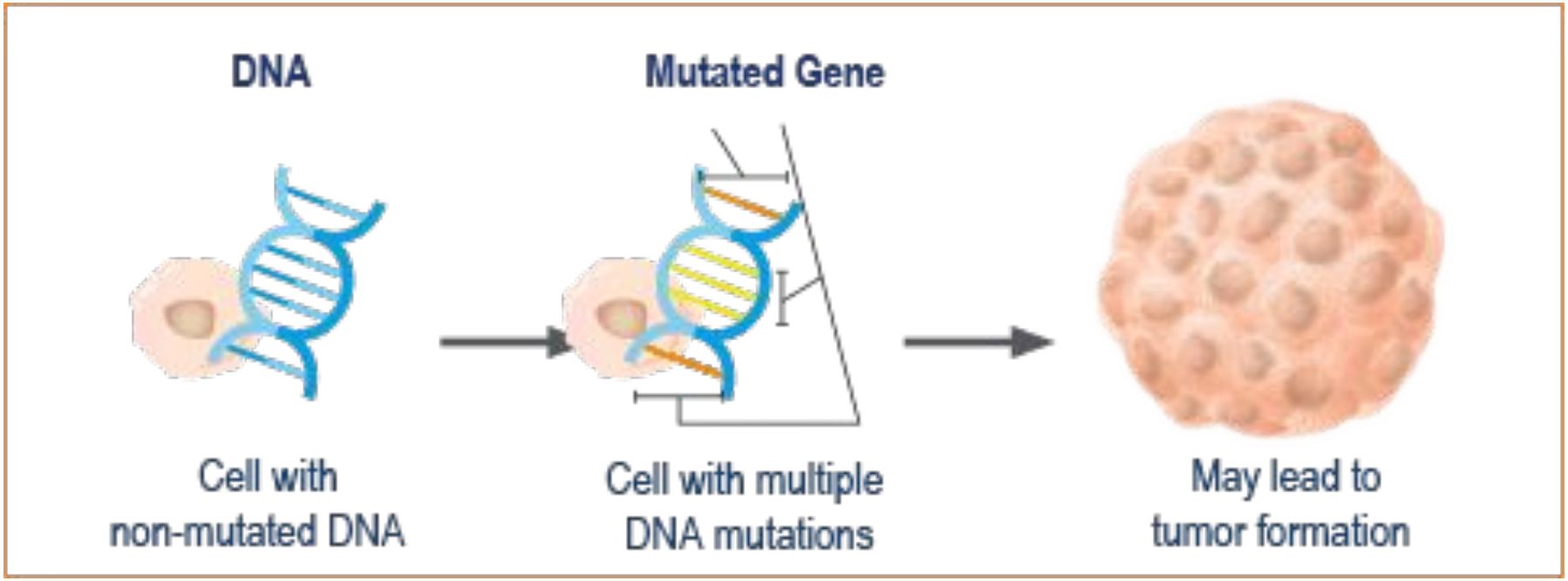
# Genetic Mutation Classification Using Natural Language Processing

Jessica Huang & Brett Gao

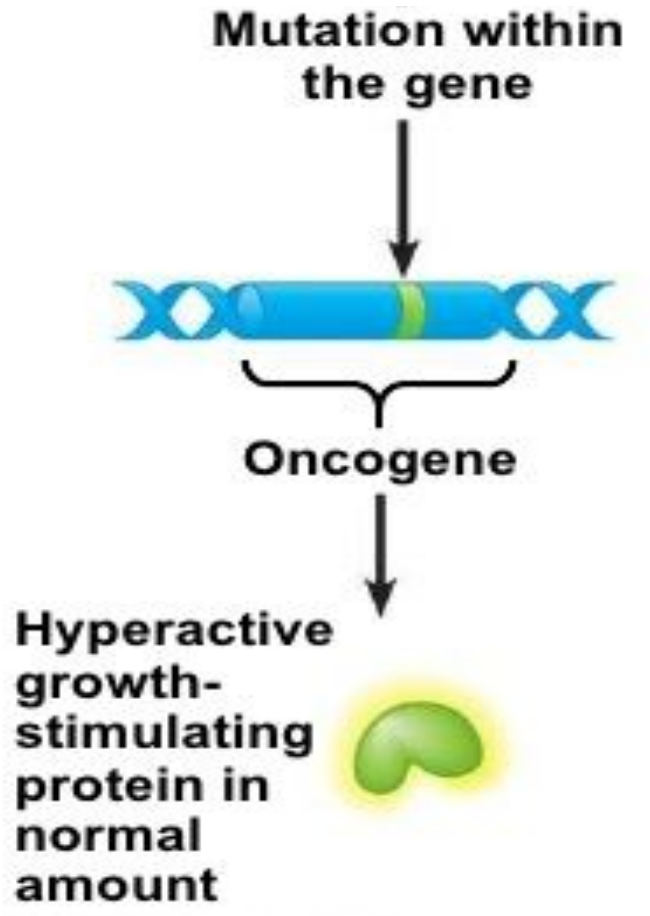
# Cancer is driven by mutations on the genome



