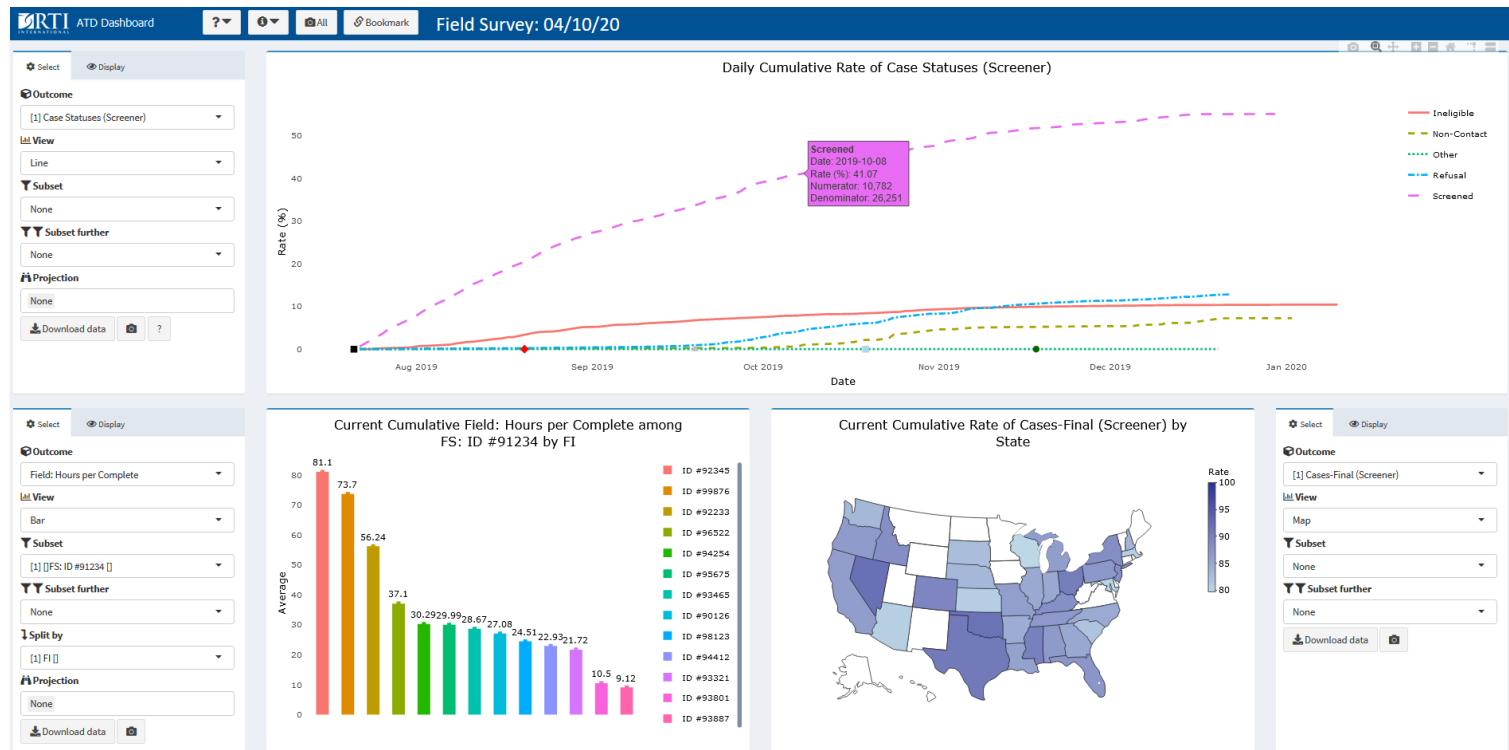


An Interactive Data Visualization Platform for Survey Data Monitoring and Paradata Exploration

