kaggle

AI Report 2023

Essays and insights from the world's largest data science and machine learning community

Authors

Abir Eltaief
Ali Jalali
Antong C.
Anya Mathur
Arya Gaikwad
Bojan Tunguz
Christof Henkel
Chuandong Tang
D. Sculley
Danial Sultanov
Dariusz Kleczek
Dave Harold Mbiazi Njanda
Diego Flores
Dmitri Kalinin
Harshit Mishra
Hoda Jalali Najafabadi
Ivaxi Sheth
Julia Elliott
Karnika Kapoor
Kobbie Manrique
Leonie Monigatti
Lezhi Li
Lorresprz
Mark McDonald
Martin Henze
Maryam Babaei
Meghana Bhange
Nghi Huynh
Parul Pandey
Patrik Joslin Kenfack
Paul Mooney
Paulina Skorupska
Phil Culliton
Piyush Mathur
Pranav Mohan Belhekar
Raghav Awasthi
Rhys Cook
Rob Mulla
Samantha Lycett
Sanyam Bhutani
Shreya Mishra
Svetlana Nosova
Theo Flaus
Trushant Kalyanpur
Will Cukierski
Xinxi Chen
Yassine Motie
Yuqi Liu
Yuxi Li
Zhengping Zhou

TABLE OF CONTENTS

TABLE OF CONTENTS

ESSAYS *[continued]*

TABLE OF CONTENTS

ESSAYS *[continued]*

07

Other Topics

Introduction

Foreword
D. Sculley

The world of AI has seen breathtaking progress over recent years, with rapid advances in the capabilities of models as large as ChatGPT, Llama, and PaLM, and as small as those that can fit on device or in a web browser. The advances of AI have not been confined only to models: we have also seen an incredible spread of knowledge and expertise across the globe, with AI experts participating in the field from every corner of the world and every walk of life.

At Kaggle, we believe that this global community of AI and ML experts – now 15 million members strong – is one of the most valuable open resources in the world today. Our community works together to learn, share, compete, collaborate, stress test, and evaluate what really works in AI and ML, and does so in a deeply rigorous fashion.

It is a great pleasure to welcome you to the 2023 Kaggle AI Report, created by our community and selected from hundreds of submissions. Each paper within this report gives a unique viewpoint on the most interesting or most important recent developments in the field of AI and ML.

Enjoy the reading, and as always, our deepest thanks to the millions of members of the Kaggle community!

Very best wishes,
D. Sculley

About the Report

Phil Culliton

The Kaggle AI Report is a collection of essays written and submitted by the Kaggle community as part of a competition, broken down into seven sections that we feel represent significant areas within the research and practice of modern ML. The submissions were evaluated and edited by a member of our community with noteworthy expertise in their section's area. Each expert selected the winner in their section, as well as a number of honorable mentions.

Sections

Section includes:

- **Generative AI:** A frontier of machine learning that is coming into new focus in the last few years, this area has combined the best of text and image research to create a fundamental shift in the usability and utility of machine learning models.
- **Text data:** Natural language processing and statistical language modeling are the backbone of some of the most exciting recent advances in AI.
- **Image / video data:** Image data was the foundation for early advances in deep learning and the past decade is a testament to the ingenuity of researchers and practitioners in this area, with ever-expanding datasets and problem formulations met with fresh, new ideas.
- **Tabular / time series data:** Tabular data problems are the most common type of problem to solve, and an area that has led to incredible research and development – some by our very own community members.
- **Kaggle competitions:** Kaggle is perhaps most famous for its competitions, curated and led with care by our team of experts, in partnership with hosts including researchers, educators, and industry giants. Competitions present cutting-edge problems and challenges that our diverse community solves in myriad exciting ways.
- **AI ethics:** Humans in the world of machine learning have always struggled to make our discoveries fair, unbiased, and equitable for other people.
- **Other:** The above six sections are extensive, but ultimately machine learning and artificial intelligence are oceans of possibility. We include a section for some fascinating essays that don't quite fit into any of the others.

The essays in our report were written as notebooks, a rich, multimedia communication form that can include text, images, video, and even runnable code. Please be sure to click on the link for each essay to fully explore the experience that the authors created.

Area Chairs

We created this report with the understanding that our community holds a sum of ability and knowledge far greater than our small team of editors. We reached out to prominent members of our community with backgrounds and proven skill in each of the report's 7 topic areas to act as judges and expert editors. These members of the community are our Area Chairs.

Sanyam Bhutani

Sanyam Bhutani drinks chai and makes content for the community at H2O.ai. When not drinking chai, you can find him hiking the Himalayas often with LLM papers. He is best known for chai, GPUs and mountains. He is the host of the most popular Kaggle podcast with top Kaggle Grandmaster interviews on Chai Time Data Science. On the internet, he's known for learning in public and "maximizing compute per cubic inch of ATX".

Christof Henkel

Christof is a Deep Learning Researcher at NVIDIA. He is particularly interested in novel deep learning architectures with respect to graphs, computer vision and audio. His background is in mathematics and he completed his PhD at the Ludwig-Maximilians-University in Munich about stochastic processes and financial markets. He started Kaggling six years ago and after participating in more than 70 competitions currently holds first rank in the world-wide competition ranking.

Martin Henze

Martin has a Ph.D. in astrophysics and an academic background in researching exploding stars in nearby galaxies. He got into the fields of data science and machine learning through Kaggle, where he became the first ever Kaggle Notebooks Grandmaster and for a time held the number one spot in the notebooks rankings. He has spoken at various conferences about his passion for effective data communication and storytelling, and he curated 100 episodes of the Hidden Gems series on underrated Kaggle Notebooks (example post and dataset). Martin is the Lead Data Scientist for the market research company YipitData.

Karnika Kapoor

Karnika Kapoor is a data scientist with a background in mechanical engineering. She is skilled in machine learning and AI, and is currently employed as a senior data scientist where she focuses on growth optimization and logistics. She is actively engaged with the Kaggle community and became a Kaggle Notebooks Grandmaster in part due to her ability to transform data into well-communicated insights. With a passion for learning, Karnika stays AI-current, offering a unique blend of technical prowess and data-driven acumen.

Rob Mulla

Rob Mulla has over 10 years experience working with data, using his skills at companies in sectors including pharmaceuticals, hospitality, energy, and sports sciences. His journey with data has led him to become a 4x Kaggle Grandmaster, where he has had the chance to participate alongside some of the best in the field. Outside of work, Rob enjoys sharing his knowledge of Python and data science on his YouTube channel, @robmulla, which has garnered a community of over 100,000 subscribers. Rob is an alumnus of Virginia Tech with a BS, Kansas State University with an MSE, and UC Berkeley where he earned his Masters in information and data science.

Parul Pandey

Parul Pandey has a background in electrical engineering and currently works as a Principal Data Scientist at H2O.ai. She is also a Kaggle Grandmaster in the notebooks category, where she composes compelling stories through the medium of notebooks. Her strength lies in analyzing data and eliciting useful insights from them with the help of powerful visuals. Parul is one of the co-authors of the Machine Learning for High-Risk Applications textbook which focuses on the responsible implementation of AI. Parul has written multiple articles focused on data science and she mentors, speaks, and delivers workshops on topics related to responsible AI.

