

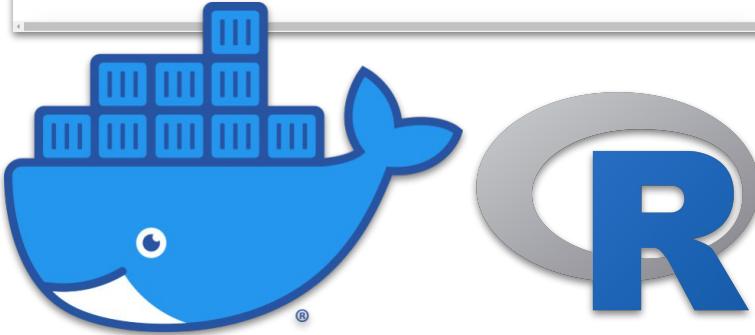
Get Started with Docker
We help developers and development teams
[Get Started](#)

READ MORE

AWS & Docker Collaborate

Speed Delivery of Modern Apps to the Cloud
AWS & Docker make life easier for Developers.

→ Read Press Release
→ Read Docker Blog
→ Download Preview Version



RStudio - lab_40_docker

File Edit Code View Plots Session Build Debug Profile Tools Help

Console Terminal

Environment History Connections Git

Global Environment

aws

Environment is empty

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Name Size Modified

Name	Size	Modified
00_data	838 B	Aug 5, 2020, 10:56 AM
00_production_model	6.5 MB	Aug 5, 2020, 10:56 AM
img	40 B	Aug 5, 2020, 10:56 AM
README.md	257 B	Aug 5, 2020, 10:57 AM
bankruptcy_report.html	2.8 KB	Aug 5, 2020, 10:56 AM
.gitignore	1.3 KB	Aug 5, 2020, 10:56 AM
lab_40_docker.Rproj		
bankruptcy_report.Rmd		
Dockerfile		

Docker for Data Science

Reproducible Machine Learning

Matt Dancho & David Curry
Business Science Learning Lab





Learning Lab Structure

- **Presentation**
(20 min)
- **Demo's**
(30 min)
- **Pro-Tips**
(15 mins)



Matt Dancho

Founder of Business Science, Matt designs and executes educational courses and workshops that deliver immediate value to organizations. His passion is up-leveling future data scientists coming from untraditional backgrounds.



David Curry

Founder of Sure Optimize, David works with businesses to help improve website performance and SEO using data science. His passion is ethical Machine Learning initiatives.

MLOps Series



- **Business Case**
 - Problem with H2O reproducibility
 - How Docker makes H2O reproducible
- **Docker**
 - Key Concepts
 - Docker Workflow
 - Working w/ GitHub
- **30-Min Demo**
 - Bankruptcy Classifier
 - MLFlow + H2O Grid Search
 - [Bonus] Run RStudio + H2O in Cloud w/ AWS
- **Keys to Your Transformation**
 - Reproducible Analysis
 - Deploying Applications



MLOps Series



A screenshot of a website page titled 'Learning Labs Pro'. It features a photo of three people working on a laptop, a brief description 'Community-Driven Data Science Courses', a profile picture of Matt Dancho, and a price of '\$19/m'. The overall theme is professional data science education.

- **Lab 39 - MLFlow**

- MLFlow Tracking API - H2O Machine Learning Artifacts



- **Lab 40 - Docker**

- Reproducible Analysis



- **Lab 41 - Drake & Renv**

- Workflow Dependencies, Caching Expensive Ops & Package Management



- **Lab 42 - Plumber APIs**

- Building APIs for apps to connect to



Learning Labs PRO

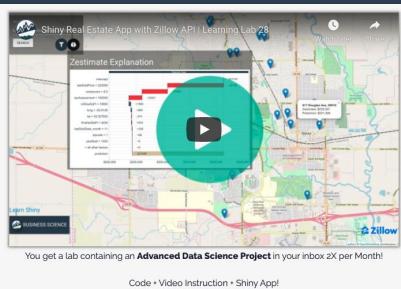
40 Labs

1.5-Hour Courses

2X Per Month

100's of Valuable Bonuses

\$349/year or \$39/month



Learn Data Science Tools



Get Labs 2X per Month on Advanced Topics

<https://university.business-science.io/p/learning-labs-pro>

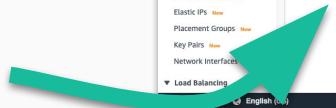
Bonuses

for LL PRO Members Today

Bonus #1

AWS

Tutorial



The screenshot displays a desktop interface with several open windows:

- RStudio - lab_40_docker:** A browser window showing the RStudio interface for a Docker container named "lab_40_docker". The global environment is empty. The file tree shows a directory structure including "OJ_data", "OJ_production_model", "README.md", "bankruptcy_report.html", "grapheR", "lab_40_docker.Rproj", and "bankruptcy_report.Rmd".
- aws Services - Resource Groups:** An AWS Management Console window showing the EC2 Dashboard. It lists instances, images, and security groups. One instance, "lab_40_docker", is selected, showing its details: Instance ID i-0c04ad64d887ea3eb7, Instance Type t2.micro, Availability Zone us-east-2c, Status running, Public DNS (IPv4) ec2-13-59-12-104.us-east-2.compute.amazonaws.com, and Private IP 13.59.12.104.
- Business Science - Data Studio:** A separate window showing a dashboard with a large yellow box containing the "H2O.ai" logo.

