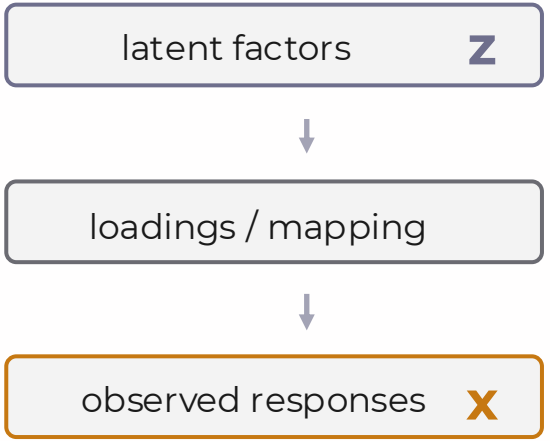
Variational Autoencoders

Konstantina Sokratous

Psychology thinks in latent variables

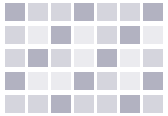
- Classical tools focus on a handful of hidden dimensions that generate many observed responses
- Measurement itself is a modeling decision, but core idea is that useful models preserve the structure we care about



Where standard tools can become strained

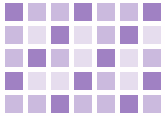
Nonlinear

Constructs may bend through task space



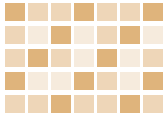
Multimodal

Choices, RTs, EMA, physiological and neural data, text



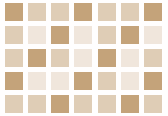
Sparse

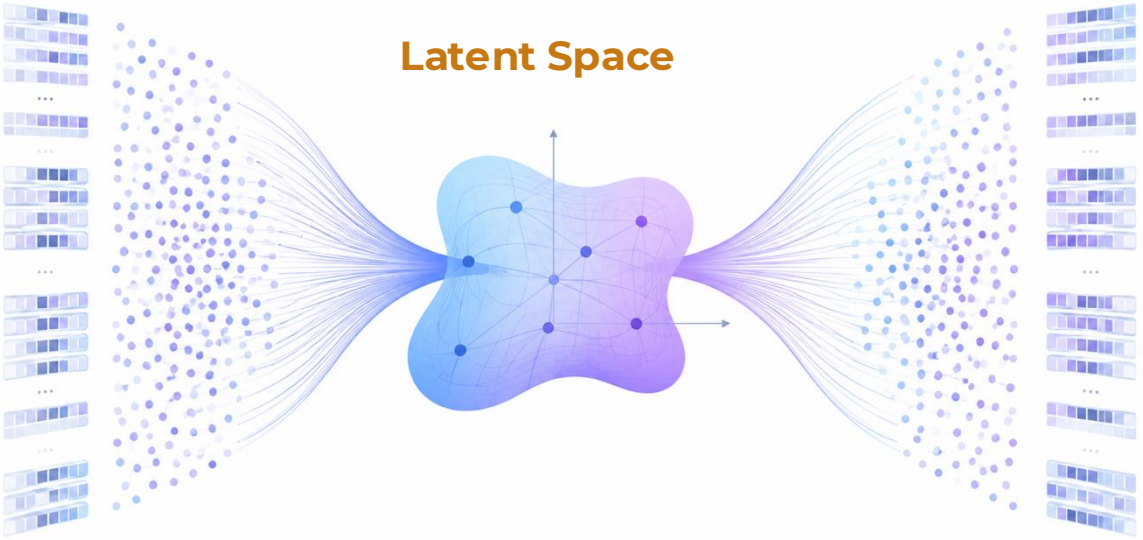
Missing entries



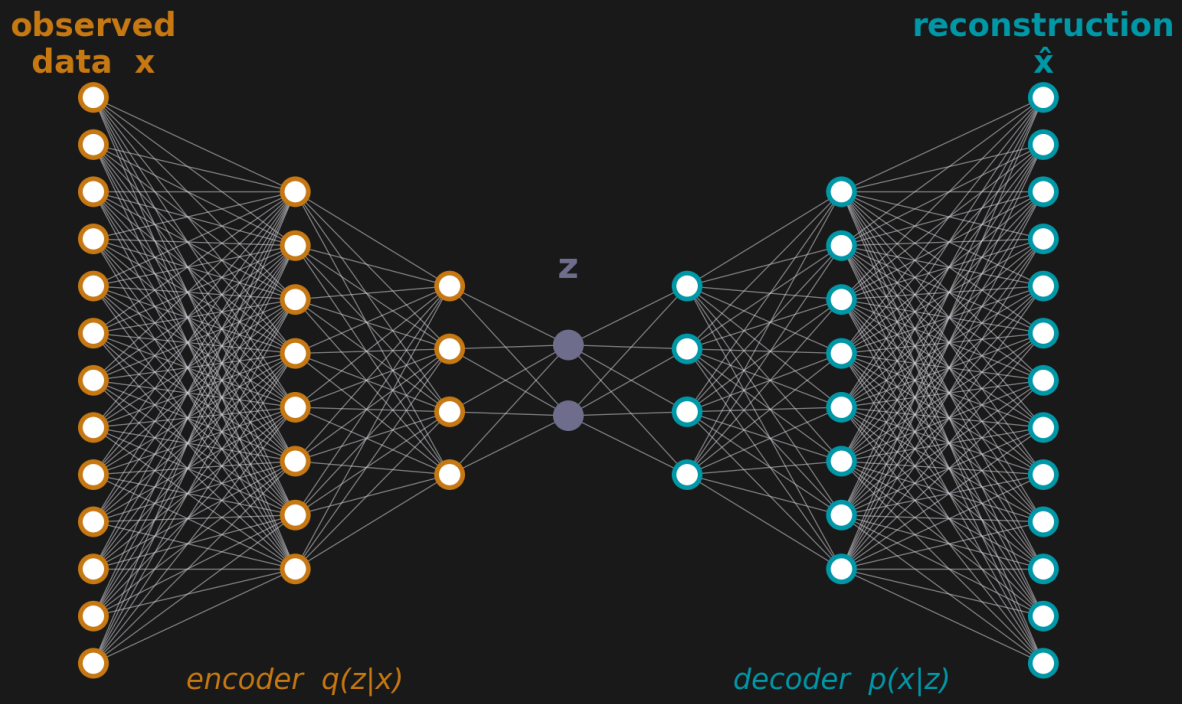
High-dimensional

Many features, fewer people, nested repeated measures

