Matthew J. Collins
1658 NW 22nd Circle
Gainesville, FL 32605
352-275-4027
mjcollinsresume@gmail.com

Education

M.S. Industrial and Systems Engineering — May 2016
University of Florida, Gainesville, FL

B.S. Civil Engineering with Environmental Option — May 1997
University of Vermont, Burlington, VT

Professional Experience

Data Engineer — January 2019 – Current
Constant Contact via acquisition of SharpSpring, Gainesville, FL

Data engineer for leading online marketing automation company building cloud-based data systems.

Key Activities

- Design and maintain user-facing data aggregation pipelines that back the SharpSpring app's custom reporting feature based on MariaDB and Elasticsearch
- Author Python data pipelines to populate a Google BigQuery data lake with all SharpSpring application and business data from Vitess, MariaDB, MongoDB, and 3rd party APIs
- Maintain Dockerized Airflow deployment running in a Google Cloud Platform Kubernetes Engine (GKE) cluster
- Standardize and approve all application database schema changes to align software developers' data designs with data analysts' reporting needs and best practices
- Mentor business analysts in software development practices and manage a business intelligence engineer

Technical Operations Manager — May 2010 – January 2019
Advanced Computing and Information Systems Laboratory, University of Florida, Gainesville, FL

Infrastructure software developer for computer engineering academic research laboratory specializing in big data, cloud, and virtualization systems.

Key Activities

- Design, develop, and maintain open-source infrastructure for the NSF-sponsored Integrated Digitized Biocollections (iDigBio) project. Used PostgreSQL, Elasticsearch, and Ceph to aggregate 120 million records of natural history museum specimens with 30 million images consuming 160 TB of storage (https://www.idigbio.org)
- Provide training and outreach to biologists in software development, data APIs, and reproducible research practices to increase use of large aggregated data in research

- Collaboratively design, deploy, and use Spark-based data analytics infrastructure with 192 threads and 576 GB of memory for mining and analyzing biodiversity data sets including formatting common data sources into parallelized data frames (http://guoda.bio)
- Maintain the facilities and configurations of over 260 servers running cloud, distributed, and virtualization environments such as Elasticsearch, VMware, XenServer, Spark, Cloudera, and Ceph
- Create collaborations across research institutions including the Encyclopedia of Life, Duke University, Royal Botanic Gardens, Kew, San Diego Super Computing Center, and NARA Institute of Science and Technology to design, present, and publish novel data processing pipelines and systems configurations
- Author publications, presentations, reports, blogs, and other technical communications to disseminate the research of the lab to a global audience including organizing symposia, hackathons, and workshops in multiple countries
- Mentor computer engineering graduate students in the technical aspects of their research and supervise master's projects

System Administrator — January 2008 – May 2010

Florida Museum of Natural History, University of Florida, Gainesville, FL

Lead system administrator for research/business environment with 300 staff and 40 Linux and Windows servers.

Key Activities

- Create, prioritize, and complete projects like virtualization and SAN storage that add new capabilities to museum staff and provide support for grant projects
- Design and maintain software development environments for team of 6 programmers and web site managers
- Supervise and mentor assistant system administrator and interns

System Administrator — April 2007 – January 2008

Bureau of Economic and Business Research, University of Florida, Gainesville, FL

System and workstation administrator for business environment with 250 staff, Linux and Windows servers, and 120 Windows workstations.

Key Activities

- Modernize aging infrastructure with virtualization for servers and Ghost workstation deployment

Lead Programmer — September 2005 – April 2007

Florida Museum of Natural History, University of Florida, Gainesville, FL

Lead developer on NSF funded BioCorder grant project. Quarter-time system administrator.

Key Activities

- Author web application for storing and sharing genetic laboratory data in PHP/PostgreSQL
- Coordinate work from programmers and principal investigators from other institutions
- Lead special projects aimed at modernizing IT infrastructure like automated build systems

Consultant System Administrator — September 2004 – December 2011

Bear Code LLC, Montpelier, FL

Part-time consulting system administrator for international software development company.

