Building, Whole-Population Data

With few exceptions, the available source for whole-building or whole-population data sets is AMI data held by various utilities. We conducted staged outreach to utilities with a target

[1] Compliant with the Federal Information Security Management Act (FISMA) Moderate baseline via NIST/FedRAMP Moderate controls and policies.



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

of achieving six substantial utility AMI data sets with representation in each of the five Building America climate zones (Baechler et al. 2010) and a stretch goal of two sets per climate zone.

We started our outreach with our TAG. From there, we reviewed EIA Form 861 2018 (EIA 2019) to identify utilities in priority regions with high AMI saturation. We reached out to these through our own networks, DOE, the TAG, the REEOs, and National Rural Electric Cooperative Association (NRECA), scheduling initial phone calls with members of the NREL technical team to go over the project and its data needs and to get a sense of the utility's interest level and likely obstacles. We followed up these initial calls with a more detailed data request.

We requested 1–2 years of anonymous AMI data for all residential and commercial AMI meters in a contiguous geographic region, such as a county or a city metro area, that has at least 3,000 commercial buildings (i.e., not industrial or agricultural in nature). We asked that the anonymous AMI data be provided with as much of the following metadata as is available:

- Building type (apartment, single family home, retail, education, office, and so on)
- Building location (e.g. zip code, nearest airport—to assign correct weather file)
- Vintage (decade built or, for commercial, major renovation)
- Size (square footage bin)
- Number of stories
- Heating type/heating fuel.

Many utilities have been able to provide location (typically zip code), a breakdown of building type by commercial or residential, and often by multifamily versus single family. A few utilities also know whether a building has gas heating or at least if it has gas service. In some cases where the electric utility is also the gas utility, we can get both gas and electricity data.

For the residential buildings, we worked with whatever was available from the above list. For the commercial buildings, having the building type and some of the other metadata is critical for us to be able to use the data effectively, and we have only seen one utility so far that had that data at the level of accuracy we wanted. We generally found that North American Industry Classification System (NAICS) codes were inconsistent and outdated. In some cases, county assessor data was usable.

In the remaining cases, we used the CoStar™ commercial real estate database ("CoStar" n.d.) to get the commercial building metadata through pairing CoStar commercial building addresses with utility commercial building addresses using code developed in-house. The address matching code uses data homogenization techniques such as replacing street suffix with U.S. Postal Service (USPS) abbreviations (USPS Postal Explorer n.d.) and fuzzy matching (such as matching without the street direction (e.g., N, S) if a direct match cannot be found. Even then, we do not get 100% match but we do get a high enough match rate in the building types we are modeling to make the data usable—generally over 60% for the building types that are in scope for this project.

In about half of cases, the utility provided us with the commercial building addresses, typically under a signed agreement that we delete them within a specified time frame. For the remaining cases, the plan had been to send an NREL data specialist to do the matching within the utility's own data environment, and then have the utility provide us with the metadata and no addresses. The current novel coronavirus pandemic has prevented this to date; instead, we have written and packaged our code for address matching and assigning commercial buildings to metadata bins and are asking utility staff to run the code with remote support from our team. In



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

all cases, we set a minimum number of buildings that must share identical metadata in accordance with utility preferences and combine metadata bins or drop buildings until it is met. This reduces the risk of re-identification.

