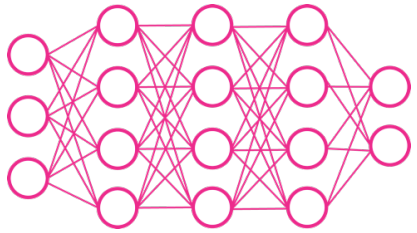
Making Deep Learning Understandable for Analyzing Sequential Data about Gene Regulation

Dr. Yanjun Qi
Department of Computer Science
University of Virginia

Tutorial @ ACM BCB-2018

BREAK 5mins ->Second Half

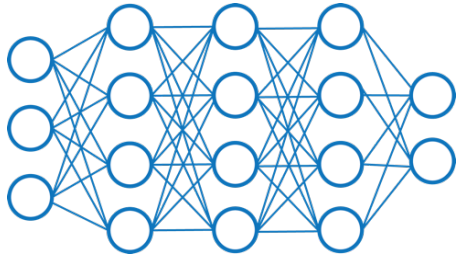
State-of-the-art: Deep Neural Networks (DNNs)



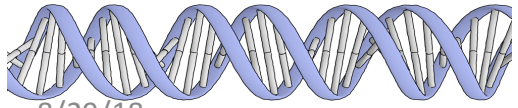
“Dog”

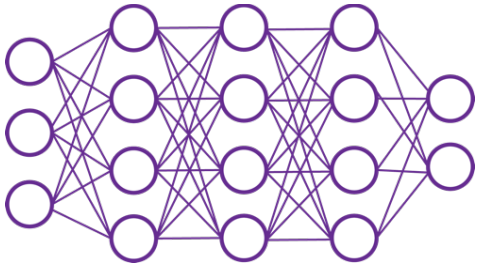
Can get overly sentimental at times, but Gus Van Sant's sensitive direction... and his excellent use of the city make it a hugely entertaining and effective film.

Full Review... | May 25, 2006



ATGCGATCAAGTCTG





“Protein-binding Site”

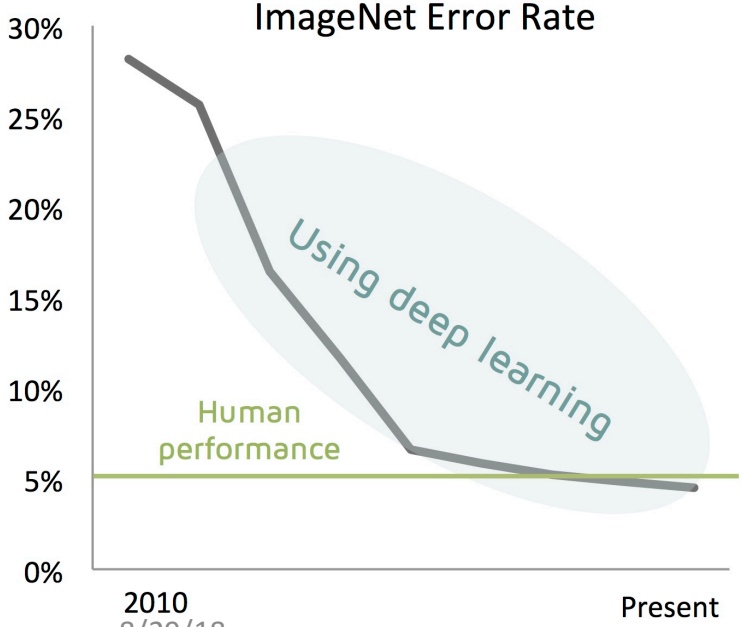
Challenge : DNNs are hard to Interpret

X

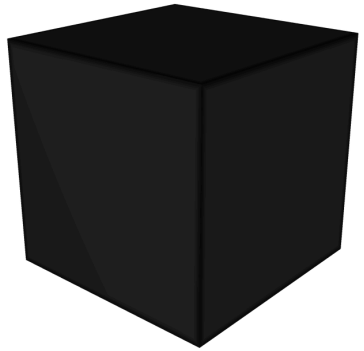
Y

“Dog”

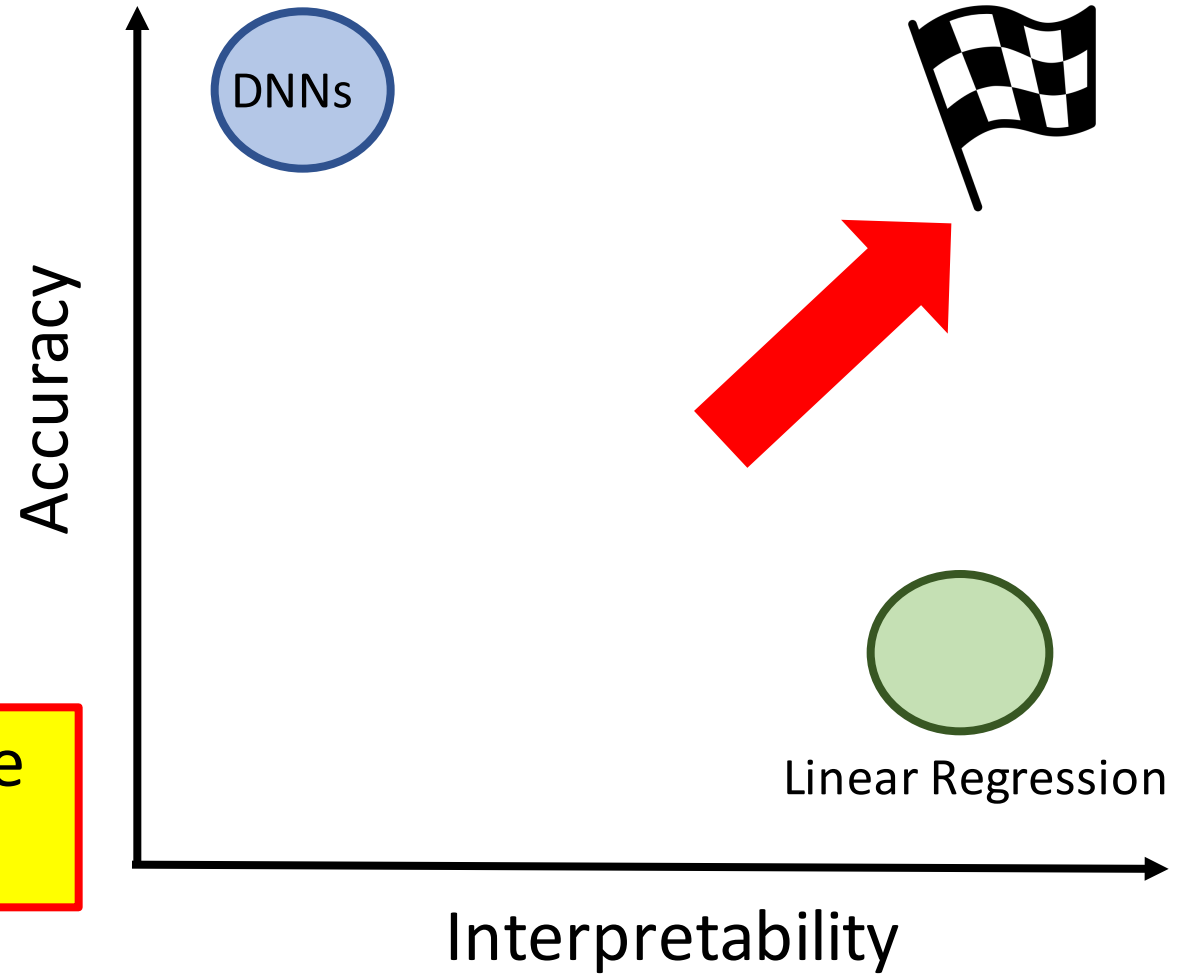
$f_1(.)$ $f_2(.)$ $f_3(.)$ $f_4(.)$



$$Y=f_4(f_3(f_2(f_1(X))))$$



Our Goal: Interpretable DNNs



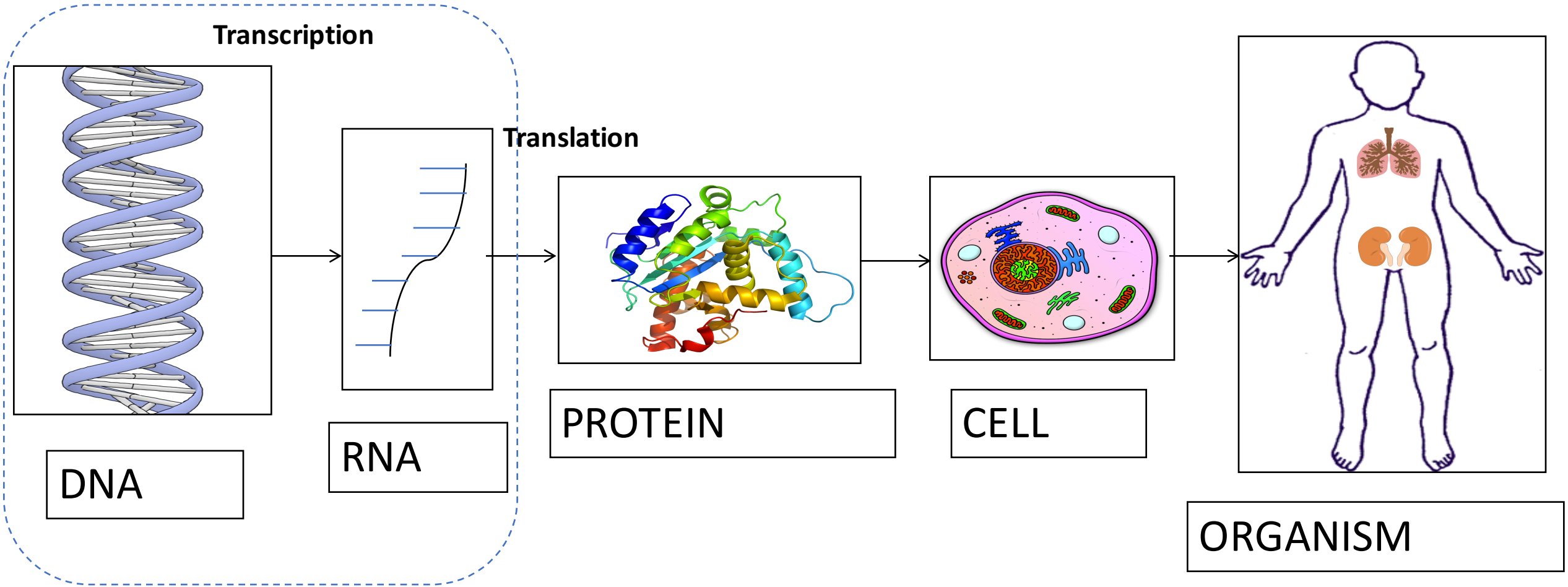
Challenge : DNNs are hard to Interpret

Today

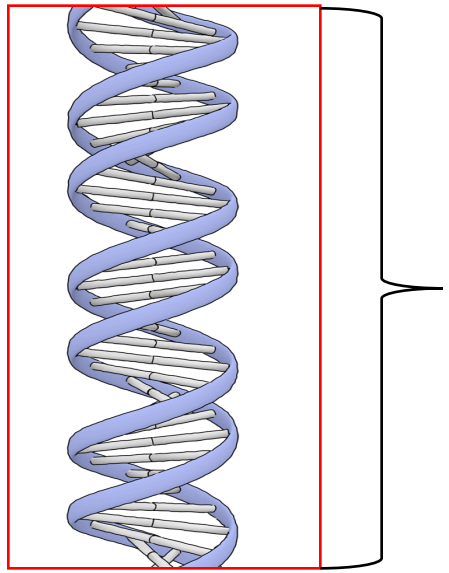
- Machine Learning: a quick review
- Deep Learning: a quick review https://qdata.github.io/deep2Read/
- Background Biology: a quick review
- Deep Learning for analyzing Sequential Data about Regulation:
  - DeepChrome
  - AttentiveChrome
  - DeepMotif

https://www.deepchrome.org

Biology in a Slide



DNA and Diseases



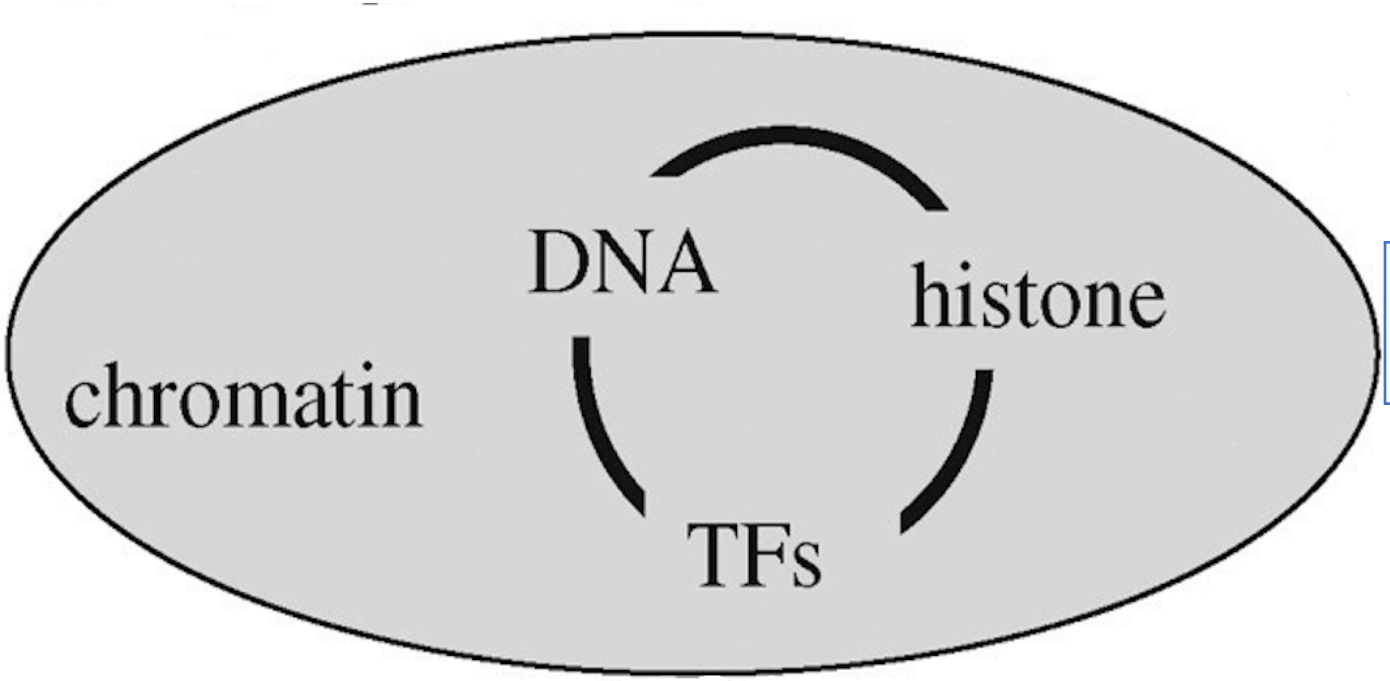
DNA

- Down Syndrome
- Parkinson's Disease
- Autism
- Muscular Atrophy
- Sickle Cell Disease

…….

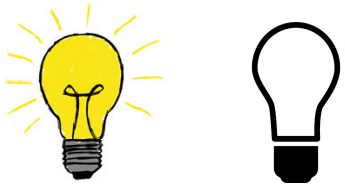
…….

Chromatin



Control

Gene Transcription



Epigenetics "Environment of the DNA"

Histone Proteins

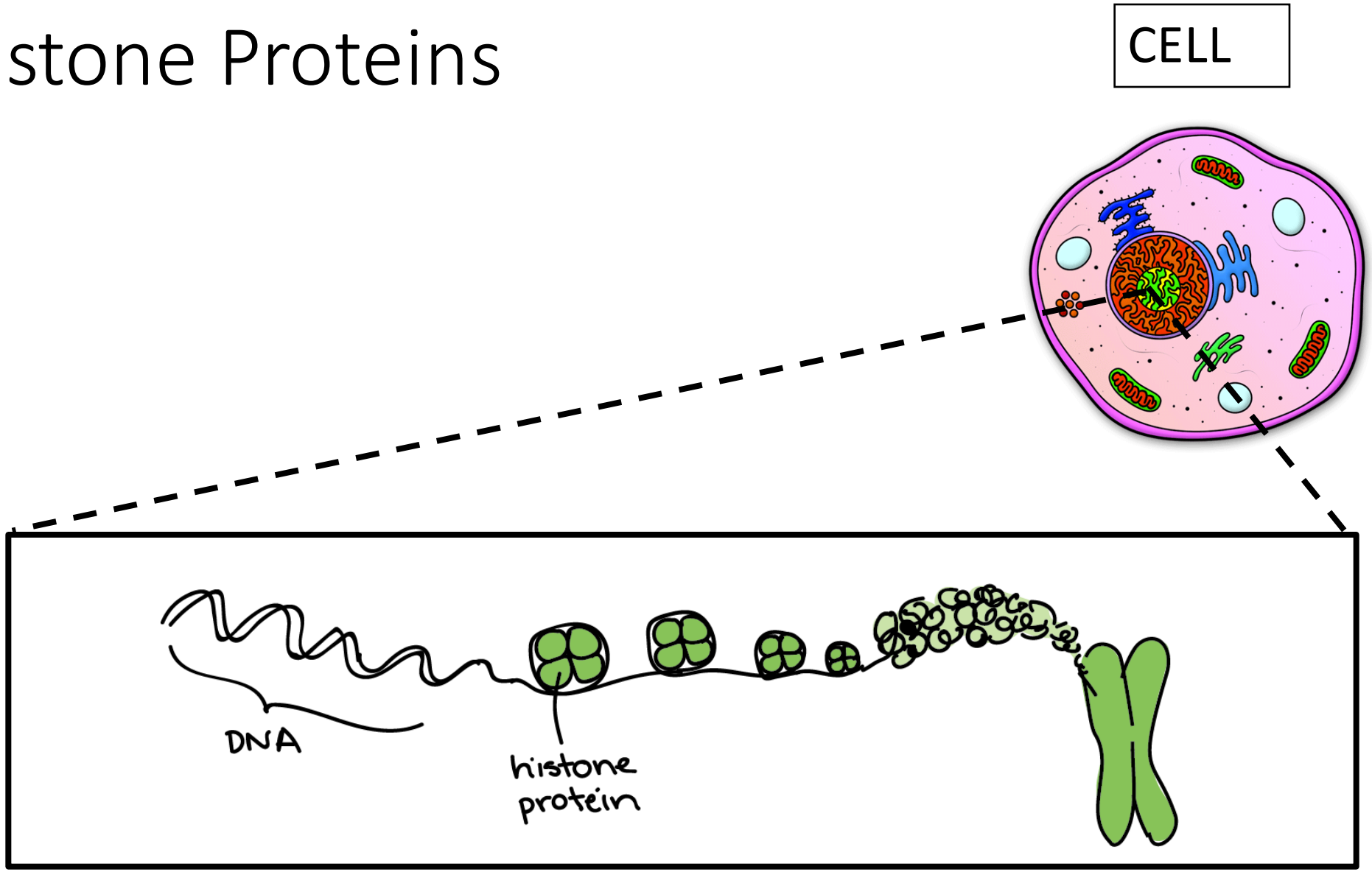
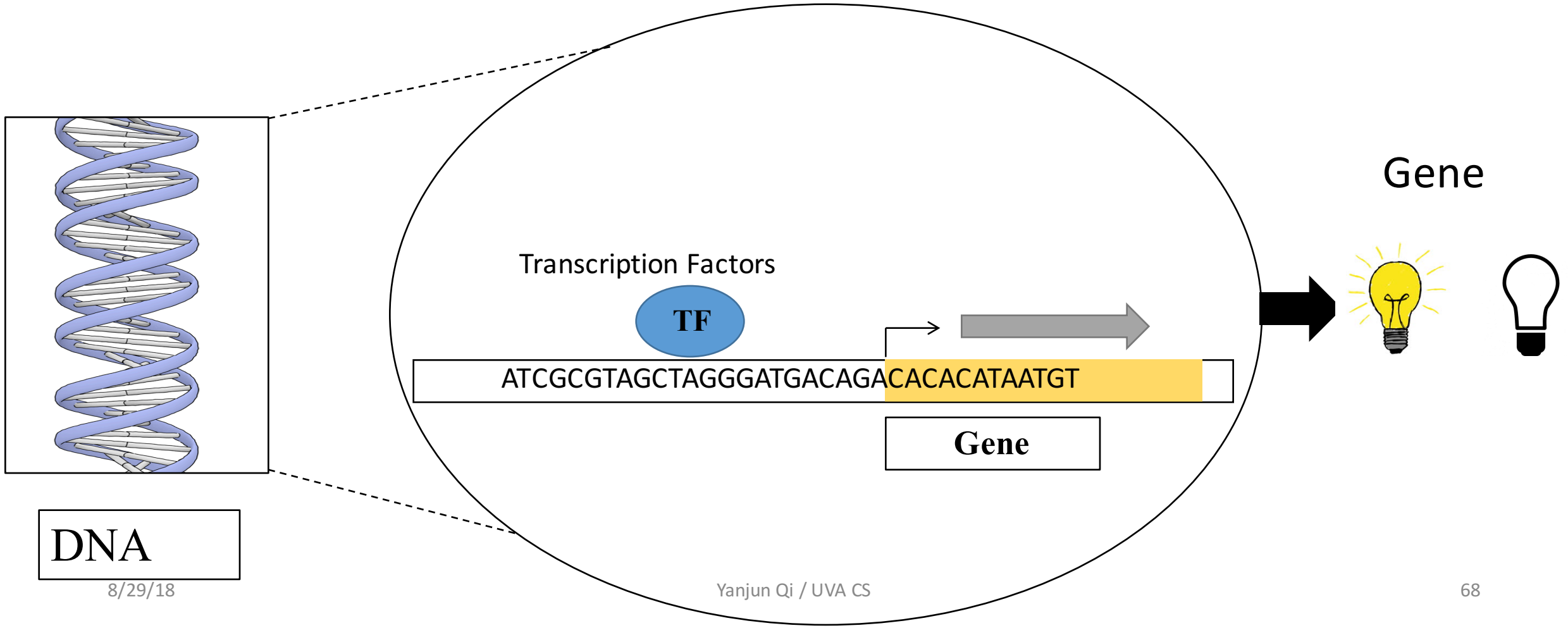
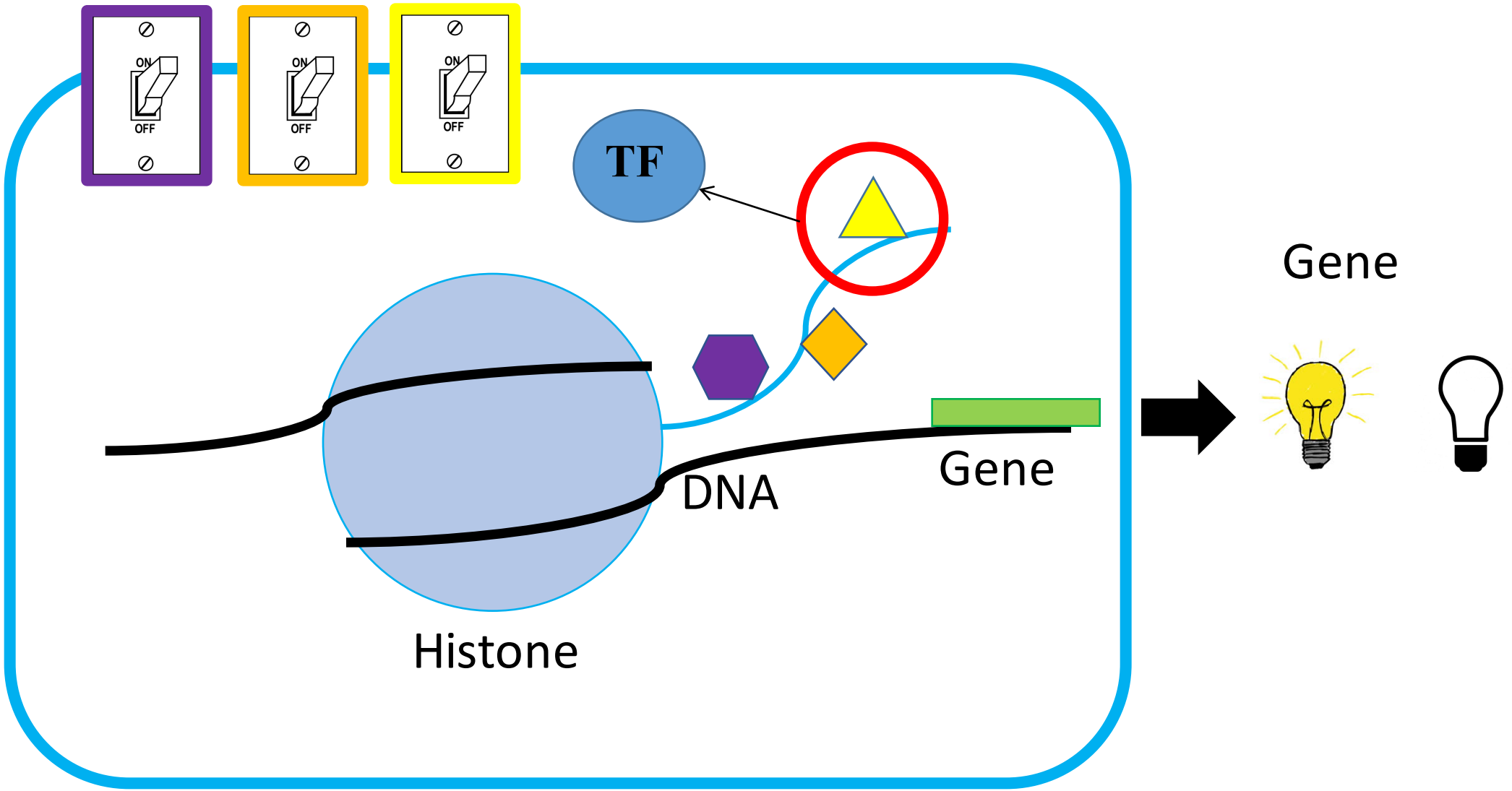


Image:https://www.khanacademy.org/science/biology/cellular-molecular-biology/intro-to-cell-division/a/dna-and-chromosomes-article

