VARIATIONAL AUTOENCODERS FOR SHARED LATENT SPACE REPRESENTATION OF STUDENT BEHAVIOR

Serena Nicoll

Generative Models in Educational Settings

Rasch Model (1960)

- Generalized Linear Model (GLM)
- Item Response Theory (IRT)
- Assumptions about data distribution

Latent Dirichlet Allocation (2003)

- Topic & Content Modeling
- Only able to model one category of data (written)

Feature Embeddings with Variational Autoencoders (2017)

- Removes assumptions about data distribution
- Modeling data in a shared, low-complexity space
- Limited to gaussian distributions in the latent space

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.
Blei et al. "Latent Dirichlet Allocation." *Journal of Machine Learning Research.* 2003.
Klingler et al. "Efficient Feature Embeddings for Student Classification with Variational Autoencoders." *Conference on Educational Data Mining.* 2017.

Can we personalize student representations within a shared latent space?

**Students have many types of actions:**

- Watching Lecture Videos

- Posting in class discussion forums

- Taking Quizzes and Exams

- Interacting with posted content

- Giving and receiving feedback

- Etc.

Quizzes

Posts

?

Videos

PURDUE UNIVERSITY®
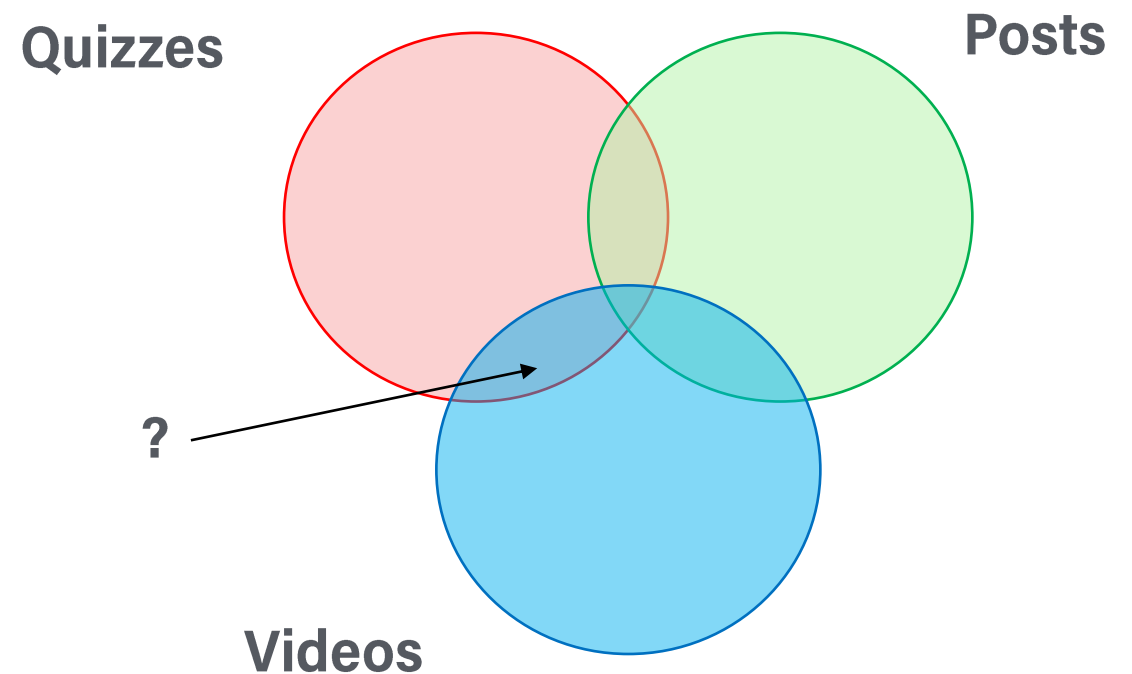
Can we personalize student representations within a shared latent space?

Accounting for Learning Differences:

- Informed by federated setting

- Data Heads: Categorical Emission

- Central VAE: Gaussian/Gaussian
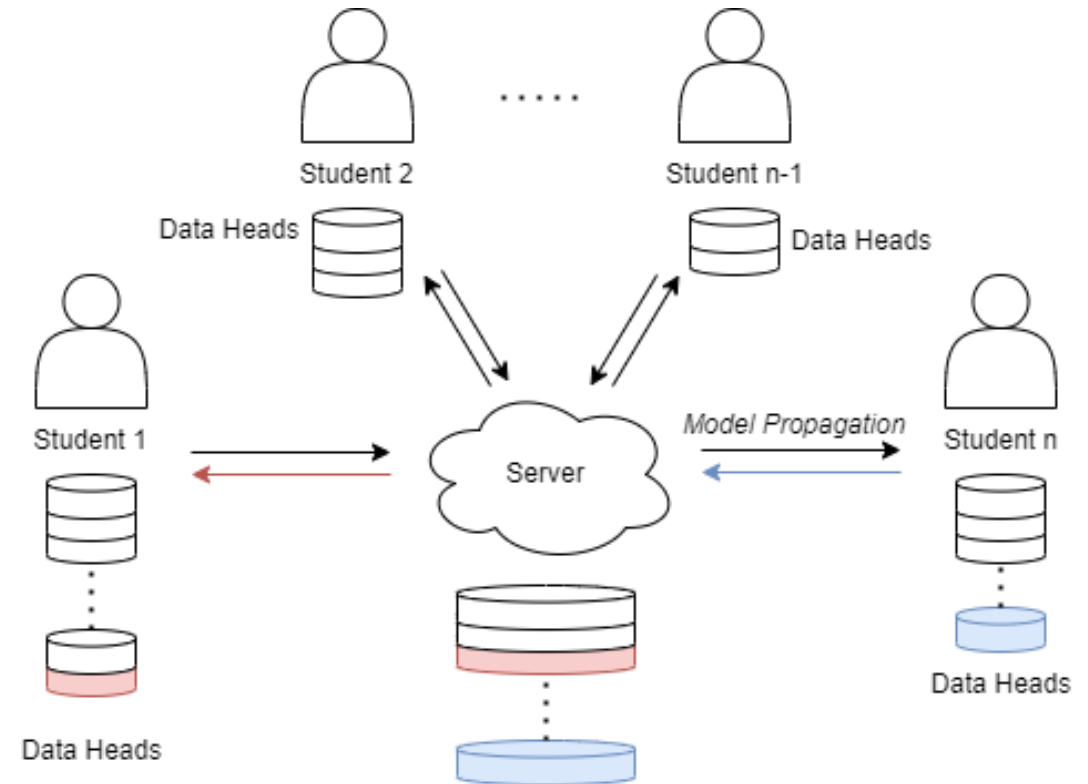
Datasets

Year	Semester	Course	Enrolled	Active	Total Actions	Forum	Video	Access
2020	Spring	A	1689	860	19408	0.671 ± 5.244	4.183 ± 7.689	17.71 ± 35.03
2020	Spring	B	2841	169	3787	2.041 ± 7.256	4.296 ± 7.207	16.07 ± 35.43
2020	Fall	C	352	133	8174	11.10 ± 33.48	5.692 ± 7.521	44.66 ± 55.17

- Online Master's Courses with global enrollment

- Emphasis on different action types per course

Push

$$\mathbb{E}_Q\big[\log P(x|z)\big] - KL(Q(Z|X)||P(Z))$$

Reconstruction Loss

KL-Divergence (sum)

- KL: Min between all components X and std. N(0, 1)

- Minimizing loss terms

- Keep distributions close to origin (and to each other)

- 0.7142

Pull
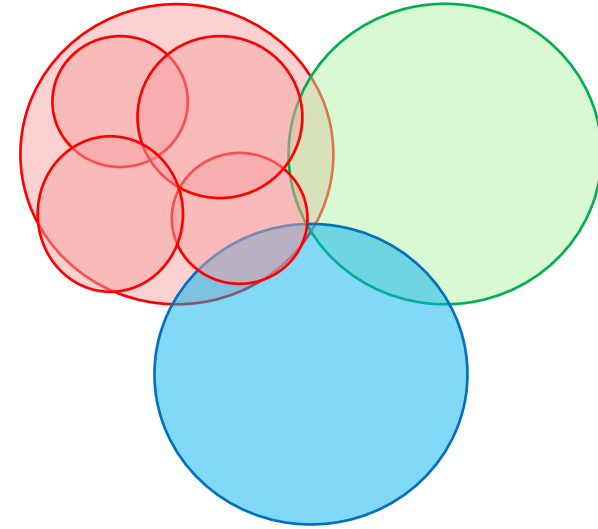
- Maximize KL-Divergence between classes

