Statistical Computing for Data Science

Day 1: Introduction to Wrangling
May 30, 2022

Who is helping you learn data science?

Instructors

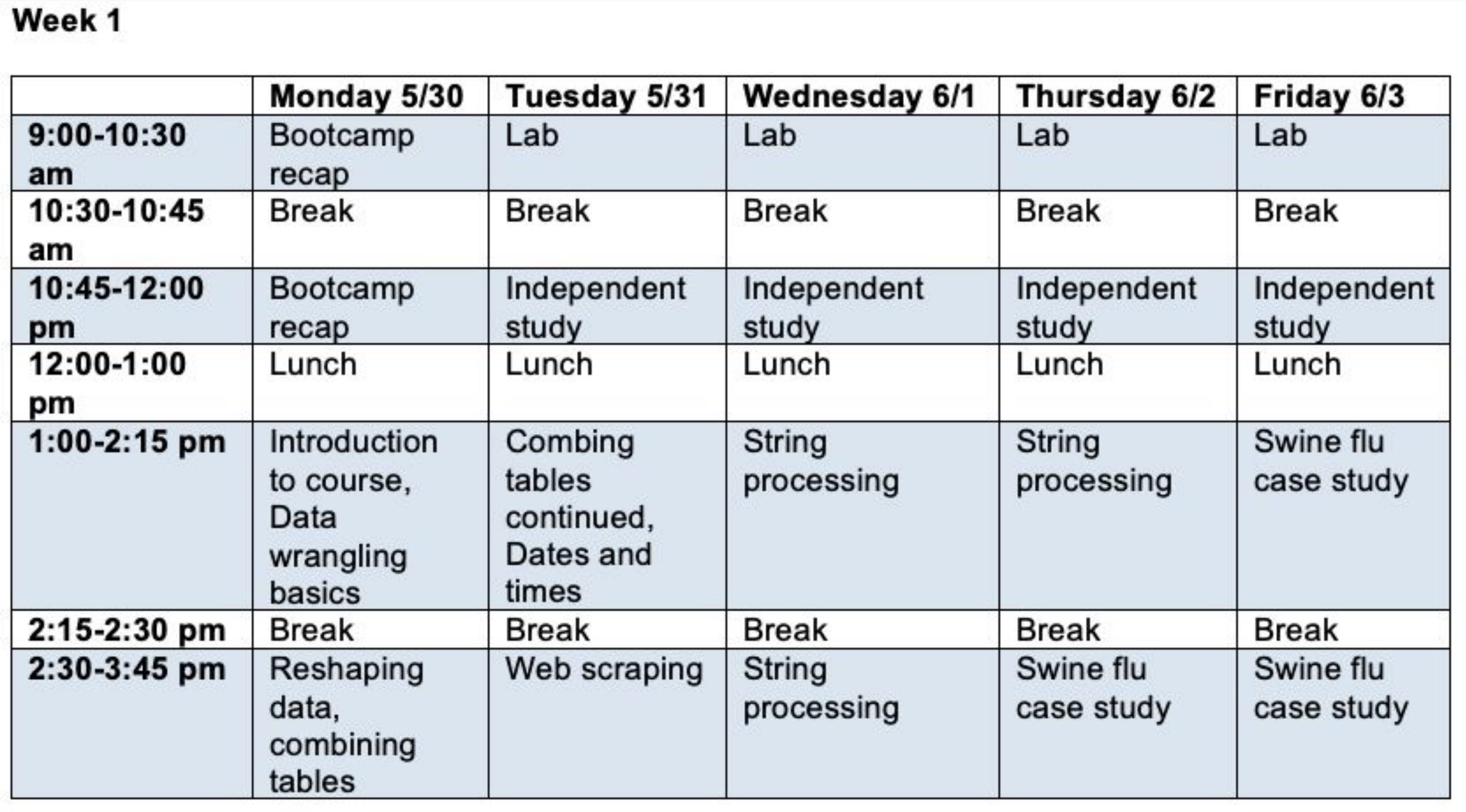
Dr. Heather Mattie

Lecturer on Biostatistics
Co-Director, Health Data Science Master’s Program
Department of Biostatistics
Harvard T.H. Chan School of Public Health
hemattie@hsph.harvard.edu

Dr. Elphas Okango

Lecturer and Researcher in Statistics, Data Science, and Actuarial Science
Stellenbosch University, South Africa
kangphas@gmail.com

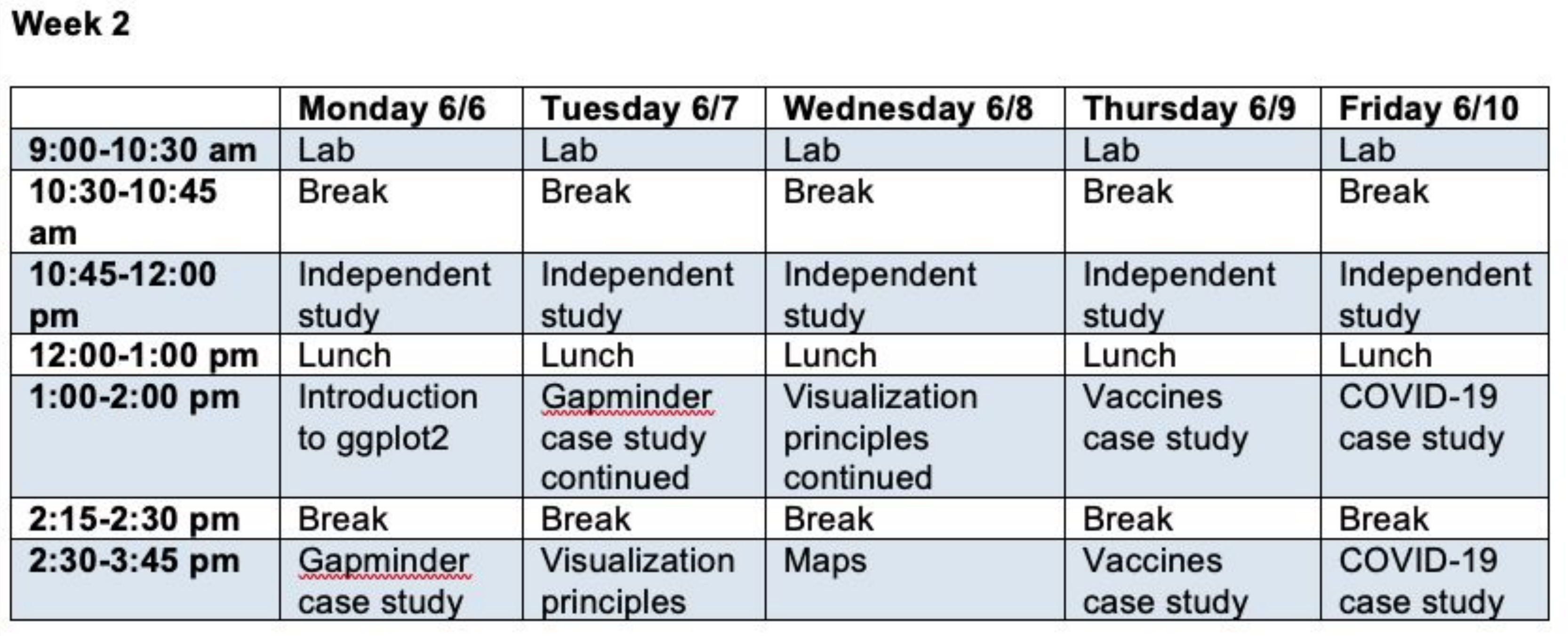
Schedule



Week 1

	Monday 5/30	Tuesday 5/31	Wednesday 6/1	Thursday 6/2	Friday 6/3
9:00-10:30 am	Bootcamp recap	Lab	Lab	Lab	Lab
10:30-10:45 am	Break	Break	Break	Break	Break
10:45-12:00 pm	Bootcamp recap	Independent study	Independent study	Independent study	Independent study
12:00-1:00 pm	Lunch	Lunch	Lunch	Lunch	Lunch
1:00-2:15 pm	Introduction to course, Data wrangling basics	Combing tables continued, Dates and times	String processing	String processing	Swine flu case study
2:15-2:30 pm	Break	Break	Break	Break	Break
2:30-3:45 pm	Reshaping data, combining tables	Web scraping	String processing	Swine flu case study	Swine flu case study

Schedule



Week 2

	Monday 6/6	Tuesday 6/7	Wednesday 6/8	Thursday 6/9	Friday 6/10
9:00-10:30 am	Lab	Lab	Lab	Lab	Lab
10:30-10:45 am	Break	Break	Break	Break	Break
10:45-12:00 pm	Independent study	Independent study	Independent study	Independent study	Independent study
12:00-1:00 pm	Lunch	Lunch	Lunch	Lunch	Lunch
1:00-2:00 pm	Introduction to ggplot2	Gapminder case study continued	Visualization principles continued	Vaccines case study	COVID-19 case study
2:15-2:30 pm	Break	Break	Break	Break	Break
2:30-3:45 pm	Gapminder case study	Visualization principles	Maps	Vaccines case study	COVID-19 case study

What is Data Science?

