

THE DATA LAKE SUMMIT

The definitive virtual conference for all things Data Lake

OCT 13-14, 2020

Scaling Data Science with Spark and R



Javier Luraschi

Software Engineer
Author of Mastering Spark with R

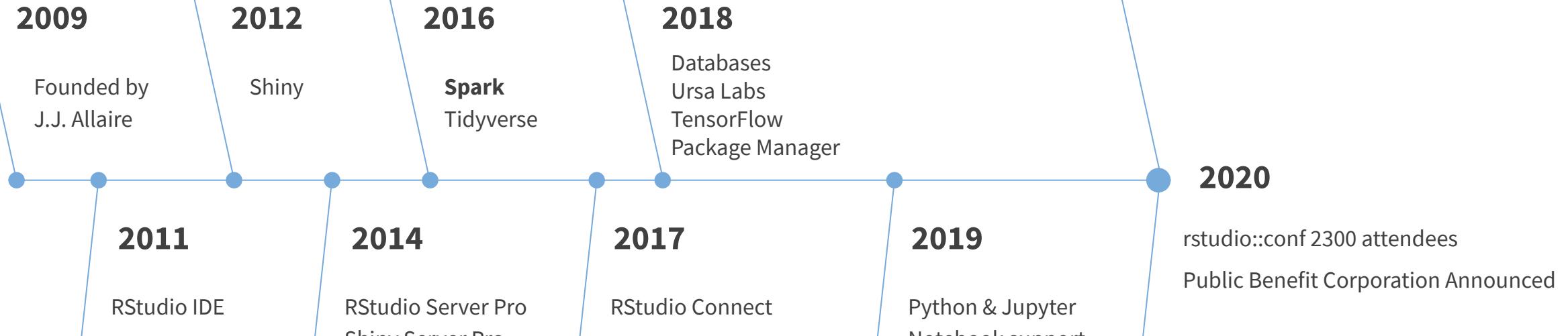
Presented by **Qubole** in collaboration with



and



About RStudio

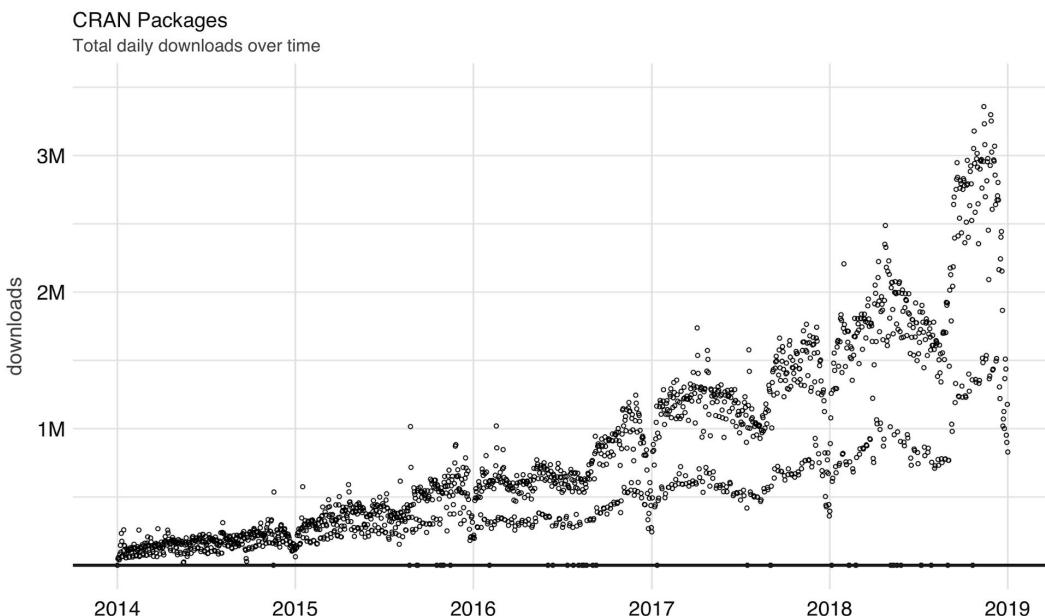


Why R for Data Science?

