Discovering Context Effects from Raw Choice Data

ARJUN SESHADRI, *STANFORD UNIVERSITY*
ALEX PEYSAKHOVICH, *FACEBOOK ARTIFICIAL INTELLIGENCE RESEARCH*
JOHAN UGANDER, *STANFORD UNIVERSITY*

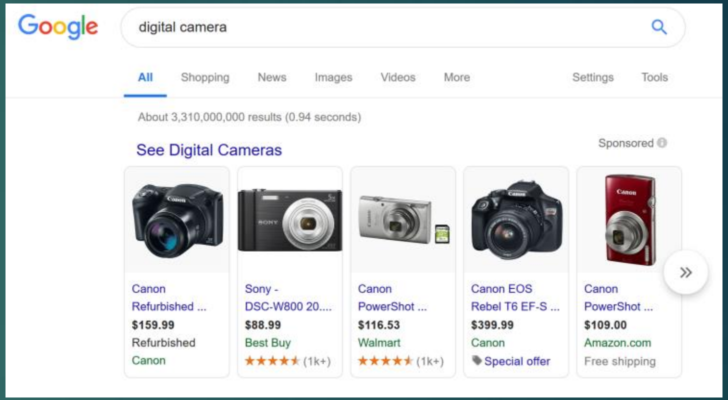
ICML 2019

Modelling in Discrete Choice

- Data of the form (x, C) where "alternative x is chosen from the set C" and C is a subset of $\mathcal{X}$, the universe of n alternatives
- Discrete choice settings are ubiquitous

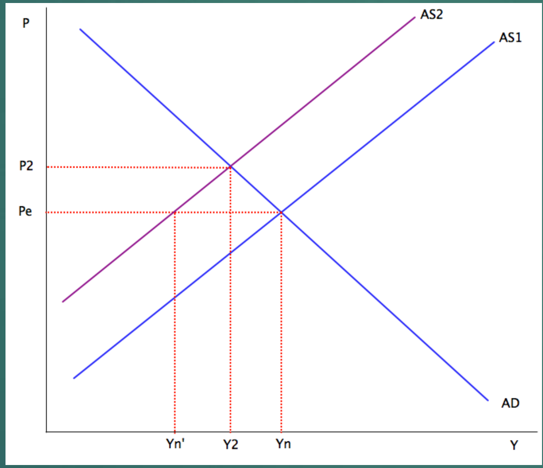
Modelling in Discrete Choice

- Data of the form (x, C) where "alternative x is chosen from the set C" and C is a subset of $\mathcal{X}$, the universe of n alternatives
- Discrete choice settings are ubiquitous



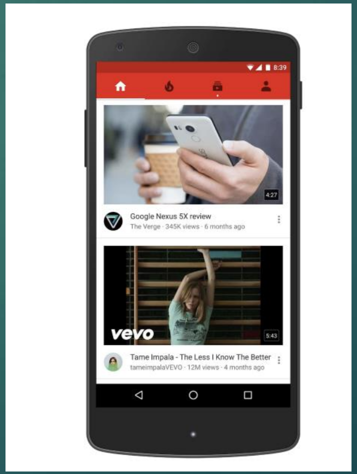
Encompasses Many Fields

Inverse reinforcement learning



Structural Modeling

Virtual Assistants



Recommender Systems

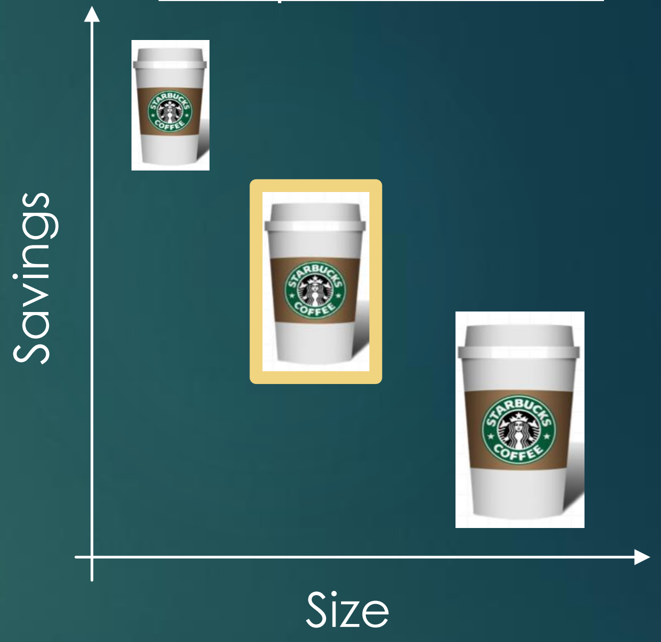
Independence of Irrelevant Alternatives (IIA)

