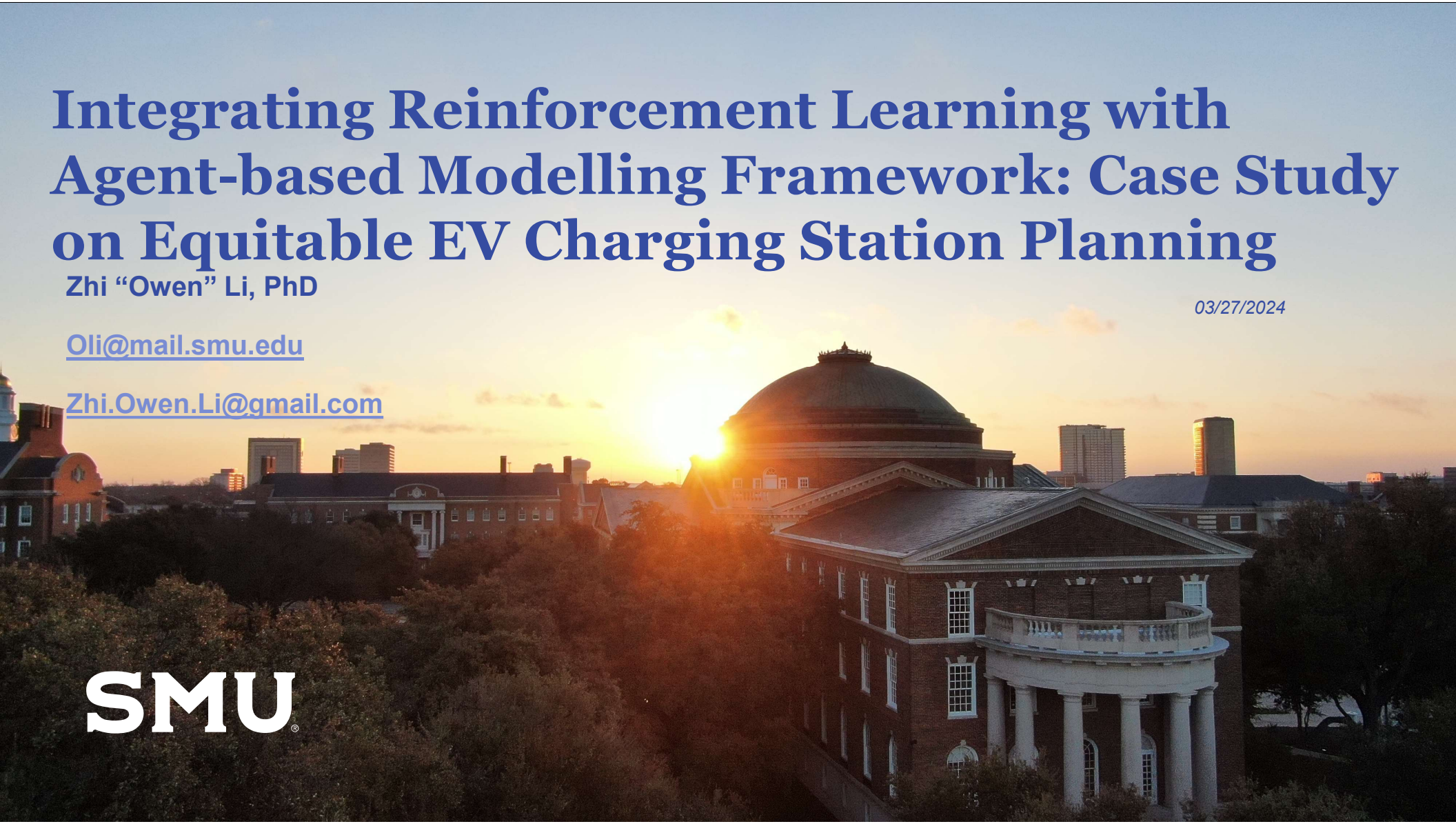
Integrating Reinforcement Learning with Agent-based Modelling Framework: Case Study on Equitable EV Charging Station Planning

Zhi "Owen" Li, PhD

Oli@mail.smu.edu

Zhi.Owen.Li@gmail.com

03/27/2024



Content

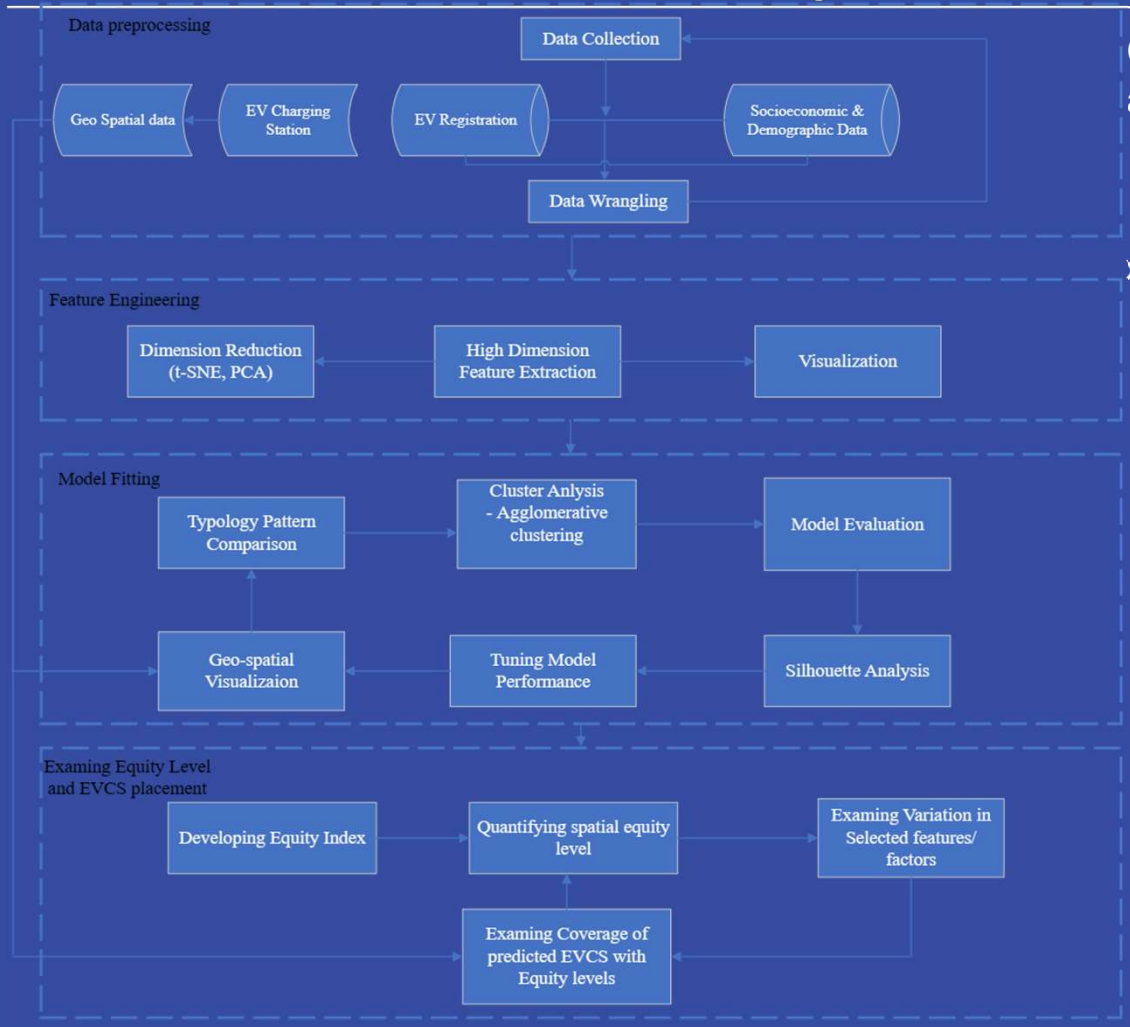
» Goal: to inform policymakers and stakeholders on enhancing equitable planning of EV infrastructures at community level.

» Sections:

- » Part I: A unified data-driven equity metric is proposed. Cluster analysis with a unified equity metric are conducted to discover typologies.
- » Part II: Reinforcement Learning is coupled with Agent-Based Modeling Framework. The proposed equity metric is employed in reward design.

» Part I: A Unified Multi-Criteria Equity Metric for Data-driven Decision Making: Case Study on Tract-level EV Charging Station Planning

Workflow for cluster analysis



Cluster analysis to find new typologies. And Policies are recommended based on each typologies.

- Data:
  - Before data wrangling, more than 400 features;
- Feature engineering:
  - After feature engineering DR, 1 principal component consists if 367 features
- Modeling fitting and evaluation:
  - Discover 3 typologies
  - Silhouette analysis for evaluation
- Examining Equity Level:
  - Data-driven Equity Metrics
  - Validation

SMU

Data-driven Unified Equity Indicator

» Why? Too many features, but most research focuses on 1 criterion

» 367 features !=> 367 criteria

» $EVCS\ Equity\ Indicator\ (EVCSEI) = tanh[(\alpha \times EVCS\_indicator - TCAI(var\_1, var\_2, \ldots . var\_k)]$

» It comprises two components: $\alpha \times EVCS\_indicator$ and $TCAI$.

» EVCS_indicator is a function of the number of EV (Electric Vehicle) charging stations

» Tract Composite Attribute Indicator (TCAI) is designed to represent a given tract's composite attribute

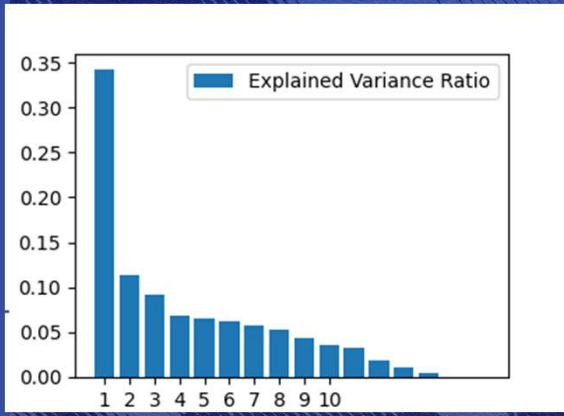
» The tanh function is the hyperbolic tangent function, which is used to map the input to a bounded range between -1 and 1.

» The alpha is a scaling factor that can be adjusted to control the sensitivity of the indicator.

Dimension Reduction and Tract Composite Attribute Indicator (TCAI)

» Tract Composite Attribute Indicator (TCAI): It is a composite measure related to the community's characteristics and attributes.

» In this study, we use PCA_PC1 (i.e. 1st principal component) as TCAI, other functions may also be considered. PCA_PC1 is derived from PCA (Principal Component Analysis) using 367 social and demographic features. PCA summarizes the variance in the data by finding linear combinations of the original features. Therefore, PCA_PC1 represents the most important dimension of variation in the data.