Key Activities

- Build and maintain web and database services in Amazon cloud environments for production web applications
- Manage and migrate services across on-premise, traditional VPS, and Amazon EC2 platforms

System Administrator, Programmer — June 2001 – August 2005

Signal Advertising, Inc., Montpelier, VT

System administrator and software developer for Internet development business.

Key Activities

- Design Linux-based infrastructure to provide ISP services at 99.99% availability (web hosting, email, database, custom applications, etc)
- Manage infrastructure and relations with multiple Tier IV data centers
- Author bespoke web applications in PHP/MySQL including e-commerce and data aggregation
- Provide development framework and core code to 3 person development team

Database Specialist — May 1997 – June 2001

Stone Environmental, Inc., Montpelier, VT

Use desktop software and scripts to manage and present environmental data for reporting.

Key Activities

- Collect, normalize, and present weather, soil, and hydrologic data in EPA reports for pesticide regulation
- Develop software to perform hydrologic modeling and spatial analysis with ArcView GIS
- Author and support application for managing municipal septic systems

Honors and Awards

EAGER: Towards the Web of Biodiversity Knowledge: Understanding Data Connectedness to Improve Identifier Practices, NSF Award 1839201. $299,973. 2018-2020

Divisional winner University of Florida Superior Accomplishment Award. 2013

Publications

Journals

Matsunaga, A, Thompson, A, Figueiredo, R, Germain-Aubrey, CC, **Collins, M**, Beaman, RS, MacFadden, BJ, Riccardi, G, Soltis, PS, Page, LM, Fortes, JAB, “A Computational- and Storage-Cloud for Integration of Biodiversity Collections”, in 2013 IEEE 9th International Conference on e-Science, Beijing, China, 2013, p. 78-87.

Ichikawa, K, U-Chupala, P, Huang, C, Nakasan, C, Liu, T-L, Chang, J-Y, Ku, L-C, Tsai, W-F, Haga, J, Yamanaka, H, Kawai, E, Kido, Y, Date, S, Shimojo, S, Papadopoulos, P, Tsugawa, M, **Collins, M**, Jeong, K, Figueiredo, RJ, Fortes, JAB, “PRAGMA-ENT: An International SDN

testbed for cyberinfrastructure in the Pacific Rim", Concurrency and Computation: Practice and Experience, vol. 29, no. 13, 2017.

James, S. A., P. S. Soltis, L. Belbin, A. D. Chapman, G. Nelson, D. L. Paul, and **M. Collins**. 2018. Herbarium data: Global biodiversity and societal botanical needs for novel research. Applications in Plant Sciences 6(2): e1024.

Thessen AE, Poelen JH, **Collins M**, Hammock J. (2018) 20 GB in 10 minutes: a case for linking major biodiversity databases using an open socio-technical infrastructure and a pragmatic, cross-institutional collaboration. PeerJ Computer Science 4:e164https://doi.org/10.7717/peerj-cs.164

Proceedings

Collins M, Nicolson N, Poelen J, Thompson A, Hammock J, Thessen A (2017) "Building Your Own Big Data Analysis Infrastructure for Biodiversity Science". Proceedings of TDWG 1: e20161. https://doi.org/10.3897/tdwgproceedings.1.20161

Collins M, Tarvin R, Kandziora M, Dahdul W, Paul D (2018) Phenomap - Challenges and Successes in Bringing Together Multiple Data Projects to Build New Visualizations of Phenotypic Information and Specimen Records. Biodiversity Information Science and Standards 2: e25698. https://doi.org/10.3897/biss.2.25698

Collins M, Yeole G, Frandsen P, Dikow R, Orli S, Figueiredo R (2018) A Pipeline for Deep Learning with Specimen Images in iDigBio - Applying and Generalizing an Examination of Mercury Use in Preparing Herbarium Specimens. Biodiversity Information Science and Standards 2: e25699. https://doi.org/10.3897/biss.2.25699

Posters

Collins M, Poelen J, Thompson A. "Whole-dataset Analyses Using Apache Spark". Poster presented at: Biodiversity Information Standards Conference (TDWG). 2015 Sep 28-Oct 1; Nairobi, Kenya.

