Data Alliance for Science

Dept. of Data Science,
University of Washington

deepa15@uw.edu
orbite@uw.edu
anqiw2@uw.edu

Proposal for the Design of the Seattle Gentrification Atlas

Prepared by:

Deepa Agrawal

Erin Orbits, J.D.

Angel Wang

Prepared for:

Bernease Herman, Data Scientist, eScience Institute

Rachel Berney, Assistant Professor in Urban Design and Planning,
College of Built Environments

Gundula Proksch, Associate Professor, Architecture Dept.

Table of contents

Statement of the Problem

Seattle is one of the fastest growing big cities in the United States and future growth is likely to drive increased demand for housing, transportation, and other social services. While this prosperity provides economic opportunities for many residents, it also exacerbates urban inequity through the process of gentrification.

Gentrification has many definitions and the gentrification process is closely studied by policymakers, developers, investors, and researchers. For example, the Center for Disease Control monitors the health effects of gentrification, and defines gentrification "as the transformation of neighborhoods from low value to high value … [having] the potential to cause displacement of long-time residents and businesses" [8]

Similarly, a gentrified neighborhood is one that moves from the bottom half to the top half in the distribution of median household income, rent, or home prices. Other demographic factors can also be considered in defining gentrification, like the distribution of residents' education levels, ethnicities, and ages.

To observe how gentrification has progressed across the city, the Seattle Gentrification Atlas will be developed. This interactive, dashboard Atlas will allow city stakeholders to identify which neighborhoods are in danger of gentrification, while simultaneously visualizing trends within each neighborhood and across the entire city. Visualizing these patterns will provide users with increased insight into the impact of socioeconomic changes in Seattle. The Gentrification Atlas will be useful to policymakers, researchers, and residents.

In order to accomplish this, a supervised machine learning model will predict whether each neighborhood should be classified as: gentrified, in danger of being gentrified, or not gentrified. The labels will be generated by applying the three definitions of gentrification to Census data over the period from 1990 to 2015. Once the labels are established and the model is trained, other neighborhood data will be evaluated for correlation. Ultimately, users will be able to explore (1) how gentrification affected each neighborhood over a twenty-five year period, (2) how other factors are related to the pattern of gentrification, and (3) how neighborhoods are likely to change by 2020.

Design Objectives

1. Conduct exploratory analysis using visualization and unsupervised learning to better understand the available neighborhood data.
2. Using three definitions of gentrification, procure the best data proxies to quantify the factors for each definition, and using the changes in that data, establish class labels for each neighborhood in five year increments.
3. Using those labels, select and train a supervised machine learning model to analyze the relationships between features from other relevant datasets.
4. Combine that data analysis with GIS-based visualizations to produce spatial maps to reflect how gentrification has changed Seattle since 1990.

Literature Review

While all research builds on earlier work to some extent, in many respects, the Seattle Gentrification Atlas project is a continuation of the Equitable Futures project administered by the University of Washington eScience Institute during the annual Data Science for Social Good (DSSG) program. The Equitable Futures project developed:

(1) a prototype for an interactive visualization tool for mapping equity indicators related to housing, income, mobility, and education on the city and neighborhood scale, and

(2) a structural equation model that attempted to measure relationships between publicly available data and underlying socioeconomic factors affecting equity. [1]

The data used in the Equitable Futures project was primarily sourced from the American Community Survey (ACS) and the City of Seattle open data portal [9]. That data, along with the earlier work is available as a resource for the current project.

Figure 1 illustrates the Equitable Futures mapping tool. Here, the neighborhoods are color coded based on how long it takes to walk to a light rail station. The neighborhood detail window includes the histogram to show how First Hill's 10.7 minute walk time compares to the rest of the neighborhoods.

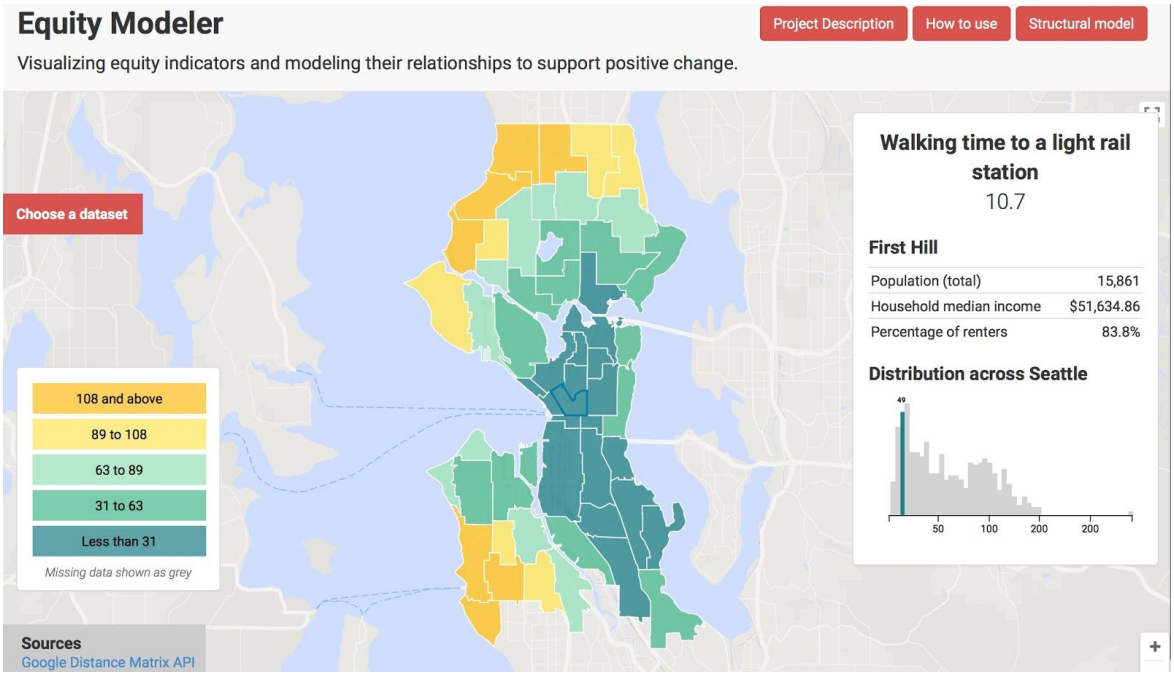

Figure 1. Equity Modeler, the Equitable Futures visualization tool for comparing Seattle neighborhoods.