Joe Murphy¹, Michael Duprey¹, Rob Chew¹, Rebecca Powell¹, Katie Lewis²

FedCASIC
April 14, 2021

¹RTI International
²U.S. Energy Information Administration



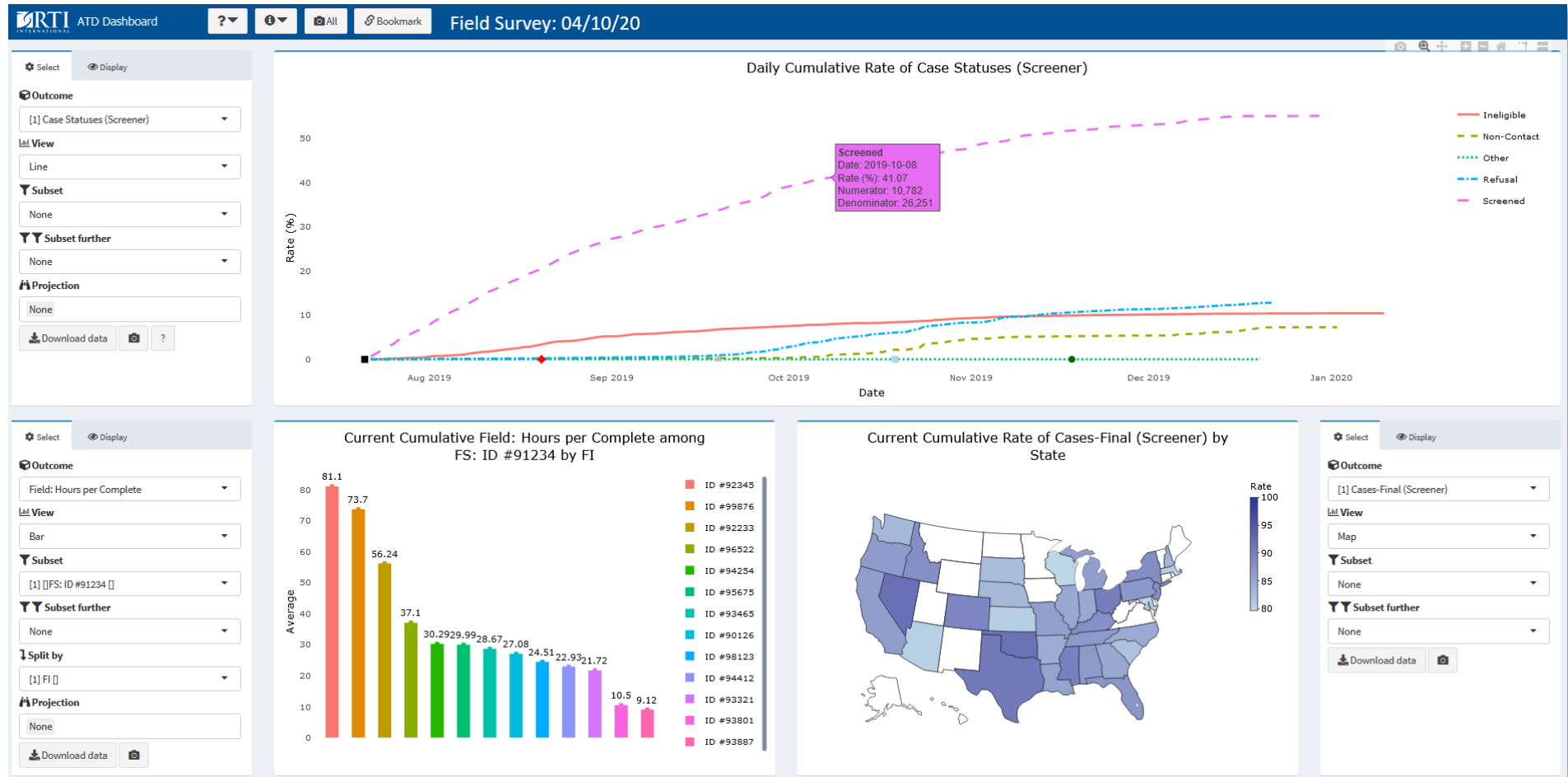
The challenge

- In the age of paradata, the amount of information available to inform decisions during data collection can be overwhelming.
- Adaptive, responsive, or tailored designs require the survey team to monitor critical-to-quality indicators to minimize total error across data sources.
- Monitoring traditionally involves disparate reports, systems, and indicators. This results in an ineffective use of resources (labor) from manually piecing the puzzle together.
- Further, the traditional tools do not facilitate real time monitoring and adaptive strategies.

