Computing in Context: Fall 2024

Lecture 10 | Intro to ML for Healthcare

Dr Sarah Ashcroft-Jones | Mailman School of Public Health

Section I : Introduction & Overview

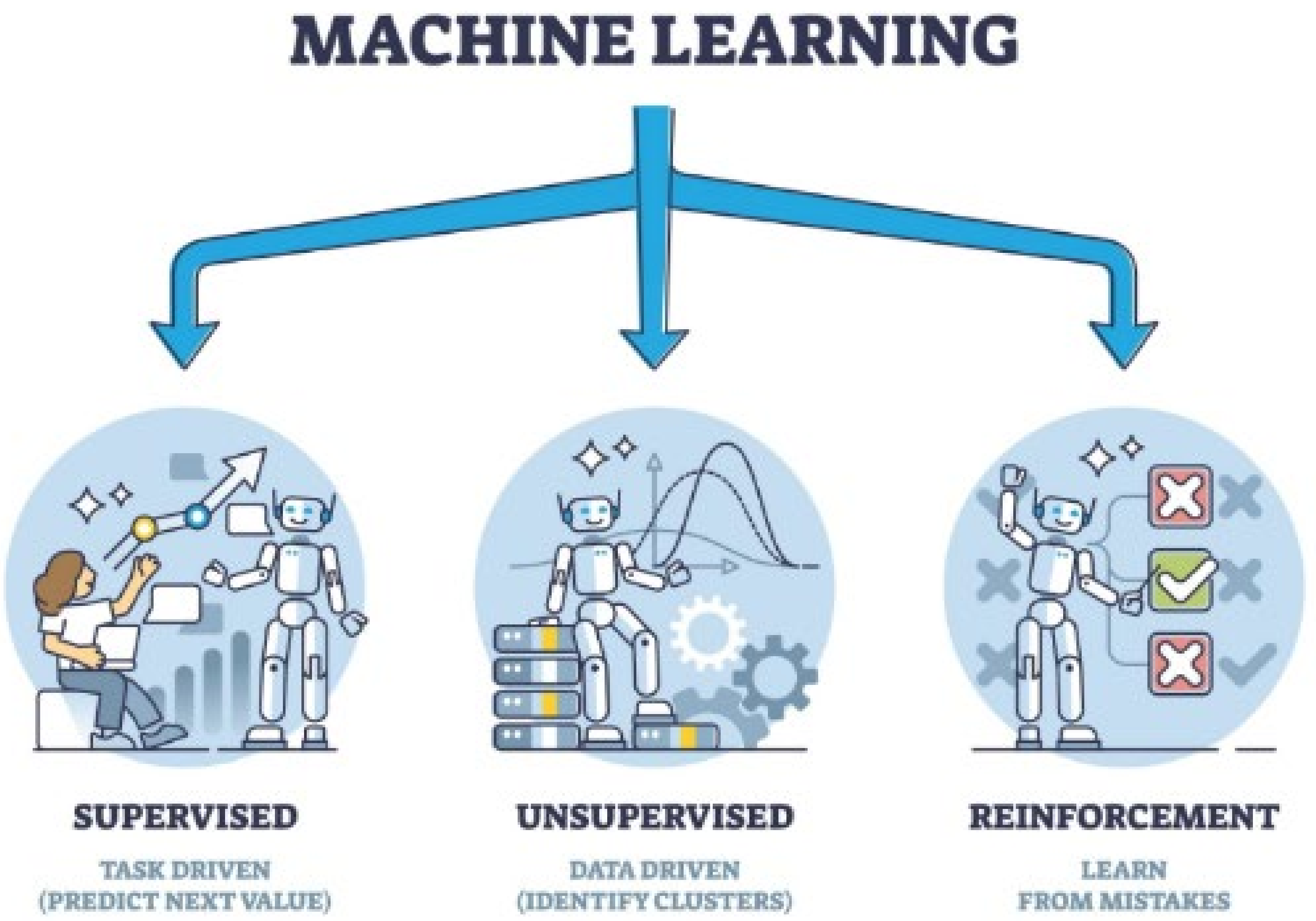
What is Machine Learning?

