BST 260
Introduction to Data Science

Lecture 1: Introduction to Course, R, & RMarkdown

Agenda

- What is Data Science?
- What is Health Data Science?
- What is a Data Scientist?
- Data Science success stories
- When Data Science goes wrong
- How do we learn Data Science?
- Who is helping you learn Data Science?
- Review of R and RMarkdown

Virtual Etiquette

- During lecture please keep yourself muted unless you would like to ask a question
- You may ask questions by unmuting yourself or through the chat or course Slack workspace
- This is a safe space and no question is dumb or pointless

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

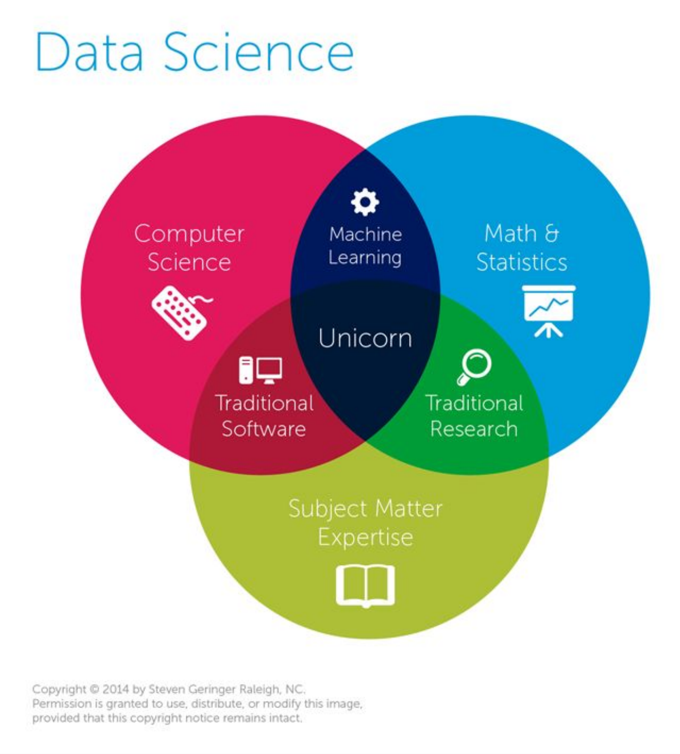
- The goal is to turn data into information and information into **insight** - Carly Florina, former CEO of Hewlett-Packard

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

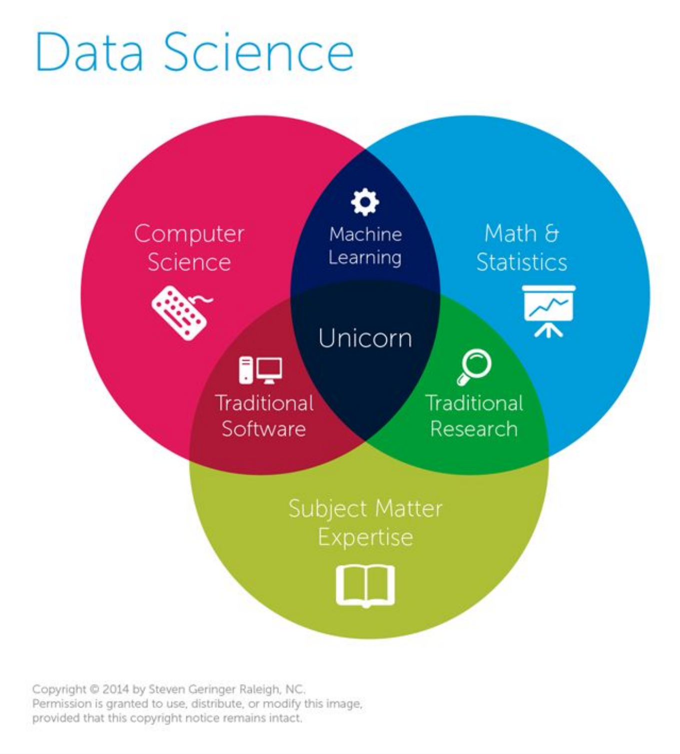
What is Data Science?

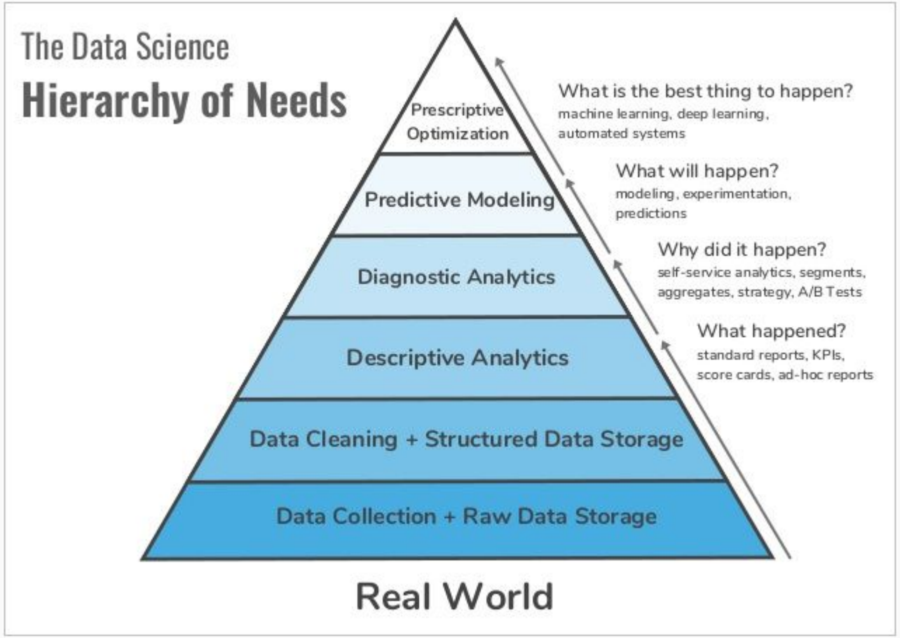
- Data Science is a **multidisciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and **insights** from structured and unstructured data - Wikipedia



What is Data Science?

