Computer Science
at the UNIVERSITY of VIRGINIA

Making Deep Learning Understandable for Analyzing Sequential Data about Gene Regulation

Dr. Yanjun Qi

2017/11/26

Roadmap

- ✧Background of Machine Learning
- ✧Background of Sequential Data about Gene Regulation
- ✧AttentiveChrome for understanding gene regulation by selective attention on chromatin

Roadmap

- ✧Background of Machine Learning
- ✧Background of Sequential Data about Gene Regulation
- ✧AttentiveChrome for understanding gene regulation by selective attention on chromatin

Machine Learning is Changing the World

How may I help you, human?

Apple Siri / Amazon Echo

IBM WATSON

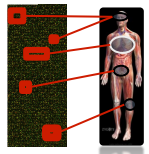
Control learning

Object recognition

Many more !

Challenge of data explosion

Molecular signatures of tumor / blood sample

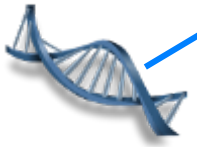


Signs & Symptoms

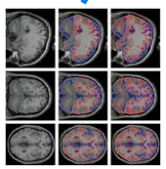
Patient Medical History & Demographics

Public Health Data

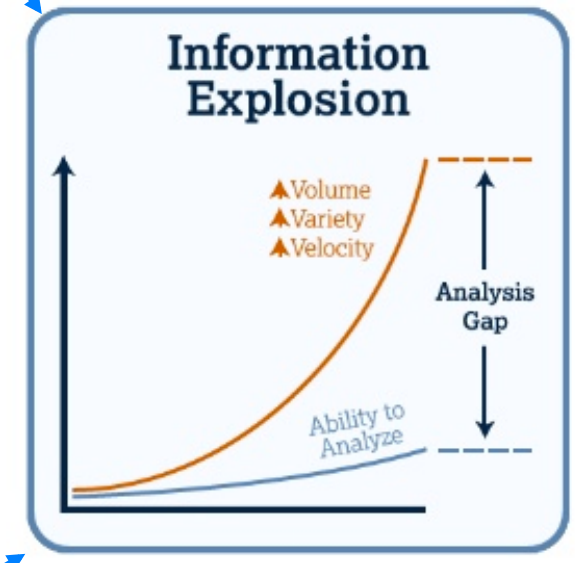
Genetic Data



Mobile medical sensor data



Medical Images

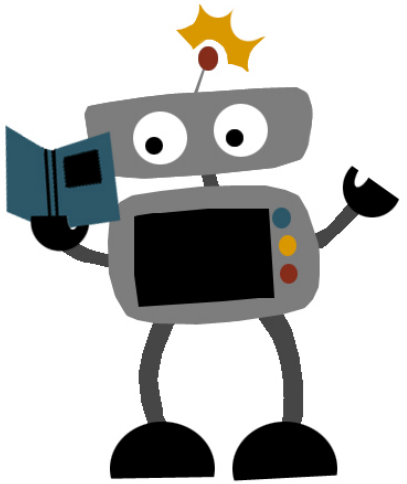


Traditional Approaches

Data-Driven Approaches

Machine Learning

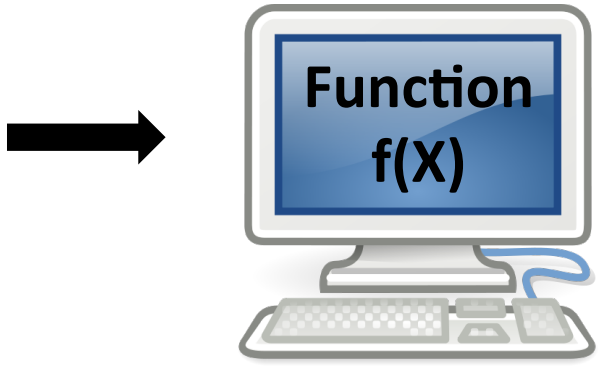
What is Machine Learning?



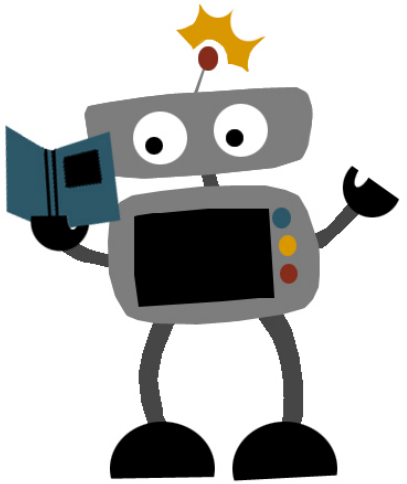
Training

Input : X

Output: Y



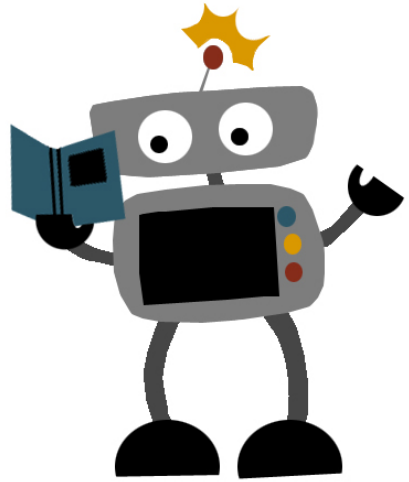
What is Machine Learning?



Testing

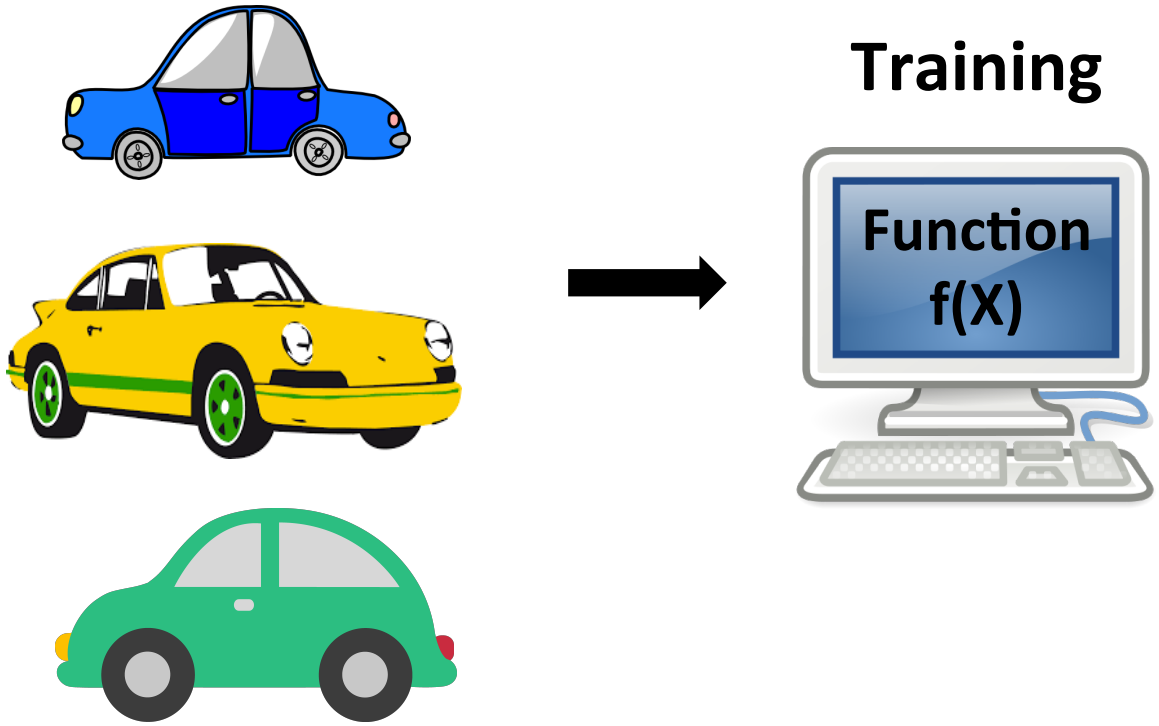
Input : X’ → **Function f(X)** → **Output:** Y’

[Example:] What is Machine Learning?



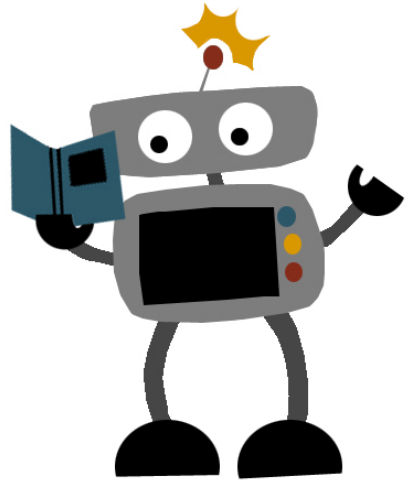
Output: CAR

Inputs:

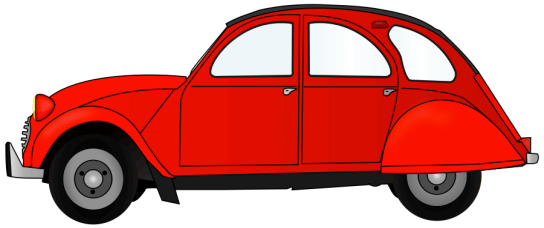


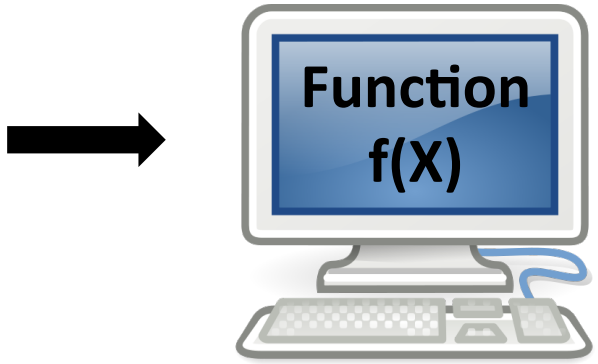
Training

[Example:] What is Machine Learning?



Testing



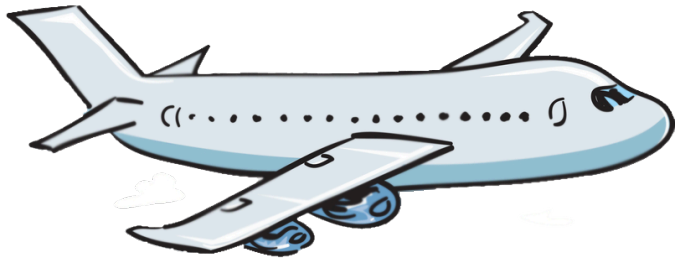


CAR

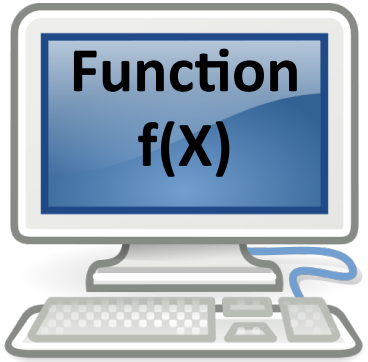
Yes
(+1)

[Example:] Classification task in Machine Learning

Class: Car



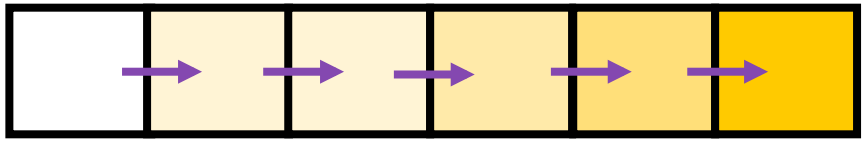
Testing



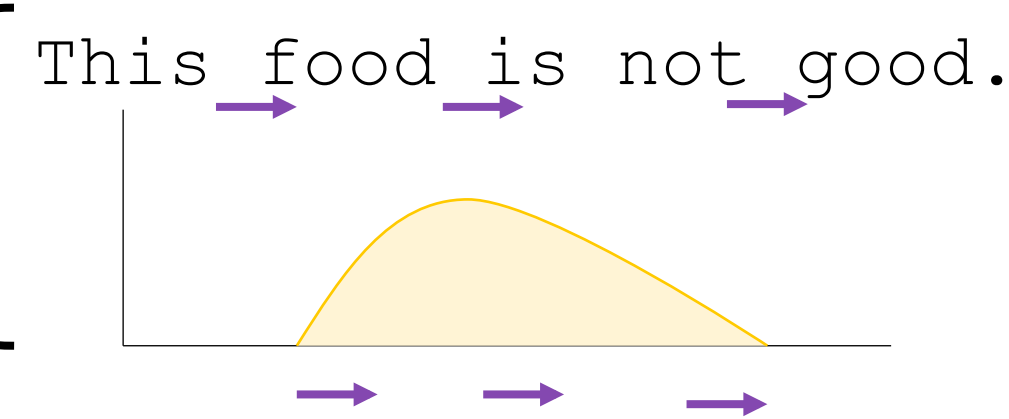
NO
(-1)

Sequential Data

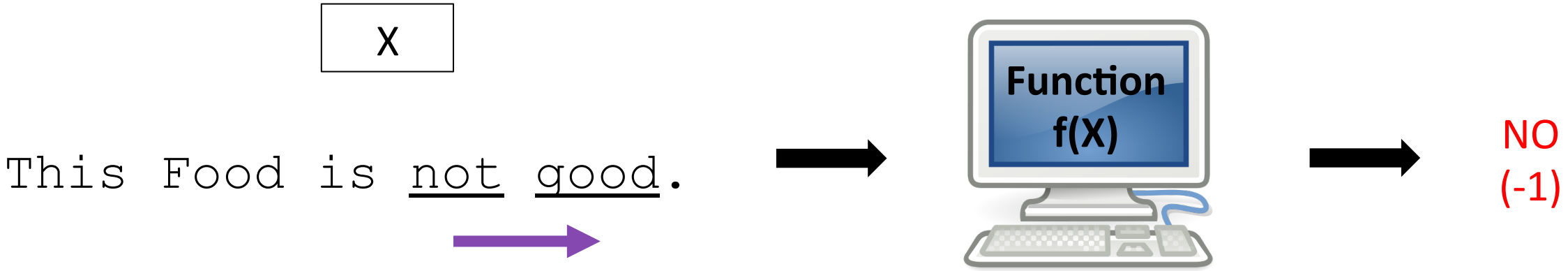
Sequential Data:



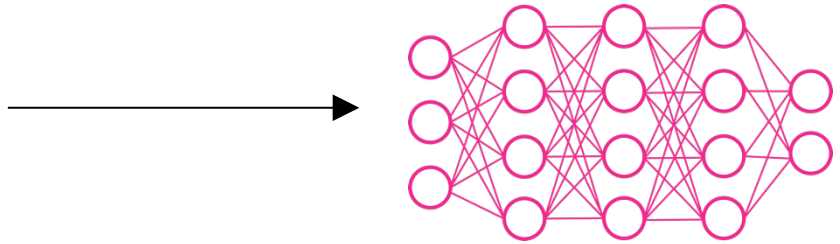
Strings, signals etc.



[Example:] Classification of Sequential Data

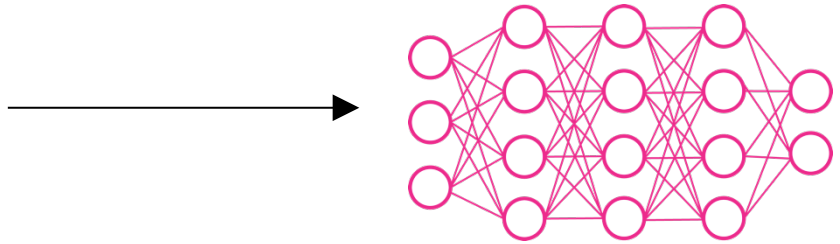


State-of-the-art Machine Learning - Deep Neural Networks (DNN)



“Dog”

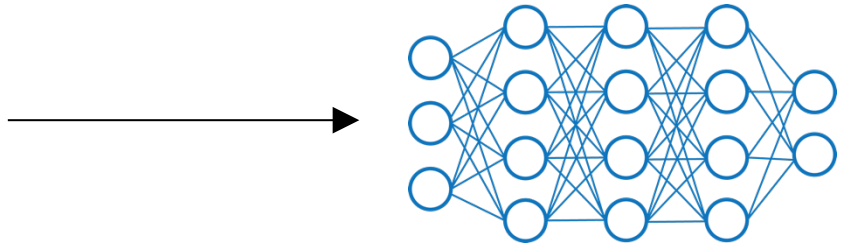
State-of-the-art Machine Learning - Deep Neural Networks (DNN)



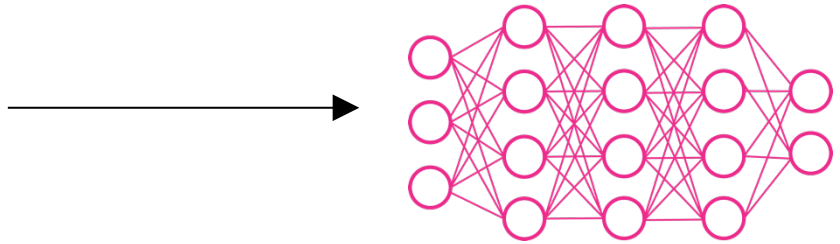
"Dog"

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



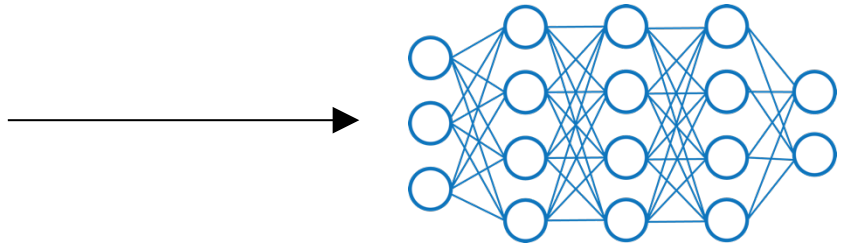
State-of-the-art Machine Learning - Deep Neural Networks (DNN)



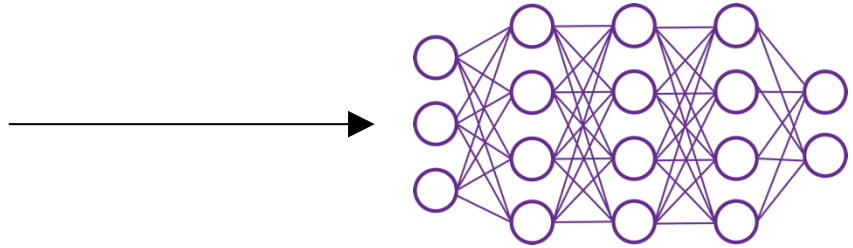
"Dog"

