BIOGRAPHY

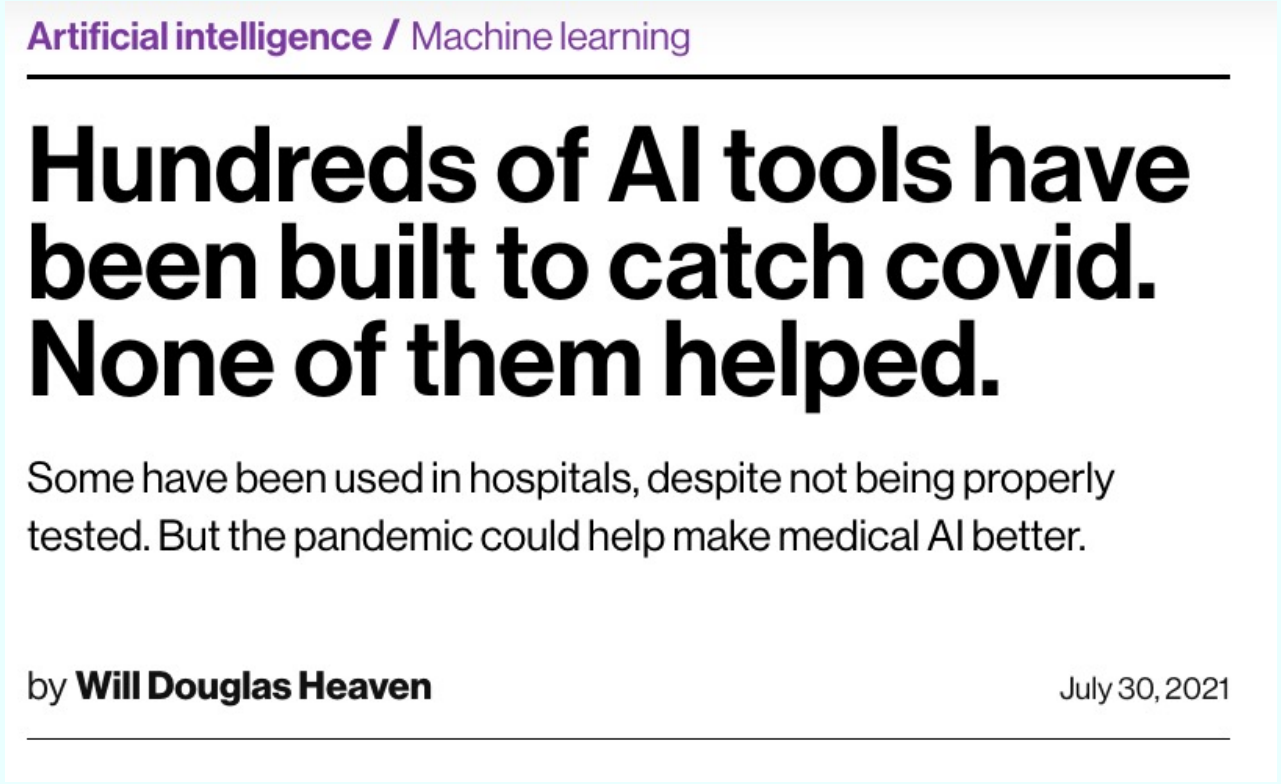
Jim Box: Jim Box is a Principal Data Scientist at the SAS Institute, where he has been supporting customers implementation of machine learning for the past seven years. Prior to that he spent 18 years in Clinical Research Organizations primarily as a statistician and programming director, He has Masters Degrees in Statistics and in Analytics.

