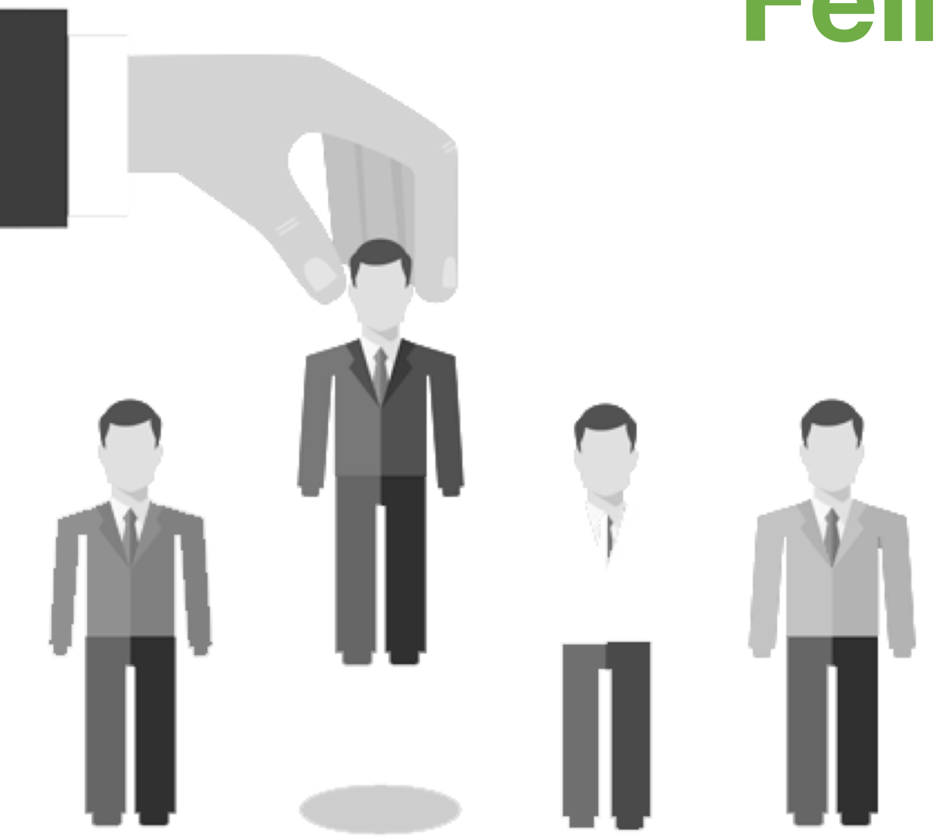
2018 Wharton **People Analytics** Conference

Biases and (Dis)agreement in Fellowship Selection Process

Insights & Strategies



Park Sinchaisri, Wharton Operations Research
Titipat Achakulvisut, Penn Neuroscience/Bioengineering

Review processes are prone to *biases*

Domains:
Employment interviews/Peer reviews in academia

Review processes are prone to *biases*

Domains:
Employment interviews/Peer reviews in academia

Existing biases of applicant's characteristics
Race, ethnic names, accents, appearances
Authors from further away in networks

Review processes are prone to *biases*

Domains:
Employment interviews/Peer reviews in academia

Existing biases of applicant's characteristics
Race, ethnic names, accents, appearances
Authors from further away in networks

Reviewer's demographics	Nature of application	Multiple evaluations/ rankings

?

Research Questions

How do applicants'/reviewers' demographics and position's characteristics affect the evaluation?	What may influence (dis)agreements among human reviewers? Can ML help?

Research Questions

How do **applicants'/reviewers' demographics** and **position's characteristics** affect the evaluation?	What may influence **(dis)agreements** among human reviewers? Can ML help?

Agenda

Data/Review Process	Empirical Methodology	Findings	Proposed Strategies
		female with exp. *citizenship bias* *reviewer skill/happiness*	*normalized scores* *optimal assignment* *machine learning*

GLOBAL HEALTH CORPS

Fellowship Review Process

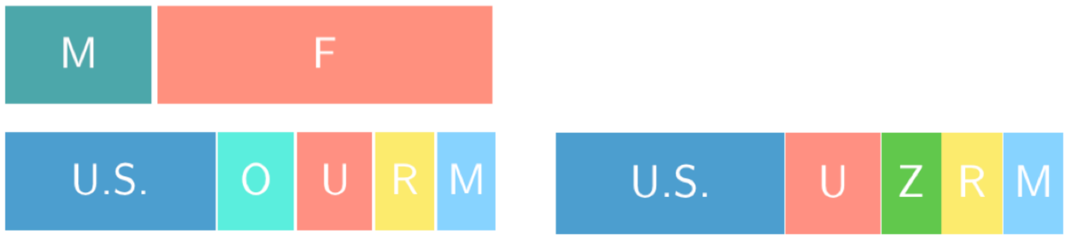
1	2	3	4	5
Reading	Reading	Reading	Interview	Interview
	2-6 reviewers	1 reviewer		

