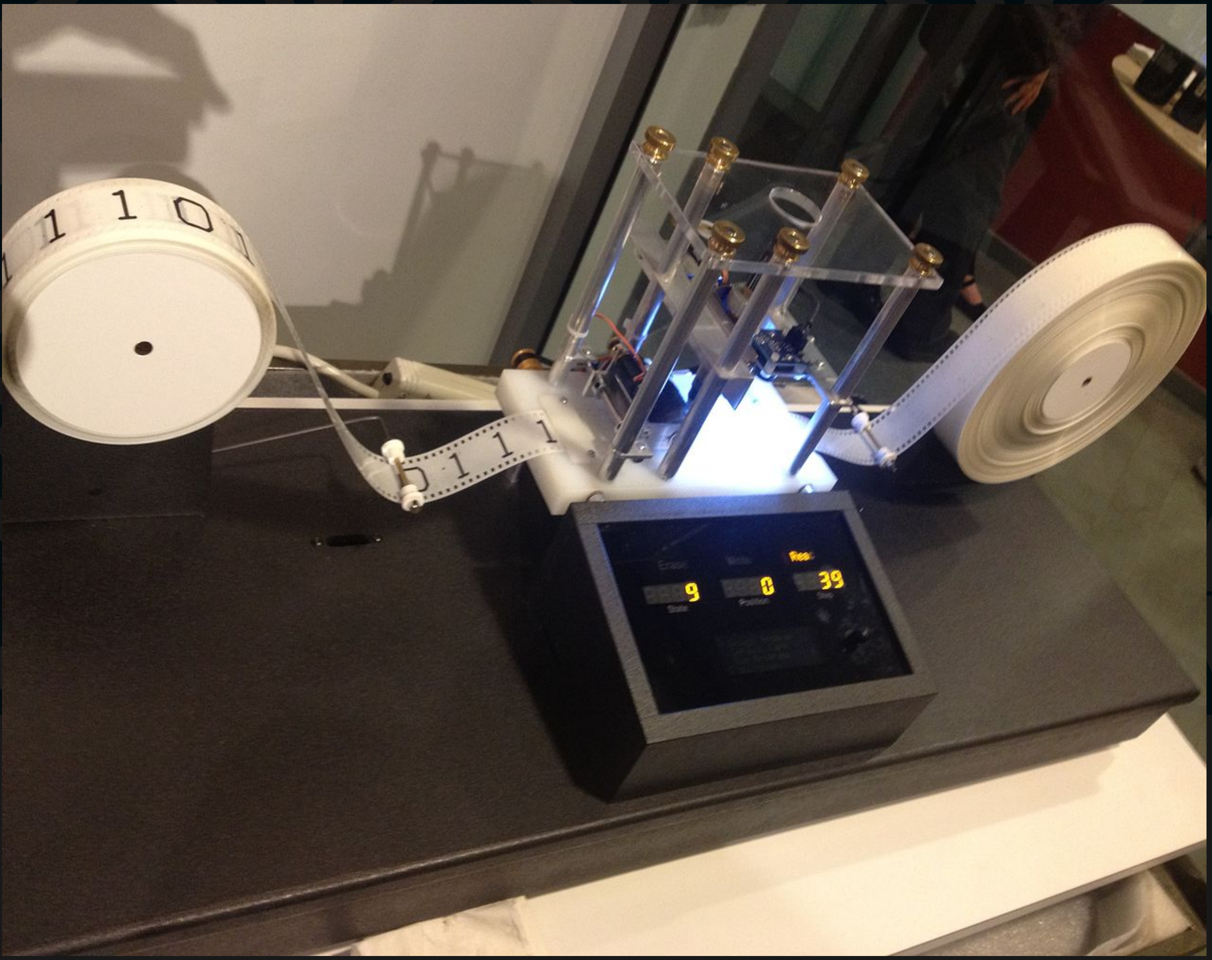


Memory Advances in Neural Turing Machines

Truyen Tran

Deakin University

truyen.tran@deakin.edu.au

truyentran.github.io

@truyenoz

letdataspeak.blogspot.com

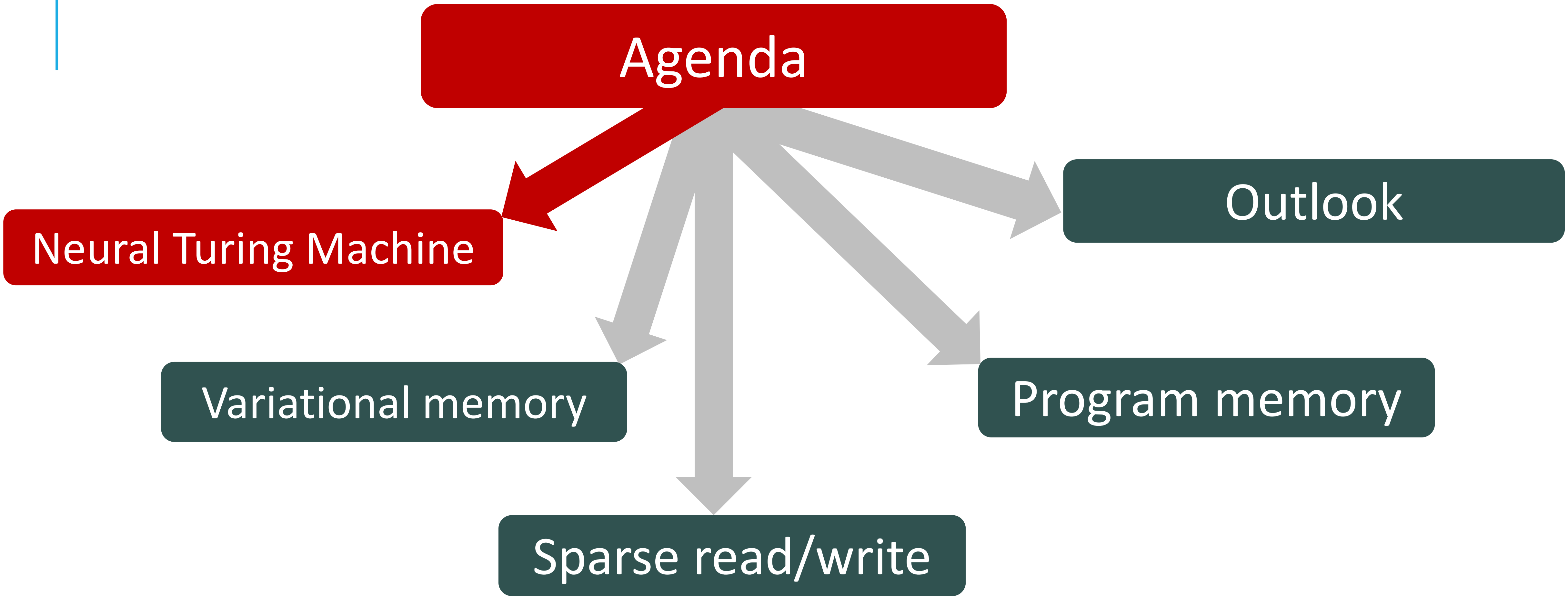
goo.gl/3jJ1O0

Hanoi, June 2019


Domain expert
Deep Learning
Knowledge-based

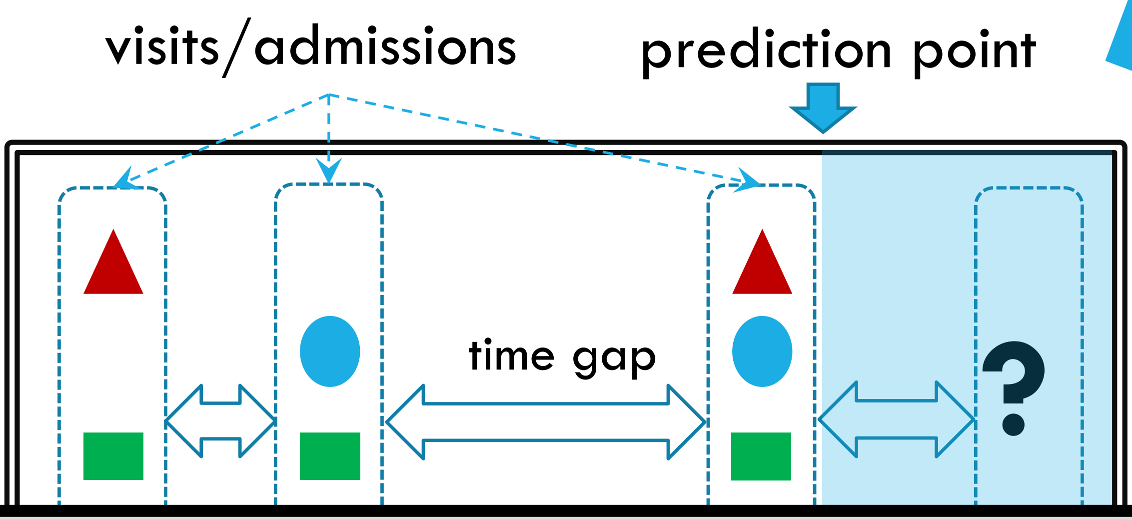
Can we learn from data a model that is as powerful as a Turing machine?

