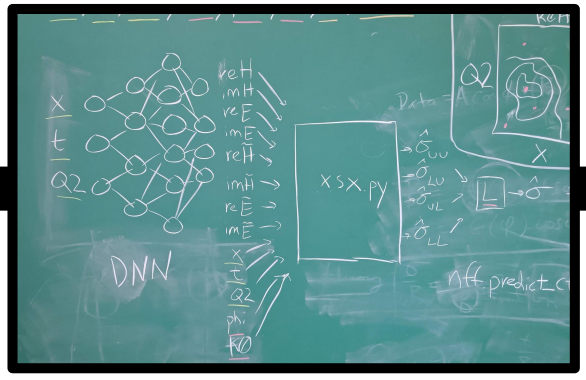
Open Data Lab Annual Report: FY19-20

The Open Data Lab Collaboration

September 23, 2019

OPEN DATA LAB



Open Data Lab Annual Report 2018-2019

The Team

Pete Alonzi came to data science by way of the particle physics community. As a result, he has great interest in making data open and usable to broad audiences. He serves as a data scientist at the DSI and is the project manager for the Open Data Lab.

Phil Bourne, Stephenson Chair of Data Science and Director of the Data Science Institute. Before coming to serve at UVA Phil had been working for three years as the associate director for data science at the National Institutes of Health.

Tim Clark, Ph.D., is a biomedical informatician and computer scientist with 28 years experience in academia, government, and industry. He is an expert in data fusion and open science applications. He serves as an Associate Professor in the UVA School of Medicine & the DSI.

Max Levinson is a Software Engineer and Cloud Developer in Public Health Services. When not building out REST apis for the NIH Data Commons Project, he can be found hacking on the Open Data Lab Project. Max is passionate about microservices development and anything Python.

Daniel Mietchen is a data scientist at the DSI. Trained as a biophysicist, he is interested in how research can be performed openly and collaboratively across traditional boundaries like disciplines, jurisdictions, economic sectors, languages, or professional versus volunteer participants.

Lane Rasberry is a Wikimedia-in-residence. His focus is using Wikimedia projects as a channel for publishing and distributing media from the DSI and UVA to an audience that engages with Wikimedia projects as an information resource and communication forum.

Project manager's note

At the beginning of 2018 the Open Data Lab was an idea, a sketch on a whiteboard. Today the Open Data Lab has over 100 users. Students at the University of Virginia have used it to conduct research published by IEEE. In just over a year a small team of dedicated individuals from the Data Science Institute at UVA produced a product that furthered the educational and research mission of the University. Today students at UVA can do things they could not do before this project launched.

I am extremely honored to have been asked to lead this effort. The work done by our collaboration is of the very best calibre and I am proud of the work we have done and will do in the future. As project manager I feel I have the easiest job on the team. I ask everyone for ideas and they put forth not just an enormous number of ideas but also brilliant ideas. The hardest part for me is to choose from all the great ideas. Without the team Phil Bourne has built my job would be impossible. With this team I do not see any limit to what we can achieve.

The first year of the Open Data Lab was one of exploration. We discovered new tools, in particular the value of Project Jupyter. We also interviewed our potential users and recognized archetypes which we can use to guide what we build. Putting all of this together we are currently developing a strategic plan for the next three years. We now look to make an impact in the community and I am excited for what is to come.

Open Data Lab Project Manager

Peter Alonzi

Executive Briefing

The 2018-2019 year was the first year for the Open Data Lab. In this time we learned many things and many users attended workshops, hosted and anlyzed data, some event produced a publication. We served 116 users, working on 25 projects, and 6 individuals contributed directly to the ODL project. Going forward we look to increase all of those numbers including users from beyond the Data Science Institute.

Here are a few highlights from 2018-2019:

- Development of a GitHub workflow for beginners useful beyond the hard sciences. This workflow was adopted by an international collaboration called the Open Greek and Latin Project [3] as well as the Archaeology Department of Monticello [4]. Details are found on our github page and in Section 4.2.

- We realized the power of Project Jupyter as a system to deliver resources to the user without excessive cognitive load. This platform is transformative and will lead to great things in the future. Details are found in Section 2.3.1.

- We studied three user archetypes for the Open Data Lab: the collaborator, the student, and the sharer. Details are round in Section 1.2.

We have accomplished a lot but now we need help. If you have an idea or want to join the team please reach out by emailing datascientist@virginia.edu (Subject Line: "I want to help the ODL").

Bug Bounty Program

We encourage close reading and critique of the Open Data Lab proceedings. If you find a bug in any of our work including this report please let us know. The best way to pass along notes is through a pull request on our github page: `https://github.com/UVA-DSI/Open-Data-Lab`.

Contents

List of Figures

List of Tables

Chapter 1

Overview

1.1 What is the Open Data Lab?

OPEN

We encourage all users to be as open as possible with every aspect of their work. That may be in opening up their data sets, publication, source code, or something else. The Data Science Institute working definition of Open:

> Openness means team members responsibly sharing their data and professional endeavors (when possible and ethical). We believe in the importance of practicing openness because advancement requires assembling a heap of known pieces into a coherent picture containing new knowledge. In the world today some of the necessary pieces are unknown due to traditional non-open information practices. Wide spread open practices are the first steps to changing the world.

This definition was developed by the DSI Open Working Group. Their phase one summary is listed in Appendix A.

DATA

We take an expansive definition of data. Everything from traditional data, to code, to workflows, to published material, and so on is considered data to us. We provide a place for all things digital data.

LAB

We provide a place where the power of computing can be brought to bear against data resources. Given the scale of data today this means colocating the data and computational resources.

Open Data Lab

The Open Data Lab is a resource to provide state of the art computing and data infrastructure to researchers, students, and sharers. It is guided by the principles of science and openness.

1.2 User Archetypes

There are many potential use cases for the Open Data Lab. In this section we describe the three cases that have been studied so far. They are: the Collaborator, someone who is working on a research project; the Student, someone who is using the Open Data Lab to learn; and the Sharer, a person with data who wants to open it up to a broader audience.

The Collaborator

This archetypal person uses the Open Data Lab to conduct research. They access data and computational resources that are colocated. This colocation facilitates lower latency and increased performance. A wide range of services can be provided globally by AWS and locally through UVA HPC resources. Sample workflow:

1. Request a user account on the Open Data Lab
2. Once per collaboration:
   (a) Load data
   (b) Provision computational resources
3. Conduct research operations
4. Register resulting products in Dataverse

The Student

This archetype uses the Open Data Lab to facilitate learning. An example would be someone who participates in a workshop where and ODL notebook instance powered by AWS SageMaker provides the working environment. Sample workflow:

1. Request a user account on the Open Data Lab
2. Logon to AWS console to launch Jupyter
3. Use Jupyter during the workshop

The Sharer

This archetype is a user who owns data and wants to make it available. There are many mechanisms for sharing the data ranging from RESTful API of S3, to a SPARQL endpoint. Sample workflow:

1. Request a user account on the Open Data Lab
2. Load data into an S3 bucket
3. Configure one of the following
   (a) SPARQL endpoint
   (b) API Gateway to access S3
   (c) S3 permissions for a SageMaker notebook
   (d) etc.

group	projectID	# members	type
Bourne-Mura	bamc	4	MSDS Capstone
CBW	cbwc	4	MSDS Capstone
Wiki	wiki	10	MSDS Capstone
Mental Health	miip	6	SYS Capstone
Women Terror Recruitment	watr	2	Presidential Fellow
Healthy Markets	hmtt	5	DSI Research
Independent Study	pmis	1	DSI Research
Grommullyang	gmmy	3	DSI Research
Linked Open Data	nept	2	External Data
Spark	sprk	17	Education
GitHub	gith	9	Education
Practice of DS	pods	60	Education
ORCI	orci	2	ODL Development
ML under	mlunder	7	Club
ML grad	mlgrad	3	Club
Rivanna	–	11	Local
Ivy	–	6	Local
ODL-education	–	26	Education Users
ODL-users	–	116	Unique Users

Table 1.1: Summary of ODL Users

1.3 User Summary

1.4 Phased project strategy

The first three phases of the Open Data Lab have been outlined. Phase 0 focused on pre investigation and decided on what technology to test in Phase 1, the closed beta. Phase 2 is an open Beta and will serve the community of Charlottesville and other associated research and educational efforts.

To realize phase 2 we need a person to take ownership of each element of the Open Data Lab that moves into phase 2. We also require a personnel roster capable of providing service at the level necessary for the users. That requires hiring and as a results the timeline is TBD.

Phase	type	start	end
0	alpha	FEB 2018	JUN 2018
1	closed beta	JUL 2018	TBD
2	open beta	TBD	–

Table 1.2: ODL phase schedule

Put another way, the current bus factor[1] for the open data lab is 1.

