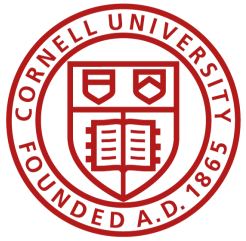


Weill Cornell Medicine

Classifying Clinically Actionable Genetic Mutations

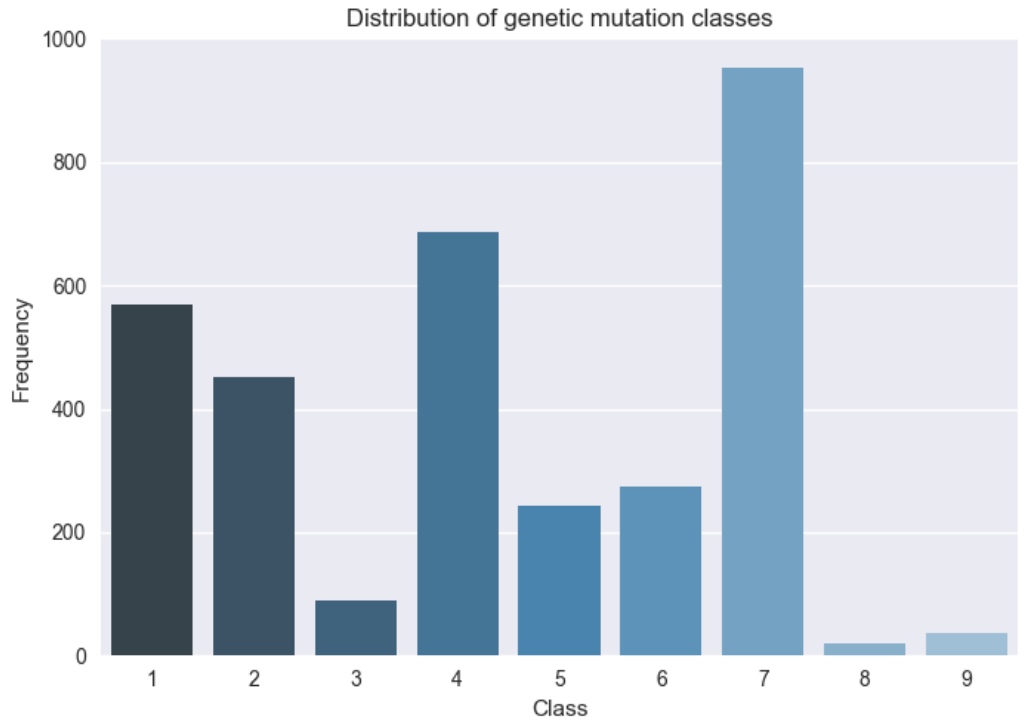
Xi Zhang, Dandi Chen, Yongjun Zhu, Chao Che, Chang Su, Sendong Zhao, Xu Min and Fei Wang.

Department of Healthcare Policy and Research, 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.
```

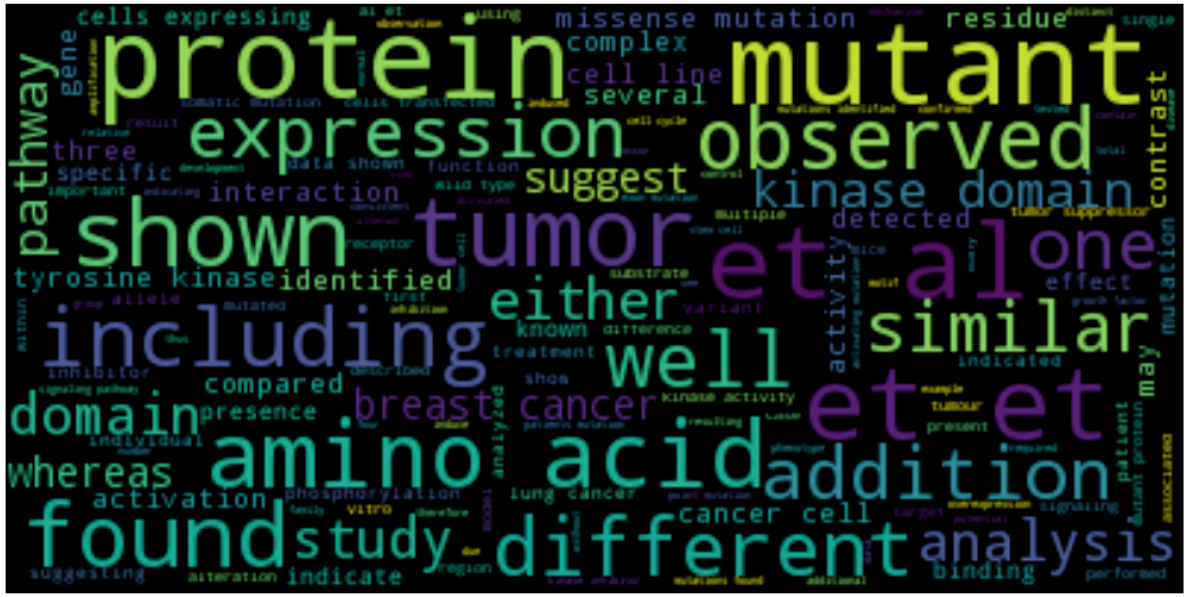
BRCA1 is the name of a sort of human gene
P1749R is the name of a sort of related mutation

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.

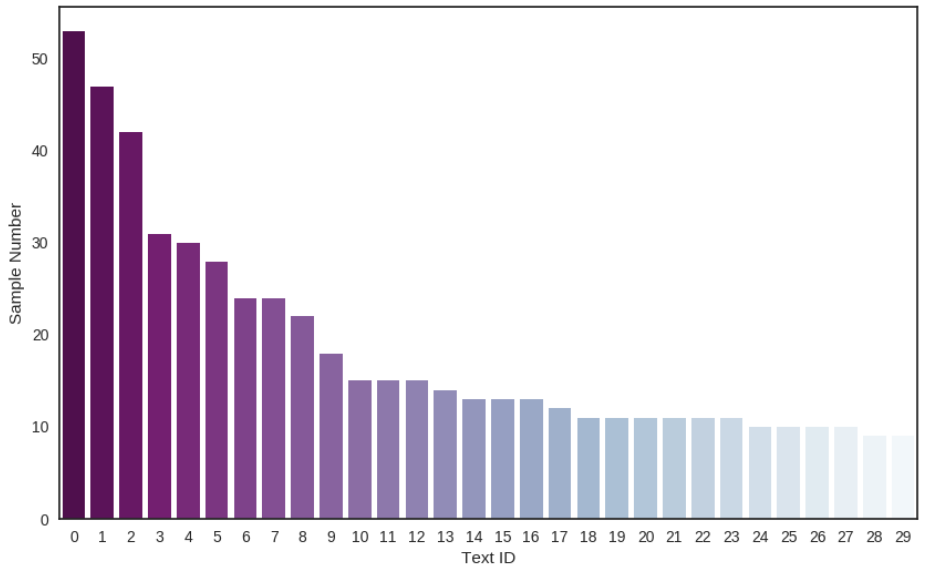


3 Basic Problems:

- ✓ 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

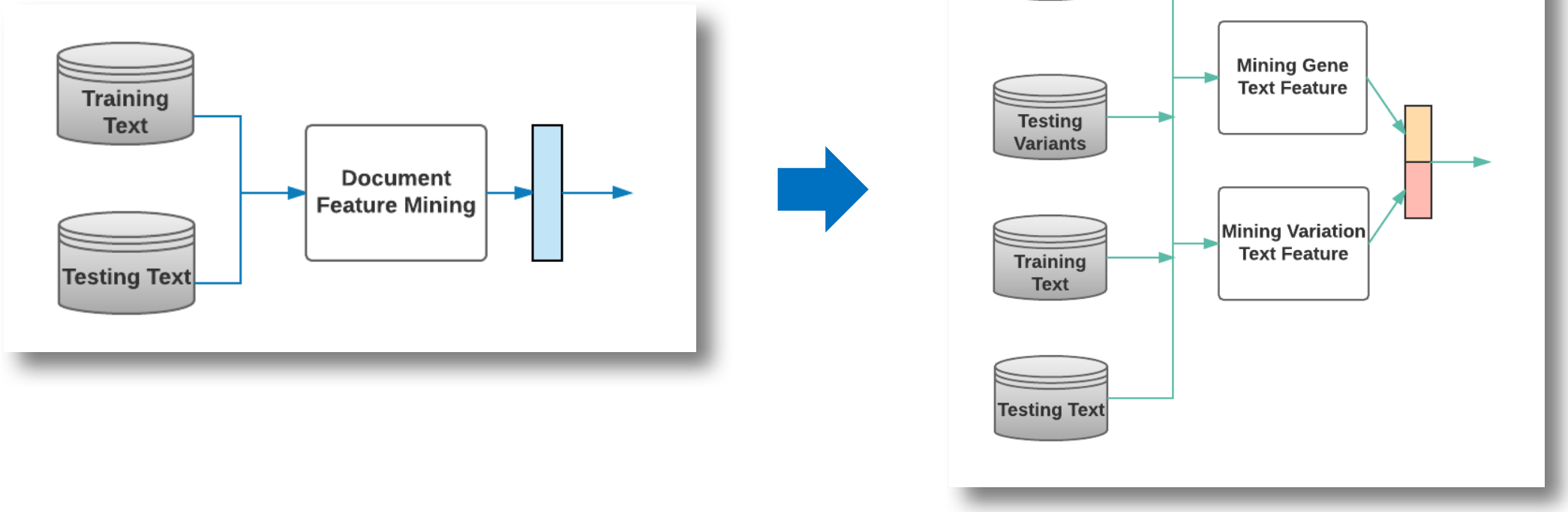
