Reproducible scientific computing using Vagrant, Ansible, and Anaconda.

Bret Davidson

NCSU Libraries

go.ncsu.edu/dsvil-sb

NCSU Libraries' Open Science Initiative

Goals

- explore open science practice at NC State
- better understand researcher needs in context

Modern Research Skills Gap

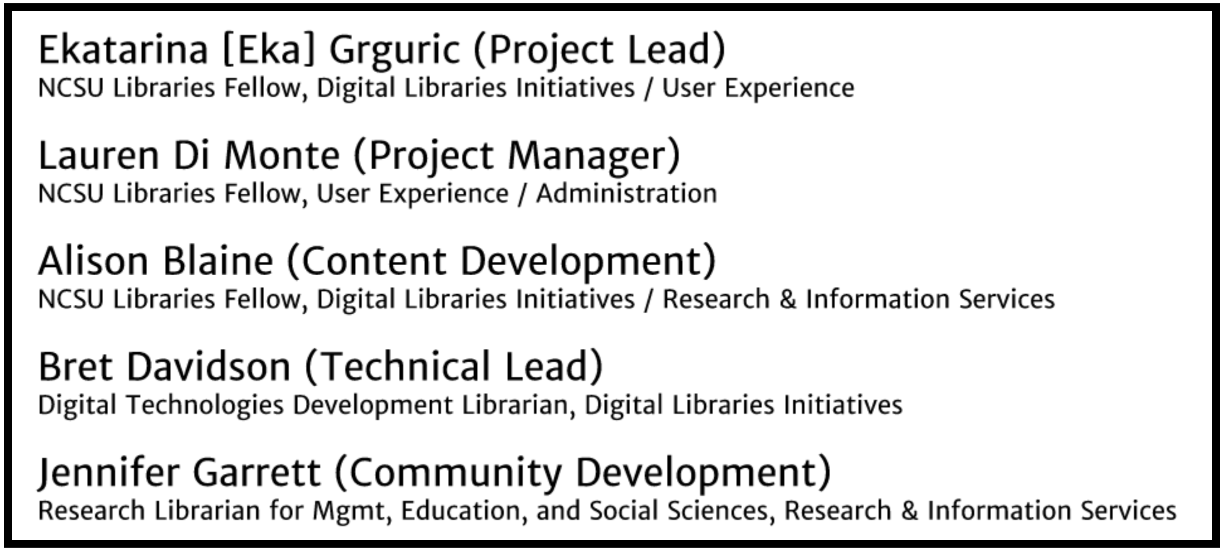
NCSU LIBRARIES
SUMMER
of
OPEN SCIENCE

Summer of Open Science

- Intro to the Command Line Interface
- Web Scraping with Python
- Understand and Build Your Scholarly Identity
- Scientific Computing with Python & Raspberry Pi
- Build Your Scholarly Website the Easy Way

SOS Planning Team

Representation from broad range of departments.



Ekatarina [Eka] Grguric (Project Lead)
NCSU Libraries Fellow, Digital Libraries Initiatives / User Experience

Lauren Di Monte (Project Manager)
NCSU Libraries Fellow, User Experience / Administration

Alison Blaine (Content Development)
NCSU Libraries Fellow, Digital Libraries Initiatives / Research & Information Services

Bret Davidson (Technical Lead)
Digital Technologies Development Librarian, Digital Libraries Initiatives

Jennifer Garrett (Community Development)
Research Librarian for Mgmt, Education, and Social Sciences, Research & Information Services

Instructors

Brittany Johnson

Eka Grguric

Lauren DiMonte

Alison Blaine

Madison Sullivan

Will Cross

Todd Stoffer

Interdisciplinary Need:
over 40 departments across ~16 colleges

Reproducible Computing

Technical workshops are ripe for disaster.

What could go wrong?

