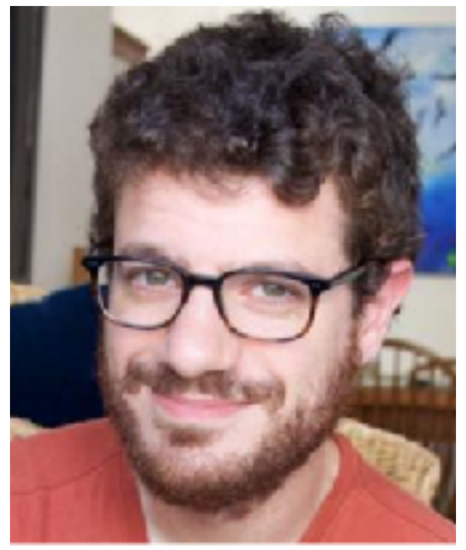
Variational Autoencoders for Recommendation

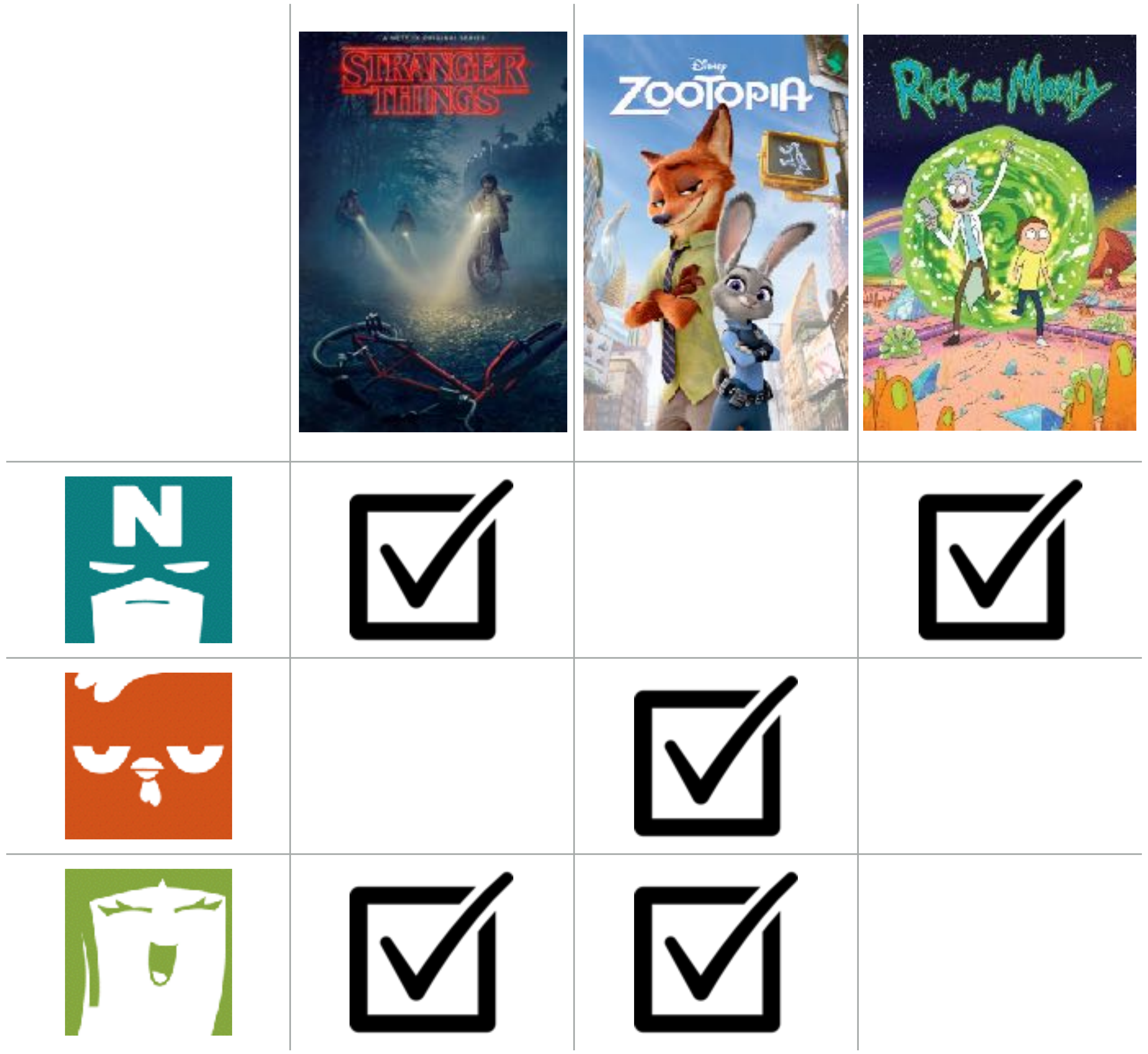
Dawen Liang
Netflix Research



Rahul Krishnan   Matt Hoffman   Tony Jebara

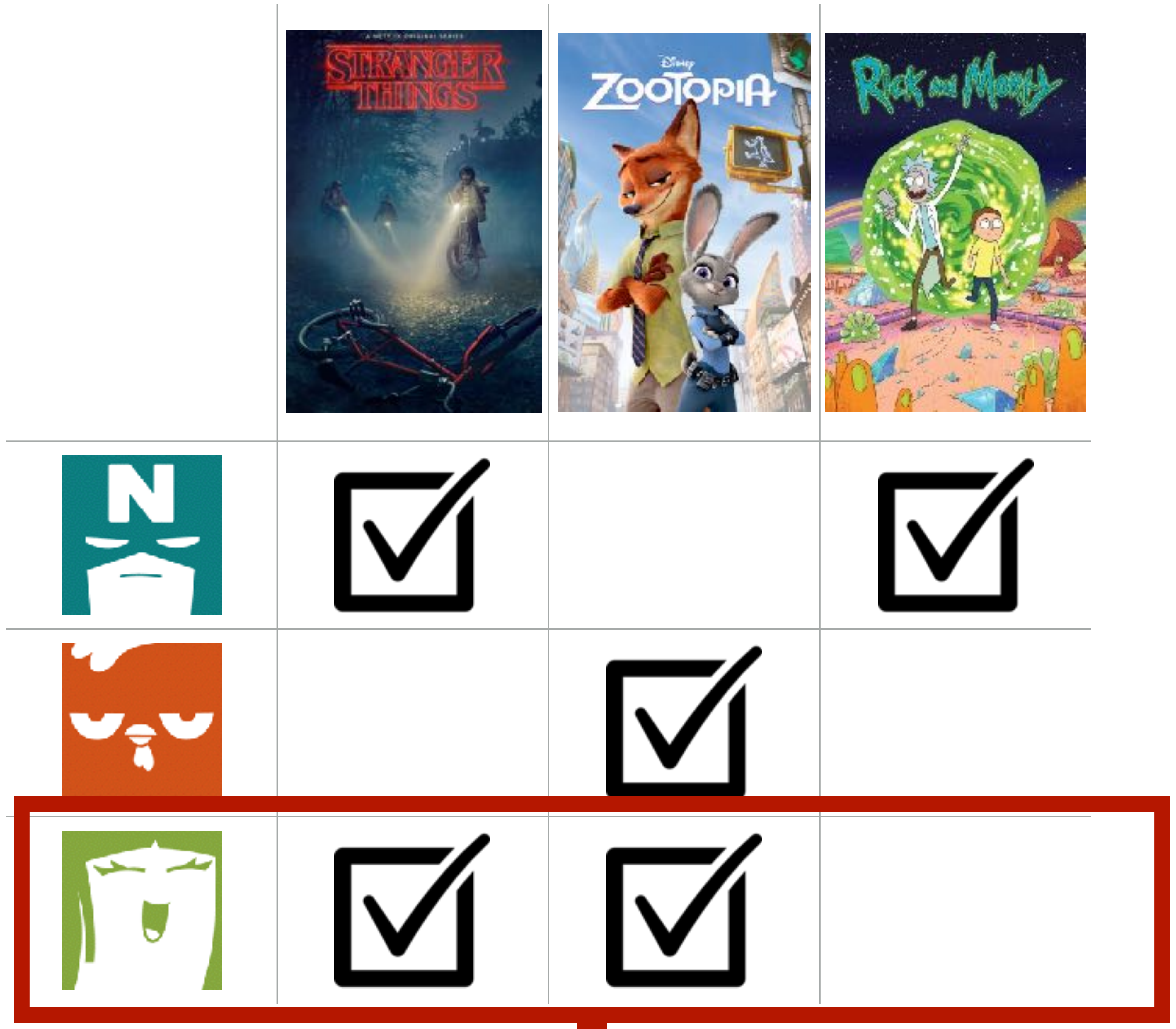
Background

- Implicit feedback data (No more rating predictions with RMSE please :)
  - In the form of user-item interaction matrix
  - Both the observed and missing entries are taken into account for modeling
  - Top-N recommender systems



Background

- Implicit feedback data (No more rating predictions with RMSE please :)
  - In the form of user-item interaction matrix
