

NeuroMod Annual Meeting
Université Côte d'Azur
Antibes, France

What are Neural Language Models the Model of?
Epistemological and Theoretical Perspectives on LLMs

Juan Luis Gastaldi

www.giannigastaldi.com

ETH zürich

July 8, 2025

Introduction

Epistemological Perspectives

Theoretical Perspectives

The Algebra Behind the Embeddings

The Structure Behind the Algebra

The Categories Behind the Structure

Take Aways

Outline

Introduction

Epistemological Perspectives

Theoretical Perspectives

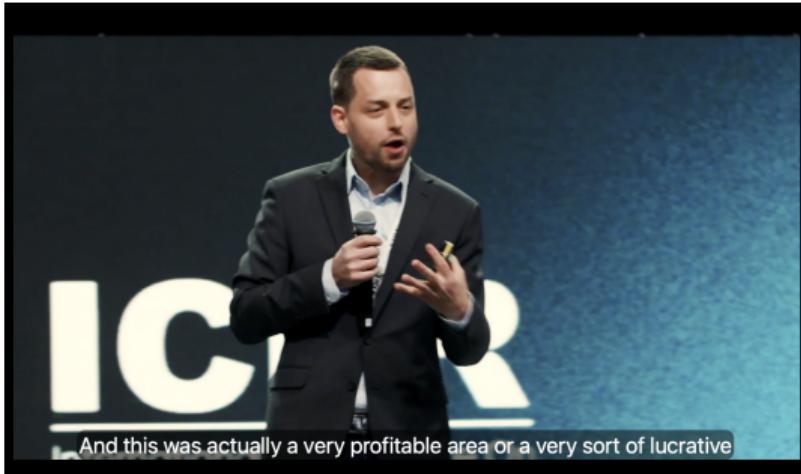
The Algebra Behind the Embeddings

The Structure Behind the Algebra

The Categories Behind the Structure

Take Aways

Empirical Saturnalia



The empirics of deep learning

(Circa 2020) the scaling era is here; deep networks are now just emergent things we have created, that have to be studied scientifically like any other physical phenomenon

It seemed like **the best way for academic research to influence the field** is to develop the biology/physics (and let's be honest, more often pop psychology) of existing large models

24

Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.



Can Large Language Models Be an Alternative to Human Evaluation?

Cheng-Han Chiang
National Taiwan University,
Taiwan
dcml0714@gmail.com

Hung-yi Lee
National Taiwan University,
Taiwan
hungyilee@ntu.edu.tw

And this was actually a very profitable area or a very sort of lucrative

The empirics of deep learning

(Circa 2020) the scaling era is here; deep networks are now just emergent things we have created, that have to be studied scientifically like any other physical phenomenon

It seemed like **the best way for academic research to influence the field** is to develop the biology/physics (and let's be honest, more often pop psychology) of existing large models

24

Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

DO LLMs HAVE CONSISTENT VALUES?

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Liat Bezalel
Tel-Aviv University
liatbezalel@mail.tau.ac.il

Gal Elidan
Google Research
Hebrew University
elidan@google.com

Amir Globerson
Google Research
Tel-Aviv University
amirg@google.com

Ella Daniel
Tel-Aviv University
della@tauex.tau.ac.il

Cheng-Han Chiang
National Taiwan University,
Taiwan
dcml0714@gmail.com

Hung-yi Lee
National Taiwan University,
Taiwan
hungyilee@ntu.edu.tw

And this was actually a very profitable area or a very sort of lucrative

The empirics of deep learning

ca 2020) the scaling era is here; deep networks are now just emergent things have created, that have to be studied scientifically like any other physical phenomenon

seemed like **the best way for academic research to influence the field** is to develop the biology/physics (and let's be honest, more often pop psychology) of existing large models

24

Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

Can Large

DO LLMS HAVE CON

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Gal Elidan
Google Research
Hebrew University
elidan@google.com

Cheng-Han Chiang
National Taiwan University,
Taiwan
dcml0714@gmail.com

Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithm Discovery for Recursive Self-Improvement through Reinforcement Learning

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com

Amir Globerson
Google Research
Tel-Aviv University
amirg@google.com

Hung-yi Lee
National Taiwan University,
Taiwan
hungyilee@ntu.edu.tw

Taro Yano
NEC
taro_yano@nec.com

Ella Daniel
Tel-Aviv University
della@tauex.tau.ac.il

Masafumi Oyamada
NEC
oyamada@nec.com

Deep learning

ca 2020) the scaling era is here; deep networks are now just emergent things have created, that have to be studied scientifically like any other physical phenomenon

seemed like the best way for academic research to influence the field is to develop the biology/physics (and let's be honest, more often pop psychology) of existing large models

And this was actually a very profitable area or a very sort of lucrative

24

Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

Empirical Saturnalia

Can Large

DO LLMS HAVE CON

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Gal Elidan
Google Research
Hebrew University
elidan@google.com

Cheng-Han Chiang
National Taiwan University,
Taiwan
dcml0714@gmail.com

Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithm Discovery for Reinforcement Learning

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com

Amir Globerson
Google Research
Tel-Aviv University
amirg@google.com

Hung-yi Lee
National Taiwan University,
Taiwan
hungyilee@ntu.edu.tw

DO LLMS “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Juyeon Heo^{1,*} Christina Heinze-Deml² Oussama Elachqar² Kwan Ho Ryan Chan^{3,*} Shirley Ren²
Udhay Nallasamy² Andy Miller² Jaya Narain²

¹University of Cambridge ²Apple ³University of Pennsylvania
jh2324@cam.ac.uk jnarain@apple.com

Seemed like the best way for academic research to influence the field is to

develop the biology/physics (and let's be honest, more often pop psychology) of existing large models

And this was actually a very profitable area or a very sort of lucrative

24

Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithm Discovery for Reinforcement Learning

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com

Do LLMs “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Juyeon Hwang, Udhay Narayanan, Jiebo Li, Cheng-Han Chiang, Hung-yi Lee, Ella Daniel, Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, Kaixiang Lin

DO LLMs HAVE CONSCIOUSNESS?

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Gal Elidan
Google Research
Hebrew University
elidan@google.com

Amir Globerson
Google Research
Tel-Aviv University
amirg@google.com

Ella Daniel
Tel-Aviv University
della@tauex.tau.ac.il

Cheng-Han Chiang
National Taiwan University,
Taiwan
dcml0714@gmail.com

Hung-yi Lee
National Taiwan University,
Taiwan
hungyilee@ntu.edu.tw

Siyan Zhao^{2*}, Mingyi Hong^{1,3}, Yang Liu¹, Devamanyu Hazarika¹, Kaixiang Lin¹ †
¹Amazon AGI, ²UCLA, ³University of Minnesota
siyanz@cs.ucla.edu, mhong@umn.edu, devamanyu@u.nus.edu
{yangliud, kaixianl}@amazon.com

And this was actually a very profitable area or a very sort of lucrative

Empirical Saturnalia

DO LLMs HAVE CONSCIOUSNESS?

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Gal Elidan
Google Research
Hebrew University
elidan@google.com

Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithm Discovery for Reinforcement Learning

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com

Amir Globerson
Google Research
Tel-Aviv University
amirg@google.com

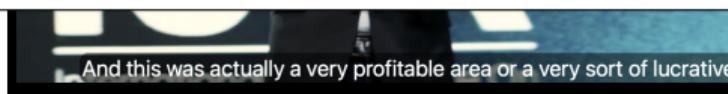
DO LLMs “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Juyeon Hwang
Udhay Narayanan
¹University of Texas at Austin
jh2324@mail.utexas.edu

Cheng-Han Chiang
National Taiwan University,
Taiwan
dcml0714@gmail.com

Hung-yi Lee
National Taiwan University,
Taiwan
hungyilee@ntu.edu.tw

DO LLMs RECOGNIZE YOUR PREFERENCES? EVALUATING PERSONALIZED PREFERENCE FOLLOWING IN LLMs



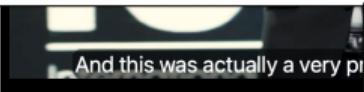
Zico Kolter, *Building Safe and*

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Empirical Saturnalia

Can Large Language Models Invent Algorithms to Improve Themselves? Algorithm Discovery for Reinforcement Learning	DO LLMs HAVE CONSCIOUSNESS?	DO LLMs “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?	
	Naama Rozen Tel-Aviv University naamarozen240@gmail.com	Yoichi Ishibashi NEC yoichi-ishibashi@nec.com	Juyeon Han Udhay Narayanan ¹ University of Texas at Austin jh2324@mail.utexas.edu
Gal Elidan Google Research Hebrew University elidan@google.com	Amir Globerson Google Research Tel-Aviv University amirg@google.com	Ella Daniel Tel-Aviv University della@tauex.tau.ac.il	DO LLMs RECOGNIZE YOUR PREFERENCES? EVALUATING PERSONALIZED PREFERENCE FOLLOWING IN LLMs
Cheng-Han Chiang National Taiwan University, Taiwan dcml0714@gmail.com	Hung-yi Lee National Taiwan University, Taiwan hungyilee@ntu.edu.tw	de ex	Language Models are Few-Shot Learners
		<p>LLMs Are Not Intelligent Thinkers: Introducing Mathematical Topic Tree Benchmark for Comprehensive Evaluation of LLMs</p> <p>Arash Gholami Davoodi¹, Seyed Pouyan Mousavi Davoudi, Pouya Pezeshkpour² ¹Carnegie Mellon University, ²Megagon Labs agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai</p> <p>Samuel K. Ryder*, Melanie Subbiah*, Pranav Shyam, Girish Sastry, Gretchen Krueger, Tom Henighan</p>	

Empirical Saturnalia

<p>DO LLMS HAVE CONSCIOUSNESS?</p> <p>Naama Rozen Tel-Aviv University naamarozen240@gmail.com</p> <p>Gal Elidan Google Research Hebrew University elidan@google.com</p>	<p>Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithm Discovery for Reinforcement Learning</p> <p>Yoichi Ishibashi NEC yoichi-ishibashi@nec.com</p> <p>Amir Globerson Google Research Tel-Aviv University amirg@google.com</p> <p>Ella Daniel Tel-Aviv University della@tauex.tau.ac.il</p>	<p>DO LLMs “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?</p> <p>Juyeon Han¹, Udhay Narayanan², ... ¹University of Illinois Urbana-Champaign jh2324@illinois.edu</p>	
<p>Can Large Language Models Recognize Personal Preferences?</p> <p>Cheng-Han Chiang National Taiwan University, Taiwan</p> <p>Hung-yi Lee National Taiwan University, Taiwan</p>	<p>When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs</p> <p>Ryo Kamoi¹, Yusen Zhang¹, Nan Zhang¹, Jiawei Han², Rui Zhang¹ ¹Penn State University, USA ²University of Illinois Urbana-Champaign, USA {ryokamoi, rmz5227}@psu.edu</p> <p>agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai</p>	<p>Language Models are Few-Shot Learners</p> <p>Yannic Kilcher¹, ... ¹Megagon Labs</p> <p>Davoudi, Pouya Pezeshkpour² ²Megagon Labs</p> <p>Gretchen Krueger¹, Tom Henighan² ¹CMU, ²OpenAI</p>	

DO LLMS HAVE CONSCIOUSNESS?	Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithm Discovery for Reinforcement Learning	DO LLMS “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?
Can Large Language Models be Zero-Shot Reasoners?	Large Language Models are Zero-Shot Reasoners	DO LLMS RECOGNIZE YOUR PREFERENCES? EVALUATING PERSONALIZED PREFERENCE FOLLOWING IN LLMS
Natural Language Processing Takeshi Kojima The University of Tokyo t.kojima@weblab.t.u-tokyo.ac.jp	Shixiang Shane Gu Google Research, Brain Team Machel Reid Google Research* Yutaka Matsuo The University of Tokyo Yusuke Iwasawa The University of Tokyo	Language Models are Few-Shot Learners Learning Mathematical Topic Tree Evaluation of LLMs
Ryo Kamoi ¹ Yusen Zhang ¹ Nan Zhang ¹ Jiawei Han ² Rui Zhang ¹ ¹ Penn State University, USA ² University of Illinois Urbana-Champaign, USA {ryokamoi, rmz5227}@psu.edu	Davoudi, Pouya Pezeshkpour ² Megagon Labs agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai	Samuel K. Ryder* Melanie Subbiah* Pranav Shyam Girish Sastry Gretchen Krueger Tom Henighan

Can Large Language Models Invent Algorithms to Improve Themselves?:
Algorithm Discovery for Reinforcement Learning

DO LLMS HAVE CONSCIOUSNESS?

Sparks of Artificial General Intelligence:
Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Natalia Lai Naai Tel-Aviv University

Can Large Language Models Invent Algorithms to Improve Themselves?

“OW” INTERNALLY WHEN THEY FOLLOW?

LLMs RECOGNIZE YOUR PREFERENCES? EVALUATING PERSONALIZED PREFERENCE FOLLOWING IN LLMs

Takeshi Kojima Shixiang Shane Gu
The University of Tokyo Google Research, Brain Team
t.kojima@weblab.t.u-tokyo.ac.jp

Machel Reid Yutaka Matsuo Yusuke Iwasawa
Google Research* The University of Tokyo The University of Tokyo

Ryo Kamoi¹ Yusen Zhang¹ Nan Zhang¹ Jiawei Han² Rui Zhang¹
¹Penn State University, USA ²University of Illinois Urbana-Champaign, USA
{ryokamoi, rmz5227}@psu.edu

Language Models are Few-Shot Learners

Learning Mathematical Topic Tree Evaluation of LLMs

Davoudi, Pouya Pezeshkpour²
Megagon Labs
davoudi@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

Samuel K. Ryder* Melanie Subbiah*
Pranav Shyam Girish Sastry
Gretchen Krueger Tom Henighan

Can Large Language Models Invent Algorithms to Improve Themselves?: PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES REASONING IN LARGE LANGUAGE MODELS	WHEN THEY FOLLOW YOUR PREFERENCES? EVALUATING PERSONALIZED PREFERENCE FOLLOWING IN LMs	
	La Sparks E. Sébastien Bubeck Eric Horvitz Harsha Nori Laura Ruis* AI Centre, UCL Hamid Palangi Microsoft Research Maximilian Mozes Cohere Juhan Bae University of Toronto & Vector Institute Marco Tulio Ribeiro Yi Zhang	Takeshi Kojima The University of Tokyo t.kojima@weblab.t.u-tokyo.ac.jp Shixiang Shane Gu Google Research, Brain Team Machel Reid Google Research* Yutaka Matsuo The University of Tokyo Yusuke Iwasawa The University of Tokyo Ryo Kamoi ¹ Yusen Zhang ¹ Nan Zhang ¹ Jiawei Han ² Rui Zhang ¹ ¹ Penn State University, USA ² University of Illinois Urbana-Champaign, USA {ryokamoi, rmz5227}@psu.edu Davoudi, Pouya Pezeshkpour ² Megagon Labs agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

Empirical Saturnalia

DO LLMS HAVE COMMON SENSE?

Natalia Tiel-Antonsen, Daniel L. Sparks, Sébastien Bubeck, Eric Horvitz, Ece Kamar, Harsha Nori

Can Large Language Models Invent Algorithms to Improve Themselves?: PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES WHEN THEY FOLLOW THE LEADER

Laura Ruis*, Hamid Palangi, Microsoft Research

Can LLMs Learn From Mistakes? An Empirical Study on Reasoning Tasks

Shengnan An^{*◇*}, Zexiong Ma^{*◇*}, Siqi Cai^{*◇*}, Zeqi Lin^{†*}, Nanning Zheng^{†◇}, Jian-Guang Lou^{*}, Weizhu Chen^{*}

◇National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center of Visual Information and Applications, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
*Microsoft ◇Peking University

Language Models are Few-Shot Learners

Takeshi Kojima, Shixiang Shane Gu, Google Research, Brain Team

Machel Reid, Yutaka Matsuo, Yusuke Iwasawa, The University of Tokyo

Ryo Kamoi¹, Yusen Zhang¹, Nan Zhang¹, Jiawei Han², Rui Zhang¹

¹Penn State University, USA ²University of Illinois Urbana-Champaign, USA
{ryokamoi, rmz5227}@psu.edu

Learning Mathematical Topic Tree via Knowledge Distillation for Few-Shot Evaluation of LLMs

Davoudi, Pouya Pezeshkpour², Megagon Labs

agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

ENCES? EVALUATING LANGUAGE MODELS FOLLOWING INSTRUCTIONS

Natalia Tiel-Antonsen, Daniel L. Sparks, Sébastien Bubeck, Eric Horvitz, Ece Kamar, Harsha Nori

Language Models are Few-Shot Learners

Machel Reid, Yutaka Matsuo, Yusuke Iwasawa, The University of Tokyo

Ryo Kamoi¹, Yusen Zhang¹, Nan Zhang¹, Jiawei Han², Rui Zhang¹

¹Penn State University, USA ²University of Illinois Urbana-Champaign, USA
{ryokamoi, rmz5227}@psu.edu

Learning Mathematical Topic Tree via Knowledge Distillation for Few-Shot Evaluation of LLMs

Davoudi, Pouya Pezeshkpour², Megagon Labs

agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

ENCES? EVALUATING LANGUAGE MODELS FOLLOWING INSTRUCTIONS

Natalia Tiel-Antonsen, Daniel L. Sparks, Sébastien Bubeck, Eric Horvitz, Ece Kamar, Harsha Nori

Language Models are Few-Shot Learners

Machel Reid, Yutaka Matsuo, Yusuke Iwasawa, The University of Tokyo

Ryo Kamoi¹, Yusen Zhang¹, Nan Zhang¹, Jiawei Han², Rui Zhang¹

¹Penn State University, USA ²University of Illinois Urbana-Champaign, USA
{ryokamoi, rmz5227}@psu.edu

Learning Mathematical Topic Tree via Knowledge Distillation for Few-Shot Evaluation of LLMs

Davoudi, Pouya Pezeshkpour², Megagon Labs

agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

Empirical Saturnalia

DO LLMS HAVE COMMON SENSE?

Naan Tian^{*}, Sparks E. E. E., Sébastien Bubeck, Eric Horvitz, Ece Kamar, Harsha Nori

Can Large Language Models Invent Algorithms to Improve Themselves?: PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES WHEN THEY FOLLOW THE RULES?

Laura Ruis^{*}, AI Centre, UCL, Hamid Palangi, Microsoft Research

Can LLMs Learn From Mistakes? An Empirical Study on Reasoning Tasks

Shengnan An^{*†‡}, Zexiong Ma^{*§}, Siqi Cai^{*○}, Zeqi Lin^{†*}, Nanning Zheng^{†○}, Jian-Guang Lou^{†*}, Weizhu Chen^{*§}

[†]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Institute of Parallel Computing, China

Can Large Language Models Reason About Goal-Oriented Tasks?

Takeshi Kojima, The University of Tokyo, t.kojima@weblab.t.u-tokyo.ac.jp

Shixiang Sha, Google Research

Filippos Bellos, Yayuan Li, Wuao Liu, Jason J. Corso, University of Michigan, Ann Arbor, Michigan, USA, {fbellos,yayuanli,wuaoliu,jjcorso}@umich.edu

Learning Mathematical Topic Tree via LLMs

Machel Reid, Google Research*, Yutaka Matsuo, The University of Tokyo, Yusuke Iwasawa, The University of Tokyo, Ryō Kamoi¹, Yusen Zhang¹, Nan Zhang¹, Jiawei Han², Rui Zhang¹

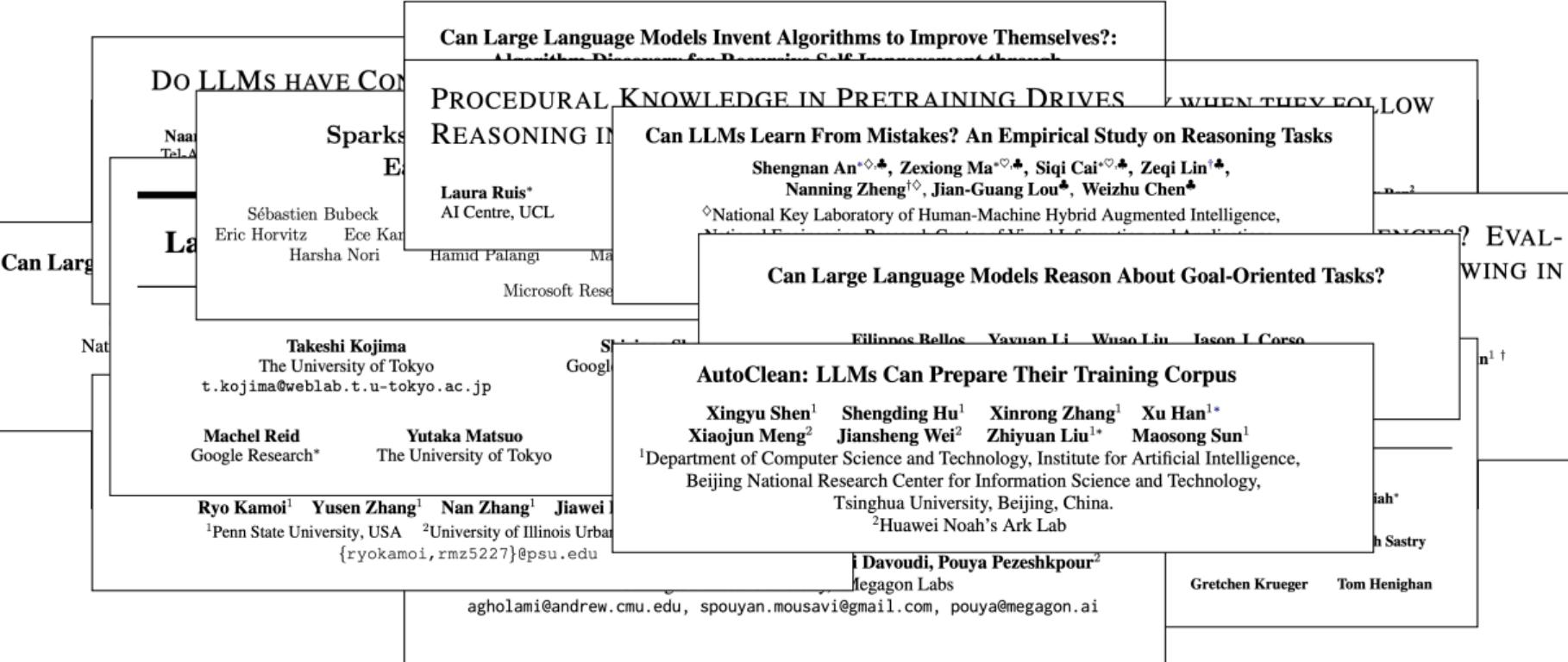
¹Penn State University, USA ²University of Illinois Urbana-Champaign, USA, {ryokamoi,rzm5227}@psu.edu

