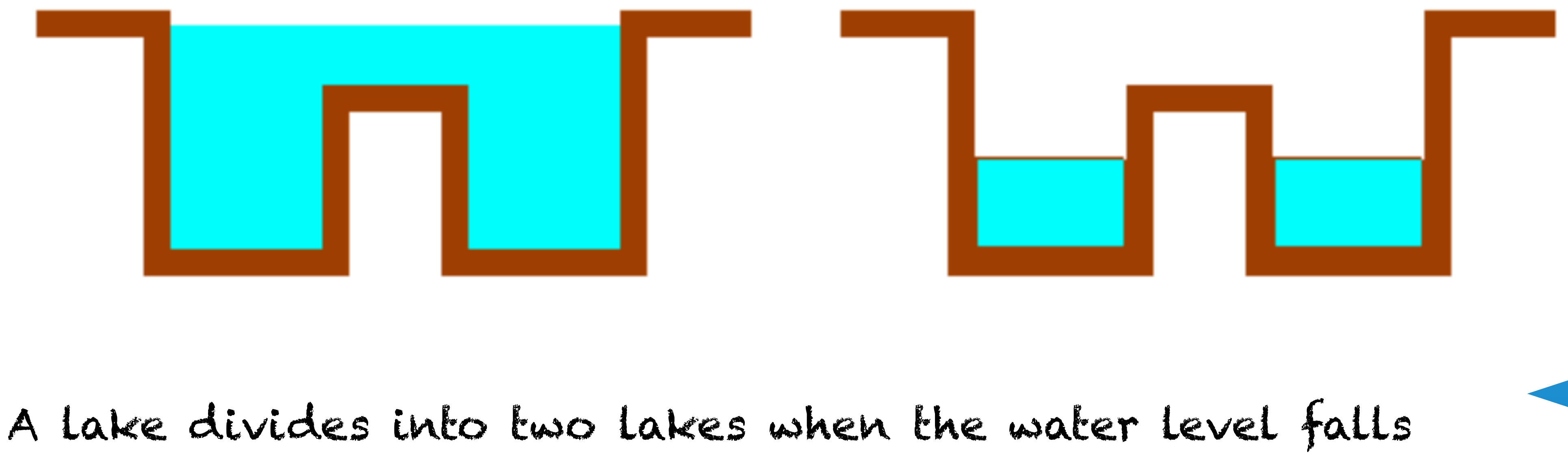
Commonsense Knowledge Representation in NLP

Xiang Lorraine Li

PhD Student at UMass Amherst,

Advised by Prof Andrew McCallum

Commonsense: Essential in AI [1]



A five-year-old would know this!

A five year old would also know the brick will fall if it's not placed correctly

Common sense is sound practical judgement

- **Concerning everyday matters**
- **Basic ability to perceive, understand, and judge**
- **Shared by ("common to") nearly all people.** ----Wikipedia

Disclaimer: we mean, roughly, what a typical five year old knows about the world, including fundamental categories like time and space, and specific domains such as physical objects and substances; plants, animals, and other natural entities; humans, their psychology, and their interactions; and society at large. We will not attempt to be precise about this, but let us indicate roughly which issues we are considering and which we are ignoring. Obviously, this body of knowledge in fact depends on place, time, culture, social standing, personal characteristics (e.g. unusual cognitive or physical abilities or disabilities), schooling, perhaps on language. We ignore all that; without embarrassment, we have in mind a 21st-century, first-world, urban, with an appropriate level of schooling.

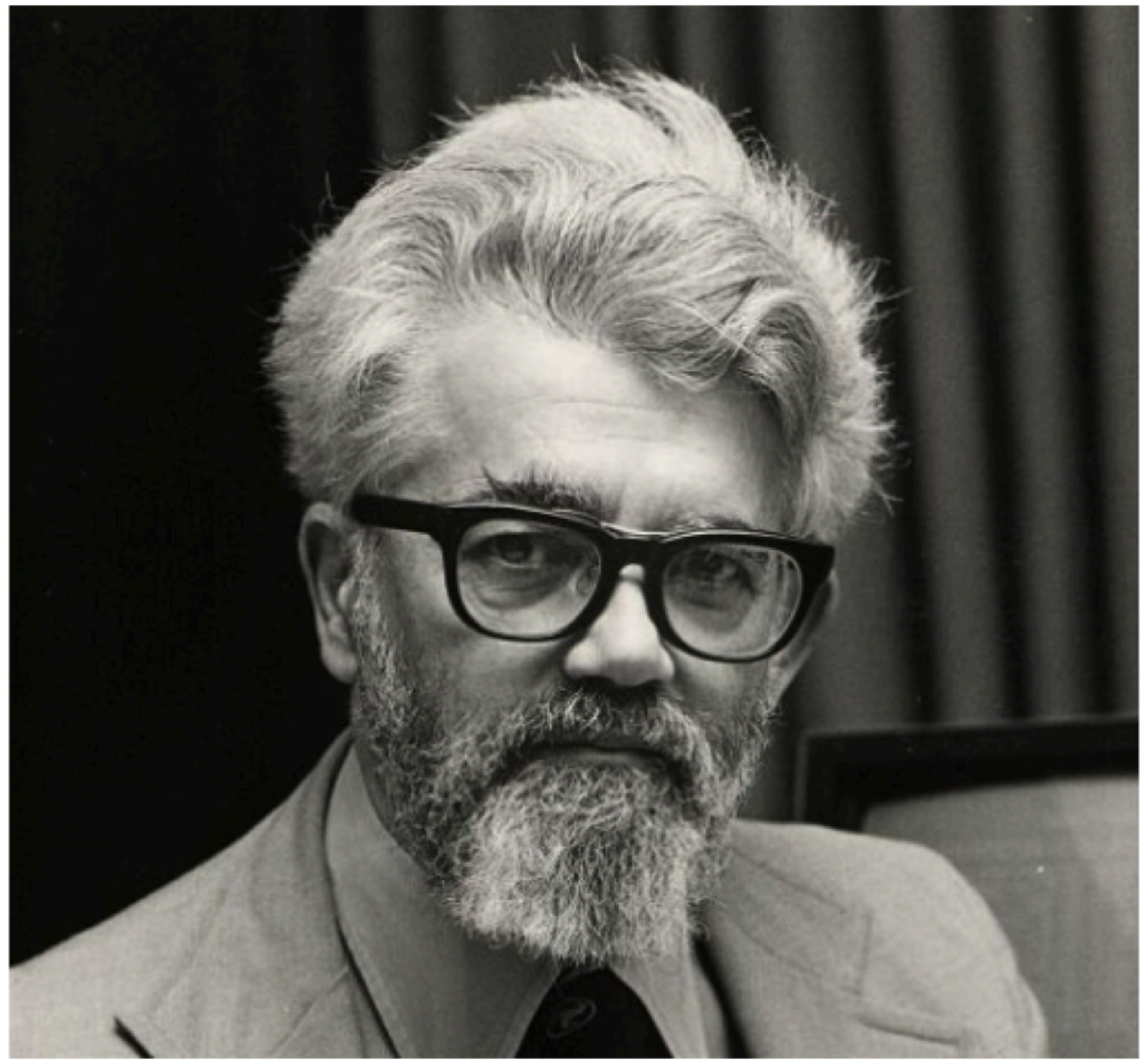
[1] https://cacm.acm.org/magazines/2015/9/191169-commonsense-reasoning-and-commonsense-knowledge-in-artificial-intelligence/fulltext?mobile=false#F1

Commonsense: Essential in AI [1]

- Natural Language Processing
  - Machine T
  - Smart Ho

[1] https://cacm.acm.org/magazines/2015/9/191169-commonsense-reasoning-and-commonsense-knowledge-in-artificial-intelligence/fulltext?mobile=false#F1

Logical Formalization of Commonsense



John McCarthy

Advise Taker

1. First, we have a predicate "*at*". "$at(x, y)$" is a formalization of "*xisaty*". Under this heading we have the premises

$$at(I, desk) \quad (1)$$

$$at(desk, home) \quad (2)$$

$$at(car, home) \quad (3)$$

$$at(home, county) \quad (4)$$

$$at(airport, county) \quad (5)$$

In order for a program to be capable of learning something it must first be capable of being told it — John McCarthy [1]

Programs with Common Sense was probably the first paper on logical AI, i.e. AI in which logic is the method of representing information in computer memory and not just the subject matter of the program. The paper was given in the Teddington Conference on the Mechanization of Thought Processes in December 1958 and printed in the proceedings of that conference. **It may also be the first paper to propose common sense reasoning ability as the key to AI.**

[1] Programs With Common Sense. John McCarthy 1959

Logical Formalization of Commonsense

Incomplete list of different logic systems.

Deductive

Propositional Logic

- There are three types of non-logical symbols: Constants, functions, and predicates. A constant symbol denotes an individual entity; a function denotes a mapping; and a predicate symbol denotes a relation.
- The logical symbols include the Boolean operators; the two quantifiers ∀ (for all) and ∃ (there exists); and variable symbols. Optionally, the equality sign = may also be included.

First-Order Logic

- Modal sentences are generally closer to a natural language expression of the fact.
- Modal sentences may be easier for a knowledge engineer to use.
- Propositional modal logics — that is, logics that have modal operators but no explicit quantifiers — are often both expressive enough for the purpose at hand and reasonably tractable, or at least decidable (Vardi, 1996).

Modal Logic

Plausible Reasoning

Non-monotonic Logic

Probabilistic Logic

Fuzzy Logic

[1] Davis, Ernest. "Logical formalizations of commonsense reasoning: a survey." *Journal of Artificial Intelligence Research* 59 (2017): 651-723.

Logical Formalization of Commonsense

But … it did not end up well.

Past failures (in 70s – 80s) are inconclusive

- -- weak computing power
- -- not much data
- -- not as strong computational models
- -- not ideal conceptualization / representations

"I was told not to speak the word commonsense…" —Yejin Choi [1]

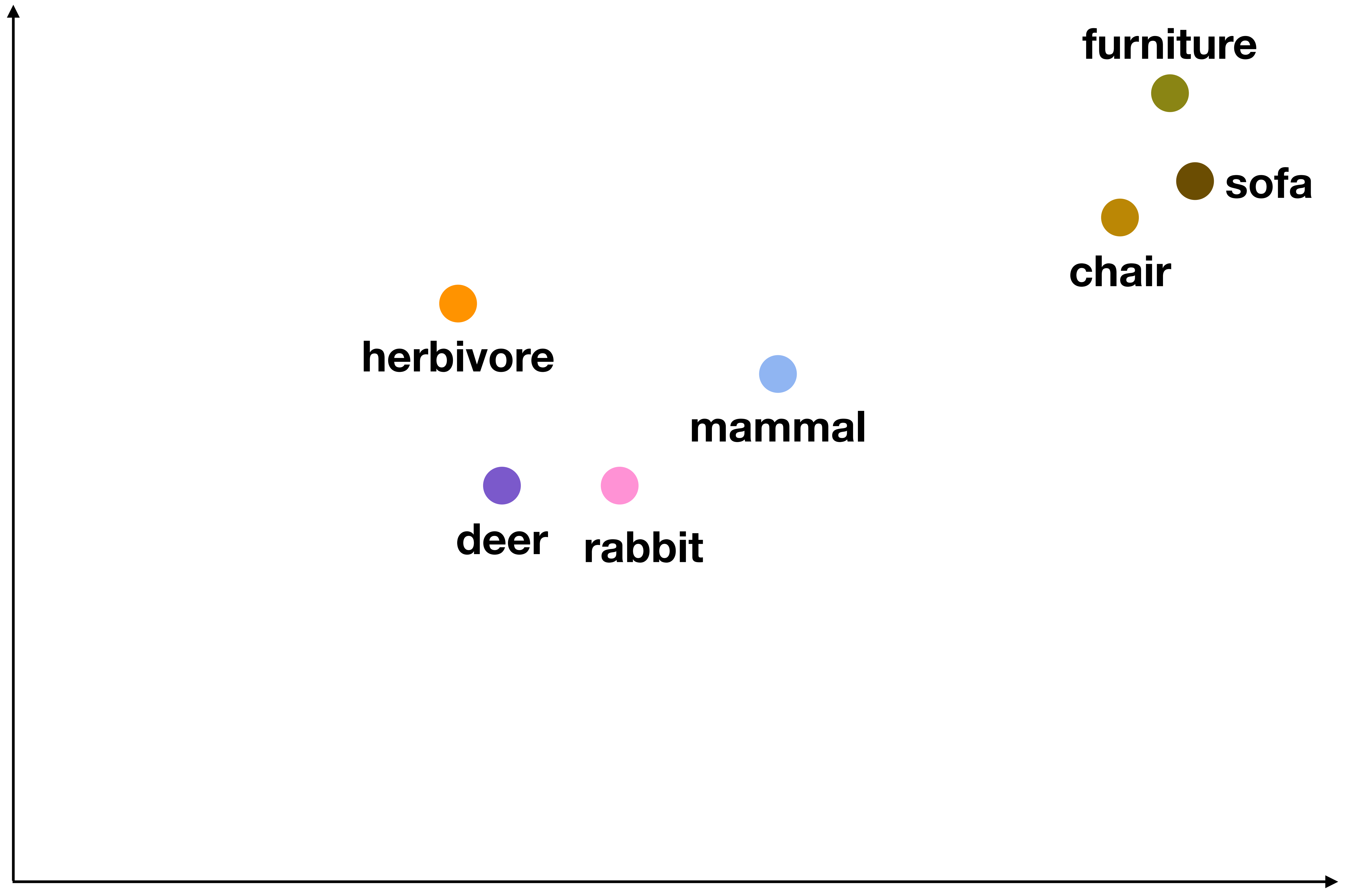
[1] ACL 2020 Commonsense Tutorial

Outline

- Commonsense Knowledge.
- Learn the Right Representation.
- Commonsense Knowledge in Pre-trained LMs.
- Benchmark Datasets for Evaluation.

