Robust evaluation of language–brain encoding experiments

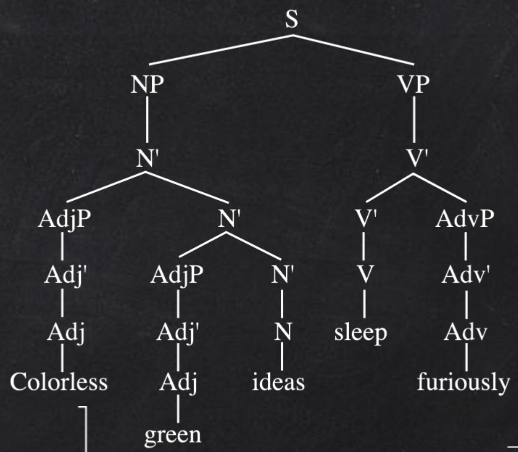
Lisa Beinborn, Samira Abnar, Rochelle Choenni
Institute for Logic, Language and Computation
Universiteit van Amsterdam

CLAP

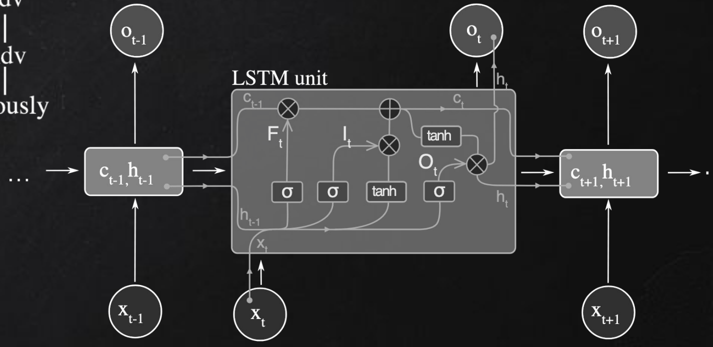
Cognitively inspired
Language Processing

HOW TO MODEL LANGUAGE?

Wiktionary
The free dictionary

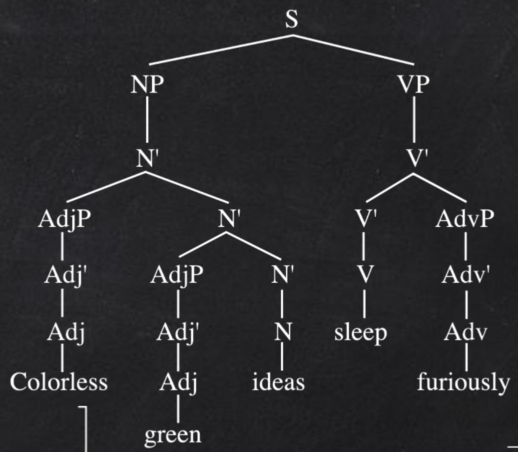


$$\begin{bmatrix} CAT & v \\ VFORM & finite \\ TNS & pres \\ SUBJ & \begin{bmatrix} CAT & np \\ PER & 1 \\ NUM & sg \end{bmatrix} \\ OBJ1 & \begin{bmatrix} CAT & np \end{bmatrix} \end{bmatrix} \rightarrow \textit{writes}$$

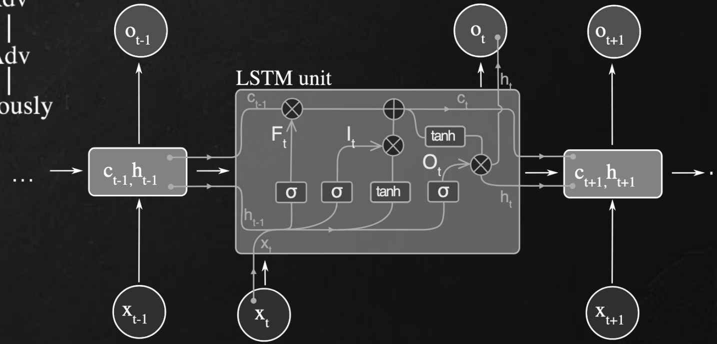


HOW TO MODEL LANGUAGE?

Wiktionary
The free dictionary



$$\begin{bmatrix} \mathit{CAT} & v \\ \mathit{VFORM} & \mathit{finite} \\ \mathit{TNS} & \mathit{pres} \\ \mathit{SUBJ} & \begin{bmatrix} \mathit{CAT} & \mathit{np} \\ \mathit{PER} & 1 \\ \mathit{NUM} & \mathit{sg} \end{bmatrix} \\ \mathit{OBJ1} & \begin{bmatrix} \mathit{CAT} & \mathit{np} \end{bmatrix} \end{bmatrix} \rightarrow \mathit{writes}$$



[4 1 5 3 1 6 8 0 3 ...]

Which one?