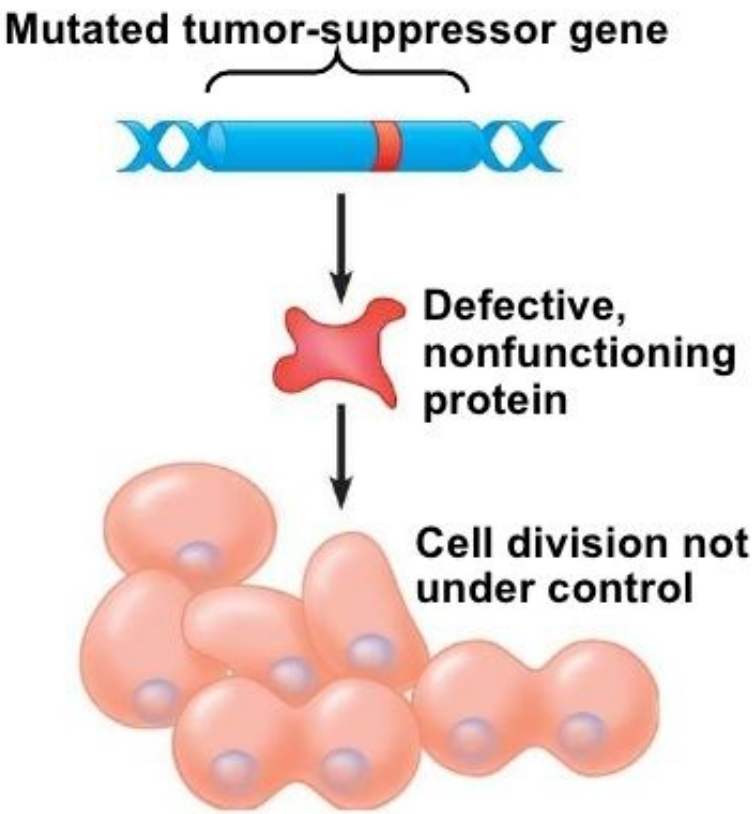

Source: Pearson Education

# Classifying mutations can help cancer treatment

*Gain-of-functions* mutation



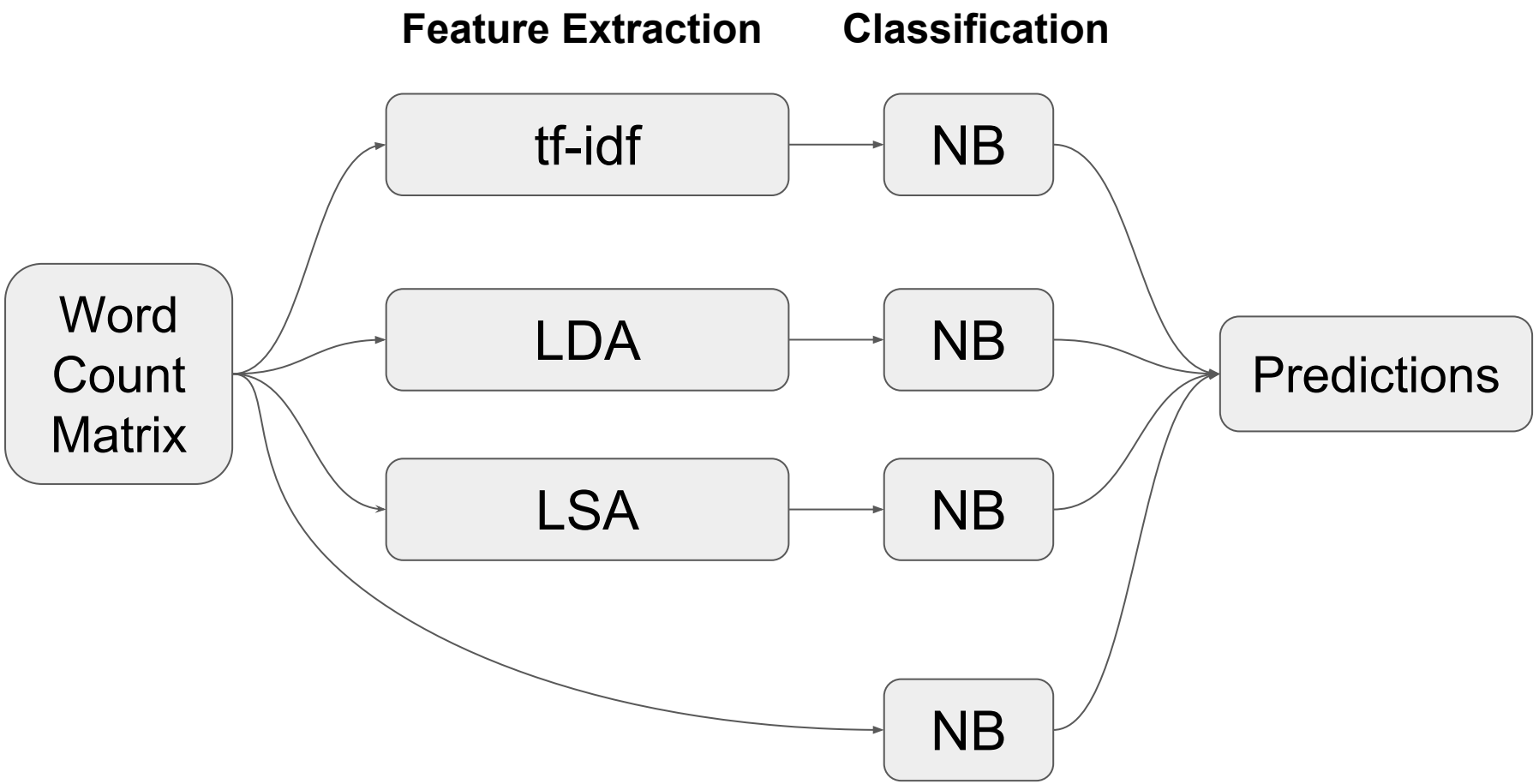

*Loss-of-functions* mutation




Source: Pearson Education

# Classifying mutations is labor intensive

# We can automate mutation classification using machine learning



# Data and preprocessing steps

| ID | Gene | Variation | Class | Text |
|---|---|---|---|---|
| 0 | FAM58A | Truncating Mutations | 1 | Cyclin-dependent kinases (CDKs) regulate a variety of fundamental cellular processes. CDK10 stands out as one of the last orphan CDKs for which no activating cyclin has been identified and no kinase activity revealed. Previous work has shown that CDK10 silencing increases ETS2… |

