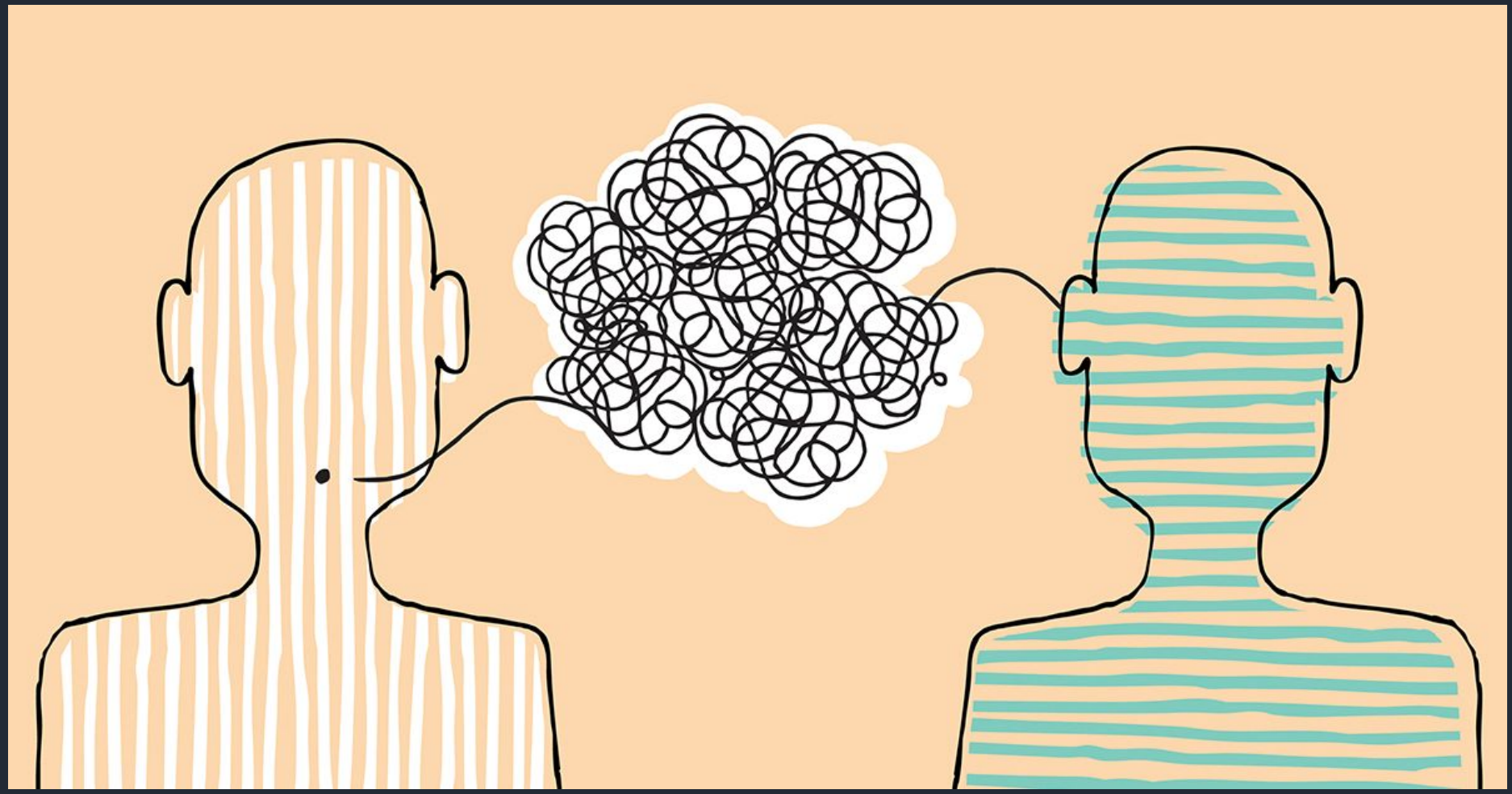
Cognitive Plausibility of Deep Language Models

Lisa Beinborn
VU Amsterdam

Nora Hollenstein
U Copenhagen

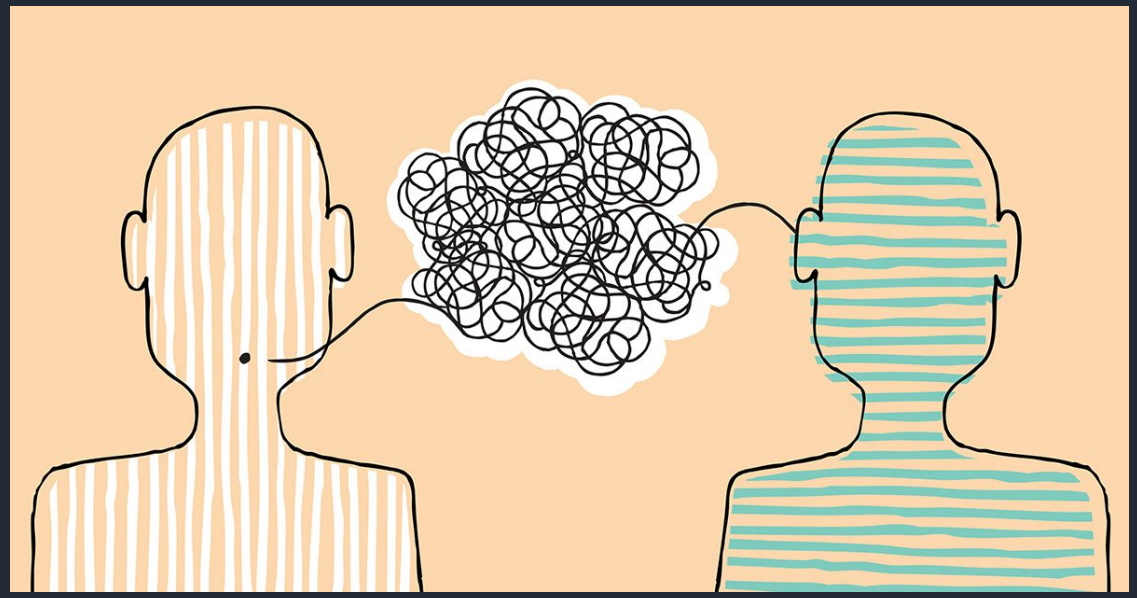
Willem Zuidema
Uv Amsterdam

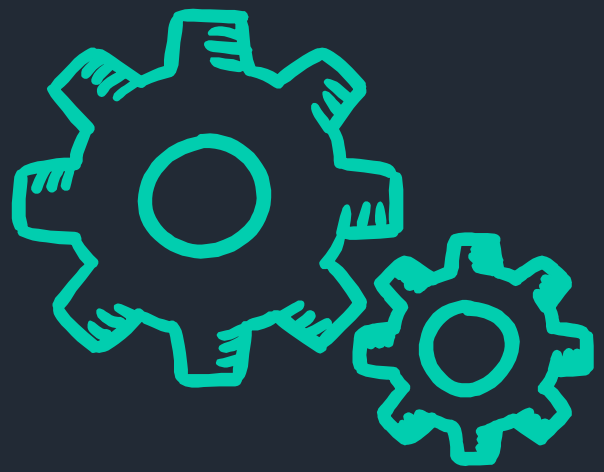


Quick RECAP

https://www.analyticsinsight.net/why-the-future-is-bright-for-natural-language-processing/

Quick Recap





https://www.analyticsinsight.net/why-the-future-is-bright-for-natural-language-processing/

When is a language model cognitively plausible?

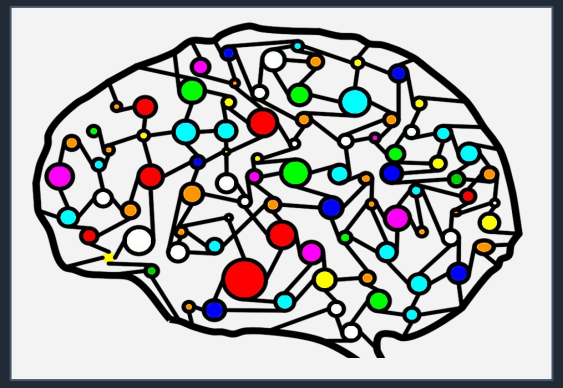
Quick RECAP

We work on the intersection of three fields:

Computational Linguistics: If it makes similar decisions as humans.

Psycholinguistics: If it exhibits similar processing patterns as humans.

Interpretability: If it uses similar intermediate representations as humans.



Evaluation paradigms

- Cognitive analysis
- Linguistics analysis
- Representational analysis

Human signals

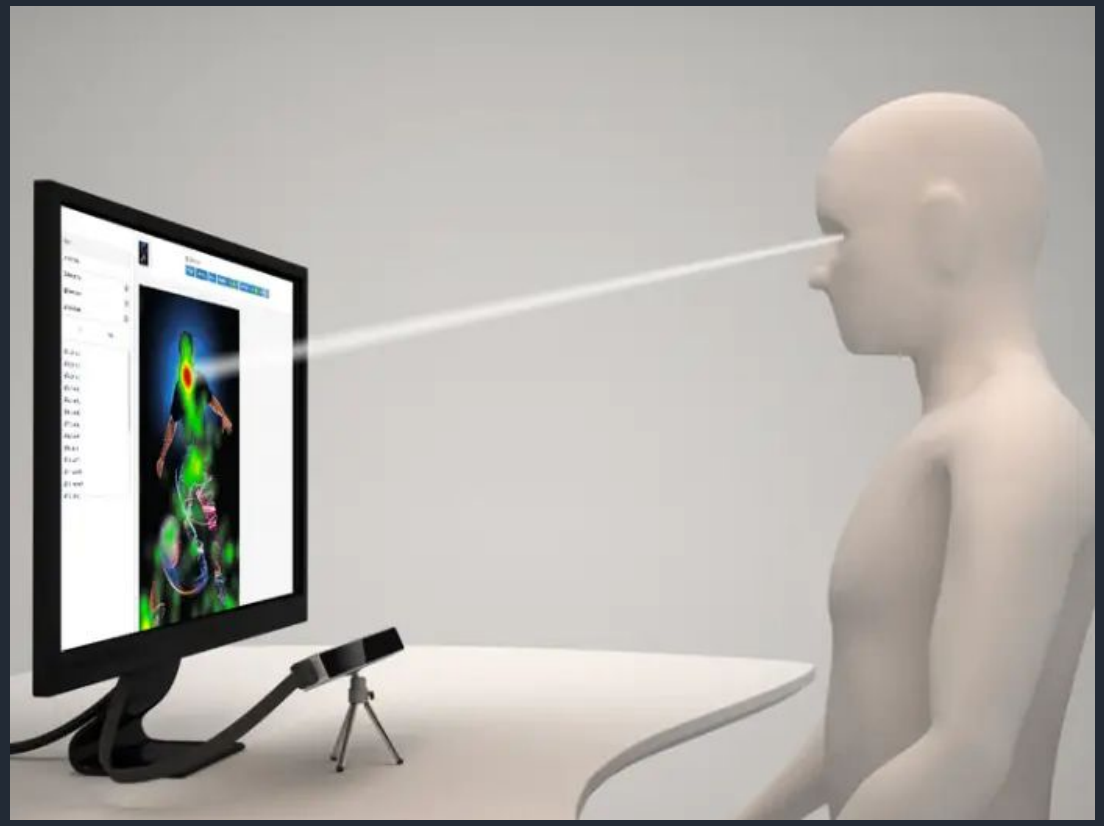
- Human language processing data comes in many forms
- Behavioral or physiological data
- Capturing conscious or unconscious cognitive processes

FOCUS TODAY:

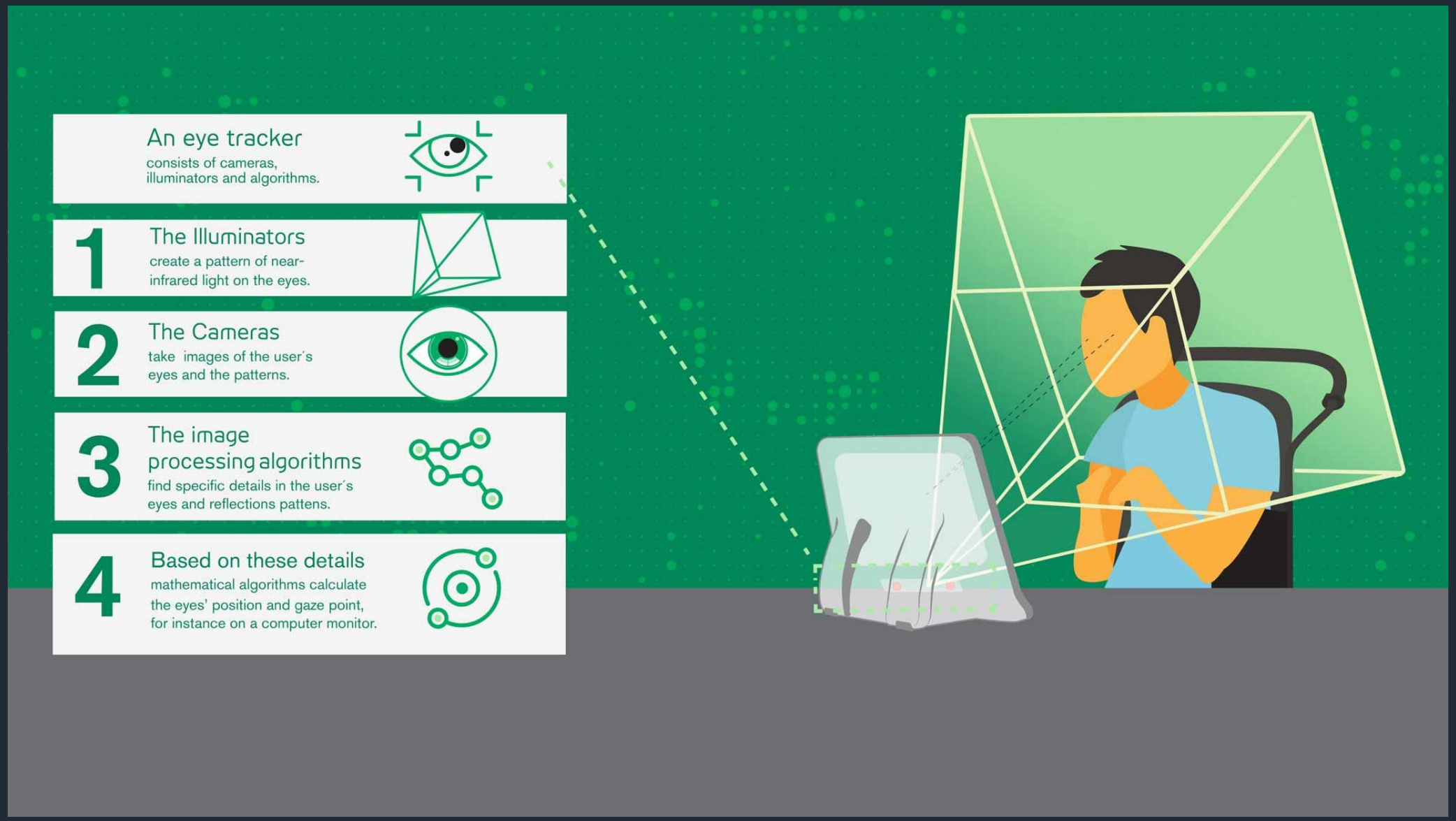
- Eye movements
- Brain activity

Eye tracking

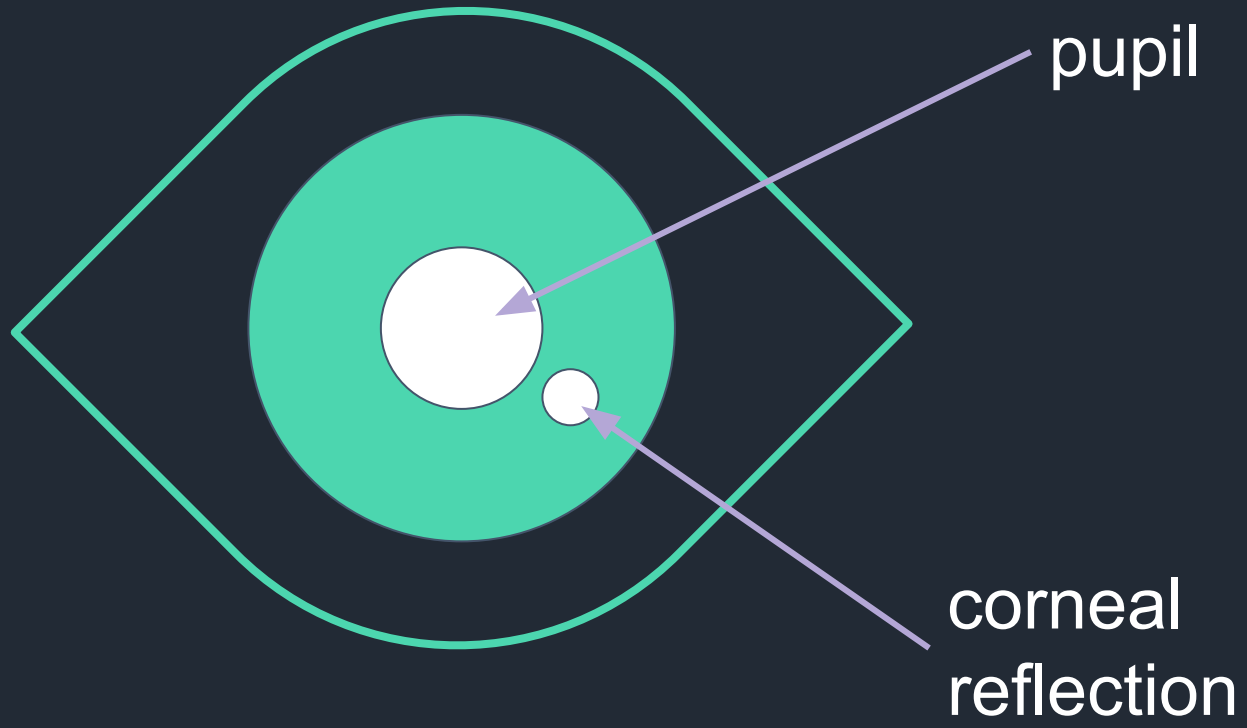
- Eye tracking is the process of measuring the point of gaze at a given time.
- An eye tracking devices measures eye positions and eye movements.



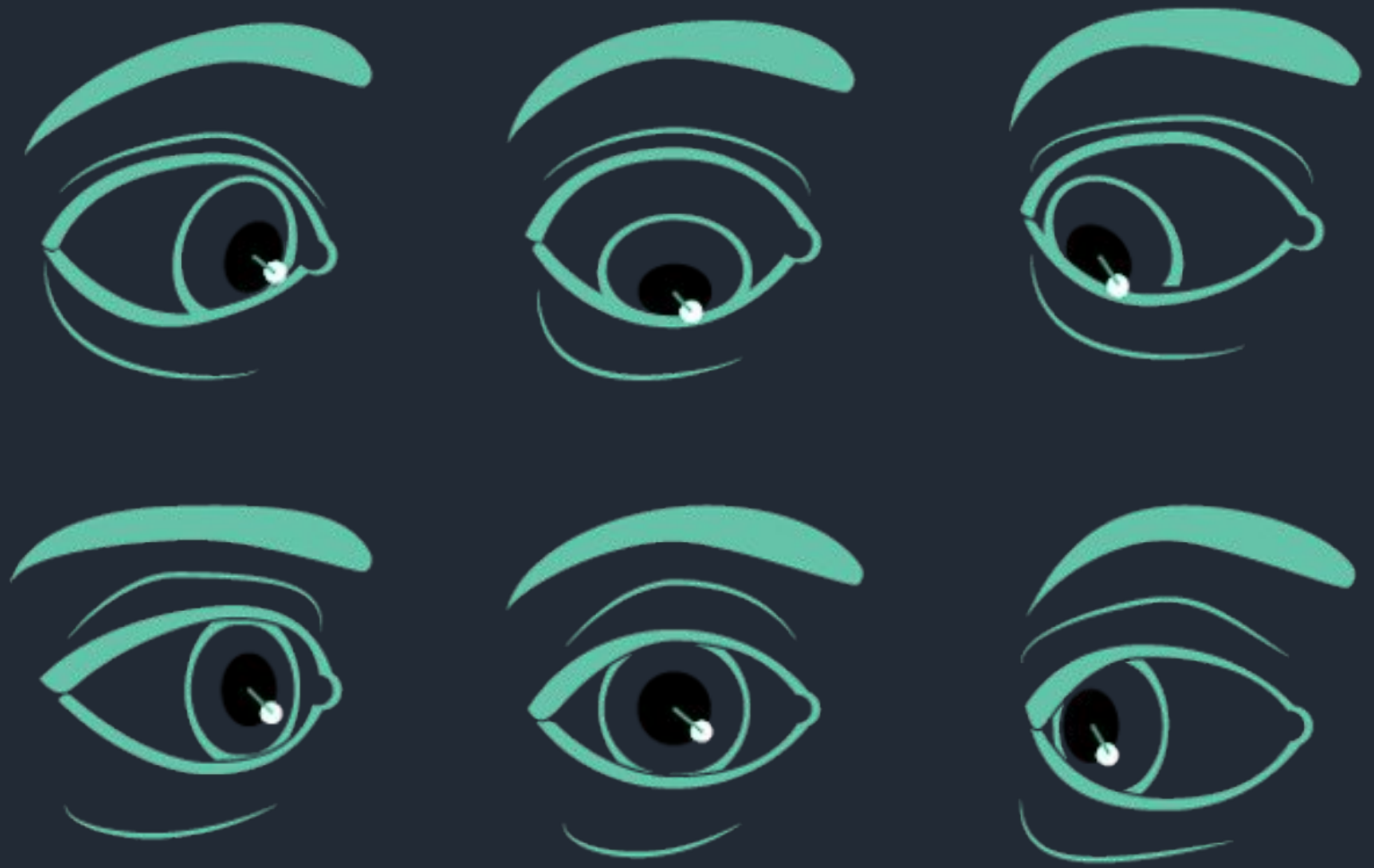
Eye Tracking Procedure



The center of the eye (pupil center) is tracked in relation to the position of the corneal reflection. The relative distance between the two areas allows the calculation of the direction of the gaze.

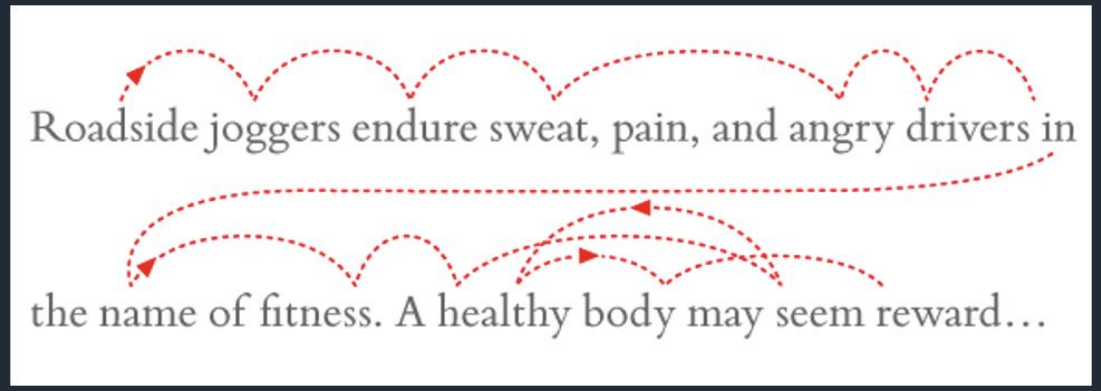


Deduction of gaze direction



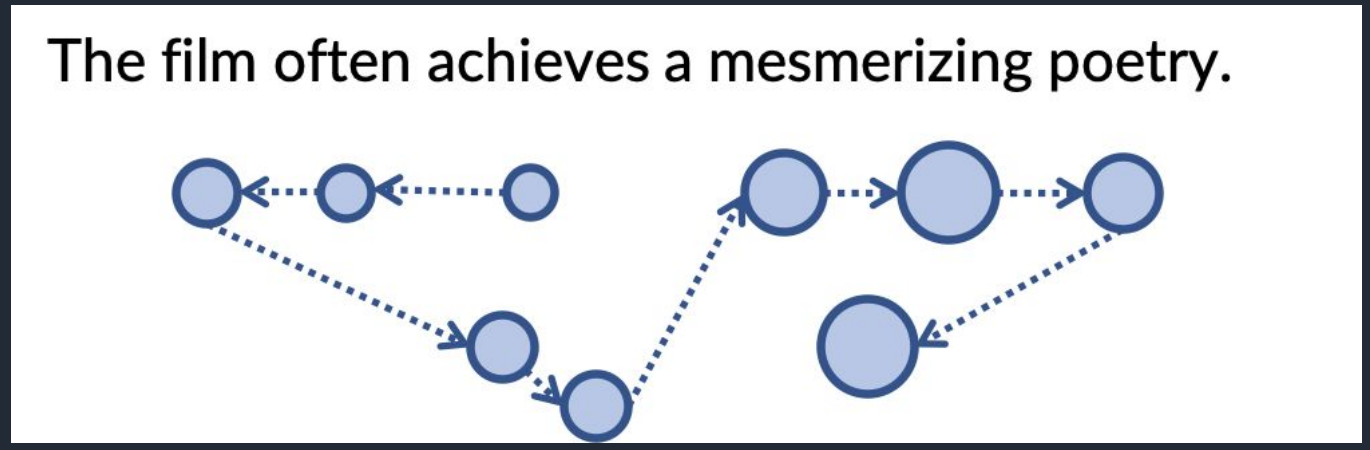
Eye movements

- Fixations - maintaining of the visual gaze on a single location
- Saccades - quick, simultaneous movement of both eyes



Eye movements in reading

- Visual processing of written text
- Eye movements during reading → during language understanding
- Indirect measure of cognitive load
- Regressions: backward saccades



READING IS NOT A SMOOTH MOVEMENT IN ONE DIRECTION!