  - Both the observed and missing entries are taken into account for modeling
  - Top-N recommender systems



$$\mathbf{x}_u = [0, 0, 1, 0, \ldots, 1, 0]$$

Variational autoencoders: Model & Inference

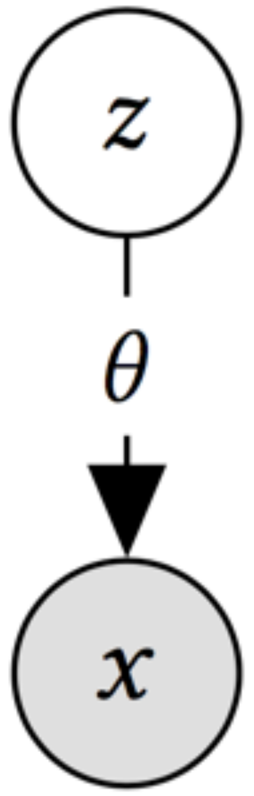
Kingma & Welling, Auto-encoding variational Bayes, 2013
Rezende et al., Stochastic backpropagation and approximation inference in deep generative models, 2014

Variational autoencoders: Model & Inference

- Model: multinomial non-linear factor analysis

For each user *u*

$$\mathbf{z}_u \sim \mathcal{N}(0, \mathrm{I}_K), \quad \pi(\mathbf{z}_u) \propto \exp\{f_\theta(\mathbf{z}_u)\},$$
$$\mathbf{x}_u \sim \mathrm{Mult}(N_u, \pi(\mathbf{z}_u)).$$



