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/

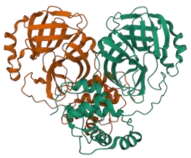
Research Interests:

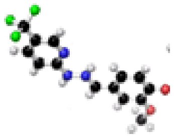
- Data Mining and Machine Learning
- Neural-Symbolic Learning
- Graph and Network
- LLM + Structured Knowledge
- AI/ML/DM Applications
  - Document Intelligence
  - Social Computing
  - Networking Physical Infrastructure

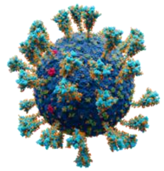
📢📢📢 **Recruiting Ph.D. students and interns!** I am actively seeking highly motivated students for Ph.D. or Research intern positions. Please feel free to email me your CV, transcripts, and brief descriptions about why you want to work with me if you are interested!

What is Data?

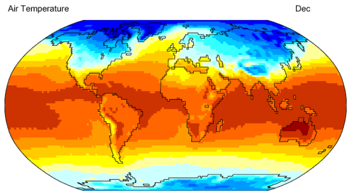
Science


Protein


Small Molecule


Virus


Brain Neural

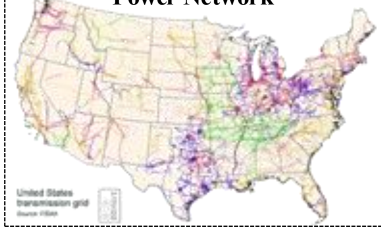


Surface Temperature of Earth

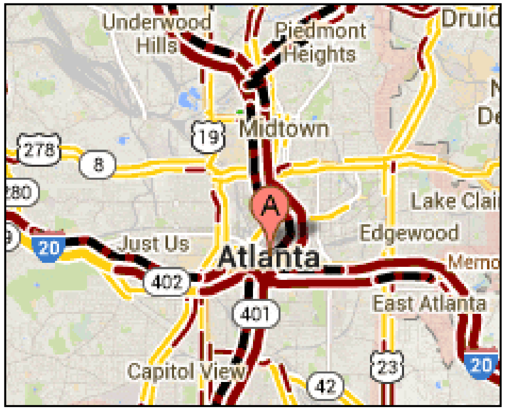
Infrastructure

Gas Network

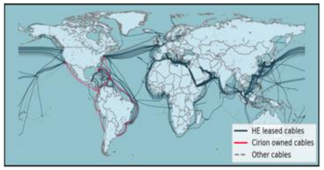
Power Network



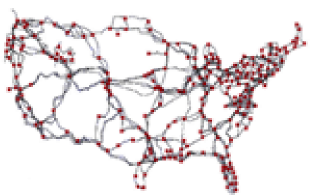
Transportation Network



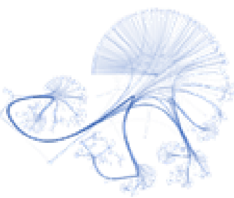
Submarine Cable



Terrestial Cable



Social Network


Citation Network

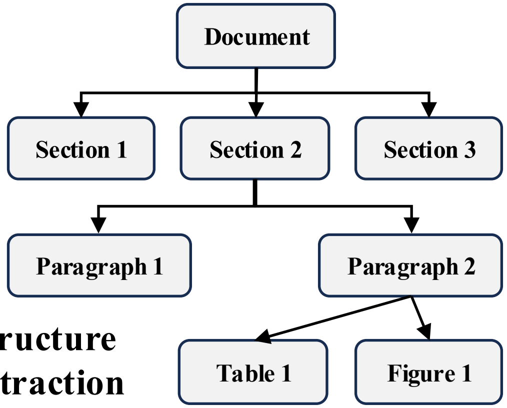

Transaction Network

User-Entity Interaction Graph

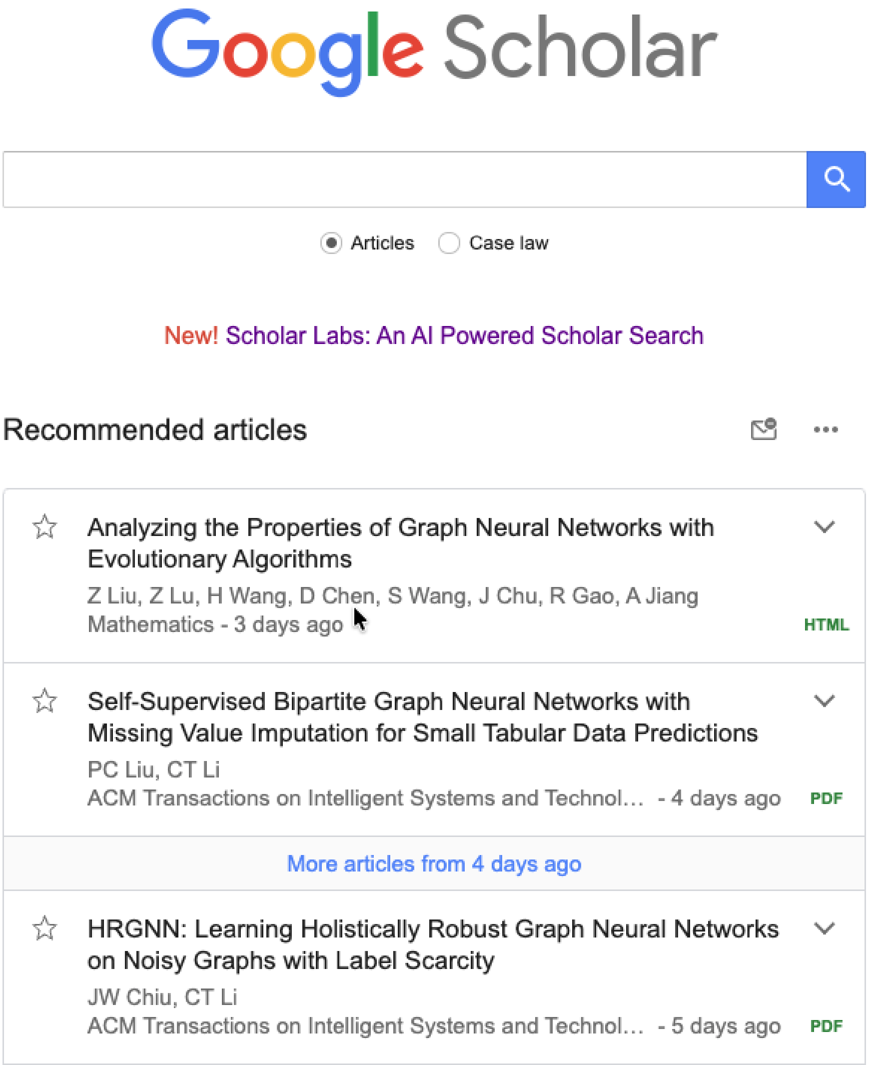

Virtual Village with AI Agents

Document

Structure Extraction



Why Analyze Data? – Paper Management



REFERENCES

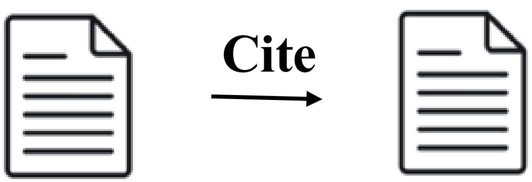
Art of Problem Solving. Aime problems and solutions, 2025. URL `https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions`. 8, 22

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

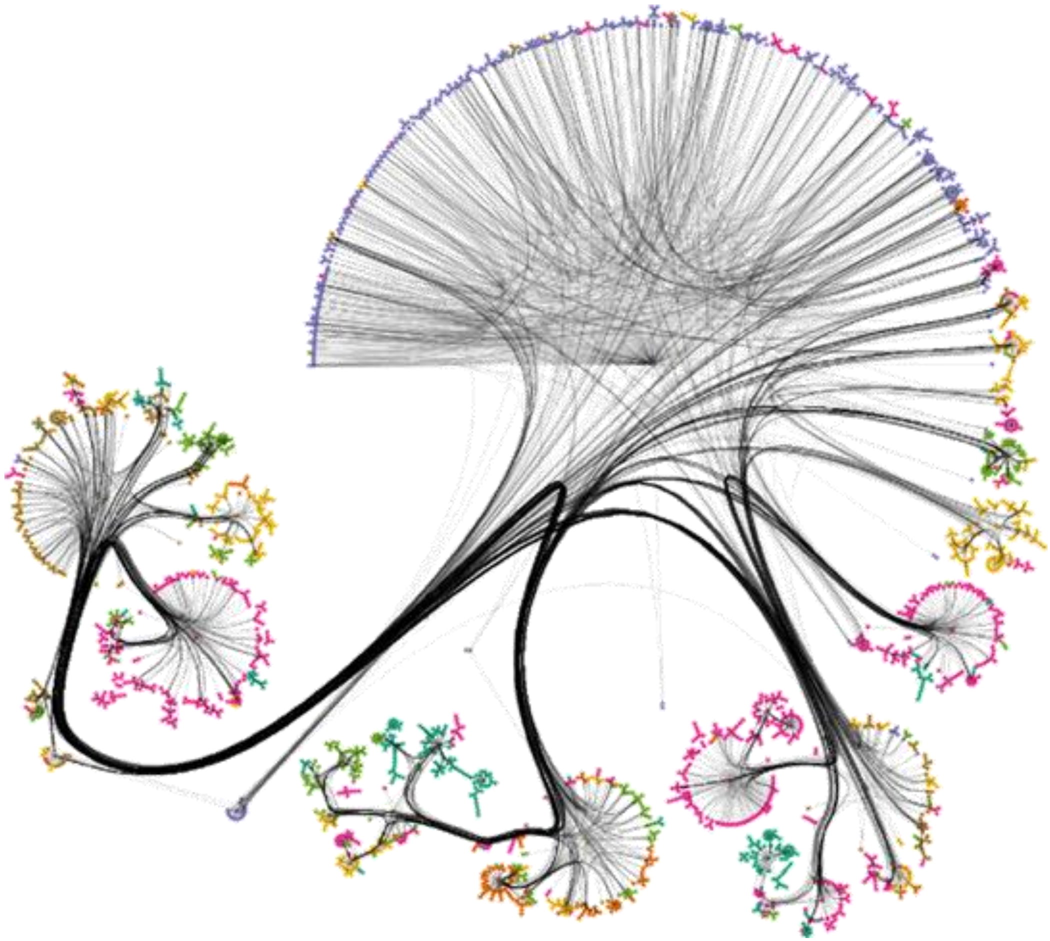
Zihao Cheng, Hongru Wang, Zeming Liu, Yuhang Guo, Yuanfang Guo, Yunhong Wang, and Haifeng Wang. ToolSpectrum: Towards personalized tool utilization for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20679–20699, 2025. 10

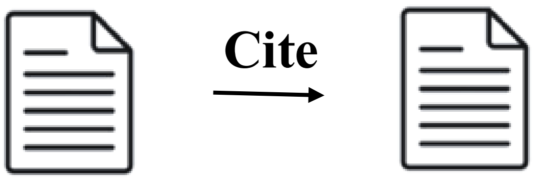
Yingfan Deng, Anhao Zhou, Yuan Yuan, Xian Zhang, Yifei Zou, and Dongxiao Yu. Pe-ma: Parameter-efficient co-evolution of multi-agent systems. *arXiv preprint arXiv:2506.11803*, 2025. 11

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



Why Analyze Data? – Paper Management





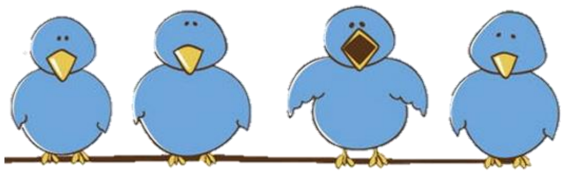
$$\frac{\sum_{e_{ij}\in\mathcal{E}} 1[y_i == y_j]}{|\mathcal{E}|}$$