Psycholinguistic evidence form reading

Fixation times depend on …

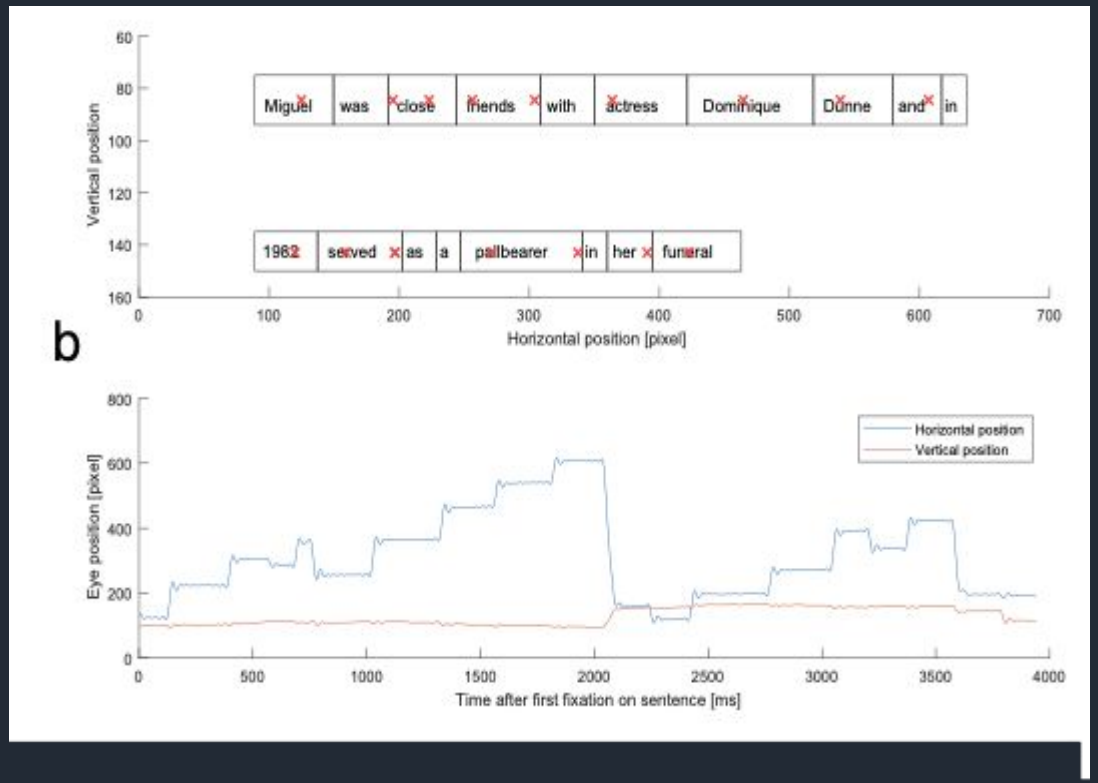
- word length
- word frequency
- word predictability
- word familiarity

QUIZ → https://forms.gle/HHy5hcZDy1i5oSpj9

See additional readings for more info on linguistic effects in eye tracking

Preprocessing

- Cleaning: Filtering fixations
- Extract word boundaries, map to x,y coordinates of the screen
- Extract features (many corpora already provide extracted features)
- Average over all readers



Recording devices

- Desktop/stationary devices
- Portable/mobile headsets
- Webcam-based eye tracking

Eye-tracking experiment design

- Define research questions
- Expected findings, hypotheses
- Type of eye-tracking device
  - Sampling rate vs. research question
  - Cost
- Naturalistic vs. controlled experiments
- Participant demographics
- Control conditions, additional tests
- Selection of (reading) materials

Eye tracking applications

- Marketing research
- Driver safety
- Medical use:
  - Diagnosis of autism spectrum disorders
  - Diagnosis of dyslexia
  - Radiology & image analysis
  - Training of physicians
- Augmentative and alternative communications
- Computer vision: object recognition
- Behavioral and decision-making studies
- Reading studies

Benefits & Challenges

+ Study early and late cognitive processing
+ Well-established word-based metrics
+ Datasets in multiple language readily available

- Large variability between subjects and populations
- Noise
- High accuracy required for reading studies

functional magnetic resonance imaging (fMRI)

fMRI

functional magnetic resonance imaging

for the purpose of **functional** neuroimaging: analyzing what the brain is doing

makes use of static **magnetic** fields in the brain

produces a 3D **image** of the brain

spatial resolution

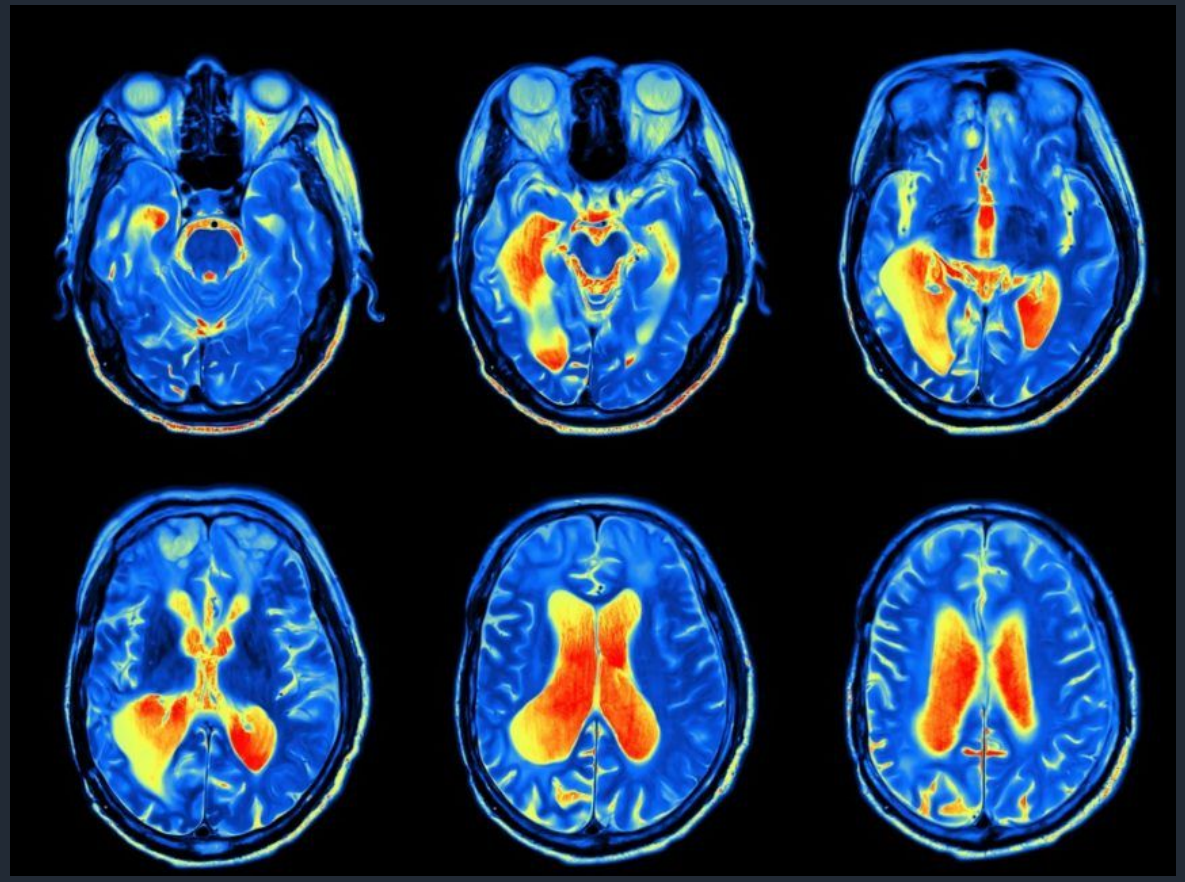
temporal resolution

fMRI Properties

- Measurement based on the properties of **oxygen-rich blood**
- MRI uses a strong, permanent, static **magnetic field** to align nuclei in the brain region being studied
- fMRI extends MRI to capture **functional changes** in the brain caused by neuronal activity
- Blood flow and blood oxygenation in the brain (= **hemodynamics**) are closely linked to **neural activity**

Blood-oxygen-level dependent (BOLD) contrast

- Hemodynamic response: changes in blood flow
- This response is related to the energy use of brain cells:
  Firing neurons need more energy to be brought in quickly
- More oxygen released from the blood to more active neurons
- → Causes a change in the relative levels of oxygen in the blood, which can be detected through their magnetic susceptibility



Hemodynamic delay in fMRI

Brain activity is not *directly* measured; instead the human hemodynamic responses to brief periods of neural activity are delayed and dispersed in time.

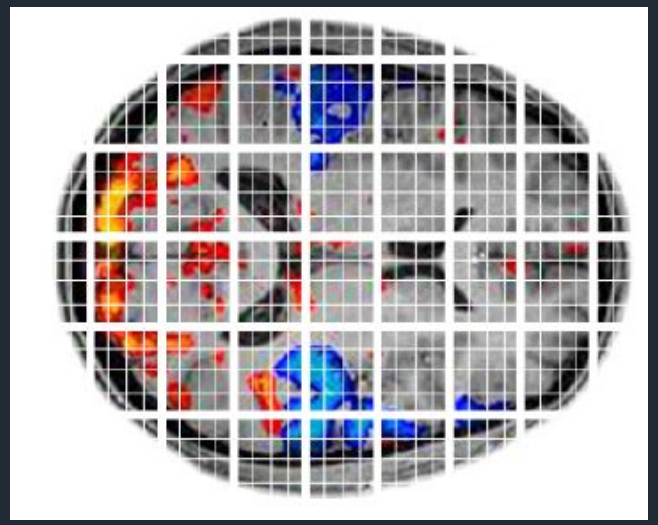
→ **Hemodynamic delay**

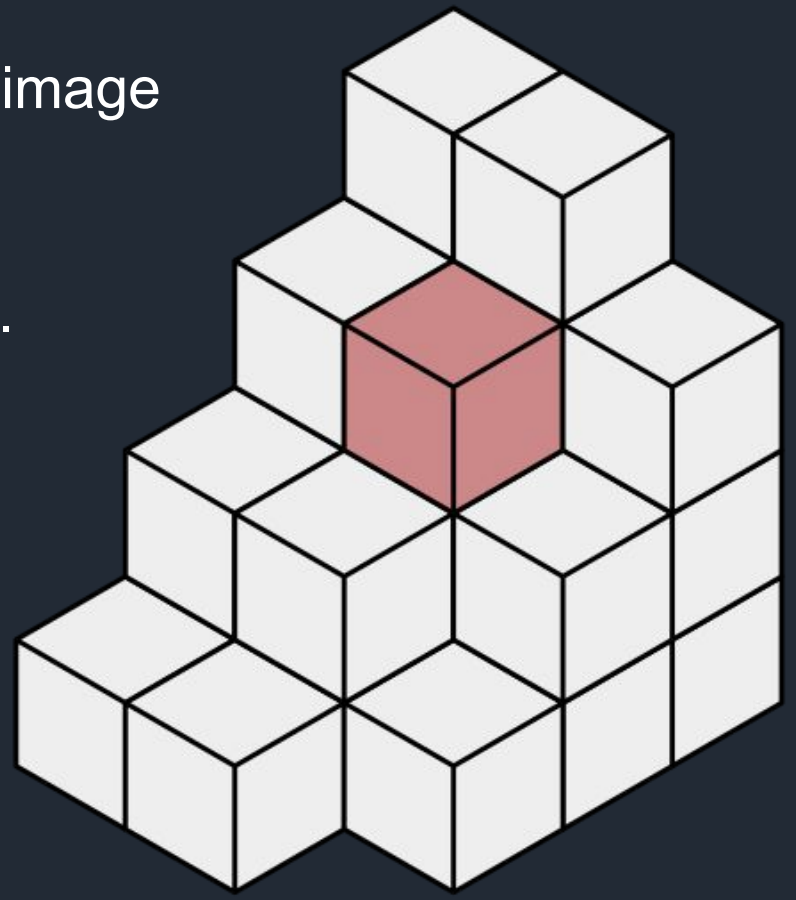
The fMRI signal measures a brain response to a stimulus with a delay of a few seconds, and it decays slowly over a duration of several seconds.

