BIOS 635: Introduction to Machine Learning

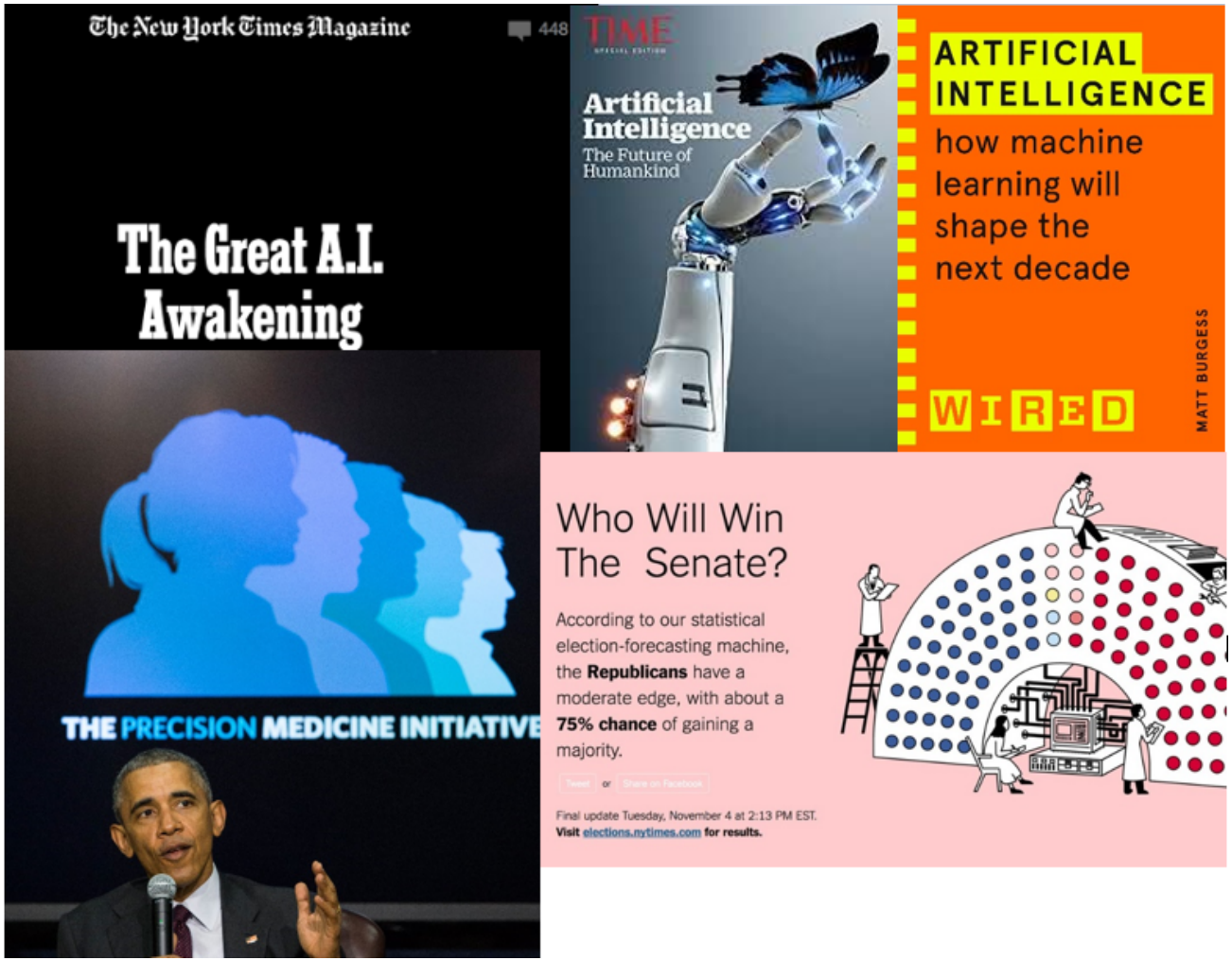
Kevin Donovan

UNC-Chapel Hill

1/19/2021

Welcome

What is machine learning?



