Towards Interpretable Neuro-symbolic AI(AGI)

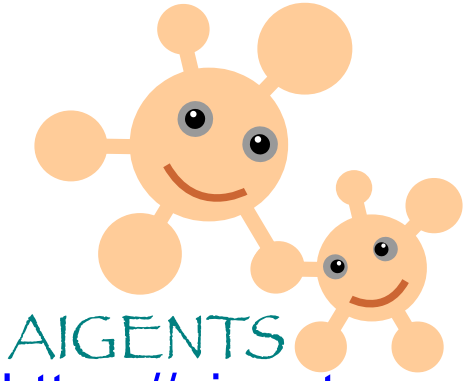
Anton Kolonin

akolonin@aigents.com

Facebook: akolonin

Telegram: akolonin

https://agirussia.org



https://aigents.com

All You Need Is ...

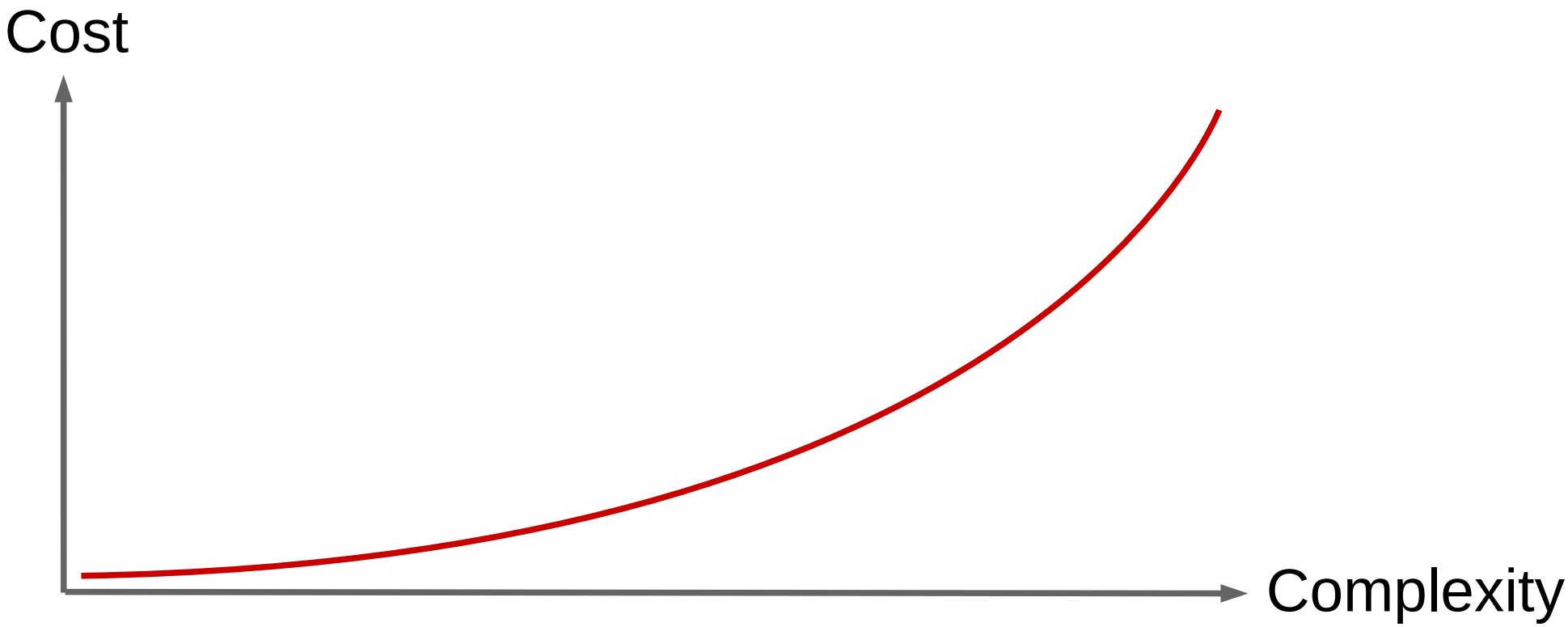
- Cost Efficiency
- Structural Learning
- Hierarchical Planning
- Handling Uncertainty
- Interpretability

A Path Towards Autonomous Machine Intelligence
Yann LeCun
https://openreview.net/pdf?id=BZ5a1r-kVsf