Gene	Mutation	Text	Class
BRCA1	T1773I	Genetic screening of the breast and ovarian cancer …	1
BRCA1	M1663L	Genetic screening of the breast and ovarian cancer …	5
		…	

Same text but different class label

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

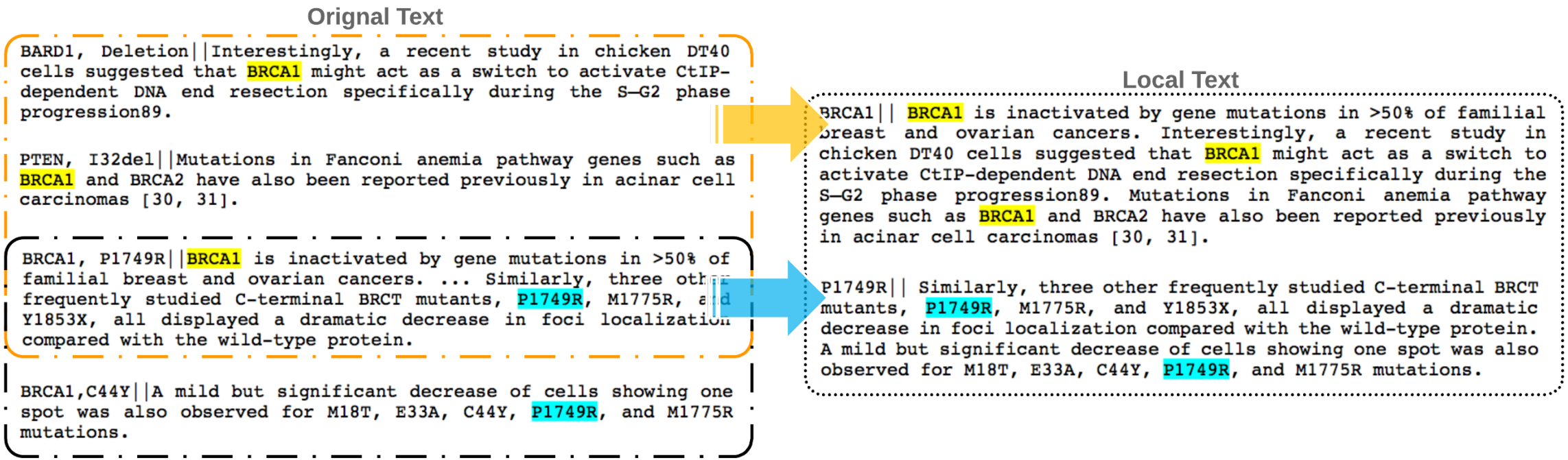
Solution to Problem-1

- Document Feature Extract ➔ Entity Text Feature Extract



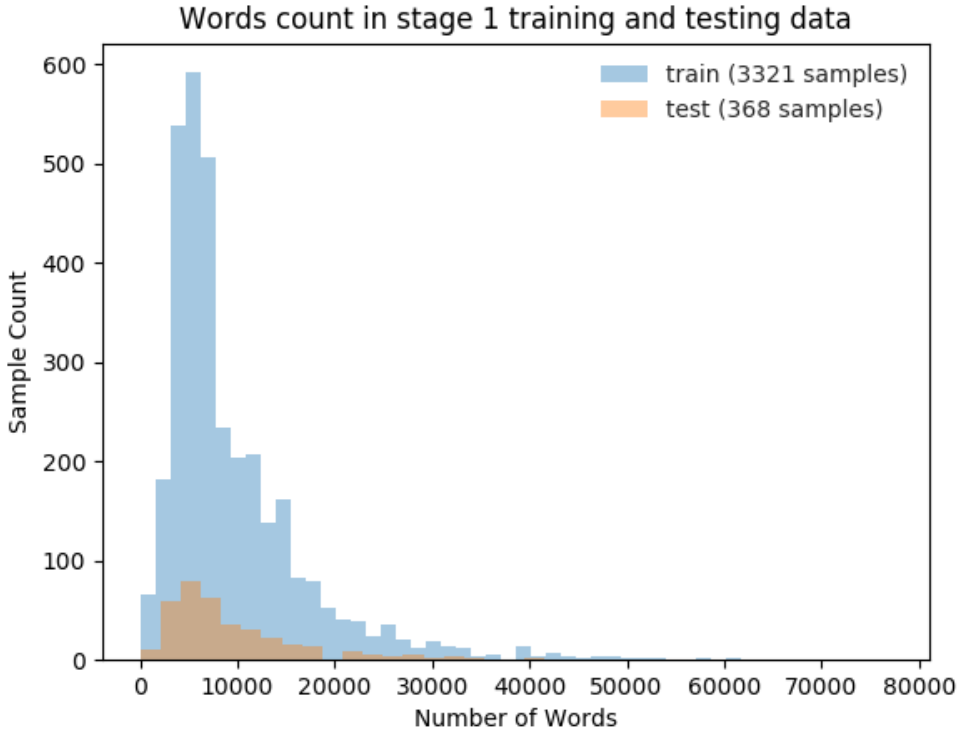
Solution to Problem-1

- Entity Text View



Problem-2

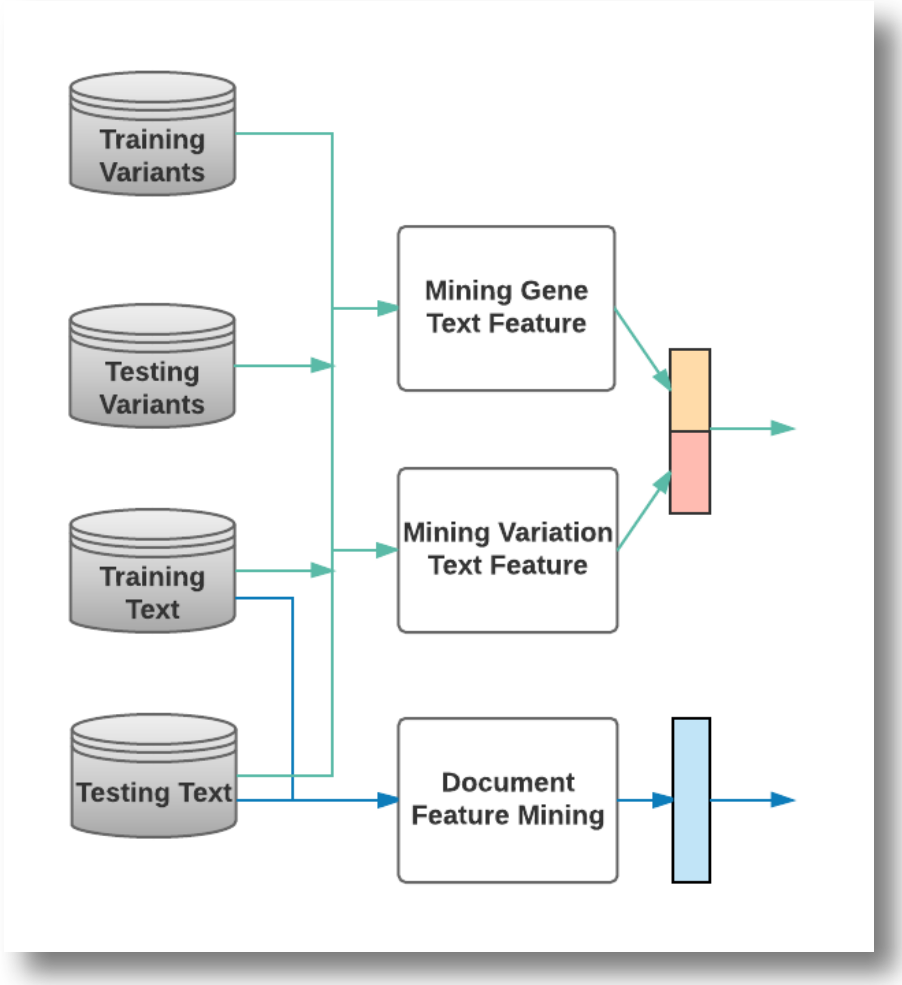
- Overall distribution of the text entry lengths



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

http://oncokb.org/#/

  - ✓ 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.

