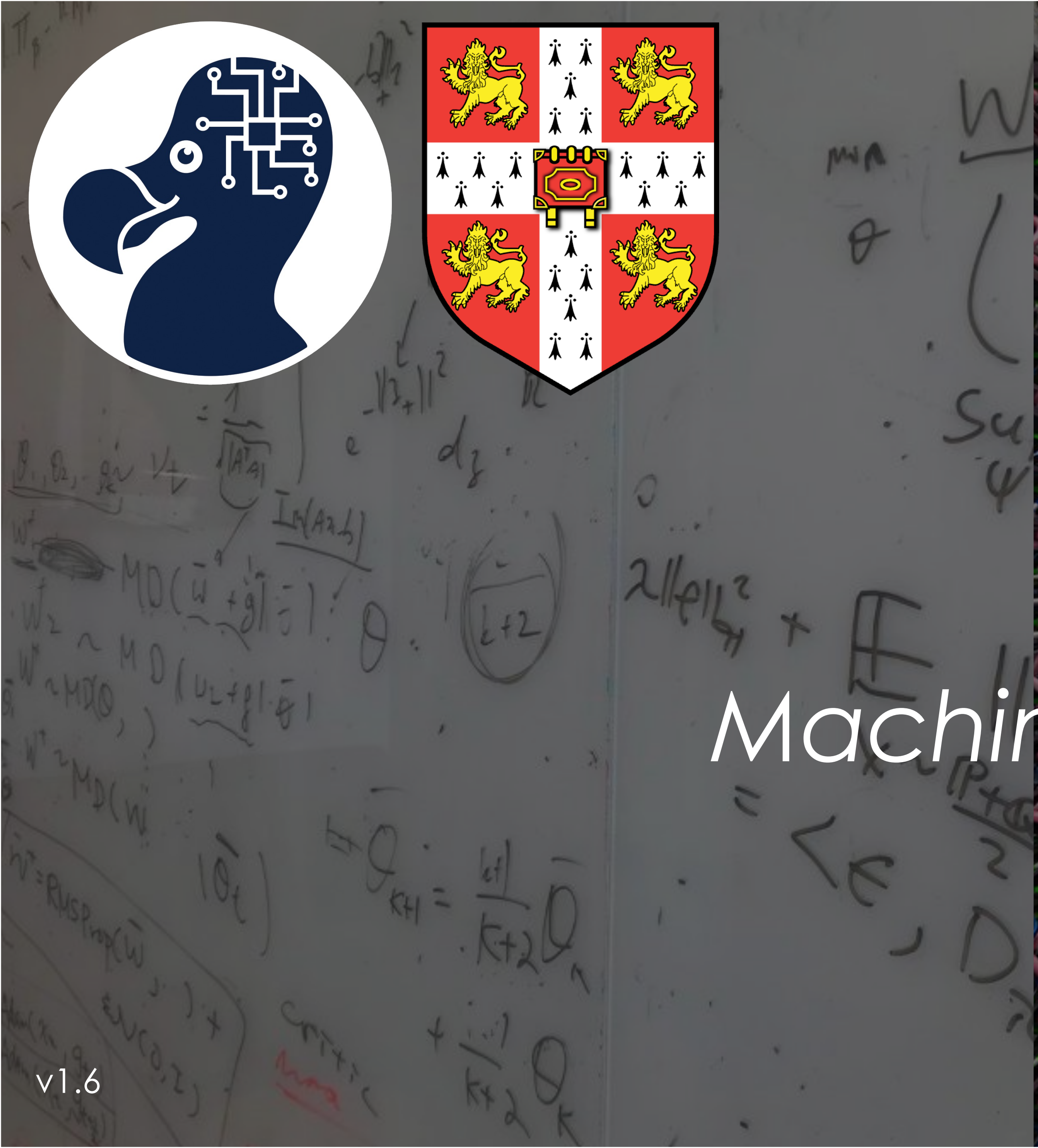
