

Weill Cornell  
Medicine

# Classifying Clinically Actionable Genetic Mutations

---

**Xi Zhang**, Dandi Chen, Yongjun Zhu, Chao Che, Chang Su, Sendong Zhao, Xu Min and Fei Wang.

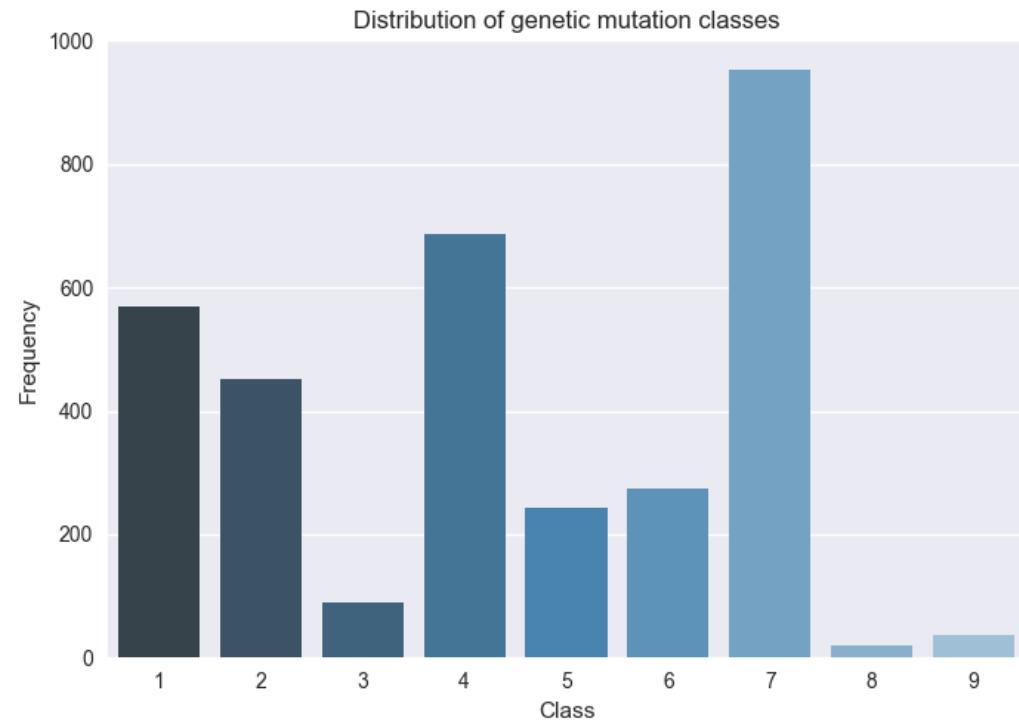
**Weill Cornell Medicine, Cornell University**

# Genetic Mutation Classification

---

**Background:** The challenge is distinguishing the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers).

- Meanings of 9 mutation classes:  
Gain-of-function,  
Likely Gain-of-function,  
Loss-of-function,  
Likely Loss-of-function,  
Neutral,  
Likely Neutral,  
Switch-of-function,  
Likely Switch-of-function,  
Inconclusive



# Dataset: An Example

---

given a **<gene, mutation, text>** sample:

BRCA1, P1749R| BRCA1 is inactivated by gene mutations in >50% of familial breast and ovarian cancers. ... Similarly, three other frequently studied C-terminal BRCT mutants, P1749R, M1775R, and Y1853X, all displayed a dramatic decrease in foci localization compared with the wild-type protein.

*BRCA1* is the name of a sort of human gene

*P1749R* is the name of a sort of related mutation

**Goal:** Predict a class label for the given **<gene, mutation, text>** based on the information:

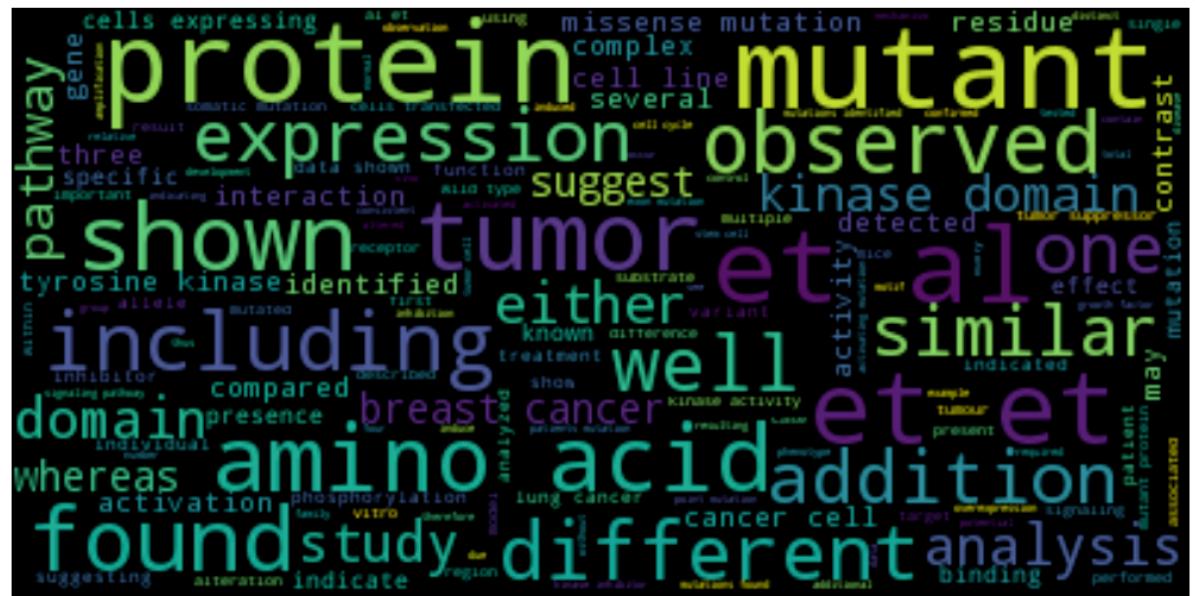
- text content
- gene/mutation name

# Dataset: Sample Size

- Training data: 3321 samples, each sample can be denoted as <gene, mutation, text>, and each text is extracted from Pubmed literatures.
  - Off-line test data: 368 samples. (Stage 1 solution)
  - Labels: 9 Classes indicating the corresponding gene mutation types.

## 3 Basic Problems:

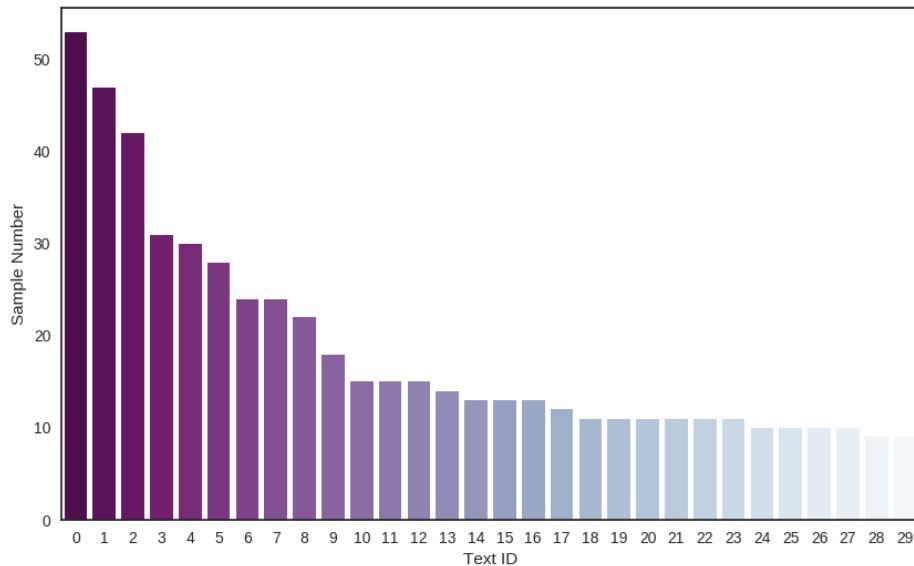
- ✓ different sample, common text;
  - ✓ long, noisy text;
  - ✓ almost no collaborative information in <gene, mutation> set.



