Generative AI Models
ECE 598 LV – Lecture 25

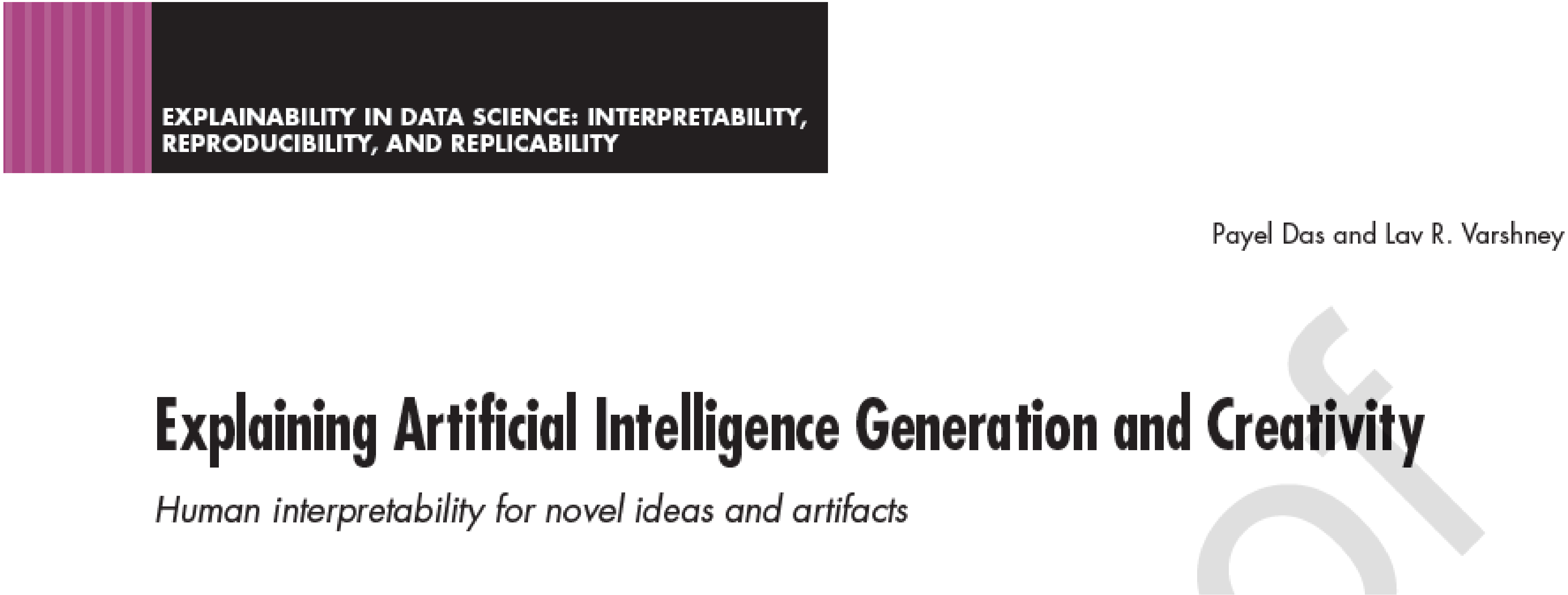
Lav R. Varshney

21 April 2022

- Group project
  - Group formation

- Interpretability

- Why?
- What?
- How?



EXPLAINABILITY IN DATA SCIENCE: INTERPRETABILITY, REPRODUCIBILITY, AND REPLICABILITY

Payel Das and Lav R. Varshney

Explaining Artificial Intelligence Generation and Creativity

Human interpretability for novel ideas and artifacts

(to appear in *IEEE Signal Processing Magazine*)

