Carnegie Mellon University
Computer Science Department

ICLR
International Conference On Learning Representations

Designing Rules to Pick a Rule: Aggregation by Consistency

Ratip Emin Berker

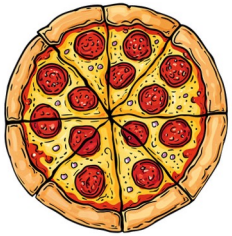
Foundations of Cooperative AI Lab (FOCAL)

Carnegie Mellon University

Joint work with Ben Armstrong, Vincent Conitzer, and Nihar Shah

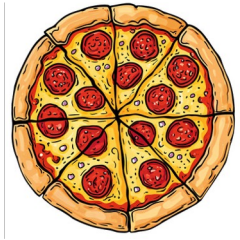
Task: Rank Aggregation

Items:



Task: Rank Aggregation

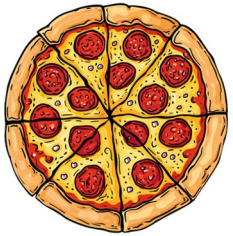
Items:



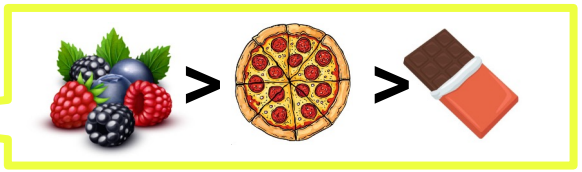
Evaluators:

Task: Rank Aggregation

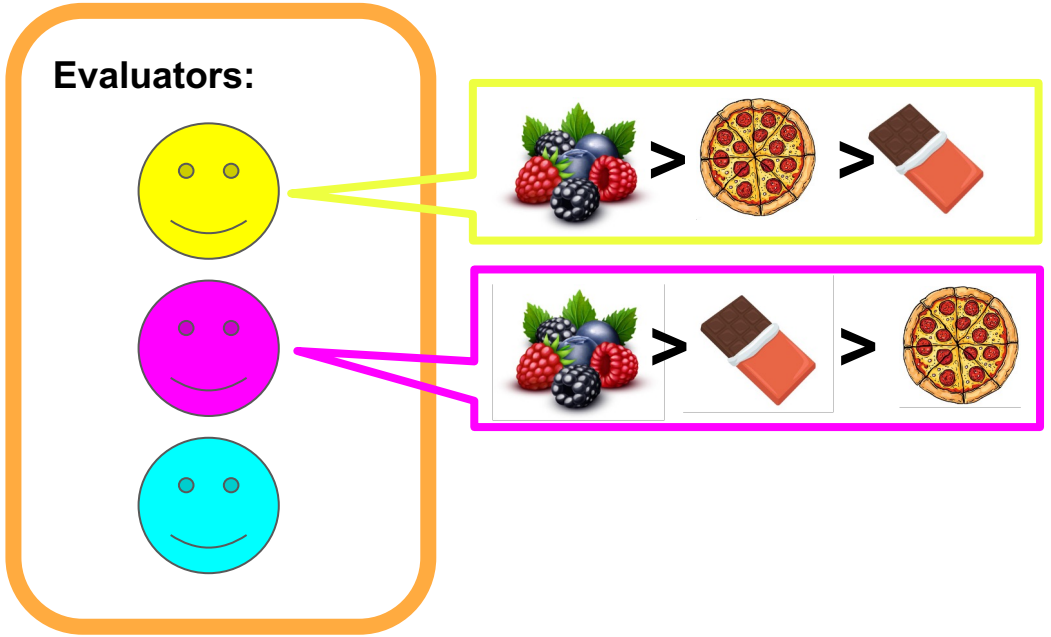
Items:



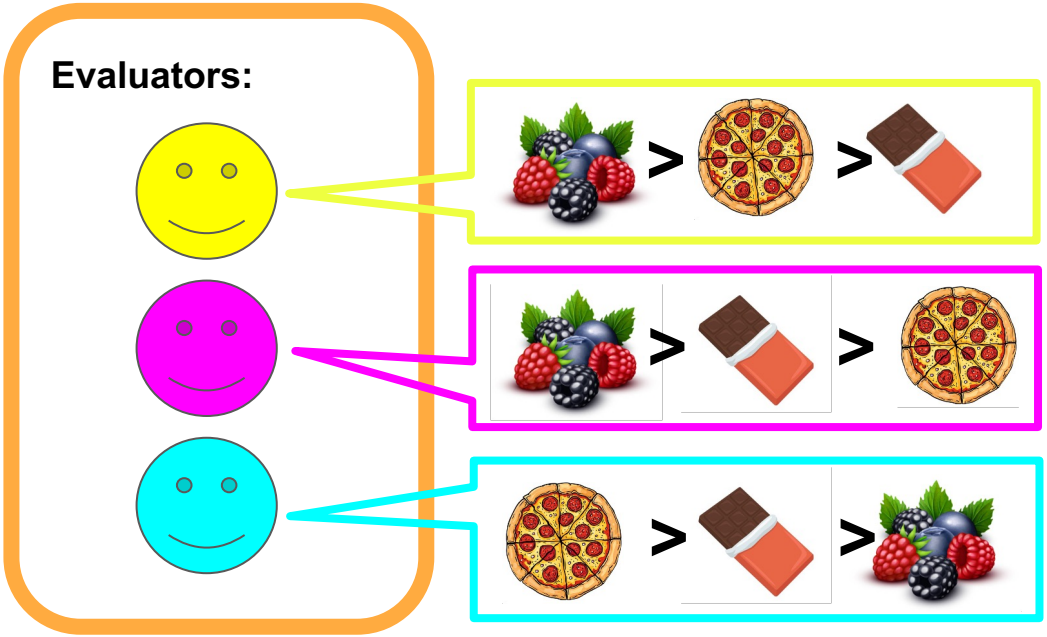
Evaluators:



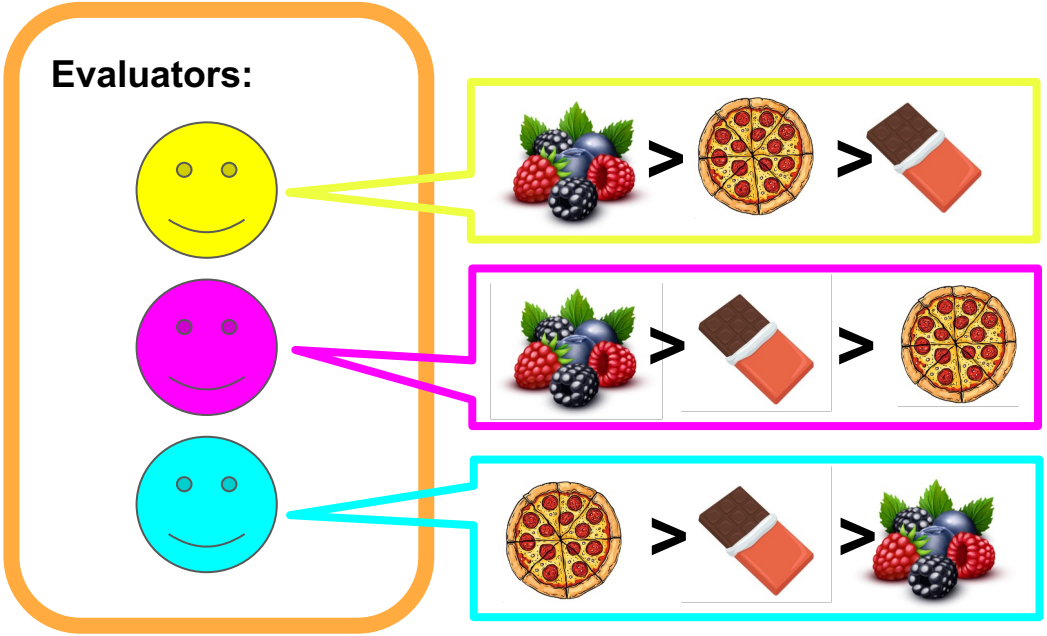
Task: Rank Aggregation



Task: Rank Aggregation



Task: Rank Aggregation



Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

Two approaches

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

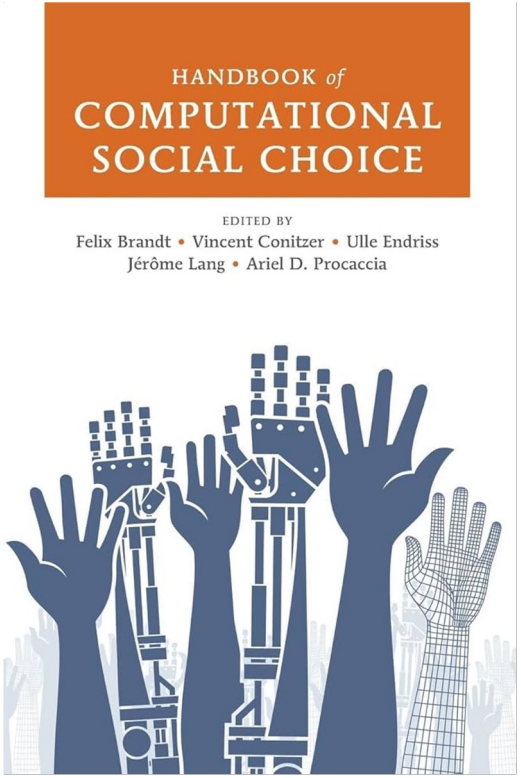
Two approaches

The "axiomatic" approach

1. Decide the desired criteria (*axioms*) you want the aggregation to satisfy
   - anonymity,
   - monotonicity,
   - independence of clones, etc.
2. Design rules that satisfy these axioms.
   - Borda score
   - Single Transferable Vote,
   - Ranked pairs* etc.

No "one rule that satisfies it all"

[Arrow, 1963; Gibbard, 1973; Satterthwaite, 1975]



Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

Two approaches

The “axiomatic” approach

1. Decide the desired criteria (*axioms*) you want the aggregation to satisfy
   - anonymity,
   - monotonicity,
   - independence of clones, etc.
2. Design rules that satisfy these axioms.
   - Borda score
   - Single Transferable Vote,
   - Ranked pairs* etc.

No “one rule that satisfies it all”

[Arrow, 1963; Gibbard, 1973; Satterthwaite, 1975]

The “statistical” approach

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

Two approaches

The "axiomatic" approach

1. Decide the desired criteria (*axioms*) you want the aggregation to satisfy
   - anonymity,
   - monotonicity,
   - independence of clones, etc.
2. Design rules that satisfy these axioms.
   - Borda score
   - Single Transferable Vote,
   - Ranked pairs* etc.

No "one rule that satisfies it all"

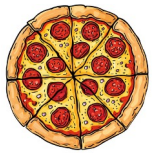
[Arrow, 1963; Gibbard, 1973; Satterthwaite, 1975]

The "statistical" approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

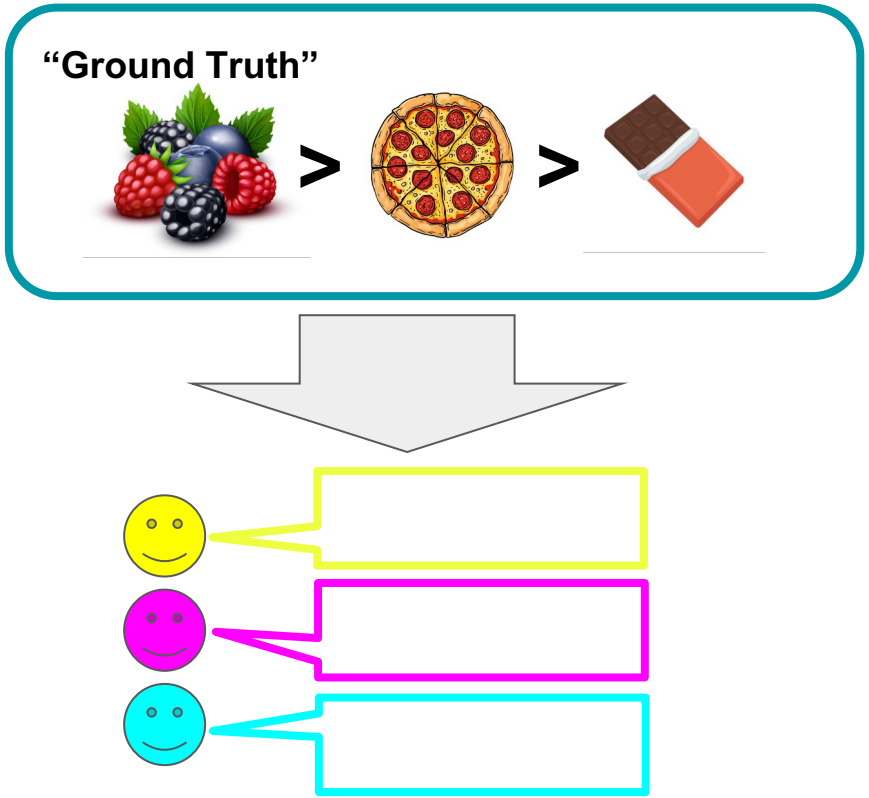
“Ground Truth”

 >  > 

The “statistical” approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.