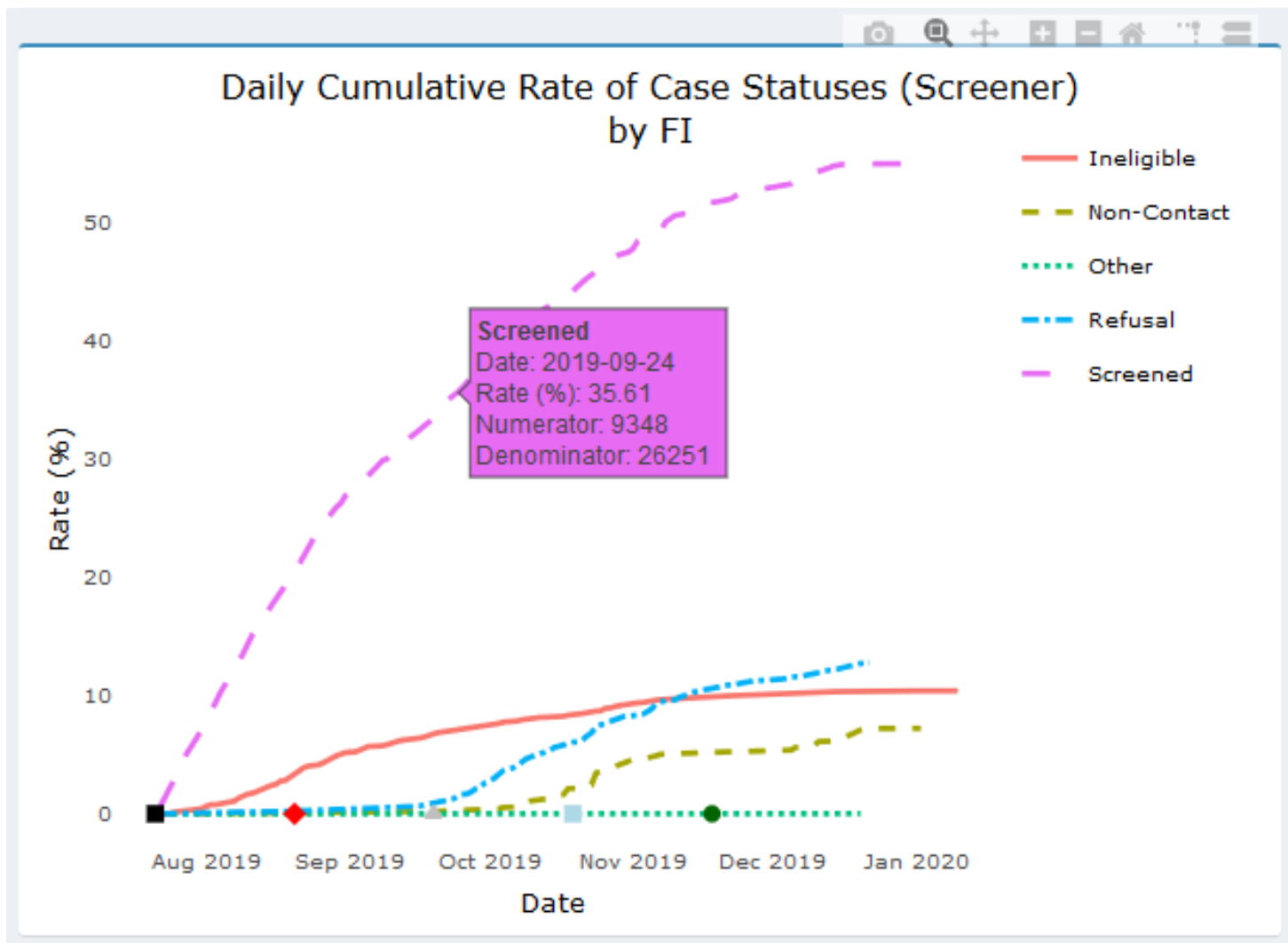
The solution

- We designed the Adaptive Total Design (ATD) Dashboard to monitor and visualize data from multiple sources to track experimental, multimode, and longitudinal survey designs in near-real time.
- Data inputs may be from various systems and may exist at multiple units of analysis.
- A data taxonomy allows only logical instantiations.
- By employing an extensible app framework for R (Shiny) the dashboard standardizes visualizations and reports.
- The dashboard is hosted on a secure, password-protected website and does not require users have any licenses or software beyond a browser. No programming skills are required to use the dashboard.

