On the Symbolic Universes of Language: from the Economy of Words to the Semantic Embeddings, a study of the Units of Culture

Thesis Discussion
Master's Degree in Data Science

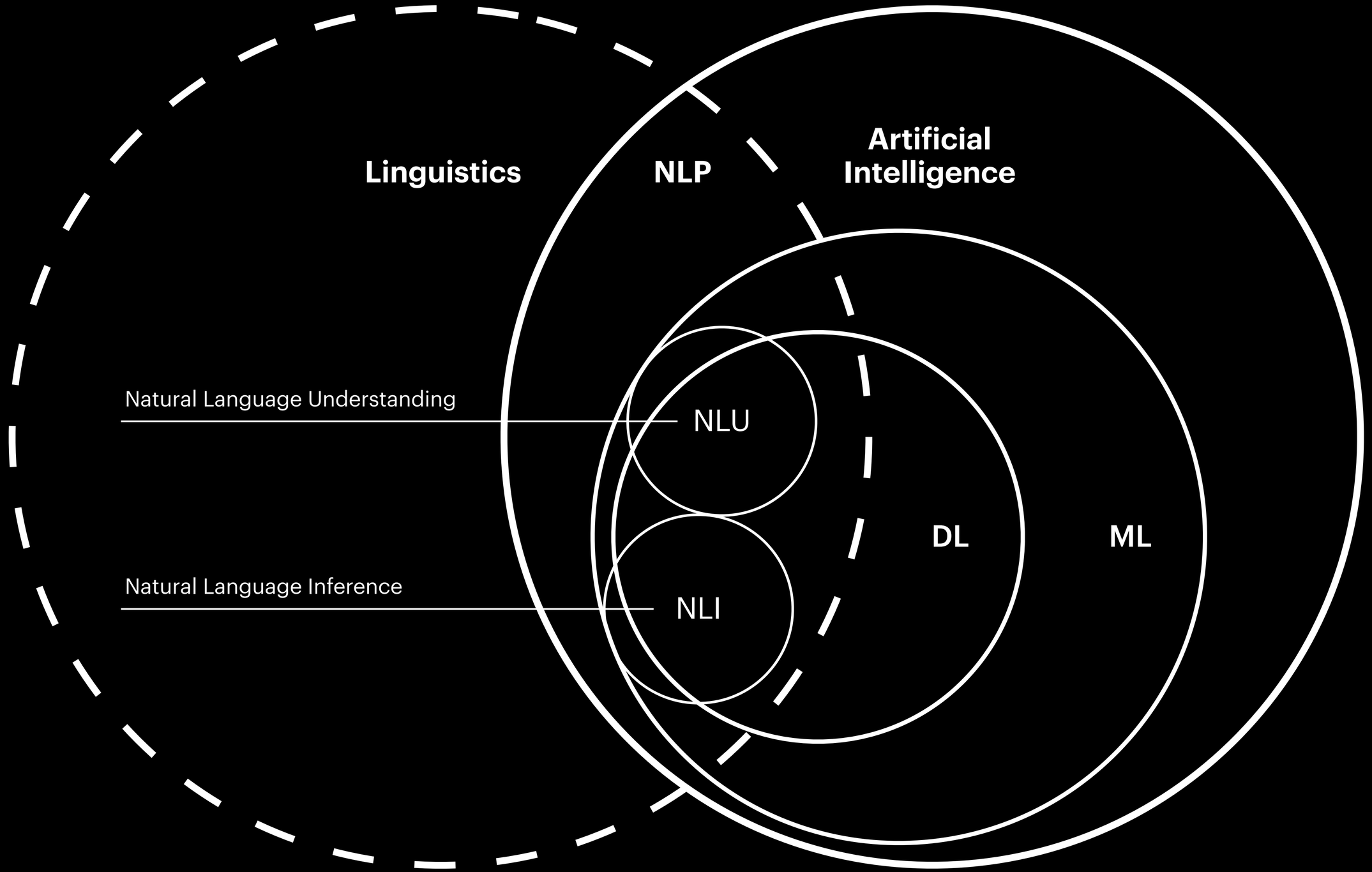
Andrea Leone

What is culture?

Passions
Traditions
Values
Religion
Beliefs
Personality
Customs
Sociability
Fashion
Lifestyle
Priming
Rituals
Language
Art
Habits
Manners
Literature
Leisure