# Problem-1

---

- Distribution of the counts of common text



Share the Save Text

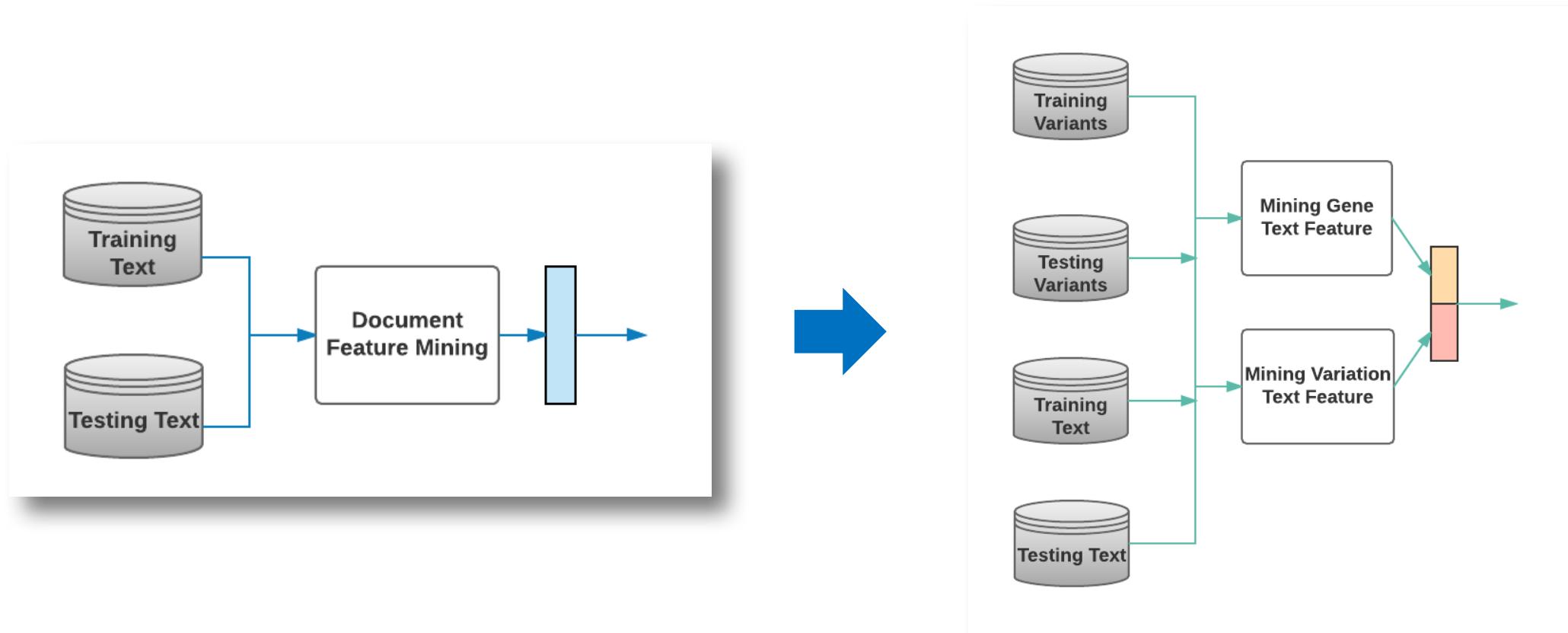
Gene	Mutation	Text	Class
BRCA1	T1773I	Genetic screening of the breast and ovarian cancer ...	1
BRCA1	M1663L	Genetic screening of the breast and ovarian cancer ...	5
...			

Same text but different class label

Different sample may share the same text entry! → Need to find evidence from other perspective.

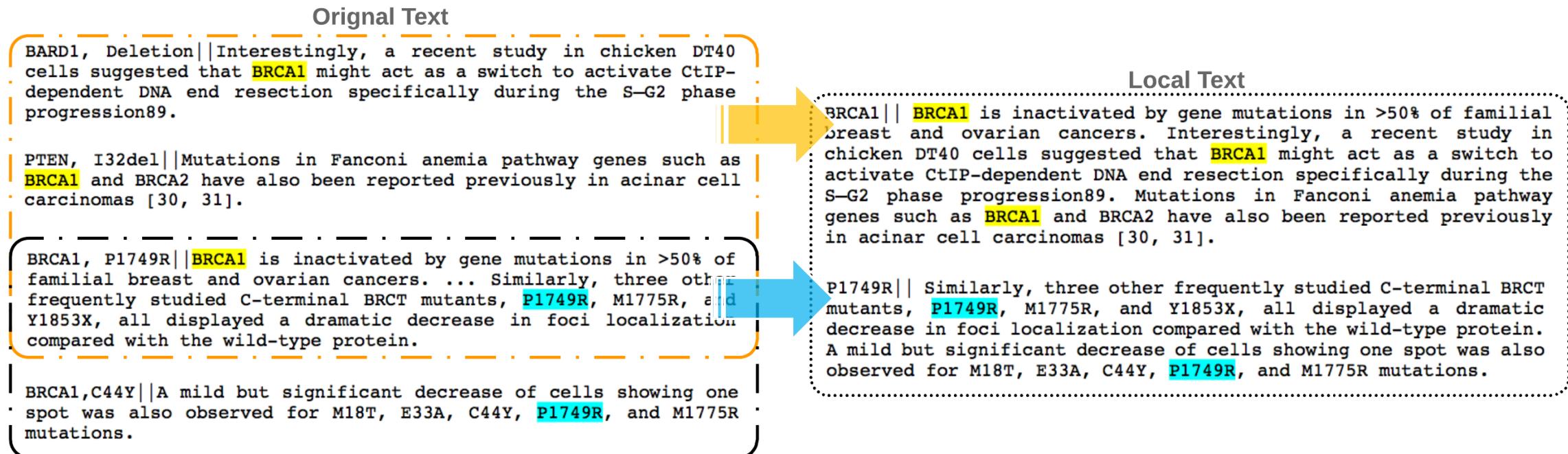
# Solution to Problem-1

- Document Feature Extract → Entity Text Feature Extract



# Solution to Problem-1

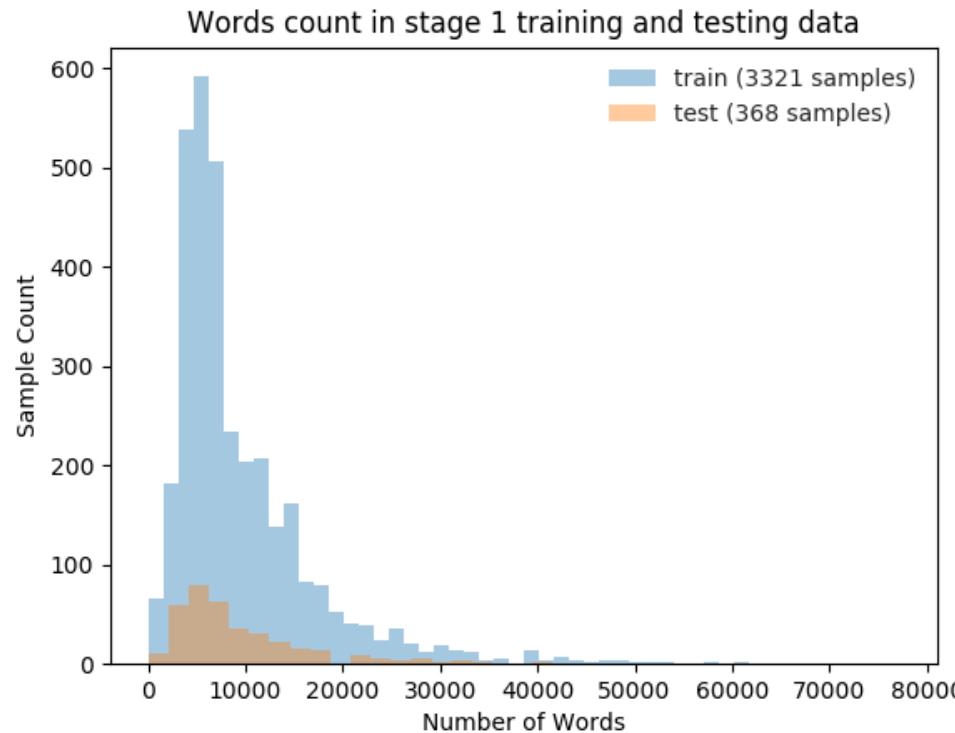
- Entity Text View



# Problem-2

---

- Overall distribution of the text entry lengths

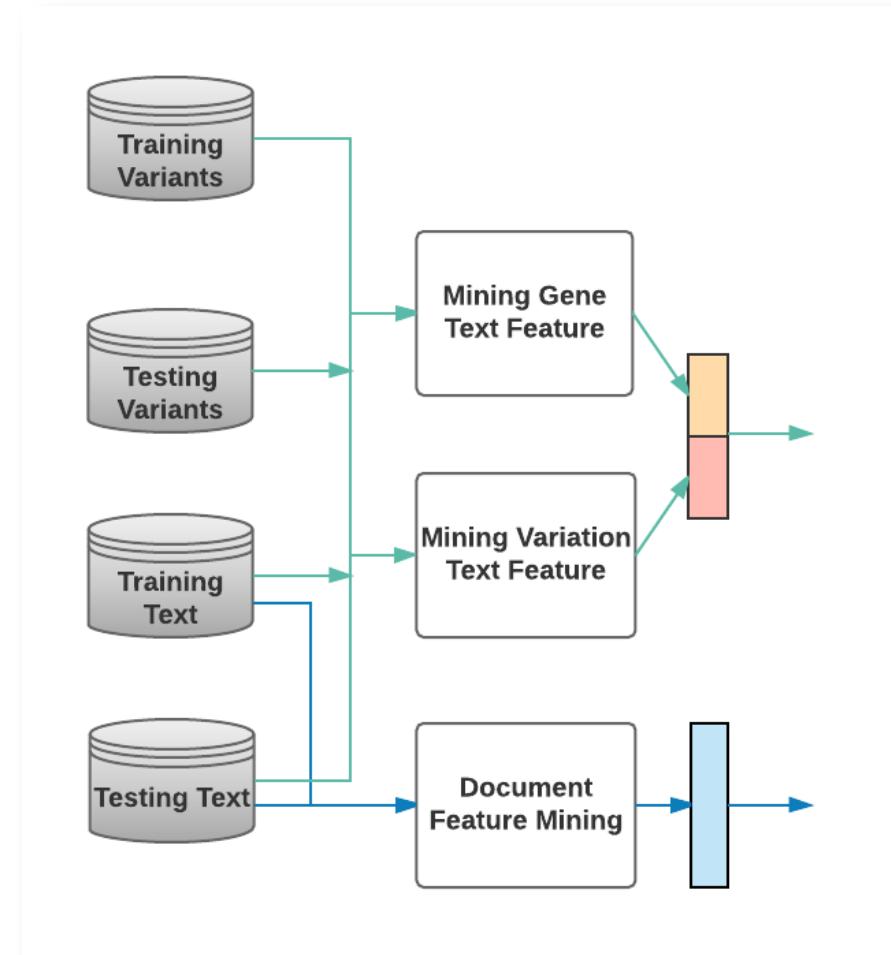


The text for each sample is **extremely long** with a large amount of noisy information, i.e., the median value of word count per text: 6,743; the maximum word count in a text: 77,202.

