Human Preference Models: Choice models

CS329H: Machine Learning from Human Preferences

Sanmi Koyejo

Stanford CS

Autumn 2023

Course Logistics Reminder

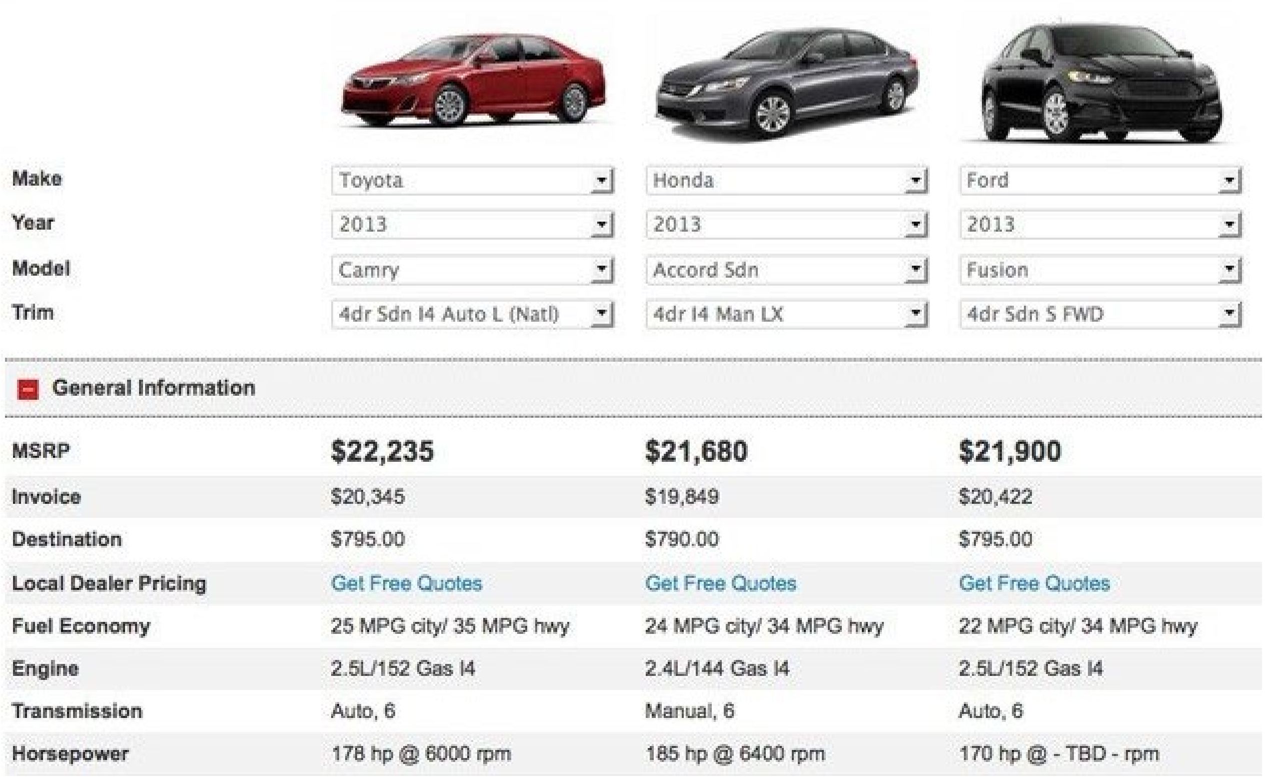
- **You are responsible** for class presentation (1x), scribes (2x) and project (proposal, report, medium, poster)
  - Group and date assignments should be complete
  - Project proposal is due Oct 9 (next Monday)
- **You are expected to peer review** class presentation (1x), scribes (2x) and project (report, medium, poster – all in a group)
  - The focus of peer review is critical feedback, not necessarily grading (we will do this)
  - When possible (i.e., timeline permits), peer review aims for improvement – presenters get to respond.
- **You will need to be in class** at least 3x (though hopefully most lectures)
  - Lecture presentation, speaker scribe day, poster day

Today: Choice Modeling

- Tools to predict the choice behavior of a group of decision-makers in a specific choice context.

Application: Marketing

What features affect a car purchase?



Make	Toyota	Honda	Ford
Year	2013	2013	2013
Model	Camry	Accord Sdn	Fusion
Trim	4dr Sdn I4 Auto L (Natl)	4dr I4 Man LX	4dr Sdn S FWD

General Information

MSRP	**$22,235**	**$21,680**	**$21,900**
Invoice	$20,345	$19,849	$20,422
Destination	$795.00	$790.00	$795.00
Local Dealer Pricing	Get Free Quotes	Get Free Quotes	Get Free Quotes
Fuel Economy	25 MPG city/ 35 MPG hwy	24 MPG city/ 34 MPG hwy	22 MPG city/ 34 MPG hwy
Engine	2.5L/152 Gas I4	2.4L/144 Gas I4	2.5L/152 Gas I4
Transmission	Auto, 6	Manual, 6	Auto, 6
Horsepower	178 hp @ 6000 rpm	185 hp @ 6400 rpm	170 hp @ - TBD - rpm

