

Data Mining

Course Overview and Logistics

<https://data-mining.github.io/winter-2026/>

CS 453/553 – Winter 2026

Yu Wang, Ph.D.

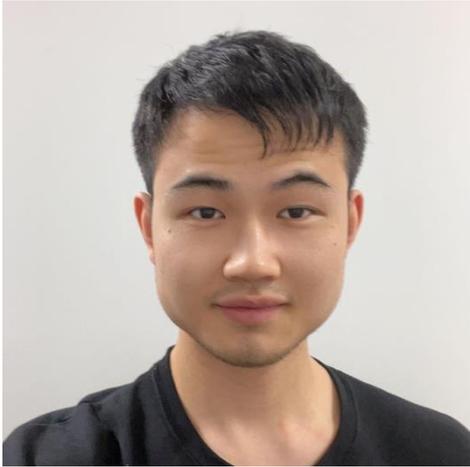
Assistant Professor

Computer Science

University of Oregon



Self-Introduction



Yu (Jack) Wang
(You)

Contact:
yuwang@uoregon.edu

<https://yuwang0103.github.io/>

Research Interests:

- Data Mining and Machine Learning
- Neural-Symbolic Learning
- Graph and Network
- LLM + Structured Knowledge
- AI/ML/DM Applications
 - Document Intelligence
 - Social Computing
 - Networking Physical Infrastructure

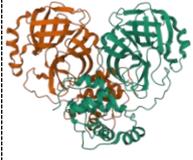


📣📣📣 Recruiting Ph.D. students and interns! I am actively seeking highly motivated students for Ph.D. or Research intern positions. Please feel free to email me your CV, transcripts, and brief descriptions about why you want to work with me if you are interested!

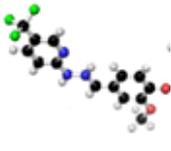


What is Data?

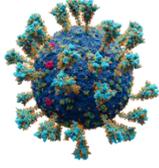
Science



Protein



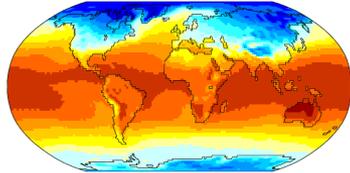
Small Molecule



Virus

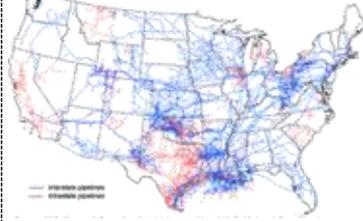


Brain Neural

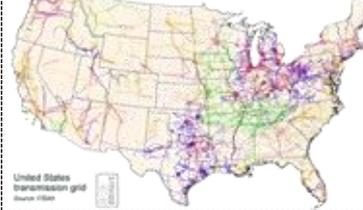


Surface Temperature of Earth

Gas Network

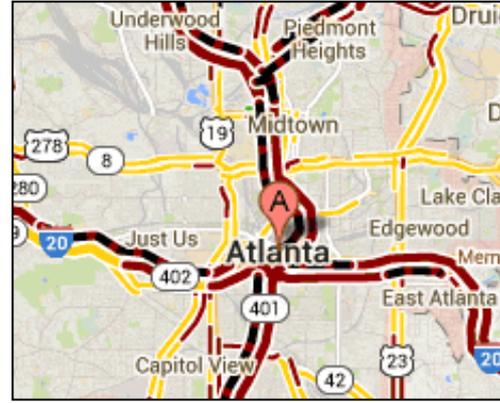


Power Network

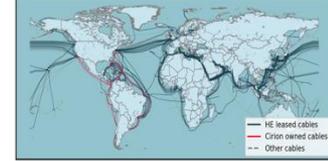


Infrastructure

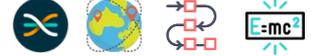
Transportation Network



Submarine Cable



Terrestrial Cable



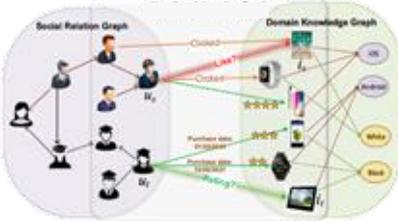
Social Network



Citation Network



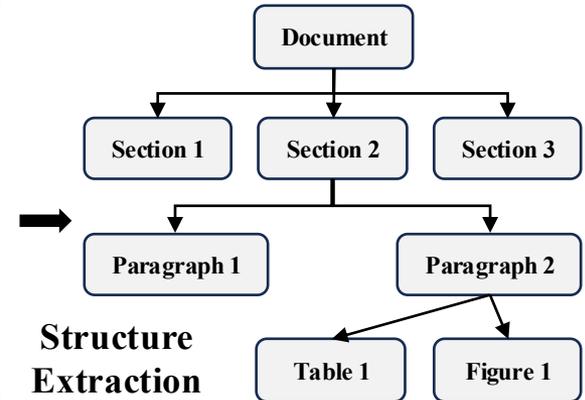
Transaction Network



User-Entity Interaction Graph



Document



Structure Extraction

Virtual Village with AI Agents





Why Analyze Data? – Paper Management

Google Scholar



Articles Case law

New! Scholar Labs: An AI Powered Scholar Search

Recommended articles



☆ Analyzing the Properties of Graph Neural Networks with Evolutionary Algorithms ▼
 Z Liu, Z Lu, H Wang, D Chen, S Wang, J Chu, R Gao, A Jiang
 Mathematics - 3 days ago HTML

☆ Self-Supervised Bipartite Graph Neural Networks with Missing Value Imputation for Small Tabular Data Predictions ▼
 PC Liu, CT Li
 ACM Transactions on Intelligent Systems and Technol... - 4 days ago PDF

[More articles from 4 days ago](#)

☆ HRGNN: Learning Holistically Robust Graph Neural Networks on Noisy Graphs with Label Scarcity ▼
 JW Chiu, CT Li
 ACM Transactions on Intelligent Systems and Technol... - 5 days ago PDF

REFERENCES

Art of Problem Solving. Aime problems and solutions, 2025. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. 8, 22

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, et al. ReSearch: Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025. 2, 4, 7, 10, 21

Zihao Cheng, Hongru Wang, Zeming Liu, Yuhang Guo, Yuanfang Guo, Yunhong Wang, and Haifeng Wang. ToolSpectrum: Towards personalized tool utilization for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20679–20699, 2025. 10

Yingfan Deng, Anhao Zhou, Yuan Yuan, Xian Zhang, Yifei Zou, and Dongxiao Yu. Pe-ma: Parameter-efficient co-evolution of multi-agent systems. *arXiv preprint arXiv:2506.11803*, 2025. 11

Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. Tool-star: Empowering llm-brained multi-tool reasoner via reinforcement learning. *arXiv preprint arXiv:2505.16410*, 2025. 2, 10

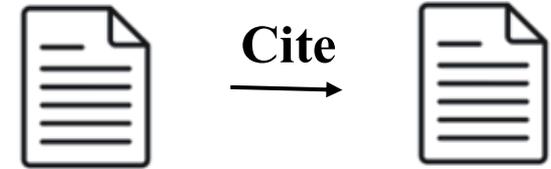
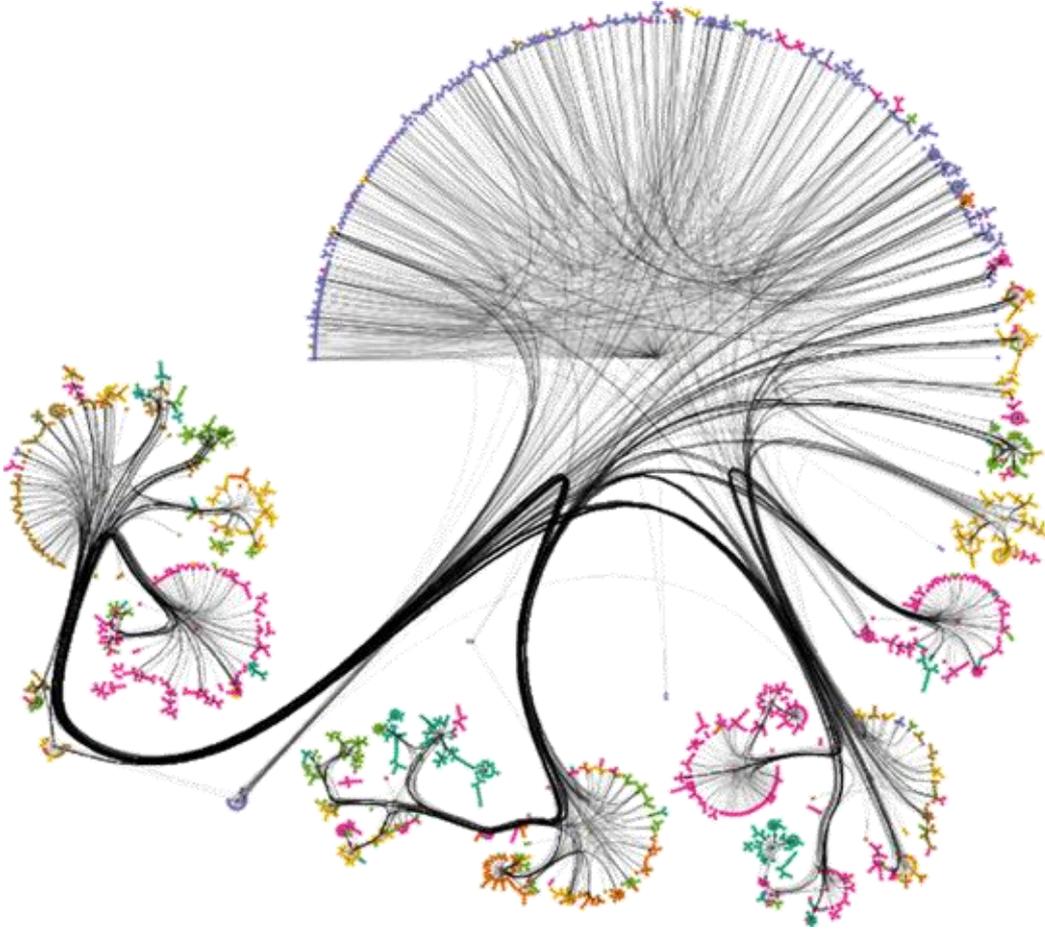


