CHRISTOPHER CHAO

chris.a.chao@gmail.com | (408) 693-0555 | LinkedIn

OBJECTIVE

Master of Data Science and Bachelor of Biological Sciences with 2 years of professional experience, seeking a challenging position where I can apply my interdisciplinary expertise in genomics, molecular biology, and data analysis to contribute to cutting-edge research and development in the field.

SKILLS

Python (Pandas, EasyOCR, ocrmypdf, nltk, NumPy, SciKit Learn, scipy, matplotlib, RE, py2neo, GeoPandas), R (ggplot2, psych, pracma), SQL, Neo4J and Cypher, Regular Expressions, HTML Web Scraping, Tableau, JSON, RNAseq, CRISPR, BLAST+, PCR

PROFESSIONAL EXPERIENCE

QQ Tech **San Francisco, CA**

Data Scientist (Founding Member) ***Jul 2022 - Present***

- ***Tools and Methods:*** *Python, Tesseract, ocrmypdf, pdfminer.six, PyPDF2 fastapi, httpx, Selenium, JSON*
- Selected by the founders of QQ Tech to be a founding member of QQ Tech's product, Lexempla, a SaaS that focuses on helping customers understand their legal documents by providing our open domain question and answering service and connecting to lawyers through messaging
- Engineering the new product from the ground up by contributing toward end-to-end architecture design and refining the application user workflows
- Developing in-house solutions for preprocessing PDFs and images and use computer vision packages, such as Google's Tesseract OCR, to extract data from input documents to generate customized hOCR confidence level files in XML/HTML format
- Implementing a generic web scraping engine using Selenium to retrieve legal related documents from publicly available sources
- Functioning as a project manager and scrum master to ensure progress of product features are delivered as planned in Jira system; presented project status in Jira metrics
- Authoring, maintaining, and enforcing Confluence pages for proper coding documentation abiding by PEP-257 and PEP-8 standards, including information on unit testing procedures using a combination of pytest and httpx+fastapi calls, and company-wide convention for the setup and use of Python IDE, PyCharm
- Creating and updating unit tests for program algorithms to ensure proper functionalities with no regressions, automatically executed by my implementation with GitHub Actions per push to the remote repo
- Established GitHub pull request and code review protocols for version control of feature, development/integration, stage/QA, and production branches
- Building Docker containers and implementing GitOps to deploy services to Google Cloud Run

SKAEL **San Francisco, CA**

Data Scientist (full-time concurrent with Masters program) ***Apr 2021 – Jul 2022***

- ***Tools and Methods:*** *Python, PaddleOCR, EasyOCR, regex, spacy, nltk, Agglomerative Clustering*
- Selected by the Chief AI Officer and former Data Science professor to be a founding team member of SKAEL's AI team, focused on building the PageAI architecture and algorithms
- Designed and implemented an image processing workflow to extract text from images using OpenCV, EasyOCR, and PaddleOCR, which resulted in over 90% accuracy of read texts
- Built the OCR workflow and naive regular expression rule-based searching, which conducts field extractions from images and PDFs, in addition to NER, part of speech tagging, and auto fillable PDF fields
- Innovated an in-house clustering algorithm to cluster documents based on semantic similarity, which is used by the customer base when sending in documents for scanning
- Documented software use cases for new employees, created tutorials and instructions for software such as ngrok (expose local web apps) and Insomnia (REST API testing), pycharm (Python IDE)
- Partnered with backend and frontend engineers to set up the data pipelines, and product managers

ThermoFisher Scientific **Pleasanton, CA**

COVID-19 Kit Manufacturing ***Apr 2020 – Apr 2021***

- Part-time employment while pursuing Masters degree in Data Science at University of the Pacific
- Handle MS2 phage control and TaqPath COVID-19 assay solutions used for real-time PCR test for in vitro qualitative detection of nucleic acid from the SARS-CoV-2
- Utilize and maintain machinery (Hamilton STAR, BioMicroLab, Tube Capper) for filling, labeling, and kitting COVID-19 test kits, with an average output of 1.5 million test reactions per day

University of Pacific **Stockton, CA**

Undergraduate Research Assistant ***Jun 2019 - Dec 2019***

- ***Tools and Methods:*** Galaxy, IQ-Tree, R language (blastdb), Python, BLAST+
- Assisted a biology professor with research on evolutionary development of ostracods
- Maintained protein sequence data with the use of Galaxy workflow management tool and manipulate (trim, edit, organize) data to produce phylogenetic trees with use of IQ-Tree software
- Ran BLAST searches with protein sequence data and compare similar sequences within NCBI database

