R CONFERENCE

GOVERNMENT & PUBLIC SECTOR

2021

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

#rstatsgov
@rstatsai

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

Conference Schedule: Day 1 (Thursday) & Day 2 (Friday)

The History of the R Conference: 7 Years of Fun & Learning

Learn More About the Organizers: The Lander Analytics Team

Thank You: All our wonderful sponsors

SPEAKER ABSTRACTS

Aaron Mannes	Jordan Jasuta Fischer	Cezary Podkul	Coline Zeballos
Jared P. Lander	Asmae Toumi	Brook Frye & Ben Witte	David Shor
Alex Gold	Dr. Wendy Martinez	Marck Vaisman	Lauren Lombardo
Madhava Jay	Tommy Jones	Surabhi Hodigere	Jasmine Ye Han
Jorge Luna	Vivian Peng	Boriana P. Pratt	Mayari Montes de Oca
Benjamin Braun	Dr. Abhijit Dasgupta	Sydney Coston	

CONFERENCE
GOVERNMENT & PUBLIC SECTOR

DAY 1 | Thursday, December 9th

R CONFERENCE

All times are EST	
8:00 a.m. - 8:50 a.m.	Virtual Breakfast & Registration hopin
8:50 a.m. - 9:00 a.m.	**Opening Remarks** Jared P. Lander, Lander Analytics @jaredlander
9:00 a.m. - 9:20 a.m.	**Big Data @DHS: Vast and Varied** Aaron Mannes, Culmen LLC supporting DHS S&T @awmannes
9:25 a.m. - 9.45 a.m.	**Ensemble NLP to classify medical conditions** Jordan Jasuta Fischer, IBM @JordanJasuta
9:50 a.m. - 10:10 a.m.	**“Data or it didn’t happen": How to make data-driven news stories happen** Cezary Podkul, ProPublica @Cezary
10:10 a.m. - 10:40 a.m.	Break & Networking
10:40 a.m. - 11:00 a.m.	**Ensuring the quality of your R packages for regulatory submissions** Coline Zeballos, Roche Pharma @colinezeballos
11:05 a.m. - 11:25 a.m.	**Using {targets}, {arrow}, Docker, Postgres and the Command Line for Medium Data** Jared P. Lander, Lander Analytics @jaredlander
11:30 a.m. - 11:50 a.m.	**Using data and R to improve substance use disorder care: the startup experience** Asmae Toumi, PursueCare @asmae_toumi
11:50 a.m. - 1:00 p.m.	Lunch & Networking
1:00 p.m. - 1:20 p.m.	**Data and Decision Making in the Legislative Process** Brook Frye & Ben Witte, New York City Council @brook_frye

DAY 1 | Thursday, December 9th

R CONFERENCE

All times are EST	
1:25 p.m. - 1:45 p.m.	**Bayesian statistics, government, and the 2020 election** David Shor, Blue Rose Research @davidshor
1:45 p.m. - 2:15 p.m.	Break & Networking
2:15 p.m. - 2:35 p.m.	**Reliably Reproducible R-Tifacts** Alex Gold, RStudio @alexkgold
2:40 p.m. - 3:00 p.m.	**The Journey Continues: Using R at a U.S. Government Agency** Dr. Wendy Martinez, Bureau of Labor Statistics @BLS_gov
3:05 p.m. - 3:25 p.m.	**Creating and Managing Your University Course with R** Marck Vaisman, Microsoft @wahalulu
3:25 p.m. - 3:55 p.m.	Break & Networking
3:55 p.m. - 4:15 p.m.	**Building Government Platforms: The architectural, operational, and political choices behind building platforms in government** Lauren Lombardo, Harvard University John F. Kennedy School of Public Policy @_laurenlombardo
4:20 p.m. - 4:40 p.m.	**PETs: Remote Data Science Unleashed** Madhava Jay, OpenMined @madhavajay
4:40 p.m. - 4:50 p.m.	Closing Remarks
5:10 p.m. - 5:40 p.m.	**TRIVIA THURSDAY!** Get your trivia on & win some prizes! Join us as we play 3 rounds - R, Data Science & The Kitchen Sink.