[Mikolov, Sutskever, Chen, Corrado and Dean 2013]
[Devlin, Chang, Lee and Toutanova 2019]

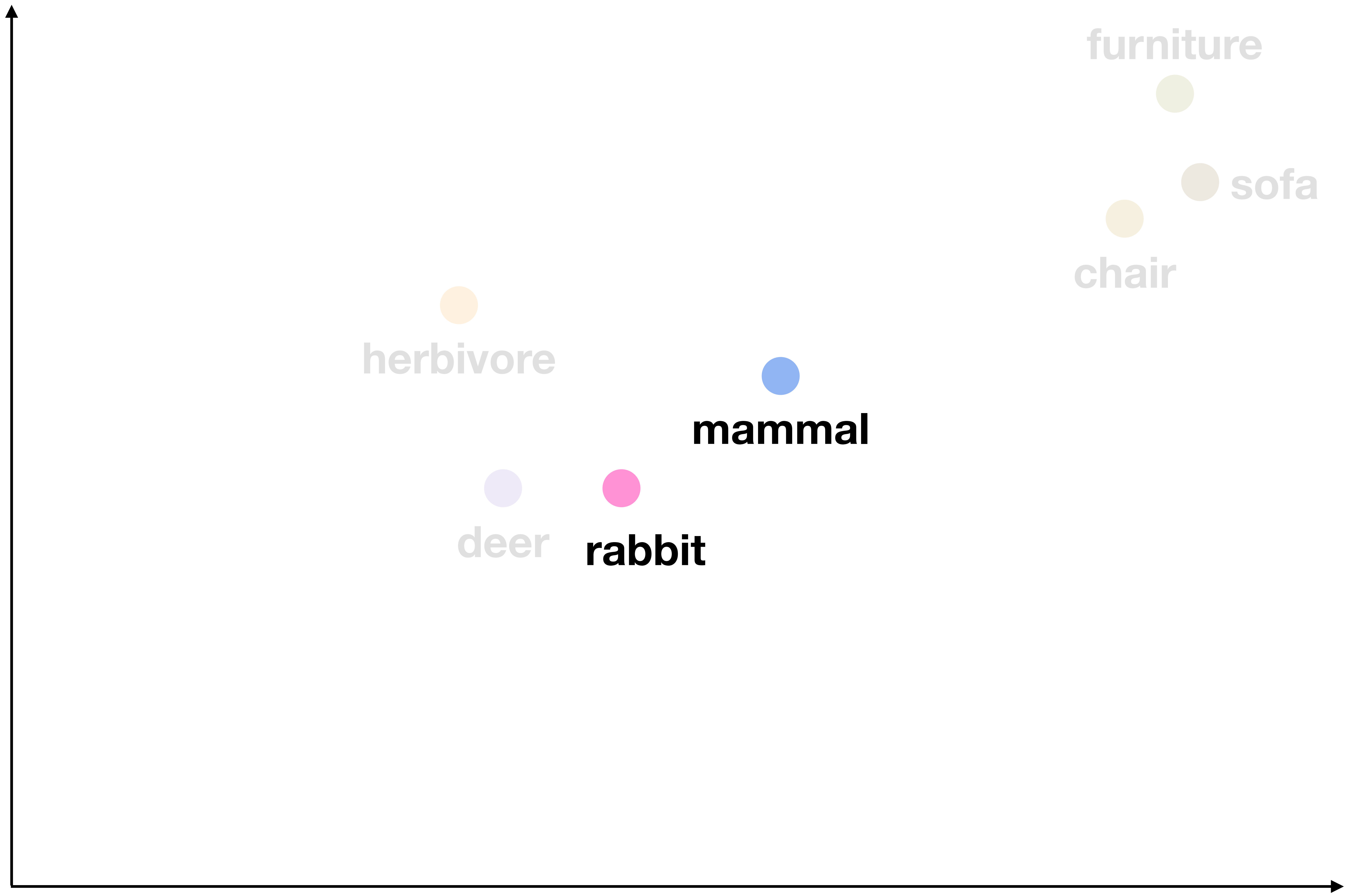
Vector Representation



[Mikolov, Sutskever, Chen, Corrado and Dean 2013]
[Devlin, Chang, Lee and Toutanova 2019]

Vector Representation

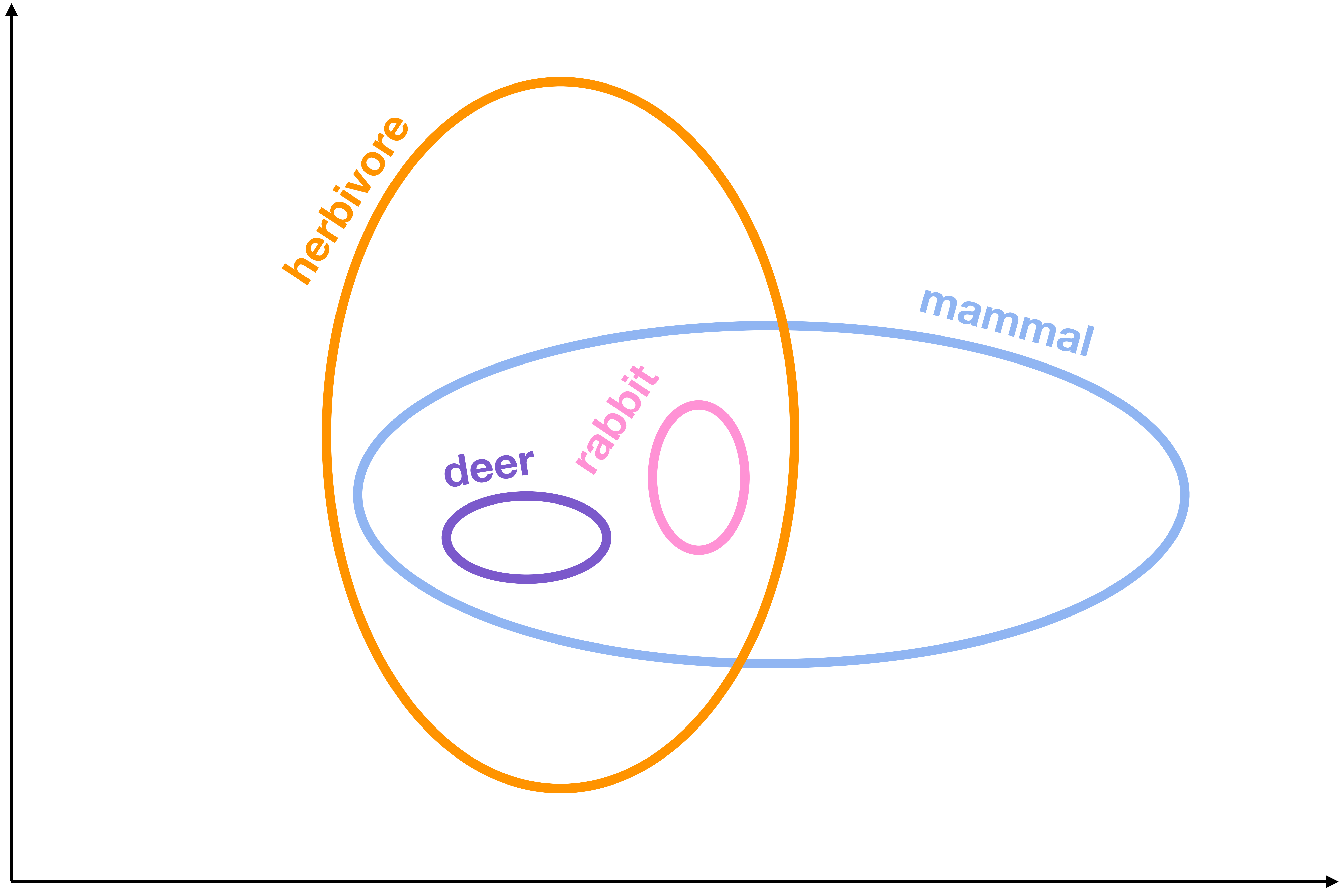
❌ Region ❌ Asymmetry



[Vilnis and McCallum 2014]

