SIGCHI

Collaborative Prediction and Ranking with Non-Random Missing Data

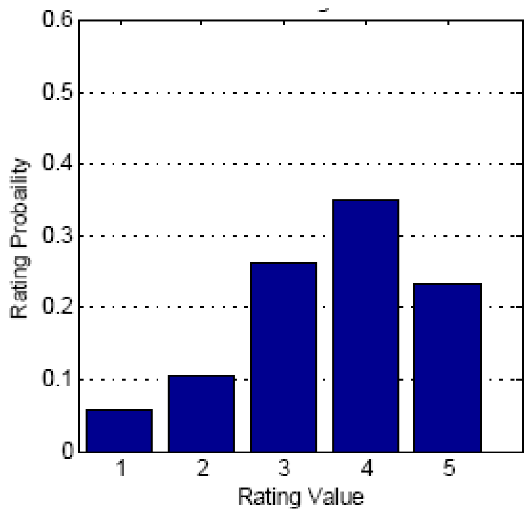
Benjamin Marlin
Department of Computer Science
University of British Columbia

Richard Zemel
Department of Computer Science
University of Toronto

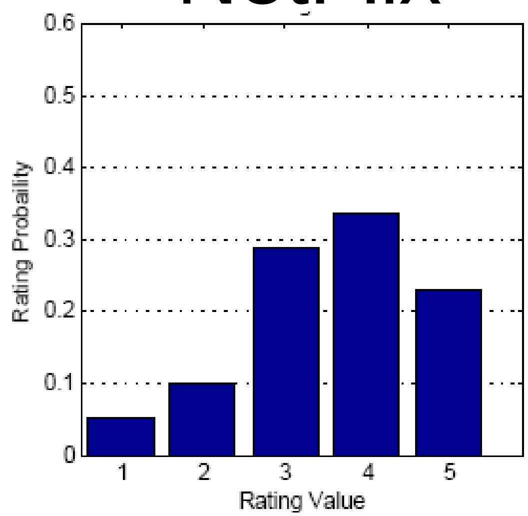
Introduction: *Observation and Question*

Observation: Many rating data sets exhibit marginal rating distributions that are skewed toward high rating values.