Bojan Tunguz

Bojan is a Machine Learning Modeler at NVIDIA. Bojan holds a Ph.D. in theoretical physics,
and has been working in machine learning and data science fields for eight years, where he has worked with real-world fintech problems. He is one of seven 4x Kaggle Grandmasters, and is the first person to be ranked in the top 10 in all four Kaggle categories simultaneously. He has nine Kaggle competition gold medals, and with his team, won the largest Kaggle competition at the time – the Home Credit Classification challenge. He has worked as a competition participant and organizer in the Human Protein Classification challenge and the mRNA Open Vaccine challenge and has been part of three distinct Nature publications (#1, #2, #3). Bojan considers himself a machine learning generalist, and has considerable experience working with a wide variety of Machine Learning problems - image, NLP, voice, tabular, etc. He is passionate about learning and exploring new applications of machine learning, especially those that can have a significant impact on real world problems. He is currently working on the application of machine learning in RNA bioinformatics and in social media analysis. He is particularly interested in fundamental issues of machine learning for tabular data.
Bojan is a voracious reader, he is passionate about tinkering with all sorts of tools and gadgets, loves digital photography, and really enjoys hiking in the woods and biking.

Bojan is happily married, and lives with his wife and three little boys in Florida.

Essays

01 Generative AI

Section Overview by Karnika Kapoor

Topic Summary

This section delves into the dynamic world of generative AI, highlighting its rapid progress, particularly over the last two years. It unveils core advancements and transformative applications across various media forms. Generative AI focuses on creating new content such as images, text, and music, driven by breakthroughs in generative adversarial networks (GANs) and large language models (LLMs).

GANs, complex neural networks for creating realistic data, and LLMs, skilled in text and language generation have been central to recent advances in the ability of AI models to produce convincingly realistic outputs.

Enhanced generative modeling algorithms, like diffusion models and normalizing flows, coupled with refined training methods, further elevate this transformation by leveraging larger datasets, more powerful hardware, and innovative applications like medical image synthesis and financial model crafting, the field is propelled to new heights.

Despite being in its early stages, generative AI holds the potential to revolutionize many fields, its versatility showcased by applications as diverse as writing, music, and data production. Looking ahead, generative AI promises even more captivating advancements, influencing various industries and warranting a comprehensive understanding of its ethics and risk management.

Trends & Predictions

The evolution of generative AI is a captivating and significant topic deserving in-depth exploration. This technology is rapidly transforming how we create and interact with content, with the potential to revolutionize numerous industries. One noteworthy trend in generative AI is its expanding use beyond content generation. For instance, AI-generated imagery is being integrated in healthcare diagnostics and personalized educational experiences.

As generative AI becomes increasingly sophisticated, it's crucial to consider the potential risks and ethical implications of its usage. These risks include creating misinformation, perpetuating bias, vulnerability to disclosing private data if misused in a training corpus, inappropriate or unattributed use of copyright or artistic intellectual property, etc. Continued research and development of generative AI should be conducted in an ethical and responsible manner.

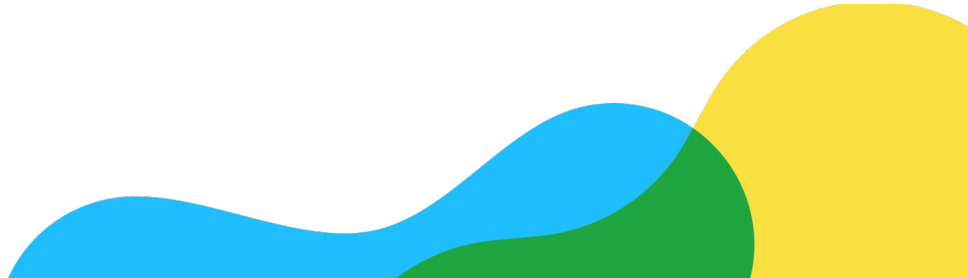
Overview of Essays

This collection of essays provides a concise history of AI's evolution, followed by an in-depth exploration of current developments in generative AI. The consensus among many essay authors is that the release of the 'Attention is All You Need' paper in 2017 marked a pivotal moment in the advancement of generative AI, with transformer architectures serving as the foundation for sophisticated models with billions to hundreds of billions of parameters (such as GPT-4). Architectural complexity and rigorous research play a central role in enhancing generative AI's performance.

Several essays delve into user-friendly generative AI solutions, catering to individuals seeking streamlined content creation.

Ethical considerations in AI, particularly related to energy consumption, are also extensively discussed.

Opinions differ on AI's impact on employment opportunities, with some embracing its transformative potential and others expressing concerns about job displacement. This duality mirrors historical patterns of technological shifts, where initial apprehensions gradually transform into productive integration. The overarching message emphasizes harnessing AI as a supportive tool rather than a disruptive force, aligning with historical trends that steer initial concerns towards productive applications.



Generative AI Essay #1

“2023 Kaggle AI Report - Generative AI”

Award Winner

By Trushant Kalyanpur

Spanning 2021 to 2023, this essay charts the profound evolution of generative AI, spotlighting strides in image synthesis, language models, and audio generation. Noteworthy innovations like GPT4, DALL-E, and ChatGPT take center stage, propelling AI-generated content into new realms. This narrative journey navigates pivotal years: 2021 witnesses DALL-E's text-to-image feats and Github Copilot's code suggestions, 2022 showcases Meta's contributions, ChatGPT's emergence, and 2023 highlights the rapid rise of multimodal AI with GPT4's unveiling.

The essay carefully addresses ethical concerns tied to AI's capabilities. Looking forward, it acknowledges ongoing research, tools like LangChain and AutoGPT, and the promise of innovative, ethically conscious AI integration.

In summary, this essay captures the dynamic trajectory of generative AI's evolution, showcasing significant achievements, addressing challenges, and envisioning a future where AI's potential is harnessed responsibly.

Link to Notebook

Generative AI Essay #2

“Understand, Generate and Transform the World”

Honorable Mention

By Yuqi Liu

This essay traces the evolution and growing influence of generative AI across many domains. With an adept blend of historical insight and technological exploration, the essay delves into the societal significance and intricate mechanisms of generative AI.

From its historical roots to the accelerated growth fueled by Deep Learning, the essay navigates pivotal models like DALL·E 2, highlighting the synergy between computational resources, dataset scale, and AI potential. Bridging Computer Vision and Natural Language Processing, the exploration showcases transformative structures such as GANs and Vision Transformers, offering a glimpse into the collaborative future of multimodal learning.

The essay candidly confronts challenges and ethical considerations, casting a balanced light on the practical applications and limitations of generative AI. Ultimately, the essay comprehensively celebrates generative AI's present impact while illuminating its promising trajectory into the future.