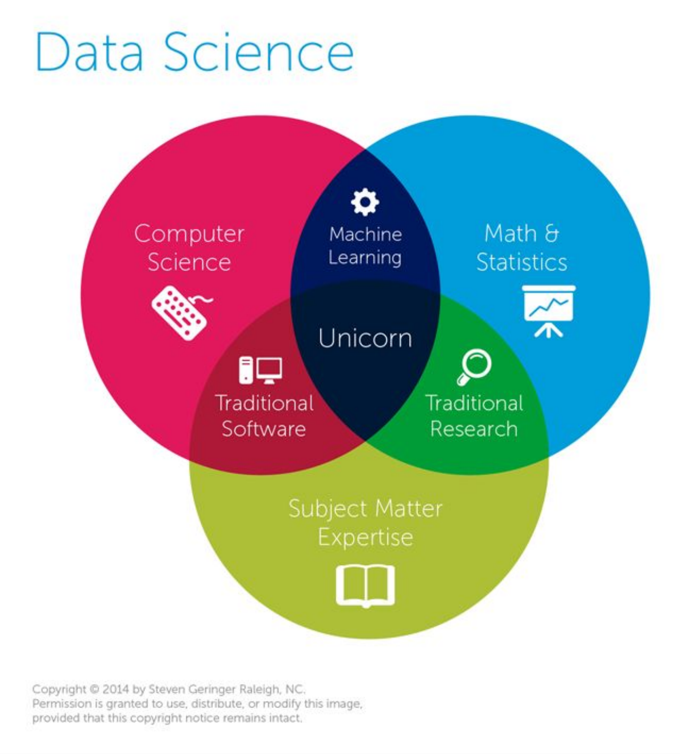
- Data Science is a **multidisciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and **insights** from structured and unstructured data - Wikipedia

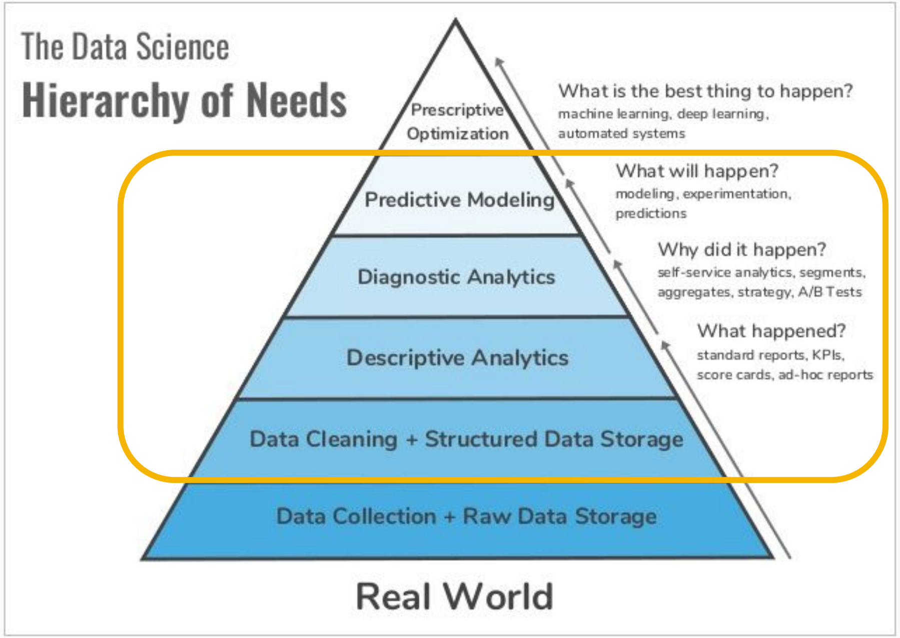




What is Data Science?

- 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



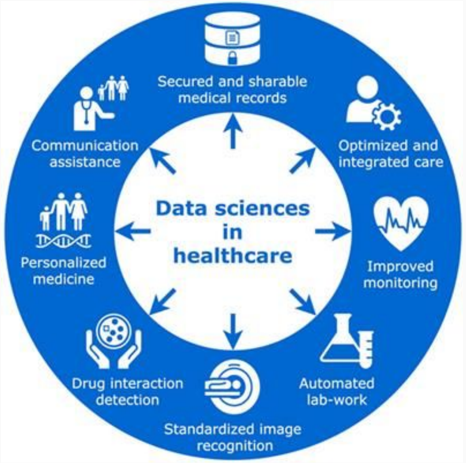


What is Health Data Science?

- “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 diagnoses
- Natural language processing



What is a Data Scientist?

- “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

- “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

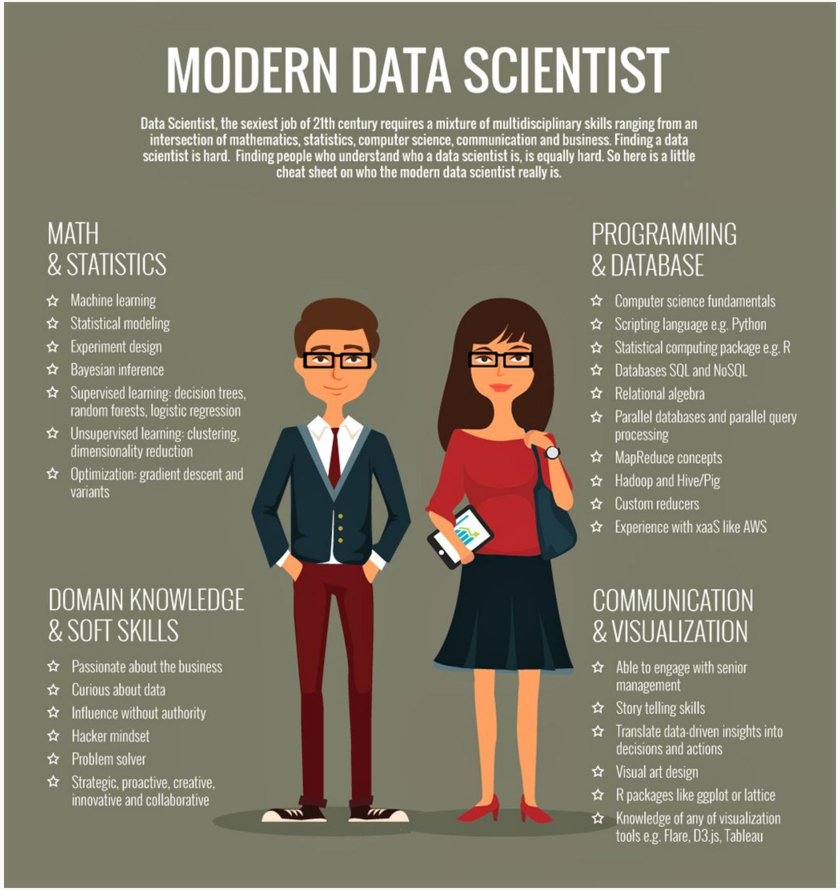
- “**As data scientists, our job is to extract signal from noise.**” —Daniel Tunkelang, Consultant / Advisor