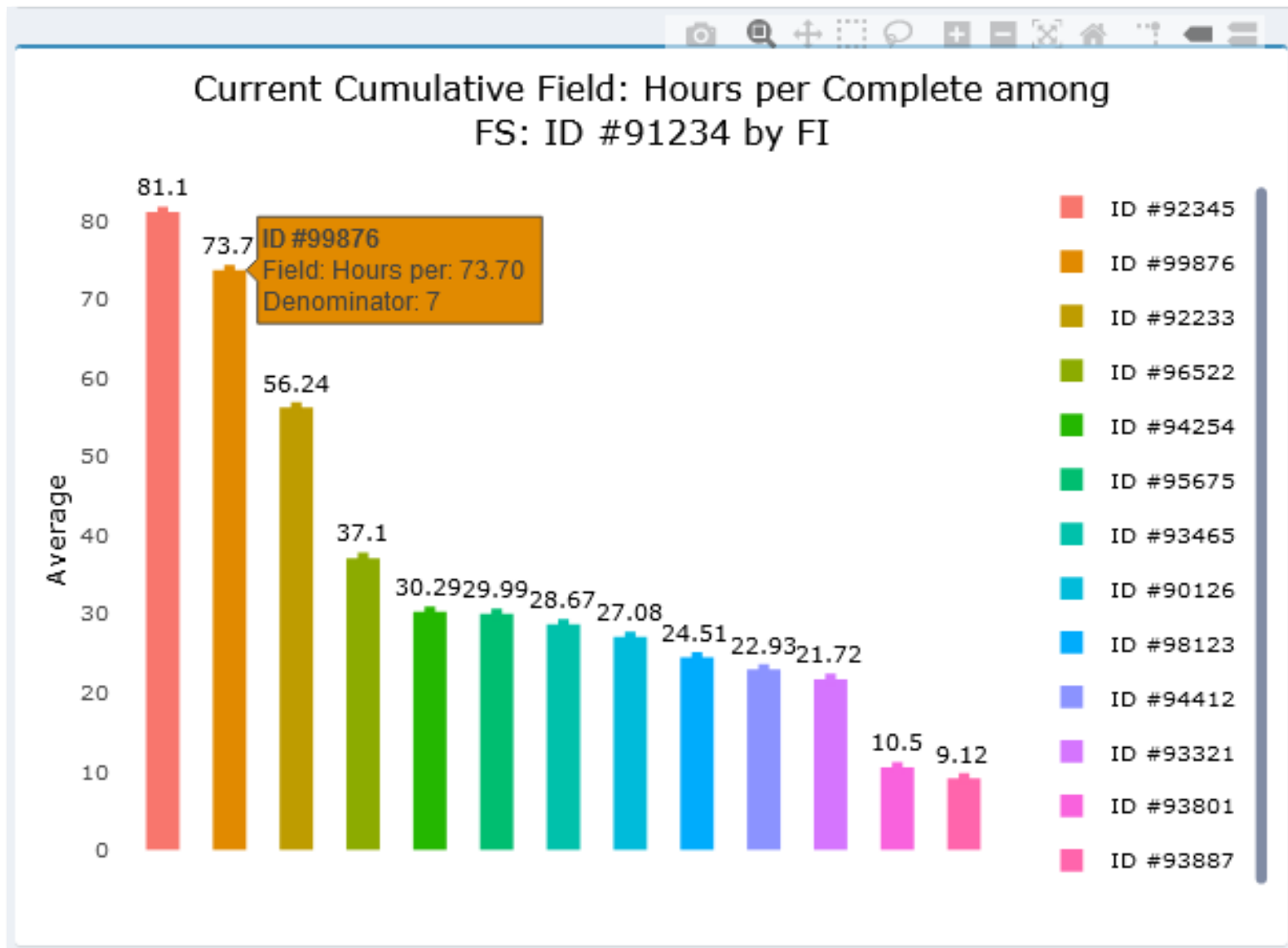
Adaptive Total Design (ATD) Dashboard



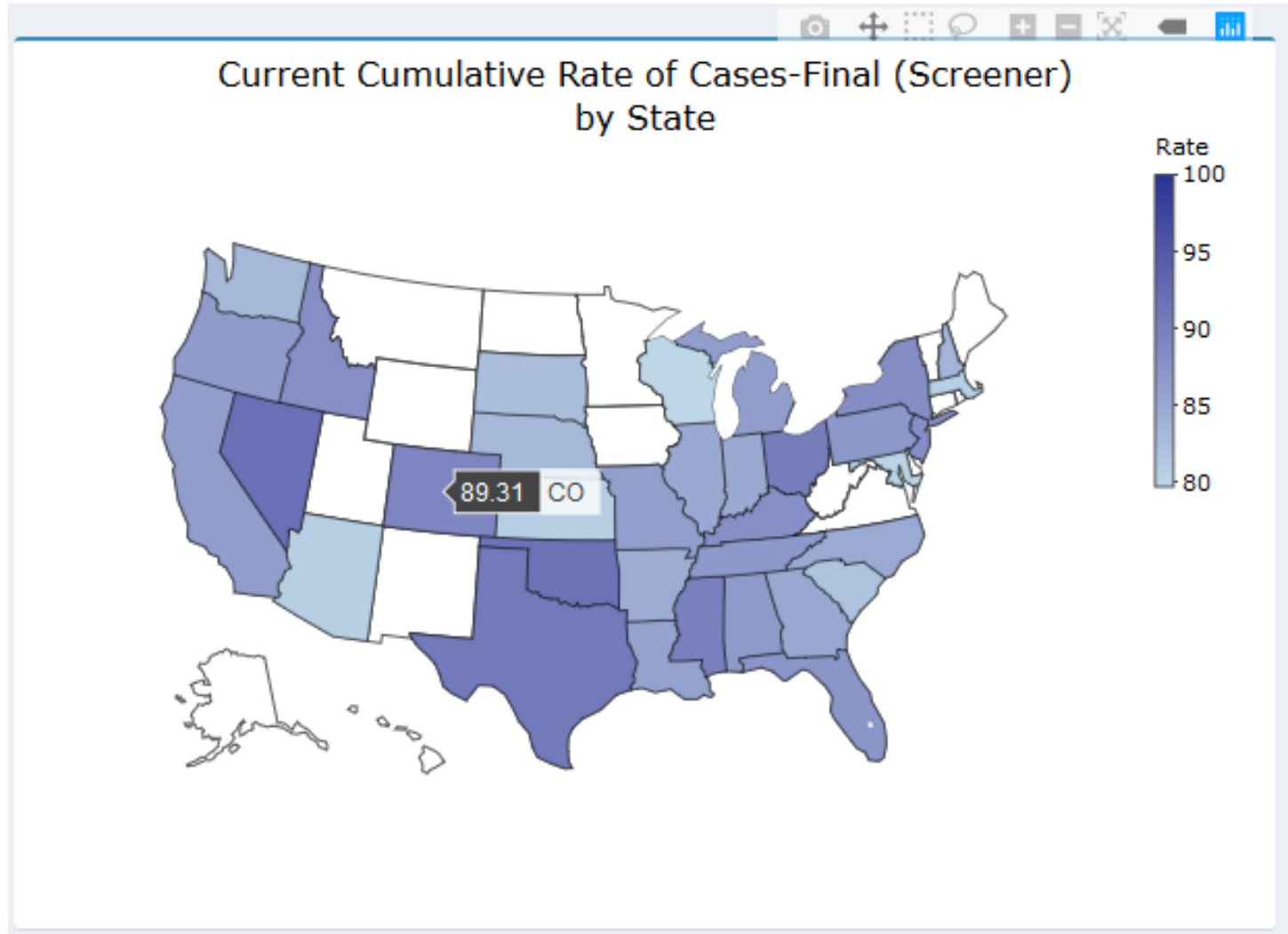
ATD Dashboard: Line view



ATD Dashboard: Bar view



ATD Dashboard: Map view



ATD Dashboard: Table view

Daily Cumulative Rate of Cases-Final (Screener)
by State

| Date | Rate (%) | State |
|------------|----------|-------|
| 2019-07-22 | 1.76 | AL |
| 2019-07-23 | 3.04 | AL |
| 2019-07-24 | 4.81 | AL |
| 2019-07-25 | 6.25 | AL |
| 2019-07-26 | 7.69 | AL |
| 2019-07-27 | 10.10 | AL |
| 2019-07-28 | 10.58 | AL |
| 2019-07-29 | 11.06 | AL |
| 2019-07-30 | 12.18 | AL |
| 2019-07-31 | 13.14 | AL |
| 2019-08-01 | 15.22 | AL |
| 2019-08-02 | 16.99 | AL |
| 2019-08-03 | 19.07 | AL |

Many standard metrics with options to customize

- For the 2020 Residential Energy Consumption Survey (RECS):

Outcomes

Survey Submissions

Survey Duration

Response Mode

Device Type

Survey Language

Responses to Survey Items

Breakoffs

Ineligibles

Undeliverables

Cost per Case

Splits

Survey Frame Variables

Region / Division / State

Data Collection Phase

Experimental Treatment Groups

Comparisons

Prior Round Response Rates

Expected Response by Geography

External Source Distributions

Input data structure

- Input data in **flat-file format** (csv file)
- Long (stacked) data
- Metadata information included in “**blocked**” sections at file head
 - Block for **critical dates** (axis markers)
 - Block for **data taxonomy**
- Format consist of **four** columns
 - ID: **id**
 - Variable name: **var**
 - Value: **val**
 - Date/time-stamp: **dt**

| id | var | val | dt |
|-----|--------------------|---------------|-----|
| ... | ... | ... | ... |
| | C:All Events | event_all_ind | |
| | C:All Cases | samp_all_ind | |
| | C:All Interviewers | intd_all_ind | |
| | C:CATI Calls | event_sub_ind | |
| | C:All Days | day_all_ind | |
| | C:Completed Cases | samp_sub_ind | |

