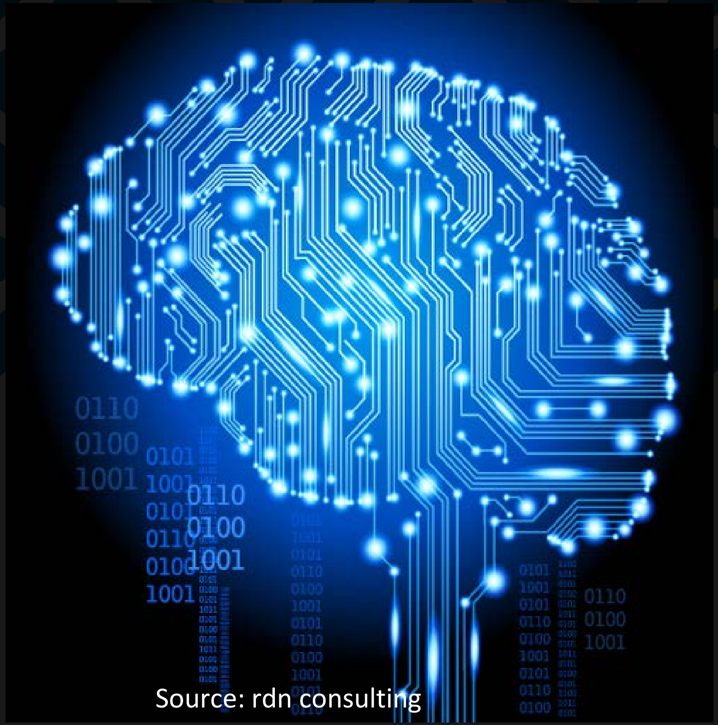
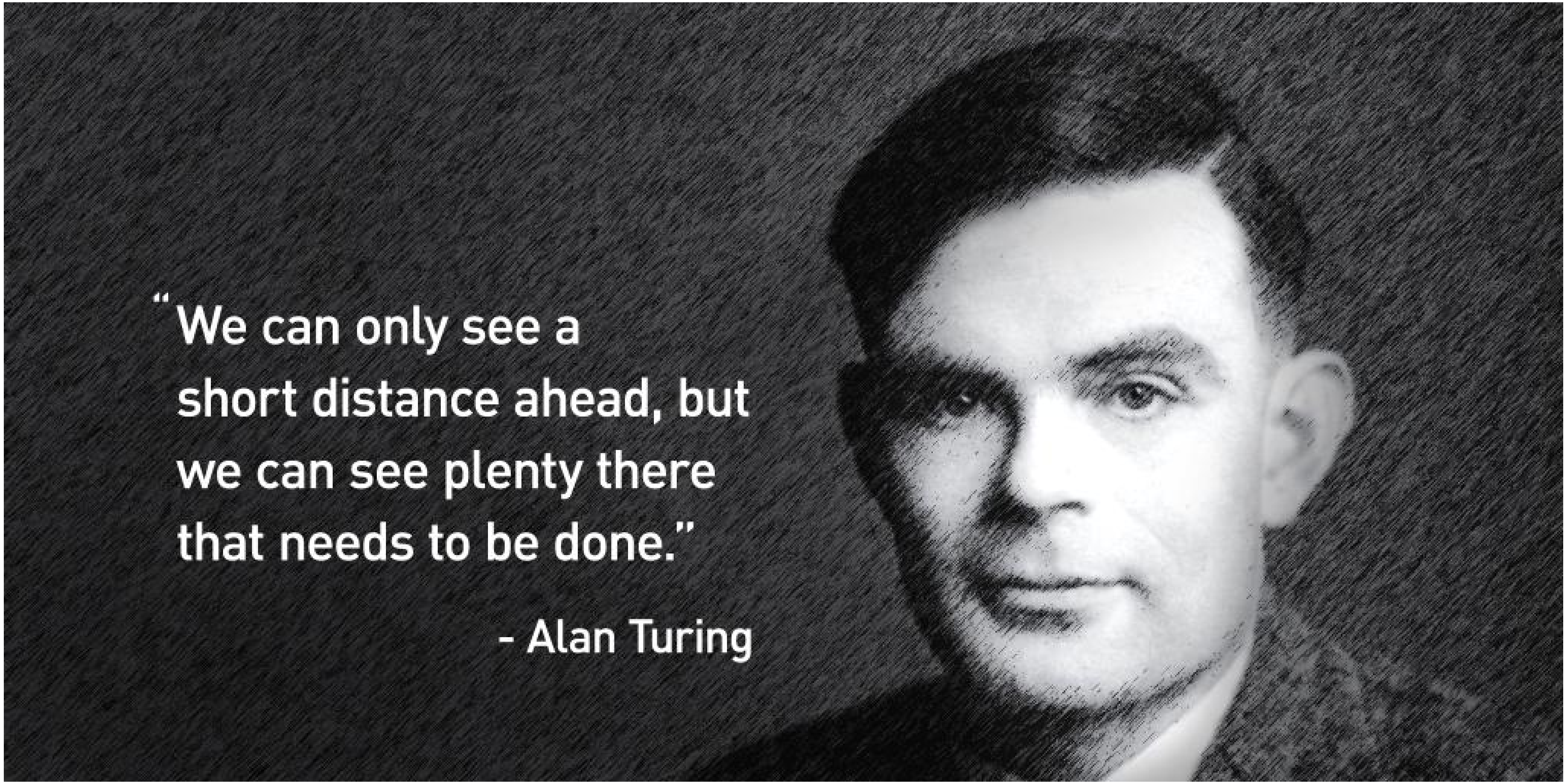
Advances in Neural Turing Machines


Source: rdn consulting

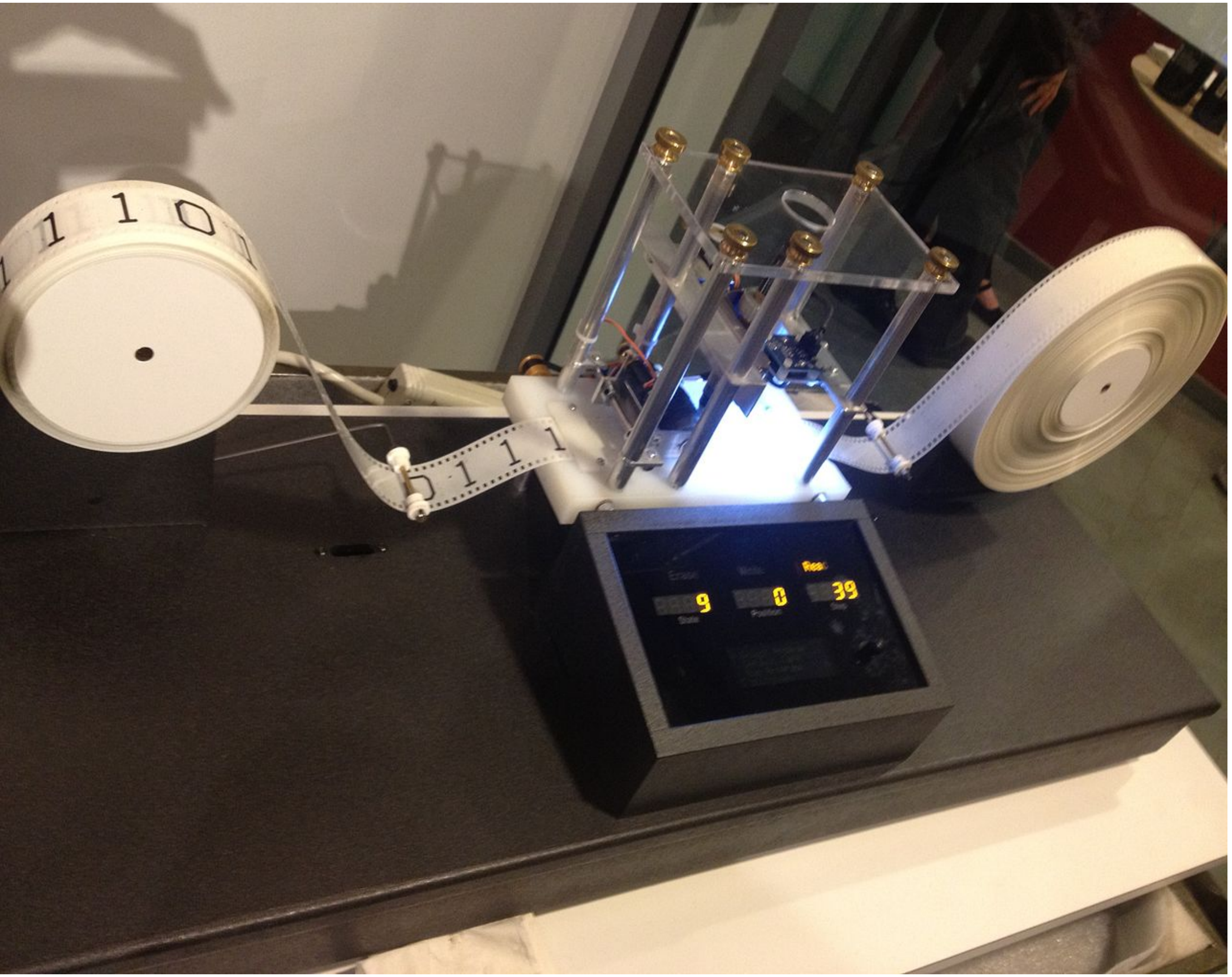
Truyen Tran
Deakin University

truyen.tran@deakin.edu.au
truyentran.github.io
@truyenoz
letdataspeak.blogspot.com
goo.gl/3jJ1O0

Aug 2018



https://twitter.com/nvidia/status/1010545517405835264



(Real) Turing machine

It is possible to invent a *single machine* which can be used to compute *any* computable sequence. If this machine **U** is supplied with the tape on the beginning of which is written the string of quintuples separated by semicolons of some computing machine **M**, then **U** will compute the same sequence as **M**.

Wikipedia

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

Agenda

Neural Turing machine (NTM)

Dual-view in sequences (KDD'18)

Bringing variability in output sequences (NIPS'18)

Bringing relational structures into memory (ICPR'18+)

Looking ahead (ACL'19, KDD'19, CVPR'19, ICML'19, NIPS'19 ?)

Let’s review current offerings

Feedforward nets (FFN)

Recurrent nets (RNN)

Convolutional nets (CNN)

Message-passing graph nets (MPGNN)

Universal transformer

…..

Work surprisingly well on LOTS of important problems

Enter the age of differentiable programming

BUTS …

No storage of intermediate results.

Little choices over what to compute and what to use

Little support for complex chained reasoning

Little support for rapid switching of tasks

Searching for better priors

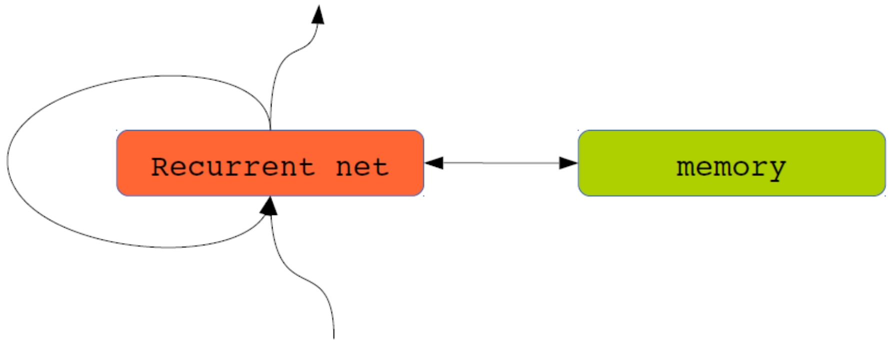
Translation invariance in CNN

Recurrence in RNN

Permutation invariance in attentions and graph neural networks

Memory for complex computation

→ **Memory-augmented neural networks (MANN)**



(LeCun, 2015)

What is missing? A memory

Use multiple pieces of information

Store intermediate results (RAM like)

Episodic recall of previous tasks (Tape like)

Encode/compress & generate/decompress long sequences

Learn/store programs (e.g., fast weights)

Store and query external knowledge

Spatial memory for navigation

Rare but important events (e.g., snake bite)

Needed for complex control

Short-cuts for ease of gradient propagation = constant path length

Division of labour: program, execution and storage

Working-memory is an indicator of IQ in human

Example: Code language model

```
FileWriter writer = new FileWriter(file);
writer.write(``This is an example'');
int count = 0;
System.out.prinltln(``Long gap'');
    ......

writer.flush();
writer.close();
```

Still needs a better memory for:

Repetitiveness

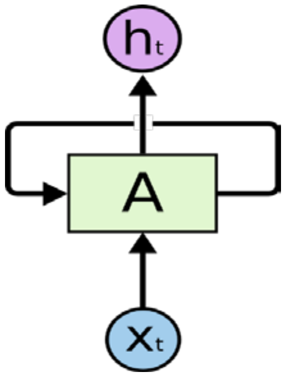
E.g. for (int i = 0; i < n; i++)

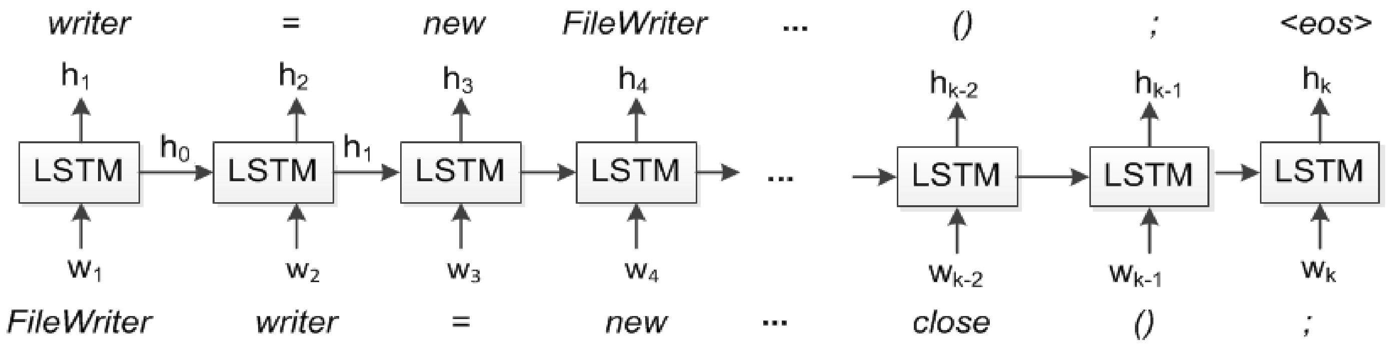
Localness

E.g. *for (int size* may appear more often that *for (int i* in some source files.