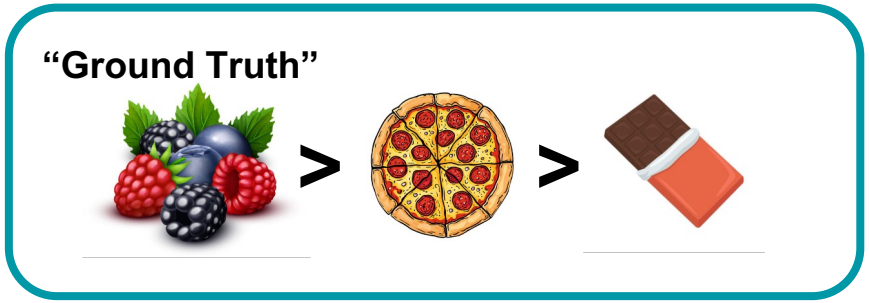
Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

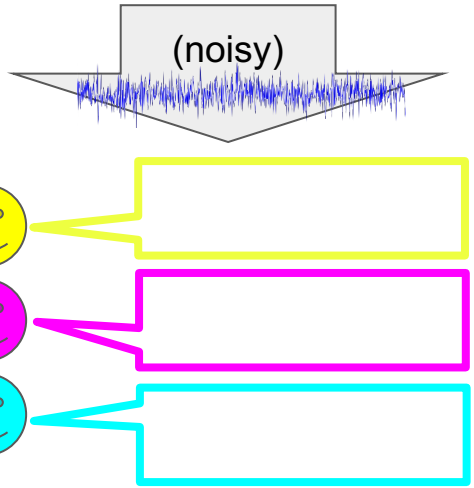


The “statistical” approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

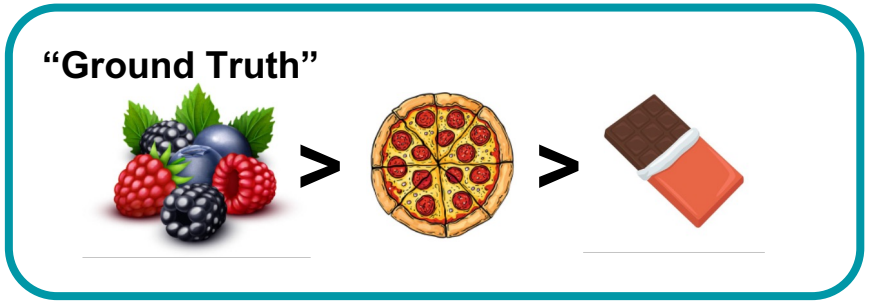


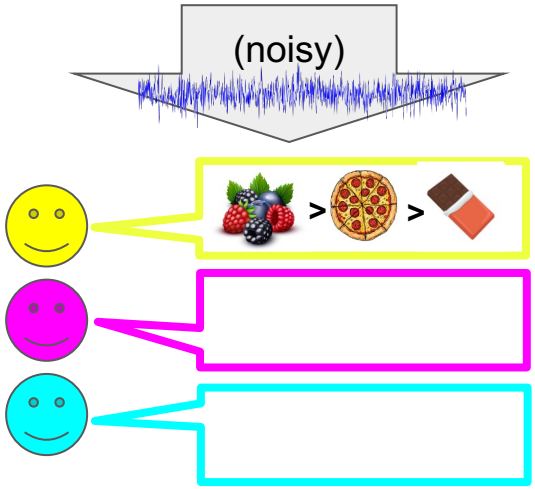


The “statistical” approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

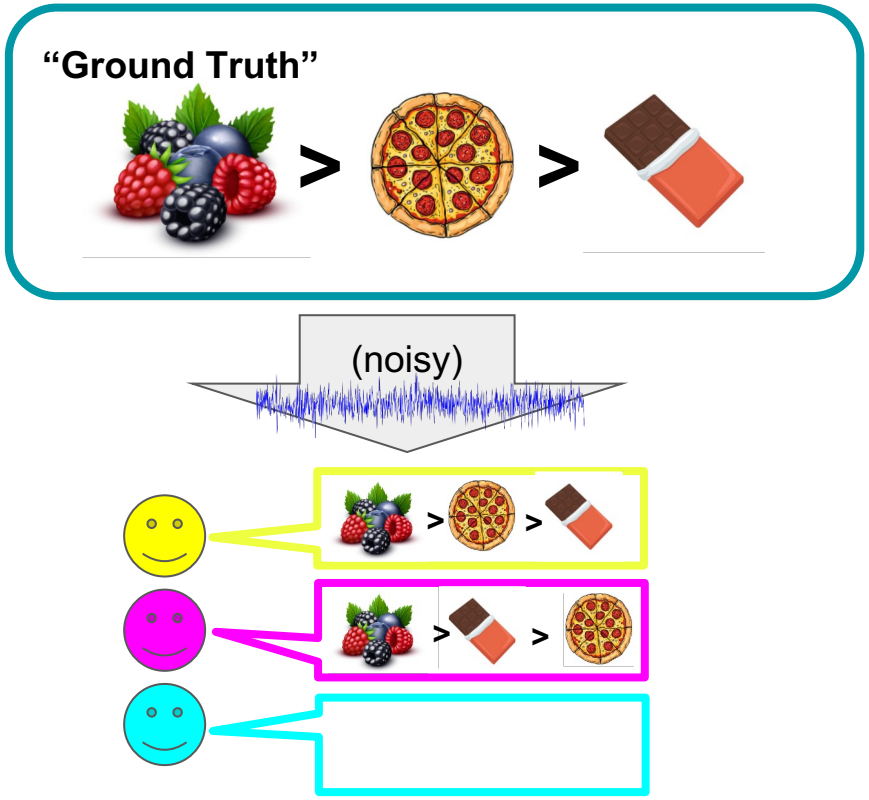




The “statistical” approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.

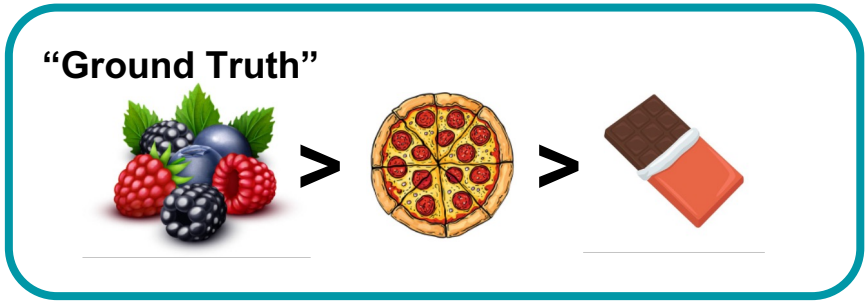
Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

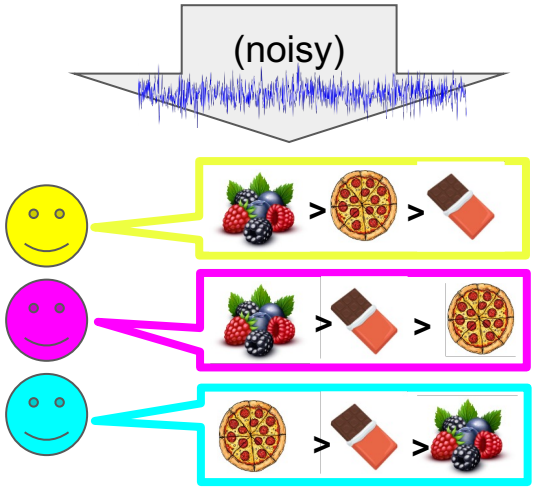


The “statistical” approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

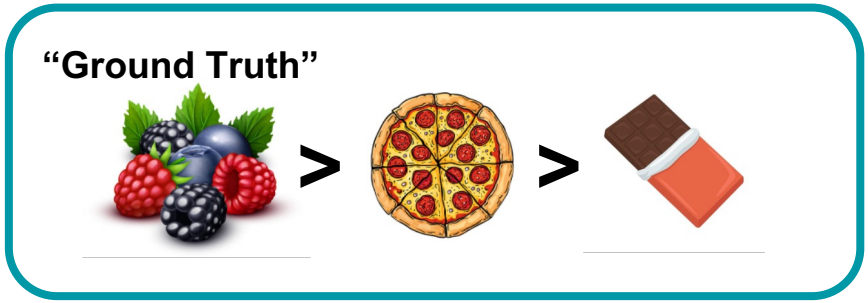


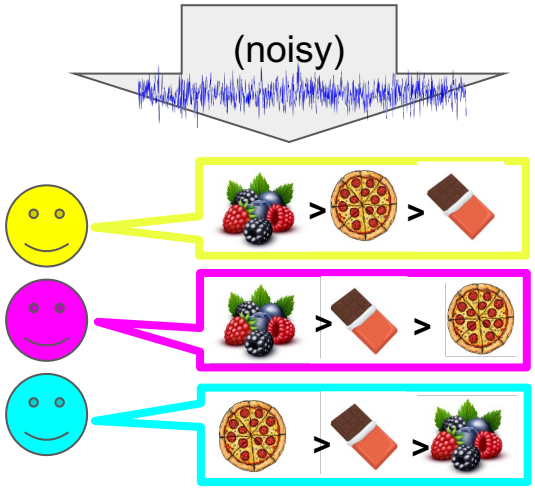


The “statistical” approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

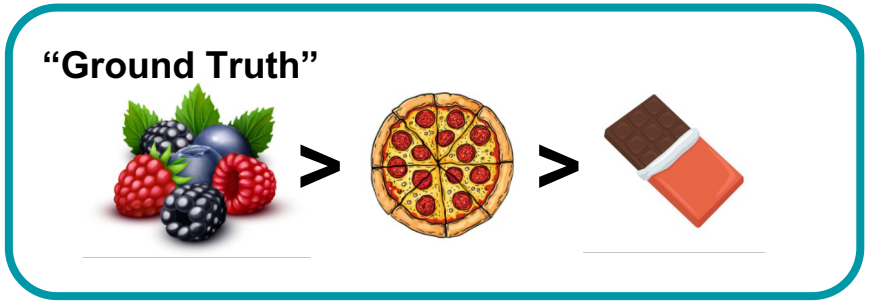


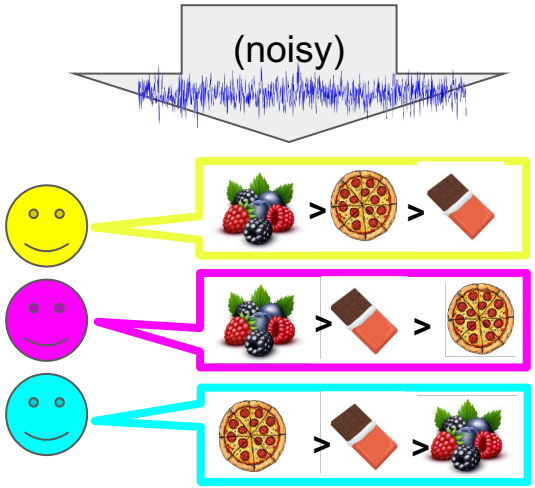


The “statistical” approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.
   - Mallows,
   - Placket Luce,
   - i.i.d. flips, etc.