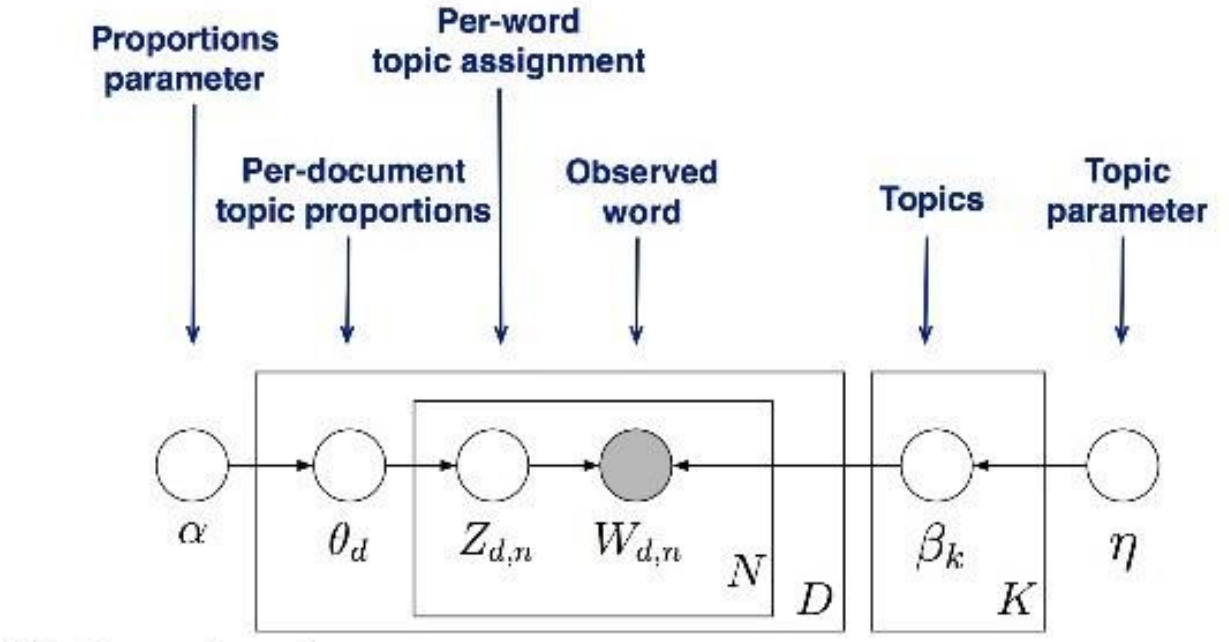
**9 possible class labels:** Gain-of-function, Likely Gain-of-function, Likely Loss-of-function, Loss-of-function, Likely Neutral, Neutral, Likely Switch-of-function, Switch-of-function, and Inconclusive

**Training set size:** 3321
**Test set size:** 986

**Vectorization of Word Counts**
**Removal of stopwords**

Source: Kaggle - *Personalized Medicine: Redefining Cancer Treatment*

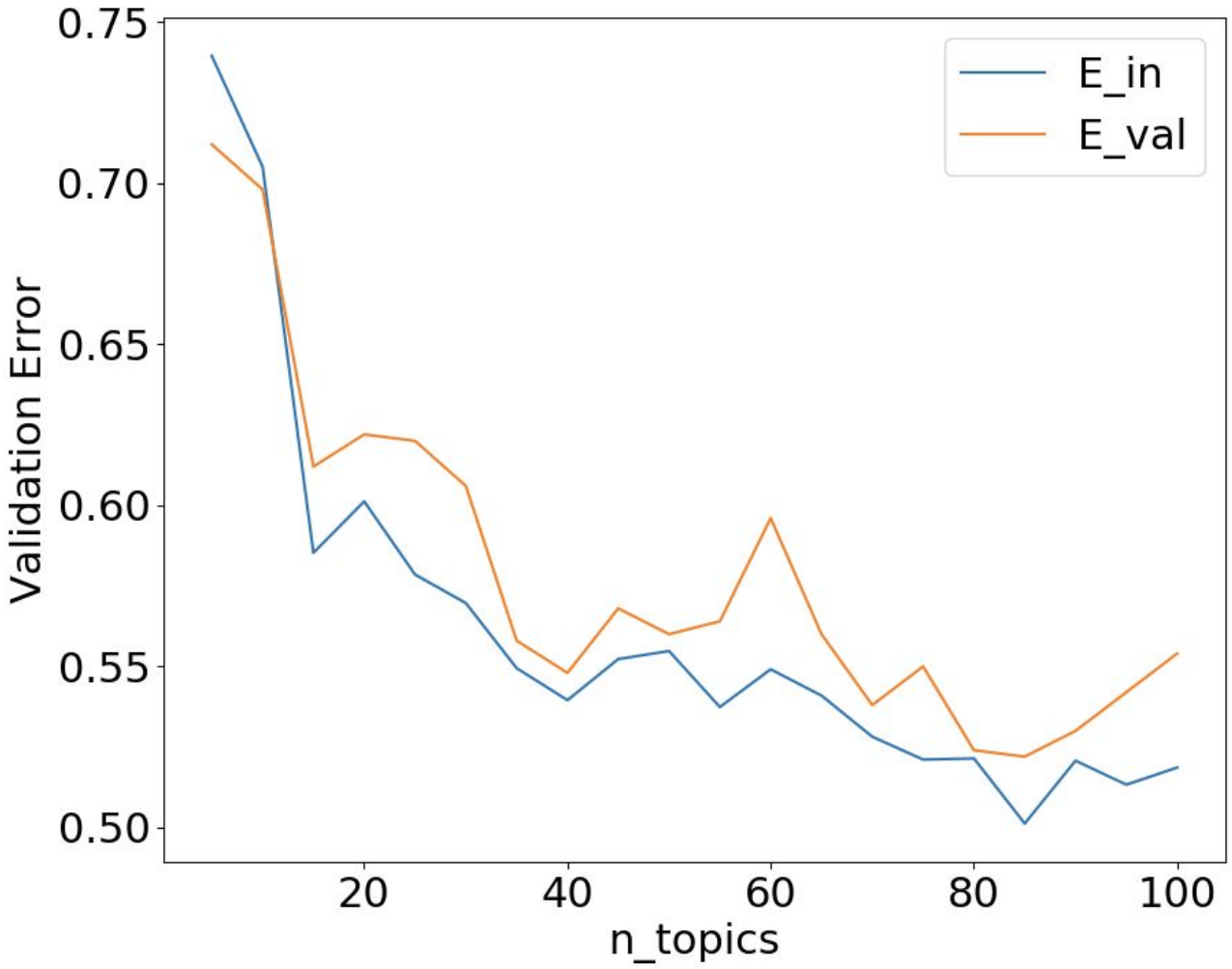
# Feature extraction using Latent Dirichlet Allocation



