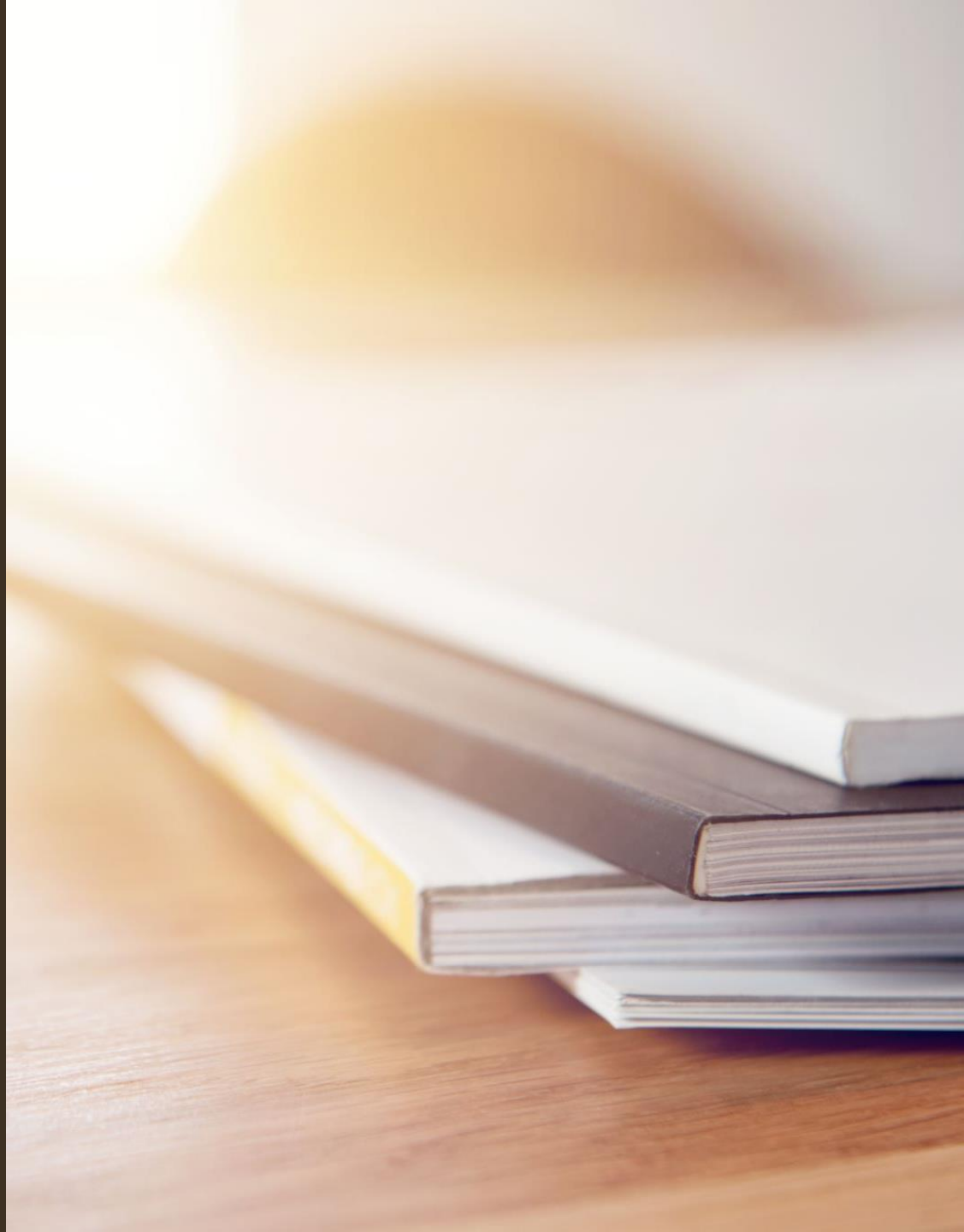
BIONLP AND ITS APPLICATIONS

Shankai Yan (NCBI/NIH)

Homepage: skyan.me

Lecture Materials:

skyan.me/lectures/online/bionlp-intro



What is BioNLP

Biomedical Natural Language Processing, Biomedical Text Mining

Knowledge extraction from text in biomedical/clinical field

Objective: Facilitate understanding and decision making in computational way

Subfield of bioinformatics

- Common: sequence data
- Difference: alphabet

Data

Data is generated by human

- Biomedical Literature

PubMed PMC

- Clinical Notes

MIMIC i2b2

- Textual Description in Knowledge Database

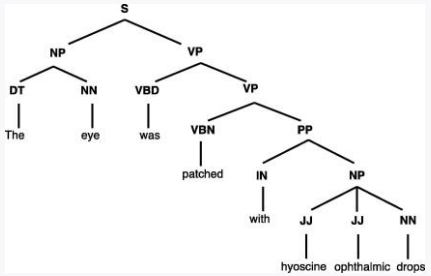
GEO Gene Expression Omnibus

human phenotype ontology

NLP Tasks

Basic Tasks:

- ❑ Tokenization (paragraph, sentence, words)
- ❑ POS Tagging (noun, verb, adj), Lemmatization, Stem
- ❑ Dependency Parsing



Advance Tasks:

- ❑ Name Entity Recognition (PubTator)
- ❑ Information Retrieval (PubMed)
- ❑ Document classification (Multi-class & Multi-label)
- ❑ Information Extraction
  - Protein-Protein-Interaction
  - Gene-Gene-Interaction
  - Drug-Drug-Interaction
  - Biological Event
- ❑ Question Answering

NER

- Online Tools
- API Calling
- Write your own code

Export Annotations

Export our annotated publications in batches of up to 100 in GET or 1000 in POST, in BioC, pubtator or JSON formats. To programmatically retrieve text-mined results in PubTator, one can use web queries as follows:

https://www.ncbi.nlm.nih.gov/research/pubtator-api/publications/export/**[Format]**?**[Type]**=**[Identifiers]**&concepts=**[Bioconcepts]**

Parameters

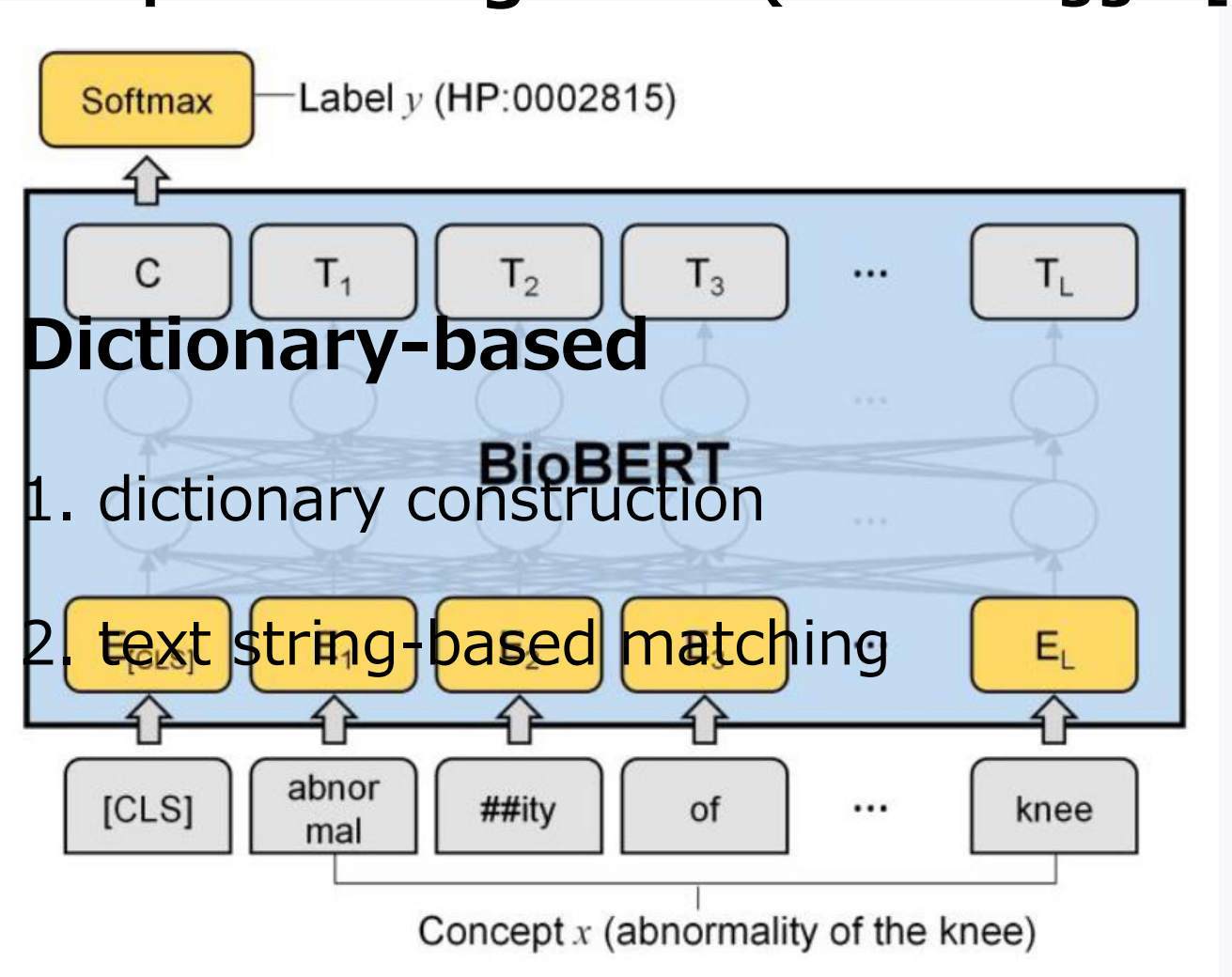
Parameter	Description
Format	**pubtator** (pubtator) **biocxml** (BioC-XML) **biocjson** (BioC-JSON)

NER Methods

Dictionary-based

ML-based

Deep Learning Model (PhenoTagger[1])



Dictionary-based

1. dictionary construction
2. text string-based matching

[1] Luo, L., **Yan, S**., Lai, P. T., Veltri, D., Oler, A., Xirasagar, S., ... & Lu, Z. (2021). PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, *37*(13), 1884-1890.