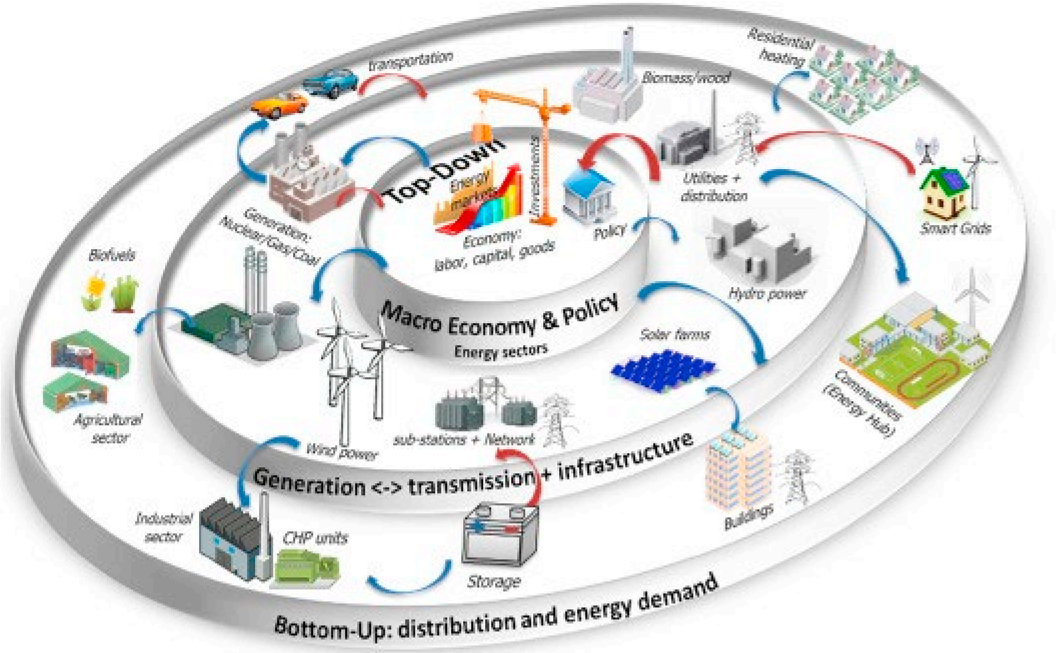
Autoguide.com

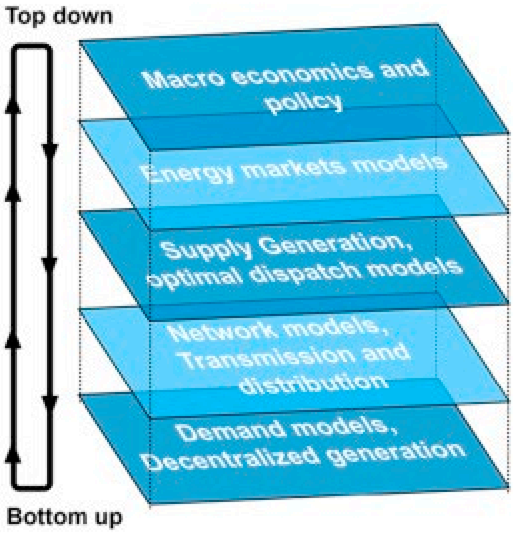
Application: Transportation

- How pricing affects route choice
- How much is a driver willing to pay

Image source: https://www.supplychain247.com/article/8_factors_to_consider_when_choosing_route_optimization_software/locus

Application: Energy Economics

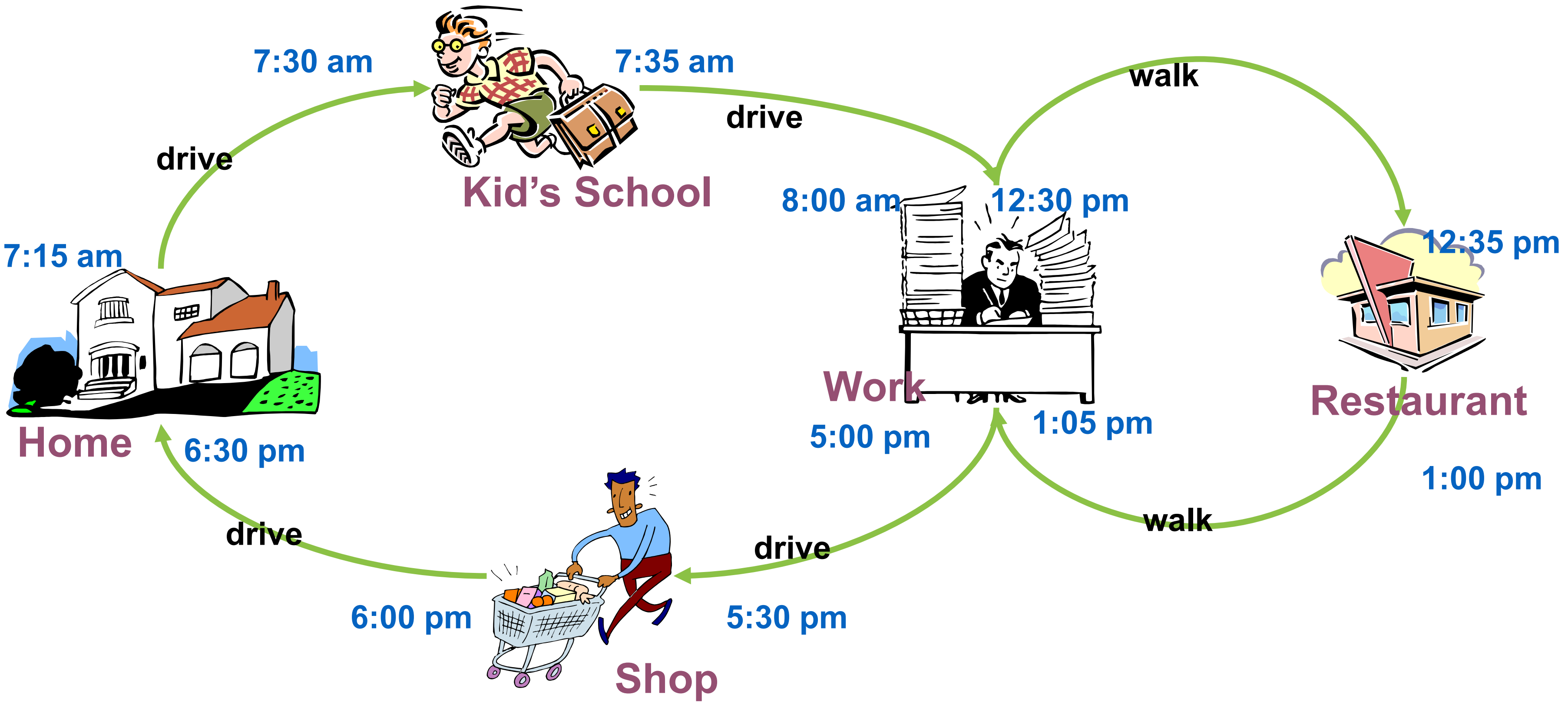




Top down
Macro economics and policy
Energy markets models
Supply Generation, optimal dispatch models
Network models, Transmission and distribution
Demand models, Decentralized generation
Bottom up

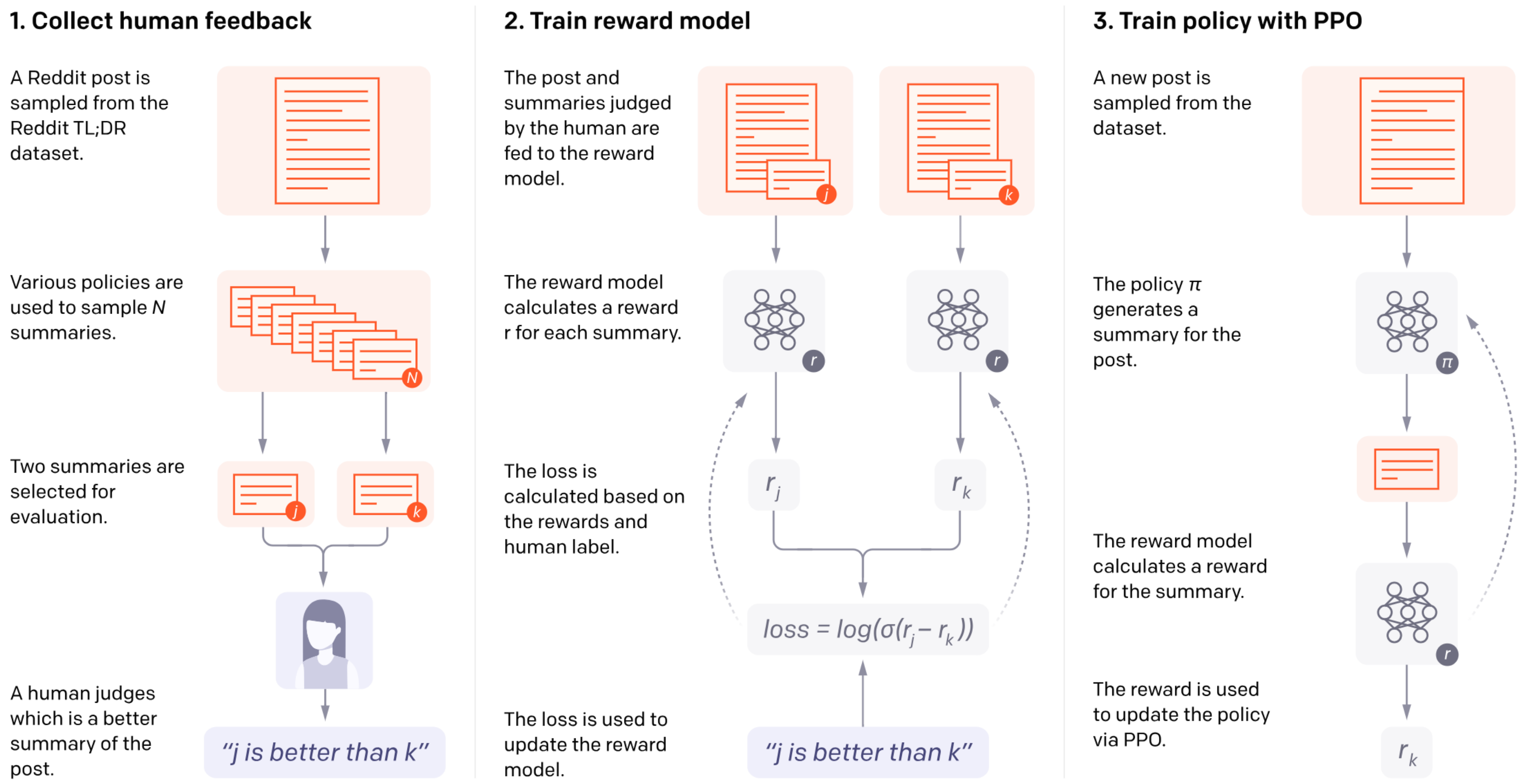
Del Granado, Pedro Crespo, Renger H. Van Nieuwkoop, Evangelos G. Kardakos, and Christian Schaffner. "Modelling the energy transition: A nexus of energy system and economic models." Energy strategy reviews 20 (2018): 229-235.