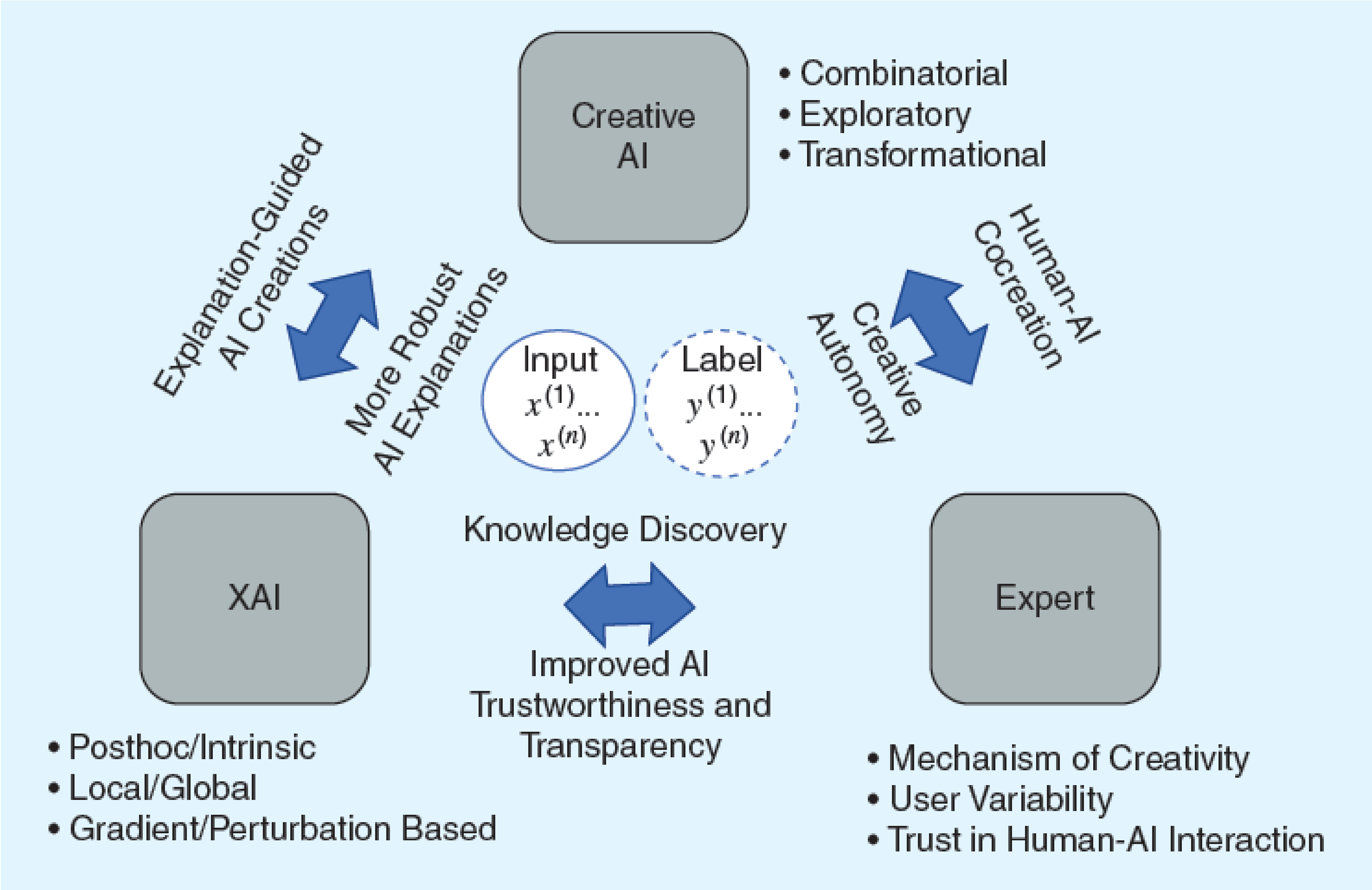
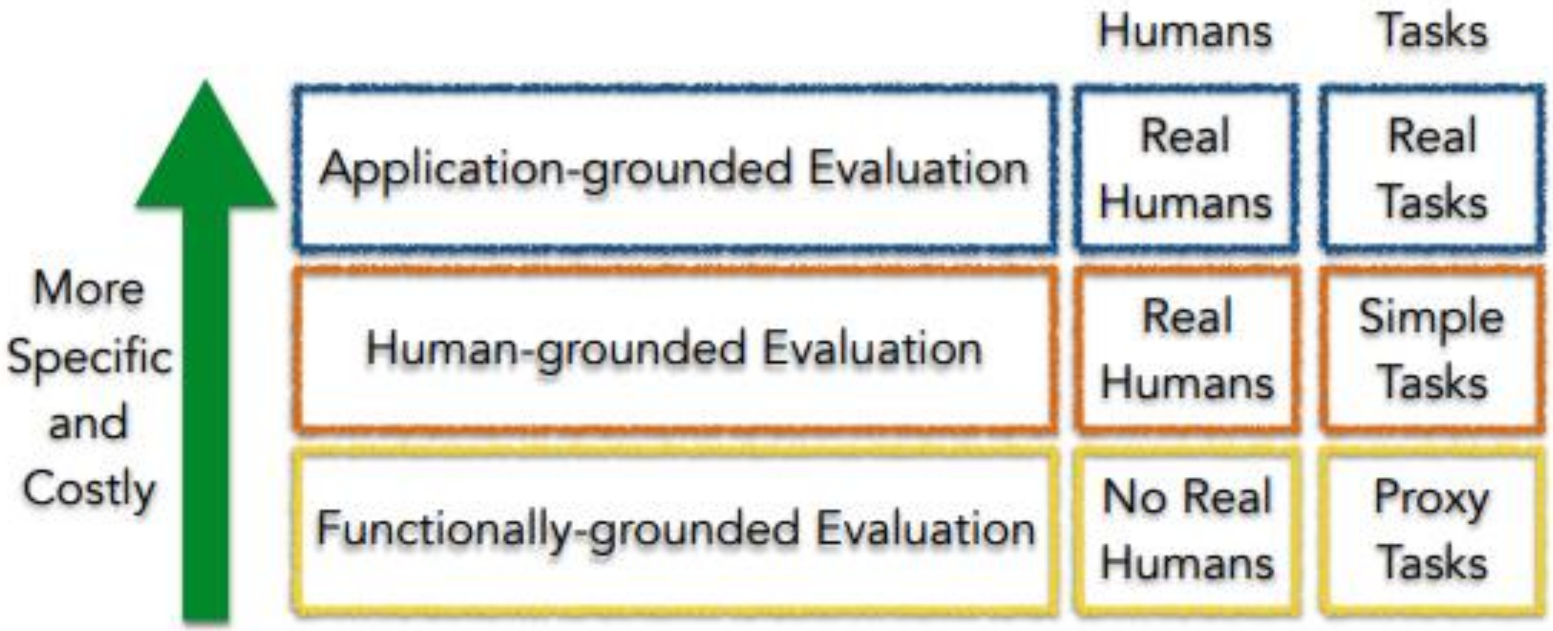


