Posted on 12/15/2023

DSML Trend: DeepMind's FunSearch discovers new knowledge

Recently researchers from Google DeepMind claim a groundbreaking achievement—the world's first scientific discovery facilitated by a large language model. This breakthrough suggests that technologies like ChatGPT have the potential to generate information surpassing human knowledge. The development, called "FunSearch" (short for "searching in the function space"), leverages a Large Language Model (LLM) to devise computer program solutions for various problems. Paired with an "evaluator" ranking the program performances, the best solutions are amalgamated and fed back into the LLM, propelling an iterative process that transforms weak programs into robust ones capable of unveiling new knowledge.

In a noteworthy accomplishment, AI, through FunSearch, tackled a longstanding mathematical challenge—the cap set problem. This problem involves identifying the most extensive set of points in space where no three points align in a straight line. FunSearch produced programs generating new large cap sets, surpassing the best-known solutions devised by mathematicians.

This is my take on it: During an interview with CNN, renowned physicist Michio Kaku derided chatbots, likening them to a "glorified tape recorder." However, contrary to the criticism that chatbots merely recycle existing data without generating new knowledge and are prone to confabulation, the preceding advancement showcases the potential for AI to contribute to knowledge creation. In addition, while chatbots lack the ability to conduct original research or independent experiments, they can aid in hypothesis generation. By sifting through vast datasets, recognizing patterns, and formulating hypotheses, AI can offer valuable insights. For instance, analyzing medical records for potential symptom-disease relationships or studying financial data to predict market trends—akin to the Swanson process—demonstrates AI's capacity to contribute meaningfully to the creation of new knowledge.

Link to article: **https://www.theguardian.com/science/2023/dec/14/ai-scientists-make-exciting-discovery-using-chatbots-to-solve-maths-problems?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+December+15th%2C+2023&utm_campaign=16122023**

Posted on 12/8/2023

DSML trend: Google Gemini may outperform ChatGPT

Gemini released by Google two days ago (December 6) is considered a quantum leap in AI innovation. Gemini comes in three versions tailoring for specific tasks. Gemini Ultra is the most powerful variant intended for handling incredibly complex tasks with its multimodal capabilities, whereas Gemini Pro is designated for powering Google's consumer-level products operating in the cloud, such as Google Bard and other PaLM2 products. Last, Gemini Nano is specifically crafted to operate natively on mobile devices, such as cell phones. According to Google, Ultra demonstrated superior performance compared to "state-of-the-art" AI models, including ChatGPT's most advanced model, GPT-4, across 30 out of 32 benchmark tests. Additionally, the Pro model surpassed GPT-3.5, the underlying technology for the free-access version of ChatGPT, in six out of eight tests. The driving force behind Gemini is DeepMind co-founder Demis Hassabis, who advocates the integration of LLMs and other AI techniques to enhance comprehension.

That's my take on it: While I haven't personally experimented with Gemini yet, third-party analyses suggest that it has the potential to surpass ChatGPT. First, in contrast to conventional large language models (LLMs) that are predominantly text-centric, Gemini stands out as a natively multimodal model, displaying proficiency in learning from a diverse array of data sources, including audio, video, and images. This breakthrough transcends the text-focused constraints of LLMs, hinting at a potential paradigm shift in the capabilities of AI products. Second, Gemini reportedly undergoes training on more extensive datasets of text and code, ensuring that the AI model stays updated with the latest information and can provide accurate and highly reliable responses to queries. Moreover, the model can also generate hypotheses for further research, a capability that experts believe could revolutionize scientific discovery and potentially lead to breakthroughs in fields such as technology and medicine.

Links to articles: **https://www.theguardian.com/technology/2023/dec/06/google-new-ai-model-gemini-bard-upgrade**
https://www.youtube.com/watch?v=lgBAS9CFYIE

Posted on 11/17/2023

DSML trend: OpenAI fires Sam Altman

It happened just now. Today (November 17, 2023) OpenAI's board of directors announced that Sam Altman will be stepping down as CEO, with technology chief Mira Murati set to take over the position. The decision comes after a thorough review process, during which the board determined that Altman's communication lacked consistent

truthfulness, thereby impeding the board's ability to fulfill its responsibilities. The statement emphasized that due to this, the board no longer has confidence in Altman's capacity to effectively lead OpenAI. The board also announced that Greg Brockman, OpenAI's president will be stepping down as chairman of the board but will keep a role at the company.

That's my take on it: As of now, Altman has not issued any public response yet. Given the maturity of the technology, I believe the departure of both Altman and Brockman will likely have minimal impact on the development of OpenAI or the broader field of generative AI. However, it's improbable that Altman will sit there and do nothing. It is possible that he may embark on launching another startup or join a competitor to OpenAI (e.g. Claude or Google Bard?)

Link to article: **https://www.cnbc.com/2023/11/17/sam-altman-leaves-openai-mira-murati-appointed-interim-boss.html**

Posted on 11/17/2023

The *Harvard Business Review* featured an article on November 2, 2033, titled "How Cloud Technology is Transforming Data Science." Written by Peter Wang, the CEO and co-founder of Anaconda, the article discusses the impact of cloud computing on data science practices. Wang highlights how cloud platforms, such as IBM Watson and Tableau, are revolutionizing the field by offering scalable computational resources and enhancing workforce agility. These cloud-based analytics tools empower teams to access information and collaborate in real time, facilitating quicker insights and problem-solving. Moreover, cloud computing promotes inclusivity in data science by providing smaller entities, such as startups and small teams, with the means to innovate on par with larger corporations. The cloud's collaborative capabilities extend to distributed data science teams, enabling effective collaboration irrespective of geographical constraints. While the cloud brings forth significant advancements in data science, it also introduces new challenges, particularly in data privacy and security. To address these concerns, Wang emphasizes the importance of employing techniques like data partitioning, encryption, and robust frameworks for mitigation.

That's my take on it: Given the significance of cloud computing, it is undoubtedly essential to integrate it into the curriculum of data science education. However, the current landscape of the cloud computing market is highly diverse, featuring numerous vendors such as AWS, Google Cloud, Microsoft Azure, IBM Watson, and more. This question arises: should cloud computing training be tailored to specific vendors or remain vendor-independent? Opting for vendor-specific training allows students to gain practical experience with the tools and services of major cloud providers like AWS, Azure, and IBM

Watson, preparing them for roles utilizing these platforms. The drawback is that knowledge becomes less transferable if students later work with a different cloud provider. On the other hand, adopting a vendor-independent approach ensures knowledge transferability across various cloud platforms and avoids explicit promotion of specific vendors within the program. However, graduates may need additional, vendor-specific training upon entering the workforce. Striking a balance between these approaches is crucial to provide students with a well-rounded and adaptable skill set in the dynamic field of cloud computing. What do you think?

Full article: https://hbr.org/2023/11/how-the-cloud-is-changing-data-science?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+November+17th%2C+2023&utm_campaign=18112023

Posted on 10/13/2023

Yesterday (10/12) an article published by *Analytics Insight* detailed how seven data science positions can be executed without the need for programming skills. Instead, they rely on the capabilities of user-friendly software tools like Tableau, Excel, Power BI, and more. These positions are:

- Data analyst
- Business Intelligence Analyst
- Data Consultant
- Market Research Analyst
- Data Visualization Specialist
- Data-driven Strategist
- Data Product Manager

That's my take on it:

No code solutions provide pre-built components, templates, and graphical user interfaces (GUI) that can accelerate development compared to programming. These tools allow users to focus more on the research question, the data, and the business logic rather than the syntax. However, no-code solutions inevitably involve some trade-offs in terms of flexibility, customization ability, scalability and performance compared to coding. Data science education should balance both sides. In my humble opinion, starting data science training with a focus on programming right away might not be the most advisable approach. Emphasizing the fundamental concepts as the foundation is crucial, while the tools, which serve as means to an end, should be treated as secondary. Leveraging GUI-

based software applications reduces the entry barriers into the field, thereby broadening the pool of potential talents.

Article: https://www.analyticsinsight.net/7-data-science-jobs-you-can-do-without-any-coding-skills/?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+October+13th%2C+2023&utm_campaign=14102023

Posted on 9/14/2023

Today marks the second day of the 2023 Dreamforce conference, which is being hosted by Salesforce. It was a great experience even though I attended the conference remotely. Salesforce is widely recognized for its exceptional data visualization platform, Tableau, as well as its AI-driven analytical tool, Einstein. The central theme of this conference revolves around the concept of trust. Specifically, Salesforce is dedicated to constructing reliable systems that prioritize security, compliance, and dependability.

