Learning & Doing Data for Good:

a conference for academics and their partners

September 9-11th, 2022

University of Washington
Seattle, WA

#UWdata4good

WEST BIG DATA INNOVATION HUB

UNIVERSITY of WASHINGTON eScience Institute
DATA SCIENCE FOR SOCIAL GOOD

Academic Data Science Alliance

THE UNIVERSITY OF BRITISH COLUMBIA
Data Science Institute
Faculty of Science

WELCOME

Thank you for joining us for this unique conference! At this forum, you will hear from current students and alumni from university-based data for good programs alongside project partners and data science professionals. This will be a space for inspiring discussions and for connecting with colleagues and partners who are motivated to learn from and meet the needs of communities and people using data for change. We gratefully acknowledge the work of Conference Director Juandalyn Burke, our Program Committee, and our Sponsors.

Land Acknowledgement

We ask for those engaging in this event to reflect on the lands on which you reside and acknowledge all of the ancestral homelands and traditional territories of Indigenous peoples who have been here since time immemorial. We are hosting this event out of the University of Washington and we would like to acknowledge the Coast Salish peoples of this land, the land which touches the shared waters of all tribes and bands within the Duwamish, Puyallup, Suquamish, Tulalip, and Muckleshoot nations. More information here: **washington.edu/diversity/tribal-relations/**

Code of Conduct

The University of Washington eScience Institute is dedicated to providing a welcoming, supportive, and inclusive environment for all people, regardless of background and identity. We do not tolerate discrimination or harassment. Any form of behavior to exclude, intimidate, or cause discomfort is a violation of the Code of Conduct. By participating in this community, participants accept to abide by the eScience Code of Conduct and accept the procedures by which any Code of Conduct incidents are resolved. Full text available here: **escience.washington.edu/code-of-conduct**

Kumospace - the Virtual Environment www.kumospace.com

There will be times during the conference when you will need to access the online web-based platform, Kumospace, to engage with conference attendees remotely. The virtual events are also noted on the conference schedule. You do not need to register or sign-up for Kumospace. You will simply join the virtual space using a Kumospace conference link. Please note: the virtual space link will be provided on Day 1 of the conference (Friday, September 9th).

COVID-19 Precautions

Per UW policy, masking is strongly recommended when indoors. We encourage you to take a COVID test prior to arriving at the event. Walk-up testing in E4 (the lot next to the golf range across from University Village) is open to the public and guests can schedule here: **www.solvhealth.com/book-online/AX1lop.** We will have tests available at the registration desk. If you need a rapid test, please allow an additional 15 minutes to complete the test before the start of the session. If you test positive, or if you have symptoms that are likely to be COVID, please self isolate and stay home or in your hotel room. Contact Sarah Stone: 425-246-4835, sstone3@uw.edu, for additional guidance.

AGENDA

Please note all times are in Pacific Daylight Time

Friday, September 9th (Day 1)

12:00 p.m.	**Location: University of Washington Husky Union Building (HUB), first floor, street level. 4001 E. Stevens Way NE, Seattle, WA 98195**
	Event registration begins *(open 12:00 - 6:00 p.m.)*
	Walking tour of University of Washington campus - optional *(12:00 - 3:45 p.m.)*
	Virtual and in-person game room - optional *(open 12:00 - 3:45 p.m.)*
4:00 p.m.	**Location: Gates Center for Computer Science & Engineering, Zillow Room. 3800 E. Stevens Way NE, Seattle, WA 98195**
	Pre-welcome *(4:00 - 4:15 p.m.)*
4:30 p.m.	Mixer: Getting to Know You in-person & virtual chat *(4:30 - 6:00 p.m.)*
7:00 p.m.	Dinner on your own, reservation sign up sheets available in person

Saturday, September 10th (Day 2)

7:30 a.m.	**Location: Gates Center for Computer Science & Engineering, Zillow Room. 3800 E. Stevens Way NE, Seattle, WA 98195**
	Event registration
	In-person and virtual poster display
8:15 a.m.	**Welcome**
8:30 a.m.	**Keynote speaker, Dr. Desmond Upton Patton**
9:30 a.m.	**Session 1 - Community-Health: Discovering Data and Modeling in Public Health**, 10 minute talks. Details on page 6
10:45 a.m.	Break
11:00 a.m.	**In-person and virtual poster session**. Details on page 8
12:00 p.m.	Lunch provided
1:00 p.m.	**Session 2** - **Transparency, Accountability and Engagement with Civic Systems**, 10 minute talks. Details on page 6

AGENDA

Saturday, September 10th (Day 2) continued

2:00 p.m. **Data for Good Partners panel**
Moderator: Katherine Chang; Panelists: Anat Caspi (UW), Rob MacDougall (Johnson County Mental Health Center), Kathryn Thomas (Yoga 4 Change)

3:00 p.m. Break

3:15 p.m. **Exploring Career Paths panel**
Moderator: Pratik Sachdeva; Panelists: James (Lamar) Foster (Bill and Melind a Gates Foundation), Natalie Gasca (California Council on Science and Technology), Caprichia Jeffers (JP Morgan Chase), Dan Morris (Google AI for Nature and Society)

4:30 p.m. Appetizers, posters, and mingling

6:30 p.m. Dinner on your own, reservation sign up sheets available in person

Sunday, September 11th (Day 3)

8:00 a.m. **Location: Gates Center for Computer Science & Engineering, Zillow Room. 3800 E. Stevens Way NE, Seattle, WA 98195**

Breakfast

In-person and virtual poster display

8:30 a.m. **Closing remarks**

8:45 a.m. **Session 3 - Climate Change and Environmental Impacts**, 10 minute talks. Details on page 7

9:45 a.m. Break

10:00 a.m. **Session 4 - Housing, Poverty, and Employment**, 10 minute talks. Details on page 7