Cite





Why Analyze Data? – Paper Management



$$\frac{\sum_{e_{ij} \in \mathcal{E}} 1[y_i == y_j]}{|\mathcal{E}|}$$

\mathcal{E} - Total Number of Edges

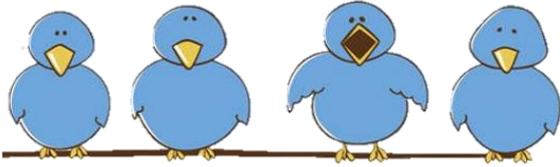
e_{ij} - Edge between node i/j

y_i - Label of i

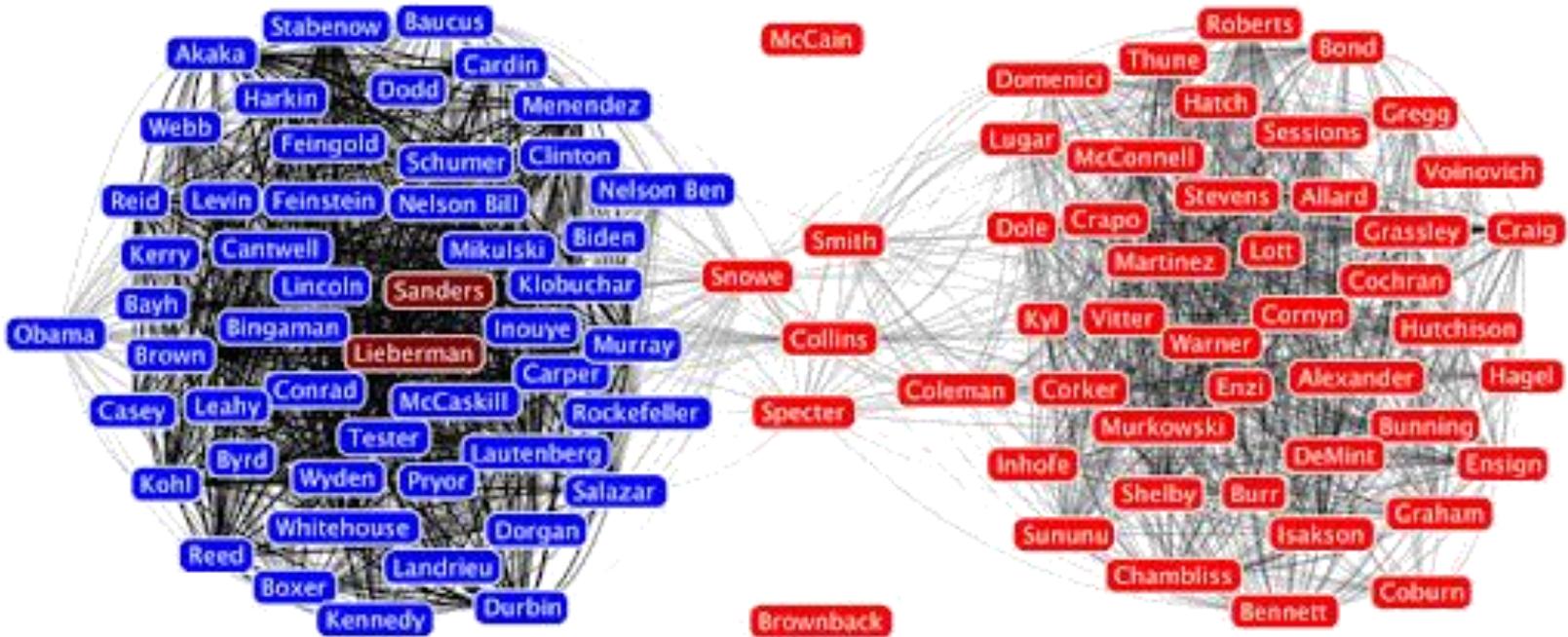
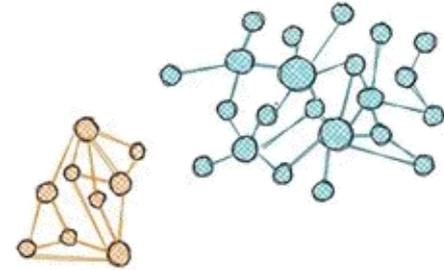




Why Analyze Data? – Paper Management



Birds of a feather flock together

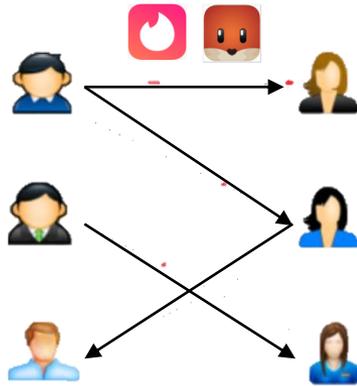


Gun Control Belief Network

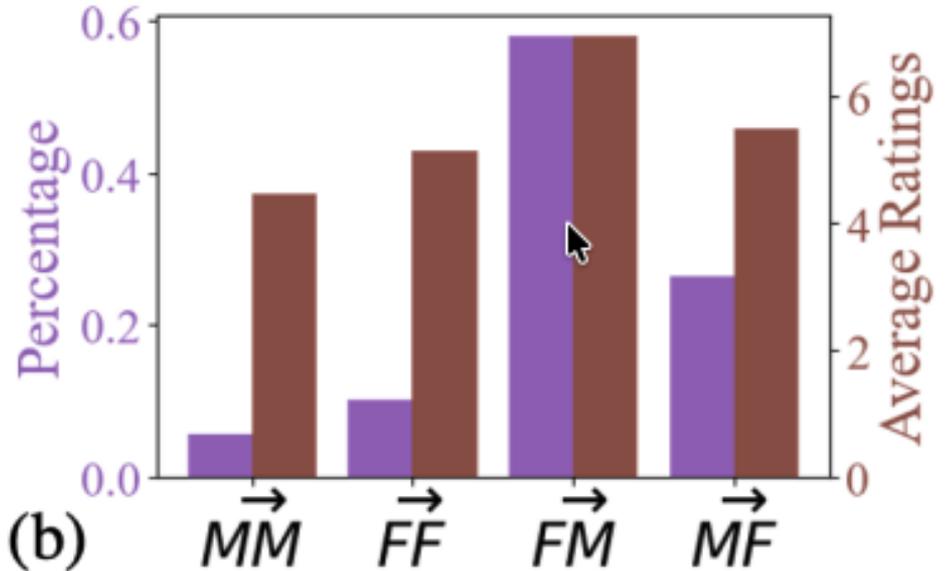
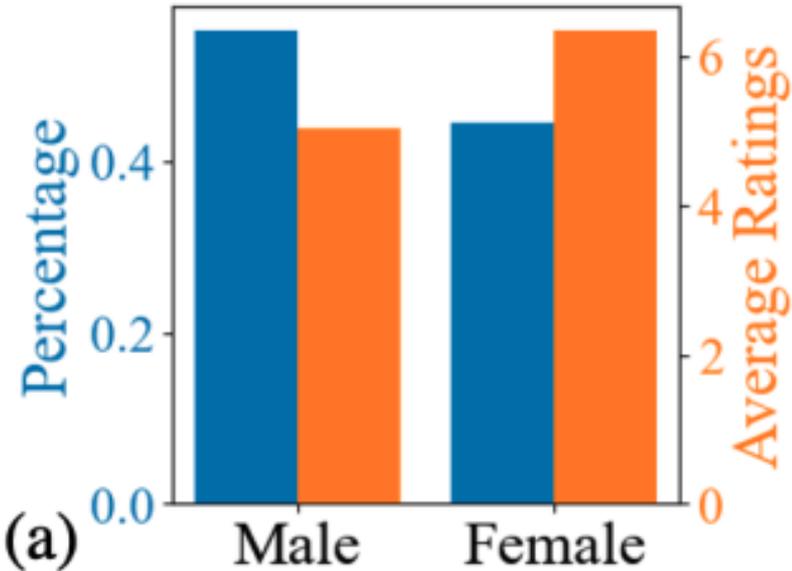
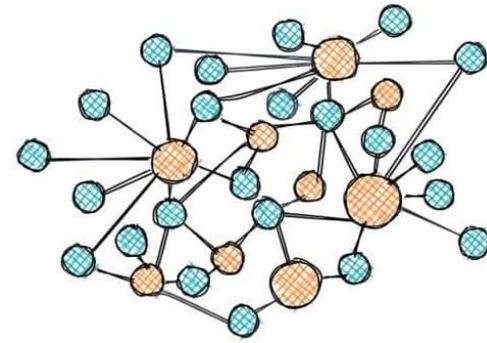