Link to Notebook

Generative AI Essay #3

“A Glimpse Into the Realm of Generative AI”

Honorable Mention

By Pranav Mohan Belhekar, Arya Gaikwad

This essay explores the innovative journey and paradigm shift that generative AI has inspired. From its resemblance to an enchanted data-powered box that generates patterns to its evolution from Boltzmann Machines to Generative Adversarial Networks (GANs), generative AI's impact on image synthesis, text generation, and more is discussed in detail. Recent achievements, including the fusion of transformers and GANs, showcase its collaborative potential in enhancing human creativity. Amid these strides, ethical considerations emerge, emphasizing impartial training data and addressing the societal implications of AI-generated content. The essay navigates the challenges, particularly deepfakes, while inviting us to witness a revolution that defines a new balance between human ingenuity and generative AI capabilities.

Link to Notebook

02

Text Data

Section Overview by Christof Henkel

Topic Summary

This section covers natural language processing (NLP) — understanding of natural language texts for classification and regression tasks, as well as text generation tasks such as summarization or translation. Insights can be gained from text data in a diverse number of languages including both common languages and programming languages. The methods can even be applied to data tables that are presented in text format.

When the entire internet is used as a training corpus for AI models, knowledge transfer using deep learning becomes an incredibly powerful technique – and the major advancements of the past few years are a testament to that.

Early approaches to small datasets often included term-frequency based feature engineering combined with non-neural-network-based machine learning methods. The rise of deep learning enabled solutions for larger datasets by learning word representations and applying machine learning models to interpret them. Deep learning also became state of the art for some small datasets through transfer learning as only a small amount of data was needed to fine-tune a pre-trained model to become relevant to new applications.

Similar to the benefits of scaling neural network architectures, transformer-based models can be trained on trillions of tokens and the resulting pre-trained model weights can either be used directly or fine-tuned to be state-of-the-art for a large and diverse number of tasks. Large language models (LLMs) are unsurprisingly the primary focus of attention for many or most AI researchers today.

Trends and Predictions

An observable trend in text-based Kaggle competitions is that instead of training a model from scratch, competitors tend to fine-tune publicly available models which have been pre-trained on a similar language or task. With text-based models becoming more and more accurate, simply because they are pre-trained on bigger datasets and hence carry more knowledge to transfer via fine-tuning, they often surpass human-level performance.

An interesting result of stronger text modeling is that they enable the ability to solve more challenging use cases.

This is evident in the rise of increasingly complex competitions hosted on Kaggle, ranging from analyzing Jupyter notebooks, to understanding and grading student essays, to problems where text data intersects with other data types like images or audio. With LLMs becoming more and more powerful, they will not only be the basis for solving text-based problems but also will be used to generate or augment training data. They might also be used to automatically judge generative models on complex tasks like abstract reasoning.

Overview of Essays

The top essays in this section provide a diverse and high-quality overview of contemporary LLMs. They notably skip over foundational NLP work that preceded LLMs, and similarly gave light treatment to the building blocks that enable their breakthrough performance on benchmarks and interactive use cases. Contributions focused primarily on the scientific merits of models, largely sidelining discussion of practicalities like hardware and the productionization of models. The focus on LLMs is not surprising, given they are currently in the center of society's attention when it comes to the applicability of artificial intelligence solving human problems.

Interestingly, the top essays cover different aspects of LLMs. The winning essay discusses the emergence of contemporary large language models as well as the most recent methods and techniques that make them possible. The other essays discuss fine-tuning and applicability of LLMs to small datasets, and the capabilities of LLMs with respect to the ability to perform reasoning. As such the top essays cover diverse and interesting topics of LLMs.

Text Data Essay #1

“Contemporary Large Language Models LLMs”

Award Winner

By Abir Eltaief

This essay explores the rise of contemporary language models, delving into the latest methodologies enabling them, and reflecting on the insights gained by the machine learning community in the past couple of years.

It also draws from the author's individual odyssey to learn the captivating realm of LLMs. The journey encompasses initial interactions with diverse GPT-based chatbots, the creation of applications fueled by LLMs, and the reading of remarkable research papers.

The essay discusses core concepts and features of LLMs. It examines the efficient utilization of extensive training data and parameters by pretrained LLMs. It further explores prompt engineering which spans from basic techniques like zero-shot, one-shot, and few-shot prompting to more advanced methods including Chain of Thought (CoT) and Reasoning & Acting (ReAct). The focus then shifts to enhancing LLMs through human feedback, exploring reinforcement learning-based fine-tuning (RLHF) for targeted optimization. The essay also investigates the augmentation of LLMs for app development, highlighting Retrieval Augmented Generation (RAG) as a framework to equip LLMs with external knowledge. The content concludes by providing referenced sources.

Link to Notebook

Text Data Essay #2

“Large Language Models: Reasoning Ability”

Honorable Mention

By Théo Flaus, Yassine Motie

This essay focuses on insights gained over the past two years of working and experimenting with LLM reasoning. After giving an overview of different types of reasoning, which are important to understand in tackling reasoning challenges, the authors explore the advancements made through Chain of Thought prompting, Tree of Thought frameworks, linguistic feedback reinforcement, interleaved reasoning and action, and handling complex mathematical reasoning.

By analyzing related papers and their findings, they provide a comprehensive overview of the architecture and progress of LLM reasoning. They put special emphasis on highlighting the lessons learned and the way forward in ensuring ethically safe and reliable AI systems.

Link to Notebook

Text Data Essay #3

“Mini-Giants: ‘Small’ Language Models”

Honorable Mention

By Zhengping Zhou, Lezhi Li, Xinxi Chen, Yuxi Li

With giant LLMs becoming expensive to train and prohibitively large to finetune for individuals or small companies, small language models are flourishing and becoming more and more competent. The authors call them "mini-giants" and argue a win-win for the open source community by focusing on small language models.

This essay presents a brief yet rich background, discusses how to attain small language models, presents a comparative study of small language models, and a brief discussion of evaluation methods. The authors discuss the application scenarios where small language models are most needed in the real world, and conclude with discussion and outlook.

Link to Notebook

03
Image / Video Data

Section Overview by Rob Mulla

Topic Summary

This section explores some of the latest advancements in computer vision, particularly as it relates to the use of image and video data. While the field of computer vision dates back to the 1960s, its evolution has been especially exciting over the last few decades. Specifically, over the past two years, there have been significant advancements not only in traditional computer vision tasks like classification and object detection, but also in emerging areas such as Vision Transformers (ViT) and few-shot learning.

In addition to discussing model architectures, this section covers standard practices used in computer vision today like preprocessing and augmentation techniques. Additionally, this section explores the practical applications of these technologies across industries like healthcare (for medical imaging), agriculture (for crop monitoring), and the automotive sector (for self-driving cars). It also covers some of the limitations that computer vision still faces.

Trends & Predictions

Computer vision can be traced back to the 1950s and 1960s, when researchers started the development of algorithms for detecting edges and patterns in images. While improvements in these areas continue, new challenges like object detection, self-supervised learning, and knowledge reasoning are being addressed with innovative model architectures and training techniques. Particularly, video data is witnessing developments in multi-object tracking, action recognition, and spatiotemporal reasoning. A noticeable area of active research in the field revolves around generalization and transformer-based architectures.