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

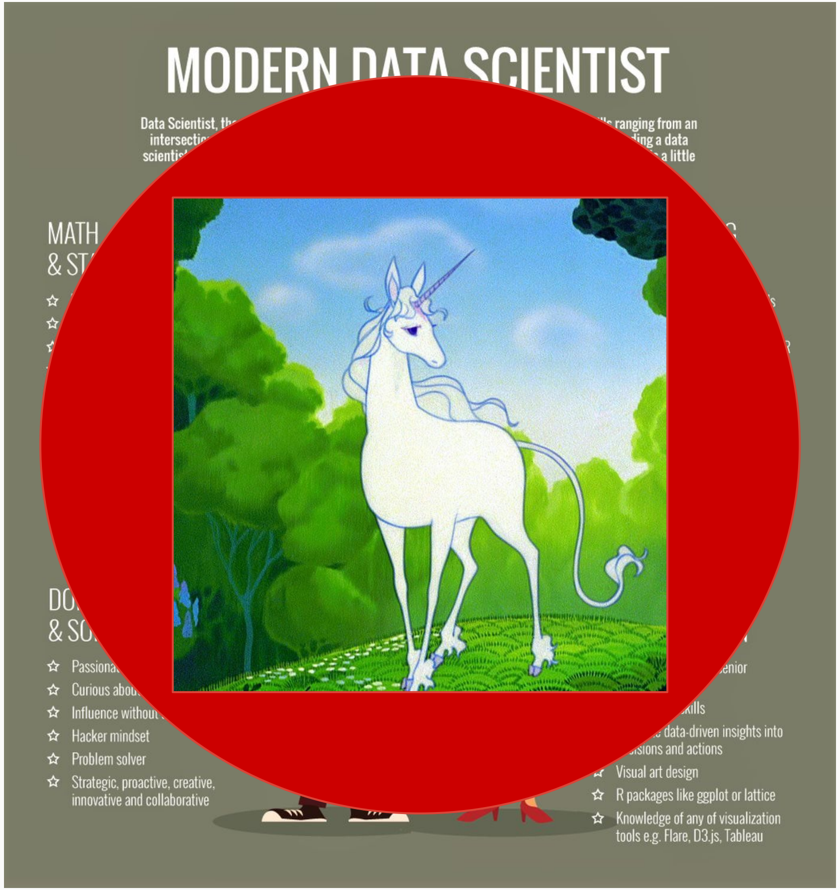
- “**What makes a good scientist great is creativity with data, skepticism and good communication skills.** Getting all of that together in the same person is difficult—because traditionally, different people follow different paths in their careers—some are more technical, others are more creative and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

What is a Data Scientist?

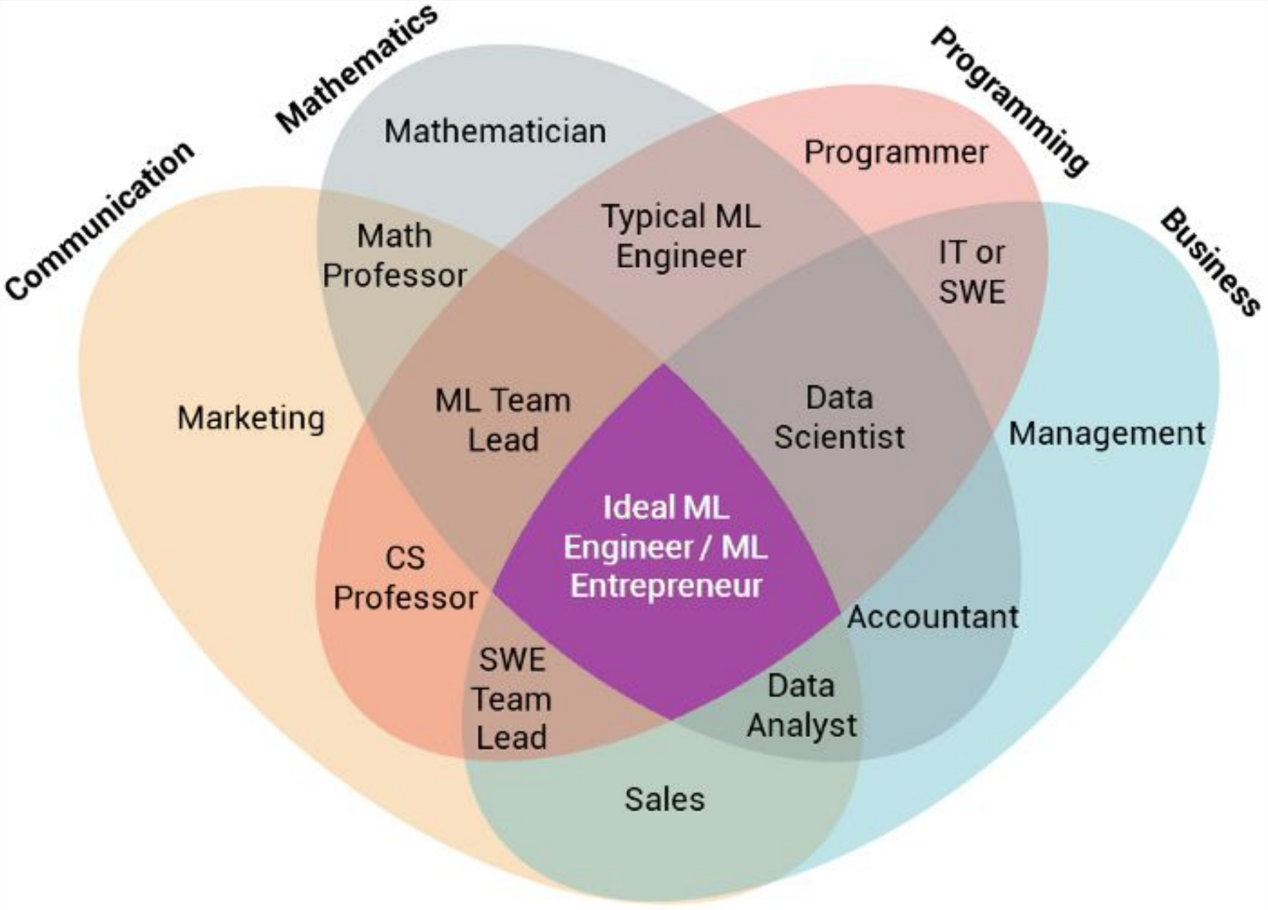


What is a Data Scientist?

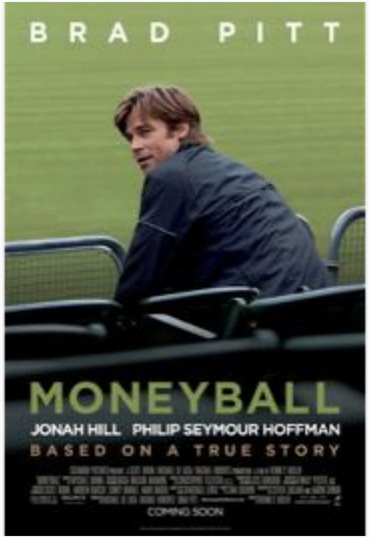
What is a Data Scientist?



What is a Data Scientist?