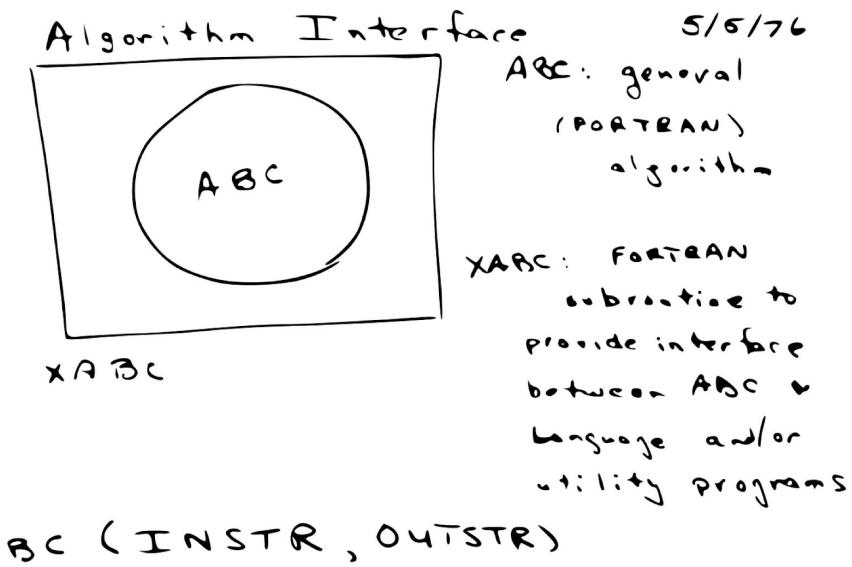


Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

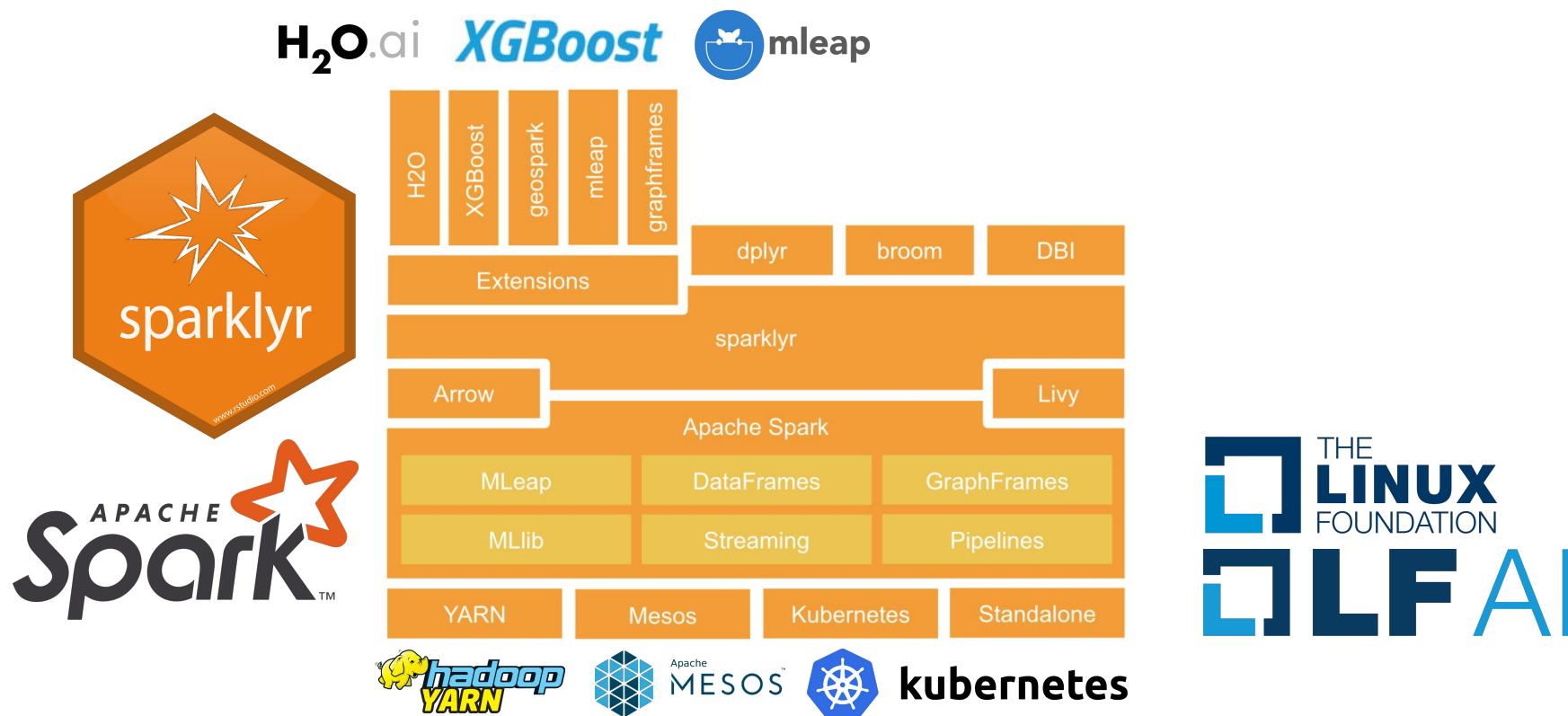


"R is a **programming language** and free software environment for **statistical computing** and **graphics**."



sparklyr: R Interface to Apache Spark

sparklyr is an open-source and modern interface to scale data science and machine learning workflows using Apache Spark™ and R.



Why sparklyr to scale Data Science?

From the very-beginning, way before Pandas, sparklyr was designed to reuse as much R code through packages like dplyr, broom, rlang, tidyverse, etc. and practices like the pipe operator, formulas, and so on.



```
library(sparklyr)
sc <- spark_connect(master = "local|yarn|mesos|spark|livy")
flights <- copy_to(sc, flights)
```



```
library(dplyr)
flights %>%
  group_by(month, day) %>%
  summarise(count = n(), avg_delay = mean(dep_delay)) %>%
  filter(count > 1000)
```

How do I use Spark with R in Qubole?

Seamless, non-disruptive RStudio Server Pro integration with Qubole — combining the power of the RStudio Server Pro with enhanced Apache Spark capabilities and near-zero cluster administration benefits provided by Qubole.

R Studio Blog

Home About Categories Tags Archives

RSS Twitter Facebook

RStudio Adds New R Features in Qubole's Open Data Lake

Samantha Toet

2020-08-03

Tags: [RStudio Server Pro](#)

The screenshot shows a blog post on the R Studio blog. The title is "RStudio Adds New R Features in Qubole's Open Data Lake". The author is Samantha Toet, and the date is 2020-08-03. The tags listed are "RStudio Server Pro". Below the post is a screenshot of the Qubole interface showing a cluster named "spark" with one active node. The sidebar includes links for "Upcoming webinars", "Categories", and a search bar.

Launch RStudio Server Pro from inside the Qubole platform