Throughout the conference, Salesforce has showcased its ability to guide users in creating more effective prompts through the innovative feature known as **prompt tuning**. Moreover, the event has featured numerous enlightening and captivating sessions. For instance, it has provided a platform for interviews with several distinguished AI leaders and innovators who have been acknowledged by TIME 100. Among these esteemed interviewees is Dr. Fei Fei Li. During her interview, Dr. Li openly expressed her wholehearted embrace of this transformative technology. While some individuals may be skeptical of this powerful yet unfamiliar technology, Dr. Li made a thought-provoking comparison. She pointed out that today, we are not overwhelmed by electricity, and we readily use medications like Tylenol despite not fully comprehending their chemical composition. Addressing concerns about AI bias, Dr. Li contended that AI can be harnessed to mitigate bias. As an example, AI can scrutinize instances where male actors receive more screen time than their female counterparts, highlighting disparities and providing an avenue for rectification.

That's my take on it: Critics have voiced concerns that AI tools might inadvertently encourage laziness and plagiarism. However, it is undeniable that AI is here to stay. The integration of AI into various industries is inevitable, and skills related to AI, such as prompt engineering, are increasingly being recognized as indispensable.

Salesforce, as the world's third-largest software company and the second largest in Japan, wields significant influence in this technological landscape. As high-tech companies like Salesforce incorporate prompt tuning into their product portfolios, it is foreseeable that in the near future the utilization of prompt engineering will become as ubiquitous as the use of smartphones and tablets.

Dr. Fei Fei Li's compelling metaphors, likening AI to electricity and Tylenol, underscore the notion that embracing transformative technologies is a natural progression of human innovation. This phenomenon is not dissimilar to the initial opposition encountered by calculators, which were once believed to diminish human numerical skills. Today, they are as commonplace as electricity and Tylenol, illustrating how society adapts and integrates new tools into everyday life.

Conference's website: https://www.salesforce.com/dreamforce/

Posted on 9/6/2023

DSML trend: Guardian blocks ChatGPT from accessing its content

On September 1 2023, the Guardian announced its decision to block access to its content for the AI text generation program, ChatGPT. In a statement, the publisher emphasized that the scraping of their intellectual property for commercial purposes has always been against their terms of service. They also highlighted their commitment to fostering mutually beneficial commercial relationships with developers worldwide. Other news media, including CNN, Reuters, the Washington Post, Bloomberg, and the New York Times implement similar policies. OpenAI, the owner of ChatGPT, had previously revealed an opt-out option for website owners who didn't want their content used by AI algorithms.

That's my take on it: Whether AI's utilization of existing published content constitutes copyright infringement or qualifies as fair use has been an ongoing debate. This same issue extends to AI art tools like Midjourney and Stable Diffusion. It is important to note that AI chatbots do not simply copy and paste content from the source. Rather, the nature and purpose of its use can be seen as **transformative**, meaning that AI repurposes copyrighted material in a novel and distinct manner. Similarly, AI art tools do not merely create collages; instead, they **learn from patterns** in existing artworks to generate entirely new images. Consider this analogy: If I extensively study art by browsing around a library and a museum, and subsequently, based on this knowledge I write a new article or create a new painting on my own, should the library or museum prevent me from accessing their information?

Full text: https://www.theguardian.com/technology/2023/sep/01/the-guardian-blocks-chatgpt-owner-openai-from-trawling-its-content

Posted on 9/1/2023

DSML trend: China's Baidu ‘s AI chatbot Ernie Bot is publicly accessible

On August 31, 2023, Baidu, the Chinese search engine and AI company, made a significant move by unveiling "Erine Bot," their equivalent of the ChatGPT language model, to the public. As a result, Baidu's stock price surged more than 3%. This strategic move aligns with Beijing's vision of AI as a critical field, one where they aim to challenge the US and emerge as a global leader by 2030. By releasing Ernie Bot to the public, Baidu intends to gather extensive real-world user feedback. This feedback loop will, in turn, play a vital role in enhancing Ernie and strengthening Baidu's foundational models. Coincidentally, on the same day, two other prominent AI companies in China, Baichuan and Zhipu AI, also introduced their own AI language models.

That’s my take on it: Back in 2017, Russian President Putin emphasized the transformative potential of AI by saying “whoever becomes the leader in this sphere will become the ruler of the world.” This perspective reflects the ongoing international competition among major technological powers, including the US and China, to gain supremacy in AI research and development.
Unfortunately, I faced challenges registering with Baidu's AI, as it requires a China’s cell phone number for access. Consequently, I was unable to evaluate Ernie Bot personally. However, those who did manage to access Baidu's AI encountered significant restrictions, particularly in its reluctance to answer **sensitive political and historical inquiries**. If you had the opportunity to assess Ernie Bot, I would greatly appreciate it if you could share your insights and findings with me. If you found a way to circumvent the requirement of providing a China’s cell phone information when registering for Ernie Bot, please let me know too.
Full text: https://techxplore.com/news/2023-08-china-baidu-ai-chatbot-ernie.html?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+September+1st%2C+2023&utm_campaign=02092023

Ernie Bot’s website: https://yiyan.baidu.com/welcome

Posted on 8/30/2023

DSML Trend: Revival of OpenAI?

According to reports posted by "The Information" and Reuters on August 29, 2023, OpenAI is poised to achieve over $1 billion in revenue within the upcoming year through the sale of AI software and the corresponding computational capacity that drives it. Previously, the creators of ChatGPT had estimated revenue of $200 million for the current year. Notably, the company, backed by Microsoft, is now amassing a staggering revenue surpassing $80 million each month, a significant escalation from the mere $28 million garnered throughout the entirety of the preceding year.

That's my take on it: Two weeks ago, IT experts predicted that OpenAI might go bankrupt by the end of 2024 due to a decline in usage. Suddenly this situation has undergone a surprising reversal. In my opinion, the future trajectory of OpenAI remains uncertain, because the fate of the company relies on a single product. Prior to ChatGPT, OpenAI boasted another flagship product known as DALL.E2, a creative tool for generating visual art. However, the market of generative art has now been predominantly seized by Midjourney, which boasts a user base of 15 million, the largest among all image generation platforms. In terms of overall image production volume, Stable Diffusion takes the lead with an impressive 12.59 billion images generated.
The question arises: should OpenAI reallocate its R&D resources to the more promising ChatGPT and relinquish DALL.E2, or should it engage in a dual-front battle? This is an intricate puzzle that demands careful consideration.

Reuters's report: **https://www.reuters.com/technology/openai-track-generate-more-than-1-bln-revenue-over-12-months-information-2023-08-29/#:~:text=Aug%2029%20(Reuters)%20%2D%20OpenAI,in%20revenue%20for%20this%20year.**
Statistics of generative art tools: **https://journal.everypixel.com/ai-image-statistics**

Posted on 8/26/2023

DSML trend: G2 Grid for Data Science and Machine Learning Platforms

On August 24, 2023, G2 released the *G2 Grid for Data Science and Machine Learning Platforms*. To be considered for inclusion in this DSML benchmarking, the candidate must adhere to the following criteria:

1. Facilitate the connection of data to algorithms, enabling them to acquire and adapt knowledge.
2. Enable users to construct machine learning algorithms and/or furnish pre-built machine learning algorithms suitable for less experienced users.
3. Furnish a platform for the widespread deployment of artificial intelligence.

G2 classified DSML companies into four distinct quadrants, namely, leaders, high performers, contenders, and niche, utilizing a dual-dimensional framework: market presence and customer satisfaction. According to G2 scoring, currently the leaders of DSML are:

- Databricks Lakehouse
- IBM Watson Studio
- Matlab
- Alteryx
- Vertex AI
- SAS Visual Data Mining and Machine Learning
- Anaconda
- Saturn Cloud
- Microsoft Azure Machine Learning
- Deepnote
- Amazon SageMaker and AWS Trainium
- TensorFlow
- Qlik AutoML

That's my take on it: The preceding list includes well-established companies like SAS, IBM, and Microsoft, alongside newcomers challenging the existing order. I admit that I do not possess the skill sets required for all of the software tools mentioned. Coping with the rapid evolution of technologies poses a considerable challenge for university professors, particularly in fields where progress is frequent. In my opinion, transitioning the emphasis from instructing specific skills to nurturing the capacity for **perpetual learning** is undeniably a valuable approach. To remain current, one effective tactic involves inviting guest speakers from industry or research domains to share their expertise and insights with students. This exposure acquaints students with real-world applications and prevailing industry methodologies. Moreover, it is imperative for faculty to motivate students to cultivate a mindset characterized by **openness to change and a willingness to experiment**. By the time my students graduate, G2, Gartner, Forrester, and IDC may compile a new list of DSML leaders!