Welcome


REMEMBER, WITH GREAT POWER
COMES GREAT RESPONSIBILITY
memegenerator.net

Welcome

About Me

PhD student in Biostatistics
UNC at Chapel Hill
Basketball and Green Bay Packers enthusiast
Hip-Hop, Jazz, Funk, Metal music fanatic

Welcome

Structure for class:

- Lecture - Tues/Thurs 9:30-10:45AM
- Office Hours - Wed 10-11AM
- Weekly homework assignments, bi-weekly article evaluations
- Midterm exam on real data analysis, final group project
- Implementation of methods taught through R

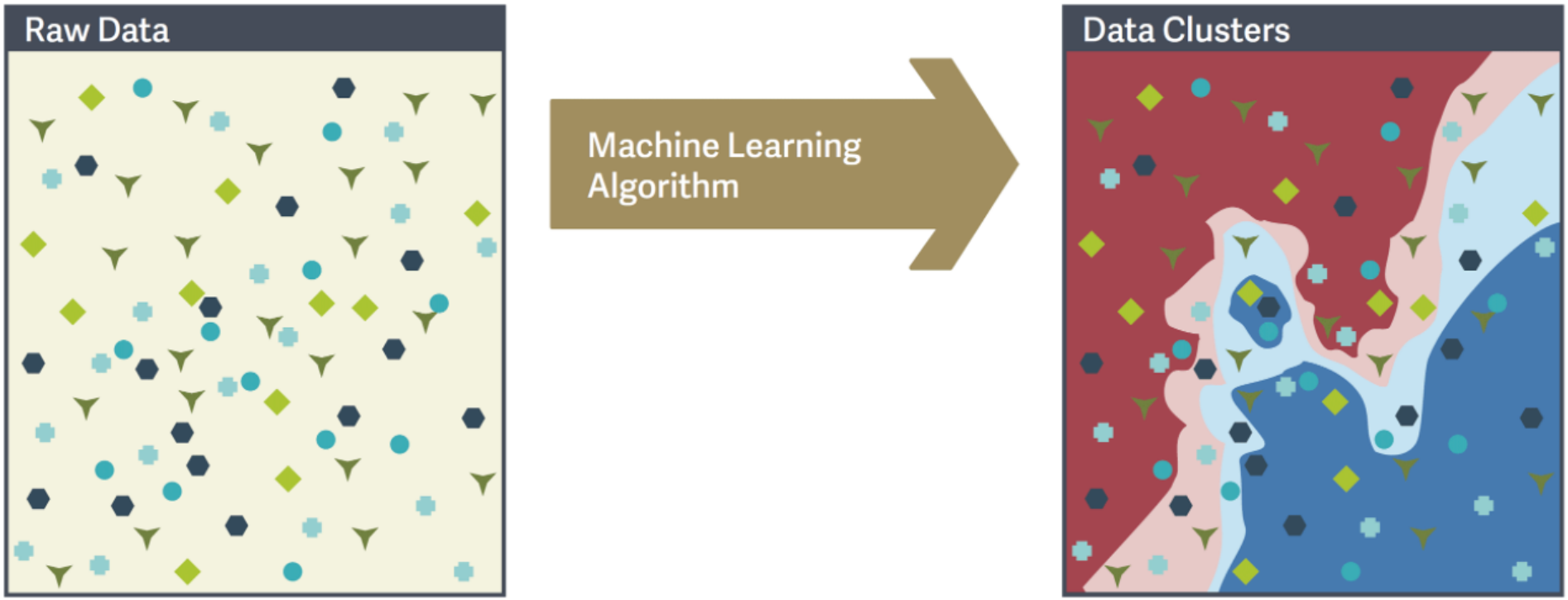
Welcome

Goals for class:

1. Primary: Promote the understanding of machine learning conceptually from a stats framework
2. Primary: Teach implementation of methods in R
3. Secondary: Promote good analytic practices, with focus on
    1. data management
    2. data visualization/tabulation
    3. exploratory analysis
    4. **reproducibility**
4. **Develop a critical eye for machine learning and its application in science and society at large**

Introduction

Machine Learning:
Development and application of methods to identify patterns in data

