Integrating Reward Maximization and Population Estimation

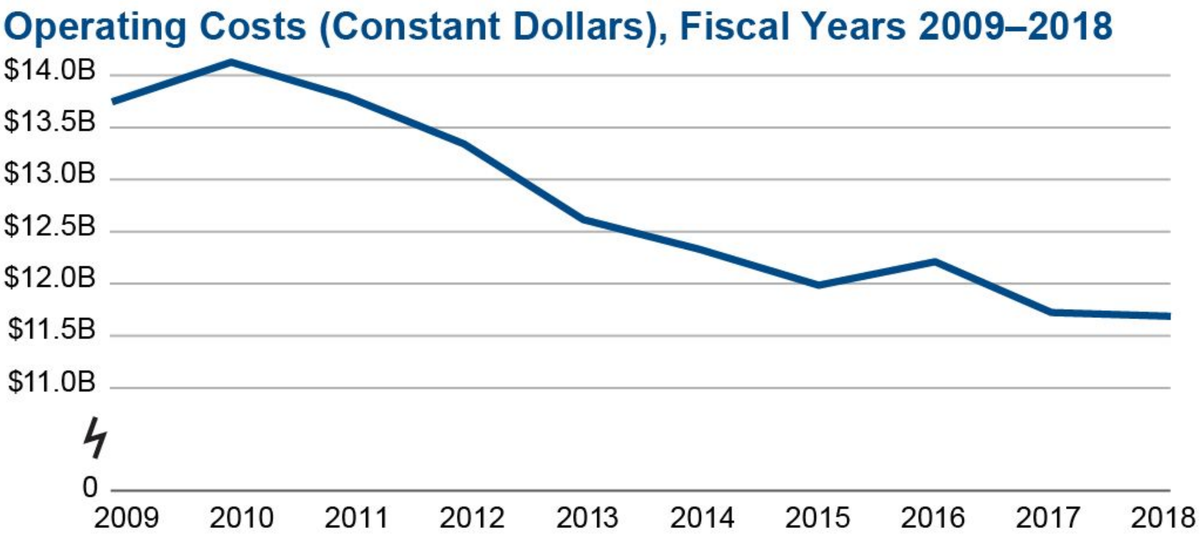
Sequential Decision-Making for Internal Revenue Service Audit Selection

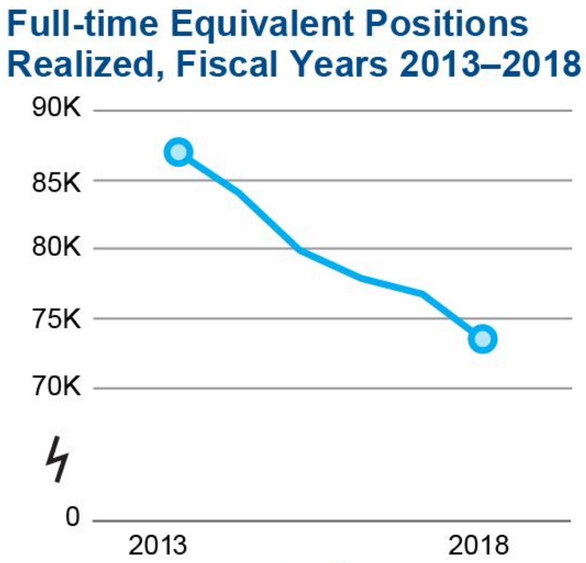
Peter Henderson[1], Ben Chugg[1], Brandon Anderson[12], Kristen Altenburger[1], Alex Turk[2], John Guyton[2], Jacob Goldin[1], Daniel E. Ho[1]

[1]Stanford University [2]IRS RAAS

All views and opinions expressed in this presentation are our own and not of any of our co-authors, nor of the Internal Revenue Service or any other company or government entity.