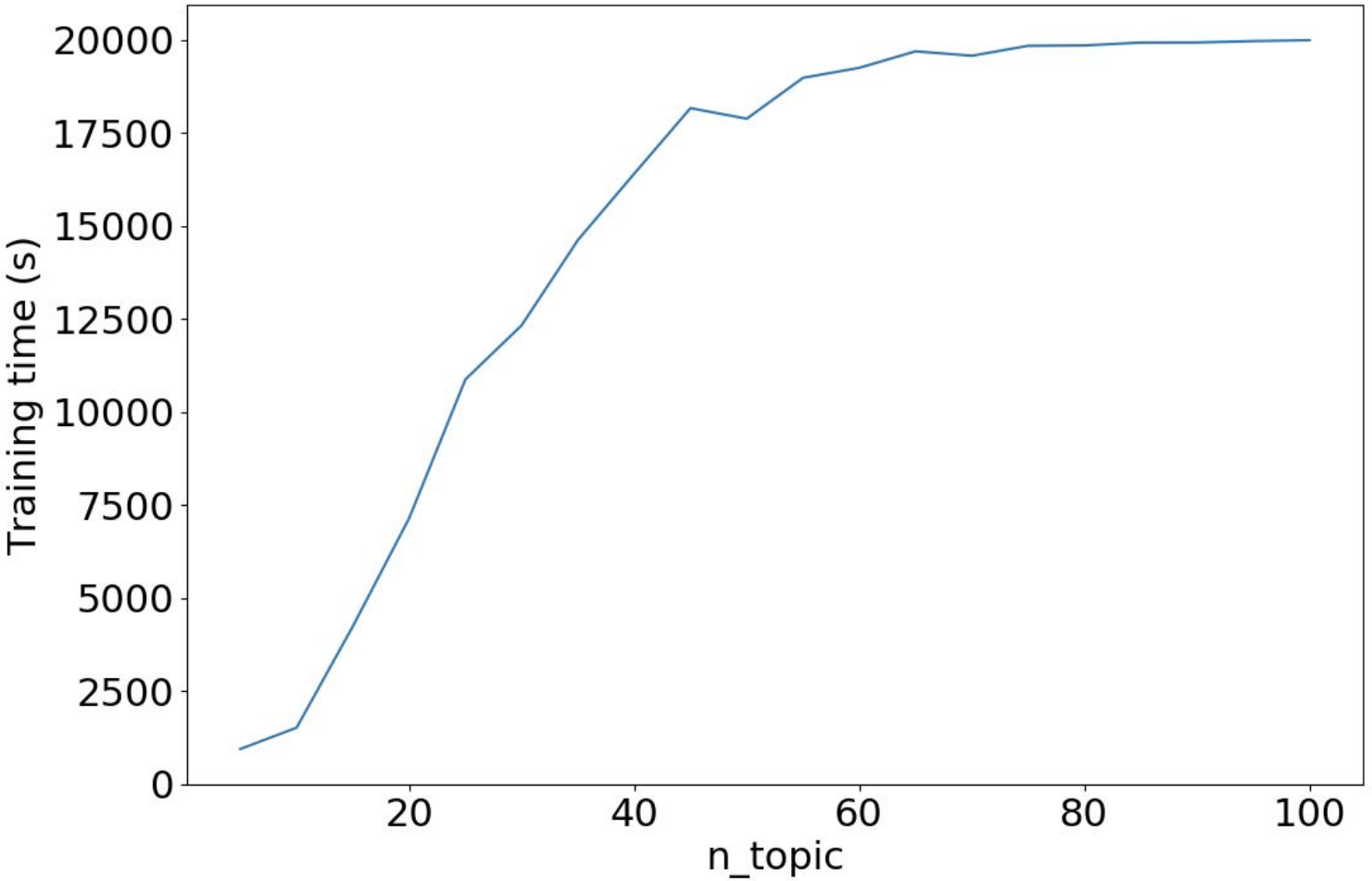

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Source: Blei, ICML 2012 tutorial