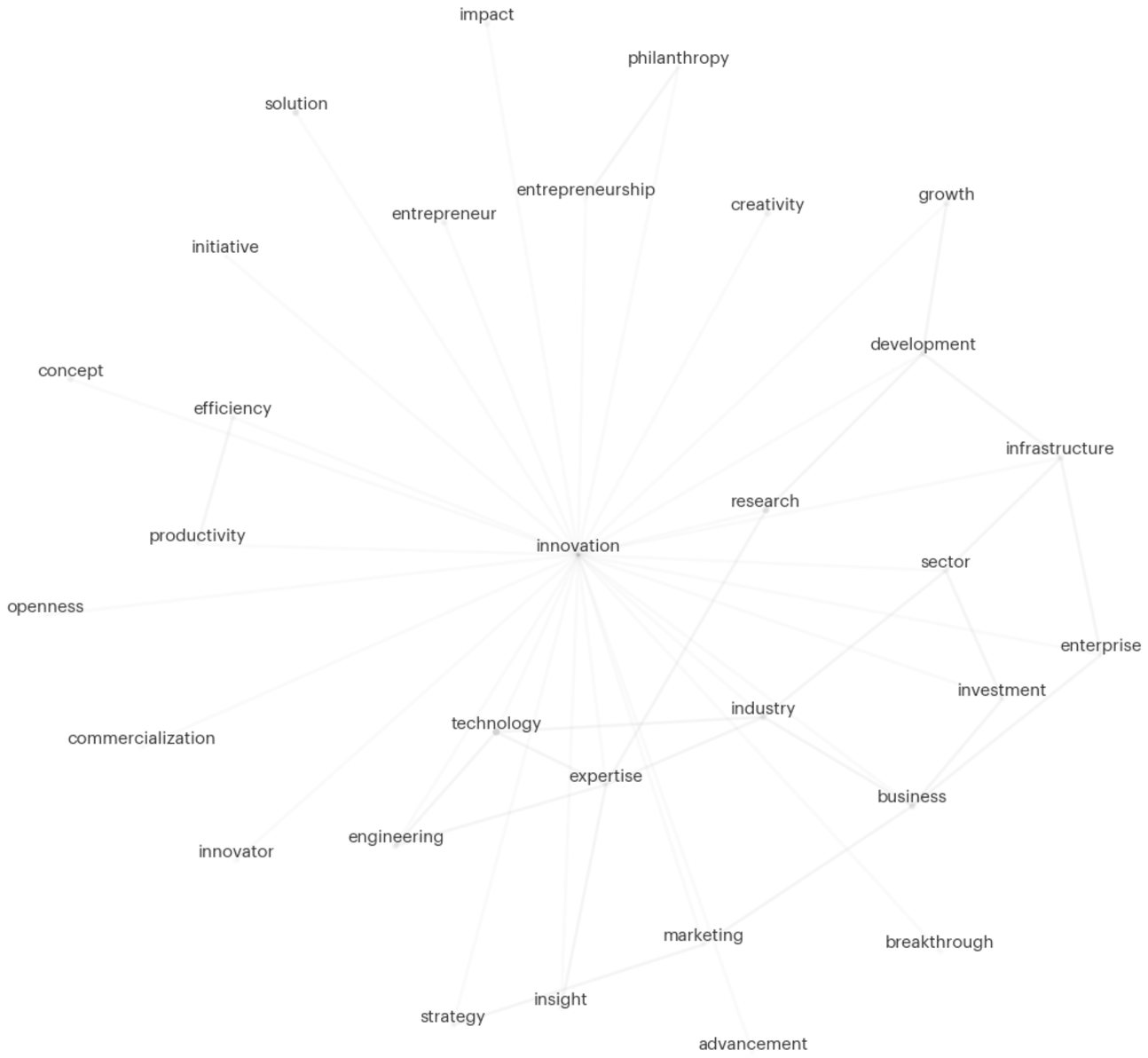
Culture denotes the conventions of a community
that guide the individuals into regular patterns of behaviour and
provide communal meanings and values we can use to describe it.

Passions
Customs
Traditions
Values
Religion
Beliefs
Personality
Sociability
Priming
Fashion
Leisure
Rituals
Lifestyle
Literature
Habits
Language
Manners
Art


Linguistics
NLP
Artificial Intelligence
Natural Language Understanding
NLU
DL
ML
Natural Language Inference
NLI


