CSC 84040: Advanced Data Mining *with AI Agents*

Lecture 01: Introduction, Distributed Techniques for Data Mining, and MapReduce

By Saptarashmi Bandyopadhyay
Email: sbandyopadhyay@ccny.cuny.edu
Assistant Professor of Computer Science
City College of New York and Graduate Center at the City University of New York

About Me

Positions & Education

- Tenure-Track Assistant Professor at CCNY/CUNY Graduate Center
  - Associate Director of CUNY CREST Entrepreneurship Pillar
  - Chair of the Research and Entrepreneurship Committee (REC)
  - Chair of the NY Agents Summit @ Manhattan
- Adjunct Assistant Professor of Computer Science at UMD, College Park
- PhD in Computer Science from UMD, College Park
- Recent PhD Student Researcher - Google AI AR & DeepMind (‘24-’25); Google X (‘23)
- Lead PhD Researcher in a DoD LTS project

Research Focus

- AI Agents
  - Energy Infrastructure, Housing Policy, Scientific Discovery and more!
- Multimodal Deep Learning (Semantic Understanding across Image, Text, Video, Audio)
- AI Alignment (Human-AI collaboration; Fixing Mistakes; Explainable AI)
- AI Agents Seminar Series: (https://sites.google.com/view/ai-agents-group)

Brief Bio

My Introduction

Federal Grant Awards and Collaborators

Community Leadership

AI Agents Open-Source Research Group

NEW YORK AI AGENTS SUMMIT
APRIL 30 – MAY 1, 2026 @ MANHATTAN

Industry Research

Research Collaborators and Grant Awards

Key Institutions, Partners, Collaborators of NSF AI-BIOME

AI-BIOME Key Partners & Collaborators

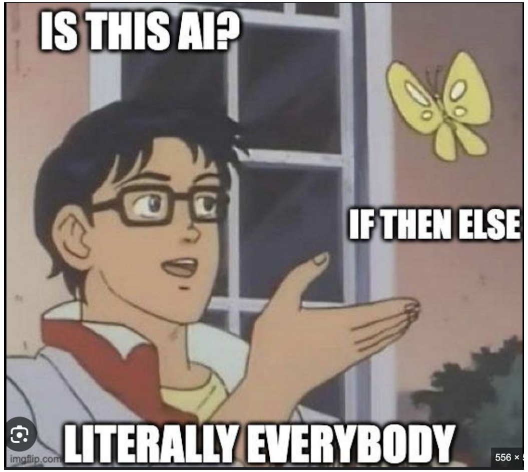
About Me

- PhD in Computer Science from the University of Maryland, College Park in Summer 2025
- Recent PhD Student Researcher at Google AI AR and Google DeepMind and as an AI Resident at Google X
- Lead PhD Researcher in a DoD LTS project
- Tenure-Track Assistant Professor of Computer Science at the City College of New York beginning Fall 2025

Brief Bio

Research Interests

- AI Agents
  - Improving capabilities of AI Agents
    - Real World Applications in Climate Conservation, Supply Chain Orchestration, Multimodal Agents, LLM Agents, Recommender Systems, Economics, AI Privacy etc
- Distributed Training (we know it as Modern Distributed Computing)
- Multimodal Deep Learning
  - Interpretable Semantic Understanding across Image, Text, Video, Audio modes
- AI Alignment
  - Human-AI collaboration; Fixing Mistakes of Humans/other AI Agents; Explainable AI



How to Reach Me

- Email: sbandyopadhyay@gc.cuny.edu
- Office: CUNY GC 4410
- Office Hours: TBD

Course Website

Brightspace, and also:

https://saptab.github.io/advanced-data-mining/

Course materials are subject to change, so check the website and BrightSpace for the most up to date information.

Note: Course materials will be on the website, but homeworks will only be on BrightSpace.