Davoudi, Pouya Pezeshkpour², Megagon Labs, agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

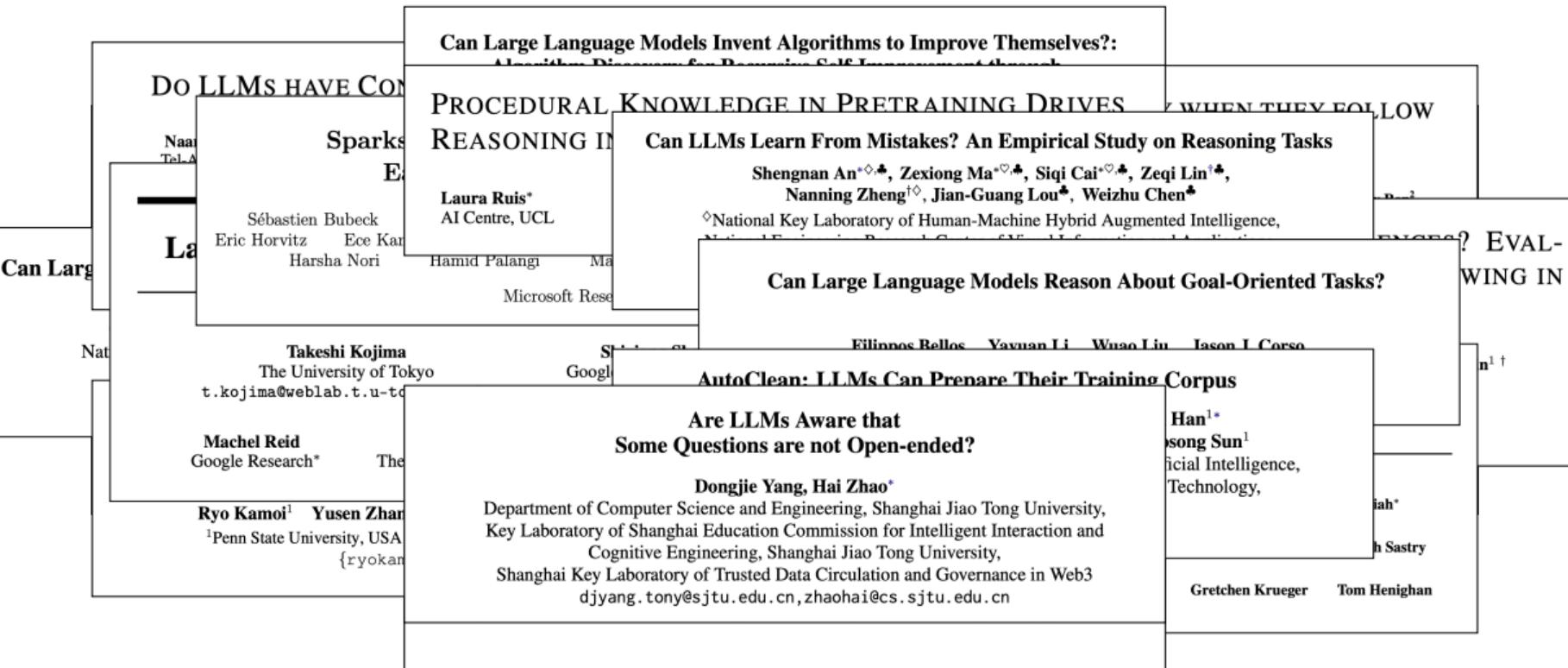
EVALUATING LANGUAGE MODELS IN A CHALLENGE ENVIRONMENT

Samuel K. Ryder*, Melanie Subbiah*, Pranav Shyam, Girish Sastry, Gretchen Krueger, Tom Henighan

Empirical Saturnalia



Empirical Saturnalia



DO LLMS HAVE CONSCIOUSNESS?

Naama Tishby^{*}, Daniel S. Sparks^{*}, Sébastien Bubeck[†], Eric Horvitz[‡], Ece Kamar[§]

Can Large Language Models Invent Algorithms to Improve Themselves?: PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES WHEN THEY FOLLOW THE LEADER?

Shengnan An^{*△▲●}, Zexiong Ma^{*○▲●}, Siqi Cai^{*○▲●}, Zeqi Lin^{†▲●}, Nanning Zheng^{†○}, Jian-Guang Lou^{‡*}, Weizhu Chen^{*○}

[△]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Tsinghua University

Can LLMs Learn From Mistakes? An Empirical Study on Reasoning Tasks

Laura Ruis^{*}, AI Centre, UCL

Do Models Reason About Goal-Oriented Tasks? (Poster)

Vayyan Li¹, Wuao Liu¹, Jason J. Corso², Han^{1*}, Song Sun¹, Ming Tang¹, Michael Sastry¹, Gretchen Krueger³, Tom Henighan³

Self-Interpretability: LLMS Can Describe Complex Internal Processes that Drive Their Decisions, and Improve with Training

Dillon Plunkett¹, Northeaster University, d.plunkett@northeastern.edu

Adam Morris², Princeton University, thatadammorris@gmail.com

Keerthi Reddy³, Independent Researcher

Jorge Morales⁴, Northeaster University

Empirical Saturnalia

Can Large Language Models Invent Algorithms to Improve Themselves? (Nan, Sparks, Bubeck, Horvitz, Kar)	DO LLMS HAVE CONSCIOUSNESS? (Sparks, Bubeck, Horvitz, Kar)	PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES REASONING IN LLMs (Laura Ruis)	Can LLMs Learn From Mistakes? An Empirical Study on Reasoning Tasks (Shengnan An, Zexiong Ma, Siqi Cai, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, Weizhu Chen)	WHEN THEY FOLLOW THE RULES (Nanning Zheng, Jian-Guang Lou, Weizhu Chen)	Are Large Language Models Reliable Judges? A Study on the Factuality Evaluation Capabilities of LLMs (Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, Shashi Bhushan TN)	EVALUATING INDEPENDENCE IN GOAL-ORIENTED TASKS (Jason J. Corso, Han Song Sun, Tongzhi Wang, Ming Tang, Tom Henighan)	Self-Interpreting Internal Processes (Dillon Plunkett, Adam Morris, Keerthi Reddy, Jorge Morales)	Corpus (Gretchen Krueger, Tom Henighan)

Can Large Language Models Invent Algorithms to Improve Themselves?: A Deep Dive into Self-Improvement in AI

DO LLMS HAVE
SELF-IMPROVEMENT ALGORITHMS?

Na

Tel. A

S

Sébastien B
Eric Horvitz

La

Can Larg

Nat

Self-Int
Internat

C

d. plu

Keerthi Reddy
Independent Researcher

Jorge Morales
Northeastern University

edu.cn

Gretchen Krueger Tom Henighan

IN THEY FOLLOW
Training Tasks

n - 2

gence,
ci

EVAL-
WING IN

Oriented Tasks?

L Corso

1. S

n¹
elligence,
ogy,

iah*

h Sastry



Outline

Introduction

Epistemological Perspectives

Theoretical Perspectives

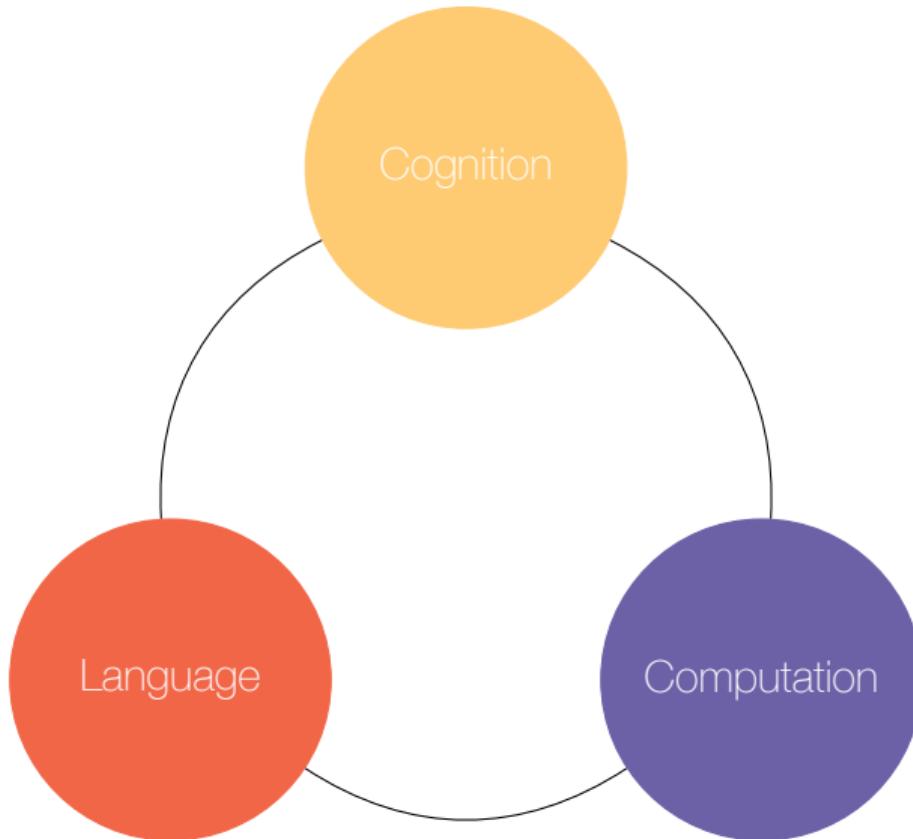
The Algebra Behind the Embeddings

The Structure Behind the Algebra

The Categories Behind the Structure

Take Aways

Chomsky's Generativist Program and the Cognitive Revolution



The New York Times

OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023



Chomsky against Abstraction in Principle

"Pick the properties that you like for a set of processors. Pick the criteria you like for success, whether in terms of performance or structure or whatever. Consider the class of all organisms, *abstracting in principle* from the existing world, that satisfy those things. And then you can ask whether they have some property of things in the material world. Do they breathe? Do they grow? Do they think? Do they talk? Do they walk? Do they enjoy themselves? Do they have moral rights?"

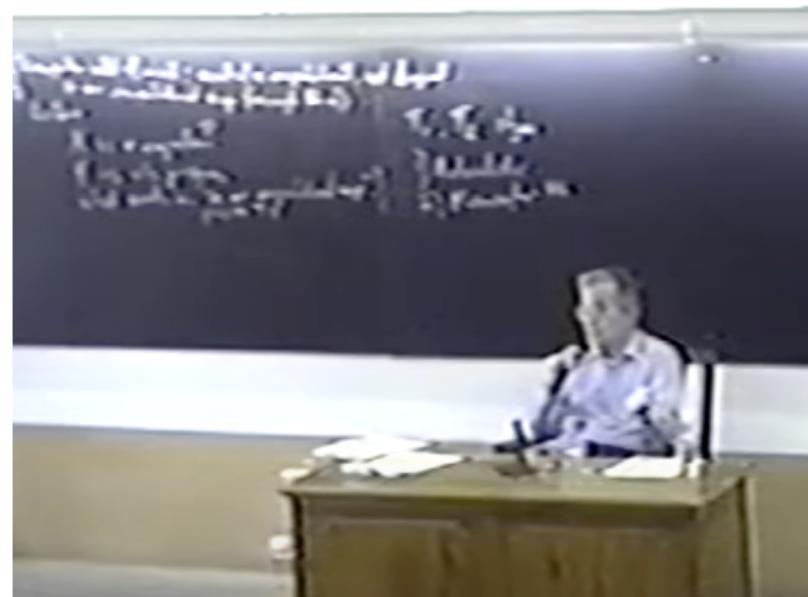
(Chomsky, 1992)



Chomsky against Abstraction in Principle

"All of these questions are stupid. And the reason they're stupid is because you've departed from naturalism. Once you've departed from naturalism, you have an algorithm for constructing stupid questions."

(Chomsky, 1992)

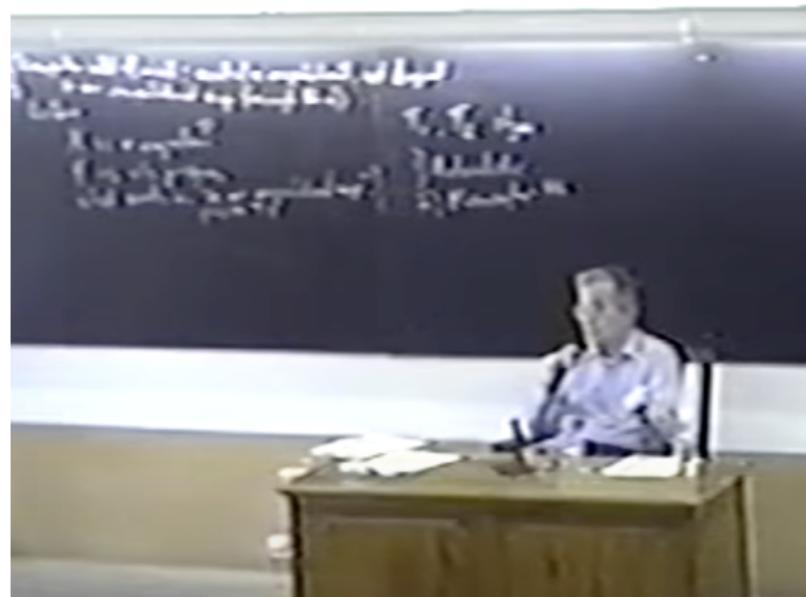


Chomsky against Abstraction in Principle

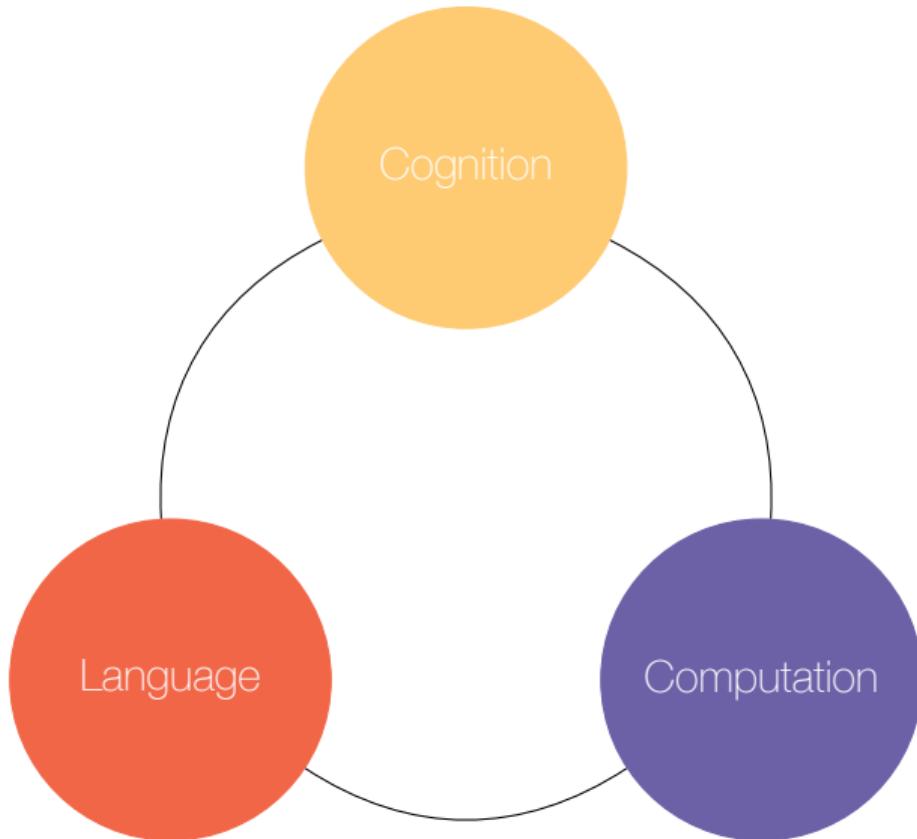
“There’s nothing wrong with principled abstraction. In fact, one might think of large areas of **mathematics** as that. **But here we have something new, principled abstraction in an empirical discipline.**”

“I don’t think we should cross that border, because **there’s no empirical claim**. It is just a question of **how to extend the metaphor.**”

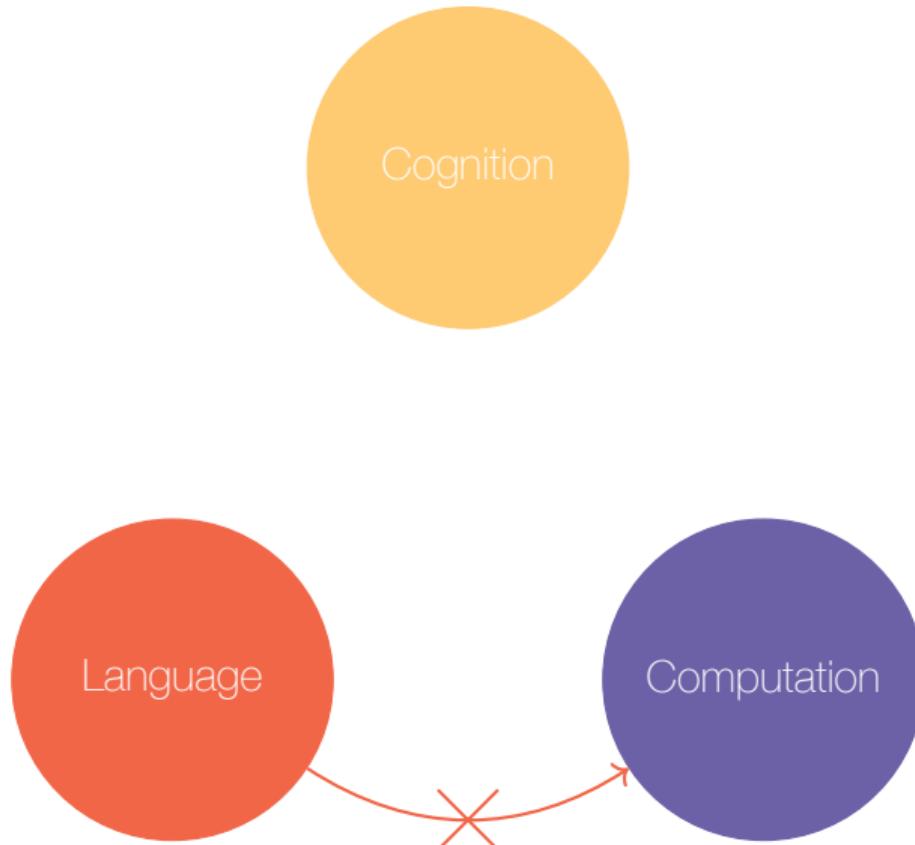
(Chomsky, 1992)



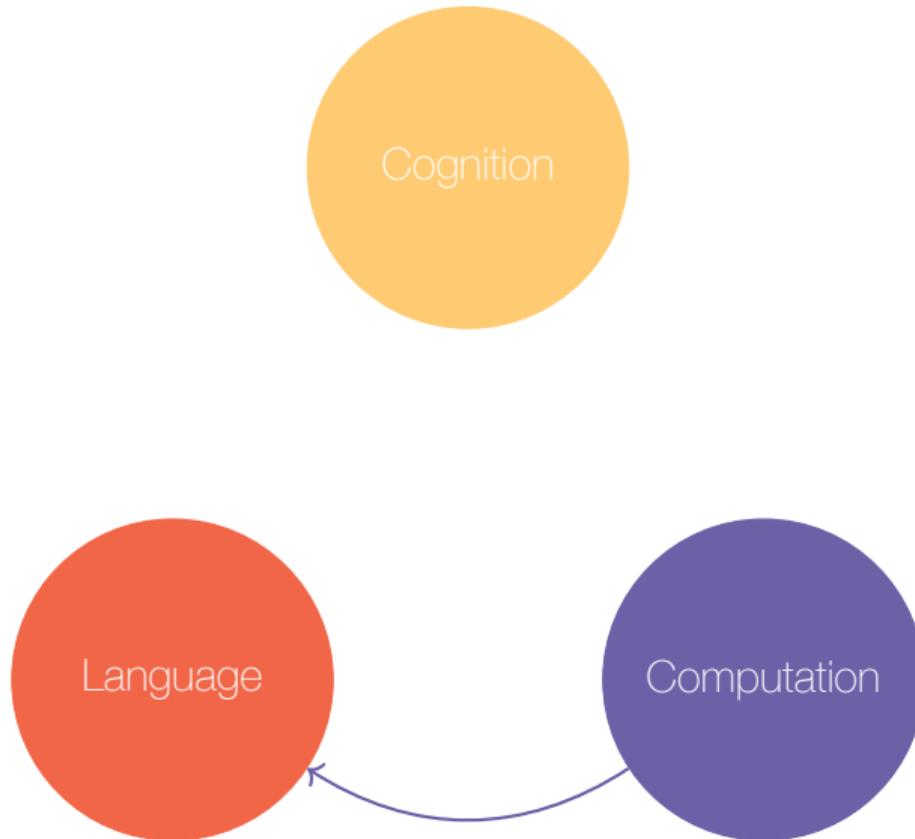
The Condition of Chomsky's Cognitive Foundations



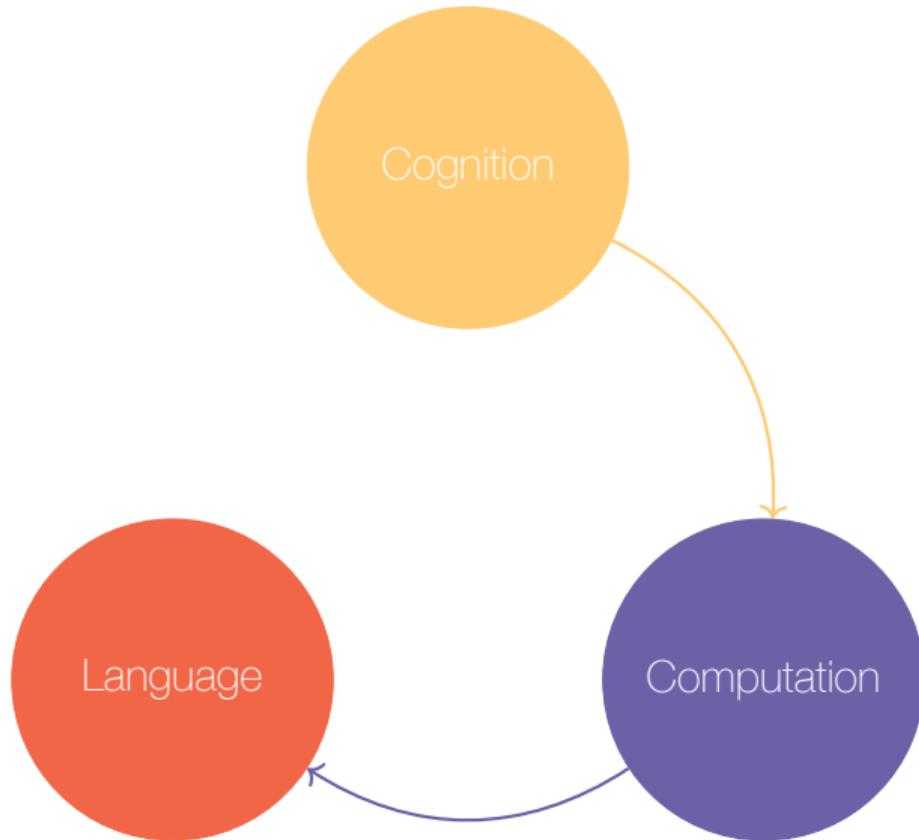
The Condition of Chomsky's Cognitive Foundations