# LDA-NB: More topics → Higher accuracy

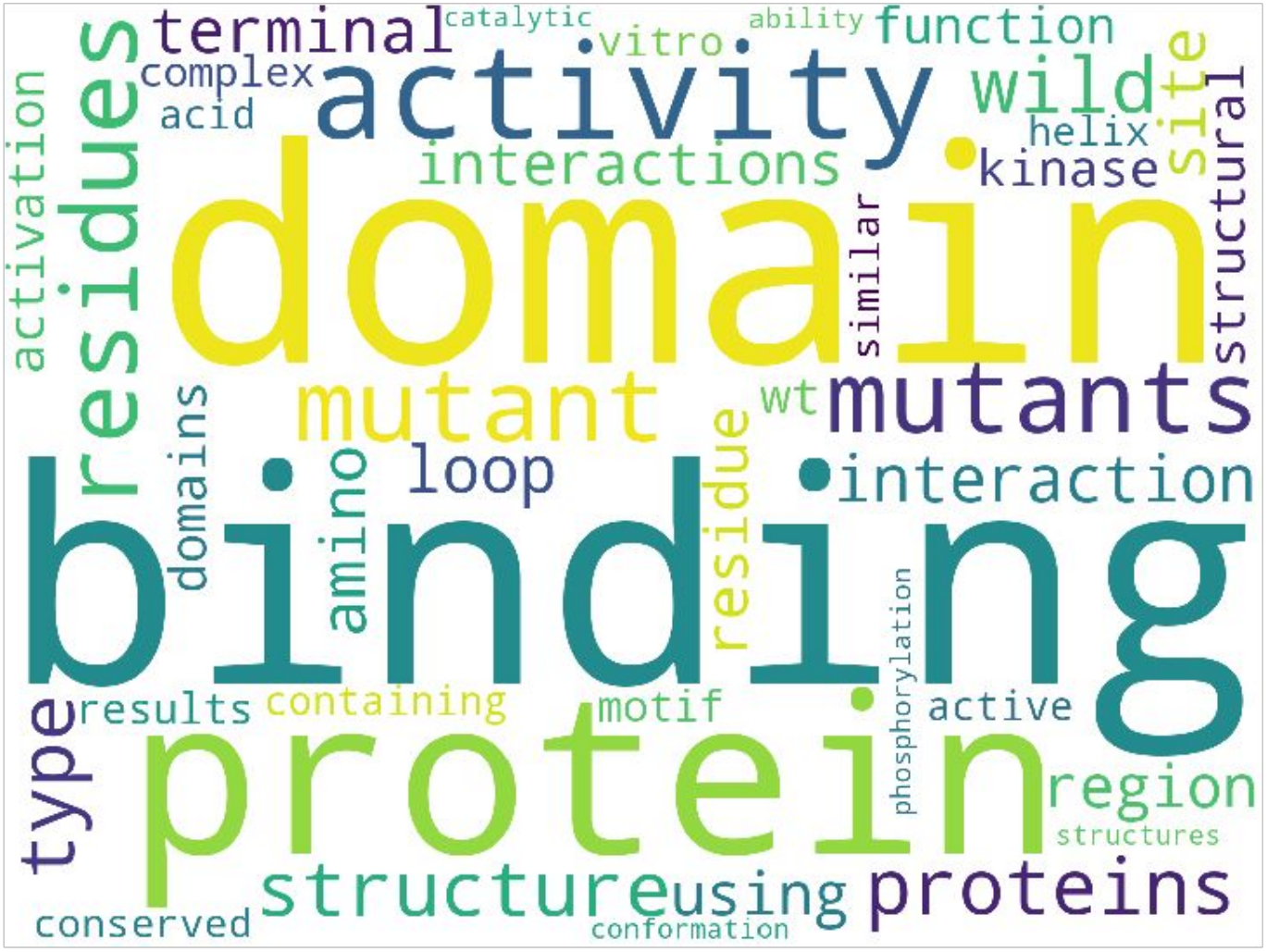


# LDA-NB: More topics → Longer to train



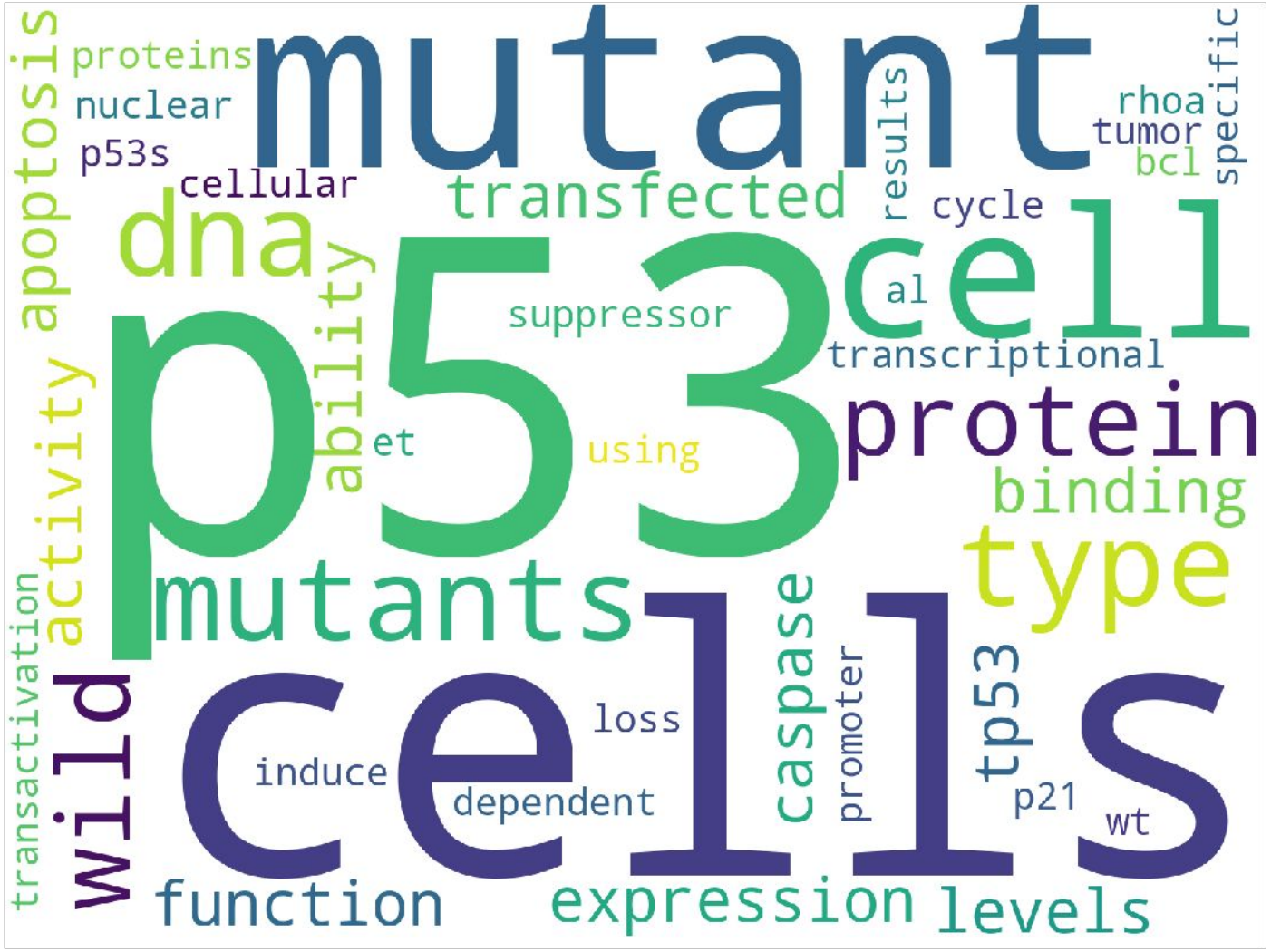
# LDA-extracted topics reflect nature of the classes