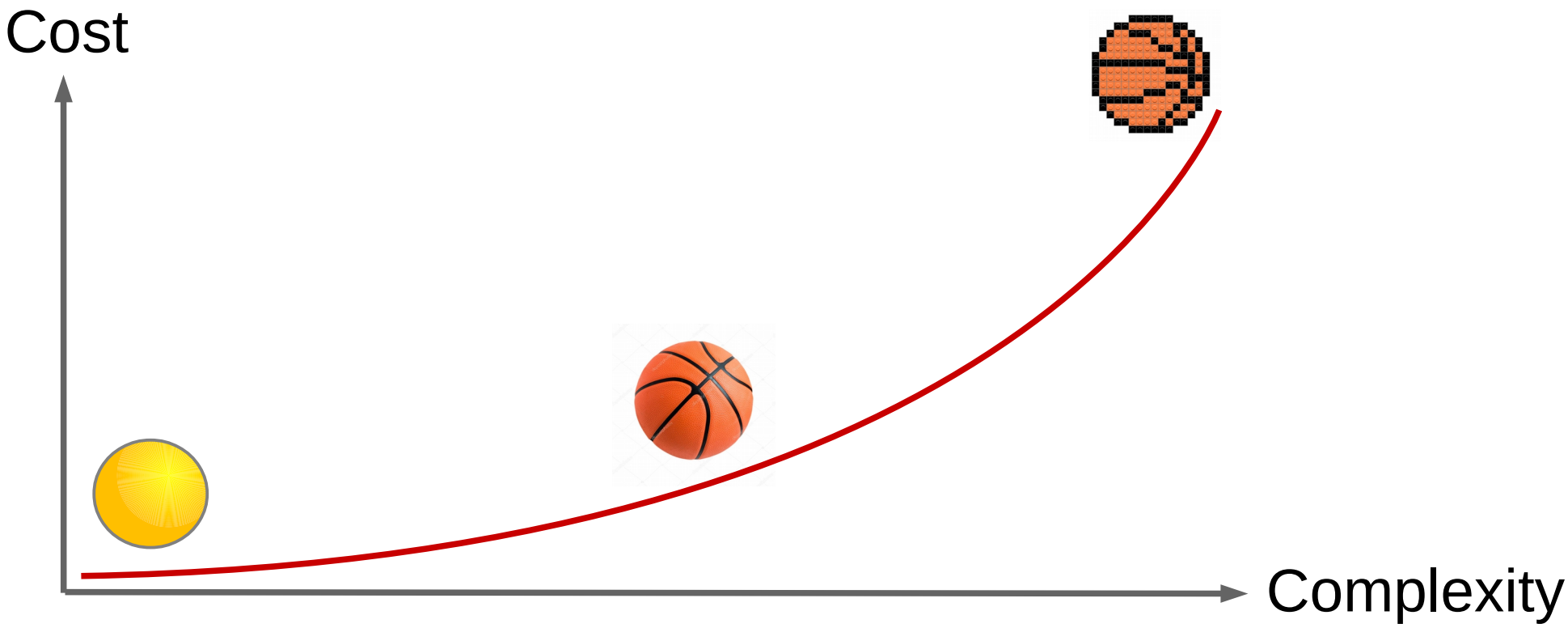
Unsupervised Learning of Temporal Abstractions with Slot-based Transformers
Anand Gopalakrishnan, Kazuki Irie, Jürgen Schmidhuber, Sjoerd van Steenkiste
https://arxiv.org/abs/2203.13573

Active Inference: The Free Energy Principle in Mind, Brain, and Behavior
Karl J. Friston
https://www.amazon.com/Active-Inference-Energy-Principle-Behavior/dp/0262045354