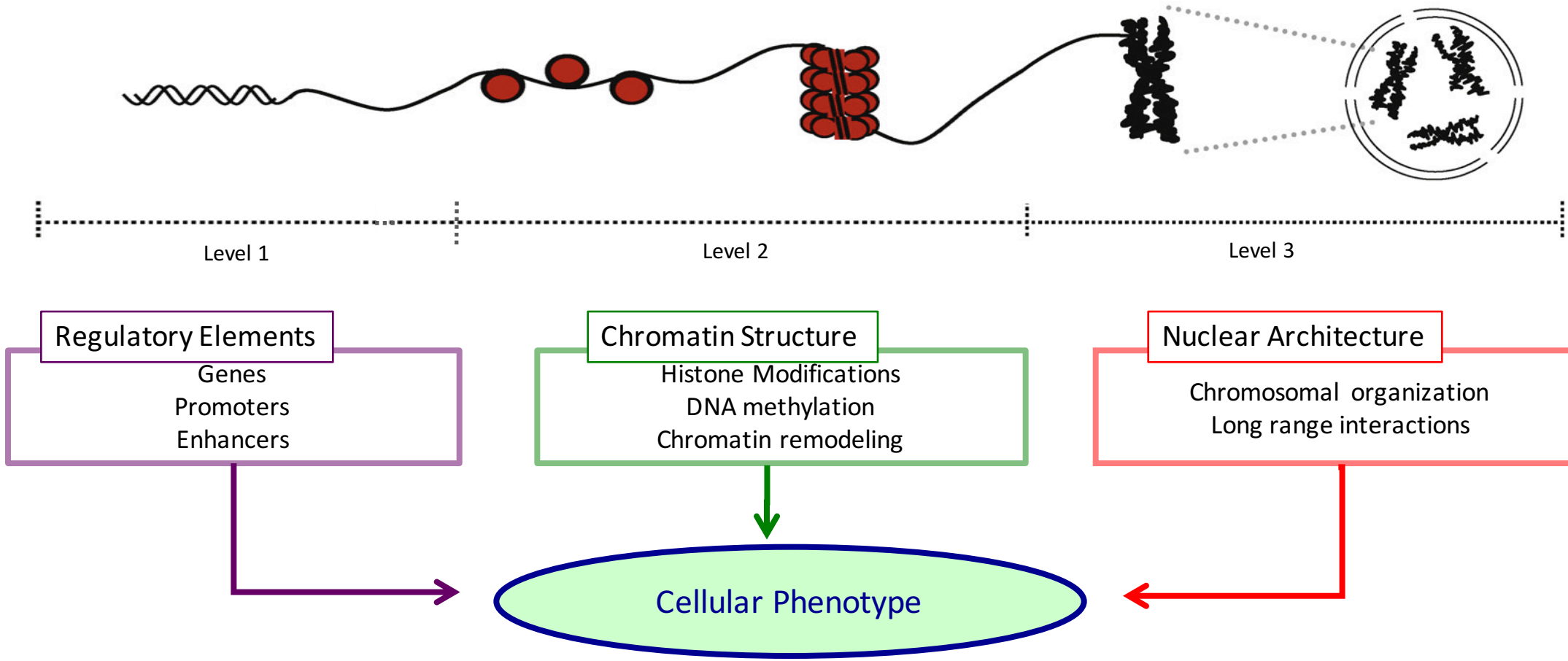
Transcription Factor Binding => Gene Transcription



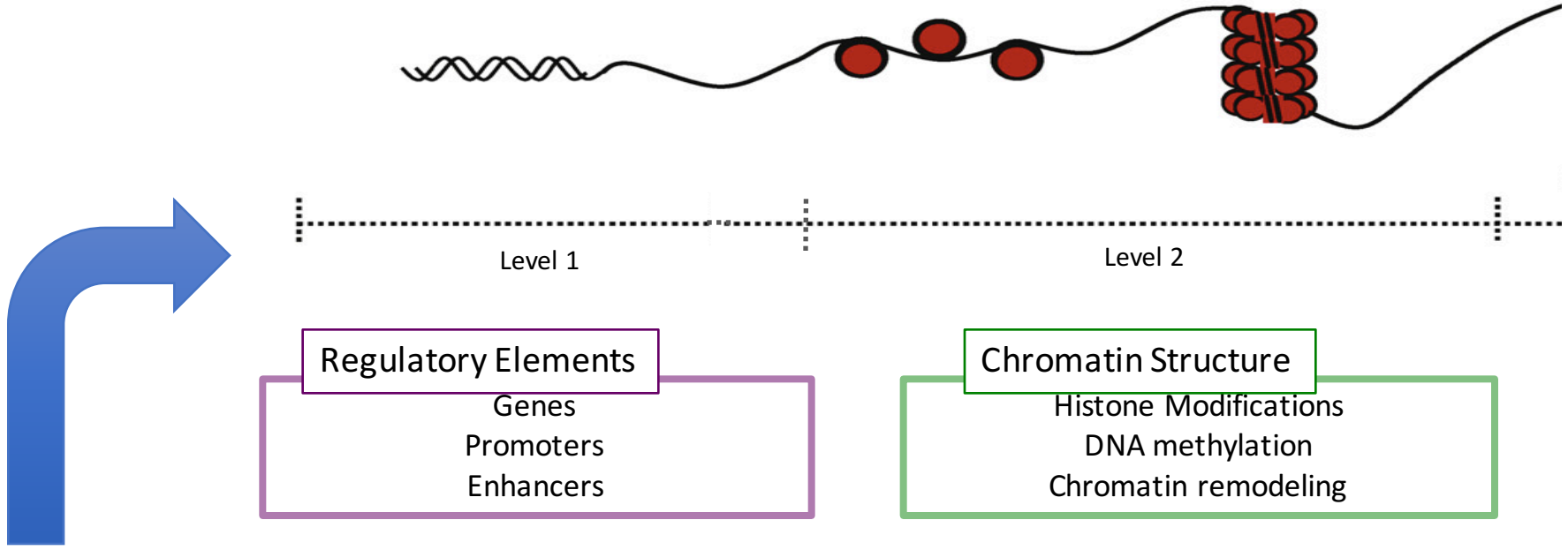
Histone Modifications (HM)



Genome Organization and Gene Regulation

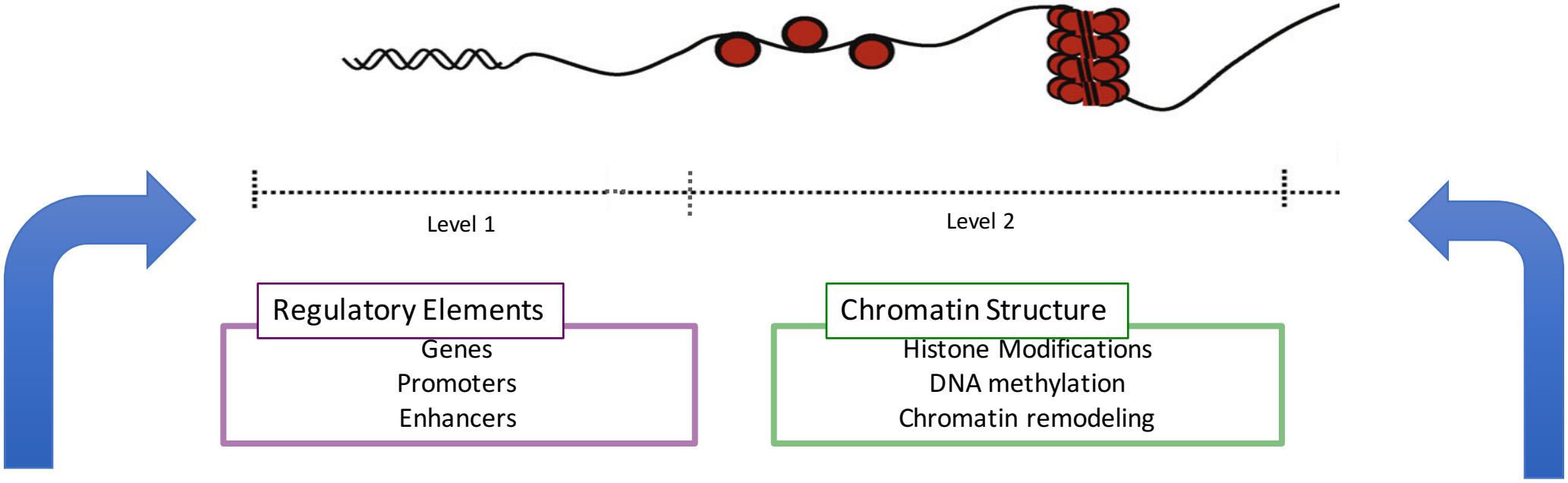


(adapted from Babu et al., 2008)



ENCODE Project (2003-Present)

Describe the functional elements encoded in human DNA



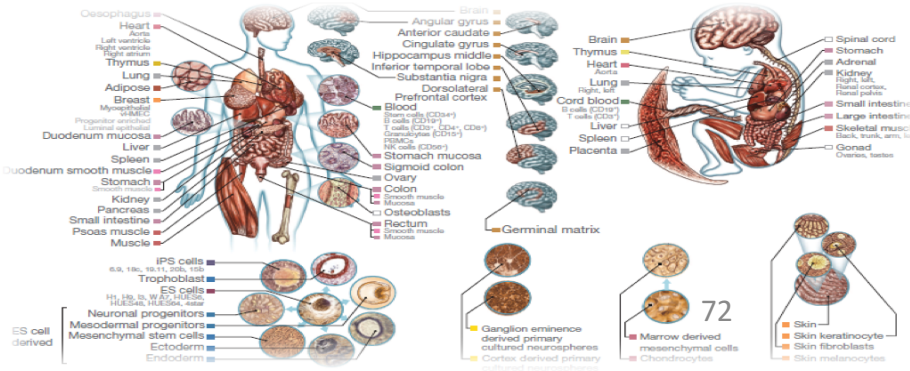
ENCODE Project (2003-)

Describe the functional elements encoded in human DNA

Roadmap Epigenetics Project (REMC, 2008-)

To produce a public resource of epigenomic maps for stem cells and primary ex vivo tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease.

Integrative analysis of 111 reference human epigenomes (Abstract)

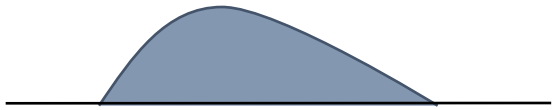


Many Important Data-Driven Computational Tasks

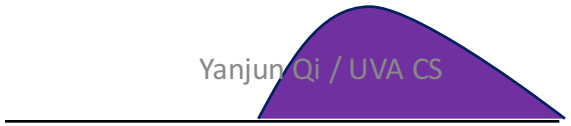
DNA Segments on Genomes

ATGCGATCAAGTCTG

TF Binding Signals



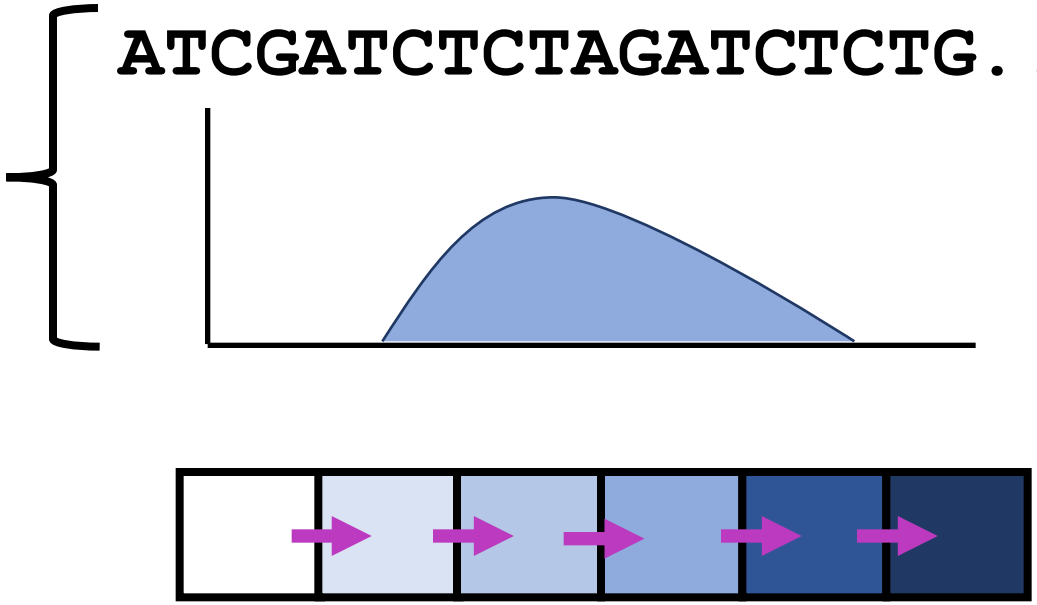
Histone Modification Signals



Gene Expression

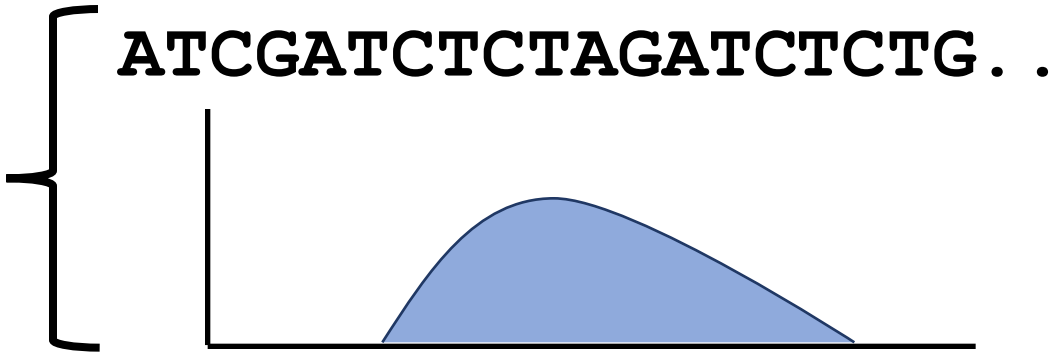
Sequential Input (X)

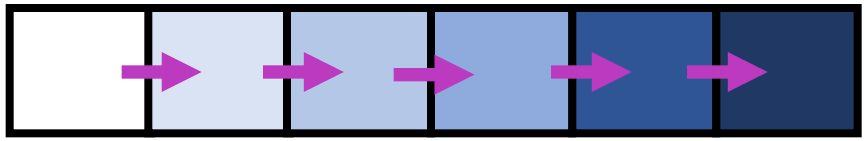
Strings, signals etc.



Sequential Input (X)

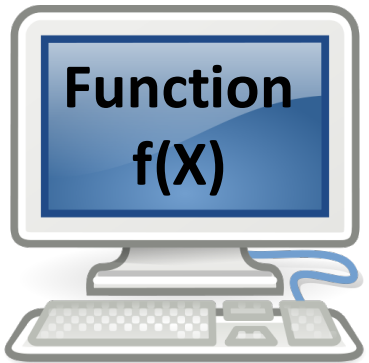
Strings, signals etc.





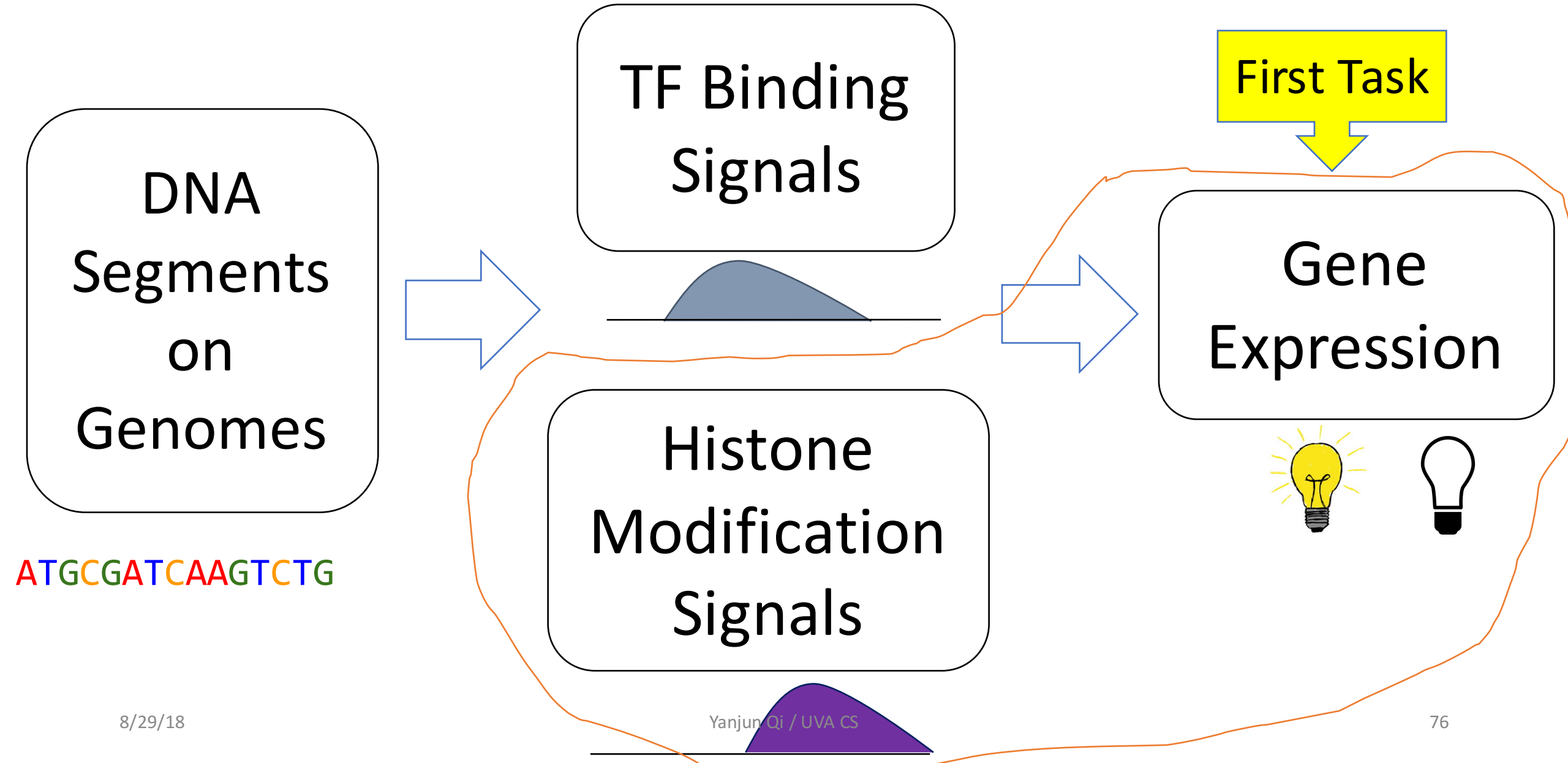
X

This Food is not good.

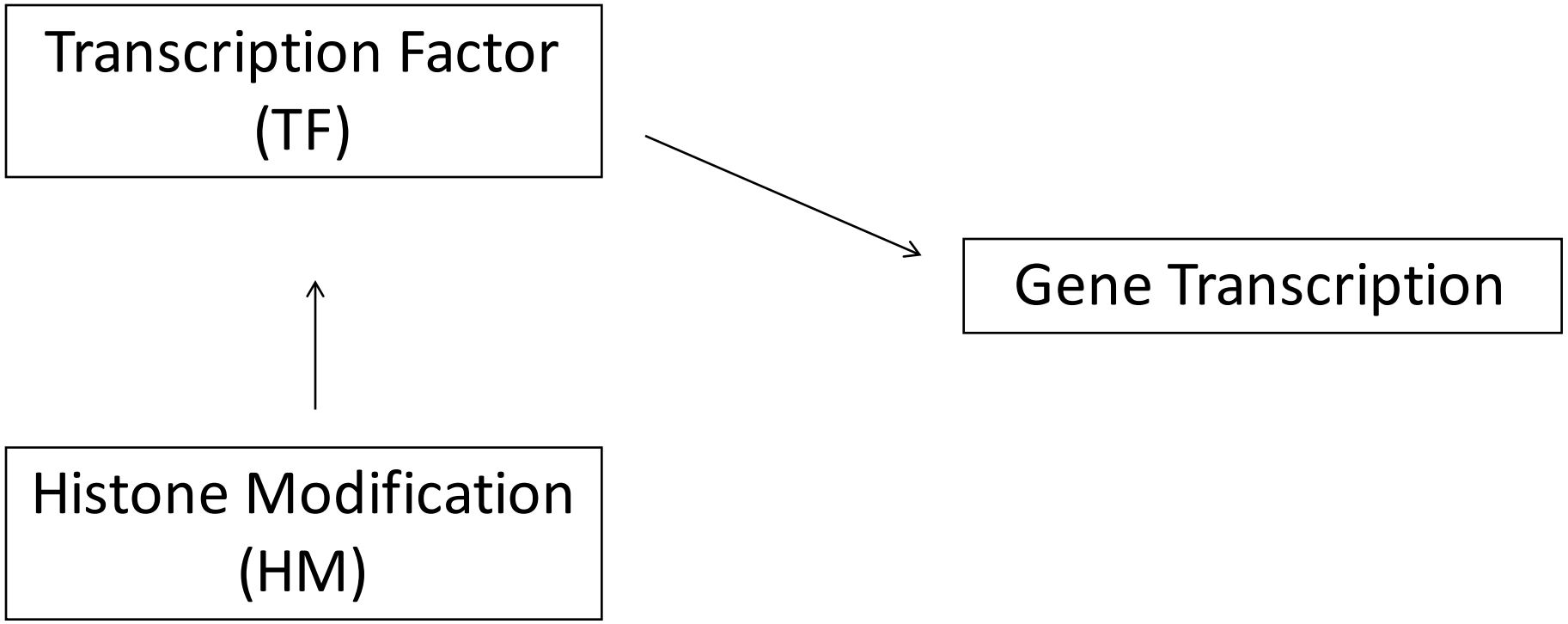


NO
(-1)

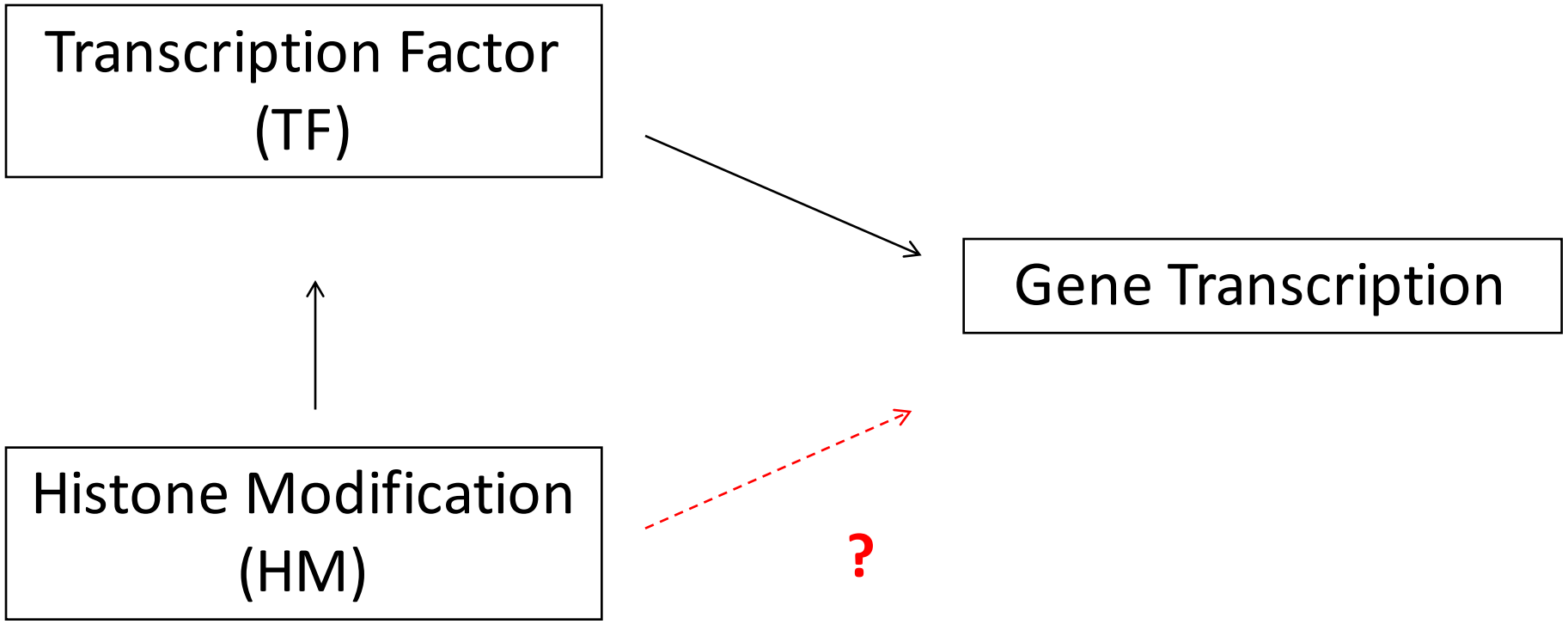
Many Important Data-Driven Computational Tasks



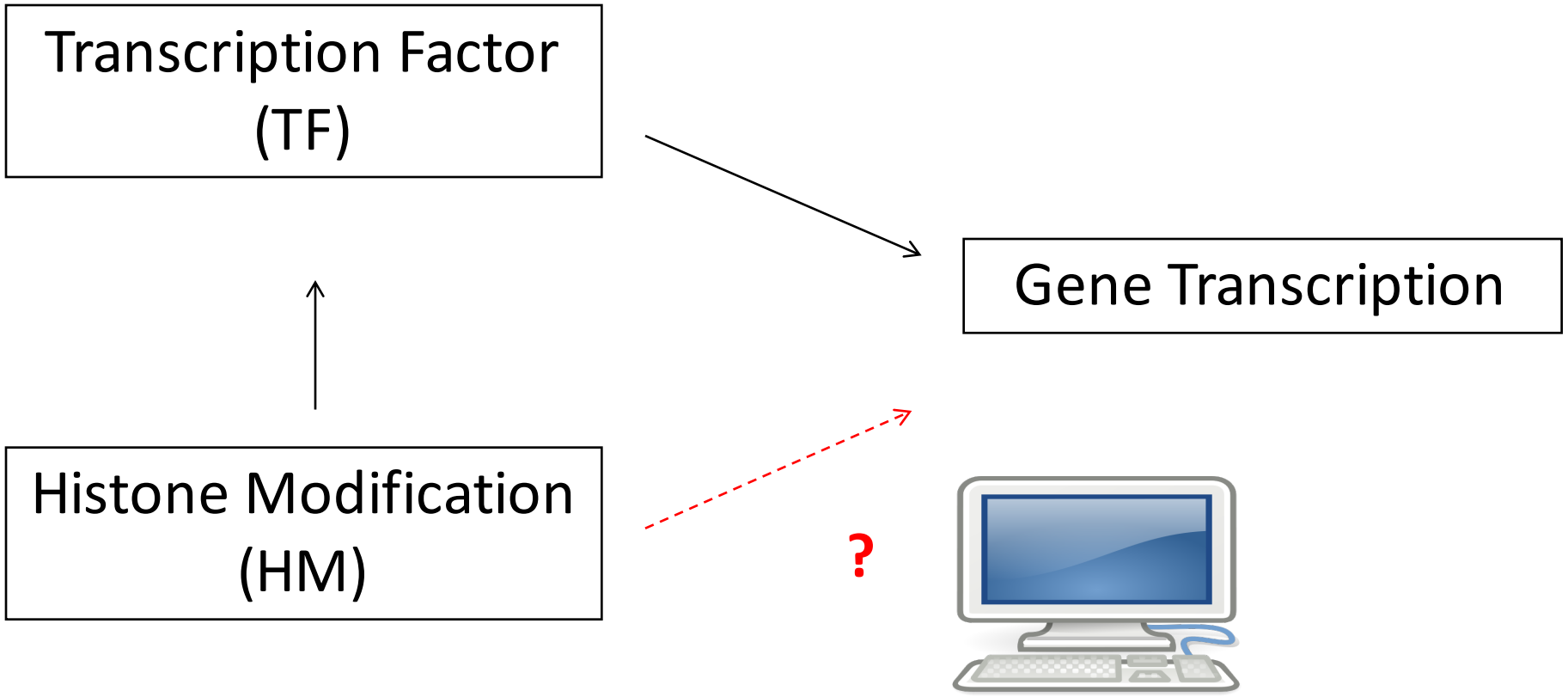
Histone Modification and Gene Transcription



Histone Modification and Gene Transcription



Histone Modification and Gene Transcription

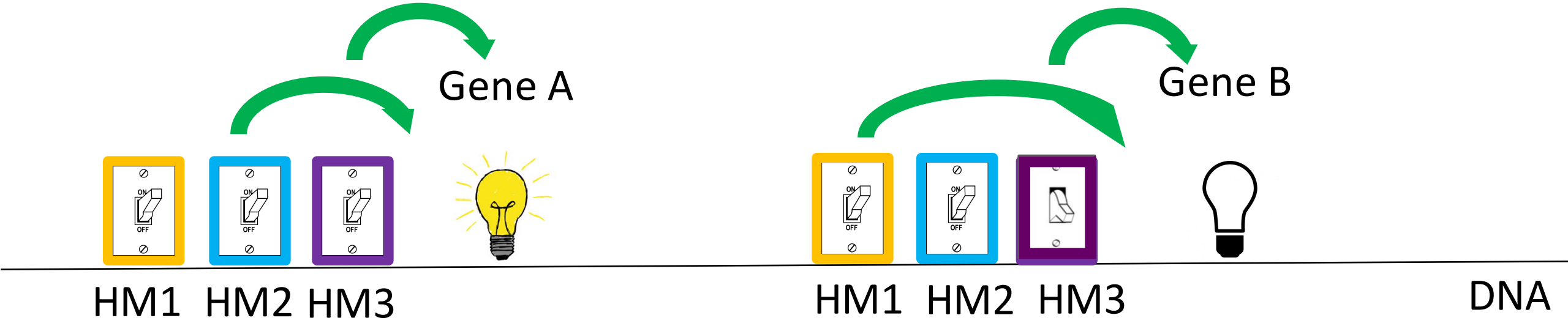


Why Studying [HM => Gene Expression] ?