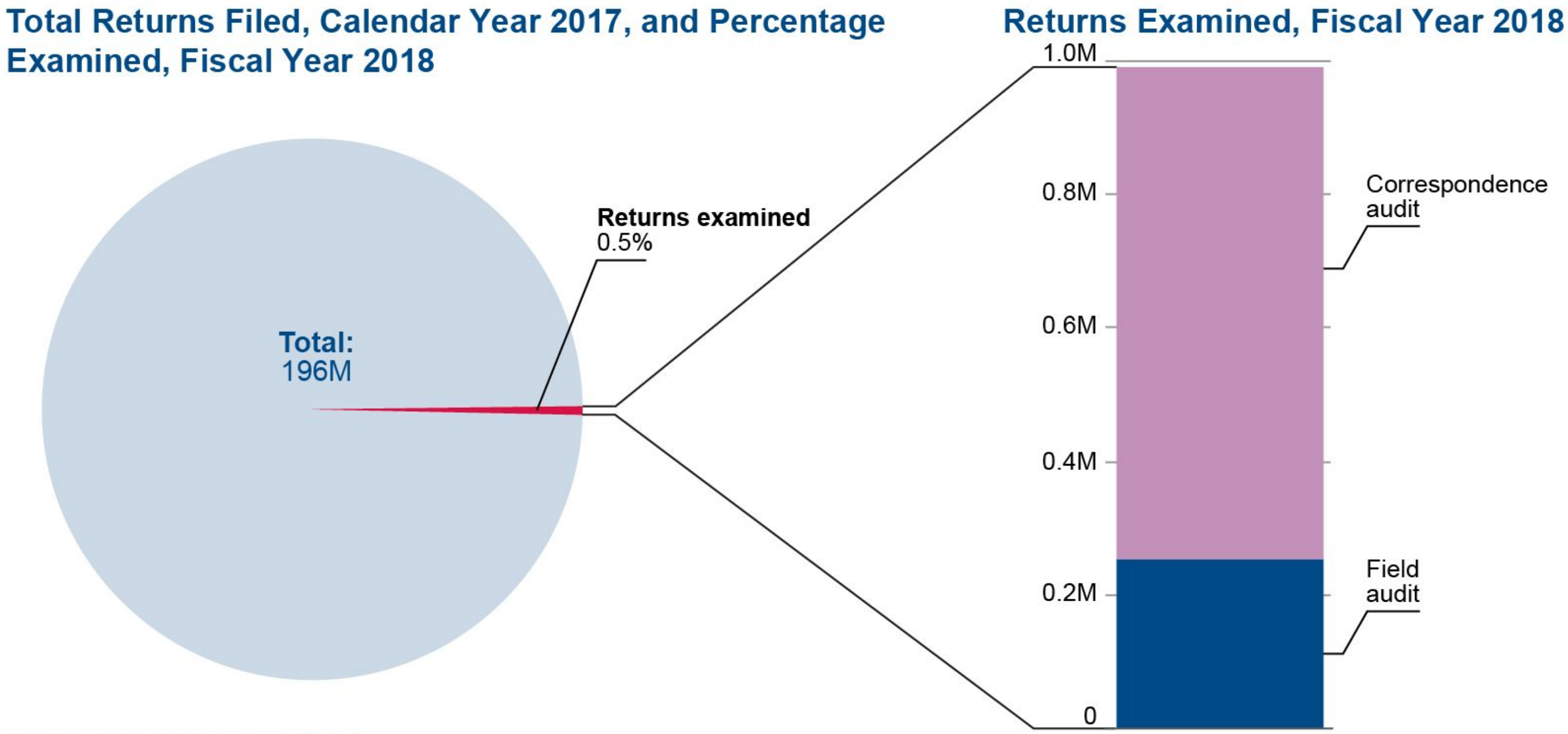
Institutional Context





Stanford

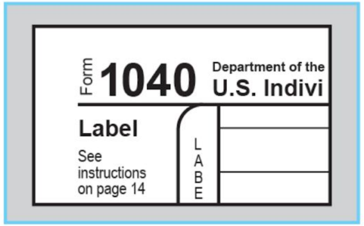
Institutional Context



SOURCE: 2018 *IRS Data Book* Table 9a

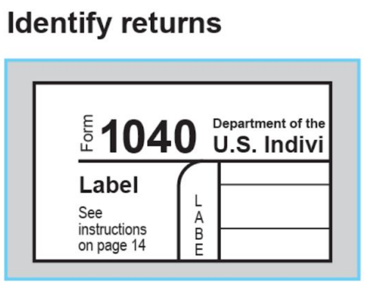
Stylized Program

Identify returns

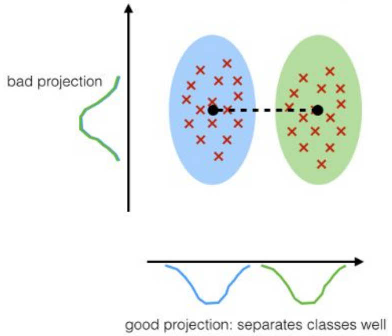

Form 1040 Department of the
U.S. Indivi
Label
See instructions on page 14
LABE

Random sample
(~15k / year, 2006-14)

Stylized Program

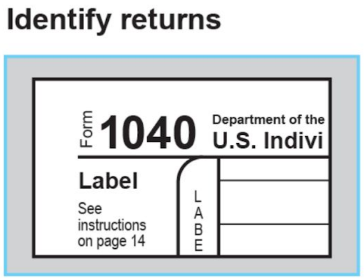


Random sample
(~15k / year, 2006-14)

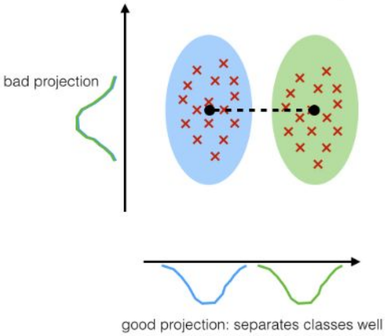


Risk model

Stylized Program



Random sample
(~15k / year, 2006-14)

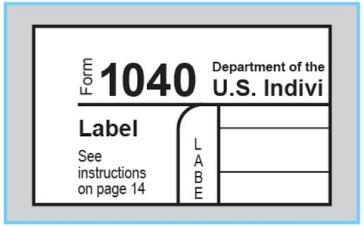


Risk model

Risk selected Op Audits
(>500k / year)

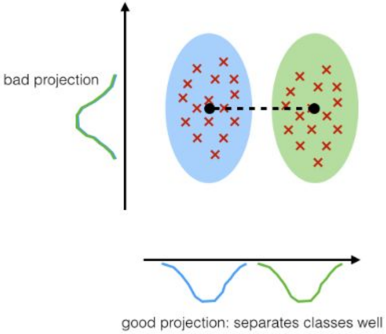
Stylized Program

Identify returns



Random sample
(~15k / year, 2006-14)

Tax Gap Estimate

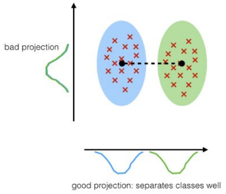


Risk model

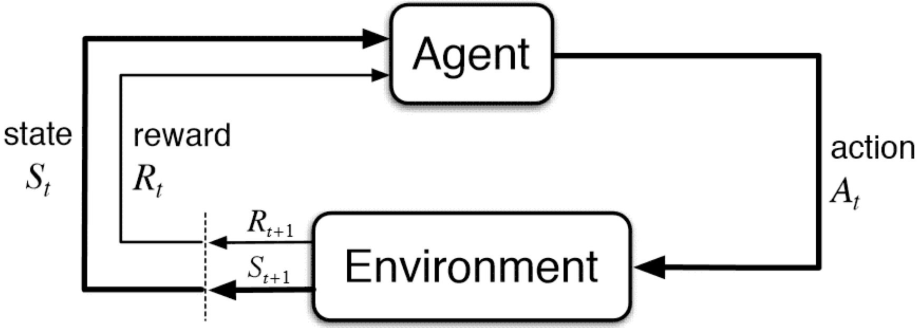
Audit returns

Risk selected Op Audits
(>500k / year)

Sequential Decision-Making (Machine Learning)

