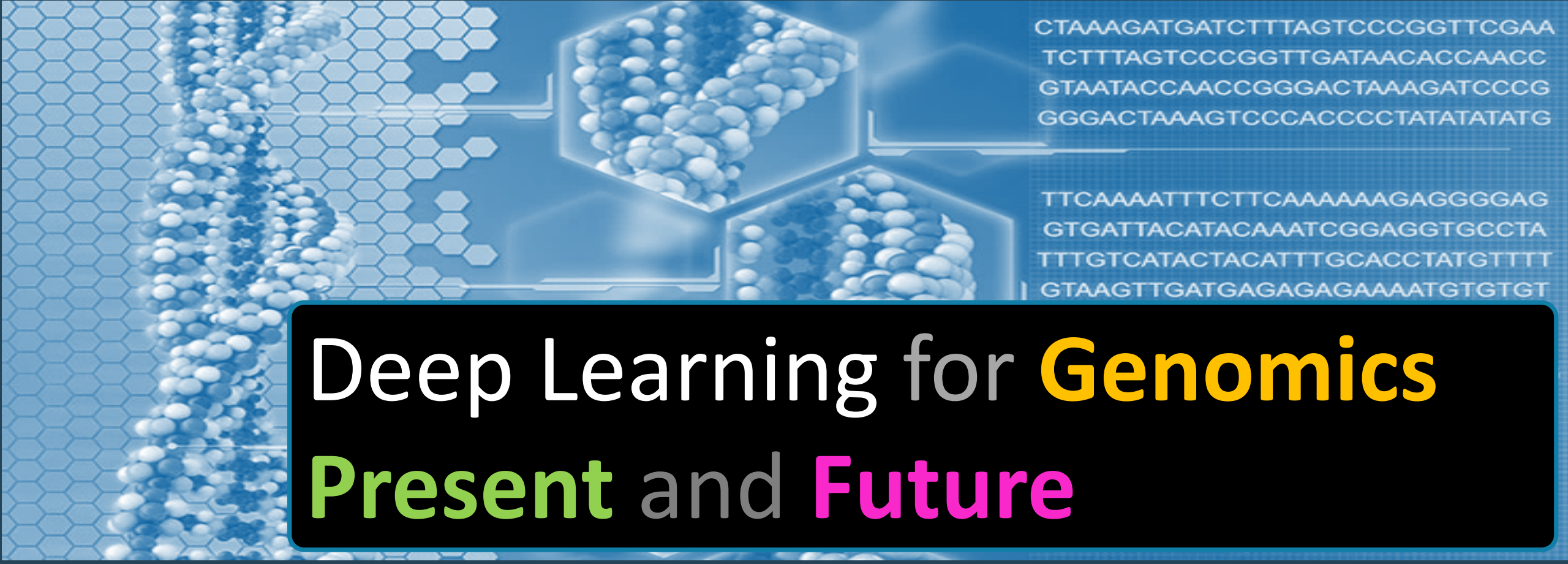
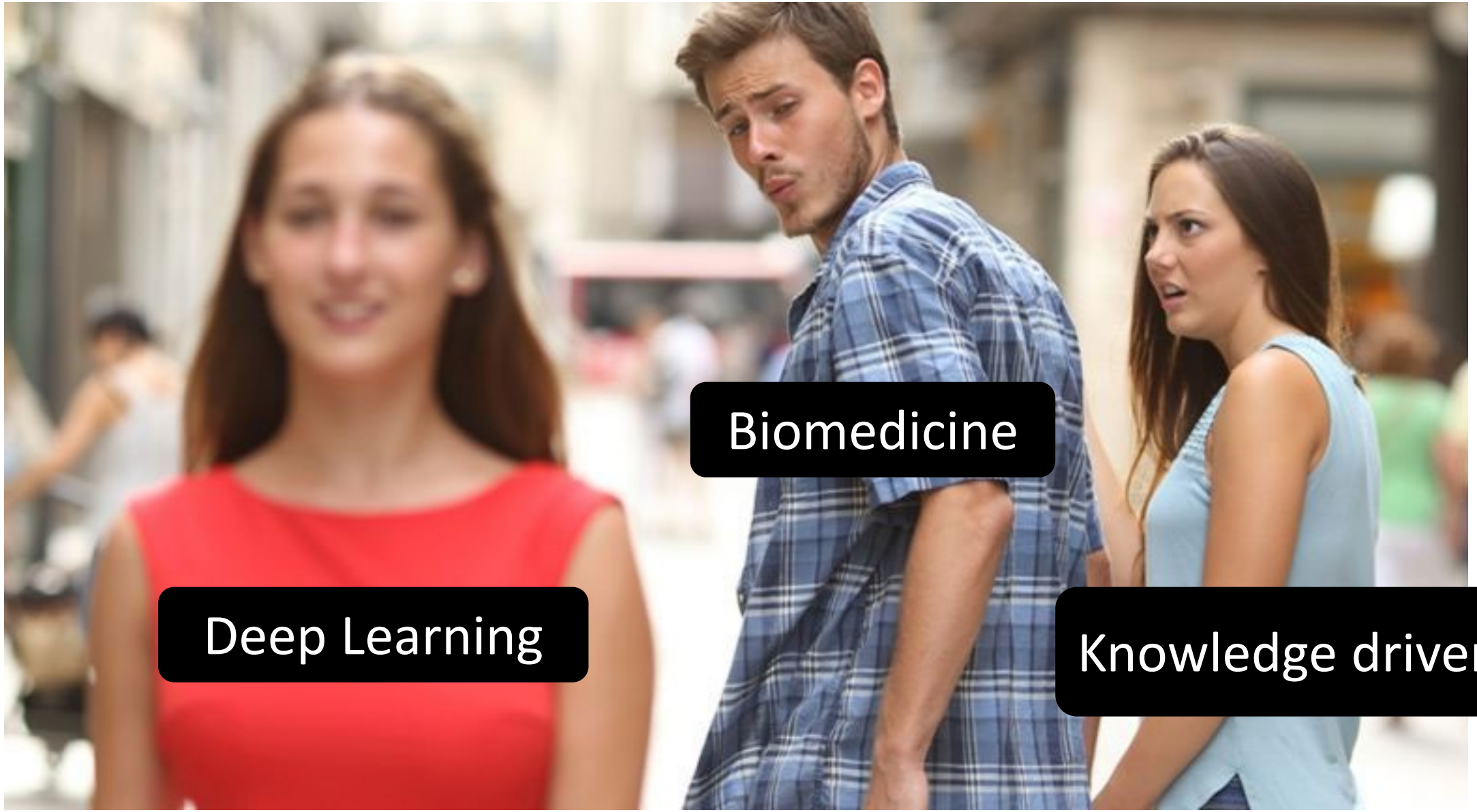


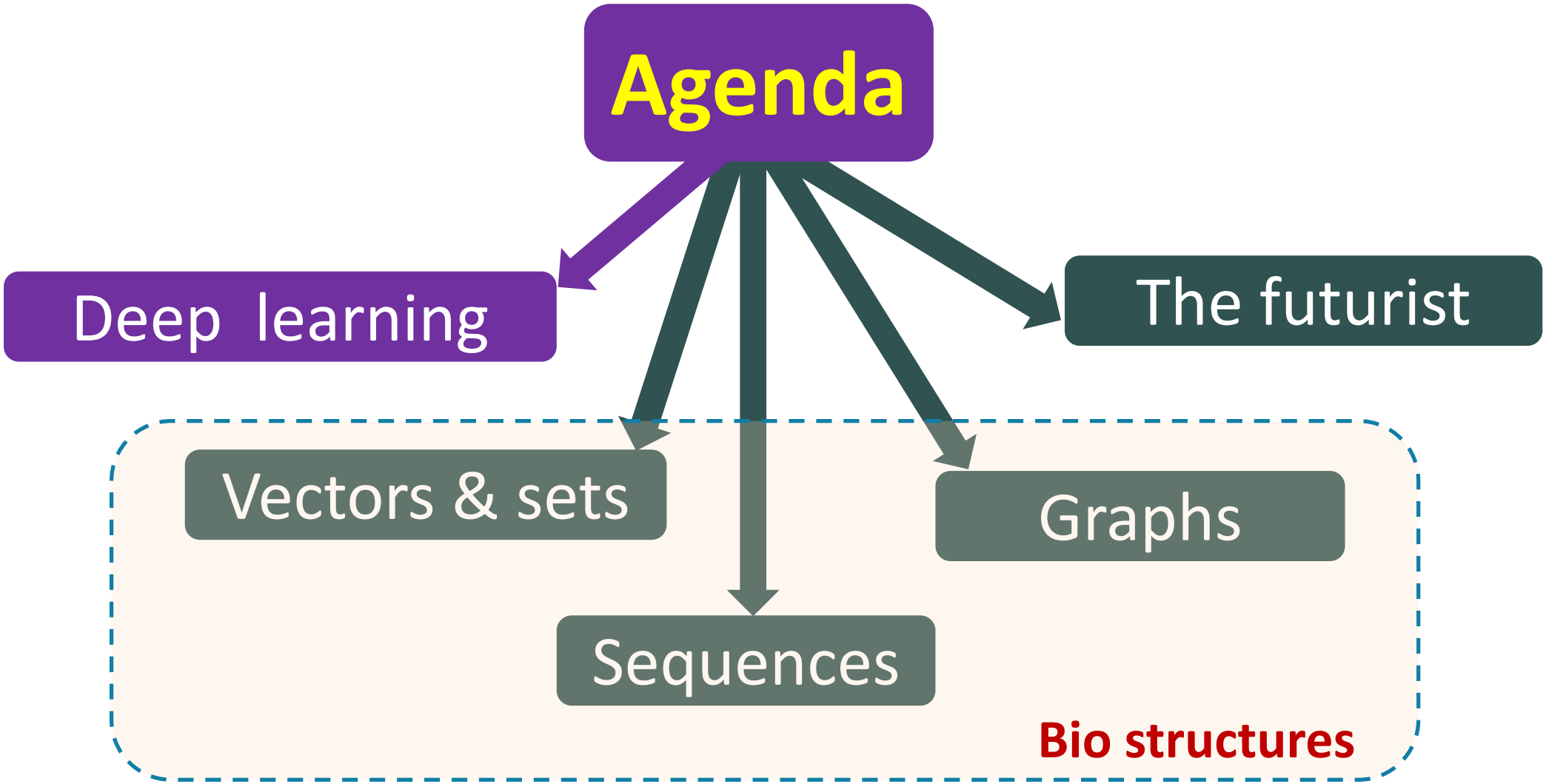
Deep Learning for **Genomics**
Present and **Future**

Truyen Tran
Deakin University

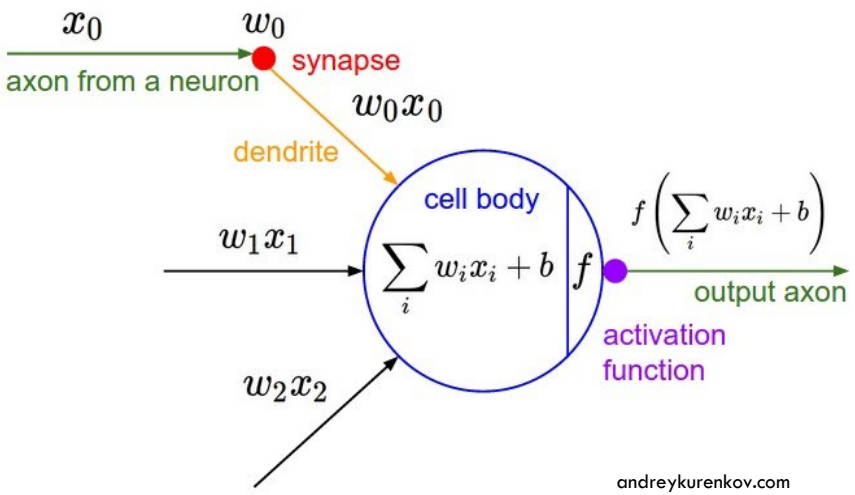
truyen.tran@deakin.edu.au
truyentran.github.io
@truyenoz
letdataspeak.blogspot.com
goo.gl/3jJ1O0