1.5 What's next for the Open Data Lab?

Starting at the end of the 2018-2019 academic cycle we have restructured the Open Data Lab Effort. We are now moving to a targeted impact strategy. What we call phase 2 (open beta) will focus on particular areas of impact. At this time the collaboration is undergoing a strategic planning effort. This effort will produce a plan to guide the Open Data Lab development over the next three years. It aligns with and inherits from the DSI plan as well as the UVA strategic plan put forth by President Ryan.

[1] `https://en.wikipedia.org/wiki/Bus_factor`

Chapter 2

Key Developments

This year saw great progress for the Open Data Lab. Early on there were weekly meetings to define goals and objectives. Those meetings lead to implementation of several areas and this chapter presents the key developments during 2018.

2.1 Phase 1 Closed Beta

The first decision made for the Open Data Lab was to implement a staged approach. Section 1.4 summarizes the whole scope. This section discusses the state of the closed beta (Phase 1). The user base for this phase is predominantly the Data Science Institute. There were 116 participants ranging from students, to faculty, to staff. There were also 26 workshop attendees ranging from a diverse selection of UVA researchers to community members. The primary goal of this phase is to test different technology solutions to anticipated needs (see section 2.3). Those range from data storage to computation to discovery to pedagogy, and so on. Of particular note was the wild success of implementing Project Jupyter. The tools developed by this project served many roles (see section 2.3.1). This phase will run until several criteria are met. The first is the establishment of a new funding model that will cover the scope of the open beta test. The second is the acquisition of new staff. Currently the Open Data Lab has a bus factor of one and phase 2 requires a higher value.

2.2 Establishment of User Base

This year saw the birth of the Open Data Lab and growth to include 116 users. Those users take various form from capstone research programs at the graduate and undergraduate level to full fleged dissertation research. Some of the users involved are dealing with datasets that now reside within the Open Data Lab. Currently those datasets and under tight restriction as we explore proper security protocols. There is also a contingent of undergrad and graduate students that are part of data science clubs at UVA who gain access to resources through the Open Data Lab.

It is important to understand that the technology behind the system is not the driver of the system. The needs of the user are the driver. Right now the closed beta format allows us to interview new users and tailor a program to them. Sometimes we get the resource wrong and adjustments have to be made. Regarding data storage the use of S3 storage from Amazon Web Services has served a broad selection of users well. Recent developments in AWS object storage technology enable users to use it as if it were block storage. As a result S3 has proved an effective solution both for large scale data storage as well as database query repositories. Providing computational resources has been guided by the user base as well. As the base grew it became clear that the notebook technology developed by Project Jupyter was highly effective and actually resulted in more people volunteering for the closed beta test. The use of that technology helped bring users into the system.

2.3 Technology Exploration

Many different technologies were tested in 2018. Many options using Amazon Web Services were explored and those services were found to be excellent. Local UVA resources were also used and in particular the UVA HPC portal developed out of the VP-IT's office is phenomenal. Collaboration is also underway with the UVA library regarding the discovery component of the Open Data Lab and the implementation of Harvard's Dataverse, known locally at UVA as Libra. For version control and sharing purposes GitHub was evaluated.

2.3.1 Amazon Web Services

Cloud computing provides agility that local computational resources do not. To that end we established a contract with Amazon Web Services (AWS) through the third-party vendor DLT solutions. This contract is collectively negotiated and takes advantage of the Internet-2 network [5].

In selecting a cloud resource provider our choice was informed by the needs of the MSDS students at the DSI. The most requested cloud service was AWS and the plurality of job postings that are interested in cloud skills prefer AWS. During 2018 the initial scope of the AWS exploration was for functionality on colocating data and computation. However the initial implementation was more sucessful than anticipated. The system was popular and as a result Pete Alonzi functionally become the sysadmin for the DSI AWS account. What started as a development project turned into an operations project as soon as a minimum viable product was established. This is great for the DSI because there is a now a new resource for the researchers. However it did pull substantial resources from the development of the Open Data Lab.

In the following sections we will breakdown the different services used by the ODL.

Service	Function	Notes
S3	Object Storage	$30/TB/yr
EC2	Compute Instances	pricing
SageMaker	Project Jupyter	popular interface
IAM	Identity/Access Manager	users and groups
API Gateway	Credentialed REST	allows trigger of lambda
Lambda	Serverless Functions	miscellaneous tasks
CloudWatch	Log/Monitor/Alarm	system monitor

Table 2.1: Summary of Amazon Web Services evaluated in 2018

S3

This service provides the cheapest usable storage solution at scale. Furthermore recent policy developments at Amazon now require that other services interact with S3 (object storage) with comparable performace to block storage. That policy aligns nicely with the ODL needs. It means that cheaper

Object Storage can be used for larger and larger datasets without yielding performance. Additionally this means only one storage solution needs to be implemented. We do not need to provision additional block storage for execution operations requiring data migration.

The S3 systems divides the data into buckets. For this test we treated a bucket as the unit holding data for a particular research project. As a result we have 22 buckets provisioned to accomodate all of our efforts. All buckets on AWS are localized to a region but must have a globally unique name. To that end for phase 1 we have adopted the following convention. Each bucket id begins with 'odl-' and is followed by the four character project id (eg: odl-hmtt).

ID	Project
odl-bamc	Bourne/Mura Capstone
odl-bamc-scratch	Bourne/Mura Capstone
odl-dome	Dominion Capstone
odl-hmtt	Healthy Markets
odl-hmtt-scratch	Healthy Markets
odl-nept	Numismatic Linked Open Data
odl-orci	Educational Open Datasets
odl-podc	DSI Communications
odl-projhects-test	DSI project configurations
odl-readonly-test	read only test
odl-scratch-test	scratch space test
odl-sp19-sys6016	Class materials
odl-spark-education	Spark Educational Materials
odl-spark19spds6003-001	Class Materials
odl-watr	Women and Terrorism Recruitment
odl-watr-scratch	Women and Terrorism Recruitment
odl-wiki	Wiki Capstone
2017-2018-capstone-plos	17/18 capstone
uva-bucket	initial test bucket

Table 2.2: Summary of S3 Buckets Provisioned for ODL 2018

For the example of 'odl-hmtt' this bucket serves the Healthy Markets project. We bill that project PTAO for the service the ODL provides. The

mechanism is to pay the DLT contract off the ODL PTAO and then do a cost transfer annually from the Healthy Markets PTAO to the ODL PTAO.

13 TB Data Transfer

When we acquired the Healthy Markets dataset we transferred it from their aws bucket to ours. However this transfer was not trivial. AWS was unwilling to change the bucket ownership from their account to ours necessitating a data transfer.

The recommendation was to use the AWS CLI to perform the copy. It functions very similarly to scp from standtard unix systems. However we discovered several interesting features.

- Some of the files were copied to our bucket using an IAM account on their AWS account. Then Healthy Markets deleted that IAM user and as a result we lost control of the files in our bucket. We had to ask our partner to reestablish the account and then update the ownership. To do that we copied the files onto themselves but using our account so they would be owned by us. You cannot change the owner of a file on S3.

- Direct bucket to bucket transfer is managed via a serverless operation and does not prototype the transfer. As a result the resources allocated automatically were not sufficient to transfer 13 TB during a human lifetime. We then setup a dedicated server via EC2 and were able to configure the system to complete the transfer at a rate of about 4 TB per day. We did incur cost for operating that server but is was not prohibitive.

- The AWS CLI is not optimized for mass transfer and to copy the data in a reasonable timescale we had to write some scripts to chunk the operation. That work is located on the ODL github page.

EC2

This service allows for provisioning of compute resources. We established lambda functions to automate the create of EC2 instances for projects given a json configurationf file. Here is a sample JSON file:

```
{
   "projectId": "open-data-lab",
   "github": "https://github.com/UVA-DSI/Open-Data-Lab.git",
   "data-bucket": "odl-hmtt",
   "scratch-bucket": "odl-scratch-test",
   "ImageId": "ami-b70554c8",
   "InstanceType": "t2.nano",
   "email": "lpa2a@virginia.edu",
   "maxNumInstances": "1"
}
```

These EC2 units are monitored by the CloudWatch system and we can configure them to turn on and off as necessary. The majority of our users use EC2 as their compute engine however they access the compute via auto-provisioning from the SageMaker service.

SageMaker (Project Jupyter)

SageMaker has been the breakout star of the 2018 ODL development phase. Our MSDS students prefer Project Jupyter as their mechanism for using computational and storage resources. To that end Peter Alonzi went to the JupyterCon in New York City in the fall of 2018[1]. At this conference Peter Alonzi was able to speak directly with the creators, developers, and users of Project Jupyter as well as the AWS developers of SageMaker. This service through the power of Project Jupyter is able to democratize the access and management of computational resources. The systems lifts a large cognative load off the user and empowers them to accomplish their goals efficiently without requiring arcane knowledge of computers. Our working metaphor is as follows:

> **The user is able to drive the car and get where they want to go without needing to first learn how to build a transmission.**

This power is the true killer feature of Project Jupyter and is the reason the system is so popular and widely used. Often people we say the ability to break code into cells and use markdown and inline plotting is the killer feature but it is really the lifting of the cognitive load.

