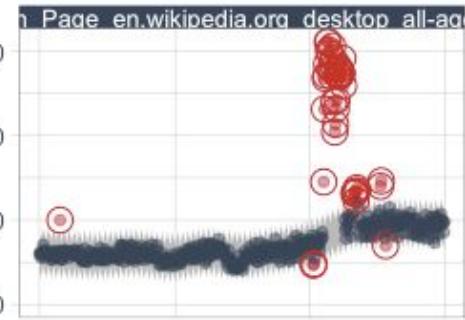


## Page Visits



# Anomaly Detection

For Time Series

Difficulty: **Intermediate**



Matt Dancho & David Curry  
**Business Science Learning Lab**





# Learning Lab Structure

- **Presentation**  
(30 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**.

# Success Story

## Masatake Hirono

- Took 201
- Completed the 10-Week Course
- **Landed a Job at one of the most Prestigious Management Consulting Firms**



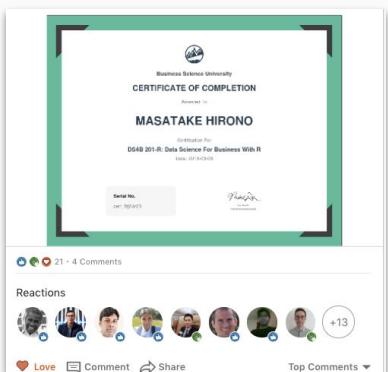
***"This course showed me how to place data analytics in real business settings."***



Masatake Hirono • 1st  
Data Scientist at 株式会社進研アド  
4d

After struggling to balance with my work for many months, I've finally completed the Business Science University DS4B 201-R: Data Science For Business With R, taught by [Matt Dancho](#). Unlike other MOOCs, this course showed me how to place data analytics in real business settings. Without this course, I would have never attempted to pay attention to business/financial impacts, generated through my analysis. His instruction turned me a more advanced data scientist and helped me find a new career opportunity. I will start to work at one of the most prestigious management consulting firms in October as a cognitive & analytics consultant. Highly recommended if you would like to use R as a professional business person!

#business\_science\_success #dataanalysis #machinelearningtraining



2d ...

How did it help you find another career opportunity? Did he place you in touch with hiring firms?

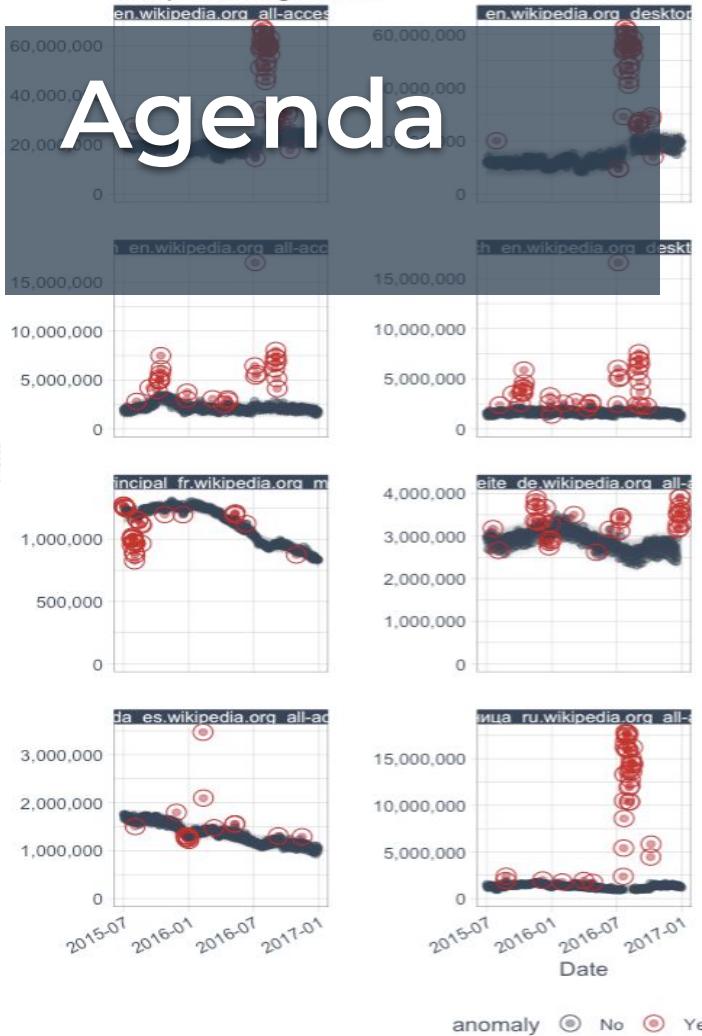
1 Reply

Masatake Hirono • 1st  
Data Scientist at 株式会社進研アド  
1d ...

Of course not. In a job interview, I was able to draw interviewer's attention because of my experiences to formulate some insight from analytics for driving business, which I had developed through his course.

