Predictive, **S**calable and **I**nterpretable knowledge tracing on structured domains

Hanqi Zhou, Robert Bamler, Charley M. Wu*, Álvaro Tejero-Cantero* (*equal contribution)

hanqi.zhou@uni-tuebingen.de

Department of Computer Science,

Cluster of Excellence “Machine Learning for Science”, Tübingen AI Center,

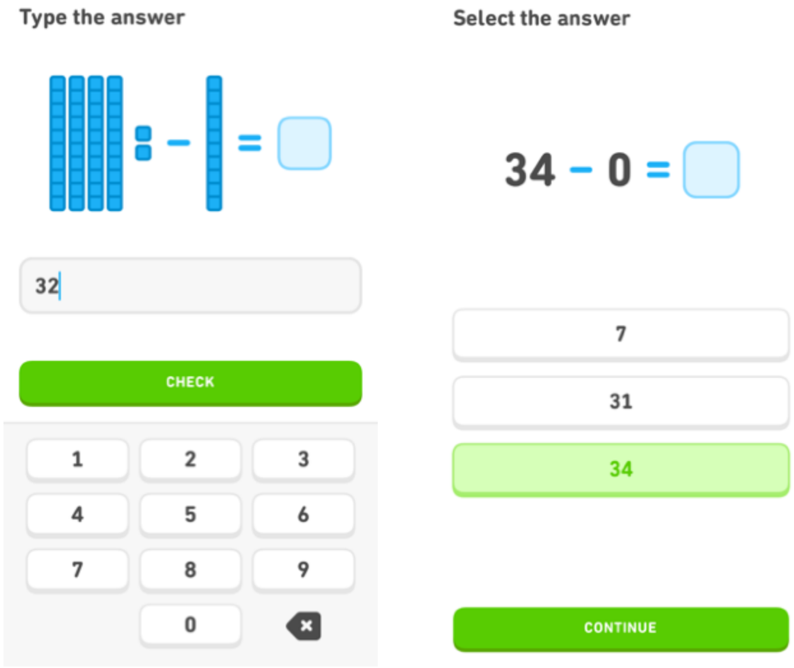
University of Tübingen

12th International Conference on Learning Representations (ICLR), 2024

Tracing knowledge over time

We aim to improve self-directed learning

- by estimating learner knowledge, and providing the right learning materials



source: Duolingo (math learning)

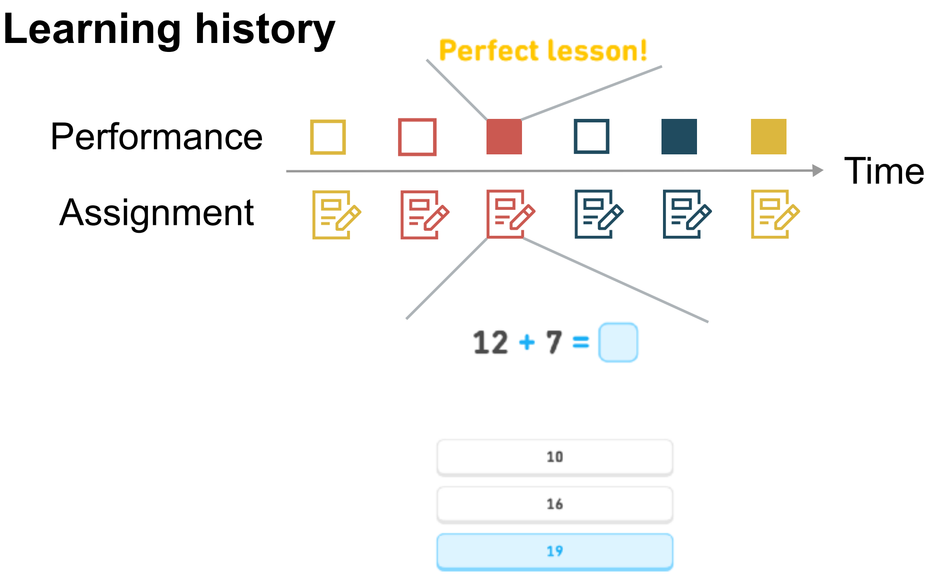
Tracing knowledge over time

- We observe learning histories of each learner

source: Duolingo (math learning)

Tracing knowledge over time

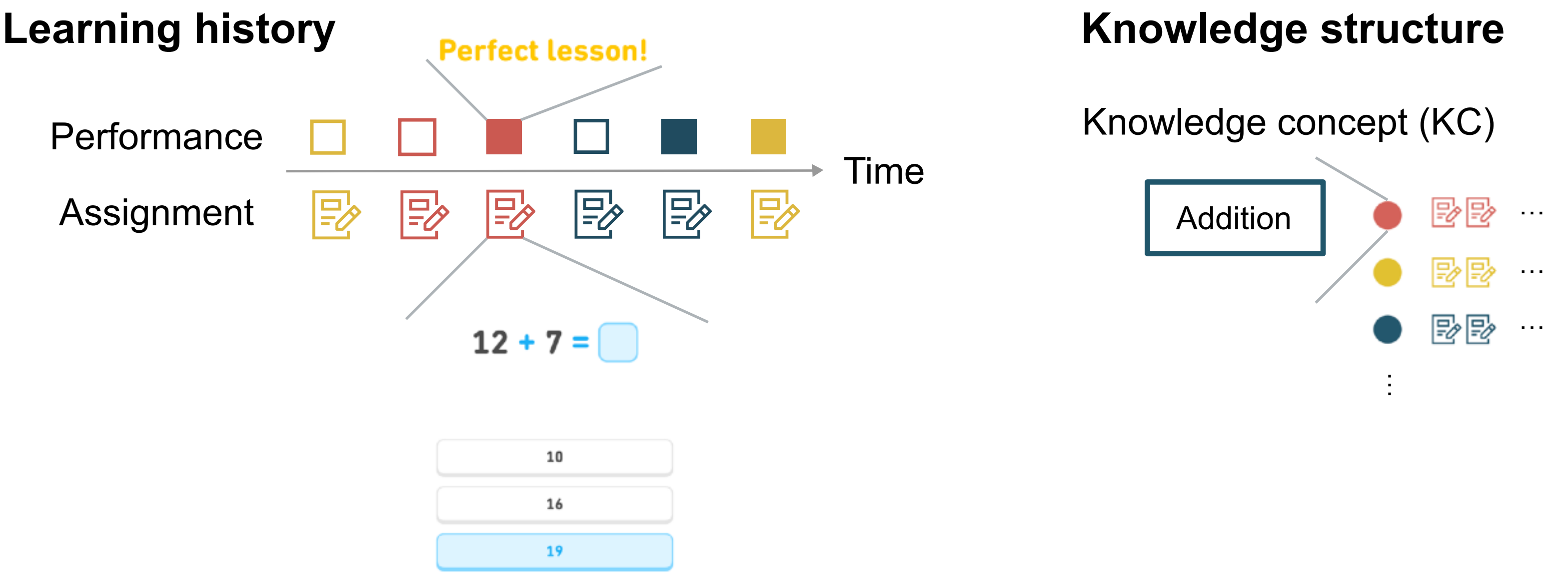
- We observe learning histories of each learner



source: Duolingo (math learning)

Tracing knowledge over time

- We observe learning histories of each learner



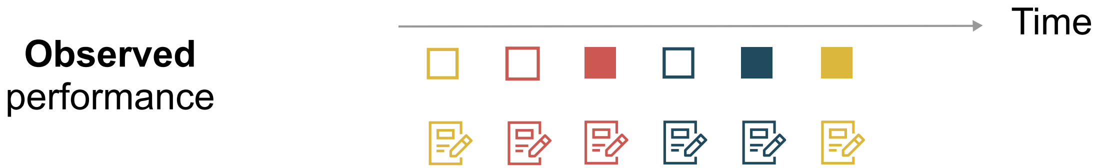
source: Duolingo (math learning)

Tracing knowledge over time

- We observe learning histories of each learner
- We aim to estimate a learner's knowledge states and predict future performance

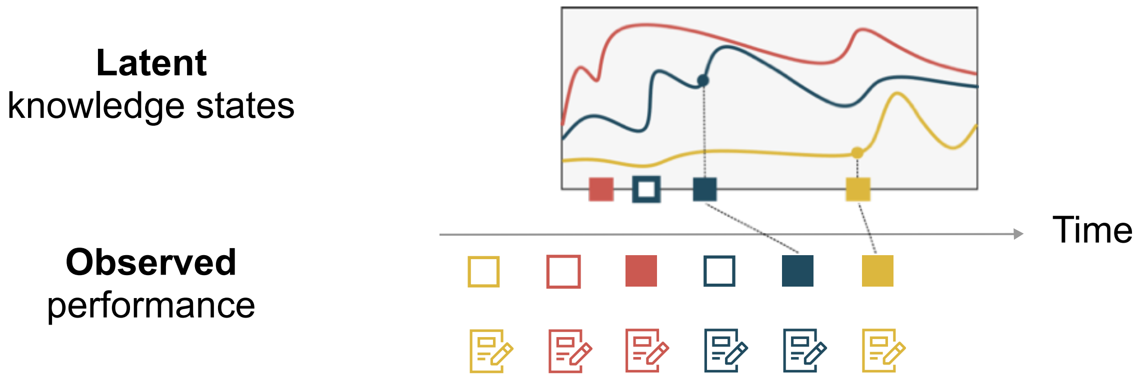
Tracing knowledge over time

- We observe learning histories of each learner
- We aim to estimate a learner's knowledge states and predict future performance



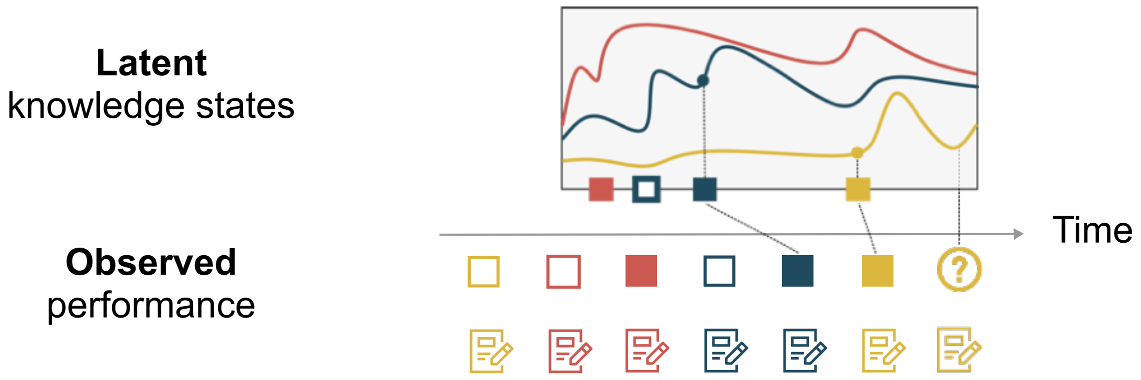
Tracing knowledge over time

- We observe learning histories of each learner
- We aim to estimate a learner's knowledge states and predict future performance



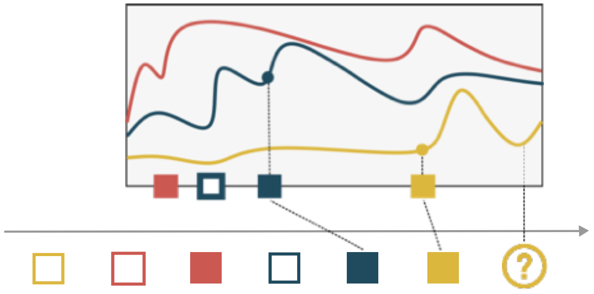
Tracing knowledge over time

- We observe learning histories of each learner
- We aim to estimate a learner's knowledge states and predict future performance



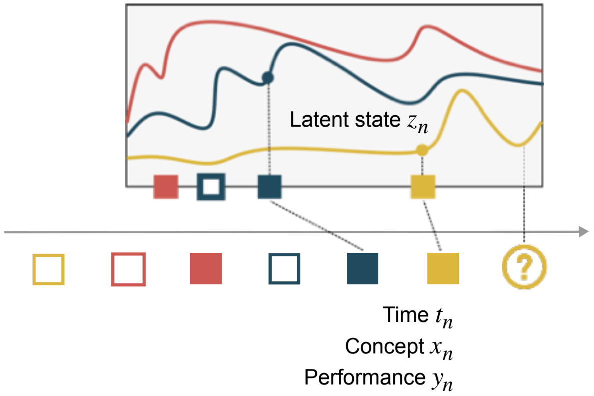
Existing models just can't keep up!