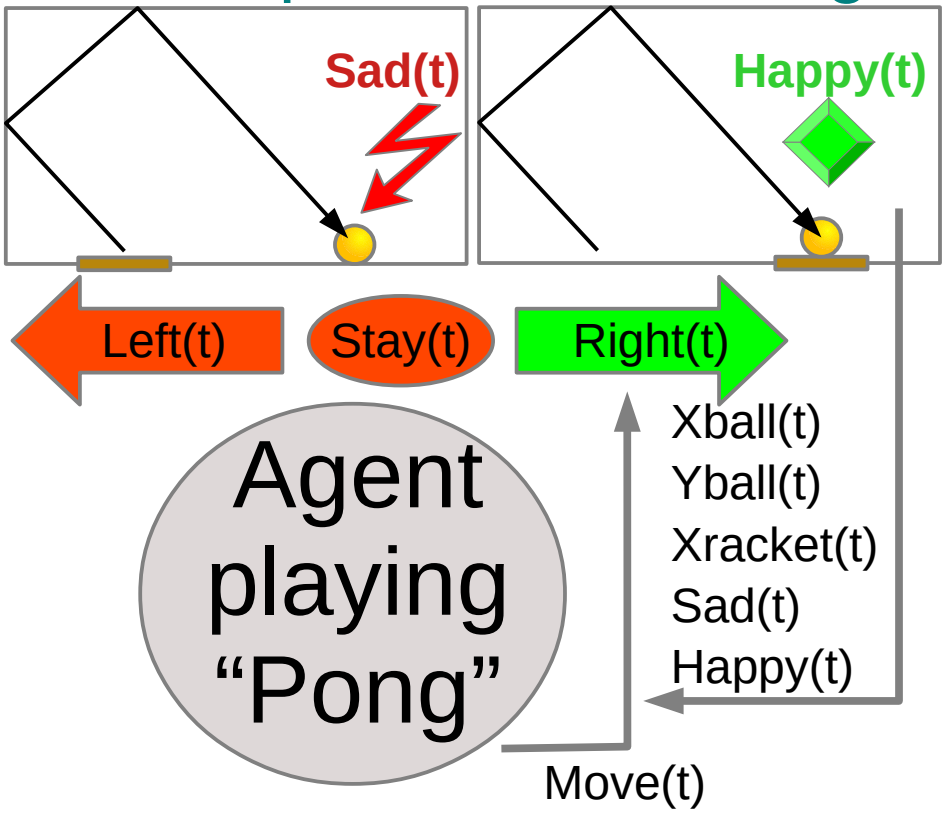
Cost of Dealing With Complexity

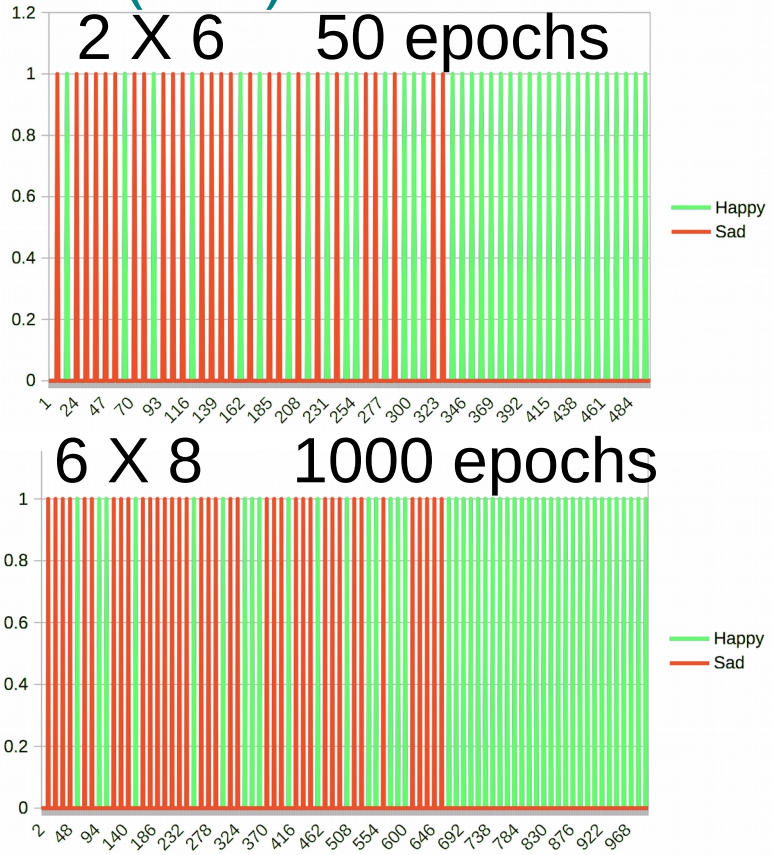


Cost of Dealing With Complexity



Identifying successful/unsuccessful sequential experiences for experiential learning with global (self)reinforcement

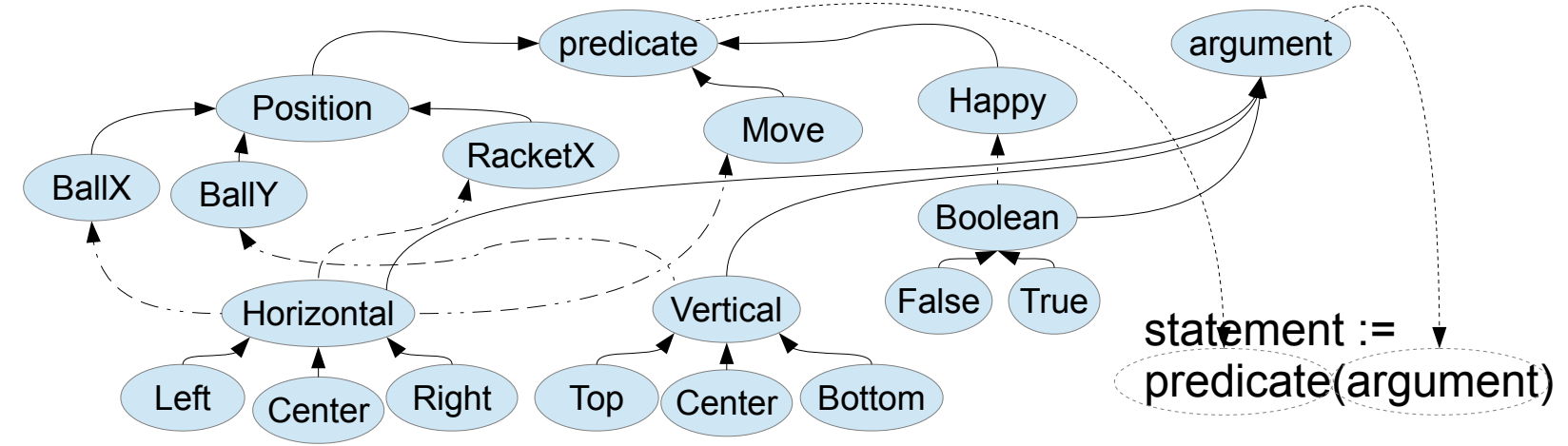


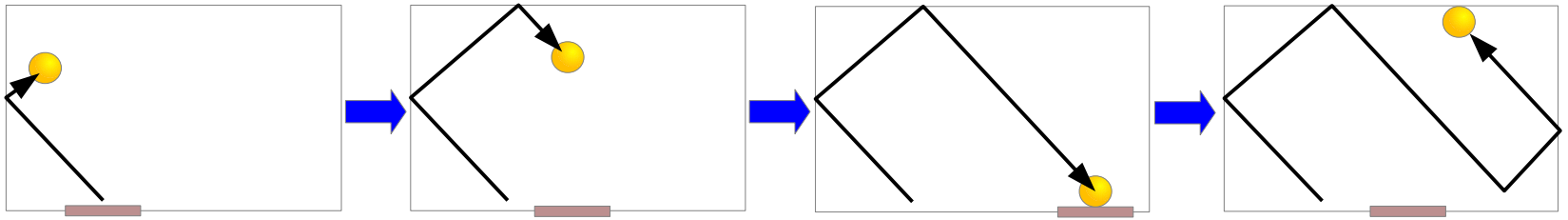


https://www.youtube.com/watch?v=2LPLhJKh95g
https://www.springerprofessional.de/neuro-symbolic-architecture-for-experiential-learning-in-discret/20008336
https://github.com/aigents/aigents-java/tree/master/src/main/java/net/webstructor/agi