- Fully determines the workhorse Multinomial Logit (MNL) Model
- Main (strong) assumption:

$$\left.\begin{matrix} x, y \in A \\ x, y \in B \end{matrix}\right\} \Rightarrow \frac{\Pr(x \text{ from } A)}{\Pr(y \text{ from } A)} = \frac{\Pr(x \text{ from } B)}{\Pr(y \text{ from } B)}$$

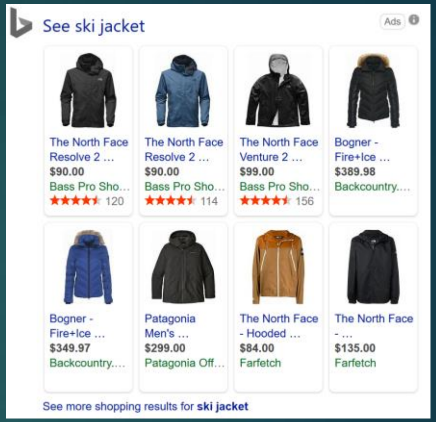
- The **Good**:
  - inferentially tractable, powerful, and interpretable
- The **Bad**:
  - When IIA does not hold, out of sample predictions are wildly miscalibrated
  - Cannot account for the wide literature on context effects (e.g. Compromise Effect)

Compromise Effect



Problems we address

- Modelling individual choice behavior
  - Behavioral economics “anomalies” are all over the place
    - Search Engine Ads (Ieong-Mishra-Sheffet ’12, Yin et al. ’14)
    - Google Web Browsing Choices (Benson-Kumar-Tomkins ’16)
  - Need to model while retaining parametric and inferential efficiency



“ad group quality”

- Statistical tests for violations of IIA
  - General, global tests are intractable (Seshadri & Ugander ‘19, Long & Freese ‘05)
  - Model based approaches challenging due to identifiability issues (Cheng & Long, ‘07)

Context Dependent Utility Model (CDM)

$$P(x \mid C) = \frac{\exp(u(x \mid C))}{\sum_{y \in C} \exp(u(y \mid C))}.$$

Universal logit model (McFadden et al., '77)

Context Dependent Utility Model (CDM)

$$P(x \mid C) = \frac{\exp(u(x \mid C))}{\sum_{y \in C} \exp(u(y \mid C))}.$$

Universal logit model (McFadden et al., '77)

Decompose the model (Batsell & Polking, '85)

$$u(x \mid C) = \underbrace{v(x)}_{\text{1st order}} + \underbrace{\sum_{y \in C \setminus x} v(x \mid \{y\})}_{\text{2nd order}} + \underbrace{\sum_{\{y,z\} \subseteq C \setminus x} v(x \mid \{y, z\})}_{\text{3rd order}} + \ldots + \underbrace{v(x \mid C \setminus \{x\})}_{|C|\text{th order}},$$

Developing the CDM

Context Dependent Utility Model (CDM)

$$P(x \mid C) = \frac{\exp(u(x \mid C))}{\sum_{y \in C} \exp(u(y \mid C))}.$$

Universal logit model (McFadden et al., ’77)

Decompose the model (Batsell & Polking, ’85)

$$u(x \mid C) = \underbrace{v(x)}_{\text{1st order}} + \underbrace{\sum_{y \in C \setminus x} v(x \mid \{y\})}_{\text{2nd order}} + \underbrace{\sum_{\{y,z\} \subseteq C \setminus x} v(x \mid \{y,z\})}_{\text{3rd order}} + \ldots + \underbrace{v(x \mid C \setminus \{x\})}_{|C|\text{th order}},$$

Truncate to 2nd order (effects are pairwise)

$$P(x \mid C) = \frac{\exp(\sum_{z \in C \setminus x} u_{xz})}{\sum_{y \in C} \exp(\sum_{z \in C \setminus y} u_{yz})}.$$