Fellowship Review Process

1	2	3	4	5
Reading	Reading	Reading	Interview	Interview
	2-6 reviewers	1 reviewer		

5778 applicants

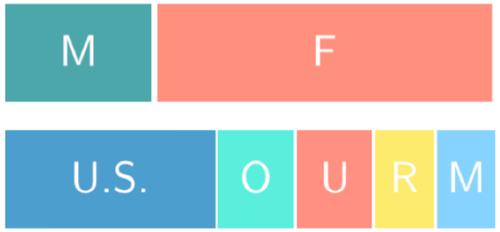
139 positions



Fellowship Review Process

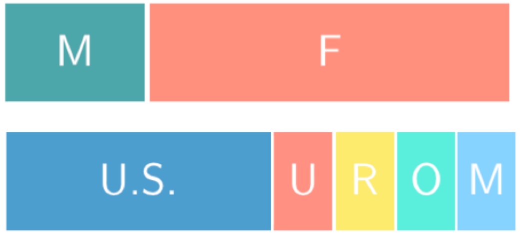
1	2	3	4	5
Reading	Reading	Reading	Interview	Interview
	2-6 reviewers	1 reviewer		

5778 applicants



139 positions

244 reviewers



1204 semifinalists

Data Pre-Processing

Text Preprocessing

Features generation

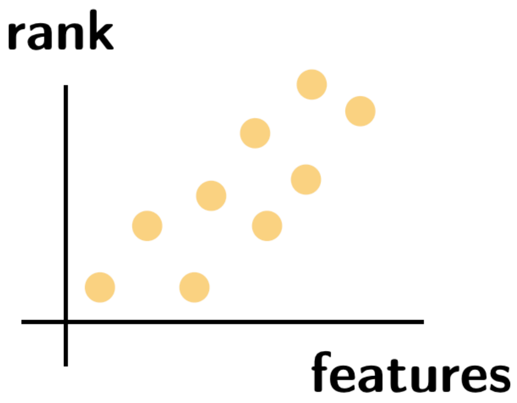
R1	R2
25	30
20	25
20	27
22	25
10	20

>

R1	R2
1.0	1.0
0.67	0.5
0.67	0.7
0.8	0.5
0.0	0.0

Normalized Score
within reviewer

$$s_{\text{norm}} = \frac{s_i - s_{min}}{s_{max} - s_{min}}$$



OLS model

Negative Binomial model

Beta model

Probit/Logit model

Roles of **Applicants' Characteristics**

	Whites	Blacks	Hispanics	Asians
	25.97%	*50.20%*	*3.46%*	*10.81%*
% selected	60.31%	51.27%	56.58%	54.79%

Roles of **Applicants' Characteristics**

	Whites	Blacks	Hispanics	Asians
	25.97%	*50.20%*	*3.46%*	*10.81%*
% selected	60.31%	51.27%	56.58%	54.79%
accept rate corrected for competition	**39.15%**	**53.94%**	**36.61%**	**25.04%**

Race of applicants do not significantly affect their scores

Roles of **Applicants' Characteristics**

	25.97%	*50.20%*	*3.46%*	*10.81%*
	Whites	Blacks	Hispanics	Asians
% selected	60.31%	51.27%	56.58%	54.79%
accept rate corrected for competition	**39.15%**	**53.94%**	**36.61%**	**25.04%**

Race of applicants do not significantly affect their scores

More favorable

Female, eligible citizenship, work experience in public health, previously applied