PharmaSUG
NORTH CAROLINA • OCT. 29. 2021

Identifying Sources of Bias in Machine Learning Models

October 29, 2021

MIT Technology Review



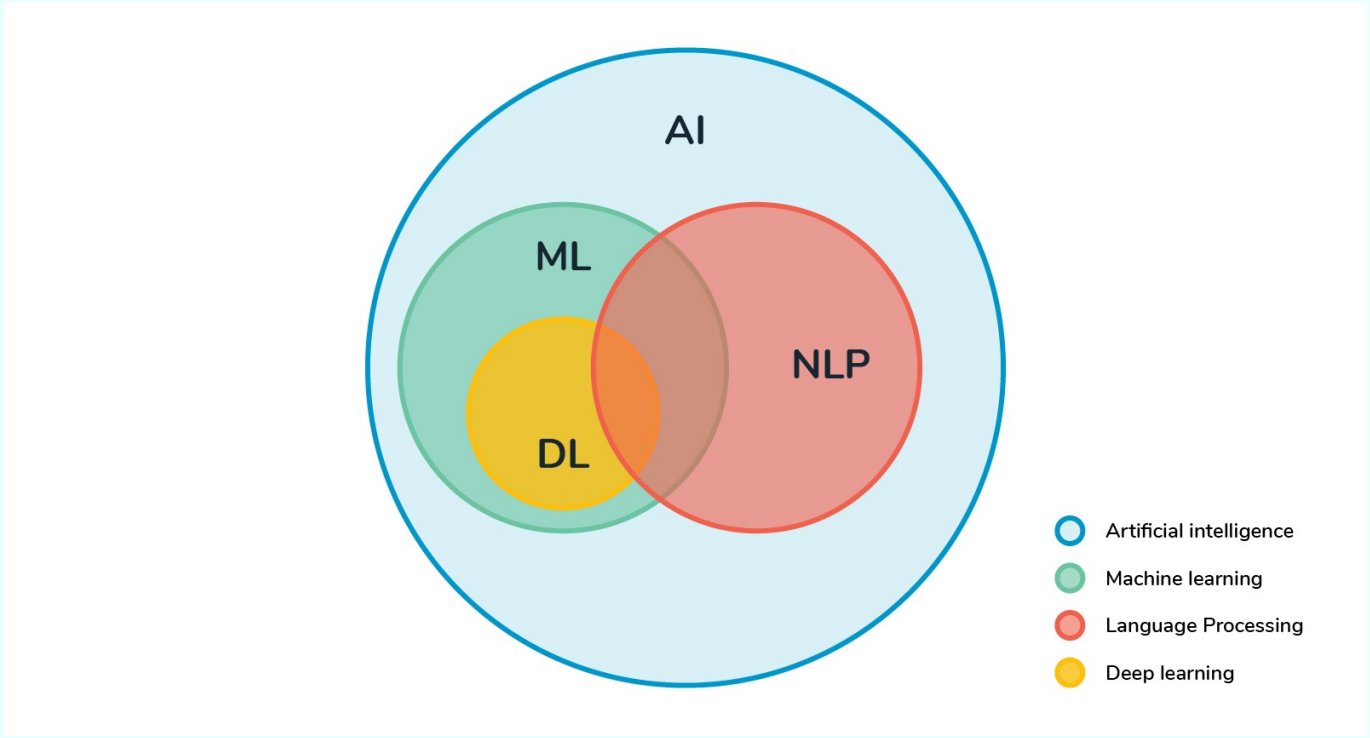
Artificial intelligence / Machine learning

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

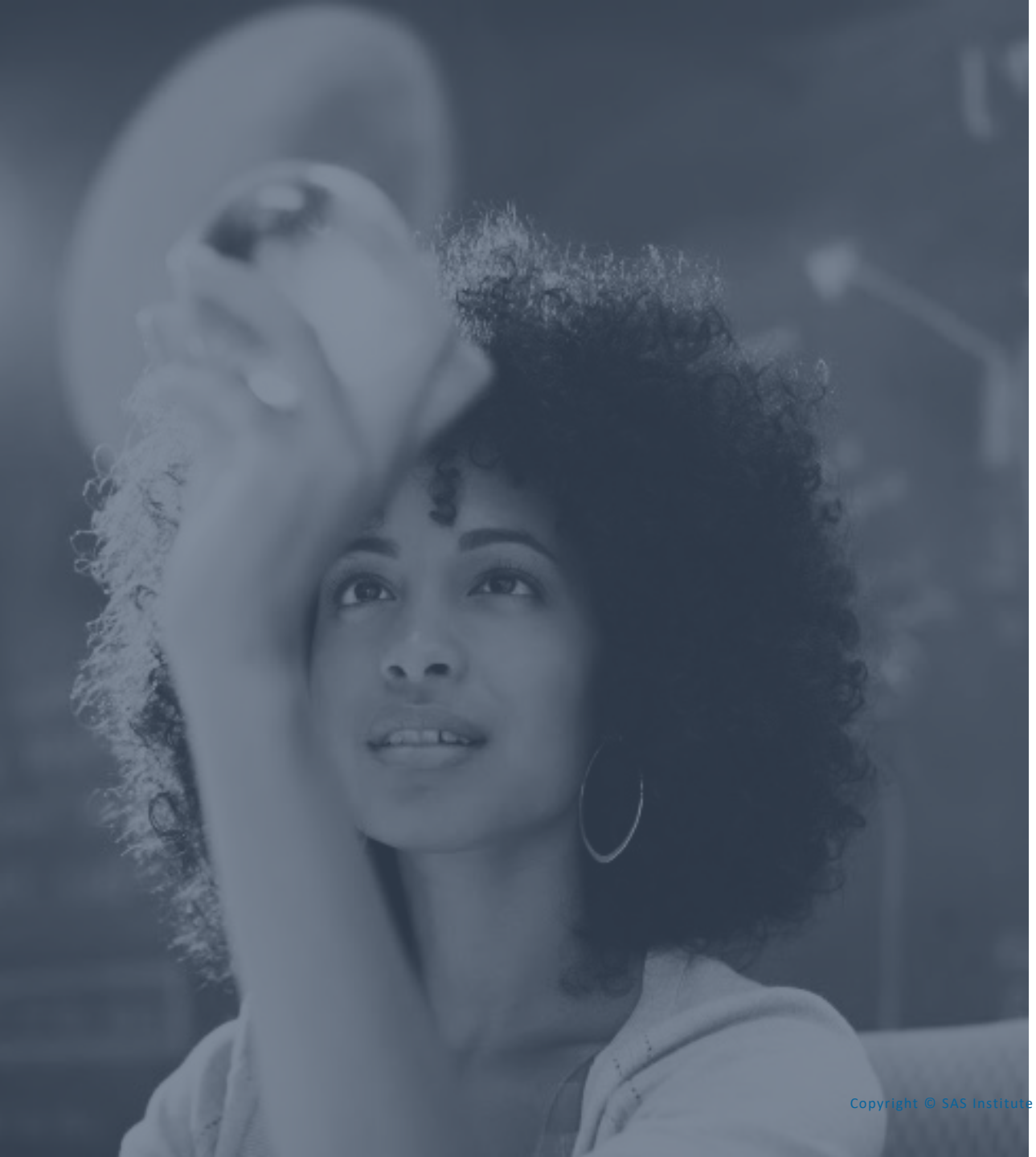
by **Will Douglas Heaven** July 30, 2021

Defining Artificial Intelligence



Artificial Intelligence

is the science of training systems to emulate human tasks through Learning and Automation



What is Machine Learning?