Explained Variance Ratio vs Principal component

Top disadvantage features to Tract Composite Attribute Indicator

Variable	Importance	Concept	Description
B28010_006E	0.160234	COMPUTERS IN HOUSEHOLD	Has one or more types of computing devices; Smartphone, tablet or other portable wireless computer or other computer; Smartphone, tablet or other portable wireless computer or other computer, no desktop or laptop
B08122_003E	0.143247	MEANS OF TRANSPORTATION TO WORK BY POVERTY STATUS IN THE PAST 12 MONTHS	100 to 149 percent of the poverty level
B15003_013E	0.136446	EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER	9th grade
B08124_006E	0.133283	MEANS OF TRANSPORTATION TO WORK BY OCCUPATION	Production, transportation, and material moving occupations
B15003_010E	0.124265	EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER	6th grade
B17026_005E	0.120407	RATIO OF INCOME TO POVERTY LEVEL OF FAMILIES IN THE PAST 12 MONTHS	1.00 to 1.24
B15003_017E	0.117345	EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER	Regular high school diploma
B08124_005E	0.116397	MEANS OF TRANSPORTATION TO WORK BY OCCUPATION	Natural resources, construction, and maintenance occupations
B08122_007E	0.114946	MEANS OF TRANSPORTATION TO WORK BY POVERTY STATUS IN THE PAST 12 MONTHS	Car, truck, or van - drove alone; 100 to 149 percent of the poverty level
B08124_013E	0.112518	MEANS OF TRANSPORTATION TO WORK BY OCCUPATION	Car, truck, or van - drove alone; Production, transportation, and material moving occupations
B28011_008E	0.111538	INTERNET SUBSCRIPTIONS IN HOUSEHOLD	No Internet access
B08124_012E	0.109789	MEANS OF TRANSPORTATION TO WORK BY OCCUPATION	Car, truck, or van - drove alone; Natural resources, construction, and maintenance occupations
B02001_007E	0.107057	RACE	Some other race alone
B19001_007E	0.102627	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2021 INFLATION-ADJUSTED DOLLARS)	$30,000 to $34,999
B28010_007E	0.101709	COMPUTERS IN HOUSEHOLD	No Computer
B15003_012E	0.100822	EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER	8th grade
B11001_004E	0.097628	HOUSEHOLD TYPE (INCLUDING LIVING ALONE)	Family households; Other family
B08126_018E	0.097342	MEANS OF TRANSPORTATION TO WORK BY INDUSTRY	Car, truck, or van - drove alone; Construction
B17026_006E	0.096159	RATIO OF INCOME TO POVERTY LEVEL OF FAMILIES IN THE PAST 12 MONTHS	1.25 to 1.49
B08126_003E	0.093237	MEANS OF TRANSPORTATION TO WORK BY INDUSTRY	Construction

Based on 1st principal component, we know how much each feature contributes to final results

SMU

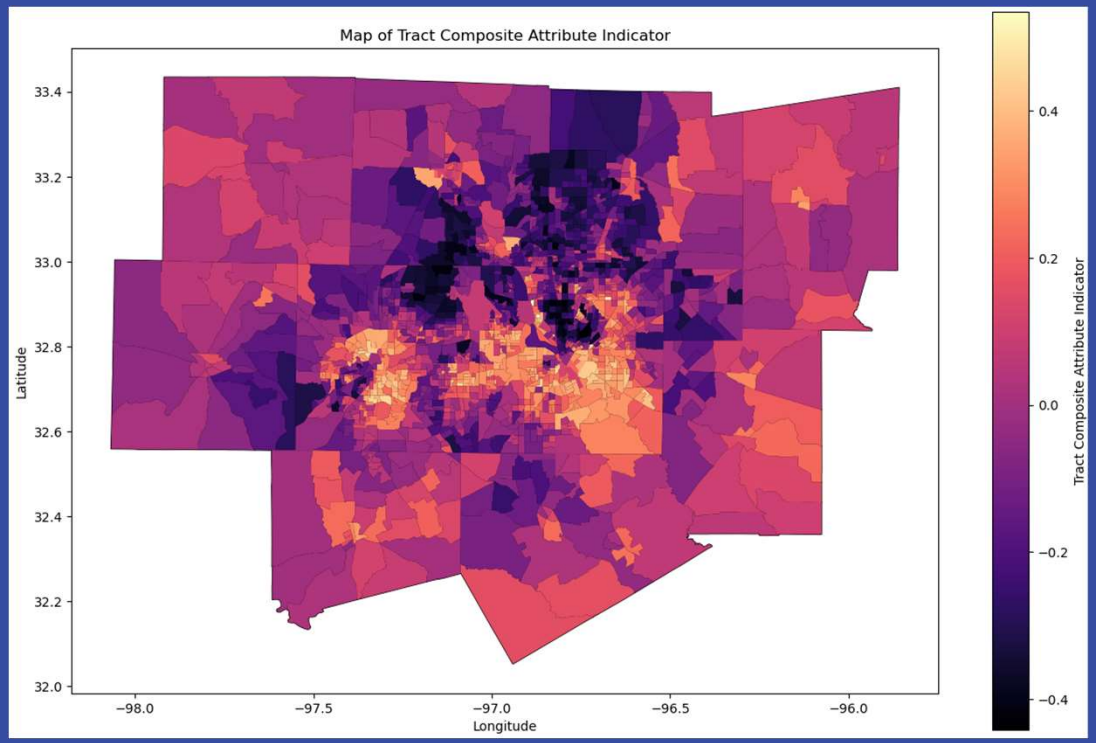
Top advantage features to Tract Composite Attribute Indicator

Variable	Importance	Concept	Description
B15003_022E	-0.169651	EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER	Bachelor's degree
B08124_009E	-0.156395	MEANS OF TRANSPORTATION TO WORK BY OCCUPATION	Car, truck, or van - drove alone: Management, business, science, and arts occupations
B19001_017E	-0.154168	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2021 INFLATION-ADJUSTED DOLLARS)	$200,000 or more
B17026_013E	-0.149859	RATIO OF INCOME TO POVERTY LEVEL OF FAMILIES IN THE PAST 12 MONTHS	5.00 and over
B08124_002E	-0.142268	MEANS OF TRANSPORTATION TO WORK BY OCCUPATION	Management, business, science, and arts occupations
B08126_009E	-0.140001	MEANS OF TRANSPORTATION TO WORK BY INDUSTRY	Finance and insurance, and real estate and rental and leasing
B08006_034E	-0.121073	SEX OF WORKERS BY MEANS OF TRANSPORTATION TO WORK	Male: Worked from home
B08122_028E	-0.118562	MEANS OF TRANSPORTATION TO WORK BY POVERTY STATUS IN THE PAST 12 MONTHS	Worked from home: At or above 150 percent of the poverty level
B08122_025E	-0.117678	MEANS OF TRANSPORTATION TO WORK BY POVERTY STATUS IN THE PAST 12 MONTHS	Worked from home
B08006_017E	-0.117366	SEX OF WORKERS BY MEANS OF TRANSPORTATION TO WORK	Worked from home
B08126_024E	-0.114856	MEANS OF TRANSPORTATION TO WORK BY INDUSTRY	Car, truck, or van - drove alone: Finance and insurance, and real estate and rental and leasing
B08124_044E	-0.112452	MEANS OF TRANSPORTATION TO WORK BY OCCUPATION	Worked from home: Management, business, science, and arts occupations
B19001_016E	-0.111902	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2021 INFLATION-ADJUSTED DOLLARS)	$150,000 to $199,999
B28011_004E	-0.107001	INTERNET SUBSCRIPTIONS IN HOUSEHOLD	With an Internet subscription: Broadband such as cable, fiber optic, or DSL
B28010_003E	-0.101950	COMPUTERS IN HOUSEHOLD	Has one or more types of computing devices: Desktop or laptop
B15003_023E	-0.101596	EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER	Master's degree
B08006_051E	-0.101146	SEX OF WORKERS BY MEANS OF TRANSPORTATION TO WORK	Female: Worked from home
B15003_024E	-0.093671	EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER	Professional school degree
B08126_099E	-0.092977	MEANS OF TRANSPORTATION TO WORK BY INDUSTRY	Worked from home: Finance and insurance, and real estate and rental and leasing
B25132_009E	-0.092244	MONTHLY ELECTRICITY COSTS	Charged for electricity: $250 or more

Leading to advantage

SMU®

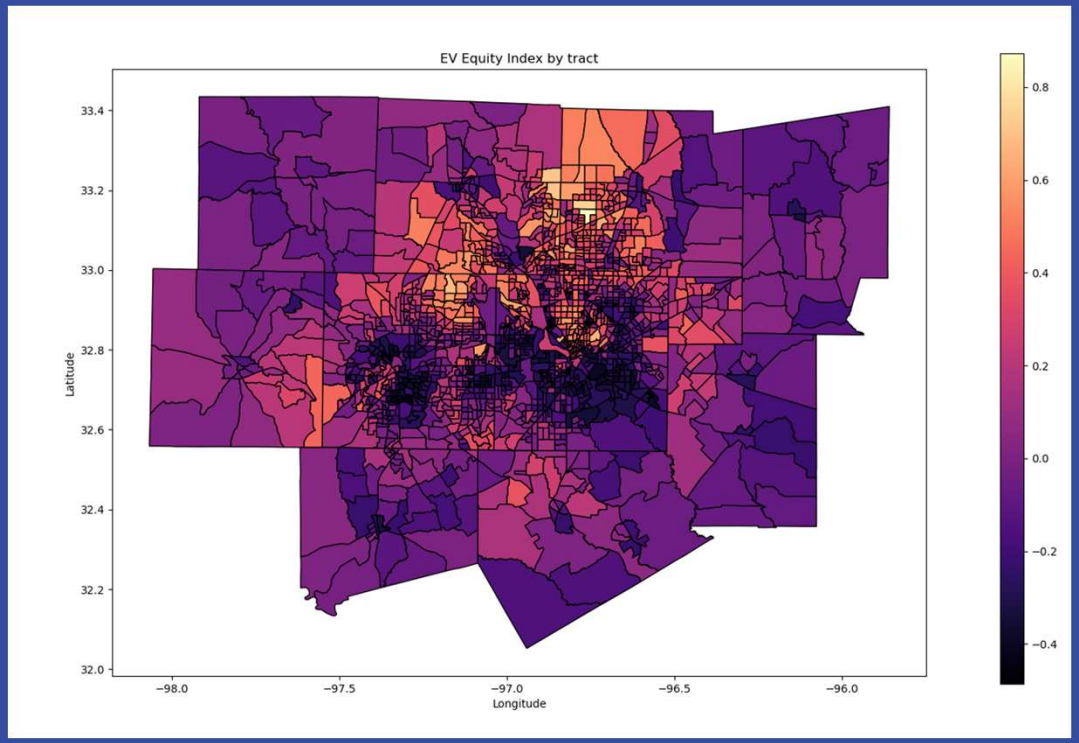
Tract Composite Attribute Indicator



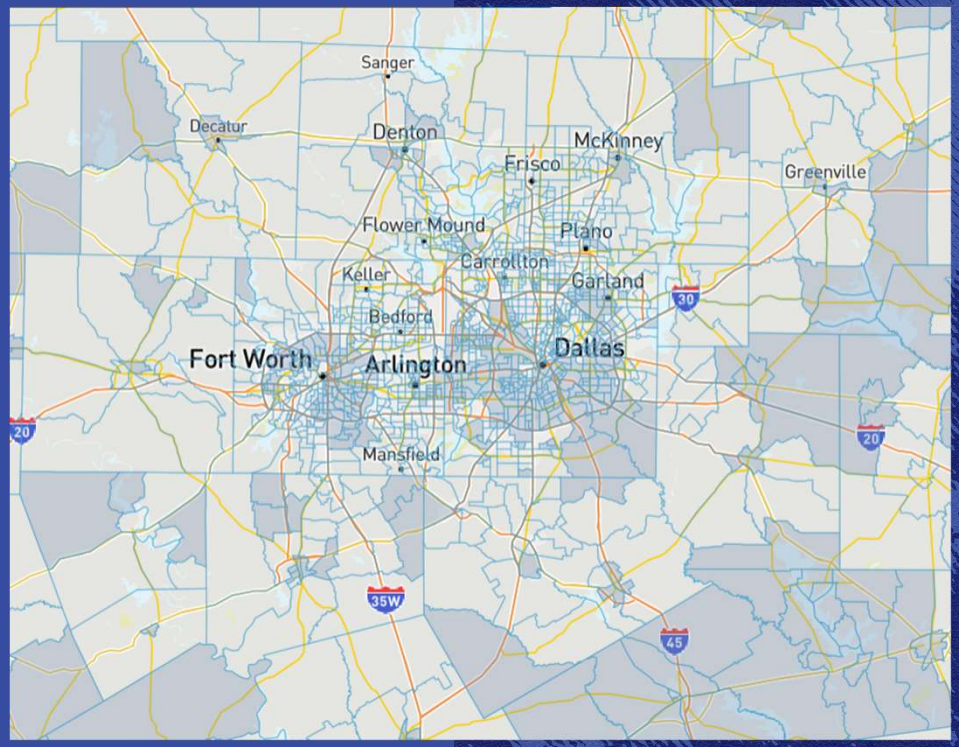
Visualization of Tract Composite Attribute Indicator