TED

Symbolic Universes

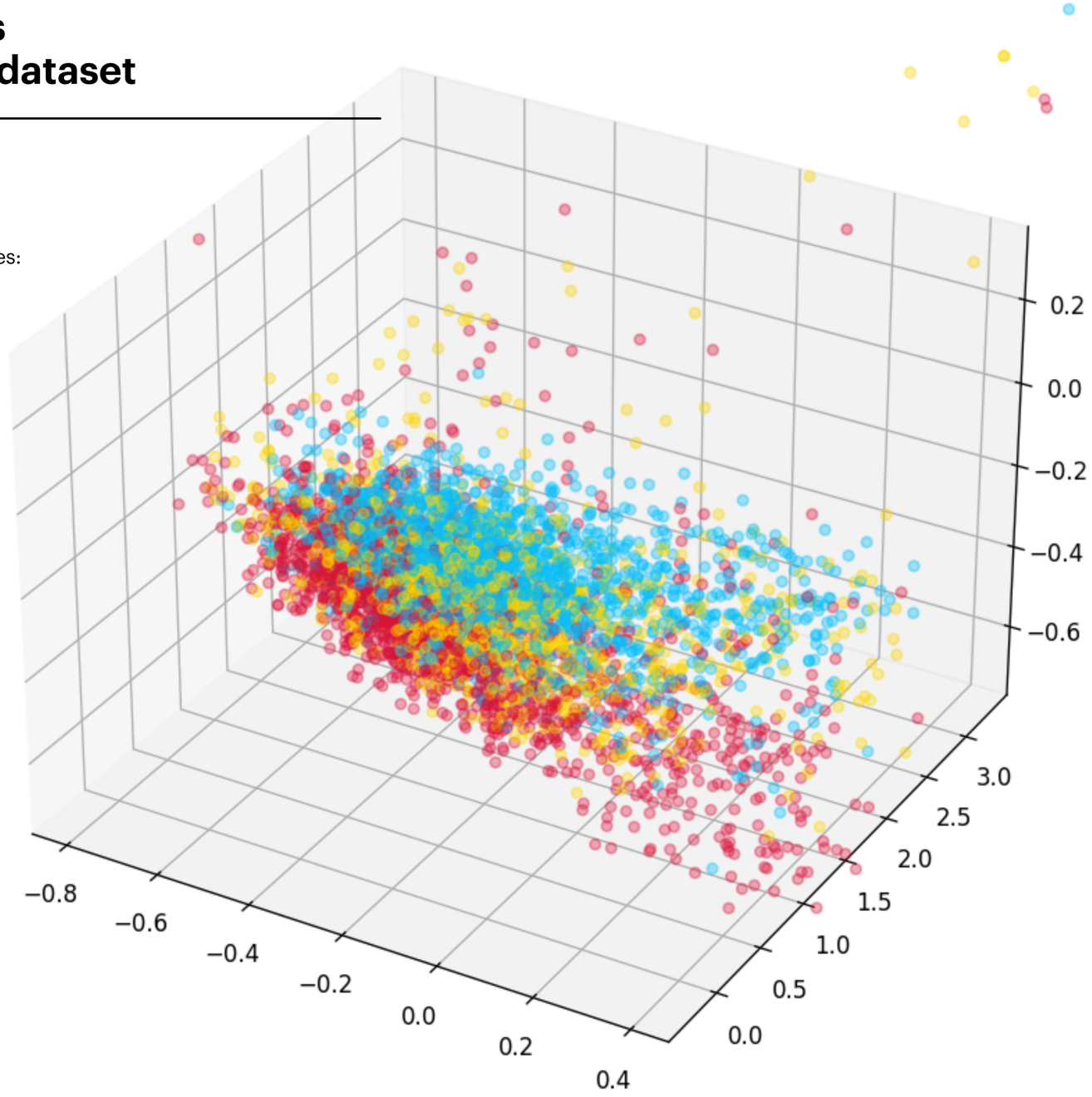

impact
philanthropy
solution
entrepreneurship
entrepreneur
creativity
growth
initiative
development
concept
efficiency
infrastructure
research
productivity
innovation
sector
openness
enterprise
investment
industry
technology
commercialization
expertise
business
engineering
innovator
marketing
breakthrough
insight
strategy
advancement

Semantic Embeddings

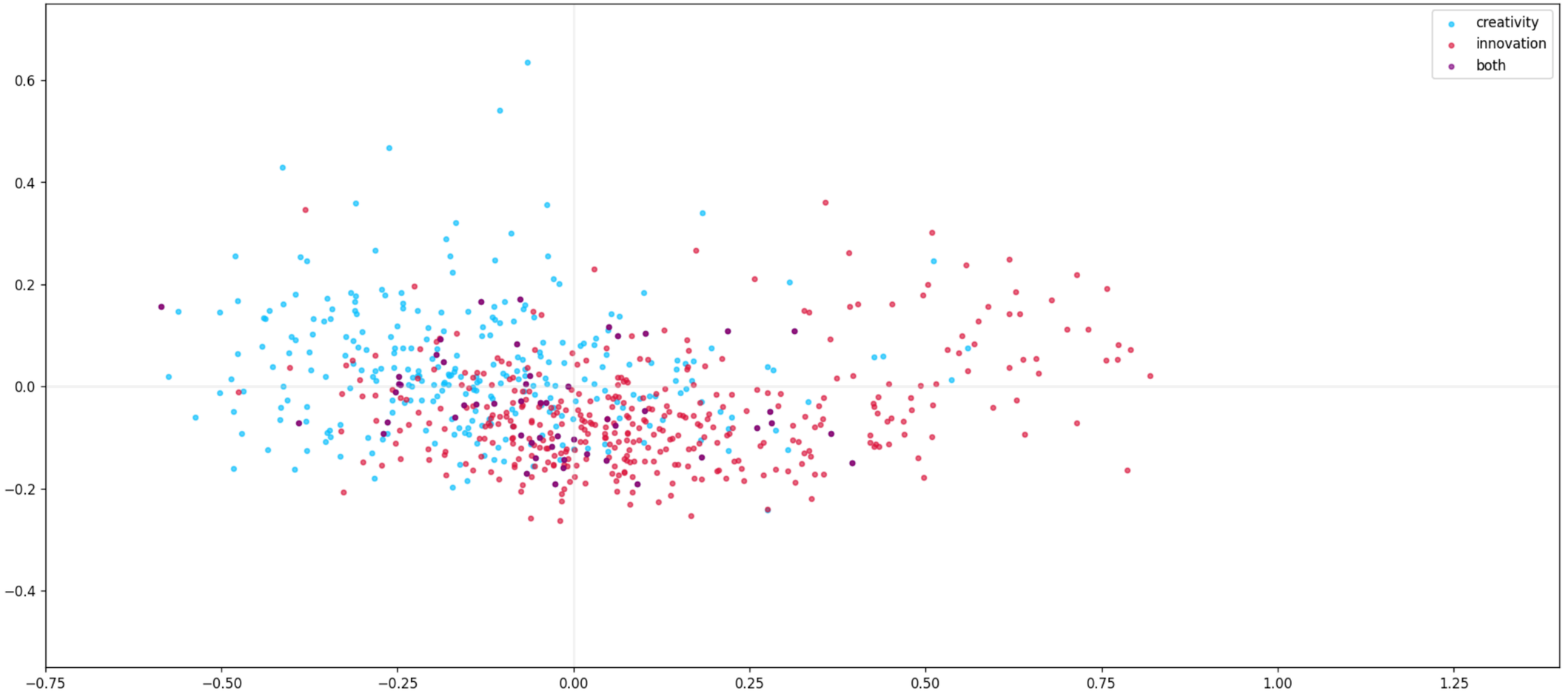
Semantic embeddings of the talks in the TED dataset