## Results

### Outreach Results: Commercial End-Use Data

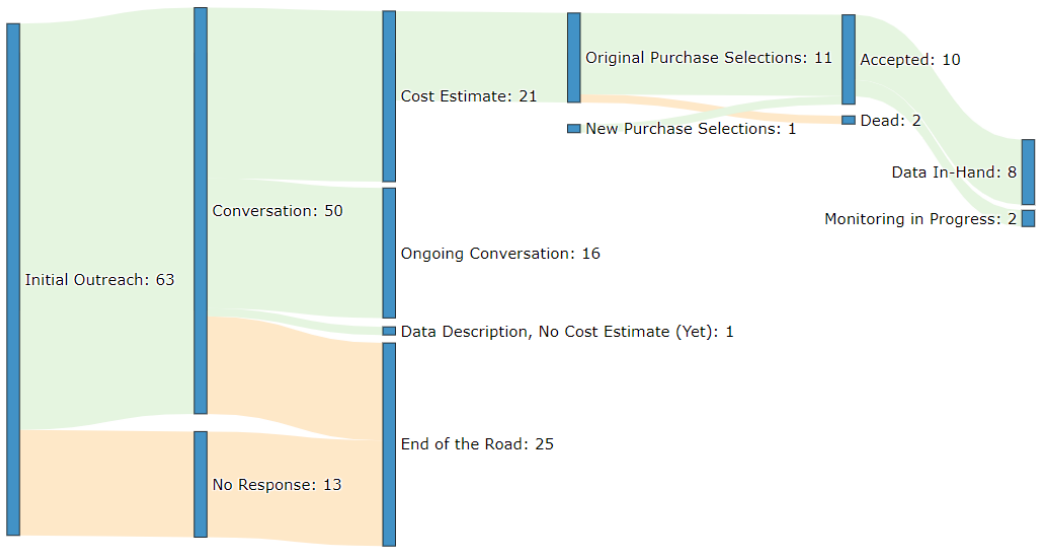


Figure 1. Outreach, contracting, and procurement results diagram for commercial end-use data

We ultimately reached out to 63 varied organizations on the topic of commercial building end-use data for the end-use load profiles project. Twenty-one provided cost estimates to us for data sets that they could provide access to. Of these, we chose 11 with which to move forward.

During the follow-ups and procurement process, a twelfth data set became available that filled gaps, and one of the original 11 declined to move forward with procurement. Another one of the original 11 identified potential data sources determined that obtaining the proper permissions to share the data set was going to be more cumbersome than expected, so the procurement of this data set was abandoned. This leaves us moving forward with 10 total data sets, and of these 10, as of June 18, we have contracts in place with 10 and data in hand from 8. The remaining two data sets for which we have contracts in place but do not yet have the data in hand are still currently collecting their data—these should be available with sufficient logged data by the end of summer 2020. These results are visualized in the Sankey diagram in Figure 1.

### Outreach Results: Whole-Building, Whole-Population

We have, to date, directly contacted 38 utilities or utility representatives (e.g., state energy offices) regarding population-scale whole-building data for this project. Of these 38, 11 have signed NDAs with us covering individual AMI data from unbiased samples (either every building in a geographic area or curated sample data) of 1,000 or more buildings. We are in



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

progress with NDAs with another four. Four data sets are currently in hand at NREL, including one that was provided for a different project and underwent an NDA modification, and two additional utilities have provided us with data that will be useful to the project, but that is not AMI data from unbiased samples of 1,000 or more buildings—for example, monthly billing data. This is visualized in Figure 2.

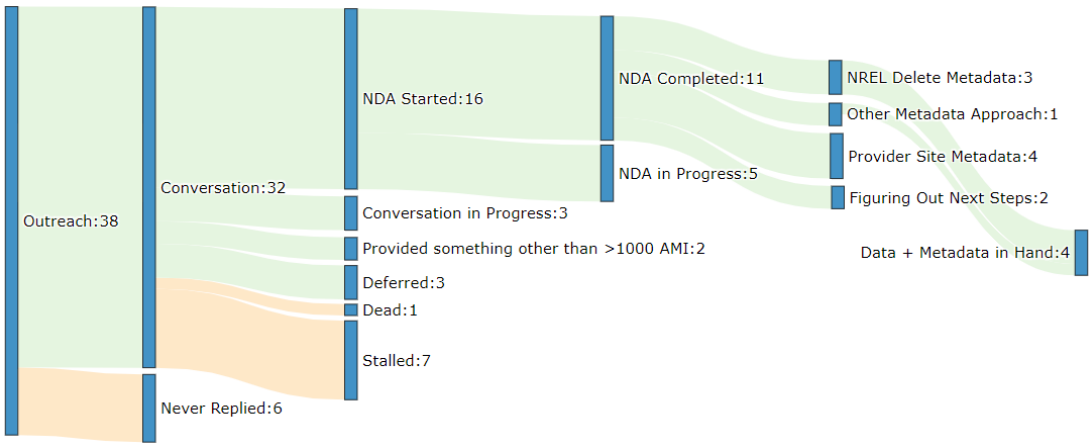


Figure 2. Outreach results diagram for whole-building, whole-population data

## The Data Set to Date: Commercial End-Use Data

The commercial end-use data that we have gathered to date is represented in Table 1 (for schedule-driven end uses) and Table 2 (for weather-driven end uses). To date, we have been able to meet or exceed our proposed minimum sample sizes in every category, with the exception of interior equipment (e.g., plug loads) in three building types: medical, food sales, and lodging.