$\mathcal{E}$ - Total Number of Edges

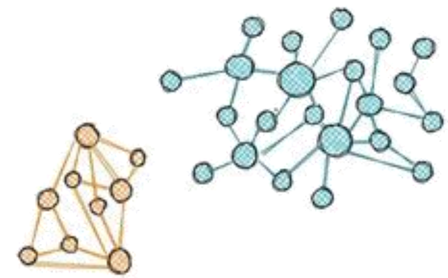
e_{ij} - Edge between node i/j

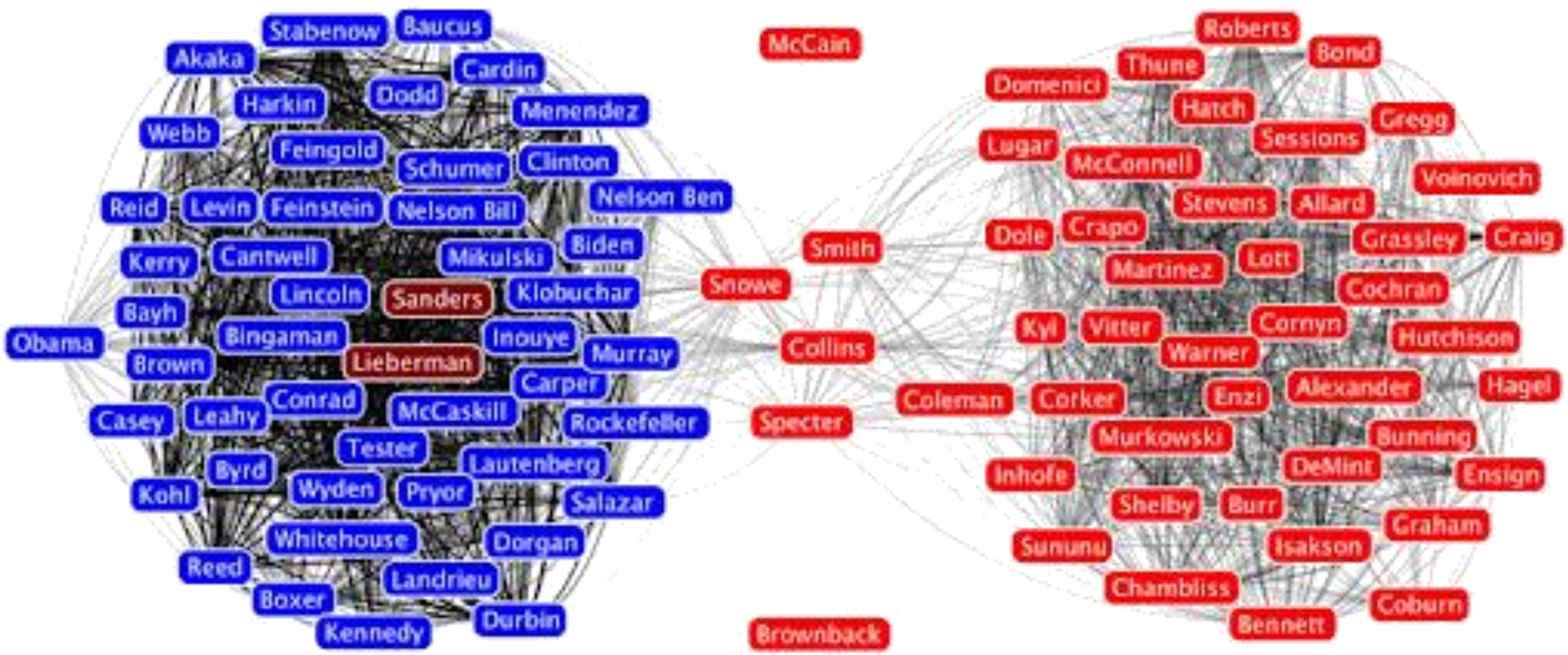
y_i - Label of i

Why Analyze Data? – Paper Management



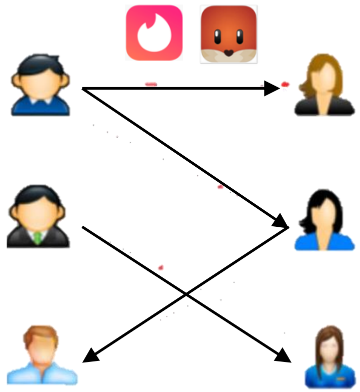
Birds of a feather flock together



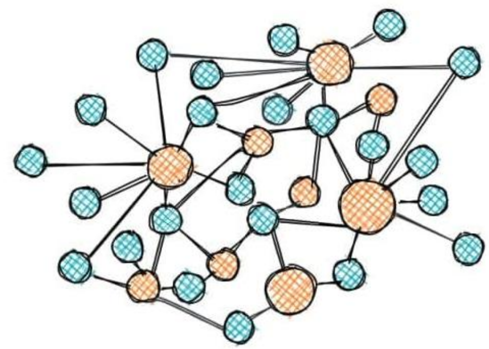


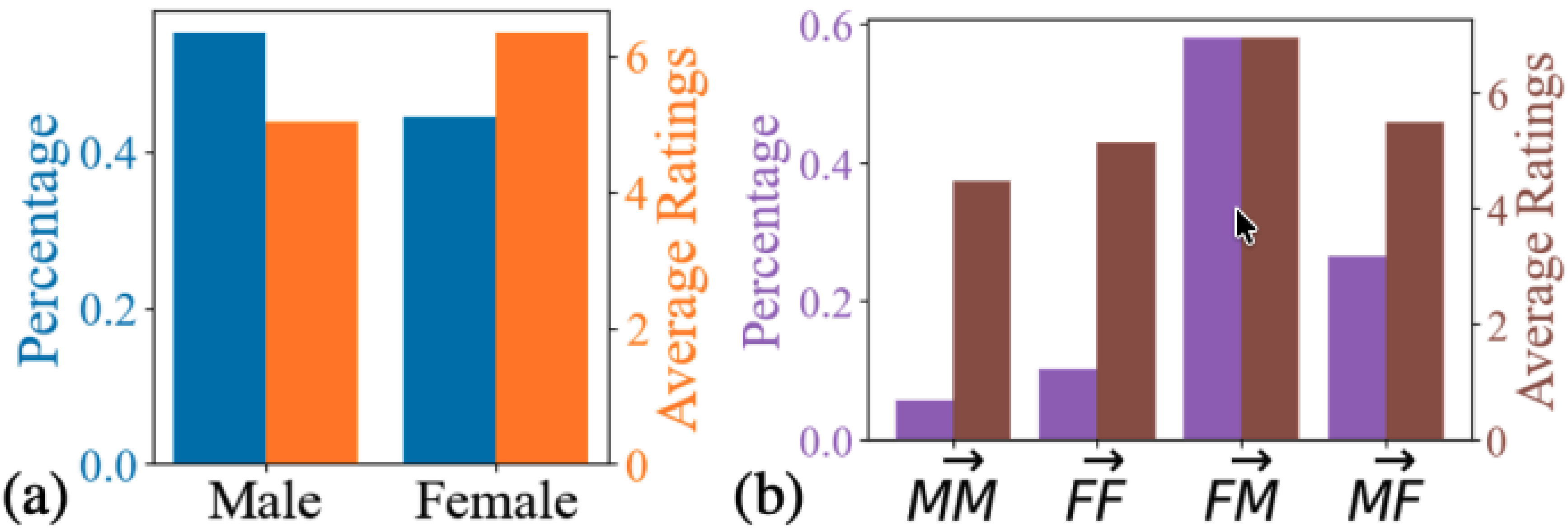
Gun Control Belief Network

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 Han[1], Sheng Liu[1], Jianwen Xie[4], Yu Zhang[2], Yejin Choi[1], James Zou†[1], Pan Lu†[1]
[1]Stanford University, [2]Texas A&M University, [3]UC San Diego, [4]Lambda

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

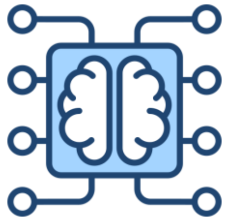
IN-THE-FLOW AGENTIC SYSTEM OPTIMIZATION FOR EFFECTIVE PLANNING AND TOOL USE

Zhuofeng Li*[1,2], Haoxiang Zhang*[1,3], Seungju Han[1], Sheng Liu[1], Jianwen Xie[4], Yu Zhang[2], Yejin Choi[1], James Zou†[1], Pan Lu†[1]
[1]Stanford University, [2]Texas A&M University, [3]UC San Diego, [4]Lambda

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.



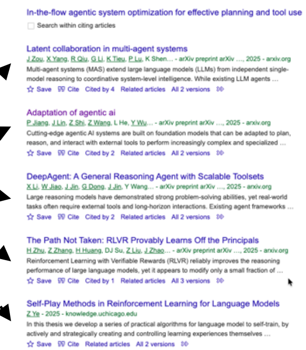
IN-THE-FLOW AGENTIC SYSTEM OPTIMIZATION FOR EFFECTIVE PLANNING AND TOOL USE

Zhuofeng Li*[1,2], Haoxiang Zhang*[1,3], Seungju Han[1], Sheng Liu[1], Jianwen Xie[4], Yu Zhang[2], Yejin Choi[1], James Zou†[1], Pan Lu†[1]
[1]Stanford University, [2]Texas A&M University, [3]UC San Diego, [4]Lambda

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

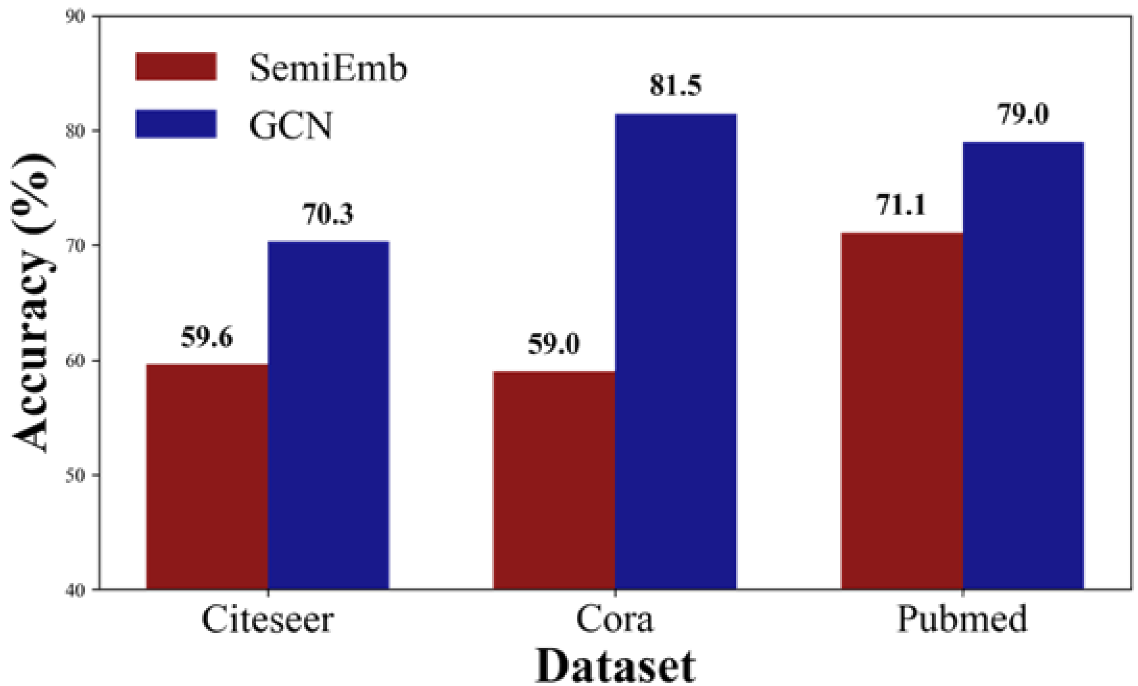
Latent collaboration in multi-agent systems
J Zou, X Yang, R Qiu, G Li, K Tieu, P Lu, K Shen… - arXiv preprint arXiv …, 2025 - arxiv.org
Multi-agent systems (MAS) extend large language models (LLMs) from independent single-model reasoning to coordinative system-level intelligence. While existing LLM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

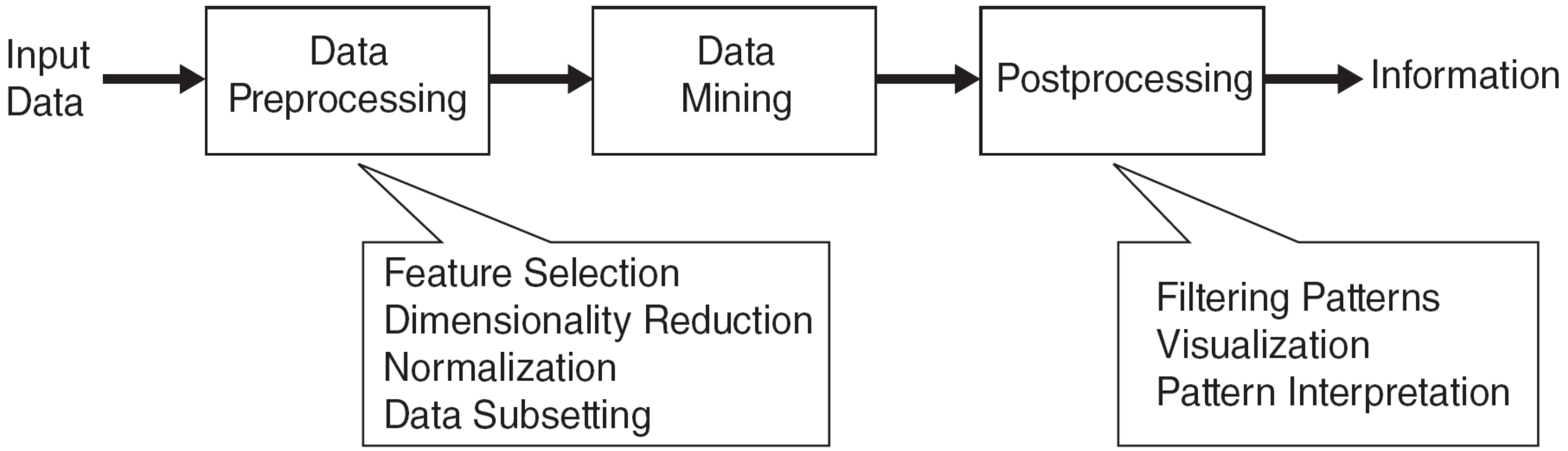
DeepAgent: A General Reasoning Agent with Scalable Toolsets
X Li, W Jiao, J Jin, G Dong, J Jin, Y Wang… - arXiv preprint arXiv …, 2025 - arxiv.org
Large reasoning models have demonstrated strong problem-solving abilities, yet real-world tasks often require external tools and long-horizon interactions. 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



What is Data Mining?



Many Definitions

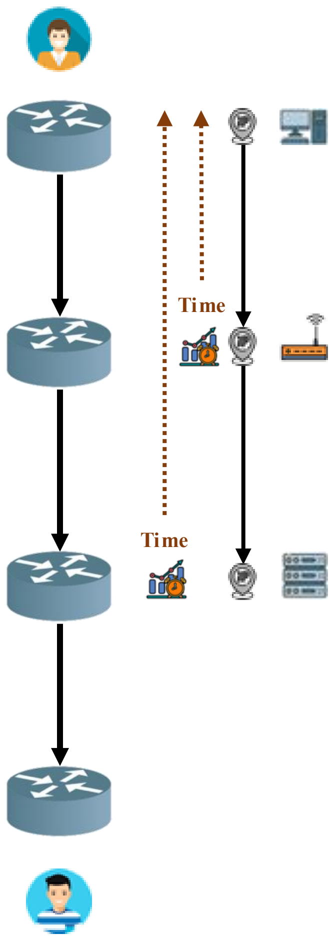
Non-trivial extraction of implicit, previously unknown and potentially useful information from data

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

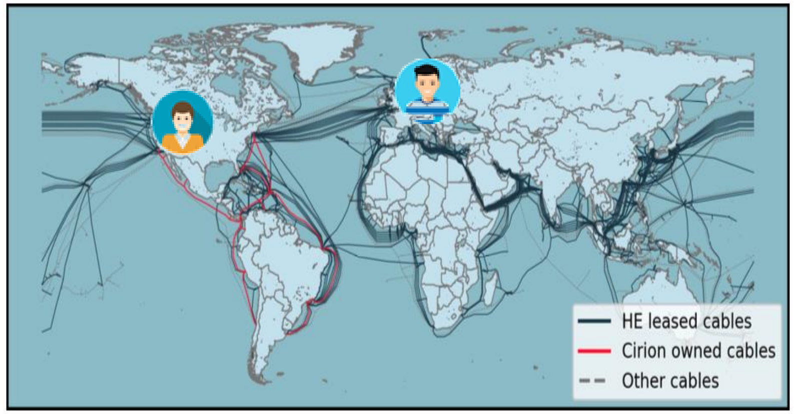
ONRG

Logical Layer

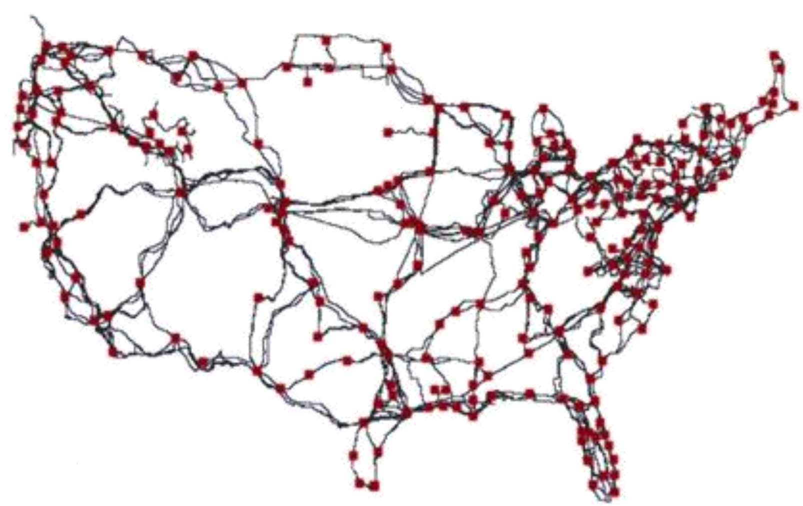


Mapping

Physical Layer



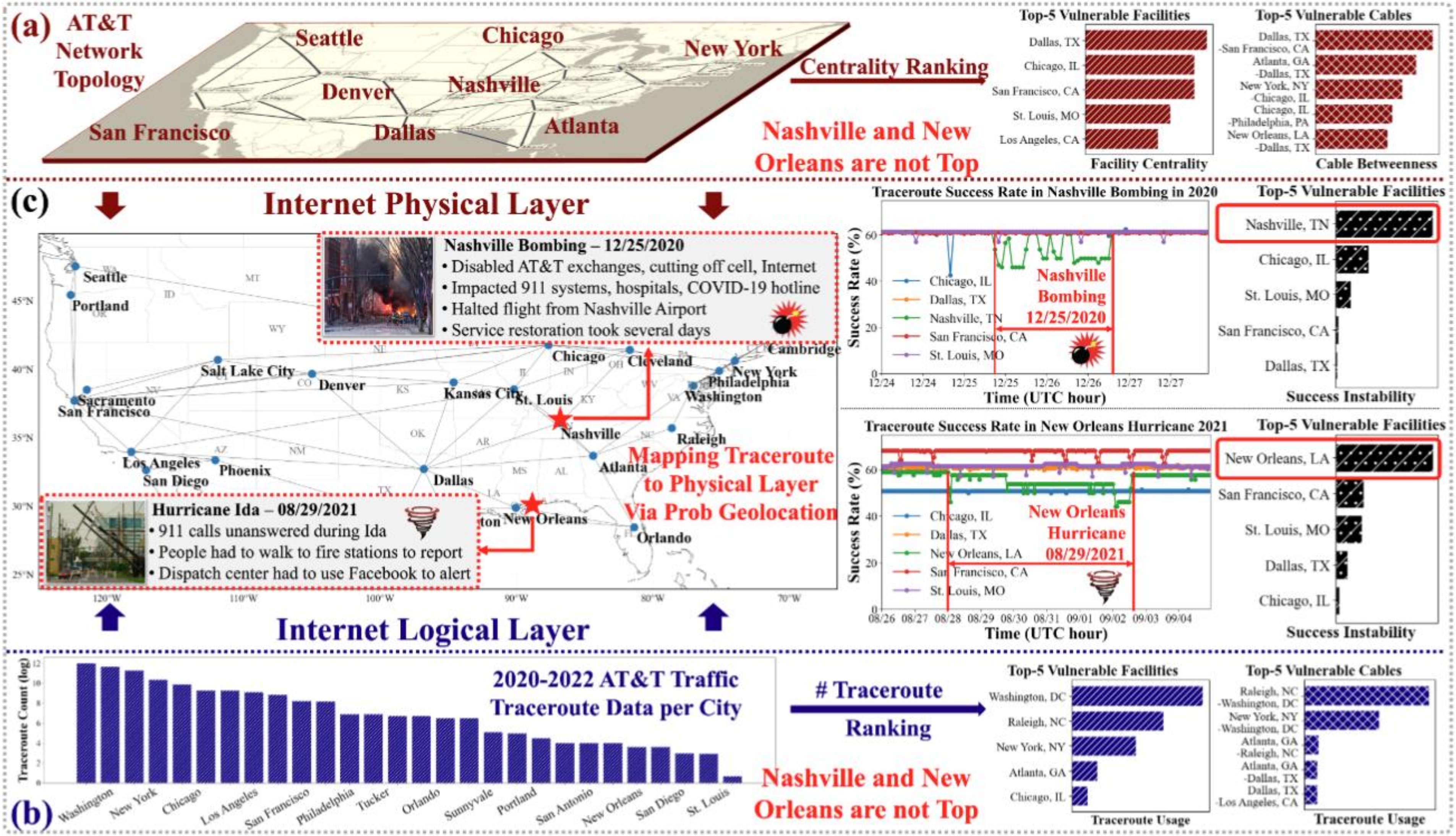
Submarine Cable Map



Terrestrial Cable Map

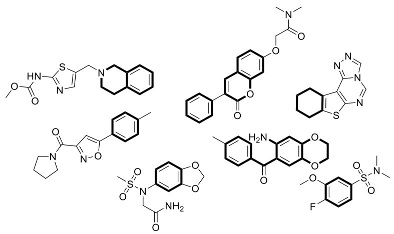
Which physical cable path does this logic signal traverse?