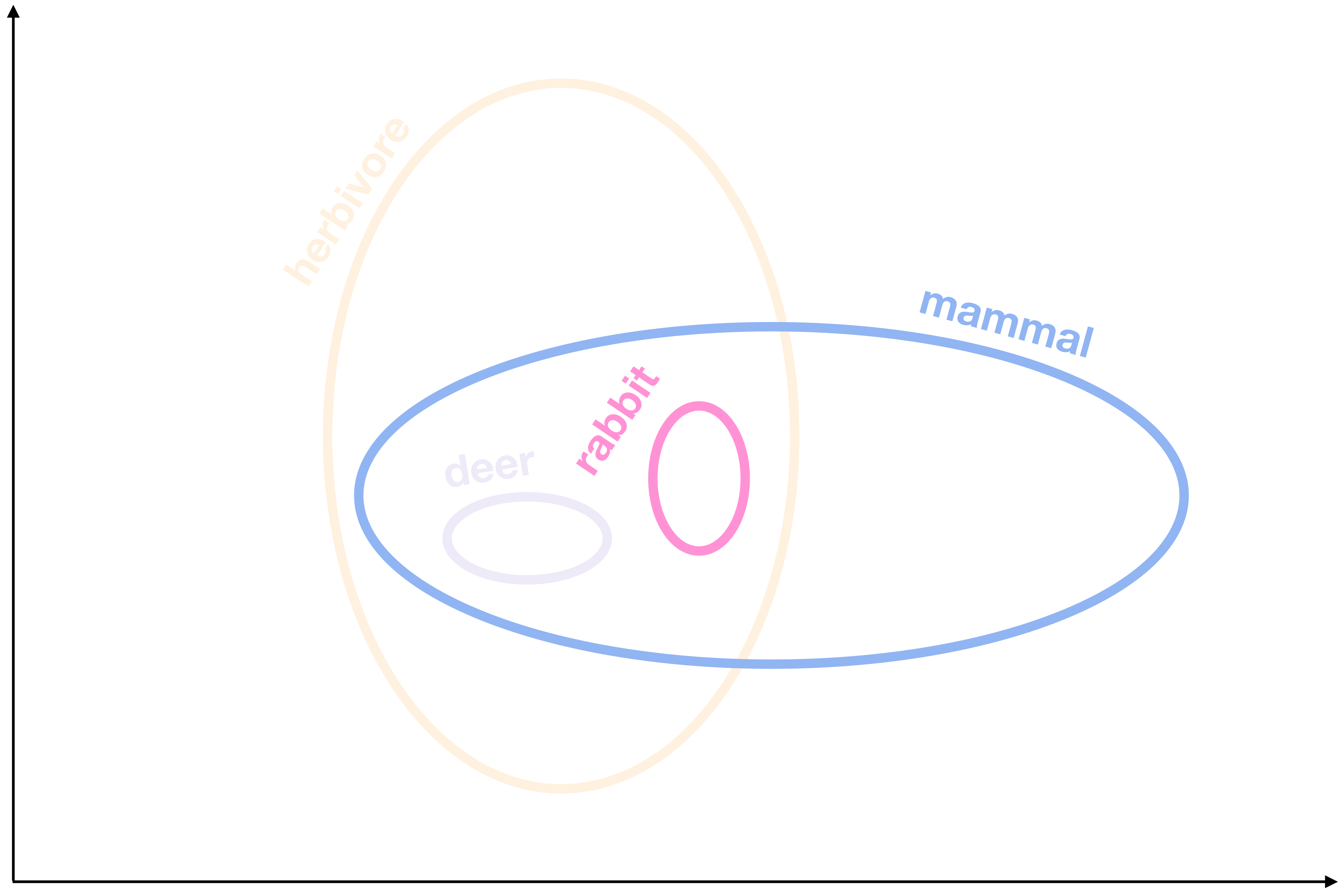
Gaussian Representation

✓ Region ✓ Asymmetry



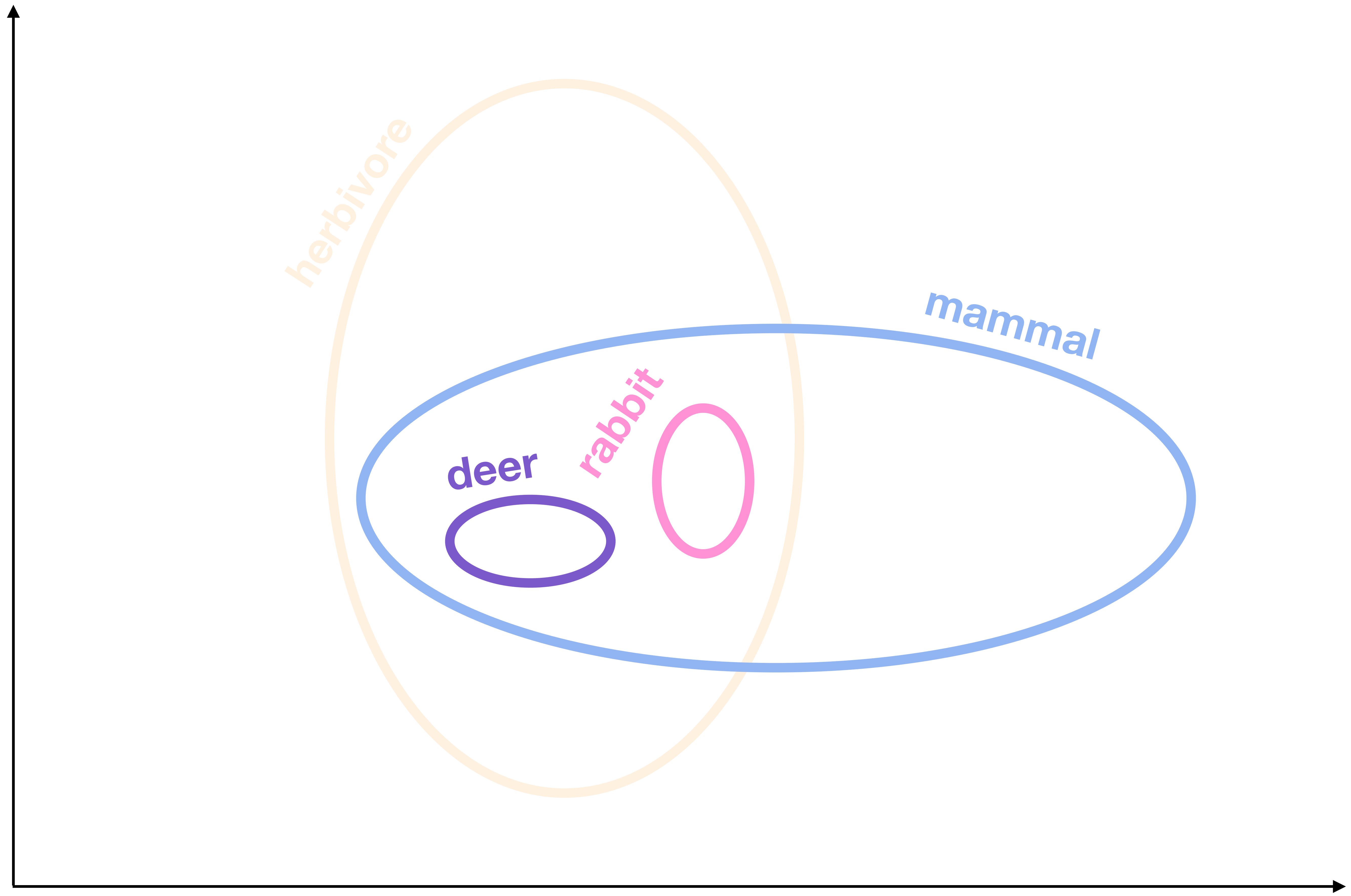
Gaussian Representation

✓ Region ✓ Asymmetry



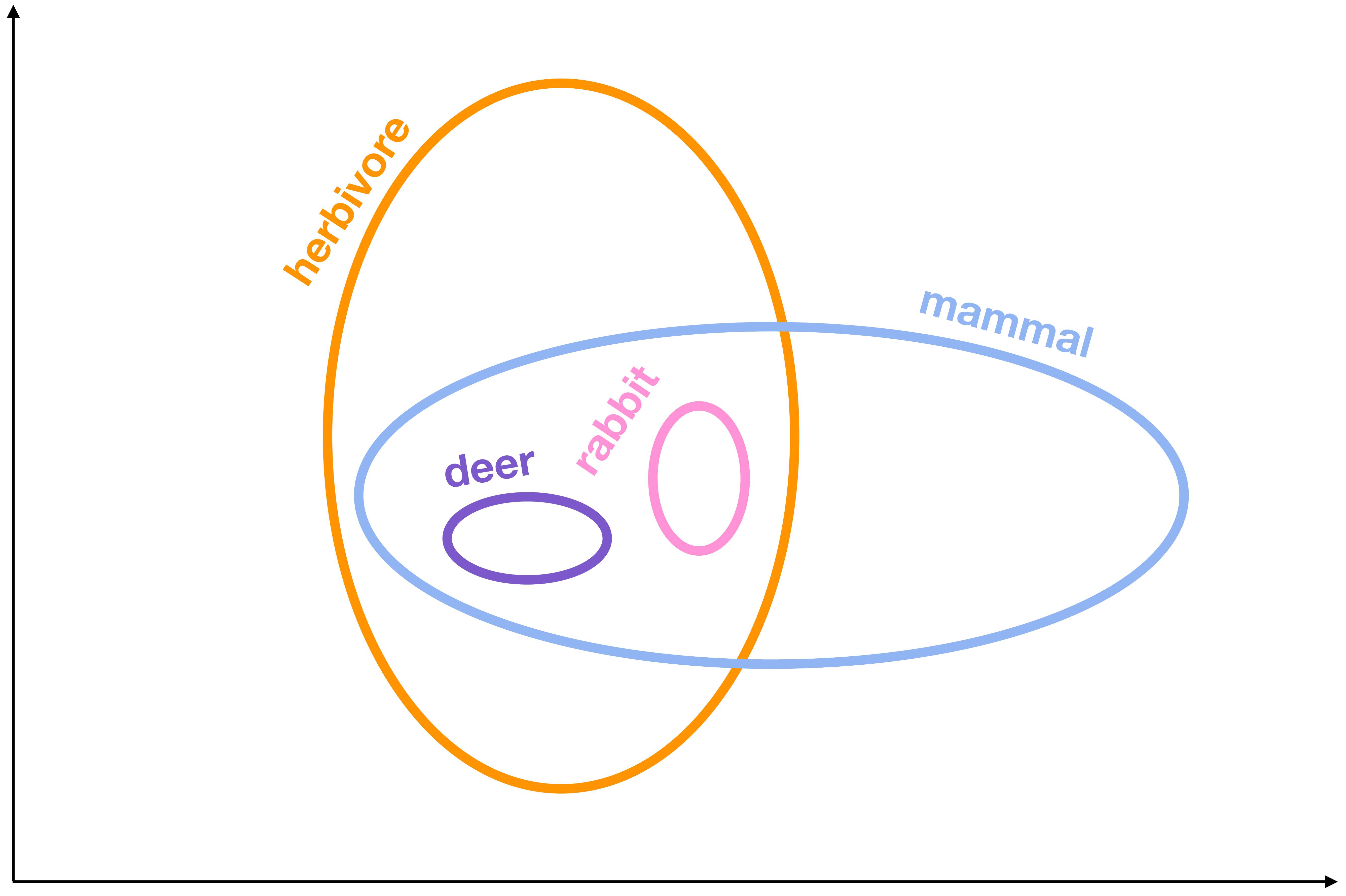
Gaussian Representation

✓ **Region** ✓ **Asymmetry** ✓ **Disjointness**



Gaussian Representation

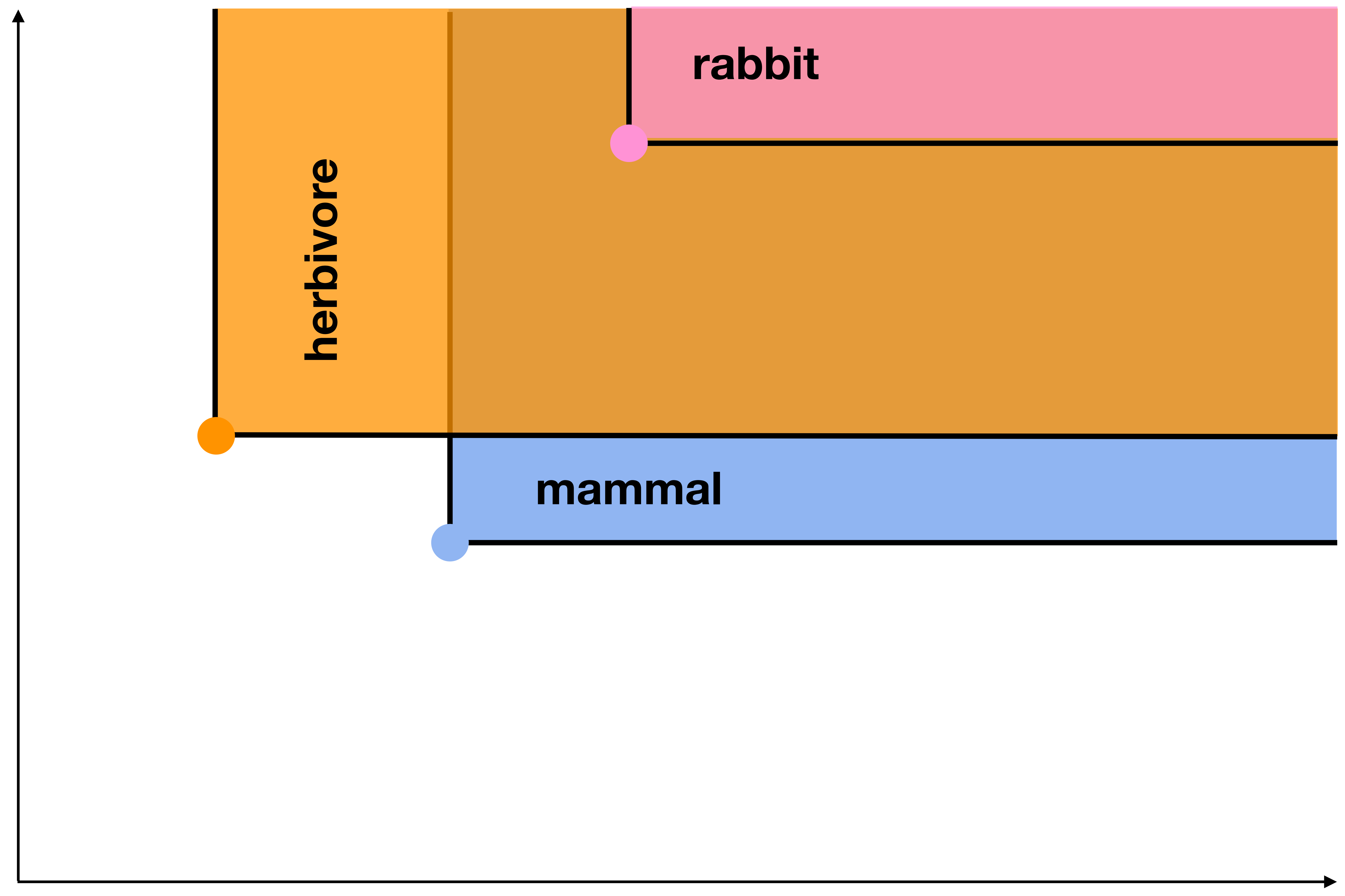
✓ Region ✓ Asymmetry ✓ Disjointness ✗ Closed under intersection



[Vendrov, Kiros, Fidler and Urtasun 2015]
[Lai and Hockenmaier 2017]

Cone Representation

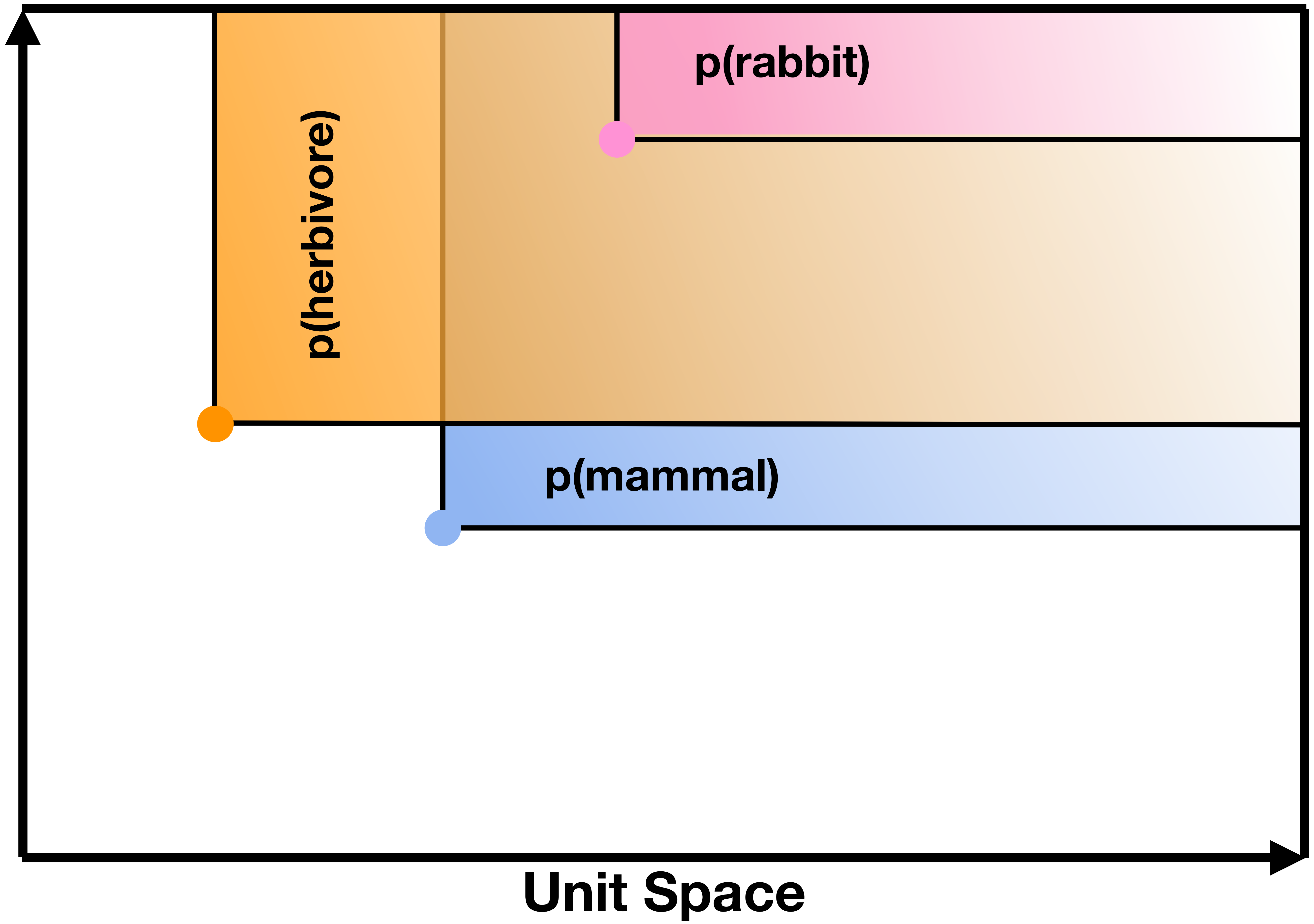
✓ Region ✓ Asymmetry ❌ Disjointness ✓ Closed under intersection



[Vendrov, Kiros, Fidler and Urtasun 2015]
[Lai and Hockenmaier 2017]

Cone Representation

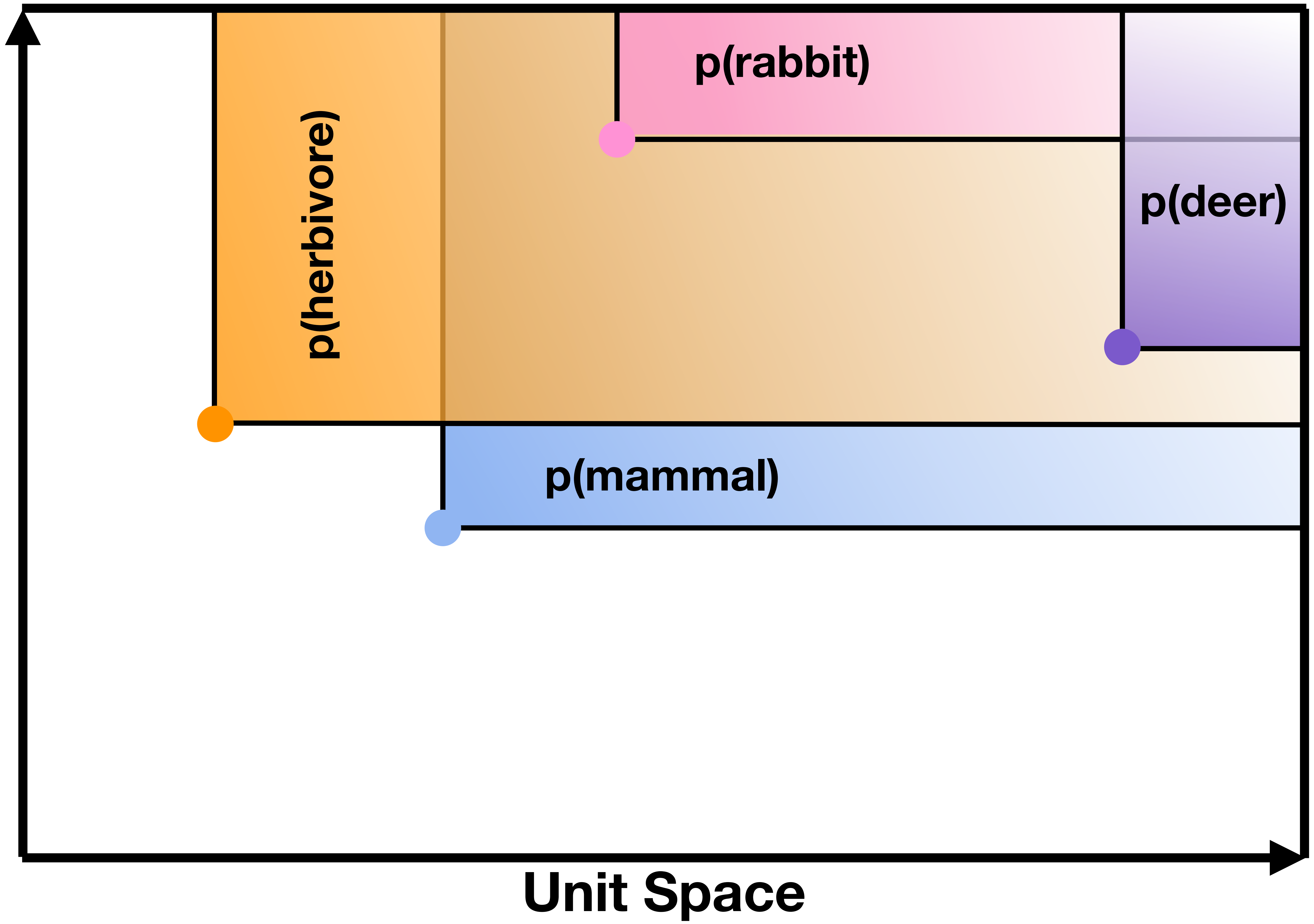
✓ Region ✓ Asymmetry ❌ Disjointness ✓ Closed under intersection



[Vendrov, Kiros, Fidler and Urtasun 2015]
[Lai and Hockenmaier 2017]