We are excited to team up with Qubole to offer data science teams the ability to [use RStudio Server Pro from directly within the Qubole Open Data Lake Platform](#). Qubole is an open, simple, and secure data lake platform for machine learning, streaming and ad-hoc analytics. RStudio and Qubole customers now have access to RStudio's out-of-the-box features and Qubole's unique managed services that supercharge data science and data exploration

The screenshot shows the Qubole website homepage. The header includes the Qubole logo and navigation links for PLATFORM, SOLUTIONS, PRICING, PARTNERS, CUSTOMERS, SERVICES AND SUPPORT, RESOURCES, COMPANY, and a search icon. The main banner features a blue background with a network graph and the text "Integration of RStudio & Qubole Platform come together at your Fingertips | Qubole". Below the banner is a "START FREE TRIAL" button. The sidebar on the right contains sections for "Blog Subscription" (with a form to enter an email address) and "Recent Posts". The central content area includes a timestamp "August 6, 2020" and a bio for Vipul Modi and Pradeep Reddy, followed by a paragraph about the integration of RStudio and Qubole.

August 6, 2020 by Vipul Modi and Pradeep Reddy

The integration of both platforms accelerates data science and scientific research with a single click access to large datasets within the RStudio integrated development environment (IDE) for data scientists.

Data scientists use RStudio for machine learning (ML), artificial intelligence (AI), and data exploration. With the vast amounts of enterprise and other sources of data that are accessible today, the volume of data to be processed for ML and exploration requires the power of cluster computational frameworks like Apache Spark. While R has

Blog Subscription

Get the latest updates on all things big data.