DAY 2 | Friday, December 10th

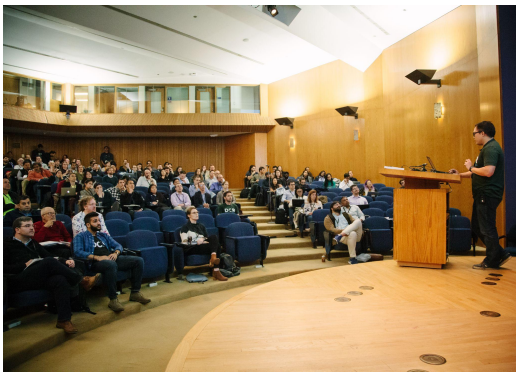
R CONFERENCE

All times are EST	
9:00 a.m. - 9:50 a.m.	Virtual Breakfast & Registration — hopin
9:50 a.m. - 10:00 a.m.	Opening Remarks
10:00 a.m. - 10:20 a.m.	**tidylda: Latent Dirichlet Allocation Using 'tidyverse' Conventions** Tommy Jones, In-Q-Tel @thos_jones
10:25 a.m. - 10.45 a.m.	**Governance Playbook for Digital Public Goods** Surabhi Hodigere, Ash Center for Democratic Governance, and Innovation at Harvard Kennedy School @surabhihodigere
10:45 a.m. - 11:15 a.m.	Break & Networking
11:15 a.m. - 11:35 a.m.	**A Data Journalist's R Toolbox** Jasmine Ye Han, Bloomberg Industry Group @JasmineHanYe
11:40 a.m. - 12:00 p.m.	**Hospital analytics approaches to help inform strategies to reduce 30-day hospital readmissions** Jorge Luna, Aetna, a CVS Health Company, Analytics & Behavior Change
12:05 p.m. - 12:25 p.m.	**Building Blocks of Design** Vivian Peng, City of Los Angeles @create_self
12:25 p.m. - 1:35 p.m.	Lunch & Networking
1:35 p.m. - 1:55 p.m.	**Running simulations in Parallel in R with doParallel package** Boriana P. Pratt, Office of Population Research, Princeton University

DAY 2 | Friday, December 10th

R CONFERENCE

All times are EST	
2:00 p.m. - 2:20 p.m.	**What works to support refugee children? Using BART for impact evaluation** Mayari Montes de Oca, Research Scientist @Mayari_MOca
2:25 p.m. - 2:45 p.m.	**We can't help if we can't communicate: Conveying data science findings to non-experts** Benjamin Braun, 202 Group @Ben_G_Braun
2:45 p.m. - 3:15 p.m.	Break & Networking
3:15 p.m. - 3.35 p.m.	**Multilingual pipelines for data analyses: R as glue** Dr. Abhijit Dasgupta, Georgetown University's Data Science and Analytics @webbedfeet
3:40 p.m. - 4:00 p.m.	**Using R Shiny to visualize stem cell treatment data over time** Sydney Coston, United States Military Academy
4:00 p.m. - 4:10 p.m.	Closing Remarks





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

Aaron Mannes
Culmen LLC supporting
DHS S&T

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

Big Data @DHS: Vast and Varied

If big data is characterized by volume, velocity, and variety, the Department of Homeland Security (DHS) is the ultimate big data organization. In its mission to protect the American people, DHS undertakes an array of diverse functions and often has to make decisions in real time. Using data analytics to enable these missions requires a blend of creativity and pragmatism. | @awmannes

Jordan Jasuta Fischer
IBM

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

Ensemble NLP to classify medical conditions

Medical terms are linguistically very specific: a letter or two can completely change the word, and prefixes and suffixes can link two words that otherwise look wildly different. As such, many typical methods of natural language processing (NLP) are ill-adapted to work with medical records and their specific vocabulary and syntax. When a government client needed to classify medical conditions for record processing, IBM built a hybrid ensemble model that incorporates both rules-based and machine learning classification, to accommodate the client's system structure while flexibly handling the nuances of medical terminology. | @JordanJasuta

Cezary Podkul
ProPublica

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

"Data or it didn't happen": How to make data-driven news stories happen

"Data or it didn't happen" is a credo we all live by. It's especially important for data journalism, but sourcing data for an investigation is rarely easy. In this talk I will walk you through how a data-driven story comes together and share some ideas for how the public sector and journalists can work together more effectively. | @Cezary

