REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
01-11-2022	Final Report	1-Oct-2016 - 31-Aug-2020

4. TITLE AND SUBTITLE
Final Report: HELIOS: Accelerated Recovery of Evolving Spatial-Temporal Dynamics

5a. CONTRACT NUMBER
W911NF-16-1-0565

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Texas Engineering Experiment Station
SRS
400 Harvey Mitchell Parkway South, Suite 300
College Station, TX 77845 -4375

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
69865-HC-DRP.5

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			James Caverlee
UU	UU	UU	UU		19b. TELEPHONE NUMBER 979-458-3870

RPPR Final Report
as of 01-Nov-2022

Agency Code: 21XD

Proposal Number: 69865HCDRP **Agreement Number: W911NF-16-1-0565**
INVESTIGATOR(S):

Name: James Caverlee
Email: caverlee@cse.tamu.edu
Phone Number: 9794583870
Principal: Y

Organization: **Texas Engineering Experiment Station**
Address: SRS, College Station, TX 778454375
Country: USA
DUNS Number: 847205572 EIN: 741974733

Report Date: 30-Sep-2020 Date Received: 01-Nov-2022

Final Report for Period Beginning 01-Oct-2016 and Ending 31-Aug-2020
Title: HELIOS: Accelerated Recovery of Evolving Spatial-Temporal Dynamics
Begin Performance Period: 01-Oct-2016 **End Performance Period:** 31-Aug-2020
Report Term: 0-Other

Submitted By: James Caverlee Email: caverlee@cse.tamu.edu
Phone: (979) 458-3870

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 8 **STEM Participants:** 15

Major Goals: The overall goal of the HELIOS project is to create new methods, algorithms, and frameworks for ``filling in the gaps'' of large rapidly evolving datasets that are characterized by noisy and missing information, and to demonstrate a significant improvement on methods of group innovation tasks built over these datasets. The detailed tasks and milestones are listed in the attached PDF report.

Accomplishments: See the attached PDF report

Training Opportunities: We have mentored many PhD and MS students as part of this project.

RPPR Final Report
as of 01-Nov-2022

Results Dissemination: We have published many papers in top-tier venues, and given talks on the topic of this project.

Towards Automated Neural Interaction Discovering for Click-Through Rate Prediction
Qingquan Song, Dehua Cheng, Eric Zhou, Jiyan Yang, Yuandong Tian, and Xia Hu
ACM SIGKDD Conference on Knowledge Discovery and Data Mining
2020

Content-Collaborative Disentanglement Representation Learning for Enhanced Recommendation
Yin Zhang, Ziwei Zhu, Yun He, James Caverlee
RecSys 2020

Multi-Channel Graph Neural Networks
Kaixiong Zhou, Qingquan Song, Daochen Zha, Na Zou, and Xia Hu
International Joint Conference on Artificial Intelligence
2020

Unbiased Implicit Recommendation and Propensity Estimation via Combinational Joint Learning (short paper).
Ziwei Zhu, Yun He, Yin Zhang, and James Caverlee
The 14th ACM Conference on Recommender Systems
2020

On Robustness of Neural Architecture Search under Label Noise
Yi-Wei Chen, Qingquan Song, Xi Liu, P S Sastry, and Xia Hu
Frontiers in Big Data, section Data Mining and Management
2020

Coupled Variational Recurrent Collaborative Filtering
Qingquan Song, Shiyu Chang, and Xia Hu
ACM SIGKDD Conference on Knowledge Discovery and Data Mining
2019

(*) Graph Recurrent Networks with Attributed Random Walks
Xiao Huang, Qingquan Song, Yuening Li, and Xia Hu
ACM SIGKDD Conference on Knowledge Discovery and Data Mining
2019

Tensor Completion Algorithms in Big Data Analytics
Qingquan Song, Hancheng Ge, James Caverlee, and Xia Hu ACM Transactions on Knowledge Discovery from Data (TKDD)

DisTenC: A Distributed Algorithm for Scalable Tensor Completion on Spark Hancheng Ge, Kai Zhang, Majid Alfifi, Xia Hu, and James Caverlee
2018 IEEE 34th International Conference on Data Engineering

Fairness-Aware Tensor-Based Recommendation
Ziwei Zhu, Xia Hu, and James Caverlee
27th ACM International Conference on Information and Knowledge Management (CIKM 2018)

Accelerated Local Anomaly Detection via Resolving Attributed Networks
Ninghao Liu, Xiao Huang, and Xia Hu
Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)

Multi-Aspect Streaming Tensor Completion
Qingquan Song, Xiao Huang, Hancheng Ge, James Caverlee, and Xia Hu
23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2017)

RPPR Final Report
as of 01-Nov-2022

An Interpretable Classification Framework for Information Extraction from Online Healthcare Forums
Jun Gao, Ninghao Liu, Mark Lawley, and Xia Hu
Journal of Healthcare Engineering, Volume 2017

Honors and Awards: James Caverlee
==========
Promoted to Full Professor, 2019
General Co-Chair for WSDM 2020
WSDM 2019 Outstanding Senior Program Committee, 2019
Associate Editor, IEEE Transactions on Knowledge and Data Engineering Associate Editor, IEEE Intelligent Systems
Associate Editor, Social Network Analysis and Mining

University-Specific Awards:
==========
Texas A&M Computer Science Undergraduate Faculty Teaching Award, 2019 Texas A&M TEES Faculty Fellow Award, 2018-2019

Xia "Ben" Hu
==========
Best Student Paper Award, IISE QCRE 2019 Best Paper Award Shortlist, WWW 2019 Adobe Data Science Research Award, 2019 JP Morgan AI Research Faculty Award, 2019 NSF CAREER Award, 2018

University-Specific Awards:
==========
TEES Young Faculty Fellow, Texas A&M Engineering Experiment Station, 2018 Dean of Engineering Excellence Award, Texas A&M University, 2019

Protocol Activity Status:

Technology Transfer: PyTen: An efficient software of distributed large-scale tensor completion recovers the missing values for real-world datasets.
https://github.com/tamu-helios/pyten

PARTICIPANTS:

Participant Type: PD/PI
Participant: James Caverlee
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co PD/PI
Participant: Xia "Ben" Hu
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

RPPR Final Report
as of 01-Nov-2022

Partners

,

I certify that the information in the report is complete and accurate:
Signature: James Caverlee
Signature Date: 11/1/22 11:55AM

NGS2 Program Update:

HELIOS Representation Learning in Big Social Data

PIs: James Caverlee and Xia “Ben” Hu

Department of Computer Science & Engineering
Texas A&M University

Recap of Methodological Approach

1 **Representation Learning Overview**

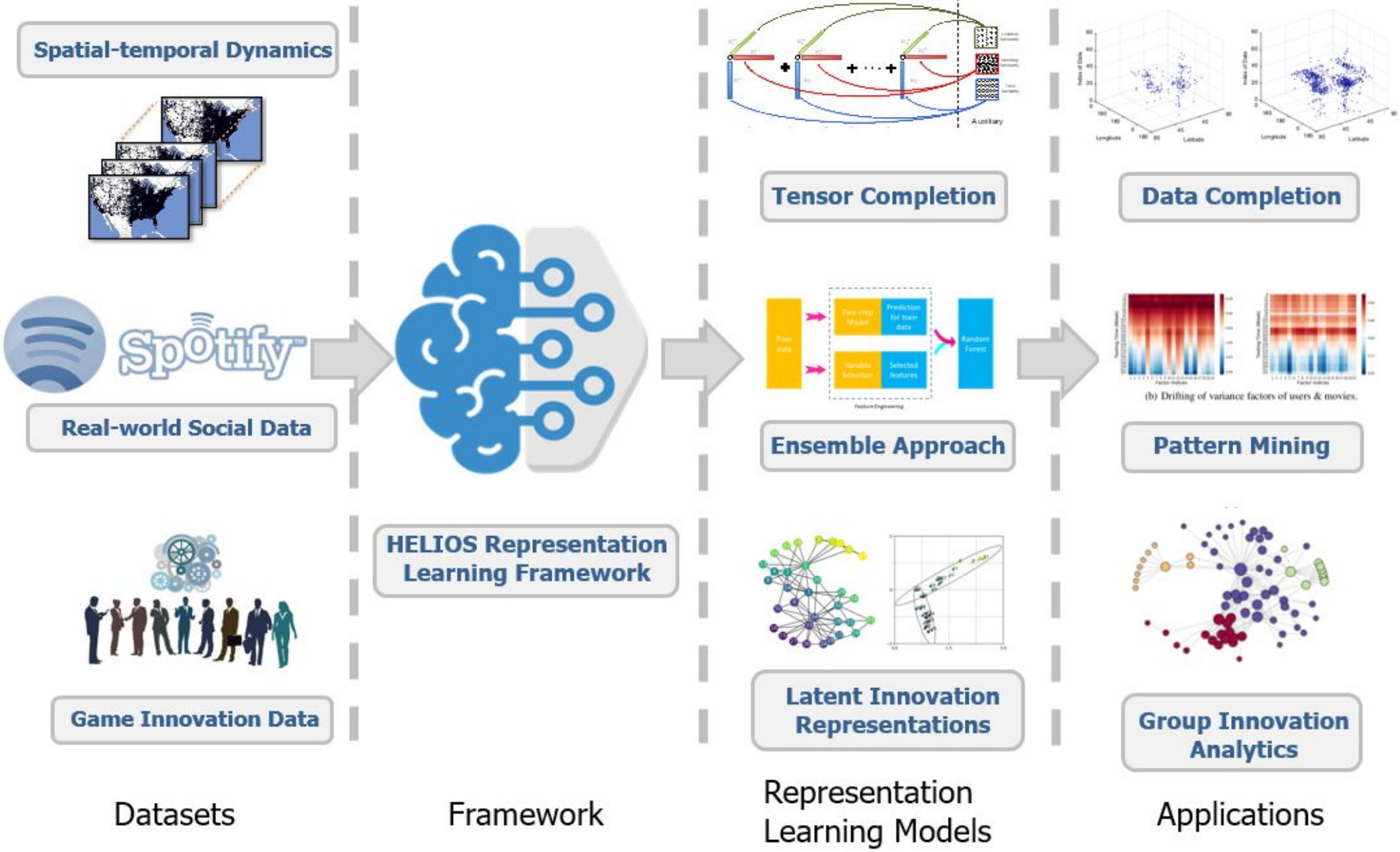
2 **Tensor Completion**

3 **Ensemble Approach**

4 **Group Innovation Representation**

Representation Learning – Overview

❖ ***Objective:*** Helios representation learning framework is designed to analyze real-world big social data via three main thrusts: tensor completion, ensemble approach, and innovation representation learning.