Why Data Mining? – Networking Infra Risk



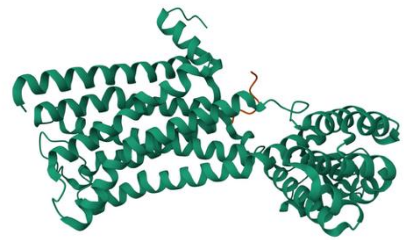
Why Data Mining? – Drug Design

Chemical Libraries

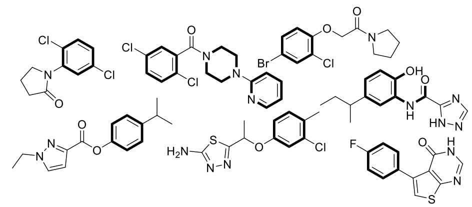


Number of Molecules: 103-106

Protein Target



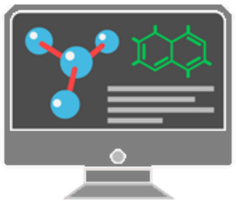
Virtual Libraries



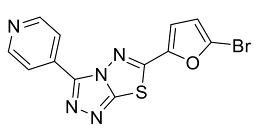
e.g., 10^9 Virtual Molecules on the REAL database in Enamine Ltd.

High Throughput Screening (HTS)

Training



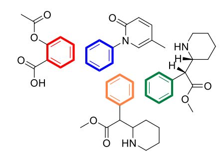
Deep Learning Models



Hit Rate: 0.05%-0.5%

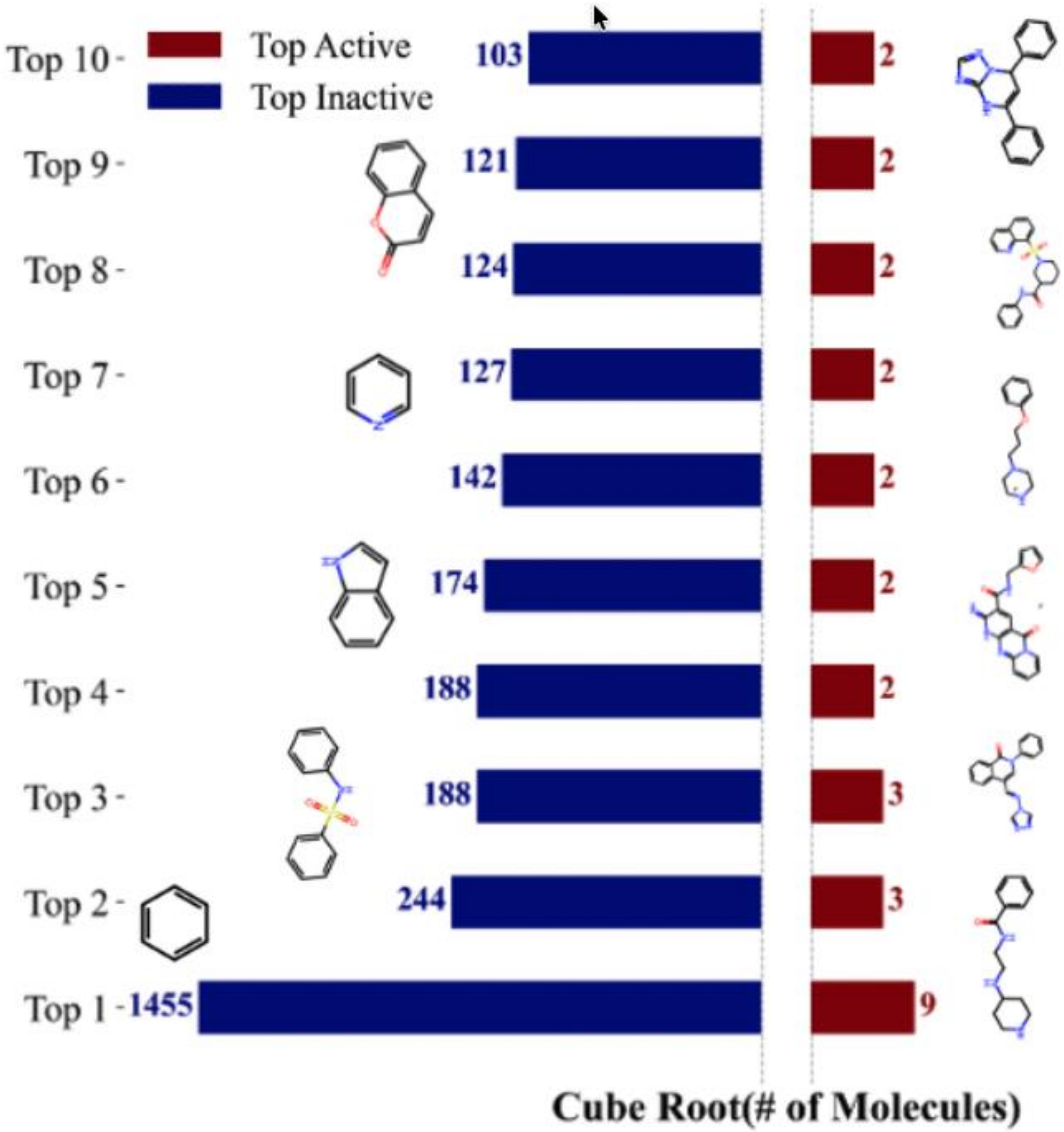
Evaluating

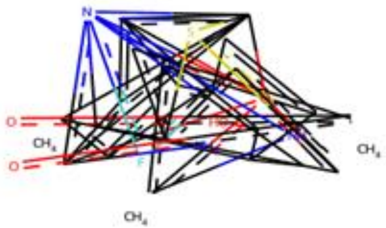
Predicted Actives

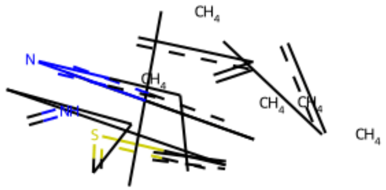


Number of Molecules: 500-1000

Why Data Mining? – Drug Design







Why Data Mining? – Commercial Perspective

- Lots of data is being collected and warehoused
  - Web data
    - 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

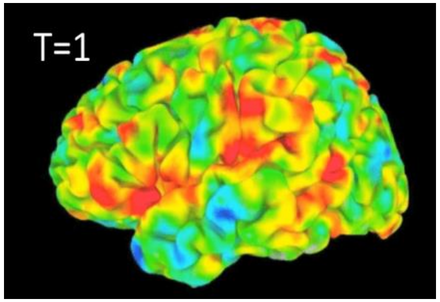
1,000 terabytes,
1,000,000,000,000,000= bytes

- 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

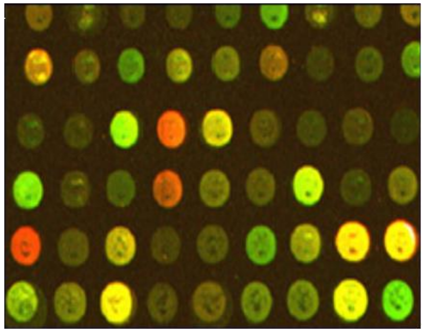
- Data collected and stored at enormous speeds
  - Remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - Telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - Scientific simulations
    - terabytes of data generated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

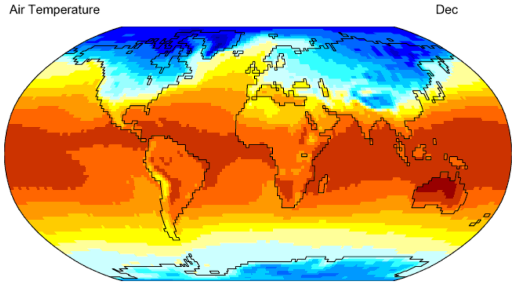
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation



fMRI Data from Brain

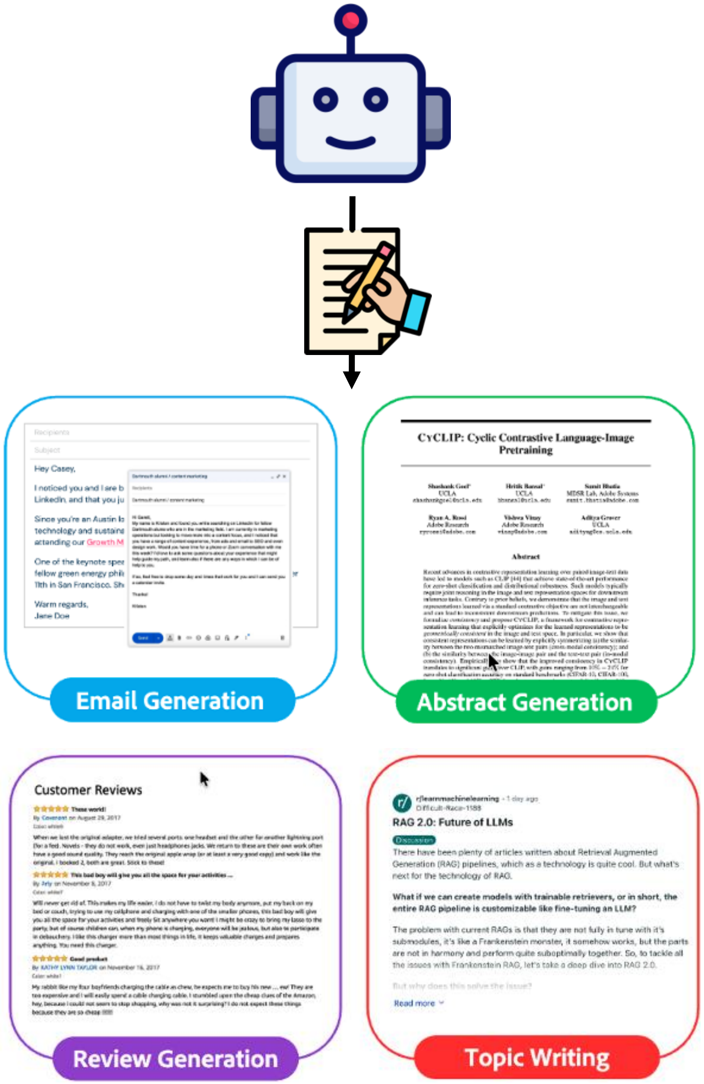

Sky Survey Data

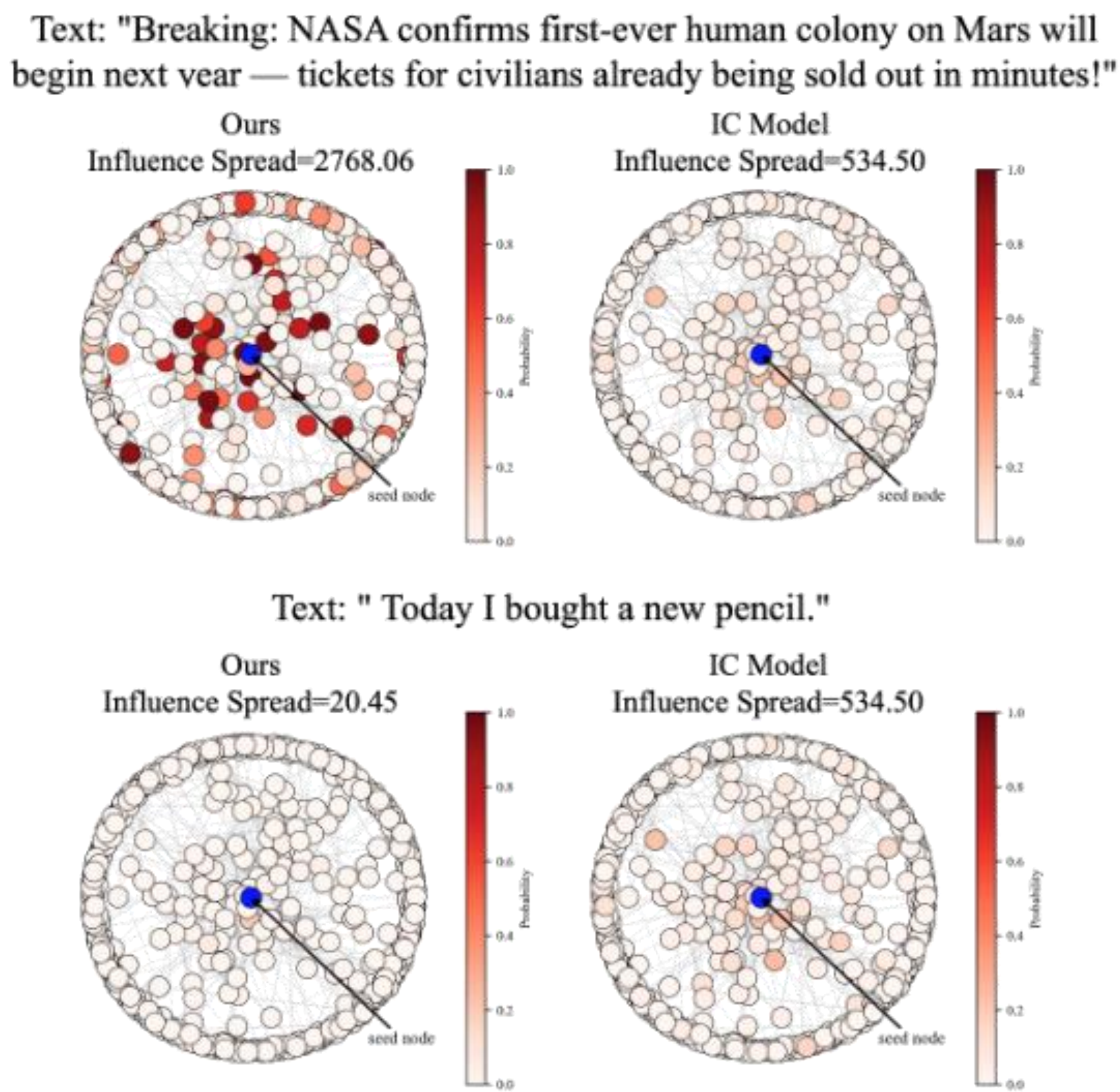

Gene Expression Data



Surface Temperature of Earth

Why Data Mining? – Social Good

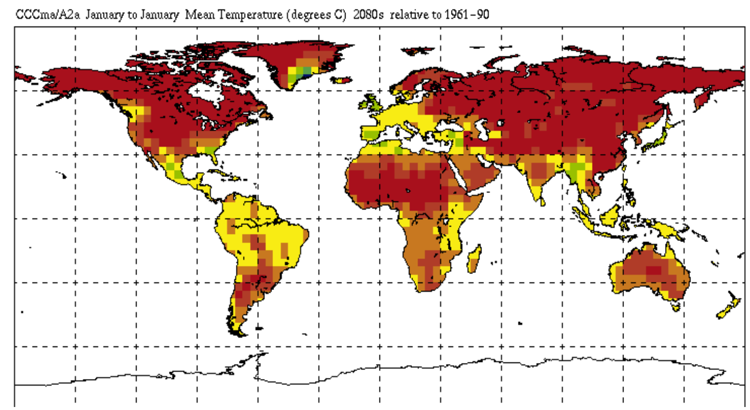




However, we have challenges – Question

What kind of data mining question you want to answer?