The color map in Figure 1 is attractive, yet well-suited to aid color-blind users, and the histogram provides a nice way to place the neighborhood in context. However, there is too much wasted space in the dashboard, the red buttons are distracting, the walk time lacks units,

the percentage of renters doesn't seem relevant to the dataset being mapped, and it would be nice to have other neighborhood details, *e.g*. the walk score or the average age.

To understand why gentrification effects neighborhoods differently, it's important to consider the historical context. Unfortunately, systematic racial discrimination was tolerated for decades in Seattle. [13] A 1975 report described how banks and other lending institutions refused to grant home or business loans or required higher interest rates and larger down payments to creditworthy residents in particular neighborhoods, otherwise known as redlining. [13] Sadly, the City's attempts to address the issue of redlining were unsuccessful. Redlining was finally outlawed in June 1977 when Governor Dixie Lee Ray signed the Fairness in Lending Act. Codified as RCW 30A.04.510, the Act prohibits financial institutions from: denying or varying the terms of a loan because of the property's location; or using lending standards with no economic basis. [12]

The City's 2016 Seattle Growth and Equity Analysis report recognizes that due to systemic discrimination, "certain populations and neighborhoods prospered at the expense of others." [2] Minorities were unable to live outside of specific neighborhoods and their neighborhoods were disenfranchised by lack of investment. [2] The Seattle Growth report also produced the map in Figure 2, showing the risk of displacement faced by modern day residents.

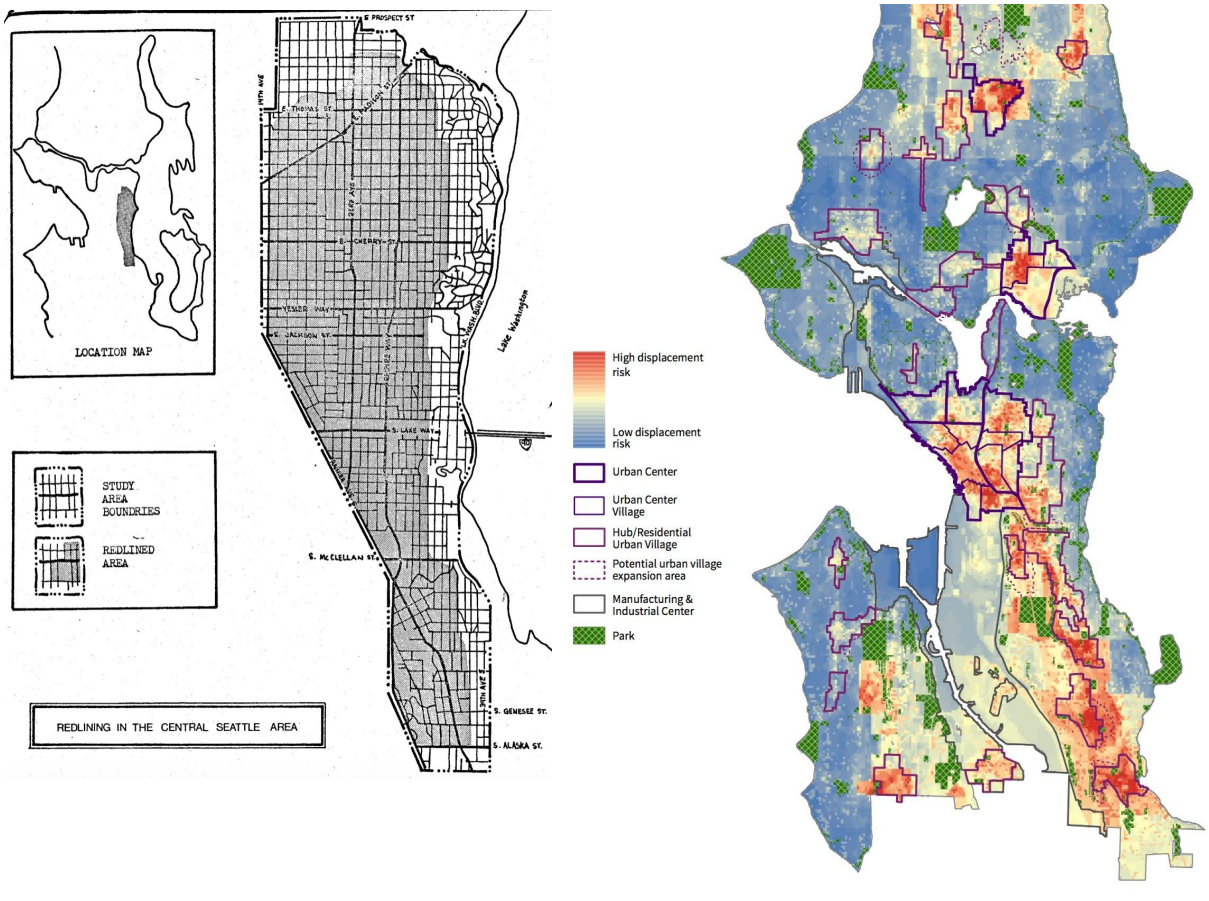

Figure 2. Left image is from the Central Seattle Community Council Federation's 1975 report, and right

image is Displacement Risk Index Map from the Seattle Growth and Equity Analysis 2016 report.