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

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

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

$$\mathrm{idf}_t = \log \frac{C}{\mathrm{cf}_t}$$

  - n-Gram. (unigram/bigram/trigram)

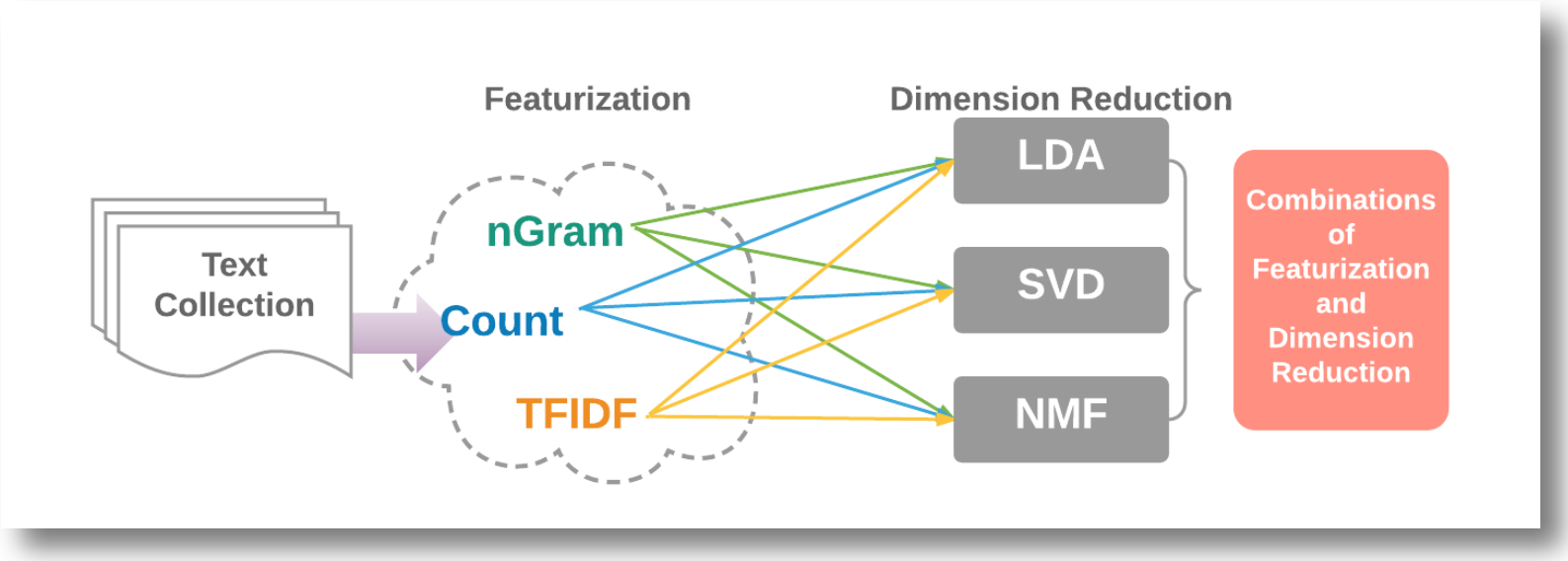
    9,473,363 dimension at maximum

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



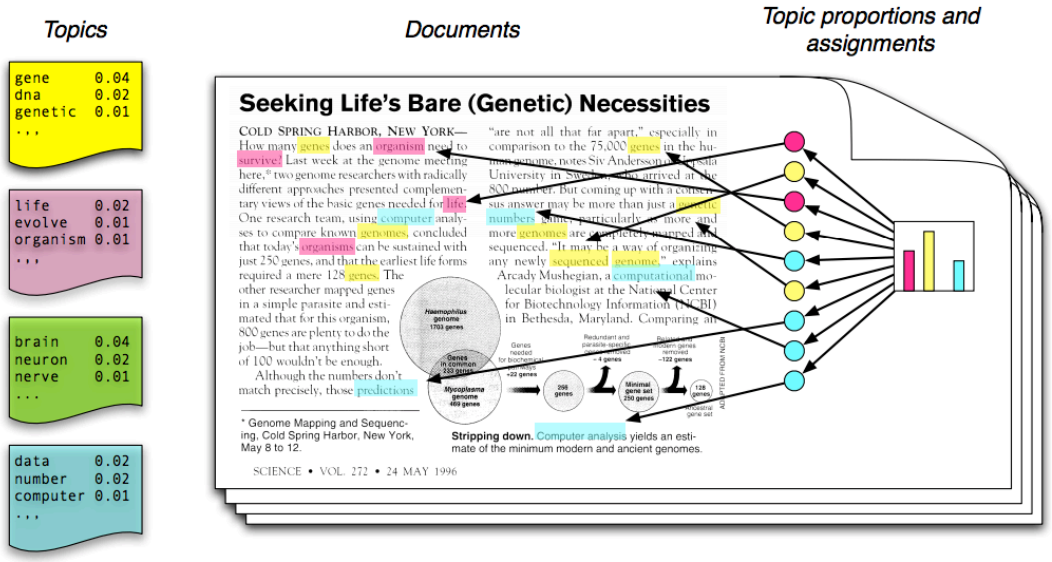
Solution to Problem-2

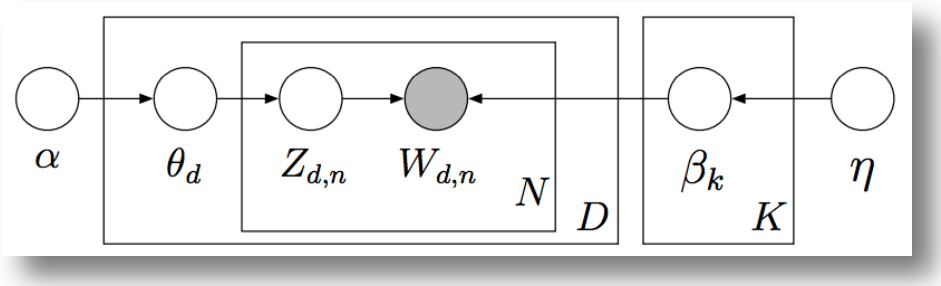
- Text Mining-Dimension Reduction
  - Latent Dirichlet Allocation (LDA)

✓ 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





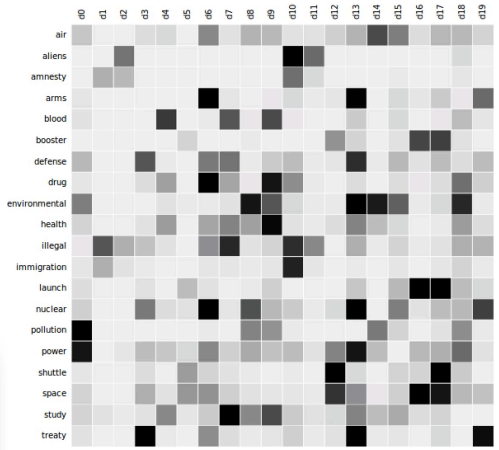
$$\prod_{i=1}^{K} p(\beta_i|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right)$$

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research*3.Jan (2003): 993-1022.

Solution to Problem-2

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

$$
\begin{array}{ccccc}
 & & X & & \\
 & & (\mathbf{d}_j) & & \\
 & & \downarrow & & \\
(\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,j} & \cdots & x_{m,n} \end{bmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow & \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} \cdots \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_l \end{bmatrix} \cdot \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix} \\
 & U & & \Sigma & V^T \\
 & & & & (\hat{\mathbf{d}}_j) \\
 & & & & \downarrow
\end{array}
$$



occurrence matrix

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

https://en.wikipedia.org/wiki/Latent_semantic_analysis

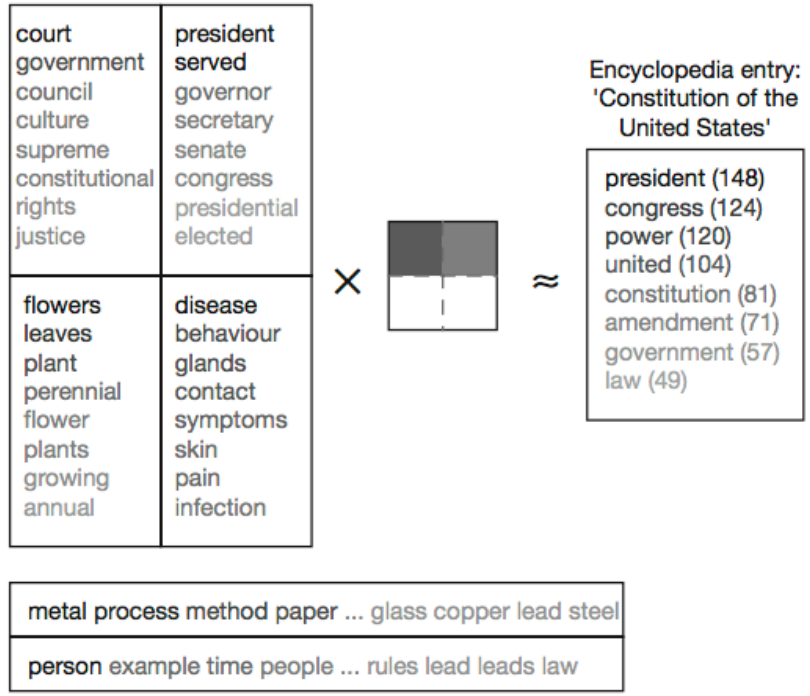
Solution to Problem-2

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

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

$$\text{with } W \geq 0 \text{ and } H \geq 0.$$

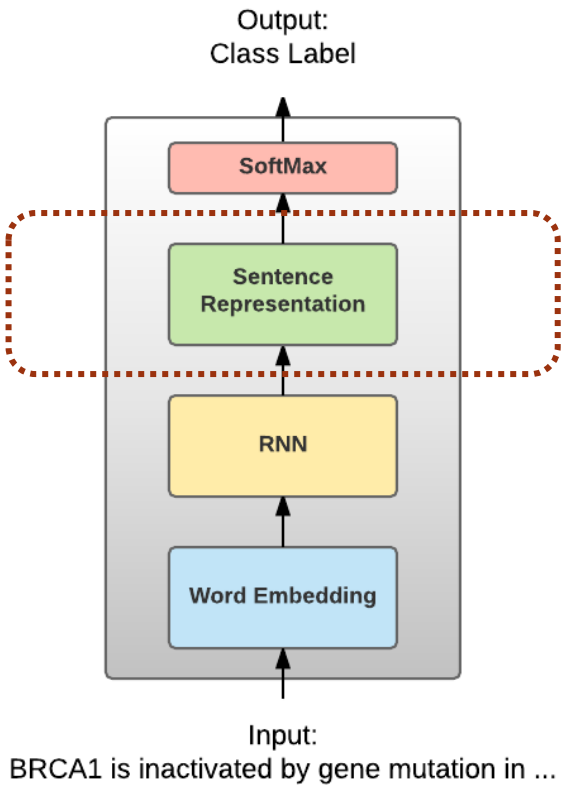
Better Interpretation



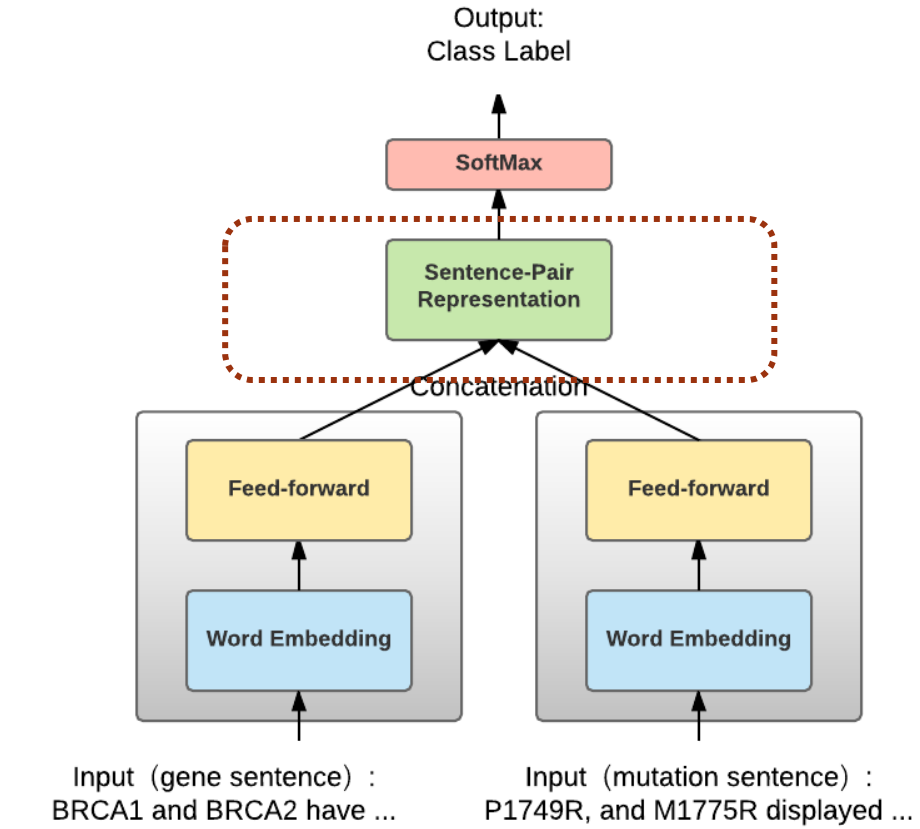
Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature*401.6755 (1999): 788-791.

