Google MLX

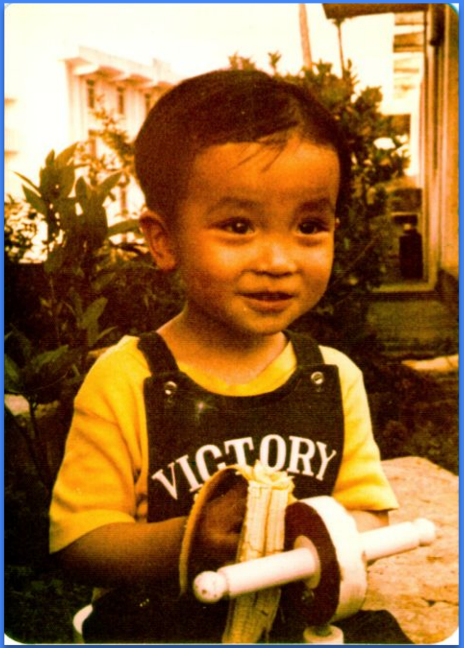
Serving Under-Represented Users in Recommenders with Focused Learning

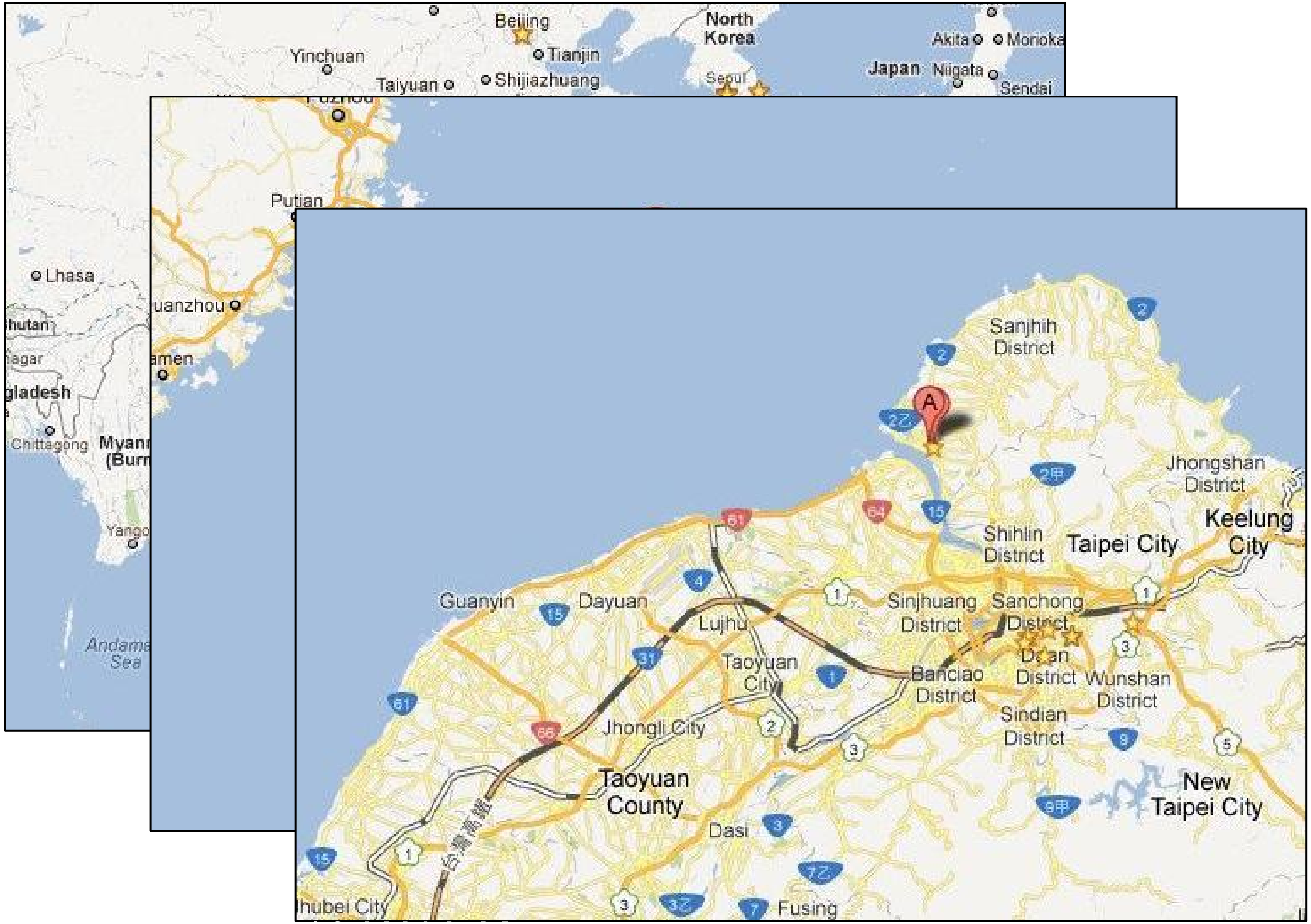
Ed H. Chi

With Alex Beutel, Zhiyuan Cheng, Hubert Pham, John Anderson

SIR research & Laser team

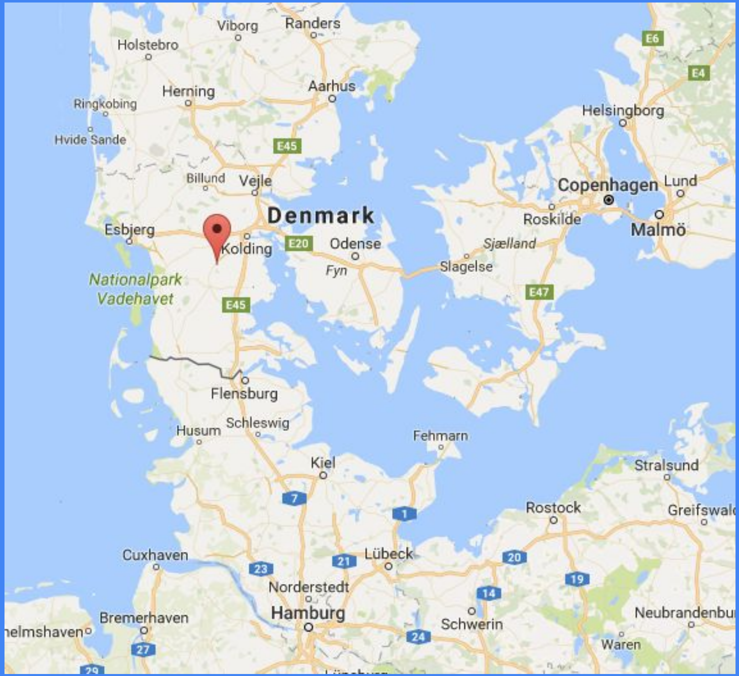
This is me ...




Beijing
North Korea
Yinchuan
Tianjin
Akita
Morioka
Japan
Niigata
Taiyuan
Shijiazhuang
Seoul
Sendai
Lhasa
Chittagong
Andaman Sea
Putian
Sanjhih District
Jhongshan District
Keelung City
Shihlin District
Taipei City
Guanyin
Dayuan
Lujhu
Sinjhuang District
Sanchong District
Taoyuan City
Banciao District
Da'an District
Wunshan District
Jhongli City
Sindian District
Taoyuan County
New Taipei City
Dasi
Fusing

Google

This is Rasmus...



LASER

Rasmus noticed lack of resolution on Danish music...

Poor regional recommendations:

- L.O.C. Danish rapper band
- D.A.D. Danish rock band
- Recommendations: dominated by generic Danish head artists

Similar phenomenon in many other regions.