The Seattle Gentrification Atlas will include a map similar to Figure 2. However, the Gentrification Atlas will be interactive, and the metrics may be based on different factors. Finally, since the interactive visualization will allow users to zoom in and pan across the map, the Gentrification Atlas will be easier to read than Figure 2.

Of course, interactive visualization tools for exploring gentrification in Seattle have been created, like the Seattle Gentrification Maps and Data project. [4] Their Seattle Gentrification Map, shown in Figure 3, was created using ACS data from the period of 2009-2013. In addition to panning and zooming, this visualization lets users click on a Census tract to display the detailed tract-level demographic data.

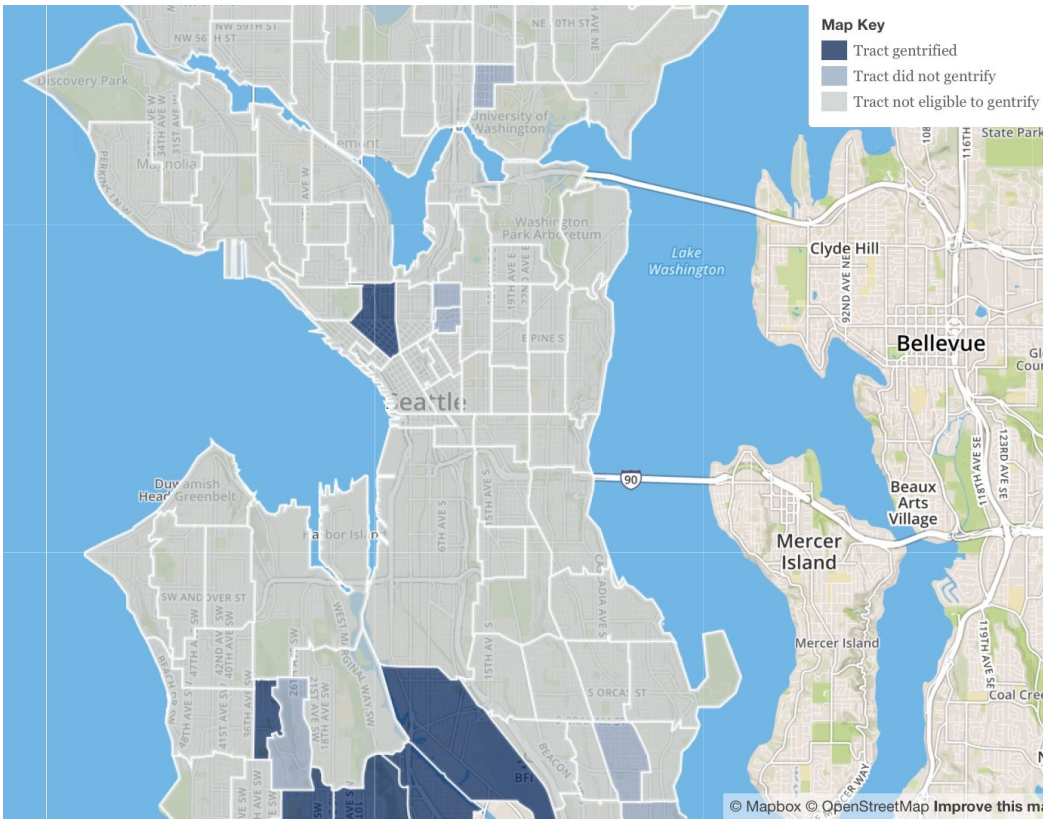

Figure 3. Seattle Gentrification Map: 2000 Census - Present [4]

However, census tracts do not necessarily correspond to neighborhoods. This is a significant weakness since residents tend to be invested in their neighborhoods, not their census tracts. Figure 3 also lacks the detail provided in the Equitable Futures visualization. The Gentrification Atlas will have far more detail than the Seattle Gentrification Map project.

Moreover, the analysis of what constitutes gentrification is somewhat simplistic. Tracts eligible for gentrification have both a median income and median home value in the bottom 40th percentile of all tracks within a city. Gentrified tracts had the median inflation, adjusted home value and percentage of adults with a bachelor degree rise to the top third of all tracts.

A more nuanced analysis of gentrification is presented in the article, Mapping Susceptibility to Gentrification: The Early Warning Toolkit. [5] Chapple investigated patterns of gentrification based on census data, and identified factors that appeared to increase the likelihood of residents being displaced by new development. [5] This project is similar in scope to the Gentrification Atlas, but Mapping Susceptibility to Gentrification analyzed the Bay Area in California. Table 1 lists the factors related to gentrification in the Bay Area: a positive direction means it was associated with more gentrification and a negative factor means less.

Variable type	Variable	Direction	Rank
Transportation	**% of workers taking transit**	**Positive**	4
	Youth facilities per 1,000	**Positive**	3
	Public space per 1,000	**Positive**	5
Amenities	Small parks per 1,000	Positive	17
Demographic	**% non-family households**	**Positive**	8
	% of dwelling units in buildings with 5+ units	**Positive**	7
	% of dwelling units in buildings with 3-4 units	**Positive**	10
	% renter-occupied	**Positive**	13
Housing	Public housing units	Positive	19
	Income diversity	**Positive**	6
Income	**% of renters paying > 35% of income**	**Positive**	11
Location	Distance to San Jose	Positive	14
Transportation	**% of dwelling units with three or more cars available**	**Negative**	2
Amenities	**Recreational facilities per 1,000**	**Negative**	1
	% married couples w/ children	**Negative**	9
Demographic	**% non-Hispanic white**	**Negative**	12
Housing	Median gross rent	Negative	18
Income	% of owners paying > 35% of income	Negative	15
Location	Distance to San Francisco	Negative	16