Tract Composite Attribute Indicator (TCAI) = 1st principal component of PCA

EVCS Equity Index- Comparable Evaluation



(a) EV Equity Indicator



(b) Disadvantaged Tract in Justice 40 Initiative

Tract Composite Attribute Indicator (TCAI) = 1st principal component of PCA

$$EVCS\ Equity\ Index\ (EVCSEI) = tanh[(\alpha \times EVCS\_index - TCAI(var\_1, var\_2, \ldots . var\_k)]$$

Cluster analysis – Discover Typologies

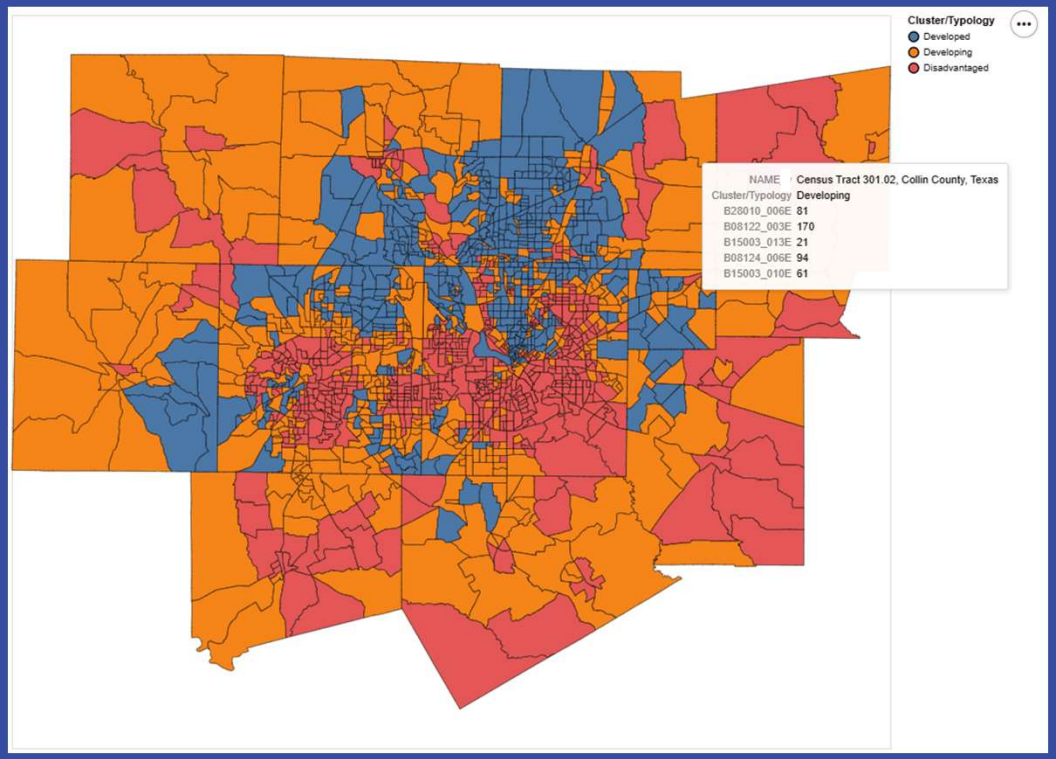
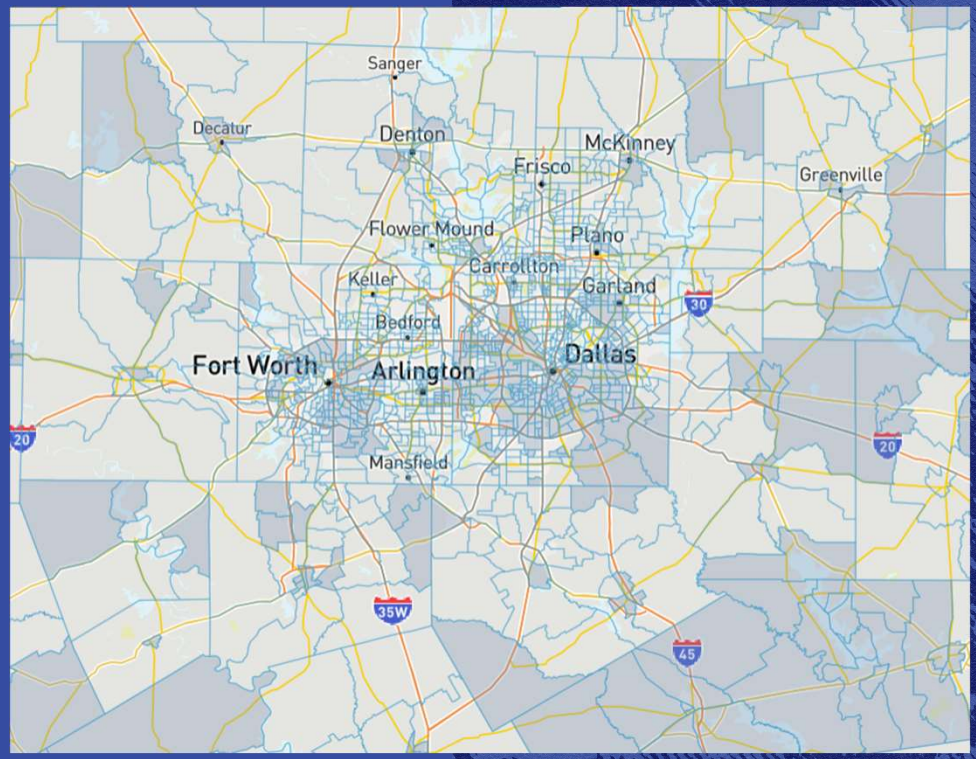


Fig. 10 Typologies (Interactive Map)



Disadvantaged Tract in Justice 40 Initiative

Developed /Advantaged Typology (Blue)
Developing Typology (Orange)
Disadvantaged Typology (Red)

SMU

Cluster analysis - Evaluation

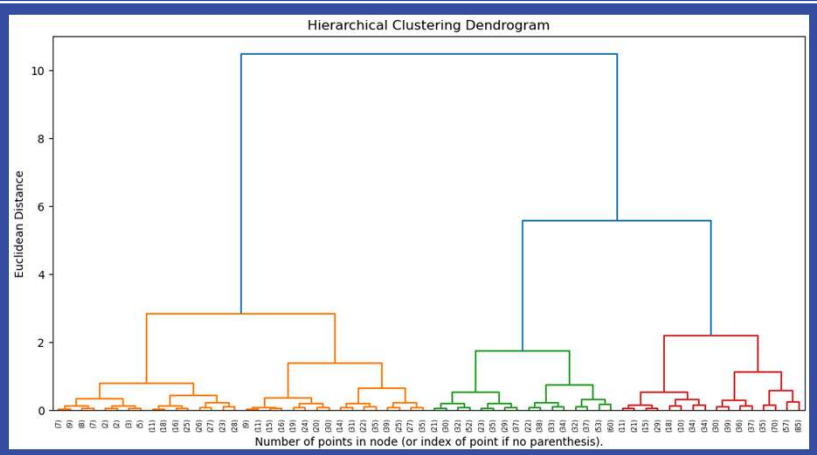


Fig. 7 Hierarchical clustering dendrogram

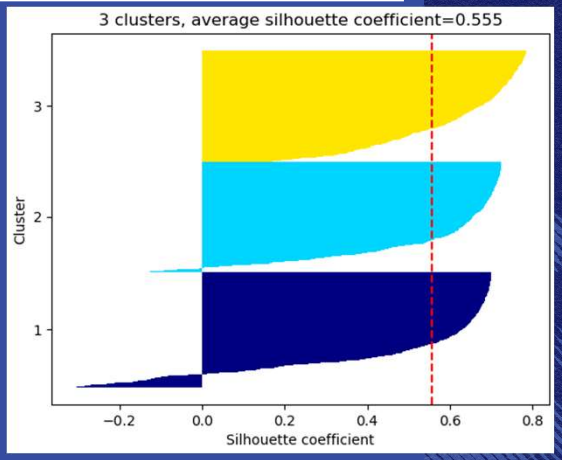


Fig. 8 Silhouette plots for 3 clusters (best results)