[1] `https://github.com/UVA-DSI/conferences/tree/master/JupyterCon18`

IAM

This services governs Identity and Access Management. It breaks down into users and groups. Users are placed into appropriate logical groups and then access policies are assigned to groups. When testing permission settings a dummy IAM user is established and given the same groups as the user being tested. In that way the sysadmin can see what the user sees and debug/test/prepare the system to the user.

API Gateway

We configured an API gateway to provide a mechanism for the users to trigger lambda functions. The users require credentialing via the standard aws authentication protocol when the post to the Gateway. Details of the system are on GitHub[2].

lambda

This services allows for serverless execution of code. Currently we use it for provisioning of EC2 instances and automation of EC2 management.

CloudWatch

This service is how we monitor the system and record logs.

[2] `https://github.com/UVA-DSI/Open-Data-Lab/blob/master/aws/api-gateway/hitapi.py`

Architecture Diagrams

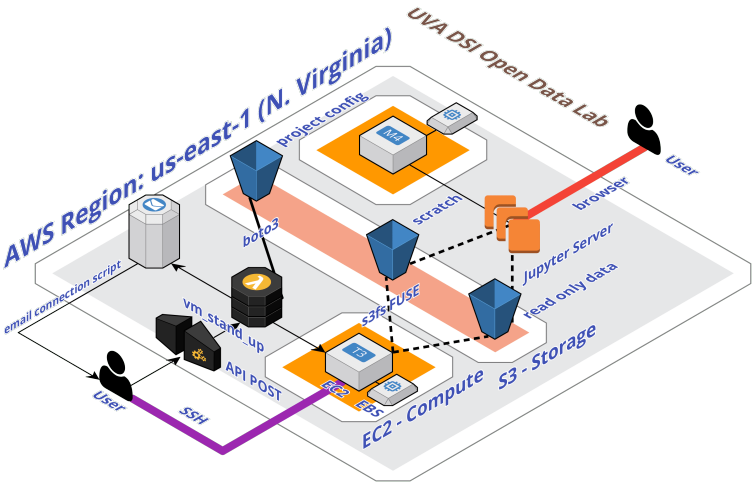


Figure 2.1: Schematic of AWS service configuration

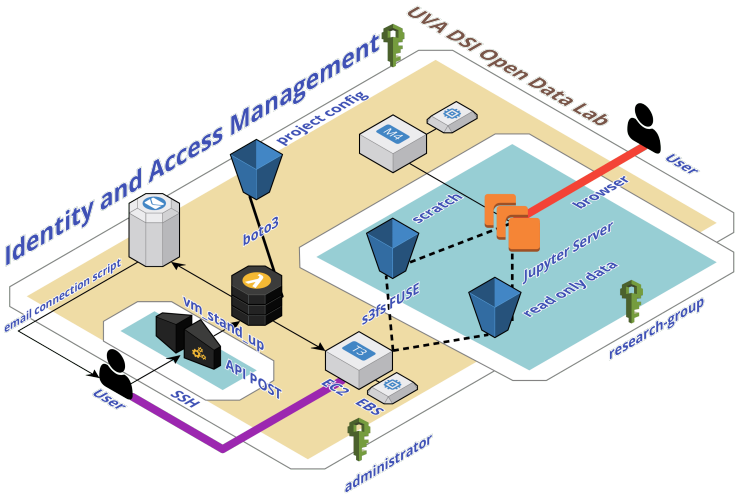


Figure 2.2: Schematic of AWS IAM configuration

Support Plan

For the first year of the ODL we elected to keep the business support plan. In the future we can have a ten percent savings by eliminating this support.

`https://aws.amazon.com/premiumsupport/compare-plans/`

2.3.2 Local UVA - Rivanna and Ivy

The local computational resources at UVA are facilitated through the office of Vice President for IT. That group is dedicated and hard working and provides great resources to the local UVA community. We have established a working relationship with them and discuss technical problems and solutions. Independently we arrived at the utility of Project Jupyter. These solutions are for UVA personnel and their collaborators and as such will not scale to later phases of the Open Data Lab project. However for the closed and open beta it is a great resource. Furthermore their technical expertise will be invaluable to the Open Data Lab regardless of phase.

2.3.3 GitHub

GitHub is the most broadly adopted cloud platform for version control. Therefore we evaluated it first. The utility for managing repositories is fully mature. The collaborative features focused around the fork and pull request paradigm are excellent. GitHub also has project level capability with issue tracking and team/permission functionality for managing permissions and progress. We have been extremely pleased with the capabilities of GitHub. The only motivation to try other solutions is for the sake of due diligence.

Concerning the acquisition by Microsoft: Many have raised the issue that GitHub may not be the appropriate solution now that Microsoft has acquired GitHub. However the recent track record of Microsoft is to not meddle with projects like GitHub but rather to protect them. Additionally most users use other Microsoft products.

We developed a workflow for beginning users on the GitHub platform. Details are given in 4.2.

2.4 Upcoming Technical Exploration

The following sections describe exploratory work that is on the schedule. There is more to be done beyond this list but not scheduled.

2.4.1 Dataverse

A framework has been outlined to use Dataverse as the discovery mechanism for the Open Data Lab. In this system a metadata entry will be made in the Dataverse containing all of the usual materials. However the final piece with the datafiles will contain pointers to the data and projects within the Open Data Lab. Dataverse is not configured for colocating computation resources with the data resources. The pilot of this test will be with the Libra project from the UVA Library. Currently that system is undergoing an upgrade and once there is a stable release exploration will commence.

2.4.2 Spark

The first scale data solution the Open Data Lab will explore is Spark. Preliminary work so far as been the development of a introductory workshop on the technology (available on the Open Data Lab github repository). A second pedagogical series will be presented early in 2019 and will lead to testing different technical solutions.

2.4.3 SPARQL Endpoint

The numismatic dataset will be accessible through a SPARQL endpoint. This exploration is in the early stage and has not matured to the point of evaluation. The next annual report will have a full breakdown of the best way to treat this form of data and delivery.

Chapter 3

Research

This section of the report includes excerpts from research publications produced by groups using the open data lab. We highlight a couple MSDS capstone projects (3.1 and 3.2) as well as a faculty research project (3.3) and the undergraduate machine learning club (3.4).

3.1 Bourne/Mura Capstone

Sean Mullane, Ruoyan Chen, Sri Vaishnavi Vemulapalli, Eli J. Draizen, Ke Wang, Cameron Mura, asnd Philip E. Bourne

The biological function of a protein stems from its 3D structure. Understanding these functions is important in biomedical research. Given the high costs of using experimental means to determine protein structures, current methods do not scale. As a result, many have attempted to apply machine learning to this problem to predict structure from amino acid sequence data.

Our work focuses on a sub-problem of this field: predicting locations of the capping motifs which terminate α-helix protein structures. These motifs have only been described empirically to date [1]. As a step toward a robust and statistically-based understanding of helix capping, we demonstrate that machine learning modules can be trained to predict helix cap positions from sequence data.

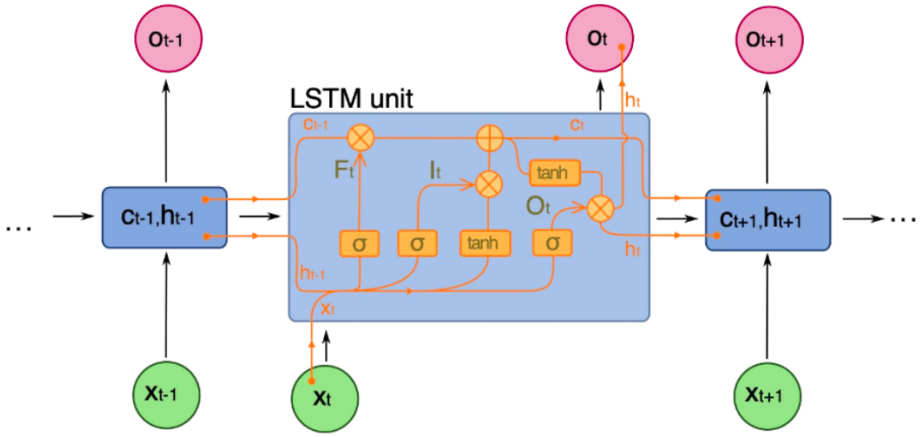


Figure 3.1: LSTM with feature vector of single amino acid residue as input to each cell

From poster presented at SIEDS 2019.