The popularity of models like Segment Anything Model (SAM) and YOLO (You Only Look Once) showcase how generalized, open source models can be leveraged and adapted to solve specific tasks. Looking ahead, I anticipate that within the next five years, tasks such as segmentation, object detection, and image classification will continue to be addressed by generalized large-scale models that can be fine-tuned for specific challenges, mirroring what we've seen occur with large language models. We will also see the intersection of computer vision and generative AI continue to progress in areas like augmented reality, deep fakes, and AI-enhanced photo and video editing. These advancements inevitably bring with them ethical and philosophical considerations that require our attention.

It's also important to note some of the limitations that research has yet to overcome. Some of these areas include the limitation of multi-modal models that incorporate image and video, as well as vision model's ability to perform in uncontrolled environments, like self-driving cars on new roads.

Overview of Essays

The essays chosen for this section cover the significant research advancements in image and video understanding in recent years. They reference breakthrough papers that have reshaped our understanding of machine learning and the types of tasks that computer vision is able to solve.

Dmitri Kalinin's essay, "Advances in AI Vision Models in the Last Two Years," is a comprehensive summary of recent advances in computer vision. It covers the recent progress in semantic segmentation, vision transformers, and continual learning methodologies, among others. It serves as a great resource for anyone desiring a deeper understanding of these subjects.

Lastly, Dron Bespilotnik's essay "Image and Video Data | Kaggle AI Report" covers some of the practical areas of working with image and video data, including preprocessing techniques, and discussing how data augmentation enhances the training performance of vision models. He describes the role of computer vision in tasks like image classification, segmentation, and question answering.

Image / Video Data Essay #1

“Advances in AI Vision Models in the Last Two Years”

Award Winner

By Dmitri Kalinin

Dmitri's essay summarizes the most recent advancements in computer vision models, spotlighting six pivotal areas: Semantic Segmentation, Vision Transformers, Few- and Zero-shot Learning, Generalization of Computer Vision Models, Continual Learning, and Human-Assisted AI in Computer Vision. While not all of these areas are necessarily new, Dmitri's essay emphasizes the latest studies, showing the newest evolution in computer vision.

Semantic Segmentation models are highlighted for their role in both autonomous driving and medical imaging sectors, and an emphasis is put on the emergence of novel models, unique loss functions, and the surge in weakly supervised strategies for large image dataset processing. Vision Transformers have emerged as a novel approach, providing breakthroughs in efficiency through image decomposition.

Few- and Zero-shot learning involves leveraging large language models with methods like label-free parameter tuning and feature extraction adapters. The Contrastive Language–Image Pre-training (CLIP) architecture has used these approaches to set new benchmarks in various tasks.

Generalizing computer vision models remains challenging, as evidenced by Human-Object Interaction (HOI) detection models underperforming on new datasets; research is being made in increasing data diversity and integrating natural language resources. Additionally, to address computational intensity in Online Continual Learning (OCL), methods like the Summarizing Stream Data (SSD) offer efficient memory usage and outperform prior models in benchmark datasets.

In addition to these advancements, Human-in-the-Loop techniques in computer vision are discussed. These methods integrate human expertise directly into the AI model's learning process, enhancing generalization, explainability, and causal understanding.

Link to Notebook

“Image and Video Data | Kaggle AI Report”

Honorable Mention

By Danial Sultanov

Dron’s report highlights the growing trend in image and video data usage. Specifically, he points to the increase in papers published in the field of computer vision and how the number of papers released has continued to increase each year following the groundbreaking AlexNet publishing in 2012. The report then covers 5 areas of computer vision: Data Preprocessing, computer vision applications, an analysis of CVPR conference submissions, CV use cases, and future perspectives for computer vision.

Data preprocessing and augmentation are discussed as essential components of computer vision pipelines. The use of a combination of real and synthetic data sources have been shown to enhance image and video processing in competitions and research.

Next, the evolution of image classification and segmentation models are covered, highlighting the rise of architectures like Visual Transformers and ConvNeXt, as well as their use with natural language processing for visual comprehension. An exploration of top papers from 2021 and 2022 to the Conference on Computer Vision and Pattern Recognition (CVPR) conference further shows that transformers and ViT topics are some of the hottest topics in the field.

In the final sections, the report discusses real-life computer vision applications and its foreseeable challenges.

Link to Notebook

04

Tabular / Time Series Data

Section Overview by Bojan Tunguz

Topic Summary

Tabular data, in the form of transactional data and records of exchange and trade, has existed since the dawn of writing. It may even precede written language. In most organizations, it is the most commonly used form of data. There is no definitive measure, but it is estimated that between 50% and 90% of practicing data scientists use tabular data as their primary type of data in their professional setting.

Time series data is, in many respects, similar to tabular data. It is often used to encode the same kinds of transactions as non-temporal tabular data, with one important distinction: inclusion of temporal information. The temporal nature of those data points becomes a major underlying feature of time-series datasets, requiring special considerations in analysis and modeling.

Tabular data, and to much lesser extent time-series data, has proven largely impervious to the deep learning revolution. Non-neural-network-based ML techniques and tools are still widely used and have stood the test of time. Nonetheless, there have been some interesting recent developments on that front as well. This remains a kind of data where a wide variety of tools and techniques are relevant, and there exists tremendous potential for further research and improvement.

Trends & Predictions

There are three main trends with machine learning for tabular data:

1. Need for unique approaches for every dataset/problem.
2. Outsized importance of data munging and feature engineering.
3. Continuing dominance of gradient boosted trees as the algorithm of choice.

The first two trends have been well known in the Kaggle circles ever since the platform launched, and the last since XGBoost was first introduced on Kaggle in 2014 (you can find a collection of winning Kaggle solutions that use XGBoost here). Even though there have been comparatively few featured tabular data competitions on Kaggle over the past couple of years, the ones that were held there reinforced these trends.

The dominance of gradient boosted trees has, until recently, not been substantially covered in the ML research literature, but over the past few years there have been more attempts to understand this phenomenon and evaluate alternative approaches. In particular, there have been many more attempts to use neural networks with tabular data, but those attempts – even when successful – come at the added expense of computational complexity. Most of that research has had minimal effect on applied ML modeling for tabular data.

It is very likely that, in the upcoming years, we will see even more research on neural networks for tabular data, as well as new innovations for gradient boosted trees. A very promising new area would be the application of generative AI to automated feature engineering and AutoML for tabular data in general.

Overview of Essays

Kaggle has its roots in hosting competitions for tabular and time series data. However, over the years, demand has shifted to other areas of ML, and the primary venue for these “classic” ML problems have become the monthly Tabular Playground Series competitions. These competitions rely on synthetic data, and are constructed in such a way as to highlight similar “real” problems that have been featured on Kaggle in the past. Overall, these competitions provide a good testing ground for various tabular and time series ML approaches, which has especially been emphasized in the winning essay in this collection.

The winning essays also incorporate information about the few "featured" tabular competitions that have taken place over the past couple of years (here, the distinction is that "featured" competitions reward large cash prizes while "playground" competitions do not).

These essays take a closer look at the special data wrangling and feature engineering techniques that have proven crucial for extracting the most amount of information from these datasets. These in-depth looks help to highlight how idiosyncratic and unique each one of those datasets and problems are.

