ADSA Data Science Community Newsletter

Data Science Community Newsletter features journalism, research papers and tools/software for September 8, 2022.

Please let us (Micaela Parker, Brad Stenger, Laura Norén) know if you have something to add to next week's newsletter. We are grateful for the generous financial support from the Academic Data Science Alliance.

HELPING STUDENTS GET MORE OUT OF VIDEO LECTURES

Hari Subramonyam from **Stanford**, **Yining Cao** from **University of California, San Diego**, and **Eytan Adar** from **University of Michigan** created a tool, **VideoSticker**, for students consuming lectures via video. The tool uses natural language processing and image recognition to extract portions of text, charts, and other images out of video feeds and place it into a digital notes document that students can interact with.

The researchers believe this will make the process of consuming an educational lecture more interactive and thus more pedagogically effective. VideoSticker, "turns the passive experience of watching video into an active one where the students learn by manipulating images and text and supplementing those elements with their own explanations and comments to cement comprehension and recall."

The team does not plan to support VideoSticker in an open source model. Instead, they intend to commercialize it.

Featured Job

See the ADSA Jobs Page for more opportunities.

Berkeley
UNIVERSITY OF CALIFORNIA

Assistant Professor — Statistics, Data Science, Neuroscience. Statistics Department & Helen Wills Neuroscience Institute — UC Berkeley. Berkeley, CA.

MICROSOFT LAST WEEK, GOOGLE THIS WEEK: AI CODE COMPLETE

Last week we wrote about CS professors **reconsidering** how they give exams because of Microsoft/ GitHub's CoPilot AI. Students using **Microsoft's** open source **CoPilot** can write many old exam problems correctly with the help of AI and little else (e.g. little help from students' coding skills).

This week we note that **Google** has a similar AI-based code completion tool. The **technical specs** are worth noting. The model — yes, just one model — is trained on 8 different languages (C++, Java, Python, Go, Typescript, Proto, Kotlin, and Dart). They got "improved or equal performance across all languages, removing the need for dedicated models." They also shared the number of parameters — approximately 500 million — which they say "gives a good tradeoff for high prediction accuracy with low latency and resource cost." They also note that humans can add to the efficacy by writing good code that makes it into the training set.

The Google developers aren't exactly working themselves out of a job. First, they were able to "reduce the coding iteration time for Googlers by 6%" which is nice, but not nearly enough to result in layoffs. Second, they got a solid suggested-code "acceptance rate of 25-34%" which leaves a lot of code that needs to be crafted by a human artisan of the trade.

Featured Job

Berkeley
UNIVERSITY OF CALIFORNIA

Assistant Professor — Statistics, Data Science. Department of Statistics — UC Berkeley. Berkeley, CA.

PRIVACY-PRESERVING CAMERAS

New **research** from a team led by **Aydogan Ozcan** at **UCLA** proves that a privacy-preserving camera is possible. Images can capture all sorts of sensitive information — social security numbers and similarly sensitive text, faces that may not have consented to facial recognition, nude bodies — and blurring these things out after the fact means the original could still be hacked or reconstructed with AI. A new camera technique can be trained to produce clear imaging only of the type of image it is trained to see, producing only fuzzy outlines of the rest.

The camera uses deep learning to produce composite images based on input from "up to seven transmissive, 3D-printed surfaces, each composed of tens of thousands of diffractive features at the scale of the wavelength of light". The deep learning model is trained on what to capture so, "when the input objects from the target optical fields appear in front of the camera lens, they are captured as high-quality images, whereas when the input objects belong to other undesired classes, they are optically erased, forming random and low-intensity patterns that become" background noise.

While unlikely to be commercialized soon, this privacy-preserving camera philosophically ups the ante for photo-based privacy. Indigenous cultures have long rejected the shutterfly selfie-happy mentality ushered in by cameras in every hand. Now there's going to be a technologically sophisticated way to adopt a consent mode for photography (e.g. anyone who hasn't consented to be the training "object" would only appear as background noise). Clearly, indigenous culture did not *need* a fancy technology to make the point that capturing a person's image should always already operate within a trust-and-respect mode. Finally the technology has caught up to that advanced ideology.

PRIVACY AND REMOTE-PROCTORED EXAMS — UNCONSTITUTIONAL?

Aaron Ogletree brought **a case** against **Cleveland State University** in a U.S. District Court for the use of a fairly standard remote exam proctoring tool. The tool required a scan of Ogletree's immediate environs to determine whether he might be surrounded by material he could use to cheat. At the time, he had tax documents out, which were recorded.