Example of Document Classification

Cancer Hallmark Annotation[2]

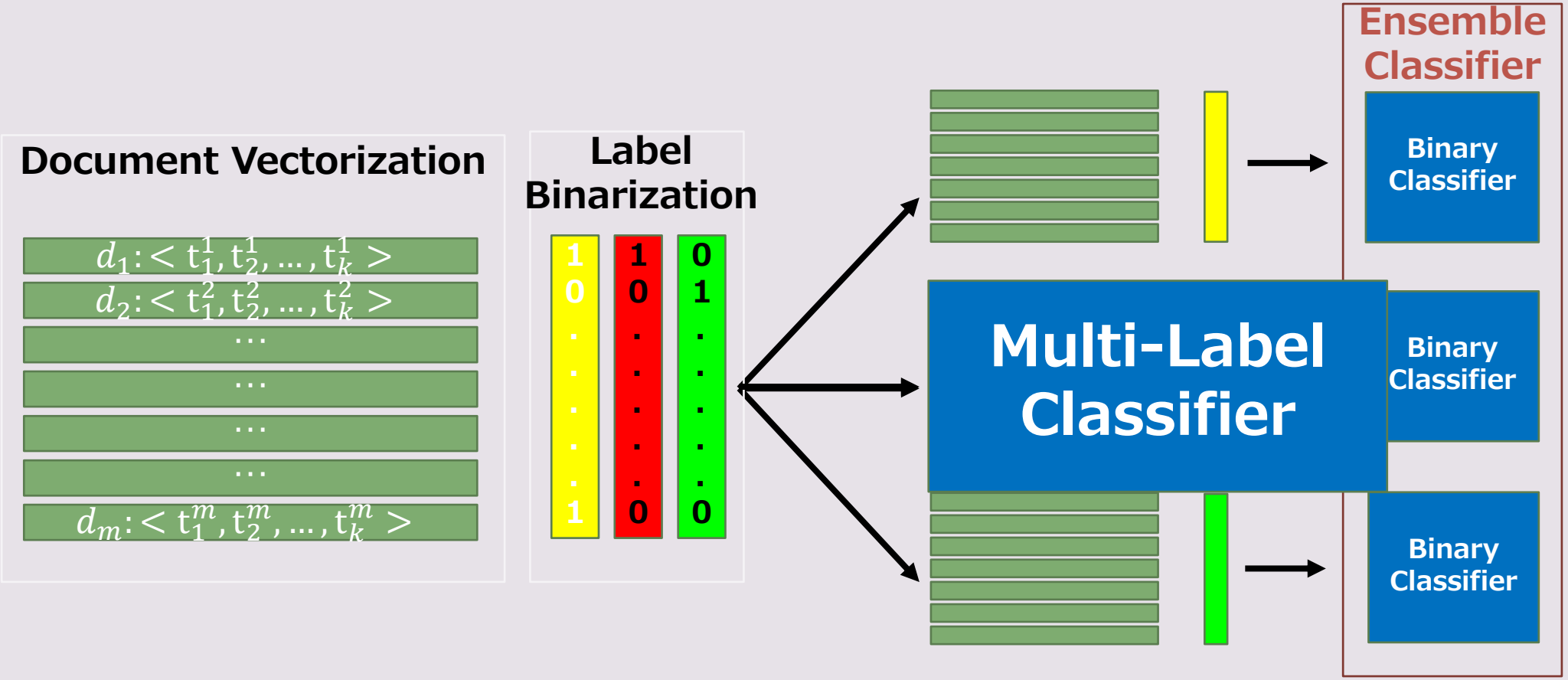
… Genes that were overexpressed in OM3 included oncogenes, cell cycle regulators , and those involved in signal transduction, whereas genes for DNA repair enzymes and inhibitors of transformation and metastasis were suppressed. In arsenic-treated cells, multiple DNA repair proteins were overexpressed. … [3]

- Activating invasion and metastasis (IM)
- Genomic instability and mutation (GI)
- Sustaining proliferative signaling (PS)
- Evading growth suppressors (GS)

[2] **Yan, Shankai**, and Ka-Chun Wong. "Elucidating high-dimensional cancer hallmark annotation via enriched ontology." *Journal of biomedical informatics* 73 (2017): 84-94.

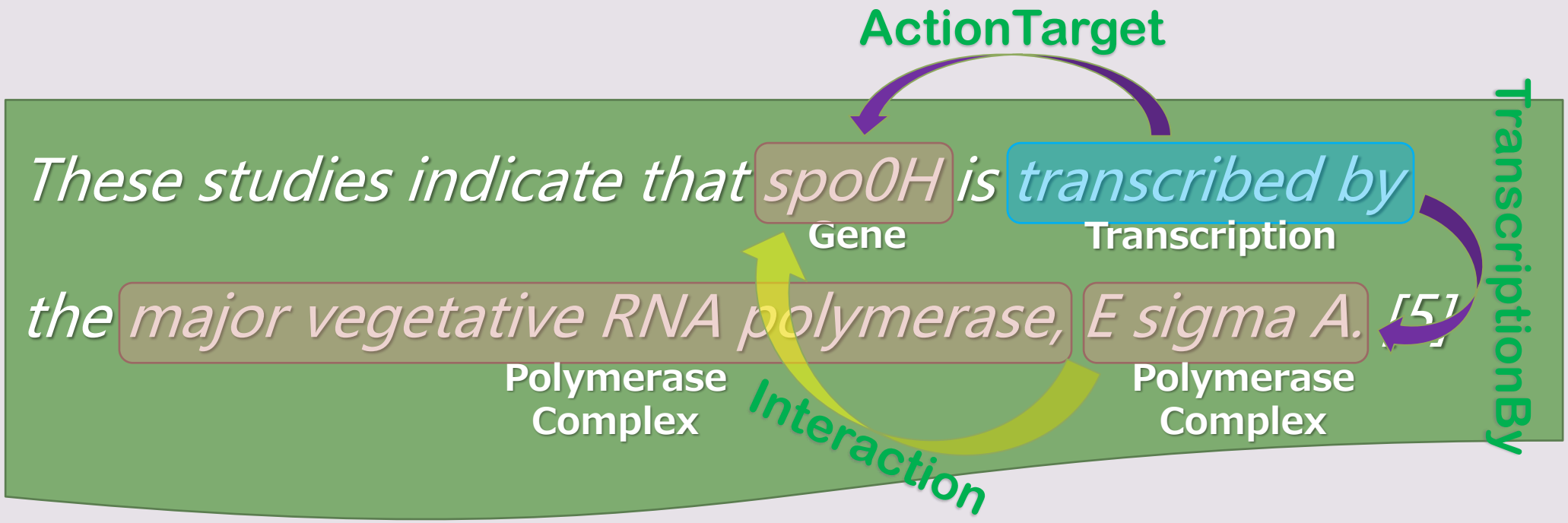
[3] Bae, Dong-Soon, et al. "Characterization of gene expression changes associated with MNNG, arsenic, or metal mixture treatment in human keratinocytes: application of cDNA microarray technology." *Environmental Health Perspectives* 110.suppl 6 (2002): 931-941.

Multi-label Classification



Example of IE

Bacteria Gene Interaction Extraction[4]



[4] Yan, Shankai, and Ka-Chun Wong. "Context awareness and embedding for biomedical event extraction." Bioinformatics 36.2 (2020): 637-643.
[5] Weir, J; Predich, M; Dubnau, E; Nair, G; Smith, I. (1991). Regulation of spo0H, a gene coding for the Bacillus subtilis sigma H factor. J. Bacteriol. vol. 173 (2) p. 521-529

TRIGGER-BASED APPROACH

- Extract lexical features from text
- Construct word matrix, edge matrix and label matrices.