A green curved arrow points from the top right towards the RStudio window, and another green curved arrow points from the bottom left towards the AWS Management Console window.





Bonuses

for LL PRO Members Today

Bonus #2

MLFLOW

Model Tracking Automation

```
28 # Iterate through Grid Search
29 run_h2o_grid_search_with_mlflow(
30
31     data      = data_prepared_tbl,
32     target    = "class",
33
34
35     # H2O Grid Search
36     h2o_init  = TRUE,
37     n_trees   = c(25, 50, 100),
38     max_depth = c(5, 10),
39
40
41     # MLFLOW
42     launch_mlflow_ui      = TRUE,
43     mlflow_tracking_uri   = "mlflow_test_2",
44     mlflow_experiment_name = "Bankruptcy Prediction API"
45 )
```

H₂O.ai

mlflow

Problem Reproducing Analysis



Version Incompatibility

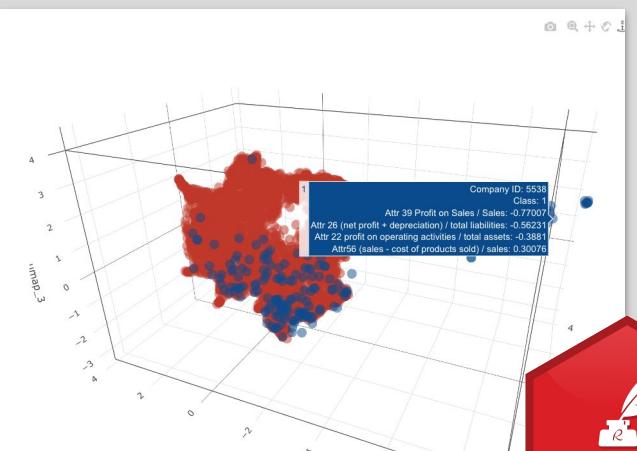
You can not load models built from
prior versions of H2O





Problem

Your Machine



Their Machine

```
69
70  ### Load Production Model
71
72 ~ ````{r}
73 path <- file.path(rprojroot::find_rstudio_root_file(),
74                      "00_production_model/PROD_H2O_MODEL1")
75
76 h2o_model <- h2o.loadModel(path)
77 ~ ````
```

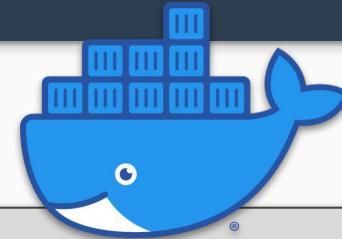
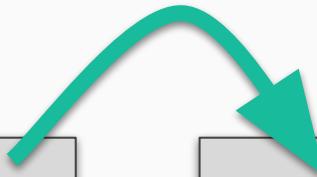
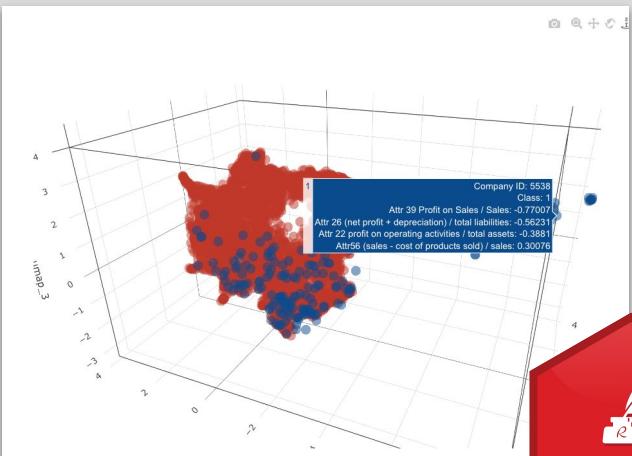
ERROR: Unexpected HTTP Status code: 412 Precondition Failed (url = http://localhost:54321/99/Models.bin)

```
water.exceptions.H2OIllegalArgumentException
[1] "water.exceptions.H2OIllegalArgumentException: Illegal argument:
water.api.FSIOException: FS IO Failure: \n accessed path :
file:/Users/mdancho/Desktop/learning_labs/lab_40_docker/00_production_
File not found"
```

