Analyzing Gender, Racial, & Ethnic Disparities & Predicting Priority Scores of NIDA Grant Portfolio

Office of Research Training, Diversity, and Disparities (ORTDD)
National Institute on Drug Abuse (NIDA)

CIVIC DIGITAL FELLOWSHIP

NIH

Alex Hayward
The University of Chicago
Molecular Engineering & Data Science

NIDA Office of Research Training, Diversity, and Disparities (ORTDD) Mission

The **Office of Research Training, Diversity, and Disparities** (ORTDD) is committed to developing a cutting-edge, diverse research workforce to address urgent public health substance use and addiction problems.

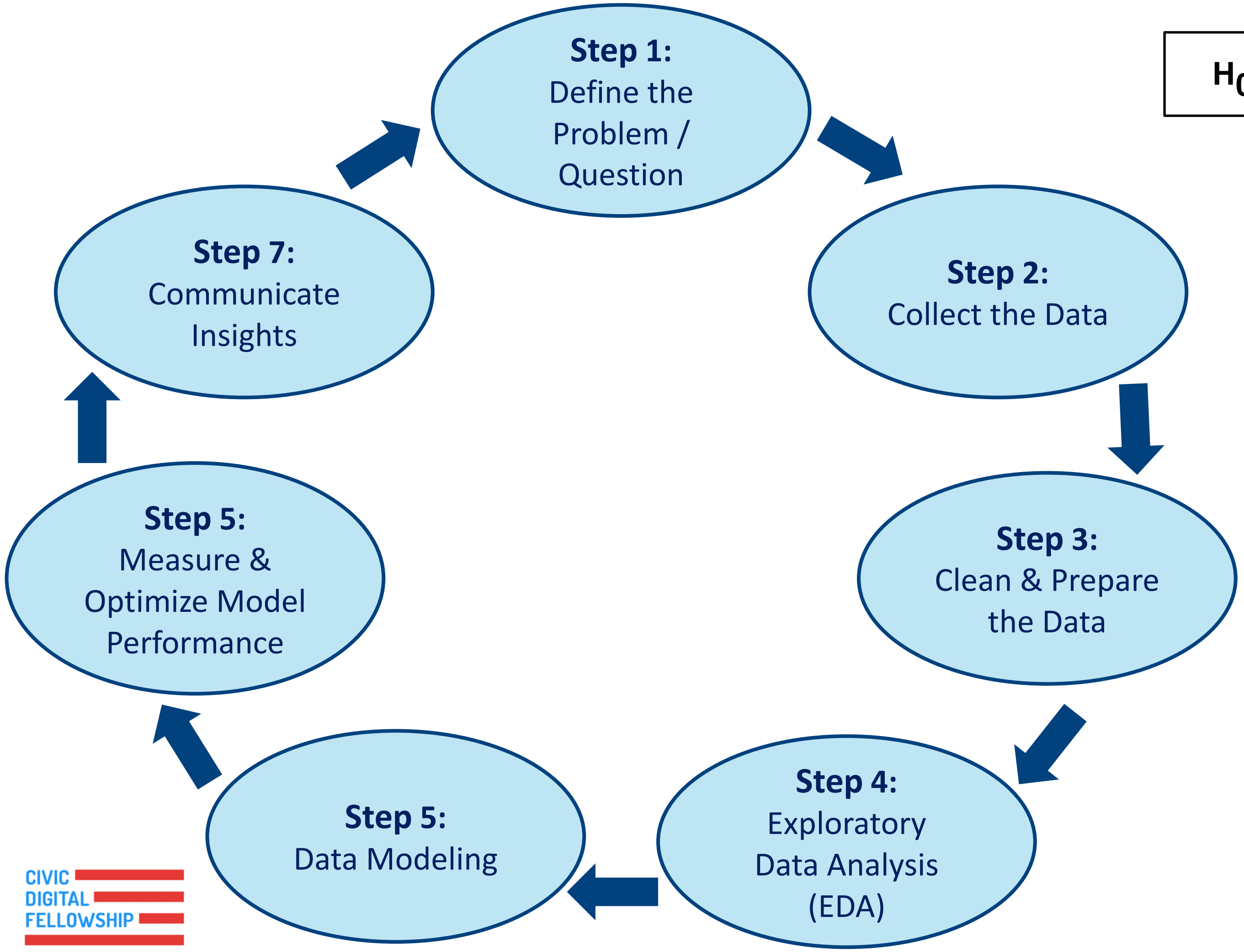
ORTDD funds **training, career development,** and **research grants** to support substance use and addiction research scientists throughout the career pipeline, with a focus on the development of **underrepresented researchers.**

Project Goals: Using 2010 – 2020 NIDA grant portfolio data,

- Analyze disparities in NIDA grant applications and awards
- Identify factors that predict whether a NIDA grant application is likely to be funded
- Derive insights to eliminate disparities and increase racial and ethnic equity in substance use and addiction science

CIVIC DIGITAL FELLOWSHIP

The Data Science Life Cycle



H_0: No disparities exist in NIDA grant applications or awards.

Research Questions

1. Do **disparities** exist in the NIDA grant application and/or award pools?
2. What factors **predict** whether a NIDA grant application is likely to be funded (i.e., receive a **Priority Score < 30**)?
3. What can we do to **eliminate disparities**?