FIGURE 1. The expected improvements due to pursuing research at the intersection of creative AI, XAI, and human–AI interaction.

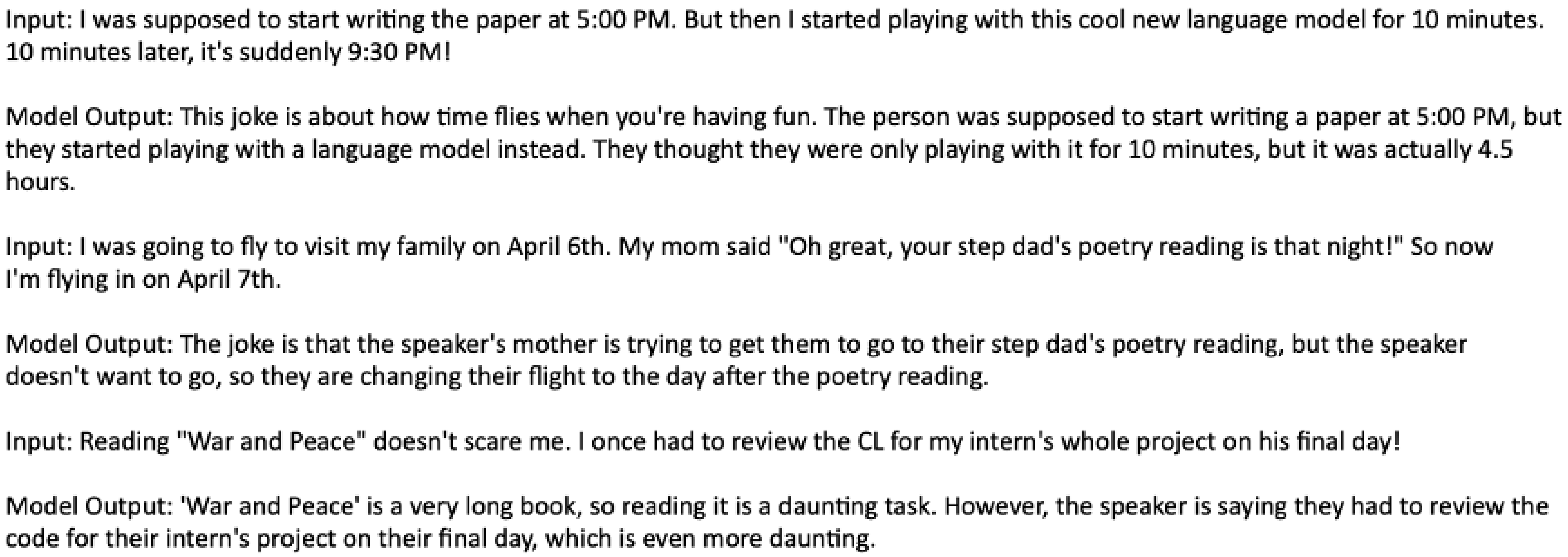


	Humans	Tasks
Application-grounded Evaluation	Real Humans	Real Tasks
Human-grounded Evaluation	Real Humans	Simple Tasks
Functionally-grounded Evaluation	No Real Humans	Proxy Tasks

Figure 1: Taxonomy of evaluation approaches for interpretability

[Doshi-Velez and Kim, 2017]

Interpreting the artifact itself



Input: I was supposed to start writing the paper at 5:00 PM. But then I started playing with this cool new language model for 10 minutes. 10 minutes later, it's suddenly 9:30 PM!

Model Output: This joke is about how time flies when you're having fun. The person was supposed to start writing a paper at 5:00 PM, but they started playing with a language model instead. They thought they were only playing with it for 10 minutes, but it was actually 4.5 hours.

Input: I was going to fly to visit my family on April 6th. My mom said "Oh great, your step dad's poetry reading is that night!" So now I'm flying in on April 7th.

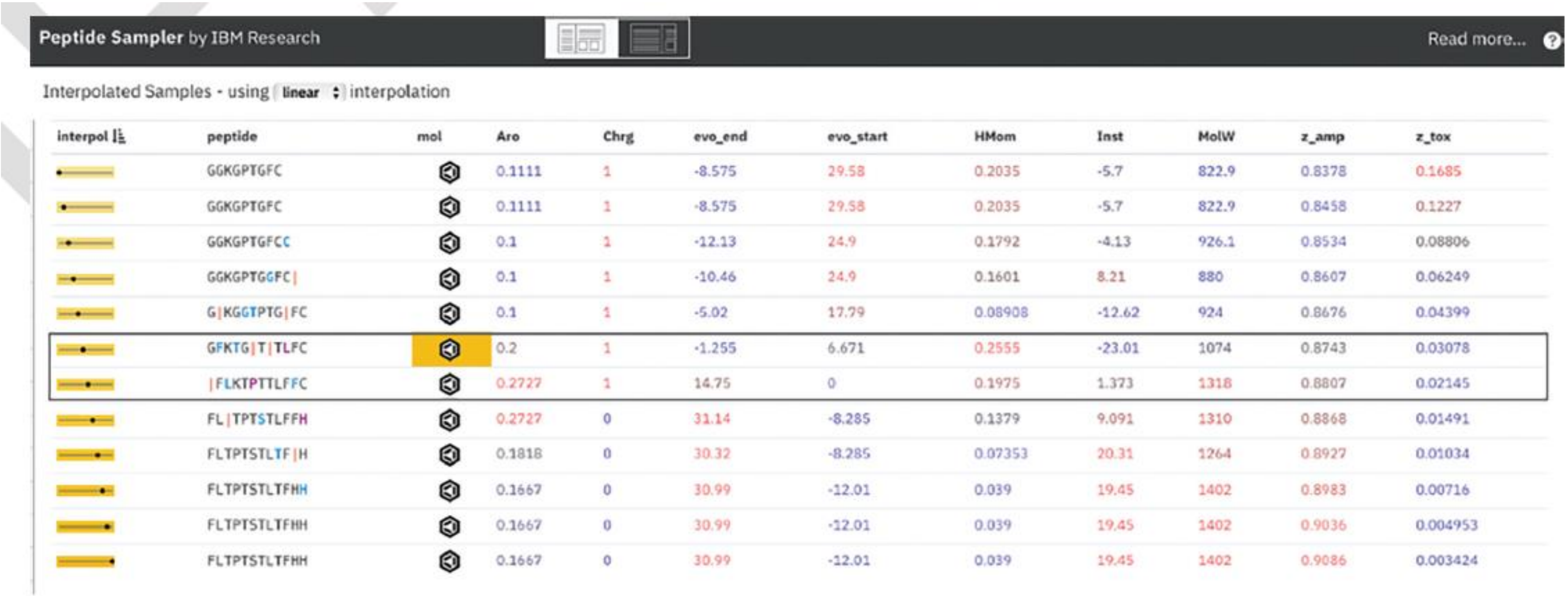
Model Output: The joke is that the speaker's mother is trying to get them to go to their step dad's poetry reading, but the speaker doesn't want to go, so they are changing their flight to the day after the poetry reading.