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Table 1. Commercial End-Use Data Outreach Results for Weather-Driven End Uses

| End Use | Proposed Minimum Sample Size[1] | Pursued Sample Size[2] | Procured Sample Size[3] |
|---|---|---|---|
| Heating | 48 | 6,218 | 4,119 |
| Cooling | 48 | 6,598 | 3,957 |
| Fans | 21 | 2,497 | 645 |
| Pumps | 21 | 500 | 95 |
| Heat Rejection | 21 | 21 | 51 |
| Humidification | 21 | 27 | 53 |
| Heat Recovery | 21 | 22 | 49 |
| Refrigeration | 21 | 1,076 | 1,080 |
| Exterior Lighting | 21 | 846 | 627 |



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Table 2. Commercial End-Use Data Outreach Results for Schedule-Driven End Uses

| Schedule-Driven | | Hospital | Outpatient | Primary School | Secondary School | Full-Service Restaurant | Fast-Service Restaurant | Retail | Strip Mall | Supermarket | Small Hotel | Large Hotel | Warehouse | Multifamily | Sm. Office | Med. Office | Large Office |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proposed Minimum Sample Size[1] | Interior Lighting | 21 | | 21 | | 21 | | 21 | | 21 | 21 | | 21 | n/a | 21 | | |
| | Plug and Process | 21 | | 21 | | 21 | | 21 | | 21 | 21 | | 21 | n/a | 21 | | |
| | Service Water Heating | 0 | | 0 | | 0 | | 0 | | 0 | 0 | | 0 | n/a | 0 | | |
| | Cooking | n/a | | n/a | | n/a | | n/a | | n/a | n/a | | n/a | n/a | n/a | | |
| Pursued Sample Size[2] | Interior Lighting | 103 | | 281 | | 760 | | 1046 | | 137 | 53 | | 270 | 20 | 337 | | |
| | Plug and Process | 2 | | 285 | | 196 | | 214 | | 4 | 5 | | 25 | 22 | 270 | | |
| | Service Water Heating | 0 | | 0 | | 316 | | 106 | | 0 | 0 | | 0 | 0 | 1 | | |
| | Cooking | 0 | | 2 | | 2618 | | 0 | | 0 | 1 | | 0 | 0 | 0 | | |
| Procured Sample Size[3] | Interior Lighting | **76** | | **178** | | **159** | | **908** | | **92** | **42** | | **132** | **74** | **133** | | |
| | Plug and Process | **4** | | **332** | | **56** | | **30** | | **13** | **2** | | **36** | **270** | **91** | | |
| | Service Water Heating | **0** | | **6** | | **58** | | **99** | | **15** | **1** | | **15** | **101** | **10** | | |
| | Cooking | **0** | | **9** | | **687** | | **18** | | **23** | **0** | | **1** | **6** | **8** | | |

[1]Proposed minimum sample size targets for successful project completion, presented at a subject matter expert webinar in August 2019. No goals were set for the cooking end use or the multifamily building type.

[2]Counts include in-hand data, data we expected to be able to get for free from other research institutions, and data to which we intended to purchase access—the potential purchase sample sizes were based on vendor rough estimates obtained during market outreach.

[3]Procured sample size includes data in hand and data that is being contracted for procurement.

= gap in coverage



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

The data we have is geographically diverse, with at least one end-use timeseries from one building in 49 states, and no one state representing more than 16% of the total buildings on which we have any data.