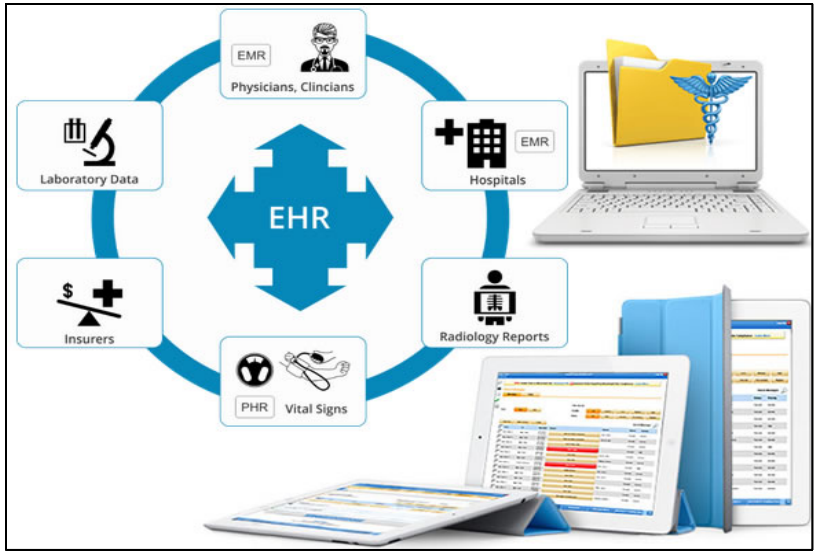
In other words, can we learn a (neural) program that learns to program from data?


Agenda
Neural Turing Machine
Outlook
Variational memory
Program memory
Sparse read/write

Example: Electronic medical records

Abstraction





Source: medicalbillingcodings.org

Modelling

Three interwoven processes:

- Disease progression
- Interventions & care processes
- Recording rules

Need memory to handle thousands of events, compute complex healthcare “grammars”, support chain of reasoning, rapid switching of tasks.

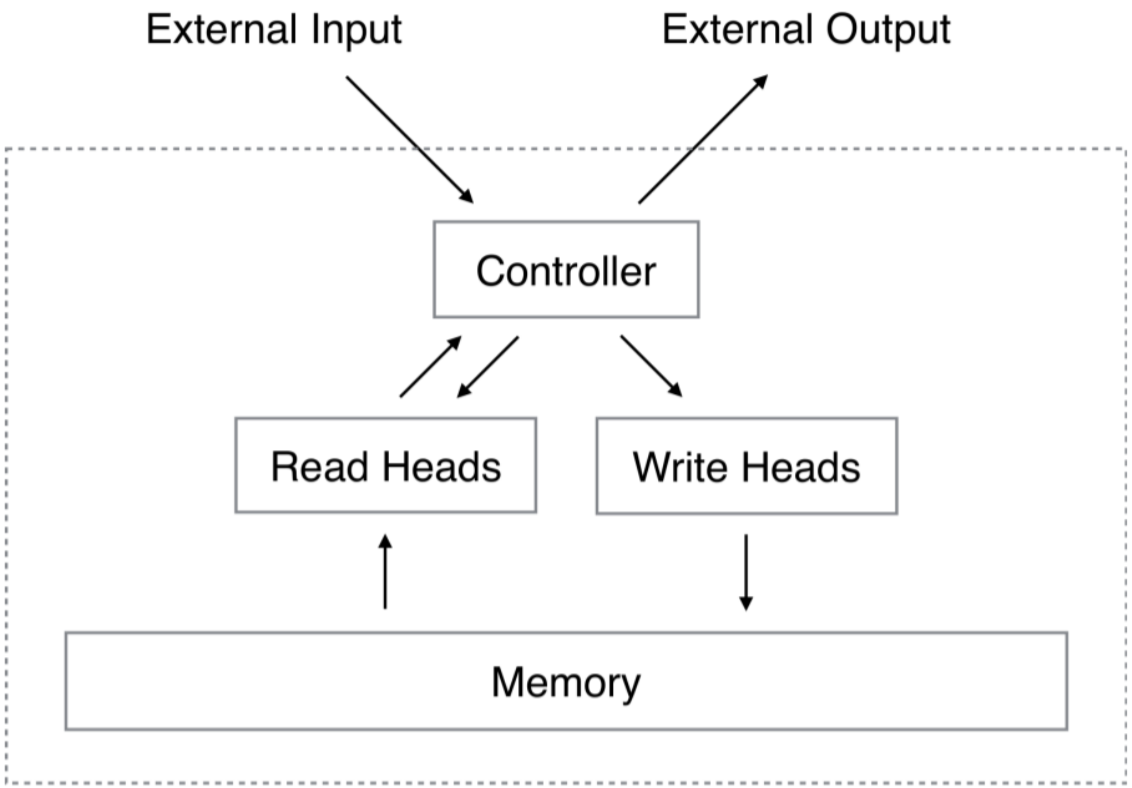
Neural Turing machine (NTM)

A controller that takes input/output and talks to an external memory module.

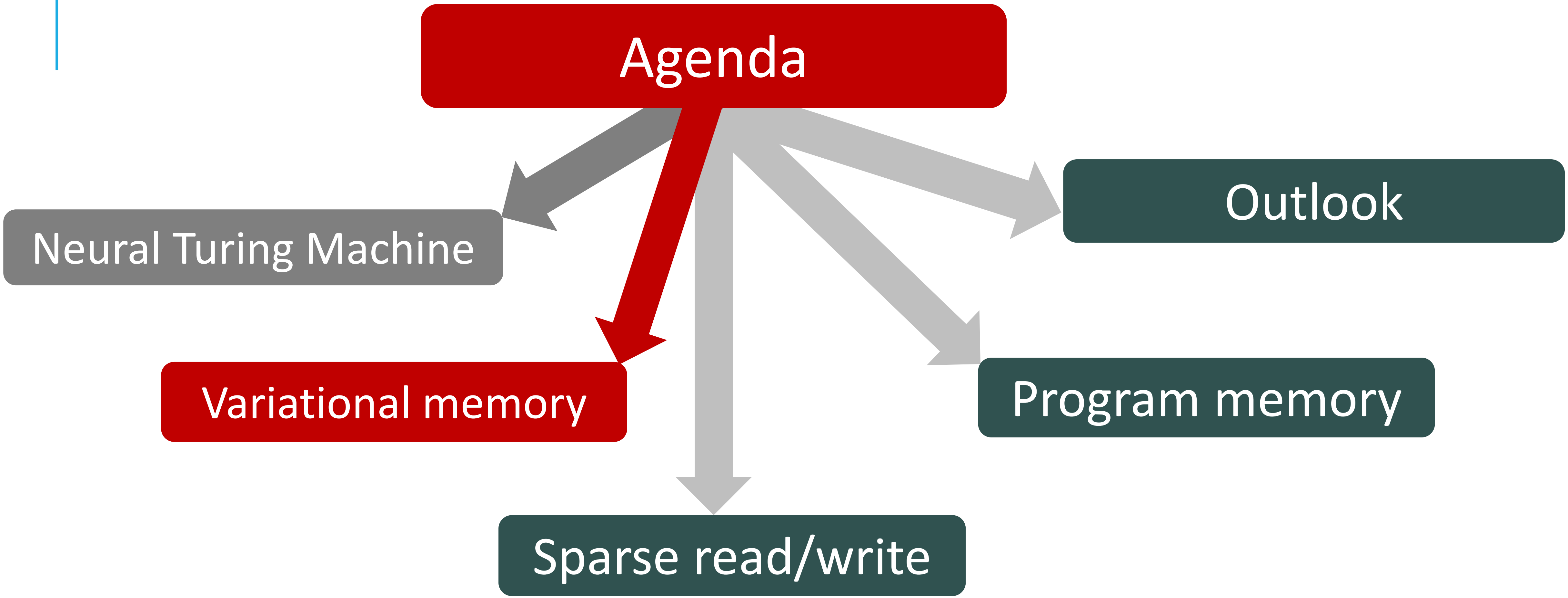
Memory has read/write operations.

The main issue is where to write, and how to update the memory state.

