Are Deep Neural Network latent spaces a good model for human brain representations?

Leila Reddy
Rufin VanRullen

COMMUNICATIONS BIOLOGY

ARTICLE

https://doi.org/10.1038/s42003-019-0438-y OPEN

Reconstructing faces from fMRI patterns using deep generative neural networks

Rufin VanRullen[1] & Leila Reddy[1]

Deep Neural Networks

- Convolutional Neural Networks (supervised networks)
  - AlexNet, VGG, GoogLeNet, ResNet
  - Excellent performance on object recognition and classification

- Unsupervised learning (e.g., **G**enerative **A**dversarial **N**etworks)
  - Latent spaces have intuitive properties.
  - What is a latent space? ~ Internal representation of a network.

Outline

- Properties of Latent spaces : intuitively, they seem to make sense.
  - Natural Language Processing (NLPs): word embeddings
  - Computer Vision: Face processing
- General Framework : DNN latent spaces a good model for brain representations?
- Latent spaces : how do we get them?
- Testing our hypothesis: fMRI brain decoding

Example 1: NLP word embeddings

- Natural language processing (NLP):
  - Create a word embedding or latent space
  - Fairly low dimensional (e.g., 300 or 500 dimensional)
  - A word is represented by a vector in this space.
  - Vector operations make sense

QUEEN – WOMAN + MAN = KING

vector $\boldsymbol{x}$ defined as:	Example 1	Example 2
$\boldsymbol{x}$ = Paris – France	Italy + $\boldsymbol{x}$ = Rome	Japan + $\boldsymbol{x}$ = Tokyo
$\boldsymbol{x}$ = bigger – big	cold + $\boldsymbol{x}$ = colder	quick + $\boldsymbol{x}$ = quicker
$\boldsymbol{x}$ = scientist – Einstein	Messi + $\boldsymbol{x}$ = midfielder	Mozart + $\boldsymbol{x}$ = violinist
$\boldsymbol{x}$ = Cu – copper	zinc + $\boldsymbol{x}$ = Zn	gold + $\boldsymbol{x}$ = Au
$\boldsymbol{x}$ = sushi – Japan	Germany + $\boldsymbol{x}$ = bratwurst	USA + $\boldsymbol{x}$ = pizza

Mikolov, T. *et al. ArXiv e-prints* arXiv:1310.4546 (2013)

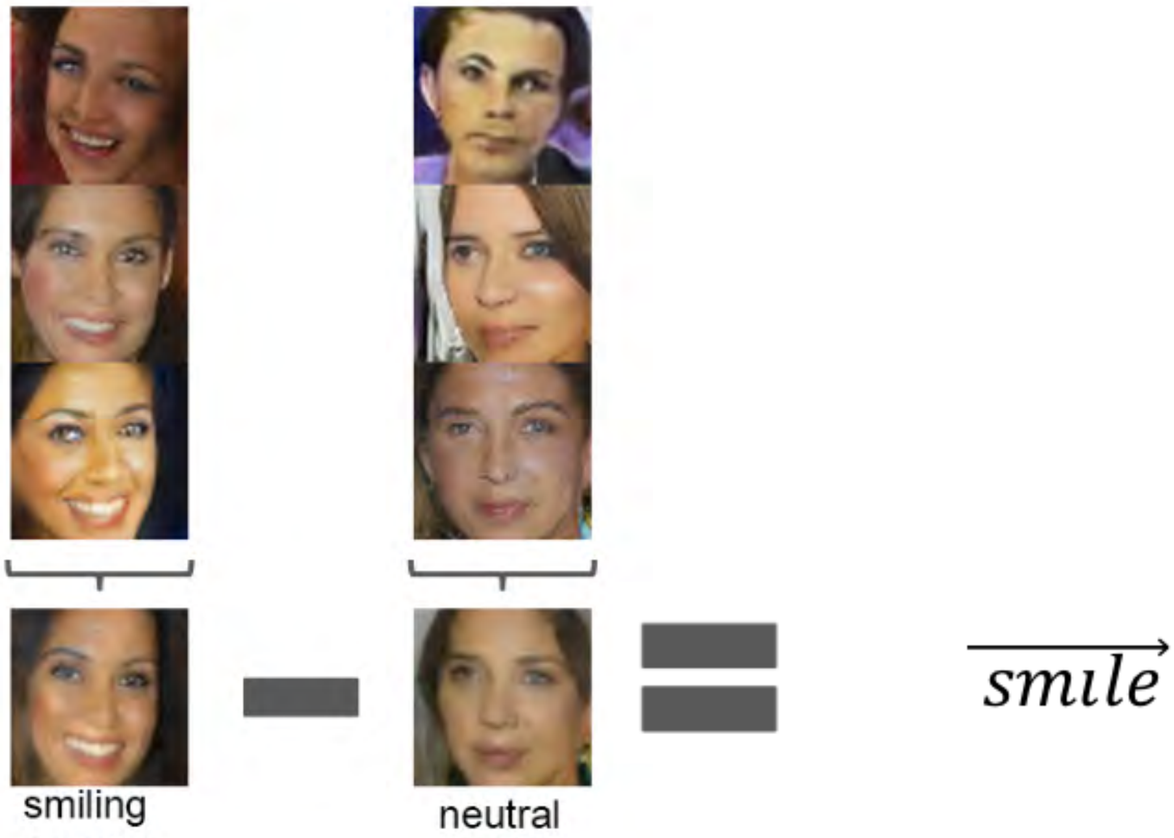
Example 2: Face latent spaces

- Computer Vision:
  - Example: a **G**enerative **A**dversarial **N**et trained on celebrity faces
  - Creates a latent space, e.g., a 500 or 1000 dimensional space
  - A point/vector in this space corresponds to a face
  - GAN: generative model ➔ generate a face from a vector
  - Perform operations on these vectors and look at the faces that are generated
  - Vector operations make sense?

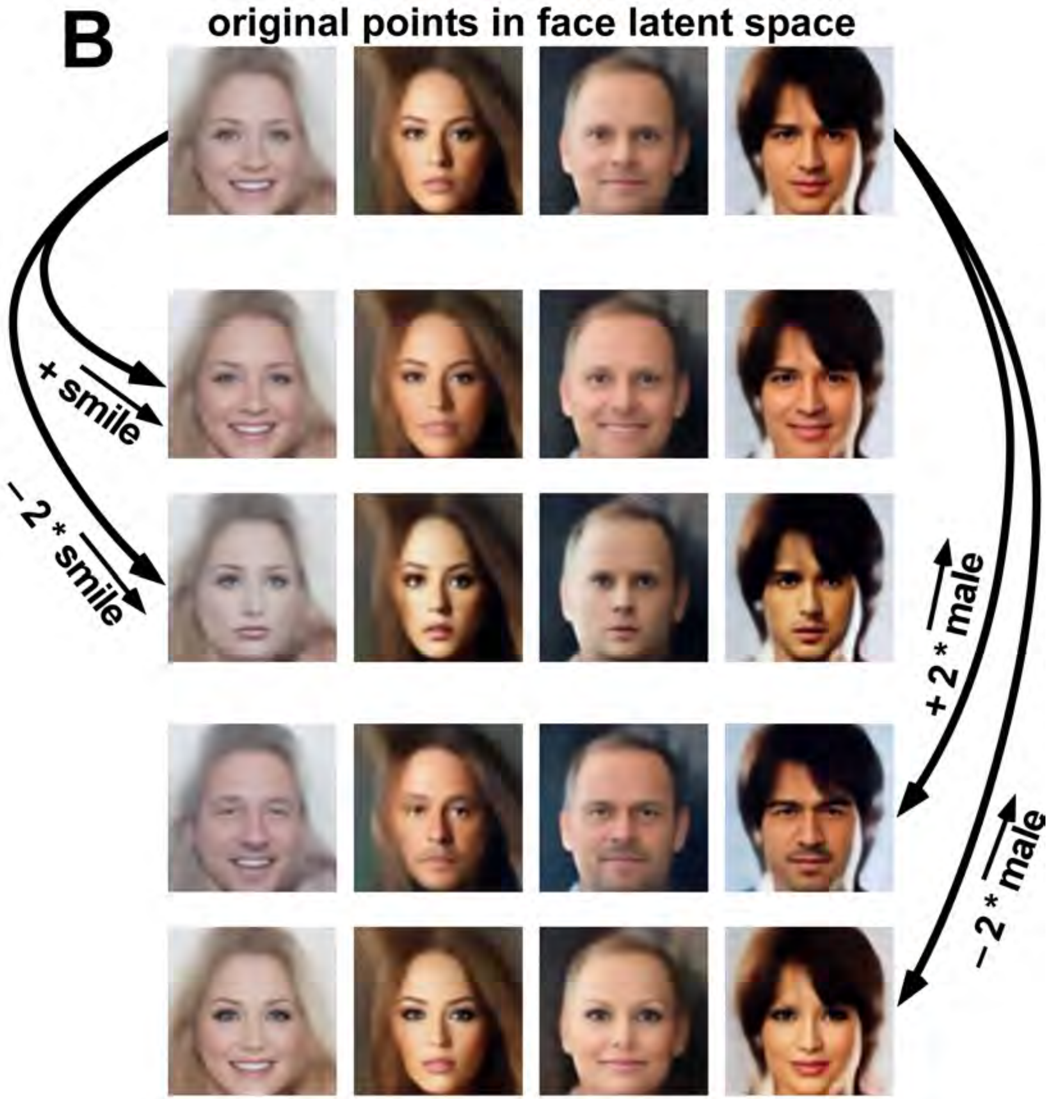
Latent space interpolations & extrapolations

White,T .ArXiv e-prints arXiv:1609.04468 (2016).