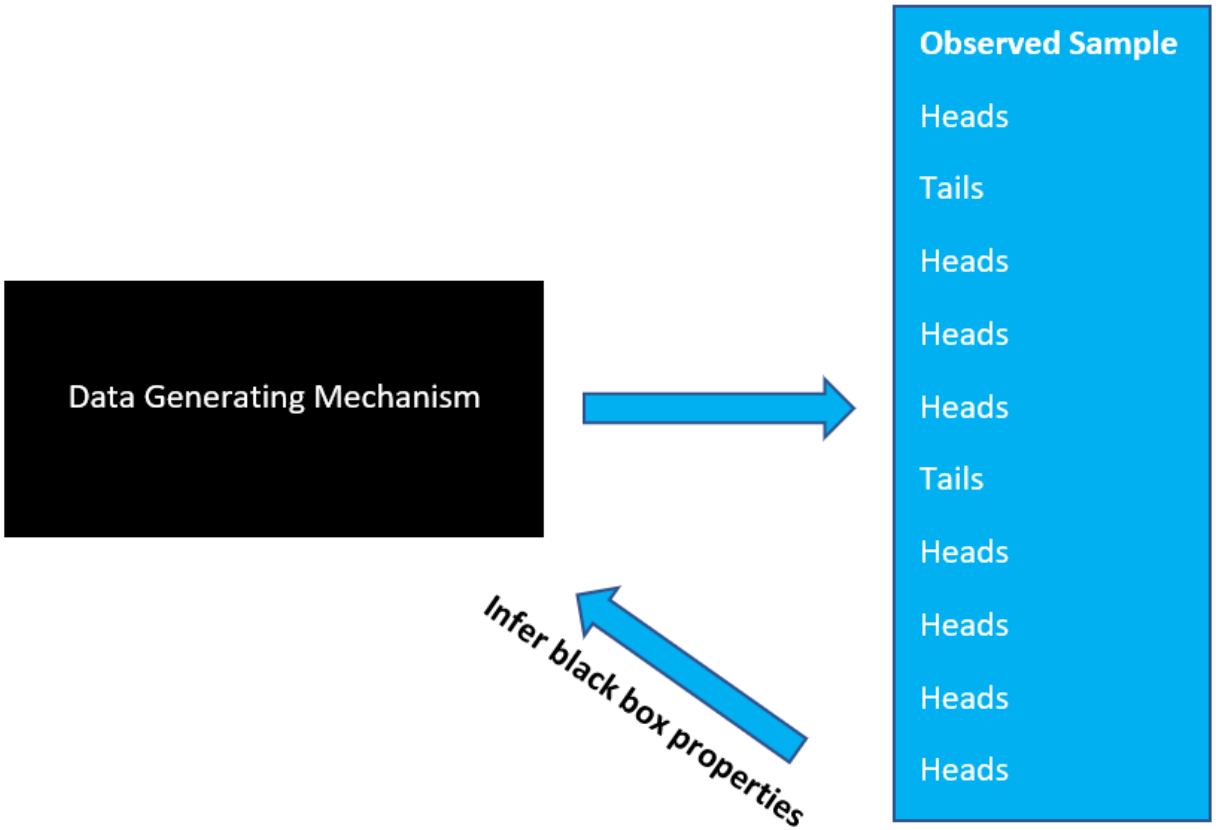


Introduction

How to identify these patterns?

- Patterns are complex, how to model?
- How much modeling do we enforce on the data?
- How to evaluate a method's performance?
- What population are we analyzing? Who is benefiting from the analysis? Who is being left out?