Example: Daily activity-travel pattern of an individual



Source: Chandra Bhat, " General introduction to choice modeling"

Application: RL and Language



1. Collect human feedback

A Reddit post is sampled from the Reddit TL;DR dataset.

Various policies are used to sample *N* summaries.

Two summaries are selected for evaluation.

A human judges which is a better summary of the post.

"j is better than k"

2. Train reward model

The post and summaries judged by the human are fed to the reward model.

The reward model calculates a reward r for each summary.

The loss is calculated based on the rewards and human label.

The loss is used to update the reward model.

$loss = log(\sigma(r_j - r_k))$

"j is better than k"

3. Train policy with PPO

A new post is sampled from the dataset.

The policy π generates a summary for the post.

The reward model calculates a reward for the summary.

The reward is used to update the policy via PPO.

https://openai.com/research/learning-to-summarize-with-human-feedback

History

- Thurstone research into food preferences in the 1920s
- Microeconomics: Random Utility Theory (1970s)
    - McFadden: Nobel prize in 2000 for the theoretical basis for discrete choice.
- Psychology: Duncan Luce and Anthony Marley
    - Luce, R. Duncan (1959). "Conditional logit analysis of qualitative choice behavior"
- Early use in marketing
    - Predict demand for new products that are potentially expensive to produce
- Early use in transportation:
    - predict usage of transportation resources
    - E.g., used by McFadden to predict the demand for the Bay Area Rapid Transit (BART) before it was built

Why are we studying choice models?

- Human preferences are often gathered by asking for choices across alternatives
- Basic choice models are the workhorse for ML from preferences (Bradley-Terry, Plackett Luce)
- Our discussion will highlight some of the key assumptions, e.g., utility and rationality
    - We will cover models originally built for discrete/finite choices, which have been extended to ML applications (conditional choices)

(Discrete) choice models

- Models designed to capture decision-process of individuals
- True utility is not observable, but perhaps can measure via preferences over choices
- **Main assumption:** utility (benefit, or value) that an individual derives from item A over item B is a function of the frequency that they choose item A over item B in repeated choices.

- **Useful Note:** “Utility” in choice models <=> “Reward” in RL

Modeling: Discrete choice

- Choices are collectively exhaustive, mutually exclusive, and finite

$$y_{ni} = \begin{cases} 1, & \text{if } U_{ni} > U_{nj} \ \forall j \neq i \\ 0, & \text{otherwise} \end{cases}$$

$$U_{ni} = H_{ni}(z_{ni})$$

- $z_{n,i}$ are variables describing the individual attributes and the alternative choices
- $H_{ni}(z_{ni})$ is a stochastic function, e.g., linear $H_{ni}(z_{ni}) = \beta z_{ni} + \epsilon_{ni}$
    - ϵ_{ni} is are unobserved individual factors

Implications of the choice model

- Only the utility differences matter

$$
\begin{aligned}
P_{ni} &= Pr(y_{ni} = 1) \\
&= Pr\left(U_{ni} > U_{nj},\ \forall j \ \neq i\right) \\
&= Pr\left(U_{ni} - U_{nj} > 0,\ \forall\ j \neq i\right)
\end{aligned}
$$

- Note that utility here is scale-free
    - May be invariant to monotonic transformations
    - Ok within a single context, but will need to normalize for comparing across datasets
    - Common approach: normalize scale by standardizing the variance

Example: Binary choice with individual attributes

- Benefit of action depends on s_n = individual characteristics

$$\begin{cases} U_n = \beta \boldsymbol{s}_n + \varepsilon_n \\ y_n = \begin{cases} 1 & U_n > 0 \\ 0 & U_n \leqslant 0 \end{cases} \\ \varepsilon \sim \text{Logistic} \end{cases} \quad \Rightarrow \quad P_{n1} = \frac{1}{1 + \exp(-\beta \boldsymbol{s}_n)}$$

- Replacing $\epsilon \sim$ Standard Normal gives the probit model

$$P_{n1} = \Phi(\beta s_n)$$

- Where $\Phi(.)$is the normal CDF

Example: Utility is linear function of variables that vary over alternatives (Bradley-Terry Model)

- The utility of each alternative depends on the attributes of the alternatives (which may include individual attributes)
- Unobserved terms are assumed to have an extreme value distribution

$$\begin{cases} U_{n1} = \beta z_{n1} + \varepsilon_{n1} \\ U_{n2} = \beta z_{n2} + \varepsilon_{n2} \\ \varepsilon_{n1}, \varepsilon_{n2} \sim \text{iid extreme value} \end{cases} \quad \Rightarrow \quad P_{n1} = \frac{\exp(\beta z_{n1})}{\exp(\beta z_{n1}) + \exp(\beta z_{n2})}$$

- Equivalently $P_{n1} = \dfrac{1}{1 + \exp(-\beta(z_{n1} - z_{n2}))}$

- Can replace noise with Standard Normal $P_{n1} = \Phi(\beta(z_{n1} - z_{n2}))$.