Kingma & Welling, Auto-encoding variational Bayes, 2013
Rezende et al., Stochastic backpropagation and approximation inference in deep generative models, 2014

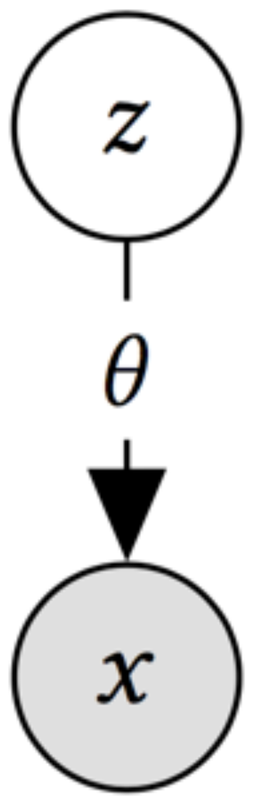
Variational autoencoders: Model & Inference

- Model: multinomial non-linear factor analysis

For each user *u*

$$\mathbf{z}_u \sim \mathcal{N}(0, \mathrm{I}_K), \quad \pi(\mathbf{z}_u) \propto \exp\{\boxed{f_\theta}(\mathbf{z}_u)\},$$
$$\mathbf{x}_u \sim \mathrm{Mult}(N_u, \pi(\mathbf{z}_u)).$$

Non-linear function



Kingma & Welling, Auto-encoding variational Bayes, 2013
Rezende et al., Stochastic backpropagation and approximation inference in deep generative models, 2014

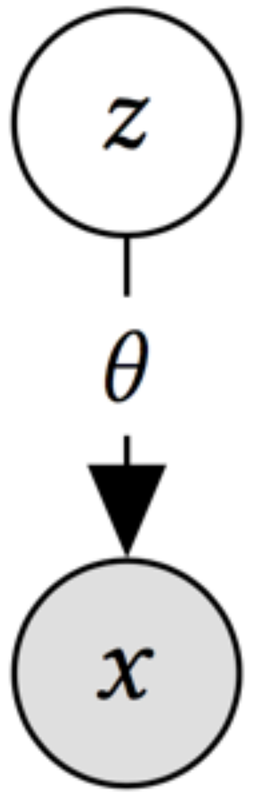
Variational autoencoders: Model & Inference

- Model: multinomial non-linear factor analysis

For each user *u*

$$\mathbf{z}_u \sim \mathcal{N}(0, \mathbf{I}_K), \quad \pi(\mathbf{z}_u) \propto \exp\{\boxed{f_\theta}(\mathbf{z}_u)\},$$
$$\mathbf{x}_u \sim \mathrm{Mult}(N_u, \pi(\mathbf{z}_u)).$$

Non-linear function



- Inference: reason about the (intractable) posterior

$$p(\mathbf{z}_u \mid \mathbf{x}_u) \approx q(\mathbf{z}_u) = \mathcal{N}(\boldsymbol{\mu}_u, \boldsymbol{\sigma}_u^2)$$

Kingma & Welling, Auto-encoding variational Bayes, 2013
Rezende et al., Stochastic backpropagation and approximation inference in deep generative models, 2014

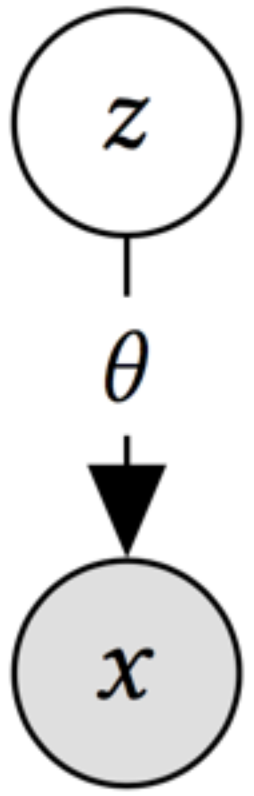
Variational autoencoders: Model & Inference

- Model: multinomial non-linear factor analysis

For each user *u*

$$\mathbf{z}_u \sim \mathcal{N}(0, \mathrm{I}_K), \quad \pi(\mathbf{z}_u) \propto \exp\{\boxed{f_\theta}(\mathbf{z}_u)\},$$
$$\mathbf{x}_u \sim \mathrm{Mult}(N_u, \pi(\mathbf{z}_u)).$$

Non-linear function



- Inference: reason about the (intractable) posterior

$$p(\mathbf{z}_u \mid \mathbf{x}_u) \approx q(\mathbf{z}_u) = \mathcal{N}(\boxed{\boldsymbol{\mu}_u}, \boxed{\boldsymbol{\sigma}_u^2})$$

Free parameters

Kingma & Welling, Auto-encoding variational Bayes, 2013
Rezende et al., Stochastic backpropagation and approximation inference in deep generative models, 2014