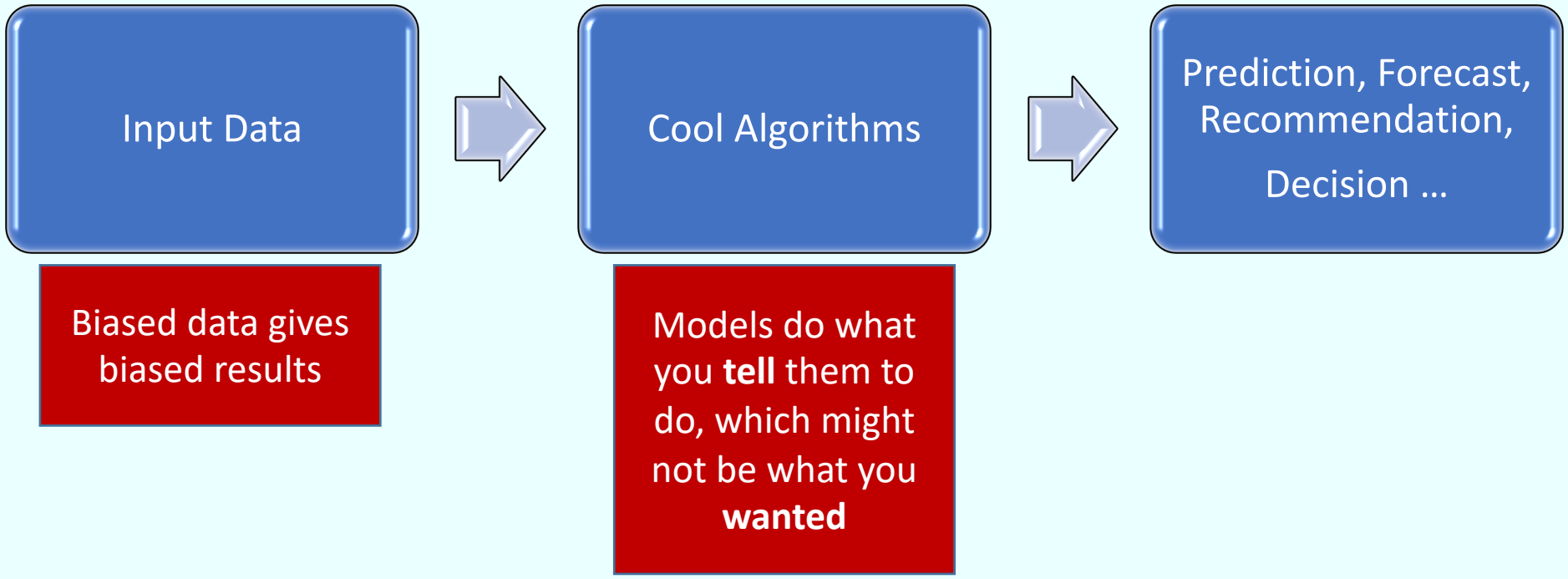
Machine Learning is a branch of artificial intelligence based on the idea that systems can **learn from data**, **identify patterns** and **make decisions** with minimal human intervention.

Copyright © SAS Institute Inc. All rights reserved.