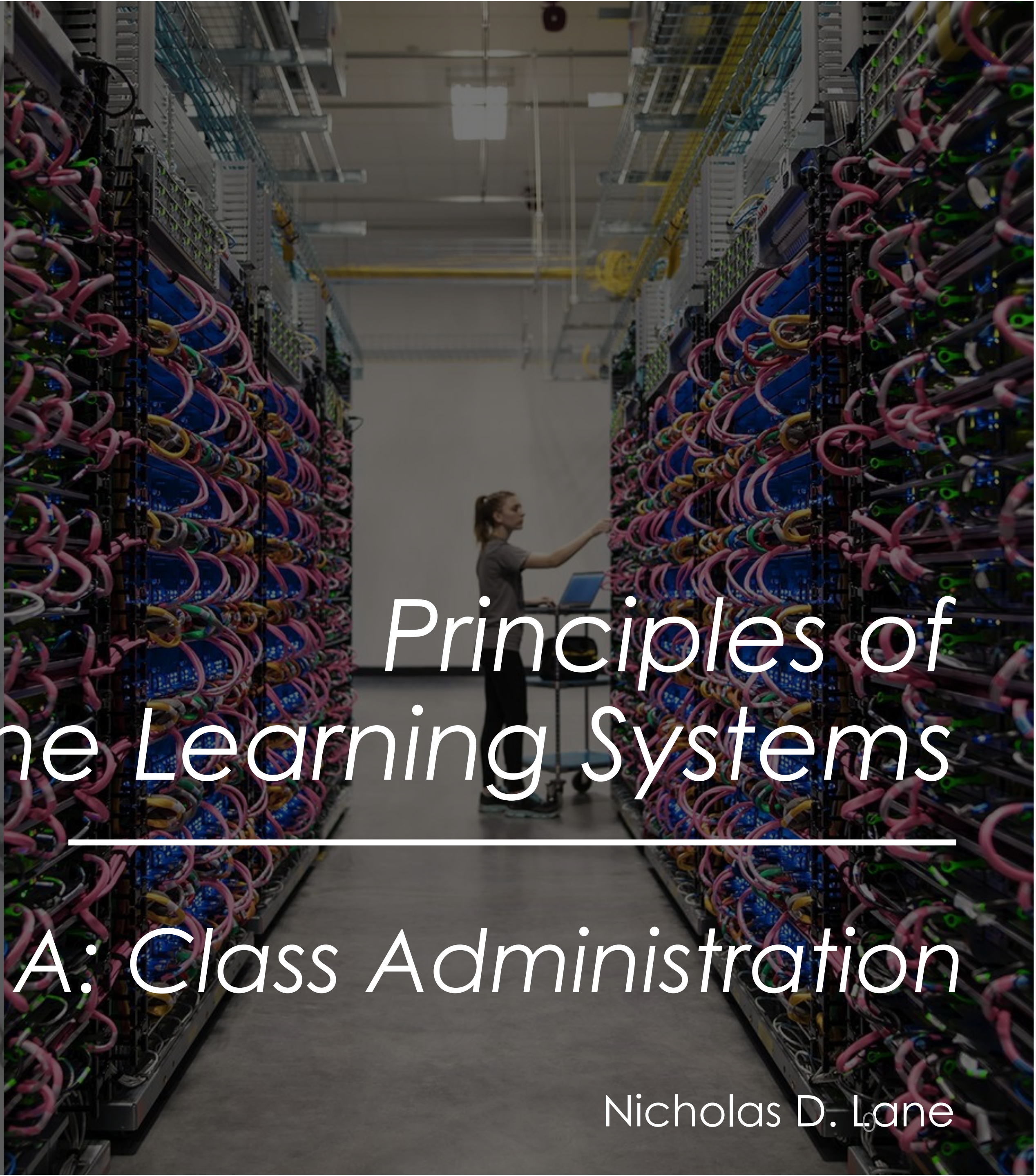