All operations are differentiable.



https://rylanschaeffer.github.io/content/research/neural_turing_machine/main.html


Agenda
Neural Turing Machine
Variational memory
Sparse read/write
Program memory
Outlook

Motivation: Dialog system

A dialog system needs to maintain the history of chat (e.g., could be hours)

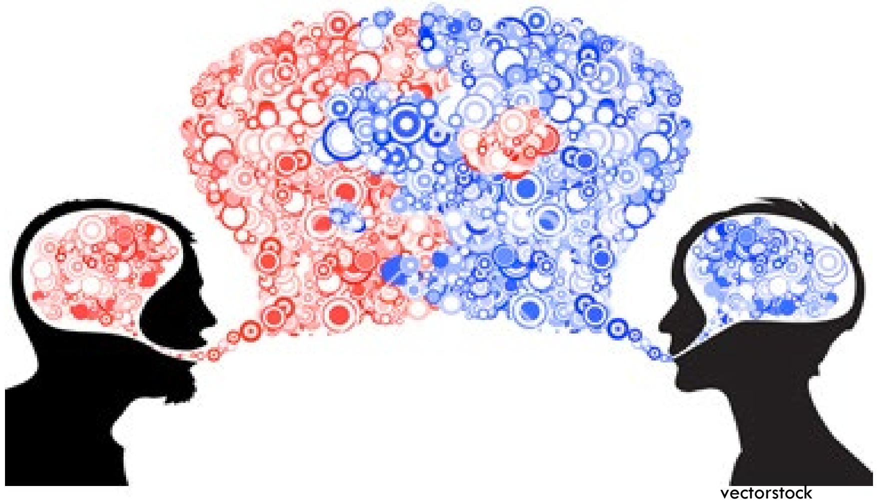
- → Memory is needed

The generation of response needs to be flexible, adapting to variation of moods, styles

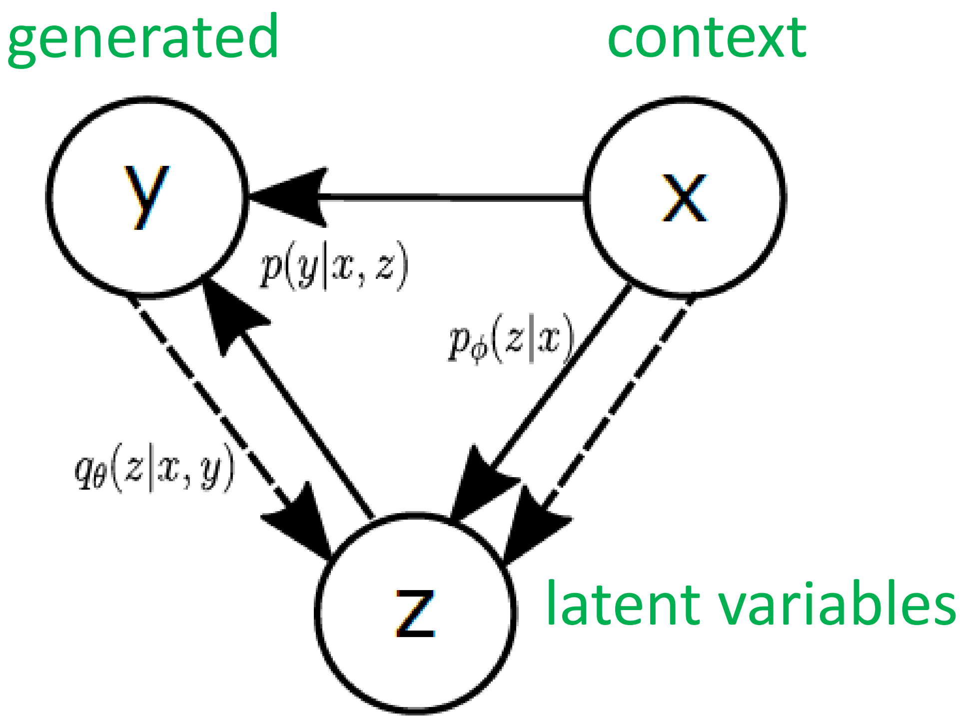
- Current techniques are mostly based on LSTM, leading to "stiff" default responses (e.g., "I see").

There are many ways to express the same thought

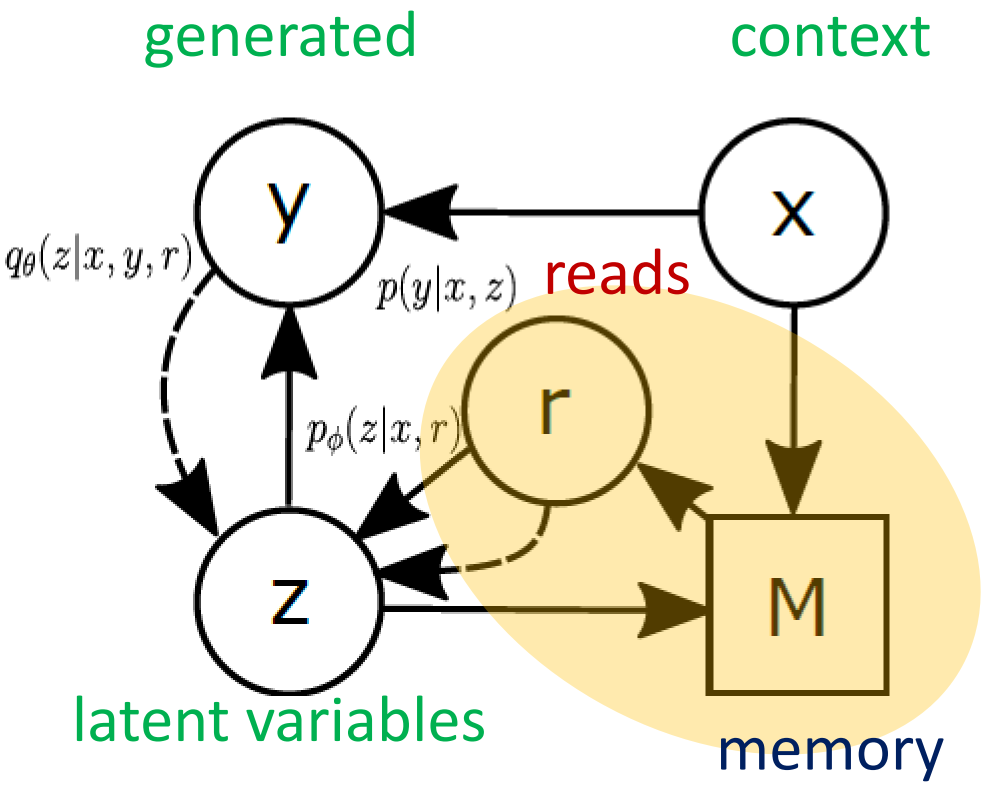
- → Variational generative methods are needed.



Variational memory encoder-decoder (VMED)