Hierarchical

Bidirectional

Contextualized

Long short-term network

Multilingual

Stacked

Attention

Character

Universal

Convolution

transformer

TASKS

Model	Avg	Single Sentence CoLA	Single Sentence SST-2	Similarity and Paraphrase MRPC	Similarity and Paraphrase QQP	Similarity and Paraphrase STS-B	Natural Language Inference MNLI	Natural Language Inference QNLI	Natural Language Inference RTE	Natural Language Inference WNLI
Single-Task Training										
BiLSTM	62.0	15.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	60.8	52.8	62.3
+ELMo	66.2	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	69.4	50.1	**65.1**
+CoVe	62.4	14.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	64.8	53.5	61.6
+Attn	60.0	15.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	51.9	51.9	55.5
+Attn, ELMo	64.8	**35.0**	90.2	68.8/80.2	**86.5**/**66.1**	55.5/52.5	**76.9**/**76.7**	61.1	50.4	**65.1**
+Attn, CoVe	60.8	14.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	53.8	52.7	64.4
Multi-Task Training										
BiLSTM	63.5	24.0	85.8	71.9/82.1	80.2/59.1	68.8/67.0	65.8/66.0	71.1	46.8	63.7
+ELMo	64.8	27.5	89.6	76.2/83.5	78.5/57.8	67.0/65.9	67.1/68.0	66.7	55.7	62.3
+CoVe	62.2	16.2	84.3	71.8/80.0	82.0/59.1	68.0/67.1	65.3/65.9	70.4	44.2	**65.1**
+Attn	65.7	0.0	85.0	75.1/**83.7**	84.3/63.6	73.9/71.8	72.2/72.1	82.1	**61.7**	63.7
+Attn, ELMo	**69.0**	18.9	**91.6**	**77.3**/83.5	85.3/63.3	72.8/71.1	75.6/75.9	81.7	61.2	**65.1**
+Attn, CoVe	64.3	19.4	83.6	75.2/83.0	84.9/61.1	72.3/71.1	69.9/68.7	78.9	38.3	**65.1**
Pre-Trained Sentence Representation Models										
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	75.1	54.1	62.3
Skip-Thought	61.5	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	74.7	53.1	**65.1**
InferSent	64.7	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	79.8	58.0	**65.1**
DisSent	62.1	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	75.2	56.4	**65.1**
GenSen	66.6	7.7	83.1	76.6/83.0	82.9/59.8	**79.3**/**79.2**	71.4/71.3	**82.3**	59.2	**65.1**

Table 3: Baseline performance on the GLUE tasks. For MNLI, we report accuracy on the matched and mismatched test sets. For MRPC and Quora, we report accuracy and F1. For STS-B, we report Pearson and Spearman correlation. For CoLA, we report Matthews correlation. For all other tasks we report accuracy. All values are scaled by 100. A similar table is presented on the online platform.

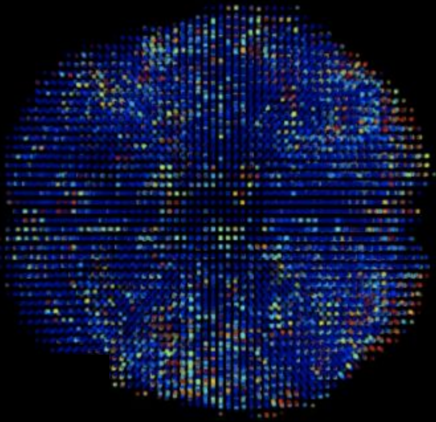
Wang et al, (2018): GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

Same same
but different?

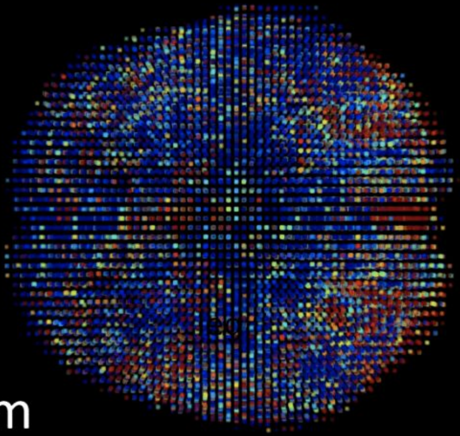
- Why is one model better than another?
- How do the representations differ?
- Which linguistic properties are encoded?
- Which phenomena cannot be modeled?
- How are they different from human language processing?

Compare computational models with the signal that we measure when humans process language.

But how?

