NEW YORK

2021

WHERE R ENTHUSIASTS AND DATA SCIENTISTS
GATHER TO EXPLORE, SHARE AND INSPIRE IDEAS

#rstatsnyc

Presented by landeranalytics

Originating in New York City with expansions in Washington D.C., and Dublin, Ireland, the R Conference hosts one of the most elite gatherings of data scientists and data professionals who come together to explore, share, and inspire ideas, and to promote the growth of open source ideals.

TABLE OF CONTENTS

SPEAKER ABSTRACTS

DAY 1 | Thursday, September 9th

R CONFERENCE NEW YORK

8:30 a.m. - 8:50 a.m. (Eastern Daylight Time)	Open Virtual Registration 
8:50 a.m. - 9:00 a.m.	**Opening Remarks** Jared P. Lander, Lander Analytics @jaredlander
9:00 a.m. - 9:20 a.m.	**Creating Production Level Data Science Code** Megan Robertson, Nike @leggomymeggo4
9:25 a.m. - 9.45 a.m.	**Set your R code free; turn it into a command-line tool** Jeroen Janssens, Data Science Workshops @jeroenhjanssens
9:50 a.m. - 10:10 a.m.	**How to Write Engaging R Tutorials** Danielle Oberdier, DiKayo Data @dikayodata
10:10 a.m. - 10:40 a.m.	Break & Networking
10:40 a.m. - 11:00 a.m.	**R for the Planet** Alexa Fredston, Ecology, Evolution, and Natural Resources, Rutgers University @AFredston
11:05 a.m. - 11:25 a.m.	**GPU Computing in R** Jared P. Lander, Lander Analytics @jaredlander
11:30 a.m. - 11:50 a.m.	**Learner Personas for Domain-Specific Data Science Educational Materials** Daniel Chen, Lander Analytics @chendaniely
11:50 a.m. - 1:00 p.m.	Lunch & Networking
1:00 p.m. - 1:20 p.m.	**Impacting Churn with Data** Caitlin Hudon, OnlineMedEd @beeonaposy

*All times are EDT

DAY 1 | Thursday, September 9th

R CONFERENCE NEW YORK

1:25 p.m. - 1:45 p.m.	**Go or Kick? Enhancing the NFL Fan Experience with Real-Time 4th Down Decision Analytics** Mike Band, NFL Next Gen Stats @MBandNFL
1:50 p.m. - 2:10 p.m.	**One team, one dream, one ggplot2 color scheme: how a group of strangers won the NFL Big Data Bowl** Asmae Toumi, PursueCare @asmae_toumi
2:10 p.m. - 2:40 p.m.	Break & Networking
2:40 p.m. - 3:00 p.m.	**AB Testing of Data Science Solutions** Reenah Nahum Muldavski, Data Science Services
3:05 p.m. - 3:25 p.m.	**Personalizing Mental Healthcare at Scale** Adam Chekroud, Spring Health @itschekkers
3:30 p.m. - 3:50 p.m.	**Against Machine Learning; For Causal Inference** Sarah Catanzaro, Amplify Partners @sarahcat21
3:50 p.m. - 4:20 p.m.	Break & Networking
4:20 p.m. - 4:40 p.m.	**5 mistakes you'll probably make with language data (and how to recover)** Rachael Tatman, Rasa @rctatman
4:45 p.m. - 5:05 p.m.	**{morphemepiece}: more meaningful tokenization for NLP** Jonathan Bratt, Macmillan Learning
5:05 p.m. - 5:15 p.m.	Closing Remarks
5:30 p.m. - 6:15 p.m.	**Join us for Happy Hour with Cointreau & The Botanist Gin!**