Example: Utility for each alternative depends on attributes of that alternative

- Unobserved terms are assumed to have an extreme value distribution
- With J alternatives

$$\begin{cases} U_{ni} = \beta z_{ni} + \varepsilon_{ni} \\ \varepsilon_{ni} \sim \text{iid extreme value} \end{cases} \quad \Rightarrow \quad P_{ni} = \frac{\exp(\beta z_{ni})}{\sum_{j=1}^{J} \exp(\beta z_{nj})}$$

- Compare to standard model for multiclass classification (multiclass logistic)
- Can also replace noise model with Gaussians

Capturing correlations across alternatives

- All the prior models use the logistic model which does not capture correlations in noise.
- This can be fixed using a joint distribution over the noise e.g.,

$$\begin{cases} U_{ni} = \beta z_{ni} + \varepsilon_{ni} \\ \varepsilon_n \equiv (\varepsilon_{n1}, \cdots, \varepsilon_{nJ}) \sim N(0, \Omega) \end{cases}$$

Estimation

- **Linear case:** maximum likelihood estimators
  - Logistic model: use (binary or multinomial) logistic regression
  - Gaussian Model: use probit regression
- More **complex function classes**: use standard ML fitting tools for (regularized) maximum likelihood, e.g., stochastic gradient descent (SGD)
- Standard tradeoffs, e.g., bias-variance tradeoff.
  - more complex utility models generally require more data
  - Most ML applications pool the model across individuals, individual differences may matter (more on this in future class)

What of measuring ordered preferences?

- Example: On a 1-5 scale where 1 means disagree completely and 5 means agree completely, how much do you agree with the following statement: “I am enjoying this class so far"
- Use ordinal regression, e.g.,

$$U_n = H_n(z_n)$$

$$y_n = \begin{cases} 1, & \text{if } U_n < a \\ 2, & \text{if } a < U_n < b \\ 3, & \text{if } b < U_n < c \\ 4, & \text{if } c < U_n < d \\ 5, & \text{if } U_n > d \end{cases}$$

- For some real numbers a, b, c, d (parameters)

Ordered Logit

- For linear utility: $U_n = \beta z_n + \epsilon, \epsilon \sim$Logistic

$$\Pr(\text{choosing 1}) = \Pr(U_n < a) = \Pr(\varepsilon < a - \beta z_n) = \frac{1}{1 + \exp(-(a - \beta z_n))}$$

$$\Pr(\text{choosing 2}) = \Pr(a < U_n < b) = \Pr(a - \beta z_n < \varepsilon < b - \beta z_n) = \frac{1}{1 + \exp(-(b - \beta z_n))} - \frac{1}{1 + \exp(-(a - \beta z_n))}$$

$$\dots$$

$$\Pr(\text{choosing 5}) = \Pr(U_n > d) = \Pr(\varepsilon > d - \beta z_n) = 1 - \frac{1}{1 + \exp(-(d - \beta z_n))}$$

- Can also replace with Gaussian for ordered probit regression

Plackett-Luce Model

- Ranking models the **sequence of choices** (Plackett and Luce in 1970s)
- Probability of choice $1, 2, \ldots, J$ is

$$\Pr(\text{ranking } 1, 2, \ldots, J) = \frac{\exp(\beta z_1)}{\sum_{j=1}^{J} \exp(\beta z_{nj})} \frac{\exp(\beta z_2)}{\sum_{j=2}^{J} \exp(\beta z_{nj})} \cdots \frac{\exp(\beta z_{J-1})}{\sum_{j=J-1}^{J} \exp(\beta z_{nj})}$$

- PL is common in biomedical literature
- aka rank ordered logit (econometrics ~1980s), or exploded logit model
- All the extensions mentioned also apply (nonlinear utility, correlated noise, …)

Modeling and estimation summary

- Choose the utility model, i.e., how the attributes and alternatives define the utility e.g., linear function of attributes with logistic noise
- Choose the response/observation model, e.g., binary, multiple choice, ordered choice.
- Fit the model using (regularized) maximum likelihood

Aside: “Revealed preference” vs “stated preference”

- **Revealed preference**: Use observed data about the choices to estimate value ascribed to items.
  - Generally offline observational data about real choices
- **Stated Preference**: Use the choices made by individuals made under experimental conditions to estimate these values
  - Generally online experimental data (can include controlled experiments)
- Revealed preference is considered a “real” choice, so can be more accurate
  - In simulated situations, participants may not respond well to hypotheticals
  - OTOH: observed data may not cover the space, hence the appeal of experiments

Exercise (in class): choice model for class(es)

- “Should you take CS 329H or not?”
  - What are the attributes/features (describe what to measure about a class)?
  - What utility model?
  - What is the observation/response model?
  - Revealed preference (observed choices) or stated preference (hypothetical)?
- “Should you take CS 329H or CS 221 or CS 229?”
  - What are the attributes/features?
  - What utility model?
  - What is the observation/response model?
  - Revealed preference or stated preference?

Exercise (in class): choice model for language

- Design a choice model to evaluate the quality of a language model?
  - What utility model?
    - What are the attributes/features?
  - What is the observation/response model?
  - Revealed preference or stated preference?
  - Who should you query?
    - Individual or pooled responses: why or why not?
  - What are some pro/cons of your design?

The ideal point model

- An embedding approach, assumes user item preference depends on distance
  - Let x_n denote a latent vector representing an individual n
  - Let v_i denote a latent vector representing choice (or item) i

$$U_{ni} = dist(x_n, v_i) + \epsilon_{ni}$$

  - Model is equivalent to choosing the “closest” item

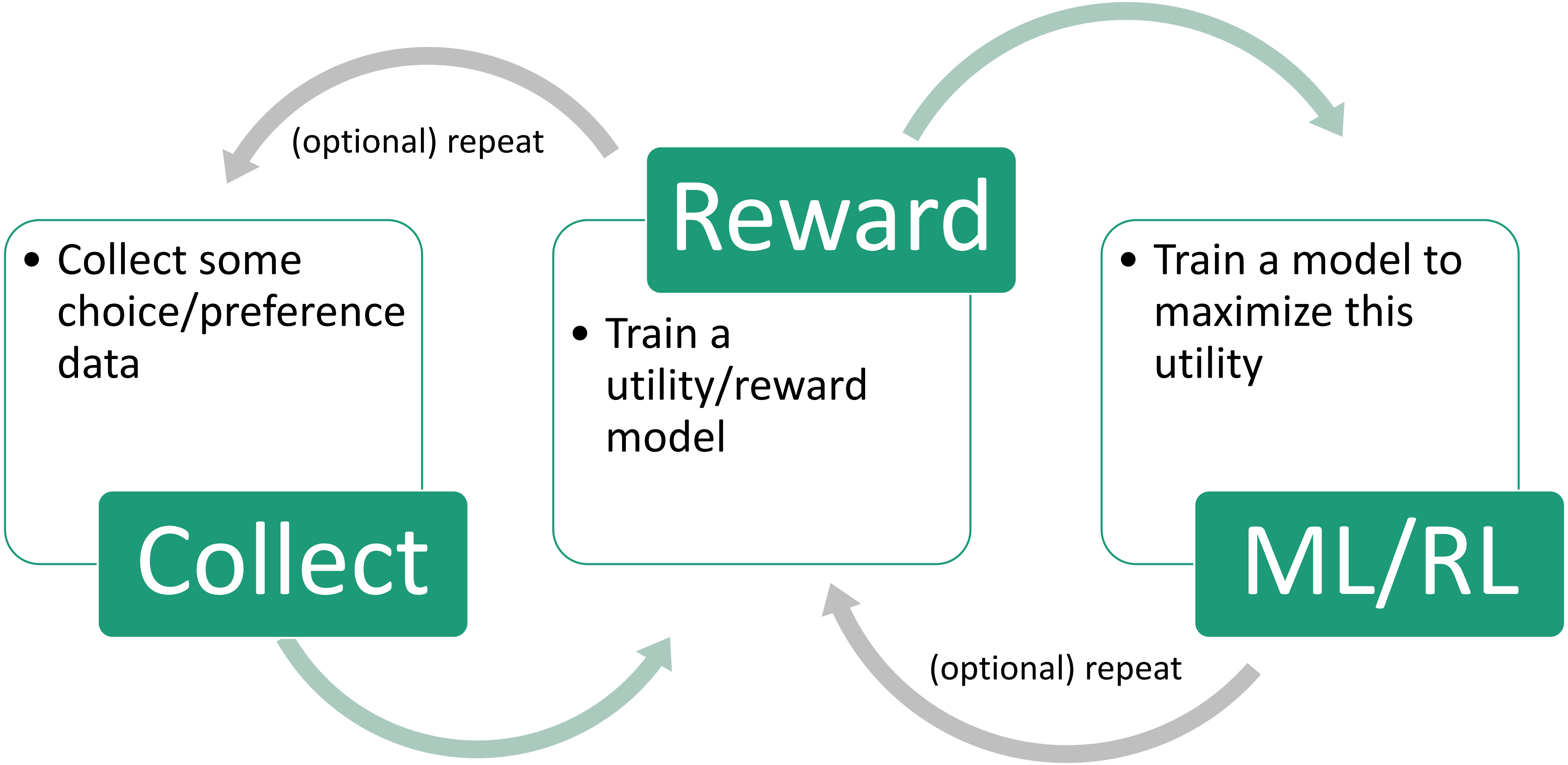
$$y_{ni} = \begin{cases} 1, & \text{if } U_{ni} > U_{nj} \ \forall j \neq i \\ 0, & \text{otherwise} \end{cases}$$

