

BST 219

Core Principles of Data Science

Lecture 1: Introduction to Course, R, & RMarkdown

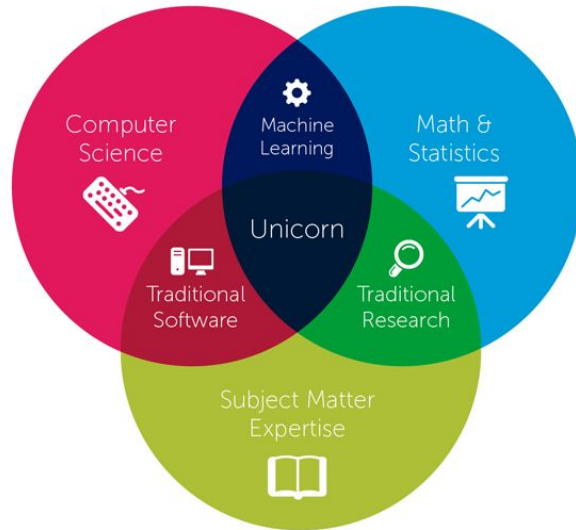
What is Data Science?

- “**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

Data Science

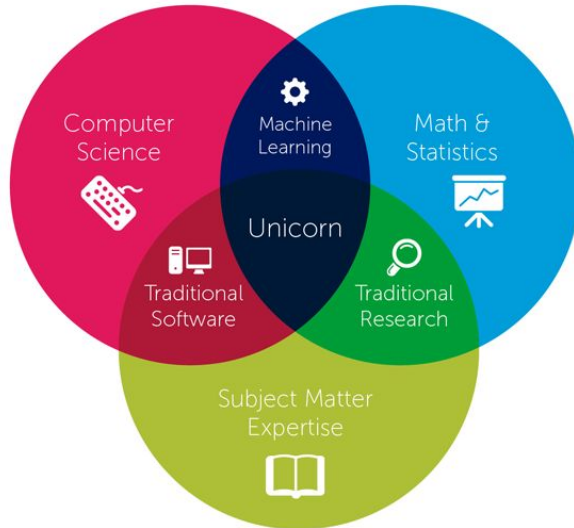


Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image,
provided that this copyright notice remains intact.

What is Data Science?

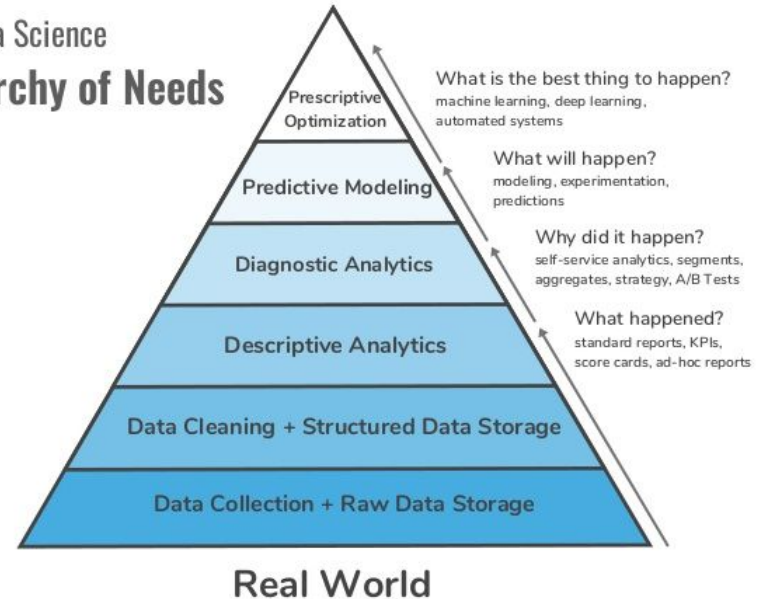
- Data Science is a **multidisciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and **insights** from structured and unstructured data - Wikipedia

Data Science



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

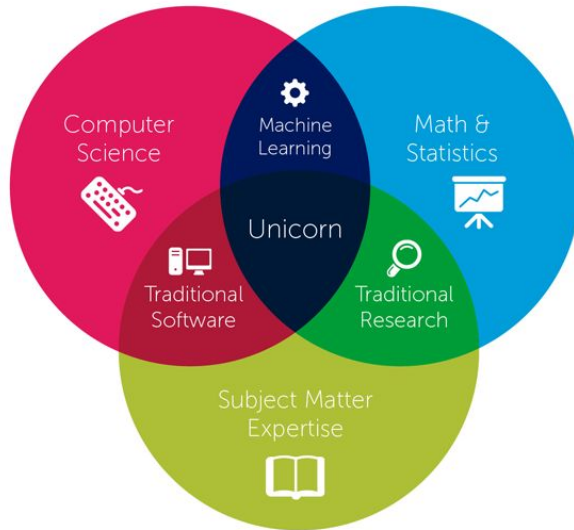
The Data Science Hierarchy of Needs



What is Data Science?

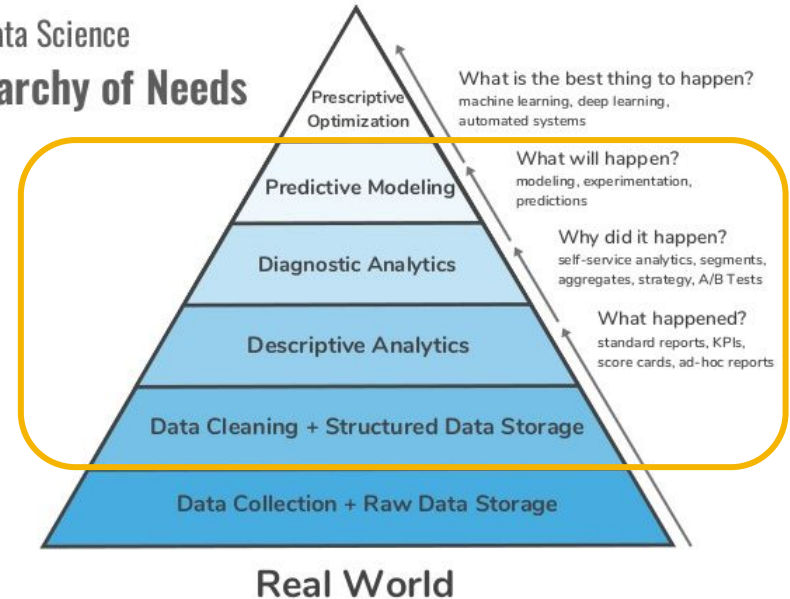
- Data Science is a **multidisciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and **insights** from structured and unstructured data - Wikipedia

Data Science



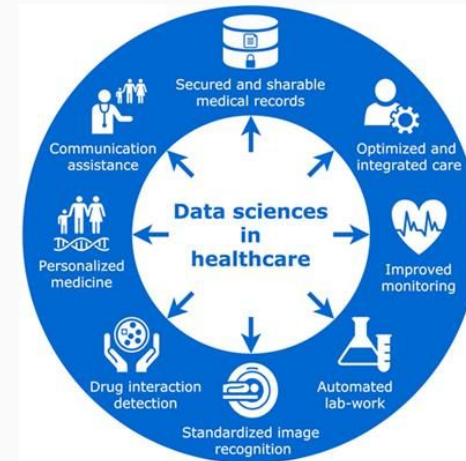
Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image,
provided that this copyright notice remains intact.