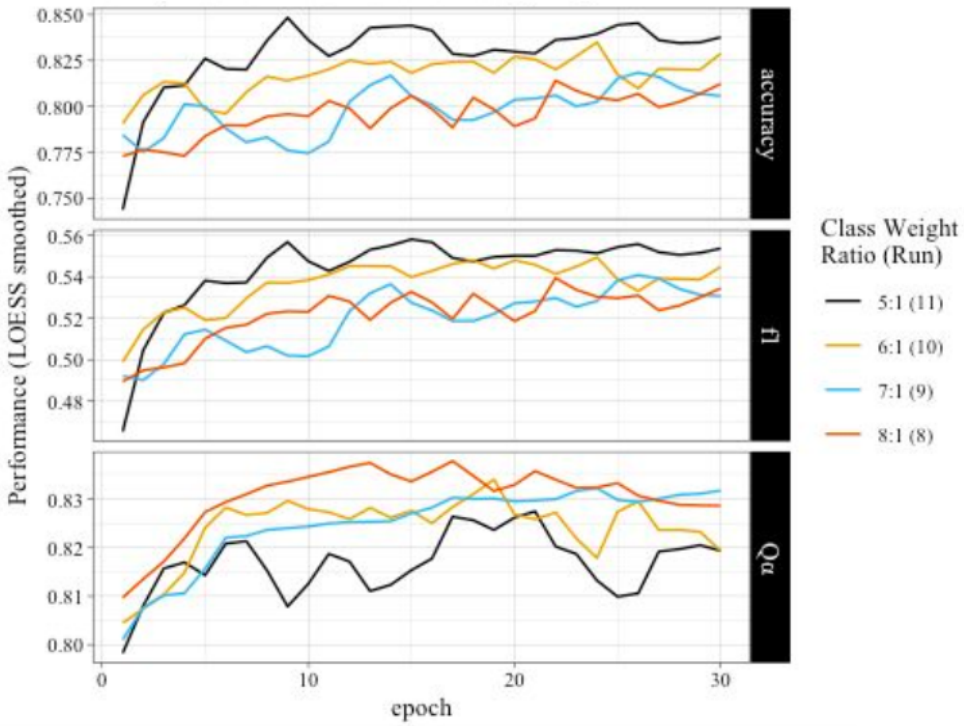


Figure 3.2: Effect of Loss Function Weighting

3.2 DSI Wiki Capstone

Charu Rawat, Arnab Sarkar, Sameer Singh, Rafael Alvarado, and Lane Rasberry

In this paper, we propose a framework to understand and detect abuse in the English Wikipedia community. We analyze multiple publicly available data sources provided by Wikipedia. We propose a web scraping methodology to extract user-level data and perform extensive exploratory data analysis to understand the characteristics of users who have been blocked for abusive behavior in the past.

We further build upon these insights to develop an abuse detection model that leverages Natural Language Processing techniques, such as character and word n-grams, sentiment analysis, and topic modeling, to generate features that are used as inputs in a model based on machine learning algorithms to predict abusive behavior.

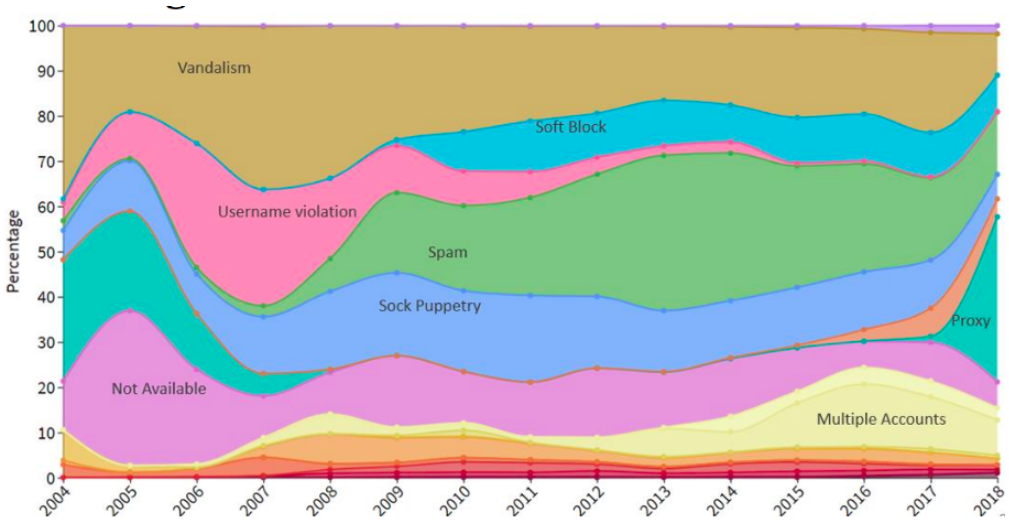


Figure 3.3: Share of block reasons over time

From poster presented at SIEDS 2019.

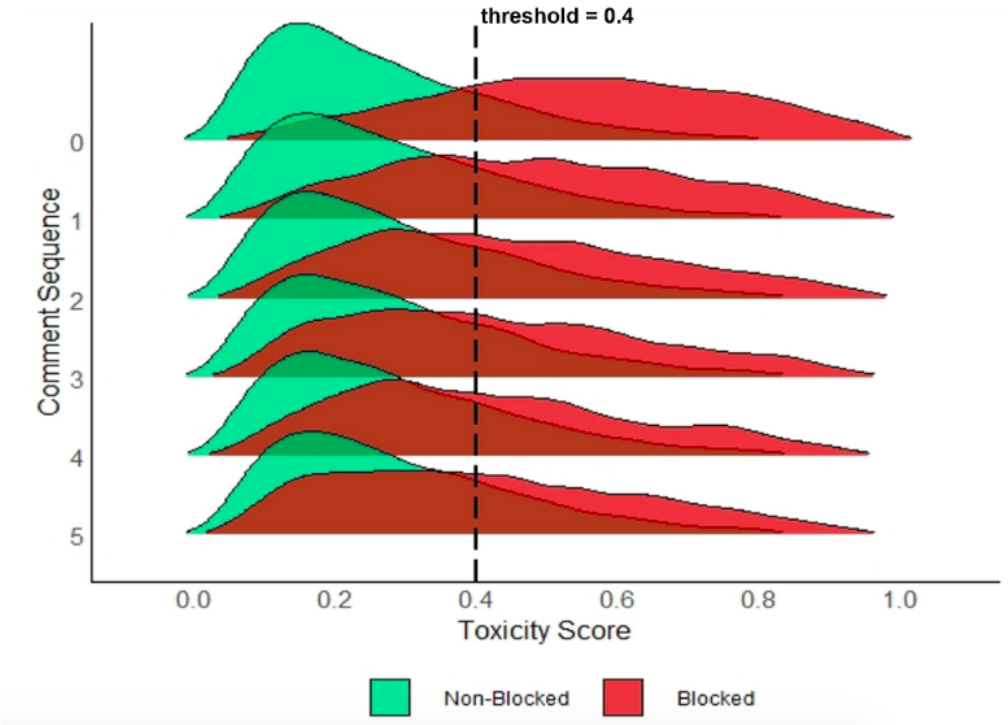


Figure 3.4: Toxicity score evaluation

3.3 Healthy Markets

Narges Tabari

In Healthy Market project, our goal is to understand the effect of high frequency trading on variables such as social media, stock market, retirement plans, and unemployment rate. To do this, we improved the structure of multiple causal models. Examples include:

- Identifying trends in high frequency trading with structural causal models that were improved by treating the time-series data as images and leveraging the power of convolutional neural networks (CNNs). We use time series data encoded as images using Gramian Angular Fields (GAF) in a classification task. We then built a structural causal model, based on the architecture of the model, that learn structural equations using CNN filter values (Shown in figure 3.5). By doing this, we are able to visually present the important patterns that the model looks for in the image that can potentially be interpreted as variations in the time-series data.

- Improving the construction and interpretability of Granger causal models with long short term memory networks (LSTMs) and hierarchical lasso penalties to find out the granger causal nature and the time lag selection between stocks of different companies

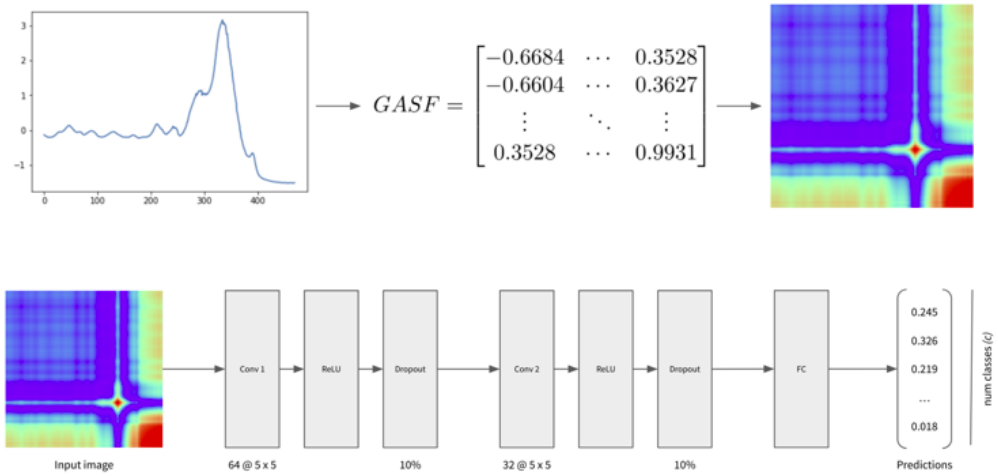


Figure 3.5: Schematic of CNN filter process

3.4 Undergraduate Machine Learning Club

Jake Grigsby, James Hanson, Hugh Jones, John Morris, Jonah Weissman, Zabihullah Yousuf