3D spatial arrangement of the static document embeddings obtained with the first three principal components (PCA).

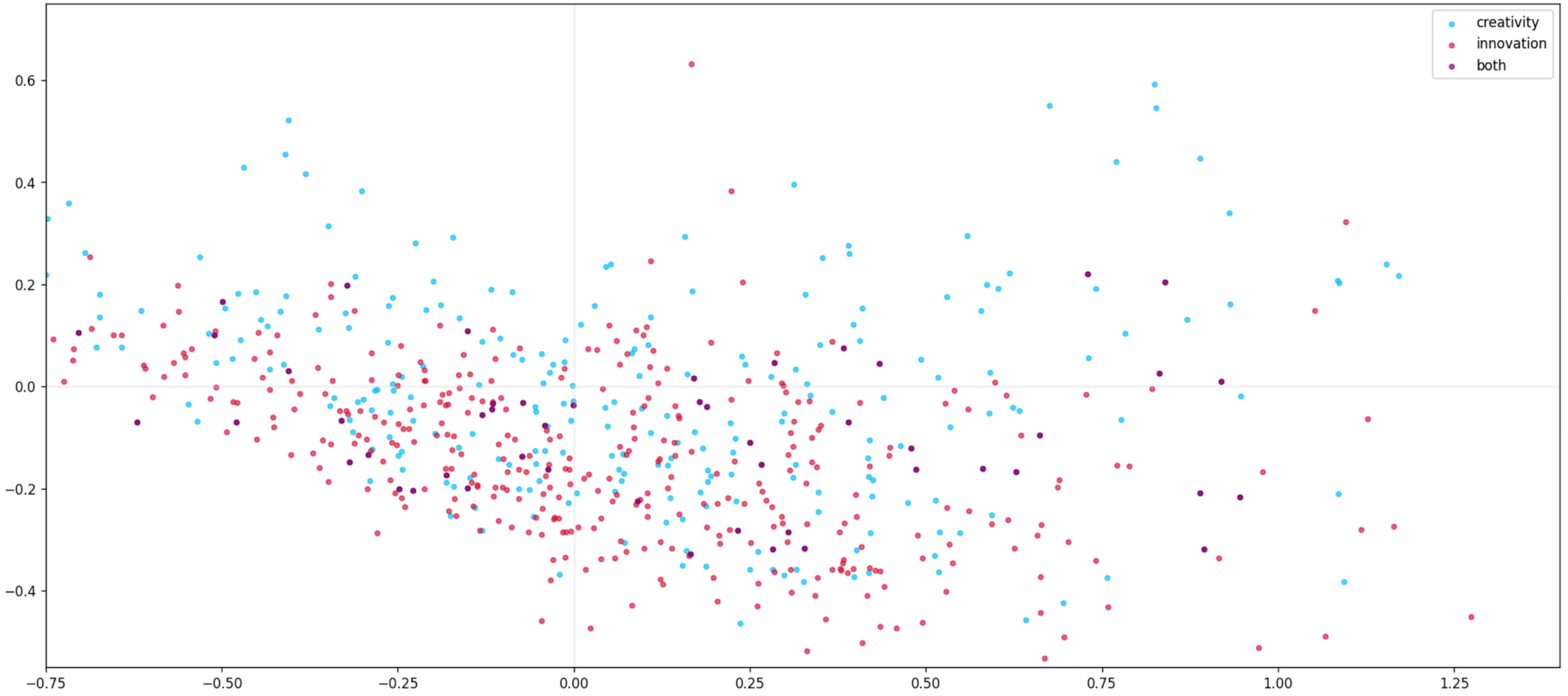
Colours represent the three macro-categories: science and innovation (blue, 34%), culture and society (red, 39.4%), economy and environment (yellow, 26.6%).