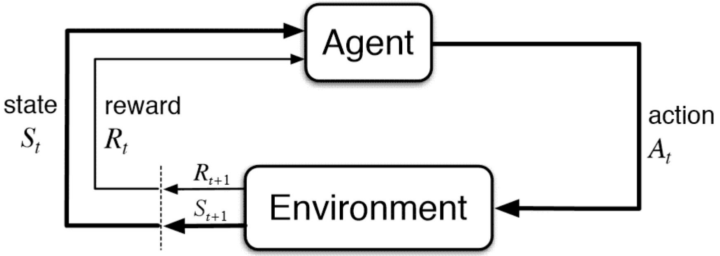


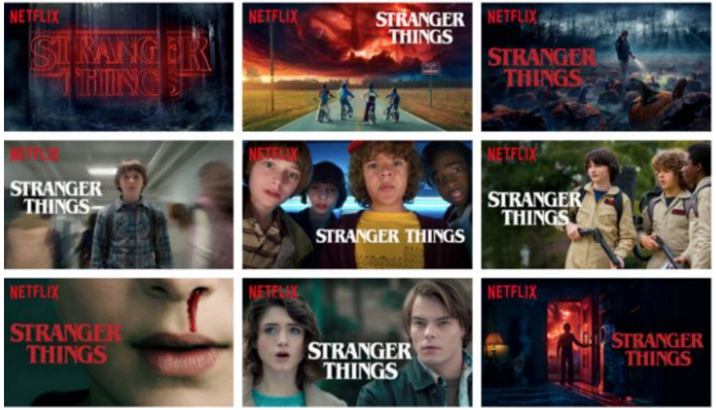
Reward-prediction Model



Sequential Decision-Making In the Real World

Example: NETFLIX



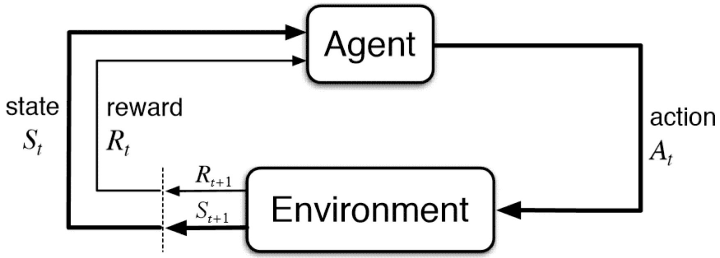


Context	User information on device (environment)
Actions	Set of movie banners to show
Reward	User engagement (click-through, minutes)
Learner	Identify policy to maximize cumulative reward

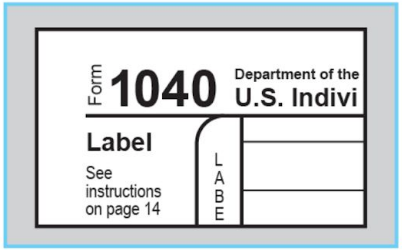
Explore new movies / preferences vs. Exploit known preferences

Sequential Decision-Making In the Real World

Example: IRS



Identify returns

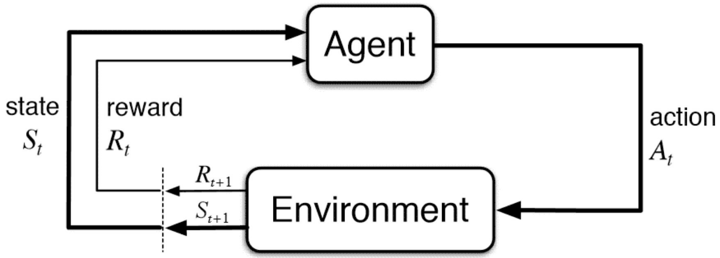


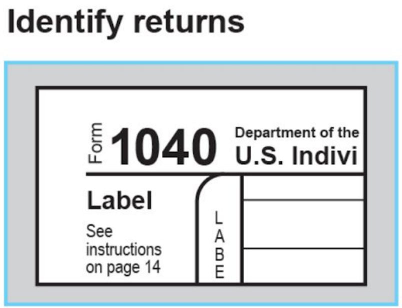
Context	Tax return information (taxpayer, stratum, etc.)
Actions	Selecting returns to audit
Reward	Detected Under-reporting
Learner	Identify policy to maximize cumulative reward

Explore forms of underreporting vs. Exploit known underreporting

Sequential Decision-Making In the Real World

Example: IRS



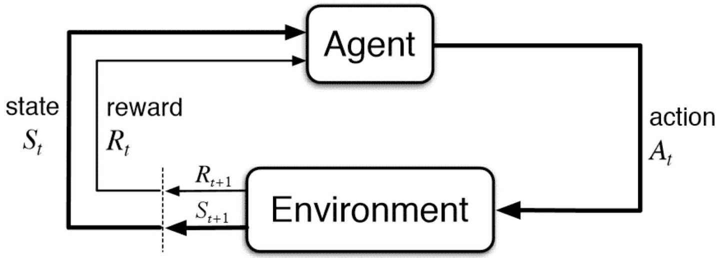


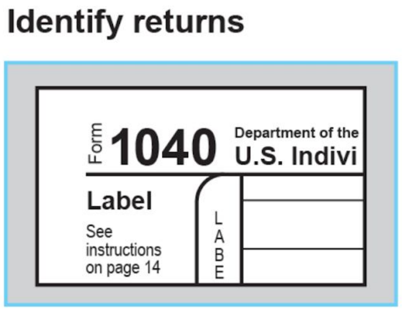
Context	Tax return information (taxpayer, stratum, etc.)
Actions	Selecting returns to audit
Reward	Under-reporting
Learner	Identify policy to maximize cumulative reward

+ Estimate unbiased population statistics (e.g., tax gap, average misreporting)

Sequential Decision-Making In the Real World

Example: IRS



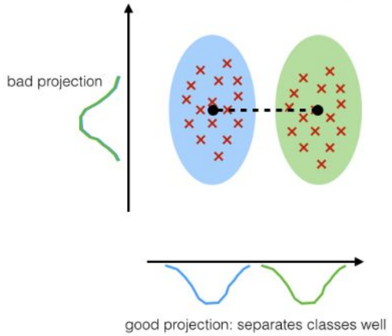


Secondary objective not typical of machine learning literature

+ Estimate **<u>unbiased</u>** population statistics (e.g., tax gap, average misreporting)

Sequential Decision-Making In the Real World

Example: IRS



Risk model

Tempting Solution: Use a regression-based risk-model to do selection and estimation, with no random sampling.