Improving health care and reducing costs



Predicting the impact of climate change

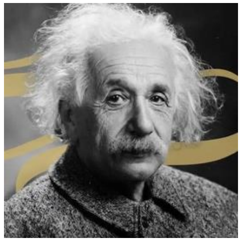
Finding alternative/ green energy sources

Reducing hunger and poverty by
increasing agriculture production

What kind of data mining question you want to answer?

Judge a man by his questions rather than his answers.

------ Voltaire



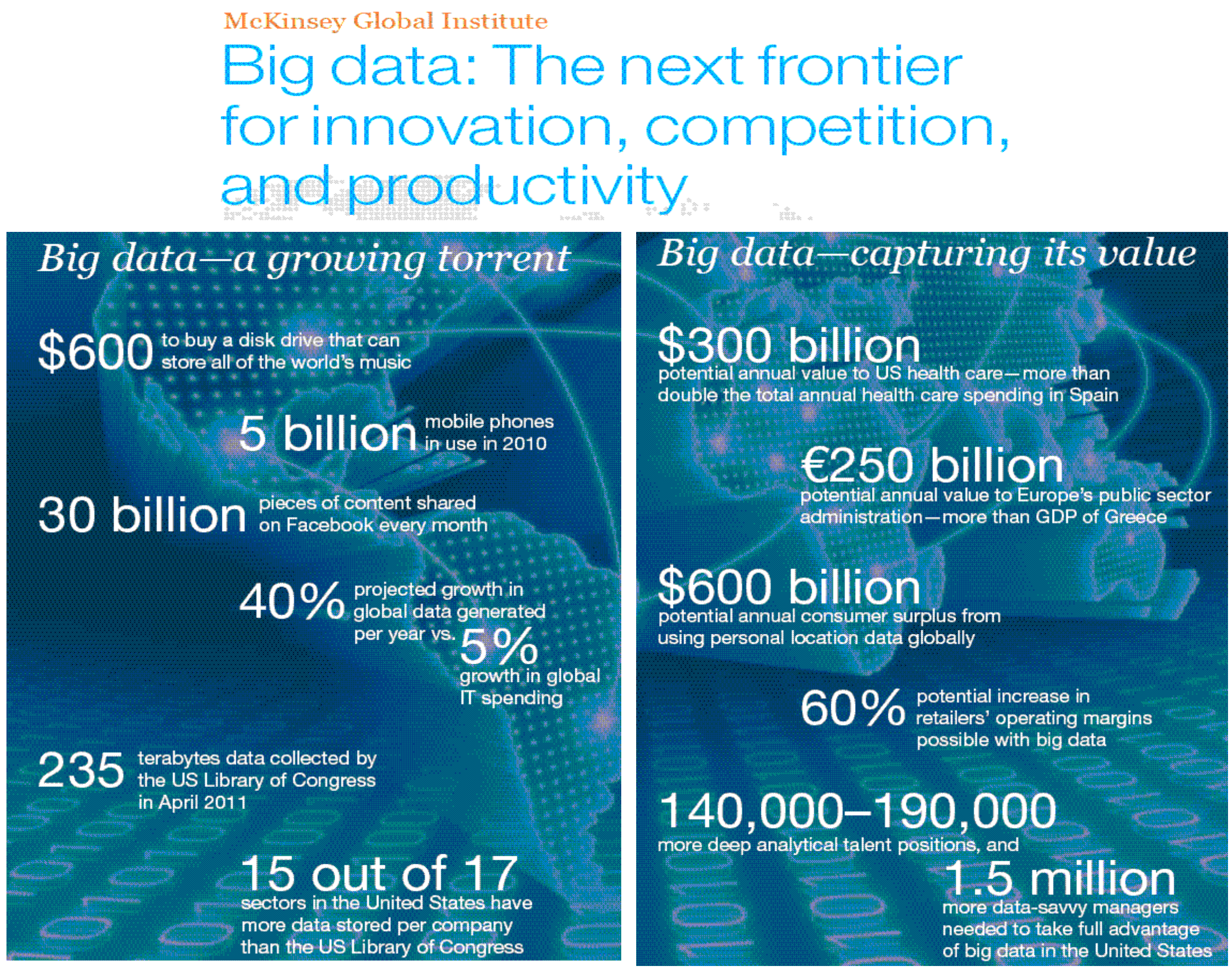
The important thing is not to stop questioning.

------ Albert Einstein

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

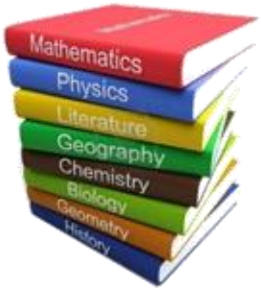
------ Confucius

Data is usually in a very large scale!



However, we have challenges – Data

Data is usually in a very large scale!

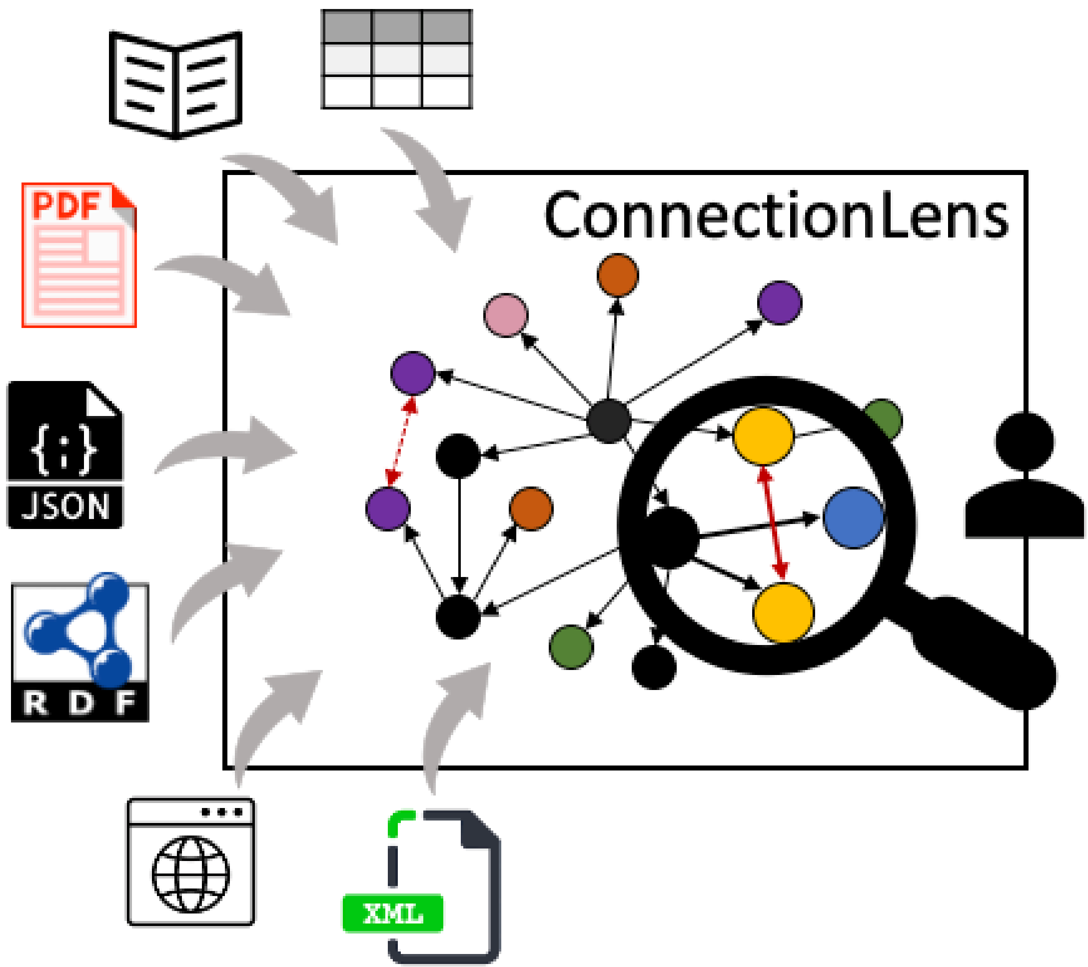
Textbook Knowledge Base		Internet Knowledge Base		Neural Knowledge Base
	→		→	
158 million books		**1.1 billion websites**		**405 billion parameters**
ISBN DB 2023		Musemind 2024		Hugging Face 2024

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

2.5 petabytes, 1 billion books

However, we have challenges – Data

Data is diverse and heterogeneous



Summary

- **Data is everywhere**

- **Data Mining brings scientific advancement and social wellness**

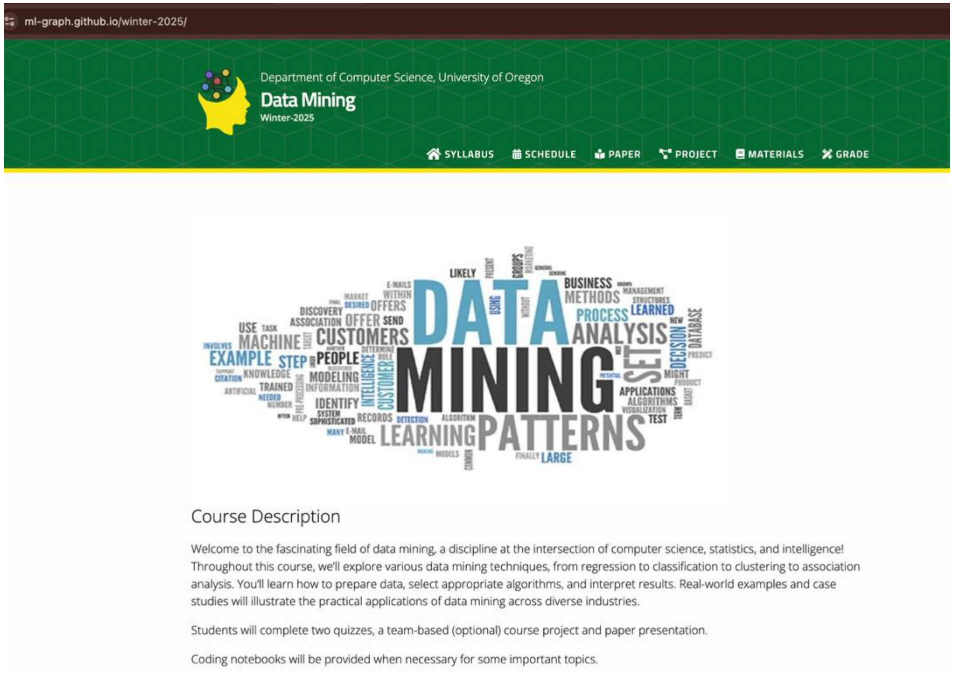
- **However, there are challenges**

  (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



https://ml-graph.github.io/winter-2025/

All information will be available on the website!

Goals

- Broad overview of Data Mining
- Data Mining Skills – Knowledge and Code
- Machine Learning Skills – Knowledge and Code
- Real-world GML/DM applications

Prerequisite

- Linear Algebra, Probability /Statistics, Calculus
- Programming – Python, PyTorch
- Curiosity – Critical Thinking
- Diligence – Hard Working

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 – Quzz

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 (%)
Quzz 1	20%	15%
Quzz 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!

Course Logistics – Quzz

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 (%)
Quzz 1	20%	15%
Quzz 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

Course Logistics – Project

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 (%)
Quzz 1	20%	15%
Quzz 2	20%	15%
Project	40%	45%
Participation	5%	5%
Paper Presentation	15%	20%
Overleaf Bonus	5%	5%

https://ml-graph.github.io/winter-2026/project/

Course Logistics – Project

Project

The project may be completed either individually or as a team; both approaches are acceptable. For team-based projects, only one team member should submit the final report and clearly specify all contributing teammates. Bonus Points will apply if you consider doing projects in the following fields with (*) or any domain beyond the following:

1. Background and Problem Formulation - 10%

- **Background - 5%**:
  - What is the general background of the problem you are working on?
    - I want to develop a better paper categorization system
- **Problem Formulation - 5%**:
  - Under the general topic, what specific problem is your project addressing?
    - I want to develop a machine learning model/algorithm to take input of the paper, output the paper topic (machine learning, computer system, human-computer collaboration, etc.)

2. Data Mining Stage - 35%

- **Data Collection and Store - 15%**:
  - What data are you looking to kick off your project? How do you collect them? What data structure do you use to represent them?
    - I collect Cora/Citeseer/Pubmed Data from somewhere (e.g., a paper, a GitHub repository, Hugging Face, etc.), and I use an adjacency list to store their connection and a matrix to store their node feature
- **Data Mining - 20%**:
  - What kind of data mining problem do you need to do and why?
    - I need to analyze the network homophily/heterophily since leveraging this property might help me develop a better machine learning model for paper classification.
  - How do you do it?
    - I calculate for every edge, the two ending points, whether they are in the same class or not, and quantify the average ratio as a homophily ratio
  - What kind of pattern do you find? How do you present your findings/analysis?
    - I find that in many paper citation networks, the homophily is pretty high. Using Number/Table/Figure, etc.

3. Machine Learning Stage - 35%

- **Machine Learning Model Design**:
  - Based on your targeted problem, what kind of machine learning model do you want to build and why?
    - I want to build a graph neural network to fully exploit the discovered homophily principle.

https://ml-graph.github.io/winter-2026/project/

Course Logistics – Paper Presentation

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 (%)
Quzz 1	20%	15%
Quzz 2	20%	15%
Project	40%	45%
Participation	5%	5%
Paper Presentation	15%	20%
Overleaf Bonus	5%	5%

https://ml-graph.github.io/winter-2026/presentation/

Course Logistics – Paper Presentation

Presentation

Paper Presentation Details

You can either collaborate with a team or present individually. The choice of topic is entirely up to you.

- Introduction and Background – What is the general impact and background of the topic?
- Motivation and Problem – What is the core research problem, and why do we study it?
- Related Work and Challenges – How did previous works address this problem, and what are some of the challenges?
- Proposed Solutions/Methods and Rationale – What are the proposed methods/techniques, and why are they proposed? What specific reasons would solving this problem require these proposed(1) methods/techniques?
- Experimental Setting, Results, and Analysis – What experiments are designed to verify the proposed method? How are results being discussed and analyzed? Are there any interesting findings?
- Conclusion and Future Work

Do not use sentences in the slides, but use bullet points and important points that you can logically chain together for your speech I will be very careful taking note of this. Please pardon me for this!

Natural Disaster Modeling

Neural-Biology Analysis

Social Network

Agentic AI

Reasoning/Planning

https://ml-graph.github.io/winter-2026/presentation/

Course Logistics – Paper Presentation – Bad Example

The provided image outlines the logistical and academic requirements for a course at the University of Oregon, likely **CS-453/553**. Classes are held on **Mondays and Wednesdays from 12:00 pm to 1:20 pm PST** in Gerlinger 302, with office hours scheduled for Wednesdays from 1:20 pm to 2:00 pm or by appointment. A specific Zoom link is also provided for virtual access.

The grading structure, labeled "Course Assessment and Grading Scale," distinguishes between undergraduate (**CS-453**) and graduate (**CS-553**) requirements. For undergraduate students, the grade is heavily weighted toward two quizzes at **20% each** (40% total) and a project worth **40%**, followed by a paper presentation at **15%** and participation at **5%**. Graduate students have a slightly different distribution, with quizzes weighted less at **15% each** (30% total), while the project and paper presentation are weighted higher at **45%** and **20%** respectively.

Both groups have the opportunity for a **5% Overleaf Bonus**. Beside the grading chart, a motivational note emphasizes that students will succeed as long as they maintain **active thinking** and **understand the content**.