Where is **outreach** most needed?
Among prospective applicants at **institutions** or stakeholders in **review**?

Methods

Data: Anonymized data on all NIDA grant applications

- FY 2010 – FY 2020
- 35,035 rows x 34 columns
- NIH Office of Extramural Research
- 1 row = 1 grant application
- New and competing awards, all NIH grant mechanisms
- Type 1 and 2

Variables: Fiscal Year, Activity Code, Priority Score, Organization Name, Age, Gender, Race, Ethnicity, Early-Stage Investigator (ESI)…

Engineered Features (Y/N): Priority Score < 30, Minority-Serving Institution (MSI), Top 50 NIH-Funded Institution…

Data were **cleaned, visualized,** and **analyzed** using

- Python (Jupyter Notebook via Anaconda)
  - Pandas, NumPy, SciPy, Scikit-Learn, Statsmodels, Matplotlib, Seaborn, Plotly

Statistical Analyses

- Data quality assurance
- Data visualization
- Hypothesis testing
- Feature importance

Machine Learning Models

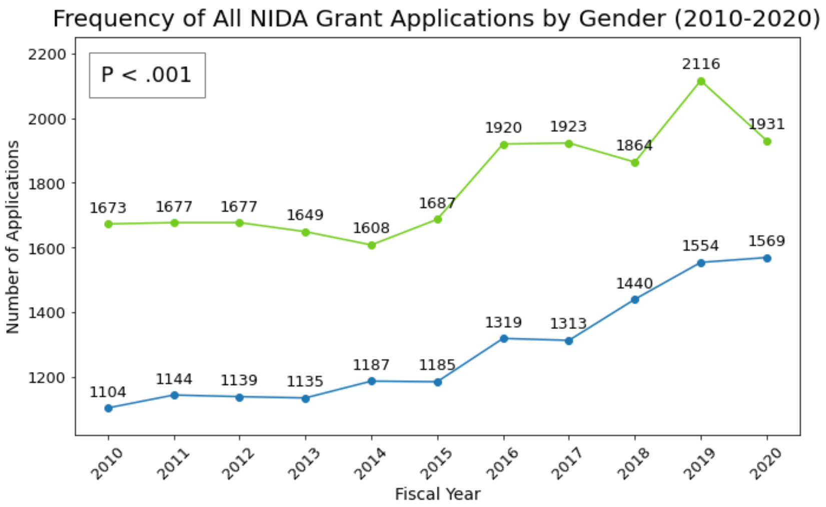
- Logistic Regression *(baseline)*
- Random Forest *(baseline)*
- XGBoost

CIVIC DIGITAL FELLOWSHIP

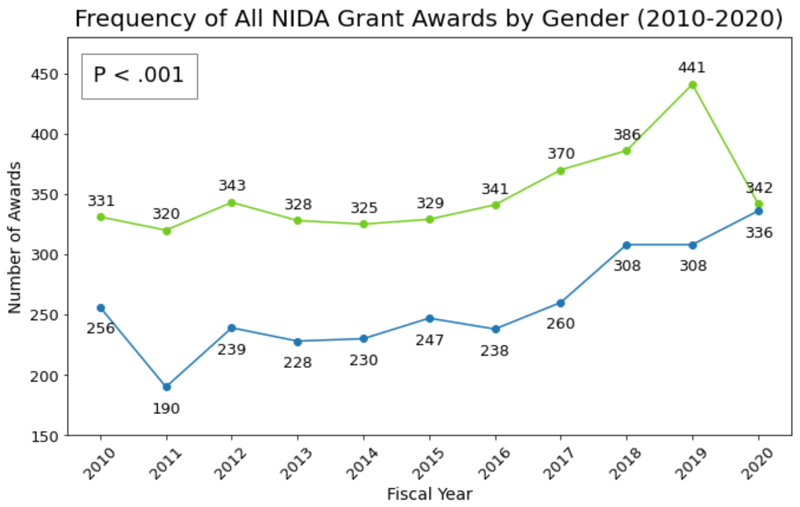
Gender Disparities in Applications & Awards Are Decreasing

New & Competing Applications/Awards, All Grant Mechanisms

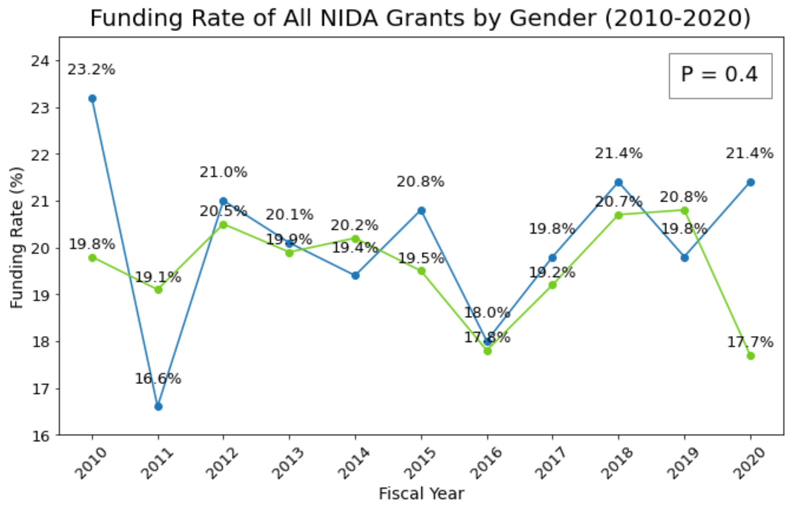
Applications



Awards



Funding Rate



Applications/Awards with ‘Unknown’ or ‘Withheld’ data for ‘Gender’ not shown.