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

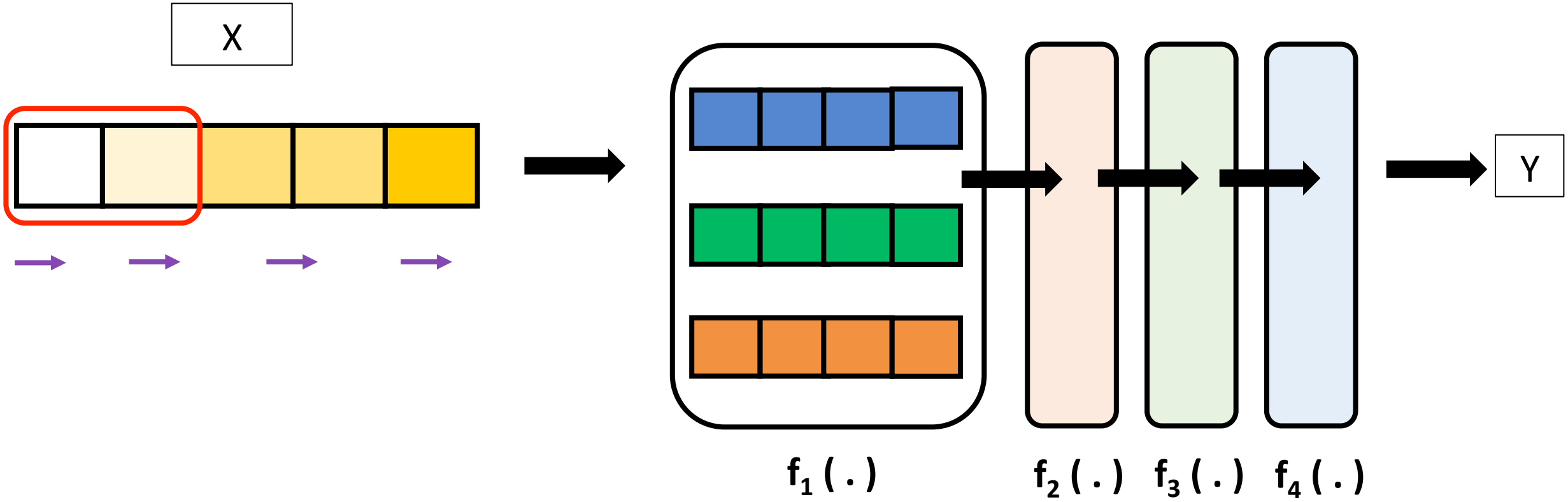


ATGCGATCAAGTCTG

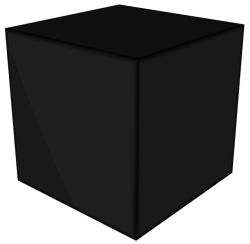


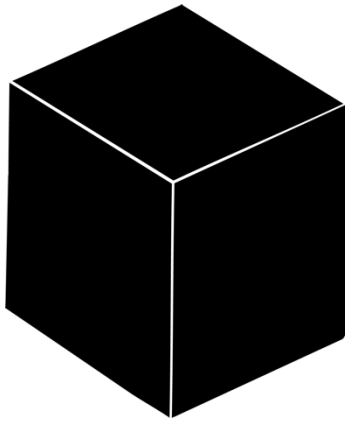
"Protein Binding Site"

Deep Neural Networks (DNN)



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

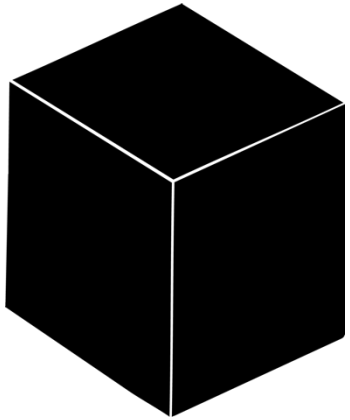




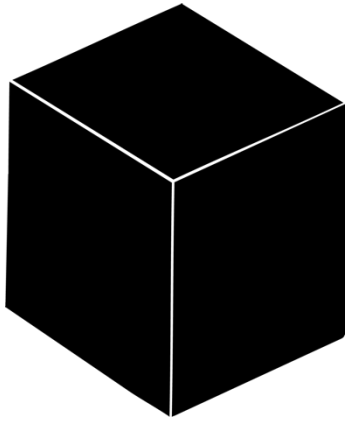
"Dog"

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.

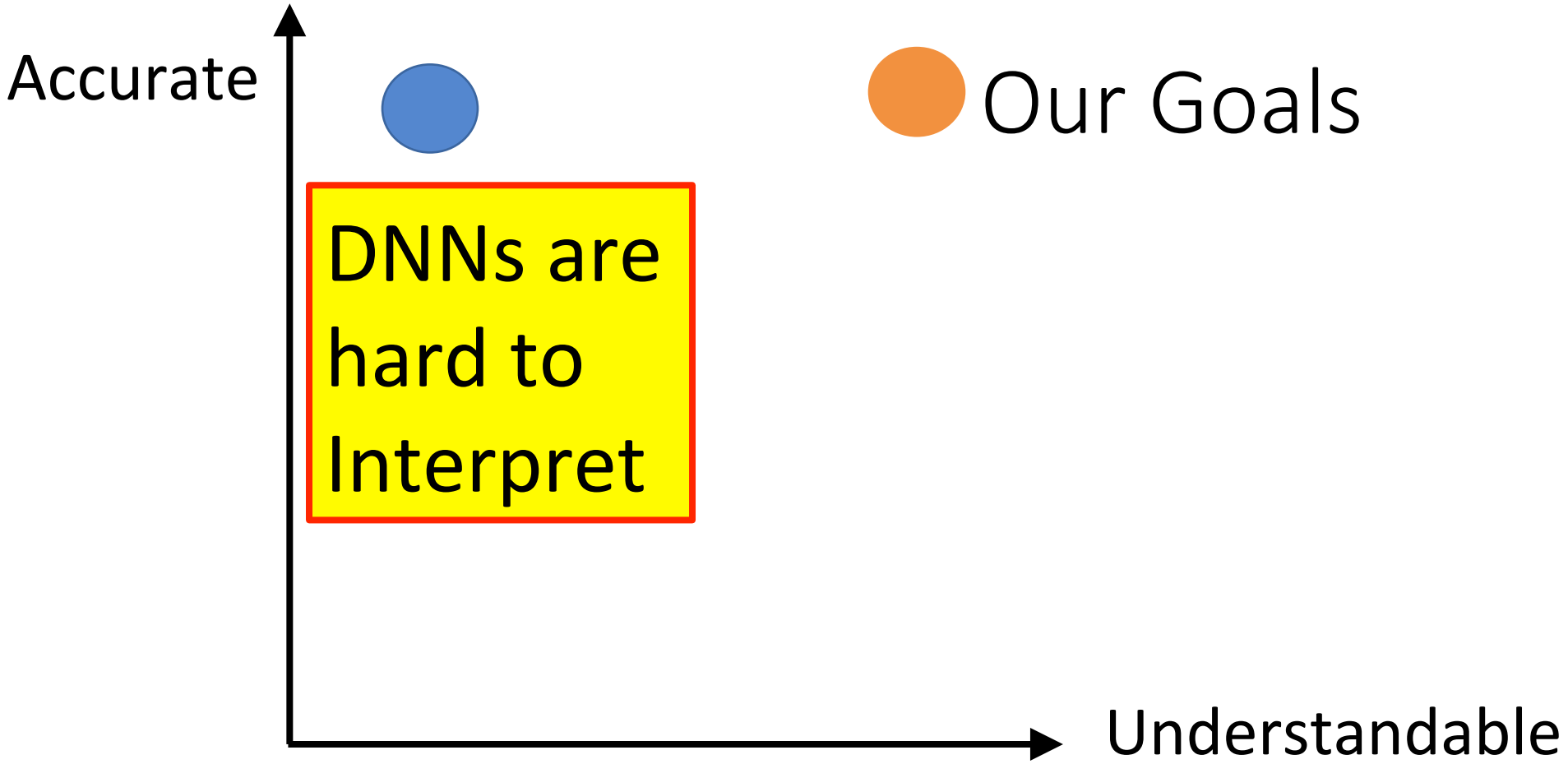
Full Review... | May 25, 2006



ATGCGATCAAGTCTG



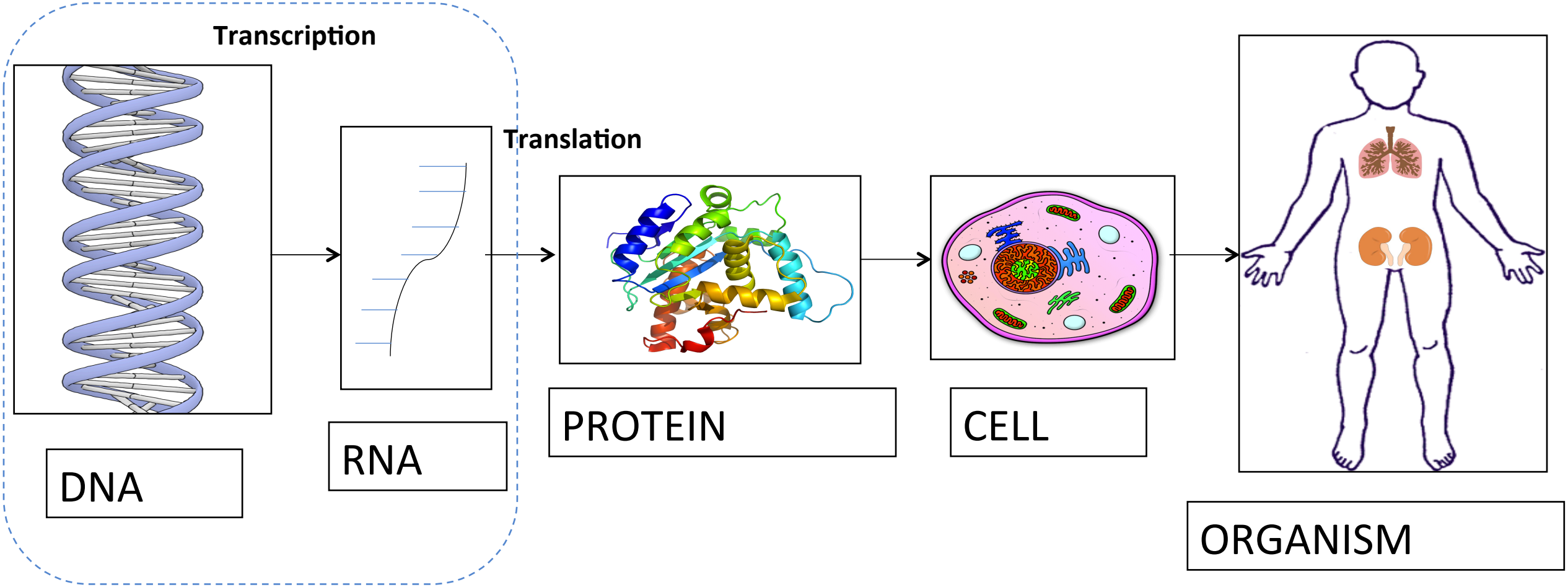
"Protein Binding Site"


Accurate
Our Goals
DNNs are hard to Interpret
Understandable

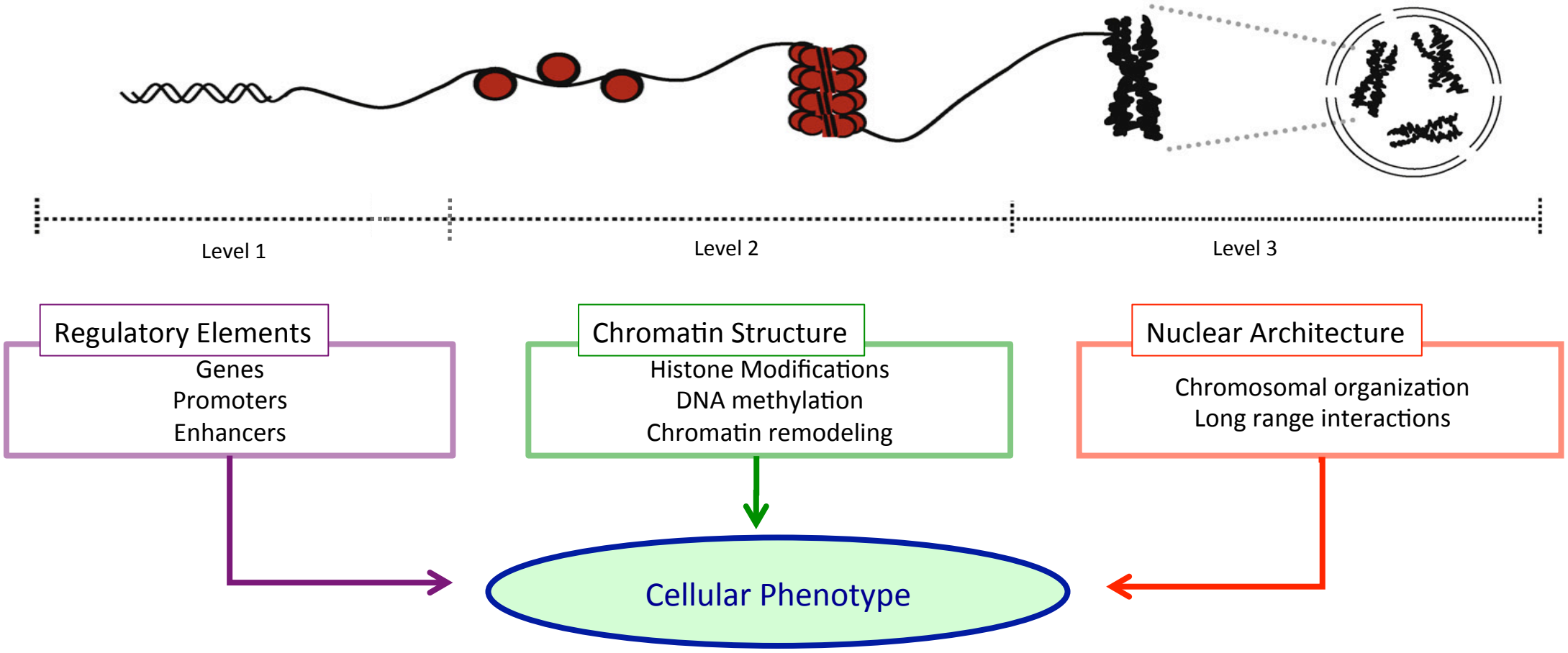
Roadmap

- ✧Background of Machine Learning
- ✧Background of Sequential Data about Gene Regulation
- ✧AttentiveChrome for understanding gene regulation by selective attention on chromatin

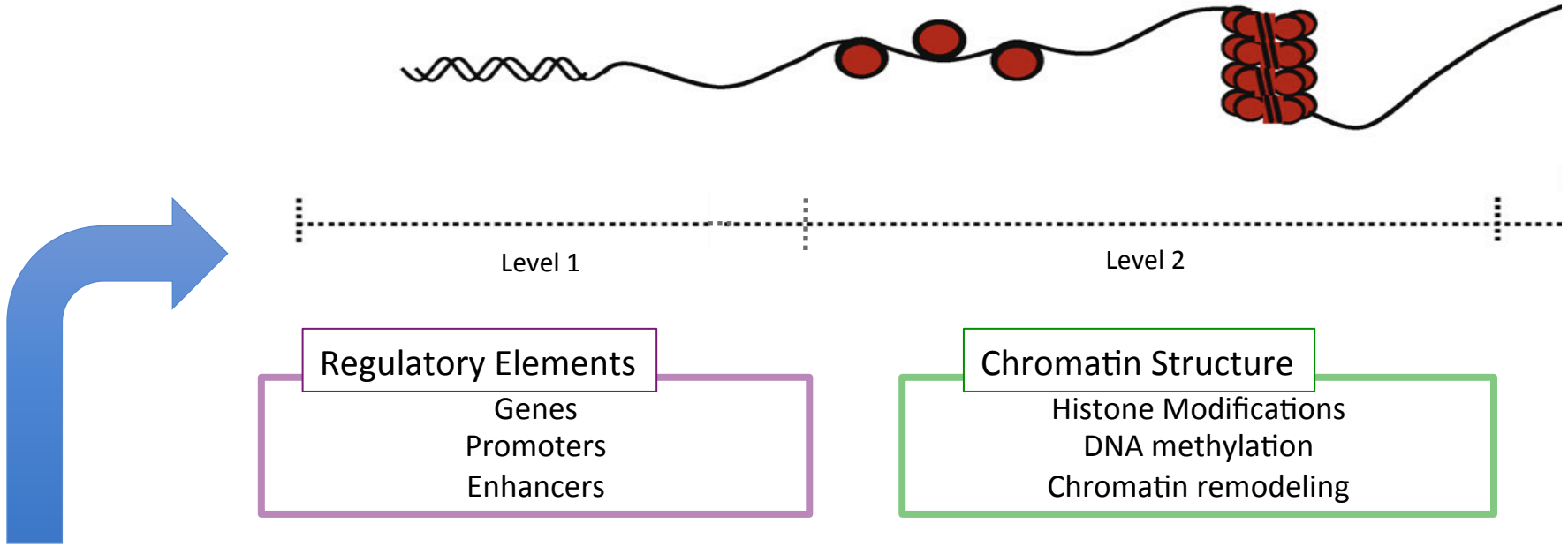
Biology in a Slide



Genome Organization and Gene Regulation

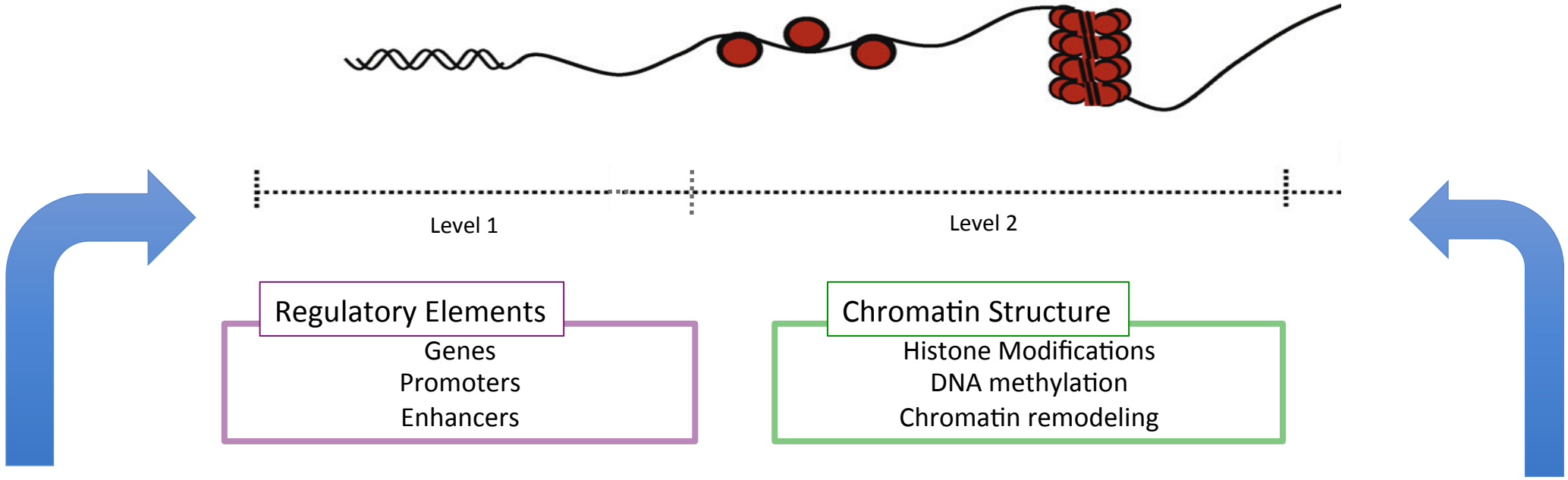


(adapted from Babu et al., 2008)



ENCODE Project (2003-Present)