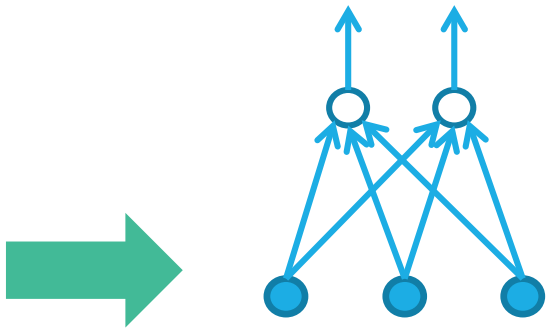
Hanoi, June 2019


Deep Learning
Biomedicine
Knowledge driven

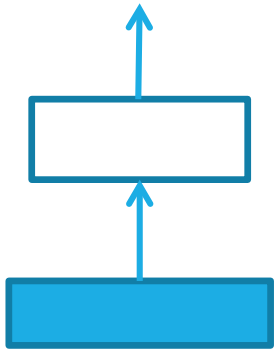

Agenda
Deep learning
The futurist
Vectors & sets
Graphs
Sequences
Bio structures

Integrate-and-fire neuron





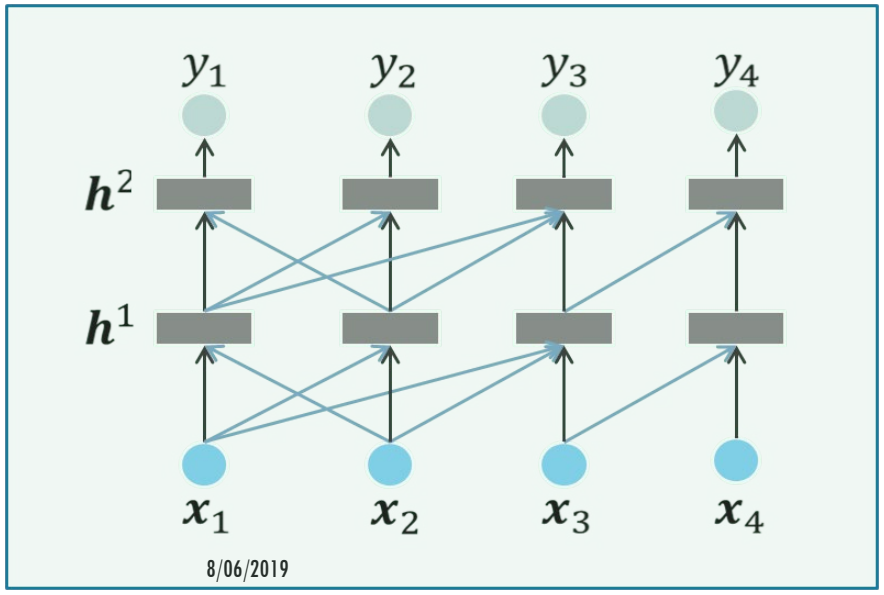
Feature detector



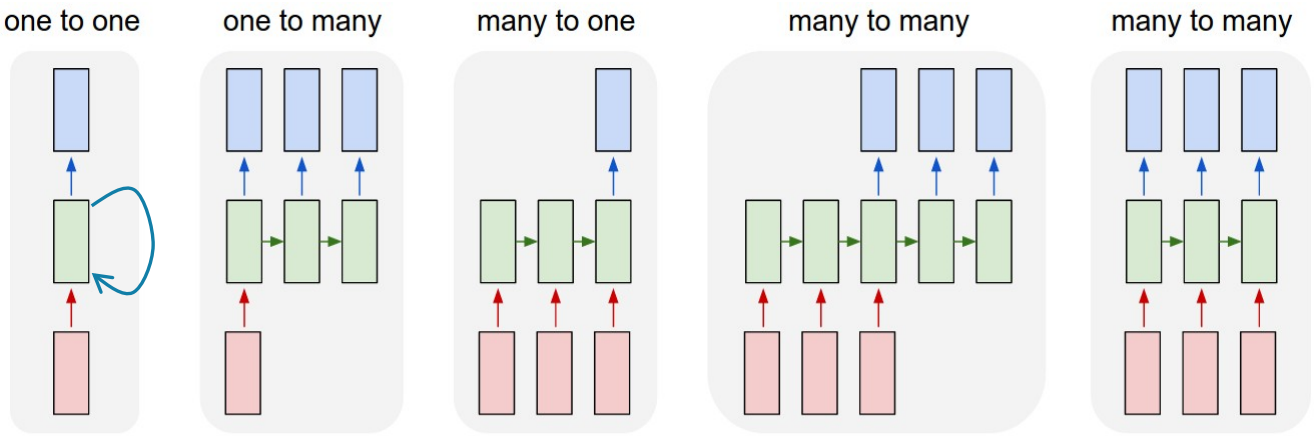
Block representation

DL in a nutshell

Relations, messaging & attention

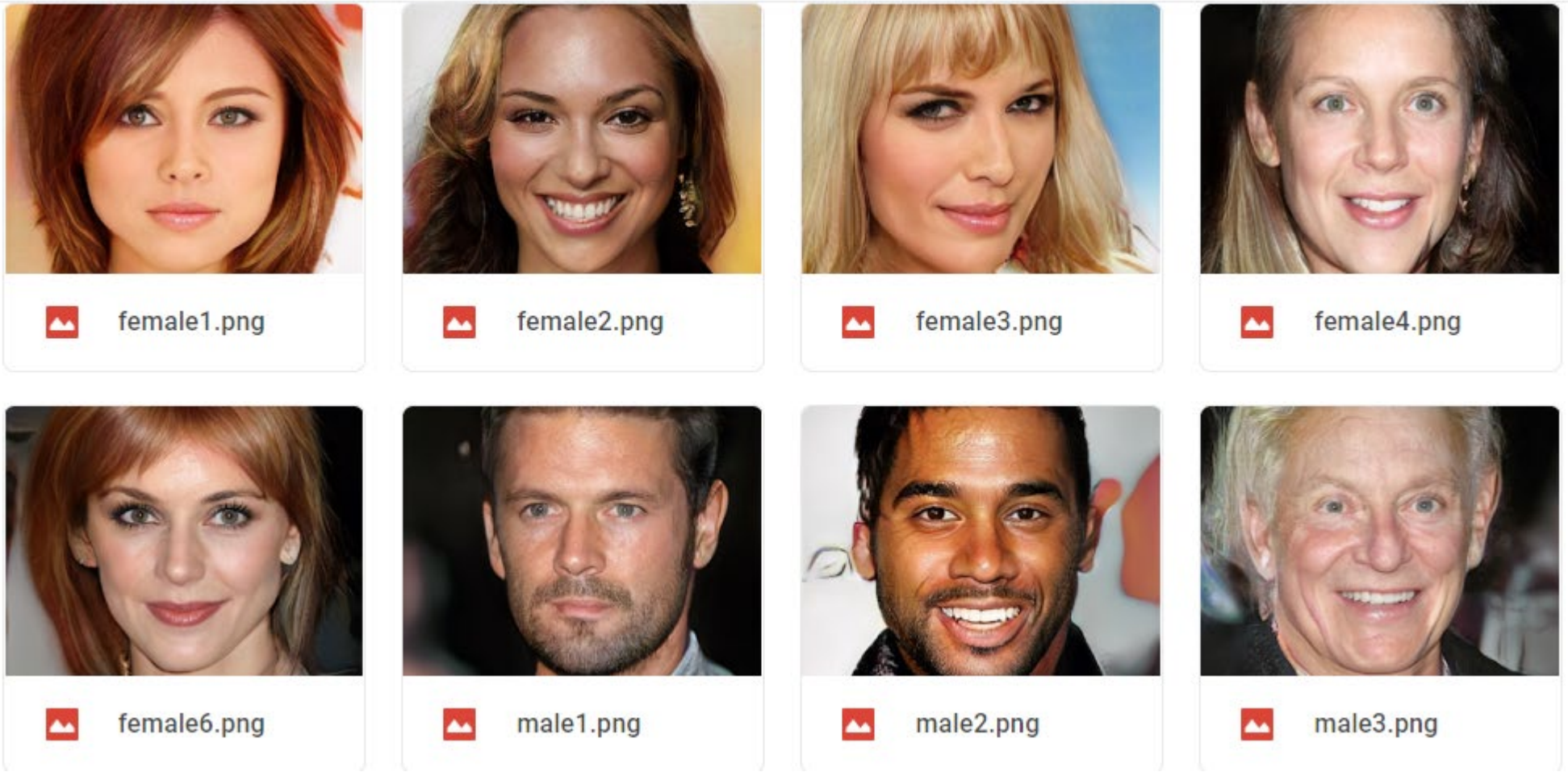


"Deep" as layered blocks



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

DL can learn from data, and fake it



Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

DL can generate sequences nicely

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

GPT-2, https://openai.com/blog/better-language-models/#fn2

What can DL do to genomics?

Deep learning offerings

Function approximation

Program approximation

Program synthesis

Deep density estimation

Disentangling factors of variation

Capturing data structures

Generating realistic data (sequences)