Vector operations in a face latent space



Vector operations in a face latent space



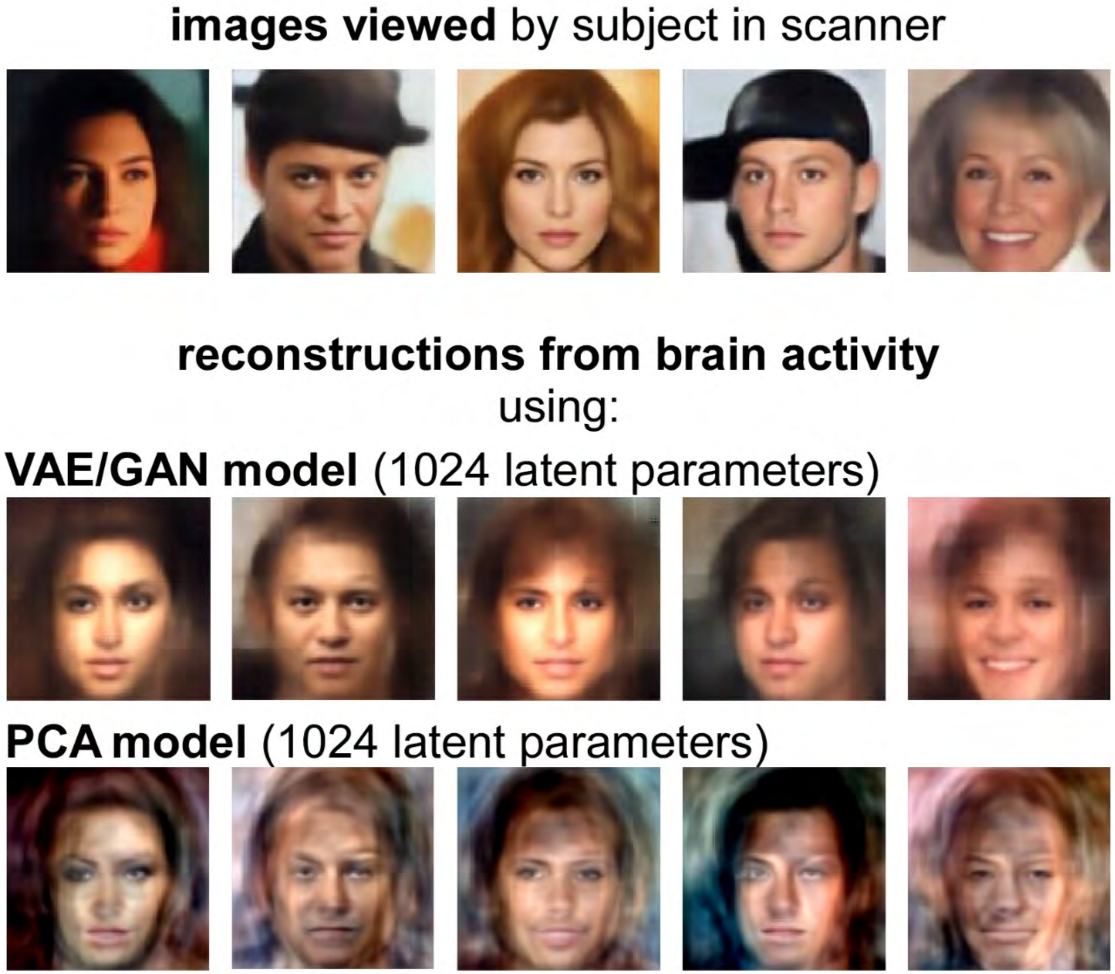
DNN latent spaces : a good model for brain representations?

General Framework:

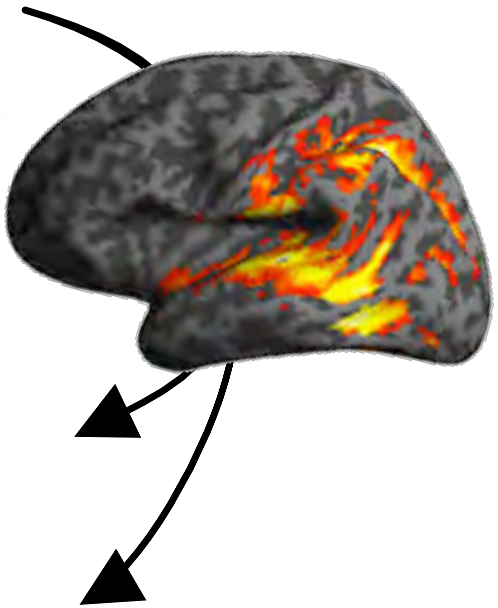
DNN latent spaces are a good model for brain representations.

- Clarification:
  - Not about one specific model, one particular dimensionality, one type of GAN….
  - A whole class of models might be similar to biological representations.
- Prediction:
  - DNN latent spaces allow for better fMRI brain decoding.

DNN Latent Spaces for fMRI Decoding



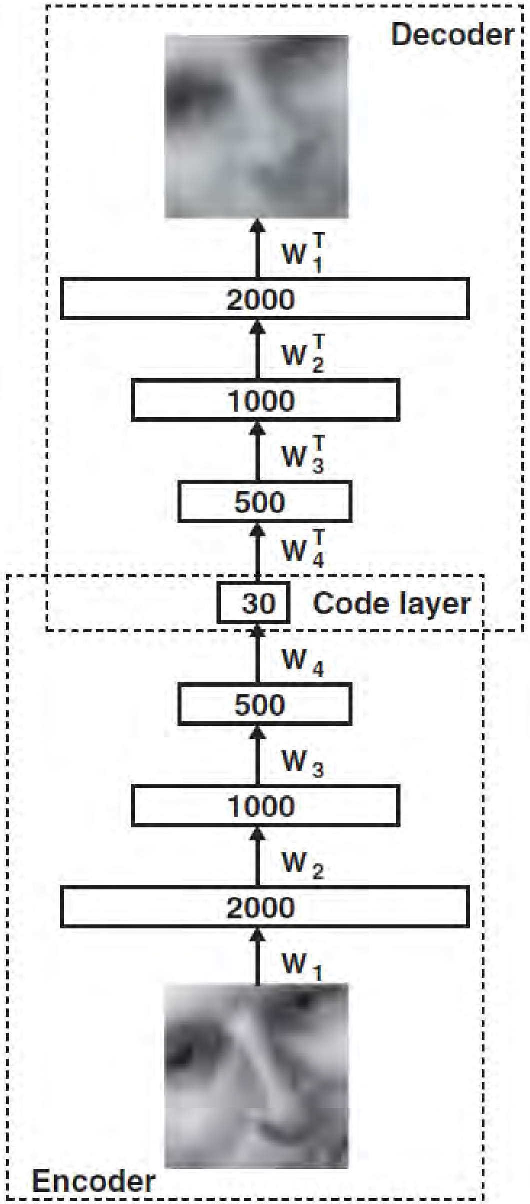
Cowen et al., 2014; Lee et al., 2016.



Successfully decode:
- **face identity**
- **face gender**
- **imagined faces**

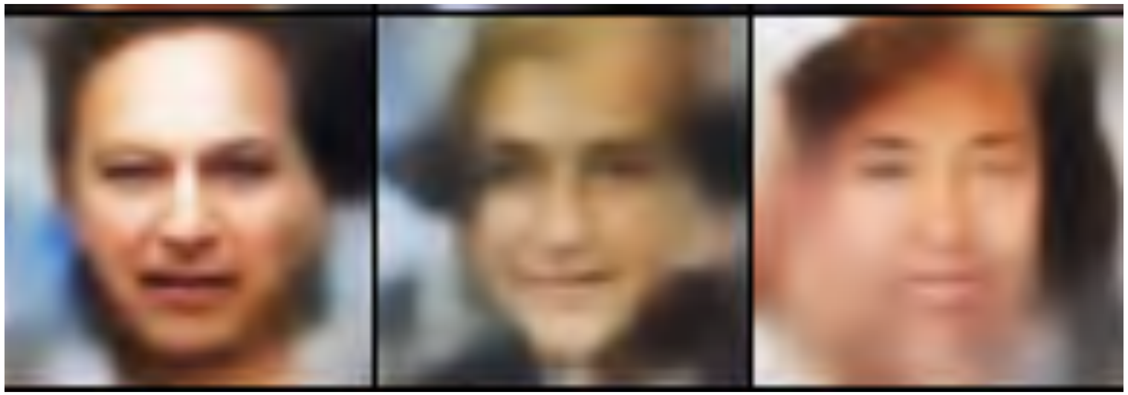
VanRullen & Reddy, 2019. *Communications Biology*

