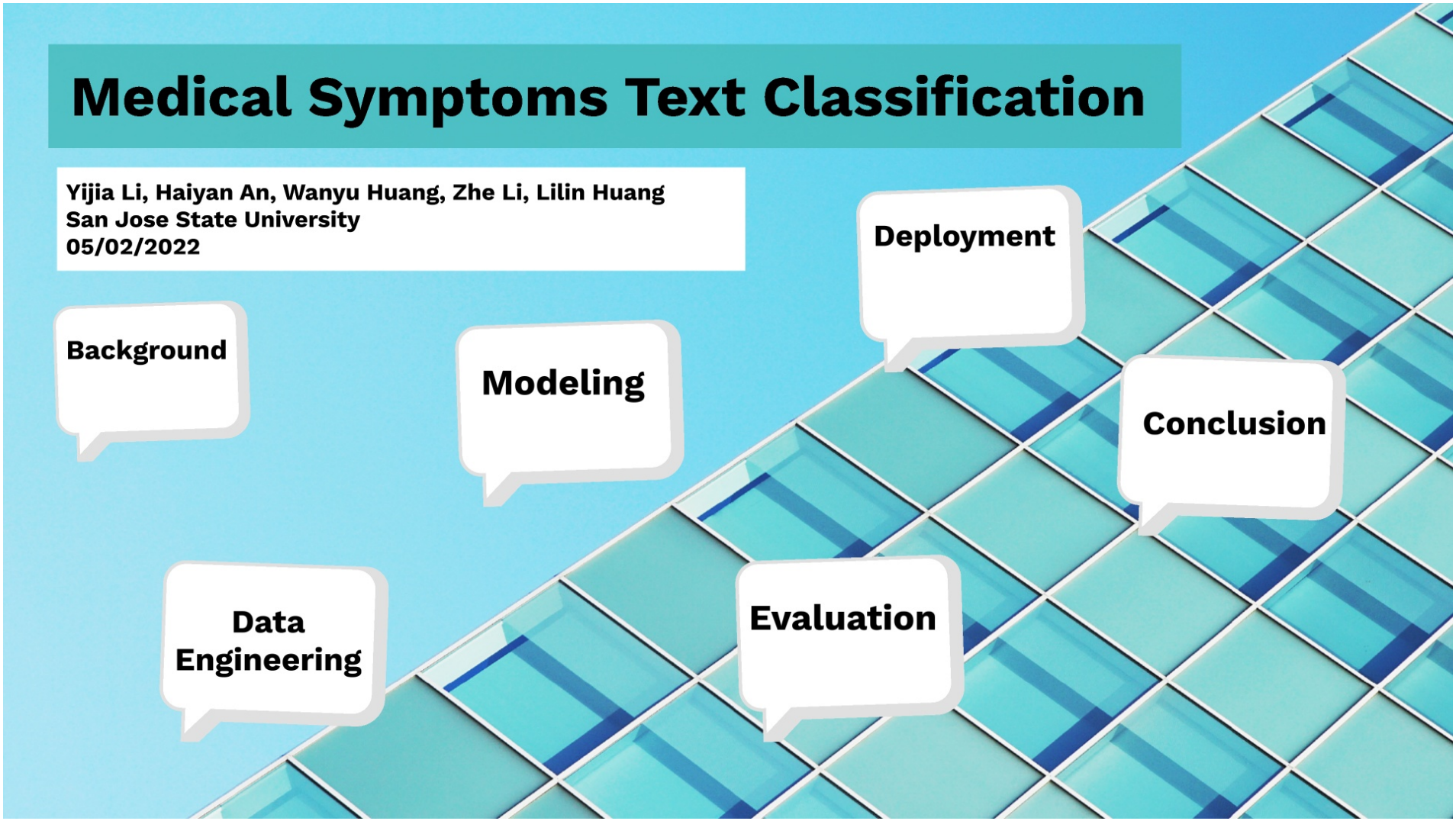
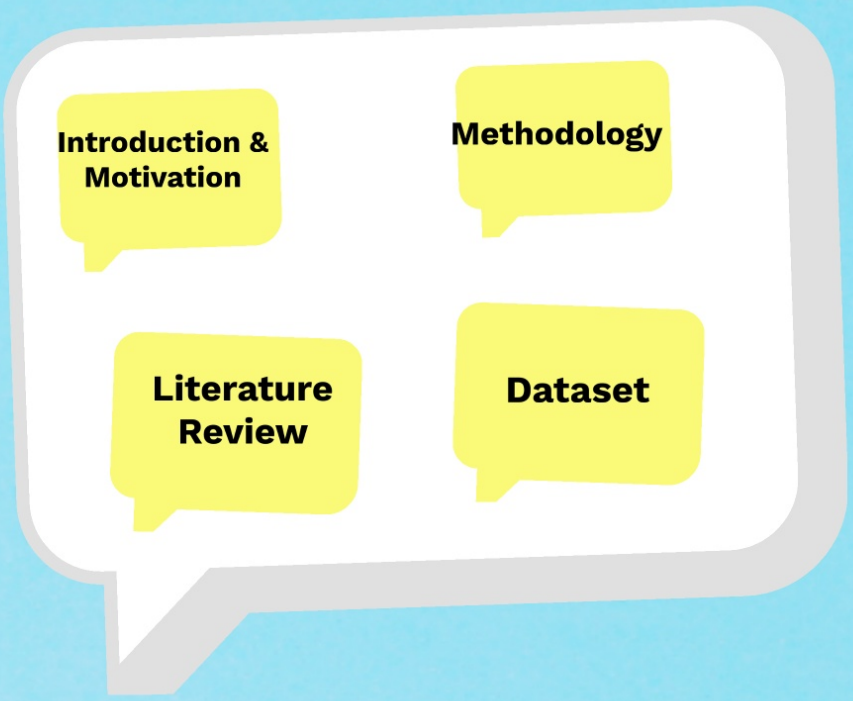