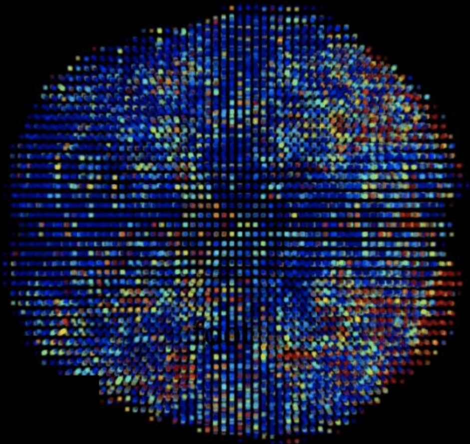


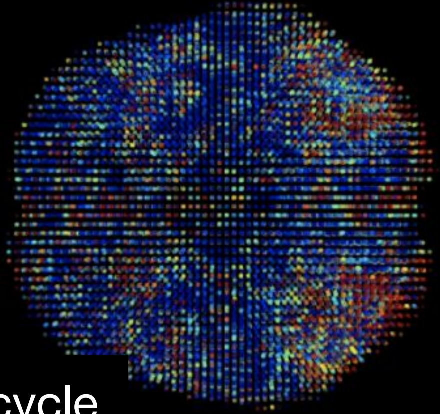
ant



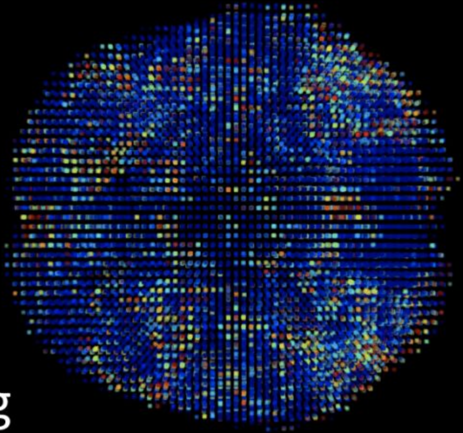
arm

?



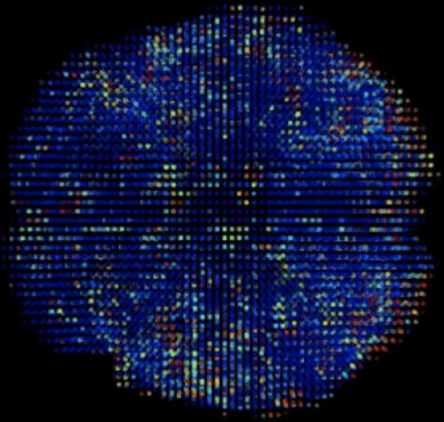


bicycle

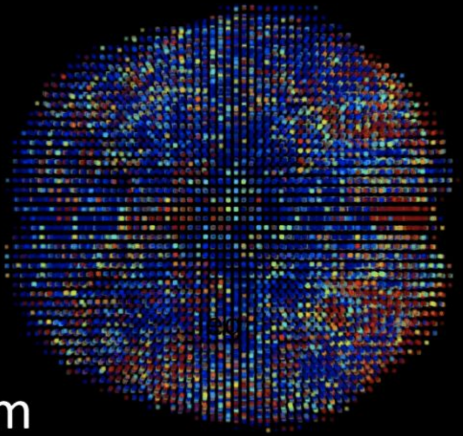


leg

Data: Mitchell et al. 2008
Visualization: Samira Abnar

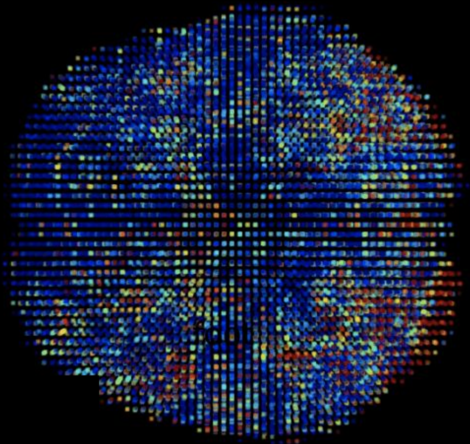


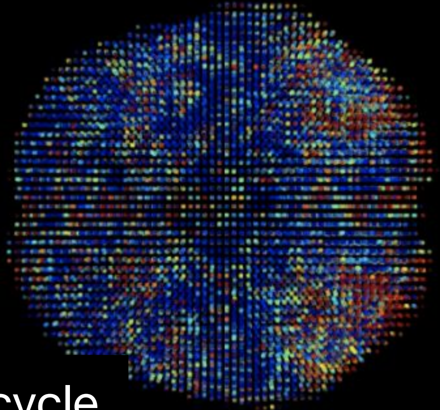
ant



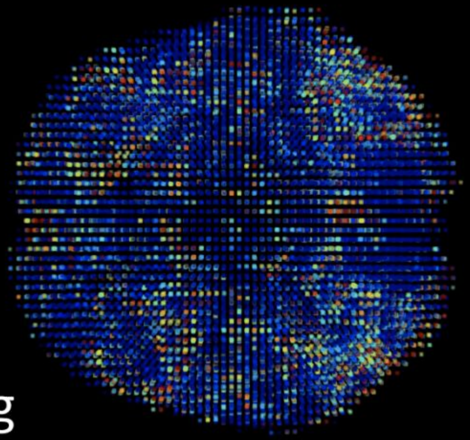
arm

?