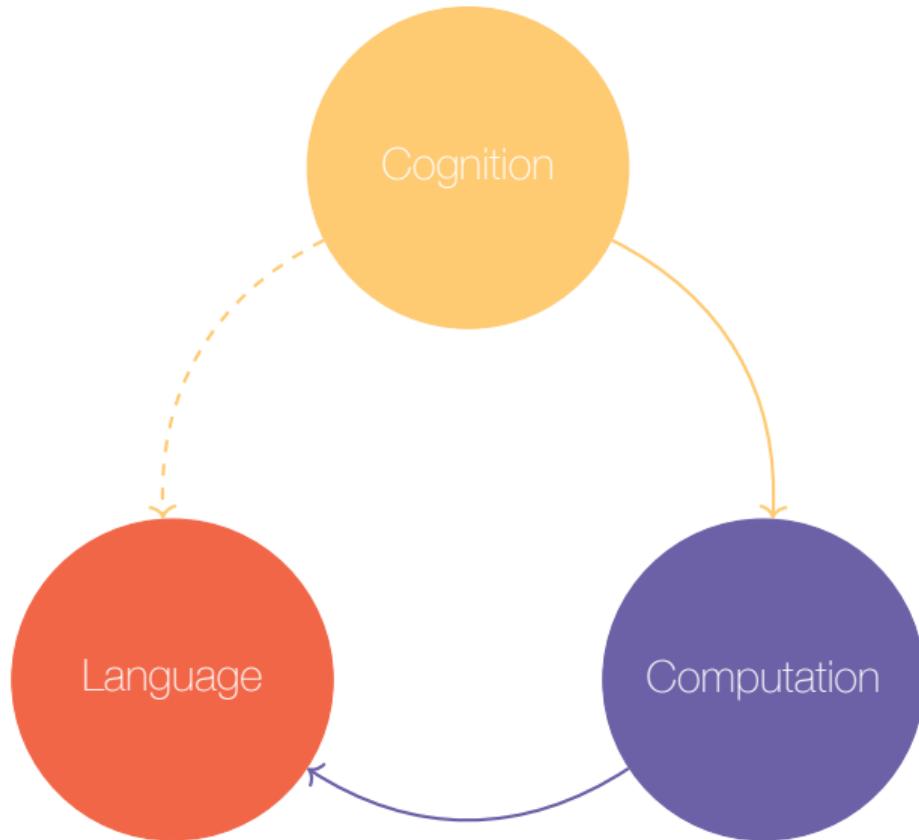
The Condition of Chomsky's Cognitive Foundations



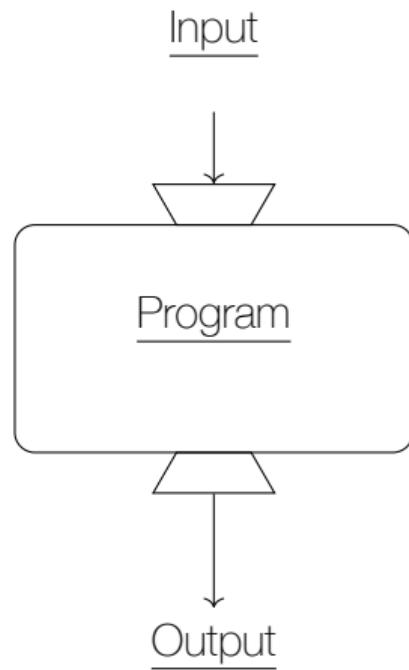
The Condition of Chomsky's Cognitive Foundations



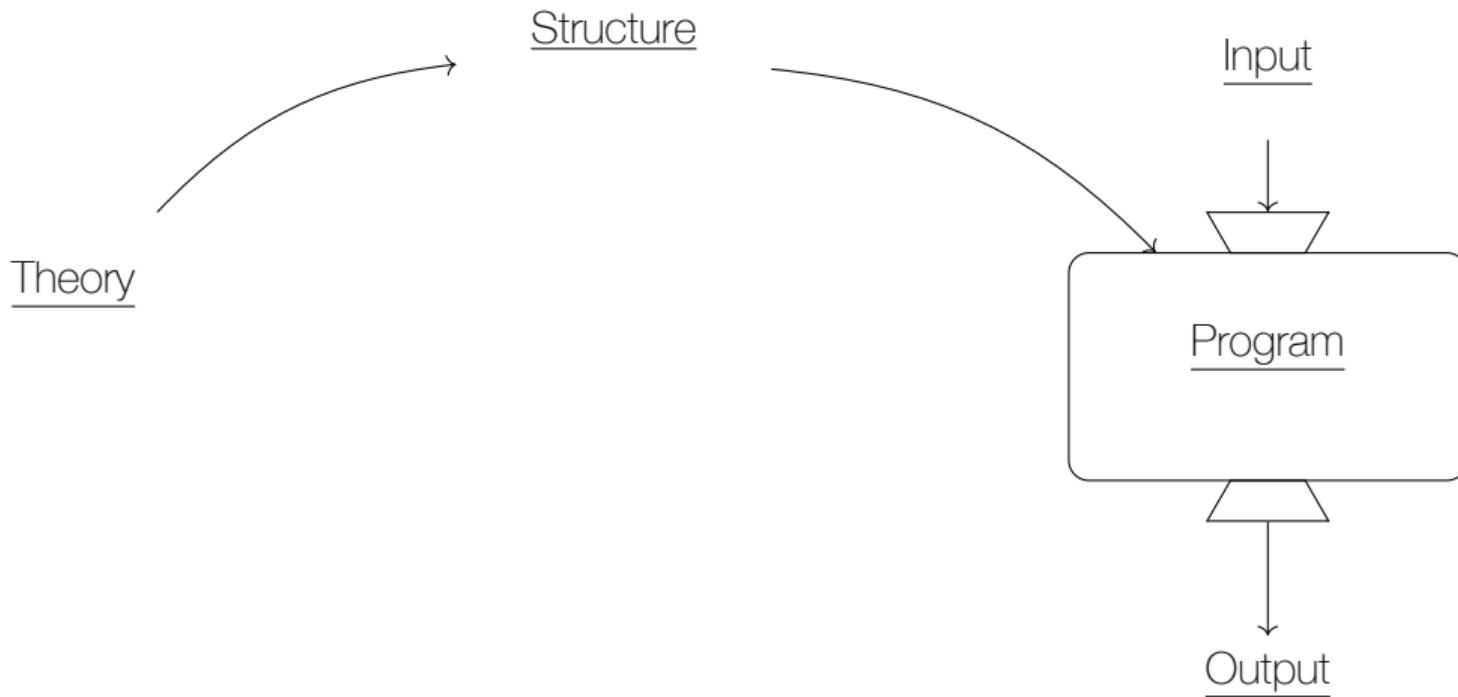
The Condition of Chomsky's Cognitive Foundations



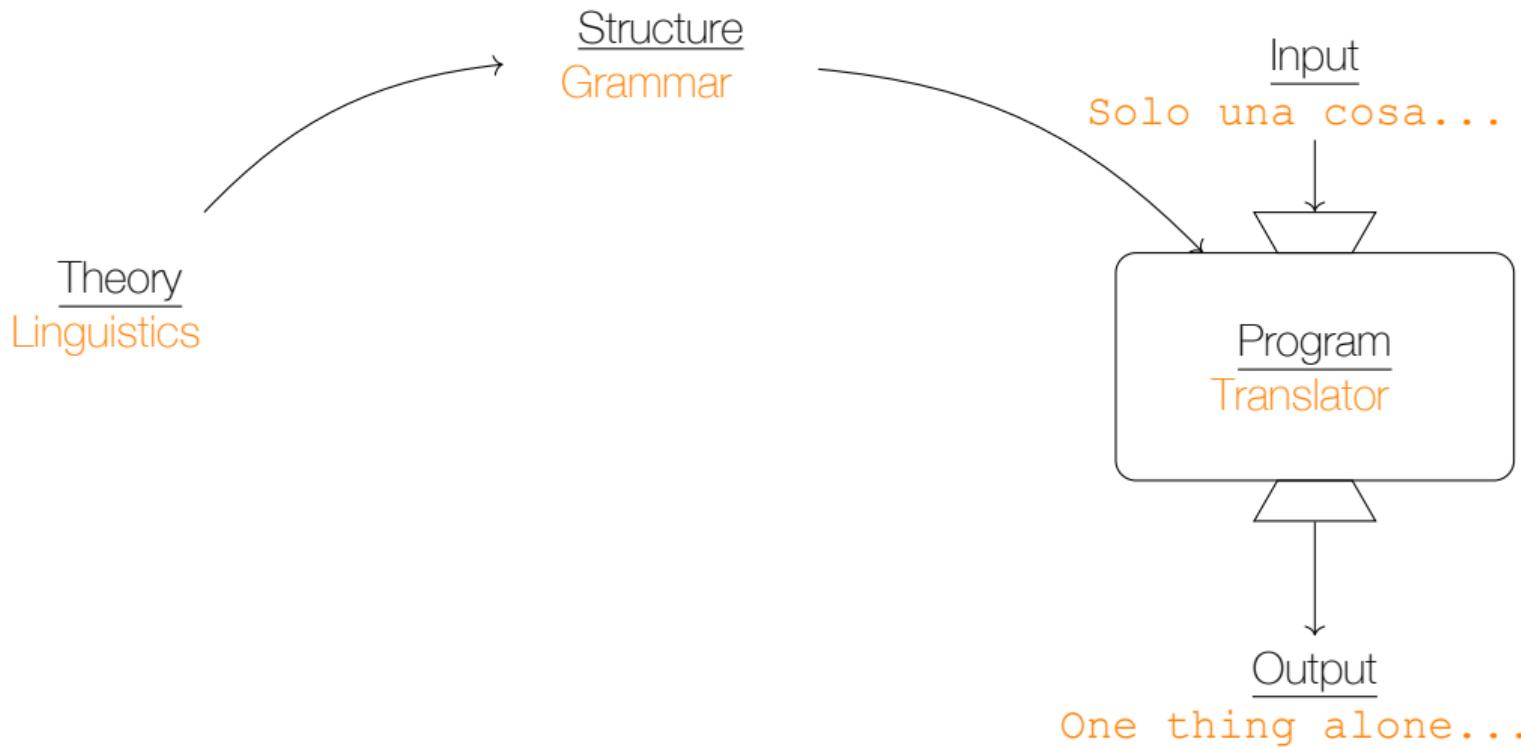
The Trap



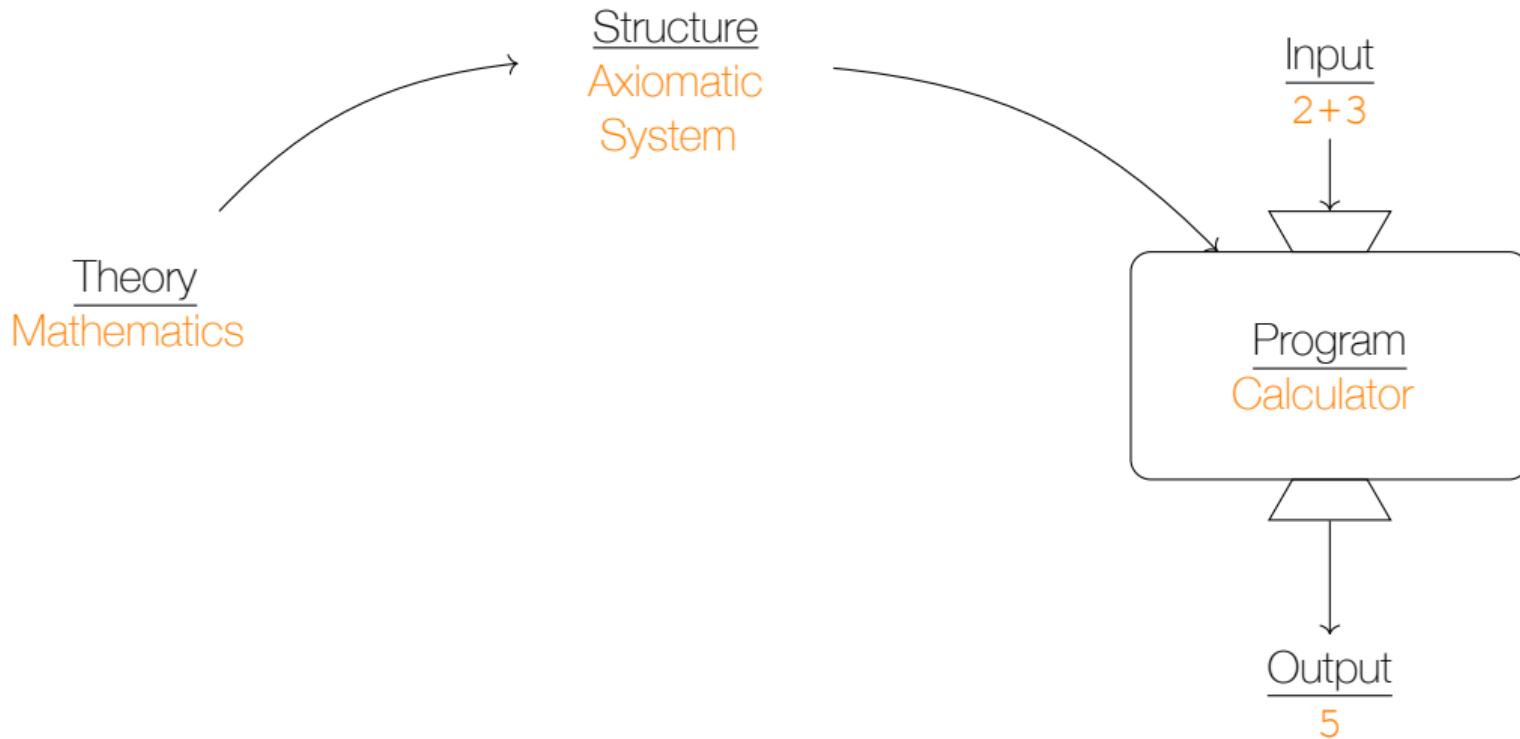
The Trap



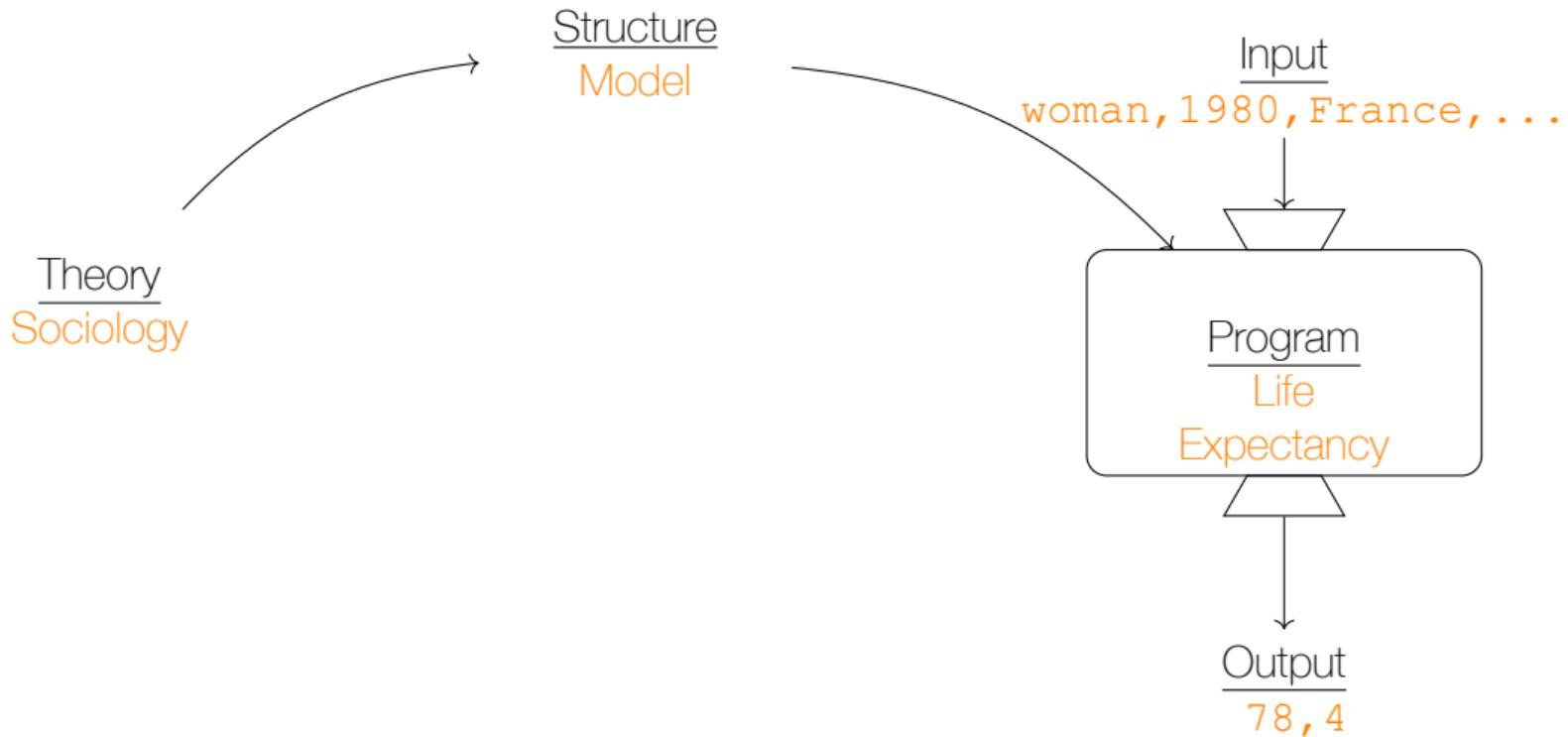
The Trap



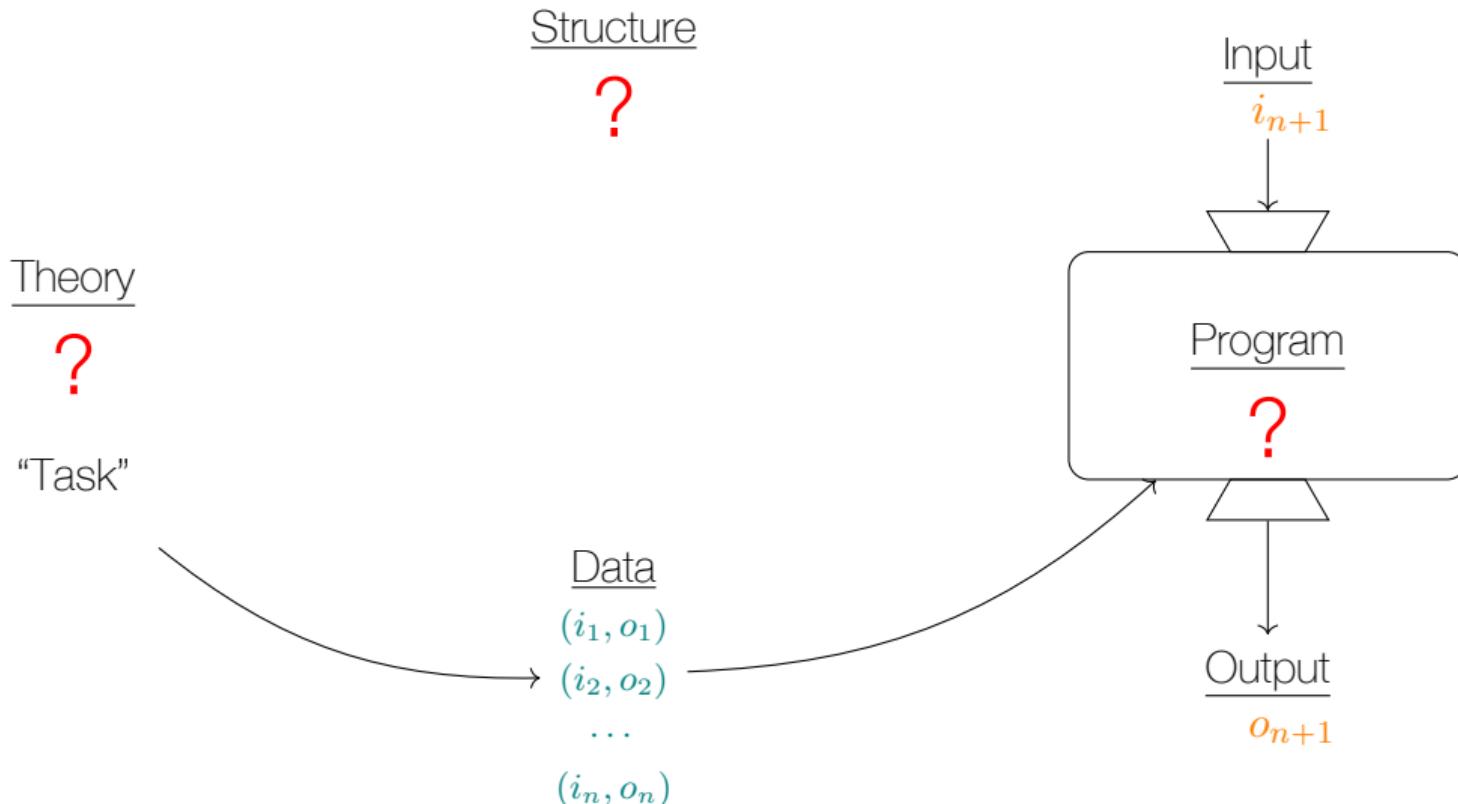
The Trap



The Trap

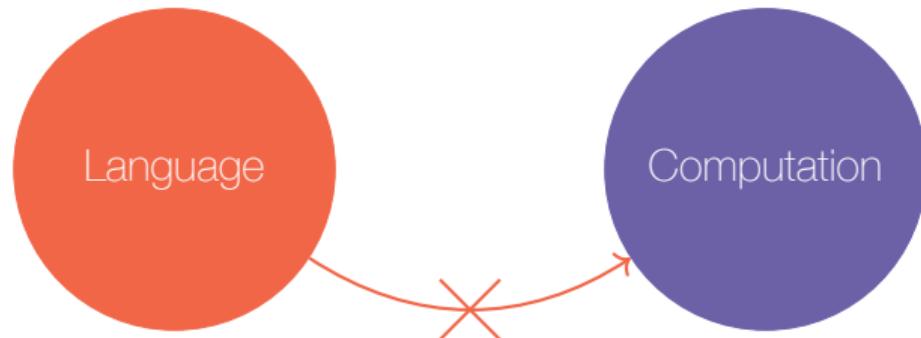


The Trap



Necessary Condition?