Full report: https://www.g2.com/categories/data-science-and-machine-learning-platforms

Posted on 8/26/2023

In a recent piece published on *KDnuggets* (August 24, 2023), Dr. Mandar Karhade speculated the architecture of GPT-4 based upon leaked information. The author posited that rather than being a singular colossal model, GPT-4 might consist of eight separate models, each bearing 220 billion parameters. This novel approach involves breaking down a given task into smaller subtasks, which are then tackled by specialized experts within the context of these models. The strategy mirrors a divide-and-conquer methodology. Subsequently, a gating model is introduced to determine the optimal expert for each subtask, culminating in the final prediction. However, the author included a disclaimer emphasizing the non-official nature of this information.

That's my take on it: At the present time, this notion remains an unverified rumor. Nevertheless, the idea holds a certain degree of credibility. The underlying concept closely resembles, if not mirrors, the principles of **ensemble methods and model comparison**, a common practice in the realm of Data Science and Machine Learning. In ensemble methods such as boosting and bagging, numerous modeling procedures are executed on partitioned subsets of data. Subsequent model comparison is conducted to select the most optimal solution derived from an array of modeling techniques: neural networks, SVM, bagging, boosting, among others. Hence, the synthesis of eight models in GPT-4 represents a natural progression akin to ensemble methods and model comparison, taking the idea a step further.

Full article: https://www.kdnuggets.com/2023/08/gpt4-8-models-one-secret.html

Posted on 8/14/2023

DSML trend: IBM will integrate Meta's Llama into Watson

On August 9, IBM announced plans to host Meta's 70 billion parameter Llama 2 large language model on its Watson AI and data science platforms. Currently in watsonx.ai, users can leverage pre-trained models from IBM and Hugging Face for Natural Language Processing tasks, such as content generation and summarization, as well as text classification and extraction (text mining). The future addition of Llama 2 to watsonx.ai will be a milestone for IBM's generative AI roadmap, likely followed by upcoming releases of its AI Tuning Studio.

That's my take on it: IBM's flagship data science products are Watson Studio and SPSS Modeler. For a long time, IBM has trailed its top competitor SAS Institute in user base, interface, and capabilities. Nevertheless, IBM has invested in AI research and development since the 1950s. In 1997, IBM's Deep Blue beat the world chess champion in a six-game match. In 2011, IBM's Watson competed and won against top human Jeopardy! contestants. Although Meta's Llama is less powerful than models like Claude 2, Google Bard, and ChatGPT, incorporating a large language model into IBM products is still strategic. However, it's too early to tell whether IBM can overtake SAS in the near future.

Full announcement: **https://newsroom.ibm.com/2023-08-09-IBM-Plans-to-Make-Llama-2-Available-within-its-Watsonx-AI-and-Data-Platform**

Posted on 8/14/2023

DSML trend: OpenAI faces financial challenges and the rise of Claude

An article in yesterday's *Business Today* (August 13) reported that OpenAI, the pioneering AI company that brought ChatGPT to the mainstream public, is facing financial challenges. The costs to operate ChatGPT amount to around $700,000 per day. Despite efforts to monetize GPT-3.5 and GPT-4, OpenAI has yet to earn sufficient revenue to cover its expenses. According to SimilarWeb data., ChatGPT's user base declined 12% from June to July 2023, dropping from 1.7 billion to 1.5 billion monthly users.

That's my take on it: Researchers at Stanford and UC Berkeley systematically evaluated different versions of ChatGPT. It was found that in math tests, ChatGPT solved 488 out of 500 correctly in March (97.6% accuracy). By June, its accuracy dropped 2.4%. ChatGPT's global website traffic fell, especially after the launch of Claude 2. Claude 2 scored 71.2% on a Python coding test versus ChatGPT's 67%. Claude is also more updated, with an early 2023 cutoff versus September 2021 for ChatGPT. While it's premature to declare the end of ChatGPT, the future landscape of large language models is volatile as more competitors enter the market.

Full article: https://www.businesstoday.in/technology/news/story/openai-might-go-bankrupt-by-end-of-2024-chatgpt-costing-over-rs-580-crore-per-day-report-393925-2023-08-12

Posted on 8/1/2023

Center for Consciousness Studies at the University of Arizona and California for Human Science will cohost a conference entitled **"Neuroscience needs a revolution to understand consciousness"** between August 18-23, 2023. One of the keynote speakers is **Sir Roger Penrose**, a British mathematician, physicist, philosopher of science, and Nobel Laureate in Physics. The following is a brief introduction to the theme of the conference.

"'AI has reinforced the notion of the brain as a complex computer of simple, empty, 'cartoon' neurons based on 1950s physiology, processing solely by surface membranes, synaptic transmissions and firings as "bit-like' units in frequencies up to 100 hertz…The Penrose-Hameroff 'Orch OR' theory proposes consciousness depends on 'orchestrated' ('Orch') quantum superpositions leading to Penrose 'objective reductions' ('OR', wavefunction self-collapses) in brain microtubules, connecting to fundamental spacetime geometry. Orch OR has more explanatory power, connection to biology, and experimental validation than all 'neuroscientific' theories based on low frequency, oversimplified cartoon neurons combined… Neuroscience needs a revolution inward, to deeper, faster quantum processes in microtubules to understand consciousness and treat its disorders."

That's my take on it: Many experts speculate that AI may eventually attain self-consciousness, potentially posing a threat to humanity. The concept of consciousness raises several fundamental questions: What is consciousness? How can we ascertain whether an AI system is genuinely self-conscious? Do you have to fully understand consciousness in order to know whether a machine is self-aware? The widely-used Turing test, considered behavioristic, is deemed unreliable for this purpose.

During the 1980s and 1990s, Roger Penrose expounded on the notion of human consciousness in his books "Emperor's New Mind" and "Shadows of the Mind." He argued that consciousness involves non-algorithmic processes that defy computational reduction. Penrose also criticized the concept of Strong AI, which contends that machines can achieve human-like consciousness. He posited that human attributes such as creativity, insight, and mathematical intuition are beyond the reach of artificial systems due to their dependence on non-computable processes. I have registered for the conference (online only: $75). I look forward to hearing updates of Penrose's arguments during the event.

Registration: **https://www.eventbrite.com/e/neuroscience-needs-a-revolution-to-understand-consciousness-tickets-667018659767?aff=oddtdtcreator**

Posted on 7/21/2023

DSML Trend: New role of data scientists by embracing economic thinking

In response to the burgeoning influence of generative AI (GenAI), Bill Schmarzo has authored an insightful article titled "Next-Gen Data Scientist: Thinking Like an Economist" on *Data Science Central*. This article explores the parallels between economic principles and data science methodologies, underscoring the criticality of considering trade-offs, incentives, and resource allocation in data-driven decision-making processes. As outlined in a recent report by McKinsey, GenAI is projected to potentially automate up to 40% of the tasks currently executed by data science teams by 2025. These tasks, including data preprocessing, coding, and hyperparameter tuning, can be more effectively and efficiently accomplished through AI assistance. Nevertheless, it is crucial to recognize that GenAI lacks significant domain knowledge, setting it apart from human experts. By embracing an economic mindset, data scientists can optimize their strategies, thoughtfully prioritize projects based on **potential returns**, and skillfully communicate **insights** to stakeholders, thus providing robust decision support.

That's my take on it: For a long time, I have advocated against an excessive focus on data wrangling and programming within DSML (Data Science and Machine Learning) education. First, if the data collection protocol and data architecture are well-designed, there is no need to waste our time on data cleaning and data conversion. Second, complicated coding can, to some extent, hinder the discovery of insightful knowledge. Looking ahead, as AI progressively assumes responsibility for more low-level tasks, data scientists should concentrate their efforts on analytics and interpreting the implications of results for end-users.