Question-answering

Information extraction

Knowledge graph construction and completion

Genomic problems

GWAS, gene-disease mapping

Binding site identification

Function prediction

Drug-target binding

Drug design

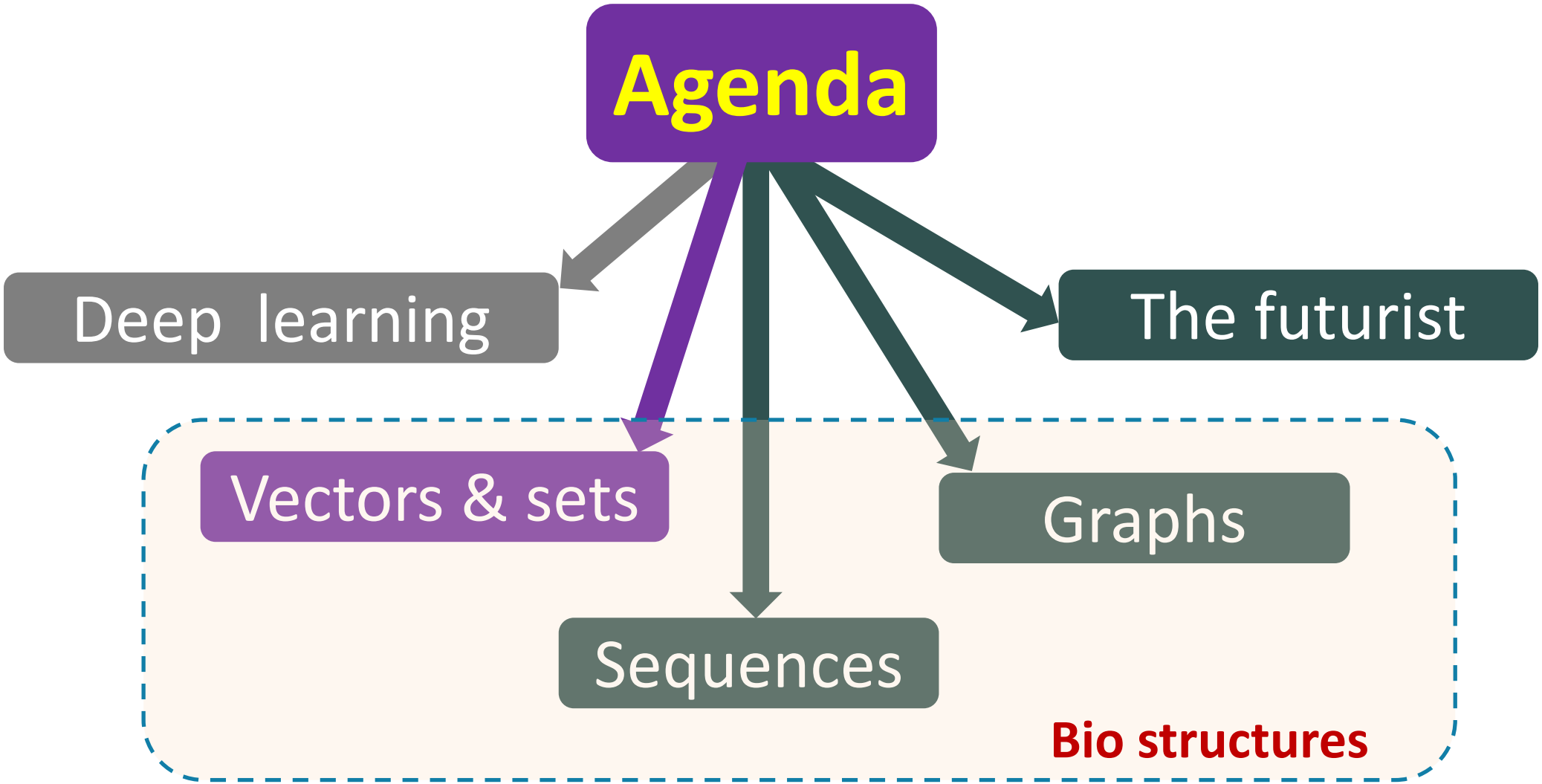
Structure prediction

Sequence generation

Functional genomics

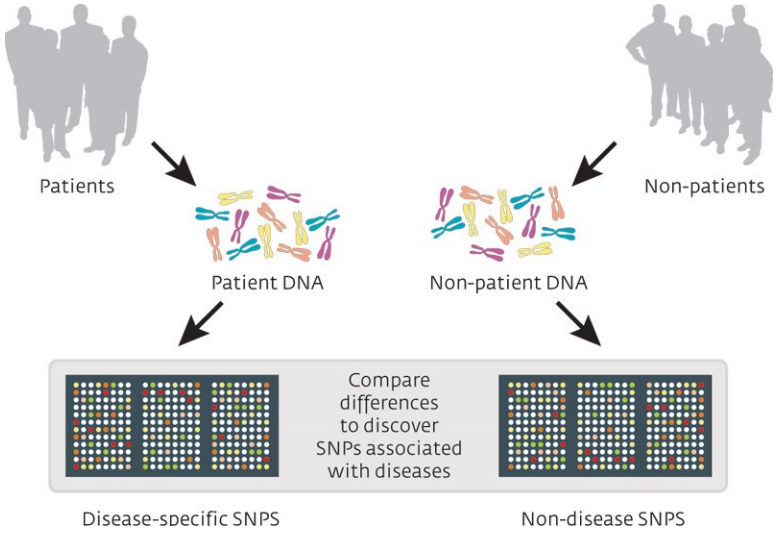
Optimizing sequences

Organizing the (knowledge about) omics universe


Agenda
Deep learning
The futurist
Vectors & sets
Graphs
Sequences
Bio structures

“Diet networks” for GWAS

#REF: Romero, Adriana, et al. "Diet Networks: Thin Parameters for Fat Genomic" *ICLR* (2017).



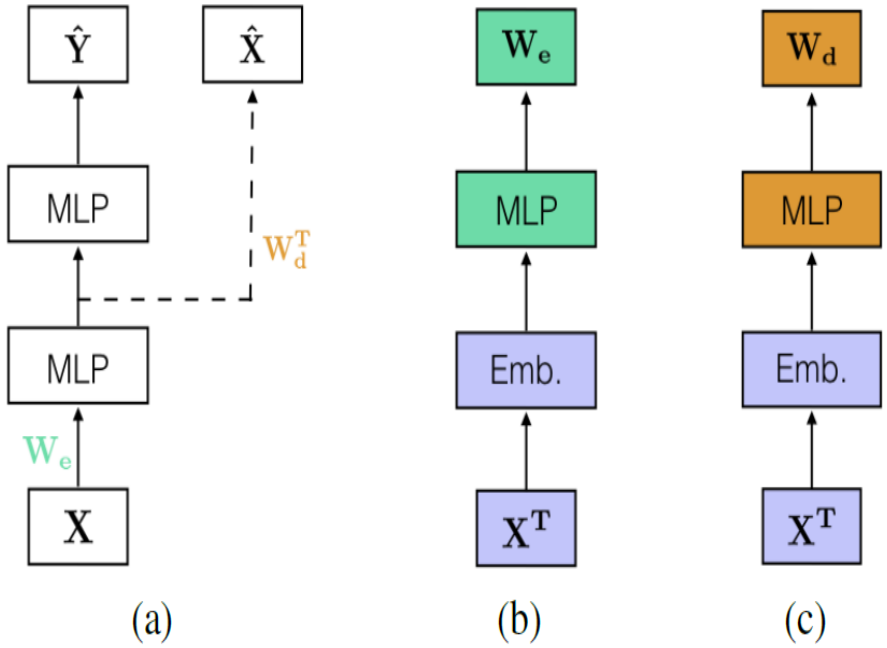
Use a “hypernet” to generate the main net.

Features are embedded (not data instance).

Unsupervised autoencoder as regularizer.

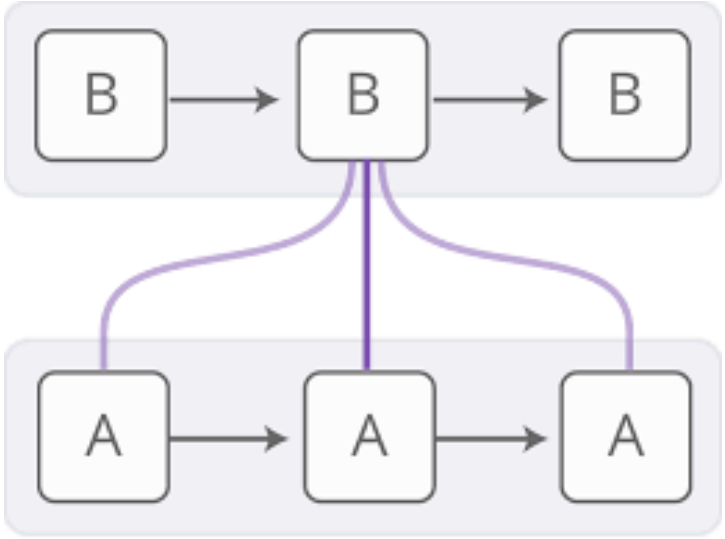
Works well on country prediction on the 1000 Genomes Project dataset.

- But this is a relatively easy problem. PCA, even random subspace can do quite well!



Images taken from the paper

Gene expression: DeepTRIAGE



http://distill.pub/2016/augmented-rnns/

Attention mechanism

DeepTRIAGE: Interpretable and Individualised Biomarker Scores using Attention Mechanism for the Classification of Breast Cancer Sub-types

Adham Beykikhoshk[1,*], Thomas P. Quinn[1,*], Samuel C. Lee[1], Truyen Tran[1], and Svetha Venkatesh[1]

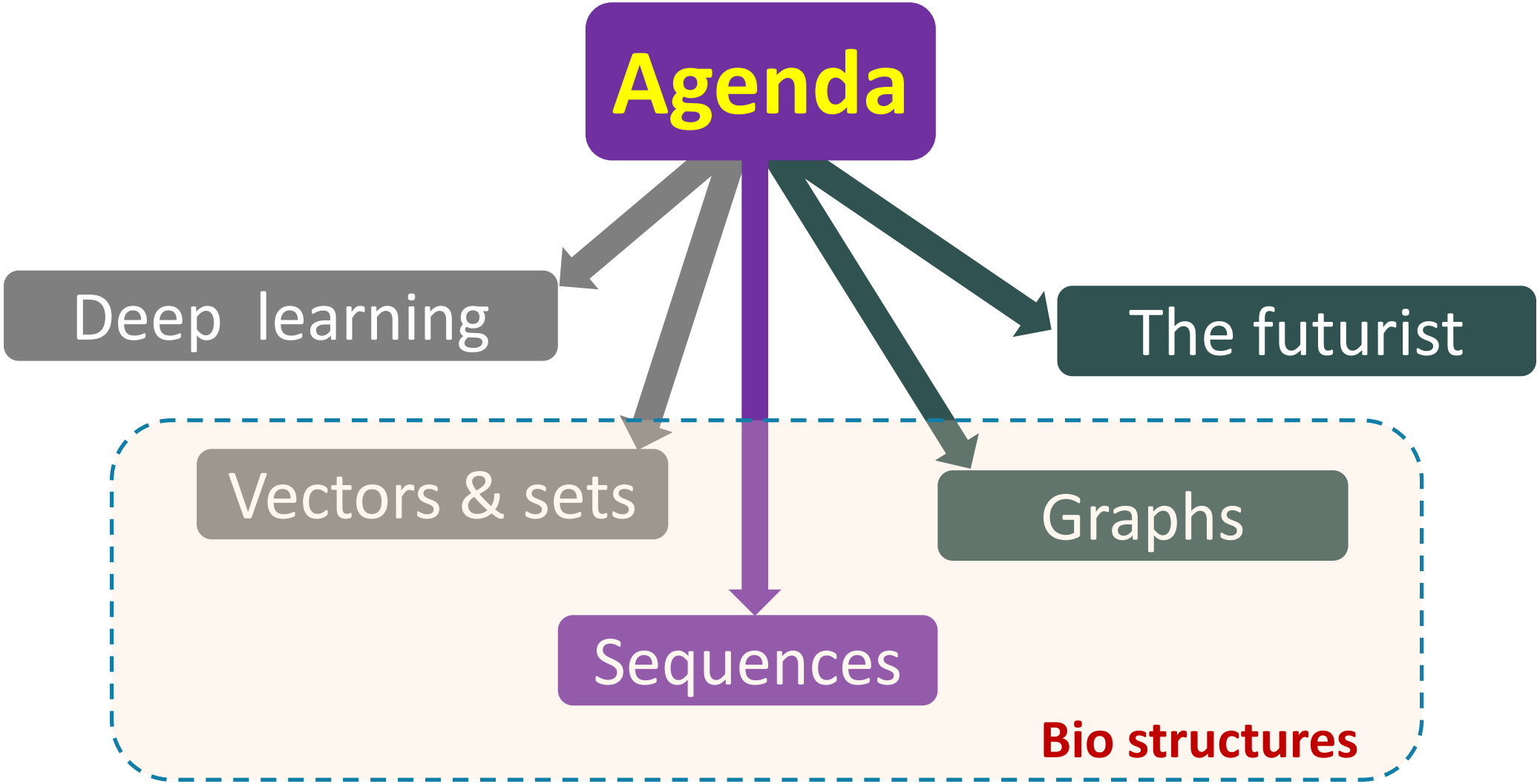
[1]Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, Australia
* adham.beyki@deakin.edu.au; contacttomquinn@gmail.com

Abstract

Motivation: Breast cancer is a collection of multiple tissue pathologies, each with a distinct molecular signature that correlates with patient prognosis and response to therapy. Accurately differentiating between breast cancer sub-types is an important part of clinical decision-making. Already, this problem has been addressed using machine learning methods that separate tissue samples into distinct groups. However, there remains unexplained heterogeneity within the established sub-types that cannot be resolved by the commonly used classification algorithms. In this paper, we propose a novel deep learning architecture, called DeepTRIAGE (Deep learning for the TRactable Individualised Analysis of Gene Expression), which not only classifies cancer sub-types with comparable accuracy, but simultaneously assigns each patient their own set of interpretable and individualised biomarker scores. These personalised scores describe how important each feature is in the classification of each patient, and can be analysed post-hoc to generate new hypotheses about intra-class heterogeneity.