The Data Science Hierarchy of Needs



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



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 21st 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

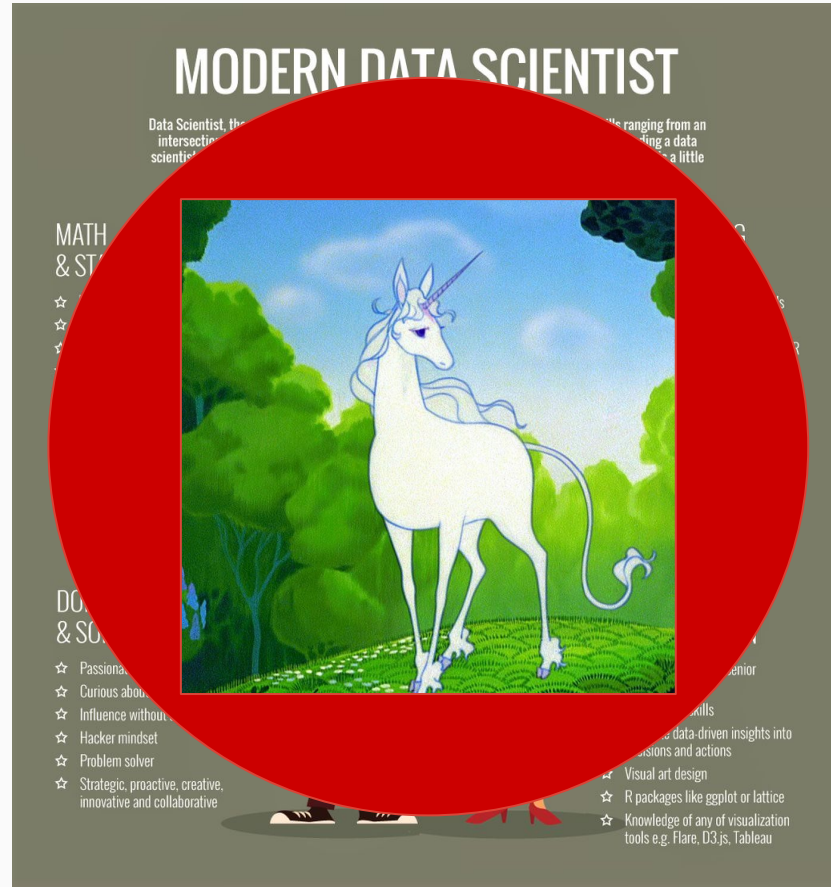
COMMUNICATION & VISUALIZATION

- ☆ 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

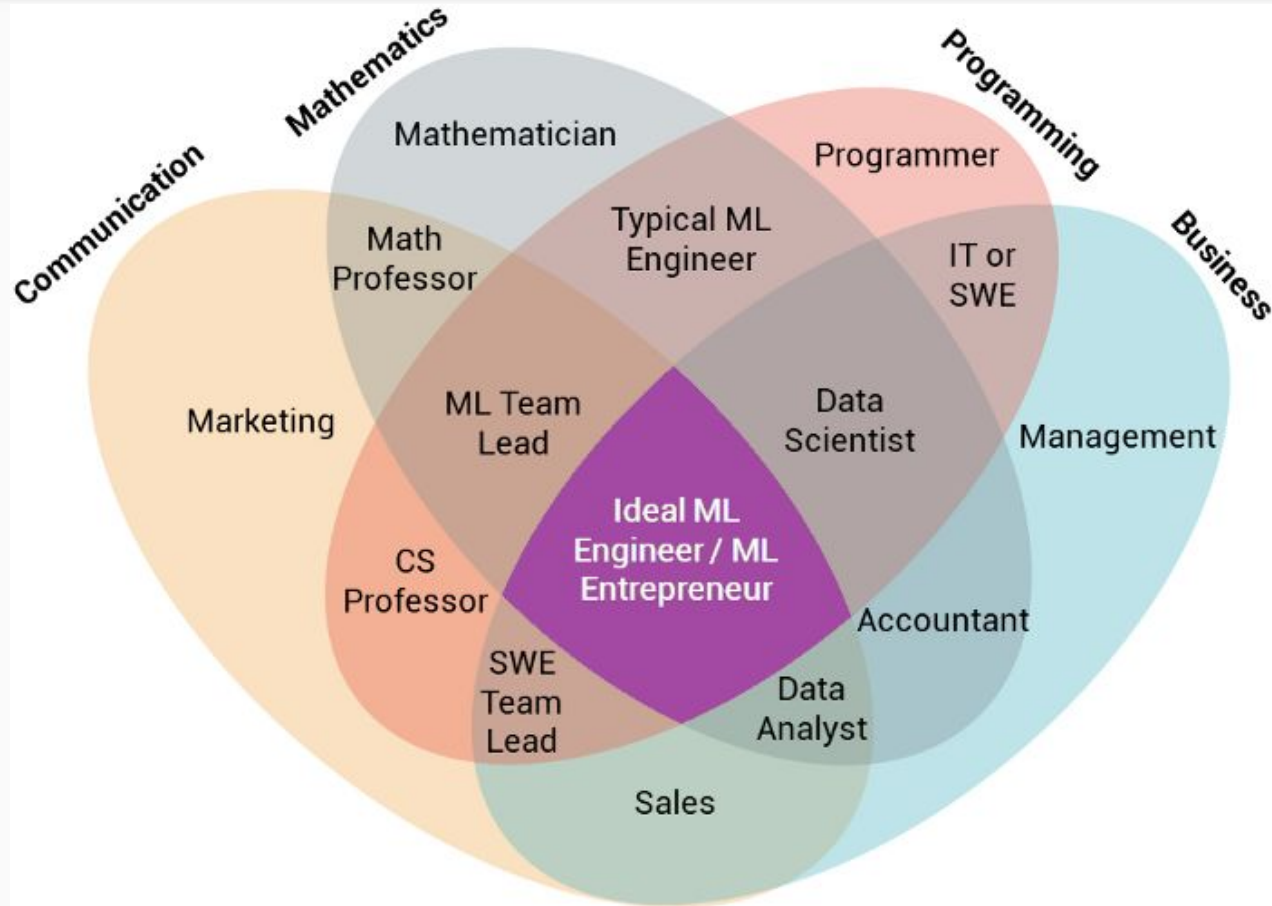
What is a Data Scientist?