- ◊ Inadequacy of distributional models
(Chomsky, 1953)
- ◊ The probability of a sentence is useless
(Chomsky, 1957, 1959)
- ◊ Limited expressive power of FSAs
(Chomsky, 1956)
- ◊ Poverty of stimulus
(Chomsky, 1959)



Necessary Condition?

- ◊ Inadequacy of distributional models
(Chomsky, 1953)

Inconclusive

- ◊ The probability of a sentence is useless
(Chomsky, 1957, 1959)

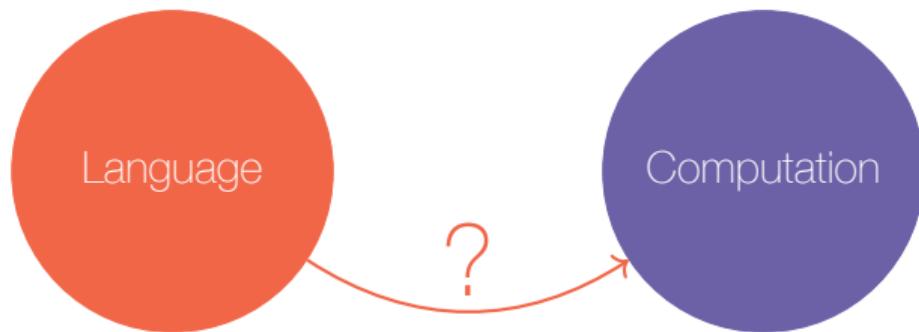
Empirically challenged

- ◊ Limited expressive power of FSAs
(Chomsky, 1956)

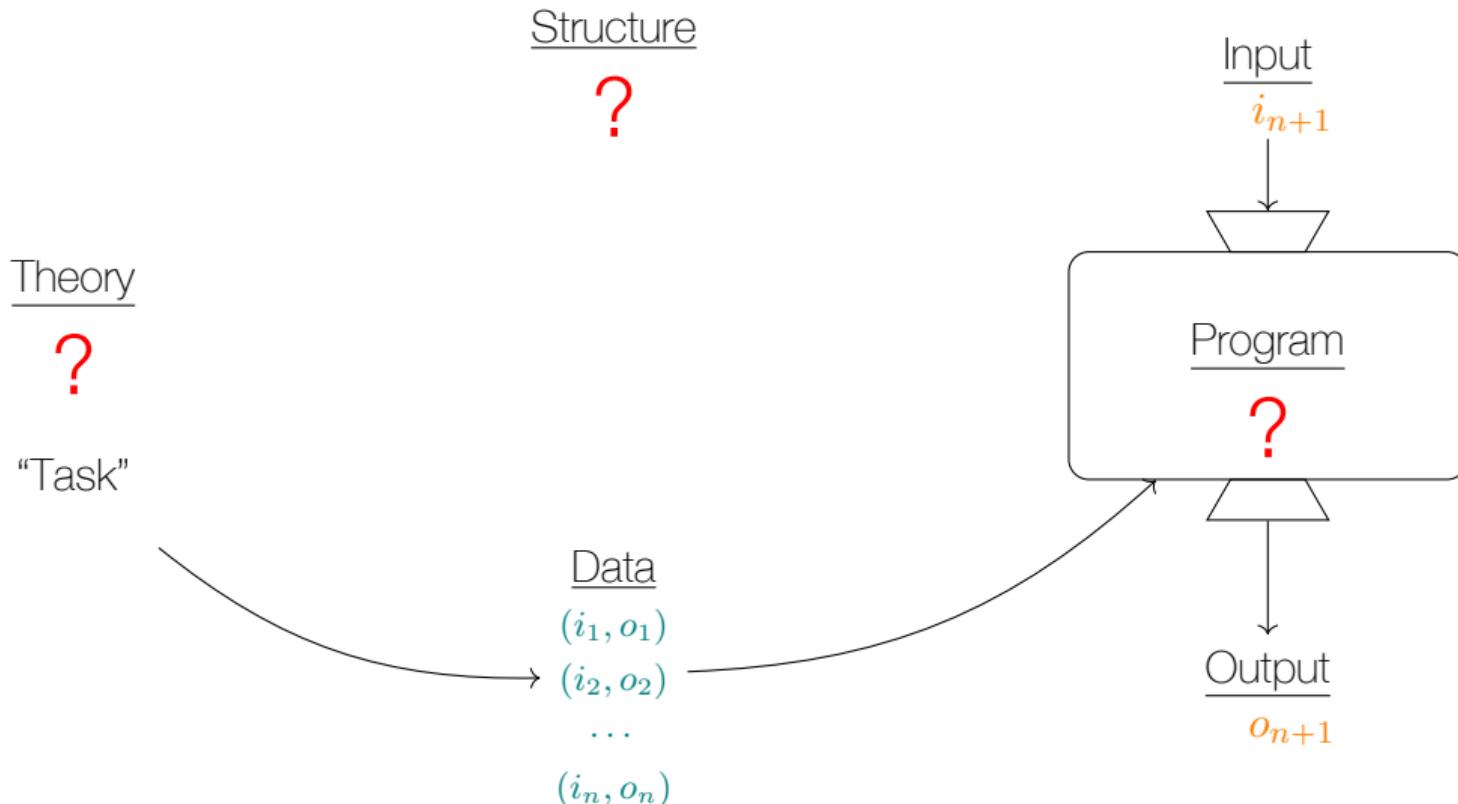
The relevance is unclear

- ◊ Poverty of stimulus
(Chomsky, 1959)

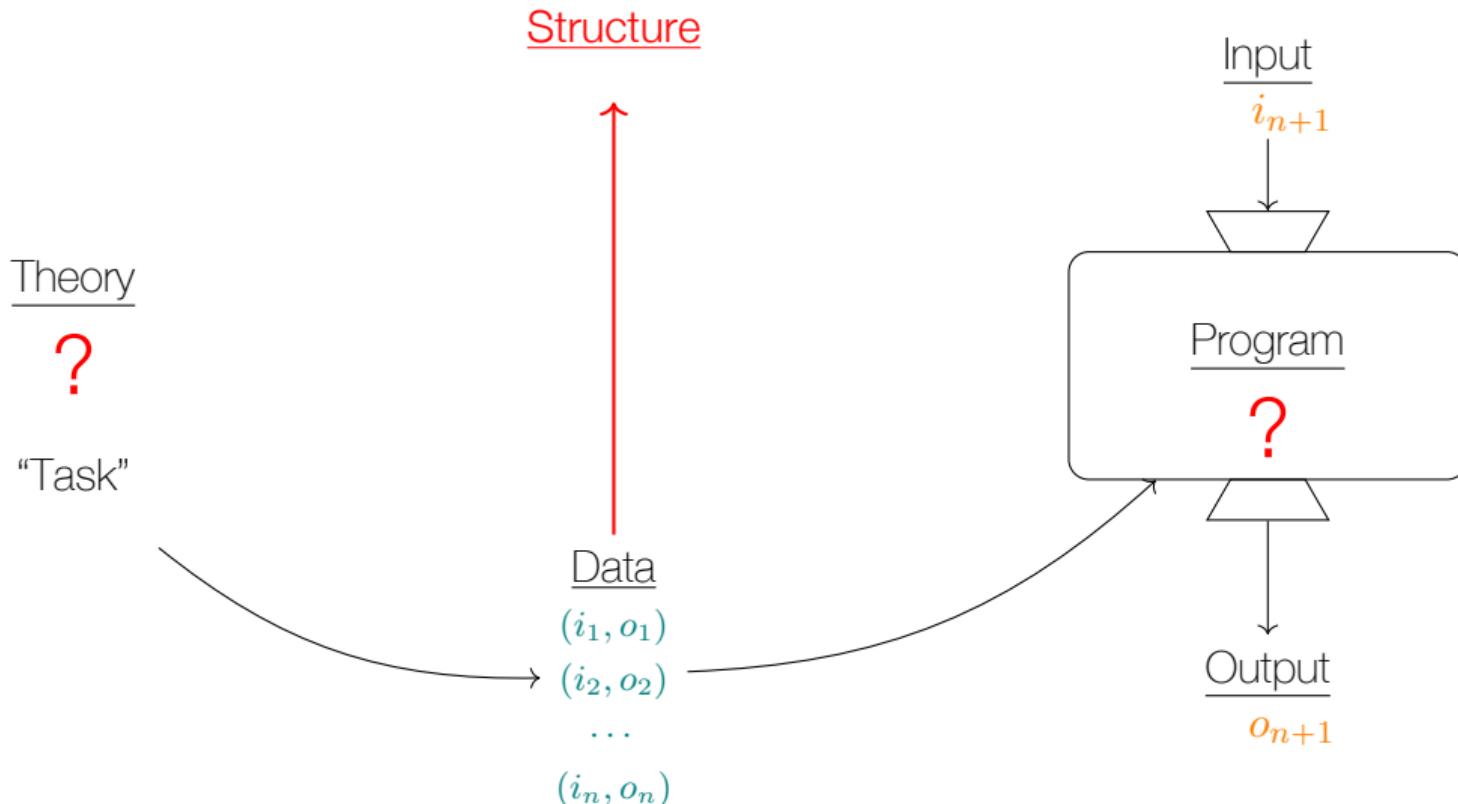
Assumes what is to be proved



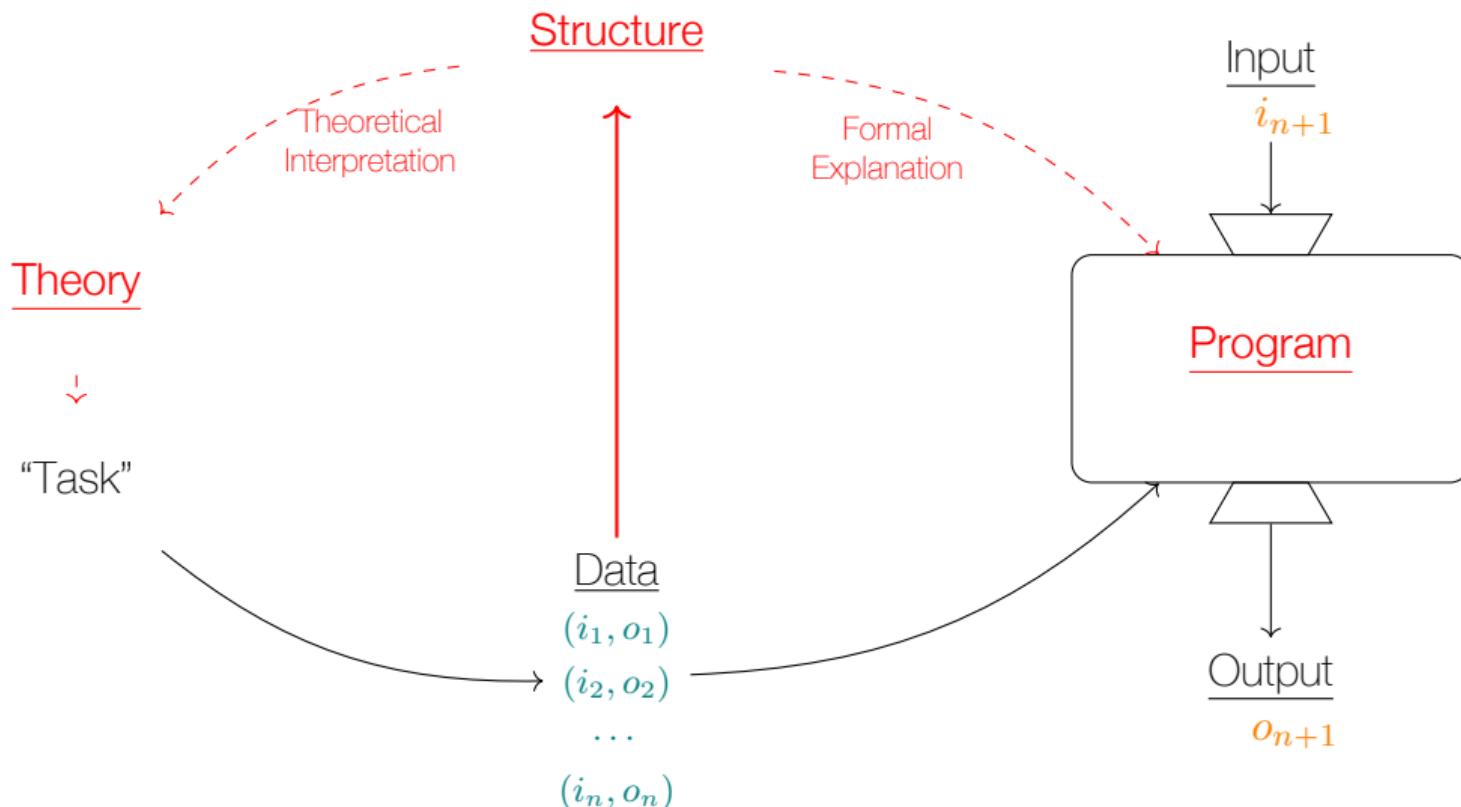
Making It Explicit



Making It Explicit



Making It Explicit



Introduction

Epistemological Perspectives

Theoretical Perspectives

The Algebra Behind the Embeddings

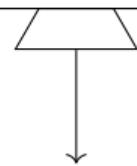
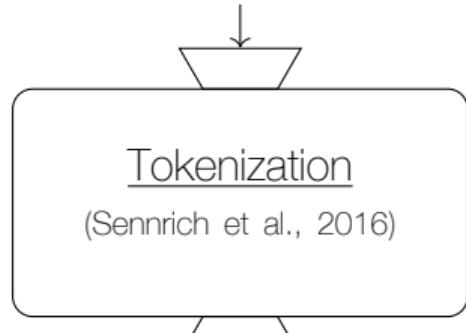
The Structure Behind the Algebra

The Categories Behind the Structure

Take Aways

Formal Explainability

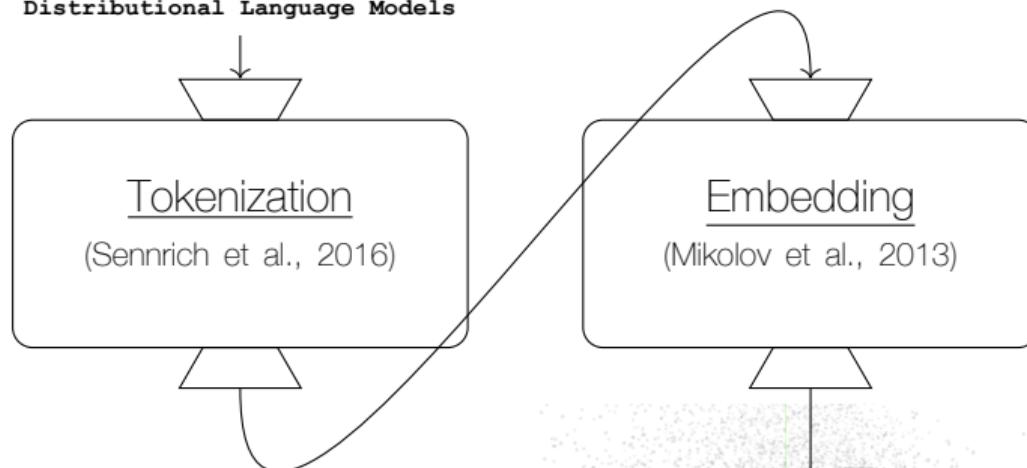
Epistemology of Machine Learning
Distributional Language Models



(<https://tiktoktokenizer.vercel.app>)

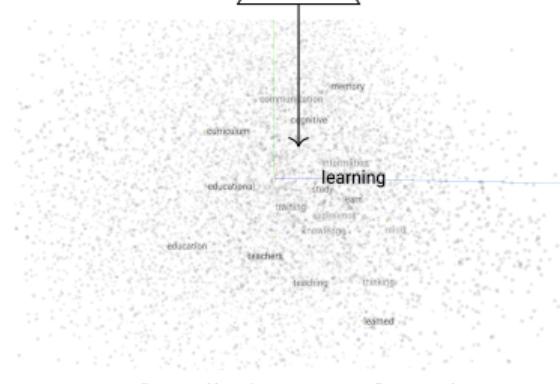
Formal Explainability

**Epistemology of Machine Learning
Distributional Language Models**



**Epistemology of Machine Learning
Distributional Language Models**

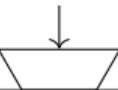
(<https://tiktoktokenizer.vercel.app>)



(<https://projector.tensorflow.org>)

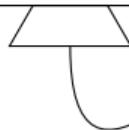
Formal Explainability

**Epistemology of Machine Learning
Distributional Language Models**



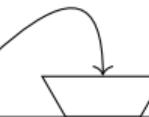
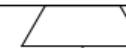
Tokenization

(Sennrich et al., 2016)



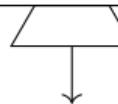
Embedding

(Mikolov et al., 2013)



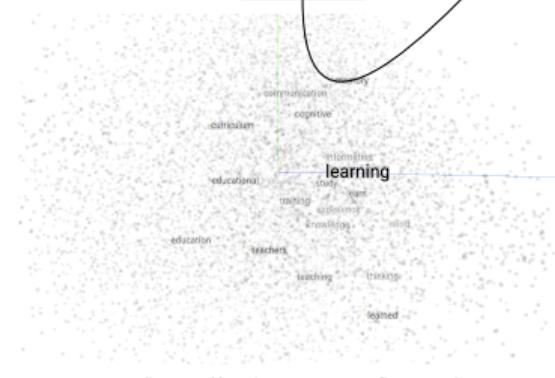
Attention

(Vaswani et al., 2017)



**Epistemology of Machine Learning
Distributional Language Models**

(<https://tiktokrizer.vercel.app>)



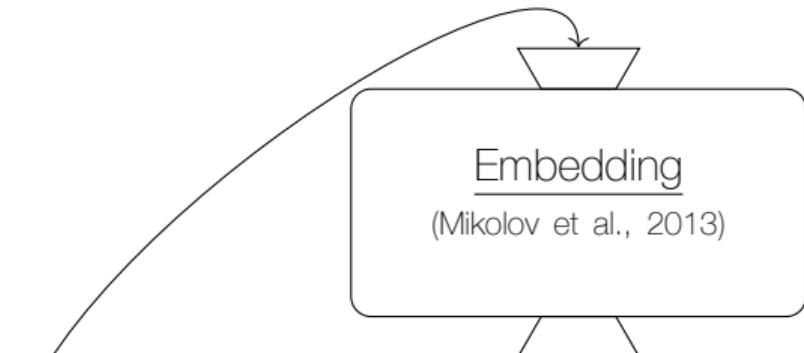
Ep
ist
em
ology
of
Machine
Learn
ing
Distribution
al
Language
Models
(<https://github.com/jessevieg/bertviz>)



Ep
ist
em
ology
of
Machine
Learn
ing
Distribution
al
Language
Models

Ep
ist
em
ology
of
Machine
Learn
ing
Distribution
al
Language
Models

Formal Explainability



**Epistemology of Machine Learning
Distributional Language Models**

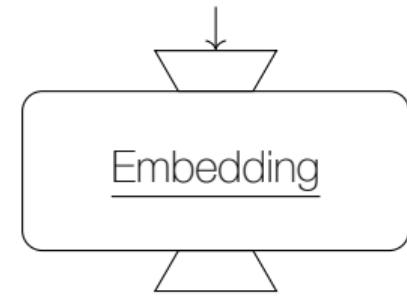
(<https://tiktoktokenizer.vercel.app>)



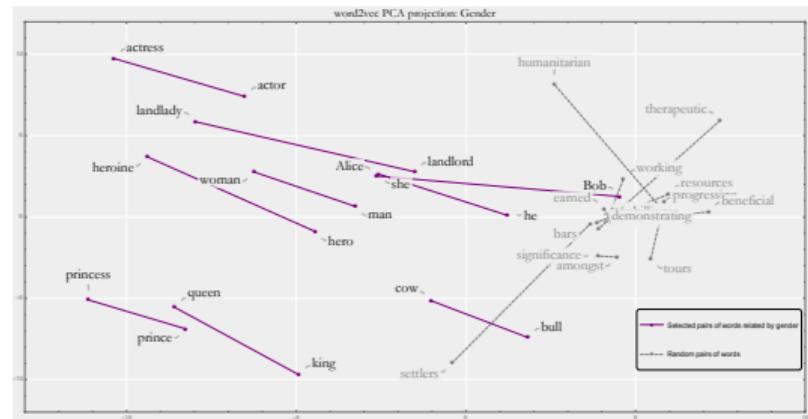
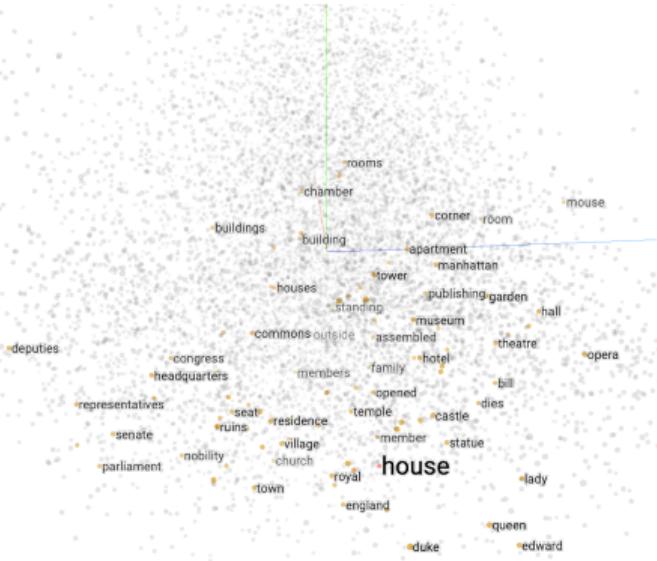
(<https://projector.tensorflow.org>)