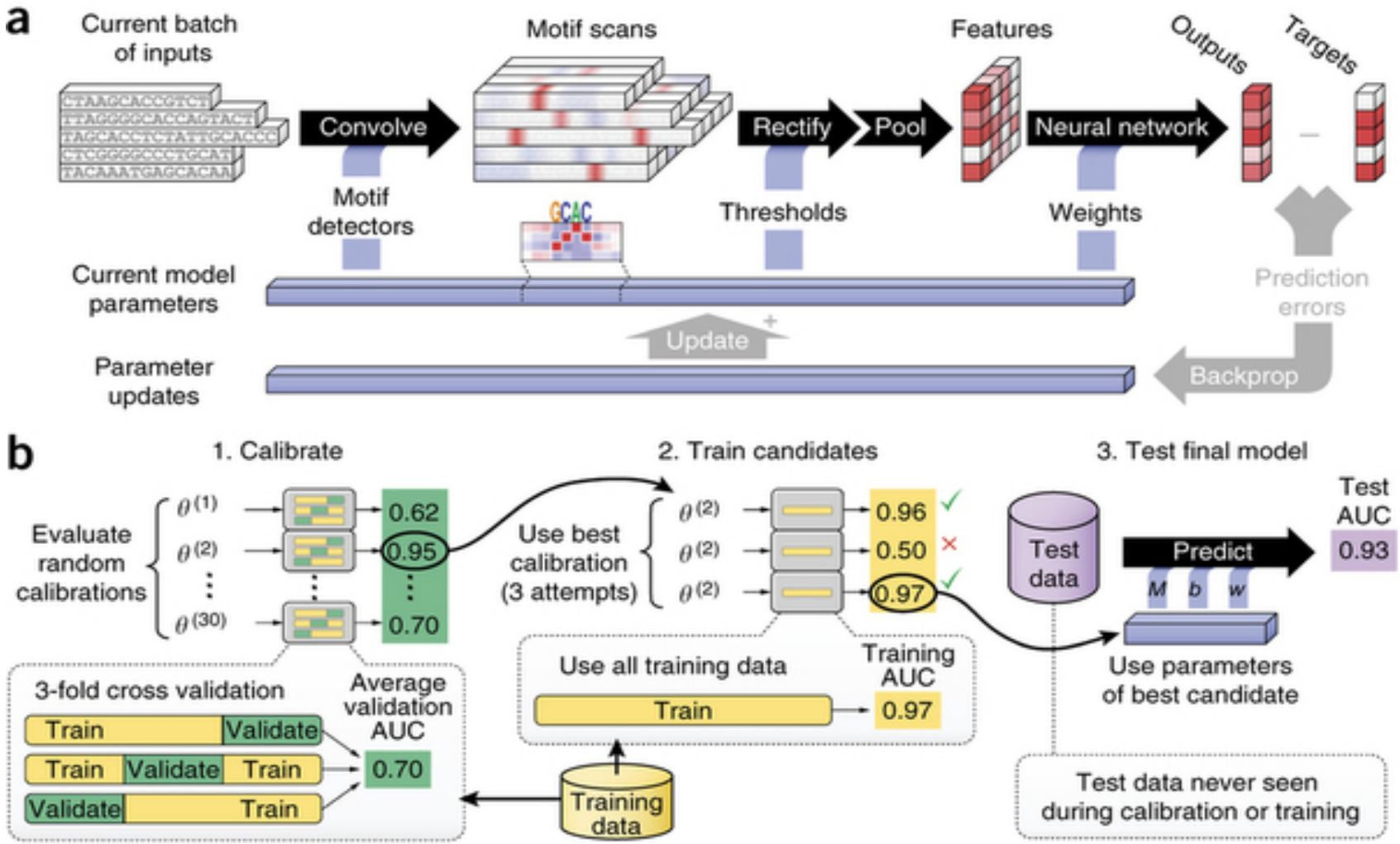
Results: We apply the DeepTRIAGE framework to classify the gene expression signatures of luminal A and luminal B breast cancer sub-types, and illustrate its use for genes and gene set (i.e., GO and KEGG) features. Using DeepTRIAGE, we find that the GINS1 gene and the kinetochore organisation GO term are the most important features for luminal sub-type classification. Through classification, DeepTRIAGE simultaneously reveals heterogeneity within the luminal A biomarker scores that significantly associate with tumour stage, placing all luminal samples along a continuum of severity.

Availability and implementation: The proposed model is implemented in Python using PyTorch framework. The analysis is done in Python and R. All Methods and models are freely available from https://github.com/adham/BiomarkerAttend.


Agenda
Deep learning
The futurist
Vectors & sets
Graphs
Sequences
Bio structures

Identifying binding sites

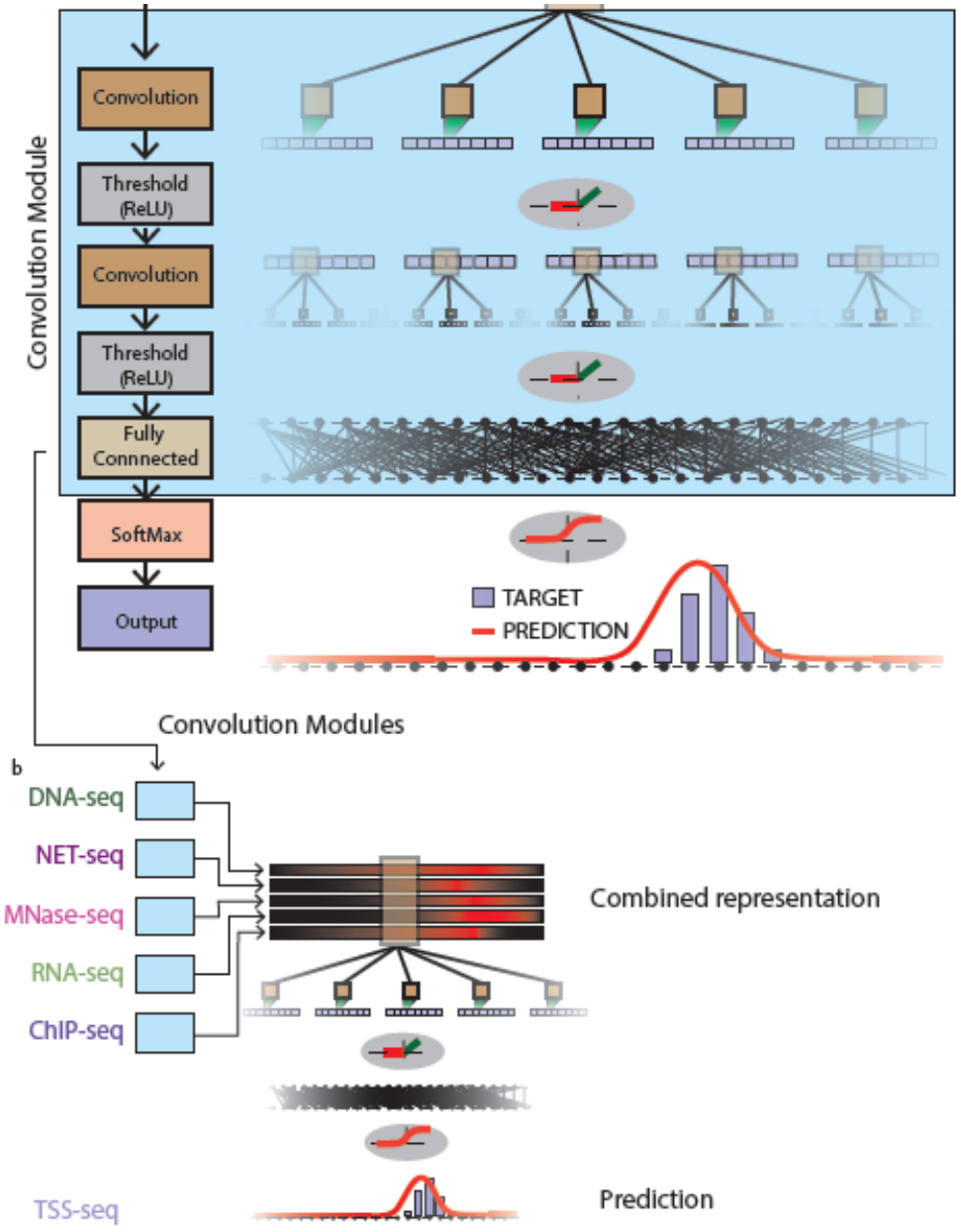
DeepBind (Alipanahi et al, Nature Biotech 2015)

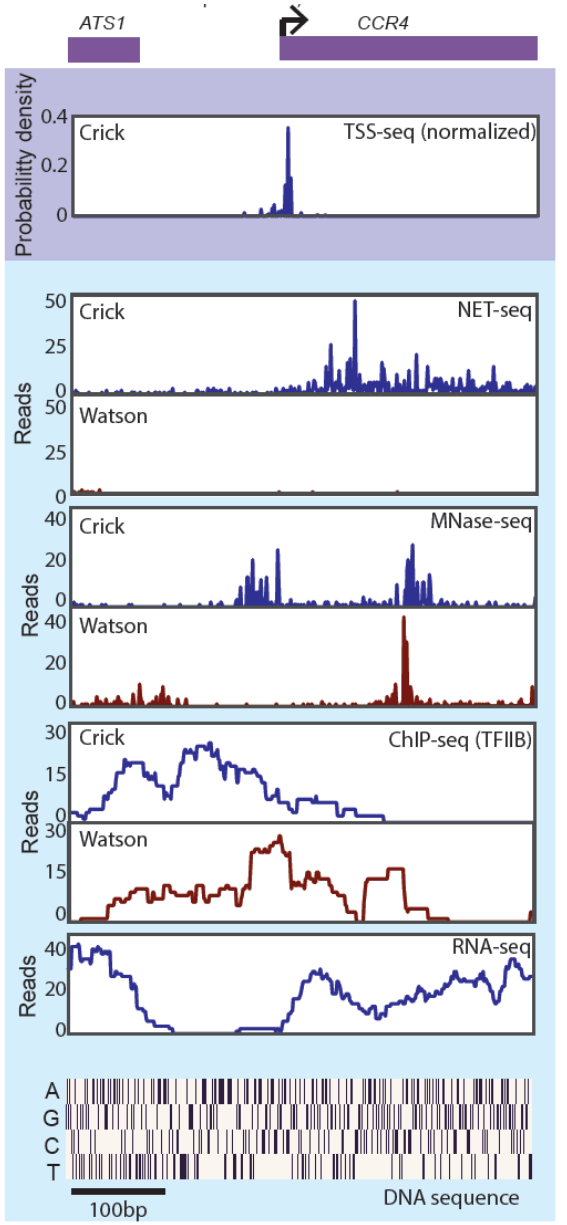


http://www.nature.com/nbt/journal/v33/n8/full/nbt.3300.html

Multiple modalities

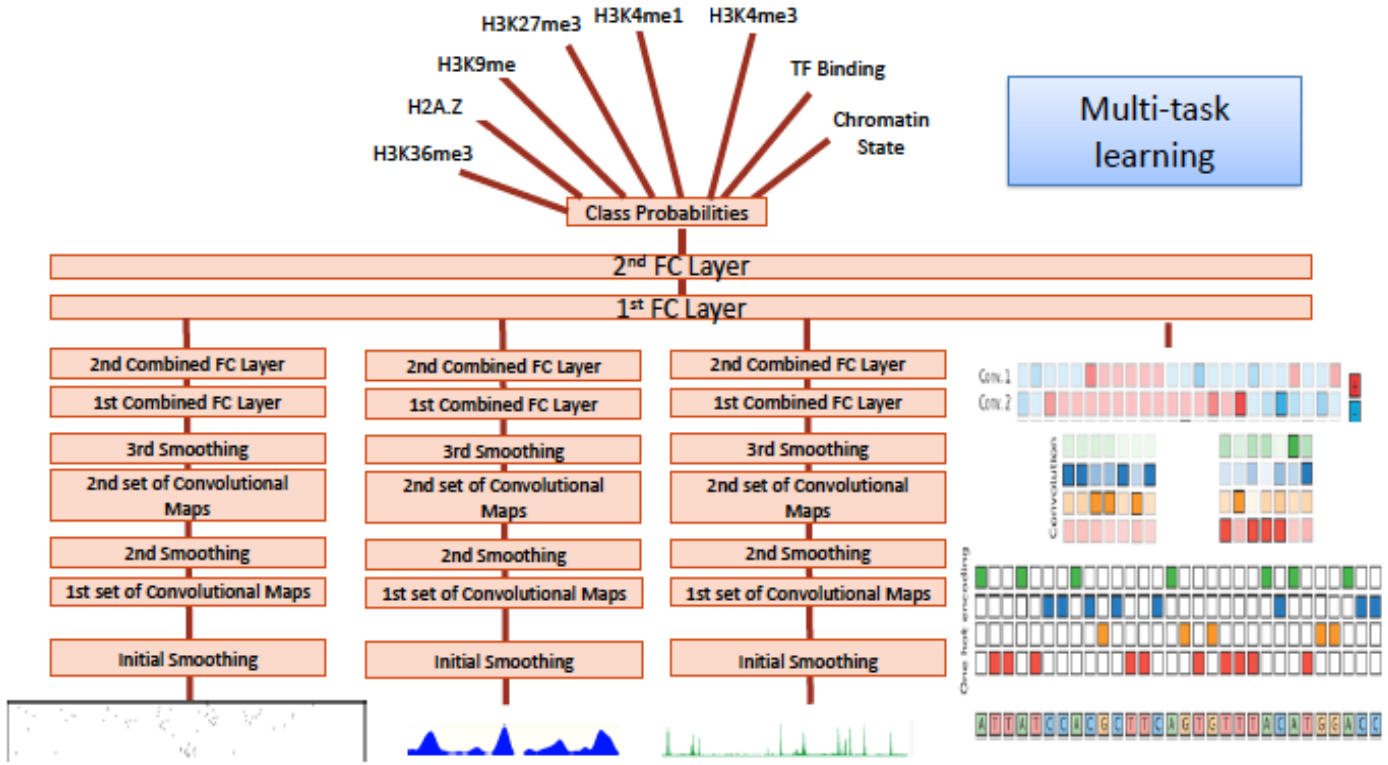
#REF: Eser, Umut, and L. Stirling Churchman. "FIDDLE: An integrative deep learning framework for functional genomic data inference." *bioRxiv* (2016): 081380.