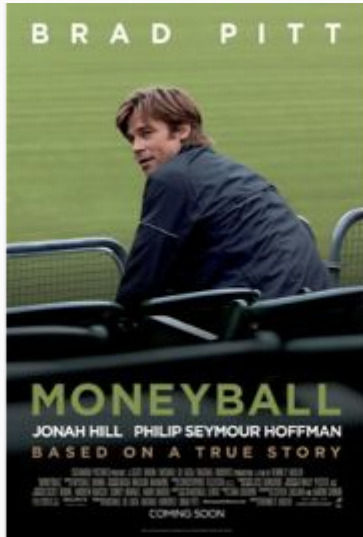
What is a Data Scientist?



What is a Data Scientist?



Data Science Success Stories



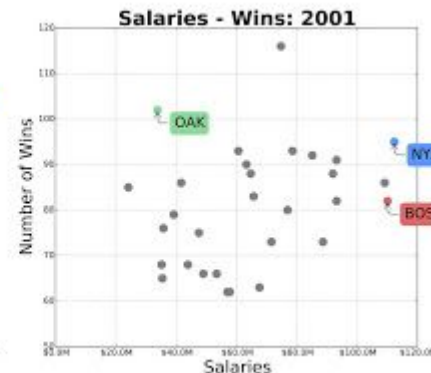
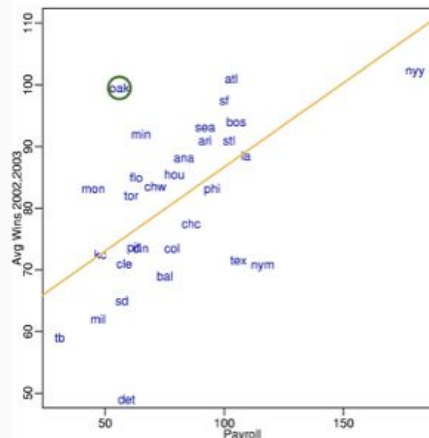
Real Data Scientist
(Paul DePodesta)



Hollywood Data Scientist

Data Science Success Stories

- Starting around 2001, the Oakland A's picked players that scouts thought were no good, but data said otherwise
- Ended up in the playoffs with one of the lowest budgets in baseball



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](#)



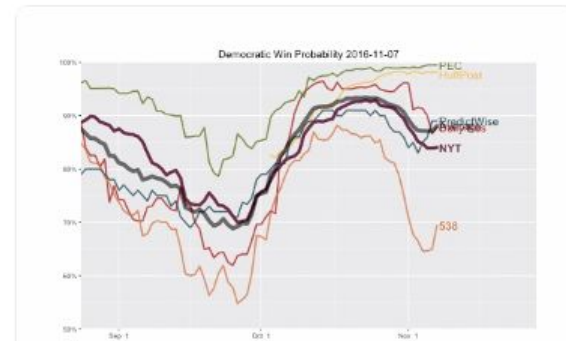
It's Tuesday, November 4th, 2008, Election Day in America. The last polls have straggled in, and show little sign of mercy for John McCain. Barack Obama appears poised for a decisive electoral victory.



Josh Katz
@jshkatz

Follow

Clinton win chances, forecast by forecast. Election Eve edition



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

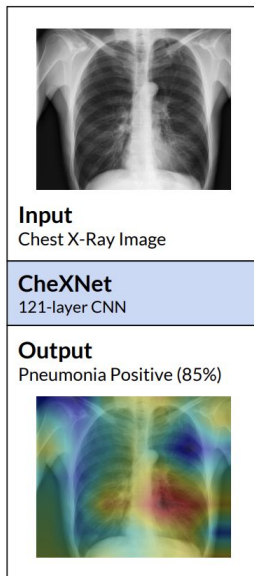
Pranav Rajpurkar^{*1} Jeremy Irvin^{*1} Kaylie Zhu¹ Brandon Yang¹ Hershel Mehta¹
Tony Duan¹ Daisy Ding¹ Aarti Bagul¹ Robyn L. Ball² Curtis Langlotz³ Katie Shpanskaya³
Matthew P. Lungren³ Andrew Y. Ng¹

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



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

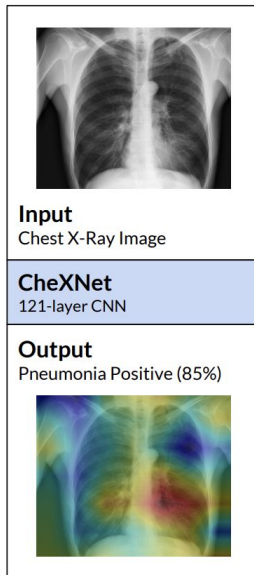
Pranav Rajpurkar^{*1} Jeremy Irvin^{*1} Kaylie Zhu¹ Brandon Yang¹ Hershel Mehta¹
Tony Duan¹ Daisy Ding¹ Aarti Bagul¹ Robyn L. Ball² Curtis Langlotz³ Katie Shpanskaya³
Matthew P. Lungren³ Andrew Y. Ng¹

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^{*}
Awni Y. Hannun^{*}
Masoumeh Haghpanahi
Codie Bourn
Andrew Y. Ng

PRANAVSR@CS.STANFORD.EDU
AWNI@CS.STANFORD.EDU
MHAGHPANAHI@IRHYTHMTECH.COM
CBOURN@IRHYTHMTECH.COM
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).