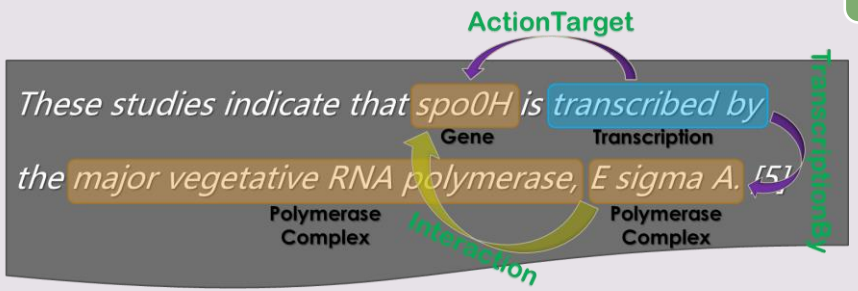
Preprocessing

Classifiers Training

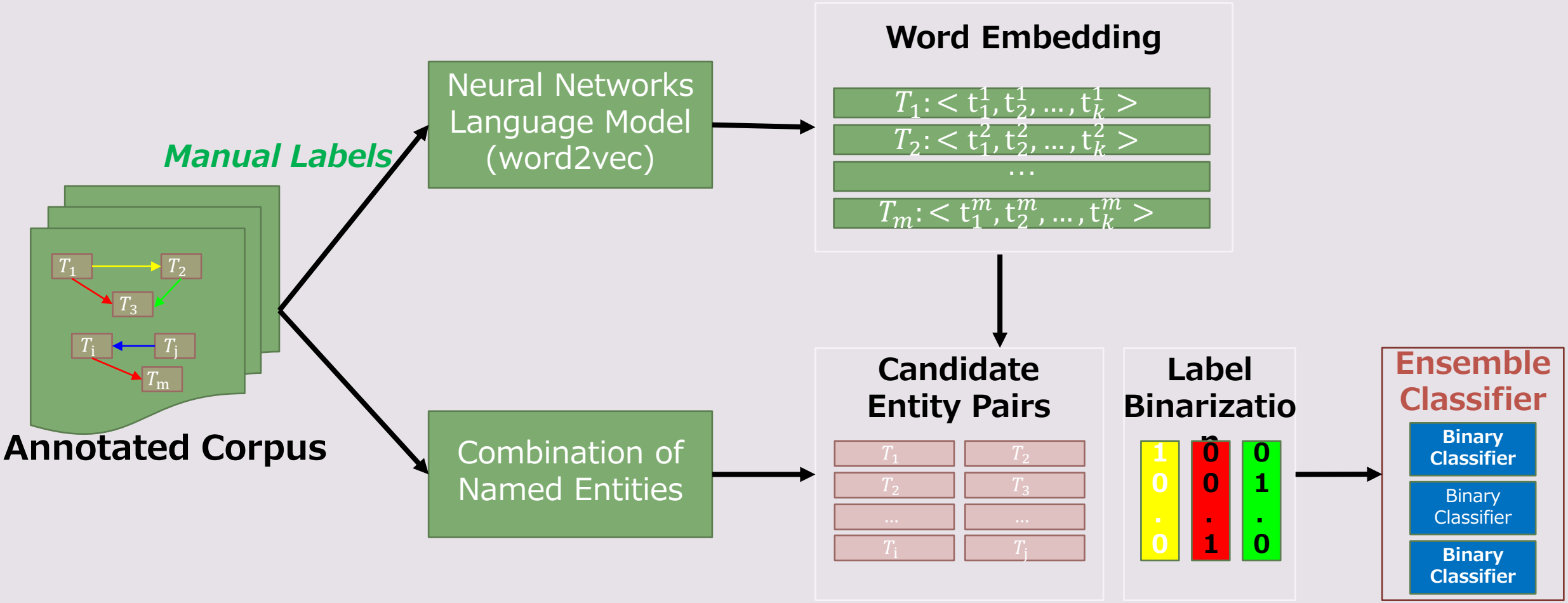
- Multiclass
- **Trigger Classification**
- **Edge Classification**

- Annotate the triggers
- Induce relation records

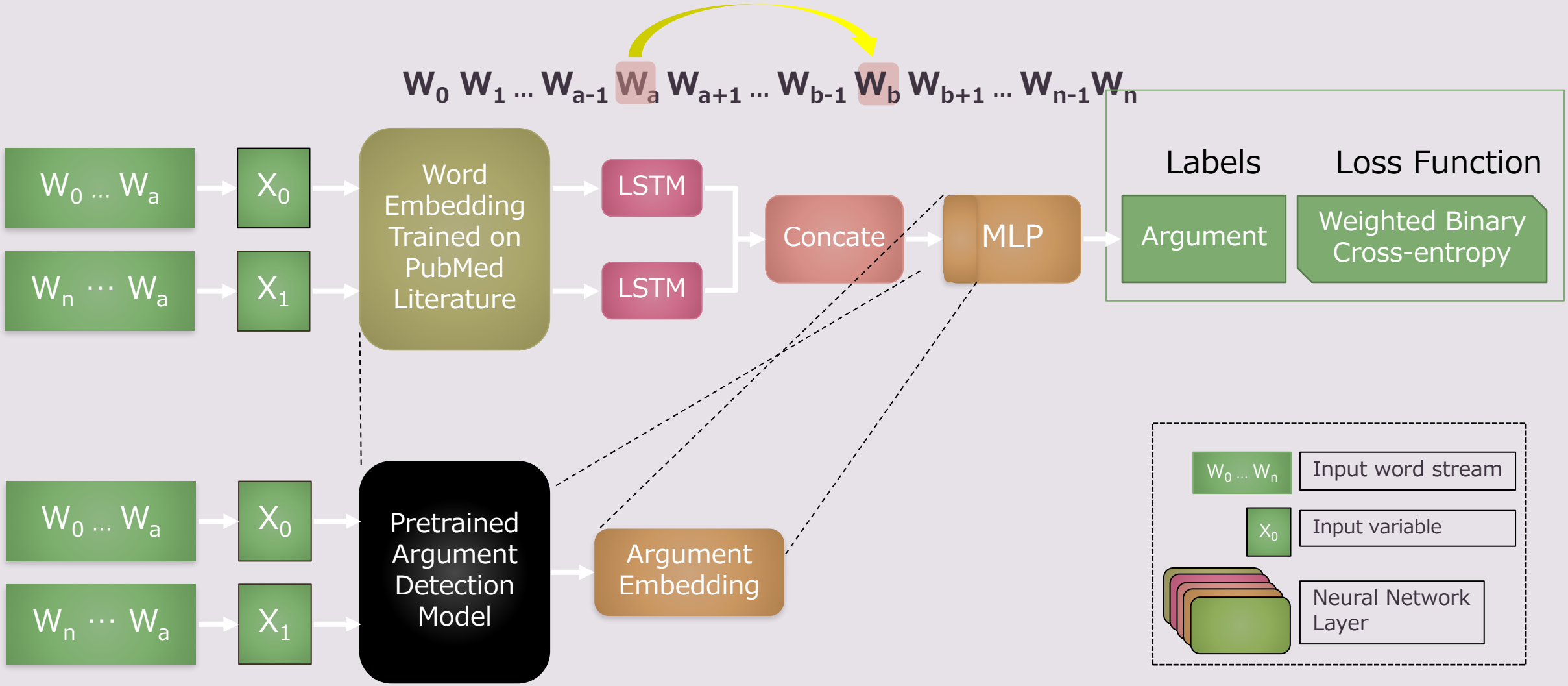
Post-processing



Non-Trigger Approach



Deep Learning Approach [4]



[4] **Yan, Shankai**, and Ka-Chun Wong. "Context awareness and embedding for biomedical event extraction." Bioinformatics 36.2 (2020): 637-643.

Difference from Conventional NLP Tasks

- Contains Biomedical Symbols, Punctuations, Terms
- The writers are from the biomedical/clinical expertise group
- Different Corpus/Context

Leaderboard

- GLUE[6]
- BLUE[7]

Task	Metrics	SOTA*	ELMo	BioBERT	Our BERT			
					Base (P)	Base (P+M)	Large (P)	Large (P+M)
MedSTS	Pearson	83.6	68.6	84.5	84.5	**84.8**	84.6	83.2
BIOSSES	Pearson	84.8	60.2	82.7	89.3	**91.6**	86.3	75.1
BC5CDR-disease	F	84.1	83.9	85.9	**86.6**	85.4	82.9	83.8
BC5CDR-chemical	F	93.3	91.5	93.0	**93.5**	92.4	91.7	91.1
ShARe/CLEFE	F	70.0	75.6	72.8	75.4	**77.1**	72.7	74.4
DDI	F	72.9	78.9	78.8	78.1	79.4	**79.9**	76.3
ChemProt	F	64.1	66.6	71.3	72.5	69.2	**74.4**	65.1
i2b2	F	73.7	71.2	72.2	74.4	**76.4**	73.3	73.9
HoC	F	81.5	80.0	82.9	85.3	83.1	**87.3**	85.3
MedNLI	acc	73.5	71.4	80.5	82.2	**84.0**	81.5	83.8
Total			78.8	80.5	82.2	**82.3**	81.5	79.2

[6] Wang, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." arXiv preprint arXiv:1804.07461 (2018).
[7] Peng, Yifan, **Shankai Yan**, and Zhiyong Lu. "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets." arXiv preprint arXiv:1906.05474 (2019).

Methods

Feature Extraction:

Tokenization, Bag-of-Words, Normalization

Tokenization, Word Embeddings, Language Model

Classification/Regression/ Model:

SVM/RandomForest

Language Model/Encoder:

LSTM/CNN/BERT

Traditional Machine Learning Methods

Handcrafted Features

Tokens

Part-of-speech

Entity type

Grammatical function tag

Distance in the parse tree

Classifical ML models

Support Vector Machine (SVM)

Bayesian Classifier

Multi-layer Perception (MLP)

Ensemble Classifiers (Random Forest, Extra Trees, etc.)