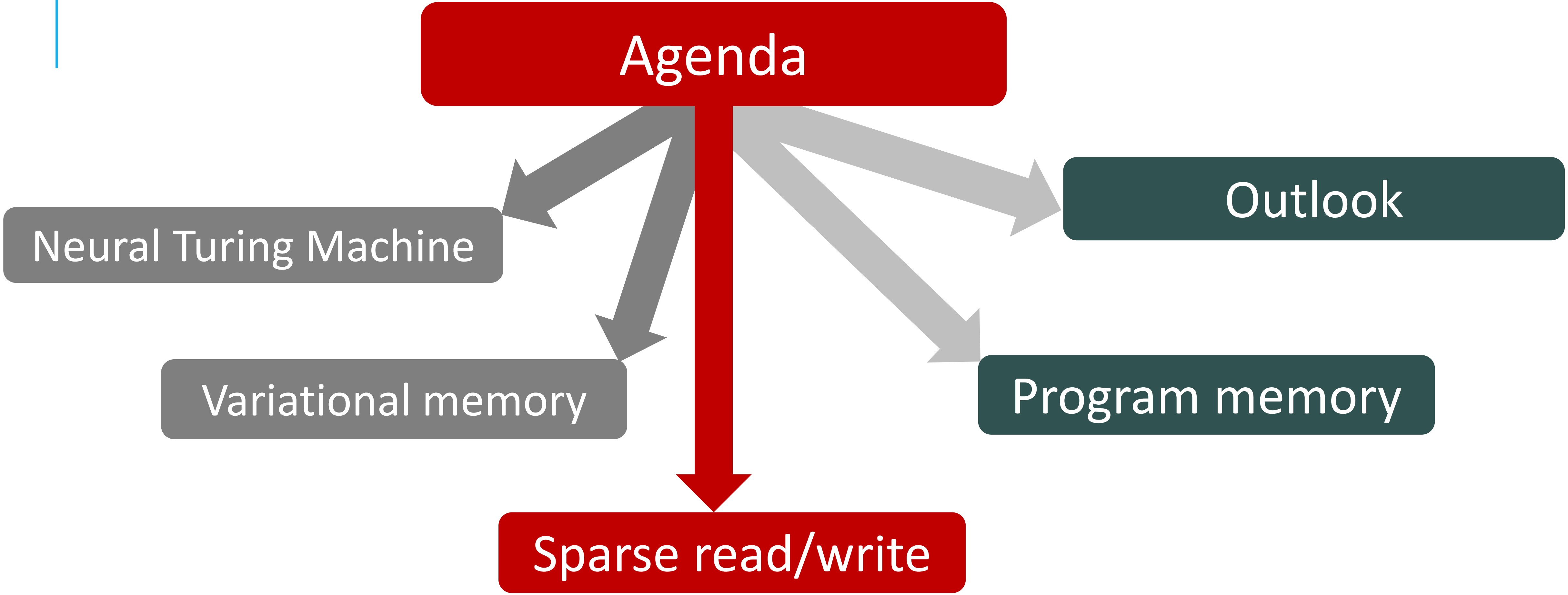
Conditional Variational Auto-Encoder



VMED

Sample response

Input context	Response
Reddit comment: What is your favorite scene in film history ? Mine is the restaurant scene in the Godfather.	**Seq2Seq:** The scene in **Seq2Seq-att:** The final **DNC:** The scene in **CVAE:** Inception god! Not by a shark /*/ Amour great /*/ Pro thing you know 3 dead **VMED (K=3):** The opening scene from history movie /*/ The scene in a shot nights! Robin movie /*/ The psycho scene in fight from
Reddit comment: What actor will win an Oscar in the next 10 years ?	**Seq2Seq:** Colin **Seq2Seq-att:** Liam Neeson **DNC:** Tom Gyllenhaal **CVAE:** Daryl and Aaron /*/ Carefully count Alfred Deniro /*/ Ponyo Joker posible **VMED (K=3):** Edward or Leo Dicaprio goes on /*/ Dicaprio will /*/ Dicaprio Tom has actually in jack on road


Agenda
Neural Turing Machine
Variational memory
Sparse read/write
Outlook
Program memory

Problems of current NTMs

Lack of theoretical analysis on optimal memory operations.

Previous works are based on intuitions:

- Location-based reading/writing; temporal linkage reading; least-used writing [Santoro et.al, Graves et.al]
- Sparse access over big memory [Rae et.al]

Very slow due to heavy memory read/write computations

Cached Uniform Writing (CUW)

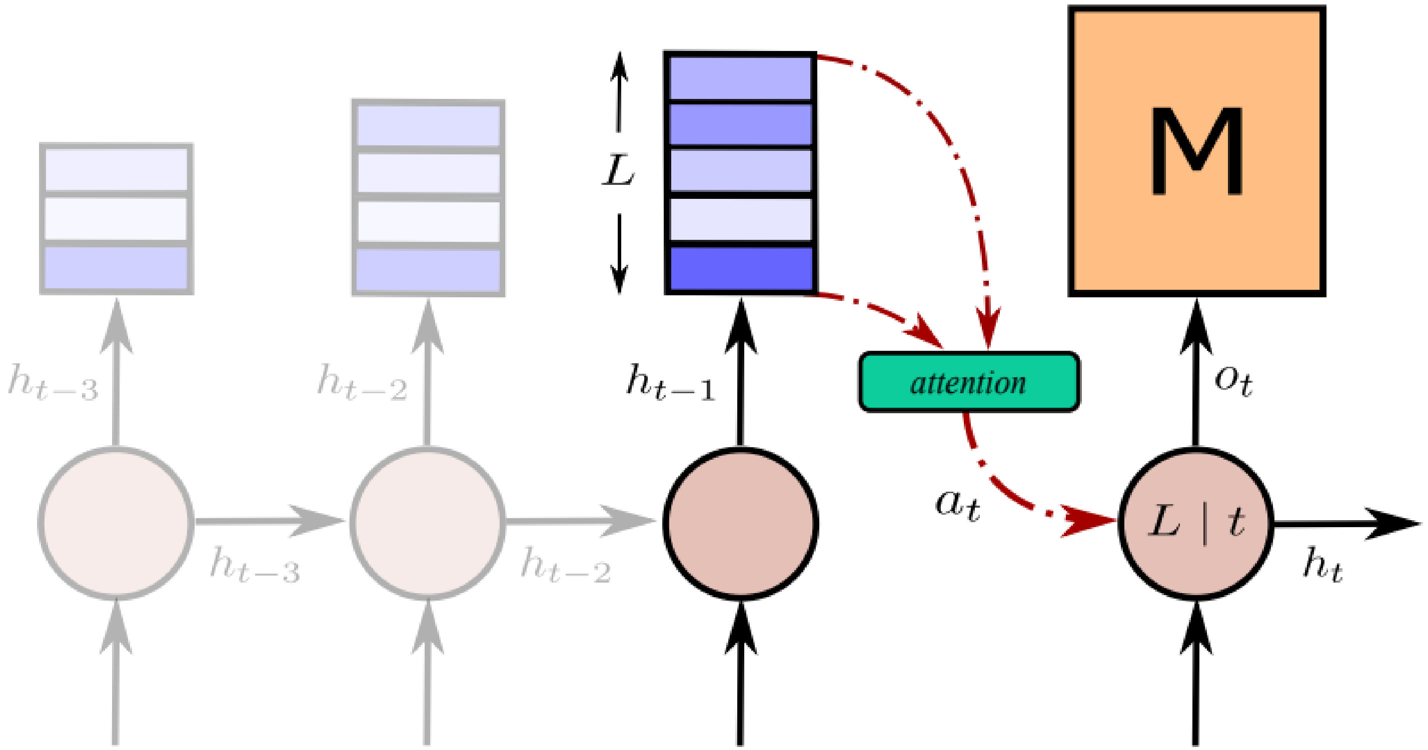


Figure 1: Writing mechanism in Cached Uniform Writing. During non-writing intervals, the controller hidden states are pushed into the cache. When the writing time comes, the controller attends to the cache, chooses suitable states and accesses the memory. The cache is then emptied.

Ablation Study

Memory-augmented Neural Networks w/wo Uniform Writing

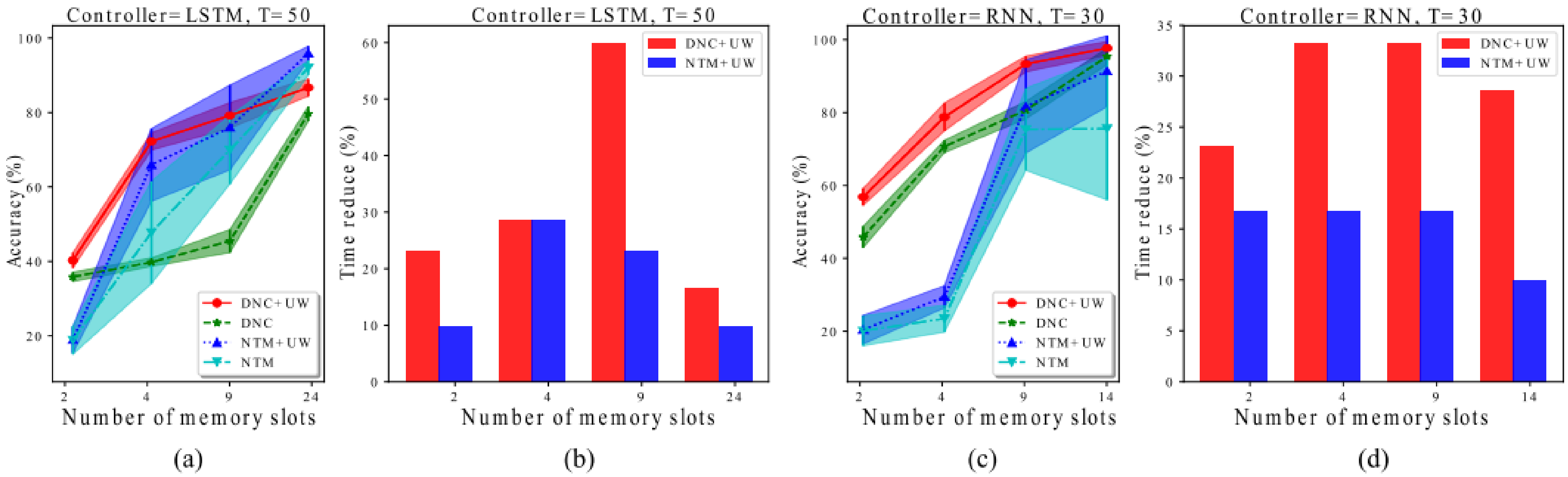


Figure 2: The accuracy (%) and computation time reduction (%) with different memory types and number of memory slots. The controllers/sequence lengths/memory sizes are chosen as LSTM/50/$\{2, 4, 9, 24\}$ (a&b) and RNN/30/$\{2, 4, 9, 14\}$ (c&d), respectively.