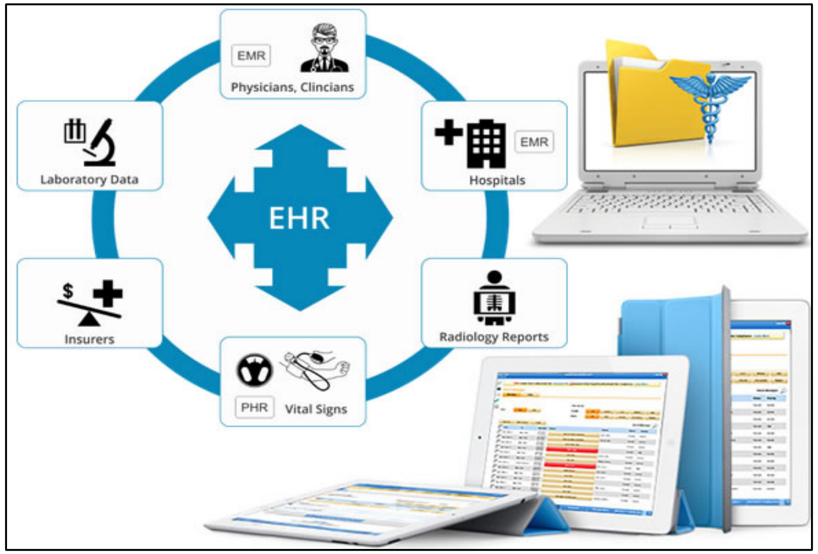
Very long sequence (big file, or char level)

$$P(s) = P(w_1) \prod_{t=2}^{k} P\left(w_t \mid \boldsymbol{w}_{1:t-1}\right)$$



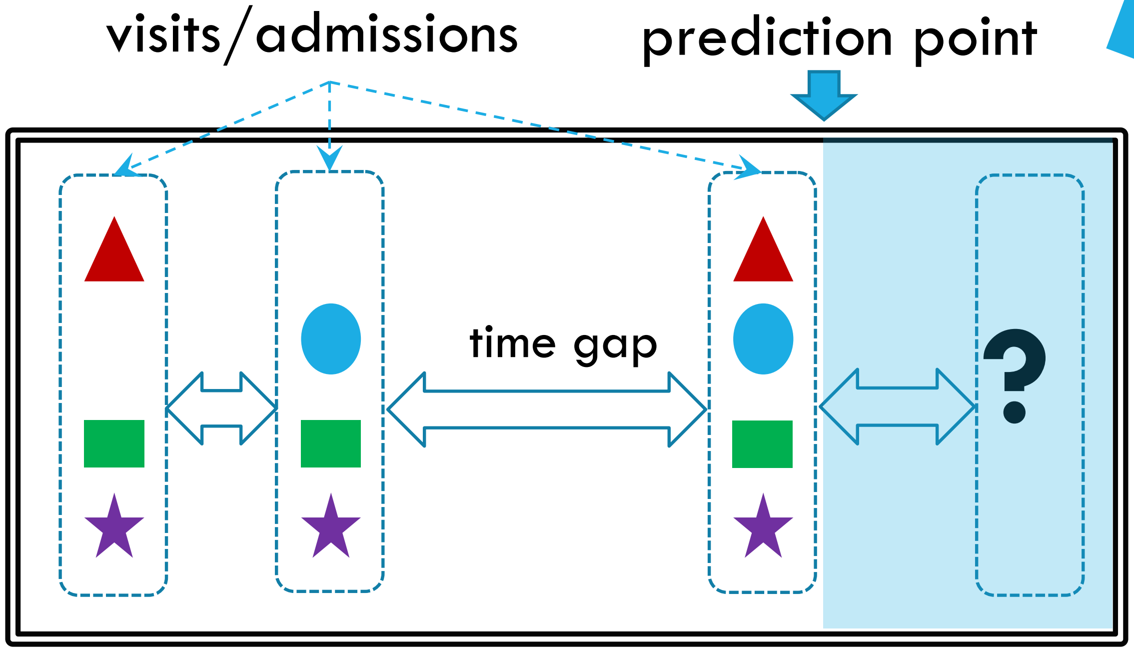


Example: Electronic medical records



Source: medicalbillingcodings.org

Abstraction



Modelling

Three interwoven processes:

- Disease progression
- Interventions & care processes
- Recording rules

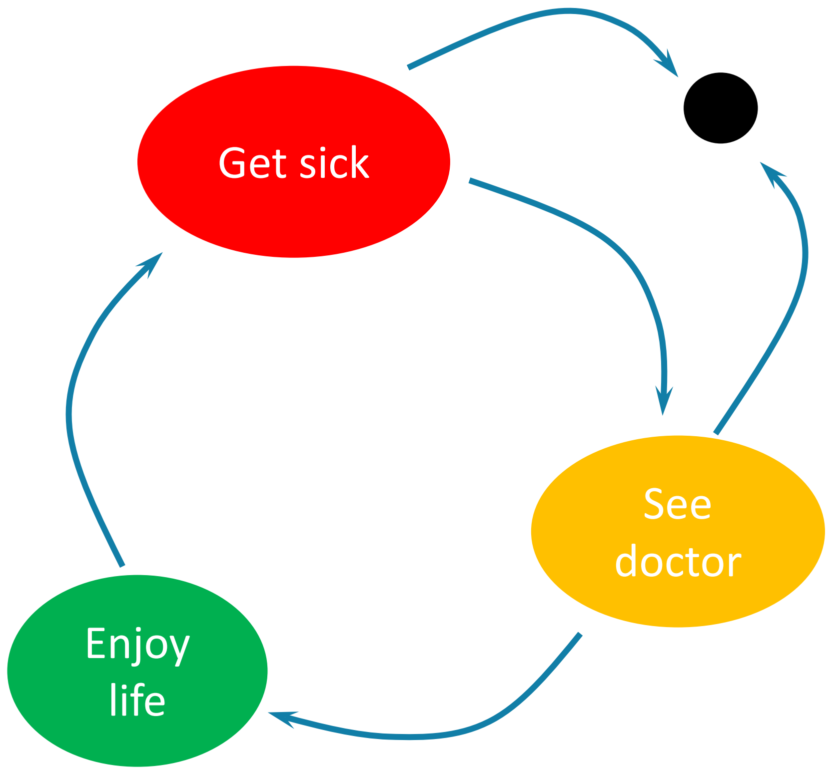
Need memory to handle thousands of events

Conjecture: Healthcare is Turing computational

Healthcare processes as executable computer program obeying hidden "grammars"

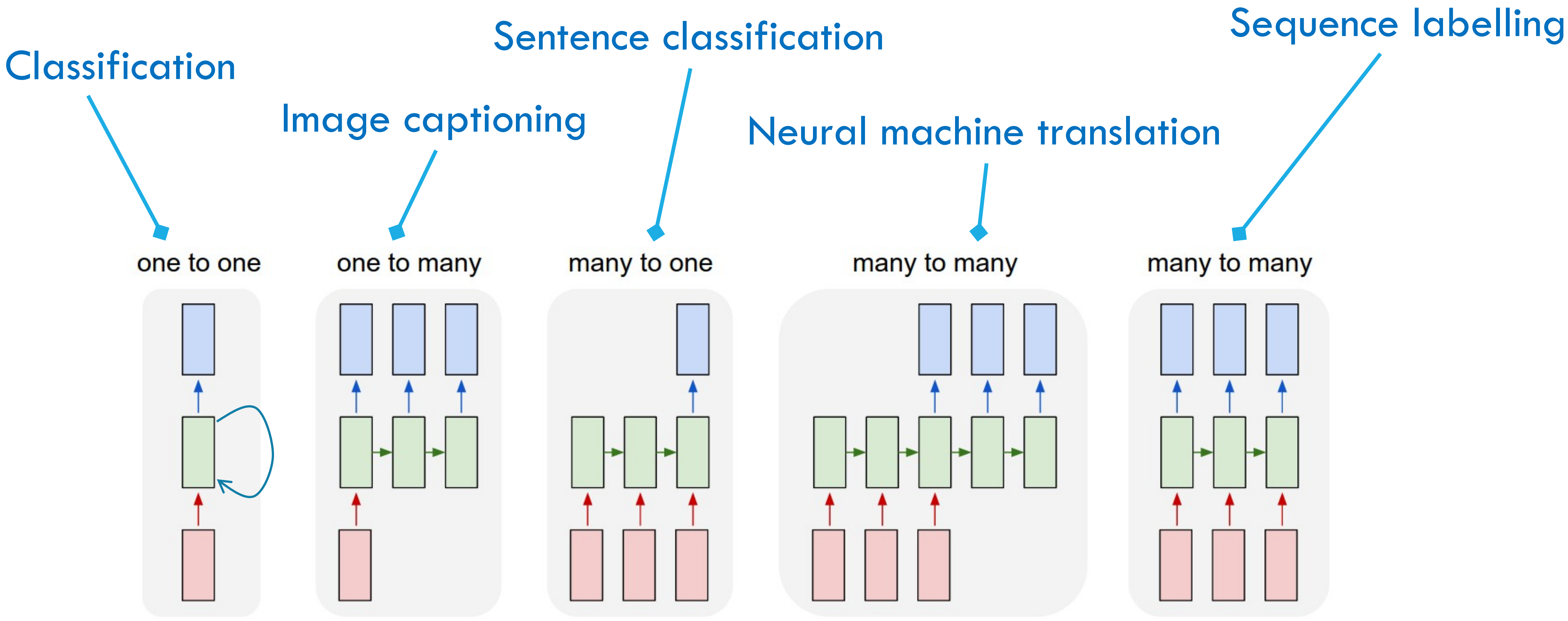
The "grammars" are learnable through observational data

With "generative grammars", entire health trajectory can be simulated.



Neural Turing machine (NTM)

RNN: theoretically powerful, practically limited



Source: http://karpathy.github.io/assets/rnn/diags.jpeg

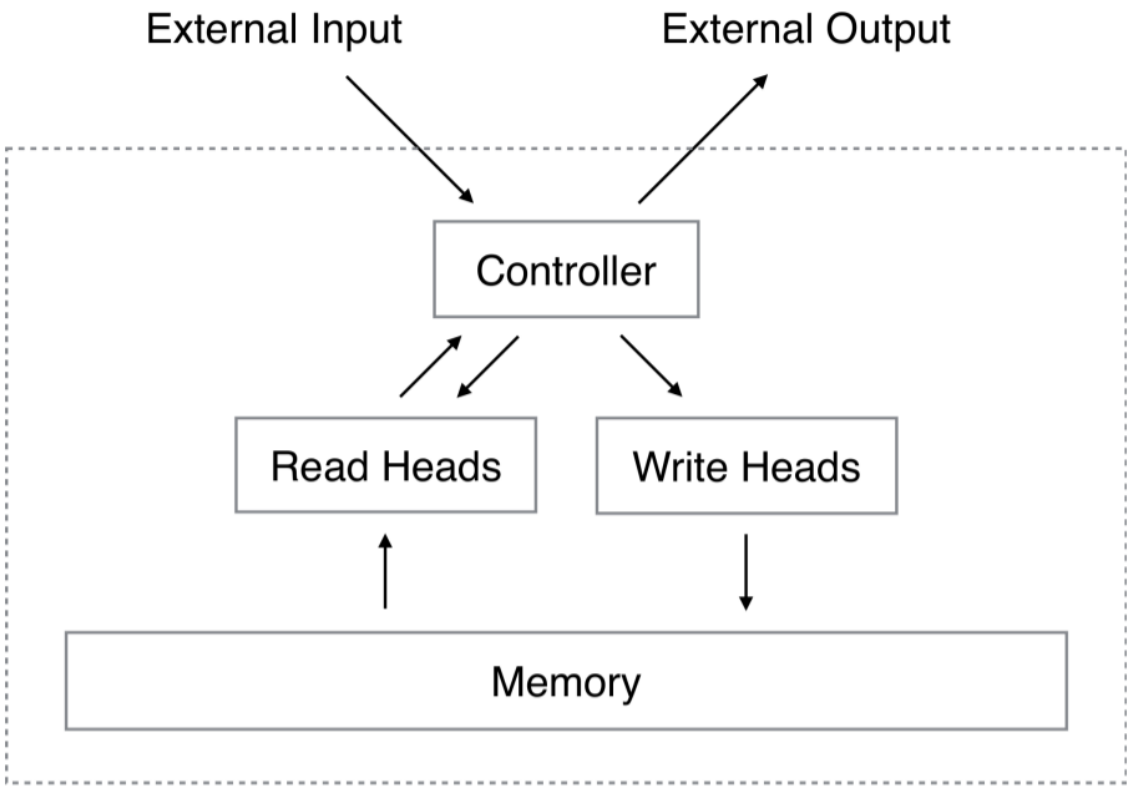
Neural Turing machine (NTM)

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

Memory has read/write operations.

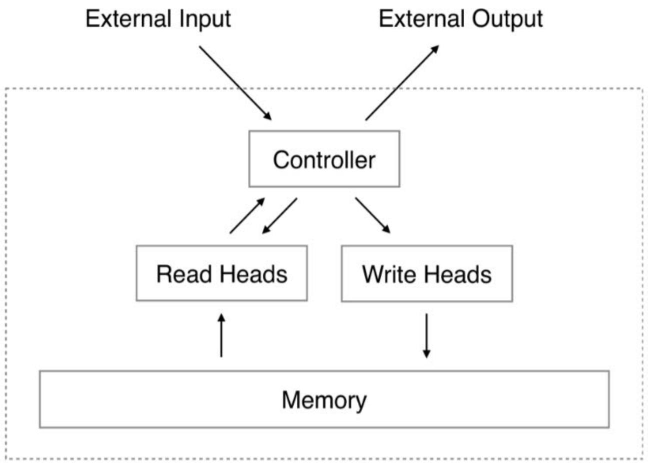
The main issue is where to write, and how to update the memory state.

All operations are differentiable.

