

# Big Data Conference 2026

Harvard University Center for Mathematical Sciences and Applications (CMSA) | September 3–4, 2026

## Thursday, September 3, 2026

08:45 AM – 09:10 AM Breakfast & Registration

09:10 AM – 09:15 AM Opening Remarks & Introductions

### 09:15 AM – 10:15 AM A Riemannian Geometry Perspective on Foundation Models

Speaker: [Rex Ying \(Yale University\)](#)

**Abstract:** In the development of foundation models and LLMs, Euclidean space has been the de facto geometric setting for machine learning architectures. However, at scale, real-world data across languages, vision, and natural sciences often exhibit inherently non-Euclidean structures—such as multi-way relationships, hierarchies, symmetries, and non-isotropic scaling. Effectively capturing these structures within Euclidean constraints remains challenging. This talk covers research moving beyond Euclidean geometry to maintain the scaling law for next-generation foundation models. Adopting non-Euclidean geometries allows foundation models to leverage data structures more efficiently, while task-aware adaptability dynamically reconfigures embeddings to match downstream applications. The presentation demonstrates the advantages of hyperbolic geometry across non-Euclidean components, LLM pre-training/fine-tuning, retrieval-augmented generation (RAG), and mixture-of-experts (MoE) models, introduces a library simplifying non-Euclidean model adaptation, and outlines promising future directions.

10:15 AM – 10:30 AM Morning Coffee Break

### 10:30 AM – 11:30 AM Toward Universal Steering and Monitoring of AI Models

Speaker: [Adit Radhakrishnan \(MIT\)](#)

**Abstract:** AI models encode vast amounts of human knowledge; understanding these internal representations is key to advancing model capabilities and safeguards. Building on advances in feature learning, this work presents an approach for extracting linear representations of semantic concepts in AI models. These internal representations enable model steering—exposing vulnerabilities and enhancing performance beyond standard prompting. Concept representations proved transferable across languages and enabled multi-concept steering, with larger models demonstrating higher steerability. Furthermore, concept representations proved more effective for monitoring misaligned content than judge models, underscoring their utility for AI safety and capability alignment.

11:30 AM – 12:45 PM Lunch Break

### 12:45 PM – 01:45 PM Crowdsourcing Mathematical Databases

Speaker: [Andrew Sutherland \(MIT\)](#)

**Abstract:** The L-functions and Modular Forms Database (LMFDB) is a major online repository of mathematical research data, housing catalogs of objects arising in the Langlands program. In collaboration with the Foundation for Science and AI Research (SAIR), a recent competition aimed to solve the inverse Galois problem over  $\mathbb{Q}$  in degree 24 (IGP24). By collecting over 50 million candidate number fields submitted by more than 100 teams, the initiative realized all 25,000 transitive permutation groups of degree 24 as Galois groups. AI tools and computer algebra systems played a central role in constructing the competition infrastructure and expanding participant capabilities.

01:45 PM – 02:00 PM Afternoon Break

02:00 PM – 03:00 PM

## Algorithmic Tools for Trading Off Sortition Ideals

Speaker: [Bailey Flanigan \(MIT\)](#)

**Abstract:** Citizens' assemblies and deliberative minipublics rely on sortition—the random selection of a small, representative group of everyday citizens to deliberate on policy issues. Sortition has become a prominent domain of computer science research, yielding algorithms now deployed in practice by governments worldwide. This talk explores the core practical challenges of sortition and details the algorithmic tools developed to navigate trade-offs among representative ideals.

03:00 PM – 03:15 PM

## Afternoon Break

03:15 PM – 04:15 PM

## Prescriptive Learning at Scale: Applied and Deployed Decision-Making

Speaker: [Chris Wiggins \(Columbia University\)](#)

**Abstract:** Many healthcare and industrial applications require real-time decision-making while concurrently learning response behaviors—such as article recommendation, treatment assignment, or paywall deployment. Addressing the prescriptive question of what to do (rather than what is true), statistical decision theory has evolved rapidly with modern online experimentation. This talk outlines 21st-century developments spanning adaptive experimentation, contextual bandits, and off-policy evaluation. Combining supervised learning with statistical sampling, these frameworks now power digital platforms and adaptive medical interventions at scale.

# Friday, September 4, 2026

08:45 AM – 09:15 AM

## Breakfast & Networking

09:15 AM – 10:15 AM

## No Generation Without Representation

Speaker: [Ariel Procaccia \(Harvard University\)](#)

**Abstract:** AI systems and democratic institutions face parallel challenges regarding representation. This talk addresses two cross-cutting questions: How AI can facilitate democratic processes that handle vast opinion spaces while guaranteeing proportional representation, and whose viewpoints are reflected when AI systems offer normative guidance. Leveraging social choice theory, formal frameworks and algorithms demonstrate that rigorous representation guarantees are both mathematically feasible and practical.

10:15 AM – 10:30 AM

## Morning Coffee Break

10:30 AM – 11:30 AM

## Biobank-Scale Genetic Discovery: From Association Testing to Global Meta-Analysis

Speaker: [Wei Zhou \(Harvard University\)](#)

**Abstract:** Genomic data linked to electronic health records offer vast opportunities for complex disease genetics, but present major statistical and computational hurdles. Within individual biobanks, models must address population structure, highly unbalanced case-control ratios, rare variants, longitudinal outcomes, and massive scale. Across biobanks, ancestral diversity, phenotype variability, and recruitment differences introduce further complexity. This presentation outlines methods developed for biobank analytics—including scalable generalized linear mixed models, survival mixed models, and region-based rare-variant tests—and highlights the Global Biobank Meta-analysis Initiative (GBMI) to demonstrate how global data integration boosts discovery power without sharing individual-level data.

11:30 AM – 12:45 PM

## Lunch Break

12:45 PM – 01:45 PM

## Using AI for Protein Structure Modeling and Design

Speaker: [Sergey Ovchinnikov \(MIT\)](#)

**Abstract:** This talk focuses on model interpretability in computational biology, specifically decomposing protein language models and structure prediction frameworks like AlphaFold. By analyzing internal representations and limitations, researchers can derive insights to build improved architectures and advance de novo protein design.

01:45 PM – 02:00 PM      Afternoon Break

02:00 PM – 03:00 PM

## **Computing with Neural Manifolds: A Multi-Scale Framework for Understanding Biological and Artificial Neural Networks**

Speaker: [SueYeon Chung \(Harvard University / Flatiron Institute\)](#)

**Abstract:** Biological and artificial neural networks process information across scales, from individual neurons to macroscopic cognitive functions. At the mesoscopic level, population activity forms neural representations whose transformation drives computation. This talk presents three approaches rooted in statistical physics, machine learning, and geometry: 1. Frameworks connecting geometric properties of neural manifolds to task execution efficiency. 2. Multi-scale analyses tracing how single-neuron properties, learning dynamics, and inter-regional transformations shape representations. 3. Extensions of efficient coding principles connecting representational geometry to task performance, providing a foundation for neuroscience-inspired artificial neural network design.