Daniel Chen

chend@vt.edu • +1 (917) 880 9254 • https://daniel.rbind.io/

Education

2015-2021 (expected)

Virginia Tech (Blacksburg, VA)

PhD: Genetics, Bioinformatics, and Computational Biology

Thesis: Data Science Education for Biomedical and Health Sciences

2012-2014

Columbia University Mailman School of Public Health (New York City, NY)

MPH: Epidemiology

Certificate: Advanced Epidemiology

Thesis: Spread of Beliefs Towards a Behavior in a Social Network

2007-2012

The Macaulay Honors College at CUNY Hunter College (New York City, NY)

BA: Psychology

Concentration: Behavioral Neuroscience

Minors: Biology, Computer Science

Research

University of Virginia, Social and Decision Analytics Division, Biocomplexity Institute

Nov 2018 - March 2019

Research Associate

Arlington, Virginia

- Build data pipeline to collect open source repositories from CRAN and PyPI to calculate cost of open source software

Virginia Tech, Social and Decision Analytics Lab, Biocomplexity Institute of Virginia Tech

Sep 2017 - Nov 2018

Research Associate, Data Engineer

Arlington, Virginia

- Co-instructor for the Data Science for the Public Good program and main point of contact for code and technical issues
- Maintain backend data storage platforms (RStudio Server, Shiny Server, PostgreSQL databases, Wikis, and Dashboards) with docker and docker-compose
- Teach and build processes for best research practices around data security, reproducibility, and replicability

Columbia University School of Nursing

May 2014 - Aug 2014

Data Analyst

New York, NY

- Analyzed survey data on mass causality preparedness that lead to 2 published papers
- Prepared hospital data to measure care pathways using a network approach

Center for Injury Epidemiology and Prevention at Columbia University 2013 - 2014

- Build an agent-based using the Repast Java framework to simulate alcohol and violence in NYC

Industry

Pandarrow

Jan 2019 - Present

Self-Employed

- Create and maintain learning materials for Pearson in Pandas (Python) and Git
- Various R projects

RStudio

Jun 2019 - Aug 2019

Education Team Intern

- Maintain the {`gradethis`} package, an autograder for R code
- Improve package API for ease of use within {`learnr`} interactive R documents and as a stand-alone code grader
- Utilize R's non-standard evaluation and metaprogramming system to check code result and code expression

Lander Analytics

Jan 2014 - Present

Data Scientist

- Teach Machine Learning and Dashboards R
- Teach Data Analysis and Programming in Python
- Various R projects

NYU Langone Medical Center

2010 - 2012

New York, NY

- Data manager and clinical trials coordinator

Funding

- 2018, Virginia Early Childhood Foundation (VCEF, Baseline Distinct Counts of Select Data Sets to Support ECIDS Establishment, PI - Aaron Schroeder, $75,000, Data Engineer
- 2017-2018, Mitre Corporation, Local Data Sources to Build a Comprehensive Community-Based Understanding of Complex National Health Problems, PI - David Higdon and Sallie Keller, $93,938, Project to continue in 2019-2020, Data Engineer
- 2016-2022, U.S. Army Research Institute (ARI), Individual and Team Performance, The Social Component of The Human Dimension: Leveraging Existing DoD Data Towards Optimized Individual And Team Performance in the Army, PI - Sallie Keller, $3,027,401, Data Engineer
- 2015-2018, U.S. Army Research Institute (ARI), The Social Component of The Human Dimension: Leveraging Existing DoD Data Towards Optimized Individual And Team Performance in the Army, PI - Sallie Keller $1.6 M Data Engineer
- 2018-2019, U.S. Army Research Institute (ARI), Towards an Integrated Data Framework for Understanding the Context of Military Environments - 1 Year Extension, PI - Sallie Keller, $286,826, Data Engineer

Book

- Pandas for Everyone. Addison-Wesley Professional. Daniel Y. Chen. 2018.
  ISBN: **978-0134546933**

Book Chapters

- Building Data Science and Literacy Communities of Practice at Academic Libraries Daniel Chen, Anne Brown. In Review. 2021.
- Systems Science and Population Health. Systems of Behavior and Population Health. Mark Orr, Kathryn Ziemer, Daniel Chen. 2017.
  ISBN: **9780190492397**

