David Lin

ddlin@berkeley.edu | (617) 412-5665 | U. S. citizen
https://tealeave.github.io/da-wei-lin-data/
5129 Vía Samuel, Yorba Linda, CA 92886

SUMMARY

Data Scientist and Machine Learning Engineer with extensive experience developing scalable pipelines for high-dimensional datasets. Proven expertise in predictive modeling, causal inference, and time-series forecasting. Adept at translating complex data into actionable insights and optimization routines applicable to dynamic environments including energy, healthcare, and bioinformatics. Passionate about leveraging advanced algorithms to forecast trends and optimize resource allocation.

Skills

Causal Inference & Biological Data Analysis:

- **Biological Data & Single-Cell Genomics:** scRNA-seq analysis pipelines; NGS panel QC & coverage modeling; CRISPR screen & RNA-seq time-series.
- Expertise in causal inference methods including propensity score matching, difference-in-differences, and instrumental variables
- Familiarity with target trial emulation concepts for robust observational study designs

Statistical Modeling & Machine Learning:

- Proficient in advanced regression techniques, ensemble methods, and predictive modeling (e.g., XGBoost, Gradient Boosting)
- **Machine Learning & Deep Learning:** Transformer architectures, foundation models, LSTM & autoencoder networks, model compression & optimization.

Programming & Data Management:

- Advanced skills in Python (NumPy, Pandas, Scikit-learn, PyTorch) and R for data analysis and model development; production-quality code, with a focus on performance and readability
- Proficient in SQL and Spark (PySpark) for managing and processing large-scale datasets, including EHR and high-dimensional clinical data

Data Visualization & Communication:

- Skilled in visualizing complex datasets using Plotly, Seaborn, and Plotnine
- Adept at transforming analytical findings into clear, actionable insights for multidisciplinary teams and non-technical stakeholders

Project Management & Collaboration:

- Experienced user of JIRA, GitHub, and Confluence for project coordination and documentation
- Proven ability to collaborate with cross-functional teams in fast-paced, dynamic environments

EXPERIENCE

Data Scientist III

Sapient Bioanalytics **Oct 2023 - present**

- **Led population-scale biomarker identification and risk score modeling for global clinical projects, working across 50,000-sample datasets with 40,000+ metabolomic features.**
- **Spearheaded capacity-building initiatives internationally, organizing workshops and mentoring emerging researchers.**
- **Developed and Optimized a Three-Stage Analytical Pipeline:**
  - **Data Cleaning and Stratification: Streamlined processing for large-scale datasets, ensuring data integrity for subsequent analyses.**
  - **Feature Engineering and Regression Analysis with PySpark:** Employed advanced statistical methods and meta-analysis to extract meaningful features and synthesize results.

- **Data Visualization and Pathway Enrichment Analysis:** Leveraged visualization tools and enrichment analyses to reveal key biological insights.
- **Metabolic Risk Score Development with XGBoost:** Implemented cutting-edge machine learning techniques, hyperparameter optimization (via Optuna), and integrated forecasting elements to dynamically track and predict emerging risk patterns.

- **Conducted in-depth metabolomic and proteomic analyses using mass spectrometry data, integrating time-series aspects where appropriate to monitor trends over treatment cycles.**

Bioinformatic Scientist III

Ambry Genetics **Dec 2019 - Oct 2023**

- Analyzed NGS panel probe coverage, sequencing run consistency, coverage uniformity, detection sensitivity and specificity for improved results.
- Contributed to the development of high-volume oncology panels including CancerNext, RNA, Exome, and Somatic panels.
- Implemented robust statistical strategies during chemistry transitions, reducing false positive rates and improving quality control for CNV calling.
- Employed GradientBoostingRegressor to model CNV counts per sample, evaluating feature importance across sequencing and coverage metrics—showcasing transferable forecasting and regression techniques.

Data Science Fellowship

The Data Incubator **Jun 2019 - Sep 2019**

Postdoctoral Researcher

UC Irvine **Jan 2014 - Dec 2019**

- Developed CRISPR gene editing and viral vector systems to investigate protein modifications and signaling events.
- Pioneered bio-screening platforms and conducted proteomic analysis, leading to a publication in *Nature Communication*.
- Developed bio-screening platform in mammalian cells for p53 reactivating compounds, resulted in a publication in **Nature Communication**
- Analyzed RNA-seq data to identify actionable targets, employing statistical and time-series approaches to monitor expression trends.

EDUCATION

University of California, Berkeley
M.S. Information and Data Science **In progress**

University of California, Irvine **Sep 2007 – Jan 2014**
Ph.D. Biological Sciences

California Institute for Regenerative Medicine (CIRM) fellowship **2009 - 2011**

AWARDS & PUBLICATIONS

Molecular Metabolism | Nutrient control of splice site selection contributes to methionine addiction of cancer **2025**

Submitted | Sphingosine and anti-neoplastic sphingosine analogs activate PP2A and inhibit nuclear import in parallel by engaging PPP2R1A and importins **2025**

Cell Chemical Biology | Discovery of compounds that reactivate p53 mutants in vitro and in vivo **2022**

Journal of Lipid Research | Lipid remodeling in response to methionine stress in MDA-MBA-468 triple-negative breast cancer cells **2021**

Data Science Fellowship | The Data Incubator - San Francisco **2019**

Journal of Biological Chemistry | Microhomology based CRISPR tagging tools for protein tracking, purification, and depletion **2019**

Methods Mol. Biol. | Isolation and characterization of methionine-independent clones from methionine-dependent cancer cells **2019**

U.S. Patent | Chembridge Small Molecules that could enhance p53 activity **2015**

Journal of Cell Science | SAM limitation induces p38 mitogen-activated protein kinase and triggers cell cycle arrest in G1 **2014**

Nature Communication | Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53 **2013**

Cell Cycle | Downregulation of Cdc6 and pre-replication complexes in response to methionine stress in breast cancer cells **2012**

Journal of Biological Chemistry | Transforming growth factor β up-regulates cysteine-rich protein 2 in vascular smooth muscle cells via activating transcription factor 2 **2008**

Genes to Cells | Identification of a putative human mitochondrial thymidine monophosphate kinase associated with monocytic/macrophage terminal differentiation **2008**