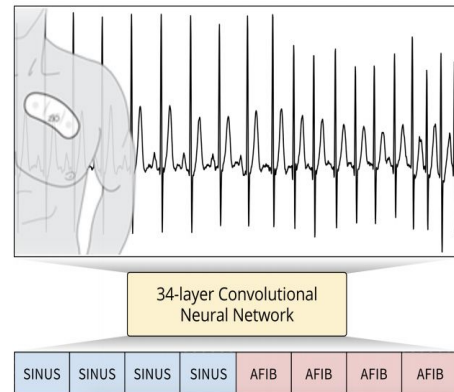


Figure 1. Our trained convolutional neural network correctly detecting the sinus rhythm (SINUS) and Atrial Fibrillation (AFIB) from this ECG recorded with a single-lead wearable heart monitor.

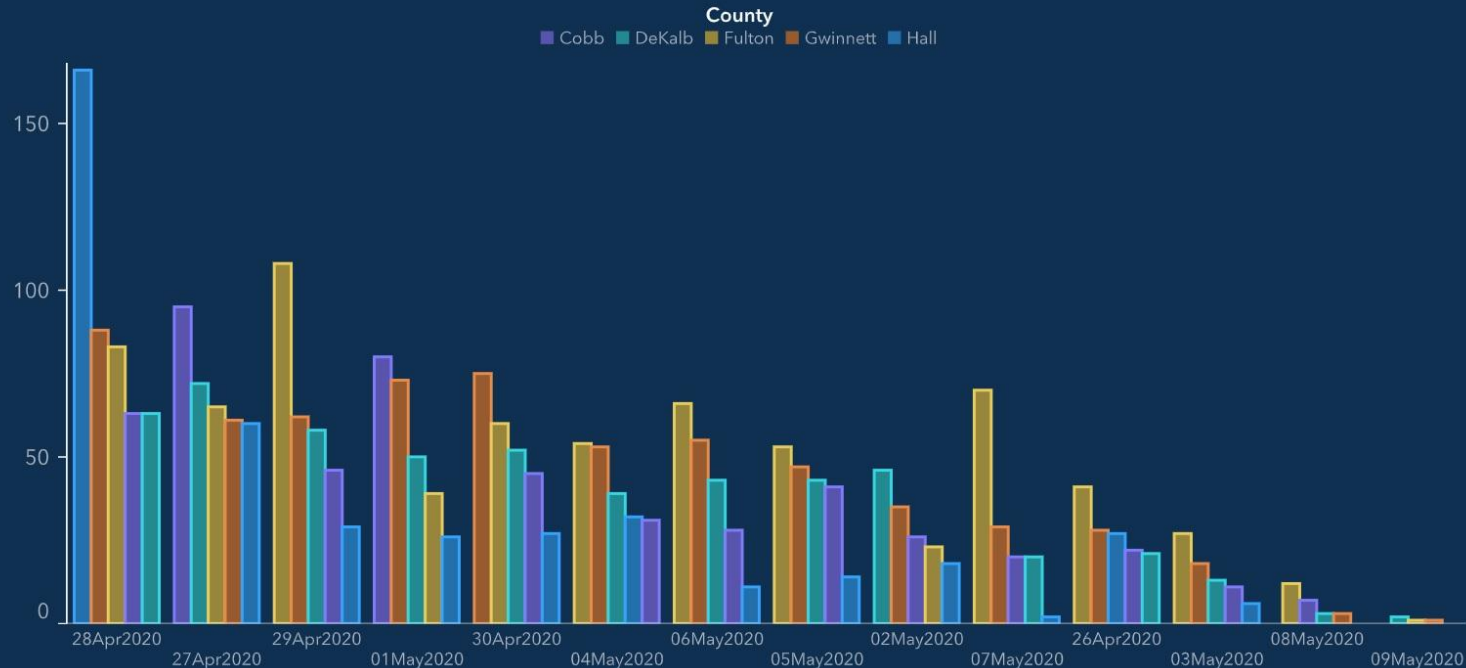
Many more examples

- 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

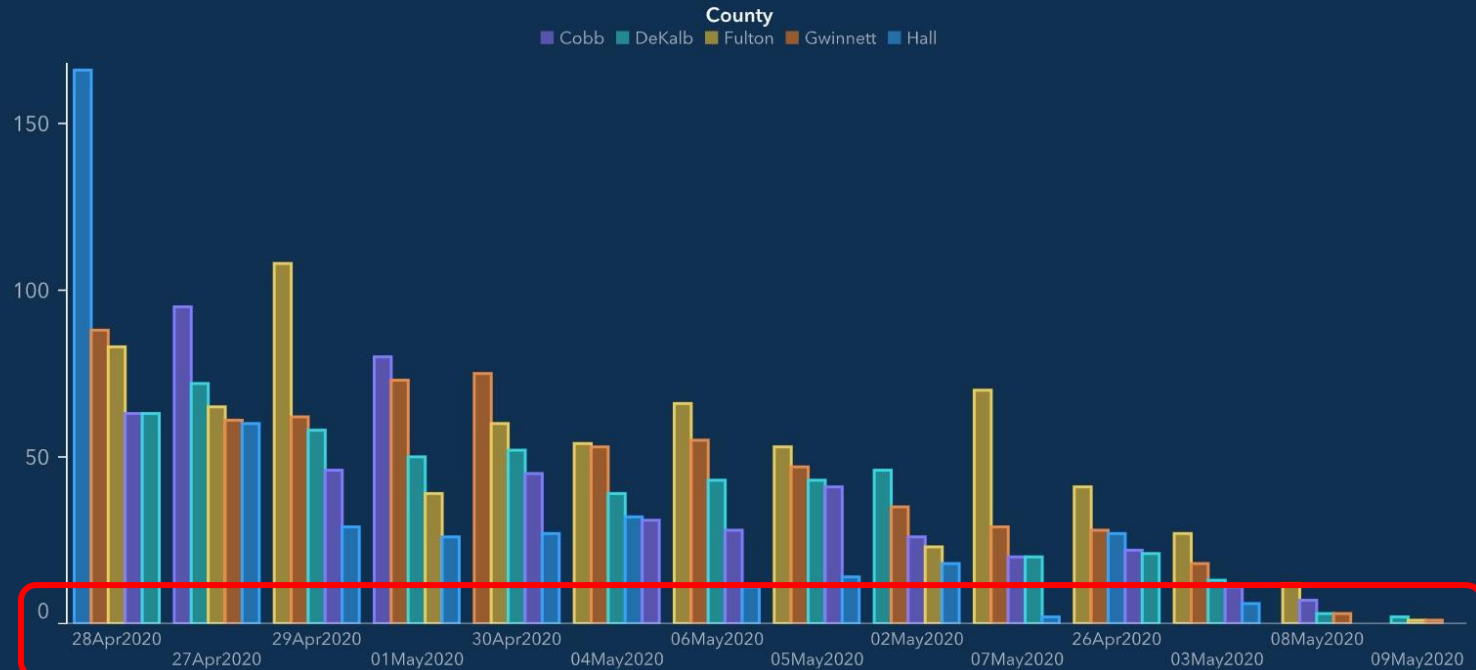
The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.