Peer-Reviewed Publications

ORCID: `http://orcid.org/0000-0003-3857-1741`

Google Scholar: **h56YmqQAAAAJ**

- Demographics, perceptions, and socioeconomic factors affecting influenza vaccination among adults in the United States. 2018. Kaja M. Abbas, Gloria J. Kang, **Daniel Chen**, Stephen R. Werre, Achla Marathe PeerJ.
  DOI: `https://doi.org/10.7717/peerj.5171`
- Harnessing the power of data to support community-based research. 2018. Sallie Keller, Stephanie Shipp, Gizem Korkmaz, Emily Molfino, Joshua Goldstein, Vicki Lancaster, Bianica Pires, David Higdon, **Daniel Chen**, Aaron Schroeder. WIREs Computational Statistics.
  DOI: `https://doi.org/10.1002/wics.1426`
- Are We Ready for Mass Fatality Incidents? Preparedness of the US Mass Fatality Infrastructure. 2015. Jacqueline A. Merrill, Mark Orr), **Daniel Y. Chen**, Qi Zhi, and Robyn R. Gershon. Disaster Medicine and Public Health Preparedness.
  DOI: `https://doi.org/10.1017/dmp.2015.135`

- Mass fatality preparedness among medical examiners/coroners in the United States: a cross-sectional study. 2014. Robyn RM GershonEmail author, Mark G Orr, Qi Zhi, Jacqueline A Merrill, **Daniel Y Chen**, Halley EM Riley and Martin F Sherman. BMC Public Health 201414:1275. DOI: https://doi.org/10.1186/1471-2458-14-1275
- Acute physiological stress promotes clustering of synaptic markers and alters spine morphology in the hippocampus. 2013. V Sebastian, JB Estil, **D Chen**, LM Schrott, PA Serrano. PloS one 8 (10), e79077

Teaching

- Data Science for the Public Good
  - Co-teach a 2.5 week long intensive **workshop** on data science tools
    - *2018*: 18 Students *2017*: 19 Students; *2016*: 8 Students; *2015*: 4 Students
    - bash, SSH tunneling, git, project templates, data import/export, code repositories, R `data.table`, functions and `apply`, `knitr` and `rmarkdown` reports, grouped and aggregated statistics, loops, data tidying, regular expressions, plotting, command line scripts, detached processes, web scraping, dashboards with `shiny`, machine learning
  - Guide research of undergraduate students and graduate fellows
    - technical and code support, data pipeline algorithms
- **The Carpentries** (Formely: **Software Carpentry** and **Data Carpentry**) (2012 - Present)
  - Community maintainer lead
  - Workshop instructor
  - Instructor trainer
  - Former Co-**maintainer** of the R **Lesson**
    - 15+ Workshops taught around the country

Online Teaching Materials

- **Git Essentials**: Online video course through **Safari Books Online** on the basics of git.
- **Pandas Data Analysis with Python Fundamentals**: Online video course through Safari Books Online on the basic data transformations using the Pandas library in Python.
- **Pandas Data Cleaning and Modeling with Python**: Online video course through Safari Books Online on the cleaning and modeling data in Pandas.
- **Cleaning Data in Python**: **DataCamp** course on diagnosing and cleaning data in Python.
- **Python for R Users**: DataCamp course on transitioning to Python from R.

Open Teaching Materials

- **Data Science for the Biomedical Sciences** (2021): Online book for teaching data science to the biomedical community
- **SDAL Training Manual**: Draft of training manual used for the lab and the DSPG program
- Software Carpentry: Programming with R. 2016. DOI: 10.5281/zenodo.57541. URL: https://doi.org/10.5281/zenodo.57541
- Software Carpentry: The Unix Shell. 2016. DOI: 10.5281/zenodo.57544. URL: https://doi.org/10.5281/zenodo.57544
- Software Carpentry: Using Databases and SQL. 2016. DOI: 10.5281/zenodo.57551. URL: https://doi.org/10.5281/zenodo.57551
- Software Carpentry: Version Control with Git. 2016. DOI: 10.5281/zenodo.57467.

URL: https://doi.org/10.5281/zenodo.57467

Invited/Accepted Presentations

2020

- “Learning Tidy Evaluation by Reimplementing {dplyr}”. DCR Conference: Government and Public Sector. https://github.com/chendaniely/rstatsdc-2020-tidyeval
- “Debunking the R vs. Python Myth”. RStudio Webinar. 2020-08-26. https://github.com/chendaniely/2020-08-26-rstudio_debunk
- “Grading Code with {gradethis}”. NYR Conference. 2020-08-14. https://github.com/chendaniely/rstatsnyc_2020-learnr_gradethis
- “Grading Code with {gradethis}”. satRday DC. 2020-03-28. https://github.com/chendaniely/2020-03-28-satRd

2019

- “Using Python with R”. DCR 2019.
  - https://github.com/chendaniely/rstatsdc_2019-python-r
- “learnr+gradethis for Data Science”. DC Meetup.
  - https://github.com/chendaniely/2019-09-17-gradethis_lightning
- “Introduction to Python”. Nonclinical Biostatistics Conference.
  - https://github.com/chendaniely/ncb-2019-python
- “Building Reproducible and Replicable Projects”. NYR Conference. 2019-05-20.

