Matthew Hawthorn

data science + mathematics

contact

+502.440.8676
hawthorn.matthew@gmail.com
github://mattHawthorn

languages

Python
+ numpy/scipy/pandas
+ scikit-learn
+ PyTorch
+ numba
+ dask
+ bokeh

_ => Scala
+ breeze
+ typeclass pattern
+ higher-kinded types

SQL
regex(pressions)?
LaTeX

computing

Linux (debian/Ubuntu)
bash shell
git

interests

record linkage
text analysis
+ entity resolution
+ distributional semantics

network analysis
+ spectral methods

unsupervised learning
+ clustering algorithms
+ dimensionality reduction

education

2015–2016 M.S. Data Science — University of Virginia

Capstone Project + Course work in Data Mining, Machine Learning, Text Mining, Computer Science, Linear Models, and Ethics

2012-2014 M.A., Mathematics — University of Louisville

Graduate course sequences in Combinatorics/Graph Theory, Probability/Statistics, Algebra, Complex Analysis; electives in Functional Analysis, Spectral Graph Theory

2003-2006 B.S., Mathematics — University of Louisville

Pure mathematics concentration. Courses in Analysis, Algebra, Combinatorics, Probability; electives in Fractal Geometry, Symbolic Logic, Coding Theory

experience

9/2017-now S&P Global Market Intelligence — Senior Data Scientist

Developed internal library for record linkage to facilitate merging of 3rd party data sets with internal data, and detection of internal duplicates. Optimized for speed with Cython and developed a modular architecture to for isolation and extensibility of blocking (search space reduction), feature generation, and classification. SQL backend for out-of-memory blocking and linking against a dynamic data set. Employed active learning to reduce bias and maximize impact of limited analyst-annotated training data.

6/2016-9/2017 Commonwealth Computer Research, Inc. — Data Scientist

Contributed to CCRi's custom text mining platform. Evaluated various clustering algorithms for speed and accuracy on clustering dense word representations. Added a custom clustering module, significantly improving downstream entity resolution tasks. Modularized portions of the code base and improved interface to Postgres. Served in a supervisory role on a research contract with the NGA to develop a tool for measuring value of data sources to analyst communities.

12/2015-1/2016 L3 Data Tactics — Data Science Intern

Developed an application to topically summarize, cluster, and visualize a corpus of RFPs (requests for proposal), to assist the company in future decisions regarding which contracts to bid on.

9/2012-5/2014 University of Louisville — Teaching Assistant

Assisted professors in teaching general education mathematics courses: lectured in recitation sections, administered tests and quizzes, tutored students, and graded assignments.

projects

summer 2017 sk-torch
Hobby project

Library to wrap PyTorch deep neural nets in an interface mimicking sklearn, allowing substitution of advanced deep learning models in place of sklearn models in ML pipelines. API allows declarative specification of optimizer, loss, input/output transformations, stopping criteria.

9/2015-4/2016 Trend detection in a large corpus of scientific documents
M.S. Capstone Project, UVA (Advisor: Rafael Alvarado)

Developed a dashboard for Battelle for the exploration of a corpus of hundreds of thousands of scientific articles from 7 commercial databases. Documents are searchable by topical similarity and semantically summarized with an LDA topic model. Frequency trends for terms of interest are visualized, and a trend detection classifier flags trending terms.

awards

5/2016 Best Paper Award, IEEE SIEDS Conference 2016
For the paper *"Revealing the landscape: Detecting trends in a scientific corpus,"* co-authored with Rafael Alvaraodo, Juan Arrivillaga, Dylan Greenleaf