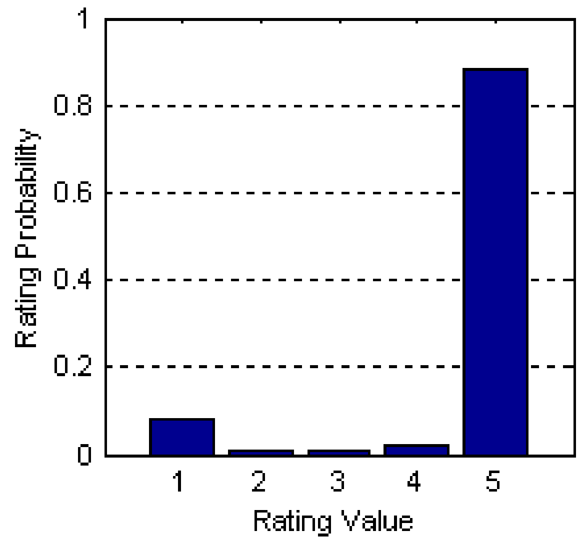
MovieLens



NetFlix

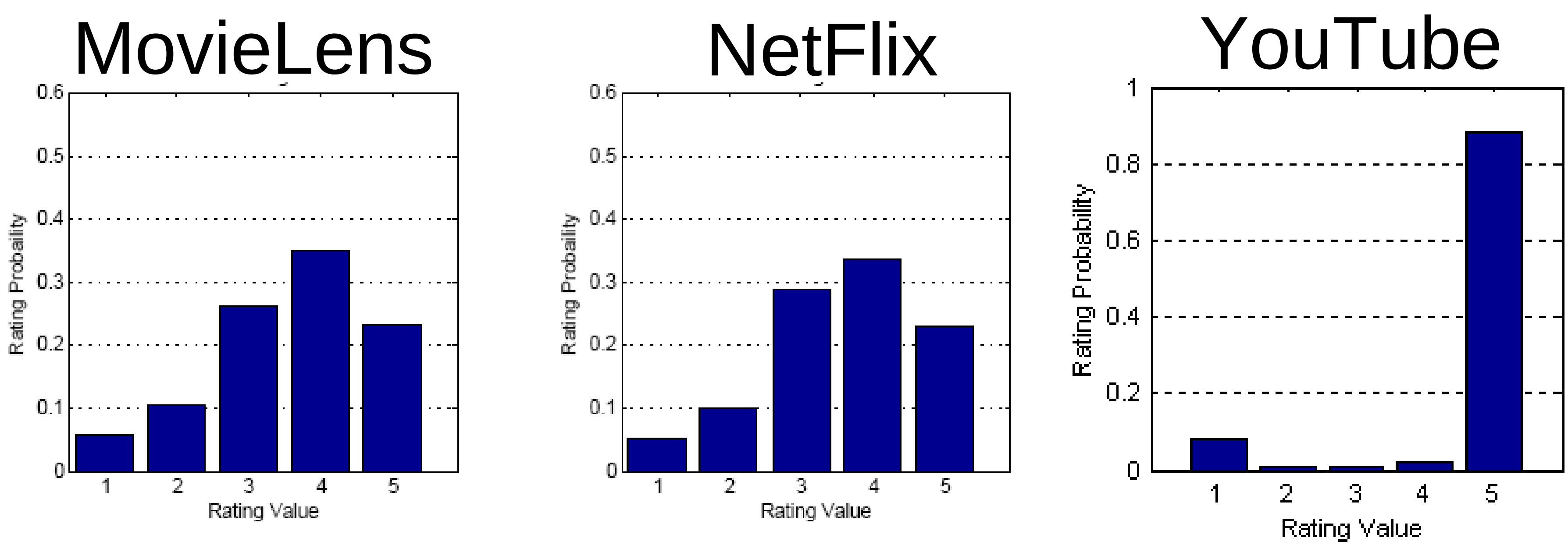


YouTube



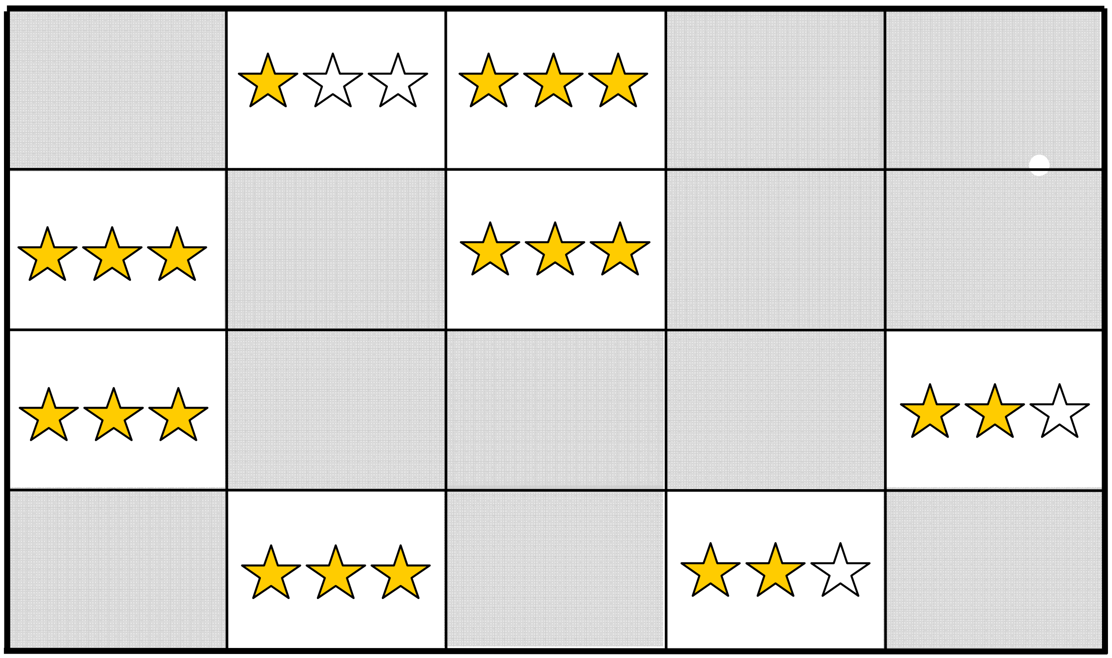
Introduction: *Observation and Question*

Question: What causes these skewed distributions?



Introduction: *Observation Processes*

Answer 1: Most people really do like most items in these data sets, and we observe a **random** sample of entries.