Input: Reading "War and Peace" doesn't scare me. I once had to review the CL for my intern's whole project on his final day!

Model Output: 'War and Peace' is a very long book, so reading it is a daunting task. However, the speaker is saying they had to review the code for their intern's project on their final day, which is even more daunting.

https://arxiv.org/pdf/2204.02311.pdf

Interpreting the VAE latent space



Interpolated Samples - using linear interpolation

interpol	peptide	mol	Aro	Chrg	evo_end	evo_start	HMom	Inst	MolW	z_amp	z_tox
	GGKGPTGFC		0.1111	1	-8.575	29.58	0.2035	-5.7	822.9	0.8378	0.1685
	GGKGPTGFC		0.1111	1	-8.575	29.58	0.2035	-5.7	822.9	0.8458	0.1227
	GGKGPTGFCC		0.1	1	-12.13	24.9	0.1792	-4.13	926.1	0.8534	0.08806
	GGKGPTGGFC|		0.1	1	-10.46	24.9	0.1601	8.21	880	0.8607	0.06249
	G|KGGTPTG|FC		0.1	1	-5.02	17.79	0.08908	-12.62	924	0.8676	0.04399
	GFKTG|T|TLFC		0.2	1	-1.255	6.671	0.2555	-23.01	1074	0.8743	0.03078
	|FLKTPTTLFFC		0.2727	1	14.75	0	0.1975	1.373	1318	0.8807	0.02145
	FL|TPTSTLFFH		0.2727	0	31.14	-8.285	0.1379	9.091	1310	0.8868	0.01491
	FLTPTSTLTF|H		0.1818	0	30.32	-8.285	0.07353	20.31	1264	0.8927	0.01034
	FLTPTSTLTFHH		0.1667	0	30.99	-12.01	0.039	19.45	1402	0.8983	0.00716
	FLTPTSTLTFHH		0.1667	0	30.99	-12.01	0.039	19.45	1402	0.9036	0.004953
	FLTPTSTLTFHH		0.1667	0	30.99	-12.01	0.039	19.45	1402	0.9086	0.003424

FIGURE 3. The Peptide Sampler visual analytic platform shows decoded sequences and their attributes during a linear interpolation between two selected sequences in the latent space of a generative autoencoder (a WAE in this example) trained on biological peptides.

GRAD-CAM: Visualizing saliency on the output

- https://keras.io/examples/vision/grad_cam/

- https://distill.pub/2017/feature-visualization/
- https://distill.pub/2018/building-blocks/

Testing with Concept Activation Vectors (TCAV)