Funding Rate

$$= \left(\frac{\text{X Awards}}{\text{X Applications}}\right) \text{x } 100\%$$

CIVIC DIGITAL FELLOWSHIP

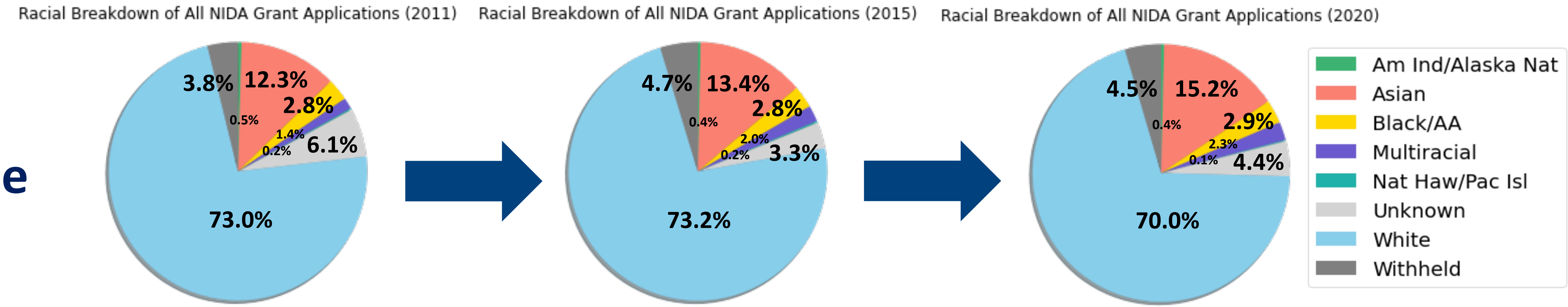
Over the Past Decade, Little Increase in Racial & Ethnic Diversity

New & Competing Applications, All Grant Mechanisms

FY 2011 | FY 2015 | FY 2020

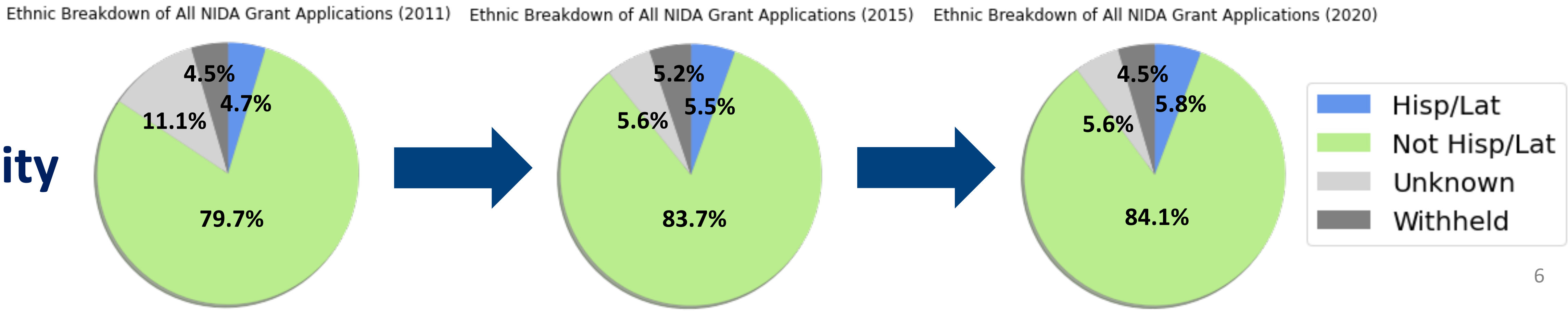
P < .001 (%)

Race



P < .001 (%)