- “**Big data is not about the data**” – Gary King, Harvard University, making the point that while data is plentiful and easy to collect, **the real value is in the analytics**.

- “For me, data science is a mix of three things: **quantitative analysis** (for the rigor necessary to understand your data), **programming** (so that you can process your data and act on your insights), and **storytelling** (to help others understand what the data means).” - Edwin Chen, Data Scientist and Blogger

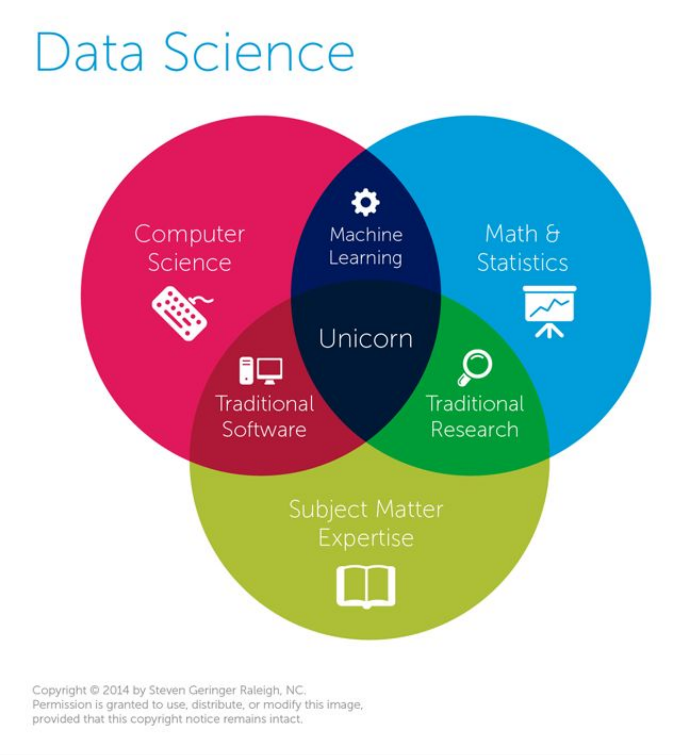
- Data Science is the field of study that combines **domain knowledge**, **expertise**, **programming skills**, and **knowledge of math and statistics** to extract meaningful insights from data - Data Robot

- The goal is to turn data into information and information into **insight** - Carly Florina, former CEO of Hewlett-Packard

- Data Science is a **multidisciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and **insights** from structured and unstructured data - Wikipedia

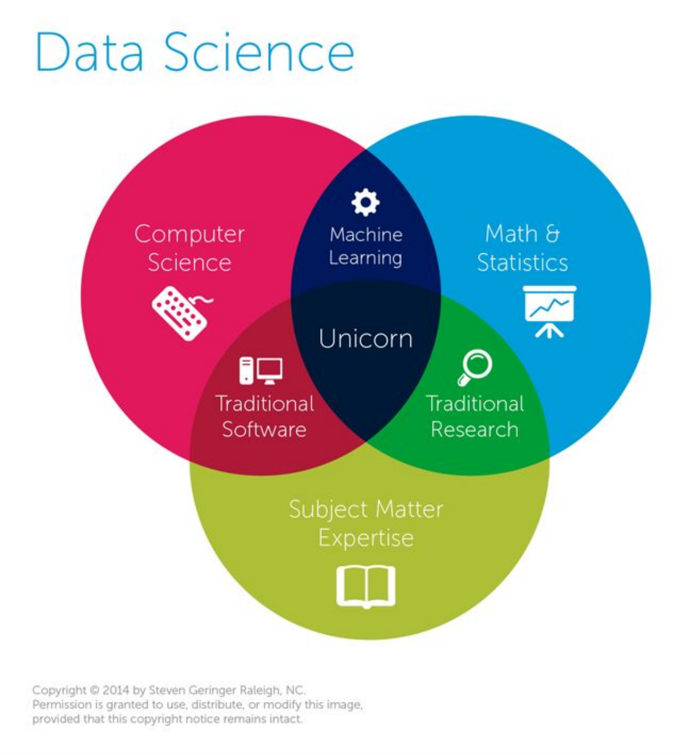
What is Data Science?

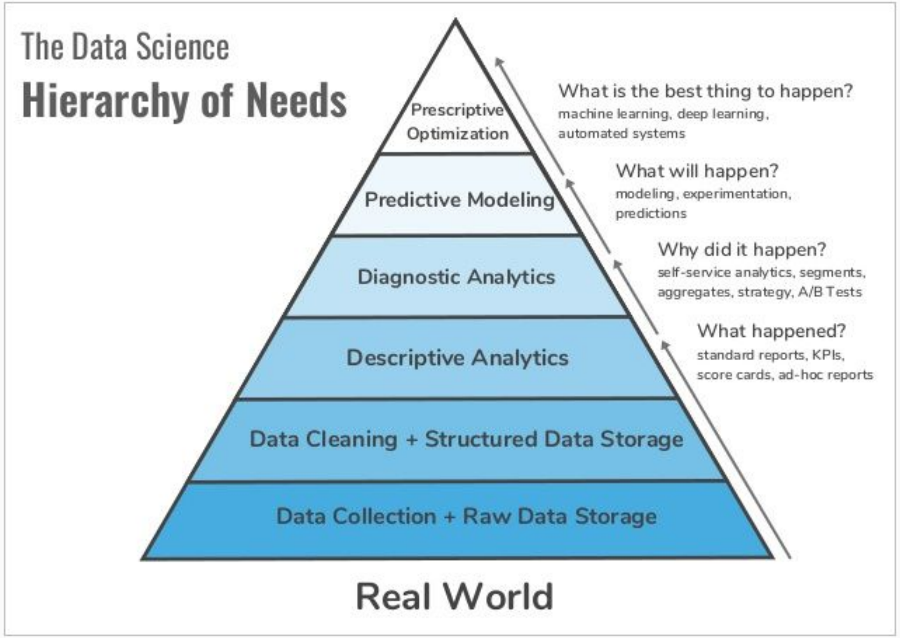
- Data Science is a **multidisciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and **insights** from structured and unstructured data - Wikipedia



What is Data Science?

- Data Science is a **multidisciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and **insights** from structured and unstructured data - Wikipedia

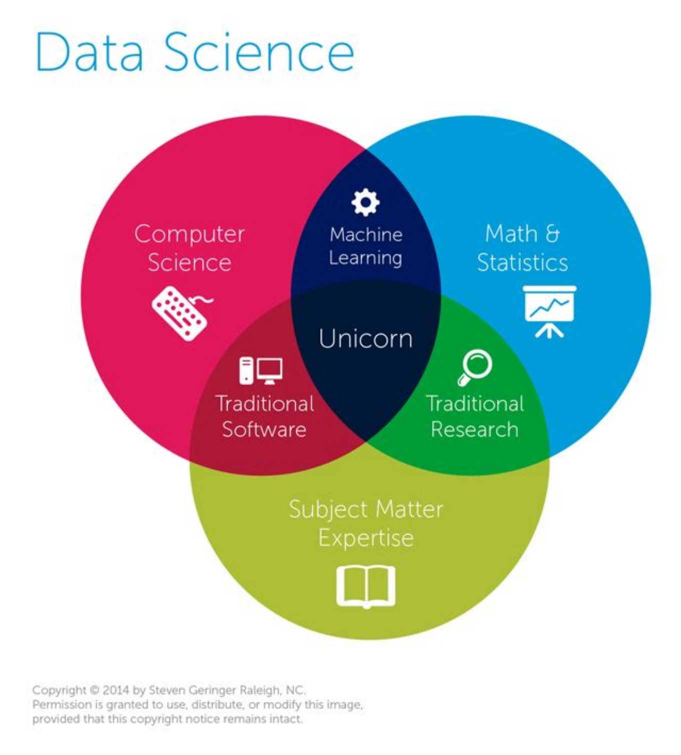




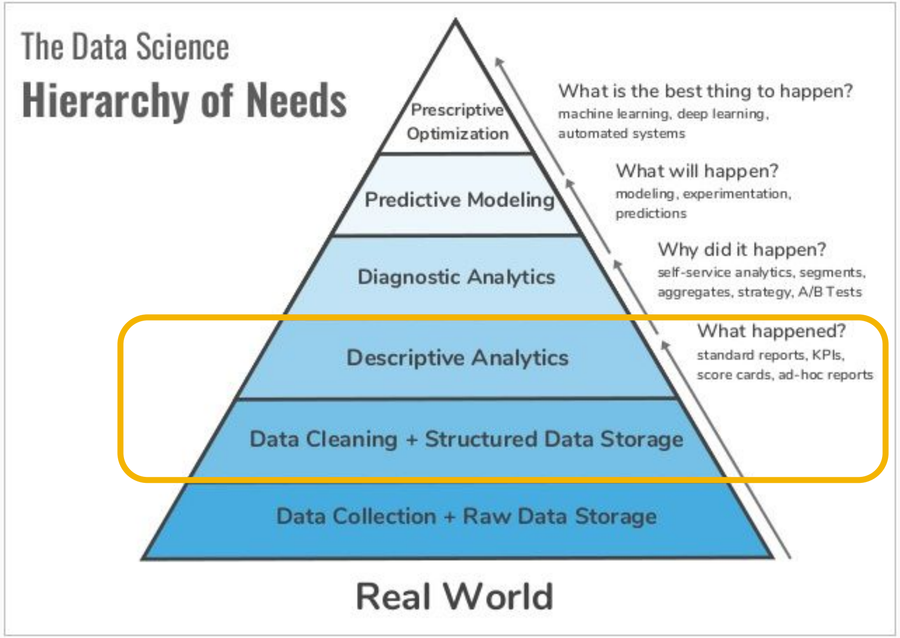
The Data Science
Hierarchy of Needs
Prescriptive Optimization
What is the best thing to happen?
machine learning, deep learning, automated systems
Predictive Modeling
What will happen?
modeling, experimentation, predictions
Diagnostic Analytics
Why did it happen?
self-service analytics, segments, aggregates, strategy, A/B Tests
Descriptive Analytics
What happened?
standard reports, KPIs, score cards, ad-hoc reports
Data Cleaning + Structured Data Storage
Data Collection + Raw Data Storage
Real World

What is Data Science?

