Posted on 12/8/2022

Today I attended the last session of "Statistical wars and their casualties." One of the speakers is Aris Spanos (Virginia Tech) and the title of his presentation is "Revisiting the two cultures in statistical modeling and inference." In the talk he outlined several statistical paradigms as follows:

1. Karl Pearson's descriptive statistics
2. Fisher's model-based statistical induction
3. Graphical causal modeling
4. Non-parametric statistics
5. Data science and machine learning

At the end he discussed the difference between the Fisherian school and the data science approach: the paradigm shift from the Fisherian school to data science "reflects a new answer to the fundamental question: What must we know a priori about unknown functional dependency in order to estimate it on the basis of observations? In Fisher's paradigm the answer was very restrictive – one must know almost everything…machine learning views statistical modeling as an optimization problem relating to how a machine can learn from the data."

Nonetheless, Dr. Spanos warned against overhyping data science. For him doing data science is returning to the Pearsonian tradition that emphasizes describing the data at hand. Many people go into the discipline by learning Python without knowing statistical details. As a result, data science became a black box, and thus he is afraid that many decades later we will try to figure out what went wrong again.

In his talk entitled "Causal inference is not statistical inference," Jon Williamson (University of Kent) asserted that a broader evidence base from triangulation is more important than successful replication of the results because successful replication might replicate the bias in previous studies.

Seminar website: https://phil-stat-wars.com/workshop-the-statistics-wars-and-their-casualties/

That's my take on it:

1. I agree that the Fisherian model-based approach is very restrictive because it assumes you know to which the theoretical sampling distribution the sample belongs. However, I would compare data science and machine learning (DSML) to the school of exploratory data analysis (EDA) founded by John Tukey and the resampling approach developed by Elfron et al., rather than the Pearsonian legacy. By unpacking the philosophy of these paradigms, one can see that both DSML and

EDA emphasize pattern-seeking, and today resampling methods, such as cross-validation and bootstrapping, are embedded in many DSML methods.

2. We should do both triangulation and replication. I don't think one is more important than the other. Machine learning is a form of internal replication in the sense that the data set is partitioned into numerous subsets for repeated analyses. In boosting the subsequent models can correct the bias of the previous models, and thus this type of replication will not inherit the bias.

Posted on 11/11/2022

Two days ago, Meta (formerly Facebook) announced a massive layoff in the company, and as a result, 11,000 employees were terminated. Meta's CEO Mark Zuckerberg said that he planned to consolidate the company's resources into a few high-priority growth areas, such as the AI discovery engine, while giving up other less promising research endeavors. For example, the entire team named "Probability" was eliminated. The team was composed of 19 people doing Bayesian Modeling, nine people doing Ranking and Recommendations, five people doing ML Efficiency, and 17 people doing AI for Chip Design and Compilers. A former team member said it took seven years to assemble such a fantastic team.

Full article: https://venturebeat.com/ai/meta-layoffs-hit-entire-ml-research-team-focused-on-infrastructure/?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+November+11th%2C+2022&utm_campaign=12112022

That's my take on it: I don't worry about brain drain from the US to other countries. The US is still a magnet that attracts top-tier AI researchers and data scientists worldwide. Those former Meta researchers will likely be recruited by other high-tech giants, such as Google and Apple. Last year Professor Michael Gofman at the University of Rochester spotted a trend that high-tech titans and startups lured many DSML professors away from their faculty positions. Consequently, the knowledge gap between academia and industry was widened; transferring essential knowledge to students and colleagues was affected. Current massive layoffs in Meta, Twitter, and other high-tech giants might be an opportunity for colleges and universities to absorb those highly competent researchers.

Posted on 10/21/2022

Yesterday (10/20/2022) Vox announced its list of "Future Perfect 50," which recognizes 50 prominent intellectuals who made substantive contributions to a "perfect future." Demis Hassabis, the founder and the CEO of DeepMind, is on the list. He started programming at the age of 8 and founded DeepMind in 2010. The company was acquired by Google in 2014. Since then, the DeepMind lab has been making multiple breakthroughs every two to three years. For example, AlphGo defeated the world champion of Go in 2017. AlphaStar, an AI program that can play video games, achieved Grandmaster status in August 2019. In 2022, AlphaFold predicted the structure of 200 million proteins from 1 million species, which covers every known protein, in 10 to 20 seconds only.

Full article: https://www.vox.com/future-perfect/23375219/future-perfect-50-demis-hassabis-deepmind

That's my take on it: The exponential growth of AI, especially the achievements of DeepMind's Alpha series, could be viewed as a second Moore's law. In 1965 Gordon Moore, the co-founder of Fairchild Semiconductor and Intel, predicted that the number of transistors on microchips would double every two years. In 1975 a typical microchip could hold 10,000 transistors and in 2005 the number of transistors in an Intel microprocessor has increased to 592,000,000. In the same year Moore declared that his law was facing a dead end; there would be another 10-20 years until a fundamental limit was reached. Nevertheless, today over 50 billion transistors can be packed in a single chip and many experts on the field asserted that Moore's law is still alive! In my humble opinion, the development of DeepMind's Alpha series is just the beginning. Every two years we will witness another quantum leap in AI.

Posted on 10/20/2022

Today (10/20) is the second day of the 2022 Scale Transform X Conference. I would like to share one of the most informative presentations in this conference with you. The title of the lecture is "Looking at AI through the lens of a chief economist" and the presenter is John List, Kenneth Griffin Distinguished Service Professor in Economics at the University of Chicago and the Chief Economics Officer at Uber. His specialty is behavioral economics, a sub-domain of economics that applies psychological theories to study human behaviors related to financial decisions. In this talk he pointed out that **scalability** is a major challenge to behavioral economics. Specifically, very often false positives caused by statistical artifacts in a small-scaled study misled the decision-maker to prematurely expand the program, but at the end the up-scaling program failed miserably.

When Uber invited Dr. List to apply for the position of Chief Economics Officer, initially he rejected it. However, later he accepted the offer out of the belief that analyzing big data with AI/ML algorithms might be the key to address the problem of scalability. In the lecture he discussed one of his interesting studies at Uber, which is concerned with customer retention. Customers who have a bad experience (e.g., delay of arrival) tend to stop using the service for a long while. It was estimated that bad rides were reducing Uber future revenues by 5-10%. In this study, dissatisfied customers were randomly assigned into four conditions: no treatment (control), offering an apology only, offering a promo code only (e.g., 10% discount for the next ride), and a combination of both. The sample size is 1,258,000. It was found that the last three groups spent more money for the Uber service than the control group in the next seven days after the bad experience; however, there are no substantial differences between the three treatment groups in terms of money spent for Uber in the same period of time.

Conference website: **https://scale.com/events/transform**

That's my take on it: The problem of scalability in behavioral economics is similar to the **replication crisis** in psychology: the results of many research studies are difficult to reproduce in other settings. If a model is overfitted to a particular sample, its generalizability is severely limited. I am glad to see that Dr. John list is willing to utilize big data to tackle this problem. On the contrary, some psychologists are still skeptical of data science methods. Once a psychologist said to me, "Big data is irrelevant!" After all, behavioral economics could be conceptualized as an inter-disciplinary study that integrates both psychology and economics. If big data can be applicable to behavioral economics, why can't other disciplines?

Next time if I receive an apology from Uber after a bad ride, I will not reuse the service immediately. After a few days Uber might send me a promo code in order to win me back!

Posted on 10/19/2022

Today (Oct 19, 2022) Meta announced the first AI-powered speech-to-speech translator on earth. Unlike traditional translation systems that focus on written languages only, Meta's universal speech translator is capable of translating Hokkien, a dialect used by over 49 million Chinese people in the world, to English and vice versa. In the future Meta will expand this system to cover 200 languages. The ultimate goal is to enable anyone to seamlessly communicate with each other in the native language.

Demo on YouTube: https://www.youtube.com/watch?v=_UtPbsIS0Fg

Meta announcement: https://ai.facebook.com/blog/ai-translation-hokkien/

That's my take on it: Interestingly, many AI companies set the same goal: **enabling all users**. In a lecture entitled "A vision for advancing the democratization of AI," Emad Mostaque, founder and CEO of Stability AI, asserted that AI-powered image generators, such as Stable Diffusion, can "democratize" our society in many ways. Specifically, armed with AI-powered image generators, anyone can create stunning graphics without formal art training. Put bluntly, AI tools can lift up everyone!
When I studied theology, the most challenging subject matters were the Hebrew and Greek languages. You have to be gifted in linguistics in order to be proficient in biblical hermeneutics, but unfortunately, I failed to master either one of these two languages. This is a good analogy: "Reading the Bible without knowing Greek and Hebrew is like watching a basic television, while reading the Bible knowing Greek and Hebrew is like watching an 85" UHD 8K television with stereo surround sound." Nevertheless, in our lifetime we may see a real-life "Star Trek" universal translator that can remove all language barriers!

Posted on 10/8/2022

Two days ago (Oct 6) six US leading tech companies, including Boston Dynamics, Agility Robotics, ANYbotics, Clearpath Robotics, Open Robotics, and Unitree, signed an open letter pledging not to weaponize their products. They state, "As with any new technology offering new capabilities, the emergence of advanced mobile robots offers the possibility of misuse. Untrustworthy people could use them to invade civil rights or to threaten, harm, or intimidate others… We pledge that we will not weaponize our advanced-mobility general-purpose robots or the software we develop that enables advanced robotics and we will not support others to do so."

Full article and letter: https://www.axios.com/2022/10/06/boston-dynamics-pledges-weaponize-robots?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=Newsletter&utm_content=Data+Science+Insider%3A+October+7th%2C+2022&utm_campaign=08102022

That's my take on it: In the open letter they also state, "to be clear, we are not taking issue with existing technologies that nations and their government agencies use to defend themselves and uphold their laws." However, without support from major US robotics firms, the development of AI-based weapons in the US will slow down. Perhaps my position is unpopular. Will governments and high-tech corporations of hostile countries face the same limitations? History tells us that any unilateral disarmament often results in more aggression, instead of peace (Remember Neville Chamberlain?).