Ideal point model: the why

- Pros:
  - Can sometimes learn preferences faster than attribute-based preference models by exploiting geometry (see refs)
- Cons:
  - Embedding assumption may be strong (can make more flexible via distance function choice)
  - However, have to select a distance function (usually use Euclidian distance in the embedding)

Jamieson, Kevin G., and Robert Nowak. "Active ranking using pairwise comparisons."
Tatli, Gokcan, Rob Nowak, and Ramya Korlakai Vinayak. "Learning Preference Distributions From Distance Measurements."

Choice models in RL (and RLHF)



Collect

- Collect some choice/preference data

Reward

- Train a utility/reward model

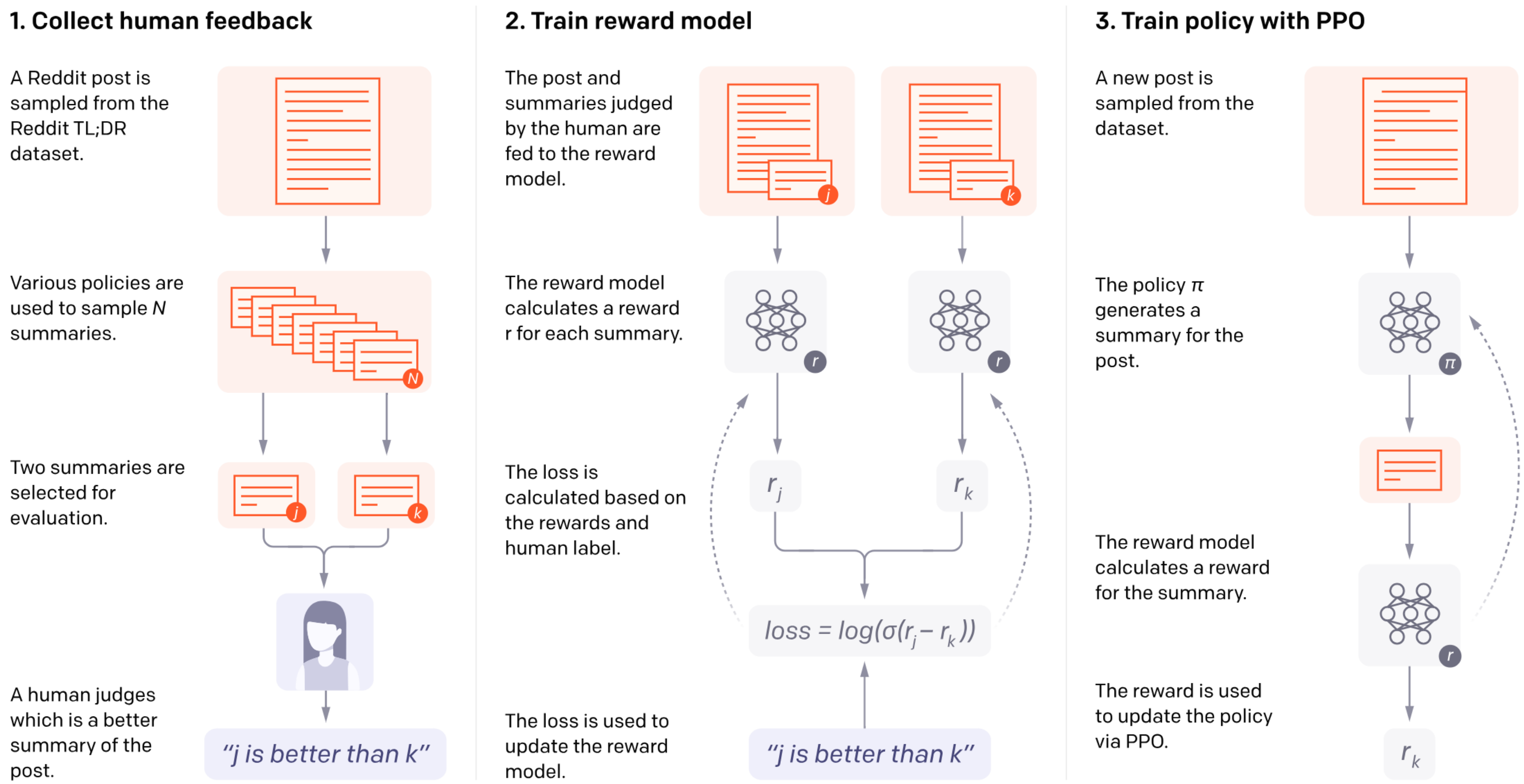
ML/RL

- Train a model to maximize this utility

(optional) repeat

(optional) repeat

Application: RL and Language (Bradley-Terry model)



1. Collect human feedback

A Reddit post is sampled from the Reddit TL;DR dataset.

Various policies are used to sample *N* summaries.

Two summaries are selected for evaluation.

A human judges which is a better summary of the post.

“j is better than k”

2. Train reward model

The post and summaries judged by the human are fed to the reward model.

The reward model calculates a reward r for each summary.

The loss is calculated based on the rewards and human label.

$loss = log(\sigma(r_j - r_k))$

The loss is used to update the reward model.

“j is better than k”

3. Train policy with PPO

A new post is sampled from the dataset.