- Data Science is a **multidisciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and **insights** from structured and unstructured data - Wikipedia



Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image, provided that this copyright notice remains intact.



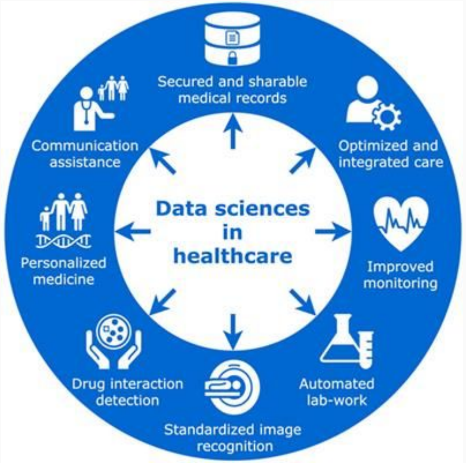
The Data Science
Hierarchy of Needs
Prescriptive Optimization
What is the best thing to happen?
machine learning, deep learning, automated systems
Predictive Modeling
What will happen?
modeling, experimentation, predictions
Diagnostic Analytics
Why did it happen?
self-service analytics, segments, aggregates, strategy, A/B Tests
Descriptive Analytics
What happened?
standard reports, KPIs, score cards, ad-hoc reports
Data Cleaning + Structured Data Storage
Data Collection + Raw Data Storage
Real World

What is Health Data Science?

- “Hiding within those mounds of data is knowledge that could **change the life of a patient**, or change the world.” – Atul Butte, Stanford School of Medicine

- Health Data Science is **data science for health / medical data**
  - Data sets might originate from observational studies, clinical trials, computational biology, electronic medical records, health care claims, genetic and genomic epidemiology, environmental health, digital phenotyping, network science and many other fields

- Precision medicine
- Medical imaging
- Predictive diagnostics
- Natural language processing


Data sciences in healthcare
Secured and sharable medical records
Optimized and integrated care
Improved monitoring
Automated lab-work
Standardized image recognition
Drug interaction detection
Personalized medicine
Communication assistance

What is a Data Scientist?

- “The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning…. Data-driven predictions can succeed—and they can fail. It is when we deny our role in the process that the odds of failure rise. **Before we demand more of our data, we need to demand more of ourselves.**” —Nate Silver, Founder and Editor-in-Chief of FiveThirtyEight

- “**Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.**” —Josh Wills, Director of Data Engineering at Slack

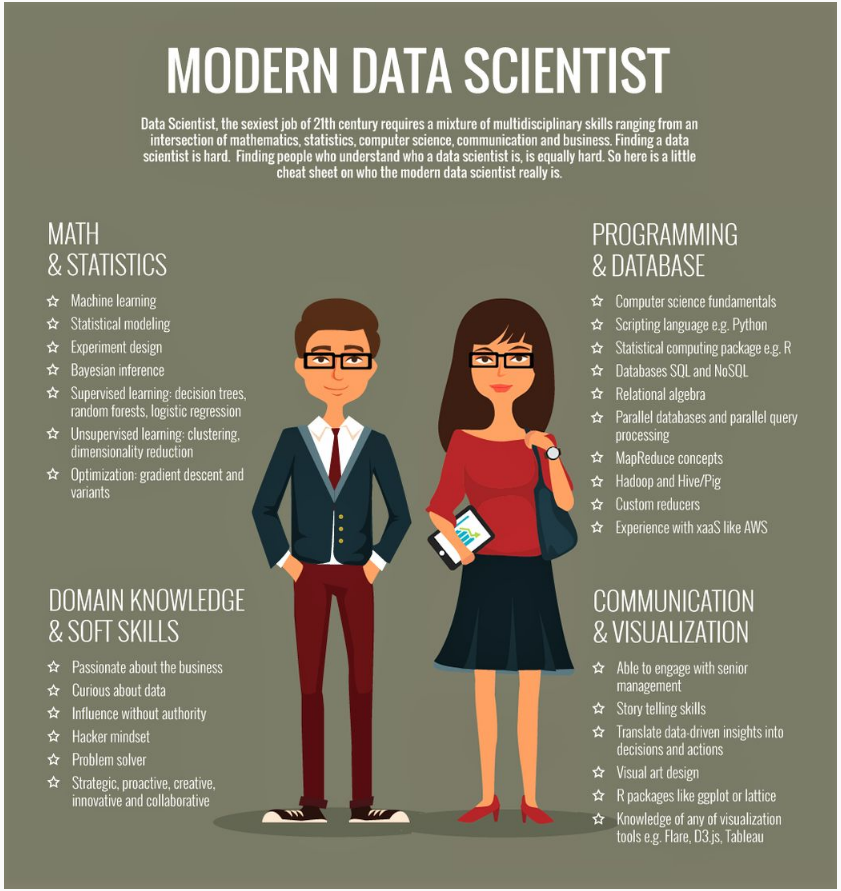
- “**As data scientists, our job is to extract signal from noise.**” —Daniel Tunkelang, Consultant / Advisor

What is a Data Scientist?

- “What sort of personality makes for an effective data scientist? Definitely curiosity…. **The biggest question in data science is ‘Why?’** Why is this happening? If you notice that there’s a pattern, ask, “Why?” Is there something wrong with the data or is this an actual pattern going on? Can we conclude anything from this pattern? A natural curiosity will definitely give you a good foundation.” —Carla Gentry, Data Scientist at Talent Analytics

- “**What makes a good scientist great is creativity with data, skepticism and good communication skills.** Getting all of that together in the same person is difficult—because traditionally, different people follow different paths in their careers—some are more technical, others are more creative and communicative. **A data scientist has to have both.**” —Monica Rogati, Independent Data Science Advisor

What is a Data Scientist?



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

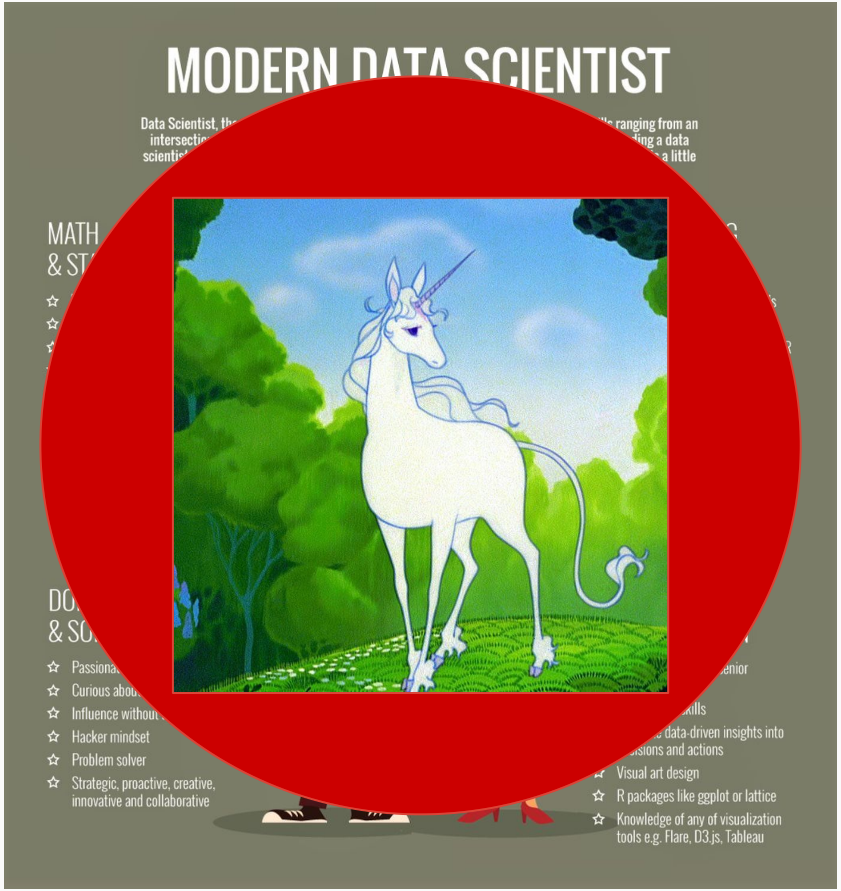
- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

