Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data

Luigi Antelmi [1] Nicholas Ayache [1] Philippe Robert [2 3] Marco Lorenzi [1]
for the Alzheimer's Disease Neuroimaging Initiative*

MLBBQ - 19 September 2025

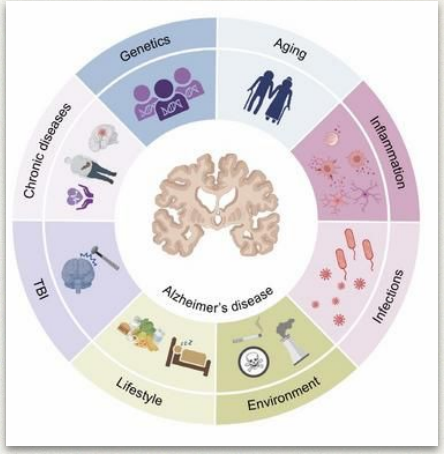
Heterogeneous Data

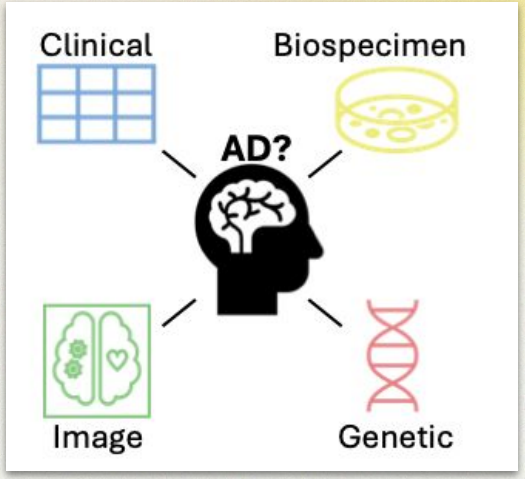
- Data that comes from different sources
- Data which exhibits non-homogeneous trends/behavior
- Data which represents complementary aspects of an underlying system

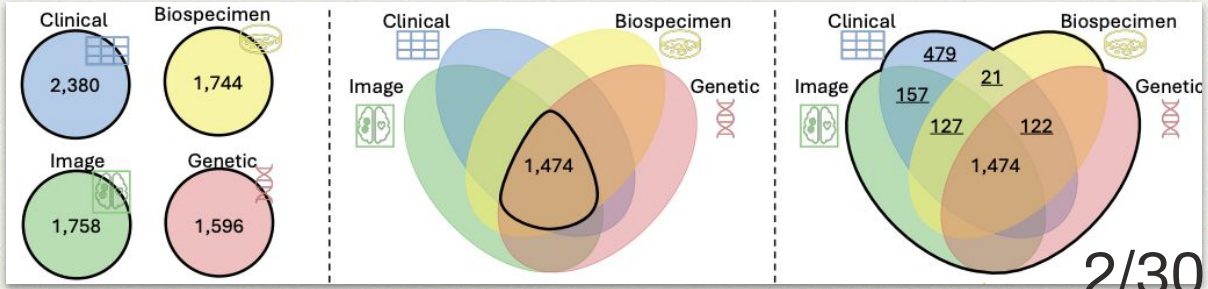
- Data types (modalities)
- Levels of measurement
- Scales/Distributions

Heterogeneous Data & AD

- Complex Pathology
  - Age, genetics, lifestyle
  - Biological processes, …
- Various representations across modalities
  - Atrophy, connectivity, plaques, CSF biomarkers, …
- Information Overlap