Principles of Machine Learning Systems

A: Class Administration

v1.6

Nicholas D. Lane

Principles of ML Systems

10 Lectures (FW26) covering:

- ML Systems Landscape
- Mapping to Hardware
- Model Compression
- Accelerators: GPUs, NPUs
- Frameworks and Run-times
- Single/Multi GPU Training
- Scalable Inference Serving
- Deep Learning Compilers
- Automated ML
- Federated Learning
- Development Practices
- MLOps related

Principles of ML Systems

3 Labs (SW02), covering:

- MCU and Model Compression: Speech Recognition
  - Lab scheduled for Oct 20th -- Submit material: Nov 2nd
- Single/Multi-GPU Training
  - Lab scheduled for Nov 10th -- Submit material: Nov 23rd
- Federated Learning: Experiments and Deployment
  - Lab scheduled for Nov 17th -- Submit material: Nov 30th

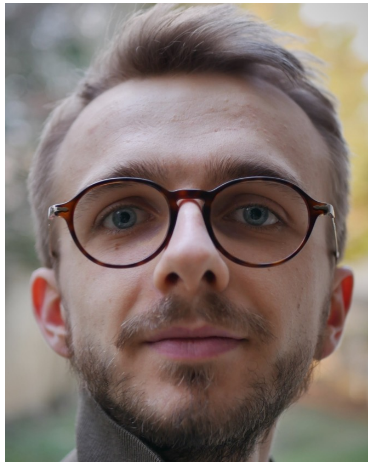
(each lab counts for 10% of your final grade; total is 30%)

- Moodle submission. zip file of Google Colab + txt file of Colab URL

Principles of ML Systems

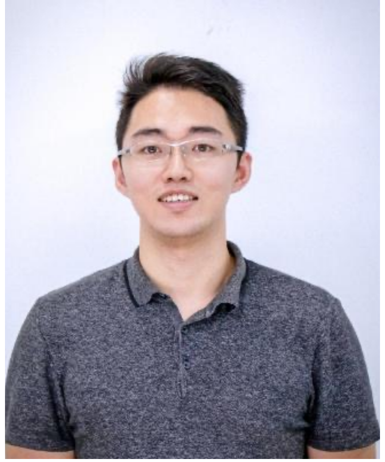
- Lectures vs. Labs
- Lectures recorded
  - Available on Moodle (maybe posted on YouTube); subtitles available
  - Past lectures also available (very similar content)
  - **You might be heard/seen in a recording**, especially if you ask questions
  - if uncomfortable with this feel free to approach us via an alt. channel
- TA sessions, and Office Hours
  - TA General+Project (weekly w/ Filip; 11am Wed mix of online/in-person)
  - TA Labs (upon appointment)
  - Office Hours (upon appointment)
- No required textbook: Lectures have optional reading lists

