Quantitative Text Analysis

Meeting 11

Petro Tolochko

Organisational

- Deadline final paper: April 30
- Final paper: application of one or several automated text analysis methods on a topic related to the Master's thesis or a topic of free choice
- Contents: Methods + Results + Discussion section of a (pseudo)-academic paper (6-8 pages)
- Commented code
- Email to: petro.tolochko@univie.ac.at

Organisational 2

- HW1 Feedback — today
- HW2 Grades/Feedback — today/tomorrow
- HW3 — uploaded today/tomorrow (deadline: two weeks from upload)

Organisational 2

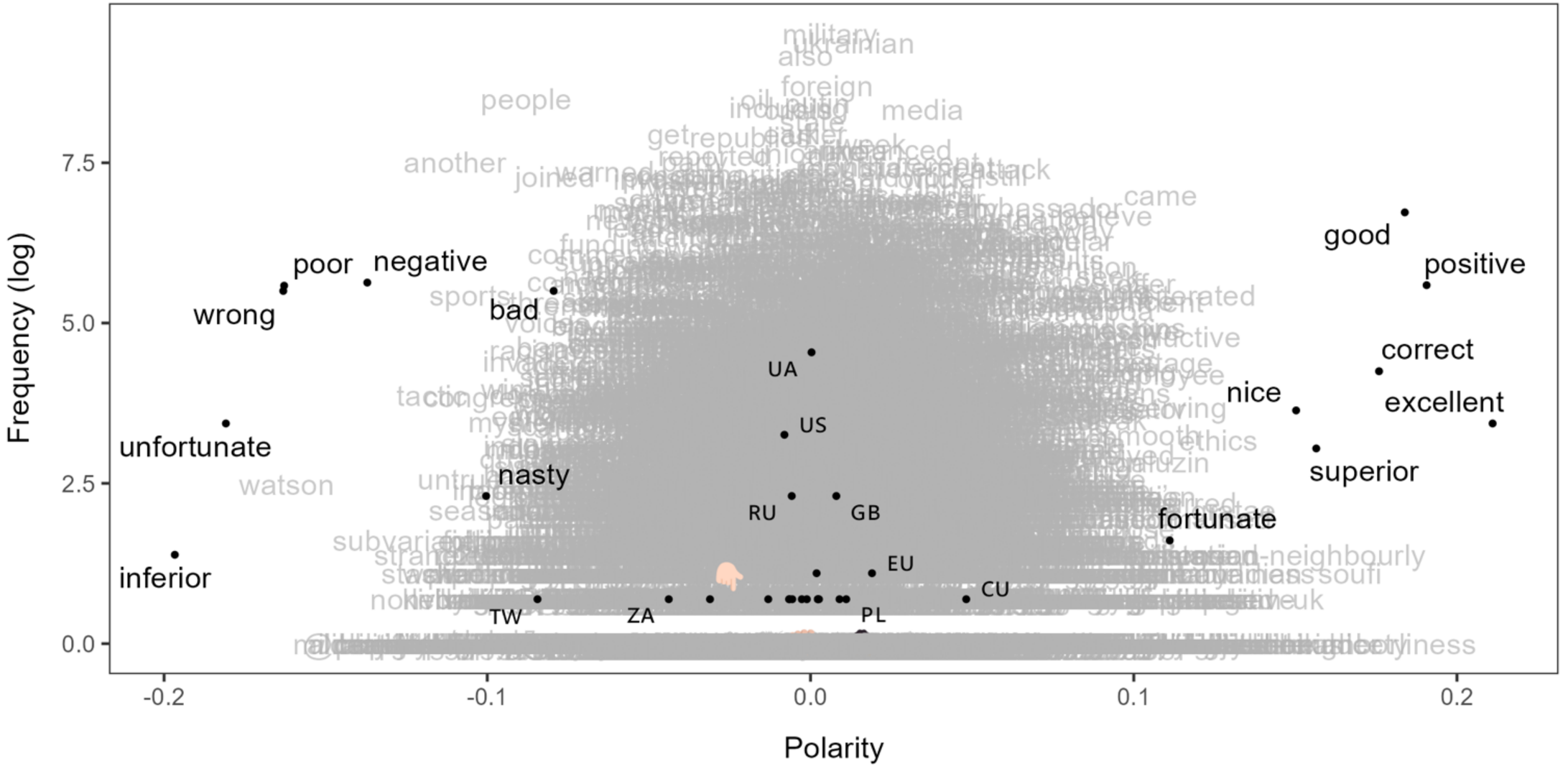
- HW1 Feedback — today
- HW2 Grades/Feedback — today/tomorrow
- HW3 — uploaded today/tomorrow (deadline: two weeks from upload)

SORRY

Questions?