*All times are EDT

DAY 2 | Friday, September 10th

R CONFERENCE NEW YORK

9:30 a.m. - 9:50 a.m. (Eastern Daylight Time)	Open Virtual Registration hopin
9:50 a.m. - 10:00 a.m.	Opening Remarks
10:00 a.m. - 10:20 a.m.	**Using Data to Make Classroom Decisions** COL. Krista Watts,United States Military Academy
10:25 a.m. - 10.45 a.m.	**Responsible AI and Ethics and Beyond: A principled approach from Microsoft** Sonia Ang, Microsoft @galleontrade
10:45 a.m. - 11:15 a.m.	Break & Networking
11:15 a.m. - 11:35 a.m.	**Why the world loves K-pop** Chrys Wu, Matchstrike @MacDiva
11:40 a.m. - 12:00 p.m.	**The Global Pandemic Ruined My Favorite Data Set** Max Kuhn, RStudio @topepos
12:05 p.m. - 12:25 p.m.	**New Directions for Apache Arrow** Wes McKinney, Ursa Computing @wesmckinn
12:25 p.m. - 12:30 p.m.	**9/11 Tribute: Never Forget, 20th Anniversary**
12:30 p.m. - 1:35 p.m.	Lunch & Networking
1:35 p.m. - 1:55 p.m	**dbcooper: Turn any database into an R package** David Robinson, Heap @drob

*All times are EDT

DAY 2 | Friday, September 10th

R CONFERENCE NEW YORK

2:00 p.m. - 2:40 p.m.	**Wrong Again! 30+ Years of Statistical Mistakes** Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University @StatModeling
2:40 p.m. - 3:10 p.m.	Break & Networking
3:10 p.m. - 3:30 p.m.	**Does Interpretable Machine Learning *Really* Matter? (Aka How Learned to Stop Worrying and Love the Bayes with rstanarm)** Jonathan Hersh, Chapman University Argyros School of Business @DogmaticPrior
3:35 p.m. - 3:55 p.m.	**Addressing missing data in R: understudied contexts, lots of missingness, hundreds of variables** Mayari Montes de Oca, NYU Global TIES for Children @Mayari_MOca
4:00 p.m. - 4:20 p.m.	**Project Robyn 3.0: Continuous & Semi-Automated Marketing Mix Model from Facebook Marketing Science** Bernardo Lares & Igor Skokan, Facebook @laresdj
4:20 p.m. - 4:30 p.m.	Closing Remarks
4:30 p.m. - 5:30 p.m.	**Join us for a live taping of SuperDataScience Podcast with Jon Krohn and special guest Drew Conway!** SuperDataScience

*All times are EDT

The backbone of the R language is its community of users, contributors and supporters. The open source ethos of this community propels the language forward with tens of thousands of add-on packages and a helpful, welcoming environment. All around the world, R users hold meetups where knowledge is shared and relationships are formed. This conference grew out of the New York Open Statistical Programming Meetup (also known as the New York R Meetup), the largest in the world, with almost 12,000 members. Topics from the meetup include data science, visualization, machine learning, deep learning and so much more. You can browse 11 years of presentations at nyhackr.org.

The **R Conference** in **New York**, **Washington D.C.**, and, soon, **Dublin**, were created to foster the local R communities and serve as fun gathering places where people can learn from their peers in an inviting setting. Because we cannot gather in person this year, we are meeting on a virtual platform designed to stream live talks and encourage great personal interactions, even remotely.

Thank you for joining us virtually. We hope to be back in-person soon, we miss you all!

Jared P. Lander
Chief Data Scientist

landeranalytics
UNLOCKING DATA TO DRIVE YOUR BUSINESS

UNLOCKING DATA TO DRIVE YOUR BUSINESS

Machine Learning & AI Solutions | Professional Training
Software Installation & Maintenance | Reporting & Dashboards

Lander Analytics is a full-service consulting firm based in New York City helping clients enhance their analytical capabilities to drive value from data. Led by Chief Data Scientist Jared Lander, our team of elite data scientists, statisticians, visual designers, published authors, professors, keynote speakers and management consultants are united by our shared talent and passion for leveraging data science to meet real world challenges.

Collectively, Lander Analytics is a recognized leader in the open source data community, hosting events like the popular annual R conferences in New York City and Washington DC, with future conferences coming to Tampa, Florida and Dublin, Ireland!