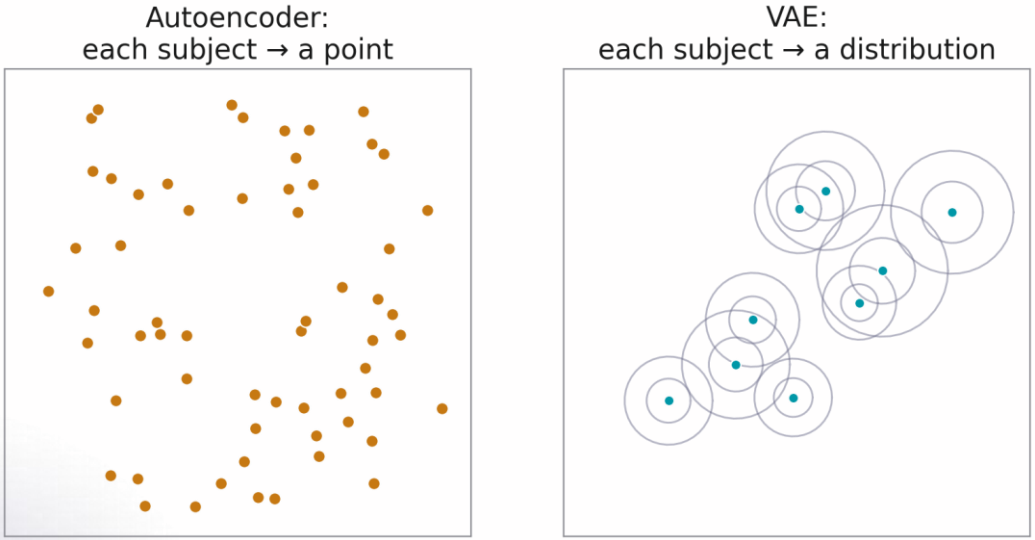




A VAE is a nonlinear latent-variable model that learns to reconstruct and simulate the data it compresses

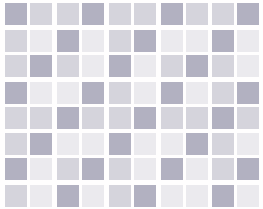

observed
data x
reconstruction
x̂
z
encoder q(z|x)
decoder p(x|z)

VAEs add uncertainty & a generative model

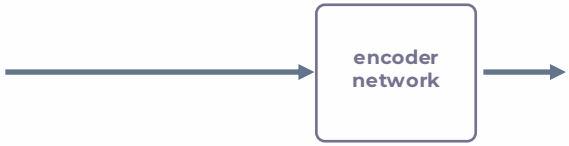


- A plain autoencoder gives each subject a single point
- A VAE gives each subject a small Gaussian $q(z|x) = N(\mu, \sigma^2)$
- That buys uncertainty per subject, a smooth space you can sample from, and built-in regularization

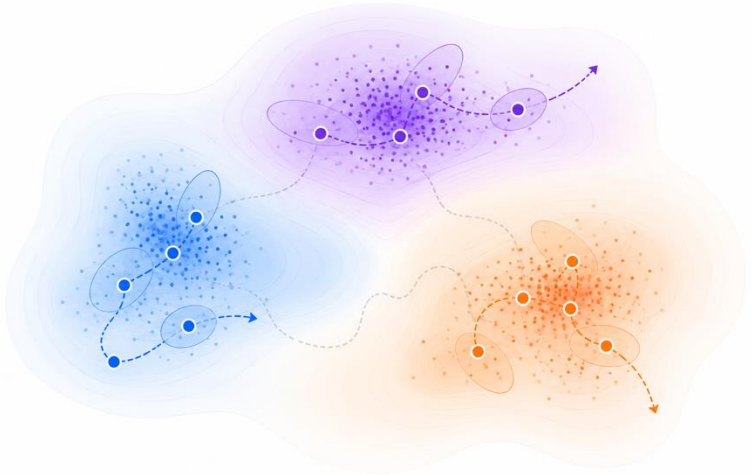
The encoder: Where in latent space could this observation live?



x

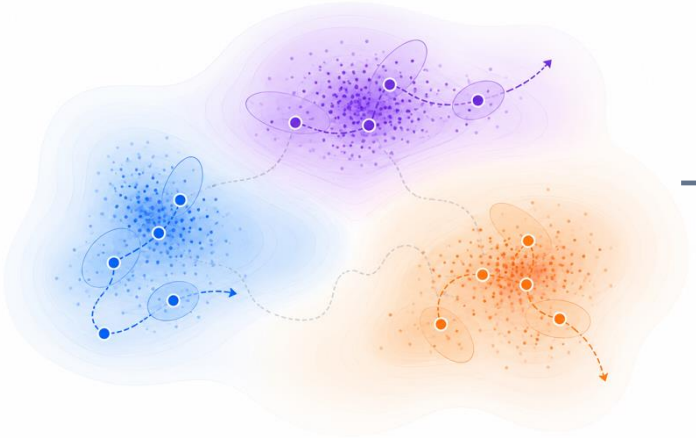


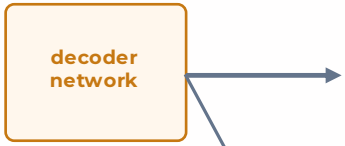
The encoder learns an approximate posterior
q(z | x)