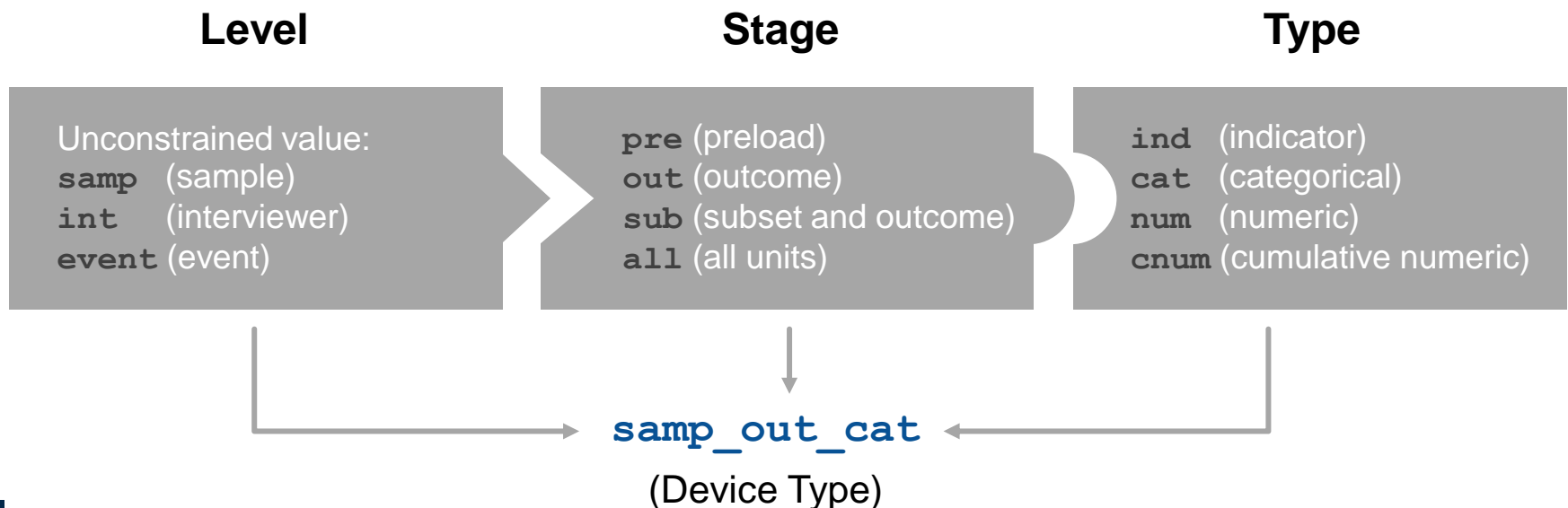
| id | var | val | dt |
|-----|-----------------|-----|------------|
| 1 | All Cases | 1 | 03-13-2017 |
| 2 | All Cases | 1 | 03-13-2017 |
| 3 | All Cases | 1 | 03-13-2017 |
| 4 | All Cases | 1 | 03-13-2017 |
| 5 | All Cases | 1 | 03-13-2017 |
| 6 | All Cases | 1 | 03-13-2017 |
| ... | ... | ... | ... |
| 2 | Completed Cases | 1 | 03-18-2017 |
| 4 | Completed Cases | 1 | 04-07-2017 |
| 5 | Completed Cases | 1 | 03-24-2017 |
| ... | ... | ... | ... |

Assigning roles to variables

Data taxonomy

- Provides **compact attribute vocabulary**, easy to input into flat file
- Provides mechanism to limit UI selection options to **only logical combinations** and invoke **variable mathematical operations**
- **Mapped** to data in csv file

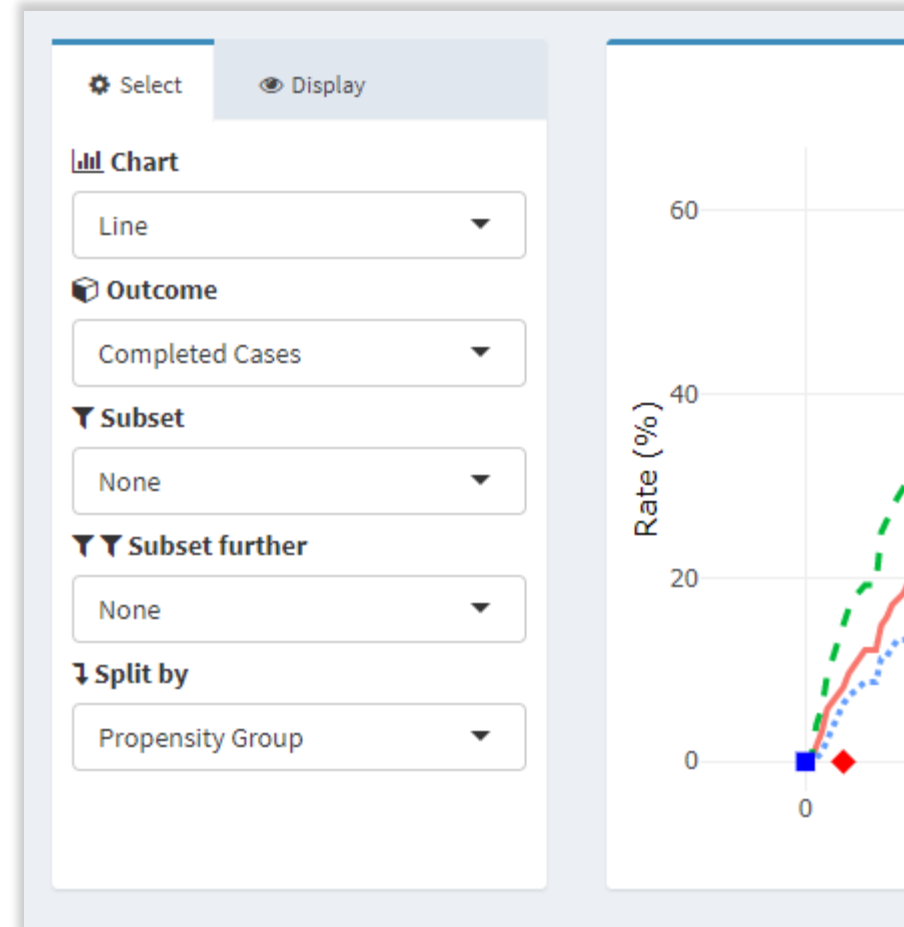
Type-class system



Data taxonomy

UI drop-down options reactively updated to allow logical combinations, conditioned on prior selections