Medical Symptoms Text Classification
Yijia Li, Haiyan An, Wanyu Huang, Zhe Li, Lilin Huang
San Jose State University
05/02/2022
Background
Modeling
Deployment
Conclusion
Data Engineering
Evaluation

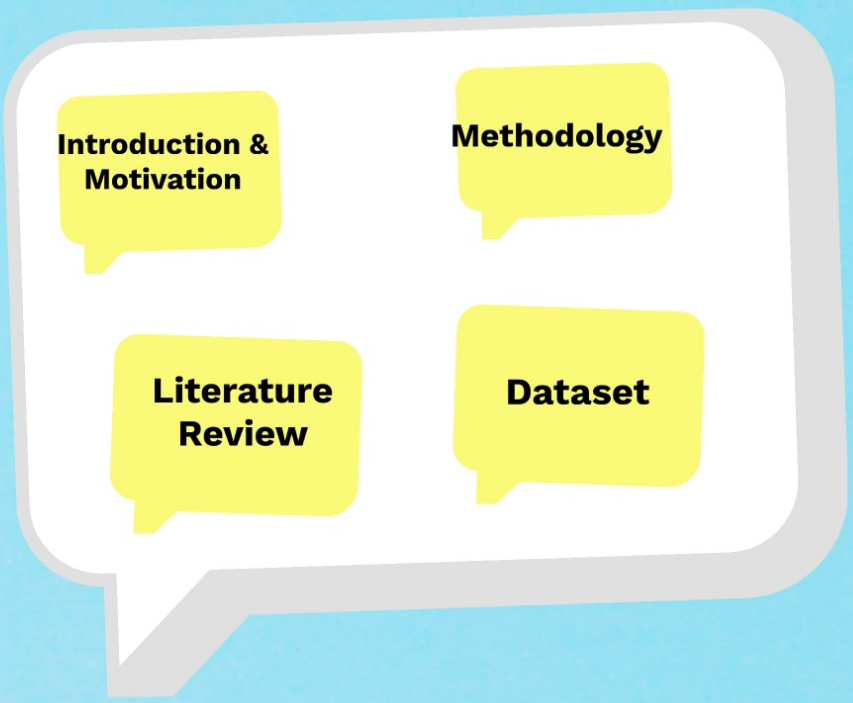
Background


Introduction & Motivation
Methodology
Literature Review
Dataset

Introduction & Motivation

- More people seeking online medical service during the Covid-19 pandemic
- Natural language processing and machine learning techniques
- Our project aims to help with the automation of the initial classification of medical symptoms, which will greatly save time in matching patients with the correct doctors.
- This is a text classification project that finds the best combination of textual feature extraction methods and multi-class classification algorithms that can be used to classify the ailments based on some phrases the patients entered.

Background


Introduction & Motivation
Methodology
Literature Review
Dataset

Literature Review

Existing works in the scope of healthcare

- Chaitrashree, K.M. and his colleagues developed an app called MedAid for those who want to self-diagnose their illness. They trained the model using 4 different machine learning algorithms with bag-of-words extracted features.
- Pendyala, V. S. and his colleagues confirmed the feasibility of the text-mining approach to the medical diagnosis problem and pointed out that the future of this sort of study should be using the Hadoop ecosystem to improvise the results

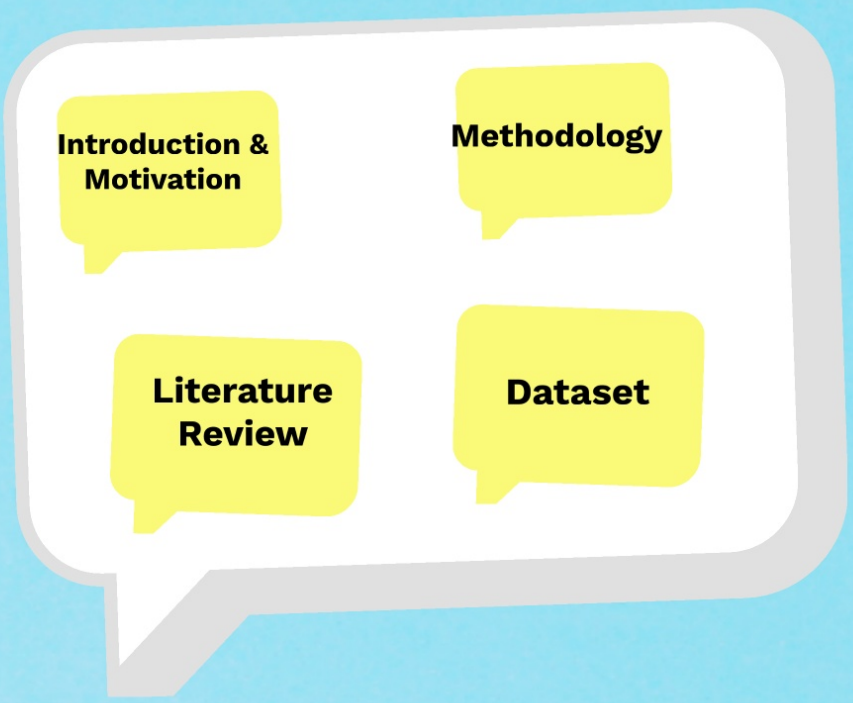
Overview elements in text classification