Solution to Problem-2

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

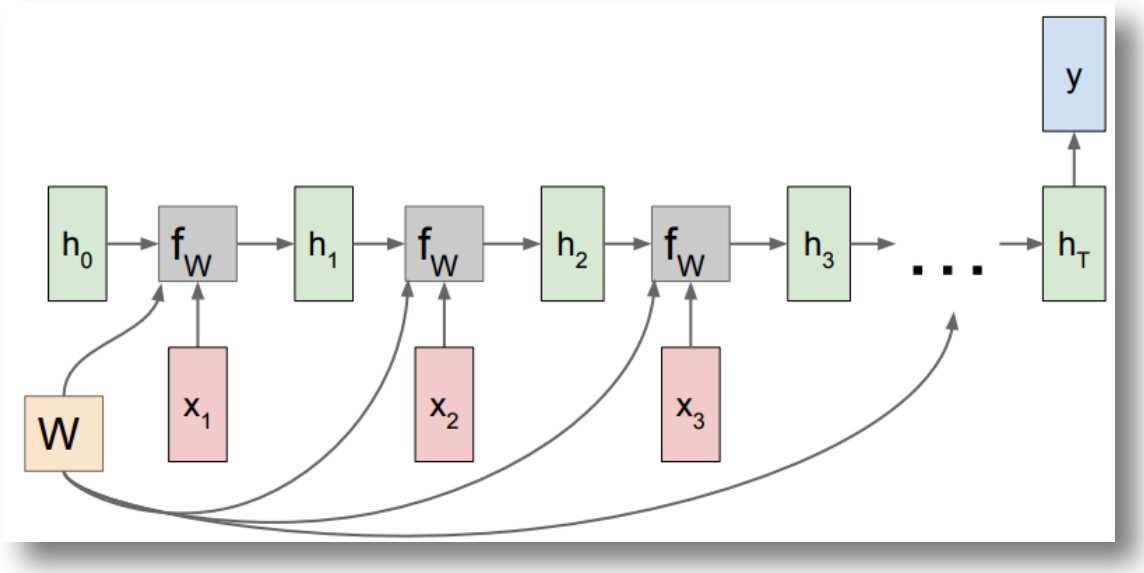


  - ✓ CNN-Entity Text View

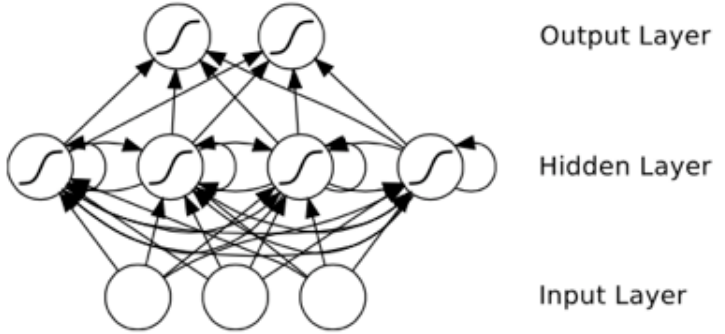


Solution to Problem-2

- Recurrent Neural Network
  - ✓ Hidden Layer: LSTM



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



fw( ) is defined as:

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

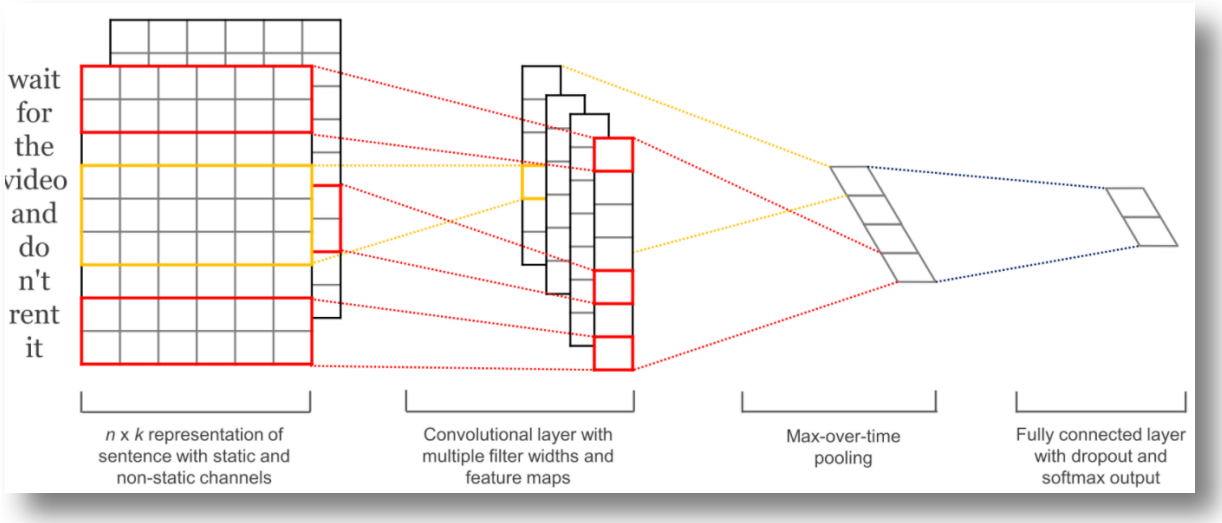
$$c_t = f_t \cdot c_{t-1} + i_t \cdot l_t$$

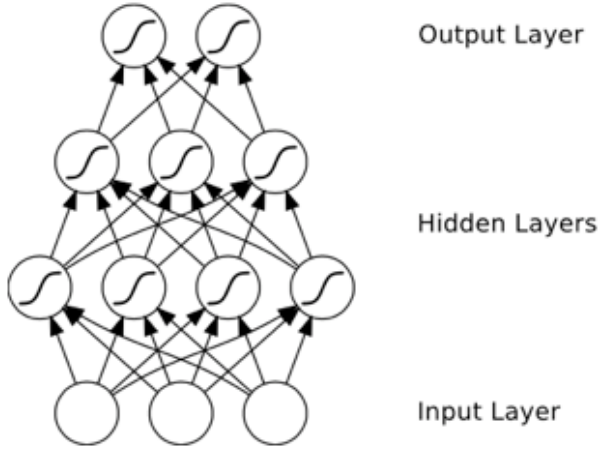
$$h_t^s = o_t \cdot c_t$$

Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

Solution to Problem-2

- Convolutional Neural Network





Max pooling