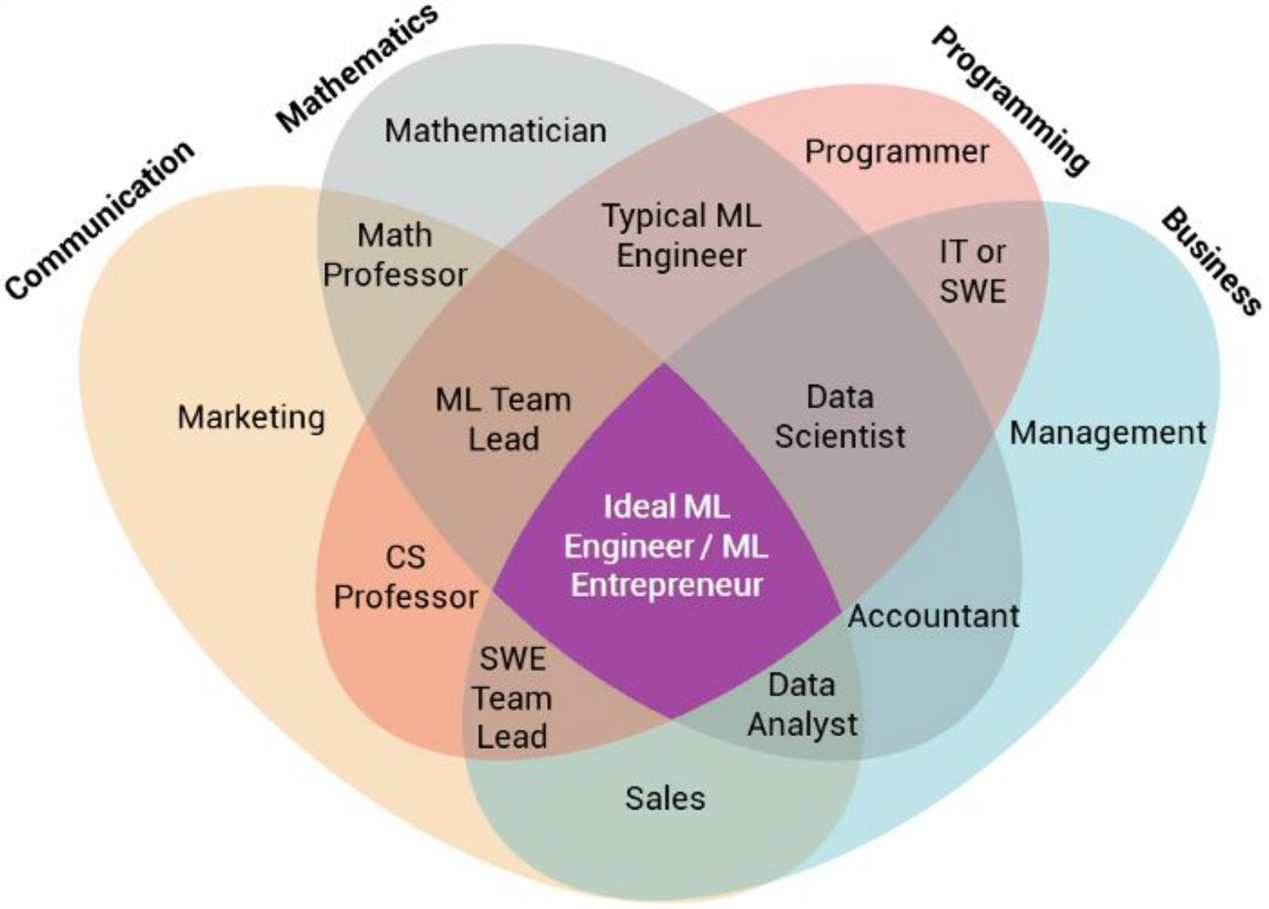
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

What is a Data Scientist?

What is a Data Scientist?



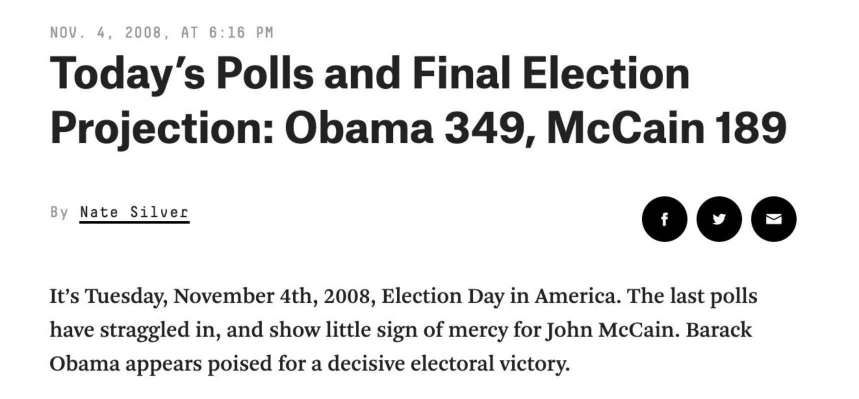
What is a Data Scientist?



Data Science Success Stories

Elections: “Nate Silver won the 2008 election”

- Predicted: 349 to 189, 6.1% difference
- Actual: 365 to 173, 7.2% difference
- While the 2016 election predictions weren’t nearly as close, Nate Silver and 538 were the least wrong by far

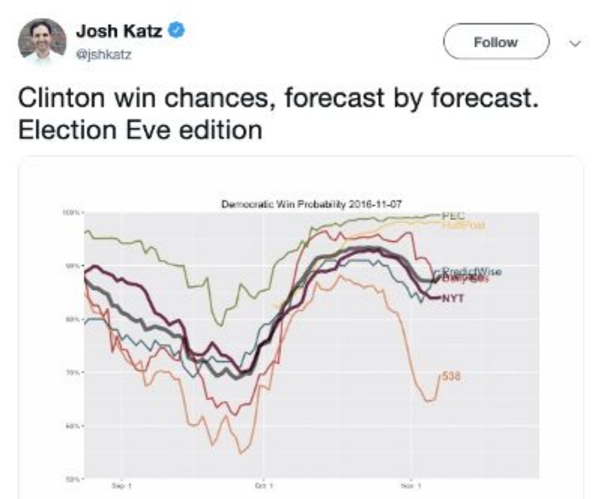


NOV. 4, 2008, AT 6:16 PM

Today’s Polls and Final Election Projection: Obama 349, McCain 189

By Nate Silver

It’s Tuesday, November 4th, 2008, Election Day in America. The last polls have straggled in, and show little sign of mercy for John McCain. Barack Obama appears poised for a decisive electoral victory.

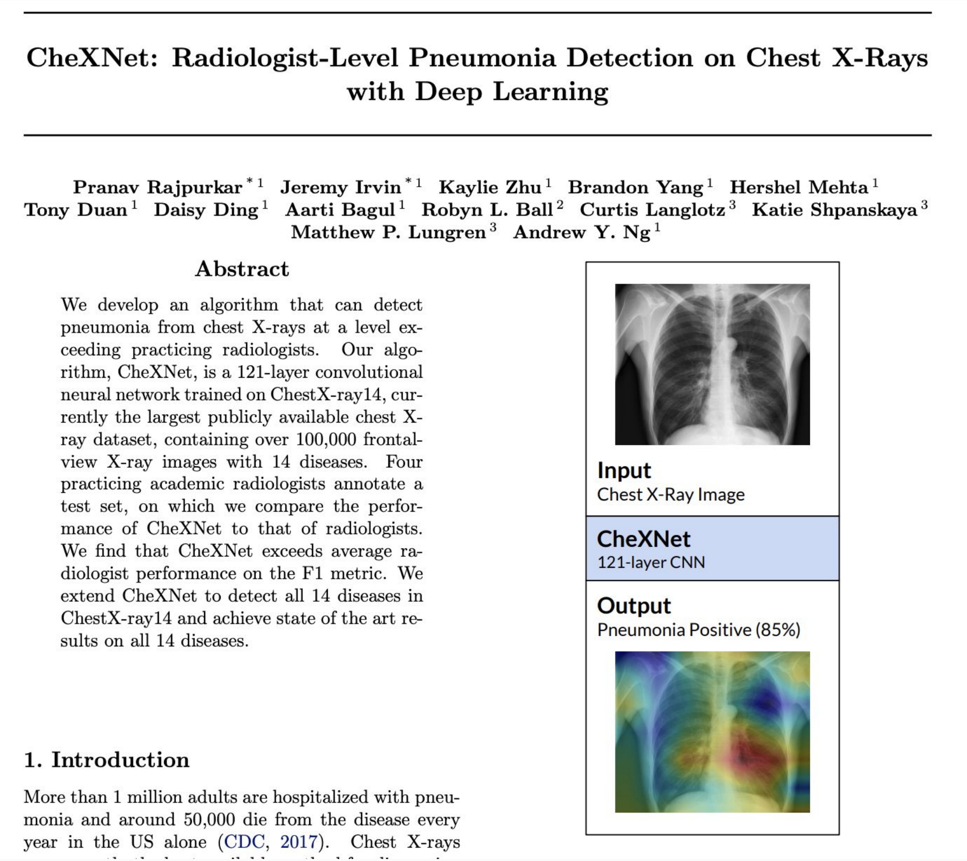


Josh Katz
@jshkatz

Follow

Clinton win chances, forecast by forecast. Election Eve edition

Data Science Success Stories



CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar [*1] Jeremy Irvin [*1] Kaylie Zhu [1] Brandon Yang [1] Hershel Mehta [1]
Tony Duan [1] Daisy Ding [1] Aarti Bagul [1] Robyn L. Ball [2] Curtis Langlotz [3] Katie Shpanskaya [3]
Matthew P. Lungren [3] Andrew Y. Ng [1]