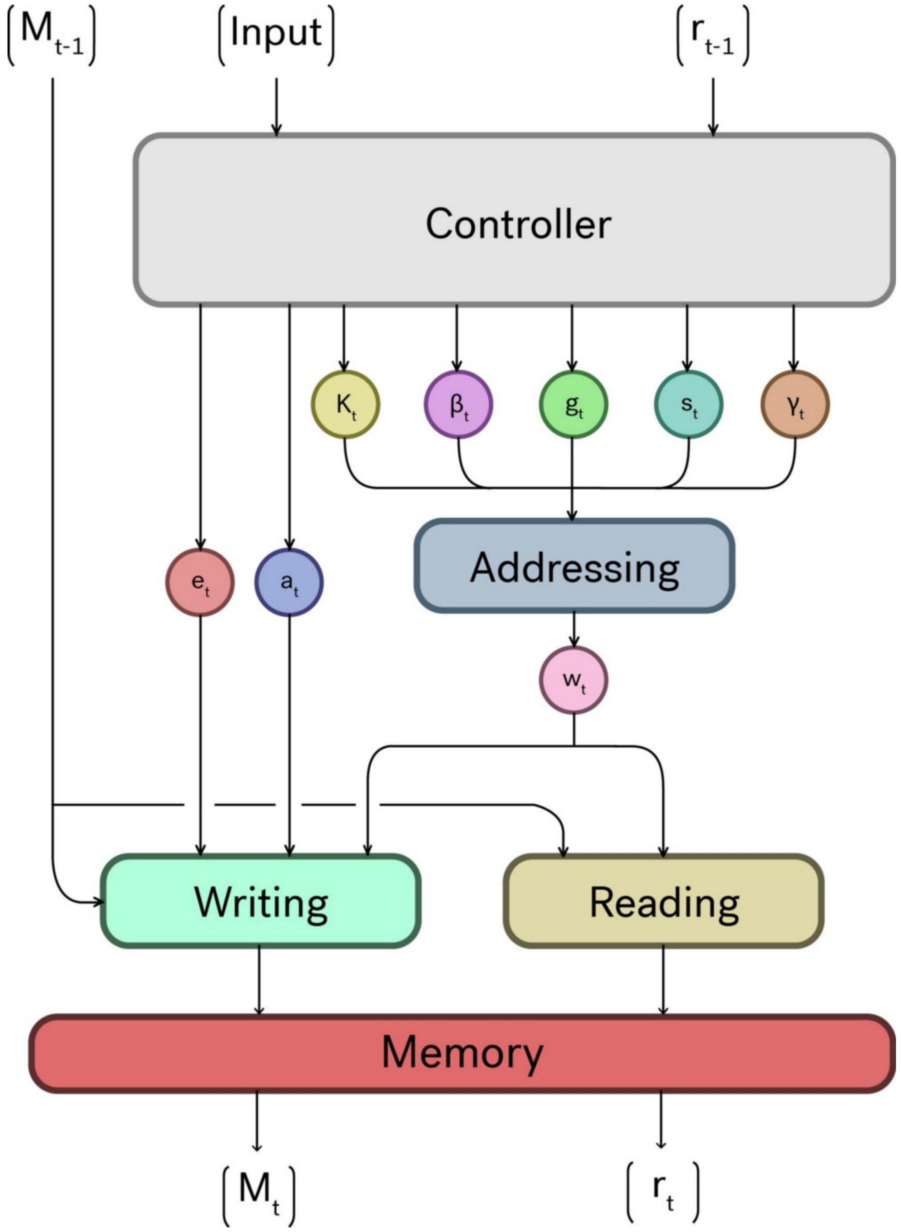


https://rylanschaeffer.github.io/content/research/neural_turing_machine/main.html

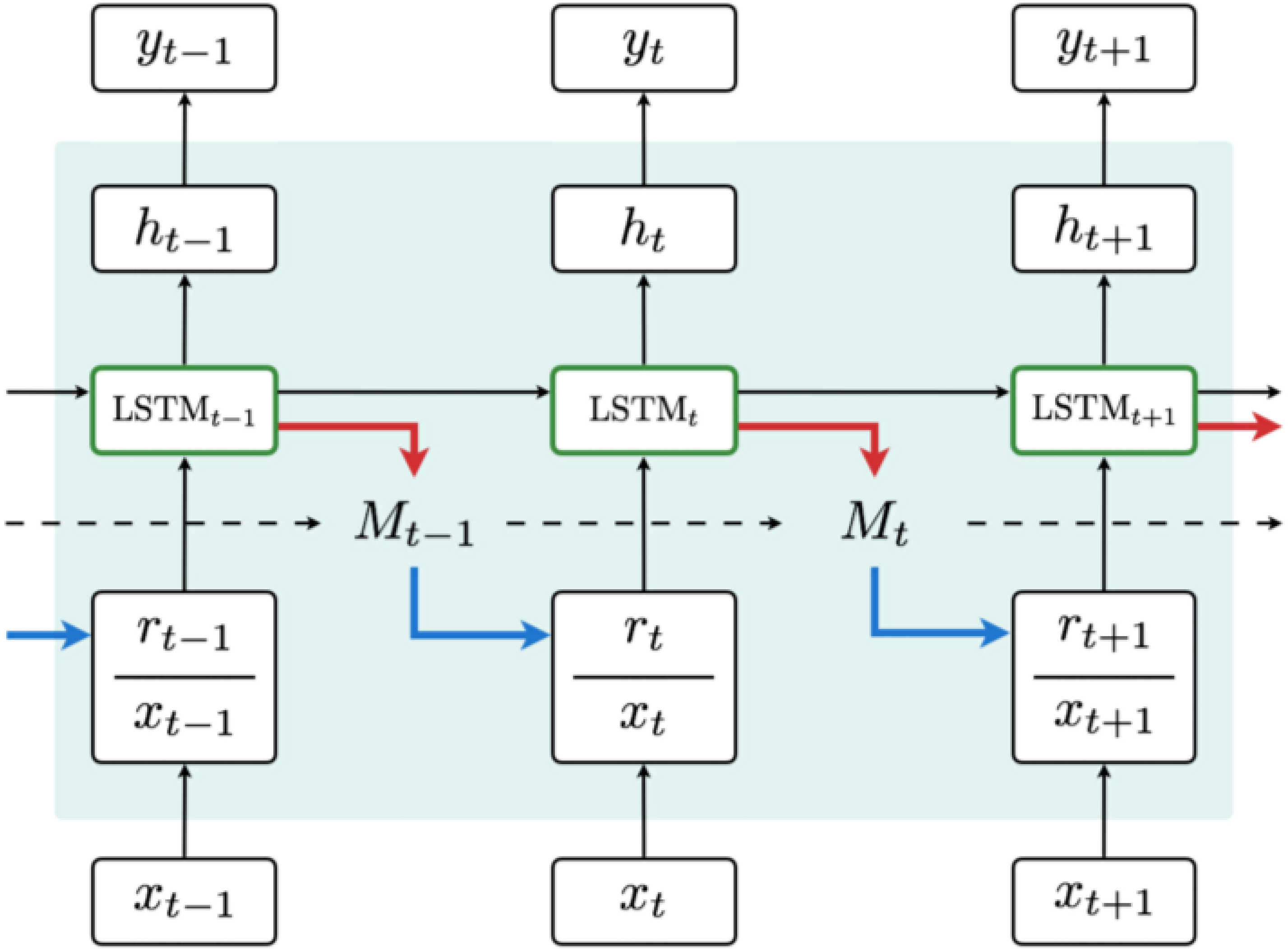
NTM operations



https://rylanschaeffer.github.io



https://medium.com/@aidangomez/the-neural-turing-machine-79f6e806c0a1

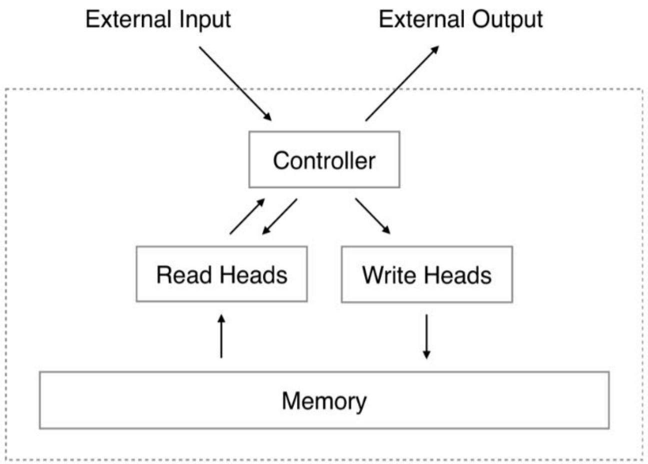


NTM unrolled in time with LSTM as controller

#Ref: https://medium.com/snips-ai/ntm-lasagne-a-library-for-neural-turing-machines-in-lasagne-2cdce6837315

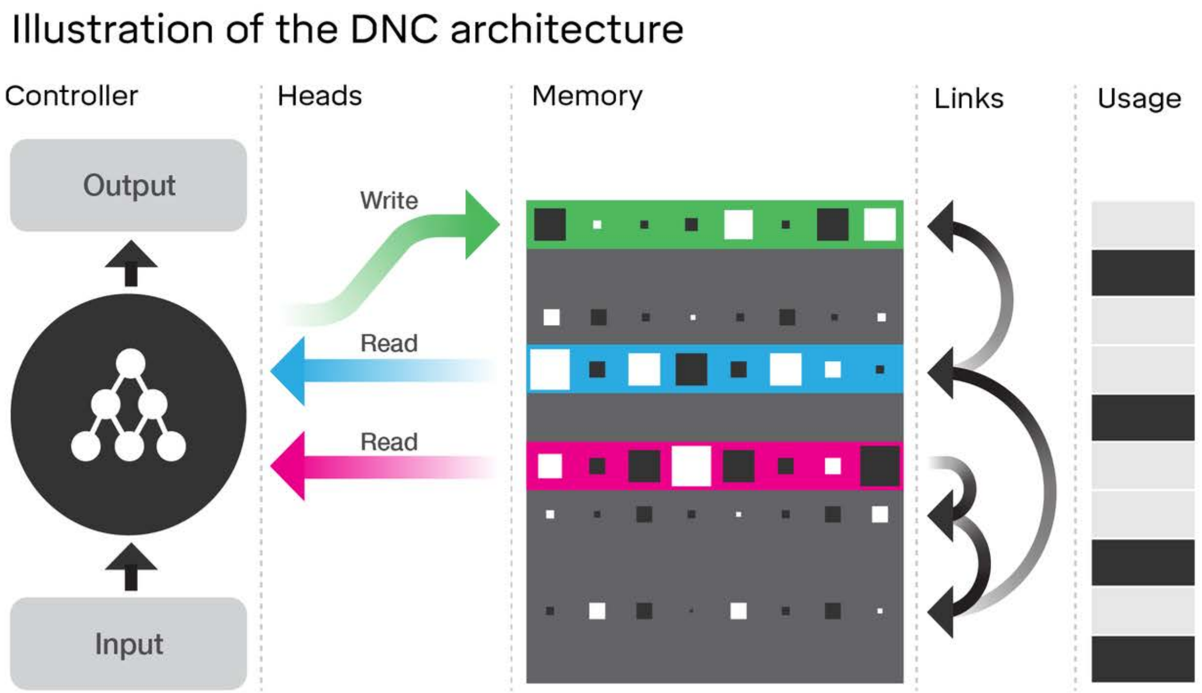
Differentiable neural computer (DNC)

2014



https://rylanschaeffer.github.io

2016



Source: deepmind.com

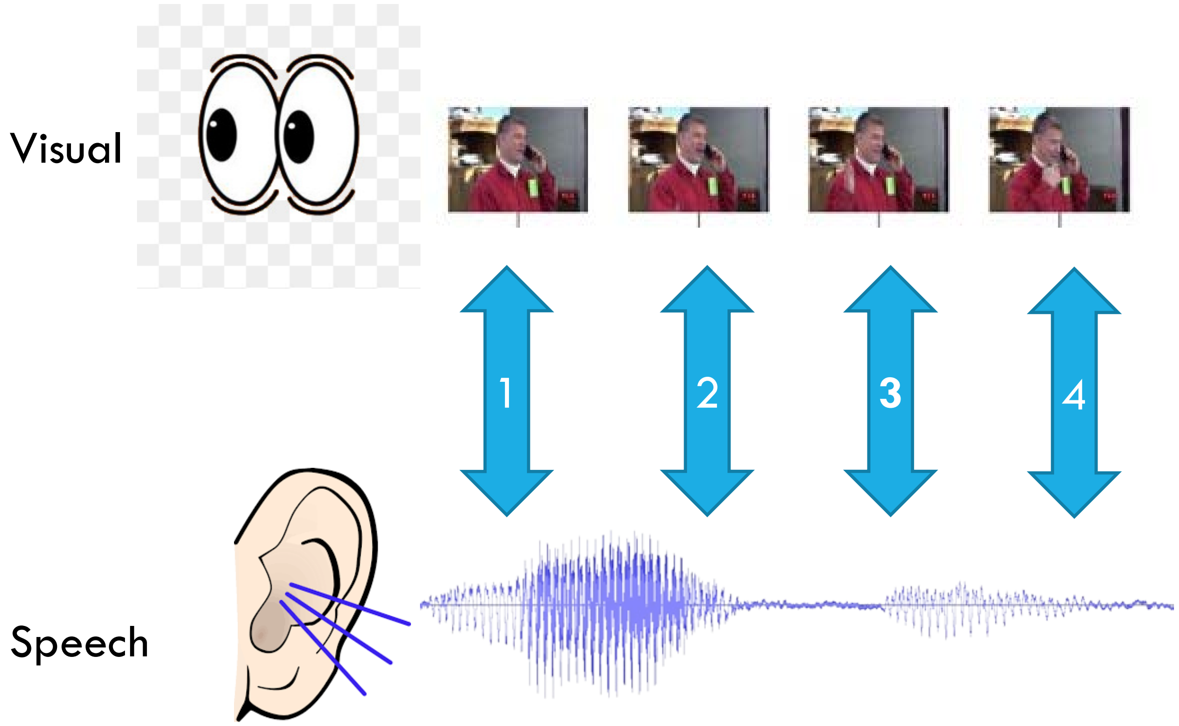
#REF: Graves, Alex, et al. "Hybrid computing using a neural network with dynamic external memory." *Nature* 538.7626 (2016): 471-476.

Dual-view sequential problems

Hung Le, Truyen Tran & Svetha Venkatesh

KDD'18

Synchronous two-view sequential learning



Asynchronous two-view sequential learning Healthcare: medicine prescription

Diagnoses

E11 | I10 | N18 | Z86 | E11

1916 | 1910 | 1952 | 1893

Procedures

Medicines

DOCU100L | ACET325

Asynchronous two-view sequential learning Healthcare: disease progression

Previous diagnoses

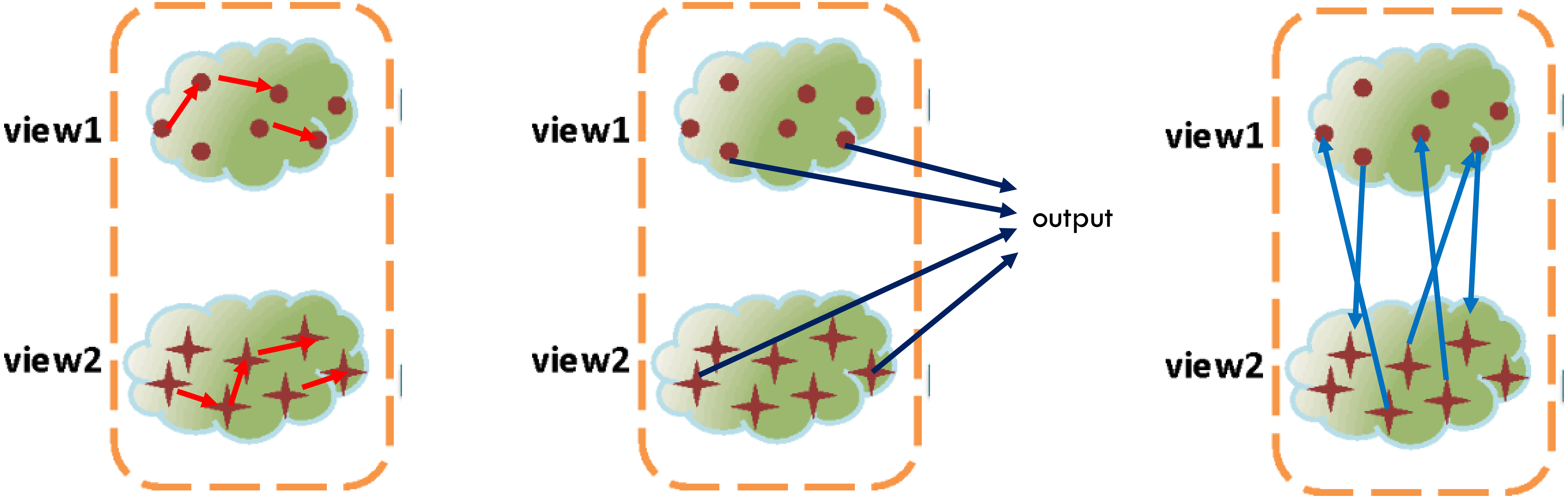
Future diagnoses ???