Task: repeat the input sequence twice

Synthetic tasks: memorize all

Model	N_h	# parameter	Copy		Reverse	
			L=50	L=100	L=50	L=100
LSTM	125	103,840	15.6	12.7	49.6	26,1
NTM	100	99,112	40.1	11.8	61.1	20.3
DNC	100	98,840	68.0	44.2	65.0	54.1
DNC+RW	100	98,840	47.6	37.0	70.8	50.1
DNC+UW	100	98,840	**97.7**	**69.3**	**100**	**79.5**
DNC+CUW	95	96,120	83.8	55.7	93.3	55.4

Table 1: Test accuracy (%) on synthetic memorization tasks. MANNs have 4 memory slots.

Synthetic tasks: memorize selectively

Model	Add		Max	
	L=50	L=100	L=50	L=100
DNC	83.8	22.3	59.5	27.4
DNC+RW	83.0	22.7	59.7	36.5
DNC+UW	84.8	50.9	71.7	66.2
DNC+CUW	**94.4**	**60.1**	**82.3**	**70.7**

Table 2: Test accuracy (%) on synthetic reasoning tasks. MANNs have 4 memory slots.

Synthetic sinusoidal generation: memorize featured points

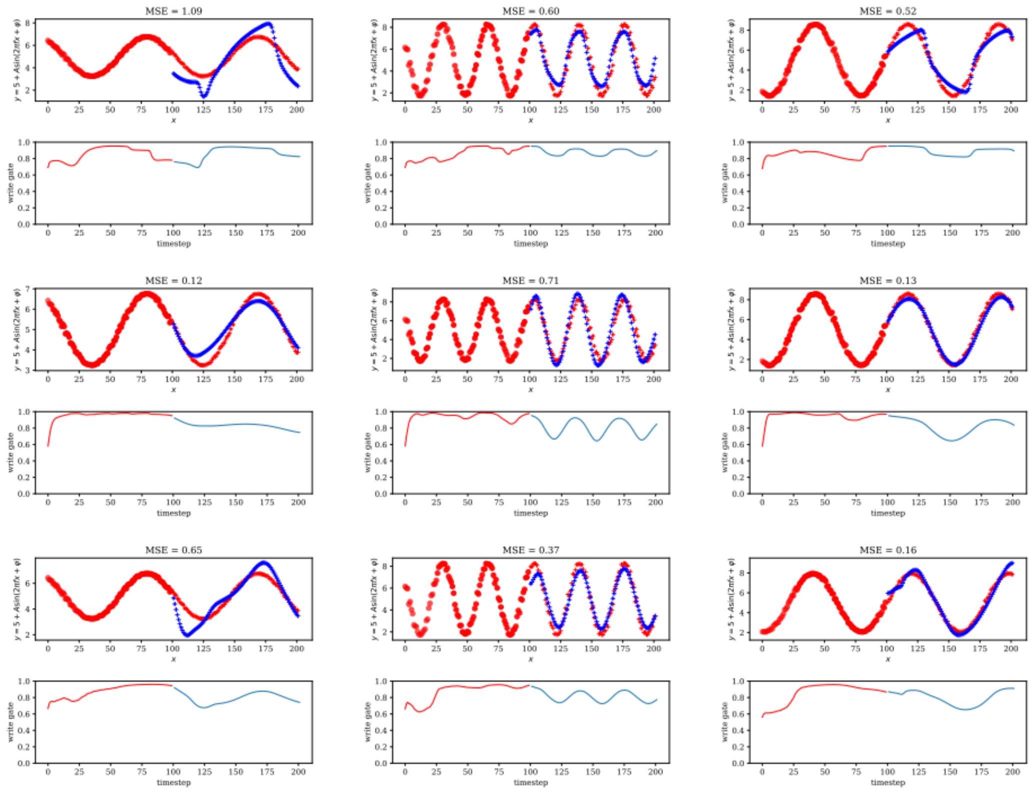


Figure 6: Sinusoidal generation with clean input sequence for DNC, UW and CUW in top-down order.

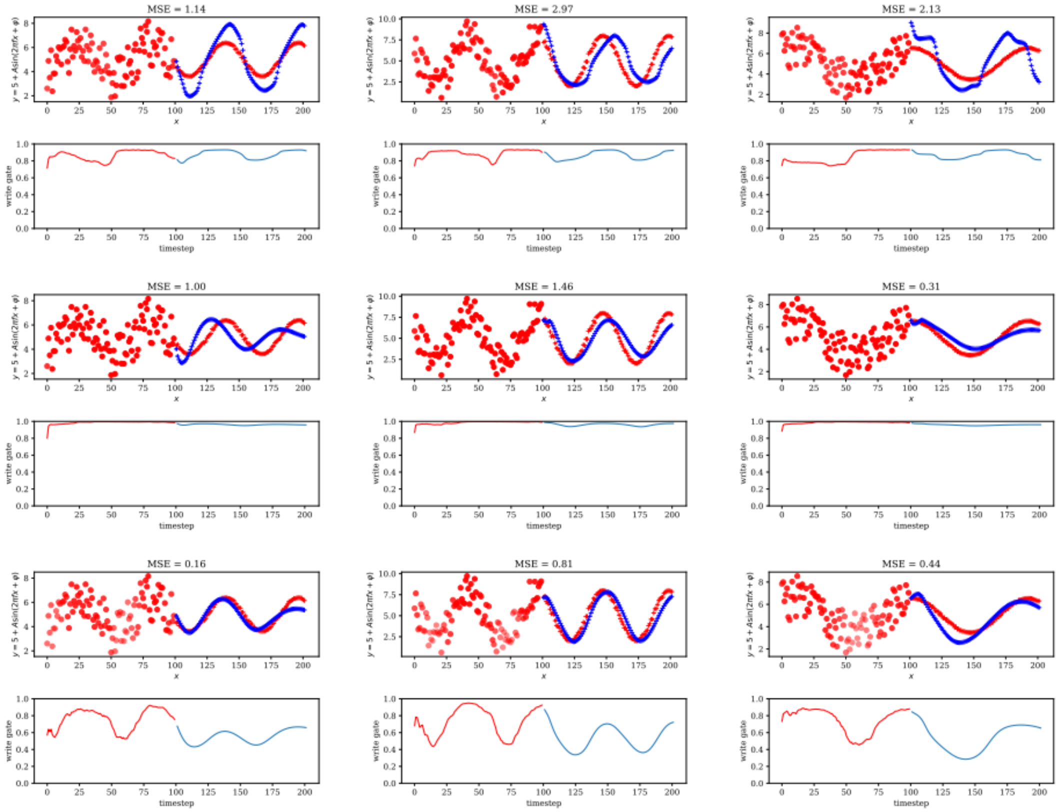


Figure 7: Sinusoidal generation with noisy input sequence for DNC, UW and CUW in top-down order.

Flatten MNIST classification

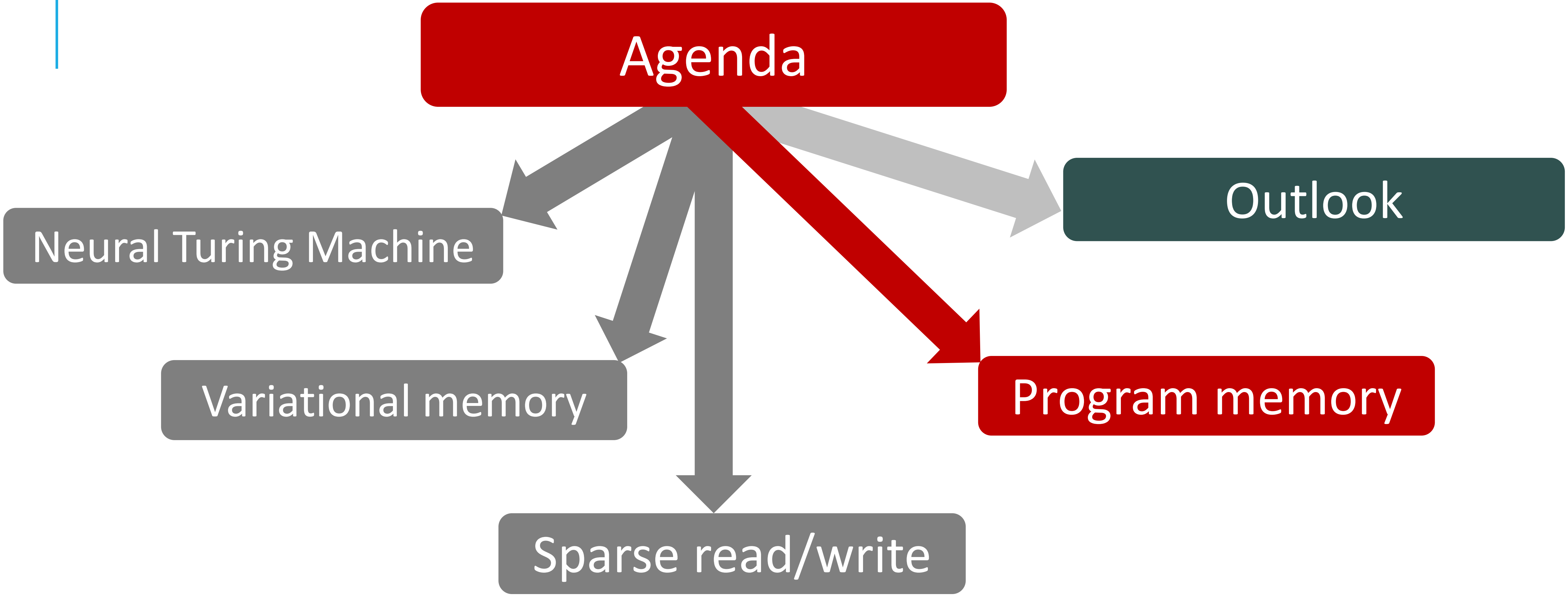
Model	MNIST	pMNIST
iRNN†	97.0	82.0
uRNN$^{\circ}$	95.1	91.4
r-LSTM Full BP*	98.4	95.2
Dilated-RNN$^{\blacklozenge}$	95.5	96.1
Dilated-GRU$^{\blacklozenge}$	**99.2**	94.6
DNC	98.1	94.0
DNC+UW	98.6	95.6
DNC+CUW	99.1	**96.3**