Data Science Success Stories



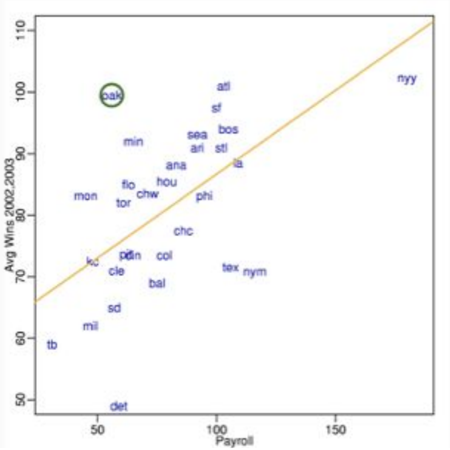
Real Data Scientist
(Paul DePodesta)

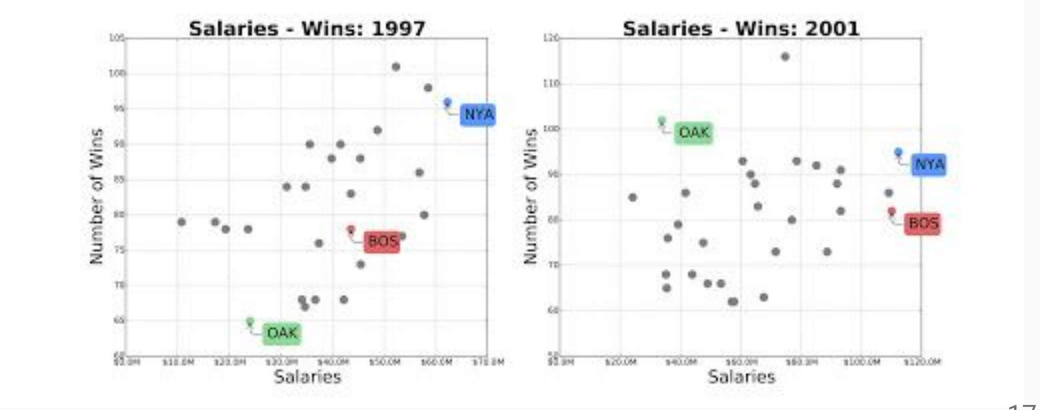
Hollywood Data Scientist

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”

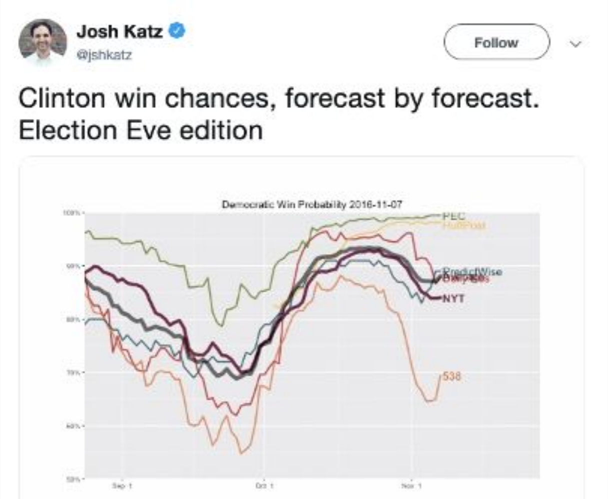
- Predicted: [349 to 189, 6.1% difference]
- Actual: 365 to 173, 7.2% difference
- While the 2016 election predictions weren’t nearly as close, [Nate Silver and 538 were the least wrong by far]


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.



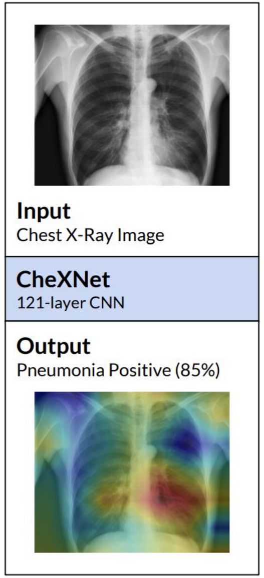
Data Science Success Stories

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

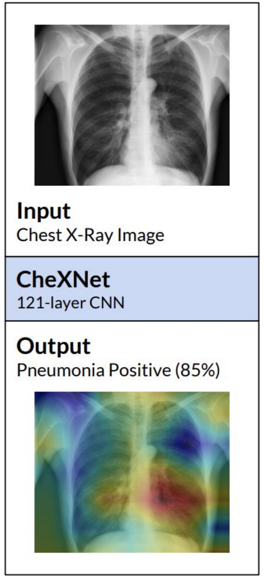
Data Science Success Stories

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).

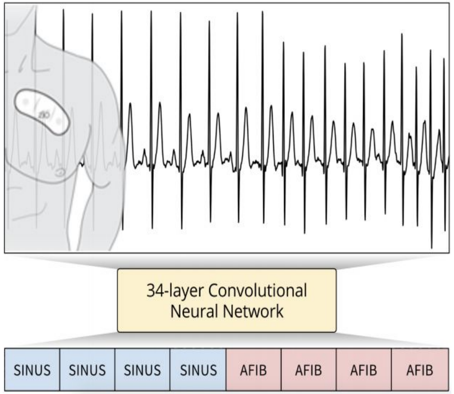


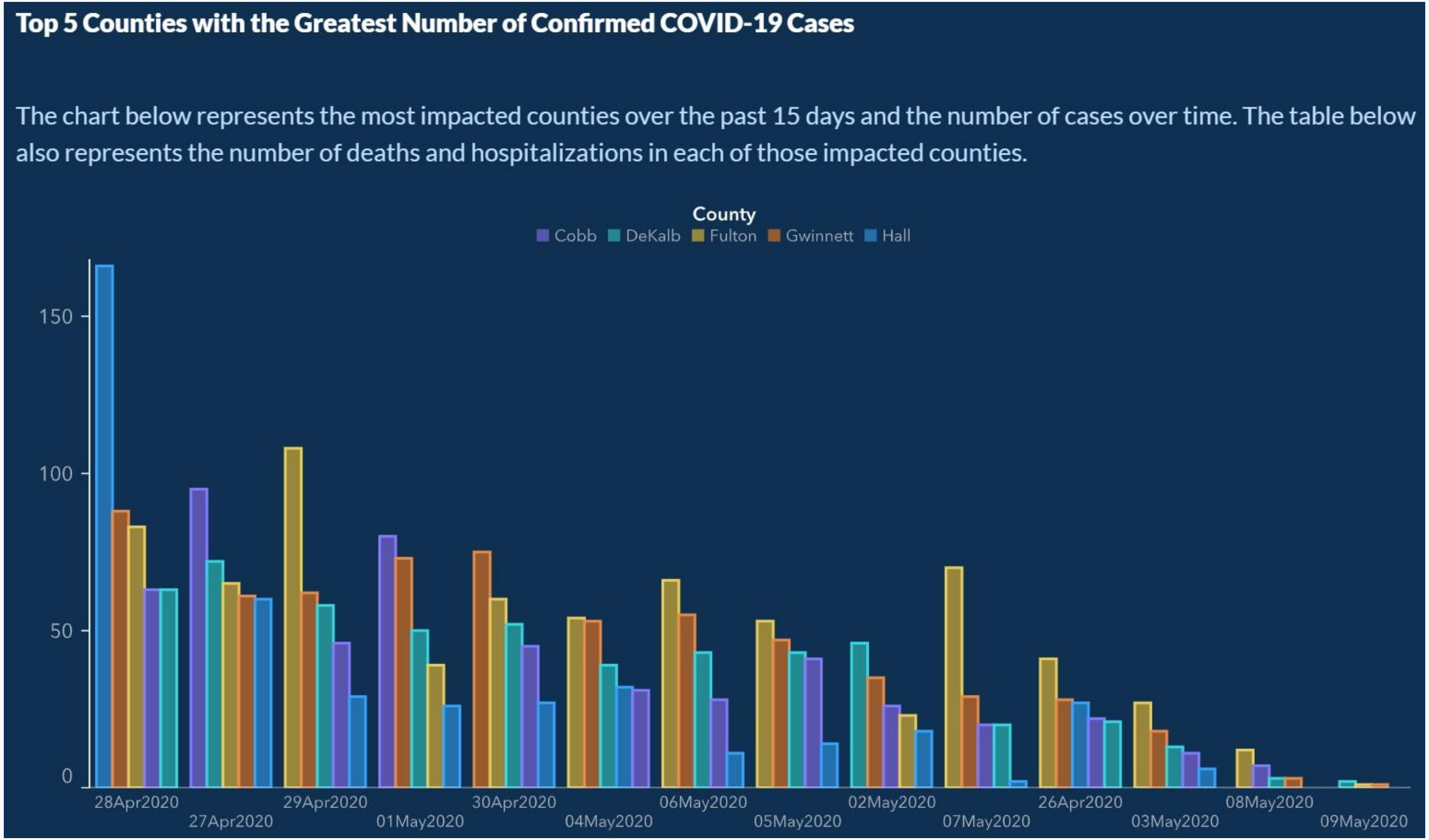
Figure 1. Our trained convolutional neural network correctly detecting the sinus rhythm (SINUS) and Atrial Fibrillation (AFIB) from this ECG recorded with a single-lead wearable heart monitor.