E11 I10 N18 Z86 E11

1916 1910 ACET325 DOCU100L

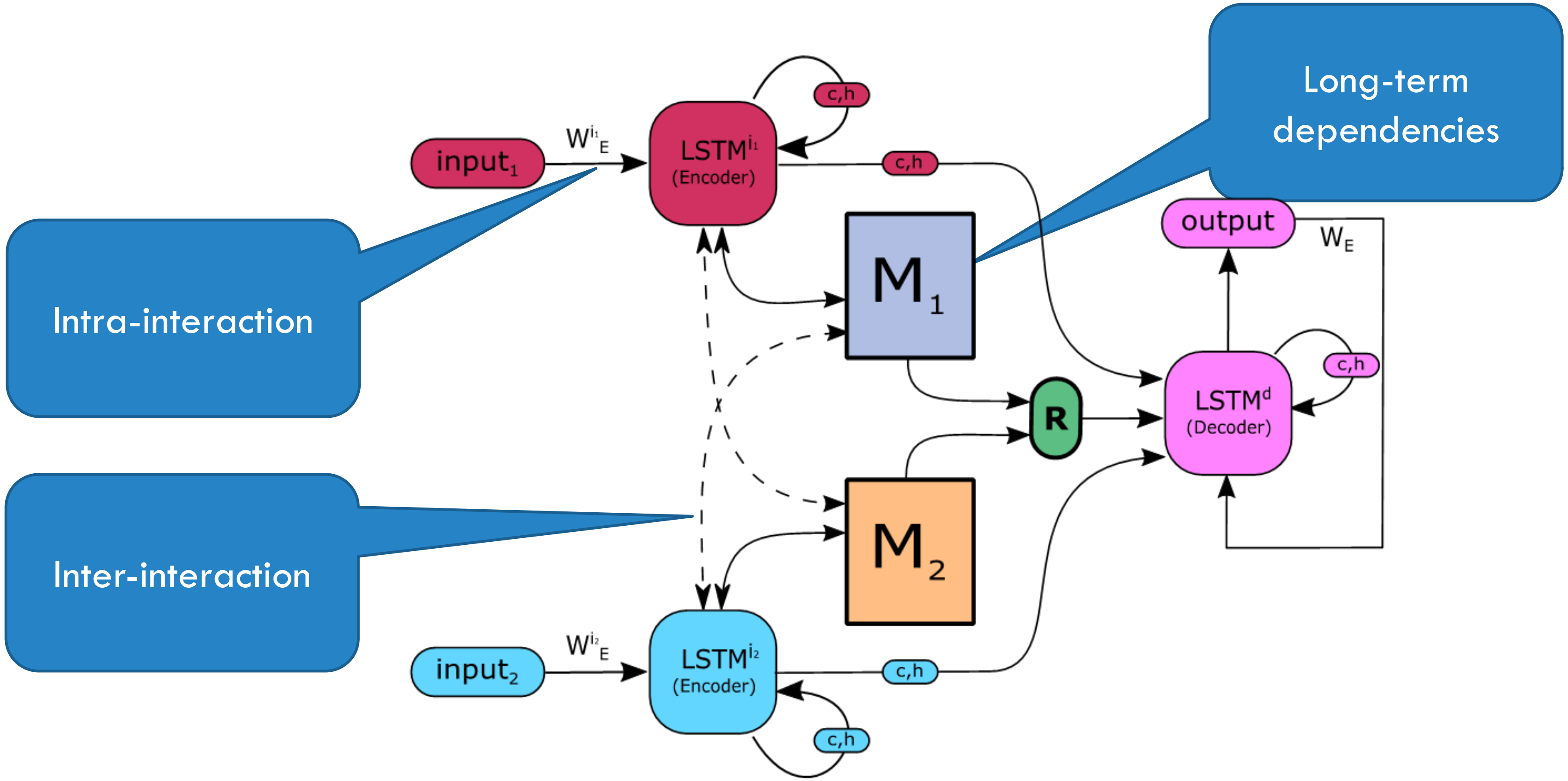
Previous interventions

Intra-view & inter-view interactions



#Ref: Le, Hung, Truyen Tran, and Svetha Venkatesh. "Dual Memory Neural Computer for Asynchronous Two-view Sequential Learning." *KDD18*.

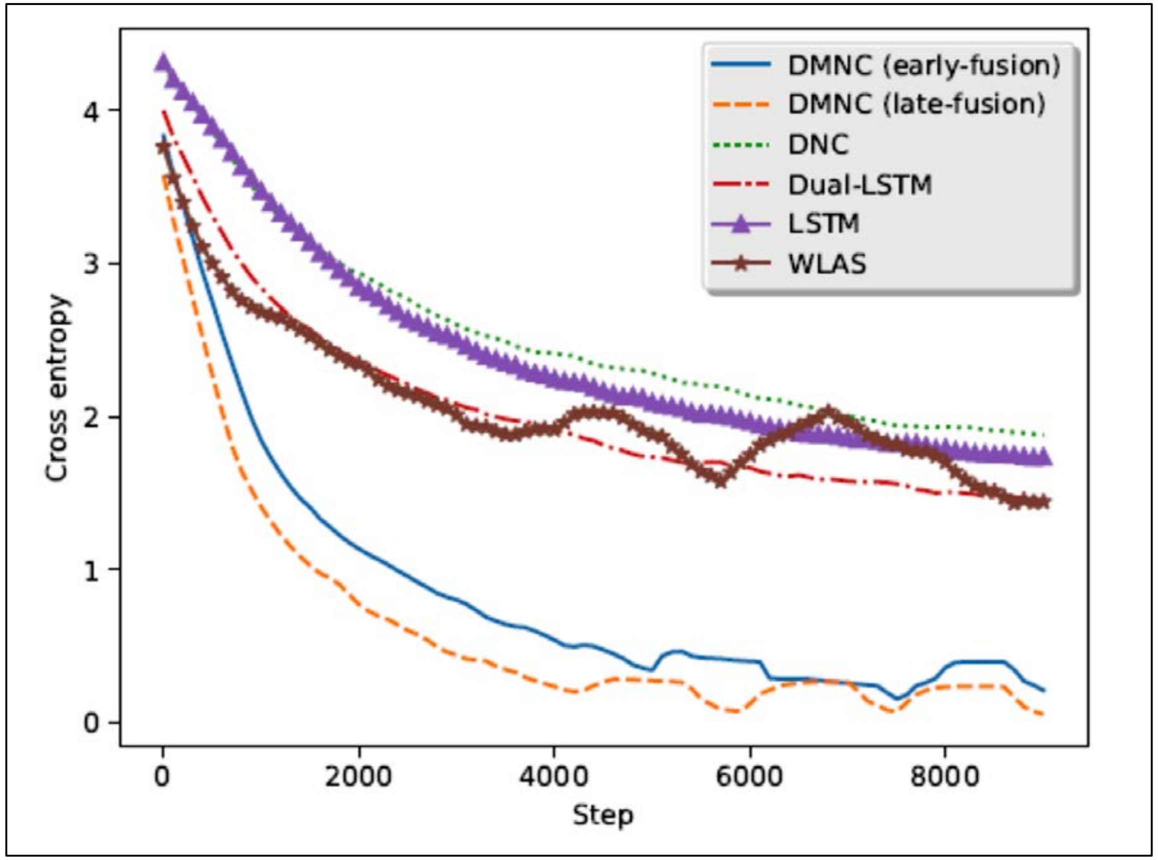
Dual architecture



***Dual Memory Neural Computer (DMNC)**. There are two encoders and one decoder implemented as LSTMs. The dash arrows represent cross-memory accessing in early-fusion mode*

Simple sum, but distant, asynchronous

$$\{y_i = x_i^1 + x_{L+1-i}^2\}_{i=1}^{L}$$



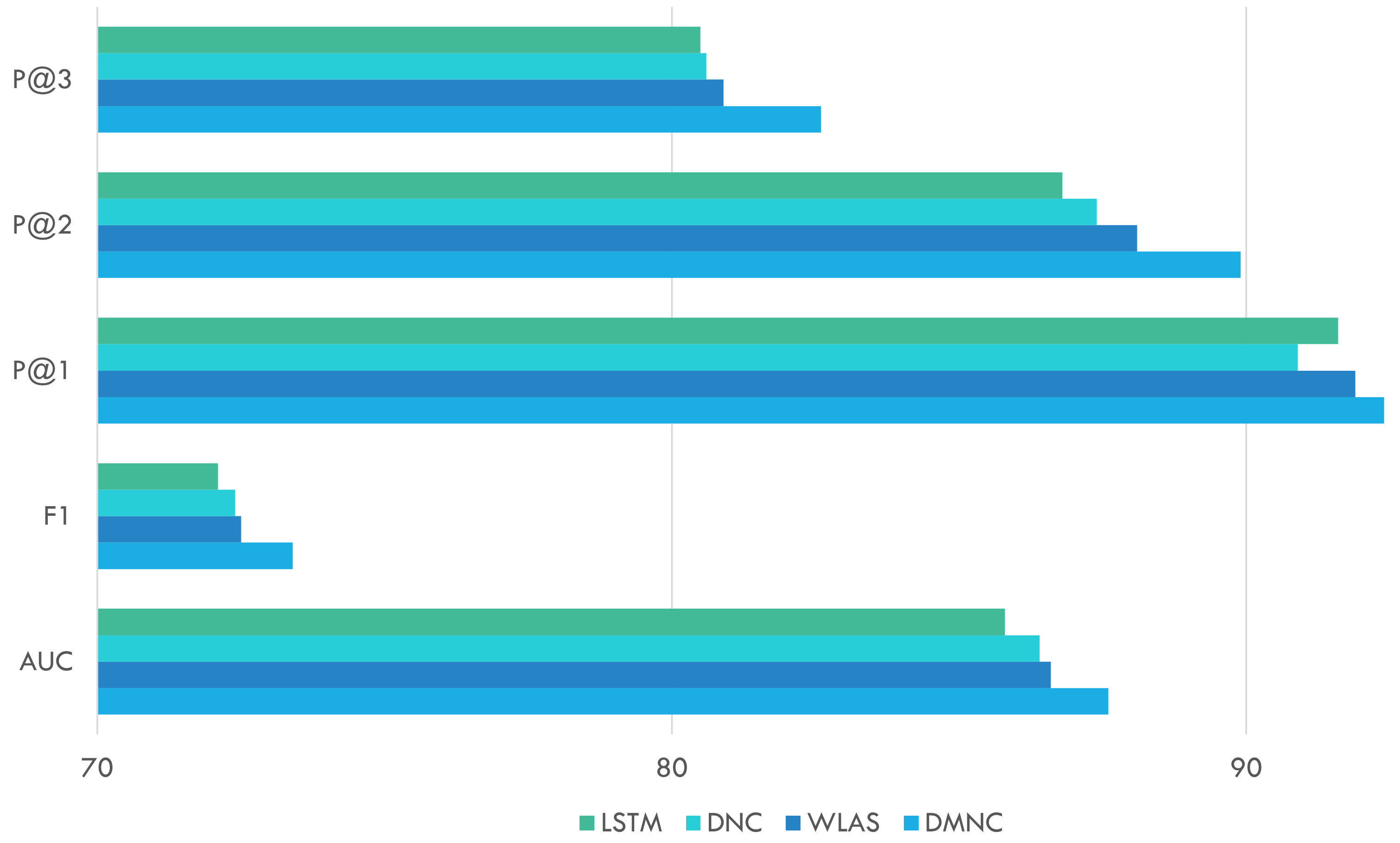
Learning curve

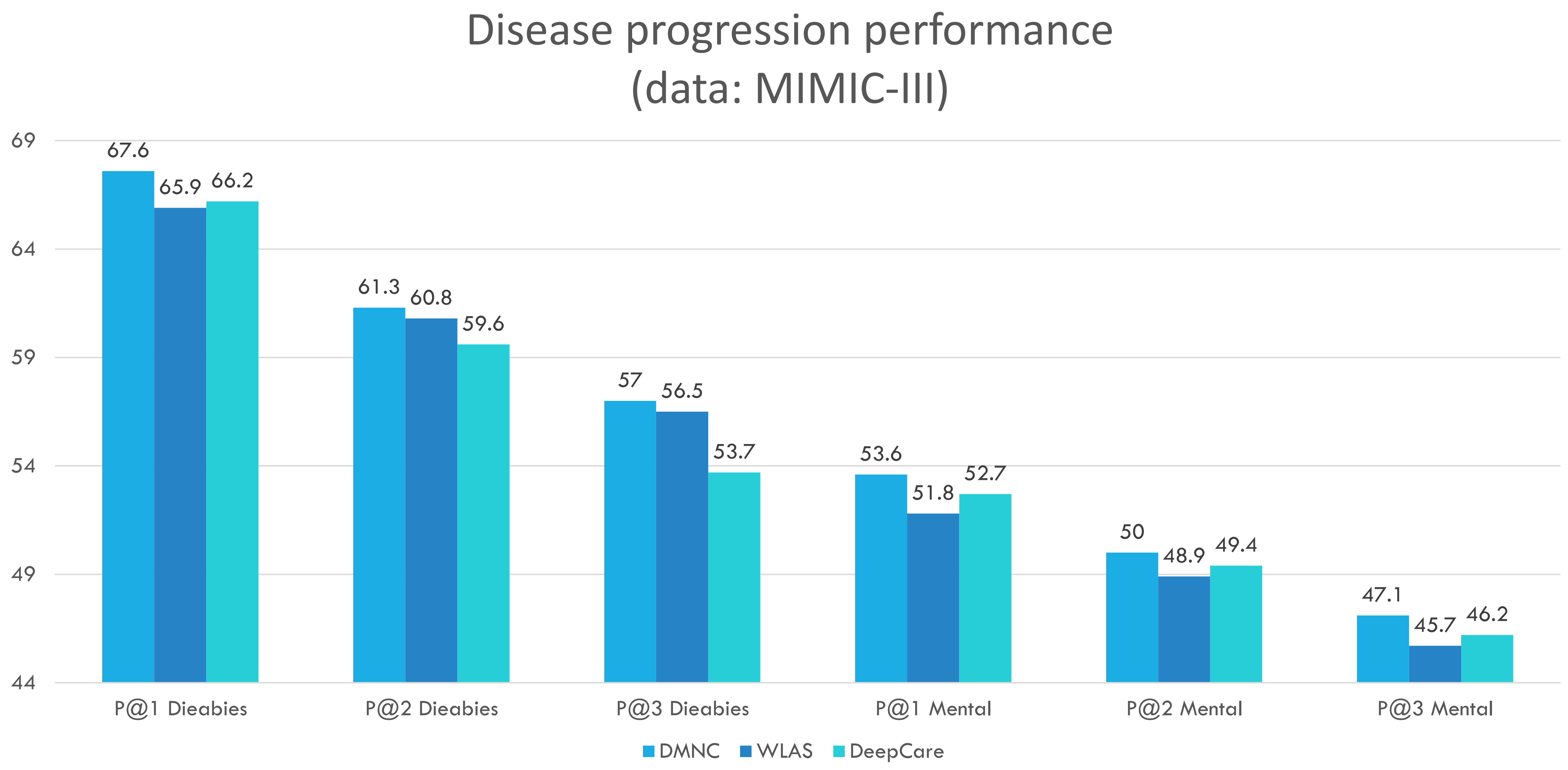
Accuracy

DMNC $\approx 99\%$

Others $<55\%$

Medicine prescription performance (data: MIMIC-III)




Disease progression performance
(data: MIMIC-III)
69
64
59
54
49
44
67.6
65.9
66.2
61.3
60.8
59.6
57
56.5
53.7
53.6
51.8
52.7
50
48.9
49.4
47.1
45.7
46.2
P@1 Dieabies
P@2 Dieabies
P@3 Dieabies
P@1 Mental
P@2 Mental
P@3 Mental
DMNC
WLAS
DeepCare

Bringing variability in output sequences

Hung Le, Truyen Tran & Svetha Venkatesh

NIPS'18

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").