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

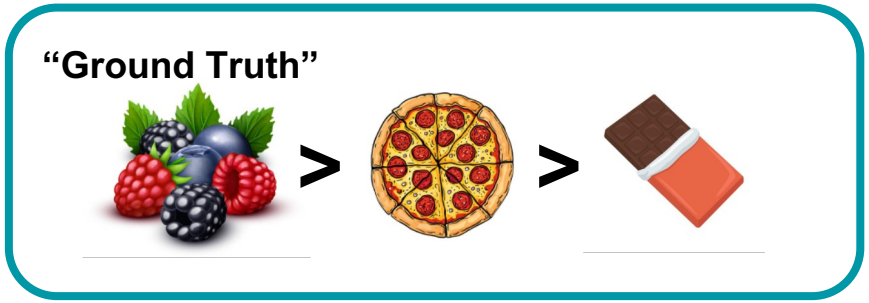


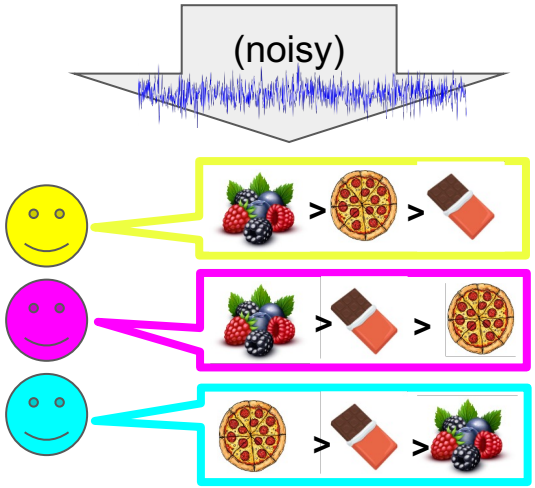


The “statistical” approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.
   - Mallows,
   - Placket Luce,
   - i.i.d. flips, etc.
2. Pick aggregate ranking that maximizes the likelihood of data

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?





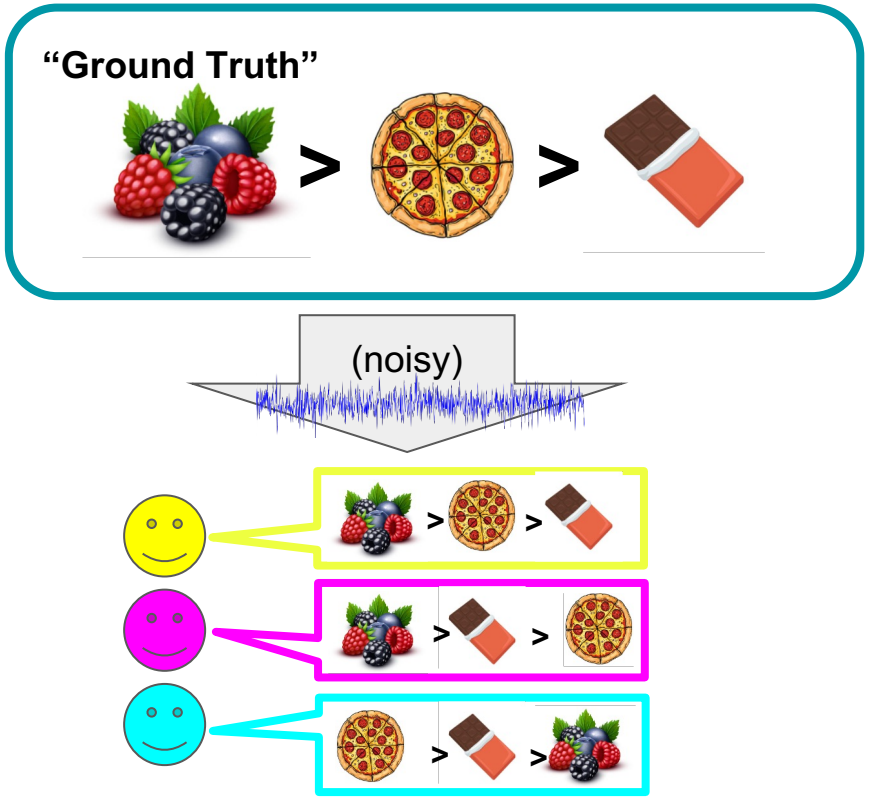
The “statistical” approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.
   - Mallows,
   - Placket Luce,
   - i.i.d. flips, etc.
2. Pick aggregate ranking that maximizes the likelihood of data

Ground truth assumption problematic

e.g., RLHF [Ge et al., 2024]

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?



The “statistical” approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.
   - Mallows,
   - Placket Luce,
   - i.i.d. flips, etc.
2. Pick aggregate ranking that maximizes the likelihood of data

Ground truth assumption problematic
e.g., RLHF [Ge et al., 2024]

Many voting rules are not MLEs
[Conitzer and Sandholm, 2005; Conitzer et al., 2009]

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

Two approaches

The "axiomatic" approach

1. Decide the desired criteria (*axioms*) you want the aggregation to satisfy
   - anonymity,
   - monotonicity,
   - independence of clones, etc.
2. Design rules that satisfy these axioms.
   - Borda score
   - Single Transferable Vote,
   - Ranked pairs* etc.

No "one rule that satisfies it all"

[Arrow, 1963; Gibbard, 1973; Satterthwaite, 1975]

The "statistical" approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.
   - Mallows,
   - Placket Luce,
   - i.i.d. flips, etc.
2. Pick aggregate ranking that maximizes the likelihood of data

Ground truth assumption problematic

e.g., RLHF [Ge et al., 2024]

Many voting rules are not MLEs

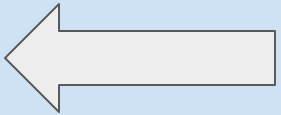
[Conitzer and Sandholm, 2005; Conitzer et al., 2009]

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

Two approaches

The "axiomatic" approach

1. Decide the desired criteria (*axioms*) you want the aggregation to satisfy
   - anonymity,
   - monotonicity,
   - independence of clones, etc.
2. Design rules that satisfy these axioms.
   - Borda score
   - Single Transferable Vote,
   - Ranked pairs* etc.



No "one rule that satisfies it all"

[Arrow, 1963; Gibbard, 1973; Satterthwaite, 1975]

The "statistical" approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.
   - Mallows,
   - Placket Luce,
   - i.i.d. flips, etc.
2. Pick aggregate ranking that maximizes the likelihood of data

Ground truth assumption problematic

e.g., RLHF [Ge et al., 2024]

Many voting rules are not MLEs

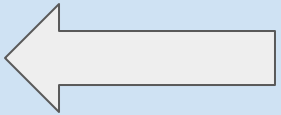
[Conitzer and Sandholm, 2005; Conitzer et al., 2009]

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

Two approaches

The "axiomatic" approach

1. Decide the desired criteria (*axioms*) you want the aggregation to satisfy
   - anonymity,
   - monotonicity,
   - independence of clones, etc.
2. Design rules that satisfy these axioms.
   - Borda score
   - Single Transferable Vote,
   - Ranked pairs* etc.



No "one rule that satisfies it all"

[Arrow, 1963; Gibbard, 1973; Satterthwaite, 1975]

The "statistical" approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.
   - Mallows,
   - Placket Luce,
   - i.i.d. flips, etc.

2. Pick aggregate ranking that maximizes the likelihood of data

Ground truth assumption problematic

e.g., RLHF [Ge et al., 2024]

Many voting rules are not MLEs

[Conitzer and Sandholm, 2005; Conitzer et al., 2009]

Question: Given individual rankings of *items* from *evaluators*, how do we come up with a ***single societal ranking***?

Two approaches

The "axiomatic" approach

1. Decide the desired criteria (*axioms*) you want the aggregation to satisfy
   - anonymity,
   - monotonicity,
   - independence of clones, etc.
2. Design rules that satisfy these axioms.
   - Borda score
   - Single Transferable Vote,
   - Ranked pairs* etc.

No "one rule that satisfies it all"

[Arrow, 1963; Gibbard, 1973; Satterthwaite, 1975]

The "statistical" approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.
   - Mallows,
   - Placket Luce,
   - i.i.d. flips, etc.
2. Pick aggregate ranking that maximizes the likelihood of data

Ground truth assumption problematic

e.g., RLHF [Ge et al., 2024]

Many voting rules are not MLEs

[Conitzer and Sandholm, 2005; Conitzer et al., 2009]

(Our) question: In a given setting, how do we pick which rule to use? In other words, what makes a ***good rule picking rule (RPR)***?

Why should we **pick a rule**, rather than just a ranking?

1. Employing a rule picking rule (RPR) leads to better interpretability by providing a formal justification of why other rules were not adopted.
2. An RPR allows “locking in a rule,” so that future repetitions of the process will employ the same rule, which can mitigate malicious behavior.
3. Different (perfectly reasonable) rules may be appropriate for different settings with different requirements, and RPRs offer a principled way of deciding this.
4. As we will see, many natural RPRs can choose from *any* set of rules, making it easy to continually incorporate novel rules into the aggregation process.

Our contributions

1. We introduce a novel framework for formally defining ***rule picking rules (RPR)***.
2. We introduce our own RPR, **Aggregation by Consistency (*AbC*)**, with the explicit goal of maximizing consistency if the data collection process was repeated.
3. We provide an axiomatic analysis of ***AbC***.
4. We prove that the problem of checking if complete agreement can be achieved for a given input is NP-complete
5. Nevertheless, we provide an implementation of ***AbC*** that is efficient in practice that (a) performs well in experiments (b) provides important insights

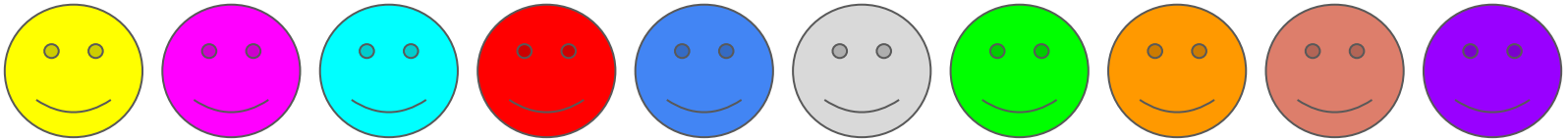
Why focus on ***consistency***?

1) **Distributed peer review**
   - Experiments commonly divide panels into two panels [Obrecht et al., 2007; Fogelholm et al., 2012; Pier et al., 2017; Bast, 2020; Lawrence and Cortes, 2014; Cortes and Lawrence, 2021; Beygelzimer et al., 2023].
   - Taking interpanel consistency as a measure of quality.
2) **Clustering**
   - Out setting is that of unsupervised learning → model selection in clustering.
   - The model that maximize consistency over repetitions maximize accuracy [Luxburg 2010].
3) **Minimum-variance unbiased estimator (MVUE)**
   - Out of all unbiased estimators, the one with min. variance achieves lowest error [Rao, 1949; 156 Chapman and Robbins, 1951].
   - Expected disagreement over two i.i.d copies is a constant factor of variance.
4) **AI Alignment**
   - Nascent interest in applying tools from social choice to AI alignment [Conitzer et al., 2024]
   - Societal goals but limited evaluators, consistency eliminates arbitrariness.

Challenge: How do you find the rule that minimizes disagreement, given only one copy of the process?