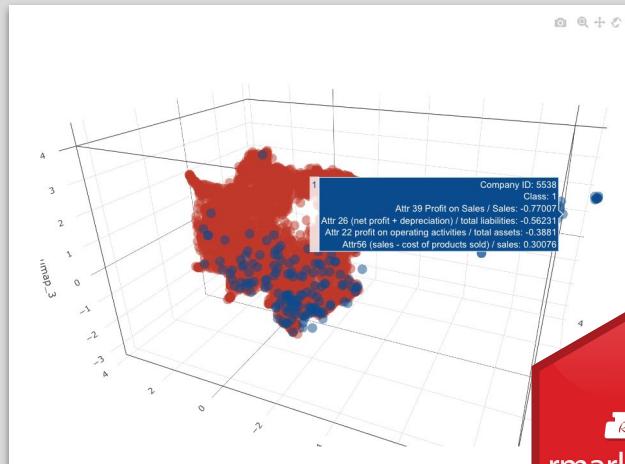


Solution

Your Machine

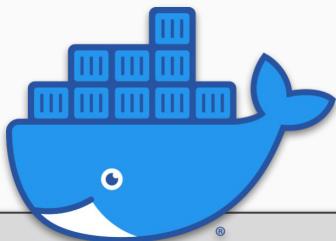
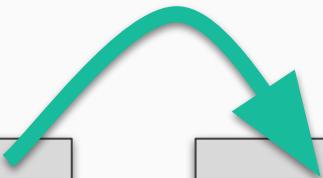
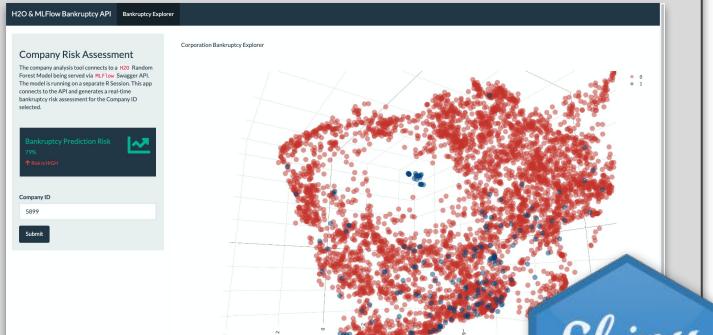


Their Machine

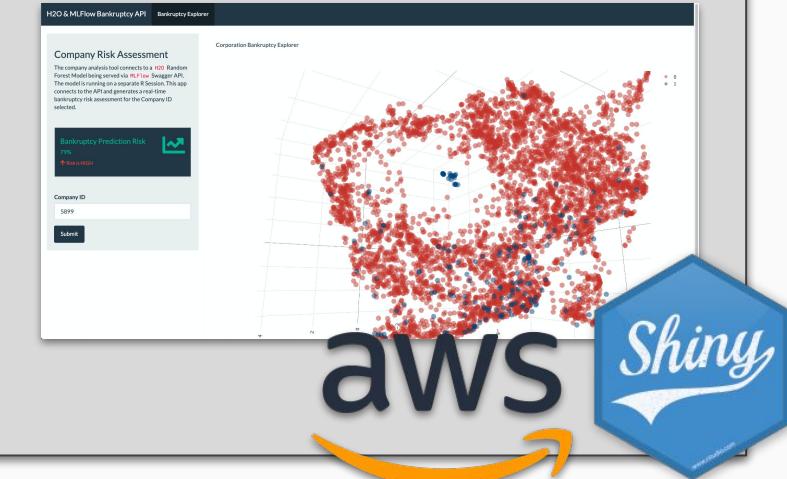


Solution

Your Machine



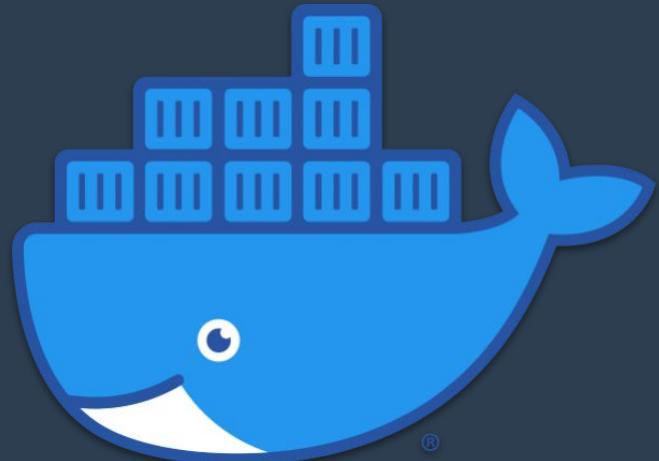
Their Server



What about Applications?

What is Docker

80/20 Concepts





What is Docker?

Compute
(AWS EC2)



Application

Files

app.R



Software

R
Shiny Server
Libraries
(shiny, tidyverse)





What is Docker?

3 Key Concepts

- 1. Image & Dockerfile**
- 2. Docker Hub**
- 3. Containers**

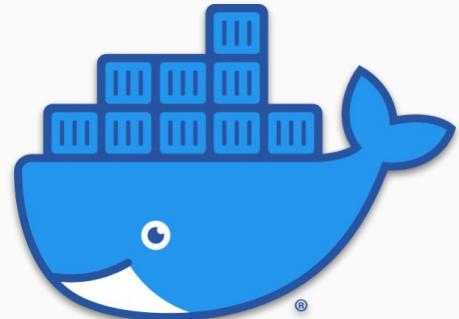




Image & Dockerfile

```
FROM rocker/verse:latest
## Install debian packages
RUN apt-get update -qq && apt-get -y --no-install-recommends install \
# usethis dependencies
build-essential \
libcurl4-gnutls-dev \
libxml2-dev \
libssl-dev \
# sf dependencies
lbzip2 \
libfftw3-dev \
libgdal-dev \
libgeos-dev \
libgs10-dev \
libgl1-mesa-dev \
libglu1-mesa-dev \
libhdf4-alt-dev \
```

docker image build



REPOSITORY	TAG	IMAGE ID	CREATED	SIZE
mdancho/h2o-verse	3.30.0.1	65d265478042	5 hours ago	4.7GB
h2o-verse	3.30.0.1	52a24e2dbbc4	16 hours ago	3.75GB
h2overse	3.30.0.1	52a24e2dbbc4	16 hours ago	3.75GB
mdancho/h2o-verse	<none>	52a24e2dbbc4	16 hours ago	3.75GB
<none>	<none>	0f3cf8c4bb04	17 hours ago	2.5GB
<none>	<none>	1eedd05b7fe9	18 hours ago	3.17GB
<none>	<none>	9bbb0857a5c	18 hours ago	3.17GB
<none>	<none>	406952c8ae4	19 hours ago	2.59GB
bsu-shiny-models	latest	2103a3a85290	5 months ago	4.59GB
mdancho/bsu-shiny-models	latest	2103a3a85290	5 months ago	4.59GB