Latent Semantic Scaling

- choose “seed words” that represent your dimensions
- calculate the polarity of words based on their proximity to seed words
    - e.g., cosine similarity
- predict polarity scores of documents by weighting word polarity scores by their frequency in the documents


Frequency (log)
7.5
5.0
2.5
0.0
-0.2
-0.1
0.0
0.1
0.2
Polarity
poor
negative
wrong
bad
unfortunate
nasty
inferior
good
positive
correct
nice
excellent
superior
fortunate
UA
US
RU
GB
EU
CU
TW
ZA
PL
people
another
military
ukrainian
also
foreign
media
watson

AMERICAN SOCIOLOGICAL ASSOCIATION

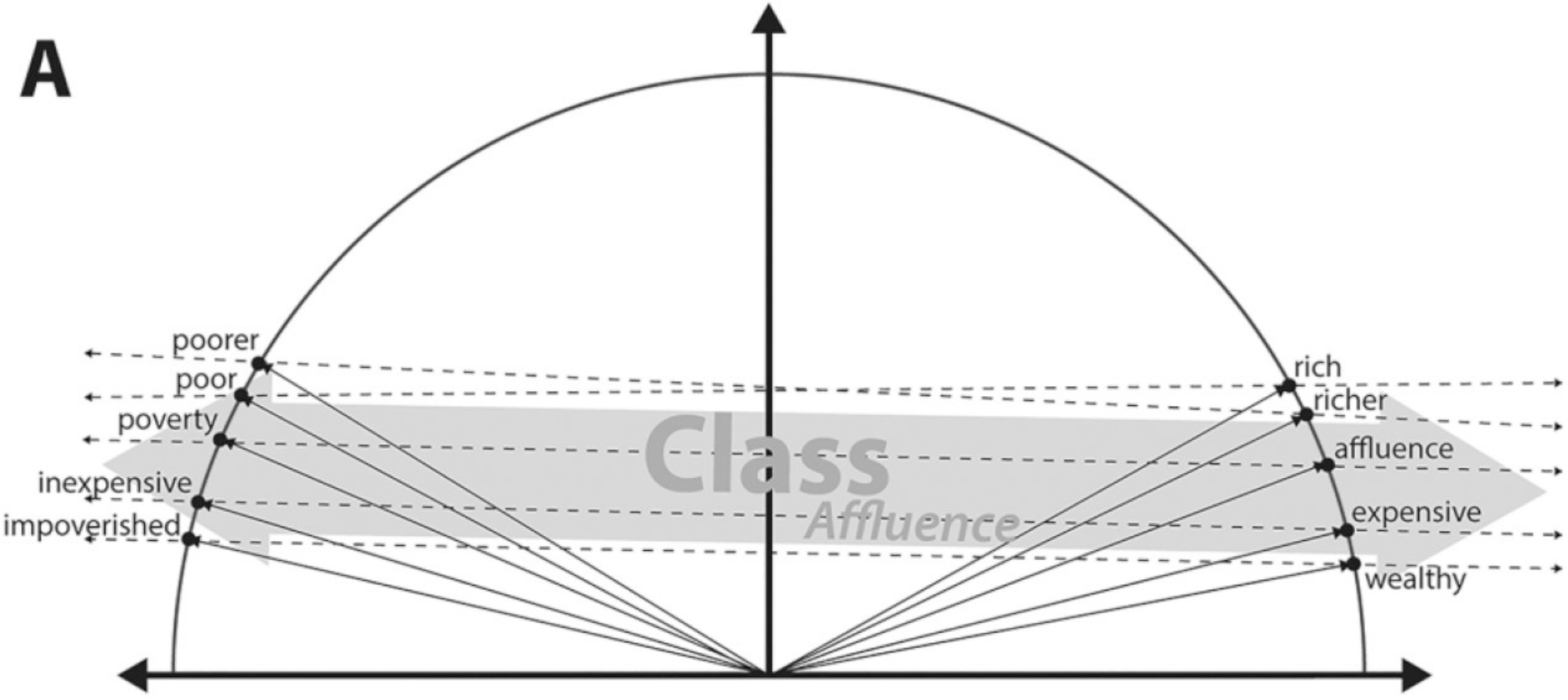
American Sociological Review
2019, Vol. 84(5) 905–949