Problems: Sequentially-learned models are known to be biased and there are no theoretically guaranteed ways to *remove* this bias in the low sample regime (yet). (Nie et al., 2018)
Lack of exploration leads to suboptimal feedback loops. (Jiang et al., 2019)

Optimize-and-Estimate Structured Bandits

Machine Learning Literature on Sequential Decision-Making (e.g., bandit algorithms that optimize for reward only)

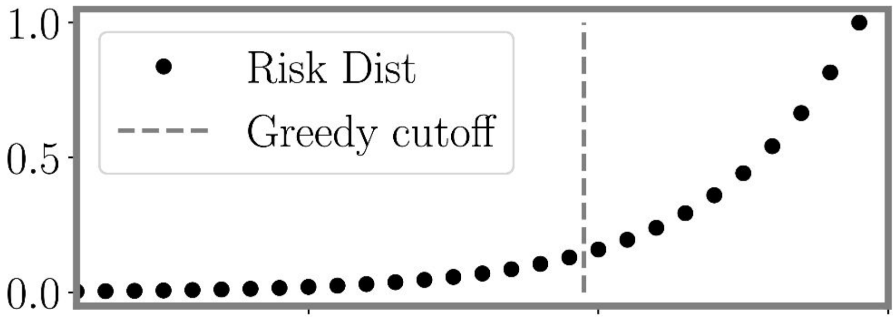
+

Sampling Literature (unbiased estimation of population statistics)

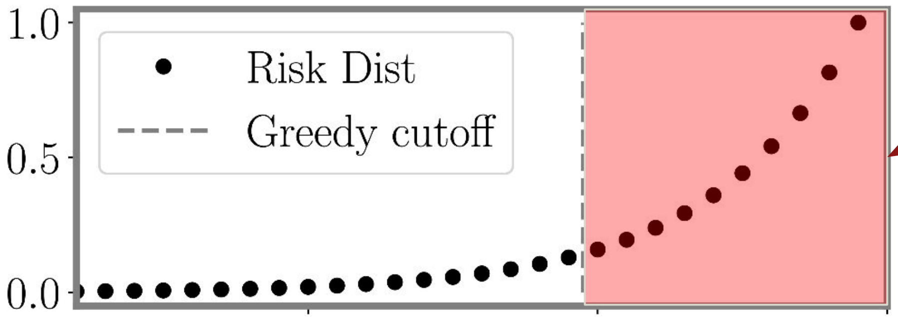
=

Optimize-and-estimate Structured Bandits

Adaptive Bin Sampling



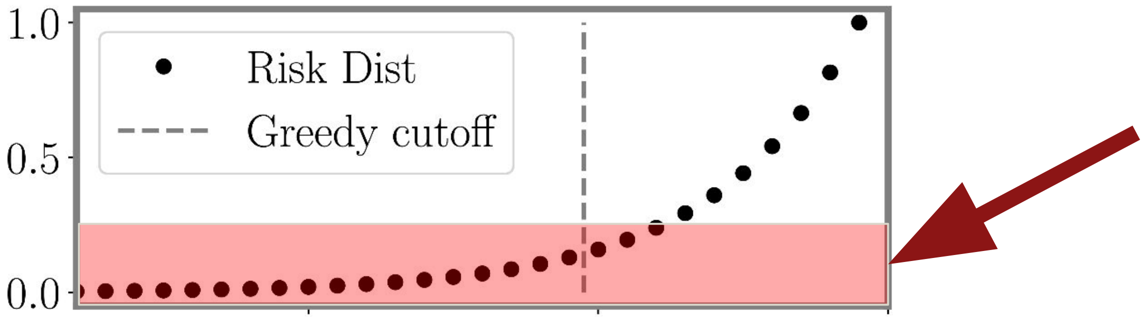
Adaptive Bin Sampling



Greedy selection
(e.g., stylized version of Op audits)