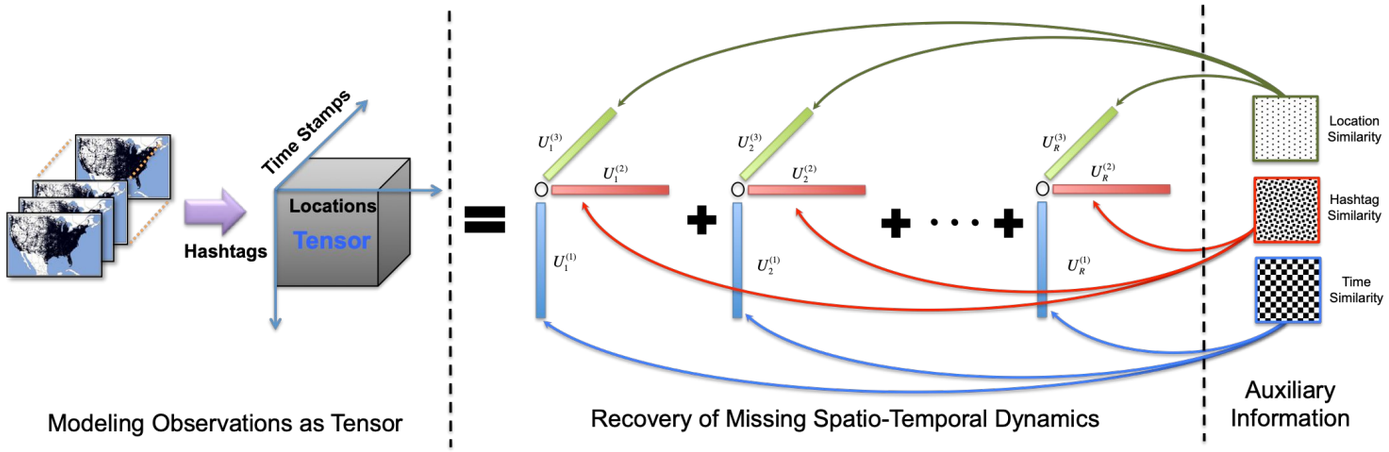


Tensor Completion (1/5)

- ***Goal:*** Tensor completion algorithms are designed to model the entangled relationships across multiple dimensions in real-world data, which is embedded into a latent representation space.
- ***Developed frameworks:***

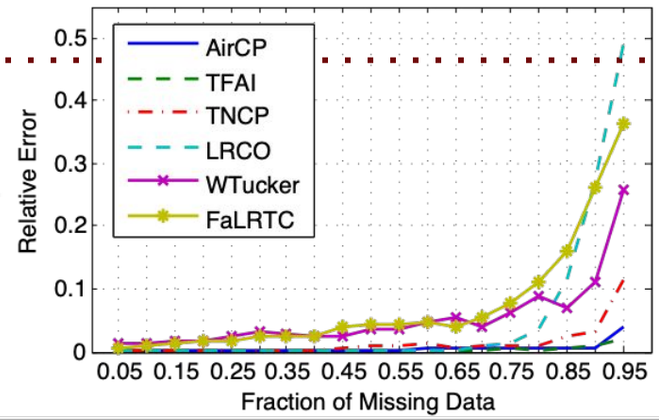
AirCP: Model the static latent relationships among locations, times, and observations of human activities, targeting at uncovering the missing data.

Contributions: (1) Accelerate this recovery through a novel parallelization for fast computation on multi-core cloud environments. (2) AirCP achieves the lowest relative error, and is an order of magnitude faster than existing approaches.



AirCP

Error



Time cost

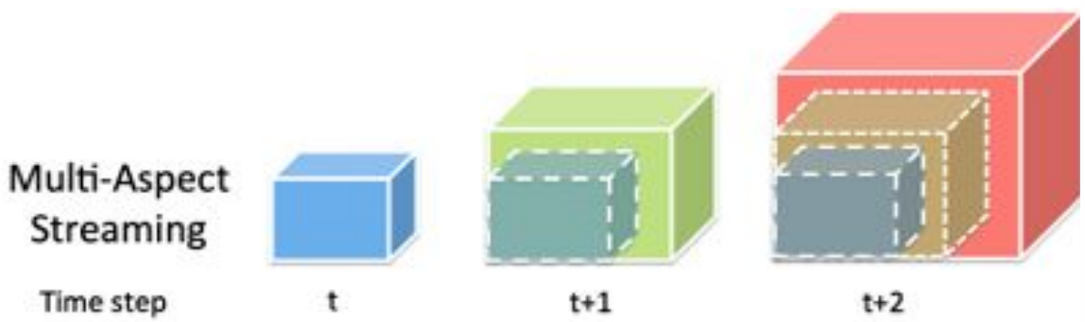
FM	AirCP	TFAI	TNCP	LRCO	WTucker	FaLRTC
20%	**3.97**	46.36	4.51	58.64	195.22	5.47
40%	4.27	49.87	**3.94**	49.61	186.13	5.62
60%	**3.89**	35.98	4.08	51.30	202.58	4.78
80%	4.76	43.75	**4.16**	60.04	183.98	4.46

Tensor Completion (2/5)

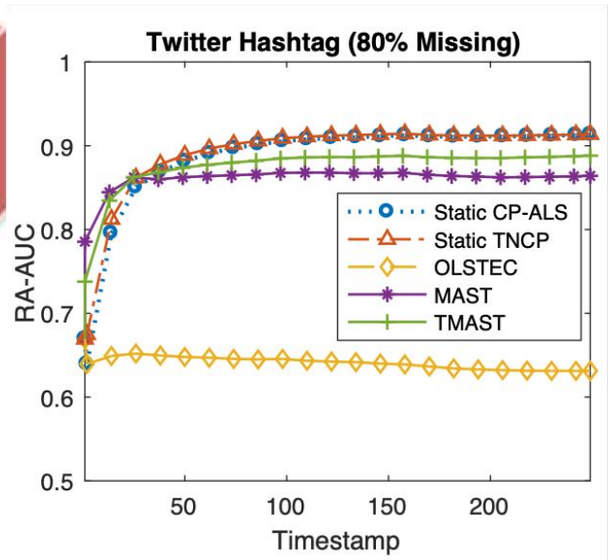
❖ ***Developed frameworks:***

MAST: Models streaming data dynamics with time. It completes the incremental tensors by tackling the uncertainty of tensor mode and complex data structure of multi-aspect streaming tensors.

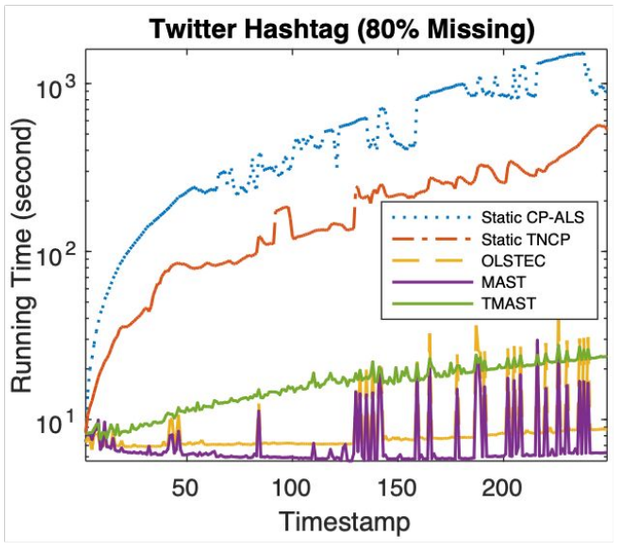
Contributions: (1) MAST has commensurate performance comparing with two static models and higher performance than the dynamic baseline method OLSTEC. (2) MAST takes much less running time than static models and outperforms dynamic baselines.



MAST



AUC comparison



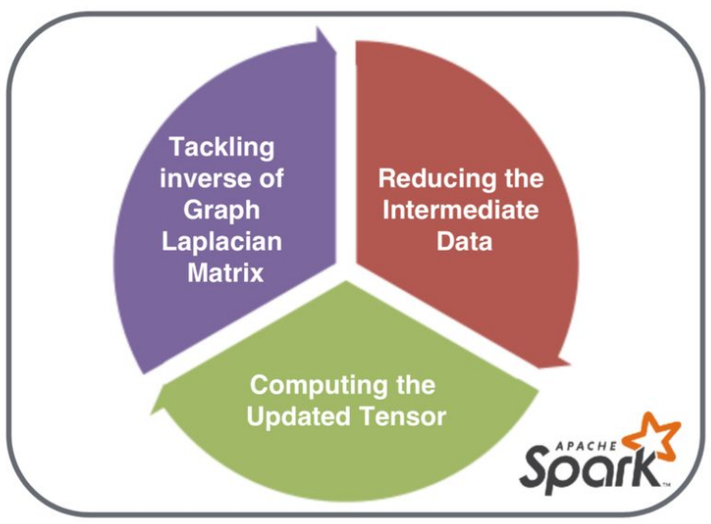
Time cost

Tensor Completion (3/5)

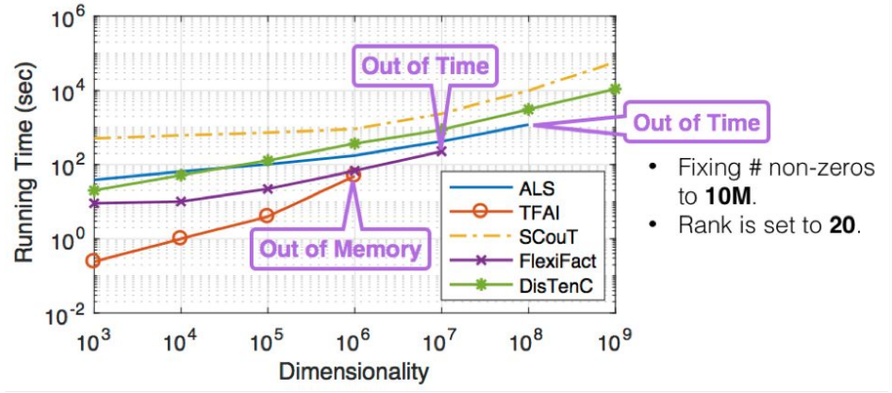
- ***Goals:*** Improve tensor completion algorithms by further considering two practical constraints: computational efficiency and algorithmic fairness over different groups of data.
- ***Developed frameworks:***

DisTenC: Speeds up the computation of tensor completion with auxiliary information based on ADMM. We implement it on the modern distributed computing architecture Spark.

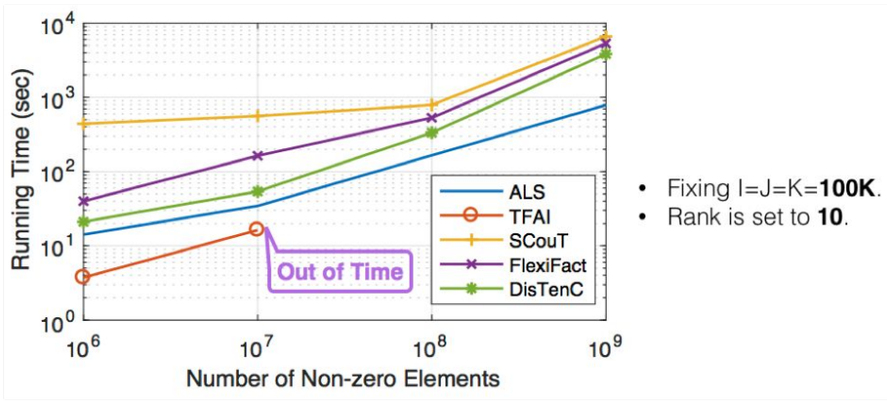
Contributions: Outperforms existing approaches with 10~1000X better scalability.



DisTenC



- Fixing # non-zeros to **10M**.
- Rank is set to **20**.