The Structure of Embeddings

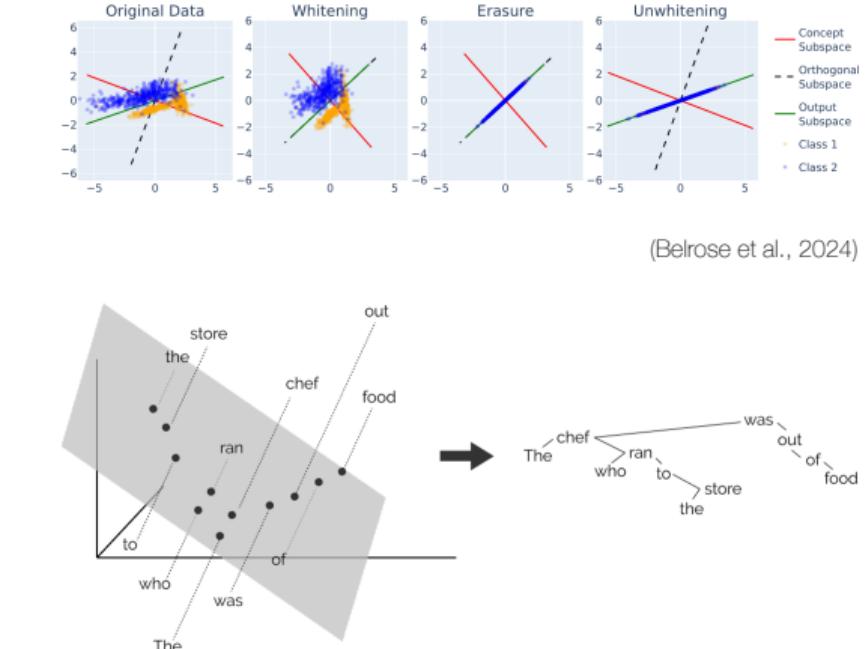
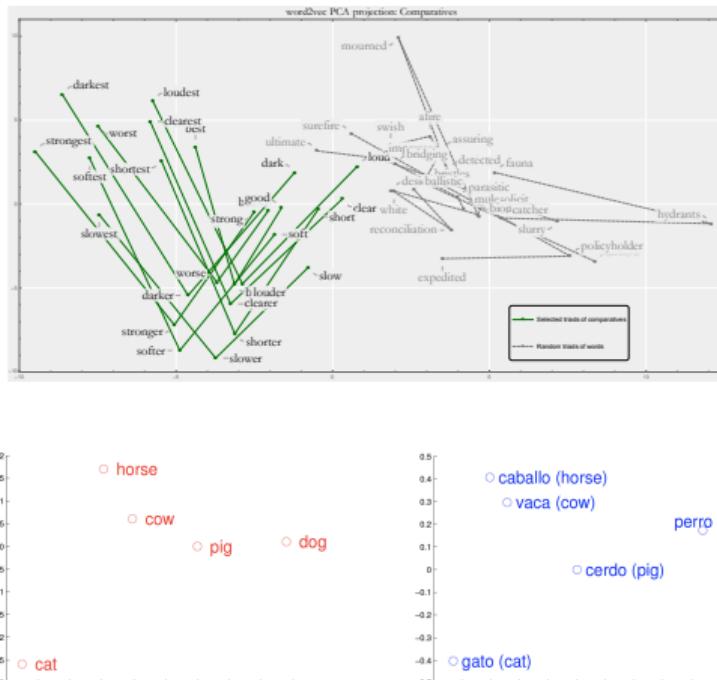
Epistemology of Machine Learning
Distributional Language Models



Embeddings: Similarity and Analogy



Embeddings: Other Applications



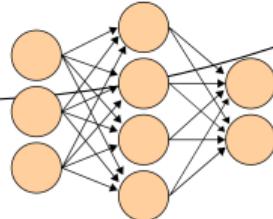
The Structure of Embeddings

Structure

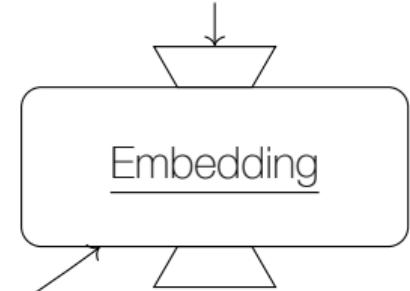
?

Data

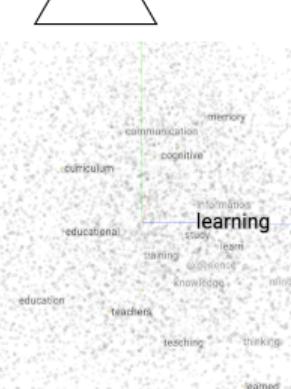
```
getElementsByClassName=function()  
{"t();{function(n,a){n.nodeType==>Element||  
(a,d,a)=>{};if(b.nodeType) return t();  
on(a){return function()<>a.type();  
return void 0=b||(a.call(),b));type.function()  
(if(e=b,apply(a[d],c),!1)=>breakValue(b);  
,b){for(var c=>b.length,=d=a.length;c>=0,  
,2),e=function(){return a.apply(b,[b,c+1]);  
}((f=b.getElementById(h))&&f.nodeType[0]>  
"")?"for(h=1,length=h-1;i>1;i--)f.innerHTML[  
""]";for(h=1,length=h-1;i>1;i--)f.innerHTML[  
""]";for(var c=>a.split(" "),c=a.length,h=0;  
on(c,s){for(var d,f=a[0],c.length,h=0;  
nodeType[0](v=b[p][[0,p]]))if(f.length>0  
||var s=j,nodeType[0][1]:o,(function()  
{var h=f[1];for(var v=a[n],da,b,v=n,i=0;  
v[n];)for(v=n);,cb=v[0],cb=v[0].value;  
v[n];)return v[n],da,b,v=n,i=0;  
v[n];)return cb.v[0].value});},class[0][0].value  
=v[n];})
```



Epistemology of Machine Learning
Distributional Language Models



Embedding

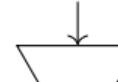


The Structure of Embeddings

Structure

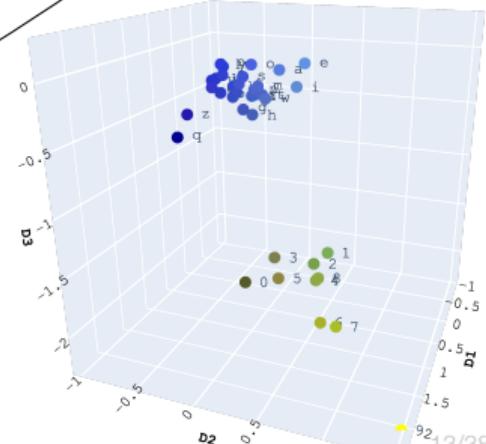
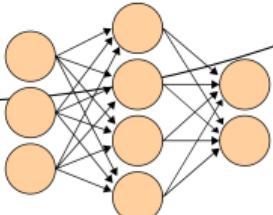
?

{-, /, 0, 1, 2, ..., 8, 9, =,
a, b, c, ..., w, x, y, z, é}



Embedding

Data



Introduction

Epistemological Perspectives

Theoretical Perspectives

The Algebra Behind the Embeddings

The Structure Behind the Algebra

The Categories Behind the Structure

Take Aways

word2vec Explained (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an **implicit**, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an **implicit, low-dimensional** factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an **implicit, low-dimensional factorization** of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an **implicit, low-dimensional factorization** of a **pointwise mutual information (pmi)**, word-context matrix.

word2vec Explained (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an **implicit, low-dimensional factorization** of a **pointwise mutual information (pmi), word-context** matrix.

word2vec Explained (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an **implicit, low-dimensional factorization** of a **pointwise mutual information (pmi), word-context matrix.**

word2vec Explained (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

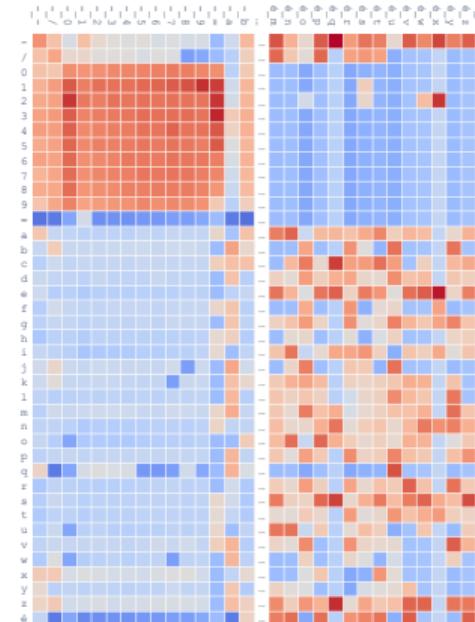
- ◊ Word2vec performs an **implicit, low-dimensional factorization** of a **pointwise mutual information (pmi)**, word-context matrix.
- ◊ The **Singular Value Decomposition (SVD)** provides an **exact solution** to this problem.

Example: Characters in Wikipedia

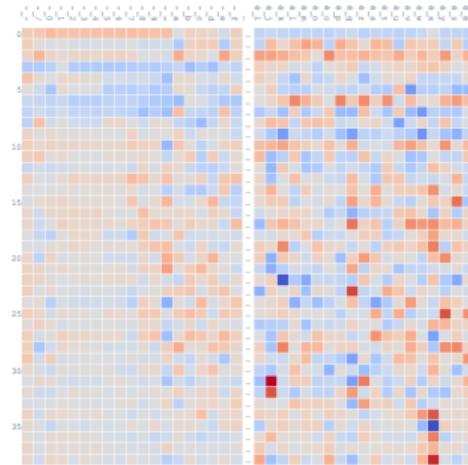
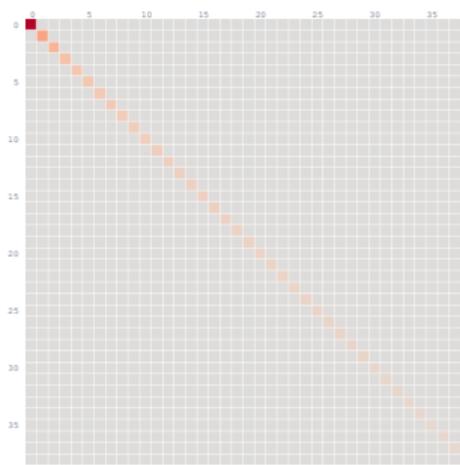
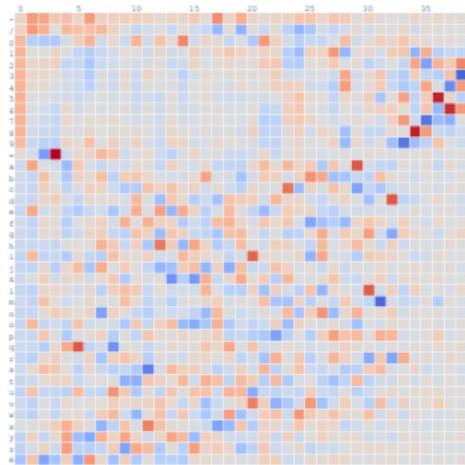
$$W = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, a, b, c, \dots, w, x, y, z, é\}$$

$$C = X \times X = \{ (-, -), (-, /), (-, 0), \dots, (é, z), (é, é) \}$$

$$\begin{aligned} M_{wc} &= \text{pmi}(w, c) \\ &= \log \frac{p(w, c)}{p(w)p(c)} \end{aligned}$$

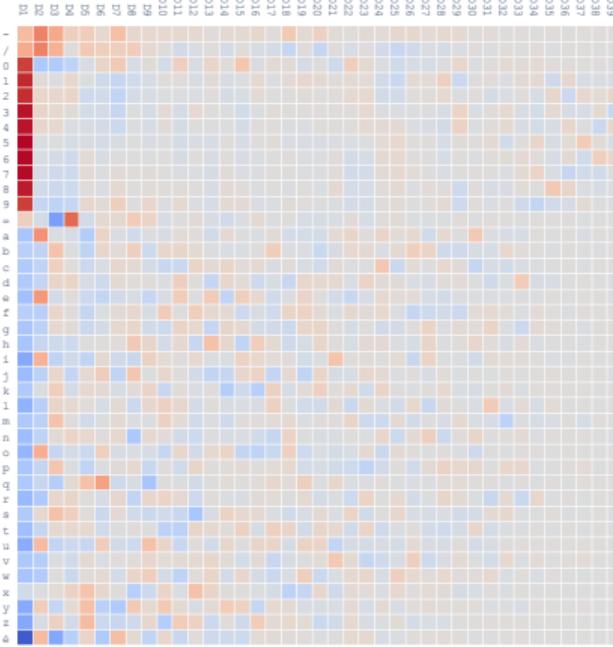


SVD of Wikipedia Character PMI Matrix

 U Σ V^T 

Truncate

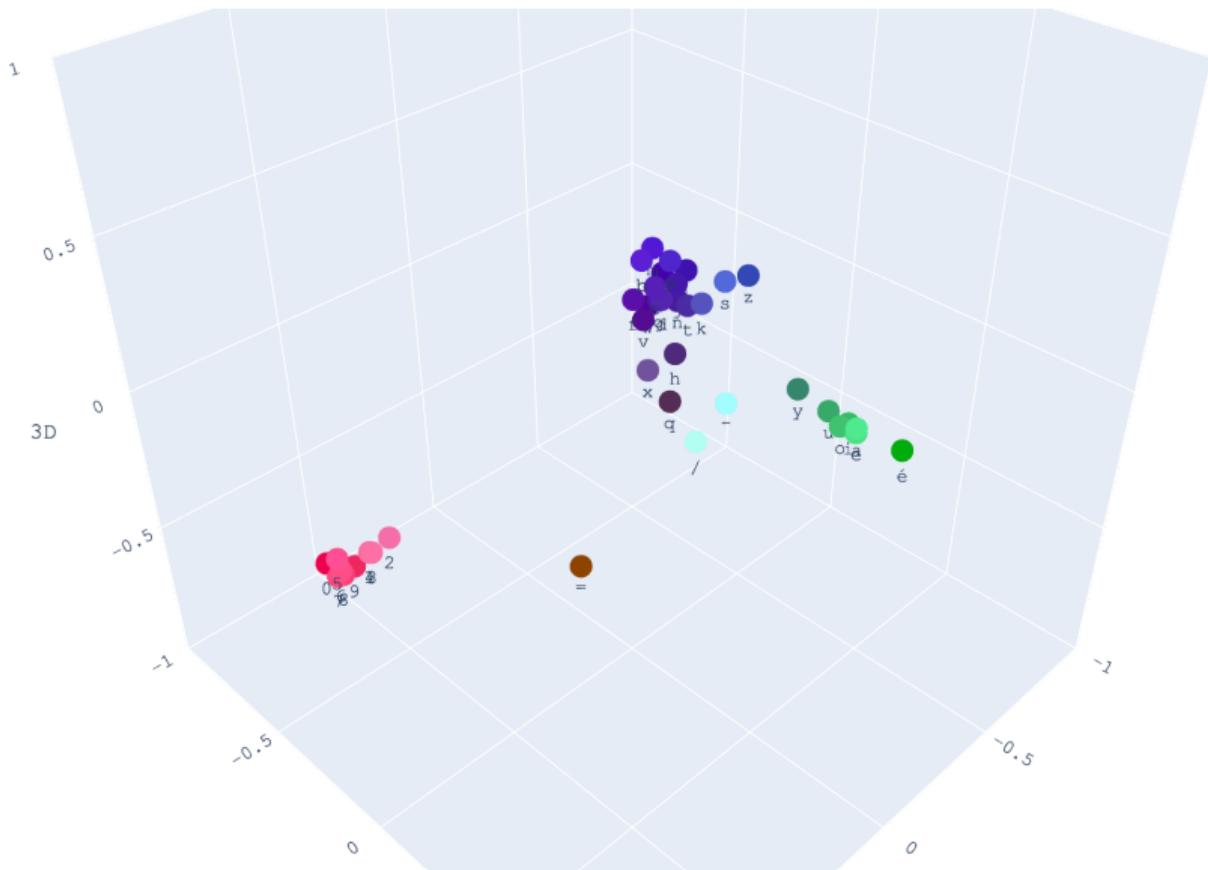
$$U \times \Sigma$$



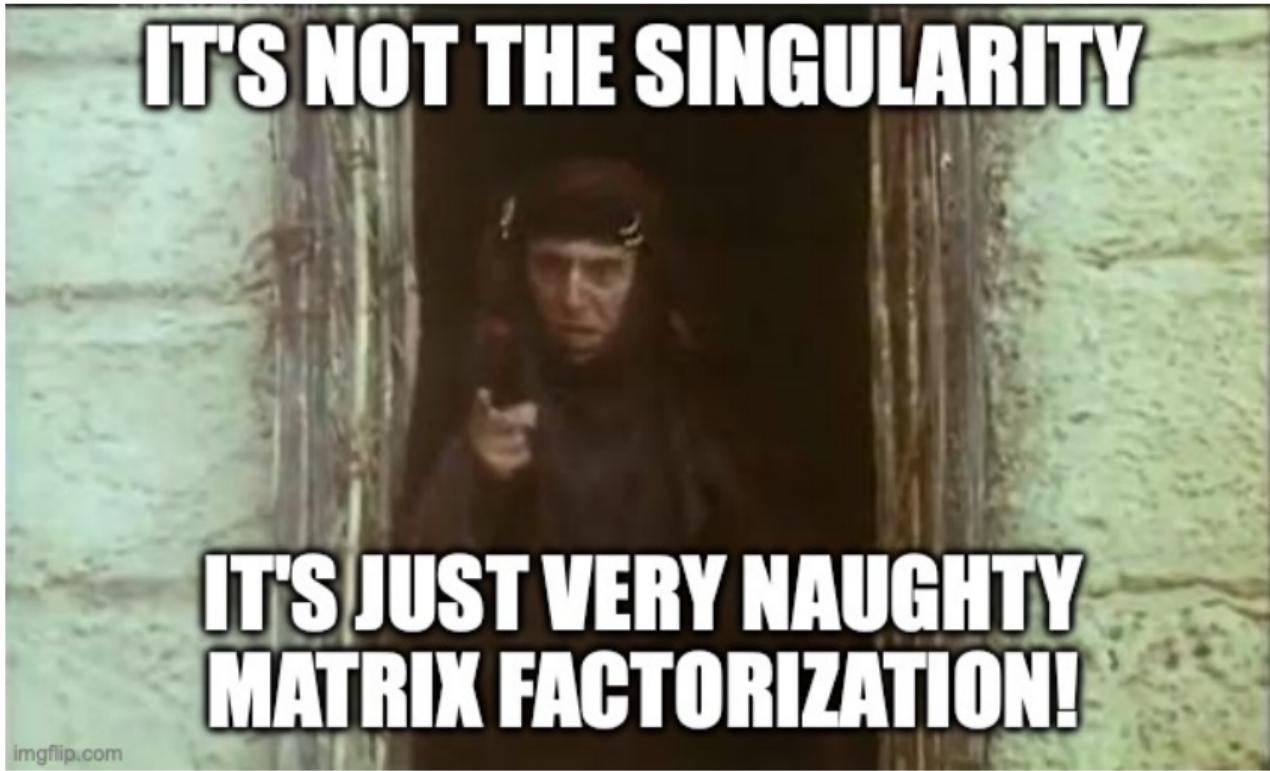
Truncate

$$\hat{U} \times \hat{\Sigma}$$



$\hat{U} \times \hat{\Sigma}$ 

What to conclude?

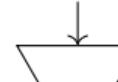


The Structure of Embeddings

Structure

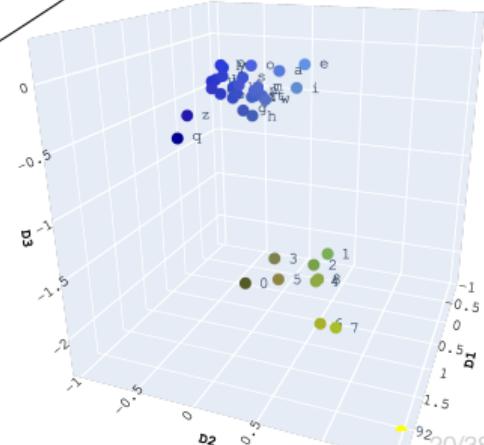
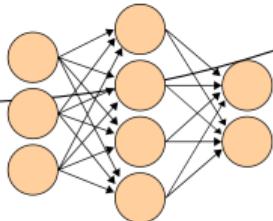
?

{-, /, 0, 1, 2, ..., 8, 9, =,
a, b, c, ..., w, x, y, z, é}



Embedding

Data

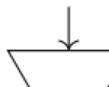


The Structure of Embeddings

Structure

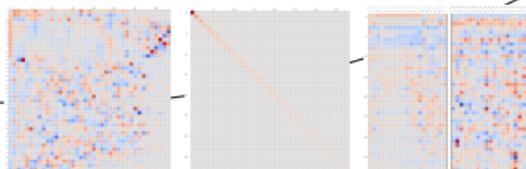


$\{-, /, 0, 1, 2, \dots, 8, 9, =,$
 $a, b, c, \dots, w, x, y, z, \acute{e}\}$

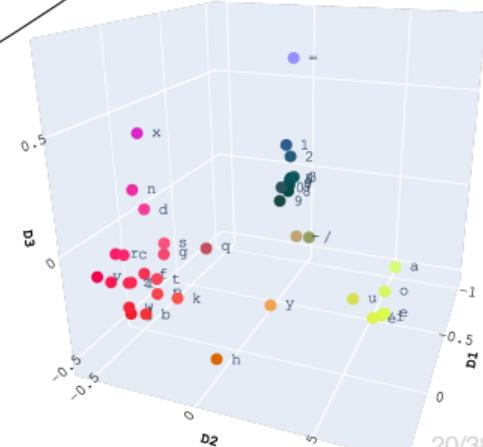


Embedding

Data



SVD



4 Why does this produce good word representations?

Good question. We don't really know.

The distributional hypothesis states that words in similar contexts have similar meanings. The objective above clearly tries to increase the quantity $v_w \cdot v_c$ for good word-context pairs, and decrease it for bad ones. Intuitively, this means that words that share many contexts will be similar to each other (note also that contexts sharing many words will also be similar to each other). This is, however, very hand-wavy.

Can we make this intuition more precise? We'd really like to see something more formal.