Coline Zeballos
Roche Pharma

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

Ensuring the quality of your R packages for regulatory submissions

Ensuring the reliability and quality of R packages used in regulatory interactions for drug approvals: a view on how Roche participates in enabling submissions in R. | @colinezeballos

Jared P. Lander
Lander Analytics

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

Using {targets}, {arrow}, Docker, Postgres and the Command Line for Medium Data

Most companies don't have big data, but rather medium data, that awkward in between where the data are too big to fit in memory but not big enough for Google-scale systems. Fortunately R has many options for working with data of this size. We will look at using the command line, {data.table} and {dplyr} to clean the data and load it into a Postgres database inside a Docker container. Then we will use {targets} to orchestrate the whole process. | @jaredlander

Asmae Toumi
PursueCare

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

Using data and R to improve substance use disorder care: the startup experience

The U.S. opioid epidemic, or opioid crisis, refers to the substantial medical, social, psychological and economic consequences due to the misuse and overdose deaths of a class of drugs called opioids. The number of drug overdose deaths increased by nearly 5% from 2018 to 2019 and has quadrupled since 1999, and over 70% of the 70,630 deaths in 2019 involved an opioid (CDC, 2021). PursueCare is a telehealth startup offering comprehensive care for opioid use disorder and other substance use disorders by combining telehealth technology, medication treatment and counseling. Asmae Toumi, the director of analytics and research at PursueCare, will talk about how data and R/RStudio's public and professional tools are being used to uncover trends, deliver care and improve outcomes.| @asmae_toumi

Brook Frye & Ben Witte
New York City Council

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

Data and Decision Making in the Legislative Process

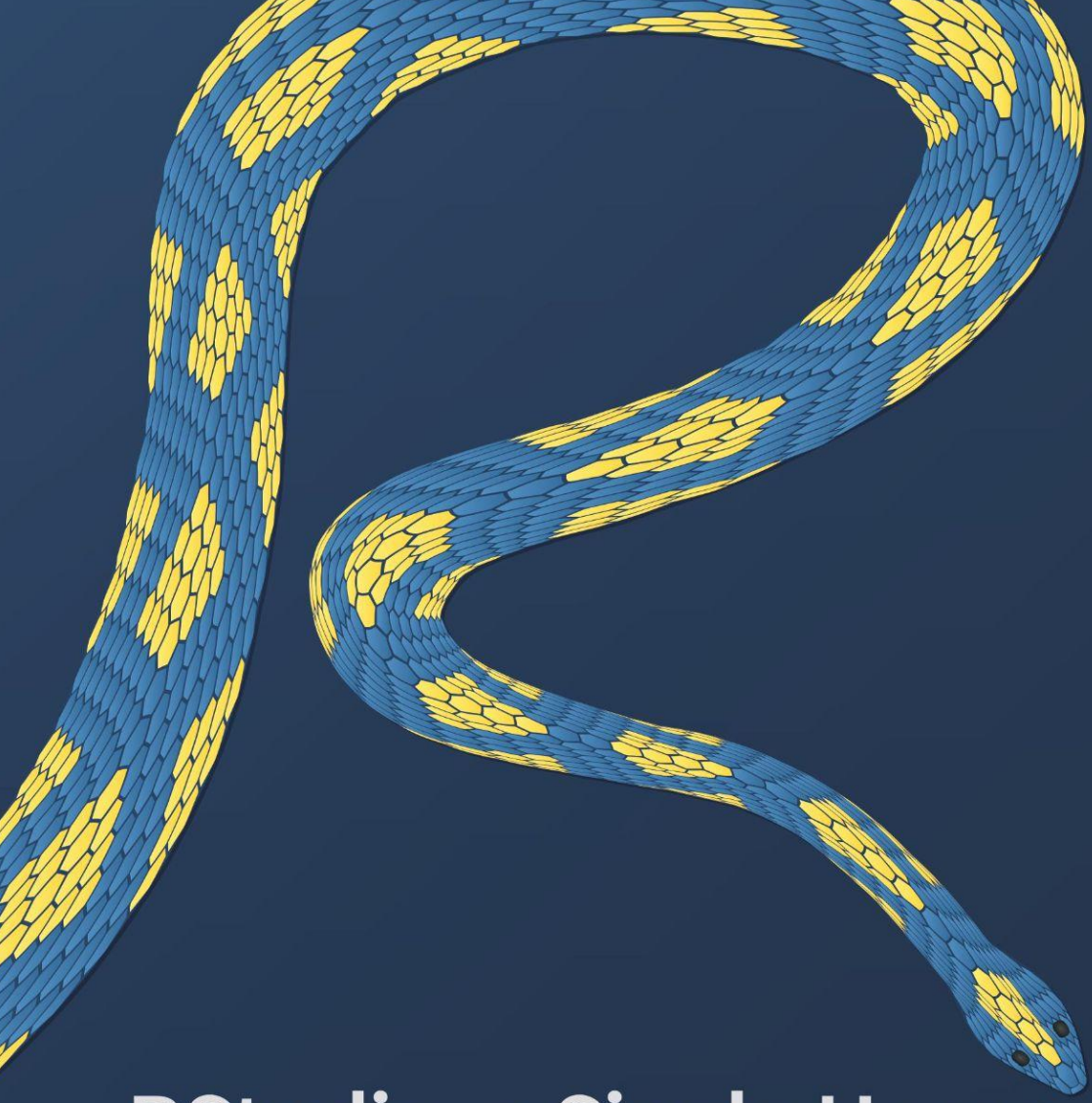
In New York City, the City Council has many functions, including oversight of the Mayor's operations and generating legislation that compliments the oversight function. We will provide a general overview of how data is used to underscore the rationale behind legislation and how the City Council works to ensure that these data feed into an overall ethos of transparency and evidence-based decision making. | @brook_frye