AGI-2021

Learning Play “Pong” at Object Level





BallY(Top)
BallX(Left)
RacketX(Left)
Happy(False)
=> **Move(Left)**

BallY(Top)
BallX(Center)
RacketX(Center)
Happy(False)
=> **Move(Right)**

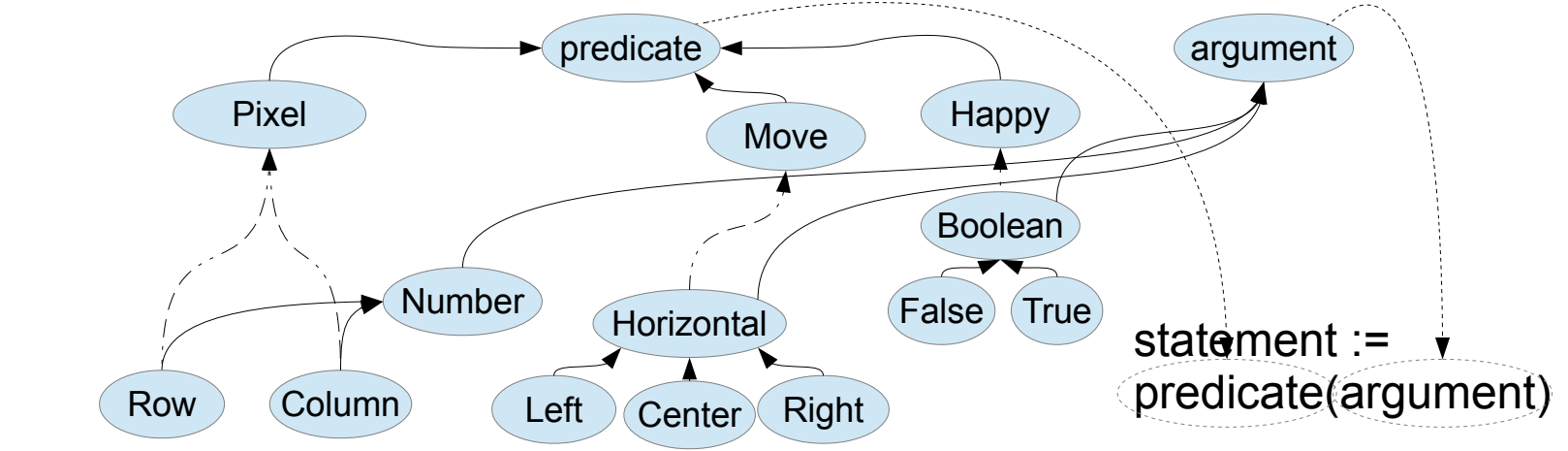
BallY(Bottom)
BallX(Right)
RacketX(Right)
Happy(False)
=> **Move(Right)**

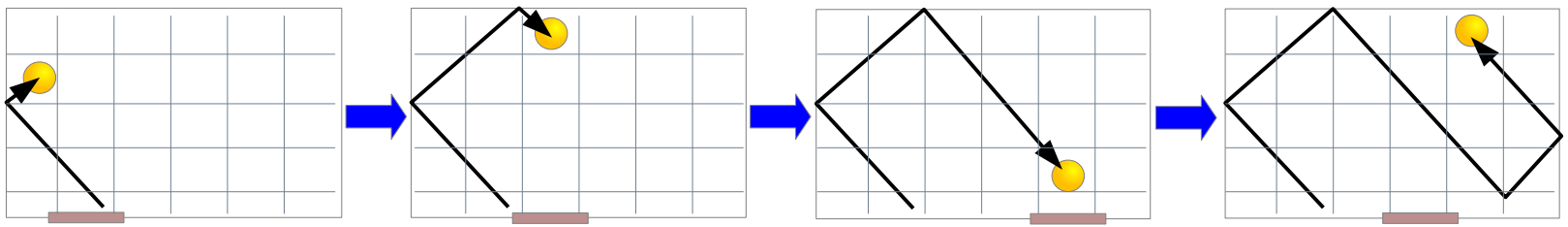
BallY(Bottom)
BallX(Right)
RacketX(Right)
Happy(True)
=> **Move(Left)**

Global Feedback

https://www.youtube.com/watch?v=2LPLhJKh95g

Learning Play “Pong” at Pixel Level





Pixel(1,0)
Pixel(4,1)
Happy(False)
=>
Move(Left)

Pixel(0,2)
Pixel(4,2)
Happy(False)
=>
Move(Right)

Pixel(3,4)
Pixel(4,4)
Happy(False)
=>
Move(Left)

Pixel(0,4)
Pixel(4,3)
Happy(True)
=>
Move(Left)

Global Feedback

https://www.youtube.com/watch?v=2LPLhJKh95g

Learning single-player “Pong” game with global feedback for successive behaviors

		Immediate feedback					Delayed feedback				
Environment	**Player Algorithm**	**2X4**	**4X6**	**6X8**	**8X10**	**Avg**	**2X4**	**4X6**	**6X8**	**8X10**	**Avg**
Functional	Sequential	89	88	88	92	89	70	73	72	85	75
Functional	SequentialAvoidance	92	90	90	93	91	67	73	81	85	77
Functional	SequentialAvoidance 0.5	**93**	**93**	**93**	93	93	80	83	81	89	83
Functional	State-Action	94	88	91	94	92	64	71	79	80	74
Functional	State-Action 0.5	93	88	87	93	90	64	68	75	83	73
Functional	Change-Action	91	86	89	92	90	64	73	76	79	73
Functional	Change-Action 0.5	93	90	90	93	92	63	69	80	84	74
Discrete	Sequential	89	88	88	92	89	70	73	72	85	75
Discrete	SequentialAvoidance	92	90	90	93	91	67	73	81	85	77
Discrete	SequentialAvoidance 0.5	93	91	88	92	91	70	76	80	83	77
Discrete	State-Action	94	88	91	94	92	64	71	79	80	74
Discrete	Change-Action	91	86	89	92	90	64	73	76	79	73