**#Business  
Science  
Success**

# Agenda



- **Business Case Study**
  - Google Analytics
  - Web Traffic
  - Anomalies
- **30-Min Demo**
  - Web traffic
  - data.table
  - anomalize
- **Anomalies**
  - 3 Types
  - Algorithms
- **Pro-Tips:**
  - Tactics to Improve Forecasts
- **Time Series Anomaly Detection**
  - Anomalize Software
  - 80/20 Concepts
  - 3 Key Functions
  - Game Plan



# Learning Labs PRO

Every 2-Weeks

1-Hour Course

Recordings + Code + Slack

**\$19/month**

*university.business-science.io*

*Lab 17*  
**Anomaly Detection with H2O  
Machine Learning**

*Lab 16*  
**R's Optimization Toolchain, Part 2  
- Nonlinear Programming**

*Lab 15*  
**R's Optimization Toolchain, Part 1  
- Linear Programming**

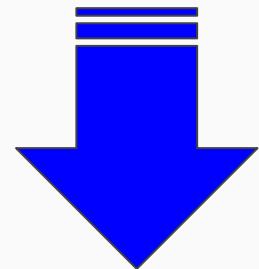
*Lab 14*  
***Customer Churn Survival Analysis***

*Lab 13*  
***Wrangling 4.6M Rows of  
Financial Data w/ data.table***

*Lab 12*  
***How I built anomalize***



**Continuous Learning**  
Jet Fuel for your Brain



**Learning Labs Pro**

Community-Driven Data Science Courses

 Matt Dancho

**\$19/m**

# Anomaly Detection

## Business Case



# Anomalous Web Traffic

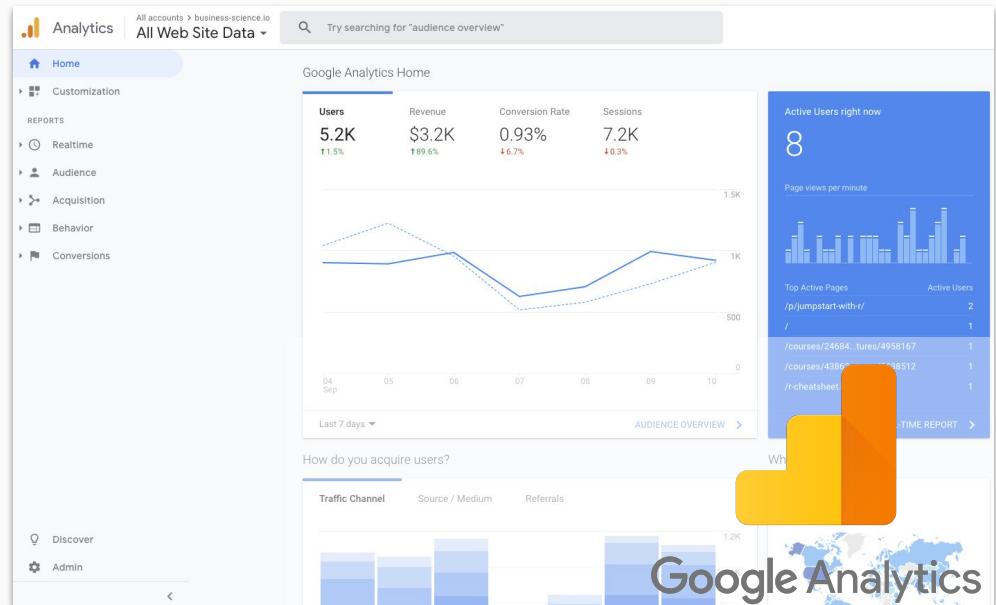
## Revenue & Web Traffic

Consumers **spend billions** online every year

Web-based companies can use web-traffic to **forecast cashflow**

Anomalies:

- Can flag important **events**
- Can impact the **forecast accuracy**

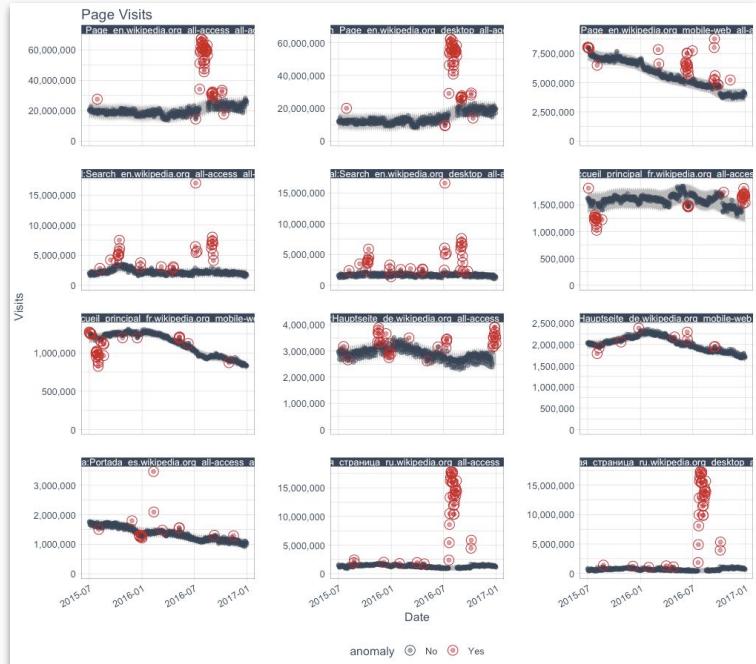




# Anomalous Web Traffic

## Key Issues

1. **1000's** of web pages
2. Scaling the data preparation (**cleaning of anomalies**) for forecasting
3. **Linking events** to anomalous data