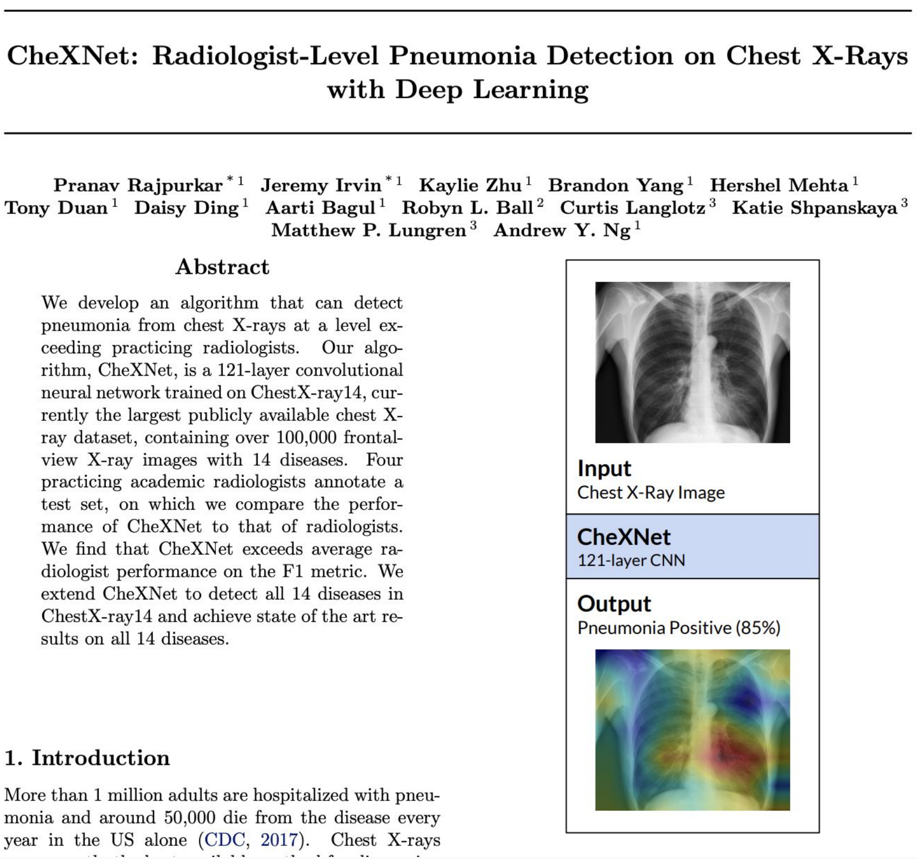
Abstract

We develop an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists. Our algorithm, CheXNet, is a 121-layer convolutional neural network trained on ChestX-ray14, currently the largest publicly available chest X-ray dataset, containing over 100,000 frontal-view X-ray images with 14 diseases. Four practicing academic radiologists annotate a test set, on which we compare the performance of CheXNet to that of radiologists. We find that CheXNet exceeds average radiologist performance on the F1 metric. We extend CheXNet to detect all 14 diseases in ChestX-ray14 and achieve state of the art results on all 14 diseases.

1. Introduction

More than 1 million adults are hospitalized with pneumonia and around 50,000 die from the disease every year in the US alone (CDC, 2017). Chest X-rays

Data Science Success Stories



CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

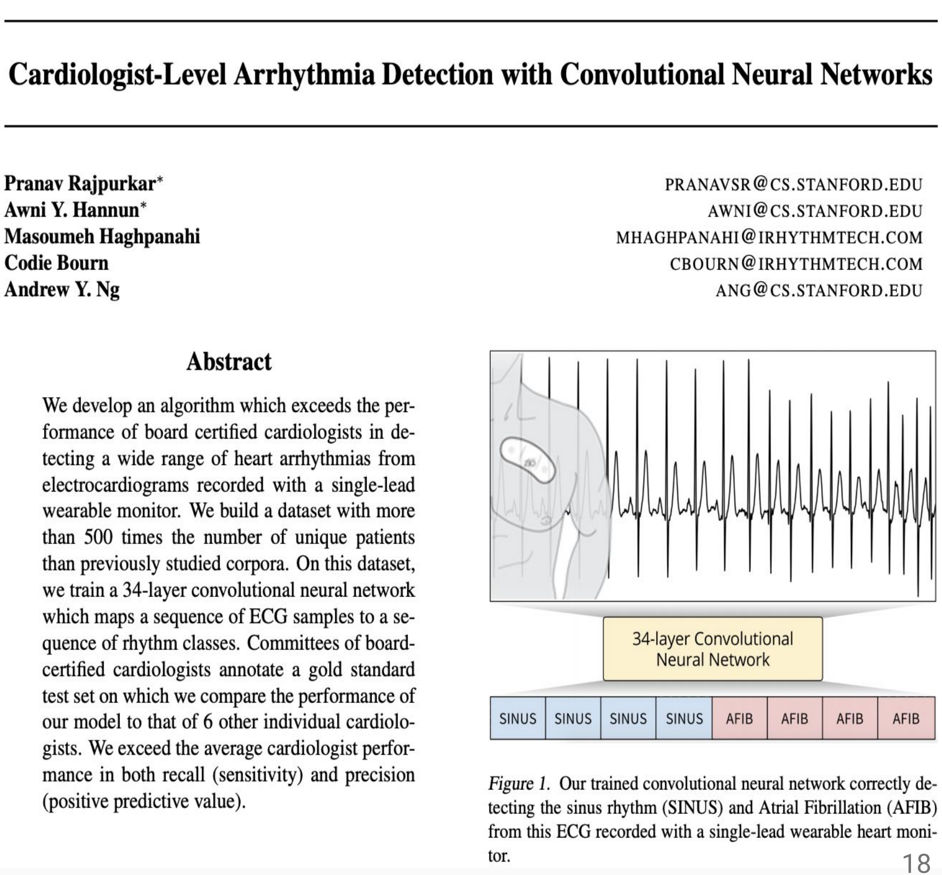
Pranav Rajpurkar [*1] Jeremy Irvin [*1] Kaylie Zhu [1] Brandon Yang [1] Hershel Mehta [1]
Tony Duan [1] Daisy Ding [1] Aarti Bagul [1] Robyn L. Ball [2] Curtis Langlotz [3] Katie Shpanskaya [3]
Matthew P. Lungren [3] Andrew Y. Ng [1]

Abstract

We develop an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists. Our algorithm, CheXNet, is a 121-layer convolutional neural network trained on ChestX-ray14, currently the largest publicly available chest X-ray dataset, containing over 100,000 frontal-view X-ray images with 14 diseases. Four practicing academic radiologists annotate a test set, on which we compare the performance of CheXNet to that of radiologists. We find that CheXNet exceeds average radiologist performance on the F1 metric. We extend CheXNet to detect all 14 diseases in ChestX-ray14 and achieve state of the art results on all 14 diseases.

1. Introduction

More than 1 million adults are hospitalized with pneumonia and around 50,000 die from the disease every year in the US alone (CDC, 2017). Chest X-rays



Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks

Pranav Rajpurkar* PRANAVSR@CS.STANFORD.EDU
Awni Y. Hannun* AWNI@CS.STANFORD.EDU
Masoumeh Haghpanahi MHAGHPANAHI@IRHYTHMTECH.COM
Codie Bourn CBOURN@IRHYTHMTECH.COM
Andrew Y. Ng ANG@CS.STANFORD.EDU

Abstract

We develop an algorithm which exceeds the performance of board certified cardiologists in detecting a wide range of heart arrhythmias from electrocardiograms recorded with a single-lead wearable monitor. We build a dataset with more than 500 times the number of unique patients than previously studied corpora. On this dataset, we train a 34-layer convolutional neural network which maps a sequence of ECG samples to a sequence of rhythm classes. Committees of board-certified cardiologists annotate a gold standard test set on which we compare the performance of our model to that of 6 other individual cardiologists. We exceed the average cardiologist performance in both recall (sensitivity) and precision (positive predictive value).