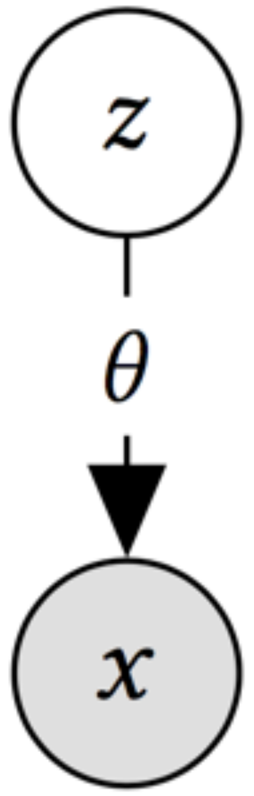
Variational autoencoders: Model & Inference

- Model: multinomial non-linear factor analysis

For each user *u*

$$\mathbf{z}_u \sim \mathcal{N}(0, \mathrm{I}_K), \quad \pi(\mathbf{z}_u) \propto \exp\{\boxed{f_\theta}(\mathbf{z}_u)\},$$
$$\mathbf{x}_u \sim \mathrm{Mult}(N_u, \pi(\mathbf{z}_u)).$$

Non-linear function



- Inference: data-dependent inference functions

Kingma & Welling, Auto-encoding variational Bayes, 2013
Rezende et al., Stochastic backpropagation and approximation inference in deep generative models, 2014

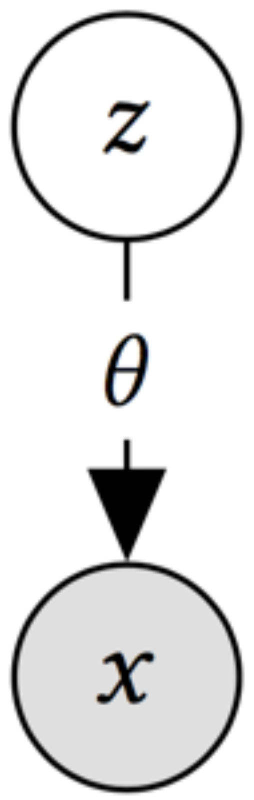
Variational autoencoders: Model & Inference

- Model: multinomial non-linear factor analysis

For each user *u*

$$\mathbf{z}_u \sim \mathcal{N}(0, \mathrm{I}_K), \quad \pi(\mathbf{z}_u) \propto \exp\{\boxed{f_\theta}(\mathbf{z}_u)\},$$
$$\mathbf{x}_u \sim \mathrm{Mult}(N_u, \pi(\mathbf{z}_u)).$$

Non-linear function



- Inference: data-dependent inference functions

$$p(\mathbf{z}_u \mid \mathbf{x}_u) \approx q_\phi(\mathbf{z}_u \mid \mathbf{x}_u) = \mathcal{N}(\mu_\phi(\mathbf{x}_u), \sigma^2_\phi(\mathbf{x}_u))$$

Kingma & Welling, Auto-encoding variational Bayes, 2013
Rezende et al., Stochastic backpropagation and approximation inference in deep generative models, 2014

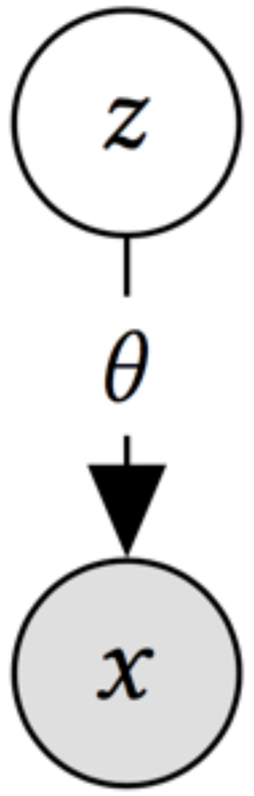
Variational autoencoders: Model & Inference

- Model: multinomial non-linear factor analysis

For each user *u*

$$\mathbf{z}_u \sim \mathcal{N}(0, \mathrm{I}_K), \quad \pi(\mathbf{z}_u) \propto \exp\{\boxed{f_\theta}(\mathbf{z}_u)\},$$
$$\mathbf{x}_u \sim \mathrm{Mult}(N_u, \pi(\mathbf{z}_u)).$$

Non-linear function



- Inference: data-dependent inference functions

$$p(\mathbf{z}_u \mid \mathbf{x}_u) \approx q_\phi(\mathbf{z}_u \mid \mathbf{x}_u) = \mathcal{N}(\boxed{\mu_\phi}(\mathbf{x}_u), \boxed{\sigma^2_\phi}(\mathbf{x}_u))$$

Non-linear function

Kingma & Welling, Auto-encoding variational Bayes, 2013
Rezende et al., Stochastic backpropagation and approximation inference in deep generative models, 2014