- Learning history $\mathscr{H}^{\ell}_{1:N} := \{t_n, x_n, y_n\}^{l}_{1:N}$ from learner ℓ
  - Interaction time t_n
  - Knowledge concept index $x_n \in \{1,\ldots,K\}$
  - Learner's performance $y_n \in 0,1$



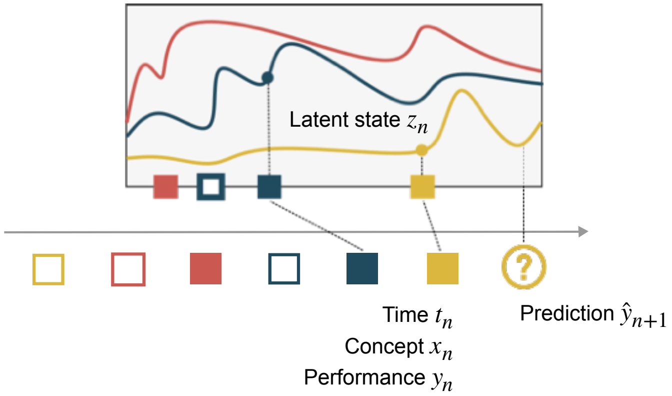
Existing models just can't keep up!

- Learning history $\mathcal{H}^{\ell}_{1:N} := \{t_n, x_n, y_n\}^{l}_{1:N}$ from learner ℓ
  - Interaction time t_n
  - Knowledge concept index $x_n \in \{1,\ldots,K\}$
  - Learner's performance $y_n \in 0,1$
- Latent knowledge states $\mathbf{z}^{\ell}_{1:N} = [z_1^{1:K}, \ldots, z_n^{1:K}]^T$



Existing models just can't keep up!

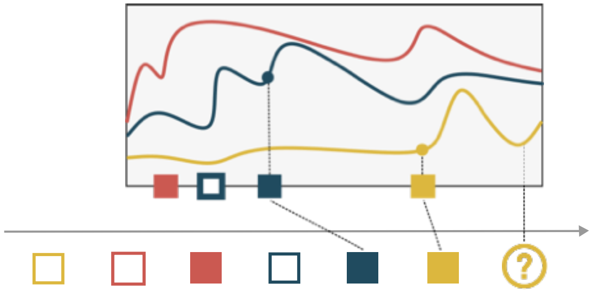
- Learning history $\mathscr{H}^{\ell}_{1:N} := \{t_n, x_n, y_n\}^{l}_{1:N}$ from learner ℓ
  - Interaction time t_n
  - Knowledge concept index $x_n \in \{1,\ldots,K\}$
  - Learner's performance $y_n \in 0,1$
- Latent knowledge states $\mathbf{z}^{\ell}_{1:N} = [z_1^{1:K}, \ldots, z_n^{1:K}]^T$
- Prediction $\hat{y}_{n+1}$
  - $p(y_{n+1} = 1) = \text{sigmoid}(z_{n+1})$



Existing models just can't keep up!

Psychological methods: multi-factor regression

$$f_\theta(z_{n+1} \mid \mathcal{H}_{1:n}, t_{n+1}, x_{n+1})$$
$$= \alpha \cdot \text{property}_{x_{n+1}}$$
$$+ \beta \cdot \text{spacing}_{\mathcal{H}}$$
$$+ \gamma \cdot \text{ability}_{\ell}$$
$$+ \dots$$



Time t_n
Concept x_n
Performance y_n
Latent states z_n
Learner ℓ

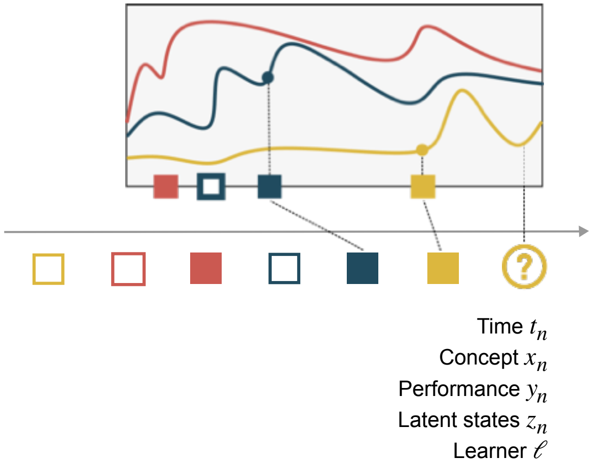
Existing models just can't keep up!

Psychological methods: multi-factor regression

$$
\begin{aligned}
&f_\theta(z_{n+1} \mid \mathscr{H}_{1:n}, t_{n+1}, x_{n+1}) \\
&= \alpha \cdot \boxed{\text{property}_{x_{n+1}}} \\
&+ \beta \cdot \text{spacing}_{\mathscr{H}} \\
&+ \gamma \cdot \text{ability}_{\ell} \\
&+ \dots
\end{aligned}
$$

KC/assignment difficulty

Correct/incorrect frequency



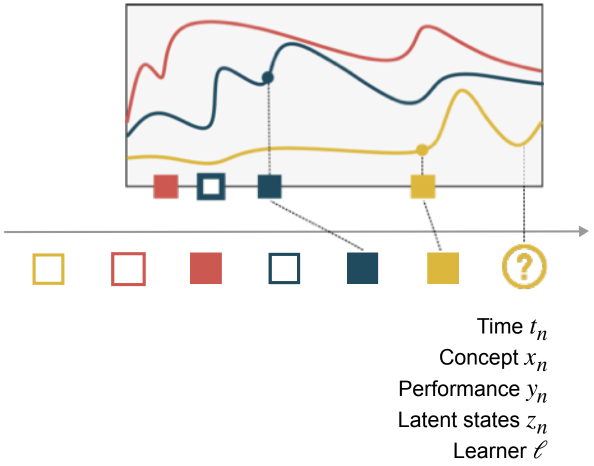
Existing models just can't keep up!

Psychological methods: multi-factor regression

$$f_\theta(z_{n+1} \mid \mathscr{H}_{1:n}, t_{n+1}, x_{n+1})$$
$$= \alpha \cdot \text{property}_{x_{n+1}}$$
$$+\beta \cdot \boxed{\text{spacing}_{\mathscr{H}}}$$
$$+\gamma \cdot \text{ability}_{\ell}$$
$$+\ldots$$

Time duration

Recency effects



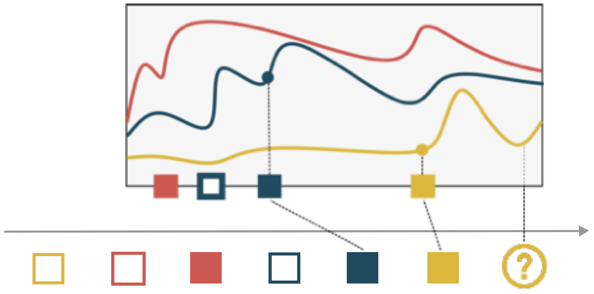
Existing models just can't keep up!

Psychological methods: multi-factor regression

$$f_\theta(z_{n+1} \mid \mathscr{H}_{1:n}, t_{n+1}, x_{n+1})$$
$$= \alpha \cdot \text{property}_{x_{n+1}}$$
$$+\beta \cdot \text{spacing}_{\mathscr{H}}$$
$$+\gamma \cdot \boxed{\text{ability}_\ell}$$
$$+\ldots$$

Memory capacity

Guessing rate



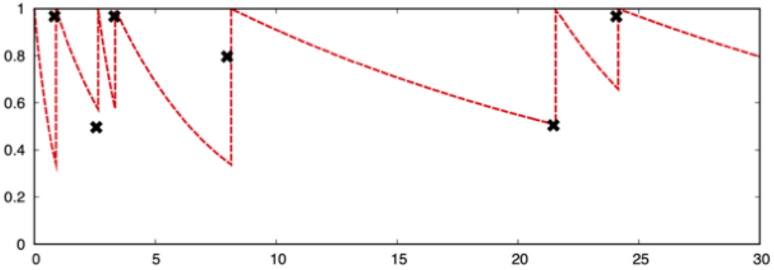
Existing models just can't keep up!

Half-life regression (HLR; Settles & Seeder, 2016; Duolingo)

$$f_\theta(z_{n+1} \mid \mathscr{H}_{1:n}, t_{n+1}, x_{n+1})$$

$$= \alpha \cdot \text{property}_{x_{n+1}}$$ Correct/incorrect frequency

$$+\beta \cdot \text{spacing}_{\mathscr{H}}$$ Time duration



(b) 30-day student-word learning trace and predicted forgetting curve

Settles, B., & Meeder, B. (2016, August). A trainable spaced repetition model for language learning. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers) (pp. 1848-1858).

Existing models just can't keep up!