- Choose **plot type** →
- Select primary **outcome of interest** →
`samp_out_cat`
- Filter to **subset** →
(e.g., `samp_sub_ind`)
- Filter to (second) **subset** →
- Separate** outcome into categories →
`samp_out_cat`



Generating and automating

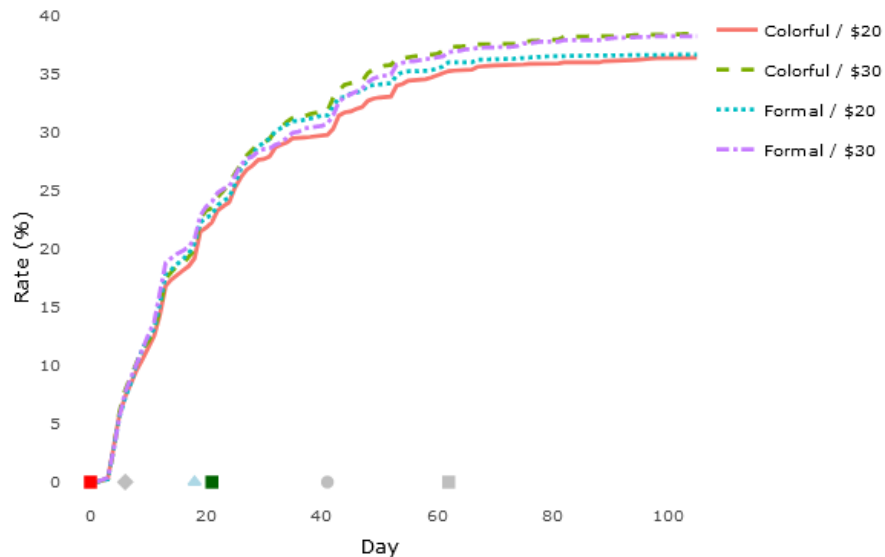
- Set up the folder structure for logs and output
- Generate entries for outcomes, titles, categories, etc. and store in SQL Server database
- Scheduled SAS programs aggregate and generate input file for each project daily
- Linux server copies input file to setup the visualization
- Server runs R Shiny and served using R Shiny Server & NGINX
- All dashboards run off the same version of codebase
- Customization determined through input file content
- Only RTI employees can access internal dashboards
- Secure client dashboards hosted on externally-facing web server

RECS Case Study (1)

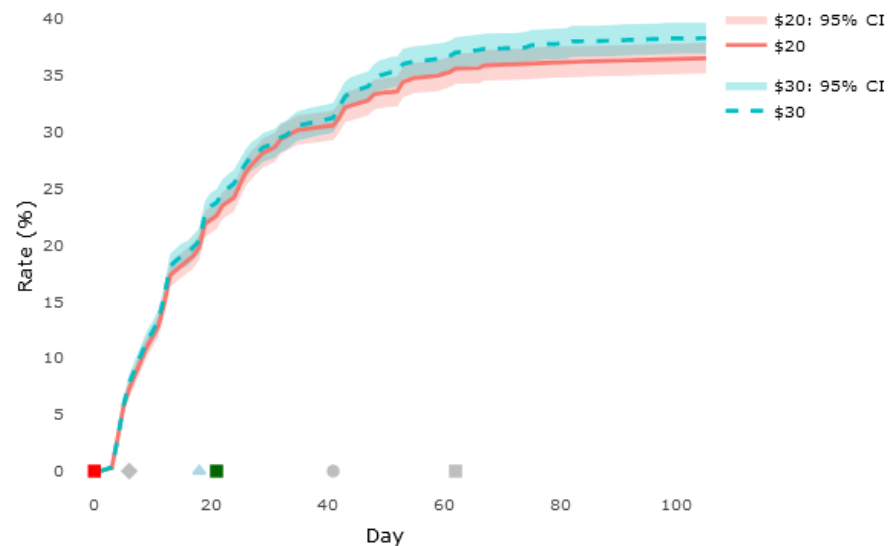
The 2020 RECS included two experiments to help identify the best web/mail protocol in terms of data quality and cost.

- The first experiment was on the promised incentive amount. In the control condition, we offered \$10 for completion (a.k.a. submission or response) by paper and \$20 for completion by web (since we want to push as many to the web as possible). The treatment condition offered \$10 for paper completion and \$30 for web to determine if the extra expense equated to a significant increase in response.
- The second experiment was on the format of the advance postcard. We tested a formal appearance vs. a colorful design.
- The experiment was done during Phase 1 of data collection, which included 20% of the total initial sample. We needed to decide around Day 40 which treatments were best and to include for all cases in Phase 2 (the remaining 80% of the initial sample).
- The postcard format generally did not have an impact on response and while there was a small difference by incentive amount, the difference was not statistically significant. The \$20 / colorful design was chosen for Phase 2.

Daily Cumulative Rate of Submissions among Phase: 1
by Experiment - Postcard / Web Incentive



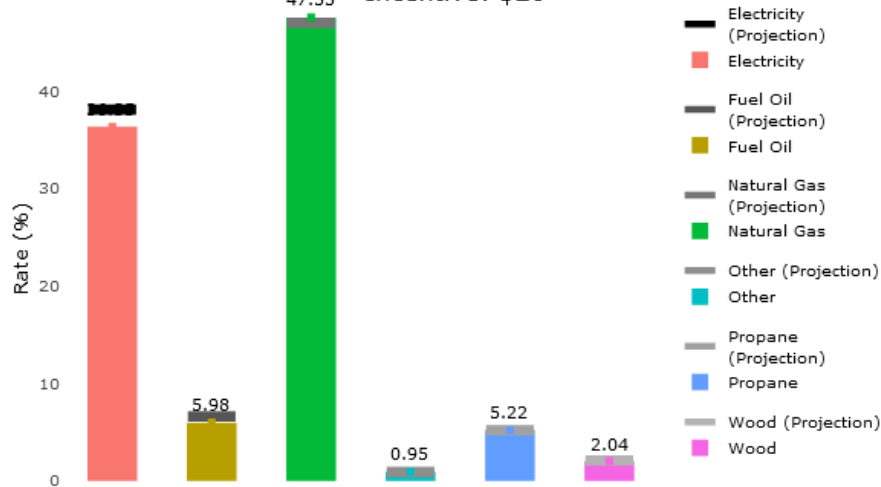
Daily Cumulative Rate of Submissions by Experiment -
Web Incentive