David Shor
Blue Rose Research

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

Bayesian statistics, government, and the 2020 election

A walk through how statistics and data science are commonly applied in politics and government. | @davidshor



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®

Alex Gold
RStudio

2:15 p.m. - 2:35 p.m. EST

Reliably Reproducible R-Tifacts

Some people get to write YORO (You Only Run Once) code, not really worrying about whether it'll run again. You probably aren't one of them. More likely, you have to be ready to re-run analyses months or years later. That's a tall order given the constant changes to the R language and package ecosystem. In this talk, you'll learn a taxonomy of reproducibility for your code, and be introduced to the foremost tools — docker and renv — for making your work environments more reproducible. | @alexkgold

Dr. Wendy Martinez
Bureau of Labor Statistics

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

The Journey Continues: Using R at a U.S. Government Agency

This presentation will continue the story that I started at last year's R Conference | Government & Public Sector. At the previous conference, I described some of my experiences – both successes and failures – using the open-source statistical computing software R at several U.S. government agencies. I described the goal of my journey, which was to get agreement from my agency to use R in the production of our official statistics. I am happy to announce that I have reached an important waypoint in this journey. R has been approved for production at the Bureau of Labor Statistics! Notice that I did not say I reached the end of my journey. This is because there is still a lot of important work ahead of us. In this talk, I will briefly recap the start of my journey, how I got to this point, and our way forward. | @BLS_gov

Marck Vaisman
Microsoft

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

Creating and Managing Your University Course with R

Did you know you can use R to create and maintain teaching materials, including slides, assignments, exams and even a website? This talk will illustrate how several R packages -including but not limited to {xaringan}, {rmarkdown}, {distill}, {xaringanThemer}, {ghclass}, and {Rexams}- are used in preparing and maintaining materials for courses I teach at Georgetown University and the George Washington University. | @wahalulu

Lauren Lombardo
Harvard University John F. Kennedy School of Public Policy

3:55 p.m. - 4:15 p.m. EST

Building Government Platforms: The architectural, operational, and political choices behind building platforms in government

Public sector organizations are increasingly turning to platforms as ways to improve service delivery while reducing costs. However, the term "platform" has been used to describe several different architectural designs and operational approaches. Decision-makers need to understand how their selected architectural design and operational approach, which are constrained by government structures, will impact their implementation of government platforms. Without a clear definition and an understanding of the technical decisions that must be made it is impossible to responsibly build and implement public sector platforms. | @_laurenlombardo

Madhava Jay
OpenMined

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

PETs: Remote Data Science Unleashed

Ever wished you could get access to more data for your data science problems without the painful and slow process of existing data access agreements?