Semantic embeddings of the talks in the TED dataset obtained with static word vectors



Semantic embeddings of the talks in the TED dataset obtained with contextualised word vectors



baseline models

Nearest Neighbours	**Nearest Centroid Classifier**
	K-Neighbours Classifier
Linear Models	**Logistic Regression**
	Ridge Classifier
	Stochastic Gradient Descent Classifier
	Linear Discriminant Analysis
	Quadratic Discriminant Analysis
Support Vector Machines	**Linear Support Vector Classifier**
	C-Support Vector Classifier
	Nu-Support Vector Classifier
Ensemble Learners	**eXtreme Gradient Boosting Classifier**
	Decision Trees Classifier
	Random Forest Classifier

baseline models	accuracy scores / static vec.	/ neural vec.
Nearest Centroid Classifier	67.98%	43.16%
K-Neighbours Classifier	72.80%	53.87%
Logistic Regression	74.08%	69.25%
Ridge Classifier	75.63%	71.50%
Stochastic Gradient Descent Classifier	74.60%	68.28%
Linear Discriminant Analysis	75.35%	69.67%
Quadratic Discriminant Analysis	72.38%	55.27%
Linear Support Vector Classifier	75.39%	68.54%
C-Support Vector Classifier	72.58%	38.36%
Nu-Support Vector Classifier	73.52%	69.53%
eXtreme Gradient Boosting Classifier	75.65%	65.83%
Decision Trees Classifier	63.37%	45.83%
Random Forest Classifier	74.56%	62.76%

baseline models	accuracy scores / static vec.	/ neural vec.
eXtreme Gradient Boosting Classifier	75.65%	65.83%
Multi-Layer Perceptron	75.21%	65.42%
Feed-Forward Neural Network	74.36%	64.54%
Convolutional Neural Network	73.94%	64.18%

transformer models	accuracy scores	
eXtreme Gradient Boosting Classifier	75.65%	65.83%
BERT	93.22%	
RoBERTa	85.19%	
DistilBERT	94.78%	
SqueezeBERT	95.35%	

Zero-Shot Learning

Zero-Shot Learning

“To get a model to do something that it was not explicitly trained to do”

Zero-Shot Learning

"To get a model to do something that it was not explicitly trained to do"

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

Mike Lewis*, Yinhan Liu*, Naman Goyal*, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer
Facebook AI
mikelewis@fb.com,yinhan@ai2incubator.com,naman@fb.com

Abstract

We present BART, a denoising autoencoder for pretraining sequence-to-sequence models. BART is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and other recent pre-training schemes. We evaluate a number of noising approaches, finding the best performance by both randomly shuffling the order of sentences and using a novel in-filling scheme, where spans of text are replaced with a sin-

masked tokens (Joshi et al., 2019), the order in which masked tokens are predicted (Yang et al., 2019), and the available context for replacing masked tokens (Dong et al., 2019). However, these methods typically focus on particular types of end tasks (e.g. span prediction, generation, etc.), limiting their applicability.