Data Science Success Stories

Many more examples

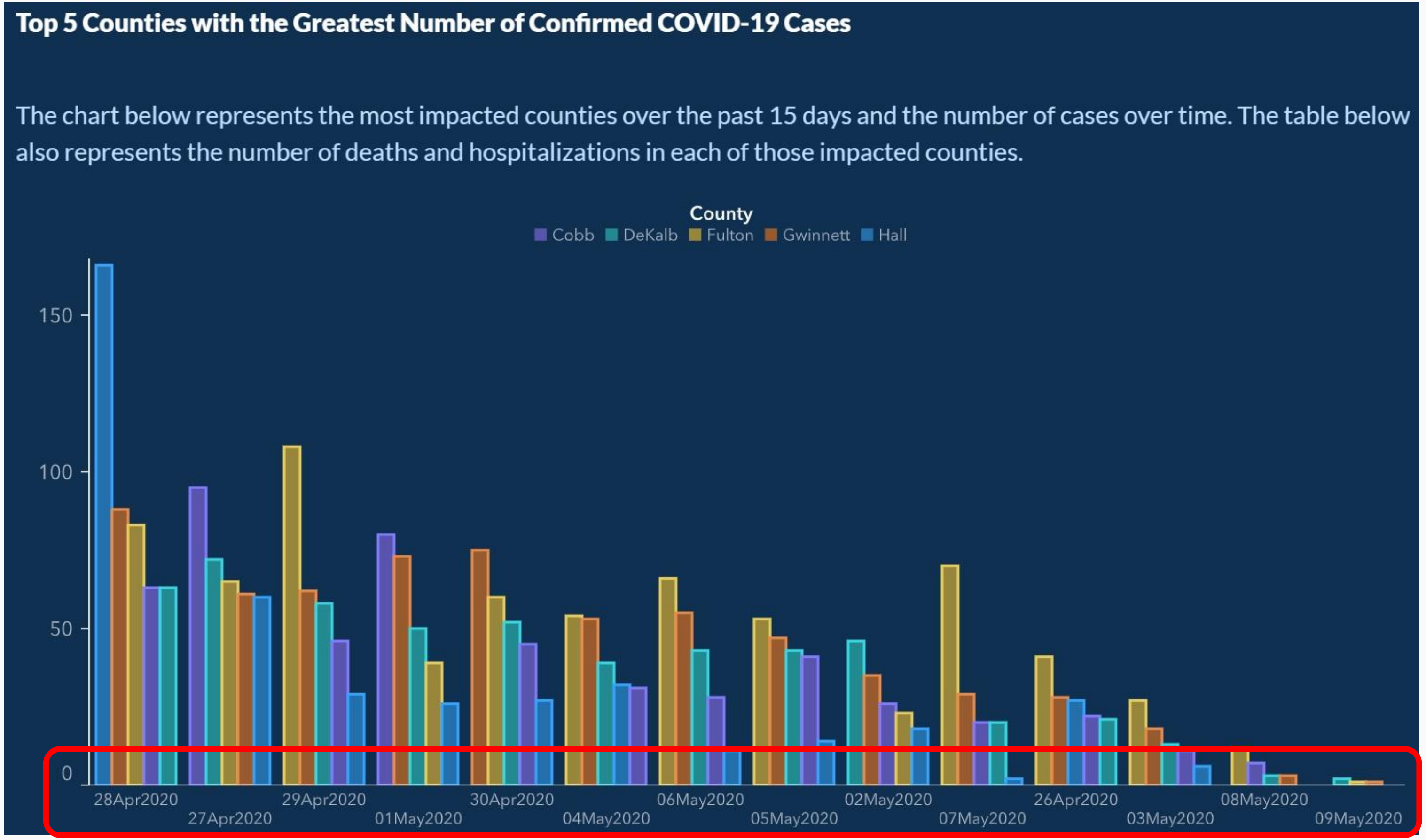
- Personalized medicine
- Medical diagnostics
- Spell checkers
- Natural language processing
- Language translators
- Etc.