When Data Science Goes Wrong

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

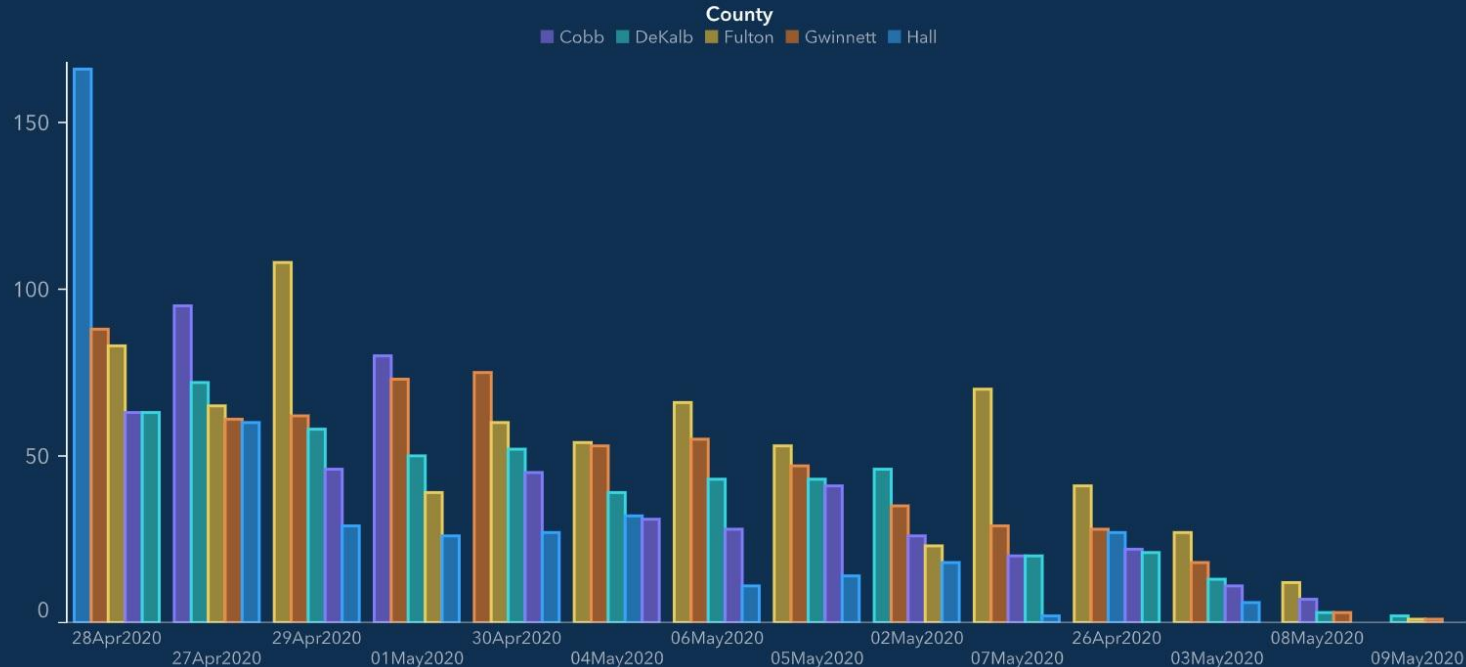
The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.



When Data Science Goes Wrong

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

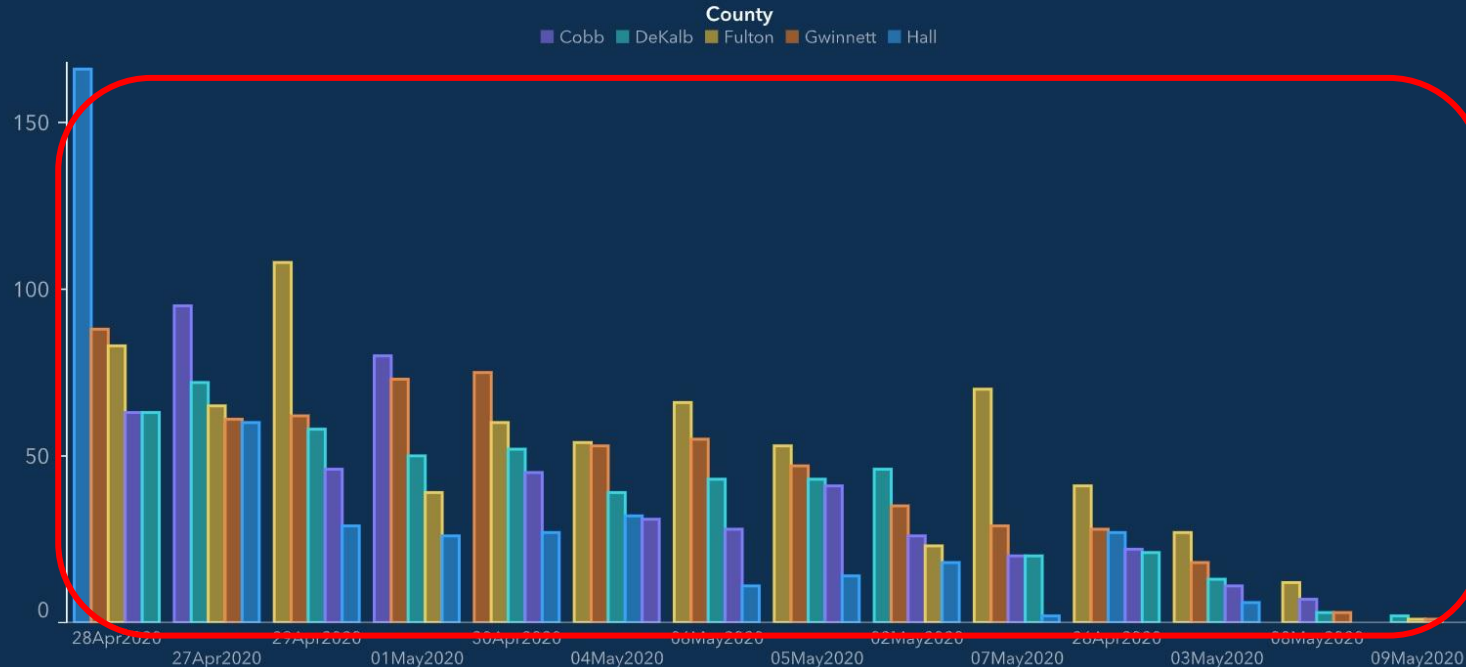
The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.