# What are Anomalies?

Recap from Lab 17

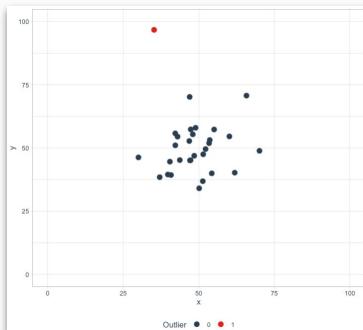


# Types of Anomalies

1

## Point Anomalies

### Single Point

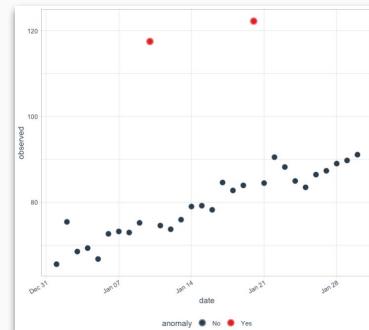


H2O Isolation Forest  
H2O K-Means

2

## Contextual

### Time Series

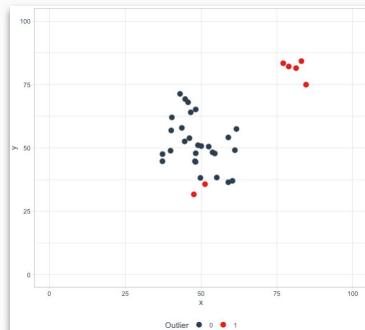


???

3

## Collective

### Cluster of Points



H2O Isolation Forest  
H2O K-Means

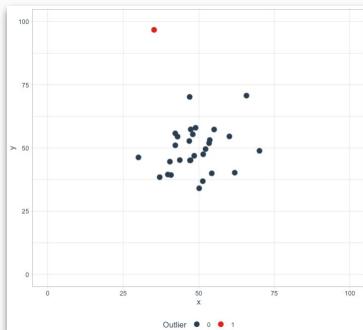


# Types of Anomalies

1

## Point Anomalies

### Single Point

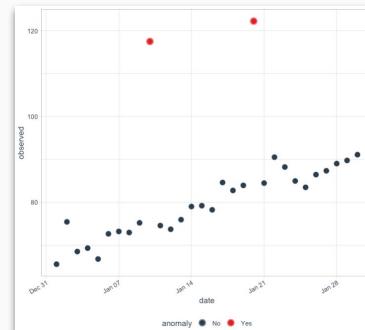


H2O Isolation Forest  
H2O K-Means

2

## Contextual

### Time Series

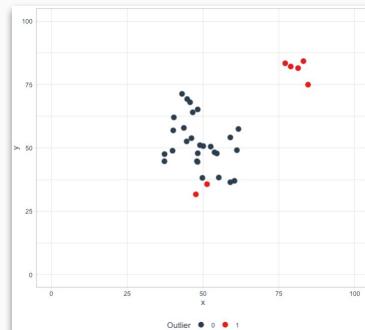


???

3

## Collective

### Cluster of Points



H2O Isolation Forest  
H2O K-Means

# Time Series Anomaly Detection Software

# Time Series Anomaly Detection



**Anomalize** enables a simple workflow for **scalable** anomaly detection for time series



Screenshot of the [anomalize](https://business-science.github.io/anomalize/) project page on GitHub:

The page shows the following details:

- anomalize 0.1.1**: Version information.
- Home**, **Function Reference**, **Vignettes**, **News**: Navigation links.
- anomalize**: The main title.
- Tidy anomaly detection**: A brief description.
- anomalize** enables a tidy workflow for detecting anomalies in data. The main functions are `time_decompose()`, `anomalize()`, and `time_recompose()`. When combined, it's quite simple to decompose time series, detect anomalies, and create bands separating the "normal" data from the anomalous data.
- Anomalous In 2 Minutes (YouTube)**: A link to a video thumbnail showing multiple time series plots with anomalies highlighted.
- Links**:
  - Download from CRAN at <https://cloud.r-project.org/package=anomalize>
  - Browse source code at <https://github.com/business-science/anomalize>
  - Report a bug at <https://github.com/business-science/anomalize/issues>
- License**: GPL (>= 3)
- Developers**:
  - Matt Dancho: Author, maintainer
  - Davis Vaughan: Author