# Solution to Problem-2

---

- Text Mining
  - ✓ Representation by Domain Knowledge
  - ✓ Representation by Word
  - ✓ Representation by Sentence
  - ✓ Representation by Document



# Solution to Problem-2

---

- Text Mining-Representation by Domain Knowledge

- ✓ Dictionary Extension

- Keywords Extraction



keywords extracted from the title of the related **PubMed** articles obtained from **Oncokb**.

- Bio-Entity Extraction

**PubTator**: Named entity recognition (NER) tools used include GeneTUKit, GenNorm and tmVar.



<http://oncokb.org/#/>



Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522, 2013.

# Solution to Problem-2

---

- Text Mining-Representation by Word

- ✓ Dictionary Construction

- noun/verb/adj./adverb. (PoS tagging)  
9,868 words in total

- novel tf-idf score: consider the term appearance in classes

$$idf_t = \log \frac{C}{cf_t}$$

- n-Gram. (unigram/bigram/trigram)  
9,473,363 dimension at maximum

Feature	dimension
Noun/verb/adj./adverb count	9868
TF-IDF <sup>c</sup>	9456
n-Gram	9,473,363
Keyword count	3379
Gene count	6987
Mutation count	3035



We may need  
dimension reduction

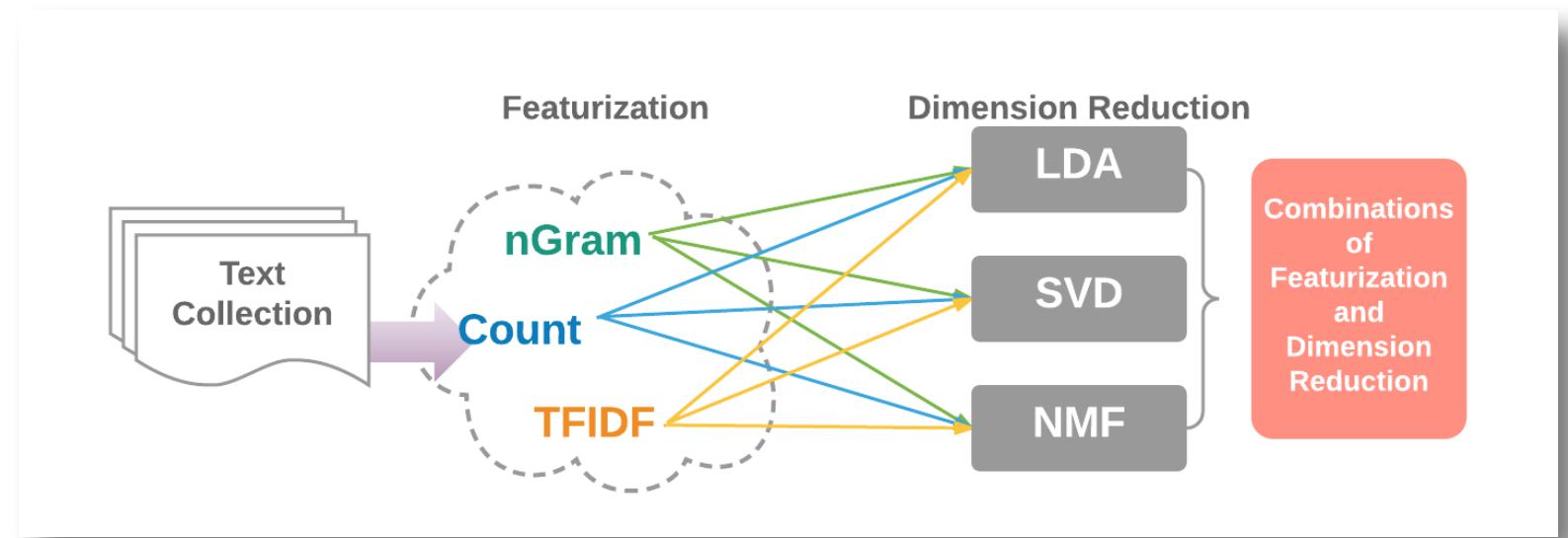
# Solution to Problem-2

- Text Mining-Dimension Reduction

- Latent Dirichlet Allocation (LDA)
- Singular Value Decomposition (SVD)
- Non-negative matrix factorization (NMF)

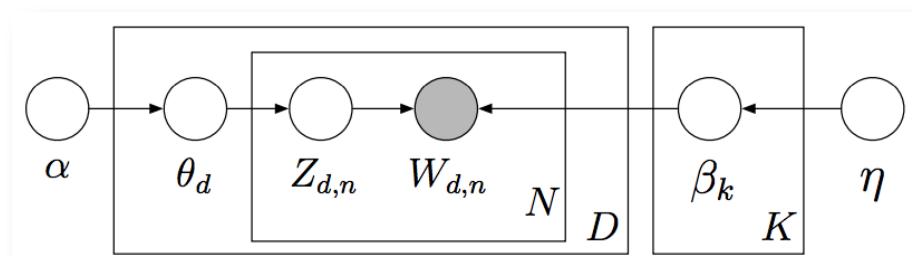
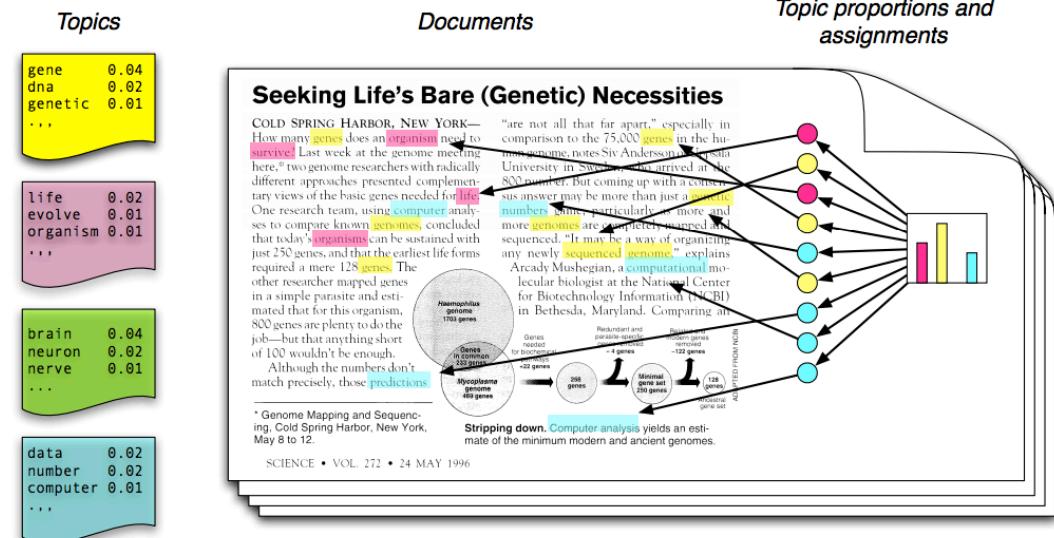
Good News:

*Can be extended to Entity-View*



# Solution to Problem-2

- Text Mining-Dimension Reduction
  - Latent Dirichlet Allocation (LDA)
    - ✓ Each topic is a distribution over words
    - ✓ Each document is a mixture of corpus-wide topics
    - ✓ Each word is drawn from one of those topics

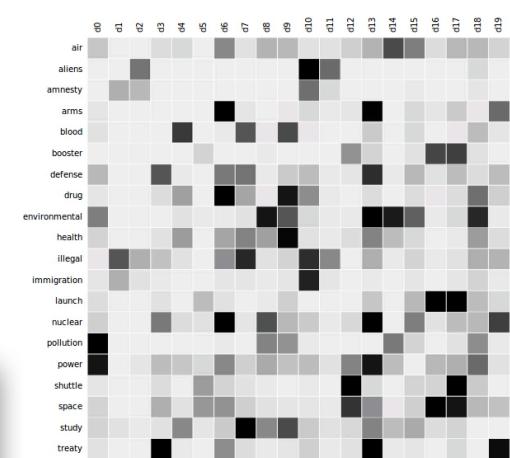


$$\prod_{i=1}^K p(\beta_i | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:k}, z_{d,n}) \right)$$

# Solution to Problem-2

- Text Mining-Dimension Reduction
  - Singular Value Decomposition (SVD)  $X = U\Sigma V^T$

$$\begin{array}{c} X \\ \downarrow \\ (\mathbf{d}_j) \\ \downarrow \\ \left[ \begin{array}{cccc} x_{1,1} & \dots & x_{1,j} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \dots & x_{i,j} & \dots & x_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,j} & \dots & x_{m,n} \end{array} \right] = (\hat{\mathbf{t}}_i^T) \rightarrow \left[ \begin{array}{c} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_l \end{array} \right] \dots \left[ \begin{array}{c} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_l \end{array} \right] \cdot \left[ \begin{array}{ccc} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{array} \right] \cdot \left[ \begin{array}{c} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_l \end{array} \right] \end{array}$$



occurrence matrix

- ✓ Low Rank Approximation: the observed term-document matrix can be approximate by an underlying low-rank matrix.  $\text{rank}(X_{m \times n}) = \min(m, n)$

