Appendix C. Description of SRHS data-scraping tools and procedures

As part of the SRHS we have explored new strategies to produce new sources of data for use by the City and other entities in developing and evaluating rental housing policy. According to the results of our survey of City data users and conversations with other community organizations and policy entities, several data features would be of high value: temporally-specific, referring to current conditions and suitable for aggregation to specific time periods; geographically granular, providing the ability to examine rent patterns in small areas or aggregate up to very specific larger areas (e.g., a particular neighborhood or urban village); flexible, with the ability to link to other data sources; and transparent, with clear information on individual properties and ongoing assessments of quality.

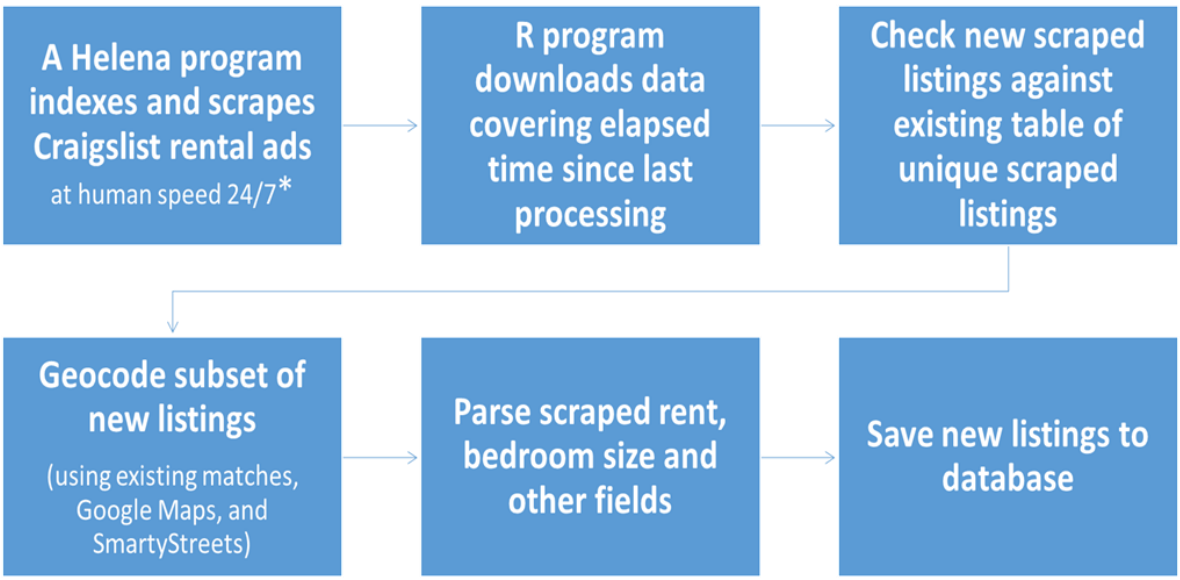
Since February 2017, the SRHS team has developed the infrastructure to compile such a data source. Specifically, we have developed methods to continually scrape data from publically-available for-rent Craigslist advertisements for the Seattle metro area. We have developed an innovative machine-learning tool developed in collaboration with UW's Computer Science department[1] and automated the system to efficiently and continually download and process advertisements, extracting advertised rents and other valuable information (e.g., unit characteristics, move-in costs, and locational information). The system features methods that utilize date-address-rent combinations to differentiate between new and repeated/continued listings. As the system collects more data, the program becomes better trained in how to classify different fields of information (e.g., rent, bedrooms, and square footage) from patterned fields in ads. As summarized in Figure C1, an R-based routine then identifies the locations of the advertised rental units from the spatial information (address or map location) embedded in ads. The program also cleans fields, adds census tract (neighborhood) identifiers, and subsets a Seattle sample and saves a datafile. Data are scraped and compiled on a daily basis, keeping the data as current as possible and allowing us to assess up-to-date trends in rent for units with specific characteristics and across specific neighborhoods. The geographic coverage of the data scraped from Craigslist has been analyzed and we have attached additional building/property characteristics from parcel data and contextual information from the census.

