Frequency
Every year

Remark
Next time: Spring 2023

Summary
This course will cover the latest technologies, platforms and research contributions in the area of machine learning systems. The students will read, review and present papers from recent venues across the systems for ML spectrum.

Content
The course will cover recent papers from the literature in the emerging area of ML systems. With abundance of data and the emergence of data science, machine learning is widely applicable in a variety of usage scenarios with high performance, accuracy and cost being key design goals. The latter not only has implications for algorithms but also platforms from software to hardware to enable collective optimization of the design metrics. The topic is inherently multidisciplinary and will cover papers from a variety of conferences in computer science subfields; e.g., ICML, NIPS, ICLR, KDD, VLDB, SIGMOD, SOSP, OSDI, SysML, ASPLOS and ISCA).

Students will understand the state-of-the-art in the emerging area of ML Systems. Specifically the cover core technologies in production ML systems including: (1) languages and paradigms for specification of large-scaling machine learning applications, (2) the convergence of analytics from relational databases to unstructured data, (3) resource management in large-scaling ML systems, (4) network stacks for ML systems, and (5) emerging ML systems accelerator architecture.

In this course students learn to read, understand, critique and present research papers.

Keywords
Machine Learning; Systems

Learning Prerequisites
Required courses
Graduate level course in architecture, databases, systems

Recommended courses
Graduate level course in architecture, databases, systems