Ying-Jung (Helen) Deweese

Phone: 805-284-3236 | Email: yingjungcd@gmail.com | Linkedin: ying-jung-deweese | US Permanent Resident

SUMMARY

Data Scientist with 5+ years of experience in end-to-end data pipeline (ETL), spatial-temporal data analysis, GCP/AWS cloud computing operations, and developing predictive machine learning models to present information for real world decision making.

TECHNICAL SKILLS

- **Programming:** Python (Scikit-learn, Geopandas, Rasterio, GDAL, CARTO, Xarray, Zarr, Dask, Matplotlib, *Pangeo*, Plotly), R (data.table , dplyr , tidyverse , ggplot2, raster, shiny), Javascript, SQL (Postgre SQL), Bash
- **Project Management:** Strategic Planning, Scrum and Agile Methodologies, Training & Mentoring
- **Geographic Tools:** ESRI Arc GIS, Google Earth Engine, QGIS, ENVI, OpenStreetMap
- **Data Science:** Regression Models, Random Forest Classification, Anomaly Detection, Clustering, Kalman filter, Time Series Analysis, Spatial Statistics, Heatmap/Leaflet Visualization, TensorFlow, Keras, PyTorch
- **Computing:** AWS, Google Cloud Platform, Azure, Databricks, BigQuery, Git/GitHub, Apache Spark, openMPI

EXPERIENCE

DESCARTES LABS **Cupertino, CA**
Applied Scientist **2021/11-Present**

- Designed a ML based crop index product for predicting 20% crop growth in U.S. over 5-year periods.
- Performed a crop disease risk assessment for preventing 5% commodity loss for packaged goods companies.
- Delivered a time series ML solution to predict crop yield for fortune 500 clients' price forecasting decision.
- Developed a deep learning approach (LSTM, Transformer) to predict crop health conditions over soybean produced states in U.S.
- Shared the state of arts cloud computing services (AWS/GCP) techniques and resources.

INDEPENDENT CONSULTANT **Cupertino, CA**
Data Scientist /Machine Learning Engineer (Part time) **2021/06-2022/06**

- Produced a normalizing strategy for GPS data to increase traveler behavior prediction accuracy by 5%.
- Reduced property evaluation time by 50% with a dashboard, including ML model output, maps, and metrics.
- Mentored two team projects from tech startups and national labs to submit paper for NeurIPS conferences.
- Organized New in ML workshop in ICML conference 2022.

BANK OF AMERICA **Charlotte, NC (Remote)**
Quantitative Risk Analyst **2021**

- Linked data curation and quality control pipelines in SQL and R, reducing data preparation time by 25%.
- Provided climate risk and natural disaster knowledge to economists for establishing climate finance models.
- Refactored a macro economic model pipeline in Python, saving 30% of processing time for forecasting GDP.
- Performed code review in gitlab from syntax to high level concepts for science teams in agile environment (Jira).

THE CLIMATE CORPORATION **San Francisco, CA**
Data Scientist **2020/01-2021/01**

- Developed multiple imagery-based strategies for establishing computer-vision models to detect crop growth.
- Generated ~1.5 m resolution field map from spatial model based on GPS tracking points for farmers to monitor field health and growth performance.
- Contributed weather features into machine learning (random forest, LSTM) models to increase accuracy by 3% for crop disease prediction.

- Established and deployed an internal Python package for calculating evaluation metrics via CI/CD approaches.
- Utilized PySpark to query and aggregate yield data at field level across U.S to test sensitivity of models.

INSIGHT DATA SCIENCE **Seattle, WA**
Data Science Fellow **2019/06-09**

- Utilized ETL (Python, PostgreSQL) to clean 10GB data from water volume, billing data in Redshift database.
- Implemented an anomaly detector approach in Python to identify outliers (5% of data) in water usage patterns.
- Provided a systematic ML approach for a data scientist to determine meters with high water usage amounts.

APPLIED PHYSICS LAB. & E-SCIENCE INSTITUTE, UNIV. OF WASHINGTON **Seattle, WA**
Post-Doc Research Associate **2018/08 – 2019/06**

- Produced seasonal trend maps with NASA researchers examining a 20% decrease in Himalayan groundwater.
- Evaluated 4 regression, distance, and Bayesian models to streamline the spatial interpolation on point-based data.
- Used python for ETL and visualization of GCM/RCM, hydrologic model output and imagery data.
- Created a pangeo based tutorial for stakeholders to learn geospatial python APIs (GDAL, geopandas).

GEOGRAPHY DEPARTMENT, UC SANTA BARBARA **Santa Barbara, CA**
Graduate Student Researcher / Teaching Assistant **2012/09 – 2017/06**

- Established a time series model in Python to estimate the groundwater recharge timing for identifying droughts.
- Applied spectral index and water usage data to estimate 20% vegetation changes and irrigation in urban areas.
- Generated an imagery preprocessing pipeline for 10m resolution imagery covering Los Angeles County.
- Managed five group projects on GIS application for natural resources management within 10 weeks.

DATA USAGE EXPERIENCE

- **Financial data**: quality assurance and control for macroeconomic data (i.e., rates, currency, GDP).
- **Geospatial data**: normalized geolocation data and produced small farm maps based on GPS tracking data.
- **Agricultural data**: ingested weather satellite data, tillage, growth stages information into models.
- **Hydrological data:** collected and cleaned water usage, rain, and streamflow gauge data for time series analysis.
- **Imagery data:** preprocessing imagery (atmospheric, radiometric, and geometric correction) for spectral index calculation (i.e., Normalized Difference Vegetation Index) based on optical and hyperspectral imagery.

EDUCATION

M.S. in Computer Science, Georgia Institute of Technology **2021-2023**
M.A./Ph.D. in Geography, **University of California Santa Barbara** **2011-2018**
B.A. in **GIS**, **National Taiwan Normal University** **2004-2009**