Collins M, Poelen J, Thompson A. "Updates on Whole-Dataset Analyses Using Spark and the GUODA Data Service". Poster presented at: 31st Annual Meeting 2016 of the Society for the Preservation of Natural History Collections (SPNHC). 2016 Jun 20-25; Berlin, Germany.

Yeole G, Sahdev S, **Collins M**, ThompsonA, Dikow R, Frandsen P, Orli S, Figueiredo R. "A Pipeline for Processing Specimen Images in iDigBio - Applying and Generalizing an Examination of Mercury Use in Preparing Herbarium Specimens using Neural Networks". Poster presented at: Biodiversity Information Standards Conference (TDWG). 2017 Oct 2-6; Ottawa, Canada.

Presentations

Paul D, Teal T, **Collins M**. Data Carpentry: "One Comprehensive Model for Teaching Biodiversity Informatics Data Management, Research Skills, and Training Instructors", in Biodiversity Information Standards Conference (TDWG), Nairobi, Kenya, 2015.

Collins M, Paul D. "Accessing Digital Collections Data Sources for Research: A Tour of iDigBio Data Services", in Island Biology 2016, II International Conference on Island Evolution, Ecology, and Conservation, University of the Azores at Angra do Heroísmo, Terceira Island, Azores, Portugal, 2016.

Collins M, "GUODA: A Unified Platform for Large-Scale Computational Research on Open-Access Biodiversity Data", in Biodiversity Information Standards (TDWG) 2016 Annual Conference, Santa Clara de San Carlos, Costa Rica, 2016.

Collins M, “Mining Whole Museum Collections Datasets for Expanding Understanding of Collections with the GUODA Service”, in 31st Annual Meeting 2016 of the Society for the Preservation of Natural History Collections (SPNHC), Berlin, Germany, 2016.

Collins M, “Demo: Text Mining Whole Museum Datasets for Expanding Understanding of Collections with the GUODA Service”, in 31st Annual Meeting 2016 of the Society for the Preservation of Natural History Collections (SPNHC), Berlin, Germany, 2016.

Collins M, Poelen J, Thessen A, Thompson A, Hammock J. ”jupyter.guoda.bio: Hosted Jupyter Notebooks with Biodiversity Datasets for Reproducible Research in R and Python”, in 32nd Annual Meeting 2017 of the Society for the Preservation of Natural History Collections (SPNHC), Denver, Colorado, 2017

Collins M, Nicolson N, Poelen J, Thompson A, Hammock J, Thessen A. “Building Your Own Big Data Analysis Infrastructure for Biodiversity Science”, in Biodiversity Information Standards Conference (TDWG), Ottawa, Canada, 2017.

Conferences and Workshops

Attendee, O’Reilly Open Source Convention, 2006

Attendee, Large Installation System Administration, 2008

Attendee, Southeast Linux Fest, 2011

Booth Attendee, Super Computing, 2011

Booth Attendee, Super Computing, 2012

Attendee, Collaborative IT Workshop, 2014

Helper, Data Carpentry Workshop National Evolutionary Synthesis Center, 2014

Instructor, Data Carpentry Workshop Integrated Digitized Biocollections, 2014

Attendee, Advancing Digitization of Biodiversity Collections Summit, 2014

Attendee, Quantitative Biology Education Summit, 2015

Organizer, Participant, Hackathon on iDigBio APIs/Services and Interoperability, 2015

Instructor, Software Carpentry Workshop University of Miami, 2015

Instructor, Software Carpentry Workshop University of South Florida, 2015

Instructor, Data Carpentry Workshop Multimedia University of Kenya, 2015

Presenter, Biodiversity Information Standards Conference (TDWG), 2015

Invited Speaker, International Workshop on Building Collaboration in Biodiversity Informatics, 2015