Describe the functional elements encoded in human DNA



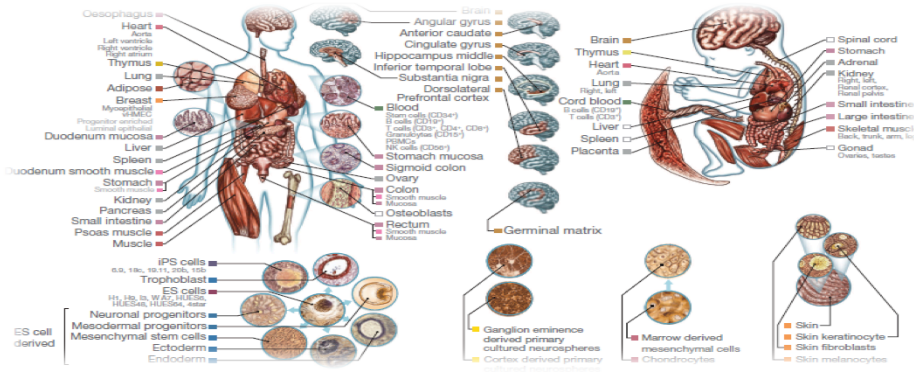
ENCODE Project (2003-)

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)

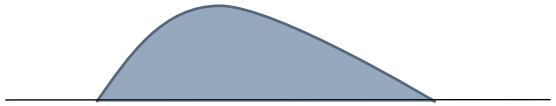


Current Available Large-Scale Data about Gene Transcription

DNA Segments on Genomes

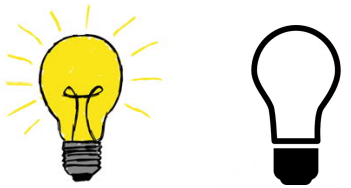
ATGCGATCAAGTCTG

TF Binding Signals



Histone Modification Signals

Gene Expression

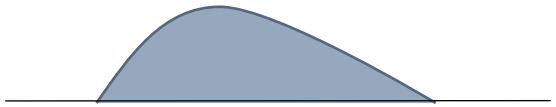


Two Important Data-Driven Computational Tasks

DNA Segments on Genomes

ATGCGATCAAGTCTG

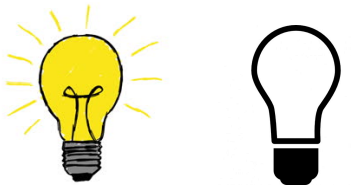
TF Binding Signals

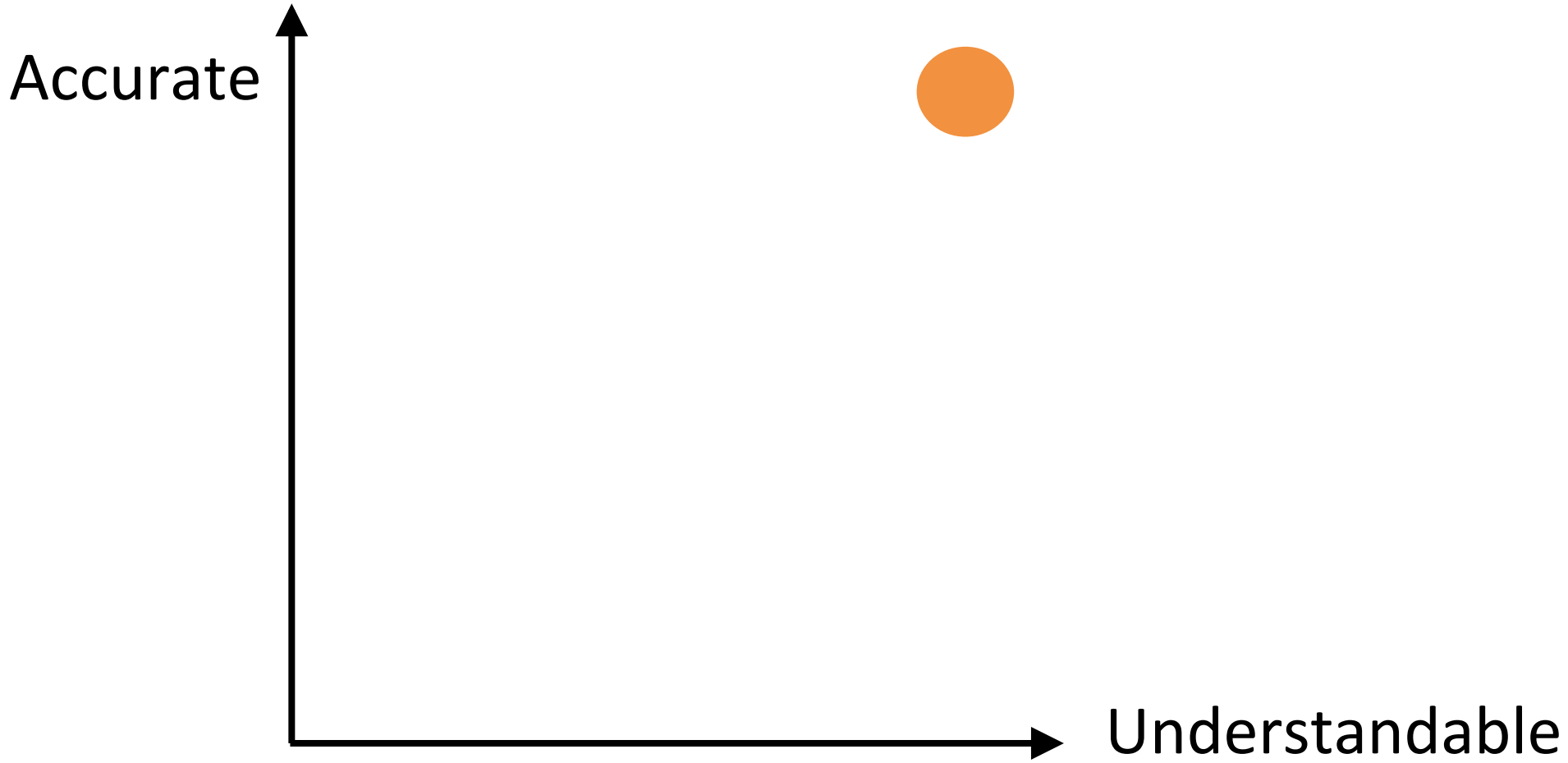


Histone Modification Signals

This Talk

Gene Expression



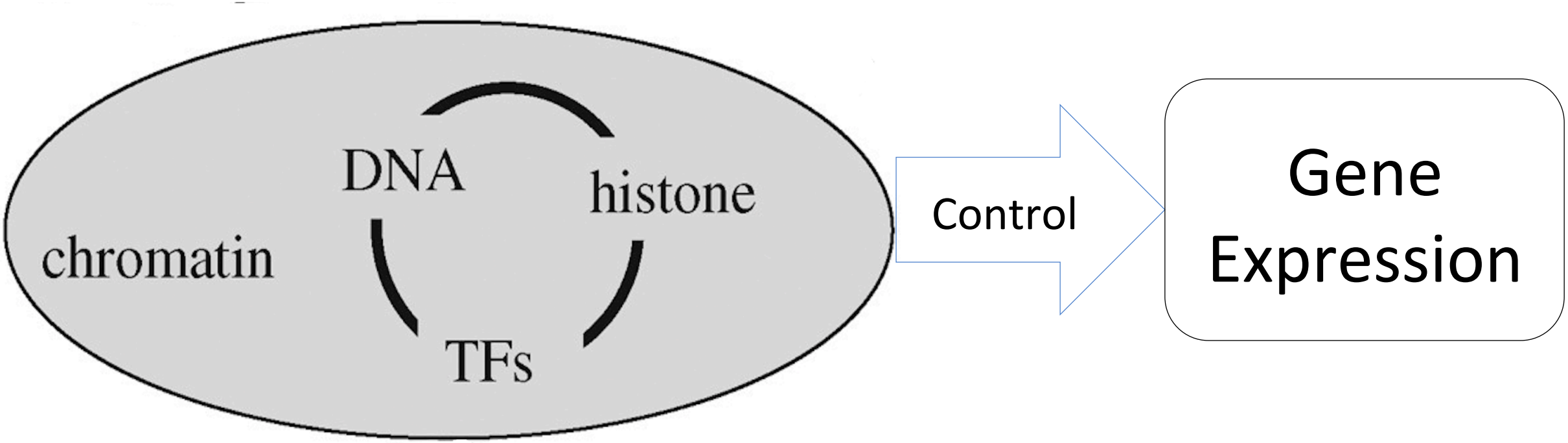


One main aim of such data analysis is to understand data and to discover knowledge.

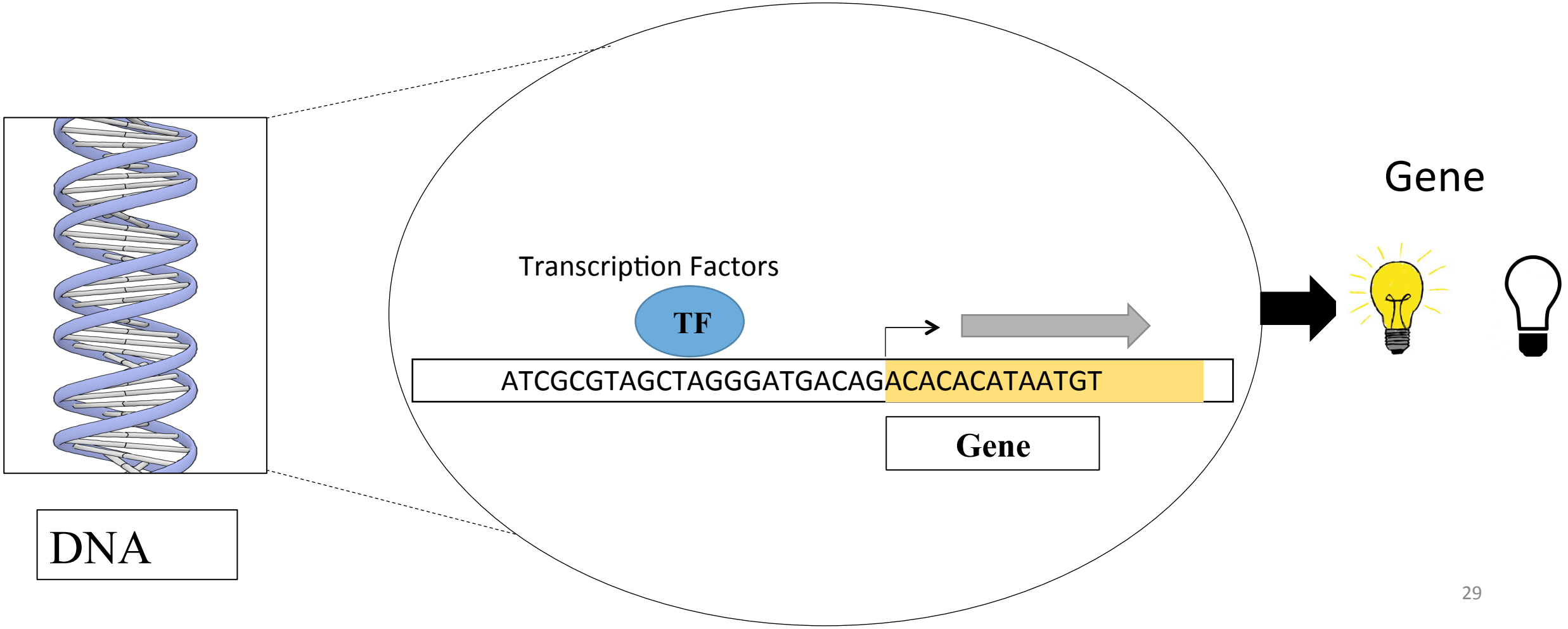
Roadmap

- ✧Background of Machine Learning
- ✧Background of Sequential Data about Gene Regulation
- ✧AttentiveChrome for understanding gene regulation by selective attention on chromatin

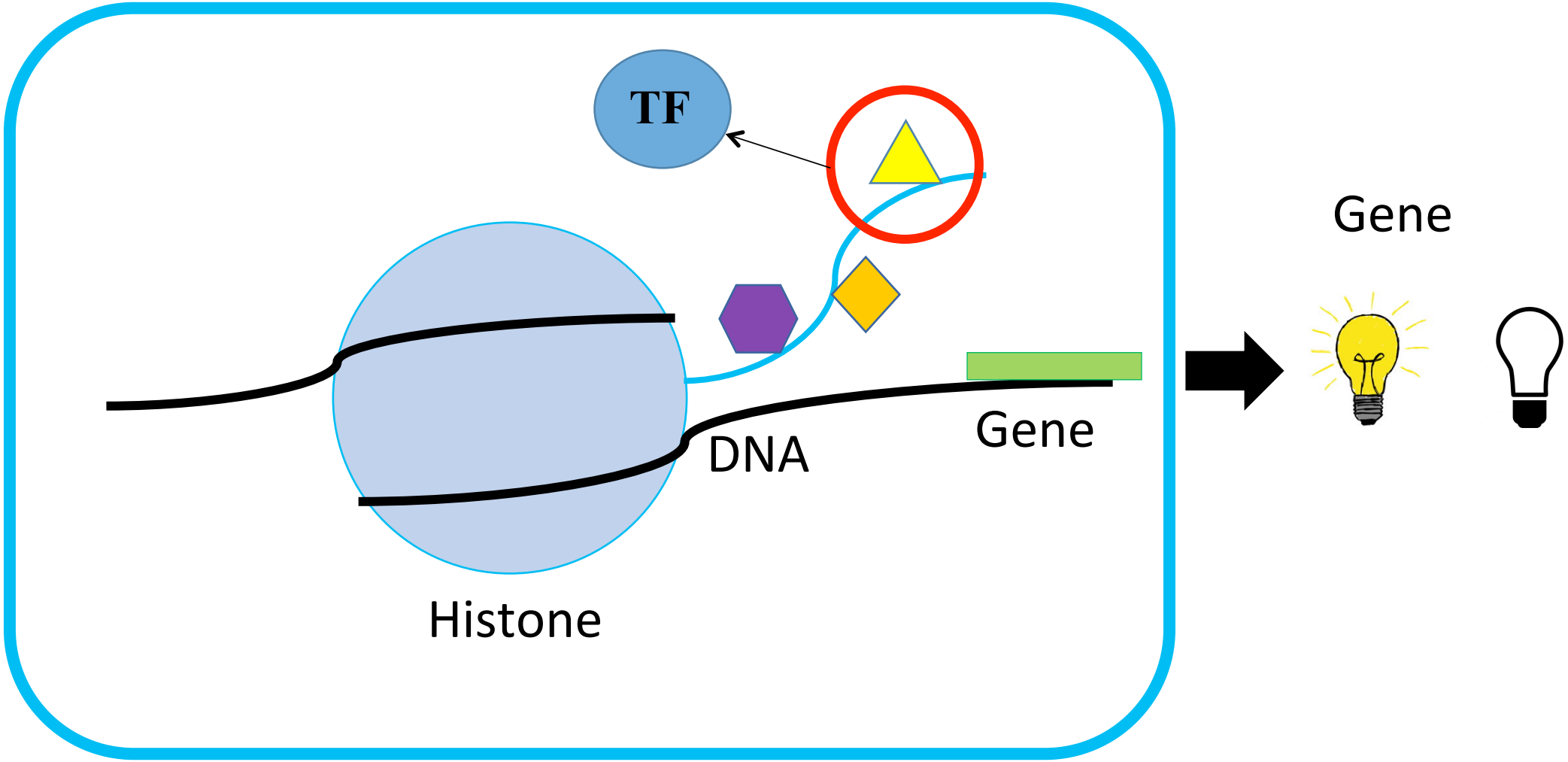
Chromatin



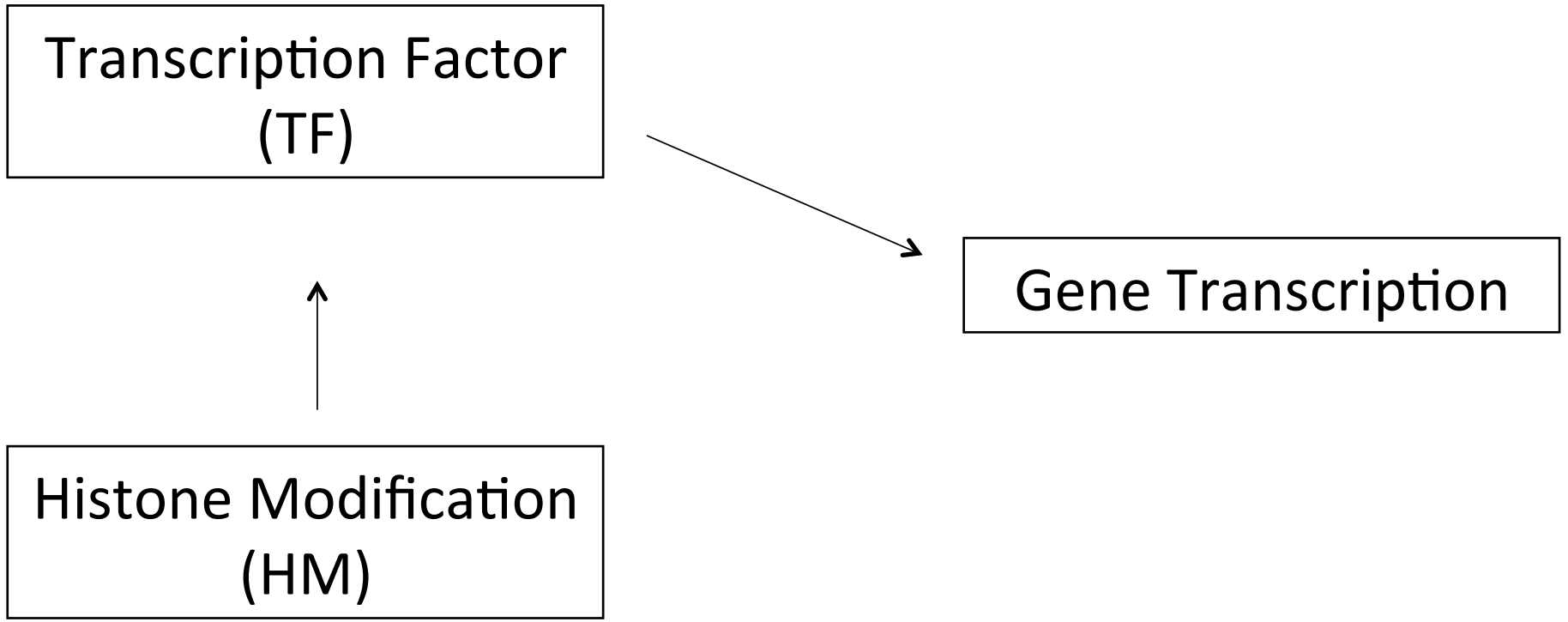
Transcription Factor Binding => Gene Transcription