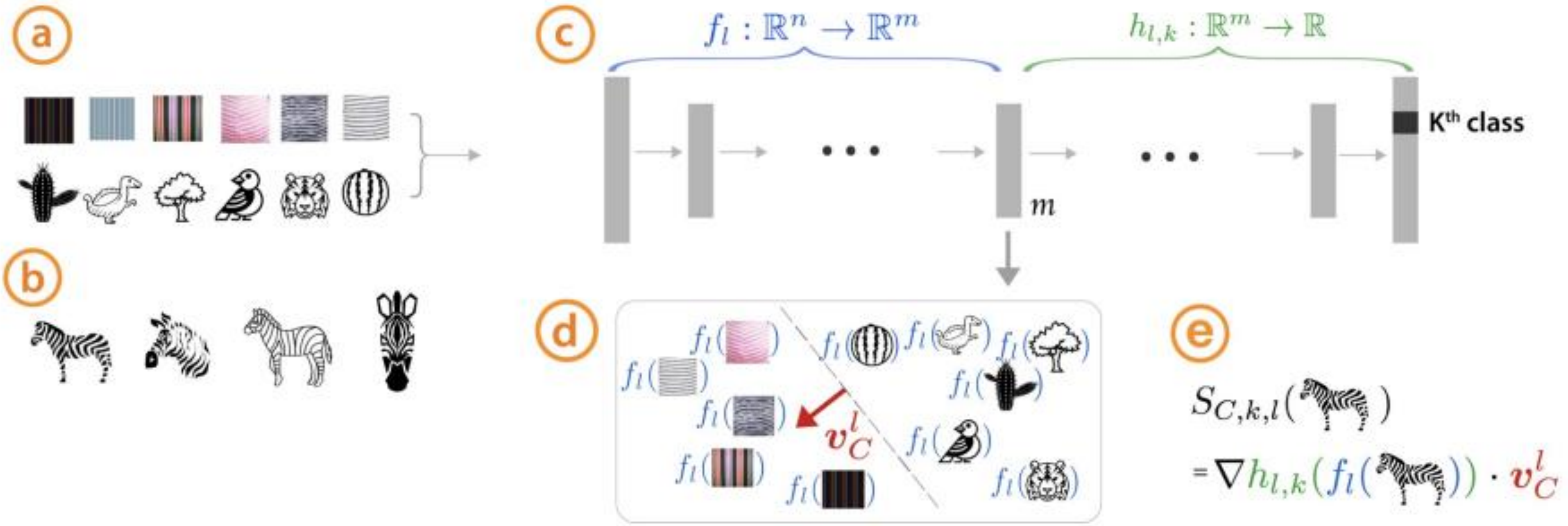

a
b
c
$f_l : \mathbb{R}^n \to \mathbb{R}^m$
$h_{l,k} : \mathbb{R}^m \to \mathbb{R}$
K^th class
m
d
v_C^l
e
$S_{C,k,l}($ $)$
$= \nabla h_{l,k}(f_l($ $)) \cdot v_C^l$

Figure 1. **Testing with Concept Activation Vectors:** Given a user-defined set of examples for a concept (e.g., 'striped'), and random examples ⓐ, labeled training-data examples for the studied class (zebras) ⓑ, and a trained network ⓒ, TCAV can quantify the model's sensitivity to the concept for that class. CAVs are learned by training a linear classifier to distinguish between the activations produced by a concept's examples and examples in any layer ⓓ. The CAV is the vector orthogonal to the classification boundary (v_C^l, red arrow). For the class of interest (zebras), TCAV uses the directional derivative $S_{C,k,l}(\boldsymbol{x})$ to quantify conceptual sensitivity ⓔ.

[Kim et al, 2018]

- https://distill.pub/2020/circuits/zoom-in/
- https://distill.pub/2020/circuits/equivariance/