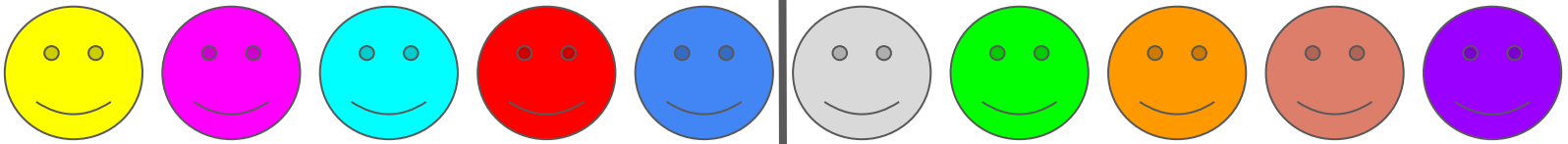
Challenge: How do you find the rule that minimizes disagreement, given only one copy of the process?

Evaluators:



Challenge: How do you find the rule that minimizes disagreement, given only one copy of the process?

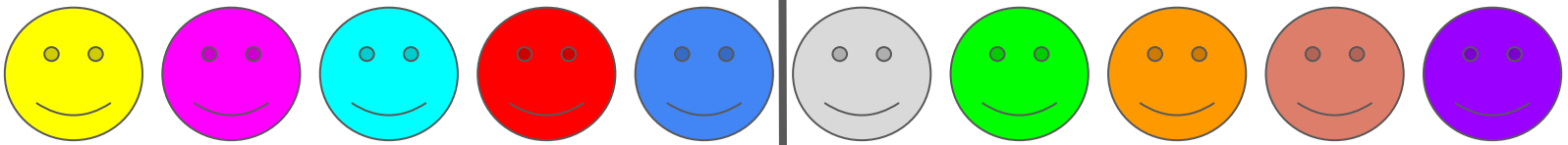
Evaluators:



Get a *random split*

Challenge: How do you find the rule that minimizes disagreement, given only one copy of the process?

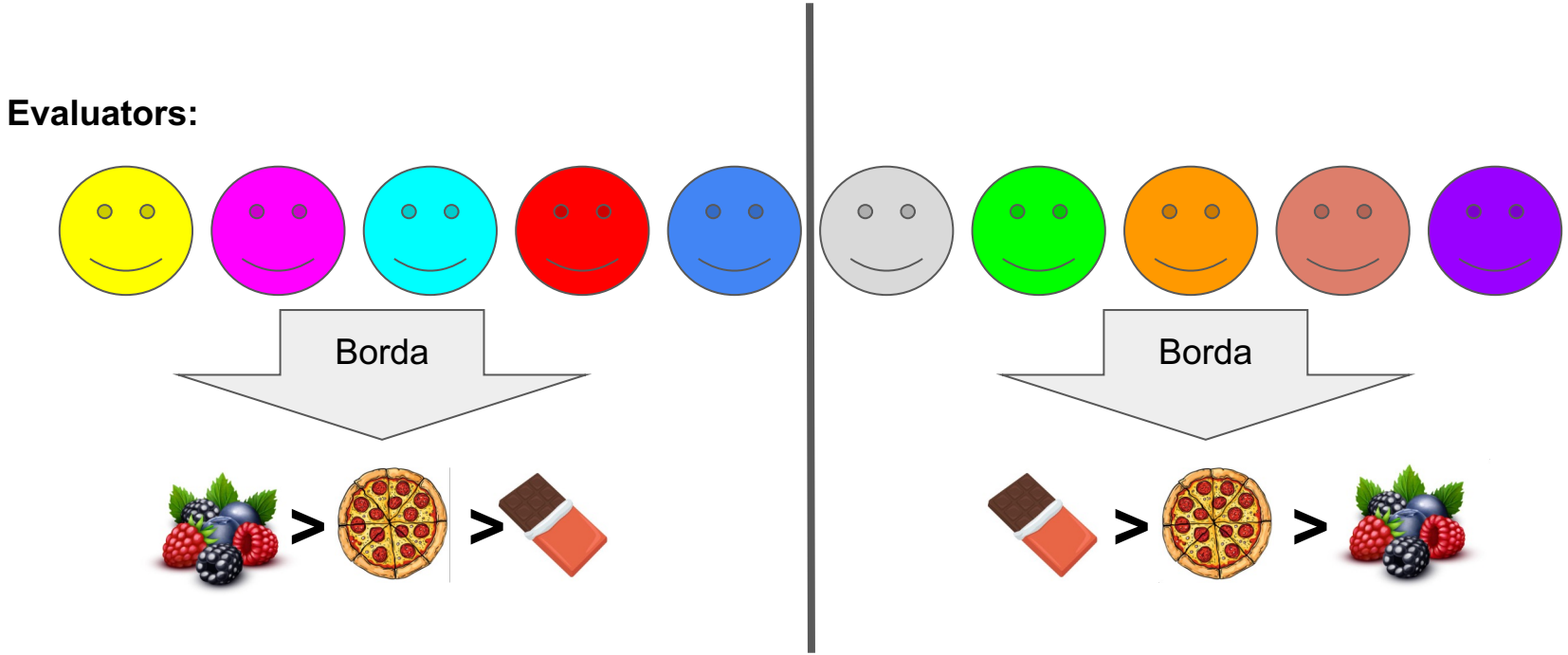
Evaluators:



Get a *random split*

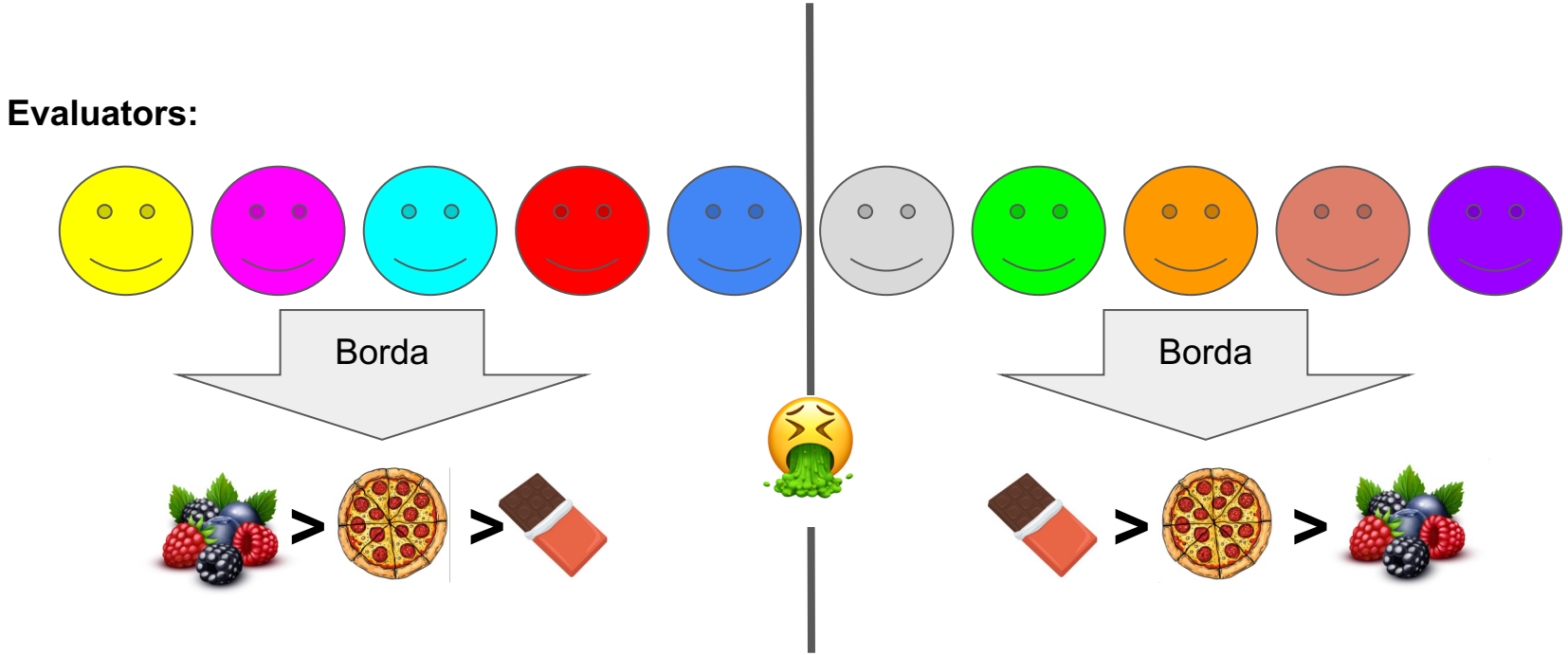
Candidate rules = {Borda, STV, Kemeny}

Challenge: How do you find the rule that minimizes disagreement, given only one copy of the process?



Candidate rules = {**Borda**, STV, Kemeny}

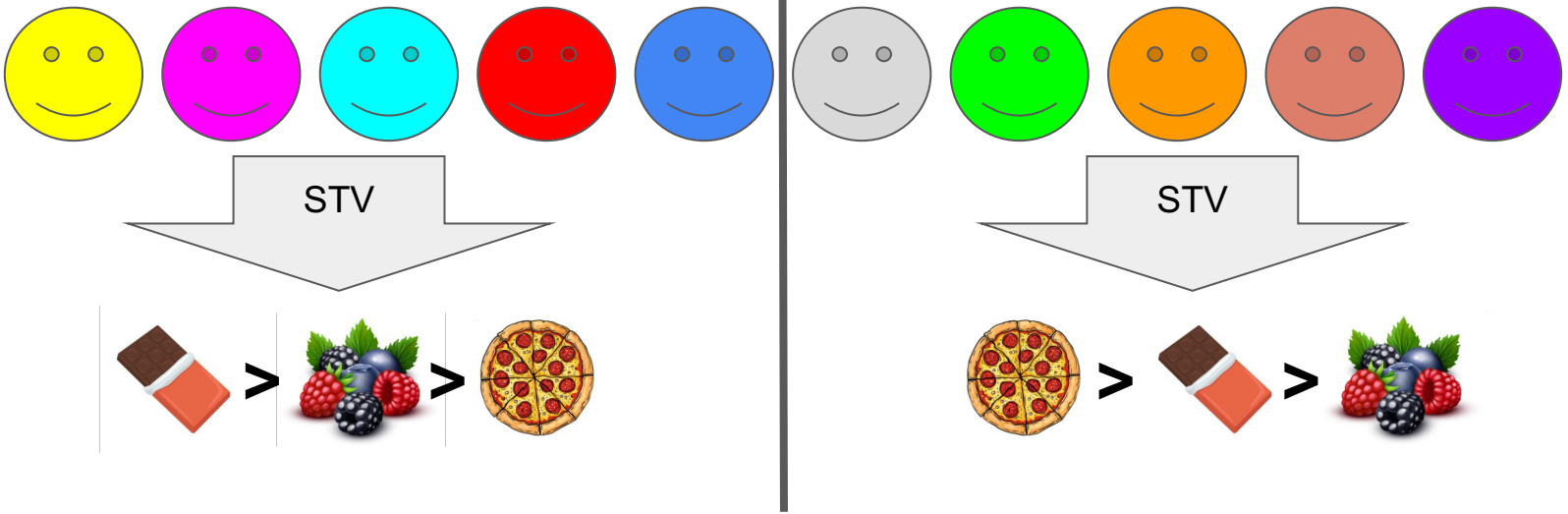
Challenge: How do you find the rule that minimizes disagreement, given only one copy of the process?