Neuroimaging - Data Tsunami

- > 1M voxels
- 100+ subjects

- Curse of dimensionality
  - Sparse feature space
  - Overfitting

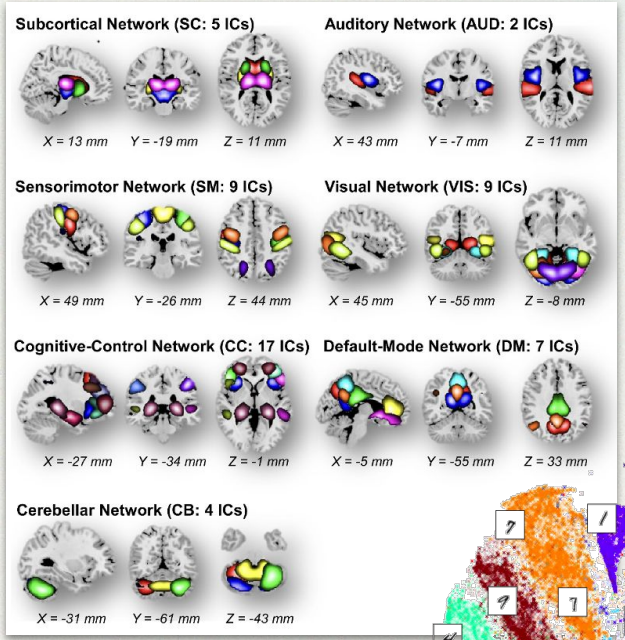
- Massive redundancy

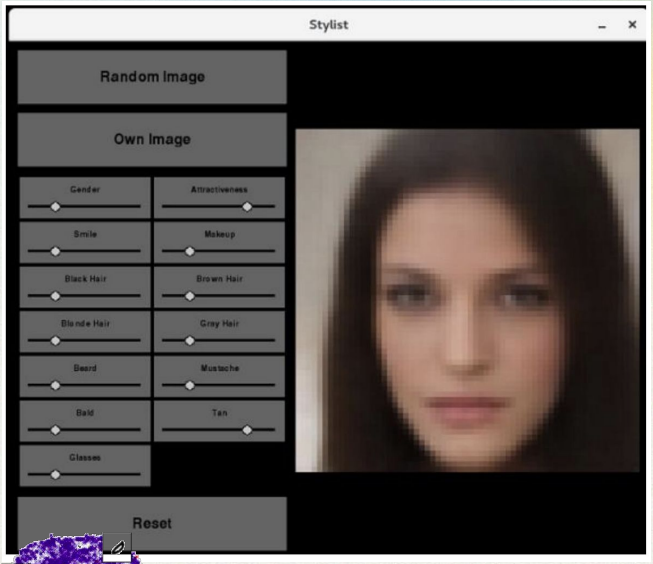
- Computational impossibility

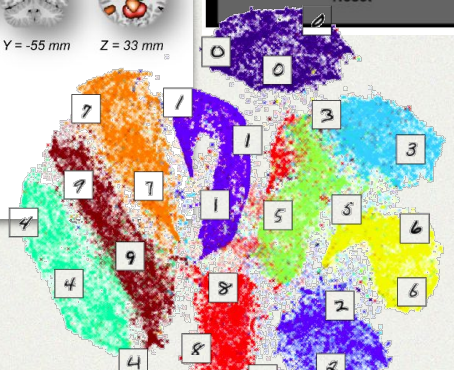
- Needle in a hackstack

Dimensionality Reduction

- Find meaningful patterns
- Enable statistical/machine learning
- Visualization + Interpretation
- De-noising







Recognition Methods

- Projects observations to low dimensional space
  - Desired qualities enforced
- Maximum Correlation
  - CCA Canonical Correlation Analysis
- Maximum covariance
  - PLS Partial Least Squares
- Minimum Regression Error
  - RRR Reduced Rank Regression
- Can’t sample from latent space
- How to handle missing data

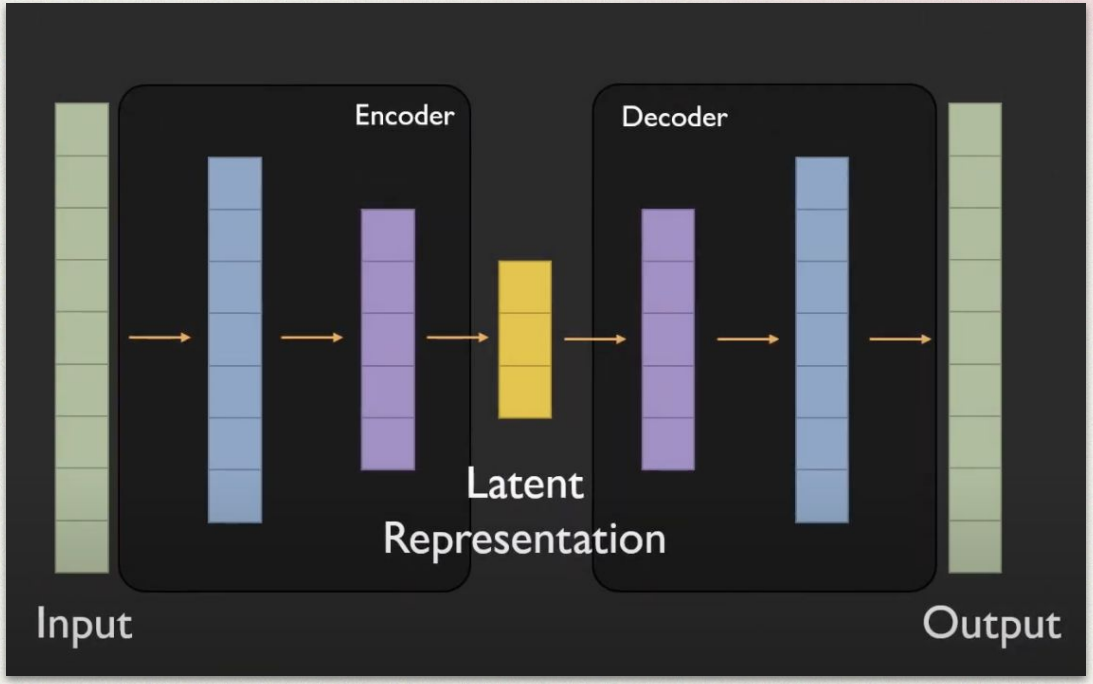
Generative Methods

- Learns a parameters of a latent ***distribution*** from which data can be sampled

- Baysian CCA
  - Doesn’t scale well
  - Lower dimension
  - Single channel
- VAEs
  - Interetability of latent space

Autoencoders

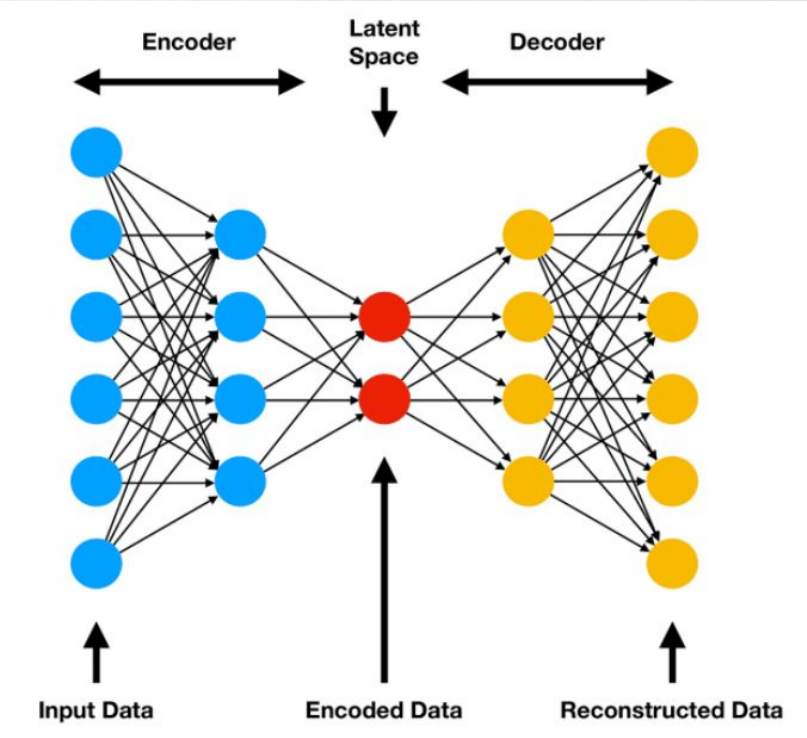
- NN that learns to **reconstruct the input**
  - Reconstruction Loss
- No labels required
  - **Self-supervised**
- Non-linear dimensionality reduction
  - hidden dimension **bottleneck**