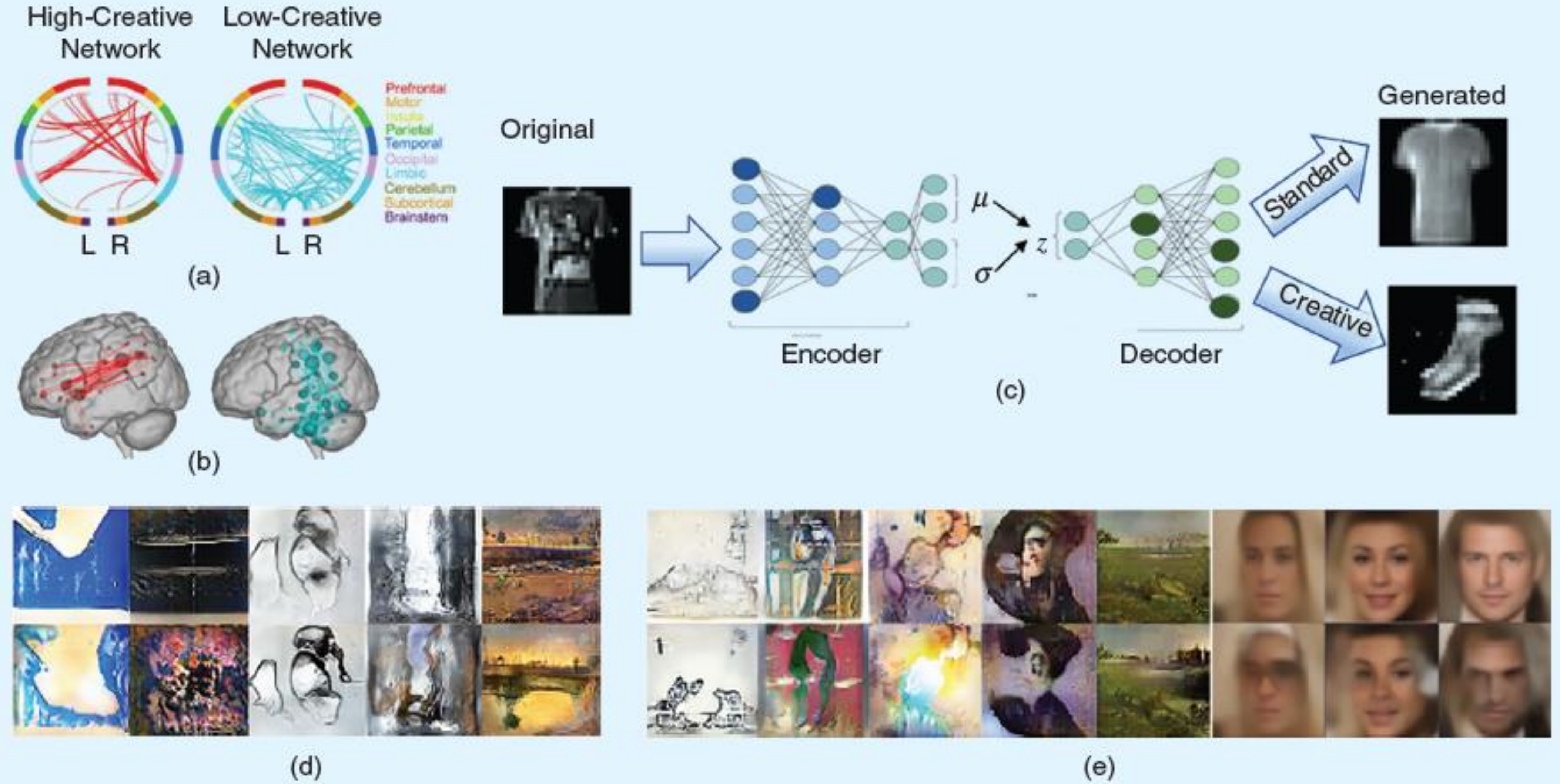