Unsupervised generative models: Auto-encoders

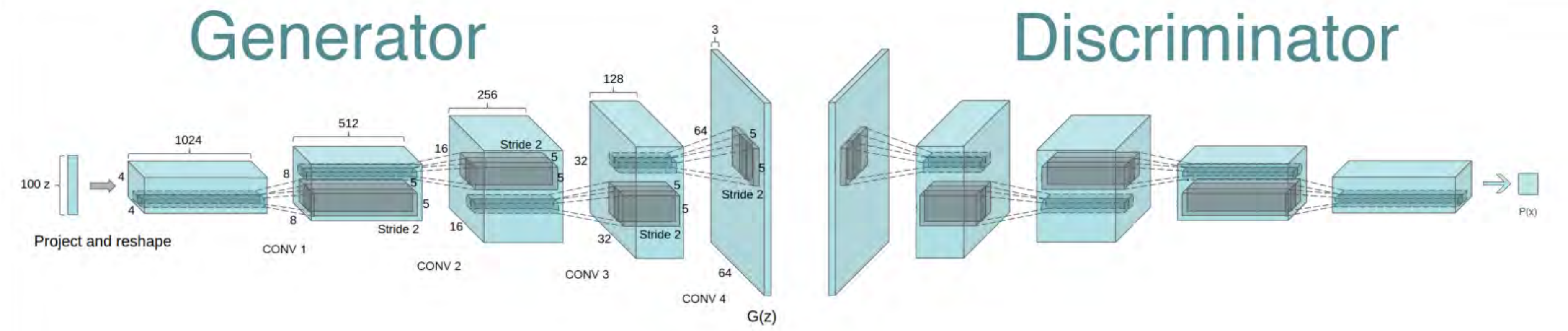


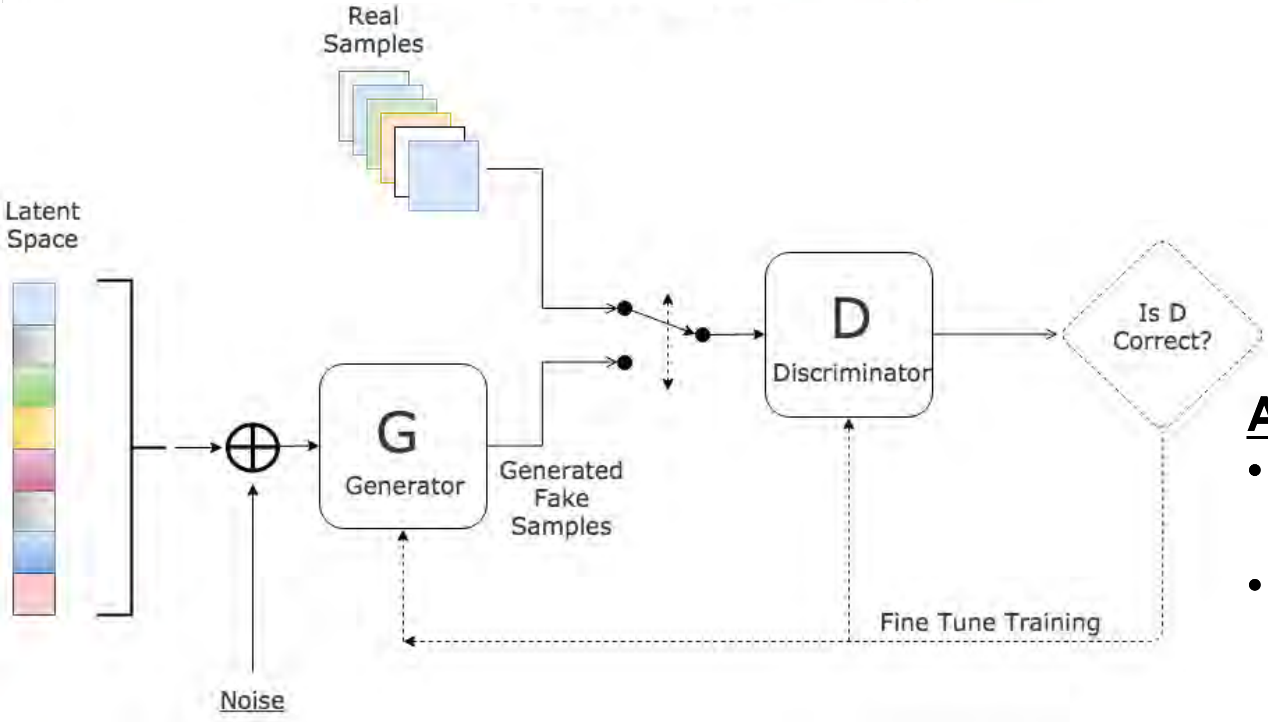
- The whole network is trained (e.g. back-prop) to minimize the reconstruction loss (MSE between input/output images)

- It needs to learn a useful feature hierarchy
- The "code" defines a "**latent space**" of efficient dimensions.
- Problem: Loss defined in pixel space: encourages blurry samples.



Generative Adversarial Networks (GAN)



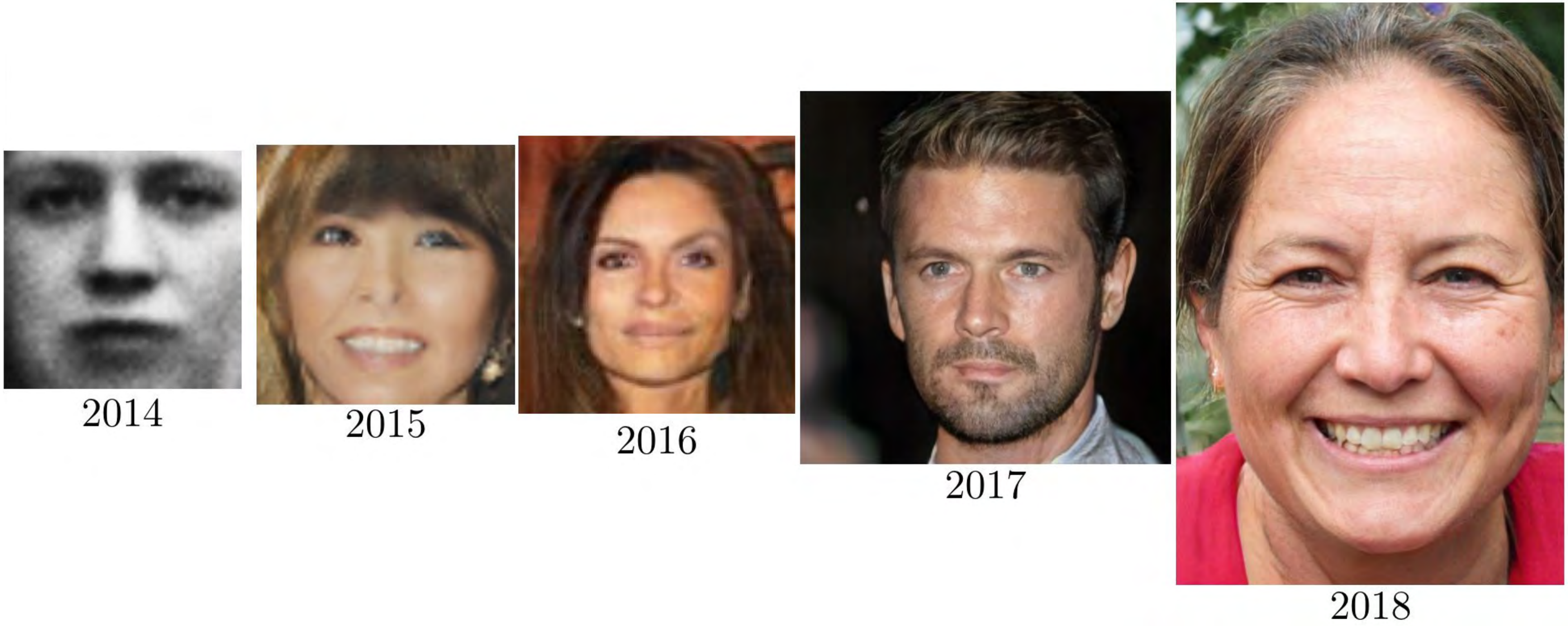


Adversarial Training:

- D is rewarded when it is right
- G is rewarded when D is wrong

Goodfellow, I. J. *et al. ArXiv e-prints arXiv:1406.2661* (2014).

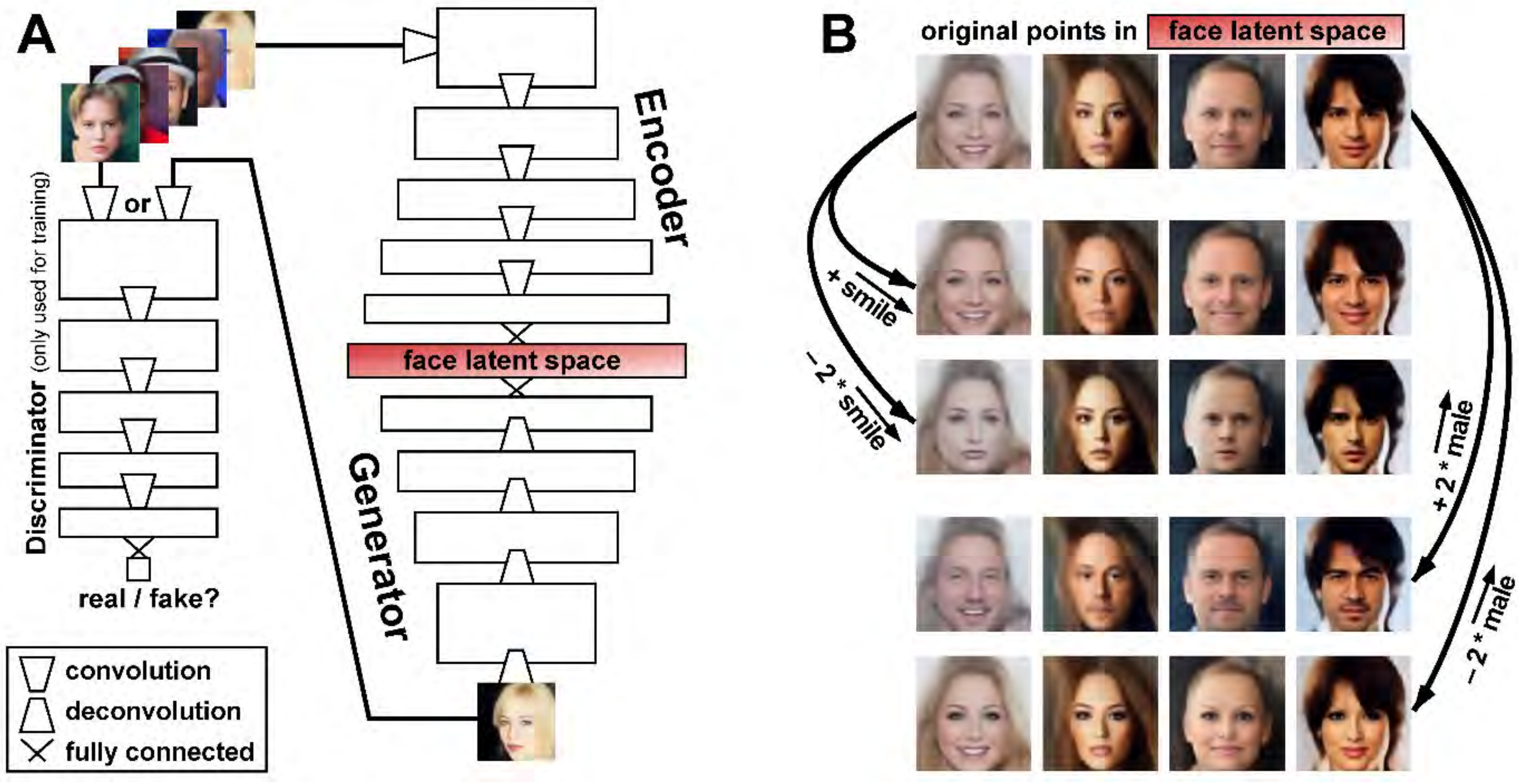
Generative Adversarial Networks (GAN)