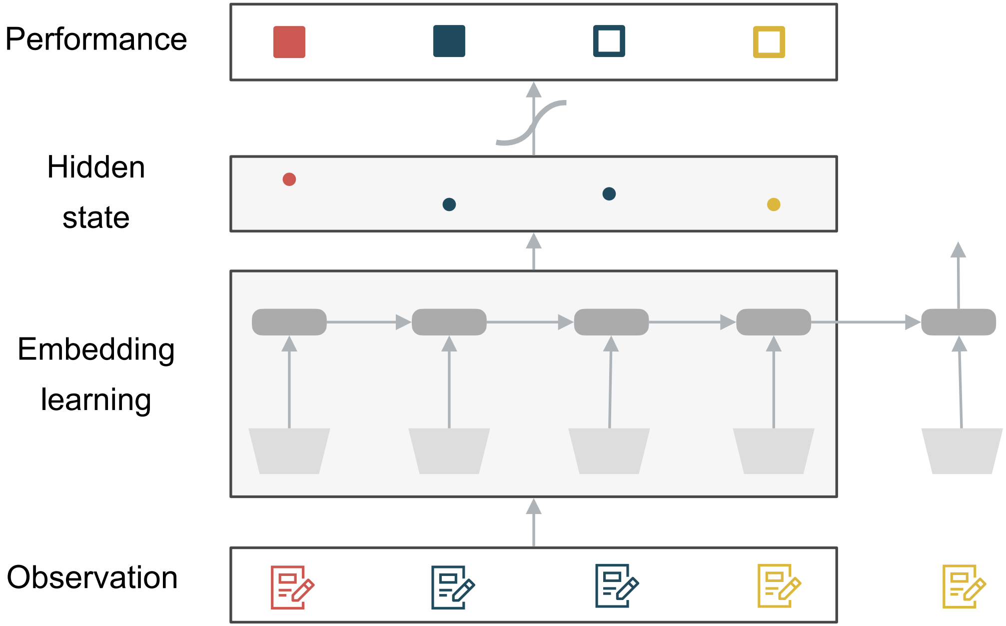
📈 Psychological methods: multi-factor regression

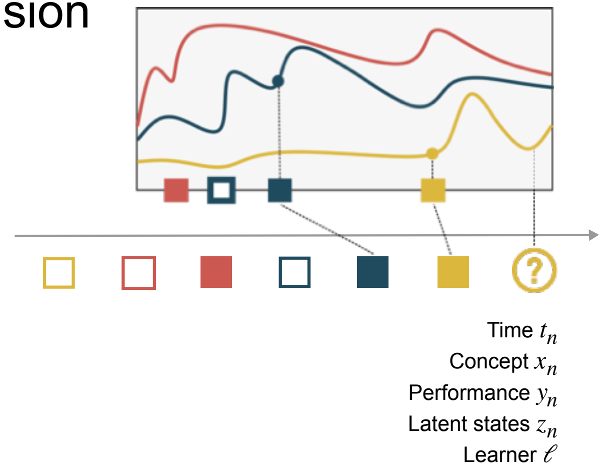
🤩 Diagnosis: explicit modeling of performance factors

😮‍💨 Inflexibility: the amount of parameters increase as the learners/concepts increase

Existing models just can't keep up!

Deep learning models: embedding learning and fusion

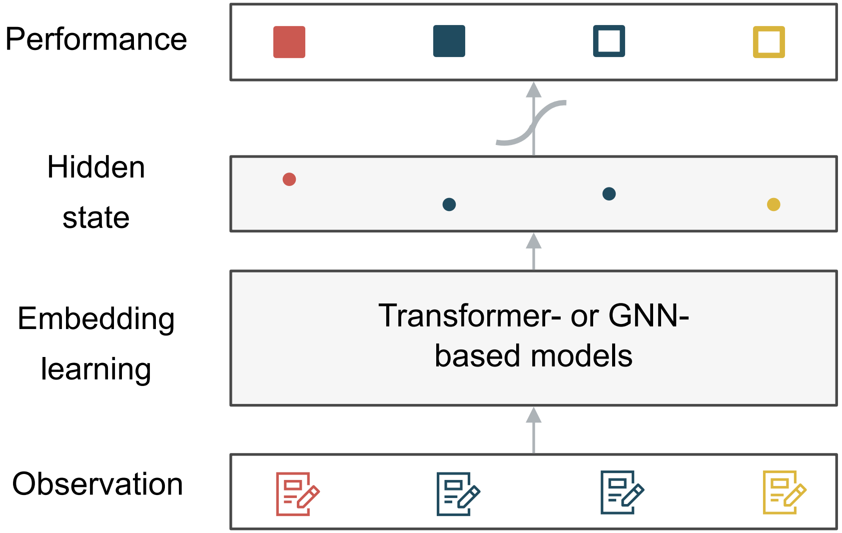


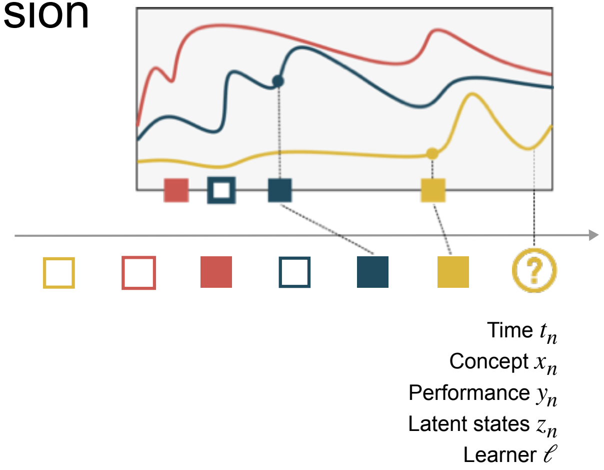


Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. Advances in neural information processing systems, 28.

Existing models just can't keep up!

Deep learning models: embedding learning and fusion





Choi, Y., Lee, Y., Cho, J., Baek, J., Kim, B., Cha, Y., ... & Heo, J. (2020, August). Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the seventh ACM conference on learning@ scale* (pp. 341-344).

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019, October). Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 156-163).

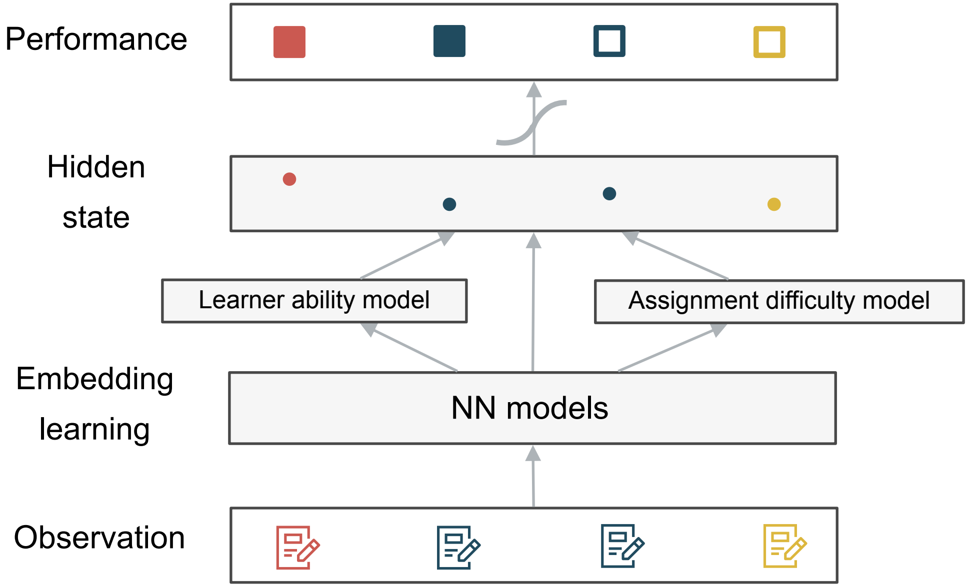
Existing models just can't keep up!

Eberhard Karls Universität Tübingen

- Deep learning models: embedding learning and fusion
- 🤩 High-capacity: can handle high-dimensional feature and large data
- 😮‍💨 Interpretability: but what do these features mean?

Existing models just can't keep up!

- A little more efforts: deep learning + psychology



$$p(y_{n+1}^{\text{correct}} \mid \mathscr{H}_{1:n}, t_{n+1}, x_{n+1})$$
$$= \alpha \cdot \text{property}_{x_{n+1}}$$
$$+\beta \cdot \text{ability}_{\ell}$$
$$+\gamma \cdot \text{spacing}_{\mathscr{H}}$$

QIKT: Chen, J., Liu, Z., Huang, S., Liu, Q., & Luo, W. (2023, June). Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 12, pp. 14196-14204).

Motivation of Probabilistic Generative Model: PSI-KT

- Computational modeling  :
  - A flexible architecture with
  - Identifiable interpretability

Motivation of Probabilistic Generative Model: PSI-KT

- Computational modeling :
  - A flexible architecture with
  - Identifiable interpretability

- Real-world scenario:
  - Small data regime and real-time adaptation

How do humans learn in structured domains?

How do humans learn in structured domains?

- Human learners **forget** over time but reinforce knowledge through practice
- Prerequisites **structure** the domain by relating KCs

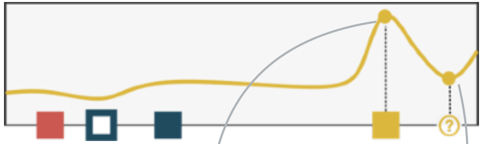
How do humans learn in structured domains?

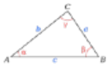
- Human learners **forget** over time but reinforce knowledge through practice
- Prerequisites **structure** the domain by relating KCs

Learning process

Knowledge states

Performance





$$c^2 = a^2 + b^2 - 2ab\cos\gamma$$

$$c^2 = a^2 + b^2 - \boxed{\ ?\ }$$

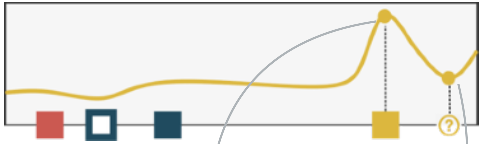
How do humans learn in structured domains?

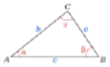
- Human learners **forget** over time but **reinforce** knowledge through practice
- Prerequisites **structure** the domain by relating concepts

Learning process

Knowledge states

Performance

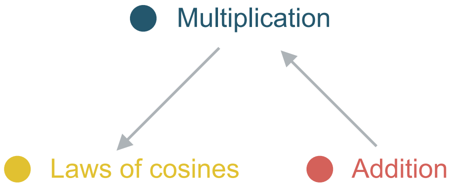




$$c^2 = a^2 + b^2 - 2ab\cos\gamma$$

$$c^2 = a^2 + b^2 - \boxed{\ ?\ }$$

Prerequisites



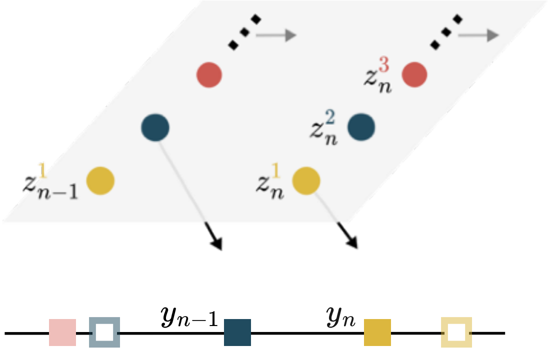
We jointly model learner traits and prerequisites

