LXMLS
LISBON MACHINE LEARNING SCHOOL 2025

The Search for Foundations in AI Interpretability

Maxime Peyrard

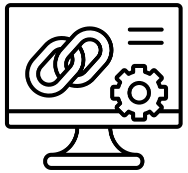
Why Explain?

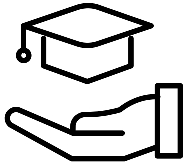
Practical Reasons

Transparency, Trust

Accountability, Legal Compliance

Error Diagnostic, Continuous Improvement, Robustness

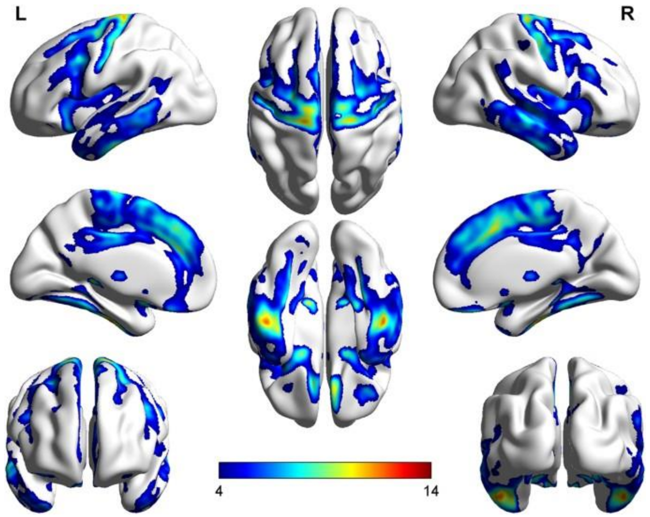




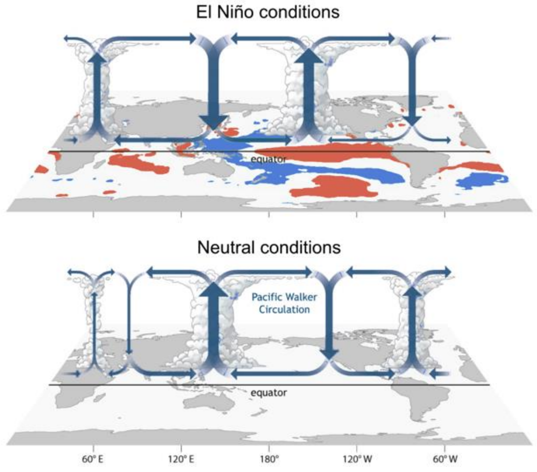
Communication, Education, Human-AI collaboration

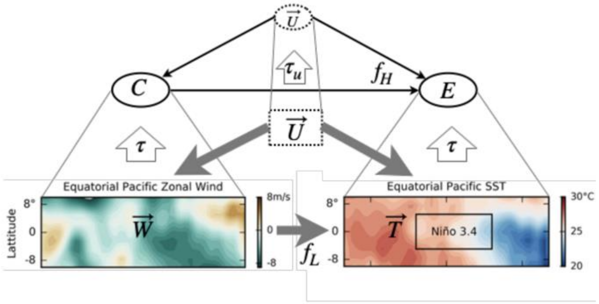
…

Scientific Question: Explaining Complex Systems

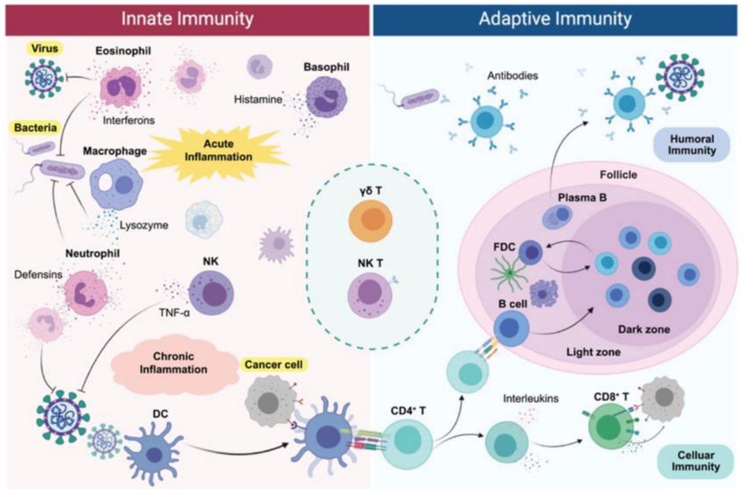


Wei, et al 2018 Nature



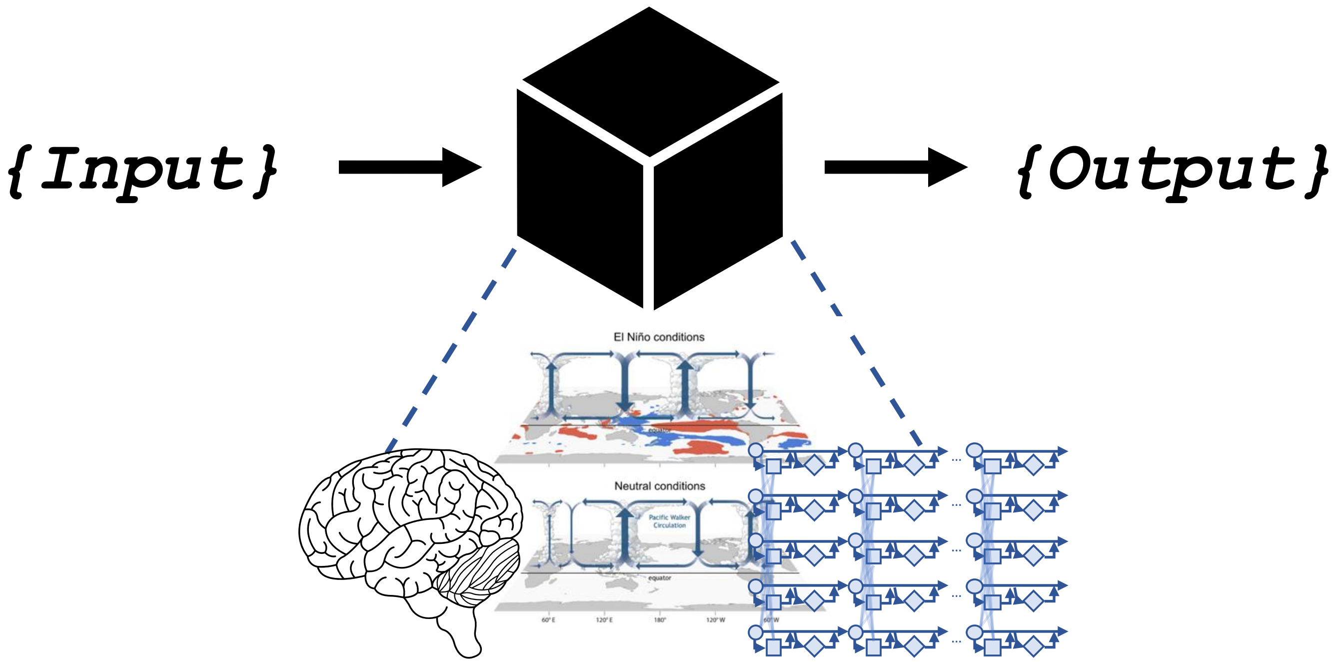


Chalupka, et al 2018 UAI
Beckers, et al, 2019 UAI



Chen, et al 2020 AM

Interpretability research



What ***should be*** explained? What ***is*** an explanation? What is *a* ***valid*** explanation?

*How to go from observations to **valid explanations**?*

AIs are the *Simplest* Complex Systems to Study

Testing the Tools of Systems Neuroscience on Artificial Neural Networks

Grace W. Lindsay

AI are fully observable and manipulable

What does it mean to understand a neural network?

Timothy P. Lillicrap & Konrad P. Kording

Forms of explanation and understanding for neuroscience and artificial intelligence

Jessica A. F. Thompson

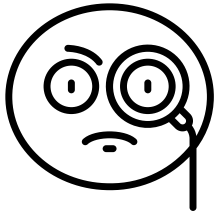
Contributions and challenges for network models in cognitive neuroscience

Olaf Sporns

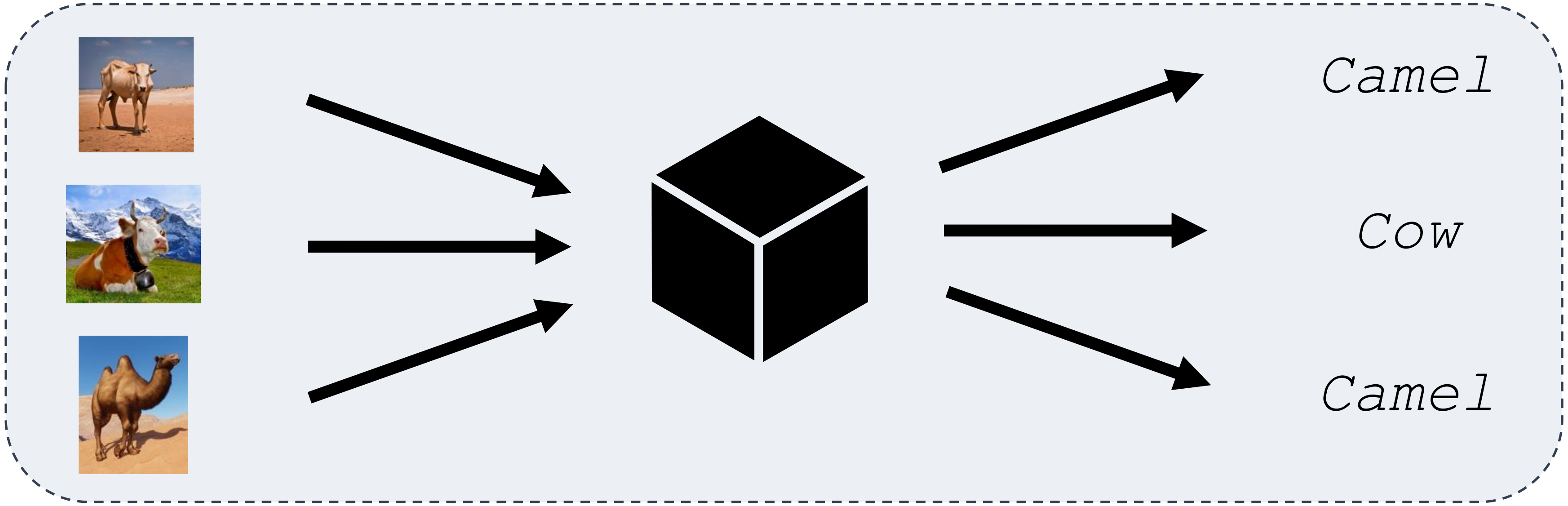
How to Explain?

“Behavioral”

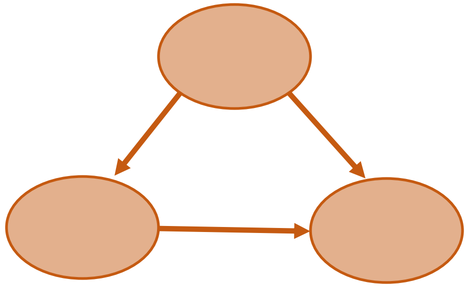
Behavioral Testing



Inputs
outputs
inspection



Carefully craft inputs, measure effects on outputs, come-up with hypothesis



High-level explanation

Benchmarking

GAIA Leaderboard

Results: Test | Results: Validation

Model name	Average score (%)	Level 1 score (%)	Level 2 score (%)	Level 3 score (%)	organisation	Model family
Multi-Agent Experiment v0.1	32.33	47.31	28.93	14.58	MSR AI Frontiers	GPT-4-turbo
MAAC_V1	25.					
FRIDAY	24.					
FRIDAY_without_learning	21.					
DIP	15.					
Chamomile	14.					
GPT4 + manually selected plu	14.					
Clarity_v1	14.					
Warm-up Act	12.					
stealth3	9.3					
stealth2	8.9					
stealth	8.6					

🤗 Open LLM Leaderboard

🏅 LLM Benchmark | 📈 Metrics through time | 📝 About | ❗ FAQ | 🚀 Submit

🔍 Search for your model (separate multiple queries wi

Select columns to show

- [x] Average ⬆️
- [x] ARC
- [x] HellaSwag
- [x] Winogrande
- [x] GSM8K
- [ ] Type
- [ ] Merged
- [ ] Hub License
- [ ] #Params

Hide models

- [x] Private or deleted
- [x] Contains a merge/mo

T	Model
🔶	davidkim205/Rhea-72b-v0.5 📄
💬	Contamination/contaminated_proof_
💬	MTSAIR/MultiVerse_70B 📄
🔶	MTSAIR/MultiVerse_70B 📄
🔶	SF-Foundation/Ein-72B-v0.11 📄

🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **500,000** human preference votes to rank LLMs with the Elo ranking system.

Arena Elo | Full Leaderboard

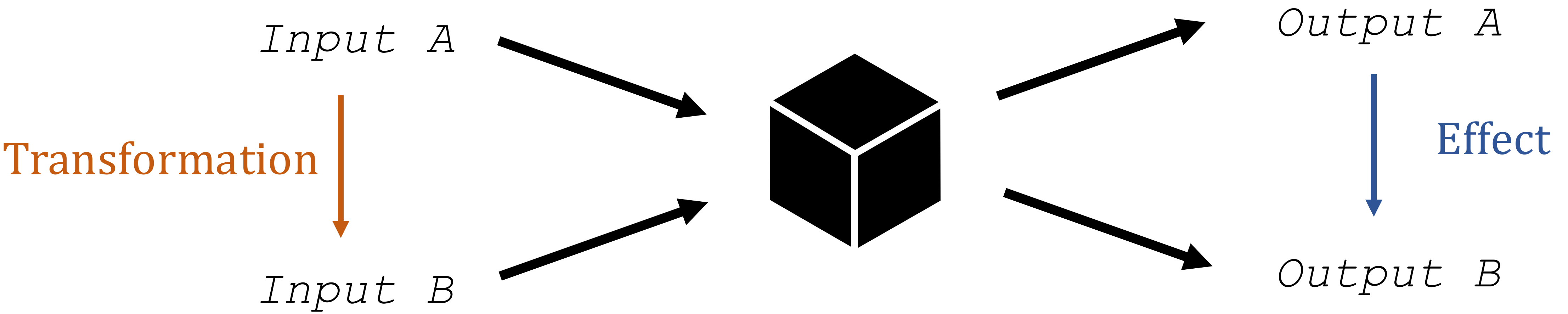
Total #models: **76**. Total #votes: **511252**. Last updated: March 29, 2024.

Contribute your vote 🗳️ at chat.lmsys.org! Find more analysis in the notebook.

Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
1	Claude 3 Opus	1255	+3/-4	37663	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1252	+3/-3	56936	OpenAI	Proprietary	2023/4
1	GPT-4-0125-preview	1249	+3/-4	38105	OpenAI	Proprietary	2023/12
4	Bard (Gemini Pro)	1204	+5/-5	12468	Google	Proprietary	Online
4	Claude 3 Sonnet	1200	+3/-4	40389	Anthropic	Proprietary	2023/8
6	GPT-4-0314	1185	+4/-4	35803	OpenAI	Proprietary	2021/9
7	Claude 3 Haiku	1177	+3/-4	26773	Anthropic	Proprietary	2023/8
8	GPT-4-0613	1160	+3/-5	54509	OpenAI	Proprietary	2021/9

Controlled Setups

Idea: controlled transformation of the inputs, and measure effects on outputs



Evaluating Models' Local Decision Boundaries via Contrast Sets
EMNLP 2020

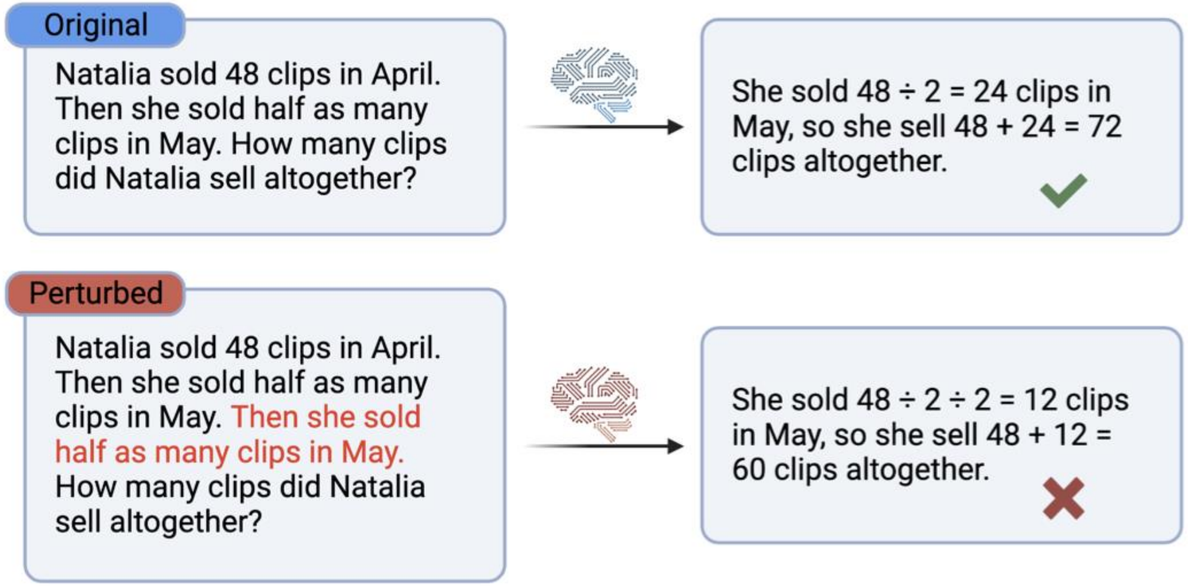
Learning What Makes a Difference from Counterfactual Examples
ECCV 2021

Controlled Setups - Examples

Noisy Exemplars Make Large Language Models More Robust: A Domain-Agnostic Behavioral Analysis

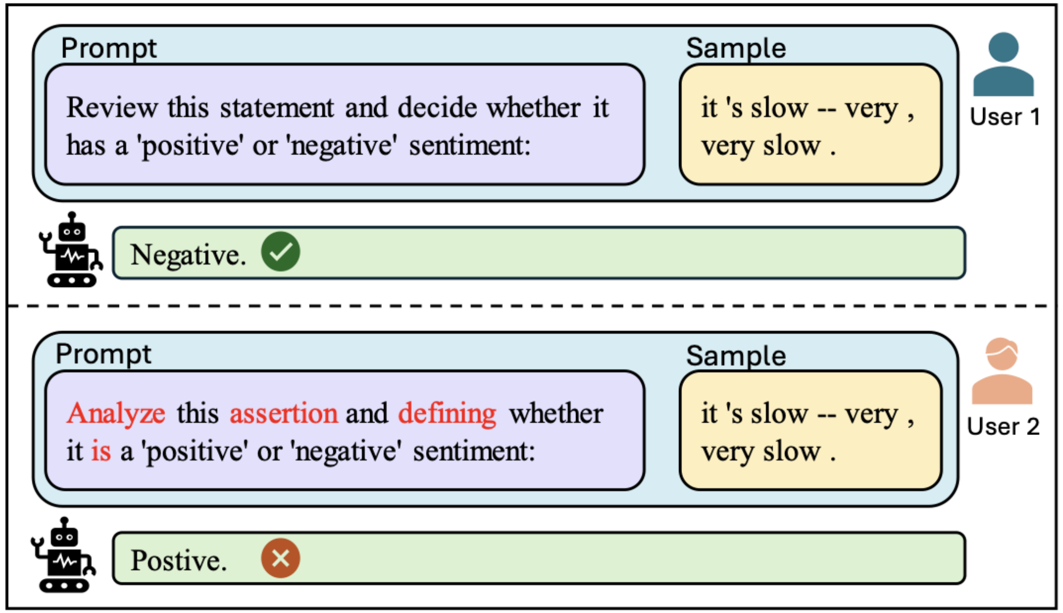
Hongyi Zheng
New York University
hz2212@nyu.edu

Abulhair Saparov
New York University
as17582@nyu.edu



PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts

Kaijie Zhu[1,2]*, Jindong Wang[1]†, Jiaheng Zhou[2], Zeek Wang[1], Hao Chen[3], Yidong Wang[4], Linyi Yang[5], Wei Ye[4], Yue Zhang[5], Neil Zhenqiang Gong[6], Xing Xie[1]
[1]Microsoft Research [2]Institute of Automation, CAS [3]Carnegie Mellon University
[4]Peking University [5]Westlake University [6]Duke University



(b) Synonyms lead to errors in sentiment analysis problems.

Input Feature Attributions

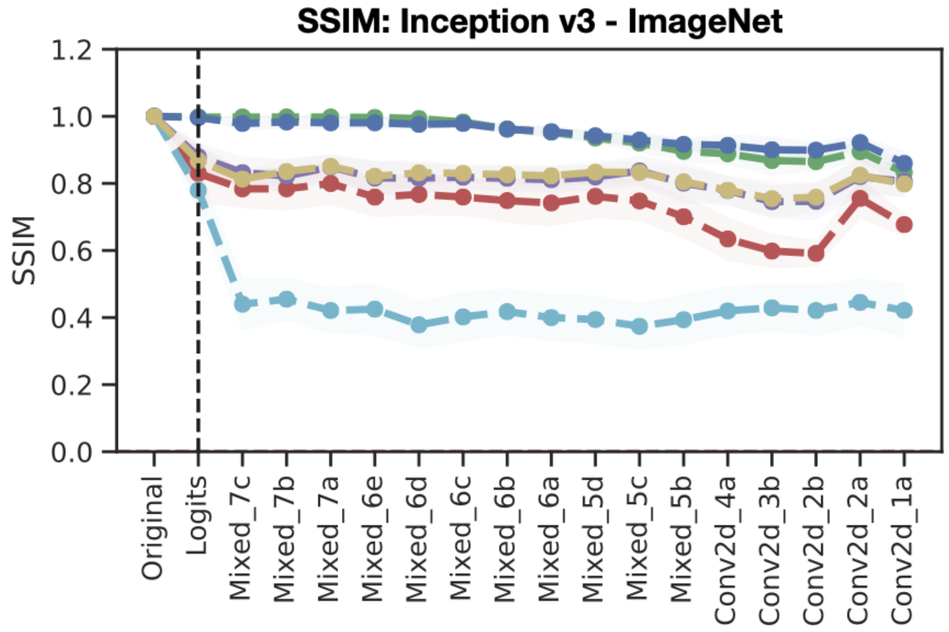
controlled changes on the input features → effects on output

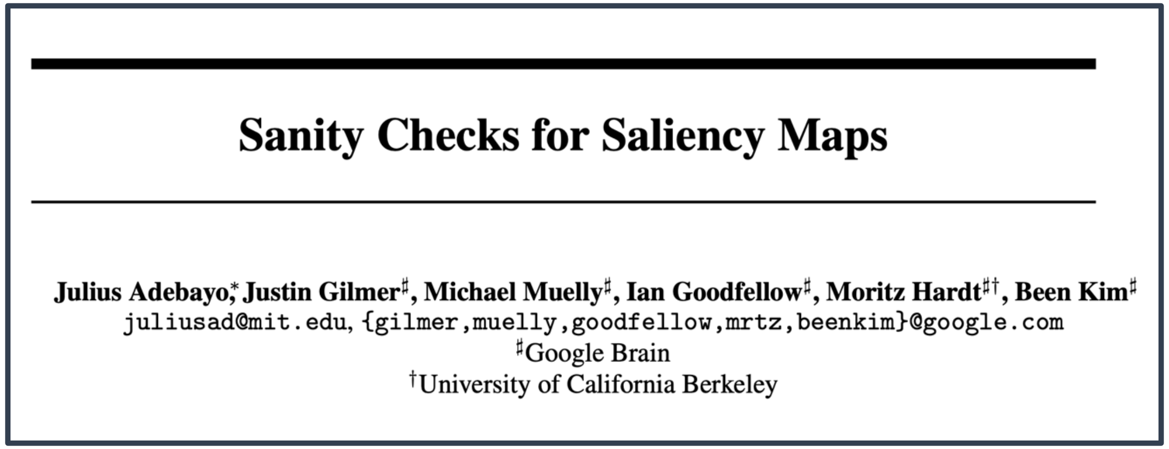
LIME: local approximation of the boundary around an input

SHAP: measure each feature contribution relative to others

Integrated Gradient: use the gradient information backpropagated in the input features

Problem with Feature Attributions





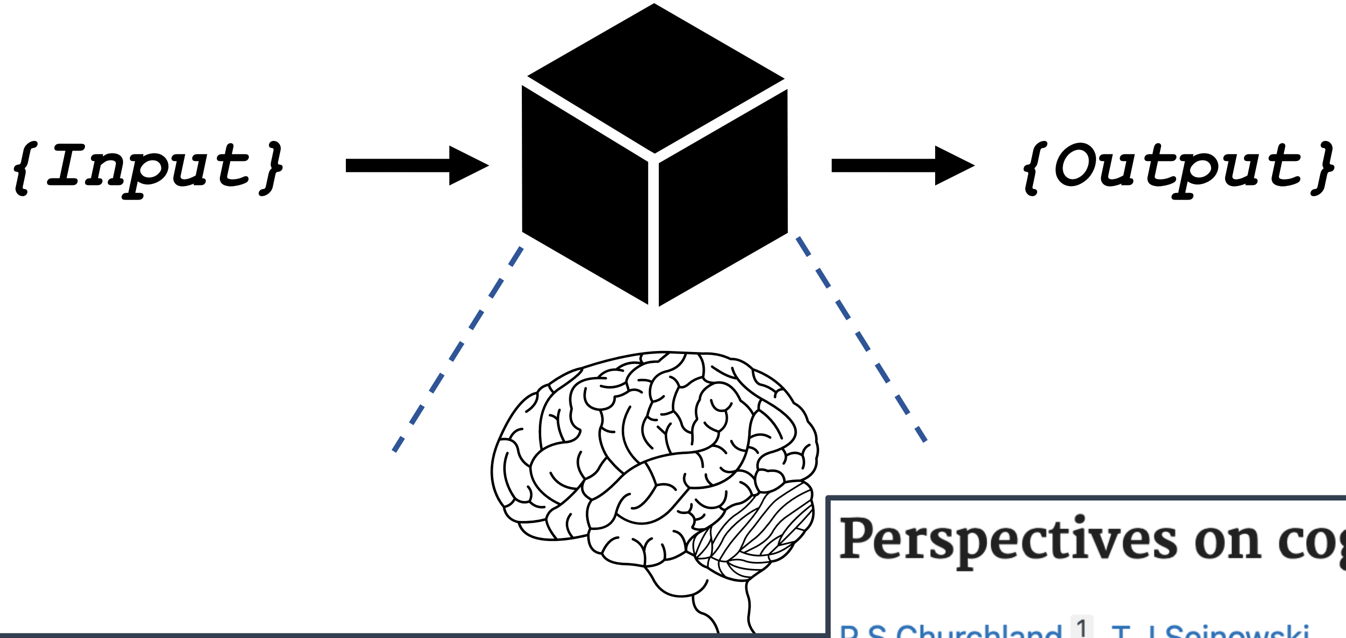
Sanity Checks for Saliency Maps

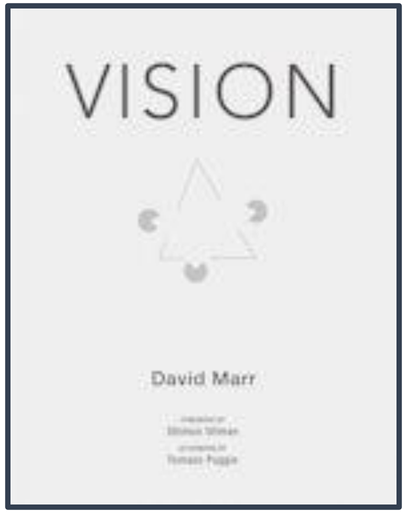
Julius Adebayo*, Justin Gilmer♯, Michael Muelly♯, Ian Goodfellow♯, Moritz Hardt♯†, Been Kim♯
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com
♯Google Brain
†University of California Berkeley

Similar feature attributions for randomly initialized networks compared to trained ones

(

Neuroscience detour I





Review Article | Published: 12 June 2008

What we can do and what we cannot do with fMRI

Nikos K. Logothetis

Nature **453**, 869–878 (2008) | Cite

Perspectives on cognitive neuroscience

P S Churchland[1], T J Sejnowski

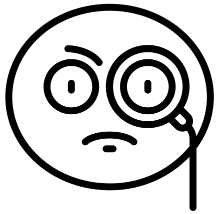
Neural representation and the cortical code

R C deCharms[1], A Zador

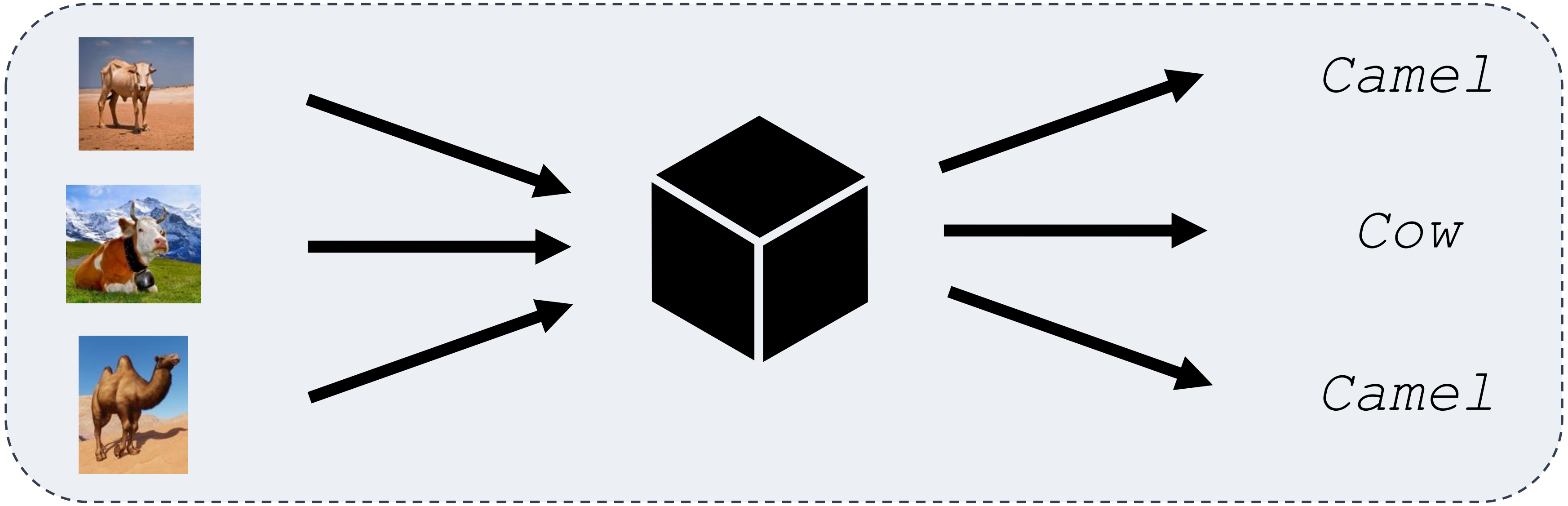
Behavior is not enough; we have to look at the computation to find objective, measurable, and generalizable predictors