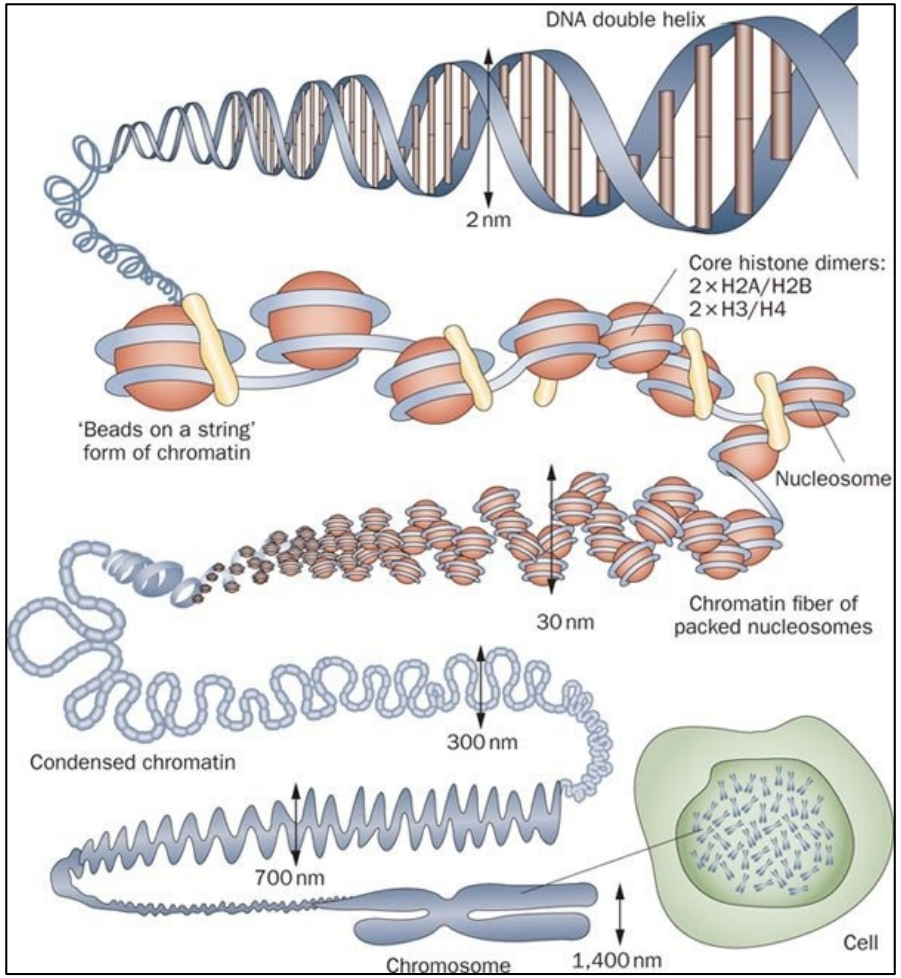
THE CHROMPUTER

Integrating multiple inputs (1D, 2D signals, sequence) to simulatenously **predict multiple outputs**

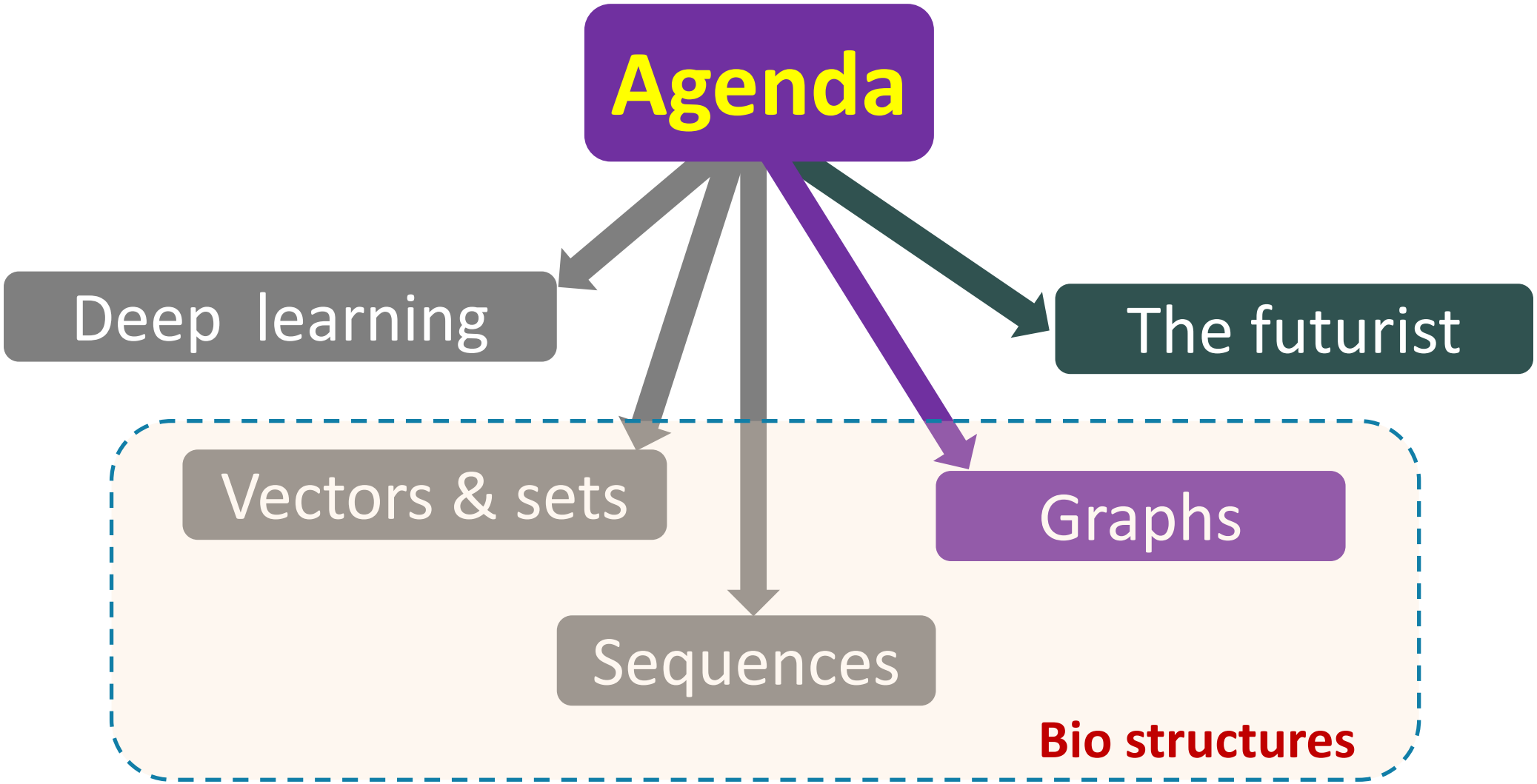


Source: https://simons.berkeley.edu/sites/default/files/docs/4575/2016-kundaje-simonsinstitute-deeplearning.pdf

