Rex W. Douglass PhD

📍 San Diego, CA (www) rexdouglass.com ✉ rexdouglass@gmail.com @RexDouglass rexdouglass

Computational social scientist with 13 years of experience executing large data science and machine learning projects from scratch. Expert in data quality, measurement, and multi-modal data (e.g. tabular, text, geospatial, and time series).

Skills

R | Python | SQL | Causal Inference | Natural Language Processing (NLP) | GIS | Machine Vision | Time Series | Linear Models | Random Forest | Gradient Boosted Trees (XGBoost , LightGBM) | Neural Networks

Positions

Director **Machine Learning for Social Science Lab - cPASS - University of California San Diego** **2016-Present**

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

- **Clean Covid Counts** - Estimates of true U.S. COVID-19 infections at a high resolution of County-Age-Day (2.5m+ obs). Challenges: Severe underreporting, non-random missingness, and measurement drift across signals. Solutions: Bayesian latent variable models integrating signals from wastewater, deaths, tests, and cases. [***Jax | Numpyro***]
- **CrisisEvents.org** - Event extraction from natural language text (10k human coded international conflict events). Challenges: Event abstraction from unstructured text | Developing a novel ontology and a way to objectively benchmark coverage, precision, and recall of our human codings. Solutions: Human labeling with a custom GUI | Developed a new benchmark called an Automated Case Study based on unsupervised clustering of sentences across a very large text corpus. [***Large Language Models (BART, MPNet) | Hierarchical clustering on graphs | SpaCy | Shiny | UMAP***]
- **Measuring the Landscape of Civil War -** Correcting measurement error in geo-referencing of places in unstructured text. Challenges: Ground truthing real world locations mentioned in unstructured text | Maps and gazetteers are heavily biased toward urban/built up locations. Solutions: Developed a unique corpus with both location names and precise military coordinates | Built a supervised ensemble to minimize geocoding error based on local geographic features. [***Gradient boosted trees (XGBoost) | Locality-sensitive hashing (LSH)***]

President **Stability Analytics Incorporated** **2016 -Present**

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

- **Machine-Vision for Crowd Counting and Demographics** - Crowd size and demographic estimates of violent protest events from images. Challenges: Concept drift across cultures and across time. Solutions: Pre-trained classifier using Google Open Images and then fine-tuned on custom imagery collected and balanced across countries and historical decades. [***Keras | Yolo | Prodigy***]
- **Natural Language Processing for Scientific Literature Discovery**- Topic discovery and citation network analysis across thousands of scientific documents. Challenges: Dirty text and difficult source documents creating severe sampling bias issues. Solutions: Heavy pre-processing pipeline and hierarchical topic clustering with cross-validated topic quality metrics. [***Structural Topic Modeling (STM)***]

Postdoctoral Scholar **Department of Mathematics - University of California San Diego** **2012-2015**

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

- **High Resolution Population Estimates from Telecommunications Data -** Infer local population counts from phone calls. Challenges: Thousands of features | Scale variant relationships | Spatial autocorrelation. Solutions: Supervised learning with iterative variable selection | Supervised land cover estimates using Openstreetmap and satellite imagery. [***Random Forest | Spatial cross-validation | Landsat satellite imagery***]
- **Machine Learning for Military Intelligence -** Unsupervised interrogation of dirty unverifiable data. Challenges: Poorly and incorrectly documented data | High dimensional categorical features | Multiple regimes. Solutions: Unsupervised hierarchical biclustering of features and observations into topics and regimes | Dimensionality reduction and clustering followed by confirmatory primary research. [***Biclustering | Random Forest | Multiple Correspondence Analysis (MCA)***]

Education

Phd Princeton University **Department of Politics** **2009-2012**

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

MA Princeton University **Department of Politics** **2007-2009**

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

BA University of Texas at Austin **2003-2007**

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