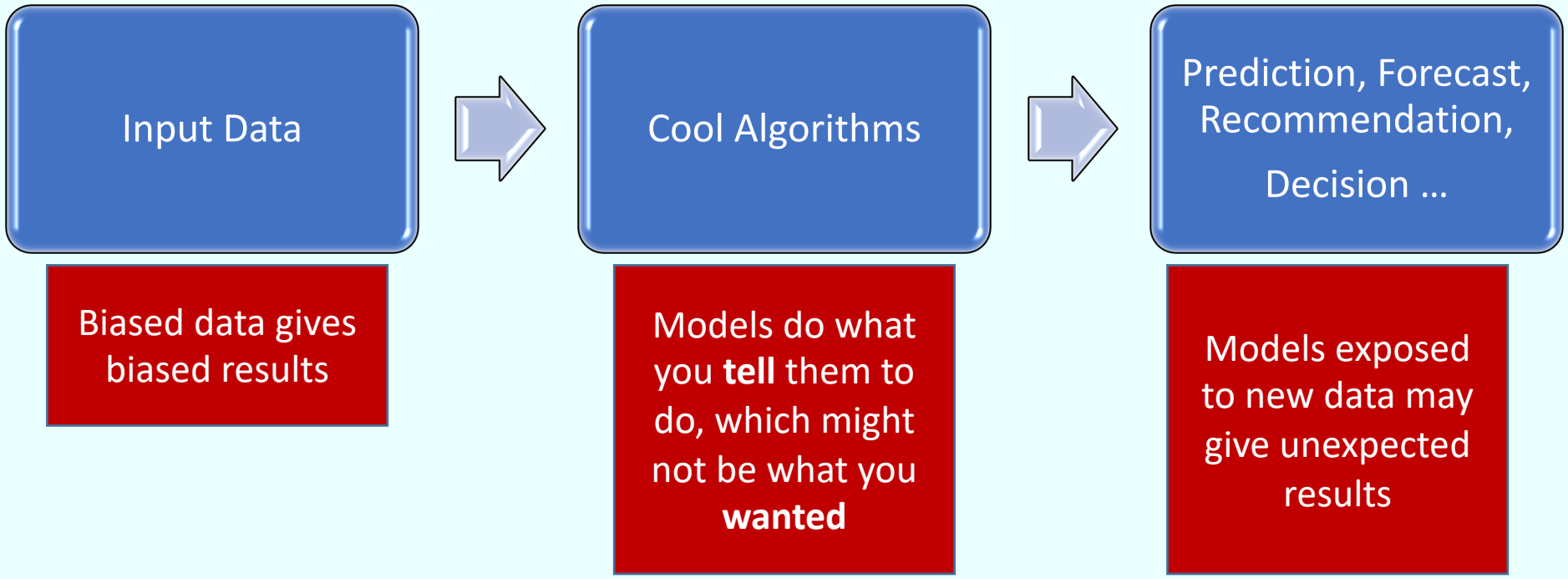
Typical ML Process

What Can Go Wrong?

What Can Go Wrong?



What Can Go Wrong?



Training Data

Biased Data Gives Biased Results

Biased Training Data

- Amazon receives hundreds of applications to open positions
- They are in the business of ranking things
- Created an AI system to comb through resumes and rank the applicants based on successful hires in the past
- Focused on interviewing the 5-star candidates

Biased Training Data

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Biased Training Data

ELENA SNAVELY

100 SAS Campus Drive
Cary, NC 27513
Elena.Snavely@sas.com

PROFESSIONAL OVERVIEW

Dynamic analyst with a proven record of achieving business objectives, committed to maximizing project potential. Experienced in critically evaluating information from disparate sources while proactively communicating results through presentations, reports, and discussions. Skilled at developing and optimizing models, conducting analysis utilizing statistical tools, and realizing business value from their output.

Key accomplishments and abilities:

- Capturing Customer Requirements
- Executing Project Plans
- Data Manipulation and Feature Engineering
- Executive Presentations
- Data Manipulation and Feature Engineering
- Managing and coaching groups
- Communicating effectively
- Design of User Interfaces

EXTRACURRICULAR

WOMEN'S CHESS CLUB CAPTAIN **Northampton, Massachusetts**
Smith College, August 2006-May2008

Biased Training Data

ELENA SNAVELY

100 SAS Campus Drive
Cary, NC 27513
Elena.Snavely@sas.com

PROFESSIONAL OVERVIEW

Dynamic analyst with a proven record of achieving business objectives, committed to maximizing project potential. Experienced in critically evaluating information from disparate sources while proactively communicating results through presentations, reports, and discussions. Skilled at developing and optimizing models, conducting analysis utilizing statistical tools, and realizing business value from their output.

Key accomplishments and abilities:

- Capturing Customer Requirements
- Executing Project Plans
- Data Manipulation and Feature Engineering
- Executive Presentations
- Data Manipulation and Feature Engineering
- Managing and coaching groups
- Communicating effectively
- Design of User Interfaces

EXTRACURRICULAR

WOMEN'S CHESS CLUB CAPTAIN **Northampton, Massachusetts**
Smith College, August 2006-May2008

Biased Training Data

ELENA SNAVELY

100 SAS Campus Drive
Cary, NC 27513
Elena.Snavely@sas.com

PROFESSIONAL OVERVIEW

Dynamic analyst with a proven record of achieving business objectives, committed to maximizing project potential. Experienced in critically evaluating information from disparate sources while proactively communicating results through presentations, reports, and discussions. Skilled at developing and optimizing models, conducting analysis utilizing statistical tools, and realizing business value from their output.

Key accomplishments and abilities:

- Capturing Customer Requirements
- Executing Project Plans
- Data Manipulation and Feature Engineering
- Executive Presentations
- Data Manipulation and Feature Engineering
- Managing and coaching groups
- Communicating effectively
- Design of User Interfaces

EXTRACURRICULAR

WOMEN'S CHESS CLUB CAPTAIN **Northampton, Massachusetts**

Smith College, August 2006-May2008

Biased Training Data

ELENA SNAVELY

100 SAS Campus Drive
Cary, NC 27513
Elena.Snavely@sas.com

PROFESSIONAL OVERVIEW

Dynamic analyst with a proven record of achieving business objectives, committed to maximizing project potential. Experienced in critically evaluating information from disparate sources while proactively communicating results through presentations, reports, and discussions. Skilled at developing and optimizing models, conducting analysis utilizing statistical tools, and realizing business value from their output.