The Team



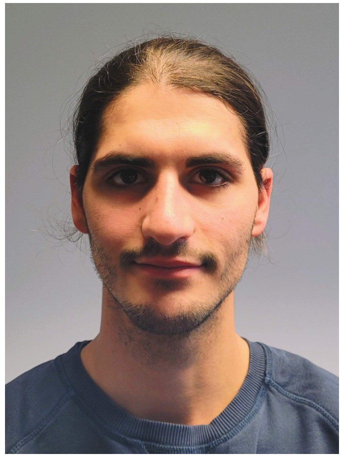
Titouan Parcollet

Filip Svoboda

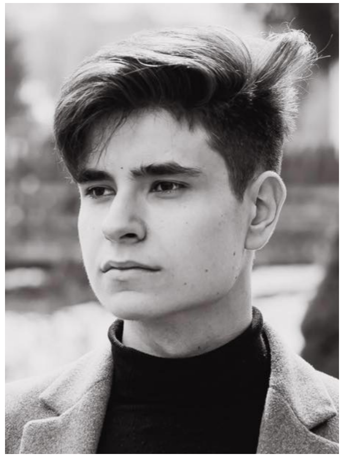


Hongxiang Fan

Nic Lane



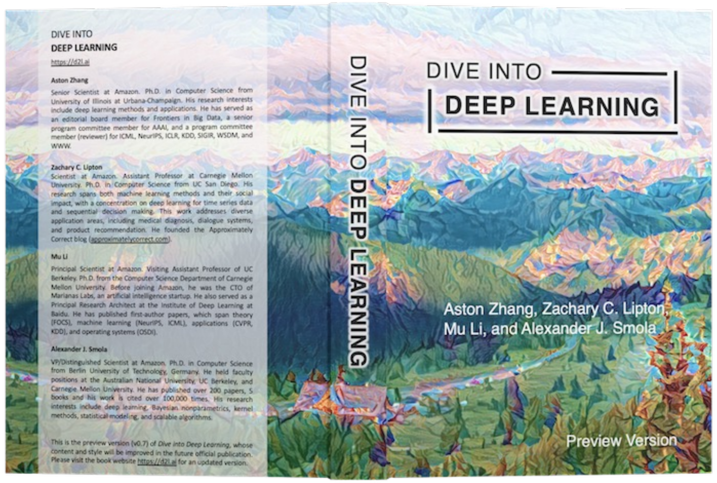
Lorenzo Sani



Alex Iacob

Pre-requisites

1. Comfortable with programming in Python
2. Undergraduate level: operating systems; computer architecture; machine learning
3. Familiarity with Deep Neural Networks (Intro Level)



DIVE INTO
DEEP LEARNING
https://d2l.ai

Aston Zhang

Senior Scientist at Amazon. Ph.D. in Computer Science from University of Illinois at Urbana-Champaign. His research interests include deep learning methods and applications. He has served as an editorial board member for Frontiers in Big Data, a senior program committee member for AAAI, and a program committee member (reviewer) for ICML, NeurIPS, ICLR, KDD, SIGIR, WSDM, and WWW.

Zachary C. Lipton

Scientist at Amazon. Assistant Professor at Carnegie Mellon University. Ph.D. in Computer Science from UC San Diego. His research spans both machine learning methods and their social impact, with a concentration on deep learning for time series data and sequential decision making. This work addresses diverse application areas, including medical diagnosis, dialogue systems, and product recommendation. He founded the Approximately Correct blog (approximatelycorrect.com).

Mu Li

Principal Scientist at Amazon. Visiting Assistant Professor of UC Berkeley. Ph.D. from the Computer Science Department of Carnegie Mellon University. Before joining Amazon, he was the CTO of Marianas Labs, an artificial intelligence startup. He also served as a Principal Research Architect at the Institute of Deep Learning at Baidu. He has published first-author papers, which span theory (FOCS), machine learning (NeurIPS, ICML), applications (CVPR, KDD), and operating systems (OSDI).

Alexander J. Smola

VP/Distinguished Scientist at Amazon. Ph.D. in Computer Science from Berlin University of Technology, Germany. He held faculty positions at the Australian National University, UC Berkeley, and Carnegie Mellon University. He has published over 200 papers, 5 books and his work is cited over 100,000 times. His research interests include deep learning, Bayesian nonparametrics, kernel methods, statistical modeling, and scalable algorithms.

This is the preview version (v0.7) of *Dive into Deep Learning*, whose content and style will be improved in the future official publication. Please visit the book website https://d2l.ai for an updated version.

DIVE INTO DEEP LEARNING
DIVE INTO
DEEP LEARNING
Aston Zhang, Zachary C. Lipton,
Mu Li, and Alexander J. Smola
Preview Version

Dive into Deep Learning
https://d2l.ai/

Assessment: Class Project

Primary assessment of the class (70%)

- Individual or Teams (sorry but Part III and MPhil students can't be in the same group)
- Repository for code, decision planning, results and write up
- Repository access provided to assessors *(myself + team)*
- Primary output assessed: written report *(submitted by moodle)*
- Repository will used understand team contributions and process
- Assume your report will be public *(unless otherwise requested)*
- Report in format of NeurIPS conference paper *(8 pages)*
- Due: Start of Lent term *(16/1/24 at 12:00 noon)*

Assessment: Class Project

Example Projects *(select projects from last year will be posted)*

- Build an application – applying some ML Sys ideas
- Explore a new direction *(novelty is not a must)*
- Leverage interesting hardware or architecture
- Detail an investigation to examine a direction of interest
- Replicating paper results
- Replicating expected system behavior *(textbook etc.)*