Why Analyze Data? – Paper Management



Dating Network



Dating Network



Why Analyze Data? – Paper Management

IN-THE-FLOW AGENTIC SYSTEM OPTIMIZATION FOR EFFECTIVE PLANNING AND TOOL USE

Zhuofeng Li^{*1,2}, Haoxiang Zhang^{*1,3}, Seungju Han¹, Sheng Liu¹, Jianwen Xie⁴, Yu Zhang², Yejin Choi¹, James Zou^{1†}, Pan Lu^{1†}

¹Stanford University, ²Texas A&M University, ³UC San Diego, ⁴Lambda

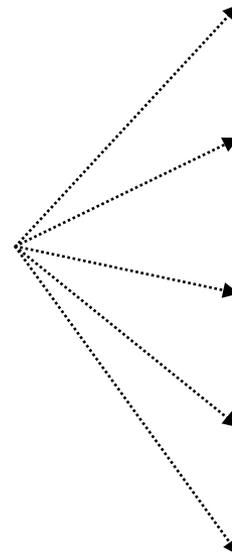


Website: <https://agentflow.stanford.edu>

Code Model Demo Visualize

ABSTRACT

Outcome-driven reinforcement learning has advanced reasoning in large language models (LLMs), but prevailing tool-augmented approaches train a single, monolithic policy that interleaves thoughts and tool calls under full context; this scales poorly with long horizons and diverse tools and generalizes weakly to new scenarios. Agentic systems offer a promising alternative by decomposing work across specialized modules, yet most remain training-free or rely on offline training decoupled from the live dynamics of multi-turn interaction. We introduce AGENTFLOW, a trainable, *in-the-flow* agentic framework that coordinates four modules (planner, executor, verifier, generator) through an evolving memory and directly optimizes its planner inside the multi-turn loop. To train on-policy in live environments, we propose *Flow-based Group Refined Policy Optimization* (Flow-GRPO), which tackles long-horizon, sparse-reward credit assignment by converting multi-turn optimization into a sequence of tractable single-turn policy updates. It broadcasts a single, verifiable trajectory-level outcome to every turn to align local planner decisions with global success and stabilizes learning with group-normalized advantages. Across ten benchmarks, AGENTFLOW with a 7B-scale backbone outperforms top-performing baselines with average accuracy gains of 14.9% on search, 14.0% on agentic, 14.5% on mathematical, and 4.1% on scientific tasks, even surpassing larger proprietary models like GPT-4o. Further analyses confirm the benefits of in-the-flow optimization, showing improved planning, enhanced tool-calling reliability, and positive scaling with model size and reasoning turns.



In-the-flow agentic system optimization for effective planning and tool use

Search within citing articles

Latent collaboration in multi-agent systems

[J.Zou](#), [X.Yang](#), [R.Qiu](#), [G.Li](#), [K.Tieu](#), [P.Lu](#), K Shen... - arXiv preprint arXiv ..., 2025 - arxiv.org
Multi-agent systems (MAS) extend large language models (LLMs) from independent single-model reasoning to coordinative system-level intelligence. While existing LLM agents ...

☆ Save Cite Cited by 4 Related articles All 2 versions

Adaptation of agentic ai

[P.Jiang](#), [J.Lin](#), [Z.Shi](#), [Z.Wang](#), L He, [Y.Wu](#)... - arXiv preprint arXiv ..., 2025 - arxiv.org
Cutting-edge agentic AI systems are built on foundation models that can be adapted to plan, reason, and interact with external tools to perform increasingly complex and specialized ...

☆ Save Cite Cited by 2 Related articles All 2 versions

DeepAgent: A General Reasoning Agent with Scalable Toolsets

[X.Li](#), [W.Jiao](#), [J.Jin](#), [G.Dong](#), [J.Jin](#), Y Wang... - arXiv preprint arXiv ..., 2025 - arxiv.org
Large reasoning models have demonstrated strong problem-solving abilities, yet real-world tasks often require external tools and long-horizon interactions. Existing agent frameworks ...

☆ Save Cite Cited by 2 Related articles All 2 versions

The Path Not Taken: RLVR Provably Learns Off the Principals

[H.Zhu](#), [Z.Zhang](#), [H.Huang](#), DJ Su, [Z.Liu](#), [J.Zhao](#)... - arXiv preprint arXiv ..., 2025 - arxiv.org
Reinforcement Learning with Verifiable Rewards (RLVR) reliably improves the reasoning performance of large language models, yet it appears to modify only a small fraction of ...

☆ Save Cite Cited by 1 Related articles All 3 versions

Self-Play Methods in Reinforcement Learning for Language Models

[Z.Ye](#) - 2025 - knowledge.uchicago.edu
In this thesis we develop a series of practical algorithms for language model to self-train, by actively and strategically creating and controlling learning experiences themselves ...

☆ Save Cite Related articles All 2 versions

Which category does this paper belong to?



Tool Learning

Agentic Learning





Why Analyze Data? – Paper Management

IN-THE-FLOW AGENTIC SYSTEM OPTIMIZATION FOR EFFECTIVE PLANNING AND TOOL USE

Zhaofeng Li^{1,2}, Haosiang Zhang^{1,2}, Seungja Han¹, Sheng Liu¹, Sheng Liu¹, Jianwen Xie¹, Yu Zhang¹, Yujin Choi¹, James Zou¹, Pan Lu¹
¹Stanford University, ²Texas A&M University, ³UC San Diego, ⁴Lambda

Website: <https://agentflow.stanford.edu>
Code Model Demo Visualize

ABSTRACT

Outcome-driven reinforcement learning has advanced reasoning in large language models (LLMs), but prevailing tool-augmented approaches train a single, monolithic policy that interleaves thoughts and tool calls under full context; this scales poorly with long horizons and diverse tools and generalizes weakly to new scenarios. Agentic systems offer a promising alternative by decomposing work across specialized modules, yet must remain training-free or rely on offline training decoupled from the live dynamics of multi-turn interaction. We introduce AGENT-FLOW, a trainable, *in-the-flow* agentic framework that coordinates four modules (planner, executor, verifier, generator) through an evolving memory and directly optimizes its planner inside the multi-turn loop. To train on-policy in live environments, we propose *Flow-based Group Refined Policy Optimization* (Flow-GRPO), which tackles long-horizon, sparse-reward credit assignment by converting multi-turn optimization into a sequence of tractable single-turn policy updates. It broadens a single, verifiable trajectory-level outcome to every turn to align local planner decisions with global success and stabilizes learning with group-normalized advantages. Across ten benchmarks, AGENT-FLOW with a 7B-scale backbone outperforms top-performing baselines with average accuracy gains of 14.9% on search, 14.0% on agentic, 14.5% on mathematical, and 4.1% on scientific tasks, even surpassing larger proprietary models like GPT-4o. Further analyses confirm the benefits of *in-the-flow* optimization, showing improved planning, enhanced tool-calling reliability, and positive scaling with model size and reasoning turns.