Ethnicity

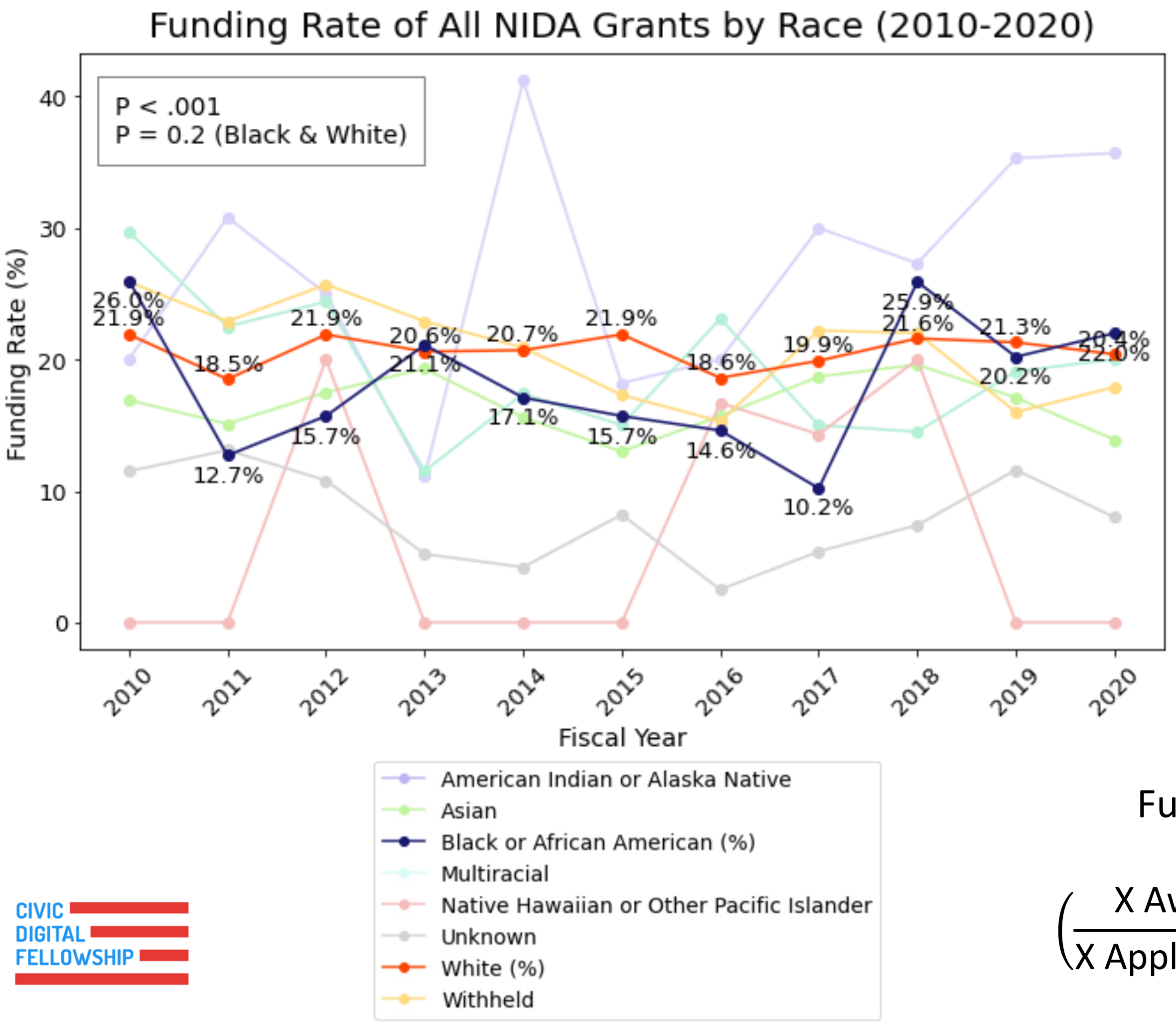


CIVIC DIGITAL FELLOWSHIP

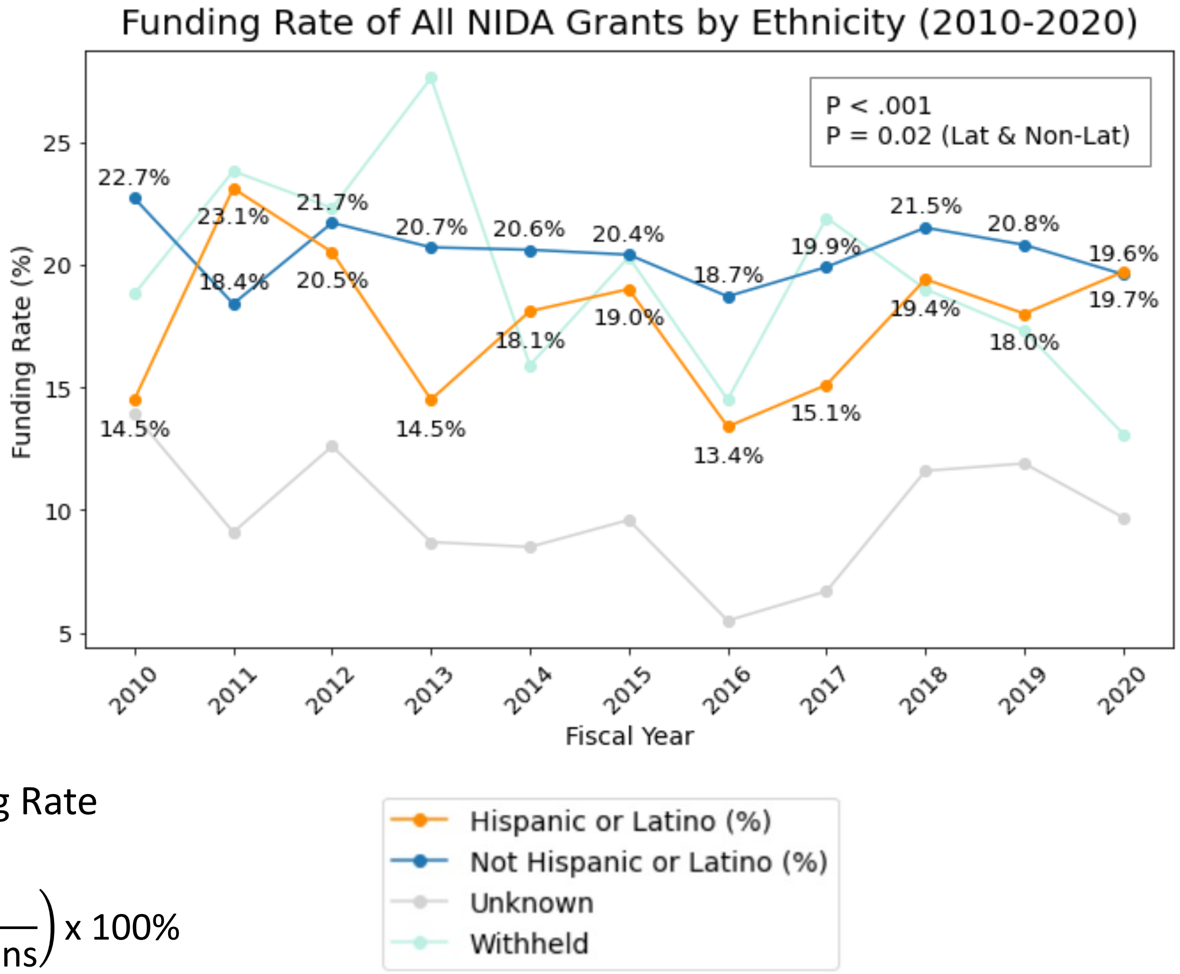
Racial & Ethnic Disparities in Funding Rates Are Decreasing