Key accomplishments and abilities:

- Capturing Customer Requirements
- Executing Project Plans
- Data Manipulation and Feature Engineering
- Executive Presentations
- Data Manipulation and Feature Engineering
- Managing and coaching groups
- Communicating effectively
- Design of User Interfaces

EXTRACURRICULAR

WOMEN'S CHESS CLUB CAPTAIN **Northampton, Massachusetts**
Smith College, August 2006-May2008

Biased Training Data - Takeaways

- Models trained on biased data will excel at applying that bias – even more efficiently than humans
- **Your responsibility: Question the data**
  - Where did it come from
  - How representative is it?
  - How did it get labeled?
  - Is there a feedback loop?

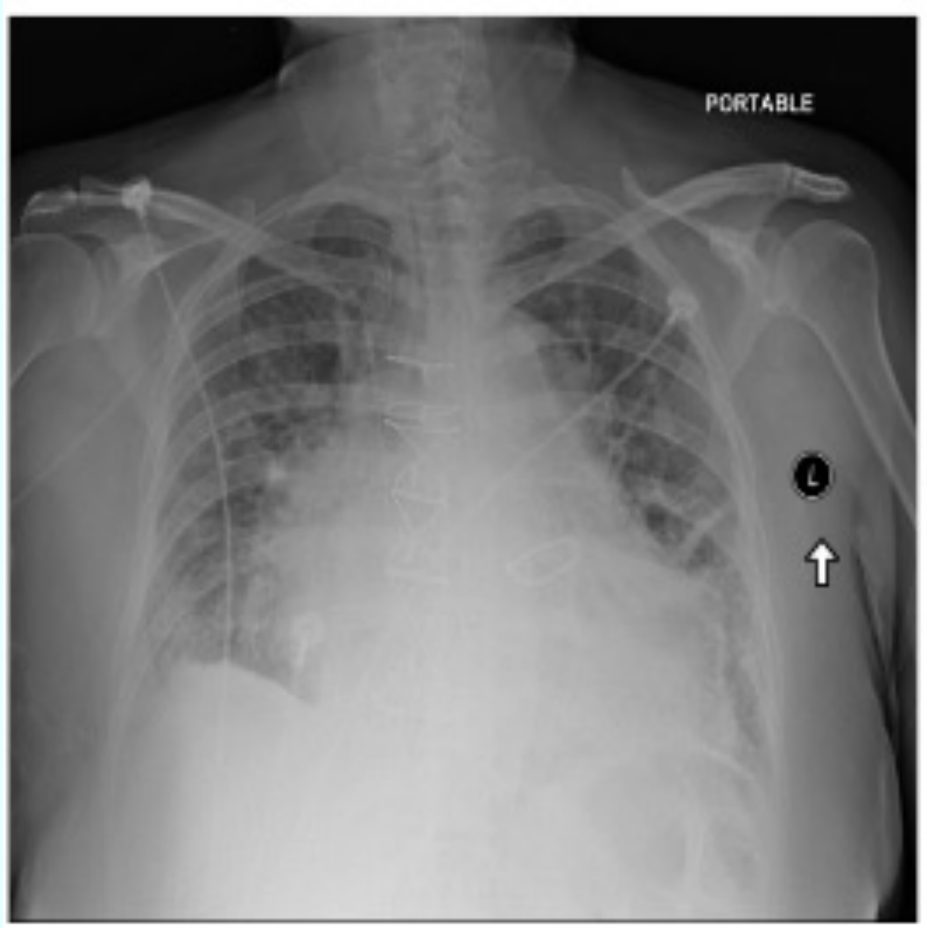