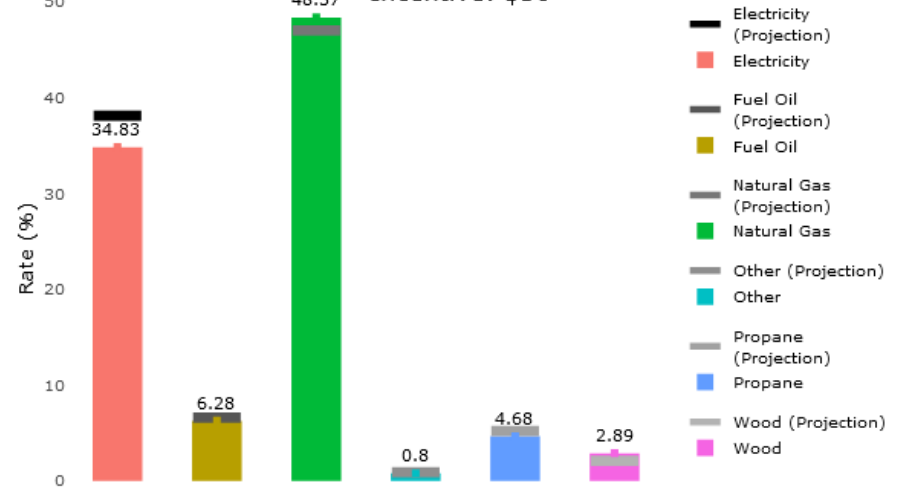
RECS Case Study (2)

Response rate was not the only consideration. We also compared data for several survey outcomes to benchmarks like the American Community Survey. Generally, results were similar by incentive treatment. For example, the projections below are the ACS values for main heating fuel used for comparison with the \$20 and \$30 conditions.

Current Cumulative Rate of Heating Fuel among Completes: Non-missing and Experiment - Web
Incentive: \$20



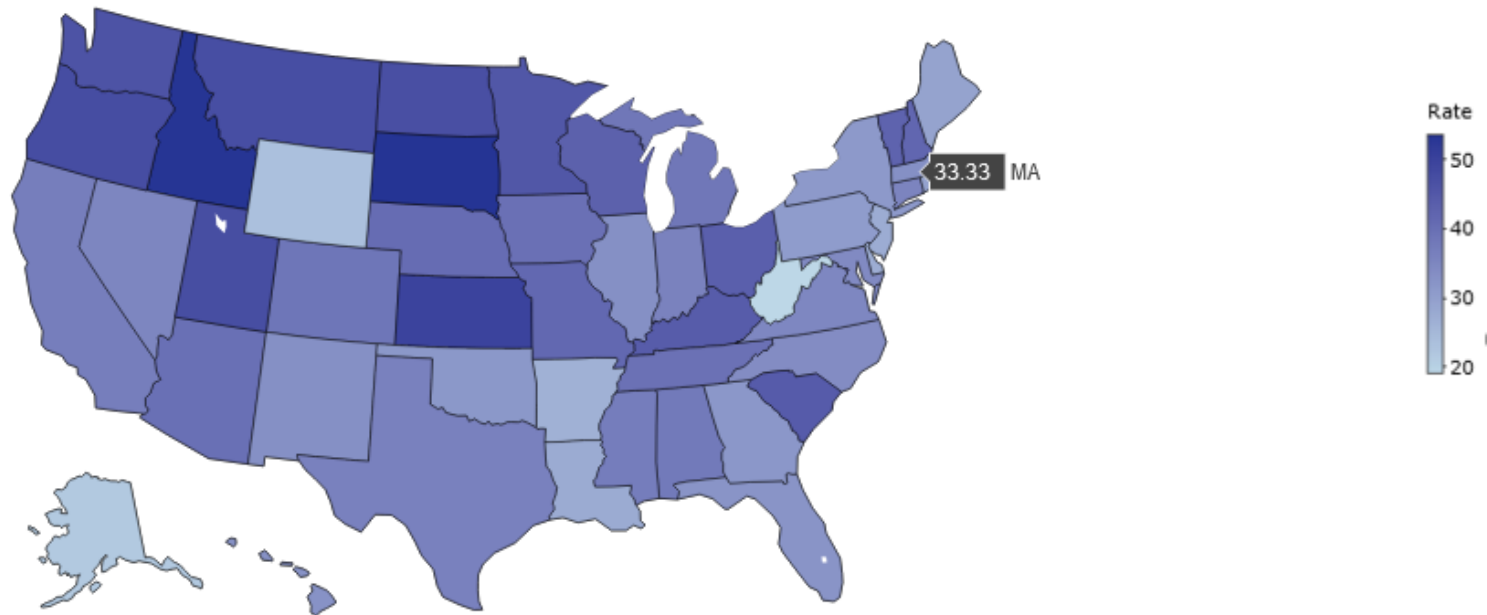
Current Cumulative Rate of Heating Fuel among Completes: Non-missing and Experiment - Web
Incentive: \$30



RECS Case Study (3)

The dashboard provides a map view to examine data at subnational levels. For example, we see here response within each state for the first phase of 2020 RECS data collection. We can use (or have used) these data to inform decisions about subsequent phases in our responsive design.