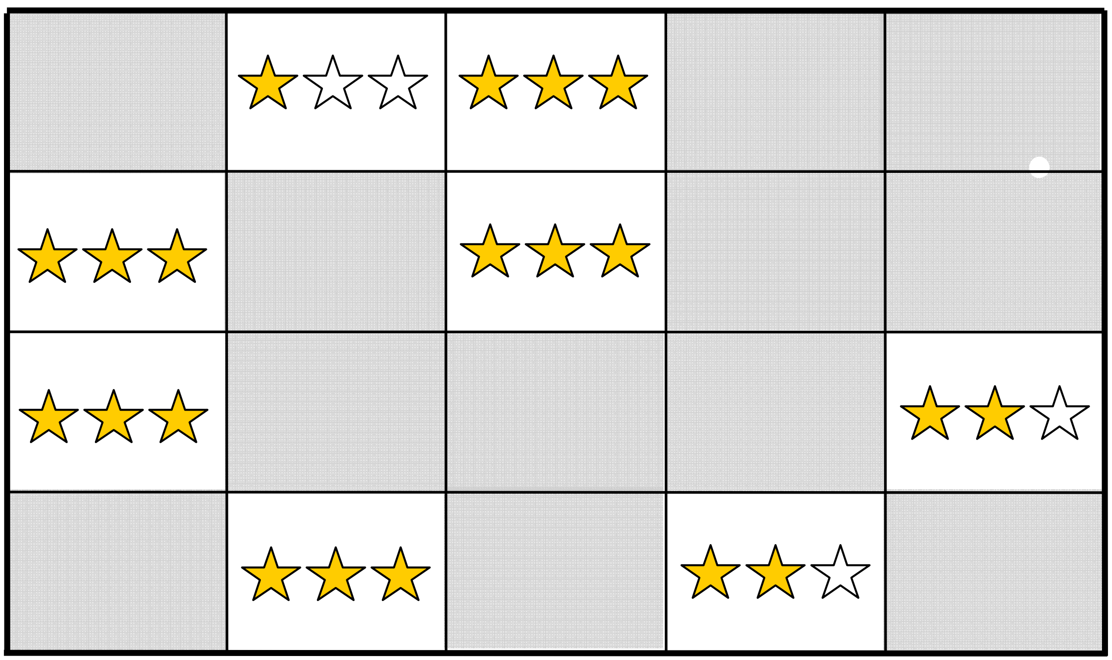


Introduction: *Observation Processes*

Answer 1: Most people really do like most items in these data sets, and we observe a **random** sample of entries.

Introduction: *Observation Processes*

Answer 2: Most people don't really like most items, but we observe a **non-random** sample where people tend to rate items they like.



Introduction: *Observation Processes*

Answer 2: Most people don't really like most items, but we observe a **non-random** sample where people tend to rate items they like.

Introduction: *Observation Processes*

My Goals for this Talk:

1. Convince you that answer #2 is the more likely answer in recommender systems.
2. Explore the implications of a non-random observation process.
3. Provide methods that can learn under a non-random observation process.
4. **Suggest future research directions.**

Talk Outline:

- Introduction
- **Missing Data Theory and Implications**
- Yahoo! LaunchCast Study
- Models and Algorithms
- Experiments and Results

Missing Data Theory: Notation

$$X^{obs}$$

	1	2	3	…	M
1	★★☆			★★☆	★★☆
2		★★☆	★☆☆		★★★
⋮	★★★		★☆☆	★★★	
N	★★☆	★★★			★★★

Observed Data Values

$$R$$

	1	2	3	…	M
1	1	0	0	1	1
2	0	1	1	0	1
⋮	1	0	1	1	0
N	1	1	0	0	1

Response Indicators

Missing Data Theory: *Notation*

X

	1	2	3	…	M
1	★★☆	★★☆	★★★	★★☆	★★☆
2	★★★	★★☆	★☆☆	★☆☆	★★★
⋮	★★★	★★☆	★☆☆	★★★	★★☆
N	★★☆	★★★	★☆☆	★☆☆	★★★

All Data Values

R

	1	2	3	…	M
1	1	0	0	1	1
2	0	1	1	0	1
⋮	1	0	1	1	0
N	1	1	0	0	1

Response Indicators

Missing Data Theory: Notation

X^{mis}

	1	2	3	…	M
1	★★☆	★★☆	★★★	★★☆	★★☆
2	★★★	★★☆	★☆☆	★☆☆	★★★
⋮	★★★	★★☆	★☆☆	★★★	★★☆
N	★★☆	★★★	★☆☆	★☆☆	★★★

Missing Data Values

R

	1	2	3	…	M
1	1	0	0	1	1
2	0	1	1	0	1
⋮	1	0	1	1	0
N	1	1	0	0	1

Response Indicators

Missing Data Theory: *Processes*

Data Model and Observation Model:

$$P(\mathbf{X},\mathbf{R}|\theta,\mu) = P(\mathbf{R}|\mathbf{X},\mu)P(\mathbf{X},|\theta)$$

Missing at Random Condition:

$$P(\mathbf{R}|\mathbf{X},\mu) = P(\mathbf{R}|\mathbf{X}^{obs},\mu)$$

- Violated if probability that user **u** will rate item **i** depends on user **u**'s rating for item **i**.