Table 1. Factors behind gentrification in the 1990s in the Bay Area [5]

Following the Mapping Susceptibility to Gentrification report, an online visualization tool called the Urban Displacement Project was developed to further understand the nature of gentrification and displacement in the Bay Area and Southern California. [6]

Like the Equitable Futures visualization (see Figure 1), the Urban Displacement Project presents several interactive maps that visualize population, housing prices, and rent. [6] But, unlike any of the other visualization tools, the Urban Displacement Project maps how tracts change over time. For example, Figure 4 shows the percent change in median home price from 2000 to 2013 for each census tract. *Id*. The Gentrification Atlas will also show change over time, and may use color to encode rates of change. Alternatively, the Atlas may use a slider to allow the user to observe changes as they scroll through the years of encoded data.

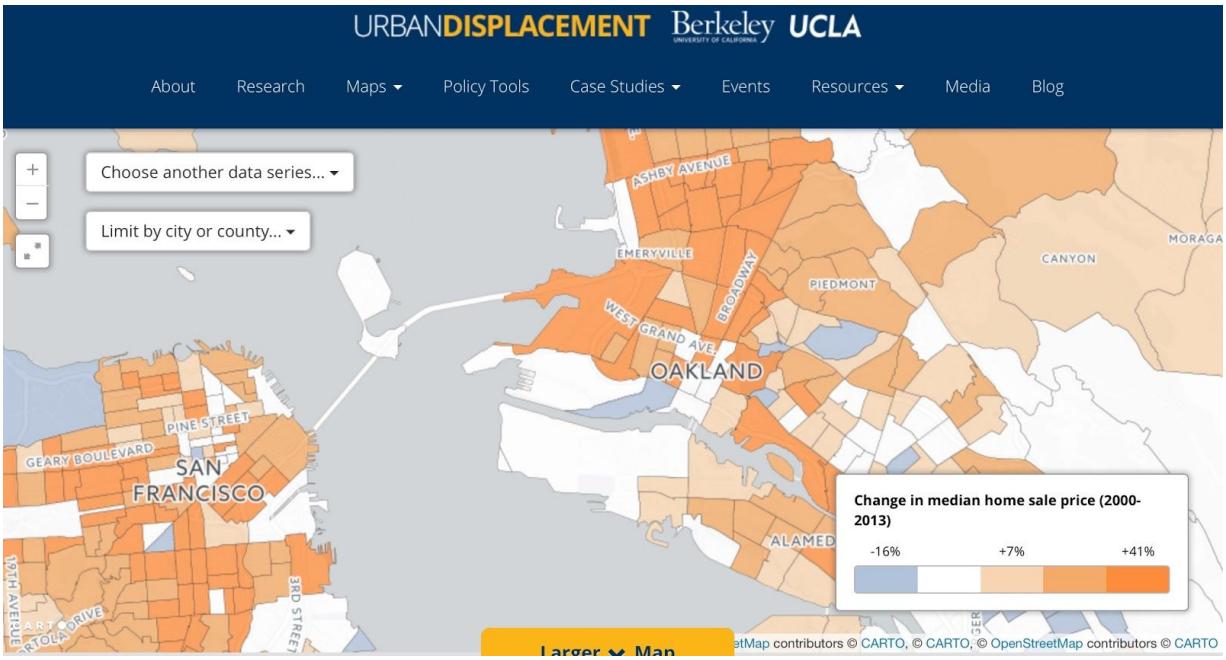

Figure 4. Urban Displacement mapping of the change in median home prices. [6]

The first map a user sees on the Urban Displacement website is shown in Figure 5. [6] Although the Displacement Typologies box arguably contains too many categories and it's difficult to discern between the medium shades of purple and orange, this is the best map among all the prior work. With a few small changes, this map could clearly communicate gentrification trends, and may be a model for one of the visualizations in the Gentrification Atlas.