- Mirończuk, M. M. claims the framework of text classification can be divided into data acquisition, labeling, feature construction, dimensionality reduction, and last and most important, training of classifiers

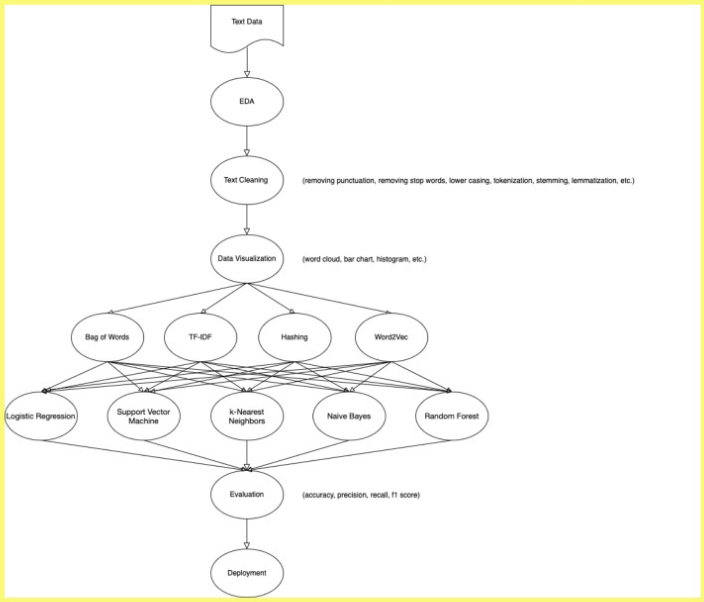
Performance among different classification

- Khursheed, M. S. used an automated process to examine the accuracy, precision, recall, and f-measure in all combinations of feature selection, term weighting, and classification algorithms on the large and diverse Arabic text dataset

