Benchmarks and Process Management in Data Science: Will We Ever Get Over the Mess?

Usama M. Fayyad
Open Insights
San Francisco, CA
usama@open-insights.com

Arno Candel
H2O.ai, Inc.
Mountain View, CA
arno@h2o.ai

Eduardo Ariño de la Rubia
Domino Data Lab
San Francisco, CA
eduardo@dominodatalab.com

Szilárd Pafka
Epoch
Santa Monica, CA
szilard.pafka@epoch.com

Anthony Chong
IKASI
Palo Alto, CA
anthony.chong@ikasi.ai

Jeong-Yoon Lee
Microsoft
Los Angeles, CA
jeongyoon.lee@microsoft.com

ABSTRACT

This panel aims to address areas that are widely acknowledged to be of critical importance to the success of Data Science projects and to the healthy growth of KDD/Data Science as a field of scientific research. However, despite this acknowledgement of their criticality, these areas receive insufficient attention in the major conferences in the field. Furthermore, there is a lack of actual actions and tools to address these areas in actual practice. These areas are summarized as follows:

1. Ask any data scientist or machine learning practitioner what they spend the majority of their time working on, and you will most likely get an answer that indicates that 80% to 90% of their time is spent on "Data Chasing", "Data Sourcing", "Data Wrangling", "Data Cleaning" and generally what researchers would refer to—often dismissively—as "Data Preparation". The process of producing statistical or data mining models from data is typically "messy" and certainly lacks management tools to help manage, replicate, reconstruct, and capture all the knowledge that goes in 90% of activities of a Data Scientists. The intensive Data Engineering work that goes into exploring and determining the representation of problem and the significant amount of "data cleaning" that ensues creates a plethora of extracts, files, and many artifacts that are only meaningful to the data scientist.
2. The severe lack of Benchmarks in the field, especially ones at big data scale is an impediment to true, objective, measurable progress on performance. The results of each paper are highly dependent on the large degree of freedom an author has on configuring competitive models and on determining which data sets to use (often Data that is not available to others to replicate results on)
3. Monitoring the health of models in production, and deploying models into production environments efficiently and effectively is a black art and often an ignored area. Many models are effectively "orphans" with no means of getting appropriate health monitoring. The task of deploying a built model to production is frequently beyond the capabilities of a Data Scientists and the understanding of the IT team.

For a typical company, a Machine Learning or Data Science expert is a major investment; yet these people are in such hot demand, that likelihood of churn is high. Typically, when a data scientist is replaced, the process pretty much starts over with a tabula rasa… In fact, I would argue most data scientists coming back to tasks they built themselves 1-2 years before are unable to reconstruct what they did.

For this panel, we have selected a unique set of experts who have different experiences and perspectives on these important problems and how they should be dealt with in real environments. It is our hope that the panel discussion will not only produce recommendations on what to do about these painful impediments to successful project deployments, but also serve as an eye opener for the research community to the importance of paying close attention to issues of Data and Model Management in KDD, as well the need to think carefully about the lifecycle of models and how they can be managed, maintained, and deployed systematically.

Without addressing these critical deployment and practice issues, our field will be challenged to grow in a healthy and sustainable way. The expert panelists for this panel along with the panel moderator: Usama Fayyad are listed below along with their biographical sketches.

CCS CONCEPTS

• Information systems→Information systems applications→Data mining
• Computing methodologies→Supervised learning by classication;
• Computing methodologies→Data and Model Management
• Social and professional topics→Professional topics→Management of computing and information systems→File systems management
• Information systems→Data management systems→Information integration

KEYWORDS

Performance benchmarks; software implementations; training speed; memory footprint; accuracy, model management, model deployment and monitoring, Data benchmarks

KDD '17, August 13-17, 2017, Halifax, NS, Canada

ACM ISBN 978-1-4503-4887-4/17/08
https://doi.org/10.1145/3097983.3120998

Panelist: Arno Candel

Dr. Arno Candel is the Chief Technology Officer at H2O.ai, the makers of the distributed and scalable open-source machine-learning platform H2O. Arno is also the main author of H2O's Deep Learning and a key contributor to H2O's GBM and DRF algorithms, and recently has been working on GPU algorithms and Driverless AI. Arno spent the last 5+ years designing and implementing high-performance machine-learning algorithms. Previously, he spent a decade in high-performance computing and ran his code on the world's largest supercomputers as a staff scientist at SLAC National Accelerator Laboratory, where he participated in US DOE scientific computing initiatives and collaborated with CERN on next-generation particle accelerators.

Arno holds a PhD and Masters summa cum laude in Physics from ETH Zurich, Switzerland. He has authored dozens of scientific papers and is a sought-after conference speaker. Arno was named "2014 Big Data All-Star" by Fortune Magazine and featured by ETH GLOBE in 2015. Follow him on Twitter: @ArnoCandel.