Roles of **Reviewer's** Characteristics

- Citizenship
- Gender
- Skillset
- Happiness

Fixed effects
regression models

Roles of **Reviewer's** Characteristics

- Citizenship
- Gender
- Skillset
- Happiness

Fixed effects regression models

Citizenship

Reviewer's Applicant's

62.7% matched

Score: +3.5%

Rank applicants of the same citizenship higher

Citizenship Bias

Citizenship

Roles of **Reviewer's** Characteristics

- Citizenship
- Gender
- Skillset
- Happiness

Fixed effects regression models

Reviewer's Applicant's

62.7% matched
Score: +3.5%

Rank applicants of the same citizenship higher

Citizenship Bias

Reviewer's Position's Country

54.6% matched
Rank: -1.5%

Harsher in ranking applicants + selecting semifinalist when reviewing for home

Roles of **Reviewer**'s

Gender

Reviewer's

26.9% male
Score: -7%

Male reviewers assign lower scores but select more semifinalists

Roles of **Reviewer's**

Gender

Reviewer's

26.9% male
Score: -7%

Male reviewers assign lower scores but select more semifinalists

Skillsets

Reviewer's

Position's Requirement

55% matched
Chance: -11%

Skilled reviewers are stricter

Roles of **Reviewer's**

Gender

Reviewer's

26.9% male
Score: -7%

Male reviewers assign lower scores but select more semifinalists

Skillsets

Reviewer's

Position's Requirement

55% matched
Chance: -11%

Skilled reviewers are stricter

Happiness

Requested

Position's Country

11 disappointed
SD: +5.3%

Disappointed reviewers tend to be less consistent/certain

(Dis)agreement among Reviewers

Metrics: mean + |diff| of ranks/scores, # overlap semifinalists, Spearman's rank correlation

Tools: t and Wilcoxon rank sum tests to compare distributions, regressions of metrics

R1 vs R2

Gender

Citizenship

Placement

Skillset

Status

(Dis)agreement among Reviewers

Metrics: mean + |diff| of ranks/scores, # overlap semifinalists, Spearman's rank correlation

Tools: t and Wilcoxon rank sum tests to compare distributions, regressions of metrics

R1 vs R2

Gender	Citizenship	Placement	Skillset	Status

Same...

Larger rank correlation

Smaller score differences

Agree on same semifinalists

(Dis)agreement among Reviewers

Metrics: mean + |diff| of ranks/scores, # overlap semifinalists, Spearman's rank correlation

Tools: t and Wilcoxon rank sum tests to compare distributions, regressions of metrics

R1 vs R2

Gender	Citizenship	Placement	Skillset	Status

Same... Larger rank correlation; Smaller score differences

Gender, Citizenship: Agree on same semifinalists

Placement, Skillset, Status: No significant differences

(Dis)agreement among Reviewers

Metrics: mean + |diff| of ranks/scores, # overlap semifinalists, Spearman's rank correlation

Tools: t and Wilcoxon rank sum tests to compare distributions, regressions of metrics

R1 vs R2

Gender | Citizenship | Placement | Skillset | Status

Same...