(Goldberg and Levy, 2014)

Introduction

Epistemological Perspectives

Theoretical Perspectives

The Algebra Behind the Embeddings

The Structure Behind the Algebra

The Categories Behind the Structure

Take Aways

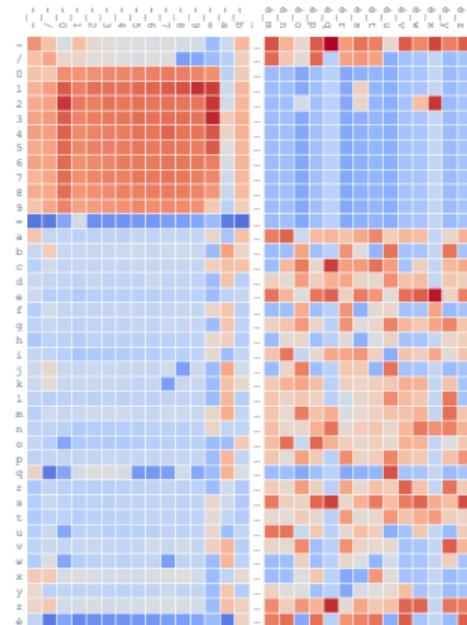
Embeddings as Functions Over Sets

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{ (-, -), (-, /), (-, 0), \dots, (\text{é}, z), (\text{é}, \text{é}) \}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$



Embeddings as Functions Over Sets

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

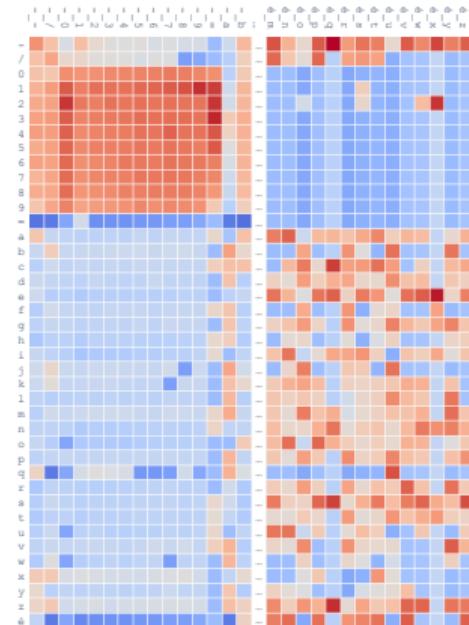
$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, z), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto \textcolor{blue}{M}(x, -)$$



Embeddings as Functions Over Sets

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, z), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

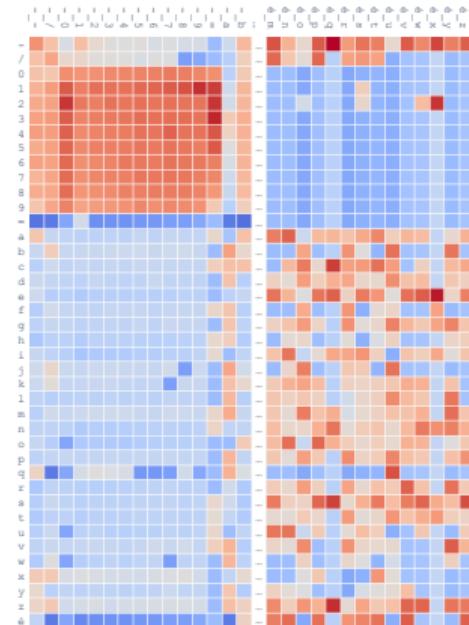
$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto \textcolor{blue}{M}(x, -)$$

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

$$y \mapsto \textcolor{red}{M}(-, y)$$



Embeddings as Functions Over Sets

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, z), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$\textcolor{red}{X} \xrightarrow{M_x} \mathbb{R}^{\textcolor{blue}{Y}}$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto \textcolor{blue}{M}(x, -)$$

$$\mathbb{R}^{\textcolor{red}{X}} \xleftarrow{M_y} \textcolor{blue}{Y}$$

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

$$\textcolor{blue}{y} \mapsto \textcolor{red}{M}(-, y)$$

Embeddings as Functions Over Sets

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, z), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto \textcolor{blue}{M}(x, -)$$

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

$$\textcolor{blue}{y} \mapsto \textcolor{red}{M}(-, y)$$

$$\begin{array}{ccc} \textcolor{red}{X} & \xrightarrow{M_x} & \mathbb{R}^{\textcolor{blue}{Y}} \\ \downarrow & & \uparrow \\ \mathbb{R}^{\textcolor{red}{X}} & \xleftarrow{M_y} & \textcolor{blue}{Y} \end{array}$$

Embeddings as Functions Over Sets

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, z), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

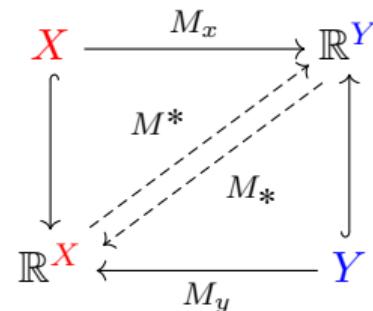
$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto \textcolor{blue}{M}(x, -)$$

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

$$y \mapsto \textcolor{red}{M}(-, y)$$

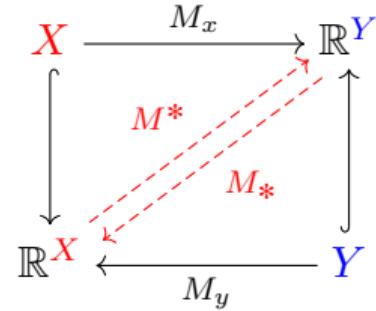


$$M^*: \mathbb{R}^{\textcolor{red}{X}} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$M_*: \mathbb{R}^{\textcolor{blue}{Y}} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

Embeddings as Functions Over Sets

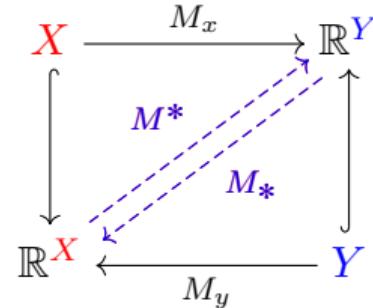
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$



Embeddings as Functions Over Sets

$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

$$M^* M_* : \mathbb{R}^Y \rightarrow \mathbb{R}^Y$$



Embeddings as Functions Over Sets

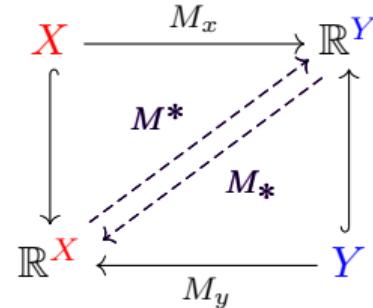
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

$$M^* M_* : \mathbb{R}^Y \rightarrow \mathbb{R}^Y$$

$$\{u_1, \dots, u_m\} \subset \mathbb{R}^X$$

$$\{v_1, \dots, v_n\} \subset \mathbb{R}^Y$$

$$\{\lambda_1, \dots, \lambda_{\min(m,n)}, 0, \dots, 0\}$$



Embeddings as Functions Over Sets

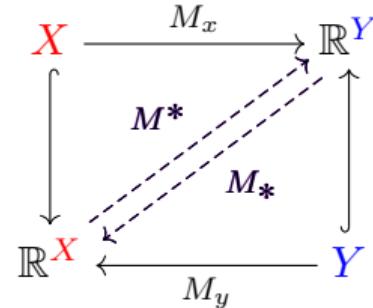
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

$$M^* M_* : \mathbb{R}^Y \rightarrow \mathbb{R}^Y$$

$$\{u_1, \dots, u_m\} \subset \mathbb{R}^X$$

$$\{v_1, \dots, v_n\} \subset \mathbb{R}^Y$$

$$\{\lambda_1, \dots, \lambda_{\min(m,n)}, 0, \dots, 0\}$$



$$U := [\color{red}{u_1}, \dots, \color{red}{u_m}]$$

$$M = U \Sigma V^T \quad V := [\color{blue}{v_1}, \dots, \color{blue}{v_n}]$$

$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

Embeddings as Functions Over Sets

$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

$$M^* M_* : \mathbb{R}^Y \rightarrow \mathbb{R}^Y$$

$$\{u_1, \dots, u_m\} \subset \mathbb{R}^X$$

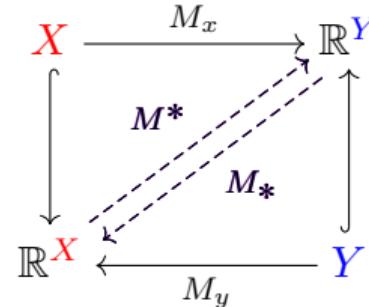
$$\{v_1, \dots, v_n\} \subset \mathbb{R}^Y$$

$$\{\lambda_1, \dots, \lambda_{\min(m,n)}, 0, \dots, 0\}$$

$$M_* M^* u_i = \lambda_i u_i$$

$$M^* M_* v_i = \lambda_i v_i$$

The u_i and v_i are (linear)
fixed points!

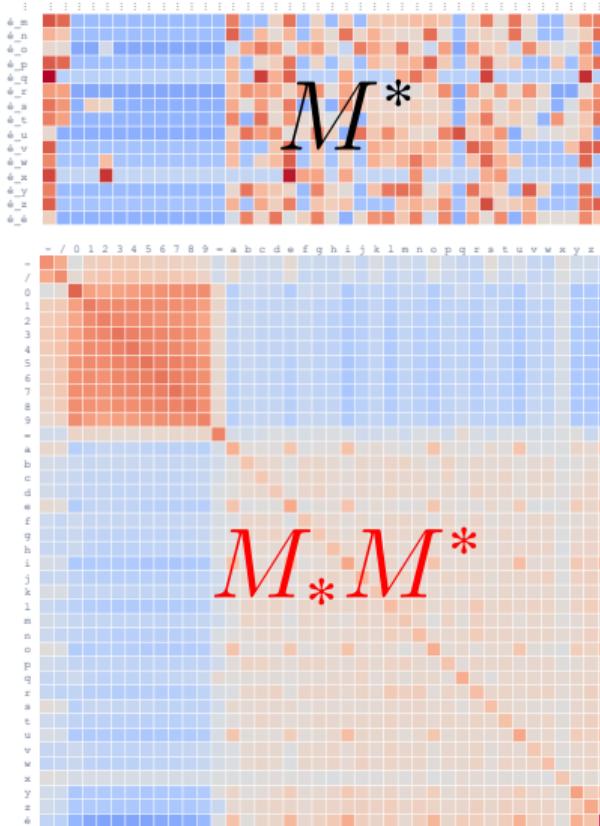
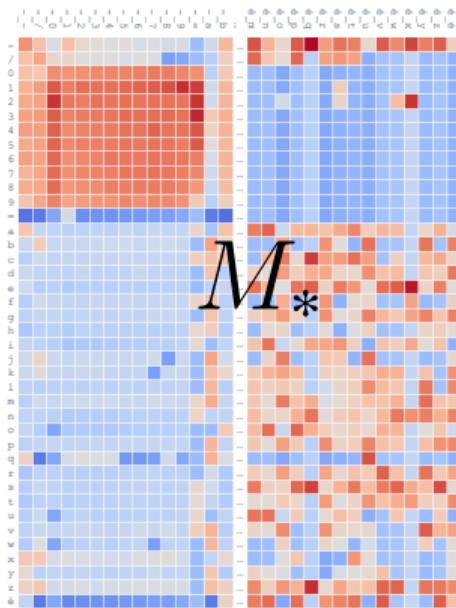


$$U := [\underline{u_1}, \dots, \underline{u_m}]$$

$$M = U \Sigma V^T \quad V := [\underline{v_1}, \dots, \underline{v_n}]$$

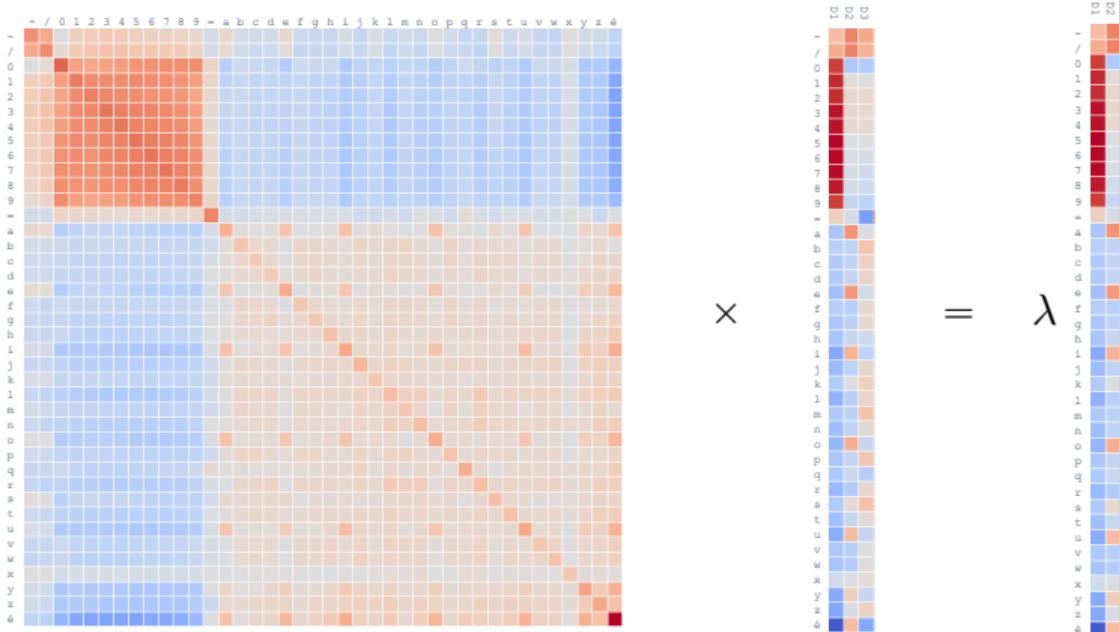
$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

$M_* M^*$ as a Covariance Matrix



Eigenvectors as Fixed Points

$$M_* M^* u = \lambda u$$

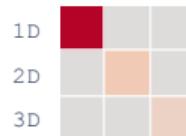


Structural Features

Eigenvectors of $M_* M^*$:



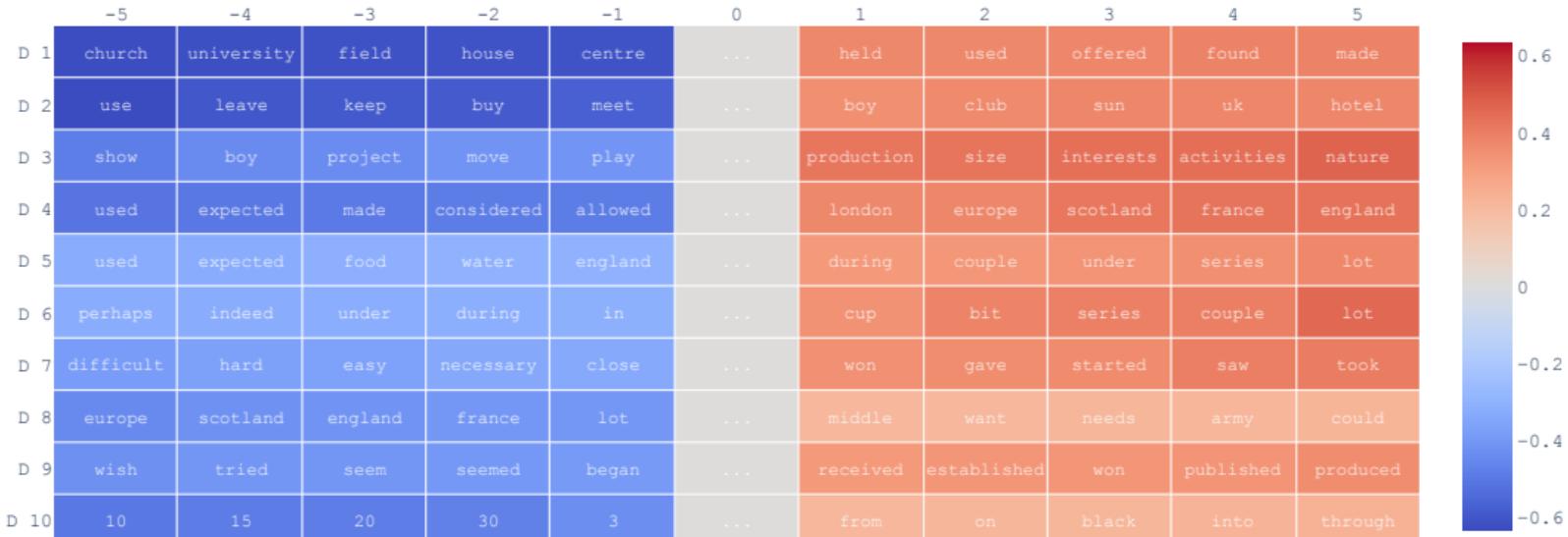
Eigenvalues of $M_* M^*$ and $M^* M_*$:



Eigenvectors of $M^* M_*$:



Words



Introduction

Epistemological Perspectives

Theoretical Perspectives

The Algebra Behind the Embeddings

The Structure Behind the Algebra

The Categories Behind the Structure

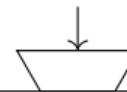
Take Aways

The Structure of Embeddings

Structure



$\{-, /, 0, 1, 2, \dots, 8, 9, =,$
 $a, b, c, \dots, w, x, y, z, \acute{e}\}$

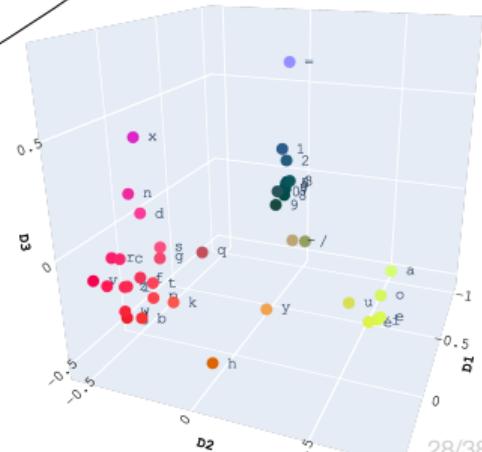


Embedding

Data



SVD

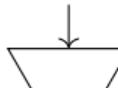


The Structure of Embeddings

Structure

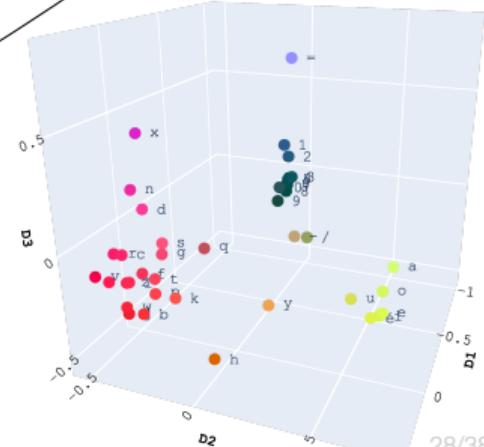
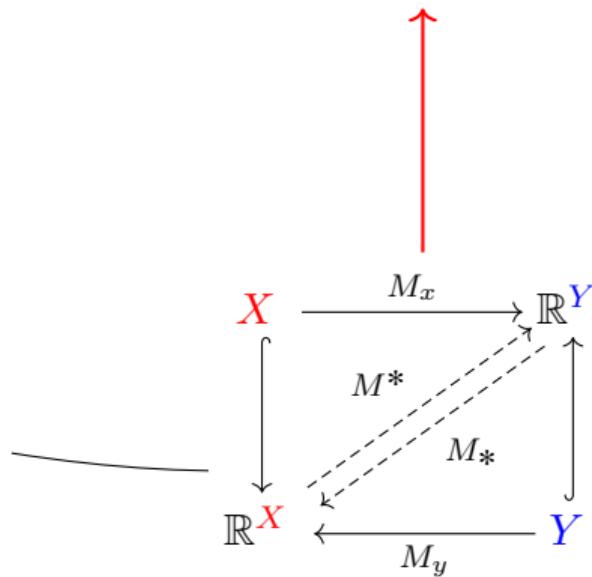


$\{-, /, 0, 1, 2, \dots, 8, 9, =,$
 $a, b, c, \dots, w, x, y, z, \acute{e}\}$



Embedding

Data



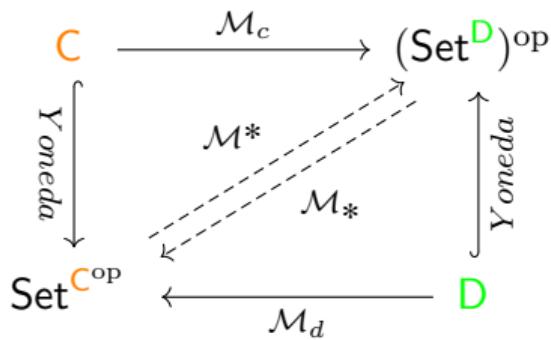
The Structure of Embeddings

Structure

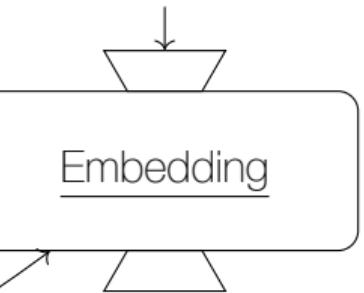
?



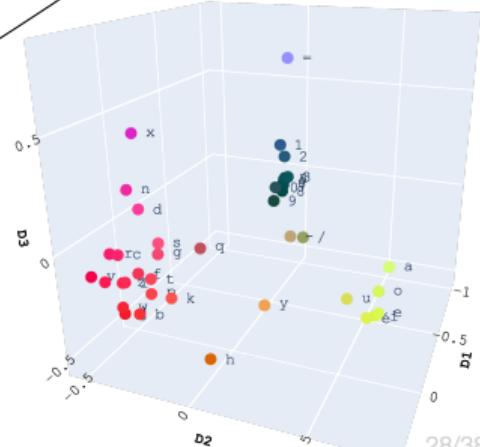
Data



$\{-, /, 0, 1, 2, \dots, 8, 9, =,$
 $a, b, c, \dots, w, x, y, z, é\}$



Embedding

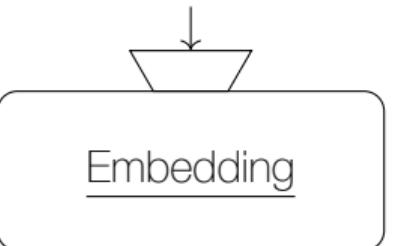


The Structure of Embeddings

Structure

?