Table 3: Test accuracy (%) on MNIST, pMNIST. Previously reported results are from (Le et al., 2015)†, (Arjovsky et al., 2016)$^{\circ}$, (Trinh et al., 2018)*, and (Chang et al., 2017)$^{\blacklozenge}$.

Document classification

Model	AG	IMDb[4]	Yelp P.	Yelp F.	DBP	Yah. A.
VDCNN•	91.3	-	95.7	64.7	98.7	73.4
D-LSTM*	-	-	92.6	59.6	98.7	*73.7*
Standard LSTM‡	93.5	91.1	-	-	-	-
Skim-LSTM‡	*93.6*	91.2	-	-	-	-
Region Embedding▲	92.8	-	***96.4***	*64.9*	*98.9*	*73.7*
DNC+UW	93.7	**91.4**	**96.4**	65.3	**99.0**	74.2
DNC+CUW	**93.9**	91.3	**96.4**	**65.6**	**99.0**	**74.3**

Table 4: Document classification accuracy (%) on several datasets. Previously reported results are from (Conneau et al., 2016)•, (Yogatama et al., 2017)*, (Seo et al., 2018)‡ and (Qui et al., 2018)▲. We use italics to denote the best published and bold the best records.


Agenda
Neural Turing Machine
Variational memory
Sparse read/write
Program memory
Outlook

Computing devices vs neural counterparts

FSM (1943) ↔ RNNs (1982)

PDA (1954) ↔ Stack RNN (1993)

TM (1936) ↔ NTM (2014)

UTM/VNA (1936/1945) ↔ NUTM--ours (2019)

The missing piece: A memory to store programs

→ Neural stored-program memory

NUTM = NTM + NSM

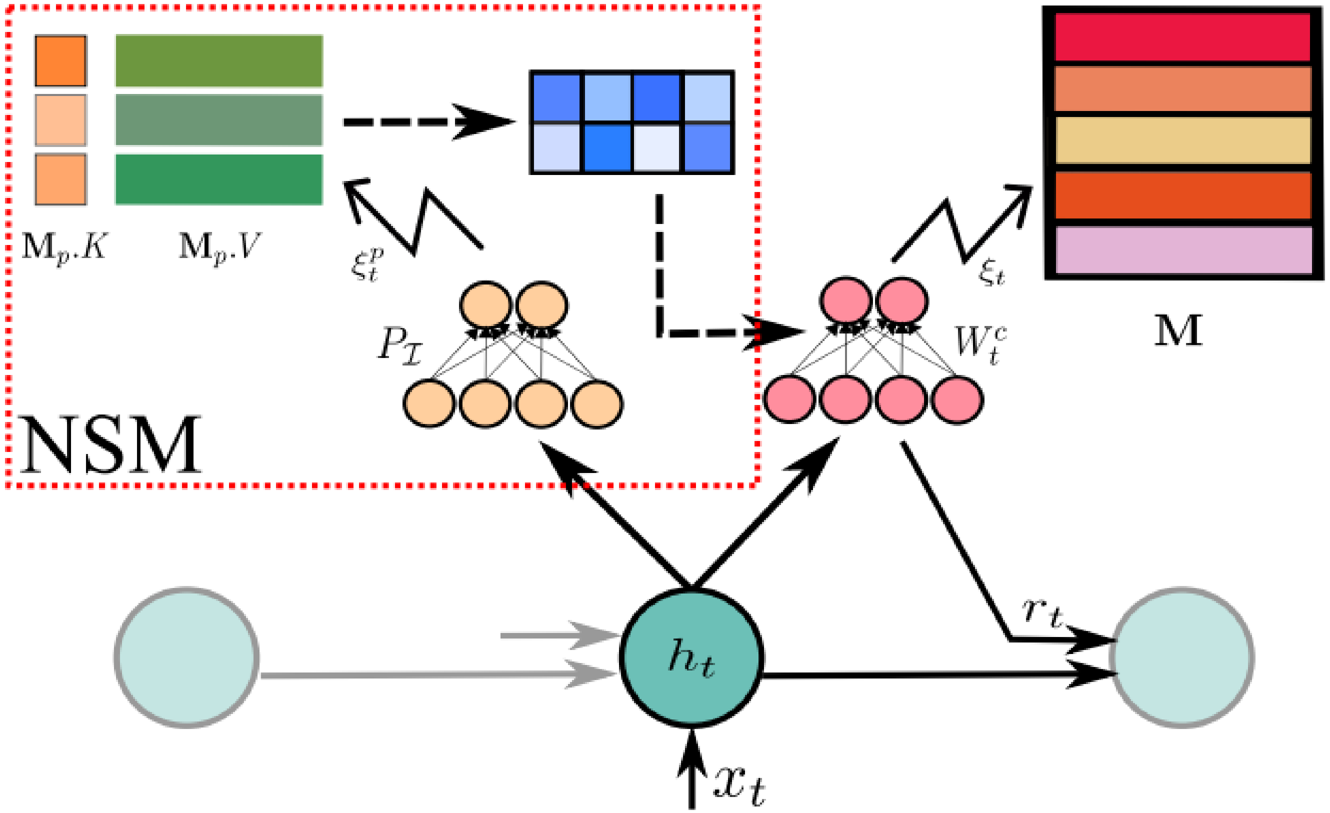


Figure 1: Introducing NSM into MANN. At each timestep, the program interface ($P_{\mathcal{I}}$) receives input from the state network and queries the program memory $\mathbf{M}_p$, acquiring the working weight for the interface network (W_t^c). The interface network then operates on the data memory $\mathbf{M}$ as normal.

Multi-level modelling

Hierarchical Regression: if the input is clustered, clustering before regression helps

Prove for low dimensions maybe available, higher dimension?

