Convergent Representations of Computer Programs in Human and Artificial Neural Networks









Shashank Srikant* Ben Lipkin* Anna Ivanova Evelina Fedorenko Una-May O'Reilly

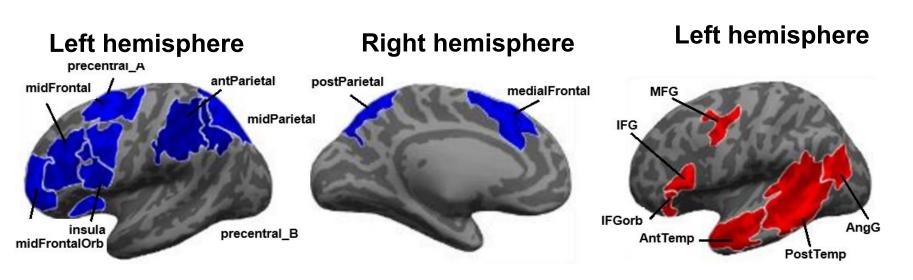








Brains ∩ Code comprehension



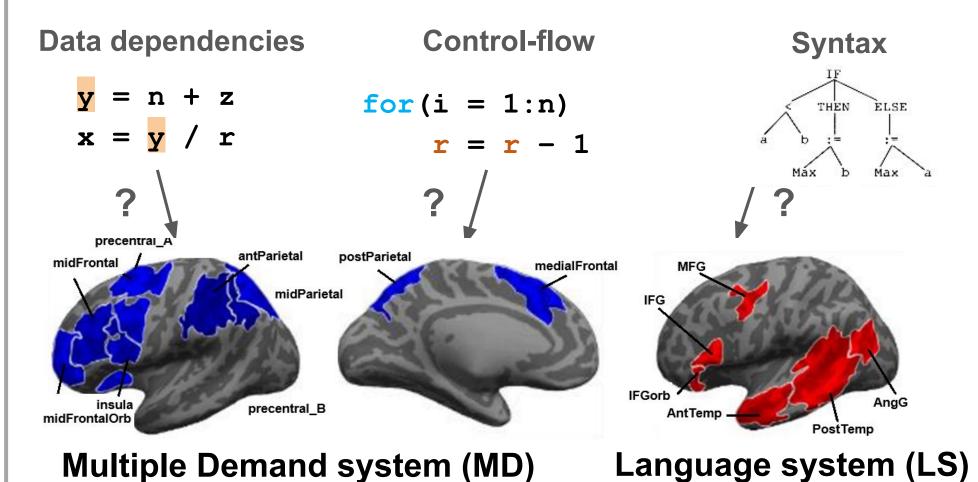
Multiple Demand system (MD)

Language system (LS)

From Ivanova et al. 2020:

MD consistently activated during code comprehension; LS inconsistent

Is it possible?



Programs from our stimuli set

height = 5 weight = 100 bmi = weight / (height*height) print(bmi)

code

Your height is 5 feet and your weight is 100 pounds. The BMI is defined as the ratio between the weight and the square of the height

of a person. What is your BMI?

sentence

nums = [10, 2, 30] prod = 1 for n in nums: prod = prod*n

big_num, small_num = 64, 16
if big_num % small_num == 0:
 print(1)
else:

string1 = "onion" new_string = string1[-2:] multiplied = new_string*4

print(multiplied)

for , math

print(prod)

if , math

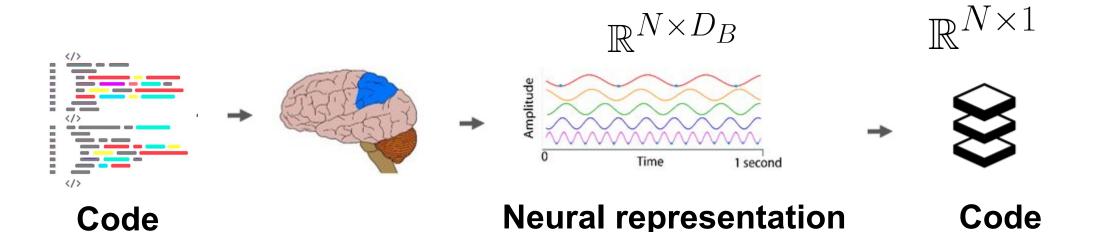
print(0)