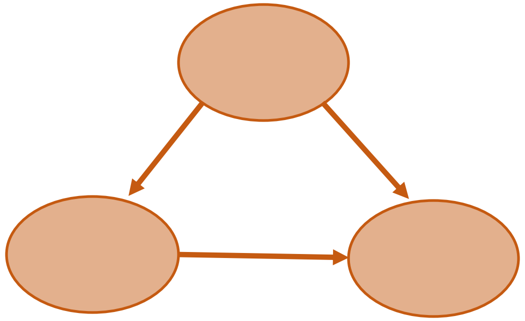
)

Behavior vs Computation

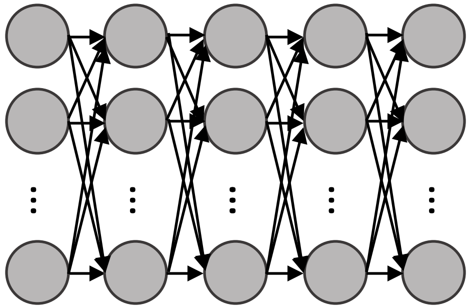


Inputs outputs inspection





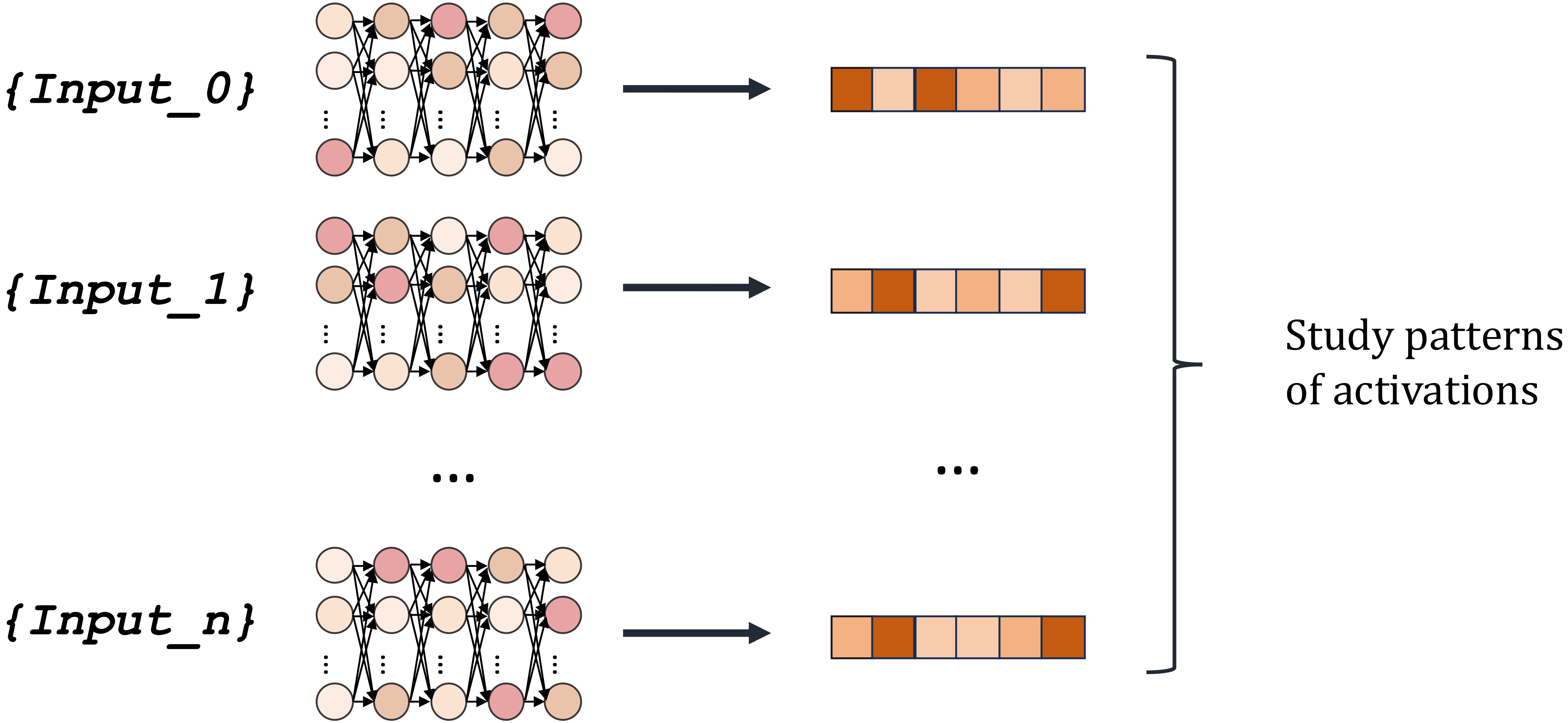
High-level explanation

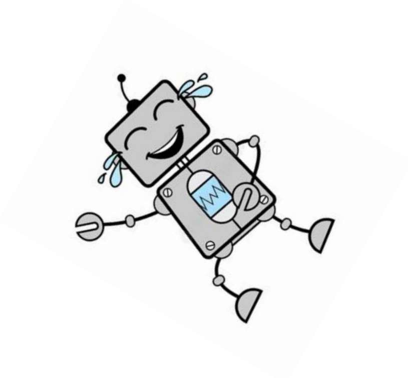


But is it consistent with the low-level implementation?

How to Explain?
“Neural Correlates”

“Neural Correlates”





Laughing Heads: Can Transformers Detect What Makes a Sentence Funny?

Maxime Peyrard, **Beatriz Borges**, **Kristina Gligorić** and **Robert West**
EPFL

Satirical

`City opens new art` `{jail}`

Non-modified chunk in funny

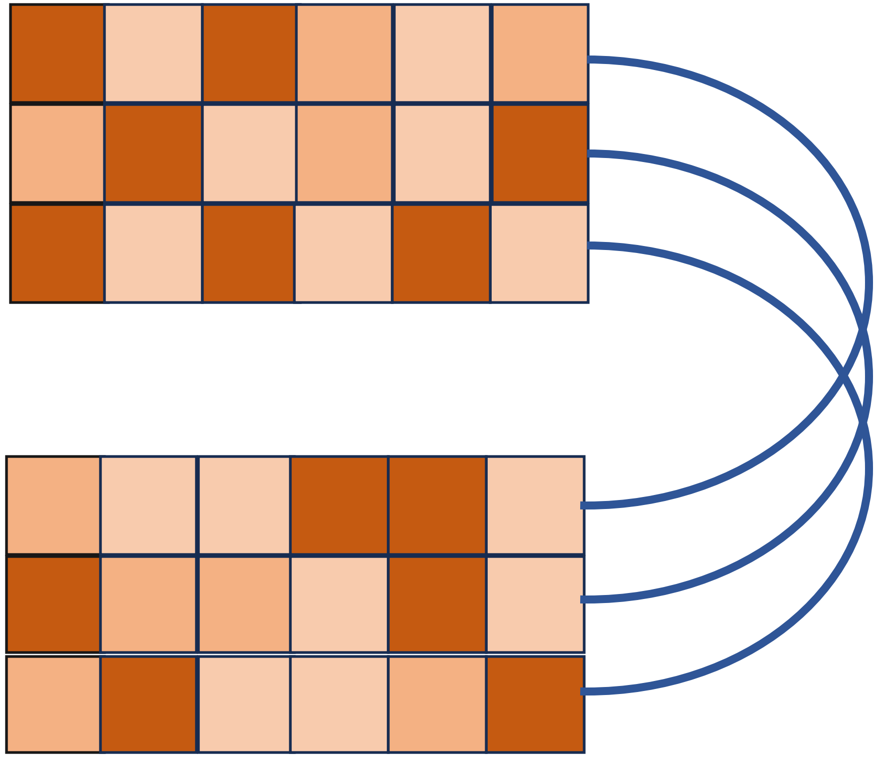
Modified chunk in funny

Serious

`City opens new art` `{museum}`

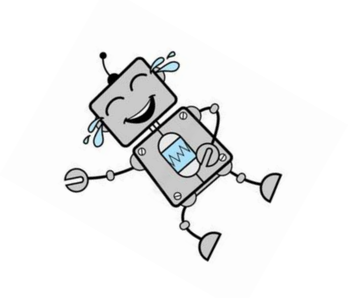
Non-modified chunk in serious

Modified chunk in serious

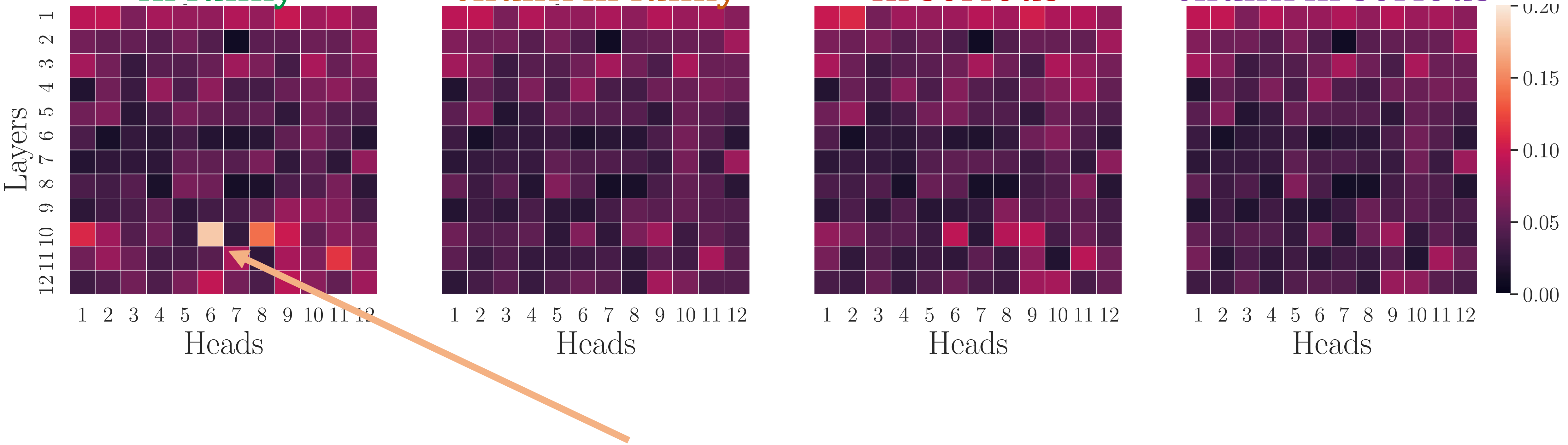


Matched comparison

Laughing Heads



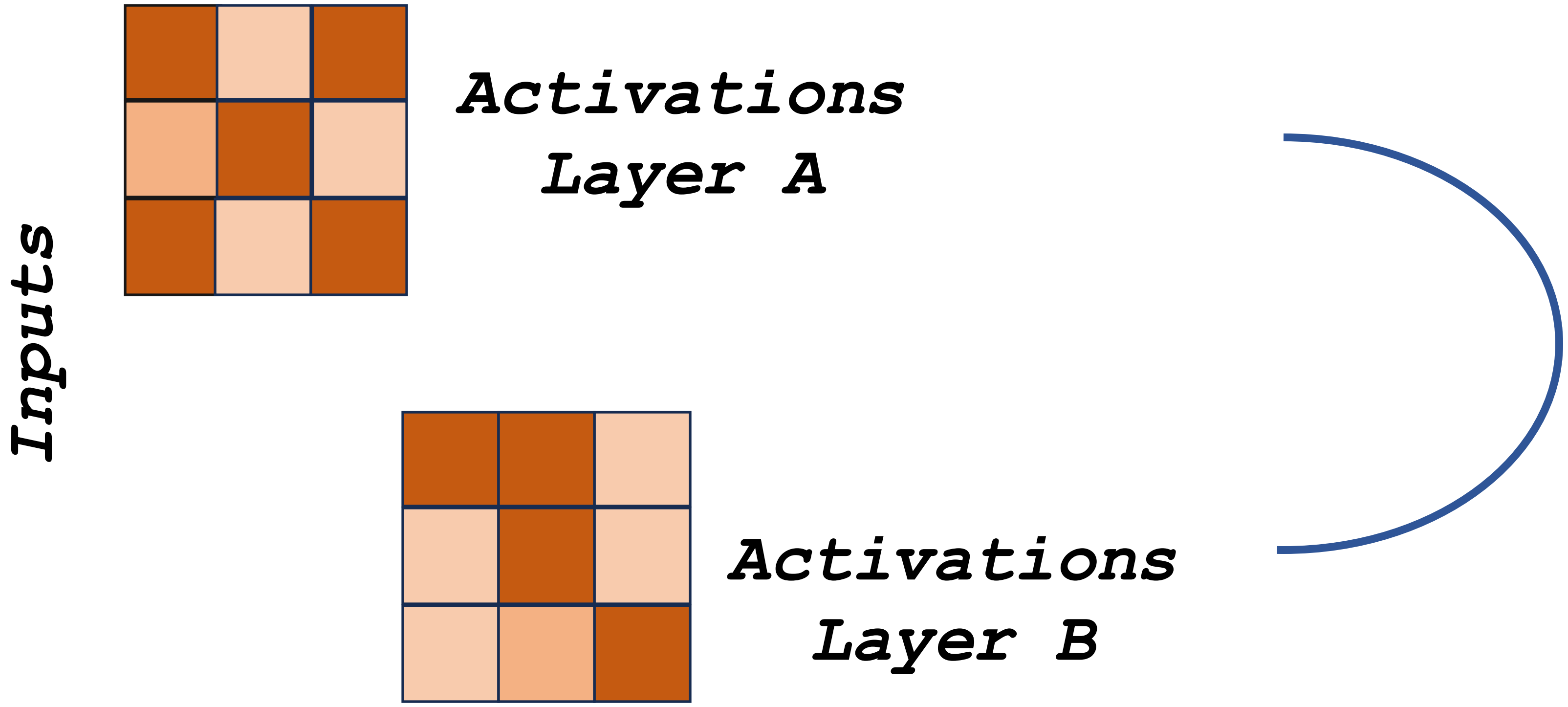
Modified chunk in funny | **Non-modified chunk in funny** | **Modified chunk in serious** | **Non-modified chunk in serious**



Surprisingly, one head attends a lot to modified chunk in funny sentence and only in this case

SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability

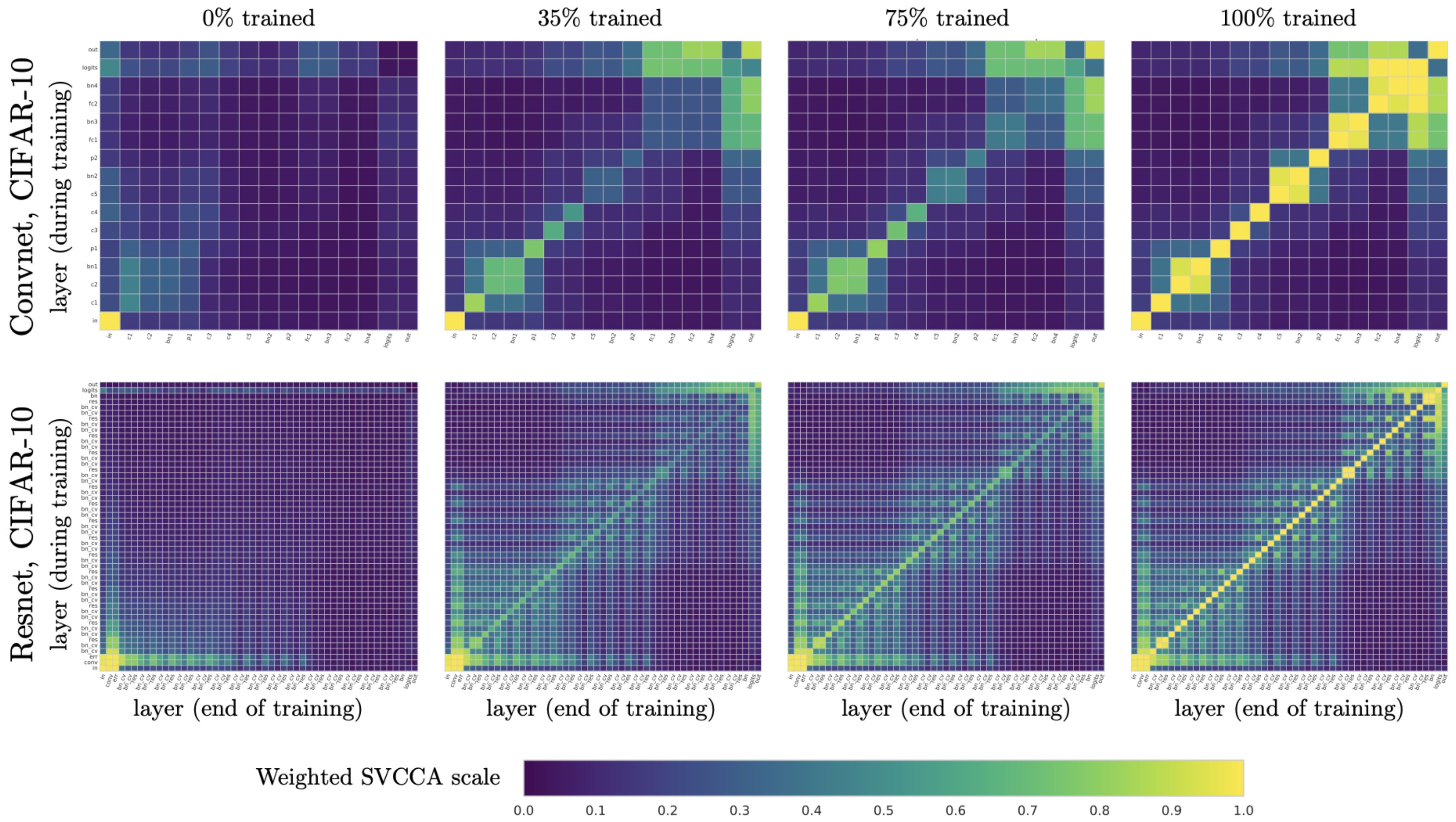
Maithra Raghu,[1,2] **Justin Gilmer,**[1] **Jason Yosinski,**[3] **& Jascha Sohl-Dickstein**[1]
[1]Google Brain [2]Cornell University [3]Uber AI Labs



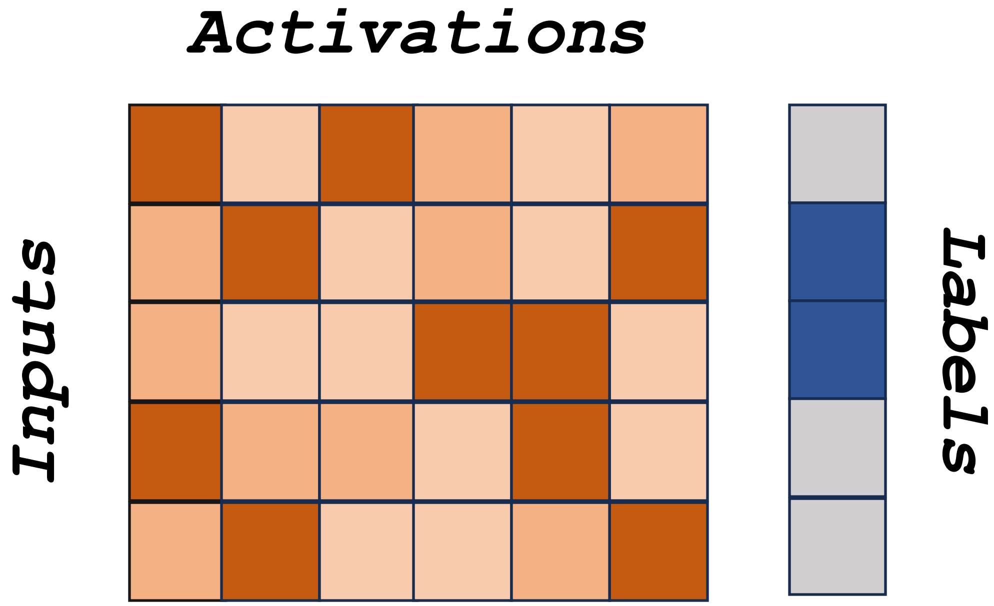
Compare layer to layer similarity with CCA

SV-CCA

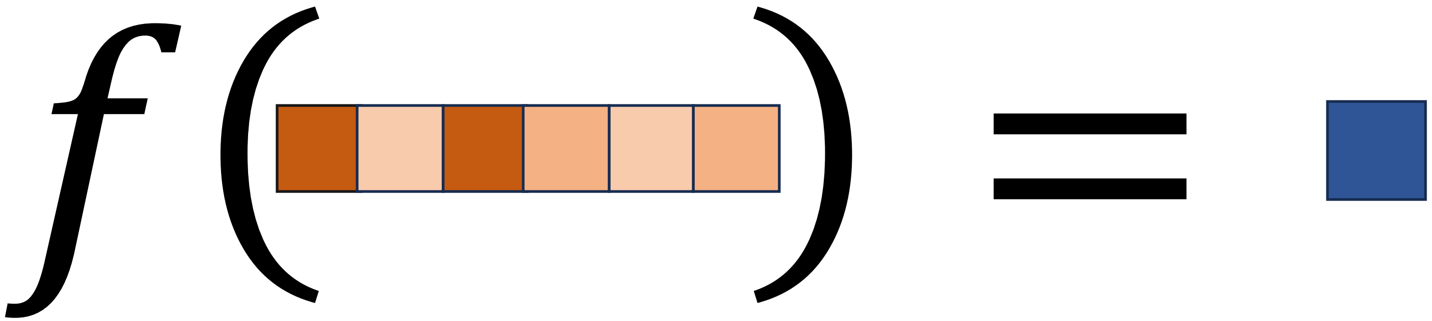
Redundant layers appearing during training → possibilities for pruning



Probes



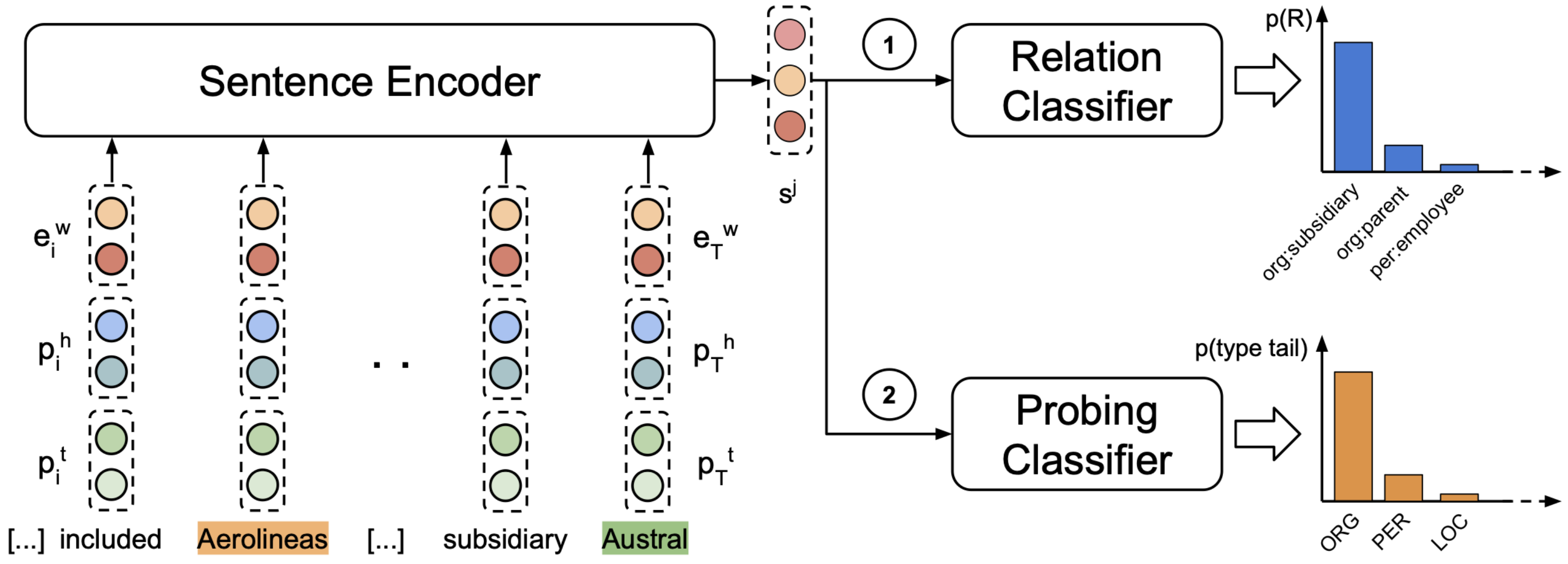
A model f that ***reliably*** predict some behavior labels from activations



Example of Probing: Linguistic Features

Labels: linguistic features

Probing Linguistic Features of Sentence-Level Representations in Neural Relation Extraction
ACL 2020



Use the activations to predict linguistic features

Problems with Probing

Probing the Probing Paradigm:
Does Probing Accuracy Entail Task Relevance?

Abhilasha Ravichander[1] **Yonatan Belinkov**[2*] **Eduard Hovy**[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Technion – Israel Institute of Technology

Probing Classifiers: Promises, Shortcomings, and Advances

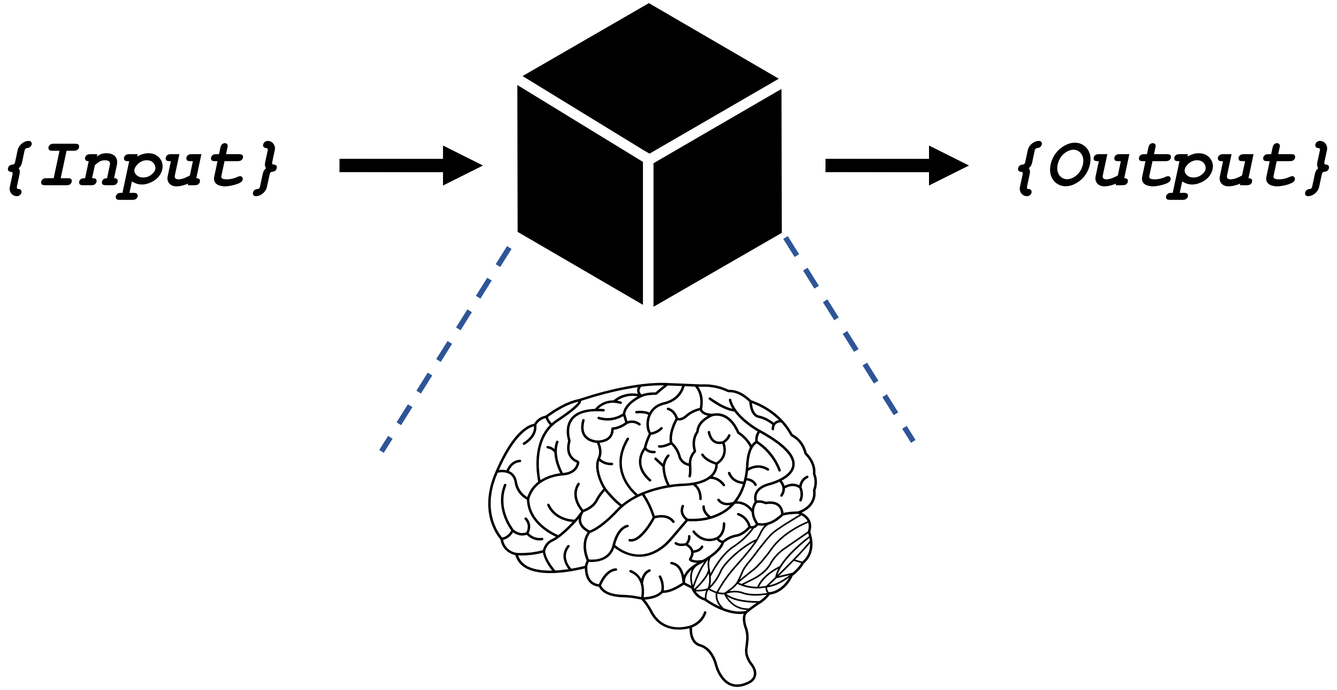
Yonatan Belinkov*
Technion – Israel Institute of Technology

An information theoretic view on selecting linguistic probes

Zining Zhu[1,2], **Frank Rudzicz**[3,4,1,2]
[1] University of Toronto, [2] Vector Institute, [3] Surgical Safety Technologies
[4] Li Ka Shing Knowledge Institute, St Michael's Hospital

(

Neuroscience detour II

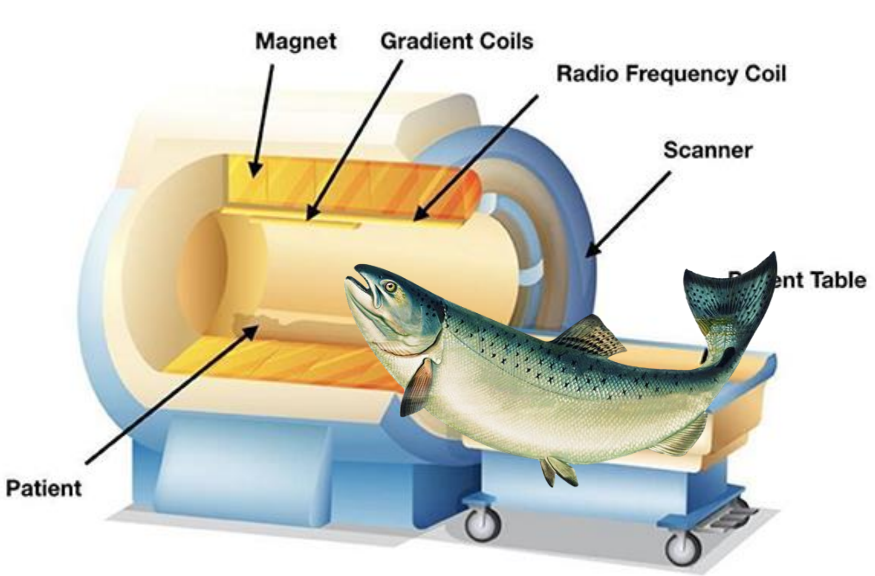


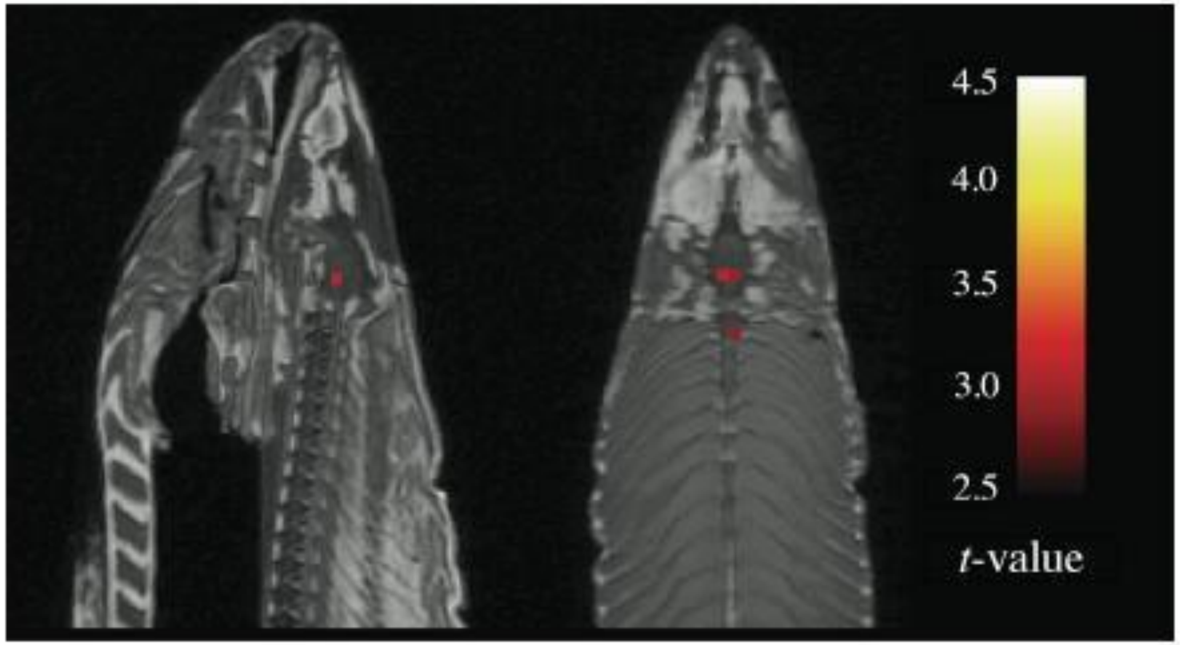
Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging

Philip A. Kragel[1,2] · Leonie Koban[1] · Lisa Feldman Barrett[3,4,5] · Tor D. Wager[1]

Very common to do "Probing" on brain activations: *"mutivariate pattern analyses", "brain signatures", …*

The Dead Salmon





Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

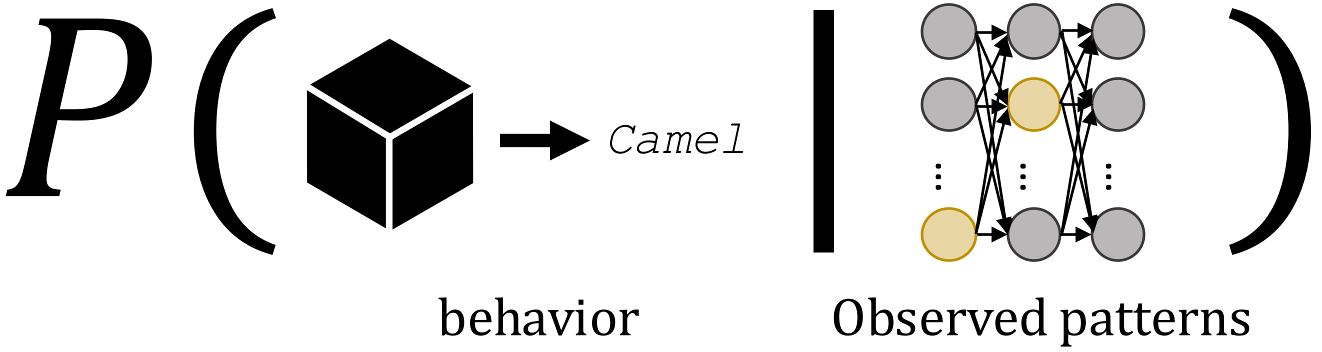
The lure of misleading causal statements in functional connectivity research

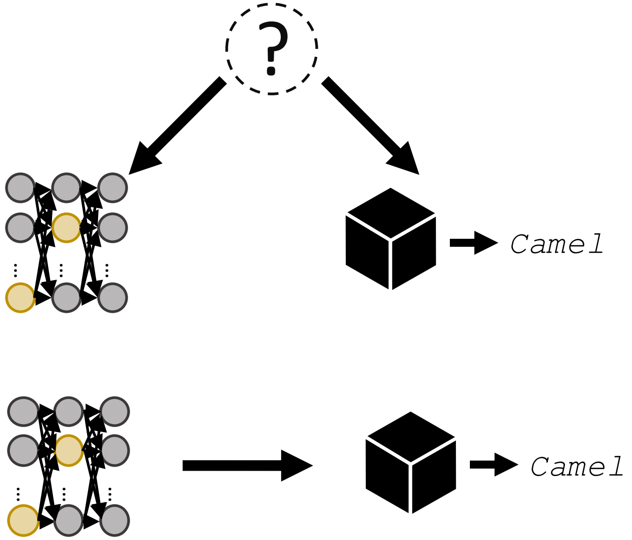
David Marc Anton Mehler, Konrad Paul Kording

)

Neural correlates fall short

Predicting is not understanding – correlations are everywhere and do not generalize





Do the observed patterns ***explain*** the behavior?

Predicting is not Understanding:

Damien Teney[1,3] **Maxime Peyrard**[2] **Ehsan Abbasnejad**[3]

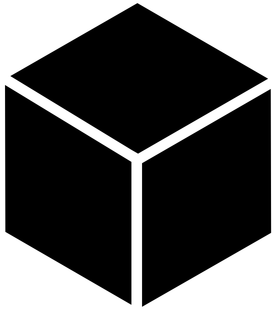
Sparse Autoencoders Can Interpret Randomly Initialized Transformers

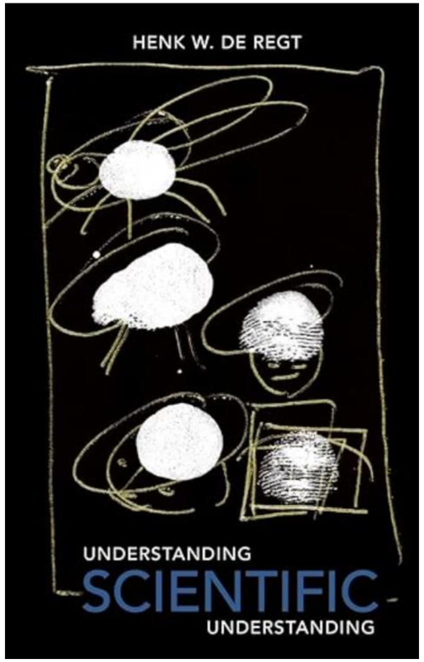
Thomas Heap, Tim Lawson, Lucy Farnik, Laurence Aitchison

How to Explain?

“Causal patterns”

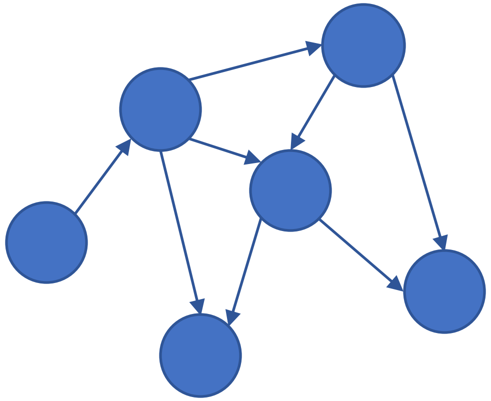
I - Causal Models

{Input} →  → *{Output}*



Why?

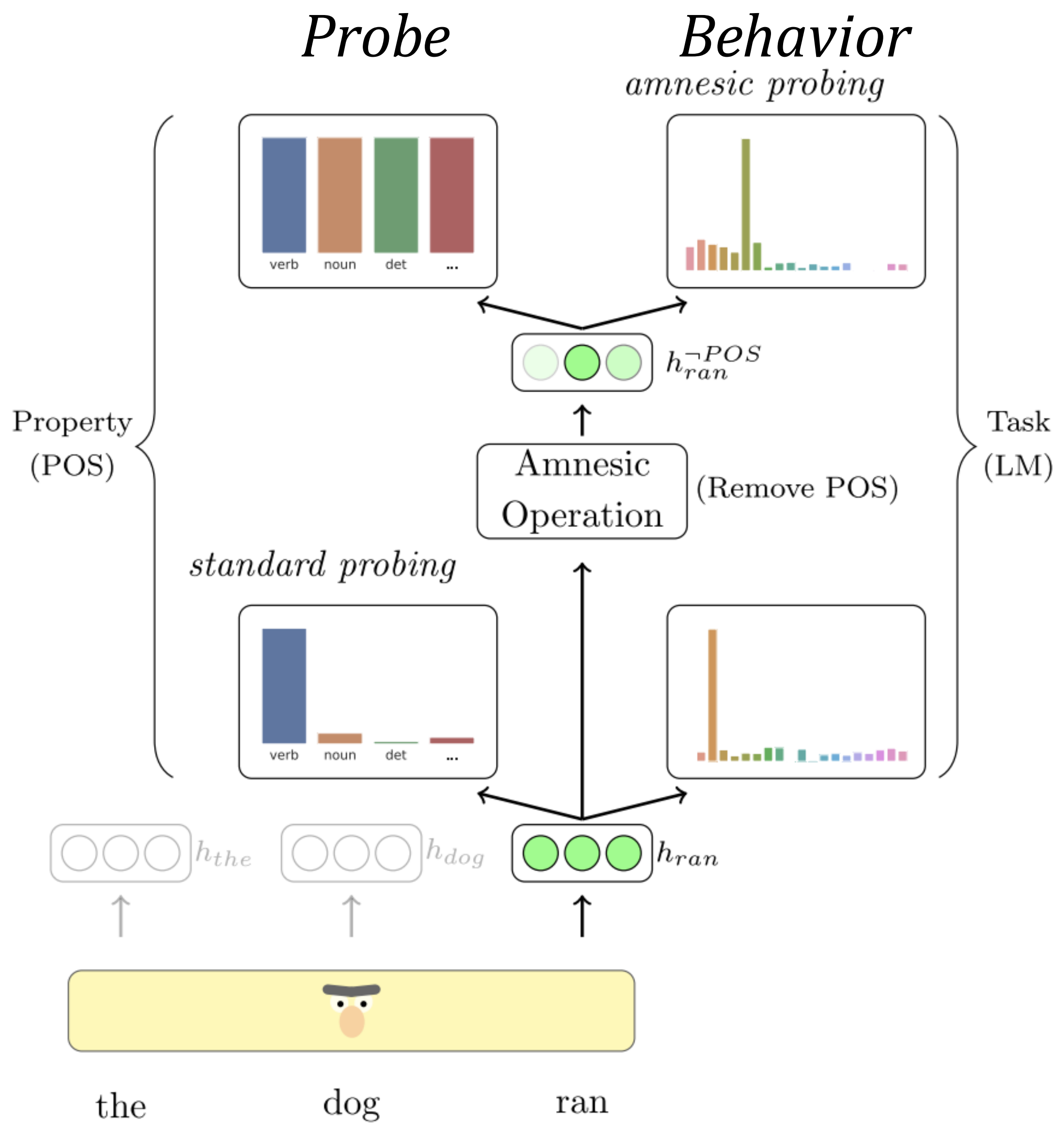
Philosophers of Science argue that **explanations** must be **causal analyses**



"causes explain their effects »

- Understanding: know the behavior in any scenario
- Control: know the impact of modifications on the system

Amnesic probing: Validating with Interventions



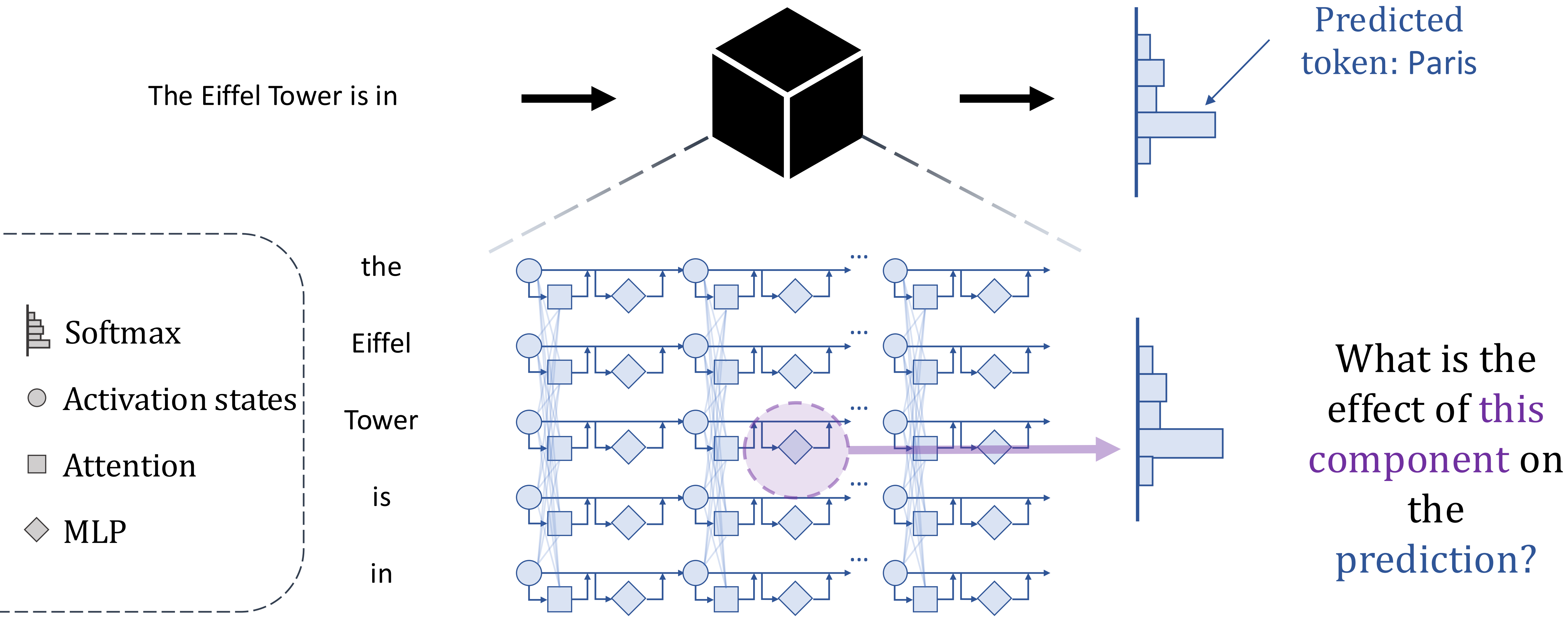
Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals

Yanai Elazar[1,2] **Shauli Ravfogel**[1,2] **Alon Jacovi**[1] **Yoav Goldberg**[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence

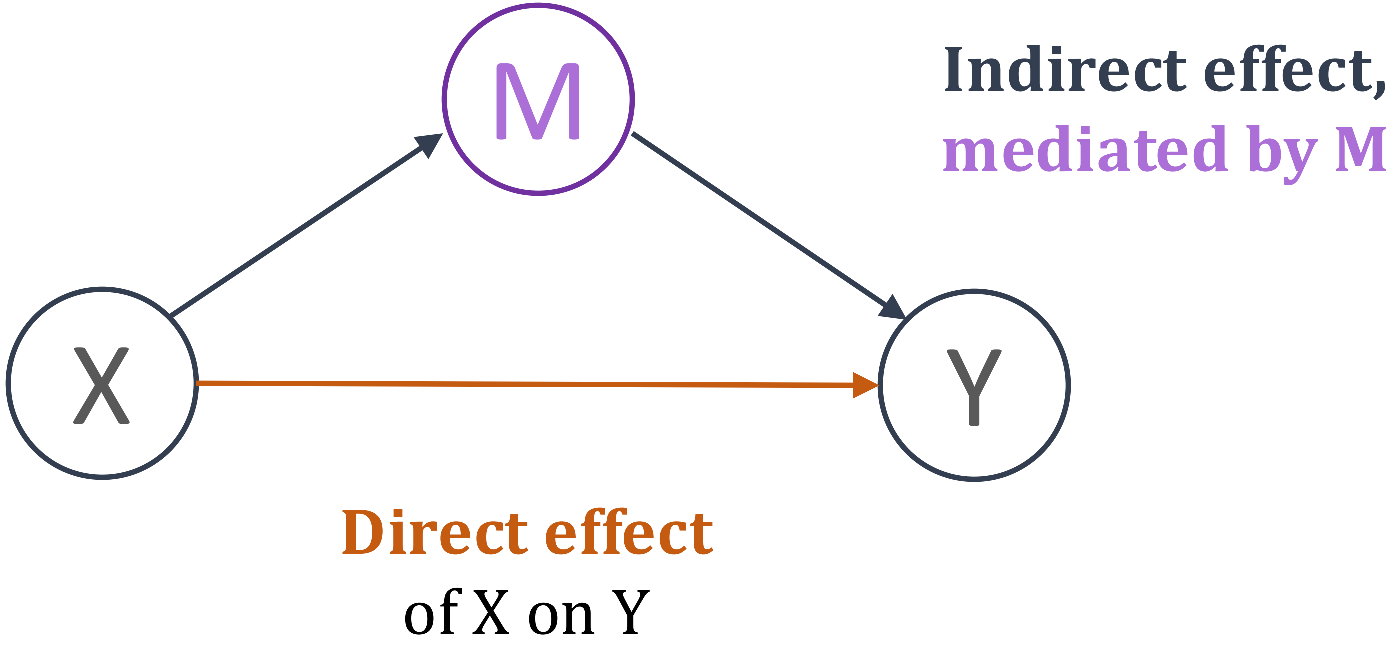
Does modifying the activation to fool the probe, removes the behavior?