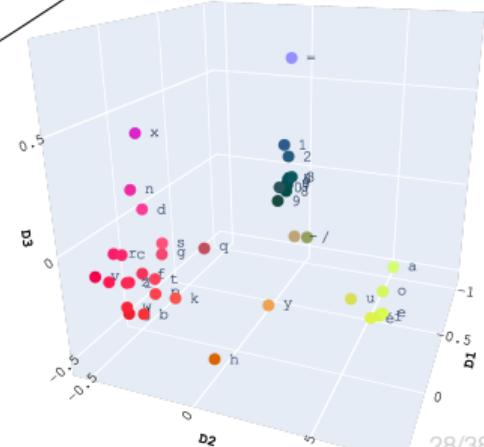
{-, /, 0, 1, 2, ..., 8, 9, =,
a, b, c, ..., w, x, y, z, é}



Data



$C^{\text{op}} \times D \rightarrow \text{Set}$



Structure

?

$$\begin{array}{ccc} \textcolor{teal}{term}_i & \textcolor{teal}{context}_i & \text{measure} \\ \downarrow & \downarrow & \swarrow \\ \textcolor{orange}{C}^{\text{op}} \times \textcolor{green}{D} & \rightarrow & \text{Set} \end{array}$$

Structure

?

$$\begin{array}{ccc} \textcolor{teal}{term}_i & \textcolor{teal}{context}_i & \text{measure} \\ \downarrow & \downarrow & \swarrow \\ \textcolor{orange}{C}^{\text{op}} \times \textcolor{green}{D} \rightarrow \textcolor{red}{Set} \end{array}$$

Structure

?

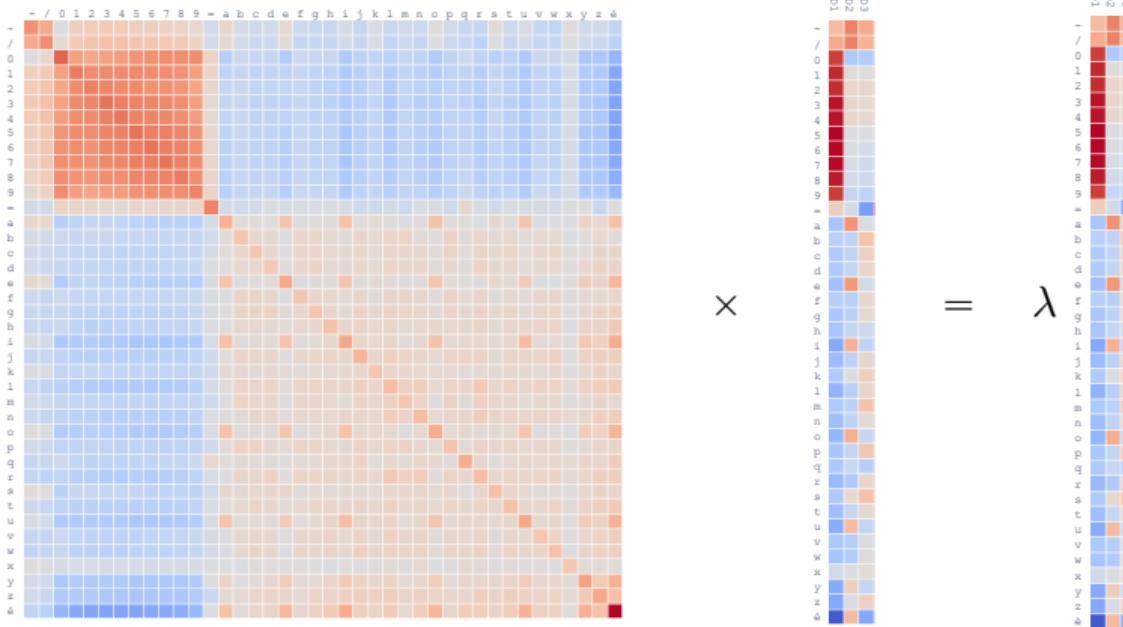
$$\begin{array}{ccc} \textcolor{teal}{term}_i & \textcolor{teal}{context}_i & \text{measure} \\ \downarrow & \downarrow & \swarrow \\ \textcolor{orange}{C}^{\text{op}} \times \textcolor{green}{D} \rightarrow \textcolor{red}{2} \end{array}$$

Structure

$$\begin{array}{c} \text{C}^{\text{op}} \times \text{D} \rightarrow 2 \\ \Downarrow \\ \mathcal{M}^*: 2^{\text{C}^{\text{op}}} \rightleftarrows (2^{\text{D}})^{\text{op}}: \mathcal{M}_* \end{array}$$

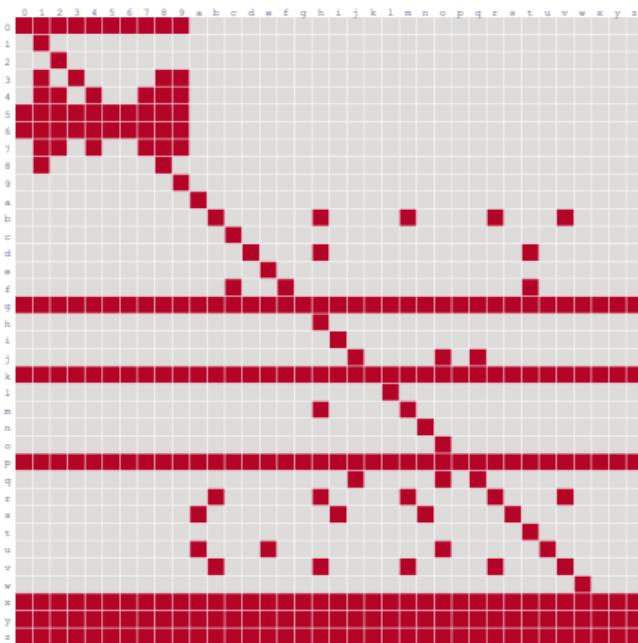
Binary Fixed Points

$$M_* M^* u = \lambda u$$



Binary Fixed Points

$$\mathcal{M}_*\mathcal{M}^*f = f$$



★

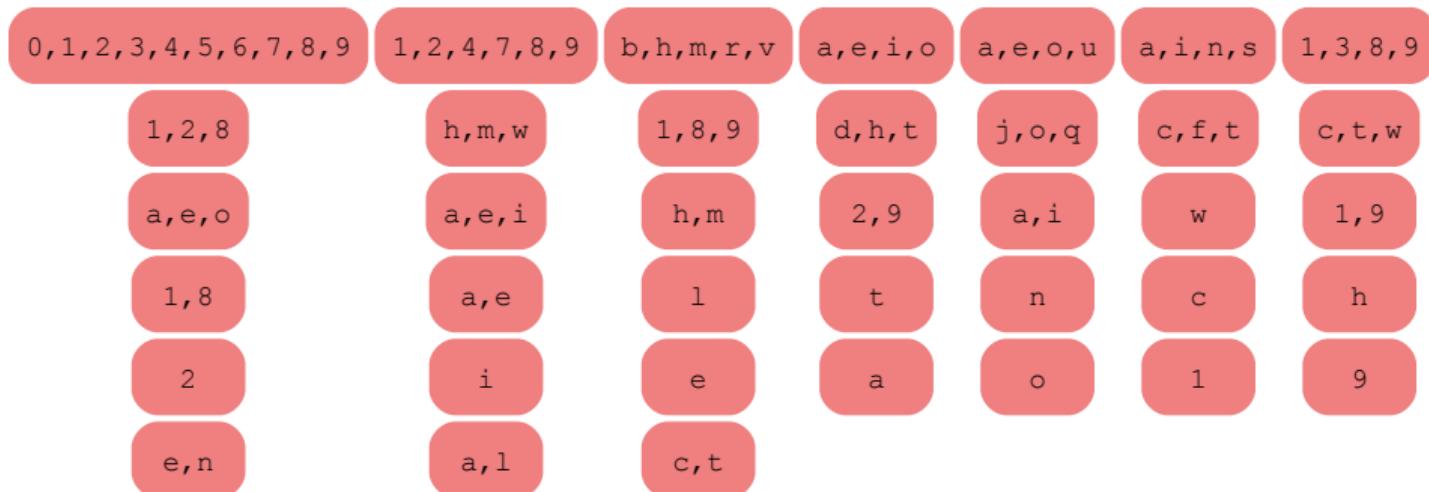


?

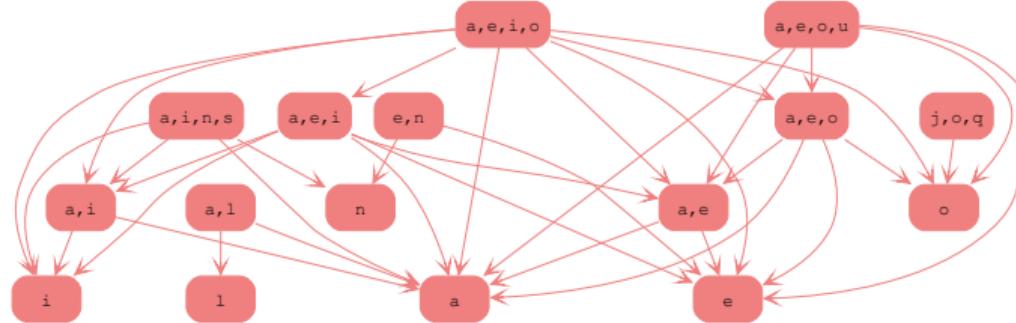
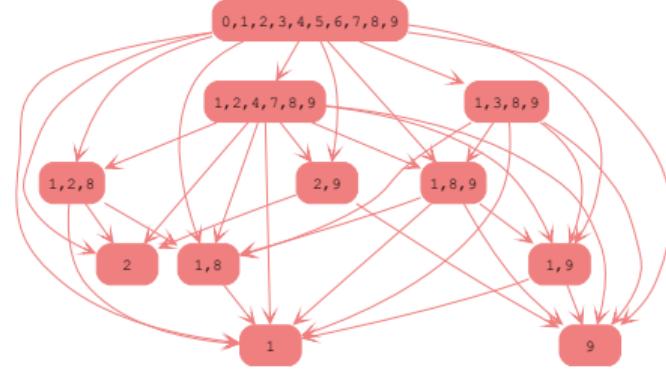
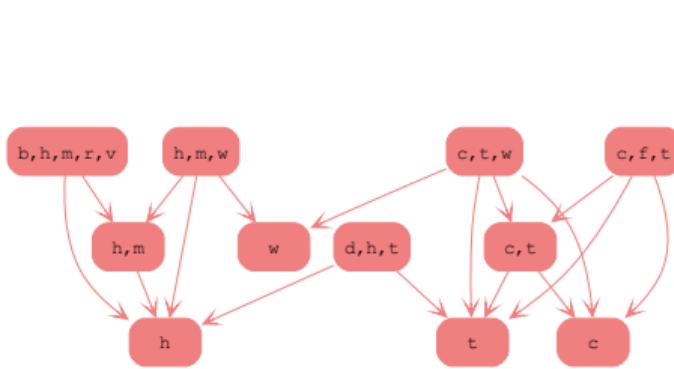


“Eigensets”

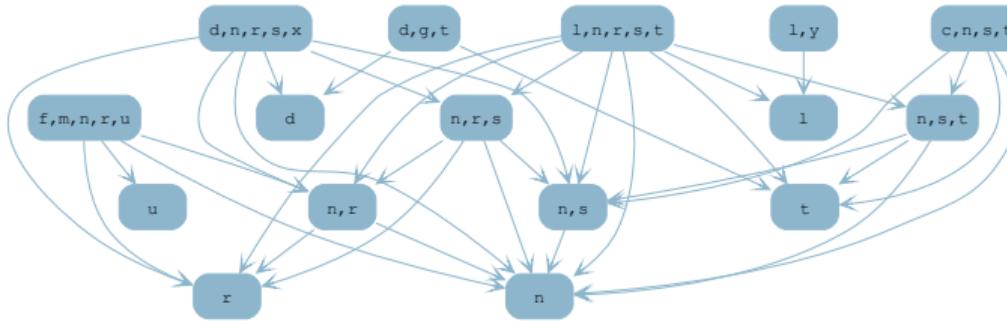
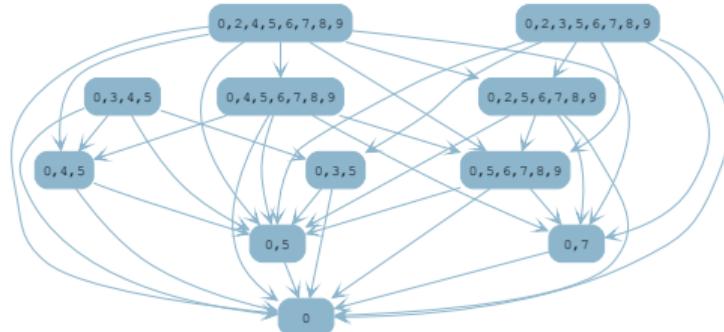
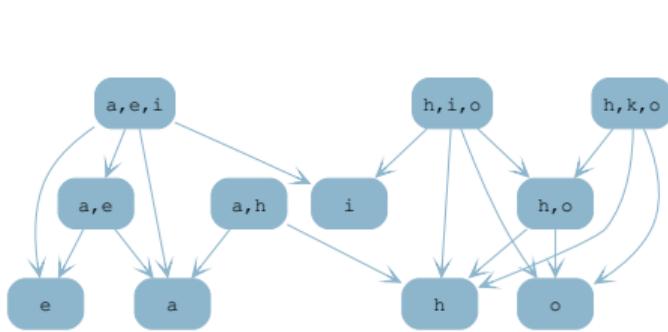
$$\mathcal{M}_*\mathcal{M}^*f = f$$



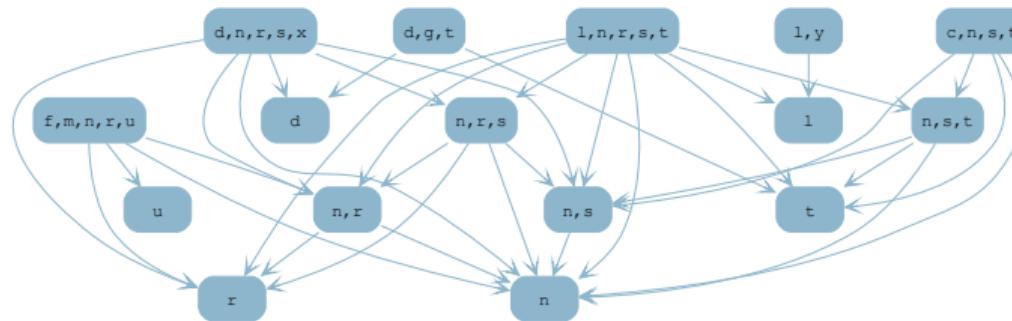
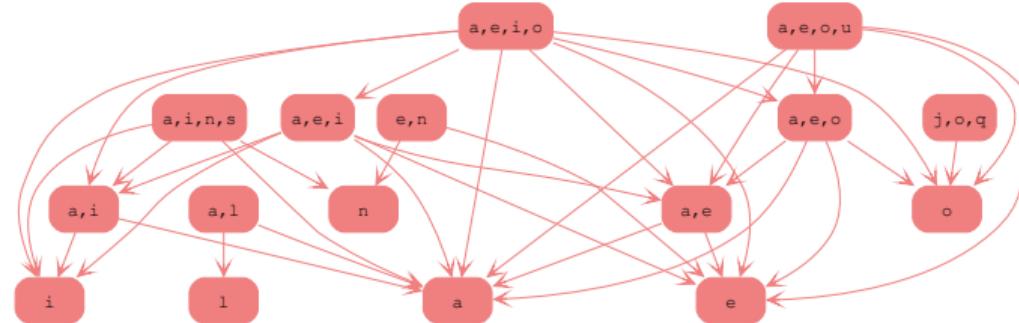
Partial Order Structure

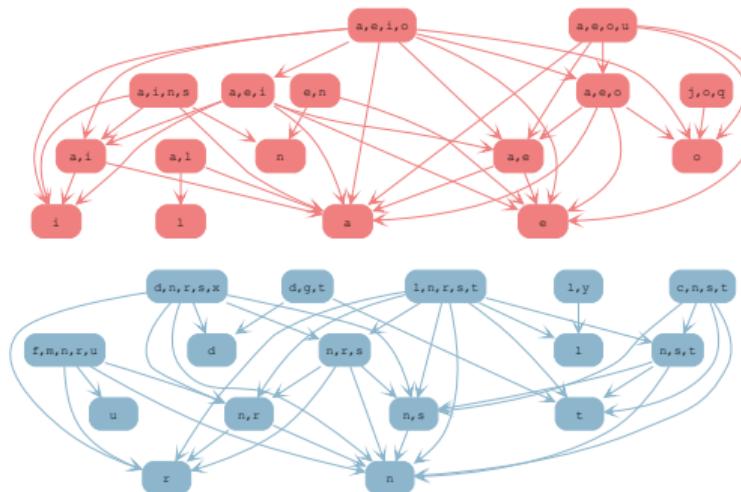


Dual Partial Order



Paring of Partial Ordered Fixed Points



Structure

$$\begin{array}{c} \text{C}^{\text{op}} \times \text{D} \rightarrow 2 \\ \Downarrow \\ \mathcal{M}^*: 2^{\text{C}^{\text{op}}} \rightleftarrows (2^{\text{D}})^{\text{op}}: \mathcal{M}_* \end{array}$$

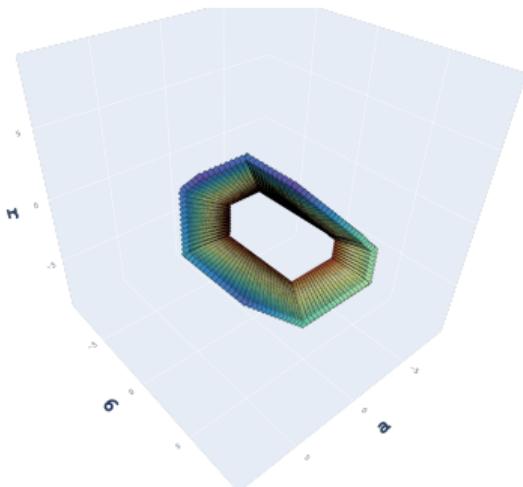
Structure

?

$$\begin{array}{ccc} \text{C}^{\text{op}} \times \text{D} & \xrightarrow{\quad} & \bar{\mathbb{R}} \\ & \Downarrow & \\ \mathcal{M}^*: \bar{\mathbb{R}}^{\text{C}^{\text{op}}} & \xleftarrow{\quad} & (\bar{\mathbb{R}}^{\text{D}})^{\text{op}}: \mathcal{M}_* \end{array}$$

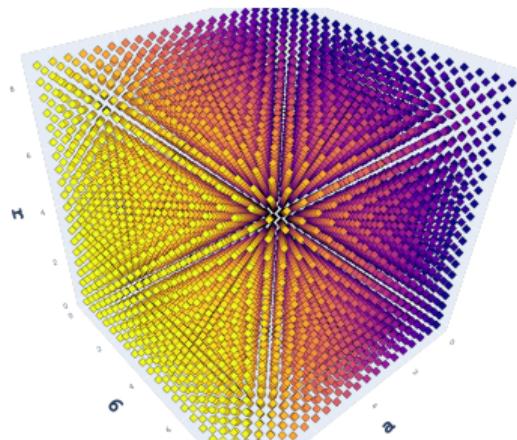
Enriching over $\bar{\mathbb{R}}$

Structure

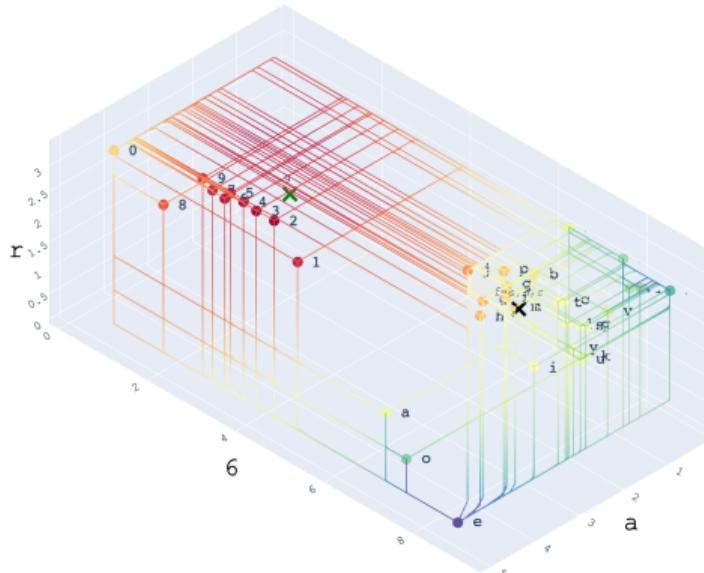


?

$$\mathcal{M}_* \mathcal{M}^*$$



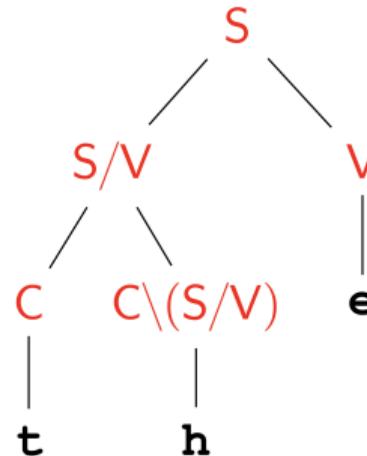
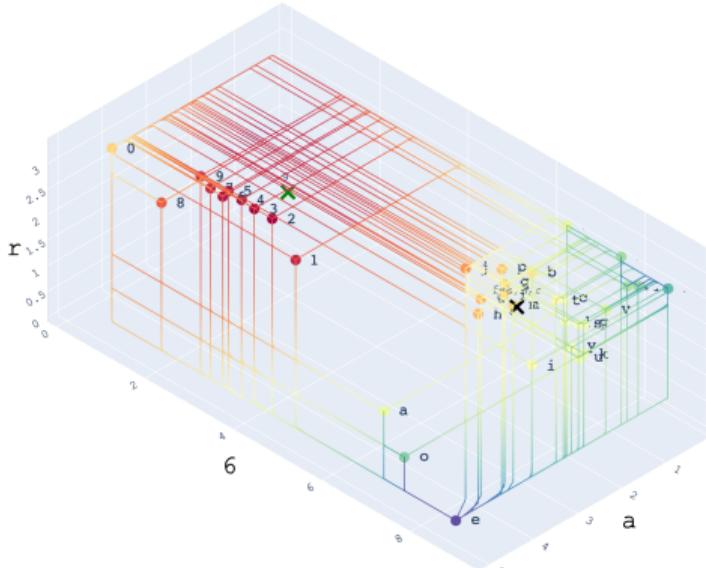
$$\begin{array}{c} \text{C}^{\text{op}} \times \text{D} \rightarrow \bar{\mathbb{R}} \\ \Downarrow \\ \mathcal{M}^*: \bar{\mathbb{R}}^{\text{C}^{\text{op}}} \rightleftarrows (\bar{\mathbb{R}}^{\text{D}})^{\text{op}}: \mathcal{M}_* \end{array}$$