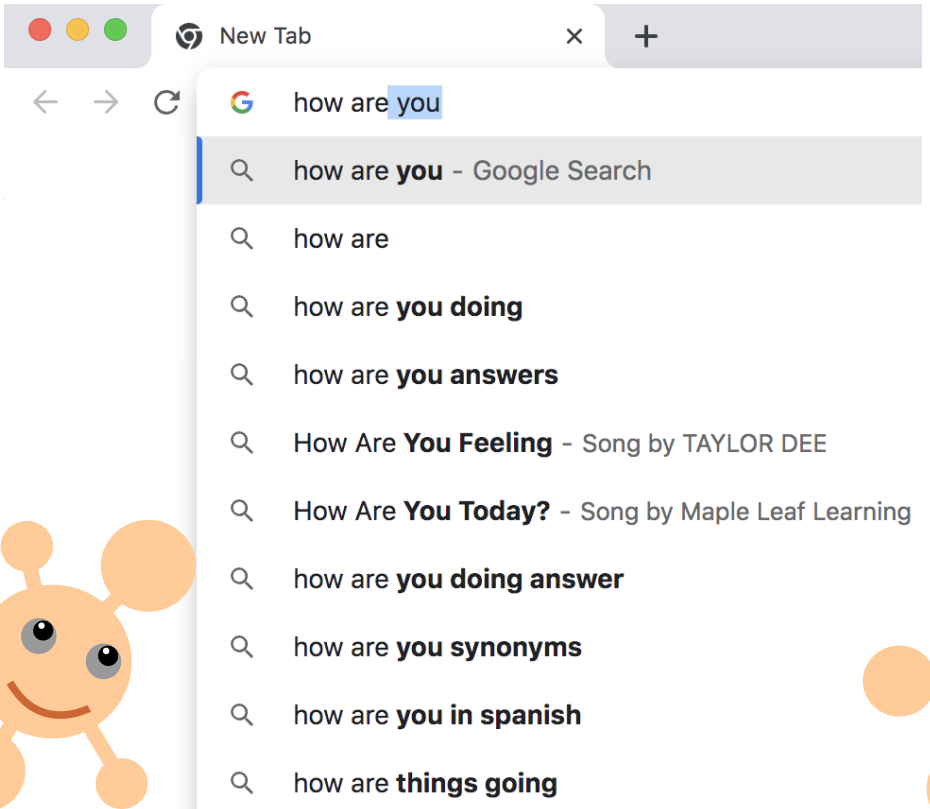
Objects:
Fast
Cheap
Real-time

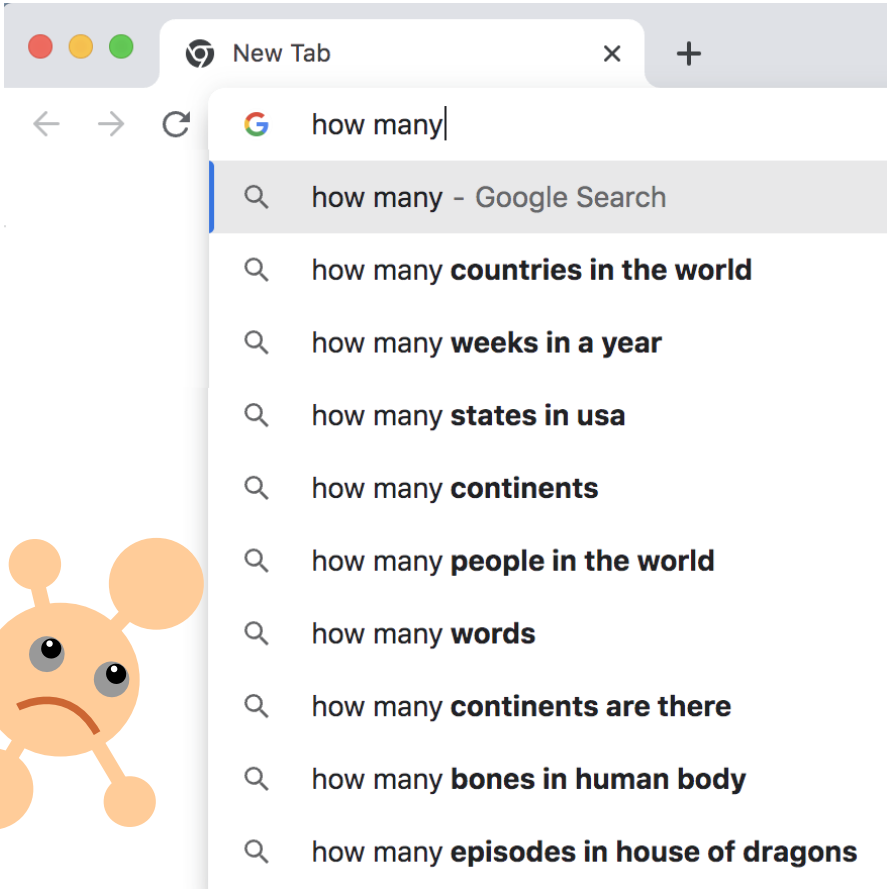
Pixels:
~~Fast~~
~~Cheap~~
~~Real-time~~

https://www.youtube.com/watch?v=2LPLhJKh95g
https://www.springerprofessional.de/neuro-symbolic-architecture-for-experiential-learning-in-discret/20008336
https://github.com/aigents/aigents-java/tree/master/src/main/java/net/webstructor/agi