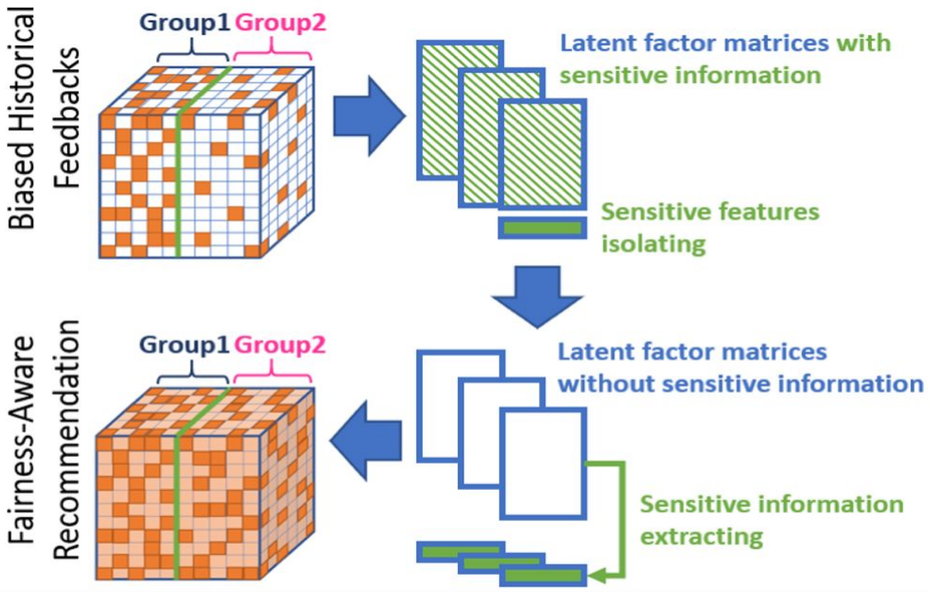
- Fixing I=J=K=**100K**.
- Rank is set to **10**.

Tensor Completion (4/5)

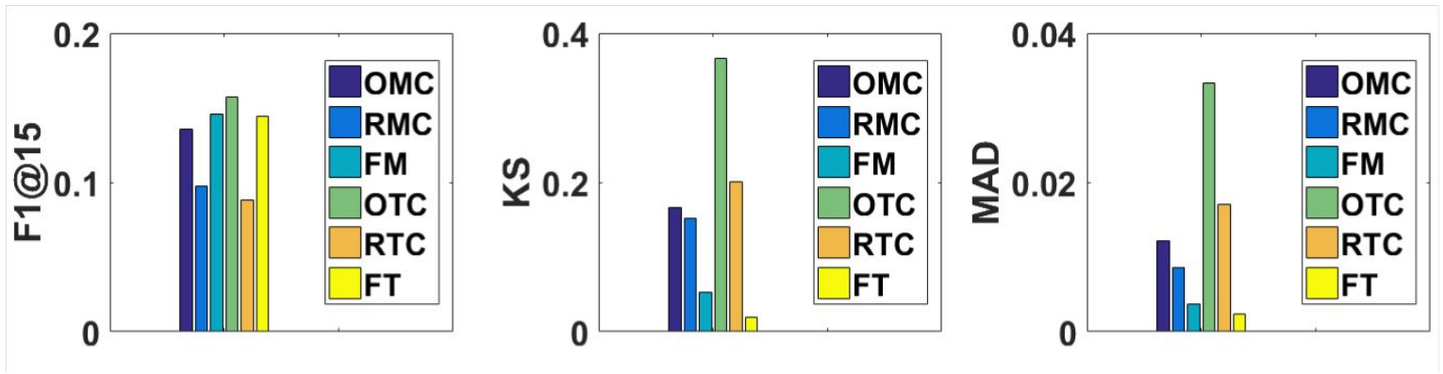
❖ ***Developed frameworks:***

Fairness-aware tensors (FT): Overcome the algorithmic bias inherent in conventional tensor-based models.

Contributions: (1) Our FT provides the best fairness enhancement result; (2) FT has a comparable recommendation performance compared with the best recommendation model among the baselines. (3) FT can effectively augment fairness while preserving recommendation utility (FT vs. FM).



Fairness-aware algorithm



Recommendation performance

Bias indicator 1

Bias indicator 2

Tensor Completion (5/5)

- Encapsulate the distributed framework DisTenC into PyTen package.

Github: https://github.com/tamu-helios/pyten

Package [Author]	Basic Operation	Basic TD	TD with Auxiliary	Dynamic TD	Tensor Completion	GUI
TensorToolbox [Bigoni, MIT]	✔	✔	✗	✗	✗	✗
ttpy [Oseledets, Skoltech]	✔	✔	✗	✔	✗	✗
scikit-tensor [Nickel, MIT]	✔	✔	✗	✗	✗	✗
PyTensor [Ji Oh Yoo, CMU]	✔	✔	✗	✔	✗	✗
PyTen [HELIOS, TAMU]	✔	✔	✔	✔	✔	✔

Tensor Completion Algorithms in Big Data Analytics

QINGQUAN SONG, HANCHENG GE, JAMES CAVERLEE, and XIA HU,
Texas A&M University

Tensor completion is a problem of filling the missing or unobserved entries of partially observed tensors. Due to the multidimensional character of tensors in describing complex datasets, tensor completion algorithms and their applications have received wide attention and achievement in areas like data mining, computer vision, signal processing, and neuroscience. In this survey, we provide a modern overview of recent advances in tensor completion algorithms from the perspective of big data analytics characterized by diverse variety, large volume, and high velocity. We characterize these advances from the following four perspectives: general tensor completion algorithms, tensor completion with auxiliary information (variety), scalable tensor completion algorithms (volume), and dynamic tensor completion algorithms (velocity). Further, we identify several tensor completion applications on real-world data-driven problems and present some common experimental frameworks popularized in the literature along with several available software repositories. Our goal is to summarize these popular methods and introduce them to researchers and practitioners for promoting future research and applications. We conclude with a discussion of key challenges and promising research directions in this community for future exploration.

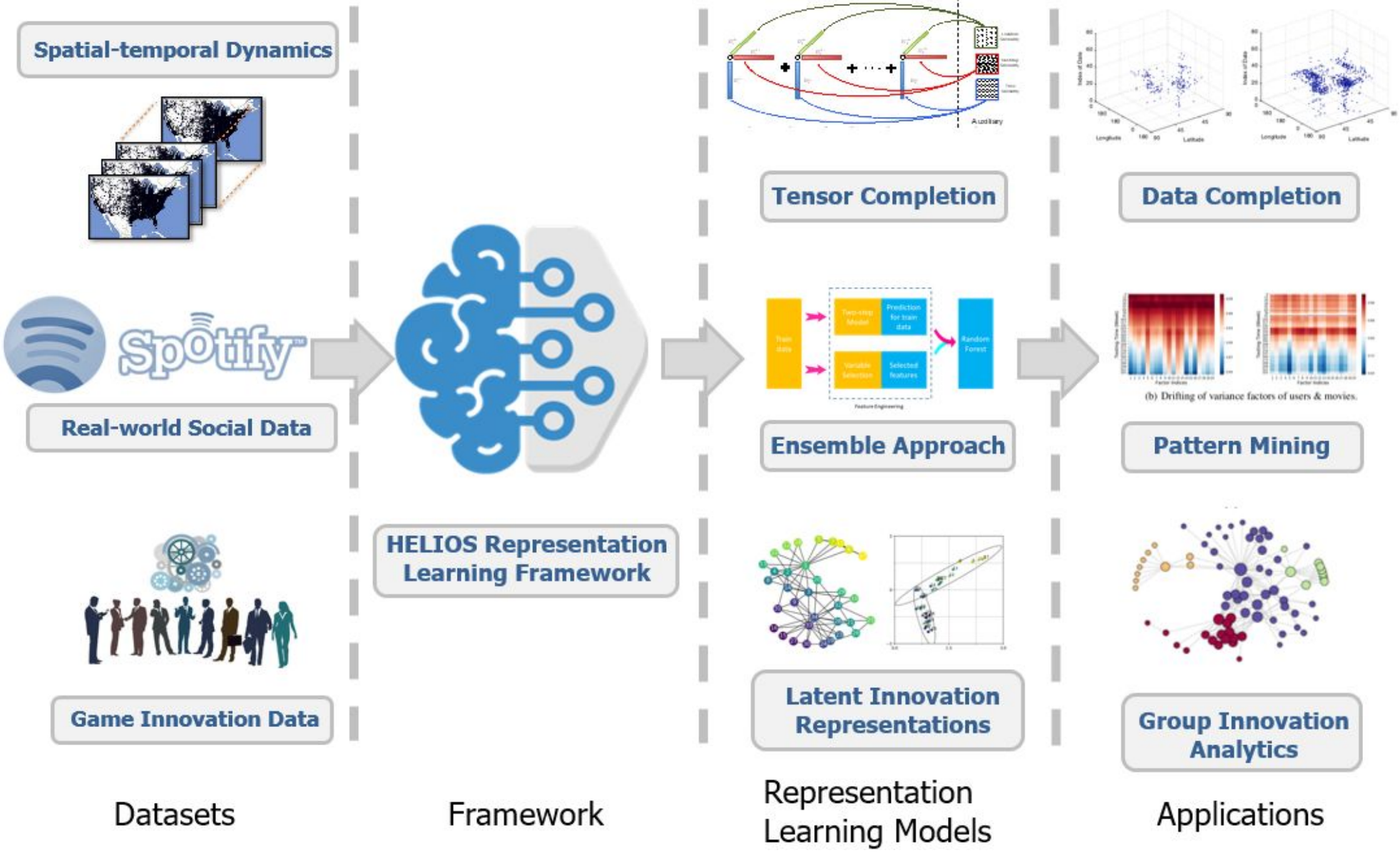
CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Mathematics of computing** → *Dimensionality reduction*; • **Computing methodologies** → *Dimensionality reduction and manifold learning*; *Factorization methods*; *Regularization*;

Additional Key Words and Phrases: Tensor, tensor completion, tensor decomposition, tensor factorization, multilinear data analysis, dynamic data analysis, big data analytics

ACM Reference format:
Qingquan Song, Hancheng Ge, James Caverlee, and Xia Hu. 2019. Tensor Completion Algorithms in Big Data Analytics. *ACM Trans. Knowl. Discov. Data* 13, 1, Article 6 (January 2019), 48 pages.
https://doi.org/10.1145/3278607

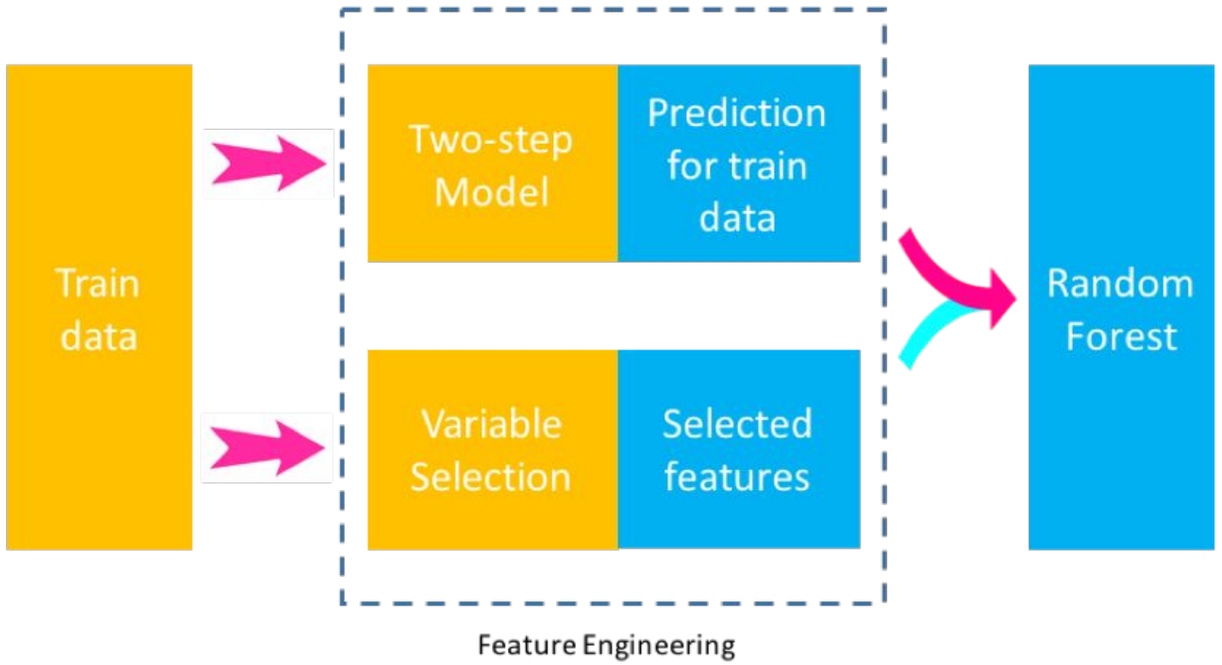
Representation Learning – Overview

❖ ***Objective:*** Helios representation learning framework is designed to analyze real-world big social data via three main thrusts: tensor completion, ensemble approach, and innovation representation learning.