R. J. A. Little and D. B. Rubin. Statistical Analysis with Missing Data. 1987.

Missing Data Theory: Learning

- The **MAR** assumption is the justification for ignoring missing data during learning:

$$
\begin{aligned}
L_{mar}(\theta|\mathbf{x}^{obs},\mathbf{r}) &= \int P(\mathbf{R}|\mathbf{X},\mu)P(\mathbf{X}|\theta)d\mathbf{X}^{mis} \\
&= P(\mathbf{R}|\mathbf{X}^{obs},\mu)\int P(\mathbf{X}|\theta)d\mathbf{X}^{mis} \\
&= P(\mathbf{R}|\mathbf{X}^{obs},\mu)P(\mathbf{X}^{obs}|\theta) \\
&\propto P(\mathbf{X}^{obs}|\theta)
\end{aligned}
$$

Missing Data Theory: Learning

- When **MAR** does not hold, the likelihood does not simplify:

$$L_{mar}(\theta|\mathbf{x}^{obs},\mathbf{r}) \;=\; \int P(\mathbf{R}|\mathbf{X},\mu)P(\mathbf{X}|\theta)d\mathbf{X}^{mis}$$

- Ignoring missing data is equivalent to using the wrong likelihood function. Parameter estimates will be “biased”.
- **One Solution: Explicitly model P(R|X,μ) and P(X|θ). Estimate μ and θ.**

Missing Data Theory: *Testing*

- Training and testing on ratings of user-selected items will not reveal any difficulties.
- Complimentary "biases" in training and testing cancel out.
- **One Solution: Collect a test set of ratings for randomly selected items and use it to test methods.**

Talk Outline:

- Introduction
- Missing Data Theory and Implications
- **Yahoo! LaunchCast Study**
- Models and Algorithms
- Experiments and Results

Yahoo! Study: *Data Collection*

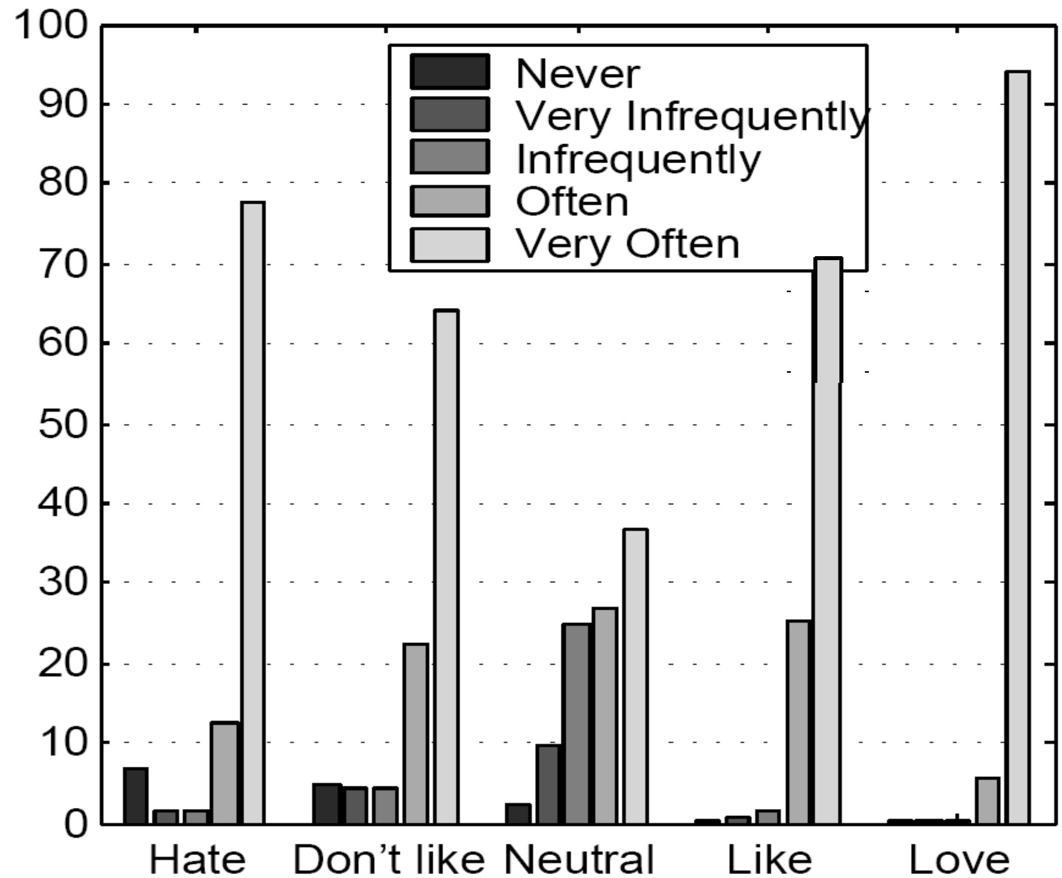
Data was collected through an online survey of Yahoo! Music LaunchCast radio users.