Course Logistics – Paper Presentation – Good Example

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 (%)
Quzz 1	20%	15%
Quzz 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

Course Logistics – Timeline

Basics

EVENT	DATE	DESCRIPTION	COURSE MATERIAL
Lecture	01/05/2026 Monday	Overview Syllabus	Course Materials: Slides
Assignment	01/05/2026 Monday	**Project released!**	[Project]
Lecture	01/07/2026 Wednesday	Logistics Basics	Course Materials: Slides
Lecture	01/12/2026 Monday	Classification KNN/Naive Bayes	Course Materials: Slides
Lecture	01/14/2026 Wednesday	Classification Decision Tree	Course Materials: Slides
Martin Luther King, Jr holiday	01/19/2026 04:30 Monday	Enjoy :)	
Lecture	01/21/2026 Wednesday	Clustering K-means, Hierarchical Clustering	Course Materials: Slides
Lecture	01/26/2026 Monday	Dimension Reduction PCA	Course Materials: Slides
Lecture	01/28/2026 Wednesday	Linear Regression Gradient Descent	Course Materials: Slides
Lecture	02/02/2026 Monday	Logistic Classification	Course Materials: Slides
Lecture	02/04/2026 Wednesday	Neural Network	Course Materials: Slides
Exam	**02/09/2026 16:00 Monday**	**Quzz 1**	Topics: Lecture 1 - Lecture 8; Closed Book

Advanced

Lecture	02/11/2026 Wednesday	Graph Mining	Course Materials: Slides
	02/11/2026 16:00 Wednesday	Presentation 1	Group: Group 1: 7-7:15 pm; Group 2: 7:15-7:30 pm; Zoom
Lecture	02/16/2026 Monday	Graph Mining	Course Materials: Slides
	02/16/2026 16:00 Monday	Presentation 2	Group: Group 3: 7-7:15 pm; Group 4: 7:15-7:30 pm; Zoom
Lecture	02/18/2026 Wednesday	Temporal Mining	Course Materials: Slides
	02/18/2026 16:00 Wednesday	Presentation 3	Group: Group 5: 7-7:15 pm; Group 6: 7:15-7:30 pm; Zoom
Lecture	02/23/2026 Monday	Spatial Cloud Point Mining	Course Materials: Video Record; Slides; Video
	02/23/2026 16:00 Monday	Presentation 4	Group: Group 7: 7-7:15 pm; Group 8: 7:15-7:30 pm; Zoom
Lecture	02/25/2026 Wednesday	Image Mining	Course Materials: Video Record; Slides; Video
	02/25/2026 16:00 Wednesday	Presentation 5	Group: Group 9: 7-7:15 pm; Group 10: 7:15-7:30 pm; Zoom
Lecture	03/02/2026 Monday	Language Mining	Course Materials: Slides
	03/02/2026 16:00 Monday	Presentation 6	Group: Group 11: 7-7:15 pm; Group 12: 7:15-7:30 pm; Zoom
Lecture	03/04/2026 Wednesday	Language Mining	Course Materials: Slides
	03/04/2026 16:00 Wednesday	Presentation 6	Group: Group 13: 7-7:15 pm; Group 14: 7:15-7:30 pm; Zoom
Lecture	03/09/2026 Monday	Review Future	Course Materials: Slides
	03/09/2026 16:00 Monday	Presentation 6	Group: Group 15: 7-7:15 pm; Group 16: 7:15-7:30 pm; Zoom
Exam	**03/11/2026 16:00 Wednesday**	**Quzz 2**	Topics: Lecture 9 - Lecture 16; Closed Book
Due	**03/20/2026 23:59 Friday**	**Project Report Due**	

Phase 1 + Quzz 1

Phase 2 + Quzz 2 + Project Report

Course Logistics – Timeline

Basics

EVENT	DATE	DESCRIPTION	COURSE MATERIAL
Lecture	01/05/2026 Monday	Overview Syllabus	Course Materials: Slides
Assignment	01/05/2026 Monday	**Project released!**	[Project]
Lecture	01/07/2026 Wednesday	Logistics Basics	Course Materials: Slides
Lecture	01/12/2026 Monday	Classification KNN/Naive Bayes	Course Materials: Slides
Lecture	01/14/2026 Wednesday	Classification Decision Tree	Course Materials: Slides
Martin Luther King, Jr holiday	01/19/2026 04:30 Monday	Enjoy :)	
Lecture	01/21/2026 Wednesday	Clustering K-means, Hierarchical Clustering	Course Materials: Slides
Lecture	01/26/2026 Monday	Dimension Reduction PCA	Course Materials: Slides
Lecture	01/28/2026 Wednesday	Linear Regression Gradient Descent	Course Materials: Slides
Lecture	02/02/2026 Monday	Logistic Classification	Course Materials: Slides
Lecture	02/04/2026 Wednesday	Neural Network	Course Materials: Slides
Exam	**02/09/2026 16:00 Monday**	**Quzz 1**	Topics: Lecture 1 - Lecture 8; Closed Book

Advanced

Lecture	02/11/2026 Wednesday	Graph Mining	Course Materials: Slides
	02/11/2026 16:00 Wednesday	Presentation 1	Group: Group 1: 7-7:15 pm; Group 2: 7:15-7:30 pm; Zoom
Lecture	02/16/2026 Monday	Graph Mining	Course Materials: Slides
	02/16/2026 16:00 Monday	Presentation 2	Group: Group 3: 7-7:15 pm; Group 4: 7:15-7:30 pm; Zoom
Lecture	02/18/2026 Wednesday	Temporal Mining	Course Materials: Slides
	02/18/2026 16:00 Wednesday	Presentation 3	Group: Group 5: 7-7:15 pm; Group 6: 7:15-7:30 pm; Zoom
Lecture	02/23/2026 Monday	Spatial Cloud Point Mining	Course Materials: Video Record; Slides; Video
	02/23/2026 16:00 Monday	Presentation 4	Group: Group 7: 7-7:15 pm; Group 8: 7:15-7:30 pm; Zoom
Lecture	02/25/2026 Wednesday	Image Mining	Course Materials: Video Record; Slides; Video
	02/25/2026 16:00 Wednesday	Presentation 5	Group: Group 9: 7-7:15 pm; Group 10: 7:15-7:30 pm; Zoom
Lecture	03/02/2026 Monday	Language Mining	Course Materials: Slides
	03/02/2026 16:00 Monday	Presentation 6	Group: Group 11: 7-7:15 pm; Group 12: 7:15-7:30 pm; Zoom
Lecture	03/04/2026 Wednesday	Language Mining	Course Materials: Slides
	03/04/2026 16:00 Wednesday	Presentation 6	Group: Group 13: 7-7:15 pm; Group 14: 7:15-7:30 pm; Zoom
Lecture	03/09/2026 Monday	Review Future	Course Materials: Slides
	03/09/2026 16:00 Monday	Presentation 6	Group: Group 15: 7-7:15 pm; Group 16: 7:15-7:30 pm; Zoom
Exam	**03/11/2026 16:00 Wednesday**	**Quzz 2**	Topics: Lecture 9 - Lecture 16; Closed Book
Due	**03/20/2026 23:59 Friday**	**Project Report Due**	

Paper Presentation

Phase 1 + Quzz 1

Phase 2 + Quzz 2 + Project Report

Course Logistics – Timeline

Basics

EVENT	DATE	DESCRIPTION	COURSE MATERIAL
Lecture	01/05/2026 Monday	Overview Syllabus	Course Materials: Slides
Assignment	01/05/2026 Monday	**Project released!**	[Project]
Lecture	01/07/2026 Wednesday	Logistics Basics	Course Materials: Slides
Lecture	01/12/2026 Monday	Classification KNN/Naive Bayes	Course Materials: Slides
Lecture	01/14/2026 Wednesday	Classification Decision Tree	Course Materials: Slides
Martin Luther King, Jr holiday	01/19/2026 04:30 Monday	Enjoy :)	
Lecture	01/21/2026 Wednesday	Clustering K-means, Hierarchical Clustering	Course Materials: Slides
Lecture	01/26/2026 Monday	Dimension Reduction PCA	Course Materials: Slides
Lecture	01/28/2026 Wednesday	Linear Regression Gradient Descent	Course Materials: Slides
Lecture	02/02/2026 Monday	Logistic Classification	Course Materials: Slides
Lecture	02/04/2026 Wednesday	Neural Network	Course Materials: Slides
Exam	**02/09/2026 16:00 Monday**	**Quzz 1**	Topics: Lecture 1 - Lecture 8; Closed Book

Advanced

Out of Town, Video Record

Lecture	02/11/2026 Wednesday	Graph Mining	Course Materials: Slides
	02/11/2026 16:00 Wednesday	Presentation 1	Group: Group 1: 7-7:15 pm; Group 2: 7:15-7:30 pm; Zoom
Lecture	02/16/2026 Monday	Graph Mining	Course Materials: Slides
	02/16/2026 16:00 Monday	Presentation 2	Group: Group 3: 7-7:15 pm; Group 4: 7:15-7:30 pm; Zoom
Lecture	02/18/2026 Wednesday	Temporal Mining	Course Materials: Slides
	02/1[illegible] 16:00 Wednesday	Presentation 3	Group: Group 5: 7-7:15 pm; Group 6: 7:15-7:30 pm; Zoom
Lecture	02/23/2026 Monday	Spatial Cloud Point Mining	Course Materials: Video Record; Slides; Video
	02/23/2026 16:00 Monday	Presentation 4	Group: Group 7: 7-7:15 pm; Group 8: 7:15-7:30 pm; Zoom
Lecture	02/25/2026 Wednesday	Image Mining	Course Materials: Video Record; Slides; Video

	02/25/2026 16:00 Wednesday	Presentation 5	Group: Group 9: 7-7:15 pm; Group 10: 7:15-7:30 pm; Zoom
Lecture	03/02/2026 Monday	Language Mining	Course Materials: Slides
	03/02/2026 16:00 Monday	Presentation 6	Group: Group 11: 7-7:15 pm; Group 12: 7:15-7:30 pm; Zoom
Lecture	03/04/2026 Wednesday	Language Mining	Course Materials: Slides
	03/04/2026 16:00 Wednesday	Presentation 6	Group: Group 13: 7-7:15 pm; Group 14: 7:15-7:30 pm; Zoom
Lecture	03/09/2026 Monday	Review Future	Course Materials: Slides
	03/09/2026 16:00 Monday	Presentation 6	Group: Group 15: 7-7:15 pm; Group 16: 7:15-7:30 pm; Zoom
Exam	**03/11/2026 16:00 Wednesday**	**Quzz 2**	Topics: Lecture 9 - Lecture 16; Closed Book
Due	**03/20/2026 23:59 Friday**	**Project Report Due**	

Phase 1 + Quzz 1

Phase 2 + Quzz 2 + Project Report

Question Time!

- **Linear Algebra**

- **Calculus**

- **Statistics/Probability**

Basics – Linear Algebra – Scalar/Vector/Matrix

Scalar

$$\mathbf{v} = 3$$

Vector

$$\mathbf{v} = \begin{bmatrix} 1 & 2 & 5 \end{bmatrix}$$

$$\boxed{\mathbf{u} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}}$$

Please note that we will use this one by default

Matrix

$$\mathbf{A} = \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix}}_{\text{3 columns}} \left.\vphantom{\begin{bmatrix} 1 \\ 0 \\ 2 \\ 3 \end{bmatrix}}\right\} \text{4 rows}$$

$$\mathbf{v} \in \mathbb{R}^{\mathbf{1 \times 3}}$$

$$\mathbf{u} \in \mathbb{R}^{\mathbf{3 \times 1}}$$

$$\mathbf{A} \in \mathbb{R}^{\mathbf{4 \times 3}}$$

Basics – Linear Algebra – Scalar/Vector/Matrix

Scalar

$$\mathbf{v} = 3$$

Vector

$$\mathbf{v} = [1 \quad 2 \quad 5]$$

Matrix

$$\mathbf{A} = \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix}}_{\text{3 columns}} \left.\vphantom{\begin{bmatrix} 1 \\ 0 \\ 2 \\ 3 \end{bmatrix}}\right\} \text{4 rows}$$

Scalar Operation

```python
a = 1
b = 2
print('a:', a)
print('b:', b)
print('a+b:', a+b)
print('a*b:', a*b)
print('a/b:', a/b)
print('a-b:', a-b)
```

✓ 0.0s

```
a: 1
b: 2
a+b: 3
a*b: 2
a/b: 0.5
a-b: -1
```

Basics – Linear Algebra – Scalar/Vector/Matrix

Scalar

$$\mathbf{v} = 3$$

Vector

$$\mathbf{v} = \begin{bmatrix} 1 & 2 & 5 \end{bmatrix}$$

Matrix

$$\mathbf{A} = \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix}}_{\text{3 columns}} \left.\vphantom{\begin{bmatrix} 1 \\ 0 \\ 2 \\ 3 \end{bmatrix}}\right\} \text{4 rows}$$

Scalar Operation with Vector

```python
def scalar_vector_ops_list(s: float, v: List[float]) -> None:
    add = [s + x for x in v]         # scalar + vector (element-wise)
    sub = [x - s for x in v]         # vector - scalar
    mul = [s * x for x in v]         # scalar * vector
    div = [x / s for x in v]         # vector / scalar

    print("== Pure Python lists ==")
    print(f"scalar s = {s}")
    print(f"vector v = {v}")
    print(f"s + v   = {add}")
    print(f"v - s   = {sub}")
    print(f"s * v   = {mul}")
    print(f"v / s   = {div}")
    print()
```

```python
if __name__ == "__main__":
    s = 2.0
    v_list = [1.0, -2.0, 3.5, 0.0]

    scalar_vector_ops_list(s, v_list)
    scalar_vector_ops_numpy(s)
```

✓ 0.0s

```
== Pure Python lists ==
scalar s = 2.0
vector v = [1.0, -2.0, 3.5, 0.0]
s + v   = [3.0, 0.0, 5.5, 2.0]
v - s   = [-1.0, -4.0, 1.5, -2.0]
s * v   = [2.0, -4.0, 7.0, 0.0]
v / s   = [0.5, -1.0, 1.75, 0.0]
```

Basics – Linear Algebra – Scalar/Vector/Matrix

Scalar

$$\mathbf{v} = 3$$

Vector

$$\mathbf{v} = [1 \quad 2 \quad 5]$$

Matrix

$$\mathbf{A} = \underbrace{\begin{bmatrix}1 & 2 & 3\\0 & 5 & 1\\2 & 3 & 7\\3 & 9 & 8\end{bmatrix}}_{3\text{ columns}} \left.\vphantom{\begin{bmatrix}1\\0\\2\\3\end{bmatrix}}\right\} 4\text{ rows}$$

Scalar Operation with Vector

```python
def scalar_vector_ops_numpy(s: float) -> None:
    import numpy as np

    v = np.array([1.0, -2.0, 3.5, 0.0], dtype=float)

    add = s + v          # broadcasting
    sub = v - s
    mul = s * v
    div = v / s

    # Useful extras
    dot = np.dot(v, v)    # dot product (v · v)
    norm = np.linalg.norm(v)

    print("== NumPy arrays ==")
    print(f"scalar s = {s}")
    print(f"vector v = {v}")
    print(f"s + v   = {add}")
    print(f"v - s   = {sub}")
    print(f"s * v   = {mul}")
    print(f"v / s   = {div}")
    print(f"v · v   = {dot}")
    print(f"||v||   = {norm:.6f}")
    print()
```

```
== NumPy arrays ==
scalar s = 2.0
vector v = [ 1.  -2.   3.5  0. ]
s + v   = [3.  0.  5.5 2. ]
v - s   = [-1.  -4.   1.5 -2. ]
s * v   = [ 2. -4.  7.  0.]
v / s   = [ 0.5  -1.    1.75  0.  ]
v · v   = 17.25
||v||   = 4.153312
```

Basics – Linear Algebra – Scalar/Vector/Matrix

Scalar

$$\mathbf{v} = 3$$

Vector

$$\mathbf{v} = [1 \quad 2 \quad 5]$$

Matrix

$$\mathbf{A} = \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix}}_{\text{3 columns}} \left.\vphantom{\begin{bmatrix} 1 \\ 0 \\ 2 \\ 3 \end{bmatrix}}\right\} \text{4 rows}$$

Scalar Operation with Matrix

```python
# ---------------------------------
# 1) Pure Python (list of lists)
# ---------------------------------
def scalar_matrix_ops_list(s: float, M: List[List[float]]) -> None:
    add = [[s + x for x in row] for row in M]   # scalar + matrix
    sub = [[x - s for x in row] for row in M]   # matrix - scalar
    mul = [[s * x for x in row] for row in M]   # scalar * matrix
    div = [[x / s for x in row] for row in M]   # matrix / scalar
```

```
scalar s = 2.0
matrix M =
[1.0, -2.0, 3.0]
[4.5, 0.0, -1.5]
s + M =
[3.0, 0.0, 5.0]
[6.5, 2.0, 0.5]
M - s =
[-1.0, -4.0, 1.0]
[2.5, -2.0, -3.5]
s * M =
[2.0, -4.0, 6.0]
[9.0, 0.0, -3.0]
M / s =
[0.5, -1.0, 1.5]
[2.25, 0.0, -0.75]
```

Basics – Linear Algebra – Scalar/Vector/Matrix

Scalar

$$\mathbf{v} = 3$$

Vector

$$\mathbf{v} = [1 \quad 2 \quad 5]$$

Matrix

$$\mathbf{A} = \left.\underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix}}_{\text{3 columns}}\right\} \text{4 rows}$$

Scalar Operation with Matrix

```
# ---------------------------------
# 2) NumPy (array-based)
# ---------------------------------
def scalar_matrix_ops_numpy(s: float) -> None:
    import numpy as np

    M = np.array([[1.0, -2.0, 3.0],
                  [4.5,  0.0, -1.5]], dtype=float)

    add = s + M          # broadcasting
    sub = M - s
    mul = s * M
    div = M / s
```

```
== NumPy arrays ==
scalar s = 2.0
matrix M =
[[ 1.  -2.   3. ]
 [ 4.5  0.  -1.5]]
s + M =
[[3.  0.  5. ]
 [6.5 2.  0.5]]
M - s =
[[-1.  -4.   1. ]
 [ 2.5 -2.  -3.5]]
s * M =
[[ 2. -4.  6.]
 [ 9.  0. -3.]]
M / s =
[[ 0.5  -1.    1.5 ]
 [ 2.25  0.   -0.75]]
```

Matrix Multiplication

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix} \times \quad \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 5 & 7 \end{bmatrix} \Rightarrow \quad \mathbf{C} = [\quad] \; ?$$

4×3 $\qquad$ 3×2

Dimensions much match!