SOTA Deep Learning Methods

Word Embedding / Language Model (Pre-trained)

Sequence to Vector

Encoder:

- *Bag of Embedding*
- *RNN/LSTM/GRU*
- *CNN*

Classifier:

- *Linear*
- *MLP*

Packages

NLTK: simple text processing

SpaCy: Pipeline

Stanza: Pipeline

Gensim: Word/Doc Embedding

Hugging Face (Transformers): SOTA Model

Text Mining on Knowledge Databases

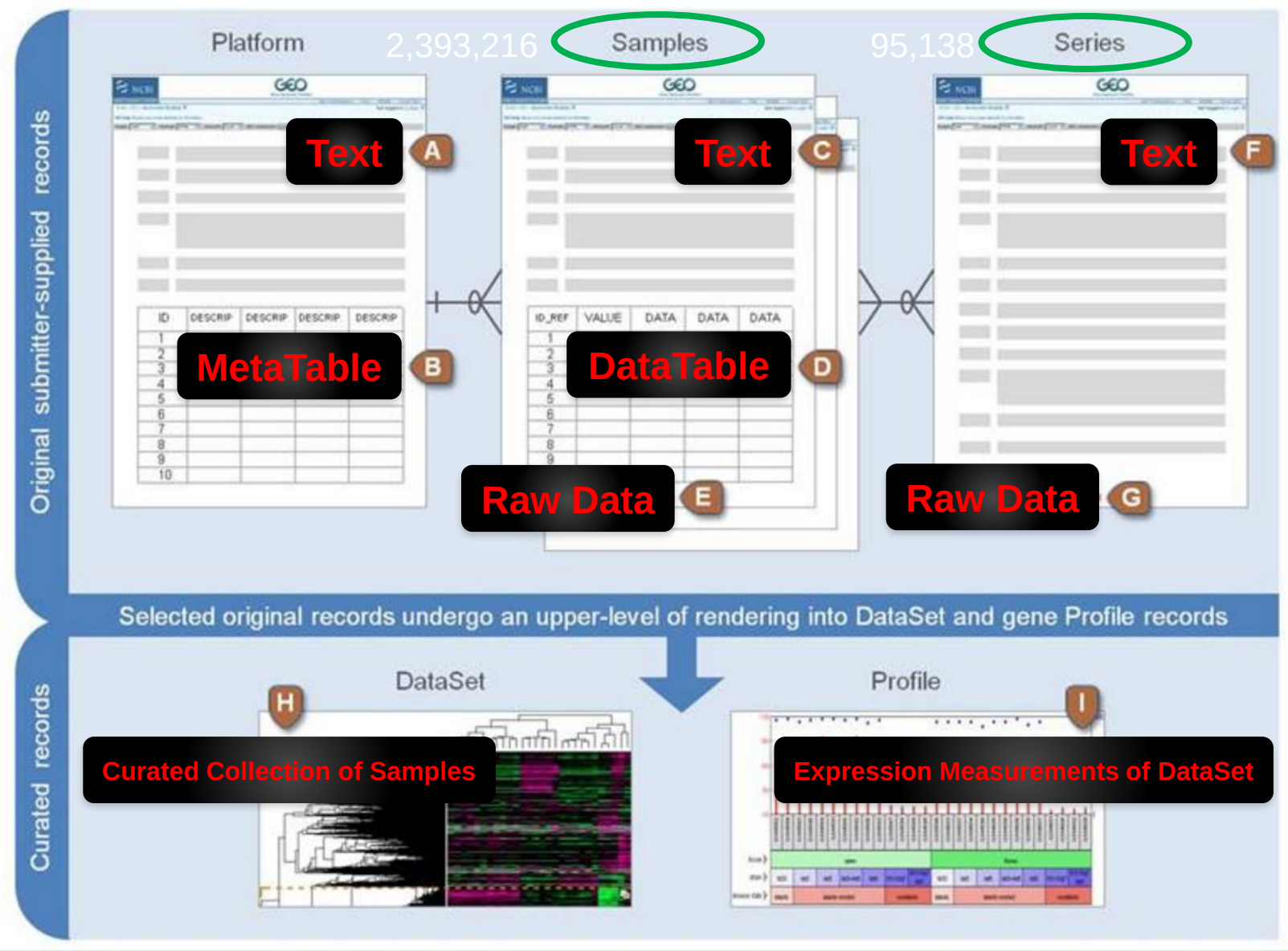
Experiment Repositories:

GEO/DECIPHER

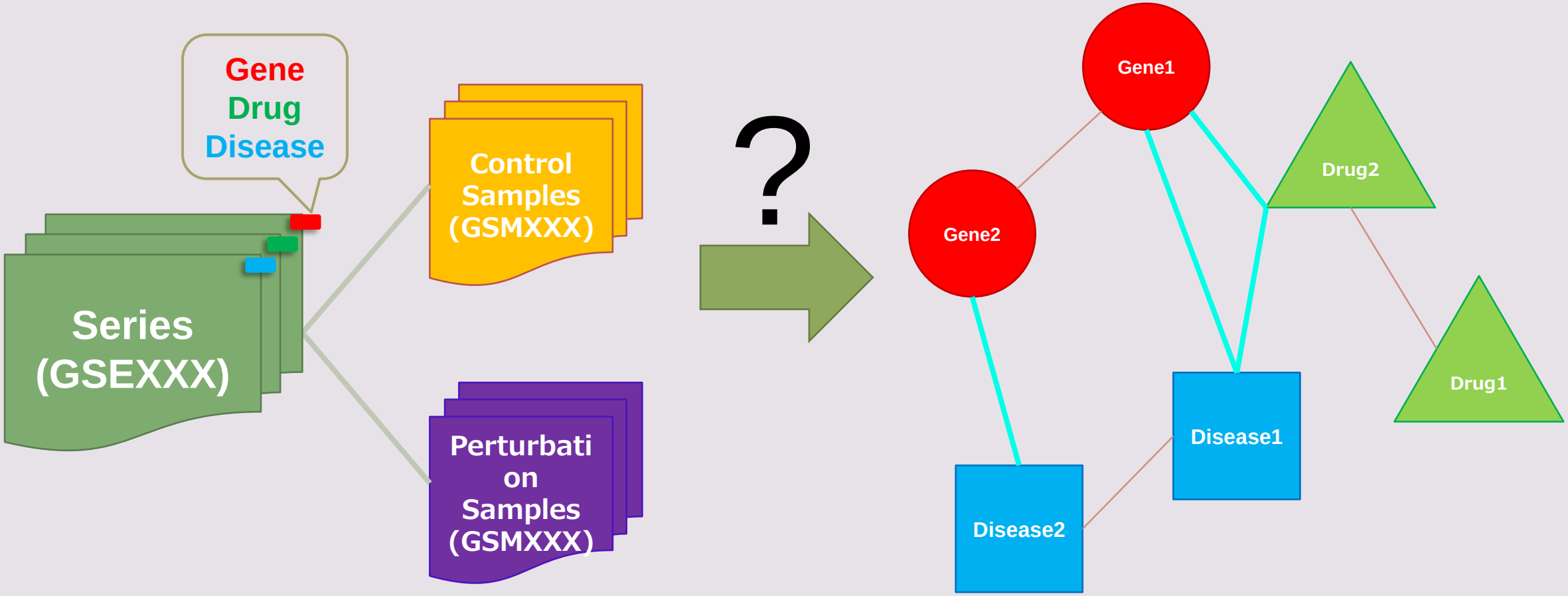
Ontology:

GO/HPO

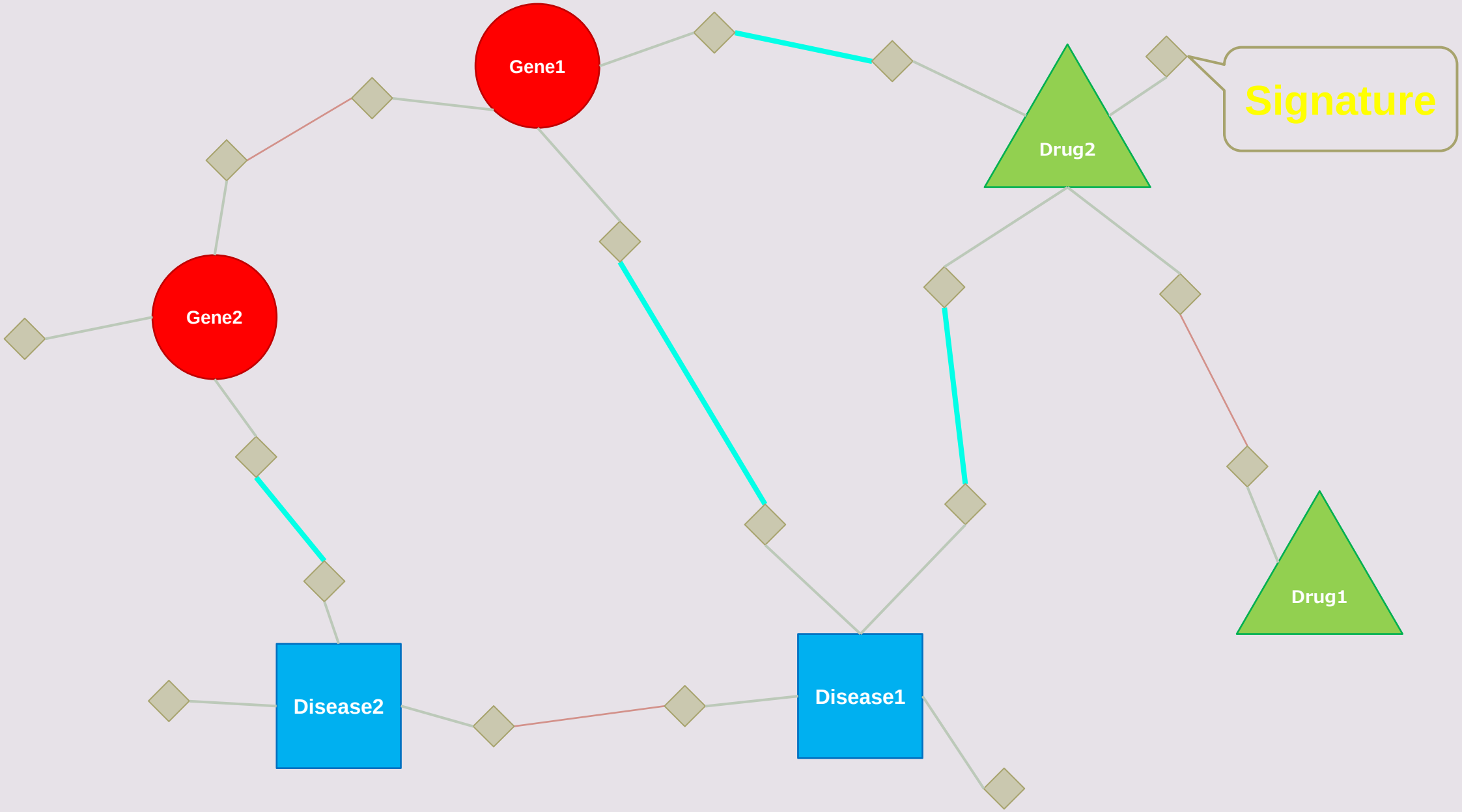
DATA DESCRIPTION IN GEO



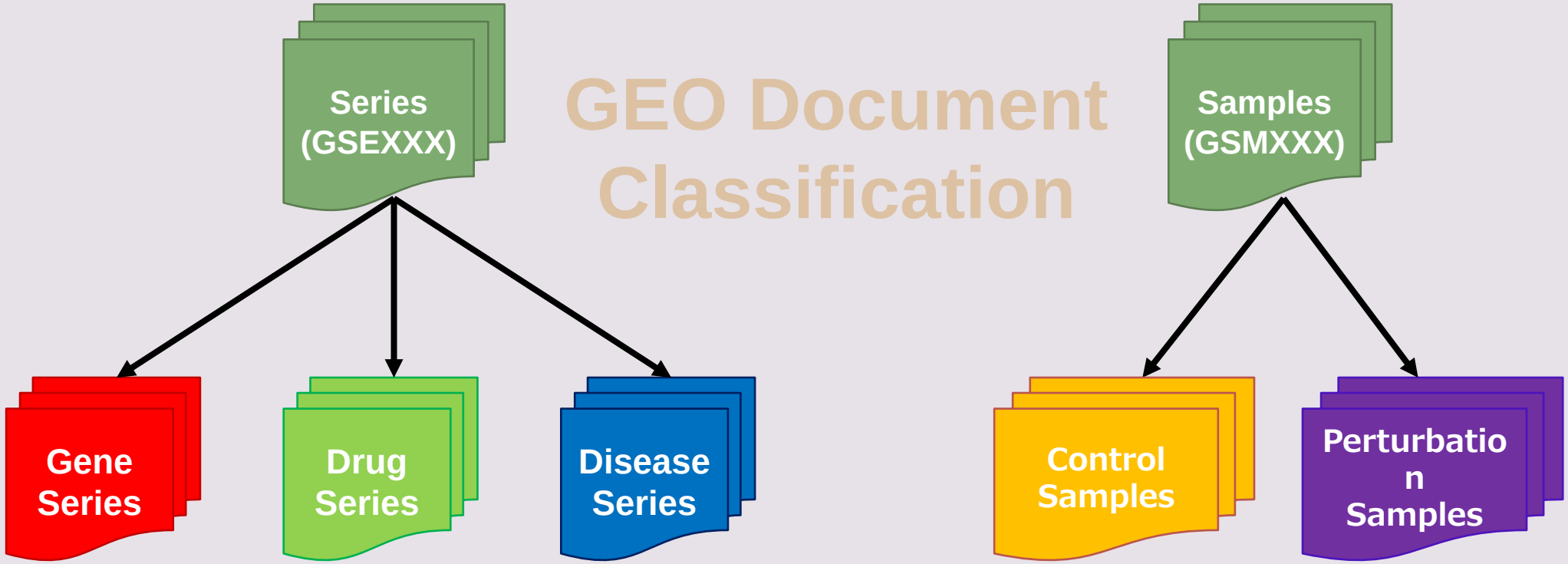
OBJECTIVE


Gene
Drug
Disease
Series
(GSEXXX)
Control
Samples
(GSMXXX)
Perturbati
on
Samples
(GSMXXX)
?
Gene1
Gene2
Drug2
Drug1
Disease1
Disease2