- 1000 songs selected at random.
- Users rate 10 songs selected at random from 1000 songs.
- Data from 5000 users.

Yahoo! Study: Survey Questions

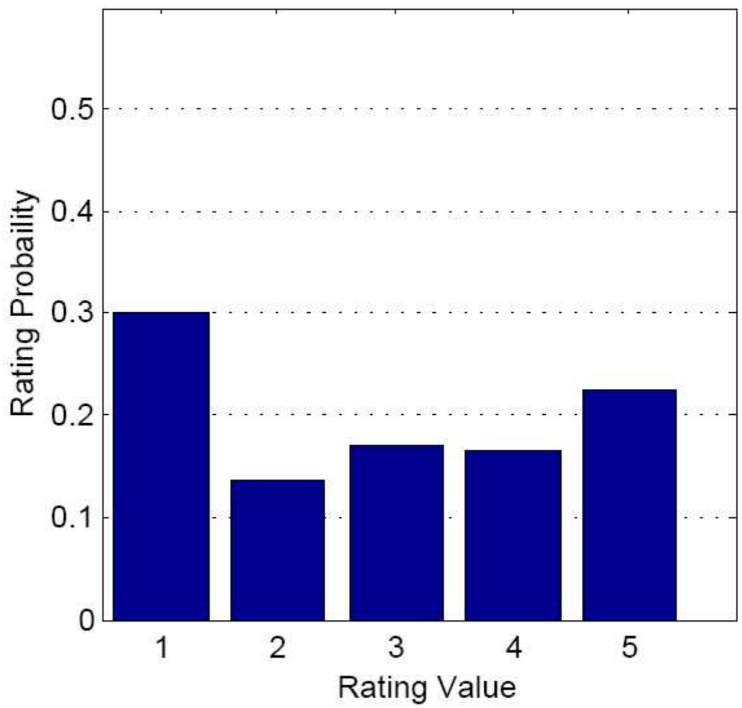
Do preferences impact choice to rate?

64.85% of users reported that their preferences **do** impact their choice to rate an item.

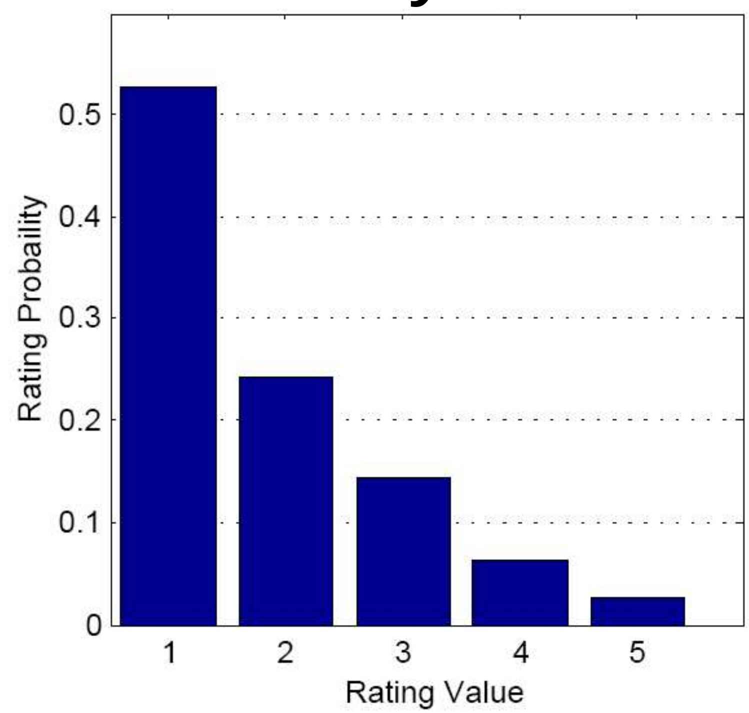


Yahoo! Study: Rating Distributions

User Selected



Randomly Selected



Talk Outline:

- Introduction
- Missing Data Theory and Implications
- Yahoo! LaunchCast Study
- **Models and Algorithms**
- Experiments and Results

Models: *Finite Mixture/CPT-v*

Probability Model:

$$P(\theta|\alpha) = \mathcal{D}(\theta|\alpha)$$

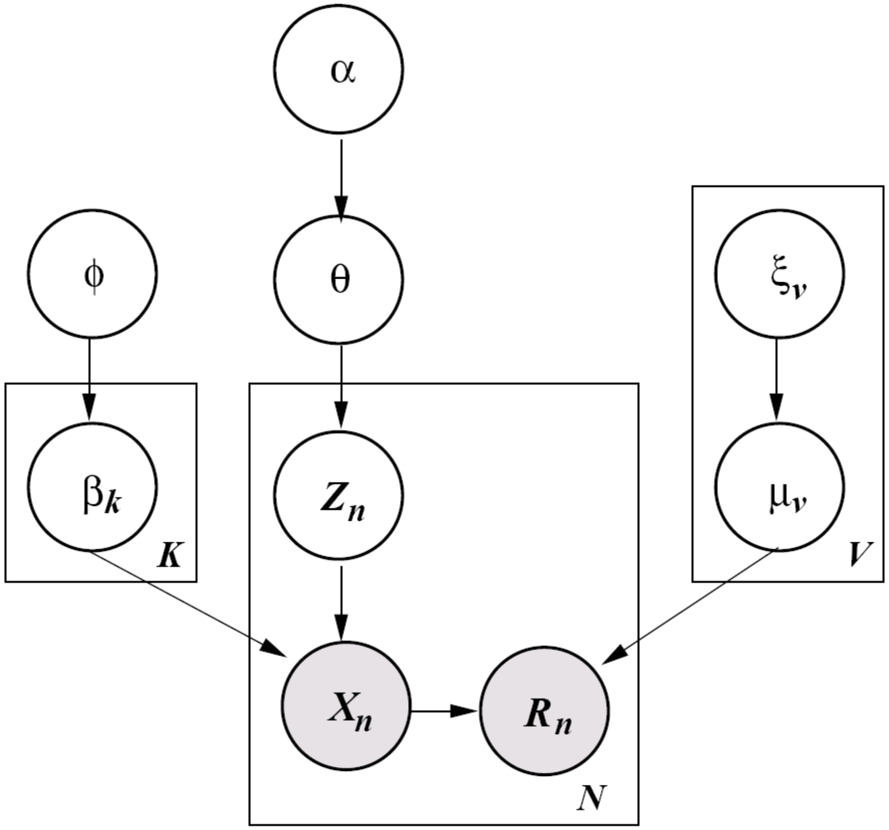
$$P(\beta|\phi) = \prod_{k=1}^{K}\prod_{d=1}^{D}\mathcal{D}(\beta_{dk}|\phi_{dk})$$

$$P(Z_n = k|\theta) = \theta_k$$

$$P(\mathbf{X} = \mathbf{x}_n|Z_n = k, \beta) = \prod_{d=1}^{D}\prod_{v=1}^{V}\beta_{vdk}^{[x_{dn}=v]}$$

$$P(\mu|\xi) = \prod_{v}\mathcal{B}(\mu_v|\xi_v)$$

$$P(\mathbf{R} = \mathbf{r}_n|\mathbf{X} = \mathbf{x}_n, \mu) = \prod_{d=1}^{D}\prod_{v=1}^{V}\mu_v^{[r_{dn}=1][x_{dn}=v]}(1-\mu_v)^{[r_{dn}=0][x_{im}=v]}$$



Models: *Finite Mixture/CPT-v*

Observation Model:

$$P(R_{dn}|X_{dn} = v, \mu) = \mu_v^{[r_{dn}=1]}(1 - \mu_v)^{[r_{dn}=0]}$$

- Simple non-random observation process where the probability of observing a rating with value *v* is Bernoulli distributed with parameter μ_v.