Statistical Inference

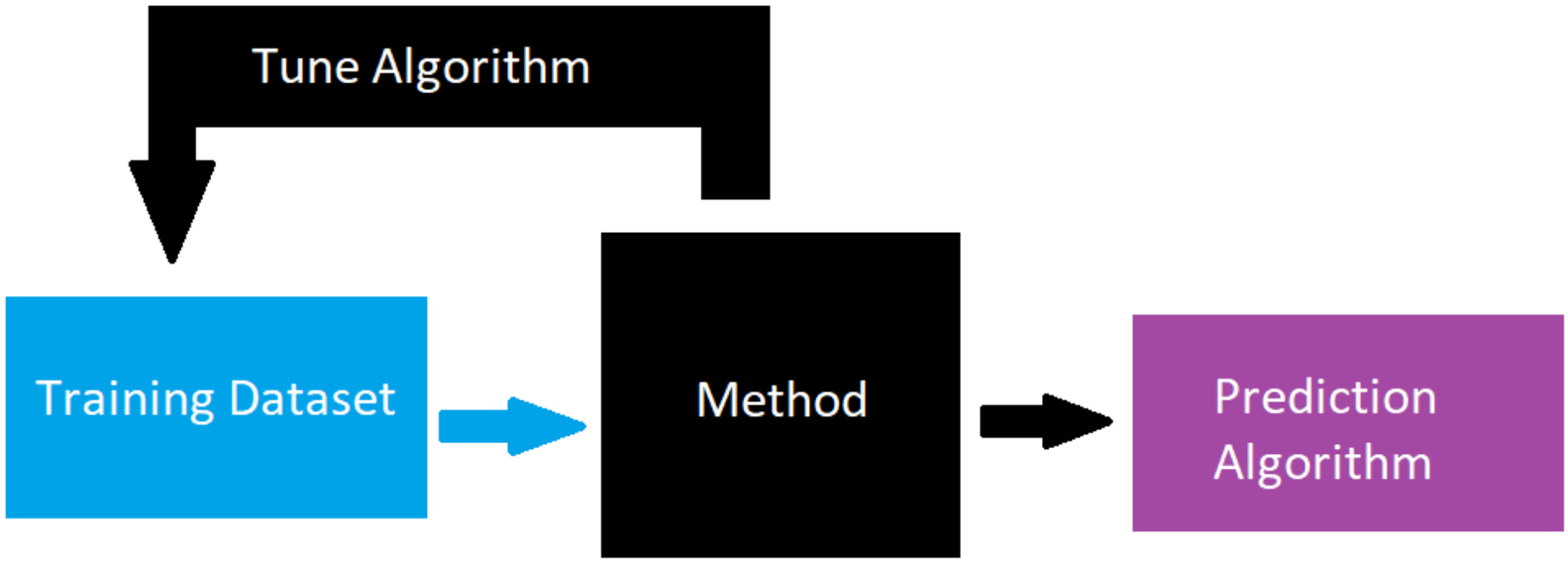


Black box may be "nuisance", may only want to predict output

Branches of machine learning

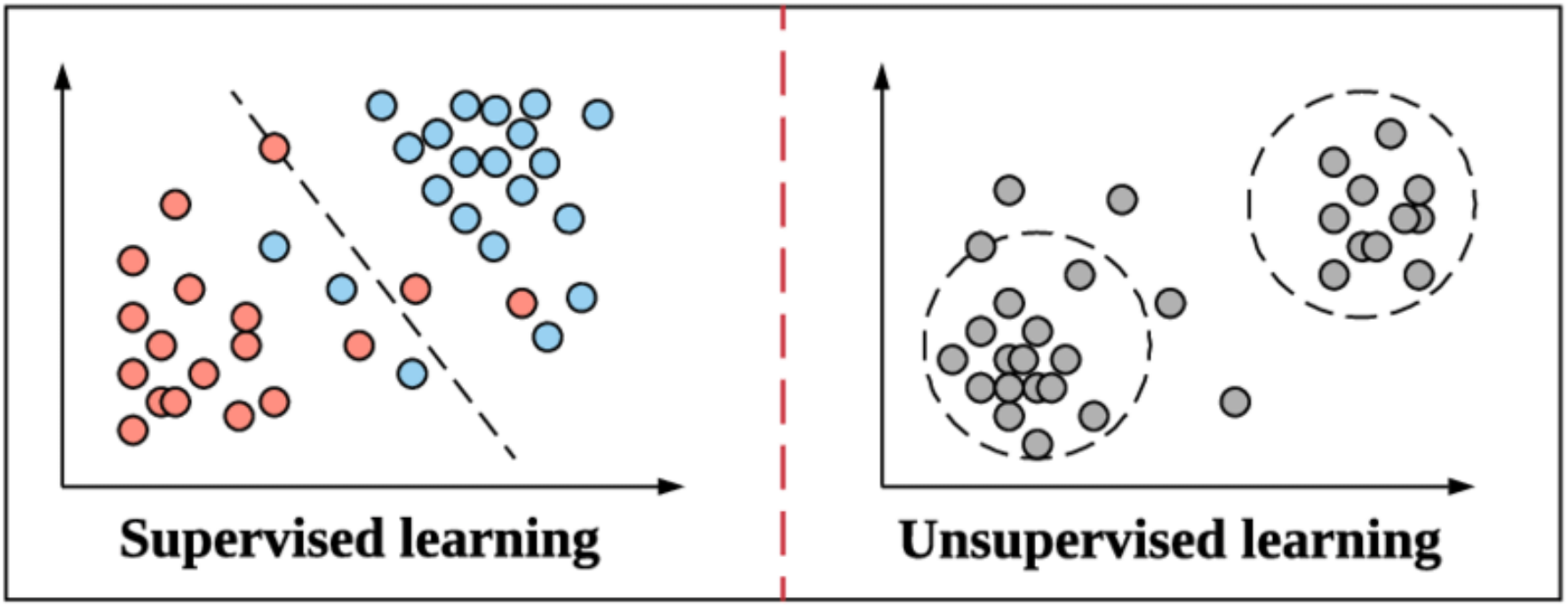
- Supervised Learning
  "Ground truth" observed for all subjects, used to "train" method
- Unsupervised Learning
  "Ground truth" not observed for any subjects, is "latent"
- Reinforcement Learning
  Algorithm "self-updates" based on data and valuation system