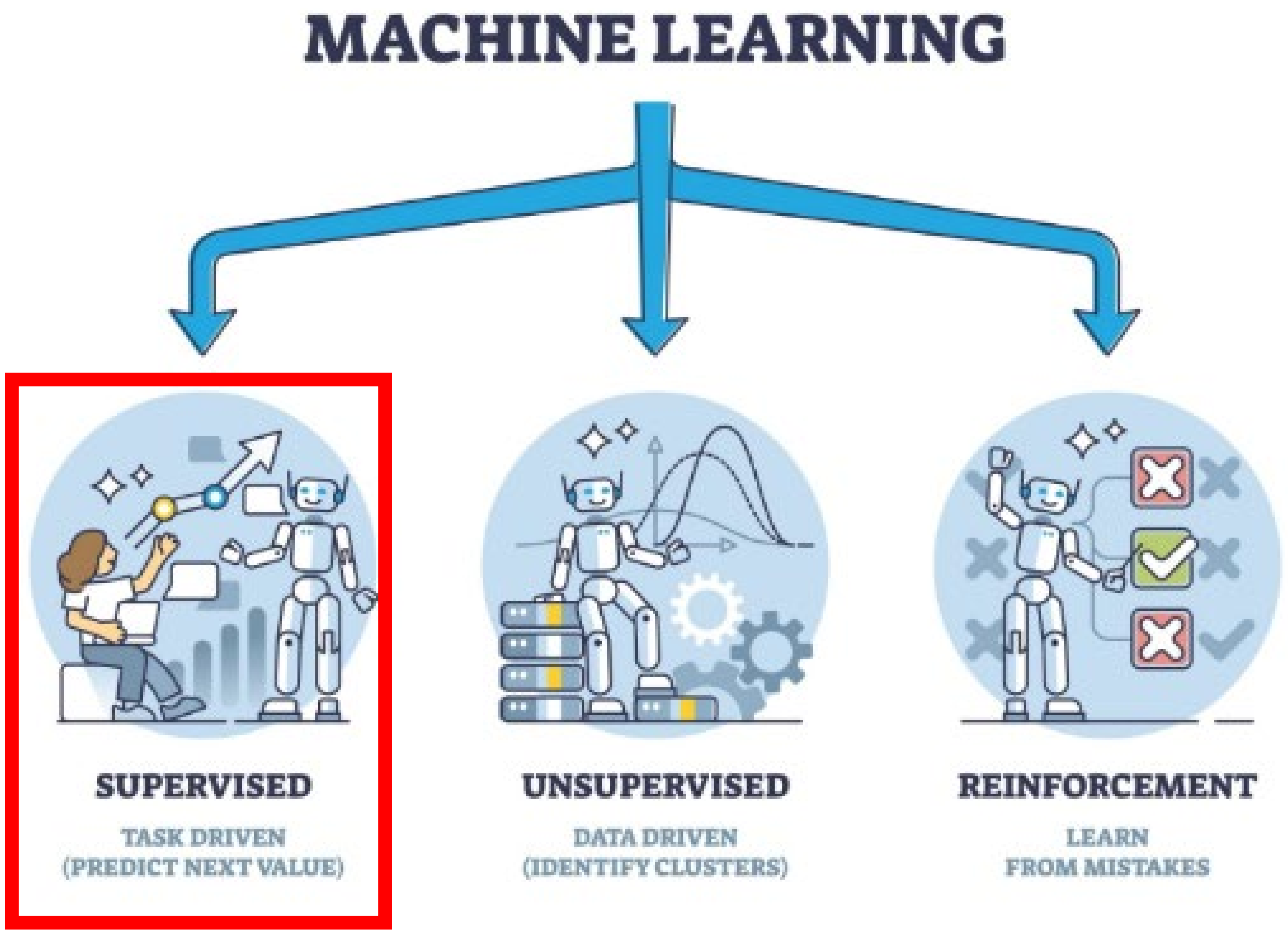
- Machine learning (ML) enables computers to learn from data.
  - teaching a computer to recognize patterns and make predictions
  - e.g. Netflix recommendation, Spotify shuffle, etc.

- ML in healthcare works on improving diagnosis and personalized treatment
  - building predictions and suggestions based on Big Data in healthcare

Key Terms (before we begin!)

- **Data:** Input variables like patient age, lab results.
- **Features:** Attributes used for predictions (e.g., glucose levels).
- **Labels:** Known outcomes (e.g., disease present or absent).