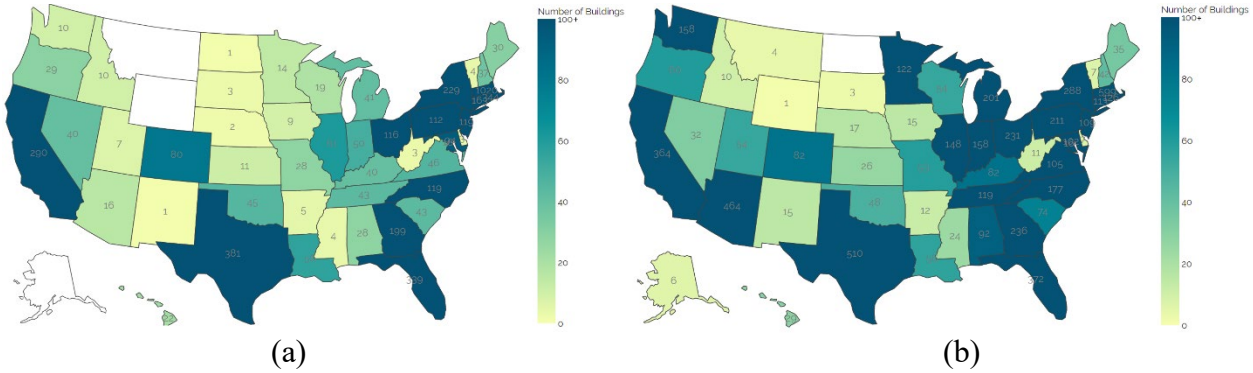


Figure 3. Geographic coverage of commercial end-use data by number of buildings with at least one end use represented in our data set. (a) represents schedule-driven loads and (b) represents weather-driven loads.

## The Data Set to Date: Whole-Building, Whole-Population

As of June 18, we have 11 NDAs completed that each cover 1,000 or more unbiased buildings of AMI data, with representation in all five Building America climate zones, as shown in Figure 5. We have six AMI data sets in hand, including the ComEd Anonymous Data Service data, the data set that was in hand from another project, and four that were gathered during our data outreach efforts. The full set of expected metadata is in hand for three of these four.

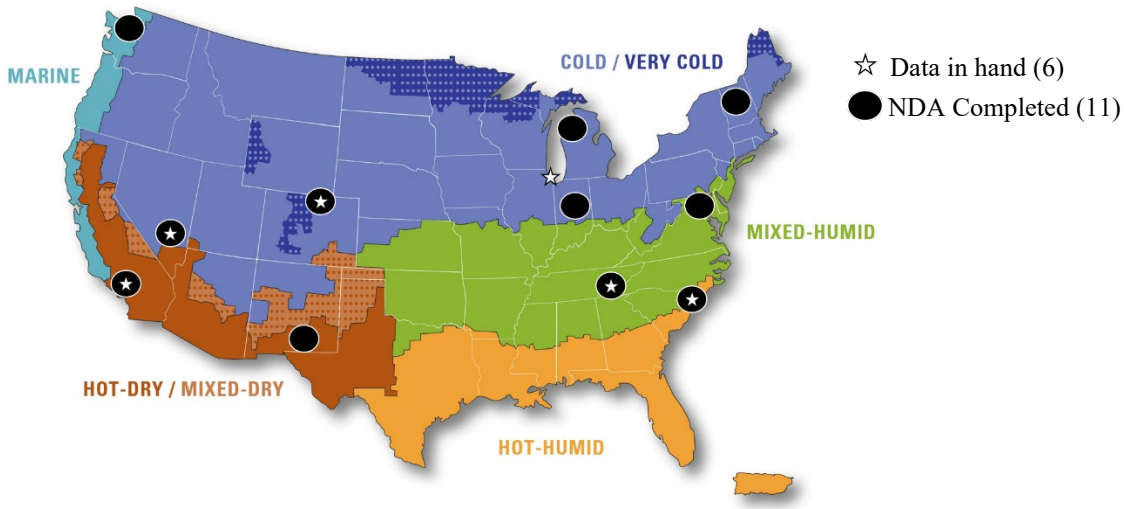


Figure 4. Whole-building, whole-building data outreach results to date visualized compared to the Building America climate zones



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

## Discussion

We were in an unparalleled position to undertake this data-gathering effort. We have the benefit of being a noncommercial national laboratory doing publicly funded work in the national interest, we have established, tested data privacy procedures that meet government standards, and we have the funding to take the time to do this thoroughly.