When Data Science Goes Wrong

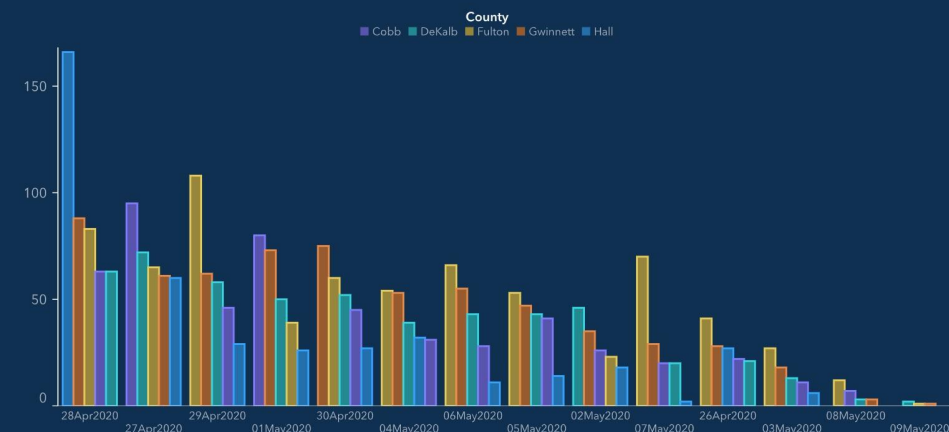
Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

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.



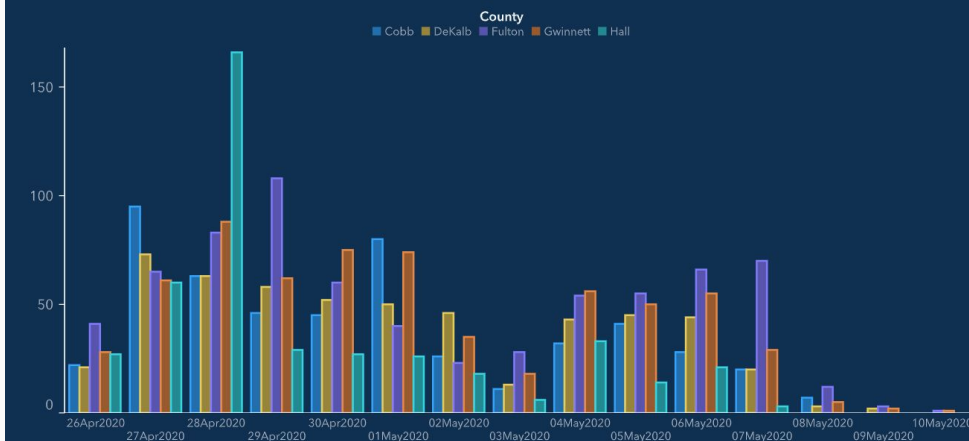
Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

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.

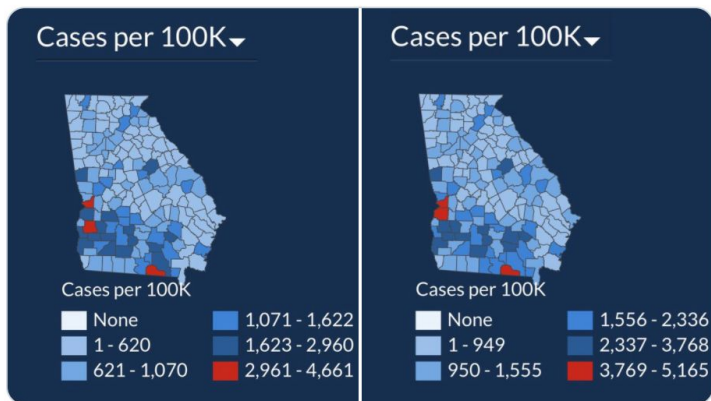


When Data Science Goes Wrong



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/

105

2.9K

13.2K



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/

58

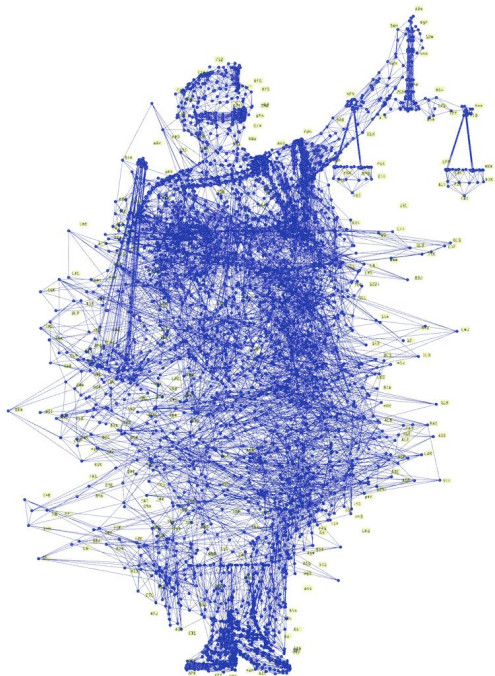
2K

10.5K



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

When Data Science Goes Wrong



- “Algorithmic bias describes **systematic and repeatable errors** in a computer system that create **unfair outcomes**, such as privileging one arbitrary group of users over others” - [Wikipedia](#)
- “High bias is a reflection of problems related to the gathering or usage of data, where systems draw **improper conclusions** about data sets” - [Margaret Rouse](#)

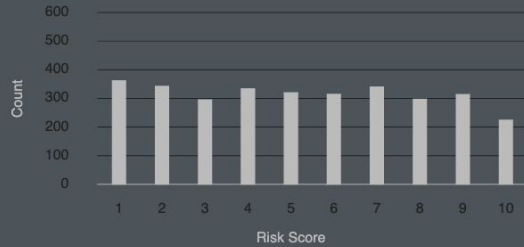
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

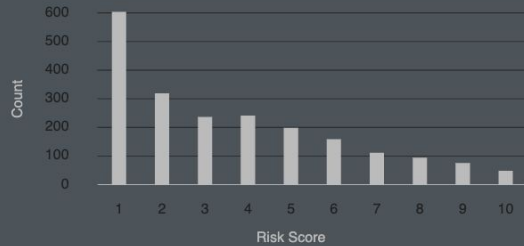
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Black Defendants' Risk Scores



White Defendants' Risk Scores



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. (Source: ProPublica analysis of data from Broward County, Fla.)