If only use this:
biased model, biased estimate

Adaptive Bin Sampling

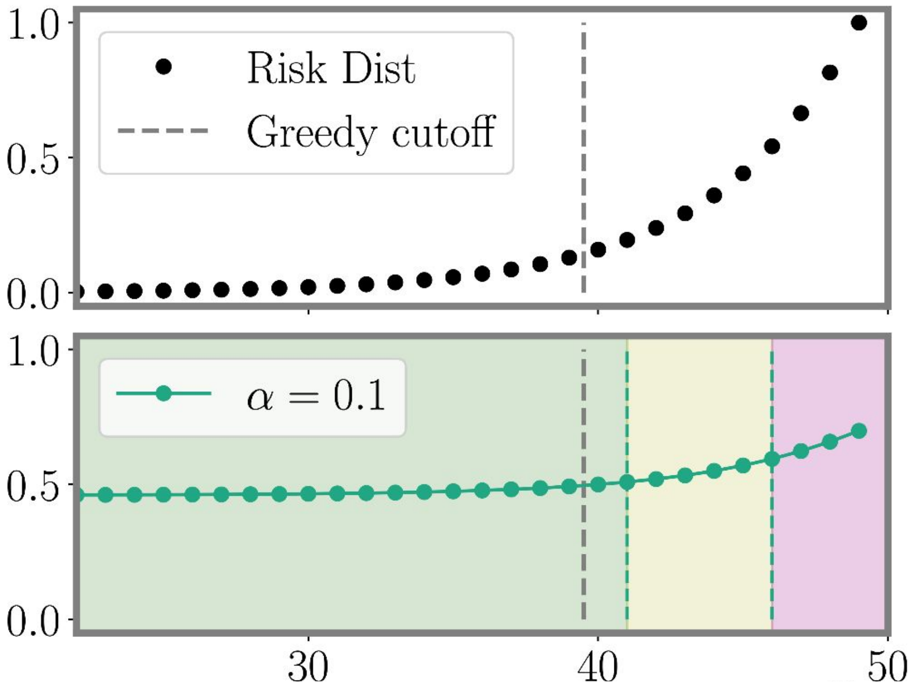


Random selection

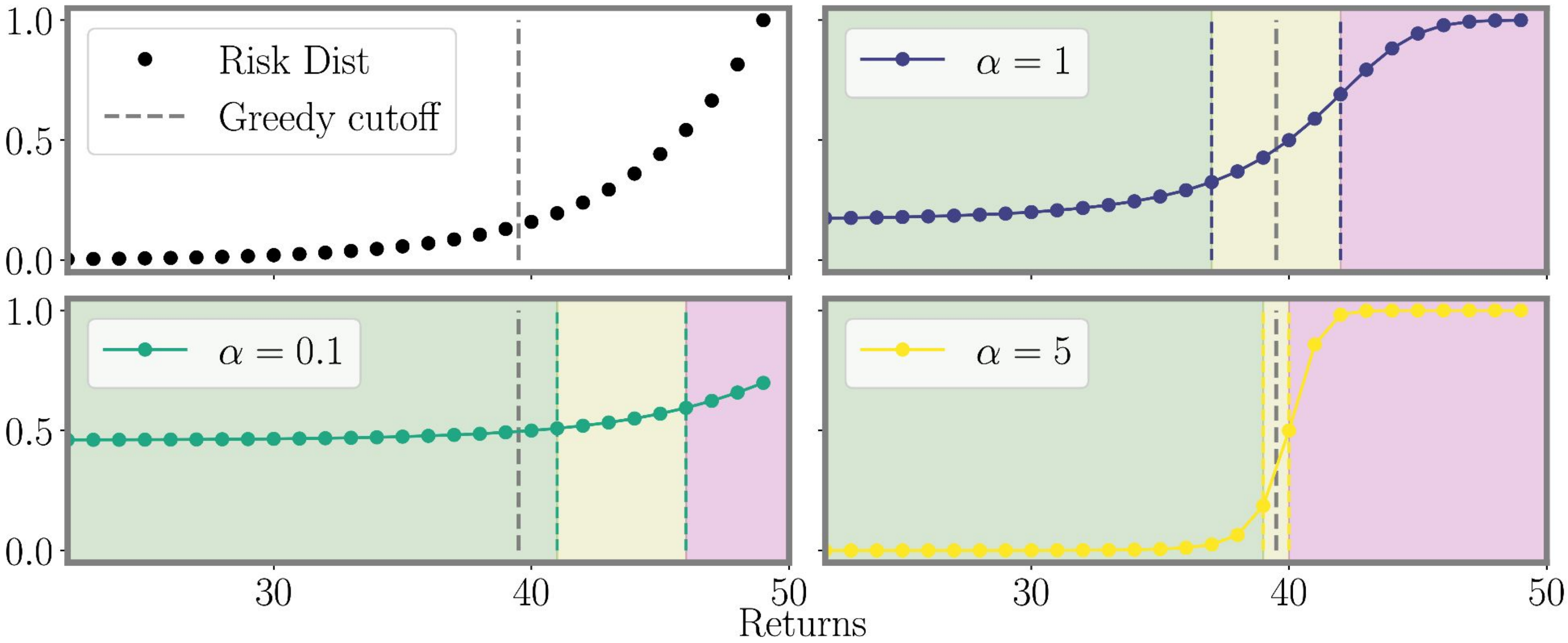
If only use this:

Unbiased estimate,
but sub-optimal and low reward

Adaptive Bin Sampling



Adaptive Bin Sampling



Adaptive Bin Sampling

Horvitz-Thompson estimator gives **unbiased** estimate.

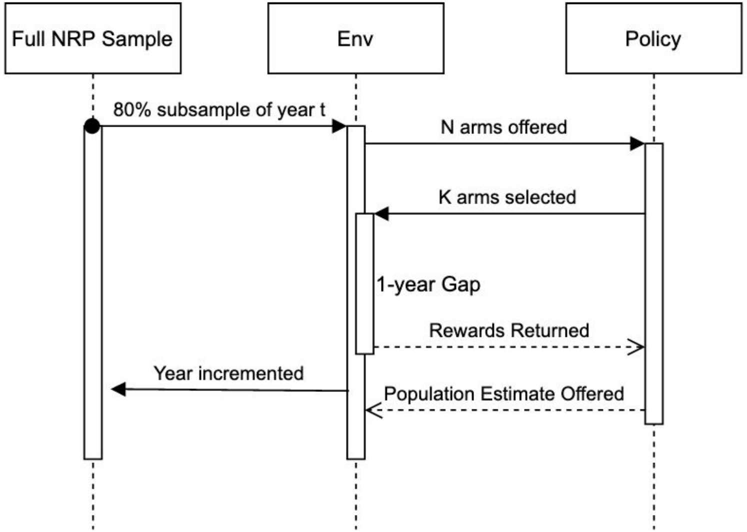
And we have **fine-grained** control over reward-variance trade-off.

$$\hat{\mu}_{HT}(t) = \frac{1}{\sum_a w_a} \sum_{a \in \mathcal{K}} \frac{w_a r_a}{p_a},$$

Experiments

For NRP data years 2006-2014

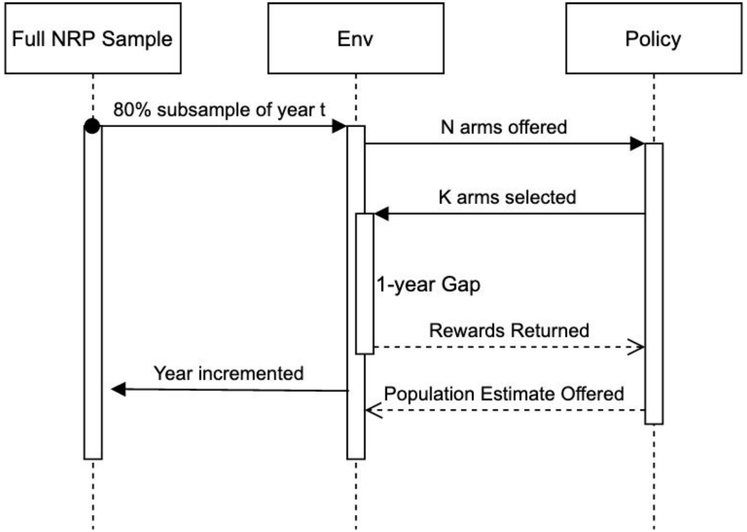
1. Take 80% subsample



Experiments

For NRP data years 2006-2014

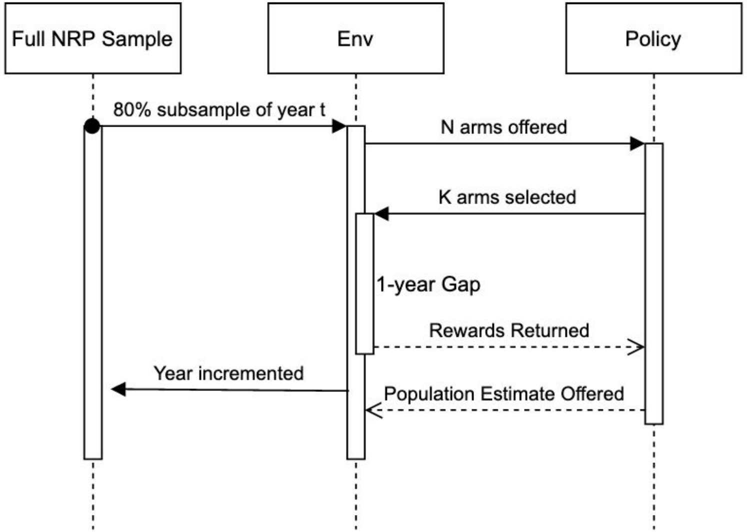
1. Take 80% subsample
2. Give selection policy ~500 covariates from tax return data for each “arm” (tax return) in the sample