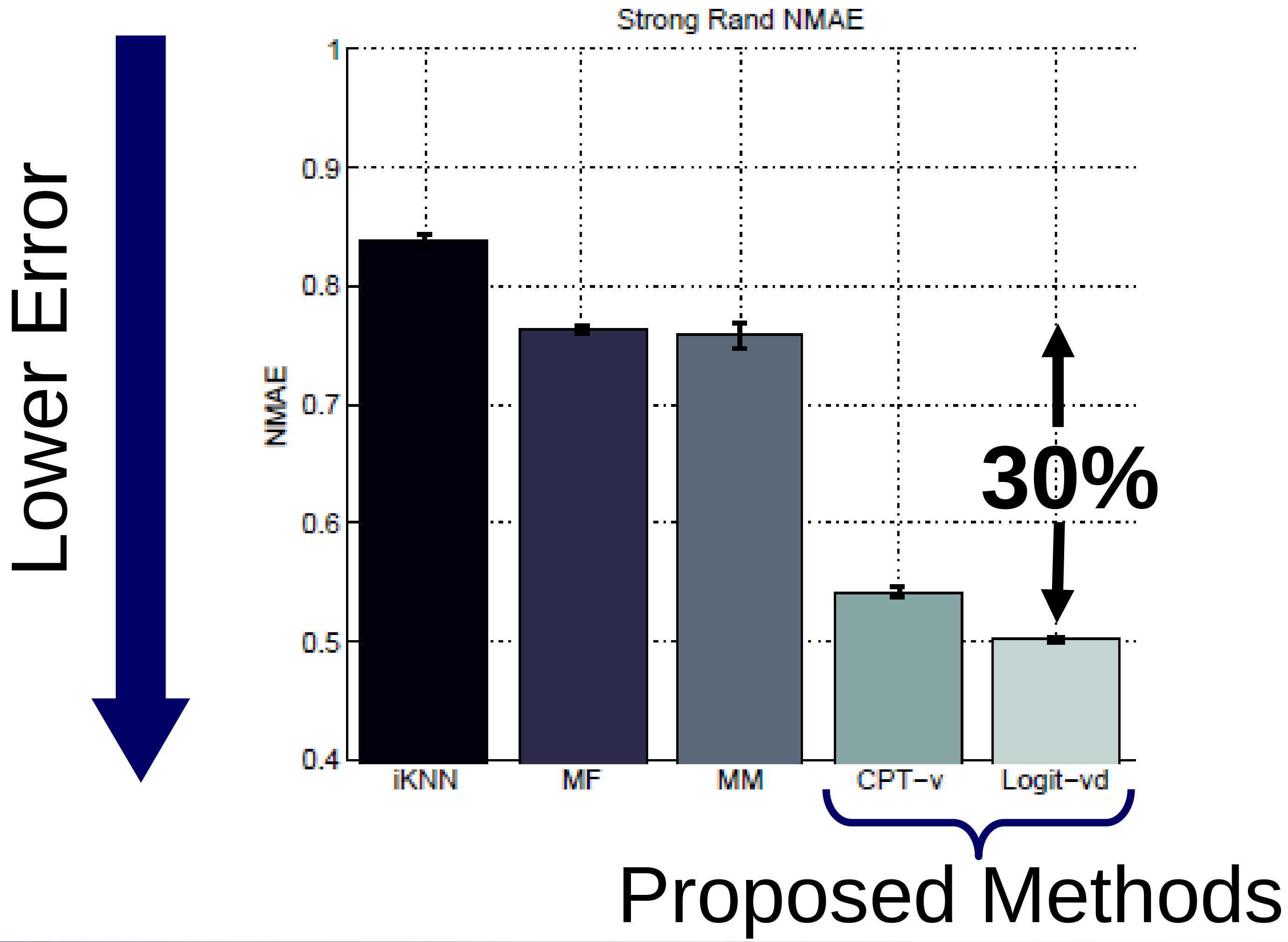
Talk Outline:

- Introduction
- Missing Data Theory and Implications
- Yahoo! LaunchCast Study
- Models and Algorithms
- **Experiments and Results**