Observed performance y

We jointly model learner traits and prerequisites

Latent
knowledge states z

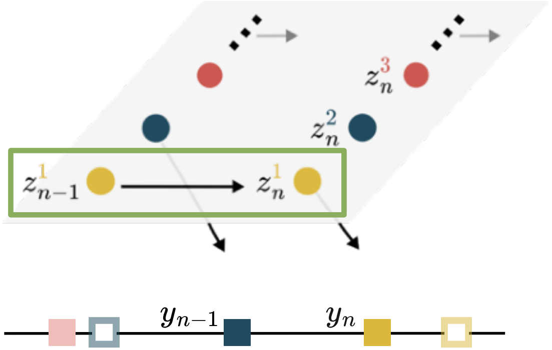
Observed
performance y



We jointly model learner traits and prerequisites

Latent knowledge states z

Observed performance y

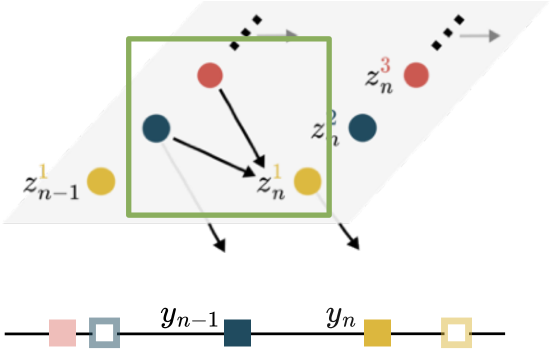


$$z_n^1 = f(z_{n-1}^1; \theta_s)$$

We jointly model learner traits and prerequisites

Latent
knowledge states z

Observed
performance y



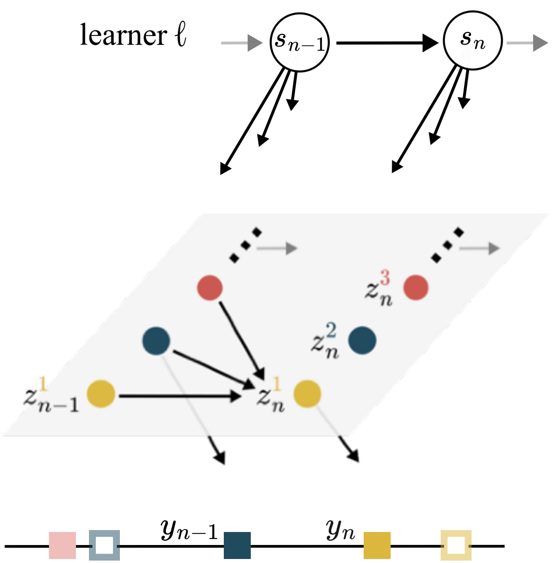
$$z^1_n = f(z^2_{n-1}, z^3_{n-1}; \theta_G)$$

We jointly model learner traits and prerequisites

Latent
cognitive traits s

Latent
knowledge states z

Observed
performance y

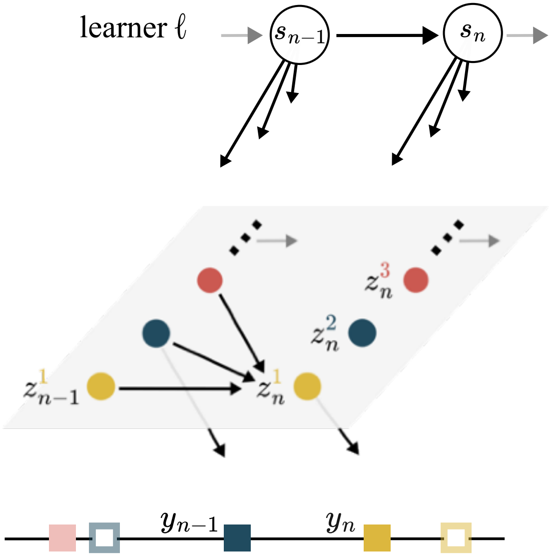


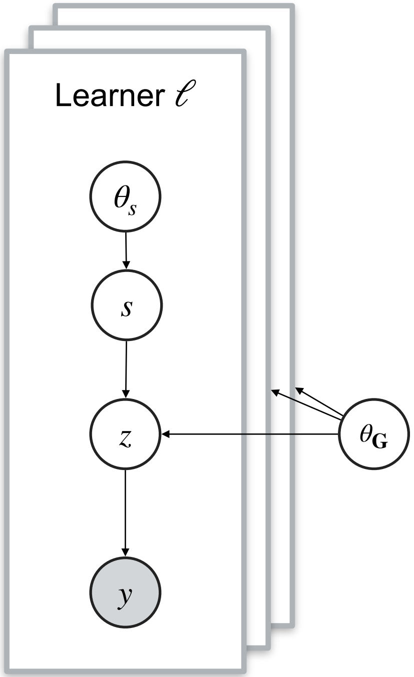
We jointly model learner traits and prerequisites

Latent
cognitive traits s

Latent
knowledge states z

Observed
performance y



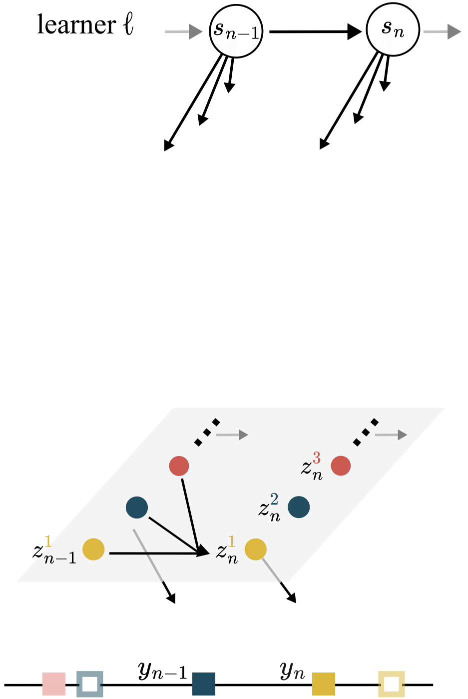


Generative model: cognitive traits

Latent cognitive traits s

Latent knowledge states z

Observed performance y

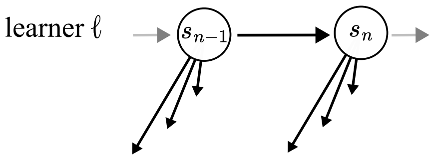


Per learner $s^\ell_n := (\alpha^\ell_n, \mu^\ell_n, \gamma^\ell_n, \sigma^\ell_n)$ for personalization

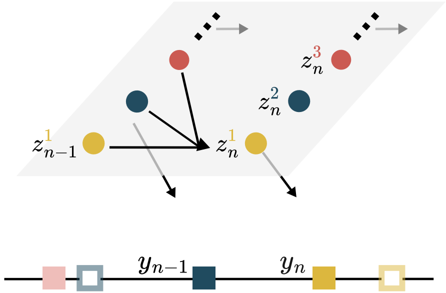
- Memory: forgetting rate α^ℓ_n, long-term consolidation μ^ℓ_n
- Structure: transfer ability γ^ℓ_n
- Noise: knowledge volatility σ^ℓ_n

Generative model: cognitive traits

Latent cognitive traits s



Latent knowledge states z



Observed performance y

Per learner $s_n^\ell := (\alpha_n^\ell, \mu_n^\ell, \gamma_n^\ell, \sigma_n^\ell)$ for personalization

- Memory: forgetting rate α_n^ℓ, long-term consolidation μ_n^ℓ
- Structure: transfer ability γ_n^ℓ
- Noise: knowledge volatility σ_n^ℓ

Evolution:

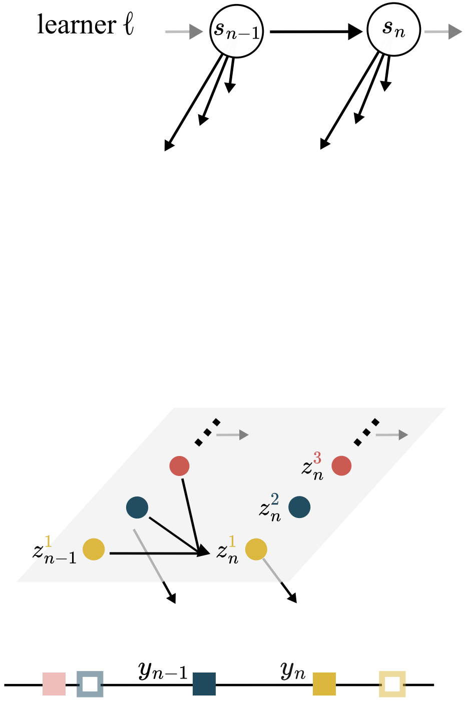
- $s_n^\ell \sim p_{\theta_s}\left(s_n^\ell \mid s_{n-1}^\ell\right) := \mathcal{N}\left(s_n^\ell \mid H s_{n-1}^\ell, R\right)$

Generative model: cognitive traits

Latent cognitive traits s

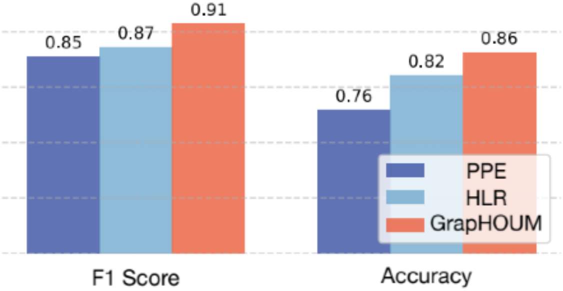
Latent knowledge states z

Observed performance y



Per learner $s_n^\ell := (\alpha_n^\ell, \mu_n^\ell, \gamma_n^\ell, \sigma_n^\ell)$ for personalization

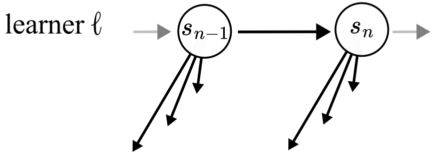
- Memory: forgetting rate α_n^ℓ, long-term consolidation μ_n^ℓ
- Structure: transfer ability γ_n^ℓ
- Noise: knowledge volatility σ_n^ℓ



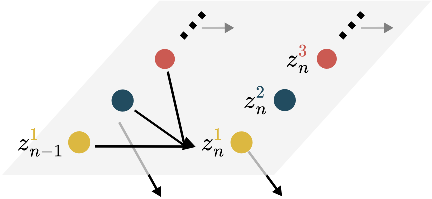
Zhou, H., Tejero-Cantero, A., & Wu, C. M (2023 CCN). The Dynamic and Structured Nature of Learning and Memory.