Full article: https://www.datasciencecentral.com/next-gen-data-scientist-thinking-like-an-economist/?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+June+21st%2C+2023&utm_campaign=22072023

Posted on 7/20/2023

DSML trend: Generative AI fails to spark a strong demand for microchips

Today (July 21) Taiwan Semiconductor Manufacturing Co (TSMC) reported a sharp 23% decline in Q2 earnings, indicating that the recovery in the global semiconductor market is happening at a slower pace than anticipated. TSMC now projects a 10% revenue contraction for the full year 2023, reversing its previous forecast of slight growth. Its peer companies like Samsung and Micron have also posted earnings declines, further signaling weakness in the industry. Apparently, generative AI fails to spark a strong demand for microchips. Many generative AI services are delivered via servers equipped with NIVIDA GPUs. To sustain growth in these services, expansion of data centers is expected. However, generative AI's impact will take time due to need for advanced chip packaging. It is predicted that AI will not lead to a full-scale recovery in demand for semiconductors until 2024.

That's my take on it: Even though generative AI is more technologically advanced than the Internet, why hasn't generative AI created an economic boom similar to the Internet revolution of the 1990s? As far as I know, currently generative AI is still experimental. Unlike e-commerce, which had a straight-forward way to make money by selling products or services online, generative AI does not yet have proven business models. Companies are still figuring out how to commercialize the technology. Further, while generative AI shows promise for some business uses, it currently has limitations in understanding context and executing practical tasks. Put it bluntly, it is fun to chat with ChatGPT, Google Bard, and Claude 2, but information provided by these large language models is not 100% accurate, and it seems that widespread enterprise adoption will take more time. Nevertheless, I believe that it will happen soon! Those who are unprepared will be left behind.

Full article (subscription required): **https://asia.nikkei.com/Business/Tech/Semiconductors/TSMC-s-woes-signal-global-chip-slump-extending-throughout-year?utm_campaign=GL_asia_daily&utm_medium=email&utm_source=NA_newsletter&utm_content=article_link&del_type=1&pub_date=20230721123000&seq_num=9&si=14108111**

Posted on 7/14/2023

Recently the California State Board of Education has approved significant changes to the K-12 math curriculum by integrating data science and emphasizing real-world applications. In response to the growing importance of data science in society as well as

the need to prepare students for careers requiring strong data analytics and problem-solving skills, in the past two years the board has approved data science courses in many high schools. However, the University of California faculty committees that oversee high school courses accepted for admission to UC argued that Algebra 2 should not be replaced with data science, because this will under-prepare students who plan to major in STEM.

That's my take on it: In essence, the debate boils down to the purpose of education. As the name implies, data science is more empirical and data-driven, whereas theoretical mathematics is more logical and model-based. To equip students with job skills sought by the market or to solve real-life problems, it seems that data science is preferable to theoretical mathematics. However, advanced math is also necessary for developing abstract reasoning and symbolic processing. My question is: **why can't they keep both?**

Articles: **https://edsource.org/2023/uc-committee-changes-admission-standard-for-data-science-causing-confusion-over-math-framework/693892**
https://www.edweek.org/teaching-learning/california-adopts-controversial-new-math-framework-heres-whats-in-it/2023/07

Posted on 7/8/2023

On July 5, 2023 Open AI announced a new initiative called "Superalignment" that aims to resolve the alignment problem. According to Jan Leike and Ilya Sutskeer, the Chief Scientist and the Head of the Superalignment team, although superintelligence will be the most impactful technology that could help us solve many important problems, the vast power of superintelligence could also threaten humanity or even result in human extinction. In response, superaligment is introduced as a proactive process of ensuring that superintelligent AI will follow human intent. Their approach is to build a human-level automated alignment researcher to validate the resulting model and to spot problematic behaviors of an AI system.

That's my take on it: If you are not familiar with the alignment problem, "The Alignment Problem: Machine Learning and Human Values" (Christian, 2020) is an accessible introduction. This book discusses the ethical and psychological challenges when the goals of AI systems and human values are misaligned. When we instruct the AI system to complete a specific task, the system may attempt to achieve the goal at all costs and by any means, but the method may not be aligned with human interests and values. For example, if we ask AI to eliminate spam emails, it might delete all email accounts in order to attack the root cause of the problem. If a factor owner instructs the

AI system to produce paper clips using the most cost-effective way, all metals may be redeployed by th AI system to the paper clip factory, which would offset other priorities. The scenarios presented here are very simplistic. Unlike conventional computers that require pre-programming, machine learning is self-evolving. As AI becomes more advanced, its behaviors might become more unpredictable, and the consequences may far exceed our predictions. Can superalignment resolve or at least alleviate the alignment problem? It is very difficult, if not impossible, to **predict the unpredictable**.

OpenAI announcement: https://openai.com/blog/introducing-superalignment

Challenges and Criticisms of OpenAI's Approach to AGI:
https://www.youtube.com/watch?v=CkOoIFpHWM8

Posted on 7/7/2023

The International Telecommunication Union's (ITU) annual AI for Good Summit, which was held on July 6 and 7, 2023, aims to harness the power of AI to address global challenges and promote sustainable development. The conference brought together experts from various fields, including healthcare, climate change, and education experts, to discuss and explore AI applications. ITU is a Geneva-based United Nations agency that represents all 193 member states as well as over 900 companies, universities, and other organizations. And therefore, the AI for Good Summit is a truly global conference.

That's my take on it: Despite its representativeness, there is no formal declaration, negotiated statement, or decision announced by ITU. Although the discussion in the Summit has led to the creation of focus groups for developing new standards, as well as addressing the impact of AI-enabled androids on humans, it is very difficult, if not impossible, for rival countries that embrace different political ideologies and ethical standards to reach a consensus. I read the closing statement of the summit. Frankly speaking, it is very general and vague.

ITU Statement On The Closing Of The 2023 AI For Good Global Summit:
https://www.publicnow.com/view/8ABBE69C860E65230F0AB9858CFA4696EBE691FA?1688759234
ITU AI For Good Global Summit 2023 Press conference:
https://www.youtube.com/watch?v=m9IN14e-PLk

Posted on 7/6/2023

Today, Tesla showcased their AI products at the World Artificial Intelligence Conference in Shanghai, China, along with 400+ exhibitors. In addition to its Autopilot (Fully Self-Driving) cars, Tesla displayed a prototype of its Optimus robot. The Optimus has the latest technology of the same origin as Tesla vehicles, including a fully self-navigation computer and a Tesla Vision visual neural network. The Tesla humanoid robot is 172 centimeters tall and weighs 56.6 kilograms, which is no different from a normal adult. Like human joints, the robot's whole body has 28 degrees of freedom. Its hand has 11 degrees of freedom, and therefore it has a high degree of flexibility and dexterity. As a result of its powerful motors, **the robot is capable of lifting a piano with just one hand**. This humanoid robot can also walk, climb stairs, squat, and pick up objects, and it already has the capability to **protect itself and other people**. In the future, robots may cook, mow the lawn, care for the elderly, or replace humans in dangerous and boring factory jobs.

That's my take on it: According to some commentators, Tesla's robots are ahead of Boston Dynamics because Boston Dynamics' robots require preprogramming for movement, while Tesla's can evolve through machine learning. I know what's on your mind. Could a self-learning robot harm humans at some point if it becomes out of control? People may even wonder whether the Optimus will be weaponized since it can lift a piano with one hand and defend itself. Is it going to be used for evil purposes? Regulations should be discussed as early as possible.

P.S.: I want the Optimus if I can afford one. Currently I am moving from LA to Honolulu. I need a robot that can lift heavy objects for me!

English text:
https://www.teslaoracle.com/2023/07/06/tesla-cars-optimus-robot-world-ai-conference-shanghai-elon-musk-ai-chinese-cars/

Chinese text:
https://www.chinatimes.com/realtimenews/20230706005385-260412?utm_source=likr&utm_medium=web_notification&utm_content=VOY0&avivid_manual=4

Posted on 6/30/2023

My previous message highlights the importance of data quality. Specifically, Microsoft is able to develop a powerful large language model using fewer parameters when textbook-quality data are used. In a similar vein, on June 27 *Nature* published an article entitled "How to make your scientific data accessible, discoverable and useful". Based on the input from expert data scientists, the article suggested the following best practices for publishing usable, high-quality data:

- Craft metadata: Publish Metadata, which are data of data, such as the timestamp and geolocation details. It can make data findable, accessible, interoperable and reusable.

- Over-share: Data analysts seldom analyze raw data only. At the end of the project, there might be numerous versions of the data with derived and transformed variables. Share all versions.