bicycle

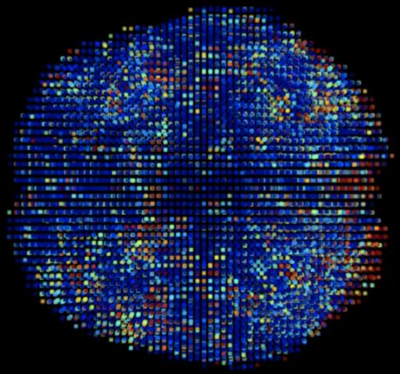


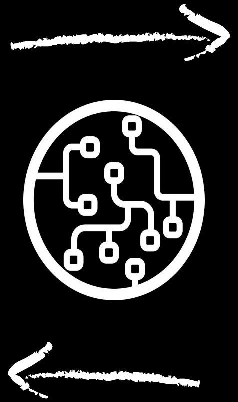
leg

Hand or foot?

Data: Mitchell et al. 2008
Visualization: Samira Abnar

Learn mapping model





cat: [2 6 8 1 3 1 0 0 …]

What's next?

Words in Context!

- *Harry had never believed he would meet a boy he hated more than Dudley.*
  [Wehbe et al. 2014]
- *A few weeks ago, a man I hardly know wrote me a really sweet love letter.*
  [Dehghani et al. 2017]
- *Alice was beginning to get very tired of sitting by her sister on the bank.*
  [Brennan et al. 2016]

DOES THAT WORK?

We don't really now...

WHAT'S THE PROBLEM?

Researchers use

- different datasets

WHAT'S THE PROBLEM?

Researchers use

- different datasets
- different encoding models

WHAT'S THE PROBLEM?

Researchers use

- different datasets
- different encoding models
- different experimental parameters

WHAT'S THE PROBLEM?

Researchers use

- different datasets
- different encoding models
- different experimental parameters
- different evaluation metrics

WHAT'S THE PROBLEM?

Researchers use

- different datasets
- different encoding models
- different experimental parameters
- different evaluation metrics
- no comparison to baseline

WHAT'S THE PROBLEM?

Researchers use

- different datasets
- different encoding models
- different experimental parameters
- different evaluation metrics
- no comparison to baseline

AND THEY ARE OFTEN NOT VERY TRANSPARENT ABOUT THE DIFFERENCES:
NO DATA, NO CODE

OUR APPROACH

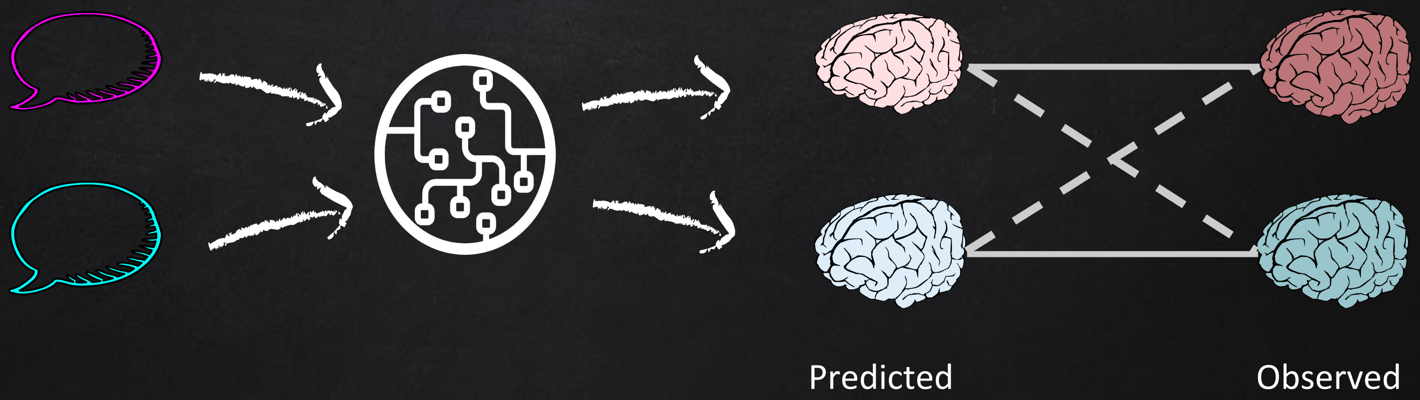
We use

- 4 datasets
- Constant encoding model
- Constant experimental parameters (if possible)
- Multiple evaluation metrics
- Comparison to a baseline with a “random language model”

We tried to standardize the procedure as much as possible and publish the experimental framework on github.