Generative model: knowledge states

Latent cognitive traits s



Latent knowledge states z



Observed performance y

Per learner $s_n^\ell := (\alpha_n^\ell, \mu_n^\ell, \gamma_n^\ell, \sigma_n^\ell)$ for personalization

- Memory: forgetting rate α_n^ℓ, long-term consolidation μ_n^ℓ
- Structure: transfer ability γ_n^ℓ
- Noise: knowledge volatility σ_n^ℓ

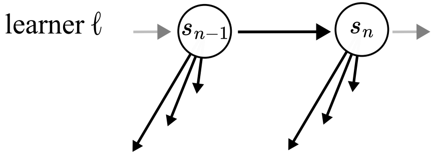
Per learner & KC $z_n^{\ell,k}$ for learning dynamics

Ornstein-Uhlenbeck process

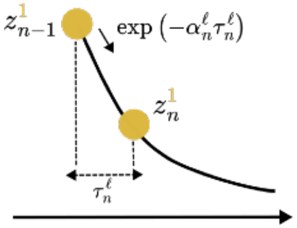
$$\mathrm{d}z^{\ell,k}/\mathrm{d}t = \alpha^\ell \left(\mu^\ell - z^{\ell,k}\right) + \sigma^\ell \eta(t)$$

Generative model: knowledge states

Latent cognitive traits s



Latent knowledge states z



Observed performance y

Per learner $s^\ell_n := (\alpha^\ell_n, \mu^\ell_n, \gamma^\ell_n, \sigma^\ell_n)$ for personalization

- Memory: forgetting rate α^ℓ_n, long-term consolidation μ^ℓ_n
- Structure: transfer ability γ^ℓ_n
- Noise: knowledge volatility σ^ℓ_n

Per learner & KC $z^{\ell,k}_n$ for learning dynamics

$$\mathrm{d}z^{\ell,k}/\mathrm{d}t = \alpha^\ell \left(\mu^\ell - z^{\ell,k}\right) + \sigma^\ell \eta(t)$$

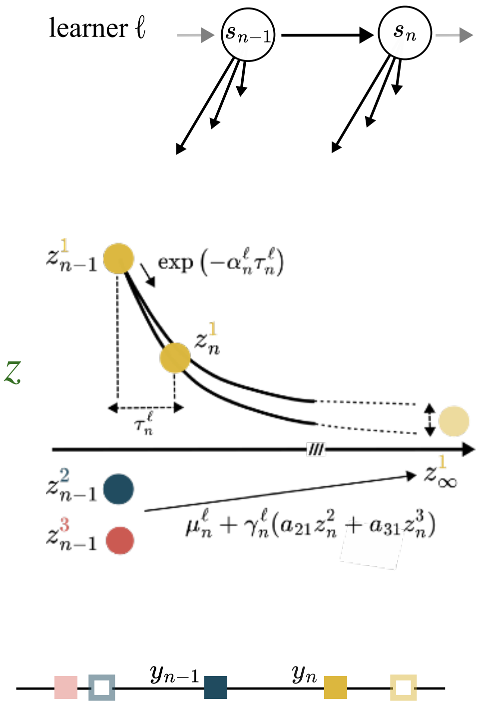
- Short-term: exponential decay
  - $z^{\ell,k}_n = z^{\ell,k}_{n-1} \exp\left(-\alpha^\ell_n \tau^\ell_n\right)$

Generative model: knowledge states

Latent cognitive traits s

Latent knowledge states z

Observed performance y



Per learner $s_n^\ell := (\alpha_n^\ell, \mu_n^\ell, \gamma_n^\ell, \sigma_n^\ell)$ for personalization

- Memory: forgetting rate α_n^ℓ, long-term consolidation μ_n^ℓ
- Structure: transfer ability γ_n^ℓ
- Noise: knowledge volatility σ_n^ℓ

Per learner & KC $z_n^{\ell,k}$ for learning dynamics

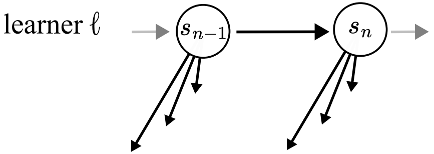
$$\mathrm{d}z^{\ell,k}/\mathrm{d}t = \alpha^\ell \left(\mu^\ell - z^{\ell,k}\right) + \sigma^\ell \eta(t)$$

- Short-term: exponential decay
  - $z_n^{\ell,k} = z_{n-1}^{\ell,k} \exp\left(-\alpha_n^\ell \tau_n^\ell\right)$
- Long-term: shifted by global structure
  - $\tilde{\mu}_n^{\ell,k} := \mu_n^\ell + \gamma_n^\ell \sum_{i \neq k} a_{ik} z_n^{\ell,i}, \qquad a_{ik} \in \theta_G$

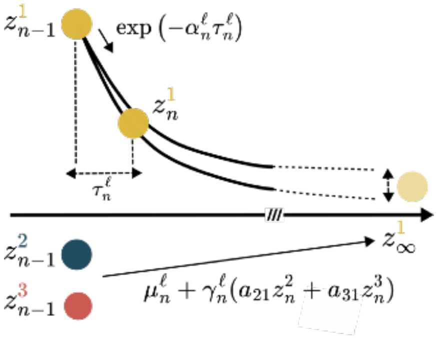
Generative model: knowledge states

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Latent cognitive traits s



Latent knowledge states z



Observed performance y

Per learner $s_n^\ell := (\alpha_n^\ell, \mu_n^\ell, \gamma_n^\ell, \sigma_n^\ell)$ for personalization

- Memory: forgetting rate α_n^ℓ, long-term consolidation μ_n^ℓ
- Structure: transfer ability γ_n^ℓ
- Noise: knowledge volatility σ_n^ℓ

Per learner & KC $z_n^{\ell,k}$ for learning dynamics

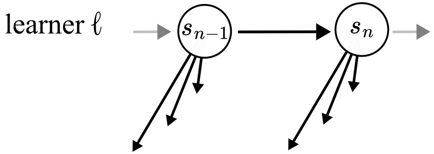
- Short-term: exponential decay
  - $dz^{\ell,k}/dt = \alpha^\ell \left(\mu^\ell - z^{\ell,k}\right) + \sigma^\ell \eta(t)$
- Long-term: shifted by global structure
  - $\tilde{\mu}_n^{\ell,k} := \mu_n^\ell + \gamma_n^\ell \sum_{i \neq k} a_{ik} z_n^{\ell,i}, \qquad a_{ik} \in \theta_G$

Evolution:

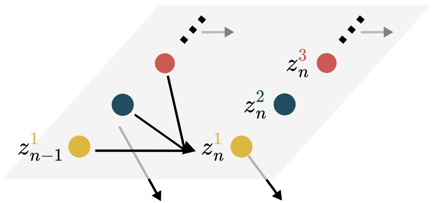
- $z_n^{\ell,k} = \underbrace{z_{n-1}^{\ell,k} \exp\left(-\alpha_n^\ell \tau_n^\ell\right)}_{\text{short-term dynamics}} + \underbrace{\tilde{\mu}_n^{\ell,k}\left(1 - \exp\left(-\alpha_n^\ell \tau_n^\ell\right)\right)}_{\text{long-term dynamics}}$

Generative model: observations

Latent cognitive traits s



Latent knowledge states z



Observed performance y

Per learner $s_n^\ell := (\alpha_n^\ell, \mu_n^\ell, \gamma_n^\ell, \sigma_n^\ell)$ for personalization

- Memory: forgetting rate α_n^ℓ, long-term consolidation μ_n^ℓ
- Structure: transfer ability γ_n^ℓ
- Noise: knowledge volatility σ_n^ℓ

Per learner & KC $z_n^{\ell,k}$ for learning dynamics

- Transient: $\mathrm{d}z^{\ell,k}/\mathrm{d}t = \alpha^\ell \left(\mu^\ell - z^{\ell,k}\right) + \sigma^\ell \eta(t)$
- Long-term: $\tilde{\mu}_n^{\ell,k\dagger} := \mu_n^\ell + \gamma_n^\ell \sum_{i \neq k} a_{ik} z_n^{\ell,i}$

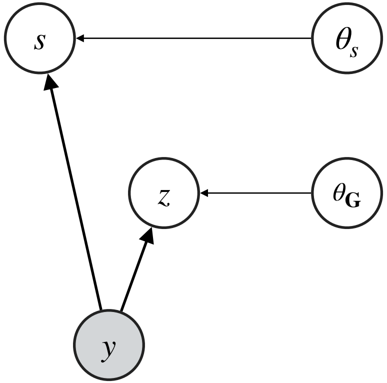
Observation

- Emission: $\hat{y}_n^\ell \sim p\left(y_n^\ell \mid z_n^{\ell,k}\right) := \mathrm{Bern}\left(\mathrm{sigmoid}\left(z_n^{\ell,k}\right)\right)$

Inference model: (continual) variational inference

Exact inference over latent variables

- Full distribution over latents
- Point estimation over parameters

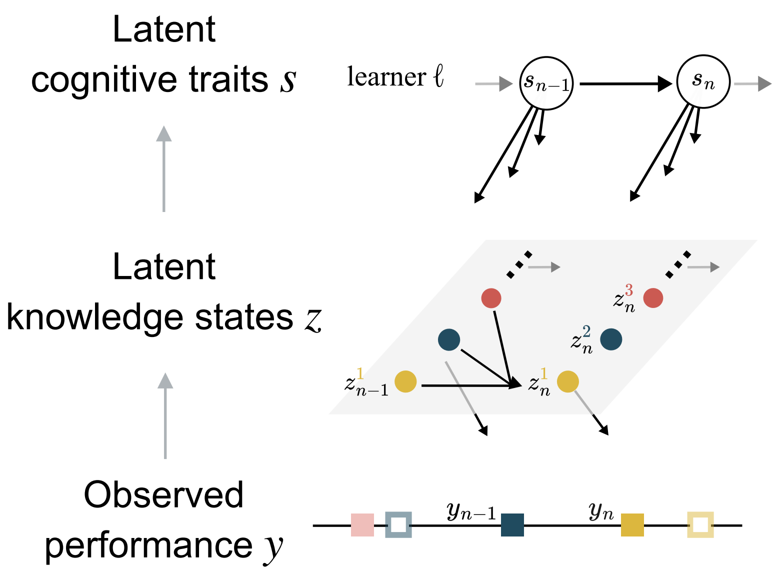


Inference model: (continual) variational inference

Exact inference over latent variables 🤯

$$p_\theta(s_{1:n}, \mathbf{z}_{1:n} \mid y_{1:n})$$

$$= \frac{p_\theta(s_{1:n}, \mathbf{z}_{1:n}, y_{1:n})}{\int_{s_{1:n}} \int_{\mathbf{z}_{1:n}} p_\theta(s_{1:n}, \mathbf{z}_{1:n}, y_{1:n})}$$



