# Calibrated Recommendations

Harald Steck,
presented by Justin Basilico
at RecSys 2018

NETFLIX

# Basic Idea

**user has played:**

70 romance movies | 30 action movies

**Calibrated recommendations:**

70% romance | 30% action

… aims to reflect: all interests of user & with correct proportions

… fairness regarding all the interests of a user

# Accuracy vs. Calibration

# Accurate vs. Calibrated Recommendations

**Accuracy as prediction objective can lead to unbalanced recommendations:**

- recommendations may
  - amplify main interests of user, and
  - crowd out the lesser interests of a user.

- 2 examples in the following (see paper for more)

# 1. Accuracy vs. Calibration (in binary classification)

**data comprised of :**

| 70 romance movies | 30 action movies |
|---|---|

if no additional information available about movies (extreme case)

**predict genre of each movie:**

| 100 % romance |
|---|

→ **accuracy:** 100% * 70 = **70** movies labeled correctly

# 1. Accuracy vs. Calibration (in binary classification)

**data comprised of :**

| 70 romance movies | 30 action movies |
|---|---|

if no additional information available about movies (extreme case)

**predict genre of each movie:**

| 70 % romance | 30 % action |
|---|---|

→ **accuracy:** 70% * 70 + 30% * 30 = **58** movies labeled correctly (in expectation)

# 2. Recommended List generated from LDA model

| **Sampling** | **Ranking** |
| --- | --- |
| 1. Sample a topic (genre) *g* for user *u*: $g \sim p(G\|u)$ 2. Sample a word (video) *i* from topic *g*: $i \sim p(I\|g)$ | Sort videos *i* according to their probabilities $p(i\|u)$ for user *u*, where $p(i\|u) = \sum_g p(i\|g) \cdot p(g\|u)$ |
| → - expected to preserve genre-proportions - reduced accuracy | → - genre-proportions not preserved - increased accuracy |

NETFLIX

# Calibration Metric

# Calibration Metric

- genre-distribution of each movie is given: $p(g|i)$
  (or other categorization)

- genre-distribution of user's play history: $\boxed{p(g|u)} = \frac{\sum_{i \in \mathcal{H}} w_{u,i} \cdot p(g|i)}{\sum_{i \in \mathcal{H}} w_{u,i}}$

  … add prior for other genres: $\bar{p}(g|u) = \beta \cdot p_0(g) + (1-\beta) \cdot p(g|u)$
  (for diversity)

- genre-distribution of recommended list: $\boxed{q(g|u)} = \frac{\sum_{i \in \mathcal{I}} w_{r(i)} \cdot p(g|i)}{\sum_{i \in \mathcal{I}} w_{r(i)}}$

# Calibration Metric

- Kullback-Leibler divergence: how similar are *p* and q ?

$$C_{\mathrm{KL}}(\boxed{p}, \boxed{q}) = \mathrm{KL}(p||\tilde{q}) = \sum_{g} \boxed{p(g|u)} \log \frac{\boxed{p(g|u)}}{\boxed{\tilde{q}(g|u)}}$$

… as to avoid q(.)=0: $\tilde{q}(g|u) = (1-\alpha) \cdot q(g|u) + \alpha \cdot p(g|u)$

- or other f-divergences (see paper)

NETFLIX

# Calibration Method

# Calibration Method

- calibration is a list-property
- recommender systems often trained via pointwise or pairwise approach

→ <u>re-ranking in post-processing step:</u> $\lambda$ determines trade-off

$$\mathcal{I}^* = \underset{\mathcal{I}, |\mathcal{I}|=N}{\arg\max}\ (1-\lambda) \cdot \boxed{s(\mathcal{I})} - \lambda \cdot \boxed{C_{\mathrm{KL}}(p, q(\mathcal{I}))}$$

… re-ranked list of items

$$s(\mathcal{I}) = \sum_{i \in \mathcal{I}} s(i)$$

… scores predicted by RecSys

NETFLIX

# Calibration Method

- calibration is a list-property
- recommender systems often trained via pointwise or pairwise approach

→ re-ranking in post-processing step: $\lambda$ determines trade-off

$$\mathcal{I}^* = \underset{\mathcal{I}, |\mathcal{I}|=N}{\arg\max}\ (1-\lambda) \cdot \boxed{s(\mathcal{I})} - \underbrace{\lambda \cdot \boxed{C_{\mathrm{KL}}(p, q(\mathcal{I}))}}$$