# Anomalize

## 80/20 Concepts & Important Operations



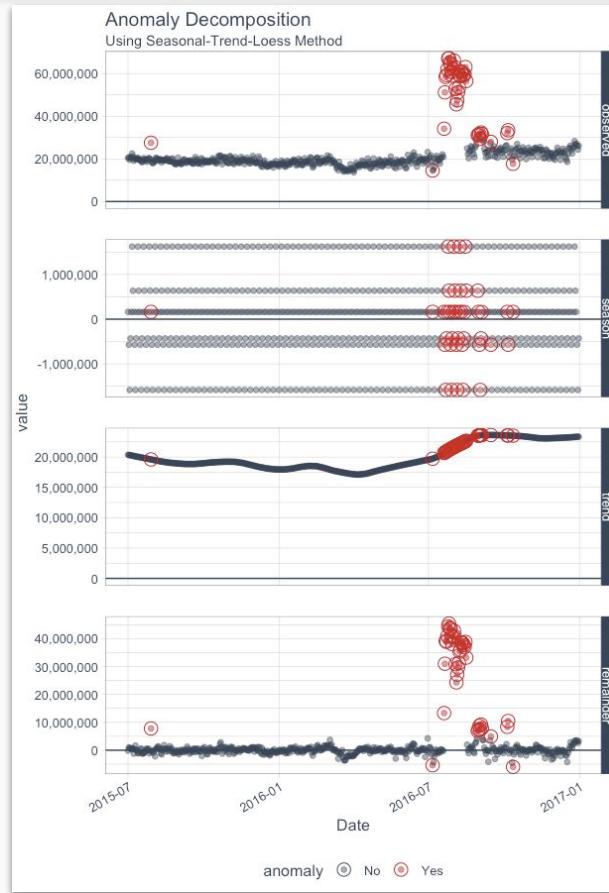
# 1. STL Method

## Algorithm Internal Process

- Uses **STL Decomposition** to decompose time series into *seasonal*, *trend* & *remainder*
- The key is the **remainders** (**residuals**)
- Uses **IQR** or **GESD** to detect anomalies

### Key Concept

Outliers have **abnormal residuals (remainders)**



Observed

Seasonal Component

Trend Component

Remaining Component (remainder)



## 2. Twitter Method

### Algorithm Internal Process

- Uses **Piecewise Medians Decomposition** to decompose time series into *seasonal, trend & remainder*
- The key is the **remainders (residuals)**
- Uses **IQR or GESD** to detect anomalies

### Key Concept

Only difference is using Piecewise Medians vs LOESS Trend



# Implementation



## 3-Step Process:

### 1. `time_decompose()`

Uses **STL or Twitter** to decompose time series into *seasonal, trend & remainder*

### 2. `anomalize()`

Uses **IQR or GESD** to detect anomalies

### 3. `time_recompose()`

Calculates outlier boundaries

```
143 wikipedia_main_page_tbl %>%
144   # Step 1 - STL Decomposition
145   time_decompose(
146     target = Visits,
147     method = "stl", # stl or twitter
148     merge  = TRUE,
149     frequency = "1 week",
150     trend    = "3 months"
151   ) %>%
152   # Step 2 - Detect Anomalies in Remainder (Residual Analysis)
153   anomalize(
154     target = remainder,
155     method = "iqr", # iqr or gesd
156     alpha  = 0.05
157   ) %>%
158   # Step 3 - Add Boundaries separating the anomaly lower and upper limits
159   time_recompose() %>%
160 
```

# Gameplan

## Workflow & Tools



# Web Forecasting

## Workflow



**data.table & ggplot2**

Exploratory Data Analysis

**anomalize**

Anomaly Detection

Data Cleaning

**parsnip, purrr &  
ggplot2**

Forecast Web Traffic

# 30-Min Demo

## Web Traffic Anomaly Detection & Forecasting

Secret Tactics for

# Forecasting with Anomalies

Use these tips to  
**increase your forecasting performance**



# Pro Tip

## Clean Your Anomalies

### Option 1

#### Flag Anomalies

Just add the “Anomaly (Y/N)” as a Flag in your model

##### Pro

Predicts well when future has anomalies that are **similar to past anomalies**

##### Con

May reduce forecasting accuracy

### Option 2

#### Clean Anomalies

Replace Anomaly Values with Trend + Seasonal Components

##### Pro

Improves Forecasting Performance

##### Con

Doesn't predict well when future has anomalies



## Option 1

## Flag Anomalies

Just add the “Anomaly (Y/N)” as a Flag in your model

```
> mape_flagged_anoms  
# A tibble: 1 x 1  
`mean(mape)`  
      <dbl>  
1        0.284
```

## Option 2

## Clean Anomalies

Replace Anomaly Values with Trend + Seasonal Components

```
> mape_cleaned_anoms  
# A tibble: 1 x 1  
`mean(mape)`  
      <dbl>  
1        0.139
```