What is the dimension of C? $(4 \times 3)(3 \times 2) \rightarrow 4 \times 2$

Matrix Multiplication

$$1 \times 1 + 2 \times 2 + 3 \times 5$$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix} \times \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 5 & 7 \end{bmatrix} \Rightarrow \mathbf{C} = \begin{bmatrix} \mathbf{20} & \\ & \\ & \end{bmatrix}$$

4×3 3×2

$$1 \times 2 + 2 \times 3 + 3 \times 7$$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix} \times \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 5 & 7 \end{bmatrix} \Rightarrow \mathbf{C} = \begin{bmatrix} 20 & \mathbf{29} \\ & \\ & \end{bmatrix}$$

4×3 3×2

Basics – Linear Algebra – Scalar/Vector/Matrix

Matrix Multiplication

$$0 \times 1 + 5 \times 2 + 1 \times 5$$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix} \times \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 5 & 7 \end{bmatrix} \Rightarrow \mathbf{C} = \begin{bmatrix} 20 & 29 \\ \mathbf{15} & \\ & \\ & \end{bmatrix}$$

$$4 \times 3 \qquad 3 \times 2$$

$$0 \times 2 + 5 \times 3 + 1 \times 7$$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix} \times \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 5 & 7 \end{bmatrix} \Rightarrow \mathbf{C} = \begin{bmatrix} 20 & 29 \\ 15 & \mathbf{22} \\ & \\ & \end{bmatrix}$$

$$4 \times 3 \qquad 3 \times 2$$

Matrix Multiplication

$$2 \times 1 + 3 \times 2 + 7 \times 5$$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix} \times \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 5 & 7 \end{bmatrix} \Rightarrow \mathbf{C} = \begin{bmatrix} 20 & 29 \\ 15 & 22 \\ \mathbf{43} & \\ & \end{bmatrix}$$

$$4 \times 3 \qquad 3 \times 2$$

$$2 \times 2 + 3 \times 3 + 7 \times 7$$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix} \times \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 5 & 7 \end{bmatrix} \Rightarrow \mathbf{C} = \begin{bmatrix} 20 & 29 \\ 15 & 22 \\ 43 & \mathbf{62} \\ & \end{bmatrix}$$

$$4 \times 3 \qquad 3 \times 2$$

Basics – Linear Algebra – Scalar/Vector/Matrix

Matrix Multiplication

$3 \times 1 + 2 \times 9 + 5 \times 8$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix} \times \quad \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 5 & 7 \end{bmatrix} \Longrightarrow \mathbf{C} = \begin{bmatrix} 20 & 29 \\ 15 & 22 \\ 43 & 62 \\ \mathbf{61} & \end{bmatrix}$$

4×3 $\qquad$ 3×2

$3 \times 2 + 3 \times 9 + 7 \times 8$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 1 \\ 2 & 3 & 7 \\ 3 & 9 & 8 \end{bmatrix} \times \quad \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 5 & 7 \end{bmatrix} \Longrightarrow \mathbf{C} = \begin{bmatrix} 20 & 29 \\ 15 & 22 \\ 43 & 62 \\ 61 & \mathbf{89} \end{bmatrix}$$

4×3 $\qquad$ 3×2

Basics – Linear Algebra – Scalar/Vector/Matrix

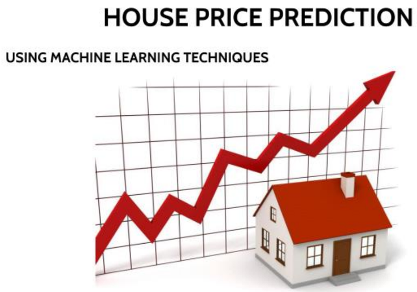
```python
def matmul_list(A: List[List[float]], B: List[List[float]]) -> List[List[float]]:
    """
    Compute C = A @ B using pure Python.
    A: m x n
    B: n x p
    C: m x p
    """
    m, n = len(A), len(A[0])
    n2, p = len(B), len(B[0])
    assert n == n2, "Inner dimensions must match"

    C = [[0.0 for _ in range(p)] for _ in range(m)]

    for i in range(m):
        for j in range(p):
            for k in range(n):
                C[i][j] += A[i][k] * B[k][j]
    return C
```

```python
def demo_numpy():
    import numpy as np

    A = np.array([
        [1, 2, 3],
        [4, 5, 6]
    ], dtype=float)

    B = np.array([
        [7,  8],
        [9, 10],
        [11, 12]
    ], dtype=float)

    C1 = A @ B          # preferred operator
    C2 = np.matmul(A, B)
    C3 = np.dot(A, B)  # works for 2D matrices
```

Basics – Physical Meaning of Matrix Multiplication in ML

House 1
Size – 1000 sqft
2 bed, 2 bath
Location: 3

House 2
Size – 2000 sqft
3 bed, 2 bath
Location: 2



House 3
Size – 1500 sqft
2 bed, 3 bath
Location: 4

Basics – Physical Meaning of Matrix Multiplication in ML

Contribution Coefficient 0.002, 1, 0.5, 1.2

House 1
Size – 1000 sqft
2 bed, 2 bath
Location: 3

1000*0.002 + 2*1 + 2*0.5 + 3*1.2
=8.6

House 2
Size – 2000 sqft
3 bed, 2 bath
Location: 2

2000*0.002 + 3*1 + 2*0.5 + 2*1.2
=10.4

House 3
Size – 1500 sqft
2 bed, 3 bath
Location: 4

1500*0.002 + 2*1 + 3*0.5 + 4*1.2
=11.3

Basics – Physical Meaning of Matrix Multiplication in ML

Contribution Coefficient 0.002, 1, 0.5, 1.2

House 1
Size – 1000 sqft
2 bed, 2 bath
Location: 3

House 2
Size – 2000 sqft
3 bed, 2 bath
Location: 2

House 3
Size – 1500 sqft
2 bed, 3 bath
Location: 4

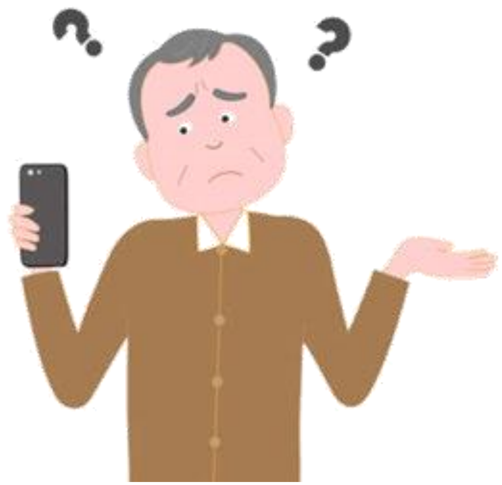
$$\mathbf{X} = \begin{bmatrix} 1k & 2k & 1.5k \\ 2 & 3 & 2 \\ 2 & 2 & 3 \\ 3 & 2 & 4 \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 0.002 \\ 1 \\ 0.5 \\ 1.2 \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{A}^{\mathrm{T}}\mathbf{X}$$

1 Basics

$$
\begin{aligned}
(\mathbf{AB})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} && (1)\\
(\mathbf{ABC}...)^{-1} &= ...\mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1} && (2)\\
(\mathbf{A}^{T})^{-1} &= (\mathbf{A}^{-1})^{T} && (3)\\
(\mathbf{A}+\mathbf{B})^{T} &= \mathbf{A}^{T}+\mathbf{B}^{T} && (4)\\
(\mathbf{AB})^{T} &= \mathbf{B}^{T}\mathbf{A}^{T} && (5)\\
(\mathbf{ABC}...)^{T} &= ...\mathbf{C}^{T}\mathbf{B}^{T}\mathbf{A}^{T} && (6)\\
(\mathbf{A}^{H})^{-1} &= (\mathbf{A}^{-1})^{H} && (7)\\
(\mathbf{A}+\mathbf{B})^{H} &= \mathbf{A}^{H}+\mathbf{B}^{H} && (8)\\
(\mathbf{AB})^{H} &= \mathbf{B}^{H}\mathbf{A}^{H} && (9)\\
(\mathbf{ABC}...)^{H} &= ...\mathbf{C}^{H}\mathbf{B}^{H}\mathbf{A}^{H} && (10)
\end{aligned}
$$



Matrix Codebook

https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

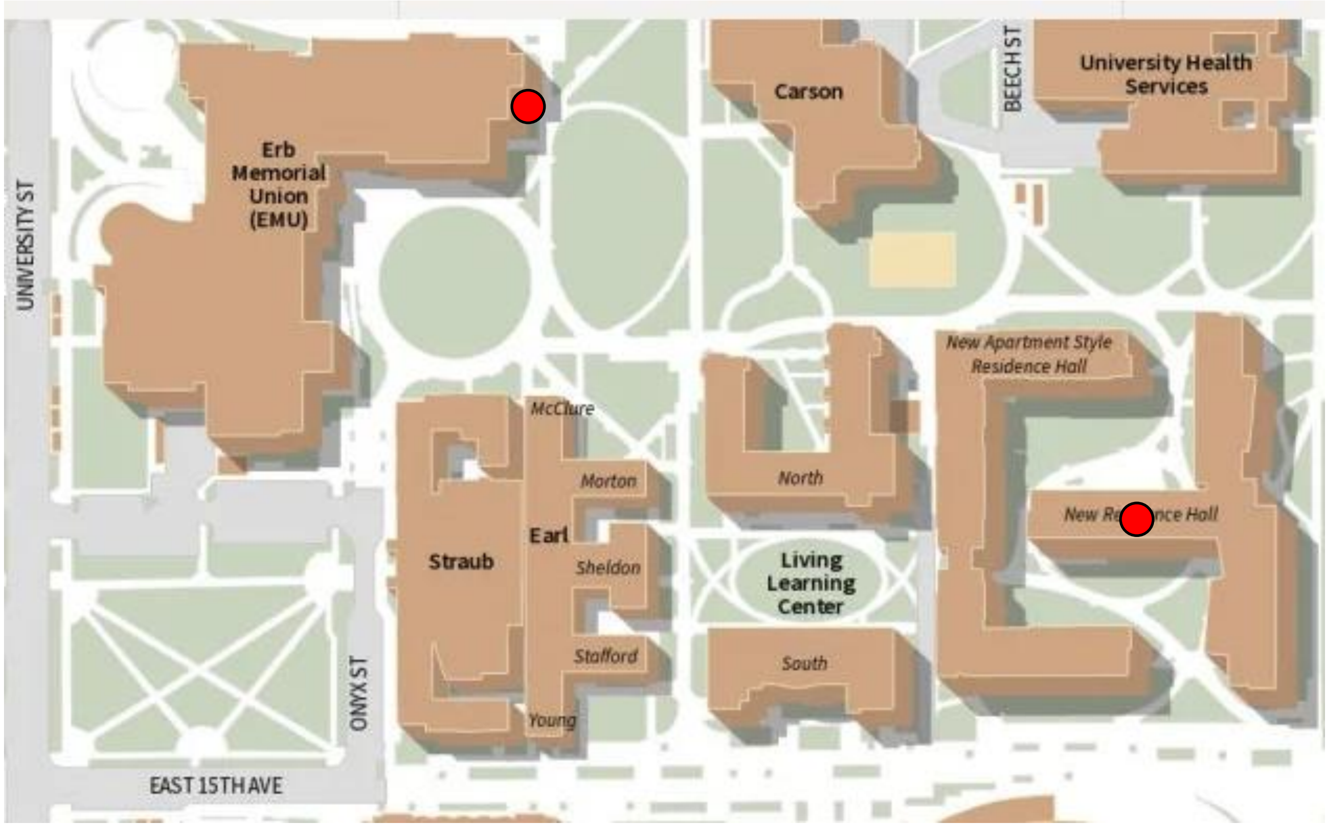
The Matrix Cookbook

[http://matrixcookbook.com]

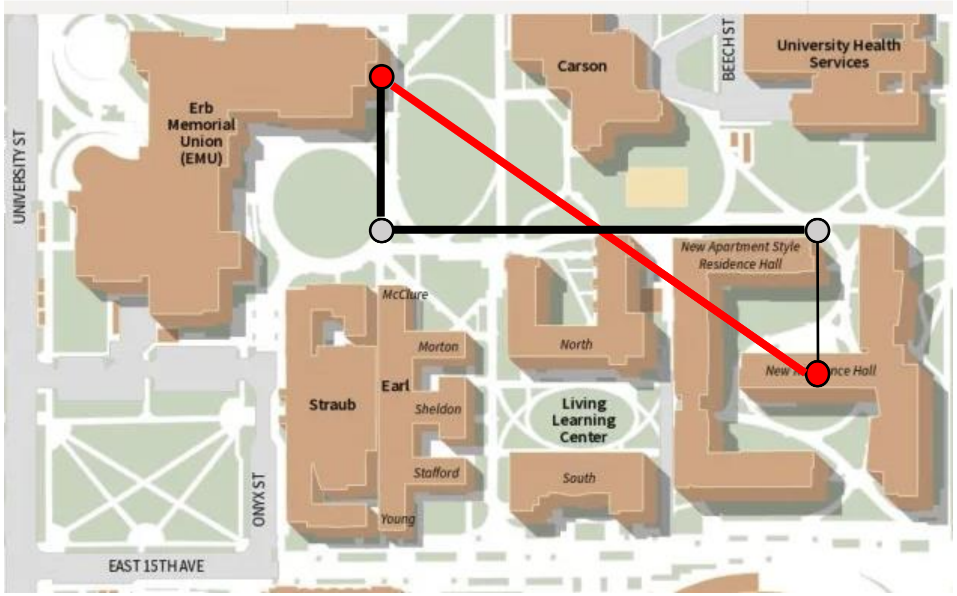
Kaare Brandt Petersen
Michael Syskind Pedersen