Variational autoencoders: Model & Inference

- Model: multinomial non-linear factor analysis

For each user *u*

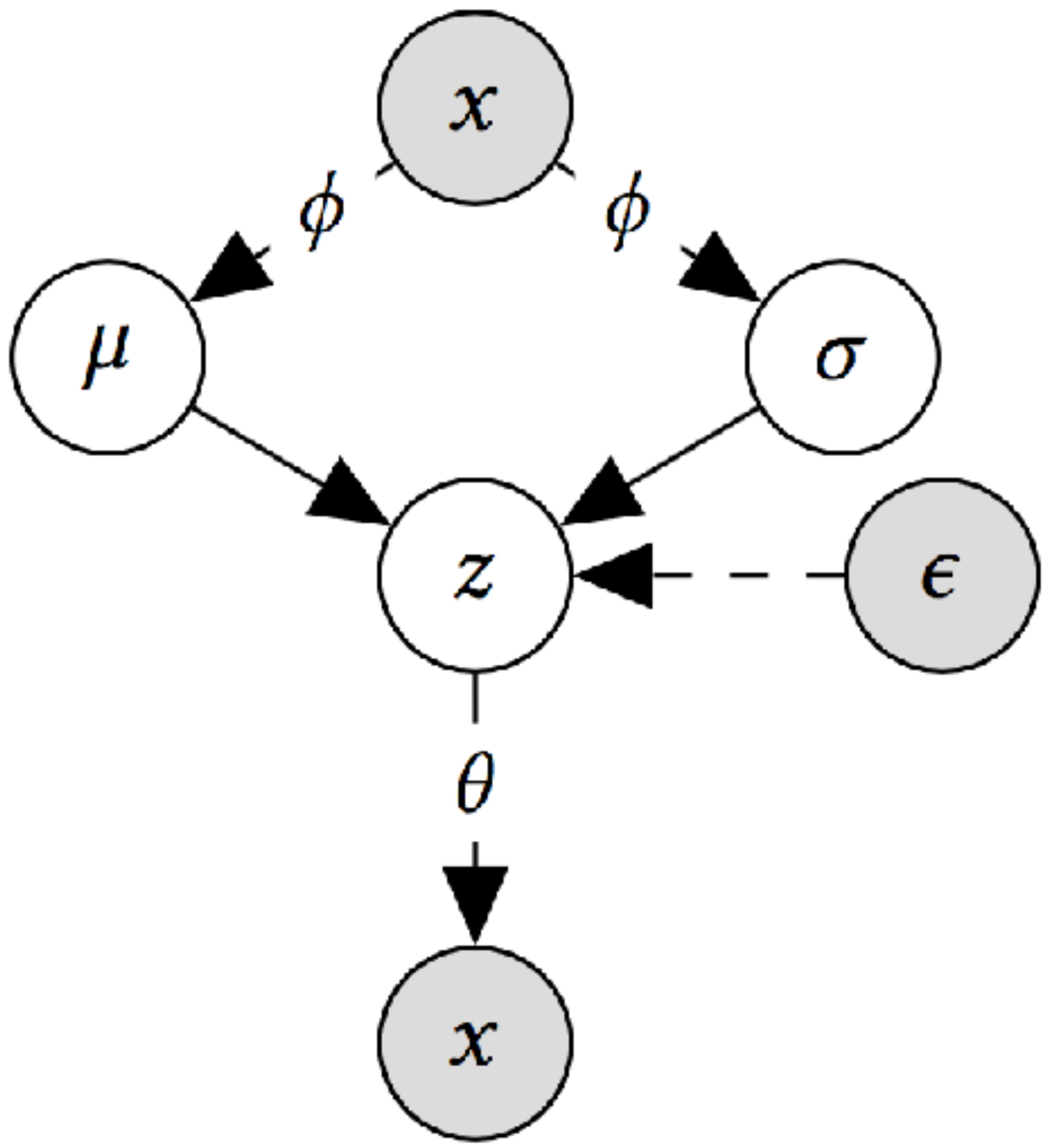
$$\mathbf{z}_u \sim \mathcal{N}(0, \mathrm{I}_K), \quad \pi(\mathbf{z}_u) \propto \exp\{\boxed{f_\theta}(\mathbf{z}_u)\},$$
$$\mathbf{x}_u \sim \mathrm{Mult}(N_u, \pi(\mathbf{z}_u)).$$

Non-linear function

- Inference: data-dependent inference functions

$$p(\mathbf{z}_u \mid \mathbf{x}_u) \approx q_\phi(\mathbf{z}_u \mid \mathbf{x}_u) = \mathcal{N}(\boxed{\mu_\phi}(\mathbf{x}_u), \boxed{\sigma^2_\phi}(\mathbf{x}_u))$$

Non-linear function



Kingma & Welling, Auto-encoding variational Bayes, 2013
Rezende et al., Stochastic backpropagation and approximation inference in deep generative models, 2014

Why VAEs
(or rather, Bayesian?)

Why VAEs
(or rather, Bayesian?)

- Generalize linear latent factor models
  - Recover Gaussian matrix factorization as a special linear case

Salakhutdinov & Mnih, Probabilistic matrix factorization, 2008

Why VAEs
(or rather, Bayesian?)

- Generalize linear latent factor models
    - Recover Gaussian matrix factorization as a special linear case
- No iterative procedure required to rank all the items given a user's watch history
    - Only need to evaluate inference and generative functions

Salakhutdinov & Mnih, Probabilistic matrix factorization, 2008

Why VAEs (or rather, Bayesian?)