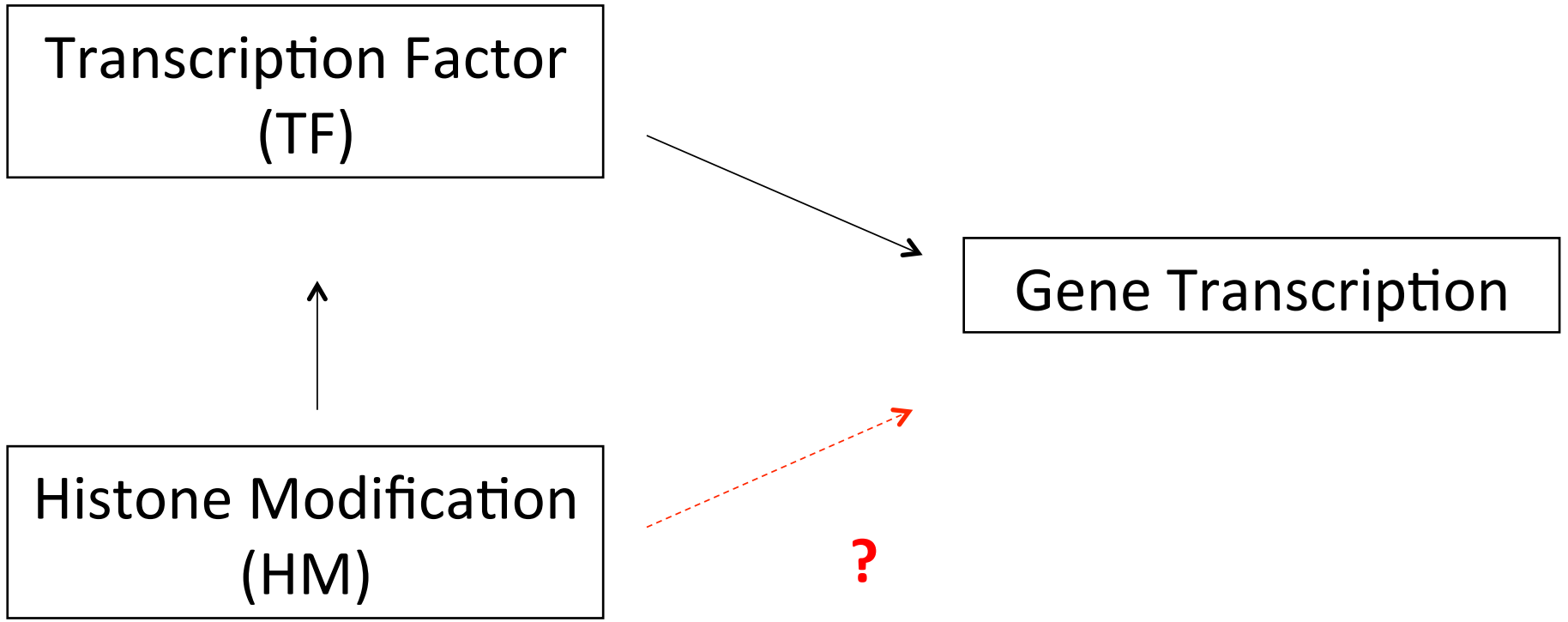
Histone Modifications (HM)



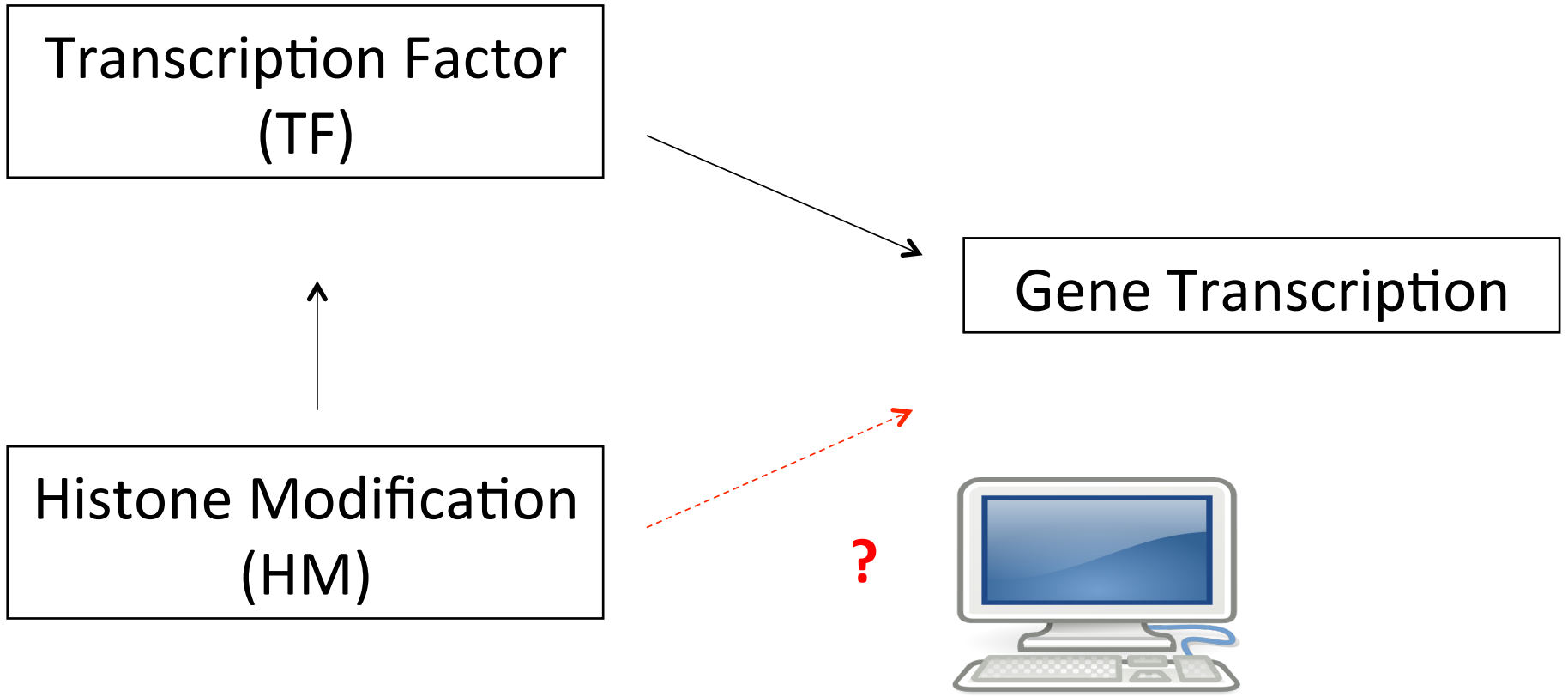
Histone Modification and Gene Transcription



Histone Modification and Gene Transcription



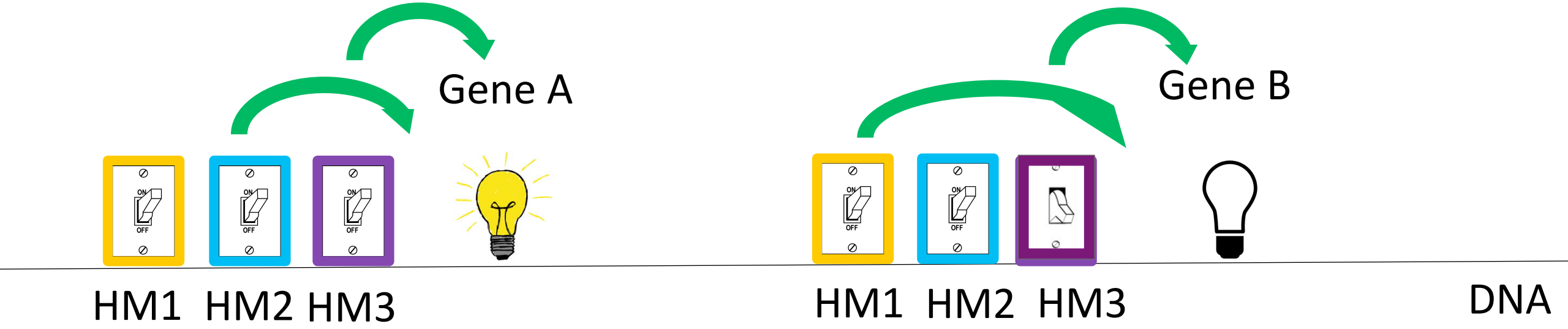
Histone Modification and Gene Transcription



Why Studying [HM => Gene Expression] ?

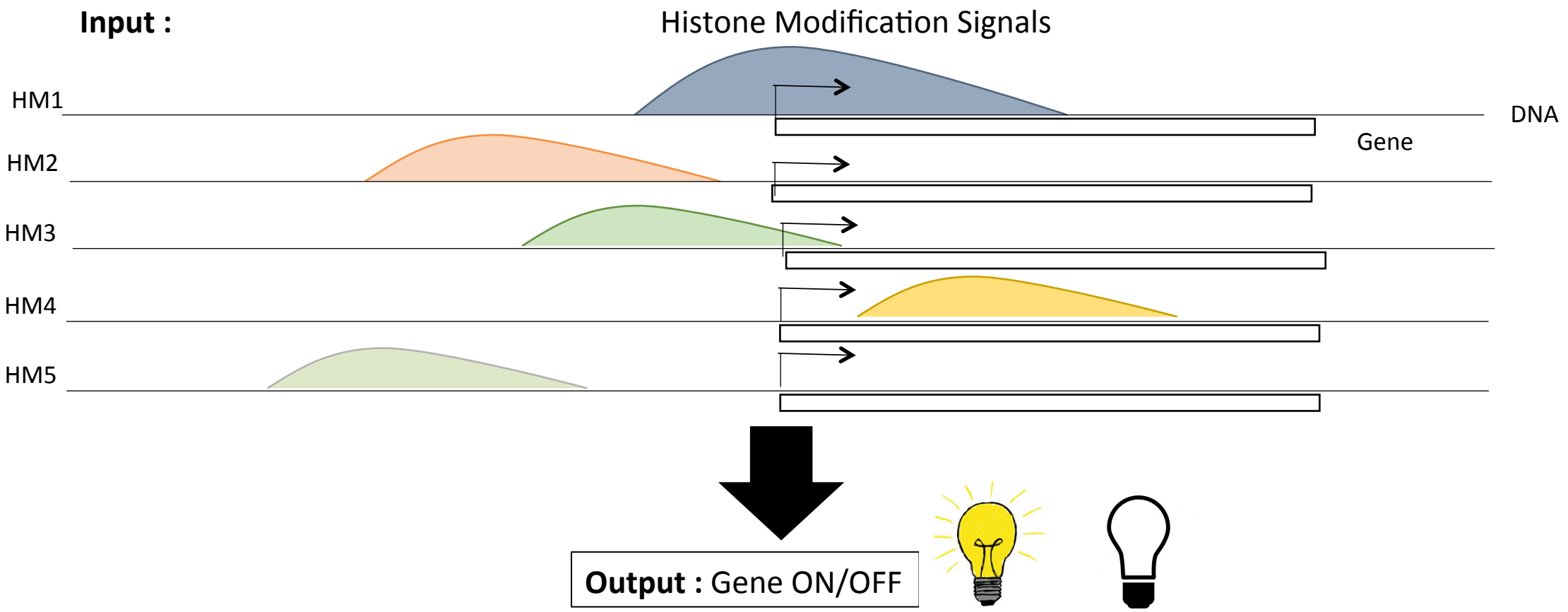
- Epigenomics:
  - Study of chemical changes in DNA and histones (without altering DNA sequence)
  - Inheritable and involved in regulating gene expression, development, tissue differentiation and suppression …
- Modification in DNA/histones (changes in chromatin structure and function)
  - => influence 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 what HMs affect which genes in what cells?

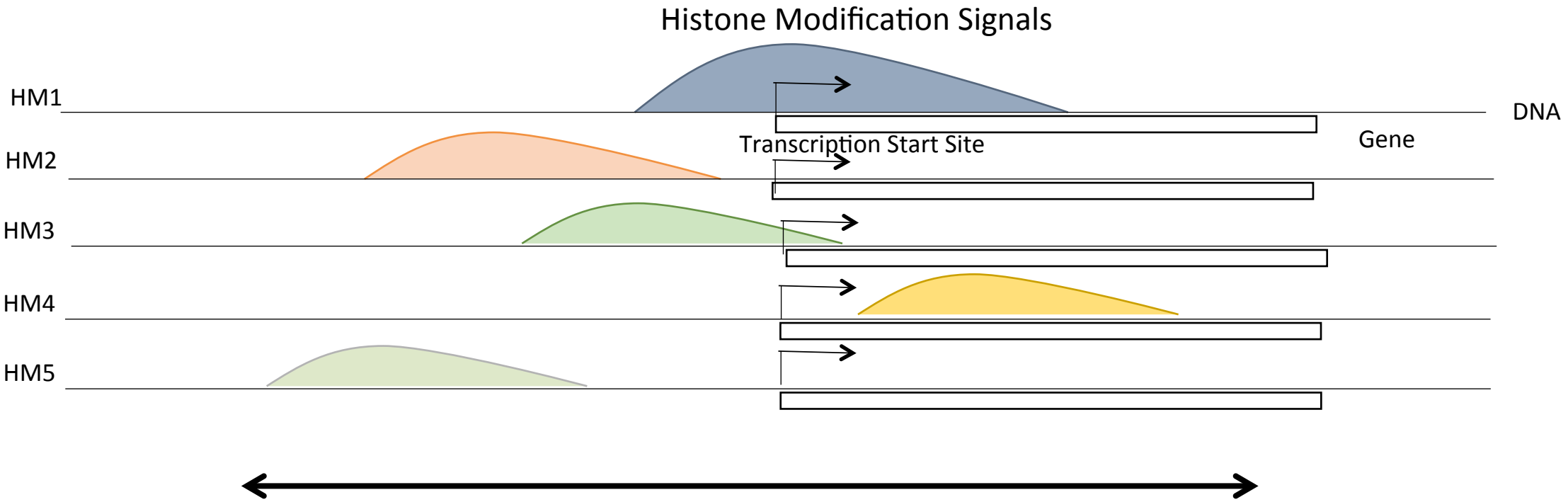


Task Formulation

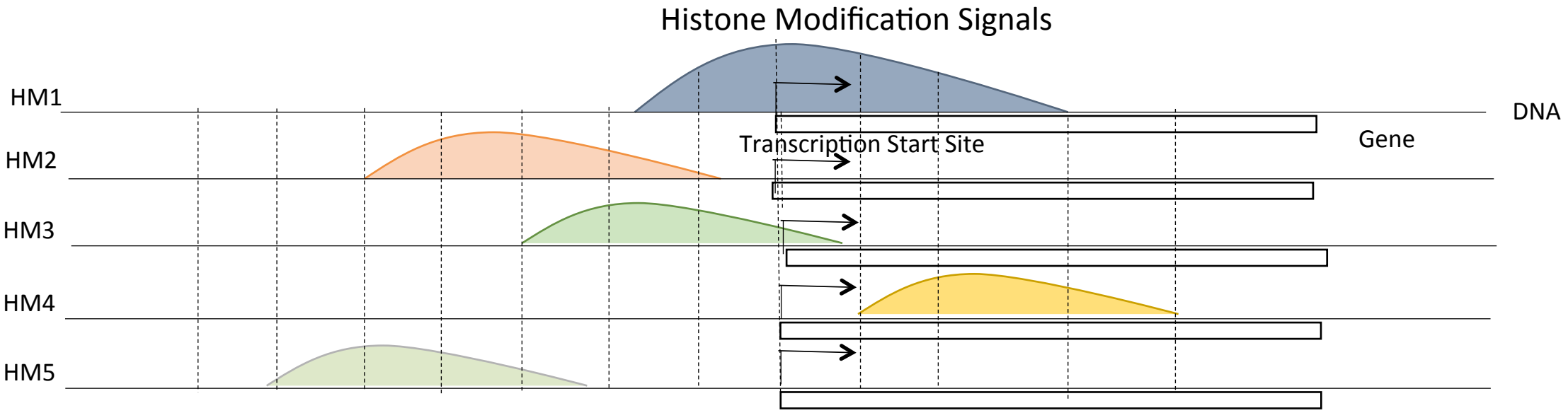
Prediction Task



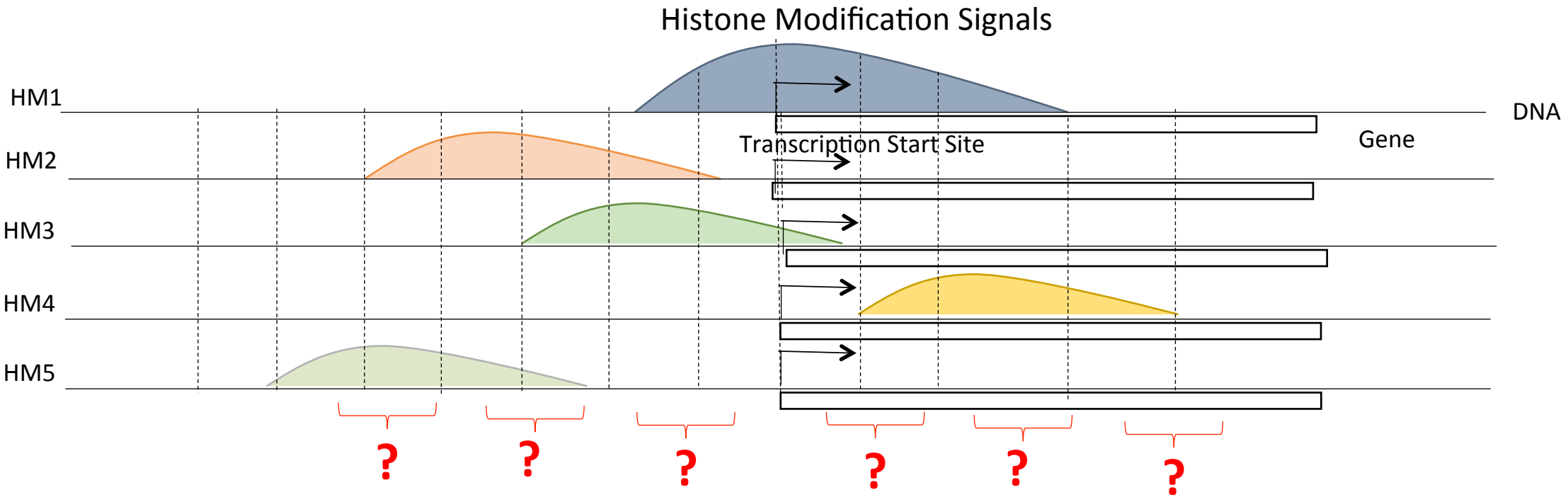
Input


Histone Modification Signals
HM1
HM2
HM3
HM4
HM5
DNA
Gene
Transcription Start Site

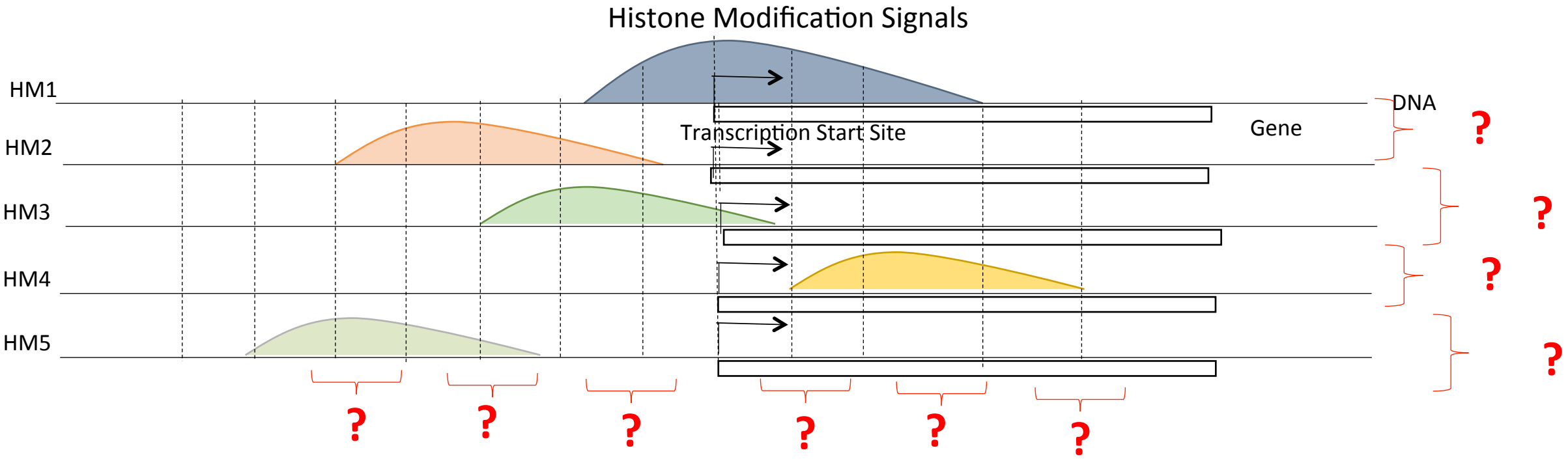
Input


Histone Modification Signals
HM1
HM2
HM3
HM4
HM5
DNA
Gene
Transcription Start Site