Salient topic for *Loss-of-Function* mutation



# LDA-extracted topics reflect nature of the classes

Salient topic for *Loss-of-Function* mutation



# LDA extracted topics reflect nature of the classes

Euclidean distance matrix calculated from class topic frequency

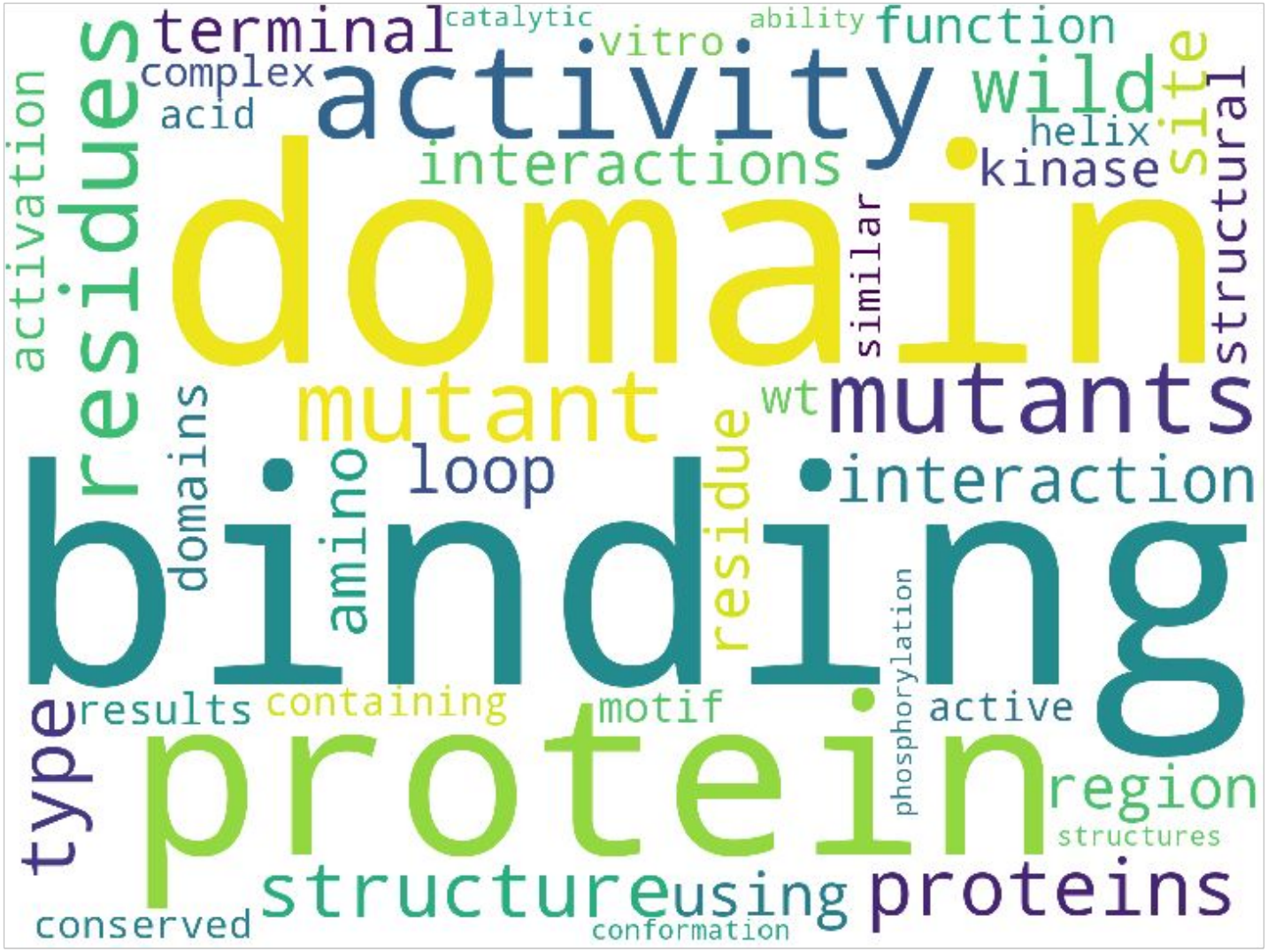
| | Gain-of-function | Loss-of-function |
|---|---|---|
| **Likely Loss-of-function** | 9.692762 | 6.012803 |
| **Likely Gain-of-function** | 5.003767 | 8.686116 |
| **Neutral** | 13.525443 | 11.601596 |
| **Loss-of-function** | 9.770549 | 0.000000 |
| **Likely Neutral** | 11.105728 | 8.943511 |
| **Inconclusive** | 11.232654 | 8.921309 |
| **Gain-of-function** | 0.000000 | 9.770549 |
| **Likely Switch-of-function** | 16.396966 | 14.706482 |
| **Switch-of-function** | 15.383589 | 13.710884 |

