**Capstone 3 Project Ideas:**

1. What are people’s opinions of psychedelic drugs, compared with other drugs, for treatment of C-PTSD or other mental illnesses? (sentiment analysis using natural language processing from several heavily-used psychedelic experience archives)
   1. Why: There is currently a resurgence of interest in psychedelics, with plant-based medicinal drugs having been decriminalized in multiple cities and even states recently. This is quickly becoming  a promising market for not just psychedelic enthusiasts but healthcare workers, entrepreneurs, and investors alike. Organizations like MAPS (Multidisciplinary Association for Psychedelic Studies) are leading advocacy efforts, often supporting reputable studies to demonstrate the benefits of psychedelics as justification for their decriminalization. While studies are crucial, there are also thousands of people who have, for decades, anonymously shared information about their experiences with psychedelics via online platforms. To my knowledge, there has never been any effort to extract summative conclusions from these reviews, which are often in narrative form. Such a project could be a genuine contribution to a growing community within the healthcare industry.
   2. Data:
      1. Scrape erowid, psychonaut wiki, reddit forums, and other psychedelic experience archives
      2. For comparison, reviews of prescription psych meds: <https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29>
      3. And <https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29>

1. How biased are news articles written for children? (comparative sentiment analysis on multiple corpora of articles)
   1. Why: People with reading levels lower than what is typically considered “8th grade” have fewer options available for quality sources of news. Of course, no news is actually totally objective, but some news articles do more than others to invite emotional attachment to certain concepts or opinions. It is possible that even if somebody does not intend to write a heavily-biased article, they may demonstrate a strong sentiment about the news topic nevertheless. My hypothesis is that this may happen more often in news written for children, because emotional content could be used to make the news more interesting in lieu of using complex grammar forms that draw advanced readers into more challenging texts. A sentiment analysis could help publishers coach reporters to improve the quality of articles for children.
   2. Data: scrape NewsELA, CommonLit, and other sources of news for children. Compare with scrapes of similarly-themed articles from NPR. Of course, NPR is not entirely unbiased, but I will use it as a benchmark because it is a publicly-funded and widely-read news source.

1. After a wildfire is detected in Northern California, when should lowlands Oakland residents prepare to be indoors or wear ventilator masks when outside? (Air Quality Index Forecasting via Satellite)
   1. Why: Too many people in Oakland are forced to live outside, and most everybody who is housed enjoys being outdoors most days. Fire season causes high anxiety, as we are often unsure throughout the late summer and early fall whether it will be safe to breathe unfiltered air. Particulate matter sensors can tell us on any given day how un/safe the air is, but we could be more well-prepared if we had better predictive models for telling us whether any given wildfire is likely to send smoke in our direction. This could help people prepare emotionally, adjust plans for outdoor activities, and ramp up mask distribution among unhoused people before the smoke is already settling into our lowlands areas.
   2. Data:
      1. NASA Wildfire Satellite Images: <https://www.earthdata.nasa.gov/learn/pathfinders/wildfires-data-pathfinder>
      2. NASA Environmental Justice Catalog (I marked those that include data intended for predicting PM2.5 levels): <https://docs.google.com/spreadsheets/d/1Pz1XqqulYctHuIfZSU7Wx1ey_NODNtRy/edit?usp=sharing&ouid=105446281087536735842&rtpof=true&sd=true>
      3. AQI levels history: <https://www.epa.gov/outdoor-air-quality-data/download-daily-data>
   3. Preliminary high-level problem decomposition:
      1. Gather data from Northern California for a particular chunk of time, the same timeframe for all data sources: wildfire locations and spread patterns, on-the-ground air quality measurements, air quality satellite measurements.
      2. Analyze image data to seek any possible correlations among wildfires -> atmospheric imagery -> AQI.
      3. Select features of wildfire tabular records and/or wildfire satellite images and/or air quality satellite images that might be used in predicting Oakland’s AQI; use these features to model and make predictions.
      4. While I’m excited about this project, I do believe this problem decomposition would require further honing in order to be viable, as it involves indexing image analysis to time-scaled tabular data.
   4. Some air quality prediction work already done in other regions: <https://www.drivendata.org/competitions/88/competition-air-quality-pm/page/423/>

**Alternatives in case my top 3 don’t work out:**

1. Is this a phishing link? (natural language processing to classify and flag potential malware)
   1. Why: I love cybersecurity, and this is a very common area for cybersecurity and data science to intersect. It’s helpful for security teams at all sorts of organizations to have programs that help them flag potentially malicious links rather than checking all reported links manually. This is a case where it’s very important to minimize false negatives, but nuance is also required because no efficiency is gained from a system that flags all links as potentially malicious. I know this is a task that could be great to practice because it’s exactly what somebody told me during an informational interview that they do with their entire job as a cybersecurity data scientist.
   2. Data: <https://www.unb.ca/cic/datasets/andmal2020.html>