MACHINE LEARNING
SUPERVISED
TASK DRIVEN
(PREDICT NEXT VALUE)
UNSUPERVISED
DATA DRIVEN
(IDENTIFY CLUSTERS)
REINFORCEMENT
LEARN
FROM MISTAKES


MACHINE LEARNING
SUPERVISED
TASK DRIVEN
(PREDICT NEXT VALUE)
UNSUPERVISED
DATA DRIVEN
(IDENTIFY CLUSTERS)
REINFORCEMENT
LEARN
FROM MISTAKES

Learning Types in ML

- **Supervised Learning:** Uses labeled data
  - (e.g., diagnosing a specific disease)

- **Unsupervised Learning:** Finds patterns without labels
  - (e.g., patient segmentation/clustering).

- **Reinforcement Learning:** Learns from outcomes to make decisions
  - (e.g., treatment recommendations).

Why Healthcare Needs Machine Learning

Big Picture:

- Rapid analysis of large datasets for faster, more accurate decisions.

- Potential to improve patient outcomes and reduce healthcare costs.

Down in the details:

- Faster Diagnoses: Analyze medical images in seconds.
- Predictive Analytics: Forecast disease progression.
- Personalized Medicine: Tailor treatments based on genetics.
- Operational Efficiency: Optimize hospital workflows.

Real world impact

EnM Endocrinology and Metabolism http://www.e-enm.org

▸ Endocrinol Metab (Seoul). 2024 Jun 10;39(3):416–424. doi: 10.3803/EnM.2023.1913

Artificial Intelligence Applications in Diabetic Retinopathy: What We Have Now and What to Expect in the Future

Mingui Kong [1], Su Jeong Song [1,2,✉]

▸ Author information ▸ Article notes ▸ Copyright and License information

PMCID: PMC11220221 PMID: 38853435

https://pmc.ncbi.nlm.nih.gov/articles/PMC11220221/#:~:text=Currently%2C%20AI%20applications%20in%20DR,microvascular%20changes%2C%20as%20previously%20mentioned.

Real world impact

ACI
Applied Clinical Informatics
Thieme

Appl Clin Inform. 2020 Sep 2;11(4):570–577. doi: 10.1055/s-0040-1715827

Implementation of Artificial Intelligence-Based Clinical Decision Support to Reduce Hospital Readmissions at a Regional Hospital

Santiago Romero-Brufau [1,2], Kirk D Wyatt [3], Patricia Boyum [1], Mindy Mickelson [1], Matthew Moore [1], Cheristi Cognetta-Rieke [4]

Author information Article notes Copyright and License information

PMCID: PMC7467834 PMID: 32877943

https://pmc.ncbi.nlm.nih.gov/articles/PMC7467834/#:~:text=Clinical%20Relevance%20Statement,period%20(%20p%20%3C%200.001).

Section II : How ML Works

ML Workflow in Healthcare: Overview

Data Collection: EHR, imaging, genomics.

Data Preprocessing: Cleaning, normalization.

Model Training: Feed data into the algorithm.

Evaluation: Validate on new data.

Deployment: Integration into clinical practice.

ML Workflow in Healthcare: Rule 1

"Garbage in Garbage out"

Rule 1: *Garbage in Garbage out*

RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE • 25 Oct 2019 • Vol 366, Issue 6464 • pp. 447-453 • DOI: 10.1126/science.aax2342

https://www.science.org/doi/full/10.1126/science.aax2342

ML Workflow in Healthcare: Rule 2

“All models are wrong, but some are useful”
George Box

Rule 2: *All models are wrong*

- **Decision Tree:** A simple, intuitive algorithm that splits data based on feature values into branches to make predictions.

- **Random Forest:** An ensemble of multiple decision trees to improve accuracy and reduce overfitting. It averages the predictions from many trees for a more robust result.