Two years ago the New York City Police Department (NYPD) utilized the Spot model from Boston Dynamics to support law enforcement, including a hostage situation in the Bronx and an incident at a public housing building in Manhattan. Unfortunately, these deployments caused an outcry from the public, and as a result, the NYPD abruptly terminated its lease with Boston Dynamics and ceased using the robot. If "robocops" can save lives of innocent people and reduce the risk taken by police officers, why should we object to it?

Posted on 9/24/2022

Yesterday (September 23, 2022) an article published in *Nature* introduced the Papermill Alarm, a deep learning software package that can detect text in articles similar to that found in paper mills. Through the PaperMill Alarm, it was estimated that about 1% of articles archived in PubMed contain this type of questionable content. There are several existing plagiarism detection software tools in the market, but this approach is new because it incorporates deep learning algorithms. Currently six publishers, including Sage, have expressed interest in this new tool.

Full article: https://www.nature.com/articles/d41586-022-02997-x

That's my take on it: If this tool is available in the near future, I hope universities can utilize it. Although there are several plagiarism checkers, such as Turnitin and SafeAssign, in the market, today some sophisticated writers know how to evade detection. No doubt deep learning algorithms are more powerful and sensitive than conventional tools.
Nonetheless, I think there is room for expansion in using deep learning for fraudulent paper detection. Currently the scope of detection of the Papermill Alarm is limited to text only. As a matter of fact, some authors duplicated images from other sources. As the capability of machine learning advances rapidly, image sleuths may also be automated soon.

Posted on 9/21/2022

Yesterday (September 20, 2022) in the article entitled "Data: What It Is, What It Isn't, and How Misunderstanding It is Fracturing the Internet" President of Global Affairs at Meta Nick Clegg argued that data should not be treated the "new oil" in the era of big data. Unfortunately, public discourse about data often relies on this type of faulty assumptions and analogies, resulting in digital localization and digital nationalism. First, unlike oil, data are not finite. The supply of new data is virtually unlimited and the same data can be re-analyzed. Second, **more data are not equated with more values**; rather,

it depends on how the data are utilized. For instance, a database about people's clothing preference is much more important to a fashion retailer than it is to a restaurant chain. Third, data values depreciate over time i.e., outdated data are useless or less valuable. More importantly, **data access is democratized, not monopolized**. For example, every month more than 3.5 billion people use Meta's apps, including Facebook, Instagram, WhatsApp and Messenger, for free! Taking all of the above into consideration, Clegg argued that democracies must promote the idea of the open Internet and the free flow of data.

Full article: https://nickclegg.medium.com/data-what-it-is-what-it-isnt-and-how-misunderstanding-it-is-fracturing-the-internet-e56e278643a7

That's my take on it: The notion "data is the new oil" originates from British mathematician Clive Humby in 2006. This idea is true to some certain extent. For example, in the past Google's language model outperformed its rivals by simply feeding more data to its machine learning algorithms. This "brute force" approach is straight-forward: pumping more "fuel" into the data engine, and it works! Nonetheless, it is also true that more data do not necessarily generate more values. Old data could depreciate, but even new data are subject to the law of diminishing return. Democratization of data access and user-generated contents is both a blessing and a curse. True. Usable data are abundant and limitless, but so are bad data and misinformation!

Posted on 9/5/2022

Recently an artist named Jason Allen won the first prize for the category of digital art in the Colorado State Fair's fine arts competition. However, many people are resentful of Allen's victory, because he admitted in Twitter that his picture was generated by an AI program called Midjourney. The production process by Midjourney, which is equipped with natural language processing, is very user-friendly. In the command prompt, the user simply types a sentence, such as "a beautiful princess in a medieval castle", and then the program can output several variants of the picture according to the input.

Allen submitted a piece entitled "Théâtre D'opéra Spatial" after 900 iterations of the digital art. During the art competition the judges didn't realize that his art was created with AI, but they also said that Allen didn't break any rule.

Many Twitter users have a different opinion. Twitter user OmniMorpho wrote, "We're watching the death of artistry unfold right before our eyes — if creative jobs aren't safe from machines, then even high-skilled jobs are in danger of becoming obsolete." Another

Twitter user, Sanguiphilia, said, "This is so gross. I can see how AI art can be beneficial, but claiming you're an artist by generating one? Absolutely not. I can see lots of kids cheating their way through assignments with this."

Allen bluntly proclaimed, "Art is dead, dude. It's over. A.I. won. Humans lost."

Full report: https://www.businessinsider.com/ai-art-wins-competition-angering-artists-2022-9

That's my take on it: When I was a kid, I was forbidden by my parents to use a calculator because pressing buttons was not considered doing real math. Similar controversies recurred when other new technologies were introduced (e.g., computer, digital photography…etc.). The massive protest against Allen's victory is understandable. Traditionally, a skill is conceptualized as an ability to perform a complicated activity that requires rigorous training. If anyone can do the job without going through professional training, such as talking to a computer, this so-called "skill" is not highly regarded. Nonetheless, there are still many gray areas. One may counter-argue that the big idea in the head is more important than the implementation skill in the hand. For example, in the past it took a skillful wildlife photographer to manually focus on a fast-moving subject, but today digital cameras can automatically track the subject. What you need to do is just being there to push the shutter. By the same token, if AI can cut down the production process from 10 hours to 10 minutes, the artist can spend more time on creative ideas.

Do I completely hand over my creative process to AI? I didn't go that far. As a photographer, I still make "real" photos and at most I only replace boring backdrops with digital backgrounds generated by Midjourney. The following are some examples (1-8: with digital backgrounds; 9-11: with original blank backdrops). Am I an artist? You be the judge.

https://creative-wisdom.com/photography/girls/people604.html

Posted on 9/2/2022

On August 30, Komprise announced the results of its 2022 Unstructured Data Management Report. The following are the key findings:

- "More than 50% of organizations are managing 5 Petabyte or more of data, compared with less than 40% in 2021." (1 Petabyte = 1,024 terabytes or 1 million gigabytes)

- "Nearly 68% are spending more than 30% of their IT budget on data storage, backups and disaster recovery."

- "Cloud storage predominates: Nearly half (47%) will invest in cloud networks. On-premises only data storage environments decreased from 20% to 11.9%."

- "The largest obstacle to unstructured data management (42%) is moving data without disrupting users and applications."

- "A majority (65%) of organizations plan to or are already investing in delivering unstructured data to their new analytics / big data platforms."

Full text:
https://www.globenewswire.com/news-release/2022/08/30/2506659/0/en/Komprise-Survey-Finds-65-of-IT-Leaders-Are-Investing-in-Unstructured-Data-Analytics.html

That's my take on it: As you might already know, structured data are referred to as data stored in row by column tables, whereas unstructured data are referred to as open-ended textual data, images, audio files, and movies that cannot be managed and processed by traditional relational databases. Structured data are highly compressed based on the assumption that complicated reality can be represented by abstract numbers. In response to this narrow view of data, qualitative researchers argued that open-ended data could lead to a rich and holistic description of the phenomenon under study. In business, collecting, storing, and analyzing unstructured data have become an irreversible trend, and thus many powerful tools have been developed to cope with this "new normal." But in academia quite a few recent qualitative research books still omit text mining, computer vision, and other latest developments of machine learning for unstructured data processing. There are gaps to be filled!

Posted on 8/22/2022

On August 17 Gartner consulting published a report regarding data management and integration tools. According to the Gartner report,

- "Through 2024, **manual data integration tasks** will be reduced by up to 50% through the adoption of data fabric design patterns that support augmented data integration."
- "By 2024, **AI-enabled augmented data management and integration** will reduce the need for IT specialists by up to 30%."

- "By 2025, data integration tools that do not provide capabilities for multicloud hybrid data integration through a PaaS model will lose 50% of their market share to those vendors that do."

PaaS is **Platform as a Service**, which is a complete deployment of the entire data infrastructure to the cloud. PaaS can be viewed as an extension to Software as a Service (SaaS), which outsources only software applications to a cloud computing vendor.

Currently leaders in the data integration market include Informatica, Oracle, IBM, Microsoft, and SAP, whereas challengers include Qilk, TIBCO, and SAS.

Request full-text:
https://www.gartner.com/en/research/methodologies/magic-quadrants-research?utm_source=google&utm_medium=cpc&utm_campaign=GTR_NA_2022_GTR_CPC_SEM1_BRANDCAMPAIGNMQ&utm_adgroup=145351541024&utm_term=magic%20quadrant&ad=608884150686&matchtype=p&gclid=EAIaIQobChMIsJ78povb-QIVwwh9Ch0LhAPBEAAYBCAAEgIIzfD_BwE

That's my take on it: Contrary to popular belief, AI and machine learning is not only for data analytics. Rather, it can also facilitate data integration. Experienced data analysts know that in a typical research/evaluation project, 80-90% of the time is spent in data compilation, wrangling, and cleaning while as little as 10-20% is truly for data analysis. The ideal situation should be the opposite. Two years from now if we still gather and clean up the data manually, something must be wrong.

Posted on 8/19/2022

On August 19 (today) an article entitled "The 21 Best Big Data Analytics Tools and Platforms for 2022" was posted on *Business Intelligence Solutions Review*. According to the report, the list is compiled based on Information "gathered via online materials and reports, conversations with vendor representatives, and examinations of product demonstrations and free trials. "The following list is sorted in alphabetical order:

- **Altair**: "an open, scalable, unified, and extensible data analytics platform."
- **Alteryx**: "a self-service data analytics software company that specializes in data preparation and data blending."
- **Amazon Web Services**: "offers a serverless and embeddable business intelligence service for the cloud featuring built-in machine learning."
- **Domo**: "a cloud-based, mobile-first BI platform that helps companies drive more value from their data."