- Images reset overnight
- Improper permissions
- Network connectivity issues
- Language Versions
- Missing packages

Instructor Challenges

- Inconsistent user environments
- Inconsistent course materials
- Provisioning is time consuming
- Difficult to collaborate

Student Challenges

- Data types and structures
- Module system
- Control Structures
- Exception Handling
- Working with file system
- Retrieve a web page with Requests
- Parse content with Beautiful Soup
- Generate a word cloud with matplotlib

Computing Tasks
vs.
Computing Environments

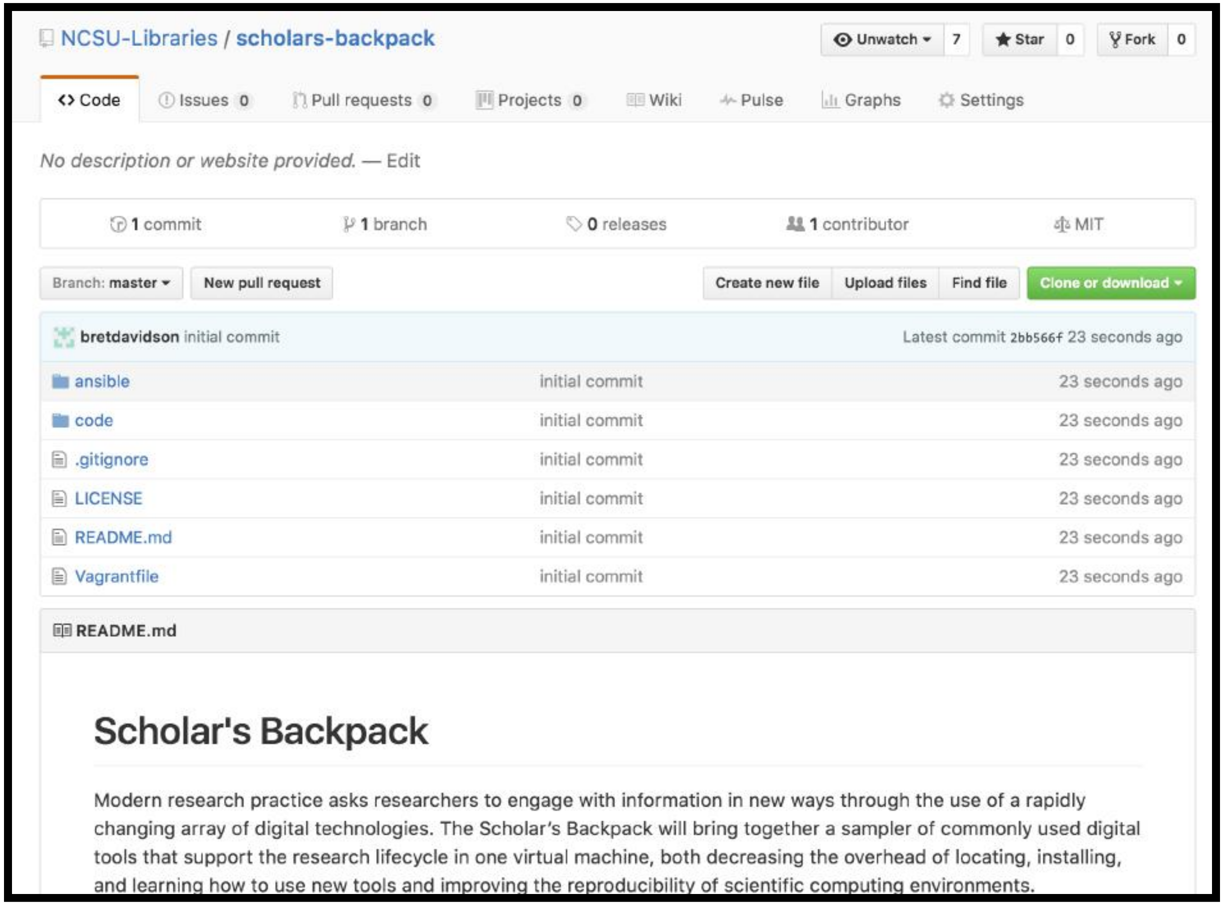
Rise of Scholarly Code

- Consistency across lab environments
- Ability to see results of code
- Consistency across time
- Ease of collaboration