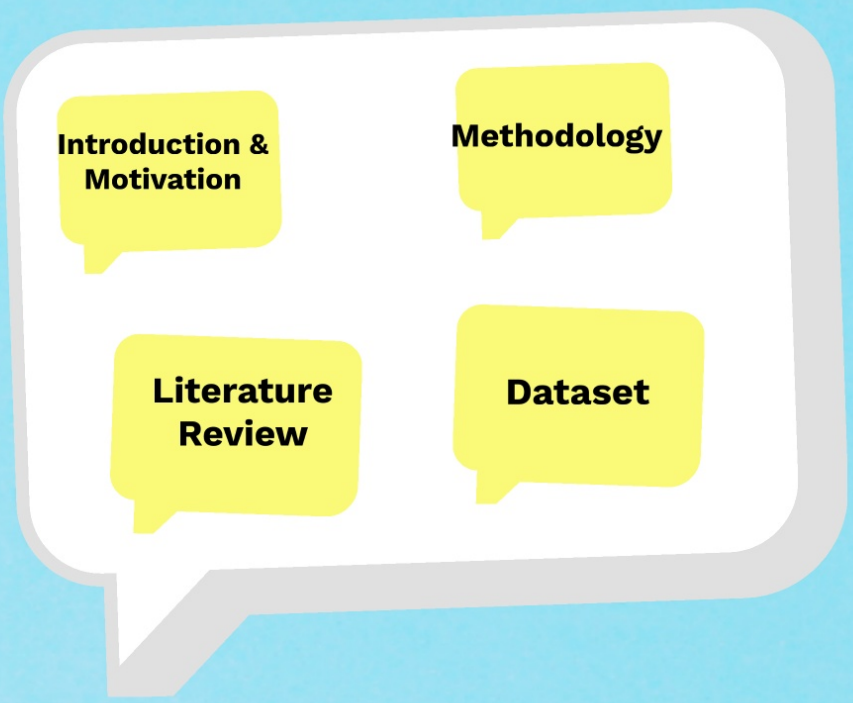
Background


Introduction & Motivation
Methodology
Literature Review
Dataset

Methodology



Background

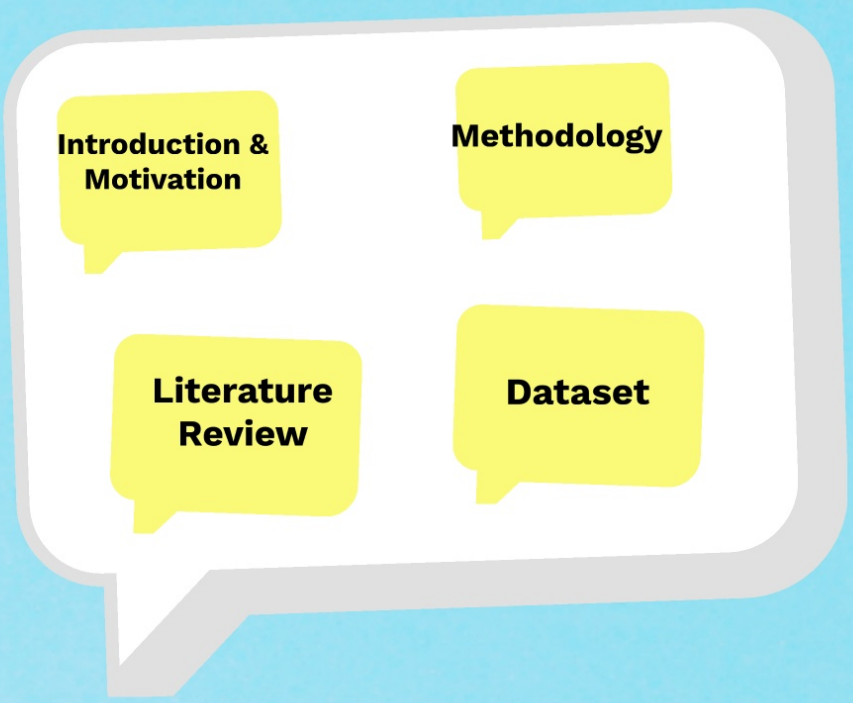

Introduction & Motivation
Methodology
Literature Review
Dataset

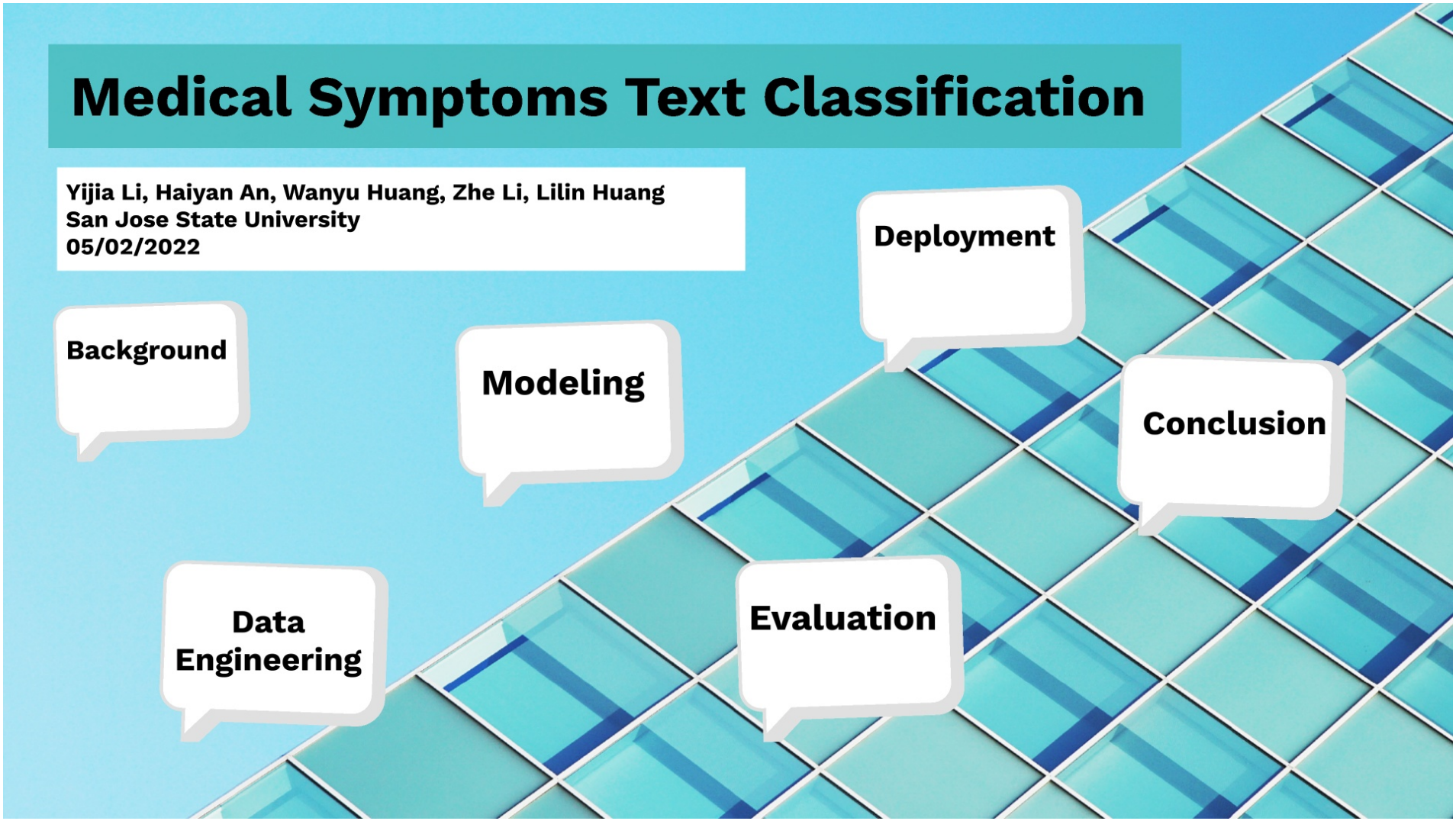
Dataset

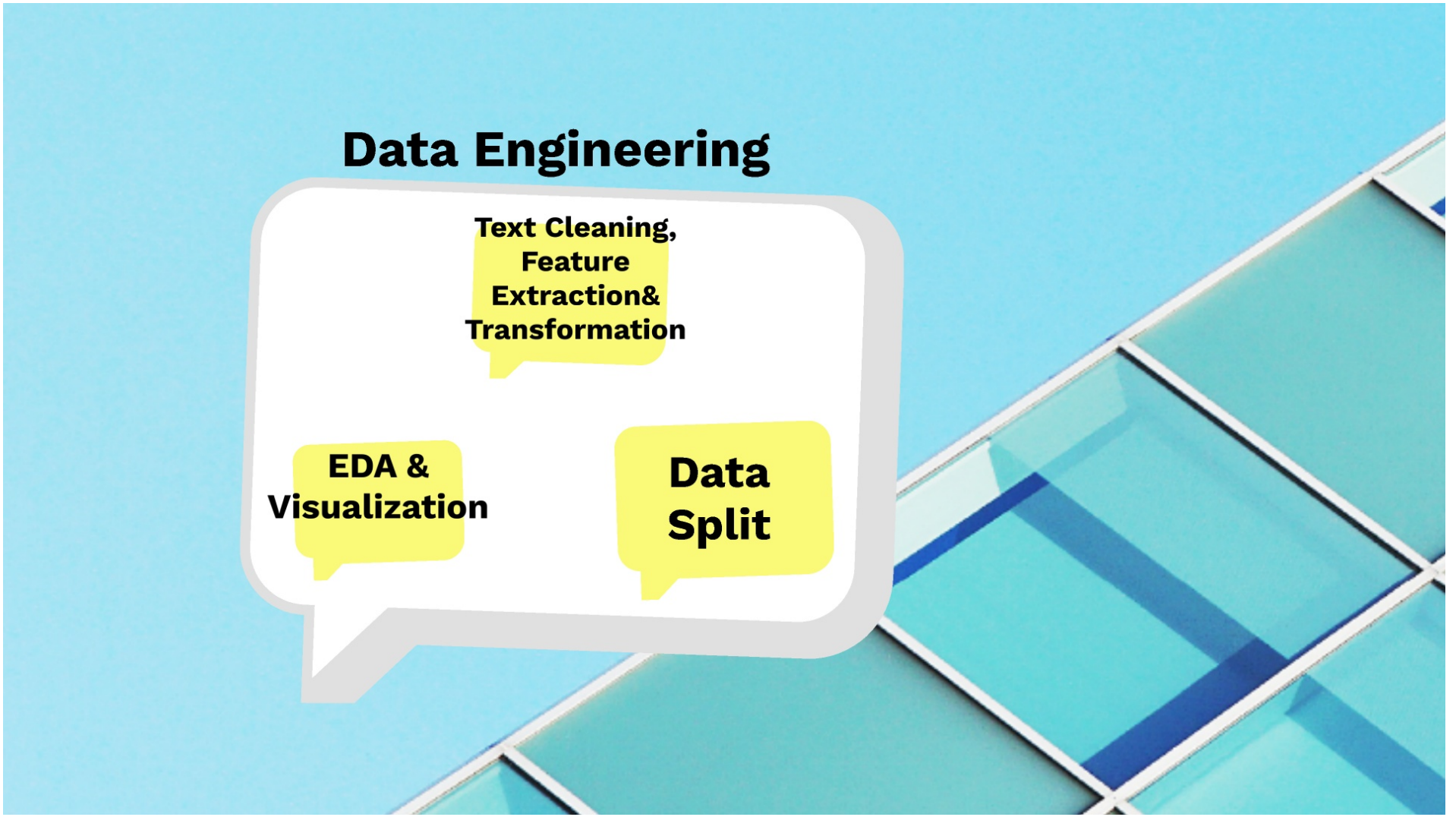
6,661 entries, 13 columns

25 different ailments: Acne, Back pain, Blurry vision, Body feels weak, Cough, Earache, Emotional pain, Feeling cold, Feeling dizzy, Foot ache, Hair falling out, Hard to breathe, Headache, Heart hurts, Infected wound, Injury from sports, Internal pain, Joint pain, Knee pain, Muscle pain, Neck pain, Open wound, Shoulder pain, Skin issue, Stomach ache.

