Lecture 1 of the

MLArchSys Seminar

Instructor: Thaleia Dimitra Doudali

Assistant Professor at IMDEA Software Institute

Universidad Politécnica de Madrid (UPM)

March 2023



About the Instructor - Thaleia Dimitra Doudali



in Greece.

Undergrad in ECE at NTUA, Athens, Greece.

PhD in CS at Georgia Tech, Atlanta, USA.

Assistant Professor at IMDEA, Madrid, Spain.



https://thaleia-dimitradoudali.github.io/

Website

About IMDEA Software Institute

- Research (Ph.D. + internships).
- Collaborative Environment.
- Fun Activities.

Learn more:

https://software.imdea.org/news.html

Open Positions: https://software.imdea.org/open_positions.html



Poster Competition – June 2022



Easter Egg Hunt – April 2022



Soft Skills Workshop – May 2022



About My Research

Research at the intersection of Machine Learning and Computer Systems Software.



About This Seminar Series

Research at the intersection of Machine Learning and Computer Systems Software.



Each lecture will go over and expand upon a specific research paper.

Today's Lecture

A Berkeley View of Systems Challenges for Al

Today's paper:

Ion Stoica, Dawn Song, Raluca Ada Popa, David Patterson, Michael W. Mahoney, Randy Katz, Anthony D. Joseph, Michael Jordan, Joseph M. Hellerstein, Joseph Gonzalez, Ken Goldberg, Ali Ghodsi, David Culler, Pieter Abbeel*

Systems	Systems <i>for</i> ML	Machine
Software		Learning

- What is behind the recent success of AI / ML?
- What are some trends in AI?
- What challenges do they create?
- What system support we need for AI?



Kleio: A Hybrid Memory Page Scheduler with Machine Intelligence

Thaleia Dimitra Doudali Georgia Institute of Technology thdoudali@gatech.edu Sergey Blagodurov Advanced Micro Devices, Inc. Sergey.Blagodurov@amd.com

Sudhanva Gurumurthi Advanced Micro Devices, Inc. Sudhanva.Gurumurthi@amd.com s, Inc. Advanced Micro Devices, Inc. Leom Abhinav.Vishnu@amd.com

Abhinav Vishnu

Ada Gavrilovska Georgia Institute of Technology ada@cc.gatech.edu



for Memory Management

LSTMs

Next Lectures	Systems Software	ML <i>for</i> Systems	Machine Learning
---------------	---------------------	-----------------------	---------------------



Structure of lectures: **23**

- Overview of the problem (cache prefetching, memory management).
- Overview of existing non-ML solutions.
- Overview of the ML-based method and how it's used to solve the problem.

No background needed!



Grades

- Seminar is worth 1 ECTS.
- Material per lecture: paper + slides.
- Grade = 20% 40% 40% per report after class.
- 1 report = Answer to few Questions.
- DUE after 1 week, before the next lecture.

Contact

• Via email: thaleia.doudali@imdea.org



https://thaleia-dimitradoudali.github.io/

Website

59' Madri	1 d	And	Alle Alle	S A	Conteres of the second	UNIVERSIDA UNIVERSIDA CAMP	D POLITÉC	Supervision of the second seco	MCORMATICOS
(Aluche))			JA		POZ	UELO DE A	ALARCÓN	
		15 min					B1	Tiempo estimado de recor	rido
140920		d	Horar e la ET	rios SI I	de salid nformáti	a cos			
		(Vige	ente de 1 o	de se	ptiembre a 3	30 de junic) - Día	s lectivos	
		De	10:05	а	13:35	cada	30	minutos	
				Α	13:55	14:	10		
		De	14:35	а	18:05	cada	30	minutos	
Lunes a		De	18:45	а	21:00	cada	45	minutos	

Teaching

🗰 Spring 2023

MLArchSys Seminar Series.

At the MUSS and EMSE Master Programs of the School of Computer Science at Universidad Politécnica de Madrid. 🖉 MUSS Link & EMSE Link

Seminar 1: Introduction to Maching Learning for Computer Architecture and Systems. 🐺 Slides ± Paper

Seminar 2: Maching Learning for Cache Prefetching. 🐺 Slides 👱 Paper

Seminar 3: Maching Learning for Hybrid Memory Management. 🗮 Slides ± Paper

Artificial Intelligence is Everywhere





Chatbot - Personal Assistants



Nurse Education in Practice Volume 66, January 2023, 103537 Rerse Boardion V Procise P

Editorial

Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?

Siobhan O'Connor * 옷 ¹ 회	, ChatGPT ^b i	
Show more 🗸		

+ Add to Mendeley 😪 Share 📑 Cite

https://doi.org/10.1016/j.nepr.2022.103537

Get rights and content







What drives the success of AI? **NEW HARDWARE**



Accelerators even in Laptops.





Supercomputers





Google Cloud

Datacenters all over the world.