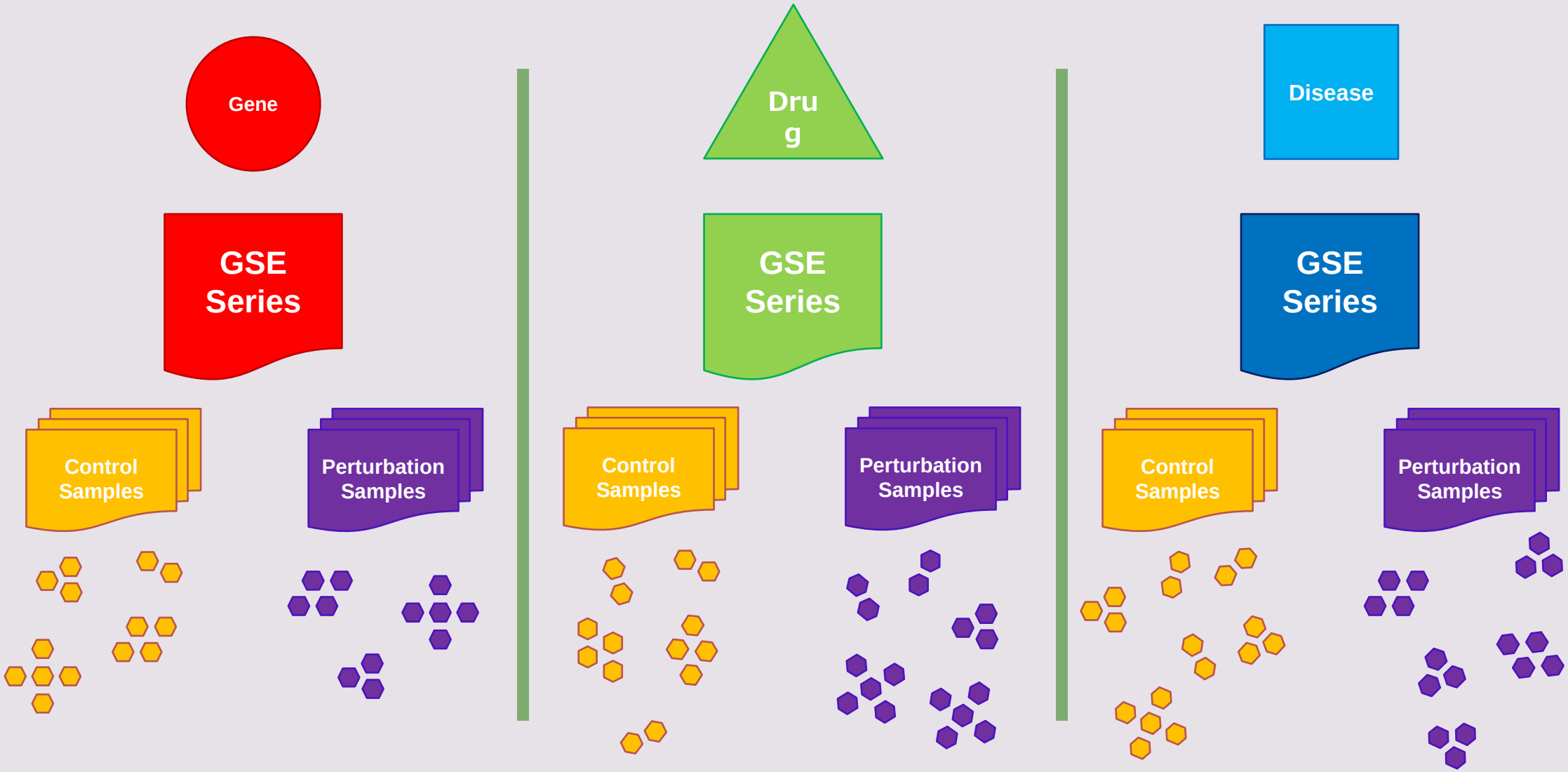
PROBLEM



METHODOLOGY

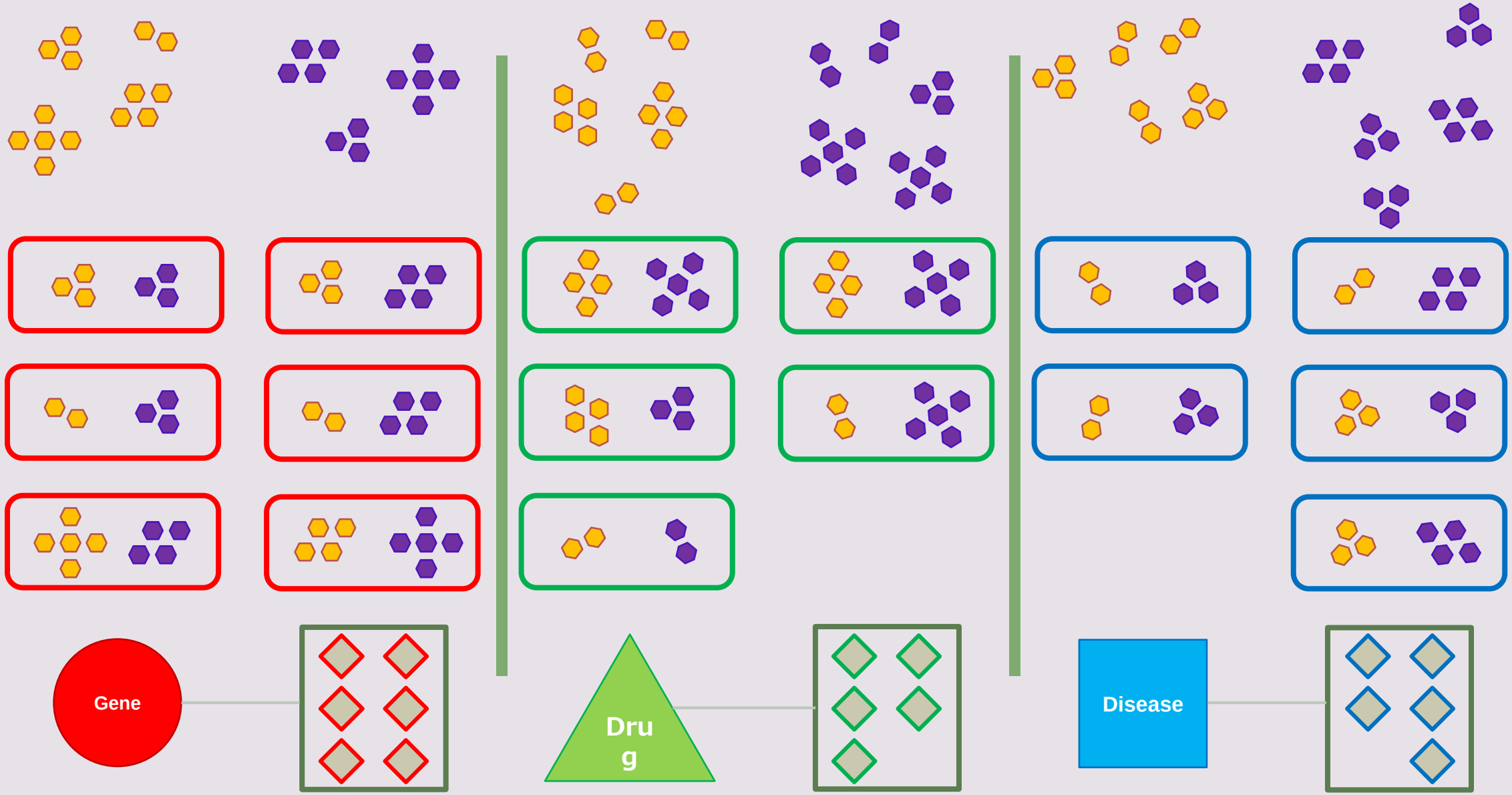


METHODOLOGY



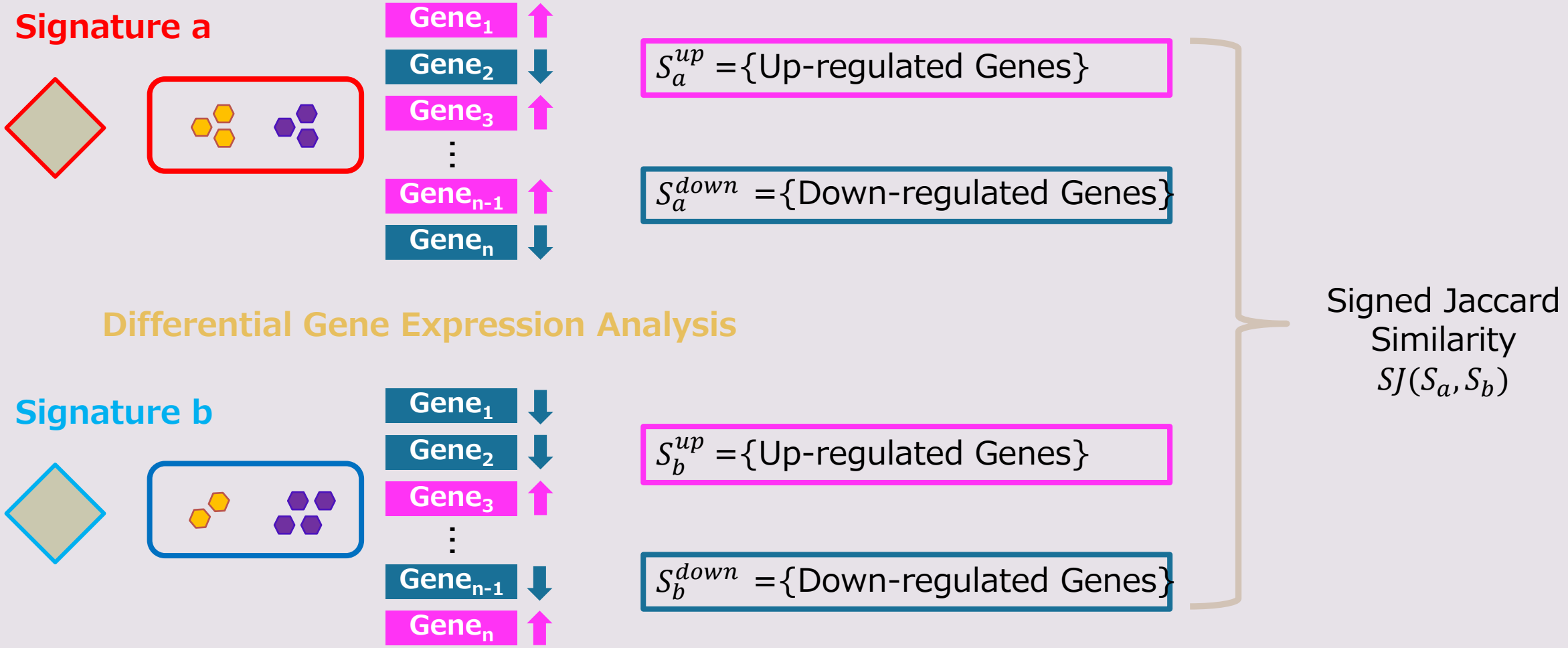
Sample Clustering

METHODOLOGY

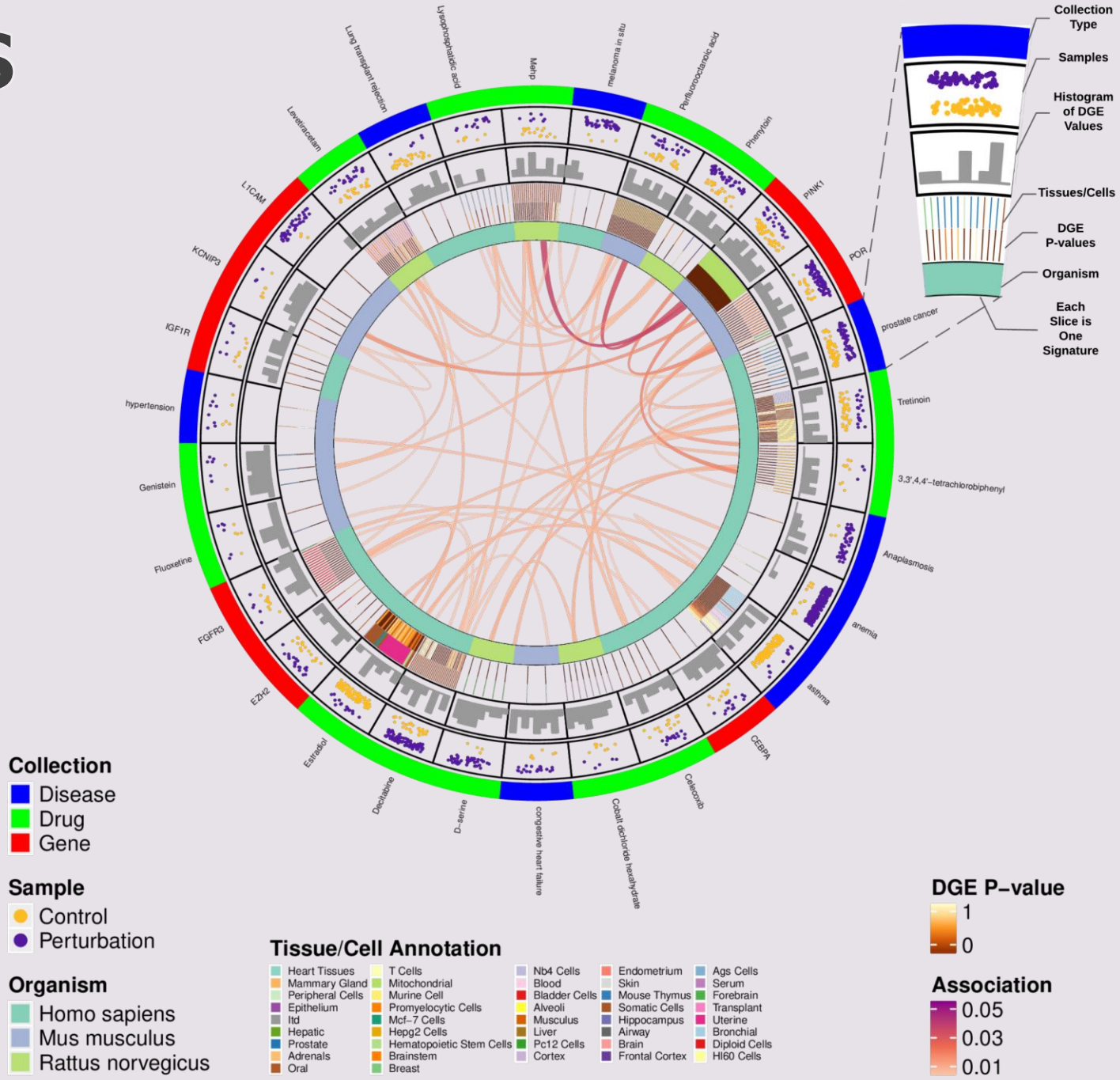


METHODOLOGY

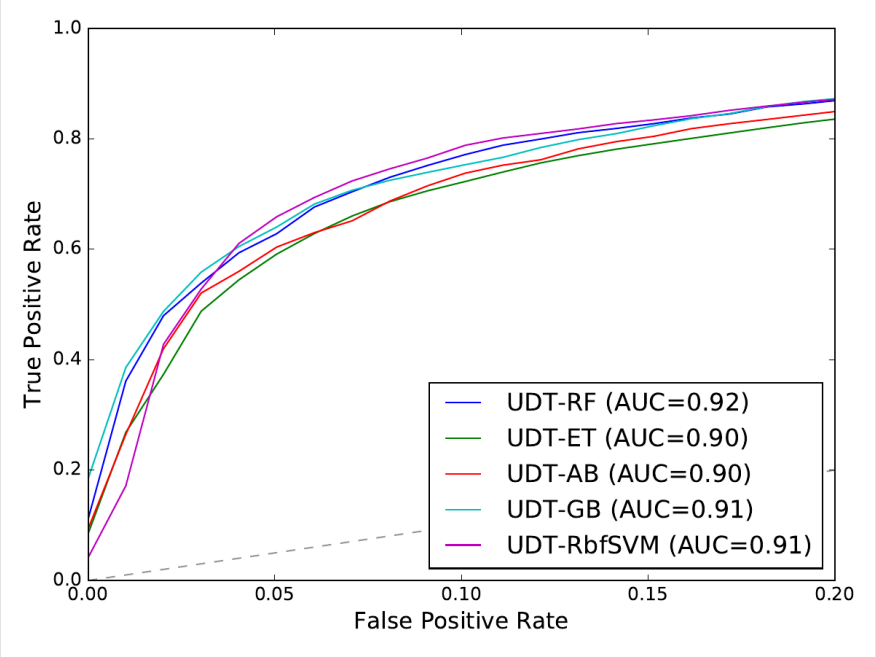
$$SJ(S_a, S_b) = \frac{J(S_a^{up}, S_b^{up}) + J(S_a^{down}, S_b^{down}) - J(S_a^{up}, S_b^{down}) - J(S_a^{down}, S_b^{up})}{2}$$



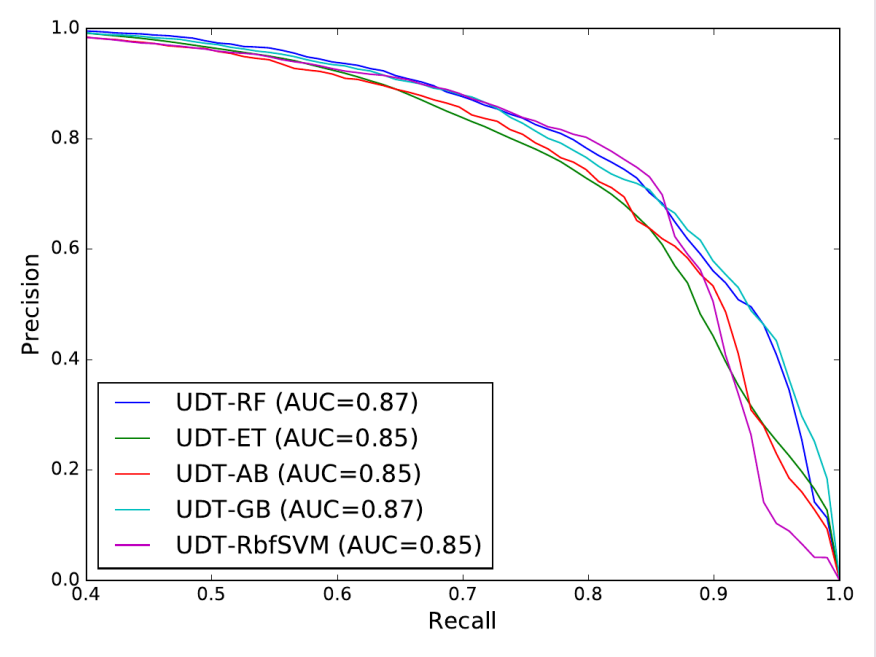
RESULTS


Collection Type
Samples
Histogram of DGE Values
Tissues/Cells
DGE P-values
Organism
Each Slice is One Signature
Lung transplant rejection
Lysophosphatidic acid
Mehp
melanoma in situ
Perfluorooctanoic acid
Phenytoin
PINK1
POR
prostate cancer
Tretinoin
3,3',4,4'-tetrachlorobiphenyl
Anaplasmosis
anemia
asthma
CEBPA
Celecoxib
Cobalt dichloride hexahydrate
congestive heart failure
D-serine
Decitabine
Estradiol
EZH2
FGFR3
Fluoxetine
Genistein
hypertension
IGF1R
KCNIP3
L1CAM
Levetiracetam
Collection
Disease
Drug
Gene
Sample
Control
Perturbation
Organism
Homo sapiens
Mus musculus
Rattus norvegicus
Tissue/Cell Annotation
Heart Tissues
Mammary Gland
Peripheral Cells
Epithelium
Itd
Hepatic
Prostate
Adrenals
Oral
T Cells
Mitochondrial
Murine Cell
Promyelocytic Cells
Mcf-7 Cells
Hepg2 Cells