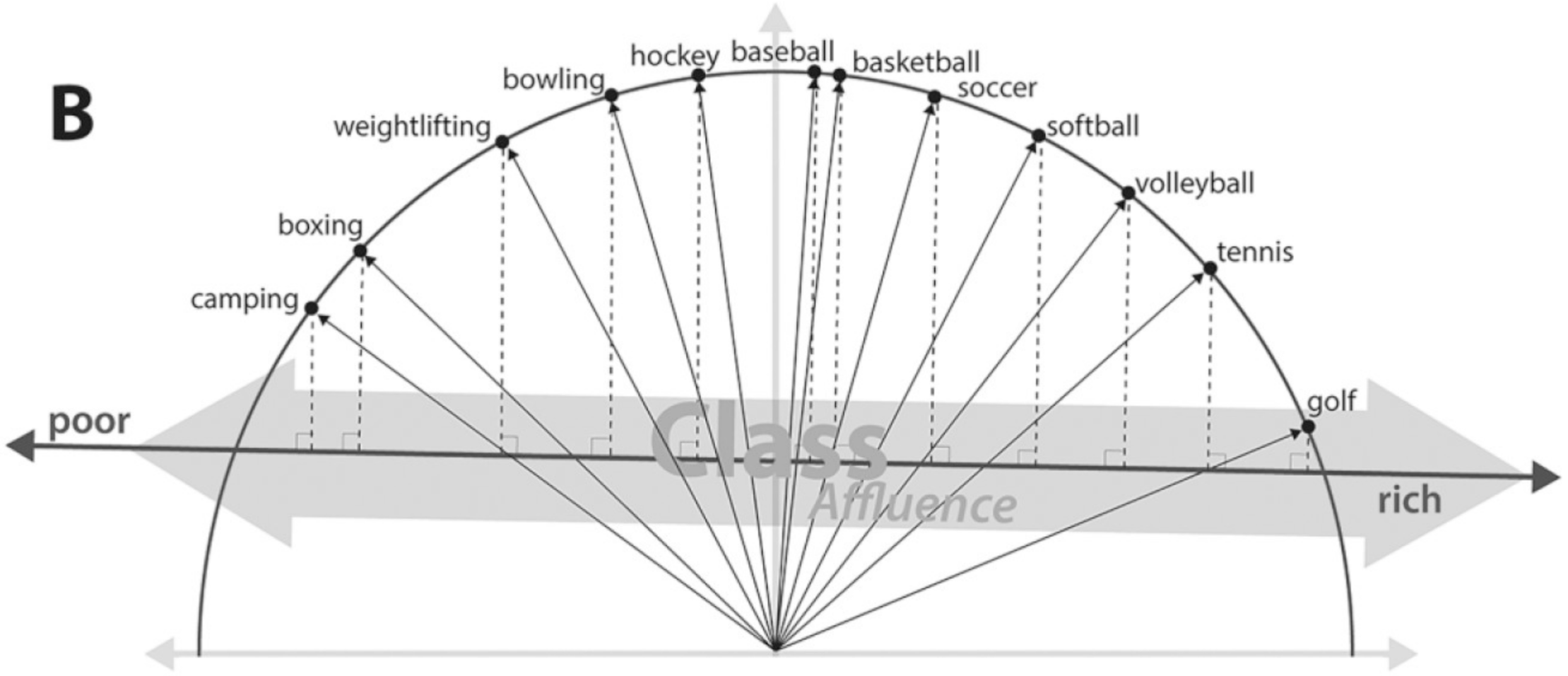
DOI: 10.1177/0003122419877135
journals.sagepub.com/home/asr