Experiments

For NRP data years 2006-2014

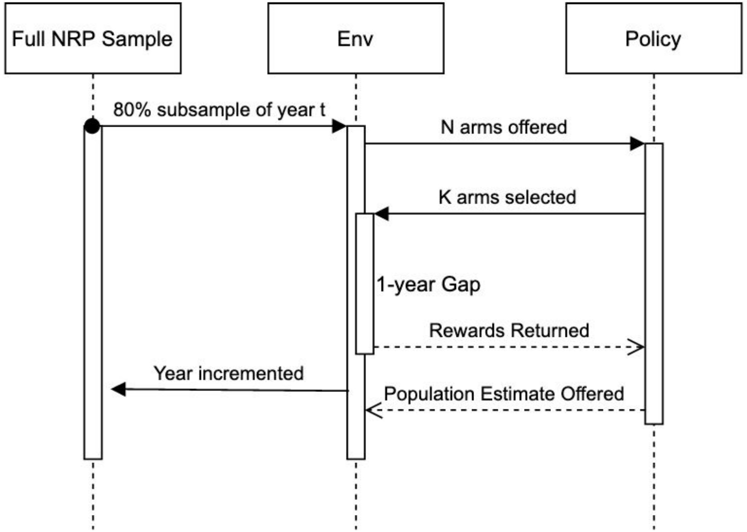
1. Take 80% subsample
2. Give selection policy ~500 covariates from tax return data for each “arm” (tax return) in the sample
3. Selection policy returns arms to audit



Experiments

For NRP data years 2006-2014

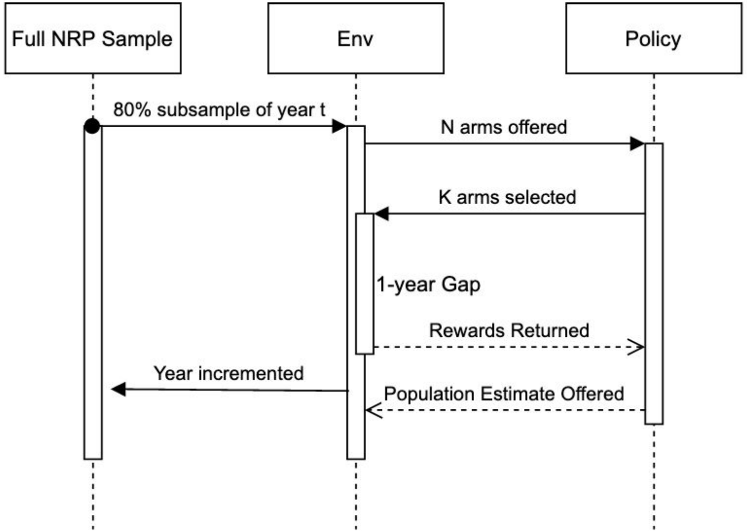
1. Take 80% subsample
2. Give selection policy ~500 covariates from tax return data for each “arm” (tax return) in the sample
3. Selection policy returns arms to audit
4. Simulate a 1 year gap



Experiments

For NRP data years 2006-2014

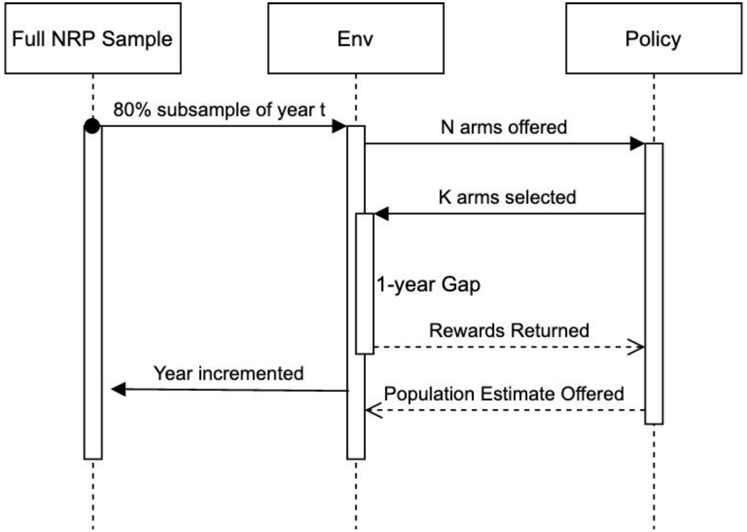
1. Take 80% subsample
2. Give selection policy ~500 covariates from tax return data for each "arm" (tax return) in the sample
3. Selection policy returns arms to audit
4. Simulate a 1 year gap
5. Return the tax adjustment (reward) after that gap



Experiments

For NRP data years 2006-2014

1. Take 80% subsample
2. Give selection policy ~500 covariates from tax return data for each “arm” (tax return) in the sample
3. Selection policy returns arms to audit
4. Simulate a 1 year gap
5. Return the tax adjustment (reward) after that gap
6. Policy makes population estimate



Experiments

Best Reward Settings

	Policy	R	μ_{PE}	σ_{PE}	μ_{NR}
Unbiased Methods	ABS-1	**$41.5M***	**0.4** ✓	31.0	**37.6%**
	ϵ-only	$41.3M*	4.3✓	37.4	38.3%
	ABS-2	$40.5M*	0.6✓	24.5	38.3%
	Random	$12.7M	1.5✓	**14.7**	53.1%

Experiments

10% (ε) random sample,
rest greedy

Best Reward Settings

	Policy	R	μ_{PE}	σ_{PE}	μ_{NR}
Unbiased Methods	ABS-1	**$41.5M***	**0.4** ✓	31.0	**37.6%**
	ϵ-only	$41.3M*	4.3✓	37.4	38.3%
	ABS-2	$40.5M*	0.6✓	24.5	38.3%
	Random	$12.7M	1.5✓	**14.7**	53.1%

Experiments

	Best Reward Settings				
	Policy	R	μ_{PE}	σ_{PE}	μ_{NR}
Unbiased Methods	ABS-1	**$41.5M***	**0.4** ✓	31.0	**37.6%**
	ϵ-only	$41.3M*	4.3✓	37.4	38.3%
	ABS-2	$40.5M*	0.6✓	24.5	38.3%
	Random	$12.7M	1.5✓	**14.7**	53.1%

Fully random sample every year,
rest greedy

Experiments

ABS can yield lower variance, similar reward, lower no-change rate, and retain unbiasedness

Best Reward Settings

	Policy	R	μ_{PE}	σ_{PE}	μ_{NR}
Unbiased Methods	ABS-1	**\$41.5M***	**0.4** ✓	31.0	**37.6%**
	ϵ-only	\$41.3M*	4.3✓	37.4	38.3%
	ABS-2	\$40.5M*	0.6✓	24.5	38.3%
	Random	\$12.7M	1.5✓	**14.7**	53.1%

Experiments

Greedy tends to perform well in highly stochastic low-sample regime (which matches our experimental setup). (Bastani et al., 2022 proved this recently.)