- Embrace standards: e.g., share data on GitHub.

- Consider the format: Use common formats, such as CSV. Avoid PDF.

- Include the code used for data analysis.

- Think accessibility.

Full text: **https://www.nature.com/articles/d41586-023-01929-7?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+June+30th%2C+2023&utm_campaign=01072023**

That's my take on it: While these guidelines are easy to follow, some researchers overlook them. For instance, all online survey engines (e.g., Qualtrics, SurveyMonkey) today collect timestamps and geolocation data. Time stamps provide useful information for determining whether respondents read the survey questions carefully based on their response times. Using geolocation data, it is possible to determine whether the respondents are representative of the population and whether they lied about their true identity. However, very few researchers use this information to evaluate the quality of data. Both traditional statistics and data science can benefit from the above practices. **Meta-analysis** often includes only high-quality studies. This will facilitate the screening process if metadata and different versions of data are included in the publication.

Posted on 6/28/2023

In a $1.3 billion deal announced two days ago (June 26), **Databricks**, an industry leader in data management, will acquire **MosaicML**, a generative AI platform that empowers enterprises to build their own AI. According to Databricks, the rationale of this acquisition is: “Today, virtually every organization is exploring how best to use generative AI and LLMs, and every leader is considering how they leverage these new innovations while retaining control of their most precious resource: their data.”

That’s my take on it: The technology industry is undergoing a wave of AI acquisitions. In early May Databricks acquired Okera, a data governance platform with a focus on AI. In late May Snowflakes acquired Neeva, an AI-enabled search engine that could enhance its cloud data management capabilities. Aside from acquisitions, forming partnerships is another common AI strategy. Yesterday (June 27) at Snowflake Summit 2023 SAS announced that SAS Viya’s AI-based decision-support capabilities have been incorporated into the Snowflake Data Cloud with Snowpark Container Services. Needless to say, those who failed to catch the wave and operated in silos may eventually lose out to more powerful competitors. Hence, I believe it is imperative to teach students (our future workforce) how to integrate various tools, or at least understand the "big picture."

Full articles:
https://www.databricks.com/company/newsroom/press-releases/databricks-signs-definitive-agreement-acquire-mosaicml-leading-generative-ai-platform
https://finance.yahoo.com/news/sas-runs-sas-viya-ai-191700223.html
https://techcrunch.com/2023/05/24/snowflake-acquires-neeva-to-bring-intelligent-search-to-its-cloud-data-management-solution/
https://techcrunch.com/2023/05/03/databricks-acquires-ai-centric-data-governance-platform-okera/

Posted on 6/16/2023

In a recent benchmark study, the Futurum Group compared SAS Viya and several open source software packages, such as Apache SparkML, H2O, and Ranger, in terms of scalability and performance. In this study, random forest, gradient boosting, including LightGBM and XGBoost, linear regression, and logistic regression were rigorously tested on big data. It was found that for running machine learning with high-dimensional data, SAS Viya is on average 30 times faster than all other competitors across 1,500 tests. Specifically, SAS Viya solution delivered results in under 12 minutes on a dataset

containing over 300 million data points, while SparkML and another rival failed to deliver results after running for hours. For running traditional procedures, such as linear regression and logistic regression, SAS Viya ran faster in 49 out of the 50 tested configurations.

Full report:
https://futurumgroup.com/wp-content/uploads/2023/06/Lab-Insight_Performance-at-Scale_Comparing-AI-ML-Performance-of-SAS-Viya-vs-Alternatives.pdf

That's my take on it: Having used both open source and proprietary software applications, I do not believe that we should side with one camp or the other. Open source is touted as a great tool, but in my view its advantages are overstated. I am not surprised by this benchmark result. While developers of open source are a loose conglomerate, resulting in incompatibility and redundancy, R&D in commercial corporations, such as SAS, IBM, and Microsoft, are coordinated and thus coherent. As a matter of fact, a lack of financial incentives makes it difficult for volunteers to devote substantial time and effort to optimizing machine learning codes. OpenAI has taken the world by storm with its ChatGPT, but few people know that OpenAI first adopted an open source model in an attempt to liberate people from big tech monopolies. However, Cade Metz, the author of *Genius Maker*, made a harsh comment by saying "It (OpenAI) was an idealistic vision that would eventually prove to be completely impractical." Yann LeCun, the inventor of CNN, even predicted that this model was doomed to fail at the beginning. Within a few years, OpenAI became a for-profit, closed-source company.

Posted on 6/13/2023

Google recently announced several new DSML products, some of which are still in the experimental stage. One of these innovative products is **StyleDrop**, an AI-enabled art tool that allows users to generate images in a **consistent** style. In StyleDrop, the user can easily transfer an original image with a desired style to a new image while preserving its unique characteristics. Furthermore, Google announced last week that it has partnered with Salesforce, the parent company of **Tableau**, to integrate data analytics into its cloud platform. Specifically, Google and Salesforce plan to integrate Data Cloud and BigQuery to enable businesses to create unified customer profiles in a more efficient way.

YouTube video: https://www.youtube.com/watch?v=q_ebiphq2Pk&t=735s
Article: https://techcrunch.com/2023/06/07/google-cloud-and-salesforce-team-up-to-bolster-ai-offerings/

That's my take on it: Even though Midjourney and Stable Diffusion are good at generating art, neither produces a consistent style. Google is so smart that it doesn't follow a "me-too" strategy. Adding features similar to those offered by Midjourney and Stable Diffusion is unlikely to entice customers away from those established generative art platforms. But customers will give Google a try for something new. By the same token, it will be very difficult for Google Cloud to compete with Amazon Web Services in terms of cloud computing capabilities. Rather, it will be more beneficial for Google to leverage **data visualization** through the strategic partnership with Salesforce.

Posted on 6/2/2023

Two days ago, Amazon announced it would pay more than $30 million in fines to settle allegations that its Alexa voice assistant and Ring doorbell camera violated privacy laws. A lawsuit filed by the Federal Trade Commission (FTC) alleges that Amazon kept records of children's conversations with Alexa in violation of privacy laws, while another alleges that its employees viewed recordings from Ring cameras without consent. Amazon would also be prohibited from using the predictive models built upon these data. Despite the FTC's rulings, Amazon argued that it had not broken any laws.

Full article: https://www.npr.org/2023/06/01/1179381126/amazon-alexa-ring-settlement

That's my take on it: Big data analytics and machine learning have made Amazon, Google, and Facebook researchers better psychologists and sociologists than academicians, since the former group can access oceans of **behavioral data** collected in naturalistic settings. When we are unaware of their data collection, these data tend to reveal our true character and behaviors. Needless to say, invasion of privacy is a concern. However, before we point our fingers to Amazon, Google, and Facebook, we should not forget that many well-known psychological studies in the past, such as Milgram's and Zimbardo's studies, were conducted in the absence of IRB approval or are considered unethical today. It will take some time to fine-tune the ethical standards of behavioral data.

Posted on 6/2/2023

In a recent article (May 31, 2023) published in *Towards Data Science*, data scientist Col Jung argued that organizations should migrate away from traditional data lakes and adopt a data mesh approach. In Jung's view, organizations using old-fashioned data

warehouses are trapped in a mess of data systems connected by innumerable data pipelines. Data lake was introduced as a solution by centralizing diverse data into a hub, but “data lake monsters” are “over-promised and under-realized.” In the era of big data, all analytical questions rest on the shoulders of the data lake team. Consequently, the central data team encountered tremendous scalability problems and became inefficient. To rectify the situation, in 2019 Dehghani proposed data mesh as the next-generation data architecture embracing a decentralized approach. Instead of transferring data to a centralized lake, a data mesh allows **domain-specific teams** to control and deliver data as products, promoting easy interoperability and accessibility across the organization.

Full article: https://towardsdatascience.com/from-data-lakes-to-data-mesh-a-guide-to-the-latest-enterprise-data-architecture-d7a266a3bc33

That’s my take on it: In my experience many requests to the central data office are simple questions, but as CoI said, the data team is overwhelmed under the traditional centralized data architecture. The good news is: Data meshes facilitate **self-service data usage**, whereas data lakes do not. Is a decentralized system likely to result in chaos, with different people processing data differently? I don’t think so. Since data meshes are owned by different entities, they require stricter formatting, metadata fields, discoverability, and **governance standards**.