This leads to difficulties in processing fMRI signals for lower-level features

Voxels

- The 3-dimensional fMRI image is built up in units called **voxels**
- Each one represents a tiny cube of brain tissue – a 3-D image building block analogous to the 2-D pixels.
- Each millimeter-sized voxel represents many brain cells.





fMRI Preprocessing

Standard **preprocessing steps** for fMRI data:

- Motion correction
- Slice timing correction
- Spatial normalization
- Spatial smoothing

fMRI Features

- 3D scan flattened to vector of voxels
- 3D image or matrix of voxels as input to convolutional neural networks

- **Voxel selection**:

  - randomly select a subset of voxels

  - information-driven (based on research question)

  - by regions of interest (known from previous research literature)

  - dimensionality reduction methods

fMRI data example

190042 voxels →

180 concepts/ examples ↓

180x190042 double

	1	2	3	4	5	6	7	8
1	105.1717	78.5471	51.3253	34.0190	29.4969	38.5325	113.2303	61.0518
2	93.2773	56.8028	33.3003	22.0770	22.0059	32.6231	19.6104	4.0131
3	-5.5998	-7.1864	-5.3221	-7.7611	-16.1646	-27.6794	7.9997	-0.2619
4	11.8732	10.8655	7.4613	7.6971	10.3729	15.2458	16.2554	9.8187
5	50.5193	29.5698	14.6296	4.3416	2.3942	14.9320	43.8499	17.9002
6	-29.6590	-24.7877	-19.7177	-11.2459	-6.0769	-11.4111	-23.5074	-10.4111
7	-84.8250	-56.2264	-38.2519	-28.7274	-27.5316	-30.6950	-13.3020	-0.0861
8	18.5263	13.3628	6.7099	0.6270	-1.7662	-2.0860	5.4359	3.1026
9	111.6956	73.2435	38.6063	24.2963	25.1311	42.0884	121.3226	56.8877
10	78.4998	50.7919	25.2342	12.1178	11.5400	23.4242	46.9230	21.2619
11	-25.6453	-13.5896	-6.7460	-7.4112	-12.5578	-21.9286	-24.9310	-13.8567
12	119.3014	83.6745	58.7098	45.7978	46.1688	54.3329	89.6363	39.7305
13	-35.2468	-21.8310	-20.8987	-22.4951	-23.9237	-26.9108	-14.5857	-2.1502

Pereira et al. (2018)

fMRI Applications

Clinical applications

- Surgical planning (identify critical area, test performance of specific tasks, e.g. language)
- Drug interventions (use fMRI to identify brain changes associated with treatment)
- Psychiatric disorders (e.g. bipolar disorders, mania, social anxiety, autism, panic disorder, …)

Research

- Mapping of brain functions and brain areas

Challenges in fMRI

- Expensive testing that requires much effort
- Physical limitations: not everyone can be scanned (claustrophobia, pregnancy, implants, pacemakers, …)
- Noise: neuroimaging data contains many artifacts – how to isolate the cognitive process being investigated?
- Interpretation: What does the data actually mean? Is it possible to draw conclusions?

fMRI experiment design

1. Define the cognitive process to examine
2. Define tasks to manipulate that process (and control conditions!)
3. Measure fMRI signals during the tasks
4. Compare fMRI data between tasks

Eye Tracking vs fMRI

Temporal resolution	High	Low
Spatial resolution	-	High
Measures brain activity?	Indirectly	Indirectly
Level of expertise	Little training	Extensive training
Cost	Accessible	Extensive research funding required
Portability	Portable devices available	Not portable

General considerations for experiment design

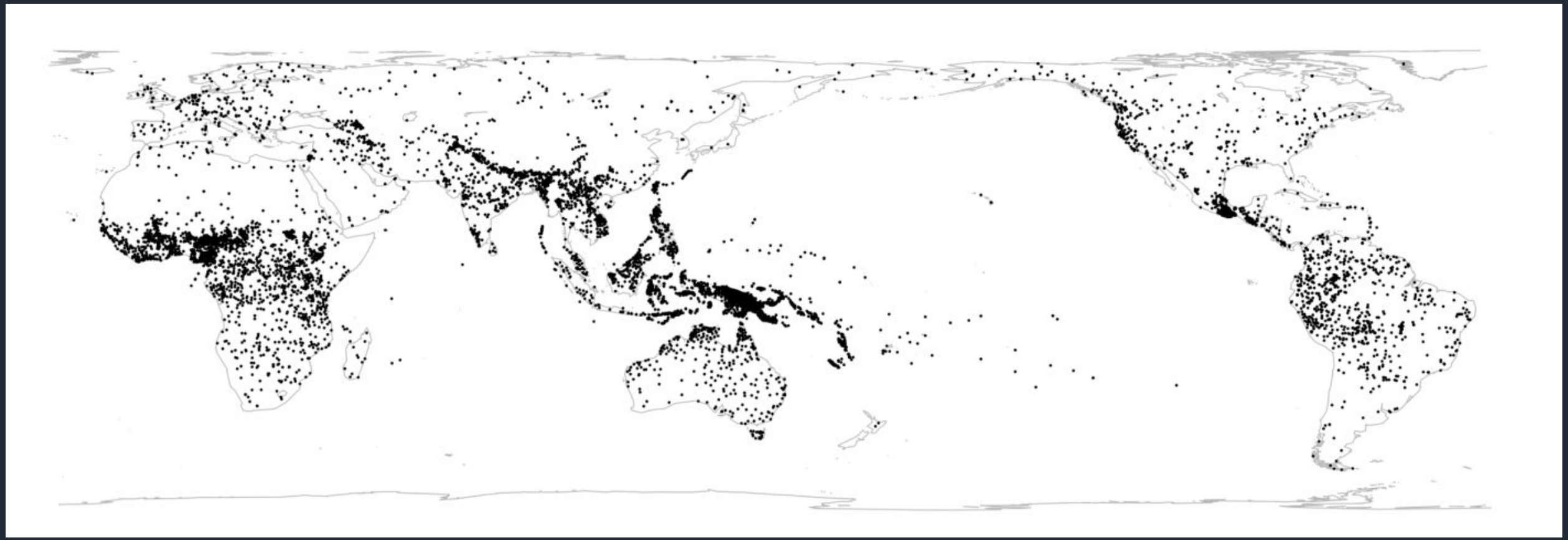
- How is the stimulus presented on the screen?
  Fontsize, background color, how many lines & words on one screen, etc.
- Which software is used to build the experiments
- How many sessions, which order, how many trials?
- Control conditions – e.g. comprehension questions, or valid/invalid trials

A Note on Ethics

Most people are not *WEIRD*… (Henrich et al., 2010)

- Western, educated, industrialized, rich and democratic (WEIRD)
- How generalizable are experimental findings?
- Challenges our understanding of human psychology and behaviour
- Basic cognitive and motivational processes vary across populations

Languages across the world



Source: https://www.comparativelinguistics.uzh.ch/dam/jcr:ffffffff-8a80-d988-0000-000026e6849b/Bickel_transcript-edited.pdf

Properties of cognitively plausible LMs

Human language processing is …

- efficient & robust
- incremental
- able to make predictions
- operates with limited memory
- is trained on limited data
- cross-linguistic

More on this topic: Keller (2010)

Evaluation paradigms

- Cognitive analysis
- Linguistics analysis
- Representational analysis

Predicting cognitive signals

How much do **computational language representations** reflect the **semantic representations** in the **human brain**?

→ Semantic maps in the human brain: https://gallantlab.org/huth2016/

Cognitive lexical semantics

Words should be defined by how the are organized in the brain (Miller & Fellbaum, 1992)

Words are defined by how they are organized in the brain

Meaning is linked to the way we mentally group all kinds of perceptions into conceptual categories

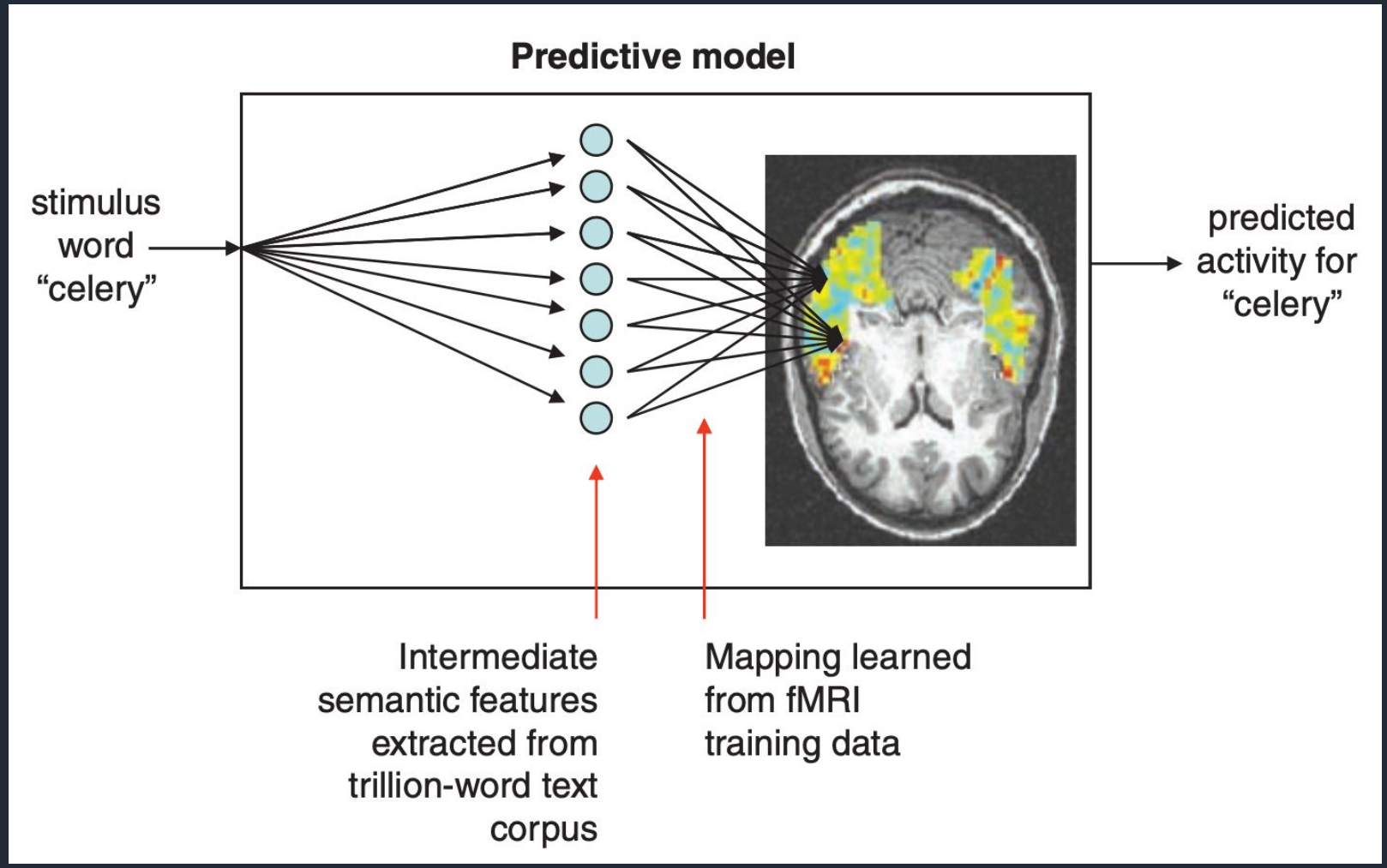
Language and cognition are inseparable: the structure of linguistic categories is assumed to reflect the structure of conceptual categories

Predicting fMRI activation of nouns

Pioneering study: Mitchell et al. (2008)

Goal: Predict neural activation when subjects are exposed to isolated word stimuli