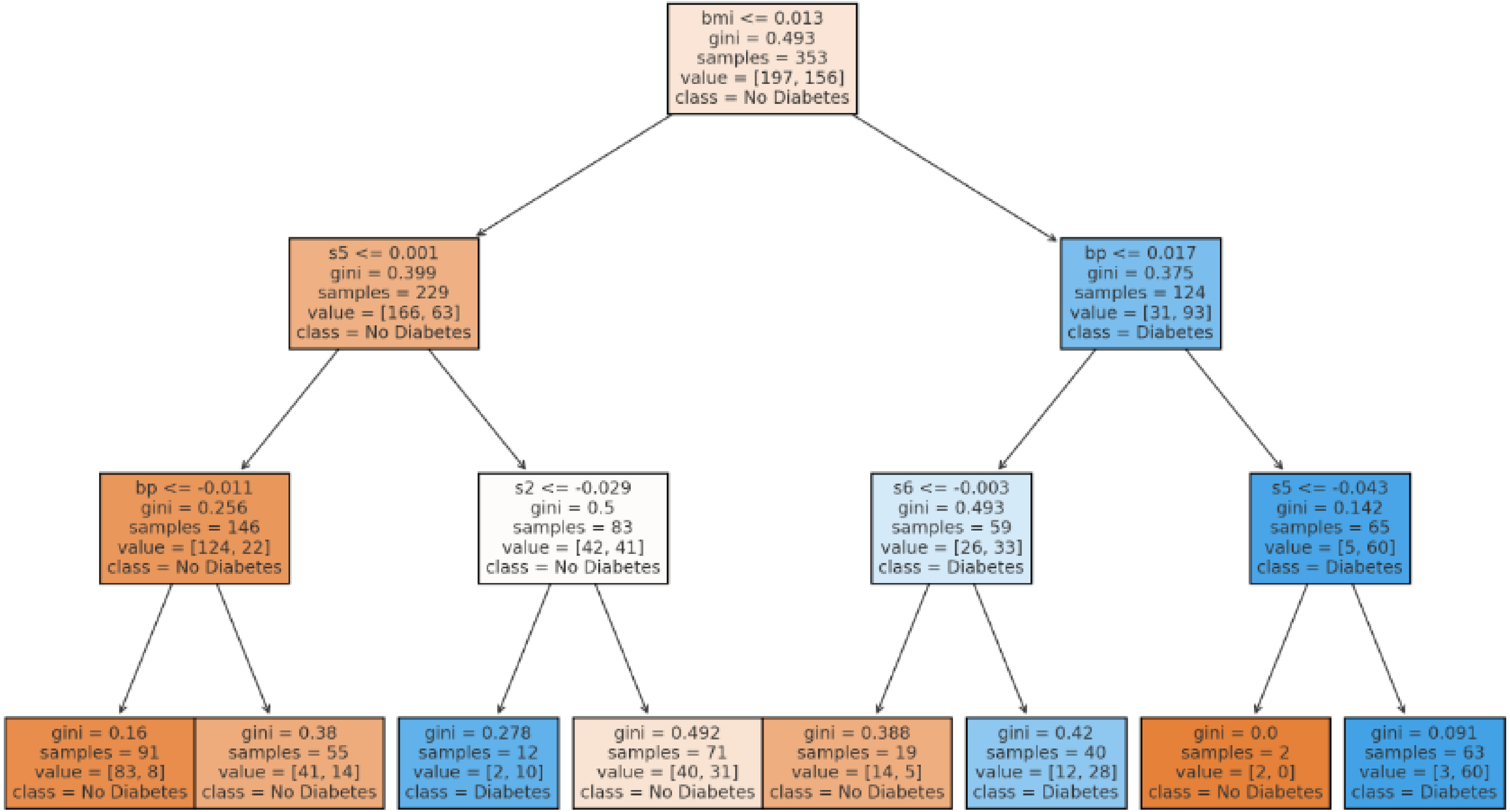
- **CNNs:** A type of deep learning algorithm specialized in processing grid-like data such as images. CNNs use layers of filters to automatically detect patterns and features in images.

Rule 2: *Some models are useful*

Algorithm	Key Strengths	Common Use Cases
Decision Tree/Random Forest	Interpretability, flexibility	Disease diagnosis, treatment recommendations
CNN (Convolutional Neural Network)	Image analysis, high accuracy	Medical imaging, pathology detection

Section III : Worked Example

Decision tree classifier in Python



Core ingredients of a "decision": Splitting Criterion

- **Gini Impurity:** Measures how often a randomly chosen element would be incorrectly labeled if randomly classified based on the distribution of labels in the dataset.
- **Entropy (Information Gain):** Measures the randomness in the dataset. The algorithm selects splits that maximize **Information Gain**, which is the reduction in entropy after the split.