Ensemble Algorithm – Fragile Family Challenge Data Analytics (1/2)

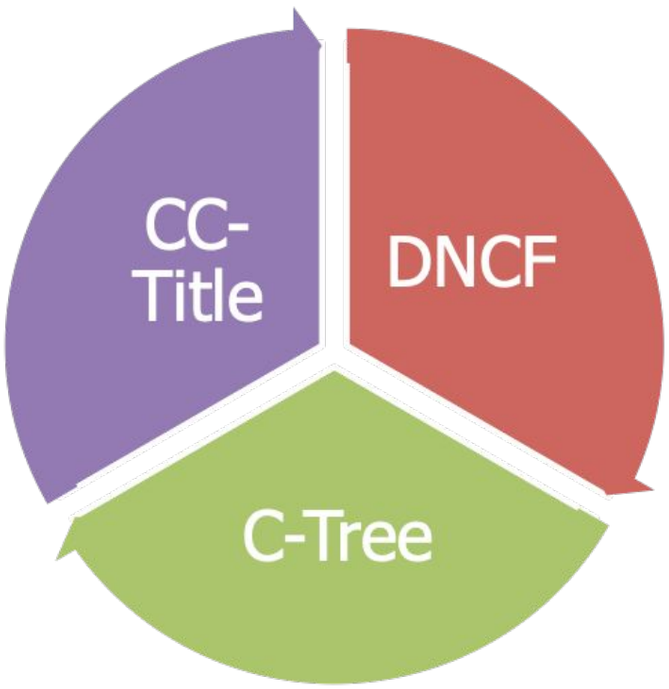
- ***Goal:*** To perform six prediction tasks posed in the Fragile Family Data Challenge related to six variables: GPA, Grit, Material, Eviction, Layoff, and Job.
- ***Dataset:*** Longitudinal social data from 4,789 American families over more than 15 years. There are more than 12,000 raw features, including relationships, health, eviction, behaviors, education, etc.
- ***Framework:*** A joint model including feature selection and regression approach of random forest, providing better features to further improve the performance.



Fragile Families & Child Wellbeing Study
PRINCETON | COLUMBIA

Ensemble Algorithm – 2018 ACM RecSys Challenge Data Analytics (2/2)

- ***Goal:*** To develop a system for the task of automatic playlist continuation.
- ***Datasets:*** Spotify 1 million playlist dataset, created by Spotify users, and includes playlist titles, track listings and other metadata.
- ***Model — TrailMix*:** Combines three different models (DNCF, CC-Title, C-True) designed to exploit complementary aspects of playlist recommendation.



- CC-Title: Context Clustering using Playlist Title
- DNCF: Decorated Neural Collaborative Filtering
- C-Tree: Constructed Tree Leveraging the Natural Tree-Structure of Each Playlist

Table 1: Results for TrailMix on Leaderboards

	R-precision	**NDCG**	**Clicks**
DNCF + CC-Title	0.1724	0.3292	2.8152
C-Tree + CC-Title	0.1981	0.3567	2.4756
TrailMix	0.2057	0.3711	2.2710

Streaming Matrix Level Representation Learning – Recommender System (1/1)

- ***Goal:*** To conduct representation learning on dynamic varying matrices, which serve as the ground model for dynamically uncovering the group innovation patterns in the next step.
- ***Task:*** Streaming Recommendation Tasks.
- ***Model — DEVAS:*** Deep Variational Streaming Recommender System.
- ***Conclusion:*** DEVAS outperforms all baselines on three benchmark datasets.

(a) Overview of six methods.

Methods \ Categories	Streaming	Temporal Involved	Probabilistic	Deep
PMF [3]			✓	
time-SVD++ [4]		✓		
sD-PMF	✓		✓	✓
sRRN [5]	✓	✓		✓
sRec [6]	✓	✓	✓	
DEVAS	✓	✓	✓	✓

(b) Comparison of RMSE results.

	Methods	Datasets MT	ML-10M	Netflix
Batch	PMF	1.5723	0.8202	0.9421
	time-SVD++	1.4630	0.7985	0.9311
Streaming	sD-PMF	1.6170	0.9017	0.9992
	sRRN	1.5646	0.8003	0.9236
	sRec	1.4831	0.8121	0.9288
	DEVAS	**1.4567**	**0.7831**	**0.9050**

Representation Learning – Innovation Representation (1/1)

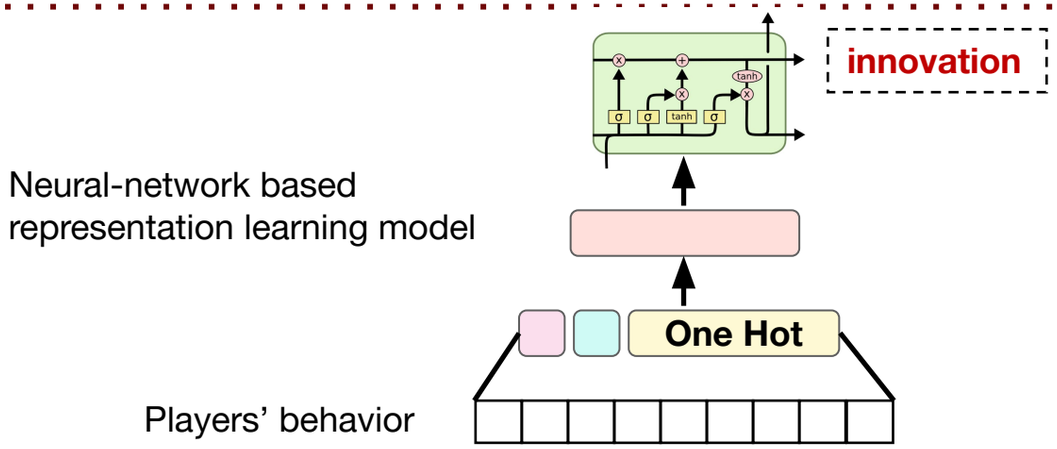
- ***Modeling Methods:*** Neural-network-based representation learning models are developed to uncover the group innovation pattern by observing players' features.
- ***Data:***

In game data, innovation is implied by the new behaviors of a group of players. There are two available game logs: ***I.*** Boomtown from the Gallup team, and ***II.*** Resource Collection from the Berkeley team.

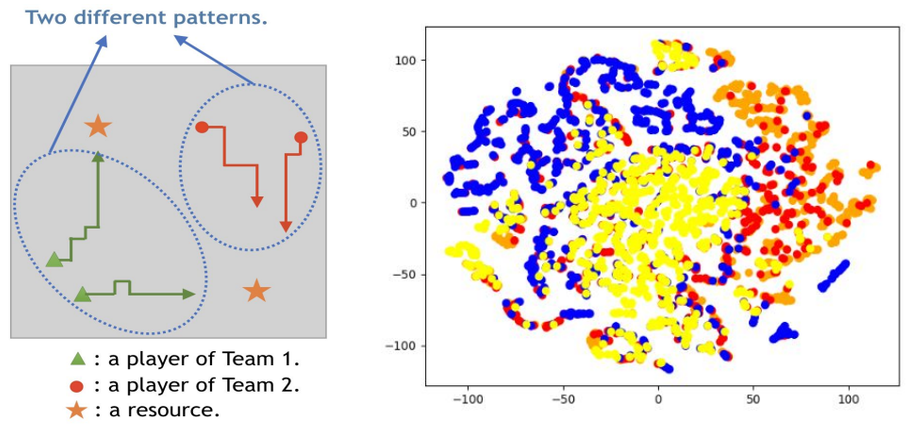
- ***Results:***

I. Our approach effectively captures the relation between players' features and group innovation. This can help following studies of group innovation in social competition.

II. We define the group innovation pattern based on players' moving trajectories. Our representation models embed group innovation into a hidden space and visualize it to facilitate understanding.



I. Representation learning models between players' behavior and innovation



II. Mapping between moving trajectory and group innovation

Approach in Cycle 3

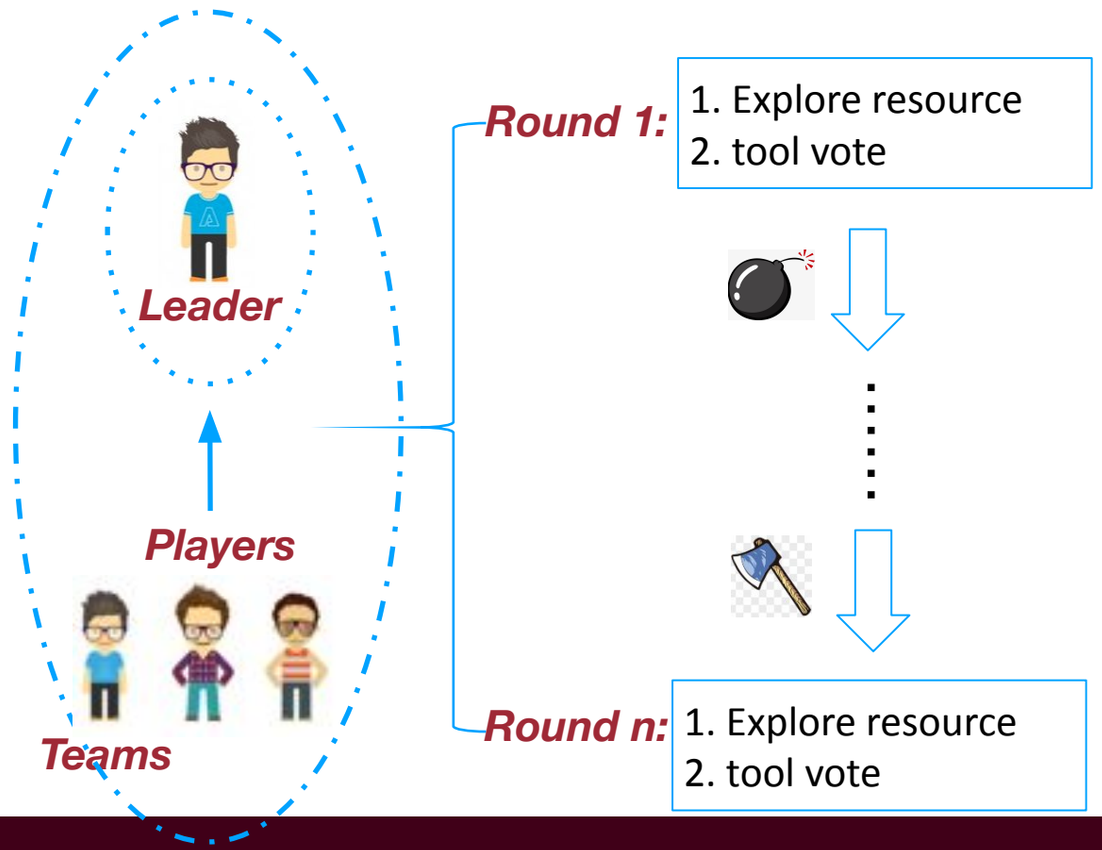
1. Gallup Data Analysis
2. Berkeley Data Analysis
3. State-of-the-art Methods

Gallup Data Analytics – Problem Statement (1/5)

Game description: *A* team consisting of several members plays the resource exploration game, while choosing tools to represent the group innovation patterns.