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

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

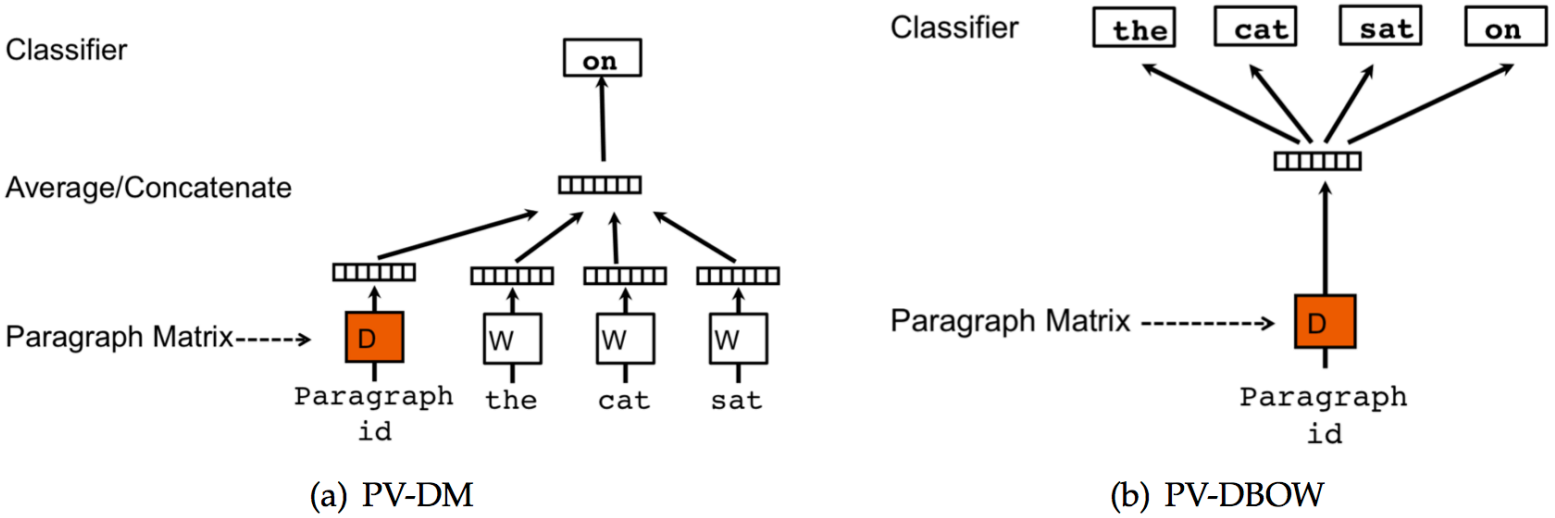
Convolution $c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b).$

Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

Solution to Problem-2

- Text Mining-Representation by Document
  - ✓ Paragraph Vector

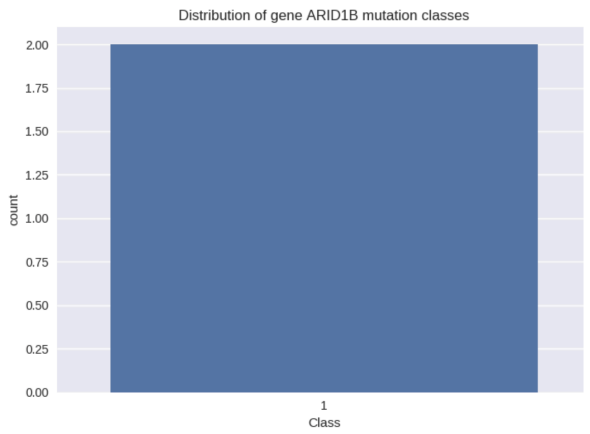
Bad News:
Can only be used to Original-View

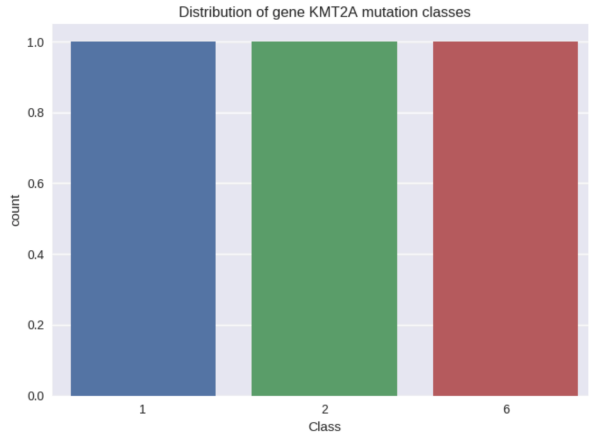


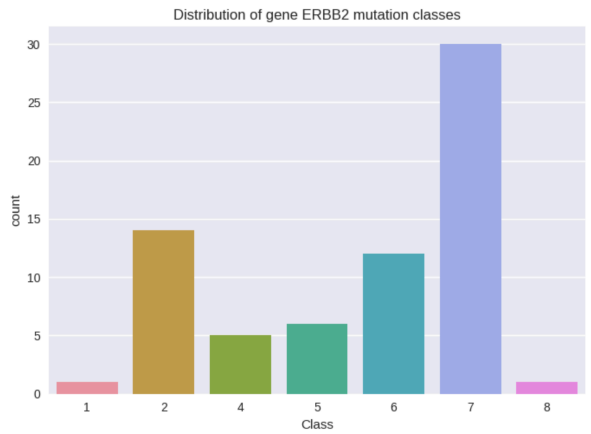
Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* , pages 1188–1196, 2014.

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

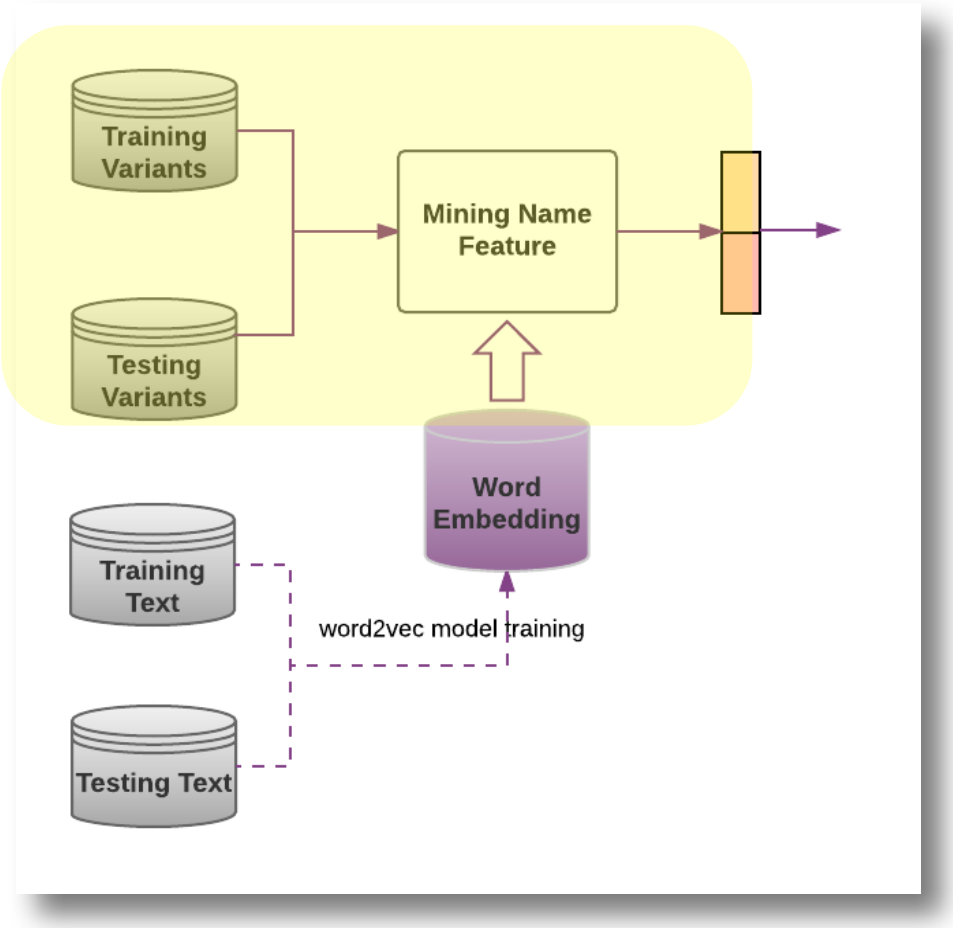