Attendee, Pacific Rim Applications and Grid Middleware Assembly PRAGMA29, 2015

Helper, Carpentry Instructor Training University of Florida, 2016

Attendee, Pacific Rim Applications and Grid Middleware Assembly PRAGMA30, 2016

Organizer, Instructor, Software Carpentry Workshop University of Florida, 2016

Presenter, The Society for the Preservation of Natural History Collections Annual Meeting, 2016

Symposium Chair, Presenter, “iDigBio Workshop”, International Conference on Island Evolution, Ecology, and Conservation, 2016

Organizer, Instructor, Software Carpentry Workshop University of Florida, 2016

Organizer, Data Carpentry Workshop University of Florida, 2016

Attendee, Pacific Rim Applications and Grid Middleware Assembly PRAGMA31, 2016

Attendee, Advancing Digitization of Biodiversity Collections Summit, 2016

Organizer, Instructor, Software Carpentry Workshop Centro de Transferencia Tecnológica y Educación Continua, 2016

Symposium Chair, Presenter, "'Big Data Analysis Methods and Techniques as Applied to Biocollections", Biodiversity Information Standards Conference (TDWG) 2016

Organizer, Instructor, Software Carpentry Workshop University of Florida, 2017

Organizing Committee, Smart Cities Student Hackathon, 2017

Host Committee, Pacific Rim Applications and Grid Middleware Assembly PRAGMA32, 2017

Helper, Carpentry Instructor Training University of Florida, 2017

Organizer, Instructor, Software Carpentry Workshop University of Florida, 2017

Helper, Data Carpentry Workshop University of Florida, 2017

Presenter, The Society for the Preservation of Natural History Collections Annual Meeting, 2017

Organizer, Instructor, Software Carpentry Workshop University of Florida, 2017

Organizer, Data Carpentry Workshop University of Florida, 2017

Organizer, Instructor, Data Carpentry Workshop Canadian Museum for Nature, 2017

Symposium Chair, Presenter, "Using Big Data Techniques to Cross Dataset Boundaries - Integration and Analysis of Multiple Datasets", Biodiversity Information Standards Conference (TDWG), 2017

Organizing Committee, Computable Evolutionary Phenotype Knowledge: a Hands-on Workshop, 2017

Presenter, Pacific Rim Applications and Grid Middleware Assembly PRAGMA34, 2018

Organizer, Instructor, Data Carpentry Workshop University of Florida, 2018

Programme Committee, biodiversity_next: 2019 Open Digital Science Week on Biological & Geological Diversity, 2019

Service

Advisor, UF Data Science and Informatics student organization (http://www.dsiufl.org/)

Founder, UF Carpentries Club (https://www.uf-carpentries.org/)

Software Carpentry instructor mentor, round 1 (Africa), round 2 (US)

Certifications

Software and Data Carpentry Instructor

Additional Qualifications

Experienced with: Amazon EC2, AWS, Google Cloud Platform, Docker, VMware vSphere, ESX, Citrix XenServer, MySQL, MariaDB, PostgreSQL, Google BigQuery, Elasticsearch, Riak, MongoDB, Ceph, Spark, Hadoop, HDFS, Cloudera, Salt configuration management, Puppet, Foreman, Python, R, Jupyter Notebook, Bash, Java, PHP, MATLAB, Git, Github, Gitlab, Subversion, Jenkins, Travis CI, Looker, Metabase, Apache, Lighttpd, Bind, HaProxy, Postfix

SMTP server, Zimbra, Redmine, Drupal, Zabbix monitoring system, Markdown, NFS, ZFS, RedHat Enterprise Linux, CentOS, Ubuntu, OpenSolaris, Windows Server, TCP/IP networks, firewalls and NAT, proxy servers, load balancers, software defined networking, DNS, IPMI, PXE, SNMP, Active Directory, Windows Terminal Server, IBM fibre channel SAN, IBM blade centers, Dell blade centers, iDataPlex, Cisco, Dell PowerConnect, Brocade, Pica8, Force10 switches.

Github account at https://github.com/mjcollin