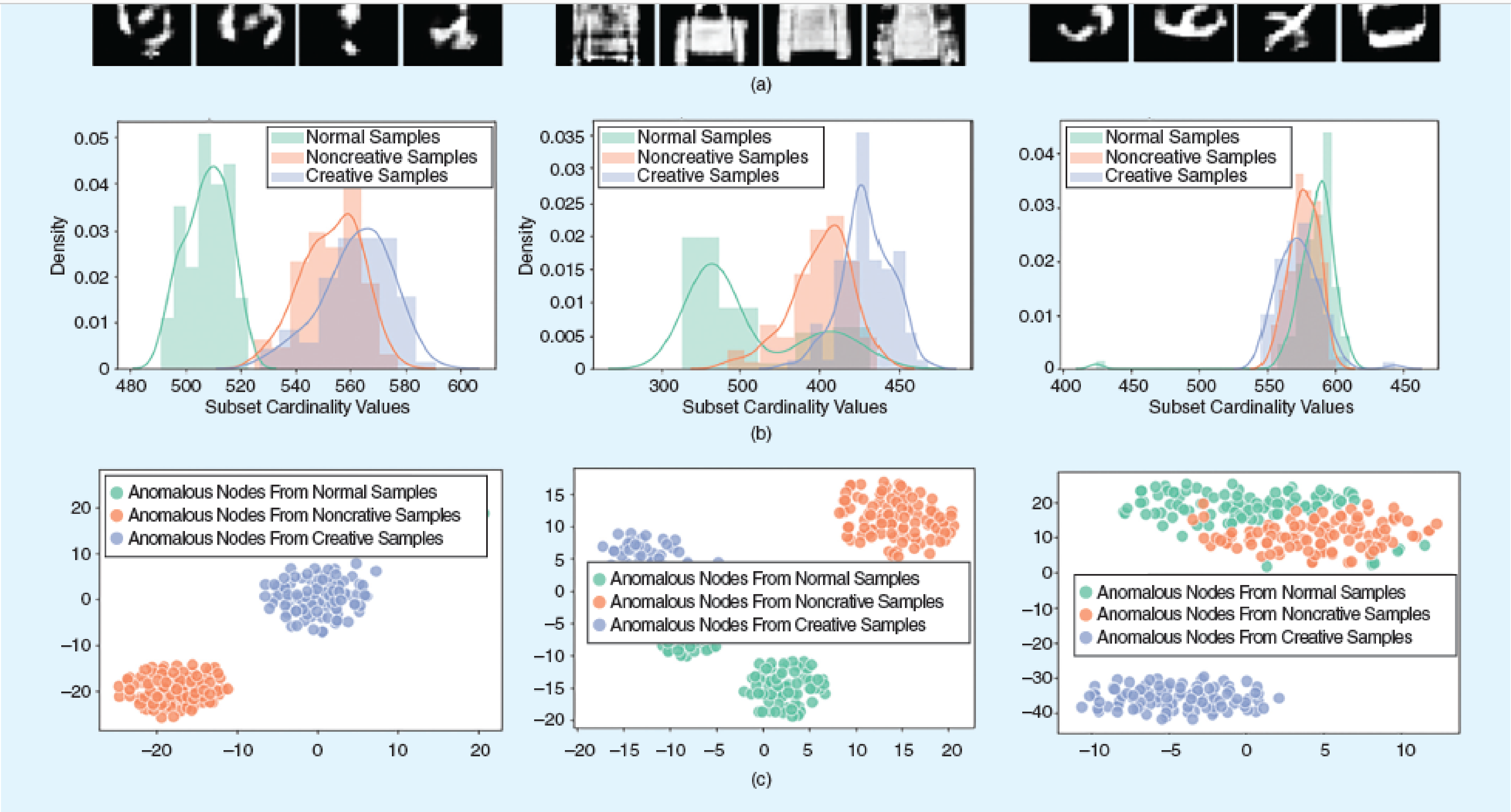