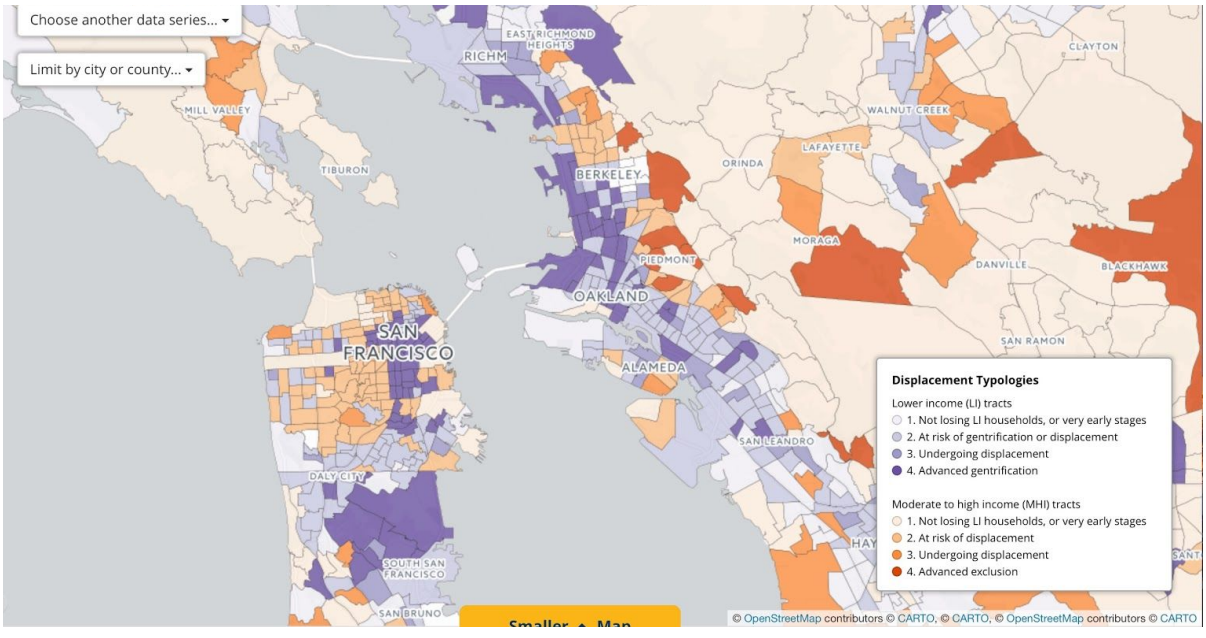

Figure 5. Map of Displacement Typologies from Urban Displacement website. [6]

Design Concept

The Seattle Gentrification Atlas project will compare the definition of gentrification from the perspectives of researchers in the fields of sociology, finance and economics, and urban planning. Those three definitions will be used to identify the neighborhoods: that have gentrified, that are currently gentrifying, and that are in the greatest danger of gentrifying. Regardless of the definition, gentrification is marked by a change over time. So, the analysis will be conducted for 5 year increments over the period from 1990 to 2015 .

Sociology

- In danger of gentrifying:
    - The block group's median household income was in the bottom 40th percentile when compared to all block groups in Seattle at the beginning of the time span.
    - The block group's median house value was in the bottom 40th percentile when compared to all block groups in Seattle at the beginning of the time span.
- Have gentrified:
    - An increase in a block group's educational attainment, as measured by the percentage of residents age 25 and over holding bachelor's degrees, was in the top third percentile of all block groups in Seattle.
    - An increase in a block group's median house value, as measured by inflation-adjusted median house value, was in the top third percentile of all block groups in Seattle.

Financial & Economics

- Currently gentrifying:
    - The block group's income levels below 40% of the median, and experienced rent increases greater than the median neighborhood did.
- Have gentrified:
    - The block group's share of neighborhoods in a metro area that moved from the bottom half to the top half in the distribution of home prices

Urban Planning

- In danger of gentrifying:
    - The block group's % of workers taking transit increased
    - The block group's youth facilities per 1000 residents increased
    - The block group's public space per 1000 residents increased
    - The block group's % non-family households increased
    - The block group's % dwelling units in building with 5+ units increased
    - The block group's % dwelling units in building with 3-4 units increased
    - The block group's % renter-occupied increased
    - The block group's income diversity increased
    - The block group's % of renters paying >35% of income increased

- The block group's % of dwelling units with three or more cars available decreased
- The block group's recreational facilities per 1000 residents decreased
- The block group's % married couples with children decreased
- The block group's % non-hispanic white decreased

Datasets

The Gentrification Atlas project will primarily analyze American Community Survey (ACS) data at the block and block group levels. Block groups are typically comprised of 600 to 3,000 residents, while census tracts are county subdivisions with 1,200 to 8,000 residents. [7] The census data is publicly available under a U.S. Work license.

Although Census data may be used to set the floor values in 1990, since the U.S. Census is only conducted every 10 years, we will be using ACS data. The Census 2000 sample represented approximately 1 in 6 households as of April 1, 2000. [7]

In contrast, the ACS represents approximately 1 in 40 households and the smaller sample sizes produce larger sampling errors. However, the ACS collects population and housing information every year instead of every 10 years. [7] The ACS is conducted by sampling a small percent of the population every month.

Variables from the Census, ACS, and Department of Motor Vehicles (DMV) records will be extracted for analysis and models will be developed based on all three definitions.