VERSION: NOVEMBER 15, 2012

Basics – Linear Algebra - Distance



Basics – Linear Algebra - Distance



$$D(a,b) = \left[(a_x - b_x)^2 + \left(a_y - b_y\right)^2\right]^{-0.5}$$

$$D(a,b) = |a_x - b_x| + |a_y - b_y|$$

Basics – Linear Algebra - Distance

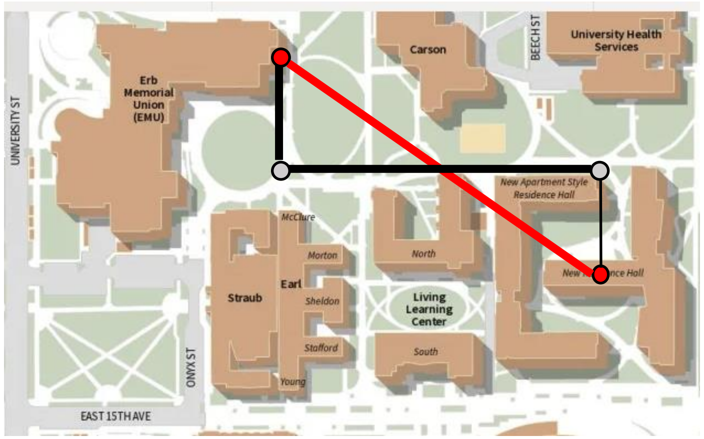
- $\|\mathbf{u}-\mathbf{v}\|^p$
- Function from a vector space to a single positive real value: $f: \mathbb{R}^d \rightarrow \mathbb{R}$
- Distance between **u** and **v**

$$\|\mathbf{u}-\mathbf{v}\|^p = \left(\sum_{i=1}^{d} |\mathbf{u}_i - \mathbf{v}_i|^p\right)^{\frac{1}{p}}$$

- Examples:

(1) Manhattan distance (L_1): $\|\mathbf{u}-\mathbf{v}\|^1 = \left(\sum_{i=1}^{d} |\boldsymbol{u}_i - \boldsymbol{v}_i|\right)$

(2) Euclidean distance (L_2): $\|\mathbf{v}\|^2 = \left(\sum_{i=1}^{d} |\mathbf{u}_i - \mathbf{v}_i|^2\right)^{\frac{1}{2}}$



Basics

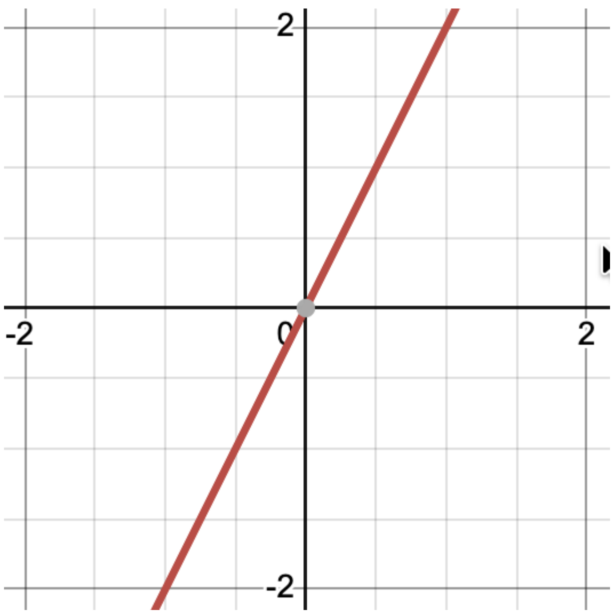
- Linear Algebra
- Calculus
- Statistics/Probability

Basics – Derivative and Gradient

$$y = 2x, \qquad \frac{dy}{dx} = 2$$

Scalar Input vs Scalar Output

How much change does the single unit change of x would cause on y?




```python
import torch

# Define input tensor with gradient tracking enabled
x = torch.tensor(3.0, requires_grad=True)

# Define function y = 2x
y = 2 * x

# Compute derivative dy/dx
y.backward()

# Access gradient
print("x =", x.item())
print("y =", y.item())
print("dy/dx =", x.grad.item())
```

✓ 0.0s

```
x = 3.0
y = 6.0
dy/dx = 2.0
```

Basics – Derivative and Gradient

$$y = 2x_1 + 3x_2, \qquad \frac{\partial y}{\partial x_1} = 2, \frac{\partial y}{\partial x_2} = 3$$

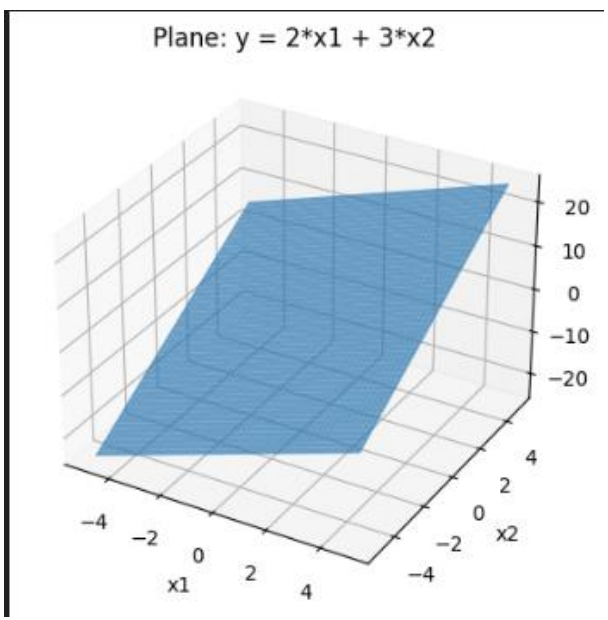
$$y = \mathbf{a}^{\mathrm{T}}\mathbf{x}$$

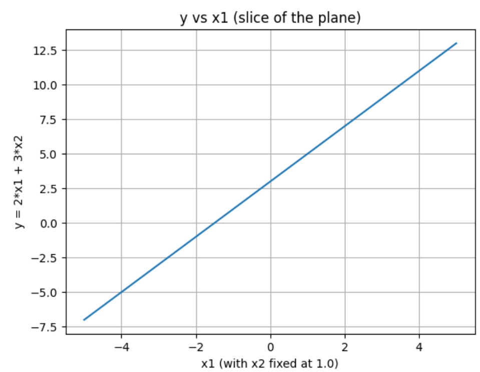
$$\mathbf{a} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

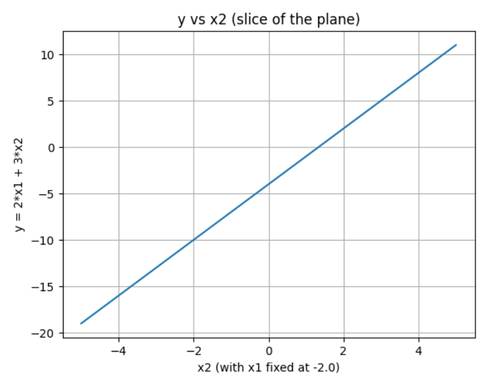
$$\nabla_x y = \boldsymbol{a}$$

Scalar Input
vs
Vector Output

How much change does the single unit change of x would cause on y?








```python
import torch

# x[0] = x1, x[1] = x2
x = torch.tensor([1.5, -0.5], requires_grad=True)

y = 2 * x[0] + 3 * x[1]
y.backward()

print("x1 =", x[0].item(), "x2 =", x[1].item())
print("y  =", y.item())
print("∂y/∂x1 =", x.grad[0].item())
print("∂y/∂x2 =", x.grad[1].item())
```

✓ 0.0s

```
x1 = 1.5 x2 = -0.5
y  = 1.5
∂y/∂x1 = 2.0
∂y/∂x2 = 3.0
```

Basics – Derivative and Gradient

$$y_1 = 2x_1 + 3x_2, \qquad \frac{\partial y_1}{\partial x_1} = 2, \frac{\partial y_2}{\partial x_2} = 3 \qquad \mathbf{a} = \begin{bmatrix} 2 & 4 \\ 3 & 3 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$y_2 = 4x_1 + 5x_2, \qquad \frac{\partial y_2}{\partial x_1} = 4, \frac{\partial y_2}{\partial x_2} = 3 \qquad \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

$$\boldsymbol{y} = \mathbf{a}^{\mathrm{T}}\mathbf{x}$$

Vector Input
vs
Vector Output

```
import torch

def f(x):
    """
    x: tensor of shape (2,) -> [x1, x2]
    returns: tensor of shape (2,) -> [y1, y2]
    """
    y1 = 2 * x[0] + 3 * x[1]
    y2 = 4 * x[0] + 5 * x[1]
    return torch.stack([y1, y2])

# Input
x = torch.tensor([1.0, 2.0], requires_grad=True)

# Compute Jacobian
J = torch.autograd.functional.jacobian(f, x)

print("Jacobian dy/dx:")
print(J)
```

```
✓ 0.0s

Jacobian dy/dx:
tensor([[2., 3.],
        [4., 5.]])
```

How much change does the single unit change of x would cause on y?

Basics – Example

```python
import numpy as np
import matplotlib.pyplot as plt
import torch
import torch.nn as nn
import torch.optim as optim
from sklearn.feature_extraction.text import CountVectorizer

# Dataset with overlapping vocabulary
ml_abstracts = [
    "We use a learning algorithm to model inflation based on training data.",
    "Neural network models predict inflation trends from economic data.",
    "Machine learning improves inflation forecasting with training-based models.",
    "This study applies a predictive learning algorithm to inflation modeling.",
    "Our model learns to predict inflation using neural networks and structured data.",
    "We evaluate algorithmic models that learn inflation patterns from data.",
    "Learning-based inflation models are trained on macroeconomic time series.",
    "The prediction model uses training data and machine learning techniques.",
    "Inflation is modeled using data-driven algorithms with predictive learning.",
    "Our learning system models inflation trends using economic indicators."
]
non_ml_abstracts = [
    "We analyze inflation using an economic theory-driven model and historical data.",
    "Inflation modeling relies on classical economic indicators and structured data.",
    "This study examines inflation dynamics using theoretical models and analysis.",
    "Inflation is analyzed through traditional models and economic frameworks.",
    "Economic data supports a theory-based approach to modeling inflation.",
    "The model explains inflation trends through historical economic data.",
    "A theory-driven framework models inflation based on macroeconomic variables.",
    "We use economic assumptions to analyze inflation using structured models.",
    "Historical data is used to model inflation without predictive algorithms.",
    "Inflation analysis is grounded in economic theory and empirical modeling."
]

texts = ml_abstracts + non_ml_abstracts
labels = [1] * len(ml_abstracts) + [0] * len(non_ml_abstracts)
```

Basics – Example

```
vectorizer = CountVectorizer(stop_words='english', binary=True)
X = vectorizer.fit_transform(texts)
feature_names = vectorizer.get_feature_names_out()
print(f"Number of keyword features: {len(feature_names)}")
print("Sample keywords:", feature_names[:10])
```

[130] ✓ 0.0s

```
Number of keyword features: 61
Sample keywords: ['algorithm' 'algorithmic' 'algorithms' 'analysis' 'analyze' 'analy
 'applies' 'approach' 'assumptions' 'based']
```

```
import torch
import torch.nn as nn

num_features = X.shape[1]          # vocabulary size (e.g. 80)
num_classes = len(set(labels))     # e.g. 2 for binary classification

model = nn.Linear(num_features, num_classes)
print(model)
# Linear(in_features=80, out_features=2, bias=True)
```

[122] ✓ 0.0s

```
Linear(in_features=61, out_features=2, bias=True)
```

Basics – Example

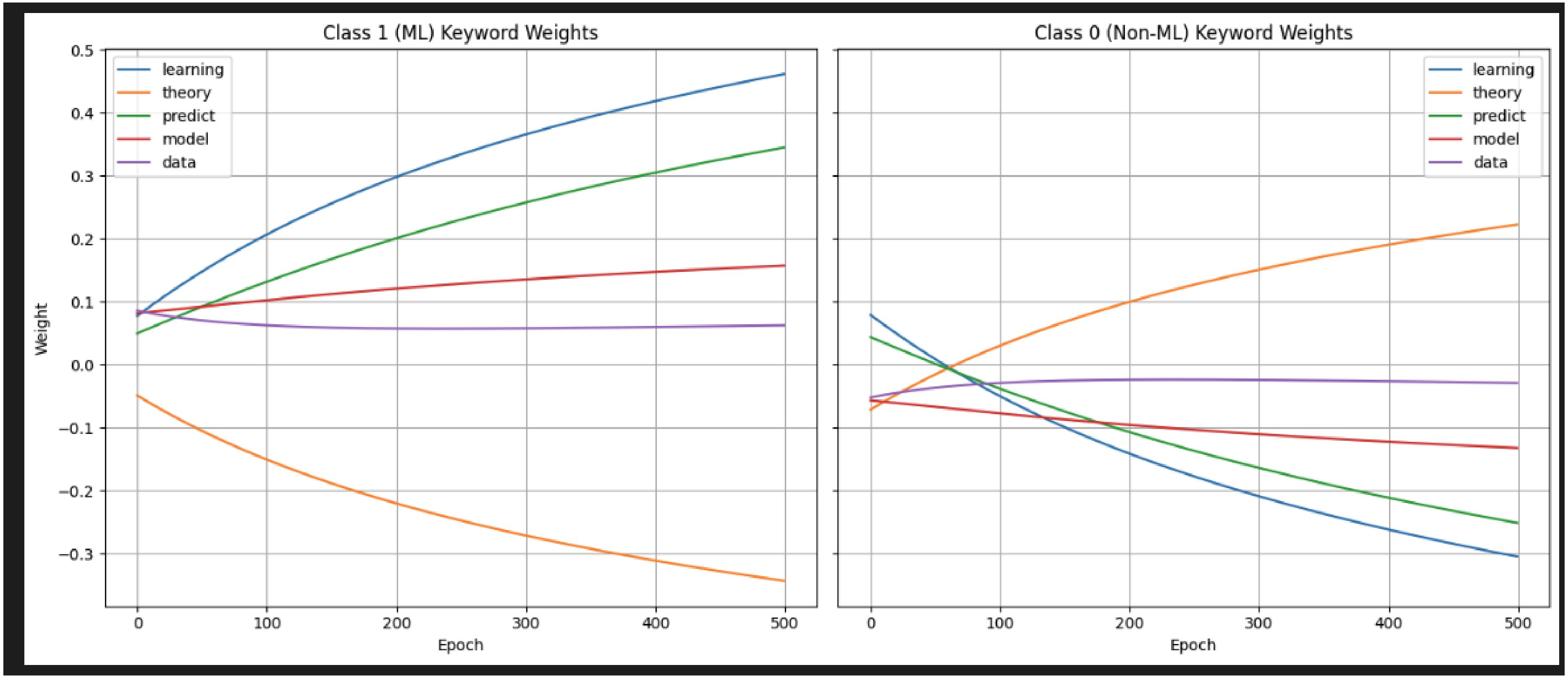
```python
from torch.utils.data import TensorDataset, DataLoader
import torch.optim as optim



# Convert feature matrix and labels to tensors
X_tensor = torch.tensor(X.toarray(), dtype=torch.float32)
y_tensor = torch.tensor(labels, dtype=torch.long)

# Group based on Y and for each group get 5 to get total 10
unique_labels = torch.unique(y_tensor)
train_indices = []
test_indices = []
for lbl in unique_labels:
    lbl_indices = (y_tensor == lbl).nonzero(as_tuple=True)[0]
    train_indices.extend(lbl_indices[:5].tolist())
    test_indices.extend(lbl_indices[5:].tolist())

X_train, X_test = X_tensor[train_indices], X_tensor[test_indices]
y_train, y_test = y_tensor[train_indices], y_tensor[test_indices]



# Define optimizer and loss function
optimizer = optim.SGD(model.parameters(), lr=0.01)
criterion = nn.CrossEntropyLoss()

# Training loop
epochs = 120
for epoch in range(epochs):
    model.train()
    optimizer.zero_grad()
    outputs = model(X_train)
    loss = criterion(outputs, y_train)
    loss.backward()
    optimizer.step()
```

```python
model.eval()
with torch.no_grad():
    logits = model(X_test)
    predictions = torch.argmax(logits, dim=1)
accuracy = (predictions == y_test).float().mean().item()
print(f"Test accuracy: {accuracy:.2f}")
```

✓ 0.0s

```
Test accuracy: 0.80
```

Basics – Example



Basics

- Linear Algebra

- Calculus

- Statistics/Probability

- **Probability → *from model to data***

A fair coin
P(Head) = 0.5

What is the probability of observing 7 heads in 10 tosses?

- **Statistics → *from data to model* *(machine learning as well)***

Data: 10 tosses → 7 heads

Question: Is the coin fair? What is p?