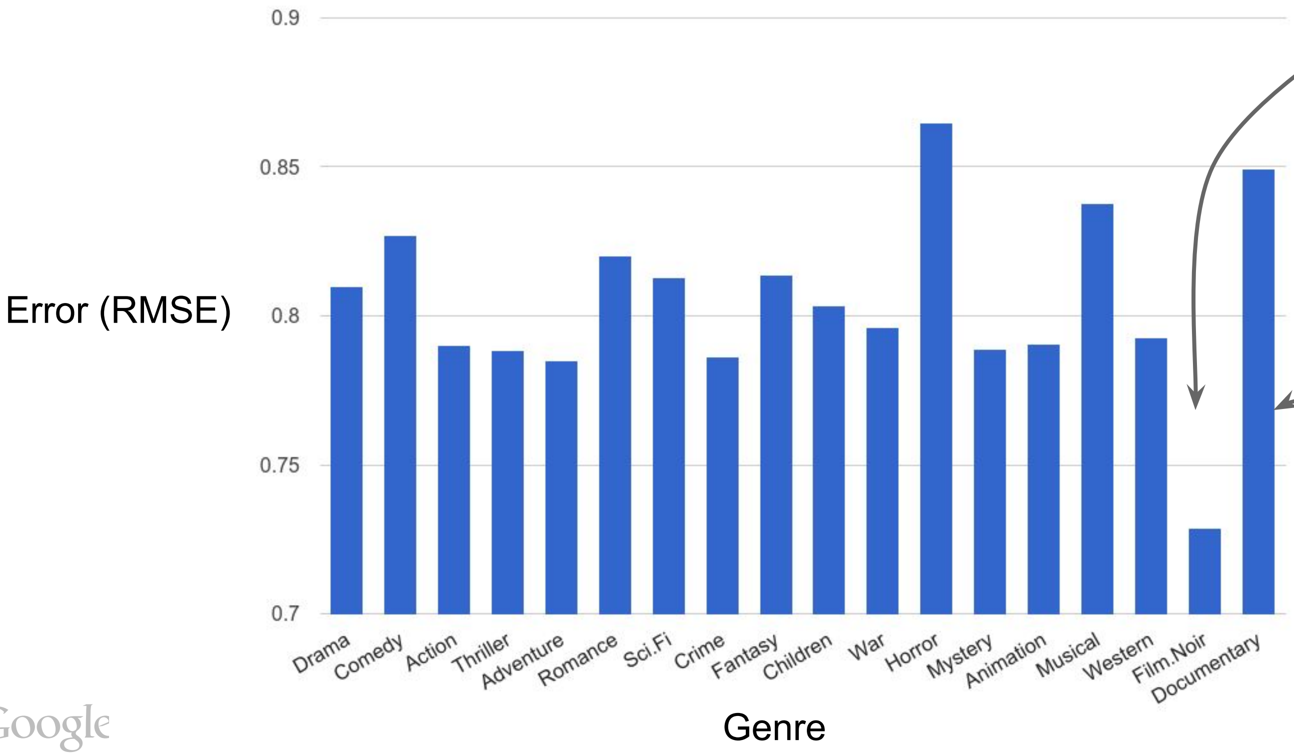
	Seeds
Rap	L.O.C.

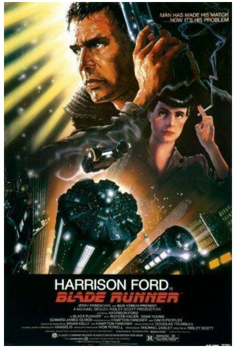
		recommended_entity Entity	Score
Rap	0	Suspekt	1.12
Rap/pop	1	Nik & Jay	1.11
Rock	2	Nephew	1.10
Pop	3	Kim Larsen	1.10
Pop	4	Kim Larsen & Kjukken	1.09
Rock	5	Dizzy Mizz Lizzy	1.09
Rock	6	TV-2	1.09
Pop	7	Thomas Helmig	1.09
Pop	8	Shu-bi-dua	1.09
Rock	9	Magtens Korridorer	1.08
Rap	10	Clemens	1.08

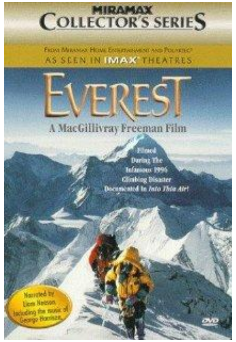
	Seeds
Rock	D.A.D.

		recommended_entity Entity	Score
Rock	0	Dizzy Mizz Lizzy	1.03
Rock	1	Gasolin	0.89
Pop	2	Shu-bi-dua	0.87
Pop/Rock	3	TV-2	0.86
Pop	4	Kim Larsen	0.86
Rock	5	D-A-D	0.85
Pop	6	Kim Larsen & Kjukken	0.85
Rock	7	Sort Sol	0.84
Pop	8	John Mogensen	0.83
Pop	9	Thomas Helmig	0.83
Rock	10	Magtens Korridorer	0.83