**h2o-verse has a
1.5 hour build time.**



Docker Hub

**h2o-verse has a
1.5 hour build time.**

```
(base) matthews-mbp-2:lab_40_docker mdancho$ docker image ls
```

REPOSITORY	TAG	IMAGE ID	CREATED	SIZE
mdancho/h2o-verse	3.30.0.1	65d265478042	5 hours ago	4.7GB
h2o-verse	3.30.0.1	52a24e2dbbc4	16 hours ago	3.75GB
h2overse	3.30.0.1	52a24e2dbbc4	16 hours ago	3.75GB
mdancho/h2o-verse	<none>	52a24e2dbbc4	16 hours ago	3.75GB
<none>	<none>	0f3cf8c4bb04	17 hours ago	2.5GB
<none>	<none>	1e0dd05b7fe9	18 hours ago	3.17GB
<none>	<none>	9bbb0857a5c	18 hours ago	3.17GB
<none>	<none>	1e0dd05b7fe9	18 hours ago	2.5GB
bsu-shiny-models	latest	b3aee4	18 hours ago	2.5GB
mdancho/bsu-shiny-models	latest	b35290	18 hours ago	2.5GB



docker push image

The screenshot shows a Docker Hub repository page for the user 'mdancho' and the repository 'h2o-verse'. At the top, there is a green curved arrow pointing from the text 'docker push image' to the 'Docker commands' section. The 'Docker commands' section contains the command 'docker push mdancho/h2o-verse:tagname'. Below this, the 'Tags' section lists the tag '3.30.0.1' with a creation timestamp of '5 hours ago'. The 'Recent builds' section is empty. At the bottom, there is a link to a 'Docker for Data Science (Lab 40)' README.

Source: <https://hub.docker.com/repository/docker/mdancho/h2o-verse>



Docker Image

docker pull mdancho/h2o-verse:3.30.0.1

The screenshot shows the Docker Hub interface for the repository 'mdancho/h2o-verse'. It includes sections for General, Tags, Builds, Timeline, Collaborators, Webhooks, and Settings. Under the General tab, there's a 'Docker commands' section with a button to 'Push image'. Below it, there's a 'Recent builds' section and a 'Readme' section. The Tags section shows one tag: '3.30.0.1'. The Builds section shows a build from 5 hours ago.

The screenshot shows the AWS Management Console with the EC2 Instances page open. An instance named 'lab_40_docker' is selected. The instance details pane shows the following information:

Attribute	Value
Instance ID	i-0c04ad64887ea3eb7
Instance State	running
Instance type	t2.micro
Private DNS	ip-172-31-33-148.us-east-2.compute.internal
Private IP	172.31.33.148
VPC ID	vpc-0747d481d
Subnet ID	subnet-276e96a
Network interfaces	eth0
IAM role	-
Key pair name	shiny_business_science
Owner	237807202019
Launch time	August 5, 2020 at 10:41:13 AM UTC-4 (less than one hour)
Termination protection	False
Lifecycle	normal
Monitoring	basic
Alarm status	None
Kernel ID	-

Source: <https://hub.docker.com/repository/docker/mdancho/h2o-verse>

Docker Container



docker container run

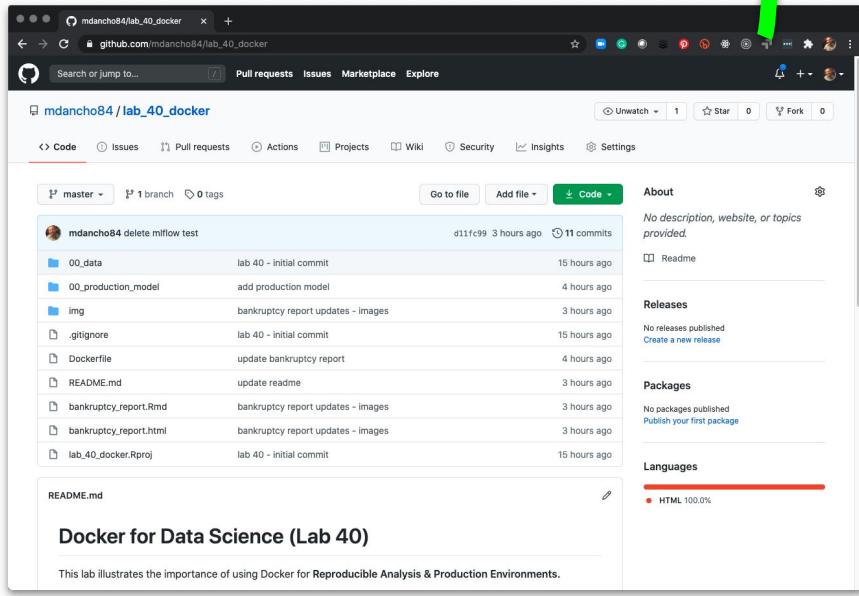
The screenshot shows the AWS EC2 Dashboard with the instance `lab_40_docker` selected. The instance details are as follows:

- Description: `i-0c04ad5d4887ea3eb7`
- Status Checks: 2/2 checks
- Instance state: running
- Instance type: t2.micro
- Public DNS (IPv4): `ec2-13-59-12-104.us-east-2.compute.amazonaws.com`
- IPv4 Public IP: `13.59.12.104`
- Elastic IPs: -
- Availability zone: us-east-2c
- Security groups: `launch-wizard-8`, `shiny_business_science`
- Scheduled events: No scheduled events
- AMI ID: `ami-0f00000000000000`
- Platform details: Linux/Windows
- Usage operation: `RunInstances`
- Source/dest. check: `T2/T3 Unintended`
- Root device type: ebs
- Root device: `/dev/sda1`
- Block devices: -
- EBS-optimized: False
- Kernel ID: -
- Elastic Graphics ID: -
- Elastic Inference accelerator ID: -
- Capacity Reservation: -

The screenshot shows a Docker container running an RStudio session. The application is a shiny web application titled "Business Science". The container's environment includes the `h2o-verse` package, which is visible in the Global Environment pane. The application interface includes a sidebar with "File", "Edit", "Code", "View", "Plots", "Session", "Build", "Debug", "Profile", and "Tools" tabs, and a main area with "Console" and "Terminal" tabs.

Source: <https://hub.docker.com/repository/docker/mdancho/h2o-verse>

GitHub

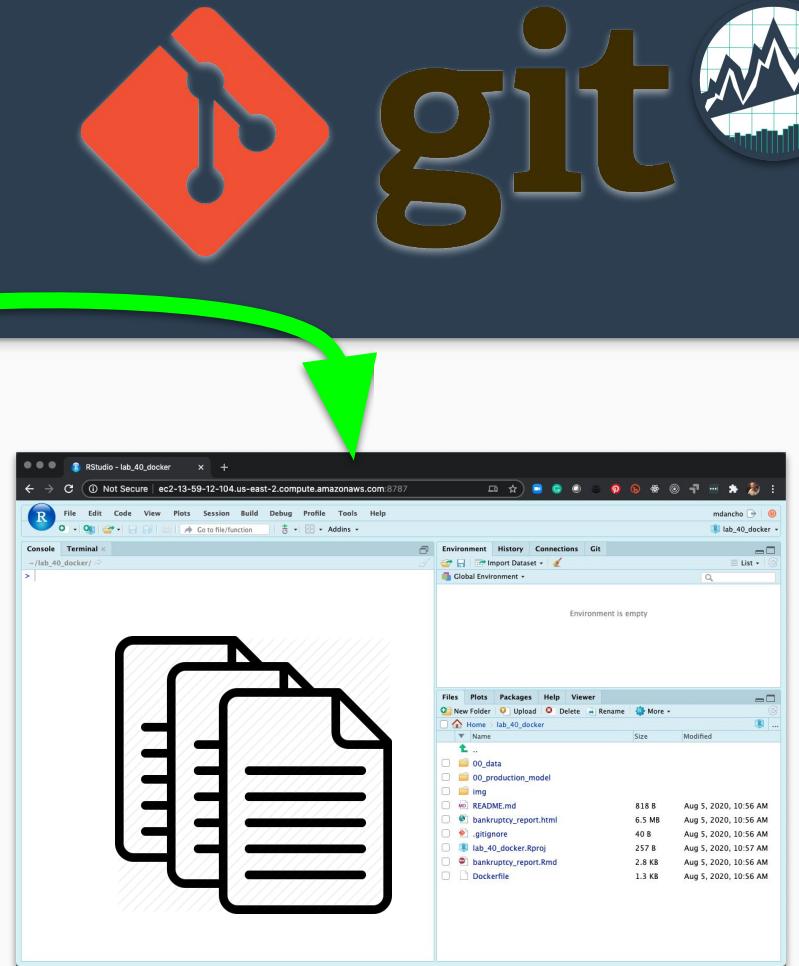


GitHub repository page for `mdancho84/lab_40_docker`. The page displays the following information:

- Code**: master branch, 1 branch, 0 tags.
- Commits**:
 - `mdancho84 delete mlflow test` (d11fc99, 3 hours ago, 11 commits)
 - `00_data` (lab 40 - initial commit, 15 hours ago)
 - `00_production_model` (add production model, 4 hours ago)
 - `img` (bankruptcy report updates - images, 3 hours ago)
 - `.gitignore` (lab 40 - initial commit, 15 hours ago)
 - `Dockerfile` (update bankruptcy report, 4 hours ago)
 - `README.md` (update readme, 3 hours ago)
 - `bankruptcy_report.Rmd` (bankruptcy report updates - images, 3 hours ago)
 - `bankruptcy_report.html` (bankruptcy report updates - images, 3 hours ago)
 - `lab_40_docker.Rproj` (lab 40 - initial commit, 15 hours ago)
- About**: No description, website, or topics provided.
- Releases**: No releases published. Create a new release.
- Packages**: No packages published. Publish your first package.
- Languages**: HTML 100.0%

Docker for Data Science (Lab 40)

This lab illustrates the importance of using Docker for Reproducible Analysis & Production Environments.



RStudio session titled `ec2-13-59-12-104.us-east-2.compute.amazonaws.com:8787`.