seq , str

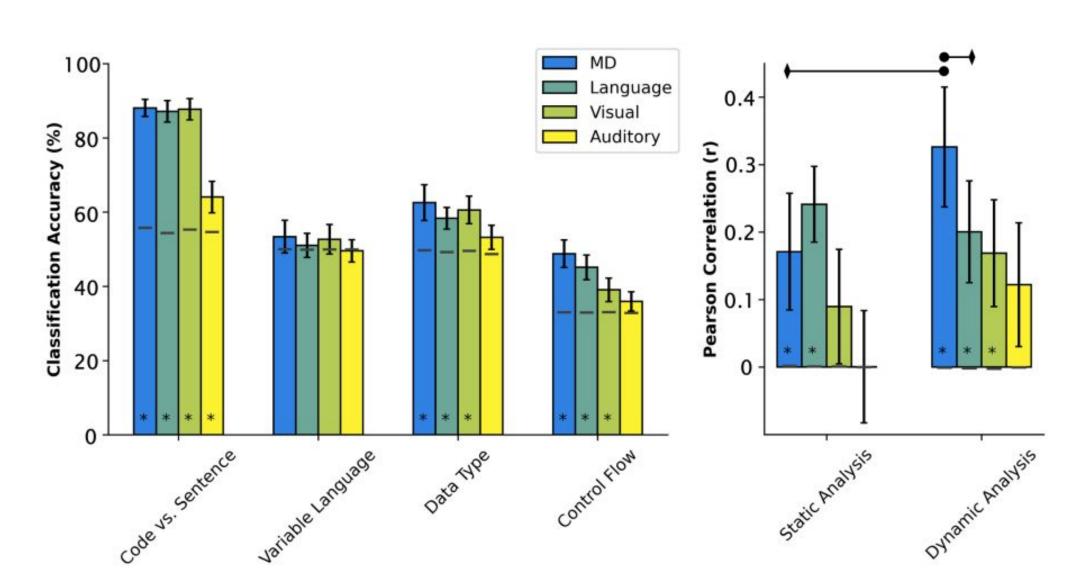
What code properties do brain representations encode?

of code

properties



Results

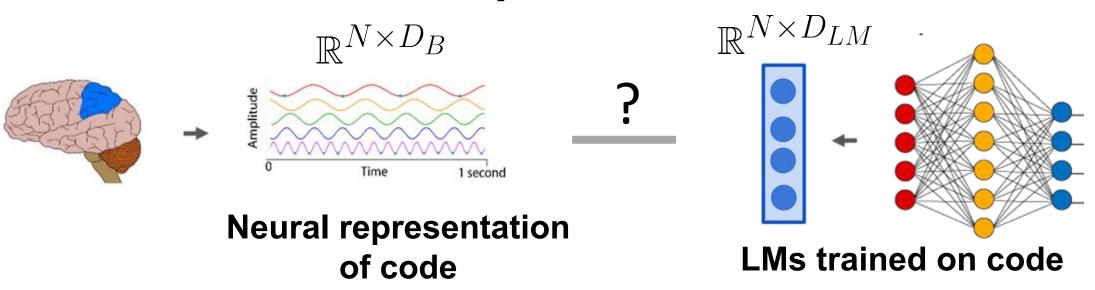


	Control Flow	Data type	Static Analysis	Dynamic Analysis
MD > null	+	+	+	+
LS > null	+	+	+	+

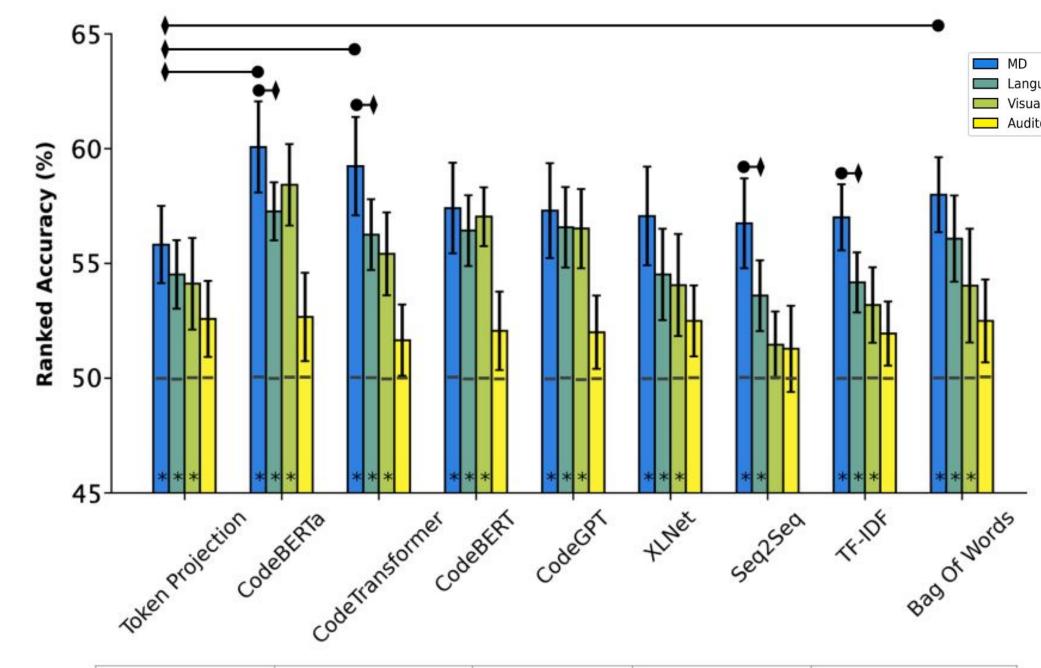
	Control Flow	Data type	Static Analysis	Dynamic Analysis
MD > LS	•		•	+

Are there more properties which are encoded in these brain systems?

How similar are machine representations and brain representations?



Results



	Token Projection	TF-IDF	seq2seq	CodeBERTa
MD > LS		+	+	+

Picture so far

code, contact