Full Rank CDM

Developing the CDM

Context Dependent Utility Model (CDM)

$$P(x \mid C) = \frac{\exp(u(x \mid C))}{\sum_{y \in C} \exp(u(y \mid C))}.$$

Universal logit model (McFadden et al., '77)

Decompose the model (Batsell & Polking, '85)

$$u(x \mid C) = \underbrace{v(x)}_{\text{1st order}} + \underbrace{\sum_{y \in C \setminus x} v(x \mid \{y\})}_{\text{2nd order}} + \underbrace{\sum_{\{y,z\} \subseteq C \setminus x} v(x \mid \{y, z\})}_{\text{3rd order}} + \ldots + \underbrace{v(x \mid C \setminus \{x\})}_{|C|\text{th order}},$$

Truncate to 2nd order (effects are pairwise)

$$P(x \mid C) = \frac{\exp(\sum_{z \in C \setminus x} u_{xz})}{\sum_{y \in C} \exp(\sum_{z \in C \setminus y} u_{yz})}.$$

Full Rank CDM

Make a low rank approximation (parameters linear in items)

$$P(x \mid C) = \frac{\exp((\sum_{z \in C \setminus x} c_z)^T t_x)}{\sum_{y \in C} \exp((\sum_{z \in C \setminus y} c_z)^T t_y)}.$$

Low Rank CDM

Developing the CDM

Context Dependent Utility Model (CDM)

$$P(x \mid C) = \frac{\exp(u(x \mid C))}{\sum_{y \in C} \exp(u(y \mid C))}.$$

Universal logit model (McFadden et al., ’77)

Decompose the model (Batsell & Polking, ’85)

$$u(x \mid C) = \underbrace{v(x)}_{\text{1st order}} + \underbrace{\sum_{y \in C \setminus x} v(x \mid \{y\})}_{\text{2nd order}} + \underbrace{\sum_{\{y,z\} \subseteq C \setminus x} v(x \mid \{y, z\})}_{\text{3rd order}} + \ldots + \underbrace{v(x \mid C \setminus \{x\})}_{|C|\text{th order}},$$

Truncate to 2nd order (effects are pairwise)

$$P(x \mid C) = \frac{\exp(\sum_{z \in C \setminus x} u_{xz})}{\sum_{y \in C} \exp(\sum_{z \in C \setminus y} u_{yz})}.$$

Full Rank CDM

Make a low rank approximation (parameters linear in items)

Other items change how features are traded off

$$P(x \mid C) = \frac{\exp((\sum_{z \in C \setminus x} c_z)^T t_x)}{\sum_{y \in C} \exp((\sum_{z \in C \setminus y} c_z)^T t_y)}.$$

r-dimensional latent feature vector
r << n items

Low Rank CDM

Developing the CDM

A Theoretical Preview

A Theoretical Preview

Identifiability

Sufficient:

Theorem. *A CDM is identifiable from a dataset $\mathcal{D}$ if $\mathcal{C}_{\mathcal{D}}$ contains comparisons over all choice sets of two sizes k, k', where at least one of k, k' is not 2 or n.*

Necessary:

Theorem. *No rank r CDM, $1 \leq r \leq n$, is identifiable from a dataset $\mathcal{D}$ if $\mathcal{C}_{\mathcal{D}}$ contains only choices from sets of a single size.*

More generally:

Theorem. *A full rank CDM is identifiable from a dataset $\mathcal{D}$ if and only if the rank of an integer design matrix $G(\mathcal{D})$, properly constructed, is $n(n-1)-1$.*

A Theoretical Preview

Identifiability

Sufficient:

Theorem. *A CDM is identifiable from a dataset $\mathcal{D}$ if $\mathcal{C}_{\mathcal{D}}$ contains comparisons over all choice sets of two sizes k, k', where at least one of k, k' is not 2 or n.*

Necessary:

Theorem. *No rank r CDM, $1 \leq r \leq n$, is identifiable from a dataset $\mathcal{D}$ if $\mathcal{C}_{\mathcal{D}}$ contains only choices from sets of a single size.*

More generally:

Theorem. *A full rank CDM is identifiable from a dataset $\mathcal{D}$ if and only if the rank of an integer design matrix $G(\mathcal{D})$, properly constructed, is $n(n-1)-1$.*

Convergence Guarantees

$$\mathbb{E}\left[\, \|\hat{u}_{\mathrm{MLE}}(\mathcal{D}) - u^\star\|_2^2 \,\right] \leq c_{B,k_{\max}} \frac{n(n-1)}{m},$$

where the expectation is taken over the dataset $\mathcal{D}$ containing m samples, where $k_{\max}$ refers to the maximum choice set size in the dataset, and $c_{B,k_{\max}}$ is a constant that depends on the structure of the design matrix $G(\mathcal{D})$.