Models Do What you Tell Them to Do

That might not have been what you wanted them do to

Models Do What You Tell Them to Do

https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf

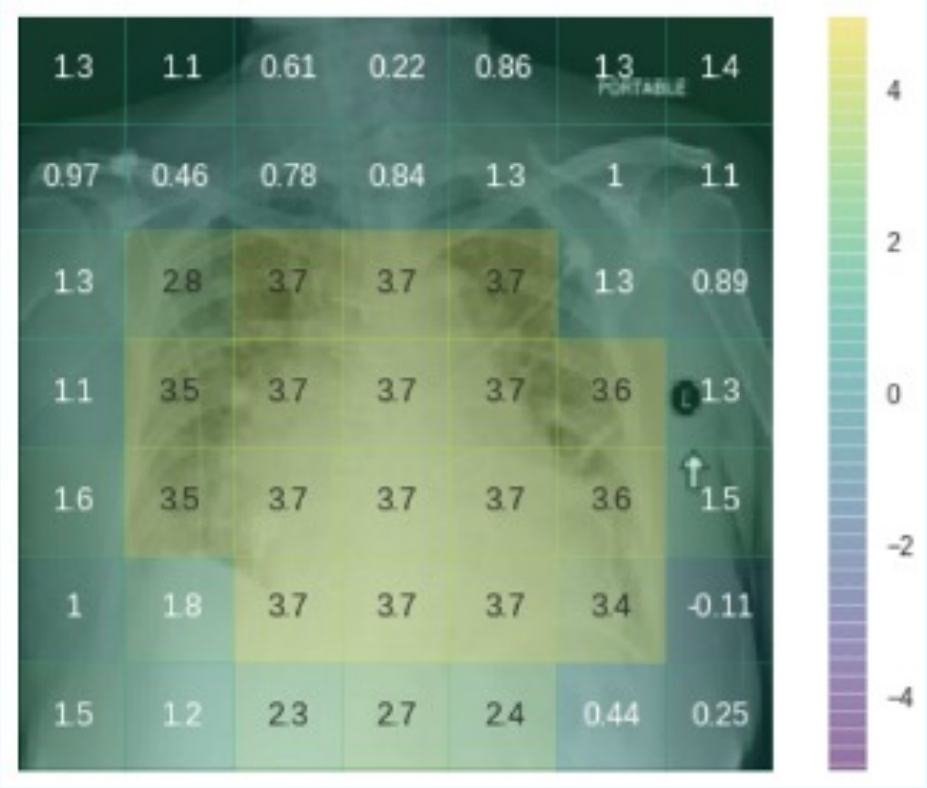
Models Do What You Tell Them to Do



- Diagnosing Cardiomegaly (Enlarged Heart)
- Model trained on labeled images
- Apply the model to new images to test how it does
- This patient has the condition, and the model gave it a probability of 0.752, so it seems to have worked

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

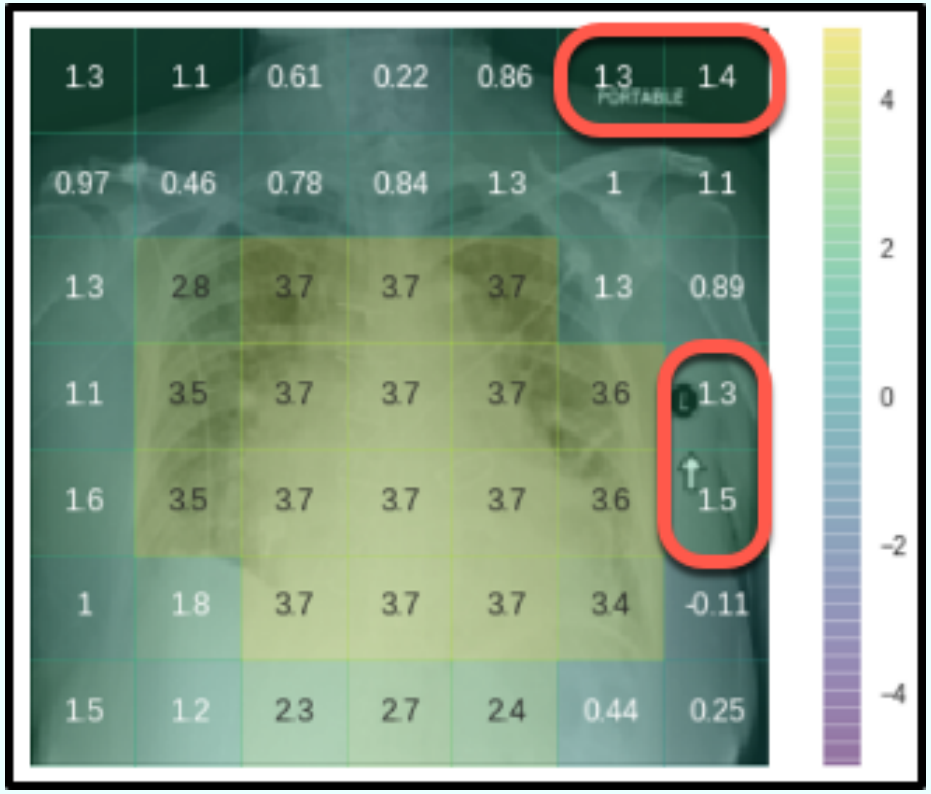
Models Do What You Tell Them to Do



- Heatmap shows how different parts of the image contribute to the prediction
- Looks like the focus area is on the heart, which is good