- Epigenomics:
  - Study of chemical changes in DNA and histones (without altering DNA sequence)
  - Inheritable and involved in regulating gene expression, development, tissue differentiation and suppression …
- Modification in DNA/histones (changes in chromatin structure and function)
  - => influence how easily DNA can be accessed by TF
- Epigenome is dynamic
  - Can be altered by environmental conditions
  - Unlike genetic mutations, chromatin changes such as histone modifications are potentially reversible => Epigenome Drug for Cancer Cells?

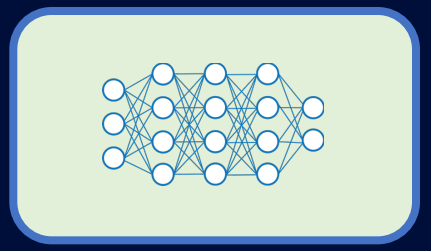
Study how HMs influence genes?

~56 Cell Types



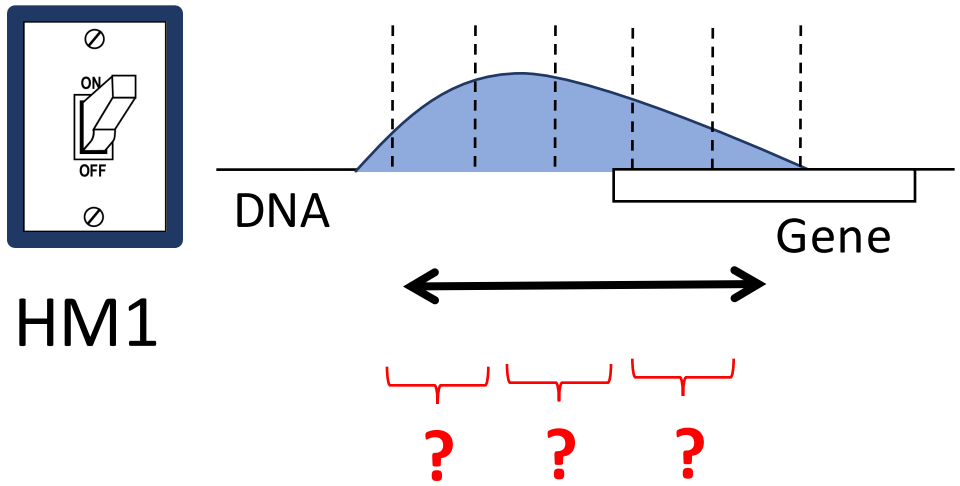
K~20 HMs

G~30,000 Genes

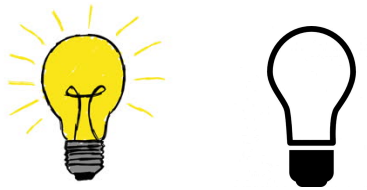


Task Formulation

Input:

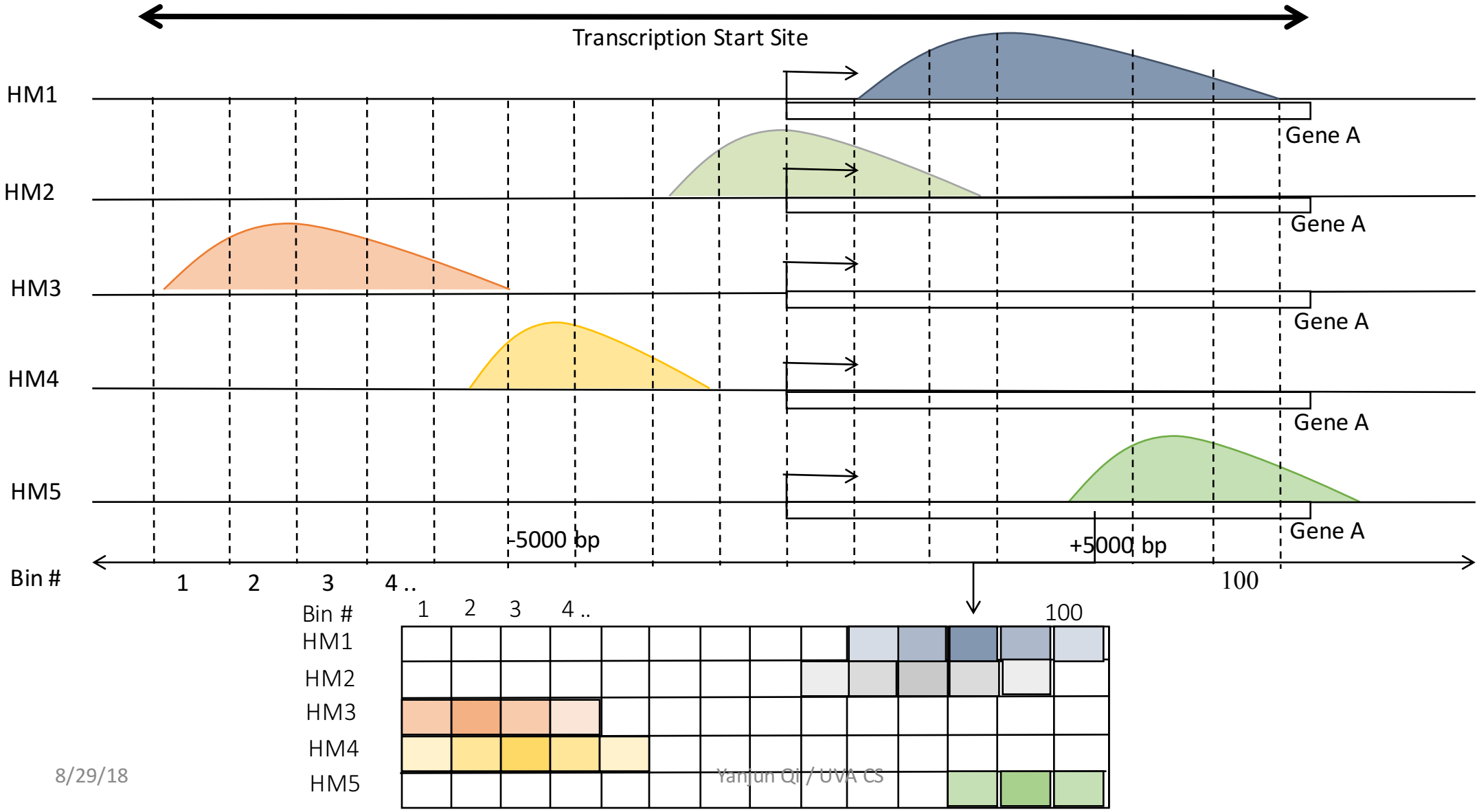


Output:

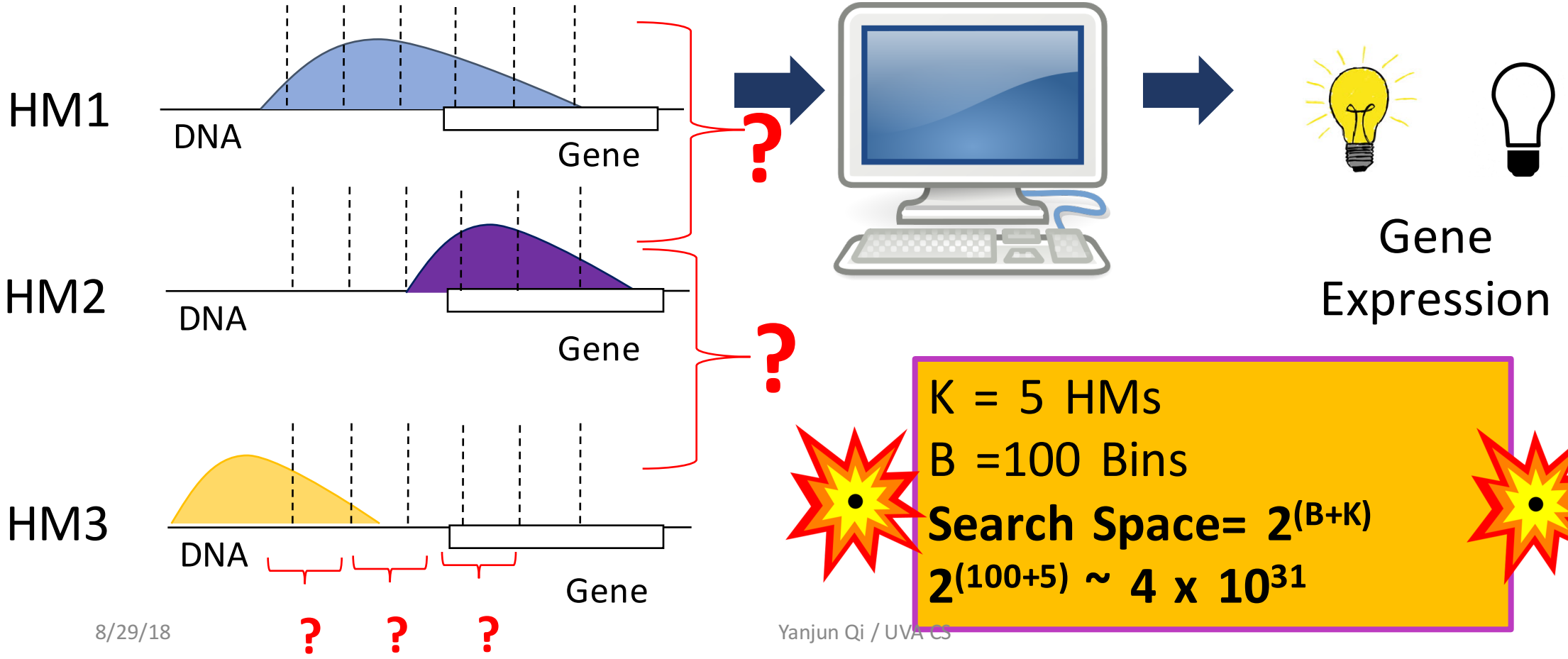


Gene Expression

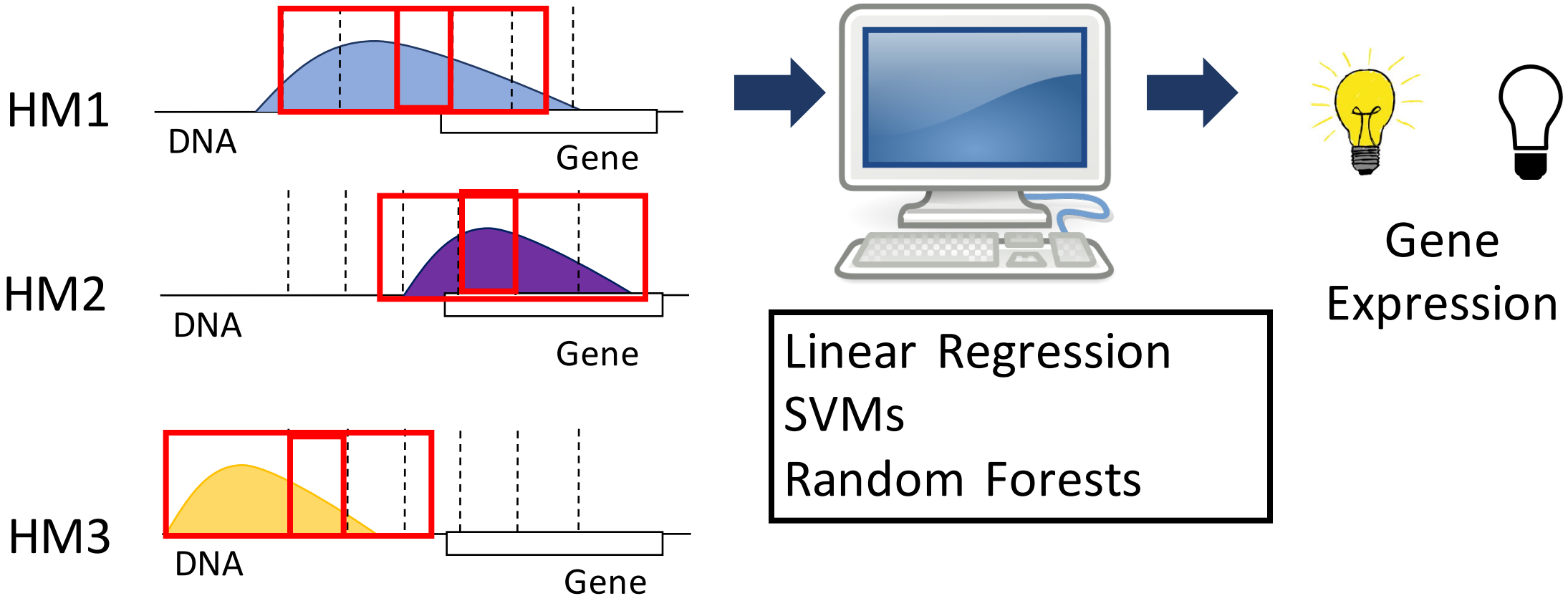
Input


Transcription Start Site
HM1
HM2
HM3
HM4
HM5
Gene A
Gene A
Gene A
Gene A
Gene A
-5000 bp
+5000 bp
Bin #
1
2
3
4 ..
100
Bin #
1
2
3
4 ..
100
HM1
HM2
HM3
HM4
HM5

Computational Challenge

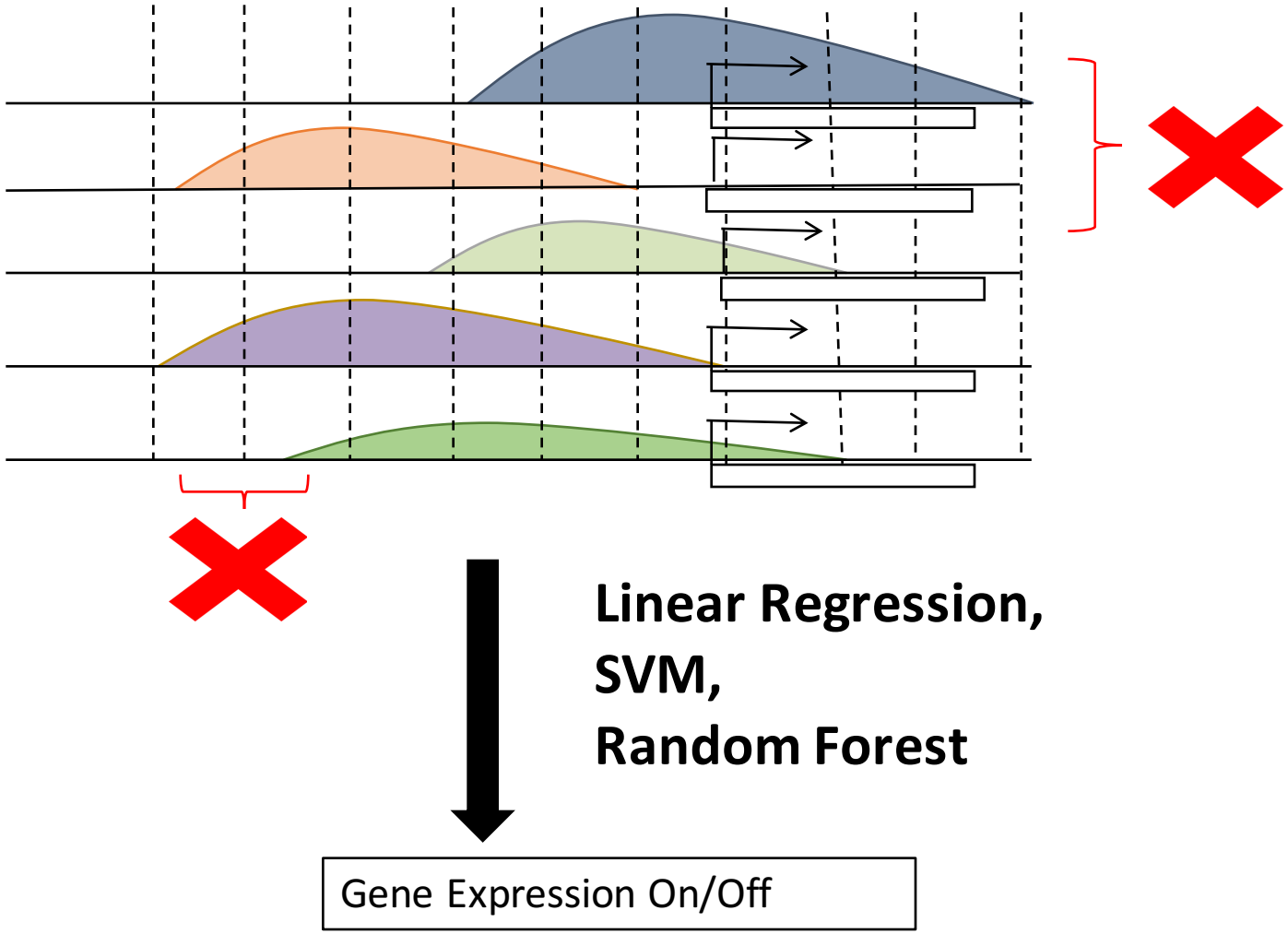


Related Work



[1] Karli´c, R. et al, Histone modification levels are predictive for gene expression. Proceedings of the National Academy of Sciences (2010)
[2] Cheng, C. et al, A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biology (2011)
[3] Dong, X. et al, Modeling gene expression using chromatin features in various cellular contexts. Genome Biology (2012)

Drawback of Related Works

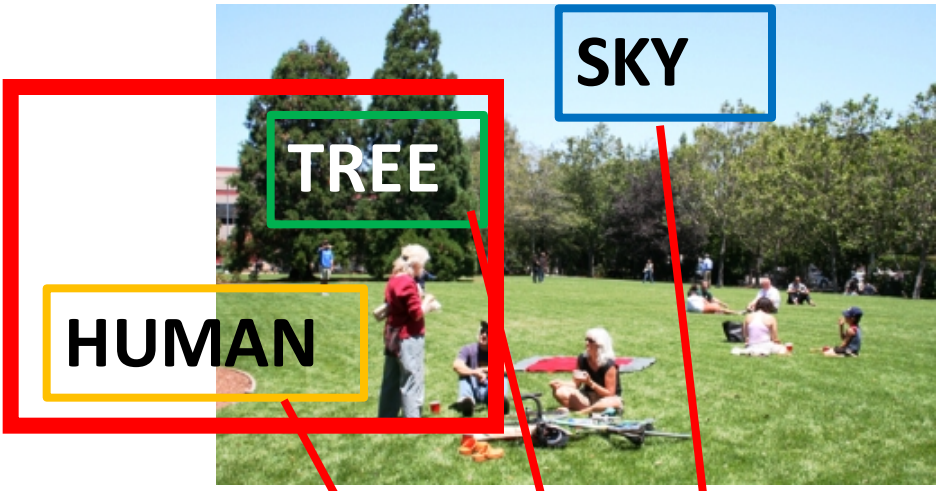


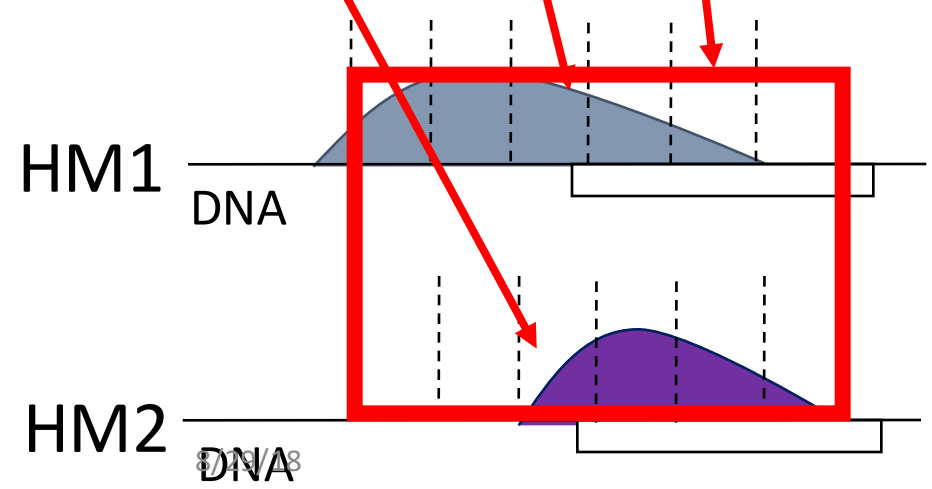
[1] Karli´c, R. et al, Histone modification levels are predictive for gene expression. Proceedings of the National Academy of Sciences (2010)
[2] Cheng, C. et al, A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biology (2011)
[3] Dong, X. et al, Modeling gene expression using chromatin features in various cellular contexts. Genome Biology (2012)

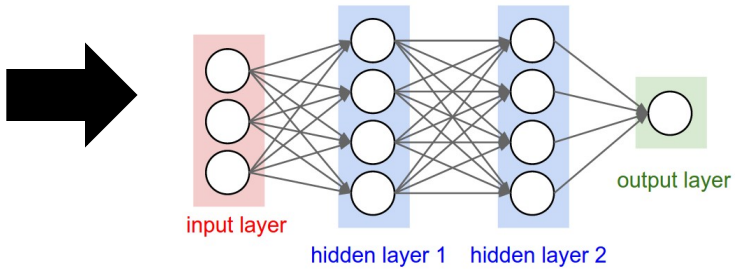
First Solution : CNN

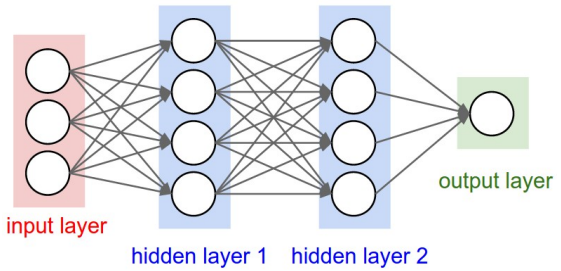
HM signals occupy a local region and look similar in different parts?

Input





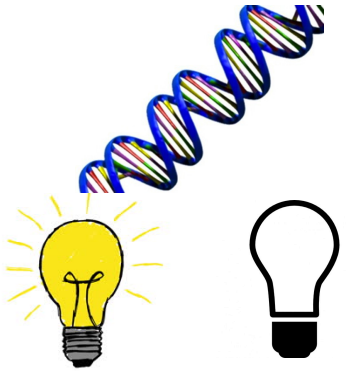




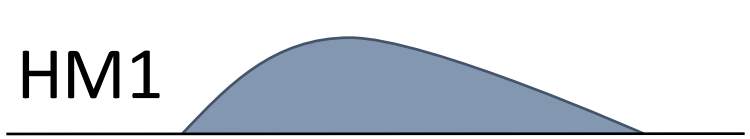
Output

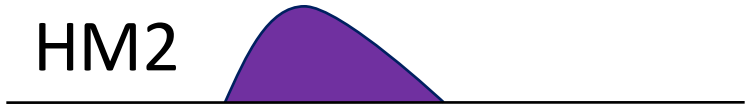
Park

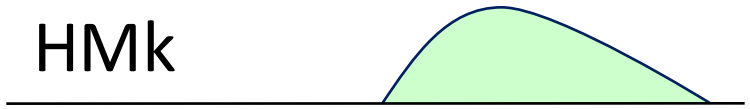
Gene

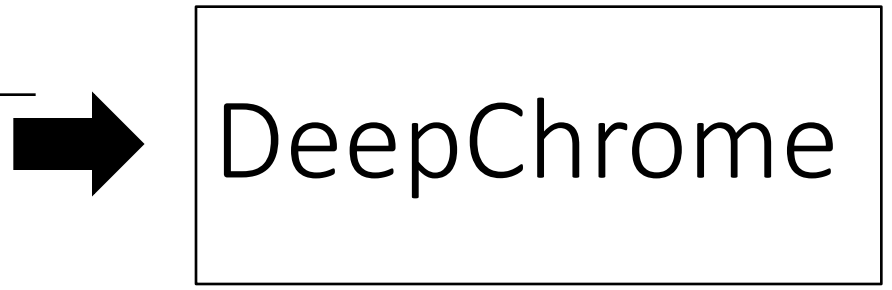


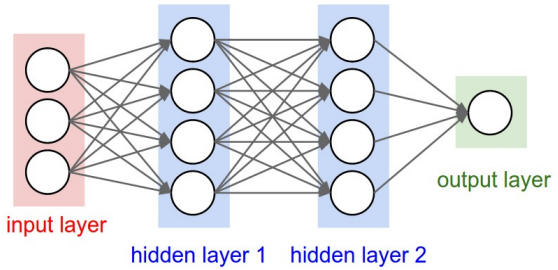
First Solution: DeepChrome : CNN

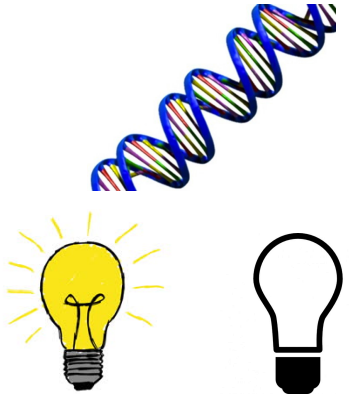




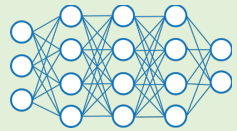




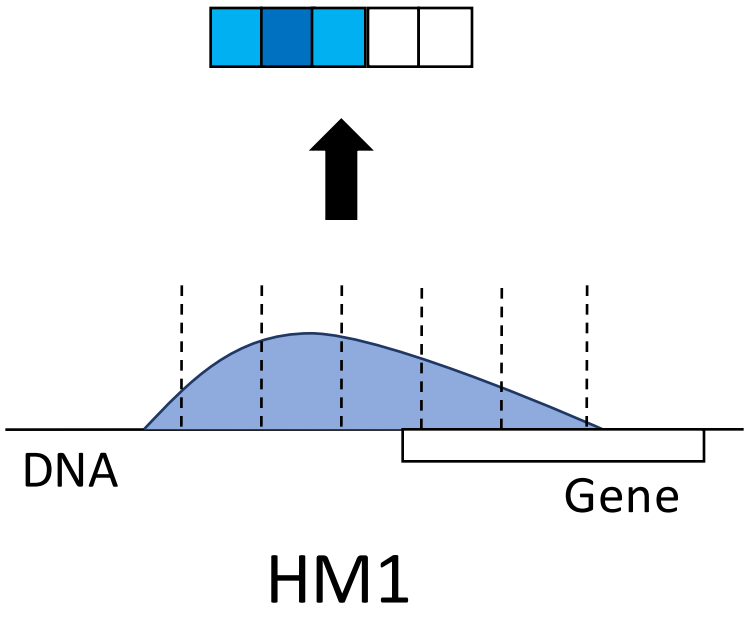


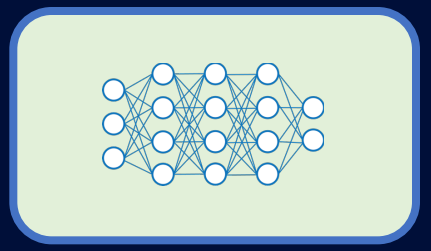


R. Singh, et al. "Deep-learning for predicting gene expression from histone modifications". *Bioinformatics.* (ECCB) (2016)