When Data Science Goes Wrong

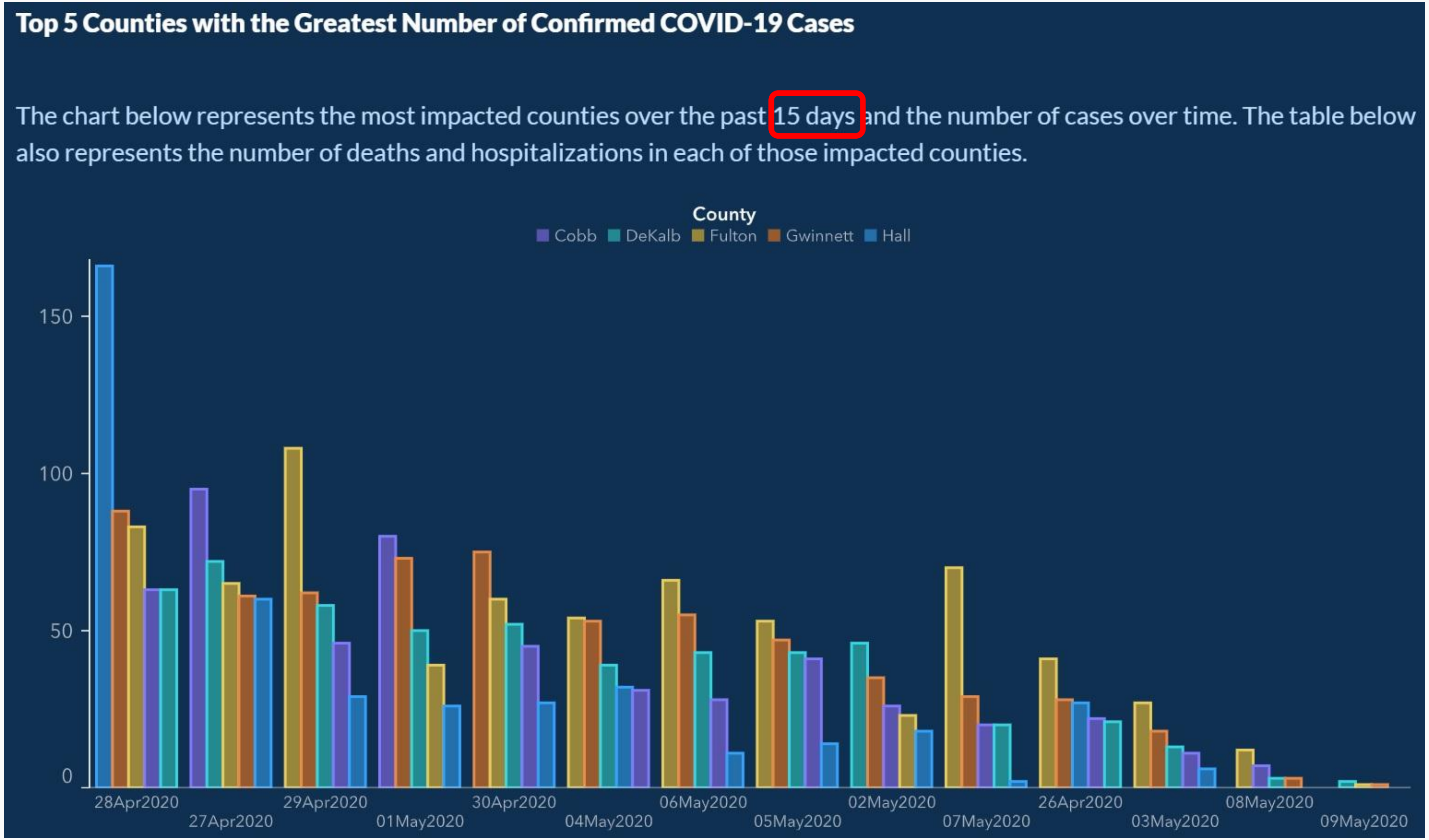


https://www.r-bloggers.com/ordering-bars-within-their-clumps-in-a-bar-chart-by-ellis2013nz/

When Data Science Goes Wrong



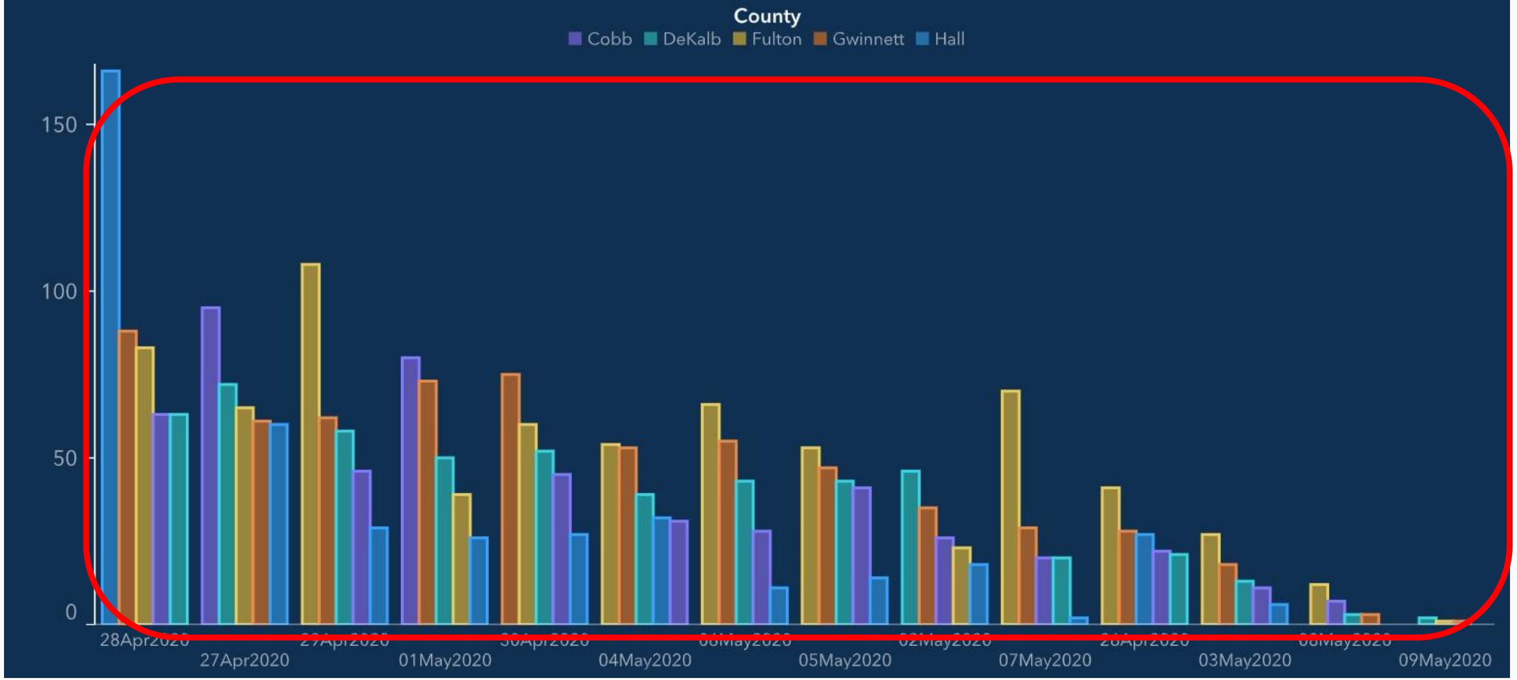
When Data Science Goes Wrong



When Data Science Goes Wrong

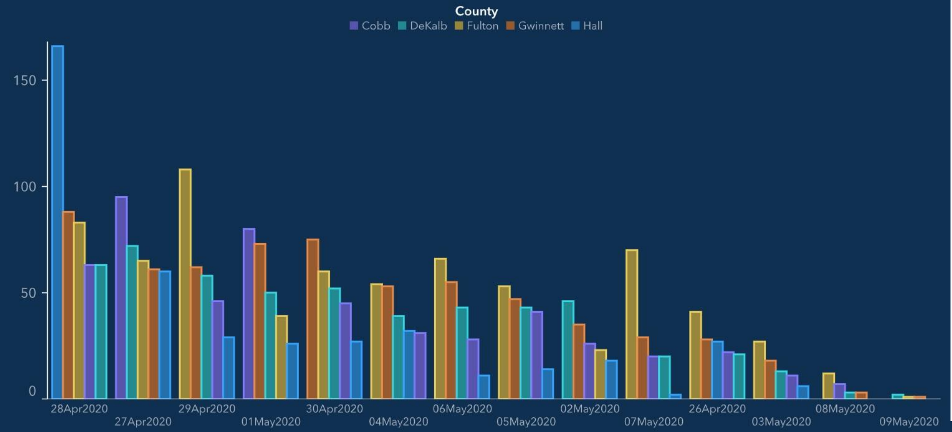
Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

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.