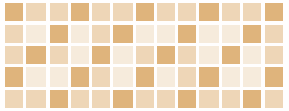


- **Input** can be a response vector, task sequence, time window, image, or neural feature vector
- **Output** is a distribution over latent possibilities

The decoder: What data would we expect from this latent position?







The decoder learns p(x | z)

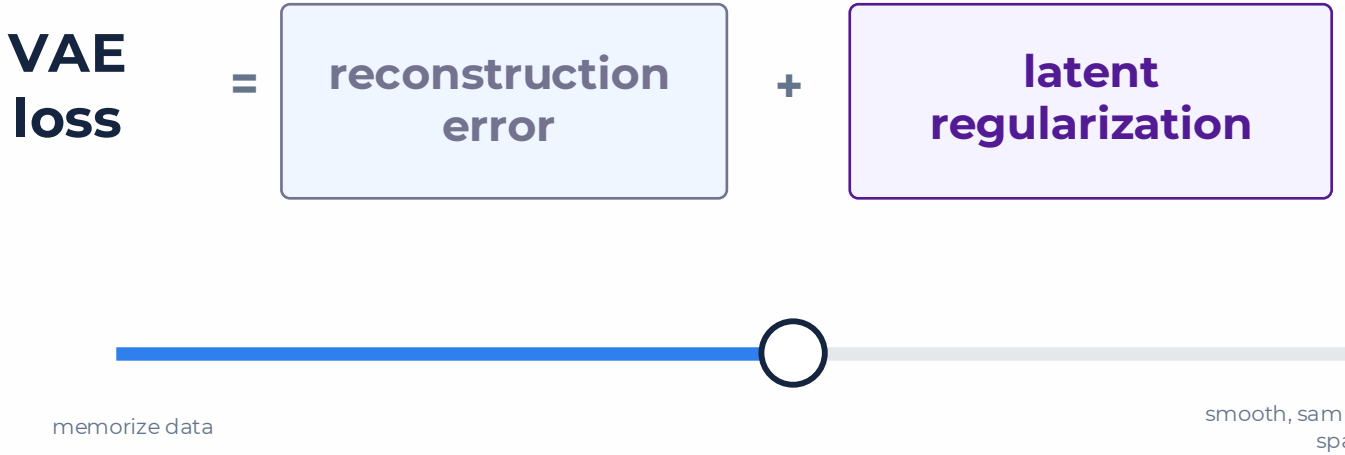
Data distribution conditional on latent variables

- **Reconstruction:** can the latent variables recreate the observed pattern?
- **Generation:** what new plausible observations does the model imply?
- **Counterfactual:** how would the data change if we move in latent space?

The training objective

ELBO & Reparameterization

The training objective tradeoff



- **Too little regularization**: excellent reconstruction, weak generative structure
- **Too much regularization**: neat latent space, poor reconstruction / posterior collapse

The training objective tradeoff

Maximize ELBO = **reconstruction** (Decode and compare to x) + **KL Penalty** (Keep q(z|x) near prior)

Fit the data

Regularize

memorize data

smooth, sampleable latent space

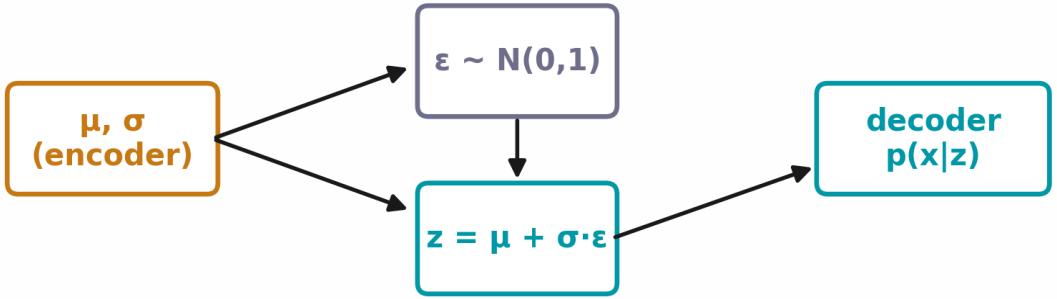
$$\log p(x) = \log \int p(x|z)\, p(z)\, dz$$

intractable integral over all latent codes

optimize the ELBO

The reparameterization trick

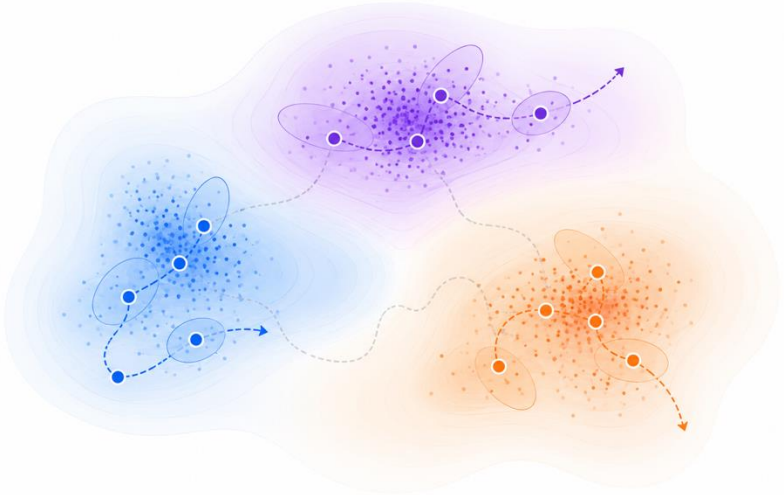
randomness pushed into ε → gradients flow through μ, σ