Key Issue:

Instability: Small changes in data can lead to different splits and different trees.

Core ingredients of a "decision": Stopping Criteria

- max_depth: Limits the depth of the tree to prevent overfitting.
- min_samples_leaf: The minimum number of samples that must be present in a leaf node.

Key Issue:

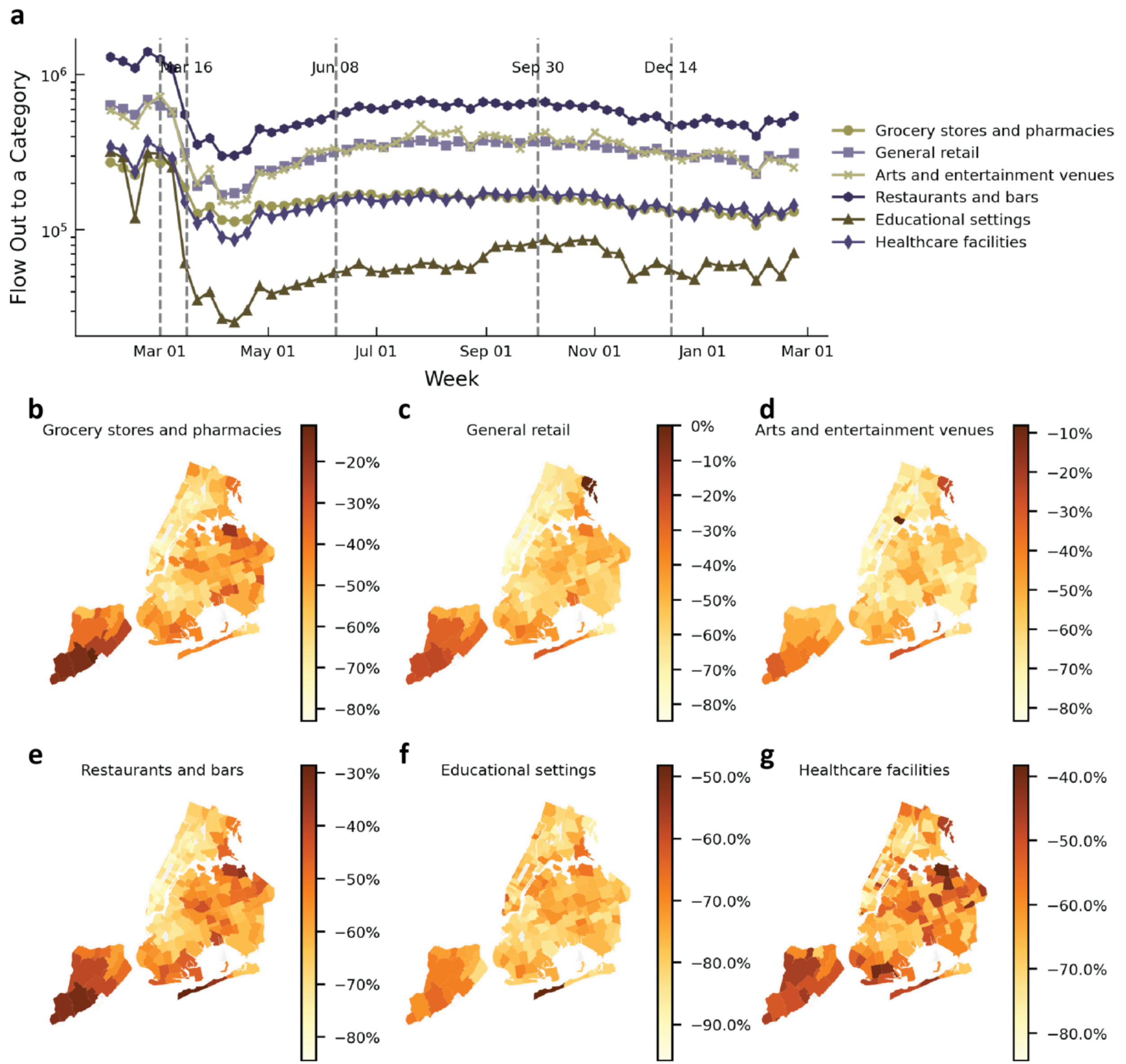
Overfitting: Decision trees can easily overfit to training data, especially if the tree is too deep.