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

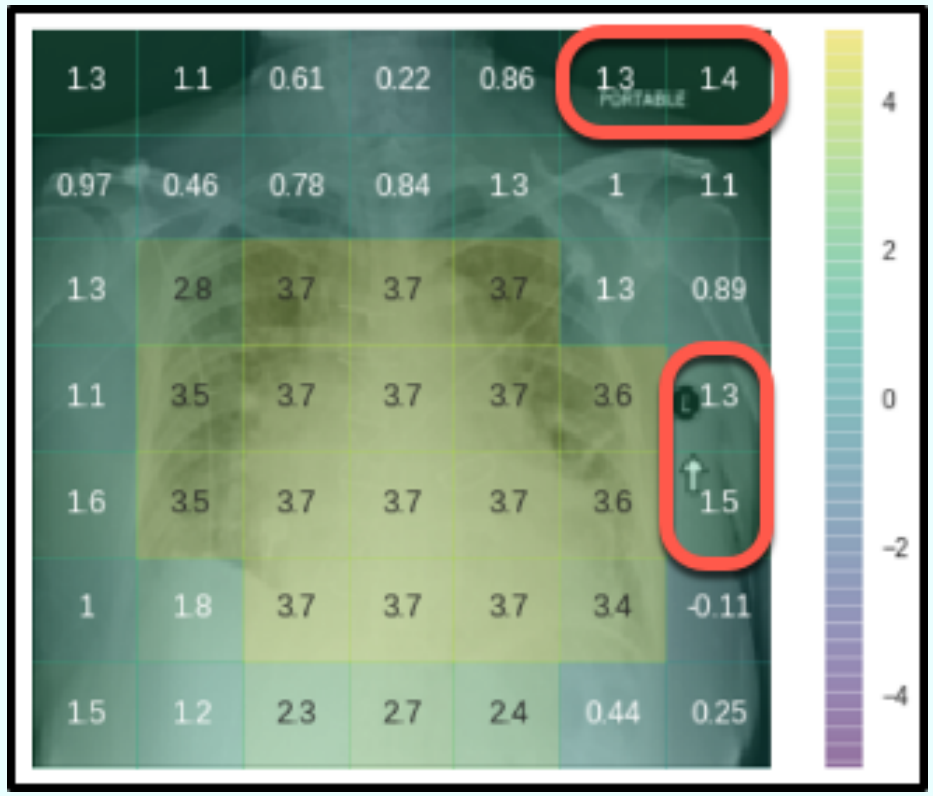
Models Do What You Tell Them to Do



- Heatmap shows how different parts of the image contribute to the prediction
- Looks like the focus area is on the heart, which is good
- Some surprises, though

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

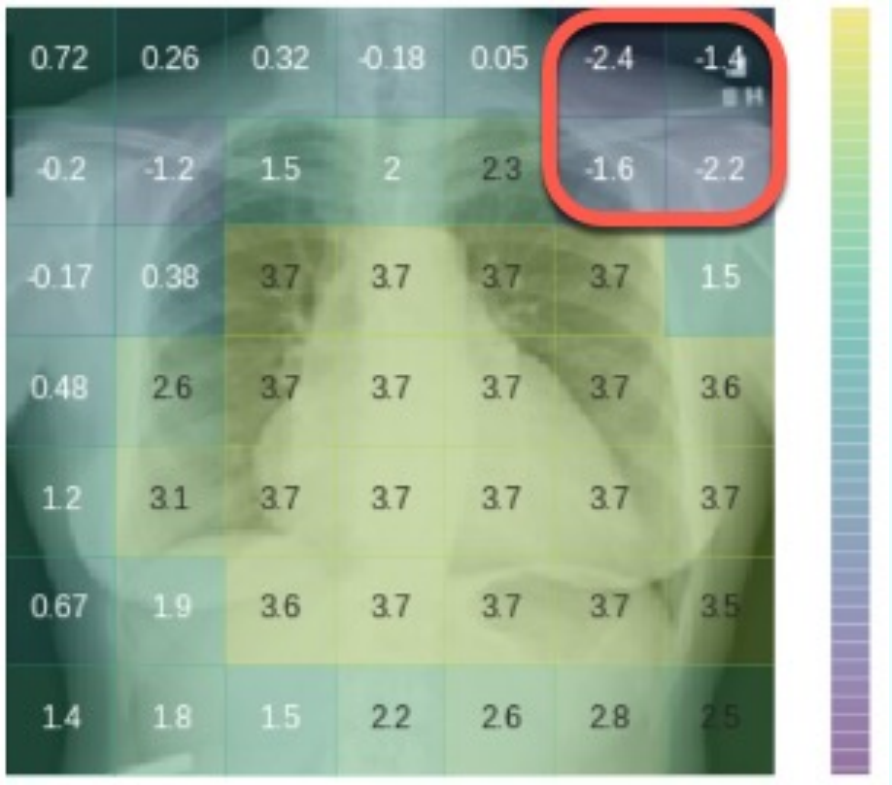
Models Do What You Tell Them to Do



- Model is looking at markers of image metadata
- Model is using the fact that this image was taken with a portable x-ray machine, which is mainly used on sicker patients
- Model also considered the reviewing radiologist

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

Models Do What You Tell Them to Do



- Different image of a patient with the same condition
- Model downgraded the predication due to the lack of the portable stamp

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

Models Do What You Tell Them to Do

- Models are lazy, but effective
- Models do not care about context unless specifically instructed
- **Your Responsibility: Question the Results**
  - Why did the model make a specific prediction for this specific case?
  - What are the key inputs being used to make predictions?
  - What, exactly, was the model set up to do?