LANDERANALYTICS.COM
INFO@LANDERANALYTICS.COM
@LANDERANALYTICS

WE'RE HIRING!

landeranalytics

Megan Robertson
Nike

9:00 a.m. - 9:20 a.m. EDT

Creating Production Level Data Science Code

A data scientist writes code throughout every stage of a project from exploratory data analysis to evaluating models and summarizing results. Once you've developed a proof of concept or minimally viable product it can be a daunting task to put it into production. How do you organize and adapt all the code that you created? What can you do to make sure the code catches errors and alerts you to them? Do you feel overwhelmed by everything you need to do? By attending this presentation you will learn tips and strategies to organize your own code during a project to make creating production code easier. You will also learn how to optimize your code to catch errors and create effective documentation. | @leggomymeggo4

Jeroen Janssens
Data Science Workshops

9:25 a.m. - 9:45 a.m. EDT

Set your R code free; turn it into a command-line tool

If your data analyses involve coding, then you know how liberating it is to use and create functions. They hide complexity, improve testability, and enable reusability. In this talk I explain how you can really set your R code free: by turning it into a command-line tool. The command line can be a very flexible and efficient environment for working with data. It's specialized in combining tools that are written in all sorts of languages (including R and Python), running them in parallel, and applying them to massive amounts of (streaming) data. Although the command line itself has quite a learning curve, turning your existing R code into a tool is, as I demonstrate, a matter of a few steps. I discuss how your new tool can be combined with existing tools in order to obtain, scrub, explore, and model data at the command line. Finally, I share some best practices regarding interface design. | @jeroenhjanssens

Danielle Oberdier
DiKayo Data

9:50 a.m. - 10:10 p.m. EDT

How to Write Engaging R Tutorials

Whether you're trapped in a quiet place with no headphones or simply don't want to pause your favorite TV show, written R tutorials often rival video tutorials in terms of efficiency and of course the ability to quickly transfer strings of code to one's own platform. But what makes a written tutorial stand out amongst all the resources out there? In this talk, I will show you how to determine the right length for a given tutorial, which R packages are best for this type of teaching and most importantly, how to make your written tutorials personal and engaging. | @dikayodata

Alexa Fredston
Ecology, Evolution, and Natural Resources, Rutgers University

10:40 a.m. - 11:00 a.m. EDT

R for the Planet

In the rapidly growing field of environmental data science, R is the language of choice for many researchers seeking to forecast the impacts of extreme weather events, chronicle global biodiversity loss, or map injustice in environmental health. In this talk, I'll showcase some of the challenges frequently encountered by environmental data scientists, and the R tools we use to solve them — from data wrangling to spatial analysis and Bayesian models. Through a series of real examples that required fitting models to messy data on biodiversity, oceans, and climate change, I'll demonstrate how ecological and environmental researchers are leveraging R to help save the planet. | @AFredston

Jared P. Lander
Lander Analytics

11:05 a.m. - 11:25 a.m. EDT

GPU Computing in R

Parallel computing has become easier and easier in R over the years thanks to packages like parallel and future. But the CPU measures cores in the single or double digits. With GPUs we can access thousands of cores, significantly speeding up our work. Taking advantage of the GPU for machine learning has never been easier thanks to torch, xgboost, catboost and Stan. We look at how to fit those models on the GPU and how to use some lower level code to perform custom operations with the GPU. | @jaredlander

Dan Chen
Virginia Tech &
Lander Analytics

11:30 a.m. - 11:50 a.m. EDT

Learner Personas for Domain-Specific Data Science Educational Materials

Finding data science learning and teaching materials is not what educators and learners will find difficult these days. Rather, finding domain-specific materials that will resonate with learners is the current challenge. In the medical sciences, many of our learners only know about spreadsheets,and treat our data as a visualization, using colors, spaces, one-off tables, and side calculations. They lack the vocabulary to talk and work with data in a programmatic manner that integrates with other data scientists.
This is a talk intended for data science educators and the education community. We adapted surveys from The Carpentries, "How Learning Works", and "Teaching Tech Together" to create a learner self-assessment survey to discover learner personas in the biomedical sciences by clustering survey results. These personas and findings were used to create a data science curriculum that is grounded in data literacy topics around spreadsheets and good data practices. | @chendaniely

Caitlin Hudon
OnlineMedEd

1:00 p.m. - 1:20 p.m. EDT

Impacting Churn with Data

"Data scientists are uniquely empowered to solve big business problems, and analyzing and understanding churn is one area which lends itself to impactful analysis. Using my recent work on churn as a case study, this talk will cover:
-How to get buy-in for ""big business problem"" projects, and how to structure analysis projects such that you're adding and delivering value at multiple checkpoints
-How to turn exploratory data analysis into useful deliverables
-How to tackle subscription churn through holistic analysis and thoughtful, actionable recommendations" | @beeonaposy