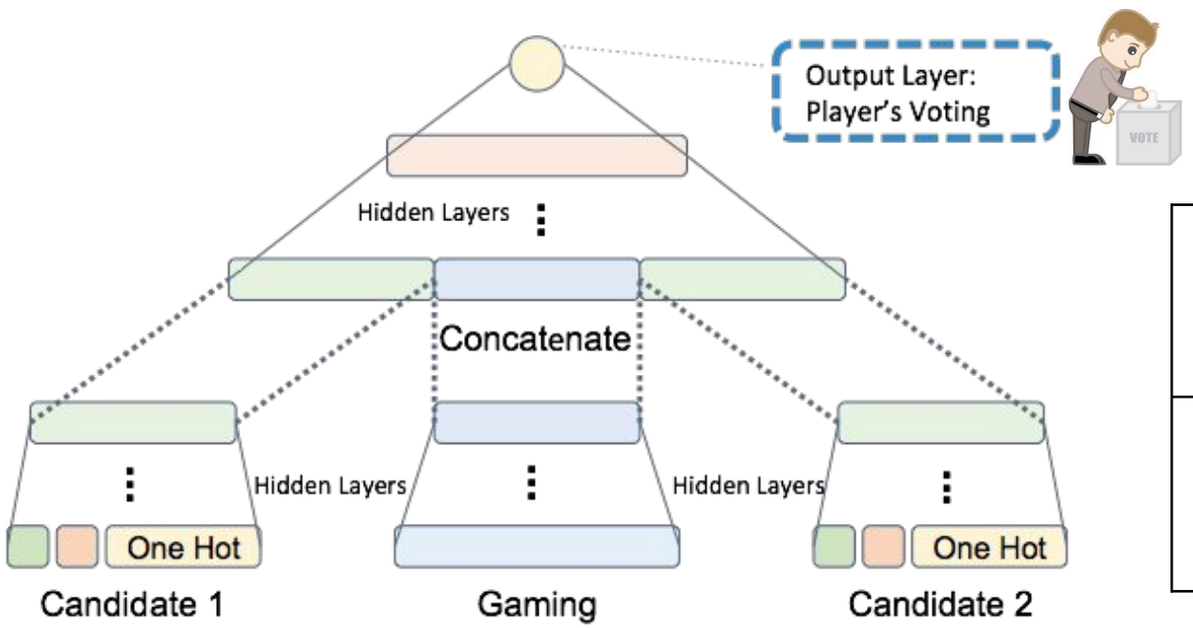
- **Time length:** Composed of n rounds of resource exploration.
- **Vote:** At the end of each round, both the leader and team members are requested to vote on expected tools used for the next round.

Innovation Analysis: We define the innovation pattern of player and group as the tool category, and formulate the Innovation analysis task as multi-class prediction problem.



Gallup Data Analytics – Player Innovation Representation (2/5)

- ***Task:*** Predict the tool selection of a player from the noisy and abundant input features, including 'player movements' and 'tool power', etc.
- ***Method:*** We propose a neural network based representation learning model accompanied with feature selection. We concatenate the input features and learn their hidden representation end-to-end.
- ***Insights:*** (1) It captures the innovation pattern in latent embedding space to improve accuracy. (2) Innovation pattern is related to informative features in the game process.

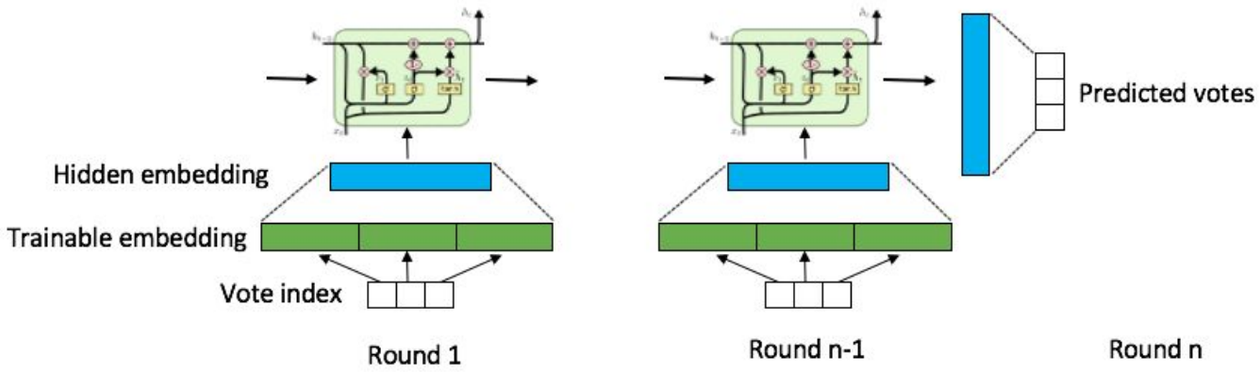


Accuracy Comparison

KNN	LR	NB	DT	RF	MLP-3	**RL**
0.71	0.72	0.67	0.67	0.73	0.73	**0.75**

Gallup Data Analytics – Sequential Player Innovation Analysis (3/5)

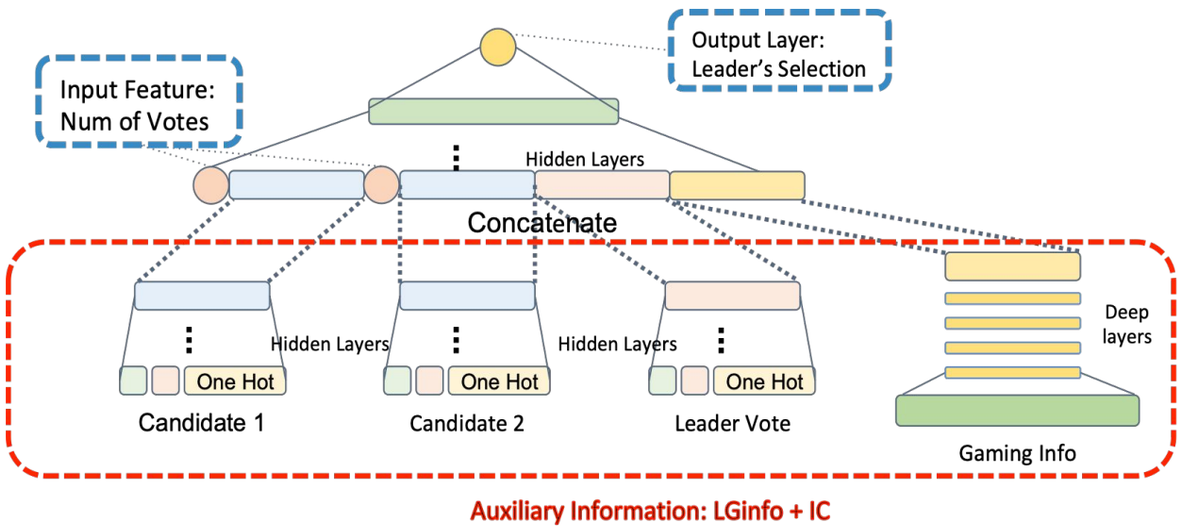
- ***Task:*** Predict the tool vote (innovation pattern) at the next round given the vote history, to understand innovation pattern preference.
- ***Method:*** We build a representation learning approach based RNN to predict the tool votes at the next round, which is essentially a multi-class classification problem.
- ***Insights:*** (1) The prediction accuracies are larger than 0.77, which means that the future votes depend highly on the vote history. (2) The prediction accuracy increases with the round memory length, meaning that the individual player has specific tool vote (innovation pattern) preference memorized in the vote history.



	Round memory length			
	1	2	3	4
Dataset size	8508	6267	4108	2013
Accuracy	77.45±1.1	80.57±2.2	85.28±6.8	**91.32±5.3**

Gallup Data Analytics – Group Innovation Representation (4/5)

- ***Task:*** Predict the tool selection of the leader, based on all the members' information.
- ***Method:*** We augment the previous model with auxiliary features, including Leader's Vote (LV), Number of each votes (NumV), Item Characters (IC), Leader's Gaming Information (LGinfo).
- ***Insights:*** (1) The selected auxiliary information improves group innovation learning. (2) The leader tends to follow other players' voting to make the decision.

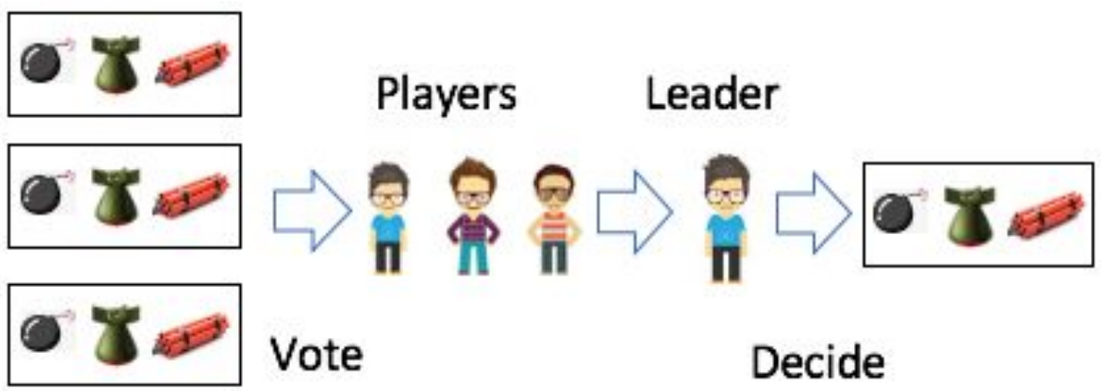


Accuracy Comparison

	Used Feature	Accuracy (mean)	Accuracy (std)
Baseline 1	LV	0.47196	0
Baseline 2	NumV	0.76988	0
Representation Learning 1	NumV + IC	0.80010	0.00768
Representation Learning 2	NumV + LGinfo	0.85649	0.01318
Representation Learning 3	**NumV + LGinfo + IC**	**0.89474**	**0.01091**