Our Approach

- Vagrant for managing operating system
- Ansible for provisioning and configuration
- Anaconda for managing environments and packages
- Workshop specific resources

github.com/NCSU-Libraries/scholars-backpack



No description or website provided. — Edit

bretdavidson initial commit Latest commit 2bb566f 23 seconds ago

ansible	initial commit	23 seconds ago
code	initial commit	23 seconds ago
.gitignore	initial commit	23 seconds ago
LICENSE	initial commit	23 seconds ago
README.md	initial commit	23 seconds ago
Vagrantfile	initial commit	23 seconds ago

README.md

Scholar's Backpack

Modern research practice asks researchers to engage with information in new ways through the use of a rapidly changing array of digital technologies. The Scholar's Backpack will bring together a sampler of commonly used digital tools that support the research lifecycle in one virtual machine, both decreasing the overhead of locating, installing, and learning how to use new tools and improving the reproducibility of scientific computing environments.

Easy!

1. Install Vagrant
2. Install VirtualBox
3. Clone project repo
4. `vagrant up`
5. `vagrant ssh`
6. Execute code!

This is reproducible computing!

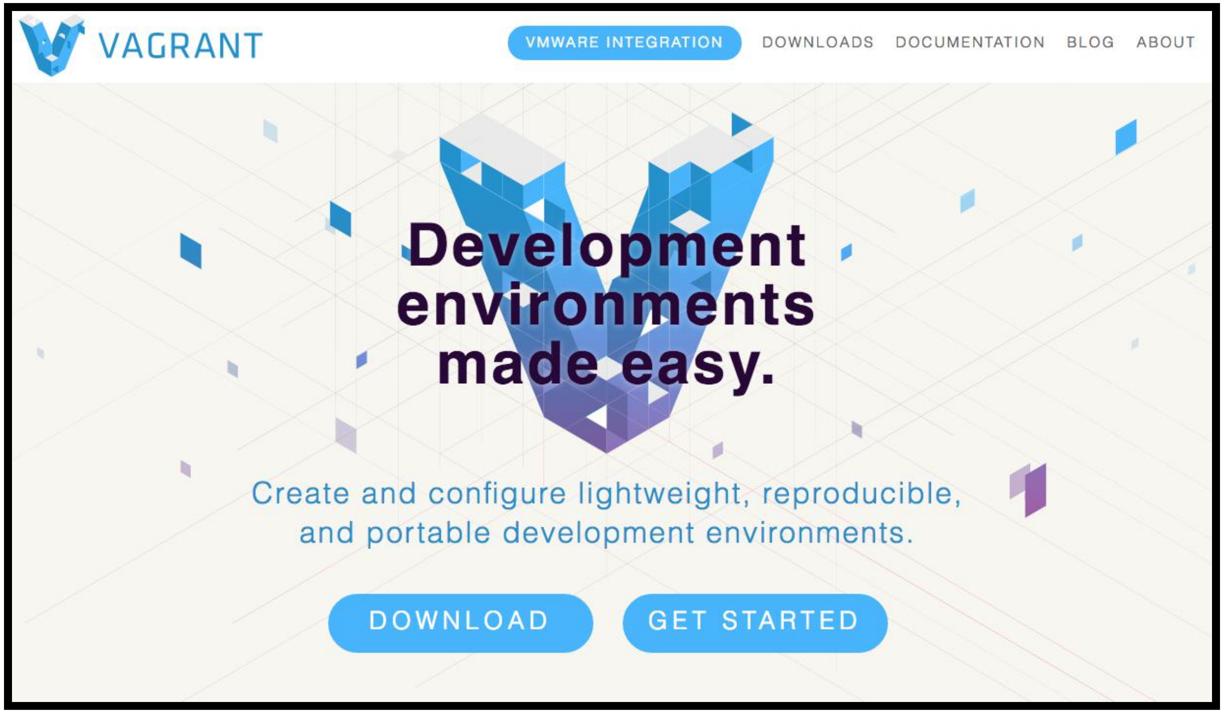
Benefits

- Consistent environment user to user
- Single target for course materials
- Faster provisioning for new workshops
- Repeatable course to course

Features

- Python
- R and R Studio
- Jupyter Notebook Server
- Example Notebooks
- Accessible from web browser

Vagrant



Development environments made easy.