Cone Representation

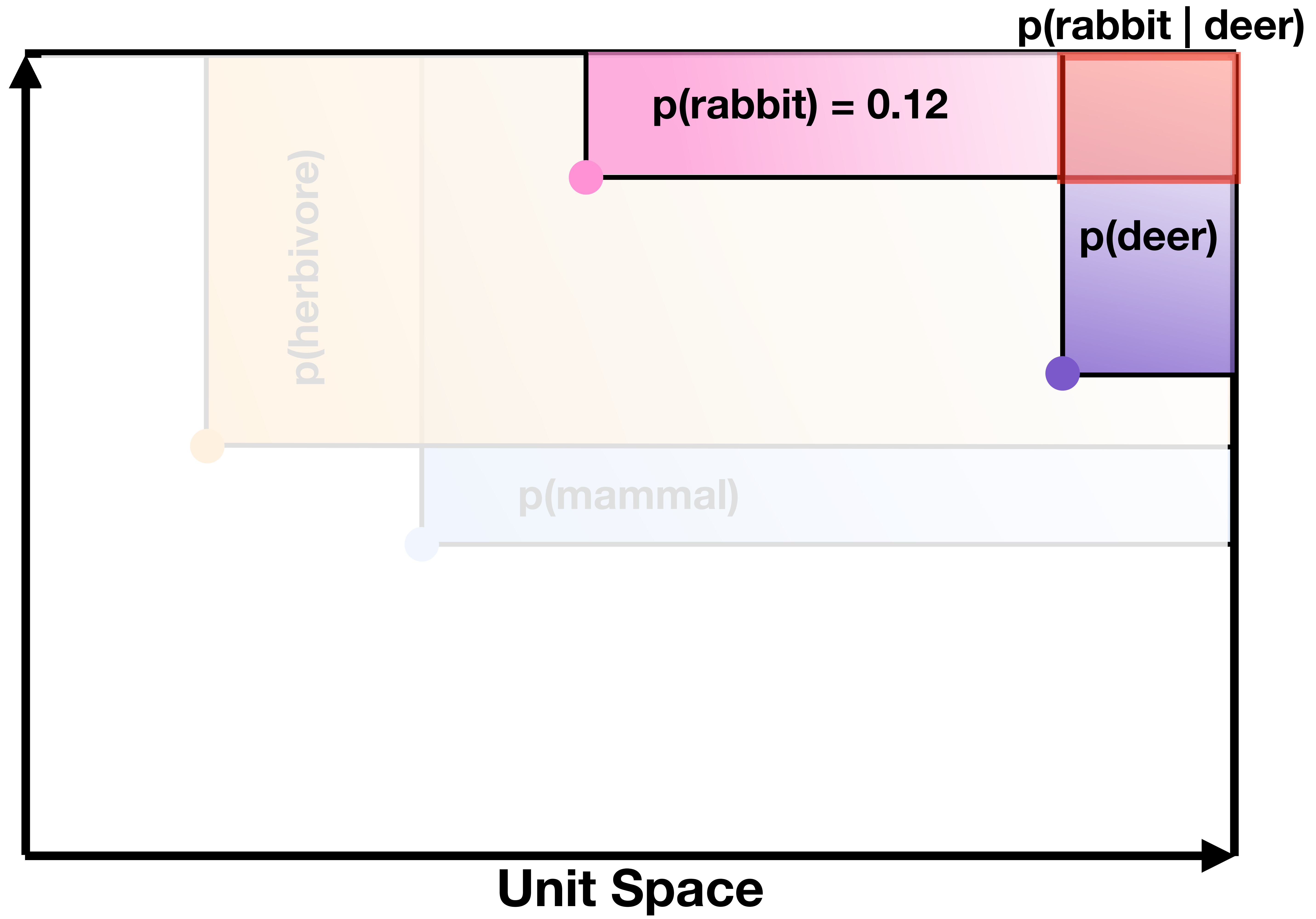
✓ Region ✓ Asymmetry ✗ Disjointness ✓ Closed under intersection



[Vendrov, Kiros, Fidler and Urtasun 2015]
[Lai and Hockenmaier 2017]

Cone Representation

✓ Region ✓ Asymmetry ❌ Disjointness ✓ Closed under intersection



[Vendrov, Kiros, Fidler and Urtasun 2015]
[Lai and Hockenmaier 2017]

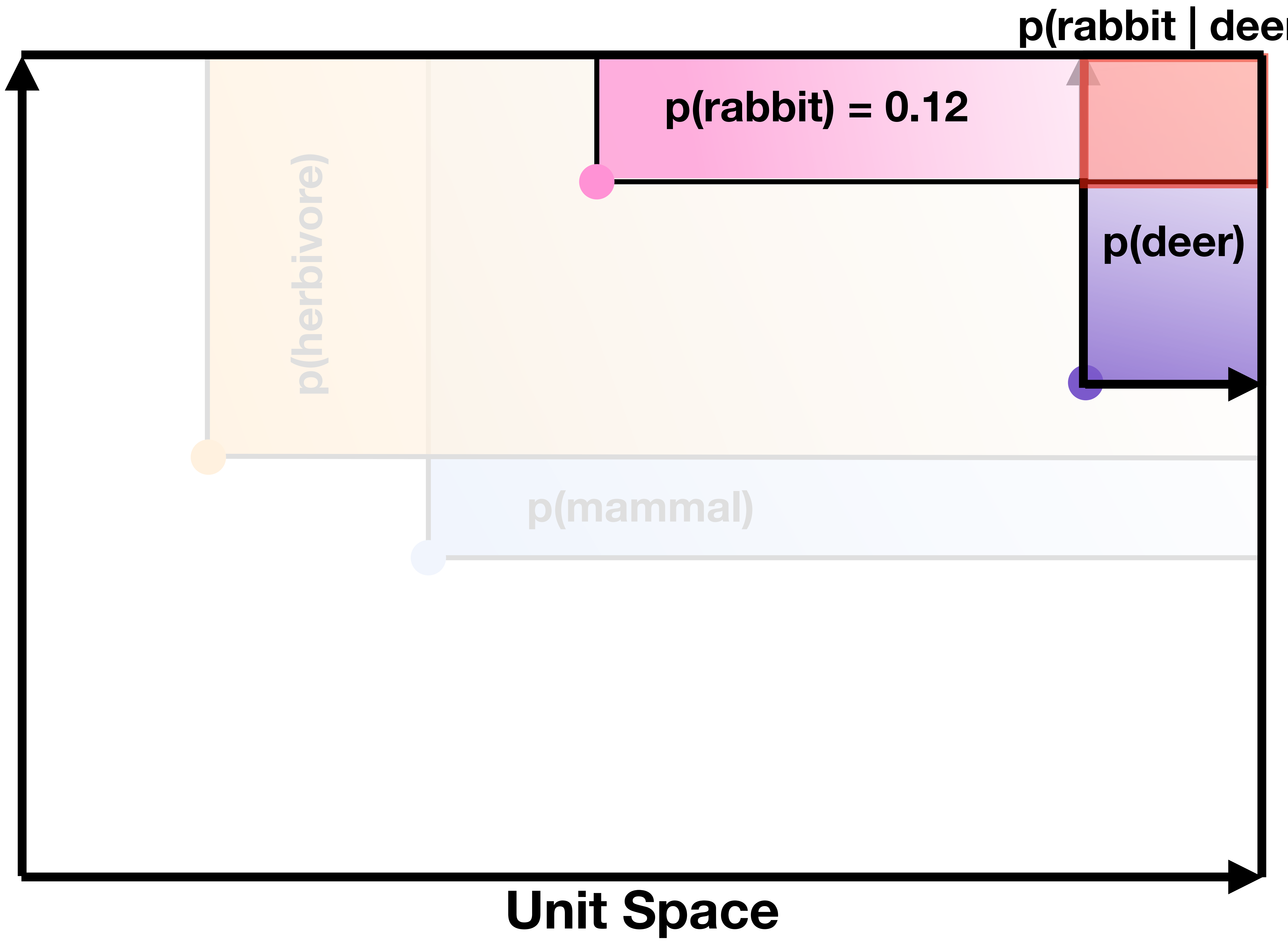
Cone Representation

✓ Region ✓ Asymmetry ✗ Disjointness ✓ Closed under intersection

[Vendrov, Kiros, Fidler and Urtasun 2015]
[Lai and Hockenmaier 2017]

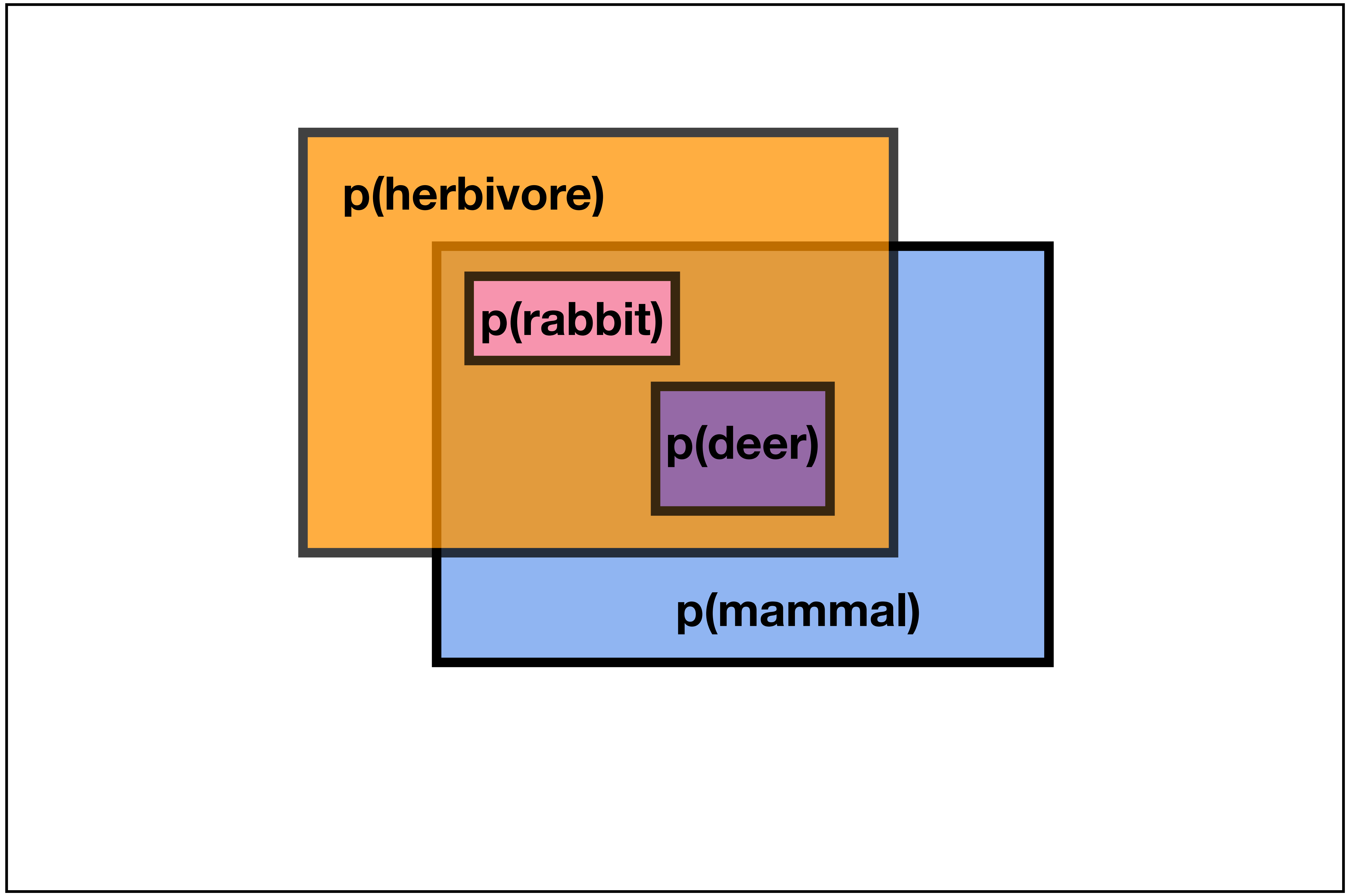
Cone Representation

✓ Region ✓ Asymmetry ✗ Disjointness ✓ Closed under intersection



Box Representation

✓ Region ✓ Asymmetry ✓ Disjointness ✓ Closed under intersection

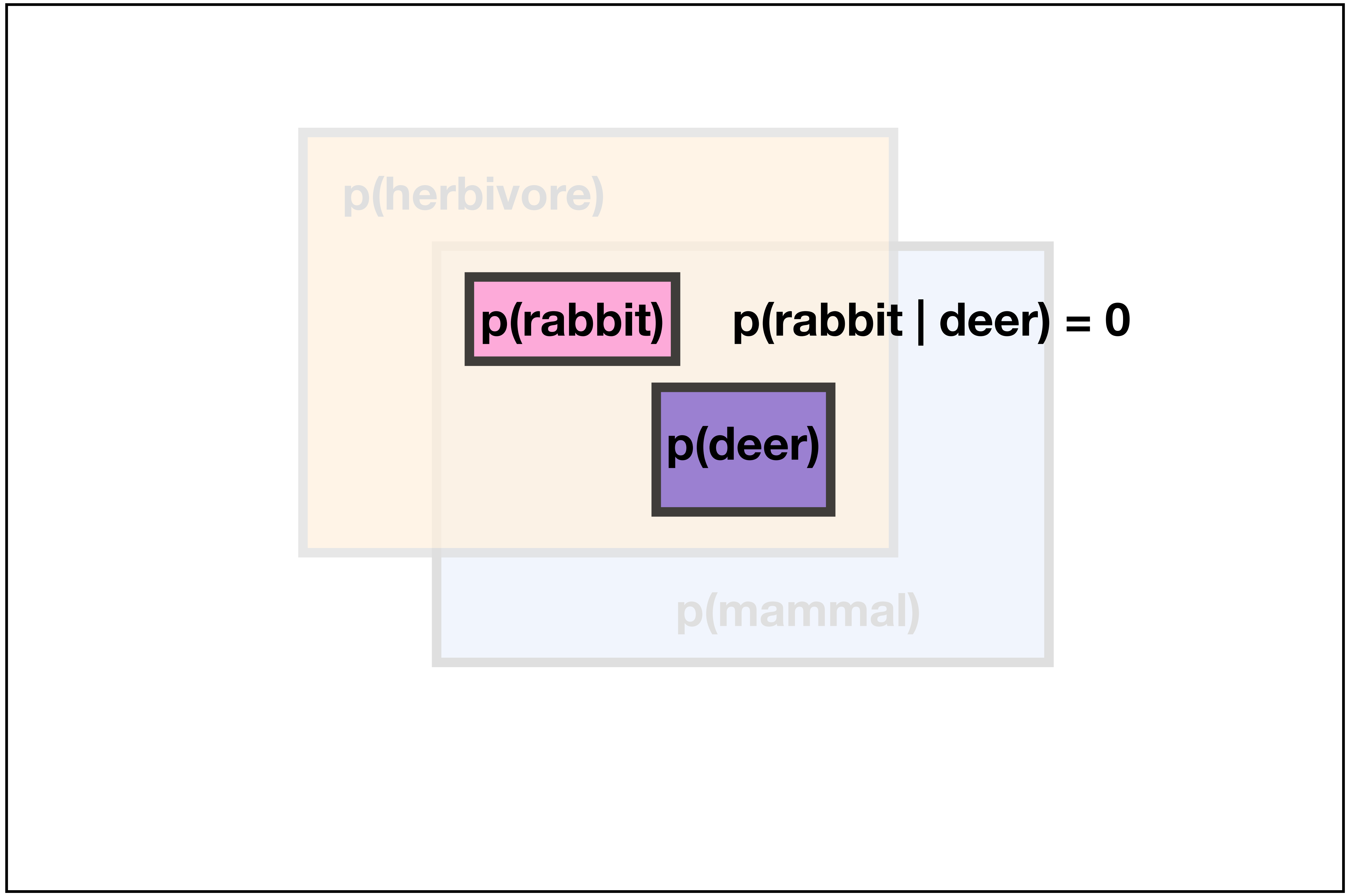


Unit Box

[Vilnis, Li, Murty and McCallum 2018]