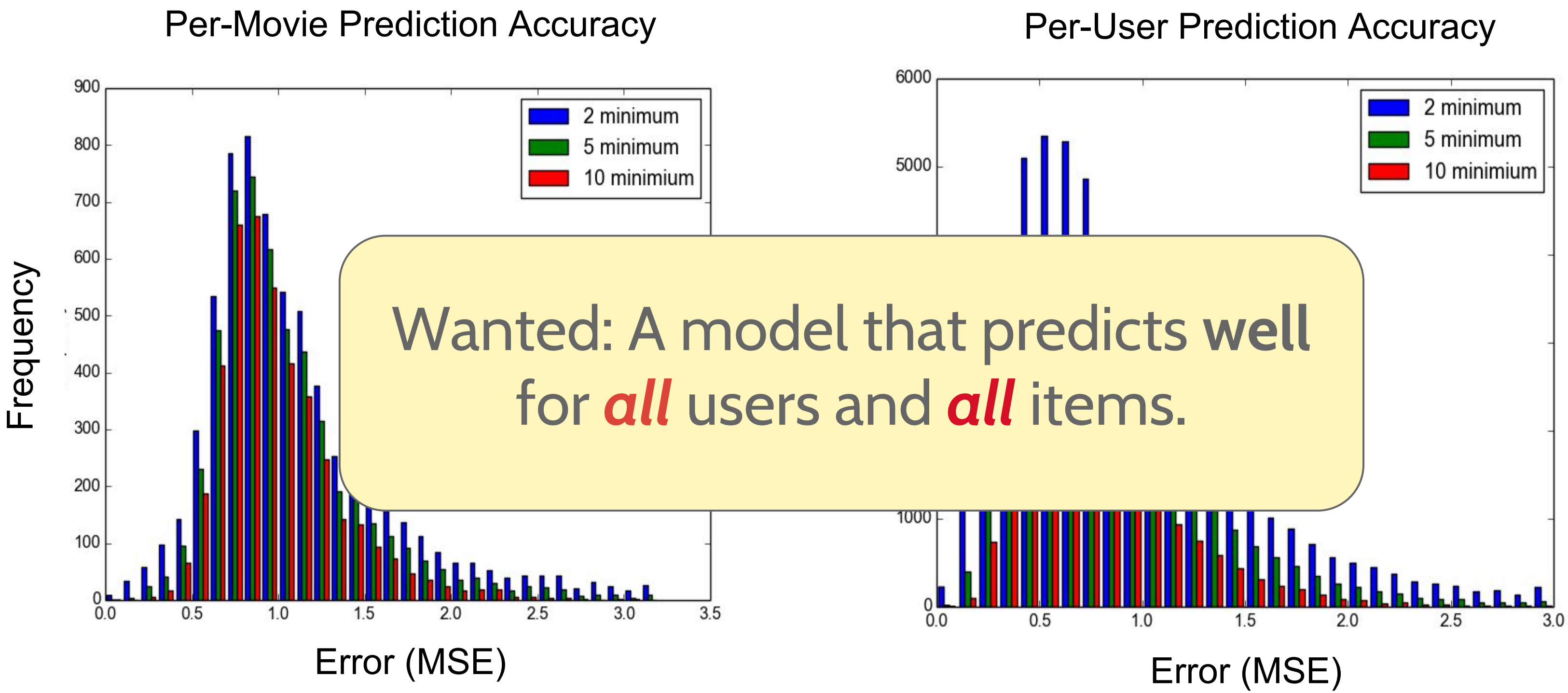
Known: Recommender quality inconsistent across movies







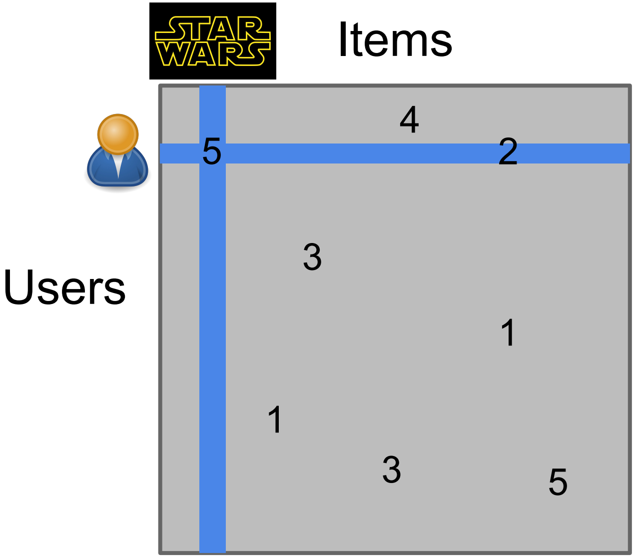
We don't represent users/items equally!