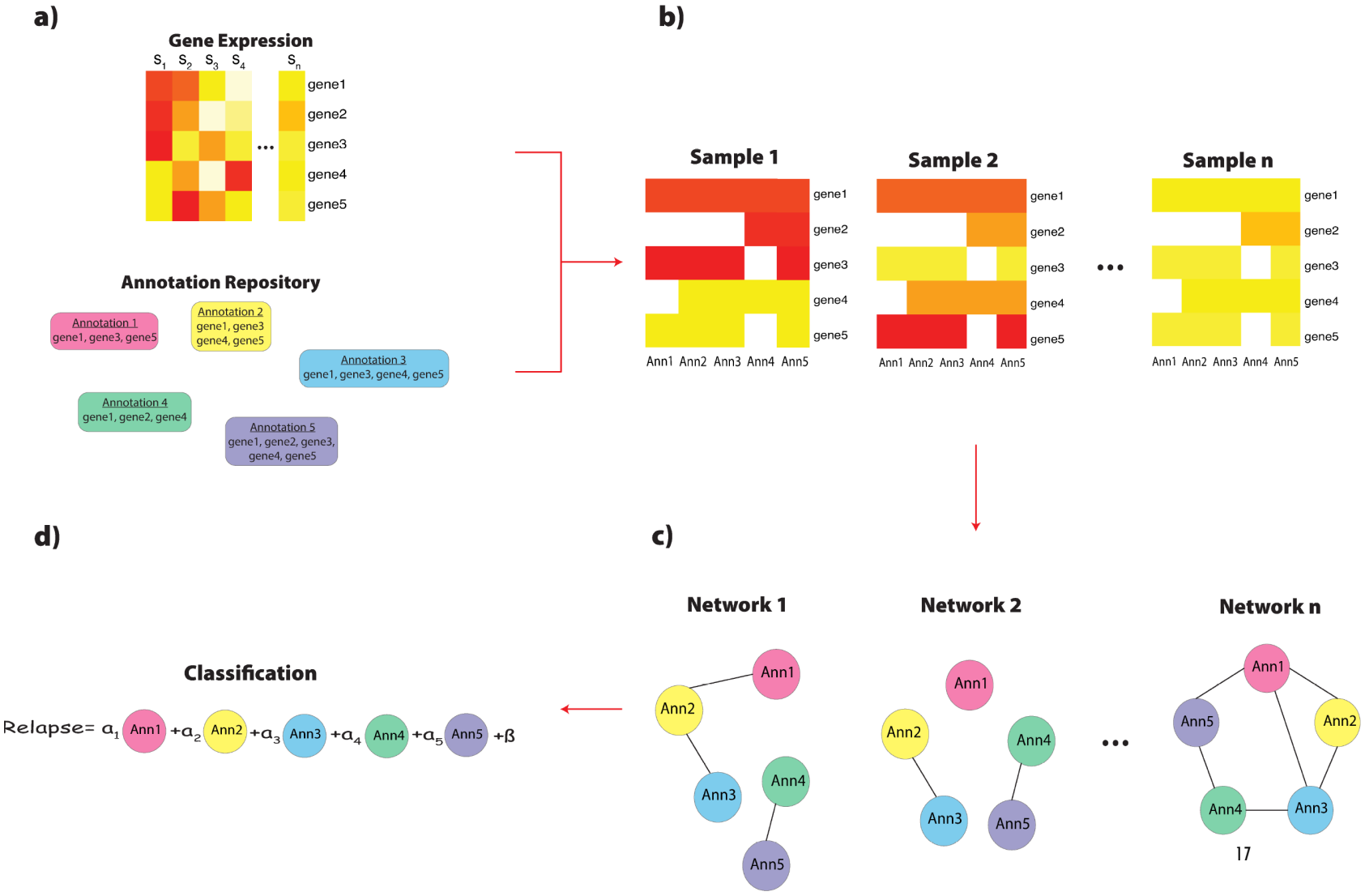
Chromatins



https://qph.ec.quoracdn.net

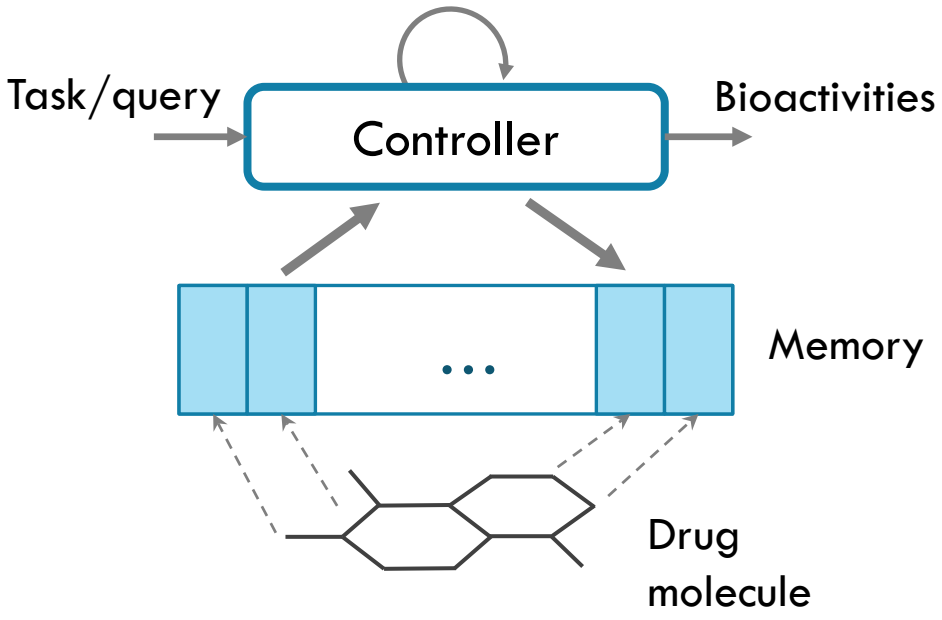

Agenda
Deep learning
The futurist
Vectors & sets
Graphs
Sequences
Bio structures

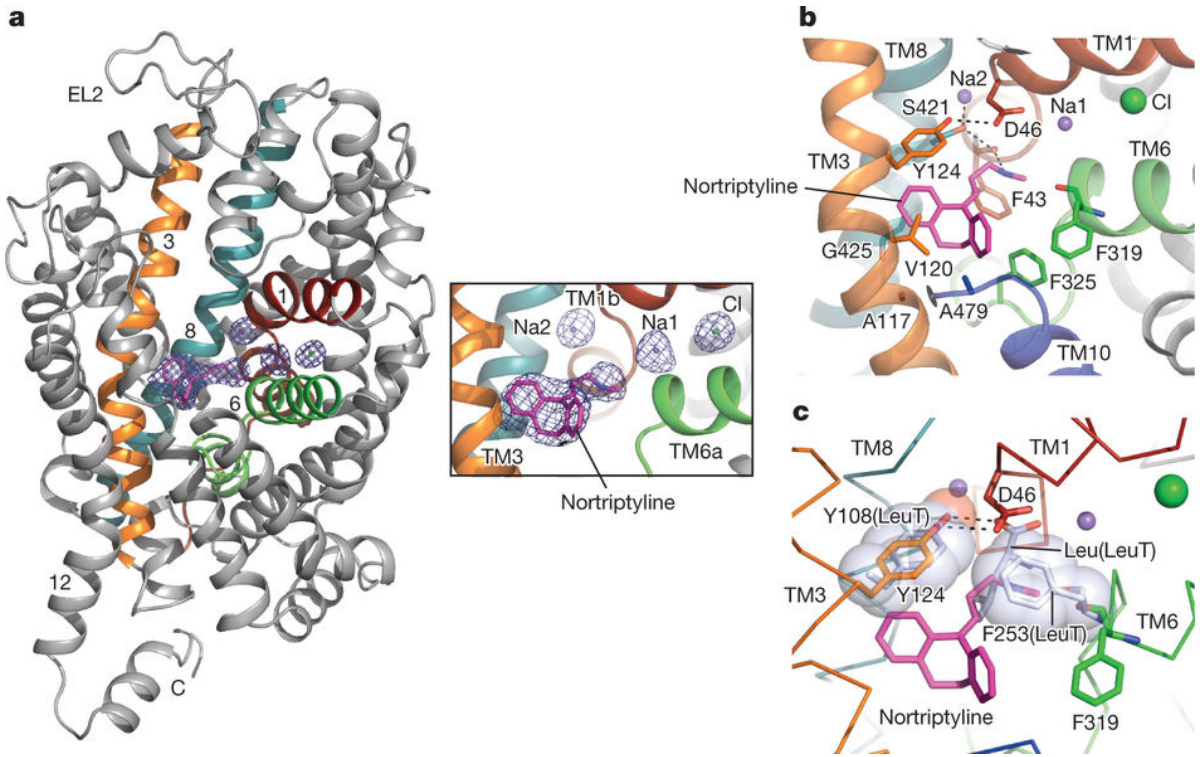
From vector to graph with PAN: Personalized Annotation Networks



Nguyen, Thin, Samuel C. Lee, Thomas P. Quinn, Buu Truong, Xiaomei Li, Truyen Tran, Svetha Venkatesh, and Thuc Duy Le. "Personalized Annotation-based Networks (PAN) for the Prediction of Breast Cancer Relapse." *bioRxiv* (2019): 534628.

Predicting molecular bioactivities as querying a graph

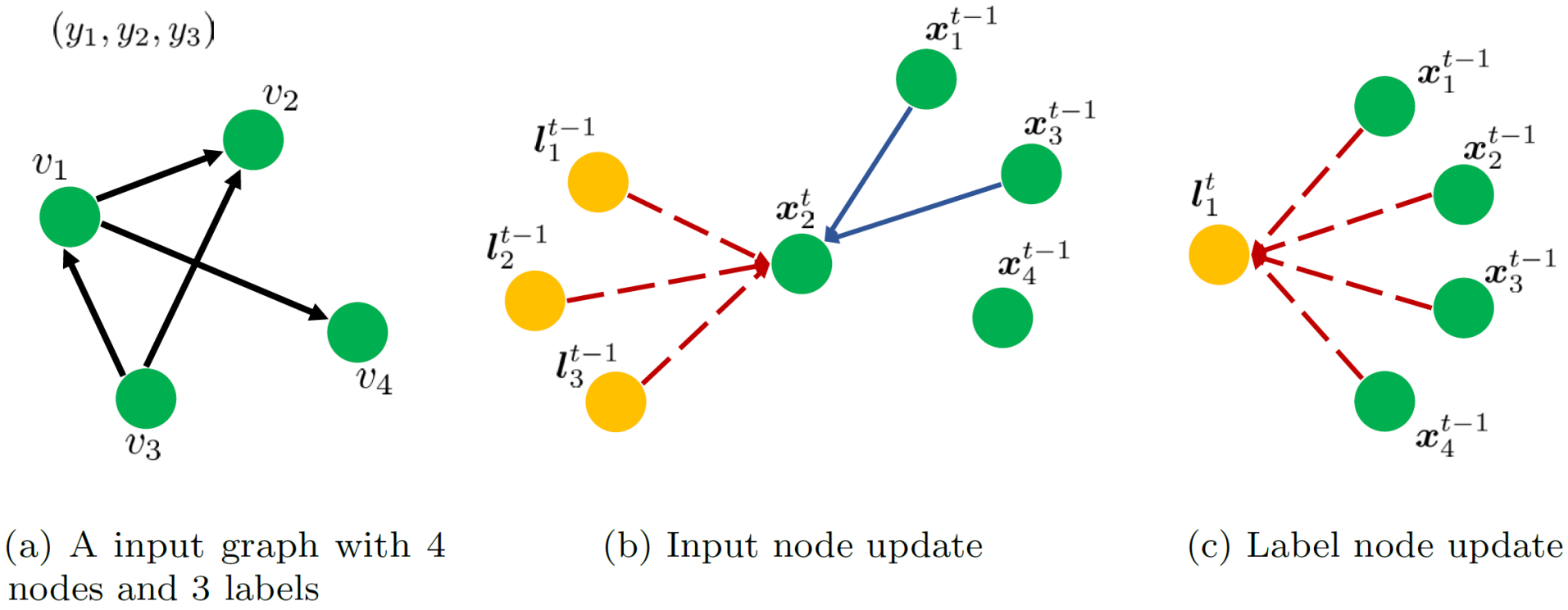




#REF: Penmatsa, Aravind, Kevin H. Wang, and Eric Gouaux. "X-ray structure of dopamine transporter elucidates antidepressant mechanism." *Nature* 503.7474 (2013): 85-90.

#Ref: Pham, Trang, Truyen Tran, and Svetha Venkatesh. "Graph Memory Networks for Molecular Activity Prediction." *ICPR'18*.

Multi-target binding for drug repurposing as graph multi-labeling



(a) A input graph with 4 nodes and 3 labels

(b) Input node update

(c) Label node update

#REF: Do, Kien, et al. "Attentional Multilabel Learning over Graphs-A message passing approach." *Machine Learning, 2019*.

Dataset	Metrics	Fingerprint		SMILES	Molecular Graph		
		SVM	HWN	GRU	WL+SVM	CLN	GAML
9cancers	m-AUC	81.94	85.95	83.29	86.06	88.35	**88.78**
	M-AUC	81.37	85.85	82.74	85.74	88.23	**88.50**
	m-F1	50.63	57.44	55.97	54.55	59.48	**62.03***
	M-F1	50.71	57.29	55.99	54.54	59.50	**62.14***
50proteins	m-AUC	79.85	77.46	79.11	81.62	82.08	**82.82**
	M-AUC	74.77	73.78	75.25	77.60	78.36	**79.35***
	m-F1	17.21	16.37	16.08	17.04	18.37	**20.47***
	M-F1	18.40	15.87	14.96	18.66	17.72	**19.83***

Table 4: The performance in the multi-label classification with graph-structured input (m-X: micro average of X; M-X: macro average). SVM and HWN work on fingerprint representation; GRU works on string representation of molecule known as SMILES; WL+BR and CLN work directly on graph representation. Bold indicates better values. (*) $p < 0.05$.

#REF: Do, Kien, et al. "Attentional Multilabel Learning over Graphs-A message passing approach." *arXiv preprint arXiv:1804.00293*(2018).

Drug-target binding as graph reasoning

Reasoning is to deduce knowledge from previously acquired knowledge in response to a query (or a cues)

Can be formulated as Question-Answering or Graph-Graph interaction:

- Knowledge base: Binding targets (e.g., RNA/protein sequence, or 3D structures), as a graph
- Query: Drug (e.g., SMILES string, or molecular graph)
- Answer: Affinity, binding sites, modulating effects

An analogy from Video Question Answering

Video as sequence of frame, but also a complex 3D graph of objects, actions and scenes