NSM is beneficial to NTM

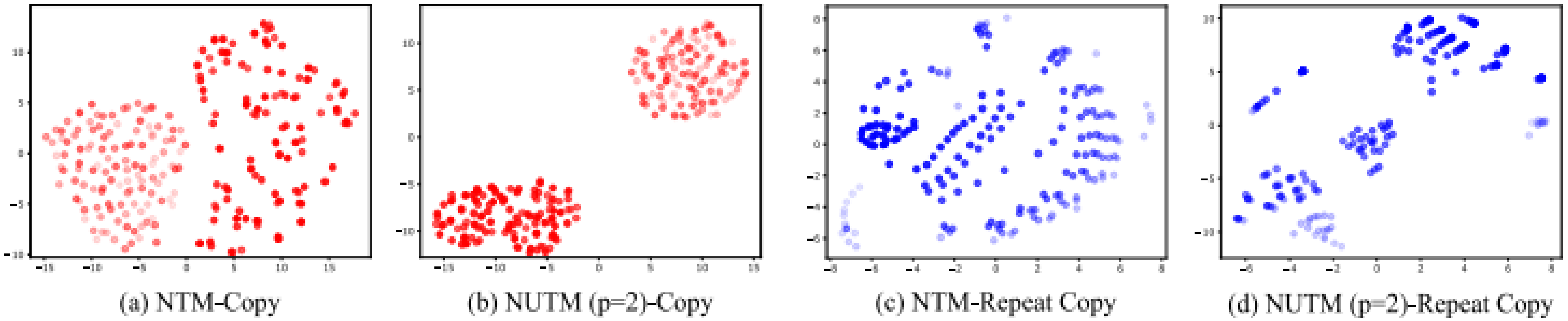
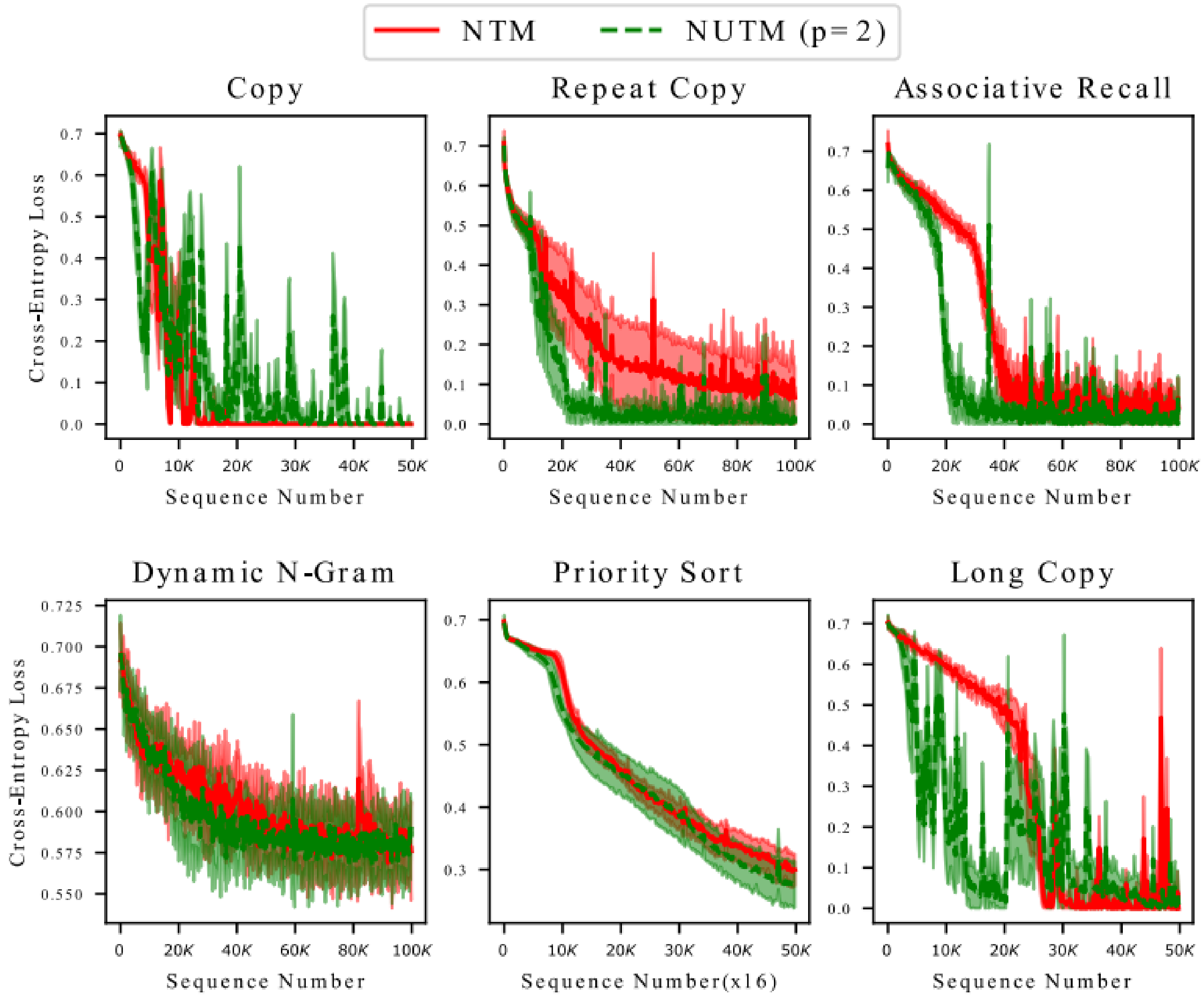


Figure 2: Visualization of the first two principal components of c_t space in NTM (a,c) and NUTM (b,d) for Copy (red) and Repeat Copy (blue). Fader color denotes lower timestep in a sequence. Both can learn clsuters of hidden states yet NUTM exhibits clearer partition.

Algorithmic single tasks



Sequencing tasks

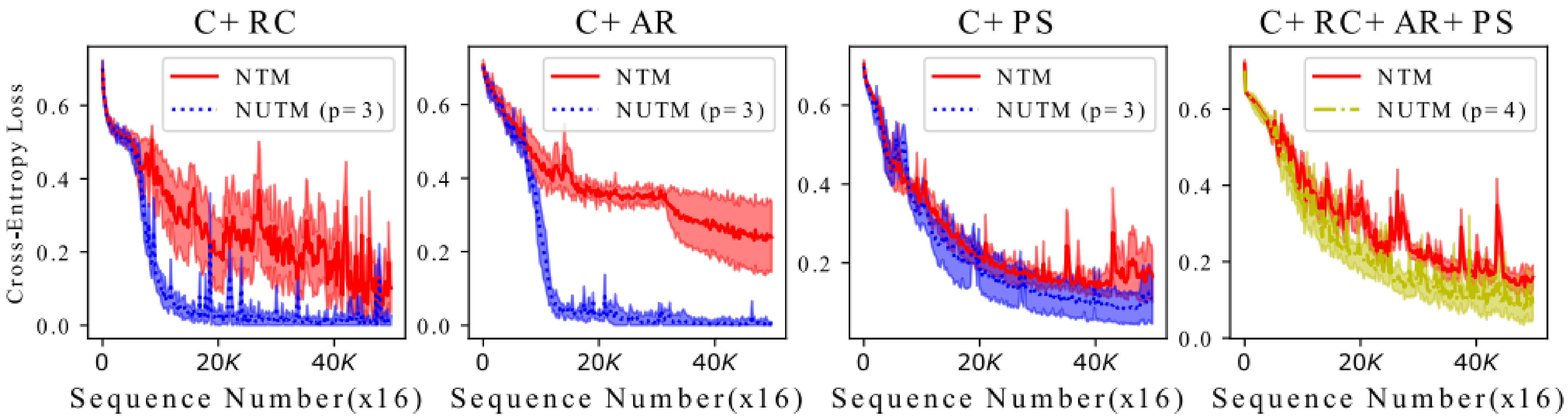


Figure 4: Learning curves on sequencing syntactic NTM tasks.

Continual Learning

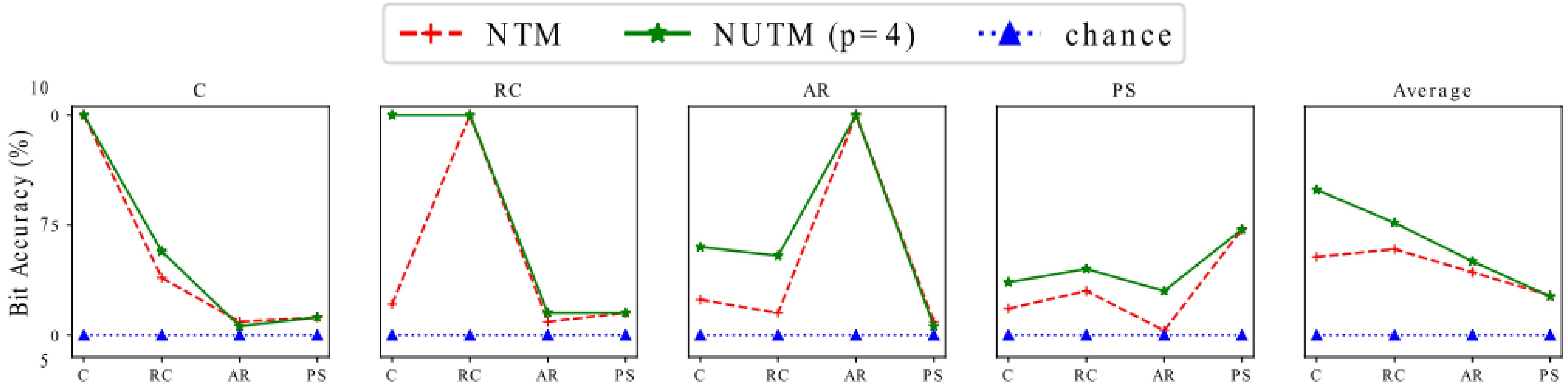


Figure 5: Mean bit accuracy for the continual algorithmic tasks. Each of the first four panels show bit accuracy on four tasks after finishing a task. The rightmost shows the average accuracy.