- These are **fake** faces (not real photographs but generated by the network)!
- https://thispersondoesnotexist.com/

Goodfellow, I. J on Twitter, 2019

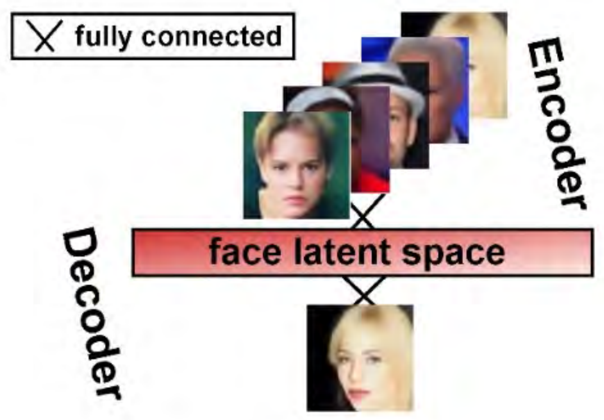
VAEGAN



- The network is trained on a celebrity dataset (200,000 images).
- The "code" defines a "**latent space**" of 1024 efficient dimensions.
- After training, the weights are frozen and the discriminator network is dropped.

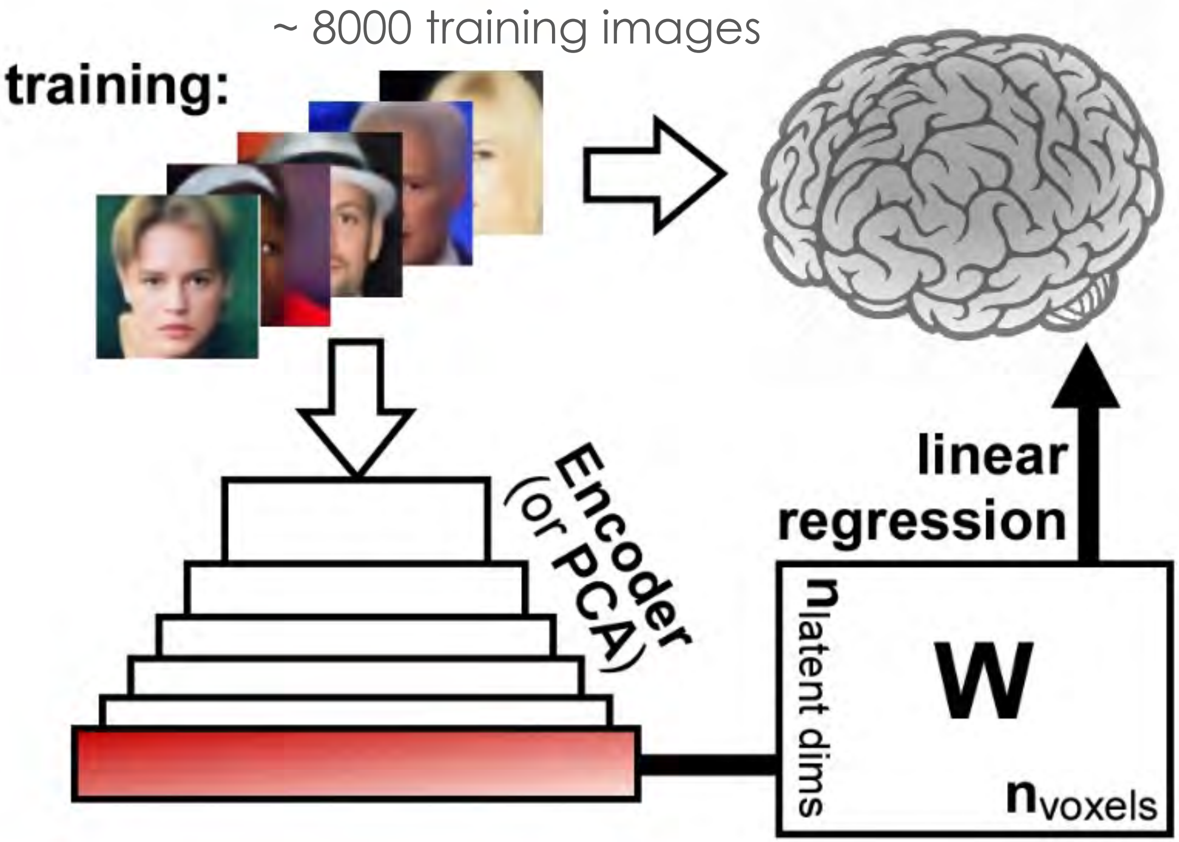
Larsen, A. B. L., Sønderby, S. K., Larochelle, H. & Winther, O. *ICML'16*).

PCA model



- Faces are encoded into a latent space of 1024 principal components (Cowen et al., 2014; Lee et al., 2016.).

Training a brain decoder (GLM)



$$Y = XW \quad \Leftrightarrow \quad X^TY = X^TXW$$

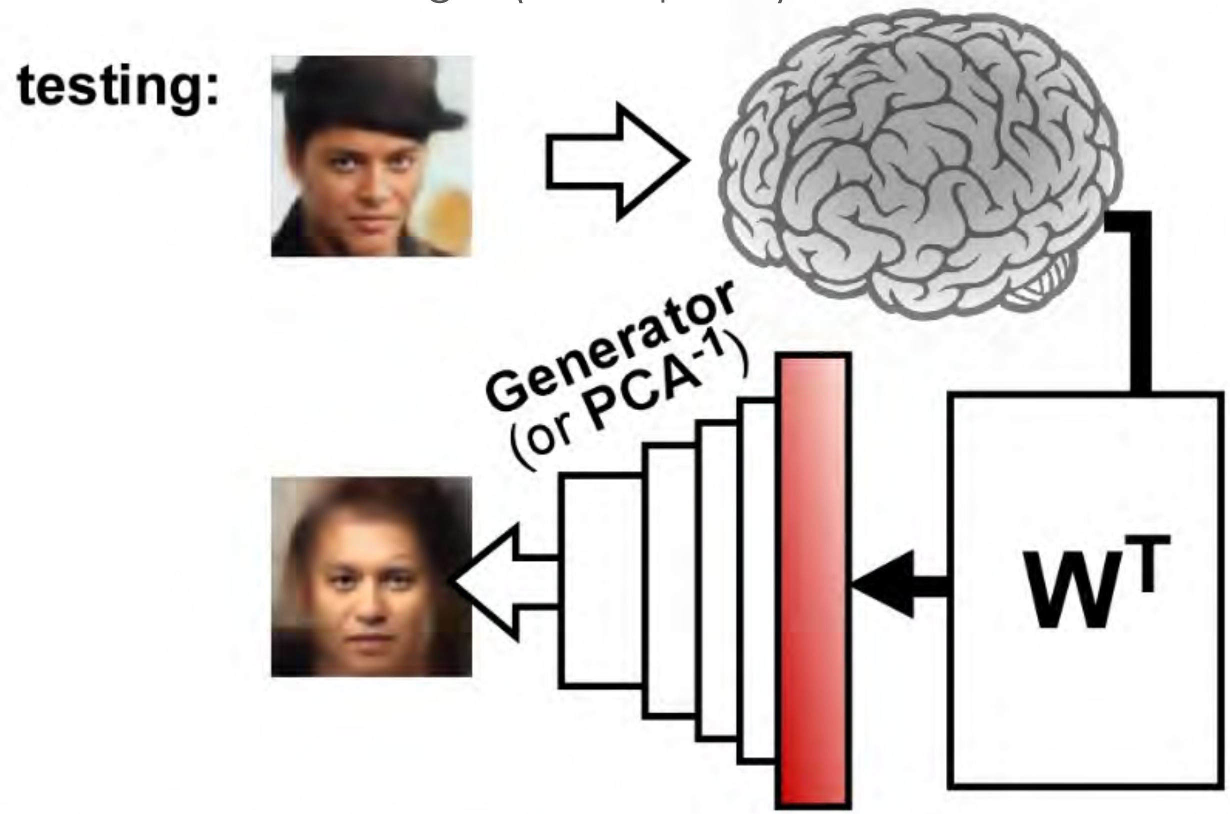
Y: $(8000, n_{voxels})$ X: $(8000, 1024)$ W: $(1024, n_{voxels})$

$$\boxed{W = (X^TX)^{-1}X^TY}$$

VanRullen & Reddy, 2019. *Communications Biology*

Testing the brain decoder (GLM)

20 test images (x 45 repeats)