A Theoretical Preview

Identifiability

Sufficient:

Theorem. *A CDM is identifiable from a dataset $\mathcal{D}$ if $\mathcal{C}_\mathcal{D}$ contains comparisons over all choice sets of two sizes k, k', where at least one of k, k' is not 2 or n.*

Necessary:

Theorem. *No rank r CDM, $1 \leq r \leq n$, is identifiable from a dataset $\mathcal{D}$ if $\mathcal{C}_\mathcal{D}$ contains only choices from sets of a single size.*

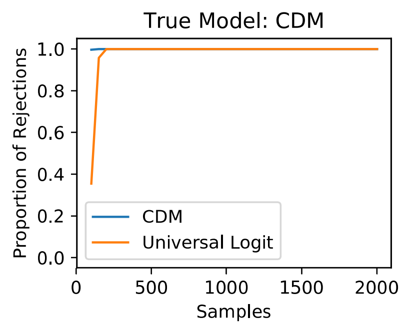
More generally:

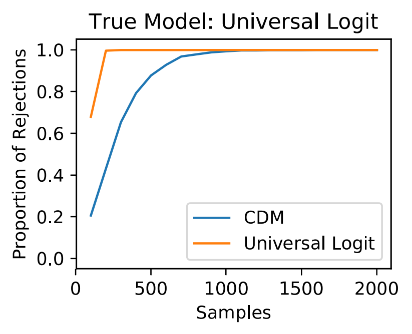
Theorem. *A full rank CDM is identifiable from a dataset $\mathcal{D}$ if and only if the rank of an integer design matrix $G(\mathcal{D})$, properly constructed, is $n(n-1)-1$.*

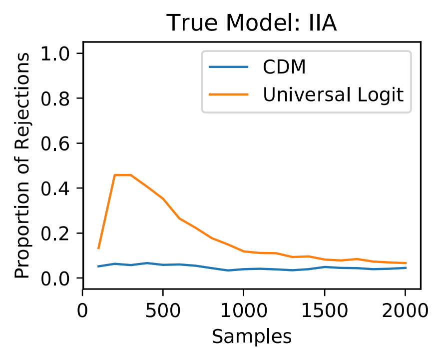
Hypothesis Testing

$$\Lambda(\mathcal{D}) = \frac{\sup_{\theta \in \Theta_{\text{Luce}} \subset \Theta_{\text{CDM}}} \mathcal{L}(\mathcal{D} \mid \theta)}{\sup_{\theta \in \Theta_{\text{CDM}}} \mathcal{L}(\mathcal{D} \mid \theta)},$$

a Likelihood Ratio Statistic where Θ_{Luce} and Θ_{CDM} respectively refer to the parameter classes of Luce and CDM Models.







Convergence Guarantees

$$\mathbb{E}\left[\, \|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^\star\|_2^2 \,\right] \leq c_{B,k_{\max}} \frac{n(n-1)}{m},$$

where the expectation is taken over the dataset $\mathcal{D}$ containing m samples, where $k_{\max}$ refers to the maximum choice set size in the dataset, and $c_{B,k_{\max}}$ is a constant that depends on the structure of the design matrix $G(\mathcal{D})$.

Unifying Existing Choice Models

Low Rank CDM

$$P(x \mid C) = \frac{\exp((\sum_{z \in C \setminus x} c_z)^T t_x)}{\sum_{y \in C} \exp((\sum_{z \in C \setminus y} c_z)^T t_y)}.$$

Unifying Existing Choice Models

Tversky-Simonson Model

$$u^{TS}(x \mid C) = w(C)^T t_x$$

(Tversky & Simonson, 1993)

Low Rank CDM

$$P(x \mid C) = \frac{\exp((\sum_{z \in C \setminus x} c_z)^T t_x)}{\sum_{y \in C} \exp((\sum_{z \in C \setminus y} c_z)^T t_y)}.$$

Unifying Existing Choice Models

Tversky-Simonson Model

$$u^{TS}(x \mid C) = w(C)^T t_x$$

(Tversky & Simonson, 1993)