- Generalize linear latent factor models
  - Recover Gaussian matrix factorization as a special linear case
- No iterative procedure required to rank all the items given a user’s watch history
  - Only need to evaluate inference and generative functions
- RecSys is more of a “small data” than a “big data” problem

Salakhutdinov & Mnih, Probabilistic matrix factorization, 2008

Training VAEs

$$\mathbb{E}_{q(\mathbf{z} \mid \mathbf{x})}\left[\log p(\mathbf{x} \mid \mathbf{z})\right] - \beta \cdot \mathrm{KL}(q(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z}))$$

Training VAEs

$$\mathbb{E}_{q(\mathbf{z}\mid\mathbf{x})}\left[\log p(\mathbf{x}\mid\mathbf{z})\right] - \beta \cdot \mathrm{KL}(q(\mathbf{z}\mid\mathbf{x})\|p(\mathbf{z}))$$

(Negative) reconstruction error

Training VAEs

$$\mathbb{E}_{q(\mathbf{z} \mid \mathbf{x})}[\log p(\mathbf{x} \mid \mathbf{z})] - \beta \cdot \mathrm{KL}(q(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z}))$$

(Negative) reconstruction error

“Regularization”

Training VAEs

$$\mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{x})}\left[\log p(\mathbf{x}\,|\,\mathbf{z})\right] - \boxed{\beta} \cdot \mathrm{KL}(q(\mathbf{z}\,|\,\mathbf{x}) \| p(\mathbf{z}))$$

(Negative) reconstruction error

“Regularization”

Training VAEs

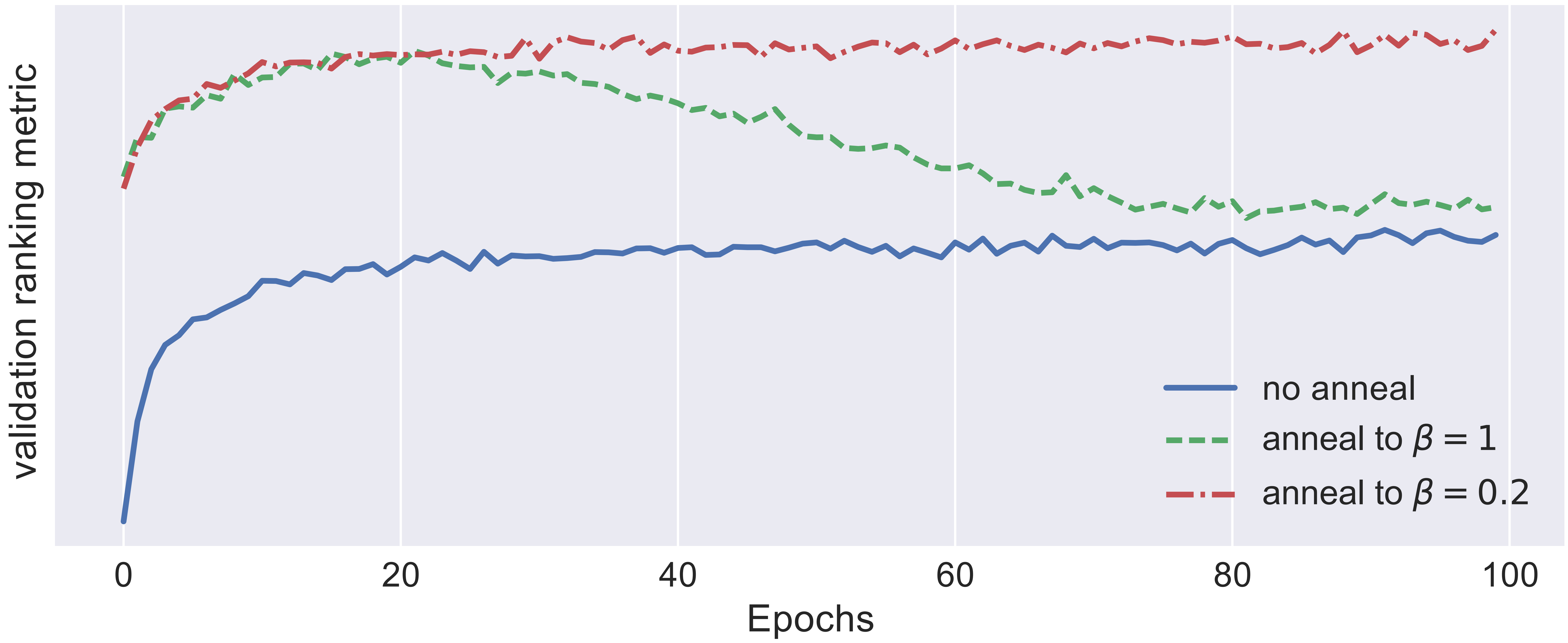
$$\mathbb{E}_{q(\mathbf{z}\mid\mathbf{x})}\left[\log p(\mathbf{x}\mid\mathbf{z})\right] - \boxed{\beta}\cdot \mathrm{KL}(q(\mathbf{z}\mid\mathbf{x})\|p(\mathbf{z}))$$

(Negative) reconstruction error

"Regularization"

- Setting $\beta < 1$ relaxes the prior constraint
  - For RecSys, we don't necessarily need all the statistical property of a generative model
  - Trading off the ability of performing ancestral sampling for better fitting the data

Selecting β



Bowman et al., Generating sentences from a continuous space, 2015

Training VAEs

$$\mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{x})}\left[\log p(\mathbf{x}\,|\,\mathbf{z})\right] - \boxed{\beta} \cdot \mathrm{KL}(q(\mathbf{z}\,|\,\mathbf{x}) \| p(\mathbf{z}))$$