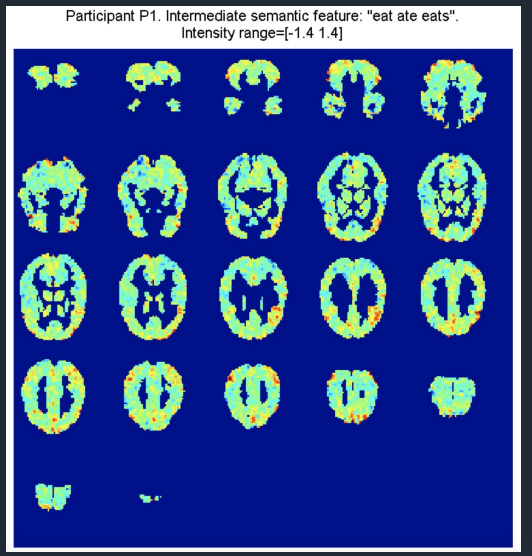
Step 1: Collect fMRI data



Predicting fMRI activation of nouns

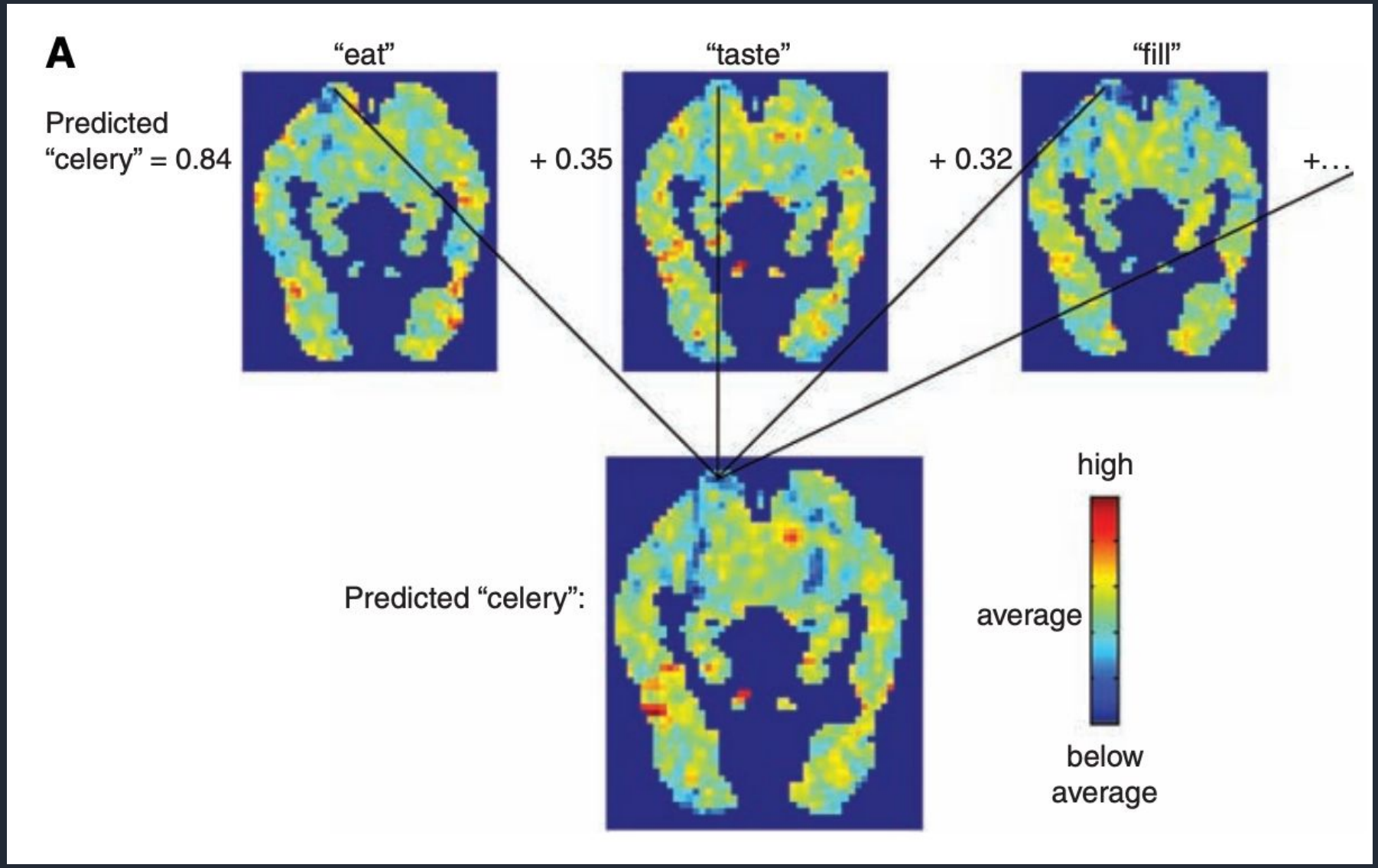
Step 2: Create corpus-based *semantic signatures*

- eat ate eats (0.837),
- taste tasted tastes (0.346),
- fill filled fills (0.315),
- see sees (0.243),
- clean cleaned cleans (0.115),
- open opens opened (0.060),
- smell smells smelled (0.059),
- touch touched touches (0.029),
- say said says (0.016),
- hear hears heard (0.000),
- listen listens listened (0.000),
- rub rubbed rubs (0.000),
- manipulate manipulates manipulated (0.000),
- run ran runs (0.000),
- push pushed pushes (0.000),
- move moved moves (0.000),
- fear fears feared (0.000),
- approach approaches approached (0.000),
- neared nears near (0.000),
- enter entered enters (0.000),
- drive drove drives (0.000),
- wear wore wears (0.000),
- lift lifted lifts (0.000),
- break broke breaks (0.000),
- ride rides rode (0.000)



Mitchell et al. (2008)

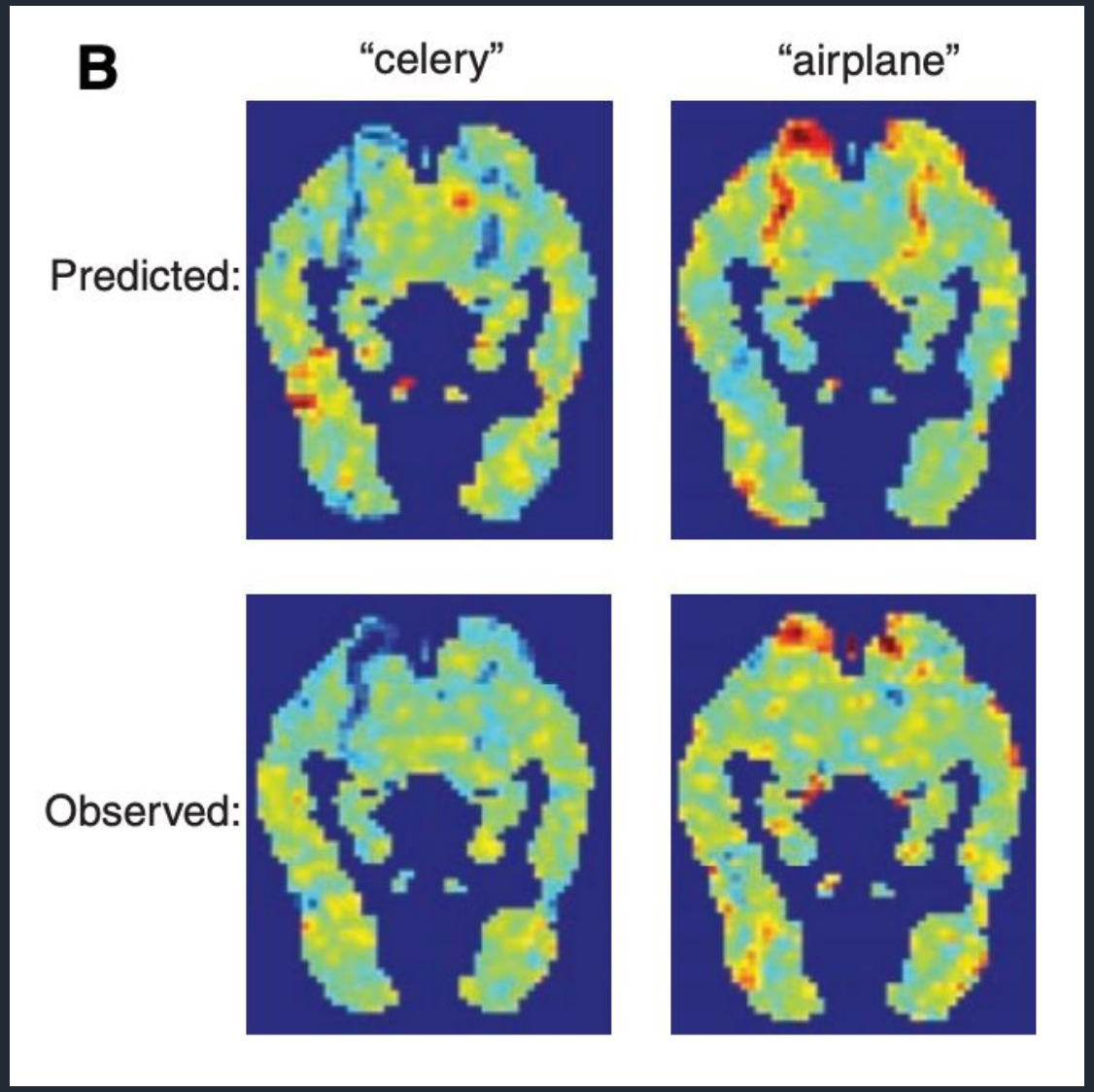
Defining a word using semantic features



Step 3: Learn a regression model to predict the neural activation of each word

Mitchell et al. (2008)

Evaluation



Step 4: Evaluate single-subject models with leave-two-out cross-validation

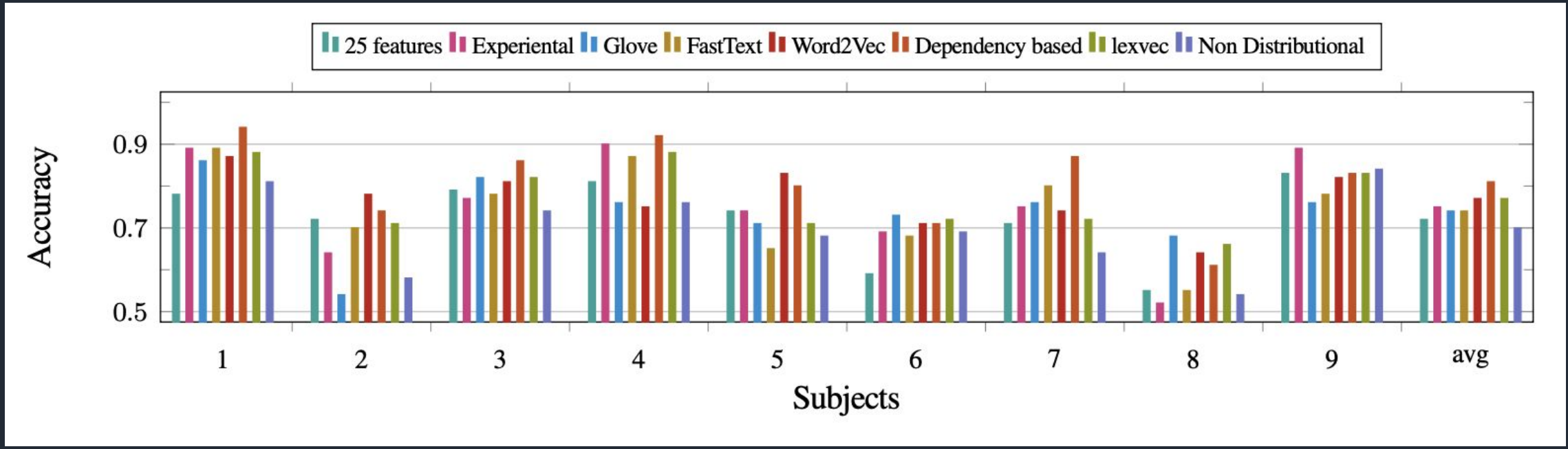
Average performance across subjects: 77% accuracy

Mitchell et al. (2008)

Word embedding evaluation

Abnar et al., (2018) use this fMRI dataset to decode cognitive signals as a method of word embedding evaluation:

- Neural word embedding models exhibit superior performance
- Error patterns deviate significantly between embedding types



Word embedding evaluation

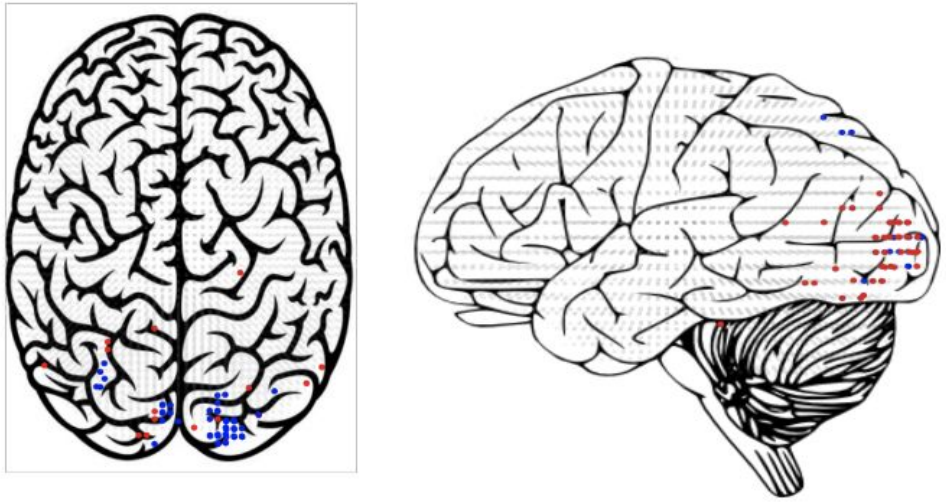


Figure 10: Most predictable voxels for dependecy based word2vec(red) and the experiential model(blue)

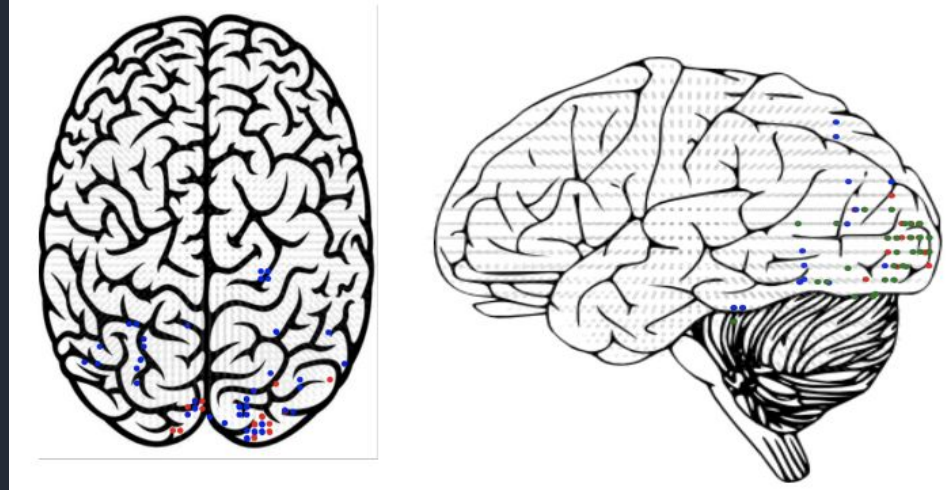


Figure 11: Most predictable voxels for dependecy word2vec(red) and word2vec(blue). Green dots are among the top 50 voxels of both models.