The winning essays also incorporate information from recent research literature on Tabular Data ML. The main finding of such literature – continuing dominance of gradient boosted trees for ML modeling – has been in line with the findings from Kaggle competitions. However, as has been noted in the winning essay in this series, there has been very little work done on exploration of feature engineering in the research literature, even though this is arguably the most valuable part of ML modeling for tabular data. This would be a very productive area of future research.

“Learnings From the Typical Tabular Modelling Pipeline”

Award Winner

By Rhys Cook

Quick, accurate and expert handling of tabular data is key in machine learning. This essay examines the recent tabular data modeling pipelines of successful Kagglers by exploring individual discussion posts and aggregated Kaggle metadata. Its goal is to extract key learnings from these recent, high performing tabular data solutions and contrast them to the latest developments in the literature.

This essay is noteworthy for its detailed and systematic analysis of all the Tabular Playground Series competitions (2023 was the third year for this series).

It methodically quantifies various models used, feature engineering techniques, and breadth of writeups for these competitions. The learnings are in line with what most long-term Kaggle competitors have observed in the past, and have reinforced the notion that these “playground” competitions are indeed representative enough of the kind of work that is required for these sorts of ML problems in the real world.

Among the main learnings in this essay are that: (1) feature engineering is one of the most important aspects of the ML modeling pipeline for tabular data; (2) gradient boosted trees dominate as the go-to algorithms for tackling these problems; and (3) ensembling methods work extremely well for increasing the predictive power of tabular-data models.

Link to Notebook

“AI Report: Time Series and Tabular Data”

Honorable Mention

By Chuandong Tang, Paulina Skorupska

Kaggle has long served as a pivotal platform for fostering advancements in machine learning, hosting a rich array of approximately 60 competitions focused on a range of data types over the past two years alone. This report narrows its scope to the intricate application of state-of-the-art ML models in time series and tabular data, primarily within the framework of supervised learning. Drawing upon diverse case studies, the report elucidates innovative feature engineering and modeling techniques that have proven effective across various sectors, from natural sciences to finance.

The strength of this essay lies in its meticulous exploration of a curated selection of recent "featured" Kaggle competitions that focus predominantly on tabular data series. The essay offers a comprehensive breakdown of highly effective strategies and intricate feature engineering methodologies employed in contests such as the American Express Default Prediction, G-Research Crypto Forecasting, Tokyo Stock Exchange Prediction, Optiver Realized Volatility Prediction, and Ubiquant Market Prediction competitions. Additionally, the essay extends its scope to encompass a range of competitions that, while not exclusively tabular in nature, manifest a rich tapestry of techniques and approaches that are emblematic of tabular data challenges.

Moreover, this essay allocates significant space to scrutinizing the impact of diverse neural network architectures and methodologies in the contemporary landscape of Kaggle's tabular competitions. It delves into the nuanced applications and performance implications of convolutional, recurrent, and transformer-based neural networks, among others, in shaping the outcomes of these cutting-edge contests. This segment serves not only as an intellectual exposition but also as a practical guide for leveraging advanced neural network paradigms in data science challenges that involve tabular data sets.

Link to Notebook

“Tabular Data in the Age of AI”

Honorable Mention

By Kobbie Manrique

The purpose of this report is to provide an overview of the recent advancements in AI techniques for tabular data. Our goal is to provide valuable insights to the Kaggle data science community and inspire future innovations. By grasping the latest AI techniques and their practical applications in tabular data analysis, data professionals can remain at the forefront of this rapidly evolving field.

This essay is noteworthy for providing an excellent historical, applied, and conceptual background on tabular data. It provided rich context for dealing with such problems, and their importance in the development of computing in particular.

Even though the essay doesn’t cover experience with Kaggle competitions and tabular datasets, it provides an interesting and unique overview of some of the most advanced recent research developments in this area. The essay is particularly noteworthy for its overviews of AutoML and explainable AI, both of which are very important for many everyday applications.

Link to Notebook

05
Kaggle Competitions

Section Overview By Sanyam Bhutani

Topic Summary

Kaggle competitions are the most meritocratic avenue for enthusiasts and veterans alike to establish their AI credentials. The leaderboard is based on an objective measurement to provide a score for each submission and therefore does not lie. It is regarded by many as one of the hardest and truest challenges in data science.

This section aims to cover the developments and observations from Kaggle competitions from the last 2 years. Competitions have always been hailed as a go-to arena for testing skills of competitors and evaluating ideas, papers and frameworks. Over the years, XGBoost, H2O-3, PyTorch, albumentations, and fastai have emerged as the trusted frameworks of the community.

The real value of competitions emerges over time, where one can observe winning solutions from older competitions becoming new baselines. Eventually the “tricks” of winners become standard practice in the next competitions: ideas like pseudo labeling, seed averaging, hill climbing - to name just a few - were “tricks” that went from being explicitly mentioned in winning solutions to now frequently appearing in many solutions.

Trends & Predictions

There are a number of trends in the types of Kaggle competitions and winning approaches which we can expect to continue in the coming years.

Recently, we have seen some repeat hosts, such as RSNA, Learning Agency, BirdCLEF. These competition series provide an opportunity for participants to build on prior approaches or become more deeply connected to similar competition themes by hosts with compelling machine learning use cases. Kaggle itself has also introduced some new, more innovative variations of competitions, for example the COVID competitions which ran in sprints, and the Stable Diffusion competition which focused on reversing a generative text-to-image model.

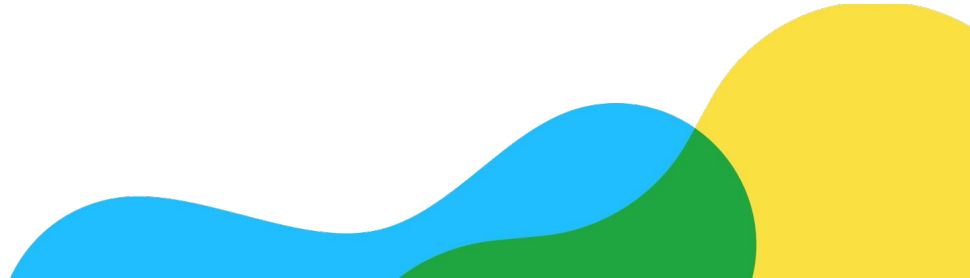
Although heavily context dependent, some model architectures have been thoroughly tested in domains: for example, the DeBERTa family, EfficientNet, and ViT have emerged as strong candidates for a wide variety of tasks. Even though these methods are computationally heavy, most recent competitions have attempted to encourage submission of small and efficient models by having an inference constraint limit, and some with a new efficiency track, where your models are allowed only to make predictions on small CPU servers.

As competitions get fiercer every year, the Kaggle team has also put extra care into beginner friendly competitions, adding incentives for sharing ideas and new competition types. It's clear we can expect more of this; Kaggle is the home of data science for a reason.

Overview of Essays

The Top essays of this category are surprisingly diverse. Instead of just focussing on winning tricks and bits, they also share the bigger picture. These essays provide an insight into emerging efficient tips and winning tricks in Kaggle competitions.