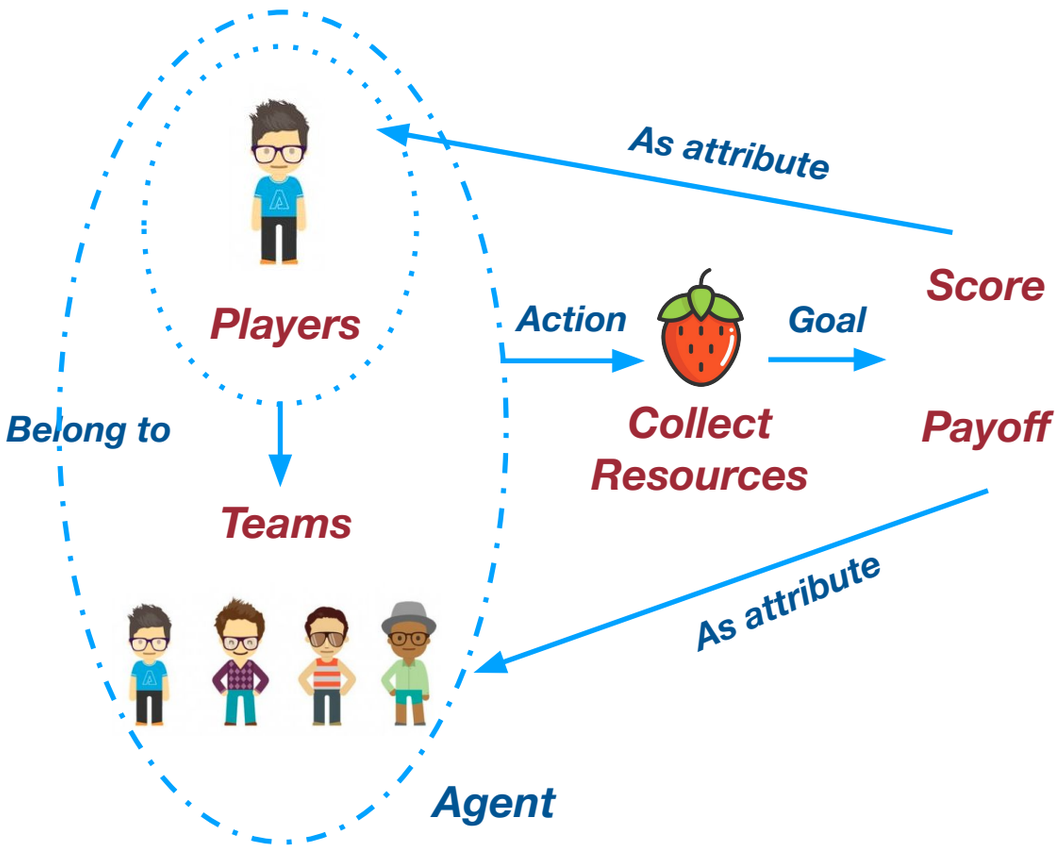
Gallup Data Analytics – Sequential Group Innovation Analysis (5/5)

- ***Task:*** Predict the tool decision (innovation pattern) of a group at the next round given the decision history.
- ***Method:*** We apply the previous RNN based model with auxiliary leader features.
- ***Insight:*** (1) The larger prediction accuracy means that the future decisions depend highly on the decision history. (2) The prediction accuracy increases with the round memory length, which means that the group innovation pattern is memorized in the decision history.



	Round memory length			
	1	2	3	4
Dataset size	2928	2176	1440	712
Accuracy	78.17±1.7	80.16±2.0	85.17±5.7	**91.62±4.9**

Berkeley Data Analytics – Game Setting (1/5)



Terminologies:

Team(s) — Fixed integer parameter (1 to 4)

- Predefined number of **teams** in one **game.**

Player(s) — Fixed integer parameter (1 to 4)

- Predefined number of **players** in each **team**

Resource(s) — Dynamically Update.

- The location of each resource is generated randomly.
- The members of a team cooperate to collect as more resources as possible.

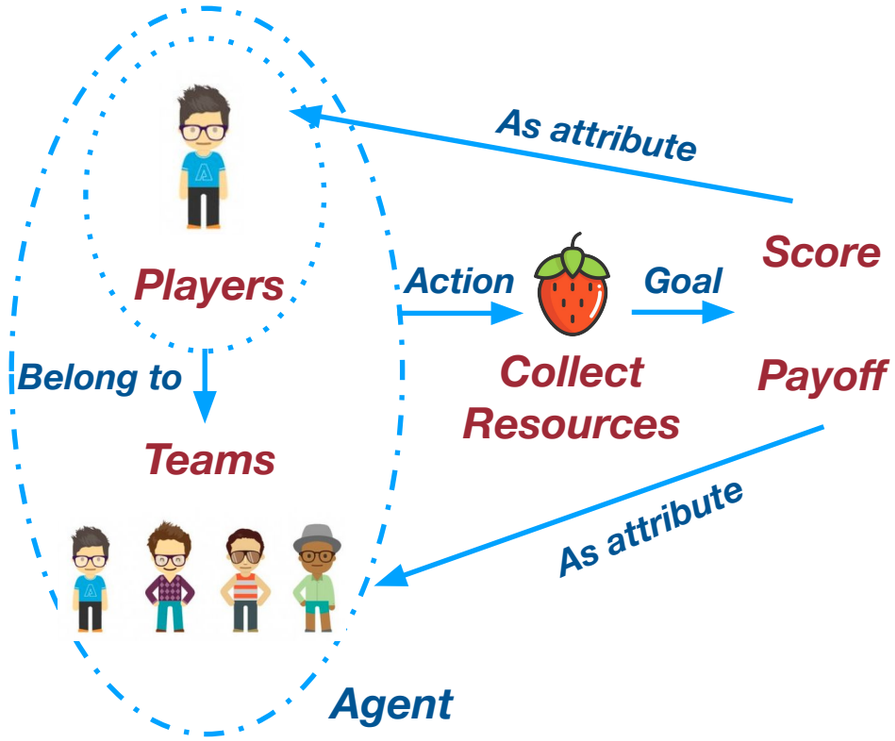
Score and Pay-off — Dynamically Update

- Score will be accumulated once one user catches a resource.
- Pay-off will be accumulated once one team catches a resource.

Berkeley Data Analytics – Problem Statement (2/5)

Data cleaning:

- Game logs capture lots of information.
- Important information after cleaning: actions with timestamps *(move/spawn resource),* movement details *(direction/speed),* current locations *(players/resources),* temporary score and pay-off.



Innovation definition and problems:

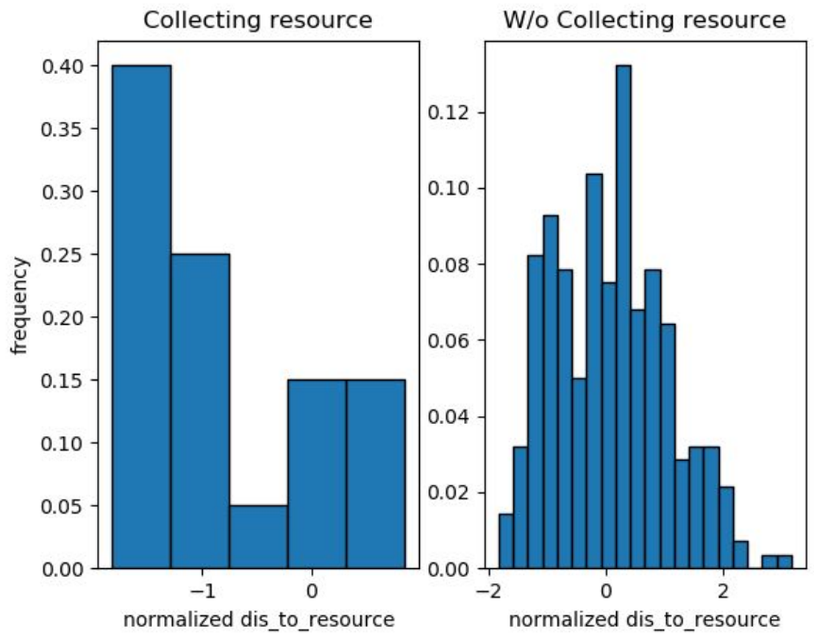
- Can we represent the innovation pattern of a player latently, and predict whether it facilitates collecting a resource?
- Can we represent the innovation pattern of a team latently, and have a *prediction* of the score/payoff?
- Can we represent the innovation pattern intuitively from the perspective of movement strategy?

Berkeley Data Analytics – Resource Collection Prediction (3/5)

- ***Task:*** Predict a player will collect a specific resource or not. We represent the innovation pattern with the combination of features, by training it to inform the resource collection.
- ***Method:*** A feature formulation and selection approach is proposed to improve prediction accuracy. We have (1) distance to resource, (2) average distance to other players, (3) direction to resource, (4) temporary score and (5) pay-off.
- ***Insights:*** (I) A specific combination showcases how well it can improve accuracy. (II) The smaller distances and the larger score indicate a good innovation pattern.

	Logistic regression	SVM	Decision tree	Adaboost	Average
(1) Dis_to_resource	***66.33***	***79.33***	67.00	67.00	69.92
(2) Dis_to_player	52.33	57.00	81.33	81.33	68.00
(3) Dir_to_resource	50.00	43.33	43.67	43.67	45.17
(4) Score	38.33	69.33	69.33	69.33	61.58
(5) Payoff	48.67	43.00	10.33	10.33	28.08
(1) + (4)	***66.33***	***79.33***	67.00	67.00	69.92
(1) + (2)	***66.33***	71.66	88.67	88.67	78.83
(2) + (4)	47.33	52.99	86.00	88.67	68.75
(1) + (2) + (4)	65.00	72.67	***89.33***	***90.00***	***79.25***
(1) + … + (5)	67.33	69.67	88.33	88.33	78.41

Prediction accuracy

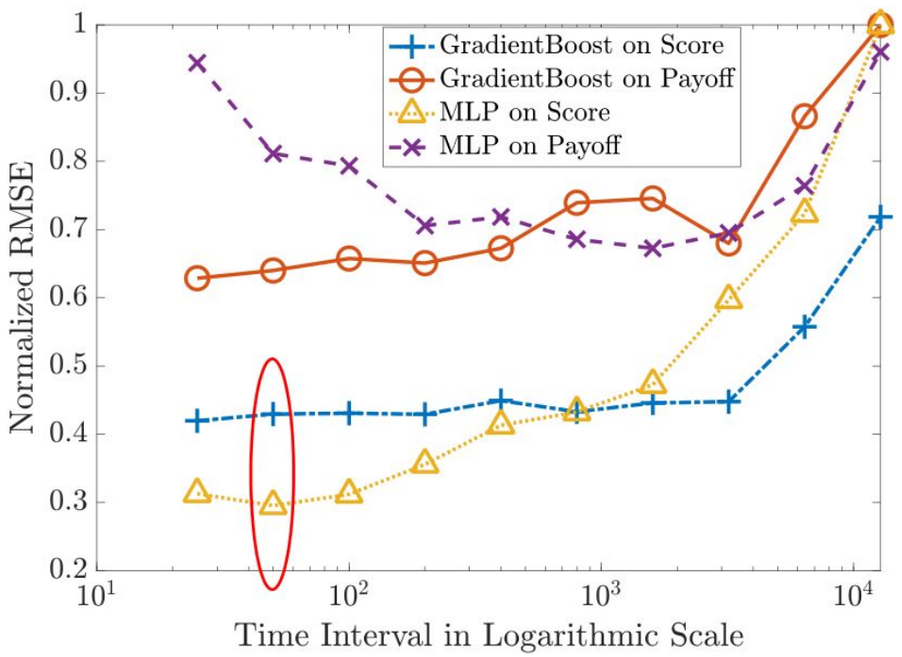


Distance distribution of users collecting resource & users failing

Berkeley Data Analytics – Group Innovation Modeling (4/5)

- ***Task:*** Predict the expected score of the entire team. We represent the innovation pattern with the combination of features, by training it to inform the score.
- ***Insight:*** We create auxiliary information to represent group innovation well, including team internal distance, distance variance, player location variance and moving area. They help improve prediction accuracy.