11:00 a.m. **Data for Good Alumni: Where are they now?**

Moderator: Juandalyn Burke; Panelists: Elba Gomez, Senior Data Analyst (Canadian Partnership Against Cancer), Habeeba Siddiqui (Mayo Clinic), Abhishek Singh (Kohl's)

12:00 p.m. Boxed lunches available

KEYNOTE SPEAKER: Dr. Desmond Upton Patton

Desmond Upton Patton studies the impact social media has on well-being, mental health, trauma, violence and grief for youth and adults of color. He leverages social work thinking, data science, qualitative methods, and community partnerships to develop strategies to support digital grief and trauma and reduce on and offline gun-related violence. Desmond is the Brian and Randi Schwartz University Professor and the thirty-first Penn Integrates Knowledge University Professor. He has joint appointments in the School of Social Policy & Practice and the Annenberg School for Communication along with a secondary appointment in the department of psychiatry in the Perelman School of Medicine.

Professor Patton's groundbreaking research into the relationship between social media and gang violence – specifically how communities constructed online can influence often harmful behavior offline – has led to his becoming the most cited and recognized scholar in this increasingly important area of social science. His early work attempting to detect trauma and preempt violence on social media led to his current roles as an expert on language analysis and bias in AI and a member of Twitter's Academic Research advisory board and Spotify's Safety Advisory Council. As a social worker, Patton realized existing gold standard data science techniques could not accurately understand key cultural nuances in language amongst predominantly black and Hispanic youth. In response, he created the Contextual Analysis of Social Media (CASM) approach to center and privilege culture, context and inclusion in machine learning and computer vision analysis. With this methodology, organizations can better foster diverse and inclusive environments and minimize employee conflict. Further, Patton's insights on creating non-biased and culturally nuanced algorithms give tech companies a holistic perspective on various business and social issues.

Before joining Penn, Patton was Professor of Social Work and Sociology at Columbia University, Senior Associate Dean at Columbia School of Social Work and Associate Director of Diversity, Equity and Inclusion at the Data Science Institute at Columbia.

Dr. Patton has been featured in The New York Times, the Chicago Tribune, USA Today, NPR, Boston magazine, ABC News, Nature, and Vice. His work was cited in an amici curiae brief submitted to the United States Supreme Court in Elonis v. United States, which examined the interpretation of threats on social media.

He holds a Bachelor of Arts in Anthropology and Political Science with honors from the University of North Carolina at Greensboro, an Master of Social Work from the University of Michigan School of Social Work, and a PhD in Social Service Administration from the University of Chicago.

PRESENTATIONS: 10-Minute Talks

Session 1: Saturday, September 10th (9:30 - 10:45 a.m.)
[Theme] Community-Health: Discovering Data and Modeling in Public Health
Session Chair: Ashley Atkins (West Big Data Innovation Hub)

Combining Content and User Data for COVID-19 Misinformation Alignment Detection
Presenter: Nazanin Jafari

Determinants of Covid Mortality: Why is the United States' COVID-19 mortality so high?
Presenters: Jiacheng Ge, Charles Hendrickson, Scout Leonard

A Multi-Label Multi-Class Approach to Assigning ICD-10 Codes from Unstructured Clinical Text Data
Presenters: Olumurejiwa Fatunde, Jennah Gosciak, Christina Last, Kennedy Odongo

Re-initiation of the Community Learning Through Data Driven Discovery Process: Revisiting Iowa's Alcohol Landscape
Presenter: Kelsey Van Selous

Supporting Communities and Community Organizations Through Data Science: A story of food rescue in central Iowa
Presenter: Matthew Voss

Improving the National Suicide Prevention Lifeline's Service through Efficient Call *Routing*
Presenters: Charles Cui, Irene Tang

Session 2: Saturday, September 10th (1:00 - 2:00 p.m.)
[Theme] Transparency, Accountability and Engagement with Civic Systems
Session Chair: Anissa Tanweer (University of Washington)

Bridging the Gap Between Computational Tools and User Adoption for Social Good: A Case Study in Political Redistricting Problems
Presenters: Rowana Ahmed, Katherine Chang

eiCompare: Making Every Vote Count
Presenters: Juandalyn Burke, Ari Decter-Frain, Hikari Murayama, Pratik Sachdeva

Identifying behavioral health conditions in Californian police records (virtual)
Presenters: Grishma Bhattarai, Salil Goyal, Alexander Lerner, Amanda Xu

Reducing the Risk of Homelessness through Prioritized Distribution of Rental Assistance Resources
Presenters: Joe Baumann, Abigail Smith, Catalina Vajiac

Session 3: Sunday, September 11th (8:45 - 9:45 a.m.)
[Theme] Climate Change and Environmental Impacts
Session Chair: Jie Ie Bak (University of British Columbia)

Smell Vancouver, Connecting citizen science odor reports to emission sources (virtual)
Presenters: Divya Bilolikar, Jacob Hutton

Heating Loads in Alaska and Beyond
Presenters: Vidisha Chowdhury, Madelyn Gaumer, Philippe Schicker

Quantifying the impact of satellite streaks in astronomical images
Presenters: Abhilash Biswas, Kilando Chambers, Ashley Santos

Digitizing paper-based electrocardiogram (ECG) files to foster deep learning based analysis of existing clinical datasets: An exploratory analysis
Presenter: Habeeba Siddiqui

Session 4: Sunday, September 11th (10:00 - 11:00 a.m.)
[Theme] Housing, Poverty and Employment
Session Chair: Dan Richard (University of North Florida)

Microestimates of Multidimensional Child Poverty in West Africa
Presenters: John Fitzgerald, Marina Vicini, Daniela Pinto Veizaga