Box Representation

✓ Region ✓ Asymmetry ✓ Disjointness ✓ Closed under intersection

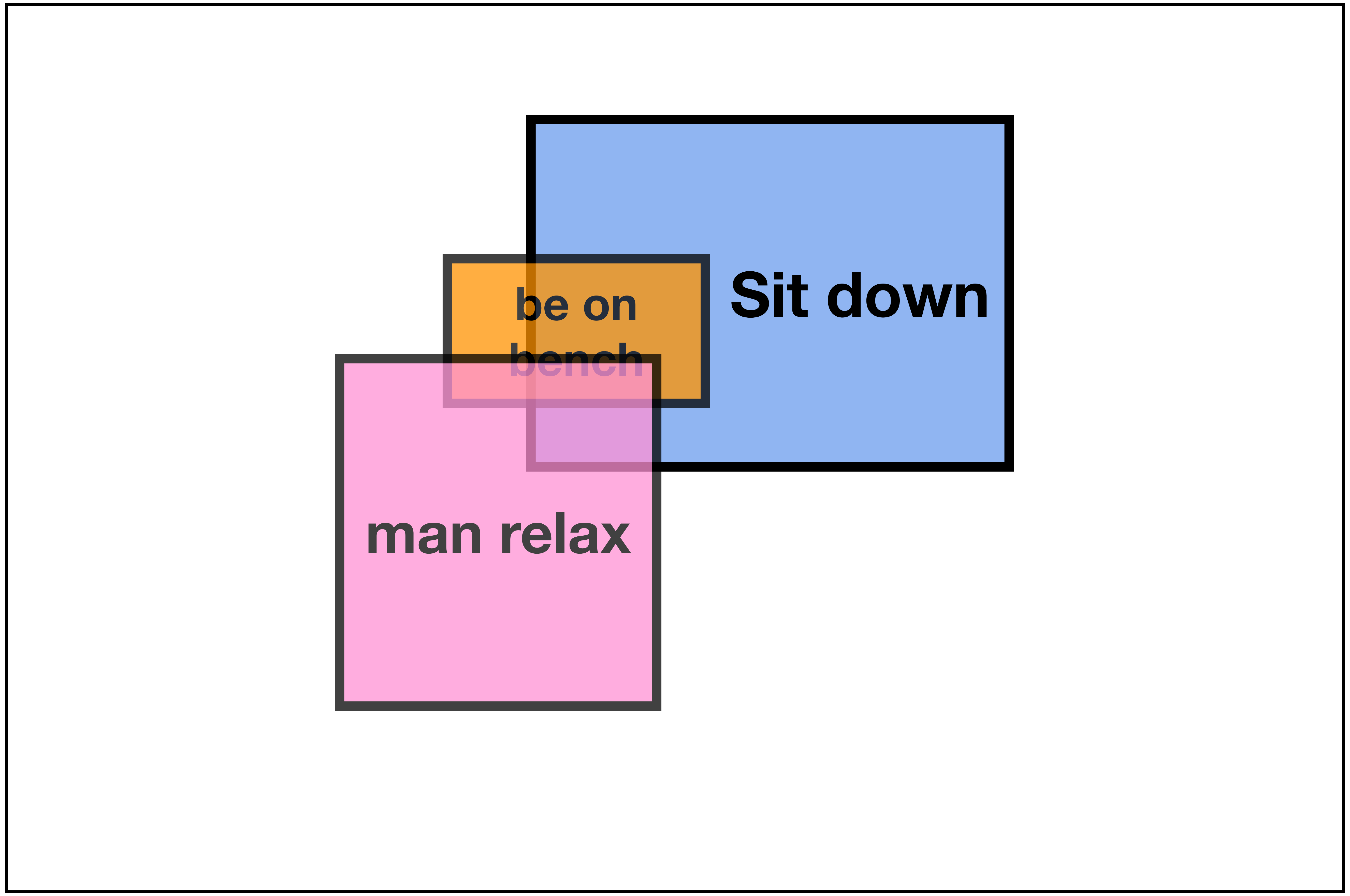


Unit Box

[Vilnis, Li, Murty and McCallum 2018]

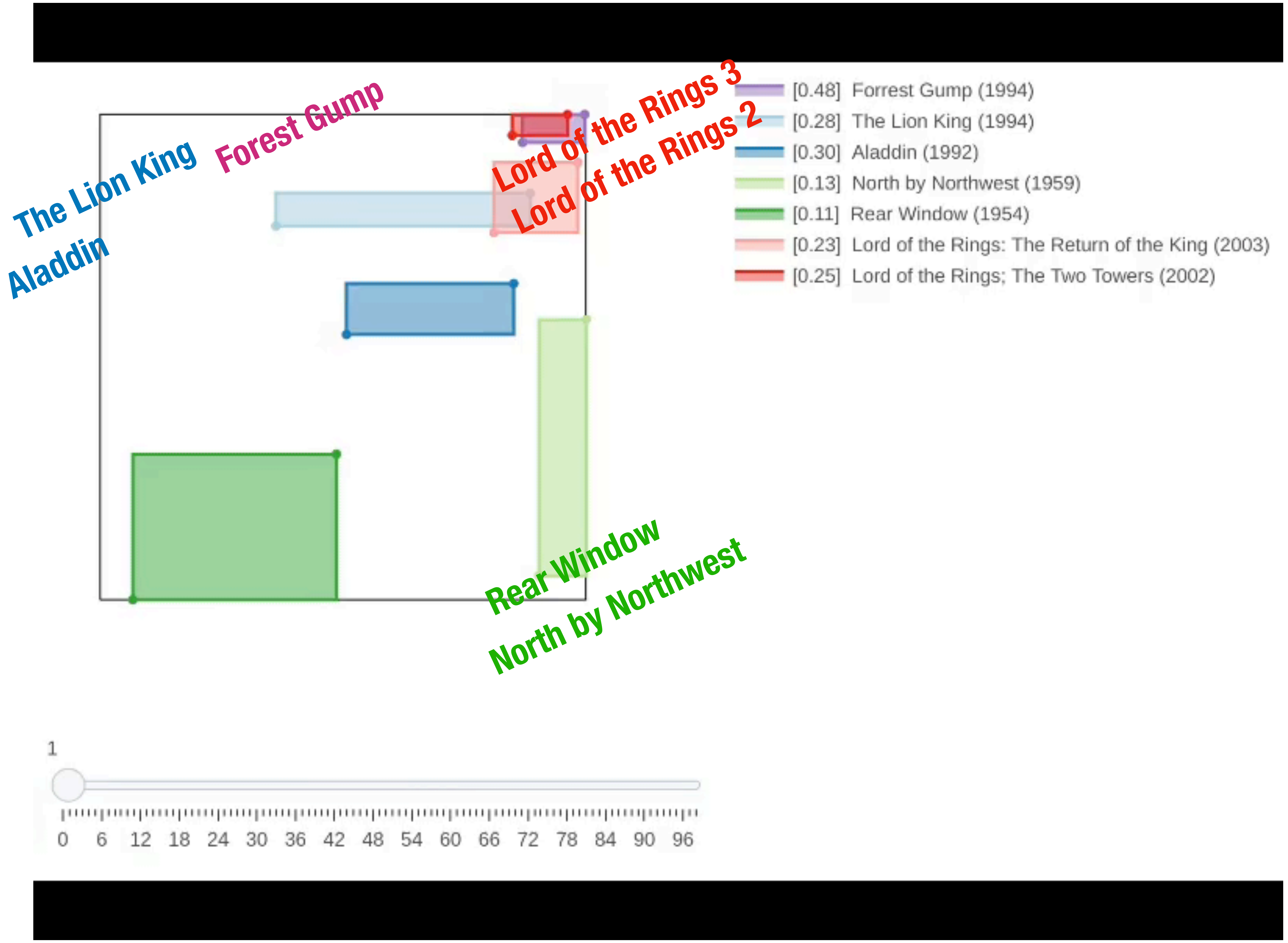
Box Representation

Common Sense



Unit Box

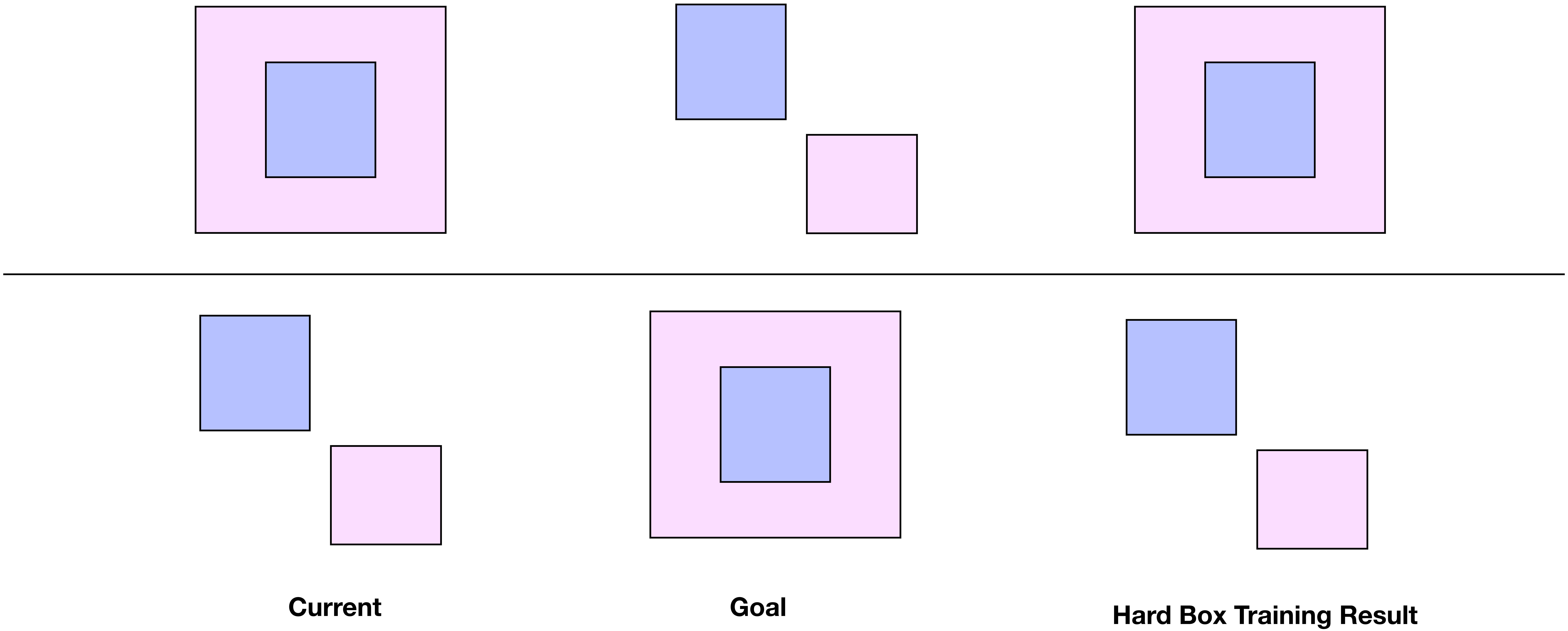
Box Representation Training Video



Video Credit: Michael Boratko.

Dasgupta, Shib Sankar, et al. "Improving Local Identifiability in Probabilistic Box Embeddings." *NeurIps* (2020).