Basics – Probability

- **Sample Space: The set of all possible outcomes**
- **Event: A subset of the sample space**
- **Probability: under certain situation, how much likelihood of event**

Space {1, 2, 3, 4, 5, 6}

“Rolling an even number” = {2, 4, 6}

Basics – Probability

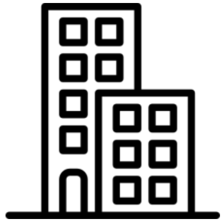
Event	Details	Formula (from English to mathematical operations)
A	Probability of A, **P(A)**	**P(A) is at or between zero and one:** $0 \le P(A) \le 1$
not A, $\mathbf{A^c}$	A^c is the complement of A	**Probability of not A = $P(A^c)$** = 1 - P(A)
A and B	A and B are independent events	**P(A and B)** = P(A)*P(B)
	A and B are dependent events	**P(A and B)** = P(A)*P(B | A) = **P(B)*P(A | B)** as 2 forms
	A and B are mutually exclusive events	**P(A and B)** = 0
A or B	A and B are independent events	**P(A or B)** = P(A) + P(B) - P(A)*P(B) conveniently expands to = **1 - [1 - P(A)]*[1 - P(B)]** or is obtained from De Morgan's Rule
	A and B are dependent events	**P(A or B)** = P(A) + P(B) - P(A)*P(B | A) as 1 of 2 forms
	A and B are mutually exclusive events	**P(A or B)** = P(A) + P(B)
A given B, **A | B**	Conditional: outcome of A given B has occurred	**P(A given B) = P(A | B)** = P(A)*P(B | A) / P(B) [Bayes' Thm] To make this formula, solve the 2 forms in "A and B" for P(A | B)

https://www.nasa.gov/wp-content/uploads/2023/11/210624-probability-formulas.pdf

Basics – Probability

$P(B=W) = 0.3$

$P(ND|W) = 0.6$



$P(B=G) = 0.5$

$P(ND|G) = 0.2$

$P(B=S) = 0.2$

$P(ND|S) = 0.05$

Basics – Probability

P(B=W) = 0.3

P(B=G) = 0.5

P(B=S) = 0.2

P(ND|W) = 0.4

P(ND|G) = 0.8

P(ND|S) = 0.95

After one earthquake, the building is not collapsed

$$P(G|ND) = \frac{0.8 * 0.5}{0.71} = 0.56 \qquad P(S|ND) = \frac{0.95 * 0.2}{0.71} = 0.27$$

$$P(T|ND) = \frac{P(ND|T)P(T)}{P(ND)}$$

$$P(W|ND) = \frac{P(ND|W)P(W)}{P(ND)} = \frac{0.4 * 0.3}{0.71} = 0.17$$

$$P(ND) = \sum_T P(ND|T)P(T)$$

$$P(ND) = \sum_T P(ND|T)P(T)$$
$$= 0.3 * 0.4 + 0.5 * 0.8 + 0.2 * 0.95 = 0.71$$

Basics – Probability

P(B=W) = 0.17 P(ND|W) = 0.4

P(B=G) = 0.56 P(ND|G) = 0.8

P(B=S) = 0.27 P(ND|S) = 0.95

After two earthquake, the building is not collapsed

$$P(G|ND) = \frac{0.8 * 0.17}{0.77} = 0.177 \quad P(S|ND) = \frac{0.95 * 0.27}{0.77} = 0.33$$

$$P(T|ND) = \frac{P(ND|T)P(T)}{P(ND)}$$

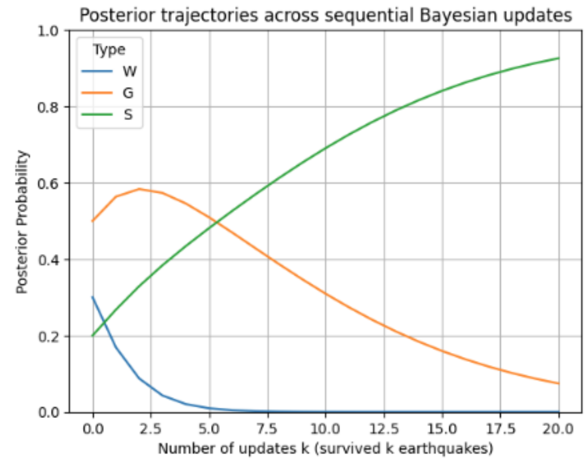
$$P(W|ND) = \frac{P(ND|W)P(W)}{P(ND)} = \frac{0.4 * 0.17}{0.77} = 0.09$$

$$P(ND) = \sum_T P(ND|T)P(T)$$

$$P(ND) = \sum_T P(ND|T)P(T)$$
$$= 0.17 * 0.4 + 0.56 * 0.8 + 0.27 * 0.95 = 0.77$$

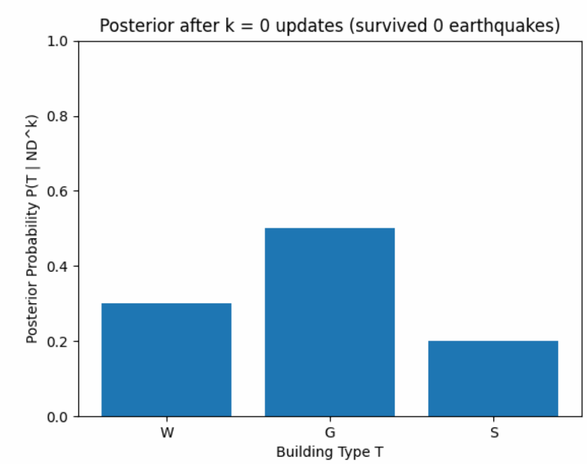
Basics – Probability

P(B=W) = 0.3



P(B=G) = 0.5

....



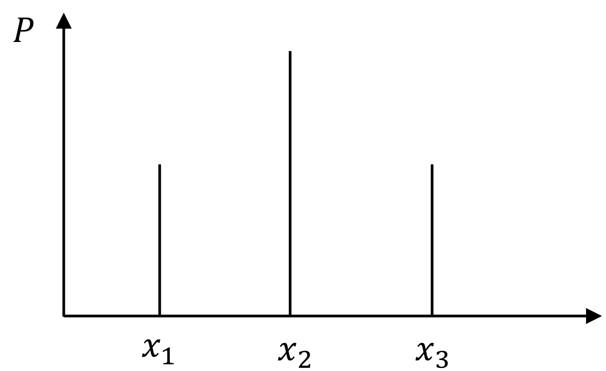
P(B=S) = 0.2

Basics – Probability

Probability of the random variable X taking the value x $\quad P_X(X = x)$

X: Type of Ball

x_1: Football	x_2: Vollyball	x_3: Basketball
5	10	5
$P_X(X = x_1)$	$P_X(X = x_2)$	$P_X(X = x_3)$

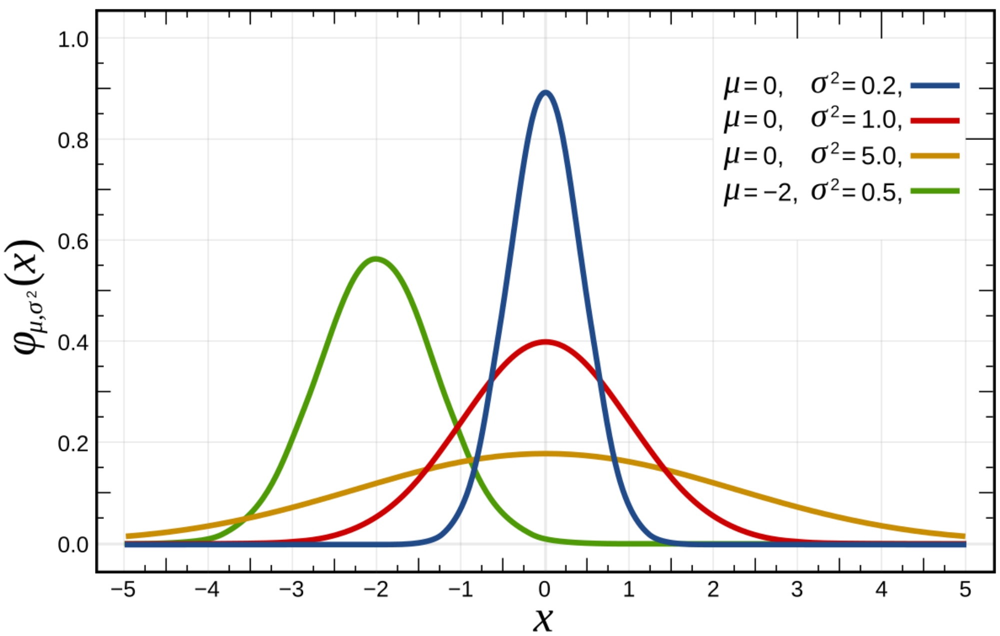


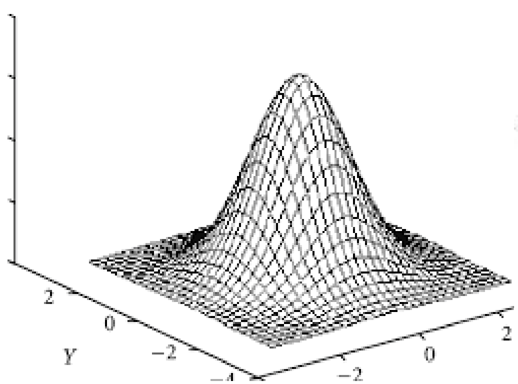
P_X

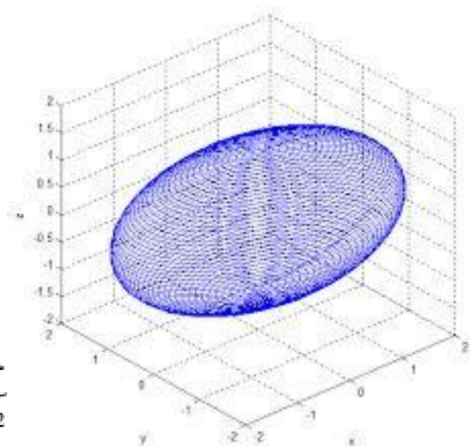
$P(X)$

Probability Distribution

Basics – Probability Density Function – Distribution







1-D Probability Density Function

2-D Probability Density Function

3-D Probability Density Function

N-D Probability Density Function

Basics – High Dimensional Random Variable

Dog

Cat

Basics – High Dimensional Random Variable

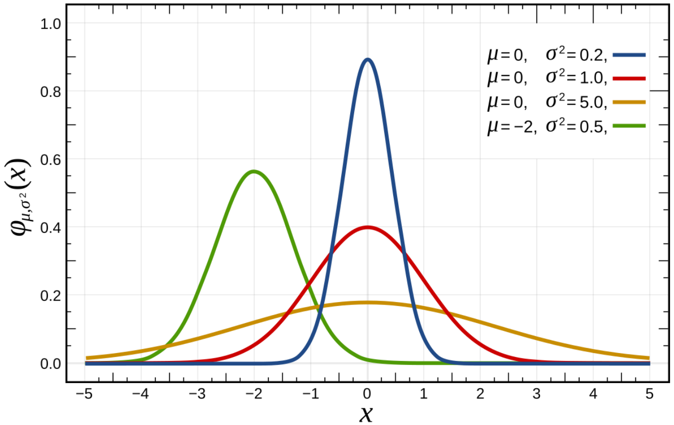
Dog – P(Dog)


Cat – P(Cat)


1. There is no concrete image/shape of the dog, everyone can come up with one of your own choice

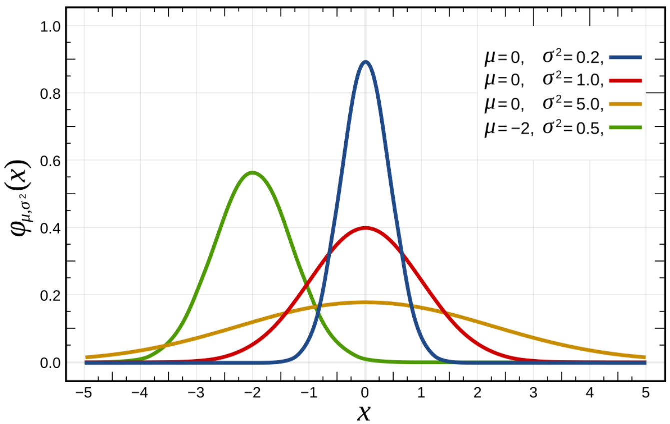
2. But somehow dog and cat image distributions are different

When you draw an image, you are actually sampling from a probability distribution!



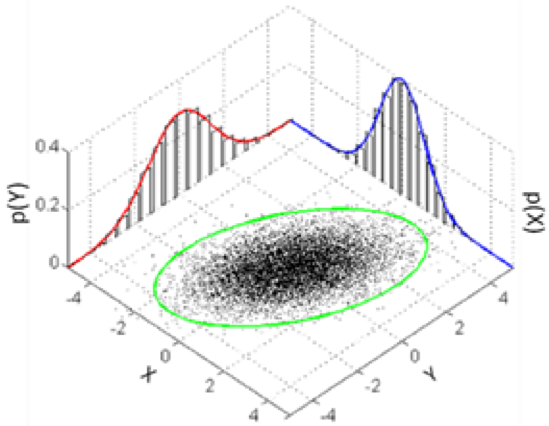
Basics – Data Distribution

1D Gaussian Distribution



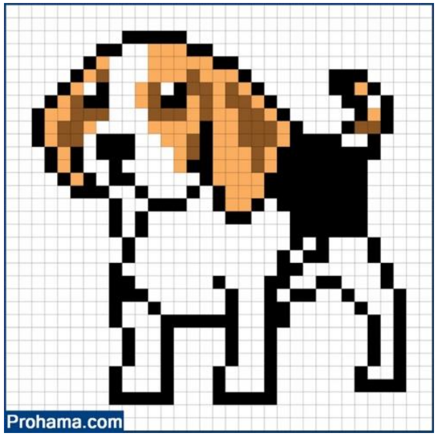
$\mathbb{R}$

2D Gaussian Distribution



$\mathbb{R}^2$

$\mathbb{R}^{256\times256}$

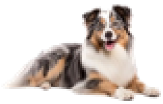


$\mathbb{R}^{256\times256}$

Basics – High Dimensional Random Variable

Probability of the random variable X taking the value x $\quad P_X(X = x)$

$X \in \mathbb{R}^{256\times256}$: Image

x_1: Image1	x_2: Image2	x_3: Image3
5	10	5
$P_X(X = x_1)$	$P_X(X = x_2)$	$P_X(X = x_3)$
		

P_X

$P(X)$

Probability Density Distribution

Basics – High Dimensional Random Variable

Probability of the random variable X, Y taking the value x, y $\quad P_{X,Y}(X = x, Y = y)$

Sampling something about Cat

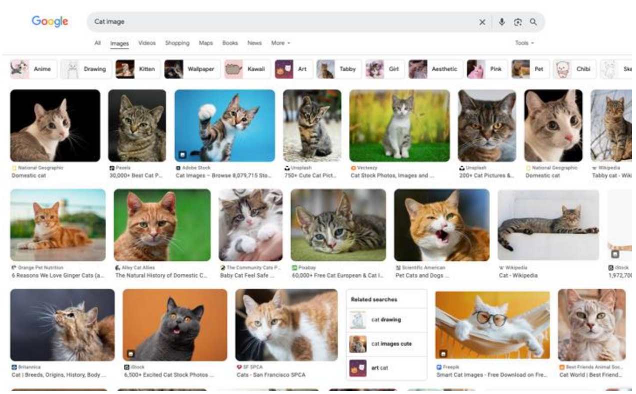
Sampling something about dog

$P_{\text{Image,Category}}(\text{Image} = \text{Image}, \text{Category} = \text{Category})$

Basics – High Dimensional Random Variable

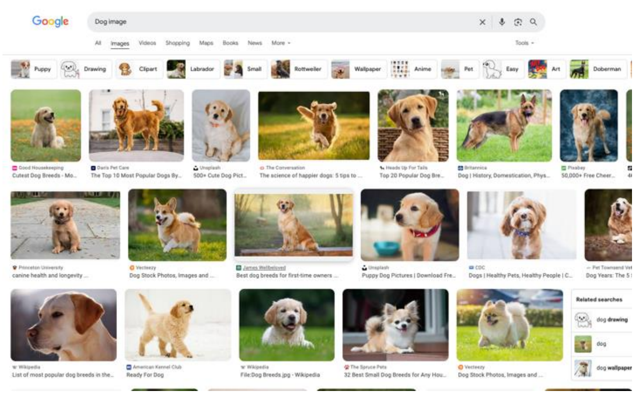
Probability of the random variable $X = x$ given $Y = y$ $\qquad P_{X|Y}(X = x|Y = y)$

Sampling something about



P(X|Y=Cat)

Sampling something about



P(X|Y=Dog)