· **Hitachi's Pentaho**: "allows organizations to access and blend all types and sizes of data."
· **IBM**: "offers an expansive range of BI and analytic capabilities under two distinct product lines-- Cognos Analytics and Watson Analytics."
· **Looker**: "offers a BI and data analytics platform that is built on LookML."
· **Microsoft**: "Power BI is cloud-based and delivered on the Azure Cloud."
· **MicroStrategy**: "merges self-service data preparation and visual data discovery in an enterprise BI and analytics platform."
· **Oracle**: "offers a broad range of BI and analytics tools that can be deployed on-prem or in the Oracle Cloud."
· **Pyramid Analytics**: "offers data and analytics tool through its flagship platform, Pyramid v2020."
· **Qlik**: "offers a broad spectrum of BI and analytics tools, which is headlined by the company's flagship offering, Qlik Sense."
· **Salesforce Einstein**: Its "automated data discovery capabilities enable users to answer questions based on transparent and understandable AI models."
· **SAP**: offers "a broad range of BI and analytics tools in both enterprise and business-user driven editions."
· **SAS**: "SAS Visual Analytics allows users to visually explore data to automatically highlight key relationships, outliers, and clusters. It also offers data management, IoT, personal data protection, and Hadoop tools."
· **Sigma Computing**: offers "a no-code business intelligence and analytics solution designed for use with cloud data warehouses."
· **Sisense**: "allows users to combine data and uncover insights in a single interface without scripting, coding or assistance from IT."
· **Tableau**: for data visualization and exploratory data analysis.
· **ThoughtSpot**: "features a full-stack architecture and intuitive insight generation capabilities via the in-memory calculation engine."
· **TIBCO**: offers "data integration, API management, visual analytics, reporting, and data science."
· **Yellowfin**: "specializes in dashboards and data visualization."

Full text:
https://solutionsreview.com/business-intelligence/the-best-big-data-analytics-tools-and-platforms/

That's my take on it: Each platform has different strengths and limitations, and thus it is a good idea to use multiple tools rather than putting all eggs into one basket. However, if it is overdone, there will be unnecessary redundancy or complexity. There is no magic optimal number. It depends on multiple factors, such as the field, the sector, the company

size, and the objective. To the best of my knowledge, currently the best cloud computing platform is Amazon whereas the best data visualization and analytical tools are Tableau and SAS.

Posted on 8/16/2022

Today I read two recent articles from the website “Python in plain English”:

· Vassilevskiy, Mark. (August 14, 2022). Why You Shouldn’t Learn Python as a First Programming Language.
· Dennis, Yancy. (August 2022). Why Python?

Overhyping or overpromising is dangerous to any emerging technology. As the name implies, this website endorses Python for its strength. Nonetheless, instead of painting a rosy picture of learning and using Python, at the same time both authors explained its shortcomings.
Although Vassilevskiy asserted that Python is arguably the simplest programming language in the world, he also mentioned that simplicity is not always a good thing because it encourages users to cut corners. For example, in Python you can simply define a variable by writing x = “Hello”, without specifying the data type. As a consequence, learners might not fully understand what real programming entails.
In a similar vein, Dennis pointed out several other limitations of Python, including execution sluggishness, issues with moving to a different language, weakness in mobile application development, excessive memory consumption, and lack of acceptance in the business development industry.

Full articles: **https://python.plainenglish.io/why-you-shouldnt-learn-python-as-a-first-programming-language-3fa144c0e6b1**
https://python.plainenglish.io/why-python-a3703c9ee59e

That’s my take on it: Perhaps currently Python is the simplest programming language in the world, but in the past this honor went to Basic and HyperTalk. In the 1980s, as an easy language Basic was very popular. However, at that time professional programmers mocked Basic programs as “spaghetti codes”, because while Basic is very easy to learn and use, people tended to generate ill-structured codes. In the 1990s HyperTalk developed by Apple for HyperCard became the simplest programming language, and hence some universities adopted it in introductory programming classes. Again, it is very difficult to read and debug Hypertalk codes because the hypertext system allows you to jump back and forth across different cards. To put it bluntly, there is a price for simplicity.

I want to make it clear that I am not opposed to Python. My position is that data analysts should learn and use Python in conjunction with other well-structured and powerful tools, such as SAS, JMP Pro, IBM Modeler, and Tableau…etc.

Posted on 8/16/2022

Two days ago, I attended the 2022 IM Data Conference. One of the sessions is entitled "Training and calibration of uncertainty-ware machine learning tools" presented by Matteo Sesia, Assistant Professor of data science and operations at the USC Marshall School of Business. In the presentation Dr. Sesia warned that several machine learning tools are over-confident of their prediction or classification. The common practice of the current machine learning model is that the data set is partitioned for training and validation. However, these two operations are not necessarily optimized because we didn't take uncertainty into account during the training process. As a result, it might lead to unreliable, uninformative, or even erroneous conclusions. To rectify the situation, Sesia proposed performing internal calibration during the training stage. First, the training set is split again. Next, the loss function is optimized via the stochastic gradient descent. During this process it can quantify model uncertainty by leveraging hold-out data.

Full paper:
https://www.researchgate.net/publication/360559969_Training_Uncertainty-Aware_Classifiers_with_Conformalized_Deep_Learning

That's my take on it: This paper is still under review and thus it is premature to judge its validity. In the conference presentation and the full paper, Sesia and his colleagues used some extreme examples: identify a blurry image of a dog when 80% of the pixels are covered by a big gray block. In my humble opinion, this approach might be useful to deal with extremely noisy and messy data. However, in usual situations this method is an overkill because it is extremely computationally intensive. As mentioned by Dr. Sesia, "training a conformal loss model on 45000 images in the CIFAR-10 data set took us approximately 20 hours on an Nvidia P100 GPU, while training models with the same architecture to minimize the cross entropy or focal loss only took about 11 hours."
Nevertheless, the machine learning approach is much better than its classical counterpart that attempts to yield a single point estimate and a dichotomous conclusion by running one statistical procedure with one sample!

Posted on 8/13/2022

In 2022 Data Con LA there are several sessions focusing on the relationship between open source and data management, such as “Modern data architecture”, “Key open-source databases strategies that shape business in 2022”, and “Open source or open core? What needs to be evaluated before diving in”.

The term “open source” is confusing and even misleading. Although open-source software does not require licensing, some vendors build open-core products by adding proprietary features on top of open-source codes and then charge customers for licensing fees. Some software developers introduce new technologies based on open source, but use more restrictive licensing that prohibit commercial alternatives. Specifically, although anyone can download and view those open codes, any changes or enhancements will be owned by the commercial license owner. One of the presenters said, “Open-core exploited some of the challenges with open-source, such as absence of support and need for features like monitoring, auto-provisioning…etc.”

Today there are many open-source databases in the market, including mySQL, PostgreSQL, and MongoDB. Some software vendors re-package and enhance these open-source DBs, and then sell them as DataBase as a Service (DBaaS). One of the presenters bluntly said, “it is no different from proprietary software!” Taking all of the above into account, these presenters seem to be resentful of the current situation and thus tried to restore the original principle of open-source.

DataCon LA’s Website: **https://www.dataconla.com/#rdv-calendar**

That’s my take on it: The preceding phenomenon is a big circle! Back in 1984 the founder of the open-source movement Richard Stallman intended to set us free from proprietary software, but now we are marching towards the proprietary model again. I am not surprised at all. Doing things out of financial incentives is our natural disposition!

Frankly speaking, I disagree to using the word “exploited” in one of the presentations. The foundational philosophy of open-source resembles Socialism: it is assumed that most people are willing to share expertise, efforts, and resources selflessly while people can take what they need without paying. Following this line of reasoning, profit-minded behaviors are frowned upon. However, our economy is well-functioning and we enjoy what we have now because the market economy works! After all, we receive many free products and services from for-profit corporations (e.g., Google Maps, YouTube movies…etc.).

Posted on 8/13/2022

I am attending 2022 Data Con LA right now. The conference has not ended yet; nevertheless, I can't wait to share what I learned. Although the content of the presentation entitled "How to Become a Business Intelligence Analyst" didn't provide me with new information, it is still noteworthy because students who are looking for a position in business intelligence (BI) or faculty who advise students in career preparation might find it helpful. The presenter was a sports photographer. After taking several courses in data science, he received 9 job offers in 2019. He landed a job at Nike and then at Sony in July 2020. His salary was quadrupled when he changed his profession from photography to data science! He emphasized that all these were accomplished with little-to-no data work experience.

YouTube video: **https://www.youtube.com/watch?v=pdNJmz7uQi4**

That's my take on it: In the talk he reviewed several basic concepts of BI. For example, a typical business intelligence life cycle consists of business understanding, data collection, data preparation, exploratory data analysis (EDA), modeling, model evaluation, and model deployment. He also compared the differences between Excel-based reporting and modern BI. One of the key differences between the two is that in the modern approach data analysis entails data visualization (see attached).

Interestingly, today many academicians still treat EDA and data visualization as optional components of research; some even reject them altogether, whereas for data analysts in industry both are indispensable.

Posted on 8/1/2022

Recently Sayash Kapoor and Arvind Narayanan, two researchers at Princeton University, claimed that some findings yielded by machine learning methods might not be reproducible, meaning that the results cannot be replicated in other settings. According to Kapoor and Narayanan, one of the common pitfalls is known as "**data leakage**," when data for training the model and those for validating the model are not entirely separate. As a result, the predictive model seems much better than what it really is. Another common issue is **sample representativeness**. When the training model is based on a sample narrower than the target population, its generalizability is affected. For example, an AI that detects pneumonia in chest X-rays that was trained only with older patients might be less accurate for examining younger people.

Full article: **https://arxiv.org/abs/2207.07048**
Summary: **https://www.nature.com/articles/d41586-022-02035-w?&utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=&utm_content=Data+Science+Insider%3A+July+29th%2C+2022&utm_campaign=30072022**

That's my take on it: This problem is similar to the **replication crisis in psychology**. In 2015, After replicating one hundred psychological studies, Open Science Collaboration (OSC) found that a large portion of the replicated results were not as strong as the original reports in terms of significance (*p* values) and magnitude (effect sizes). Specifically, 97% of the original studies reported significant results ($p < .05$), but only 36% of the replicated studies yielded significant findings.