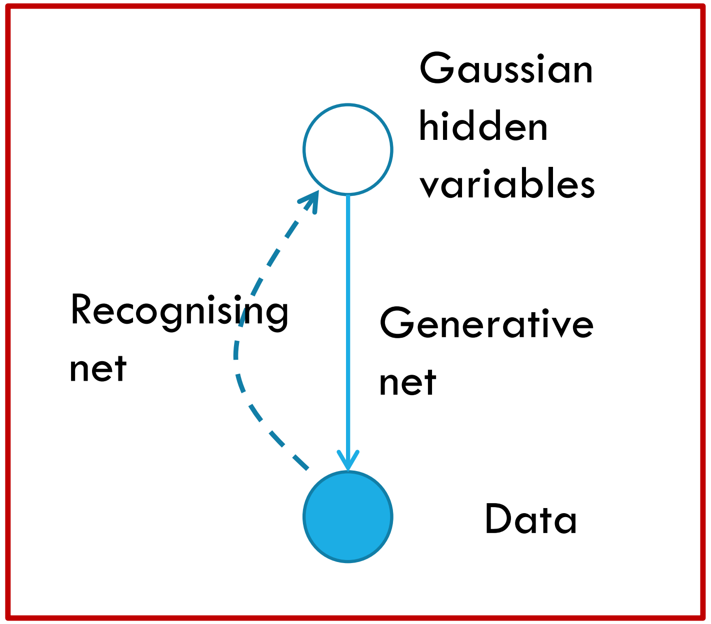
There are many ways to express the same thought

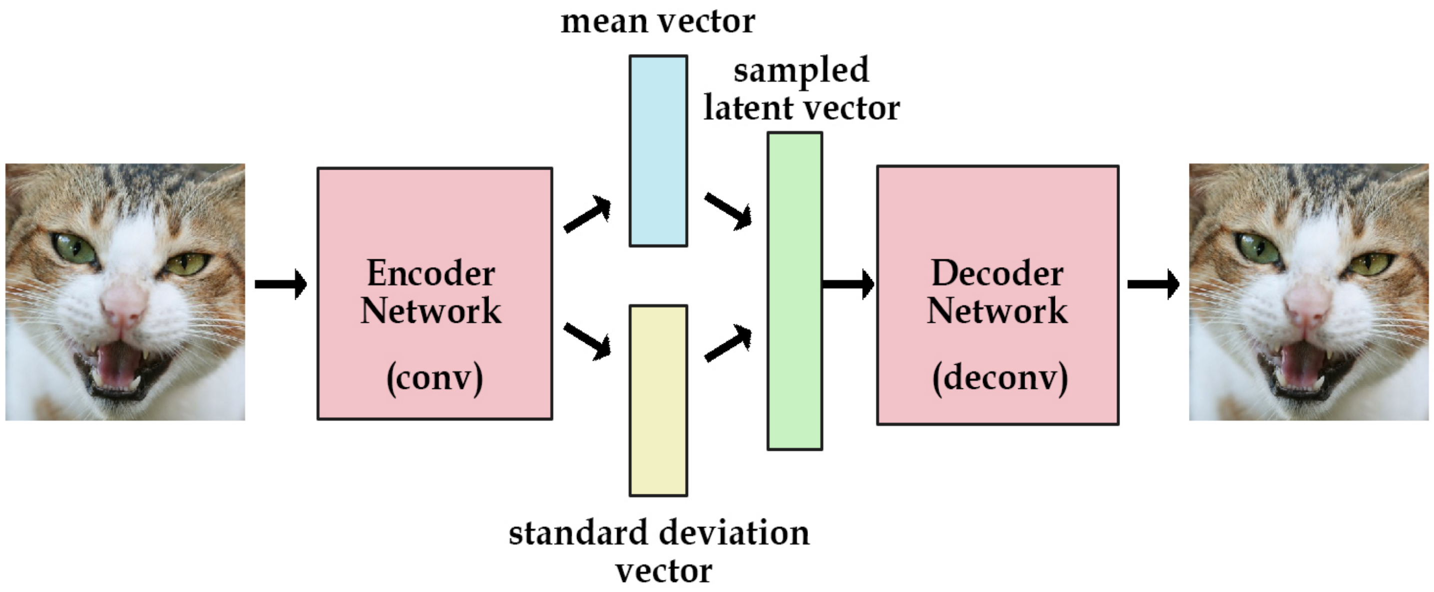
- → Variational generative methods are needed.

Variational Auto-Encoder (VAE)

(Kingma & Welling, 2014)

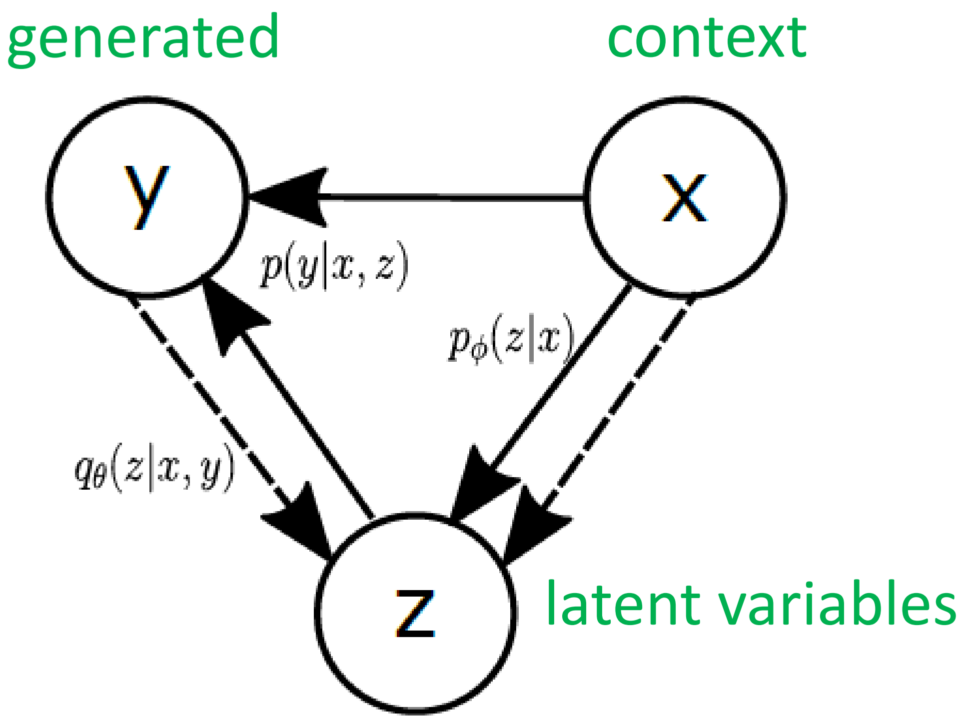
Two separate processes: generative (hidden → visible) versus recognition (visible → hidden)



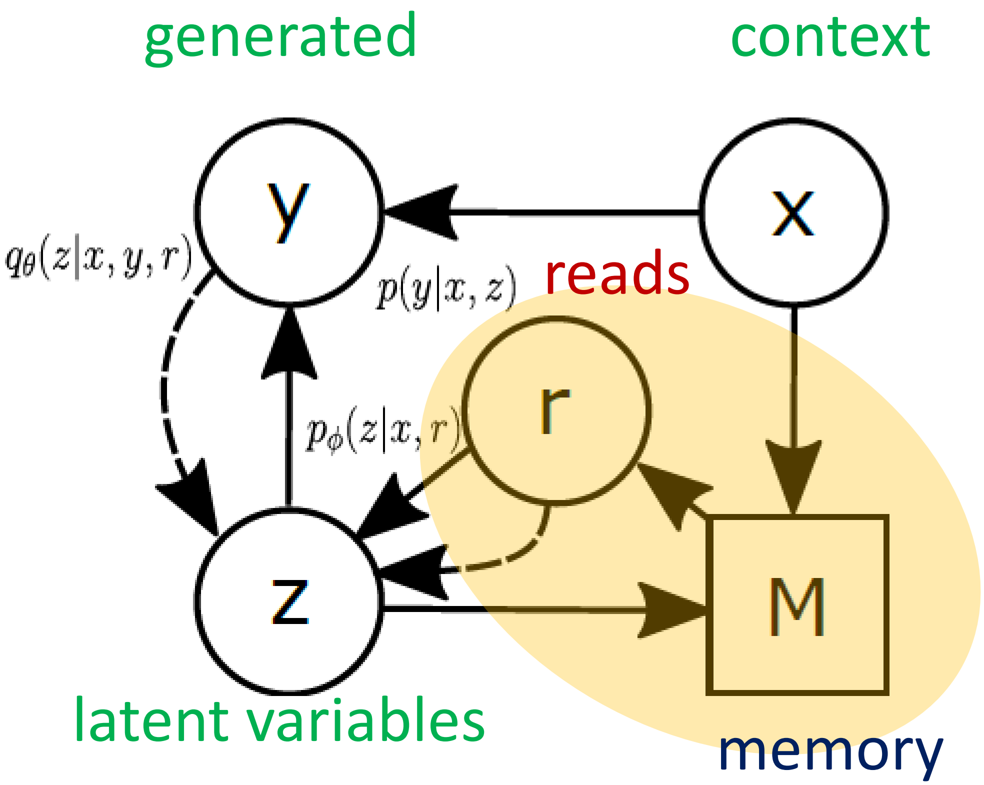


http://kvfrans.com/variational-autoencoders-explained/

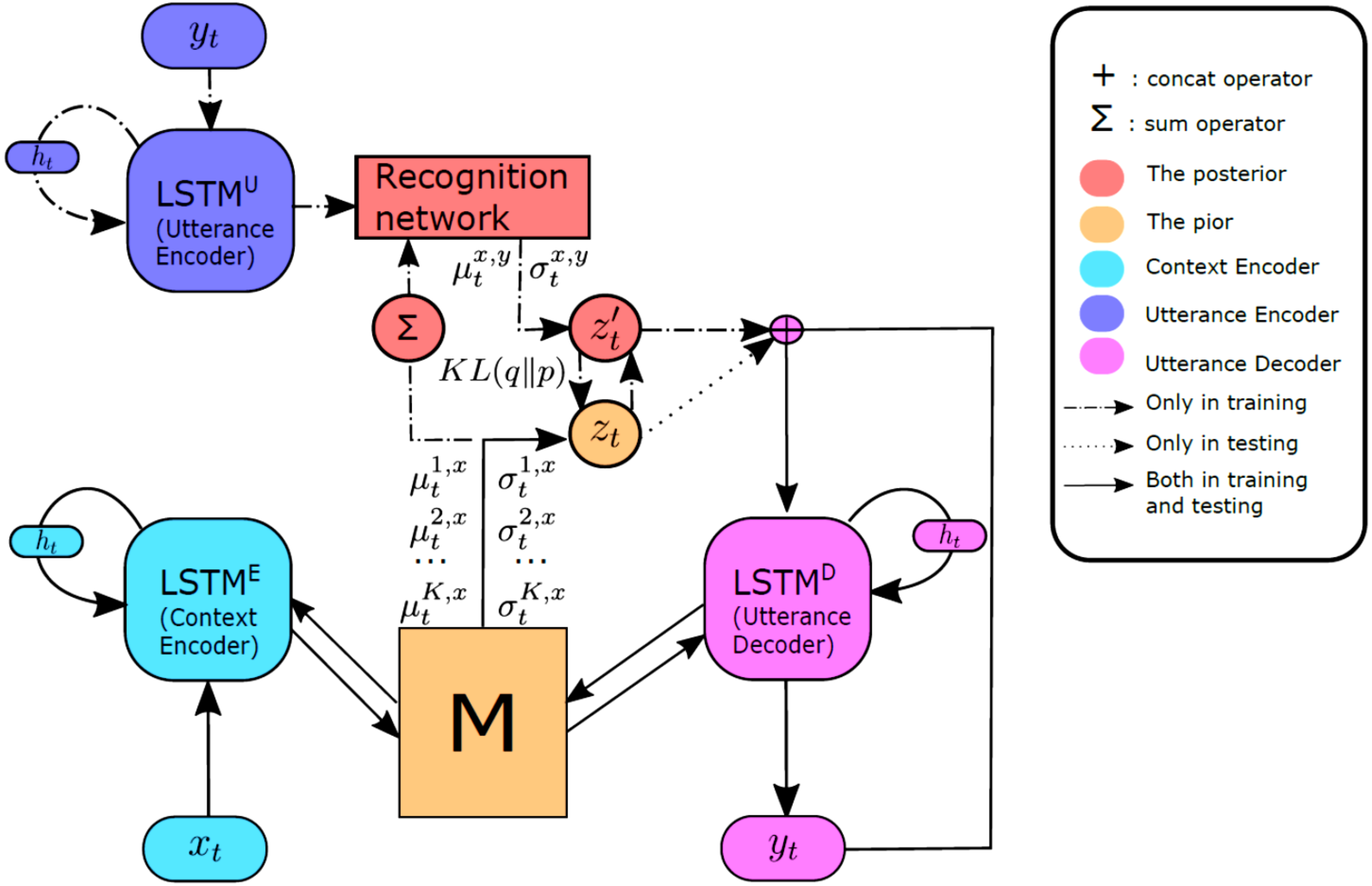
Variational memory encoder-decoder (VMED)