https://www.youtube.com/watch?v=Dp6iICL2dVI

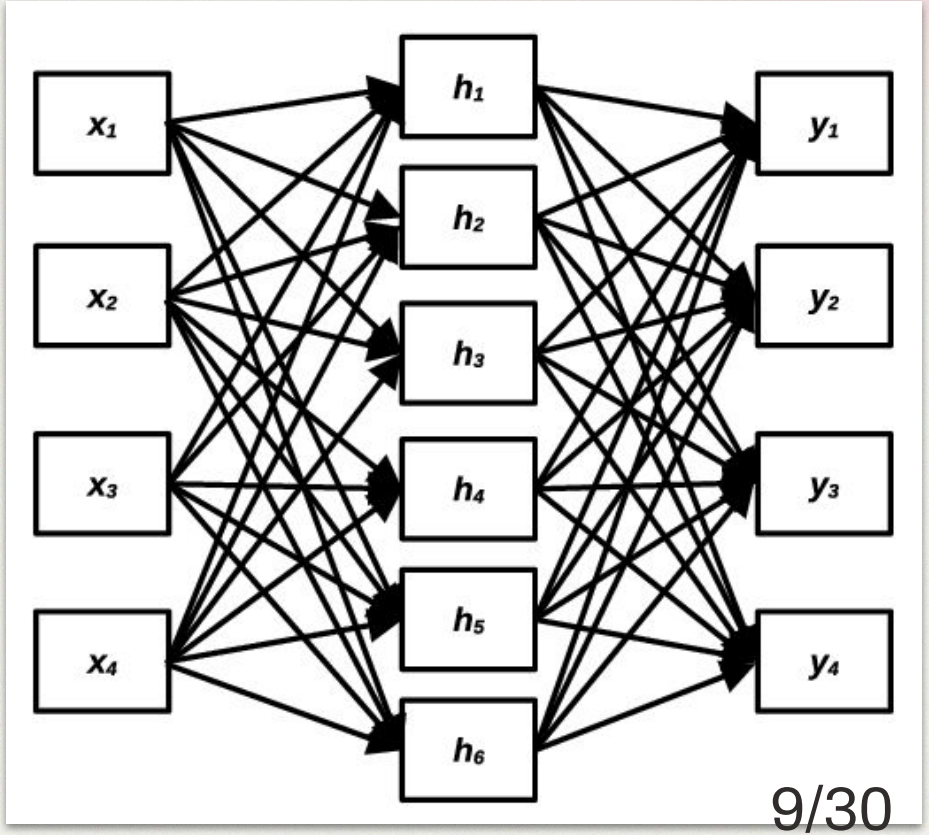
Undercomplete Autoencoder

- NN with bottle neck at hidden dimension
- Forces dimensionality reduction
- Facilitates a latent representation of input
  - used as ‘features’



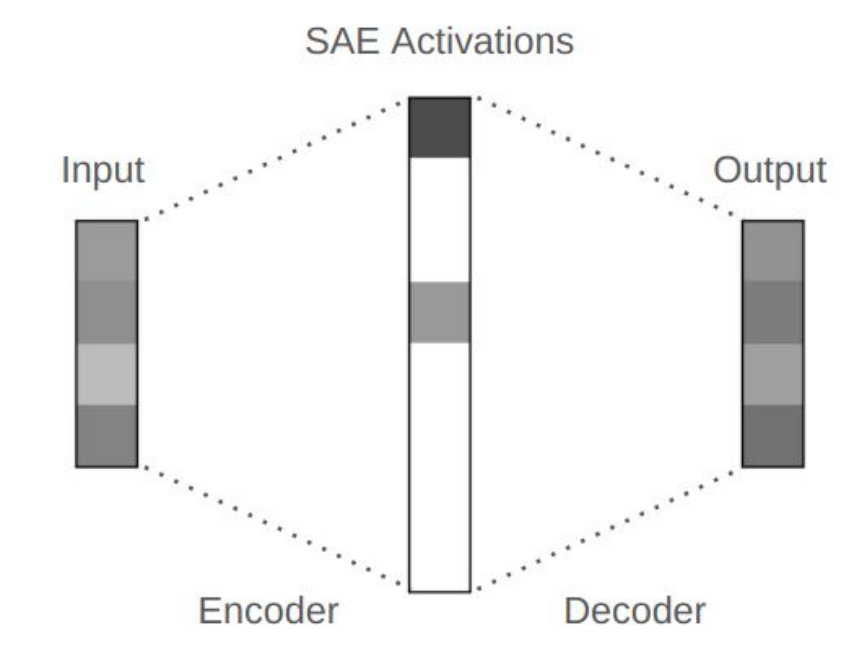
Overcomplete Autoencoder

- NN where hidden dimension is ***greater*** than input dimension
- Model can learn detailed representation
  - Risks copying input to output
- Can be used in sparse AEs