Best Reward Settings

	Policy	R	μ_{PE}	σ_{PE}	μ_{NR}
Biased Methods	Greedy	**$43.6M***	16.4 ✗	8.8	**36.5%**
	UCB-1	$42.4M*	15.3 ✗	9.4	38.6%
	ϵ-Greedy	$41.3M*	**6.1** ✗	**7.5**	38.3%
	UCB-2	$40.7M*	15.6 ✗	10.21	40.7%

Experiments

Best Reward Settings

	Policy	R	μ_{PE}	σ_{PE}	μ_{NR}
Biased Methods	Greedy	**$43.6M***	16.4 ✗	8.8	**36.5%**
	UCB-1	$42.4M*	15.3 ✗	9.4	38.6%
	ϵ-Greedy	$41.3M*	**6.1** ✗	**7.5**	38.3%
	UCB-2	$40.7M*	15.6 ✗	10.21	40.7%

Use regression model for both selection and population estimate. Means biased prediction, but slightly more reward and lower variance

Experiments

Best Reward Settings

	Policy	R	μ_{PE}	σ_{PE}	μ_{NR}
Biased Methods	Greedy	**$43.6M***	16.4 ✗	8.8	**36.5%**
	UCB-1	$42.4M*	15.3 ✗	9.4	38.6%
	ϵ-Greedy	$41.3M*	**6.1** ✗	**7.5**	38.3%
	UCB-2	$40.7M*	15.6 ✗	10.21	40.7%

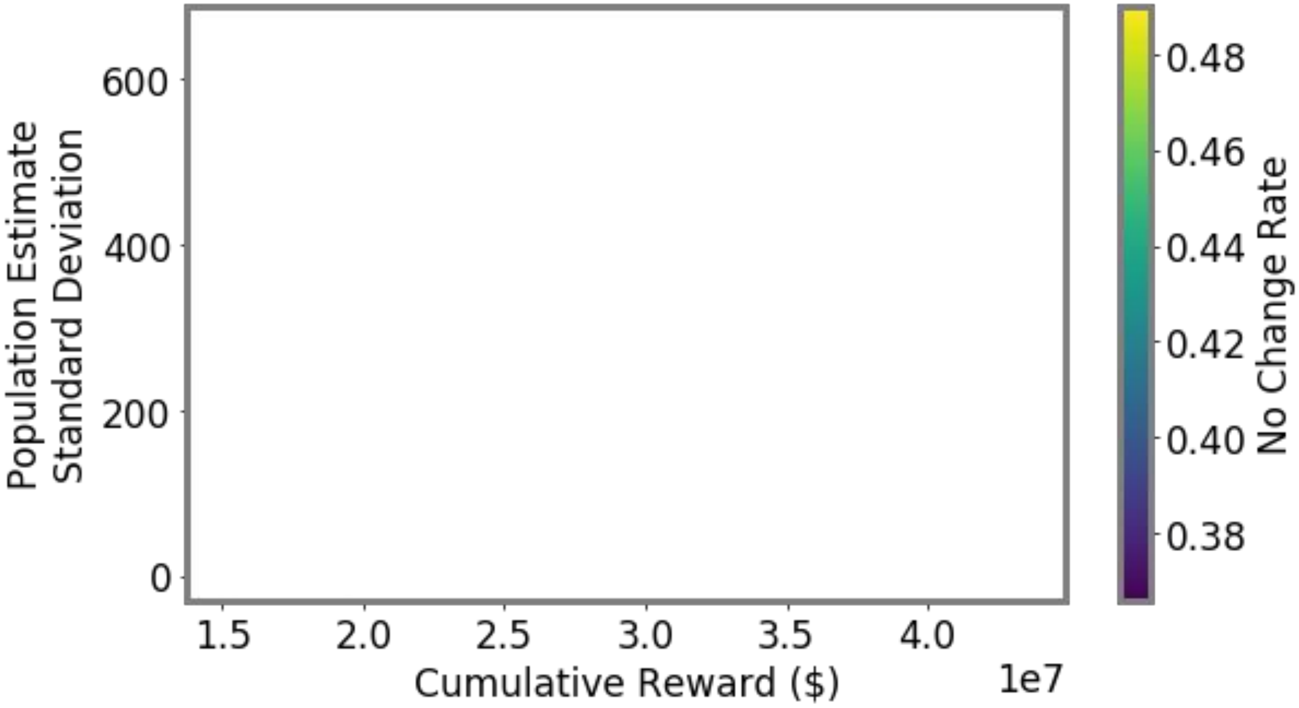
Even some randomness,
reduces bias of model-based estimate,
but not guaranteed.