This semester, the Undergrad ML group built the foundations of a new Deep Reinforcement Learning platform. At its center is an implementation of Proximal Policy Optimization with Random Network Distillation. This technique, first published by OpenAI in late 2018, augments the PPO learning algorithm with the 'curiosity' to find novel states - greatly increasing performance in environments with sparse rewards by giving the agent continuous intrinsic ones. Our version is currently geared towards vision-based video game tasks; so far it supports any pixels-only environment from both Gym and Gym-Retro, but can be modified to work with any similar RL package. While the original paper came with an open source implementation, our version is more general purpose and written in TensorFlow's eager execution mode, which will make it easier to extend, debug and maintain going forward. It is also very scalable: we use synchronous gradient descent to process small amounts of data (a single rollout) on each worker, which eliminates the need for a GPU and lets us take full advantage of CPU clusters.

So far we have primarily tested the agent on Sonic The Hedgehog. While not as difficult an exploration environment as famous benchmarks like Montezuma's Revenge, Sonic is an interesting problem because of its very inconsistent difficulty. Long stretches of running right and jumping are broken up by obstacles that require significant long-term planning, creating bottlenecks that tend to destabilize training and require the agent to try new strategies over and over again. Our agent performs surprisingly well, even with just a few parallel workers. Figure 3.6 is an example chart of the training process, which shows the map and the agent's deaths, as well as the internal and external rewards.

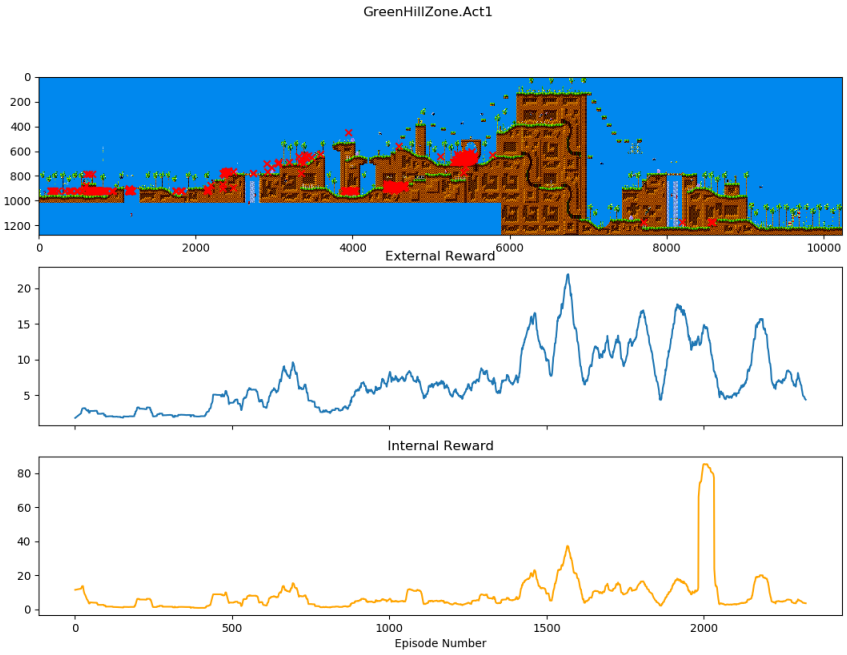


Figure 3.6: Example training process including rewards

Chapter 4

Education

A key component of openness is making resources usable. This idea falls in line with the idea that openness and accessibility are part of the same mission. As a result the educational component of the Open Data Lab is vital to the suces of the project. There are two main thrusts in this endeavor. The first, which was piloted in 2018, is the production of educational materials and methods. The second part is the development of communication paradigms. This year, two workshops were produced and delivered as part of the closed beta test. The first focused on the scale data protocol spark and the second the version control tool GitHub.

4.1 Spark Workshop

This workshop was designed as an introduction to spark. The goals were:

- Teach how to get started
- Build comfort
- Teach how to get answers to further questions

The topics covered included linking to a spark context, reading in data via dataframes, manipulating the data, and making a fundamental calculation. To power the workshop the attendees were given credentials on a Amazon SageMaker notebook. One of the features of this approach is that the whole workshop has access to the same environment. Everyone sees the same implementation of the software and hardware. There is no cumbersome overhead

in getting set up. The requirements are a web browser and access to the internet. Furthermore the single notebook environment leads to a very useful pedagocial capability. When a student encounters and error in their code the instructor can load their notebook on the main display in the room. In real time and in full view of everyone the instructor can debug and teach the whole class. This is a vast improvement over the current popular method of hovering over a single learner's station. It enables every person in the room to see what is going on and maintain their level of engagement. This experience was very positive for the learners and the feedback to this approach was superlative. Previous versions of spark training was done with Databricks resources and there were several drawbacks precipitating the switch to Amazon.

- The environment is not shared between the workshop participants and the instructor
- Every learner independently established their own cluster and there is substantial lag
- Materials must be imported in Databricks format (.dbc) instead of more universal jupyter notebook format (.ipynb).

Resources:

- Databricks based workshop can be found at:

  https://github.com/alonzi/spark

- Amazon Sakemaker based workshop can be found at:

  https://github.com/alonzi/spark-intro

- Next generation materials will be incorporated into the Open Data Lab repository at: https://github.com/UVA-DSI/Open-Data-Lab

The cost to operate and instructional environment is $0.0464 per hour. For this workshop we ran the environment for one week at a cost of approximately $10.

4.2 GitHub Workshop

The Open Data Lab was invited to present GitHub to the Archaeology Department of the Thomas Jefferson Foundation (aka Monticello). We developed a workshop to explain the fundamentals of version control and present a workflow for beginning users. The different user archetypes were also discussed. One of the major burdens to version control use is that it comes from the computer superuser community. Most of the software is developed using a terminal based interlace (CLI). However in today's research world many one computer superusers interact with code and other materials that benefit from a version control workflow. The major benefit from GitHub is the browser based interface. This implementation shifts substantial pieces of cognative load off the user. This shift enables the user to focus on developing their work rather than on the bookkeeping of version controling their work. At the same time it makes it easy for the developers to take advantage of the version control benefits. There is substantial room to further develop materials for different user archeytypes. This workshop focused on a research group. We will strive to identify other archetypes and develop materials to suit those needs. This workshop was taught from the GitHub repository itself. That was a natural fit given the subject matter. But it also demonstrated several very useful pieces of GitHub as a teaching medium, which will be discussed in 4.3. Resources are found at the Open Data Lab github page[1].

4.3 Using GitHub as a Teaching Medium

Both of the workshops taught under the Open Data Lab project used GitHub as the repository for materials. This has several benefits.

- GitHub provides a URL and free hosting for resources
- Subsequent changes to the materials are stored under version control thus allowing the actual materials presented to be recovered
- Any learner who wants to suggest improvements to materials can implement a pull request

The decision to put the materials in GitHub was one of necessity since the Open Data Lab GitHub page serves as the repository for all Open Data

[1] https://github.com/UVA-DSI/Open-Data-Lab

Lab resources. GitHub by default provides a URL for every item stored in the repository and presents the README file of a repository rendered automatically from markdown. Wikipedia has demonstrated the success of using markdown for content presentation but to enumerate some key features here. The document is organized, hyperlinkable, figures are easily embeded, and it seamlessly renders text alongside code and mathematical formulae.

4.4 Plans for 2019-2020

The plans for the future of Open Data Lab education efforts are part of the upcoming strategic plan. The educational component runs across all goals. As we finalize the areas in which to make an impact educational tools will be developed to support those efforts. Most likely this will take the form of content on github and workshop materials.

Chapter 5

Data sets

The Open Data Lab is in the business of hosting data sets with various levels of openness. We encourage all users to make their data as open as practicable. Currently there is a purchased data set reffered to as 'Healthy Markets' in the ODL (it contains financial information). And we are bringing a Numismatic data set online. There are also various data sets for student research projects hosted on the open data lab.

At the end of 2018-2019 academic year the total amount of data in the lab was 13.6 TB. This is predominantly from the Healthy Markets Dataset. Figure 5.1 shows the data usage overtime.

5.1 Healthy Markets

The Healthy Markets dataset was purchased by the University under liscense for use by all members of the university. The Open Data Lab is responsible for storage and providing access to the dataset. Narjessadat Seyeditabari (Narges Tabari) is responsible for facilitating research on the dataset and is the primary user of the dataset.