Example in Python:

Example in my work:

Human mobility in times of crisis:

- Pandemic case study
- Across NYC zipcodes and boroughs
- Who went where, when, and why?
- Behavioral measures and agency



Example in my work:

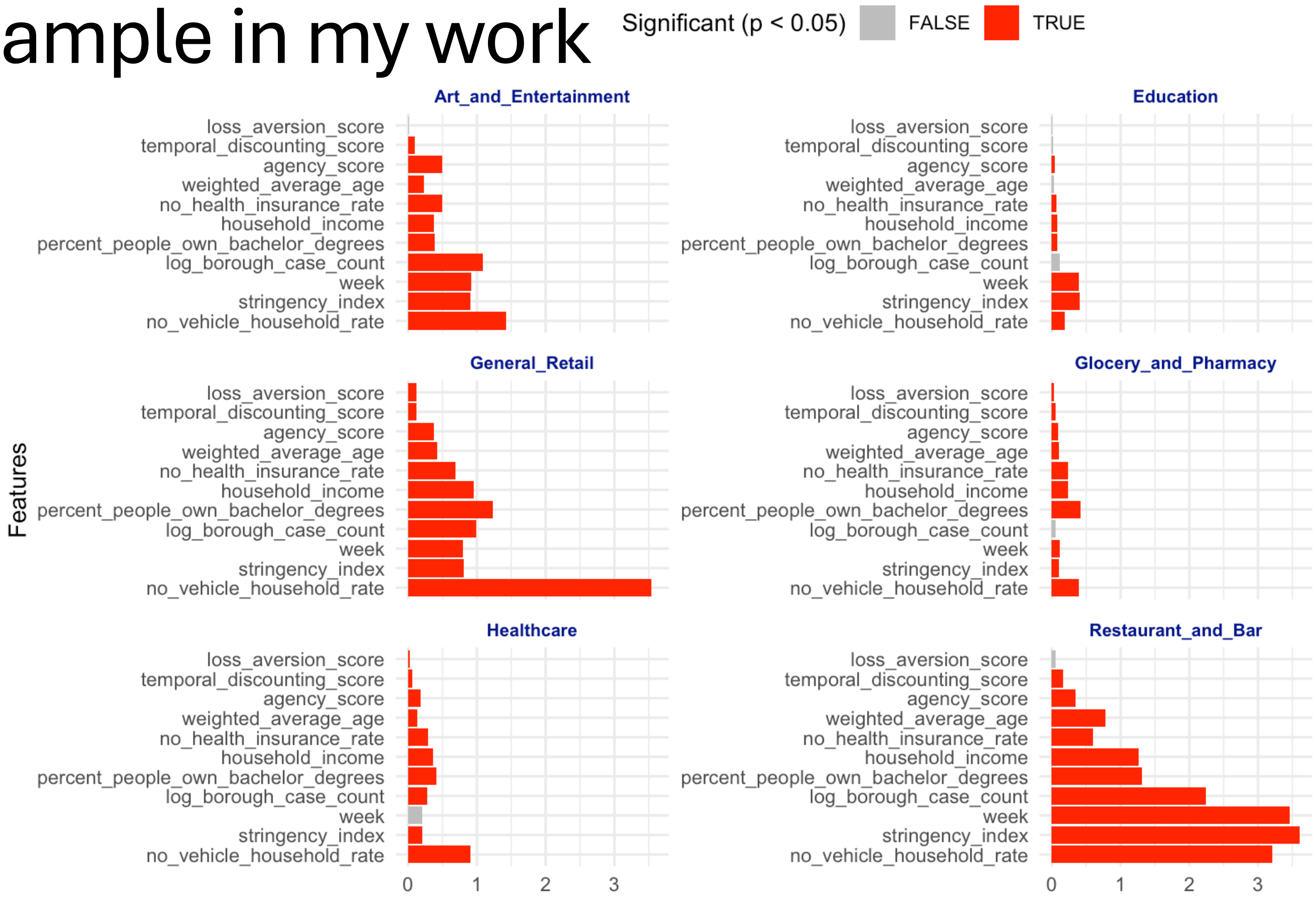
Random Forests:

ML algorithm that combines multiple independent decision trees, each trained on a random subset of the data.