Conditional Variational Auto-Encoder



VMED


y_t
h_t
LSTM^U (Utterance Encoder)
Recognition network
$\mu_t^{x,y}$
$\sigma_t^{x,y}$
Σ
z'_t
$KL(q\|p)$
z_t
$\mu_t^{1,x}$ $\mu_t^{2,x}$ ... $\mu_t^{K,x}$
$\sigma_t^{1,x}$ $\sigma_t^{2,x}$ ... $\sigma_t^{K,x}$
h_t
LSTM^E (Context Encoder)
M
LSTM^D (Utterance Decoder)
h_t
x_t
y_t
+ : concat operator
Σ : sum operator
The posterior
The pior
Context Encoder
Utterance Encoder
Utterance Decoder
Only in training
Only in testing
Both in training and testing

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

Sample response (2)

LJ user: Your very pretty and your is very cute. May i add ya to my friends list ?	**Seq2Seq:** I! Add and too you back **Seq2Seq-att:** I did may! Away back **DNC:** Sure you added **CVAE:** Sure ! I can add you /*/ Thank cool of surely /*/ Thank hi of sure! **VMED (K=3):** Thank cool for sure /*/ Add definitely! Sure surely friends /*/ Thank girl go ahead
Cornell movies: Nobody can murder someone with their feelings . I mean that is crazy ?	**Seq2Seq:** Yes you are **Seq2Seq-att:** Really not is it **DNC:** Managing the truth **CVAE:** Sure not to be in here /*/ Oh yes but i know! /*/ That to doing with here and **VMED (K=4):** You are right! /*/ That is crazy /*/ You can't know Jimmy
OpenSubtitle: I'm obliged for your hospitality. I appreciate it and your husband too.	**Seq2Seq:** That is have got coming about these **Seq2Seq-att:** May you not what nothing nobody **DNC:** Yes i am taking this **CVAE:** No . that for good! And okay /*/ All in the of two thing /*/ Sure. Is this! **VMED (K=3):** I know. I can afford /*/ I know nothing to store for you pass /*/ I know. Doing anymore you father

Bringing relational structures into memory

Trang Pham, Truyen Tran & Svetha Venkatesh

ICPR'18+

NTM as matrix machine

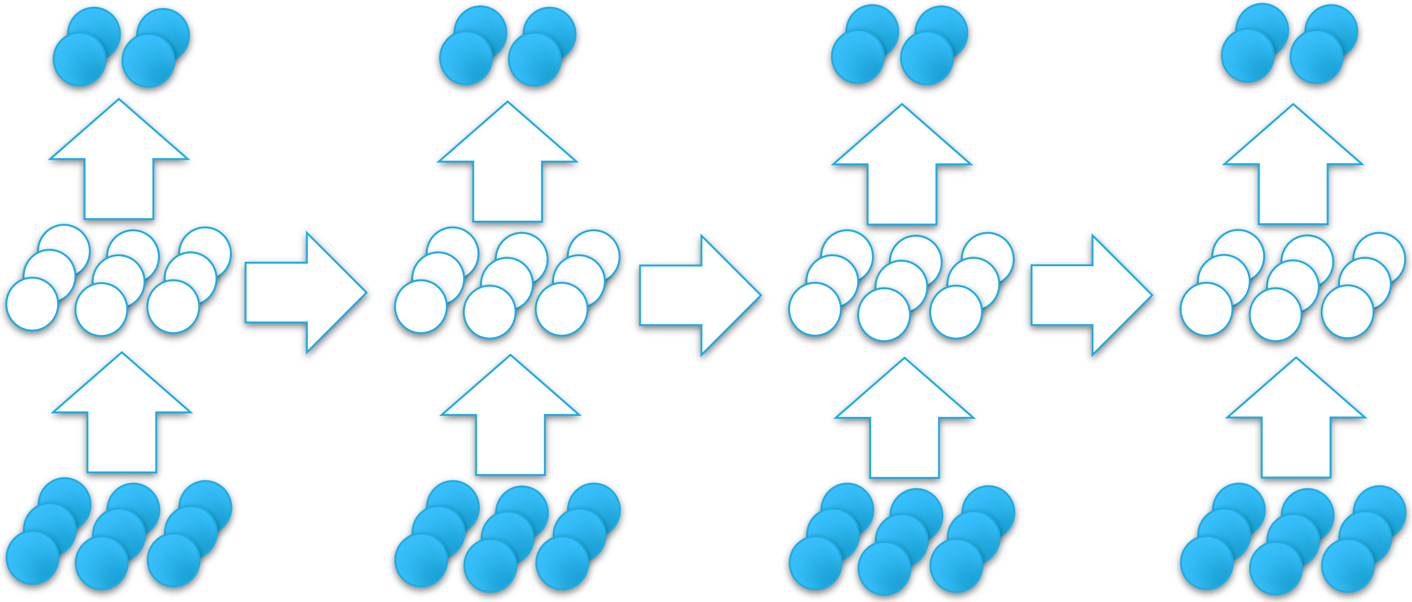
Controller and memory operations can be conceptualized as matrix operations

- **Controller is a vector changing over time**
- **Memory is a matrix changing over time**

#REF: Kien Do, Truyen Tran, Svetha Venkatesh, "Learning Deep Matrix Representations", *arXiv preprint arXiv:*1703.01454

Recurrent dynamics

$$\boxed{H_t = \sigma(U_x^\intercal X_t V_x + U_h^\intercal H_{t-1} V_h + B)}$$



Idea: Relational memory

Independent memory slots not suitable for relational reasoning

Human working memory sub-processes seem inter-dependent

Transformation

Relational structure

$$\boxed{H_t = \sigma(U_x^\intercal X_t V_x + U_h^\intercal H_{t-1} V_h + B)}$$

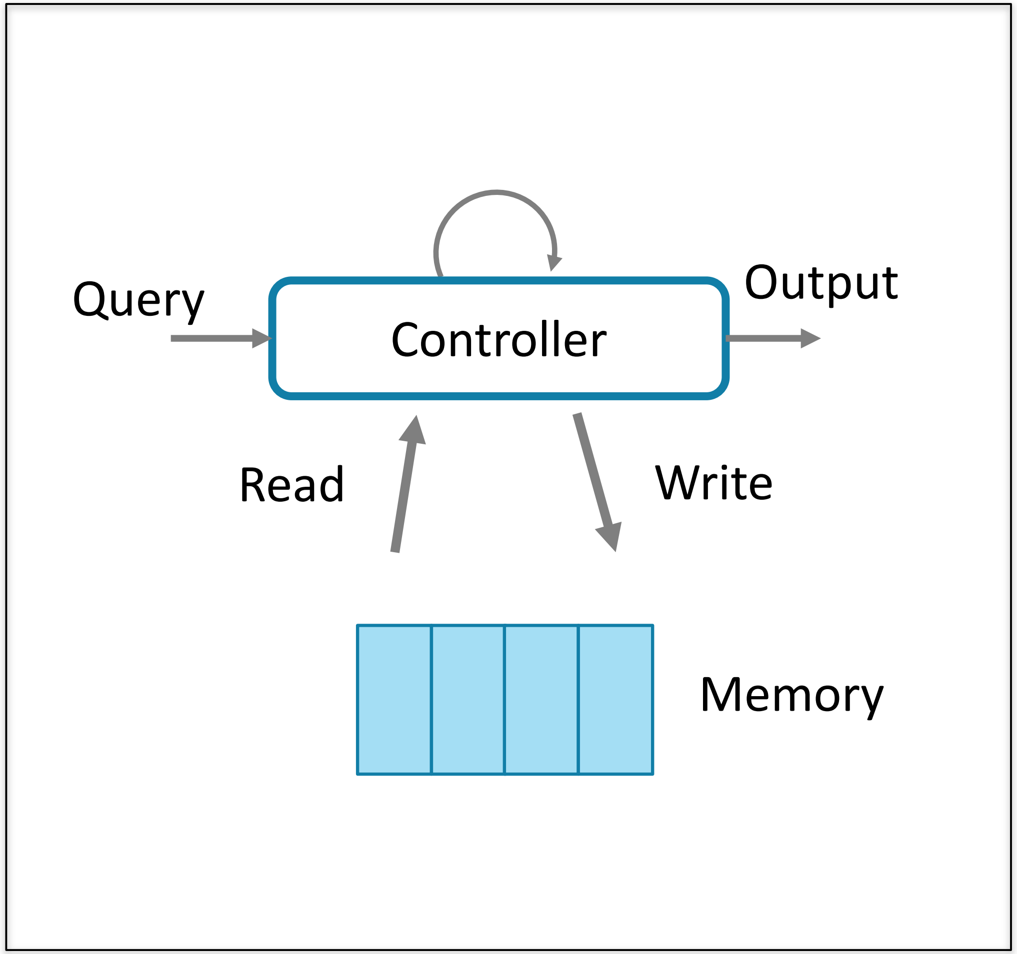
New memory proposal

New information

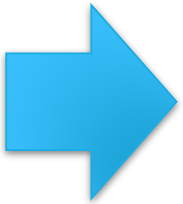
Old memory

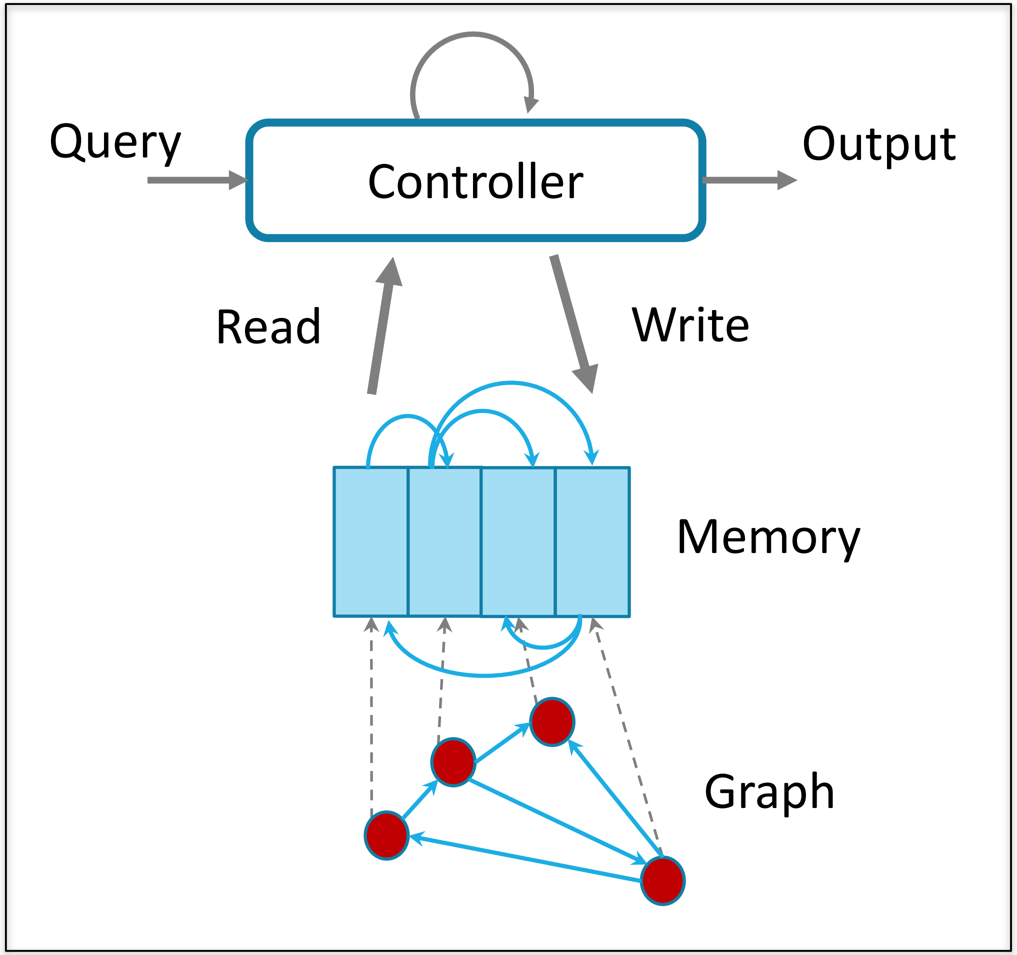
Time-aware bias

Relational Dynamic Memory Network (DMNN)