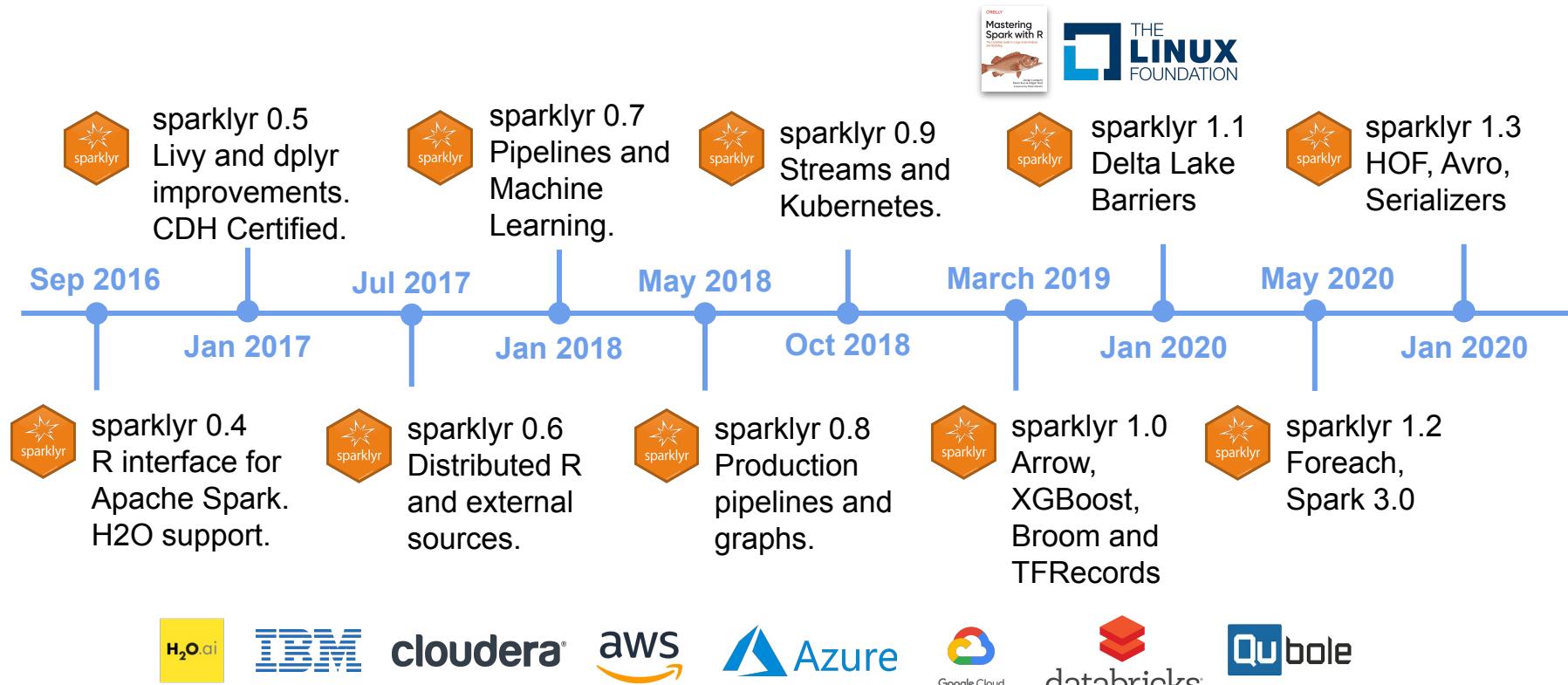
Email Address *

SUBSCRIBE NOW

Recent Posts

Catching up with sparklyr?

Since 2016, RStudio and the R community have been developing sparklyr. Starting with support for dplyr, Spark's MLlib and H2O; to structured streaming, Apache Arrow, XGBoost, Delta Lake and joining the Linux Foundation.



How do I use Spark from R?

```
spark_install()  
sc <- spark_connect(master = "local")  
  
cars <- spark_read_csv(sc, "cars", "input/")  
  
summarize(cars, n = n())  
dbGetQuery(sc, "SELECT count(*) FROM cars")  
  
ml_linear_regression(cars, mpg ~ wt + cyl)  
  
ml_pipeline(sc) %>%  
  ft_r_formula(mpg ~ wt + cyl) %>%  
  ml_linear_regression()  
  
spark_context(sc) %>% invoke("version")  
spark_apply(cars, nrow)  
  
stream_read_csv(sc, "input/") %>%  
  filter(mpg > 30) %>%  
  stream_write_json("output/")  
  
# Install local Spark  
# Connect to Spark cluster  
  
# Read data in Spark  
  
# Count records with dplyr  
# Count records with DBI  
  
# Perform linear regression  
  
# Define Spark pipeline  
# Add formula transformation  
# Add model to pipeline  
  
# Extend sparklyr with Scala  
# Extend sparklyr with R  
  
# Define Spark stream  
# Add dplyr transformation  
# Start processing stream
```

Where can I learn more about Spark and R?