Improving services for homeless youth in the UK (virtual)
Presenters: Hannah Olson-Williams, Aryan Verma, Alina Voronina

Exploratory Measures for Analysis of Local Housing Needs
Presenter: Romina Tafazzoli

ADUniverse: Evaluating the Feasibility of (Affordable) Accessory Dwelling Units in Seattle
Presenters: Yuanhao Niu, Adrian Tullock

Exploring Public Datasets on Successful Employment for Persons with Disabilities in Iowa
Presenter: Harun Célik

PRESENTATIONS: Posters

Posters are numbered: 1-2 (for virtual only); 3-14 (for in-person & virtual)

1. Upstream Migration of River Herring (virtual only)
Presenters: Neeharika Karanam, Sora Ryu

2. Detecting Fake News on Twitter with Graph Neural Networks (virtual only)
Presenter: Jonathan Cook

3. Wholesale Local Food Benchmarking
Presenters: Nabil Idris, Nayha Hussain, Maxwell Skinner, Muskan Tantia

4. Building Households and Families out of Individual Level Administrative Data
Presenters: Zhoawen Guo, Eliot Stanton

5. Designing Relational Database System for the Self-Sufficiency Standard: An alternative method for representing the cost of living
Presenters: Azizakhon Mirsaidova, Priyana Patel, Cheng Ren, Hector Sosa

6. Feeding Northeast Florida: Finding Data-Driven Insights in the Fight Against Hunger
Presenter: Abhishek Singh

7. Identifying Communities with opportunities for positive youth development
Presenters: Abhishek Pancholi, Minakshi Sharma

8. Improving the National Suicide Prevention Lifeline's Service through Efficient Call Routing
Presenters: Charles Cui, Irene Tang

9. Targeting houses for retrofit in the West Midlands, UK
Presenters: Li-Lian Ang, Meghna Asthana, Mike Coughlan, Shriya Kamat Tarcar

10. Scaling up observations on plant phenology using remote sensing and machine learning
Presenter: Kailun (Lucas) Jin

11. Predicting Students at Risk of Becoming NEET (Not in Education, Employment or Training)
Presenters: Rachel Humphries, Pranjusmrita Kalita, Abhijeet Mulgund, Vanshika Namdev

12. Identifying damaged roofs in Baltimore City for demolition inspection
Presenter: Chae Won Lee

13. Evaluating the Effectiveness and Equity of Court Interventions to Reduce Involvement with the Criminal Justice System
Presenter: Meltem Ozcan

14. Gerrymandered? Using New Metrics to Investigate Old Standards
Presenter: Janani Thoguluva

ABSTRACTS: Talks

Session 1 - [Theme] Community-Health: Discovering Data and Modeling in Public Health

CoMID: Combining Content and User Data for COVID-19 Misinformation Alignment Detection
Presenter: Nazanin Jafari
Because users control whether misinformation on social media is amplified or diminished, an important approach to understanding and mitigating the spread of misinformation is by recognizing whether a given social media post aligns with the false information (i.e., agreeing with it) or tries to combat it by providing counter-information or disagreeing with it. In this paper, we present CoMID, a method that detects whether a tweet agrees or disagrees with a misinformation claim, based on the tweet content and the tweet author's propensity to spread misinformation. We calculate the propensity of the user based on their past tweets and profile description. We evaluate this method on our newly introduced dataset, "COVID-Myths", and compare it to existing state-of-the-art content-only and user & content based methods.

Determinants of COVID-19 Mortality at the U.S. County-Level
Presenters: Jiacheng Ge, Charles Hendrickson, Scout Leonard
More than one million Americans have died from COVID-19 since the start of the pandemic; because of challenges in access to COVID-19 testing and indirect deaths resulting from the pandemic, it has been shown that additional COVID-19 mortality is not accounted for. Excess death is a value that represents the difference between expected death outcomes and actual death outcomes for a given time and place. In 2020 and 2021, the COVID-19 pandemic contributed to increased global mortality that was not always accounted for in the COVID-19 mortality data. In this project, we model expected mortality at the U.S. county level for the 500 most populous U.S. counties to generate their excess death estimates. Furthermore, we use policy, health, vaccine, and socioeconomic features to investigate their effects on U.S. county-level excess mortality. Results from this project will reveal factors that contributed to high and low levels of excess death during the COVID-19 pandemic, which can provide guidance in policy implementation and improve pandemic response in the future.

A Multi-Label Multi-Class Approach to Predicting ICD-10 Codes from Unstructured Clinical Text Data
Presenters: Olumurejiwa Fatunde, Jennah Gosciak, Christina Last, Kennedy Odongo
Clinical decisions based on manual review of nurse and physician notes are often subjective (varying across physicians), inconsistent (varying for each physician over time), inefficient, and error-prone. Standardized disease classifications such as the International Classification of Diseases (ICD) can be used to standardize clinical workflows such as diagnosis, referral, and follow-up to improve patient care and outcomes. Typically, professional coders are employed to manually convert clinical notes to ICD codes; however, obtaining the required resources can be cost-prohibitive in low-resource settings. Furthermore, manual coding is at odds with the nature of operations in emergency rooms, where patient care is unscheduled and often time-sensitive. In collaboration with a non-profit hospital in Pakistan, we explore automated methods to convert nurse and physician notes into ICD-10 codes. We use methods from Natural Language Processing, formulating the problem as a multilabel-multiclass classification problem. Preliminary results based on recall indicate that model-based classification can identify at least 73% of the ICD-10 codes associated with a given patient visit, improving on the efficiency of manual classification and outperforming simpler automated coding methods such as frequency-based

prediction or minimum-textual-distance-based prediction.