The essays from this section can alternatively be looked at as “best practice” essays, and, depending on the topic of your interest, you can find battle tested ideas here.



Kaggle Competitions Essay #1

“Towards Green AI”

Award Winner

By Leonie Monigatti

Kaggle is usually known as an “ensembling playground” to the outside world: Kaggle competitors often combine a variety of methods and models in order to increase their score without needing to balance the computational costs of their solutions. To counter this trend, Kaggle has been awarding special prizes to solutions that are both accurate and performant. This report shares learnings from Kaggle competitions concerning efficient models and efficient modeling practices, in particular.

In this essay the author tells the story of the significance of striking a balance between predictive performance and inference time to diminish the carbon footprint of deep learning models.

The report "Towards Green AI" shines a light on the pivotal challenge of our times: crafting deep learning models that deliver on performance without exacting a heavy carbon toll. Driven by Kaggle's visionary Efficiency Prize, the study delves deep into the heart of techniques like pruning, low-rank factorization, and quantization, scrutinizing their true potential in the real-world AI landscape.

Link to Notebook

“How to Win a Kaggle Competition”

Honorable Mention

By Dariusz Kleczek

This essay dives deep into the minds of Kaggle winners, and the author uses LLMs to systematically extract and analyze structured data from a myriad of Kaggle competition writeups (dedicated discussion posts where winners of Kaggle competitions describe their solutions).

It distills wisdom and ideas from the most coveted methods and strategies.

From the nuances of data augmentation to the might of gradient boosted decision trees, this report paints a comprehensive picture of what it takes to clinch that coveted top spot. A brilliant blend of data-driven insights and personal experiences, this essay is an essential guide for anyone interested in climbing the leaderboard.

Link to Notebook

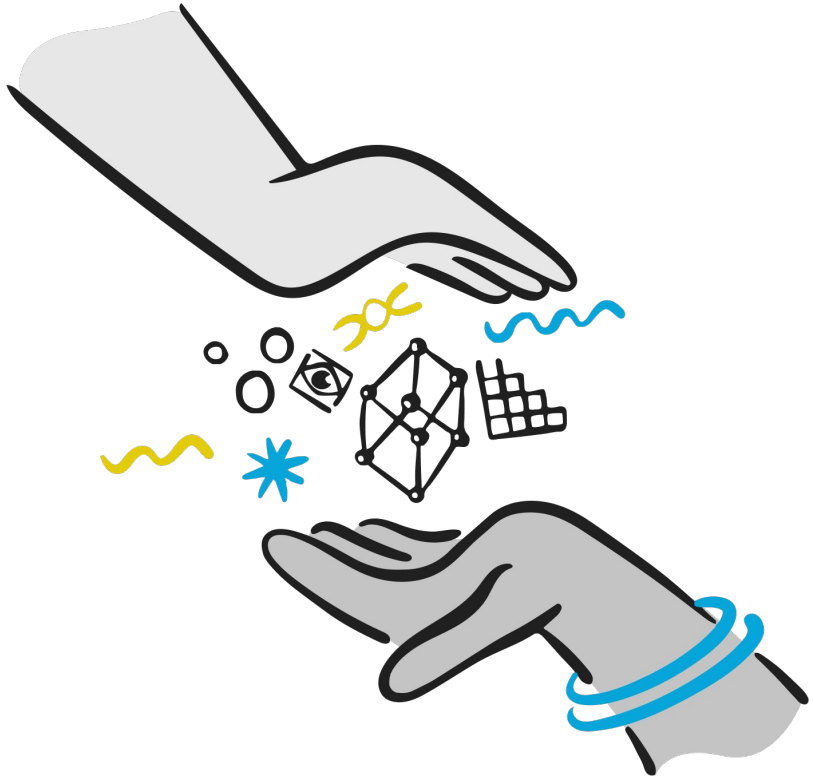
“Kaggle AI Report: Medical Imaging Competitions”

Honorable Mention

By Nghi Huynh

Medical competitions have historically been some of the most popular on Kaggle. These competitions involve different imaging modalities, such as MRI, CT scans, and X-rays. This report offers an in-depth analysis of Kaggle competitions centered around medical imaging, aiming to unearth the prevailing methodologies and architectures employed by the machine learning community. The study meticulously dissects competitions into specific categories, including Object Detection, Classification, and Segmentation. Furthermore, it delves into the nuances of the most favored Deep Learning models.

Link to Notebook



06

AI Ethics

Section Overview by Parul Pandey

Topic Summary

The section is dedicated to the continued study of AI ethics. It encompasses not only a detailed discussion of ethical principles in AI, but also an exploration of the tools and strategies that can be employed for monitoring, understanding, and mitigating risks in high-stakes machine-learning applications. We sample the complex debates surrounding ethical considerations in AI, broadening our focus to include the means by which potential risks in critical machine-learning scenarios can be managed.

Machine learning (ML), a branch of artificial intelligence (AI), has become a crucial tool in both corporate and governmental sectors worldwide, influencing areas from consumer products to security measures. Its utilization spans high-risk decisions in areas such as employment, parole, lending, security, and other high-impact realms across global economies and governments. With the recent surge in generative AI, the stakes are now even higher, amplifying both opportunities and challenges. Its widespread adoption over the past years testifies to its potential; however, it also underscores the inherent risks to users, organizations, and the public at large. Addressing these challenges is essential for both organizations and the general populace to genuinely harness the potential of this innovative technology.

Trends & Predictions

The study of AI ethics is not merely an academic endeavor but a societal imperative. As AI continues to shape the world, ensuring its ethical deployment becomes paramount to harness its benefits while safeguarding collective values. A growing consensus has emerged that ethics cannot be an afterthought; instead, they must be an integral part of AI system design. However, there's a palpable need for globally accepted standards on AI ethics. Some authoritative guidance is appearing to emerge, like the ISO's technical standards for AI and the NIST (National Institute for Standards and Technology) AI Risk Management Framework. This framework highlights key characteristics of trustworthy AI systems, including validity, reliability, safety, security, resiliency, transparency, accountability, explainability, interpretability, bias management, and enhanced privacy. To implement these characteristics, it offers actionable guidance for organizations in four areas: map, measure, manage, and govern.

Several emerging trends in this field warrant attention. One such trend is the continuous auditing of AI systems, especially those in critical sectors. Such systems might undergo ethical audits to ensure they adhere to established guidelines. It's anticipated that future AI systems will adopt an 'ethics-by-design' approach, emphasizing ethical considerations from the onset. For instance, Meta decided to approach the release of LLaMA-2 with strong focus on responsibility, providing resources and best practices for responsible development of products powered by large language models.

Moreover, as AI becomes more ingrained in daily life, there's an expected increase in public involvement. People are likely to have a more pronounced role in AI's ethical considerations, leading to more public forums, consultations, and potential referendums on significant AI deployments. In essence, the landscape of AI adoption and risk management is continuously changing, making it crucial to stay vigilant and proactive when addressing the ethical ramifications of these technologies.

Overview of Essays

The AI Ethics section garnered numerous insightful submissions, highlighting how rapidly this area is evolving. The essays received covered a diverse set of topics, reflecting the depth and breadth of this important discipline.