Console (Terminal):
~/lab_40_docker/

Global Environment: Environment is empty

Files (File Browser):

- Home / lab_40_docker
 - ..
 - 00_data
 - 00_production_model
 - img
 - ... (redacted)
 - README.md
 - bankruptcy_report.html
 - .gitignore
 - lab_40_docker.Rproj
 - bankruptcy_report.Rmd
 - Dockerfile

Environment: Global Environment

Source: https://github.com/mdancho84/lab_40_docker

30-Min Demo

lab_40_docker - master - RStudio

File Insert Run Environment History Connections Git Tutorial

00_lab_40_setup.R Dockerfile 01_docker_cli_scripts README.md bankruptcy_report.Rmd

131 add_markers(opacity = 0.5)

132

133

134

5:11 Lab 40: Docker for Data Science R Markdown

~/Desktop/learning_labs/lab_40_docker/

```
> # Plotly Visualization ----  
> plot_data_tbl %>%  
+   plot_ly(x = ~ umap_1, y = ~ umap_2, z = ~ umap_3,  
+           color = ~ class, colors = c('#BF382A', '#0C4B8E'),  
+           hovertemplate = ~ tooltip) %>%  
+   add_markers(opacity = 0.5)  
>
```

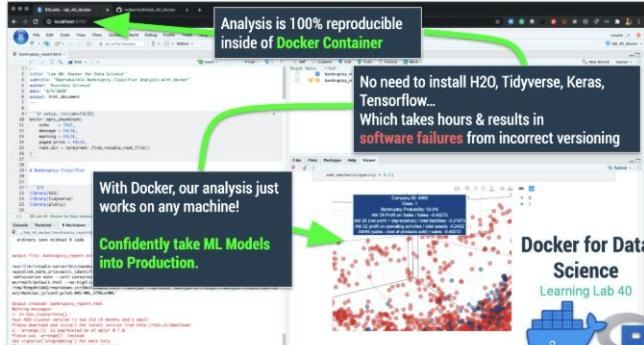
Reproducible Bankruptcy Cl... Libraries Data Bankruptcy Classification F... Bankruptcy Classification F... H2O Prediction Analysis Load Production Model Make Predictions Bankruptcy UMAP Visualiz... Apply UMAP Plotly Visualization

Lab 40: Docker for Data Science

Reproducible Bankruptcy Classifier Analysis with Docker

Business Science
8/5/2020

Reproducible Bankruptcy Classifier Analysis



Analysis is 100% reproducible inside of Docker Container

No need to install H2O, Tidyverse, Keras, Tensorflow... Which takes hours & results in software failures from incorrect versioning

With Docker, our analysis just works on any machine!

Confidently take ML Models into Production.

Docker for Data Science Learning Lab 40

Libraries

```
library(h2o)  
library(tidyverse)  
library(plotly)
```

Not in container

Your transformation

- Build & deploy apps with AWS & Docker



Shiny vs Tableau

5-Year Trend going on.

- Docker +4000%
- AWS +400%

3 Business Examples

1. Application Library
2. Full Stack Web App
3. ML-Powered Dashboard

The screenshot shows a web browser window with the URL business-science.io/business/2020/03/09/shiny-vs-tableau.html. The page title is "Part 6 - R Shiny vs Tableau (3 Business Application Examples)". It was written by Matt Dancho on March 9, 2020. The main content features a section titled "R Shiny vs Tableau" with the subtitle "3 Shiny Business Application Examples". Below this, there is a screenshot of a "Stock Analyzer" application built with R Shiny. The application interface includes a sidebar with "Favorites" (GOOGL, AMZN) and a main area showing a line chart for PEAK with three moving average lines (short, medium, long). The sidebar also has sections for "Stock List (Click Once to Analyze)" and "Last Analysis". To the right of the main content, there is a sidebar with "Get Articles in Your Inbox", "Search for Articles", "Find Articles By Category" (with categories like Learn R, Finance, Marketing, Machine Learning, Data Science), and "Download Cheat Sheets". There are also social sharing icons for LinkedIn, Facebook, Twitter, and Email, and a "41 Shares" counter.

Source: <https://www.business-science.io/business/2020/03/09/shiny-vs-tableau.html>

1. Application Library

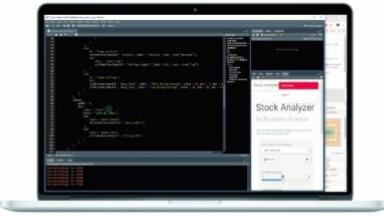


Apps by Business Science x +
apps.business-science.io

Apps by Business Science App Gallery Search Apps Submit Clear

Business Science University.

Build Applications. Generate Business Value. [Enroll Now.](#)



Featured Application

Stock Analyzer



Application Highlights

- Multi-User Application secured with Authentication & HTTPS Encryption
- Hosted on AWS
- Powered by MongoDB and Bootstrap

[Try the Stock Analyzer](#) [Build It in DS4B 202A-R](#)

ALL BUSINESS FINANCE HUMAN RESOURCES MARKETING SALES



[Business](#) [Shiny](#) [AWS](#)


[Finance](#) [Shiny](#) [AWS](#) [MongoDB](#) [Auth](#)


[Human Resources](#) [Shiny](#) [H2O](#) [Bootstrap](#)


[Application Library](#)

2. Full-Stack Web App



Stock Analyzer by Business Science

This is the multi-user application completed in our [Shiny Developer with AWS Course \(DS4B 202-R\)](#)

Favorites



Show/Hide Clear Favorites

Stock List (Pick One to Analyze)

PEAK, Healthpeak Properties Inc.