FIGURE 2. (a) and (b) Depictions [(a): circle plots; (b): glass brains] of high- and low-creative networks in human brains with their highest-degree nodes. The circle plot colors correspond to brain lobes. Adapted from [35]. (c) The depiction of a VAE model with our neuroinspired creative decoder. Normally, a small fraction of neurons in each hidden layer are low active (dark color). Inspired by the neural basis of creativity, we activate those "low-active" (task-negative) neurons to induce coupling between task-positive and task-negative neurons during "creative" decoding. (d) and (e) The samples generated by the proposed creative decoding method. The base generative models used were ARTGAN trained on (d) the WikiART data set and VAE trained on (e) the CelebA data set. Adapted from [34]. L: left hemisphere; R: right hemisphere.



(a)

(b)

(c)

FIGURE 5. (a) Examples of images that are human annotated as creative. The samples were generated by the "creative" decoding of a VAE model trained on a Modified National Institute of Standards and Technology (MNIST), Fashion MNIST (FMNIST), and combined (COMBO) (MNIST + FMIST = COMBO) data set. Adapted from [34]. (b) and (c) The activations characterization. (b) The subset cardinality distributions for anomalous subsets for different types of generated samples. (c) t-distributed stochastic neighbor embedding over anomalous subset nodes for the creative decoder activations under generation of normal, noncreative, and creative samples. Adapted from [40].