Federal Judge J. Philip Calabrese ruled in Ogletree's favor on Fourth Amendment grounds. This case

could have far-reaching implications for both remote exam proctoring and any other surveillance of students that takes places in their homes or rooms. The judge wrote: "students have a subjective expectation of privacy in their houses, and especially in their bedrooms and society recognizes that expectation as reasonable." He was unmoved by the argument that many schools use remote proctoring and most students are OK with it, writing, "although the record shows that no student, other than Mr. Ogletree, ever objected to the scans, the facts also implicate the core places where society, to the extent it can agree on much these days, recognizes reasonable and legitimate privacy interests — namely, the home. Though schools may routinely employ remote technology to peer into houses without objection from some, most, or nearly all students, it does not follow that others might not object to the virtual intrusion into their homes or that the routine use of a practice such as room scans does not violate a privacy interest that society recognizes as reasonable, both factually and legally."

He went on to note that the ubiquity of a technology does not negate the Fourth Amendment either. This could have implications for *any* monitoring of students at home, not just photo and video-based monitoring. "The ubiquity of a particular technology or its applications does not directly bear on that analysis [....] At the Fourth Amendment's 'very core' lies 'the right of a man to retreat into his own home and there be free from unreasonable governmental intrusion'." Based on this decision, both students and employees should be afforded privacy at home. This is a key privacy ruling. Please consider what your remote proctoring software does and whether you want to put your university in a position to be sued this semester.

PHDS IN FRANCE MUST TAKE RESEARCH ETHICS OATH

France has **instated** a law that requires PhDs conducting research to take an ethical oath as part of their thesis defense. A draft version published in *Science* reads: "I pledge, to the greatest of my ability, to continue to maintain integrity in my relationship to knowledge, to my methods and to my results." It will not be legally binding and ethics scholar **Boudewijn de Bruin** has criticized it for being too generic, but symbolically, it is an interesting step.

AI RAPPER/AVATAR CUT SHORT — TOO CRASSLY STEREOTYPICAL

While it is completely unsurprising that an AI trained to **generate a rapper/avatar** turned out to be grossly stereotypical, it is surprising that nobody in the product-launch chain of command at **Capitol Records** saw this coming. In this very newsletter we have featured many scholars, regulators, and other intellectual leaders calling for algorithmic fairness assessments. Please, if you're planning to release AI, use the available tool sets and perhaps a colleague from the social sciences, to conduct an impact assessment (algorithmic impact and social impact, please).

The silver lining is that the project — which probably wasn't cheap — was cut short after only a week.

Featured Job

Assistant Teaching Professor — Statistics, Data Science. Department of Statistics — UC Berkeley. Berkeley, CA.

WEARABLES UPDATE

MIT researchers have **created** a chip-less wearable that uses a gallium nitride film that will generate surface acoustic waves to send data to smartphones. The idea is that smaller wearables will be tolerated better by users. A team at **Texas A&M** has **shown** that another novel material — molybdenum disulfide — is capable of combining with a gelatin substance to behave like a flexible hydrogel. The upshot is that a tattoo-like patch can be applied to the skin to constantly monitor motion, potentially useful in monitoring patients Parkinson symptoms (TBD).

NPR science reporter **Ari Daniel** recently visited **Epicore Biosystems**, a startup in Cambridge, MA, that

makes a personalized sweat sensor to monitor hydration, developed in partnership with **Gatorade**. During the **on-air demo** the company's chief scientific officer, **A.J. Aranyosi**, provided 100 ounces of his own sweat in the name of data collection. Not an appealing image, even with the non-visual presentation. "I've been the heaviest sweater since we founded the company," Aranyosi told NPR. Elsewhere, engineers at **EPFL** have (with no Aranyosi juice) advanced the **state of the art for cortisol detection** in a sweat-reading wearable. Cortisol is the body's primary stress-response hormone and an important metabolic regulator.

Apple is **launching** a 7-year study to determine if its watches can reduce the use of blood thinners in the at-risk population for heart disease by more accurately identifying which people are at the highest risk of stroke. Using blood thinners does reduce the risk of heart disease, but it increases the risk of stroke.