Evaluation methods

1. Pairwise evaluation



Evaluation methods

1. Pairwise evaluation
2. Voxelwise evaluation

IDEA: Not all voxels are related to language processing.

Evaluate the prediction for every voxel individually.

Evaluation methods

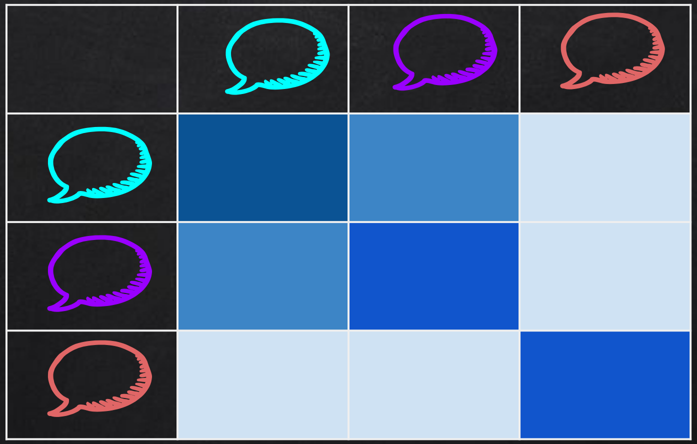
1. Pairwise evaluation
2. Voxelwise evaluation
3. Representational similarity analysis

Idea: directly compare relations between stimuli

No more prediction!

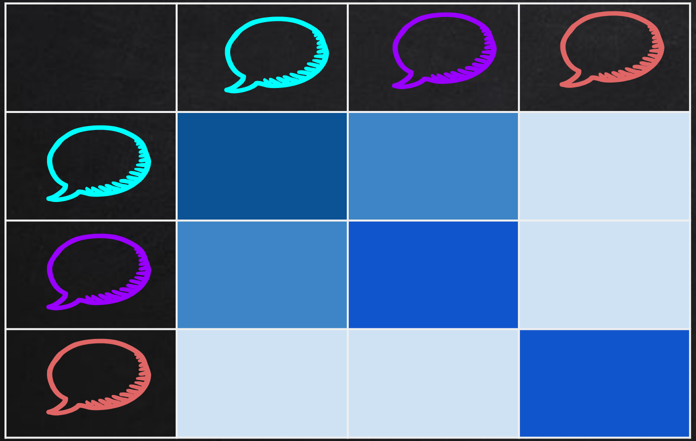
Evaluation methods

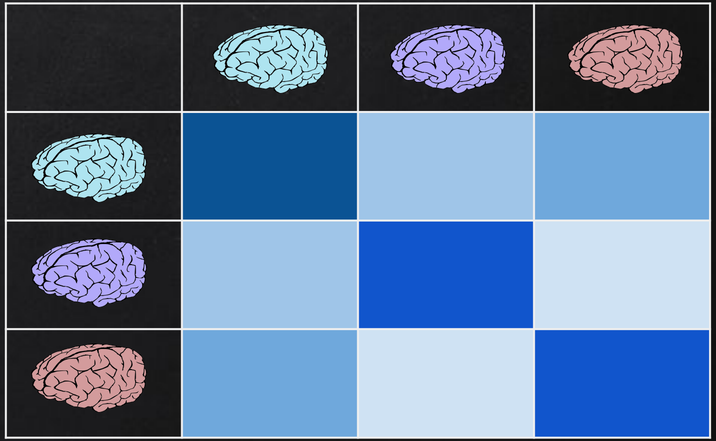
1. Pairwise evaluation
2. Voxelwise evaluation
3. Representational similarity analysis



Evaluation methods

1. Pairwise evaluation
2. Voxelwise evaluation
3. Representational similarity analysis





Evaluation methods

1. Pairwise evaluation
2. Voxelwise evaluation
3. Representational similarity analysis

Each method can be realized with different parameters.

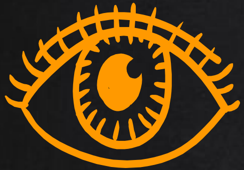
We compare their effects.