- Using distributions from VAE
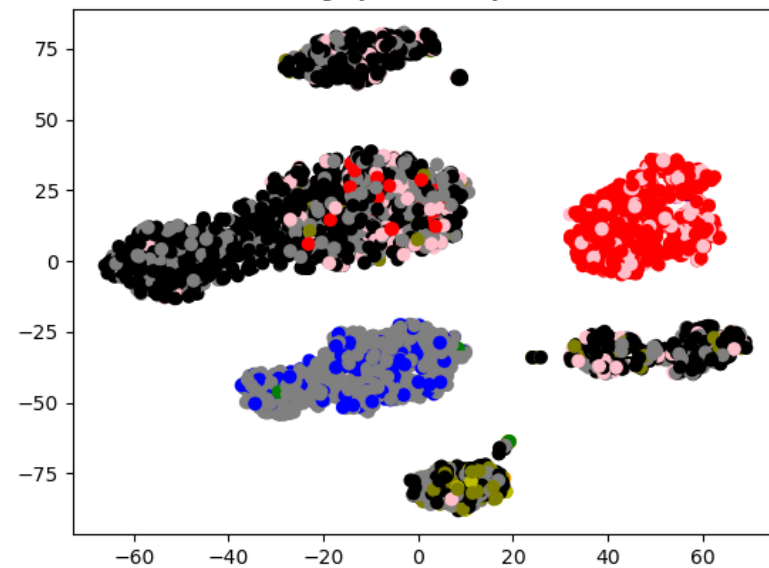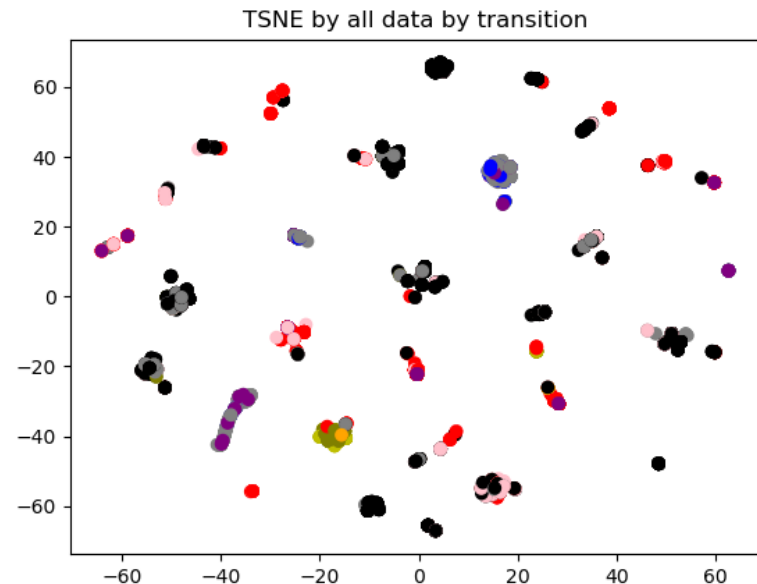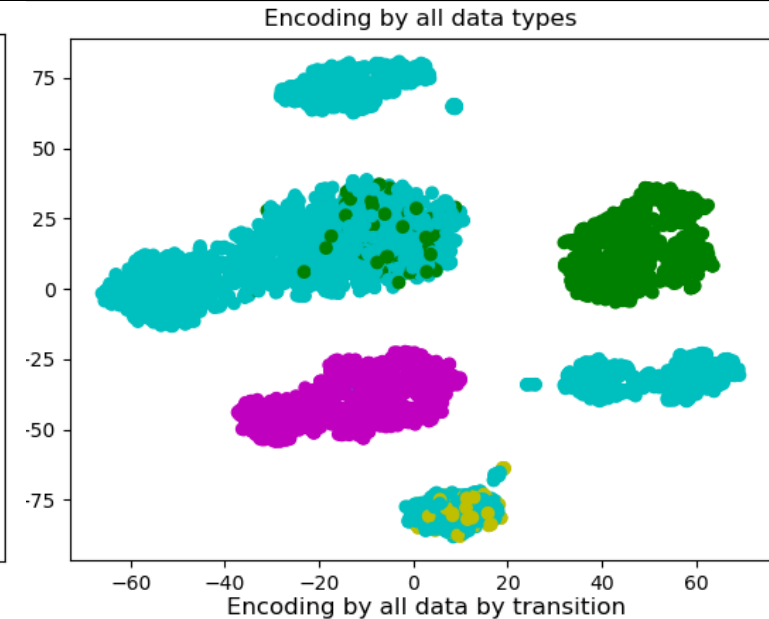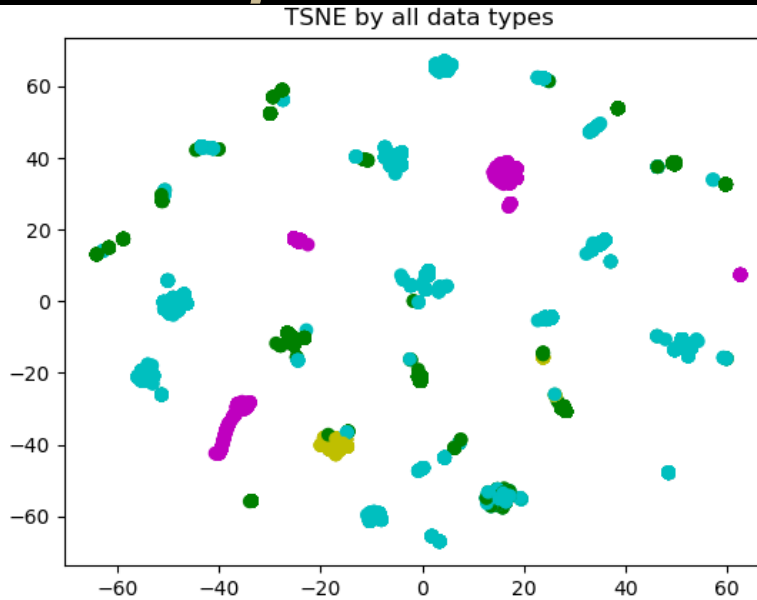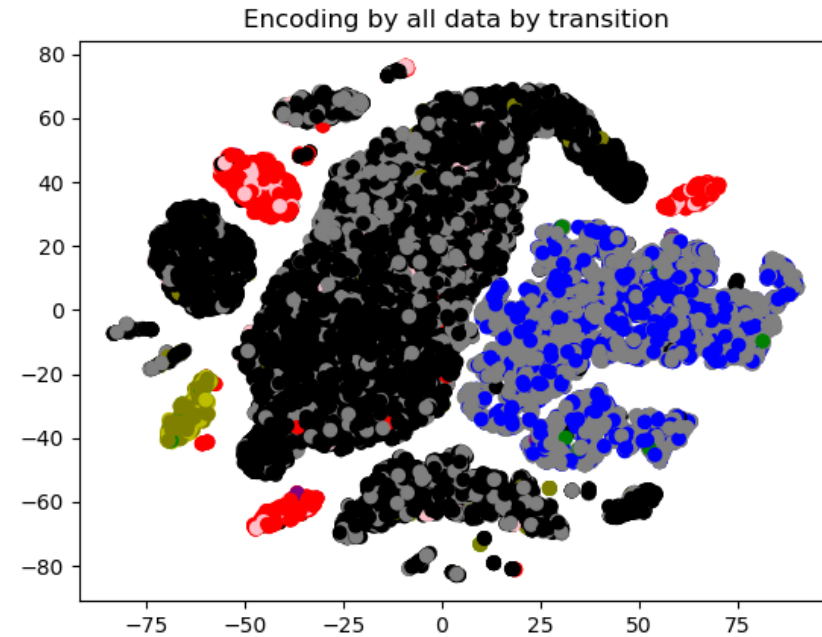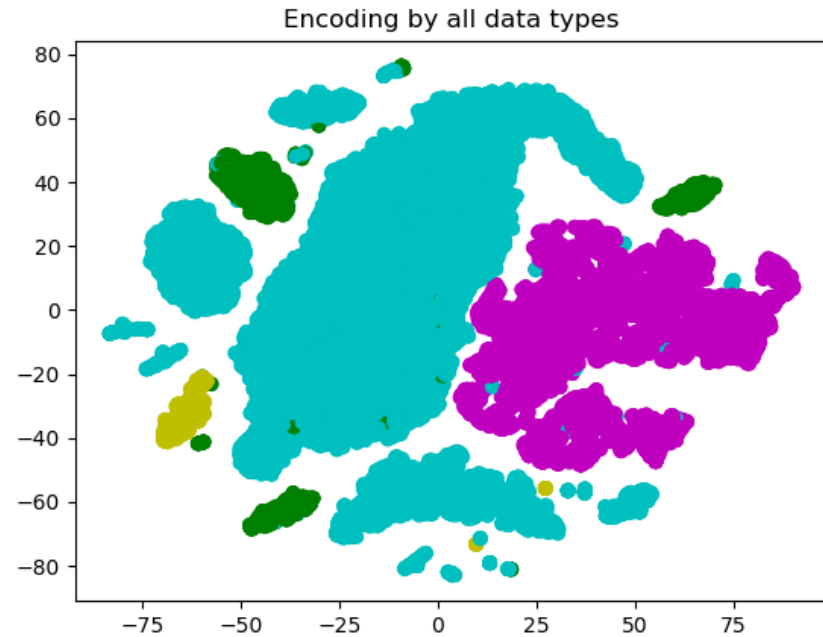
- Practically: Discriminative Step

Encodings

- Traditional Autoencoder
  - Vs
- Variational AE + Contrast

Downstream Uses

- Knowledge Tracing:
  - ACC: 0.7143

Encoding by all data types

Encoding by all data by transition

Future Work:

- Investigating "meaning" of latent variables

- Further refining discriminative step to encourage divergence within clusters

THANK YOU

Questions?