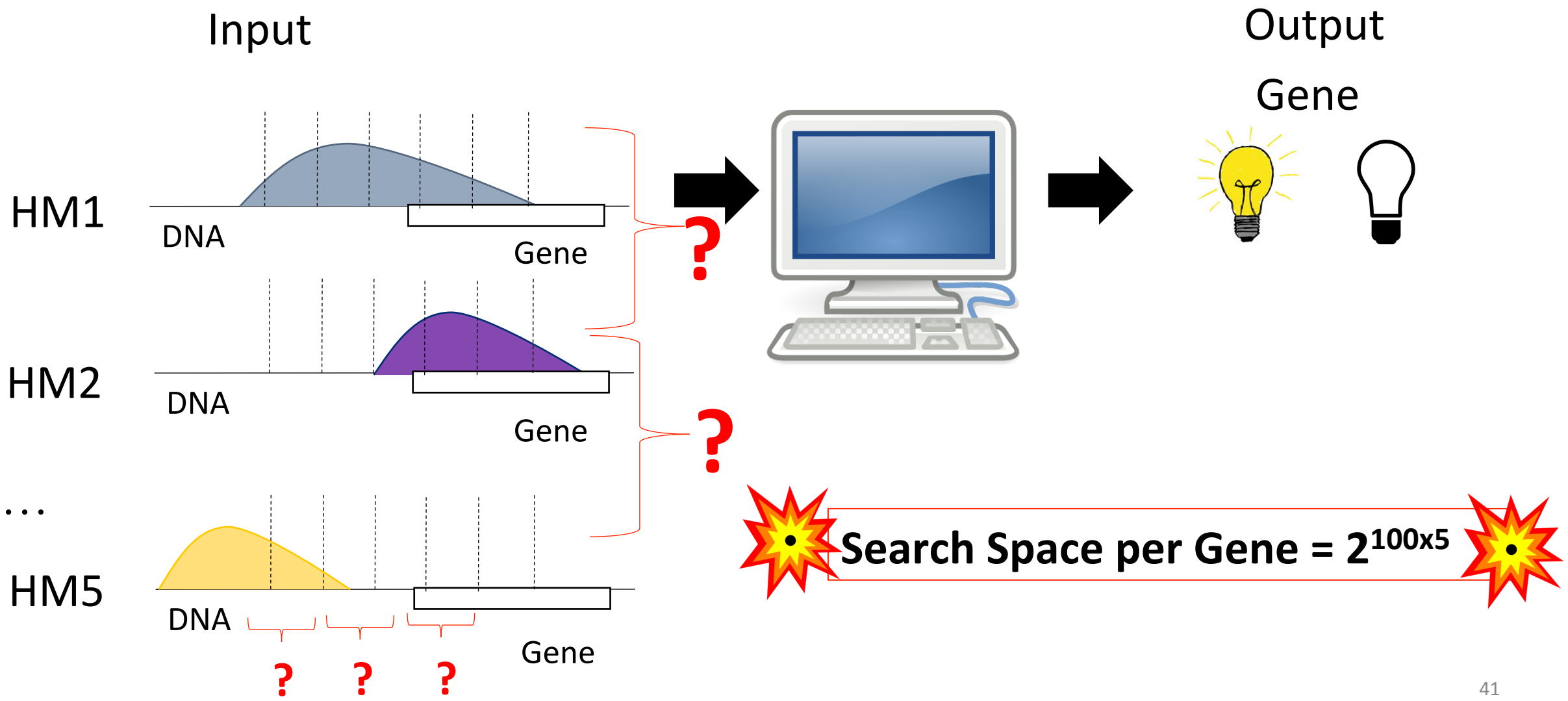
Challenge


Histone Modification Signals
HM1
HM2
HM3
HM4
HM5
DNA
Gene
Transcription Start Site
?
?
?
?
?
?

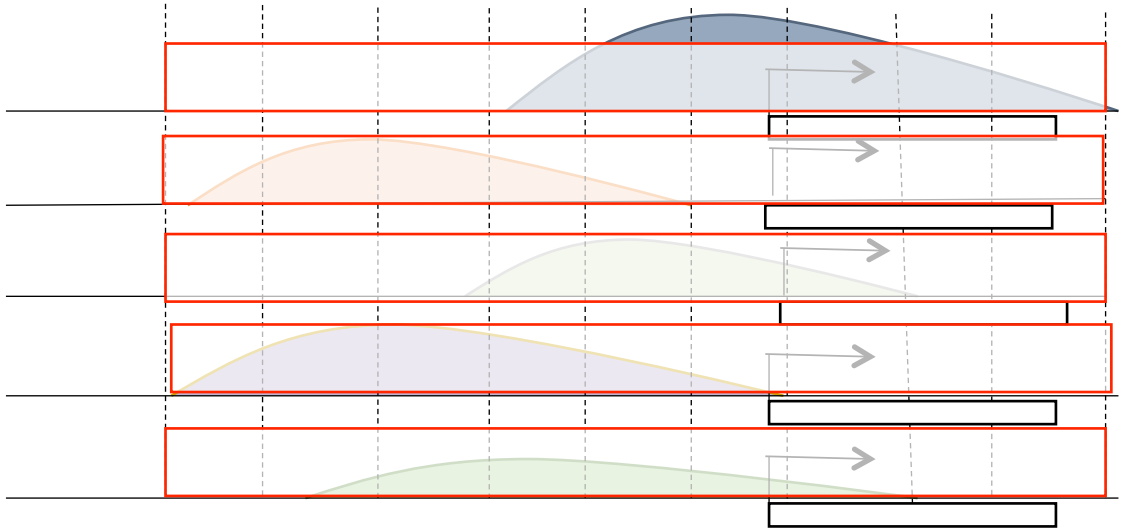
Challenge


Histone Modification Signals
HM1
HM2
HM3
HM4
HM5
Transcription Start Site
Gene
DNA
?
?
?
?
?
?
?
?
?
?
?

Computational Challenge



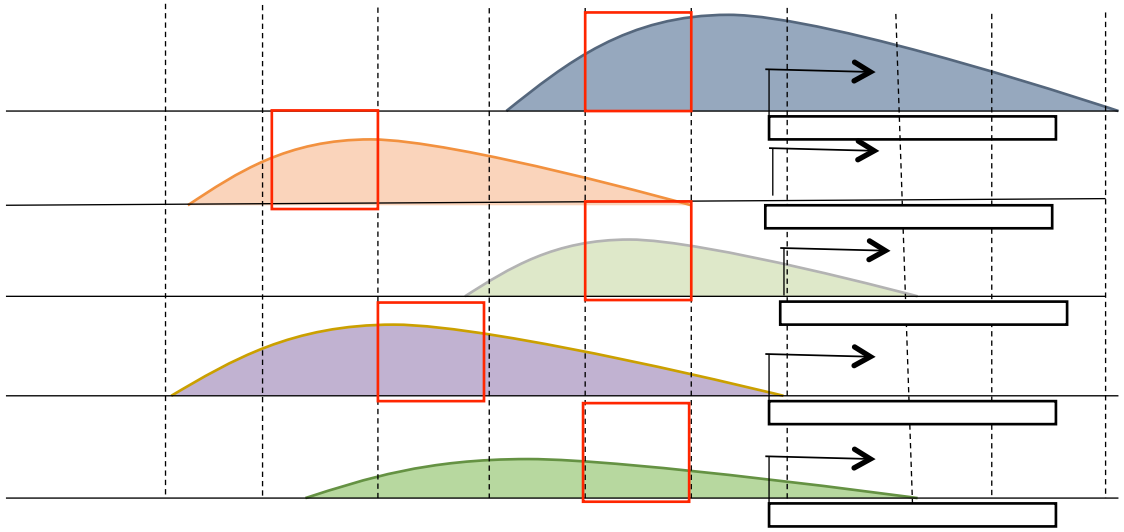
Related Work



Linear Regression,
SVM,
Random Forest

Gene ON/OFF

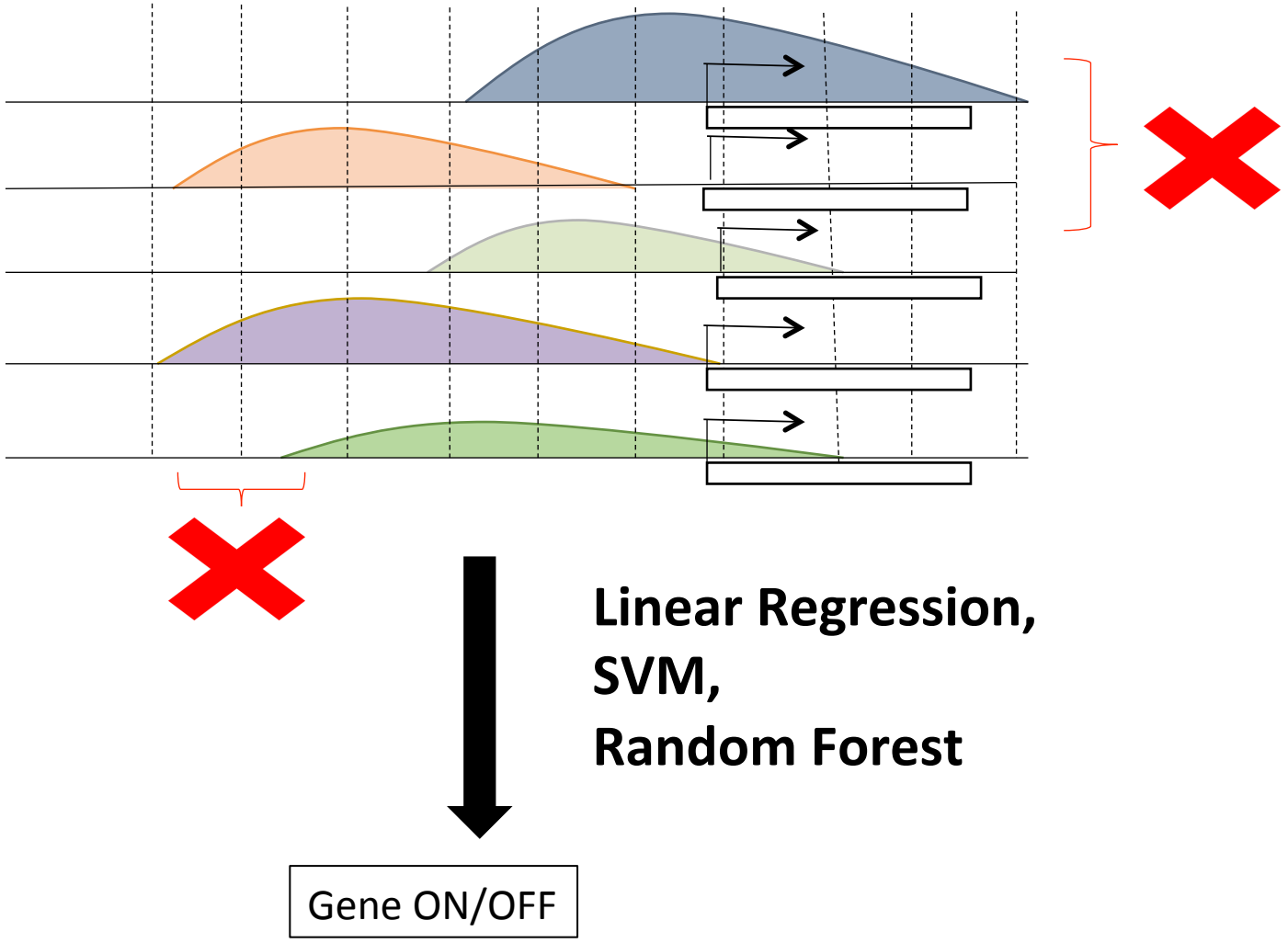
Related Work



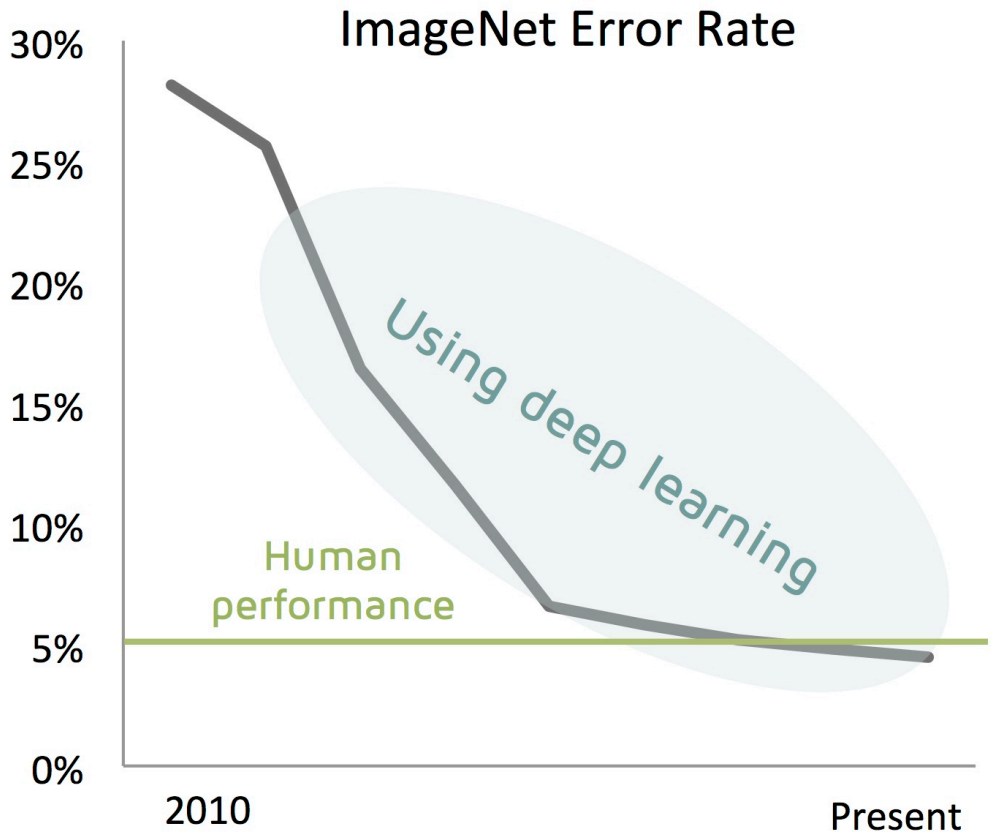
Linear Regression,
SVM,
Random Forest

Gene ON/OFF

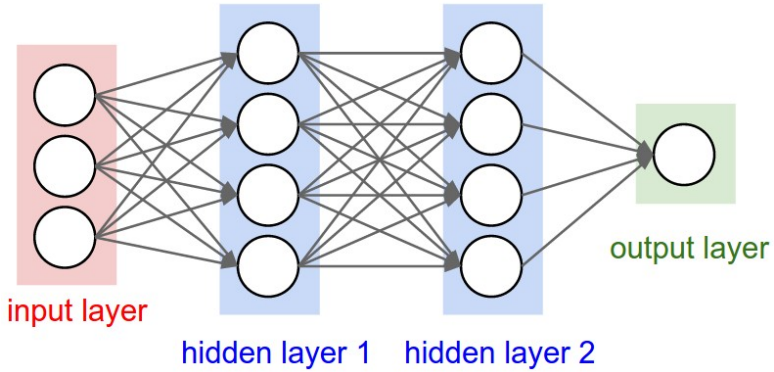
Drawback of Related Works



Deep Learning to Rescue:

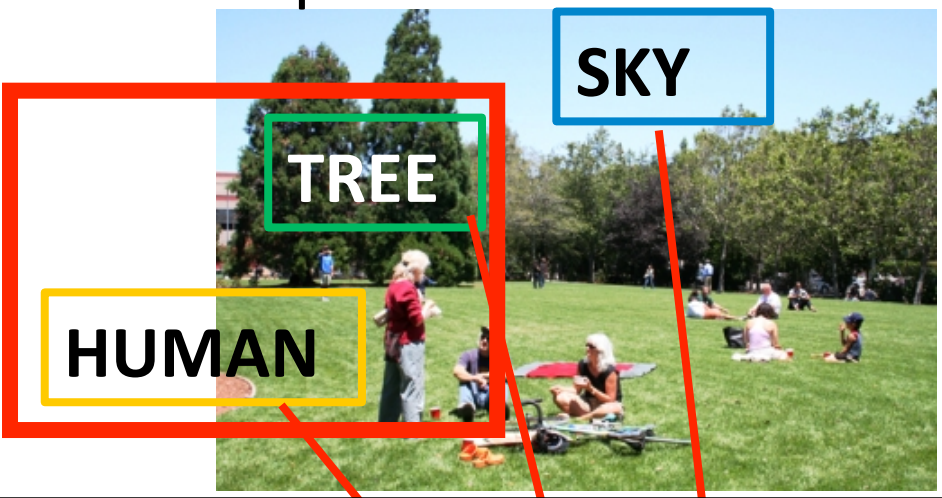


Deep Neural Network (DNN)

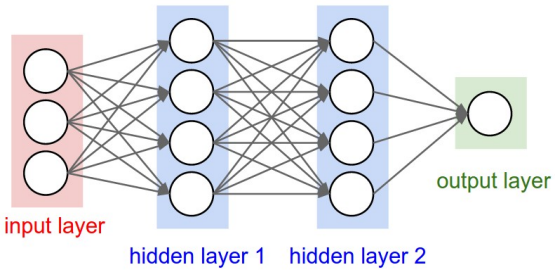


Deep Learning to Rescue:

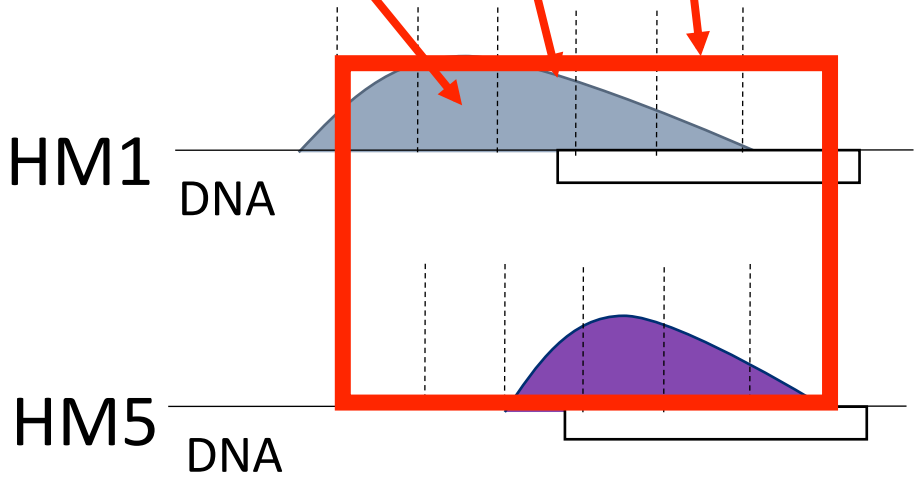
Input

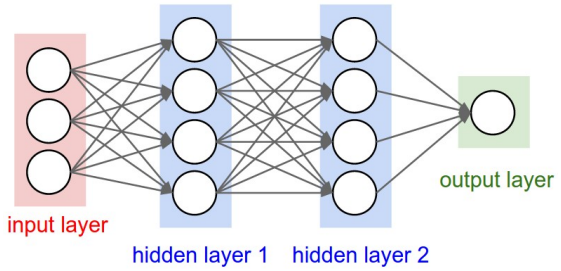


Output

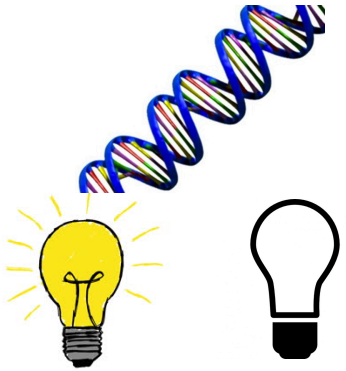


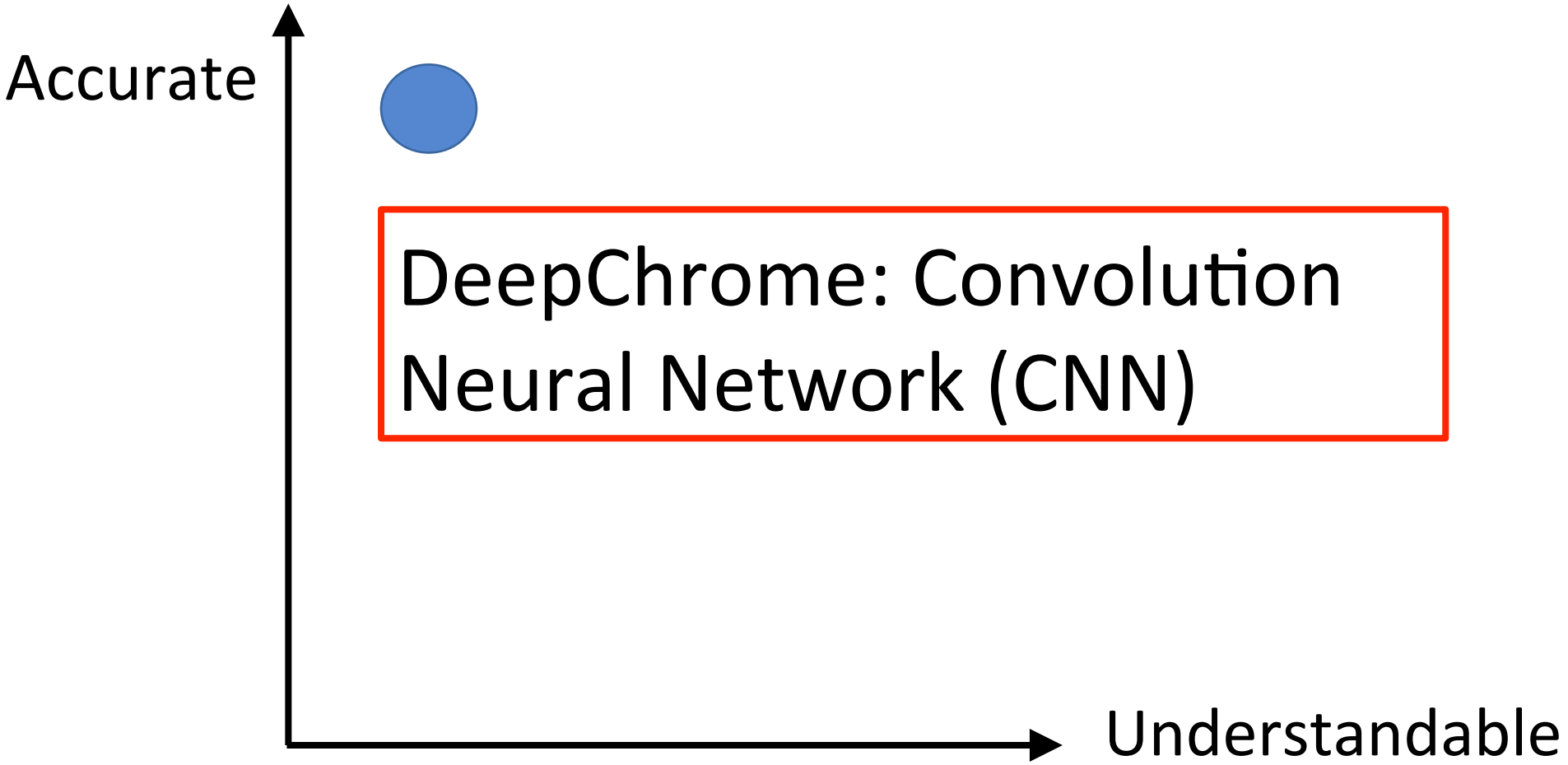
Park





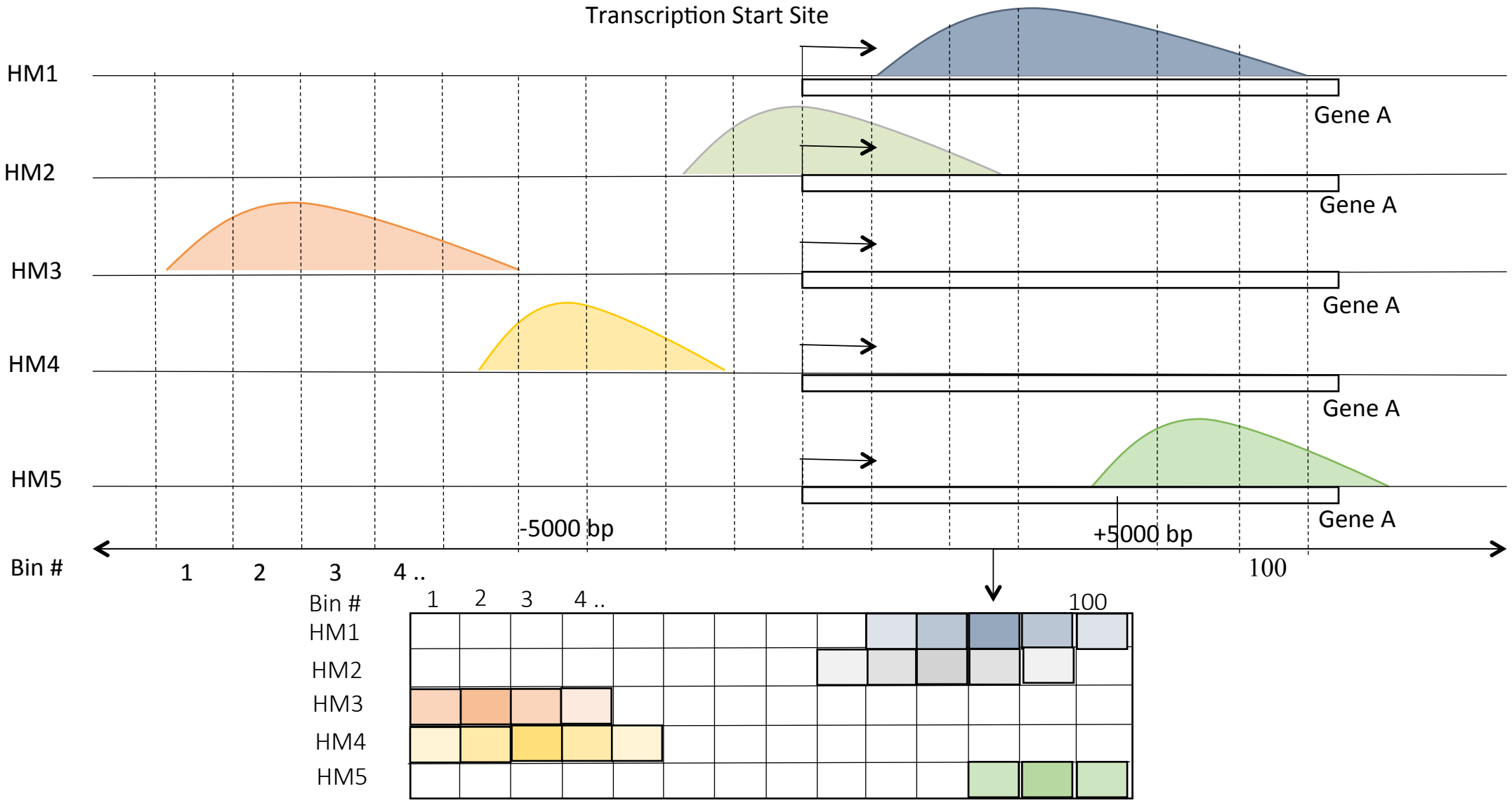
Gene



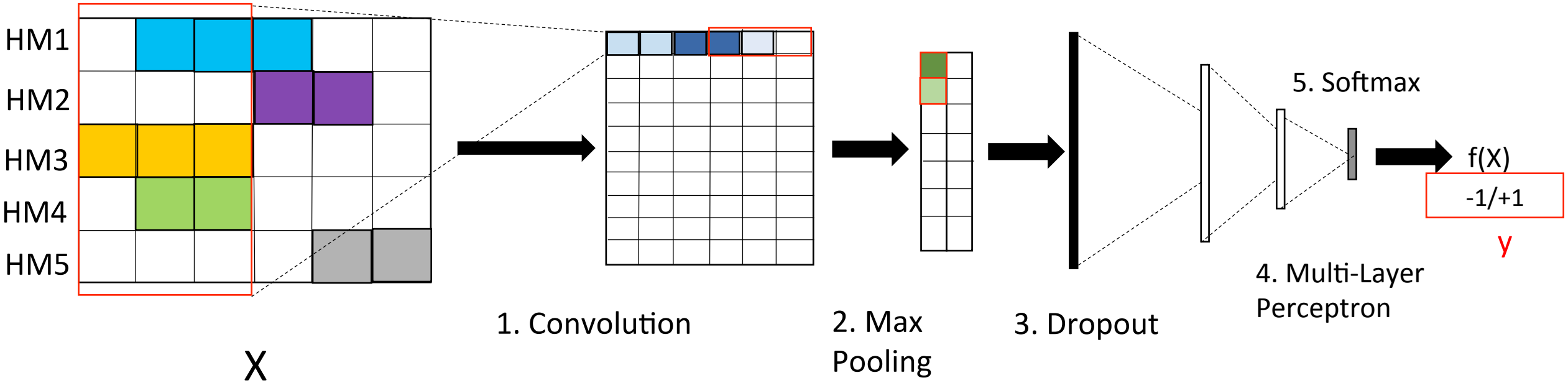


R. Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. "DeepChrome: Deep-learning for predicting gene expression from histone modifications". *Bioinformatics.* (ECCB) (2016)

Data

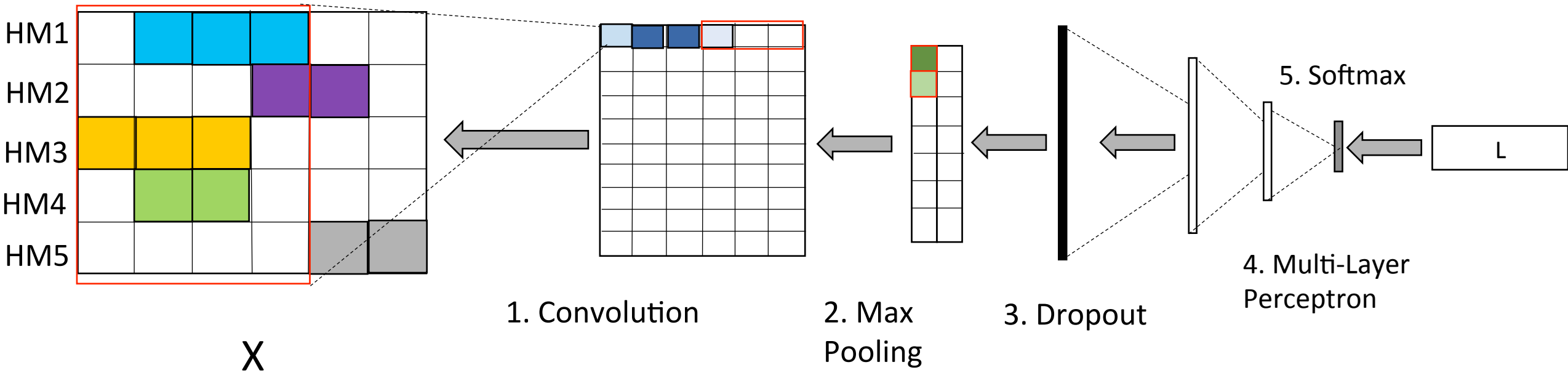

Transcription Start Site
HM1
HM2
HM3
HM4
HM5
Gene A
Gene A
Gene A
Gene A
Gene A
-5000 bp
+5000 bp
Bin #
1
2
3
4 ..
100
Bin #
1
2
3
4 ..
100
HM1
HM2
HM3
HM4
HM5

DeepChrome-Convolutional Neural Network (CNN)



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

DeepChrome-Convolutional Neural Network (CNN)



Back-propagation:

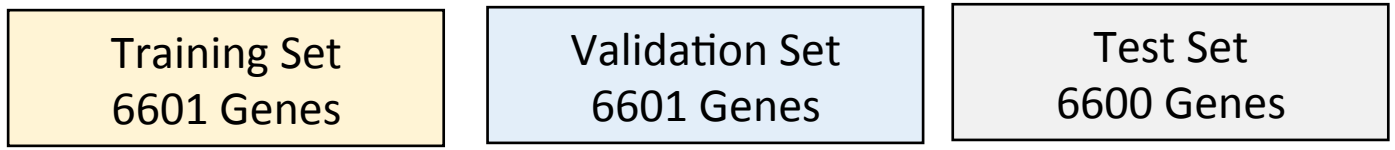
$$\Theta \leftarrow \Theta - \eta \frac{\partial L}{\partial \Theta}$$

Experimental Setup

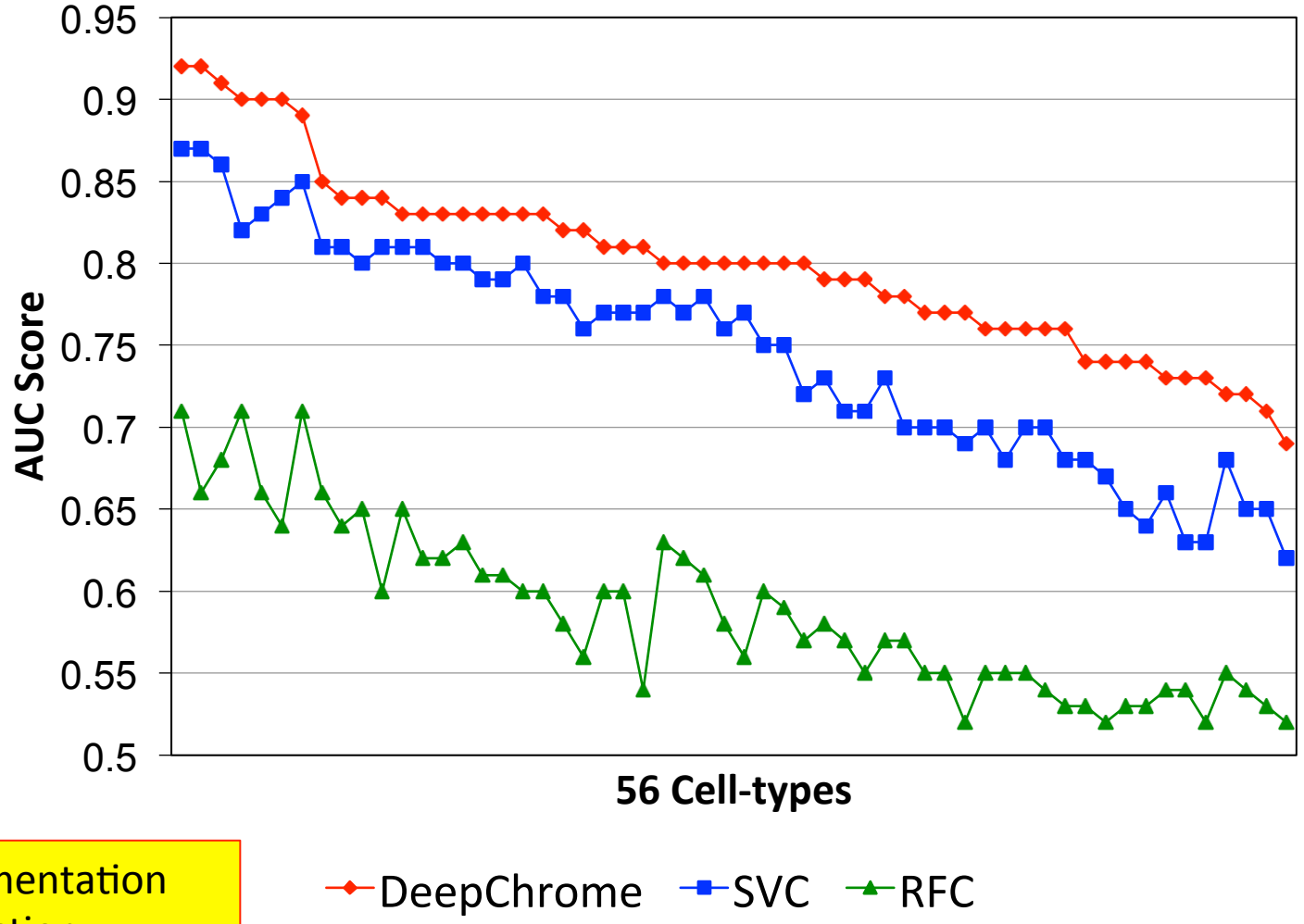
- **Cell-types:** 56
- **Input (HM):** ChIP-Seq Maps (REMC)
- **Output (Gene Expression):** Discretized RNA-Seq (REMC)

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

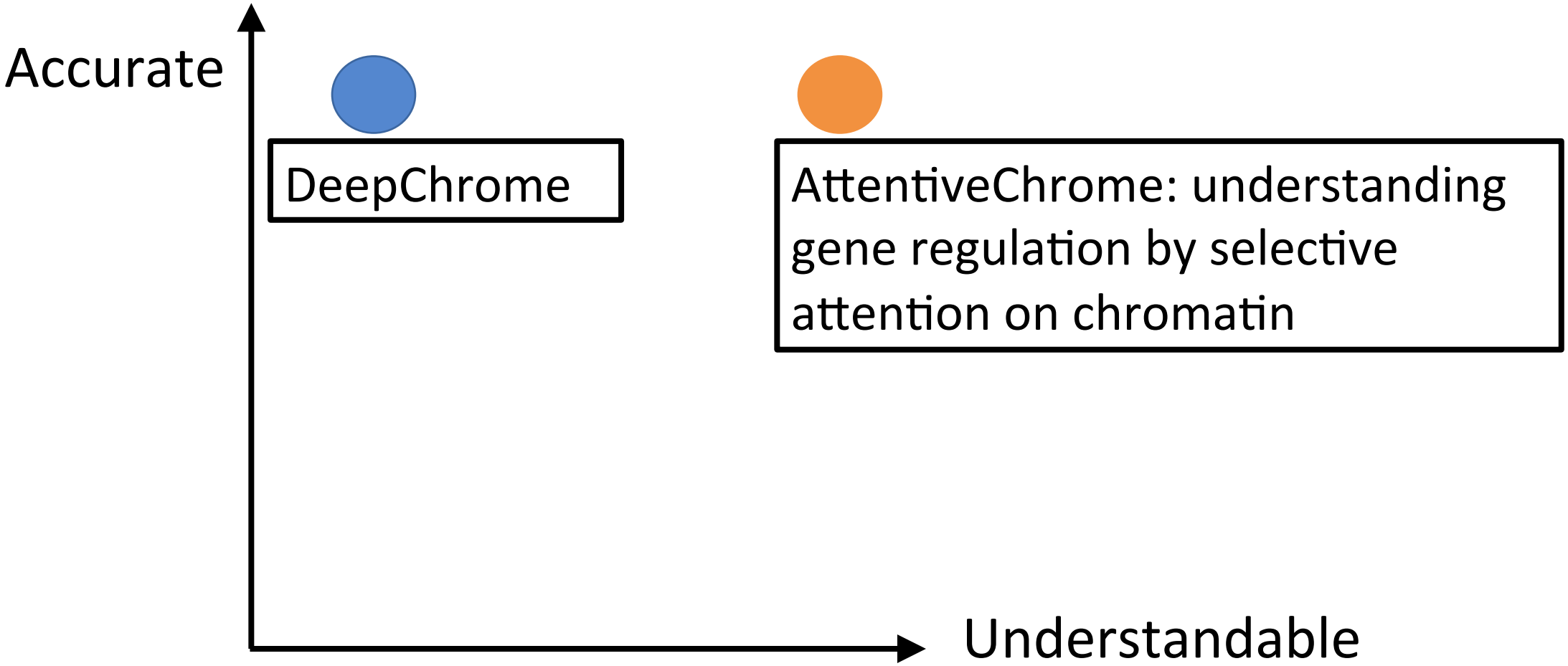
- **Baselines:** Support Vector Classifier (SVC) and Random Forest Classifier (RFC)



Results: Accuracy



- First deep learning implementation for gene expression prediction.
- But hard to interpret.