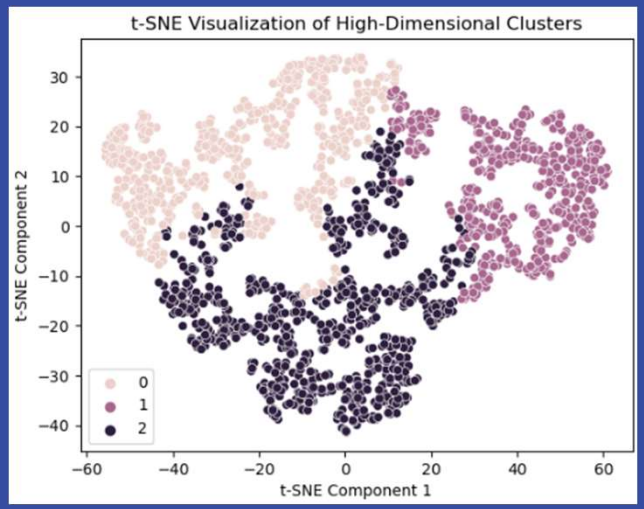
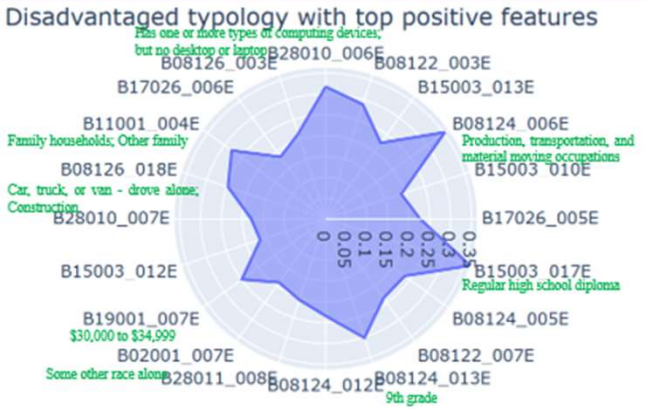
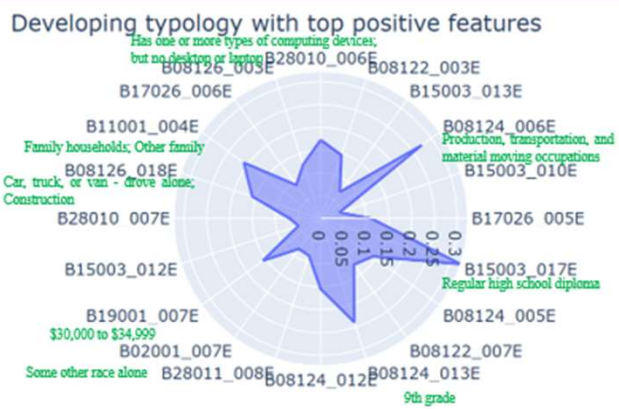
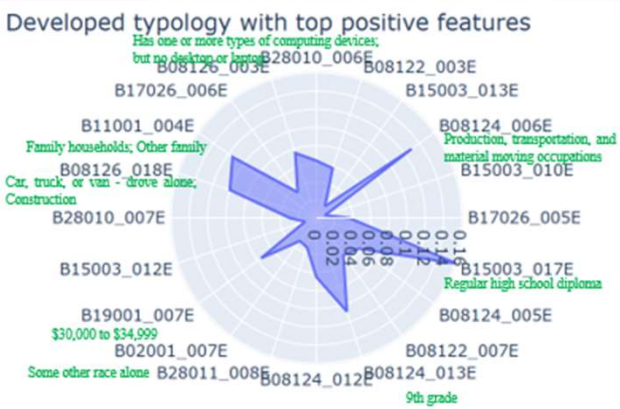


Fig. 9 Visualization of High-Dimensional Clusters

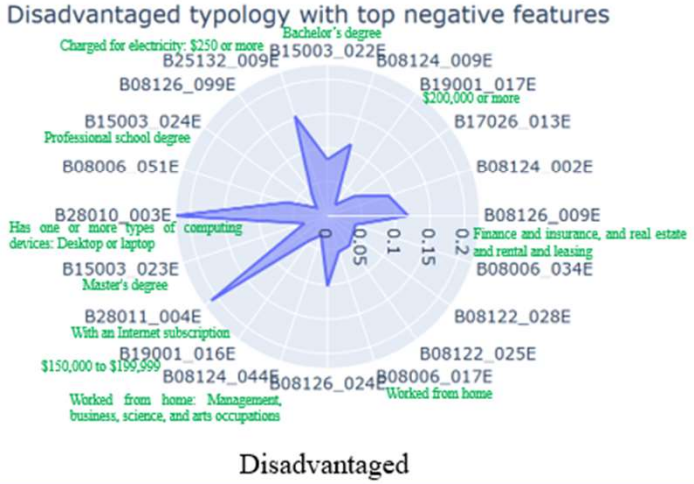
Radar Plot for each typology (Cluster)



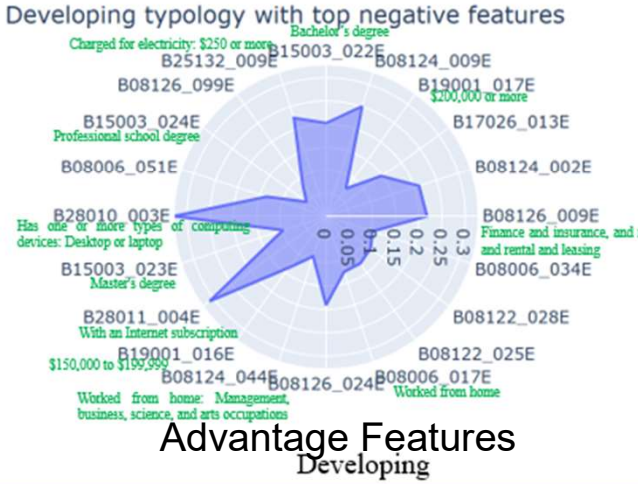




Disadvantage Features

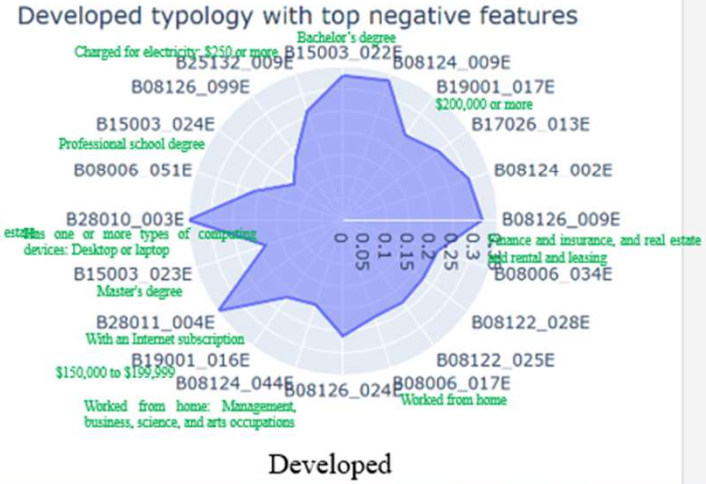


Disadvantaged



Advantage Features

Developing



Developed

Decision making for each typology

Typology	Count	EVCS count			Population			EV count			Accessibility	
		mean	std	sum	mean	std	sum	mean	sum	EV Adoption rate	Population Per EVCS	EV count Per EVCS
Developed	583	0.744	2.08	434	4527	1803	2639414	21.4	12454	0.47%	6081	13.6
Developing	528	0.718	2.07	379	4389	2021	2317389	64.0	33807	1.46%	6114	89.2
Disadvantaged	593	0.489	1.85	290	4362	1445	2586537	6.7	3944	0.15%	8919	11.2

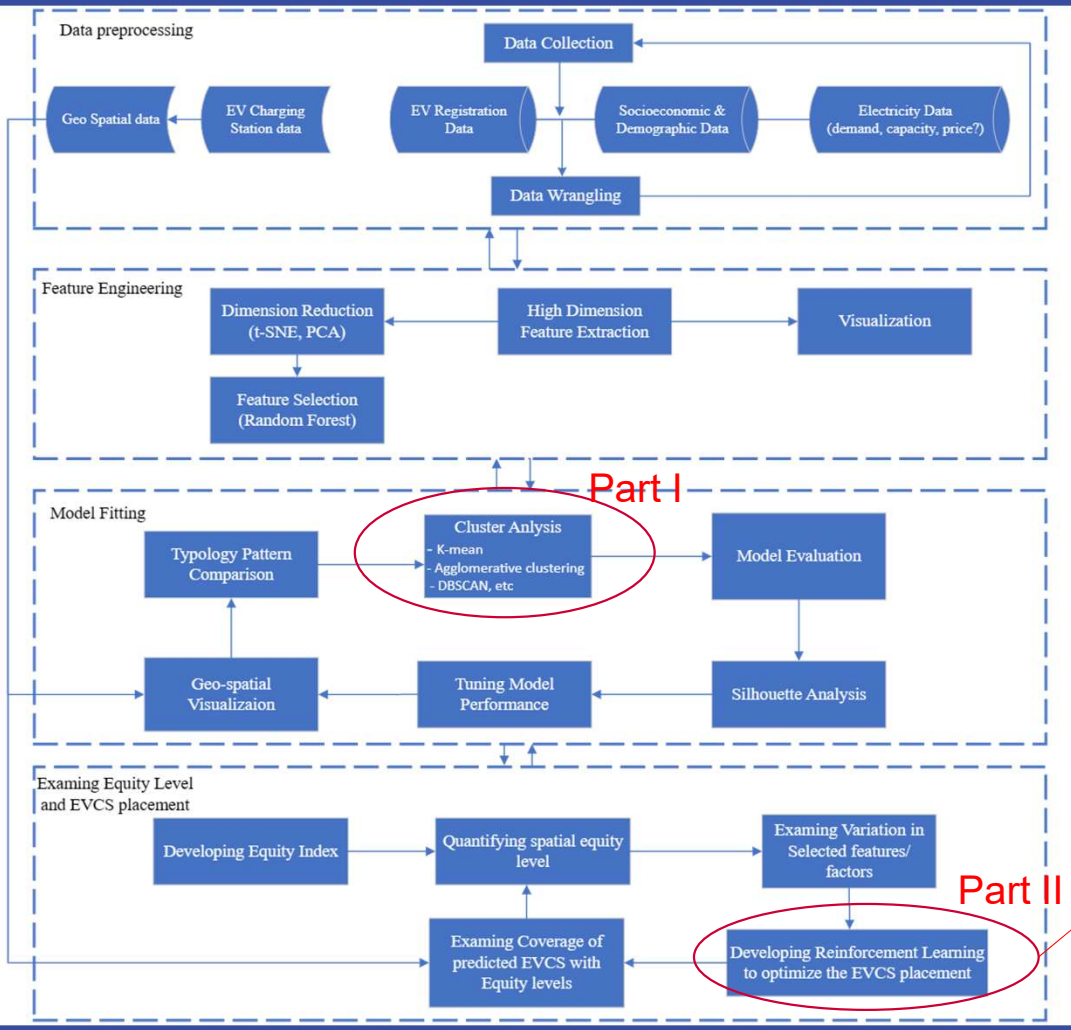
To improve the equity of EV adoption and EVCS planning:

- Developing communities, with the highest EV adoption rates but lower accessibility to EVCS, require additional EV infrastructure to accommodate their growing EV ownership.
- Disadvantaged communities, despite having sufficient accessibility to EVCS, exhibit the lowest EV adoption rates, necessitating prioritized efforts by policymakers to encourage EV registration and usage.
- Developed/advantaged communities, characterized by advantaged socioeconomic demographics, show adequate accessibility to EVCS. However, it exhibits low EV adoption rates, potentially influenced by factors such as remote work trends (as shown in radar plots) and concerns regarding EVs.[*]

SMU

- Part II: Integrating Reinforcement Learning with Agent-based Modelling Framework for Equitable EV Charging Station Planning
  - Part I: A novel data-driven Equity metric is proposed.
  - Part II: The proposed equity metric is employed in reward and states.