# LDA-NB outperforms on test set mutation classification

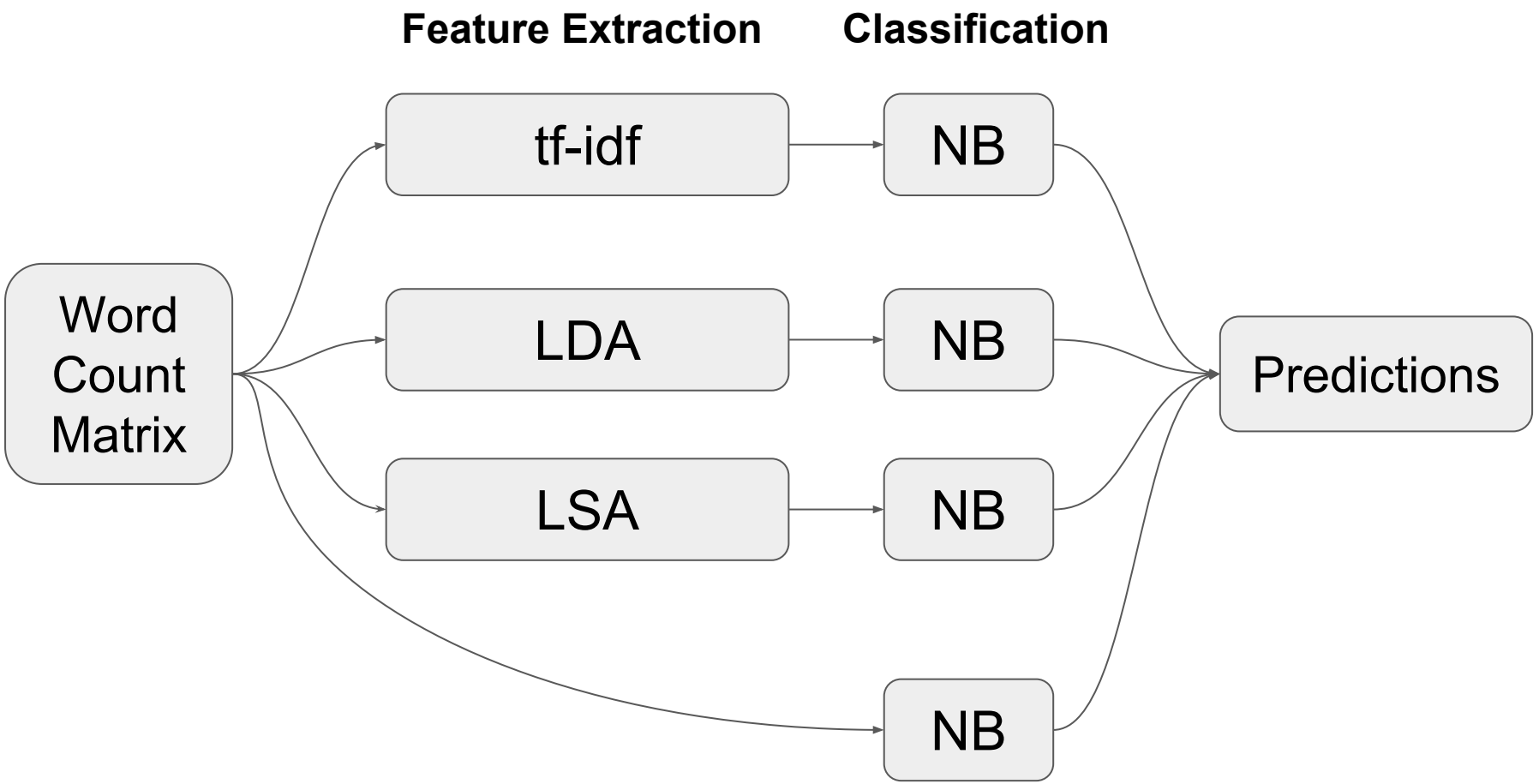
$$-\sum_{c=1}^{M} y_{o,c} \log p_{o,c}$$

| Model | Multi Class Log Loss Score |
|---|---|
| Naive Bayes | 13.15962 |
| tf-idf Naive Bayes | 2.50483 |
| LDA Naive Bayes | 1.63213 |
| LSA Naive Bayes | 9.13054 |

# Questions?


terminal
catalytic
ability
function
vitro
activity
wild
complex
acid
site
helix
interactions
kinase
residues
domain
activation
similar
structural
mutant
wt
mutants
domains
loop
interaction
amino
residue
binding
type
results
containing
motif
active
phosphorylation
protein
region
structures
structure
using
proteins
conserved
conformation

# Questions?


Feature Extraction
Classification
tf-idf
NB
Word Count Matrix
LDA
NB
Predictions
LSA
NB
NB