Analyze

Short Moving Average (Days)

40

Long Moving Average (Days)

120

Time Window (Days)

366

Apply & Save

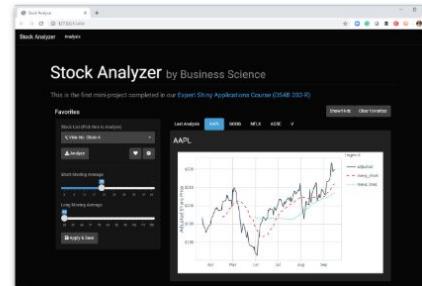
Last Analysis GOOGL AMZN

PEAK



Application Architecture

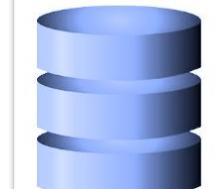
Course DS4B 202A-R
For Data Scientists & Programmers



Tidyquant API



User Data

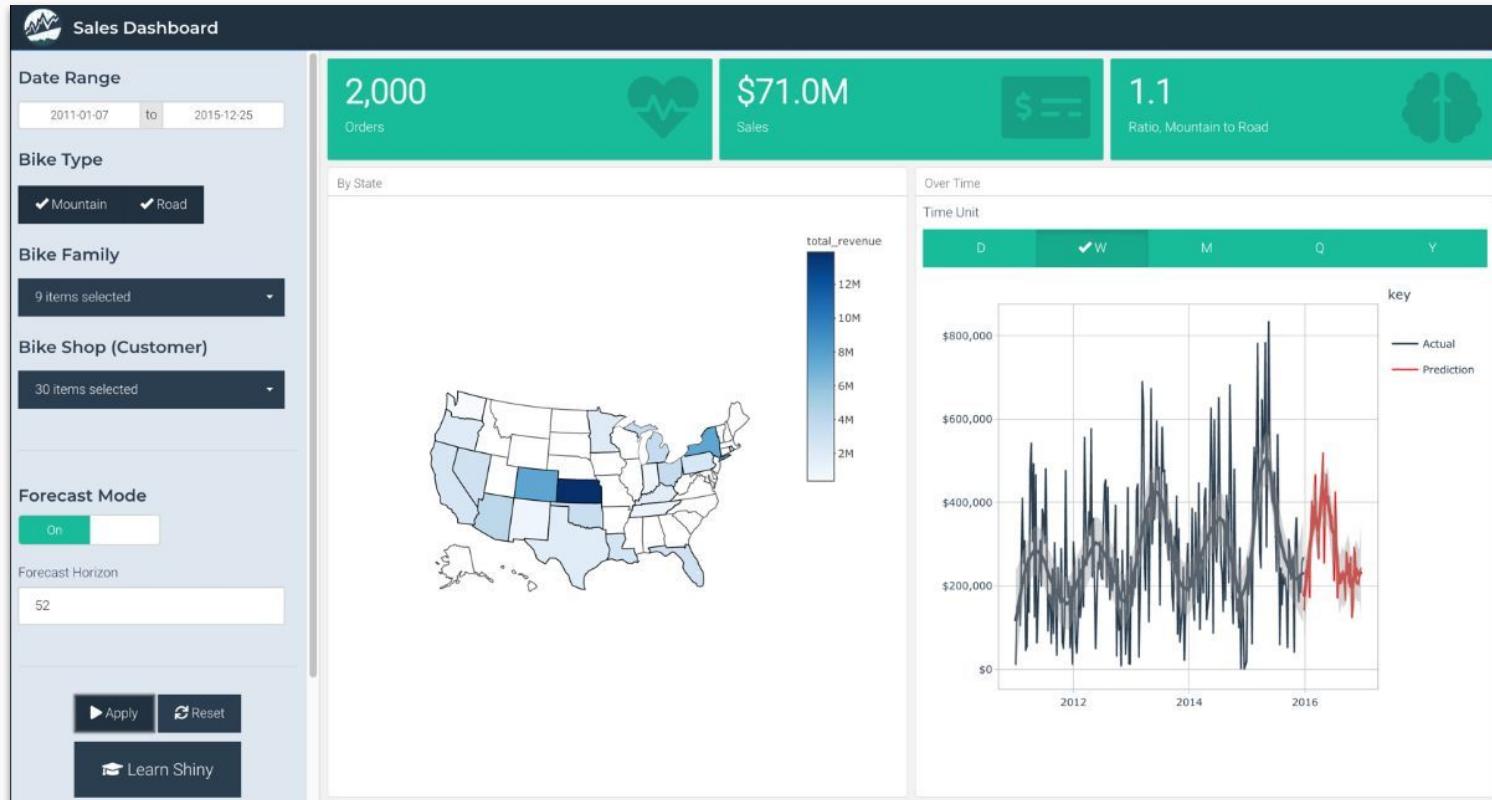


mongoDB Atlas

[Stock Analyzer](#)



3. ML-Powered Dashboard



[ML-Powered Dashboard](#)

4-Course R-Track System



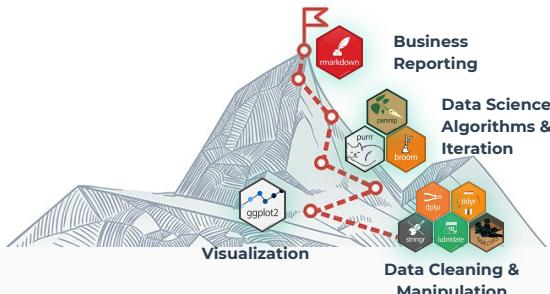
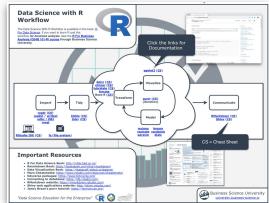
Business Analysis with R (DS4B 101-R)