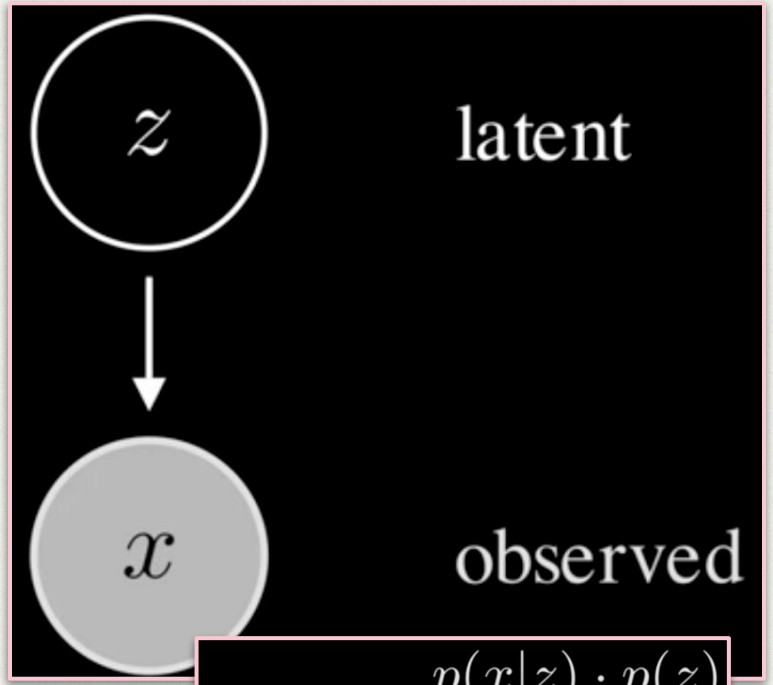


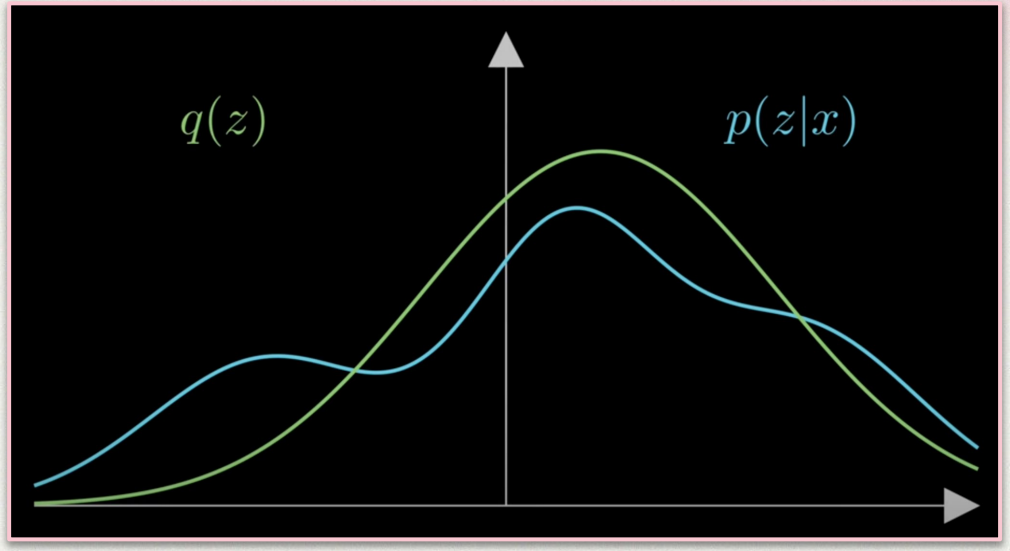
Sparse Autoencoders

- Force the hidden dimension to have as many zeros as possible
- Still can learn meaningful features even if overcomplete
- Can learn meaningful parts-based representations



Variational Inference





$$p(z|x) = \frac{p(x|z) \cdot p(z)}{p(x)}$$

$$p(x) = \int \cdots \int p(x, z_1, \ldots, z_d)\, dz_1 \cdots dz_d$$

Evidence Lower Bound

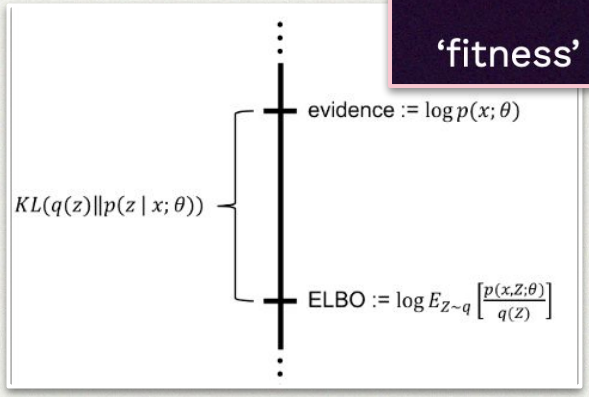
$$q(z) \approx p(z|x)$$

$$KL\left[q(Z) \,||\, p(Z \mid X)\right] = \log p(X) - \mathbb{E}_{Z \sim q(Z)} \log \left[\frac{p(Z, X)}{q(Z)}\right]$$

'fitness' of q(Z)