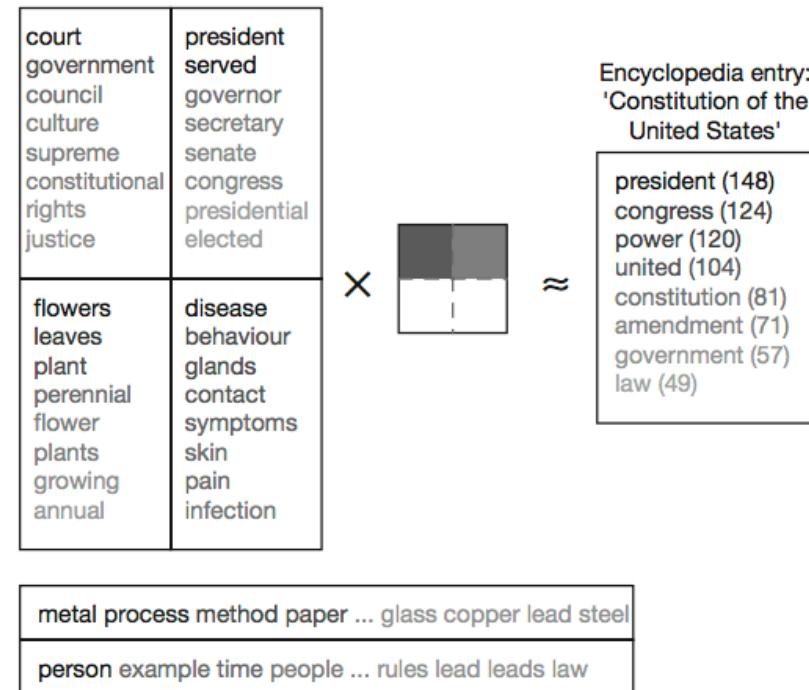
# Solution to Problem-2

- Text Mining-Dimension Reduction
  - Non-negative matrix factorization (NMF)

$$\underbrace{X(:, j)}_{j\text{th document}} \approx \sum_{k=1}^r \underbrace{W(:, k)}_{k\text{th topic}} \quad \underbrace{H(k, j)}_{\text{importance of } k\text{th topic in } j\text{th document}}$$

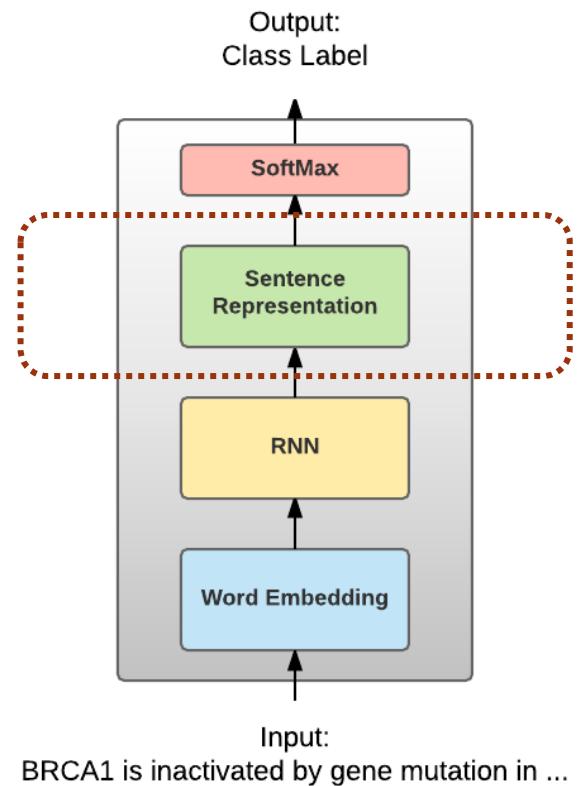
with  $W \geq 0$  and  $H \geq 0$ .

Better Interpretation

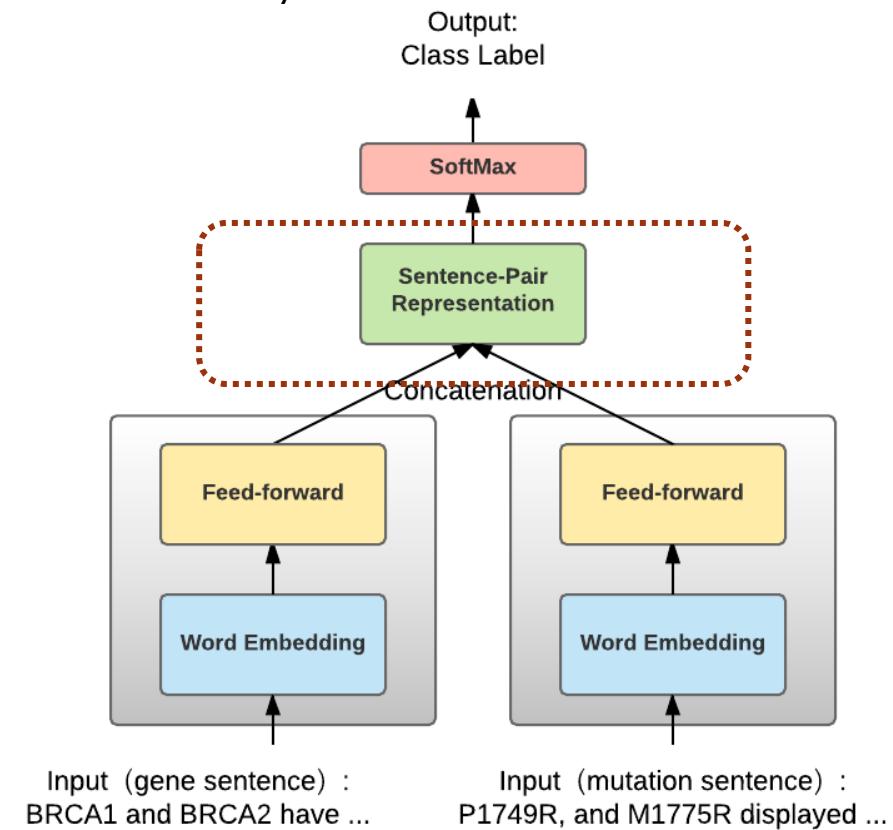


# Solution to Problem-2

- Text Mining-Representation by Sentence
  - ✓ RNN-Original Text View

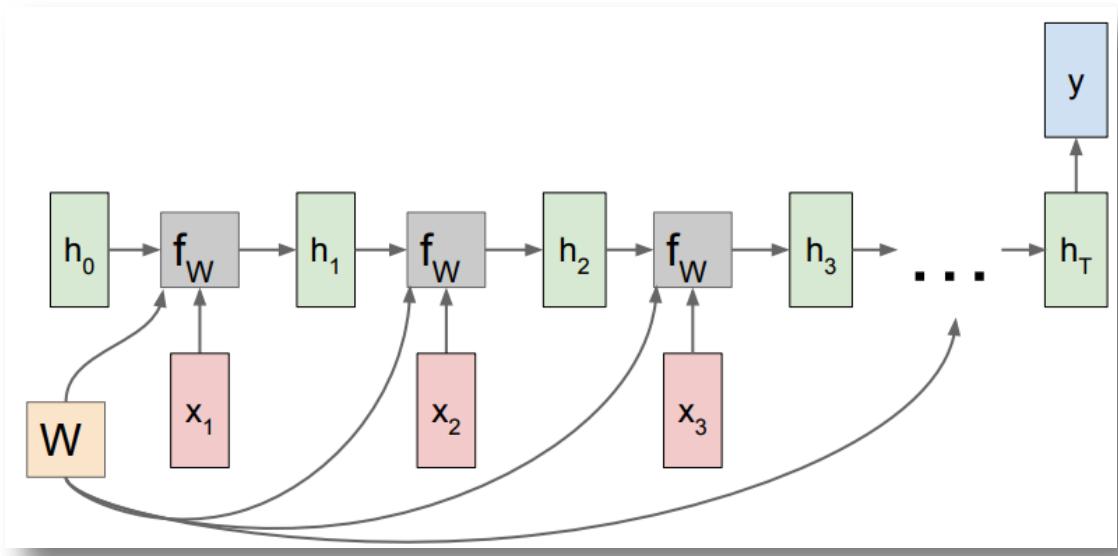


- ✓ CNN-Entity Text View

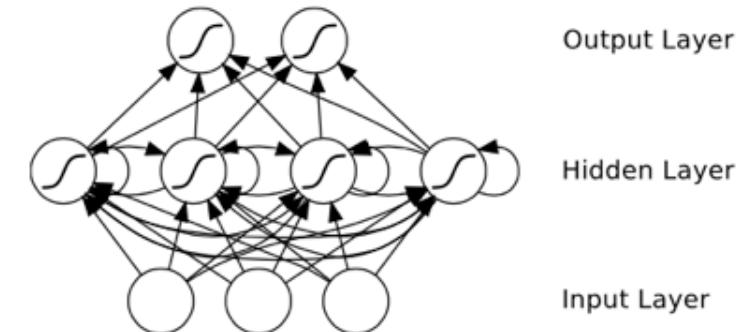


# Solution to Problem-2

- Recurrent Neural Network
  - ✓ Hidden Layer: LSTM



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



$f_W(\cdot)$  is defined as:

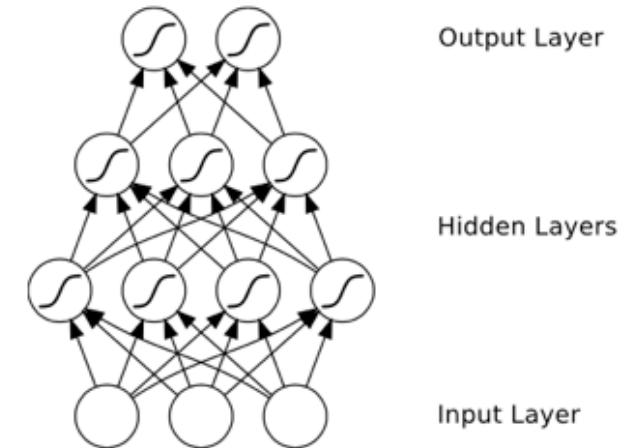
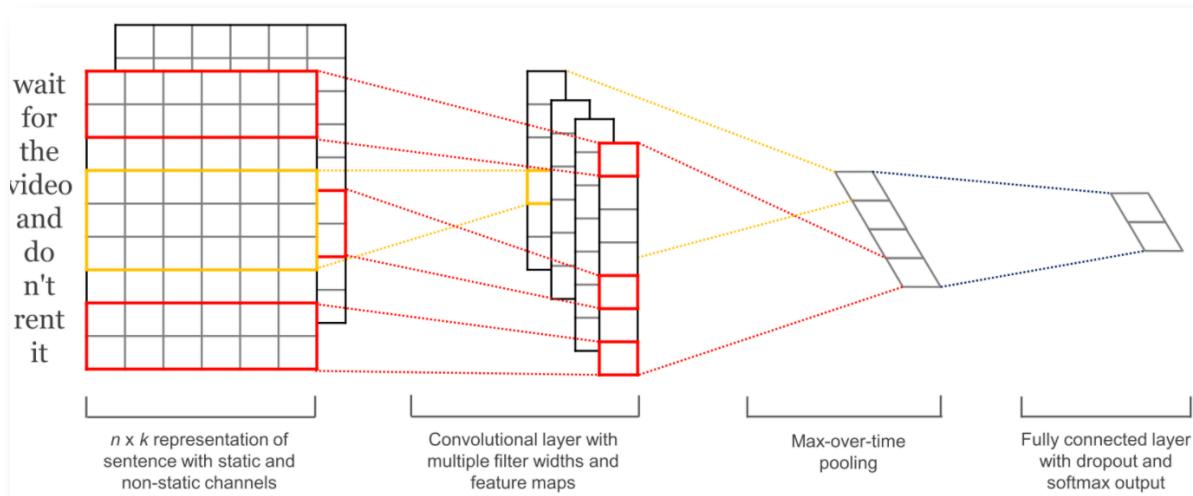
$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ l_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W \cdot \begin{bmatrix} h_{t-1} \\ e_t \end{bmatrix}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot l_t$$

$$h_t^s = o_t \cdot c_t$$

# Solution to Problem-2

- Convolutional Neural Network



Max pooling

$$\mathbf{c} = [c_1, c_2, \dots, c_{n-h+1}],$$

$$\hat{c} = \max\{\mathbf{c}\}$$

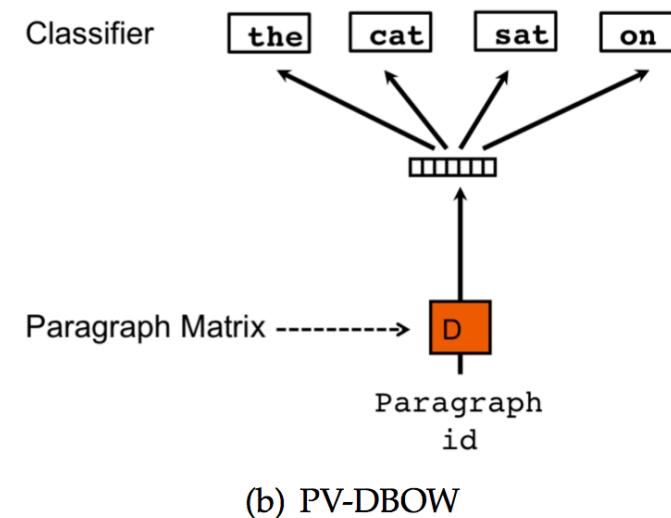
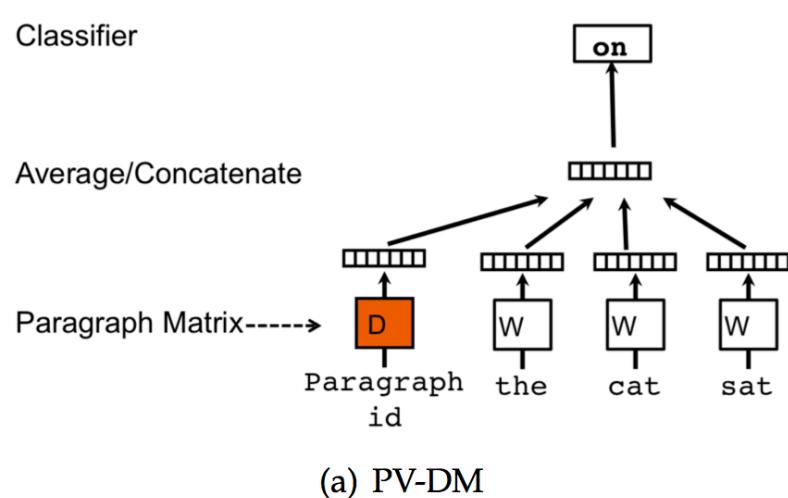
Convolution  $c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b).$

# Solution to Problem-2

- Text Mining-Representation by Document
  - ✓ Paragraph Vector

Bad News:

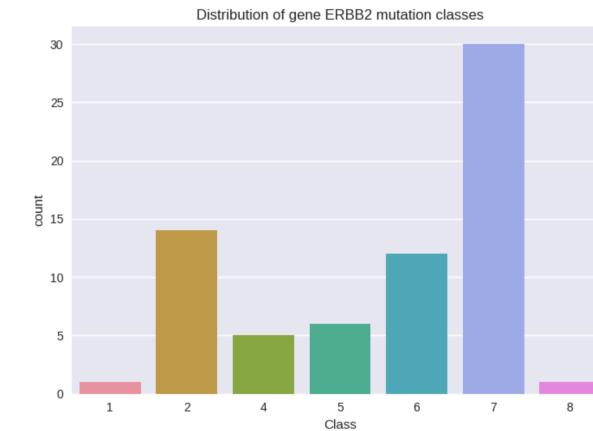
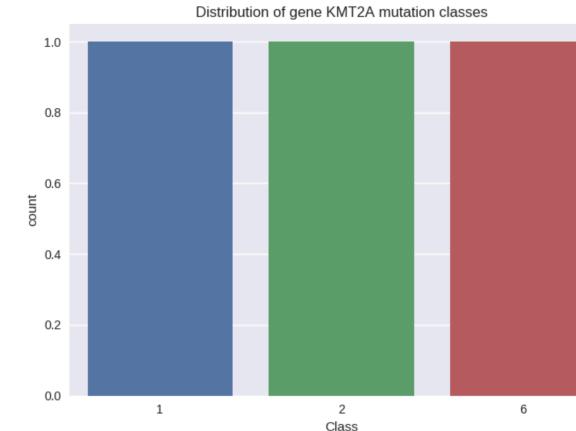
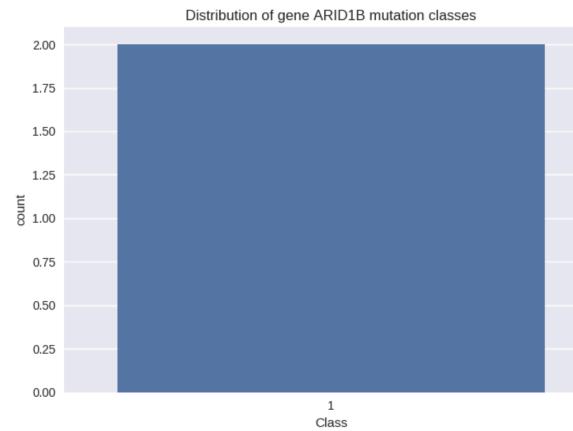
Can only be used to Original-View



# Problem-3

---

- Gene distribution over classes



**However,** we cannot use above distribution information of gene, since:

there are only a few overlapped samples in training/test set

- ✓ overlap of train/test gene: 9.7%
- ✓ overlap of train/test variation: 0.17%

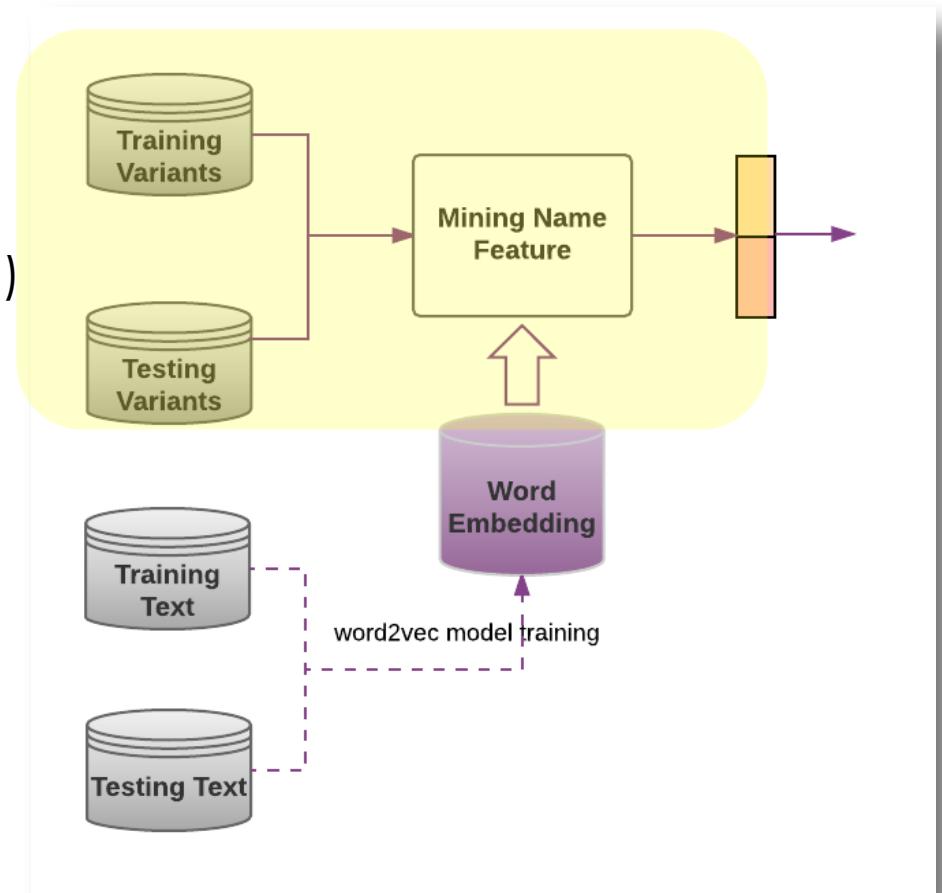
Only name text can be used

# Solution to Problem-3

- Name Mining
  - ✓ Character Level n-Gram Encoding  
(n=8, 9,473,363 dim, then dimension reduction)