- Information-theoretic connections
  - Maximum entropy discrimination & Information bottleneck principle
- Recent work on understanding the trade-offs in learning latent variable models with VAEs
  - Variational lossy autoencoders, β-VAE, deep variational information bottleneck (hopefully many to come in ICLR)

Jaakkola et al., Maximum entropy discrimination, 2000
Tishby et al., The information bottleneck method, 2000
Chen et al., Variational lossy autoencoders, 2016
Higgins et al., β-VAE: Learning basic visual concepts with a constrained variational framework, 2016
Alemi et al., Deep variational information bottleneck, 2016

Evaluation:
Strong generalization

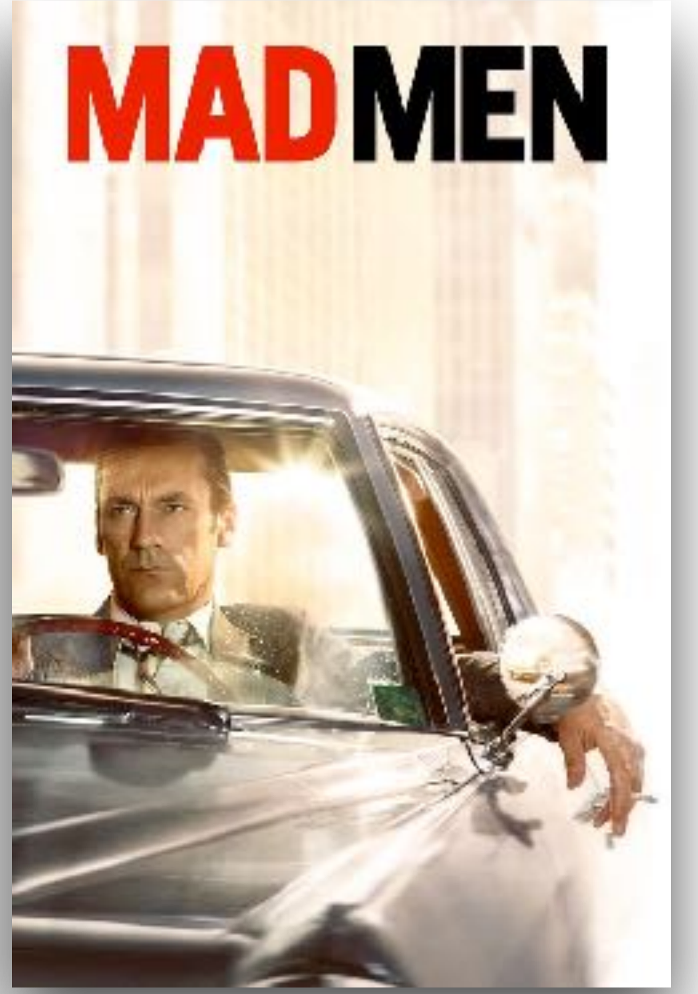
Marlin, Collaborative filtering: A machine learning perspective, 2004

Evaluation:
Strong generalization

Held-out user:
Not used in the training

Marlin, Collaborative filtering: A machine learning perspective, 2004

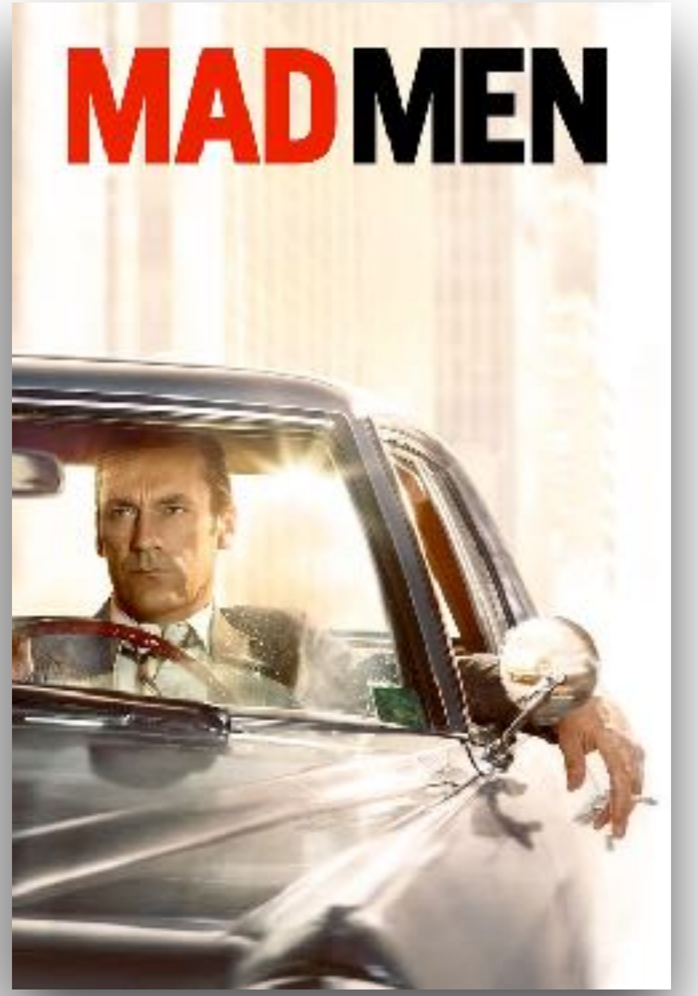
Evaluation:
Strong generalization



Held-out user:
Not used in the
training

Marlin, Collaborative filtering: A machine learning perspective, 2004

Evaluation: Strong generalization



Held-out user:
Not used in the training

“Fold-in” set:
- Learn necessary user-level representation
- Obtain predicted ranking