Re-initiation of the Community Learning Through Data Driven Discovery Process: Revisiting Iowa's Alcohol Landscape

Presenter: Kelsey Van Selous

Alcohol is the most frequently misused substance in Iowa, with binge drinking rates 1.37 times higher than the national average, and alcohol related deaths on the rise (Governor's Office of Drug Control Policy, 2021; Iowa Department of Public Health, 2022). Excessive alcohol use is also linked to chronic illness, violence, automobile accidents, property damage, and cancer, costing the state nearly $2 billion a year (Sacks et al., 2015). Considering Iowa made record breaking revenue of over $415 million in liquor sales this year (Iowa Alcoholic Beverages Division, 2021), alcohol related harms are likely to increase in the next several years. This presentation is an example a Data Science for the Public Good project that has grown through re-initiation of the Community Learning Through Data Driven Discovery Process with new stakeholders and communities.

Supporting Communities and Community Organizations Through Data Science: A story of food rescue in central Iowa

Presenter: Matthew Voss

This Data Science for the Public Good at Iowa State project looks at the non-profit organization Eat Greater Des Moines (EGDM) and its food rescue efforts. EGDM takes donations of surplus food from grocery and convenience stores, restaurants, and other locations and transports it to food pantries, non-profits, schools, housing locations, and other organizations that can distribute food to those that need it. In the project, the team used data provided by EGDM and other sources to identity where food rescue currently happens, where it can be expanded, and what areas can benefit most from food rescue. The team also built a data pipeline and dashboard that is sustainable for EGDM and will be used by the organization moving forward to support their food rescue efforts. Successes from the project can be applied in other similar community environments.

Improving the National Suicide Prevention Lifeline's Service through Better Call Routing

Presenters: Charles Cui, Irene Tang

We partnered with Vibrant Emotional Health, which is the non-profit organization that routes the National Suicide Prevention Lifeline (NSPL)'s incoming calls among approximately 200 call centers across the nation. Based on the first six digits of the caller's phone number, also known as the "exchange code," the NSPL's routing table logic bounces callers around to different call centers until either their call gets picked up, or until they hang up without speaking to a counselor. We applied machine learning concepts to redesign the order of call centers assigned to each exchange code such that the routing logic would then maximize the percentage of calls that get picked up in a timely manner. First, we built models that predict the likelihood that a potential incoming call would get picked up by any given call center. Second, referencing the best-performing among these models, we generated the routing table logic that maximizes call answer rates across exchange codes. The NSPL is one of the largest crisis hotlines in the nation, and creating an efficient call routing system supports its clinical aim to reduce individuals' psychological distress and risk of suicide.

Session 2 - [Theme] Transparency, Accountability and Engagement with Civic Systems

Bridging the Gap Between Computational Tools and User Adoption for Social Good: A Case Study in Political Redistricting Problems

Presenters: Rowana Ahmed, Katherine Chang

Recent years have seen a proliferation of open source computational and data science tools available to a diverse user base, and this talk presents one example on how technical tools can be adopted for social good impact. The translation of technical tools for a broad user base to enable thoughtful adoption and impact is an important step towards bridging the data analytics-to-outcome "last mile" consideration. Developing Ensemble Methods for Initial Districting Plan Evaluation is a 2021 University of Washington Data Science for Social Good (DSSG) project that examined computational methods to evaluate proposed political district maps for outliers, which may indicate gerrymandering. Gerrymandering, the manipulation of legislative district boundaries for personal or partisan gain, is a fundamental threat to democracy, and a diverse set of stakeholders are involved in addressing its process and effects. Our DSSG project explored GerryChain, a Python library designed to study gerrymandering, through a series of detailed state-level case studies and the development of a comprehensive user's guide designed to increase access to the GerryChain toolset. The project deliverables aimed to democratize the process of redistricting by increasing participation in the process through empowering citizen groups, activists, and non-partisan map drawing commissions to use GerryChain to inform and assist with their specific local redistricting problems.

eiCompare: Making Every Vote Count

Presenters: Juandalyn Burke, Ari Decter-Frain, Hikari Murayama, Pratik Sachdeva

A more representative democracy benefits everyone. The Voting Rights Act (VRA) of 1965 was established to provide every citizen the right to vote in a fair electoral process. Specifically, the VRA prohibits discriminatory voting practices such as voting dilution from being enacted on racial and language minorities. Since the United States election process does not require collecting information about the race or ethnicity of the voting population in an election, it is hard to prove two of the criteria needed to conclude that voting dilution is occurring. The first criteria is Racially Polarized Voting (RPV). RPV occurs when groups of different races or ethnicities have divergent candidate preferences. The second criteria is that the racial majority is able to outvote and bloc the racial minority from electing their preferred candidate. eiCompare is an R software package that proves these criteria. This package uses aggregate data to infer individual racial and ethnic identifications and compares three methods of ecological inference (Goodman's ecological regression, King's ecological inference, and RxC) to detect RPV and determine if voting dilution is occurring. We added several features to the tool including: geocoding capabilities, improved functionality for estimating the racial identities of voters, improved visualization of ecological inference results, parallel processing, and the ability to conduct a performance analysis with historical voting data. With the new improvements of the eiCompare, we believe that the package may be more accessible to future voting rights cases and readily used to identify areas where racially polarized voting and voter dilution occur.

Identifying behavioral health conditions in Californian police records

Presenters: Grishma Bhattarai, Salil Goyal, Alexander Lerner, Amanda Xu

Law enforcement officers seriously injured/killed 3600+ people in California between 2016 and 2020. Journalists from the California Reporting Project have collaborated with lawyers and academics to request police reports about uses of force, in an effort to hold law enforcement accountable. Through

preliminary analysis, they have found that mental health or substance abuse is a common factor in cases of use of force by some jurisdictions. However, the analysis process is labor intensive and slow. Our aim is to assist journalists by automatically identifying police documents that contain mentions of behavioral health conditions. We separately analyzed narratives and forms. On forms, we processed key-value pairs using keyword search to identify relevant fields. For narrative reports, we created a text cleaning pipeline and then implemented deep clustering and transformer models in sequence to perform binary classification. We use text preprocessing, deep clustering, and transformer approaches to ultimately classify documents from three jurisdictions in California.