Some essays covered the big picture of AI ethics, discussing the broader challenges and opportunities it presents. They touch upon how we need to shape AI in a way that is responsible and considers everyone in society, from ensuring fairness in its decisions to thinking about its impact on the environment.

Others focus on more specific areas. For example, some discuss the latest advancements in AI, like the exciting world of generative AI, and the ethical questions these new technologies bring. Others take a closer look at research in the AI world, analyzing the ethical considerations in AI studies and publications.

But despite their different angles, all the essays converge on a shared sentiment: AI holds immense promise for our future, but we need to approach its development and use with care, thought, and a strong sense of responsibility.

AI Ethics Essay #1

“Exploring the Landscape of AI Ethics”

Award Winner

By Patrik Joslin Kenfack, Meghana Bhange, Maryam Babaei, Ivaxi Sheth, Dave Harold Mbiazi Njanda

This winning essay delves into the critical urgency of AI’s ethical implications in our ever-evolving digital era. The essay talks about central principles, crucial for instilling trust in AI, including privacy, data protection, transparency, explainability, fairness, accountability, safety, robustness, and even environmental considerations. This essay offers a comprehensive exploration of each principle, presenting contemporary efforts and strategies to seamlessly integrate them within AI’s lifecycle.

Notably, the essay offers a dual lens, analyzing each guideline from societal impacts to technical advancements. Each guideline is dissected not only for its societal importance and value, but also for the technical innovations that aid its implementation. In doing so, the essay bridges the perceived gap between ethical principles and actionable technological solutions. Through this comprehensive approach, the essay makes a compelling case for these principles as the foundation for fostering genuine trust amongst all stakeholders throughout the AI system lifecycle.

Link to Notebook

AI Ethics Essay #2

“Developments in AI and Ethics in the Past 2 Years”

Honorable Mention

By Antong C.

The essay covers the evolution of AI ethics over the past two pivotal years, highlighting enduring challenges and the continuous efforts to address them. During this period, the AI domain has witnessed significant growth and heightened interest in its capabilities. This rapid advancement, however, has accentuated the critical need for robust AI ethics.

Emphasizing regulatory facets over philosophical ones, the essay provides an overview of AI ethics, delves into its core principles, pinpoints application gaps, and suggests potential remedies.

Designed as an introductory piece rather than a deep dive, the essay seeks to amplify awareness of AI ethics and champions a future where AI is more inclusive, beneficial, and ethically grounded.

Link to Notebook

AI Ethics Essay #3

“Ethical AI Is All We Need!!”

Honorable Mention

By Shreya Mishra, Piyush Mathur, Raghav Awasthi, Anya Mathur, Harshit Mishra

In this essay, the authors underscore the essentiality of ethics for building trust in AI systems and fostering sustainable advancement in AI research. The paper sets out to investigate and assess the ethical dimensions embedded in research publications, focusing specifically on AI articles from 2007 to 2023.

Leveraging informative data visualizations, the authors probe into four principal dimensions: “Contextual Topics Learning,” “Development Maturity Learning,” “Gender Representation Learning,” and “Sentiments Learning” in the context of AI publications. Using illustrative examples, the authors emphasize the tangible implications of AI systems, aiming to spotlight real-world ethical challenges and considerations.

Link to Notebook

07
Other Topics

Section Overview by Martin Henze

Topic Summary

The rapid growth that the field of Machine Learning has been experiencing over the recent years has led to an explosion of tools, techniques, and applications. In our final category - simply called “Other” - we cover those aspects of the ML landscape that may not fit neatly into any of the previous, more specific categories. This long tail of topics is reflected in a wide variety of topics chosen by the essayists. In turn, the multitude of fascinating subjects showcases the exceptional range and diversity of expertise within the Kaggle community.

With one major exception, the topics of the essays have little in common with one another. Participants analyze such diverse topics as optimization algorithms, graph networks, theoretical physics, robotics, healthcare, the future of work, data-centric AI, mathematical research, autonomous cars, and many more. Together, they offer a panoramic snapshot of the wide-ranging influence of Machine Learning at this pivotal moment in time.

Trends & Predictions

Participants notably focused on the increasing importance of ML algorithms in the medical and healthcare fields.

From humble beginnings of applying computer vision tools in assisting with diagnosis in radiology or tomography, the power of ML to save lives and cure diseases is being increasingly leveraged by medical professionals and researchers. Modern applications include, among others, genetic sequencing (with NLP techniques), robotic surgery, medical teaching, drug discovery, and of course the extremely relevant area of vaccine research. This trend is also reflected in a large number of healthcare-related Kaggle competitions in recent years, including those focused on COVID-19. As ML-driven tools are becoming more mature and more training data is being built through real-world experience, the medical and healthcare fields will likely see significant efficiency gains through automation in the near future.

Another exciting trend are multi-modal ML applications powered by growing repositories of pretrained models. This is relevant for generative AI applications utilizing agent systems, but also for tailored projects that draw on diverse data sources. In the coming years, models and applications that can derive insights from multi-sensory inputs in a human-like manner will likely play an increasingly important role.

Overview of Essays

This section will present a collection of the best and most interesting essays, including the winning contribution. They will cover the subjects of optimization algorithms, graph networks, theoretical physics and string theory, as well as the healthcare sector.

Any anthology in such a vibrant and fast-moving field can only aspire to offer a tasting menu of some of the delicacies that the research and praxis have to offer. This selection of essays represents the various ways in which advancements in ML transcend the boundaries between the traditional categories in the field and exert a growing interdisciplinary influence. The evergreen research area of optimizers lies at the foundations of learning methodologies. Thinking in graph networks can provide fresh perspectives on challenging problems in many areas. And, the areas of medical research and theoretical physics illustrate the increasing impact of ML techniques on topics ranging from improving our daily lives to deepening our understanding of the fundamental laws of nature. It is a golden age of ML progress, and the following essays contain many nuggets of valuable insights for the curious reader.

Other Topics Essay #1

“Optimization Algorithms in Deep Learning”

Award Winner

By Svetlana Nosova

This winning essay presents a well-structured and exceptionally accessible overview of one of the key ingredients of modern ML: optimizers. The author eases the audience into the report by providing great background information on the fundamental technique of gradient descent, along with its stochastic and mini-batch variations. Gradient descent and backpropagation are at the heart of the training process for all neural network models. Gradient descent also powers the boosted tree ensembles, which remain the gold standard for tabular data challenges.

Through a concise description of the important momentum-based and adaptive methods, the author draws an elegant arc to what is arguably the most popular optimizer at present: Adam. The default choice in many ML libraries, the Adaptive Moment Estimation that Adam provides allows for more efficient learning rates and faster convergence. But progress is ongoing, and the essay concludes by highlighting two recent developments: the versatile Momentum Models (MoMo) approach alleviates the need for learning rate schedulers. In contrast, the Scalable Stochastic Second-Order Optimizer (Sophia) is considering second derivatives of the complex objective landscape in which we are aiming to find the minimum.