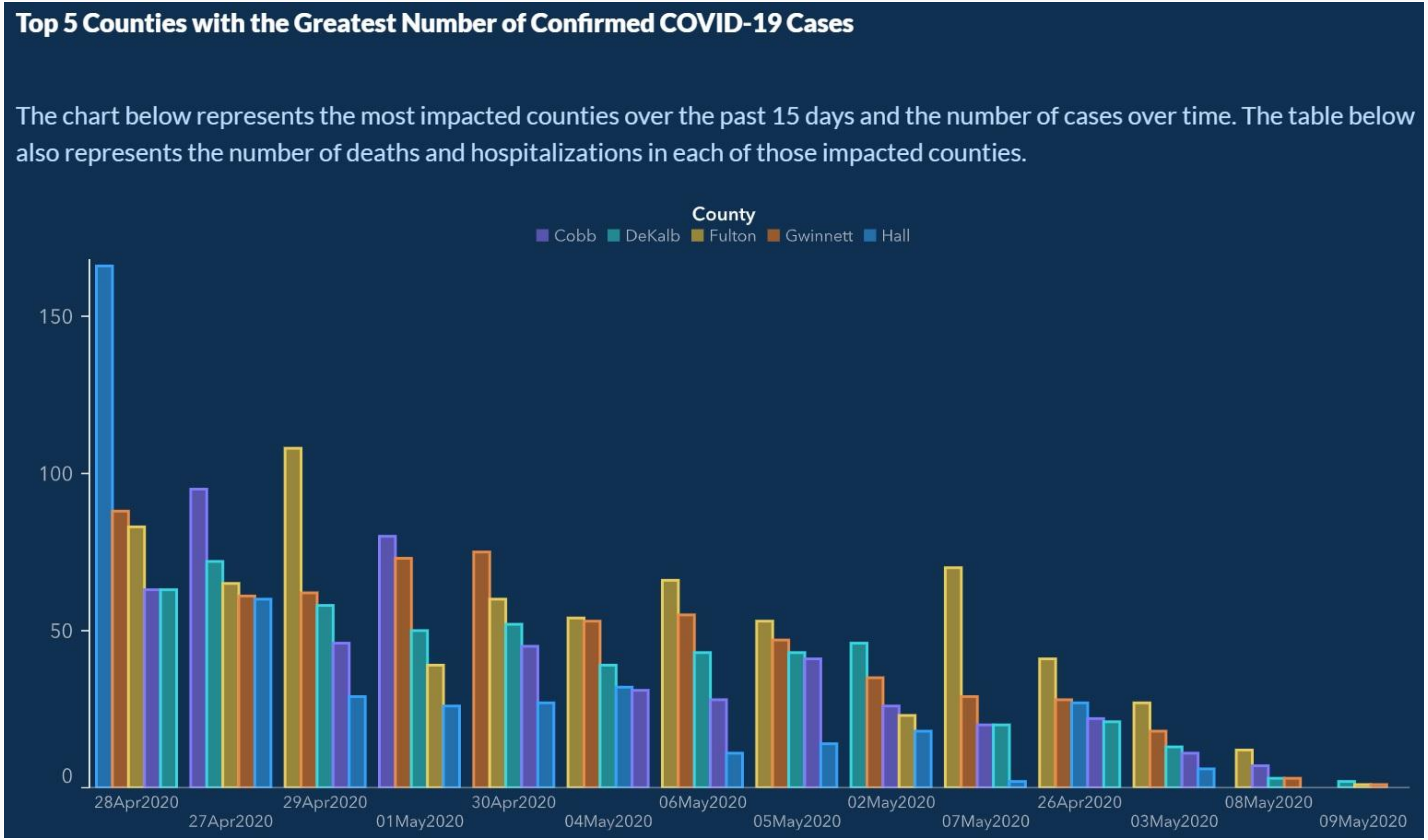
Figure 1. Our trained convolutional neural network correctly detecting the sinus rhythm (SINUS) and Atrial Fibrillation (AFIB) from this ECG recorded with a single-lead wearable heart monitor.

Data Science Success Stories

Many more examples

- Personalized medicine
- Medical diagnostics
- Spell checkers
- Natural language processing
- Language translators
- Etc.

When Data Science Goes Wrong

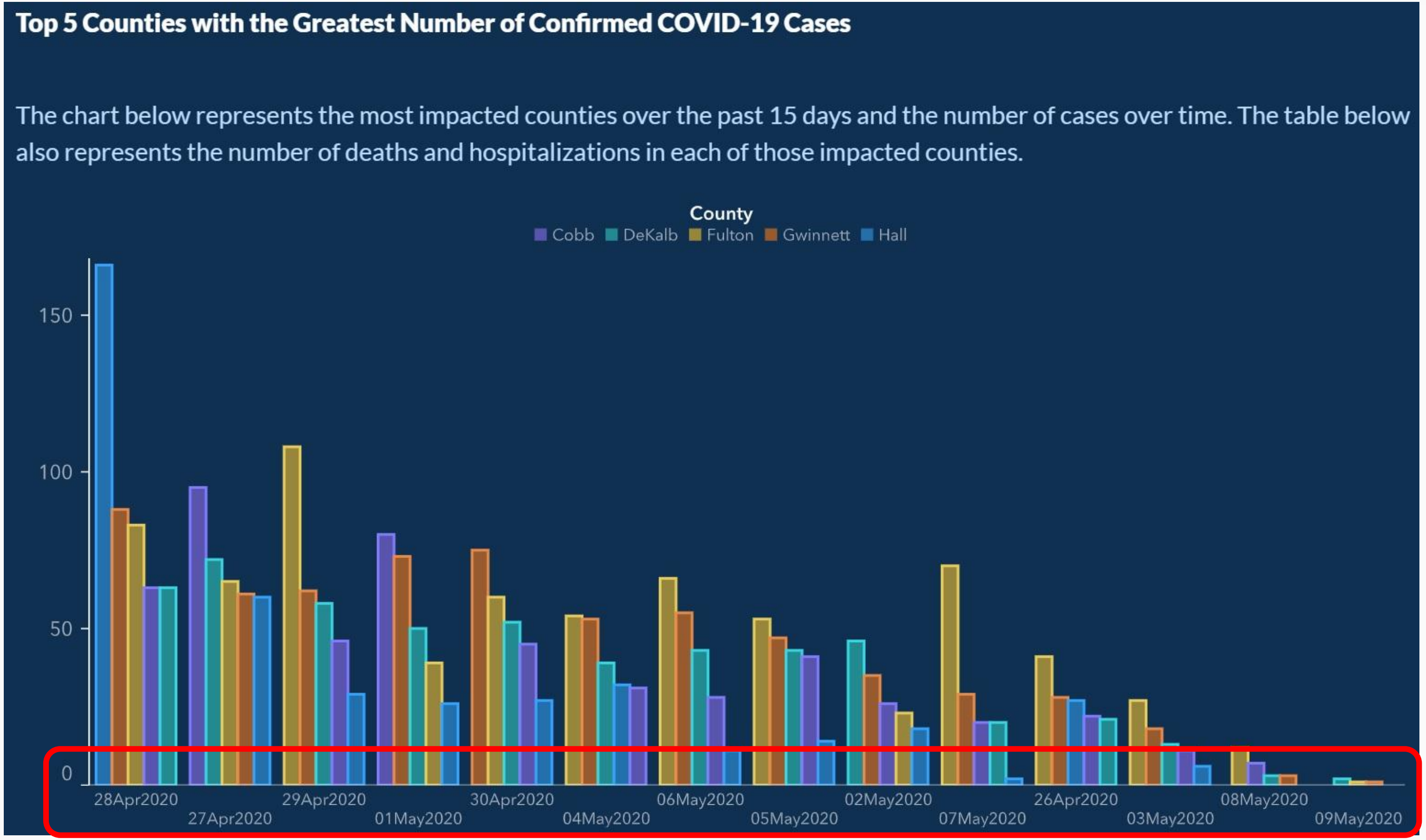


Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

https://www.r-bloggers.com/ordering-bars-within-their-clumps-in-a-bar-chart-by-ellis2013nz/