Decision Making Process Diagram

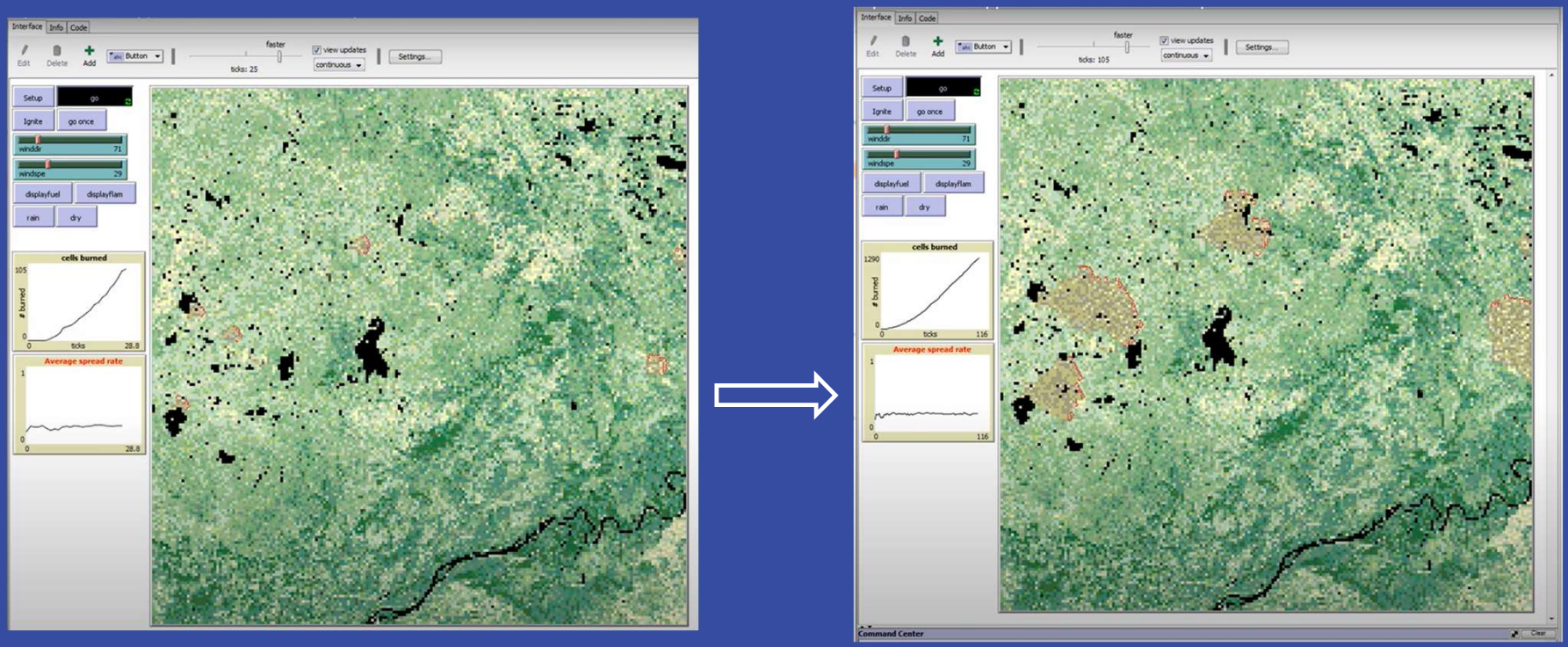


- Data Collection:
  - Dataset: Social Demographic dataset, Electricity dataset, EV registration dataset;
  - Geospatial mapping: Geospatial shapefile, EV charging station location;
- Feature Engineering: the input data are high dimensional; we want to extract the most impactful features/attributes. We deploy different dimension reduction techniques, to transform and select the features per the importance factor.
- Modelling fitting:
  - conduct the cluster analysis with consideration of the top important features;
  - Evaluate the clustering results via silhouette analysis;
  - Tuning hyperparameters to improve the results;
  - Visualize the clusters in the maps;
  - Compare different clustering results via patterns;
- Quantifying spatial equity level:
  - comparison between typologies pattern and EVCS placements;
  - develop equity metrics;
  - visualize the variations in equity level
- Develop reinforcement learning algorithms to optimize the decision making:
  - Design rewards and states based on the equity metrics;
  - Simulate the decision making process;
  - Discover and understand underlying trending, and make recommendation for decisionmakers

SMU

What is Agent Based Modelling

Agent-based modeling (ABM) is a computational modeling technique used to simulate the behavior of complex systems by representing individual agents and their interactions within an environment. In ABM, agents are autonomous entities with specific behaviors, attributes, and decision-making capabilities. These agents can interact with each other and their environment, leading to emergent phenomena at the macroscopic level.



Example - Wildfire Spread

SMU

Classic Agent based modeling framework

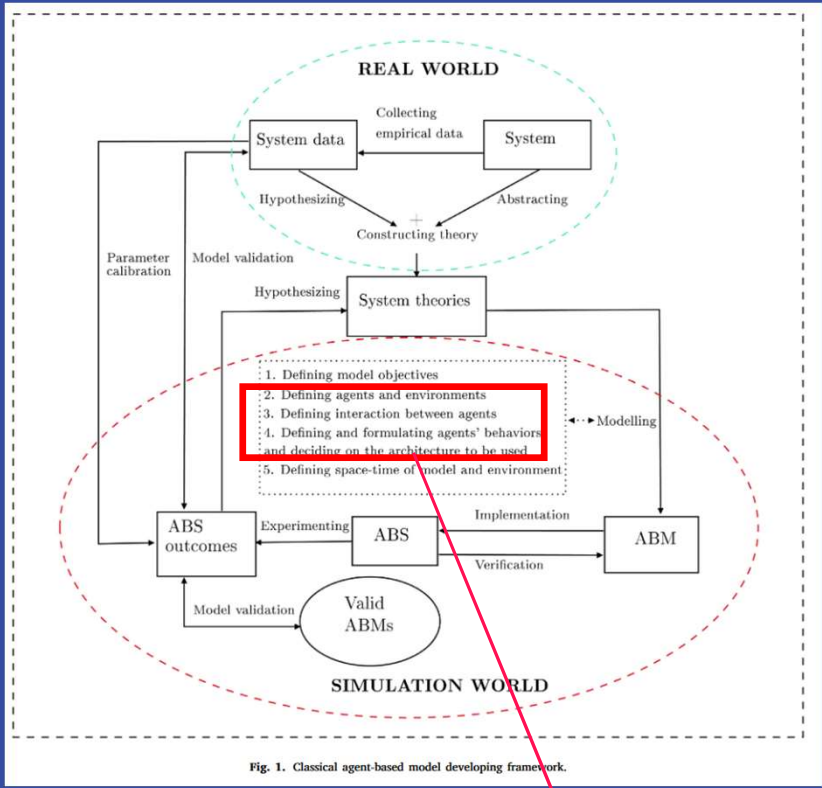
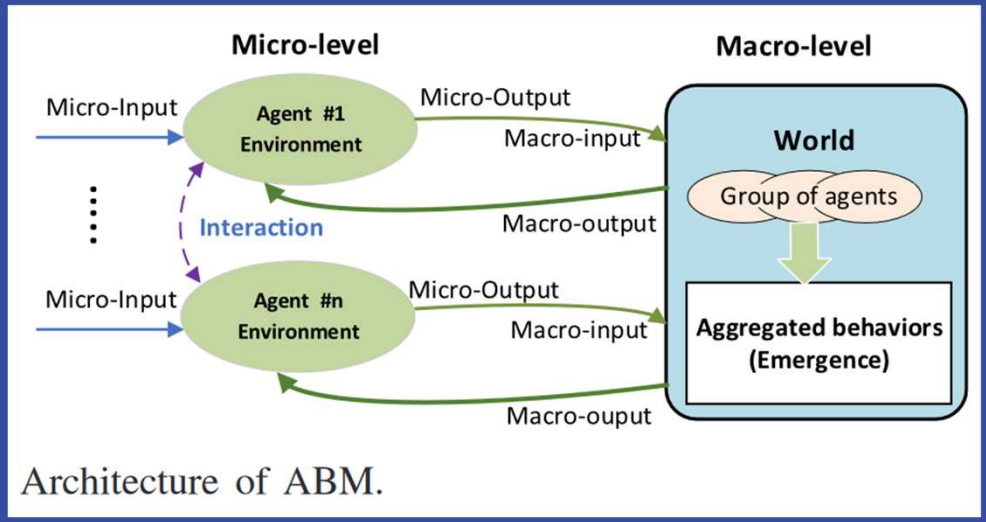


Fig. 1. Classical agent-based model developing framework.



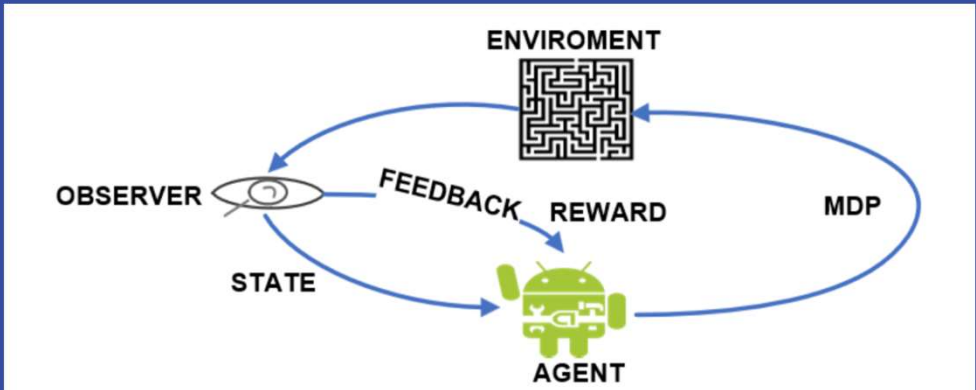
Architecture of ABM.

Keypoint: Define physic/theory-based rules/policies

SMU

What is Reinforcement Learning

» RL aims to use observations from interaction with the environment to take actions that will maximize the so-called cumulative reward. RL allows agents learn to automatically determine the optimal behavior under a specific context to optimize its performance. The reward feedback is required for the agent to learn its behavior in the RL algorithm.