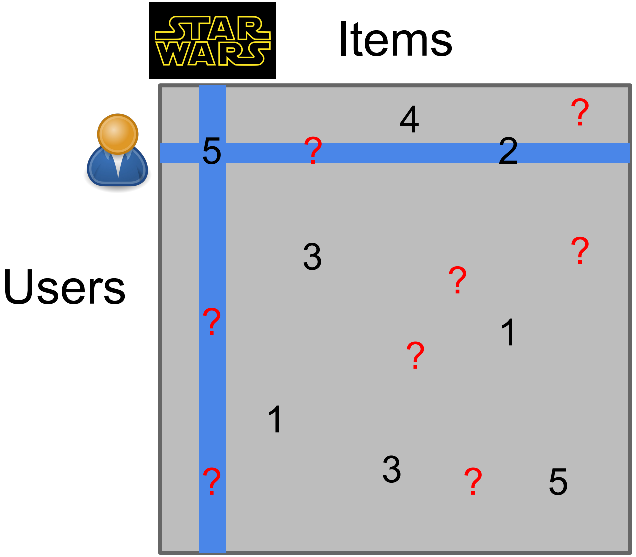
Wanted: A model that predicts **well** for ***all*** users and ***all*** items.

Why does this happen?

The Recommendation Problem



The Recommendation Problem



Given: Observed (user, item) ratings

Find: A model that predicts the missing ratings well

Used throughout Google:

Google play Google+ YouTube

Power Law of Observations

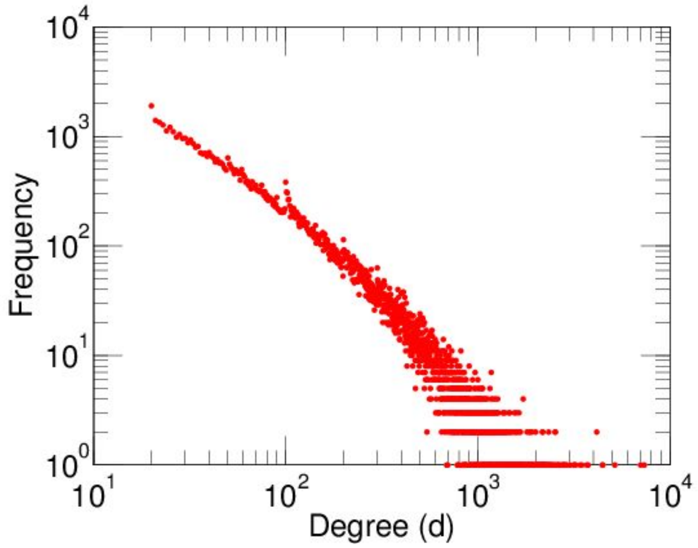
RMSE considers **all** observations equally:

$$\sqrt{\frac{1}{|\mathcal{X}|}\sum_{(i,j)\in\mathcal{X}}(\mathcal{X}_{i,j}-\langle u_i, v_j\rangle)^2}$$

Therefore it values users and movies with more ratings far more than others with less ratings.

That is, “Globally optimal” is more focused on popular, mainstream movies than niche ones.

User Degree Distribution



Images from http://konect.uni-koblenz.de/networks/movielens-10m_rating

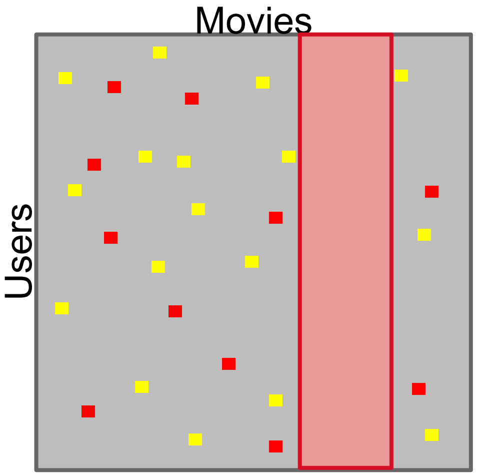
Approach

1. **Focus Selection -** Where should the additional models focus?
2. **Focused Learning -** How can learn a new model to improve prediction on a subset of the data?