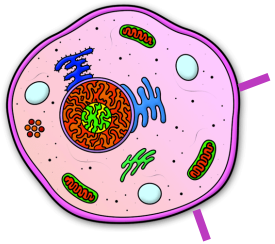


Input (X)

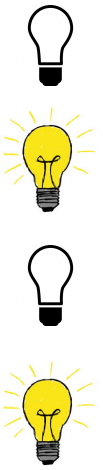




Output (Y) Labels

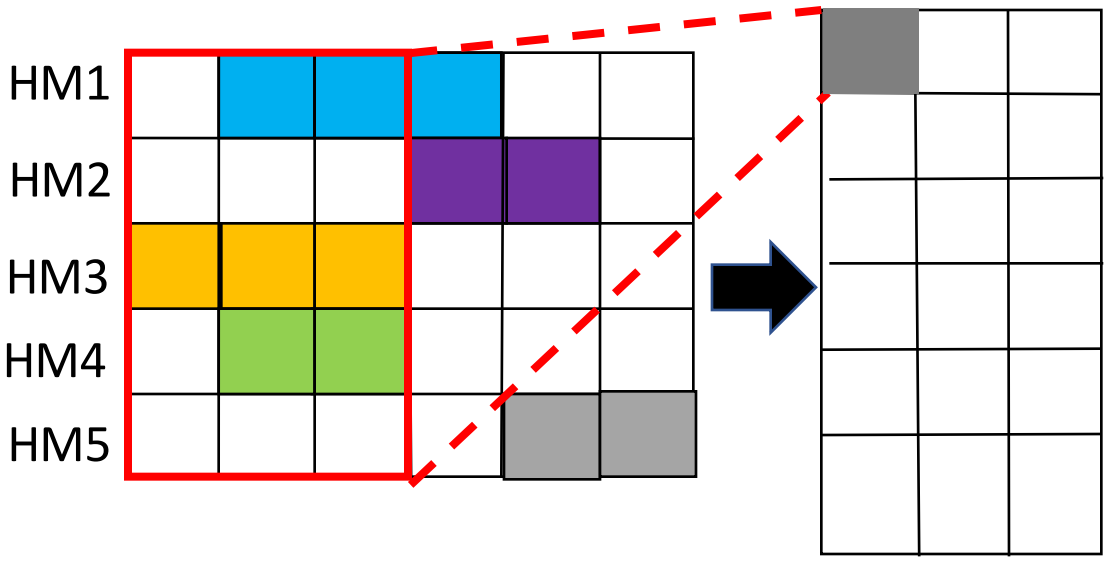


Genes	Gene Expression (RPKM)	Y Labels
RUNX1	1.296	0
SMAD2	14.902	1
MYC	3.805	0
PAX5	15.066	1
……	…….	……



Threshold = 10.245 (Median)

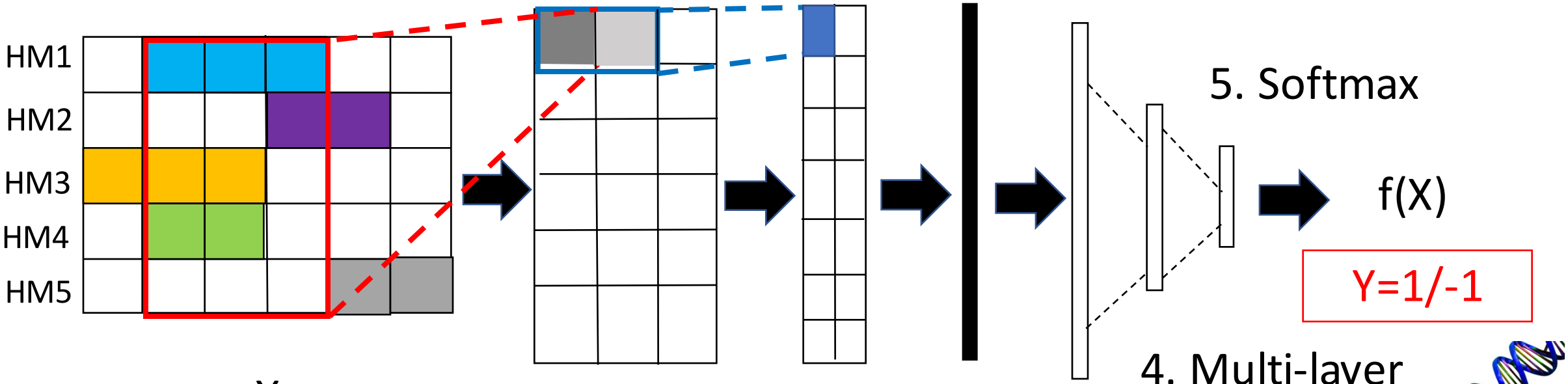
DeepChrome: Convolutional Neural Network (CNN)



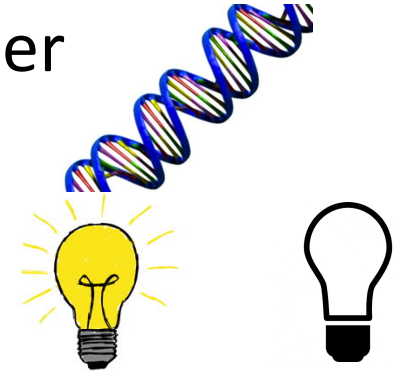
X

1. Convolution

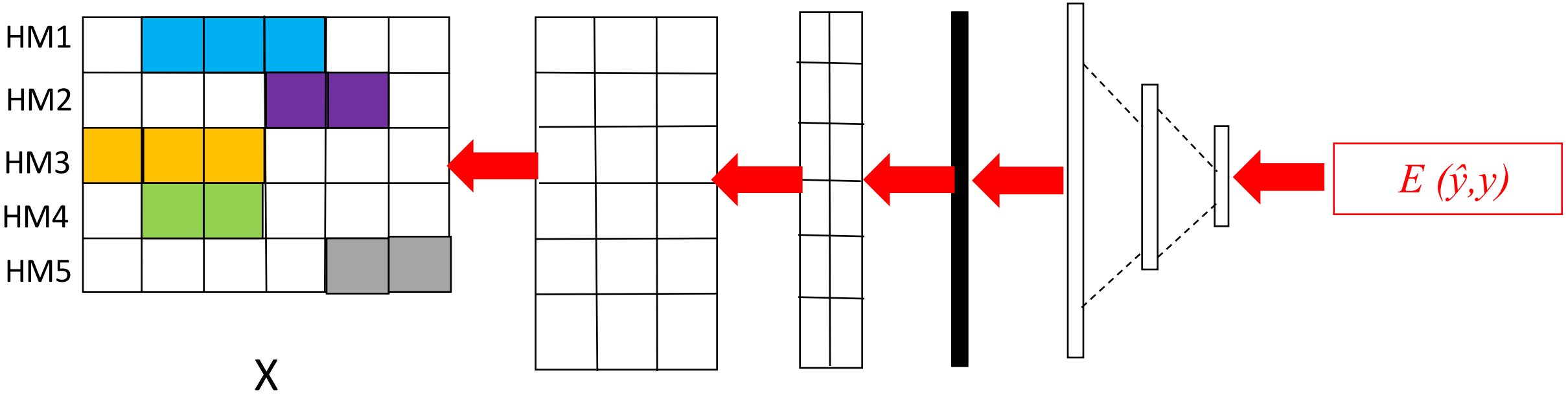
DeepChrome: Convolutional Neural Network (CNN)



$$E = \sum_{n=1}^{N_{samp}} loss(f(X^{(n)}), y^{(n)})$$



DeepChrome: Convolutional Neural Network (CNN)



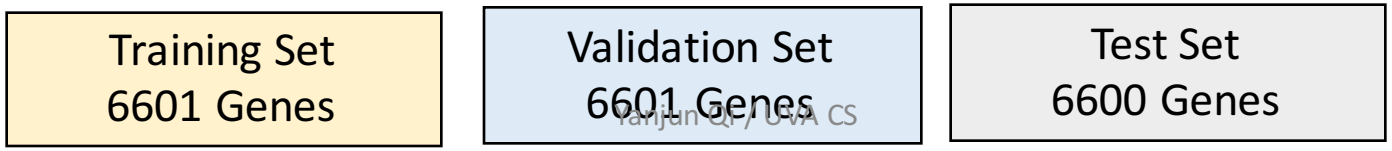
Back-propagation: $\Theta \leftarrow \Theta - \eta \frac{\partial E}{\partial \Theta}$

Experimental Setup

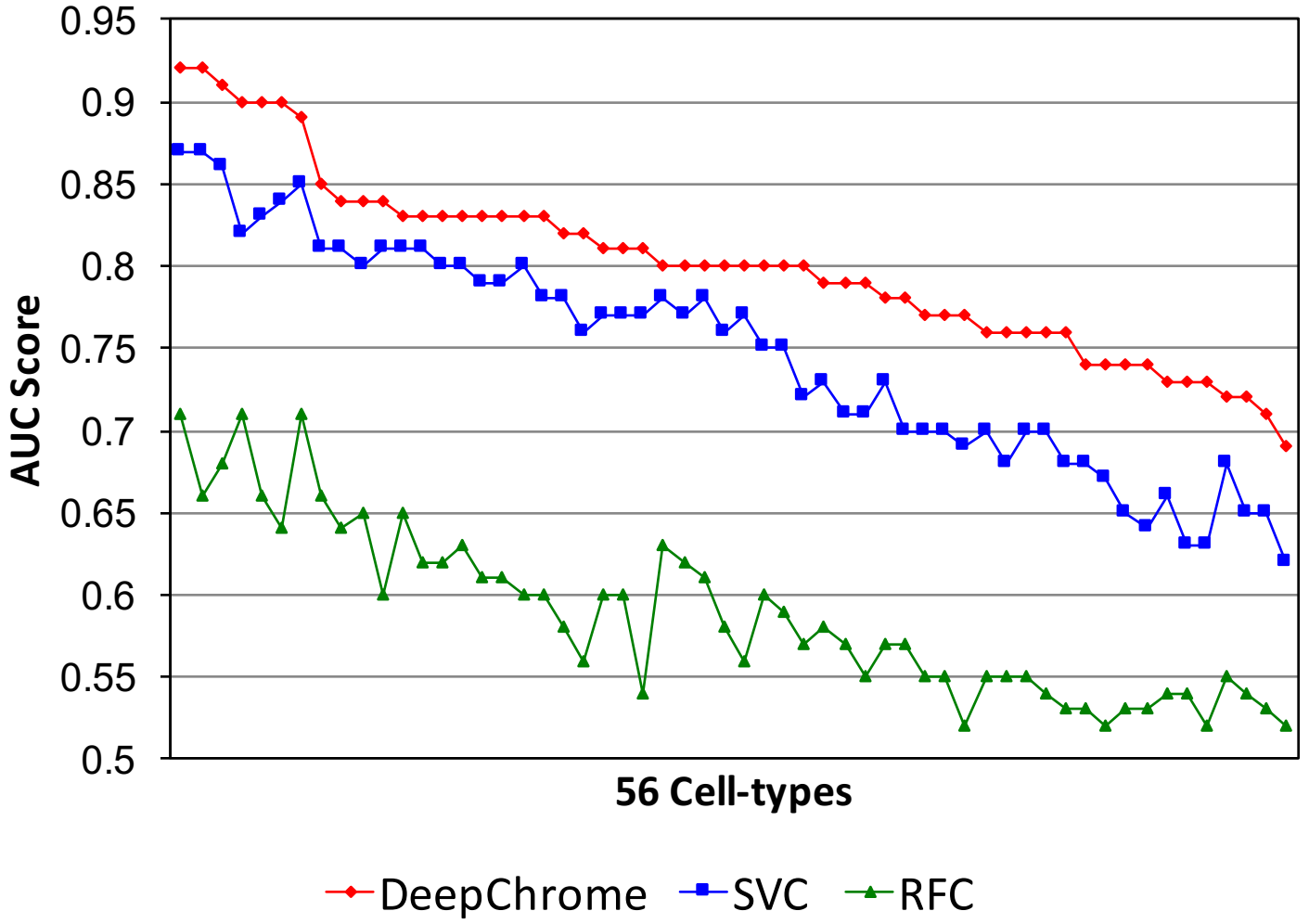
- Roadmap Epigenetics Project (REMC)
- **Cell-types:** 56
- **Input (HM):** ChIP-Seq Maps / 5 Tier-1 HMs

Histone Mark	Functional Category
H3K27me3	Repressor
H3K36me3	Structural Promoter
H3K4me1	Distal Promoter
H3K4me3	Promoter
H3K9me3	Repressor

- **Output (Gene Expression):** Discretized RNA-Seq
- **Baselines:** Support Vector Classifier (SVC) and Random Forest Classifier (RFC)

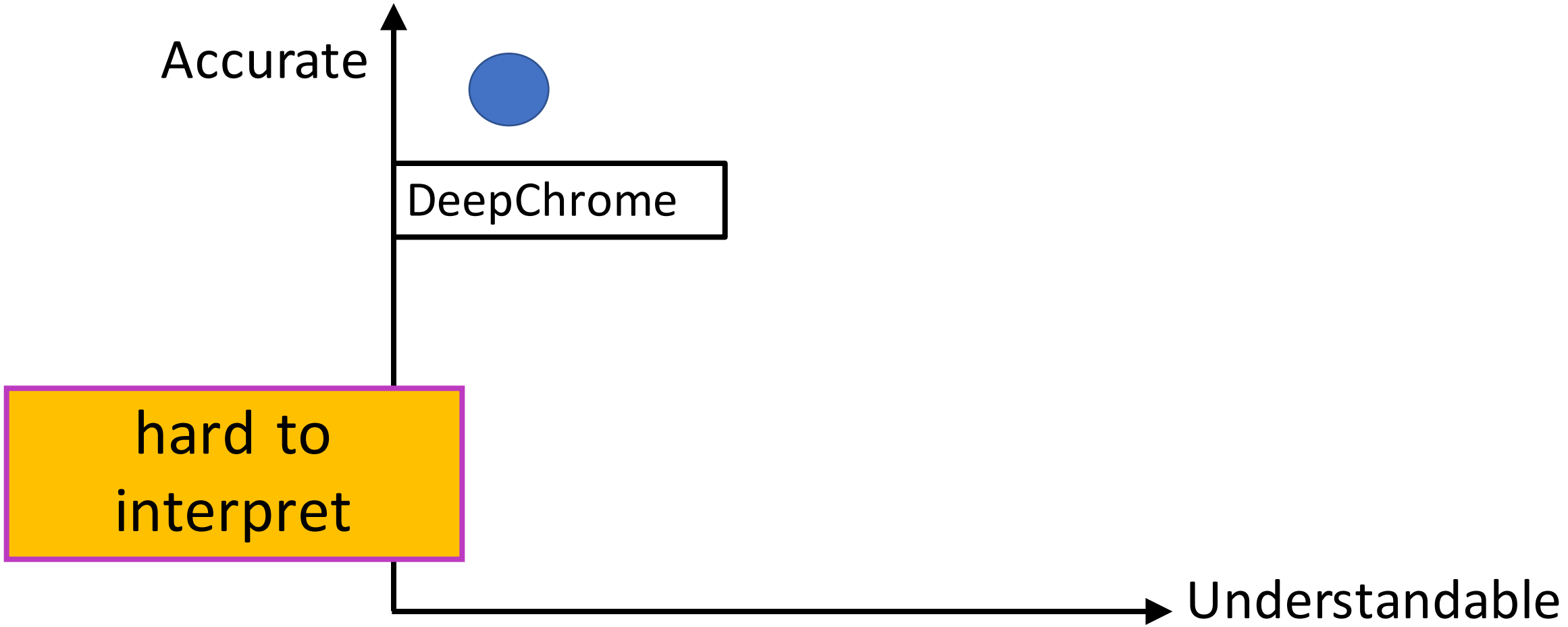


Results: Accuracy



https://www.deepchrome.org

Summary of tools

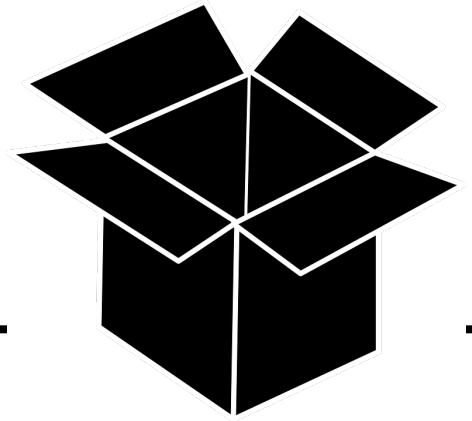


R. Singh, et al. Deep-learning for predicting gene expression from histone modifications". *Bioinformatics.* (2016)

Solution: Interpretability by Hierarchical Attention

Input

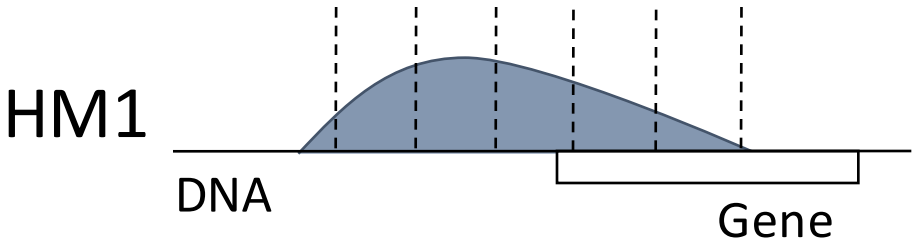
Attention Mechanism

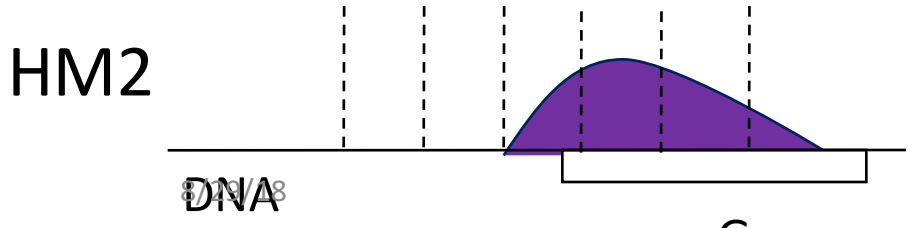


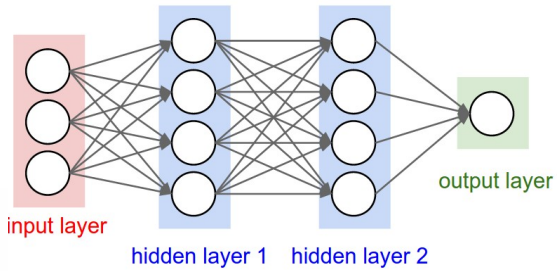
Output

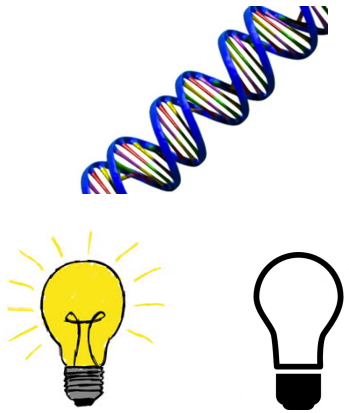
Park

Gene









Solution: Interpretability by Hierarchical Attention

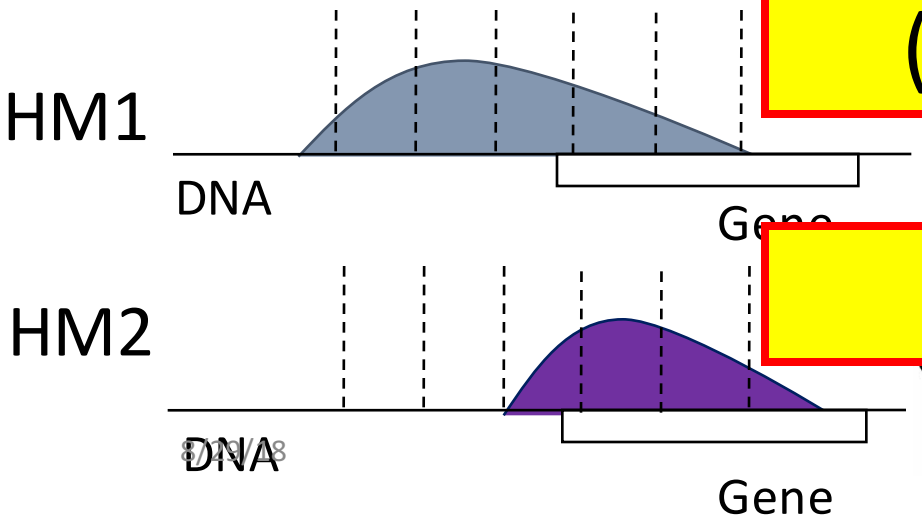
Input

Output

Attention Mechanism

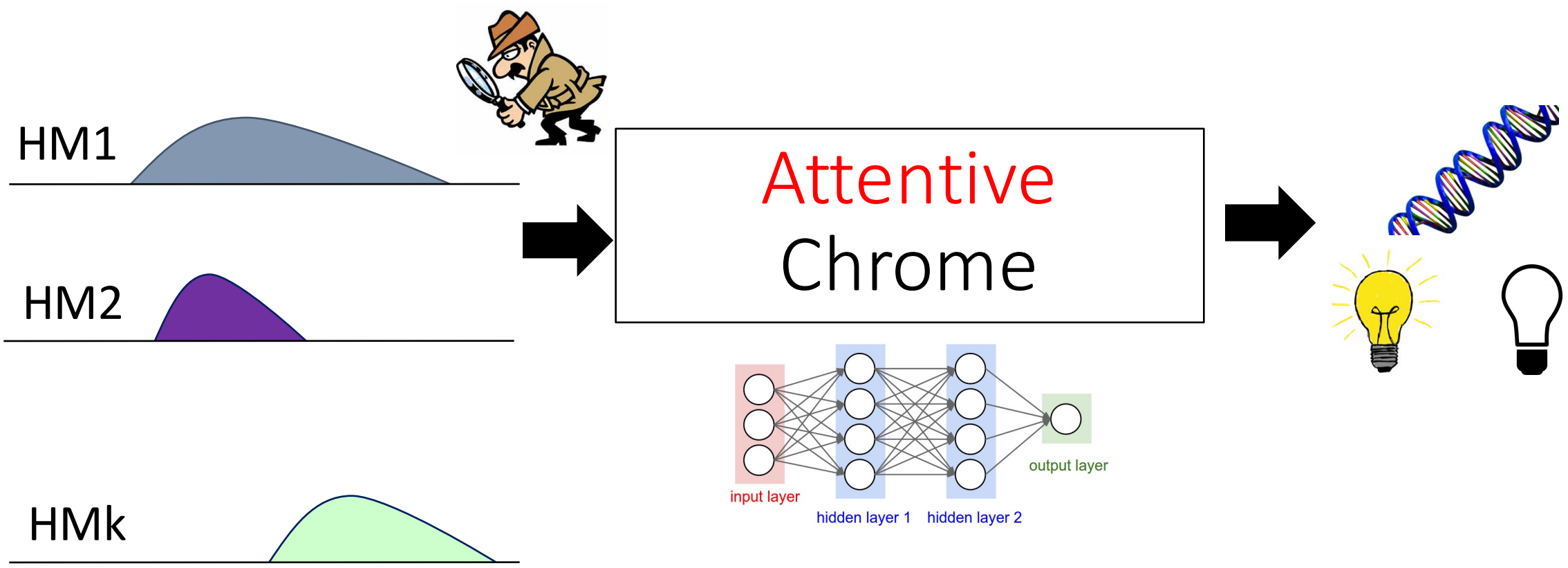
Park

Gene



(1) What positions are important?

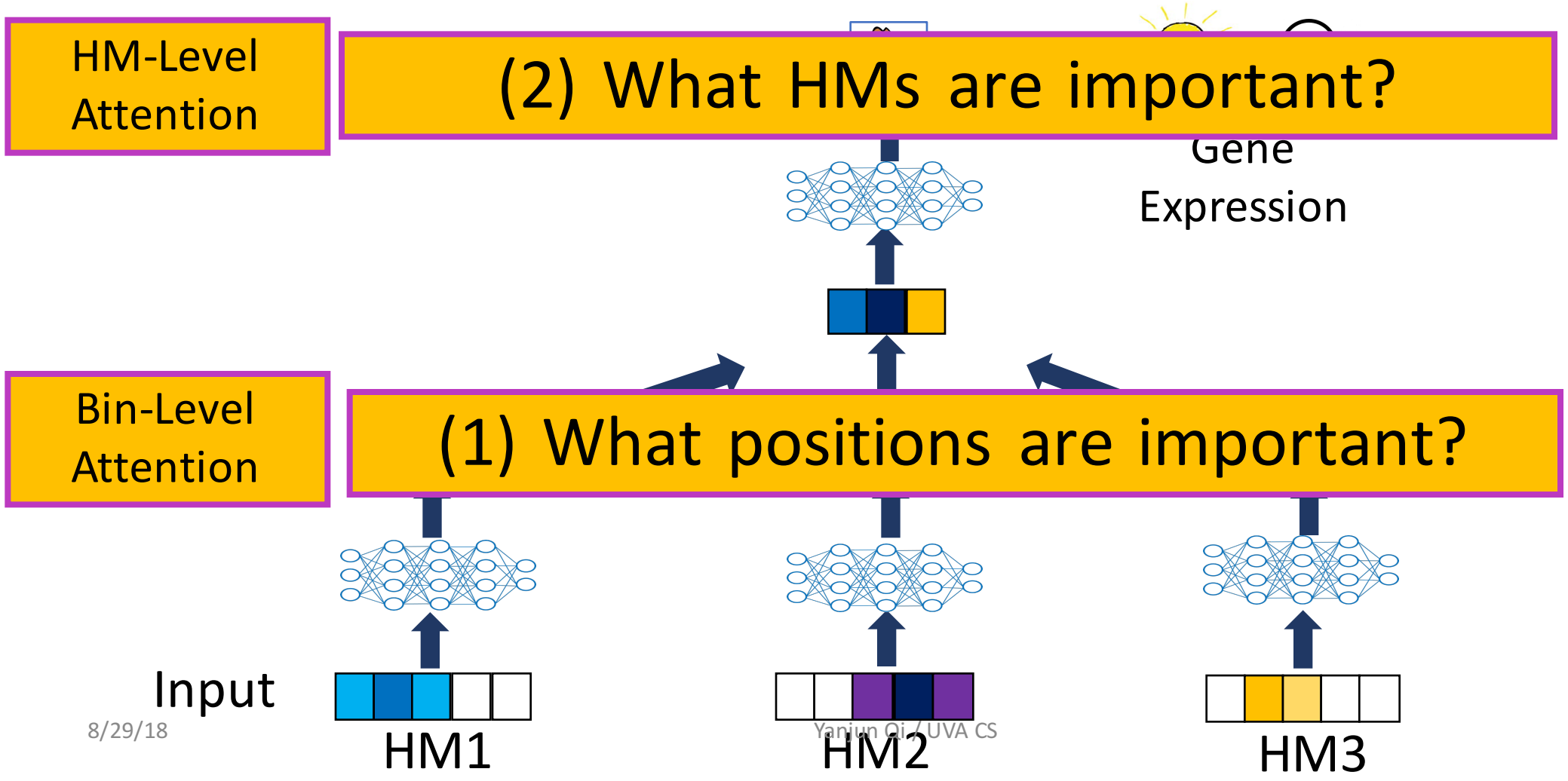
(2) What HMs are important?