IN-THE-FLOW AGENTIC SYSTEM OPTIMIZATION FOR EFFECTIVE PLANNING AND TOOL USE

Zhaofeng Li^{1,2}, Haosiang Zhang^{1,2}, Seungja Han¹, Sheng Liu¹, Sheng Liu¹, Jianwen Xie¹, Yu Zhang¹, Yujin Choi¹, James Zou¹, Pan Lu¹
¹Stanford University, ²Texas A&M University, ³UC San Diego, ⁴Lambda

Website: <https://agentflow.stanford.edu>
Code Model Demo Visualize

ABSTRACT

Outcome-driven reinforcement learning has advanced reasoning in large language models (LLMs), but prevailing tool-augmented approaches train a single, monolithic policy that interleaves thoughts and tool calls under full context; this scales poorly with long horizons and diverse tools and generalizes weakly to new scenarios. Agentic systems offer a promising alternative by decomposing work across specialized modules, yet must remain training-free or rely on offline training decoupled from the live dynamics of multi-turn interaction. We introduce AGENT-FLOW, a trainable, *in-the-flow* agentic framework that coordinates four modules (planner, executor, verifier, generator) through an evolving memory and directly optimizes its planner inside the multi-turn loop. To train on-policy in live environments, we propose *Flow-based Group Refined Policy Optimization* (Flow-GRPO), which tackles long-horizon, sparse-reward credit assignment by converting multi-turn optimization into a sequence of tractable single-turn policy updates. It broadens a single, verifiable trajectory-level outcome to every turn to align local planner decisions with global success and stabilizes learning with group-normalized advantages. Across ten benchmarks, AGENT-FLOW with a 7B-scale backbone outperforms top-performing baselines with average accuracy gains of 14.9% on search, 14.0% on agentic, 14.5% on mathematical, and 4.1% on scientific tasks, even surpassing larger proprietary models like GPT-4o. Further analyses confirm the benefits of *in-the-flow* optimization, showing improved planning, enhanced tool-calling reliability, and positive scaling with model size and reasoning turns.

In-the-flow agentic system optimization for effective planning and tool use

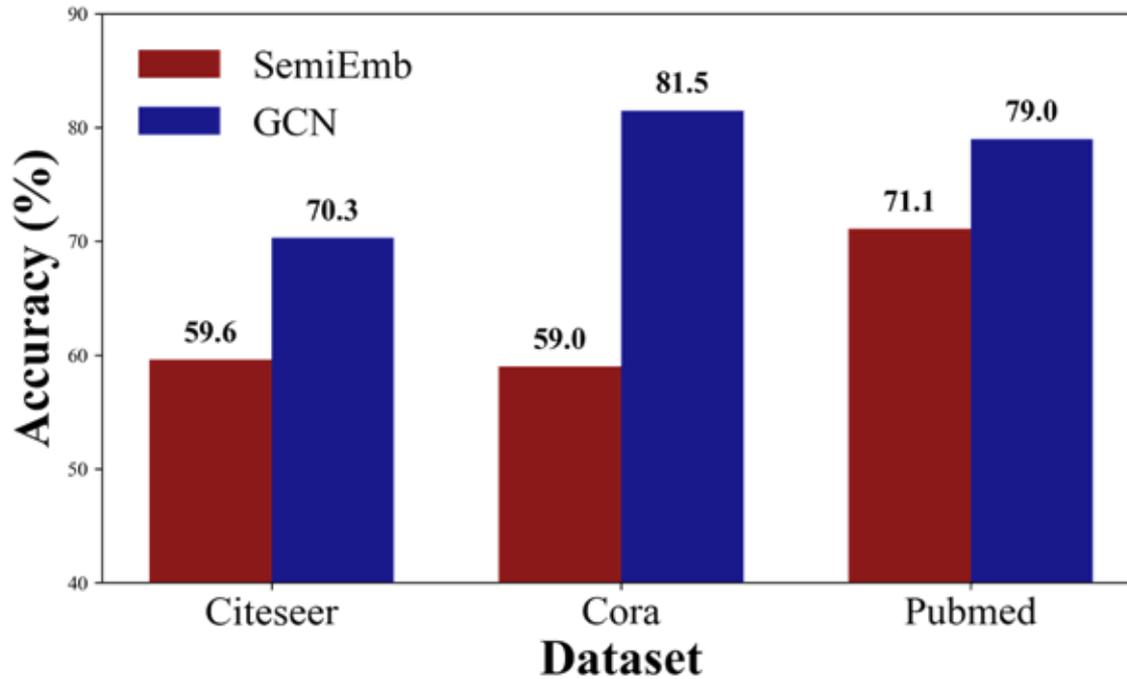
Latent collaboration in multi-agent systems
J.Zou, X.Zhou, S.Liu, G.Li, C.Tan, P.Lu, H.Shen, ... arXiv preprint arXiv: ... 2025 - arxiv.org
Multi-agent systems (MAS) extend large language models (LLMs) from independent single-model reasoning to coordinative system-level intelligence. While existing LLM agents ...
arXiv: 2501.01001v2 [cs.LG] 2025-01-04. Related articles: AI-2 versions: 10

Adaptation of agentic AI
F.Pan, L.Liu, Z.Wu, Z.Wang, Y.Hu, Y.Liu, ... arXiv preprint arXiv: ... 2025 - arxiv.org
Cutting-edge agentic AI systems are built on foundation models that can be adapted to plan, reason, and interact with external tools to perform increasingly complex and specialized ...
arXiv: 2501.01001v2 [cs.LG] 2025-01-04. Related articles: AI-2 versions: 10

DeepAgent: A General Reasoning Agent with Scalable Toolsets
S.Li, Y.Kuan, J.Zou, S.Zhang, L.Zou, Y.Wang, ... arXiv preprint arXiv: ... 2025 - arxiv.org
Large reasoning models have demonstrated strong problem-solving abilities, yet real-world tasks often require external tools and long-horizon interactions. Existing agent frameworks ...
arXiv: 2501.01001v2 [cs.LG] 2025-01-04. Related articles: AI-2 versions: 10

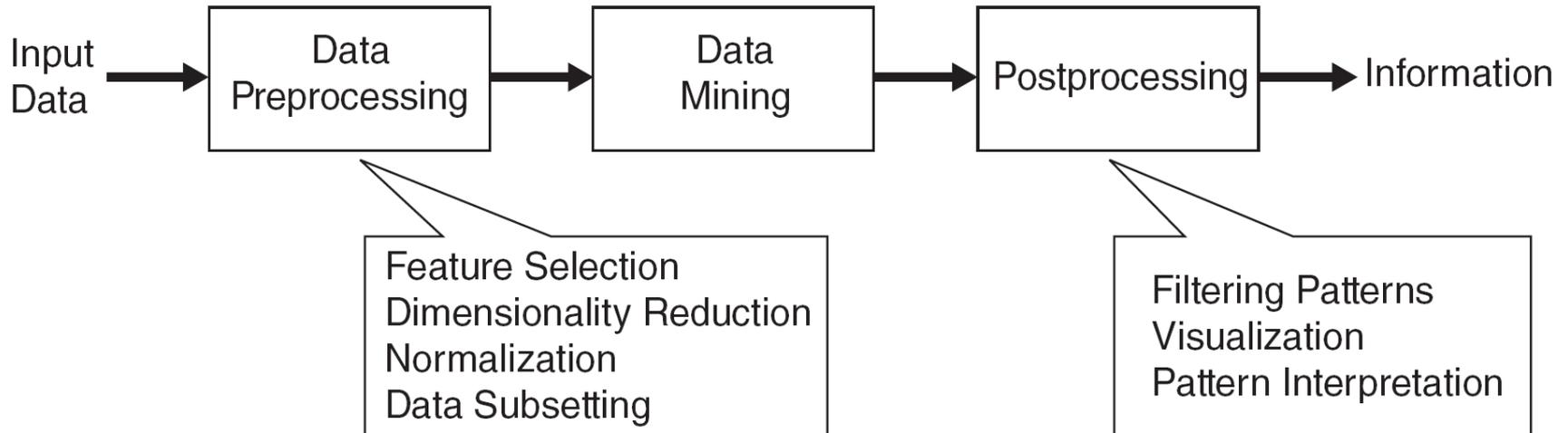
The Path Not Taken: RLVR Provably Learns Off the Principals
S.Liu, Z.Zhang, H.Huang, D.Liu, Z.Liu, J.Zhou, ... arXiv preprint arXiv: ... 2025 - arxiv.org
Reinforcement Learning with Verifiable Rewards (RLVR) reliably improves the reasoning performance of large language models, yet it appears to modify only a small fraction of ...
arXiv: 2501.01001v2 [cs.LG] 2025-01-04. Related articles: AI-2 versions: 10

Self-Play Methods in Reinforcement Learning for Language Models
Z.Liu, 2025 - knowledgecommons.org
In this thesis we develop a series of practical algorithms for language model to self-train, by actively and strategically creating and controlling learning experiences themselves ...
arXiv: 2501.01001v2 [cs.LG] 2025-01-04. Related articles: AI-2 versions: 10





What is Data Mining?