It has been a substantial undertaking. The NREL technical team spent over 2,000 person-hours on this work, in addition to time from our NDA and procurement teams, our partners at LBNL and our TAG members who helped with outreach, and time at the various potential data providers' organizations. Based on our records, we estimate that a typical successful outreach consisted of three phone calls and 24 emails, from initial outreach through having all data and metadata in hand and checks conducted to ensure we understood the data sets. This time accounting includes some amount of time to set up data infrastructure on our end, but the majority of it is time to organize and conduct the outreach and discussions. Unsuccessful and incomplete outreach efforts are not included and range from very short to just as long or longer. We were at times tracking over 200 potential data sources, with as many as 75 in ongoing conversations simultaneously.

Several themes emerged during the outreach and discussion process. We received a wide range of responses when we broached the subject of data sharing. One common—though not universal—theme was uncertainty regarding what data an organization had, what shape it was in, and the effort that would be needed to extract and compile it. Another was the difficulty of securing permissions for data sharing, even when our project was within formal use guidelines. Several states and other jurisdictions have recently passed new regulations covering energy data sharing or passed these during our outreach period. Examples include the Seattle City Council (ACLU 2018) and the State of California (CA DOJ 2018), MissionData provides an overview of the policies in place in this space (Ay, Kier, and King, n.d.). Some organizations—generally investor-owned utilities (IOUs)—told us they never provide data to third parties unless it is required by their regulators. Others were excited to participate—they saw this as an opportunity for their data to be used as it was intended to be, in some cases for the first time. Most were somewhere in between.

For many utilities, by the time we were having conversation, they were at least tentatively on board with pursuing permission to share anonymized AMI data for the project. But the commercial metadata needs, which required us to access commercial customer addresses in most cases, were a roadblock for many utilities. At the least, this caused an extra step in the generally weeks- to months-long road to secure permissions to access the data. Being able to do the metadata pairing in the utility's own data environment allowed us access to several data sets that we would not have had otherwise. In some cases, we set up meetings between an NREL lawyer and a lawyer at the utility to work out the details so that we could move forward. All of this added time to the process. The mean elapsed time between initial outreach and having all expected data and metadata in hand was 210 days for AMI data (range: 62 to 353, n = 3), and 181 days for contracted commercial end-use data (range: 121 to 295, n = 9). Commercial end-use data that did not require a contract (e.g., because it came from a Building America partner or a national lab, and was therefore no cost to our project) averaged 118 days (range: 60 to 266, n = 7). These values represent only the AMI data sets, contracted commercial submeter data, and



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

substantial no-cost submeter data that were procured through outreach conducted by our project team and that are fully in hand as of June 18; because a number of outreach efforts are ongoing, the final statistics will likely be longer.

Three types of organizations seemed most likely to ultimately provide data: those who saw a direct value to themselves from what we would do with the data, those who had a strong drive or mission to support energy research, and, in a few cases within the commercial end-use data outreach, those who saw their data as a business asset and its sale as a business venture. In most organizations, we were only successful thanks to having an advocate on the inside who believed in the value of our project and wanted to see their organization's data used for it.

Among utilities, we generally found co-ops and municipal utilities were more likely to participate than IOUs. Their regulatory situation was simpler, and the people who stood most to benefit from our project results had more direct access to the relevant decision makers. The IOUs who did participate generally did so as part of a larger collaboration with NREL.

On the commercial end-use data side, the obstacles were in some cases quite different. Initial estimates of what data was available were often inaccurate. It could be difficult for companies to allocate appropriate resources to get us the data. And they had their own permissions issues. In a few cases, they either were the real estate owner, or their contracts indicated that they owned the data, so they were able to make the decision themselves whether to give our team access. In other cases, they were comfortable sharing anonymized data within the bounds of their agreements. In the remaining cases, they had to get permission from their clients, and in some cases, additionally, the owners of the buildings being monitored. This slowed down and sometimes halted the process.

Another hurdle was that this type of data sharing was unprecedented with most of the organizations we spoke to. This often meant additional time was required to label and convert the data into a format that was transferable and understandable by others. We had several cases of organizations building data export tools into their software to share it with us—this was not a capability the software was built to have.