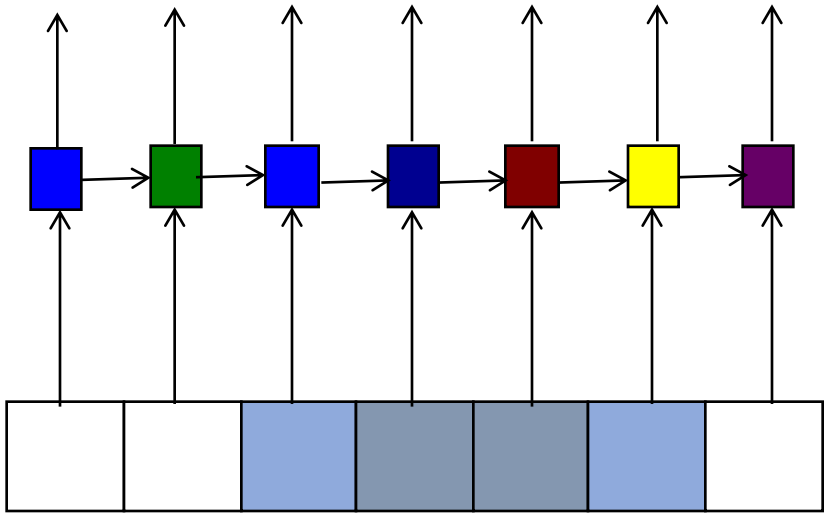
R. Singh, et al. "Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin". *NIPS (2017)*

AttentiveChrome

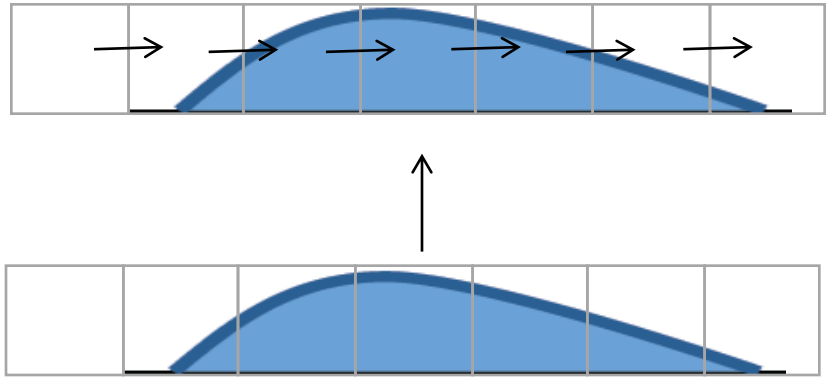
[NIPS 2017]



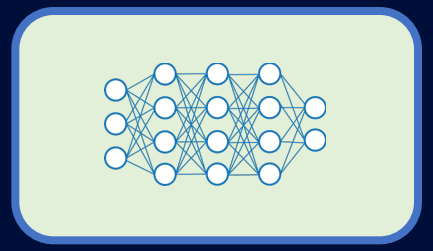
Multiple Recurrent Neural Networks (Hierarchical RNNs) to model each HM and the Combination of all HMs : for example on HM1



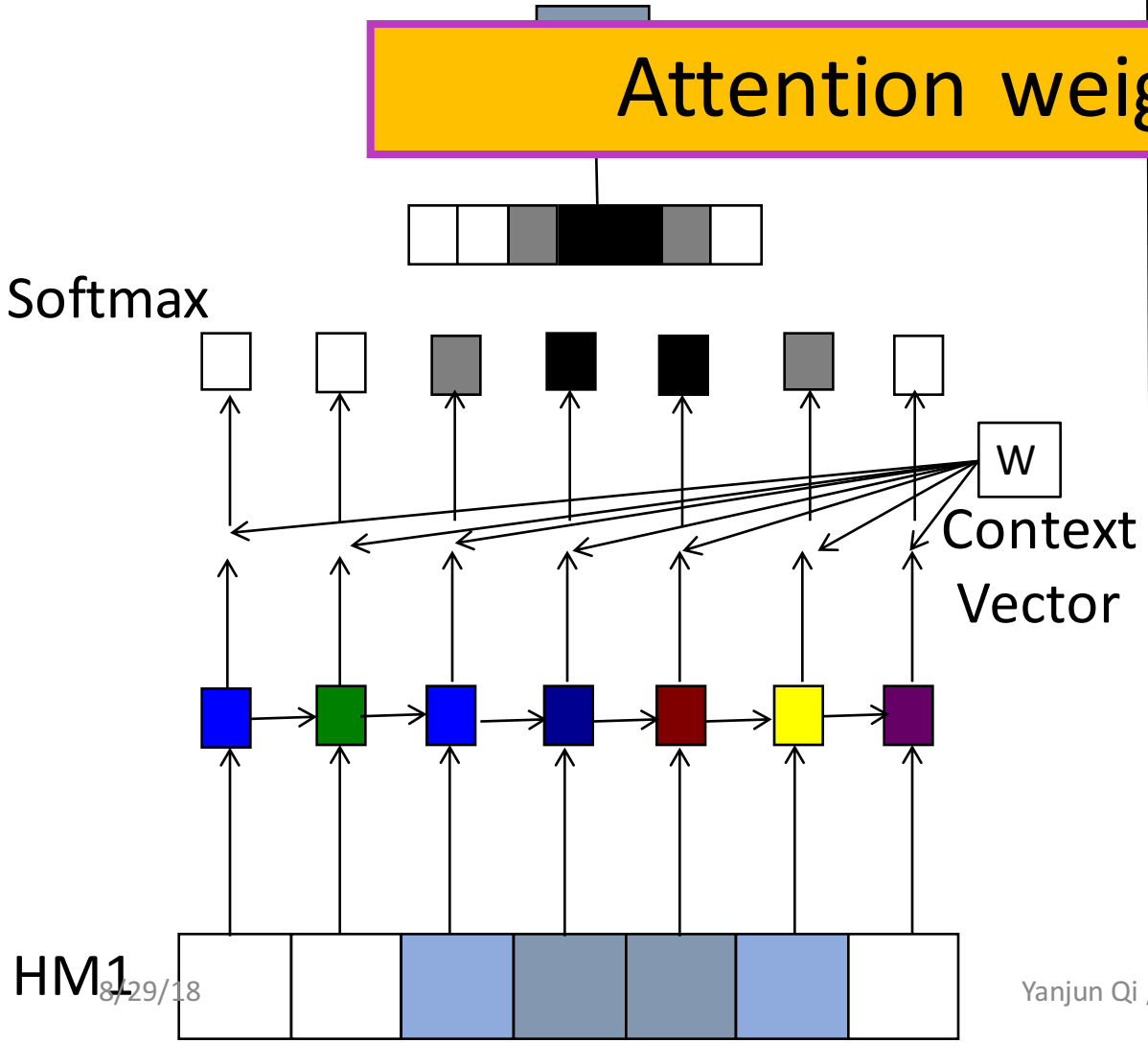
HM1

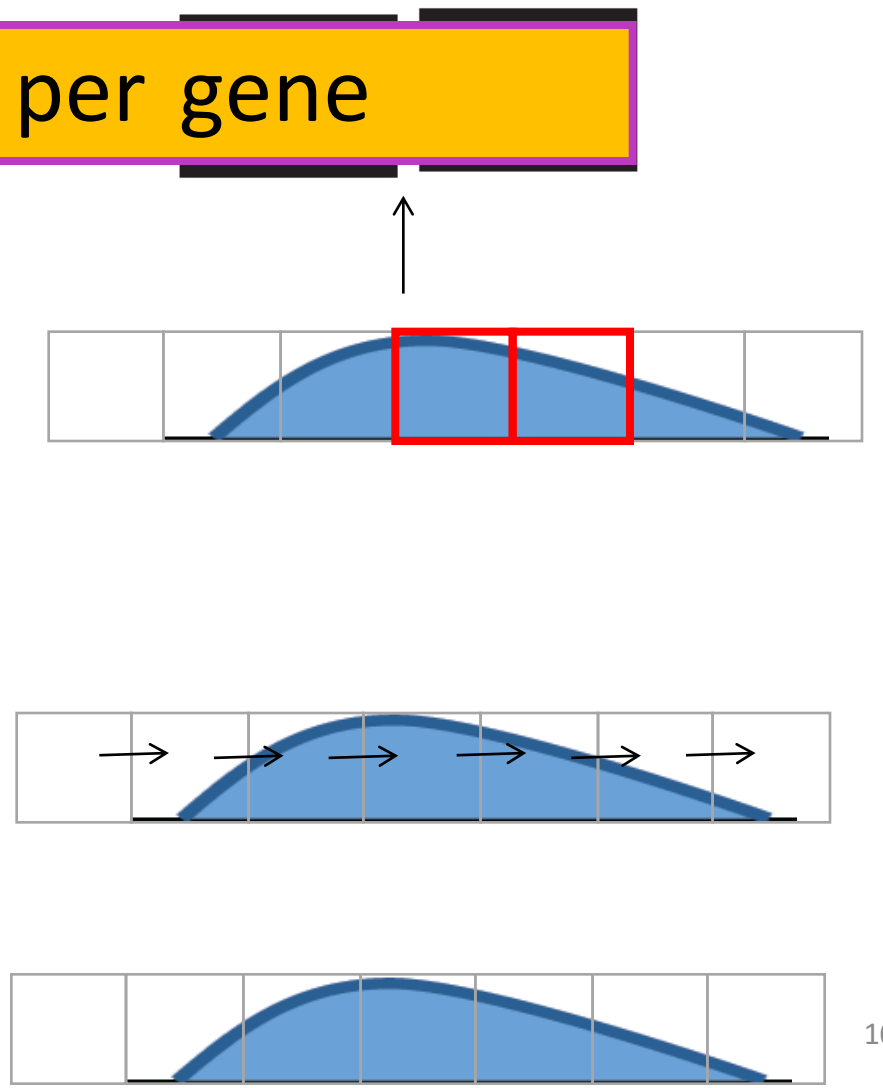


Attention Mechanism

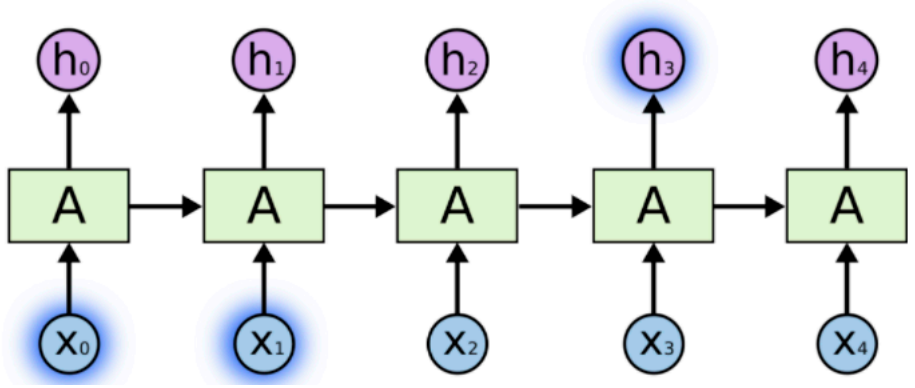


Attention weights per gene





Using Attention to Select RNN per-unit outputs



$$h_t = f_W(h_{t-1}, x_t)$$

new state

old state

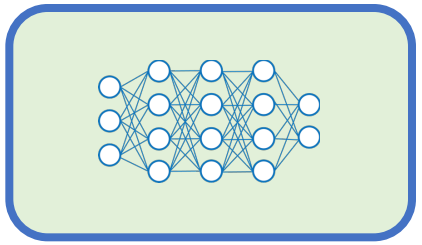
input vector at some time step

some function with parameters W

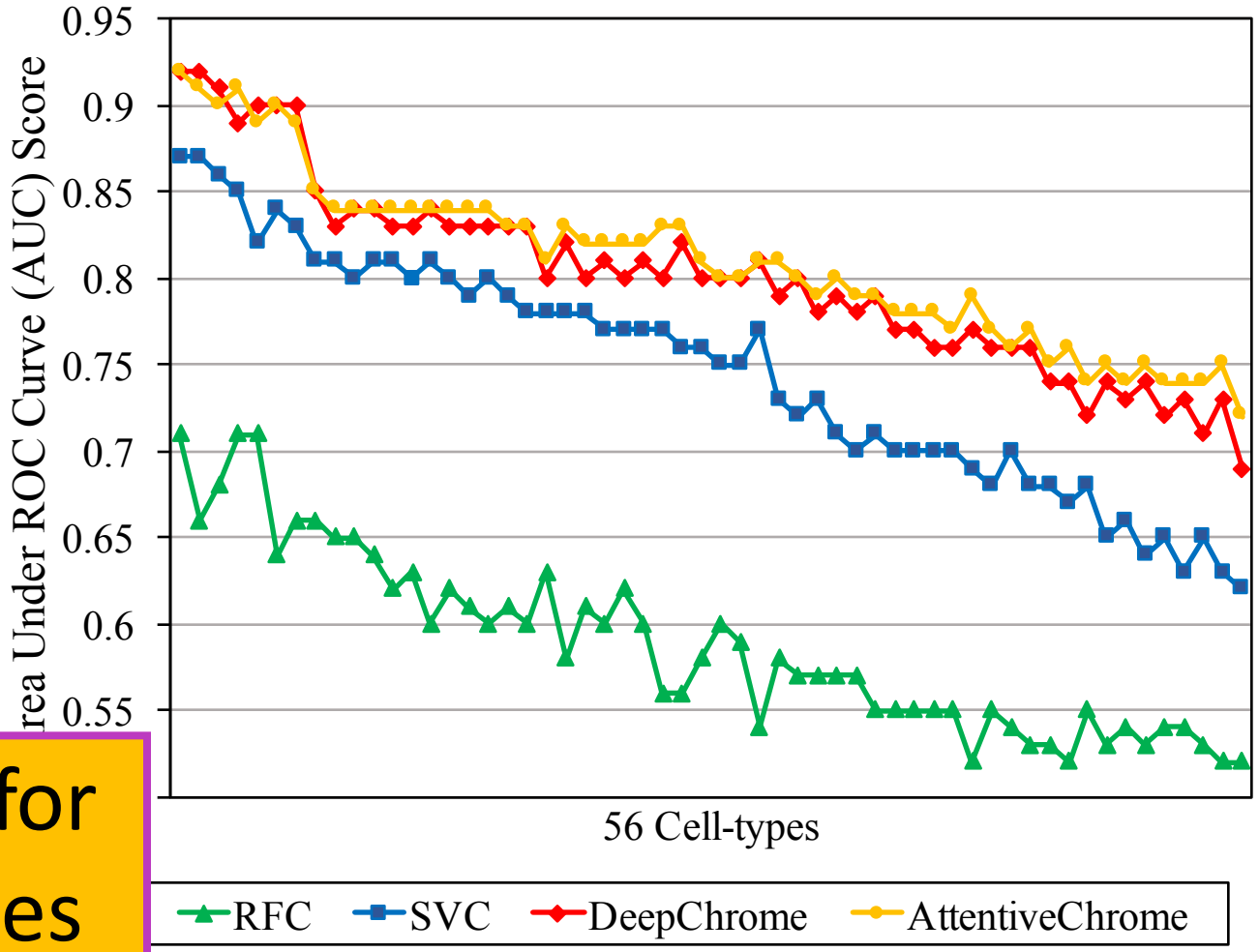
$$\alpha_t^j = \frac{exp(\mathbf{W}_b \mathbf{h}_t^j)}{\sum_{i=1}^{T} exp(\mathbf{W}_b \mathbf{h}_i^j)}$$

W_b is learned

Image Credits from Christopher Olah



Prediction



Improvement for 49/56 Cell-types

Bin-Level Visualization

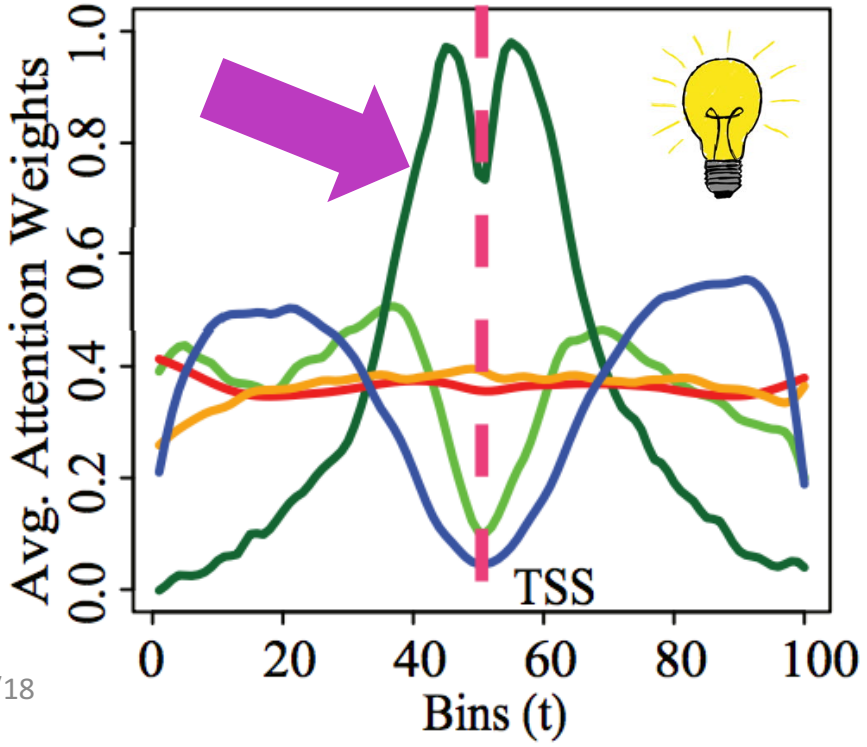
(1) What positions are important?

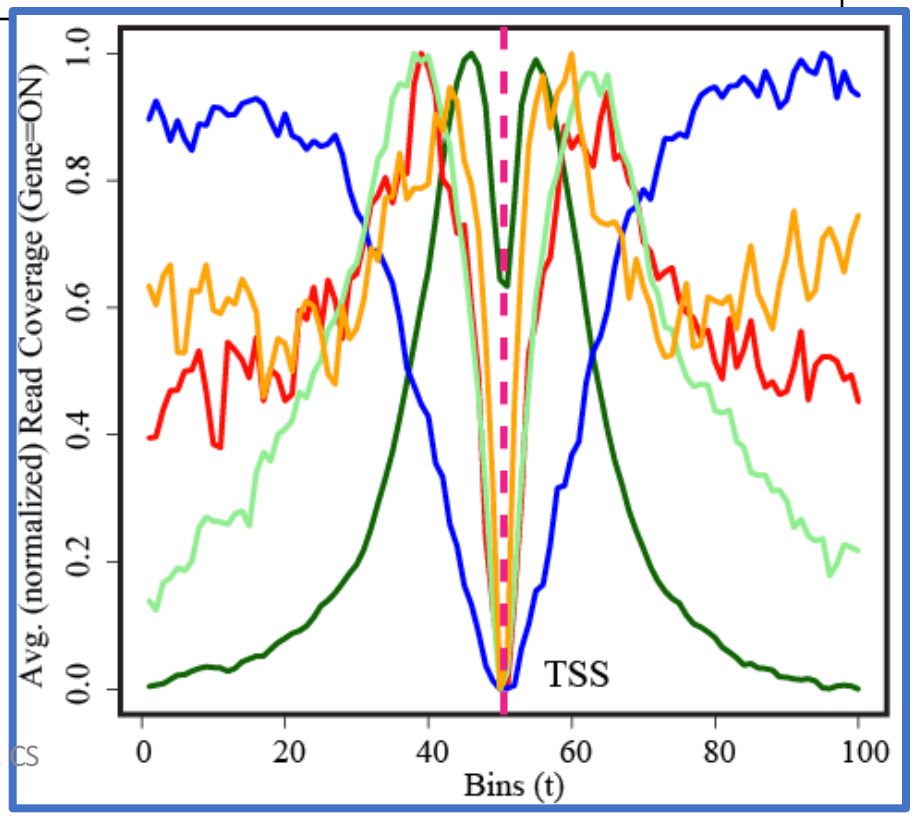
CELL TYPE: GM1

Repressors

Promoters

H3K27me3 H3K9me3 H3K4me1 H3K4me3 H3K36me3

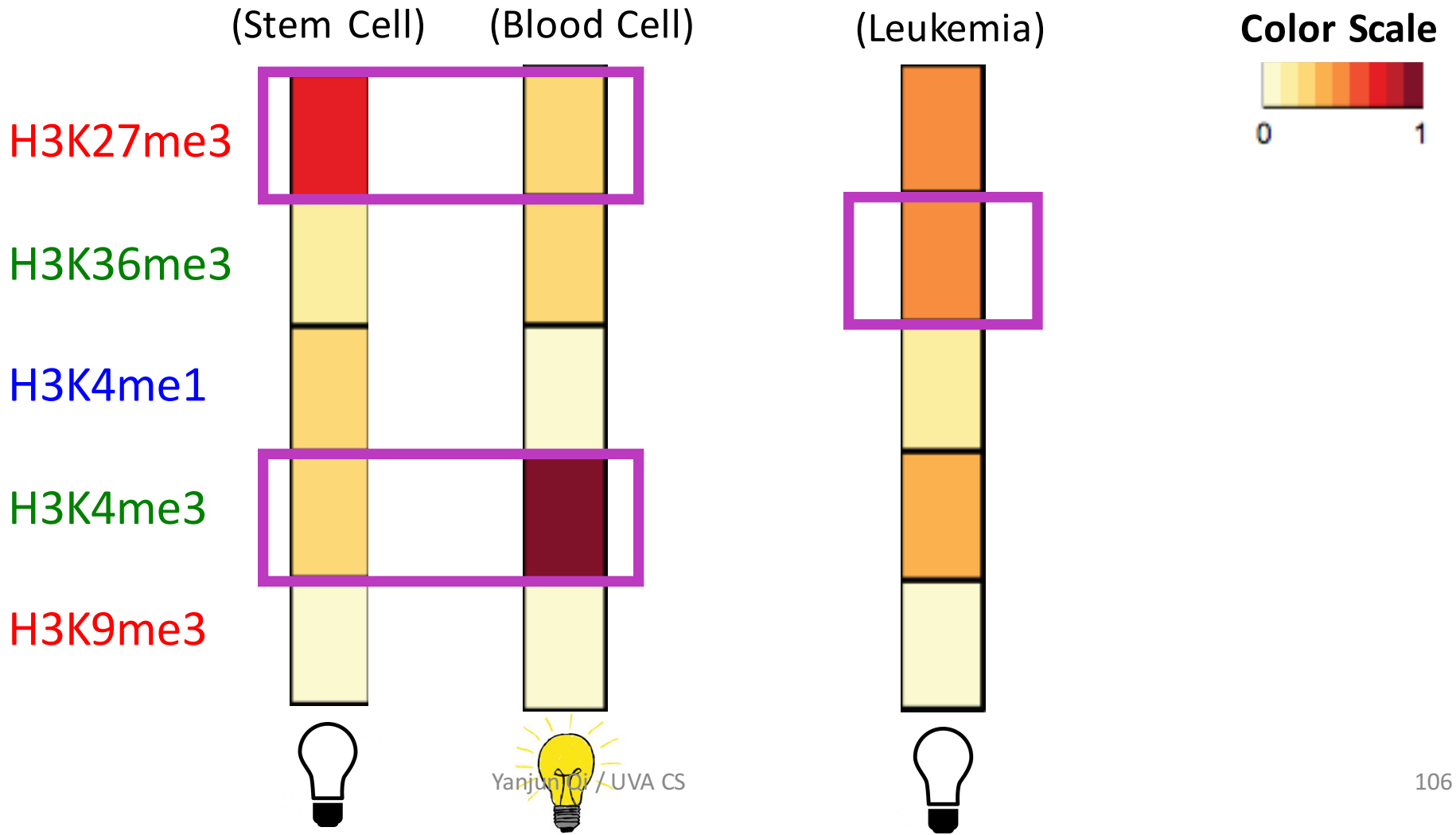




HM-Level Visualization

(2) What HMs are important?

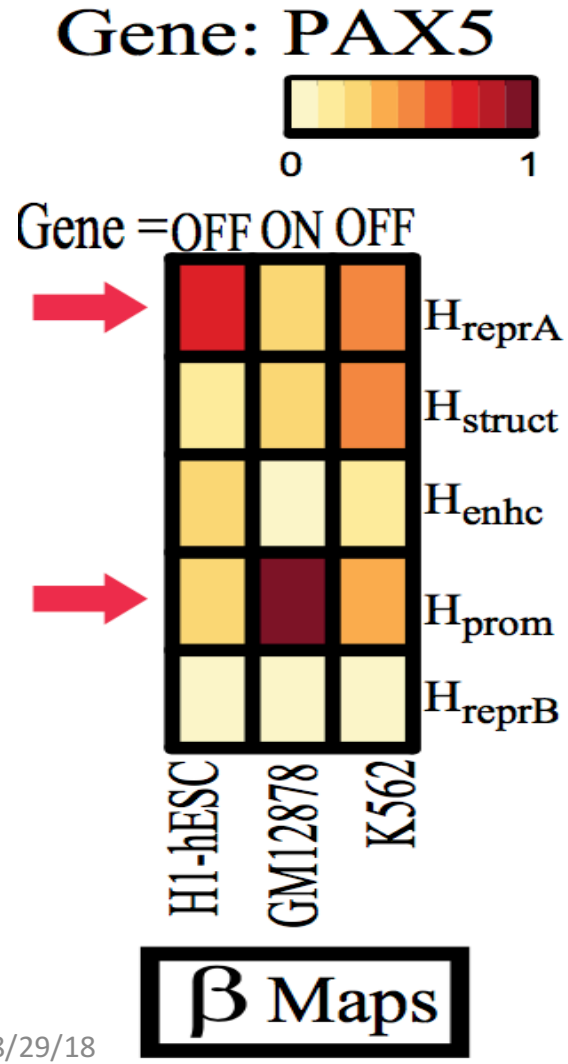
Cell Types:



PROMOTER
DISTAL PROMOTER
REPRESSOR

Gene: PAX5

Results: HM level attention



- An important differentially regulated gene (PAX5) across three blood lineage cell types:
  - H1-hESC (stem cell),
  - GM12878 (blood cell),
  - K562 (leukemia cell).
- Trend of its global weights (beta) Verified through the literature.

(2) What HMs are important?

Validation of Attention Weights (using one extra HM signals)

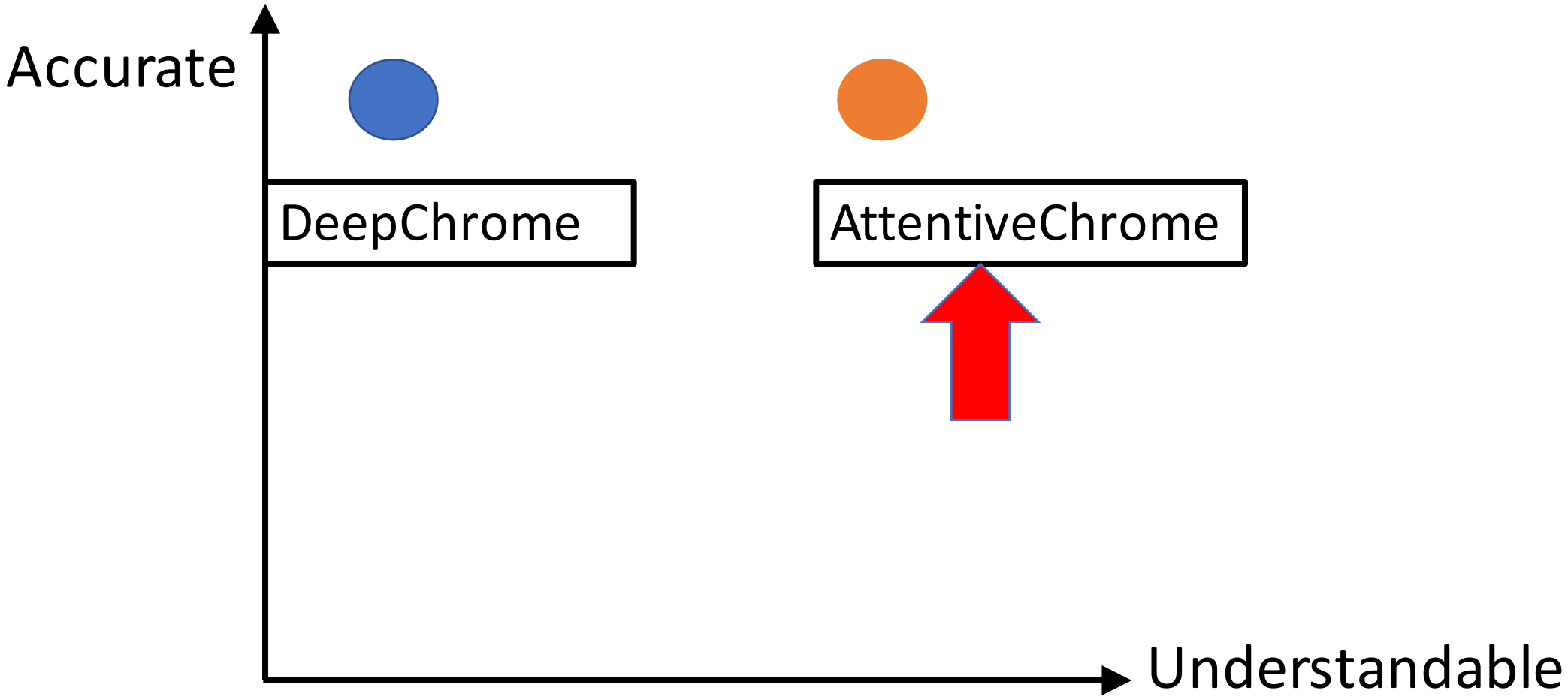
Table 3: Pearson Correlation values between weights assigned for H_{prom} (active HM) by different visualization techniques and H_{active} read coverage (indicating actual activity near "ON" genes) for predicted "ON" genes across three major cell types.