New & Competing Applications, All Grant Mechanisms

Race



Ethnicity



Funding Rate

$$= \left(\frac{\text{X Awards}}{\text{X Applications}}\right) \text{x } 100\%$$

CIVIC DIGITAL FELLOWSHIP

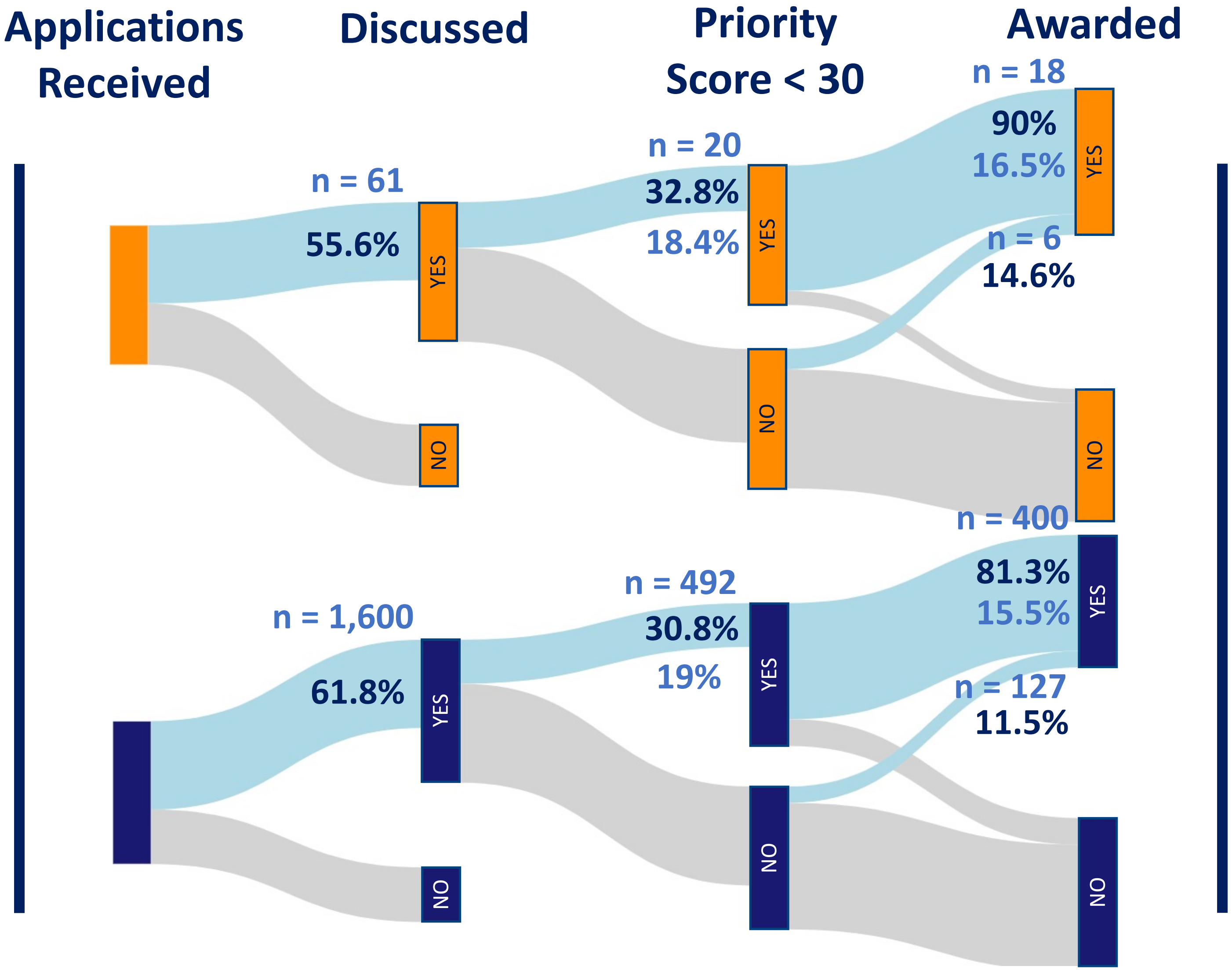
NIDA Grant Review Outcomes by Race (FY 2020)

New & Competing Applications, All Grant Mechanisms

Total n Applications
% of X Race Applications

Applications from Black Applicants
N = 109
2.9%

Applications from White Applicants
N = 2,588
70%



Awards to Black Applicants
N = 24
Funding Rate = 22%

Awards to White Applicants
N = 527
Funding Rate = 20.4%

Predictive Modeling Problem Definition

Objective: Using FY 2010 – FY 2020 NIDA Research Project Grant (RPG) application data, predict whether a NIDA RPG application is likely to be funded, or receive a **Priority Score < 30**.