Since 2016, RStudio and the R community have been developing sparklyr. Starting with support for dplyr, Spark's MLlib and H2O; to structured streaming, Apache Arrow, XGBoost, Delta Lake and joining the Linux Foundation.

THE LINUX FOUNDATION PROJECTS



About Get Started Learning Community 🐾 Q

Scale Data Science with Spark and R

About

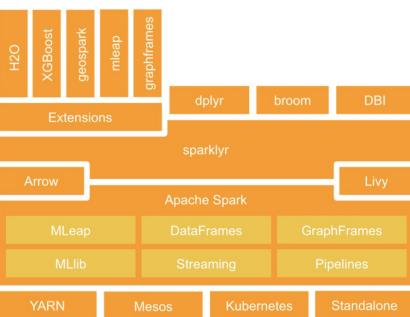
sparklyr is an open-source and modern interface to scale data science and machine learning workflows using **Apache Spark™**, R, and a rich **extension ecosystem**.

It enables using Apache Spark with ease using R by providing access to core functionality like installing, connecting and managing Spark and using Spark's **MLlib**, **Structured Streaming** and **Spark Pipelines** from R.

Supports well-known R packages like **dplyr**, **DBI** and **broom** to reduce the cognitive overhead from having to re-learn libraries.

And enables a rich-ecosystem of extensions to use in Spark and R: **XGBoost**, **MLeap**, **GraphFrames**, **H2O**, and optionally enable **Apache Arrow** to significantly improve performance.

Through Spark, this allows you to scale your Data Science workflows in **Hadoop YARN**, **Mesos**, **Kubernetes** or **Apache Livy**.



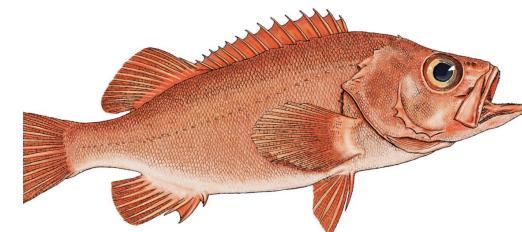
The diagram illustrates the sparklyr ecosystem. At the center is Apache Spark. It connects to various components: **Extensions** (H2O, XGBoost, geospark, mleap, graphframes), **Arrow**, **Livy**, **DataFrames**, **GraphFrames**, **MLlib**, **Streaming**, and **Pipelines**. These components further connect to **dplyr**, **broom**, and **DBI**. The entire stack sits atop **YARN**, **Mesos**, **Kubernetes**, and **Standalone** environments.

sparklyr.ai

O'REILLY®

Mastering Spark with R

The Complete Guide to Large-Scale Analysis and Modeling



Javier Luraschi,
Kevin Kuo & Edgar Ruiz
Foreword by Matei Zaharia

therinspark.com

THE DATA LAKE SUMMIT

The definitive virtual conference for all things Data Lake

OCT 13-14, 2020

Thank you!



@dfalbel



@yl790



@javierluraschi



@zkajdan



@samantha_toet

The screenshot shows the RStudio AI Blog homepage. It features a sidebar with categories like Audio Processing, Bayesian Modeling, Cloud, DeepOps, Data Management, Distributed Computing, Explainability, Image Recognition & Image Processing, Maths, Natural Language Processing, Packages/Releases, Privacy & Security, and Probabilistic ML/DL. The main content area displays five recent blog posts with their titles and small thumbnail images:

- Sept. 7, 2020: Introducing sparklyr.flint: A time-series extension for sparklyr
- Sept. 1, 2020: An introduction to weather forecasting with deep learning
- Aug. 24, 2020: Training ImageNet with R
- Aug. 16, 2020: Deepfake detection challenge from R
- July 31, 2020: FNN-VAE for noisy time series forecasting

blogs.rstudio.com/ai

[Section Divider Title]

THE DATA LAKE SUMMIT

The definitive virtual conference for all things Data Lake

OCT 13-14, 2020

Thank you

- Share this session on social media using **#DataLakeSummit**
- Interact with other attendees on Slack at **datalake-summit.slack.com**