Data Science For Business with R (DS4B 201-R)

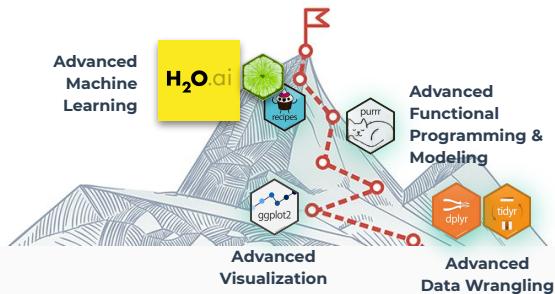
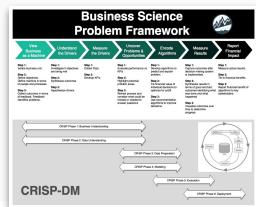
Web Apps & Shiny Developer (DS4B 102-R + DS4B 202A-R)

Project-Based Courses with Business Application

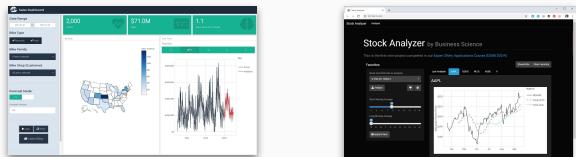
Data Science Foundations
7 Weeks



Machine Learning & Business Consulting
10 Weeks



Web Application Development
12 Weeks



Key Benefits

Frontend + Backend + Production Deployment

Frontend for Shiny

- Bootstrap

Backend for Shiny

- MongoDB Atlas Cloud
- Dynamic UI
- User Authentication
- Security

Production Deployment

- AWS
- EC2 Server
- SSL & HTTPS Encryption

Shiny Apps for Business (DS4B 202A-R)



Web Application Development
6 Weeks



DS4B 202A-R: Expert Shiny Developer with AWS

Learn how to build Scalable Data Science Applications using R, Shiny, and AWS Cloud Technology.

Matt Dancho

15% OFF PROMO Code: **learninglabs**

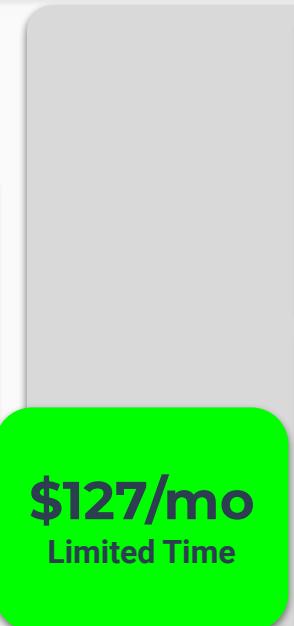


R-TRACK BUNDLE

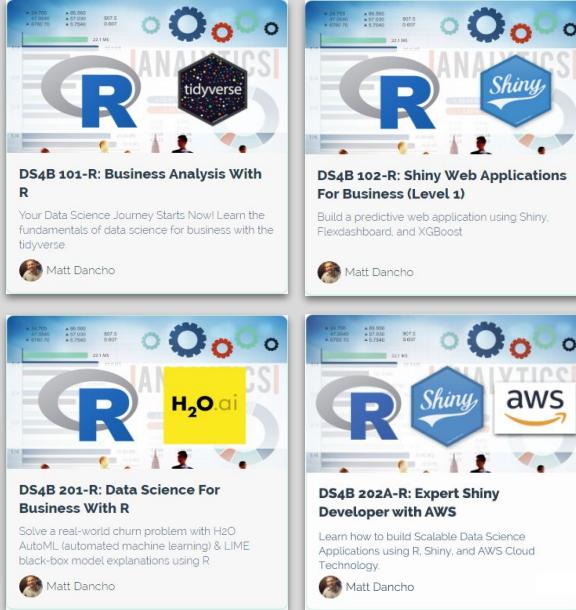
4-Course Bundle - Machine Learning + Expert Web Applications (R-Track)

Go from Beginner to Expert Data Scientist & Shiny Developer in Under 6-Months

4 Course Bundle ~~\$1,500~~



<input type="radio"/>	Paid Course 15% COUPON DISCOUNT	\$1,596 \$2,356.60
<input checked="" type="radio"/>	12 Low Monthly Payments 15% COUPON DISCOUNT 12X Payment Plan	12 payments of \$149/m 12 payments of \$126.65/m



DS4B 101-R: Business Analysis With R
Your Data Science Journey Starts Now! Learn the fundamentals of data science for business with the tidyverse.

DS4B 102-R: Shiny Web Applications For Business (Level 1)
Build a predictive web application using Shiny, Flexdashboard, and XGBoost.

DS4B 201-R: Data Science For Business With R
Solve a real-world churn problem with H2O AutoML (automated machine learning) & LIME black-box model explanations using R.

DS4B 202A-R: Expert Shiny Developer with AWS
Learn how to build Scalable Data Science Applications using R, Shiny, and AWS Cloud Technology.

Career acceleration awaits

university.business-science.io