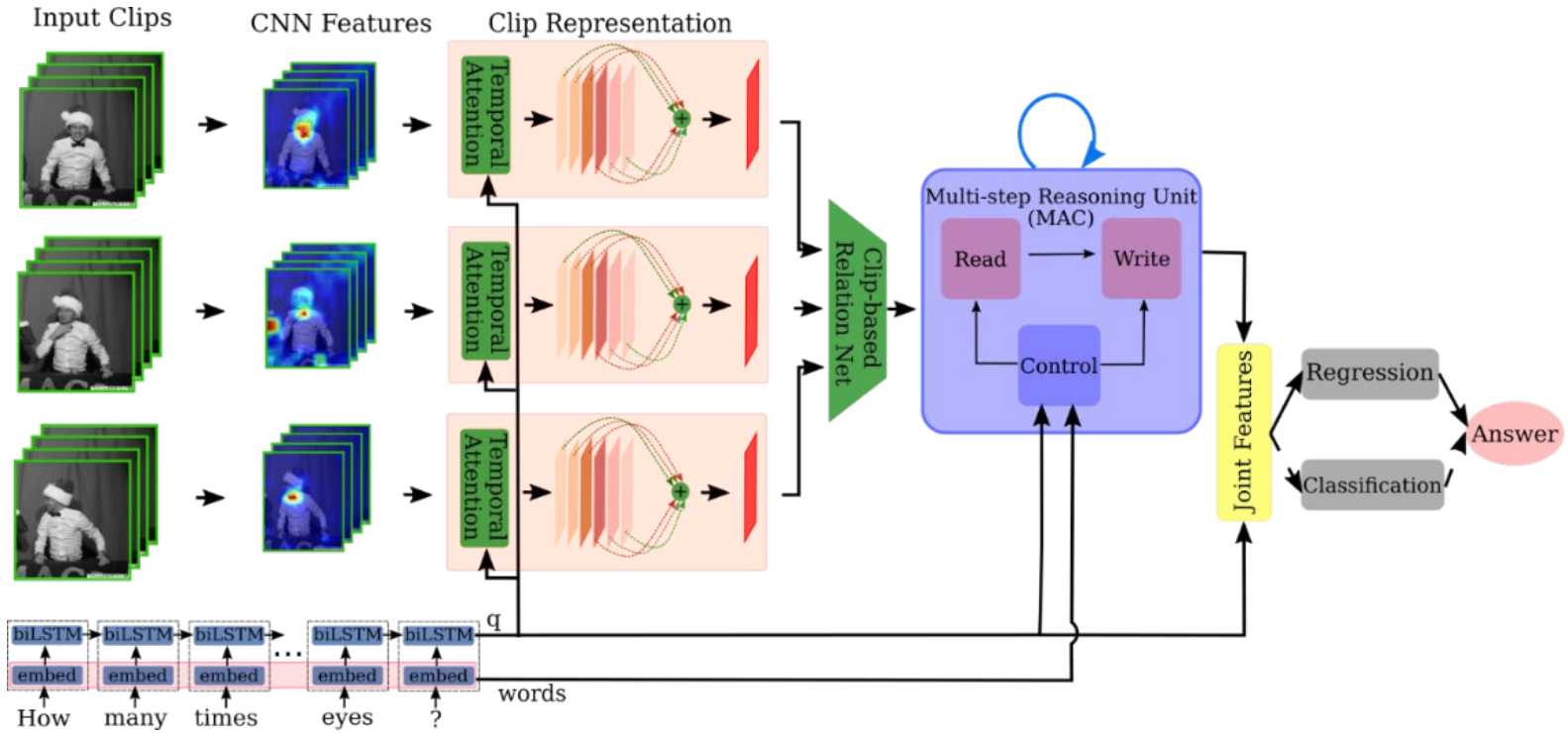
- → **Protein, RNA**

Question as sequence of words, but also a complex dependency graph of concepts

- → **Protein, drug**

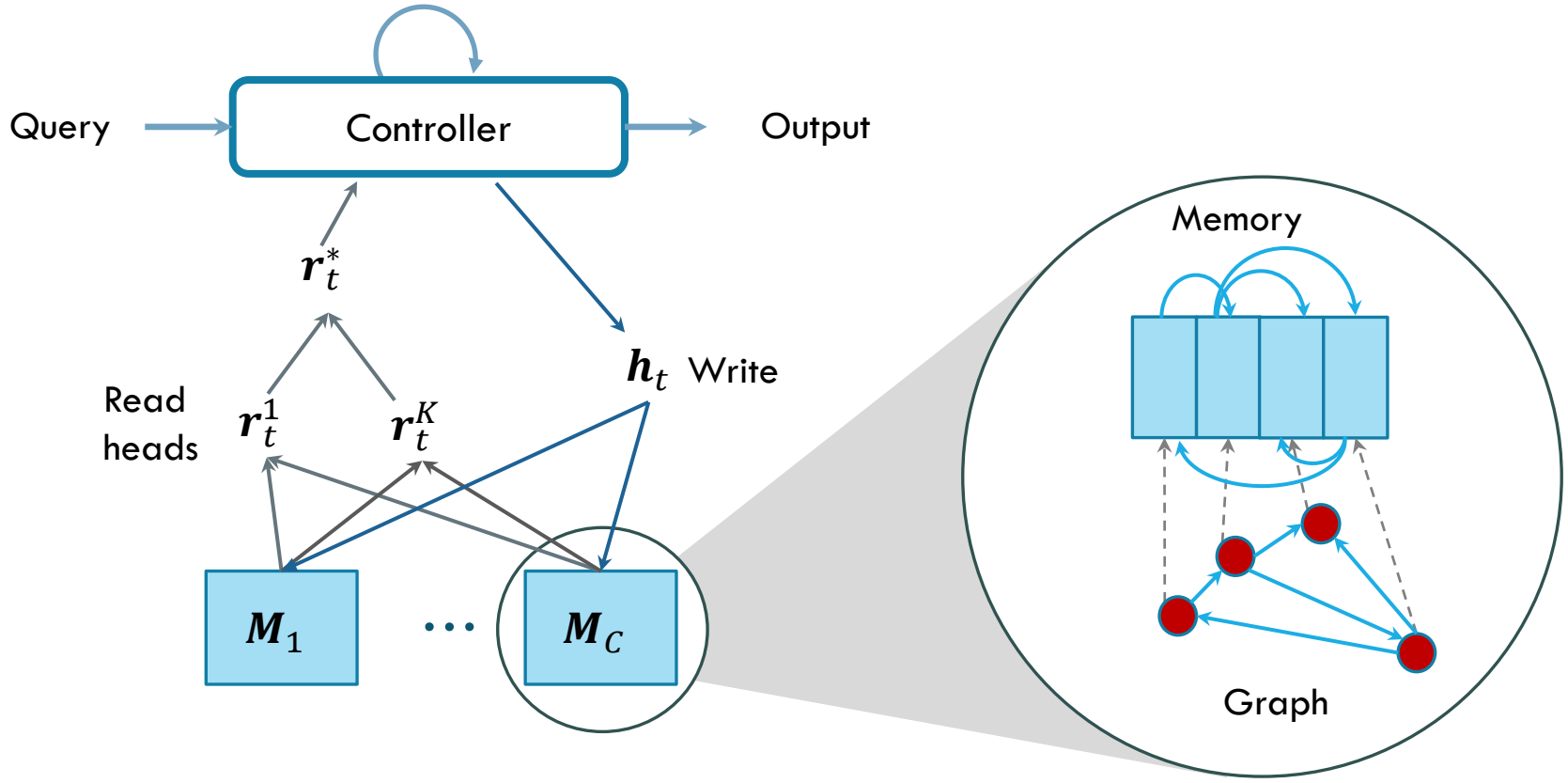
Answer as facts (what and where) and deduced knowledge.

- → **Affinity, binding sites, modulation effect**



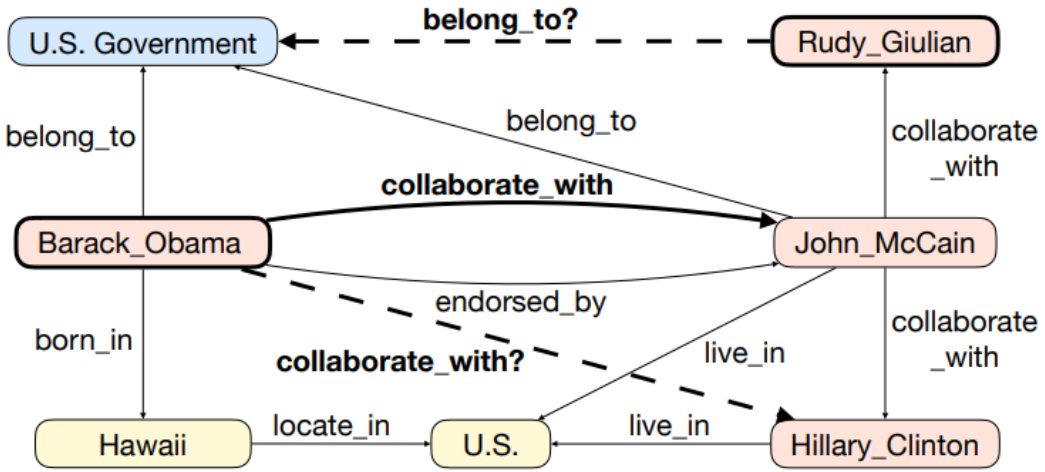
#Ref: Minh-Thao Le, Vuong Le, Truyen Tran, Learning to Reason with Relational Video Representation for Question Answering, *In Submission 2019*.

Drug-drug, drug-target & protein-protein as graph-graph interaction

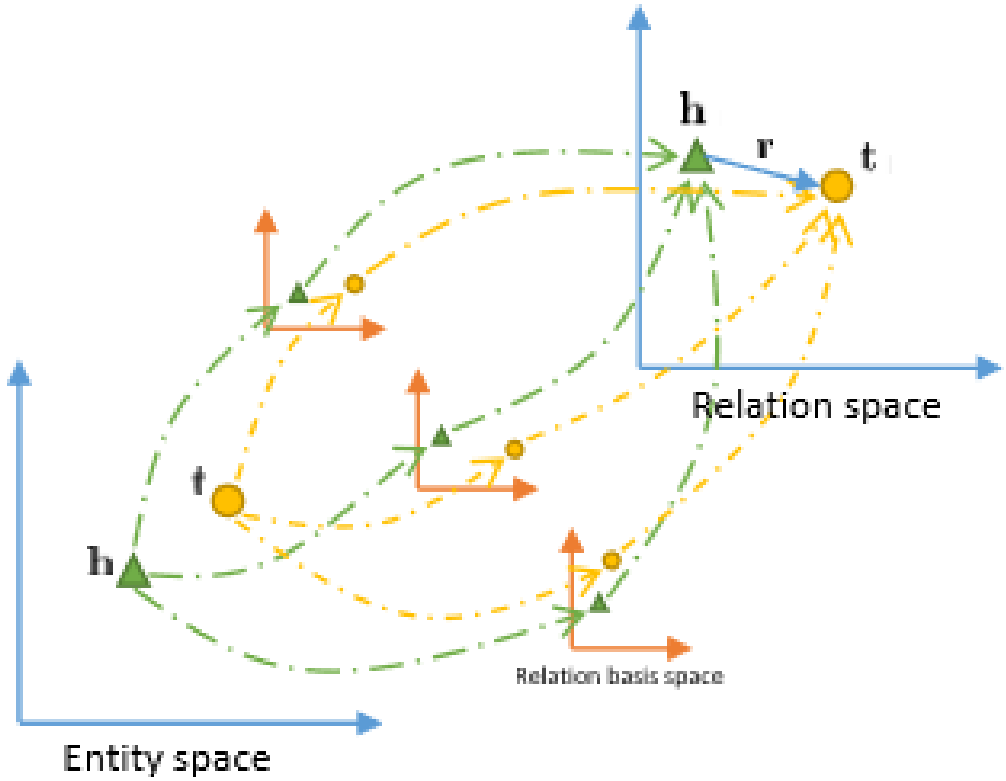


Pham, Trang, Truyen Tran, and Svetha Venkatesh. "Relational dynamic memory networks." *arXiv preprint arXiv:1808.04247*(2018).

Inferring (bio) relations as knowledge graph completion



https://www.zdnet.com/article/salesforce-research-knowledge-graphs-and-machine-learning-to-power-einstein/



Do, Kien, Truyen Tran, and Svetha Venkatesh. "Knowledge graph embedding with multiple relation projections." *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018.

Drug design as structured machine translation, aka conditional generation

Can be formulated as structured machine translation:

- Inverse mapping of (knowledge base + binding properties) to (query) → One to many relationship.

Representing graph as string (e.g., SMILES), and use sequence VAEs or GANs.