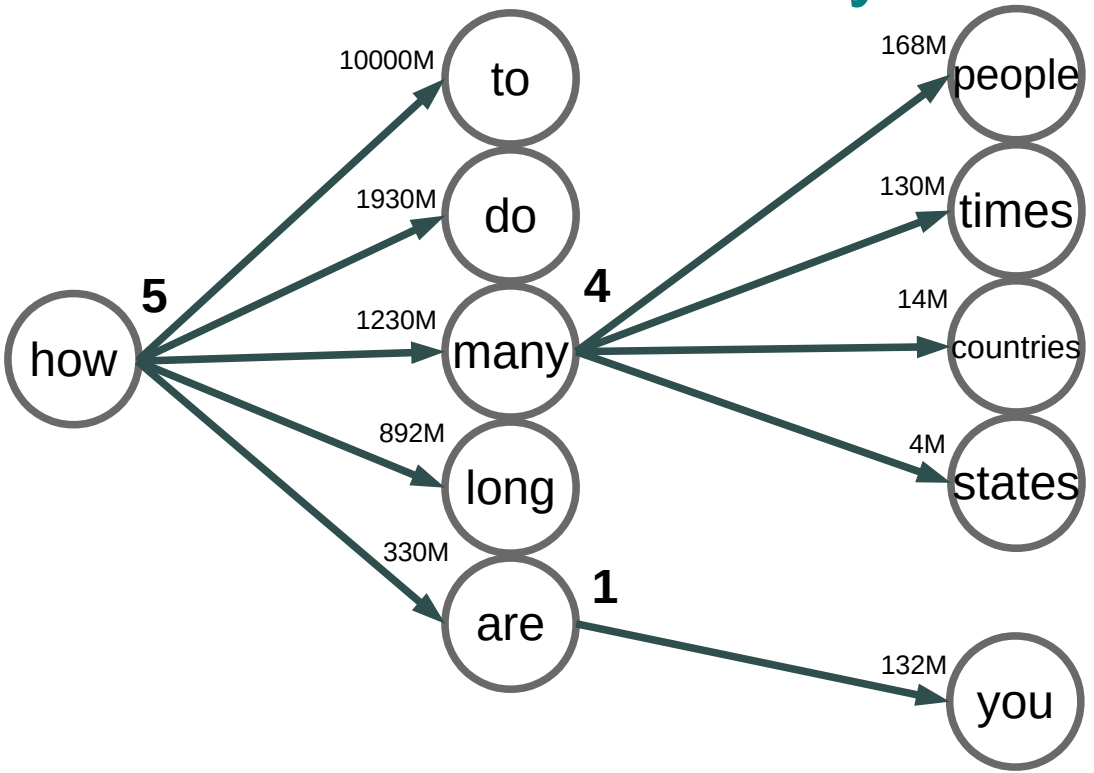
AGI-2021

Minimizing Uncertainty





Unsupervised Learning for Text Segmentation based on Probability and Uncertainty Measures



Metrics/Indicators:

Mutual Information[1]
Conditional Probability[1,2]
Transition Freedom[2,3]

[1] https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=6983&context=etd

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655800/

[3] **Karl Friston. The free-energy principle: a unified brain theory?** https://www.nature.com/articles/nrn2787

Unsupervised Text Segmentation (Tokenization)