Inference model: (continual) variational inference

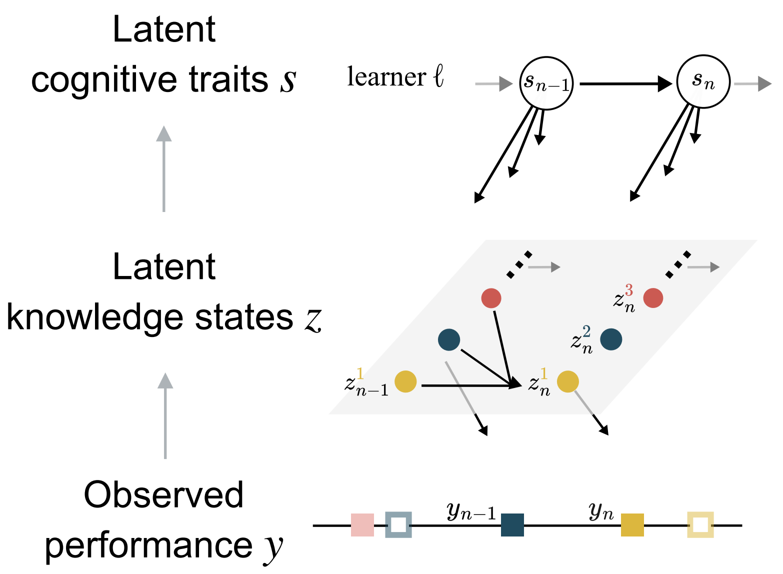
We need help from NN-based

Approximate Bayesian Inference 😎

$$q_\phi\left(z_{1:n}, s_{1:n} \mid y_{1:n}\right) = q_\phi\left(z_{1:n}\right) q_\phi\left(s_{1:n}\right)$$

$$\updownarrow$$

$$p_\theta(s_{1:n}, \mathbf{z}_{1:n} \mid y_{1:n})$$

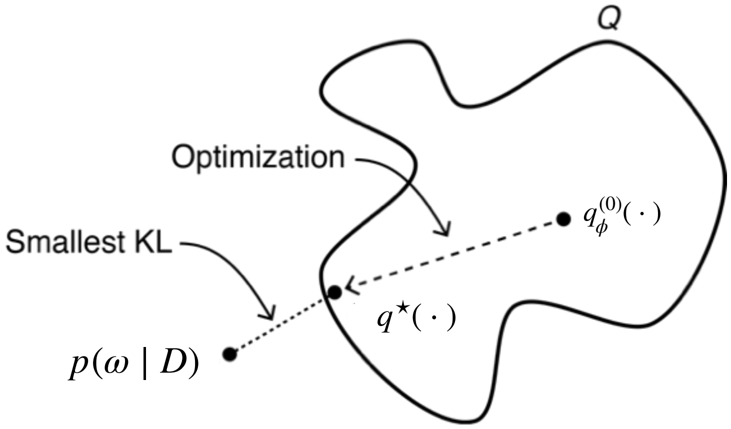


Inference model: (continual) variational inference

We need help from NN-based

Approximate Bayesian Inference 😎

$$q^\star(\omega) = \arg\min_{q(\cdot)\in\mathcal{Q}} \mathrm{KL}[q_\phi(\omega) \| p_\theta(\omega \mid D)]$$



Latent states $\omega := \{s_{1:n}, \mathbf{z}_{1:n}\}$

Observation $D := \mathcal{H}_{1:n}$

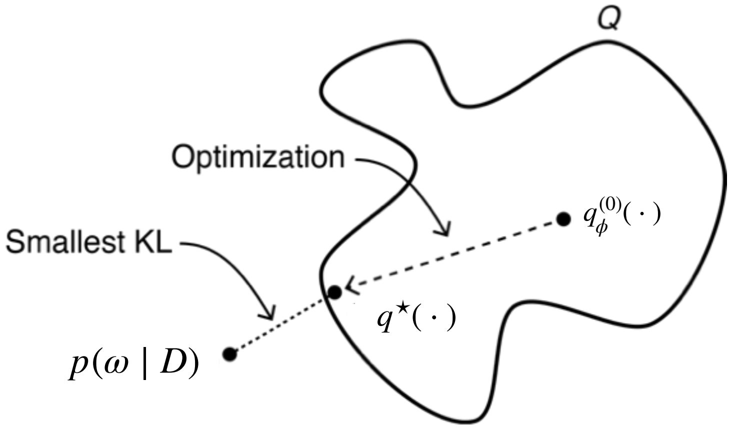
Figure source: https://gregorygundersen.com/blog/2021/04/16/variational-inference/

Inference model: (continual) variational inference

We need help from NN-based

Approximate Bayesian Inference 😎

$$q^\star(\omega) = \arg\min_{q(\cdot)\in\mathcal{Q}} \mathrm{KL}[q_\phi(\omega) \| p_\theta(\omega \mid D)]$$



$$q^\star(\mathbf{z}_{1:n}) q^\star(s_{1:n}) = \arg\min_{q(\cdot)\in\mathcal{Q}} \mathrm{KL}[q_\phi(\mathbf{z}_{1:n}) q_\phi(s_{1:n}) \| p_\theta(s_{1:n}, \mathbf{z}_{1:n} \mid y_{1:n})]$$

Latent states $\omega := \{s_{1:n}, \mathbf{z}_{1:n}\}$

Observation $D := \mathscr{H}_{1:n}$

Figure source: https://gregorygundersen.com/blog/2021/04/16/variational-inference/

Inference model: (continual) variational inference

- ELBO

$$\mathrm{KL}\left[q_\phi(\omega) \| p_\theta(\omega \mid D)\right] = \int_\omega q_\phi(\omega) \log \frac{q_\phi(\omega)}{p_\theta(\omega \mid D)} = \mathbb{E}_q\left[\log \frac{q_\phi(\omega)}{p_\theta(\omega \mid D)}\right]$$

$$= \underbrace{-\mathbb{E}_q[\log p_\theta(D \mid \omega)] + \mathbb{E}_q[\log \frac{q_\phi(\omega)}{p_\theta(\omega)}]}_{-\mathrm{ELBO}} + \log p(D)$$

$$L_{\mathrm{ELBO}}(\phi, \theta) = \mathbb{E}_{q_\phi(\omega)}\left[\log p_\theta(D \mid \omega)\right] - \mathrm{KL}\left[q_\phi(\omega) \| p_\theta(\omega)\right]$$

Latent states $\omega := \{s_{1:n}, \mathbf{z}_{1:n}\}$

Observation $D := \mathscr{H}_{1:n}$

Inference model: (continual) variational inference

- ELBO for fixed learning histories

$$L_{\text{ELBO}}(\phi, \theta) = \mathbb{E}_{q_\phi(\omega)}\left[\underline{\log p_\theta(D \mid \omega)}\right] - \underline{\text{KL}\left[q_\phi(\omega) \| p_\theta(\omega)\right]}$$

$$= \mathbb{E}_{q_\phi(z_{1:n}, s_{1:n})}\left[\underline{\log p_\theta(y_{1:n} \mid z_{1:n}, s_{1:n})} \underline{- \log(q_\phi(z_{1:n}, s_{1:n}) - p_\theta(z_{1:n}, s_{1:n} \mid z_0, s_0))}\right]$$

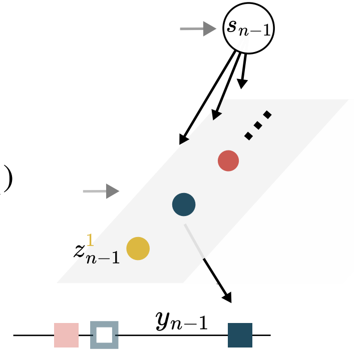
Inference model: (continual) variational inference

- ELBO for online data

Cognitive traits $s_{n-1} \sim q_{\phi_{n-1}}(s_{n-1} \mid y_{1:n-1})$

Knowledge states $\mathbf{z}_{n-1} \sim q_{\phi_{n-1}}(\mathbf{z}_{n-1} \mid y_{1:n-1})$

Performance y_{n-1}



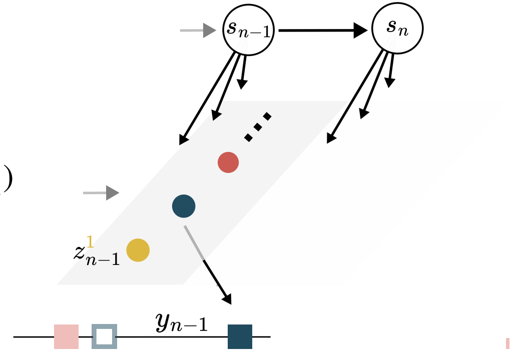
Inference model: (continual) variational inference

- ELBO for online data

Cognitive traits $s_{n-1} \sim q_{\phi_{n-1}}(s_{n-1} \mid y_{1:n-1})$

Knowledge states $\mathbf{z}_{n-1} \sim q_{\phi_{n-1}}(\mathbf{z}_{n-1} \mid y_{1:n-1})$

Performance y_{n-1}



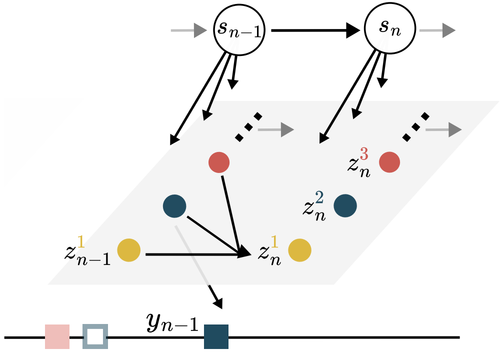
Inference model: (continual) variational inference

- ELBO for online data

Cognitive traits $s_n \sim p_{\theta_{n-1}}(s_n \mid s_{n-1})$

Knowledge states $\mathbf{z}_n \sim p_{\theta_{n-1}}(\mathbf{z}_n \mid \mathbf{z}_{n-1}, s_n)$

Performance y_{n-1}



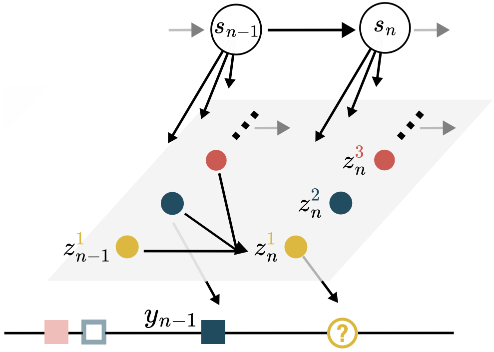
Inference model: (continual) variational inference

- ELBO for online data

Cognitive traits $s_n \sim p_{\theta_{n-1}}(s_n \mid s_{n-1})$

Knowledge states $\mathbf{z}_n \sim p_{\theta_{n-1}}(\mathbf{z}_n \mid \mathbf{z}_{n-1}, s_n)$

Performance $\hat{y}_n$



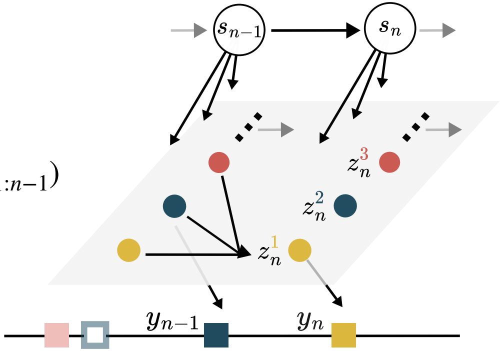
Inference model: (continual) variational inference

- ELBO for online data

Cognitive traits $p_{\theta_{n-1}}(s_n \mid s_{n-1})\ q_{\phi_{n-1}}(s_n \mid y_{1:n})$

Knowledge states $p_{\theta_{n-1}}(\mathbf{z}_n \mid \mathbf{z}_{n-1}, s_n)\ q_{\phi_{n-1}}(\mathbf{z}_{n-1} \mid y_{1:n-1})$