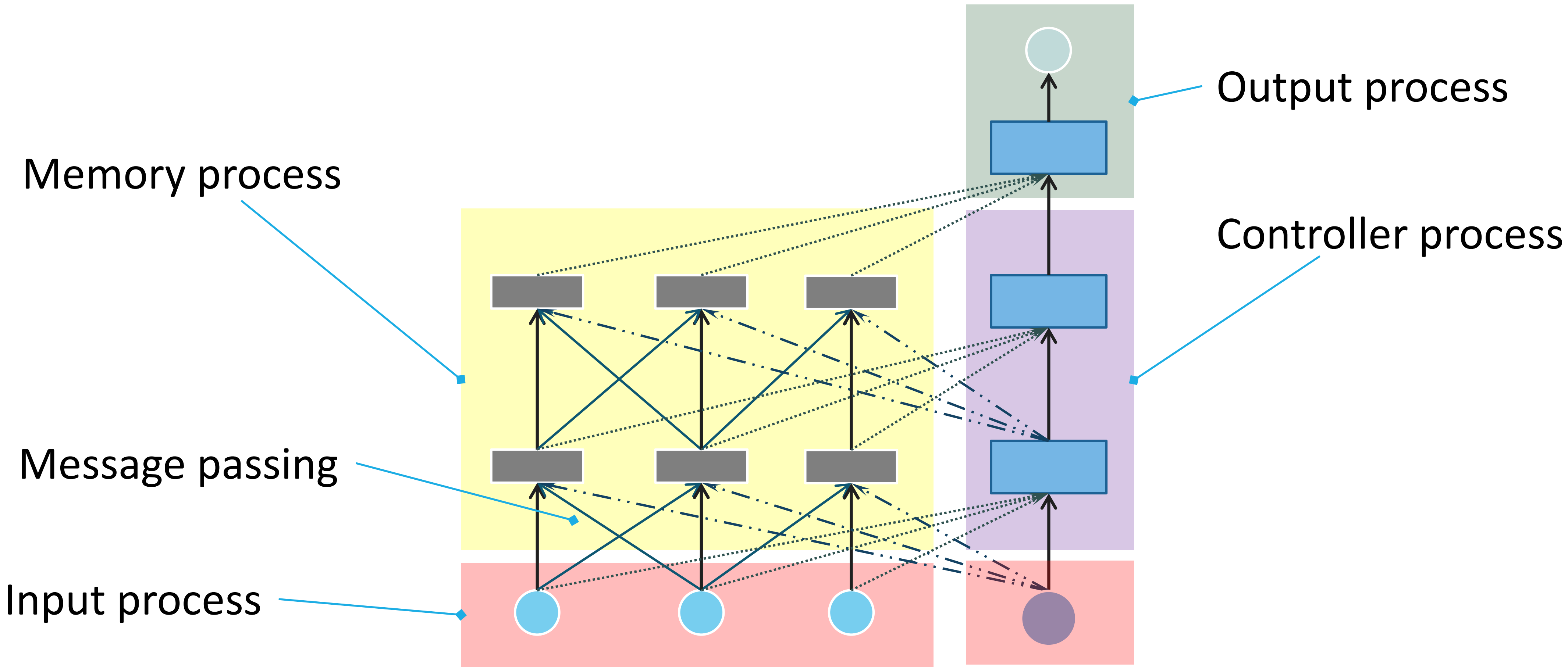
NTM





Relational Dynamic Memory Network

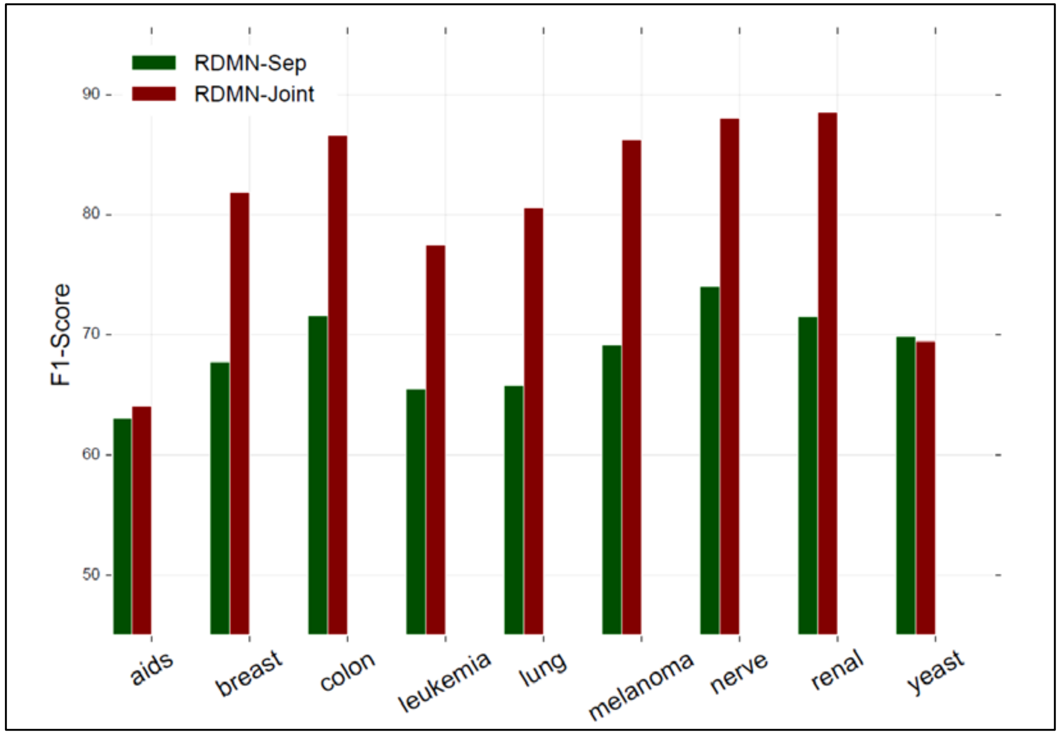
RDMN unrolled

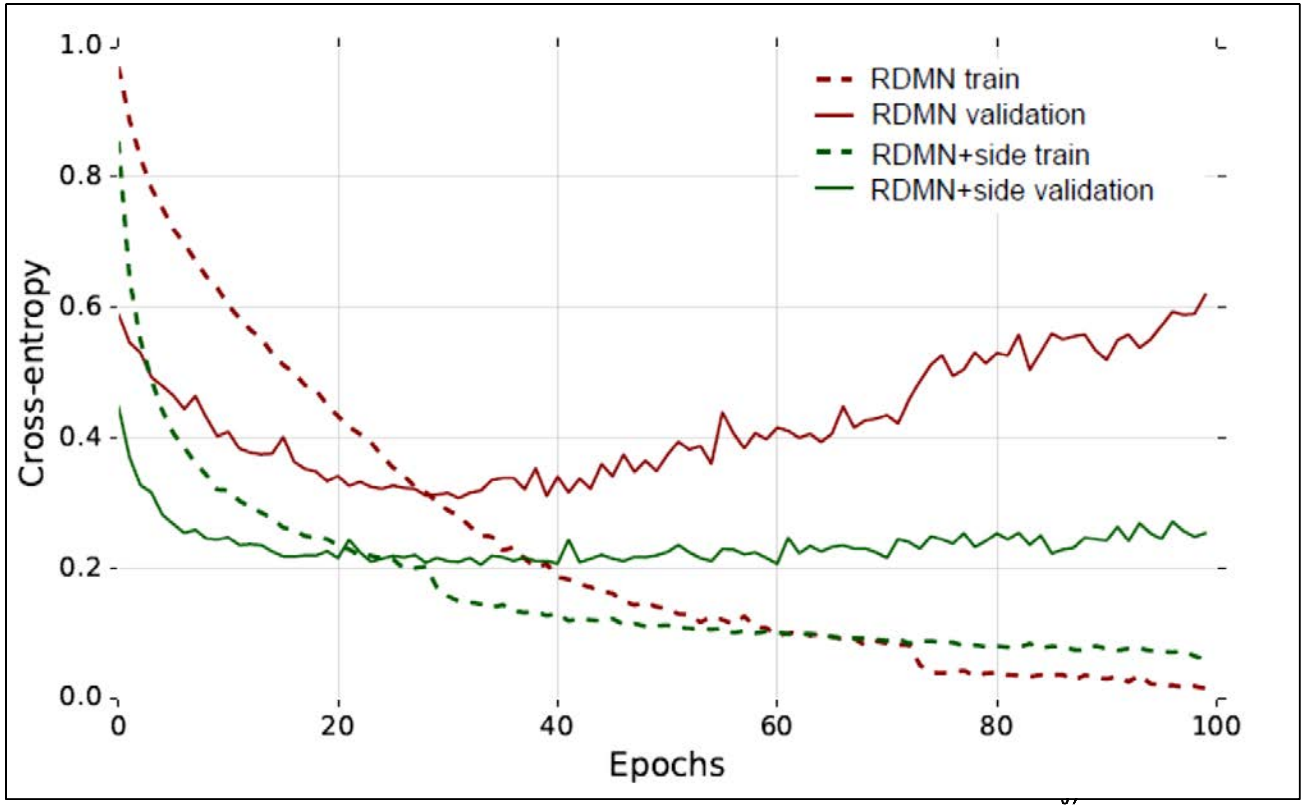


Drug-disease response

Molecule → Bioactivity

Model	MicroF1	MacroF1	Average AUC
SVM	66.4	67.9	85.1
RF	65.6	66.4	84.7
GB	65.8	66.9	83.7
NeuralFP [19]	68.2	67.6	85.9
MT-NN [51]	75.5	78.6	90.4
RDMN	77.8	80.3	92.1



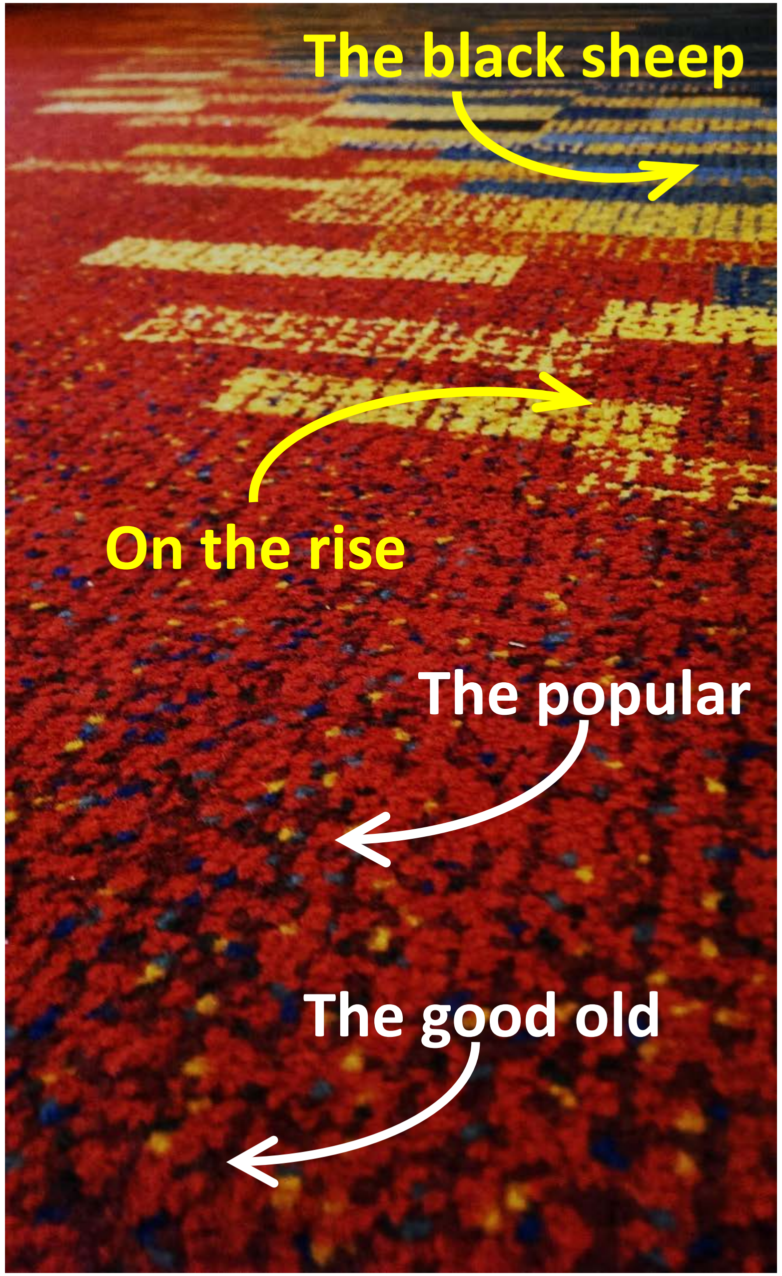


Chemical reaction

Molecules → Reaction

	CCI900		CCI800	
	AUC	F1-score	AUC	F1-score
Random Forests	94.3	86.4	98.2	94.1
Highway Networks	94.7	88.4	98.5	94.7
DeepCCI [38]	96.5	92.2	99.1	97.3
RDMN	96.6	92.6	99.1	97.4
RDMN+multiAtt	97.3	93.4	99.1	97.8
RDMN+FP	97.8	93.3	99.4	98.0
RDMN+multiAtt+FP	98.0	94.1	99.5	98.1
RDMN+SMILES	98.1	94.3	99.7	97.8
RDMN+multiAtt+SMILES	**98.1**	**94.6**	**99.8**	**98.3**

Looking ahead



- **Cognitive architecture | Unified Theory of Cognition**
- Quantum ML/AI
- Theory of consciousness (e.g., Penrose's microtubes)
- Value-aligned ML

- Reinforcement learning, imagination & planning
- Deep generative models + Bayesian methods
- **Memory & reasoning**
- Lifelong/meta/continual/few-shot/zero-shot learning
- Universal transformer

- Attention
- Batch-norm
- ReLU & skip-connections
- Highway nets, LSTM/GRU & CNN

- Representation learning (RBM, DBN, DBM, DDAE)
- Ensemble
- Back-propagation
- Adaptive stochastic gradient

Better memory theory

Sparse writing

Explaining memory operations

Dynamic memory structure (other than a fixed-size matrix)

- E.g., Differentiable pooling (NIPS’18)

Loading long-term/episodic mem into working mem

- walk(man, dog; day1); walk(woman, dog; day2) → couple(man, woman)

A grand unified theory of memory?

- May be **Free-Energy Principle** by Karl Friston

Intelligence as emergence?

- We are just little bit better than apes in each intelligence dimension, but far more intelligent overall.

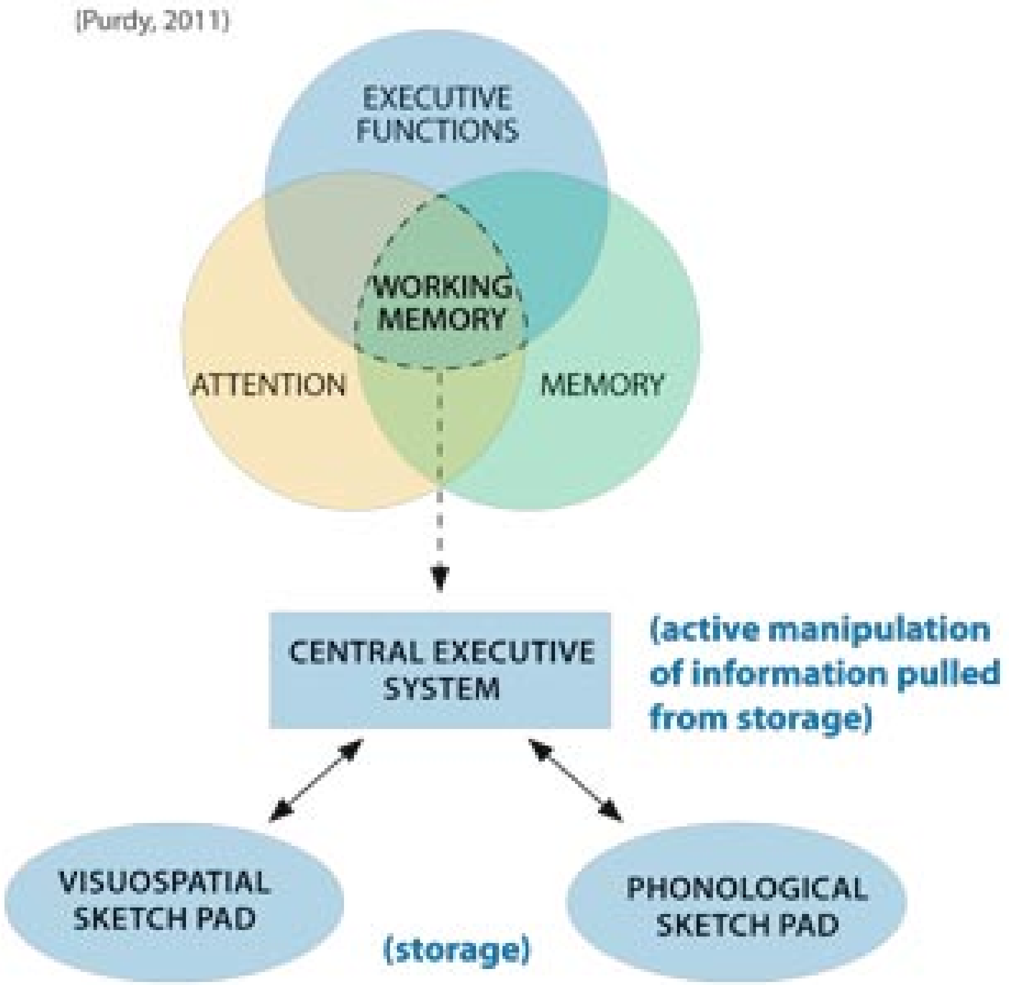
Memory types

Short-term/working (temporary storage)

Episodic (events happened at specific time)

Long-term/semantic (facts, objects, relations)

Procedural (sequence of actions)



http://www.rainbowrehab.com/executive-functioning/

Applications of memory

Rare events

Video captioning

QA, VQA

Machine translation

Machine reading (stories, books, DNA)

Business process continuation

Software execution

Code generation

Graph as sequence of edges

Event sequences

Graph traversal

Algorithm learning (e.g., sort)

Dialog systems (e.g., chat bots)

Reinforcement learning agents

Multi-agents with shared memory

Learning to optimize

Memory-supported intelligence

Reasoning with working memory (NTM style)

Meta-learning with episodic memory

Meta-remembering of memory operations

Learning to plan with procedural memory

Learning world knowledge with semantic memory

Learning to navigate with spatial memory

Learning to socialize with collective memory and memory of others (???)