Few-shot learning

Model	Persistent memory[1]	5 classes			10 classes		
		2^{nd}	3^{rd}	5^{th}	2^{nd}	3^{rd}	5^{th}
MANN (LRUA)*	No	82.8	91.0	94.9	-	-	-
MANN (LRUA)	No	82.3	88.7	92.3	52.7	60.6	64.7
NUTM (LRUA)	No	**85.7**	**91.3**	**95.5**	**68.0**	**78.14**	**82.8**
Human*	Yes	57.3	70.1	81.4	-	-	-
MANN (LRUA)*	Yes	≈ 58.0	-	≈ 75.0	≈ 60.0	-	≈ 80.0
MANN (LRUA)	Yes	66.2	73.4	81.0	51.3	59.2	63.3
NUTM (LRUA)	Yes	**77.8**	**85.8**	**89.8**	**69.0**	**77.9**	**82.7**

Table 7: Test-set classification accuracy (%) on the Omniglot dataset after 100,000 episodes of training. * denotes available results from [3] (some are estimated from plotted figures).

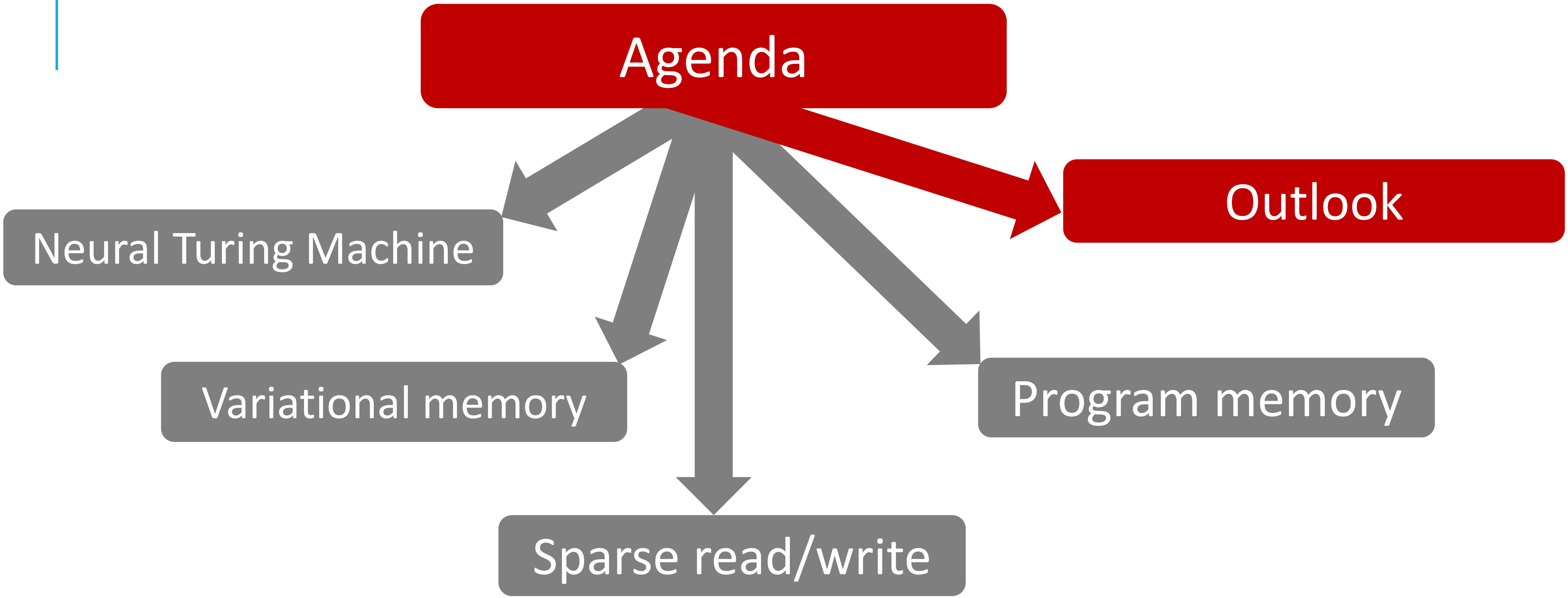
Question answering (bAbI dataset)

DNC[12]	SDNC[20]	ADNC[9]	DNC-MD[8]
16.7 ± 7.6	6.4 ± 2.5	6.3 ± 2.7	9.5 ± 1.6

Table 3: Mean and s.d. for bAbI err

Task	bAbI Best Results	bAbI Mean Results
1: 1 supporting fact	0.0	0.0 ± 0.0
2: 2 supporting facts	0.2	0.6 ± 0.3
3: 3 supporting facts	4.0	7.6 ± 3.9
4: 2 argument relations	0.0	0.0 ± 0.0
5: 3 argument relations	0.4	1.0 ± 0.4
6: yes/no questions	0.0	0.0 ± 0.0
7: counting	1.9	1.5 ± 0.8
8: lists/sets	0.6	0.3 ± 0.2
9: simple negation	0.0	0.0 ± 0.0
10: indefinite knowledge	0.1	0.1 ± 0.0
11: basic coreference	0.0	0.0 ± 0.0
12: conjunction	0.0	0.0 ± 0.0
13: compound coreference	0.1	0.0 ± 0.0
14: time reasoning	0.3	0.2 ± 0.1
15: basic deduction	0.0	2.6 ± 0.8
16: basic induction	49.3	52.0 ± 1.7
17: positional reasoning	4.7	18.4 ± 12.7
18: size reasoning	0.4	1.6 ± 1.1
19: path finding	4.3	23.7 ± 32.2
20: agent's motivation	0.0	0.0 ± 0.0
Mean Error (%)	3.3	5.6 ± 1.9
Failed (Err. >5%)	1	3 ± 1.2

Table 9: NUTM ($p = 4$) bAbI best and mean errors (%).


Agenda
Neural Turing Machine
Variational memory
Sparse read/write
Program memory
Outlook

Memory for graphs & relational structures

Turing machine to design machine learning algorithms

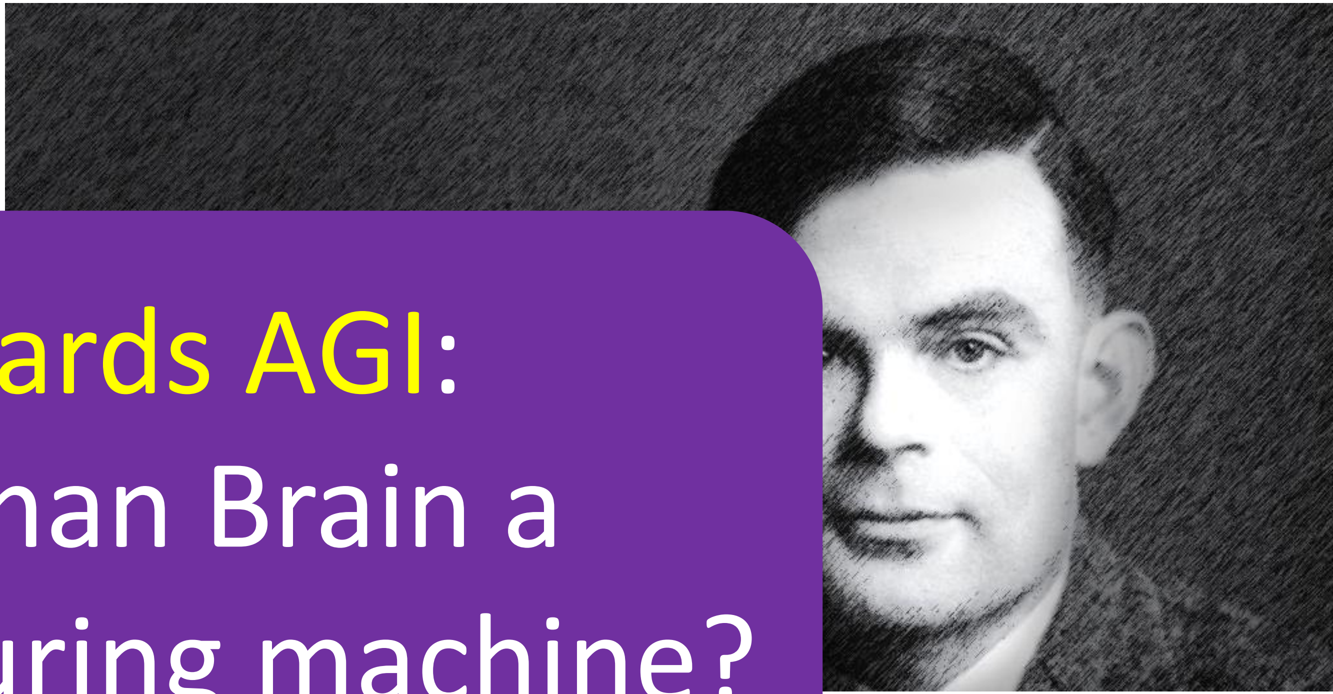
Memory-suppo

Imaginative me

Social memory: theory of mind, others

Full cognitive architectures

Theoretical analysis



Towards AGI:
Is Human Brain a (super-)Turing machine?