The data set is 13 TB and is stored in an S3 bucket named 'odl-hmtt'. This bucket is private to users of the Open Data Lab and is accessed through a SageMaker instance. To copy the data from it's source at Healthy Markets we used the AWS CLI and it took several days with a total cost of $1 for REST actions and $17 for the EC2 instance to manage the transfer. The current cost to store the data set is $290 per month. These costs are covered by the Healthy Markets PTAO.

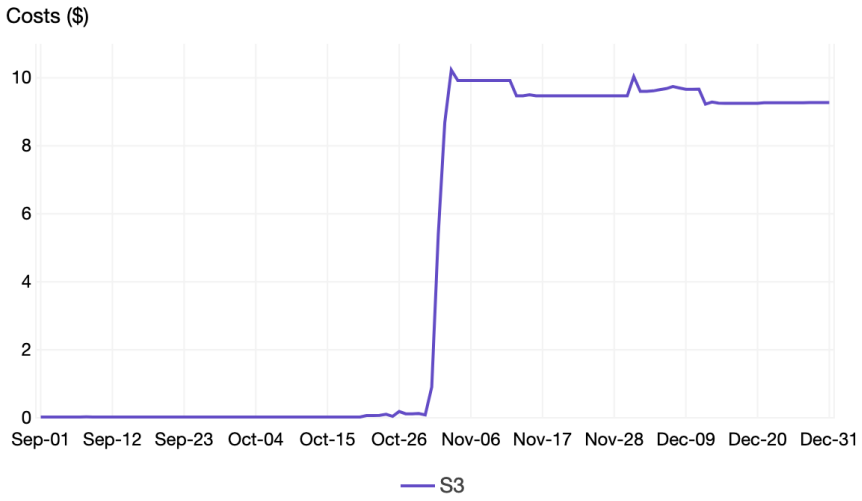


Figure 5.1: Healthy Markets daily cost for S3 service (2018)

Section 3.3 contains the report on research activities with this data set.

5.2 Numismatic

Dr. Ethan Gruber from the American Numismatic Society is in possession of a Linked Data Set and wants to make it open with the Open Data Lab. He is interested in establishing a SPARQL endpoint for the dataset. An S3 bucket (odl-nept) has been provisioned to store the data for the project and we will be making the data set open in 2020.

Chapter 6

Financial Report

6.1 Budget

The budget for the first year of the open data lab has a scope consisting of a purchase order for AWS time and personnel detailed from the Data Science Institute.

Source	Time (FTE)	Time (hours)	USD
	Income		
DSI Funds			10,000
HM Funds 2018			2,250
Data Scientist	1/2		
Clark Lab Dev		80	
DSI Staff		100	
Total Income	1/2	180	12,250
	Outlays		
AWS 2018			897
AWS 2019			8,775
Personnel Assigned	1/2		
Personnel Temporary		180	
Total Outlays	1/2	180	9,672
Net	0	0	2,578

Table 6.1: ODL Income and Outlays for 2018-2019

6.1.1 Income

- Cash: The DSI allocated $10,000 for computation and storage from AWS from the period of April 1, 2018 - May 31, 2019. These funds are drawn from PTAO[1].
- Cash: The Healthy Markets research project used ODL services for computation and storage on AWS. The ODL is authorized to transfer cost in proportion to the Healthy Markets PTAO[2].
- Personnel: The DSI allocated 50% of a staff Data Scientist's time to the ODL project (1/2 FTE). The Clark lab contributed 80 hours of software developer time.

6.1.2 Outlays

- Cash: During FY2018Q1/2 the ODL outlay for AWS was $897.
- Cash: During FY2019Q3/4 the ODL outlay for AWS was $9,672.
- Personnel: The ODL used the 1/2 FTE from the staff Data Scientist as well as a small ammount of time in meetings from various collaborators. AWS code development was contributed by Clark lab software developer.

6.1.3 FTE analysis

Development progress was made before the start of the fall semester in August of 2018. However the role of the only person on FTE detail to the ODL shifted to become a support role for the research and teaching mission of the DSI. As a result for the development progress on the ODL came to a halt. Now we have a good estimate on the needs for computational support for the DSI. If the DSI grows that requirement would increase. We recommend reaching out to Bryan Wright from the Physics Department to serve on the hiring committee for future positions in this area.

Additionally the ODL never achieved a bus factor greater than 1 for any component. As a result there were substantial wait times and often service was delayed due to the FTE being already allocated.

[1] 147242.102.LC00112.30002
[2] 147412.107.DR03397.30002

6.2 AWS usage

The main driver of AWS cost is the usage of compute resources by research groups (see 6.4). This was a surprising finding. Our belief was the data storage would be the driver. However only one project involved a data set greater than 1 TB. There were 22 S3 buckets created and six contain more than 10 GB. Two buckets contain more data than is easily processed on a laptop. Those are the wikipedia capstone bucket at 400 GB. And the healthy markets bucket with 13.3 TB.

The main driver of the compute cost was the use of ml.p2.xlarge instances through sagemaker. Those instances cost $1.26/hour at time of writing. They are Tesla K80 GPUs from Nvidia with 61 GB of conventional memory and 12 GB of GPU memory.