Graph VAE & GAN

- Model nodes & interactions
- Model cliques

Sequences

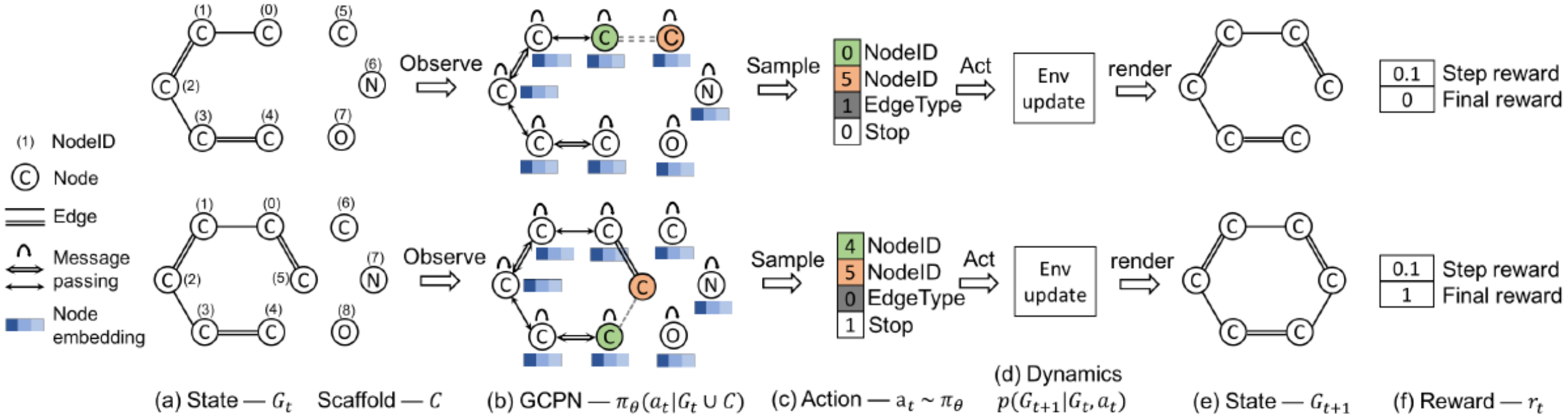
- Iterative methods

Reinforcement learning

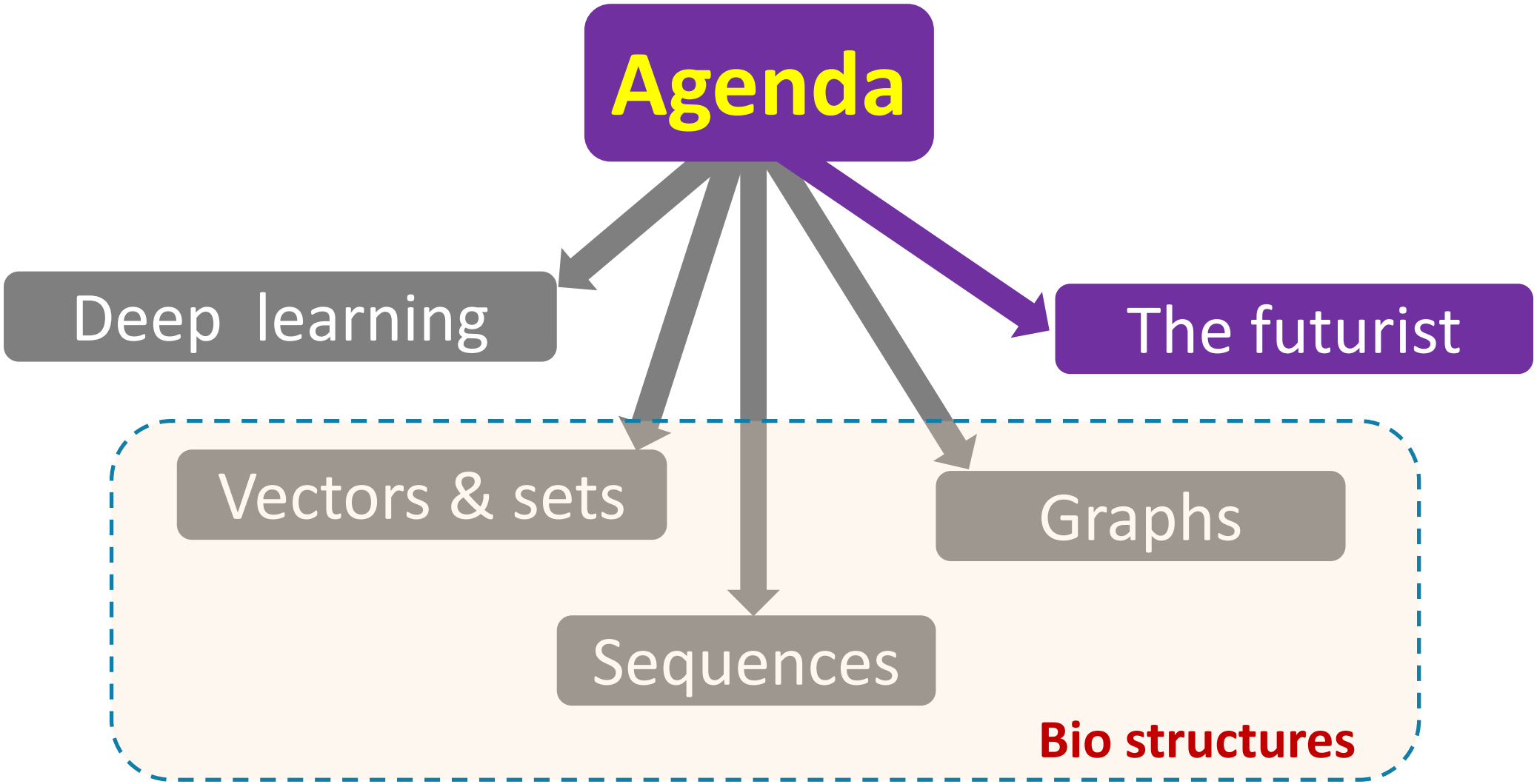
- Discrete objectives

Any combination of these + memory.

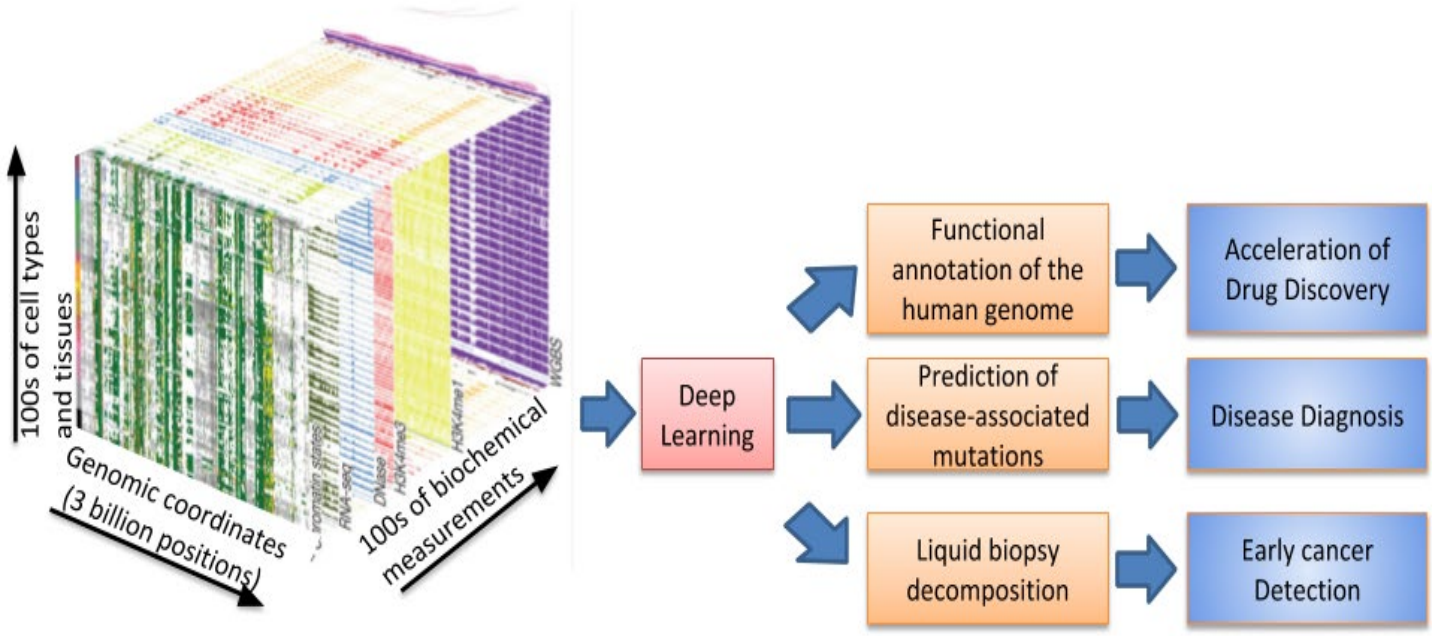
Drug design as reinforcement learning



You, Jiaxuan, et al. "Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation." *NeurIPS* (2018).


Agenda
Deep learning
The futurist
Vectors & sets
Graphs
Sequences
Bio structures

Opportunities for Deep Learning in Genomics



- Genetic diagnostics
- Refining drug targets
- Pharmaceutical development
- Personalized medicine
- Better health insurance
- Synthetic biology

https://towardsdatascience.com/opportunities-and-obstacles-for-deep-learning-in-biology-and-medicine-6ec914fe18c2
https://www.oreilly.com/ideas/deep-learning-meets-genome-biology

Deep learning versus genomics

Bertolero, M. A., Blevins, A. S., Baum, G. L., Gur, R. C., Gur, R. E., Roalf, D. R., ... & Bassett, D. S. (2019). The network architecture of the human brain is modularly encoded in the genome. *arXiv preprint arXiv:1905.07606*.

Neuron ↔ Nucleotide, amino acid (building bricks)

Neural networks ↔ Chemical/biological networks (the house)

Message passing ↔ Signalling (the communication)

Neural programs ↔ Proteins/RNAs (the operating machines)

Neural Turing machine ↔ DNA (data + instruction + control)

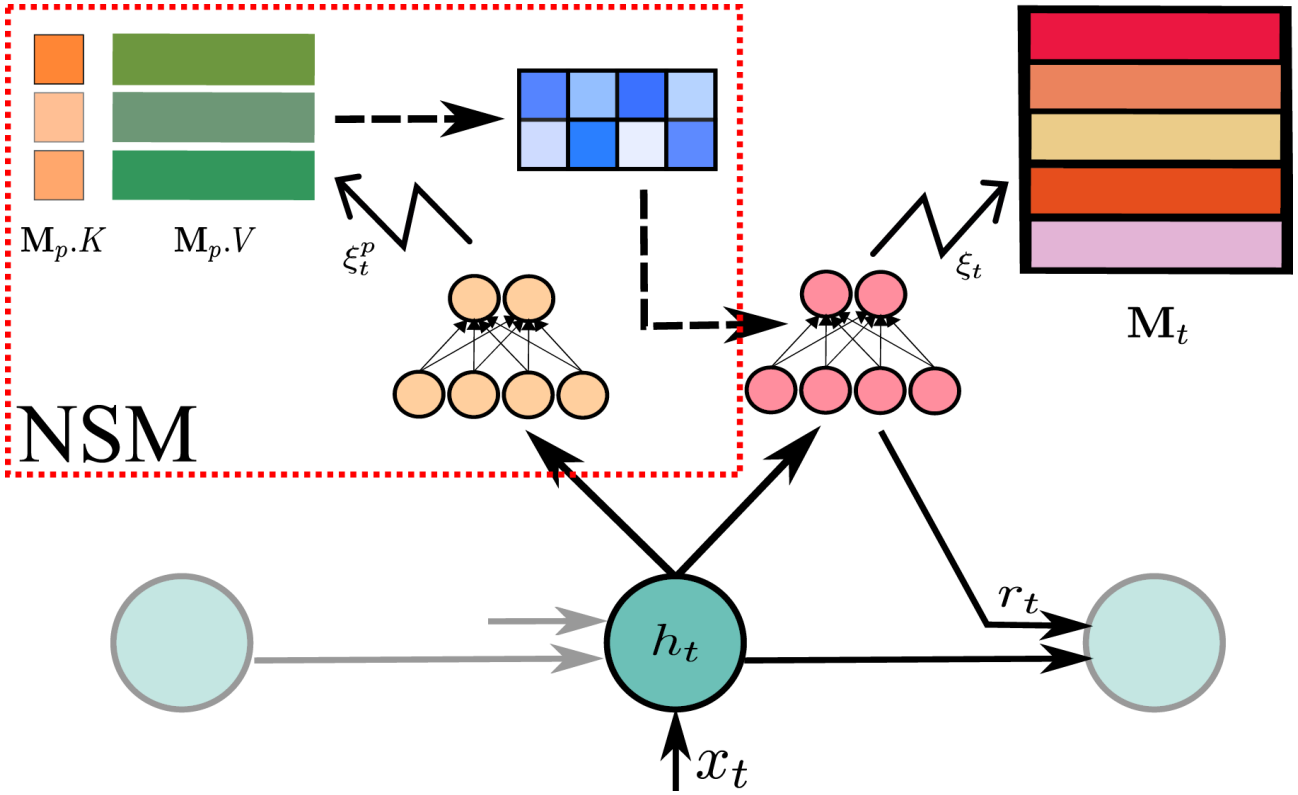
Neural universe ↔ Omics universe (the computational universe)

Learning over time ↔ Co-evolution (adaptation)

Super Neural Turing machine ↔ DNA + Evolution (data + program + adaption)

Living bodies as multiple programs interacting

- We need new (neural) capabilities:
  - Truly Turing machine: programs can be stored and called when needed.
  - Can solve BIG problem with many sub-modules.
    - → Composionality
  - Can reason given existing structures and knowledge bases



Neural Stored-program Memory

The team


DEAKIN
UNIVERSITY AUSTRALIA
Worldly


THE UNIVERSITY OF
SYDNEY


GARVAN
INSTITUTE


Barwon
Health


Epworth
HealthCare