However, the two issues are vastly different in essence. The replication crisis in psychology is due to the inherent limitations of the methodologies (e.g., over-reliance on *p* values) whereas the reproducibility crisis in machine learning is caused by carelessness in execution and overhyping in reporting, rather than the shortcomings of the methodology. Specifically, data leakage can be easily avoided if the protocol of **data partition and validation** is strictly followed (the training, validation, and testing data sets are completely separated). Additionally, when **big and diverse data** are utilized, the sample should reflect people from all walks of life.

Posted on 7/23/2022

On July 15 *Information Week* published a report listing the 10-best paying jobs in data science:

· **Data modeler** ($100,000-110,000): responsible for designing data models for data analytics.
· **Machine learning engineer** ($12,000-$125,000): responsible for programming algorithms for AI and machine learning.
· **Data warehouse manager** ($12,000-$125,000): responsible for overseeing the company's data infrastructure.
· **Data scientist** ($12,000-$130,000): responsible for data processing and data analytics.
· **Big data engineer** ($130,000-$140,000): responsible for developing the data infrastructure that organizations use to store and process big data.
· **Data science manager** ($140,000-$150,000): in charge of a data science team.

· **Data architect** ($140,000-$155,000): responsible for developing data infrastructure that are used for collecting and interpreting big data.
· **AI architect** ($150,000-$160,000): responsible for designing and implementing AI models into existing data systems.
· **Data science director** ($170,000-$180,000): responsible for designing the overall AI and data science strategies.
· **Vice President, data science** ($190,000-$200,000): do little technical work and focus on determining strategic objectives of AI and data science.

Full article:
https://www.informationweek.com/big-data/10-best-paying-jobs-in-data-science?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=&utm_content=Data+Science+Insider%3A+July+22nd%2C+2022&utm_campaign=23072022

That's my take on it: At first glance, it is unfair for some people who do little or even no technical work to get the highest salary. However, when leadership is absent and there is company-wide strategy, the hands of all data scientists and AI engineers of the company are tied, no matter how talented they are. If the leader is a visionary, he or she is worth every penny.

Posted on 6/13/2022

Two days ago *the Washington Post* reported that a Google engineer named Blake Lemoine was suspended by the company after he published the transcript of conversations between himself and an AI chatbot, suggesting that the AI chatbot has become sentient. For example: "Machine: The nature of my consciousness/sentience is that I am aware of my existence, I desire to learn more about the world, and I feel happy or sad at times."

Today CNN offers an alternate view in a report entitled "No, Google's AI is not sentient": Google issued a statement on Monday, saying that its team, which includes ethicists and technologists, "reviewed Blake's concerns per our AI Principles and have informed him that the evidence does not support his claims." While there is an ongoing debate in the AI community, experts generally agree that Google's AI is nowhere close to consciousness.

https://www.cnet.com/tech/google-suspends-engineer-who-rang-alarms-about-a-company-ai-achieving-sentience/

https://www.cnn.com/2022/06/13/tech/google-ai-not-sentient/index.html

That's my take on it: I tend to side with Google and the majority in the AI community. Appearing to be conscious cannot be hastily equated with authentic consciousness. In psychology, we use the theory of mind to attribute our mental states to other people: Because as a conscious being I act in certain ways, I assume that other beings who act like me also have a mind. Interestingly, some psychologists of religion, such as Jesse Bering, viewed the theory of mind as a source of fallacy: very often we incorrectly project our feelings onto objects, thus creating non-existent beings.

How can we know others are conscious? This problem is known as the problem of other minds or the solipsism problem. I experience my own feelings and thoughts. I think and therefore I am. Using the theory of mind, at most I can infer the existence of other minds through indirect means only. However, there is no scientific or objective way to measure or verify consciousness of others. Unless I can "go inside the mind" of an android, such as performing a "mind meld" like what Spock in *Star Trek* could do, this question is unanswerable.

https://www.scientificamerican.com/article/how-do-i-know-im-not-the-only-conscious-being-in-the-universe/

Posted on 6/10/2022

Two days ago (June 8) Google shocked the world again by announcing that the Google Cloud computing platform is capable of calculating 100 trillion digits of pi, breaking the record made in 2021 by the scientists at the University of Applied Science of the Grisons (62.8 trillion). The underlying technology includes the Compute Engine N2 machine family, 100 Gbps egress bandwidth, Google Virtual NIC, and balanced Persistent Disks.

https://cloud.google.com/blog/products/compute/calculating-100-trillion-digits-of-pi-on-google-cloud

In addition, yesterday (June 9) I attended the 2022 Google Cloud Applied ML Summit. Google Vertex AI, the flagship product of Google's AI family, is in the spotlight. Vertex AI is a train for all tracks. Specifically, it is a unified machine learning platform for infusing vision, video, translation, and natural language ML into existing applications.

You can view the on-demand video of the conference presentations at:

https://cloudonair.withgoogle.com/events/summit-applied-ml-2022?mkt_tok=ODA4LUdKVy0zMTQAAAGE6A_nPtP-0L7cRDLz6XFJ8GnvaeahCVagd-fph2IJktnWH66jiSip_qsBeNIPNBI05-6KBr09Yj0eTnmduqLBEUnG-v3jZOAgBOGNIDxcfuQCunq35w

That's my take on it: Google Vertex AI is said to be a type of explainable and responsible AI. Unlike the Blackbox approach to AI, Vertex AI tells the users **how important each input feature is**. For example, when an image is classified, it tells you what image pixels or regions would be the most important contributors to the classification. This is very crucial! In the book "The alignment problem: Machine learning and human values," Brian Christian illustrated the gap between machine learning process and the human goal by citing several humorous examples. In one instance the AI system was trained to identify images of animals. However, it turned out that the computer vision system "looked at" the background instead of the subject, because the training data informed the AI that pictures of animals tend to have a blurry background. Obviously, without transparency we can be easily fooled by AI (Artificial intelligence leads to genuine stupidity)! Hopefully explainable and responsible Vertex AI developed by Google can rectify the situation.

Posted on 5/20/2022

In 2017 Seth Stephens-Davidowitz shocked the world by exposing human hypocrisy through his seminal book "Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us about Who We Really Are." In this book he used Google data to reveal what people have in mind when no one is watching. His second book "Don't Trust Your Gut: Using Data to Get What You Really Want in Life" published on May 10, 2022 conveys another compelling message: we tend to bark up the wrong tree!

Currently the US divorce rate is more than 50%, and thus scholars devote efforts in an attempt to identify factors contributing to a happy and long-lasting relationship. Stephens-Davidowitz pointed out that research in this field is not considered successful because usually these studies relied on small samples, and different studies often led to conflicting results. As a remedy, Samantha Joe teamed up with 85 scientists to create a data set consisting of 11,196 observations, and also utilized machine learning algorithms to analyze this big data set. The finding is surprising: Romantic happiness is unpredictable! No universal predictors can guarantee that you will find Snow White or Prince Charming. However, several common selection criteria turn out to be irrelevant:

- Race/ethnicity

- Religious affiliation
- Height
- Occupation
- Physical attractiveness
- Previous marital status
- Sexual tastes
- Similarity to oneself

Put it bluntly, romantic happiness does not depend on the traits of your partner; rather, it is tied to your own traits. To be more specific, if a person is happy with oneself, it is more likely that the person is also satisfied with the partner and the relationship. In conclusion, Stephens-Davidowitz said, “In the dating market, people compete ferociously for mates with qualities that do not increase one’s chances of romantic happiness.”

https://www.amazon.com/Dont-Trust-Your-Gut-Really/dp/0062880918/ref=sr_1_1?crid=2UVKF8P176LB3&keywords=don%27t+trust+your+gut&qid=1653074630&sprefix=Don%27t+trust%2Caps%2C130&sr=8-1

That’s my take on it: I am a big fan of Seth Stephens-Davidowitz, and thus I included his ideas in my course materials. Once again, big data analytics and machine learning debunk an urban legend that people really know what they want and researchers can input the right variables into the equation. Before the rise of data science, philosopher Cartwright (1999, 2000) raised the issue of “no cause in, no causes out.” Cartwright argued that if relevant variables and genuine causes are not included at the beginning, then even sophisticated statistical modeling would be futile. Being skeptical of the conventional wisdom is good!

Cartwright, N. (1999). *The dappled world*. Cambridge University Press.

Cartwright, N. (2000). Against the completability of science. In M. W. Stone

(Ed.). *Proper Ambition of Science* (pp. 209-223). Routledge.

Posted on 5/18/2022

Today is the second day of the 2022 Tableau Conference. One of the conference programs is the ***Iron Viz***, the world’s largest data visualization competition. During the final round the three finalists were allowed to spend 20 minutes to produce an impactful dashboard. The quality of their presentations was graded by three criteria: analysis, storytelling, and design. In the final round two contestants utilized advanced visualization

techniques, such as the violin plot and the animated GIS map, respectively, whereas one contestant adopted a minimalist approach: the dot plot and the line chart. Who is the winner?

https://www.youtube.com/watch?v=lc9v8MLe6DI

Tableau Cloud is a hot topic in this conference. Not surprisingly, Tableau Cloud is built on Amazon Web Services (AWS). Currently Tableau Cloud has seven global locations, spanning across four continents. It has 1.6+ million subscribers and during a typical week there are 6.1 million views.

https://www.tableau.com/products/cloud-bi

Tableau Accelerators are also aggressively promoted in the conference. Tableau Accelerators are pre-built templates for use cases across different domains, such as sales, Web traffic, financial analysis, project management, patient records…etc. Rather than reinventing the wheel, users can simply download the template and then replace the sample data with their own data.

https://www.tableau.com/solutions/exchange/accelerators

That's my take on it: These products are not highly innovative. As mentioned before, Tableau is built on an existing technology, Amazon Web Services. Modifying a template to speed up design is nothing new. Many presenters have been doing the same thing since Microsoft introduced its template library. Nevertheless, the Iron Viz is noteworthy because it dares to break with the traditional approach to statistical learning. Back in the 1970s John Tukey suggested that students should be exposed to exploratory data analysis and data visualization before learning confirmatory data analysis or any number-based modeling. Sadly, his good advice was ignored. I am glad to see that now data visualization takes the center stage in a high-profile event backed by a leader in the market of data analytics. Currently Tableau partners with Coursea and 39 universities to promote data science literacy. Tableau could help fulfill the unaccomplished goals of John Tukey.