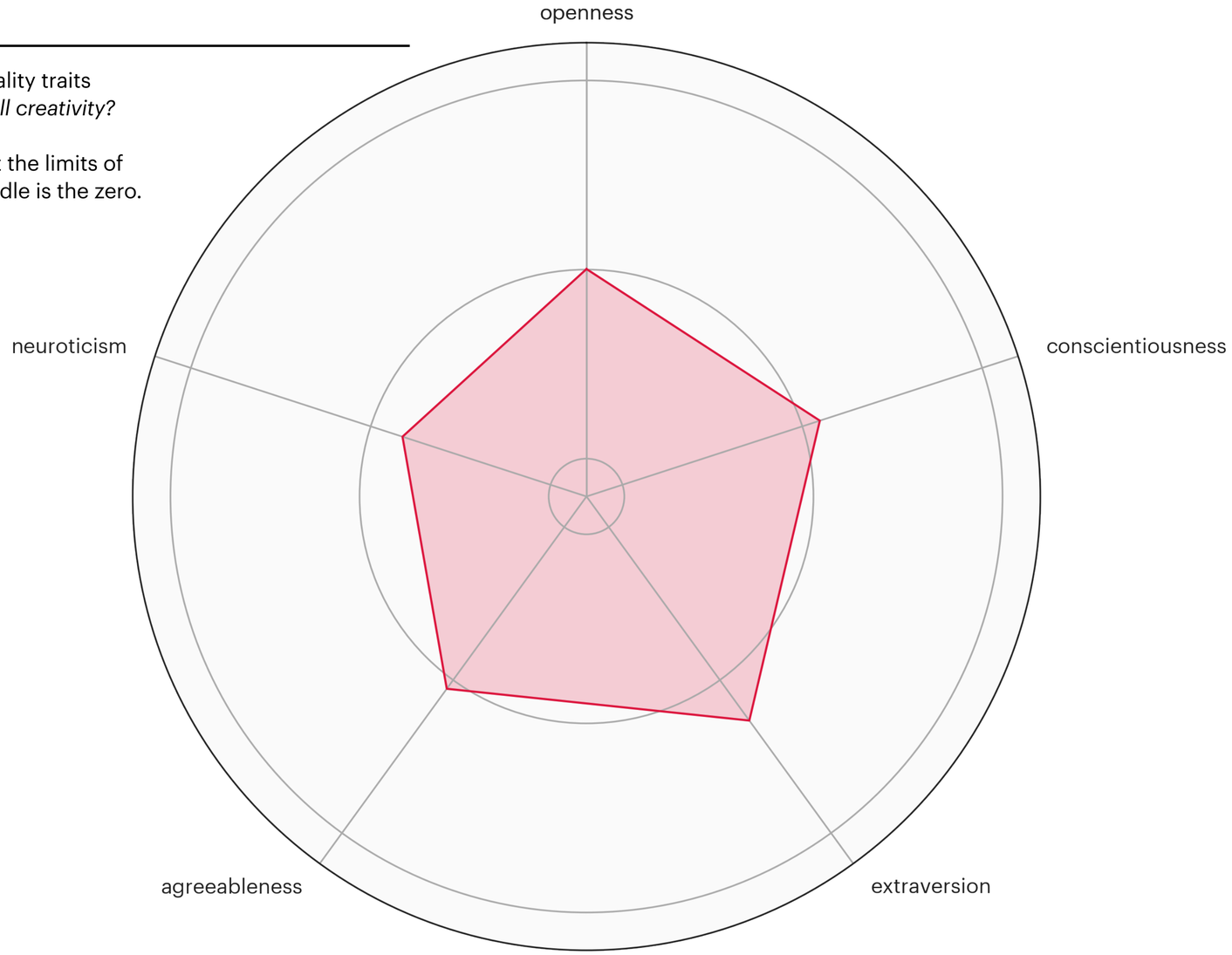
In this paper, we present BART, which pre-trains a model combining Bidirectional and Auto-Regressive Transformers. BART is a denoising autoencoder built with a sequence-to-sequence model that is applicable to a very wide range of end tasks. Pretraining has two stages (1) text is corrupted with an arbitrary noising function, and (2) a sequence-to-sequence model is learned to reconstruct the original text. BART uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as

personality traits	polarised indicators / positive	/ negative
Openness	inventive, curious	consistent, cautious
Conscientiousness	efficient, organised	extravagant, careless
Extraversion	outgoing, energetic	solitary, reserved
Agreeableness	friendly, compassionate	critical, rational
Neuroticism	sensitive, nervous	resilient, confident

Estimation of the personality traits

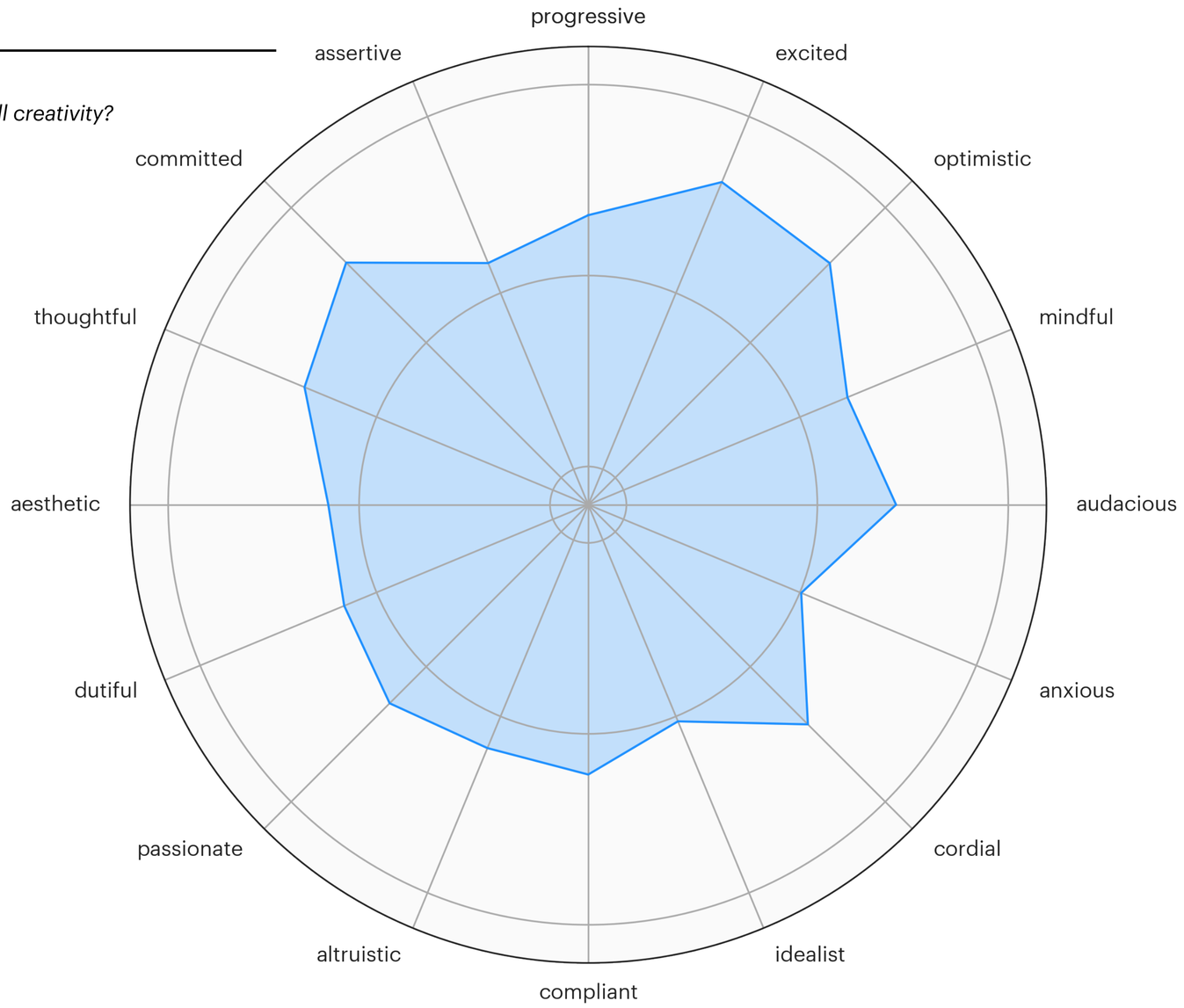
Radial plot showcasing the Big Five personality traits of Sir Ken Robinson's TED talk *Do schools kill creativity?*