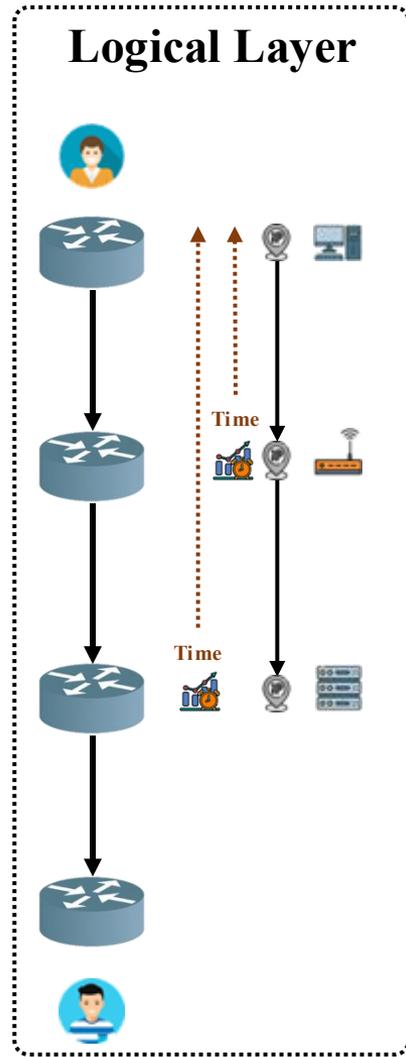
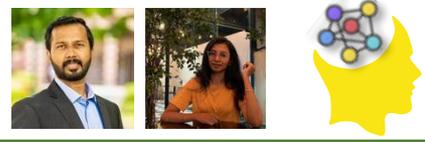


Many Definitions

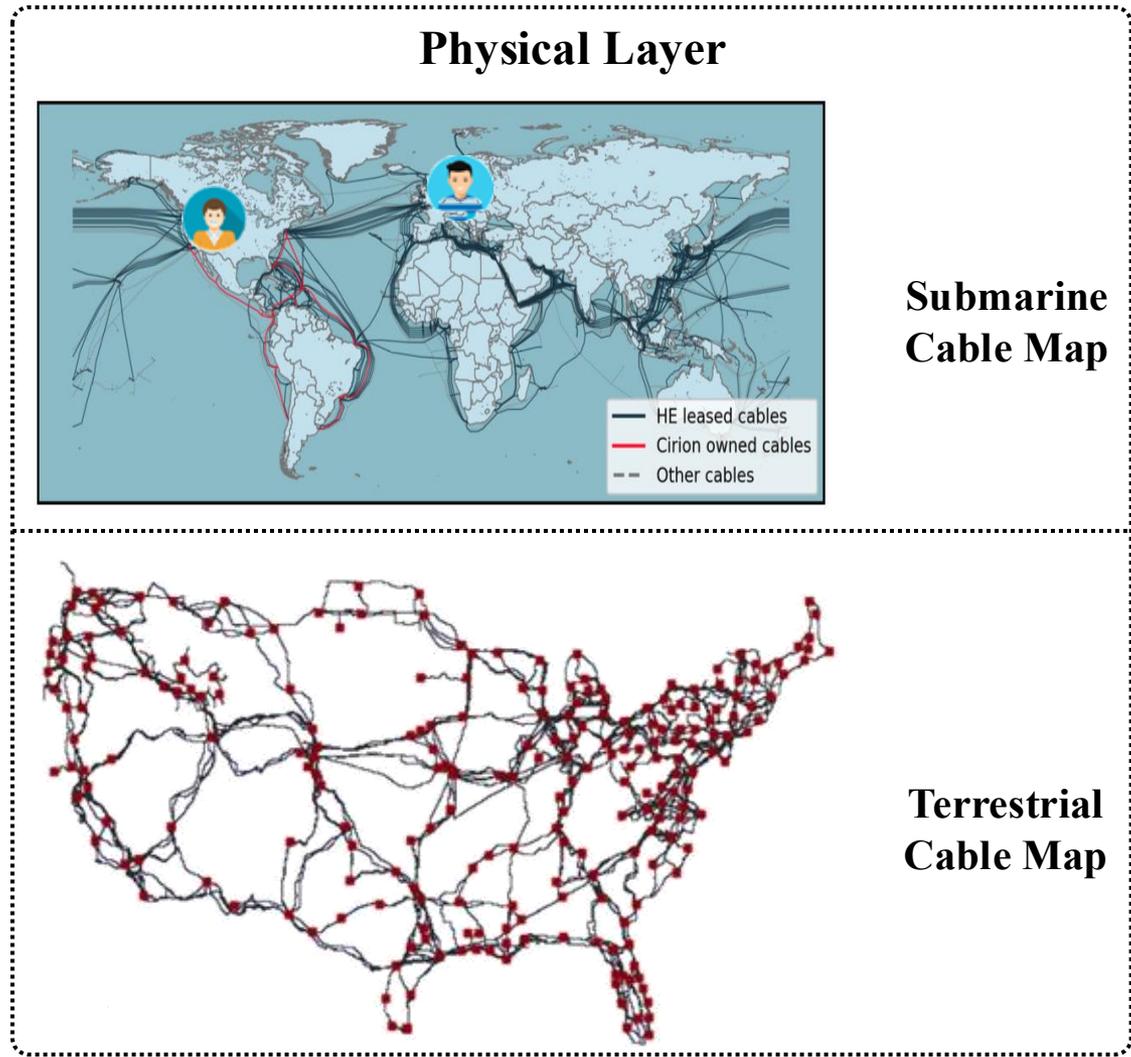
Non-trivial extraction of implicit, previously unknown and potentially useful information from data

Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

Why Data Mining? – Networking Infra Risk **ONRG**



Mapping

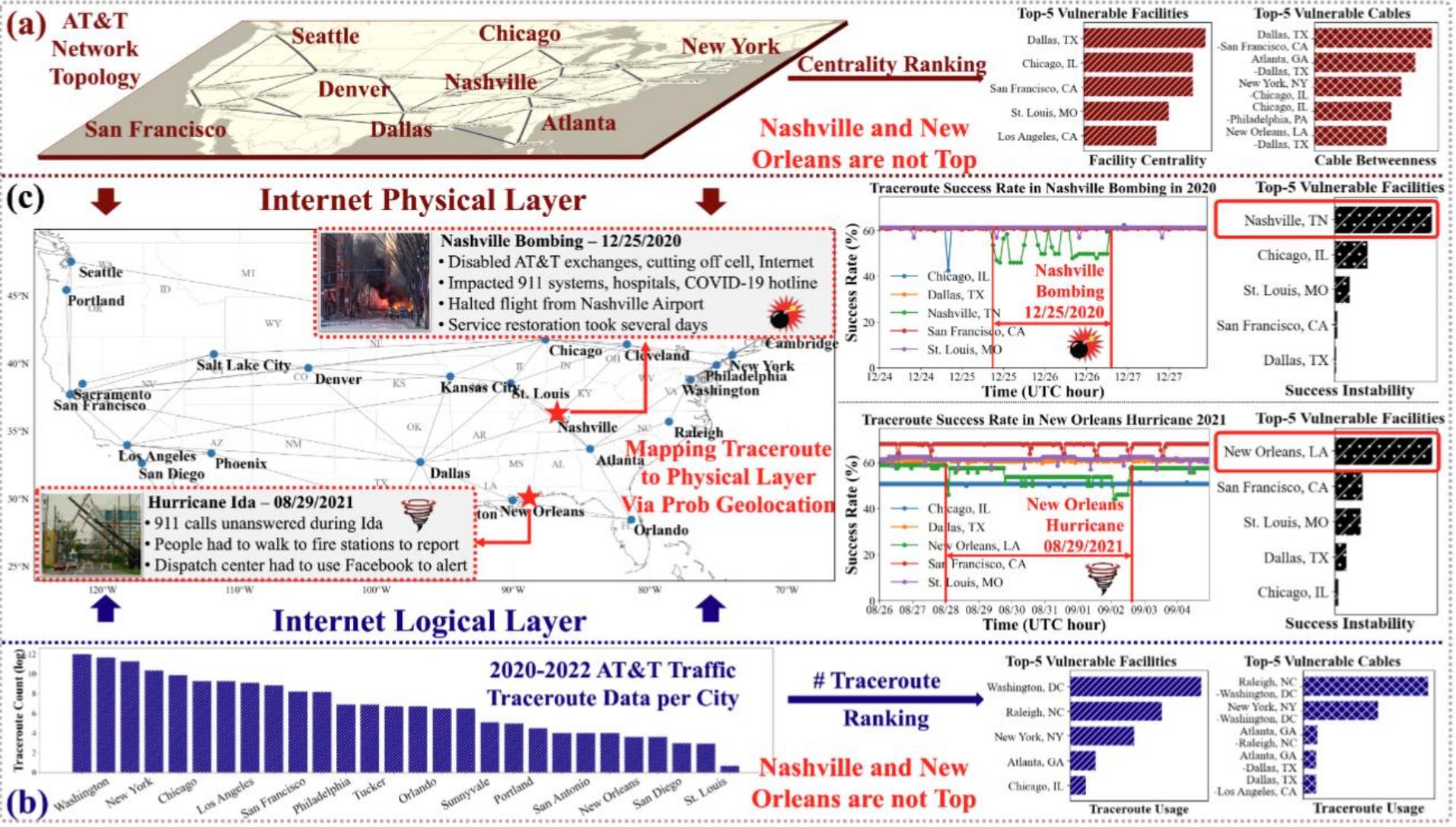


Submarine Cable Map

Terrestrial Cable Map

Which physical cable path does this logic signal traverse?