Create and configure lightweight, reproducible, and portable development environments.

Create and configure lightweight, **reproducible**, and portable development environments.

Usage

- Easy installation through binary package
- Configured via **plain text file**
- Single command: `vagrant up`

Ansible

"Automation engine" for provisioning and configuration management.

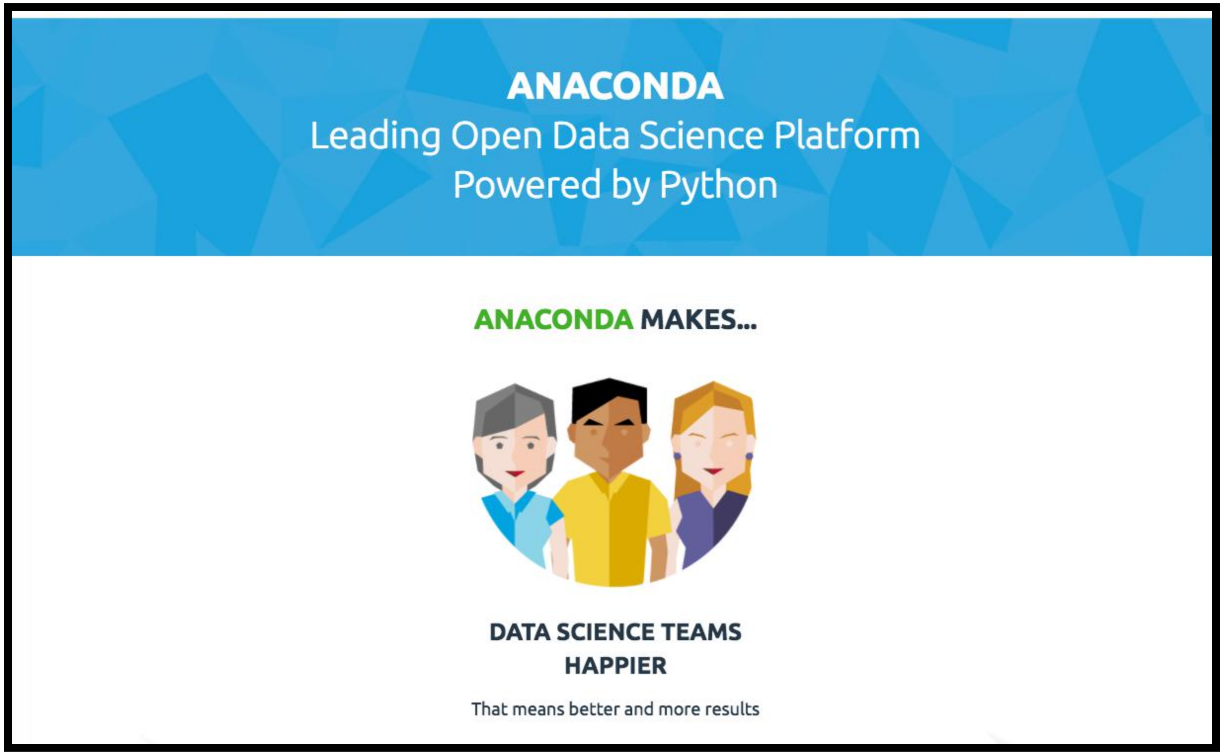
Provisioning

- Anaconda
- Python & R
- Software packages
- Jupyter Notebooks

Configuration

- Start Jupyter notebook server
- Set environment variables
- Set default login directory

Anaconda



ANACONDA
Leading Open Data Science Platform
Powered by Python

ANACONDA MAKES...