Architecture

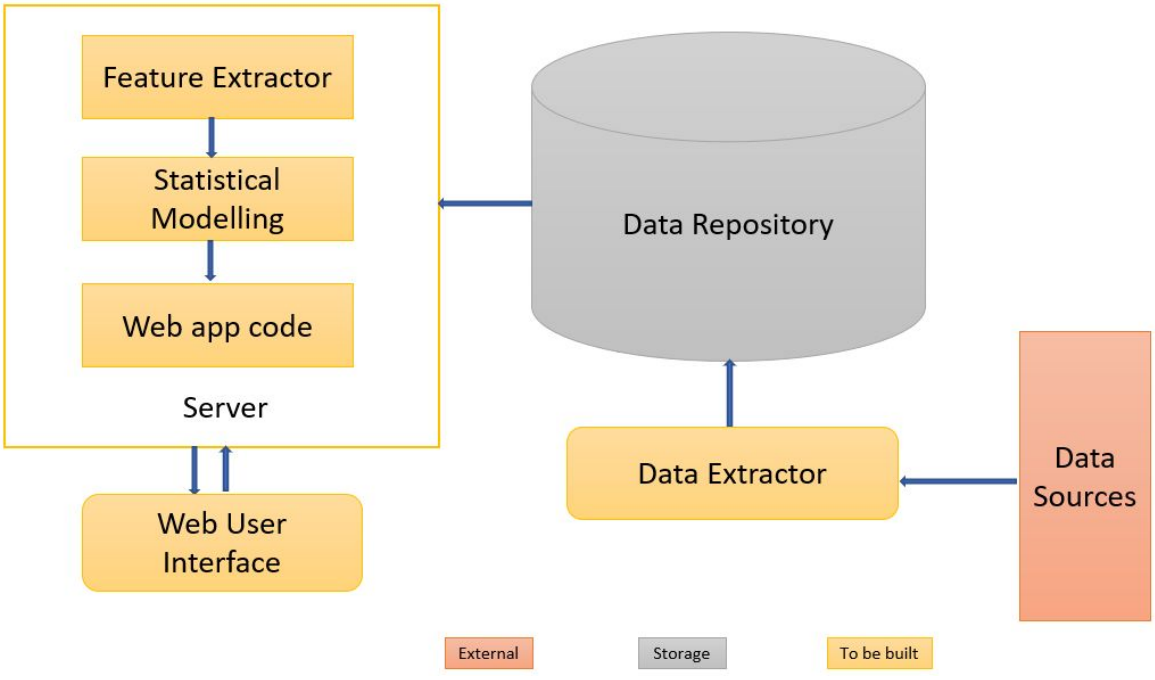

Figure 6: Architecture diagram for the Seattle Gentrification Atlas project

Components

Data extractor

The data extractor module will be built using Python. This module would contain the extraction algorithms and api calls for different selection criteria. The extractor will invoke the APIs as needed, retrieve the data from different sources(census data, ACS, DMV records), clean the data and upload the data into the database. The final product of the extractor will be dataset containing columns like median income, house prices, modes of transit, rental information, and etc.

Data Repository

The data repository will contain all data extracted from the Data Extractor in CSV file format. Since the size of these CSV files is not expected to be large, these CSV files will be stored in the team GitHub repository here. We will also create a data dictionary which would contain the metadata describing the contents of the data files such as description of each dataset, descriptions of each column, and etc.

Statistical model

We expect to primarily use Python packages to build the statistical model for gentrification prediction. In order to analyze and examine the relationships and weights among features about the neighborhoods and its gentrification process, supervised and unsupervised classification/ clustering models will be trained to predict whether a neighborhood is in danger of gentrification.

With the limitation of data sample size (<10k), the statistical model will examine the following components:

1. Clustering: Explore available data and start with clustering to segregate neighborhoods with similarity and assign them into clusters using Kmeans.
2. Classification: Based on all three gentrification definitions, we can label the historical data first and classify neighborhoods using linearSVC, Kneighbors Classifier of SVC ensemble classifier.
3. Unsupervised clustering approach will be used to increase the accuracy of supervised classification prediction by using cluster labels as independent variables in the supervised machine learning algorithm

Web page

Web page will be created using d3 or tableau. We are currently in process of vetting both the technologies. The web page will be an interactive spatial mapping tool with gentrification information in Seattle. The neighborhoods will be color coded per the gentrification status across the time period 1990-2015. In addition, a drop down menu containing the predictors for gentrification may be added to allow users to view the predictor values for each neighborhood.

User will have the capability to select a neighborhood and view the values associated with each predictor. Following design considerations will be taken while developing the UI: ease of use, accessibility, responsiveness and consistency.

Deliverables

Interactive spatial mapping tool with gentrification information in Seattle. This tool allows residents, researchers, policymakers and others to easily access and view this data. Users select from a list of available data layers to display the data for a given "neighborhood cluster," a geographic type used in Seattle, the neighborhood's demographic composition and gentrification status for the neighborhood.

Project Management

Communication and Coordination with Sponsor

For major milestone meetings with the sponsor, all members will attend the meetings in person (if possible). We will all contribute questions, feedback, and status reports for contact with the sponsor. But, for routine questions or issues Erin Orbits will be the point contact for communication by phone or email. Our team is currently sharing a Google Drive folder with the project sponsor to coordinate the sharing of documents. Depending on the needs of the project and the wishes of the sponsor, the team GitHub repository could also be shared to keep the sponsor apprised of our work.

Our communication plan for next 5 months includes:

- Weekly in person meetings on Thursdays with members of the sponsor group
- Using a private Slack channel for intra-team status updates and questions
- Scheduling online meetings through Skype or Google Hangouts as needed on the weekends to coordinate project development
- Using a GitHub repository to organize code, data and documentation updates

Team Qualifications

Deepa is studying Data Science at the University of Washington and expects to earn her Masters of Science (MS) degree in March 2018. She has 10 years of Software Development Life Cycle experience in Requirement gathering, Design, Development, Data Modelling and Business/Data analysis. Multi domain experience in Ecommerce, Insurance, Mobile, Online gaming and Avionics. Planned responsibilities for the capstone project include data acquisition, data processing, maintaining data, query execution and contributing to machine learning model.

Erin is studying Data Science at the University of Washington and expects to earn her MS degree in March 2018. Prior to enrolling in this program, she was a full-time attorney. Erin is looking forward to applying her statistical analysis, machine learning, and data visualization skills to the Seattle Gentrification Atlas project. Her responsibilities for the Seattle Gentrification Atlas project include the user interface design, code documentation, communication with the sponsor, and project presentation.

Angel is an aspiring data scientist with four-year proficiency in data analysis, statistical modelling, project management, and translating statistical results into business recommendations. Industry experience in R and Python for writing machine-learning algorithms, SQL for database querying, Hive/Hadoop for distributed computing, and Tableau for generating visualizations. Angel will be the one who is responsible to leverage machine learning algorithms to build supervised learning model and utilize her background in Statistics to better understand the data.