Reward feedback of RL; MDP is Markov decision process

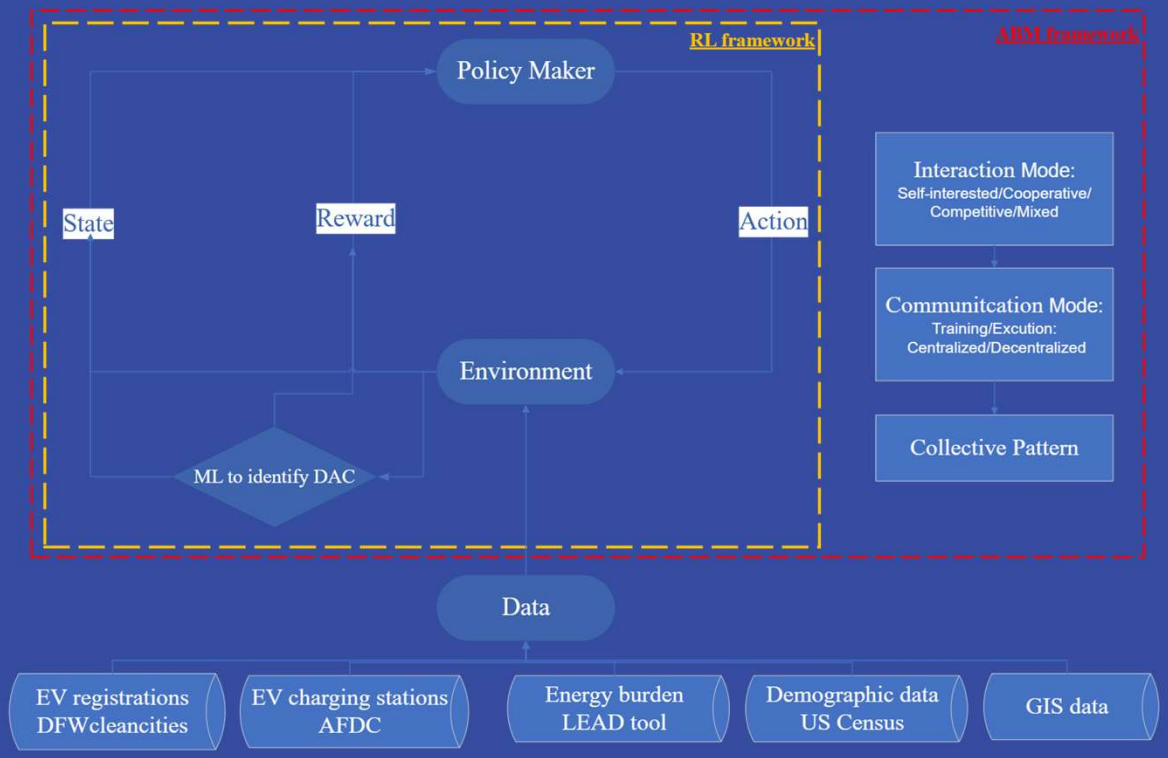
SMU

Agent-Based Modeling vs Reinforcement Learning

Aspect	Agent Based Modeling (Theory-driven)	Reinforcement Learning (Data-driven)
Focus	Focus on agents' collective behavior generating system behavior	Focus on individual agent
Interpretability	Each step and parameter interpretable, based on theories and laws	Parameters often uninterpretable, lacking physical significance
Factors influencing phenomena	Use factors known to influence social, physical, or natural phenomena	Capable of discovering latent factors influencing outcomes
Computational requirements	Iteration always required (computationally expensive)	Iteration primarily required during training (faster for prediction)
Data requirements	Requires less data	Requires large amounts of data
Sensitivity to missing data or attributes	Sensitive to missing data or attributes	Not sensitive to missing data
Model type	Model-based	Model-based and model-free
Agent's behavior	Deterministic and robust policies required (One-off models)	Agent learns complex policies in high-dimensional action spaces (Flexible and can be reused)

Integrating Reinforcement Learning with ABM Framework

- Agents will learn how to make decision to improve the disadvantaged community by interaction with environment
- The collective/aggregate behavior will help understand the global/systematic phenomenon/trend/mechanism



RL:
Train each agent (policymaker) to take actions

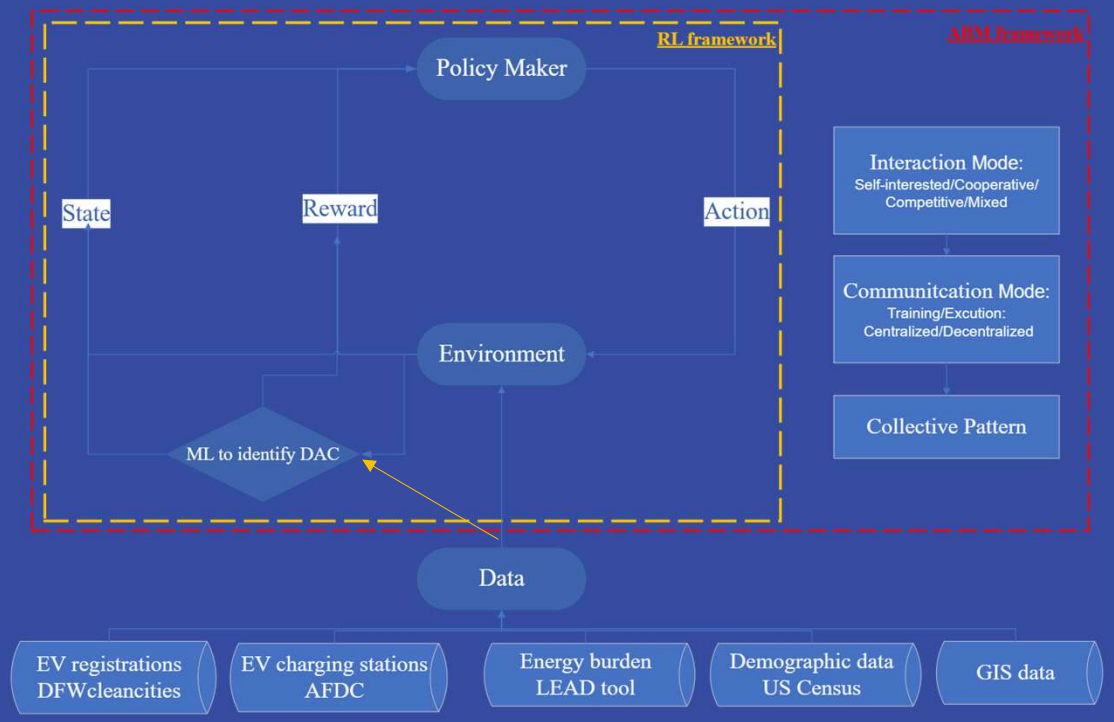
ABM:
Collect agent's behavior and interaction

- Community Policymaker to make decision
- Action impacts the environment
- Environment consists of two parts
  - Unchangeable Attributes of community
  - State of community

SMU

Custom Environment with OpenAI Gym

- Actions:
  - Discrete(6)
  - 0: Maintain current status;
  - 1: EV+1;
  - 2: EVCS+1;
  - 3: Energy Burden: -1*0.01%;
  - 4: Total households: +10;
  - 5: Education: +1*1%



After each action, the agent's current state, along with other unchangeable attributes (GIS, Race population, etc), is inputted into the Typology Machine Learning (ML) module. The output is the isDAC (Disadvantaged Communities).

SMU

Custom Environment with OpenAI Gym

» Observations(State):

» composite_space: combination of discrete and continuous components

» Dict({ "isDAC": Discrete(2),

» "EV": Discrete(1000),

» "EVCS": Discrete(100),

» "EnergyBurden": Box(0, 10, shape=(1,)),

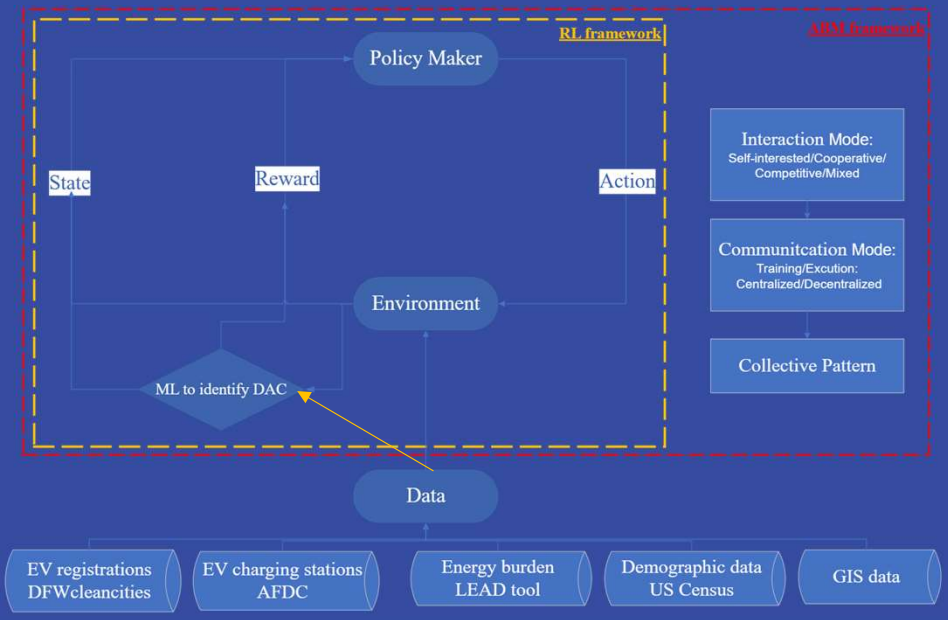
» "Households": Discrete(8000),

» "Education": Box(0, 100, shape=(1,))})

» Terminal State:

» When isDAC is False

» Meaning the community is not disadvantaged



After each action, the agent's current state, along with other unchangeable attributes (GIS, Race population, etc), is inputted into the Typology Machine Learning (ML) module. The output is the isDAC (Disadvantaged Communities).

SMU

Reward Design

- Rewards:
  - ML module based on proposed equity metrics is employed to identify three typologies (or DAC)
    - Input: EV, EVCS, Energy, Education, social demographic, and other data.
    - Output: isDAC
  - If not isDAC: +1000
  - If isDAC: -1
- Discount Factor(Gamma):
  - Smaller gamma makes agent move to the terminal state faster

$$U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot R_{t+2} + \cdots + \gamma^{n-t} \cdot R_n$$

SMU

Value vs Policy-based Learning

Value-based Learning (DQN)

state s

Conv

feature

Dense

$Q(s, \text{“left”}; \mathbf{w}) = 2000$

$Q(s, \text{“right”}; \mathbf{w}) = 1000$

$Q(s, \text{“up”}; \mathbf{w}) = 3000$

Action-Value Function

Policy-based Learning

state s

Conv

Dense

Softmax

“left”, 0.2

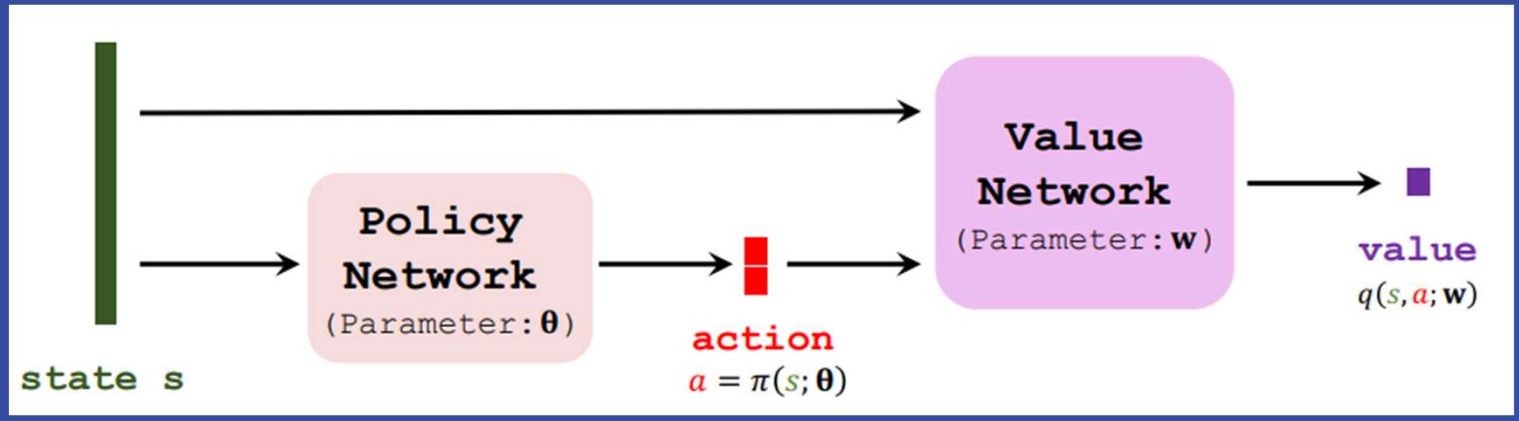
“right”, 0.1

“up”, 0.7

Probability Mass Function of Action

SMU

Discrete vs Continuous Control



Add a policy network(actor) in between. It output continuous action.