evidence

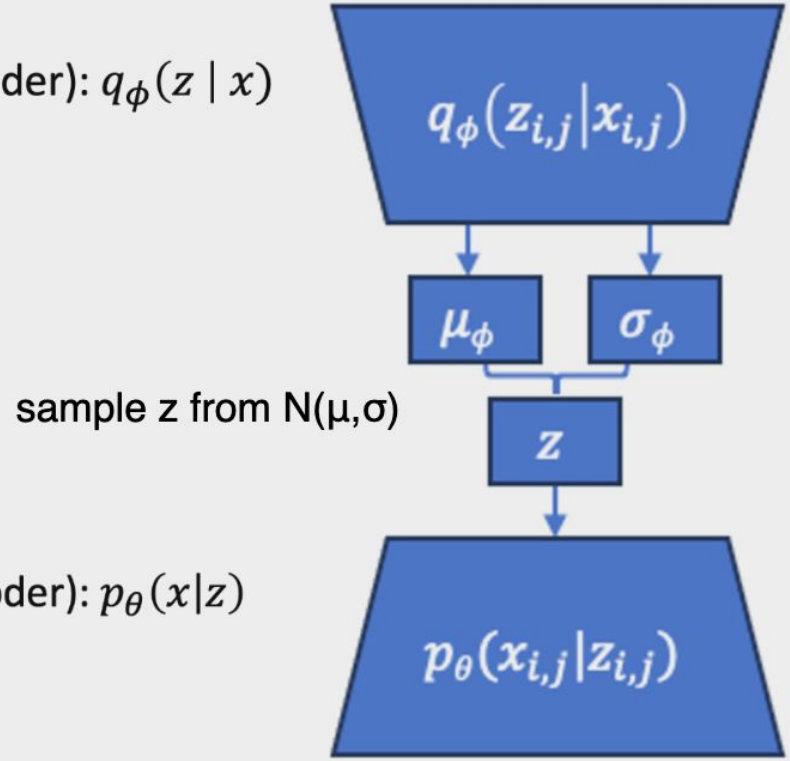
evidence lower bound



Variational Autoencoder

Inference Network (Encoder): $q_\phi(z \mid x)$

- Learn $\mu_\phi(x), \sigma_\phi(x)$



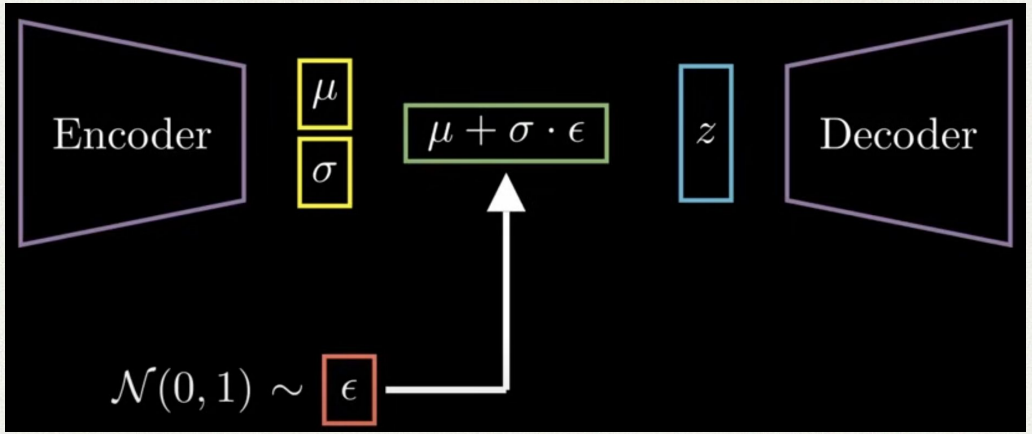
Generative Network (Decoder): $p_\theta(x|z)$

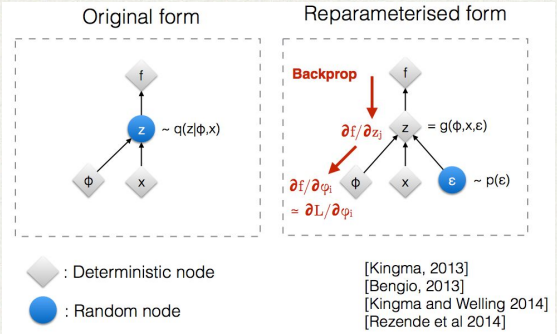
- Learn $\mu_\theta(z), \sigma_\theta(z)$

Reparameterization Trick

- VAE loss function requires random sampling of z
  - Random node

- Take a random point e
  - Scale by sigma
  - Shift by mu





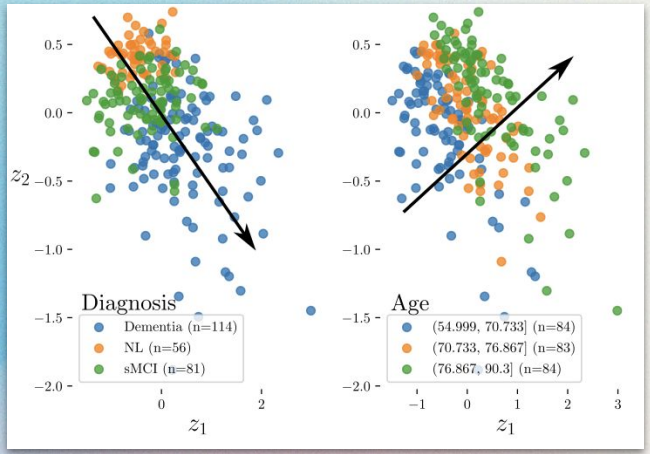
https://www.youtube.com/watch?v=qJeaCHQ1k2w

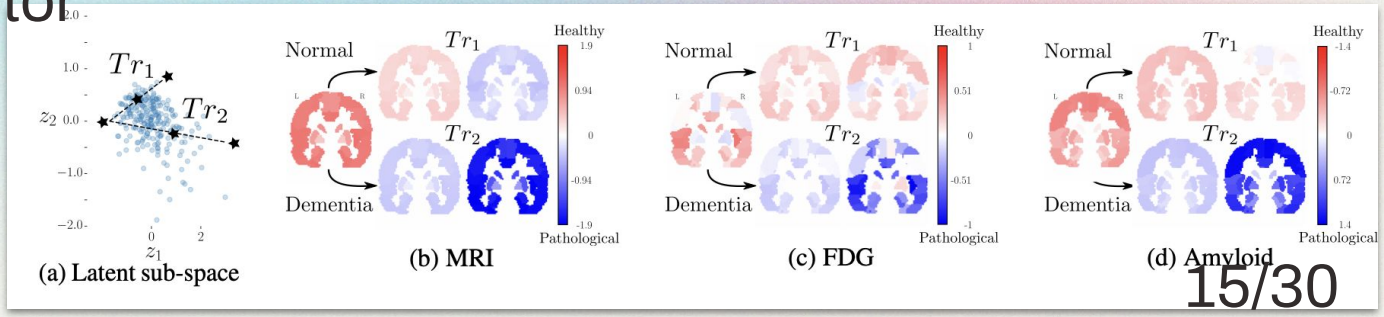
Paper

Sparse Multi-Channel VAE

- Multi-input, multi-output VAE
  - Can reconstruct any data channel from any other channel

- Finds shared, interpretable latent factors that explain all the data
  - *Healthy aging* factor
  - *Disease* factor





Sparse Multi-Channel VAE

Input

$$\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_C\} \quad \mathbf{x} \in \mathbb{R}^d$$

encoder

$$q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right)$$

$$\mathbf{z} \sim p\left(\mathbf{z}\right), \quad \textit{Latent variable distribution}$$

$$\mathbf{x}_c \sim p\left(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c\right), \qquad \text{for } c \text{ in } 1 \ldots C,$$

Channels (modalities)

decoder

Likelihood of observing $\boldsymbol{x}$ given z

$$\mathbf{z} \in \mathbb{R}^l$$

$$\mathbf{x}_c \in \mathbb{R}^d \quad l < d$$

$\boldsymbol{\theta}$ - generative parameters (decoder)

$\boldsymbol{\varphi}$ - variational parameters (encoder)

Each modality, c, has its own likelihood distribution from family $\boldsymbol{Q}$, *parameterized by* $\boldsymbol{\theta}$.