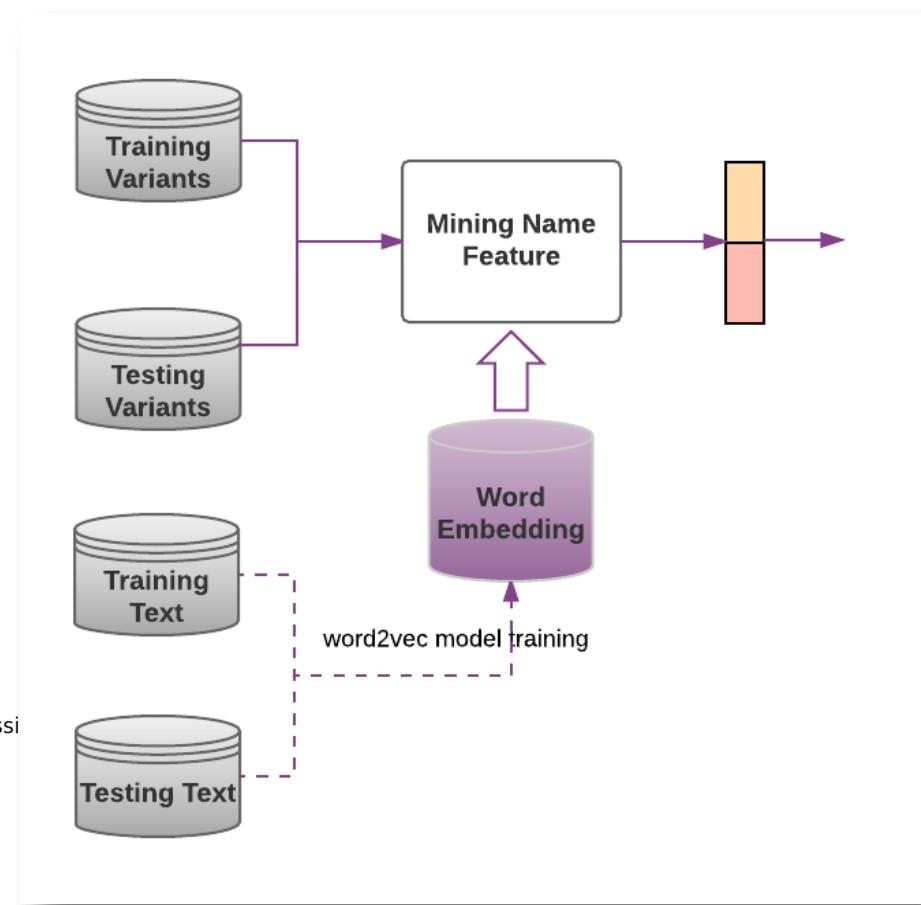
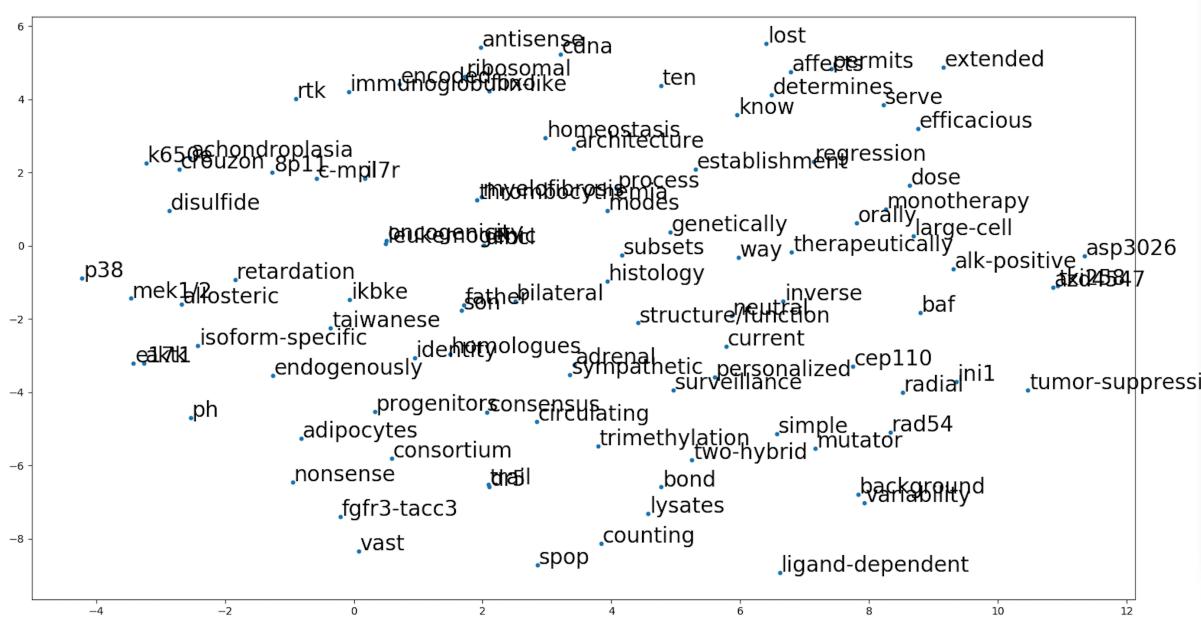
Gene	Mutation
TET2	Y1902A
MTOR	D2512H
KIT	D52N
SPOP	F125V
ETV1	Amplification
.....	

text of <gene, mutation> name



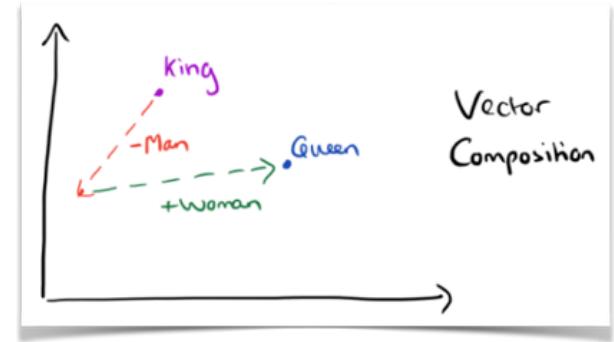
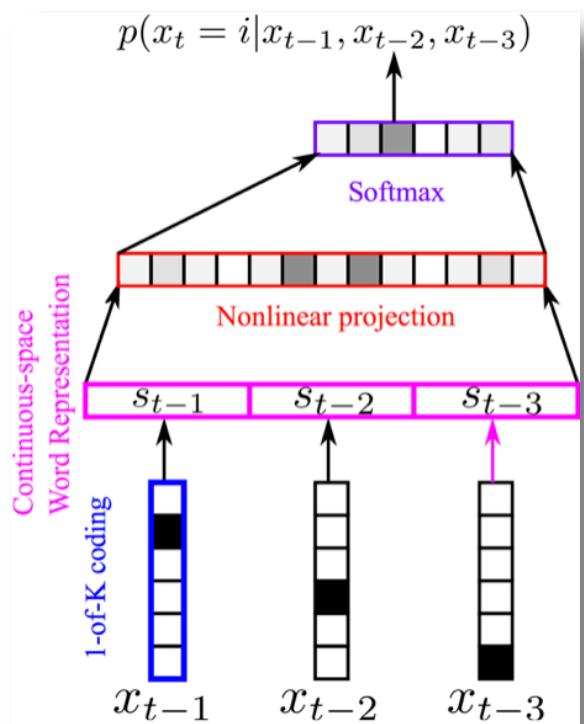
# Solution to Problem-3

- Name Mining
  - ✓ Word Embedding (word2vec)
  - ✓ Embedding Distribution



# Solution to Problem-3

- Name Mining
  - Word Embedding (word2vec)