Two Petty Theft Arrests



VERNON PRATER

LOW RISK

3



BRISHA BORDEN

HIGH RISK

8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

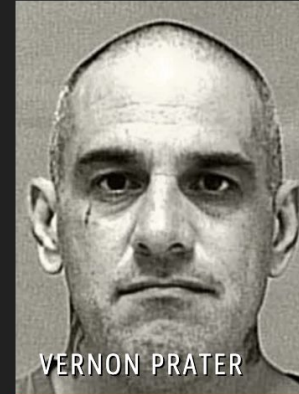
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

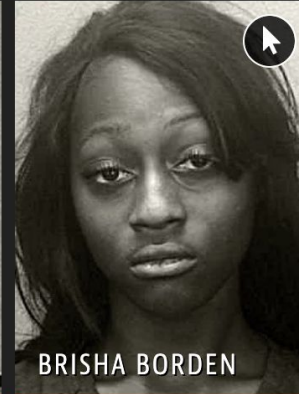
Two Petty Theft Arrests



VERNON PRATER

LOW RISK

3



BRISHA BORDEN

HIGH RISK

8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Perpetuating Gender-Based Employment Discrimination

Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs

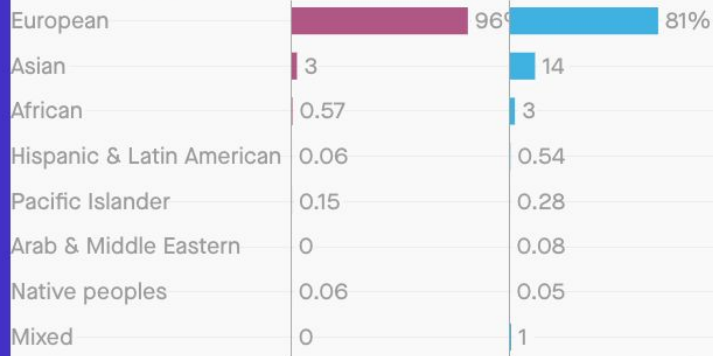
Amazon scraps secret AI recruiting tool that showed bias against women

81% of participants in genome-mapping studies were of European descent.

Little progress is being made to improve diversity in genomics

Share of samples in genetic studies, by ancestry

■ 373 studies, up to 2009 ■ 2,511 studies, up to 2016



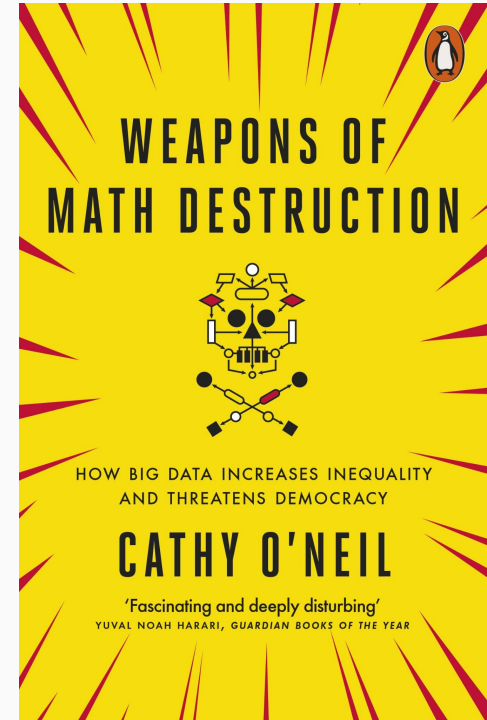
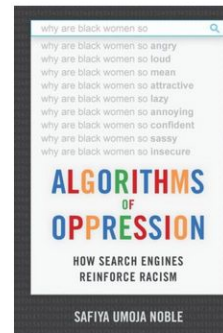
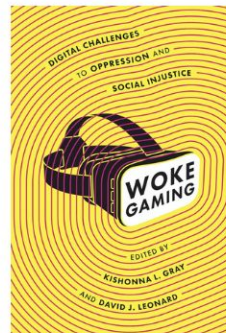
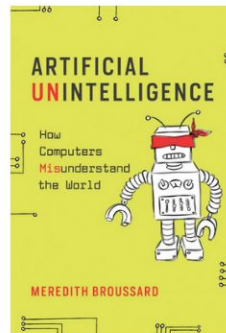
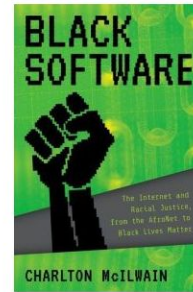
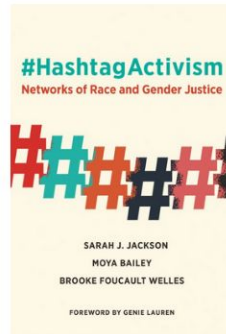
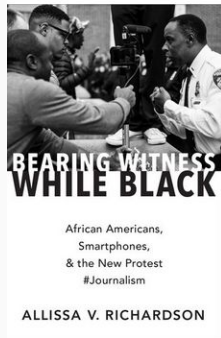
ATLAS | Data: Popejoy & Fullerton, Nature, 2016

Share

There is some precedent for the government to step in to ensure diversity during the data-gathering phase of new health policies and medical treatments. In 1993, the US Congress compelled the National Institutes of Health to bring more diversity to the medical studies it funded. It's not clear Congress or the NIH can solve this problem alone; more than 20 years later, 81% of genomics research is still from those of European descent. And furthermore, a 2015 study found that 2% of the more than 10,000 NIH-funded cancer studies include enough minority groups to be statistically significant. The study points to multiple potential causes, including inadvertent incentives in the NIH's funding structure, but the simplest is a lack of diversity in the medical field itself, and the propensity for non-white researchers to be funded less often.

Algorithmic Bias Reading List

- For more examples and reading [here is an amazing list](#)



How can we avoid algorithmic bias?

“... human beings cannot overcome all forms of bias. But slowing down and learning what those traps are—as well as how to recognize and challenge them—is critical.”