Candidate rules = {**Borda**, STV, Kemeny}

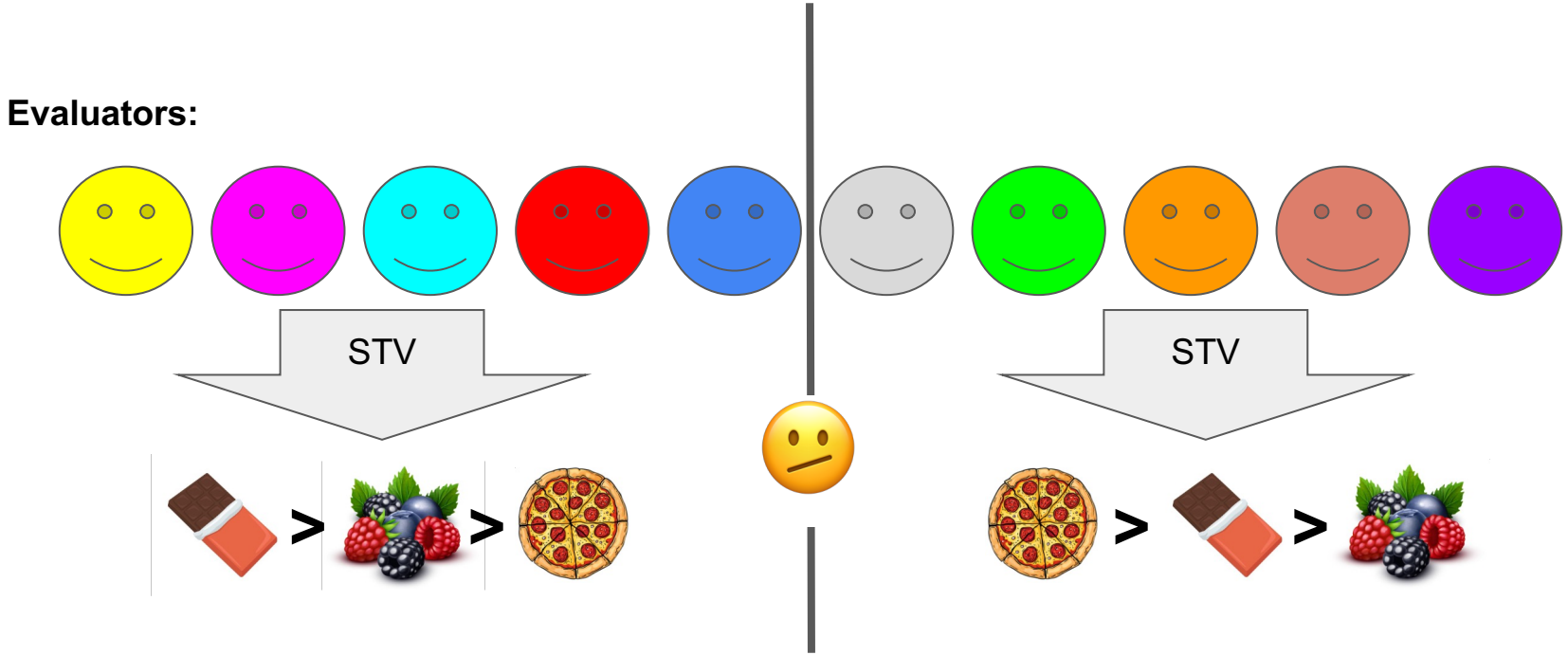
Challenge: How do you find the rule that minimizes disagreement, given only one copy of the process?

Evaluators:



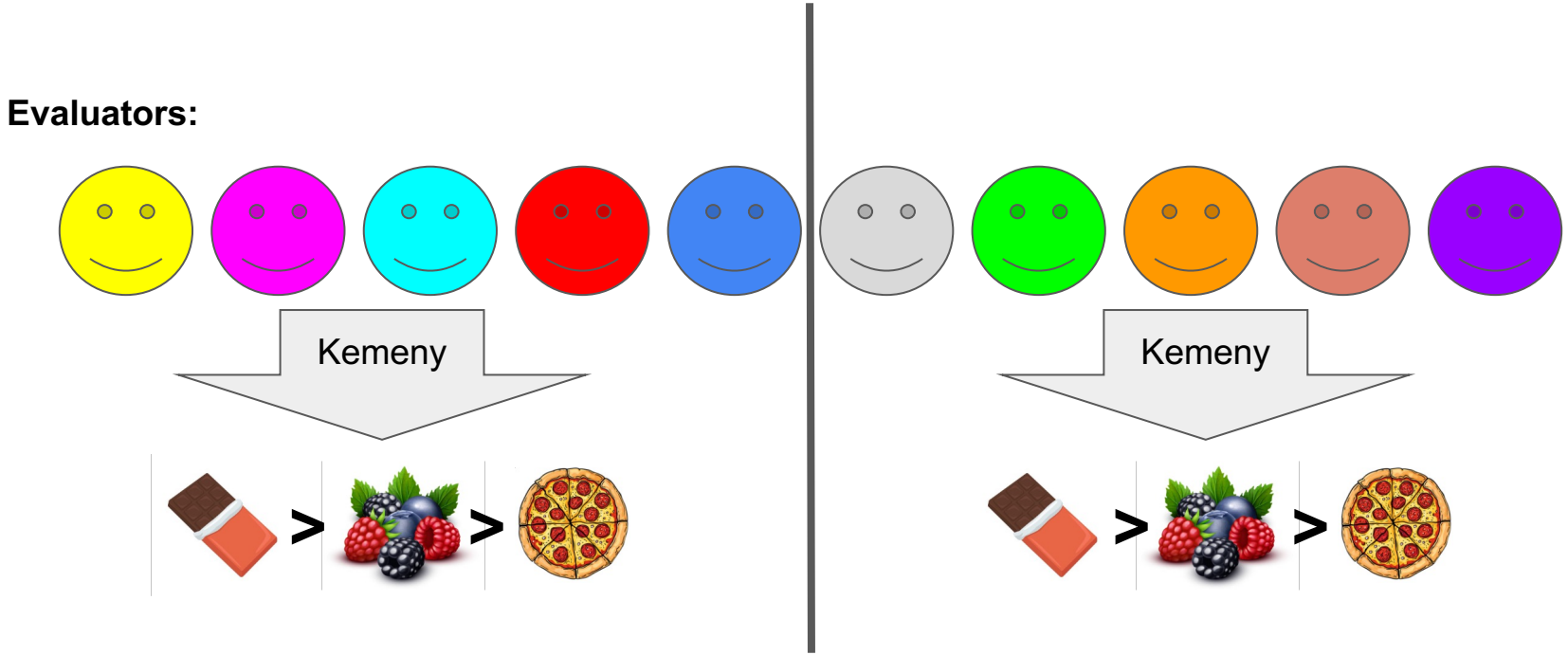
Candidate rules = {Borda, **STV**, Kemeny}

Challenge: How do you find the rule that minimizes disagreement, given only one copy of the process?



Candidate rules = {Borda, STV, Kemeny}

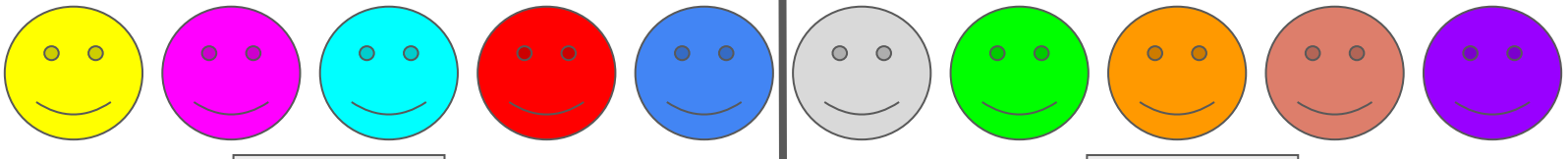
Challenge: How do you find the rule that minimizes disagreement, given only one copy of the process?



Candidate rules = {Borda, STV, **Kemeny**}

Challenge: How do you find the rule that minimizes disagreement, given only one copy of the process?

Evaluators:



Kemeny

Kemeny

Candidate rules = {Borda, STV, Kemeny}

AbC outputs Kemeny

AbC algorithm (informal)

Algorithm 1: Aggregation by Consistency (*AbC*)

Input: A set of evaluations over items, a set of acceptable ("candidate") rules

AbC algorithm (informal)

Algorithm 1: Aggregation by Consistency (*AbC*)

Input: A set of evaluations over items, a set of acceptable ("candidate") rules

1. Split the evaluators uniformly at random into two groups, considering each group as a copy of the process (in line with the peer review experiments above);

AbC algorithm (informal)

Algorithm 1: Aggregation by Consistency (*AbC*)

Input: A set of evaluations over items, a set of acceptable ("candidate") rules

1. Split the evaluators uniformly at random into two groups, considering each group as a copy of the process (in line with the peer review experiments above);
2. Choose the rule (among candidate rules) that minimizes the disagreement of the outputs when applied separately to the two groups;

AbC algorithm (informal)

Algorithm 1: Aggregation by Consistency (*AbC*)

Input: A set of evaluations over items, a set of acceptable ("candidate") rules

1. Split the evaluators uniformly at random into two groups, considering each group as a copy of the process (in line with the peer review experiments above);
2. Choose the rule (among candidate rules) that minimizes the disagreement of the outputs when applied separately to the two groups;
3. Use the chosen rule on the entire data to obtain the final output.

AbC is agnostic to input/output type of the rules that it picks among!

- Only needs a measure of disagreement comparing two outputs (for Step 2)
- Can be used with various output types (ranking, single winner, multi-winner, *etc.*)
- Can be used various evaluation types (complete rankings, partial rankings, *etc.*)

In this paper, we focus on rules that input & output rankings.

In practice, ***AbC*** can pick from **infinitely many** rules

For example, all **positional scoring rules:**

- Parametrized by a vector $s = (s_i)_{i \in [m]}$, where $1 = s_1 \geq s_2 \geq \cdots \geq s_m = 0$.
- Each evaluator "gives" score s_i to their i^{th} ranked item
- Items are are ranked according to total score
- Capture many well known rules
  - Borda
  - Plurality
  - Veto

For a given random split ***AbC*** can find the optimal scoring rule by running a constraint optimization on (s_i) (e.g., using SGD or simulated annealing).

Model

- A set of *voters* $N = \{1,2,\dots,n\}$ and a set of *alternatives* A with $|A| = m$.
- Each voter $i \in N$ has a strict and complete *ranking* $\sigma_i \in \mathcal{L}(A)$ over the alternatives.
- A *preference profile* $\boldsymbol{\sigma} \in \mathcal{L}(A)^n$ consists of the rankings of all voters and represents an election instance.
- A *social welfare function* f maps each profile $\boldsymbol{\sigma}$ to a weak ranking $f(\boldsymbol{\sigma}) \in \mathcal{R}(A)$ over alternatives.
  - e.g. positional scoring rules rank each alternative by their decreasing total score.
  - Set of all positional scoring rules: F_S

Introducing: Rule Picking Rules

Definition: A rule picking rule (RPR) is a function Z that given a set of social welfare functions (called *candidate rules*) F and a profile $\boldsymbol{\sigma}$, outputs a subset of candidate rules $Z(\boldsymbol{\sigma}, F) \subseteq F$.

- $|Z(\boldsymbol{\sigma}, F)| > 1$ means a tie between winner rules.

- **Example:** Welfare-maximizing RPRs.
  - We say an RPR Z is *welfare-maximizing* if there exists a utility function $u: \mathcal{L}(A) \times \mathcal{R}(A) \to \mathbb{R}$ such that $Z(\boldsymbol{\sigma}, F) = \operatorname{argmax}_{f \in F} u(\boldsymbol{\sigma}, f(\boldsymbol{\sigma}))$.
  - (Chooses the rule that maximizes social welfare as defined by u).

Aggregation by Consistency (*AbC*) as an RPR

Given two rankings r_1 and r_2, the *Kendall-Tau distance* between them is

$KT(r_1, r_2)$= # of pairwise disagreements between r_1 and r_2.

Given $\boldsymbol{\sigma}$ consider the following random process:

- Initiate $N_1 = N_2 = \emptyset$.
- For each $i \in N$, uniformly pick a $j \in \{1,2\}$ and let $N_j \leftarrow N_j \cup \{i\}$.
- For $j \in \{1,2\}$, let $\boldsymbol{\sigma}^{(j)}$ be the restriction of $\boldsymbol{\sigma}$ to the voters in N_j.

Aggregation by Consistency (AbC) is the RPR defined as:

$$AbC(\boldsymbol{\sigma}, F) = \operatorname{argmin}_{f \in F} \mathbb{E}[KT(f(\boldsymbol{\sigma}^{(1)}), f(\boldsymbol{\sigma}^{(2)})]$$

(*AbC* picks the candidate rule(s) that minimize expected disagreement.)