What drives the success of AI? **NEW SOFTWARE**





Software for Big Data Processing

Software for Machine Learning Pipelines

The AI Landscape

THE 2023 MAD	(MACHINE LEARNING	ARTIFICIAL INTELLIG	ENCE & DATA) LANDSCAPE
---------------------	-------------------	----------------------------	------------------------

INFRASTRUCTURE	ANALYTICS	MACHINE LEARNING & ARTIFICIAL INTELLIGENCE	APPLICATIONS – ENTERPRISE			
STORAGE MPP DBs DMALAKES/ DAGE double for the formation of the formation	BIPLATFORMS Subset of the subset of the sub	DATA SCIENCE DATA SCIENCE PLATFORMS Proteir CONSTRUCTION Proteir CONSTRU	SALES MARKETING USTORE REPERENCE USTORE			
Analysee Antonymer Excelling and antonymer Excelling Excelling a	DATA ANALYST PLATFORMS Mode III Microsoft alteryx J pyromid III Concerner II Parton GUMUS # Antake Discourse II Parton	DATA GENERATION ALABELING SCOLE () HIVE yourd' againstance:	LIGGAL PARTNERSHIPS CONSISTENT CO			
Brogess Wankinge Contract Program (Contract Program) Contract Program (Co	anodet → test * 1 ≠ A < E	Stand Control	CODE & APPLICATIONS - HORIZONIAL CODE A CODE			
DATA TRANSFORMATION ⊕ Census € CATALOG	ANALYTICS Amage: Amage: Amage		Image: Lot Moderne III works Image: Lot Moderne IIIIIIII works Image: Lot Moderne IIII works<			
*torr Ministration Chemical Riftwary @ southerer Chemical Anouncouncil Security Creation Castor Domyst Ensure @ southerer Chemical Anouncouncil Castor Domyst Ensure @ southerer Chemical Anouncouncil Anouncouncil Castor Domyst Ensure @ southerer Chemical Anouncouncil Anouncouncil Security Control Anouncouncil @ data Prestaver Chemical Prestaver Anouncouncil Anouncouncouncil Anounc	If Heap of advant Sorecoor & exabear fullstory exases.rt Sorecoor & exabear detable AppDynamics detable Luabase		APPLICATIONS - INDUSTRY			
ATTENDATE Parter Ganera Control Color Parter Ganera Control Color ATTENDATE Parter Ganera Control Color Control Color	Image: Source of the source	Ormanication Construction	Ansmalk Grouternal Control Landow Waternal Ansmalk Grouternal Ansmalk			

											OPEN	SOURCE INFRASTRUCTU						
											OTEN	Sooner Introducto						
FRAMEWORKS	FORMAT -	QUERY / DATA FLOW -	DATA -	DATABASES		OLAP 7	ORCHESTRATION -	INFRA-	DATA OPS	STREAMING &	STAT TOOLS &	MLOPS & AI INFRA	AI FRAMEWORKS & LIBRARIES	AI MODELS &	SEARCH	LOGGING & MONITORING -	VISUALIZATION	COLLABORATION
The sport of the Flink	ICEBERO	tr 34 (2)	ACCESS	P man	🎙 MongoDB. 😂 redis	CiciHouse	Reache Airflow	STRUCTURE	M MARQUEZ	MESSAGING	LANGUAGES 🐣 python"	miflow scantow	TensorFlow Optorch Keras fast at OperCV . fearn Webart	ARCHITECTURES	Sol	Selastic Kibana A serery	3 co Superset matpl-tlib	1
	@ Parquel	Spark y	Amundaan	CockroachDB	Anne () influxdata'	epinot	PREFECT	1	g prest, expectations	Spak Records	R Scala Chamby (Spondos	ADV -	Autoria Crytech anter astal	TINIDIA Google Research Bort Bort Card Congentation Congent Towney	area a		0 1	BeakerX Pupyter
	0	presto 📎 🖉 trino 🏟 overal	(FE)		🛔 💇 🛑 oriak	3	ANTITATION	Ø	Open (1990)	Flink 2 STORM BRabbitMQ	(SoPy (Soudio 🕂 julia		PyTorch Lightning PyTorch Lightning アンデン ビーズ のxnet Chainer の	O Desching Company RoBERTa DistilBERT	III matternation 12 com	logstash Oronerous Provention	re dash TensorDoard	
CDRP Cocker	anne Mb	Arries Davis Character	🛔 magda .		ecoumulo . Milvus	-	11.91		tokeFS	Apacha Rocketti) 🌀 Samza 🖬 👬 debezium	Como			XLNet BLEEM Google Al (RIFFUSION) COMeto Al COMeto Al	M HANDOURI JANK	Muenta CGrafana Wetten	Queine bakeb	Apache Zeppelin
HELIX @dask -RAY	COLTA LAKE	Dest S Dest & Hawd	Wckan	ksqlDB	ttot os Fabrist	Ball sevens	ante Aleren Lines	B MESOS	Project Nesals	restantive / Cause 🔊 mortphis 🏟 hyteres	Cano occume into into a	Odeepstecks, Stangthan	EBUILD HOROVOD Corrently Spork H200 WSINGA	CoMetgAl Google Research Drama	Search 🖛 tantivy	p	Gradoun Doken (() () () () () () () () () ()
	~			1 Avadion		(bill money	Lob Women L	V									1	

DATA SOURCES & APIs	DATA & AI CONSULTING				
	Quantum Hack BGOX Deloitte. IBM ALCONSUMING Communication LeewayHertz Scalorn Image: Communication of the communication of th				

3 Interactive version: MAD.firstmarkcap.com

Trends and Challenges in Al



The trends in AI create challenges and opportunities for new systems.