Mike Band
NFL Next Gen Stats

1:25 p.m. - 1:45 p.m. EDT

Go or Kick? Enhancing the NFL Fan Experience with Real-Time 4th Down Decision Analytics

Before any given 4th down play, an NFL head coach must decide between keeping the offense on the field and going for it, or calling for the special teams unit to attempt a field goal or punt. Nearly every team has at least one staff member crunching the numbers in these situations over the course of a game. Our team at Next Gen Stats, in collaboration with Amazon Web Services, are taking 4th down and two-point decision analytics to the next level. Powered by a series of machine learning models, the Next Gen Stats Decision Guide analyzes crucial coaching decisions in real-time. Should the team go for it, or kick? Let's see what the numbers say… | @MBandNFL

spring health x R CONFERENCE

Mental healthcare needs to be better, faster, and cheaper

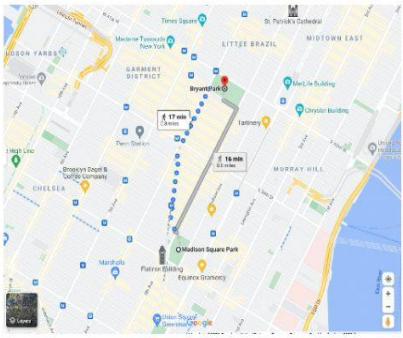
Traditional Care	Other Digital Mental Health Solutions	Spring Health
		
Long wait times	Shorter wait times	Shorter wait times
Trial and error	Trial and error	Right care, right time

Spring Health helps people find the right care for them, get appointments the next day, and gets your employer to pay for it (anonymously)

We're growing our Data Science and Business Intelligence teams

to take mental healthcare from this….

to this.



Senior Director of Data
Manager, Data Science
Senior Data Scientist
Data Scientist
Senior Data Analyst

- Awesome team & strong mission
- Over 5X annual revenue growth
- 300-400 employees, HQ in NYC
- Diverse, empathetic, and fast team
- We love R

Visit https://jobs.lever.co/springhealth

Asmae Toumi
PursueCare

1:50 p.m. - 2:10 p.m. EDT

One team, one dream, one ggplot2 color scheme: how a group of strangers won the NFL Big Data Bowl

How does a group of strangers with minimal football strategy knowledge go on to win sport's biggest data science competition during a pandemic? Asmae Toumi will share her and her group's process, lessons learned, and how they leveraged the tidyverse, tidymodels ecosystem, and other R packages to gain a competitive edge. | @asmae_toumi

Reenah Nahum Muldavski
Chief Data Scientist

2:40 p.m. - 3:00 p.m. EDT

AB Testing of Data Science Solutions

In recent years data science models have increased the efficiency and value of products across many industries. Data science models in many cases affect the product backend logic in a fundamental manner (such as directing customers in queues), and classic "split testing" is hard or even impossible to implement. AB testing methodology for frontend features is solid and well defined, however, the methodology for testing the value of backend enhancements is not as firm and well applied. In this talk we will cover experimental design best practices for testing and measuring the value of data science models.

Adam Chekroud
Spring Health

3:05 p.m. - 3:25 p.m. EDT

Personalizing Mental Healthcare at Scale

Depression is the world's leading cause of disability, and almost 1 in 4 people will suffer some kind of mental illness each year. However, most people don't get a diagnosis, don't get treatment, or don't fully recover. Adam Chekroud will talk about how Spring Health uses data to improve mental healthcare at scale, and how statistics help drive better outcomes throughout the process. | @itschekkers

Sarah Catanzaro
Amplify Partners

3:30 p.m. - 3:50 p.m. EDT

Against Machine Learning; For Causal Inference

Nearly every day, data teams and venture capitalists implicitly express their priorities and outlook by making decisions on how to allocate budget or capital to advance different technology initiatives. In the past 5 years, both groups have prioritized machine learning and business intelligence initiatives, by investing the tools and platforms to support these projects. They have not; however, invested in tools and platforms to advance causal inference. In this talk, we will discuss why investments in causal inference may have a higher ROI. We'll then study the evolution of the MLOps stack to identify opportunities to unlock increased investment in causal inference and expand adoption in industry. | @sarahcat21