Data Scientists are limited to the data their organization has painstakingly acquired a copy of. Data which often requires phone calls, contract negotiations, lawyers and special onsite security policies just to access. Getting to analyze personal data can take anywhere from weeks to months even if you understand the whole process.

Privacy Enhancing Technologies (PETs) are bringing that time down to seconds while giving data subjects even stronger privacy guarantees.

In this talk we will: Examine the privacy problem and the field of Privacy Enhancing Technologies (PETs), See how Syft's Automatic Differential Privacy and Secure Multi-Party Compute feels like magic, Hear about OpenMined's free online Privacy Focused Data Science Courses, Learn how to participate in Federated Networks and fuel tomorrow's life changing discoveries, & Discover why being a nonprofit foundation is key to OpenMined's Mission. | @madhavajay

DAY 2 | Friday, December 10th

R CONFERENCE

Tommy Jones
In-Q-Tel

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

tidylda: Latent Dirichlet Allocation Using 'tidyverse' Conventions

tidylda implements the Latent Dirichlet Allocation (LDA) topic model in a way that is fast, flexible, and most importantly tidy. Wait. Who needs another LDA implementation though? Tommy will talk us through what makes tidylda so unique and provide examples to stir your imagination on new ways you can use topic modeling in your own work. | @thos_jones

Surabhi Hodigere
Ash Center for Democratic Governance, and Innovation at Harvard Kennedy School

10:25 a.m. - 10:45 p.m. EST

Governance Playbook for Digital Public Goods

Inspired by the open-source movement, Digital Public Goods are not only non-rivalrous, but sharing them across jurisdictions could lower costs, speed adoption, and create standards to facilitate cooperation and trade. However, the joint management of any resource between sovereign entities—particularly of key infrastructure for the maintenance of public goods and services offered by the state—carries with it significant questions of governance. A team of researchers based at the Ash Center within the Harvard Kennedy School are publishing a report that proposes five governance best practices for DPGs—Codifying a Mission, Vision and Value Statement, Drafting a Code of Conduct, Designing Governance Bodies, Ensuring Stakeholder Voice and Representation, and Engaging External Contributors. These five recommendations seek to nurture institutions that will create public value, possess legitimacy, and maintain the necessary support and operational capacity. | @surabhihodigere

Jasmine Ye Han
Bloomberg Industry Group

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

A Data Journalist's R Toolbox

Some reporters chose journalism because they hate numbers. Data journalists are a group of story-tellers who like numbers and can code. And R is one of the most popular languages among them. This talk will introduce how a data journalist uses R, from web scraping, analysis, creating graphics to just automating the boring stuff. | @JasmineHanYe

PolicyViz

A Data Communication Company

Helping you do a better job processing, analyzing, sharing, and presenting your data.

We offer data visualization and presentation skills and design consulting services, as well as training workshops in data visualization and presentation skills. Tools workshops include Excel, PowerPoint, R, & Tableau.

www.policyviz.com

Jorge Luna
Aetna, a CVS Health Company, Analytics & Behavior Change

11:40 a.m. - 12:00 P.M. EST

Hospital analytics approaches to help inform strategies to reduce 30-day hospital readmissions

Reducing preventable hospital readmissions is a national priority for government payers, private payers, providers, and policymakers. Machine learning has emerged as a critical tool in seeking to improve health care, lower costs and generate more value for patients. This short talk will discuss the following key questions that hospitals analytics teams will consider when developing ML and predictive modeling to reduce avoidable readmissions:

· What are the hospital business problems that motivate the need for machine learning and predictive modeling?
· What specific events, outcomes, or quantities need to be predicted?
· What is the specific population for which a prediction of the event/outcome/quantity is needed?
· At what point or stage of the phenomenon should the readmission prediction be made? On pre-admission, admission, 24 hours post-discharge, within 7-days of discharge, etc.?
· If predictions were made available, how will the predictions be used? What are common interventions linked to predictions?
· What are recent trends in hospital analytics & analytic partnership?