How to Reach Me

- Email: sbandyopadhyay@gc.cuny.edu
- Office: CUNY GC 4410
- Office Hours: TBD

Data Mining: Current backbone of AI

- Energy Efficiency
- Memory Efficiency
- Reliability
- Energy
- Robustness
- Generalizability
- Scale

The Bitter Lesson: Data <-> AI

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

In computer chess, the methods that defeated the world champion, Kasparov, in 1997, were based on massive, deep search. At the time, this was looked upon with dismay by the majority of computer-chess researchers who had pursued methods that leveraged human understanding of the special structure of chess. When a simpler, search-based approach with special hardware and software proved vastly more effective, these human-knowledge-based chess researchers were not good losers. They said that ``brute force" search may have won this time, but it was not a general strategy, and anyway it was not how people played chess. These researchers wanted methods based on human input to win and were disappointed when they did not.

A similar pattern of research progress was seen in computer Go, only delayed by a further 20 years. Enormous initial efforts went into avoiding search by taking advantage of human knowledge, or of the special features of the game, but all those efforts proved irrelevant, or worse, once search was applied effectively at scale. Also important was the use of learning by self play to learn a value function (as it was in many other games and even in chess, although learning did not play a big role in the 1997 program that first beat a world champion). Learning by self play, and learning in general, is like search in that it enables massive computation to be brought to bear. Search and learning are the two most important classes of techniques for utilizing massive amounts of computation in AI research. In computer Go, as in computer chess, researchers' initial effort was directed towards utilizing human understanding (so that less search was needed) and only much later was much greater success had by embracing search and learning.

In speech recognition, there was an early competition, sponsored by DARPA, in the 1970s. Entrants included a host of special methods that took advantage of human knowledge---knowledge of words, of phonemes, of the human vocal tract, etc. On the other side were newer methods that were more statistical in nature and did much more computation, based on hidden Markov models (HMMs). Again, the statistical methods won out over the human-knowledge-based methods. This led to a major change in all of natural language processing, gradually over decades, where statistics and computation came to dominate the field. The recent rise of deep learning in speech recognition is the most recent step in this consistent direction. Deep learning methods rely even less on human knowledge, and use even more computation, together with learning on huge training sets, to produce dramatically better speech recognition systems. As in the games, researchers always tried to make systems that worked the way the researchers thought their own minds worked---they tried to put that knowledge in their systems---but it proved ultimately counterproductive, and a colossal waste of

The Bias-Variance Tradeoff

In statistical models, significantly lowering bias inherently drives up variance.

- **Reducing Bias:** Bias occurs when data selection is targeted or restricted; the solution is to *randomly sample data* to significantly reduce bias.
- **Reducing Variance:** To combat the resulting high variance, the solution is to *reduce the learning rate*.

A lower learning rate provides more "learning time". For example, if a CS student and a Psychology student are paired, variance is high due to different academics, but giving them time to communicate reduces that variance.

Applying this reduced learning rate technique solved the notorious high-variance problem in the *REINFORCE* algorithm, establishing the backbone of the modern generative AI industry.

We use tools like Weka to analyze statistical data distributions to actively monitor and correct high bias or variance in our models.

Navigating Large Search Spaces in Scientific Datasets

The Scale of Molecular Spaces: The design space of small molecules and oligomers exceeds 10^{60} configurations; even a simple hexapeptide (6-mer) yields 20^6 discrete variants.

The Expensive Evaluation Bottleneck:

- Physical synthesis, cell plating, and spectroscopic characterization take hours to days.
- Experimental throughput is strictly budget-constrained (N total evaluations).

The CS/DS Mission: We cannot brute-force physical discovery through scale; agents must maximize sample efficiency to locate global performance peaks with minimal physical trials.

Scientific Discovery at CUNY CCNY, ASRC and NYC

- 10+ years experience **building & running core facilities** with 1000s of users from academia and industry.
- Embedded in **the nations' largest urban university system.**
- Connecting AI Agents in the Real World
  - Scientific Discovery with **life sciences** and other domains of **physical science & engineering.**
  - **Energy Orchestration with Smart Grids**
  - **Autonomous Vehicles (Drones, Cars..)**
  - **Quantum Control**
  - **Semiconductor Chip Design**
  - **Financial Portfolio Optimization**