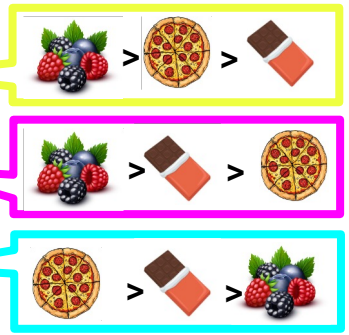
Aggregation by Consistency (AbC) is the RPR defined as:

$$AbC(\boldsymbol{\sigma}, F) = \text{argmin}_{f \in F}\, \mathbb{E}[KT(f(\boldsymbol{\sigma}^{(1)}), f(\boldsymbol{\sigma}^{(2)})]$$

(*AbC* picks the candidate rule(s) that minimize expected disagreement.)

$F = \{f_p, f_v\}$ (plurality and veto)

$$\boldsymbol{\sigma} = \begin{cases} k \times \\ k \times \\ k \times \end{cases}$$



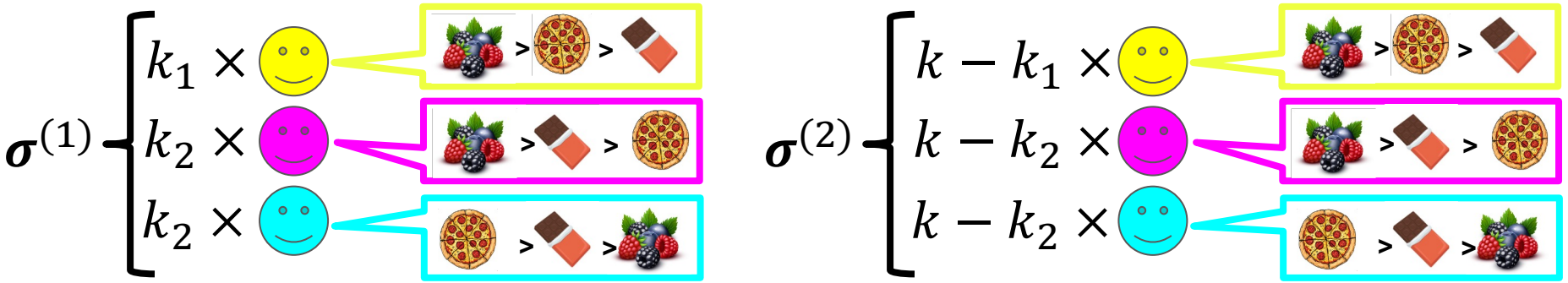
Aggregation by Consistency (AbC) is the RPR defined as:

$$AbC(\boldsymbol{\sigma}, F) = \text{argmin}_{f \in F}\, \mathbb{E}[KT(f(\boldsymbol{\sigma}^{(1)}), f(\boldsymbol{\sigma}^{(2)})]$$

(*AbC* picks the candidate rule(s) that minimize expected disagreement.)

$F = \{f_p, f_v\}$ (plurality and veto)

$$AbC(\boldsymbol{\sigma}, F) = \{f_p\}$$



With prob. ≈1, both sides will have $f_p(\boldsymbol{\sigma}^{(j)})$: > > ($KT$=0)

With prob. ≈1, $f_v(\boldsymbol{\sigma}^{(1)})$ and $f_v(\boldsymbol{\sigma}^{(1)})$ will disagree on all pairs of candidates (KT=3).

Axioms for Rule Picking Rules

Main idea: Without assuming a source for the data, see if the RPR reacts desirably to changes that data.

Axiom #1: **Reversal Symmetry**

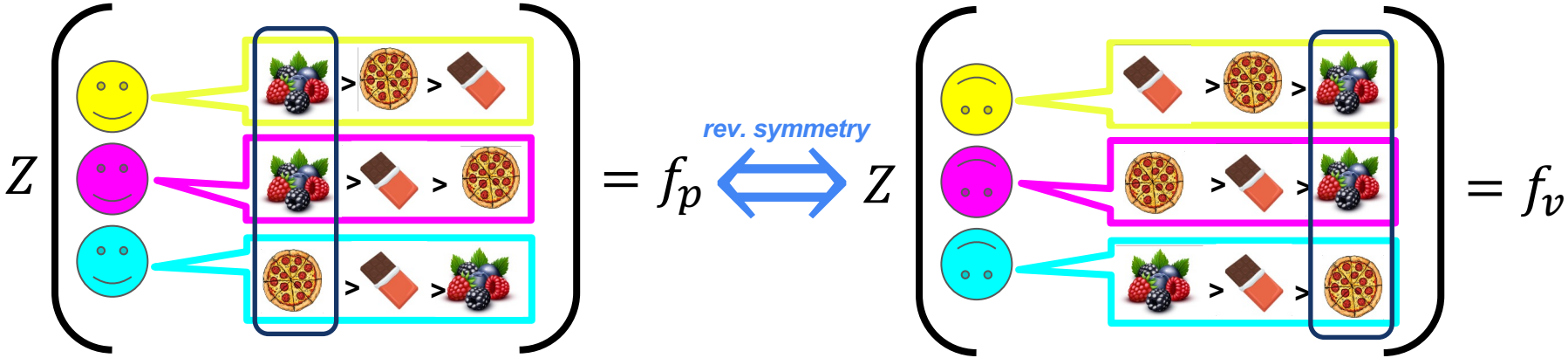
- Given a positional scoring rule f_s with scoring vector $s = (s_1, s_2, \ldots, s_m)$, its ***reverse*** is defined as $\mathrm{rev}(f_s) \equiv f_{s'}$ for $s' = (1 - s_m, 1 - s_{m-1}, \ldots, 1 - s_1)$
  - E.g., the reverse of plurality is veto, the reverse of Borda is Borda.
- The reverse of a profile is simply the profile with all rankings flipped.
- An RPR Z satisfies ***reversal symmetry*** if $\mathrm{rev}\big(Z(\boldsymbol{\sigma}, F)\big) = Z(\mathrm{rev}(\boldsymbol{\sigma}), F)$ for all $F \subseteq F_S$ such that $\mathrm{rev}(F) = F$.

Axioms for Rule Picking Rules

Main idea: Without assuming a source for the data, see if the RPR reacts desirably to changes that data.

Axiom #1: **Reversal Symmetry**

- An RPR Z satisfies ***reversal symmetry*** if $\mathrm{rev}\big(Z(\boldsymbol{\sigma}, F)\big) = Z(\mathrm{rev}(\boldsymbol{\sigma}), F)$ for all $F \subseteq F_S$ such that $\mathrm{rev}(F) = F$.



Axioms for Rule Picking Rules

Main idea: Without assuming a source for the data, see if the RPR reacts desirably to changes that data.

Axiom #1: **Reversal Symmetry**

- An RPR Z satisfies ***reversal symmetry*** if $\mathrm{rev}\big(Z(\boldsymbol{\sigma}, F)\big) = Z(\mathrm{rev}(\boldsymbol{\sigma}), F)$ for all $F \subseteq F_S$ such that $\mathrm{rev}(F) = F$.

Proposition: AbC satisfies reversal symmetry.

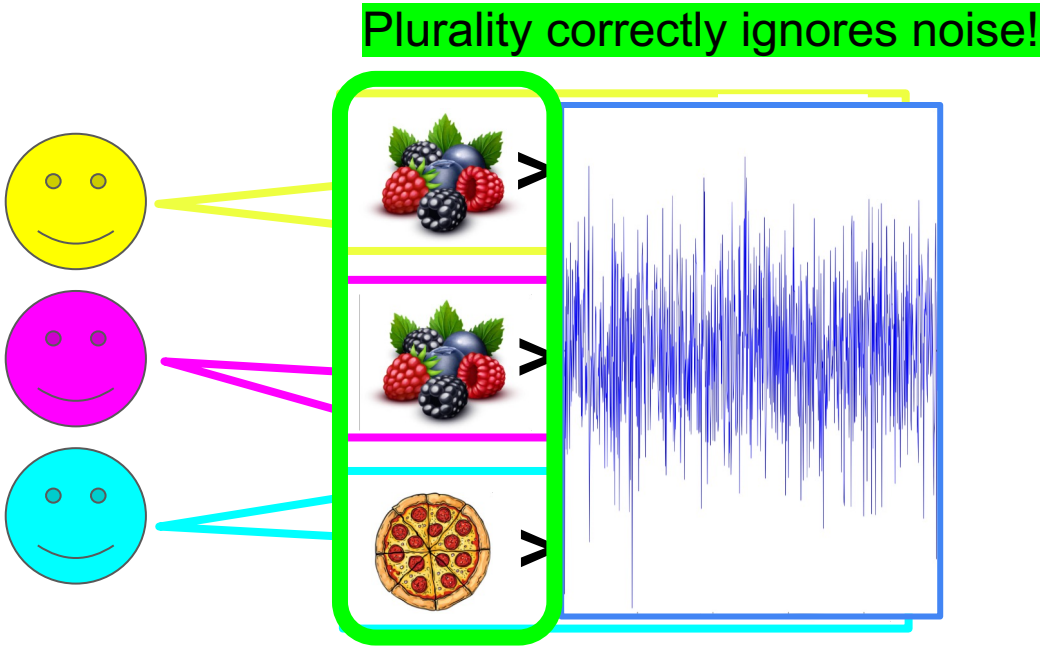
Proof (sketch):

1. $KT(r_1, r_2) = KT(\mathrm{rev}(r_1), \mathrm{rev}(r_2))$
2. $\mathrm{rev}(f_s(\sigma)) = \mathrm{rev}(f_s)(\mathrm{rev}(\sigma))$

Axioms for Rule Picking Rules

Axiom #2: **Shuffling consistency**

Main idea: we would like to “shuffle” a profile such that certain positions no longer give any information. Then, a reasonable RPR should ignore those positions.



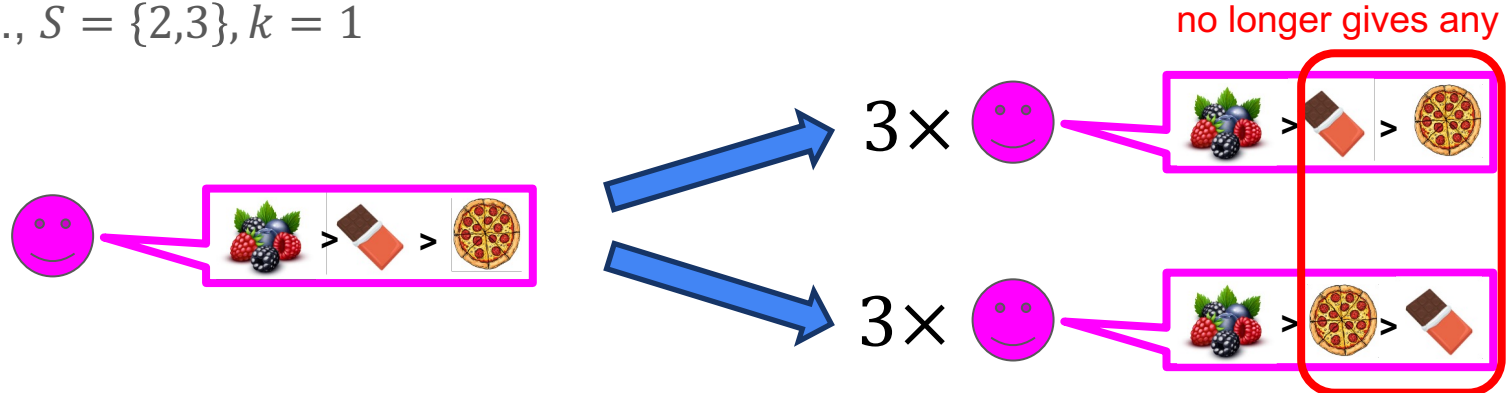
Axioms for Rule Picking Rules

Axiom #2: **Shuffling consistency**

Main idea: we would like to "shuffle" a profile such that certain positions no longer give any information. Then, a reasonable RPR should ignore those positions.