References

[1] B. Herman, et al., “Data Science for Urban Equity: Making Gentrification an Accessible Topic for Data Scientists, Policymakers, and the Community,” Bloomberg Data for Good Exchange Conference, Chicago, IL, USA, Sept. 24, 2017. [Online]. Available: https://export.arxiv.org/pdf/1710.02447

[2] “Seattle 2035 Growth and Equity Analyzing Impacts on Displacement and Opportunity Related to Seattle’s Growth Strategy,” Seattle Office of Planning and Community Development, pp. 1-68, May 2016. [Online]. Available: http://www.seattle.gov/Documents/Departments/OPCD/OngoingInitiatives/SeattlesComprehensivePlan/FinalGrowthandEquityAnalysis.pdf

[3] L. Freeman, “Displacement or Succession? Residential Mobility in Gentrifying Neighborhoods,” *Urban Affairs Review*, vol. 40, no. 4, pp 463-491, March 1, 2005. [Online]. Available: https://doi.org/10.1177/1078087404273341

[4] Seattle Gentrification Maps and Data, *GOVERNING*, [Online]. Available: http://www.governing.com/gov-data/seattle-gentrification-maps-demographic-data.html [Accessed October 25, 2017]

[5] K. Chapple, “Mapping Susceptibility to Gentrification: The Early Warning Toolkit,” Center for Community Innovation, Berkeley, CA, May 2009. [Online]. Available: http://communityinnovation.berkeley.edu/reports/Gentrification-report.pdf

[6] M. Zuk and K. Chapple, Urban Displacement Project, 2015. [Online]. Available: http://www.urbandisplacement.org [Accessed Nov. 7, 2017]

[7] American Community Survey, Esri Demographics, [Online]. Available: http://doc.arcgis.com/en/esri-demographics/reference/essential-vocabulary.htm [Accessed Nov. 7, 2017]

[8] Health Effects of Gentrification, Centers for Disease Control, updated August 21, 2013 [Online]. Available: https://www.cdc.gov/healthyplaces/healthtopics/gentrification.htm [Accessed Nov. 10, 2017]

[9] City of Seattle Open Data portal, 2017, [Online]. Available: https://data.seattle.gov

[10] Redlining, Seattle City Council, referencing Planning and Urban Development Committee Meeting, September 1, 1976. Event ID 3621, Seattle City Council Legislative Department Audio Recordings, 4601-03, [Online]. Available: http://www.seattle.gov/cityarchives/exhibits-and-education/seattle-voices/redlining [Accessed November 10, 2017]

[11] Redlining in Seattle, Seattle City Council, [Online], Available: http://www.seattle.gov/cityarchives/exhibits-and-education/online-exhibits/redlining-in-seattle [Accessed November 10, 2017]

[12] Financial Institutions Disclosure Act - Fairness In Lending Act, House Bill No. 323, Ex Session 1, chap. 301, June 21, 1977, [Online], Available: http://leg.wa.gov/CodeReviser/documents/sessionlaw/1977ex1c301.pdf?cite=1977%20ex.s.%20c%20301%20§%2012

[13] "Redlining and Disinvestment in Central Seattle: How the Banks are Destroying our Neighborhoods," Central Seattle Community Council Federation, July 1975, [Online]. Available: http://clerk.seattle.gov/~F_archives/documents/Doc_11219.pdf

Appendix A:

Résumés of Team Members

DEEPA **AGRAWAL**

Redmond, WA 98053 | C: 425-449-3295 | deepa15@uw.edu | https://www.linkedin.com/in/deepa-agrawal/

PROFILE

- Master of Science, Data Science at University of Washington starting Sept'16.
- 10+ years of SW experience: Requirements, Design, Development, Data Modelling, Business/Data analysis.
- Multi domain experience: Ecommerce, Insurance, Mobile, Online gaming and Avionics.
- Lead experience: led a team of 6 resources in a data warehousing project.
- International work experience: US, Canada and India.

EDUCATION /DESIGNATIONS

Master of Science: Data Science (Current GPA: 3.87) **University Washington,** Seattle **Class of 2018**

Engineering (BTech - IT) **National Institute of Technology,** India **June 2005**

AINS : Associate in General Insurance **The Institutes,** United States **May 2014**

SKILLS

Database: SQLite, SQL Server, SQL Azure **Tools:** SSMS, SSRS, MySQL Workbench, Apache Drill, JDBC

Data Visualization tools: Tableau, Power BI **Languages**: SQL, R, C#, Python

PROFESSIONAL EXPERIENCE (2005-2017)

Microsoft Intern in Azure Networking – Redmond **Jun'17-Sept'17**

As a data science intern, responsible for analyzing Ping Mesh data to improve networking coverage and real-time alerting.

Implementing statistical models to programmatically detect spikes and trends in latency across various dimensions.

Working on fine tuning the existing Stochastic Gradient Descent algorithm to get better confidence level.

Microsoft via Wipro Limited – Redmond **Mar'15-Apr'16**

As a senior engineer, responsible for data analysis of test failures and overall health of the triage pipeline.

Created live dashboards using Power BI to monitor Windows test runs.

Berkley North Pacific (BNP) – Bellevue **Aug'13-Dec'14**

As a Quality Analyst, owned multiple Lines of Business: General Liability, Commercial Auto, Commercial Property, Business Owners Policy and Umbrella coverage.

Used data analysis techniques to validate business rules and identify issues in policy pricing.

Cognizant Technology Solutions – Pune **Mar'10-Aug'10**

A strategic three-way reconciliation for the JPMC to match Balance Sheet Mark to Market, Realized/Unrealized Profit and Loss between numerous Risk systems, Sub Ledgers and General Ledger systems. As a Test Lead, managed a team of 6 resources to complete end to end ETL testing.

Wipro Technologies – Bangalore, Pune, Toronto **Sept'05-Jan'10**

HiPlus for Aviva Hibernian, UK: An Insurance website used by Hibernian Direct Ltd (HDL) for motor, home and travel insurance products.

Smoke Detection Function for Airbus (A-340), Germany: As a software engineer owned testing and verification of the SDF (Smoke Detection Function) Module.

Erin Orbits

Attorney and Data Scientist

orbite@uw.edu | cell (425) 444-6269 | github.com/orbitse

Objective: To utilize and expand my statistical analysis, programming, and data visualization skills working on a Data Science Team where I can also apply my legal knowledge.