EXAMPLE PIPELINE

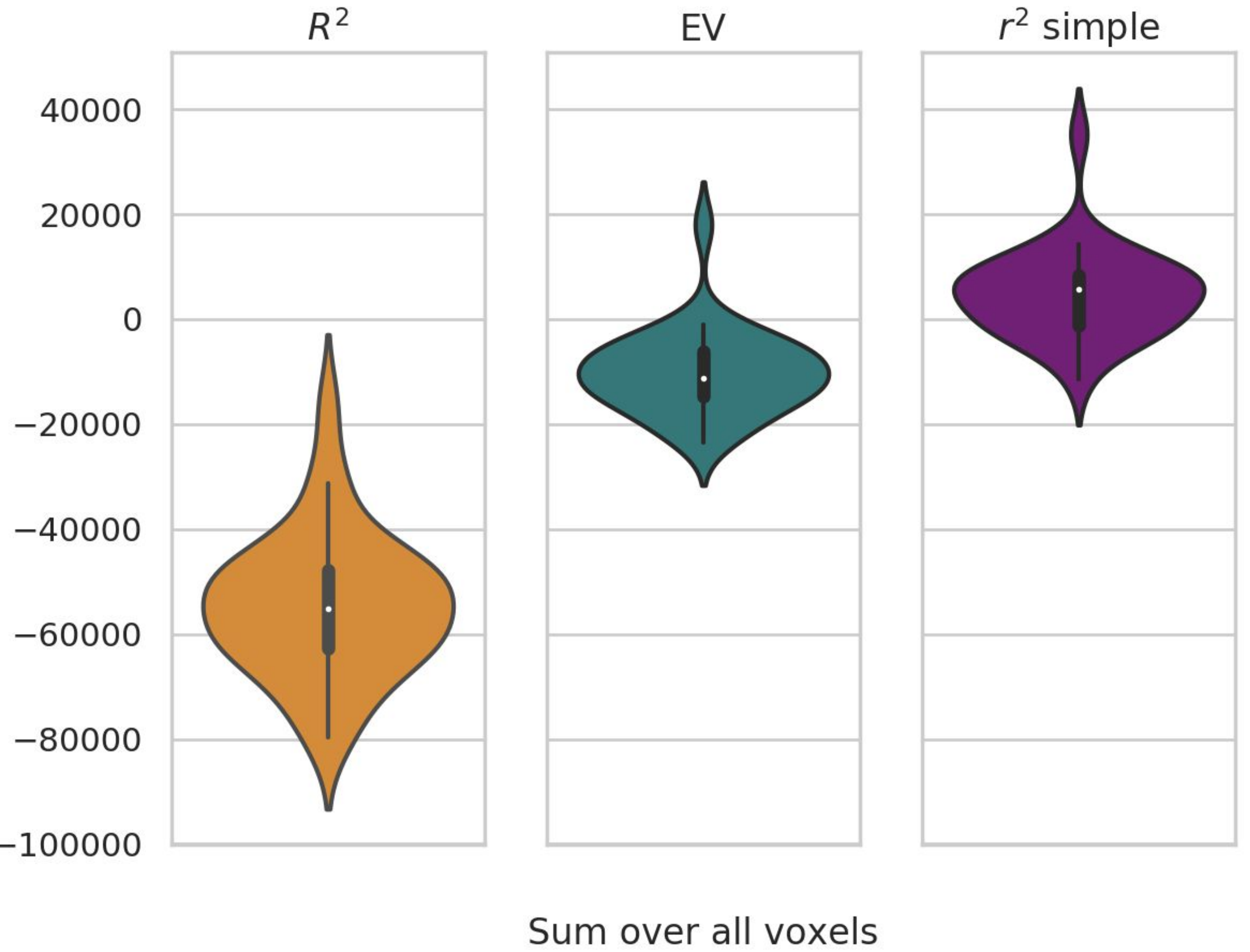
```
# Set the components
mitchell_reader = WordsReader(data_dir=mitchell_dir)
mapper = RegressionMapper()
stimuli_encoder = ElmoEncoder(save_dir)
random_encoder = RandomEncoder(save_dir)

# Try different language models
for encoder in [stimuli_encoder, random_encoder]:

    # Set up the pipelines
    mitchell_pipeline_name = "Words" + encoder.__class__.__name__
    mitchell_pipeline = SingleInstancePipeline(mitchell_reader, encoder, mapper,
                                               mitchell_pipeline_name, save_dir=save_dir)
    mitchell_pipeline.voxel_selection = "none"

    mitchell_pipeline.pairwise_procedure("Mitchell_pairwise_noVS")
    mitchell_pipeline.run_standard_crossvalidation("Mitchell_CV_noVS")
    mitchell_pipeline.runRSA("Mitchell_RSA")
```

TRY IT YOURSELF:

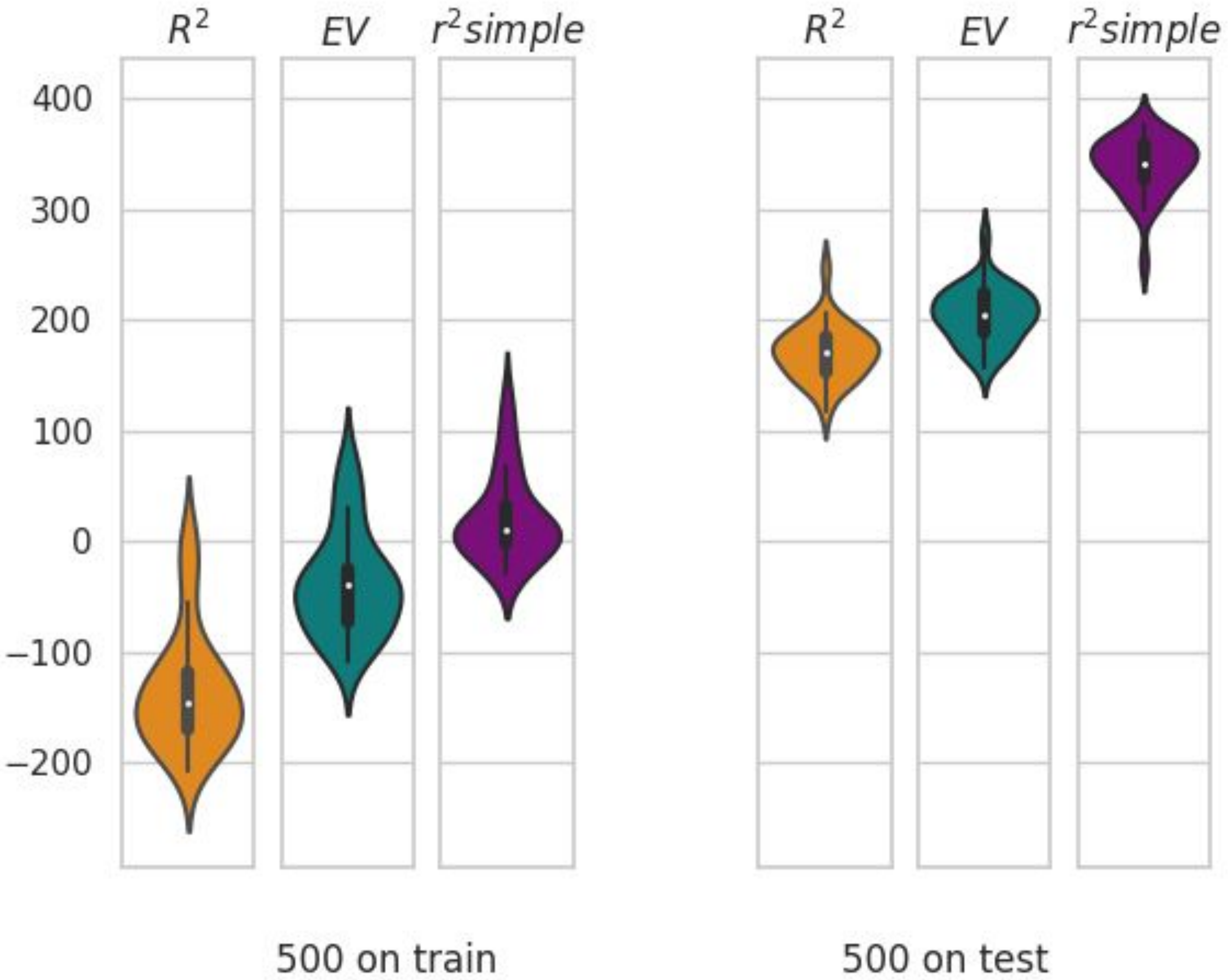
https://github.com/beinborn/brain-lang



A PEAK AT THE RESULTS?



Voxel-wise results for the data by Dehghani et al. (2017)



Voxelwise results with selected voxels for the data by

Dehghani et al. (2017)

So?

Language-brain encoding is hard.
Many crucial design decisions:
preprocessing, language model, encoding parameters, …

1. Make these decisions transparent and reproducible.
2. Analyze your hypothesis on several datasets with the same metric.
3. Compare to reasonable baselines.
4. Do not oversell your results! A tiny signal is already impressive.

Questions?