MORE ENERGY EFFICIENT CHIPS

A team from **Stanford** and **University of California, San Diego** has **made** an important incremental improvement in the efficiency of chips. "A massive amount of…energy goes towards moving the data between the compute unit (where the data is processed) and the memory unit (where the data is stored)," so the team aimed to reduce movement between the units. Existing design techniques already aim to cut down on the distance between the units. The improvement here was to "integrate resistive random-access memory (RRAM) to retain data even when the power is off and compute-in memory (CIM) to enable AI computing directly in the memory unit."

Not exactly a revolution, but an important incremental advance nonetheless. We need all the energy reduction we can get.

WHERE COLLEGES HAVE OPENED, COVID CASES UP 37%

Where colleges have already opened, **data show** that local COVID case rates have jumped 37%. Like last year, there are big divergences in the ways campuses are handling pandemic precautions. Some are mandating masks, others are not, some are only mandating them in "required" areas like classrooms and elevators.

The current report did not take campus mask mandates into account so we'll have to wait for more data to see how the mandates are working this time around. It's also unclear if students will be able to participate in class remotely if they test positive on these campuses.

In COVID-related **research**, a team of medical and public health experts at the **Broad Institute** and **Harvard Medical School** with a couple AI researchers on loan from **Uber AI** have developed a model to predict which variants will lead to COVID surges.

SPORTS GAMBLING — BIG QUESTIONS ABOUT ADDICTION

Gambling is the addiction with the highest suicidality rate and the highest recidivism rate. It's often seen as less dangerous, less harmful than drug addiction…but it has certainly destroyed enough lives, companies, and families to be taken seriously. The growth in gambling apps like JackPocket and sports betting legalization has addiction researchers like **Timothy Fong** of **UCLA Gambling Studies Program** **worried**. Fong is concerned that sports betting is perhaps more tantalizing because it can be perceived as a game of skill rather than a game of chance. People who bet on sports often know about the game, the players, the coaches, and contextual factors like the weather. This ties into the bettor's "illusion of control," a key factor for all problematic gamblers as well as their sense of intelligence. Losing in sports betting, then, is not perceived as a matter of chance. A loss could undermine the bettor's sense of self, with dire consequences. A recent **paper** by a team of 11 researchers from Spain, Canada, and the US found that, "sports betting, relative to non-sports betting, has been more strongly linked to gambling problems and cognitive distortions related to illusion of control, probability control and interpretive control." They also found a higher severity of gambling disorder among sports bettors.

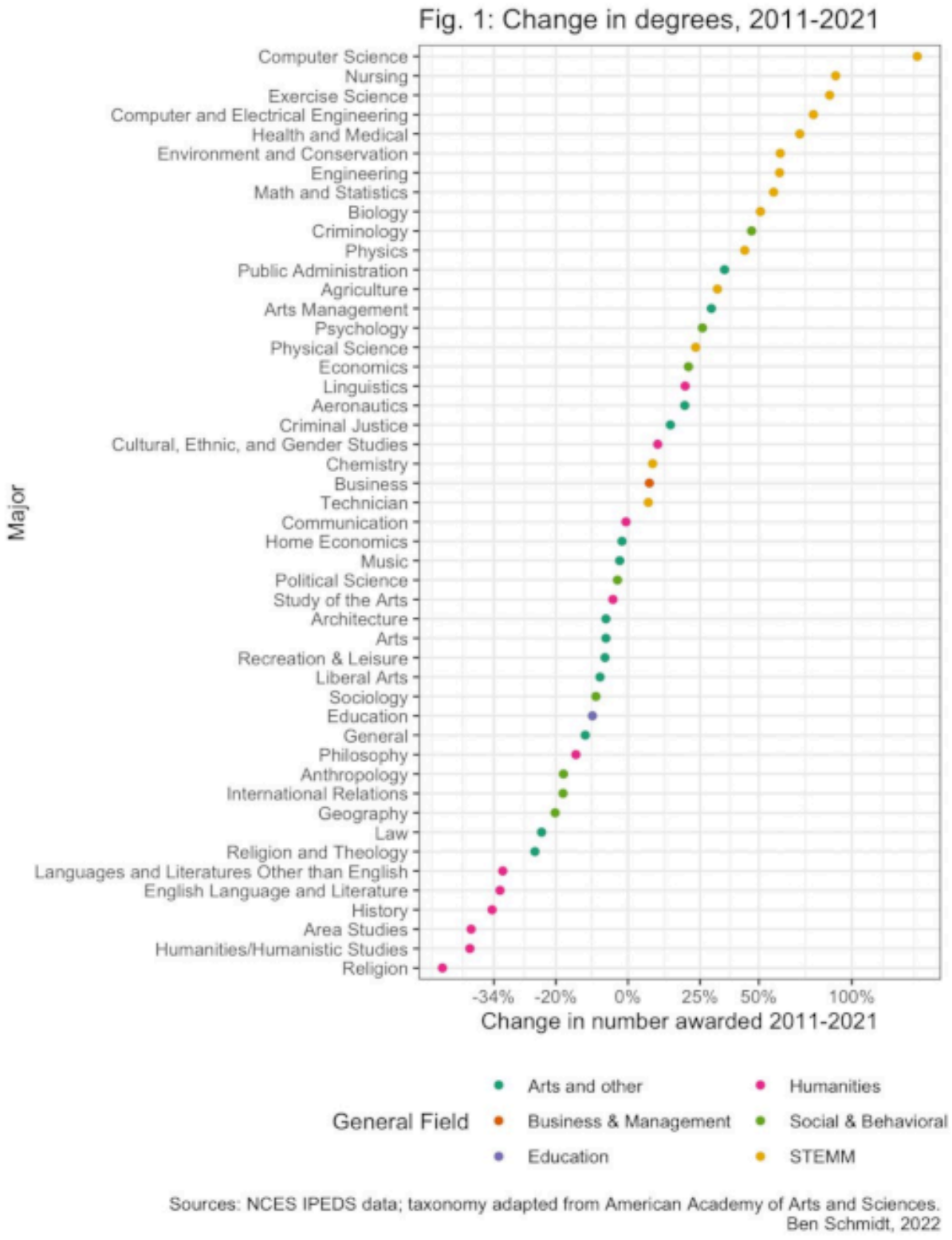
This comes at a time when many US states and now universities are allowing — sometimes even encouraging — sports betting. It's not just fun and games.