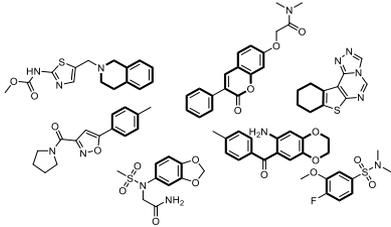
Why Data Mining? – Networking Infra Risk ONRG





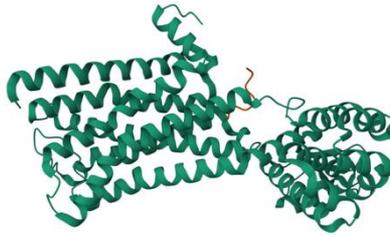
Why Data Mining? – Drug Design

Chemical Libraries

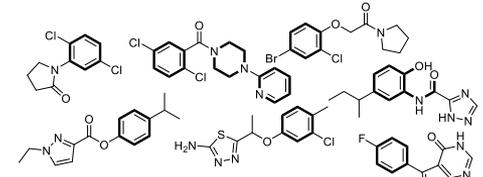


Number of Molecules: 103-106

Protein Target



Virtual Libraries



e.g., 10^9 Virtual Molecules on the REAL database in Enamine Ltd.



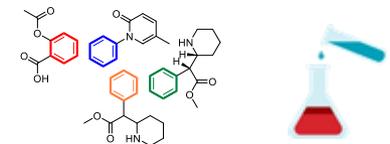
High Throughput Screening (HTS)

Training



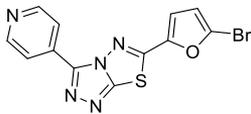
Deep Learning Models

Predicted Actives



Number of Molecules: 500-1000

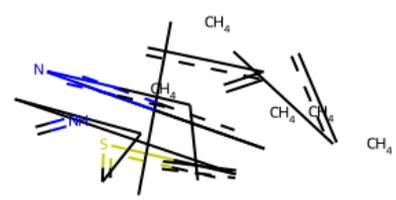
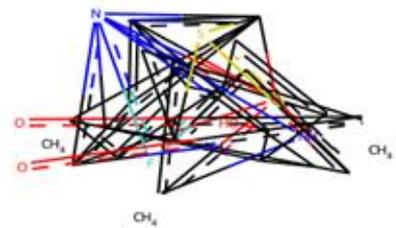
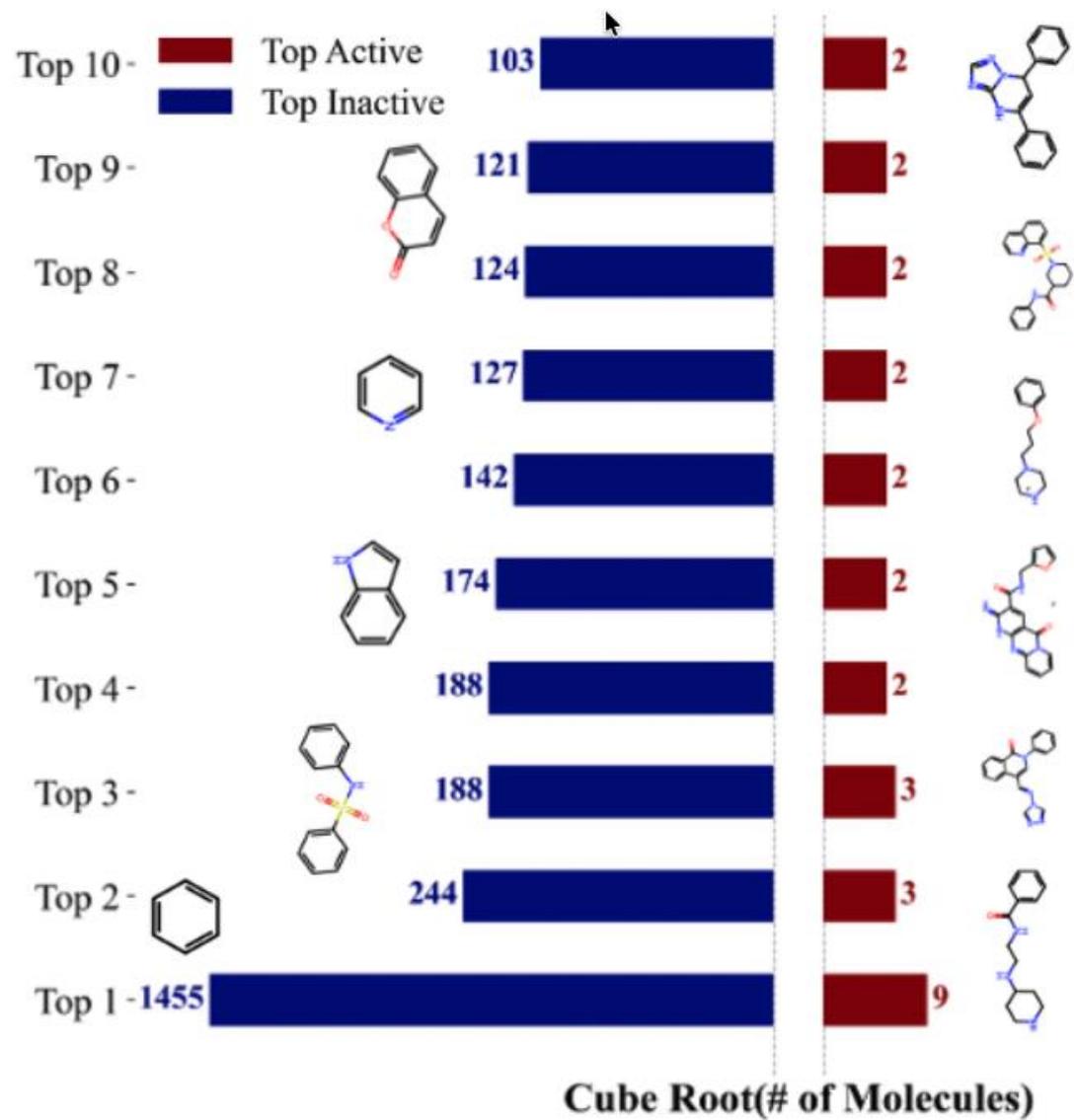
Evaluating



Hit Rate: 0.05%-0.5%



Why Data Mining? – Drug Design





Why Data Mining? – Commercial Perspective

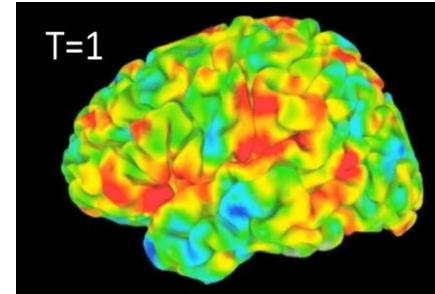
- Lots of data is being collected and warehoused
 - Web data **1,000 terabytes,**
1,000,000,000,000,000= bytes
 - Google has Peta Bytes of web data
 - Facebook has billions of active users
 - purchases at department/
grocery stores, e-commerce
 - Amazon handles millions of visits/day
 - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
 - Provide better, customized services for an edge (e.g. in Customer Relationship Management)





Why Data Mining? – Scientific Perspective

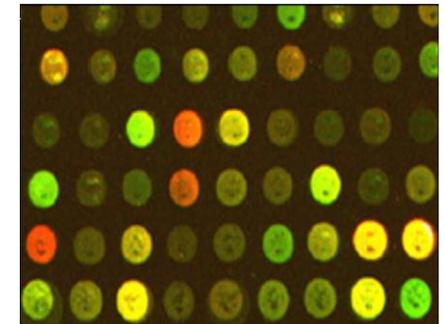
- Data collected and stored at enormous speeds
 - Remote sensors on a satellite
 - NASA EOSDIS archives over petabytes of earth science data / year
 - Telescopes scanning the skies
 - Sky survey data
 - High-throughput biological data
 - Scientific simulations
 - terabytes of data generated in a few hours
- Data mining helps scientists
 - in automated analysis of massive datasets
 - In hypothesis formation