Followed by value network(Critic), it output action-values function

SMU

Temporal Difference (TD) Learning

- Observe a transition (s_t, a_t, r_t, s_{t+1}).
- TD target: $y_t = r_t + \gamma \cdot \max_a Q(s_{t+1}, a; \mathbf{w})$.
- TD error: $\delta_t = Q(s_t, a_t; \mathbf{w}) - y_t$.
- Update: $\mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \delta_t \cdot \frac{\partial Q(s_t, a_t; \mathbf{w})}{\partial \mathbf{w}}$.

TD methods update value estimates after each time step, rather than waiting until the end of an episode as in Monte Carlo methods.

- Shortcoming 1: Waste of Experience
  - discard a transition after using it. It is a waste
- Shortcoming 2: Correlated Updates
  - Consecutive states, s_t and s_{t+1}, are strongly correlated

- Solution: experience replay

TD with Experience Replay

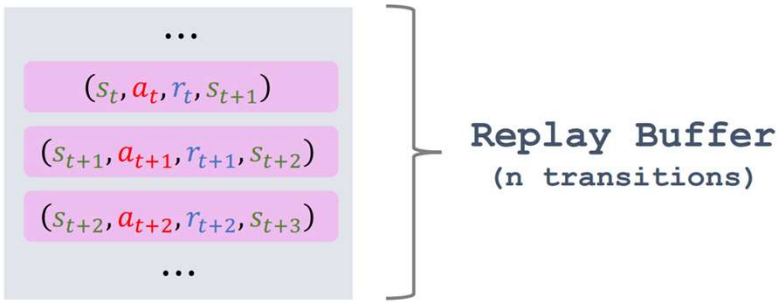
- Find $\mathbf{w}$ by minimizing $L(\mathbf{w}) = \frac{1}{T}\sum_{t=1}^{T} \frac{\delta_t^2}{2}$.
- Stochastic gradient descent (SGD):
  - Randomly sample a transition, (s_i, a_i, r_i, s_{i+1}), from the buffer.
  - Compute TD error, δ_i.
  - Stochastic gradient: $\mathbf{g}_i = \frac{\partial \delta_i^2/2}{\partial \mathbf{w}} = \delta_i \cdot \frac{\partial Q(s_i, a_i; \mathbf{w})}{\partial \mathbf{w}}$
  - SGD: $\mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \mathbf{g}_i$.

SMU

Experience Replay

Experience Replay

- **A transition:** (s_t, a_t, r_t, s_{t+1}) .
- Store recent n transitions in a **replay buffer**.



Experience Replay

- **A transition:** (s_t, a_t, r_t, s_{t+1}) .
- Store recent n transitions in a **replay buffer**.
- Remove old transitions so that the buffer has at most n transitions.
- Buffer capacity n is a tuning hyper-parameter [1, 2].
  - n is typically large, e.g., $10^5 \sim 10^6$.
  - The setting of n is application-specific.

SMU

Target Network & Double DQN

» In RL, bootstrapping means “using an estimated value in the update step for the same kind of estimated value”.

» TD learning use bootstrapping making DQN overestimate action-values. (Why?)

» • Problem: DQN trained by TD overestimates action-values.

» • Solution 1: Use a target network to compute TD targets. (Address the problem caused by bootstrapping.)

» • Solution 2: Use double DQN to alleviate the overestimation caused by maximization.

Use a transition, (s_t, a_t, r_t, s_{t+1}), to update $\mathbf{w}$.

- TD target: $y_t = r_t + \gamma \cdot \max_a Q(s_{t+1}, a; \mathbf{w})$.
- TD error: $\delta_t = Q(s_t, a_t; \mathbf{w}) - y_t$.
- SGD: $\mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \delta_t \cdot \frac{\partial Q(s_t, a_t; \mathbf{w})}{\partial \mathbf{w}}$.

Problem of Overestimation

- TD learning makes DQN overestimate action-values. (Why?)
- **Reason 1:** The maximization.
  - TD target: $y_t = r_t + \gamma \cdot \max_a Q(s_{t+1}, a; \mathbf{w})$.
  - TD target is bigger than the real action-value.
- **Reason 2:** Bootstrapping propagates the overestimation.

Target Network

- Target network: $Q(s, a; \mathbf{w}^-)$
  - The same network structure as the DQN, $Q(s, a; \mathbf{w})$.
  - Different parameters: $\mathbf{w}^- \neq \mathbf{w}$.
- Use $Q(s, a; \mathbf{w})$ to control the agent and collect experience:

$$\{(s_t, a_t, r_t, s_{t+1})\}.$$

- Use $Q(s, a; \mathbf{w}^-)$ to compute TD target:

$$y_t = r_t + \gamma \cdot \max_a Q(s_{t+1}, a; \mathbf{w}^-).$$

Update Target Network

- Periodically update $\mathbf{w}^-$.
- Option 1: $\mathbf{w}^- \leftarrow \mathbf{w}$.
- Option 2: $\mathbf{w}^- \leftarrow \tau \cdot \mathbf{w} + (1 - \tau) \cdot \mathbf{w}^-$.

SMU

exploration vs exploitation

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

model: $p(s', r|s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\},$

transition function

$$p(s'|s, a) \doteq \Pr\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r|s, a).$$

reward function $r(s, a, s') \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathcal{R}} r \frac{p(s', r|s, a)}{p(s'|s, a)}.$

Deep Q –Learning/DQN

Action space type summary for RL algos

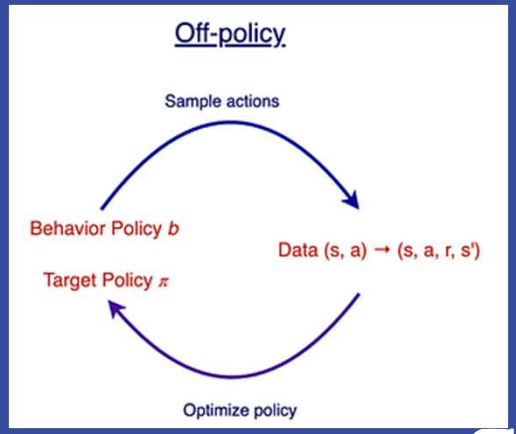
Name	Box	Discrete	MultiDiscrete	MultiBinary	Multi Processing
ARS 1	✓	✓	✗	✗	✓
A2C	✓	✓	✓	✓	✓
DDPG	✓	✗	✗	✗	✓
DQN	✗	✓	✗	✗	✓
HER	✓	✓	✗	✗	✓
PPO	✓	✓	✓	✓	✓
QR-DQN 1	✗	✓	✗	✗	✓
RecurrentPPO 1	✓	✓	✓	✓	✓
SAC	✓	✗	✗	✗	✓
TD3	✓	✗	✗	✗	✓
TQC 1	✓	✗	✗	✗	✓
TRPO 1	✓	✓	✓	✓	✓
Maskable PPO 1	✗	✓	✓	✓	✓

DQN

Space	Action	Observation
Discrete	✓	✓
Box	✗	✓
MultiDiscrete	✗	✓
MultiBinary	✗	✓
Dict	✗	✓..

» Model free

» Off policy



SMU

Why DQN

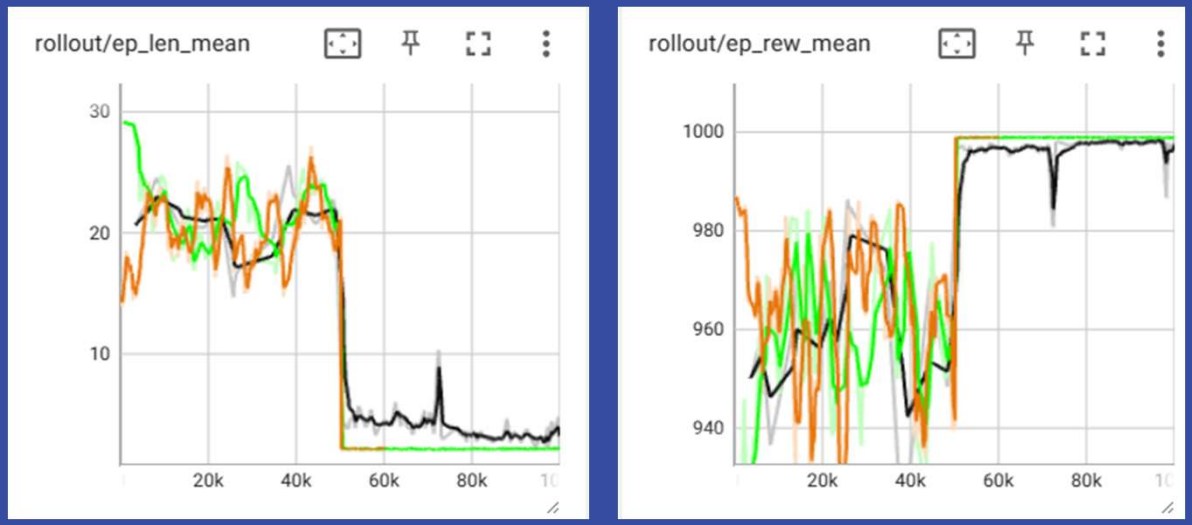
DQN is the most popular and well-established algorithm in reinforcement learning

1. **Versatility**: DQN can handle a wide range of tasks, from simple grid-world problems to complex real-world applications. Its flexibility makes it suitable for various domains, including gaming, robotics, finance, and more.
2. **Efficiency**: DQN employs **deep neural networks to approximate the Q-values**, enabling it to handle **high-dimensional state spaces efficiently**. This makes it suitable for tasks with complex observations.
3. **Sample Efficiency**: DQN utilizes **experience replay**, where past experiences are stored in a replay buffer and sampled randomly during training. This technique improves sample efficiency by reducing the correlation between consecutive samples and stabilizing training.
4. **Off-Policy Learning**: DQN uses an **off-policy learning** approach, meaning it learns from a separate behavior policy while still updating its target policy. This allows for more stable and efficient learning compared to on-policy methods.
5. **Temporal Difference Learning**: DQN leverages temporal difference learning to **update its Q-values iteratively at each step of the episode** based on the difference between predicted and target Q-values. This enables it to learn from delayed rewards and make long-term decisions.
6. **Deep Learning Capabilities**: DQN benefits from **the expressive power of deep neural networks**, allowing it to learn complex mappings between states and actions. This enables it to generalize well across different states and situations.