Scraping from online ads offers a potentially important complement to existing sources of data on rental market dynamics. A recent national study shows that rent levels derived from data scraped from Craigslist advertisements have strong geographic coverage relative to that of several other sources of commercial data derived from landlord surveys and

[1] For more information on tool developed for scraping, see Sarah Chasins and Rastislav Bodik, "Skip blocks: reusing execution history to accelerate web scripts." *Proceedings of the ACM on Programming Languages* 1. OOPSLA (2017): 51.

transactions.[2] Our analysis of rental housing dynamics in the Seattle area indicate that about 70% of the over four thousand landlords who responded to our SRHS Landlord Survey reported using Craiglist to advertise their units, and the tool was used by large majorities of landlords offering a wide range of unit types, including accessory-dwelling units, single-family homes, and multiplexes that are often excluded from commercial sources (see Appendix B, Figures 86-91).

Figure C1. Summary of approach to scraping and processing online for-rent advertisements.



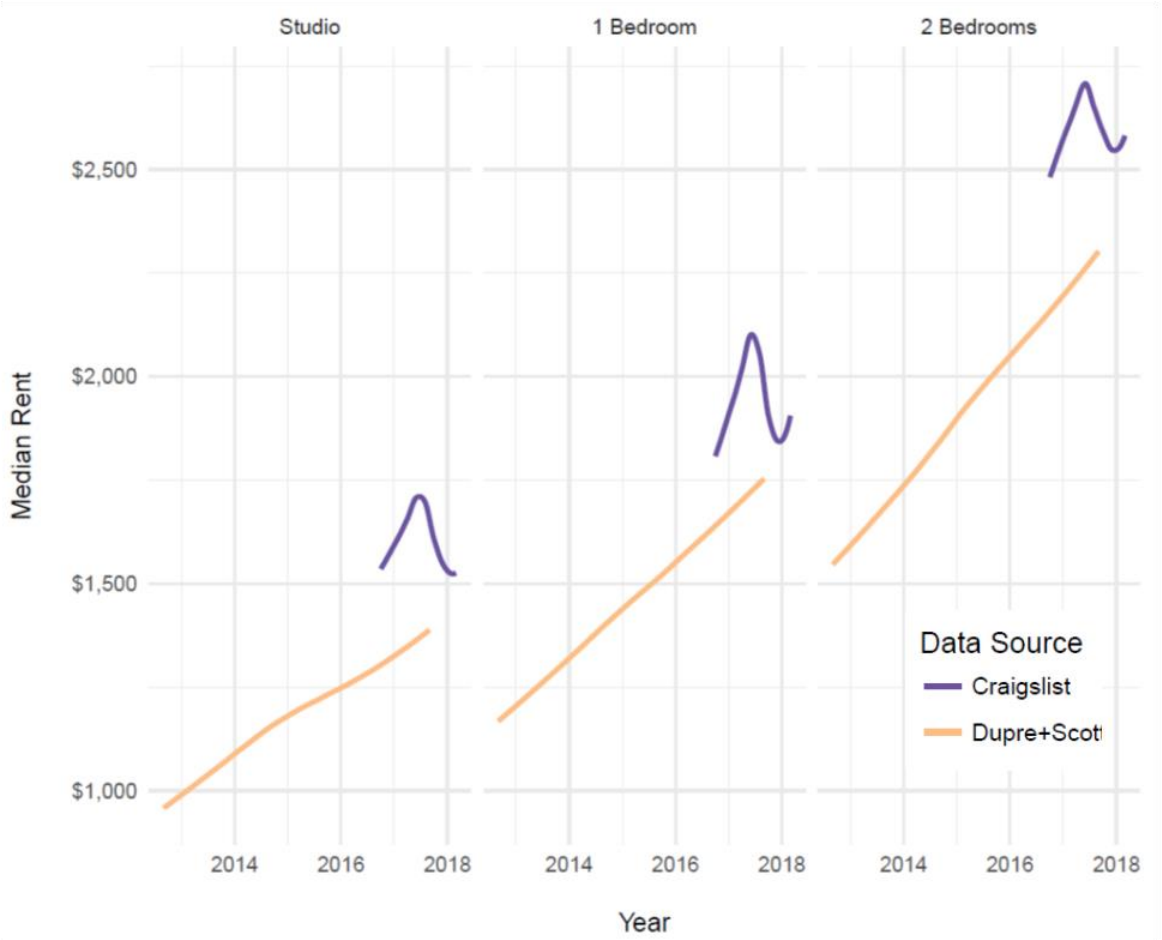
Data scraped from online ads are especially useful for characterizing the housing costs faced by current housing searchers in that they provide information on rents advertised for currently-available units. Depending on the trend in rents, these advertised rents may slightly over- or underestimate the level of contract rents estimated in commercial data. In areas, such as Seattle, with generally increasing rents, advertised rents are likely to be higher than average rents for currently-leased units. In this sense, data from online advertisements can be considered leading indicators of rent trends.