Problem: z is sampled from q(z|x), and you can't backpropagate through a random draw

Fix: write $z = \mu + \sigma \cdot \varepsilon$ with $\varepsilon \sim N(0,1)$. The randomness lives in ε, gradients flow cleanly through μ and σ.

Latent space is a learned map of similarity

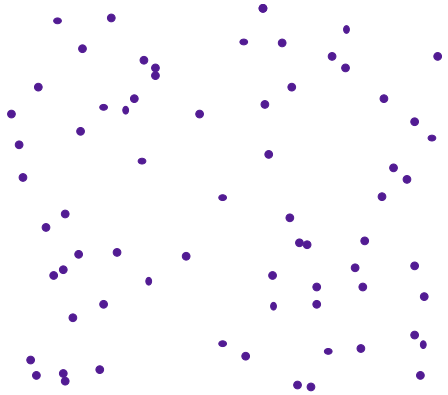


- **A subject-level map**: similar response profiles cluster together
- **A trial-level map:** strategies or states as trajectories
- **A stimulus/item-level map**: item properties in terms of geometry
- **A model-behavior map:** AI systems can be compared as behavioral agents

> Psychological interpretation starts after the map is learned, with validation.

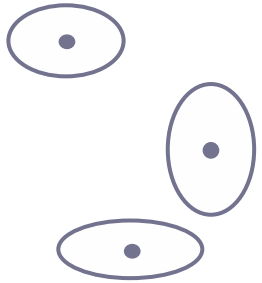
Why the latent prior matters

Prior p(z)



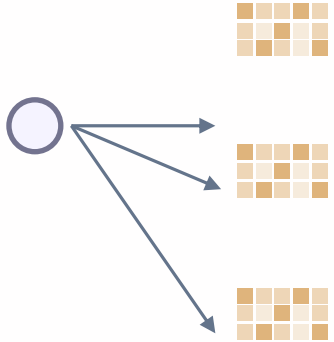
You can start with a standard normal. It encourages organized, sampleable space

Posterior q(z|x)



Each observation gets uncertainty around its latent location

Sample → decode



New plausible observations come from points in latent space

Three “knobs” to keep in mind

β (KL Weight)

β-VAE to weigh the KL divergence term

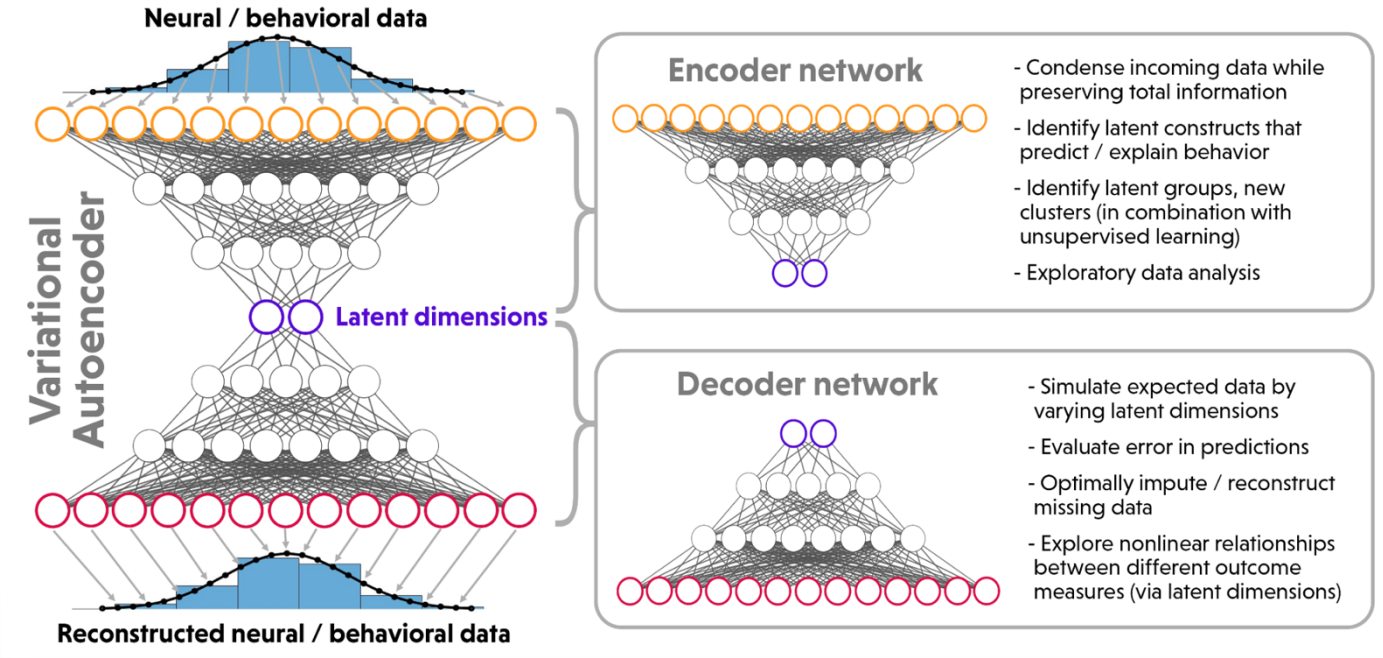
If the decoder too strong, z gets ignored,
the KL term vanishes