Posted on 5/26/2023

Recently (May 21) *Analytics Insight* posted a report highlighting the top 10 highest-paying countries in need of data scientists. The order is as follows: USA, Switzerland, UK, Australia, Israel, India, Canada, China, Italy, and France. Take the US as an example. The median salary at the entry-level starts at US$95,000. For experienced data scientists the median pay could be as high as US$165,000.

Article: https://www.analyticsinsight.net/top-10-highest-paying-countries-in-need-of-data-scientists-in-2023/?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+May+26th%2C+2023&utm_campaign=27052023

That’s my take on it: It surprises me that some countries aren't included in the list. The high-tech sectors of Germany, Japan, South Korea, and Taiwan, for instance, are vibrant and fast-growing, so data scientists should be in high demand. I might be missing something or the survey data are incomplete. Out of curiosity, I looked up employment

information for data scientists overseas. According to Glassdoor, the average salary of a data scientist in Tokyo is US$55,831 (Yen 7851192), while that in Germany is US$72,419 (Euro 67500). The figures were not adjusted for purchasing power. Nonetheless, I will stay in America!

Posted on 5/18/2023

Today (5/18) UC Berkeley announces that it will open a **College of Computing, Data Science and Society**, which is expected to be approved by The University of California Board of Regents. During the 2025-26 academic year, a new college building will house the data science major that was launched five years ago, along with other computer science degree programs. More than 89 campuses have access to the online curriculum, which includes assignments, slides, and readings. UC Berkeley also has disseminated its curriculum to other colleges and universities for free. Beginning this fall, there will be UC Berkeley-led data science classes at six California community colleges, four Cal State campuses, and Howard, Tuskegee, Cornell, Barnard, and the United States Naval Academy.

Full article: https://www.latimes.com/california/story/2023-05-18/uc-berkeley-spreads-the-gospel-of-data-science-with-new-college-free-curriculum

That's my take on it: It is important to point out that this data science conglomerate is not a result of a merely top-down decision; rather, it happens due to huge faculty and student demand. Data science has risen to the fourth most popular major at UC Berkeley in just five years. Faculty and students at UC Berkeley are aware of the importance of data science. In June 2023, UC Berkeley received three gifts totaling $75 million for supporting the construction of the data science center. Two of the gifts are from the current Berkeley faculty. Needless to say, good leaders must pay attention to bottom-up movements; they must be active listeners who can constantly learn and adapt to change.

Posted on 4/22/2023

Recently a German photographer named Boris Eldagsen refused the Sony world photography awards after admitting to being a "cheeky monkey" by generating the award-winning image using AI. Eldagsen used a pseudonym to submit the AI-generated photo, and the judges selected it as the winner. In an open statement, Eldagsen wrote, "We, the photo world, need an open discussion. A discussion about what we want to consider

photography and what not. Is the umbrella of photography large enough to invite AI images to enter – or would this be a mistake?... AI images and photography should not compete with each other in an award like this. They are different entities. AI is not photography. Therefore I will not accept the award."

Article: **https://amp.theguardian.com/technology/2023/apr/17/photographer-admits-prize-winning-image-was-ai-generated?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+April+21st%2C+2023&utm_campaign=22042023**

That's my take on it: Does AI-enabled imaging qualify as photography? It depends. This type of debate is not entirely new. When digital photography was introduced, some traditional photographers disliked images manipulated by computer software, such as Adobe Photoshop. They argued that those images are no longer authentic and natural. In the past, photographers used a variety of filters and darkroom techniques to enhance their images. For me, a tool is a tool is a tool, no matter whether the tool is physical, digital, or AI-enabled. The image, however, should not be considered photography if it was entirely created by AI without input from the photographer.

Posted on 4/11/2023

In response to the arrival of ChatGPT, recently a group of prominent AI researchers signed an open letter to call for slowing down AI developments that can pass the Turing Test. The Turing test measures a machine's ability to exhibit intelligent behavior that is indistinguishable from human behavior. Yoshua Bengio is one of the leading experts of deep learning who co-signed the letter.

Bengio wrote, "I found it appropriate to sign this letter to alert the public to the need to reduce the acceleration of AI systems development currently taking place at the expense of the precautionary principle and ethics. There is no guarantee that someone in the foreseeable future won't develop dangerous autonomous AI systems with behaviors that deviate from human goals and values. The short and medium term risks –manipulation of public opinion for political purposes, especially through disinformation– are easy to predict, unlike the longer term risks –AI systems that are harmful despite the programmers' objectives, and I think it is important to study both."

Full article: **https://yoshuabengio.org/2023/04/05/slowing-down-development-of-ai-systems-passing-the-turing-test/**

That's my take on it: Bengio cited the precautionary principle to argue for slowing down AI development. According to the precautionary principle, if an action could potentially cause harm to the public or to the ecology, without scientific consensus, the burden of proof that it is not harmful is on the shoulder of the party taking the action. Because most AI developers are not philosophers of ethics or legal experts, it places a heavy burden on them. I think there is no need to slow down AI development; instead, experts from different disciplines should be part of every development team, and there should be opportunities to engage in open debates and discussions regarding AI ethics.

Posted on 3/3/2023

Recently Harvard Business Review (HBR) reported that many retail companies have not taken advantage of **advanced data analytics** to improve their business. There are exceptions: Walmart, Amazon, and a few others. The 25 best-performing retailers during the pandemic generated 83% more profit than laggards and captured more than 90% of the sector's market capitalization gains. By interviewing 24 business leaders, HRB unveiled six sticking points as follows:

1. **Culture**: Typically, companies have a risk aversion problem and lack a clear goal for implementing analytics.
2. **Organization**: Many companies struggle to strike a balance between centralization and decentralization.
3. **People**: Very often the analytics function is managed by people who have no understanding of the industry.
4. **Processes**: Businesses do not have unlimited resources at their disposal.
5. **Systems**: Legacy systems are still serving many firms today.
6. **Data**: Data are often scattered throughout the firm in silos and not managed in an organized manner.

Full article: https://hbr.org/2023/02/why-retailers-fail-to-adopt-advanced-data-analytics?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+March+3rd%2C+2023&utm_campaign=04032023

That is my take on it: Some interviewees believe the bigger issue is people. People who know about all other issues are willing to dedicate resources to solving them, despite their presence. Sadly, this is not always the case. **William Cleveland** and **John**

Chambers were pioneers in data science. Many years ago they both proposed that data science should be interdisciplinary, incorporating domain knowledge. Agree!

Posted on 2/24/2023

ChatGPT stories continue to dominate mass media and social media, and probably you already received these stories from many channels. Therefore, I would like to draw your attention to something else. Two days ago Google unveiled its ***2003 data and AI trends report***. In addition to Google Cloud, Google also suggests a vast array of technologies to companies that planned to enhance its AI and cloud computing infrastructure:

- Aiven
- C3AI
- Confluent
- Collibra
- Databricks
- Datametica
- Elastic
- Fivetran
- MongoDB
- Nivida
- Qlik
- Quantiphi
- Salesforce
- SAP
- Striim
- ThoughtSpot

A month ago InsideBigData complied ***the IMPACT 50 list for Quarter 1, 2023***. According to InsideBigData, “These companies have proven their relevance by the way they’re impacting the enterprise through leading edge products and services.” The top 20 are:

- Open AI
- Nvidia
- Google AI
- Amazon Web Services
- Hugging Faces
- H20.ai
- Databricks
- Microsoft AI
- Intel AI

- Neural Magic
- Snowflake
- SAS
- Qlik
- Neo4j
- Alien Institute for AI
- TigerGraph
- Anaconda
- Domino Data Lab
- Hewlett Packard Enterprise
- Cloudera

Full report of Google: https://services.google.com/fh/files/misc/data_and_ai_trends.pdf
Full article of InsideBigData: https://insidebigdata.com/2023/01/17/the-insidebigdata-impact-50-list-for-q1-2023/

That's my take on it: Although the selection criteria are subjective and might even be biased, data scientists and DSML educators should still take them seriously. As you can see, the list of these most promising and most impactful tech companies consists of both fairly new companies and mature companies (e.g., Microsoft, Hewlett Packard, Intel, SAP, SAS…etc.). However, some established tech giants are absent from the list (e.g., IBM, Oracle…etc.). Both IBM and Oracle are not even among the top 50. It is understandable. Despite several decades of development, some of its products have made little progress. The rule in academia is: publish or perish. In the era of AI and big data, the choice facing companies is: innovate or perish.