Viz. Methods	H1-hESC	GM12878	K562
α Map (LSTM-α)	0.8523	**0.8827**	**0.9147**
α Map (LSTM-α, β)	**0.8995**	0.8456	0.9027
Class-based Optimization (CNN)	0.0562	0.1741	0.1116
Saliency Map (CNN)	0.1822	-0.1421	0.2238

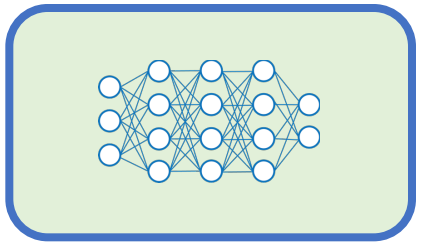
- Additional signal - H3K27ac (H-Active) from REMC
- Average local attention weights of gene=ON correspond well with H-active
- Indicating AttentiveChrome is focusing on the correct bin positions

https://www.deepchrome.org

Summary of tools

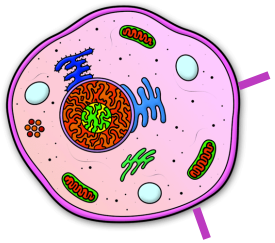


R. Singh, et al. "Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin". *NIPS (2017)*



Where are we heading?

Changing Task : Classification →Regression



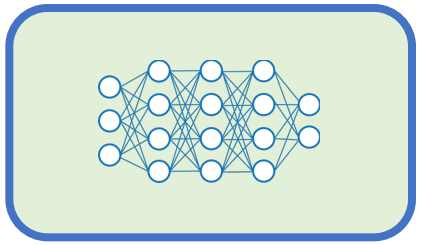
Genes	Gene Expression (RPKM)	Y log(RPKM)
RUNX1	1.296	01126
SMAD2	14.902	1.1737
MYC	3.805	0.5803
PAX5	15.066	1.779
……	…….	……

1.770

Gene Expression

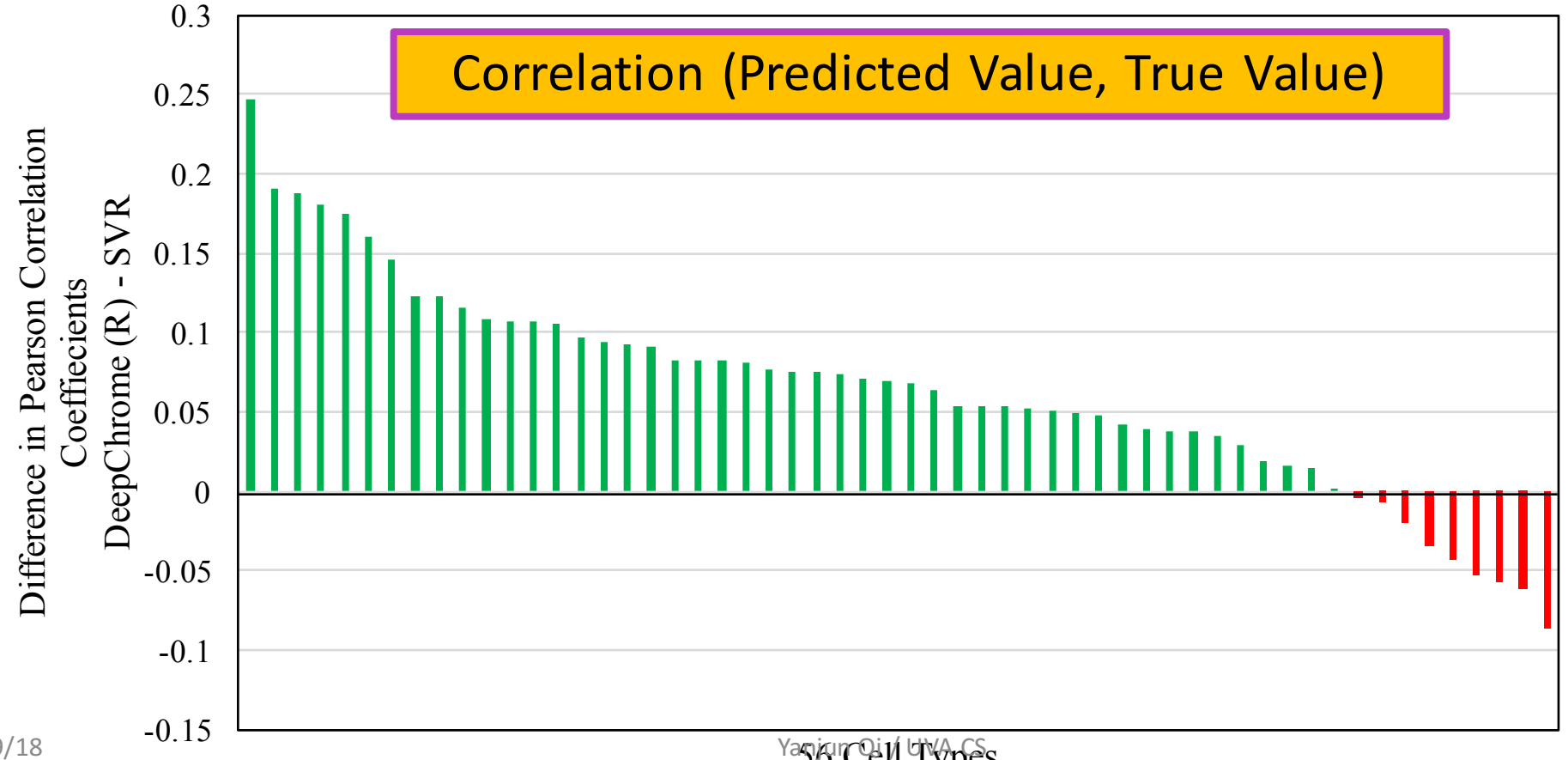
Mean Square Error Loss

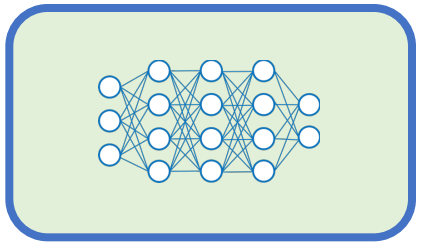
$(Y - f(X))^2$



Where are we heading?

Changing Task : Classification →Regression

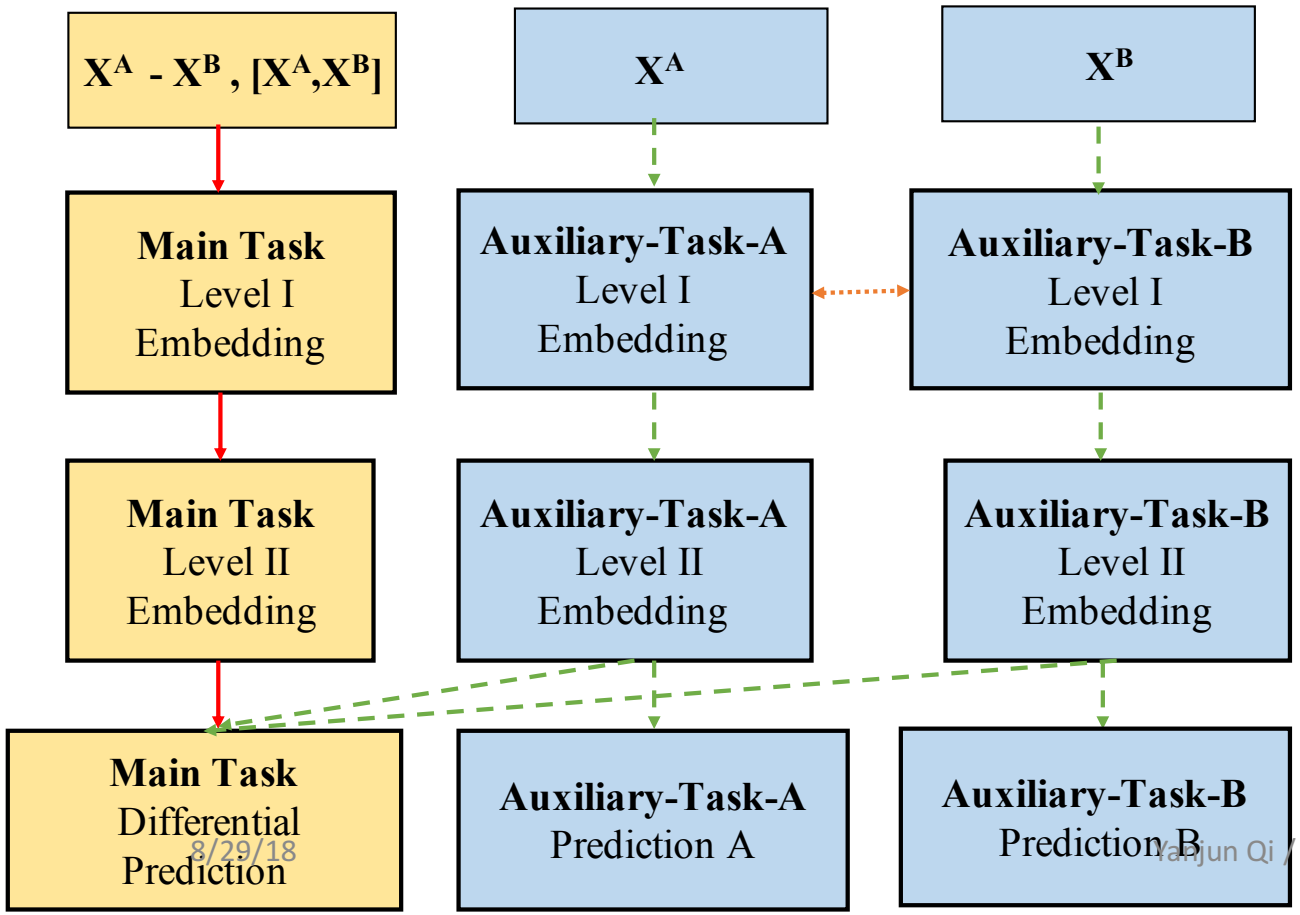




Where are we heading?

DeepDiff: Deep-learning for predicting Differential gene expression from histone modifications

Changing Task : Cell-Specific → Cross Cell



1 **Main Task**: Differential gene expression prediction

2 **Cell-Specific Auxiliary**: Auxiliary-Task-A and Auxiliary-Task-B cell type specific prediction

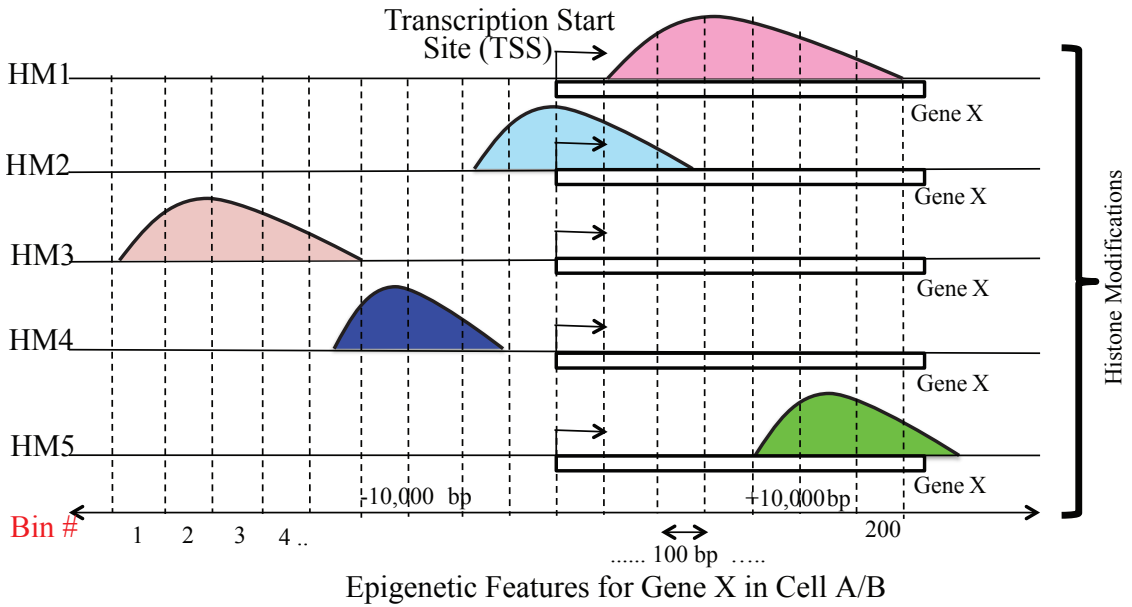
3 **Siamese Auxiliary**: Siamese contrastive loss

DeepDiff Variations	Objective Loss
1 Raw:d, Raw:c, Raw	ℓ_{Diff}
2 Aux	$\ell_{Diff}+\ell_{CellAux}$
1 + 2 Raw+Aux	$\ell_{Diff}+\ell_{CellAux}$
2 + 3 Aux+Siamese	$\ell_{Diff}+\ell_{CellAux}+\ell_{Siamese}$
1 + 2 + 3 Raw+Aux+Siamese	$\ell_{Diff}+\ell_{CellAux}+\ell_{Siamese}$

Where are we heading?

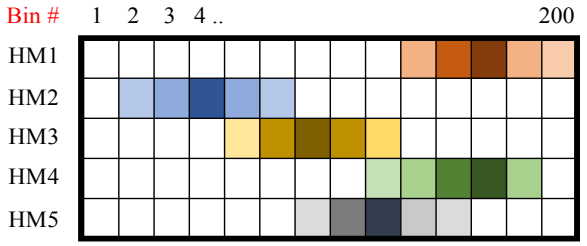
DeepDiff: Deep-learning for predicting Differential gene expression from histone modifications

Changing Task : Cell-Specific → Cross Cell

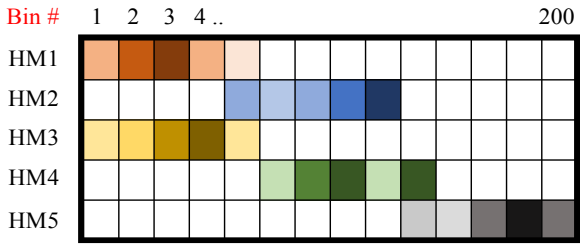


Epigenetic Features for Gene X in Cell A/B

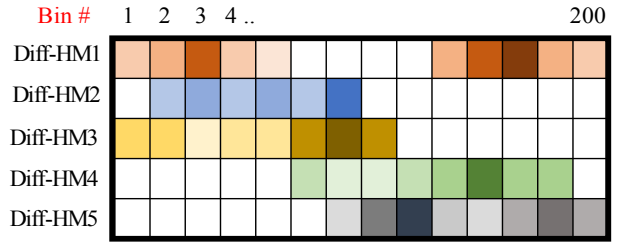
(a)



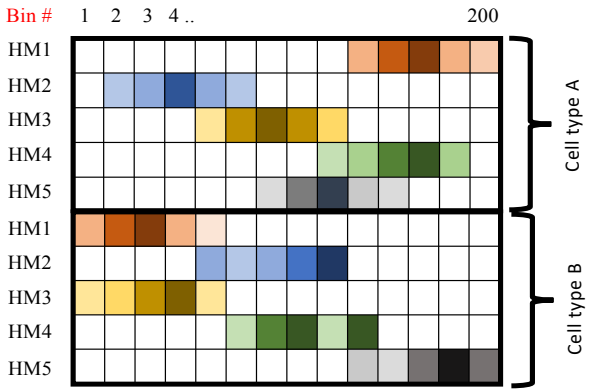
Raw Input Matrix for Gene X in Cell-type A: $\mathbf{X^A}$



Raw Input Matrix for Gene X in Cell-type B: $\mathbf{X^B}$



Raw Difference Matrix for Gene X: $\mathbf{X^A - X^B}$



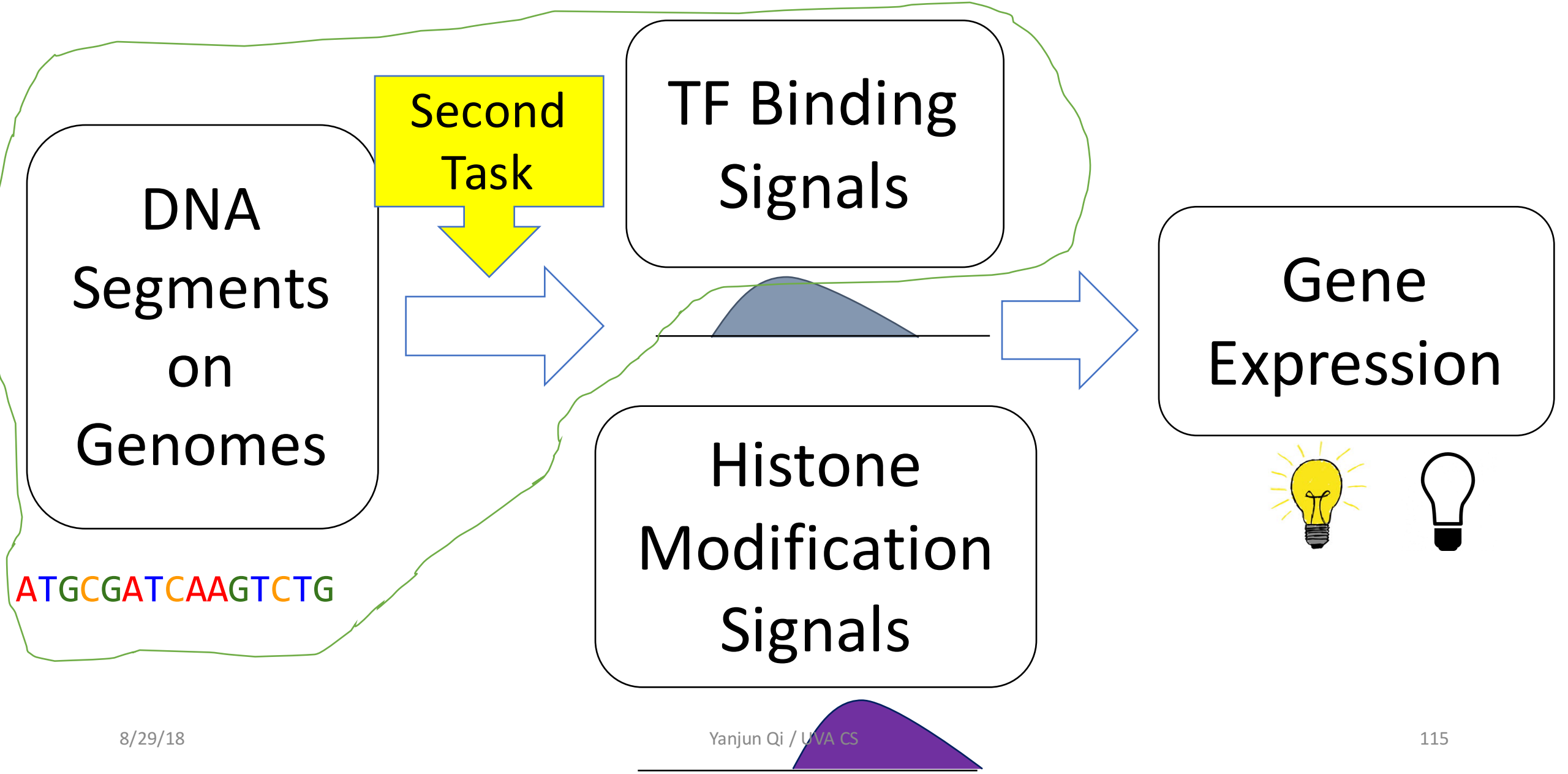
Raw Concat Matrix for Gene X: $\mathbf{[X^A, X^B]}$

(b)

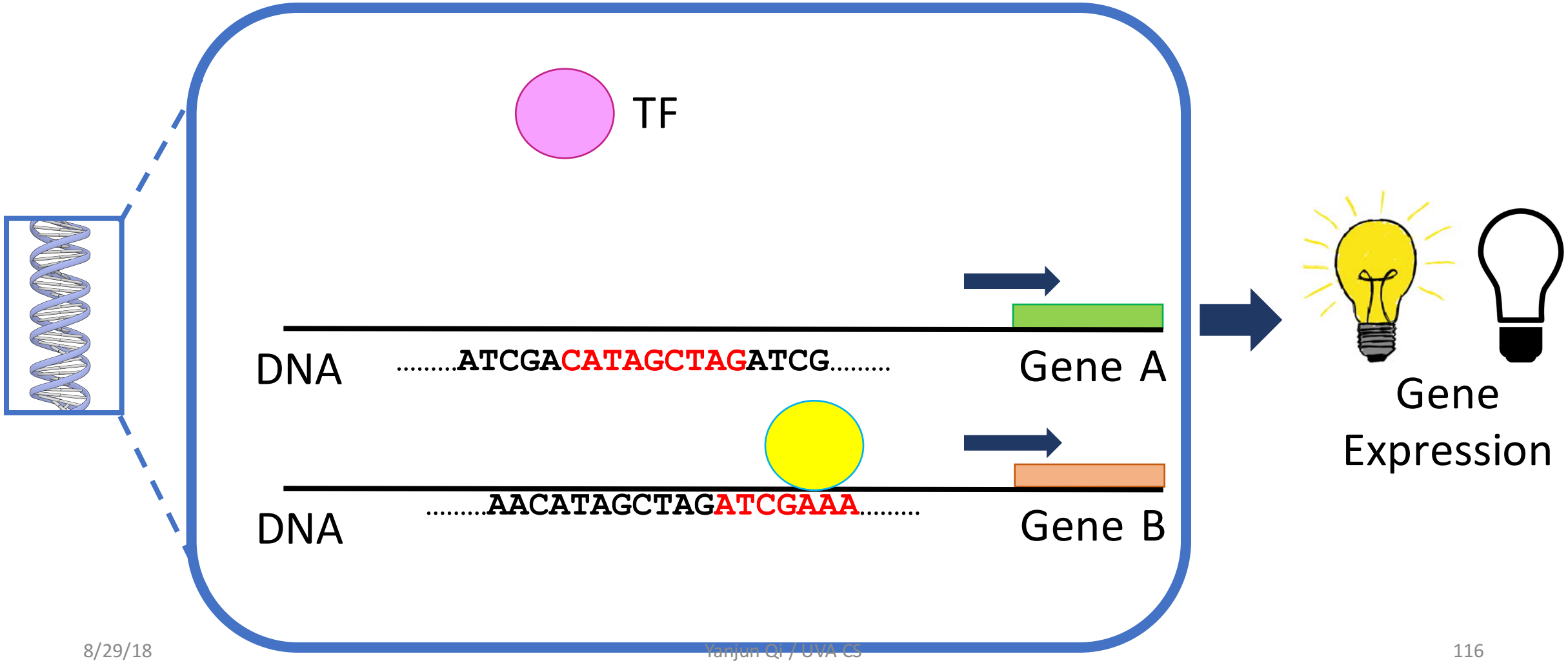
Today

- Machine Learning: a quick review
- Deep Learning: a quick review https://qdata.github.io/deep2Read/
- Background Biology: a quick review
- Deep Learning for analyzing Sequential Data about Regulation:
  - DeepChrome
  - AttentiveChrome
  - DeepMotif https://www.deepchrome.org

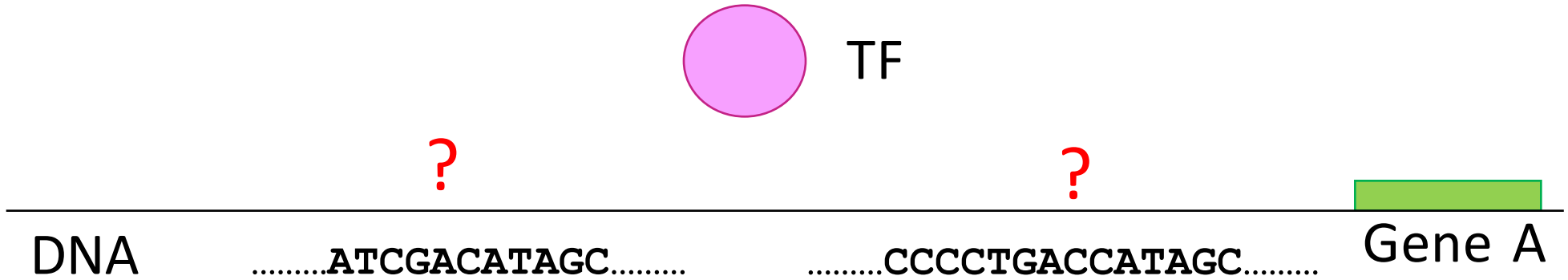
Many Important Data-Driven Computational Tasks



Transcription Factors



TF-Binding Site?

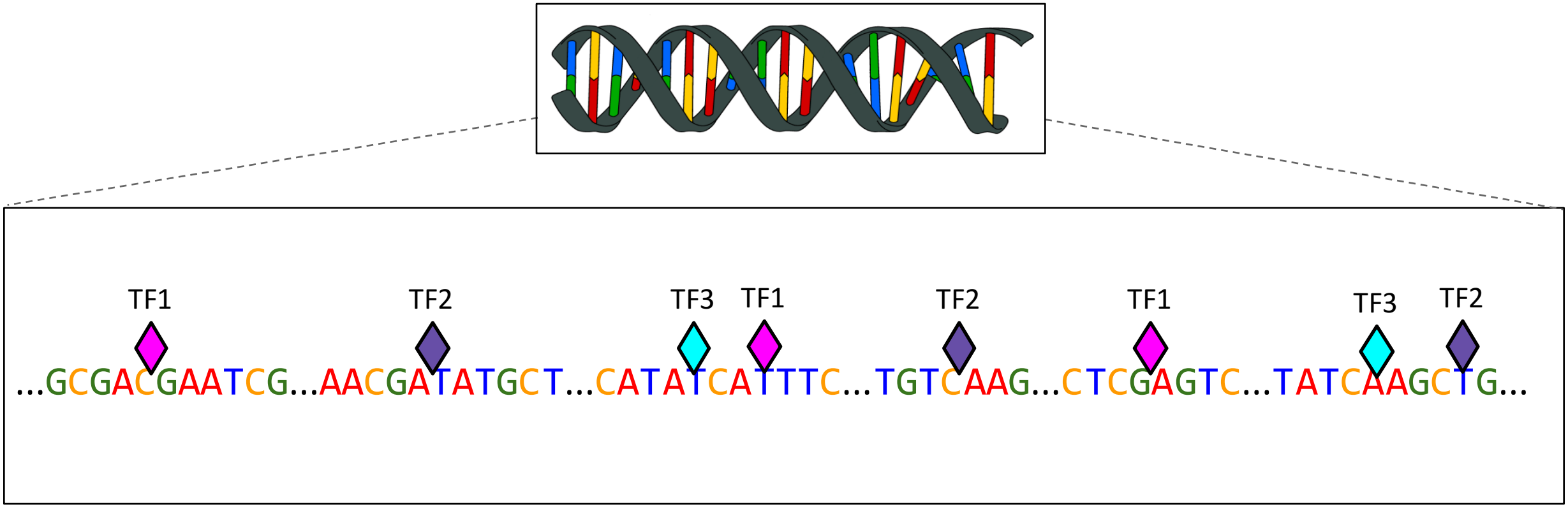


"TF-Binding Site?"