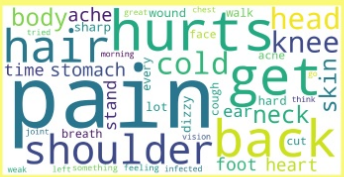
Background

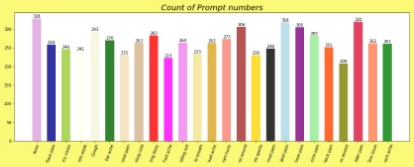

Introduction & Motivation
Methodology
Literature Review
Dataset

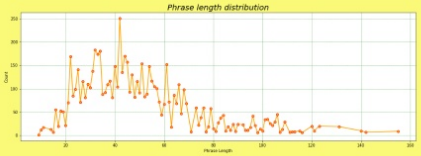

Medical Symptoms Text Classification
Yijia Li, Haiyan An, Wanyu Huang, Zhe Li, Lilin Huang
San Jose State University
05/02/2022
Background
Modeling
Deployment
Conclusion
Data Engineering
Evaluation

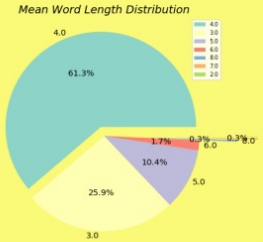

Data Engineering
Text Cleaning, Feature Extraction& Transformation
EDA & Visualization
Data Split