As we are in the process now of getting data in hand, we are still discovering surprises in the data we receive. Sometimes end uses are included that were not originally mentioned. Sometimes there are fewer buildings' data than expected, or a complete lack of units or other necessary descriptors. Additionally, each data set comes in its own format with its own fields and its own definitions. It is critical for us to understand what each data set does and does not cover. Our process in the initial 2-week quality check window helps us understand what we have and where the gaps are, both within each data set and across the body of the data we have collected. Because we are pursuing large quantities of data and not restricting ourselves to highly curated data sets, we expect to be dealing with data gaps, spurious data, and other typical anomalies found in raw data. These are known unknowns. The larger challenges are the unknown unknowns, such as which 10% of a population is absent from a data set that contains AMI for 90% of a utility territory, or how buildings were selected for a lighting study, both of which can introduce bias. Working with data that was not collected for the specific purpose it is being used for has inherently high risk of unknown unknowns.

Many of the transformations currently underway in the energy industry would benefit from access to data such as what we worked to collect. The following recommendations could go a long way toward overcoming barriers and benefitting many stakeholders. One recommendation



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

would be an industry standard for data labeling of building energy consumption and operation data that defines terms, similar to Project Haystack for IoT and BAS data (Project Haystack n.d.). This would make it easier to work with data across sources. Another recommendation is to lower barriers to access anonymous data. More utilities could consider an approach similar to ComEd's Anonymous Data Service, making their AMI available for everyone. Or at least available to some, as the California Public Utilities Commission's Energy Data Request Program for universities (PG&E n.d.). More nonutility data holders could give their customers the option to allow researchers to access their data, as is done in ecobee's Donate Your Data program (ecobee n.d.).

An alternate approach is to enable researchers to query data remotely to compute aggregated, nonpersonally identifiable information from the data rather than the querying the data locally after acquiring it. This requires a well-organized, well-labeled, well-maintained data set, and can include differential privacy mechanisms to eliminate the risk re-identification through a linkage attack. The team is aware of at least one such endeavor underway, funded by DOE and led by Recurve and NREL, as detailed in the forthcoming ACEEE Summer Study paper *Differential Privacy for Expanding Access to Building Energy Data* (Young, Paré, and Bergmann 2020).

In the longer term, we would like insights from AMI data to be available to the entire electricity research and operation industry, for the entire country. One avenue for this could be if Residential Energy Consumption Survey or Commercial Buildings Energy Consumption Survey or other EIA efforts could include statistics taken from AMI data rather than strictly survey information.

## Conclusions

There are dozens, if not hundreds, of organizations collecting and storing building energy or operation data—for their own business purposes. They are understandably wary of who they provide data to and for what purposes. Yet this data, collectively, has the ability to make our industry's work of transforming the nation's energy system easier and cheaper. Over the past year, we have engaged in a process to see how robust of a data set we could put together for our specific project's needs, focusing on commercial end-use building and whole-building whole-population data. With everything going for us, as of June 18 we have in hand at least one end use of data from 10,268 commercial buildings and 1,000+ nominally unbiased AMI meters from 6 utilities after over 2,000 person-hours of NREL technical team work and hundreds more from TAG members, REEOs, NDA personnel, and others. The effort continues.

As new regulations clarify who should have access to what data under what circumstances, we look toward to a future where industry guidelines are available for data standardization across organizations and direct data access is not required to gain the insights the data can provide.

## References

ACEEE. 2019. "Data Access Summary." 2019. https://database.aceee.org/state/data-access.

ACLU. 2018. "Seattle City Council Adopts Nation's Strongest Law to Protect Utility Customer Personal Data." ACLU of Washington. August 6, 2018. https://www.aclu-