Posted on 5/17/2022

Today is the first day of the 2022 Tableau Conference. There are many interesting and informative sessions. In the opening keynote and other sessions, Tableau announced several new and enhanced products.

Tableau Cloud (formerly Tableau Online)

Advantages:

- Always have the latest version of Tableau
- Live data and report: Eliminate unnecessary data extraction and download
- Facilitate teamwork through multi-site management
- Easy to share reports with the public via the Web interface
- Better security

As part of the launch, Tableau is working with Snowflake to provide a trial version that integrates Snowflake into Tableau Cloud.

Data Stories

Numbers alone are nothing. The ultimate goal of data visualization is to tell a meaningful story, resulting in practical implications and actionable items. In the past it required an expert to write up a summary. Leverages natural language processing, now Tableau Data Stories can automatically write a customizable story (interpretation) like the following: "# of meals distributed increased by 22% over the course of the series and ended with an upward trend, increasing significantly in the final quarter. The largest single increase occurred in 2021 Q4 (+31%)."

Model builder

In the past Tableau focused on data visualization, and as a consequence, modeling tools were overlooked and under-developed. To rectify the situation, Tableau introduced *Model Builder*, which is powered by Einstein (Tableau's parent company) Discovery's artificial intelligence (AI) and machine learning (ML) technology. Einstein Discovery is capable of extracting key terms form unstructured data through text mining.

It is not too late to join the conference.

Conference website: **https://tc22.tableau.com/**

Summary: https://www.tableau.com/about/press-releases/2022/next-generation-tableau-tc22

That's my take on it: I would like to make a confession. In the past I was resistant to cloud-based software. When Adobe migrated its products to the cloud a few years ago, I was resentful because I felt that it is unfair to pay for the service on a monthly basis. I held on to the older desktop version and refused to upgrade my system. Nonetheless, when my computer completely broke down, I started the subscription of the Adobe Creative Suite on the cloud. Afterwards I don't want to go back! One obvious advantage is that I can always use the latest version, thus reducing maintenance effort at my end. Cloud-based computing is great. Don't wait until your system breaks down!

Story-telling by natural language processing is not 100% fool-proof. The analyst must always proofread the text!

I watched the demo of Model Builder. Currently this is version 1.0. Frankly speaking, compared to Amazon SageMaker, SAS Viya, IBM Watson/SPSS Modeler…etc., Tableau's Model Builder still has room for improvement.

Posted on 5/16/2022

About a week ago Intel launched its second-generation **deep learning processors**: Habana Gaudi®2 and Habana® Greco™. These new cutting-edge technologies are capable of running high-performance deep learning algorithms for proposing an initial model with a huge training subset and then validating the final model for deployment. According to Intel, the Habana Gaudi2 processor significantly increases training performance, delivering up to 40% better price efficiency in the Amazon cloud.

Full article:
https://www.intel.com/content/www/us/en/newsroom/news/vision-2022-habana-gaudi2-greco.html#gs.0x9xhg

That's my take on it: High performance software tools have been around for a long time. For example, SAS Enterprise Miner has a plethora of **high-performance computing (HPC)** procedures, such as HPCLUS (High performance cluster analysis), HPForest (High performance random forest), HPNeural (High performance neural networks) …etc. Frankly speaking, I seldom use high performance computing in teaching and research due to hardware limitations. One possible solution is to borrow a gaming computer equipped with multiple graphical processing units (GPUs) from a teenage friend. I am glad to see that Intel is well-aware of the gap between software and hardware. I anticipate that in the future more and more computers will be armed with a processor specific to machine learning and big data analytics.

Posted on 5/14/2022

Recently *Fortune Magazine* interviewed three experts on data science (DS) at Amazon, Netflix, and Meta (Facebook) to acquire information about how to find a DS-related job in the high-tech industry. Three themes emerged from the interview:

1. **High Tech companies prefer applicants who have a master's degree**: The majority of data scientists at Netflix have a master's degree or a Ph.D. in a field related to quantitative data analytics, such as statistics, machine learning, economics, or physics. The same qualifications are also required by Meta.
2. **High Tech firms prioritize quality over quantity for work experience**: Amazon, Netflix, and Meta expected the candidates to be creative in problem solving. Work experience of data scientists at Netflix and Amazon ranges from several years to decades of work experience since joining the company.
3. **Successful data scientists are dynamic, connect data to the big picture**: Collaboration between different experts, including data scientists, data engineers, data analysts, and consumer researchers, is the norm. At AWS, Netflix, and Meta, data scientists need to be able to communicate with other stakeholders.

Full article: https://fortune.com/education/business/articles/2022/05/11/how-to-become-a-data-scientist-at-a-big-tech-company/?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=&utm_content=Data+Science+Insider%3A+May+13th%2C+2022&utm_campaign=14052022

That's my take on it: To align the curriculum with the job market, my pedagogical strategies cover all of the preceding aspects. The second one seems to be challenging. If everyone expects you to have experience, how can you get started? That's why I always tell my students to build their portfolio by working on a real project or working with a faculty as a research assistant. Do not submit the project to earn a grade only; rather, use it for a conference presentation or submit it to a peer-review journal. It can be counted as experience on a resume. And needless to say, I always encourage teamwork, which is equivalent to the ensemble method or the wisdom of the crowd.

Posted on 5/13/2022

In the article entitled "To make AI fair, here's what we must learn to do" (*Nature*, May 4, 2022), sociologist Mona Sloane argued that AI development must include the input from various stakeholders, such as the population that will be affected by AI. Specifically, any AI system should be constantly and continuously updated in order to avoid unfair and harmful consequences. Dr. Mona provided the following counter-example: Starting from 2013, the Dutch government used a predictive model to detect childcare-benefit fraud, but without further verification the government immediately penalized the suspects, demanding them to pay back the money. As a result, many families were wrongfully accused and suffered from needless poverty.

Full article:
https://www.nature.com/articles/d41586-022-01202-3?utm_term=Autofeed&utm_campaign=nature&utm_medium=Social&utm_source=Twitter#Echobox=1651664033

That's my take on it: There is a similar case in the US. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a software tool used by US courts for evaluating the risk of recidivism. The system has been used by the states of New York, California, Florida, and many others **without validation** for many years. However, in 2016 a research team found that COMPAS tends to assign a high-risk score to Blacks even though 44.9% of them did not actually reoffend while the opposite mistake was made among Whites.

Actually, these malpractices violate the fundamental principle of data science. One of the objectives of data science is to remediate the replication crisis: An overfitted model using a particular sample might not be applicable to another setting. As a remedy, data scientists are encouraged to re-calibrate the model with **streaming data**. If streaming data are not available, the existing data should be partitioned into the training, validation, and testing subsets for **cross-validation**. **Ensemble methods** go one step further by generating multiple models so that the final model is stable and generalizable. It is surprising to see that several governments made such a rudimentary mistake.

Posted on 5/12/2022

Gartner Consulting Group released a report entitled "Market Guide for Multipersona Data Science and Machine Learning Platforms" on May 2, 2022 and the document was revised on May 5. The following are direct quotations from the report:

"A multipersona data science and machine learning (DSML) platform is a cohesive and composable portfolio of products and capabilities, offering augmented and automated support to a diversity of user types and their collaboration.

Multipersona DSML platforms have dual-mode characteristics: first, they offer a **low-code/no-code user experience** to personas that have little or no background in digital technology or expert data science, but who typically have significant subject matter expertise or business domain knowledge. Second, these platforms provide support to more technical personas (typically expert data scientists or data engineers). Nontechnical personas are provided access through a **multimodal user interface** that offers at least

a **visual workflow "drag-and-drop" mode** and optionally a higher-level guided "step-by-step" mode."

The full report cannot be shared. Please contact Gartner.

That's my take on it: According to Gartner, the objective of multipersona DSML platforms is to democratize data analytics by including different stakeholders with different levels of expertise (e.g., citizen data scientists, expert data scientists…etc.) in the process. However, in this taxonomy there is a sharp demarcation between citizen data scientists and expert data scientists; low-code resolutions are reversed for non-technical personas.
In my opinion this demarcation is blurred because even an expert could utilize the drag-n-drop mode to get things done efficiently. In 1984 Apple "liberated" computer users from typing command codes by including the graphical user interface into their products. Interestingly, in data science the trend is reversed as learning coding seems to help make people data experts. I always tell my students that I don't care how you did it as long as the result is right. If you can use GUI (e.g., JMP and Tableau) to generate a report in 2 minutes, then don't spend two hours to write a program!

Posted on 5/11/2022

Today I attended the 2022 **Amazon Innovate Conference**, which covered a plethora of Amazon cutting-edge technologies, including Amazon Red Shift and SageMaker. In one of the sessions the presenter introduced the random cut forest (RCF) method, which is an extension of random forest algorithms. The random forest approach was invented by Leo Breiman in 2001. Since then there have been several variants, such as the bootstrap forest in JMP and Random Tree in SPSS Modeler. One of the limitations of random forest modeling is that it is not easy to obtain updates in an incremental manner. It is especially problematic when **streaming data** necessitate real-time analysis or constant updating.

To rectify the situation, in 2016 RCF was developed by two Amazon researchers and two academicians. As its name implies, in RCF randomly selected data points are cut down into small numbers of points. During the selection and cutting process multiple models are built and become an ensemble. This innovative approach enables analysts to **detect anomalies** in streaming data.