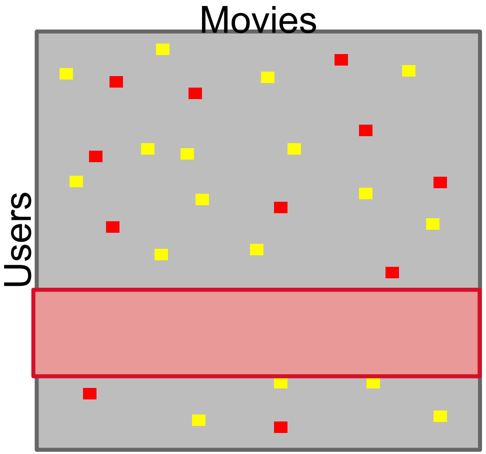
Approach

1. **Focus Selection** - Where should the additional models focus?
2. **Focused Learning** - How can learn a new model to improve predictions on a subset of the data?

Subset of columns



Subset of rows



Approach

1. Focus Selection - Where should the additional models focus?
2. **Focused Learning -** How can learn a new model to improve prediction on a subset of the data?

Distinct training and *focused* validation data

Optimize: $$\min_{\lambda, r, \vec{w}} \mathrm{WALS}_{\mathcal{X}, \mathcal{X}^{(f)}}(\lambda, r, \vec{w})$$

Simple grid search

Results: Focused by Movie Spectra

Global	Focused	% Improvement	Focused λ	Unfocused λ
1.1138	1.0212	8.314	15	150
1.0032	0.9408	6.223	15	150
0.9862	0.9450	4.174	15	60
0.9299	0.9053	2.648	15	60
0.9060	0.8908	1.679	15	60
0.8764	0.8711	0.602	30	60
0.8576	0.8552	0.279	30	60
0.8333	0.8325	0.098	30	60
0.8090	0.8090	0.003	30	60
0.7953	0.7953	0.000	30	30

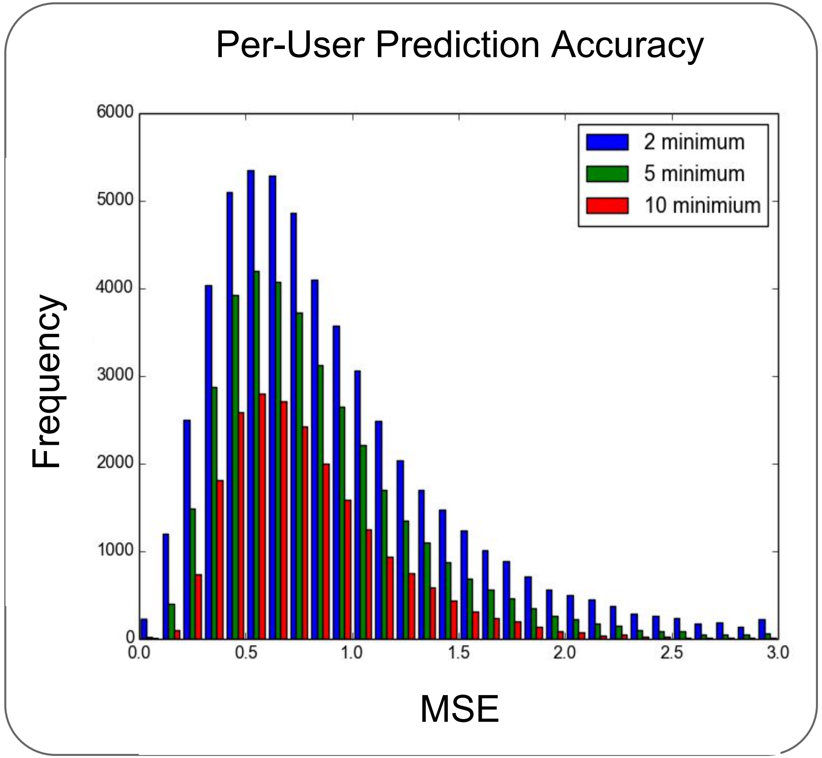
Results: Focused by User

Percentile	Global	Focused	% Improvement
0%-1%	1.9989	1.9664	1.629
0%-10%	1.4192	1.4107	0.596
10%-20%	1.1177	1.1112	0.581
20%-30%	0.9824	0.9769	0.553
30%-40%	0.8834	0.8782	0.587

Conclusion

1. "Globally optimal" is not best for everybody.
2. Learn additional models focused on problematic regions.

"Myth of the average user!"



Contact me:

edchi@
go/sir

Fight the long-tail with different representations!