Larger rank correlation

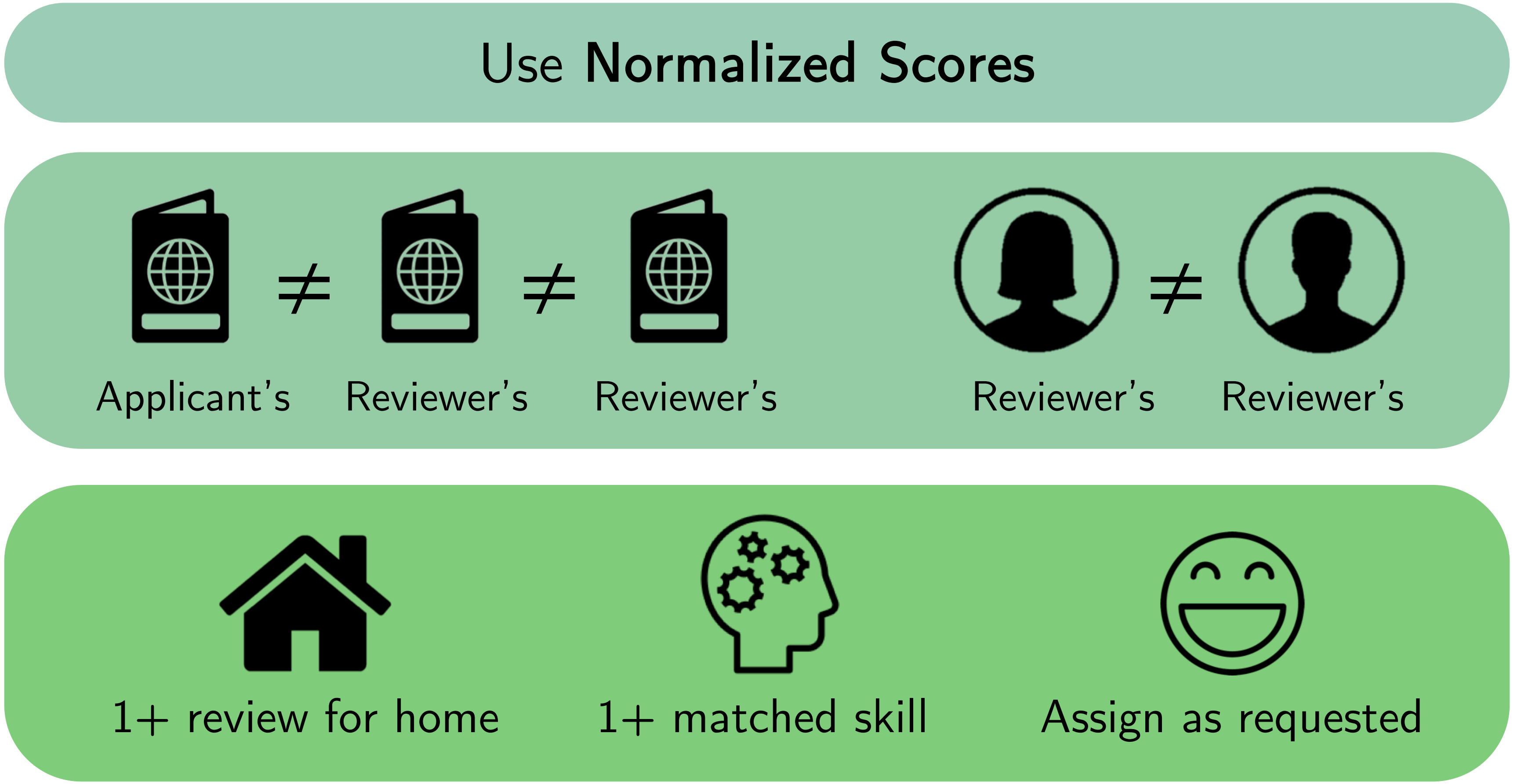
Smaller score differences

Agree on same semifinalists

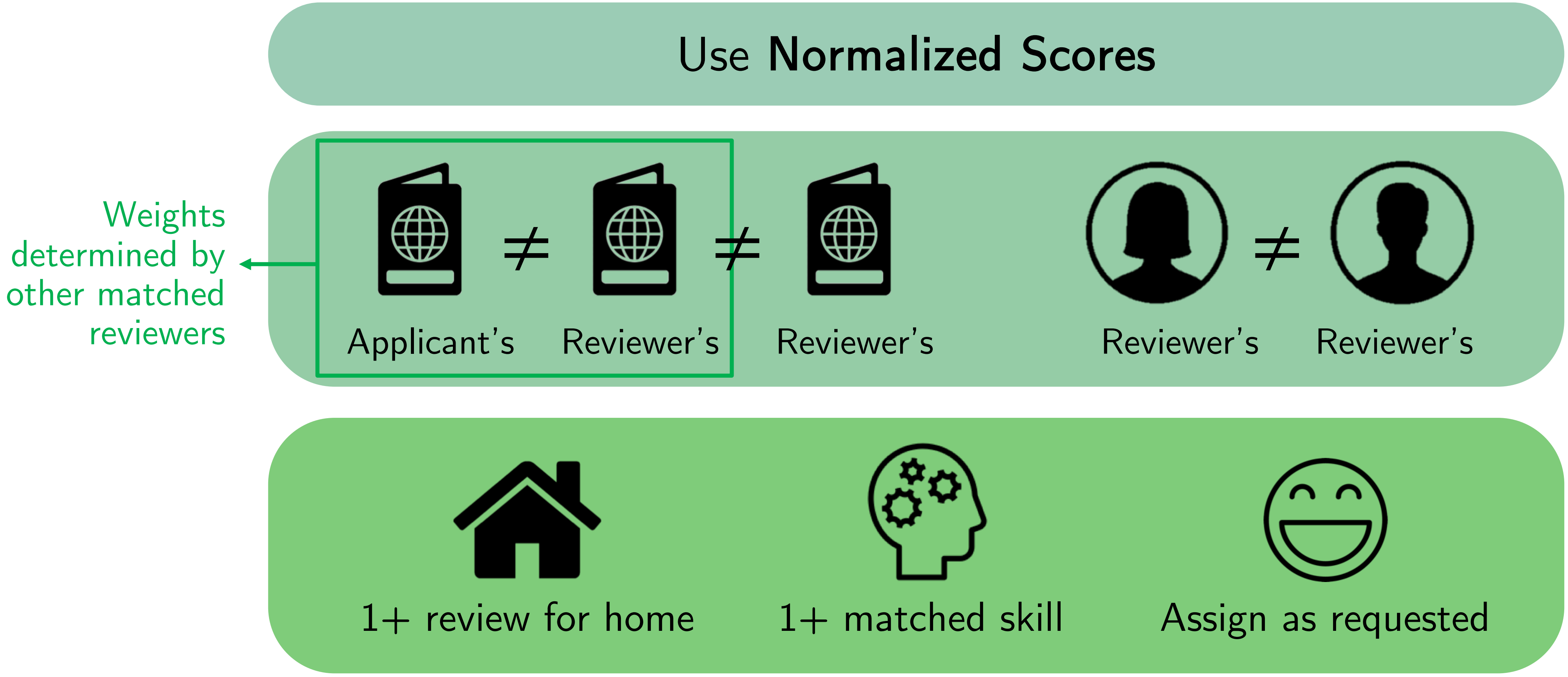
Disagree more (trend)