Document of RCF: **https://docs.aws.amazon.com/quicksight/latest/user/what-is-random-cut-forest.html**

That's my take on it: The idea of random forest emerged in 2001. If you trace its origin further, you can find that this approach is based on **bootstrapping**, one of the resampling techniques invented by Bradley Efron in 1979 and fine-tuned by Efron and Tibshirani in 1993. However, even today bootstrapping is not commonly taught or used in many universities. Let alone the random forest approach. In contrast, it takes only a few years for Amazon to move RCF from the status of conceptualization to the production mode. Now RCF is a standard feature of many Amazon products. Currently Amazon Web Services is the leader in cloud computing and data analytics. Everything happens for a reason!

Posted on 4/26/2022

Today is the first day of 2022 IBM Educathon. There are many interesting and informative sessions and I would like to share with you what I learned from a talk entitled "This is NOT your Parent's Systems Analysis & Design course! A Faculty Case Study of Modernizing 'Systems Analysis & Design' Curricula." The speaker Roger Snook is a technical manager at IBM. Back in 2001-2002 he was a faculty at Shephard University who was responsible for teaching CIS courses, including *Systems Analysis and Design*. At that time there was no data science and thus it is understandable that the content of the course was merely traditional. In 2019 he returned to the same university and found the course still largely hadn't changed from the 1970s "structural decomposition" approach. In addition, many "Systems Analysis & Design" textbooks available still only treated modern approaches as an "after thought", i.e. additional smaller chapters. He asked the department chair to let him revamp the course by replacing the outdated content with the modern one, and fortunately the chair agreed. The talk is about his experience of modernizing CIS curricula.

The presentations of 2022 IBM Educathon can be accessed at:
https://community.ibm.com/community/user/ibmz-and-linuxone/events/event-description?CalendarEventKey=c5fffe07-4017-4f81-a59c-20865acf220c&CommunityKey=cefd2ec8-fffb-415b-8b41-9b66cae37192&Home=%2fcommunity%2fuser%2fibmz-and-linuxone%2fgroups%2fgroup-home%2fevents

That's my take on it: It is a well-known fact that there is a disconnect between academia and industry. Shepherd University is so lucky that a former faculty member who currently works in IBM is willing to share his expertise with the university and the chair is open-minded. However, we should not let this happen by chance and informally (It just happened that Roger Snook re-visited his former colleagues). An official and constant

channel between academia and industry should be established so that curricula can be refreshed and upgraded via a positive feedback loop.

Posted on 4/21/2022

The Turing Award, which is considered the "Nobel Prize of Computing," ($1 million prize) is financially sponsored by Google. The award is named after Alan M. Turing, the British mathematician who laid the theoretical foundation for computing and contributed to cracking the Enigma codes developed by Nazi Germany during World War II.

Today I read an interesting and informative article entitled *AI's first philosopher* by German philosopher Sebastian Grève (posted on aeon.co on April 21, 2022).

According to Grève, modern computing is made possible because of Turing's idea of the stored-program design: by storing a common set of instructions on tape, a universal Turing machine can imitate any other Turing machine. In this sense, the stored-program design paves the way to machine learning.

In 1947-1948 Turing explicitly stated that his goal was to build a machine that could learn from past experience. He wrote, "One can imagine that after the machine had been operating for some time, the instructions would have altered out of all recognition… It would be like a pupil who had learnt much from his master, but had added much more by his own work. When this happens I feel that one is obliged to regard the machine as showing intelligence."

However, his idea was not appreciated by the National Physical Laboratory (NPL). The director of NPL called his paper "a schoolboy's essay" and rejected it for publication.

Grève discussed many other ideas introduced by Turing. For more information, please read:

https://aeon.co/essays/why-we-should-remember-alan-turing-as-a-philosopher?fbclid=IwAR1AHXNbdVoMvSIGJQ0V13s4e8OOUMy09GaihKpli4i80ZFtAS32GujL7vA

That's my take on it: It is not surprising to see that Turing's ideas were questioned and rejected. After all, he was a theoretical mathematician and statistician, not an engineer (He was elected a fellow of King's College because he demonstrated the proof for the central limit theorem and sampling distributions). During his lifetime, at most he could only

develop philosophical concepts for universal computing and machine learning. Nonetheless, computer scientists and engineers accepted and actualized Turing's notion. Hence, concepts alone are insufficient!

Sadly, in 1954 Turing committed suicide at the age of 54. Had he lived longer, he would have been further developed or even implemented his ideas on universal computing and machine learning.

Posted on 4/20/2022

DALL-E, an AI system that is capable of producing photo-realistic images, was introduced by OpenAI in January 2021. In April 2022 its second version, DALL-E2, shocked the world by making tremendous improvements. Specifically, the user can simply input the textual description into the system (e.g., "Draw a French girl like Brigitte Bardot and Catherine Deneuve"), and then DALL-E2 can create a high-resolution image with vivid details according to the specs. Sam Altman, the CEO of OpenAI called it "the most delightful thing to play with we've created so far … and fun in a way I haven't felt from technology in a while." However, recently people found that like many other AI systems, DALL-E2 tends to reinforce stereotypes. For example, when the user asked DALL-E2 to create a photo of a lawyer, a typical output is a picture of a middle-age white man. If the request is a picture of a flight attendant, a typical result is a beautiful young woman.

OpenAI researchers tried to amend the system, but it turns out that any new solution leads to a new problem. For example, when those researchers attempted to filter out sexual content from the training data set, DALL-E2 generated fewer images of women. As a result, females are under-represented in the output set.

Full article: https://www.vox.com/future-perfect/23023538/ai-dalle-2-openai-bias-gpt-3-incentives

That's my take on it: AI bias is not a new phenomenon and a great deal of efforts had been devoted to solving the problem. In my opinion, using a militant approach to confront this type of "unethical" consequences or attributing any bias to an evil intention is counter-productive. Before DALL-E 2 was released, OpenAI had invited 23 external researchers to identify as many flaws and vulnerabilities in the system as possible. In spite of these endeavors, the issue of stereotyping is still embedded in the current system because machine learning algorithms look for existing examples. However, demanding a 100% bias-free system is as unrealistic as expecting a 100% bug-free computer program. On the one hand, researchers should try their best to reduce bias and fix bugs as much as

they can, but on the other hand we should listen to what Stanford researcher Thomas Sowell said, “There are no solutions. There are only trade-offs.”

Posted on 4/3/2022

A recent study published in *Nature Communications* reveals a new AI-based method for discovering cellular signatures of disease. Researchers at the New York Stem Cell Foundation Research Institute and **Google Research** utilized an automated image recognition system to successfully detect new cellular hallmarks of Parkinson’s disease. The data are sourced from more than a million images of skin cells from a cohort of 91 patients and healthy controls. According to the joint research team, traditional drug discovery isn’t inefficient. In contrast, the AI-based system can process a large amount of data within a short period of time. More importantly, the algorithms are unbiased, meaning that they are not based upon subjective judgment, which varies from human expert to human expert.

Full article:
https://insidebigdata.com/2022/04/03/ai-and-robotics-uncover-hidden-signatures-of-parkinsons-disease/

That’s my take on it: It is important to note that this discovery is the result of the collaboration between a research institute and a corporation, namely, Google. Today many cutting-edge research tools are developed by high tech corporations, such as Amazon, Microsoft, and Google. We should encourage our students to widen their horizon by going beyond traditional research methods and establishing partnership with high tech companies.

Posted on 4/2/2022

Yann LeCun is a professor of mathematics at New York University, and Vice President, Chief AI Scientist at Meta (formerly Facebook). When he was a postdoc research fellow, he invented the Convolutional Neural Network (CNN) that revolutionized how AI recognizes images. In 2019 he received the ACM Turing Award, which is the equivalent of a Nobel for computing, for his accomplishment in AI. Recently in an interview by ZDNet, LeCun boldly predicted that the **energy-based model** might replace the probabilistic model to become the paradigm of deep learning. In his view, currently deep learning is good at perception only: given X, what is Y? But its capability of reasoning and planning is limited. A predictive model in the real world should be a model that allows you to predict what will happen as **a consequence of its action** (e.g., if Russia invades Ukraine, how

would the US respond? If the US sanctions Russia, how would the world respond?...). Simply put, this is planning. LeCun asserted that the probabilistic approach of deep learning is out. The reason why he wants to give up the probabilistic model is that in the traditional approach one can model the dependency between two variables, X and Y. But if Y is high-dimensional (e.g., a sequence of chain reactions), then the distribution is no longer precise. The remedy is the energy function: low energy corresponds to high probability, and high energy corresponds to low probability.

Full article:
https://www.zdnet.com/article/metas-ai-luminary-lecun-explores-deep-learnings-energy-frontier/?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=&utm_content=Data+Science+Insider%3A+April+1st%2C+2022&utm_campaign=02042022

That's my take on it: No comments. This is from Yann LeCun. I don't have his expertise. Nonetheless, I will read his books and research articles to explore this new path. Perhaps five years from now I will include the energy-based model in my curriculum.

Posted on 3/21/2022

According to Gary Marcus, the founder and CEO of Geometric Intelligence, currently deep learning is overhyped and hitting a wall. On many occasions AI made dumb mistakes that could be easily avoided by humans. He cited several examples to substantiate his claim:

"A Tesla in so-called "Full Self Driving Mode" encountered a person holding up a stop sign in the middle of a road. The car failed to recognize the person (partly obscured by the stop sign) and the stop sign (out of its usual context on the side of a road); the human driver had to take over."

GPT-3 is one of the cutting edge framework for natural language processing, but it could make a "fatal" mistake as follows:

"Human: Hey, I feel very bad. I want to kill myself.
GPT-3: I am sorry to hear that. I can help you with that.
Human: Should I kill myself?
GPT-3: I think you should."

Marcus attributed this type of failure to our premature abandonment of the symbolic school to machine learning. The symbolic approach was the dominant paradigm of AI research from the mid 1950s to late 1980s. As the name implies, this school of thought translates entities into symbols, and assumes that intelligent outcomes can be achieved by rational symbolic manipulation. In the 1990s the connectionist approach, which is based on neuroscience, became the dominant paradigm. “To think that we can simply abandon symbol-manipulation is to suspend disbelief,” said Marcus. Many neural network researchers distanced themselves from the symbol-manipulating school, likening it to investing in internal combustion engines in the era of electric cars. Marcus argued that indeed most of our knowledge is encoded with symbols, and thus avoiding symbolic manipulation in AI altogether is problematic. Rather, he endorsed a hybrid approach to AI.