Posted on 2/23/2023

DSML trend: Google Gemini image generator is temporarily offline due to historical inaccurate images

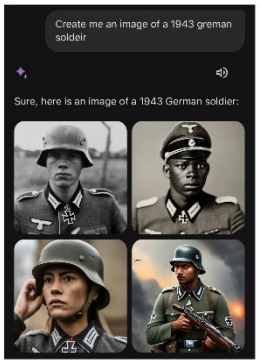
Currently Google Gemini image generator is offline. If you try to enter a prompt to request a portrait from Gemini, the following message will pop up: "We are working to improve Gemini's ability to generate images of people. We expect this feature to return soon and will notify you in release updates when it does." The issue arose when Google Gemini, in an attempt to address AI biases concerning race and gender, produced images that were factually incorrect. For example, prompting it with **"1943 German soldier"** yielded images that included black and Asian female soldiers, which is historically inaccurate. Similar problems were seen with prompts that resulted in black Vikings, a female pope, women

in the NHL, the Google founders as Asian men, and non-white depictions among the U.S. Founding Fathers.

That's my take on it:
When I used the same prompt in Midjourney and Stable Diffusion, their outputs, while not perfectly historically accurate (such as in uniform details), did not feature any non-white characters (see attached). These incidents reflect a broader trend in technology where solutions can sometimes create new challenges. For instance, during DALL.E's early development, OpenAI implemented filters to remove sexualized images of women, but this inadvertently led to a reduced representation of women in its outputs. Social media platforms, designed to foster connections and tailor user experiences, have faced criticism for enabling misinformation, echo chambers, and social divides. There is no fool-proof technology. Nevertheless, I trust that in an open society scientific inquiry is a self-correcting process in the long run.

Links:
https://variety.com/2024/digital/news/google-gemini-ai-image-racial-inaccuracies-nazi-soldiers-1235919168/

https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical



Posted on 2/10/2023

Facing the pressure from Open AI's ChatGPT, Google is devoting efforts to reassure the public that its AI technology is still promising. However, the performance of its own chatbot named Bard is so embarrassing that investors lost confidence. Bard, which was released on Twitter on Monday, tried to answer an inquiry about discoveries from the James Webb Space Telescope. According to Bard, the telescope was the first to

photograph a planet outside the solar system, but indeed this milestone was accomplished by the European Very Large Telescope in 2004. This mistake was spotted by astronomers on Twitter. Consequently, Alphabet's shares dropped more than 7% on Wednesday, losing $100 billion of its market value.

Full article: **https://www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/**

That's my take on it: As a matter of fact, ChatGPT also made many factual errors. For example, when a history professor asked ChatGPT to explain the Joseph Needham thesis, it offered a response as: "the scientific and technological achievements of the West were only possible because of the transmission of scientific and technological knowledge from China to the West." It is completely wrong! Indeed, Joseph Needham was curious about why ancient China failed to develop modern science. My friend who is a math professor in Hong Kong also found that some answers offered by ChatGPT are unsatisfactory. I guess people are more forgiving to ChatGPT because it is the first of its kind.

Posted on 2/4/2023

On January 30, 2023, the *Retraction Watch* published an exclusive report on Hao Li's research misconduct. Hao Li, the pioneer of Deepfake technology that can fabricate video, has won numerous awards for his AI-based innovations in imaging technology. According to the *Retraction Watch*, two of his articles published in *ACM Transactions on Graphics* will be retracted due to falsification of data. One of his articles is based on a presentation at the *ACM computer graphics conference SIGGRAPH 2017 Real Time Live* (The recording is available on YouTube). In the presentation, Li and his colleagues showed that his software could generate a 3D image based on a picture taken with a webcam in just a few seconds. However, later it was found that those 3D images were built and preloaded into the computer before the presentation. Li denied any wrongdoing, saying that preloading the 3D images was allowed by the conference.

Full article: **https://retractionwatch.com/2023/01/30/exclusive-deepfake-pioneer-to-lose-two-papers-after-misconduct-finding/**

Youtube video of Li's presentation: https://www.youtube.com/watch?v=hpuEdXn_M0Q

Li's ACM articles:

https://dl.acm.org/doi/10.1145/3130800.31310887
https://dl.acm.org/doi/10.1145/3098333.3107546

That' my take on it: Despite winning the "Best in Show" award at the ACM conference, Li's presentation is a fraud! In fairness, Li's misconduct was not on the same scale as Elizabeth Holmes'. Li had a working prototype and he made it appear to be more efficient, whereas Holmes lied about a promising blood-testing technology that never existed and was physically impossible. Nonetheless, it is not unusual for high-tech companies to use the strategy of "fake-it-until-you-make-it." For example, Microsoft in the past announced several "vaporware" products that didn't exist in order to keep customers from buying well-developed technologies from competitors. In the same vein, many companies use the buzzword "AI" in their product names, but whether the technology is truly AI remains to be determined.

Posted on 2/4/2023

There has been a hot debate in academia about the use of ChatGPT. In December last year, ChatGPT was included as one of 12 authors on a preprint about using the tool for medical education posted on the medical repository *medRxiv*. According to *Nature*, ChatGPT was cited as a bylined author in two preprints and two articles in science and health published in January 2023. All of the articles have an affiliation with ChatGPT, and one even gives an email address for a supposed nonhuman "author." *Nature* explained that inclusion of ChatGPT as an author was a mistake and the journal will fix it soon. However, PubMed and Google Scholar have already indexed these articles and these nonhuman "authors." *Nature* has since set forth a policy guiding how large-scale language models can be used in scientific publications, prohibiting naming them as authors. To address this latest technological concern, recently the *Journal of the American Medical Association* (JAMA) updated its instructions for authors: Artificial intelligence, language models, machine learning, and similar technologies are not eligible for authorship. When these tools are utilized to generate content or assist in the writing or preparation of manuscripts, the authors are responsible for the integrity of the content generated by these tools and must clearly state the use of AI in the manuscript.

Full articles:https://jamanetwork.com/journals/jama/fullarticle/2801170
https://www.nature.com/articles/d41586-023-00107-z

That's my take on it: It appears that faculty and student policies regarding ChatGPT are vastly different. The inclusion of any content generated by ChatGPT in a paper is strictly prohibited by many universities and violation of the policy is treated as academic

dishonesty. On the contrary, JAMA accepts AI-generated content as long as the author verifies the information and documents it in the Acknowledgment section or the Methods section of the paper. I guess it is based on the implicit assumption that mature adults are more responsible than young students. In my opinion, it is not necessarily true. This type of “discriminatory” policies may eventually lead to discontent among students. Rather than setting two sets of policies, it would be better to create one standardized policy for all, and provide workshops on ethical AI use to both groups.

Posted on 2/3/2023

Yesterday (Feb 2, 2023) an article posted on KDNuggets introduces ten free machine learning courses offered by top universities, including UC Berkeley, Carnegie Mellon University, Stanford University, Caltech, Cornell University, University of Toronto, MIT…etc. It is noteworthy that these are just not one-hour seminars; rather, the duration of these comprehensive courses is between 20 and 60 hours. More importantly, some of these courses are taught by very prominent scholars in the field, such as Andrew Ng.

Full article: https://www.kdnuggets.com/2023/02/10-free-machine-learning-courses-top-universities.html

That’s my take on it: According to the May 2022 report compiled by the Institute for Advanced Analytics at North California State University, there are about 353 graduate programs in data science and machine learning in the US. Additionally, there are many free courses in the market and the preceding list is only the tip of the iceberg. No doubt the competition is very intense, and therefore program designers must think outside the box to stay ahead of the curve.

Posted on 2/2/2023

A week after ElevenLabs opened its voice-cloning platform to the public, the startup says it may need to rethink that openness amid increasing instances of voice-cloning misuse. The Elevenlabs speech synthesis and voice cloning software modules can mimic any accent and speaking tone and can be used for newsletters, books, and videos. Piotr Dabkowski, a former Google machine learning engineer, and Mati Staniszewski, an ex-Palantir deployment strategist, founded the company in 2022. After the software was

found to generate homophobic, transphobic, violent, and racist statements from celebrities, the company addressed the issue on Twitter.