51%

Improvement

# Data Science Transformational

Skills that are needed to do what we just did



# Web Forecasting

## Step-By-Step



### **data.table & ggplot2**

Exploratory Data Analysis

**101 & Lab 13**

### **anomalize**

Anomaly Detection

Data Cleaning

**Lab 18**

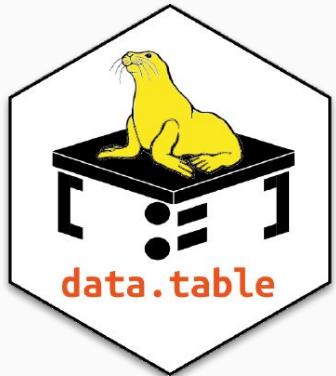
### **parsnip, purrr & ggplot2**

Forecast Web Traffic

**101 & 201**



# data.table - wrangling 80M rows

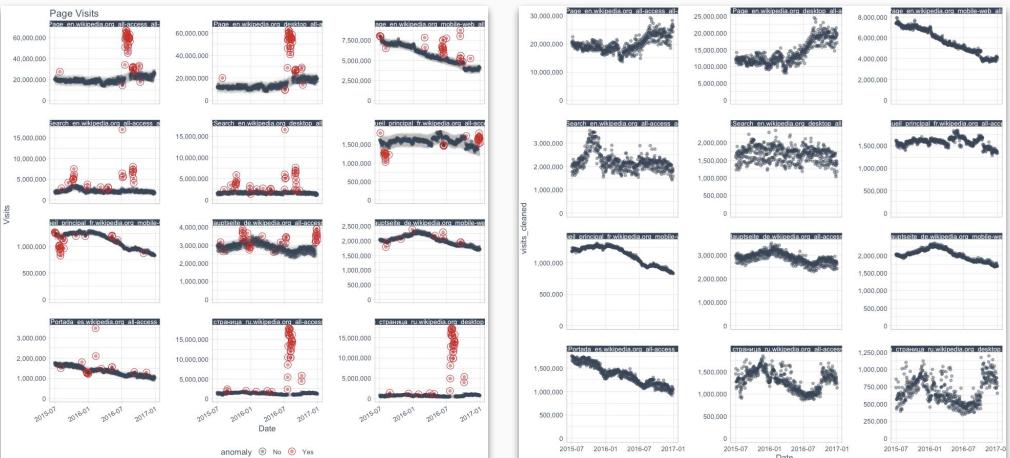
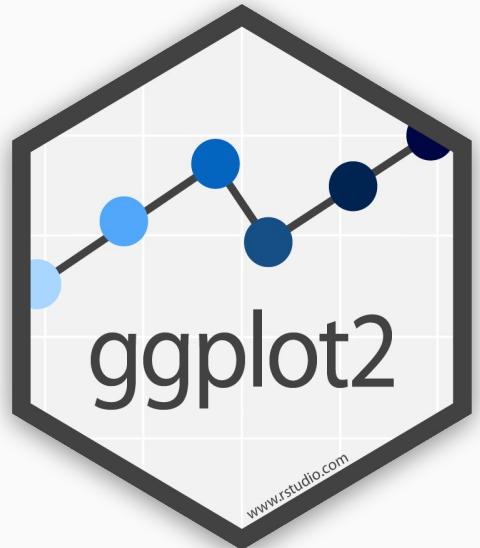


Pro-Tip:  
Learn *dplyr* first



```
26 # 3.2 Setup in Long Format ----  
27 page_visits_dt <- melt(  
28   page_visits_dt,  
29   id.vars     = c("Page"),  
30   measure.vars = setdiff(names(page_visits_dt), "Page")  
31 )  
32  
33 names(page_visits_dt) <- c("Page", "Date", "Visits")  
34 # page_visits_dt$date <- anytime::anydate(page_visits_dt$date)  
35 setkey(page_visits_dt, "Page")  
36  
37 page_visits_dt %>% glimpse()  
38  
39  
40 # 3.3 Prep & Functions ----  
41 page_visit_counts_dt <- page_visits_dt %>%  
42   .[, .(visits_sum    = sum(Visits, na.rm = TRUE),  
43         visits_mean   = mean(Visits, na.rm = TRUE),  
44         visits_median = median(Visits, na.rm = TRUE),  
45         visits_count  = sum(Visits > 0)),  
46         keyby = .(Page)] %>%  
47   .[, ratio := visits_mean / (visits_median + 1)]  
48  
49 page_visit_counts_dt %>% glimpse()
```

101 & 201  
Lab 13



```

246 # AFTER CLEANING ----
247 pages_most_visited_anom_tbl %>%
248   ggplot(aes(Date, visits_cleaned)) +
249   geom_point(alpha = 0.5, color = palette_light()[[1]]) +
250   facet_wrap(~ Page, ncol = 3, scales = "free_y") +
251   expand_limits(y = 0) +
252   scale_y_continuous(labels = scales::comma) +
253   theme_tq()
254