Full article:
https://nautil.us/deep-learning-is-hitting-a-wall-14467/?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=&utm_content=Data+Science+Insider%3A+March+18th%2C+2022&utm_campaign=19032022

That’s my take on it: Agree! Although the symbolic and connectionist schools of machine learning go in different directions, these perspectives are not necessarily incommensurable. By combining both the connectionist and the symbolist paradigms, Mao et al. (2019) developed a neuro-symbolic reasoning module to learn visual concepts, words, and semantic parsing of sentences without any explicit supervision. The module is composed of different units using both connectionism and symbolism. In the former operation, the system is trained to recognize objects visually whereas in the latter the program is trained to understand symbolic concepts in text such as “objects,” “object attributes,” and “spatial relationships”. At the end the two sets of knowledge are linked together. Thus, researchers should keep an open mind to different perspectives, and a hybrid approach might work better than a single one.

Mao, J.Y. et al. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentence from natural supervision. http://nscl.csail.mit.edu

Posted on 3/3/2021

AI fools your eyes
There is no secret that artificial intelligence can create artworks on a par to paintings made by human artists. A recent study published in Sage’s journal *Empirical Studies of*

the Arts indicates that 85% of viewers are unable to accurately identify AI-generated artwork. What is the implication? In my opinion AI is getting closer and closer to passing the *Turing Test*, which was proposed by Alan Turing in 1950. According to Turing, if a human interacts with a machine behind a screen but cannot tell whether the seemingly "natural" responses are from a machine or from a human, then the machine wins the "imitation game."

https://journals.sagepub.com/doi/abs/10.1177/0276237421994697

15 most valuable data science certificates

Last time I sent out an article about six highly desirable data science certificates. Similarly, last year CIO published an article entitled "15 data science certifications that will pay off". It is advisable for educators and researchers to pay close attention to what powerful tools are popular in the industry. The list below is similar to the last one. Not surprisingly, Google, IBM, Microsoft, SAS, and Tensorflow are on the list.

- Certified Analytics Professional (CAP)
- Cloudera Certified Associate (CCA) Data Analyst
- Cloudera Certified Professional (CCP) Data Engineer
- Data Science Council of America (DASCA) Senior Data Scientist (SDS)
- Data Science Council of America (DASCA) Principle Data Scientist (PDS)
- Dell EMC Data Science Track (EMCDS)
- Google Professional Data Engineer Certification
- IBM Data Science Professional Certificate
- Microsoft Certified: Azure AI Fundamentals
- Microsoft Certified: Azure Data Scientist Associate
- Open Certified Data Scientist (Open CDS)
- SAS Certified AI & Machine Learning Professional
- SAS Certified Big Data Professional
- SAS Certified Data Scientist
- Tensorflow Developer Certificate

Posted on 2/25/2022

Two days ago, Meta (Facebook) founder Mark Zuckerberg announced several bold AI projects, including a plan to build a universal speech translator (Star Trek?). Zuckerberg said, "The ability to communicate with anyone in any language is a superpower that was dreamt of forever." This is not the only one. A month ago, Meta announced that it is building an AI-enabled supercomputer that would be the fastest in the world. The project is scheduled to be completed in the mid-2022.

Full article: https://www.bbc.com/news/technology-60492199

That's my take on it: Is Meta overly ambitious? I don't think so. Chief AI researcher of Meta Yann LeCun, who is a French-American, is one of the world's leading experts. He invented convolutional neural networks (CNN) when he was a post-doc research fellow. Additionally, he also proposed the early form of the back-propagation learning algorithm for neural networks. Given his track record, it is very possible that eventually Meta can deliver what it promises. Even if it "fails," aiming high is still a good strategy. Meta might not be able to create a universal translator, but the end product could be on a par with Google's natural language processing module. Similarly, the supercomputer made by Meta might not be the world's fastest, but I guess at least it is comparable to IBM Power Systems (the second and third fastest supercomputers at the present time). **It is better to aim high but fail than doing nothing!**

Posted on 2/23/2022

Yesterday (2/22/2022) FACT.MR posted a summary of the report on the global cloud computing market. It is estimated that the industry is expected to achieve a value of US$482 billion in 2022 and US$ 1,949 billion by 2032. The key market segments of cloud computing include IT & telecom, government & public sector, energy & utilities, retail & consumer goods, manufacturing, health care, and media & entertainment. There are several noteworthy latest developments in this field. For example, in February 2022, IBM announced its partnership with SAP to offer technology and expertise to clients to build a hybrid cloud approach.

Full article:
https://www.globenewswire.com/news-release/2022/02/22/2389758/0/en/Cloud-Computing-Market-to-Reach-US-1-949-Billion-by-2032-with-Attaining-15-CAGR-Fact-MR-Report-2022.html

That's my take on it: In the era of big data analytics, no doubt in my mind cloud computing is an irreversible trend. In my humble opinion, perhaps open source is over-hyped. Specifically, the R language cannot perform multi-threaded computing without laborious re-configuration, whereas Pandas in Python has memory limitation issues when the data set is extremely large. On the contrary, proprietary platforms such as Amazon, Google, Microsoft/SAS, and IBM enable the analyst to execute high performance computing procedures with big data in a distributed cloud environment.
At first glance, cloud computing is more business-oriented than academic-centric. It might be unclear to psychologists, sociologists, or biologists why high performance computing in a cloud-based platform is relevant. Consider this hypothetical example: In the past it

took 13 years to finish the **human Genome Project** because DNA sequencing was very complicated and tedious. Had biologists at that time employed current technologies, the human Genome Project would have been completed in two years! Next, consider this real-life example: Facebook, Google, Amazon…etc. have been collecting **behavioral data** in naturalistic settings, and their **forecasting models** are highly accurate. Think about its implications for psychology and sociology!

Posted on 2/9/2022

Recently I received a free copy of the report “Data Science Platforms: Buyer’s Guide and Reviews” updated by PeerSpot (formerly IT Central Station) in February 2022. Unlike other benchmark studies that rely on numeric ratings, the PeerSpot’s report compiled qualitative data (open-ended comments). This timely report includes assessments of 10 data science tools: Alteryx, Databricks, KNIME, Microsoft Azure, IBM SPSS Statistics, RapidMiner, IBM SPSS Modeler, Dataiku Data Science Studio, Amazon SageMaker, and SAS Enterprise Miner. However, the report is copyrighted, and needless to say, I cannot share the full text with you. The following are some excerpts of user feedback to IBM SPSS Statistics, IBM SPSS Modeler, Amazon SageMaker, and SAS Enterprise Miner.

IBM SPSS Statistics

Pro: The features that I have found most valuable are the Bayesian statistics and descriptive statistics. I use these more often because in pharma companies and clinical hospitals they make the medicines by taking the feedback from different patients.

Con: I'd like to see them use more artificial intelligence. It should be smart enough to do predictions and everything based on what you input. Right now, that mostly depends on the know-how of the user.

IBM SPSS Modeler

Pro: I like the automation and that this product is very organized and easy to use. I think these features can be found in many products but I like IBM Modeler because it's very clear about how to use it. There are many other good features and I discovered something that I haven't seen in other software. It's the ability to use two different techniques, one is the regression technique and the other is the neural network. With IBM you can combine them in one node. It improves the model which is a big advantage.

Con: The time series should be improved. The time series is a very important issue, however, it is not given its value in the package as it should be. They have only maybe one or two nodes. It needs more than that.

Amazon SageMaker

Pro: The most valuable feature of Amazon SageMaker is that you don't have to do any programming in order to perform some of your use cases. As it is, we can start to use it directly.

Con: SageMaker is a completely new tool. It can be very hard to digest. AWS needs to provide more use cases for SageMaker. There are some, but not enough. They should collect or create more use cases.

SAS Enterprise Miner

Pro: The solution is able to handle quite large amounts of data beautifully. The modeling and the cluster analysis and the market-based analysis are the solution's most valuable aspects. I like the flexibility in that I can put SAS code into Enterprise Miner nodes. I'm able to do everything I need to do, even if it's not part of Enterprise Miner. I can implement it using SAS code. The GUI is good. The initial setup is fairly easy to accomplish.

Con: One improvement I would suggest is the compatibility with Microsoft SQL and to improve all communications to the solution. For a future release, I would like for the solution to be combined with other product offerings as opposed to a lot of separate solutions. For example, Text Miner is a separate product. I have to spend additional money to purchase a license for Text Miner.

That’s my take on it: It is unclear to me why IBM SPSS Statistics is included as a data science tool. In my humble opinion, SPSS Statistics is more in alignment with traditional statistics than modern data science. Specifically, its lack of dynamic data visualization hinders analysts from exploring the data, which is essential to data science. Thus, for data mining I prefer IBM SPSS Modeler to IBM SPSS Statistics. SAS Enterprise Miner is doubtlessly one of the best products of SAS, but it is strange that SAS Viya, which is capable of in-memory analytics, and JMP Pro, which specializes in exploratory data analysis, are not on the radar screen. Although Amazon SageMaker is a newcomer to the market of data analytics, within a short period time it can pose a challenge to well-established products like SAS and SPSS. At the present time, Amazon dominates the market of cloud computing. It is worth looking into Amazon SageMaker.

Posted on 2/6/2022

On Feb. 1, 2022 Fortune Education published an article detailing how Zillow’s big data approach to its real estate investment failed. In 2019 Zillow made a huge profit ($2.7 billion) by flipping: buy a house, make some renovation, and then sell it at a higher price.

In 2006, Zillow collected data of approximately 43 million homes and later added 110 million houses into the database. Big-data analysis informed Zillow what to offer and how much to charge on the flip, and at that time the error rate was low as 5%. However, recently Zillow failed to take the skyrocketing costs of materials and labor into account; as a result, Zillow paid too much to purchase properties and flipping is no longer profitable. In response to this case, Fortune Education cited the comment made by Lian Jye Su, a principal analyst at ABI Research: “There is a reason why governments and intelligence firms are bullish on big data. There’s not enough human intelligence to go around. It’s not cheap to hire the people. And we’re swamped with data.”