This essay expertly optimizes the balance between beginner-friendly descriptions and technical details. Its language is clear and its narrative is engaging. It leaves the reader well prepared to dive into the treasure trove of references to learn more.

Link to Notebook

Other Topics Essay #2

“Applications of Artificial Intelligence and Machine Learning Models Within the Biosciences”

Honorable Mention

By Samantha Lycett

This concise article studies the various ways in which ML approaches were used in the response to the COVID-19 pandemic. Against a backdrop of describing the healthcare and bioscience responses to the virus, the author analyzes the relevant ML contributions through the examples of Kaggle datasets and competitions.

Especially during the early pandemic years, the Kaggle community was a microcosm of the challenges that the ML field was facing under considerable time pressure. Institutions and community members published large volumes of COVID-19 data on Kaggle. This led to fine-tuned NLP models tailored to COVID-19 literature, which are now finding increasing adoption by researchers.

early-stage time-series forecasting challenges for infection numbers, a medical imaging competition to detect COVID-19 abnormalities in chest radiographs, and an urgent mRNA Vaccine Degradation Prediction to help bring a vaccine to mass production.

The author then links those events to the larger picture in biosciences and specifically to the novel methodology of using language models for biological sequence data, where molecules act like letters and structures like proteins correspond to sentences. The success of the AlphaFold project for 3D protein modeling is a key example for the successful synergy of ML and bioscience research.

Link to Notebook

“Applying AI/ML to Theoretical Physics”

Honorable Mention

By Lorresprz

This work manages to describe recent challenges in the complex field of string theory in a way that is remarkably accessible to ML practitioners. String theory, with its manifolds and compactified dimensions, is arguably one of the more complex areas in a complex field. But, its aspirations to construct a unified “theory of everything” make it a leading contender for rewriting our understanding of physics and of the entire universe. In the already decades-long endeavor to study string theory, ML has only recently arrived at the scene and might play an increasingly pivotal role in the future.

A main subject of this essay is the research to use ML to study problems involving spacetimes with more dimensions than four. String theory predicts extra dimensions that are hidden away (aka compactified), but their geometries (aka manifolds) affect those physical particles that we can observe. Both supervised and reinforcement learning can help to filter through a myriad of possible manifolds to discover viable candidates for geometries that are consistent with observational results.

As the author writes, at present more and more theoretical physicists are incorporating ML-based approaches into their research projects of increasingly diverse natures. Similar to a breakthrough like AlphaFold in biology, the ML revolution might contribute to a quantum leap in our understanding of the universe.

Link to Notebook

Other Topics Essay #4

“Kaggle AI Report: Medical Data”

Honorable Mention

By Diego Flores

In Diego Flores essay "Lessons learned from working with medical data", the author explores the use of AI technologies in the medical space, and they discuss some of the challenges and lessons learned. The report highlights important concepts such as federated learning and how it relates to privacy and security, and they discuss the outsized importance of model explainability in medical settings as well. Large language models (LLMs) can be used to summarize both medical records and biomedical research, and specialized tools such as ClinicalGPT are presented to the reader along with additional LLM-based applications.

This report does a great job of introducing both the challenges and the triumphs that AI researchers have encountered while working with medical data.

Link to Notebook

“Graph Learning and Complex Networks”

Honorable Mention

By Hoda Jalali Najafabadi, Ali Jalali

Relational graphs are a central concept in understanding the structure of complex systems like social networks or traffic flow patterns. Following a gentle introduction to the world of graph theory, the authors of this review article take the reader on a tour of graph applications.

This essay touches on the areas of particle physics, anomaly and fraud detection, transportation and traffic, protein folding (where we meet AlphaFold again), cheminformatics and computational materials science, brain and computational neuroscience, drug-drug interaction, text data, and robotics and multi-agent systems. Not content with this dizzying variety, the article contains an additional plethora of well-organized references and resources for the curious reader.

In a joyously fractal way, this essay represents a microcosm of the category we chose to call “Other”. Just like graphs can teach us about path finding in decentralized robots, physical particle dynamics, insurance fraud, neurological disorders, or disease contagions, so are the modern ML methods becoming relevant to more and more aspects of our world. As the graphs of our lives and endeavors are extending towards the future, the connections to be built and cemented through new ML technologies are bound to further accelerate the rate of human progress in ever new and exciting ways.

Link to Notebook

Wrap-up

Conclusion
Phil Culliton

In this effort, our community wrote hundreds of essays covering a broad array of topics, and then experts from our community selected the best. The result is a collective perspective on the rapid advancements of AI, shedding light on the most salient topics in modern machine learning.

Many of the essays discussed the recent progress and potential of generative AI to revolutionize multiple industries and massively broaden access, making it the front runner for shaping the future of AI. Ethics and “green” AI will command even more attention as scale, audience, and capabilities expand. Making sense of this future will require validation through competitions, improved benchmarks, and other tools capable of incorporating diverse and representative feedback.

We are excited by the potential of this report to highlight insights and advancements in our field, as told through the collective mindshare of the Kaggle community. As machine learning becomes more accessible, sampling from the diverse opinions of its practitioners is an effective strategy to understand an ever-evolving discipline.

Credits

We extend a tremendous thank you to the people that made this report possible, including:

Bojan Tunguz - Area Chair
Camille McMorrow - Design
Christof Henkel - Area Chair
Claudia Sanchez - Design
D. Sculley - Supervising Author
Jephunney Nduati - Design
Julia Elliott - Coordinator
Karnika Kapoor - Area Chair
Kinjal Parekh - Coordinator
Mark McDonald - Copyeditor
Martin Henze - Area Chair
Parul Pandey - Area Chair
Paul Mooney - First author
Phil Culliton - First author
Raphael Kerley - Design
Rob Mulla - Area Chair
Sanyam Bhutani - Area Chair
Sara Wolley - Coordinator
Siddhita Upare - Illustrations
Will Cukierski - Copyeditor

Citation

Paul Mooney*, Phil Culliton*, Abir Eltaief, Ali Jalali, Antong C., Anya Mathur, Arya Gaikwad, Bojan Tunguz, Christof Henkel, Chuandong Tang, Danial Sultanov, Dariusz Kleczek, Dave Harold Mbiazi Njanda, Diego Flores, Dmitri Kalinin, Harshit Mishra, Hoda Jalali Najafabadi, Julia Elliott, Ivaxi Sheth, Karnika Kapoor, Kobbie Manrique, Leonie Monigatti, Lezhi Li, Lorresprz, Mark McDonald, Martin Henze, Maryam Babaei, Meghana Bhange, Nghi Huynh, Parul Pandey, Patrik Joslin Kenfack, Paulina Skorupska, Piyush Mathur, Pranav Mohan Belhekar, Raghav Awasthi, Rhys Cook, Rob Mulla, Samantha Lycett, Sanyam Bhutani, Shreya Mishra, Svetlana Nosova, Theo Flaus, Trushant Kalyanpur, Will Cukierski, Xinxi Chen, Yassine Motie, Yuqi Liu, Yuxi Li, Zhengping Zhou,
D. Sculley.

(2023). AI Report 2023. Kaggle.
Kaggle.com / AI-Report-2023

*Equal first authors

Thanks

kaggle