Rachael Tatman
Rasa

4:20 p.m. - 4:40 p.m. EDT

5 mistakes you'll probably make with language data (and how to recover)

Language is fundamentally different from other types of data, and it's inevitable that you'll run into some language-specific issues. This talk will cover some of the most common types of errors I've seen data analysts and machine learning engineers make with language data, from ignoring the differences between text genres to treating text as written speech to assuming that all languages work like English. We'll also talk about ways to avoid these common mistakes (and recover gracefully if you've already made them). @rctatman

Jonathan Bratt
Macmillan Learning

4:45 p.m. - 5:05 p.m. EDT

{morphemepiece}: more meaningful tokenization for NLP

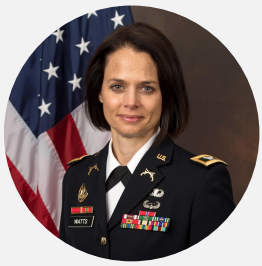
Modern language models use tokenizers based on subword-level vocabularies. Words not present in the vocabulary are broken into subword tokens. This subword tokenization is generally unrelated to the morphological structure of the word.

It is intuitively appealing to consider a tokenizer that uses a morpheme-level vocabulary to split words into meaningful units. Implementing such a tokenizer, while conceptually straightforward, presents a number of practical challenges.

We present an approach to solving these challenges and introduce {morphemepiece}, an R package that implements a new tokenization algorithm for breaking down (most) words into their smallest units of meaning.

DAY 2 | Friday, September 10th

R CONFERENCE NEW YORK



COL. Krista Watts
United States Military Academy

10:00 a.m. - 10:20 a.m. EDT

Using Data to Make Classroom Decisions

The nature of how classes are executed at the United States Military Academy makes it a unique opportunity to assess the effectiveness of different pedagogical approaches. Core classes generally have dozens of small sections, often with students randomly assigned to sections, facilitating the opportunity for cluster randomized trials. I will discuss several recent studies including sectioning by demonstrated aptitude for a subject and use of technology in the classroom.

Sonia Ang
Microsoft

10:25 a.m. - 10:45 a.m. EDT

Responsible AI and Ethics and Beyond: A principled approach from Microsoft

Microsoft is championing how to bring change to society in the use and implementation of AI and its impact has been felt from a social–technological point of view. A stark reminder was when it launched the Twitter chatbot, Tay, who ended up with bigoted rhetoric, hence a realization to consider the human element when designing AI systems. Along with innovation comes a responsibility to make sure that the future is secured. We need to take a thoughtful approach to ensure we create a future we want to see and not one we fear. This presentation revolves around an ethical framework with five core principles of fairness, reliability, safety, privacy and security, underpinned with transparency and accountability. | @galleontrade

RStudio, a Single Home for R and Python

Build and share your data science work in R and Python

Many Data Science teams today are bilingual, leveraging both R and Python in their work. With RStudio, you can develop, collaborate, manage and share your data science work in R and Python, all on single infrastructure.

LEARN MORE AT RSTUDIO.COM/PYTHON

RStudio®

Chrys Wu
Matchstrike

11:15 a.m. - 11:35 a.m. EDT

Why the world loves K-pop

This talk explores the many facets of K-pop, starting with groups like BTS and Blackpink and diving further into fandom to see trends and influences. In this “fun with R” / “hobby R” talk we’ll also explore things like release and promotion schedules, connections, and influences so we can enjoy and understand more about how this genre has been shaping popular culture. | @MacDiva

Max Kuhn
RStudio

11:40 a.m. - 12:00 p.m. EDT

The Global Pandemic Ruined My Favorite Data Set

My beloved Chicago train ridership data, like so many other things, was severely impacted by the pandemic. If I want to build models, what should I do? This talk will describe some approaches for mitigating the effect that the pandemic had on L train ridership. | @topepos

Wes McKinney
Ursa Computing

12:05 p.m. - 12:25 p.m. EDT

New Directions for Apache Arrow

Wes McKinney shares an update about recent developments in Apache Arrow, the multi-language toolbox for accelerated data interchange and in-memory processing. Wes introduces some new directions for the Arrow project and discusses why Ursa Computing, which he founded last year, has joined forces with BlazingSQL and the pioneers of RAPIDS and other open source projects to form Voltron Data. | @wesmckinn