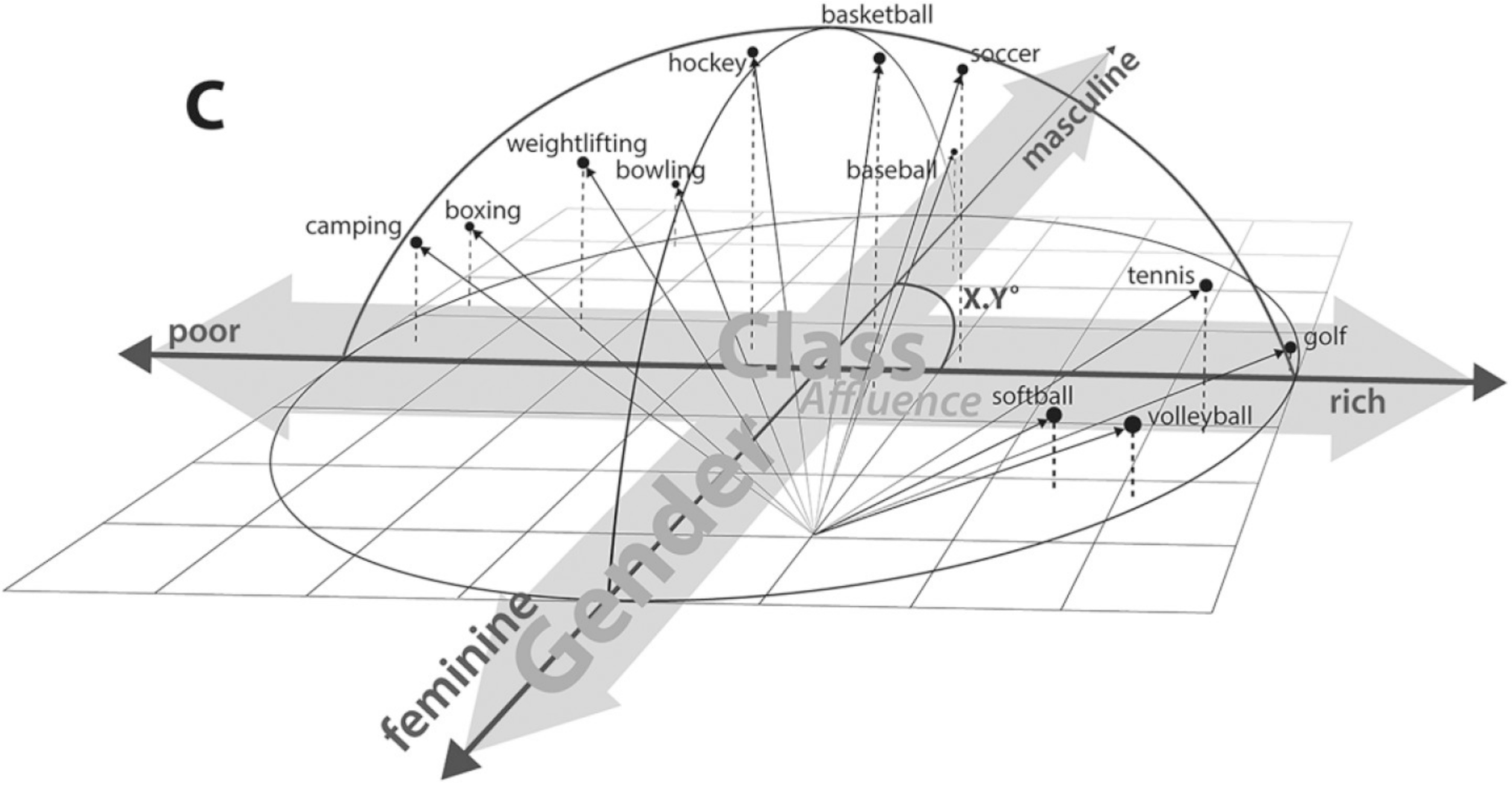
SAGE

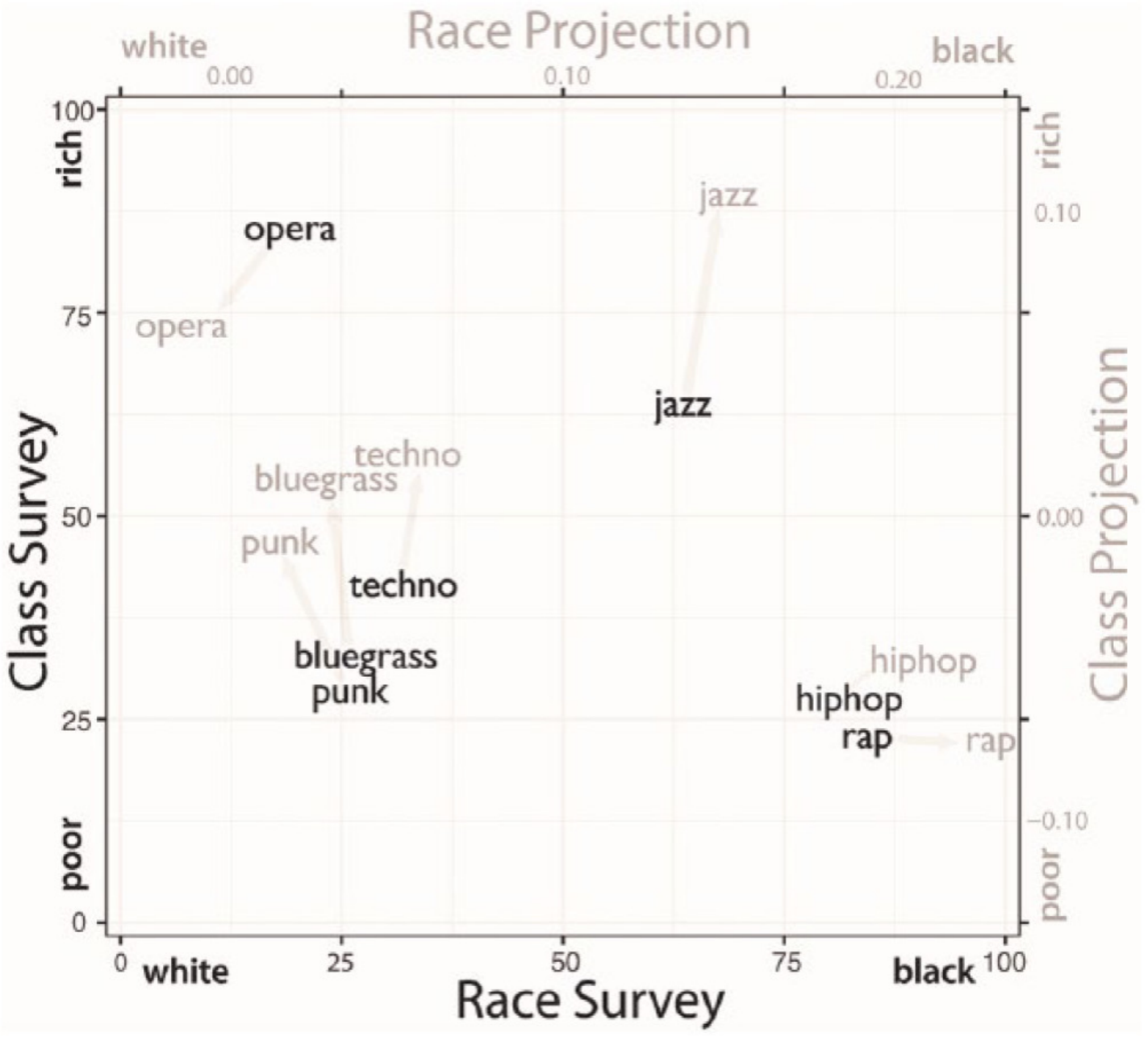
The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings

Austin C. Kozlowski,[a] Matt Taddy,[b] and James A. Evans[a,c]


A
poorer
poor
poverty
inexpensive
impoverished
rich
richer
affluence
expensive
wealthy
Class
Affluence


B
hockey
baseball
basketball
bowling
soccer
weightlifting
softball
volleyball
boxing
tennis
camping
golf
poor
Class
Affluence
rich


C
basketball
hockey
soccer
masculine
weightlifting
bowling
baseball
camping
boxing
tennis
X.Y°
poor
Class
Affluence
golf
softball
volleyball
rich
Gender
feminine


Race Projection
white
black
0.00
0.10
0.20
rich
poor
Class Survey
100
75
50
25
0
opera
opera
jazz
jazz
techno
bluegrass
punk
techno
bluegrass
punk
hiphop
hiphop
rap
rap
0
white
25
50
75
black
100
Race Survey
rich
poor
Class Projection
0.10
0.00
−0.10

