Rex W. Douglass Ph.D.

Location: Austin, Tx (Remote)
Portfolio: www.rexdouglass.com
Email: rexdouglass@gmail.com
GitHub: github.com/rexdouglass

Principal Applied Scientist — LLM Information Extraction & Machine Vision

16+ years owning end-to-end ML systems (LLM IE, document AI, industrial machine vision). Repeated wins in **accuracy**, **cost**, and **throughput**: 95%+ field-level extraction at scale, 47% cost reduction on thousand-page legal documents, and labeling automation that cut human clicks by 1,000×. Deep focus on **data quality** and **measurement** across tabular, text, geospatial, video, and time series.

Skills/Stack

Large Language Models (LLMs) / Generative AI | Python | R | SQL | Causal Inference | Natural Language Processing (NLP) | GIS | Machine Vision | Time Series | Linear Models | Random Forests | Boosted Trees (XGBoost, LightGBM) | Neural Networks

Experience

Principal Applied Scientist — **RIOS Intelligent Machines** — **2025–**

Overhauled and automated the entire image labeling pipeline for training object detectors for edge devices/robotics applications. Built three in-house tools for my team:

- **Stratified Video Frame Sampler** - automatically discovered rare events in 10k hours of footage [*fastdup | UMAP]*
- **Tool-Assisted Image Labeler** - 1,000x reduction in annotation clicks (200k→200) with real-time in-browser prediction and confirmation [*Autoencoder | CUDA Random Forest]*
- **Fully Automated Object Labeler** - automated labeling of moving products and materials [*DINOv3, Depth-Anything-V2*]

Senior Applied Scientist — **Microsoft** — **2022–2025**

Built our group's LLM infrastructure for large-scale experimentation from proof of concept through end of pilot. Developed multiple information extraction pipelines in different customer domains including legal, accounting, and product delivery.

- **Custom LLM Infrastructure** - standardized onboarding new information extraction use cases to just 2 artifacts (custom prompt template + JSON codebook) [*Azure OpenAI Service | Azure AI Doc Intelligence | Azure SQL | Azure DevOps]*
- **Supplier Invoices** - 95%+ accuracy across hundreds of extraction fields using detailed structured output codebooks
- **Construction Permits** - 47% cost reduction in processing thousand-page legal permit requirements with better accuracy

Director — **Machine Learning for Social Science Lab - UCSD** — **2016–2022**

Responsible for full research design and technology stack on the center's $5M in external research grants. Led a large team of postdoctoral, graduate, and undergraduate researchers.

- **CrisisEvents.org** - extracted 10k+ crisis events from natural language text using human coders and large language models [*Large Language Models (LLaMA) | Hierarchical clustering on graphs | spaCy | Shiny | UMAP*]
- **Measuring the Landscape of Civil War -** corrected measurement error in geo-referencing of places in unstructured text [*Gradient-boosted trees (XGBoost) | Locality-sensitive hashing (LSH)*]
- **Clean Covid Counts** - estimated true U.S. COVID-19 infections at a resolution of County-Age-Day (N=2.5M+) [*JAX | NumPyro*]

President (Consulting Vehicle) — **Stability Analytics Incorporated** — **(Concurrent) 2016–2025**

Responsible for the technology stack, research design, and sales of $500K+ in corporate research contracts.

- **Machine Vision for Crowd Counting and Demographics** - crowd size and demographic estimates of violent protest events from images [*Keras | YOLO | Prodigy*]
- **Natural Language Processing for Scientific Literature Discovery**- topic discovery and citation network analysis across thousands of scientific documents [*Structural Topic Modeling (STM)*]

Postdoctoral Scholar — **Department of Mathematics - UCSD** — **2012–2015**

Responsible for the machine learning stack and data acquisition and management for a team of mathematicians.

- **High-Resolution Population Estimates from Telecommunications Data -** inferred local population counts from phone calls and satellite imagery [*Random Forest | Spatial cross-validation | Landsat satellite imagery*]
- **Machine Learning for Military Intelligence -** unsupervised interrogation of dirty unverifiable declassified military intelligence [*Biclustering | Random Forest | Multiple Correspondence Analysis (MCA)]*

Education

Ph.D. — **Princeton University - Department of Politics** — **2009–2012**

Dissertation work on quantitative approaches to military intelligence, including wartime interviews in Kabul, Afghanistan and declassified historical archival work in the U.S. [*Item Response Theory (IRT) Models | PostGIS | PostgreSQL | QGIS | R | Stata]*

M.A. — **Princeton University - Department of Politics** — **2007–2009**

General exam in Quantitative Methods & Game Theory, American Politics, and International Relations

B.A. — **University of Texas at Austin** — **2003–2007**

Led large scale document digitization project on nuclear weapons proliferation across 7 archives in U.S., U.K., and India