Undergraduate Research Assistant ***Aug 2019 - Dec 2019***

- ***Tools and Methods:*** Restriction Digest, Transformation, PCR, Agarose Gel Electrophoresis, SDS-PAGE, EMSA, Gel Extraction and Visualization
- Assisted a genetics professor with research on binding sites of the *six* gene
- Analyzed shifts in electrophoretic mobility shift assays to see specific if binding was present in specific strands of DNA and identified two possible locations for binding
- Helped lead research group in carrying out tasks such as setting gels, pipetting, using the PCR machine, and visualizing gels under UV

EDUCATION

University of the Pacific **Stockton, CA**

Master of Data Science *Graduated May 2022*

Bachelor of Science in Biological Sciences *Graduated Dec 2019*

CHRISTOPHER CHAO

chris.a.chao@gmail.com | (408) 693-0555 | LinkedIn

PROJECTS

Sacramento Kings/Fanatics - Merchandising and Ticket Sales Insights ***Spring 2022***

- ***Tools and Methods:*** *Python (Pandas)*
- Partnered with Sacramento Kings and Fanatics to perform analytics on merchandise and ticket data to provide recommendations for advertising-matching items to specific demographics of customers
- Techniques include email and events analysis via natural language processing, customer segmentation using both geographic and needs-based analysis, and market basket analysis to predict complementary items

Sentiment Analysis - Twitter Posts Sentiment for Video Games 2022 ***Fall 2021***

- ***Tools and Methods:*** *Python (nltk, Pandas)*
- Two-phase project testing sentiment hypothesis relating to three video games released in Fall 2021
- Extracted Twitter posts with TwiPy relating to and classified their sentiments: positive, neutral, or negative
- Performed a secondary test post game release comparing their rating and score given by users against the pre-release sentiments from Twitter and identified that pre-release sentiments on Twitter is a good indicator of what the video game's post-release score will be

Regular Expressions - Syllabus Parser ***Spring 2021***

- ***Tools and Methods:*** *Python (RE, Pandas), Regular Expressions*
- Created a syllabus parser that utilizes regular expressions to extract 12 different fields from any syllabus
- Fields extracted are: course name, course ID, units, instructor's name, office hours, instructor's email, instructor's phone number, day of class, class time, semester, year, and whether the class is online or not
- Course ID, email, phone number, semester, and year can be extracted with greater than 90% accuracy

World Happiness Report – Clustering Countries by Similarity ***Fall 2020***

- ***Tools and Methods:*** Python, Pandas, NumPy, Matplotlib, sklearn.cluster, sklearn.hierarchy, GeoPandas, K-means unsupervised clustering, KNN supervised clustering
- Identified countries with high happiness factors and provided insights to improve social health, generosity, government trust, family social support, etc. for countries with lower ratings
- Created a random centroid K-means clustering algorithm to cluster countries into five groups
- Built dendrograms and associations with sklearn.cluster and sklearn.hierarchy algorithms to cluster countries by similarity and mapped the clusters using GeoPandas
- Utilized sklearn.neighbors to predict countries' geographical region based on the UN Statistics Division, with characteristic inputs such as life expectancy, economy, freedom, etc.

Cancer Biology Grant Proposal - Effects of Adenomatous Polyposis Coli Methylation ***Fall 2019***

- ***Tools and Methods:*** CRISPR Knock-out, Western Blot, Selective Methylation, EMSA
- Proposed a method in understanding the canonical wnt-signaling pathway by methylating different combinations of the two APC promoters

Transposon Mutagenesis to Identify Hormogonia Development Proteins ***Spring 2019***

- ***Tools and Methods*****:** RNAseq, Transposon Mutagenesis, Selection, Restriction Digest, Polymerase Chain Reaction (PCR), Agarose Gel Electrophoresis, Pure DNA Quantification, BLASTn and BLASTp, Immunofluorescence/Fluorescent Lectin Staining, Fluorescence Microscopy
- Identified a hypothetical protein in species *Nostoc punctiforme* by using forward genetics (transposon mutagenesis) to knock-out genes to identify hormogonia development proteins
- Utilized expression levels of three sigma factors with identified protein to construct biological pathway of hormogonia development regulation of *Nostoc punctiforme*

Bioinformatics Grant Proposal - Adenomatous Polyposis Coli Expression ***Fall 2018***

- ***Tools and Methods:*** BLAST, RNAseq, CRISPR, PCR, Restriction Digest, Agarose Gel Electrophoresis, Transformation, Screening, Plasmid Miniprep, GST Tag/Affinity Chromatography, Alternative Splicing
- Used Basic Local Alignment Search Tool (BLAST) to identify human homolog of adenomatous polyposis coli cells, which was then alternatively spliced, purified and expressed
- Proposed method of using non-homologous end-joining CRISPR to synthesize non-functional APC protein and to monitor the effects of cells in tissues other than colon and rectum

INTERESTS

Genetics, Biopsychology, Social Psychology, Consumer Analytics, Operational Efficiency, Astronomy