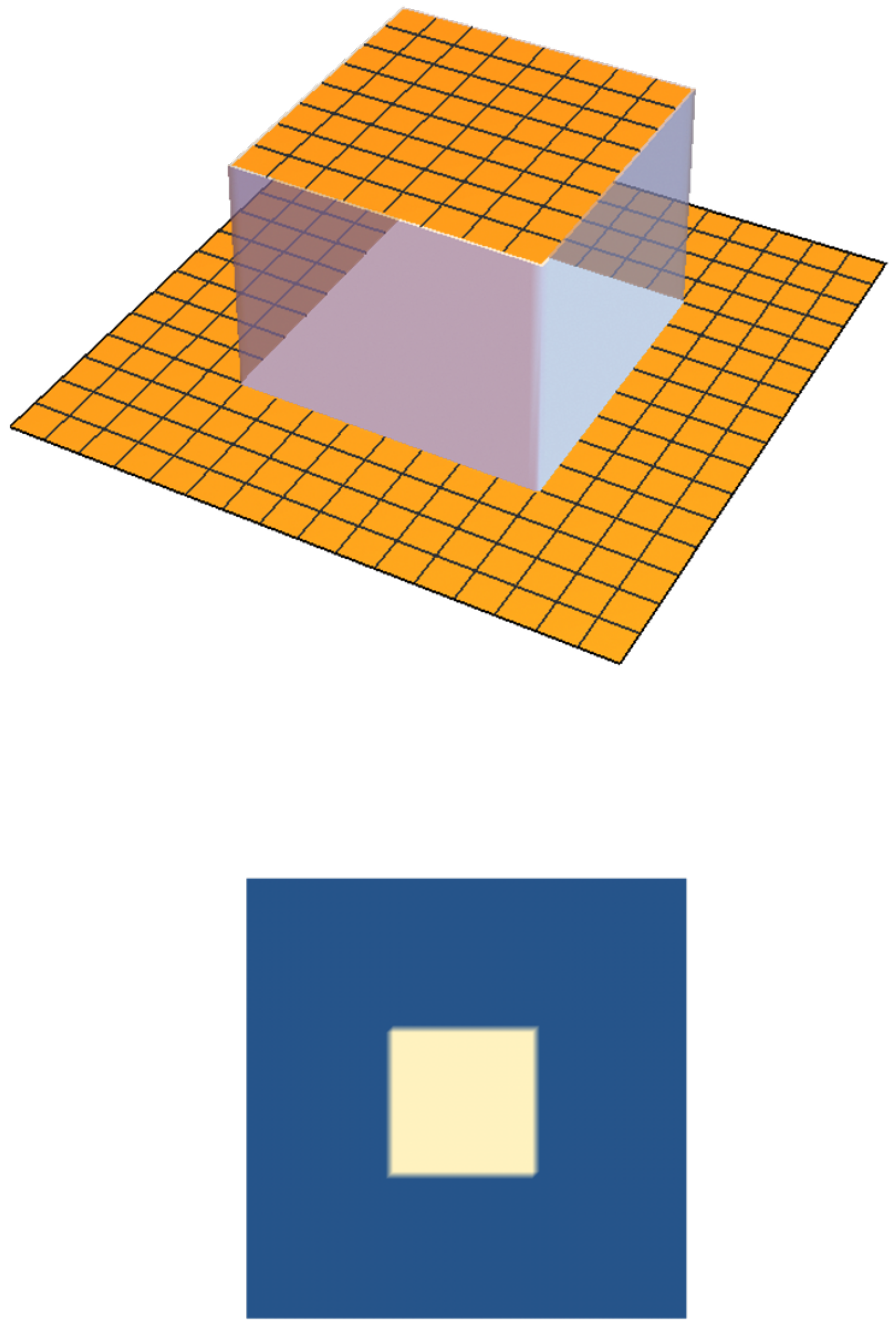
Box Training Difficulty



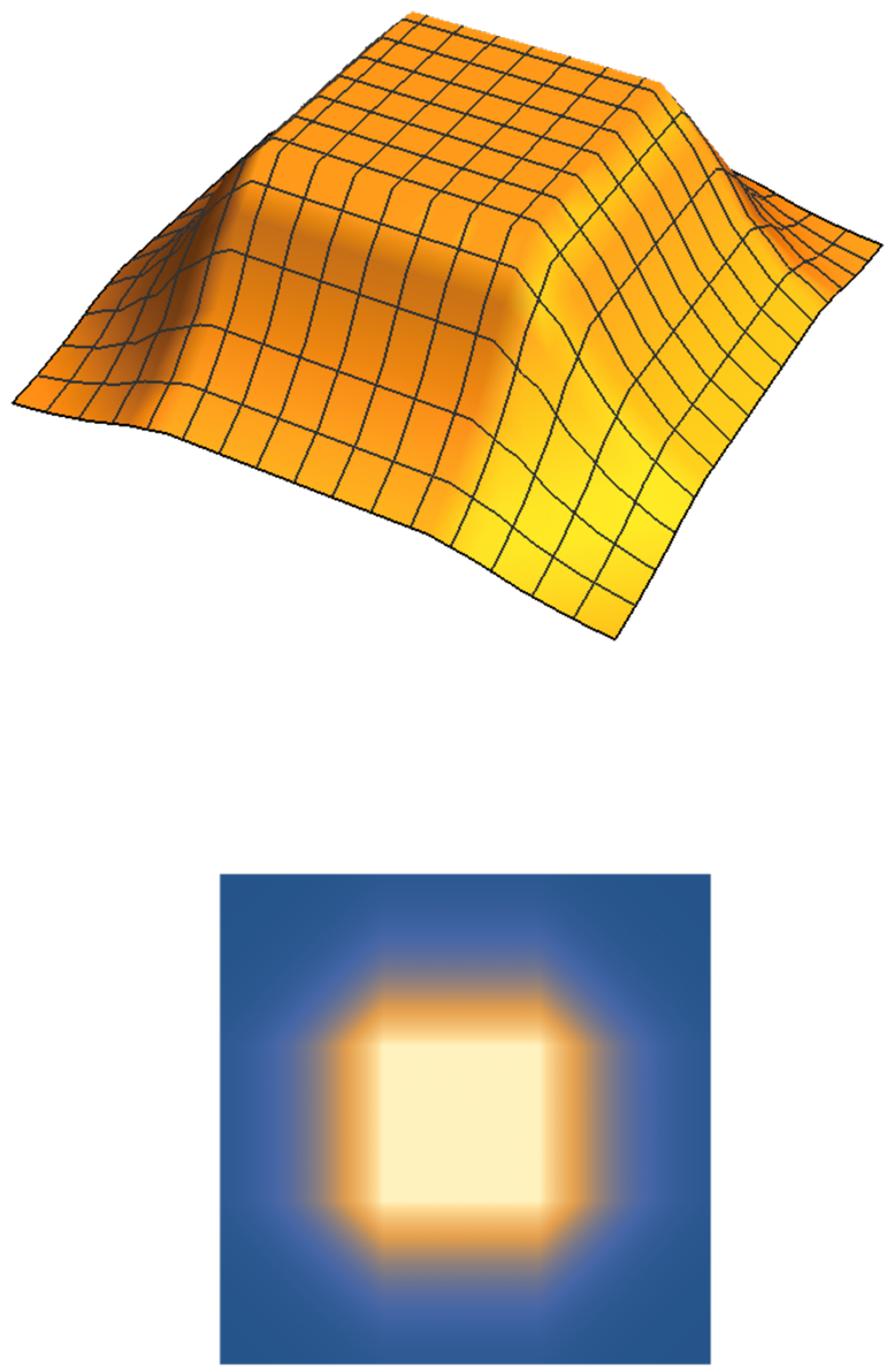
Li, Xiang, et al. "Smoothing the geometry of probabilistic box embeddings." *International Conference on Learning Representations*. 2019.

Dasgupta, Shib Sankar, et al. "Improving Local Identifiability in Probabilistic Box Embeddings." *NeurIps* (2020).

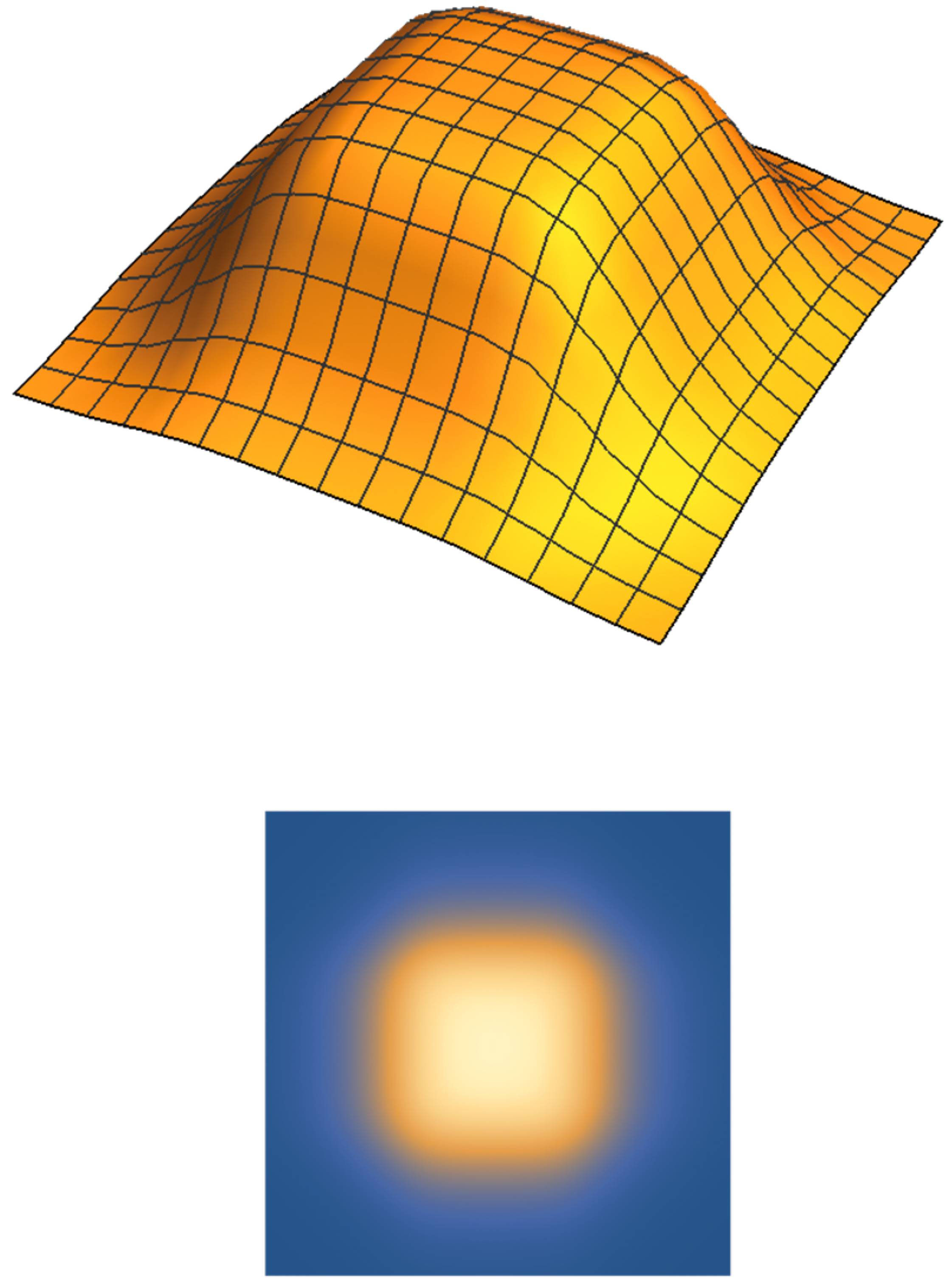
Box Training Loss



Hard Box (ACL 2018)

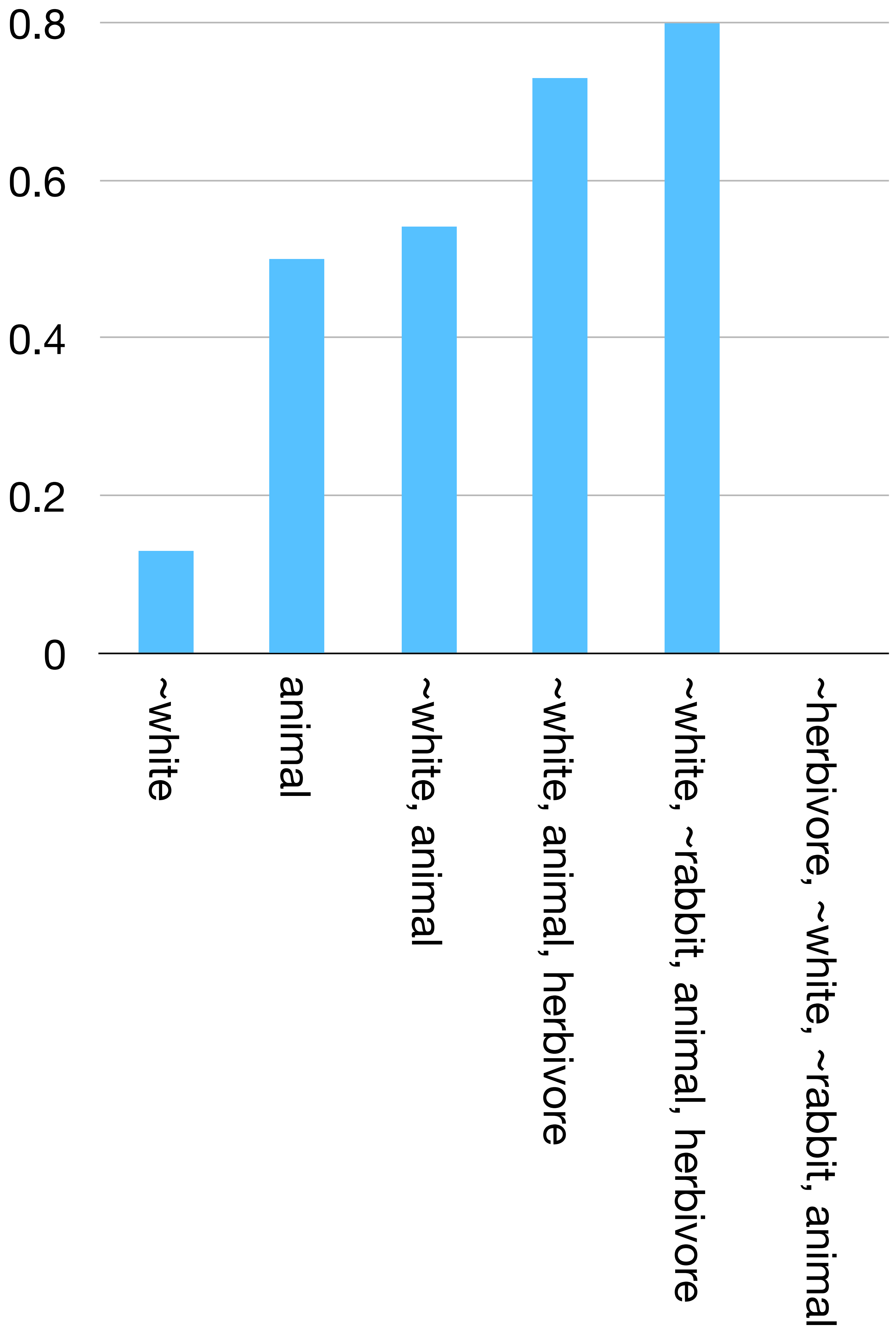


Smoothed Box (ICLR 2019)



Gumbel Box (NeurIps 2021)

Words examples



	Deer
P(deer)	0.11
~white	0.13
animal	0.50
~white, animal	0.54
~white, animal, herbivore	0.73
~white, ~rabbit, animal, herbivore	0.80
~herbivore, ~white, ~rabbit, animal	0.00

Sentence examples

- Flickr dataset is an entailment dataset containing 45 million image captions.

- Examples

x	p(x)	y	p(y)	p(x|y)
person walk	0.11516	blond woman walk down sidewalk	1.6E-04	1.0
person wear clothing	0.43036	adult dance on floor	3.9E-04	0.9
man play percussion instrument	0.00347	drummer	3.4E-03	0.51
man wear jacket	0.03077	snow on ground	5.1E-04	0.31
in basement	4.3E-04	hold instrument	5.9E-03	0.0067

Lai, Alice, and Julia Hockenmaier. "Learning to predict denotational probabilities for modeling entailment." *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Vol. 1. 2017.

Outline

- Commonsense Knowledge.
- Learn the Right Representation.
- Commonsense Knowledge in Pre-trained LMs.
- Benchmark Datasets for Evaluation.

Do Pre-trained LMs **Already** Capture Commonsense Knowledge?

OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

A robot wrote this entire article. Are you scared yet, human?

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below

New York Times Opinion @nytopinion · Jul 30

"GPT-3 is capable of generating entirely original, coherent and sometimes even factual prose," writes @fmanjoo. "And not just prose — it can write poetry, dialogue, memes, computer code and who knows what else."

Opinion | How Do You Know a Human Wrote This?

Machines are gaining the ability to write, and they are getting terrifyingly good at it.

nytimes.com

Do Pre-trained LMs **Already** Capture Commonsense Knowledge?