```

101 & 201



# ggplot2 & purrr

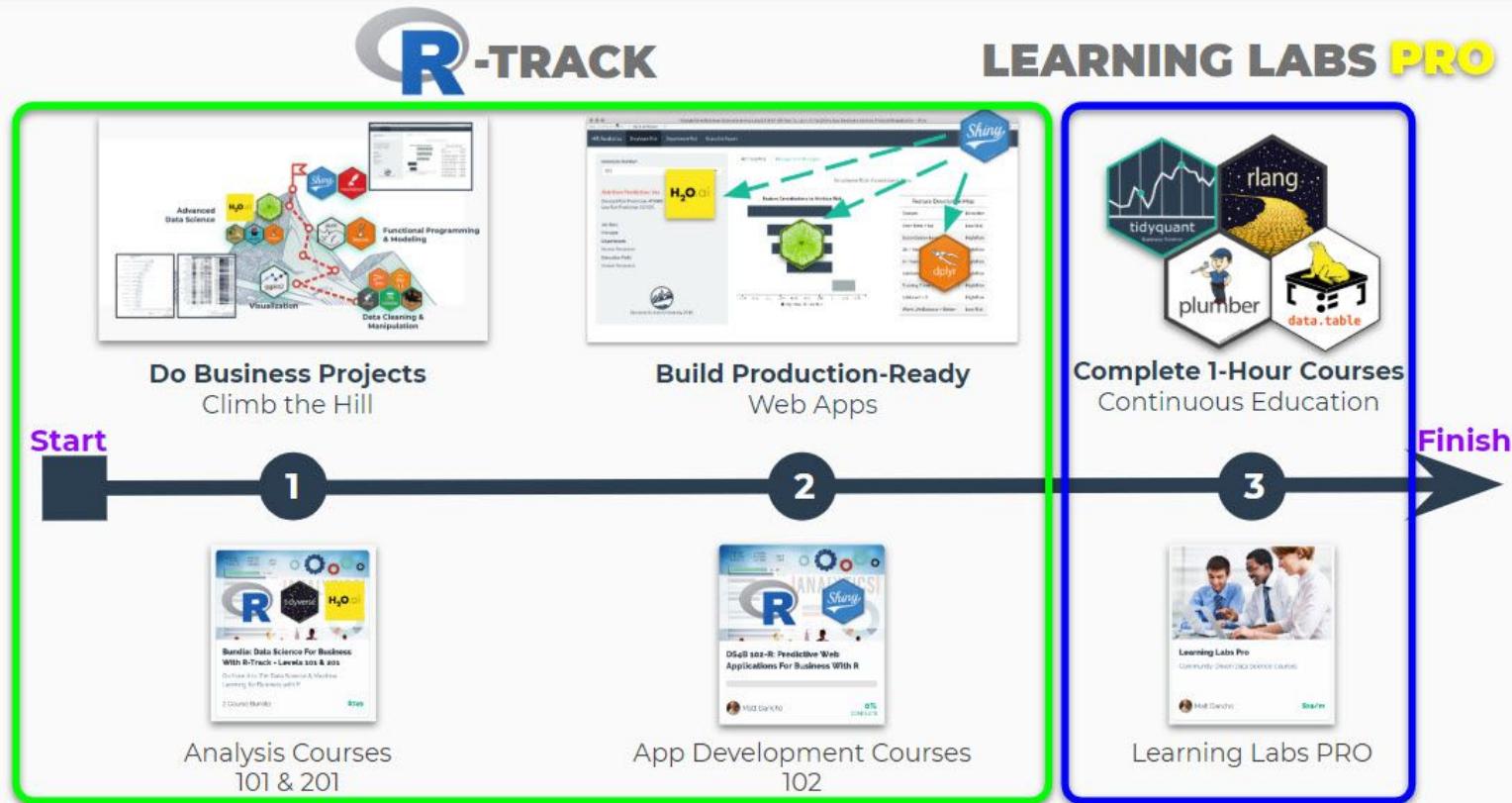


```
21     })) %>%
22
23     # Apply model to training set
24     mutate(model = map(split, .f = function(split) {
25         trn_tbl <- split %>%
26             training() %>%
27                 timetk::tk_augment_timeseries_signature() %>%
28                     select(-diff)
29
30         model <- linear_reg(mixture = 0.5) %>%
31             set_engine("glmnet") %>%
32                 fit.model_spec(visits_cleaned ~ . - Date, data = trn_tbl)
33
34         return(model)
35     })) %>%
```

# **Business Science University**

Our program that will TRANSFORM YOU in weeks, not years.

# The program that will deliver YOUR Transformation



Everything is **Taken Care of** For You in Our Platform

# 3-Course R-Track System



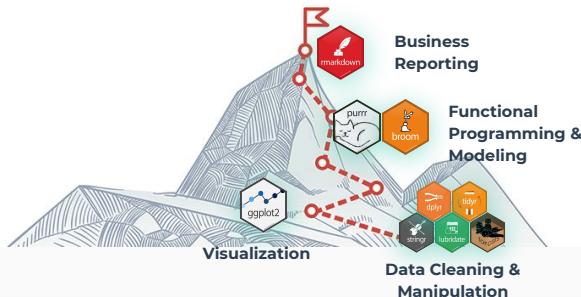
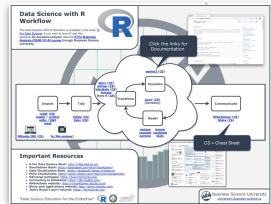
## Business Analysis with R (DS4B 101-R)