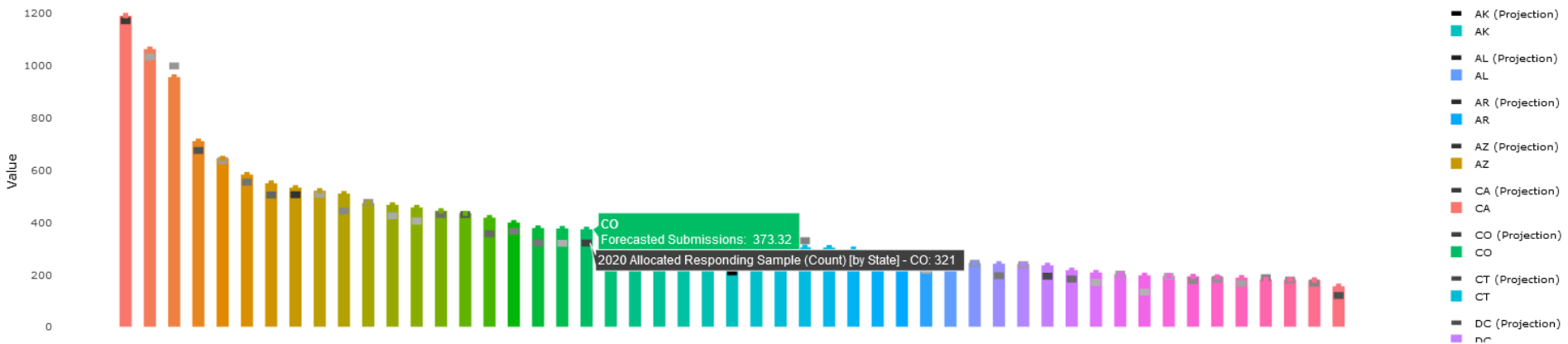
Current Cumulative Rate of Submissions among Phase: 1 and Experiment - Web Incentive: \$20 by State



RECS Case Study (4)

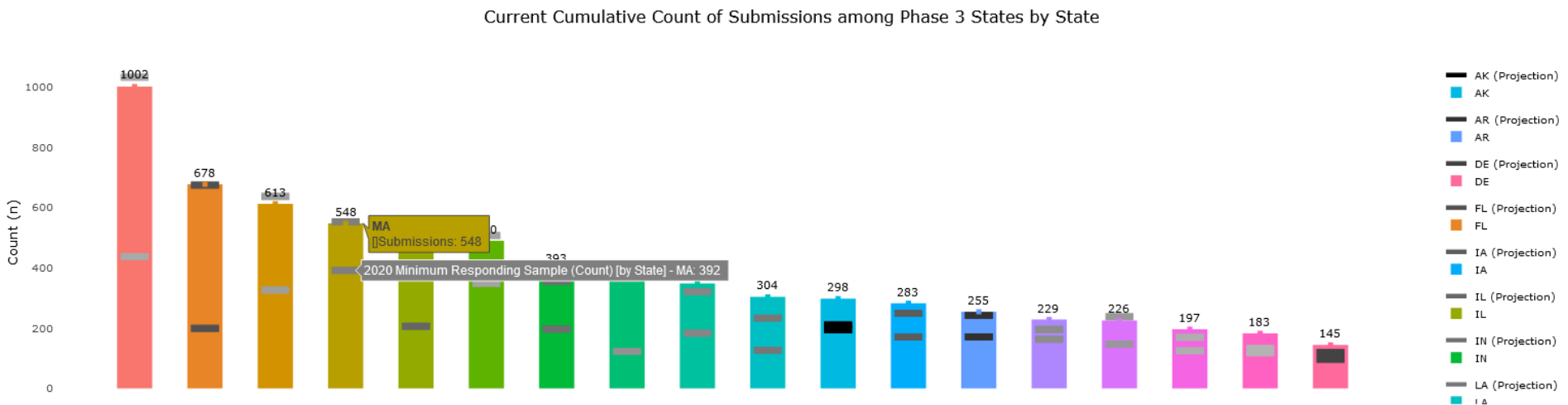
We also used the dashboard to compare forecasted completes to the target allocated for each state. The forecast was updated daily based on actuals to date and expected future counts of complete based on previous data collections. Tracking the forecast against targets led us to consider the approach for Phase 2 and whether the protocol would yield the required number of completed interviews for certain states. We determined that additional sample would be required in some states to exceed the minimum targets needed for precise estimates and approach the desired (allocated) number of respondents for other analytic purposes.

Current Cumulative Total Forecasted Submissions by State



RECS Case Study (5)

Sample was added in several states*, concurrent with Phase 2, to assure adequate respondent counts in particular states. The chart below reflects the count of submissions by state as of Day 70 in the Phase 2 collection period. We see the counts for each state exceeding the minimum target (lower line) and approaching or exceeding the allocated target (upper line). Monitoring and adjusting in reaction to the early phases of data collection allowed us to field the best protocol in terms of quality and cost while meeting targets across domains and geographies.



*Additional sample referred to as “Phase 3” in the chart above. Phase 3 was concurrent with Phase 2.

Summary and future directions

- The ATD Dashboard has allowed us to monitor important survey trends and make decisions to optimize later data collection periods.
- It represents a vast improvement over producing and reviewing pages of static tables where important information may go undetected.
- We hope to expand the utility of the dashboard by building in automatic alerts for the users signifying when actuals may be deviating from expectation and intervention is needed.
- We are looking to leverage the extensive R libraries available to introduce other functionality, such as county level and other maps.
- We also seek to enable hourly or more frequent updates to bring the dashboard even closer to a real-time reflection of data collection progress.

Thank you!

Contact

For dashboard inquires:

Joe Murphy

jmurphy@rti.org

For RECS-specific inquires:

Katie Lewis

katie.lewis@eia.gov