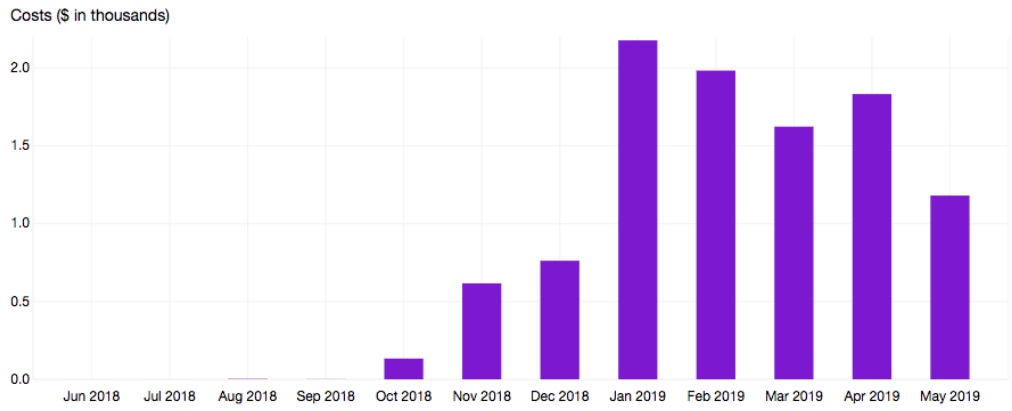


Figure 6.1: AWS Total Costs

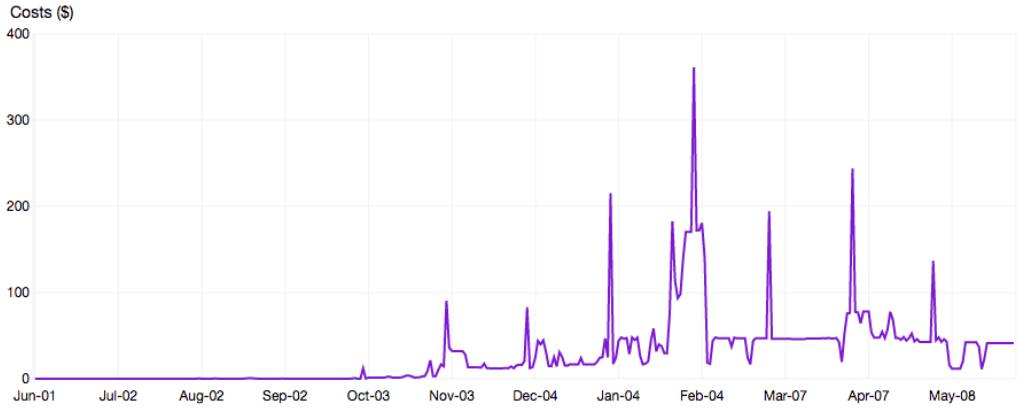


Figure 6.2: AWS Total Costs, daily

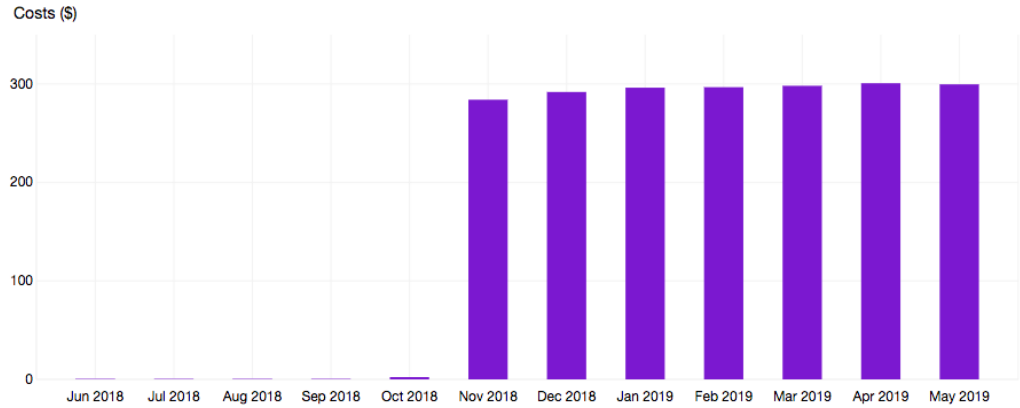


Figure 6.3: AWS S3 Costs

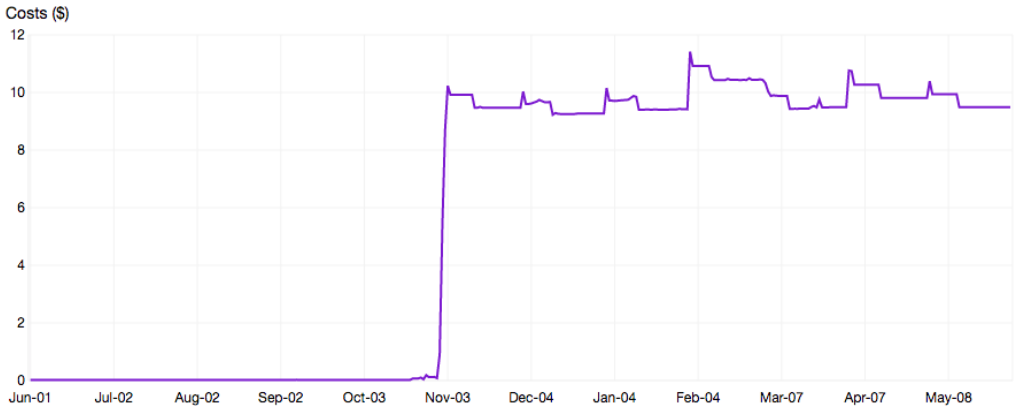


Figure 6.4: AWS S3 Costs, daily

6.3 Local UVA HPC usage

The Open Data Lab assists researchers in getting access to the local UVA HPC resources. For many of the tasks our users have there is no direct charge from local HPC and we do not track this usage. For more sustained efforts we do track the usage broken down by allocation.

The undergraduate UVA Machine Learning Club used up their entire initial allocation of 5,000 SUs. They have been issued a new allocation of 100,000 SUs. The results from one of their projects can be seen in section 3.4

The graduate UVA Machine Learning Club has used over half of their intial allocation. They have 3,335 SUs used and 1,665 SUs remaining.

6.4 Sustained Support

Many researchers are interested in using the Open Data Lab for continued hosting of their data sets to satisfy data management plan requirements of funding agencies. For example a federal agency that requires a seven year data commitment. In order for the open data lab to support this effort a budget model, at least in part, must provide for multi year sustained support.

6.5 Funding Sources

The ODL is funded by the DSI. The support for the Healthy Markets project is fronted by the ODL and then cost transferred from the ODL PTAO to the HM PTAO.

Chapter 7

Conclusion

The 2018-2019 year was the first year for the Open Data Lab. In this time we learned many things and many users attended workshops, hosted and anlyzed data, some event produced a publication. We served 116 users, working on 25 projects, and 6 individuals contributed directly to the ODL project. Going forward we look to increase all of those numbers including users from beyond the Data Science Institute.

Here are a few highlights from 2018-2019:

- Development of a GitHub workflow for beginners useful beyond the hard sciences. This workflow was adopted by an international collaboration called the Open Greek and Latin Project [3] as well as the Archaeology Department of Monticello [4]. Details are found on our github page and in Section 4.2.

- We realized the power of Project Jupyter as a system to deliver resources to the user without excessive cognitive load. This platform is transformative and will lead to great things in the future. Details are found in Section 2.3.1.

- We studied three user archetypes for the Open Data Lab: the collaborator, the student, and the sharer. Details are round in Section 1.2.

We have accomplished a lot but now we need help. If you have an idea or want to join the team please reach out by emailing datascientist@virginia.edu (Subject Line: "I want to help the ODL").

Appendix A

Open Working Group Report

DSI Open Working Group Summary

The Open working group convened several meetings during the first quarter of 2019. Pete Alonzi served as the chair of the group and the members were. Claudia, Cathy, Daniel, Lane, Samuel, Ellie, Narges, and Tim. **The mission of the group was to begin untangling the definition of 'OPEN' and present guidance for the DSI to decide on a way forward**. In this document we present the findings of the group including a recommendation of a DSI elevator pitch on our definition of open as well as a decision tree tool to help someone put their work into the context of open. There were also several documents produced over the course of the work and they are linked at the end of the document.

Summary of Findings

This working group started from the document on defining open from the DSI retreat in December 2018. The group quickly identified that everyone brought to the table a different definition of open and in particular this was complicated by everyone applying open to a different context. That variety led to the taxonomy document to classify different types of 'data' and frame how open would apply to them. In particular one example is that the Medical community has several legal and ethical obligations regarding their data. **When discussing openness care must be taken to ensure the audience does not extrapolate to conclusions which will shut down discussion.** To that end we identified that there is a critical need for a brief and universal explanation of our core principles surrounding openness.
Two suggestions are put forth from the working group for a way forward to facilitate this crucial communication step.

1. **Elevator pitch.** The goal is to take a few paragraphs to frame the discussion of openness and deliver key core principles. This enables us to get the ideas across and stop the audience from running to conclusions. By carefully selecting how we describe the principles the audience will understand our position and how to extend it.

2. **Decision tree.** The goal here is to produce a tool that is easy to understand, easy to use, and easy to remember. Having such a tool will enable us to engage people in a little exercise to understand how their work fits into the world of open.

The working group presents a draft of the elevator pitch and the decision tree for further discussion by the DSI. We believe that as a team the DSI refining these tools will help to unite ourselves and enable us to take the next steps. Going forward we suggest a major revision to the Open UVA document and then the SDS leading a global discussion in this space.

There are also substantially more details in the notes from the meetings. All are encouraged to seek out members of the working group and have further conversations.

Proposed Action Plan

1. DSI refinement of Elevator Pitch and Decision Tree and Project Form
2. Potential formation of new Open working group to focus on a new mission
    a. Group to refine elevator pitch
    b. Group to refine the decision tree
    c. Group to analyze and refine project form
3. DSI/SDS development of the Open UVA document with consultation from UVA at large and outside scholars

Elevator Pitch

Openness means team members responsibly sharing their data and professional endeavors (when possible and ethical). We believe in the importance of practicing openness because advancement requires assembling a heap of known pieces into a coherent picture containing new knowledge. In the world today some of the necessary pieces are unknown due to traditional non-open information practices. Wide spread open practices are the first steps to changing the world.

<< finish with a quote beethoven, jefferson, nobel prize winner, etc.>>

- "Educate and inform the whole mass of the people... They are the only sure reliance for the preservation of our liberty." ~Thomas Jefferson
- "There should be only one repository of research in the world, to which the artist would donate their works in order to take what they would need." ~ after Beethoven

Decision Tree

We developed several avenues for thinking about a decision tree for open. In fact due to the breadth of what open covers one decision tree is not enough and several are needed to span the space. For instance you can try to help someone understand how they fit in by thinking about their most recent data set. Or you could pick a particular kind of data and work through the implications of that kind. Here we present one example (mockup) of a decision tree.

A decision tree to help someone think about sharing their data. In this case we ask the audience to think of the last data set they used. Then we pose four simple questions:

1. Does it contain sensitive information?
2. Is it digital?
3. Is there an established platform to share this type of data?
4. What are you waiting for?