2018

- “Functions: using, writing, and non-standard evaulation!!”. satRday DC. 2018-12-08.
  - https://github.com/chendaniely/satRdays-dc2018-functions
- “Python Properties” Lightning Talk. PyData DC. 2018-11-17.
  - https://github.com/chendaniely/pydatadc18_lightning
- “Structuring Your (Data Science/Analysis) Projects”. DCR. 2018-11-15.
  - https://github.com/chendaniely/rstatsdc_2018-structure
- “R and Python” Lightening talk. PyData London 2018. 2018-04-26.
  - https://github.com/chendaniely/pydata_london_2018_lightning
- “Doing Data Science”. NYR. 2018-04-22.
  - https://github.com/chendaniely/rstatsnyc_2018-data_science

2017

- “Back to Basics: R you markDown?”. New York Open Statistical Programming Meetup. 2017-04-15.
  - https://github.com/chendaniely/2017-08-15-meetup-r_you_markDown
- “So You Want to be a Data Scientist?”. NYR 2017. 2017-04-22.
  - https://github.com/chendaniely/rstatsnyc_2017-data_scientist

2016

- “Python useRs”. PyData DC 2016. 2016-10-09.
  - https://github.com/chendaniely/2016-pydata-dc-python_useRs
- “Testing”. NYR. 2016-04-08

  - https://github.com/chendaniely/2016-04-08-rstatsnyc_testing
- “Pandas for Everyone” Lightning talk. SciPy 2016.
  - https://github.com/chendaniely/2016-scipy-lightning-pandas_for_everyone

2015

- “Shiny Primer”. Statistical Programming DC Meetup. 2015-04-22.
  - https://github.com/chendaniely/2015-04-15-SPDC-shiny
- “Interactive Ebola Plots”. NYR 2015. 2015-05-02.
  - https://github.com/chendaniely/2015-04-25-rstatsnyc-ebola
- “Literate Programming” Lightning talk. SciPy 2015. 2015-07-09.
  - https://github.com/chendaniely/2015-scipy-lightning-literate
- “Python Properties in MANN”. Lightning talk. SciPy 2015. 2015-07-10.
  - https://github.com/chendaniely/2015-scipy-lightning-mann-properties
- “Open Science and Ebola”. Health Communication and Informatics Lab. 2014-11-21.
  - https://github.com/chendaniely/2014-11-21-talk-cumc-dbmi-hci

Workshops

2020

- Data Science for Biomedical Sciences workshop.
  - https://github.com/chendaniely/ds4biomed-oct20
- Business of Healthcare. Data Science for Biomedical Sciences workshop.
  - https://github.com/chendaniely/ds4biomed-20201209-business
- CarpentryCon@Home at Home Git workshop (2 sessions)
  - 2020-09-09: https://github.com/chendaniely/2020-09-09-CCatHome-git-dan
  - 2020-07-16: https://github.com/chendaniely/2020-07-16-CCatHome-git-dan
- Virginia Tech Software Carpentry Git Lesson. 2020-08.17.
  - https://github.com/chendaniely/2020-08-17-git-dan
- NYR Conference. Git for Data Science workshop.
  - https://github.com/chendaniely/2020-08-12-nyr_git-dan
- Learn Python through Data Processing in Pandas Tutorial. Scipy 2020.
  - https://github.com/chendaniely/scipy-2020-pandas
- Plumber and Docker workshop. ODSC East 2020 Conference. 2020-04-15.
  - https://github.com/chendaniely/odsc-east-2020-plumber_docker
- Forecasting andf Timeseries workshop. ODSC East 2020 Conference. 2020-04-15.
  - https://github.com/chendaniely/odsc-east-2020-forecasting_timeseries
- Introduction to R: Cleaning and Processing Data. ODSC East 2020 Conference. 2020-04-14.
  - https://github.com/chendaniely/odsc-east-2020-intro_r

2019

- Pandas Tutorial. PyBay 2019. 2019-08-15.
  - https://github.com/chendaniely/pybay_2019-pandas_tutorial
- Pandas Tutorial. SciPy 2019. 2019-07-09.
  - https://github.com/chendaniely/scipy-2019-pandas
- Git for Data Science. NYR 2019

  - https://github.com/chendaniely/2019-05-09-rstatsnyc_git
- Pandas Tutorial. PyCon 2019.
  - https://github.com/chendaniely/pycon_2019-pandas_tutorial