Reducing the Risk of Homelessness through Prioritized Distribution of Rental Assistance Resources

Presenters: Joe Baumann, Abigail Smith, Catalina Vajiac

Each year, close to 14,000 individuals face an eviction in Allegheny County. The current process is reactive, requiring those facing eviction to call, missing out on many who need help. In this project, we show that using machine learning models helps to make the process more proactive by giving assistance to those who will likely have the most need. We partnered with the Allegheny County Department of Human Services (ACDHS) to help them make more informed, data-driven decisions for distributing their limited resources (e.g., rental assistance) to reduce entry into homelessness. We believe that the results of this effort hold the potential to considerably improve the current processes and to allocate resources more efficiently and more equitable to ultimately minimize entry into homelessness in Allegheny County.

Session 3 - [Theme] Climate Change and Environmental Impacts

Smell Vancouver: Identifying regional clusters of citizen science reported odors using an interactive web application

Presenters: Divya Bilolikar, Jacob Hutton

Smell Vancouver is a community science initiative focused on mapping the odor profile of Metro Vancouver using citizen reports. We were interested in determining regional variability in reported odors in the study area, and if this variability was associated with differences in local industrial activity. We applied several different spatial clustering algorithms to identify regional clusters of odor reporting, then examined the reported odors and causes in those clusters. We present the results of the clustering analysis in a custom built web application, which is intended to provide citizen scientists the ability to query the data themselves. We hope that our results will provide community stakeholders the tools to understand the odor profile of their region and increase participation in future citizen science efforts.

Heating Loads in Alaska and Beyond

Presenters: Vidisha Chowdhury, Madelyn Gaumer, Philippe Schicker

Alaska - and the wider Arctic region - have experienced accelerated effects of climate change over the past decade and are in much greater need of decarbonization than the rest of the world. About 75% of the energy consumption in the Arctic region can be attributed to commercial and residential heating. However, the transition to renewable energy sources faces a major challenge in these areas - the absence of comprehensive and accurate heating load estimates. This project pioneers a geospatial-first methodology that combines remote sensing and machine learning techniques to quantify large-

scale Alaskan heating loads with a high granularity. It uses open-source geospatial datasets in Google Earth Engine (GEE) to extract building features such as height, area, year-of-built, heating and cooling degree days. These variables are used along with heating load projections from the AK Warm simulation software to train models that predict hourly heating loads on the Railbelt utility grid. By doing so, this project significantly informs the decarbonization efforts currently underway in Alaska and develops geospatial capacity in this space.

Quantifying the impact of satellite streaks in astronomical images
Presenters: Abhilash Biswas, Kilando Chambers, Ashley Santos
Artificial satellites in Low-Earth orbit (LEO) reflect the sun's light and leave bright streaks in astronomical images which negatively impacts astronomical research. The quantification and analysis of satellite streaks in astronomical images is crucial to understand the problem better, draw more attention towards it, and motivate policy actions. However, current efforts in this regard is largely limited to manual analyses of specific images by a few researchers. Our project aims to generalize this process by facilitating large scale analysis of diverse astronomical images containing streaks from telescopes around the world. We have created a python library called Satmetrics that can ingest a wide variety of images in FITS format, detect streaks in them, and return various properties of those streaks such as mean pixel intensity and width. Satmetrics is a starting point for generating information about satellite streaks that will help astronomers study the problem better, aid satellite operators in adopting better brightness mitigation strategies, and provide a firm evidence base for future policy actions.

Digitizing paper-based electrocardiogram (ECG) files to foster deep learning based analysis of existing clinical datasets: An exploratory analysis
Presenter: Habeeba Siddiqui
Recently, a deep learning model was developed and validated for detecting left ventricular dysfunction based on a standard 12-lead ECG. However, this model largely depends on the availability of digital ECG data: 10 s for all 12 leads sampled at 500 Hz stored as a numeric array. This limits the ability to validate or scale this technology to institutions that store ECGs as PDF or image files ("paper" ECGs). Methods do exist to create digital signals from the archived paper copies of the ECGs. The primary objective of this study was to evaluate how well the AI-ECG model output obtained using digitized paper ECGs agreed with the predictions from the native digital ECGs for the detection of low ejection fraction. To address this objective, deep learning models that utilizes digitized data from a 12-lead ECG snapshot were needed. Two models were evaluated, Model A using data from a single lead with full 10-s recording (lead II) only and Model B using data from 3 leads with 10-s recordings (leads II, V1 and V5) in addition to 9 leads with partial (2.5-s) recordings. In a test sample of 10 patients with varying ECG features, Models A and B obtained intraclass correlation coefficients of 0.95 (95% CI: 0.82 to 0.99) and 0.58 (95% CI: 0.00 to 0.87). In an exploratory examination of model diagnostic performance to detect low ejection fraction, Model A achieved and AUC of 0.71 while Model B achieved an AUC of 0.91. Our study demonstrates an agreement between deep learning model predictions obtained from digitized paper-based ECGs and native digital ECGs and provides some insight into potential expandability of ECG-based deep learning models including the importance of captured duration (10-s vs. 2-5-s recordings) and ECG vectors (precordial leads vs. limb leads).