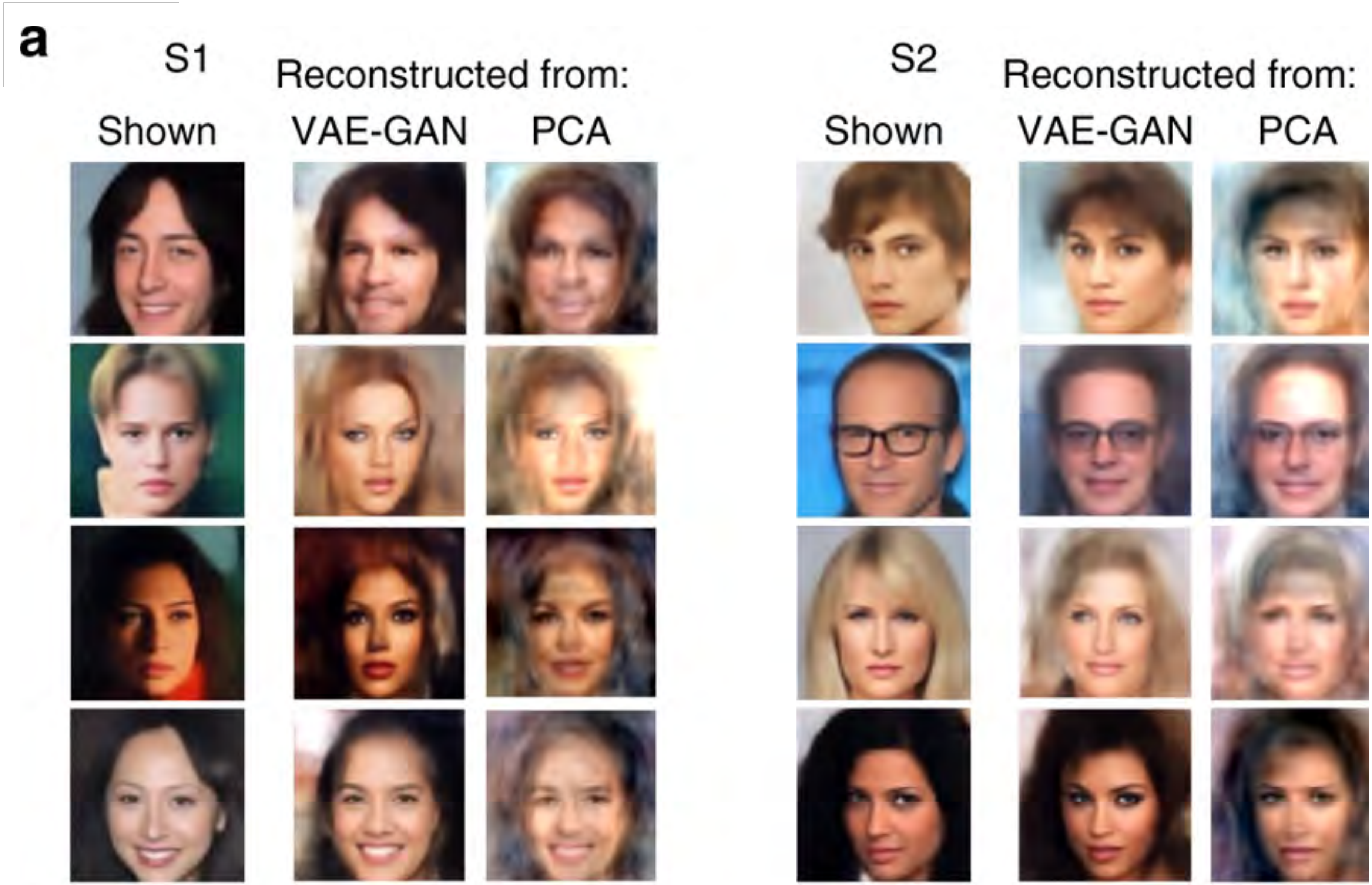


$$\underset{(20,n_{voxels})}{Y} = \underset{(20,1024)}{X}\,\underset{(1024,n_{voxels})}{W} \quad \Leftrightarrow \quad YW^T = XWW^T$$