Hematopoietic Stem Cells
Brainstem
Breast
Nb4 Cells
Blood
Bladder Cells
Alveoli
Musculus
Liver
Pc12 Cells
Cortex
Endometrium
Skin
Mouse Thymus
Somatic Cells
Hippocampus
Airway
Brain
Frontal Cortex
Ags Cells
Serum
Forebrain
Transplant
Uterine
Bronchial
Diploid Cells
HI60 Cells
DGE P-value
1
0
Association
0.05
0.03
0.01

GSE CLASSIFICATION RESULT



ROC Curve



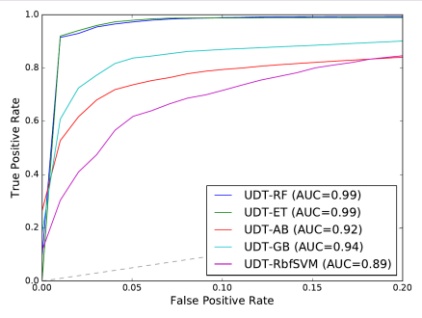
PRC Curve

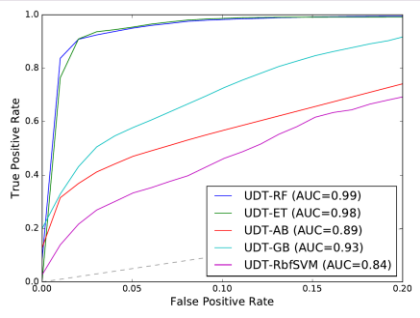
	UDT-RF[2]	UDT-ET	UDT-AB	UDT-GB	UDT-RbfSVM
Accuracy	0.68	0.63	0.65	0.68	0.56
Precision	0.81	0.83	0.79	0.82	0.62
Recall	0.77	0.69	0.75	0.75	0.58
F score	0.79	0.75	0.77	0.78	0.60

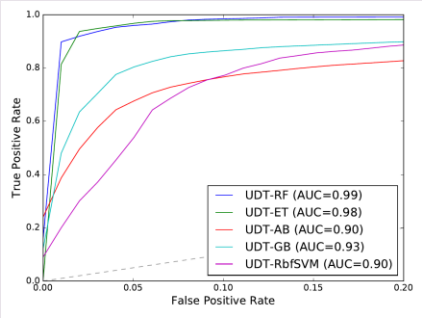
[2] Yan, S. and Wong, K.-C. (2017). Elucidating high-dimensional cancer hallmark annotation via enriched ontology. Journal of Biomedical Informatics, 73, 84–94

GSM CLASSIFICATION RESULT

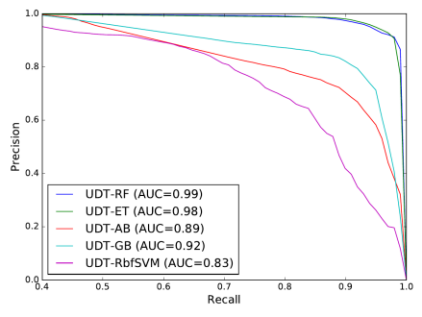
ROC Curve

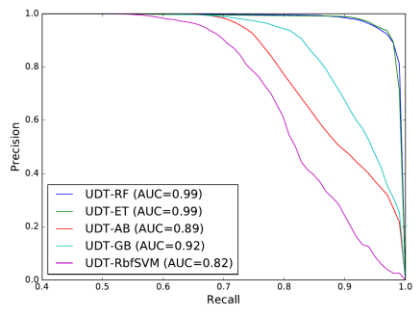


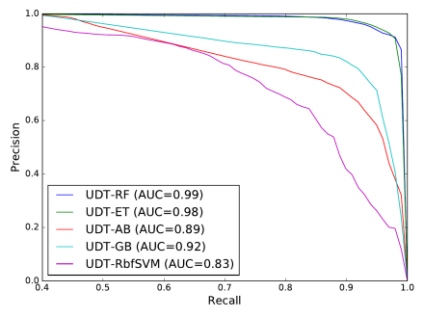




PRC Curve





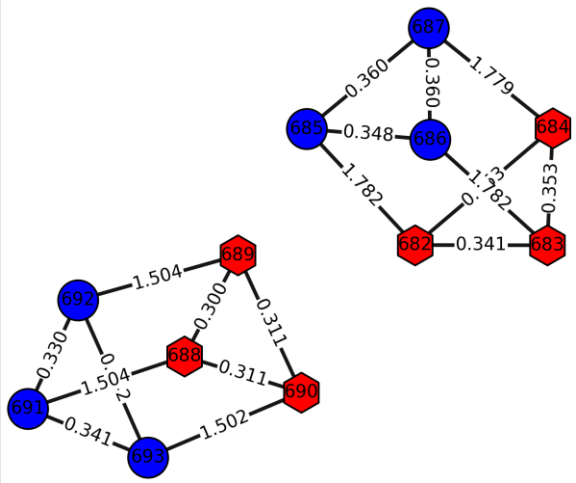


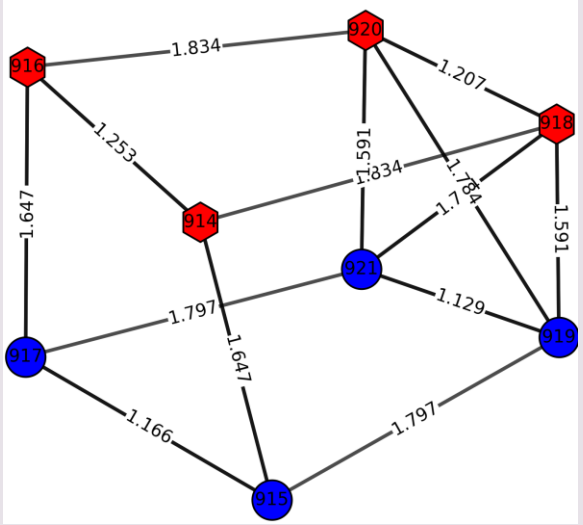
	UDT-RF[1]	UDT-ET	UDT-AB	UDT-GB	UDT-RbfSVM
Accuracy	0.96	0.96	0.85	0.87	0.66
Precision	0.95	0.96	0.85	0.89	0.51
Recall	0.94	0.94	0.75	0.76	0.37
F score	0.94	0.95	0.79	0.80	0.30