David Robinson
Heap

1:35 p.m. - 1:55 p.m. EDT

dbcooper: Turn any database into an R package

If your organization uses a database, then you have a lot to gain by building an R package to make interfacing with that database easy and intuitive. Packages like DBI and odbc handle the creation of the database connection, and dbplyr lets you translate dplyr syntax to SQL. But there’s a missing layer in between, such that working with a connection object still doesn’t feel like exploring and joining tables in memory. In this talk I’ll introduce the dbcooper package, which wraps any database connection to turn it into a set of R functions, making it easy to create a database-specific R package. dbcooper makes the management of connections transparent so that you engage with the database through prefixed functions, as well as generating autocomplete-friendly accessors for each table for fast exploratory data analysis. I’ve used this general approach as the foundation of a data science ecosystem at several companies, and I’ll show an example of using the package to explore a public BigQuery database. | @drob

Andrew Gelman
Columbia University

2:00 p.m. - 2:40 p.m. EDT

Wrong Again! 30+ Years of Statistical Mistakes

One of the benefits of a long career is that it gives us an opportunity to reflect upon all the ways our thinking has changed. In this talk I'll go over several places where my thinking has changed, for each considering why I previously took a stance that I currently disagree with, and where I anticipate my views might change further. I hope this discussion will be useful in helping each of you to introspect on your own past and future intellectual development. | @StatModeling

Jonathan Hersh
Chapman University Argyros School of Business

3:10 p.m. - 3:30 p.m. EDT

Does Interpretable Machine Learning *Really* Matter? (Aka How Learned to Stop Worrying and Love the Bayes with rstanarm)

After decades spent trying to teach computers to think, we now face the problem that AI and ML models often know more than they can communicate to us about why they make certain predictions. Interpretable machine learning, such as LIME or Shapely values, tries to shift that balance, by presenting a view towards the inner working of our complex models. My collaborator Selina Carter built some machine learning models and I did some interpretable AI, and I somehow convinced her to let me run a randomize controlled trial with 685 employees at a large firm, with half of them receiving an interpretable AI treatment. Now before you go all Andrew Gelman, I want to say that YES of course I used Bayes to analyze the data. I hadn't used Bayes since the JAGS days and I want to say rstanarm is fantastic and the people who created it should be showered with praises. Why am I talking here? They're the ones who should be celebrated. | @DogmaticPrior

Mayari Montes de Oca
NYU Global TIES for Children

3:35 p.m. - 3:55 p.m. EDT

Addressing missing data in R: understudied contexts, lots of missingness, hundreds of variables

Mayarí will talk about the importance of addressing missing data and of her practical experience when handling the non-response from a longitudinal study with Syrian refugee populations. After a rapid overview of multiple imputation, she will talk about some of the main challenges of addressing non-response comprehensively for data that caters to multiple analyses and researchers; in contexts where little is known about the DGP of each variable, missing data is prevalent, and hundreds of potentially important features are collected. She will share some of the software and settings that were helpful throughout the process, such as the Boruta, randomForest, mice, and miceadds packages. Lastly, she will illustrate the importance of transparency in analysis results regarding the uncertainty brought in by the missing data. | @Mayari_MOca

Bernardo Lares & Igor Skokan
Facebook

4:00 p.m. - 4:20 p.m. EDT

Project Robyn 3.0: Continuous & Semi-Automated Marketing Mix Model from Facebook Marketing Science

Robyn is an experimental, semi-automated and open-sourced Marketing Mix Modeling (MMM) package from Facebook Marketing Science. It uses various machine learning techniques (Ridge regression with cross validation, multi-objective evolutionary algorithm for hyperparameter optimisation, time-series decomposition for trend & season, gradient-based optimisation for budget allocation etc.) to define media channel efficiency and effectivity, explore adstock rates and saturation curves. It's built for granular datasets with many independent variables and therefore especially suitable for digital and direct response advertisers with rich data sources. | @laresdj

THANK YOU TO OUR SPONSORS!

Gold

spring health

RStudio®

Silver

consortium

Bronze

Visiting Nurse Service of New York

Supporting

Pearson

Springer

MANNING

CRC Press
Taylor & Francis Group
CHAPMAN & HALL

Vibe

COINTREAU
THE ART OF THE MIX

THE BOTANIST
ISLAY DRY GIN