Supervised and Unsupervised:
Training the method



Can we then **test** the method on the same data?
How training done when we **don't see the outcome**?

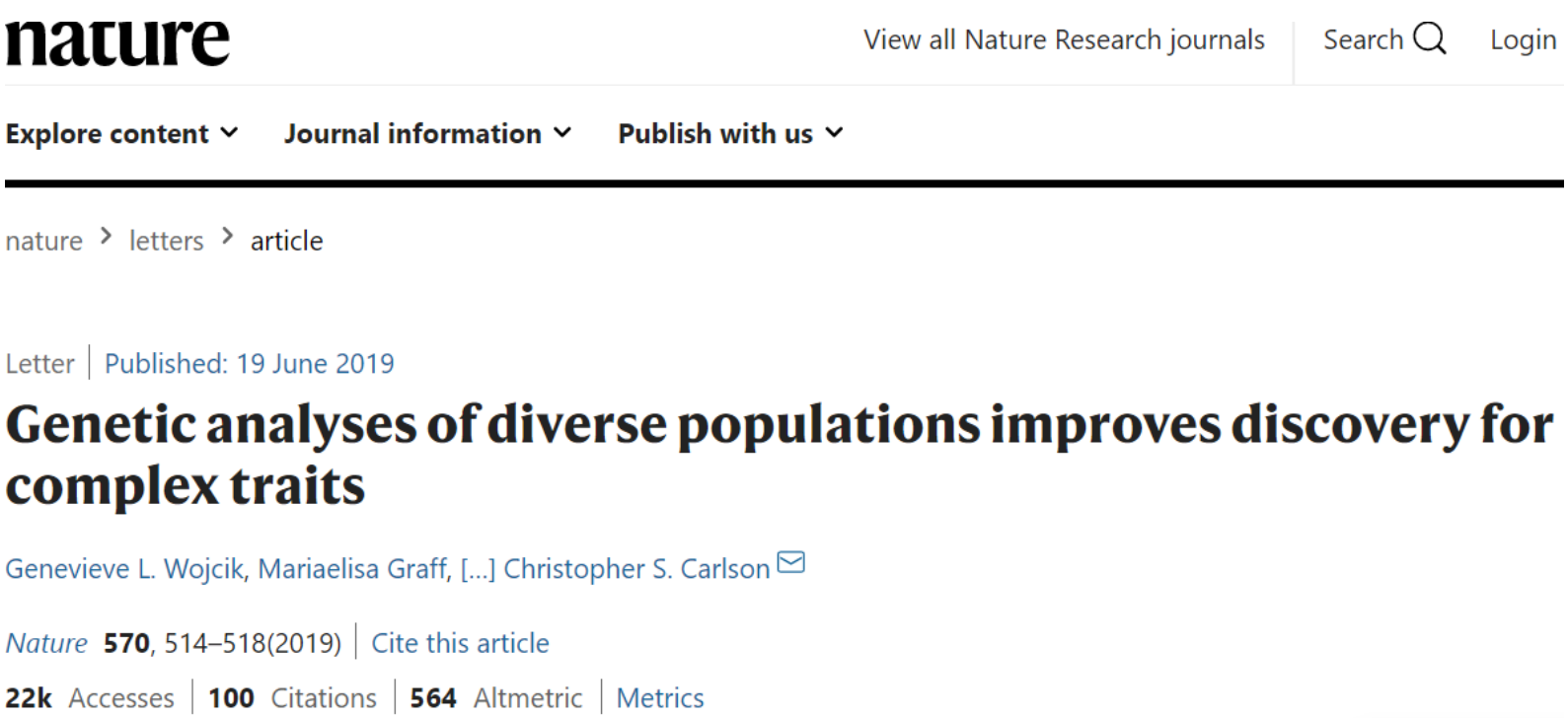
Supervised and Unsupervised



Introduction

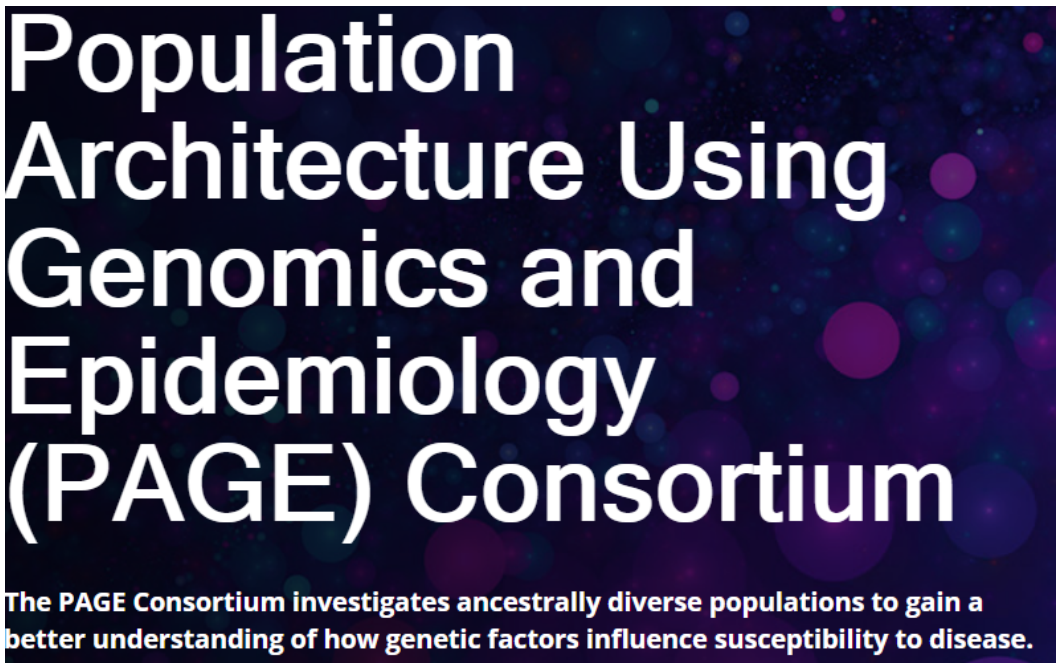
Patterns found **completely dependent** on training data

- What population is this training data from?
- What communities are represented, which are not?
- Why are these communities not represented?
- Who benefits from insights learned from method?