DATA SCIENCE TEAMS
HAPPIER

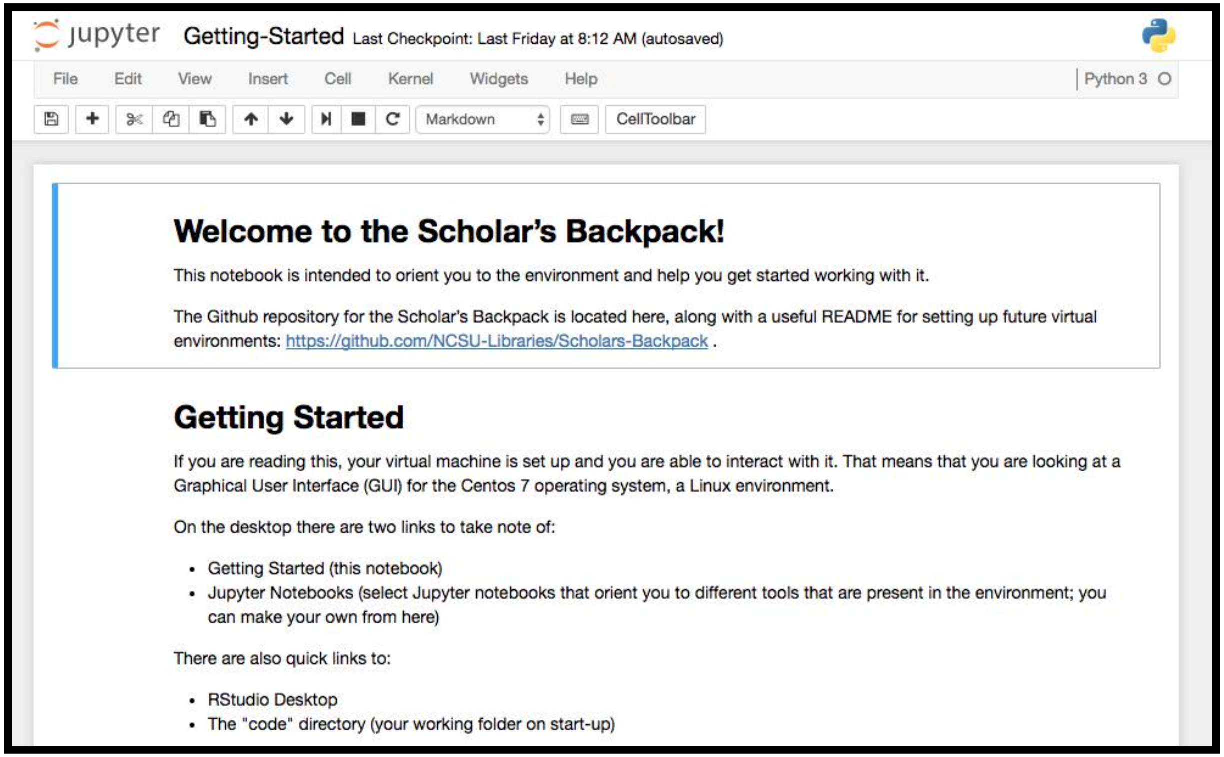
That means better and more results

Python Packages

astropy, beautifulsoup4, conda, flask, jupyter, matplotlib, numpy, nltk, pandas, pillow, pip, pytest, qt, requests, scipy, scikit-learn, seaborn, sqlite, etc.

R Packages

r, essentials, formatr, ggplot2, irkernel, knitr, kernsmooth, maps, markdown, mass, matrix, nnet, rbokeh, recommended, spatial, tidyr, etc.



Welcome to the Scholar's Backpack!

This notebook is intended to orient you to the environment and help you get started working with it.

The Github repository for the Scholar's Backpack is located here, along with a useful README for setting up future virtual environments: https://github.com/NCSU-Libraries/Scholars-Backpack .