- Yaël Eisenstat , Former CIA officer, national security advisor to vice president Biden, integrity operations head at Facebook

- Several ways to avoid bias
 - Data management
 - Choice of algorithm
 - Transparency (reproducibility)
 - Diverse data science teams

How can we avoid algorithmic bias?

“... human beings cannot overcome all forms of bias. But slowing down and learning what those traps are—as well as how to recognize and challenge them—is critical.”

- Yaël Eisenstat , Former CIA officer, national security advisor to vice president Biden, integrity operations head at Facebook

- Several ways to avoid bias

- Data management
- Choice of algorithm
- **Transparency (reproducibility)**
- Diverse data science teams



Today we will be focusing on making our analyses reproducible using RMarkdown

How can we avoid algorithmic bias?

“... human beings cannot overcome all forms of bias. But slowing down and learning what those traps are—as well as how to recognize and challenge them—is critical.”

- Yaël Eisenstat , Former CIA officer, national security advisor to vice president Biden, integrity operations head at Facebook

- Several ways to avoid bias

- Data management
- Choice of algorithm
- **Transparency (reproducibility)**
- Diverse data science teams

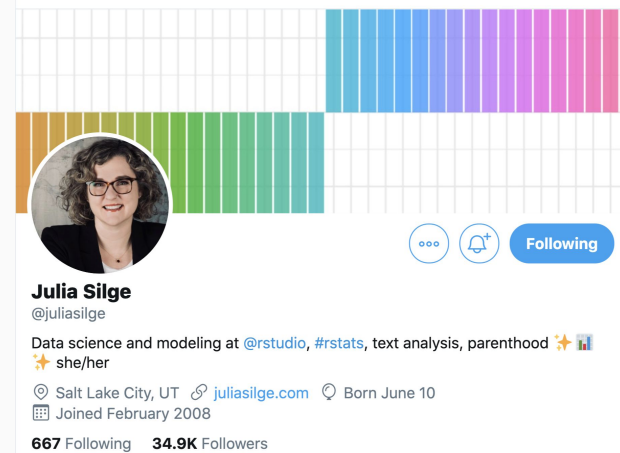


Today we will be focusing on making our analyses reproducible using RMarkdown

- “Data analysts who don’t organize their transformation pipelines often end up not being able to repeat their analyses, so the advice I would give to myself is the same advice often given to traditional scientists: **make your experiments repeatable!**” —Mike Driscoll, Founder & CEO at Metamarkets

R/RStudio Resources

- [RStudio website](#)
- [Hadley Wickham's Twitter](#)
- [Hadley Wickham's GitHub](#)
- [Hadley Wickham's Shiny](#)
- [R for Data Science](#)
- [Julia Siegel's Twitter](#)
- [David Robinson's Twitter](#)
- [Text Mining with R](#)



Who is helping you learn data science?

Instructor

Heather Mattie

Lecturer on Biostatistics
Co-Director, Health Data Science Master's Program
Department of Biostatistics
hemattie@hsph.harvard.edu

Office hour: Tuesdays 2:30-3:30 pm EST
Room: Building 1, 4th floor, room 421A



Teaching Fellows

Rebecca Astatke

rastatke@g.harvard.edu

Luke Benz

lukebenz@g.harvard.edu

Carmen Rodriguez
Cabrera

crodriguezcabrera@g.harvard.edu

Jerry Chang

chiarui_change@g.harvard.edu

Office Hours

Note: all office hours will be held in-person unless stated otherwise via a Canvas announcement

Day	Time	Location
Tuesday	1:30-2:30 pm	TBD
Tuesday	2:30-3:30 pm	Building 1, 4th floor, room 421A
Wednesday	10:30-11:30 am	TBD
Wednesday	1:00-2:00 pm	TBD
Thursday	2:00-3:00 pm	TBD

Note: lab will be held in-person unless stated otherwise via a Canvas announcement. Lab will not be held every single week. If lab will not be held a Canvas announcement will be sent out.

Day	Time	Location
Friday	9:45-11:15 am	TBD

Grading

- Homework
 - 4 assignments
 - 40% of final grade
 - You are welcome to discuss the course material and homework questions with others, but the work you turn in must be your own. Be sure to cite any sources you use.
- Take-home Midterm
 - 25% of final grade
 - Questions that require writing code and short answers
 - You are not allowed to work on or discuss this assignment with other students
- Final Project
 - 35% of final grade
 - Will work in teams of 3-4 people
- If taking course pass/fail, must earn final grade of **70% or more to pass**
- If auditing course you do not need to submit any assignments

Homework

- Real-world/public health/medical focus
- Scrape and wrangle/clean messy data
- Explore data
- Visualize data
- Perform statistical analyses
- Make predictions
- Communicate results



- Will be written in R using RMarkdown and submitted via private Github repositories
 - One repository per student per assignment
 - Only you and the teaching staff will have access to files in your repository
 - Must also submit html file
 - Points will be deducted if we are unable to knit your RMarkdown file when grading
- Can use 2 late days per assignment

Lab Sessions

- We will have labs centered around examples related to data and code presented in class
- Examples and code will help with homework assignments and the midterm
- Labs will start this week and will be held approximately every 1-2 weeks - check Canvas and the [course website for the schedule](#)

Final Project

- Teams of 3 - 4 students
- Choose a project from list of 3
 - Will be given a dataset and prompt and will be tasked with data wrangling, data visualization, statistical analyses, and communicating your findings
- [Project details and resources are available on the course website](#)
- A TF will be assigned to each team to give advice throughout the project
 - Assigned in beginning of November

Multiple modes of communication:

- Canvas site
- Course [website/GitHub repository](#)
- [Slack workspace](#)

Course Expectations

- You are encouraged, but not required, to attend lecture
 - Each lecture will be recorded and available on the Canvas course site
- Participation is not required or included as part of your final grade, but is highly encouraged
 - We all learn from each other
 - After awhile I start to dislike the sound of my own voice
- Attending a weekly lab session is recommended but not required
 - The session will be recorded and available on the Canvas course site
- Break time - we'll take a 5 minute break around the middle of each lecture (45-50 minutes in)

Action Items

- Download and install [R and RStudio](#)
 - Make sure you download R first
- Create a [GitHub account](#)
 - Remember what your username is - you'll need it to complete the survey below
- Complete this [survey](#)
 - **We need this information in order to email you your homework scores and comments**
- Explore the [course website](#)
- Optional but encouraged: introduce yourself with a short video or some text on the Canvas “Discussions” tab
- If needed: [Harvard's VPN](#)