- At each split, randomly choose a subset of features to consider (no overfit)
- When making a prediction, aggregate the results from all trees by taking the majority vote (classification) or averaging (regression)

N.B: Provides a built-in mechanism to identify the most important features in the data based on how frequently they are used in the decision trees

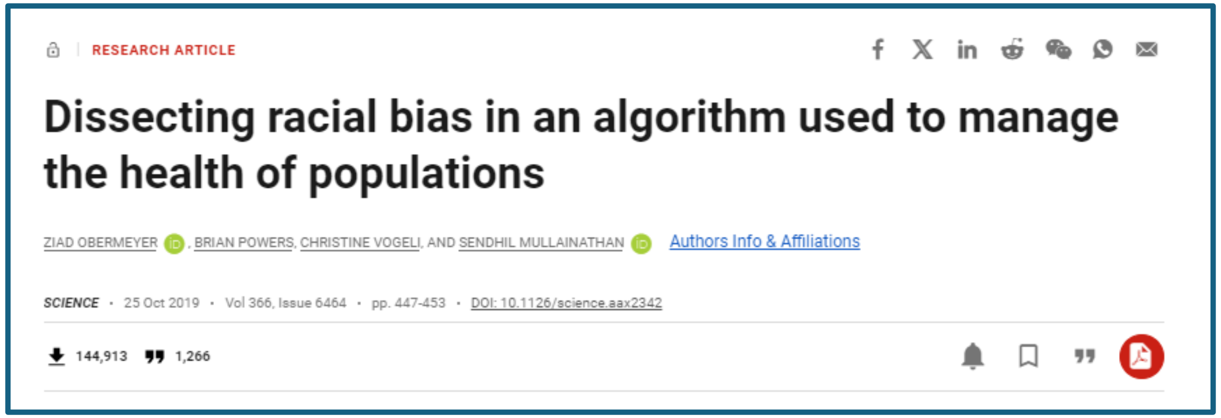
Example in my work



Section IV : Discussion

Discussion points:

- How could ML be applied in underserved healthcare systems?
- What ethical concerns resonate most with you?
- Where can you imagine bias emerging in ML models?



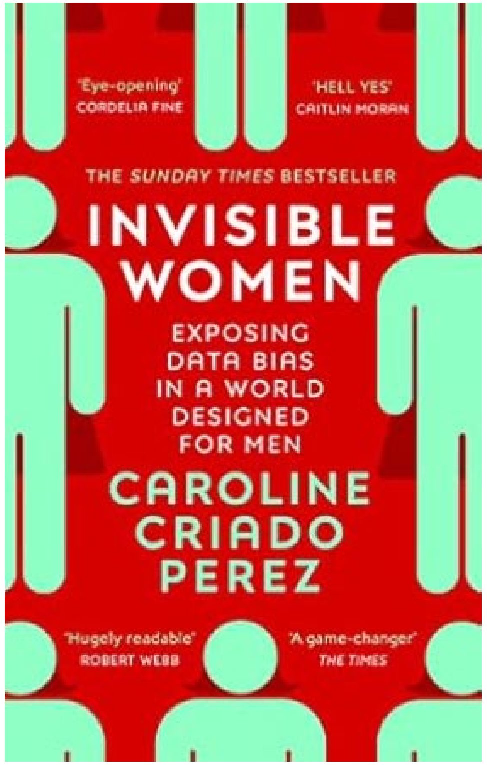
RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE • 25 Oct 2019 • Vol 366, Issue 6464 • pp. 447-453 • DOI: 10.1126/science.aax2342

144,913 1,266



Free Resources
(aka you don’t need a paid bootcamp)

- Courses:
  - AI and Medical Diagnosis Coursera (applied, light touch),
  - ML Engineering (more abstract and heavy),

- Books:
  - Little book of ML (short and sweet),
  - Big Book of ML (advanced),
  - DeepLearning (overview),
  - d2l (comprehensive DL overview),

- General/Misc:
  - LLMs (cool website for explanatory visualizations of LLMs and genAI)

Questions?