- Name Mining

  ✓ Character Level n-Gram Encoding

  (n=8, 9,473,363 dim， then dimension reduction)

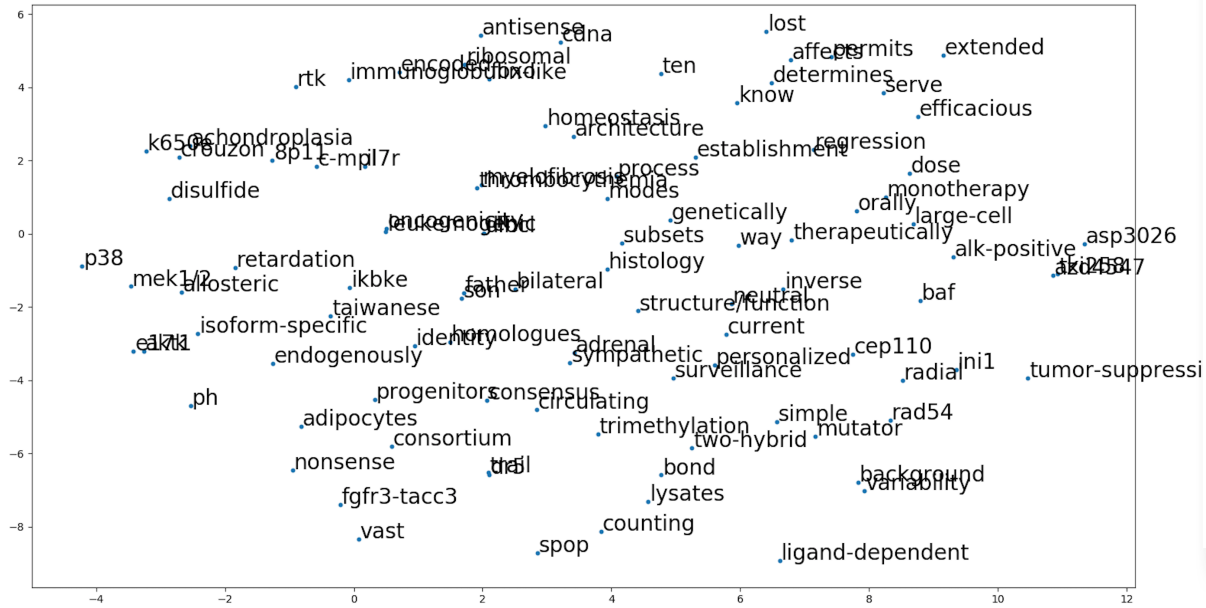
Gene	Mutation
TET2	Y1902A
MTOR	D2512H
KIT	D52N
SPOP	F125V
ETV1	Amplification
……	

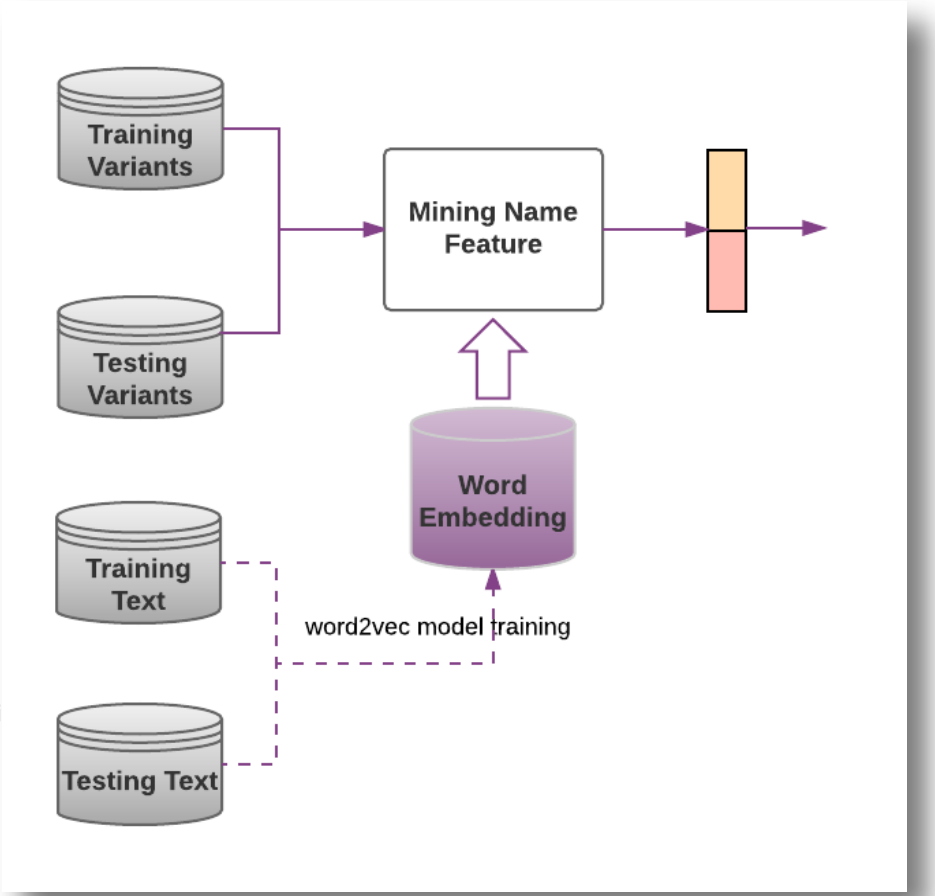
text of <gene, mutation> name



Solution to Problem-3

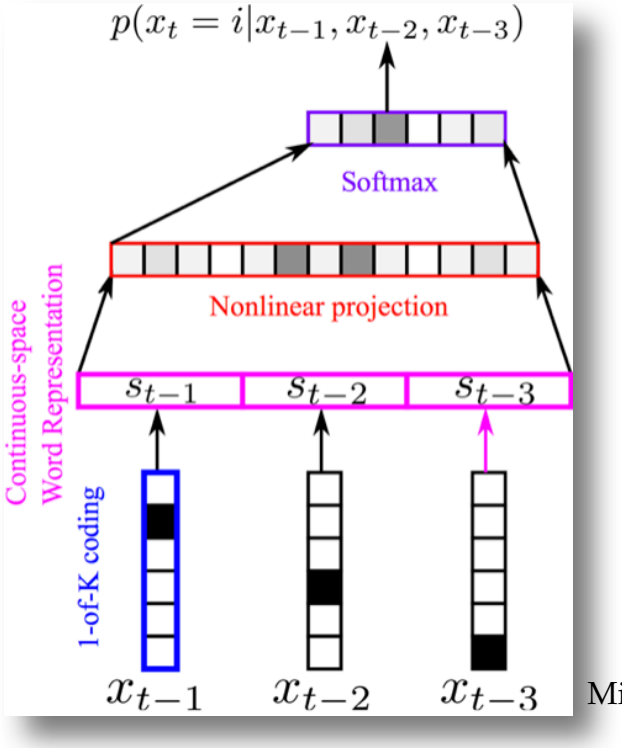
- Name Mining
  - ✓ Word Embedding (word2vec)
  - ✓ Embedding Distribution

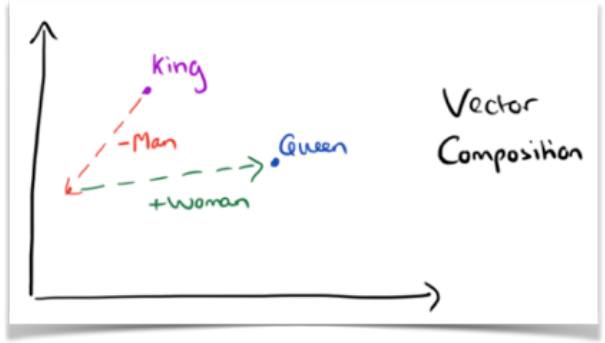




Solution to Problem-3

- Name Mining
  - Word Embedding (word2vec)





$$p(x_t|x_{t-n},\ldots,x_{t-1}) = f_{x_t}(x_{t-n},\ldots,x_{t-1})$$

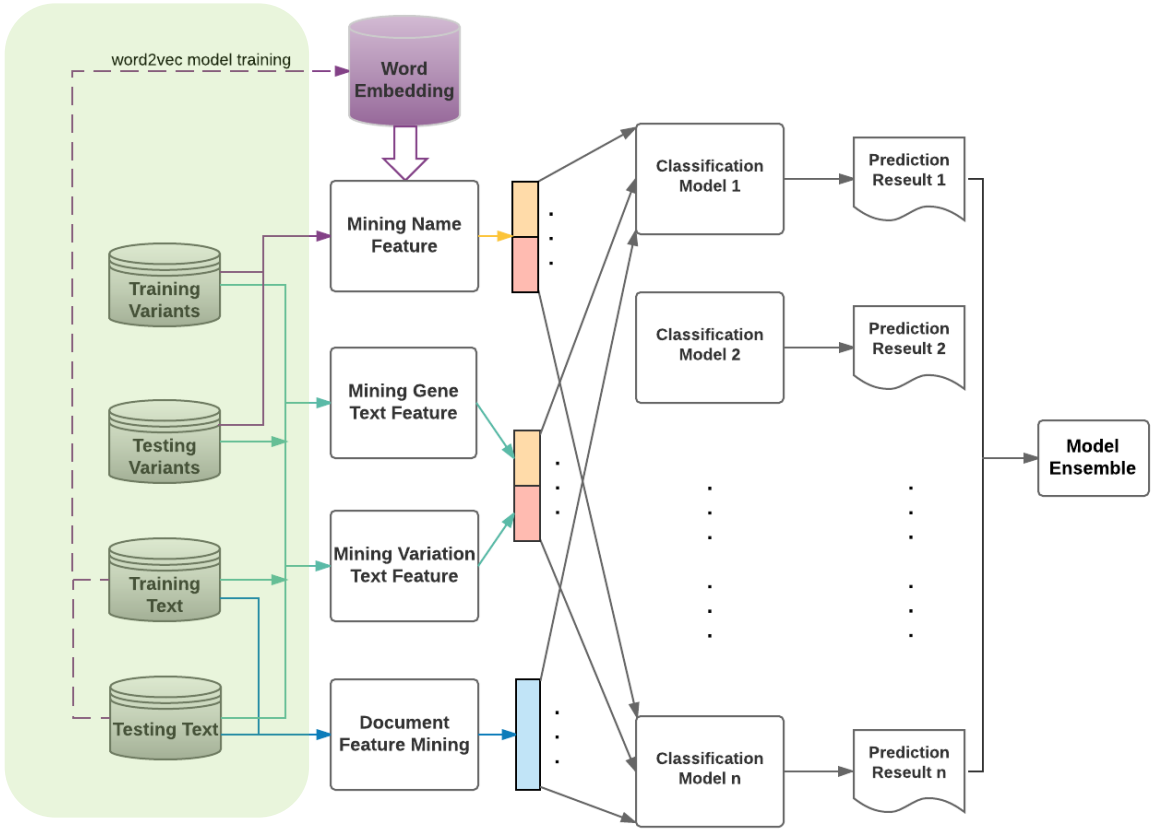
Softmax:

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

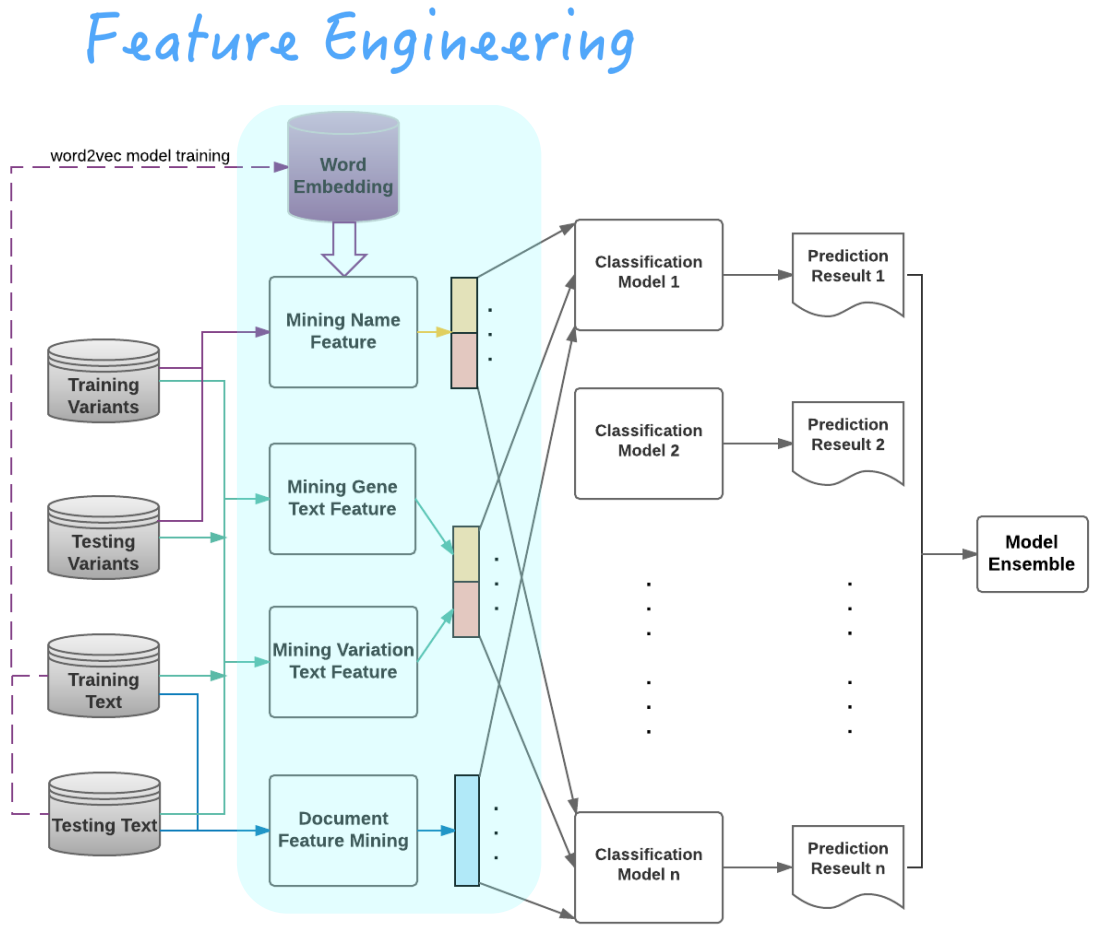
Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

Architecture Overview

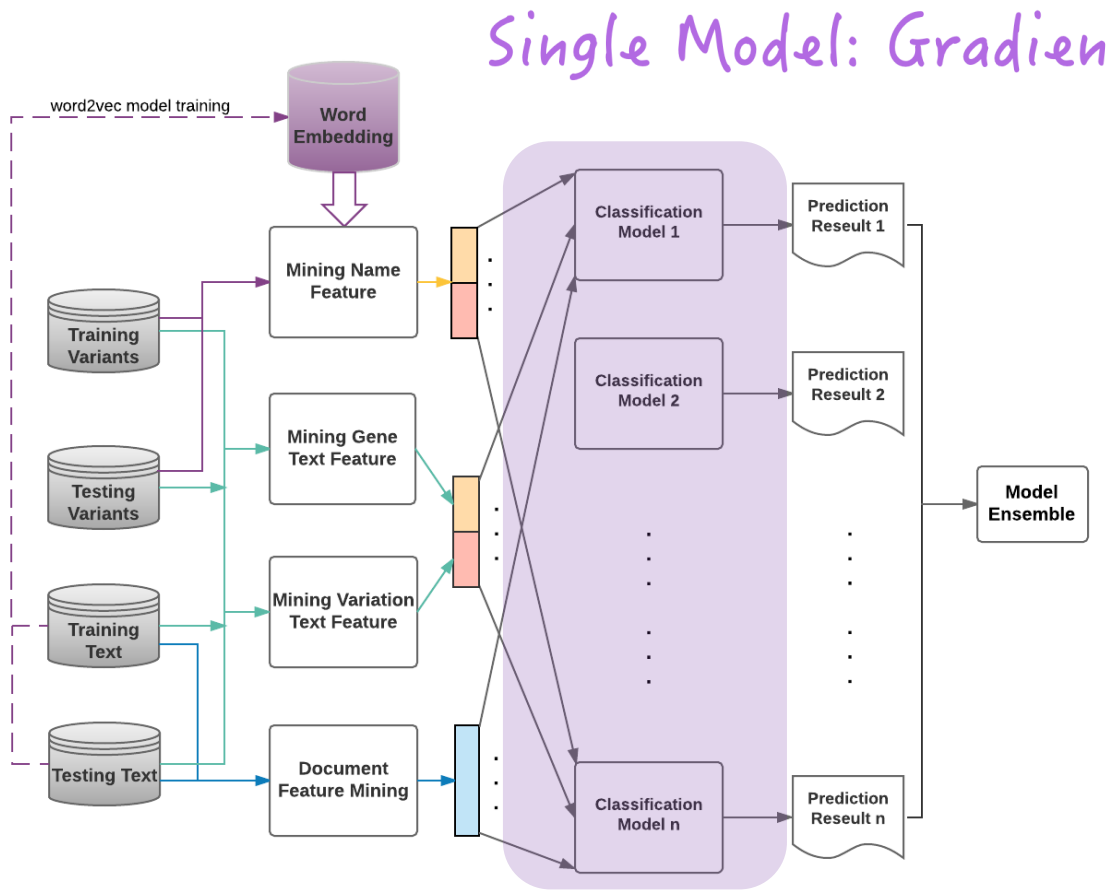
Input Data



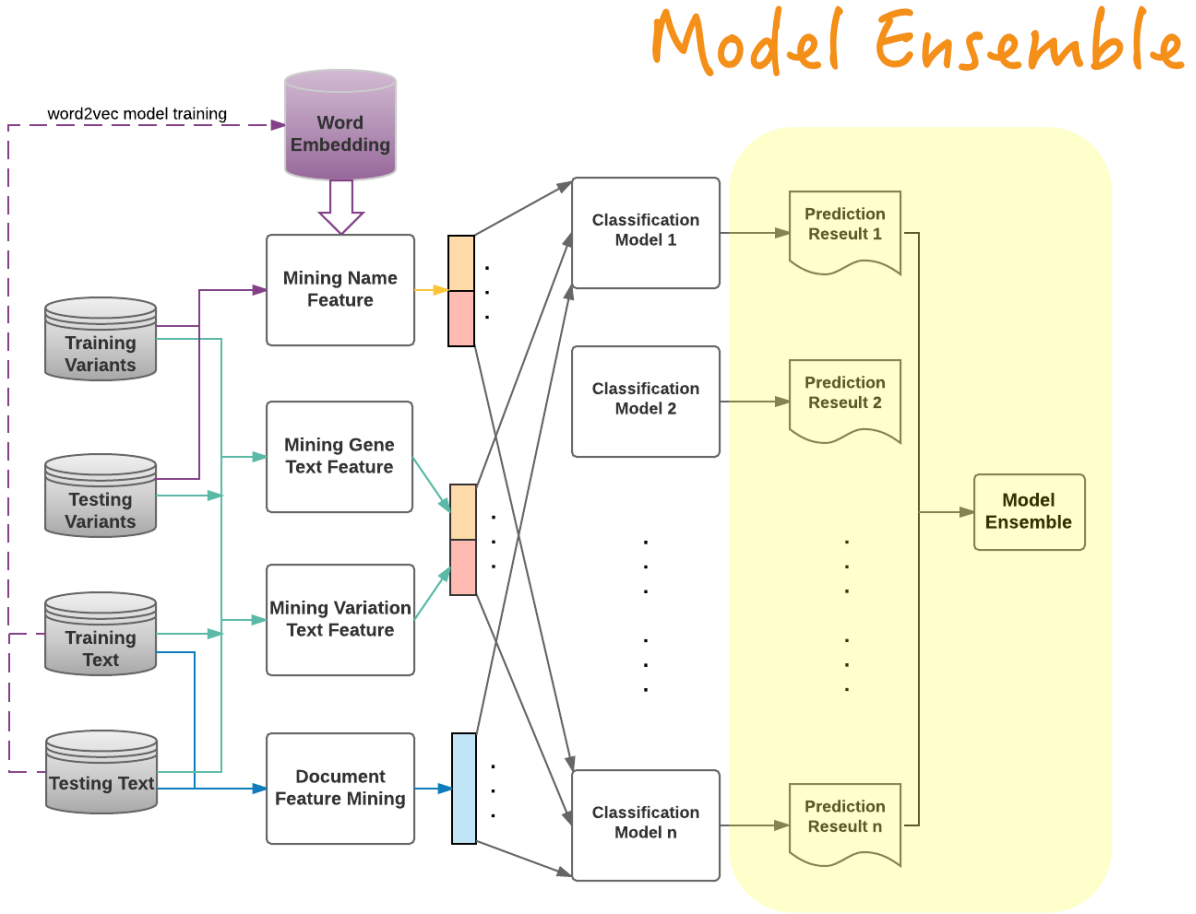
Architecture Overview



Architecture Overview



Architecture Overview



Experimental Settings

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

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

Results of Single Model

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

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

XGBoost

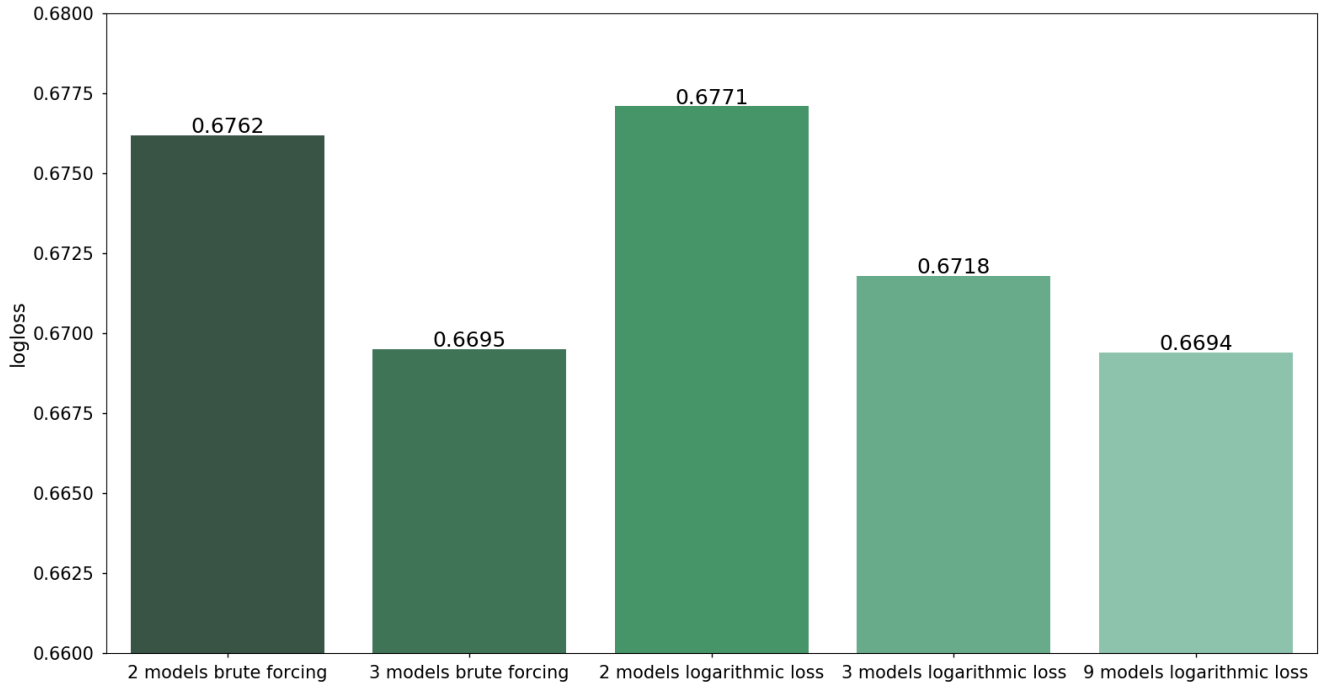
Table 2: Results of GBM model in terms of logloss on 5-fold cross validation and stage1 test set