ADVANCED SCIENCE RESEARCH CENTER
THE GRADUATE CENTER
CITY UNIVERSITY OF NEW YORK

The City College of New York

Why Use AI Agents Now? AI-BIOME as a Catalyst

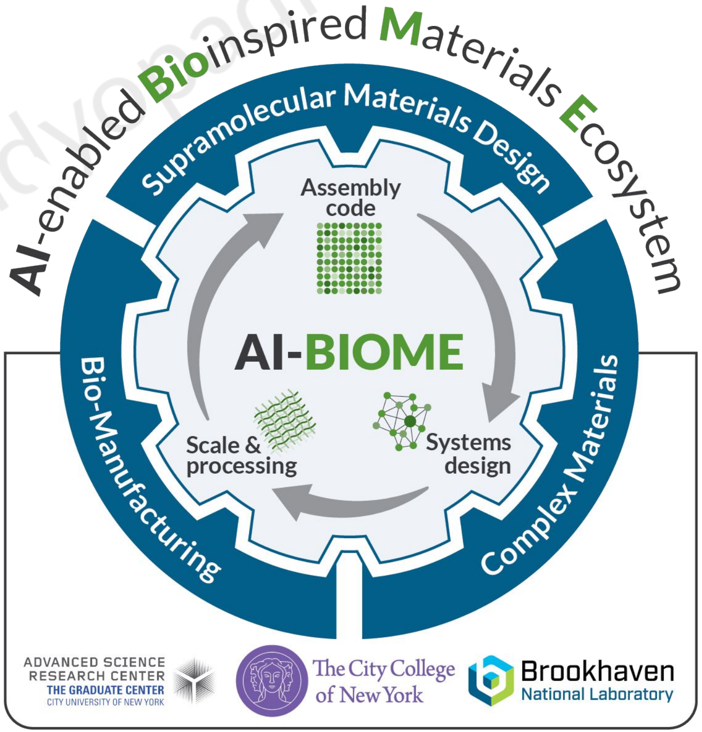
$18.1M NSF Programmable Cloud Laboratory (PCL): A self-driving node for bio-derived, supramolecular, and bioinspired materials.

Autonomous Experimentation: Orchestrating the bio-manufacturing and materials design ecosystem.

National Access: Enabling remote users with a scalable, reproducible synthesis-and-characterization pipeline.

AI-BIOME Key Partners & Collaborators

Brookhaven National Laboratory · EMPIRE AI · SIMONS FOUNDATION · CUNY I-CORPS · Columbia University · MIT · NYU · Memorial Sloan Kettering Cancer Center · LaGUARDIA COMMUNITY COLLEGE · COMMUNITAS AMERICA · NVIDIA · ILLUMINATIONSPACE · NanoBioNYC THE GRADUATE CENTER CITY UNIVERSITY OF NEW YORK · moz://a ai · ASRC cat · opentrons · GOTHAM FOUNDRY · LIFE SCIENCE CARES · NY ROBOTICS



AI-BIOME aim is to revolutionize bioinspired technologies & materials discovery

NanoBioNYC
THE GRADUATE CENTER
CITY UNIVERSITY OF NEW YORK

- **NSF NRT** NanoBioNYC program (2022–2028)
- **NSF CREST** Interface Design and Engineered Assembly of Low Dimensional Systems (IDEALS) (2016–2026)
- **NSF REU programs** NanoNY (Columbia/CUNY) (renewed); B^3 CCNY
- **$10 M** NYS-funded Industry Engagement Center for Advanced Technology (CAT)
- **$31M research funding** in bioinspired materials from NSF, DoD, Foundations at ASRC
- **$42 M** NYC EDC Gotham Foundry Materials Innovation Hub

GOTHAM FOUNDRY

Gotham Foundry Launch event (Fall 2025)