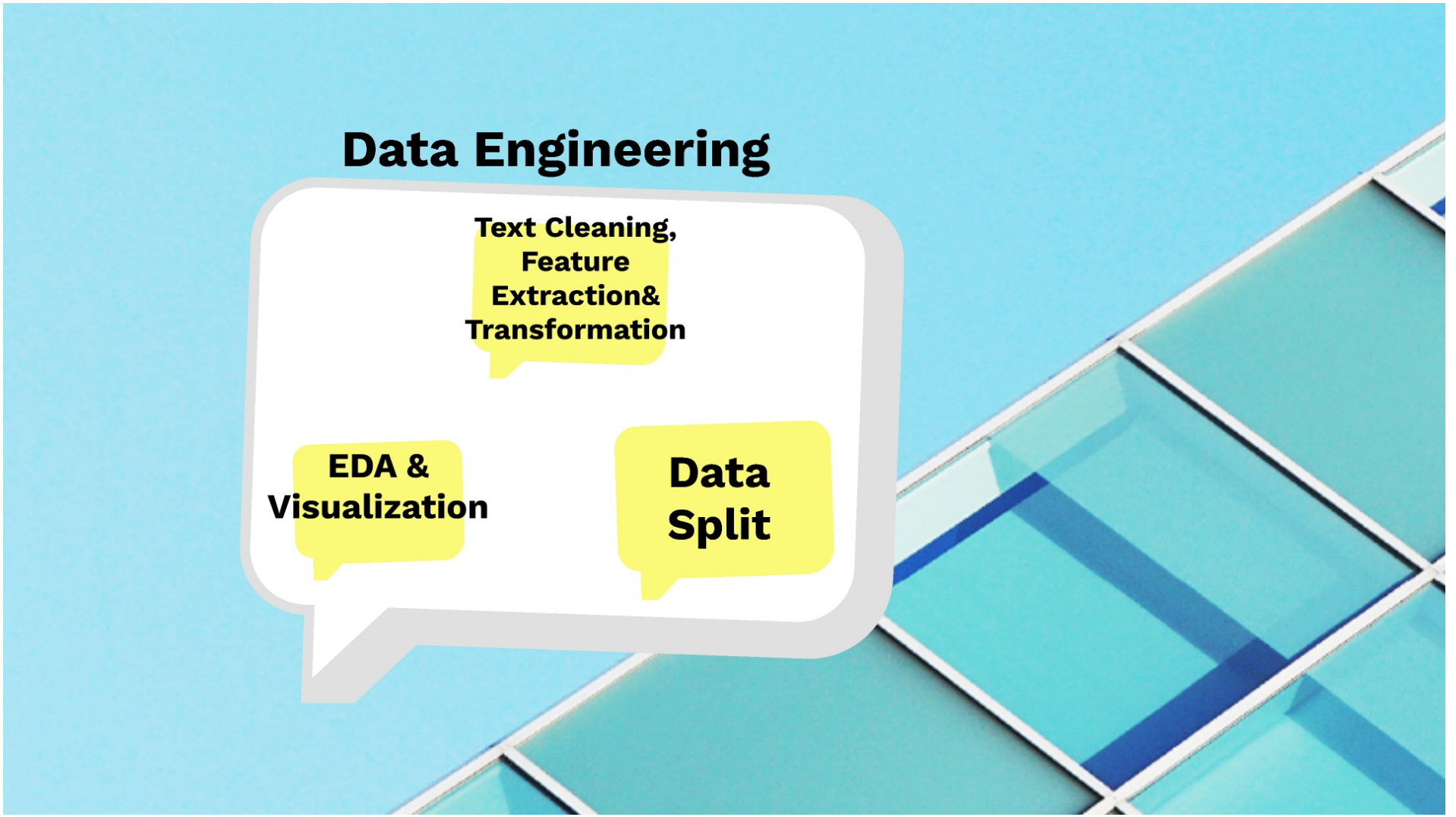
EDA & Visualization





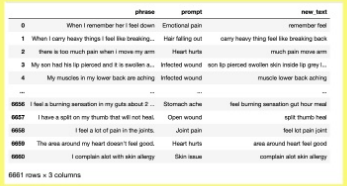





Data Engineering
Text Cleaning, Feature Extraction& Transformation
EDA & Visualization
Data Split

Text Cleaning, Feature Extraction, Feature Transformation

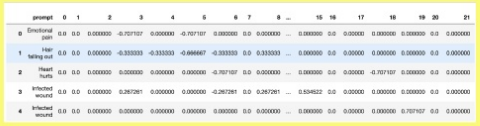
Text Cleaning: tokenization, lowercasing, removing punctuation and stop words, and lemmatization



	phrase	prompt	new_text
0	When I remember her I feel down.	Emotional pain	remember feel
1	When I carry heavy things I feel like breaking,...	Hair falling out	carry heavy thing feel like breaking back
2	there is too much pain when i move my arm	Heart hurts	much pain move arm
3	My son had his lip pierced and it is swollen a...	Infected wound	son lip pierced swollen skin inside lip grey l...
4	My muscles in my lower back are aching	Infected wound	muscle lower back aching
...	...	...	...
6656	I feel a burning sensation in my guts about 2 ...	Stomach ache	feel burning sensation gut hour meal
6657	I have a split on my thumb that will not heal.	Open wound	split thumb heal
6658	I feel a lot of pain in the joints.	Joint pain	feel lot pain joint
6659	The area around my heart doesn't feel good.	Heart hurts	area around heart feel good
6660	I complain alot with skin allergy	Skin issue	complain alot skin allergy