1. Where is there a leak? (classifying images of pipes)
   1. Why: Any organization tasked with maintaining infrastructure needs reliable tools for diagnosing decay. Visual analysis can provide faster problem identification than manual inspection. I live in earthquake territory and in an area notorious for poorly-funded public works. I am therefore interested in any strategies for improving the convenience of keeping our basic needs met and our structures safe.
   2. Data: <https://universe.roboflow.com/underwaterpipes/underwater_pipes_orginal_pictures>

1. Zine categorization:
   1. Why: Zines have long been an important way of sharing information, ideas, and news among people who do not have access to more formal routes of publishing. Luckily, many historical zines are being digitized and archived. Many archives, however, are not searchable or have extremely limited search features. Natural language processing could be used to categorize and tag content for easier access. There are several nonprofit and community organizations that would love to receive the results of this work.
   2. Data:
      1. <https://www.sproutdistro.com/catalog/zines/accountability-consent/accounting-for-ourselves>
      2. <https://anarchist-archive.org/>
      3. <https://theanarchistlibrary.org/special/index>
      4. <https://crimethinc.com/>
      5. More archives exist!

1. Is there breast cancer, and if so, what stage is it? (Image processing and classification with regression)
   1. Why: A family member of mine was just diagnosed with breast cancer. I’m grateful to doctors who have been able to provide such clear information about levels of risk and would like to better understand how they’re using computer vision to provide these clear descriptions and inform effective action plans. This would not be an original contribution to the field, but it could be a first step for me in working more with healthcare-related computer vision, which is a major interest of mine.
   2. Data: <https://computervisiononline.com/dataset/1105138703>

1. Diagnosing Alzhimers with Images:
   1. Why: Half of my grandparents and great- grandparents experienced Alzhimers or other conditions that lead to extreme memory loss. I’m at a high risk for this myself, and I would love to learn more about the role data scientists are playing in its prevention. This project is not original, but this could be a first step in working toward sharing meaningful contributions with the healthcare community.
   2. Data: <https://portal.conp.ca/dataset?id=projects/preventad-open>
   3. Work already done on a similar topic:
      1. <https://www.drivendata.org/competitions/65/clog-loss-alzheimers-research/>
      2. <https://towardsdatascience.com/alzheimer-diagnosis-with-deep-learning-a-survey-265406fa542a>
      3. <https://medium.com/analytics-vidhya/predicting-alzeheimers-disease-using-u-net-algorithm-6a6bb7267188>
      4. <https://towardsdatascience.com/detecting-precursors-of-alzheimers-by-utilizing-deep-learning-a6de0ee0e2d2>

**Capstone 2 Project Ideas (I chose to go with the Instacart project):**

1. What change in revenue might a business expect if its website is designed to direct shoppers to key pages?

Dataset: UCI Machine Learning Repository online shopper browsing activity + shopping results including revenue:

<https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>

Similar: Market Basket Analysis:

<https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis>

2. What courses should CourseERA highlight, recommend to learners, or otherwise promote? What course characteristics can CourseERA recommend to curriculum designers to increase learner engagement? How would this likely impact enrollment?

Dataset: CourseERA Course Details + Reviews: <https://www.kaggle.com/datasets/yasirabdaali/coursera-free-courses-dataset>

3. Crypto APIs: How have security-focused cryptocurrencies (XMR, XIN, MOB) grown in comparison with fiat currency and bitcoin in the past? Given current trends, what can we predict about security-focused crypto growth in the coming year?

<https://min-api.cryptocompare.com/?ref=apilist.fun>

<https://coinmarketcap.com/api/?ref=apilist.fun>

<https://www.worldcoinindex.com/apiservice?ref=apilist.fun>

<https://docs.gemini.com/rest-api/?ref=apilist.fun#rate-limits>

<https://docs.coinapi.io/?ref=apilist.fun#md-docs>

4. What products should Instacart recommend to a shopper? What impact might these recommendations have on sales?

Datasets: Instacart basket contents: <https://www.kaggle.com/c/instacart-market-basket-analysis/data>

5. What are risk factors for alcohol abuse in adults (with whom could early interventions be targeted via mental healthcare providers, apps, etc.)?