Commonsense KB relations => Natural language template => Using LMs to query / score

- LAMA: Petroni et al. (EMNLP 2019)

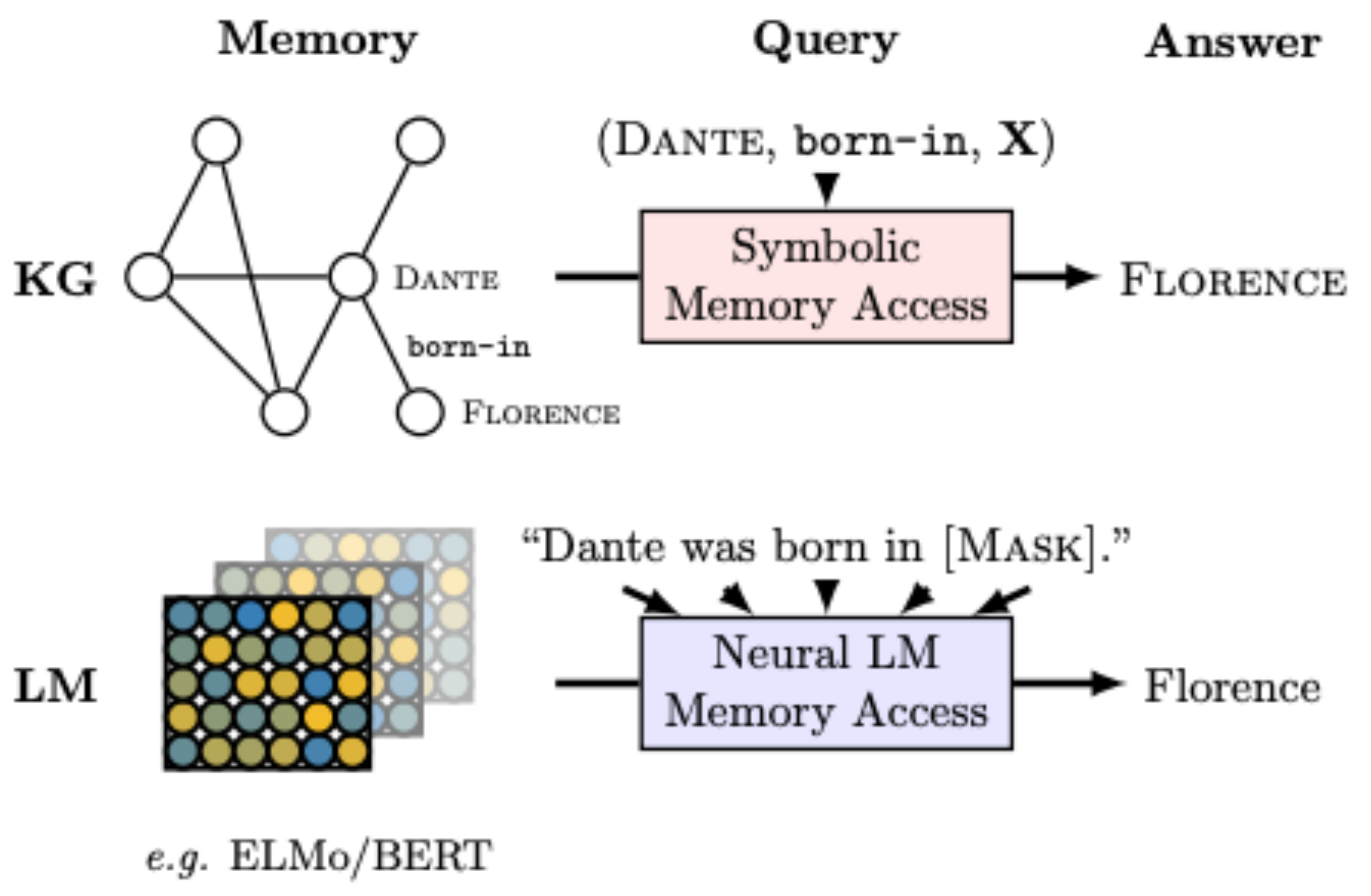


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

- Feldman et al. (EMNLP 2019)

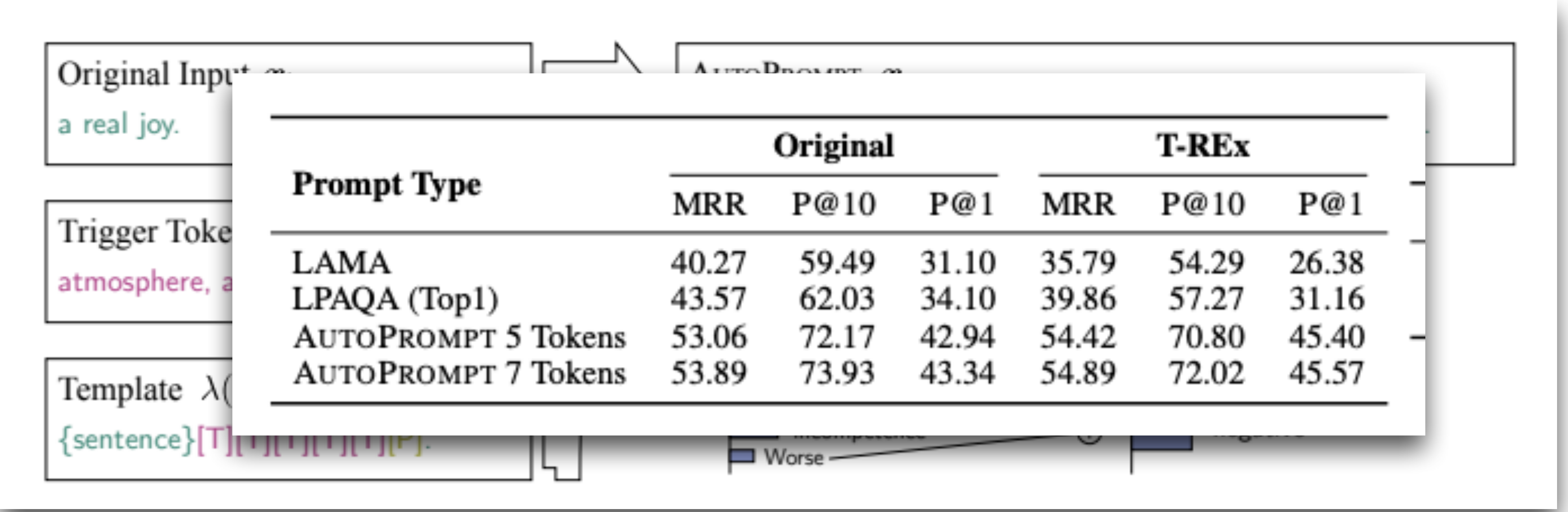
Candidate Sentence S_i	$\log p(S_i)$
"musician can playing musical instrument"	-5.7
"musician can be play musical instrument"	-4.9
"musician often play musical instrument"	-5.5
"a musician can play a musical instrument"	$\mathbf{-2.9}$

Table 1: Example of generating candidate sentences. Several enumerated sentences for the triple `(musician, CapableOf, play musical instrument)`. The sentence with the highest log-likelihood according to a pretrained language model is selected.

Adopted from ACL Commonsense tutorial: Knowledge in LMs.

Does the prompt matter?

- Yes! It matters! AutoPrompt (Shin et al., EMNLP 2020)
    - Generating gradient guided prompt.



Prompt Type	Original MRR	Original P@10	Original P@1	T-REx MRR	T-REx P@10	T-REx P@1
LAMA	40.27	59.49	31.10	35.79	54.29	26.38
LPAQA (Top1)	43.57	62.03	34.10	39.86	57.27	31.16
AUTOPROMPT 5 Tokens	53.06	72.17	42.94	54.42	70.80	45.40
AUTOPROMPT 7 Tokens	53.89	73.93	43.34	54.89	72.02	45.57

Do Pre-trained LMs **Already** Capture Commonsense Knowledge?

Properties of Concepts (Weir et al., 2020)

1. Do pre-trained LM correctly **distinguish concepts associated with a given set of properties?**
2. Can pre-trained LMs be used to **list the properties associated with given concepts?**

Adopted from ACL Commonsense tutorial: Knowledge in LMs.

Do Pre-trained LMs **Already** Capture Commonsense Knowledge?

Properties of Concepts (Weir et al., 2020)

1. Do pre-trained LM correctly **distinguish concepts associated with a given set of properties?**

- **A ___ has fur.**
- **A ___ has fur, is big, and has claws.**
- **A ___ has fur, is big, and has claws, has teeth, is an animal, eats, is brown…**

Adopted from ACL Commonsense tutorial: Knowledge in LMs.

Do Pre-trained LMs **Already** Capture Commonsense Knowledge?

Properties of Concepts (Weir et al., 2020)

1. Do pre-trained LM correctly **distinguish concepts associated with a given set of properties?**

- Good performance, RoBERTa > BERT
- Perceptual (e.g. visual) < non-perceptual (e.g. encyclopaedic or functional).
- Highly-ranked incorrect answers typically apply to a subset of properties.

Adopted from ACL Commonsense tutorial: Knowledge in LMs.

Do Pre-trained LMs **Already** Capture Commonsense Knowledge?

Properties of Concepts (Weir et al., 2020)

1. Can pre-trained LMs be used to **list the properties associated with given concepts?**
   - Low correlation with human elicited properties, but coherent and mostly "verifiable by humans"

Context	**Human**		**RoBERTa-L**	
	Response	**PF**	**Response**	p_{LM}
(Everyone knows that) a bear has ___ .	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

Adopted from ACL Commonsense tutorial: Knowledge in LMs.

Can we trust knowledge from LMs?

- **LMs also generate fictitious facts!**

Distributionally-related:

Barack's Wife Hillary:
Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV[*] **Nelson F. Liu**[†§] **Matthew E. Peters**[§]
Matt Gardner[§] **Sameer Singh**[*]

[*] University of California, Irvine, CA, USA
[†] University of Washington, Seattle, WA, USA
[§] Allen Institute for Artificial Intelligence, Seattle, WA, USA
{rlogan, sameer}@uci.edu, {mattg, matthewp}@allenai.org, nfliu@cs.washington.edu

Syntactically-similar:

Negated and Misprimed Probes for Pretrained Language Models:
Birds Can Talk, But Cannot Fly

Nora Kassner, Hinrich Schütze
Center for Information and Language Processing (CIS)
LMU Munich, Germany
kassner@cis.lmu.de

Adopted from ACL Commonsense tutorial: Knowledge in LMs.

Outline

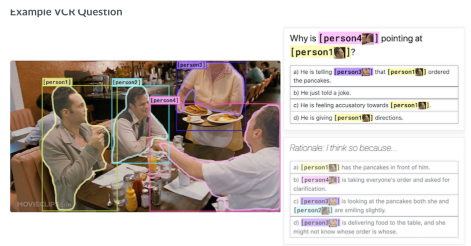
- Commonsense Knowledge.
- Learn the Right Representation.
- Commonsense Knowledge in Pre-trained LMs.
- Benchmark Datasets for Evaluation.

Benchmark Evaluation Dataset

- **Besides probing the model using commonsense knowledge bases, are there standard commonsense benchmark datasets to evaluate the model?**
  - **Question Answering.**
  - **Natural Language Inference.**
  - **Coreference Resolution.**
  - …

https://leaderboard.allenai.org/

Benchmark Evaluation Dataset

	Task Type	Domain	Example	Gap to Human Perfomance
HELLA SWAG	Multi Choice QA	Grounded commonsense.		95.6% - 93.85% = 1.75%
Abductive NLI	Multi Choice Selection	Abductive Reasoning	*Obs1: Jenny was addicted to sending text messages.* *Obs2: Jenny narrowly avoided a car accident.* *Hyp1:* Since her friend's texting and driving car accident, Jenny keeps her phone off while driving. *Hyp2:* **Jenny was looking at her phone while driving so she wasn't paying attention.**	92.90% - 89.70% = 3.2%
Cosmos QA	Multi Choice QA	Reading Comprehension	**Example 1.** *Paragraph:* It's a very humbling experience when you need someone to dress you every morning, tie your shoes, and put your hair up. Every menial task takes an unprecedented amount of effort. It made me appreciate Dan even more. But anyway I shan't dwell on this (I'm not dying after all) and not let it detact from my lovely 5 days with my friends visiting from Jersey. *Question:* What's a possible reason the writer needed someone to dress him every morning? *Option1:* The writer doesn't like putting effort into these tasks. *Option2:* **The writer has a physical disability.** *Option3:* The writer is bad at doing his own hair.	94% - 91.79% = 2.3%
Physical IQa	I			= 4.77%
Social IQa	I			= 4.95%
WinoGrande	I			2.72%
VCR	Multi Choice QA	Vision & Language		85% - 77.79% = 2.3%

Example HellaSwag Question

A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She

- a) rinses the bucket off with soap and blow dries the dog's head.
- b) uses a hose to keep it from getting soapy.
- **c) gets the dog wet, then it runs away again.**
- d) gets into the bath tub with the dog.

https://leaderboard.allenai.org/

Benchmark Evaluation Dataset

	Task Type	Domain	Example	Gap to Human Perfomance
HELLA SWAG	Multi Choice QA	Grounded commonsense.		95.6% - 93.85% = 1.75%
Abductive NLI	Multi Choice Selection	Abductive Reasoning		92.90% - 89.70% = 3.2%
Cosmos QA	Multi Choice QA	Reading Comprehension		94% - 91.79% = 2.3%
Physical IQa	Multi Choice QA	Naive Physical reasoning	You need to break a window. Which object would you rather use? a) a metal stool … c) a bottle of water	94.9 … .13% = 4.77%
Social IQa	Multi Choice QA	Social commonsense		88.1% - 83.15% = 4.95%
WinoGrande	Multi Choice Selection	Coreference resolution		94% - 91.28% = 2.72%
VCR	Multi Choice QA	Vision & Language		85% - 77.79% = 2.3%

Generative Evaluation

https://leaderboard.allenai.org/

Benchmark Evaluation Dataset

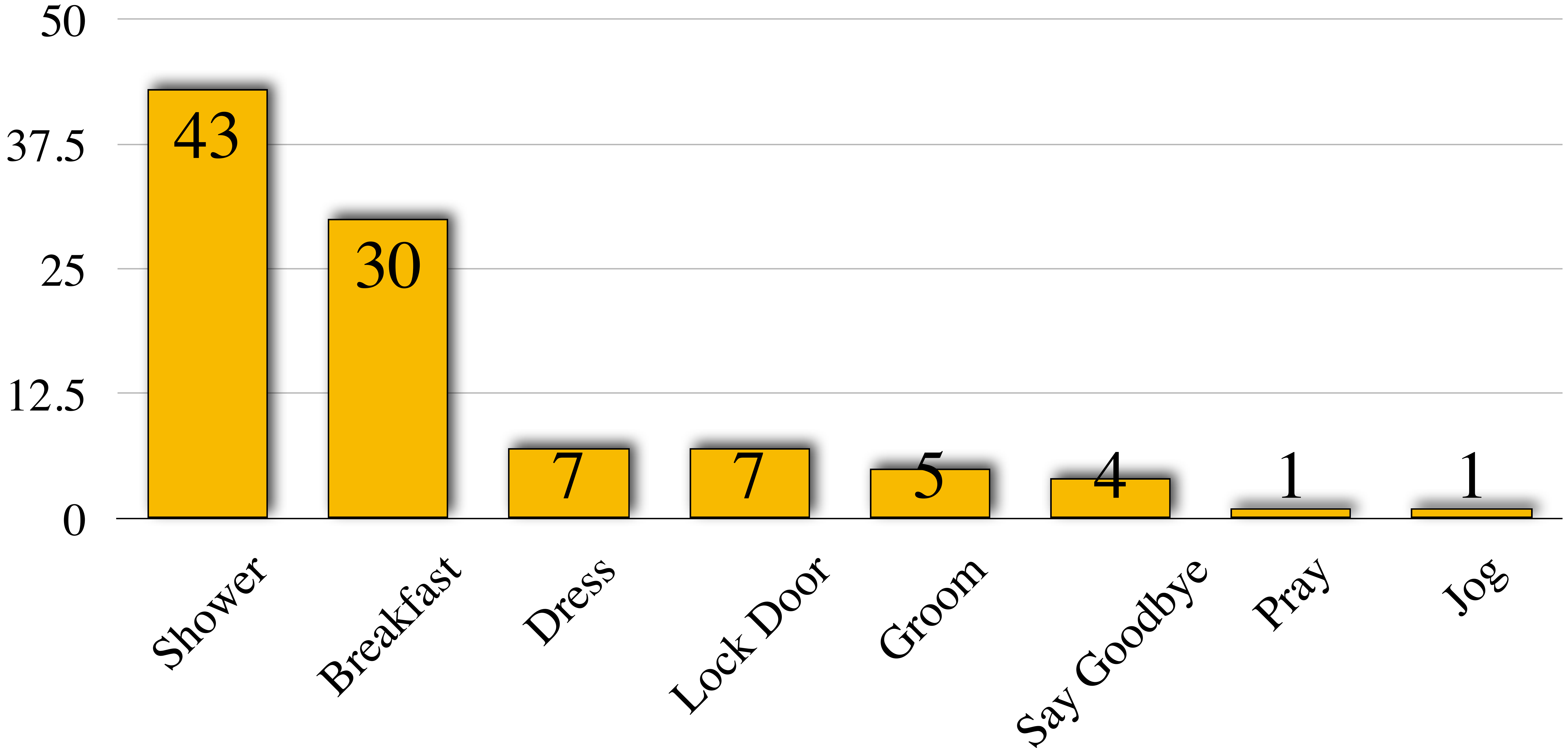
Generative Evaluation

✦ **ProtoQA (EMNLP 2020)**: dataset that captures prototypical situation.