6661 rows × 3 columns

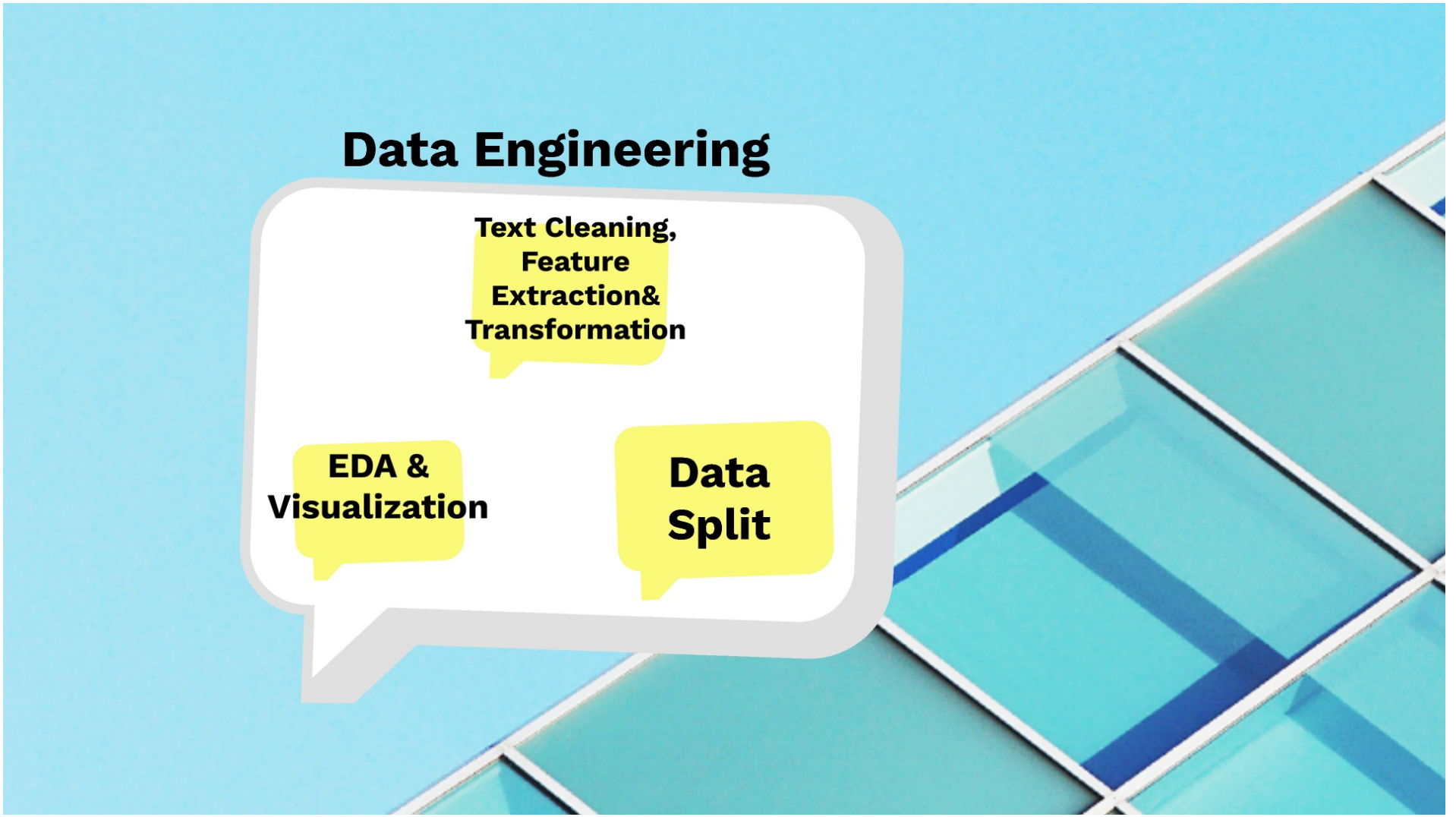
Feature Extraction: Bag of Words, TF-IDF, Word2Vec, and hashing



prompt	0	1	2	3	4	5	6	7	8	...	15	16	17	18	19	20	21
0 Emotional pain	0.0	0.0	0.000000	-0.707107	0.000000	-0.707107	0.000000	0.0	0.000000	...	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000
1 Hair falling out	0.0	0.0	0.000000	-0.333333	-0.333333	-0.666667	-0.333333	0.0	0.333333	...	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000
2 Heart hurts	0.0	0.0	0.000000	0.000000	0.000000	0.000000	-0.707107	0.0	0.000000	...	0.000000	0.0	0.000000	-0.707107	0.000000	0.0	0.000000
3 Infected wound	0.0	0.0	0.000000	0.267261	0.000000	0.000000	-0.267261	0.0	0.267261	...	0.534522	0.0	0.000000	0.000000	0.000000	0.0	0.000000
4 Infected wound	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.000000	0.000000	0.707107	0.0	0.000000