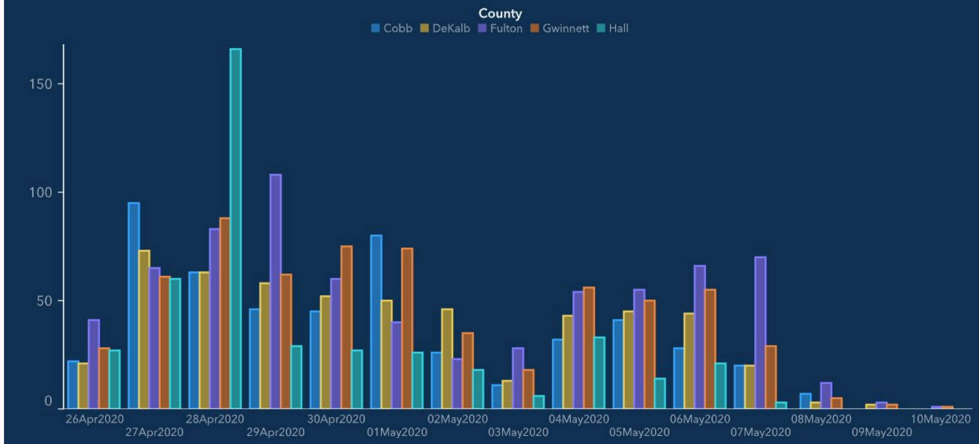
Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

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

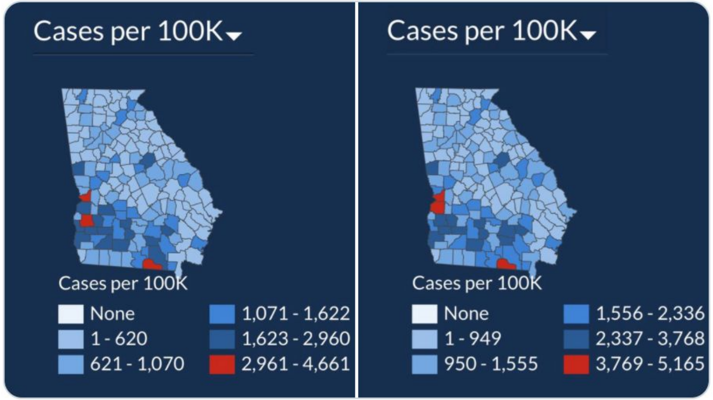
The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.



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/

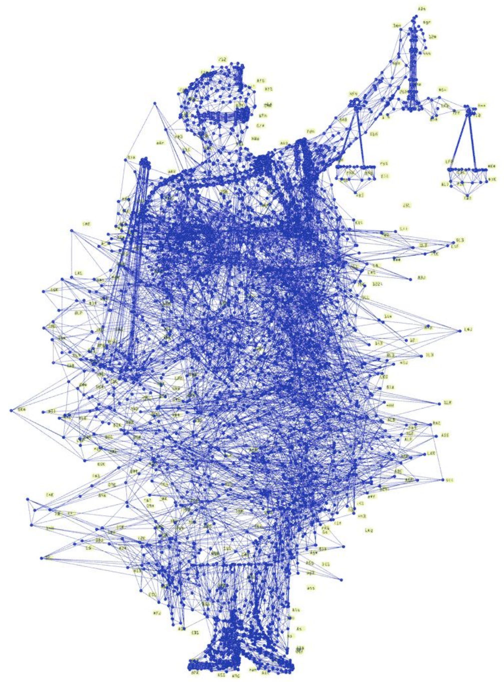
105 2.9K 13.2K

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

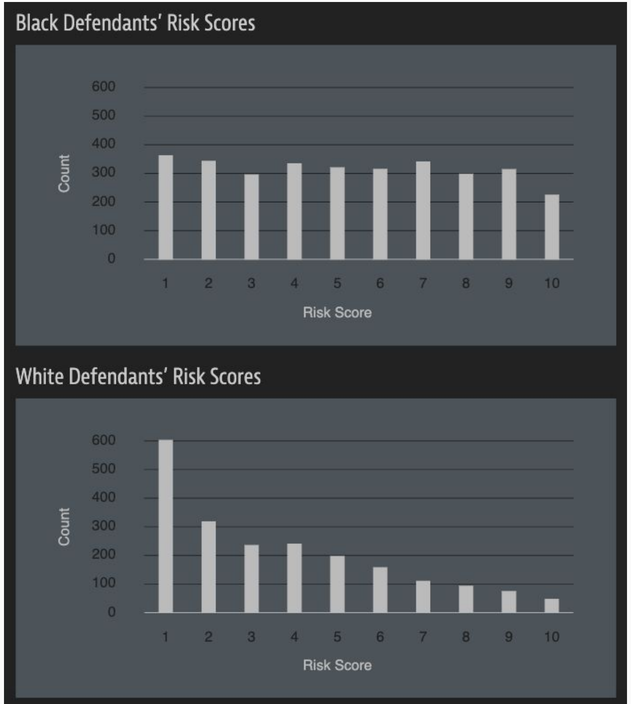
- “High bias is a reflection of problems related to the gathering or usage of data, where systems draw **improper conclusions** about data sets” - Margaret Rouse

Machine Bias

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

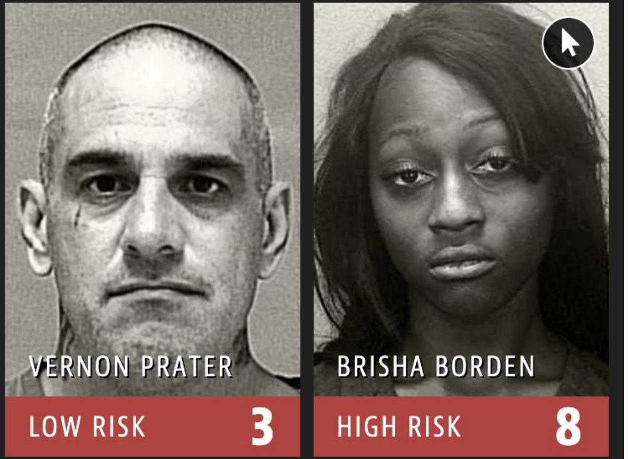
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016



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



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.

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Machine Bias

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

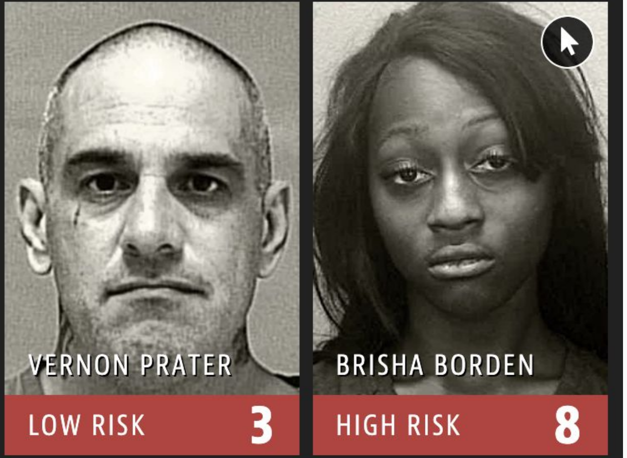
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

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



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.