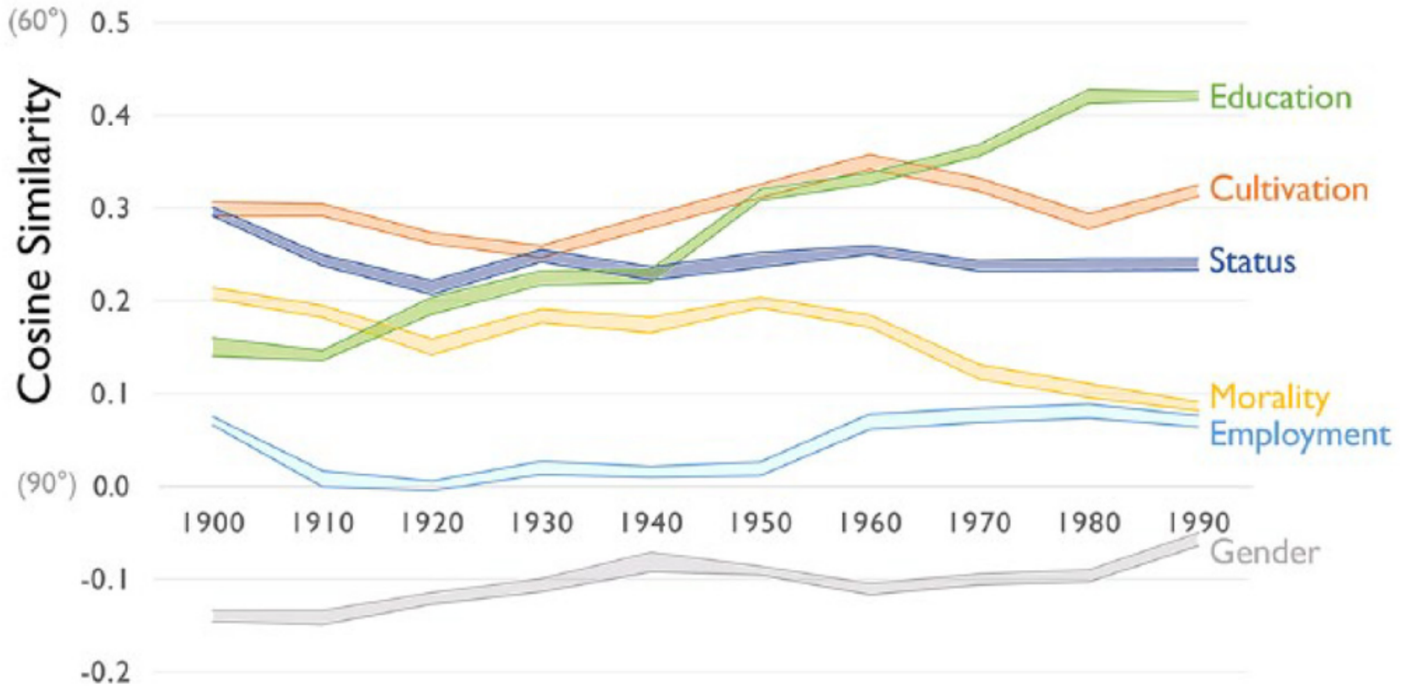


Figure 5. Cosine Similarity between the Affluence Dimension and Six Other Cultural Dimensions of Class by Decade; 1900 to 1999 Google Ngrams Corpus

Note: Bands represent 90 percent bootstrapped confidence intervals produced by subsampling.