… adding several calibration-categorizations is straightforward

# Calibration Method

Equivalent greedy optimization problem (see paper):

$$\mathcal{I}^* = \arg\max_{\mathcal{I}, |\mathcal{I}|=N} (1-\lambda) \cdot \boxed{s(\mathcal{I})} + \lambda \cdot \boxed{\sum_g p(g|u) \log \sum_{i \in \mathcal{I}} w_{r(i)} \tilde{q}(g|i)}$$

submodular function:

greedy optimization is (1-1/e) optimal,

also for each length n<*N*

# Related Concepts

# Related Concepts

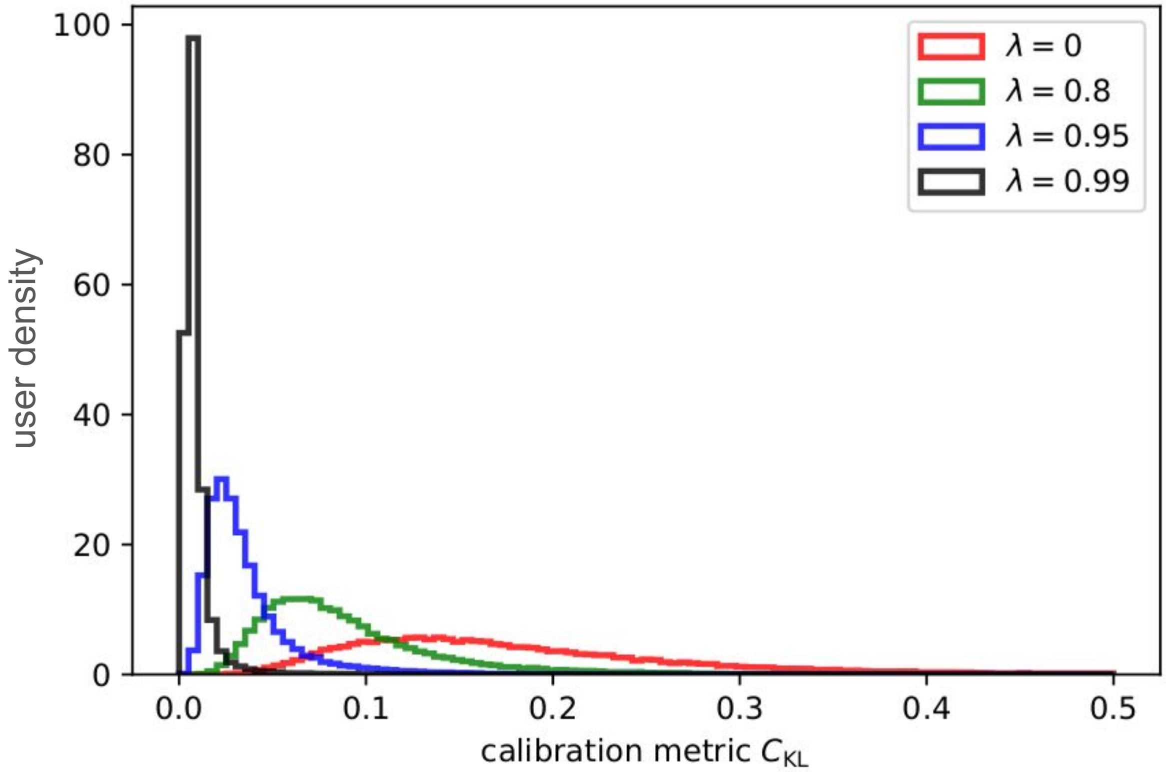
**Fairness:** - typically refers to persons or groups within a population
- several fairness criteria besides <u>calibration</u> exist:
  - equal(ized) odds, equal opportunity, statistical parity

**Diversity:** - minimal similarity or redundancy among items [majority of literature]
- proportionality in search results [Dang, Croft 2012]
- new metric to capture three properties [Vargas et al. 2014]
- focus on submodularity [Teo et al. 2016]