Performance y_n comes



Inference model: (continual) variational inference

- ELBO for online data

$$L_{\mathrm{ELBO}}(\phi, \theta) = \mathbb{E}_{q_\phi(\omega_{1:n})}\left[\underline{\log p_\theta(D_{1:n} \mid \omega_{1:n})}\right] - \mathrm{KL}\left[\underline{q_\phi(\omega_{1:n}) \| p_\theta(\omega_{1:n})}\right]$$

$$L_{\mathrm{ELBO}}(\phi_n, \theta_n) = \mathbb{E}_{q_{\phi_n}(\omega_n)}\left[\underline{\log p_{\theta_n}(D_n \mid \omega_n)}\right] - \mathrm{KL}\left[\underline{q_{\phi_n}(\omega_n) \| q_{\phi_{n-1}}(\omega_{n-1})}\right]$$

Overall

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Latent inference:

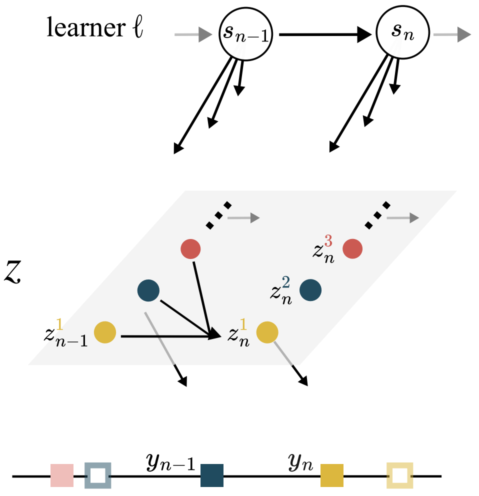
Online variational inference

$$L_{\text{ELBO}}(\phi,\theta) = \mathbb{E}_{q_\phi(\omega)}\left[\log p_\theta(D \mid \omega)\right] - \mathrm{KL}\left[q_\phi(\omega) \| p_\theta(\omega)\right]$$

Latent cognitive traits s

Latent knowledge states z

Observed performance y



Per learner $s_n^\ell := (\alpha_n^\ell, \mu_n^\ell, \gamma_n^\ell, \sigma_n^\ell)$

- forgetting rate α_n^ℓ, long-term consolidation μ_n^ℓ
- transfer ability γ_n^ℓ
- knowledge volatility σ_n^ℓ

Per learner & KC $z_n^{\ell,k}$

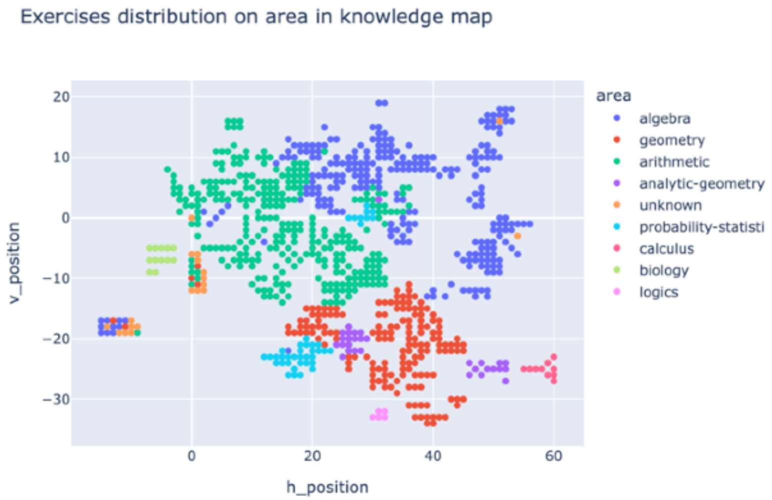
- $\mathrm{d}z^{\ell,k}/\mathrm{d}t = \alpha^\ell\left(\mu^\ell - z^{\ell,k}\right) + \sigma^\ell \eta(t)$
- $\tilde{\mu}_n^{\ell,k^\dagger} := \mu_n^\ell + \gamma_n^\ell \sum_{i \neq k} a_{ik} z_n^{\ell,i}$

Observation

- Emission:
  $\hat{y}_n^\ell \sim p\left(y_n^\ell \mid z_n^{\ell,k}\right) := \mathrm{Bern}\left(\mathrm{sigmoid}\left(z_n^{\ell,k}\right)\right)$

KT datasets

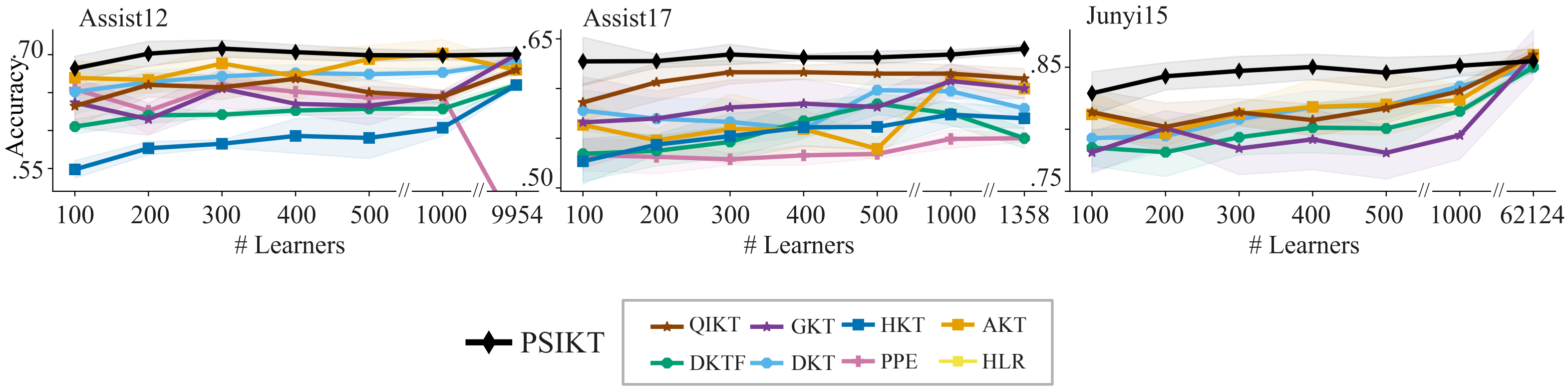
- Assistment 2012
- Assistment 2017
- Junyi 2015
  - Pre-college mathematics study



	user_id	exercise	problem_type	problem_number	topic_mode	suggested	review_mode	time_done	time_taken	time_taken_attempts	correct
0	12884	time_terminology	analog_word	1	False	False	False	1420714810324490	4	3&1	False
1	239464	multiplication_1	0	6	False	False	False	1403098400836660	2	2	True
2	147359	adding_decimals_0.5	0	6	False	False	False	1418890695540340	16	16	True
3	158155	multiplication_1	0	3	False	False	False	1400469444264040	2	2	True
4	147151	subtraction_2	subtraction-2	10	True	True	False	1382650905730160	4	4	True

PSI: Achieving predictive accuracy on limited data

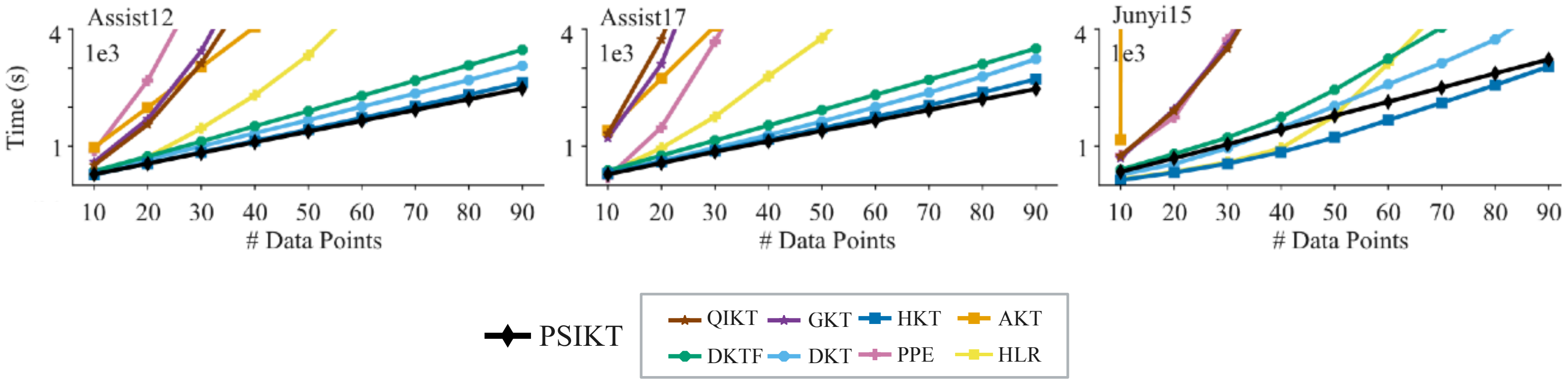
- Within-learner 10-step prediction performance as a function of cohort sizes



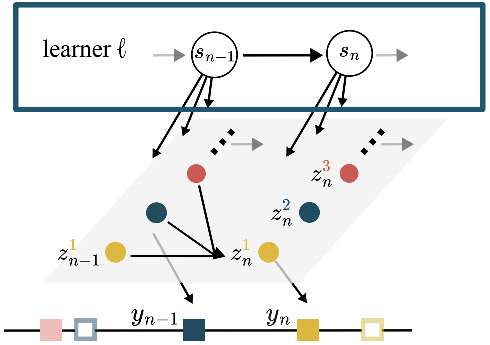
PSI: Scaling efficiently with more interactions

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- Cumulative training time of continual learning



- **Cognitive traits:** forgetting rate, consolidation memory, transfer ability, volatility

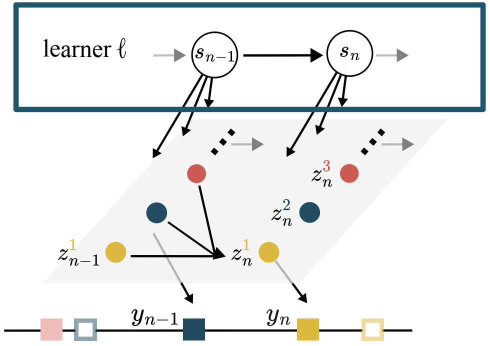


PSI: Modeling interpretable reps. from the ground up

- **Cognitive traits:** forgetting rate, consolidation memory, transfer ability, volatility

Representation capacity

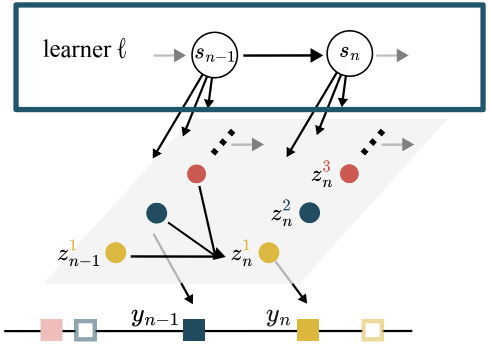
- *Specific* to each learner
- *Consistent* across data splits
- *Disentangled* across dimensions