NEW PROGRAMS, FOLLOW THE MONEY

Click through to access **a structured spreadsheet** of New Programs and money moving around in academic data science.

DATA VISUALIZATION OF THE WEEK

By **Benjamin Schmidt**, **quote-tweeted** by **Yann LeCun** on August 24, 2022



Relatedly, nearly 2 in 5 college graduates **regret** their choice of major, but only 24% of engineering majors are in the regretful group. The **last word** goes to **Thomas Dietterich** who tweeted, "Our hardest societal problems (including, but not just, #natsec) require strong social science + humanities expertise coupled with technical expertise," while linking to a thread about education by security researcher, **Nadya**

Bliss.

Deadlines

Studies/Surveys

2022 marks the sixth year of the #dataviz state of the industry survey

"Missing Data: Why Don't We All Take the 2022 DVS SOTI Survey? ... The **Data Visualization Society** is running its annual survey. Contribute your data this month, and invite people you know."

Education Opportunities

I'm chairing the hiring committee for the @FlatironCCA "Flatiron Research Fellow" postdoc fellowship this year

"If you're on the astro postdoc job market this year (or if you know someone who is), read on!"

RFPs

Apply Now: $300,000 for Data Systems to Safeguard Human Rights

"The Data and Society Accelerator Program is run annually through a cohort model by the **Patrick J. McGovern Foundation**." The deadline for submissions is September 19.

Tools & Resources

Language Models have taken #NLProc by storm. Even if you don't directly work in NLP, you have likely heard and possibly, used language models. But ever wonder who came up with the term "Language Model"?

Twitter, Delip Rao from September 6, 2022

"Recently I went on that quest, and I want to take you along with me."

Interested in recruiting undergraduates to work with you on a research project?

Twitter, Alex Tamkin from September 1, 2022

"I recently wrote up a guide"

Next Gen Stats: New advanced metrics you NEED to know for the 2022 NFL season

NFL, News, The Next Gen Stats Analytics Team from September 1, 2022

"Next Gen Stats is excited to give you a sneak peek into two new metrics we are launching this season, inspired by submissions from each of the last two **Big Data Bowl** contests: Coverage Classification and Expected Return Yards."

About Us: The Data Science Community Newsletter was founded in 2015 in the Moore-Sloan Data Science Environment at NYU's Center for Data Science. We continue to be supported by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation through the Academic Data Science Alliance. Our archive of newsletters is at https://academicdatascience.org/resources/newsletter. Our mailing address is 1037 NE 65th St #316; Seattle, WA 98115.