Target Variable: Whether a NIDA grant application receives a Priority Score < 30

Binary Classification Rule

- 1 = Priority Score < 30
- 0 = Priority Score ≥ 30 / Not Discussed

Features:

- Fiscal Year
- Gender
- Race
- Ethnicity
- Early-Stage Investigator (ESI)
- Minority-Serving Institution (MSI)
- Top 50 NIH-Funded Institution
- A1
- Application Type Code
- Activity Code (RPGs Only)

CIVIC DIGITAL FELLOWSHIP

Building & Optimizing Machine Learning Models

Machine Learning Models

1. Logistic Regression *(baseline)*
2. Random Forest *(baseline)*
3. XGBoost

Preprocessing Data

- Clean and normalize data
- Feature engineering
- One-hot encoding
- Balance classes
- Split data into train (80%) and test (20%) sets

Building & Optimizing Models

- Tune hyperparameters
- Fit classifier on train set
- Make predictions on test set
- Measure and optimize model performance
- Analyze feature importance

Logistic Regression & Random Forest

Baseline Machine Learning Models

1 = Priority Score < 30

0 = Priority Score ≥ 30 / ND

Logistic Regression

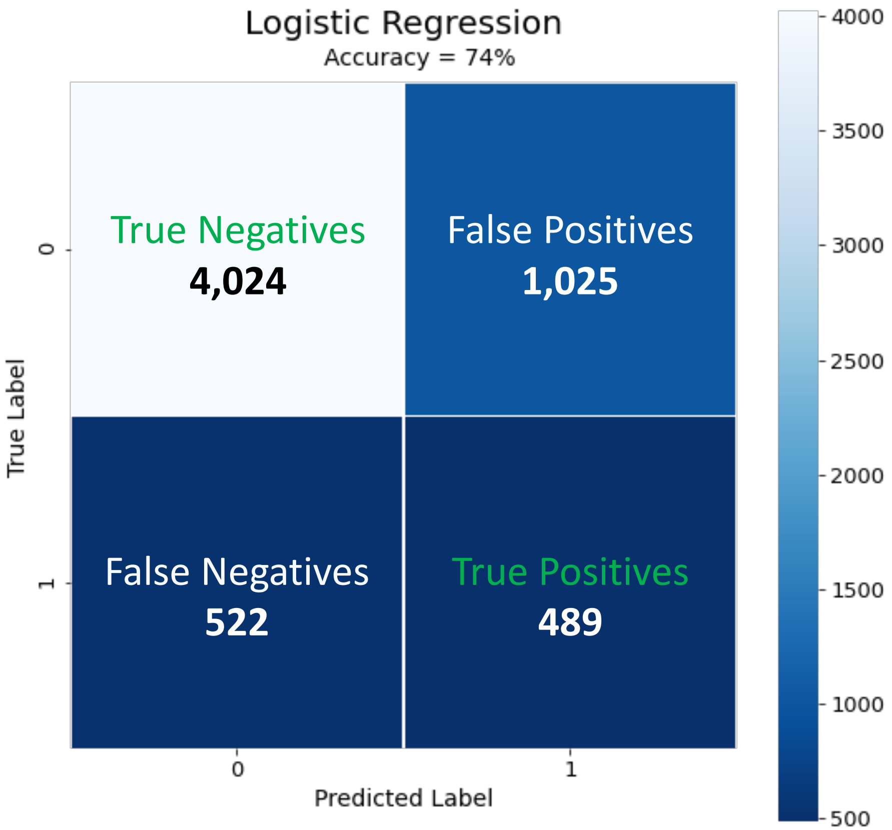
Accuracy = 74%

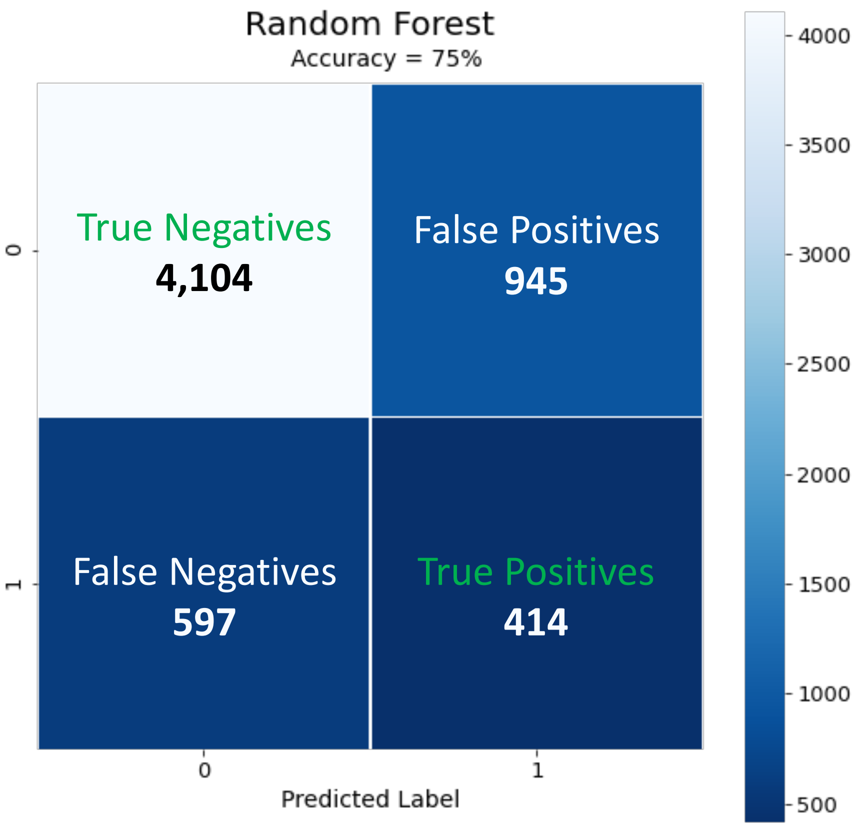
Precision = 79%

Recall = 74%

F-1 Score = 76%

Logistic Regression *predicts the probability that a sample belongs to a certain class using the sigmoid function.*





Random Forest

Accuracy = 75%

Precision = 78%

Recall = 75%

F-1 Score = 76%

Random Forest *builds many decision trees using random subsets of features and outputs the classification predicted by most trees.*

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN} \qquad \text{F-1 Score} = 2 \times \frac{\text{Precision x Recall}}{\text{Precision + Recall}}$$