- **Cognitive traits:** forgetting rate, consolidation memory, transfer ability, volatility

Representation capacity

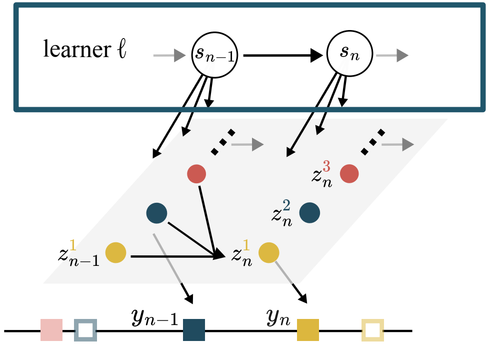
- *Specific* to each learner: $\mathrm{MI}(s;\ell) = \mathrm{H}(s) - \mathrm{H}(s \mid \ell)$
- *Consistent* across data splits
- *Disentangled* across dimensions



- **Cognitive traits:** forgetting rate, consolidation memory, transfer ability, volatility

Representation capacity

- *Specific* to each learner
- *Consistent* across data splits:
  $\mathbb{E}_{\ell_{\text{sub}}} \mathrm{MI}\left(s^{\ell}; \ell_{\text{sub}}\right) := \mathbb{E}_{\ell_{\text{sub}}}\left[\mathrm{H}(s \mid \ell) - \mathrm{H}\left(s \mid \ell_{\text{sub}}\right)\right]$
- *Disentangled* across dimensions

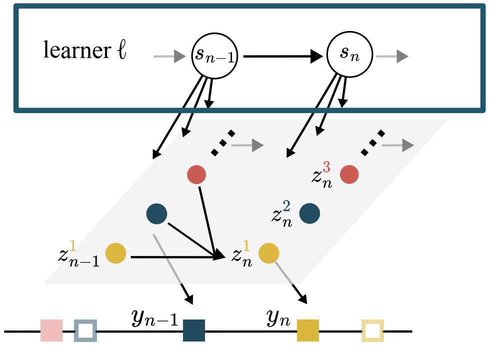


- **Cognitive traits:** forgetting rate, consolidation memory, transfer ability, volatility

Representation capacity

- *Specific*
- *Consistent* across data splits
- *Disentangled* across dimensions:

$\mathrm{KL}(s\|\ell) := \mathrm{H}(s)_{\text{full}} - \mathrm{H}(s \mid \ell)_{\text{diag}}$



PSI: Modeling interpretable reps. from the ground up

- **Cognitive traits:** forgetting rate, consolidation memory, transfer ability, volatility

Representation capacity

- *Specific* to each learner
- *Consistent* across data splits
- *Disentangled* across dimensions

Metric	Dataset	Baseline	PSI-KT
Specificity $\mathrm{MI}(s;\ell)\uparrow$	Assist12	**8.8**	$\underline{8.4}$
	Assist17	**10.1**	$\underline{10.0}$
	Junyi15	$\underline{13.5}$	**14.4**
Consistency^{-1} $\mathbb{E}_{\ell_{\text{sub}}}\mathrm{MI}(s^{\ell};\ell_{\text{sub}})\downarrow$	Assist12	$\underline{12.2}$	**7.4**
	Assist17	**6.4**	**6.4**
	Junyi15	$\underline{7.7}$	**5.0**
Disentanglement $D_{\text{KL}}(s\|\ell)\uparrow$	Assist12	$\underline{2.3}$	**7.4**
	Assist17	$\underline{0.6}$	**8.4**
	Junyi15	$\underline{5.0}$	**11.5**

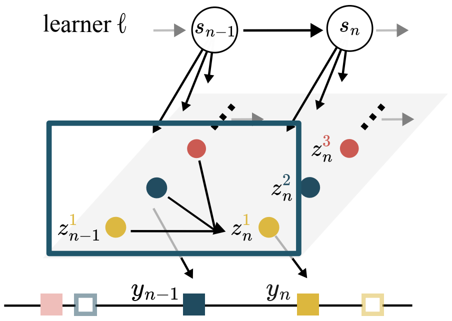
Bold indicates the better model.
PSI-KT vs. the best baseline model.

PSI: Modeling interpretable reps. from the ground up

- **Knowledge structure:** prerequisites

Structure correctness:

- Human-annotated ground-truth
- Learners' progress



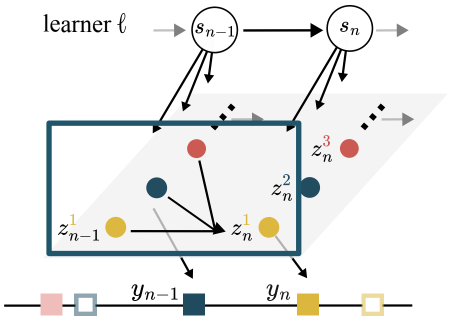
PSI: Modeling interpretable reps. from the ground up

- **Knowledge structure:** prerequisites

Structure correctness:

- Human-annotated ground-truth
- Learners' progress

Metric	MRR ↑	JS expert ↑	JS crowd ↑	nLL ↓
Dataset		Junyi15		
Best Baseline	.0082	.0015	.0047	**3.03**
PSI-KT	**.0086**	**.0019**	**.0095**	4.11

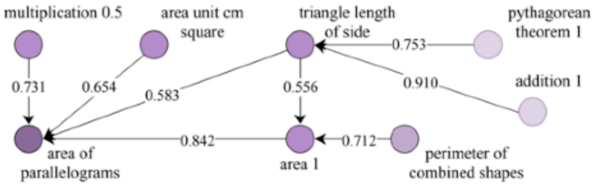


EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- **Knowledge structure:** prerequisites

Structure correctness: causally support in human learning

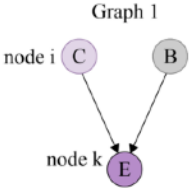
Inferred graph



Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological review*, *116*(4), 661.

Behavioral data

Background info (e.g., learner's ability, other concepts) causes correctness on k



Correctness on i causes correctness on k

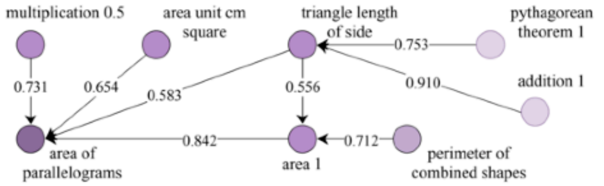
$$\text{support}_{i\to k} := \log P\left(\mathscr{H} \mid G_{i\to k}\right) - \log P\left(\mathscr{H} \mid G_{i\to k}\right)$$

PSI: Modeling interpretable reps. from the ground up

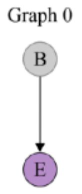
- **Knowledge structure:** prerequisites

Structure correctness: causally support in human learning

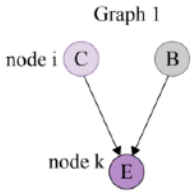
Inferred graph



Behavioral data



Background info (e.g., learner's ability, other concepts) causes correctness on k



Correctness on i causes correctness on k

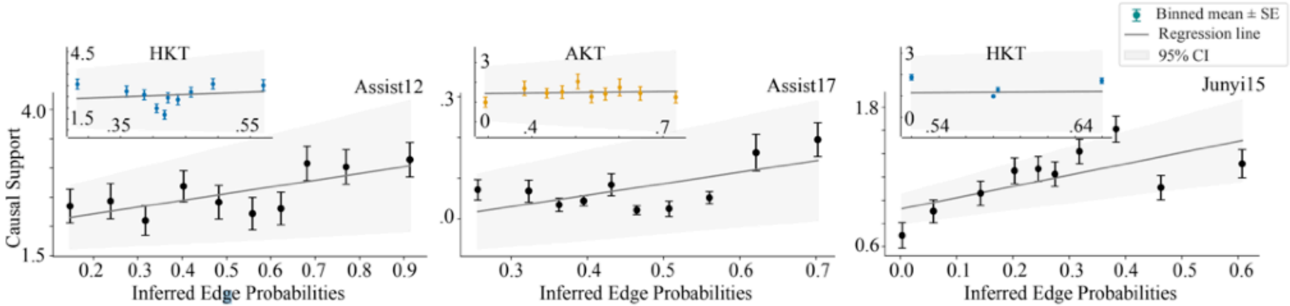
$$\text{support}_{i \to k} := \log P\left(\mathscr{H} \mid G_{i \to k}\right) - \log P\left(\mathscr{H} \mid G_{i \not\to k}\right)$$

PSI: Modeling interpretable reps. from the ground up

- **Knowledge structure:** prerequisites

Structure correctness: causally support in human learning

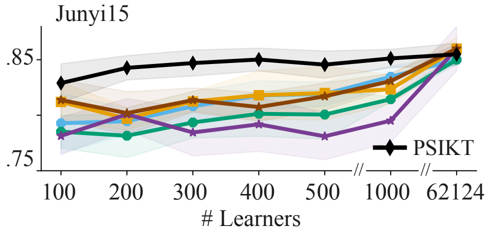
$$\text{support}_{i \to k} := \log P\left(\mathscr{H} \mid G_{i \to k}\right)$$
$$-\log P\left(\mathscr{H} \mid G_{i \to k}\right)$$

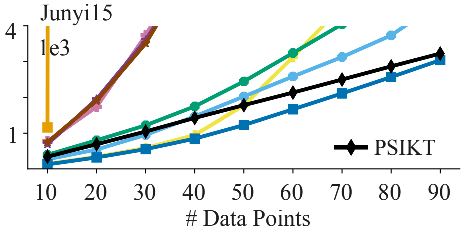


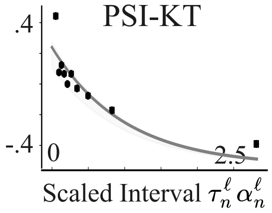
Summary

PSI-KT

Predictive, Scalable, Interpretable







Knowledge tracing for future intelligent tutoring systems

Scan Me

Acknowledgement

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Robert Bamler

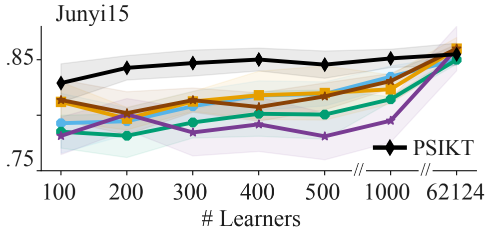
Charley Wu

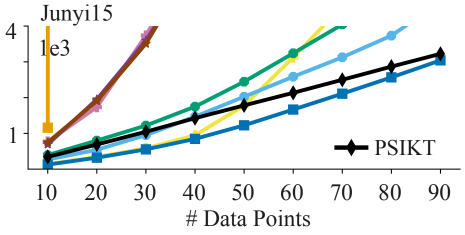
Álvaro Tejero-Cantero

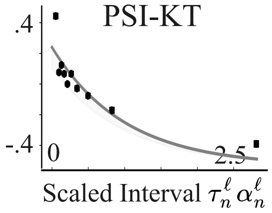
Summary

PSI-KT

Predictive, Scalable, Interpretable







Knowledge tracing for future intelligent tutoring systems

Scan Me