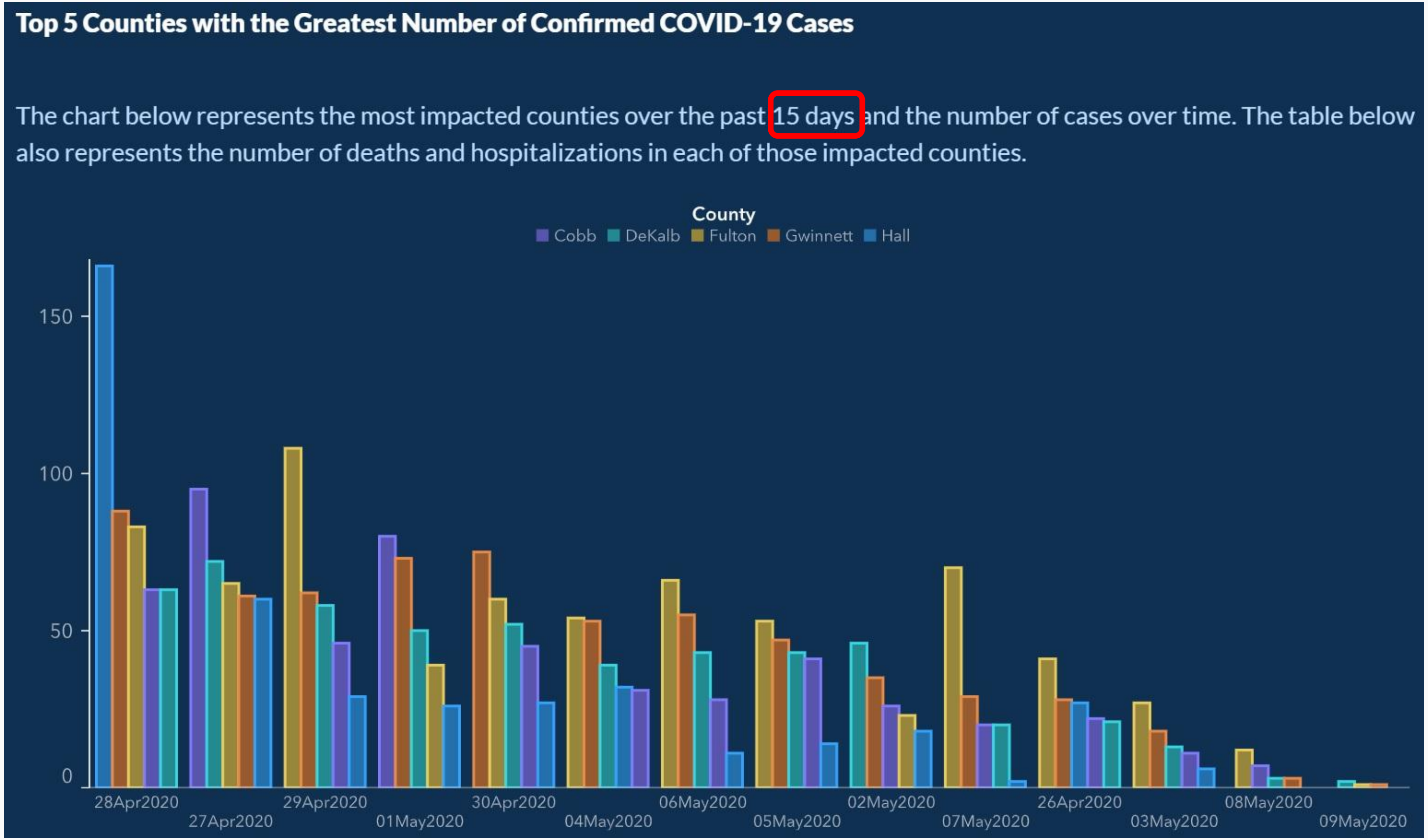
When Data Science Goes Wrong



Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

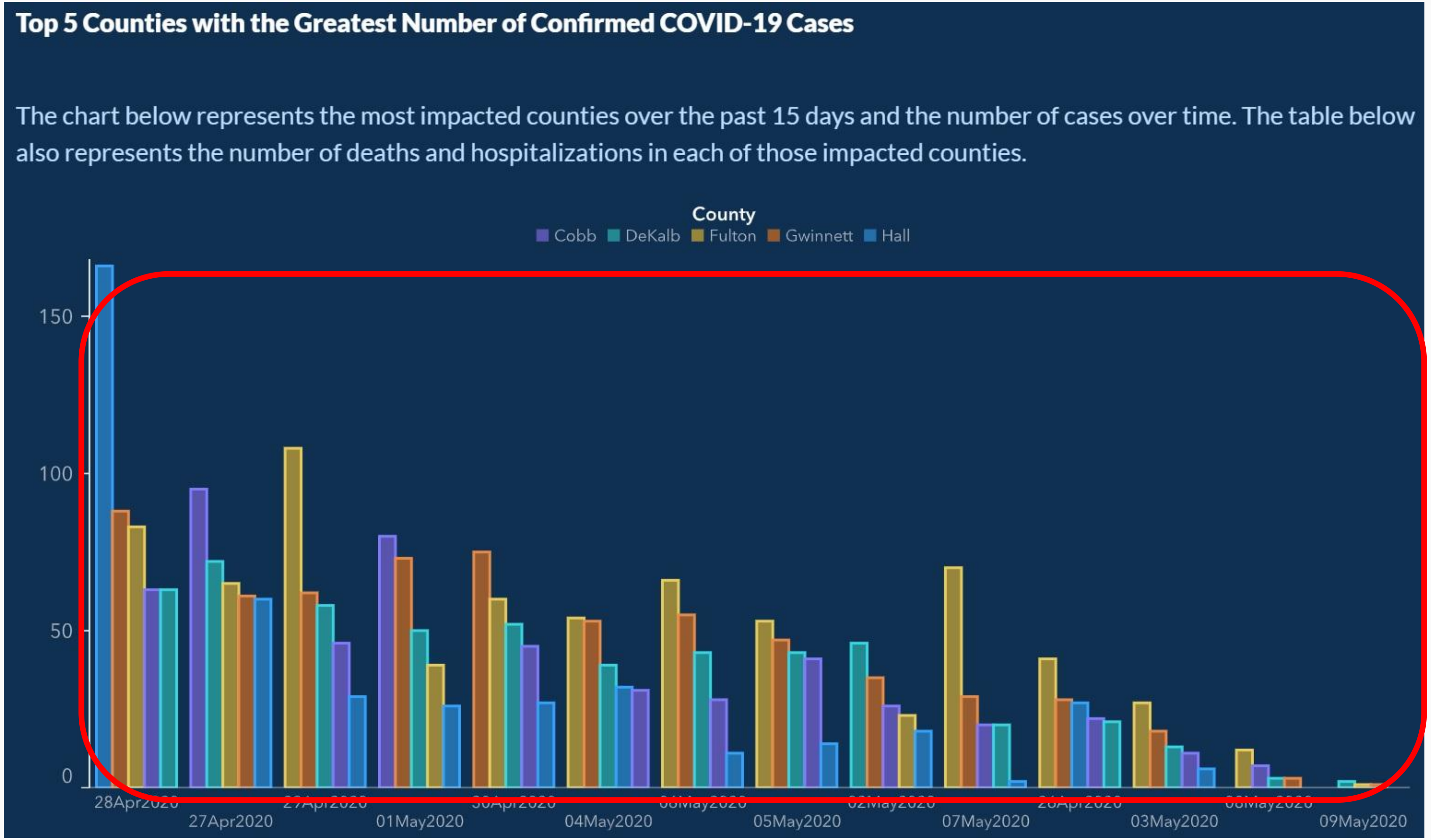
When Data Science Goes Wrong



Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

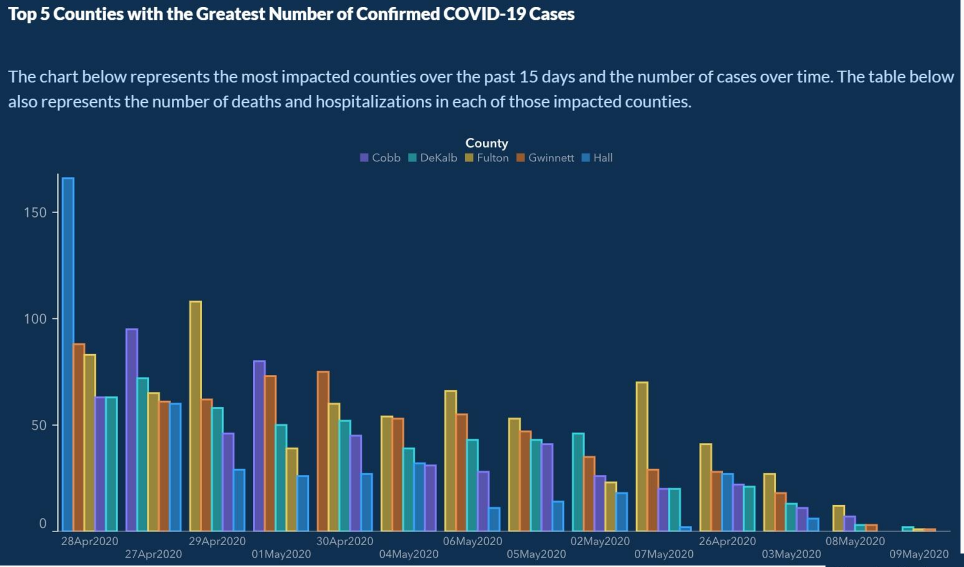
The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

When Data Science Goes Wrong



Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

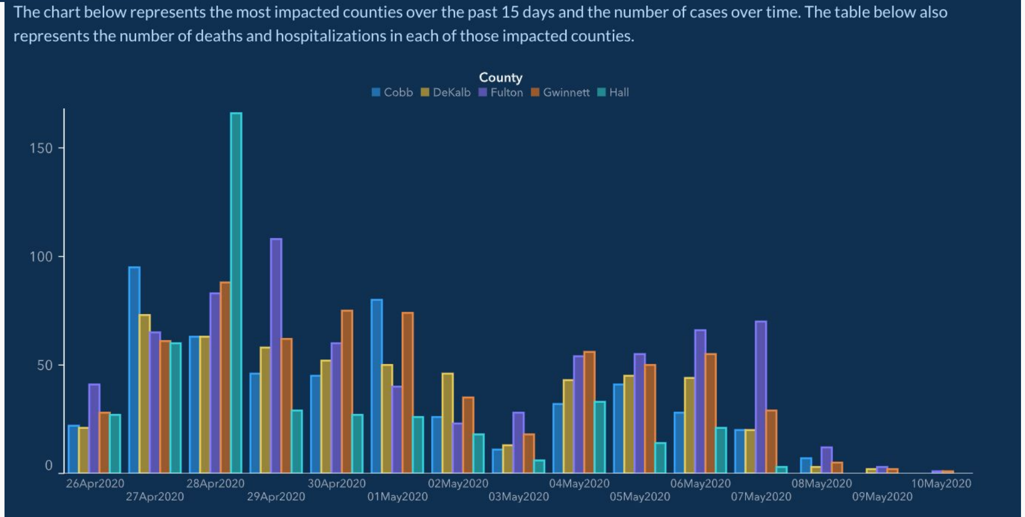
The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.



Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

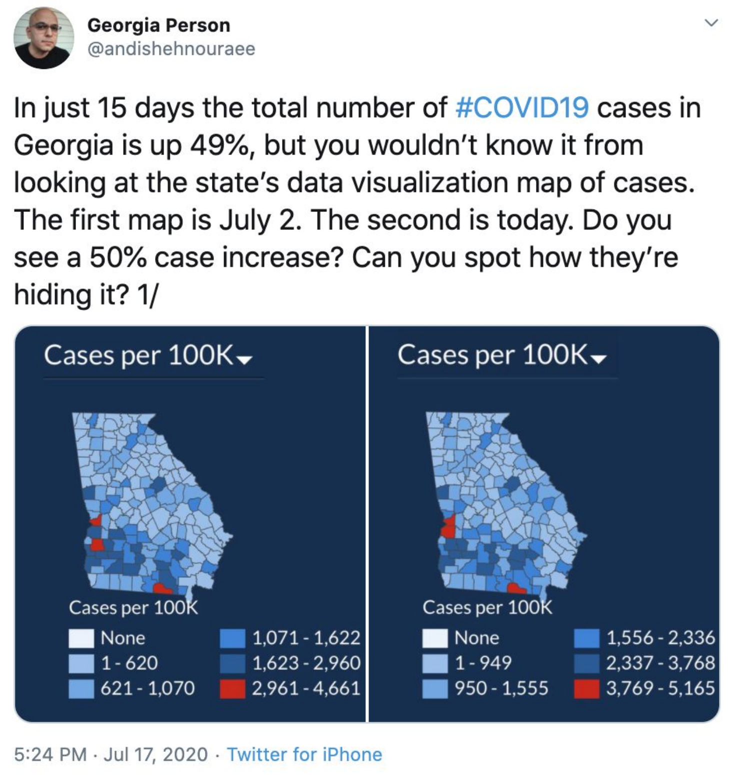
The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

Corrected version



The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

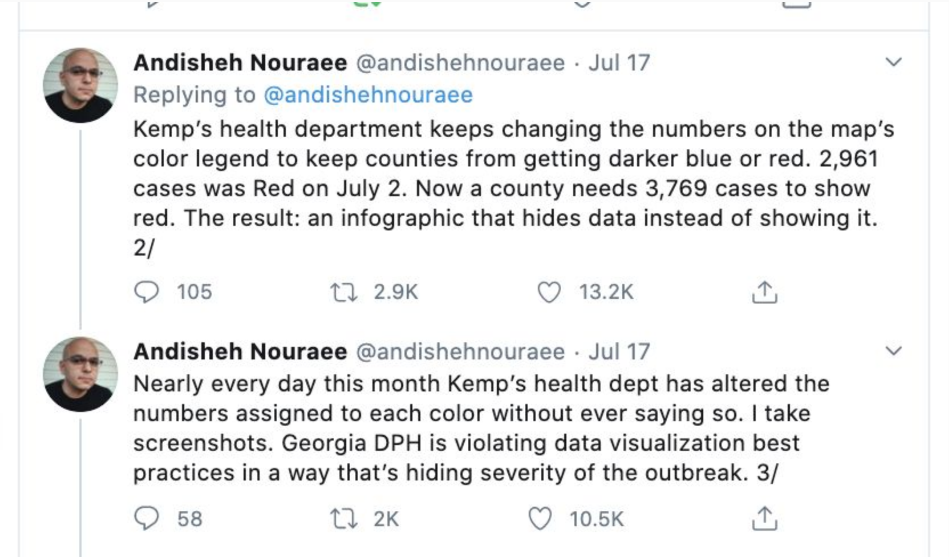
When Data Science Goes Wrong



Georgia Person
@andishehnouraee

In just 15 days the total number of #COVID19 cases in Georgia is up 49%, but you wouldn't know it from looking at the state's data visualization map of cases. The first map is July 2. The second is today. Do you see a 50% case increase? Can you spot how they're hiding it? 1/

5:24 PM · Jul 17, 2020 · Twitter for iPhone



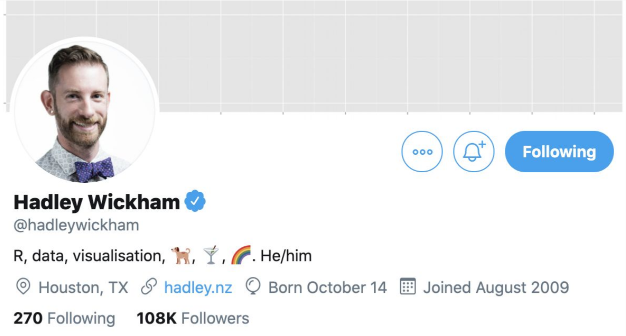
Andisheh Nouraee @andishehnouraee · Jul 17
Replying to @andishehnouraee
Kemp's health department keeps changing the numbers on the map's color legend to keep counties from getting darker blue or red. 2,961 cases was Red on July 2. Now a county needs 3,769 cases to show red. The result: an infographic that hides data instead of showing it. 2/

Andisheh Nouraee @andishehnouraee · Jul 17
Nearly every day this month Kemp's health dept has altered the numbers assigned to each color without ever saying so. I take screenshots. Georgia DPH is violating data visualization best practices in a way that's hiding severity of the outbreak. 3/

https://twitter.com/andishehnouraee/status/1284237474831761408

R/RStudio Resources

- RStudio website
- Hadley Wickham's Twitter
- Hadley Wickham's GitHub
- Hadley Wickham's Shiny
- R for Data Science
- Introduction to Data Science
- Julia Siege's Twitter
- David Robinson's Twitter
- Text Mining with R

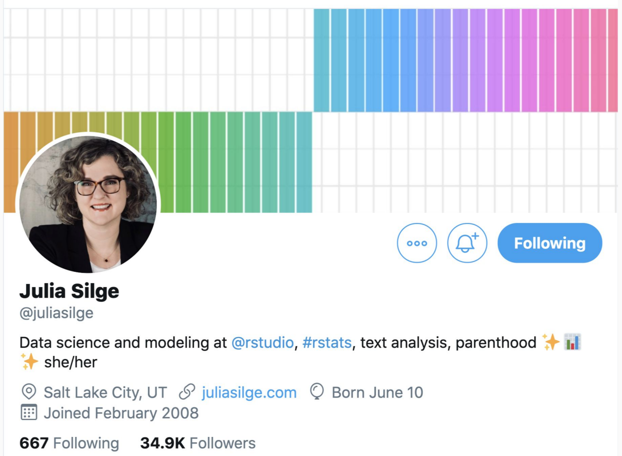


Hadley Wickham
@hadleywickham

R, data, visualisation, 🐕, 🍸, 🌈. He/him

Houston, TX hadley.nz Born October 14 Joined August 2009

270 Following **108K** Followers

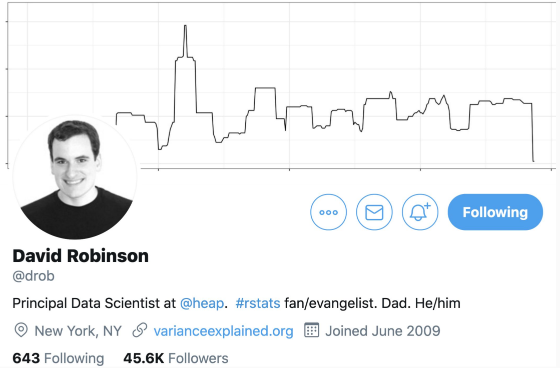


Julia Silge
@juliasilge

Data science and modeling at @rstudio, #rstats, text analysis, parenthood ✨📊✨ she/her

Salt Lake City, UT juliasilge.com Born June 10
Joined February 2008

667 Following **34.9K** Followers



David Robinson
@drob

Principal Data Scientist at @heap. #rstats fan/evangelist. Dad. He/him

New York, NY varianceexplained.org Joined June 2009

643 Following **45.6K** Followers

Lectures and Labs

- Real-world/public health/medical focus
- Scrape and wrangle/clean messy data
- Explore data
- Visualize data
- Communicate results
- Will be written in R using RMarkdown
- All course material is available in this GitHub repository