National Science Foundation

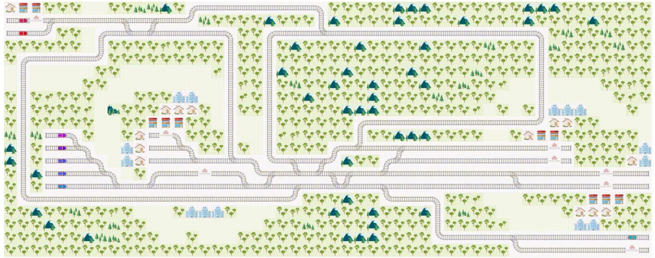
SIMONS FOUNDATION

ALFRED P. SLOAN FOUNDATION

Office of Naval Research Science & Technology

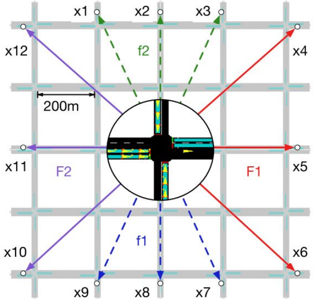
Recent publications in *Nature Materials* (2021, 2025) *Nature* (2022, 2023); *Science* (2023); *Chem* (2021, 2022, 2025); *Matter* (2026); *Advanced Materials* (2020, 2021, 2022, 2022); *JACS* (2021, 2022, 2023, 2025); *Angewandte Chemie* (2021, 2023, 2024, 2026) *P.N.A.S. (2021, 2022, 2023, 2024).*

Multi-Agent AI Applications



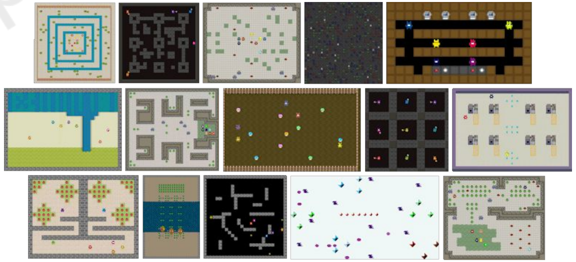
Train scheduling

[Mohanty et al., 2020]



Traffic signal control

[Chu et al., 2020]



Games

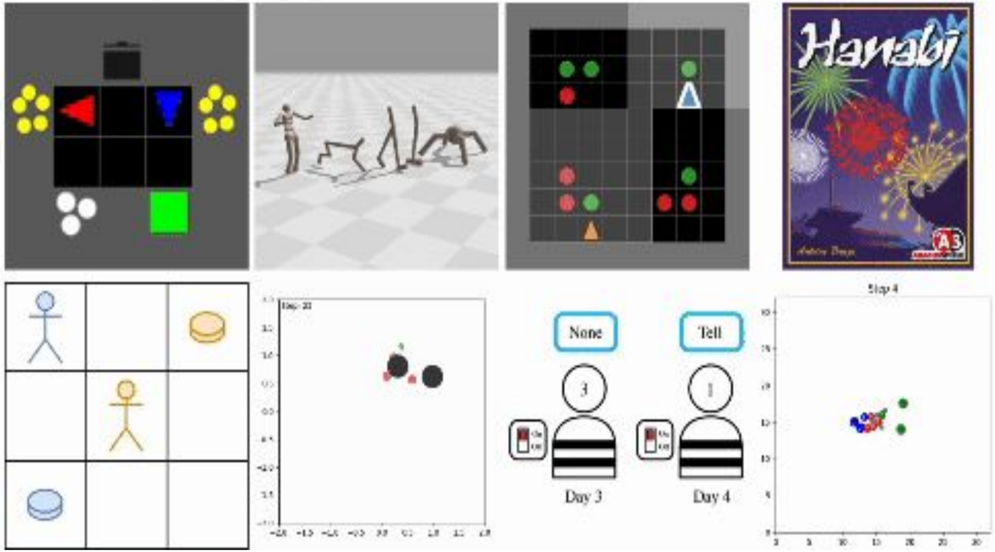
[Leibo et al., 2021]