Low Rank CDM

$$P(x \mid C) = \frac{\exp((\sum_{z \in C \setminus x} c_z)^T t_x)}{\sum_{y \in C} \exp((\sum_{z \in C \setminus y} c_z)^T t_y)}.$$

Batsell-Polking Model

$$\log \frac{\Pr(x \mid C)}{\Pr(y \mid C)} = \alpha_{xy} + \sum_{z \in C \setminus \{x,y\}} \alpha_{xz}$$

(Batsell & Polking, 1985)

Unifying Existing Choice Models

Tversky-Simonson Model

$$u^{TS}(x \mid C) = w(C)^T t_x$$

(Tversky & Simonson, 1993)

Low Rank CDM

$$P(x \mid C) = \frac{\exp((\sum_{z \in C \setminus x} c_z)^T t_x)}{\sum_{y \in C} \exp((\sum_{z \in C \setminus y} c_z)^T t_y)}.$$

Batsell-Polking Model

$$\log \frac{\Pr(x \mid C)}{\Pr(y \mid C)} = \alpha_{xy} + \sum_{z \in C \setminus \{x,y\}} \alpha_{xz}$$

(Batsell & Polking, 1985)

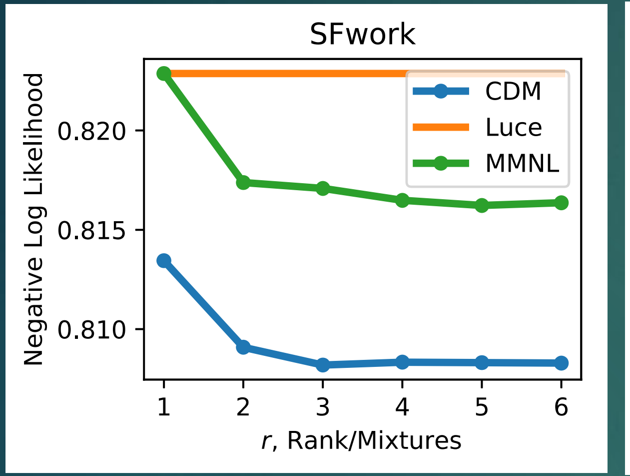
Blade-Chest Model

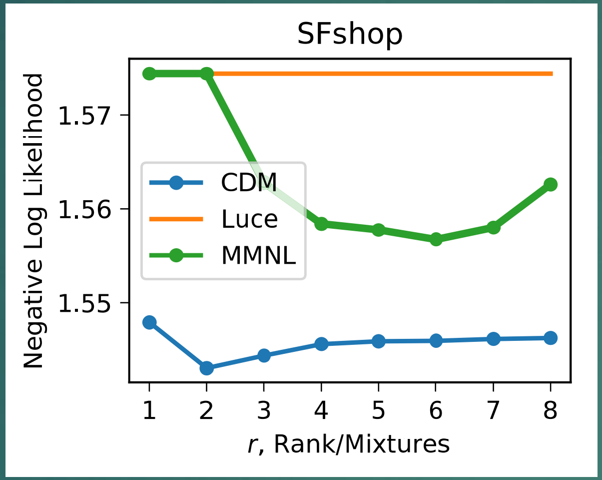
$$\Pr(x \mid \{x, y\}) = \frac{\exp(t_x^T c_y)}{\exp(t_x^T c_y) + \exp(t_y^T c_x)}$$

(Chen & Joachims, 2016)

An Empirical Preview: Performance and Interpretability

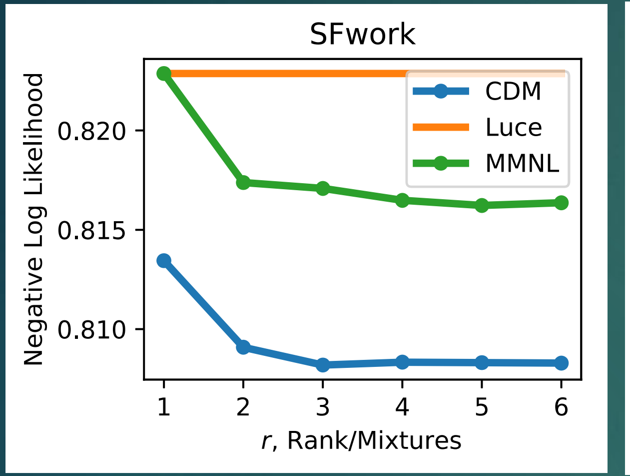
An Empirical Preview: Performance and Interpretability