Neural Language Modeling

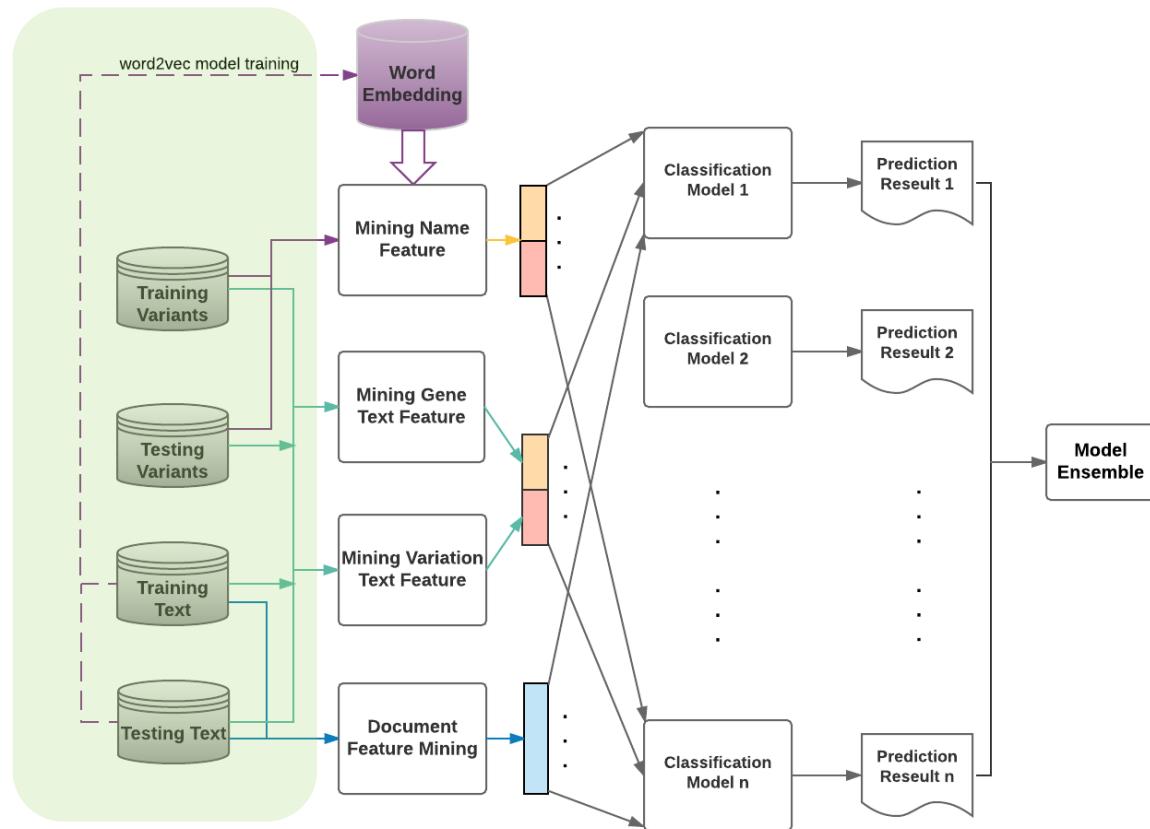
$$p(\mathbf{x}_t | \mathbf{x}_{t-n}, \dots, \mathbf{x}_{t-1}) = f_{\mathbf{x}_t}(\mathbf{x}_{t-n}, \dots, \mathbf{x}_{t-1})$$

Softmax:

$$p(x_t = i | x_{t-n}, \dots, x_{t-1}) = \frac{\exp(y_i)}{\sum_{j=1}^{|V|} \exp(y_j)}$$