- Given a profile $\boldsymbol{\sigma}$ and any set of indices $S \subseteq \{1,2,\dots,m\}$, $\pi^k(\boldsymbol{\sigma}, S)$ as follows:
  1. Create $k \cdot m!$ copies of each vote
  2. Have each permutation of S equally represented in each vote.

E.g., $S = \{2,3\}, k = 1$



Axioms for Rule Picking Rules

Axiom #2: **Shuffling consistency**

Main idea: we would like to “shuffle” a profile such that certain positions no longer give any information. Then, a reasonable RPR should ignore those positions.

- Given a profile $\boldsymbol{\sigma}$ and any set of indices $S \subseteq \{1,2,\dots,m\}$, $\pi^k(\boldsymbol{\sigma}, S)$ as follows:
  1. Create $k \cdot m!$ copies of each vote
  2. Have each permutation of S equally represented in each vote.
- An RPR Z satisfies *plurality-shuffling-consistency* if $\forall \boldsymbol{\sigma}, \exists k \geq 0$, such that

$$Z\left(\pi^k\left(\boldsymbol{\sigma}, \{2,3,\dots,m\}\right), F\right) = \{f_p\}$$

i.e., given a profile where all but first position is shuffled, Z ignores the noise.

Axioms for Rule Picking Rules

Axiom #2: **Shuffling consistency**

- An RPR Z satisfies *plurality-shuffling-consistency* if $\forall \boldsymbol{\sigma}, \exists k \geq 0$, such that

$$Z\left(\pi^k\left(\boldsymbol{\sigma}, \{2,3,\ldots,m\}\right), F\right) = \{f_p\}$$

i.e., given a profile where all but last position is shuffled, Z ignores the noise.

Proposition: *Any* welfare-maximizing RPR fails plurality-shuffling-consistency.

- Recall: Z is welfare-maximizing if $Z(\boldsymbol{\sigma}, F) = \operatorname{argmax}_{f \in F} u(\boldsymbol{\sigma}, f(\boldsymbol{\sigma}))$.

Theorem: AbC satisfies plurality-shuffling-consistency.

Proof (sketch):

- For each pair of alternatives, write their difference of plurality scores on a single side of the split as a binomial variable
- Use upper/lower tail bounds + linearity of expectation.

Axioms for Rule Picking Rules

Axiom #3: **Union consistency**

- An RPR Z satisfies union consistency if given two profiles $\boldsymbol{\sigma}_a$ and $\boldsymbol{\sigma}_b$, whenever $Z(\boldsymbol{\sigma}_a, F) \cap Z(\boldsymbol{\sigma}_b, F) \neq \emptyset$, we have

$$Z(\boldsymbol{\sigma}_a + \boldsymbol{\sigma}_b, F) = Z(\boldsymbol{\sigma}_a, F) \cap Z(\boldsymbol{\sigma}_b, F)$$

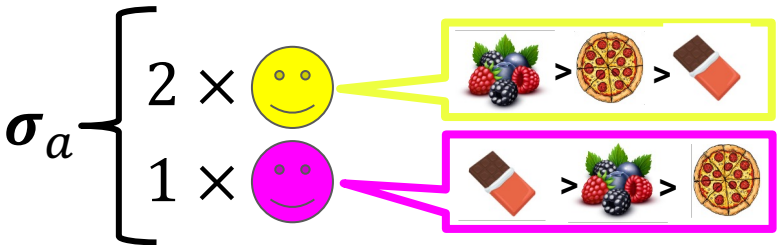
- Inspired by an axiom for social choice functions simply called consistency.
- AbC ***fails*** union consistency! 🥶

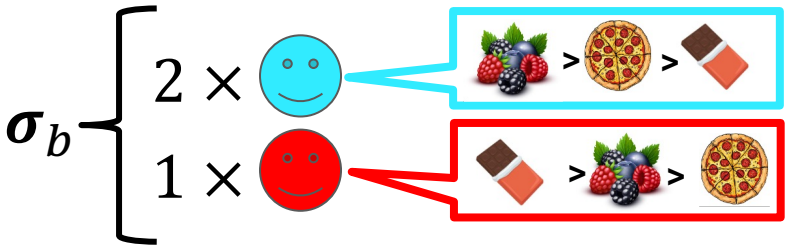
Theorem: No RPR can satisfy all three of reversal symmetry, plurality-shuffling-consistency, and union consistency.

PSC: $Z\left(\pi^k(\boldsymbol{\sigma}_a)\right) = Z\left(\pi^k(\boldsymbol{\sigma}_b)\right) = \{f_p\}$

Rev. symmetry: $Z\left(\pi^k(\boldsymbol{\sigma}_a) + \pi^k(\boldsymbol{\sigma}_b)\right) = \{f_p, f_v\}$

Proof (sketch):





~Nice~ Properties of ***Aggregation by Consistency***

1) Satisfies important natural axioms for rule picking rules, e.g.,
    - reversal symmetry
    - shuffling consistency
2) **Preserves** fundamental social choice axioms if all candidate rules do
    - Smith criterion, Condorcet consistency, majority winner, pairwise majority consistency, unanimity, …

~Nice~ Properties of ***Aggregation by Consistency***

1) Satisfies important natural axioms for rule picking rules, e.g.,
    - reversal symmetry
    - shuffling consistency
2) **Preserves** fundamental social choice axioms if all candidate rules do
    - Smith criterion, Condorcet consistency, majority winner, pairwise majority consistency, unanimity, …

AbC thereby obtains the benefits of the ***axiomatic approach***.

Complexity of ***AbC***

$\mathrm{PerfPos}$: Given a split $\boldsymbol{\sigma}_a, \boldsymbol{\sigma}_b$, does there exists a positional scoring rule $f_s \in F_S$ that achieves perfect agreement, *i.e.*, $KT\left(f_s(\boldsymbol{\sigma}_a), f_s(\boldsymbol{\sigma}_b)\right) = 0$?

Theorem: $\mathrm{PerfPos}$ is NP-Complete (reduction from 3SAT).

- Implies minimizing disagreement is hard to approximate to any mult. factor.

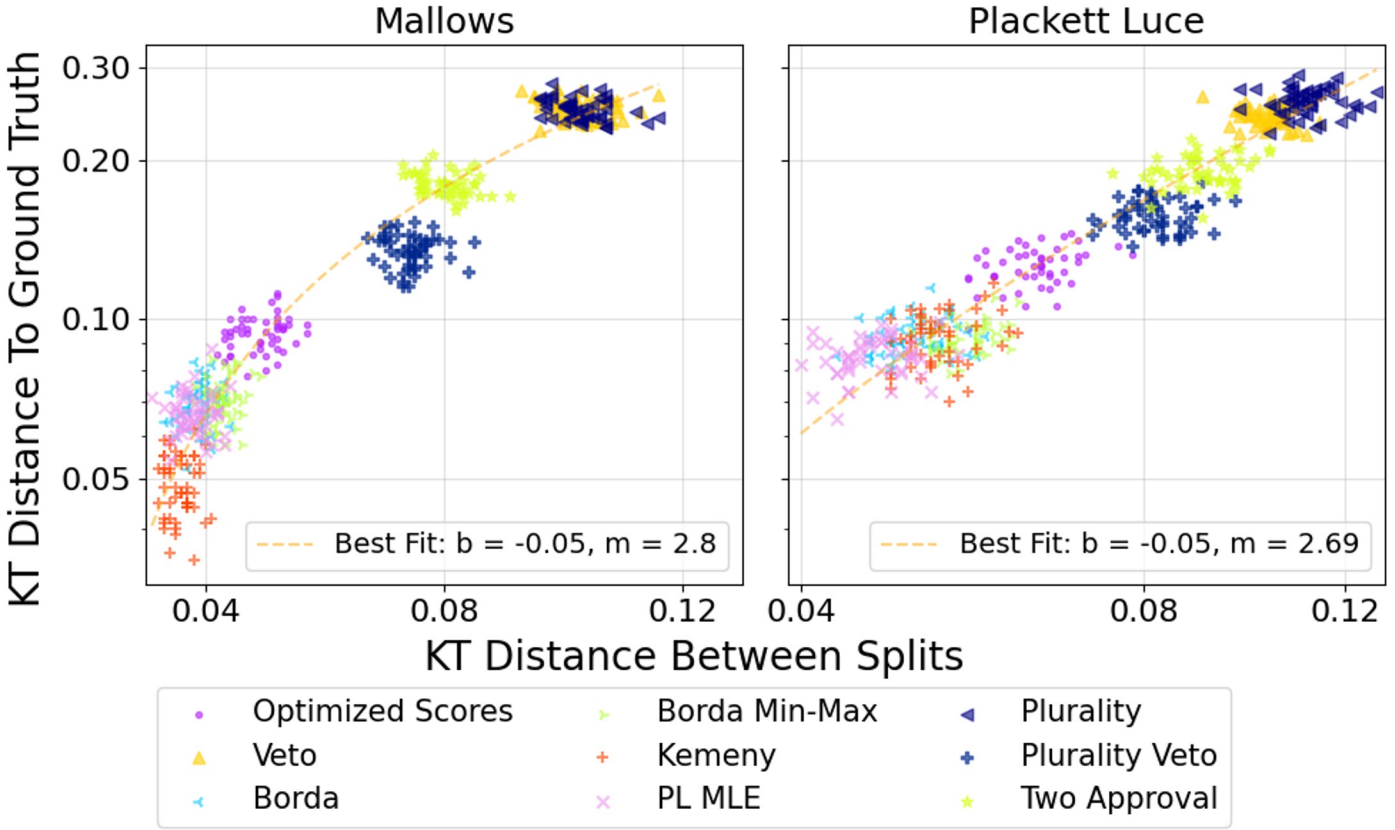
- In practice, AbC can be implemented using **Monte-Carlo sampling**
  - Sample a split, measure disagreement for each score
  - After sampling “enough” splits, return the rule with the average disagreement
  - Can optimize over all scoring rules using SGD/simulated annealing.

Aggregation by Consistency on known distributions

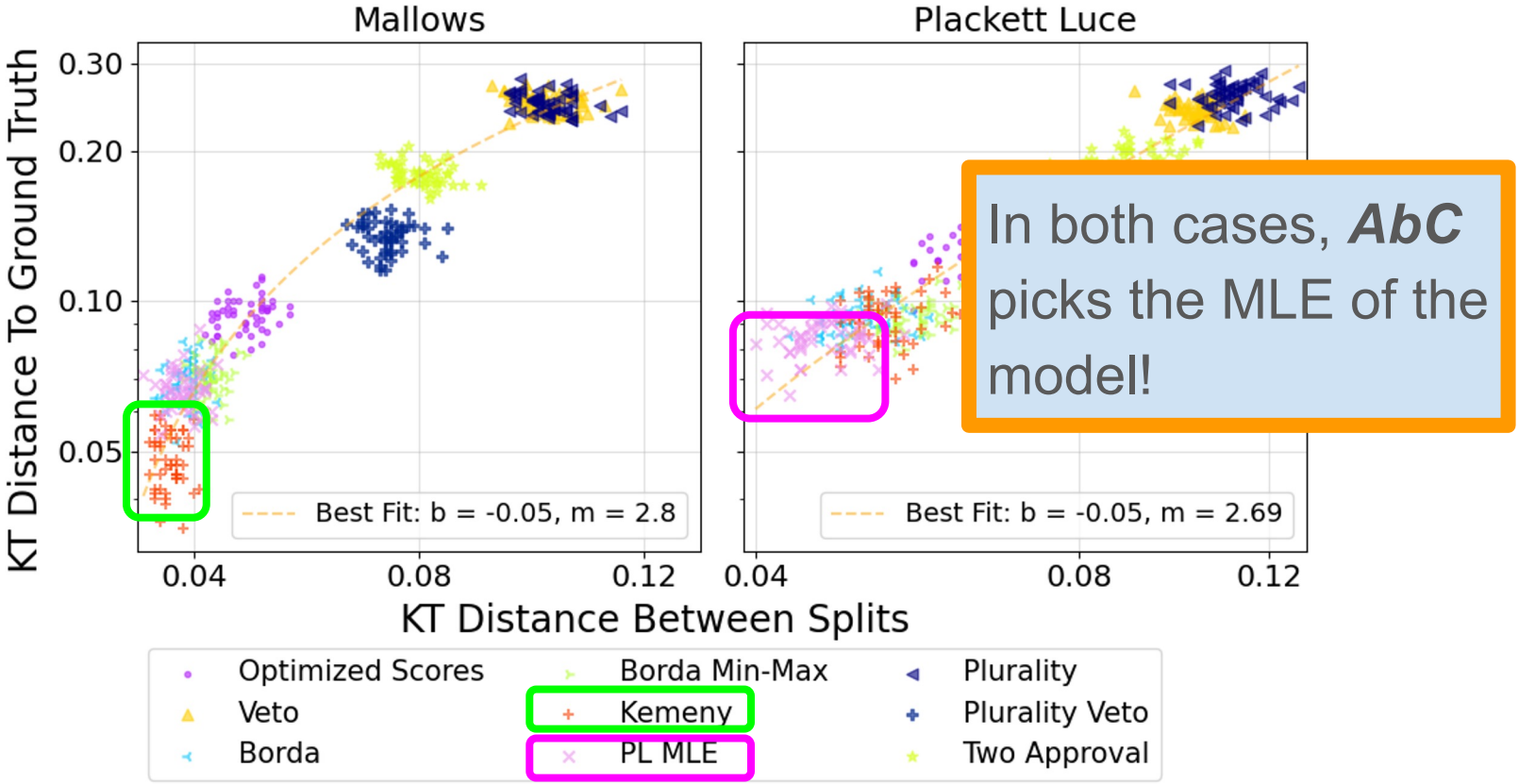
Q: AbC does not assume a ground truth. But what happens on data that is in fact coming from a ground truth?