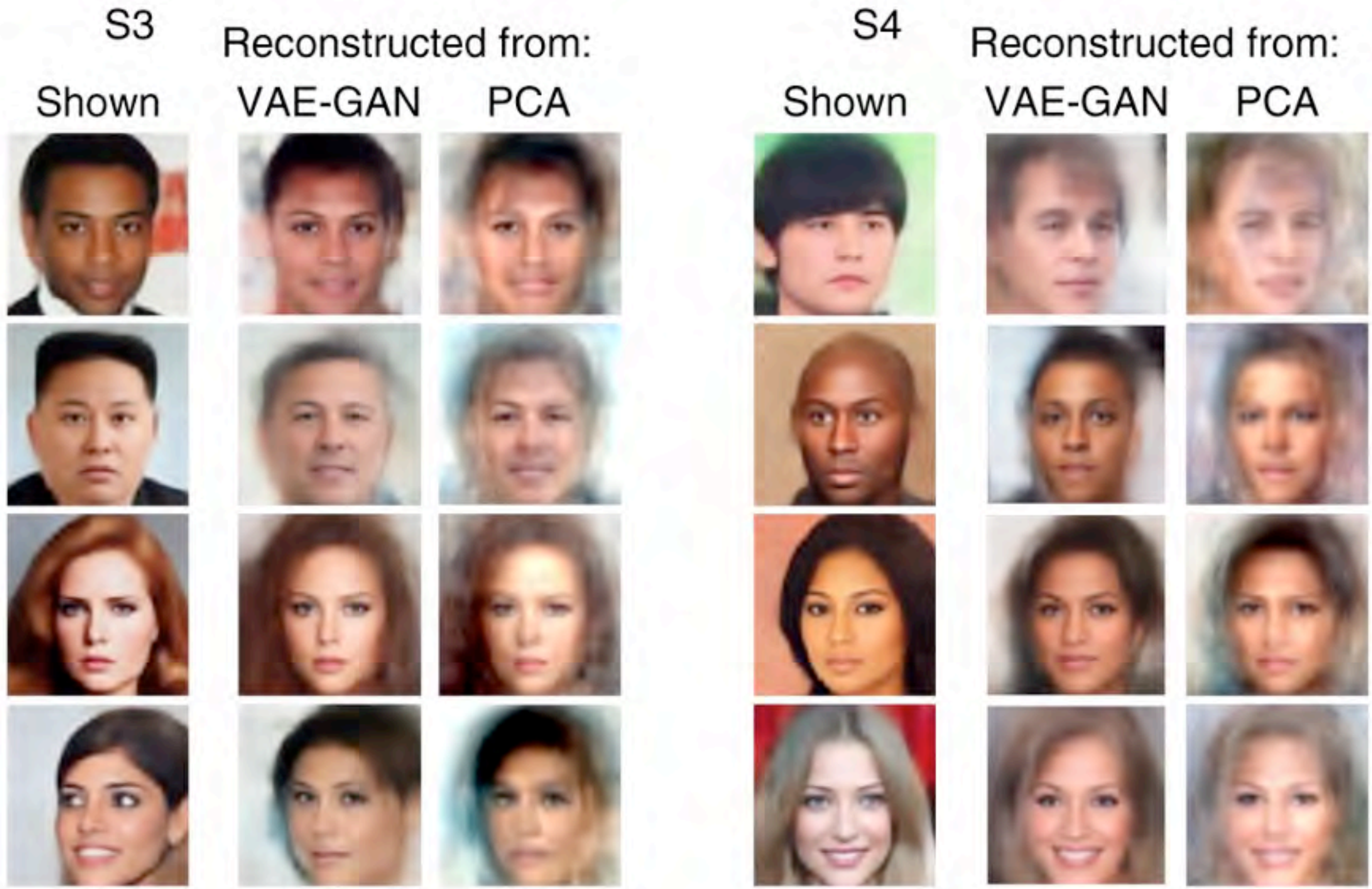
$$\boxed{X = YW^T(WW^T)^{-1}}$$

VanRullen & Reddy, 2019. *Communications Biology*

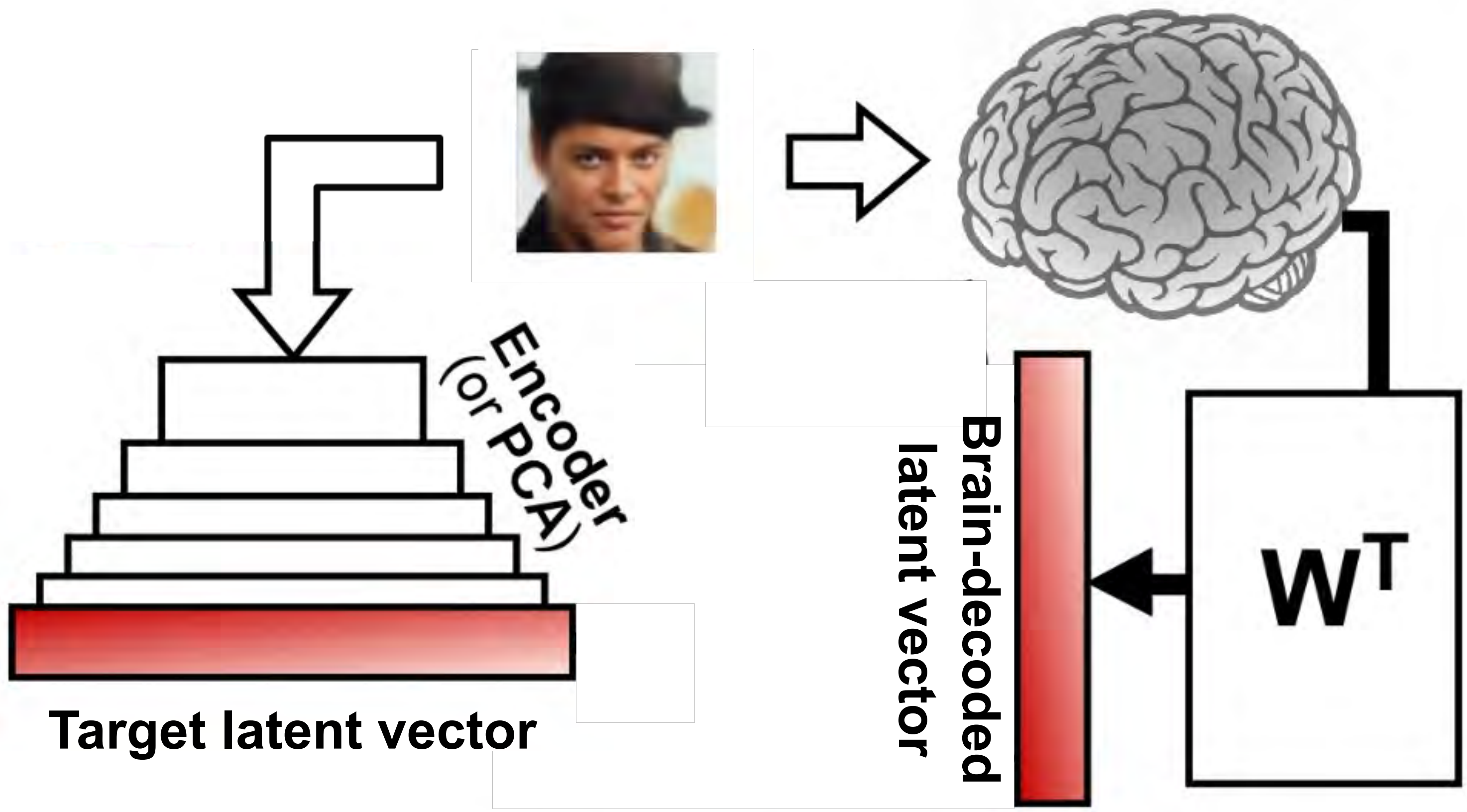
Face Decoding and Reconstruction



Face Decoding and Reconstruction

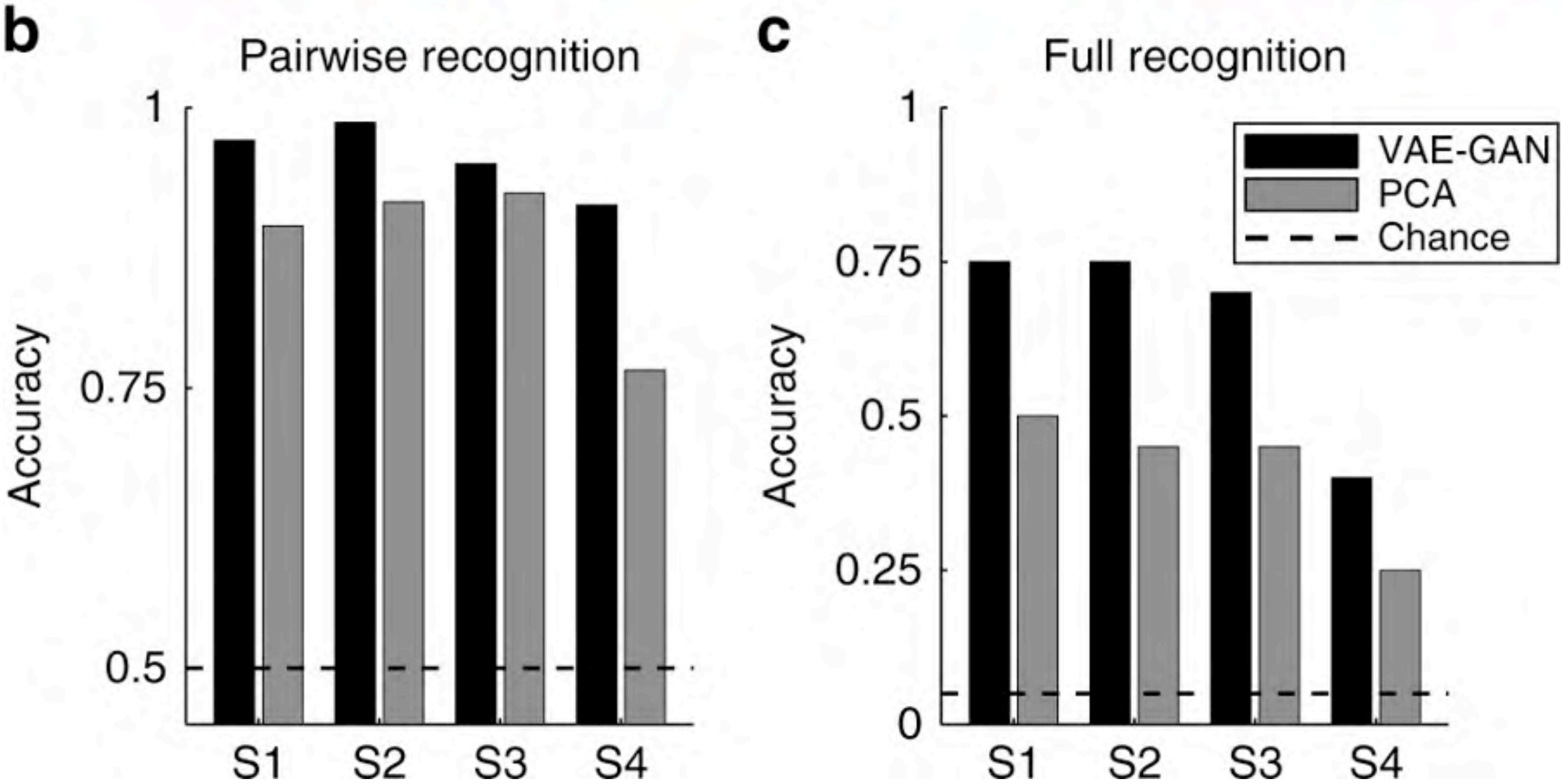


Face Decoding and Reconstruction



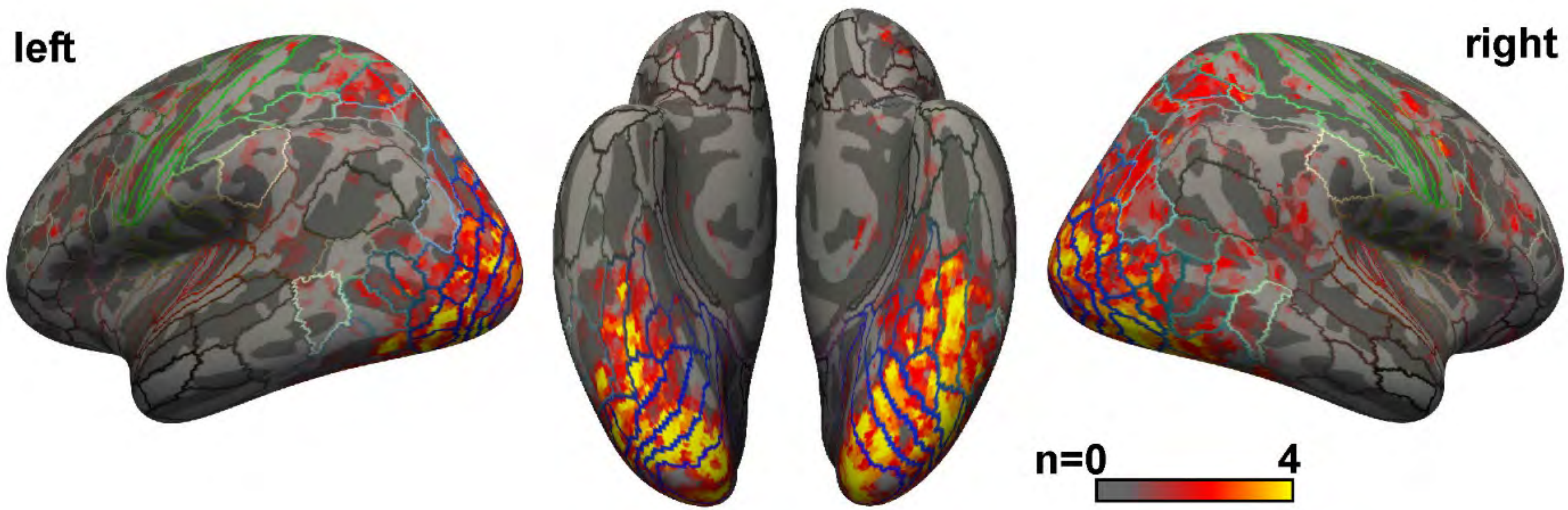
VanRullen & Reddy, 2019. *Communications Biology*

Face Decoding and Reconstruction



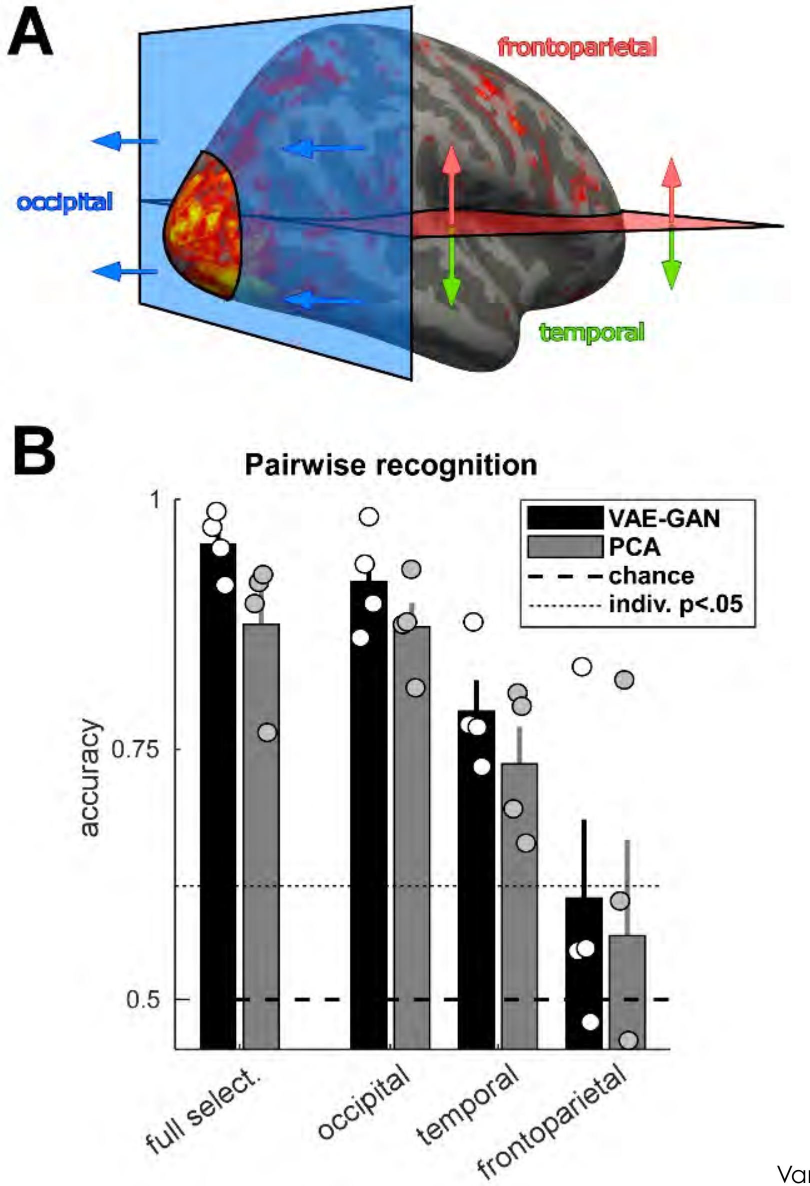
VanRullen & Reddy, 2019. *Communications Biology*