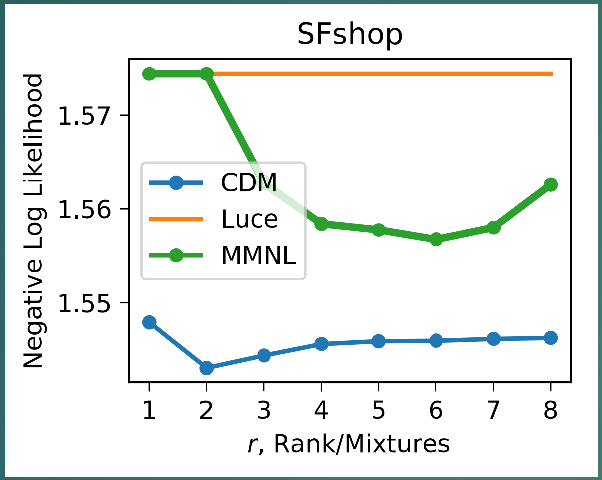




- **Transportation Preferences** (Koppelman & Bhat, '06)
  - Survey of transportation choices for residents in various San Francisco neighborhoods
  - Low Rank CDMs significantly outperform MNL and MMNL

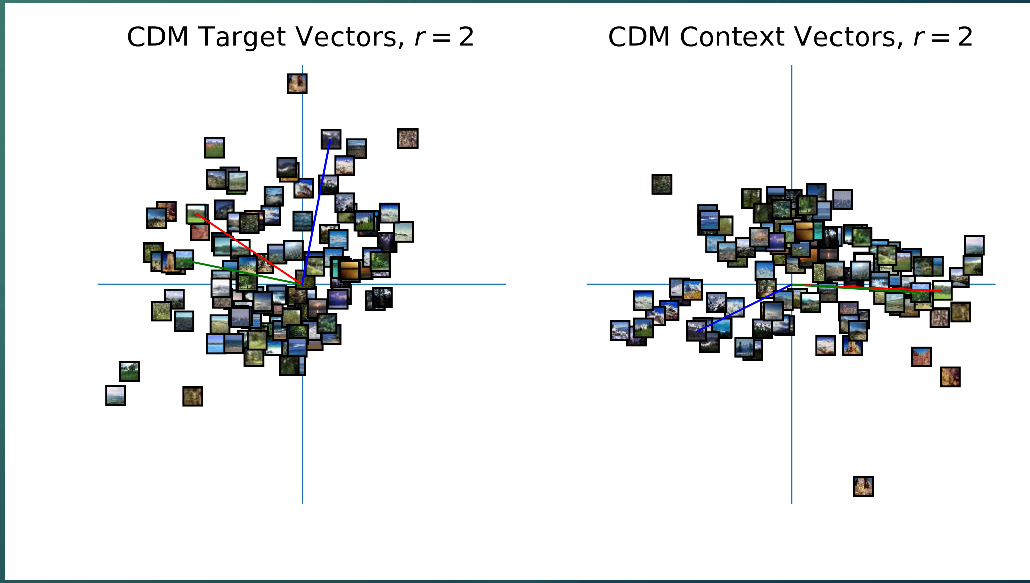
An Empirical Preview: Performance and Interpretability





- **Transportation Preferences** (Koppelman & Bhat, '06)
  - Survey of transportation choices for residents in various San Francisco neighborhoods
  - Low Rank CDMs significantly outperform MNL and MMNL

- **Not Like the Other** (Heikinheimo & Ukkonen, '13)
  - Individuals are shown triplets of nature photographs
  - asked to choose photo most unlike the other two
  - CDM illustrates intuitive property of dataset: similar items have negative target-context inner product
    - Induces grouping by similarity in both target and context vectors



Conclusions

- CDM models context effects with efficiency guarantees and enables practical tests of IIA
- Can be easily applied to many pipelines by modifying "the final layer"
- Simultaneously brings both:
  - Machine Learning rigor to Econometrics models (identifiability, convergence)
  - Econometrics modeling (choice set effects) into Machine Learning research

Thanks!!

Discovering Context Effects from Raw Choice Data

Arjun Seshadri, Alex Peysakhovich, and Johan Ugander

Poster: **Pacific Ballroom #234**