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

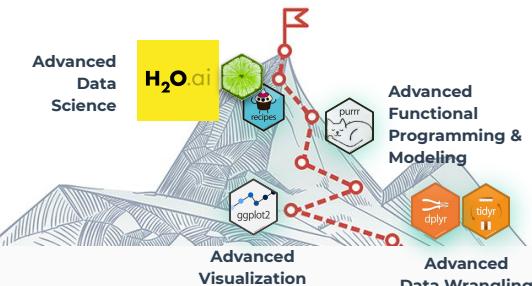
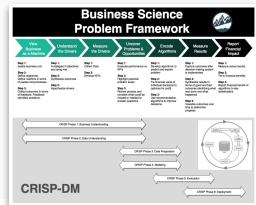
## R Shiny Web Apps For Business (DS4B 102-R)

### Project-Based Courses with Business Application

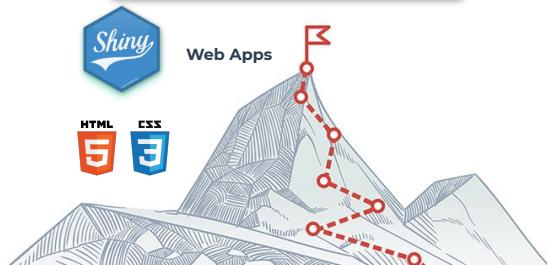
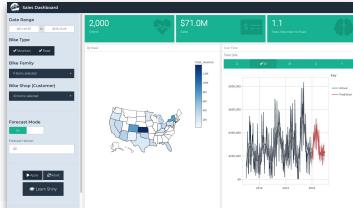
Data Science Foundations  
**7 Weeks**



Machine Learning & Business Consulting  
**10 Weeks**



Web Application Development  
**4 Weeks**

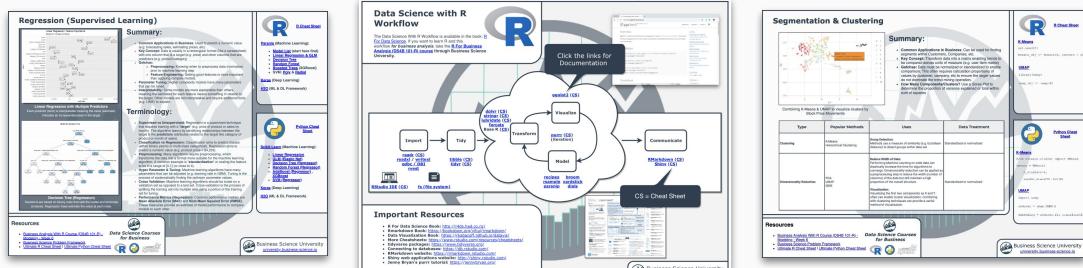


# Key Benefits

- Fundamentals - Weeks 1-5 (25 hours of Video Lessons)
  - Data Manipulation (dplyr)
  - Time series (lubridate)
  - Text (stringr)
  - Categorical (forcats)
  - Visualization (ggplot2)
  - Programming & Iteration (purrr)
  - 3 Challenges
- **Machine Learning - Week 6 (8 hours of Video Lessons)**
  - Clustering (3 hours)
  - Regression (5 hours)
  - 2 Challenges
- Learn Business Reporting - Week 7
  - RMarkdown & plotly
  - 2 Project Reports:
    1. Product Pricing Algo
    2. Customer Segmentation

# Business Analysis with R (DS4B 101-R)

Data Science Foundations  
**7 Weeks**



# Key Benefits

## End-to-End Churn Project

Understanding the Problem & Preparing Data - Weeks 1-4

- Project Setup & Framework
- Business Understanding / Sizing Problem
- Tidy Evaluation - rlang
- EDA - Exploring Data -GGally, skimr
- Data Preparation - recipes
- Correlation Analysis
- 3 Challenges

## Machine Learning - Weeks 5, 6, 7

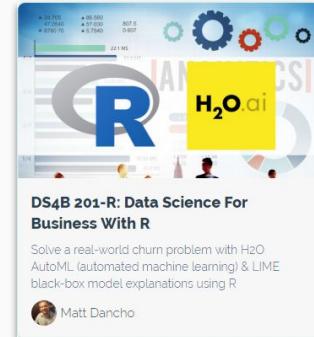
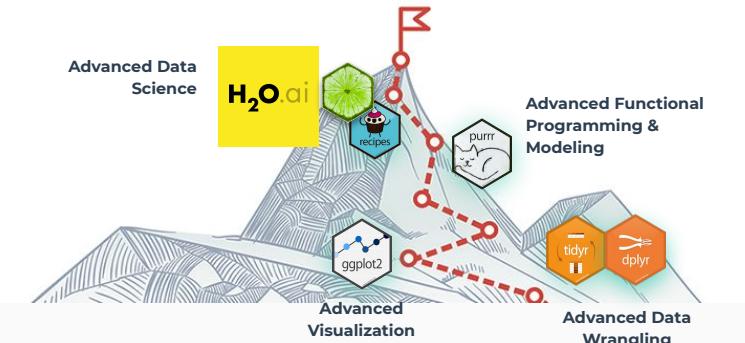
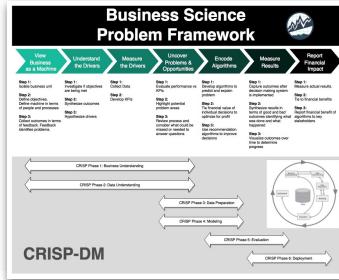
- H2O AutoML - Modeling Churn
- ML Performance
- LIME Feature Explanation