CIVIC DIGITAL FELLOWSHIP

XGBoost

Optimized Machine Learning Model

1 = Priority Score < 30

0 = Priority Score ≥ 30 / ND

Accuracy

78%

Precision

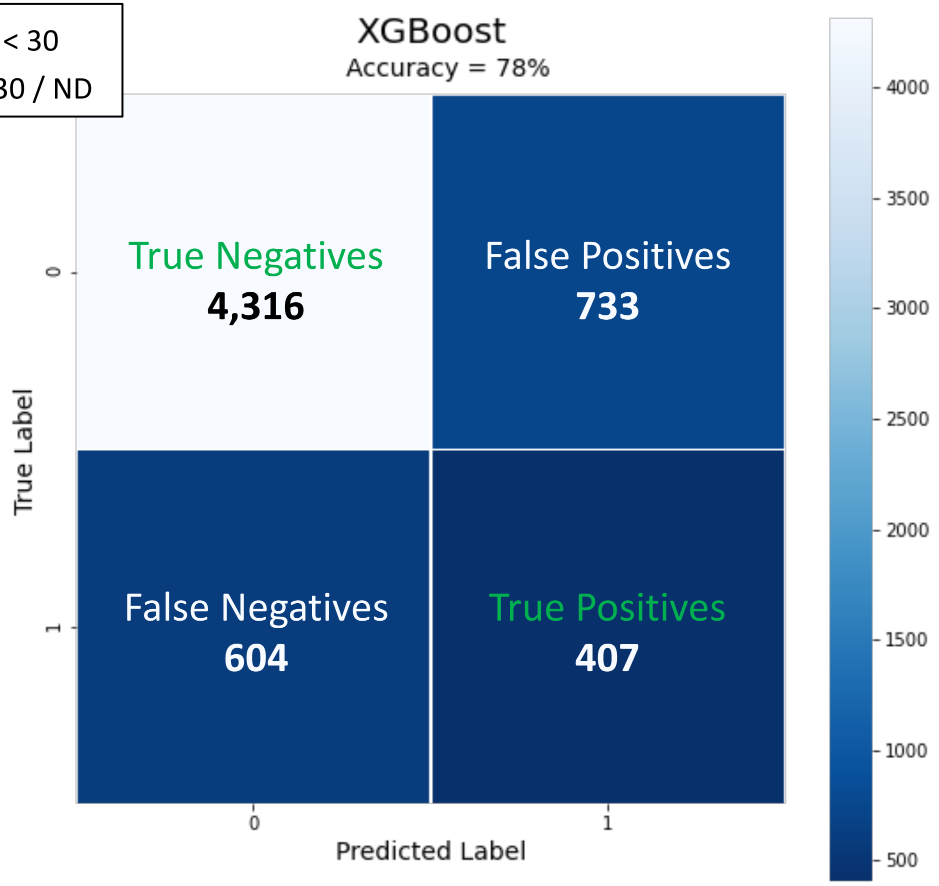
79%

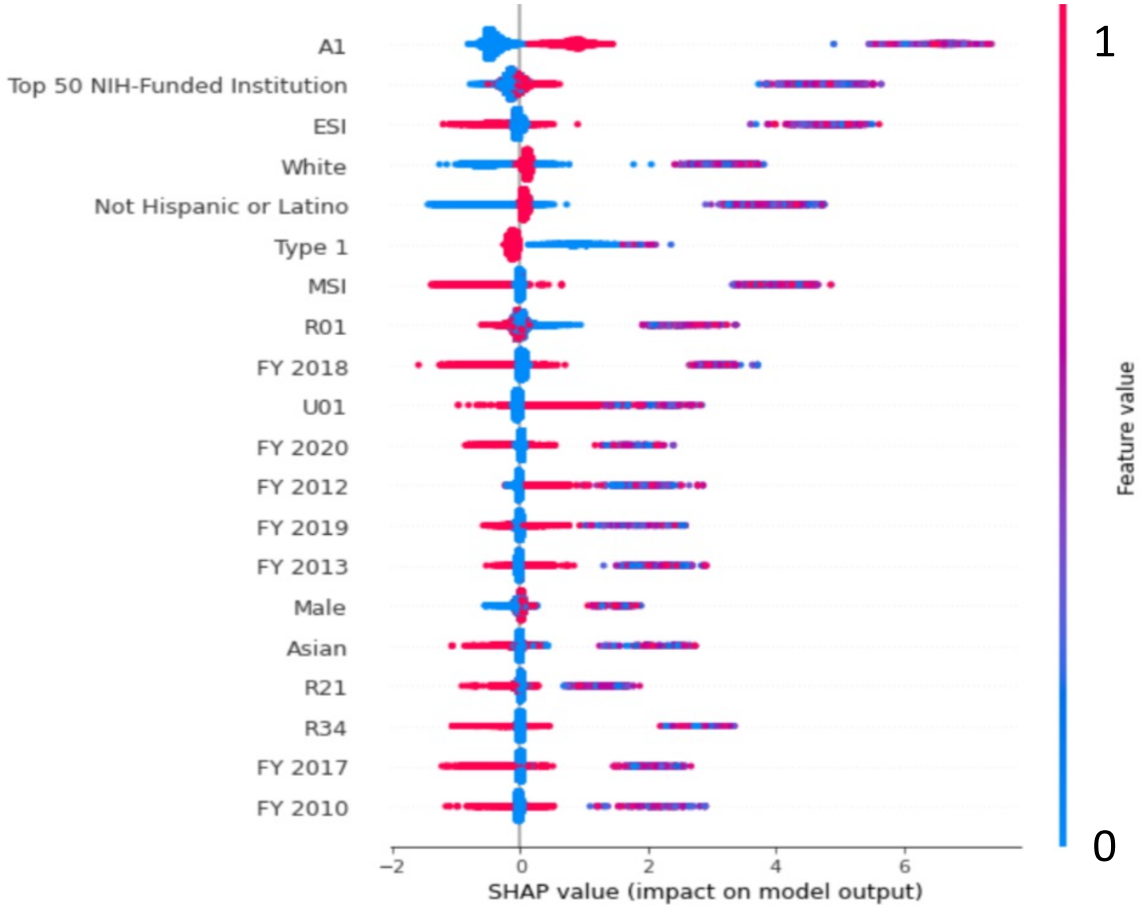
Recall

78%

F-1 Score

78%





XGBoost (Extreme Gradient Boosting) *uses gradient boosted decisions trees to achieve high speed and performance.*

The ***SHAP (Shapley Additive exPlanations)*** *value explains how much each feature contributes –* ***positively*** *or* ***negatively*** *– to the model's predictions.*

Conclusions

1. **Gender disparities** in NIDA grant awards are **almost eliminated**, with increasing applications from women and gender equity in funding rates playing a crucial role.

2. Over 11 years, **little increase in racial and ethnic diversity** in NIDA grant applications and awards.

3. **No significant disparities in funding rates** between Black and White applicants and disparities in funding rates between Latinx and non-Latinx applicants are decreasing.

4. There are proportionally more **Early-Stage Investigators (ESIs)** and no significant disparities in ESI funding rates among Black and Latinx applicants compared to White and non-Latinx applicants, respectively.

5. **Resubmission, institution, ESI, race, ethnicity, activity code,** and **gender** strongly **influence** whether a NIDA RPG application receives a **Priority Score less than 30**.

CIVIC DIGITAL FELLOWSHIP

What Can We Do?