Posterior Collapse

Decode Likelihood

Bernoulli for binary items, Gaussian for continuous
Use the appropriate link function

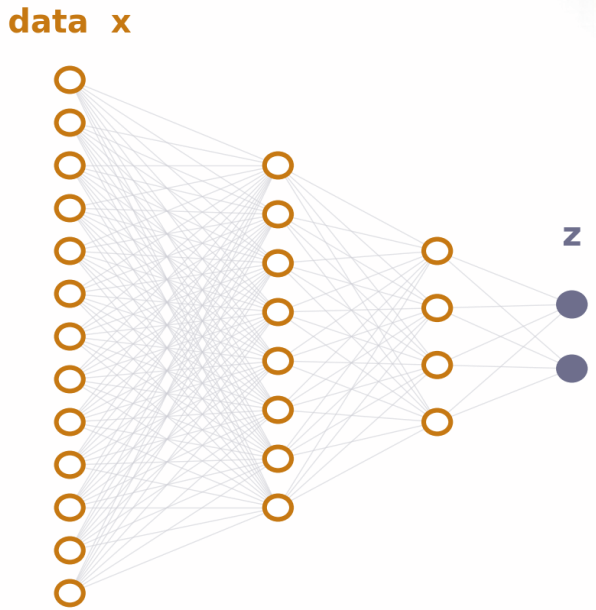
Affordance visual



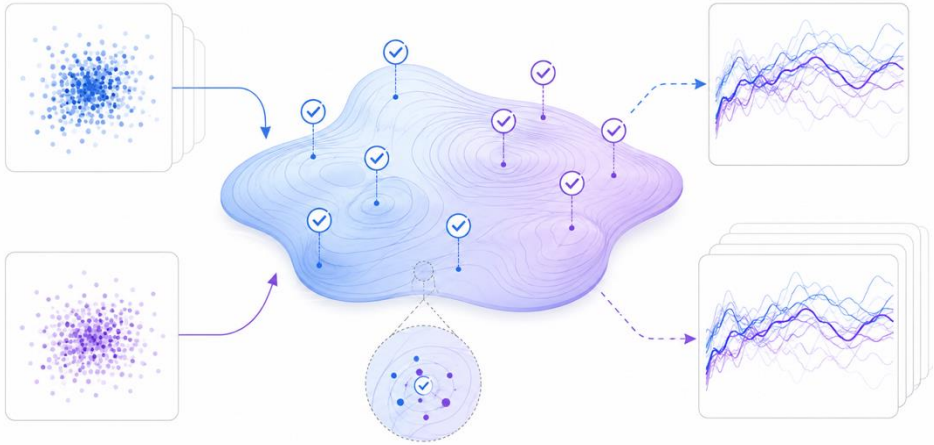
Affordance 1 – nonlinear dimensionality

> How many latent dimensions are needed to explain the response pattern?

- Learn **nonlinear mappings** from latent variables to observed responses
- **Best use-case:** when you suspect curved manifolds, interactions, or multimodal outcomes



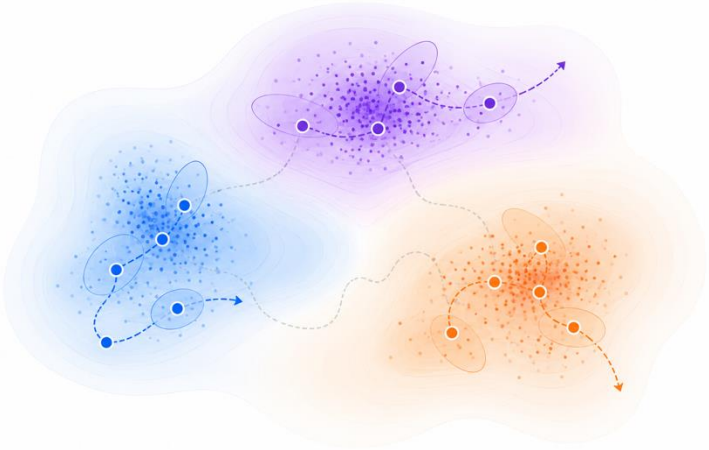
Recover known structure



1. **Simulate data with known K**
2. **Fit VAE sweep across K**
3. **Select K by held-out metrics**
4. **Check stability & external validity**