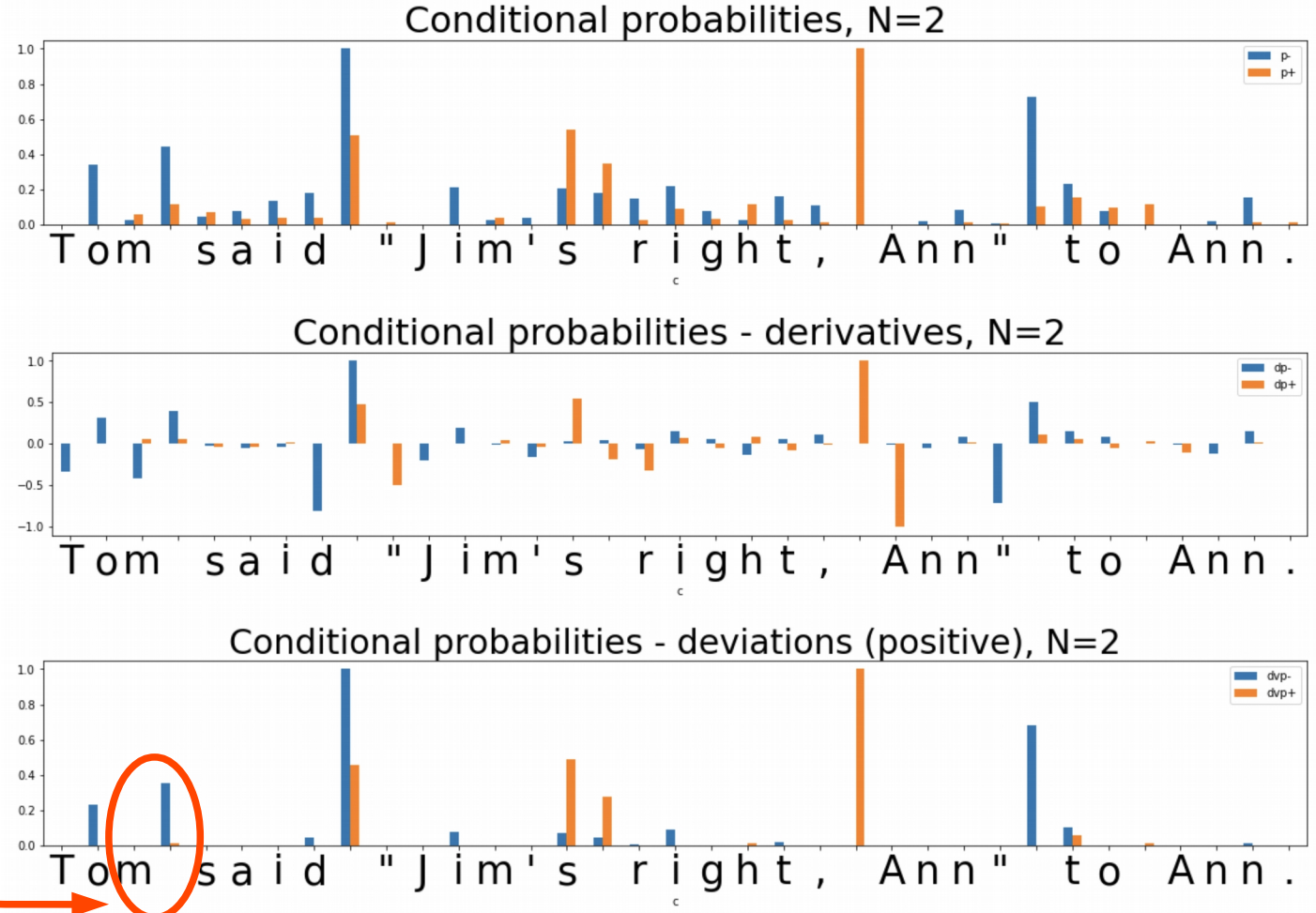
Metrics/Indicators:

Ngram (Character) Conditional Probability

(of Transition)

$P(Ngram_{n+1})/P(Ngram_n)$

P(“m_”)/P(“m”)

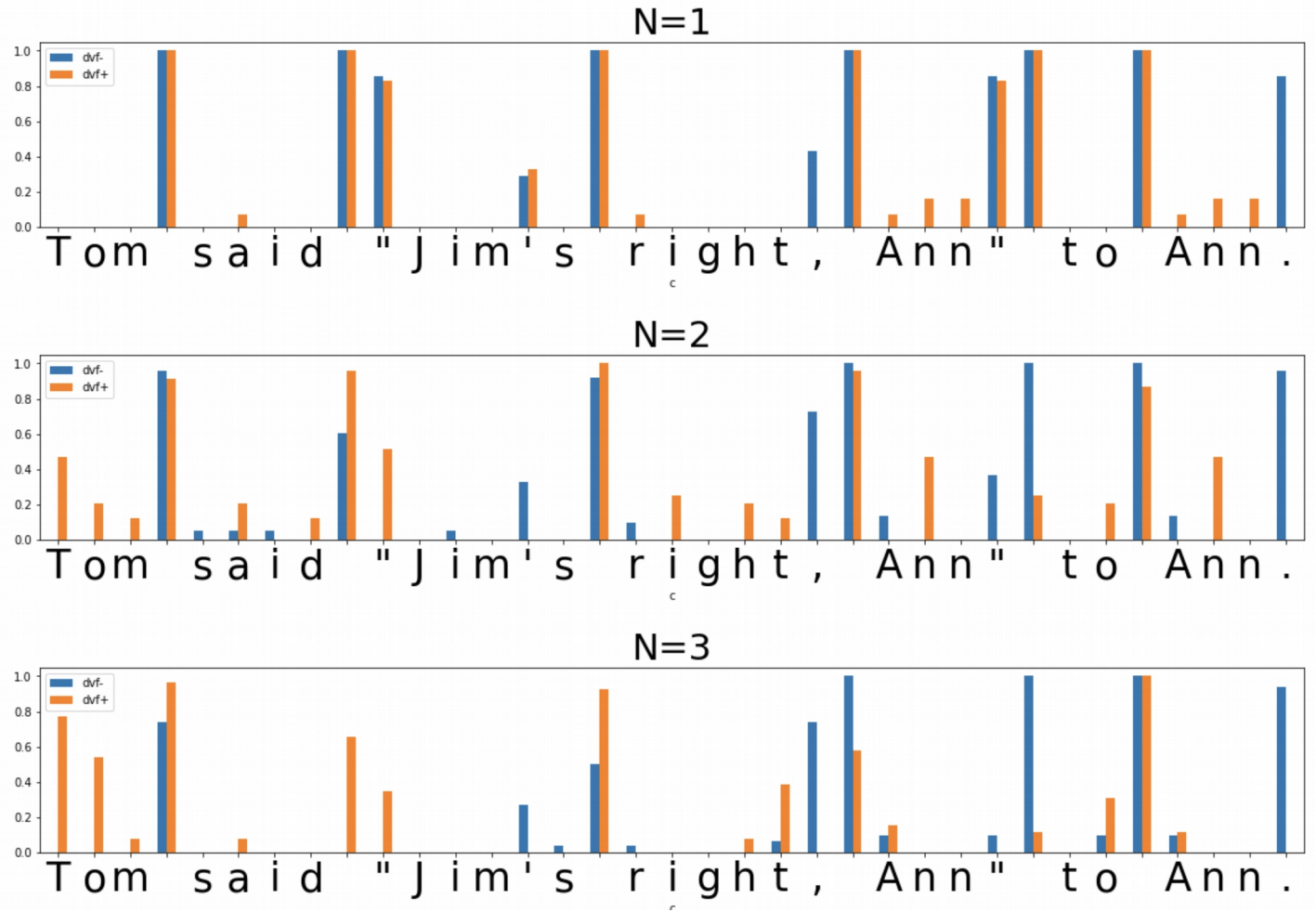


Unsupervised Text Segmentation (Tokenization)