The smaller and the larger circles represent the limits of the spectrum (-1 and +1); the one in the middle is the zero.



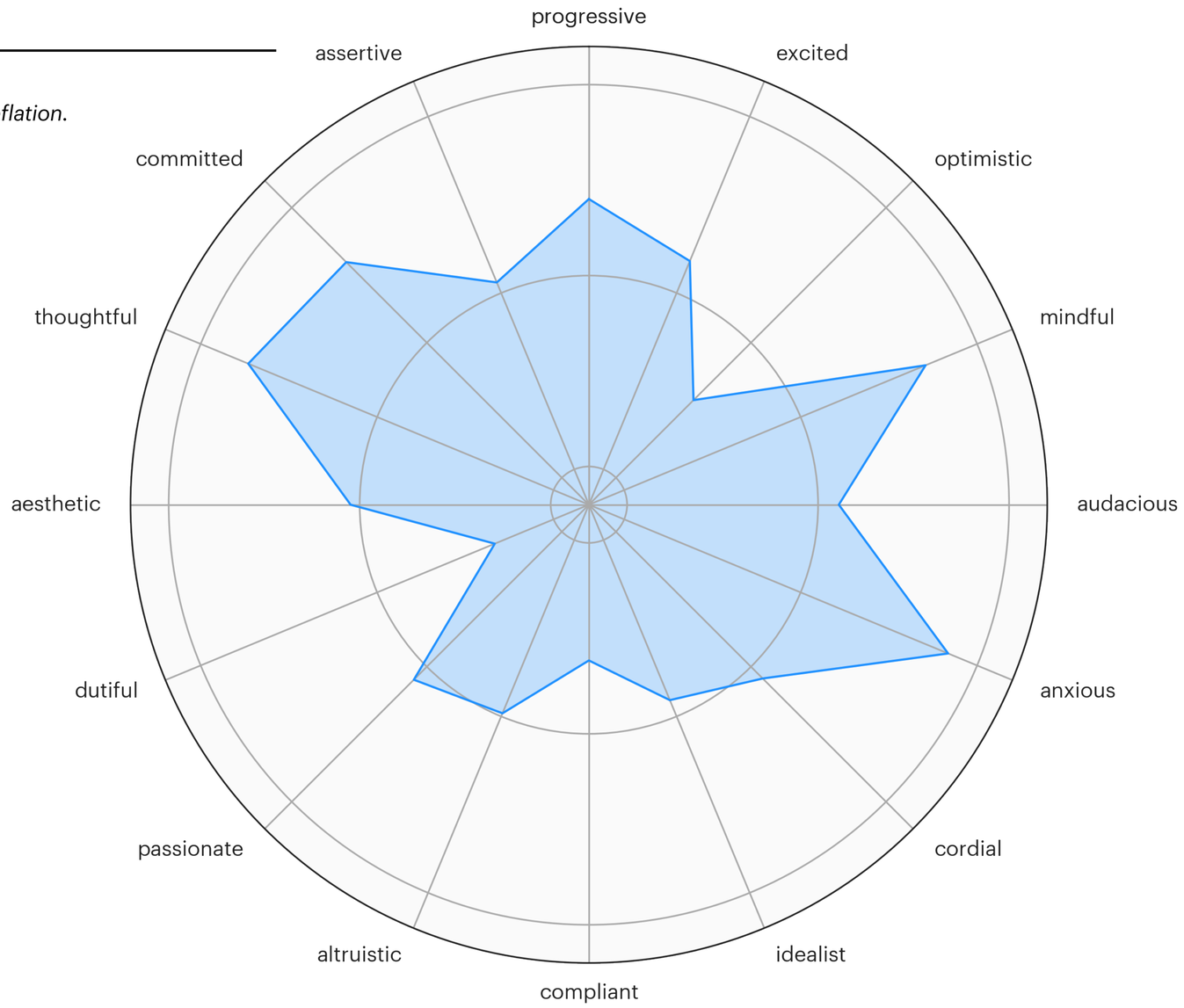
Estimation of the units of culture

Radial plot showcasing the cultural insights
of Sir Ken Robinson's TED talk *Do schools kill creativity?*