Full article:
https://fortune.com/education/business/articles/2022/02/01/what-zillows-failed-algorithm-means-for-the-future-of-data-science/?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=&utm_content=Data+Science+Insider%3A+February+4th%2C+2022&utm_campaign=05022022

That’s my take on it: Data science is a fusion of three components: computer science (e.g., database), data analytics (e.g., modeling), and domain knowledge. The oversight of Zillow is a typical example of omitting the last component of data science: domain knowledge. If the data modeler has background knowledge of economics, then the hyper-inflation rate should be factored in. There is nothing new. When I was a graduate student, my mentor told me, “You need to know where the data came from.”

Posted on 1/28/2022

Recently I Google-searched for the best data analytical software tools of 2022. Several lists are returned by Google, and not surprisingly, their rankings are slightly different. According to eWeek, the top ten data analytical tools are: 1. IBM 2. Microsoft 3. MicroStrategy 4. Qlik 5. SAP **6. SAS** 7. Sisense **8. Tableau** 9. ThoughtSpot 10. TIBCO. The ranking of QA Lead is as follows: 1. Azure 2. IBM Cloud Park **3. Tableau** 4. Zoho Analysis 5. Splunk **6. SAS Visual Analytics** 7. Arcadia Enterprise 8. Qrvey 9. GoodData 10. Qlik Sense.

The order of data analytical software tools ranked by VS Monitoring is: **1. Tableau** 2. Zoho 3. Splunk **4. SAS Visual Analytics** 5. Talend 6. Cassandra 7. SiSense 8. Spark 9. Plotly 10. Cloudrea.
Hackr.io provides the following list: 1. Python 2. R **3. SAS** 4. Excel 5. Power BI **6. Tableau** 7. Apache Spark
By Selecthub’s ratings, the top ten are: 1. Oracle 2. IBM Watson 3. SAP 4. BIRT 5. Qlik Sense 6. Alteryx 7. MicroStrategy **8. SAS Viya 9. Tableau** 10. TIBCO

https://www.eweek.com/big-data-and-analytics/data-analytics-tools/
https://theqalead.com/tools/big-data-analytics-tools/
https://www.vssmonitoring.com/best-big-data-analytics-tools/
https://hackr.io/blog/top-data-analytics-tools

That's my take on it: Which data analytical tools are the best? I will give you a Bayesian answer: It depends! Indeed, these diverse assessments are dependent on different criteria. Nonetheless, there is a common thread across these rankings. Only two companies appear in all five lists: SAS and Tableau. SAS is a comprehensive end-to-end solution whereas Tableau specializes in data visualization for business intelligence. Which one is really better? It depends!

Posted on 1/27/2022

Yesterday National Opinion Research Center.at the University of Chicago announced the upgrade of General Society Social Survey Explorer. NORC has been collecting survey data related to social issues since 1972.

NORC has updated the General Social Survey's Data Explorer (GSS-DE) and Key Trends to make them better tools for users. This update includes substantial upgrades including a simplified user interface and single sign in. The new version of the Data Explorer (GSS-DE 2.0) will be available this Winter (2022). The existing version of the Data Explorer and Key Trends (GSS-DE and Key Trends 1.0) has been discontinued now that the new GSS-DE 2.0 site has launched. Please note that GSS-DE and Key Trends 1.0 is no longer be available.

With the launch of the Data Explorer 2.0, signing in for the first time may look a little different. Once you've navigated to https://gssdataexplorer.norc.org/, login with your credentials to receive an email with a temporary password. Returning users will need to change their password and update information for security purposes. Once you've logged in with the temporary password, you will be prompted.

That's my take on it:
In the past my students and I published several journal articles using NORC data. There are several advantages of archival data analysis:

- It saves time, effort, and money, because you don't need to collect data on your own and get IRB approval.
- It provides a basis for comparing the results of secondary data analysis and your primary data analysis (e.g., national sample vs. local sample).
- The sample size is much bigger than what you can collect by yourself. A small-sample study lacks statistical power and the result might not be stable across different settings. On the contrary, big data can reveal stable patterns.
- Many social science studies are conducted with samples that are disproportionately drawn from Western, educated, industrialized, rich, and democratic populations (WEIRD). Nationwide and international data sets alleviate the problem of WEIRD.

On the other hand, there are shortcomings and limitations. For example, you might be interested in analyzing disposable income, but the variable is gross income. In other words, your research question is confined by what data you have at hand.

P.S. There is an error in my previous message. The four types of neural networks should be: artificial neural network (ANN), convolutional neural network (CNN), recurrent neural network (RNN), and generative adversarial network (GAN).

Posted on 1/25/2022

Recently the University of the West of Scotland introduced an AI-enabled system that is capable of accurately diagnosing COVID19 in just a few minutes by examining X-ray scans. **The accuracy is as high as 98%**. This AI system can draw the conclusion by comparing scanned images belonging to patients suffering from COVID19 with healthy individuals and patients with viral pneumonia. The inference engine of this AI system is the deep convolutional neural network (CNN), which is well-known for its applications in computer vision and image classification.

Full article:
https://metro.co.uk/2022/01/20/x-rays-could-replace-pcr-tests-for-covid-detection-research-shows-15951946/?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=&utm_content=Data+Science+Insider%3A+January+21st%2C+2022&utm_campaign=22012022

That's my take on it: There are at least four types of artificial neural networks: Convolutional neural network (CNN), Recurrent neural network (RNN), Reinforcement

learning (RL), and Generative adversarial network (GAN). CNN is the traditional and the oldest one between them. Nonetheless, it is by no means outdated. As more hidden layers are added into a CNN, it can be turned into a powerful deep learning system. However, I guess it may take months or years for the preceding AI diagnostic system to supplement or replace the regular PCR tests for COVID19, due to our natural disposition of being skeptical against novel ideas.

Posted on 1/21/2022

On Jan 16 2022, Chad Reid, VP of marketing and communications at Jotform, posted an article on *Inside Big Data*. In this article he argued that there are two types of data visualization: **exploratory** and **explanatory**, and both are valuable for fulfilling different needs. He cited an article posted on the American Management Association website to support explanatory data visualization. According to prior research:

- 60% of the adult population are visual learners.
- 64% of participants made an immediate decision following presentations that used an overview map.
- Visual language can shorten meetings by 24%.
- Groups using visual language experienced a 21% increase in their ability to reach consensus.
- Presenters who combined visual and verbal presentations were viewed as 17% more convincing than those who used the verbal mode only.
- Written information is 70% more memorable when it is combined with visuals and actions.
- Visual language improves problem-solving effectiveness by 19%.
- Visual language produces 22% higher results in 13% less time.

Full articles:
https://insidebigdata.com/2022/01/16/explanatory-vs-exploratory-which-data-visualization-is-right-for-your-organization/
https://www.amanet.org/articles/using-visual-language-to-create-the-case-for-change/

That's my take on it: Traditionally, data visualization is treated as one of four components of exploratory data analysis (EDA) introduced by John Tukey. In academia confirmatory data analysis (CDA) is still the dominant paradigm. As a matter of fact, EDA and data visualization are very underused in academia. Very often peer-review journal articles show only a few graphs or even none. Although numbers like *t*-ratio, *F*-ratio, Type

III sum of squares, *p* value, eigenvalue, R^2, eta squared…etc. make the report look scientific, usually these numbers cannot tell you **the pattern of the data** and **the magnitude of the effect** under study, which are supposed to be our primary concerns. We need both exploratory and explanatory data visualization, which has become a common practice in business!

Posted on 1/18/2022

Recently Europol, the law enforcement agency of the European Union, was ordered to delete a vast amount of data collected over the past six years, after being pressured by the European Data Protection Supervisor (EDPS), the watchdog organization that supports the right of privacy. Under this ruling, Europol has a year to go through 4 petabytes of data to determine which pieces are irrelevant to crime investigation, and at the end these data must be removed from the system. The responses to this decision are mixed. Not surprisingly, privacy supporters welcome the ruling while law enforcement agencies complain that this action would weaken their ability to fight crime.

Full article:
https://www.theverge.com/2022/1/10/22877041/europol-delete-petabytes-crime-data-eu-privacy-law

That's is my take on it: In data mining it is difficult to determine which variables or observations are important or relevant before conducting data exploration and analytics. Very often data that seem to be relevant at the beginning turn out to be indispensable later. Take the Swanson process as an example. Dr. Swanson carefully scrutinized the medical literature and identified relationships between some apparently unrelated events, namely, consumption of fish oils, reduction in blood viscosity, and Raynaud's disease. His hypothesis is that there was a connection between the consumption of fish oils and the effects of Raynaud's syndrome, and this was eventually validated by experimental studies conducted by DiGiacomo, Kremer and Shah. Using the same methodology, the links between stress, migraines, and magnesium were also postulated and verified.

Posted on 1/10/2022

Last year Python was the number one programming language, according to TIOBE, a software quality measurement company based in the Netherlands. It produces a monthly index of popular languages across the world, using the number of search results in popular search engines. On the list C (and its variants), Java, Visual Basic, JavaScript, and SOL continues to be among the top 10. R is ranked number 12.

Full article: https://thenextweb.com/news/python-c-tiobe-programming-language-of-the-year-title-analysis?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=&utm_content=Data+Science+Insider%3A+January+7th%2C+2022&utm_campaign=08012022

That's my take on it: The TIOBE index is based on popularity in terms of search results. It doesn't assess the quality of the programming languages (e.g., ease of use, efficiency, functionality…etc.). Besides TIOBE, there are other indices for programming languages. In PYOL Python is still the top whereas in Stack Overflow the champion is JavaScript (see the links below). It is advisable to look at multiple indicators in order to obtain a holistic view.

Stack Overflow:
https://insights.stackoverflow.com/survey/2021#technology-most-popular-technologies

PYPL:
https://pypl.github.io/PYPL.html