Structure

$$\begin{array}{c}
 \textcolor{orange}{C}^{\text{op}} \times \textcolor{green}{D} \rightarrow \bar{\mathbb{R}} \\
 \Downarrow \\
 \mathcal{M}^*: \bar{\mathbb{R}}^{\textcolor{orange}{C}^{\text{op}}} \rightleftarrows (\bar{\mathbb{R}}^{\textcolor{green}{D}})^{\text{op}}: \mathcal{M}_*
 \end{array}$$

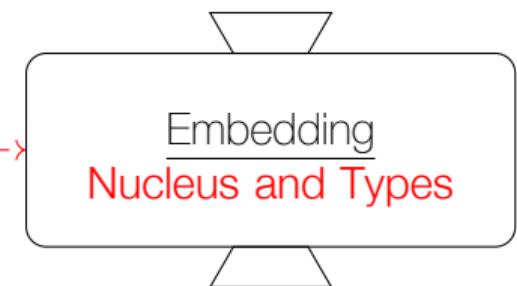
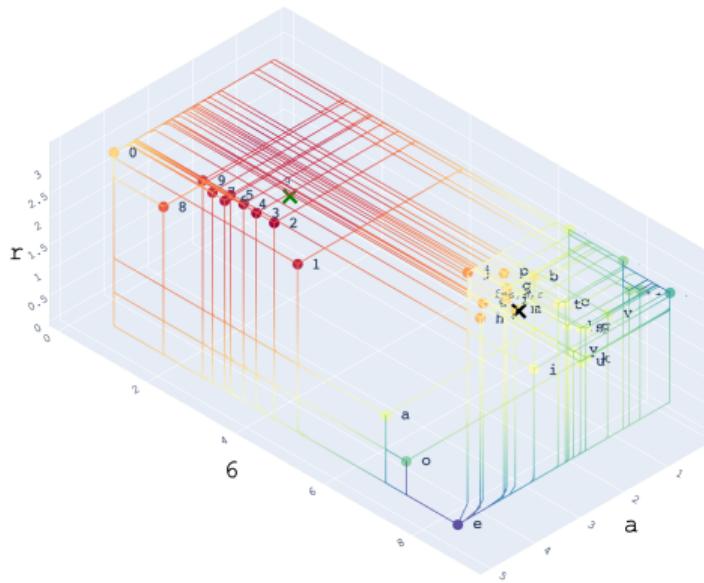
Enriching over $\bar{\mathbb{R}}$

Structure



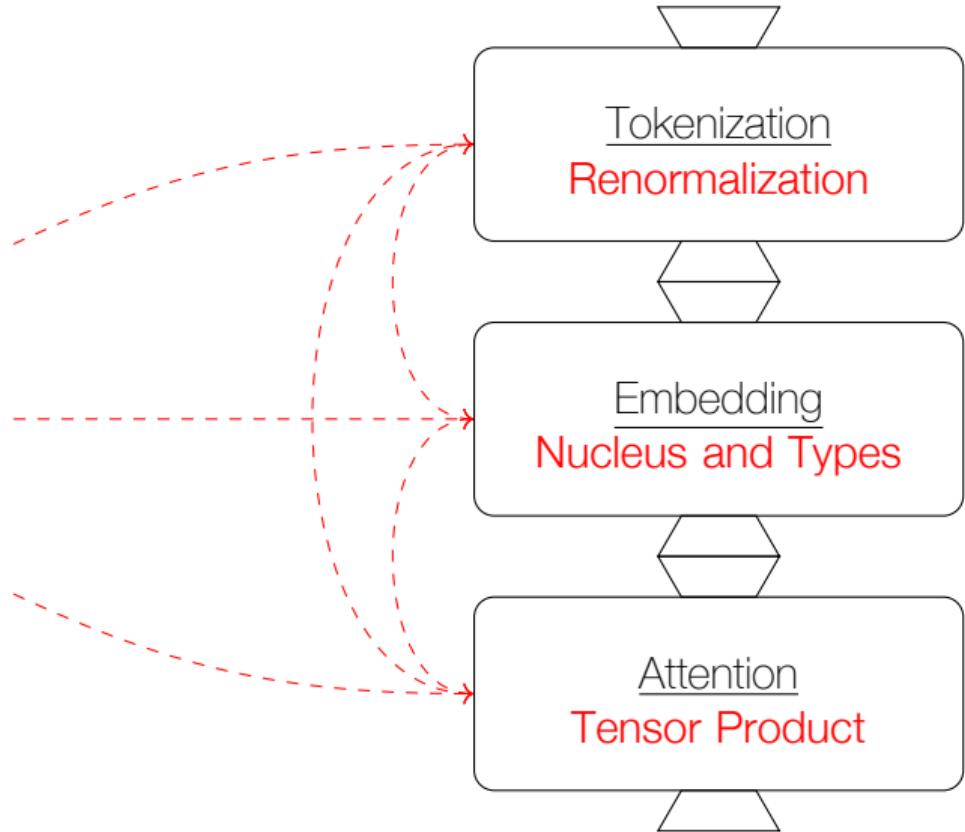
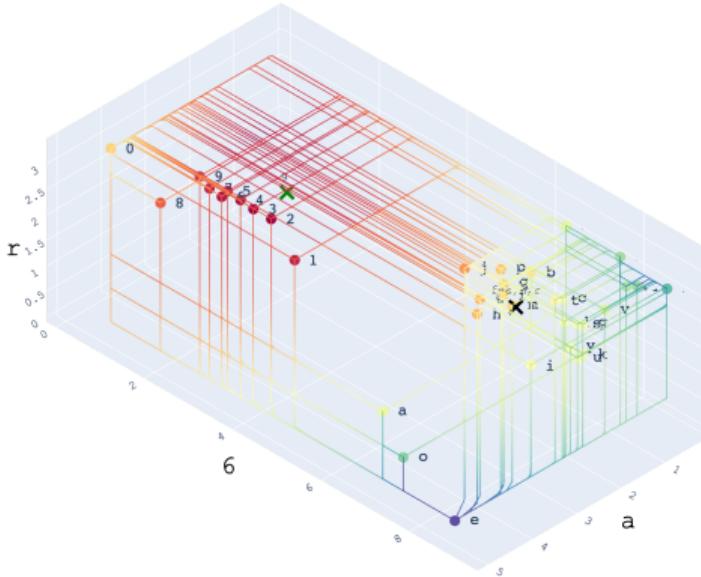
$$\begin{array}{c} \textcolor{orange}{C}^{\text{op}} \times \textcolor{green}{D} \rightarrow \bar{\mathbb{R}} \\ \Downarrow \\ \mathcal{M}^*: \bar{\mathbb{R}}^{\textcolor{orange}{C}^{\text{op}}} \rightleftarrows (\bar{\mathbb{R}}^{\textcolor{green}{D}})^{\text{op}}: \mathcal{M}_* \end{array}$$

Structure



Formal Explainability

Structure



Outline

Introduction

Epistemological Perspectives

Theoretical Perspectives

The Algebra Behind the Embeddings

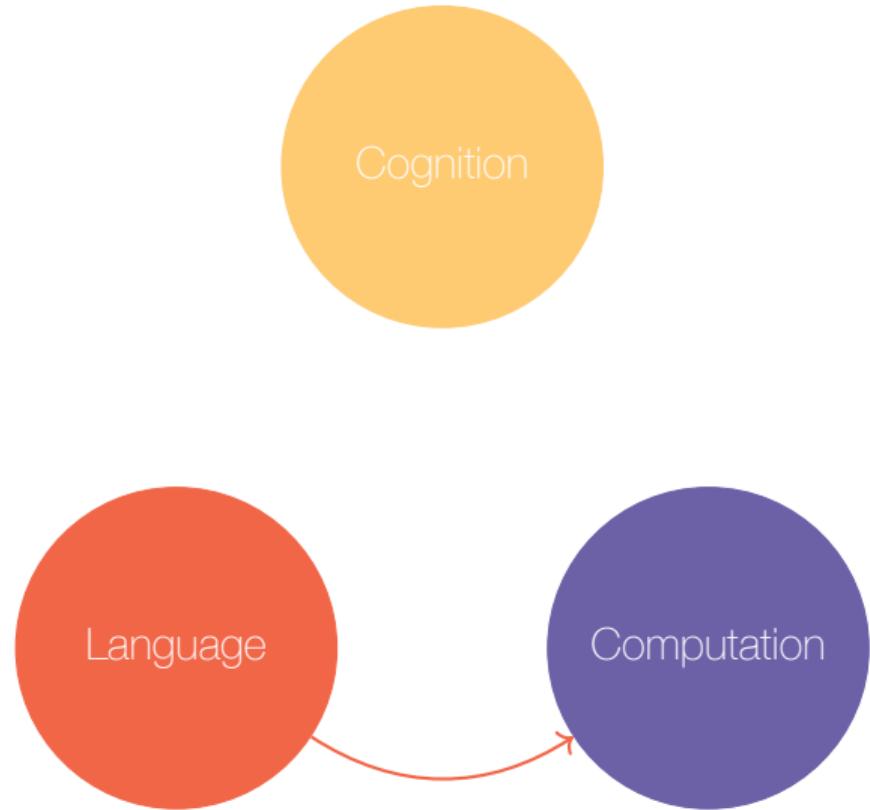
The Structure Behind the Algebra

The Categories Behind the Structure

Take Aways

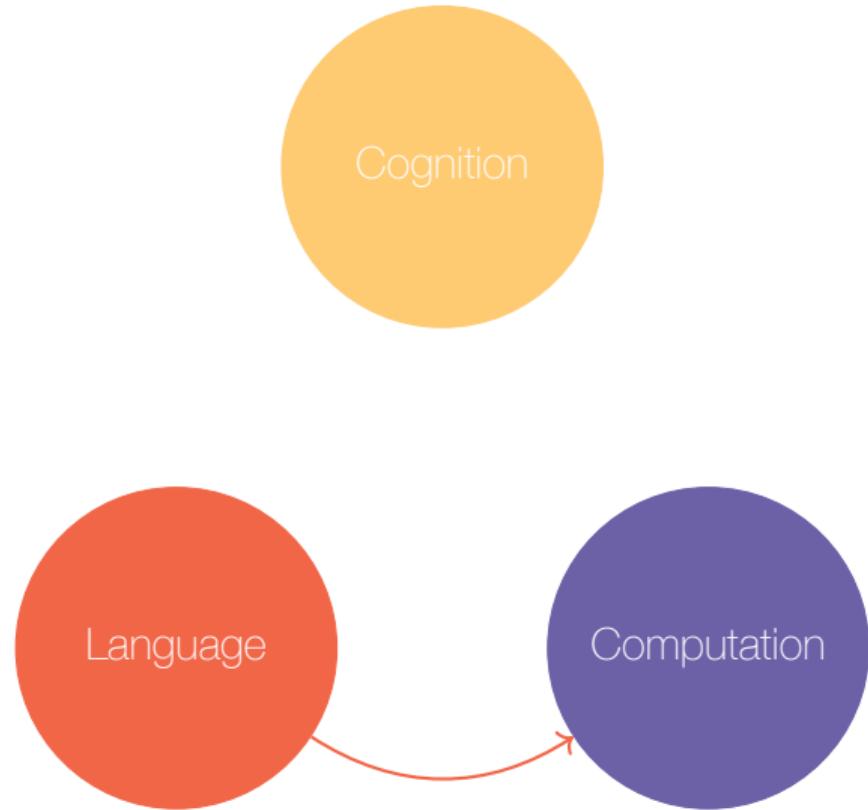
Take Aways

- ◊ A **formal** approach to data analysis can contribute to inferring **symbolic language** models **from** linguistic **data**.



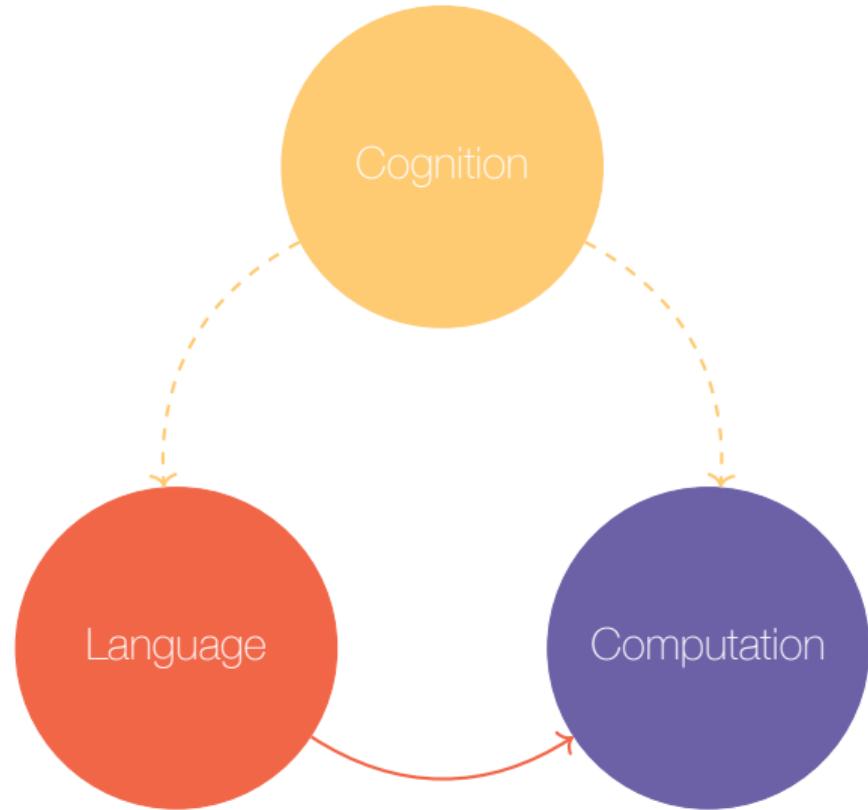
Take Aways

- ◊ A **formal** approach to data analysis can contribute to inferring **symbolic language** models **from** linguistic **data**.
- ◊ Resulting models are, a priori, **models of the data**.



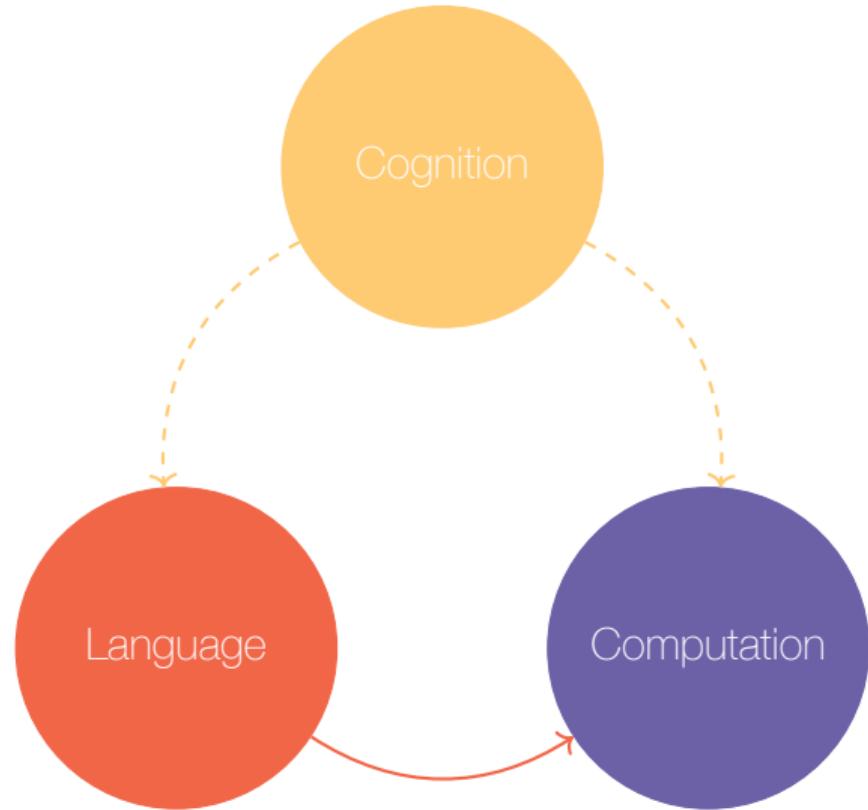
Take Aways

- ◊ A **formal** approach to data analysis can contribute to inferring **symbolic language** models **from** linguistic **data**.
- ◊ Resulting models are, a priori, **models of the data**.
- ◊ The **cognitive content** of such models is **suspended**, and cannot be restored without raising the **problem of the data**.



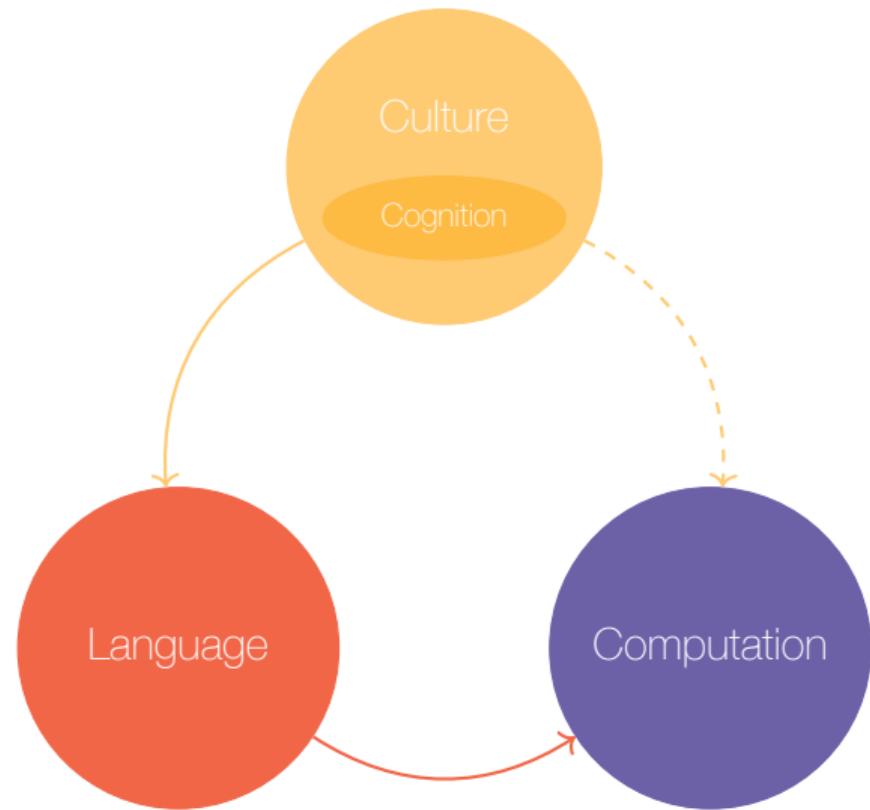
Take Aways

- ◊ A **formal** approach to data analysis can contribute to inferring **symbolic language** models **from** linguistic **data**.
- ◊ Resulting models are, *a priori*, **models of the data**.
- ◊ The **cognitive content** of such models is **suspended**, and cannot be restored without raising the **problem of the data**.
- ◊ The **scale** of the data for such models **exceeds the individual scale**.

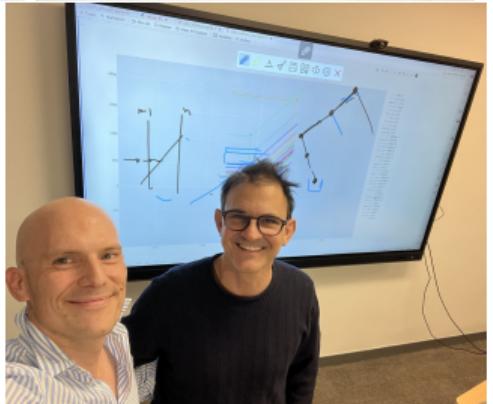


Take Aways

- ◊ A **formal** approach to data analysis can contribute to inferring **symbolic language** models **from** linguistic **data**.
- ◊ Resulting models are, *a priori*, **models of the data**.
- ◊ The **cognitive content** of such models is **suspended**, and cannot be restored without raising the **problem of the data**.
- ◊ The **scale** of the data for such models **exceeds the individual scale**.
- ◊ **Cultural conditions** of data production become **constitutive** in the relation between cognitive contents and language models.



Collaborations



J. Terilla (CUNY), T.-D. Bradley (SandboxAQ), L. Pellissier (Paris-Est Créteil), Th. Seiller (CNRS), S. Jarvis (CUNY)

Reference Papers

- ◊ Gastaldi, J. L. (2021). Why Can Computers Understand Natural Language? *Philosophy & Technology*, 34(1), 149–214. <https://doi.org/10.1007/s13347-020-00393-9>
- ◊ Gastaldi, J. L., & Pellissier, L. (2021). The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*, 46(4), 569–590. <https://doi.org/10.1080/03080188.2021.1890484>
- ◊ Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024). The structure of meaning in language: Parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*. <https://api.semanticscholar.org/CorpusID:263613625>

References I

- Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2024). Leace: Perfect linear concept erasure in closed form. *Proceedings of the 37th International Conference on Neural Information Processing Systems*.
- Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024). The structure of meaning in language: Parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*.
<https://api.semanticscholar.org/CorpusID:263613625>
- Chomsky, N. (1953). Systems of syntactic analysis. *Journal of Symbolic Logic*, 18(3), 242–256.
<https://doi.org/10.2307/2267409>
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3), 113–124. <https://doi.org/10.1109/TIT.1956.1056813>
- Chomsky, N. (1957). *Syntactic structures*. Mouton; Co.
- Chomsky, N. (1959). *Language*, 35(1), 26–58. Retrieved July 7, 2025, from <http://www.jstor.org/stable/411334>
- Chomsky, N. (1992, November). Language and the “cognitive revolutions” [Delivered November 23–27, 1992].
- Delétang, G., Ruoss, A., Grau-Moya, J., Genewein, T., Wenliang, L. K., Catt, E., Cundy, C., Hutter, M., Legg, S., Veness, J., & Ortega, P. A. (2023). Neural networks and the chomsky hierarchy.
<https://arxiv.org/abs/2207.02098>
- Gastaldi, J. L. (2021). Why Can Computers Understand Natural Language? *Philosophy & Technology*, 34(1), 149–214.
<https://doi.org/10.1007/s13347-020-00393-9>
- Gastaldi, J. L., & Pellissier, L. (2021). The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*, 46(4), 569–590.
<https://doi.org/10.1080/03080188.2021.1890484>
- Girard, J.-Y. (2011, September). *The blind spot*. European Mathematical Society.

References II

- Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.
- Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2177–2185.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Le, Q., & Strohmann, T. (2013). *Learning representations of text using neural networks. NIPS deep learning workshop 2013 slides*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.
- Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the ACL*, 1715–1725.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb0d053c1c4a845aa-Paper.pdf

NeuroMod Annual Meeting
Université Côte d'Azur
Antibes, France

What are Neural Language Models the Model of?
Epistemological and Theoretical Perspectives on LLMs

Juan Luis Gastaldi

www.giannigastaldi.com

ETH zürich

July 8, 2025