- Implemented a number of SWFs and ran them on Mallows and PL
- Both models have a “ground truth”, so we can compare each method’s distance to ground truth (error) with disagreement over a random split.
- AbC picks the method with the lowest disagreement
- **Our hope:** The lowest disagreement score also corresponds to the one closest to ground truth

Aggregation by Consistency on known distributions

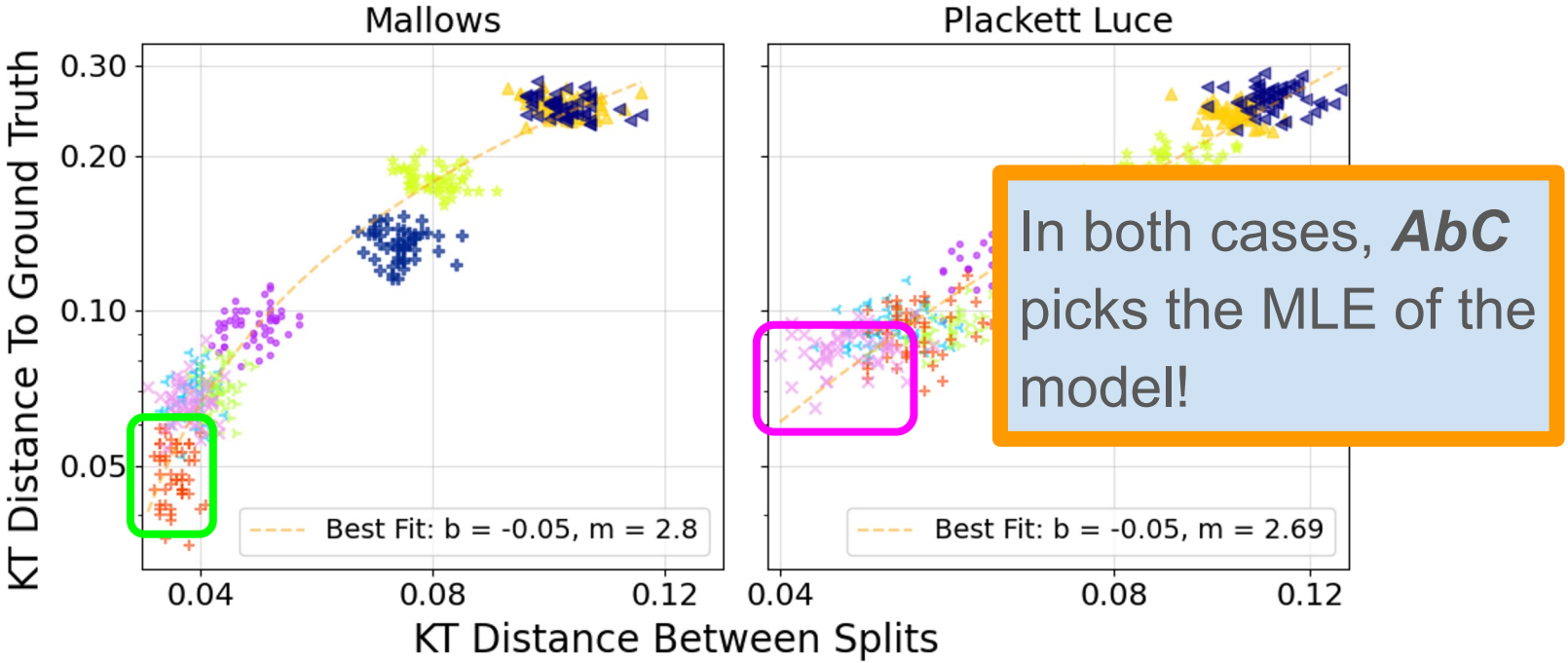


Aggregation by Consistency on known distributions



In both cases, ***AbC*** picks the MLE of the model!

Aggregation by Consistency on known distributions



In both cases, ***AbC*** picks the MLE of the model!

AbC thereby obtains the benefits of the ***statistical approach***.

Two approaches

The "axiomatic" approach

1. Decide the desired criteria (*axioms*) you want the aggregation to satisfy
   - anonymity,
   - monotonicity,
   - independence of clones, etc.
2. Design rules that satisfy these axioms.
   - Borda score
   - Single Transferable Vote,
   - Ranked pairs* etc.

No "one rule that satisfies it all"

[Arrow, 1963; Gibbard, 1973; Satterthwaite, 1975]

The "statistical" approach

1. Treat evaluator rankings as noisy estimates of a *ground truth*.
   - Mallows,
   - Placket Luce,
   - i.i.d. flips, etc.
2. Pick aggregate ranking that maximizes the likelihood of data

Ground truth assumption problematic

e.g., RLHF [Ge et al., 2024]

Many voting rules are not MLEs

[Conitzer and Sandholm, 56 2005; Conitzer et al., 2009]

AbC achieves "best of both worlds"

Demo: ***AbC*** can be used to improve real-world processes

Demo: ***AbC*** can be used to improve real-world processes

- Astronomy peer review dataset provided by Kerzendorf et al. [2020]

Demo: ***AbC*** can be used to improve real-world processes

- Astronomy peer review dataset provided by Kerzendorf et al. [2020]
- Score data rather than rankings, AbC still applicable

Demo: ***AbC*** can be used to improve real-world processes

- Astronomy peer review dataset provided by Kerzendorf et al. [2020]
- Score data rather than rankings. AbC still applicable

Ar. Mean	Min	Max	Median	Harm. Mean	Geo. Mean	Midrange
0.389 ± 0.016	0.453 ± 0.013	0.418 ± 0.013	0.397 ± 0.017	0.401 ± 0.016	0.393 ± 0.016	0.403 ± 0.016

Table 1: Mean and St. Dev. over 1000 splits of KT distance between rankings produced by several metrics on review scores.

Demo: ***AbC*** can be used to improve real-world processes

- Astronomy peer review dataset provided by Kerzendorf et al. [2020]
- Score data rather than rankings. AbC still applicable

Ar. Mean	Min	Max	Median	Harm. Mean	Geo. Mean	Midrange
0.389 ± 0.016	0.453 ± 0.013	0.418 ± 0.013	0.397 ± 0.017	0.401 ± 0.016	0.393 ± 0.016	0.403 ± 0.016

Table 1: Mean and St. Dev. over 1000 splits of KT distance between rankings produced by several metrics on review scores.

- In practice, proposals with a “champion” reviewer are prioritized [Nierstrasz, 2000]

Demo: ***AbC*** can be used to improve real-world processes

- Astronomy peer review dataset provided by Kerzendorf et al. [2020]
- Score data rather than rankings. AbC still applicable

Ar. Mean	Min	Max	Median	Harm. Mean	Geo. Mean	Midrange
0.389 ± 0.016	0.453 ± 0.013	0.418 ± 0.013	0.397 ± 0.017	0.401 ± 0.016	0.393 ± 0.016	0.403 ± 0.016

Table 1: Mean and St. Dev. over 1000 splits of KT distance between rankings produced by several metrics on review scores.

- In practice, proposals with a “champion” reviewer are prioritized [Nierstrasz, 2000]

Demo: ***AbC*** can be used to improve real-world processes

- Astronomy peer review dataset provided by Kerzendorf et al. [2020]
- Score data rather than rankings. AbC still applicable

Ar. Mean	Min	Max	Median	Harm. Mean	Geo. Mean	Midrange
0.389 ± 0.016	0.453 ± 0.013	0.418 ± 0.013	0.397 ± 0.017	0.401 ± 0.016	0.393 ± 0.016	0.403 ± 0.016

Table 1: Mean and St. Dev. over 1000 splits of KT distance between rankings produced by several metrics on review scores.

- In practice, proposals with a “champion” reviewer are prioritized [Nierstrasz, 2000]
- However, using AbC on peer review data suggest focusing (arithmetic/geometric) mean would provide better agreement!

Demo: ***AbC*** can be used to improve real-world processes

- Astronomy peer review dataset provided by Kerzendorf et al. [2020]
- Score data rather than rankings. AbC still applicable

Ar. Mean	Min	Max	Median	Harm. Mean	Geo. Mean	Midrange
0.389 ± 0.016	0.453 ± 0.013	0.418 ± 0.013	0.397 ± 0.017	0.401 ± 0.016	0.393 ± 0.016	0.403 ± 0.016

Table 1: Mean and St. Dev. over 1000 splits of KT distance between rankings produced by several metrics on review scores.

- In practice, proposals with a “champion” reviewer are prioritized [Nierstrasz, 2000]
- However, using AbC on peer review data suggest focusing (arithmetic/geometric) mean would provide better agreement!

Demo: ***AbC*** can be used to improve real-world processes

- Astronomy peer review dataset provided by Kerzendorf et al. [2020]
- Score data rather than rankings. AbC still applicable

Ar. Mean	Min	Max	Median	Harm. Mean	Geo. Mean	Midrange
0.389 ± 0.016	0.453 ± 0.013	0.418 ± 0.013	0.397 ± 0.017	0.401 ± 0.016	0.393 ± 0.016	0.403 ± 0.016

Table 1: Mean and St. Dev. over 1000 splits of KT distance between rankings produced by several metrics on review scores.

- In practice, proposals with a “champion” reviewer are prioritized [Nierstrasz, 2000]
- However, using AbC on peer review data suggest focusing (arithmetic/geometric) mean would provide better agreement!
- Can be used to improve methods in elections, RLHF, etc.

Other empirical results for AbC:

1. Peer review dataset for the ALMA observatory proposals
   - Can be used to study the impact of various aggregation functions they used.
   - E.g. [Meyer et al 2018] considers adding min/max outlier rejection to Borda for evaluating ALMA proposals, but AbC shows this significantly hurts consistency.
2. Real-life elections from Preflib
   - We compare various rules to the rule that was used in practice (usually instant runoff) in terms of consistency
3. Olympics data
   - Given the gold/silver/bronze medals won by each country in each race, how do we give an overall ranking? E.g., 3 Golds better than 5 silvers?
   - Can treat each race as a voter, the medals are its top-3 ranked alternatives
   - Rules such as leximax (order by gold first, then silver, etc.) are still captured by positional scoring rules.

Thank you for your attention!