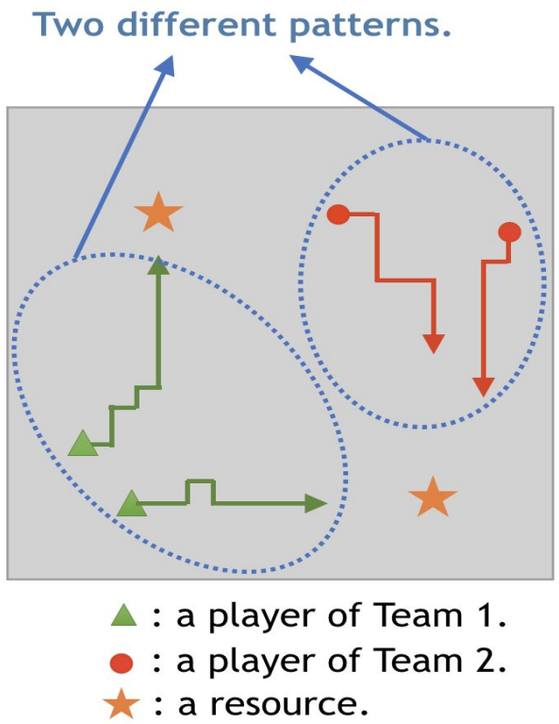
RMSE	Before Adding Auxiliary	After Adding Auxiliary
Score by MLP	48.40	45.52 (***6.0%***)
Payoff by GradientBoot_ls	55.98	42.70 ***(23.7%)***

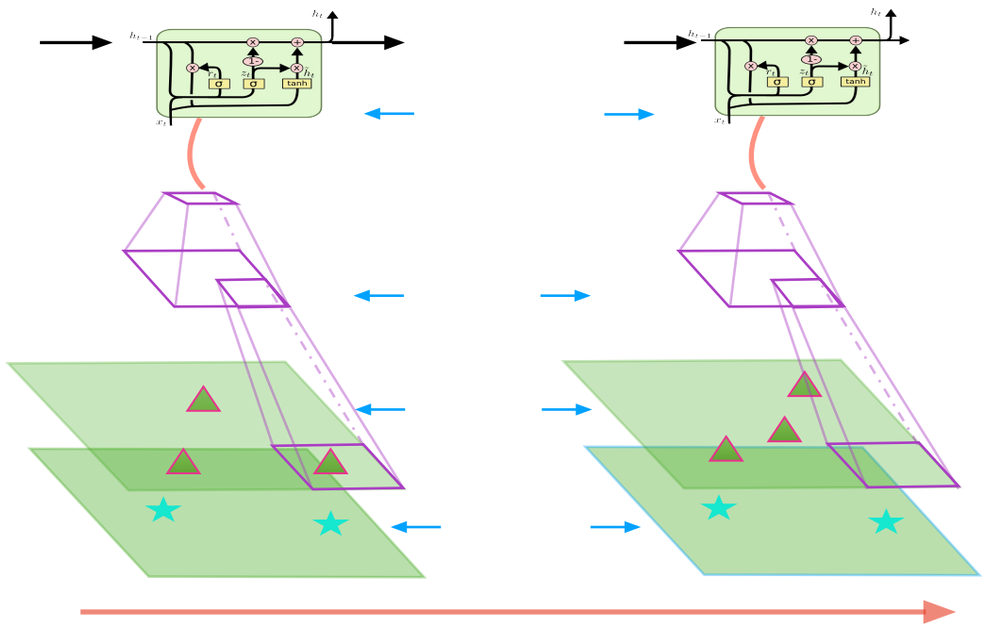


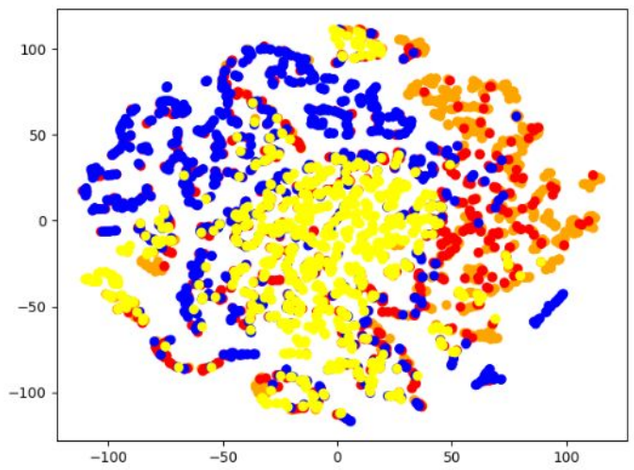
RMSE iteration during the training for two considered methods: MLP and GradientBoost

Berkeley Data Analytics – Group Innovation Representation (5/5)

- ***Definition:*** Group patterns correspond to the team strategies for collecting resources, which is reflected by the relative movement of players to resources.
- ***Task:*** Represent the group innovation pattern in high dimensional space.
- ***Method:*** Construct the video of player movement, propose architecture of CNN+LSTM to learn the pattern representation.
- ***Results:*** The learned representation shows the pattern of each group well.







Representation visualization of four teams

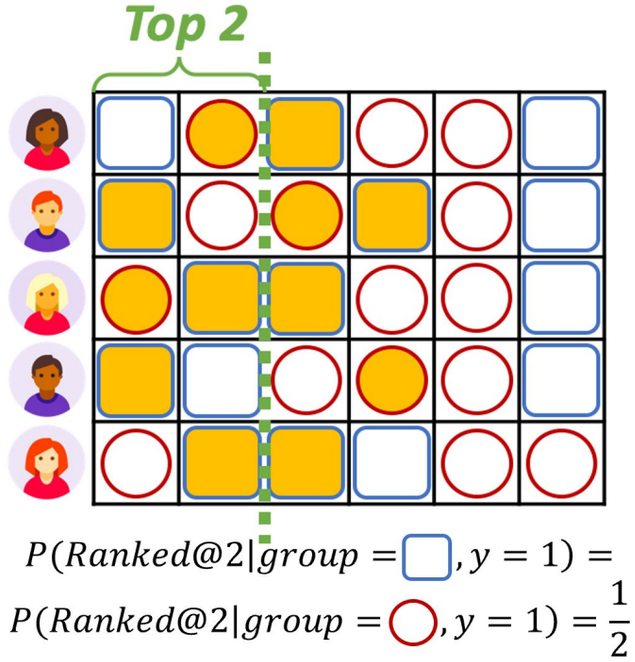
State-of-the-art Methods (1/3)

❖ **On Robustness of Neural Architecture Search under Label Noise**

- Published in Frontiers in Big Data, section Data Mining and Management.
- In this paper, we systematically explore the robustness of Neural Architecture Search under label noise. We demonstrate that the performance degradation under symmetric label noise can be mitigated by the use of robust loss functions.

❖ **Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems**

- Published in SIGIR Conference on Research and Development in Information Retrieval, 2020.
- Study the equal of opportunity problem in the personalized ranking recommendation task.



Debiasing Inequality of Opportunity in Personalized Ranking Recommendation

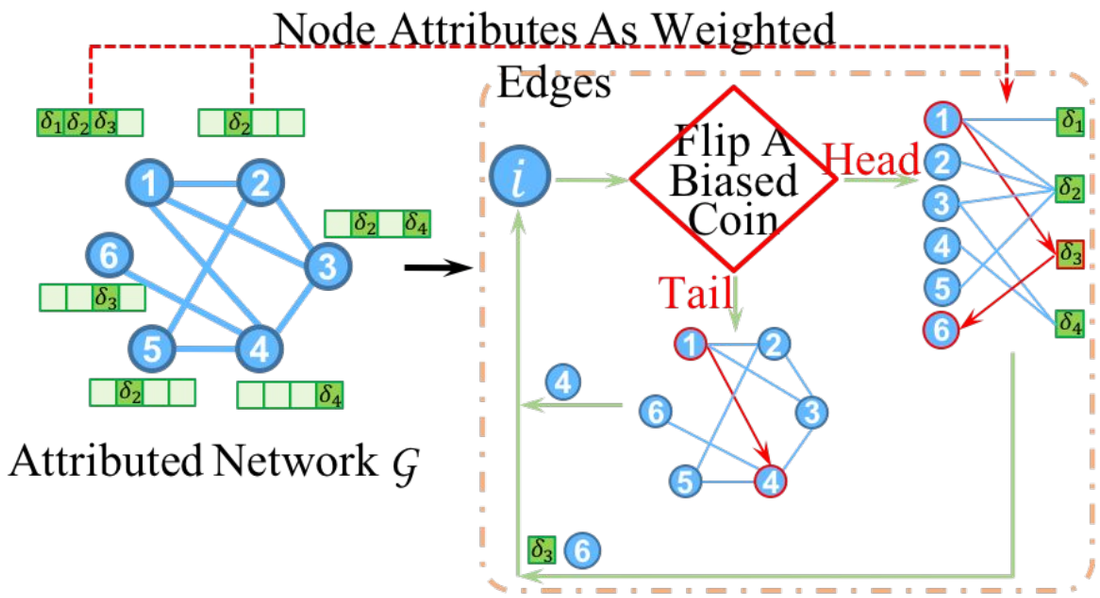
State-of-the-art Methods (2/3)

❖ **Coupled Variational Recurrent Collaborative Filtering.**

- Published in Twenty-fifth Conference on Knowledge Discovery and Data Mining (KDD).
- We focus on the problem of streaming recommender system and handle the data dynamicity and complexity in a streaming manner. By conjoining the complementary advantages of probabilistic models and deep neural networks, the proposed collaborative filtering algorithm performs favorably against the state-of-the-art methods.

❖ **Graph Recurrent Networks**

- Published in Twenty-fifth Conference on Knowledge Discovery and Data Mining (KDD).
- We apply random walks to attributed networks. We tailor graph neural networks to embed order information in attributed random walks.