Session 4 - [Theme] Housing, Poverty, and Employment
Microestimates of Multidimensional Child Poverty in West Africa
Presenters: John Fitzgerald, Marina Vicini, Daniela Pinto Veizaga

ABSTRACTS: Talks

Measurement and analysis of the geographic distribution of children in poverty allow governments and other organisations to both design novel policies to eliminate such poverty, and monitor the impact of implemented policies. However, estimations of child poverty are currently available only at country or state level. This paper creates a complete and publicly available set of micro-estimates of the distribution of child poverty across 9 low- and middle-income countries in West Africa at 5km2 resolution. Estimates of prevalence, severity and specific poverty dimensions such as water, sanitation, housing and education have been computed using DHS survey data as ground truth and applying Machine Learning models aggregating geographical, demographic and economic data. Prediction intervals are provided to facilitate responsible downstream use. These methods and maps provide tools to study, guide interventions, monitor and evaluate policies, and track the elimination of child poverty in low and middle income countries.

Improving Services for Homeless Youth in UK

Presenters: Hannah Olson-Williams, Aryan Verma, Alina Voronina

Youth homelessness is a social problem and homelessness services help young people to develop into independent individuals. We predict complexity score as a measure of support a young person would require and optimally allocate different interventions based on individual characteristics. To judge the impact of different interventions on the course and employability outcomes of young people, we use mediation analyses.

Exploratory Measures for Analysis of Local Housing Needs

Presenter: Romina Tafazzoli

Effective housing policy development requires awareness of the types and extent of local housing needs and the barriers to private sector investment. Accordingly, local decision-makers seek reliable and current data sources to help set their housing policy priorities. This process might face hardships due to time and budget limits, inflexible tabulation formats, lack of detail, and time sensitivity in data sources. In this project, we explored how much and what types of data are available to describe local housing markets and fill the gaps in demand estimation. We obtained, reorganized, categorized, and analyzed data sets from various housing-related sources and developed a dashboard of housing-related indicators summarized and visualized in a unified format.

ADUniverse after 3 years: From PoC (Proof of Concept) to Production

Presenters: Yuanhao Niu, Adrian Tullock

ADUniverse, the affordable housing tool developed by UW DSSG fellows in the summer of 2019, was formally launched by the Seattle government in September 2020. The production version maintained most of the core functionalities from the PoC version. Users can search their properties on the map and check the feasibility of building an accessory dwelling unit. Pre-approved plans and construction details were added to the public website as designers pitched in. Rather than providing financial calculators in the PoC, the city website opted to offer more descriptive evidence about affordability. We reviewed the evolutions of the ADUniverse project and the educational aspects of the DSSG program for its participants.

Exploring Public Datasets on Successful Employment for Persons with Disabilities in Iowa

Presenter: Harun Célik

Data discovery is a vital step in shaping the direction of pertinent policies for persons with disabilities in the public setting. The "Successful Employment for Persons with Disabilities in Iowa" project was designed to help policy-makers and public professionals in Iowa to better access information about Iowans with disabilities through data discovery, exploration, and analysis at the county level. The project additionally focused on assessing the impact of publicly funded services and expenditures to the employment and livelihood of persons with disabilities in Iowa. This information was ultimately presented on a web-page with interactive Tableau dashboards.

ABSTRACTS: Posters

1. Upstream Migration of River Herring (virtual only)

Presenters: Neeharika Karanam, Sora Ryu

Accurate and efficient stock assessment methods of commercially relevant fish species are extremely important toward sustainable fisheries management. Currently used manual technologies are highly inefficient, time consuming and not incredibly accurate. We propose our end-to-end platform that automates detection and counting of herring fish species moving upstream in image and video data for efficient fishery management. We achieved a 0.94 F1 score and the highest mAP score of 0.95 at mAP@0.5 for the detection model via fine-tuning YOLO with a dataset consisting of herring and non-herring fish species. The model also achieved a higher IoU score which makes tighter bounding boxes and improves detection during overlap of fishes.

2. Detecting Fake News on Twitter with Graph Neural Networks (virtual only)

Presenter: Jonathan Cook

Misinformation takes the form of a false claim under the guise of fact. It is essential to protect social media against misinformation by means of effective detection and analysis. To this end, we formulate misinformation propagation as a dynamic graph, allowing us to extract the temporal evolution patterns and geometric features of the propagation graph based on Temporal Point Processes (TPPs). TPPs provide the appropriate modelling framework for a list of stochastic, discrete events. In this context, that is a sequence of social user engagements. Furthermore, we forecast the cumulative number of engaged users based on a power law. Such forecasting capabilities can be useful in assessing the potential virality of specific misinformation. By jointly considering the geometric and temporal propagation patterns, our model has achieved comparable performance with state-of-the-art baselines on two well known datasets.

3. Wholesale Local Food Benchmarking

Presenters: Nabil Idris, Nayha Hussain, Maxwell Skinner, Muskan Tantia

Iowa is currently in need of a data process/platform that will provide more localized and up-to-date information on regional food systems, specifically information around price-points for local products, including specialty and niche crops. The Iowa State Farm Food and Enterprise Development (FEED) is frequently asked for benchmarks on pricing of products both in retail and wholesale spaces. This occurs both within feasibility studies for new farm and food businesses and market assessments, as well as appropriate price points for selling to schools and early care sites, hospitals, universities, and other institutions. There is a need for additional data and research on the potential sales point for these

wholesale products when many of our specialty crop producers across the state are operating in direct-to-consumer retail spaces. While data is available from the AMS and USDA (including the Agricultural Census), there is limited aggregation of sales for these products at the local level. While this data is supposed to be updated weekly, some producers have expressed that this has not been consistent and is too generalized and not consistent with what they are seeing.