Vivian Peng
City of Los Angeles

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

Building Blocks of Design

When we think about design, it's common to jump immediately to thinking about what colors to choose or what graphs to make. The design process starts further back, by getting to know your audience at a foundational level – what motivates, challenges, and inspires them. At a time when we are overloaded by information, and desensitized to numbers, how do we develop data tools and visualizations that create an impact? | @create_self

Boriana P. Pratt
Office of Population Research, Princeton University

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

Running simulations in Parallel in R with doParallel package

Simulations are often run to benchmark a method using data where the results are known or to compare a few methods on a nicely structured (simulated) data. Simulating data in R is not hard. If you have to simulate many different datasets, tweaking some parameters, how to automate such a process to run multiple times and maximize the use of computer or server resources. In this talk I will show how I was able to run multiple simulations at the same time using the doParallel package to run a few R threads simultaneously (from within R) to simulated multiple datasets with genetics data under different scenarios.

Mayari Montes de Oca
Research Scientist

2:00 p.m. - 2:20 p.m. EST

What works to support refugee children? Using BART for impact evaluation

Rigorous evidence of what works to support refugee children is scarce and challenging to attain. During this talk, Mayarí will share with us the strategy that she used to study the impact on children's reading skills, of attending a remedial support program, brought to Syrian refugees by the IRC and NYU. She will share her experience working with machine learning and statistical frameworks that can be helpful to 1) leverage the information available in understudied contexts and to 2) better account for the problem of self-selection into different dosage levels, under a causal framework. In this talk you will also learn about the data challenges of conducting research with vulnerable populations and the R tools that were helpful in the process. | @Mayari_MOca

Benjamin Braun
202 Group

2:25 p.m. - 2:45 p.m. EST

We can't help if we can't communicate: Conveying data science findings to non-experts

The end-user understanding how something works is just as important as the result. Concepts and techniques that come naturally to us as Data Scientists can be totally perplexing to non-expert consumers . . . and that's when critical analysis gets lost in translation. We can't help if we can't communicate, explores how we bridge the gap between Data Science practitioners and the government executives who use our findings to develop policy. The talk will explore two use-cases—one failure and one success—from the speaker's decade-plus of US Federal Government experience to establish a set of best practices in conveying data science findings to non-experts.

Why it's important:
Government needs our help, but we can't help if we can't communicate. If we can't convey what our findings mean and why they are important, essential decisions will be made without the data or analysis needed to back them up.

Participants who attend this session will leave with:
· A set of best practices for ensuring data science findings and products remain accessible, relevant, and actionable
· A deeper understanding of how government executives make decisions
· Where data science fits in to the decision-making process | @Ben_G_Braun

Dr. Abhijit Dasgupta
Georgetown University's Data Science and Analytics

3:15 p.m. - 3:35 p.m. EST

Multilingual pipelines for data analyses: R as glue

We live in a multilingual computational environment, where each language provides certain advantages in terms of developed packages and capabilities. Often, we are faced with utilizing multiple languages to create efficient data analytic workflows. In this talk, I'll describe some experiences in integrating languages utilizing R as the backbone and glue.| @webbedfeet

Sydney Coston
United States Military Academy

3:40 p.m. - 4:00 p.m. EST

Using R Shiny to visualize stem cell treatment data over time

This presentation will introduce an R Shiny app that my partner and I have created to examine the trend of negative outcomes after stem cell treatments over time. The dataset used is from Sloan Kettering Hospital and includes 5 years (20 quarters) of de-identified data from adults and children. The app allows the user to see the proportion of patients who experienced each negative outcome (toxicity) per quarter for the data set of their choice. This app reveals concerning trends in certain toxicities over time.

landeranalytics is a proud member of

R

consortium

THANK YOU TO OUR SPONSORS!

Gold

RStudio®

GEORGETOWN UNIVERSITY
The Graduate School of Arts & Sciences
Master of Science in Data Science & Analytics

Silver

R consortium

Bronze

PolicyViz
A Data Communication Company

Supporting

Pearson

Springer

CRC Press
Taylor & Francis Group
CHAPMAN & HALL

MANNING