

I am a data scientist driven by a strong desire to find, analyze, and predict patterns in data with a commitment and focus on reproducibility and scalability. My training has strengths in software development, predictive models, social science, as well as quantitative and statistical reasoning, giving me a unique ability to synthesize and manage quantitative data for high level insights.

Current position

Data Scientist

Mattersight

Computer languages + applications

Analytics R (packages: (d)plyr, ggplot2, lme4, ranger, shiny, testthat), Python (modules: Pandas, ElementTree, NumPy), SQL, MapReduce (basic)

Modeling GLM regression models (both simple and hierarchical models), tree-based models, cluster-based models, model ensembles, cross validation, hidden markov models, null hypothesis significance testing

Systems + productivity AWS, version control (git), test-driven development, Mac OS X, Linux/UNIX

Markup + web \LaTeX , (X)HTML, CSS, javascript, PHP **Hardware** microcontrollers, PhaseSpace motion capture

Software

MocapGrip – a complete data processing and analysis pipeline R package github.com/jonkeane/mocapGrip

An R package, that contains a full pipeline for motion capture data collection, from initial data processing to statistical modelling and report generation. The package includes functions to check human annotated data for consistency and errors, as well as a flexible analysis and reporting system for use by novice users.

Articulatory Model of Handshape (AMOHS) – Python module github.com/jonkeane/amohs

A Python implementation of phonological and phonetic properties of handshapes in sign languages, as well as a 3D rendering mechanism to evaluate the full impacts of theoretical specifications and assumptions. This module was designed to be as extensible as possible, and has already found use in a number of other projects.

Pyelan – Python module github.com/jonkeane/pyelan

A Python implementation for reading and generating files compatible with the ELAN annotation software. This allows for quick and sophisticated processing or creation of annotations and the underlying data associated with them (eg video, audio, motion capture). Pyelan is a critical dependency for other packages, for example, it is used in the clipping software (github.com/jonkeane/fflipper) I developed that vastly speeds up video data clipping and processing.

Experience

2016
summer

Data Science Fellow [Data Science for Social Good 2016 Summer Fellowship](#) University of Chicago

Worked on a four person team in partnership with the Metropolitan Nashville Police Department to develop an Early Intervention System (EIS) to predict officers at high risk of having an adverse incident. [project webpage](#) | [github repo](#)

Our model correctly identified 80% of officers who went on to have an adverse incident, while only intervening on 30% of officers in the department. A drastic improvement over using the current state of the art in most police departments: which would have needed to intervene on 67% of the department for the same level of accuracy.

- Designed and lead development of our ETL pipeline (evaluating between a number of frameworks: drake, luigi, etc.).
- Lead data visualization efforts for public presentations, internal stake-holders, and our partner police department.
- Architected a database for efficient and department-independent storage of police records data needed to generate EIS predictions. Contributed to python-based feature engineering, model training, and prediction generation.

2014–2016

Research Coordinator and Post-Doctoral Fellow [Center for Sign, Gesture, and Language](#) University of Chicago

Generated and statistically tested hypotheses about human communication (both linguistic and gestural), including numerous collaborations across a team of leading linguists and gesture researchers.

- Developed a full, automated pipeline (MocapGrip) for data collection using motion capture and video equipment: collection, synchronization, verification, data processing, human annotation, data extraction, exploratory data analysis, data visualization, statistical analysis, and report generation. This pipeline introduced numerous efficiencies, including speeding up initial data preparation by 4x and data preparation for analysis by at least 10x.
- Responsible for statistical analysis of quantitative data collected for various projects. Techniques included: clustering, hierarchical linear and logistic regression models (AKA mixed effects models), power analysis.

- Developed and implemented an innovative and distributed psycholinguistics experiment using R+shiny for stimuli display and data collection, hosted using Amazon Web Services.
- Directly managed undergraduate annotators; managed experimental setup, data ingestion, and statistical analysis for graduate student projects. Contributed to the development and implementation of annotation systems.

2009–2014 *Researcher and Lab Manager* [Chicago Language Modeling](#) and [Sign Language Linguistics](#) Labs University of Chicago
Generated and statistically tested hypotheses about linguistic data using computational approaches to data collection and analysis. Worked on projects both from conception, as well as in a post-hoc analyst and statistics consultant role.

- Developed numerous distributed (semi-crowd sourced) annotation methods to enrich collected data. These methods took advantage of as much latent structure as possible to minimize the number person work-hours that were needed to fully annotate our data. These process cut annotation time in half compared to more traditional transcription and annotation methods.
- Managed undergraduate annotators: workflow development and implementation, reliability, deliverables, *etc.*
- Developed reliable processing pipelines for video, motion capture, and metadata associated with data collection.
- Developed an SQL database and front-end for annotated sign language data, which then facilitated further, more detailed annotation, and data analysis.
- Established and taught a source code version control system for the lab.
- Developed online psycholinguistic experiments using a variety of methodologies: subjective (Likert) rating, forced choice, free response, AB testing as well as in-person psycholinguistic experiments using video and motion capture.
- Developed software to extract annotation data from software, and allow for more flexible extension of that data. This includes the development of fast, easy to use, user-facing video clipping software (eg [pyelan](#) and [flipper](#), see above) that was based on open source technologies already being used in the lab. This software turned a manual, supervised task into an automatic, batch process, saving hundreds of person work-hours.
- Developed hardware needed for various experiments, including a microcontroller-based button box, and equipment for the reliable and efficient setup and placement of motion capture markers. The button box, for example, gave the lab additional flexibility needed in our lab, at $1/10$ the standard cost for off the shelf equivalents.

2004–2008 *Network Administrator & Print Design Specialist* [Florida Technology Transfer Center](#) University of Florida
Supported and managed computer infrastructure for 20-25 end users and managed desktop support personnel.

- Configured and managed all servers including active directory, web, file, and email servers. Successfully planned, purchased, configured, and implemented rollouts of Microsoft Exchange, as well as a near hot swap backup system for file and active directory servers.
- Responsible for typesetting and graphic design of various publications released by the center. Designed logos and identities for special projects and events.

Education

Certified [Software](#) and [Data](#) carpentry instructor.

2014 PhD (with honors) in Linguistics *Advisors* Jason Riggle and Diane Brentari University of Chicago
Dissertation Towards an articulatory model of handshape: What fingerspelling tells us about the phonetics and phonology of handshape in ASL.

2007 BA (Magna cum Laude) in Linguistics *Advisor* D. Gary Miller University of Florida

Awards + funding

2013–2015 *NSF Doctoral Dissertation Research Improvement Grant* Coarticulation and the phonetics of fingerspelling (BCS1251807)

2013–2014 *Mellon Humanities Dissertation-Year Fellowship*

2013 *Nominated for best student presentation* at TISLR 11 for *Towards an articulatory model of handshape*

References + publications

Reference contact information or letters are available by request. My publications, talks, and posters can be found at pubs.jonkeane.com.

Human languages

English (native); American Sign Language (fluent); Swedish (inter.); German (reading); Japanese and Hebrew (basic)