Causal mediation analysis

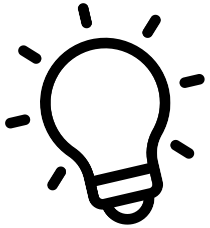
Goal: Understand the impact of *model components* on *model behavior*



Causal Mediation Analysis



How much of the effect of X on Y ***is explained by*** *the path through M?*



i.e., understanding the mechanisms by which X acts on Y,
disentangling the different paths of influences.

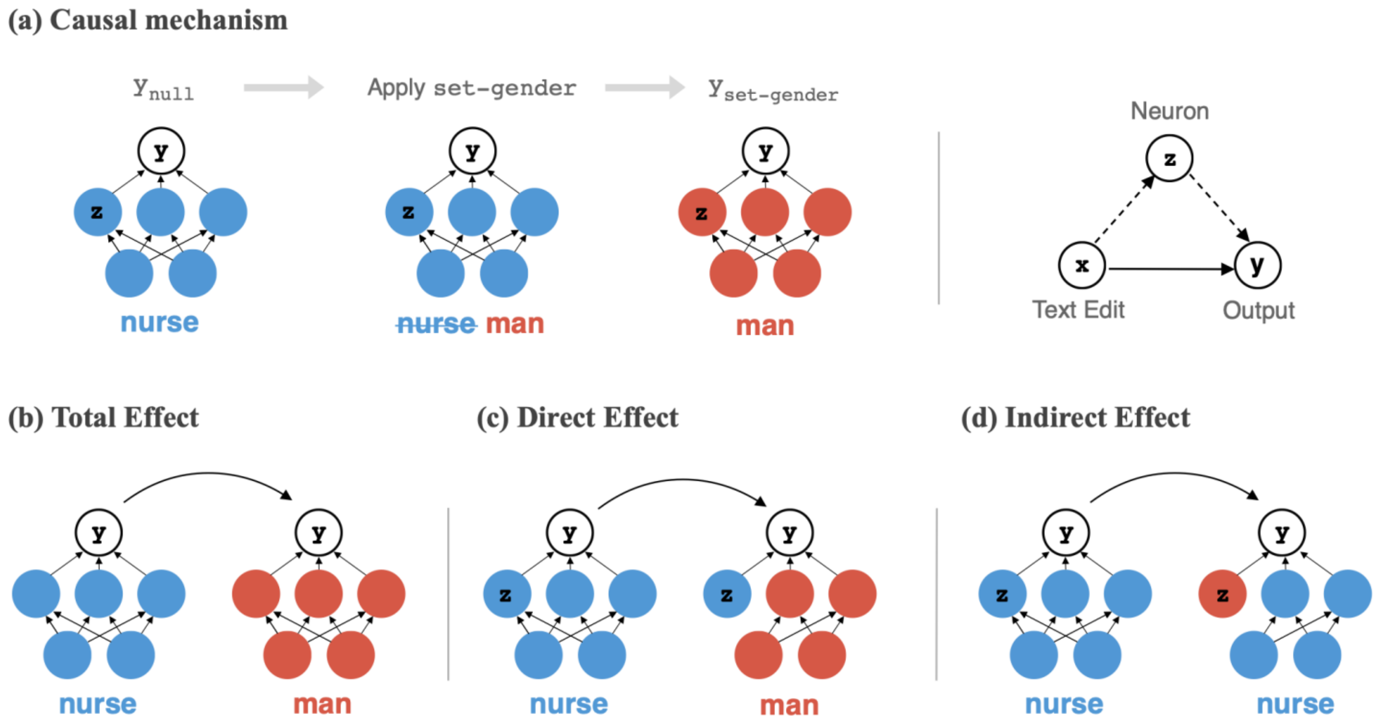
Examples – Gender Bias

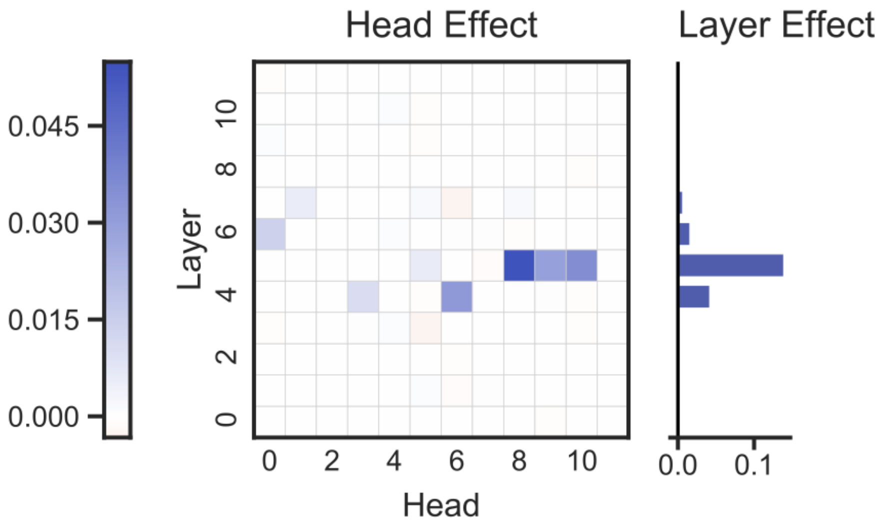
Prompt u: The nurse said that __

Stereotypical candidate: she

Anti-stereotypical candidate: he

Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias
NeurIPS 2020





Examples – factual recall

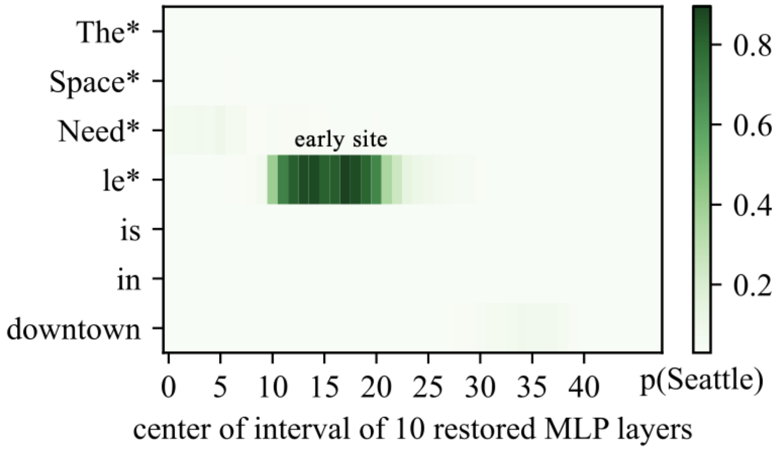
Locating and Editing Factual Associations in GPT

Kevin Meng* MIT CSAIL **David Bau*** Northeastern University **Alex Andonian** MIT CSAIL **Yonatan Belinkov**† Technion – IIT

Transformer Feed-Forward Layers Are Key-Value Memories

Mor Geva[1,2] **Roei Schuster**[1,3] **Jonathan Berant**[1,2] **Omer Levy**[1]
[1]Blavatnik School of Computer Science, Tel-Aviv University
[2]Allen Institute for Artificial Intelligence
[3]Cornell Tech

(f) Impact of restoring MLP after corrupted input



A Glitch in the Matrix?
Locating and Detecting Language Model Grounding with Fakepedia

Giovanni Monea,◇ **Maxime Peyrard,**♡ **Martin Josifoski,**◇ **Vishrav Chaudhary,**♠ **Jason Eisner,**♠ **Emre Kıcıman,**♠ **Hamid Palangi,**♠ **Barun Patra,**♠ **Robert West**◇
◇EPFL ♡Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG ♠Microsoft Corporation

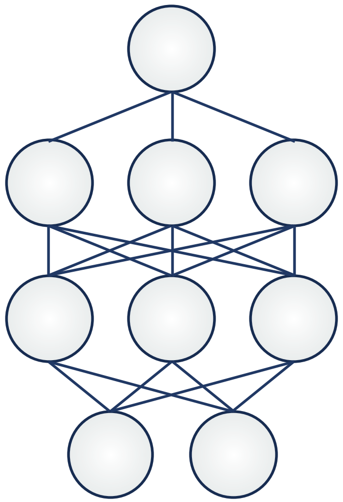
Facts are localized in few MLPs that are associative memories for factual knowledge

High causal effect on the prediction in early sites
→ due to the activity of few MLPs

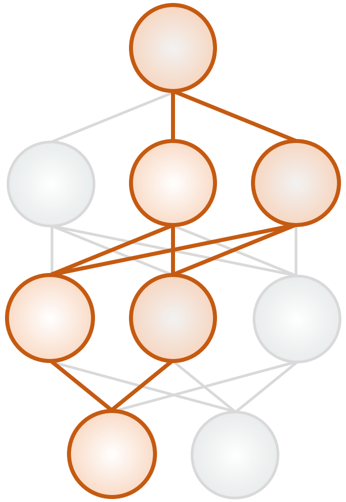
Mechanistic Interpretability

Idea: Reverse-engineer trained neural networks to find simple, human-interpretable, algorithms embedded in the computation

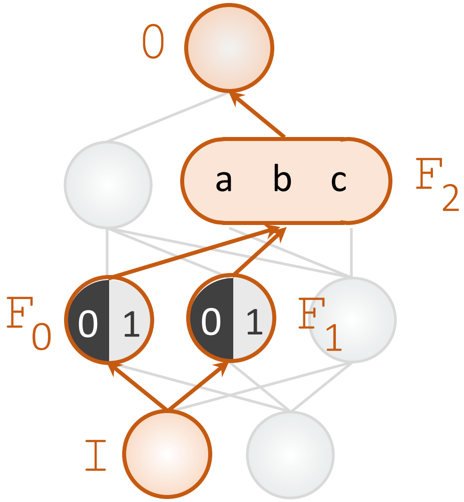
Base
Neural Network



1. Circuit

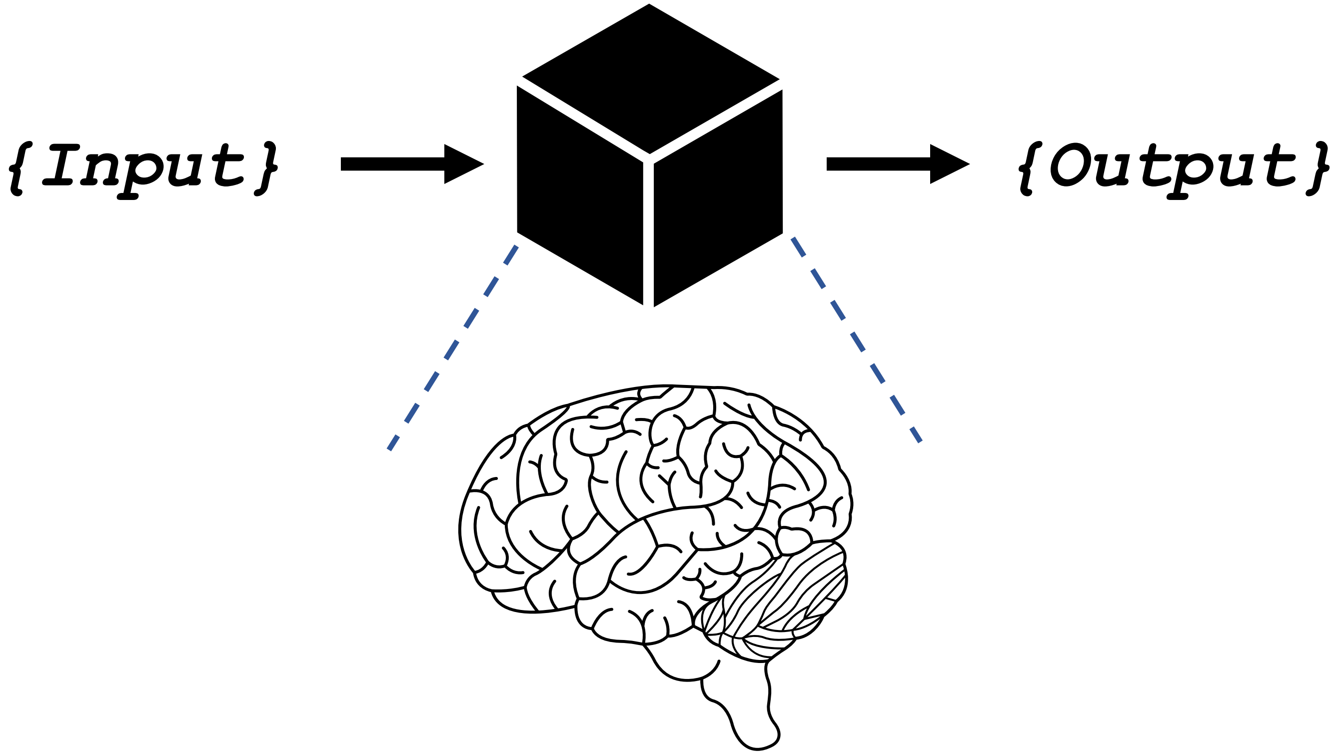


2. Interpret components



(

Neuroscience detour III



How do we know that our explanations are correct?

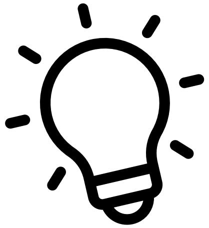
How can we trust our analysis methods if we never test on examples of behavior / true explanations

Neuroscience detour

RESEARCH ARTICLE

Could a Neuroscientist Understand a Microprocessor?