NETFLIX

# Experiments (on MovieLens 20 million data)

# Calibration Metric: across users




**Baseline model (wMF):**
many users receive uncalibrated rec's.

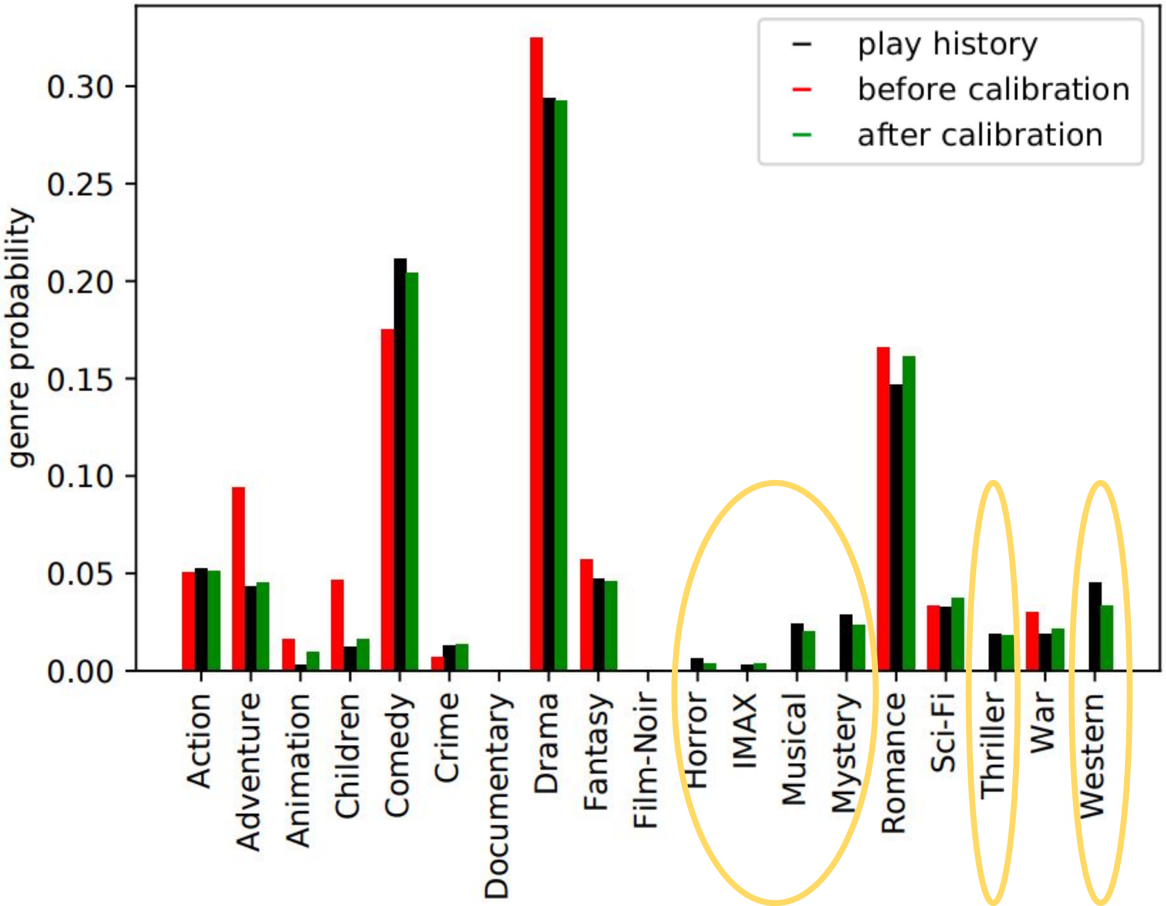
**After re-ranking:**
rec's are much more calibrated (smaller $C_{\mathrm{KL}}$)

# Calibration-Accuracy Tradeoff

| calibration | recall @10 | recall @50 | $C_{KL}$ @10 | $C_{KL}$ @50 |
|---|---|---|---|---|
| none ($\lambda = 0$) | 0.209 | 0.464 | 0.677 | 0.185 |
| $\lambda = 0.2$ | 0.209 | 0.464 | 0.465 | 0.171 |
| $\lambda = 0.5$ | 0.199 | 0.464 | 0.274 | 0.141 |
| $\lambda = 0.8$ | 0.170 | 0.463 | 0.128 | 0.092 |
| $\lambda = 0.9$ | 0.146 | 0.460 | 0.084 | 0.061 |
| $\lambda = 0.95$ | 0.121 | 0.453 | 0.065 | 0.037 |
| $\lambda = 0.99$ | 0.090 | 0.417 | 0.054 | 0.009 |
| $\lambda = 0.999$ | 0.082 | 0.339 | 0.054 | 0.005 |

- Calibration can be improved a lot without degrading accuracy much.
- Extreme calibration reduces accuracy considerably.