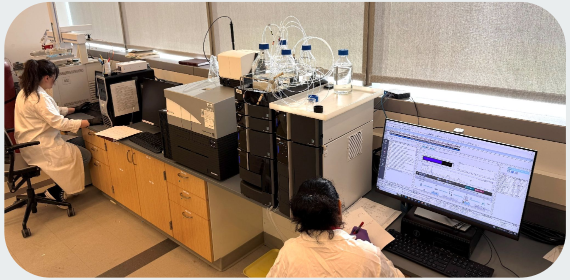
[Mohanty et al., 2020] Mohanty et al., 2020. Flatland-RL: Multi-Agent Reinforcement Learning on Trains. https://arxiv.org/abs/2012.05893. 2020.
[Chu et al., 2020] Chu et al., 2020. Multi-Agent Deep Reinforcement Learning for Large-Scale Traffic Signal Control. IEEE Transactions on Intelligent Transportation Systems. 2020.
[Leibo et al., 2021] Leibo et al., 2021. Scalable Evaluation of Multi-Agent Reinforcement Learning with Melting Pot. ICML. 2021.

Multi-Agent AI Environments for Games

- We implement JAXMARL, the fastest Multi-Agent AI training library
- Distributed AI Agents for 8 popular Multi-Agent AI environments
- Only 17 lines of code to train 12x faster than PyTorch libraries

```python
import jax
from jaxmarl import make

key = jax.random.PRNGKey(0)
key, key_reset, key_act, key_step = jax.random.split(key, 4)

# Initialise and reset the environment.
env = make('MPE_simple_world_comm_v3')
obs, state = env.reset(key_reset)

# Sample random actions.
key_act = jax.random.split(key_act, env.num_agents)
actions = {agent: env.action_space(agent).sample(key_act[i]) \
            for i, agent in enumerate(env.agents)}

# Perform the step transition.
obs, state, reward, done, infos = env.step(key_step, state, actions)
```





Front end: HTE synthesis + formulation + sample preparation

- 3,000 ft² CCNY high-throughput experimentation facility (HTE)
- Tecan liquid handler/plate reader
- 2 Mantis V4 liquid handlers
- High-throughput ultra-performance liquid chromatography–mass spectrometry (UPLC/MS) system with 8-column changer and autosampler
- High-throughput gas chromatography–mass spectrometry (GC/MS) with autosampler
- Planned autonomous modules: spin coaters, sample dryers, light-scattering, reflectometry, etc.

Back end: high-end multimodal characterization

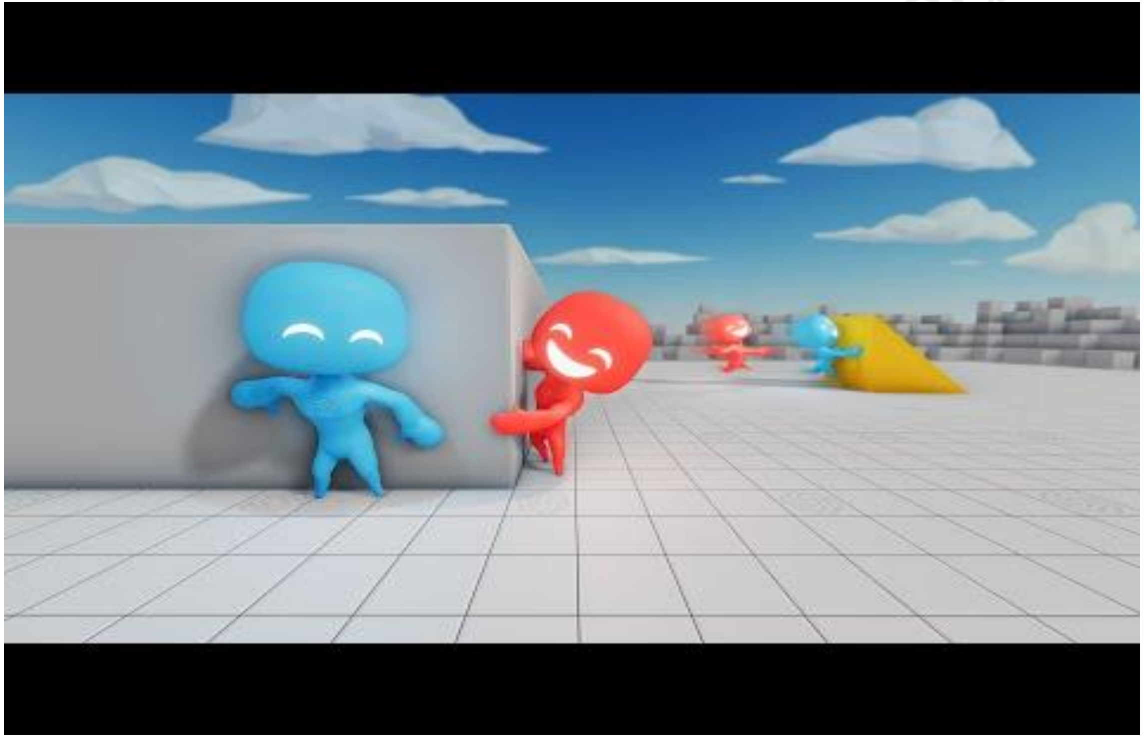
- 200,000 ft² research center and user facilities provide world-class material characterization instrumentation with autonomous data collection and interpretation capabilities.
- **Surface science core:** AFM, Bio-AFM, Spinning disk confocal microscope, XPS, and TOF-SIMS
- **Imaging core:** Cryo-EM, TEMs, SEM-FIB, environmental SEM, confocal Raman, and XRD.
- **Photonics core:** Ultrafast/supercontinuum laser, photoluminescence/pulsed/CW/ imaging spectrometers, and ellipsometer.