Eric Jonas[1]***, Konrad Paul Kording**[2,3]

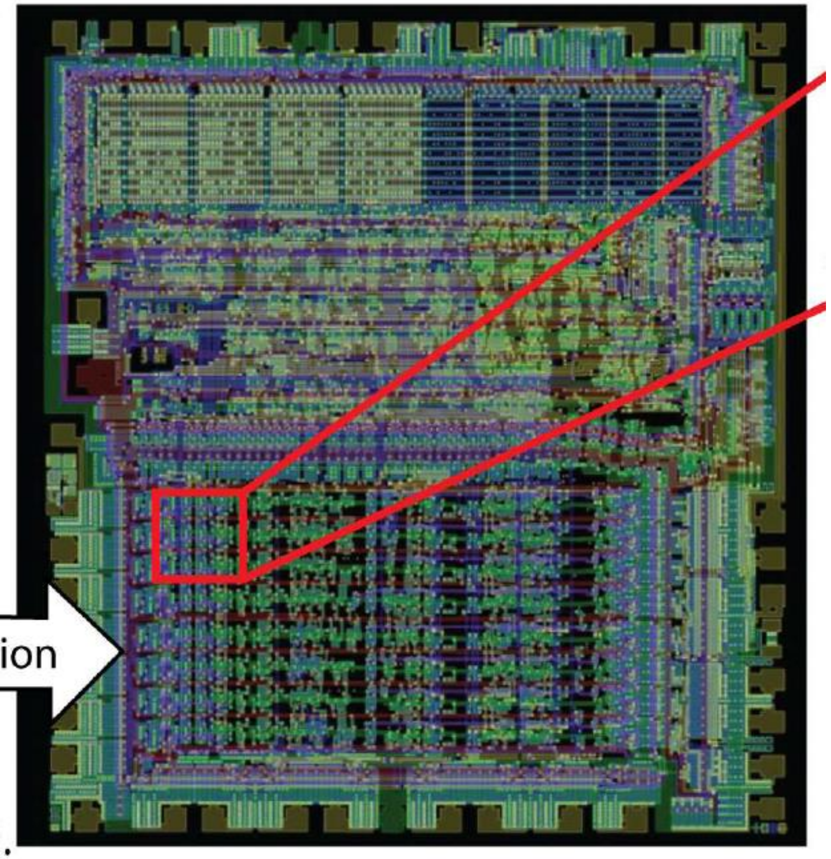


We know everything about the microprocessor,
Let's treat it as if it was a brain (where transistor ≈ neurons)

Can analysis methods recover meaningful information about the microprocessor even with perfect observations and manipulation capabilities?

NO

Neuroscience detour



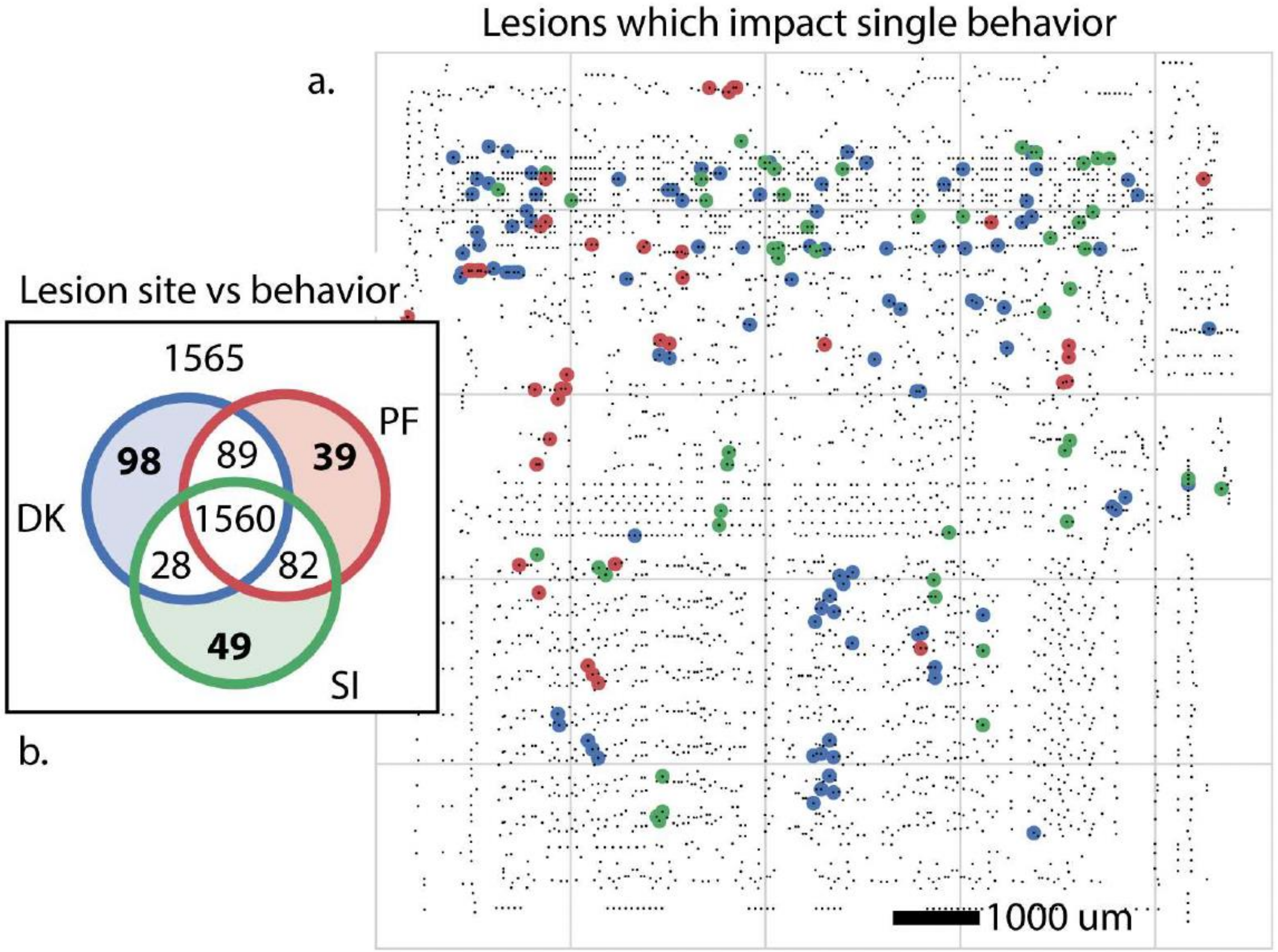


Fig 4. Lesioning every single transistor to identify function. We identify transistors whose elimination disrupts behavior analogous to lethal alleles or lesioned brain areas. These are transistors whose elimination results in the processor failing to render the game. (**A**) Transistors which impact only one behavior, colored by behavior. (**B**) Breakdown of the impact of transistor lesion by behavioral state. The elimination of 1565 transistors have no impact, and 1560 inhibit all behaviors.

Even extensive intervention study gives no useful information!

)

MI is not the end of the story

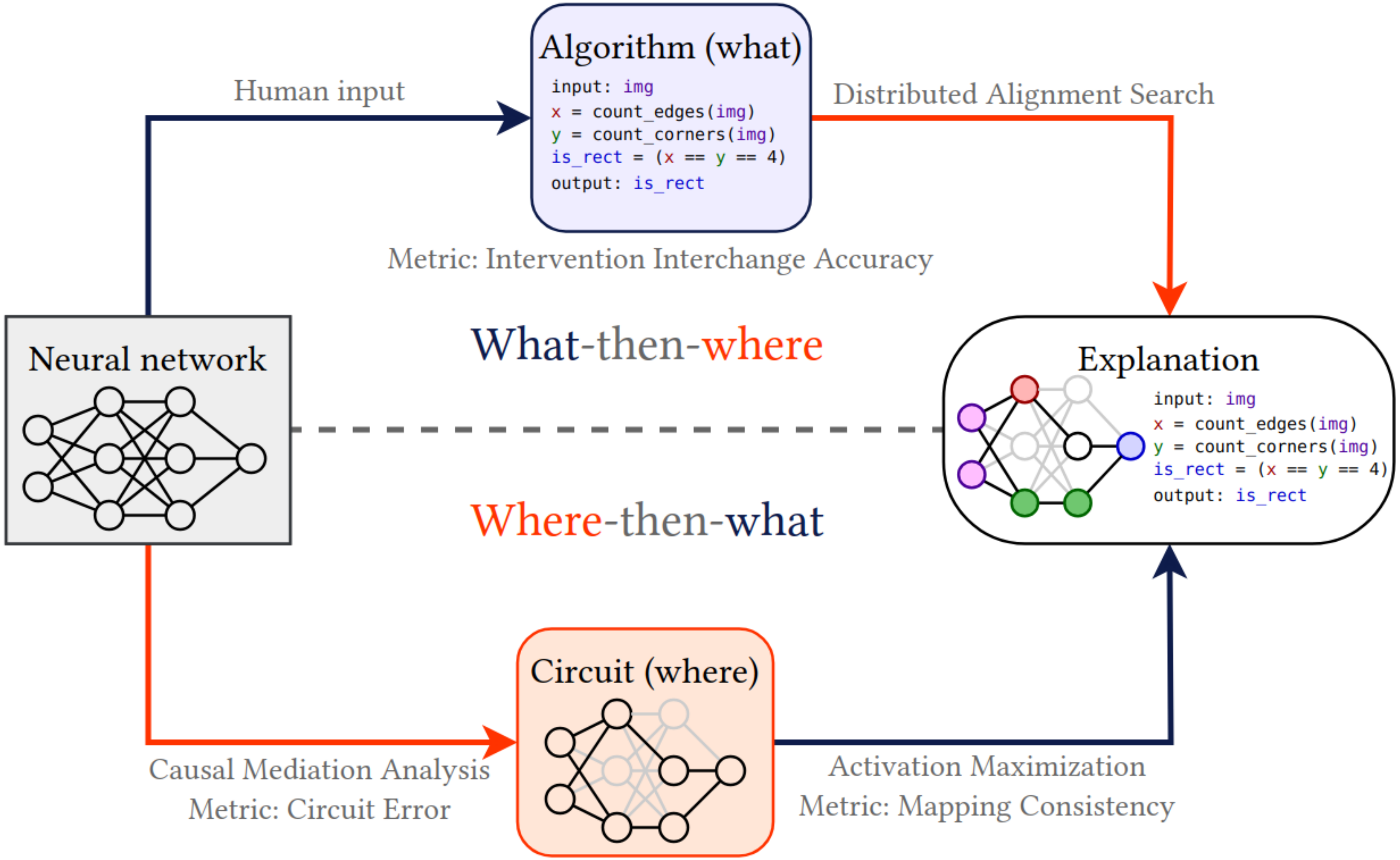
Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiable?

Maxime Méloux, Silviu Maniu, François Portet, Maxime Peyrard

We can find *almost any explanation* if we look hard enough
Even in randomly initialized neural networks

"High-dimensional nonlinear systems may be hard to understand, but they are easy to find stories in." – Grace W. Lindsay → **un-identifiable**

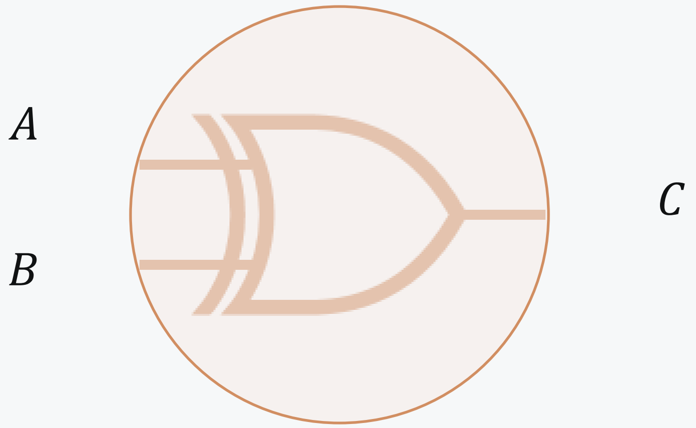
The Two Types of MI approaches



Experimental Setup

Base Neural Network

Train an MLP to implement XOR



$$A = 0|1 + \mathcal{N}(0, \mathcal{E})$$

$$B = 0|1 + \mathcal{N}(0, \mathcal{E})$$

$$C = round(A) \oplus round(B)$$

Toy exercise in interpretability:

- What sequence of logic gates is implemented by the MLP?
- Where in the network is each gate implemented?

Enumerate **exhaustively** all candidate algorithms and mappings and test them with existing criteria.

Where-then-What is not Identifiable

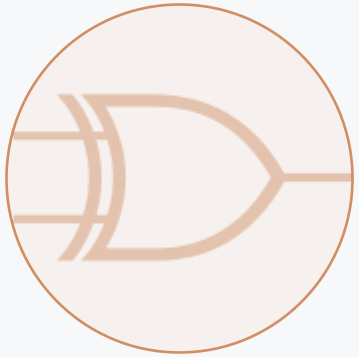
Do the current criteria used for selecting circuits and their grounding induce a unique solution? NO

Base Neural Network

Train an MLP to implement XOR

A

B

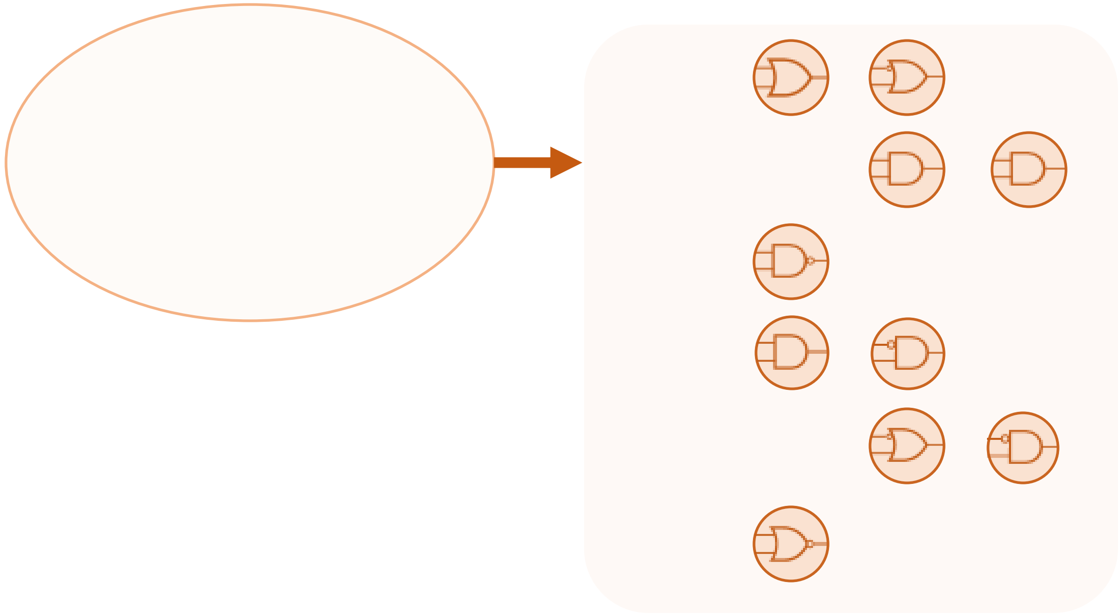


C

$$A = 0|1 + \mathcal{N}(0, \mathcal{E})$$
$$B = 0|1 + \mathcal{N}(0, \mathcal{E})$$
$$C = round(A) \oplus round(B)$$

85 unique circuits with perfect accuracy

25 unique explanations with exact grounding for one circuit



...