Toward a full cognitive architecture

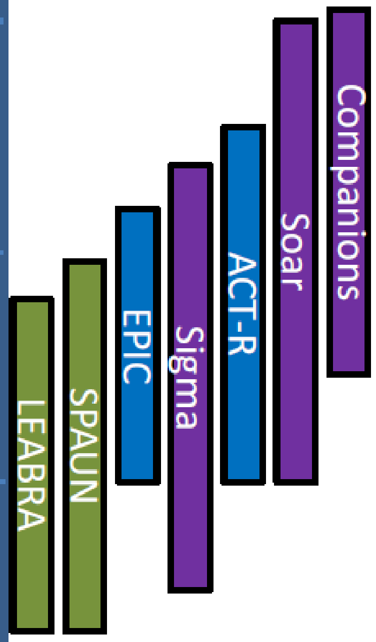
Cognitive Architecture

- Fixed computational structures underlying intelligent behavior
  - Representations of knowledge
  - Memories that hold knowledge
  - Processors that manipulate knowledge
- Supports end-to-end behavior
  - Includes integration with perception and action
- General across tasks
  - Architectural mechanisms are reused across every task and subtask
  - Task-specific knowledge guides task behavior
- Complete
  - No "escape" to additional specialized programming

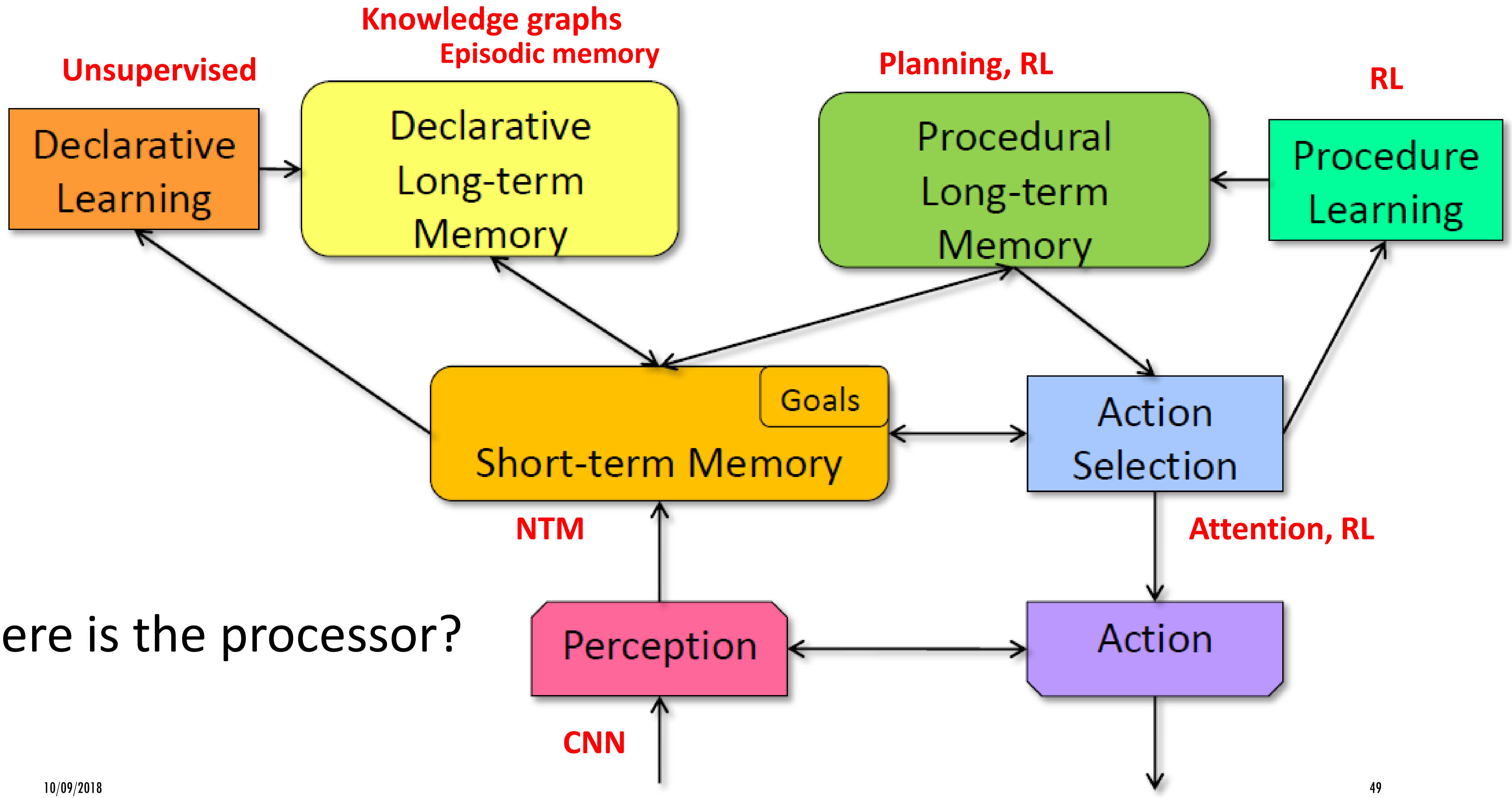
#REF: A Cognitive Architecture Approach to Interactive Task Learning, John E. Laird, University of Michigan

Newell's Time Scale of Human Action

Scale (sec)	Time Units	System	Band
10^7	months		
10^6	weeks		Social
10^5	days		
10^4	hours	Task	
10^3	10 min	Task	Rational
10^2	minutes	Task	
10^1	10 sec	Unit task	
10^0	1 sec	Operations	Cognitive
10^{-1}	100 ms	Deliberate act	
10^{-2}	10 ms	Neural Circuit	
10^{-3}	1 ms	Neuron	Biological
10^{-4}	100 µs	Organelle	

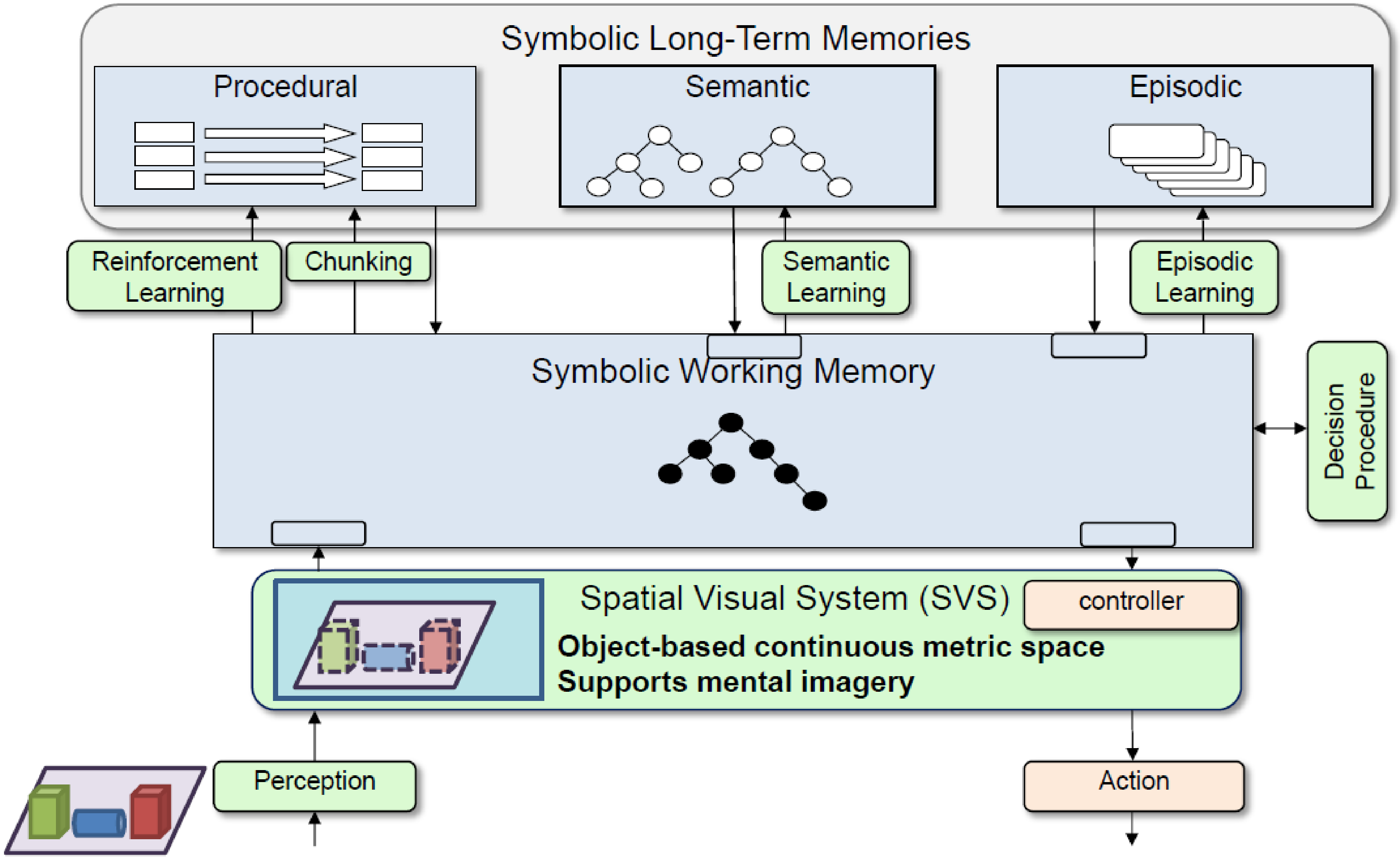


#REF: A Cognitive Architecture Approach to Interactive Task Learning, John E. Laird, University of Michigan

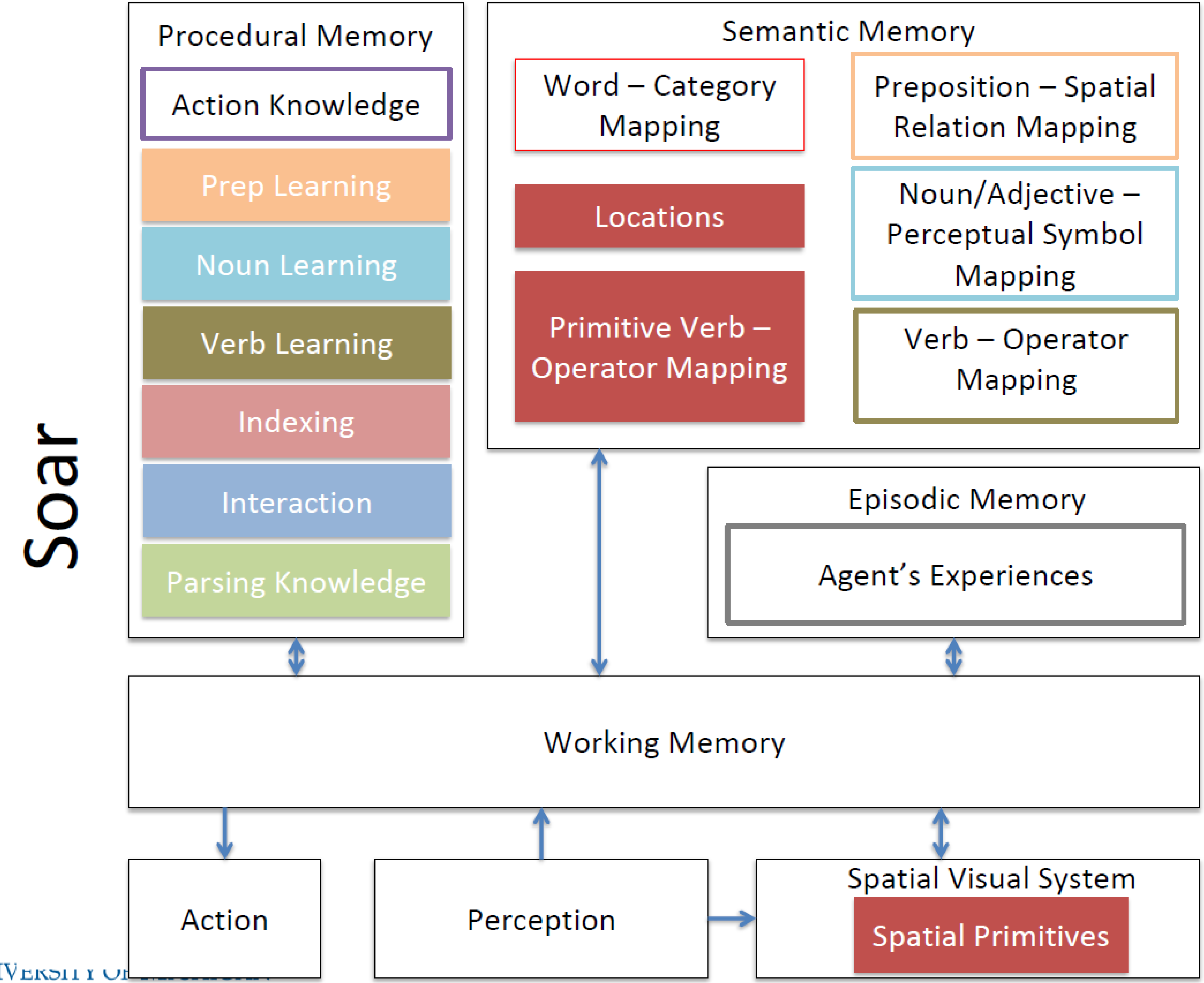


Where is the processor?

Soar Structure



#REF: A Cognitive Architecture Approach to Interactive Task Learning, John E. Laird, University of Michigan



Team @ Deakin (A2I2)





Thanks to many people who have created beautiful graphics & open-source programming frameworks.

References

Memory–Augmented Neural Networks for Predictive Process Analytics, A Khan, H Le, K Do, T Tran, A Ghose, H Dam, R Sindhgatta, *arXiv preprint* arXiv:1802.00938

Learning deep matrix representations, K Do, T Tran, S Venkatesh, *arXiv preprint* arXiv:1703.01454

Variational memory encoder-decoder, H Le, T Tran, T Nguyen, S Venkatesh, *arXiv preprint* arXiv:1807.09950

Relational dynamic memory networks, Trang Pham, Truyen Tran, Svetha Venkatesh, *arXiv preprint* arXiv:1808.04247

Dual Memory Neural Computer for Asynchronous Two-view Sequential Learning, H Le, T Tran, S Venkatesh, *KDD'18*

Dual control memory augmented neural networks for treatment recommendations, H Le, T Tran, S Venkatesh, *PAKDD'18*.