## Return-On-Investment - Weeks 7, 8, 9

- Expected Value Framework
- Threshold Optimization
- Sensitivity Analysis
- Recommendation Algorithm

# Data Science For Business (DS4B 201-R)

Machine Learning & Business Consulting  
**10 Weeks**



# Key Benefits

## Learn Shiny & Flexdashboard

- Build Applications
- Learn Reactive Programming
- Integrate Machine Learning

## App #1: Predictive Pricing App

- Model Product Portfolio
- XGBoost Pricing Prediction
- Generate new products instantly

## App #2: Sales Dashboard with Demand Forecasting

- Model Demand History
- Segment Forecasts by Product & Customer
- XGBoost Time Series Forecast
- Generate new forecasts instantly

# Shiny Apps for Business (DS4B 102-R)



Web Application Development  
**4 Weeks**

The collage includes:

- A "Data Science with R" course page featuring a "Predictive Pricing App" dashboard.
- A "Flexdashboard Apps" section showing a dashboard with a map of the US and time series plots.
- A "Shiny Apps" section showing a dashboard with a scatter plot and a bar chart.
- A "Themes, Dashboards, & Examples" section showing a dashboard with multiple panels and a sidebar.
- A "Business Analytics" section showing a dashboard with a map and a bar chart.
- A "Machine Learning" section showing a dashboard with a scatter plot and a bar chart.
- A "Data Science for Business" section showing a dashboard with a map and a bar chart.



The collage includes:

- A "Predictive Pricing App" dashboard showing a map of the US and time series plots.
- A "Sales Dashboard with Demand Forecasting" dashboard showing a map of the US and time series plots.
- A "Machine Learning" dashboard showing a scatter plot and a bar chart.
- A "Data Science for Business" dashboard showing a map and a bar chart.
- A "Business Analytics" dashboard showing a map and a bar chart.

**DS4B 102-R: Shiny Web Applications for Business (Level 1)**

Build a predictive web application using Shiny, Flexdashboard, and XGBoost.

Matt Dancho

# Success Story

## Masatake Hirono

- Took DS4B 201-R
- Completed the 10-Week Course
- **Landed a Job at one of the most Prestigious Management Consulting Firms**



***"This course showed me how to place data analytics in real business settings."***



Masatake Hirono • 1st  
Data Scientist at 株式会社進研アド  
4d

After struggling to balance with my work for many months, I've finally completed the Business Science University DS4B 201-R: Data Science For Business With R, taught by [Matt Dancho](#). Unlike other MOOCs, this course showed me how to place data analytics in real business settings. Without this course, I would have never attempted to pay attention to business/financial impacts, generated through my analysis. His instruction turned me a more advanced data scientist and helped me find a new career opportunity. I will start to work at one of the most prestigious management consulting firms in October as a cognitive & analytics consultant. Highly recommended if you would like to use R as a professional business person!

#business\_science\_success #dataanalysis #machinelearningtraining

2d ...

How did it help you find another career opportunity? Did he place you in touch with hiring firms?

1 Reply

Masatake Hirono • 1st  
Data Scientist at 株式会社進研アド  
1d ...

Of course not. In a job interview, I was able to draw interviewer's attention because of my experiences to formulate some insight from analytics for driving business, which I had developed through his course.

**#Business  
Science  
Success**

# 15% OFF PROMO Code: learninglabs



## R-TRACK BUNDLE

**Bundle - DS For Business + Web Apps (Level 1): R-Track - Courses 101, 102,**

3 Course Bundle

0% COMPLETE

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

**Get started now!**

	Paid Course 15% COUPON DISCOUNT	\$1,149 \$976.65	<a href="#">Enroll</a>
<input type="radio"/>	3 Monthly Payments 15% COUPON DISCOUNT 3X Monthly	3 payments of \$449/m	3 payments of \$381.65/m
<input type="radio"/>	6 Low Monthly Payments 15% COUPON DISCOUNT 6X Payment Plan	6 payments of \$224.67/m	6 payments of \$198.90/m
<input type="radio"/>	12 Low Monthly Payments 15% COUPON DISCOUNT 12X Plan	12 payments of \$125/m	12 payments of \$106.25/m

# Begin Learning Today

[university.business-science.io](http://university.business-science.io)