The latent space has its own distribution, from family $\boldsymbol{Q}$, *parameterized by* $\boldsymbol{\varphi}$.

Loss Function (ELBO)

Make appx. posterior close to prior

$$\mathcal{L} = \mathbb{E}_c \left[L_c - \mathcal{D}_{KL}\left(q\left(\mathbf{z}|\mathbf{x}_c\right) || \, p\left(\mathbf{z}\right) \right) \right]$$

$$\text{where} \quad L_c = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)} \left[\sum_{i=1}^{C} \ln p\left(\mathbf{x}_i|\mathbf{z}\right) \right].$$

Reconstruction loss

When c = C -> self reconstruction term (standard VAE)

When c ≠ C -> cross reconstruction terms

Comparison w/ VAE

- Latent space isn’t a concatenated vector
  - They are all separate but learn jointly

- Not the same as ‘stacked’ VAEs
  - Stacked = disjoint latent space

Gaussian-Linear Case

$\mathbf{z} \in \mathbb{R}^l$

$\mathbf{x}_c \in \mathbb{R}^{d_c}$

$$q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c) = \mathcal{N}\left(\mathbf{z}|\mathbf{V}_c^{(\mu)}\mathbf{x}_c, diag(\mathbf{V}_c^{(\sigma)}\mathbf{x}_c)\right), \quad (6)$$

encoder

$$\mu_z = V_c^{\mu} x_c$$
$$\sigma_z^2 = V_c^{\sigma} x_c$$

$$p(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c) = \mathcal{N}\left(\mathbf{x}_c|\mathbf{G}_c^{(\mu)}\mathbf{z}, diag(\mathbf{g}_c^{(\sigma)})\right), \quad (7)$$

decoder

$$\boldsymbol{\theta}_c = \{\mathbf{G}_c^{(\mu)}, \mathbf{g}_c^{(\sigma)}\}$$

$\boldsymbol{\theta}$ - generative parameters (decoder)

$$\boldsymbol{\phi}_c = \{\mathbf{V}_c^{(\mu)}, \mathbf{V}_c^{(\sigma)}\}$$

$\boldsymbol{\varphi}$ - variational parameters (encoder)

$\boldsymbol{G}$ - generative matrix
- Decoder weights
- Linear map that transforms z to x

$$G \in \mathbb{R}^{d_c \times l}$$

V_c^{μ} - inference matrix
- Encoder weights
- Linear map that transforms x to z

$$V_c^{\mu} \in \mathbb{R}^{l \times d_c}$$

$\boldsymbol{g}$ - noise parameter
- Variances of each channel not explained by z

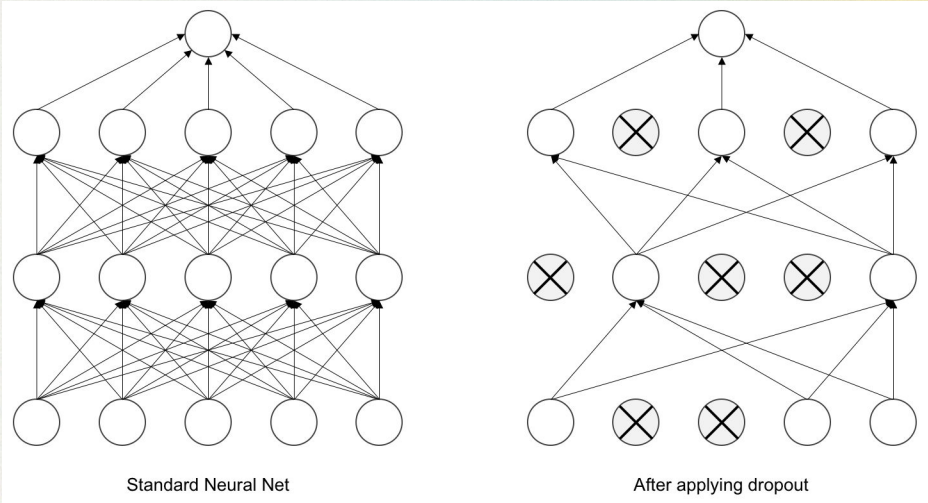
$$g \in \mathbb{R}^{d_c}$$

V_c^{σ} - uncertainty parameter
- Model's uncertainty about inferred z

$$V_c^{\sigma} \in \mathbb{R}^{l \times d_c}$$

Regularization via Dropout

- How to choose the latent space size?
  - Heuristic method
    - Vary l and look for loss inflection
  - Automatic Sparsity
    - Variational dropout



- Start with an ‘overcomplete’ latent space
  - Let model prune dimensions

Experiments

Synthetic Data

- Train sparse/non-sparse on different channels with various latent dimension sizes

Attribute description	Iteration list
Total channels (C)	2 3 5 10
Channel dimension (d_c)	32
Latent space dimension (l)	1 2 4 10 20
Samples (training and testing)	100 1000
Signal-to-noise ratio (snr)	10 1
Seed (re-initialize $\mathbf{R}_c$)	1 2 3 4 5

Medical Imaging Data

- Compare their sparse model to various baselines
  - AD prediction on latent features
- Imputation & Generation
  - Pathological aging
  - Cross modality generation