No significant differences

Optimal Reviewer Assignment



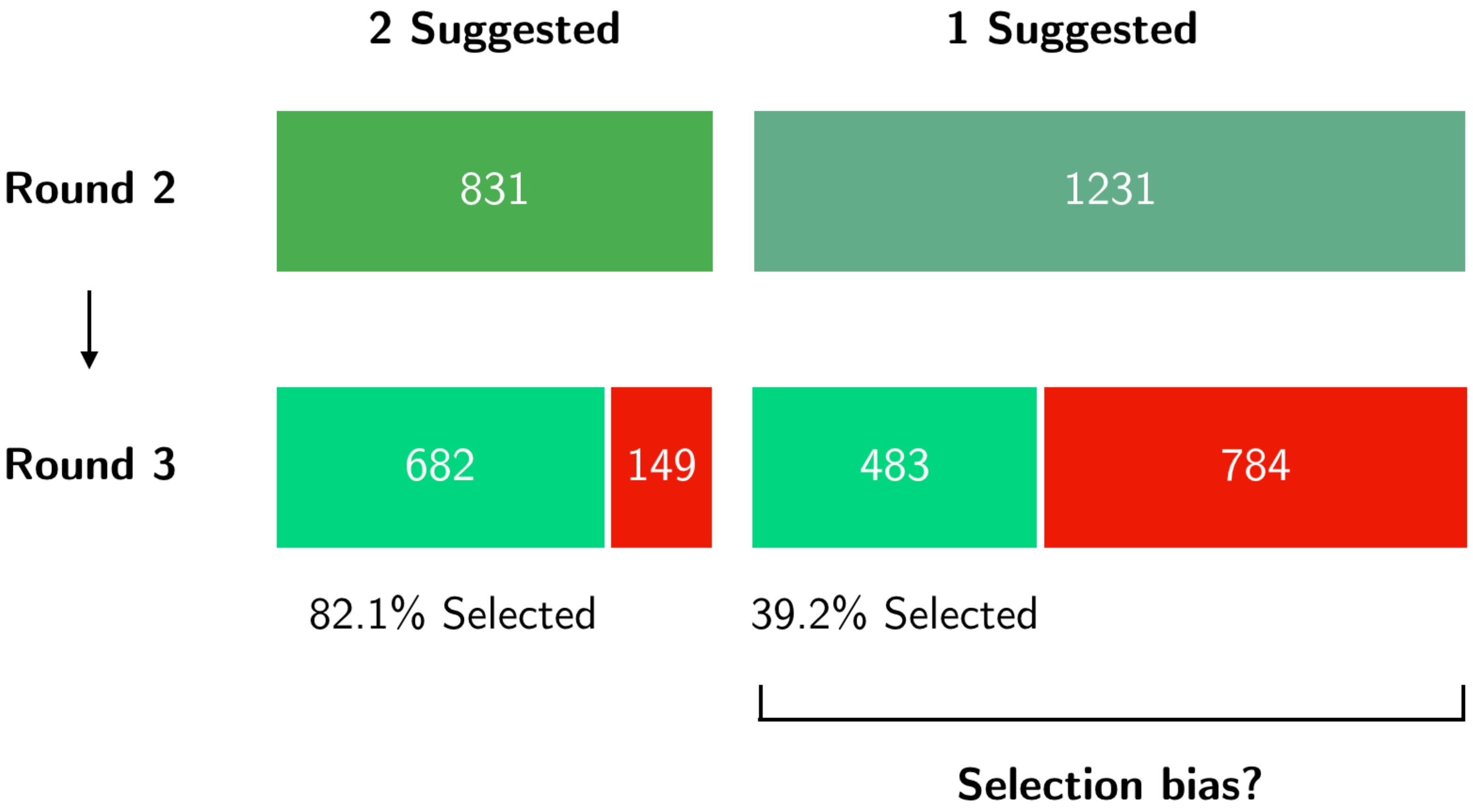
Optimal Reviewer Assignment



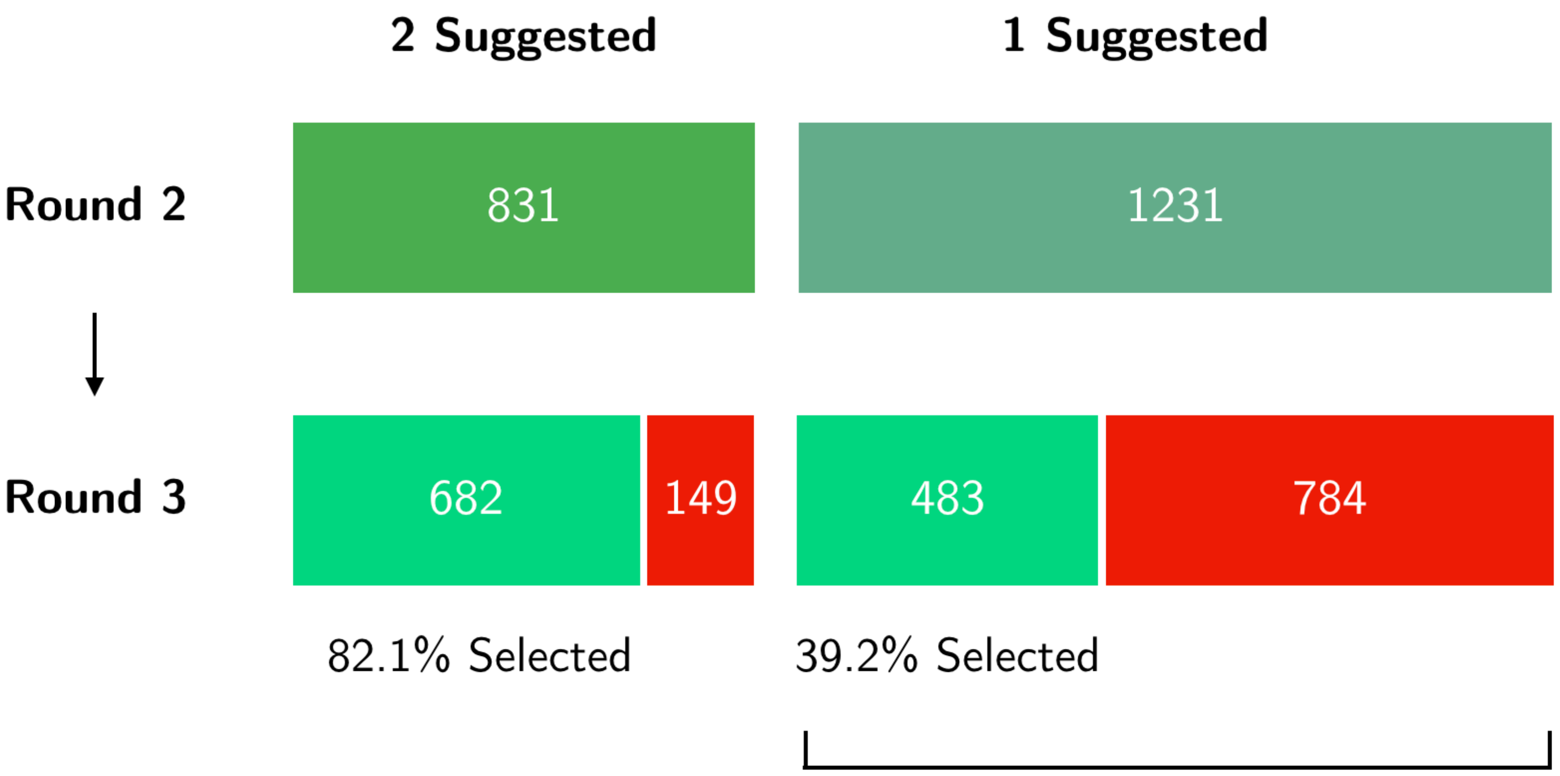
Round 3 Selection

	2 Suggested	1 Suggested
Round 2	831	1231

Round 3 Selection



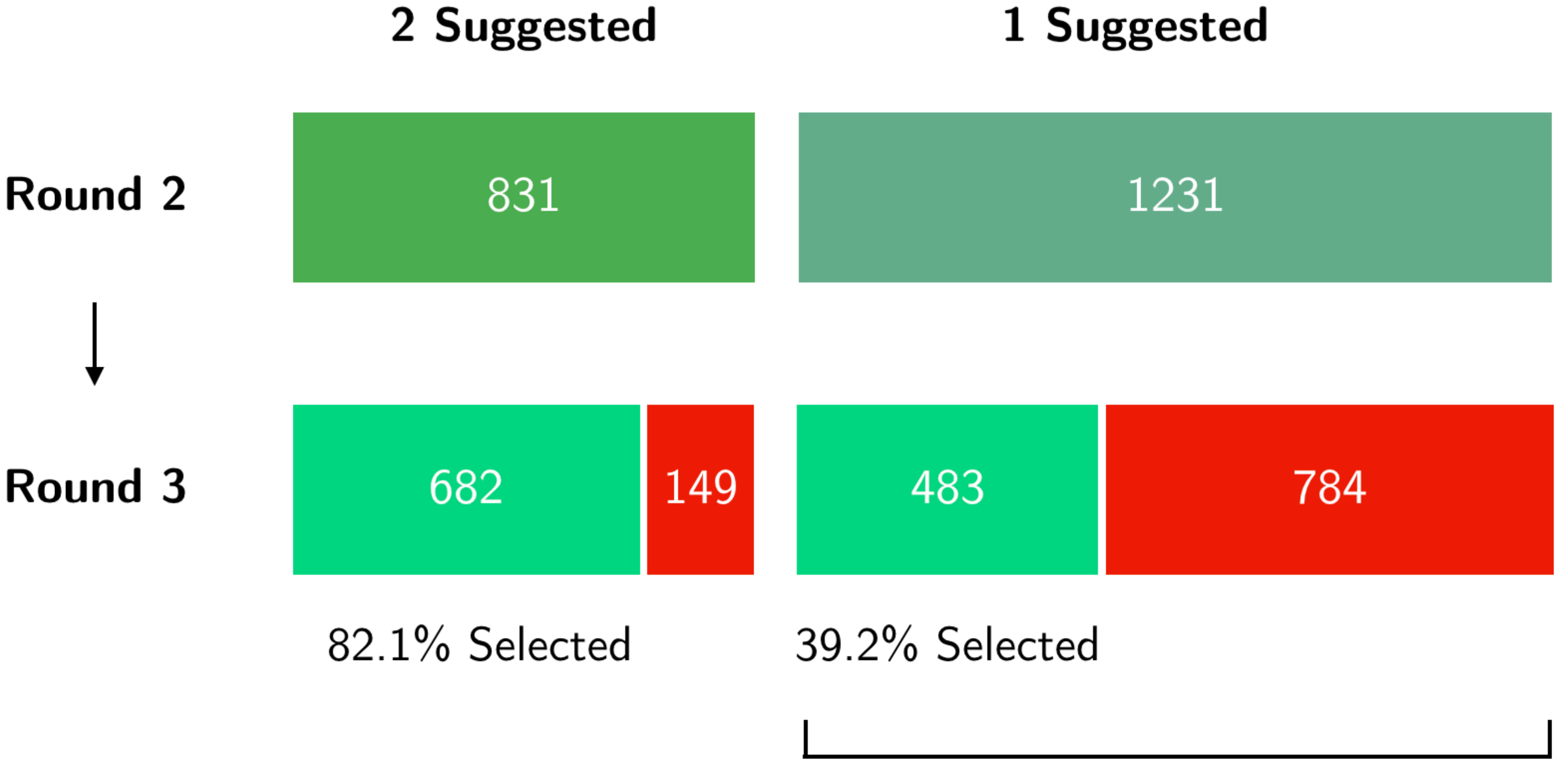
Round 3 Selection



Selection bias? No selection bias

Maximum of normalized scores predicts selection

Round 3 Selection

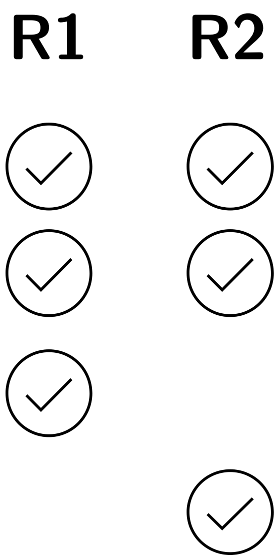