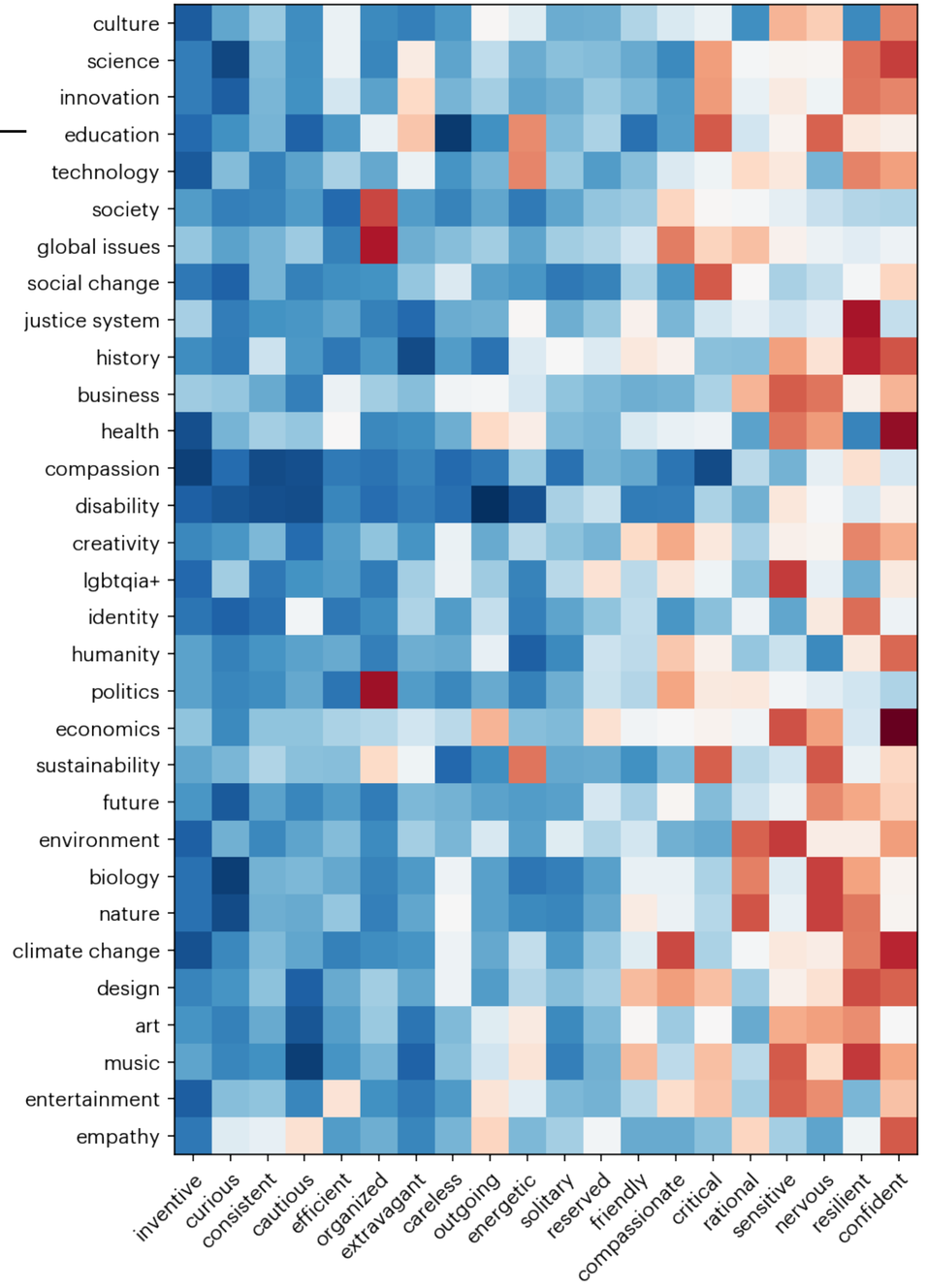
Estimation of the units of culture

Radial plot showcasing the cultural insights
of Cathie Wood's report on *Inventories & Deflation*.



Exploration of the personality traits

Diverging heatmap of the personality indicators associated to a selection of topics across TED talks. Shades of blue indicate higher scores, reds the lower.



On the Symbolic Universes of Language: from the Economy of Words to the Semantic Embeddings, a study of the Units of Culture