Datasets

Synthetic Dataset

- Datasets: $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_C\}$

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}; \mathbf{I}_l),$$
$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}; \mathbf{I}_{d_c}),$$
$$\mathbf{G}_c = diag\left(\mathbf{R}_c \mathbf{R}_c^T\right)^{-1/2} \mathbf{R}_c,$$
$$\mathbf{x}_c = \mathbf{G}_c \mathbf{z} + snr^{-1/2} \cdot \boldsymbol{\epsilon},$$

- *z:* vector drawn from normal distribution
  - *l* dimensional
- ϵ: noise vector
  - d_c dimensional
- R_c: random matrix with *l* orthogonal columns
- G_c: decoding matrix for channel c
- *snr*: signal to noise ratio controlling the noise
- d_c: dimensionality of each channel

ADNI Dataset

- Randomly selected 504 subjects
- Clinical channel
  - age; results to mini-mental state examination, adas-cog, cdr, and faq tests; scholarity level
- MRI, Amyloid-PET, FDG-PET
  - Averaged into 90 AAL regions
- Baselines:
  - MCVAE
    - Their non-sparse model
  - IVAE
    - Stack of VAEs
  - VAE
    - Vanilla VAE

Synthetic Experiments

- Here they're looking at the choice for latent dimension
- For their sparse model, you can see that a lot of the parameters for the latent and generative got pruned out compared to the non-sparse.

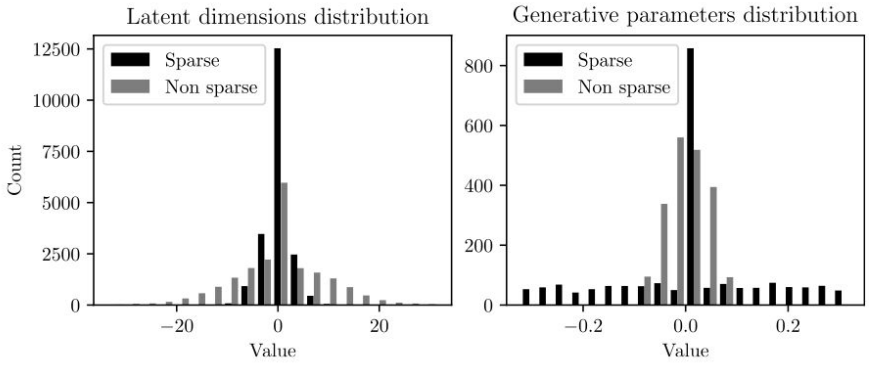


Figure 1: Effect of variational dropout on a synthetic experiment modeled with the Multi-Channel VAE. As expected, the minimum amount of non-zero components of $\mathbf{z}$ (left) and generative parameters $\mathbf{G}$ (right) is obtained with the sparse model.

Benchmark

- Looking at the reconstruction error for their baselines on synthetic data
- Sparse model generally has lower reconstruction loss
- Number of channels seems to have an important role

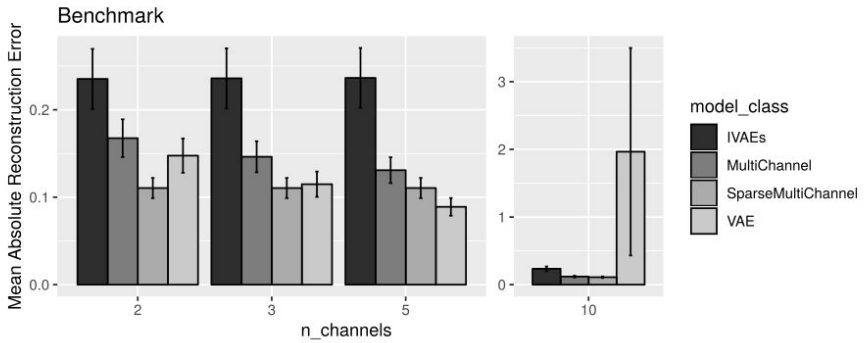


Figure 3: Testing benchmark of four variational methods applied to the multi-channel scenarios in Tab. 1 (cases snr $= 10$, $l_{\text{fit}} = l$). Sparse Multi-Channel models performs consistently better than non-sparse Multi-Channel ones.

Dropout Rates

- Looking at dropout rates for their sparse model on synthetic data
- Dropout rate high for useless dimensions
  - Low for ground truth dimensions

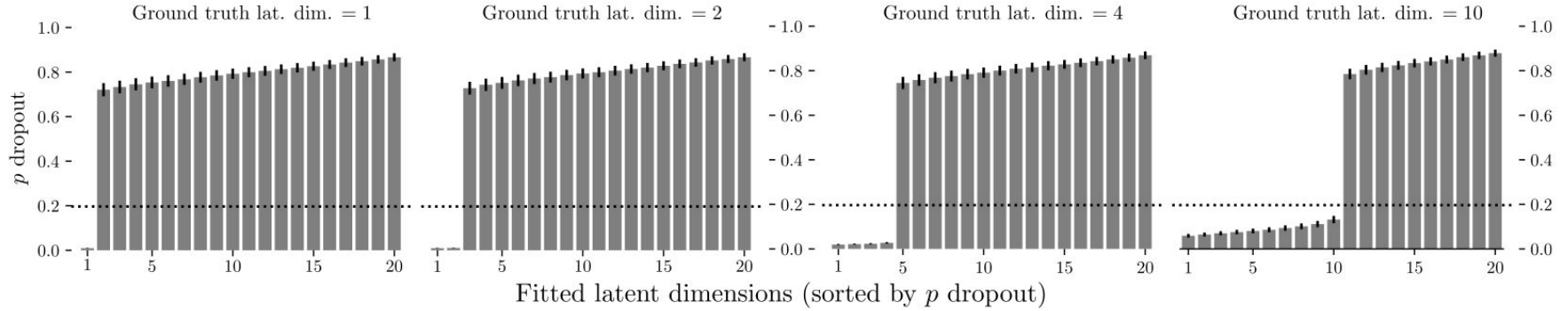


Figure 2: Estimated dropout rates for the latent dimensions when the initial latent dimensions of the Sparse Multi-Channel VAE was set to $l_{\text{fit}} = 20$ on data generated with respectively $l = 1, 2, 4$, and 10 latent dimensions.