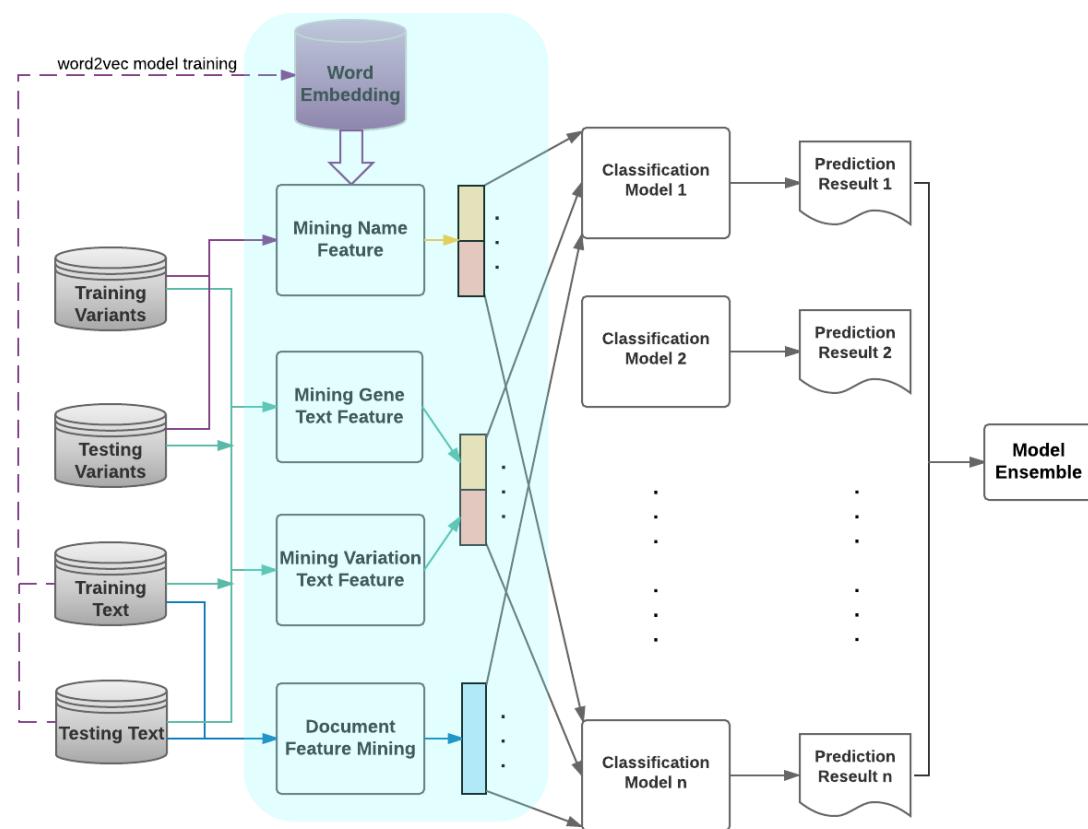
# Architecture Overview

## Input Data



# Architecture Overview

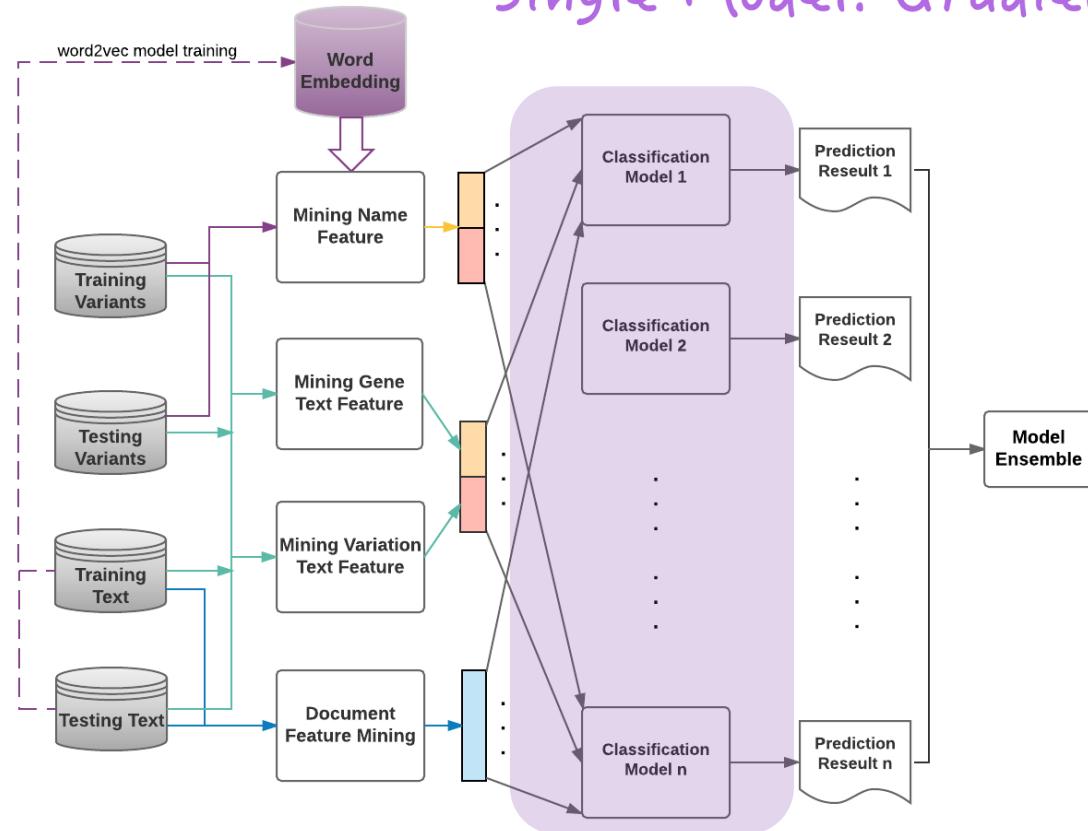
## Feature Engineering



# Architecture Overview

---

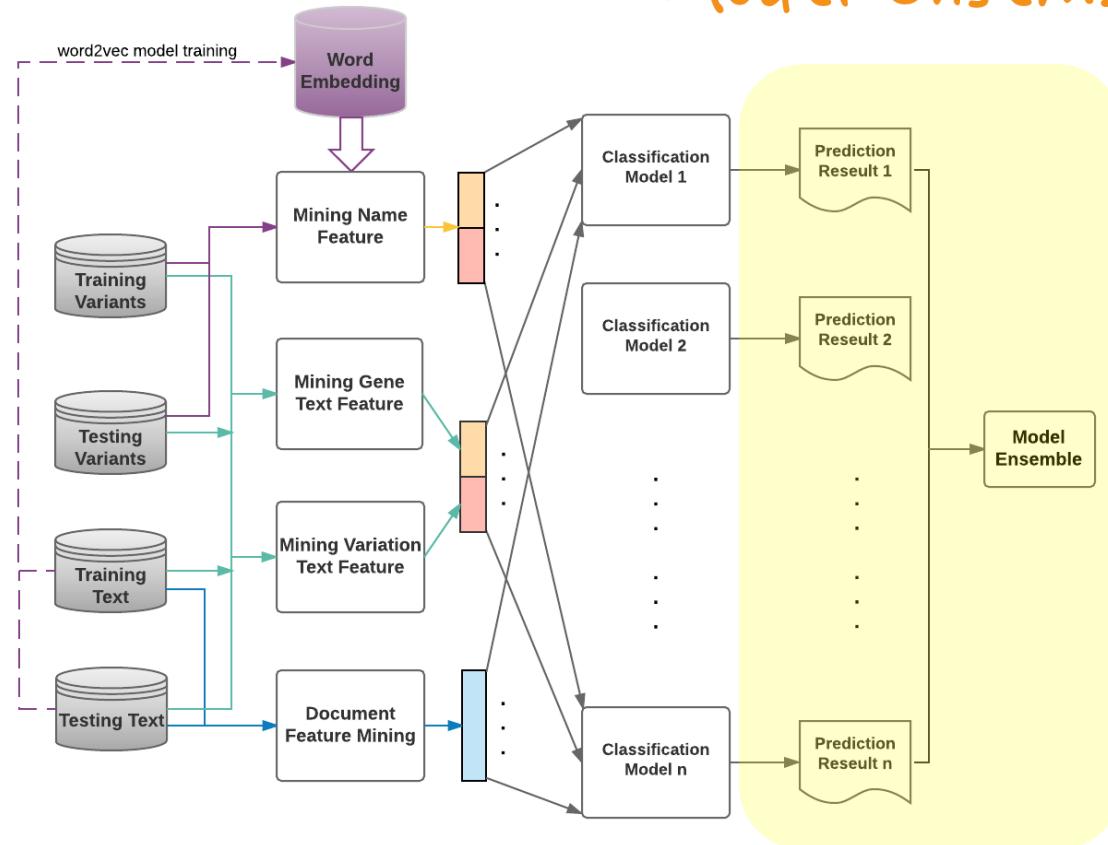
## Single Model: Gradient Boosting



# Architecture Overview

---

## Model Ensemble



# Experimental Settings

---

- ✓ Off-line Test
  - Stage-1 Test (368 samples)
  - 5-fold cross validation
  
- ✓ Metrics

$$\text{Logloss} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M (y_{i,j} \log(p_{i,j}))$$

# Results of Single Model

---

Table 1: Results of GBDT model in terms of logloss on 5-fold cross validation and stage1 test set

<b>Model Id</b>	<b>5-fold cv</b>	<b>Stage1 test</b>
GBDT_1	0.7068	0.5997
GBDT_2	0.6930	<b>0.5638</b>
GBDT_3	<b>0.6870</b>	0.5743
GBDT_4	0.6901	0.5657

XGBoost

Table 2: Results of GBM model in terms of logloss on 5-fold cross validation and stage1 test set

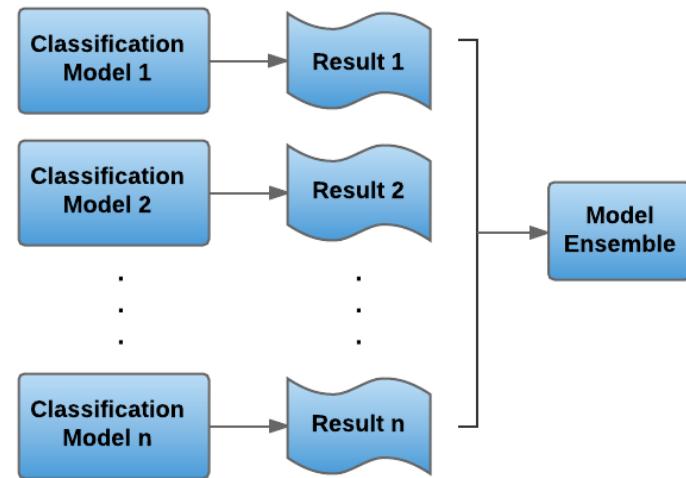
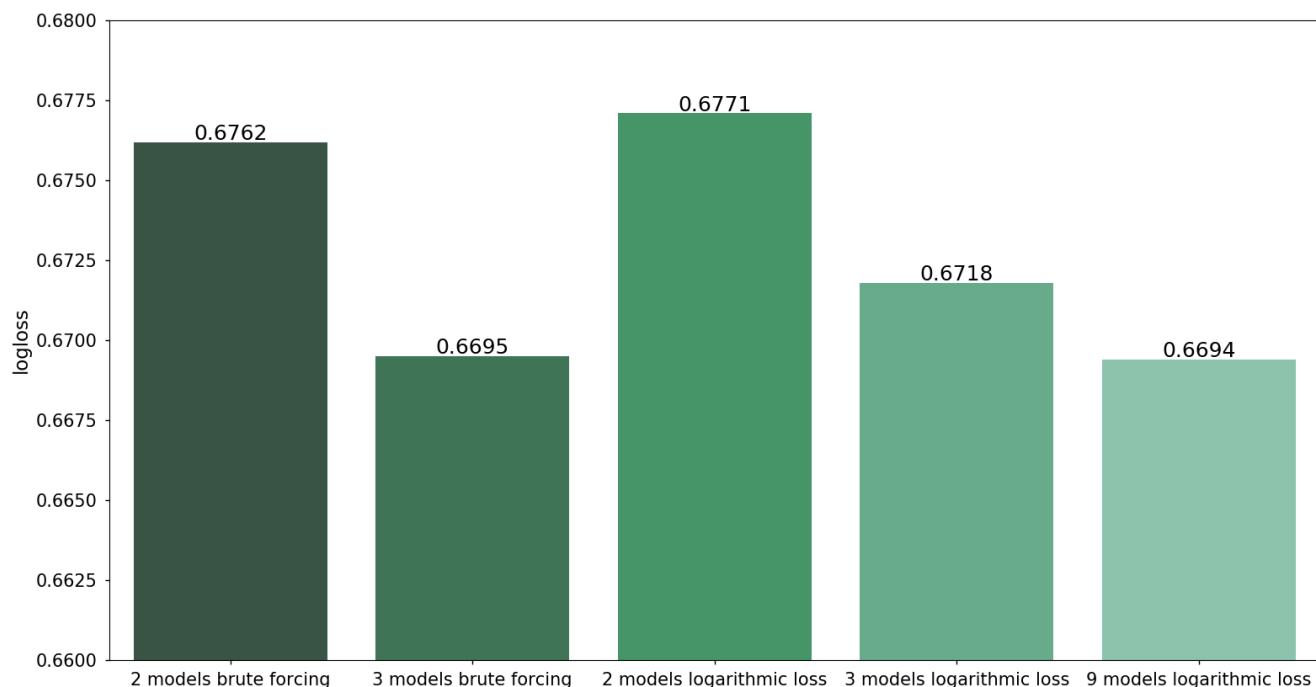
<b>Model Id</b>	<b>5-fold cv</b>	<b>Stage1 test</b>
GBM_1	0.7005	0.6090
GBM_2	0.7121	0.6152
GBM_3	<b>0.6967</b>	0.6139
GBM_4	0.7028	0.6178
GBM_5	0.7001	<b>0.6006</b>

LightGBM

# Results of Ensemble Model

---

- 2/3/9 model ensemble
  - ✓ Brute Force Grid Search
  - ✓ Logarithmic Loss Minimization



# Results of Ensemble Model

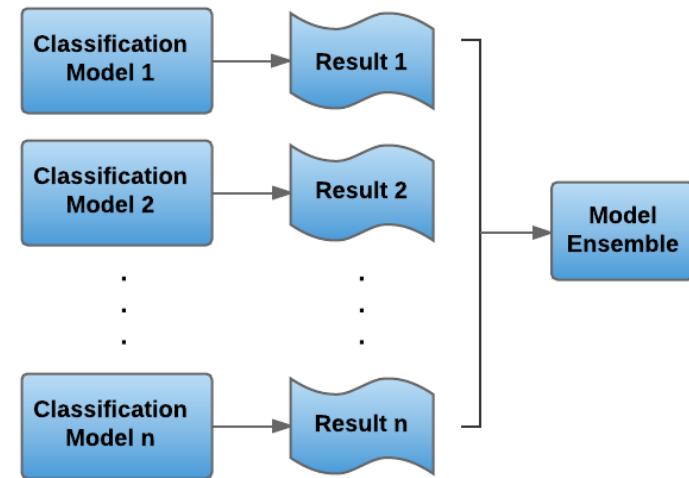
---

- A Variant of Logarithmic Loss Minimization

$$\text{Logloss} = -\sum_{j=1}^M \frac{1}{N_j} \sum_{i=1}^{N_j} (y_{i,j} \log(p_{i,j}))$$

Table 7: Results of the ensemble 9 models by logarithmic loss minimization

Method	5-fold cv
LOGLOSS_MIN	0.6711
LOGLOSS_MIN_CL	0.6695



# An Interface

Weill Cornell Medicine

## Prediction of Gene-Variation Type

Gene Name, e.g., TGFBR2

Variation Name, e.g., Deletion

Predict

Well done! Successful prediction.

Type	Prediction Probability	Ground Truth
1	0.802710146617	1
2	0.0153880442257	0
3	0.000548553247686	0
4	0.176723589183	0
5	0.000539179914472	0
6	0.000621710422534	0
7	0.00254607929569	0
8	0.000522774675983	0
9	0.000399922417731	0

Prediction Probability Visualization

Type	Prediction Probability
1	0.802710146617
2	0.0153880442257
3	0.000548553247686
4	0.176723589183
5	0.000539179914472
6	0.000621710422534
7	0.00254607929569
8	0.000522774675983
9	0.000399922417731

The demo link:

<http://34.207.90.132:8000/demo/>

Prediction Results

# Summary

---

- We developed a comprehensive pipeline to perform gene mutation type classification.
- Our solution includes various ways of extracting features from two views of text mining: original text mining and local text mining as well as name mining.
- Predictive models are build for model ensemble. The empirical study shows that we achieve off-line performance of 0.55 on stage1 test data, 0.66 on 5- fold cross validation.

# Thank you!