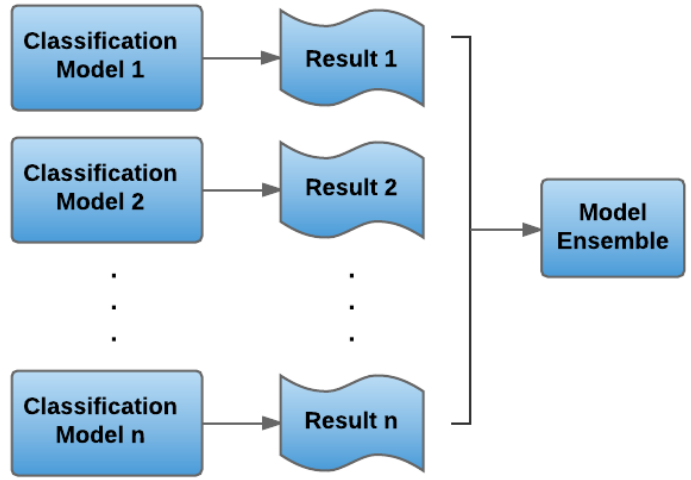
Model Id	5-fold cv	Stage1 test
GBM_1	0.7005	0.6090
GBM_2	0.7121	0.6152
GBM_3	**0.6967**	0.6139
GBM_4	0.7028	0.6178
GBM_5	0.7001	**0.6006**

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

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

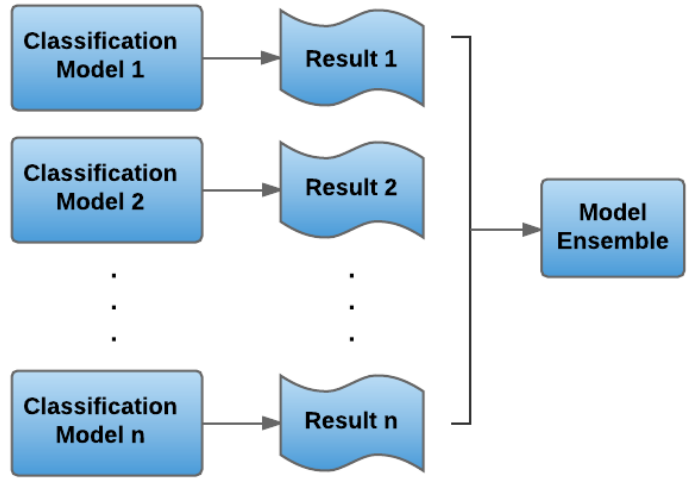
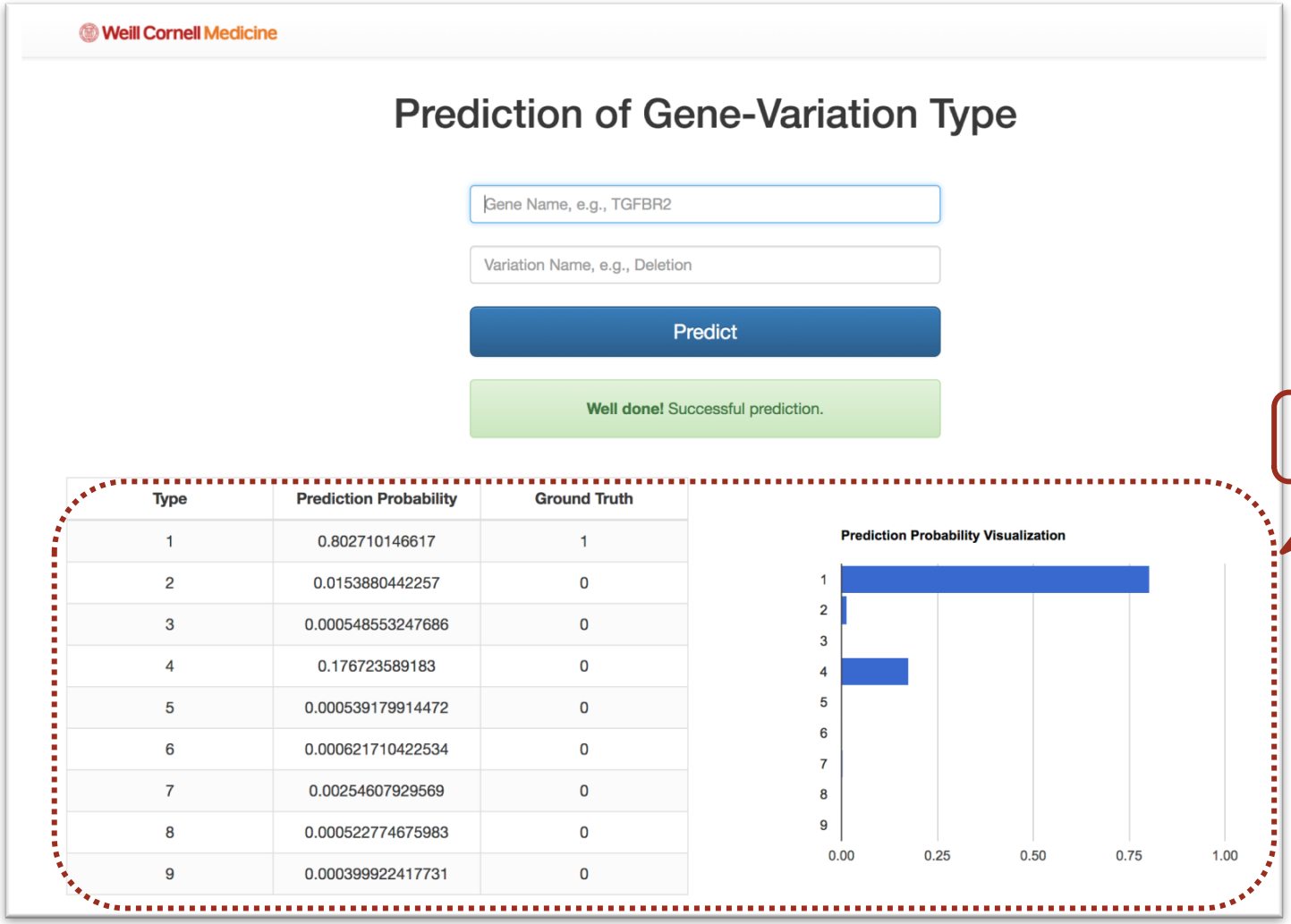


Table 7: Results of the ensemble 9 models by logarithmic loss minimization

Method	5-fold cv
LOGLOSS_MIN	0.6711
LOGLOSS_MIN_CL	0.6695

An Interface



Prediction Results

The demo link:

http://34.207.90.132:8000/demo/

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!