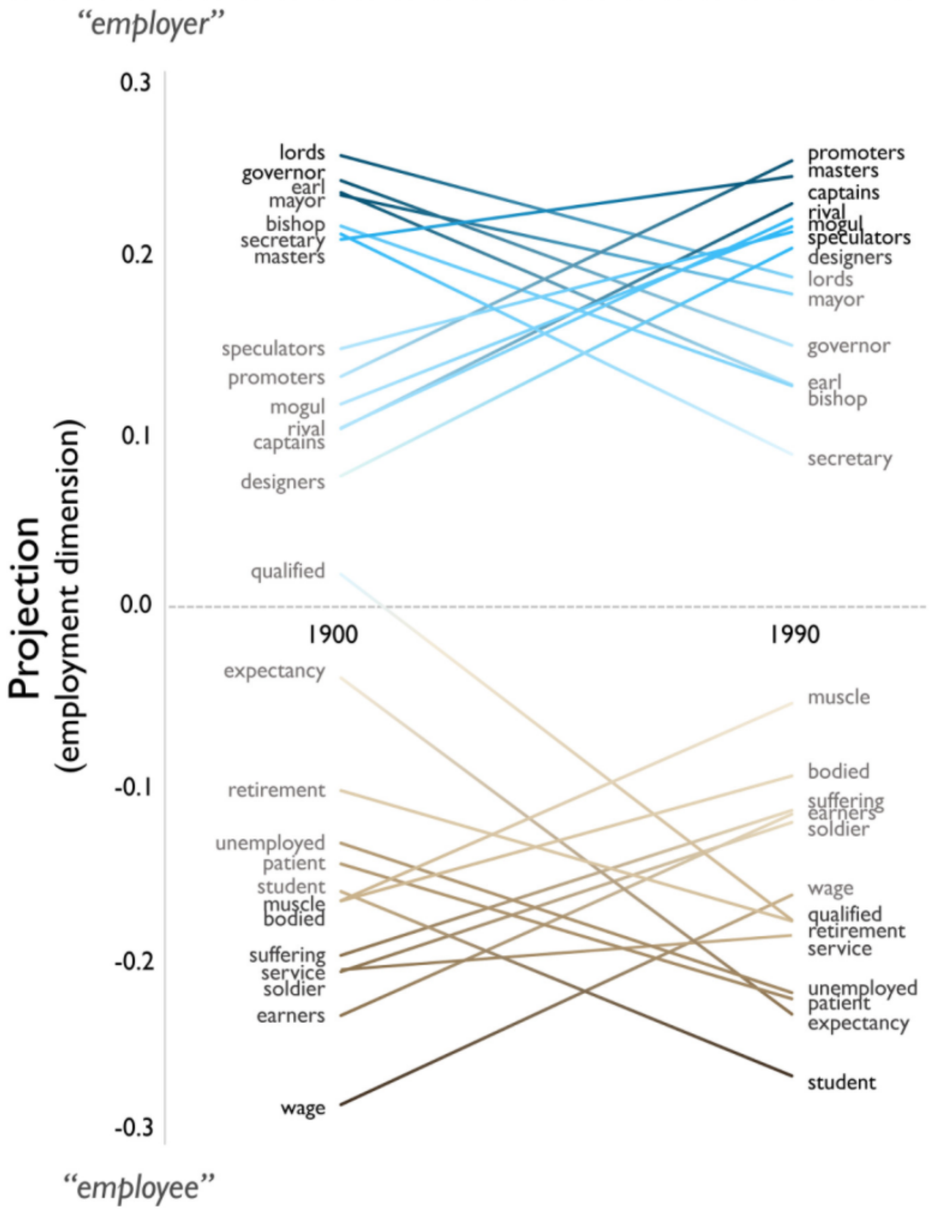


Figure 10. Words That Project High and Low on the Employment Dimension of Word Embedding Models Trained on Texts Published at the Beginning and End of the Twentieth Century; 1900–1919 and 1980–1999 Google Ngrams Corpus

Transformer Models

Language Model

- probability distribution over a sequence of words
- e.g., Markov Chains language model

Large Language Model

- probability distribution over a sequence of words
- Trained on the large chunk of the internet
- ~100TB of text data
- (Rumours) GPT-4 cost ~$200m to train
- Next-word prediction task

Large Language Model

- probability distribution over a sequence of words
- Trained on the large chunk of the internet
- ~100TB of text data
- (Rumours) GPT-4 cost ~$200m to train
- ***Next-word prediction task***

Transformer Models

- A lot more complicated than what we've seen…

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Self-Attention module

- Handles long-term dependencies
- Pays “attention” to which parts of the input sequence are important
- The self-attention mechanism allows each position in the input sequence to attend to all other positions, weighing the importance of different positions during the computation
- capture long-range dependencies and relationships between words or tokens in the sequence effectively