4. Building Households and Families out of Individual Level Administrative Data
Presenters: Zhoawen Guo, Eliot Stanton
Administrative data is collected at an individual level, but poverty and many other social outcomes are measured on a household level. This research project focused on grouping individuals into households and families out of the Washington Merged Longitudinal Administrative Dataset (WMLAD), a compilation of administrative records from six state agencies. Using point-in-time and longitudinal approaches on address and last name data, we designed and implemented four different definitions of a household or family unit to capture a variety of household types reflecting our complex society. The definitions, code, and resulting household groupings will be used for future research at a household level. Additionally, this project restructured and organized relevant data from WMLAD into a relational database to make this administrative data user-friendly for future researchers.

5. Designing Relational Database System for the Self-Sufficiency Standard: An alternative method for representing the cost of living
Presenters: Azizakhon Mirsaidova, Priyana Patel, Cheng Ren, Hector Sosa
The Official Poverty Measure (OPM) sets eligibility for critical benefits (e.g., food assistance, child care subsidies, or housing vouchers). Many families, however, cannot afford their basic needs and are not considered “in need” by the OPM and cannot access these supports. The Self-Sufficiency Standard (SSS) was created by the Center for Women’s Welfare (CWW) at UW to provide an alternative to the OPM by defining the income working families need to meet their basic necessities without public or private assistance. However, the current data is spread across “state” and “year”, which makes it difficult for researchers to conduct deep analyses. The following project seeks to answer how we can store the SSS data to increase efficiency for stakeholders to manipulate the data, extract meaningful information, and conduct further analyses. Our team created a relational database using SQLAlchemy and Python to hold the SSS for the available states. Our database includes a primary table with the SSS based on the family household type and several secondary tables, such as the cost of broadband and cellphone(s). The research also aims to increase the transparency and accessibility of data for stakeholders with varying technical backgrounds through robust documentation.

6. Feeding Northeast Florida: Finding Data-Driven Insights in the Fight Against Hunger
Presenter: Abhishek Singh
As a part of the 2019 Florida Data Science for Social Good program, the DSSG team partnered with Feeding Northeast Florida (FNEFL), a leading food bank in the Northeast Florida region. FNEFL’s mission is to improve the quality of life of Northeast Florida by addressing food insecurity, poverty, and poor health through providing nutritious foods and other essential goods to those in need in collaboration with community partners. Given FNEFL’s vast region of operation and limited staff constraints, FNEFL approached FL-DSSG with questions like - which areas are in the most critical situation and based on the data how can FNEFL prioritize their efforts? With this regard, the DSSG

team helped consolidate the data by carrying out analysis to convert data at a census tract level to a zip code level to support FNEFL's reports with uniform comparison between demographic and distribution data to deliver data-driven insights like how does their distribution compare to census information about areas in need. The team also transformed distribution data from FNEFL's traditional SAP reports to analyze the impact made by FNEFL and identify opportunities. As an outcome, the FL-DSSG team helped FNEFL's grants team with an interactive dashboard to get area-specific insights to enrich grants. The team also supported FNEFL's strategy decisions by devising a ranking algorithm for critical zones(food deserts) with machine learning to prioritize efforts.

7. Identifying Communities with opportunities for positive youth development

Presenters: Abhishek Pancholi, Minakshi Sharma

GameFace seeks to be a place that transforms the prime time for juvenile crime into productive development spaces. The social problem that GameFace addresses are (1) lack of physical health, mentorship, and coaching for youths who are not getting enough exercise, (2) lack of educational funding for families due to poverty or single-parent homes, large class sizes in schools decreasing the individual attention given to students, and the drop-off learning/retention rate after students come from summer break, and (3) lack of character development and morality/integrity/ethics in youth. To identify the areas where GameFace can offer their programs, we collected data from census, Florida dept. of juvenile justice, Florida Dept. of Health, Duval county public schools, churches with food banks and Food Deserts and performed statistical analysis and created dashboards on tableau.

8. Improving the National Suicide Prevention Lifeline's Service through Efficient Call Routing

Presenters: Charles Cui, Irene Tang

We partnered with Vibrant Emotional Health, which is the non-profit organization that routes the National Suicide Prevention Lifeline (NSPL)'s incoming calls among approximately 200 call centers across the nation. Based on the first six digits of the caller's phone number, also known as the "exchange code," the NSPL's routing table logic bounces callers around to different call centers until either their call gets picked up, or until they hang up without speaking to a counselor. We applied machine learning concepts to redesign the order of call centers assigned to each exchange code such that the routing logic would then maximize the percentage of calls that get picked up in a timely manner. First, we built models that predict the likelihood that a potential incoming call would get picked up by any given call center. Second, referencing the best-performing among these models, we generated the routing table logic that maximizes call answer rates across exchange codes. The NSPL is one of the largest crisis hotlines in the nation, and creating an efficient call routing system supports its clinical aim to reduce individuals' psychological distress and risk of suicide.

9. Targeting houses for retrofit in the West Midlands, UK

Presenters: Li-Lian Ang, Meghna Asthana, Mike Coughlan, Shriya Kamat Tarcar

Our partners, the West Midlands Combined Authority (WMCA) and Pure Leapfrog, aim to combat climate change by increasing the energy efficiency of houses and usage of renewable energy sources. The problem is that information on the energy efficiency of houses only exists for 40% of the 1.2 million houses in the West Midlands, UK. We generated three key insights for our partners to strategise where and what kinds of interventions are most effective: (1) predict the energy efficiency and heating type of houses using a random forest and similarity quantification model, (2) estimate solar PV output from remote sensing data and (3) determine the impact of converting houses with non-electric heating to electric heating.