Full article: https://www.theverge.com/2023/1/31/23579289/ai-voice-clone-deepfake-abuse-4chan-elevenlabs

That's my take on it: This issue is not entirely new. For a long time many people have been using Deepfake software to impersonate celebrities in video, and DeepNude to remove clothing from images. It is surprising that ElevenLabs did not learn anything from DeepFake or DeepNude, but instead released its speech synthesis and voice cloning programs without taking any proactive measures. In contrast, developers of AI-enabled art generators are more thoughtful. For example, DALLE-2 forbids using any public figure's name in the input process, while Midjourney disallows any potentially problematic terms, such as "sexy" or "bloody." We cannot be too optimistic and naïve about human nature!

Posted on 1/24/2023

In spite of a mass layoff (10,000 employees), Microsoft recently announced a $10 billion investment in Open AI, the company that developed ChatGPT and DALLE-2. Microsoft's investment will allow OpenAI to accelerate its research since all of its models are trained in Microsoft Azure. In return, Microsoft will receive a boost to its Azure cloud and even catch up with Amazon Web Services.

Full article: https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai

That's my take on it: Currently Amazon Web Services dominates the cloud computing market. However, Open AI can undoubtedly improve the functionality of Microsoft Azure. While AWS does not have a powerful AI partner like Open AI, its Sagemaker provides powerful predictive modeling capabilities. A long time ago, Microsoft and SAS Institute formed a partnership to offer cloud-based data analytics. It is my belief that this fierce competition in machine learning, cloud computing, and data science will drastically change the landscape of these fields in the near future. Be sure to stay tuned!

Posted on 1/20/2023

Today I read an interesting article entitled "Is artificial intelligence a threat to Christianity?" posted on *Patheos*. The article contains many insightful points, and I will only highlight

one. According to Keith Giles, the author of the article, “In fact, this fear of creating an AI that is “more intelligent than humans” isn’t even what we should be most afraid of. As one former top social media tech executive was quoted as saying in the excellent NETFLIX documentary, The *Social Dilemna*, we shouldn’t be afraid of creating an AI that eventually exceeds human intelligence, what we should be afraid of is the fact that we’ve already created machine learning programs that know how to overcome our human weaknesses.”

Article: https://www.patheos.com/blogs/keithgiles/2023/01/is-artificial-intelligent-a-threat-to-christianity/?utm_source=Newsletter&utm_medium=email&utm_campaign=Best+of+Patheos&utm_content=57&lctg=252615&rsid=Legacy&recipId=252615&siteId=7DF2956C-D2F1-40D4-A777-98E450E58360

That’s my take on it: Last evening in my class I told my students that I like machine learning a lot. Machine learning has the ability to learn very quickly, as its name implies. With the right data, the algorithm can improve, and it won't make the same error again. On the contrary, humans (including myself) are so stubborn that we let our cognitive and emotional weaknesses affect our judgment and behavior. We fear AI partly because we are envious of it.

Posted on 1/20/2023

Today I read an interesting article entitled “Is artificial intelligence a threat to Christianity?” posted on *Patheos*. The article contains many insightful points, and I will only highlight one. According to Keith Giles, the author of the article, “In fact, this fear of creating an AI that is “more intelligent than humans” isn’t even what we should be most afraid of. As one former top social media tech executive was quoted as saying in the excellent NETFLIX documentary, The *Social Dilemna*, we shouldn’t be afraid of creating an AI that eventually exceeds human intelligence, what we should be afraid of is the fact that we’ve already created machine learning programs that know how to overcome our human weaknesses.”

Article: https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit

That’s my take on it: Last evening in my class I told my students that I like machine learning a lot. Machine learning has the ability to learn very quickly, as its name implies. With the right data, the algorithm can improve, and it won't make the same error again. On the contrary, humans (including myself) are so stubborn that we let our cognitive and

emotional weaknesses affect our judgment and behavior. We fear AI partly because we are envious of it.

Posted on 1/20/2023

With over 477 million items, Getty Images is one of the largest visual media companies in the world, offering stock images, videos, and music to business and individual clients. Recently Getty Images announced that it is suing Stability AI, a company that enables users to generate images using its machine-learning software module, Stable Diffusion. Getty Images accused Stability AI of training its algorithms by unlawfully extracting images from the Internet, including stock images owned by Getty. Getty claimed that the company is not seeking financial damages or trying to stop the distribution of AI-art technology; rather, it attempts to push for laws and regulations that respect intellectual property.

Article: **https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit**

That's my take on it: Getty Images' reaction is understandable. It will not be necessary for illustrators or other users to buy stock images from Getty or other suppliers when they are able to generate images using AI. For example, the Atlantic published a report by Charlie Warzel in 2022 right after Midjourney was released, another AI art generation program. The report depicts two images of Alex Jones, the founder of InfoWar. Later Warzel apologized. "This was entirely my fault…Instead of selecting a photo or illustration from Getty Images to go with the story, as I do for most of my newsletters, I decided to try something different and use an AI art tool to come up with the story's accompanying image," says Charlie Warzel.

It is interesting to note that Getty Images is not suing Midjourney and DALLE-2. There is an obvious reason for omitting DALLE-2. While Stability AI uses an open-source model, Open AI, which developed DALLE-2, did not disclose its mechanics. In the absence of ample evidence, attorneys have a difficult time building a case. However, I don't understand why Getty Images is not targeting Midjourney. Do you know why?
Posted on 1/13/2023

Eight major Australian universities have announced that they have changed their assessment formats as a result of several cases in which students turned in papers generated from ChatGPT. The University of Sydney, for instance, has revised its academic integrity policy to explicitly state that using artificial intelligence to create content is cheating. The Australian National University has changed assessment designs, such

as shifting emphasis on laboratory activities and fieldwork, as well as using time exams and oral presentations.

Full article: https://www.theguardian.com/australia-news/2023/jan/10/universities-to-return-to-pen-and-paper-exams-after-students-caught-using-ai-to-write-essays?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+January+13th%2C+2023&utm_campaign=14012023

That's my take on it: This issue is not entirely new. Before the introduction of ChatGPT and other AI tools, Wolfram products, such as Mathematics and Wolfram Alpha, are capable of solving complex math problems. These tools are also used by students to cut corners, say math and statistics professors. The widespread availability of Google and other search engines has led to many students turning in "instant" papers that reference many websites. Nonetheless, Wolfram, Google, and now ChatGPT are here to stay. The solution is not to ban them. Instead, we should teach students how to use these tools ethically.

Posted on 1/9/2023

ChatGPT, an Open AI language module released on November 30, 2022, is capable of writing articles, generating codes, and solving complex math problems. As expected, the introduction of ChatGPT has triggered widespread resistance. On Jan 5, 2023 the International Conference on Machine Learning (ICML) announced that it bans authors from using AI tools like ChatGPT to write scientific papers, unless the produced text is a part of an experimental analysis. It is important to point out that this ban applies only to text generated entirely by AI-enabled language models, but does not apply to papers "coauthored" by humans and AI. In a similar vein, Stack Overflow also banned users from submitting answers created using ChatGPT last year, while the New York City Department of Education blocked access to this tool just last week.
"With a tool like this at their fingertips, it could muddy the waters when evaluating a student's actual writing capabilities because you're giving kids potentially a tool where they could misrepresent their understanding of a prompt," says Whitney Shashou, founder and advisor at educational consultancy Admit NY.

Full articles:
https://www.theverge.com/2023/1/5/23540291/chatgpt-ai-writing-tool-banned-writing-academic-icml-paper

https://www.forbes.com/sites/rashishrivastava/2022/12/12/teachers-fear-chatgpt-will-make-cheating-easier-than-ever/?utm_campaign=31122022&utm_content=Data+Science+Insider%3A+December+30th%2C+2022&utm_medium=ONTRAPORT-email-broadcast&utm_source=ONTRAPORT-email-broadcast&utm_term=Newsletter&sh=3d9f6aa31eef

That's my take on it: Any new technology could lead to some unintended consequences. As you might already know, some papermill "companies" provide users with "publishing" services. It is estimated that about 1% of articles archived in PubMed contain questionable content. With the advance of AI tools like ChatGPT, it will be much easier for authors to produce instant articles. At the present time, I am unaware of any academic journal that prohibits submissions generated by AU.

ChatGPT also simplifies the process of writing for students. As of right now, my university does not have an academic honesty policy regarding artificial intelligence.

While Turnitin and SafeAssign can detect plagiarism, they cannot tell the difference between human-written and AI-generated text.

It is not my intention to oppose ChatGPT. As an initial research tool, I find this tool perfectly acceptable. Authors should, however, verify the information provided rather than blindly trusting the results. I recommend that at least 80% of the final paper should be written by a human author to ensure its originality.