Claim recoverability, robustness, and usefulness, not identifiability

Affordance 2 – Representation learning



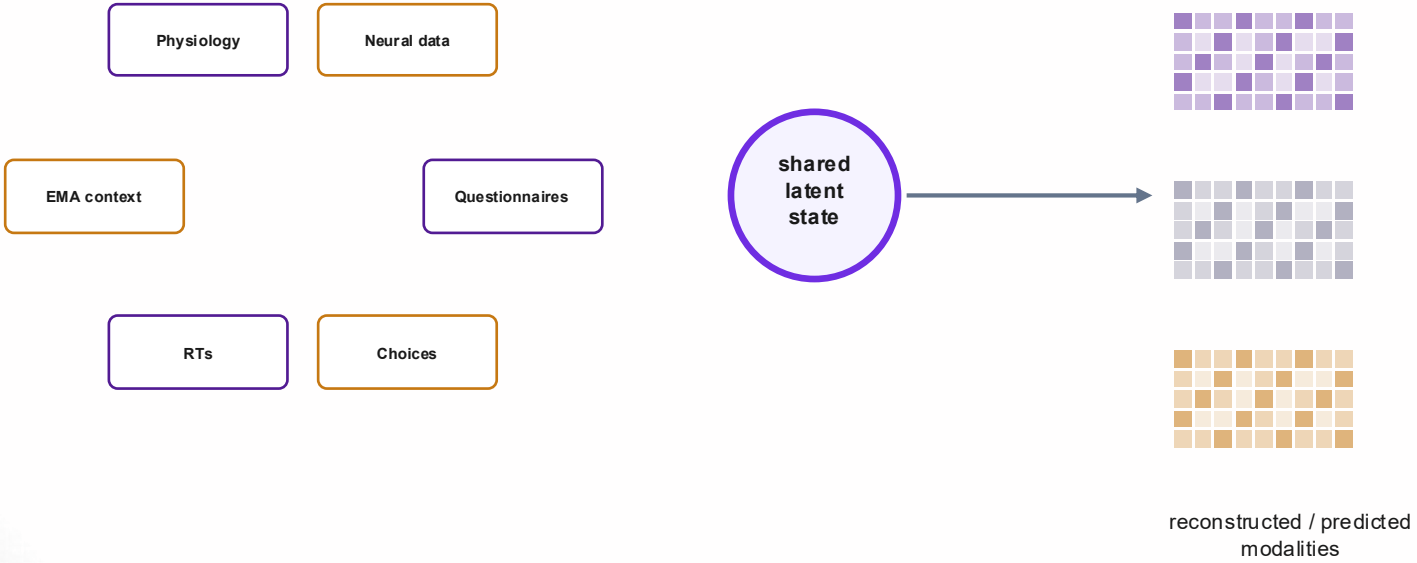
- Predict outcomes
- Cluster profiles
- Track dynamics

Individual differences in the latent space

Embed every subject, then read the space two ways:

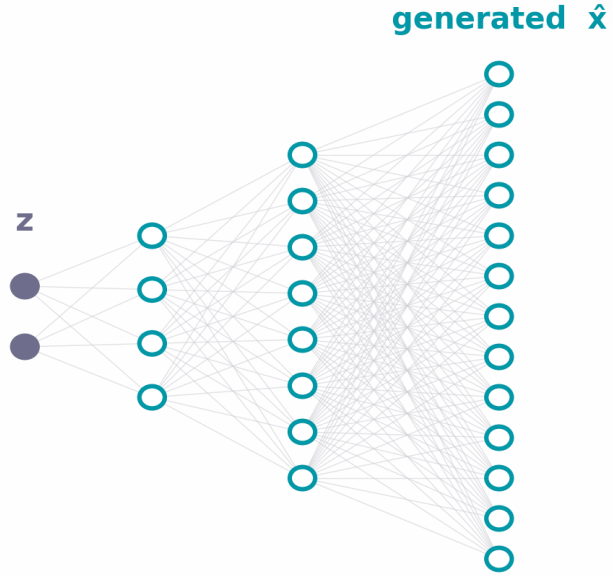
- colour it by an outcome to see what a latent dimension tracks
- cluster the codes to surface latent subtypes of people
- then correlate codes with external measures

Affordance 3 – Joint & Multimodal Modeling

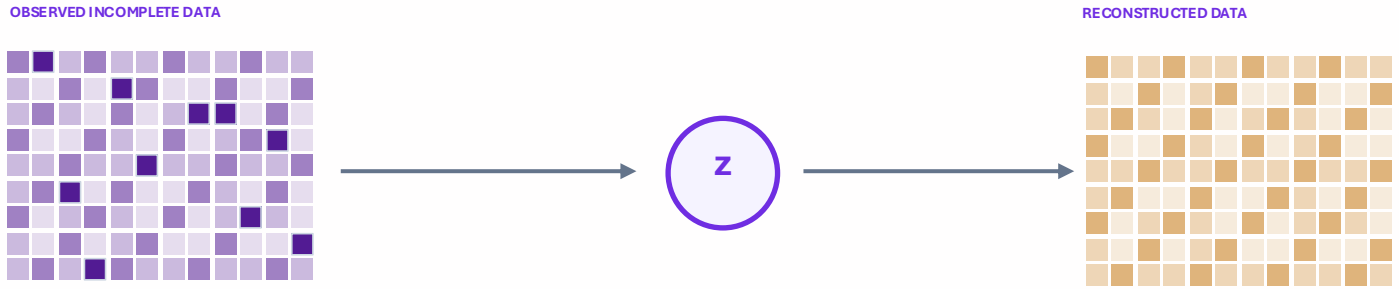


This is especially useful when the scientific question is about the joint distribution, not one outcome at a time