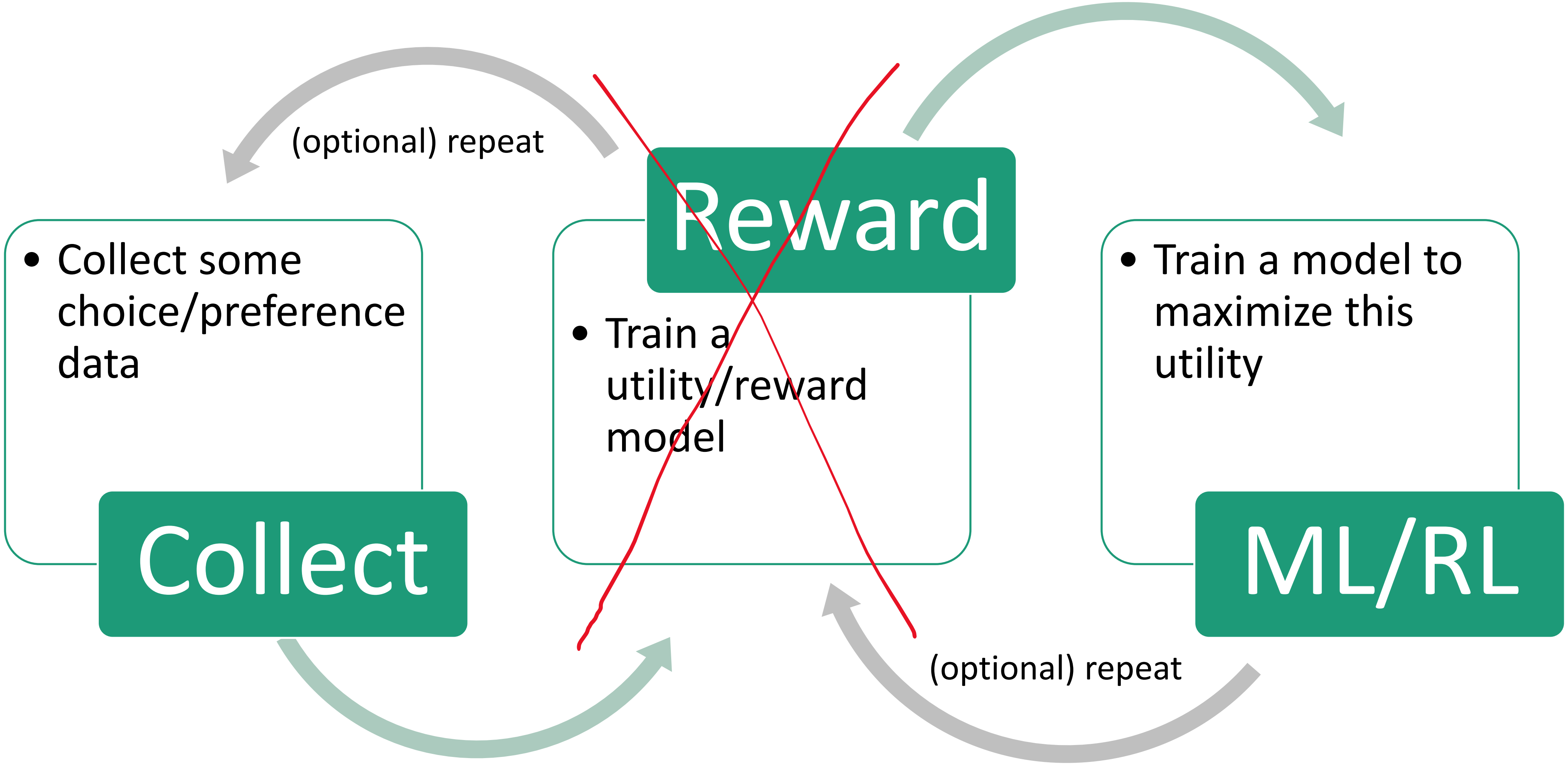
The policy π generates a summary for the post.

The reward model calculates a reward for the summary.

The reward is used to update the policy via PPO.

https://openai.com/research/learning-to-summarize-with-human-feedback

Choice models in ML (recommender systems, bandits, Direct Preference Optimization)



Collect

- Collect some choice/preference data

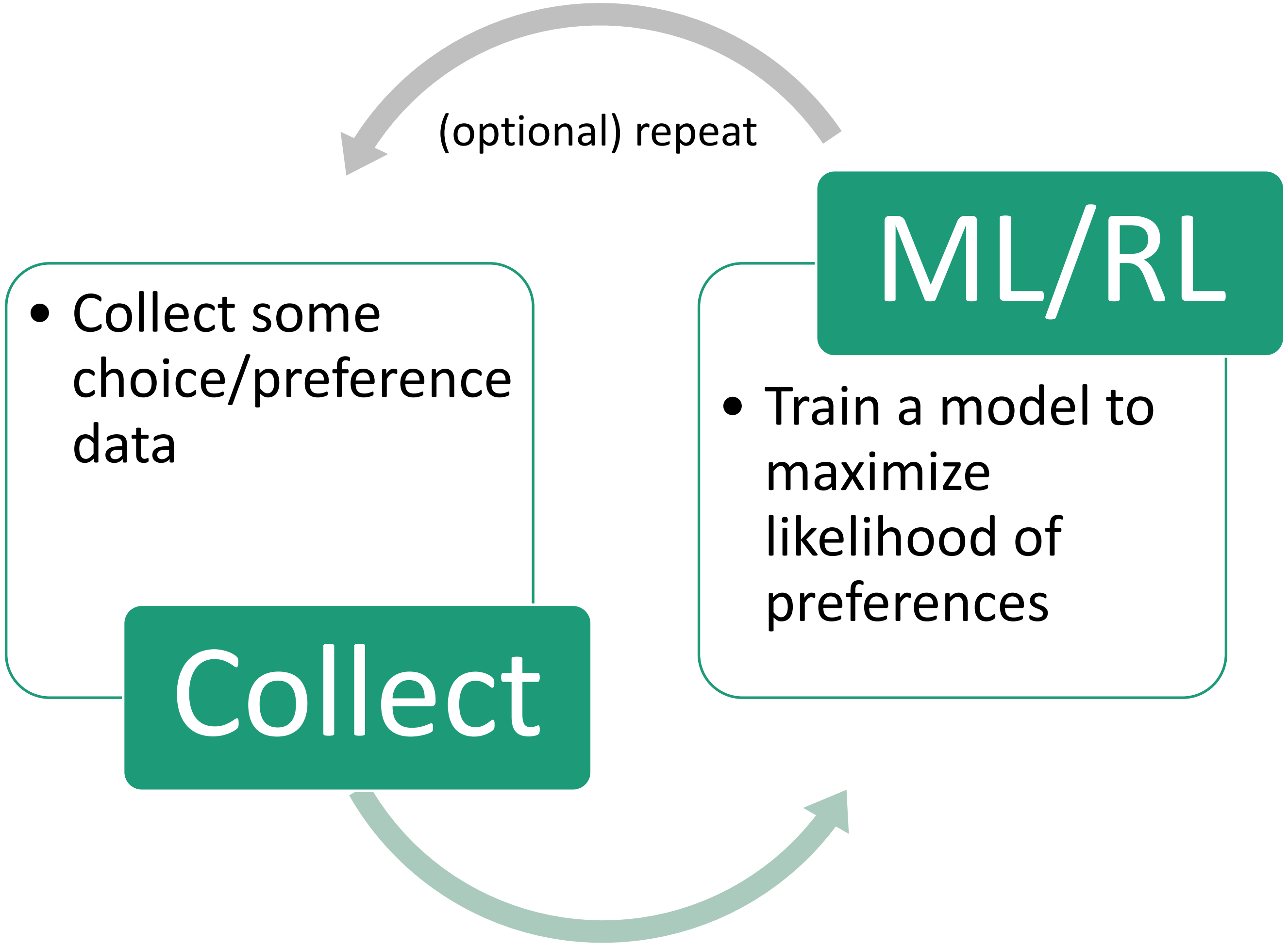
~~Reward~~

- ~~Train a utility/reward model~~

ML/RL

- Train a model to maximize this utility

Choice models in ML (recommender systems, bandits, DPO)



(optional) repeat

Collect

- Collect some choice/preference data

ML/RL

- Train a model to maximize likelihood of preferences

DPO: Bradley Terry model

- Given prompt x and completions y_1 and y_2 the choice model gives the preference

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}$$

- Which corresponds to the loss function (negative log likelihood):

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

- In fine-tuning, we optimize the language model (using RL)

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y \mid x)}\left[r_\phi(x, y)\right] - \beta \mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \mid\mid \pi_{\mathrm{ref}}(y \mid x)\right]$$

- With the solution $r(x, y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} + \beta \log Z(x).$

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. "Direct preference optimization: Your language model is secretly a reward model."