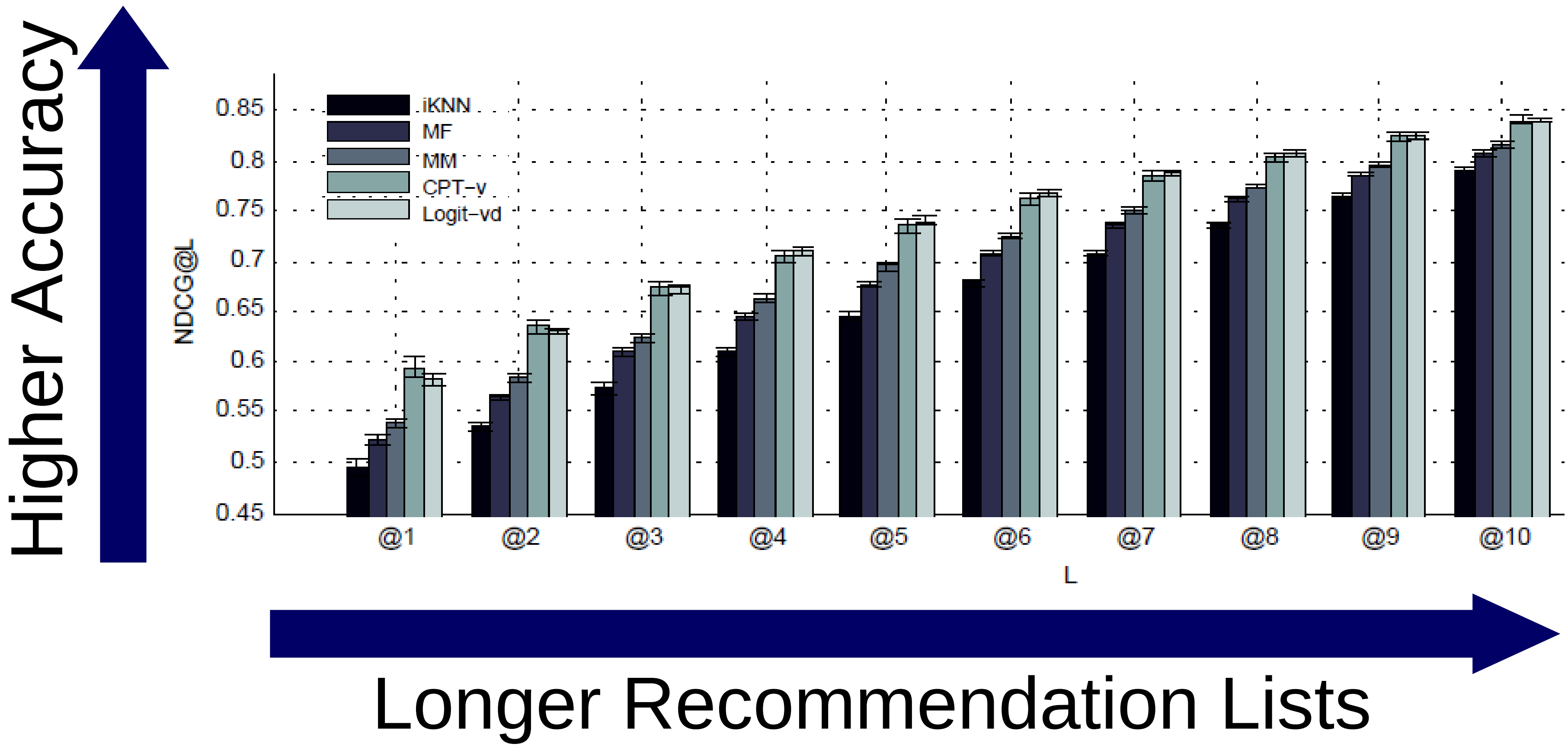
Experiments: *Protocol*

1. Train models on ratings for user selected items collected during normal interaction.
2. Test models on ratings for randomly selected items collected during survey.
3. Evaluate prediction and ranking using MAE and NDCG.
4. We consider iKNN, SVD, MM/MAR, MM/CPT-v, MM/Logit-vd.

Results: *Prediction - NMAE*



Results: *Ranking - NDCG*



Conclusions:

- We believe non-random observation processes are a reality for recommender systems.
- Treating missing data as if it were MAR results in poor performance on the rating prediction and ranking tasks we really care about.
- Simple NMAR models can be combined with standard complete data models to yield improved performance on both tasks.

Future Directions:

- Much room for testing existing prediction and ranking methods on the Yahoo! data set.
- Combining CPT-v and Logit-vd with other data models (LDA/Aspect models).
- Deriving more flexible observation models for the discrete as well as continuous cases.
- Generalizing observation models to include rating-scale usage models.

Future Directions:

- Re-visiting the debate about side information in the non-random observation process setting.

- Developing and testing models for the alternative factorization P(X|R)P(R).

- Developing methods that can side-step these issues instead of meeting them head on.

- Instrumenting software and devices to collect rich, implicit feedback and forget about ratings completely.

Acknowledgements:

• **We would like to thank:** Malcolm Slaney, Ron Brachman and David Pennock at Yahoo! Research. Todd Beaupre and Mike Mull at Yahoo! Music. Sam Roweis at NYU.