✓ Multiple correct answers.

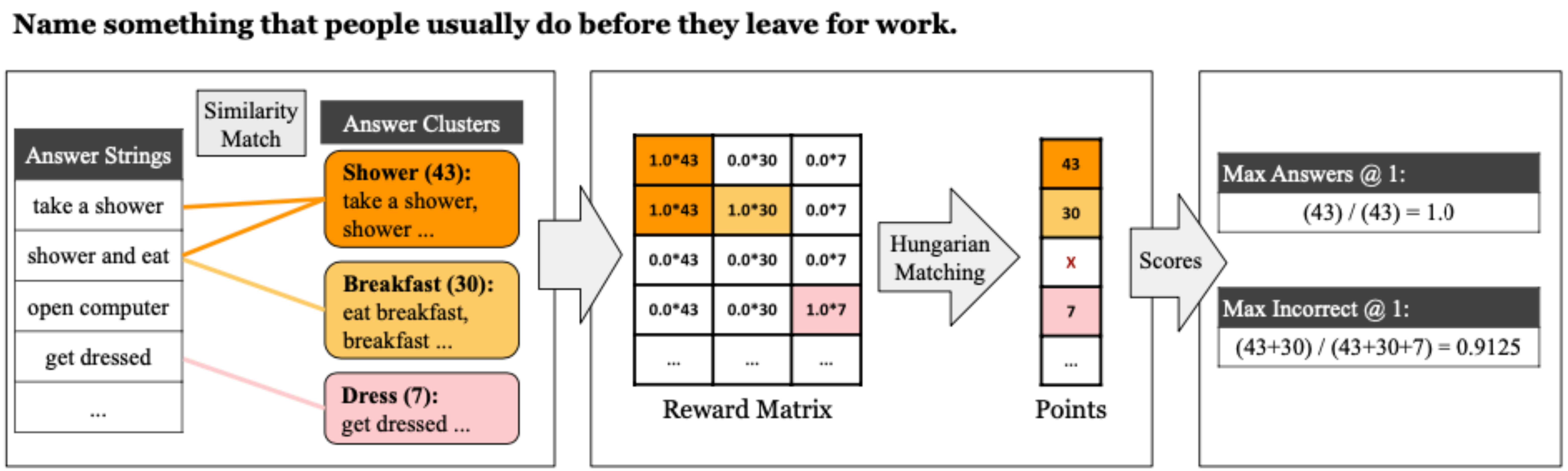
✓ Scores for each answer.

Name something that people usually do before they leave the house for work?

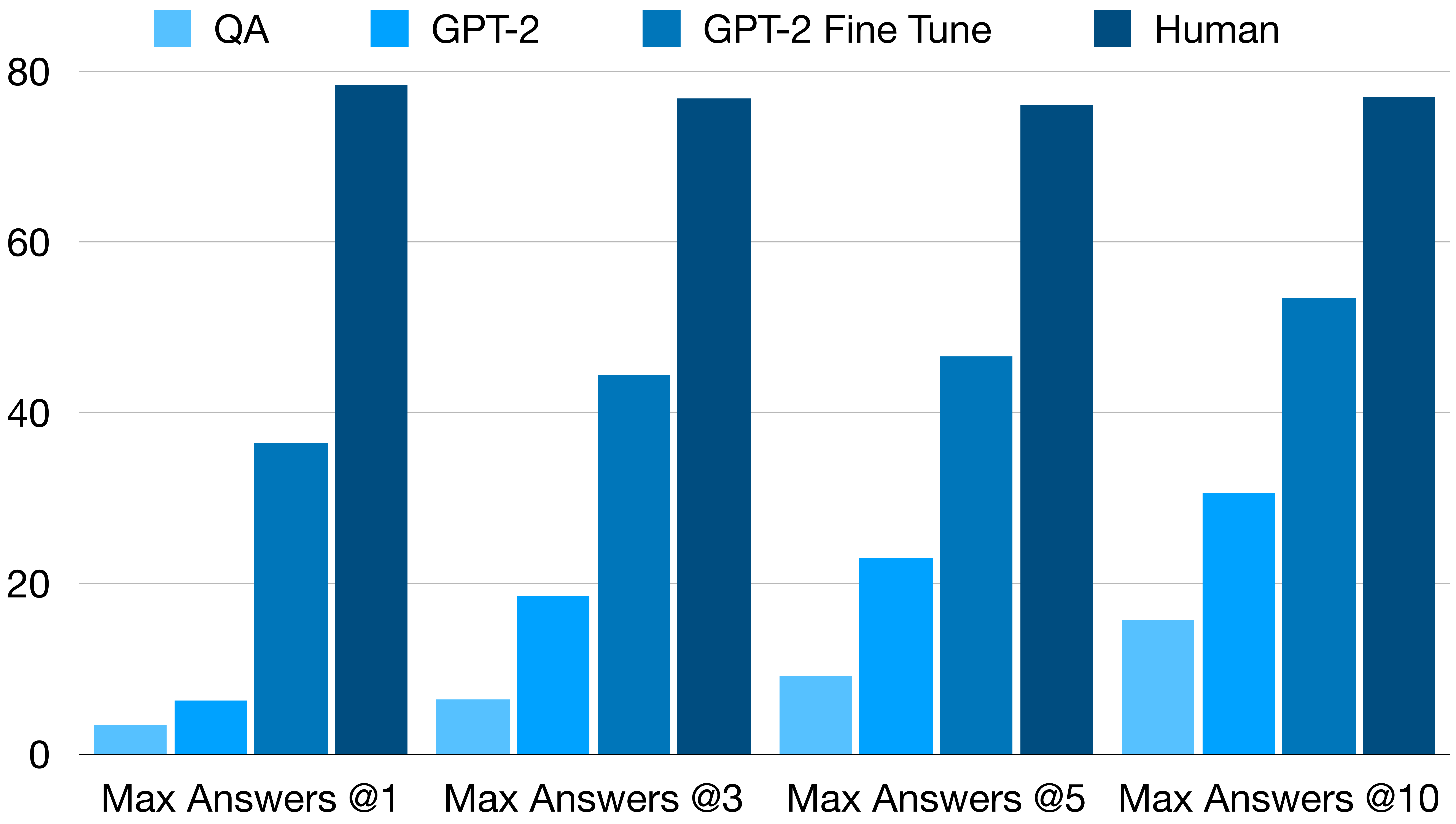


ProtoQA (EMNLP 2020)

- ✦ **Generative Evaluation**
  - ✓ Evaluate <u>multiple</u> correct answers generative by the model.
  - ✓ Reward models with <u>correct ranking of answer list</u>.
  - ✓ Reward models with <u>higher coverage of answer list</u>.



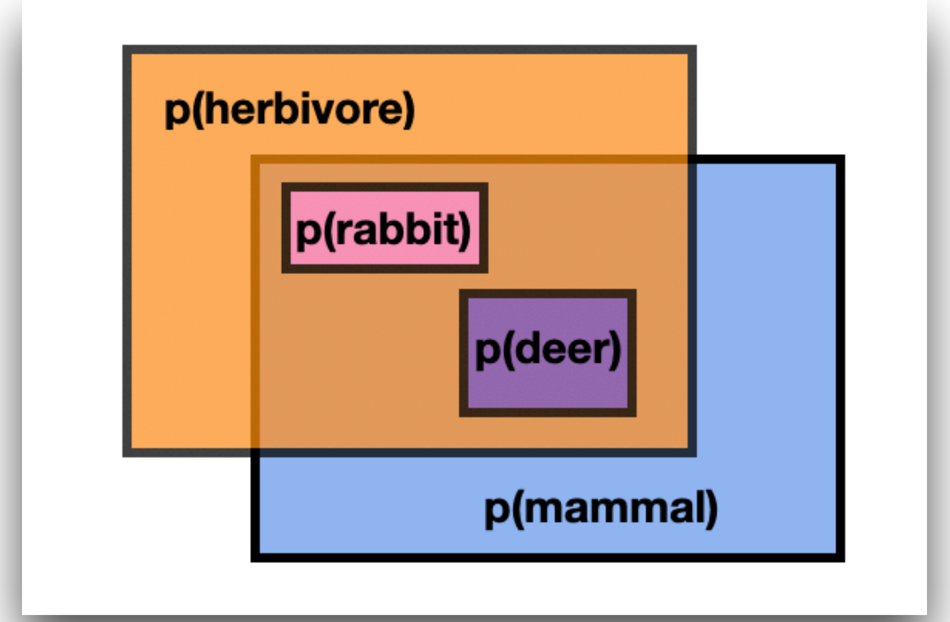
Results



Numbers reported are percentage of perfect score, i.e. answering with a list with an element from each answer cluster in decreasing order would yield 100.

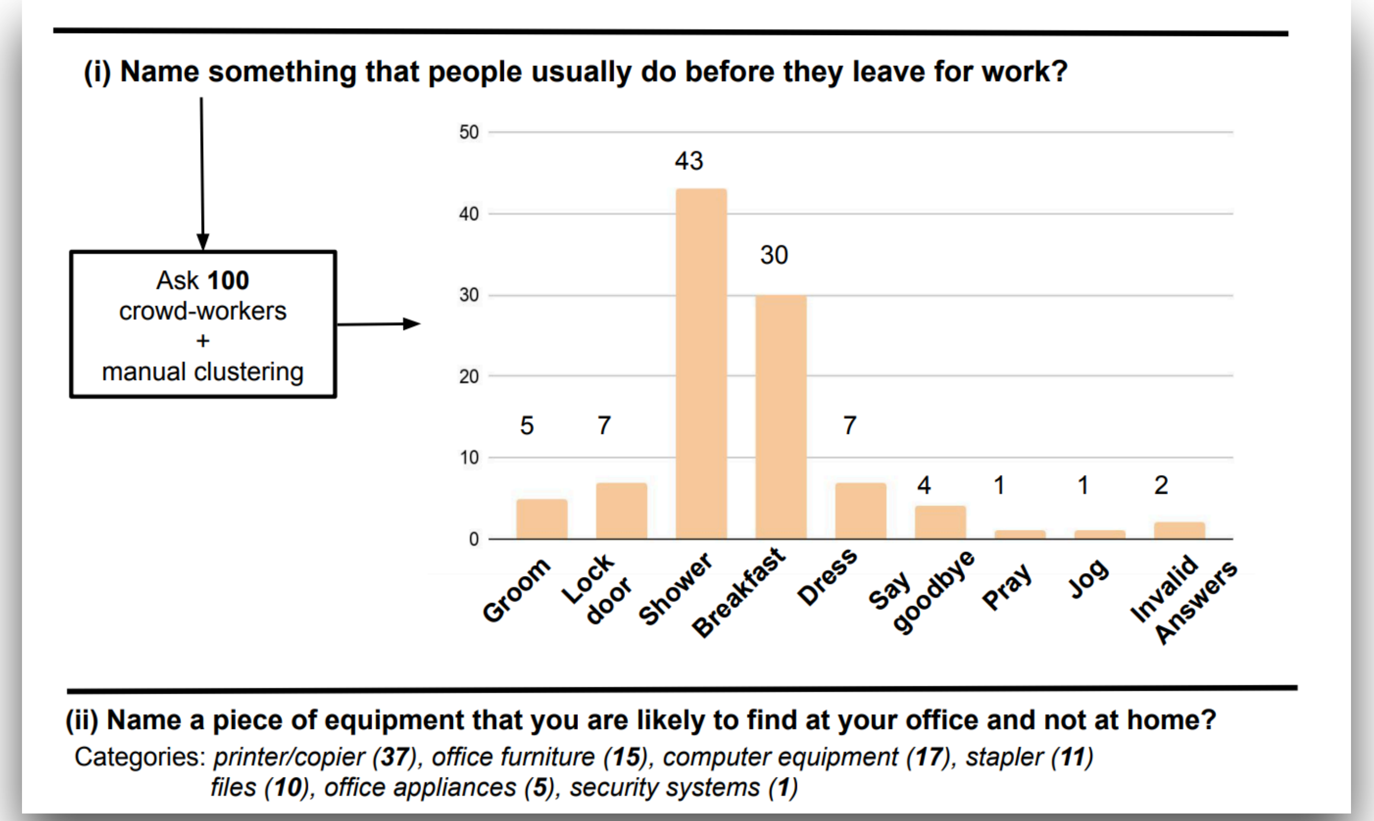
Summary

- Commonsense Knowledge.
- Learn the Right Representation.
- Commonsense Knowledge in Pre-trained LMs.
- Benchmark Datasets for Evaluation.



Candidate Sentence S_i	$\log p(S_i)$
"musician can playing musical instrument"	-5.7
"musician can be play musical instrument"	-4.9
"musician often play musical instrument"	-5.5
"a musician can play a musical instrument"	$\mathbf{-2.9}$

Table 1: Example of generating candidate sentences. Several enumerated sentences for the triple (musician, CapableOf, play musical instrument). The sentence with the highest log-likelihood according to a pretrained language model is selected.

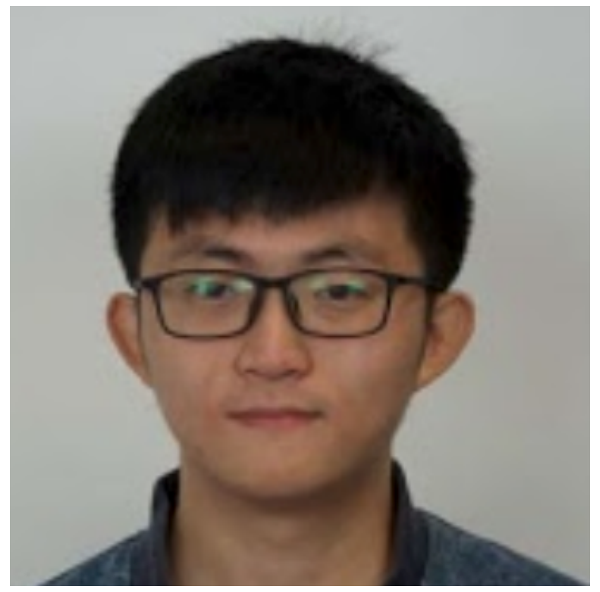


Thanks to all the collaborators!

Dan Le

Shikhar Murty

Shib Sankar Dasgupta



Dongxu Zhang

Rajarshi Das

UNIVERSITY OF MASSACHUSETTS AMHERST • 1863

Luke Vilnis

Michael Boratko

Tim O'Gorman

Andrew McCallum