ATCGAATCCG

CCCTCTATCG

Task: Sequence Based Functional Annotation Tasks

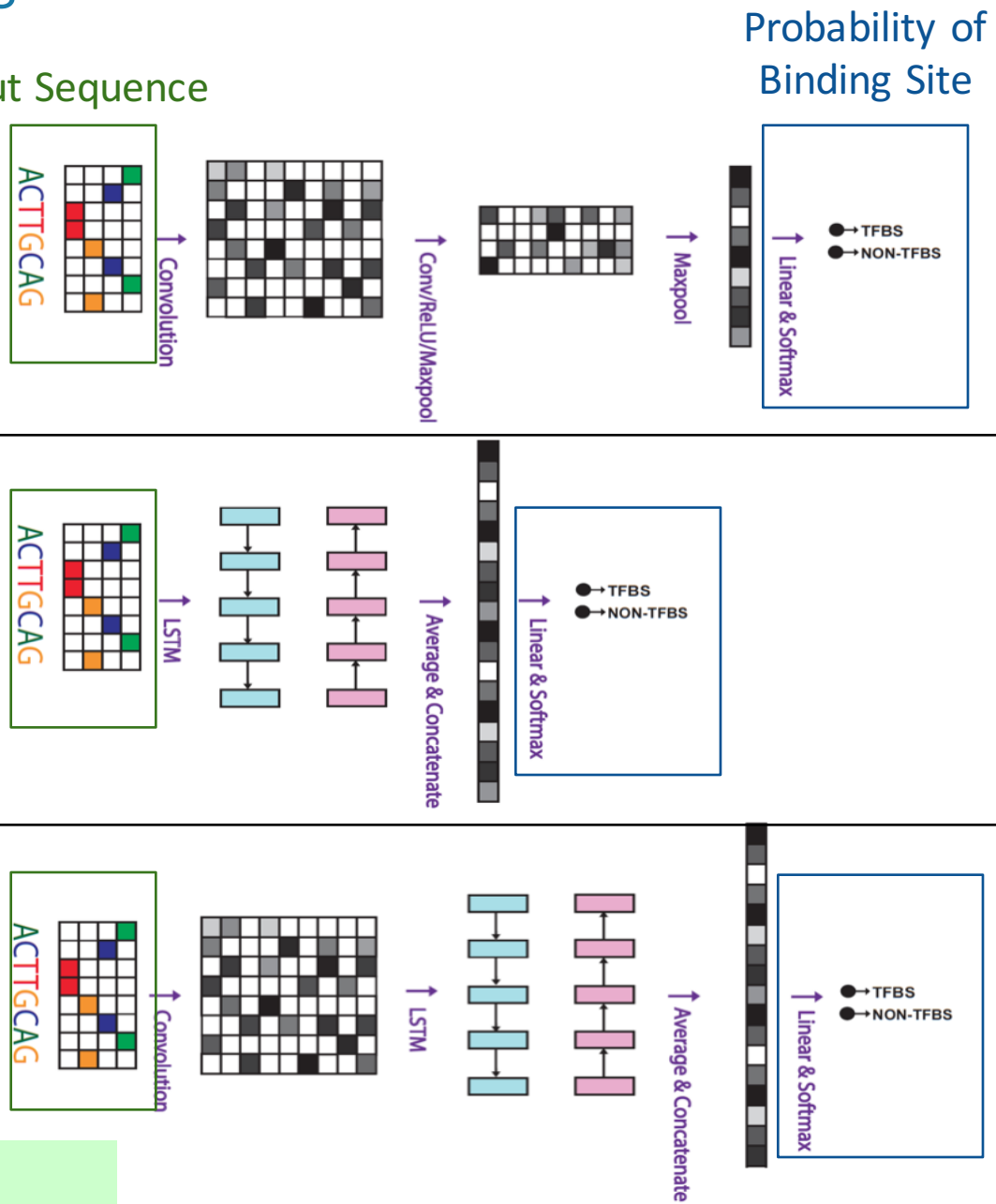


Literature: Various DNN Tools

1. **Convolutional (CNN)**

2. **Recurrent (RNN)**

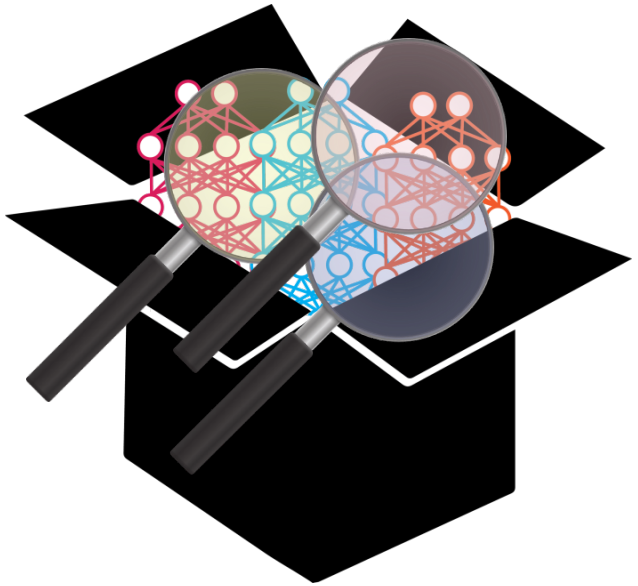
3. **Convolutional-Recurrent (CNN-RNN)**

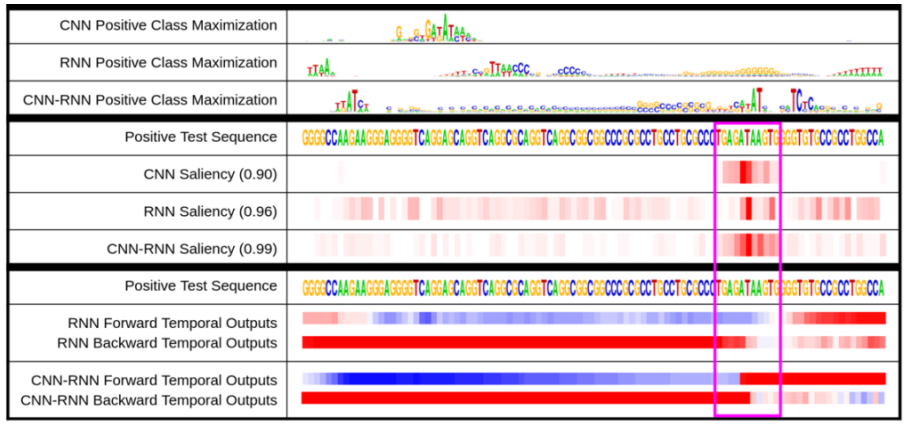


DeepSEA, DeepBind, BASSET, DanQ, ….

Deep Motif Dashboard: Understand DNNs by Post Analysis

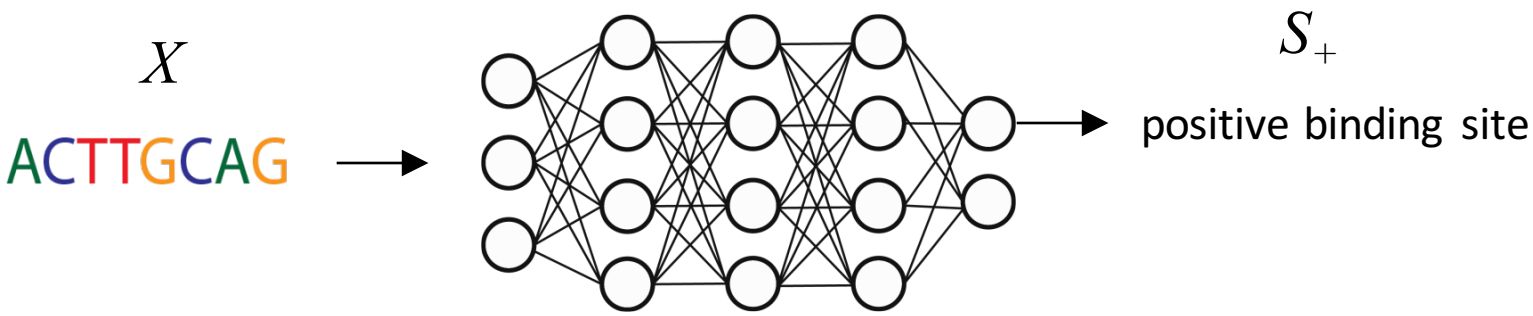
Lanchantin, Singh, Wang & Qi - Pacific Symposium on Biocomputing, 2017





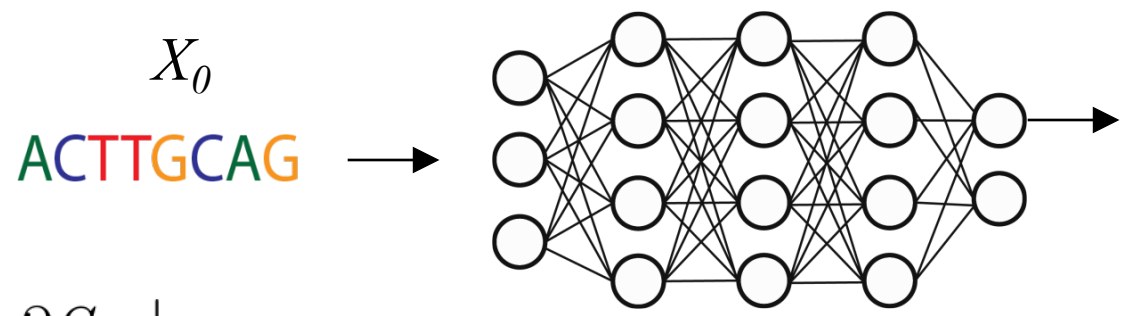
1. Saliency Maps - recommending on CNN kind
2. Temporal Output Values - recommending on RNN kind
3. Class Optimization - recommending on CNN kind

1. Saliency Map



Which nucleotides are most important for my current-sample classification?

1. Saliency Map



$$w = \left.\frac{\partial S_+}{\partial X}\right|_{X_0} = \text{“saliency map”}$$

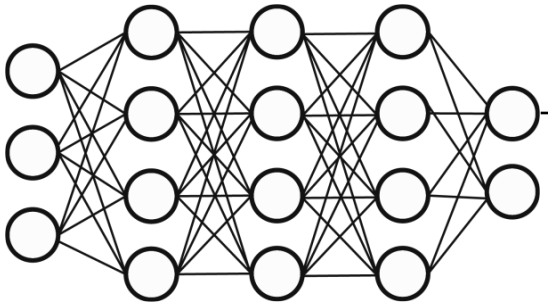
Quiz: What is gradient?

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, ICLR 2013

1. Saliency Map

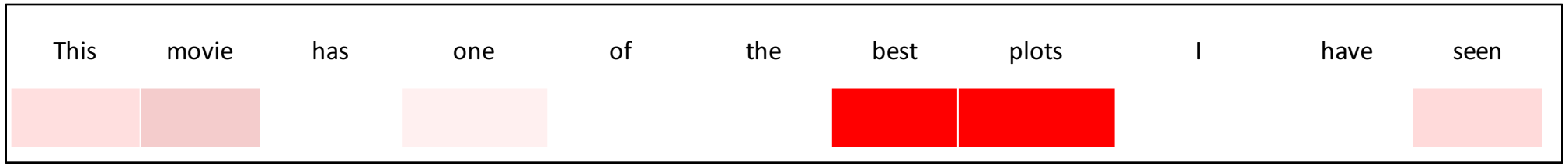
X

This movie has one of the best plots I have seen



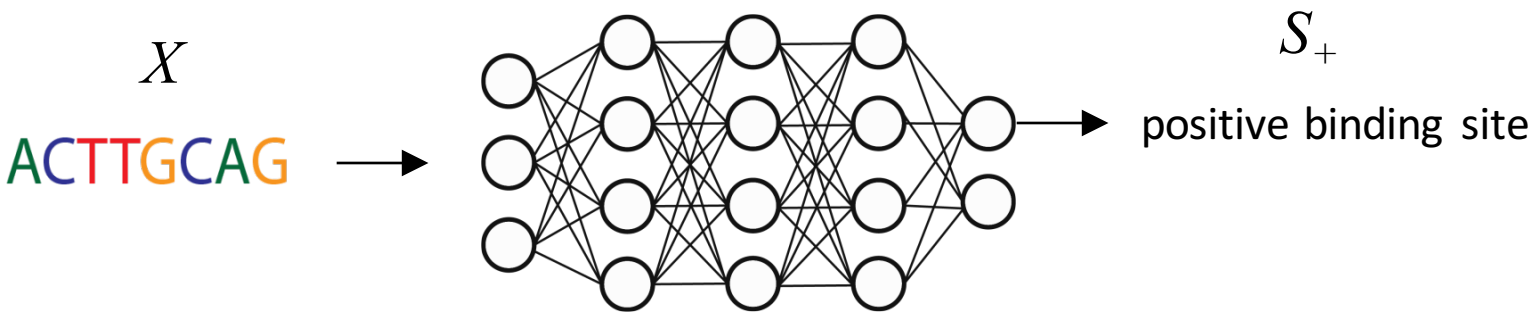
S_+

Positive sentiment



= important for classification

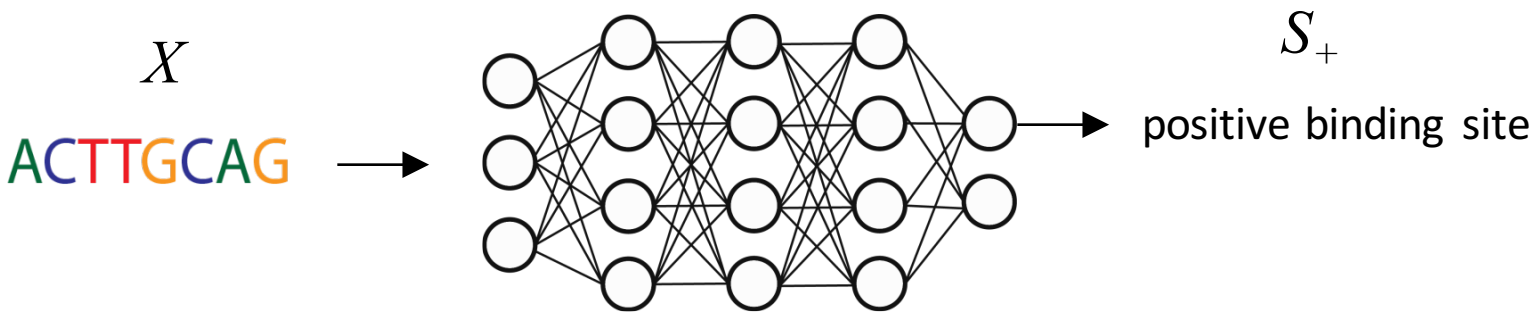
1. Saliency Map



Positive Test Sequence	TGCTCGCATCCTATTGGCCACGTTAGTCACATGGCCCCACCTGGCTGCAAAGCACGCTGGGAAACGTAGTCTTTCTT
Saliency Map	

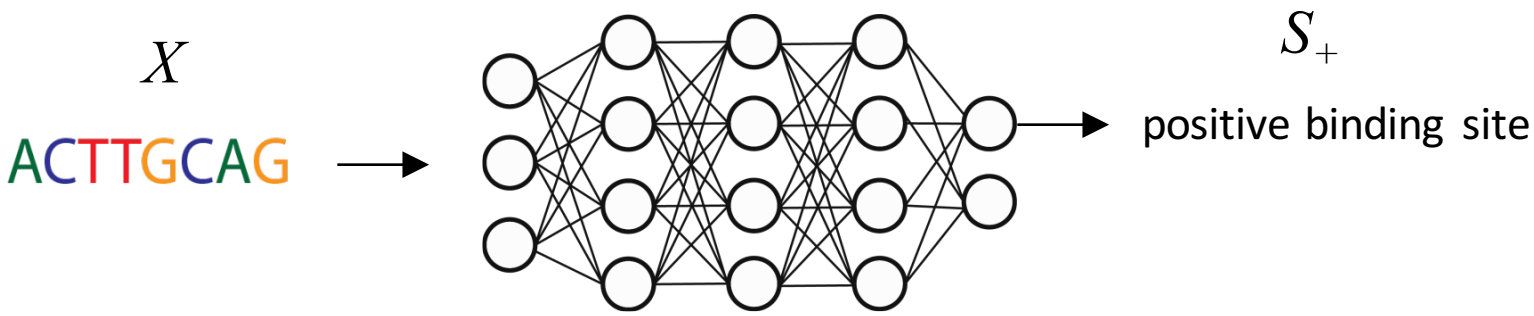
= important nucleotide for prediction

2. Temporal Output Values



What are the model's predictions at each timestep of the DNA sequence?

2. Temporal Output Values

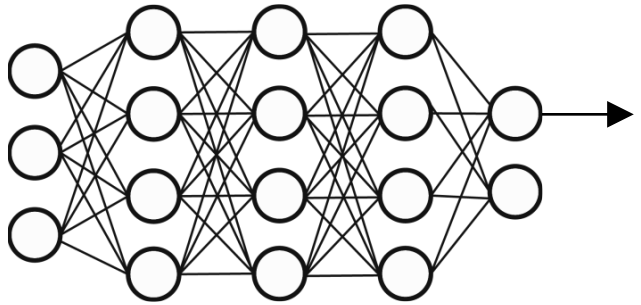


Check the RNN's prediction scores when we vary the input of the RNN starting from the beginning to the end of a sequence.

2. Temporal Output Values

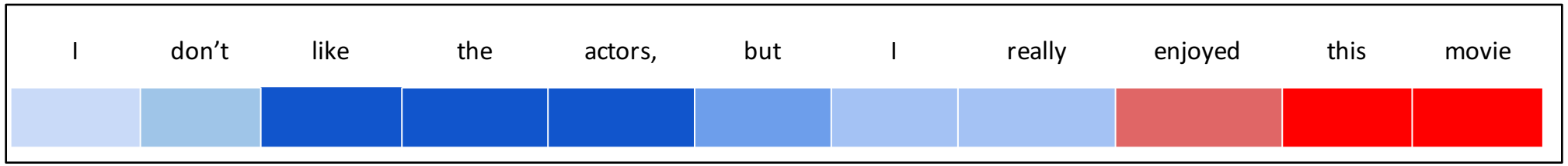
X

I don't like the actors, but I really enjoyed this movie



S_+

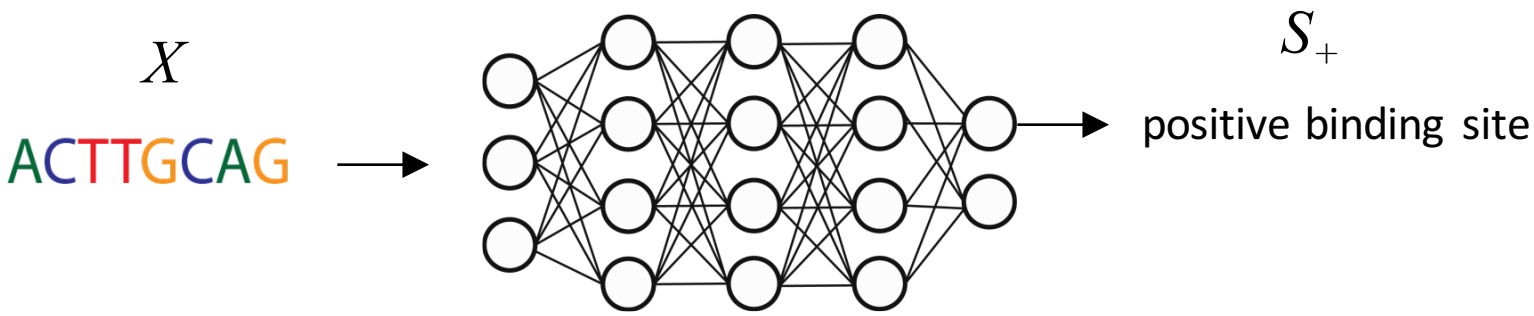
positive sentiment



= negative sentiment

= positive sentiment

2. Temporal Output Values

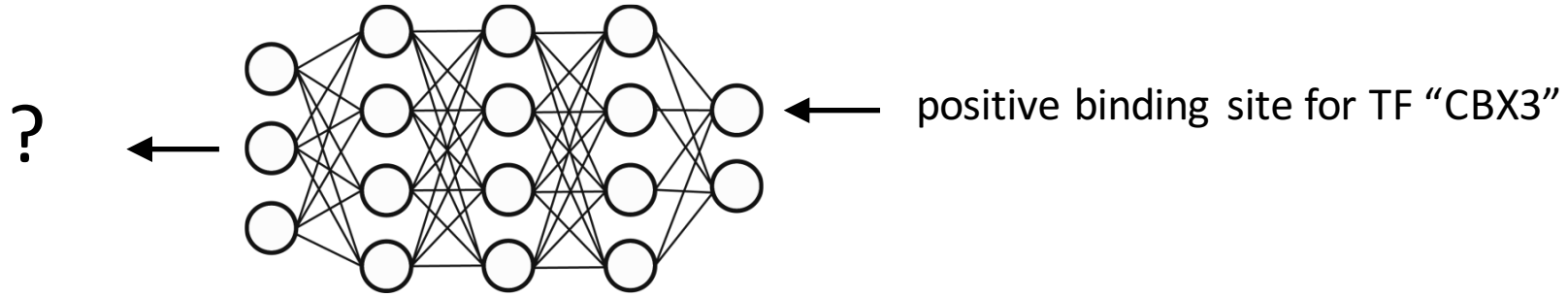


Positive Test Sequence	CTTCTGCTCGCATCCTATTGGCCACGTTAGTCACATGGCCCCACCTGGCTGCAAAGCACGCTGGGAAACGTAGTCTTTCTT
RNN Forward Output	
RNN Backward Output	

= negative binding site prediction

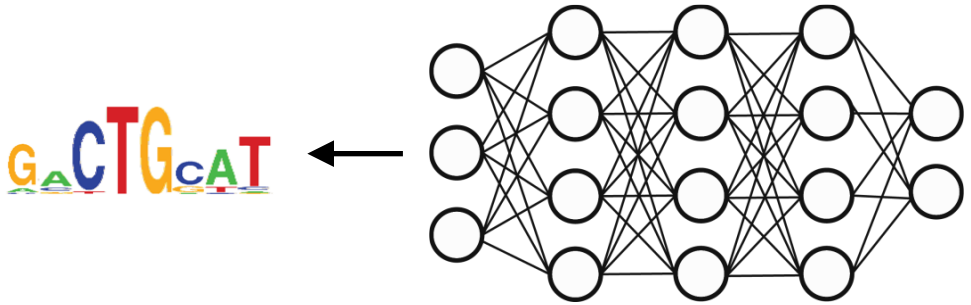
= positive binding site prediction

3. Class Optimization



For a particular TF, what does the optimal binding site sequence look like?