On the decoder side

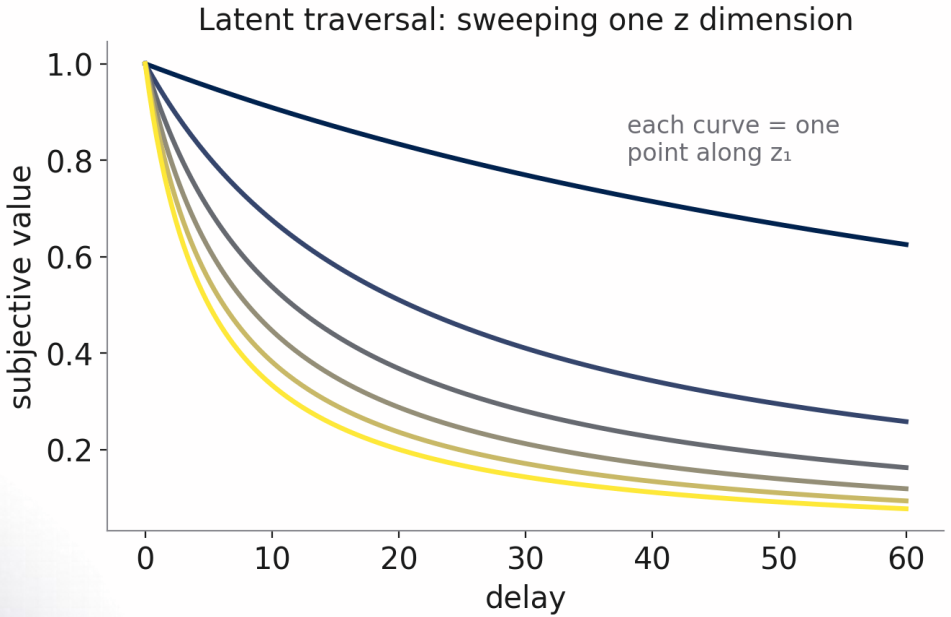


Affordance 4: Reconstruction & imputation



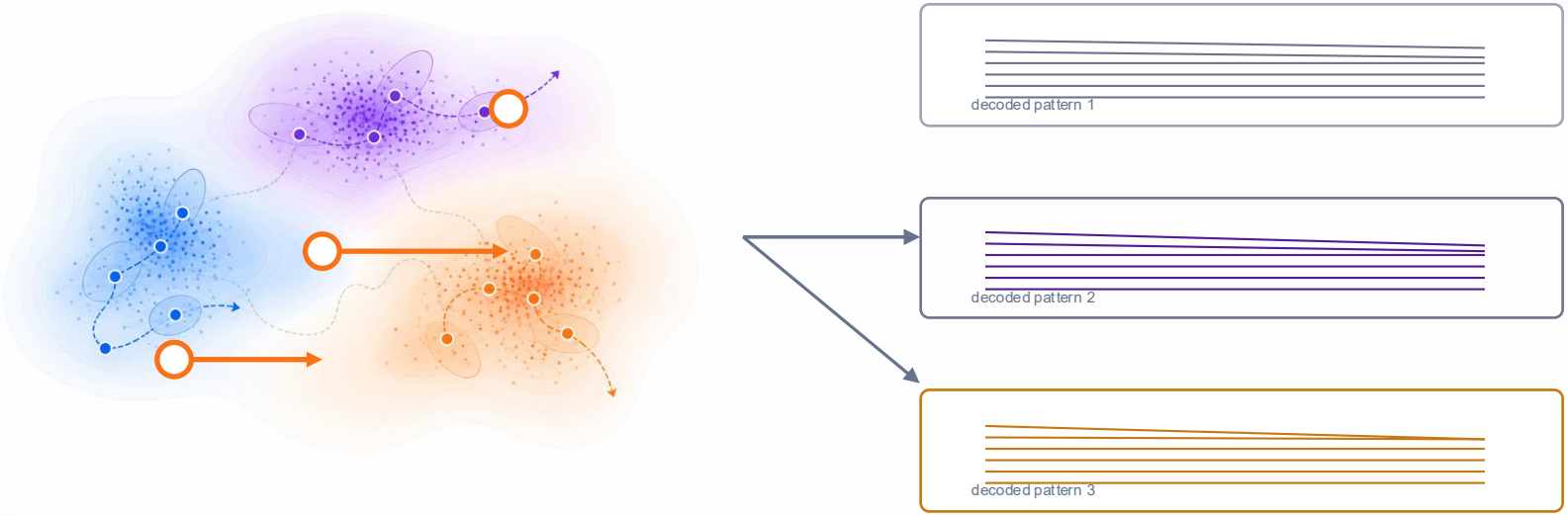
- Impute missing data
- Denoise noisy behavioral or neural features
- Check model misfit by inspecting reconstruction errors

Affordance 5 – Latent Traversals

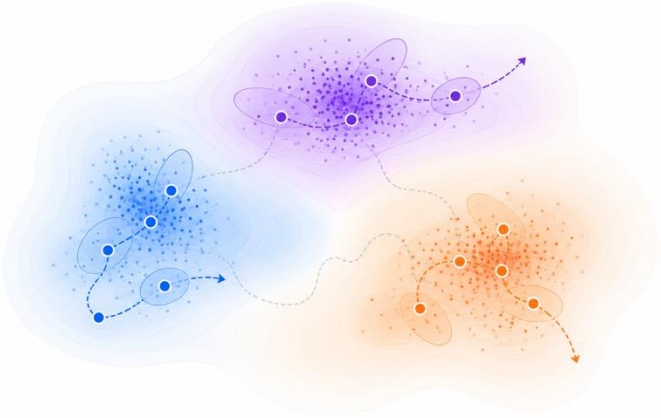


- Hold every dimension fixed but one, sweep it, and decode.
- The generated behavior shows what that dimension means, here
- This is how you attach psychological meaning to an otherwise abstract z.

Counterfactual Exploration



Interpreting latent dimensions requires extra work



- decode traversals
- correlate with outcomes
- map item/task loadings
- test invariance
- check seed stability

> Name the dimension only after showing what changes when you move along it.

Validation checklist for psychology applications

1. **Recoverability** — Can the approach recover known K and known nonlinear structure in simulation?
2. **Stability** — Do conclusions hold across seeds, splits, architectures, priors, and hyperparameters?
3. **Predictive validity** — Do latent representations predict external outcomes or future behavior?
4. **Construct validity** — Do decoded patterns and correlates make psychological sense?
5. **Comparative value** — What does the VAE add beyond PCA, EFA, IRT, SEM, or task models?