TECHNICAL SKILLS

Statistics

Bayesian inference

Descriptive Statistics

Hypothesis Testing

Machine Learning Tools

SciKit Learn

TensorFlow

Machine Learning Algorithms

Gradient Descent

K-means

K nearest neighbors

Logistic Regression

Primary Component Analysis

Ridge Regression

Data Management

AWS: EC2, EMR, Redshift

Azure

Hadoop, Hive

SQL, SQL Server

Programming Languages

Python, R, Java

EDUCATION

University of Washington (Sept. 2016 – Present)
Masters of Science in Data Science, exp. Mar. 2018
Current cumulative GPA 3.7
Practiced applied statistics and experimental design, machine learning techniques, MapReduce, parallel processing, and user-centered data visualization in ggplot2, Matplotlib, Bokeh, d3.js and Tableau

Seattle University Law School (July 2004 – May 2007)
Juris Doctor, *magna cum laude*
Awarded President's Scholarship; CALI Award for highest grade in Appellate Advocacy; and competed on the Moot Court Team, winning Best Advocate Award in national law competition

Whitman College (Sept. 2000 – May 2003)
Bachelor of Arts, with distinction, Economics
Awarded President's Scholarship; Elected Student Body President

RECENT WORK EXPERIENCE

Attorney, Private Practice (Nov. 2012 – Sept. 2016)
Specialized in statutory construction and interpretation; conducted research, drafted briefs, and consulted

Deputy Prosecuting Attorney, Pierce Co. (Feb. 2008 – Oct. 2012)
Tried 40+ cases, including a 3rd Strike, 1st Degree Assault jury trial that resulted in a guilty verdict; Argued appellate cases including the appeal of a juvenile court order dismissing a deferred disposition before the Court of Appeals: State v. D.P.G., 169 Wn. App. 396, 280 P.3d 1139 (2012); Exercised judgment in thousands of negotiations

SELECTED DATA SCIENCE PROJECTS

Data Visualization Tool	**Image Classifier Model**
Cleaned, standardized, and analyzed WA Dept. of Revenue data on open and closed businesses in Seattle, then merged that data with geolocation data and GIS shapefiles to create an interactive tool for analyzing small business trends in Tableau.	Used TensorFlow to build a convolutional neural network, extracted features from 10,000 photos of birds, and used a logistic regression model for identifying 200 species of birds in an GPU EC2 instance before saving the predictions in a CSV file.

Anqi Wang| 588 Bell St Unit 2705S, Seattle, WA 98121 | anqiw2@uw.edu | (206) 790-5569

PROFESSIONAL SUMMARY

Aspiring Data Scientist with four-year proficiency in data analysis, statistical modelling, project management, and translating statistical results into business recommendations. Industry experience in R and Python for writing machine-learning algorithms, SQL for database querying, Hive/Hadoop for distributed computing, and Tableau for generating visualizations.

WORK EXPERIENCE

Expedia, Inc., Bellevue, WA — Jun 2017 – Sep 2017
Data Science Intern

- Loaded and modeled Expedia booking data in Hadoop, and processed the data for analysis to answer customer service experience questions.
- Developed, implemented and tested statistical models using R (caret, party, tree, mboost, e1071, rpart, etc.) and Python (SciKit-Learn, TensorFlow, Scrapy, etc.) to show how booking variables impact customer service contact behavior and recommended business solutions to deflect customer service contact and provided effortless customer experience.
- Project: Created predictive customer service call models utilizing 23M Expedia 2016 booking transaction data. Three predictive models were built and tested in R and Python: identifying likely callers, predicting caller window and caller need. Proposed initiatives on service call reduction and customer segmentation, using personalized experience based on insights from analysis. Customer service calls have been reduced by 4% after implementing the model for test and learn in September 2017.

UW Medicine, Seattle, WA — Sep 2016 – Jun 2017
Analyst Intern

- Conducted statistical analysis on employee on-boarding survey data to generate predictive analysis trends using Excel and R. Produced and presented reports, using PowerPoint and Word, to senior leadership in HR to improve employee satisfaction.
- Provided administrative support and managed record maintenance based on established retention schedules for HR.

Wells Fargo, Seattle, WA — Jan 2015 – Jun 2016
Financial Advisor/Licensed Private Banker

- Managed client relationships by structuring portfolios catering to clients' financial status, objective, risk tolerance, tax exposure and investment goals.
- Leveraged computational analytics to devise accurate projections of portfolio returns for clients.

Expeditors International of Washington, Inc., Seattle, WA — Jan 2014 – Dec 2014
Business Analyst

- Analyzed transportation and insurance trends in data sets using SQL and Excel to improve annual operational efficiency by 20%.
- Proposed solutions aligned with customer needs and increased satisfaction by 7% in Q4 2014.

EDUCATION

University of Washington, Seattle, WA — Expected: Mar 2018
M.S. Data Science — Current GPA: 3.91
Relevant coursework: Data Visualization & Experimental Design, Data Management for Data Science, Statistical Machine Learning, Scalable Data Systems & Algorithms.
Project: Built a visualization application (TravelViz) to assist users in making informed travelling decisions based on area of interest, neighborhood, time of year and time of day. The application data, over 8GB, was sourced from 2.7M Yelp reviews. Using R, I applied statistical modeling and ANOVA to analyze trends between user reviews and business star rating, and Natural Language Processing to extract keywords from reviews. Visualizations from the analyzed data were generated using Tableau.
Link: http://cp6863.axshare.com/#c=2.

University of Washington, Seattle, WA — Dec 2013
B.S. Statistics & Economics, Minor in Mathematics — GPA: 3.55

QUALIFICATIONS

- Programming Languages: R, Python, SQL, JavaScript, HTML, CSS
- Cloud computing platform: AWS (Redshift, EC2, S3), Azure
- Computing Framework: Hive/Hadoop, Spark
- Data Visualization Tool: Tableau