Panelist: Eduardo Ariño de la Rubia

Eduardo Arino de la Rubia is chief data scientist at Domino Data Lab. Eduardo is a lifelong technologist with a passion for data science who thrives on effectively communicating data-driven insights throughout an organization. He is a graduate of the MTSU Computer Science department, completed graduate studies in negotiation, conflict resolution, and peacebuilding from CSUDH, the General Assembly's Data Science program, and the Johns Hopkins Coursera Data Science specialization. You can follow him on Twitter as @earino.

Panelist: Szilárd Pafka

Szilard studied Physics in the 90s and has obtained a PhD by using statistical methods to analyze the risk of financial portfolios. Next, he has worked in a bank quantifying and managing market risk. About a decade ago he moved to California to become the Chief Scientist of a credit card processing company doing everything data (analysis, modeling, data visualization, machine learning, data infrastructure etc). He is also the founder/organizer of several meetups in Los Angeles (R, data science etc) and the data science community website datascience.la. He is teaching data science and machine learning in graduate programs at CEU (Europe) and UCLA (California).

Panelist: Anthony Chong

Anthony is the founder/CEO of IKASI. Prior to founding IKASI he helped start the social media advertising company Adaptly, where he was the first employee. Anthony was the Head of Optimization (Data Science), where he was responsible for building their Data Science team, tackling problems ranging from automated ad content and bidding optimization to sentiment analysis. He helped grow the company from three people in an apartment to over 80 employees. Anthony received his BS from the California Institute of Technology, in Computer Science. Anthony serves on the Board of Directors of the Caltech Alumni Association.

Panelist: Jeong-Yoon Lee

Jeong is Technical Evangelist at Microsoft, where he promotes and helps adoption of new Machine Learning technologies among enterprises and communities. He is also Science Advisor at Conversion Logic, where he served as Chief Data Scientist and developed marketing analytics platforms with cutting edge ML models for enterprise clients. Prior to Conversion Logic, Jeong was Lead Applied Science Engineer at Demand Media, where he deployed Machine Learning pipelines for titling, pricing and deduplication algorithms for eHow.com and eNom.com. Jeong was also Co-Founder of Neofect, a smart rehabilitation solution company, and Micro ML, a Machine Learning solution provider for embedded IoT devices.

As an avid competitor, Jeong has participated in over 70 Data Science competitions, won 6 times including KDD Cup 2012 and 2015, finished Top 10 8 times including the Deloitte, AARP, Criteo competitions, and was ranked Top 10 at Kaggle in 2015.

Jeong earned his Ph.D. in Computer Science and M.S. in Electrical Engineering from University of Southern California. He earned his B.S. in Electrical Engineering from Seoul National University.

MODERATOR: Usama Fayyad

Usama is CEO of Open Insights in Silicon Valley which he reactivated after leaving his position as Chief Data Officer and Group Managing Director at Barclays in London (2013-2016). He is served as Interim CTO for Stella.AI a Mountain View, CA and is acting Chief Operations & Technology Officer for MTN's new division: MTN2.0 aiming to extend Africa's largest telco into new revenue streams beyond Voice & Data.

In 2010 Usama was appointed as founding Executive Chairman of OASIS-500 by King Abdullah II of Jordan to build a tech startup investment fund and accelerator in Jordan and MENA Region. He was also Chairman, Co-Founder and CTO of Blue Kangaroo, a mobile search engine service for offers based in Silicon Valley 2011-2013. In 2008, Usama founded Open Insights, a US-based data strategy, technology and consulting firm that helps enterprises deploy data-driven solutions that effectively and dramatically grow revenue and competitive advantage.

Usama became the world's first Chief Data Officer at Yahoo!'s when he served as CDO & EVP (2004-2008) after Yahoo! acquired his second startup: DMX Group, a data mining and data strategy consulting and technology company specializing in Big Data Analytics and Data Science for Fortune 500 clients. In 2000 Usama left his leadership position at Microsoft to co-found and serve as Chairman and CEO digiMine (Audience Science). At Microsoft, he spent 5 years leading the data mining and exploration group at Microsoft Research and headed the data mining products group for Microsoft's server division.

Usama held a leadership role at NASA's Jet Propulsion Laboratory (1989-1996) where his work garnered him the Lew Allen Award for Excellence in Research from Caltech, as well as a U.S. Government medal from NASA.

Usama has published over 100 technical articles on data mining, Artificial Intelligence, machine learning, and databases. He holds over 30 patents, is a Fellow of the Association for Advancement of Artificial Intelligence and a Fellow of the ACM. He has edited two influential books on data mining and served as editor-in-chief on two key industry journals. Usama earned his Ph.D. in engineering in AI/Machine Learning from the University of Michigan, Ann Arbor. He holds two BSE's in Engineering, MSE Computer Engineering and M.Sc. in Mathematics. He is active in the academic community with several adjunct professor posts and is the only person to receive both the ACM's SIGKDD Innovation Award (2007) and Service Award (2003). He serves on several private/public boards