Getting Started

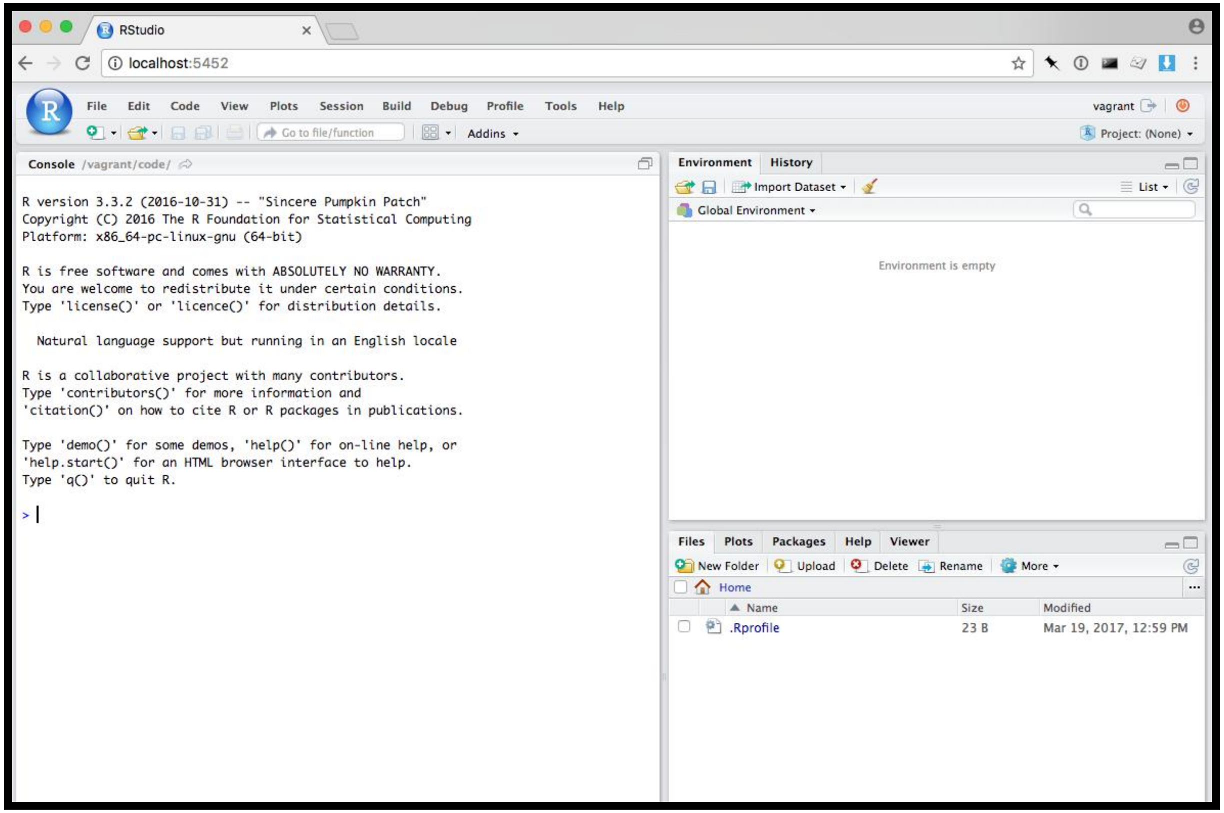
If you are reading this, your virtual machine is set up and you are able to interact with it. That means that you are looking at a Graphical User Interface (GUI) for the Centos 7 operating system, a Linux environment.

On the desktop there are two links to take note of:

- Getting Started (this notebook)
- Jupyter Notebooks (select Jupyter notebooks that orient you to different tools that are present in the environment; you can make your own from here)

There are also quick links to:

- RStudio Desktop
- The "code" directory (your working folder on start-up)



```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Ongoing Work

- Embedded use in curriculum
- Additional open source contributions

Summary

Open Science represents a new framework for research and provides an opportunity for libraries to engage researchers in new ways.

NCSU Libraries has done workshops and outreach around this framework and there is evidence of strong interest across disciplines.

We are redeploying existing technical resources and cutting edge technology in ways that used to be difficult or impossible.

This approach has helped us identify a new leadership role for libraries in open research support.

Thanks!

bret_davidson@ncsu.edu

github.com/NCSU-Libraries/scholars-backpack

go.ncsu.edu/dsvil-sb