Selection bias? No selection bias

Maximum of normalized scores predicts selection

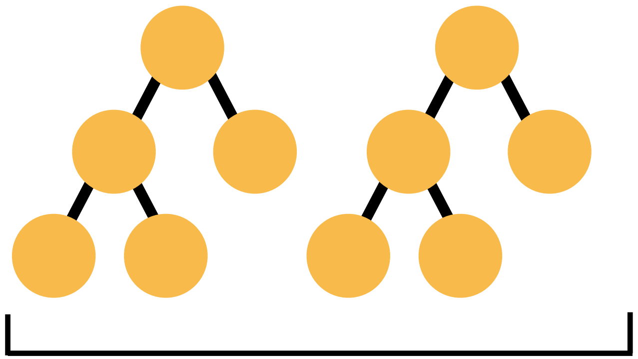
Selects applicants from two recommenders then by ranking of normalized scores

Data-Driven Selection in Round 3



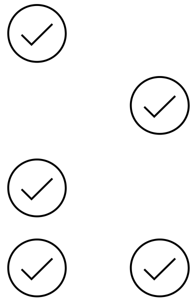
Score Ranking

2 reviewers
and normalized
score



Random Forest Ensemble

Learn selection
probability from
30% of data



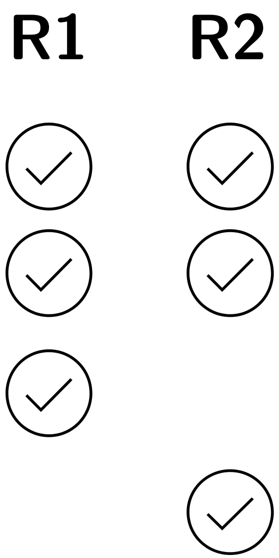