[2] Yan, S. and Wong, K.-C. (2017). Elucidating high-dimensional cancer hallmark annotation via enriched ontology. Journal of Biomedical Informatics, 73, 84–94

CLUSTERING RESULT

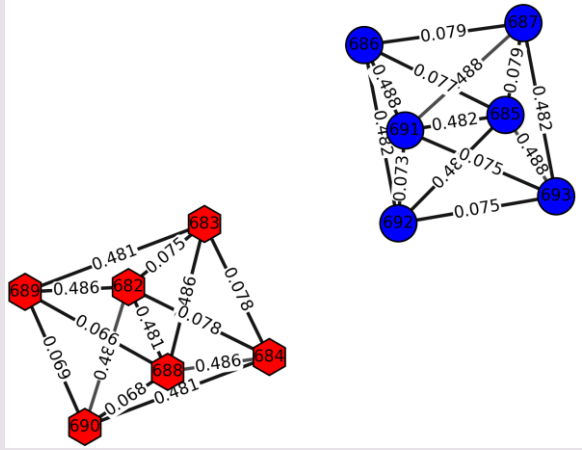
Euclidean Distance



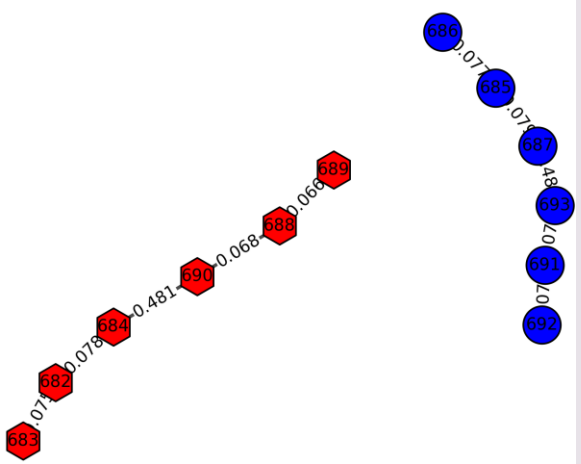


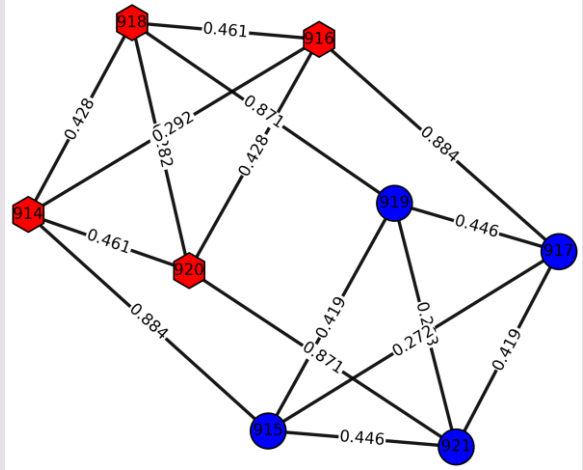
Our Distance Metric

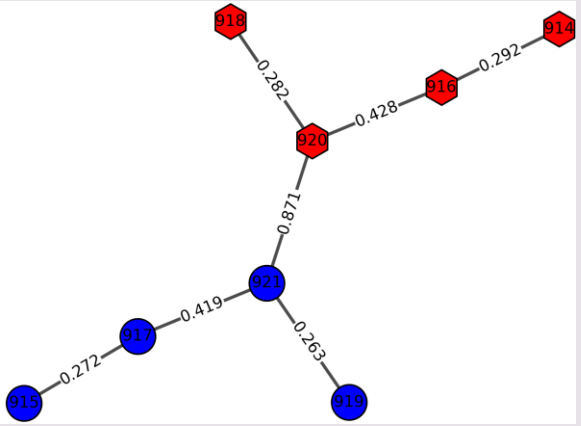
KNN Graph



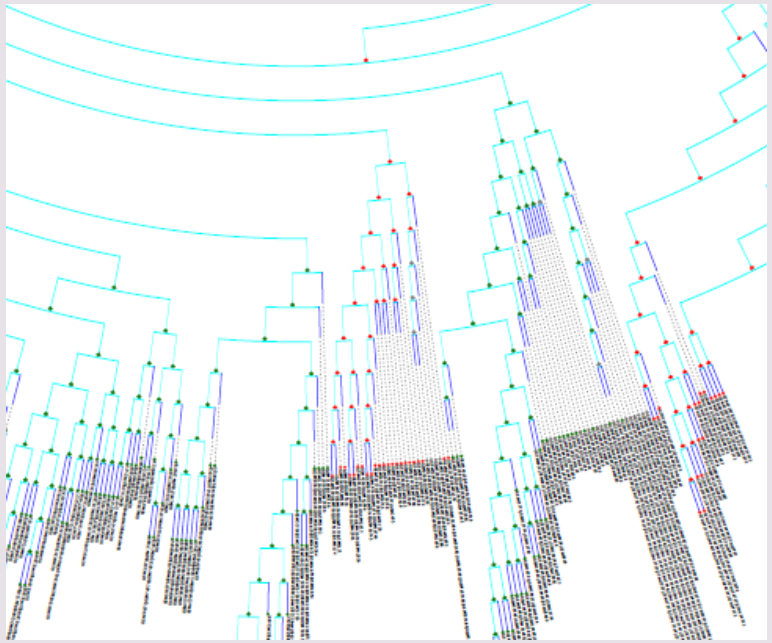
Minimum Spanning Tree

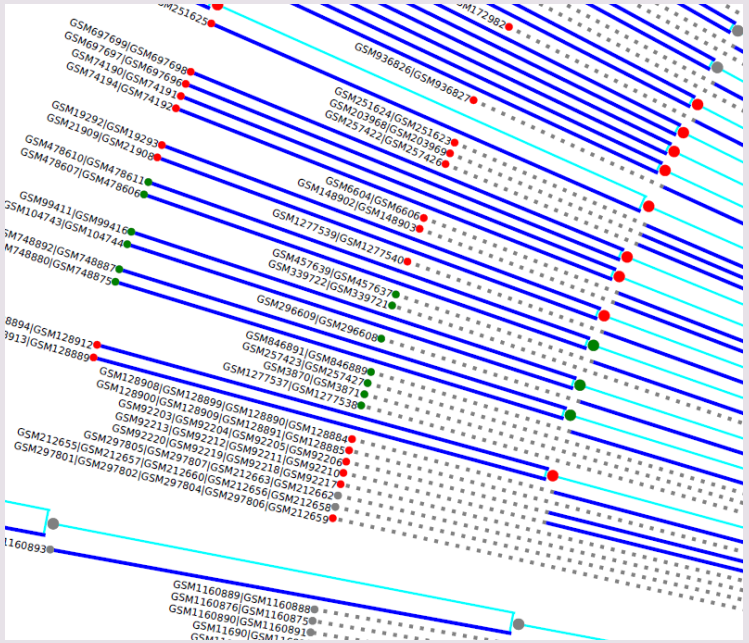


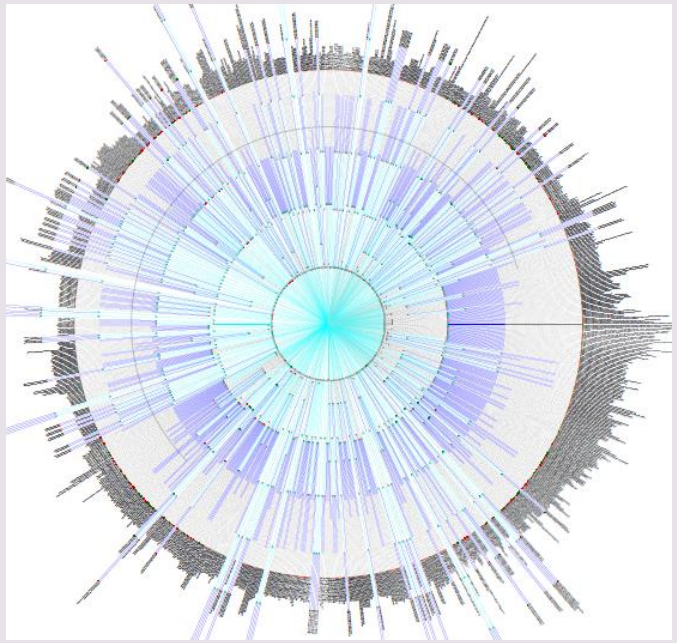




CLUSTERING RESULT







Thank You