ADNI Experiments

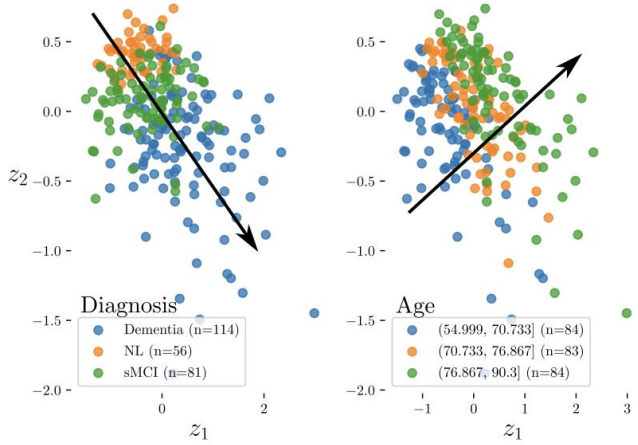


Figure 4: Stratification of the ADNI subjects (test data) in the sparse latent subspace inferred from the first two least dropped out dimensions. In the same subspace it is possible to stratify subjects in the test-set by disease status (left) and by age (right) in almost orthogonal directions. Classification accuracy for these subjects is given in the fifth numeric column of Tab. 3.

- Here they trained their sparse model on the ADNI dataset.
- Model used variational dropout to select most important latent dimensions
- Encoded hold-out subjects to latent space
  - Plotted top 2 dimensions
- Colored data points
  - Age
  - Diagnosis

ADNI Experiments

- Use trained encoder to sample latent data
  - Move along age axis centered at 1 diagnosis
  - HC aging, AD aging
- Use trained decoder to generate new data
  - Push points along the trajectory through decoder
- Generate missing images from different modalities
- Found plausible aging per group (atrophy/none, etc.)

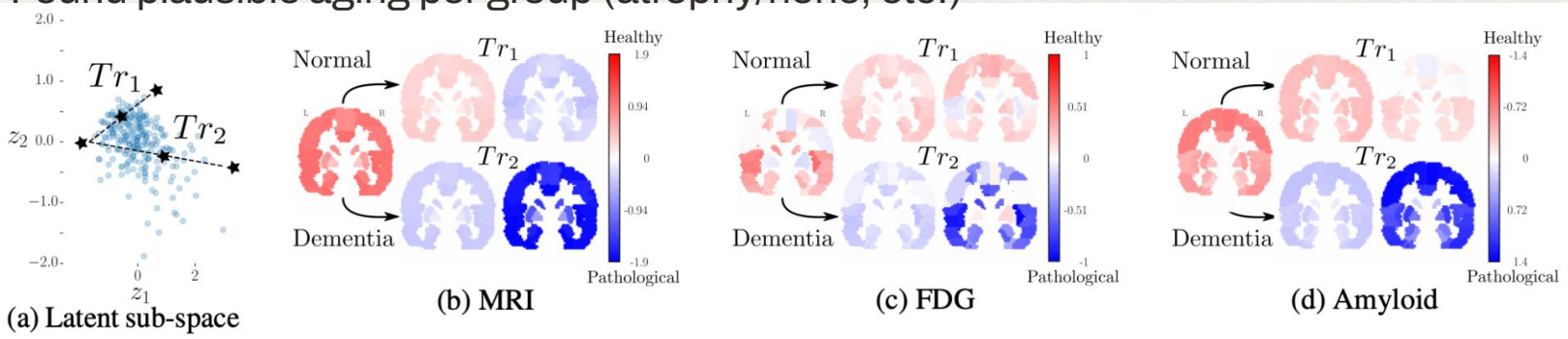


Figure 5: Generation of imaging data from trajectories in the latent space. (a) Normal aging trajectory (Tr_1) *vs* Dementia aging trajectory (Tr_2) in the latent 2D sub-space (*cfr.* Fig. 4). Stars indicate the sampling points along trajectories. The trajectories share the same origin. MRIs (b), FDG (c), and Amyloid PET (d). All the trajectories show a plausible evolution across disease and healthy conditions.

ADNI Experiments

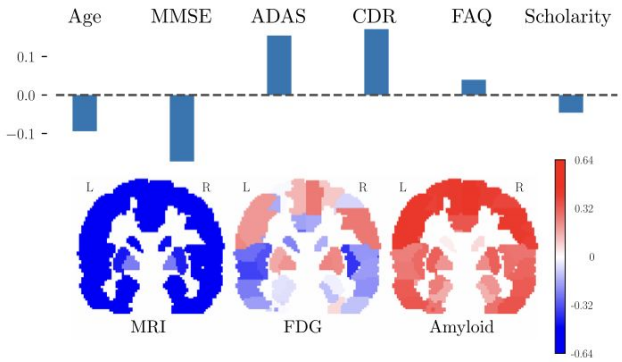


Figure 6: Generative parameters $\mathbf{G}_c^{(\mu)}$ (*cfr.* Eq. (7)) of the four channels associated to the least dropout latent dimension in the sparse multi-channel model. (Top) Clinical channel parameters. (Bottom) Imaging ch. parameters.

- Here they're looking at a biomedical interpretation of what each latent dimension represents
- Dimensions with lowest dropout rate
- Plot the weights of the G matrix
  - Rows: observed features
  - Cols: latent features

Resources/References

1. https://deeplearning.cs.cmu.edu/S21/document/recitation/recitation9.pdf
2. https://www.youtube.com/watch?v=Dp6iICL2dVI
3. http://doi.org/10.15439/2023F865
4. https://www.surfertoday.com/surfing/what-is-a-tsunami#google_vignette
5. https://www.eurekalert.org/multimedia/1054477
6. https://web.stanford.edu/class/bios221/Pune/Lectures/Lecture_Day1_heterogeneity.pdf
7. https://arxiv.org/abs/2410.08245

Discussion & Questions!