“Target” set:
Report ranking metrics (Recall@K, NDCG@K) on

Marlin, Collaborative filtering: A machine learning perspective, 2004

Empirical studies

	ML-20M	**Netflix**	**MSD**
# of users	136,677	463,435	571,355
# of items	20,108	17,769	41,140
# of interactions	10.0M	56.9M	33.6M
% of interactions	0.36%	0.69%	0.14%
# of held-out users	10,000	40,000	50,000

Quantitative results

- Multi-VAE^{PR}: Partially Regularized VAE with multinomial likelihood
- Multi-DAE: Denoising autoencoder with multinomial likelihood
- Baselines:
  - WMF & SLIM: linear collaborative filtering methods
  - CDAE: Non-linear neural network based method

	Recall@20	Recall@50	NDCG@100
Mult-VAE^{PR}	**0.395**	**0.537**	**0.426**
Mult-DAE	0.387	0.524	0.419
WMF	0.360	0.498	0.386
SLIM	0.370	0.495	0.401
CDAE	0.391	0.523	0.418

ML20M (s.e. ~0.002)

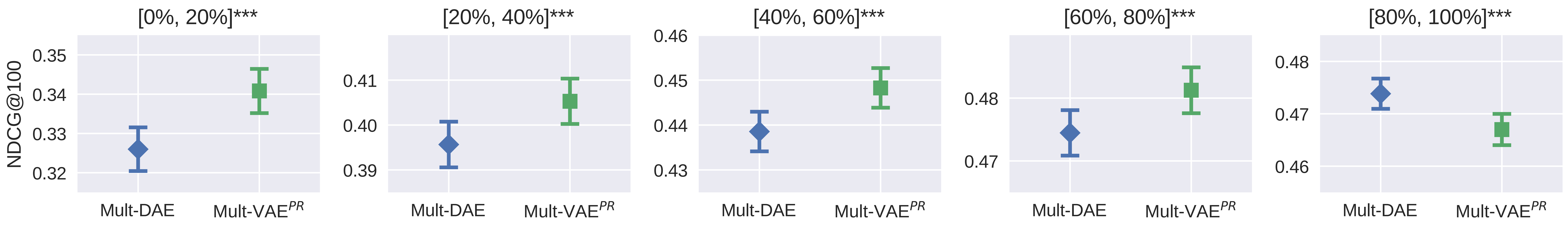
	Recall@20	Recall@50	NDCG@100
Mult-VAE^{PR}	**0.351**	**0.444**	**0.386**
Mult-DAE	0.344	0.438	0.380
WMF	0.316	0.404	0.351
SLIM	0.347	0.428	0.379
CDAE	0.343	0.428	0.376

Netflix Prize (s.e. ~0.001)

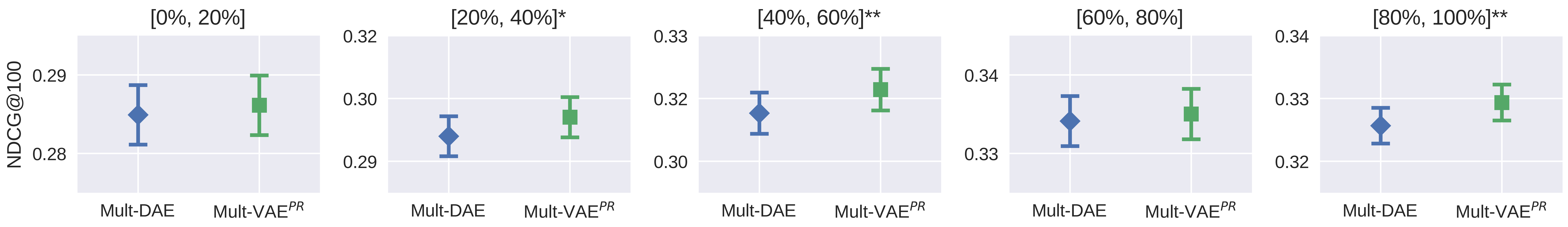
Hu et al., Collaborative filtering with implicit feedback datasets, 2008
Ning & Karypis, SLIM: Sparse linear methods for top-N recommender systems, 2011
Wu et al., Collaborative denoising auto-encoders for top-N recommender systems, 2016

Why Bayesian? (cont.)

ML20M: each user has watched at least 5 movies



MSD: each user has listened to at least 20 songs



User activity: Low → High

Conclusion

- We extend VAEs to collaborative filtering for implicit feedback
- We introduce a regularization parameter for the learning objective to trade-off the generative power for better predictive recommendation performance
- Besides competitive empirical performance, we also identify when and why a principled Bayesian approach performs better

Thanks!

- We extend VAEs to collaborative filtering for implicit feedback
- We introduce a regularization parameter for the learning objective to trade-off the generative power for better predictive recommendation performance
- Besides competitive empirical performance, we also identify when and why a principled Bayesian approach performs better

Netflix tech blog: https://medium.com/netflix-techblog