Voxels Selected for Brain Decoding



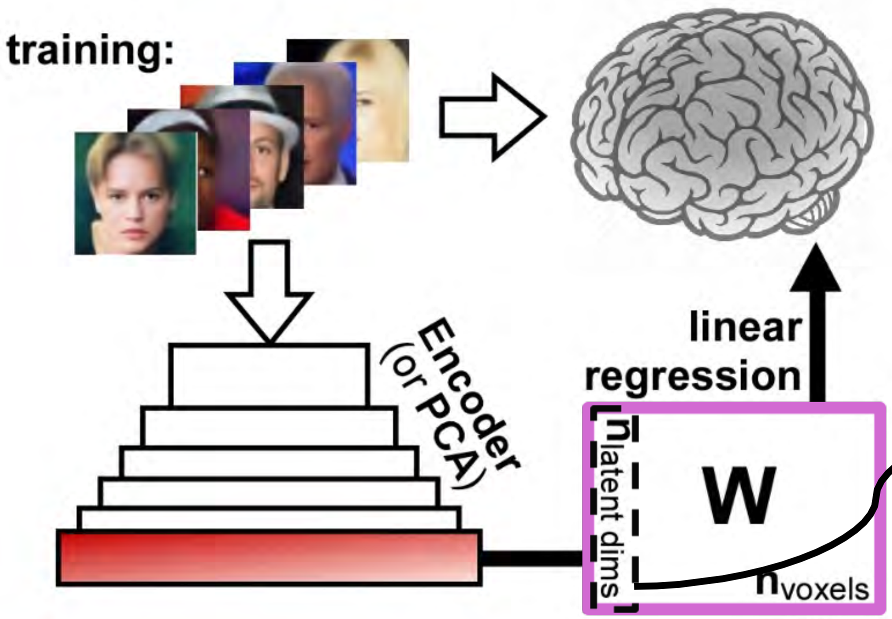
VanRullen & Reddy, 2019. *Communications Biology*

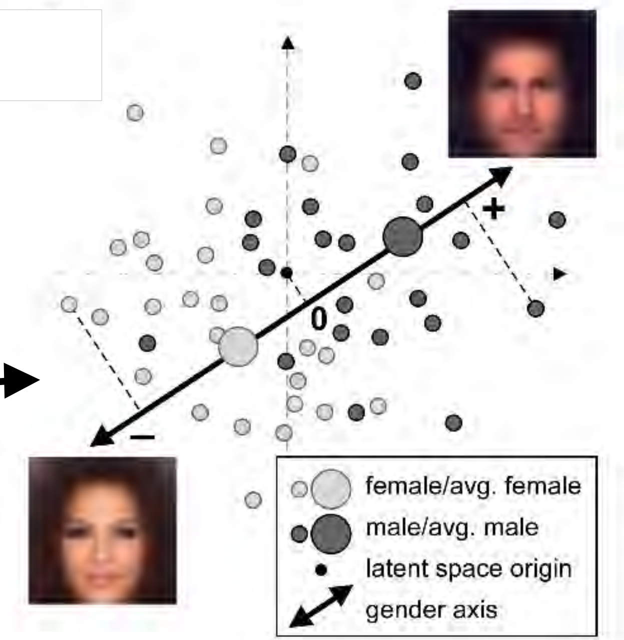
Contribution of Different Brain Regions



VanRullen & Reddy, 2019. *Communications Biology*

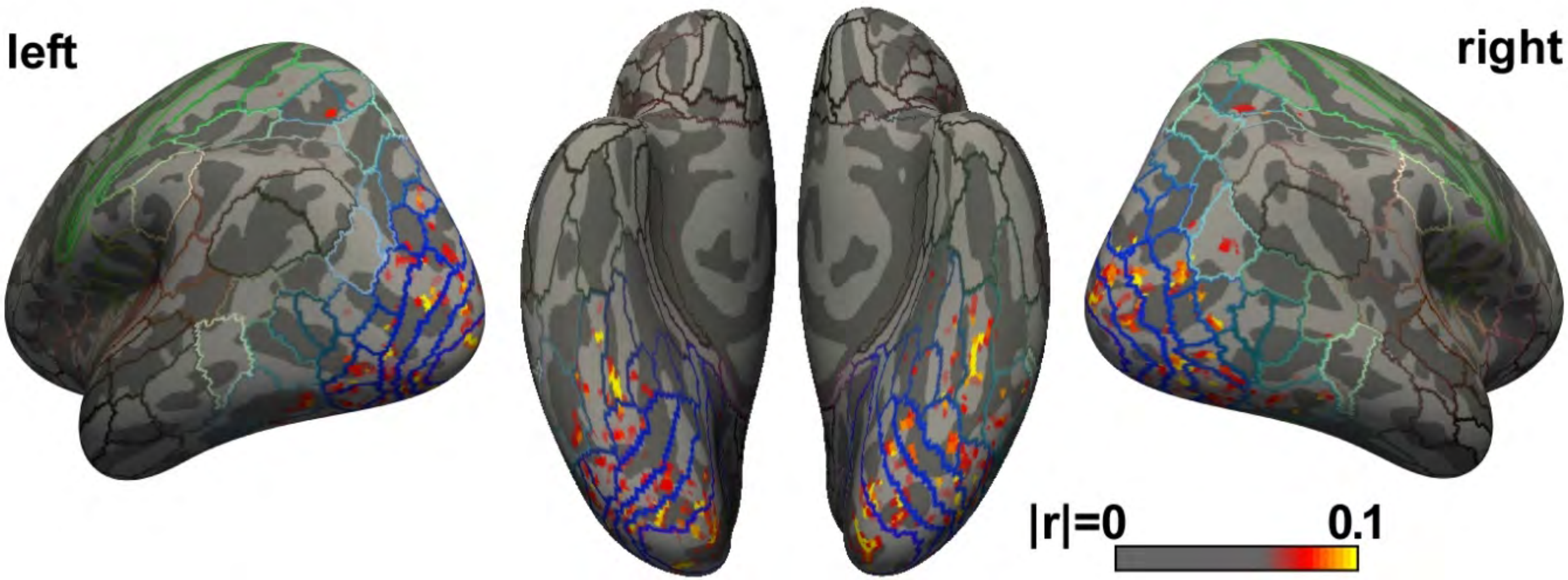
W matrix : a treasure trove for exploring face representations





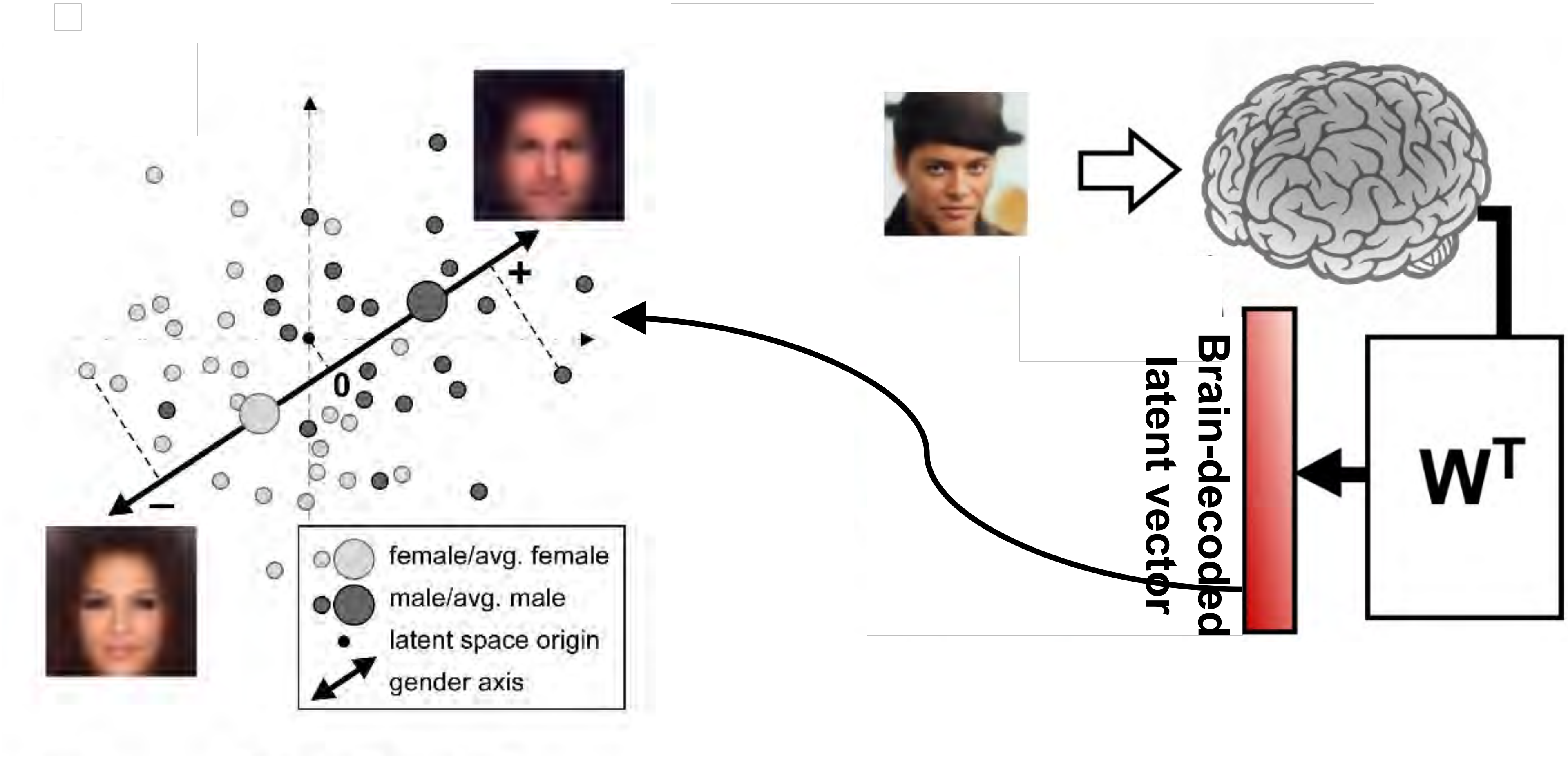
VanRullen & Reddy, 2019. *Communications Biology*