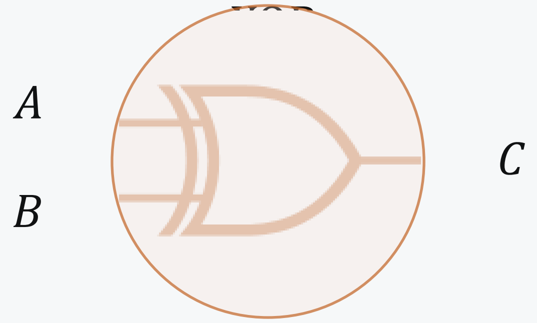
...

What-then-where is not Identifiable

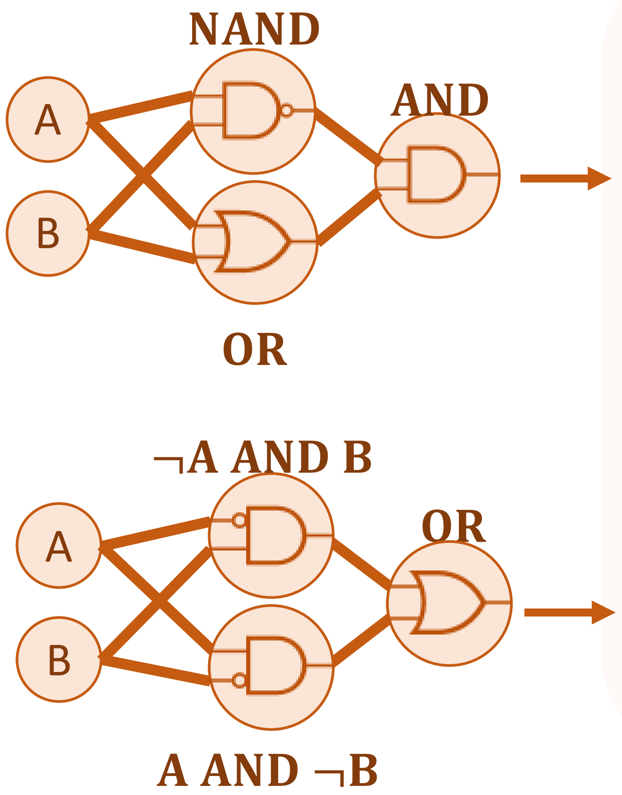
Do the current criteria used for assessing causal alignment of an explanatory algorithm guarantee a unique solution? NO

Base Neural Network

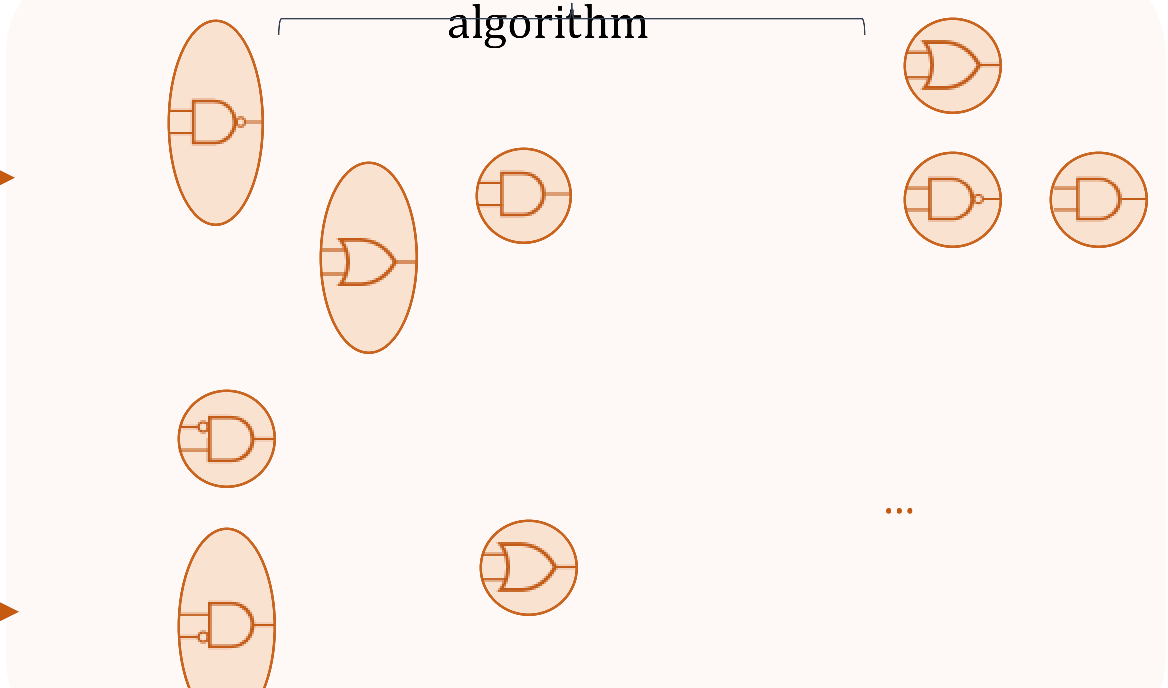
Train an MLP to implement



$A = 0|1 + \mathcal{N}(0, \mathcal{E})$
$B = 0|1 + \mathcal{N}(0, \mathcal{E})$
$C = round(A) \oplus round(B)$

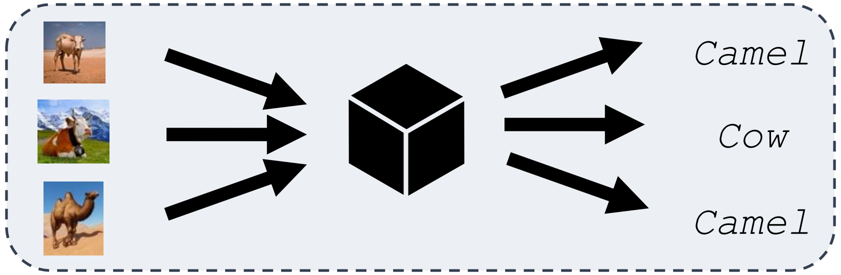


Example of **2 perfect mappings** for one algorithm

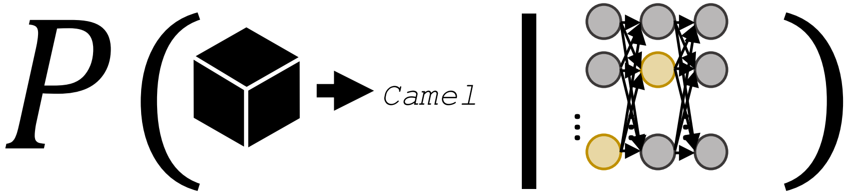


159 perfect mappings (IIA=1)

Identifiability Issues → Generalization Issues



Behavior: Many explanations compatible with observed behavior. *Which one matches the computation? (Which one generalizes?)*



behavior

Observed patterns

Computational Correlate: Many causal mechanisms compatible with observed correlations. *Which one generalizes?*

Model behavior

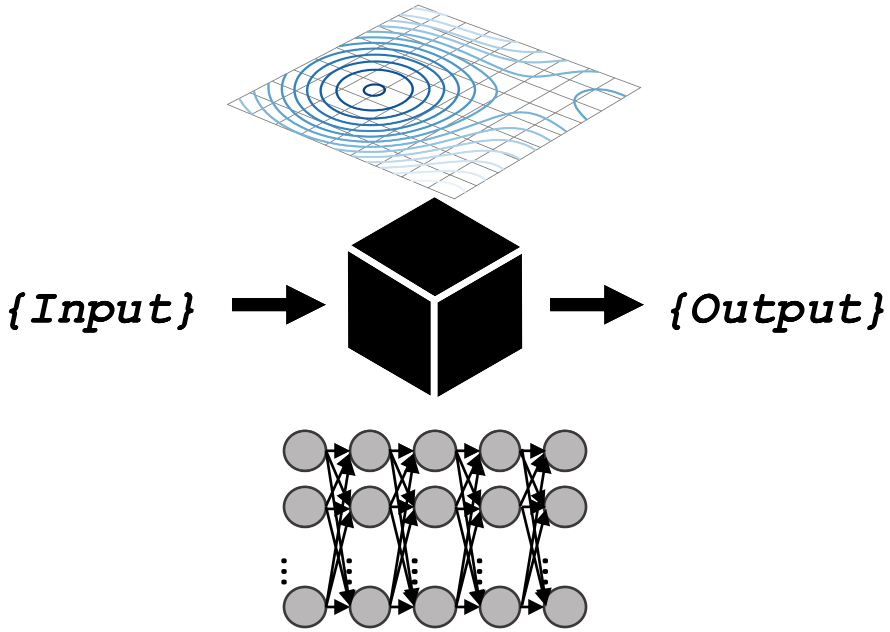
perturbed patterns

Computational Causal Mechanisms: Many causally aligned explanations!! *(Which one generalizes?)*

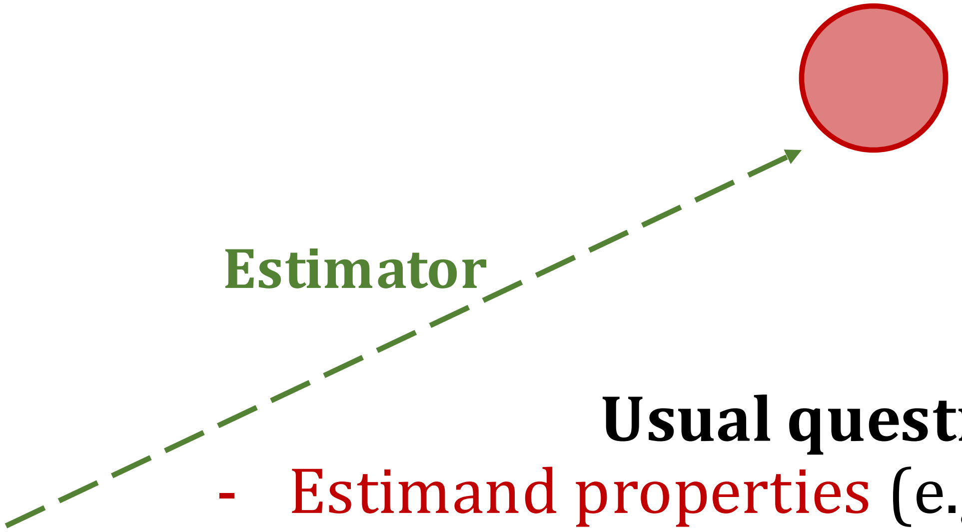
What to do?

Instrumentalism: Statistical (Causal) Inference on computational data

Distribution over computational traces



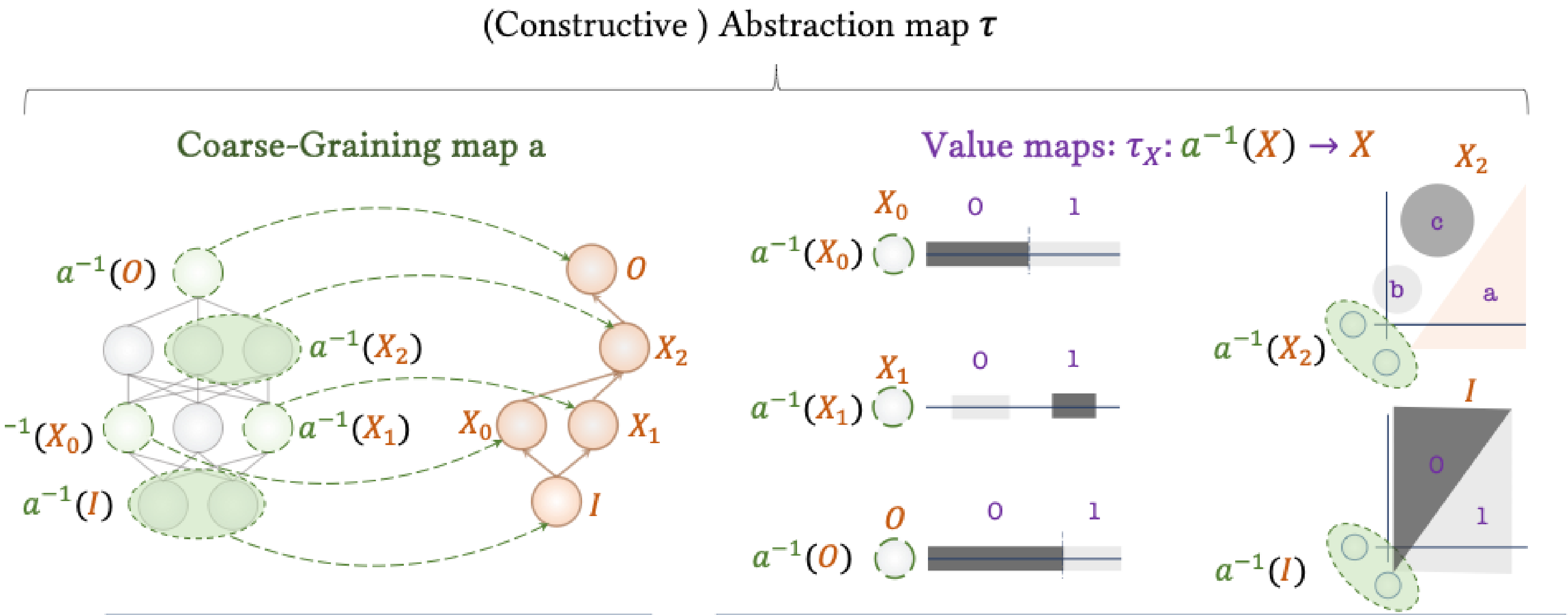
Target property of interest
(estimand)



Usual questions:

- Estimand properties (e.g., identifiability)
- Estimators' properties:
  Bias, variance, consistency, …
- Distributional properties:
  Generalization, uncertainty

Computational Summarization aka Causal Abstraction



Specifies which micro-variable maps to which macro-variable

Specifies how the micro-states (values of the micro-variables) define the macro-states (values of macro-variables)

Abstracting Causal Models

Sander Beckers **Joseph Y. Halpern**

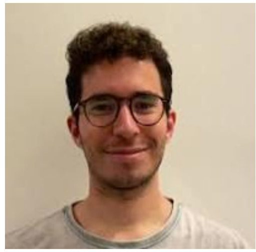
Causal Abstraction:
A Theoretical Foundation for Mechanistic Interpretability

Atticus Geiger*♢, Duligur Ibeling♠, Amir Zur♢, Maheep Chaudhary♢, Sonakshi Chauhan♢, Jing Huang♠, Aryaman Arora♠, Zhengxuan Wu♠, Noah Goodman♠, Christopher Potts♠, Thomas Icard*♠

The high-level model $\mathcal{A}$ is a **causal abstraction** of the low-level implementation $\mathcal{L}$ *if the variables in $\mathcal{A}$ play the same causal role as their associated low-level variables.*

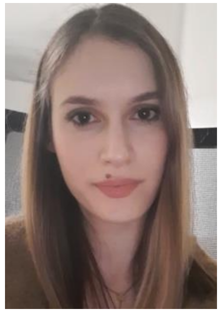
Collaborators: Thank you!

Damien Teney



Giovanni Monea

Robert West



Marija Sakota

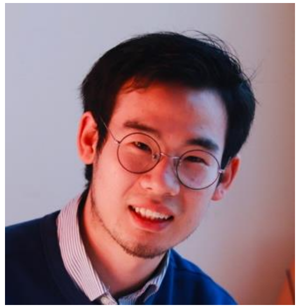
Martin Josifoski

Emre Kiciman

Wei Zhao

Jason Eisner

Debjit Paul



Saibo Geng

Kristina Gligoric

Maxime Méloux

Francois Portet

Fei liu

Thank you!
Questions?

Contact: maxime.peyrard@univ-grenoble-alpes.fr