l.beinborn@uva.nl

Back-up slides

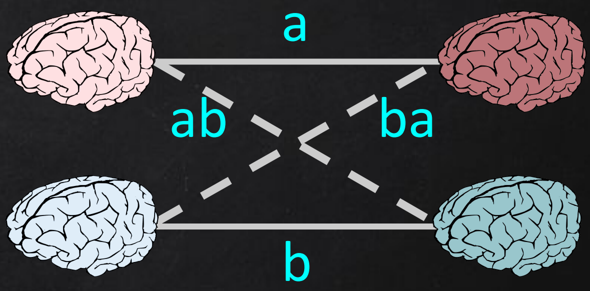
Pairwise evaluation

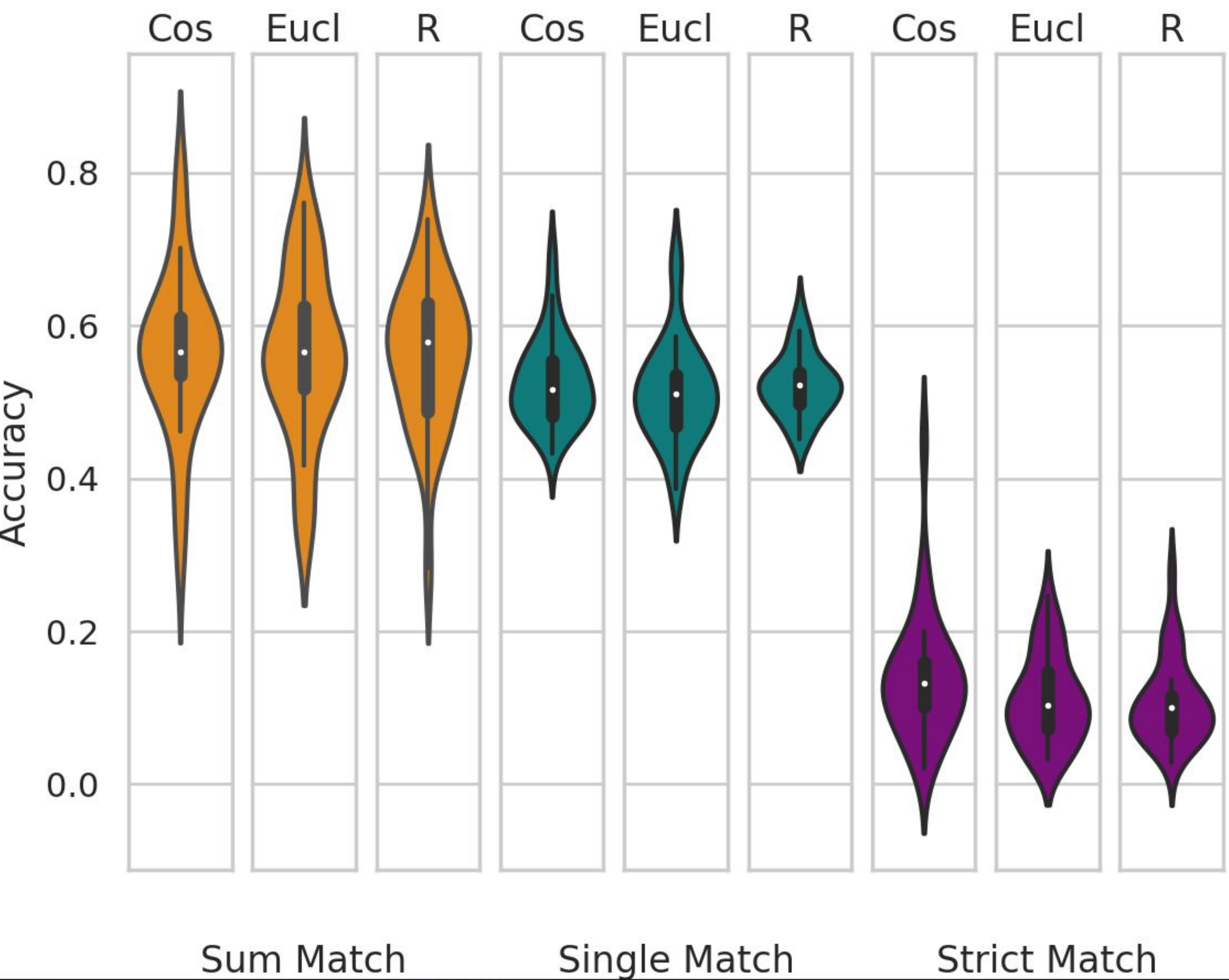
Match definition:

Sum: (a +b) < (ab + ba)

Single: a < ab

Strict: (a < ab) & (b < ba)





Results for the data by Dehghani et al. (2017)

Why?

Across datasets?

	Encoding Model (Random LM)				
	Match	Words	Stories	Alice	Harry
Cos	Sum	.67 (.54)	.57 (.53)	.54 (.53)	.50 (.49)
	Single	.60 (.53)	.53 (.53)	.53 (.51)	.49 (.49)
	Strict	.26 (.13)	.14 (.02)	.28 (.27)	.25 (.24)

RSA Results

	Words	Stories	Alice	Harry
Pearson	0.41	0.19	0.06	0.06
Spearman	0.09	0.22	0.02	0.03

RSA Results – Random LM

	Words		Stories		Alice		Harry	
Pearson	0.41	**0.44**	0.19	**0.21**	0.06	**0.02**	0.06	**0.03**
Spearman	0.09	**0.05**	0.08	**0.09**	0.03	**0.01**	0.00	**0.01**