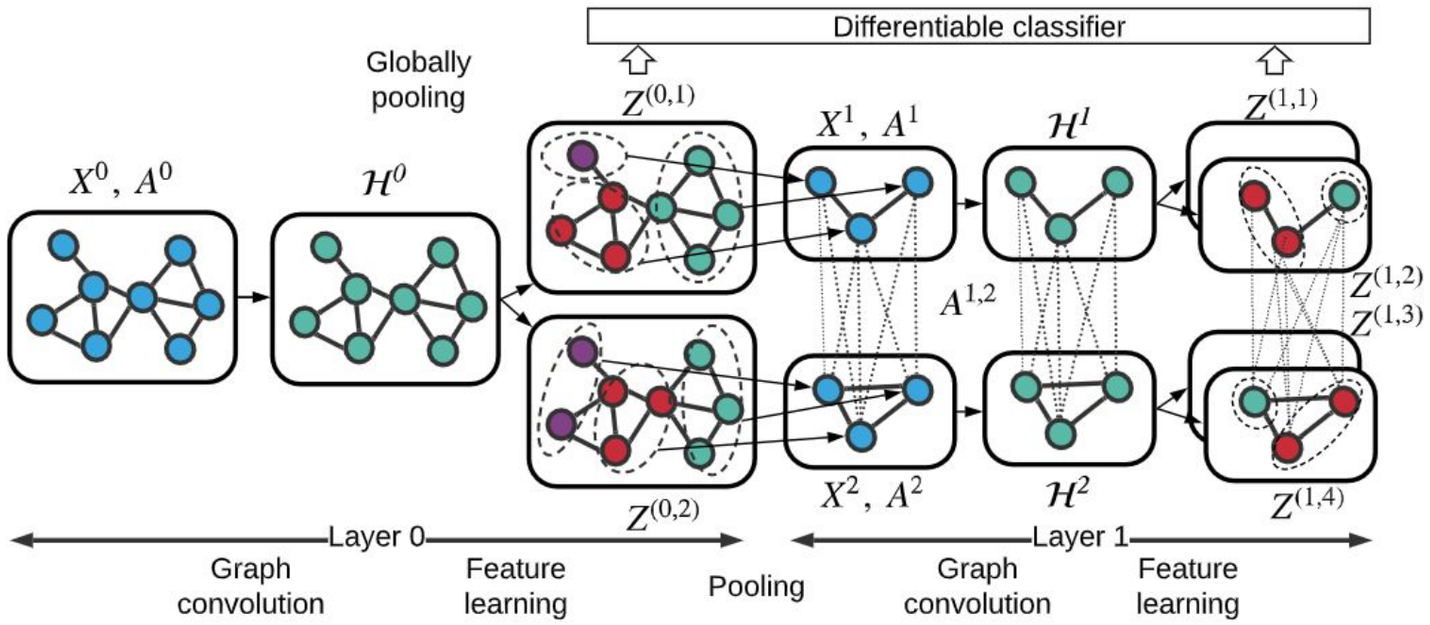


Graph Recurrent Networks

State-of-the-art Methods (3/3)

❖ **Multi-channel graph convolutional networks**

- Published in International Joint Conference on Artificial Intelligence (IJCAI 2020).
- Classify the graph-structured data to predict their associated labels. We Propose multi-channel graph neural networks (MuchGNN) to learn the graph embedding, via generating a set of coarse-grained structures with distinct characteristics to learn graph structure hierarchically.



Approach in Cycle 4

1 **Objectives Overview**

2 **Baseline Setup**

3 **Mutual Information Analysis**

4 **Sequential Group Innovation Analysis**

5 **Future Plans**

Objectives Overview

- ***Game description in Boomtown:*** A team consisting of several members plays the resource exploration game several rounds. Each round is associated with a binary variable to indicate the group innovation.
- ***Table data format:*** Each row represents the game information for a team at a specific round, including (1) X: game features (2) Y: binary innovation variable.
- ***Research problem:*** We formulate the Innovation analysis task as a binary classification problem. We target at:
  - Advancing model development to improve prediction accuracy.
  - Capturing the meaningful input features that are informative for group innovation.
  - Studying the interaction among input features, and comparing our results to pre-registration plan.

X | **Y**

Round	density	playernum	grmot1	grmot2	conformity	nPlayers	unanimity	unanimous	innovation
1	0.06666667	865.714286	NA	NA	NA	7	0.85714286	NA	NA
2	0.06666667	865.714286	0	0	0	7	0.85714286	0	0
3	0.06666667	865.714286	0	0	0	7	0.85714286	0	-1
4	0.19047619	596.285714	0.14285714	0.14285714	0	7	0.85714286	0	0

Baseline Setup

- ***Task:*** We apply traditional machine learning approaches for the binary classification problem.
- ***Data Cleaning:*** We clean the data by deleting invalid samples associated with NAN value, deleting uninformative features (e.g., group ID), filling empty features. We have 3312 samples, each of which is associated with 40 features and an innovation variable.
- ***Insights:*** (1) The group innovation could be captured by the baselines having classification accuracy larger than 0.5. (2) AdaBoost achieves the best performance as 79.2, via aggregating several weak classifiers (e.g., decision tree).
- ***Following methodologies:*** This result encourage us to explore more advanced models to improve the prediction accuracy.

Table 1: Mean accuracies in percent of baseline models.

Models	KNN	LR	NB	DT	SVC	RF	MLP	AdaBoost
Accuracy	64.5	78.0	76.7	74.2	78.8	78.1	78.6	79.2

Mutual Information Analysis

- ***Task:*** To analyze the relationships between game features and the group innovation level, we calculate mutual information metrics (1) between input features and group innovation variable; and (2) among input features.
- ***Insights:*** (1) The most informative features for group innovation include: Inmot 1, Inmot 2, Grmot 1, Grmot 2, which represent the individual or group motivation to innovate. (2) The feature pairs with larger value of mutual information: (tools, round), (tool, risk), (tool, Prb), since the tool selection of a player depends on the round number and tool risk/probability of success.
- ***Following methodologies:*** These results are in line with human knowledge. This encourages us to develop advanced interpretation model to explain the group innovation pattern.

Table 2: The mutual information between all features and innovation. Higher values represent the corresponding features are more related to innovation

settingsNum	**competition**	**timeUncertainty**	**support**	**nRound**	**nConnected**	**chat_per_round**	**tools**
0.0049	0.0010	0.0115	0.0000	0.0029	0.0030	0.0000	0.0487
inmot1	**inmot2**	**tolerance**	**round**	**group**	**pressure**	**toolsCPT**	**toolsEUT**
0.1235	0.1770	0.0083	0.0317	0.0362	0.0000	0.0161	0.0000
toolsPT	**toolsCPTEXP**	**risk**	**prb**	**structureHie**	**structureCel**	**structureNet**	**centralization**
0.0000	0.0261	0.0281	0.0418	0.0000	0.0000	0.0047	0.0000
leaderWeight	**compStrong**	**complexity**	**density**	**playernum**	**grmot1**	**grmot2**	**conformity**
0.0002	0.0156	0.0192	0.0091	0.0359	0.1001	0.1679	0.0022
nPlayers	**unanimity**	**framing_0**	**framing_1**	**framing_2**	**unanimous_-1**	**unanimous_0**	**unanimous_1**
0.0000	0.0000	0.0097	0.0000	0.0000	0.0044	0.0147	0.0044

Sequential Group Innovation Analysis

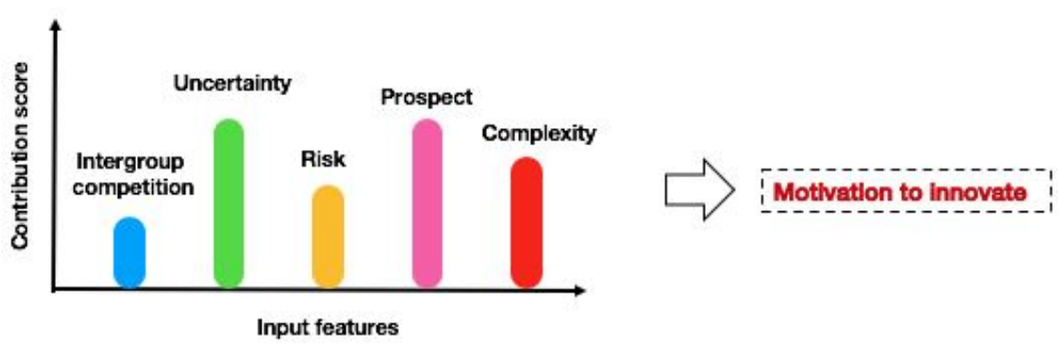
- ***Task:*** Following the previous experiences, we analyze the time dependency of group innovation.
- ***Method:*** We develop representation learning model based on an LSTM encoder to capture the sequential information from the previous rounds, and apply it to predict the innovation variable at the next round.
- ***Insights:*** (1) The prediction accuracy is improved by leveraging a short memory length of rounds. The group innovation has a short-term dependency on the group's historical behavior. (2) The more the memory rounds, the lower the accuracy is obtained, possibly since innovation preference with increasing rounds becomes too complex to be modeled well.

Table 1: Experimental results

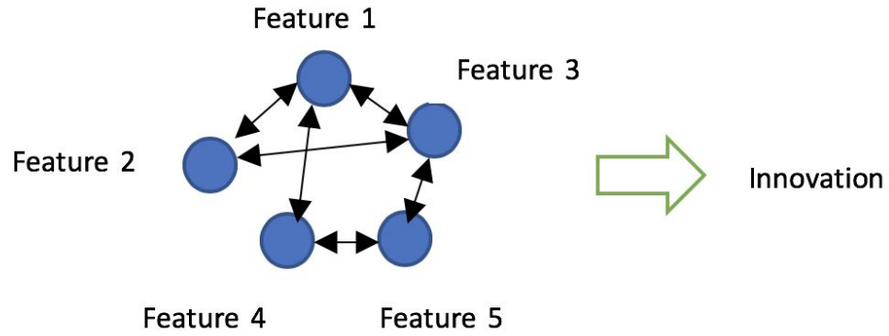
Target Round	2	3	4	5	6	7	8	9	10	11	12	13
Accuracy	0.82	0.75	0.69	0.85	0.74	0.75	0.74	0.75	0.61	0.58	0.54	0.54

Future Plans

- ***Feature selection:*** We plan to conduct feature selection to choose informative features and their combinations, in order to better capture innovation patterns and improve prediction accuracy. These crucial features can also explain group innovation pattern.
- ***Model interpretation:*** We try to explain the group innovation classification results, to identify the important features indicating the innovation pattern.
- ***Graph neural networks:*** We apply graph neural networks to model interactions between features explicitly, and then aggregate these features as well as their interaction weights to predict innovation pattern. The explicit weights help to explain the correlation between features, which may also improve innovation classification.



Model Interpretation



Graph neural networks

Risks and Mitigation

❖ ***Feature selection***

- ***Risk:*** Given the ~40 input features, traditional feature selection may not determine the informative features effectively.
- ***Mitigation:*** We apply human knowledge to narrow down the search space of input features, to facilitate the following feature selection process.

❖ ***Graph neural networks***

- ***Risk:*** It may be hard to construct the edge connections between input features in the table data, since there is little evidence to indicate the relation of features.
- ***Mitigation:*** We compute the mutual information and set a threshold to determine edge connection. Or we apply a graph attention layer to learn the connection weight automatically.

Outside Transition

❖ ***Technology transfer and commercialization***

PyTen: An efficient software of distributed large-scale tensor completion recovers the missing values for real-world datasets. Link for this package: https://github.com/tamu-helios/pyten

Commercialization: We transfer developed frameworks to identify outliers from graph and boundary graphs with applications in cyber security, manufacturing, etc.

❖ ***Further development***

1. CAREER: Human-Centric Big Network Embedding Period of Performance: 2018-2023, funded by National Science Foundation
2. III: Small: Collaborative Research: A General Feature Learning Period of Performance: 2017-2020, funded by National Science Foundation
3. III: Small: Collaborative Research: Modeling and Managing Extremist Group Influence in Massive Social Media Networks: 2019-2022, funded by National Science Foundation

Representation Learning – Overview

❖ ***Objective:*** Helios representation learning framework is designed to analyze real-world big social data via three main thrusts: tensor completion, ensemble approach, and innovation representation learning.