Random selection

29
27
27
25
24

Maximum Average Score

Measure overlap between ranking model and selection in round 3

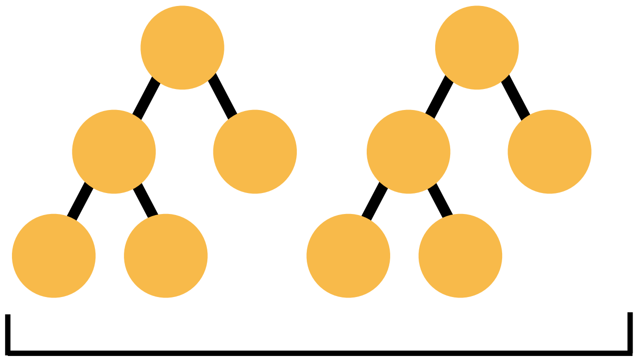
Data-Driven Selection in Round 3



Score Ranking

2 reviewers
and normalized
score

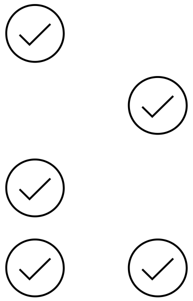
73.4%



Random Forest Ensemble

Learn selection
probability from
30% of data

77.3%



Random selection

39.7%

29
27
27
25
24