Datasets: CDC Behavioral Risk Factor Surveillance System <https://www.cdc.gov/BRFSS/>

California Health Interview Survey: <https://healthpolicy.ucla.edu/chis/Pages/default.aspx>

<https://computervisiononline.com/dataset/1105138635>

**Other: I don’t want to discard the work I did searching and considering so many ideas, because I may want to refer to that in the future, but these are currently pretty far on the back burner:**

- What are the demographics of people utilizing telehealth services, and how might providers increase access and use of telehealth?

Dataset: Medicare telehealth trends: <https://data.cms.gov/summary-statistics-on-use-and-payments/medicare-service-type-reports/medicare-telehealth-trends/data>

- How might Resident Advisor be able to increase sales by promoting tickets for DJs who are popular, niche, rising, etc?

<https://www.kaggle.com/datasets/dgaitsgo/resident-advisor-top-1000-djs>

- When developing curriculum features for upcoming app updates, how can Duolingo use emerging knowledge about optimal spaced repetition to spiral curriculum and present review of vocabulary and language features at a pace that maximizes learner retention and transfer?

* Dataset: <https://paperswithcode.com/dataset/duolingo-spaced-repetition-data>
* Context: The value of spaced repetition is nothing new, but until recently, there has been limited understanding of what characteristics of spacing optimize learning (number of repetitions, complexity of content, time between repetitions, change over time in what types of practice are needed to move learning from short- to long-term memory and become transferable). Improving our understanding of these details could greatly enhance learners’ experiences on duolingo but also in all sorts of digital and even in-person learning environments.
* Relatedly, google datasets provides access to a wide range of data related to duolingo’s business model (market size, number of users, market share, revenue, growth financially and by use, etc.). I believe Duolingo teaches language well, and I’d like to see them continue to succeed, so I would be interested in taking a different approach and analyzing details of their business performance for insights.

538: Lots of sports, politics, culture datasets: <https://data.fivethirtyeight.com/>

Oakland Abandoned Cars

<https://data.oaklandca.gov/Infrastructure/Abandoned-Autos-CCD7-Starting-7-1-17/kuxb-xxnt>

Public APIs: [https://github.com/public-apis/public-apis](https://github.com/public-apis/public-apis#anti-malware)

Stock Images:

<https://api-reference.shutterstock.com/#overview>

<https://www.pexels.com/api/>

<https://cocodataset.org/#home>

<http://places.csail.mit.edu/>

UCI datasets: <https://archive.ics.uci.edu/ml/datasets.php>

ComputerVision: <https://computervisiononline.com/datasets?page=4>

ImageNet (1million+ images from 100+ categories): <https://image-net.org/download.php>

MIT Indoor Scenes labeled images for classification: ​​<https://web.mit.edu/torralba/www/indoor.html>

Future Project Idea: Is this plant an invasive species or native to the East Bay? (classification based on varied images of native and invasive plants)

1. Why: As a child, plant identification was one of my favorite hobbies. Since moving from Minnesota to California, I have occasionally felt dissociated from my surroundings because I’m not familiar with many of the plants. In particular, I don’t have a clear sense of whether any “natural” areas actually represent old growth or whether they’ve been developed by humans. Beyond this personal application, there are several reasons why others would need to know about the presence of invasive plants. Property owners can avoid high water costs in this drought-ridden area by steering clear of or removing invasive species when purchasing new lots.
2. Data:
   1. Invasive Plants: <https://map.dfg.ca.gov/metadata/ds1121.html>
   2. Could also scrub here: <https://wric.ucdavis.edu/information/info_spec_weed.htm>
   3. Native Plants: <https://www.calflora.org/entry/observ.html?track=c#srch=t&cols=p&nphoto=t&taxon=Cortaderia+selloana>
   4. And: <https://calisphere.org/collections/150/>
   5. And: <https://www.cch2.org/portal/collections/index.php>
   6. And: <https://ebcnps.org/ebrare-plant-database/>
   7. Map of which plants are native, where, in California: <https://wildlife.ca.gov/Data/GIS/Vegetation-Data>

Future Project Idea: Which incidents are police going to respond to today? (Time series analysis predicting rates of responses to various categories of reported alleged crimes, i.e. violent, vandalism, etc.)

1. Why: Many city elections hinge on questions of crime and policing, and many nonprofit organizations have dedicated efforts to improve, reduce, or abolish policing. But apart from documentation of police harassing Black and houseless neighbors, it’s often not clear exactly what they do. This analysis could support policy advocacy efforts by bringing clear analysis to how huge portions of city budgets are being spent.
2. Data: <https://data.oaklandca.gov/Public-Safety/CrimeWatch-Maps-Past-90-Days/ym6k-rx7a/data>