A novel data science gateway for visual exploration and analysis of surveys and image collections, with examples from analysis of Facebook ads from Russia

Ilya Zaslavsky
San Diego Supercomputer Center, University of California San Diego

Efficient and user-friendly analysis of surveys and image collections is indispensable in many disciplinary domains. While the precise meaning of "survey" varies by domain, it typically involves systematic collection and organization of observations and human appraisal of the collected data to develop a "lay of the land" for further in-depth analysis. Besides the social sciences where questionnaire surveys are common, biologists carry out ecological and biodiversity surveys, geoscientists conduct mineral and fossil surveys, astronomers assemble sky surveys, while hydrographic or geological surveying has been a common activity of teams of explorers of Earth frontiers. Survey data often present a mixture of structured and unstructured information, including measurement results organized in a known data model, labels, notes, narratives, annotations, links to relevant resources, images, and other media types. Such surveys are often conducted by research teams and require collaborative exploration and management of the collected data. An additional common requirement is the ability to visually examine the data using various aggregated data views, moving from exploring general patterns to individual cases, transcending spatial and temporal scales, connecting with other sources that may provide relevant information, and gaining insights from other researchers. Exploration of survey data and media collections typically requires specialized desktop software and expects that users are well-versed in statistical analysis techniques. Steep learning curves and other limitations of existing tools for survey data exploration complicate teaching statistical methods to undergraduate students, especially in the social sciences where students often come to statistics classes with little math background are anxious about the use of quantitative approaches.

Addressing these challenges, we developed a novel system for online collaborative analysis of surveys and image datasets and tested it in several research projects and undergraduate research methods courses. We also integrated the system with Jupyter notebooks, creating a data science gateway for survey and collection analysts, focused on analytical tasks in social sciences and humanities, in particular. In this gateway, visual data exploration is supported and reinforced by additional analytical functionality, as results of computations and models implemented as Jupyter notebooks are seamlessly reintroduced into the visual data exploration interface. We argue that such "deeper" integration of Jupyter notebooks, a data science framework of choice in many research projects and curricula, with online visual analytics is required for efficient exploration of social media collections and similar data.

The system is called Survey Analysis via Visual Exploration (SuAVE); it is available at http://suave.sdsc.edu. The SuAVE platform helps communicate scientific knowledge and develop insights through guided exploration of relationships between different variables while engaging a broad audience in sharing their findings. Working with SuAVE, users can visualize the entire collection or survey content, sort and filter items by any combination of metadata elements, and seamlessly navigate from the big picture to individual items. The platform integrates visual, statistical and cartographic analyses and presents an intuitive framework for exploratory data analysis, enabling faceted browsing, Google Maps-like navigation over a gallery of images or survey respondents, and animating transitions between different data views. It supports survey-specific data types, including multiple-choice and

open-ended questions, questions with multi-value responses, as well as location variables, which enable automated geocoding and mapping. SuAVE allows users to annotate distribution patterns and/or outliers discovered in the data and share such annotations with peers, thus supporting collaborative analysis of surveys by distributed groups of researchers. In addition to presenting several statistical views of the data, such as crosstab and qualitative comparative analysis (QCA) views, and supporting analysis of conditional frequencies in multi-dimensional contingency tables, SuAVE interfaces with the R statistical package. The essential added functionality is the integration of SuAVE with Jupyter notebooks. Users can pass survey parameters to a notebook on a Jupyterhub server to compute additional variables and statistical models, perform image processing, run classification algorithms, do text analysis, access external data resources, and perform other operations. The computed variables are then automatically added to a new version of the survey dataset for visual analysis. This feature opens new data analysis and integration possibilities in fields that require simultaneous exploration of multifaceted data and images.

SuAVE has been already used in several fields, including public opinion polls and electoral surveys, visual arts and humanities, biology and ecology, geosciences and urban management and planning, archaeology and public health, and for publishing ethnographic collections. It has also been used in 3 undergraduate classes at UCSD, which introduced social science students to research methods and data analysis techniques, and to Web-based GIS. Using SuAVE, students learned to evaluate outliers and unusual data patterns and to conduct deviant case analysis in large surveys.

Many SuAVE applications are accessible from the project website. In this presentation, we will demonstrate a recent application, which enables interactive online analysis of Facebook and Instagram advertisements posted by operatives of the Russian Internet Research Agency (IRA) during 2015-2017. In May 2018 the advertisements were publicly released by Democrats on the Permanent Select Committee on Intelligence of the US House of Representatives (https://democrats-intelligence.house.gov/facebook-ads/). National news media have published multiple reviews of these ads. These reviews emphasized that the ads meant to sow discord in the US political system by focusing on critical divisive points in the US political discourse, while at the same time noting contradictory messaging as the ads were created to target different population groups at different times. Additional analysis with SuAVE reveals a number of interesting patterns of the broad and systematic social media campaign. With SuAVE one can analyze the ads in several ways. In particular, one can examine how different target groups were addressed over time, what types of ads were pushed stronger than others (based on their costs, number of impressions, number of clicks), which groups were targeted for different types of ads (based on target interests and geographic location of the audience, and on audiences to be excluded), and the multitude of political and cultural viewpoints of the US electorate impersonated by the operatives. All these factors are included in the SuAVE application that enables joint analysis of advertisement metadata, textual content, and images. Authenticated users can compute additional variables (e.g., the click-through rate, the number of clicks on an ad divided by ad impressions, or the number of times the ad is shown) and add them to analysis, using the SuAVE-Jupyterhub interface. Availability of this dataset via an easy-to-use online visual interface lets the public to analyze the data by themselves and make their own conclusions.