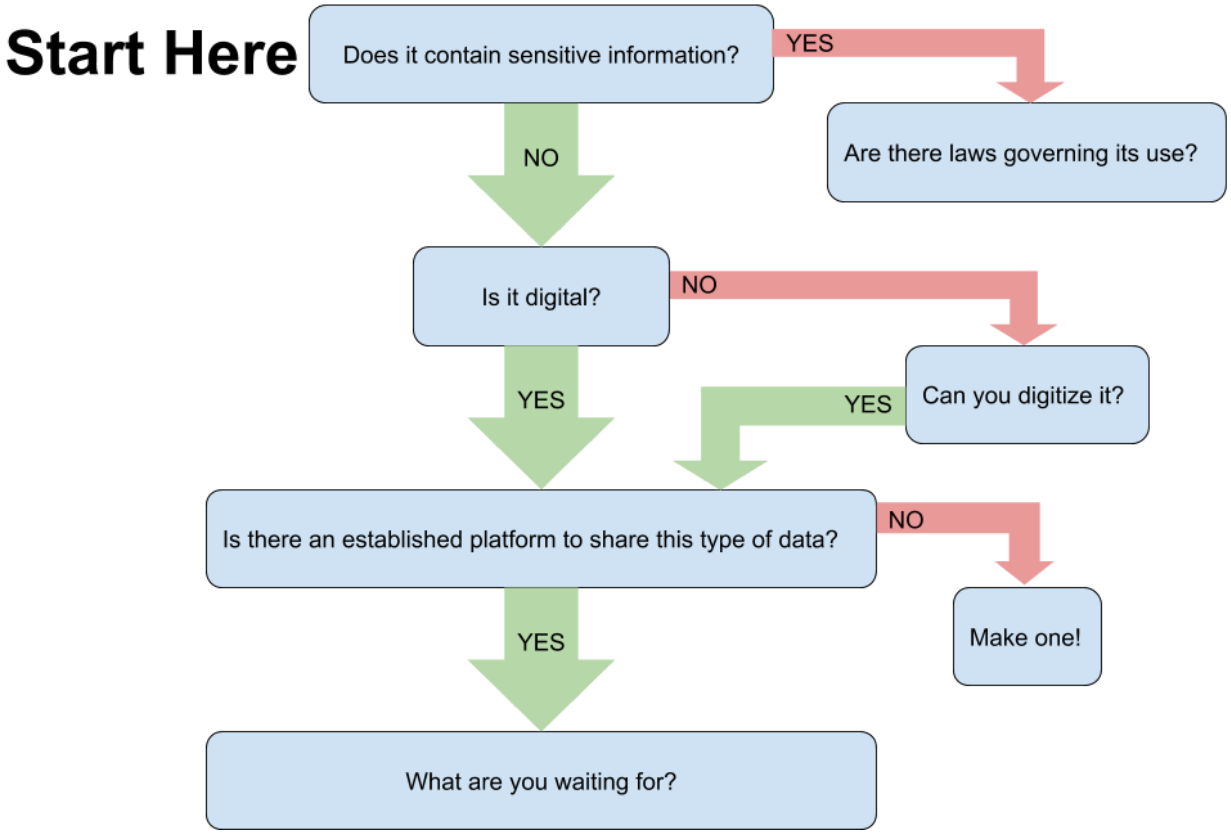
<<

philosophical underpinning are points that guide the tree and the presenter uses to guide the discussion from the tree. they are explicitly on the tree because we want to hide the complexity from the audience. but rather guide them into it, lower the barrier to entry, lower the cognitive load.

1. opens the door to the discussion of legal and ethical responsibilities
2. starts the conversation about the necessity of digitization to openness
3. makes the audience think about established practice in their space
4. makes the audience look at themselves as the agent of change and openness

>>

Imagine the last data set you used.



Data Science Institute project log

Faculty, staff, students, researchers, and anyone else doing projects at DSI may report their projects here. This log seeks to create a record of what people do, what parts of our projects meet the DSI values of openness, diversity and ethics, and what parts do not.
This form should take 5 minutes or less to complete.

https://docs.google.com/forms/d/e/1FAIpQLSd5zT1vC6qUkbxtBLgcwEci9vvstQ6datzw2NGFI-1H7LJagQ/viewform

Documents of Record

- DSI Open Thoughts - from december 2018 retreat
- Open Working Group Meeting Notes
- Taxonomy Document
- FAIR diagram
- Position Paper
- Decision tree notes
- Decision tree drawing
- Open Hierarchy Diagram

Links

- Working group notes- https://docs.google.com/document/d/1XAxVxDqzNJwl4Krbeh_8sDLg55DesNiKqE9SXbo9uW4/edit#heading=h.tr07zg3i85mu
- Open uva - https://docs.google.com/document/d/1szvWOvdxQp84uE4Fsdk0wAIsUWddgYXQmAwFDk0AycM/edit
- Open taxonomy - https://docs.google.com/document/d/1q4lg2jkM1srCm5F9aN6HM2hmJjEOhmF8uAuudY-O8qs/edit
- Dsi retreat worksheet on open - https://docs.google.com/document/d/1BtjJMd1W5TL3fDtAJIVHExZgGs92qVv9hxZnMAW0DZk/edit
- Open pyramid - https://drive.google.com/file/d/1pDH8VdLK2ZkLGmI3u4aasQb2ZvE1t5Hl/view
- Open decision tree - https://docs.google.com/drawings/d/1LteSNaKTlYhN07k3DdJax75IGu6QOqvq_WZiy_Ltdh4/edit
- Decision tree notes - https://docs.google.com/document/d/1mFNu1wtXnqPtotWHd44soZeOmWoZM-BOn-li1Ie0kgs/edit#heading=h.pzeqa4un0rqe
- Position paper - https://docs.google.com/document/d/1EHeGYvUybGoeHBaXcBKe4Vw-PnRRBlB-47Y5-eYNUU8/edit#heading=h.klww9ixted5m
- Fair diagram - https://drive.google.com/file/d/1SEUG4Op2PQXoSbkkOK4gcf_4UxyspjEq/view

Appendix B

References

Bibliography

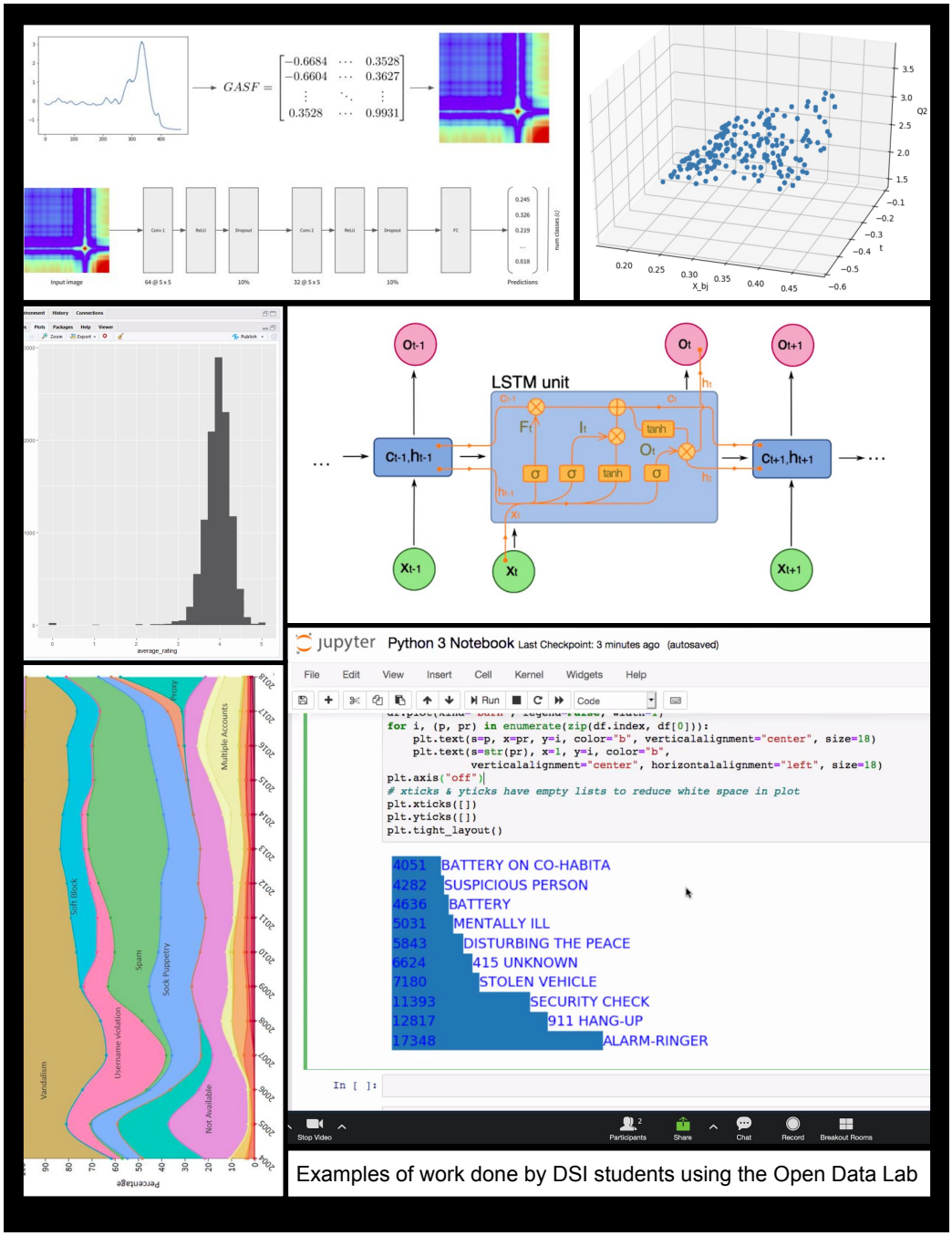
[1] Aurora, R. and Rose, G.D. 1998. "Helix capping." Protein Science 7(1), pp. 21-38.

[2] https://github.com/jakegrigsby/supersonic

[3] https://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/

[4] https://www.monticello.org/site/research-and-collections/monticello-archaeology

[5] https://www.internet2.edu/products-services/cloud-services-applications/amazon-web-services/



Examples of work done by DSI students using the Open Data Lab