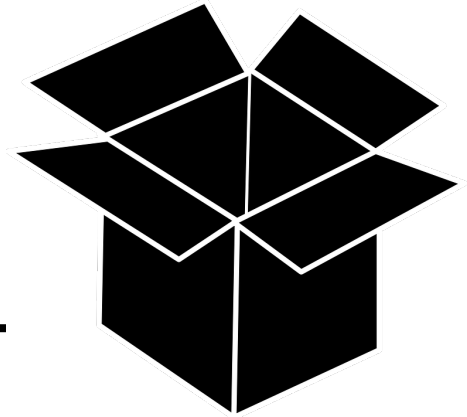
Goal: one DNN both accurate and interpretable

Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, Yanjun Qi, "Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin", to appear at (NIPS-17)

Interpretability by Attention

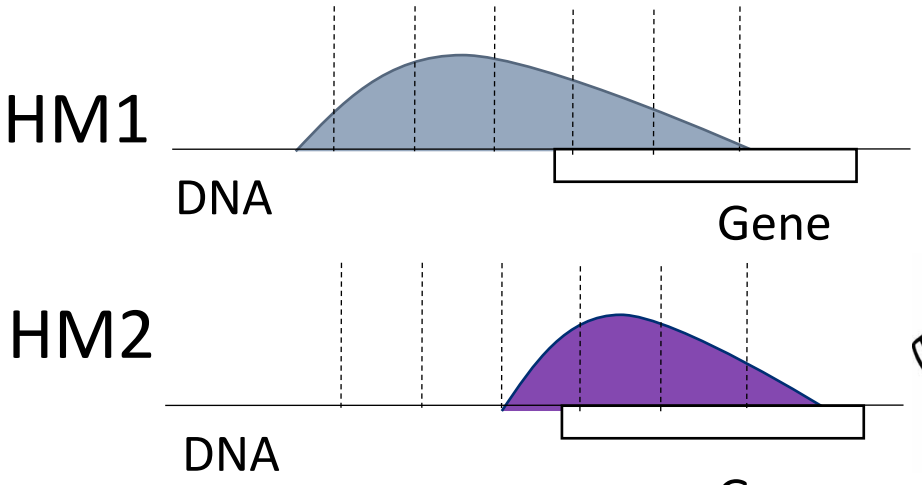
Input

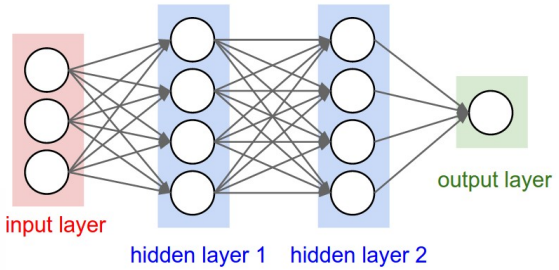
Attention Mechanism



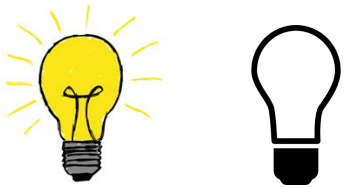
Output

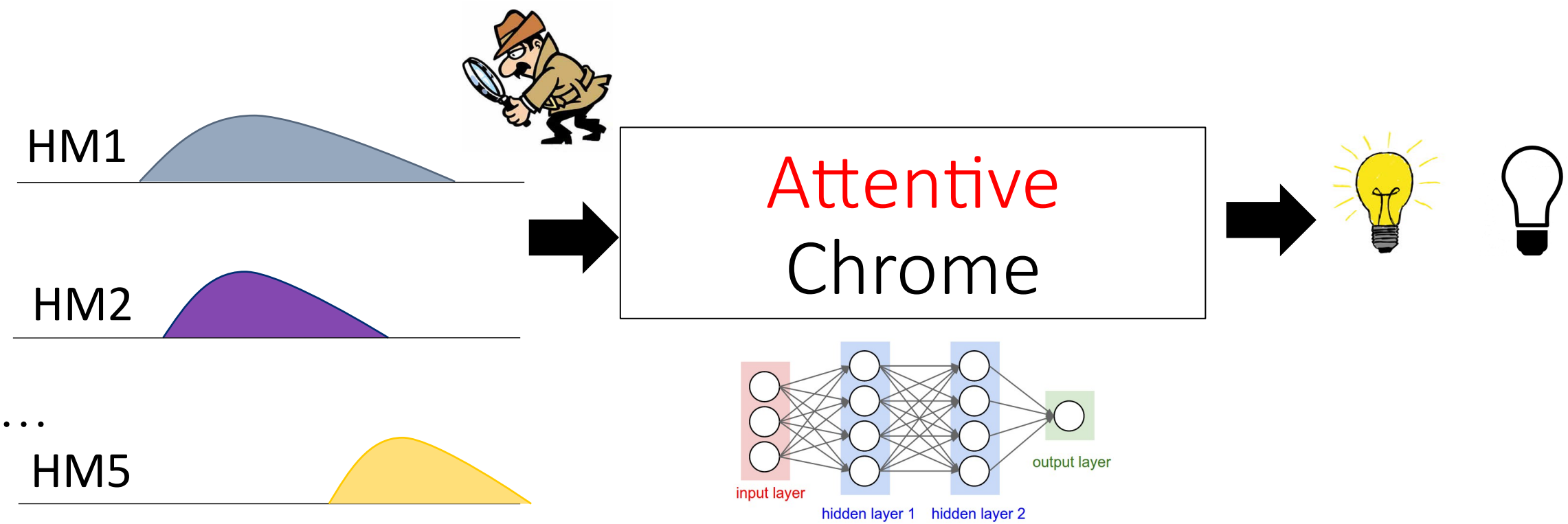
Park





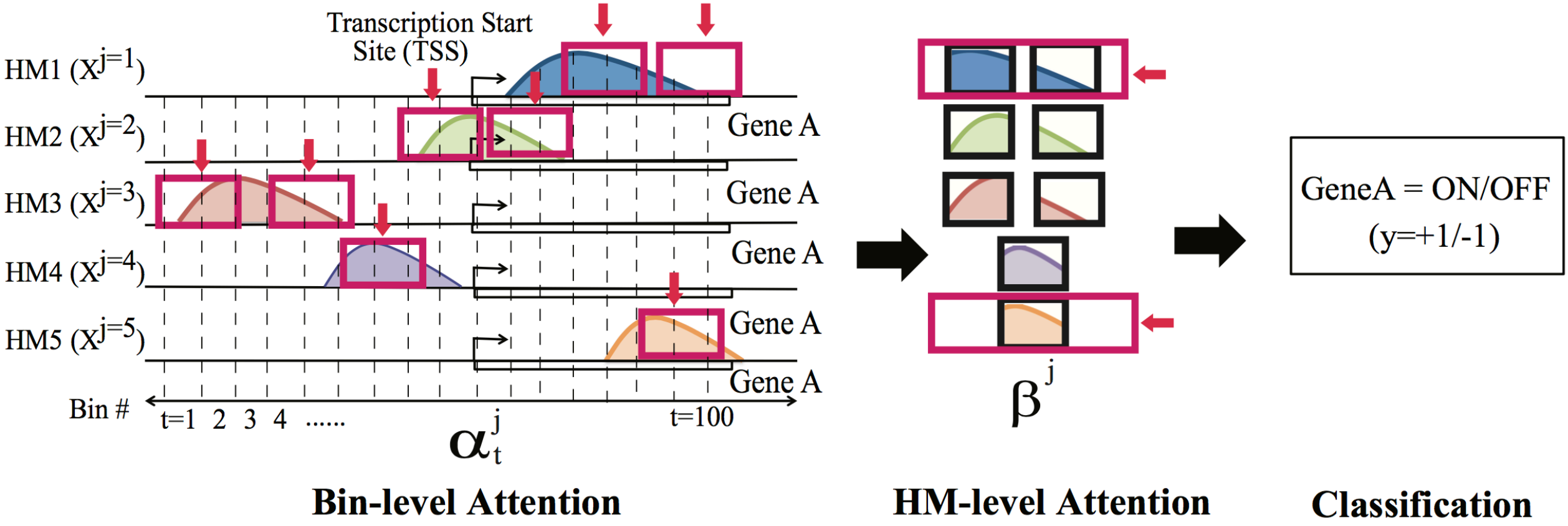
Gene



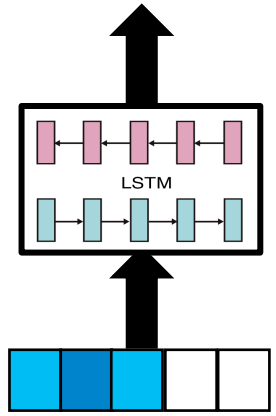


Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, Yanjun Qi, "Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin", to appear at (NIPS-17)