Metrics/ Indicators:

Transition
Freedom
Deviation

(varying “N”)



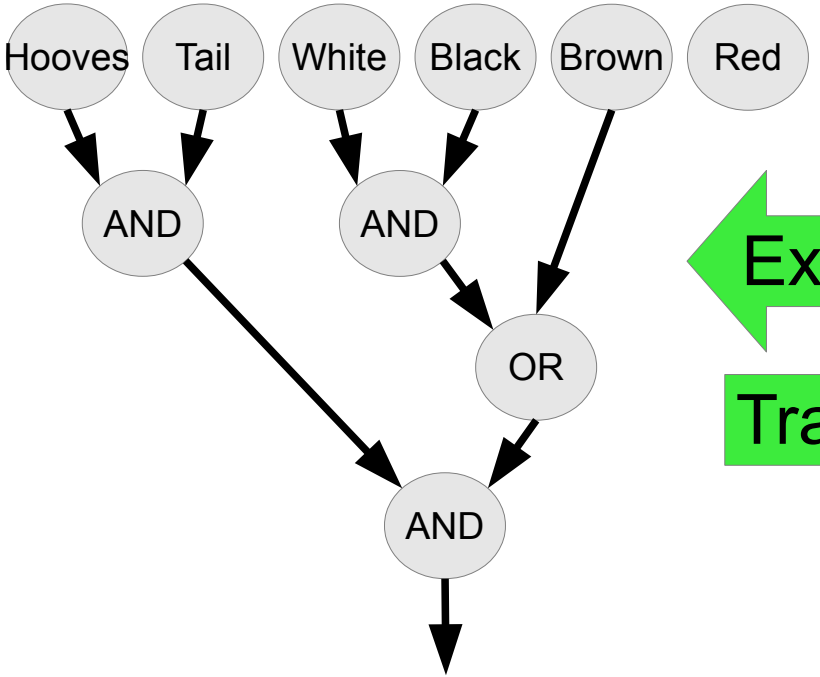
Results – Freedom-based Tokenization against Lexicon

Language	Tokenizer	Tokenization F1	Lexicon Discovery Precision
English	Freedom-based	**0.99**	**0.99** (vs 1.0)
English	Lexicon-based *	0.99	-
Russian	Freedom-based	**1.0**	**1.0** (vs 1.0)
Russian	Lexicon-based *	0.94	-
Chinese	Freedom-based	**0.71**	**0.92** (vs 0.94)
Chinese	Lexicon-based *	0.83	-

** Lexicon-based Tokenization - greedy/beam search on word length (optimal) or frequency*

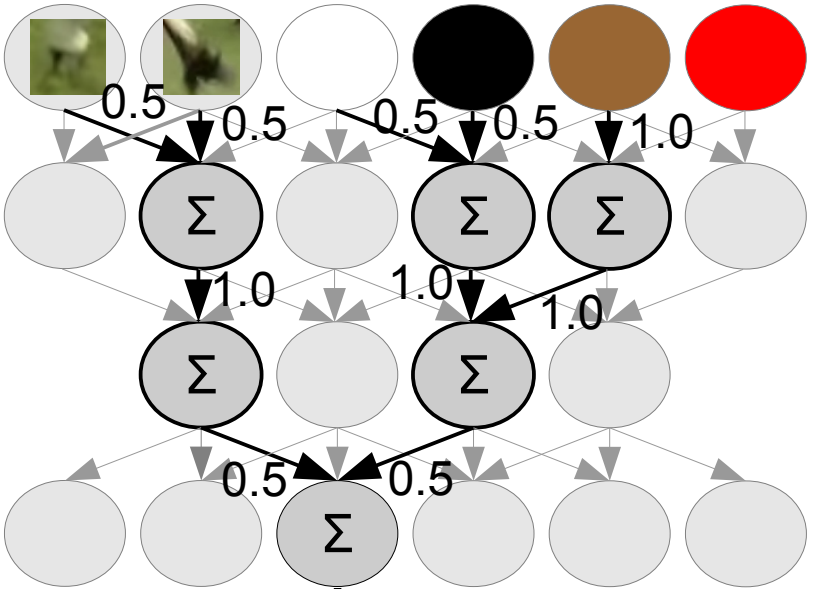
https://arxiv.org/abs/2205.11443 **EMNLP-2022**

Neuro-Symbolic Integration for Interpretable AI



Explain

Transfer



(Hooves AND Tail) AND
((White and Black) OR Brown)

=> Horse

https://www.springerprofessional.de/neuro-symbolic-architecture-for-experiential-learning-in-discret/20008336

Thank You and Welcome!

Anton Kolonin

akolonin@aigents.com

Facebook: akolonin

Telegram: akolonin

AGI IN RUSSIAN

https://agirussia.org

Novosibirsk State University

*THE REAL SCIENCE

AIGENTS

https://aigents.com