As shown in Figure C2, trends in median rents scraped from online ads are similar to estimates released by Dupre+Scott. However, two deviations are noteworthy. First, the scraped data, with greater temporal specificity, clearly show the seasonal nature of rent trends; rents tend to increase in summer months when mobility rates are relatively high, and decrease in winter. Second, median rents reflected in the scraped data are higher

[2] Boeing, Geoff, and Paul Waddell. "New insights into rental housing markets across the united states: web scraping and analyzing craigslist rental listings." *Journal of Planning Education and Research* 37.4 (2017): 457-476.

than those reflected in the Dupre+Scott data. Again, these higher levels, in part, reflect a contrast between advertised rent and contract rents for currently-leased apartments. However, two other factors are likely at play here as well. First, the Dupre+Scott data are based on self-reports by property owners and managers and may be subject to recall bias or other threats to reliability. Perhaps more importantly, as described earlier, the data scraped from online ads likely reflects a broader, more representative cross-section of rental units of varying types, building characteristics, and locations.

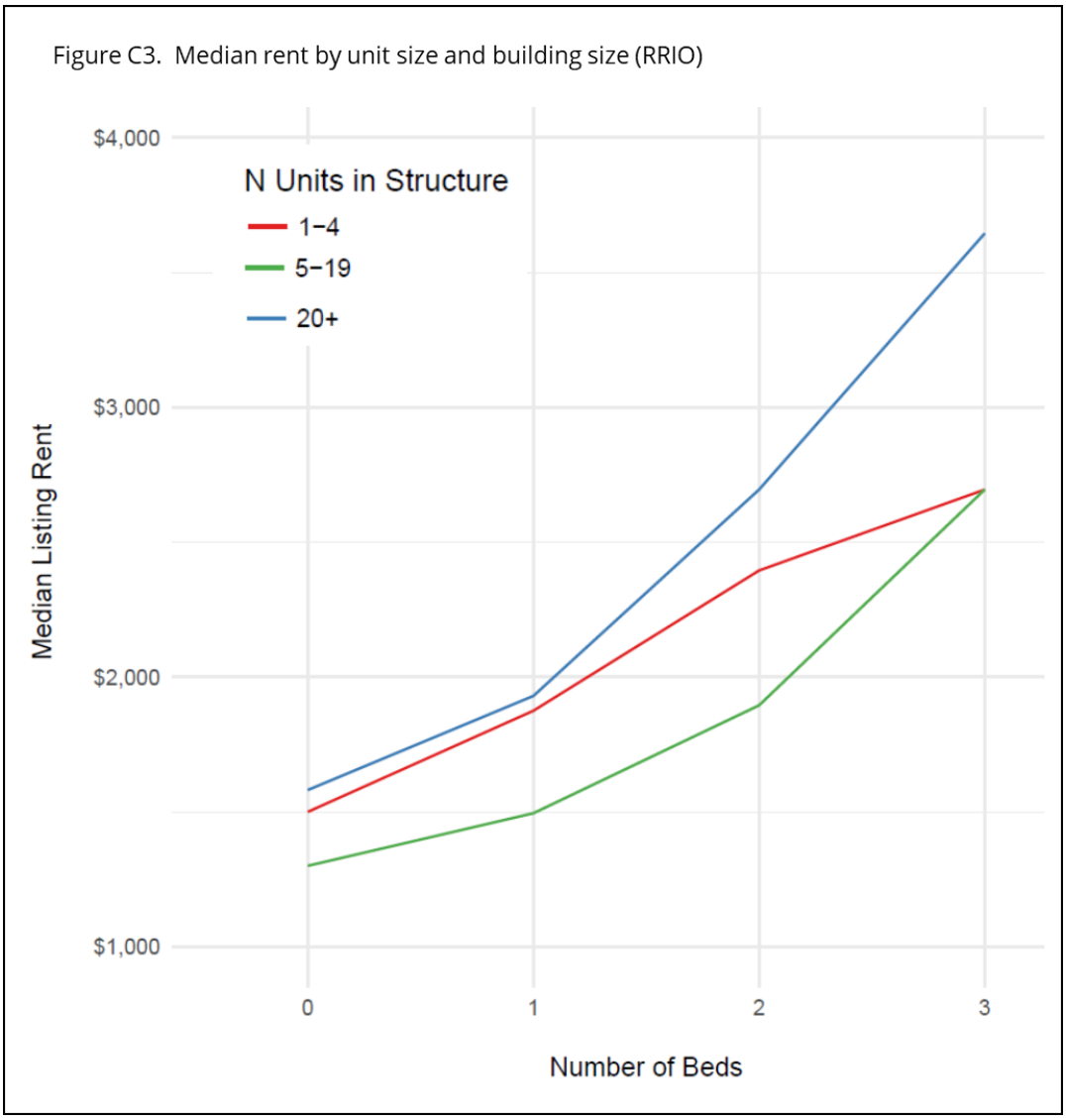
Figure C2. Smoothed Trends in Median Rent for Seattle, 2013-2018.



One of the strong advantages of the data scraped from online ads is that they are highly extensible. Specifically, because these data are available at the unit level, it is possible to attach them to other sources to provide a more complete picture of local housing dynamics. As an illustration, we attached the data scraped from Craigslist to the records of individual units found in data collected under the City's Rental Registration and Inspection

Ordinance (RRIO). This combination allows us to gain much more information about the units advertised on Craigslist, including their modification history and the type of building in which they are located. Figure C3 provides one example of the type of analysis made possible from this type of data merge, showing median rent levels by both unit size and size of building. Given the heavy overrepresentation of large apartment buildings and the uneven provision of unit-level data, this type of analysis is difficult, if not impossible, to accomplish with commercial data, highlighting the potential value of the tools presented here.

Figure C3. Median rent by unit size and building size (RRIO)



The SRHS team continues to perform statistical comparisons of rent patterns as derived from multiple sources of data and is integrating data on current rental listings with property records in King County tax parcel data. This will enable us to assess differences in rent and availability by specific features of the building in which rental properties are located (e.g., year built, number of units, etc.). In addition, this data integration will allow us

to analyze rents of apartment complexes separate from condos and other residential properties, thereby increasing the comparability of sub-samples of scraped listings to other data sources. We can also see how housing units differ according to the type of land parcels for which the King County last record classified the listing's address. For instance, those listings associated with recent parcel records for "Vacant Lots" have dramatically higher-than-average rents, highlighting the fact that very recent developments tend to have the highest rents of available units.