2018

- Cleaning and Tidying Data in Pandas. PyData DC 2018. 2018-11-16.
  - https://github.com/chendaniely/pydatadc_2018-tidy

2017

- Pandas Tutorial. SciPy 2017.
- https://github.com/chendaniely/scipy-2017-tutorial-pandas

2016

- Pandas Tutorial. PyData Carolinas 2016. 2016-09-16.
  - https://github.com/chendaniely/2016-pydata-carolinas-pandas
- Introduction and Intermediate Git. Data Intensive Biology training program.
  - https://dib-training.readthedocs.io/en/pub/
  - 2016-04-06: https://github.com/chendaniely/2016-04-06-dib-git-intro
  - 2016-04-13: https://github.com/chendaniely/2016-04-06-dib-git-intro-intermediate
- Git Workshop. Network Dynamics and Simulation Science Laboratory. 2016-02-08.
  - https://github.com/chendaniely/2016-01-22-ndssl

2014

- Python workshop. Complex System Approaches to Population Health. 2014-03-30.
  - https://github.com/chendaniely/csaph-python
- R workshop. Virginia Biocomplexity institute. 2014-10-09
  - https://github.com/chendaniely/2014-10-08-vbi

Workshop Assistant

- Introduction to R for Clinicians for the R/Medicine 2020 Virtual Conference
  - https://github.com/chendaniely/intro-to-r-for-clinicians-rmed2020
- Introduction to Machine Learning with the Tidyverse. rstudio::conf 2020. 2020-01-27
  - https://github.com/chendaniely/intro-to-ml-tidy
- Learn Bayesian Data Analysis (BDA) and Markov chain Monte Carlo (MCMC) computation using Stan. 2017-08-23.
  - https://github.com/chendaniely/2017-08-23-stan_class

Presentations (Oral and Poster) and Publiched Abstracts

- “Changing People’s Minds: Understanding Social Diffusion Dynamics Using Networked Cognitive Systems”. 2016-09-01
  - https://github.com/chendaniely/gbcb_seminar_presentation_1
- “Doing Data Science”. Biocomplexity Institute Symposium. 2017. Poster.
  - https://github.com/chendaniely/2017-bi_day_poster-doing_data_science

- "Understanding Social Diffusion Dynamics Using Networked Cognitive Systems". Society for Epidemiologic Research 2017. Poster.
  - https://github.com/chendaniely/2017_ser_poster
- "Spread of Ideas in Social Networks". New York Presbyterian Hospital: Value Institute. 2017-04-20.
  - https://github.com/chendaniely/2017-04-20-nyp_value_institute
- "Understanding Social Diffusion Dynamics Using Networked Cognitive Systems". Biocomplexity Institute Symposium. 2016-11-01.
  - https://github.com/chendaniely/2016-11-01-bi_symposium

Book Reviewer

- Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and The Cloud. Paul Deitel, Deitel & Associates, Inc. Harvey Deitel. 2020. Pearson

Technical Experience

MANN Multi-Agent Neural Network

- Continuation of Master's Thesis Project.
- Agent-based simulation model
- Measure spread of beliefs towards a behavior in a social network

SDAL Computational and data storage/management infrastructure

- Co-maintain and develop docker containers that hold all of the software and services used in the lab

Open Source

pyprojroot: Finding project directories in Python (data science) projects Python implementation of the R {`rprojroot`} and {`here`} packages

Mesa: Agent-based modeling in Python. Contributed to the documentation and provided an example model based on the conference proceeding paper.

EnTICE3: Communication tool to generate tailored infographics from electronic health data to improve patient literacy. It is based on a **style guide** and has a supporting **grader** to help analyze the data for the tool's clinical trial assessment

Zika Open Data Dashboard: Dashboard to quickly glance at all the data curated by the CDC during the Zika outbreak in 2016. The dashboard makes it easy to assemble together all the curated datasets to view and download. Extends on a **similar project** during the 2014 West Africa Ebola outbreak.

pylens: Python wrapper for the LENS neural network simulator used in the MANN project

sdalr: R package of commonly used functions for all members of the lab

- Project Templates
  - **Computational Project Cookie Cutter**: Template for analytics projects
  - **Pweave document**: LaTeXtemplate for literate programming in Python
  - **Knitr document**: `knitr` template for literate programming in R
  - **SDAL Paper Template** LaTeXtemplate for writing papers

Programming Languages

R: Primary language for data cleaning, statistical modeling, and machine learning

Python: Computer simulations, web scraping, and other general data collection programs. Primarily use `pandas` for data related work.

Basic knowledge of **SQL**, **C++**, **Java**, **HTML/CSS**

System administration: **Bash**, **Docker**, **VirtualBox**