Source:https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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

TECHNOLOGY

How Algorithms Can Bring Down Minorities' Credit Scores

Analyzing people's social connections may lead to a new way of discriminating against them.

A German company called Kreditech, for instance, asks loan applicants to share information from their social-media networks, which they can comb for details about their friends. Being connected to someone who's already paid back a loan with the company is "usually a good indicator," the company's chief financial officer told *Financial Times*.

In India and Russia, FICO—the company behind the popular FICO credit scores—is partnering with startups like Lenddo to capture information about users from their cellphones. Lenddo uses locations reported by applicants' phones to figure out whether they really live and work where they say they do, and then analyzes an applicant's network to figure out "if they are in touch with other good borrowers—or with people with long histories of fooling lenders," *Bloomberg* reports.

Source:
https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/

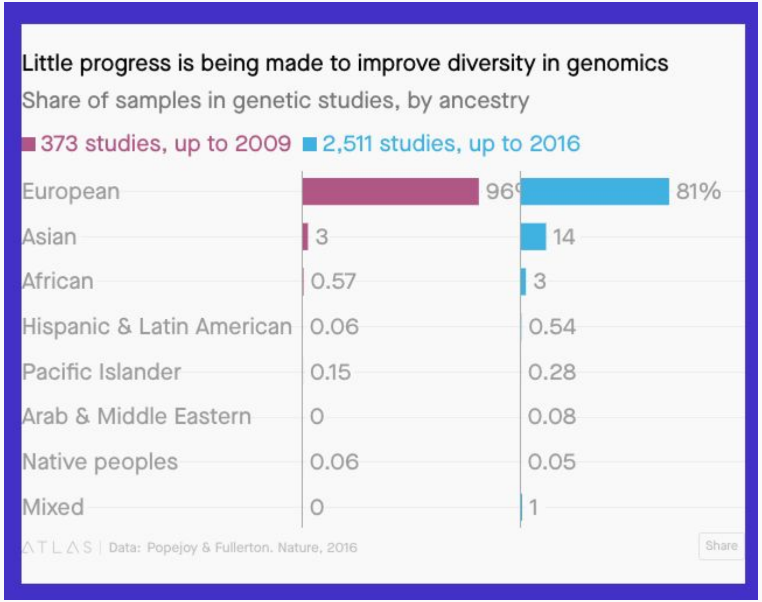
How algorithms can create inequality in health care, and how to fix it

by Matt Wood, University of Chicago

analyze data in the electronic medical records. One of the projects they're working on is to help decrease the length of stay for patients, because it's in everyone's best interest to have patients go home as soon as they're ready to leave. The thought was if we can identify patients who are most likely to be discharged early, we can assign a case manager to make sure there are no further blockages or barriers that could prevent them from leaving the hospital in a timely manner.

The data analytics group initially developed the algorithms based on clinical data, and then they found that adding the zip code where the patient lives improved the accuracy of the model identifying those people who would have shorter lengths of stay. The problem is when you add a zip code, if you live in a poor neighborhood or a predominantly African-American neighborhood, you were more likely to have the longer length of stay. So, the algorithm would have led to the paradoxical result of the hospital providing additional case management resources to a predominantly white, more educated, more affluent population to get them out of the hospital earlier, instead of to a more socially at-risk population who really should be the ones that receive more help.

Source: https://medicalxpress.com/news/2018-12-algorithms-inequality-health.html



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

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.

Source:https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/

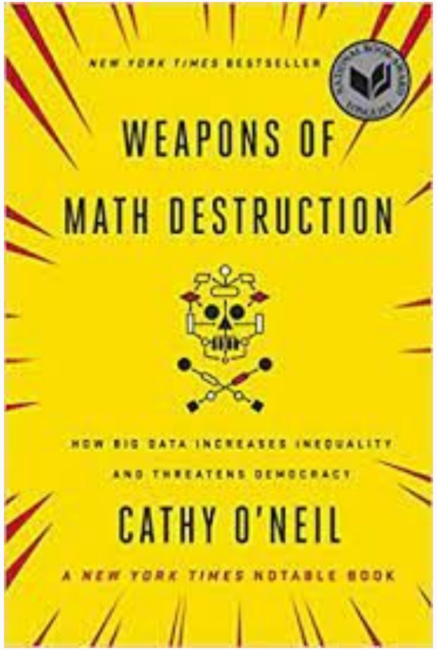
APPLICATION, MACHINE LEARNING, REPORT, SUMMARY

IBM's Watson for Oncology: A Biased and Unproven Recommendation System in Cancer Treatment?

Source: https://paulvanderlaken.com/2017/09/12/ibms-watson-for-oncology-a-biased-and-unproven-recommendation-system-in-cancer-treatment/

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]
- [Introduction to Data Science]
- [Introduction to Data Science course GitHub repository]
- [Julia Siege's Twitter]
- [David Robinson's Twitter]
- [Text Mining with R]

Who is helping you learn data science?

Instructor	Teaching Assistants
Heather Mattie Lecturer on Biostatistics Co-Director, Health Data Science Master's Program Department of Biostatistics hemattie@hsph.harvard.edu Office hour: Mondays 8-9pm, Tuesdays 1-2pm EST 	Jane Liang ● Office hour: Wednesdays 1-2pm EST Andy Shi ● Office hour: Fridays 1-2pm EST Kareem Carr ● Office hour: Mondays 11:30-12:30pm EST Santiago Romero-Brufau ● Office hour: Mondays 7-8pm EST Lara Maleyeff ● Office hour: Thursdays 8-9am EST Labs: Thursdays 8-9:30pm, Fridays 8:45-10:15am EST

Grading

- Homework
  - 5 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
  - Mix of multiple choice questions and 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
  - May work individually or as part of a team of up to 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, with a maximum of 6 late days total for the course

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 other week - check Canvas and the course website for the schedule

Final Project

- Teams of 2 - 4 students
- Choose your own data and project
  - Must include at least 1 type of analysis per team member
  - A Shiny app counts as a type of analysis

- Part 1: describe your project question and plan for answering it
- Part 2: present code, visualization, results and conclusions
  - RMarkdown file with knitted html file in a GitHub repository with README file
  - 2-minute screencast
    - A few will be shown during the last lecture of the course on December 16th
  - Website showcasing project

- Project details and resources are available on the course website and syllabus
  - Includes deadlines, links and examples of past projects
- A TA will be assigned to each team to give advice throughout the project
  - Assigned in beginning of November

Course Communication

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

- I will start the recording 10-15 minutes before class (~9:30am EST) and end the recording 10-15 minutes after class (~11:30am EST) to give time for questions and to get to know your classmates

- <u>Participation</u> is not required or included as part of your final grade, but is <u>highly encouraged</u>
  - Online courses are much more engaging with participation
  - We all learn from each other
  - <u>After a while I start to dislike the sound of my own voice</u>

- Attending a weekly lab session is recommended but not required
  - One 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