ABSTRACTS: Posters

10. Scaling up observations on plant phenology using remote sensing and machine learning
Presenter: Kailun (Lucas) Jin
Shifts in the timing of plant flowering are a key signal of an ecosystem's response to environmental change. This matters both in natural, agricultural, and urban settings: the timing of plant flowering affects the synchronization between plants and pollinators; it influences the exposure of crops to weather extremes during flowering, the most sensitive stage of development, with implications for later yield; and it impacts the timing of pollen allergies. A consistent and large-scale measure tracking the timing of flowering across years would be of great use for assessing risks of plant-pollinator desynchronization, risks of subpar crop yields, or shed light on the connection between allergies, urban vegetation, and climate. Current efforts, ranging from local to regional scales, do not quite achieve this goal: ground observations, while expanding, remain limited in their spatial and temporal coverage; process-based or statistical models are not flexible enough to capture local acclimation or adaptation to environmental factors; and remotely sensed estimates of plant development stages are often disconnected from the ground. Hence, crop studies still rely on a time-invariant harvesting calendar to infer the flowering period for crops. The goal of this project would be to create a large-scale, high-resolution proxy for the timing of flowering, by leveraging spectral reflectance data from satellite imagery, ground observations on plant phenology, and predictive models and approaches from machine learning.

11. Predicting Students at Risk of Becoming NEET (Not in Education, Employment or Training)
Presenters: Rachel Humphries, Pranjusmrita Kalita, Abhijeet Mulgund, Vanshika Namdev
Although the education system in the UK is flexible for post-16 students, a significant portion end up not in education, employment, or training (NEET). A NEET status is often a precursor for issues later in life such as unemployment, homelessness, and depression. To assist groups that can provide needed interventions, we developed methods to identify students who are likely to become NEET based on their educational history. Our tools will help these organizations prioritize their aid and ensure help goes to the highest risk students.

12. Identifying damaged roofs in Baltimore City for demolition inspection
Presenter: Chae Won Lee
The City of Baltimore has seen a roughly 50% population decline over the last 50 years which has left behind vacant buildings that have deteriorated. Roof damage is an acute form of blight that is especially difficult for the Department of Housing and Community Development to identify and address given its low visibility from the ground. We run several models (deep learning, pixel count, logistic regression, decision tree) on structured and unstructured (high resolution aerial pictometry) data to answer the question: every year, when new aerial data arrives, for all residential block lots in Baltimore City, can we identify the 1,000 blocklots with the highest level or roof damage to prioritize for demolition inspection in the next year?

13. Evaluating the Effectiveness and Equity of Court Interventions to Reduce Involvement with the Criminal Justice System
Presenter: Meltem Ozcan
Kansas City Missouri Municipal Court (KCMO-MC) has observed that probation terms assigned to

probationers are often left incomplete and a large number of probationers subsequently return to the court with new cases. In this project we aimed to help the court develop mechanisms to evaluate the outcomes and effectiveness of their interventions in order to reduce individuals' future involvement with the criminal justice system. Our approach to tackling this issue was two-fold: 1) setting up an infrastructure that allows the court to experiment with various probation conditions to test the effectiveness of their practices, and 2) building a machine learning pipeline that makes it possible to compare outcomes of pilot programs across different risk groups to evaluate the effectiveness and equity of the program. Here, we focus on predicting the risk of individuals receiving low intensity probation sentences returning to the court with a new case. Together, these components allow the court to determine which interventions work best for which individuals (or alternatively, do not work), and make the necessary adjustments to improve outcomes for the individuals in the system to increase probation completion rates and reduce recidivism.

14. Gerrymandered? Using New Metrics to Investigate Old Standards
Presenter: Janani Thoguluva
Gerrymandering is a multi-faceted issue that requires much research in order to understand, let alone solve. Our team analyzed numerical and geographical census data using statistical models in order to make sense of what U.S. Courts think is most important when ruling on redistricting cases. We built upon popular existing metrics as well as creating new ones to include in our model, based on aspects such as county and district lines, compactness, race, partisanship, and more. We examined the outcomes of our model through case studies to discuss the importance of each metric in context.

OTHER INFO

Scheduled Activities

We have planned a few hosted and non-hosted activities for you to experience prior to the start of the conference if you wish to participate.

On Friday, September 9th between 12:00 - 3:30 p.m. (PDT), you will have the option of attending the following activities hosted by the LDDG Conference staff:

- A walking tour of the University of Washington (UW) campus
- Access to the UW HUB Gaming Area (with options to play pool, bowl, and play games)
- A self-guided scavenger hunt across campus

Other Activities (arranged and purchased by you if interested)

You will have to arrange transportation and pay for these activities:

Seattle Bouldering Project - Fremont Location
3535 Interlake Ave N, Seattle, WA 98103
(206) 430-7757
Seattleboulderingproject.com

Day passes can be purchased at the front desk in the gym:

- Adult (24+) Day Pass : $20
- Young Adult (14 – 23) Day Pass: $18
- Youth (13 & under) Day Pass : $13

Agua Verde Paddle Club - kayak and paddle board rentals
1307 NE Boat St, Seattle, WA 98105 - Below on the water side
(206) 632-1862
Aguaverdepaddleclub.com

The rentals for kayaking and paddle boarding are on a first-come, first-serve basis. Please call the Agua Verde Paddle Club for any additional questions you may have.

Getting Around Seattle and the UW Campus

Sound Transit - for Seattle Metro bus routes and Link Light Rail train schedules, and for planning your travel around the city. Soundtransit.org

Rideshare services - Uber and Lyft are available in the city of Seattle. Download the Uber or Lyft app to access via your smart phone.
UW Transportation - The UW's Transportation Services provides info if you are planning to walk or bike around campus, as well as shuttles that are accessible around campus and to select destinations like South Lake Union. Transportation.uw.edu

Contact info for the conference hotels:

Residence Inn by Marriott, Seattle University District
4501 12th Ave NE, Seattle, WA 98105
(206) 322-8887

University Inn
4140 Roosevelt Way NE, Seattle, WA 98105
(206) 632-5055

Watertown Hotel
4242 Roosevelt Way NE, Seattle, WA 98105
(206) 826-4242