3. Class Optimization



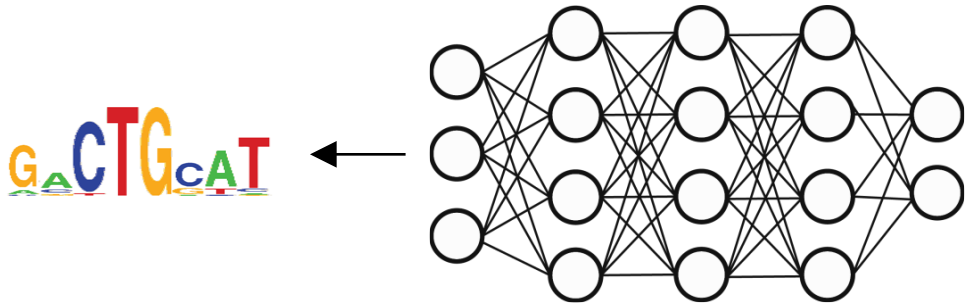
positive binding site for TF “CBX3”

$$\arg\max_X S_+(X) + \lambda \|X\|_2^2$$

Where X is the input sequence and the score S_+ is probability of sequence X being a positive binding site

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, ICLR 2013

3. Class Optimization



positive binding site for TF “CBX3”

Optimal binding site for TF “CBX3”	

Visualization Methods

Sequence Specific

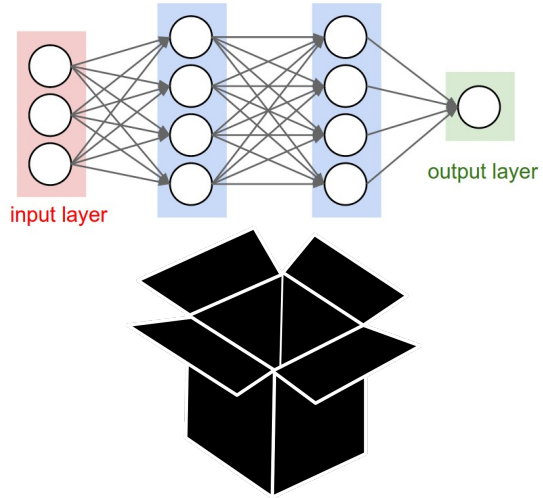
1. Saliency Maps – (CNN kind)
2. Temporal Output Values – (RNN kind)

TF Specific

3. Class Optimization – (CNN kind)

code available at: **deepmotif.org**

Related Work to Post-Understand DNN



- Deconvolution
- Perturbation-based
- Backpropagation-based
- Difference to Reference
- Influence based

Temporal Output Values

Saliency Map

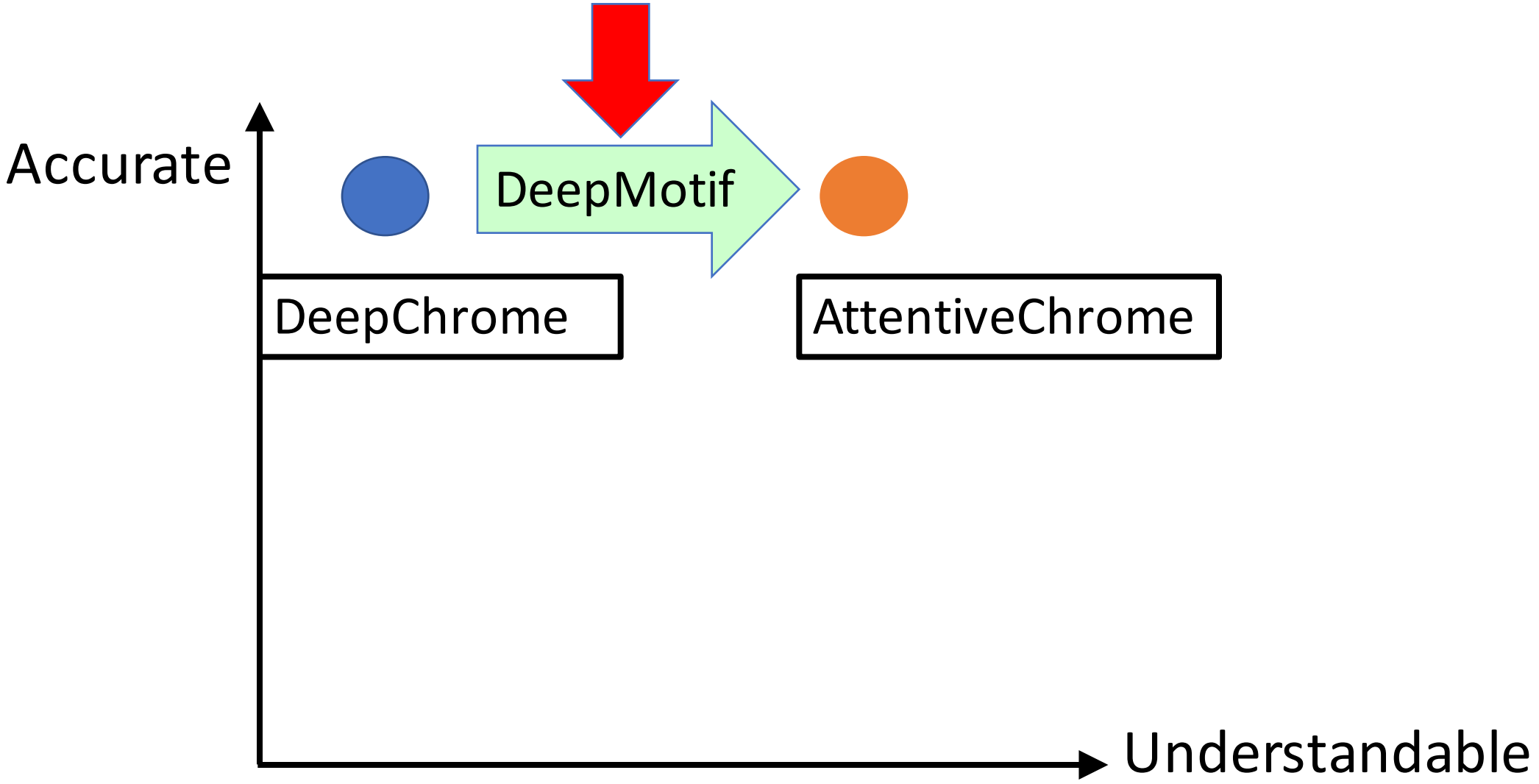
Class Optimization

DeepLift

Influential Function / ICML27 Best Paper

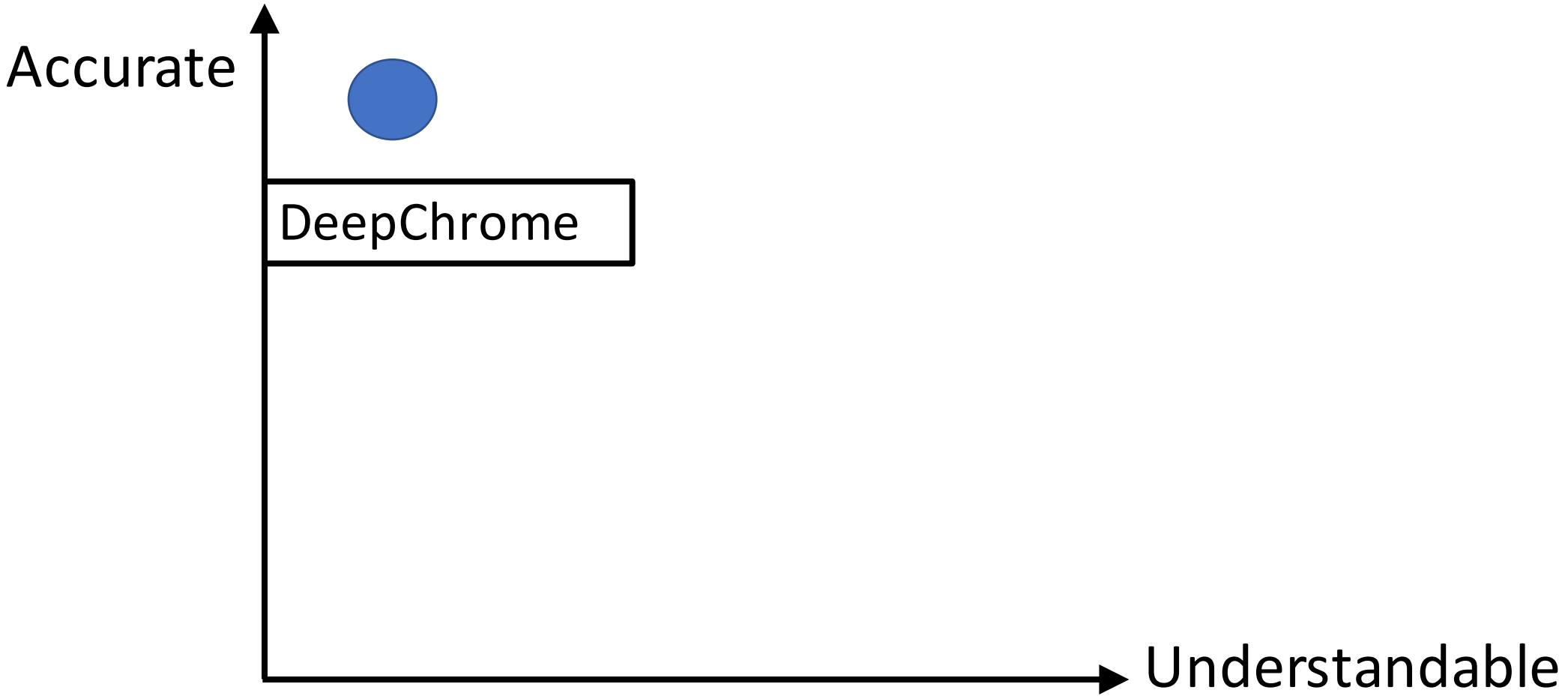
https://www.deepchrome.org

Summary of tools



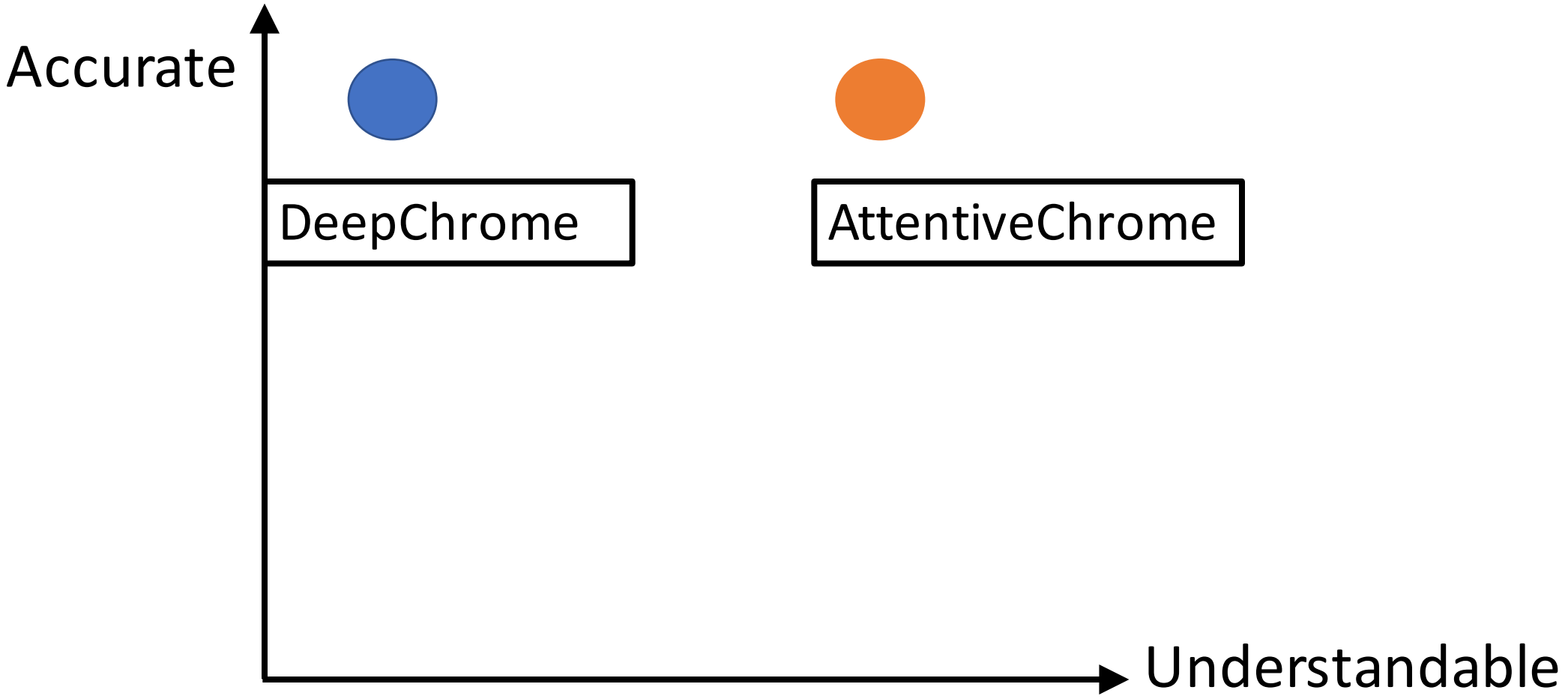
https://www.deepchrome.org

Recap

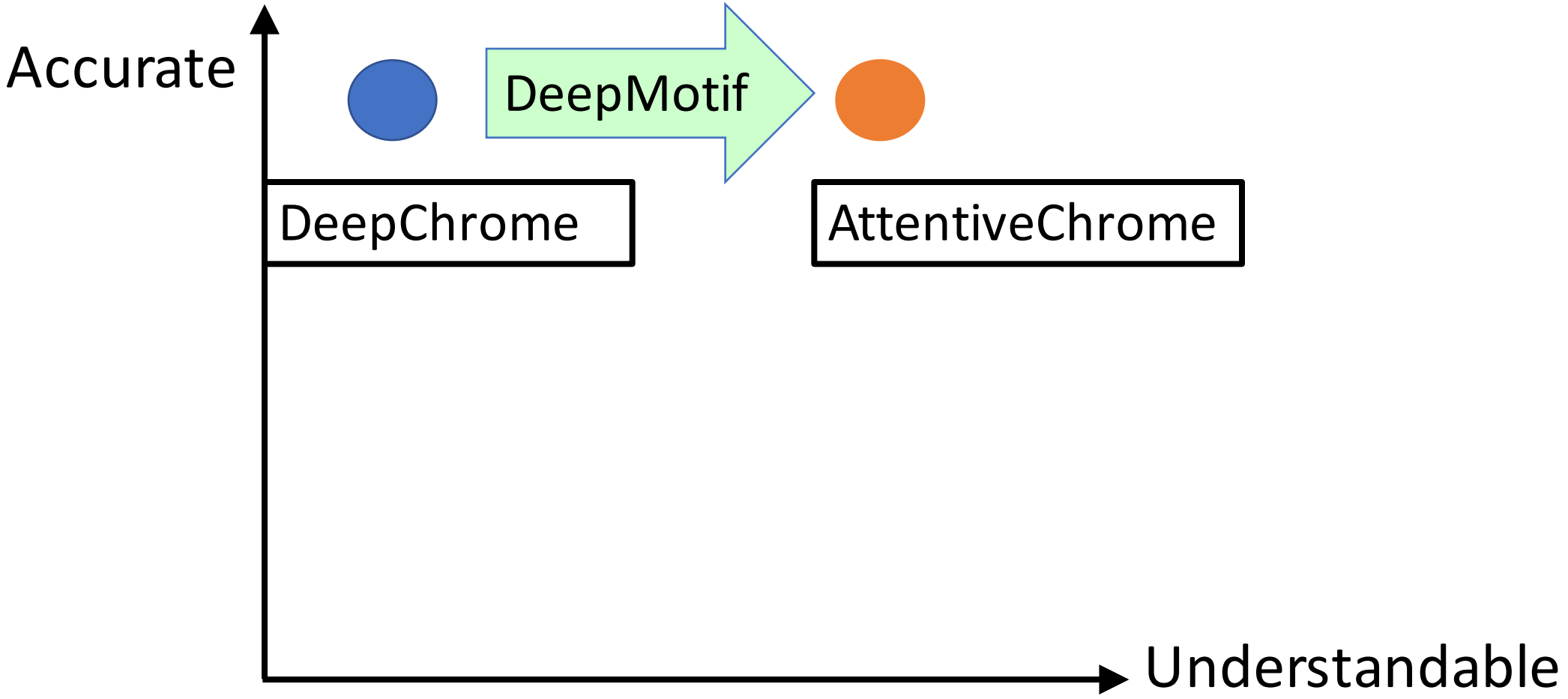


https://www.deepchrome.org

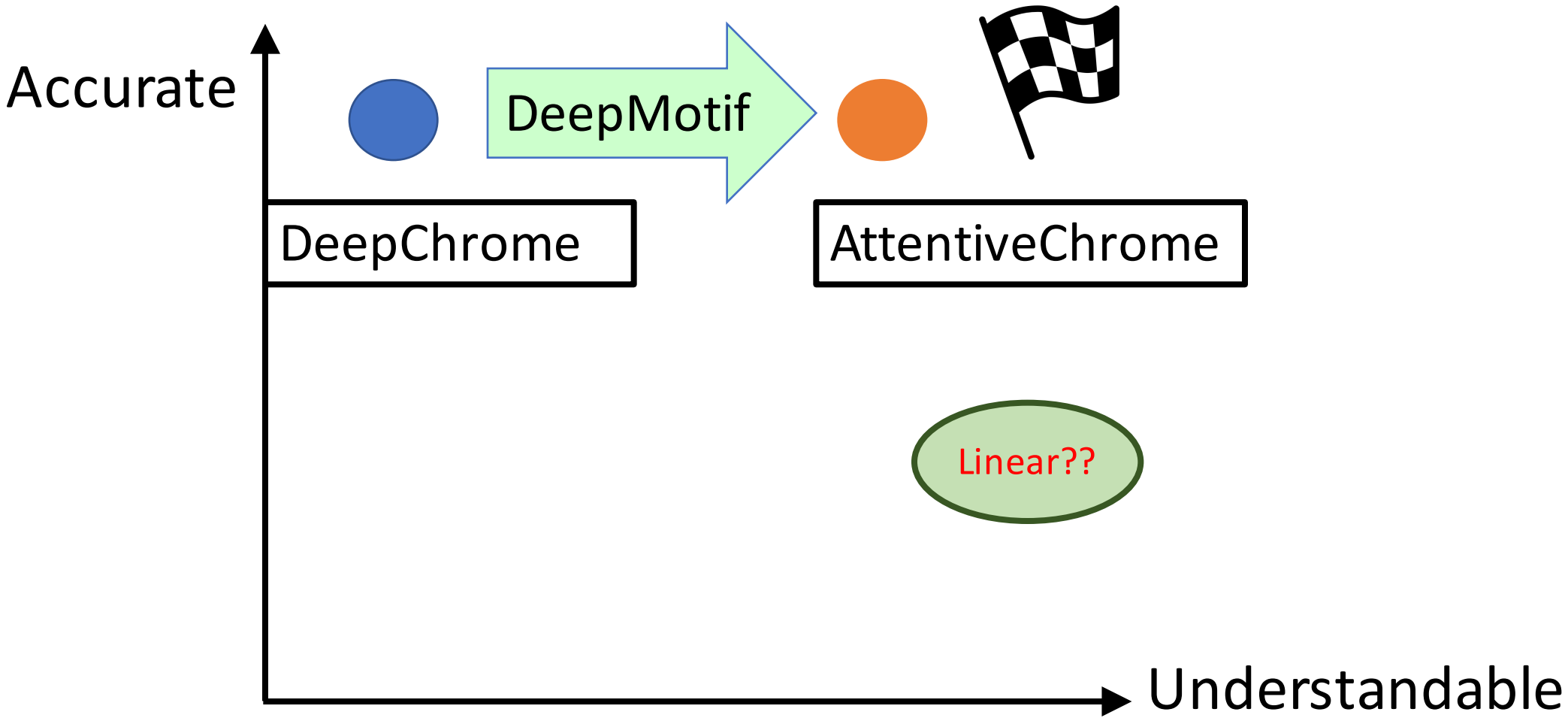
Recap



Recap


Accurate
DeepMotif
DeepChrome
AttentiveChrome
Understandable

Recap


Accurate
DeepMotif
DeepChrome
AttentiveChrome
Linear??
Understandable

Acknowledgements


Ritambhara Singh


Jack Lanchantin


Arshdeep Sekhon


Beilun Wang

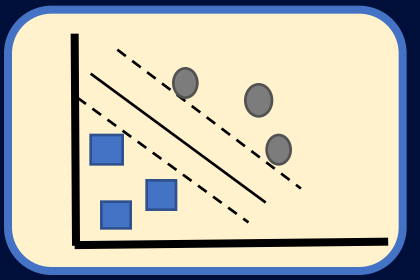
UVA Department of Biochemistry and Molecular Genetics: Dr. Mazhar Adli


Thank you

More Tools: A Fast and Scalable Tool to Classify Biological Sequences

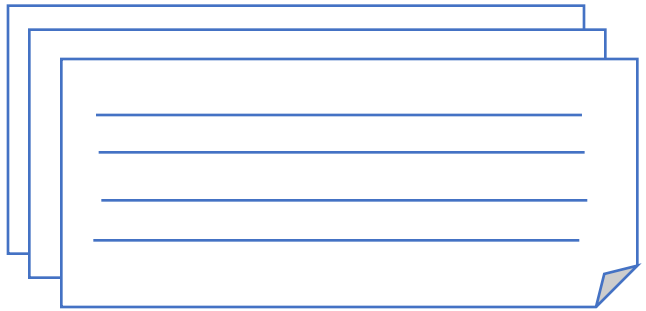
https://github.com/QData/iGakco-SVM

String Kernel + SVM



String Kernel Function

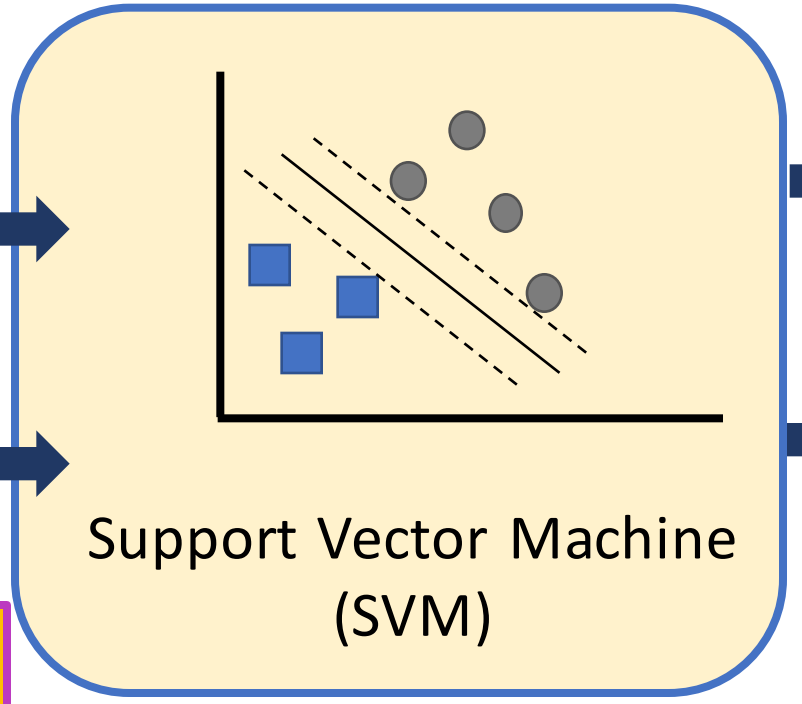
K(.)



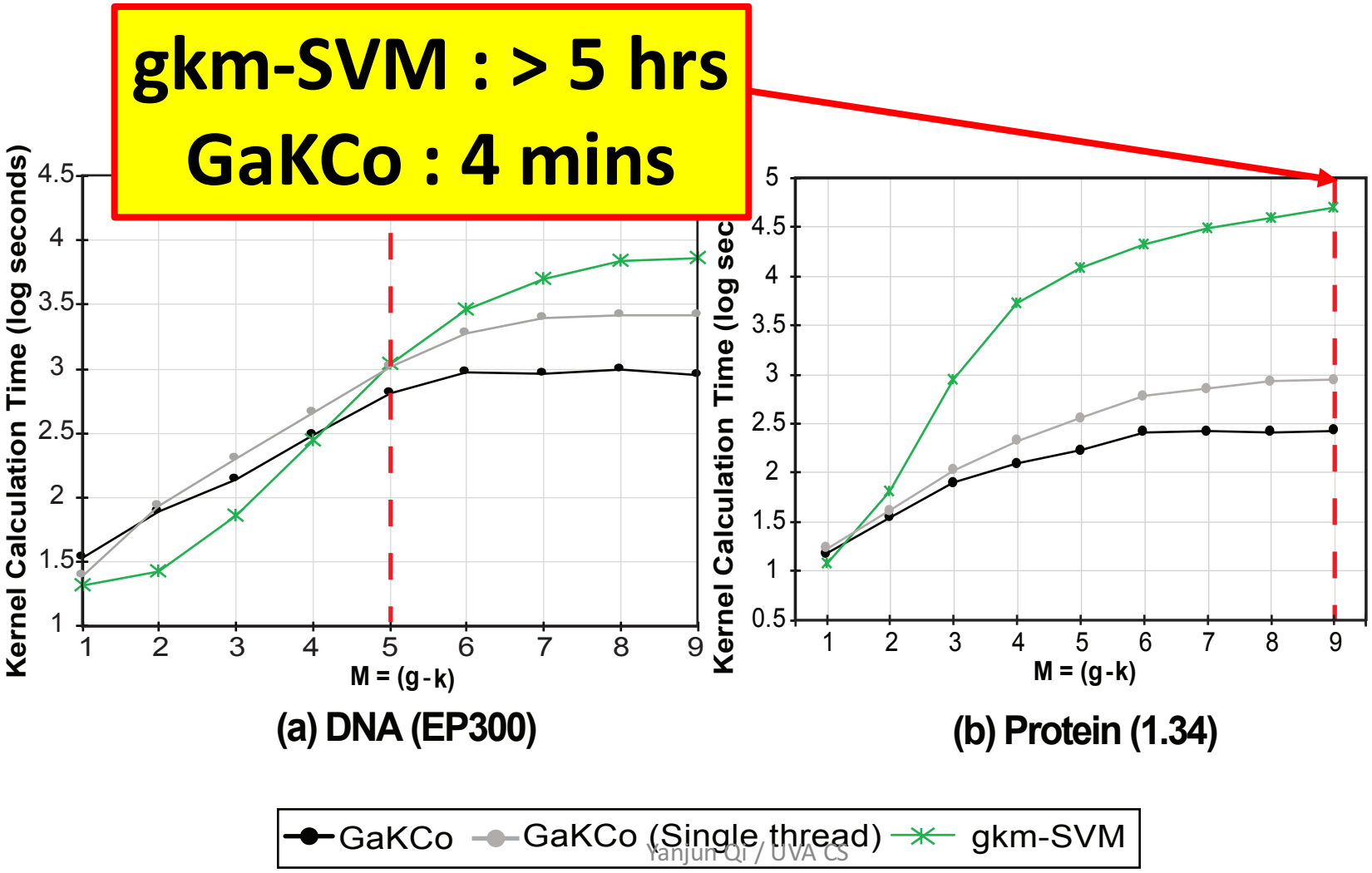
S=TCGAATCCG
T=GCTGAATCG

S=QGGHAKKQQ
T=KKHAVQQVV

$K(S,T)=\varphi(S).\ \varphi(T)$



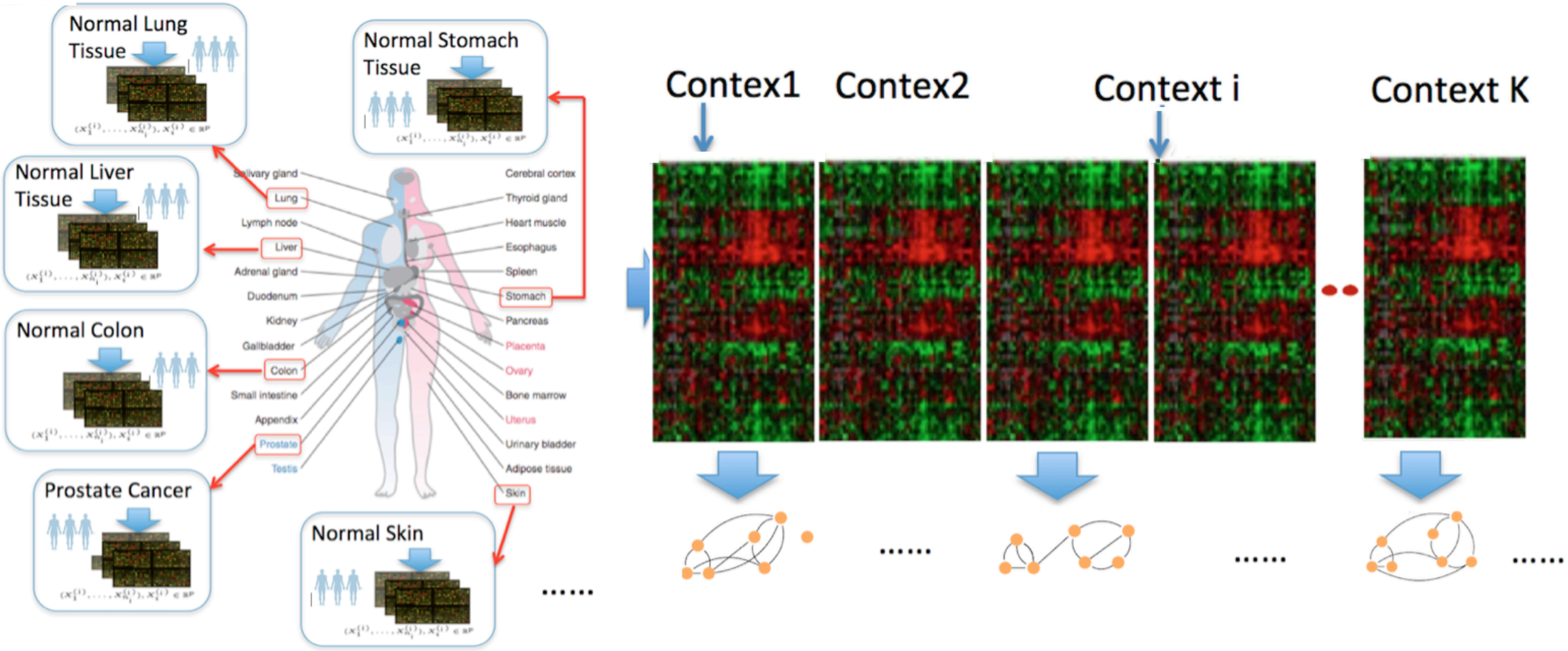
Our Tool Scales well with increasing ∑ and m



More Tools: Extracting graphs from data

https://www.jointggm.org

Motivation: Graphs vary across contexts



Limitation of Previous Methods : Storage

e.g., calculate the gradient

$$\Sigma = \mathrm{Cov}(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

$$\Sigma = \mathrm{Cov}(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

$$\Sigma = \mathrm{Cov}(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

When K contexts= 91, p nodes= 30K

$O(Kp^2)$ in memory

Double type: 65 TB

Limitation of Previous Methods: Speed

Suppose they have the same iteration number T

K = 91, p= 30K

Traditional Optimization Method

---- Block Coordinate Descent : $O(K^3 p^4)$ / Itera

more than 2 billion years

Current Optimization: ADMM based

---- Still needs SVD for each covariance matrix

SVD for the matrices needs $O(Kp^3)$ / Itera → 3.5 days

Our Tools

- Fast and scalable estimators for joint graph discovery from heterogeneous samples
- Parallelizable algorithms
- Sharp convergence rate (sharp error bounds)

More details at: http://www.jointggm.org/