Take-home messages

1. **VAEs are latent-variable models**

   They compress data into probabilistic latent dimensions

2. **The decoder is the extra power**

   It enables reconstruction, imputation, simulation, and counterfactual exploration

3. **Psychology needs validation**

   Use simulations, baselines, seed stability, and external outcomes

4. **Best perk: nonlinearity**

   VAEs extend classic latent-variable modeling to messy, high-dimensional behavioral data

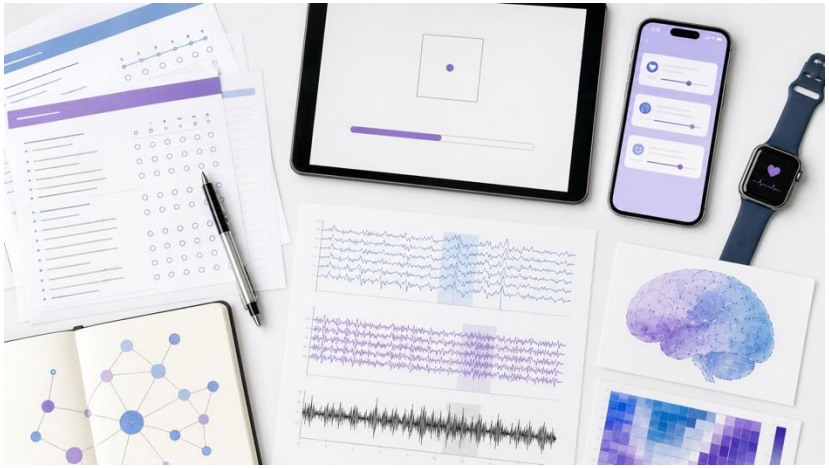
Thanks!

Do you have any questions?

ksokratous@missouri.edu

lazyneuron.com

Application: EMA, clinical states, and addiction science

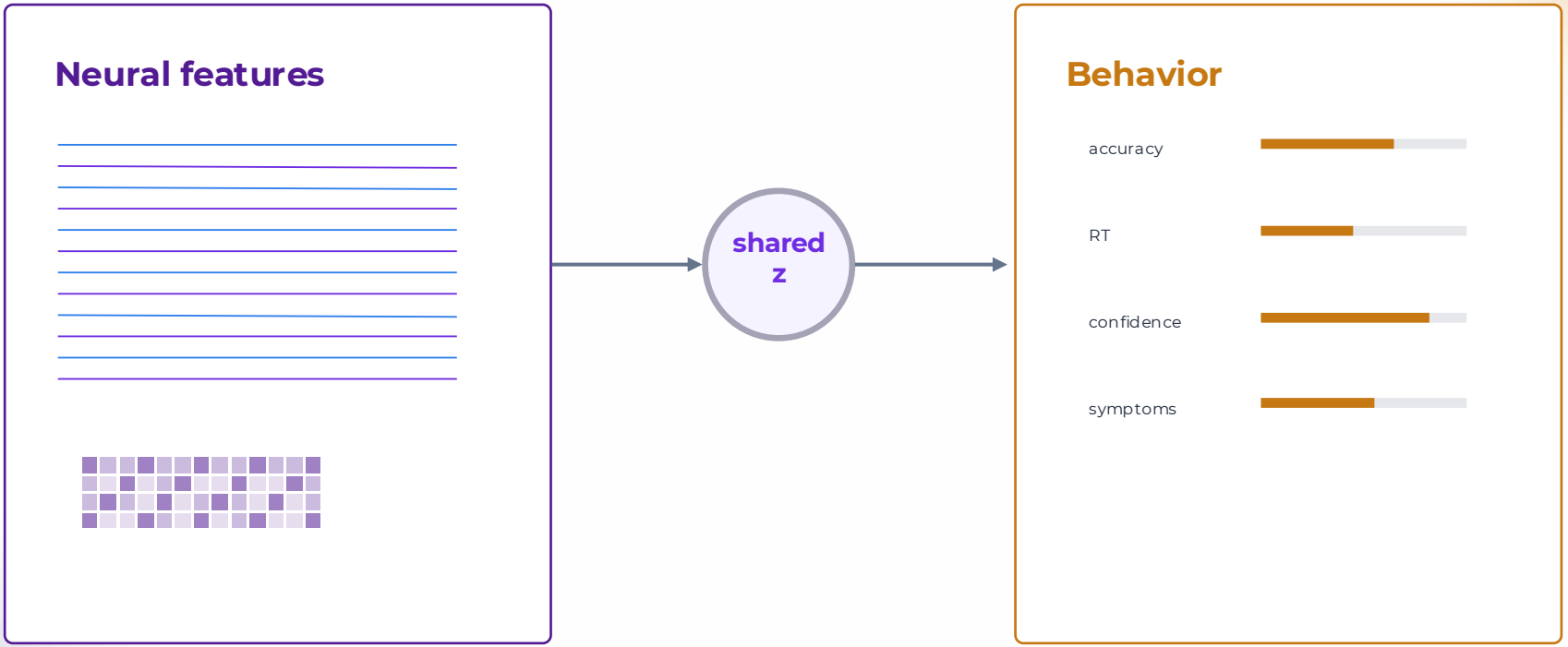


Example latent state

- mood
- craving
- context
- availability
- drinking
- transport risk

> **Model the joint pattern then predict clinically meaningful outcomes**

Application: neural + behavioral data



Application: AI systems as behavioral subjects