Control + compute layer

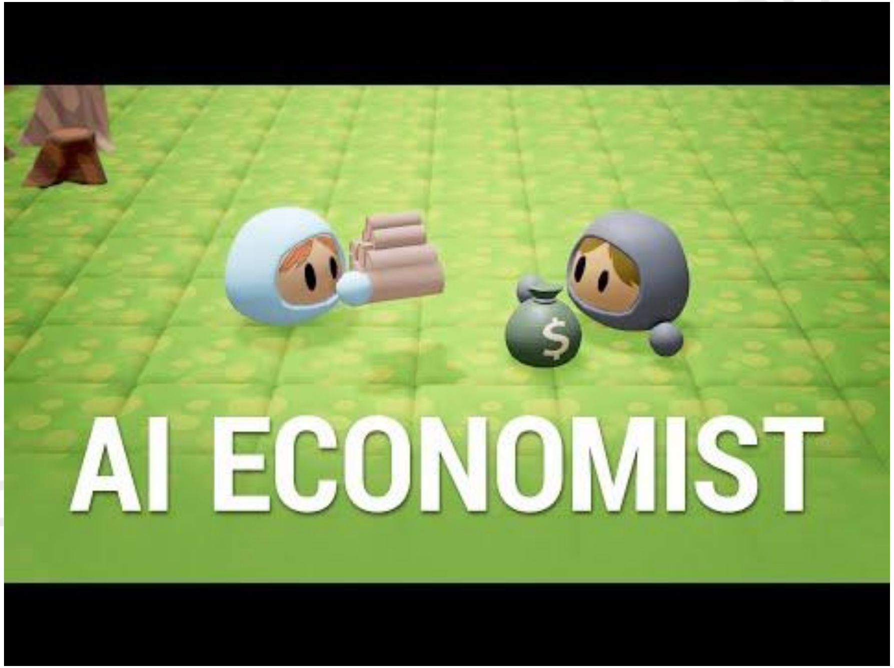
- ASRC high-performance computing (HPC)+ broader CUNY AI hardware ecosystem
- ASRC HPC: 1 PetaByte storage, 10 Intel Xeon Emerald Rapids Silver 24-core CPUs, 20 NVIDIA RTX 6000 Ada Generation GPUs for a total of 363,520 CUDA cores.
- 10-30 (up to 120) Dell R640 CPU nodes are scheduled to be added from the Simons Foundation CPU gift to the GC CUNY.
- Empire AI and CUNY HPCC systems.

AI Agents: Overview and Applications

Multi-agent RL for Hide-and-Seek (OpenAI, 2019)



Tax Policy with Multi-agent RL (Zheng et. al, 2020)



When AI Agents Fail

Claude Code operated a vending machine.

It stocked it with tungsten cubes, live fish, and a playstation, losing hundreds of dollars!

AI Agents are still very much an open problem.

Questions?