Hyperparameter Tuning Example:

» Fine tune one agent's hyperparameters

» Use the best hyperparameters combination for other agents



Run	Smoothed
DQN\lr0.001_bf100000_bat128_ex0.1_gm0.8_t0.5_ntv10000\DQN_1	997.8
DQN\lr0.01_bf100000_bat256_ex0.1_gm0.8_t0.5_ntv100\DQN_2	998.9
DQN\lr0.1_bf100000_bat256_ex0.1_gm0.8_t0.5_ntv100\DQN_2	998.9

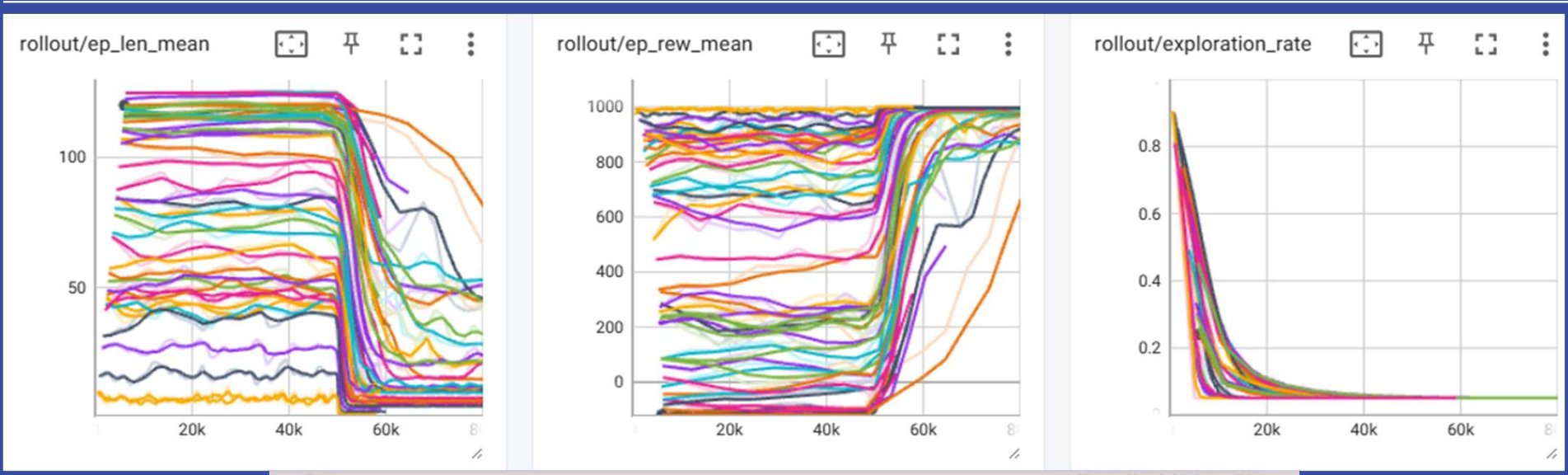
SMU

Best hyperparameters

- lr =1e-1
- Lr_schedule = cosine_annealing_schedule
- gamma = 0.8
- Replay buffer size = 1_000
- Batch size = 256
- Target network update interval = 100
- Exploration fraction = 0.05
- Exploration_initial_eps = 1.0,
- Exploration_final_eps = 0.05,
- tau = 1
- total_timesteps = 100_000
- Early_stop_mode = "StopTrainingOnNoModelImprovement"
- max_no_improvement_evals=5
- min_evals=50

SMU

Tracts training curves



Run	Smoothed	Value	Step
predDAC_DQN\id48085030101_lr0.1_bf100000_bat256_ex0.05_gm0.8_t0.5_ntv100\DQN_1	995.7	995.7	79,565
predDAC_DQN\id48085030901_lr0.01_bf100000_bat256_ex0.1_gm0.8_t0.5_ntv100\DQN_1	994.7	994.7	79,661
predDAC_DQN\id48085031008_lr0.01_bf100000_bat256_ex0.1_gm0.8_t0.5_ntv100\DQN_1	995.7	995.7	79,566
predDAC_DQN\id48113000410_lr0.01_bf100000_bat256_ex0.1_gm0.8_t0.5_ntv100\DQN_1	990.5	990.5	79,781
predDAC_DQN\id48113004700_lr0.01_bf100000_bat256_ex0.1_gm0.8_t0.5_ntv100\DQN_1	995.8	995.8	79,634
predDAC_DQN\id48113005300_lr0.01_bf100000_bat256_ex0.1_gm0.8_t0.5_ntv100\DQN_1	880.2	848.1	79,769
predDAC_DQN\id48113005500_lr0.1_bf1000_bat256_ex0.05_gm0.8_t1_ntv10\DQN_1	971.7	973	57,879
predDAC_DQN\id48113006301_lr0.1_bf100000_bat256_ex0.05_gm0.8_t0.5_ntv100\DQN_1	989.4	988.3	79,627
predDAC_DQN\id48113006702_lr0.1_bf100000_bat256_ex0.05_gm0.8_t0.5_ntv100\DQN_1	909.6	959.6	78,563
predDAC_DQN\id48113010001_lr0.01_bf100000_bat256_ex0.1_gm0.8_t0.5_ntv100\DQN_1	494.6	662	64,603
predDAC_DQN\id48113010001_lr0.1_bf100000_bat256_ex0.05_gm0.8_t0.5_ntv100\DQN_1	929.1	949	78,535
predDAC_DQN\id48113010101_lr0.1_bf100000_bat256_ex0.05_gm0.8_t0.5_ntv100\DQN_1	-120	-120	6,000

SMU

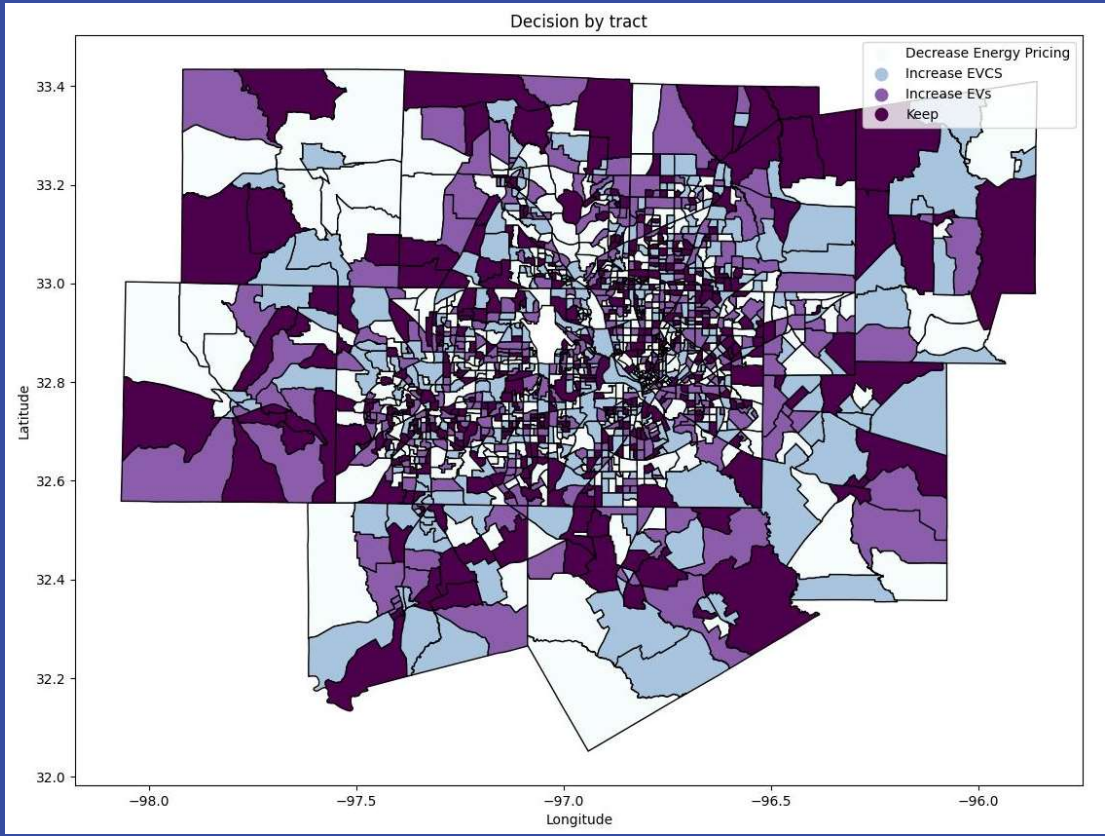
Actions taken by communities: Examples

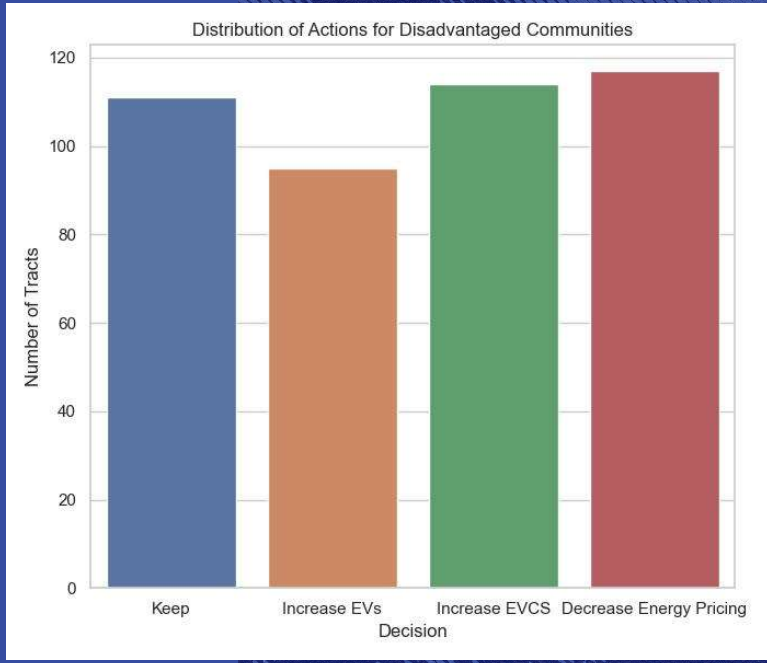
```
id:  48439105009
Actions: ['EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+
1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1']
Rewards: [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, 1000]
==================================
id:  48113013901
Actions: ['Education:+1%', 'Education:+1%', 'Education:+1%', 'Education:+1%', 'Education:+1%', 'Education:+1%', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1']
Rewards: [-1, -1, -1, -1, -1, -1, -1, -1, -1, 1000]
==================================
id:  48085031008
Actions: ['Education:+1%', 'Education:+1%', 'Education:+1%', 'Education:+1%', 'Education:+1%']
Rewards: [-1, -1, -1, -1, 1000]
==================================
id:  48113010001
Actions: ['Education:+1%', 'Education:+1%', 'Education:+1%', 'EV:+1', 'EV:+1', 'Education:+1%', 'Education:+1%', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'E
V:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1',
'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1']
Rewards: [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, 1000]
==================================
id:  48251130212
Actions: ['energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:
-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-
0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-
0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-0.01%', 'energy_burden:-
0.01%']
Rewards: [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, 1000]
==================================
id:  48367140412
Actions: ['Education:+1%']
Rewards: [1000]
==================================
id:  48257050602
Actions: ['Education:+1%', 'Education:+1%', 'Education:+1%', 'Education:+1%', 'Education:+1%', 'Education:+1%']
Rewards: [-1, -1, -1, -1, -1, 1000]
==================================
id:  48113017400
Actions: ['EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+
1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1', 'EV:+1']
Rewards: [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, 1000]
==================================
```


SMU

Pattern (work in progress)

The final pattern shows the collective behaviors of DFW communities. It provides higher level insights for policymakers





Remark and Future Work

» The proposed RL-ABM effectively integrates the strengths of both modeling techniques, maximizing their advantages for more effective analysis and decision-making

» It offers policymakers valuable insights and aids in understanding trends and patterns, facilitating informed decision-making processes.

» The framework serves as a foundational model applicable across disciplines, extending beyond the confines of this project's specific topic. Its versatility makes it a valuable tool for addressing a wide range of challenges and scenarios across various domains."

» Future work will consider incorporating alternative modes of interactions among agents, such as cooperative, competitive, or mixed modes, to better capture real-world dynamics and behaviors beyond the traditional self-interested paradigm.