Experiments

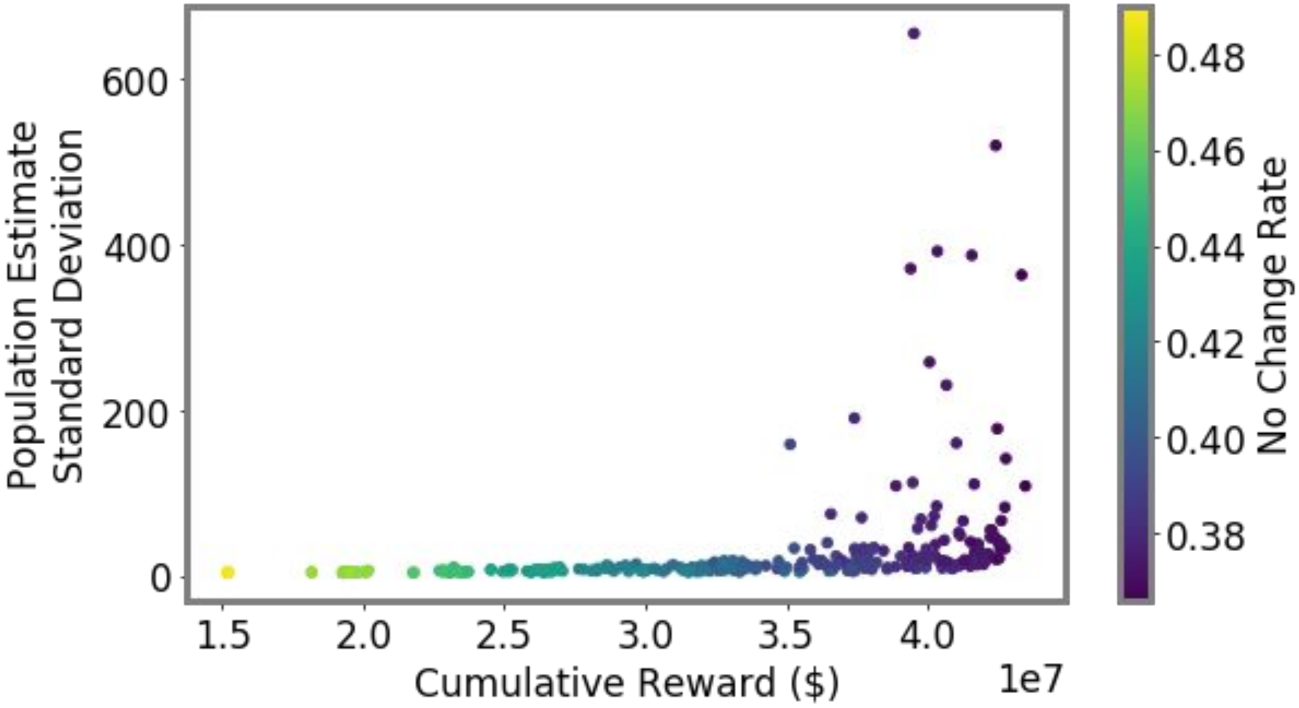
Best Reward Settings

	Policy	R	μ_{PE}	σ_{PE}	μ_{NR}
Unbiased Methods	ABS-1	**$41.5M***	**0.4** ✓	31.0	**37.6%**
	ϵ-only	$41.3M*	4.3✓	37.4	38.3%
	ABS-2	$40.5M*	0.6✓	24.5	38.3%
	Random	$12.7M	1.5✓	**14.7**	53.1%
Biased Methods	Greedy	**$43.6M***	16.4 ✗	8.8	**36.5%**
	UCB-1	$42.4M*	15.3 ✗	9.4	38.6%
	ϵ-Greedy	$41.3M*	**6.1** ✗	**7.5**	38.3%
	UCB-2	$40.7M*	15.6 ✗	10.21	40.7%

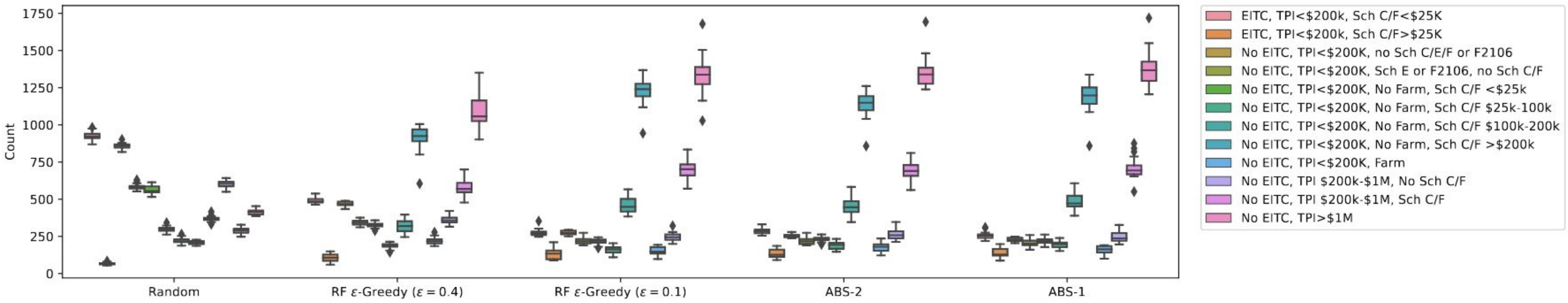
ABS Enables Formal Tradeoff Between Precision and Reward



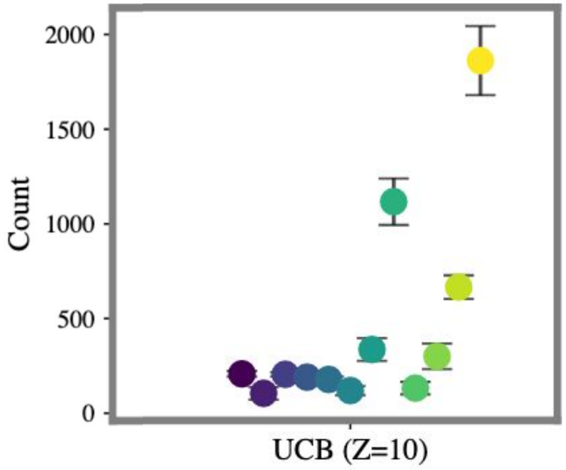
ABS Enables Formal Tradeoff Between Precision and Reward



More optimal methods sample higher incomes



But heteroskedasticity can also drive sampling higher incomes



EITC	TPI	C/F	E/2016	Farm
Y	<$200K	<$25K	N	N
Y	<$200K	>$25K	N	N
N	<$200K	N	N	N
N	<$200K	N	Y	N
N	<$200K	<$25K	N	N
N	<$200K	$25K-100K	N	N
N	<$200K	$100K-200K	N	N
N	<$200K	>$200K	N	N
N	<$200K	N	N	Y
N	$200K-$1M	N	N	N
N	$200K-$1M	Y	N	N
N	>$1M	N	N	N

Takeaways

1. Unbiased estimation of population (e.g., average misreporting) **can still yield returns almost as high as greedy selection, with careful sampling and HT estimation**.
   a. Suggests that a **unified optimize-and-estimate program** could be better and be more efficiently optimized.
2. Model-based population mechanisms are not guaranteed to be unbiased, but bias in practice can be reduced with some randomness.
3. More optimal methods tend to sample higher incomes in our experiments.
4. But heteroskedasticity also drives sampling of higher-incomes in uncertainty-based methods.