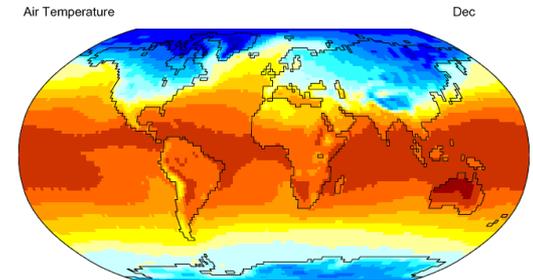
fMRI Data from Brain



Sky Survey Data



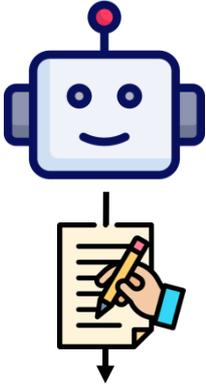
Gene Expression Data



Surface Temperature of Earth

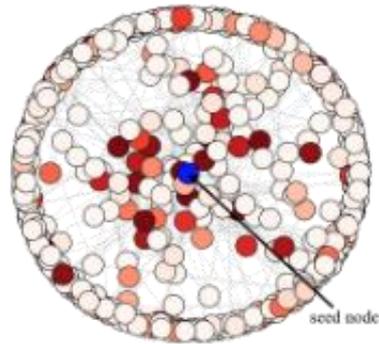


Why Data Mining? – Social Good

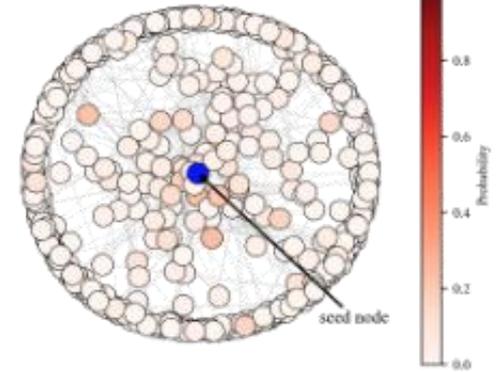


Text: "Breaking: NASA confirms first-ever human colony on Mars will begin next year — tickets for civilians already being sold out in minutes!"

Ours
Influence Spread=2768.06



IC Model
Influence Spread=534.50



Recipients

Subject

Hey Casey,

I noticed you and I are b LinkedIn, and that you ju

Since you're an Austin b technology and sustaine, attending our Q&A M

One of the keynote spe fellow green energy shi lth in San Francisco. Sh

Warm regards,

Jane Doe

Email Generation

CyCLIP: Cyclic Contrastive Language-Image Pretraining

Nishant Gup† UCCLA, nishantg@uccla.edu
Hank Han† UCCLA, hhan@uccla.edu
Saurabh Mehta† MSR LA, saurabh.mehta@microsoft.com
Ayan A. Bhowik† Adobe Research, ayankb@adobe.com
Vishay Viny† Adobe Research, viny@adobe.com
Ashley Coen† UCCLA, ashcoen@uccla.edu

Abstract

Recent advances in contrastive representation learning over paired image-text data have led to notable gains in CLIP (and their variants) across a wide range of performance for zero-shot classification and distributed embeddings. Such models typically require pre-training in the image and text modalities separately, often using separate relevance tasks. Country is a good fit, we demonstrate that the image and text modalities trained in a shared contrastive objective are not interchangeable and can be used to overcome downstream problems. To mitigate this issue, we build a consistent and principled CyCLIP framework for cross-modal representation learning that explicitly optimizes for the shared representation in the generatively consistent in the image and text space. In particular, we show that consistent representations can be further refined by incorporating the consistency between the cross-modal image and text (cross-modal consistency) and the consistency between the image and text (cross-modal consistency). Experiments show that the proposed consistency in CyCLIP outperforms in zero-shot classification and zero-shot generation tasks. We also provide the source code for the proposed framework at <https://github.com/ucclalab/cyclip>.

Abstract Generation

Customer Reviews

★★★★★ These work!

By [Cristian](#) on August 28, 2017
Color: white

When we lost the original adapter, we tried several parts and the other for another lighting part (the "chip" board) that do not work, and just happened back. This means to think any other one will have a good amount of quality. They reach the original appearance for at least a very good price and work like the original. I highly recommend. Thank you very much.

★★★★★ This had they will give you all the space for your activities...

By [Joy](#) on November 8, 2017
Color: white

Will never get old of this. Makes my life easier. I do not have to hold my body anymore, and my back on my feet or neck. Trying to use my old one and changing with one of the newer models. This had they will give you all the space for your activities and finally for anyone you want I might be easy to bring my laptop to the party. A lot of people think you want my phone is changing, someone will be talking, but also in perspective is okay. I like this change more than most things in life. It keeps colorful changes and jumps anything. Thank this change.

★★★★★ Good product

By [KIMBERLY LYNNE TAYLOR](#) on November 18, 2017
Color: white

My rabbit like my four legged friends changing the cable as they he expects me to buy this new... and they are his companion and will really spend a cable changing cable. I wanted one that does the same thing of the original. My rabbit would not seem to be changing only one and I surprised! I do not expect these things because they are so cheap!!!!

Review Generation

🔗 [r/learnmachinelearning](#) · 1 day ago
DIT:Gib:Racc-1593

RAG 2.0: Future of LLMs

Discussion

There have been plenty of articles written about Retrieval Augmented Generation (RAG) pipelines, which as a technology is quite cool. But what's next for the technology of RAG.

What if we can create models with trainable retrievers, or in short, the entire RAG pipeline is customizable like fine-tuning an LLM?

The problem with current RAGs is that they are not fully in tune with it's submodules, it's like a Frankenstein monster, it somehow works, but the parts are not in harmony and perform quite suboptimally together. So, to tackle all the issues with Frankenstein RAG, let's take a deep dive into RAG 2.0.

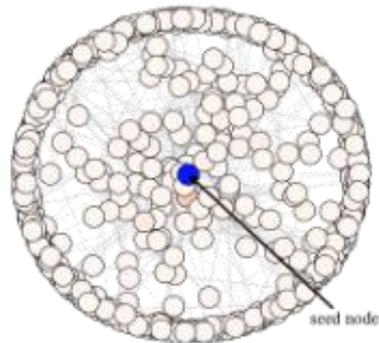
But why does this solve the issues?

Read more →

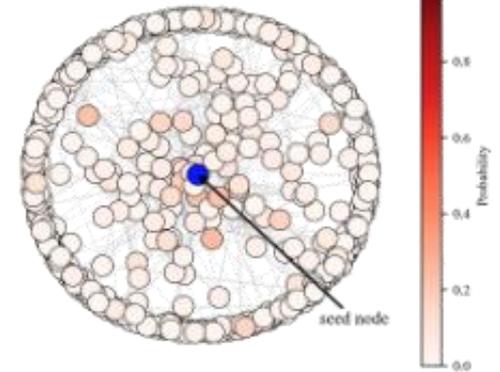
Topic Writing

Text: " Today I bought a new pencil."

Ours
Influence Spread=20.45



IC Model
Influence Spread=534.50



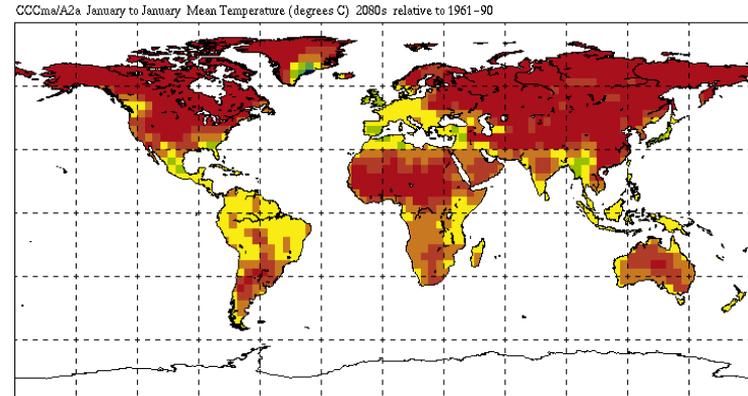


However, we have challenges – Question

What kind of data mining question you want to answer?



Improving health care and reducing costs



Predicting the impact of climate change



Finding alternative/ green energy sources



Reducing hunger and poverty by increasing agriculture production



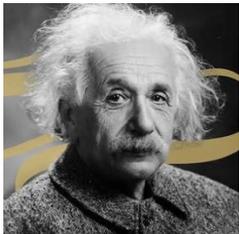
However, we have challenges – Question

What kind of data mining question you want to answer?