# Genre-Distribution for a User

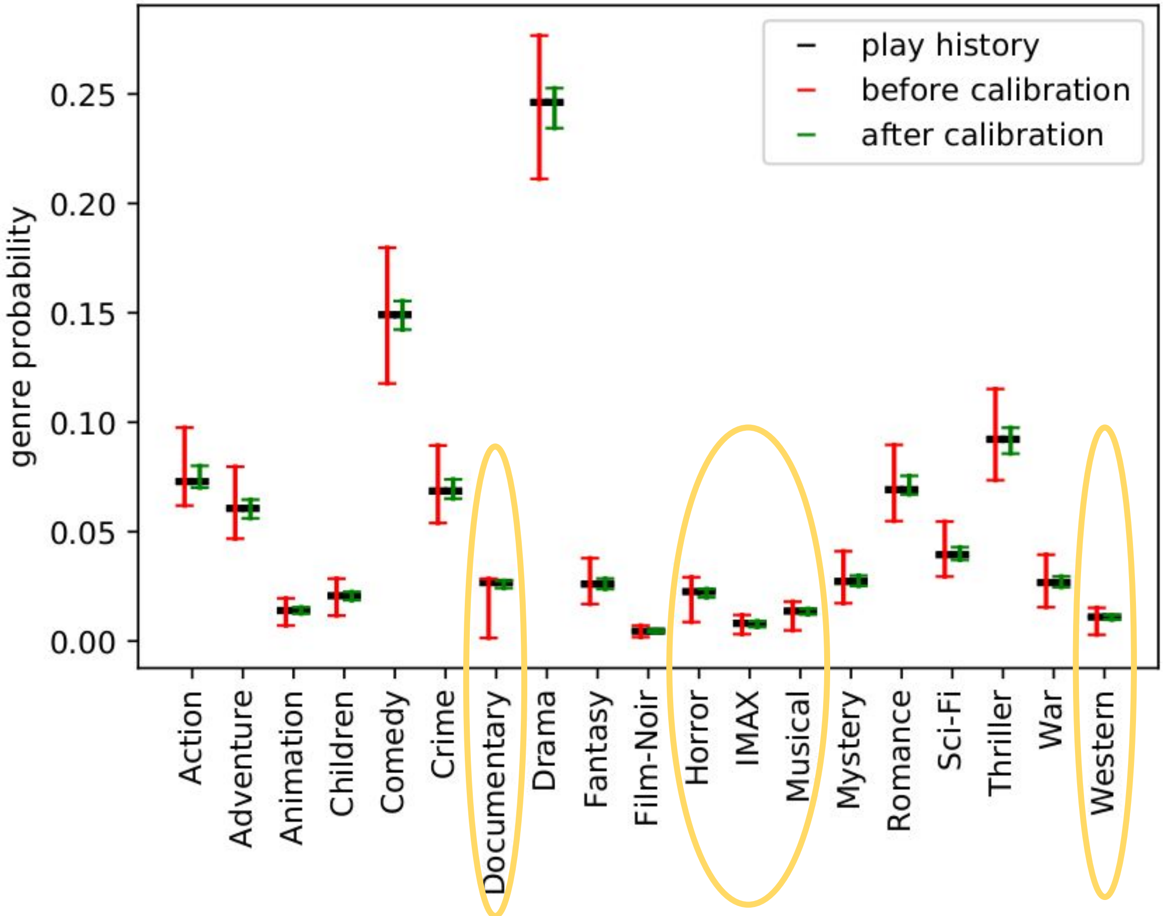



Example: a user with very uncalibrated rec's:

- Without calibration, lesser interests of user are absent from rec's.
- After calibration, all genres are recommended with approx. correct proportions.

NETFLIX

# Genre-Distribution Averaged over 10% of Users




Average over **10% of users** with least calibrated rec's:

- results similar to previous slide
- for details, see paper

NETFLIX

# Summary

# Summary

Motivation:

unbalanced recommendations can result from training recommender-models

- on limited amounts of data,
- towards accuracy-metrics.

Calibration-Approach combines two aspects:

1. aimed at **fairness / proportionality** regarding all interests of a user.
2. **submodular function** in post-processing step:
   - efficient optimization,
   - (1-1/e) optimality guarantee.

# Thank you.

NETFLIX