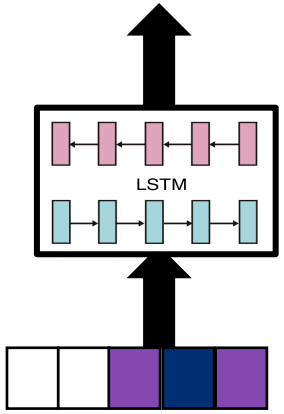
Two Levels of Attention



Input

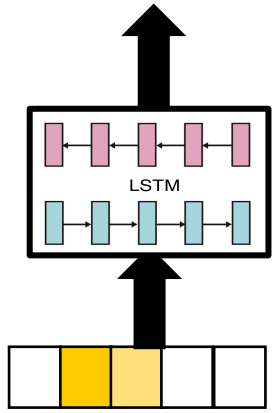


HM1

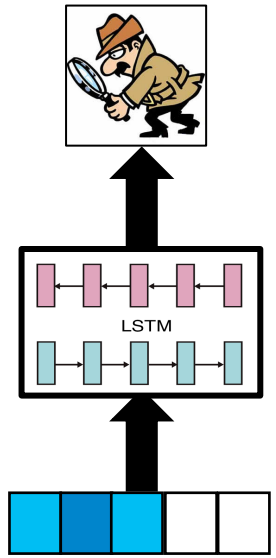


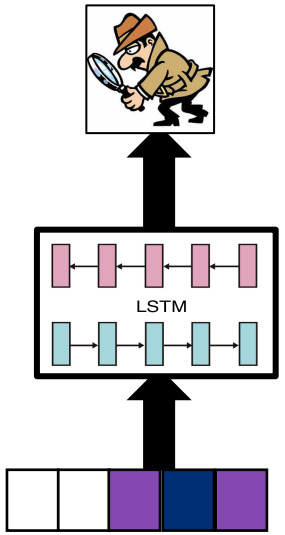
HM2

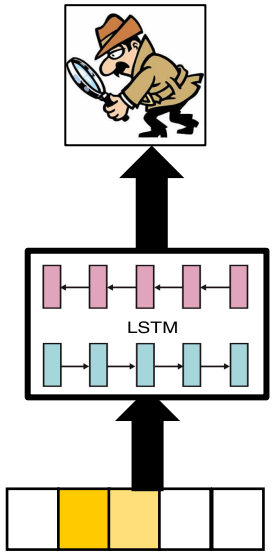
LOCAL LEVEL



HM5







Input

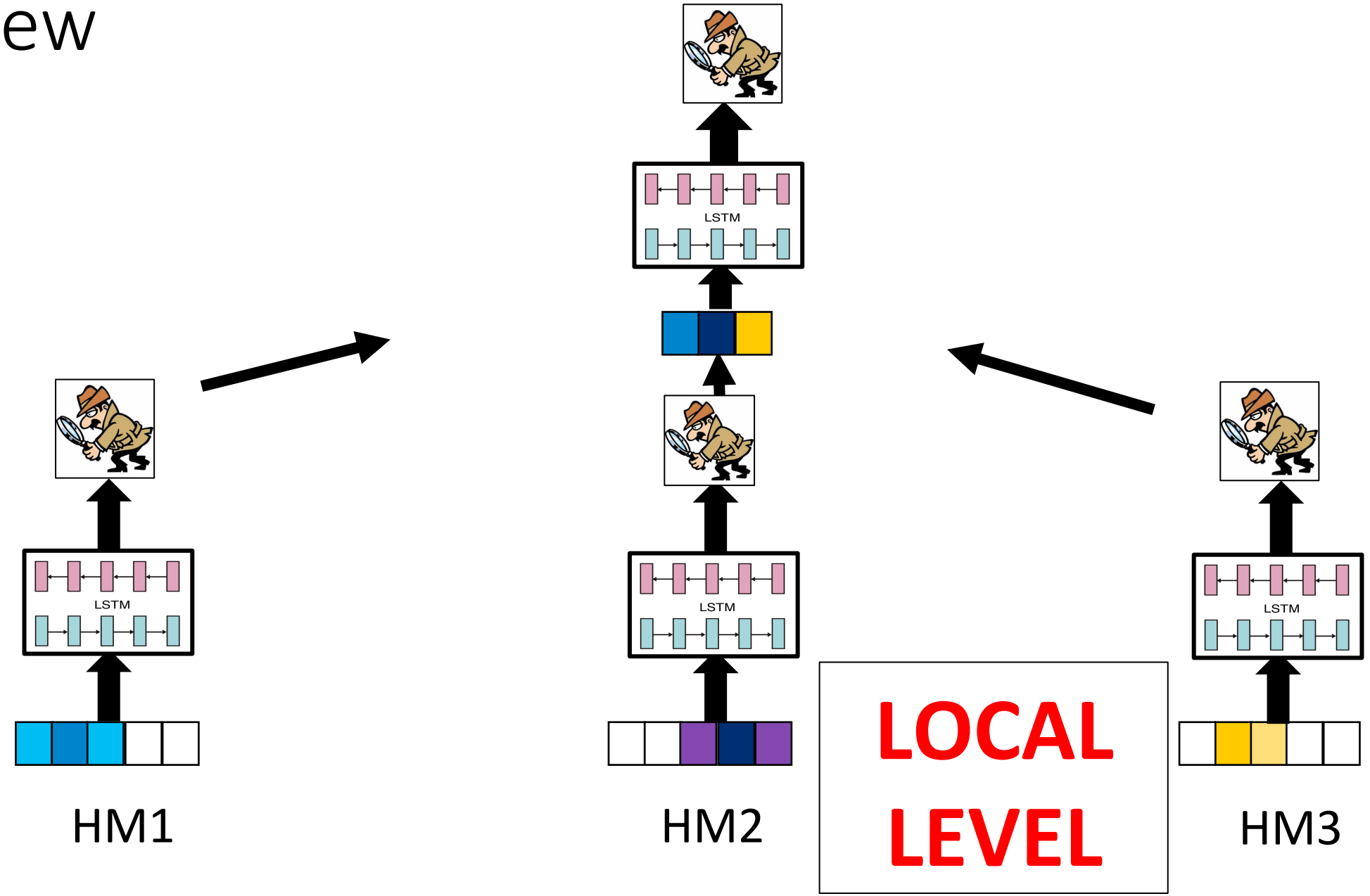
HM1

HM2

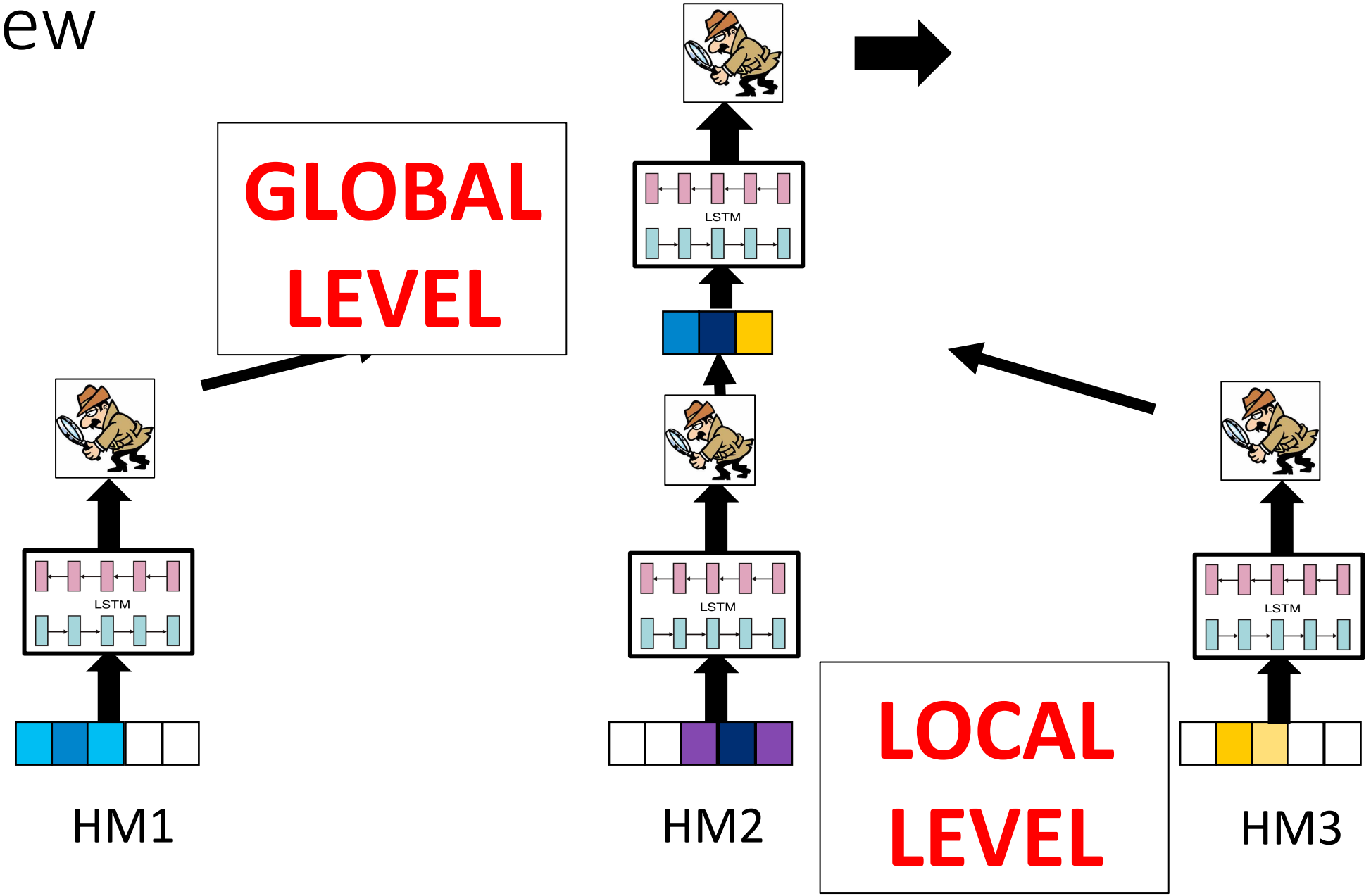
LOCAL LEVEL

HM5

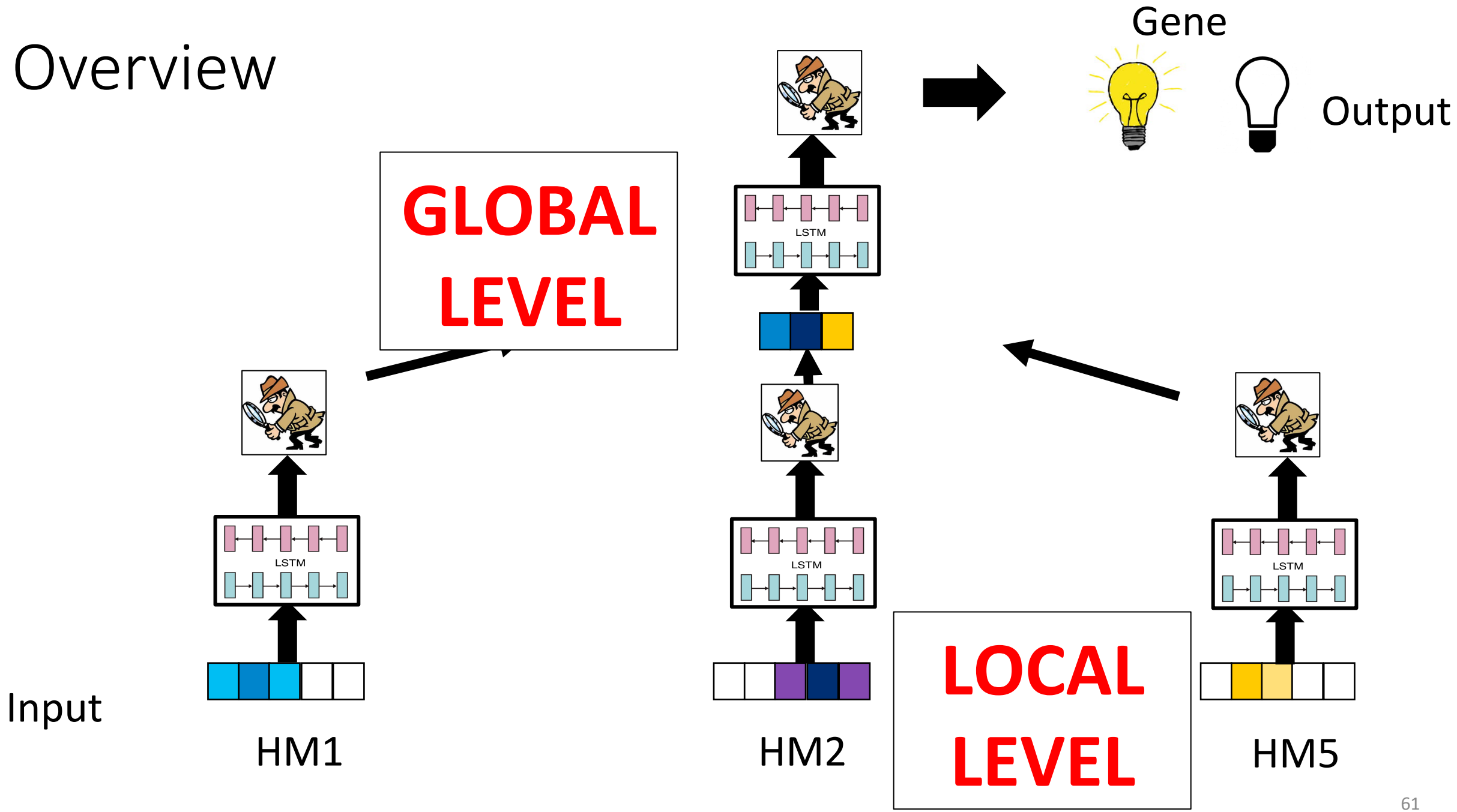
Overview



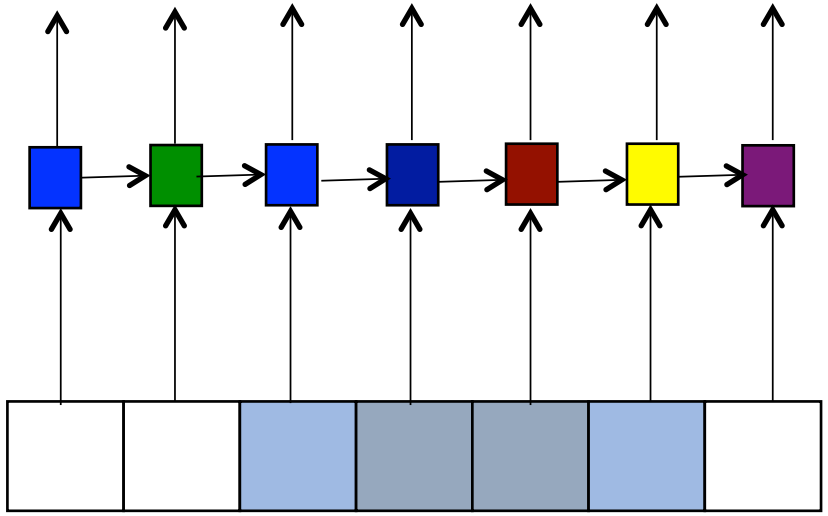
Overview

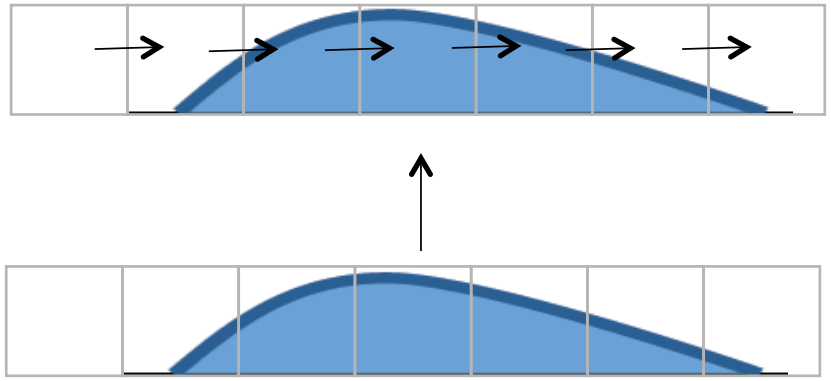

GLOBAL LEVEL
LSTM
LSTM
LSTM
LSTM
LOCAL LEVEL
Input
HM1
HM2
HM3

Overview


Gene
Output
GLOBAL LEVEL
LSTM
LSTM
LSTM
LSTM
LOCAL LEVEL
Input
HM1
HM2
HM5

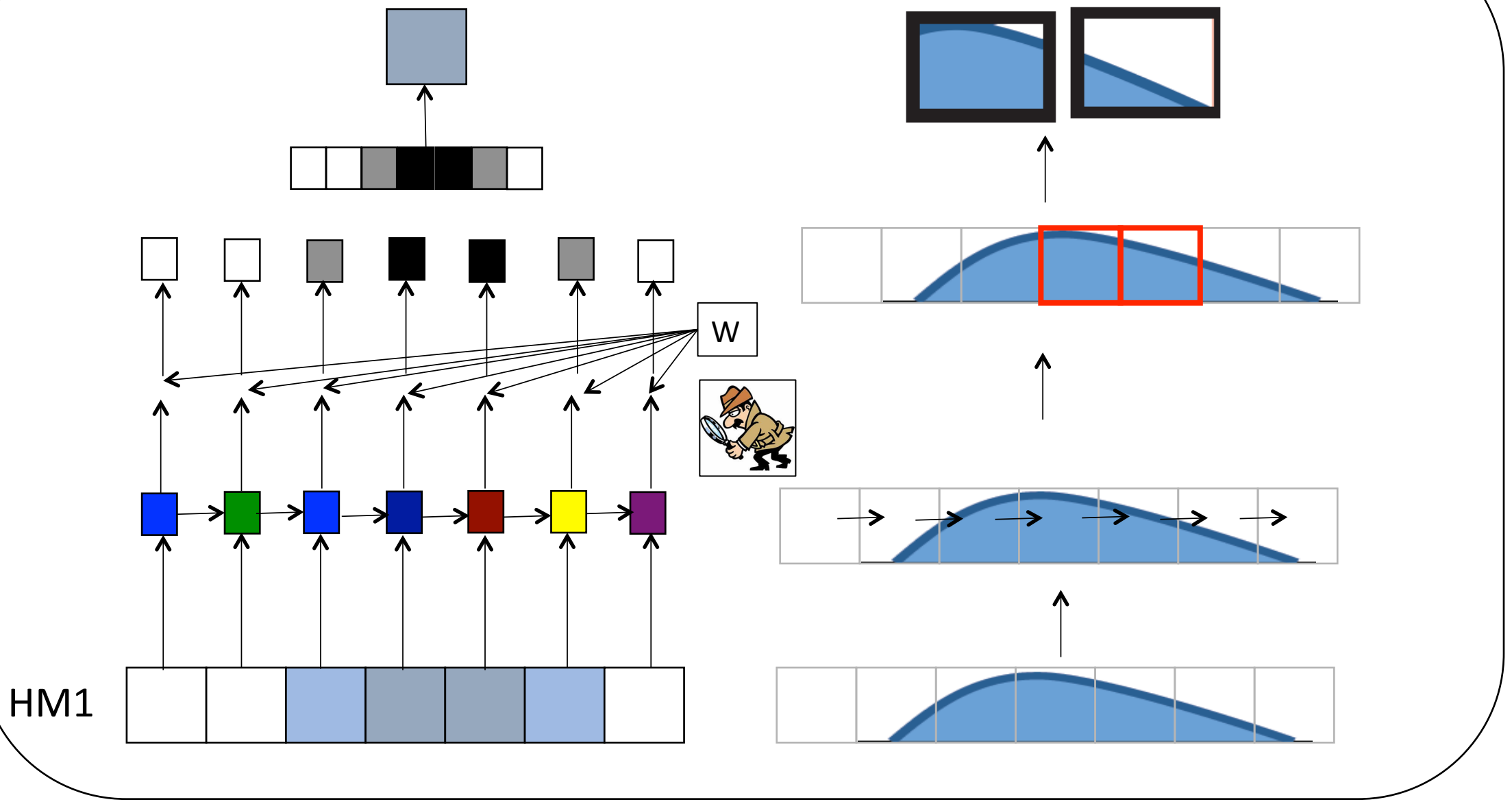
Recurrent Neural Network (RNN)





HM1

Attention Mechanism



Versus Baselines

Computational Study	Unified	Non-linear	Bin-Info	Representation Learning		Prediction	Feature Inter.	Interpretable
				Neighbor Bins	Whole Region			
Linear Regression ([14])	×	×	×	×	✓	✓	×	✓
Support Vector Machine ([7])	×	✓	Bin-specific	×	✓	✓	✓	×
Random Forest ([10])	×	✓	Best-bin	×	✓	✓	×	×
Rule Learning ([12])	×	✓	×	×	✓	×	✓	✓
DeepChrome-CNN [29]	✓	✓	Automatic	✓	✓	✓	✓	×
AttentiveChrome	✓	✓	Automatic	✓	✓	✓	✓	✓

Experiments: Prediction Performance

- Same setup as DeepChrome
- AttentiveChrome is as accurate as (slightly better than) DeepChrome

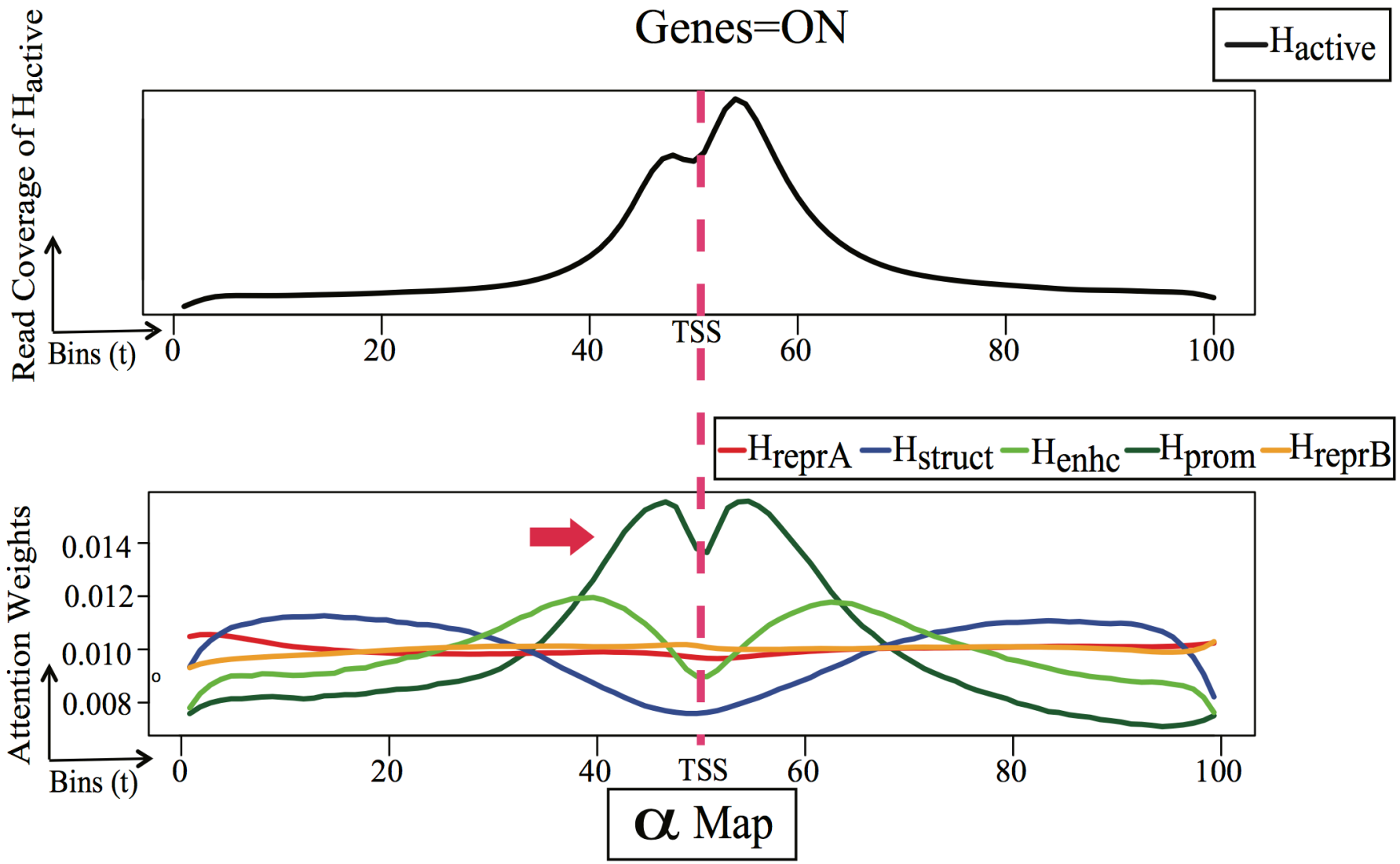
	Baselines		Our Model
Models	**DeepChrome (CNN) [3]**	**LSTM**	**AttentiveChrome**
Mean	0.8008	0.8052	**0.8115**
Median	0.8009	0.8036	**0.8123**
Max	**0.9225**	0.9185	0.9177
Min	0.6854	0.7073	**0.7215**
Improvement over DeepChrone [3] (out of 56 cell types)		36	**49**

Experiments: Interpretability

- Local-level (HM-level) Attention
- Global-level (HM interactions) Attention

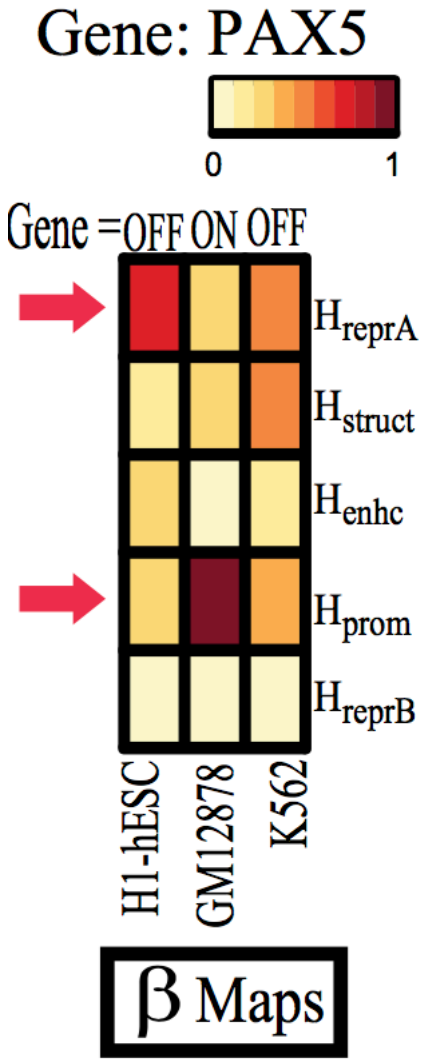
Histone Modification (HM) Mark	Renamed as	Functional Category
H3K4me3	H_{prom}	Promoter mark
H3K4me1	H_{enhc}	Distal Enhancer mark
H3K36me3	H_{struct}	Structural mark
H3K9me3	H_{reprA}	Repressor mark
H3K27me3	H_{reprB}	Repressor mark

(1) Visualization of Local Attention Weights (Learned from Data)



- 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

(2) Visualization of Global Attention Weights (Learned from Data)



- 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.

(3) Comparison with State-of-Art Deep-Visualization Methods

Correlation between local-level (HM-level) attention weights and the additional signal - H3K27ac (H-Active) from REMC

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

Summary

code available at: **deepchrome.org**

- **Attentive DeepChrome**
  - Both accurate and interpretable
  - Novel implementation of deep attention mechanism
  - Importance analysis at both HM and HM-HM level

References

- Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. “DeepChrome: Deep-learning for predicting gene expression from histone modifications”. *Bioinformatics.* (ECCB) (2016)
- Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, Yanjun Qi, "Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin", to appear at NIPS (2017)

Acknowledgements

Ritambhara Singh

Jack Lanchantin

Arshdeep Sekhon

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


Thank you