Autonomous Driving

CHALLENGES

- Quick response.
- Unexpected conditions.
- Continually adapt.
- Learn new skills.



Challenges: Design AI systems that learn continually by interacting with a dynamic environment, while making decisions that are timely, robust, and secure.



Personal Assistants

Challenges: Design AI systems that enable personalized applications and services, yet do not compromise users' privacy and security.

CMS Allows Orgs to Share and Sell Medicare, Private Claims Data

CMS has finalized a healthcare reform rule that would allow qualified entities to share and sell Medicare and private payer claims data and analyses.



CHALLENGES

- Share data to solve a common problem.
- Protect private data.

Challenges: Design AI systems that can train on datasets owned by different organizations without compromising their confidentiality.



Al across Organizations





Google Cloud

CHALLENGES

- Massive amounts of data.
- Need for new hardware/software solutions customized to AI needs.

Al-specific architectures

Challenges: Develop custom hardware and software solutions, to address the performance and storage needs of future AI applications.

Trends and Challenges in Al



The trends in AI create challenges and opportunities for new systems.





Acting in Dynamic Environments

R1: Continual Learning

Autonomous Driving

Learns from interaction with environment

Reinforcement Learning

Neural Networks

Systems Research: Build systems that can faithfully simulate the real-world environment, as the environment changes continually and unexpectedly, and run faster than real time.

Quick response.

Continually adapt.

Learn new skills.

•

Unexpected conditions.

Acting in Dynamic Environments

R2: Robust Decisions

Microsoft chatbot is taught to swear on Twitter

() 24 March 2016

<

The AI was taught to talk like a teenager

By Jane Wakefield Technology reporter

A chatbot developed by Microsoft has gone rogue on Twitter, swearing and making racist remarks and inflammatory political statements.

Systems Research: Build systems to detect the source of data, protect against noisy or even malicious data.

Acting in Dynamic Environments

R3: Explainable Decisions

Medical Diagnosis

- Which X-ray feature led to the diagnosis?
- Explanation meanigful to humans.
- Medical diagnosis may raise legal issues.

Explainable AI Tools

Systems Research: Build systems that record and faithfully replay the computations that led to a particular decision (diagnostics).

R4: Secure Enclaves

Google Cloud

App App Operating System Hypervisor CPU Host

No control "where" the code runs.

Enclave: a secure execution environment, usually enforced by hardware.

Systems Research: Build systems that use enclaves to ensure data confidentiality, user privacy and decision integrity.

Secure Al

R5: Adversarial Learning

Data Poisoning during training.

Evasion attack during prediction.

Systems Research: Build systems that are robust against adversarial inputs both during training and prediction (e.g., decision making).

Secure Al

R6: Shared Learning on Confidential Data

CMS Allows Orgs to Share and Sell Medicare, Private Claims Data

CMS has finalized a healthcare reform rule that would allow qualified entities to share and sell Medicare and private payer claims data and analyses.

Share data across hospitals to identify and predict epidemics.

Systems Research: Build systems that can learn across multiple data sources, protecting private data and user confidentiality.

AI-specific architectures

R7: Domain Specific Hardware

Systems Research: Build Systems that efficiently use new hardware technologies, like accelerators, new types of memory and storage devices.

R8: Composable AI Systems

Systems Research: Build systems that allow the composition of models and actions in a modular and flexible manner.

Al-specific architectures

R9: Cloud-edge Systems

Cloud Computing vs Edge Computing

- Smaller devices.
- Different, specialized hardware.
- Less storage.
- Closer to data generation.

Systems Research: Build systems that (1) leverage the edge to reduce latency and (2) leverage the cloud to share data and models across edge devices.

Read in 3 passes:

- 1st pass [10 mins]: Quick pass. Read abstract, introduction, conclusion.
- 2nd pass [1 hour]: Read the full paper to understand the problem and the solution.
- 3rd pass [x hours]: Read again and challenge the choises. Is it well motivated? Well designed? Well evaluated?

Resources:

- <u>http://svr-sk818-web.cl.cam.ac.uk/keshav/papers/07/paper-reading.pdf</u>
- <u>https://sosp19.rcs.uwaterloo.ca/diversity/slides/rebecca.pdf</u>
- http://muratbuffalo.blogspot.com/2013/07/how-i-read-research-paper.html

Report Due March 21 at 18.00

Send report via email at: thaleia.doudali@imdea.org

Answer / expand upon these 4 questions:

- 1. What drives the recent success of AI / ML?
- 2. Which 1 of the 4 trends in AI you find most important to you and why?
- 3. If you had to solve 1 of the challenges, which one you would choose and why?
- 4. What are 2 things you will remember from this paper?

Your Answers

TRENDS

If you had to solve 1 of the challenges, which one you would choose and why?

CHALLENGES