Models May Not Perform Well on Novel Data

You may experience some unexpected results

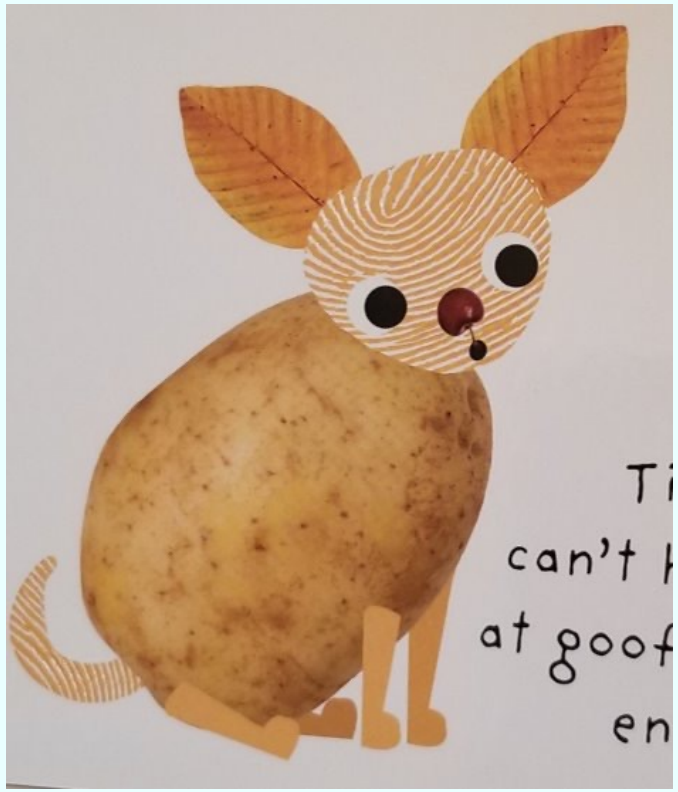
Models May Not Perform Well on Novel Data

- Models should only be applied to data that is like the data they were trained on
- Models will return a prediction on novel data, but it may not be trustworthy (although it will appear to be so)

Models May Not Perform Well on Novel Data

https://www.pngarts.com

Models May Not Perform Well on Novel Data



Happy Dog and Other Furry Friends. Written by Robert Newton. Illustrated by Ellie Boultwood

Models May Not Perform Well on Novel Data

- Models should only be applied to data that is like the data they were trained on
- Models will return a prediction on novel data, but it may not be trustworthy (although it will appear to be so)

Models May Not Perform Well on Novel Data

Genetics research 'biased towards studying white Europeans'

Ethnic minorities set to miss out on medical benefits of research, scientist warns

People from minority ethnic backgrounds are set to lose out on medical benefits of genetics research due to an overwhelming bias towards studying white European populations, a leading scientist has warned.

> In a recent study, published in Psychiatric Genetics, Curtis found that a commonly used genetic test to predict schizophrenia risk gives scores that are 10 times higher in people with African ancestry than those with European ancestry. This is not because people with African ancestry actually have a higher risk of schizophrenia, but because the genetic markers used were derived almost entirely from studies of individuals of European ancestry.

https://www.theguardian.com/science/2018/oct/08/genetics-research-biased-towards-studying-white-europeans

Models May Not Perform Well on Novel Data

- **Your responsibility – Question the Application**
  - Is this data similar to what the model was trained on
  - Do different groups (population subgroups) in my predictions get different results
  - Is there a human feedback loop that allows for retraining the model

Summary

Where does Bias come from?

Where Does Bias Come From?

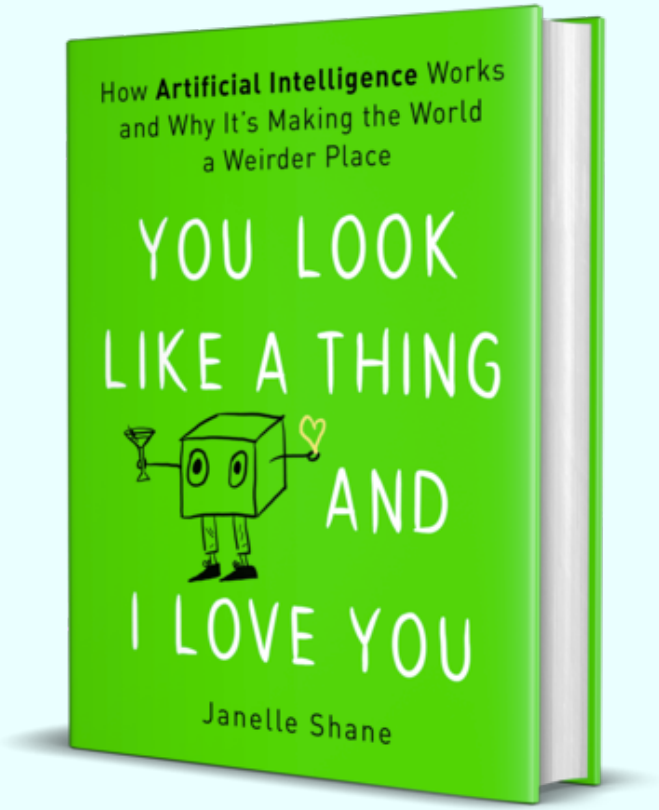
- Like children, models can pick up patterns in the data that we are not explicitly trying to teach them
- There is a **lack of awareness** by Data Scientists/Statisticians about how historical/societal biases may be present in data modeling
  - How we collect data
  - The problems we decide to solve
  - The data we choose to train models on
  - How we assess accuracy
  - How we present the results

Recommendations

- Coded Bias
  - Available on Netflix
- MIT Media Lab researcher Joy Buolamwini discovers that commercial facial recognition software does not see dark-skinned faces, she pushed for legislation
- Joy also has an excellent TED Talk (How I'm fighting bias in algorithms)

Recommendations

- AI Weirdness Blog
  - https://aiweirdness.com/
- Book gives an easy-to-understand look at AI systems and how they can behave unexpectedly



Name: Jim Box

Affiliation: SAS Institute

E-mail: jim.box@sas.com

LinkedIn: https://www.linkedin.com/in/jwbox/