Maximum Average Score

70.3%

Discussion and Future Research

Age

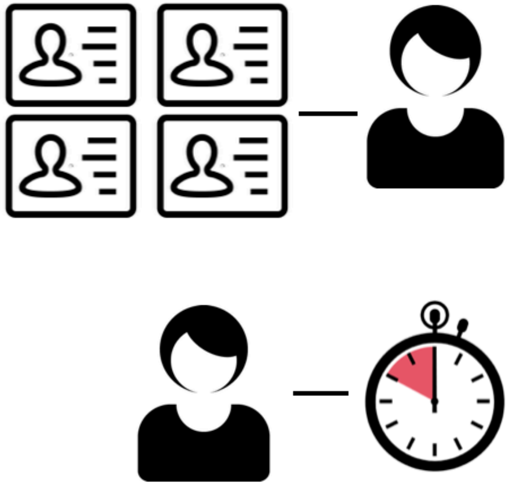
Language

$0/1 > [0,1)$

Features improvement

1	2	3	4	5

Round 3 quality checking



Review details

Conclusion

Insights

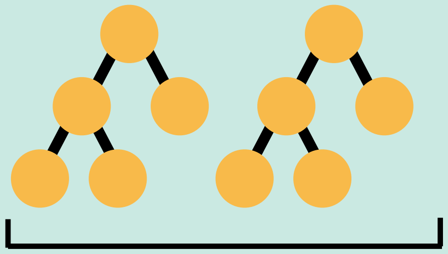
Applicants
Reviewers
Interactions

Proposed Strategies

Applicant-Reviewer

Reviewer-Reviewer

Use Normalized Score



ML for ranking
can replace
human reviewer

Round 2

Round 3