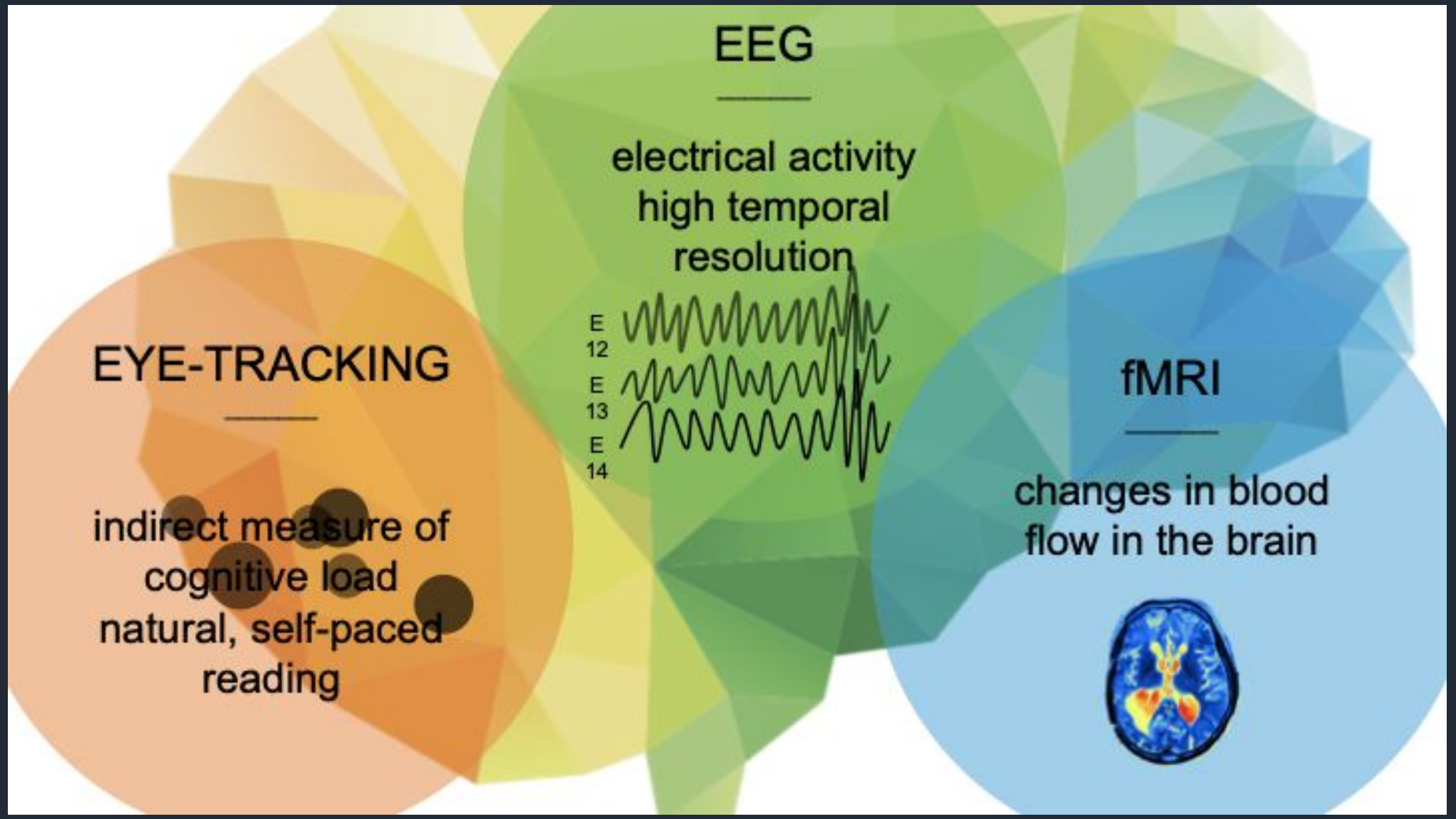
Abnar et al. (2018)

Generalizing to multiple types of signals

Challenges of using human language processing data

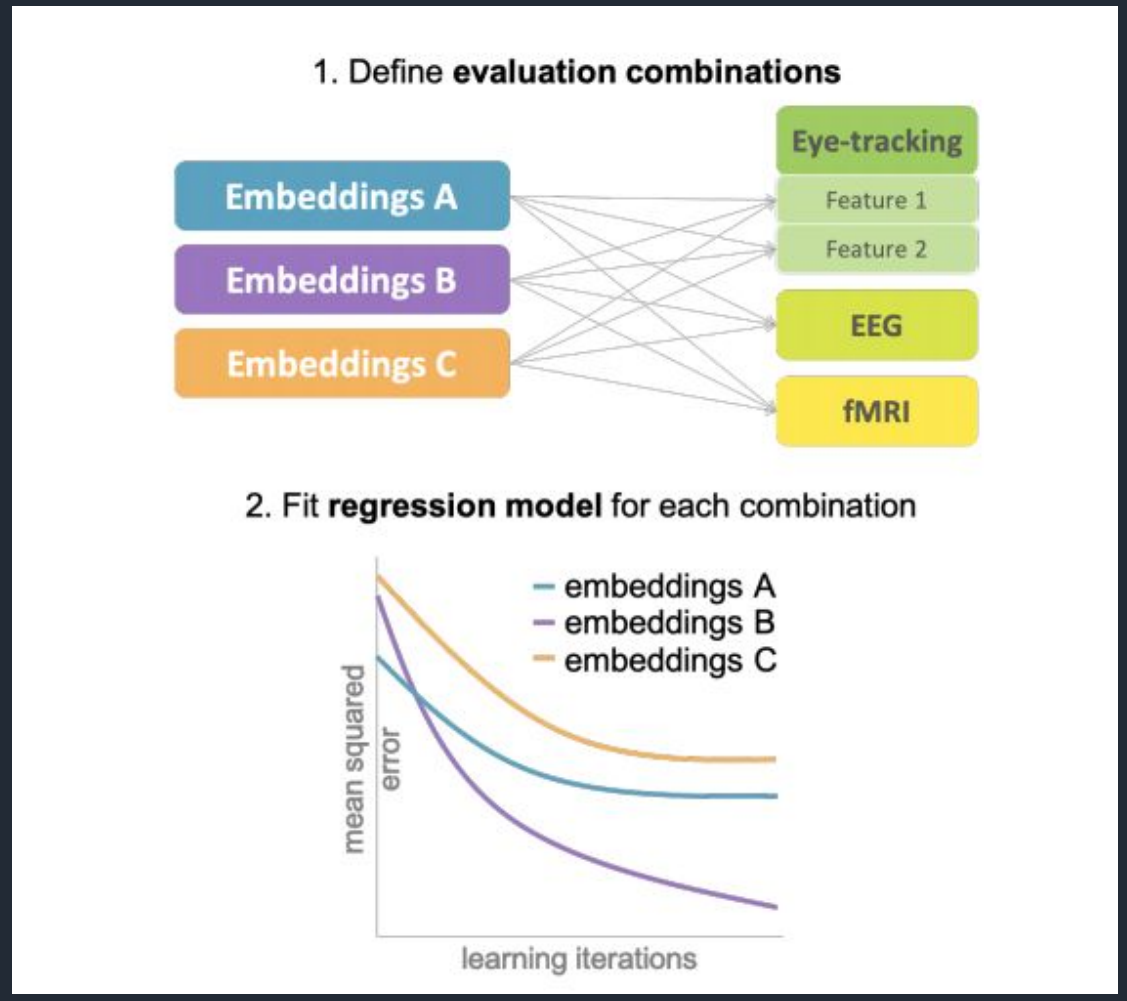
- ↦ lack of data
- ↦ heterogeneity of data
- ↦ mental lexical representation
- ↦ variance between subjects

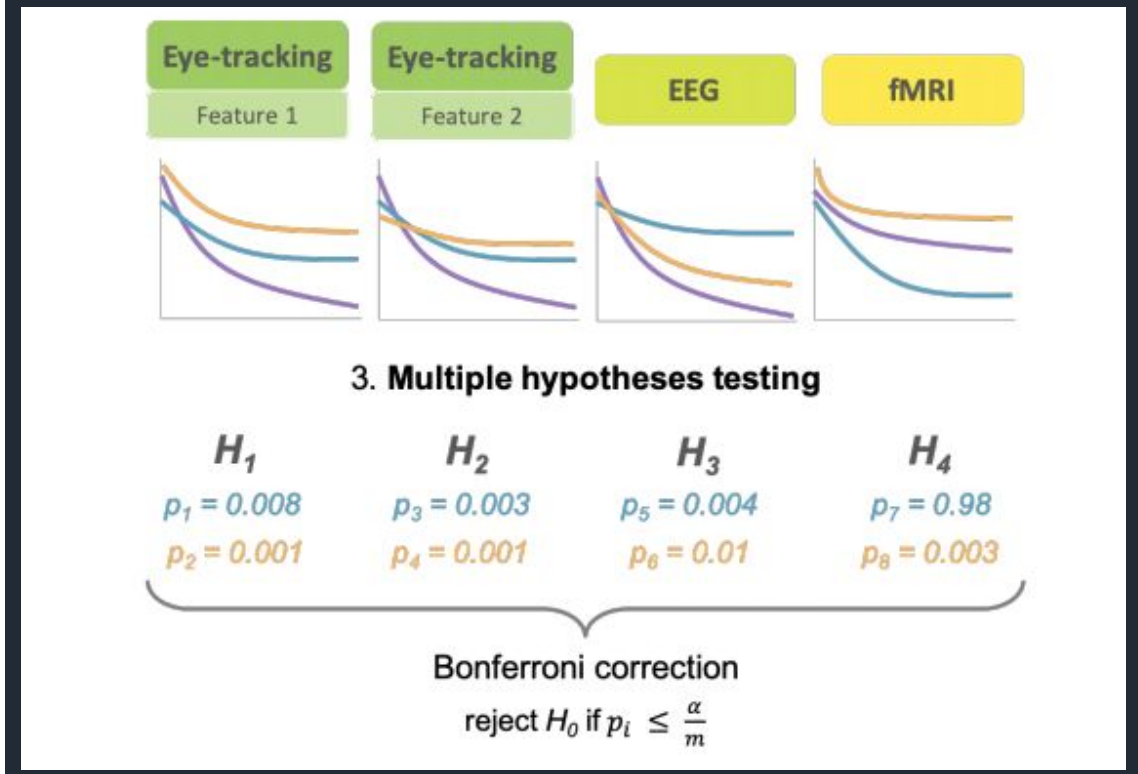
Generalizing to multiple types of signals



Hollenstein et al. (2019)