Gender activation map



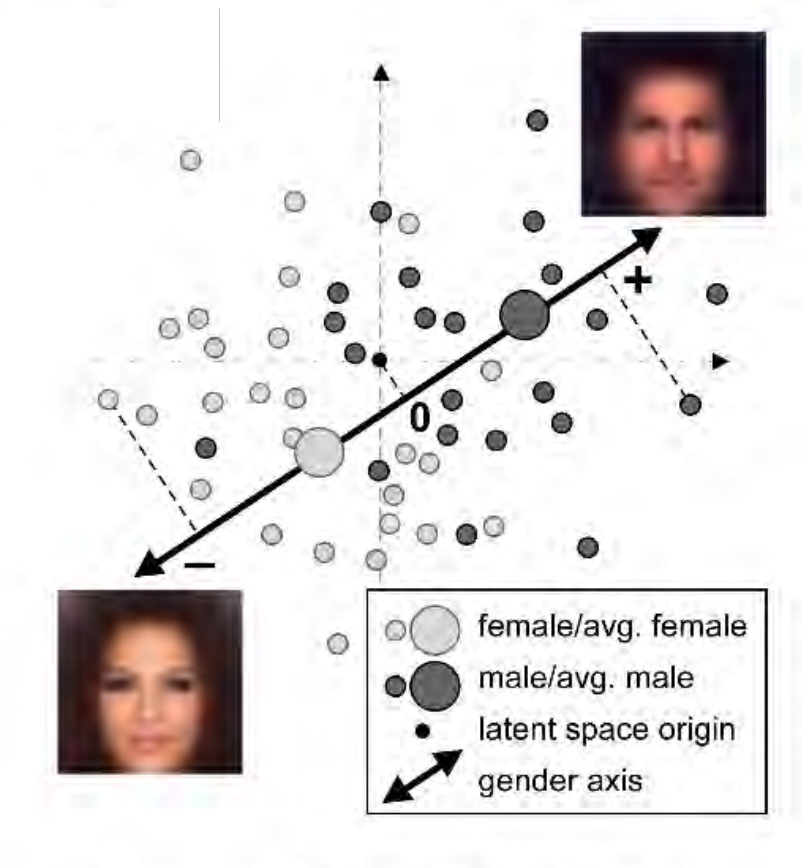
VanRullen & Reddy, 2019. *Communications Biology*

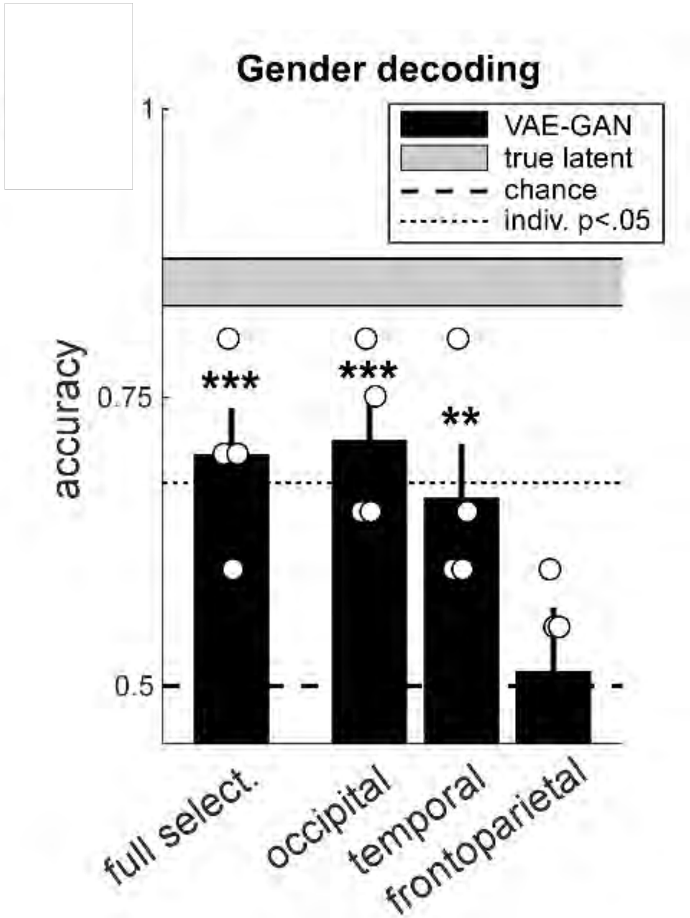
W matrix : a treasure trove for exploring face representations



VanRullen & Reddy, 2019. *Communications Biology*

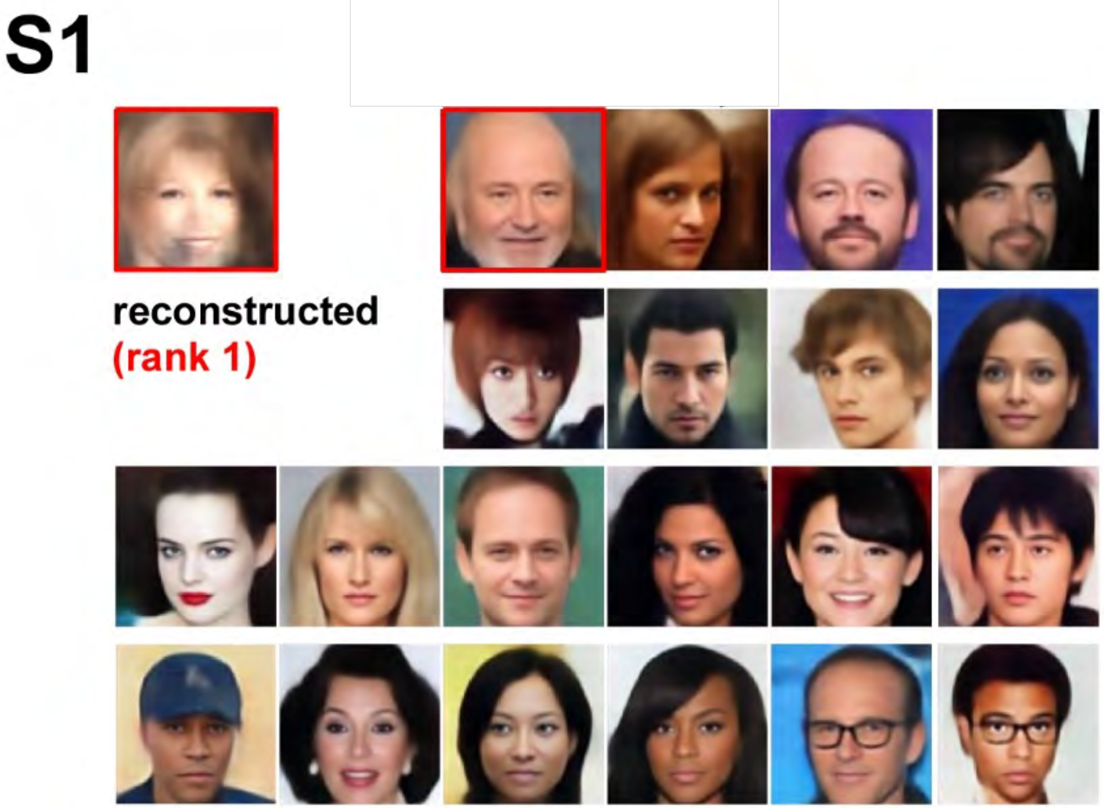
Gender Classification

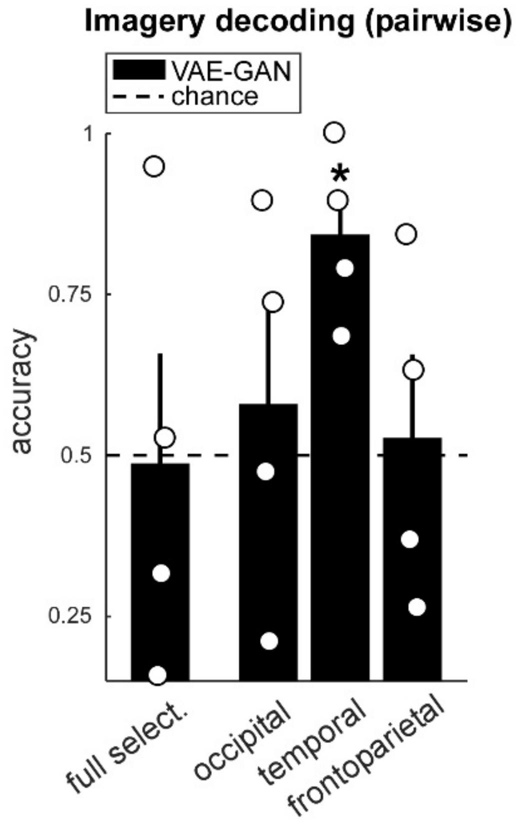




VanRullen & Reddy, 2019. *Communications Biology*

Imagery Decoding





VanRullen & Reddy, 2019. *Communications Biology*

Conclusions

- Superior decoding and reconstructions for face identity, face gender, and imagined faces in the VAEGAN latent space
  - Compared to the PCA space
  - Compared to the state-of-the-art in the literature

⇒ The VAEGAN latent space is a better representation space for linear brain decoding of faces.

- The VAEGAN latent space(and similar network spaces) is topologically similar to the face space in the brain?

  ⇒ Both the artificial and biological neural nets "unfold" the complexity of the face representation space, making it more linear (e.g., DiCarlo & Cox, 2007).