Encoding for self-attention

- Query
- Key
- Value

Encoding for self-attention

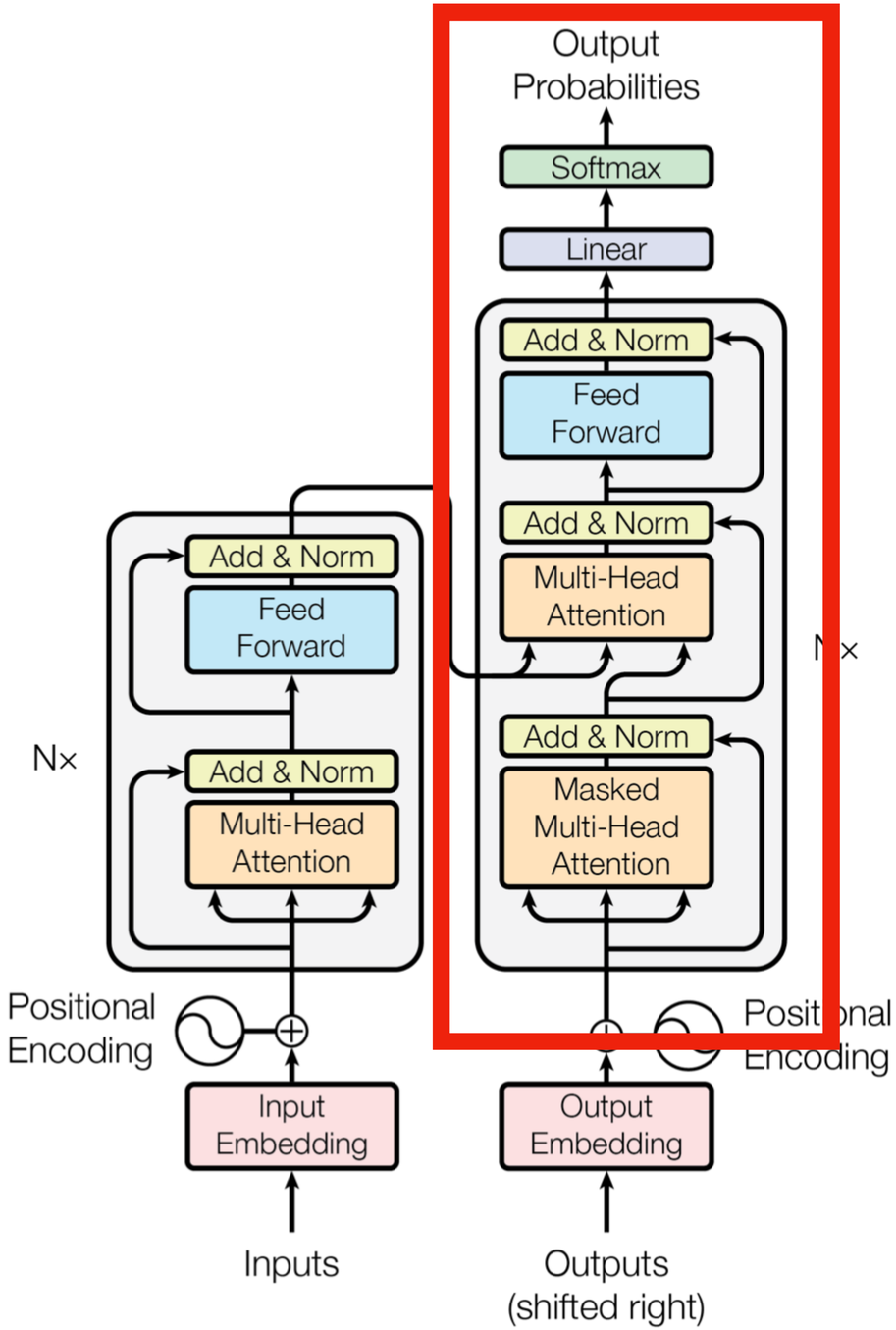
- For each token in the sequence, the model calculates query, key, and value vectors by applying learned linear transformations to the token embeddings
- The model computes the attention weights by measuring the similarity between the query vector of a token and the key vectors of all tokens in the sequence
- The attention weights obtained in the previous step are used to weight the value vectors of each token
- The resulting weighted sum is the transformed representation of the original token

	I	Love	Ice-Cream
I	.76	.89	.54
Love	.89	.46	.88
Ice-Cream	.54	.88	.54

	I	Love	Ice-Cream
I	.76	.89	.54
Love	.89	.46	.88
Ice-Cream	.54	.88	.54

Value

Encoder



Decoder

Figure 1: The Transformer - model architecture.

Open vs. Closed source

- Ethics
- Performance vs. Reproducibility
- Use cases
    - Do you ***need*** it?

Open Source

- Huggingface:
    - https://huggingface.co/
    - Many pre-trained models and datasets
- spaCy:
    - https://spacy.io/
    - Ecosystem for NLP

Open Source

- Better in Python
- But:
    - spaCyR
    - https://github.com/chainsawriot/grafzahl
        - Huggingface wrapper