DPO: Bradley Terry model

- For the optimal model (plug in $r^*(x, y)$), this is equivalent to:

$$p^*(y_1 \succ y_2 \mid x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

- Thus, we can improve the language model π_θ by starting with $\pi_{ref}(x, y)$ and optimizing directly

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. "Direct preference optimization: Your language model is secretly a reward model."

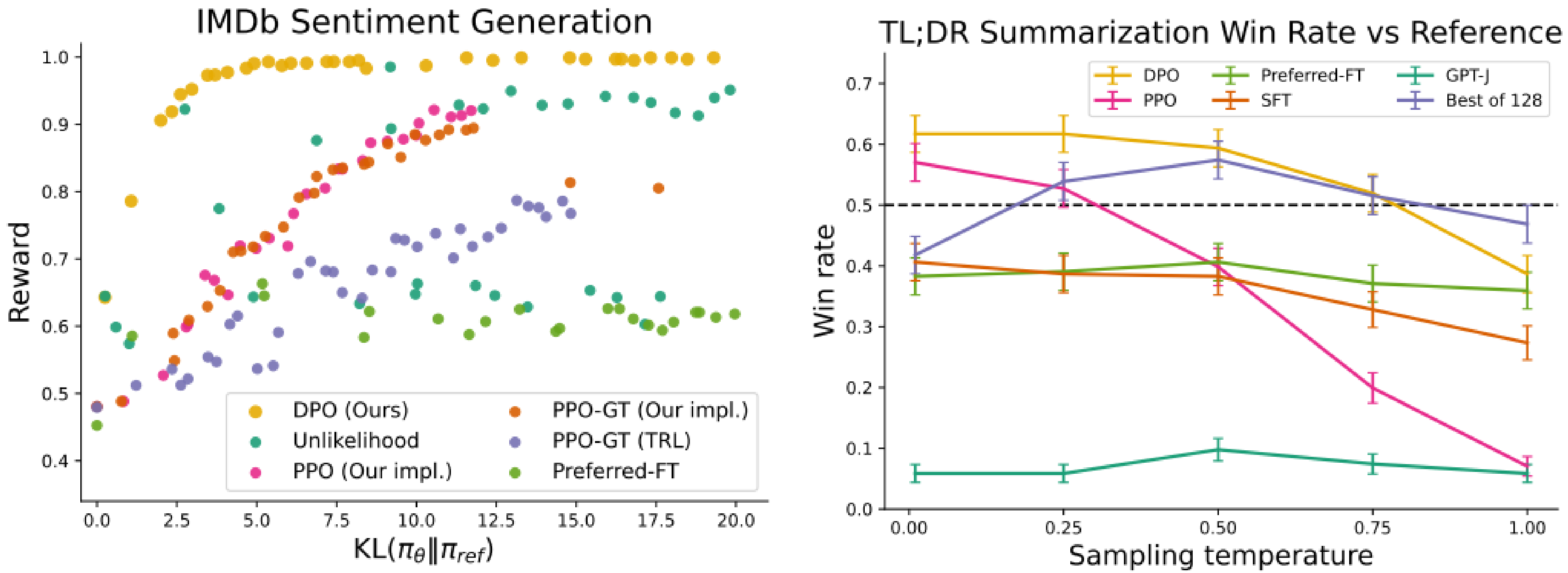


Figure 2: **Left.** The frontier of expected reward vs KL to the reference policy. DPO provides the highest expected reward for all KL values, demonstrating the quality of the optimization. **Right.** TL;DR summarization win rates vs. human-written summaries, using GPT-4 as evaluator. DPO exceeds PPO's best-case performance on summarization, while being more robust to changes in the sampling temperature.

Should your ML application use an explicit utility/reward model?

- Pro:
  - Reward models can be re-used (in principle)
  - Reward model can be examined to infer properties of human(s), and measure the quality of the preference model(s)
  - Reward model(s) add useful inductive biases to the training pipeline
- Cons:
  - The extra step of reward modeling can introduce (unnecessary?) errors
  - Reward model optimization can be unstable (e.g., in RLHF, as argued by DPO)

Some criticisms of choice modeling more broadly

- Real-world choices often appear to be highly situational or context-dependent e.g., way choice is posed, emotional states, other factors not well modeled.
  - Arguably what is exploited by marketing. Related to framing effects (more later).
  - A partial rebuttal: In principle, can always add more context to the model.
- Many choices are intuitive rather than rational, so utility optimization models do not apply
  - Please have limited attention and cognitive capability, especially for less salient choices
  - Default choices are powerful, e.g., in 401K, or opt-in organ donors

Q/A

- What are some key assumptions in (discrete) choice models?
  - Rationality (existence of a utility function that determines choices)
  - Parametric model for utility and choice noise
  - Finite set of choices, and explicit alternatives
- How does one apply discrete choice models to ML/RL applications with changing context (input)
  - Model utility via generic models (e.g., deep neural networks)
- What are some criticisms of discrete choice models?
  - Humans display context-dependent choices
  - Humans often make intuitive (or irrational) choices

What is not covered

- Details of estimation, analysis
    - Maximum likelihood is generally equivalent to standard classification/ranking
    - Existing analysis (though often interesting) is mostly for linear (or simpler) utilities
    - Many of the interesting theoretical questions are for active querying settings
- Beyond discrete choice models
    - with equivalent alternatives ($U_1 > U_2$, $U_1 \approx U_3$)
    - Continuous “choices” e.g., pricing, demand/supply
    - Dynamic discrete choice (for time varying choices) $\approx$ RL.
- Experimental design for “stated preferences”
    - How to design a survey to measure alternatives, conjoint analysis
- Active querying (future discussion)

Summary

- **Today:** Overview of discrete choice models
  - Basics of discrete choice and rationality assumptions
  - Benefits and criticisms of discrete choice
  - Some special cases and applications of discrete choice models to ML
- **Next Lecture:** Preference Models and Reward Models
  - (student lecture, suggested papers are up)

References

- Train, K. (1986). Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand. MIT Press. ISBN 9780262200554. Chapter 8.
- McFadden, D.; Train, K. (2000). "Mixed MNL Models for Discrete Response" (PDF). Journal of Applied Econometrics. 15 (5): 447–470.
- Luce, R. D. (1959). Individual Choice Behavior: A Theoretical Analysis. Wiley.
- Additional:
    - Ben-Akiva, M.; Lerman, S. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand. Transportation Studies. Massachusetts: MIT Press.
    - Park, Byeong U.; Simar, Léopold; Zelenyuk, Valentin (2017). "Nonparametric estimation of dynamic discrete choice models for time series data" (PDF). Computational Statistics & Data Analysis. 108: 97–120. doi:10.1016/j.csda.2016.10.024.
    - Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. "Direct preference optimization: Your language model is secretly a reward model." arXiv preprint arXiv:2305.18290 (2023).