1. Implement racial and ethnic equity **initiatives to increase applications** from underrepresented researchers, especially Black and Latinx researchers.

2. Increase **outreach** about NIDA funding opportunities and ORTDD programs to **Minority-Serving Institutions (MSIs), especially HBCUs.**

3. **Measure and disseminate ORTDD program impact** to increase applications and representation of BIPOC in substance use and addiction science.

4. **Further analyze** the causes of low applications and awards among underrepresented researchers, including potential biases and barriers within review and institutions.

Project Challenges, Achievements & Next Steps

Challenges

- Anonymized data vs. unique identifiers
- Class imbalance (80% / 20%)
- Duplicates
- Low N values

Achievements

- Presented to the Director of NIDA
- Preparing to publish my project's findings
- Invited to continue interning with ORTDD part-time

Next Steps

- Odds ratio analysis
- NLP on abstracts to identify textual predictors of Priority Score < 30
- Track resubmission rates
- Analyze causes of low applications from underrepresented researchers

Thank You!

Questions?

NIDA ORTDD Team

Lindsey Friend

Albert Avila

Wilson Compton

Angela Holmes

Julie Huffman

Isabela Lopes

Ernestine Lenteu

NIH ODSS Team

Jaqueline Cattell

Allissa Dillman

Erin Walker

Coding it Forward Team

Rachel Dodell

Ariana Soto

DJ Jain

Johncarlo Cerna

Sarah MacHarg

Regine De Guzman

Civic Digital Fellowship Mentor

Sherry Shenker

CIVIC DIGITAL FELLOWSHIP