Judge a man by his questions rather than his answers.

----- Voltaire



The important thing is not to stop questioning.

----- Albert Einstein



He who asks a question is a fool for five minutes; he who does not ask a question remains a fool forever.

----- Confucius

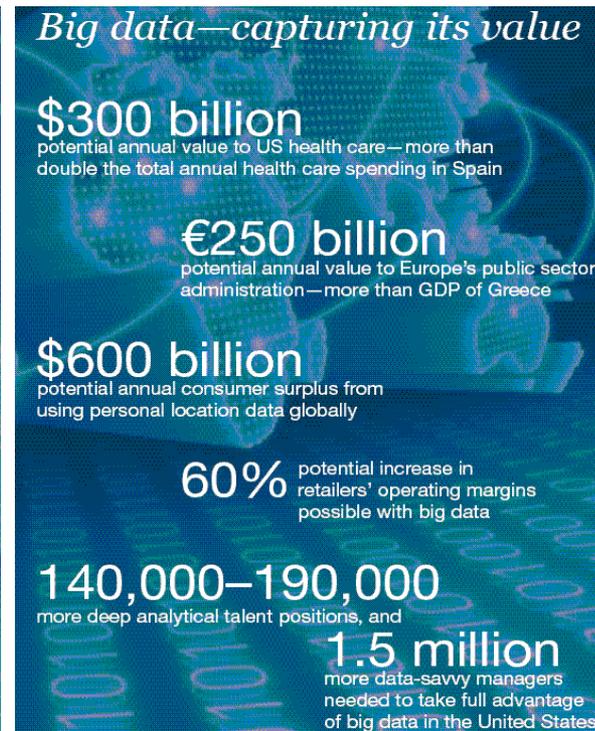
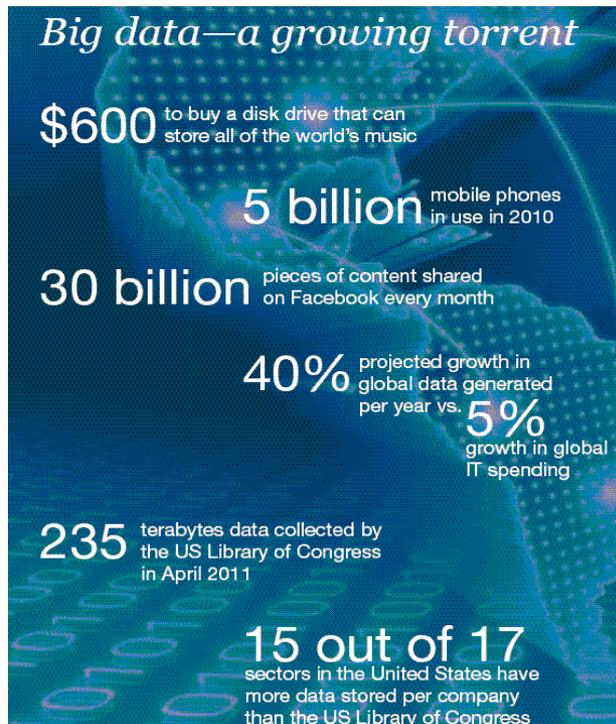


However, we have challenges – Data

Data is usually in a very large scale!

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

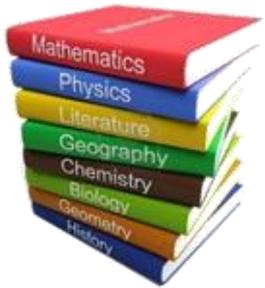




However, we have challenges – Data

Data is usually in a very large scale!

**Textbook
Knowledge Base**



158 million books

[ISBN DB 2023](#)

**Internet
Knowledge Base**



1.1 billion websites

[Musemind 2024](#)

**Neural
Knowledge Base**



405 billion parameters

[Hugging Face 2024](#)



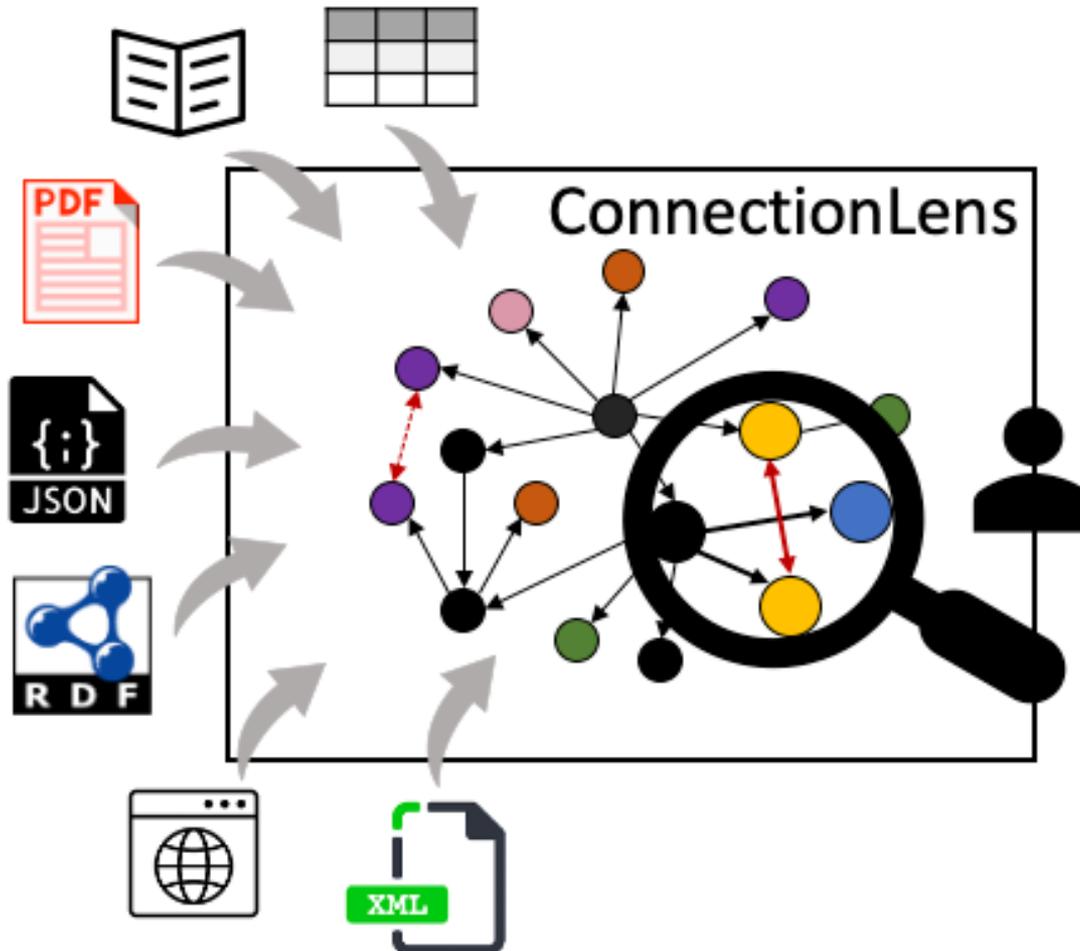
 **2.5 petabytes, 1 billion books**

- **We remember meanings, not details.**
- **We forget on purpose.**
- **Tiny active memory, Larger long-term memory.**



However, we have challenges – Data

Data is diverse and heterogeneous





- **Data is everywhere**
- **Data Mining brings scientific advancement and social wellness**
- **However, there are challenges**
 - (1) What are good questions to ask?
 - (2) Data is scattered around the world, how to find them?
 - (3) Data is very large-scale, how to analyze them efficiently, space/time?
 - (4) Data is very heterogeneous and specialized

This is the reason for taking data mining!

Question Time!





Course Logistics - Time

Times:

- **Classes:** Monday/Wednesday 12:00-1:20 pm PST, Gerlinger 302
- **Office hours:** Wednesday 1:20-2:00 pm PST, other time by appointment
- **Zoom:** <https://uoregon.zoom.us/j/4052006678>



Course Logistics – Quizz

Times:

- **Classes:** Monday/Wednesday 12:00-1:20 pm PST, Gerlinger 302
- **Office hours:** Wednesday 1:20-2:00 pm PST, other time by appointment
- **Zoom:** <https://uoregon.zoom.us/j/4052006678>

Components:

Course Assessment and Grading Scale

Category	CS-453 (%)	CS-553 (%)
Quizz 1	20%	15%
Quizz 2	20%	15%
Project	40%	45%
Participation	5%	5%
Paper Presentation	15%	20%
Overleaf Bonus	5%	5%

- As long as you are **active thinking** and **understand the content**, you will be good

Question Time!