# Naive Bayes classifier (NB)

**Baye's Theorem**

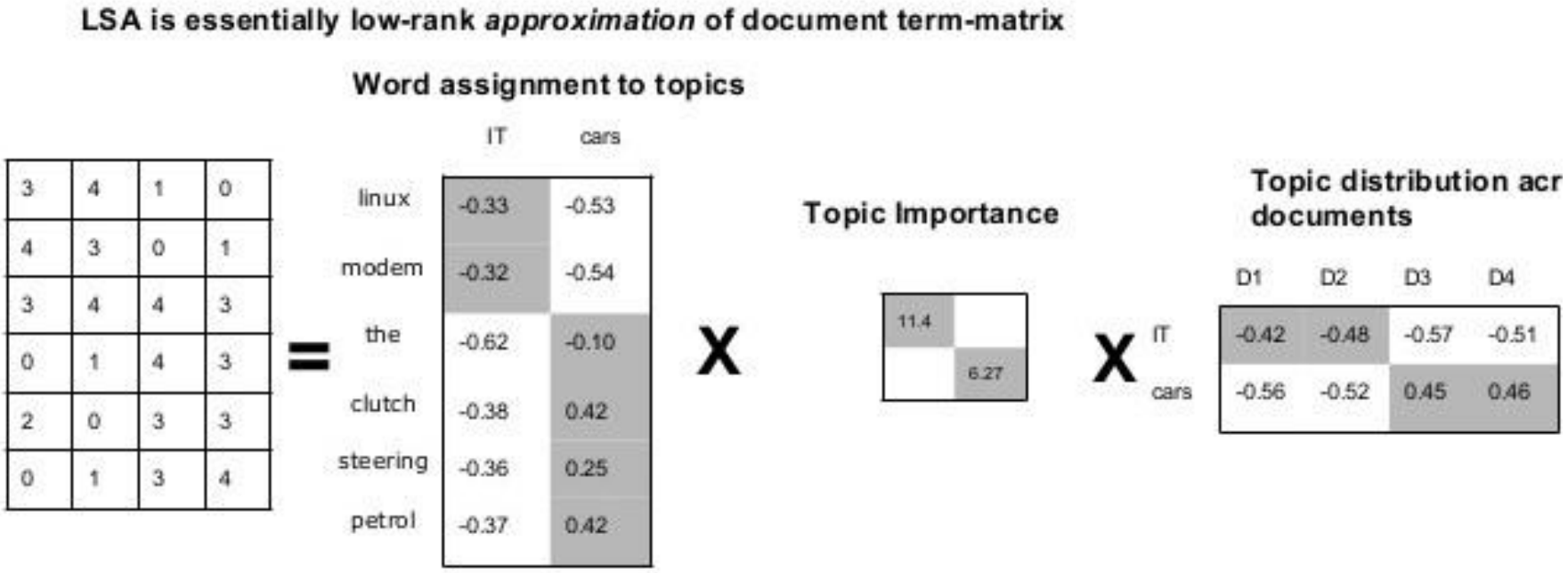
$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

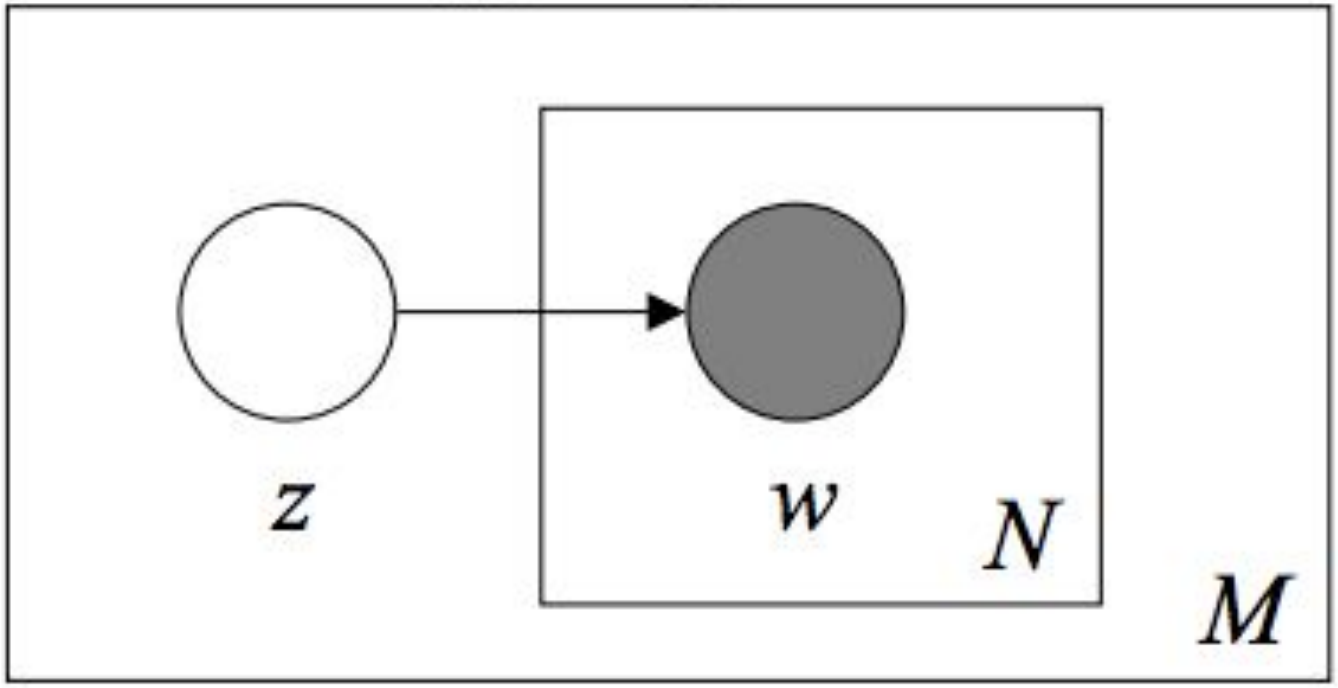
$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

**Multinomial NB**

# Feature extraction using Latent Semantic Analysis



# Naive Bayes classifier (NB)




$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z)$$