Pre-Course survey response summary

(pulled Oct 1, 2023; evening) ~50% response rate

What is your expectation from the class?

"… in-context learning, online/active/reinforcement learning, and per-user model training."

"… to apply human alignment to AI applications"

"Learn how to effectively use human preference data to update an existing model"

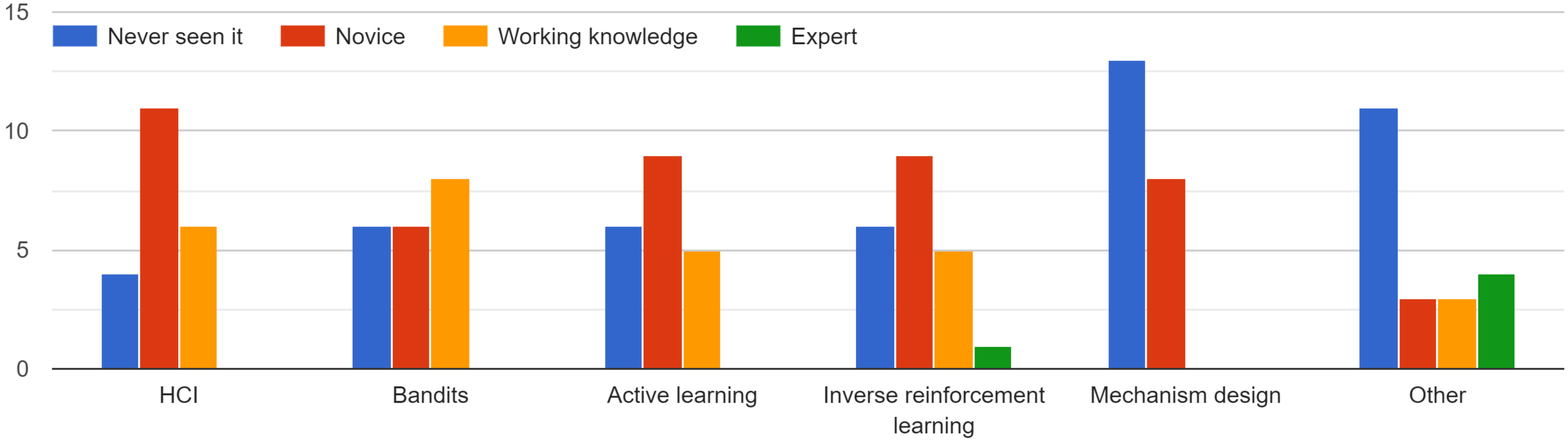
"Starting research…", "… Start from the basics."

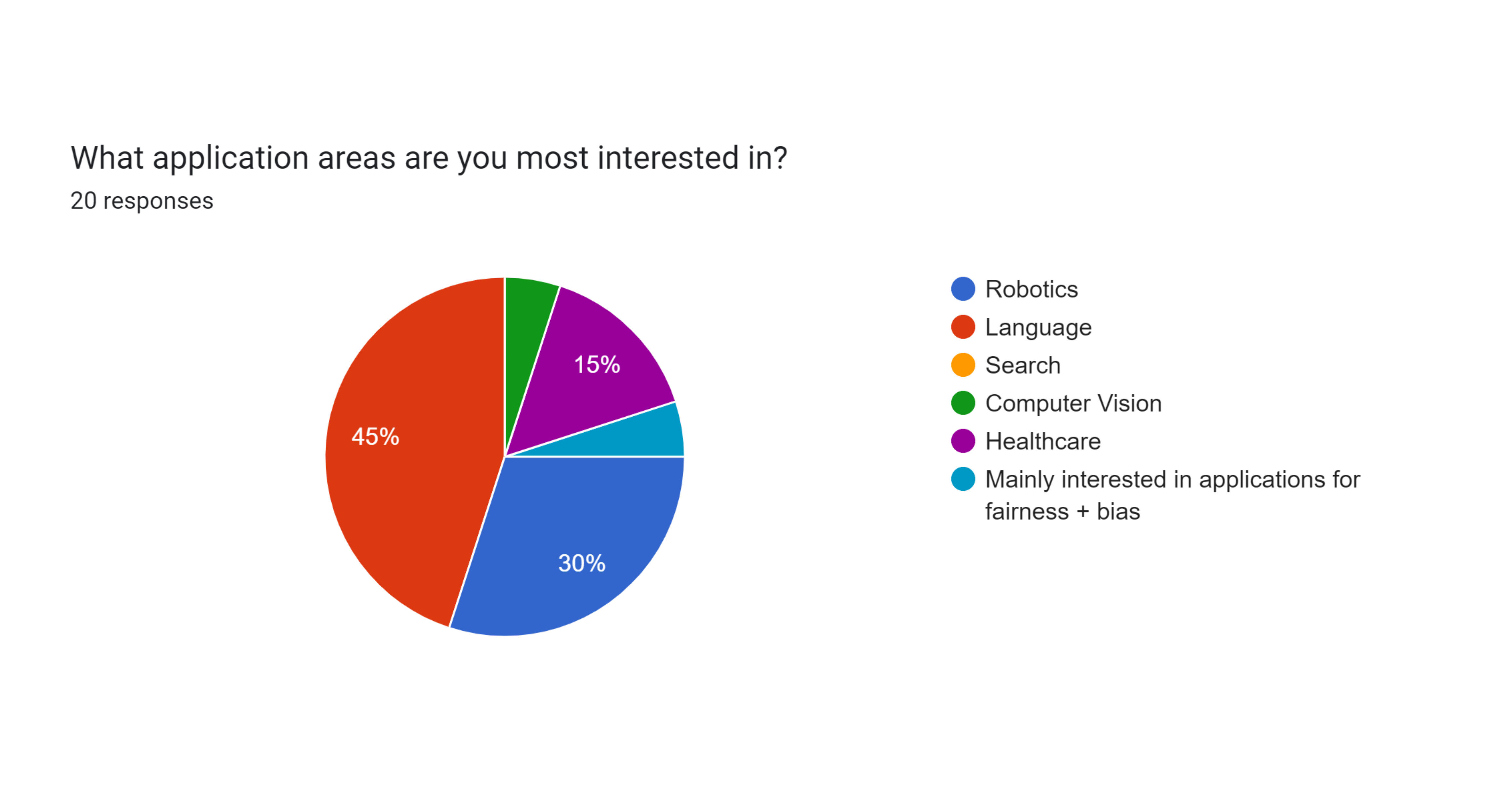
"To learn and discuss the literature"

"mechanism design for eliciting human feedback", "connections to other disciplines (social choice theory, political science, psych, etc.)"

Pre-Course survey response

What is your level of experience with the following topics (never seen it, novice, working knowledge, expert):





What application areas are you most interested in?

20 responses