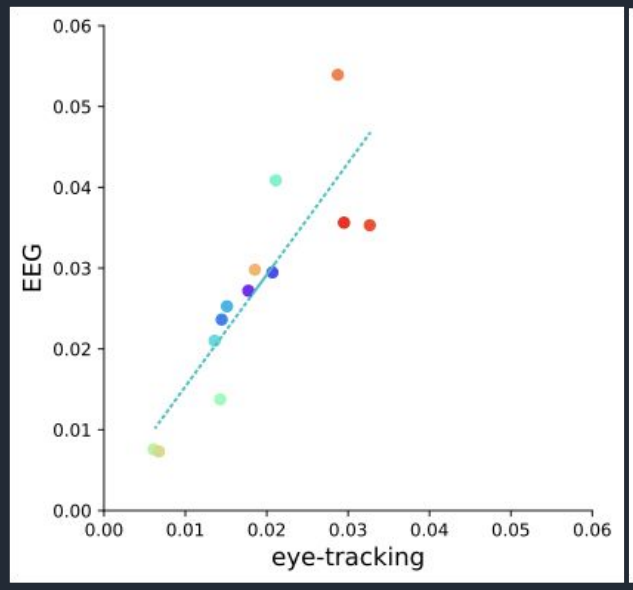
Cognitive evaluation framework

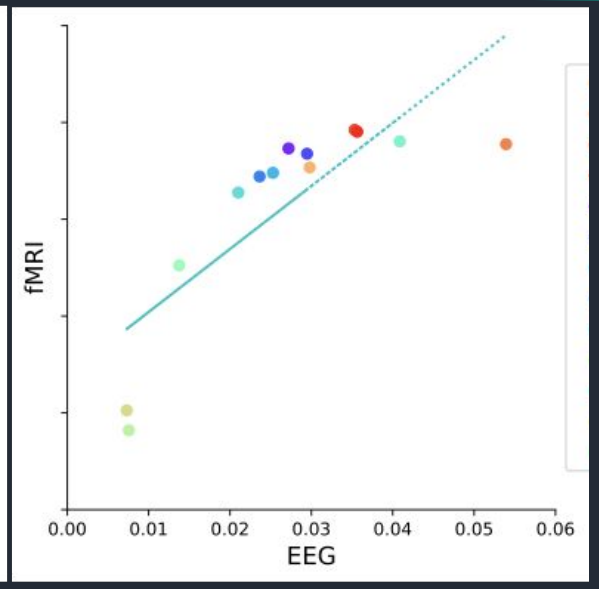


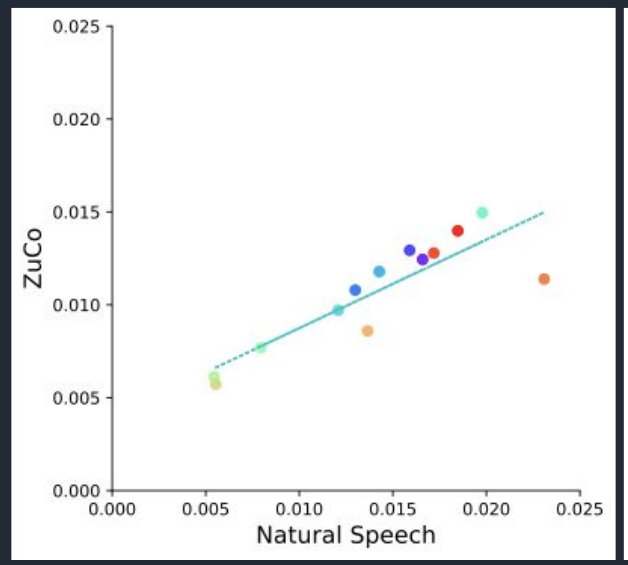


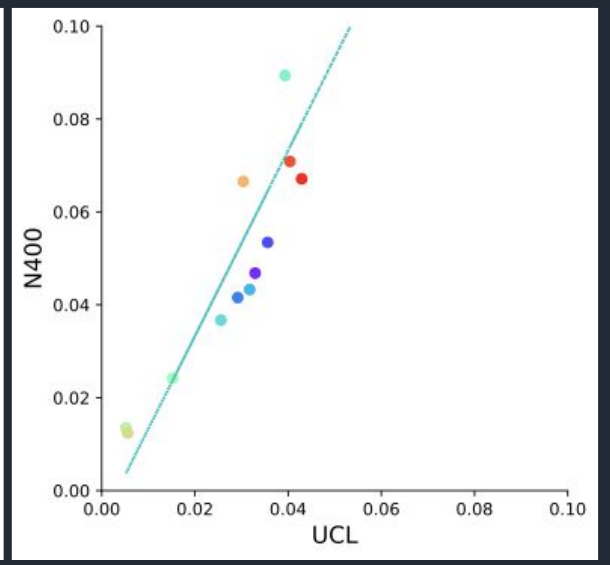
Hollenstein et al. (2019)

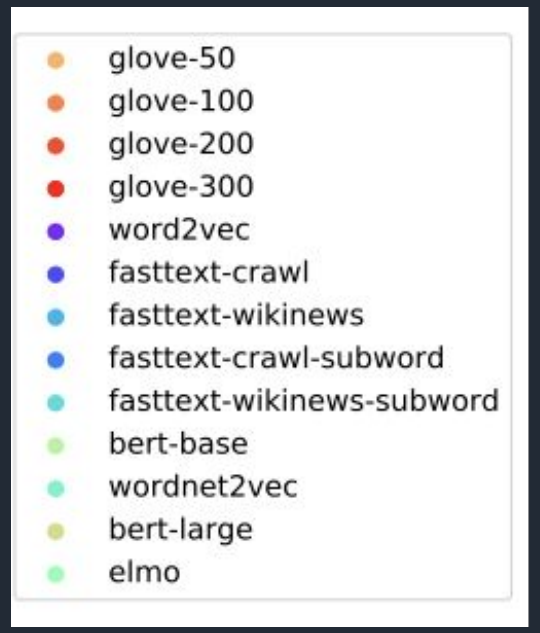
Correlation across & within modalities





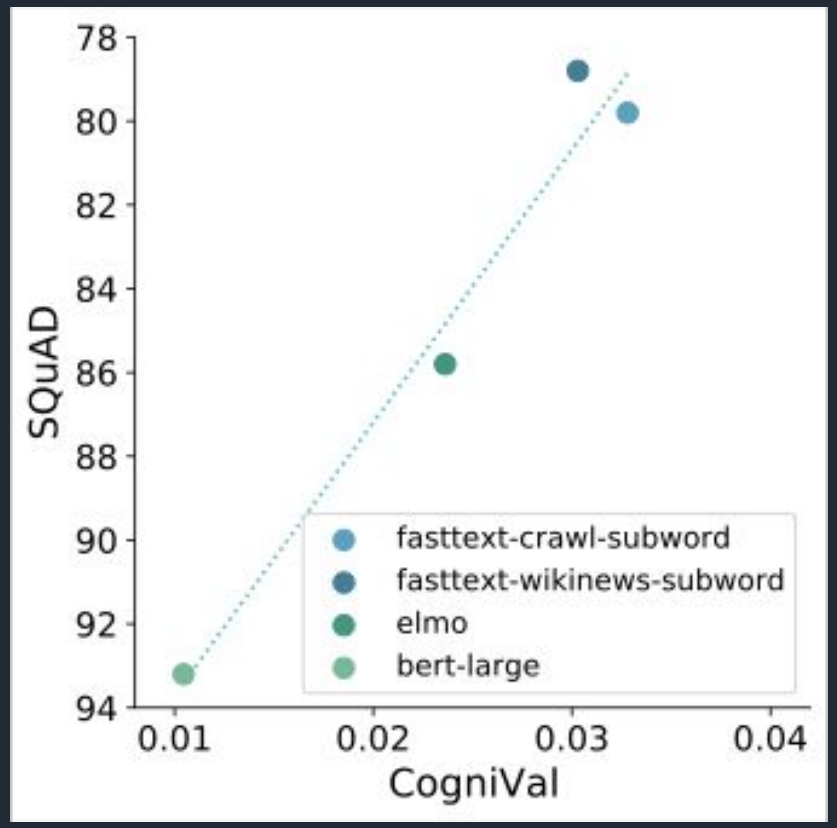


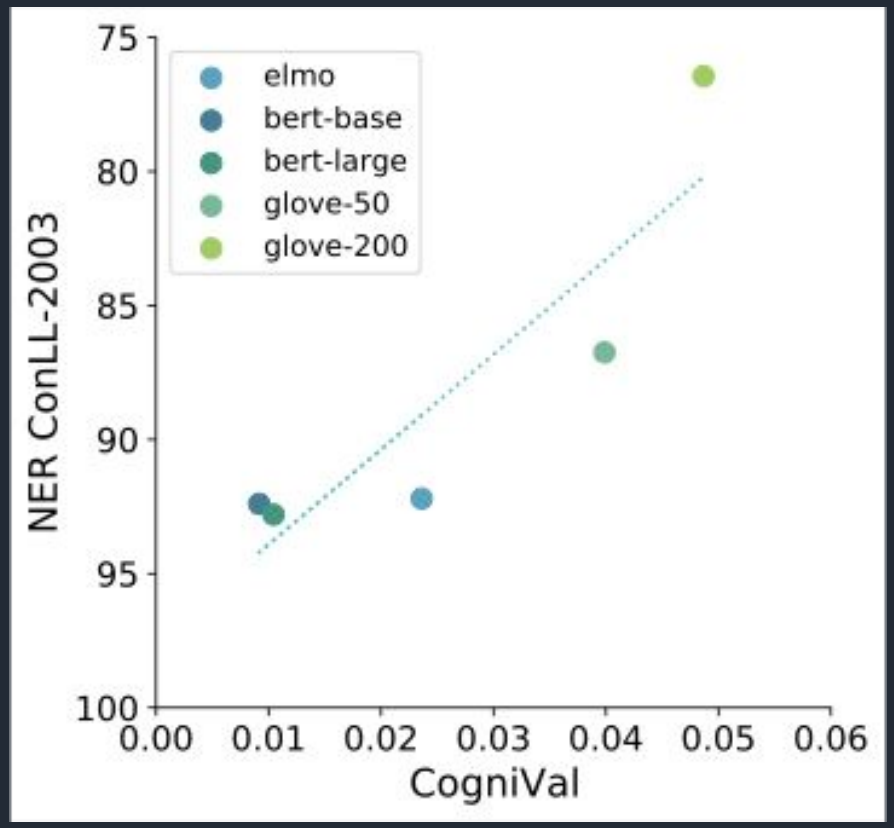




Hollenstein et al. (2019)

Correlation with extrinsic results





Hollenstein et al. (2019)

Probing “human-like” linguistic capabilities of LMs

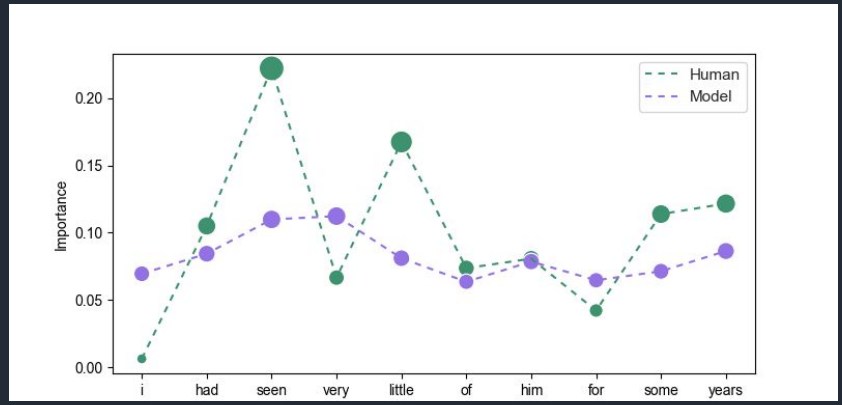
What BERT is not: Diagnostic tests for language models based on psycholinguistic experiments

Context	$\text{BERT}_{\text{LARGE}}$ predictions
A robin is a ___	*bird, robin, person, hunter, pigeon*
A daisy is a ___	*daisy, rose, flower, berry, tree*
A hammer is a ___	*hammer, tool, weapon, nail, device*
A hammer is an ___	*object, instrument, axe, implement, explosive*
A robin is not a ___	*robin, bird, penguin, man, fly*
A daisy is not a ___	*daisy, rose, flower, lily, cherry*
A hammer is not a ___	*hammer, weapon, tool, gun, rock*
A hammer is not an ___	*object, instrument, axe, animal, artifact*

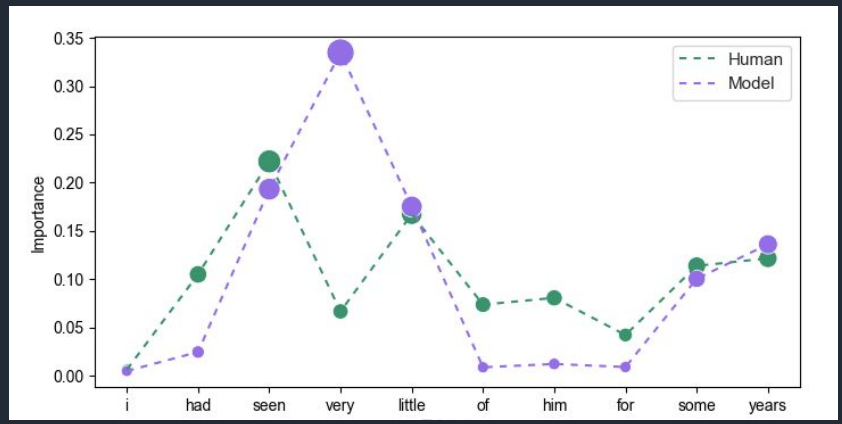
Table 13: $\text{BERT}_{\text{LARGE}}$ top word predictions for selected NEG-136-SIMP sentences.