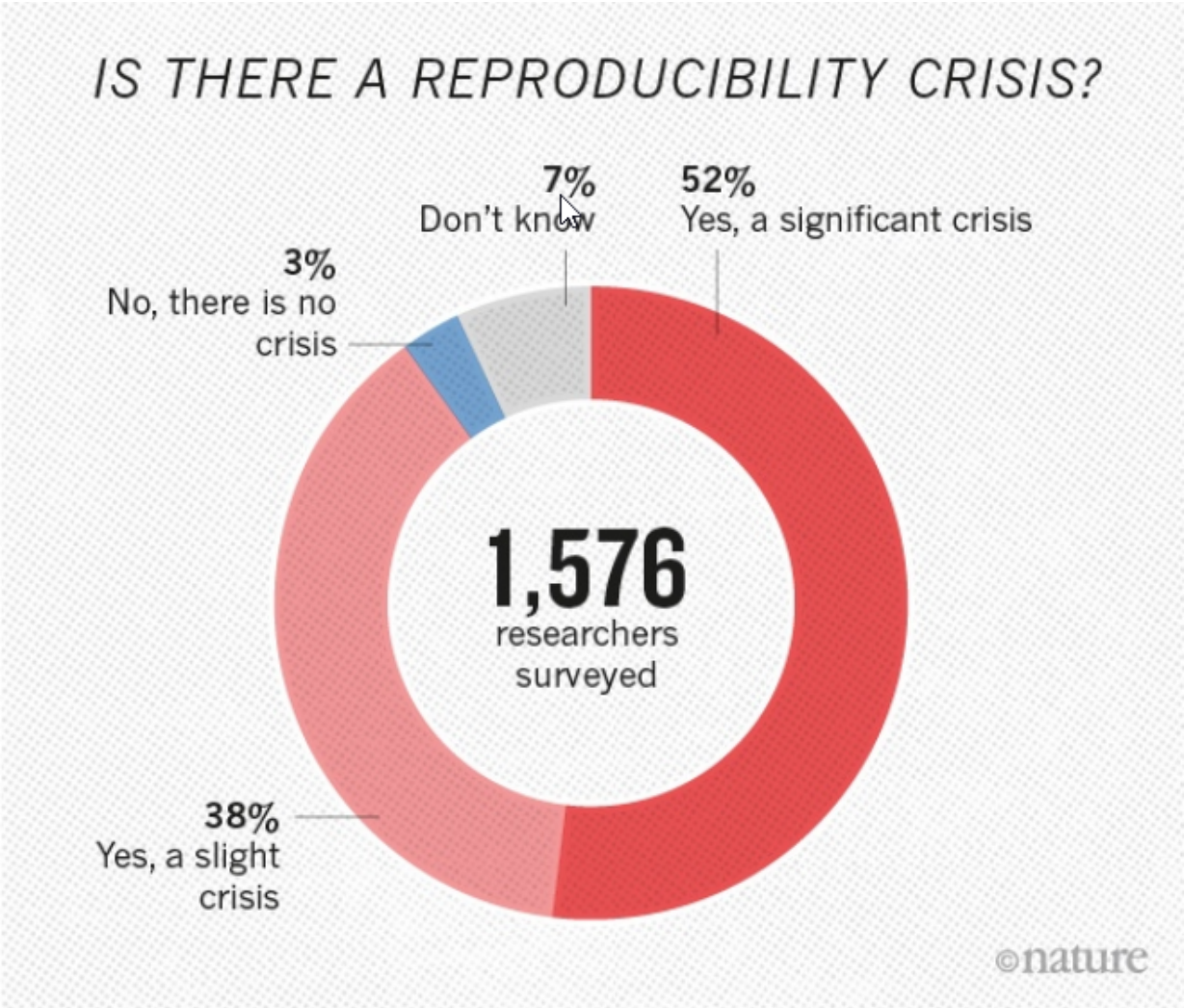


Introduction

- Many communities largely left behind in medical machine learning research
- Reflects systemic inequities in other parts of society at large
- As data analysts, need to combat and consider in our research



Reproducibility in the Computing Age



Introduction

Evidence shows low reproducibility of scientific research despite

1. Extensive professional specialization
2. Well-developed methodology for study design and data analysis
3. Explosion in open source statistical software and computing tools
4. Software supported by extensive documentation

How can we turn the tide on the reproducibility issue?

Introduction

Song of the session:
Pomp & Pride by Toots and the Maytals