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

wa.org/news/seattle-city-council-adopts-nation%E2%80%99s-strongest-law-protect-utility-customer-personal-data.
Ay, Michael Murr, Laur A Kier, and Bob King. n.d. "Unlocking Innovation with Smart Policy," 11.
Baechler, Michael C., Jennifer Williamson, Theresa Gilbride, Pam Cole, Marye Hefty, and Pat M. Love. 2010. "Building America Best Practices Series, Volume 7.1 - High-Performance Home Technologies: Guide to Determining Climate Regions by County." PNNL-17211.
Barth-Jones, Daniel C. 2012. "The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now," 19.
Bianchi, Carlo, and Amanda D. Smith. 2019. "Localized Actual Meteorological Year File Creator (LAF): A Tool for Using Locally Observed Weather Data in Building Energy Simulations." *SoftwareX* 10 (July): 100299. https://doi.org/10.1016/j.softx.2019.100299.
CA DOJ. 2018. "California Consumer Privacy Act (CCPA)." State of California - Department of Justice - Office of the Attorney General. October 15, 2018. https://oag.ca.gov/privacy/ccpa.
ComEd. n.d. "Anonymous Data Service." Accessed March 10, 2020. https://www.comed.com/SmartEnergy/InnovationTechnology/Pages/AnonymousDataService.aspx.
"CoStar." n.d. CoStar. Accessed March 16, 2020. https://www.costar.com.
ecobee. n.d. "Donate Your Data Smart Wi-Fi Thermostats." Accessed March 17, 2020. https://www.ecobee.com/donate-your-data/.
EIA. 2019. "Annual Electric Power Industry Report, Form EIA-861 Detailed Data Files Final 2018 Data." U.S. Energy Information Administration. October 1, 2019. https://www.eia.gov/electricity/data/eia861/.
Fenaughty, Karen, Danny Parker, and Eric Martin. 2017. "Phased Deep Retrofit Project: Real-Time Measurement of Energy End-Uses and Retrofit Opportunities." In *2017 International Energy Program Evaluation Conference*, 14. Baltimore, MD: Florida Solar Energy Center.
Frick, Natalie Mims, Eric Wilson, Janet Reyna, Andrew Parker, Elaina Present, Janghyun Kim, and Tianzhen Hong. 2019. "End-Use Load Profiles for the U.S. Building Stock: Market Needs, Use Cases, and Data Gaps."
McKibbin, Anne. 2014. "Unleashing the Power of Big Data on Efficiency? Not So Fast." In *ACEEE Summer Study Proceedings*, 11. https://doi.org/10.13140/2.1.4786.6240.
Narayanan, Arvind, and Vitaly Shmatikov. 2008. "Robust De-Anonymization of Large Sparse Datasets." In *IEEE Symposium on Security and Privacy*, 15. https://doi.org/10.1109/SP.2008.33.
Navigant. 2018. "Res 1 Baseline Load Shape Study." http://ma-eeac.org/wordpress/wp-content/uploads/RES-1-FINAL-Comprehensive-Report-2018-07-27.pdf.
NEEA. 2019. "Residential Building Stock Assessments." Northwest Energy Efficiency Alliance (NEEA). April 4, 2019. https://neea.org/data/residential-building-stock-assessment.
NEEP. 2017. "Advanced Metering Infrastructure: Utility Trends and Cost-Benefit Analyses in the NEEP Region."



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

NREL. n.d. "ResStock Analysis Tool." Accessed March 16, 2020. https://www.nrel.gov/buildings/resstock.html.

Parks Associates. 2019. "23% of US Broadband Households Reporting Intentions to Purchase a Smart Thermostat in 2019." October 22, 2019. http://www.parksassociates.com/blog/article/ses2020-2.

Pecan Street. n.d. "Dataport." Pecan Street Inc. Accessed March 10, 2020. https://www.pecanstreet.org/dataport/.

PG&E. n.d. "Energy Data Request Portal." Accessed March 12, 2020. https://pge-energydatarequest.com/.

Project Haystack. n.d. "Home." Accessed March 10, 2020. https://project-haystack.org/.

USPS Postal Explorer. n.d. "C1 Street Suffix Abbreviations | Postal Explorer." Postal Addressing Standards Appendix C1 Street Suffix Abbreviations. Accessed June 19, 2020. https://pe.usps.com/text/pub28/28apc_002.htm.

Young, McGee, Marc-Antoine Paré, and Harry Bergmann. 2020. "Differential Privacy for Expanding Access to Building Energy Data." In *ACEEE Summer Study on Energy Efficiency in Buildings Proceedings*. VIRTUAL.



This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.