Ettinger (2020)

Attention

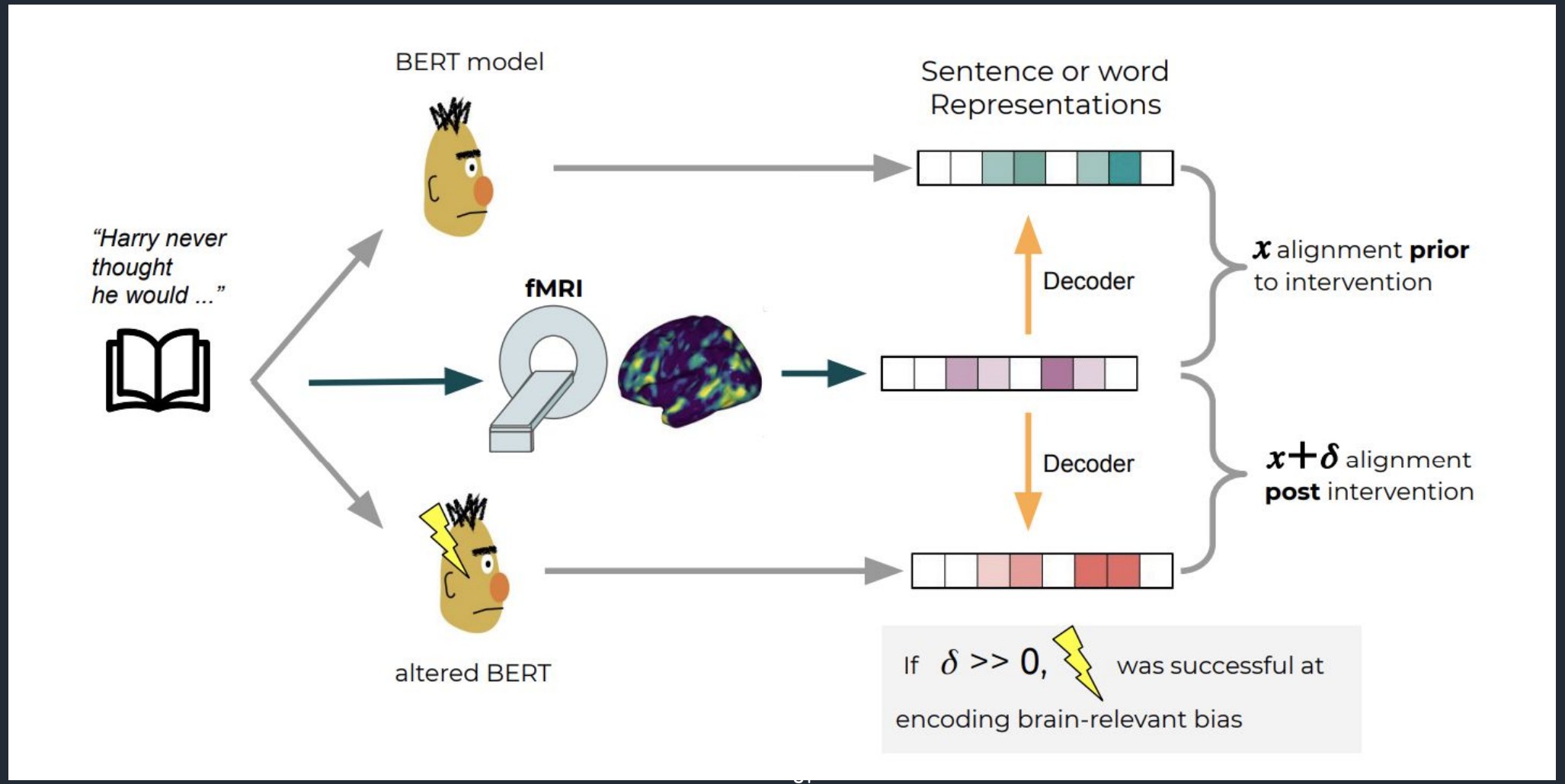


Saliency



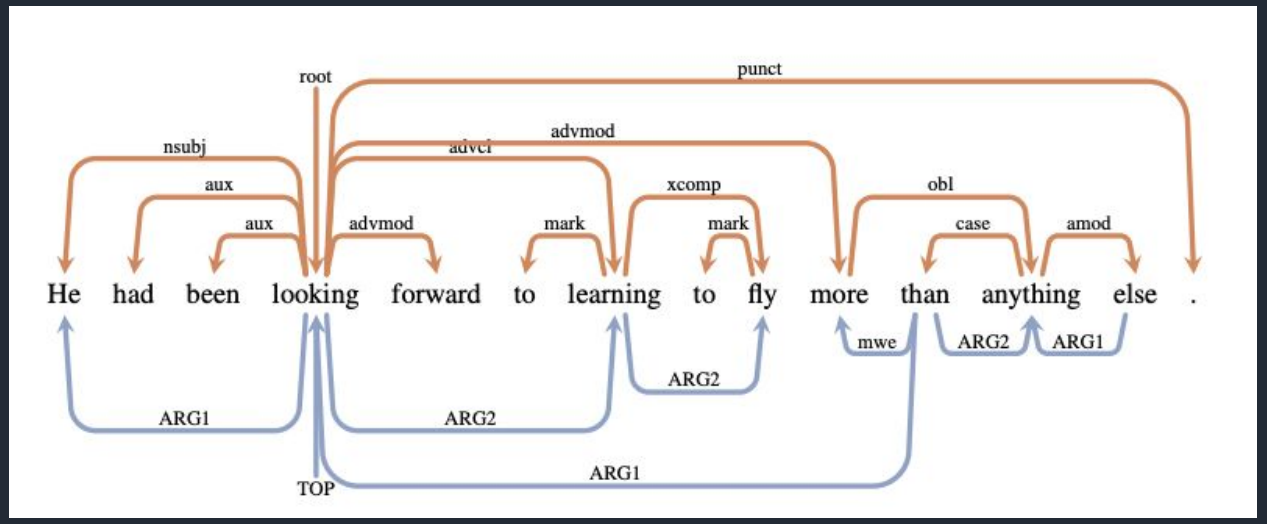
Hollenstein & Beinborn (2021)

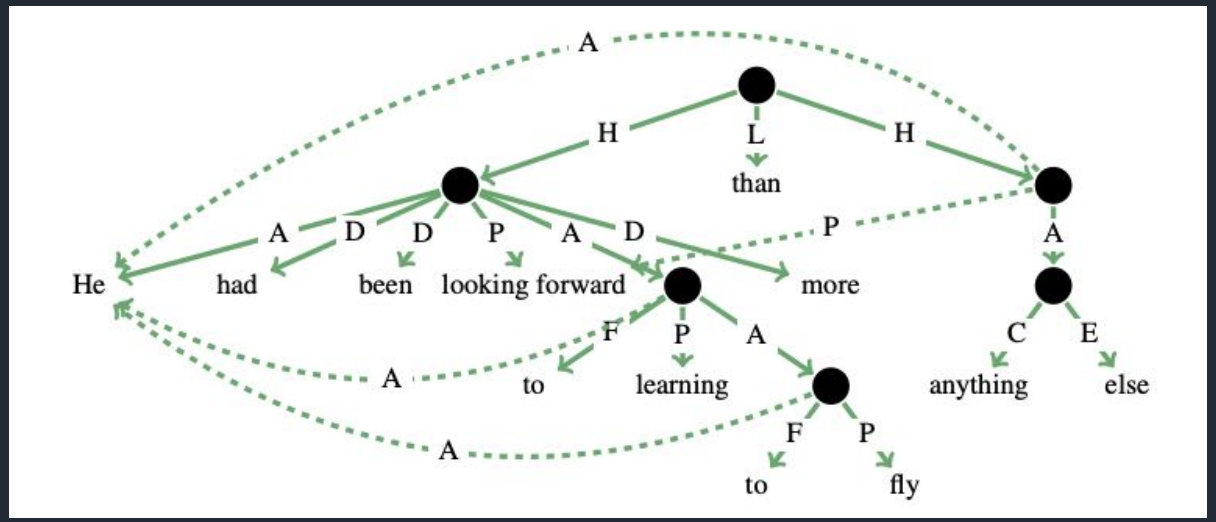
Modifying LMs for better alignment with brain recordings



Abdou et al. (2021)

Modifying LMs for better alignment with brain recordings





Cognitive data sources: 2 fMRI datasets

Syntax/semantic formalisms to inject linguistic structure to BERT embeddings - guide BERT's attention according to the annotations of these formalisms

Results: Improved alignments for both datasets - the models that aligned most with the brain performed best at a range of subject-verb agreement syntactic tasks

Abdou et al. (2021)

Modifying LMs for better performance on NLP tasks

Inducing human bias into NLP models:

- Using eye movement features to improve semantic and syntactic text analysis models

NLP task	Earliest reference
Part-of-speech tagging	Barrett et al. (2016a)
Sentiment analysis	Mishra et al. (2017b)
Named entity recognition	Hollenstein & Zhang (2019)
Relation detection	Hollenstein et al. (2019a)
Sarcasm detection	Mishra et al. (2016)
Multiword expressions	Rohanian et al. (2017)
Referential/non-referential *it*	Yaneva et al. (2018)
Coreference resolution	Cheri et al. (2016)
Sentence compression	Klerke et al. (2016)
Predicting misreadings	Bingel et al. (2018)
Predicting native language	Berzak et al. (2017)
Predicting language proficiency	Kunze et al. (2013)
Dependency parsing	Strzyz et al. (2019)
Text summarization	Xu et al. (2009)

Overview of NLP tasks which have been enhanced with eye tracking features (Hollenstein et al. 2020)

Modifying LMs for better performance on NLP tasks

Inducing human bias into NLP models:

- Using brain activity to fine-tune language models

Fine-tuning BERT with two types of brain activity improves performance on many NLP tasks (Schwartz et al. 2019).

Metric	Vanilla	MEG	Joint
CoLA	57.29	57.63	**57.97**
SST-2	93.00	**93.23**	91.62
MRPC (Acc.)	83.82	83.97	**84.04**
MRPC (F1)	88.85	**88.93**	88.91
STS-B (Pears.)	**89.70**	89.32	88.60
STS-B (Spear.)	**89.37**	88.87	88.23
QQP (Acc.)	90.72	**91.06**	90.87
QQP (F1)	87.41	**87.91**	87.69
MNLI-m	83.95	**84.26**	84.08
MNLI-mm	84.39	84.65	**85.15**
QNLI	89.04	**91.73**	91.49
RTE	61.01	**65.42**	62.02
WNLI	53.52	**53.80**	51.97

Tutorial: Predicting eye-tracking features

Tutorial: Prerequisites

- Code: https://github.com/beinborn/ESSLLI2021/code/tutorial1
- Jupyter Lab or Google Colab
- Python version >3.7

To use Google Colab change URL of the Jupyter Notebook to:
https://colab.research.google.com/github/beinborn/ESSLLI2021/blob/main/code/tutorial1/esslli_tutorial1.ipynb

Questions?

Resources & Readings

- Linguistic effects in eye-tracking:
  Clifton Jr, Charles, Adrian Staub, and Keith Rayner. "Eye movements in reading words and sentences." *Eye movements* (2007).

- Collection of cognitive data sources useful for NLP:
  https://github.com/norahollenstein/cognitiveNLP-dataCollection

- *Nilearn*: Machine learning for neuroimaging in Python
  https://nilearn.github.io/

- Semantic maps in the human brain: https://gallantlab.org/huth2016/

- Keller, Frank. "Cognitively plausible models of human language processing." *Proceedings of the ACL 2010 Conference*. 2010.

References

- Abdou, Mostafa, et al. "Does injecting linguistic structure into language models lead to better alignment with brain recordings?." *arXiv preprint arXiv:2101.12608*. 2021.

- Abnar, Samira, et al. "Experiential, Distributional and Dependency-based Word Embeddings have Complementary Roles in Decoding Brain Activity." *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics* (CMCL). 2018.

- Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. "Most people are not WEIRD." *Nature*. 2010.

- Hollenstein, Nora, et al. "CogniVal: A Framework for Cognitive Word Embedding Evaluation." *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. 2019.

- Mitchell, Tom M., et al. "Predicting human brain activity associated with the meanings of nouns." *Science*. 2008.