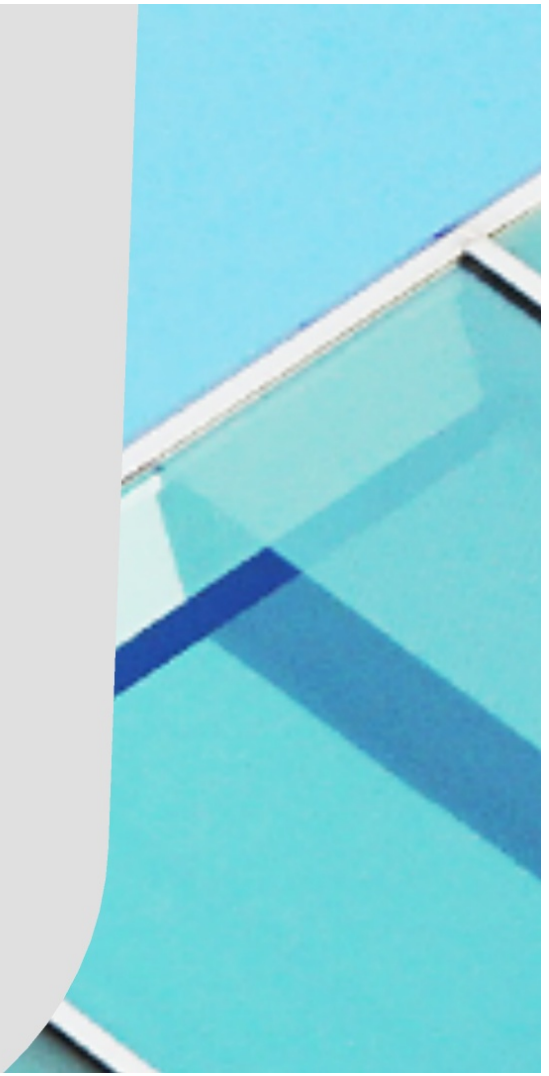
Dimension Reduction using PCA with 99% variance to mitigate curse of dimensions - our datasets are sparse after extraction

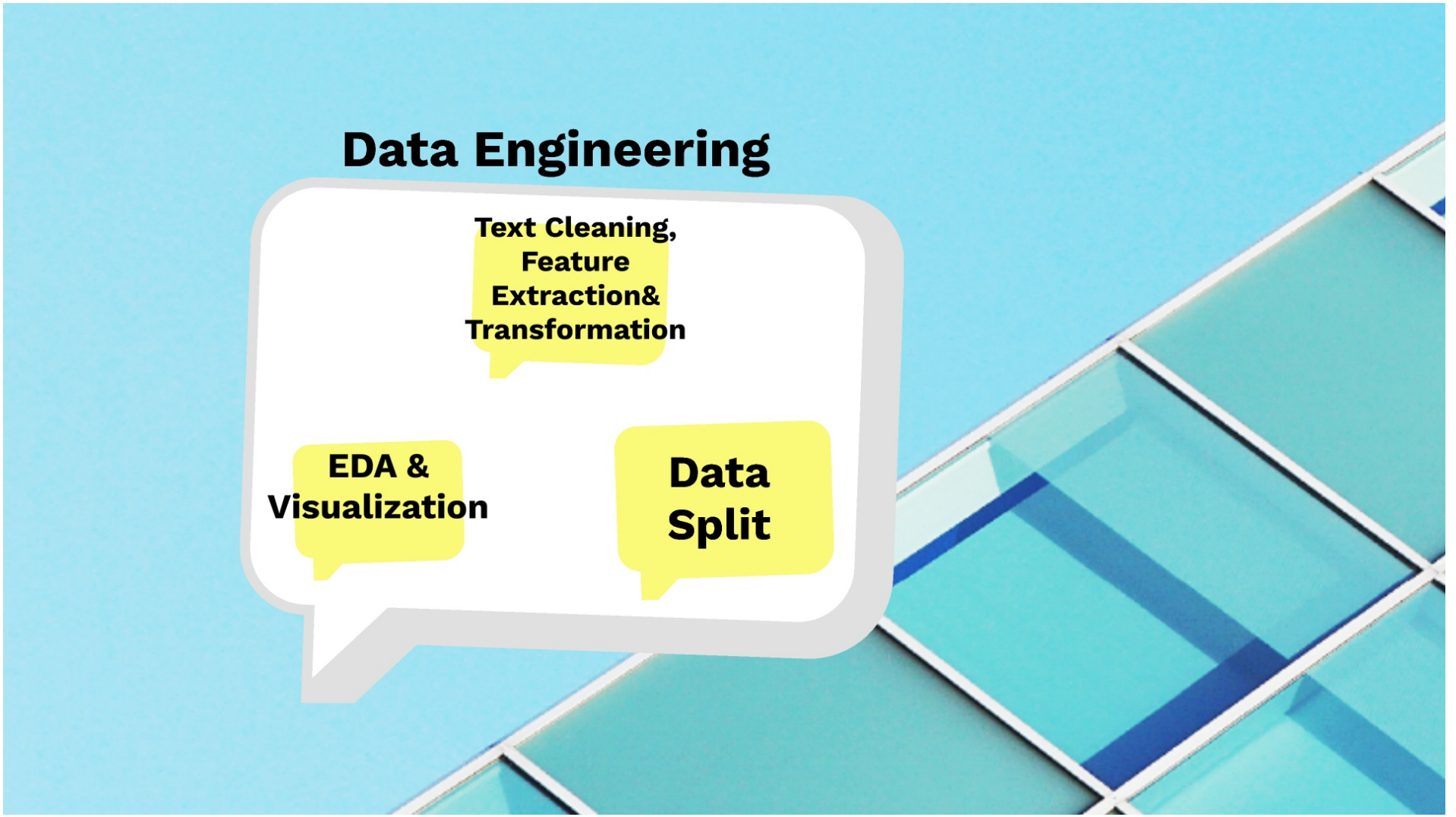
Subsets	# of Features	# of Reduced Features
Bag of Words	953	464
TF-IDF	953	507
Hash Vectorizer	75	69
Word2Vec	100	8

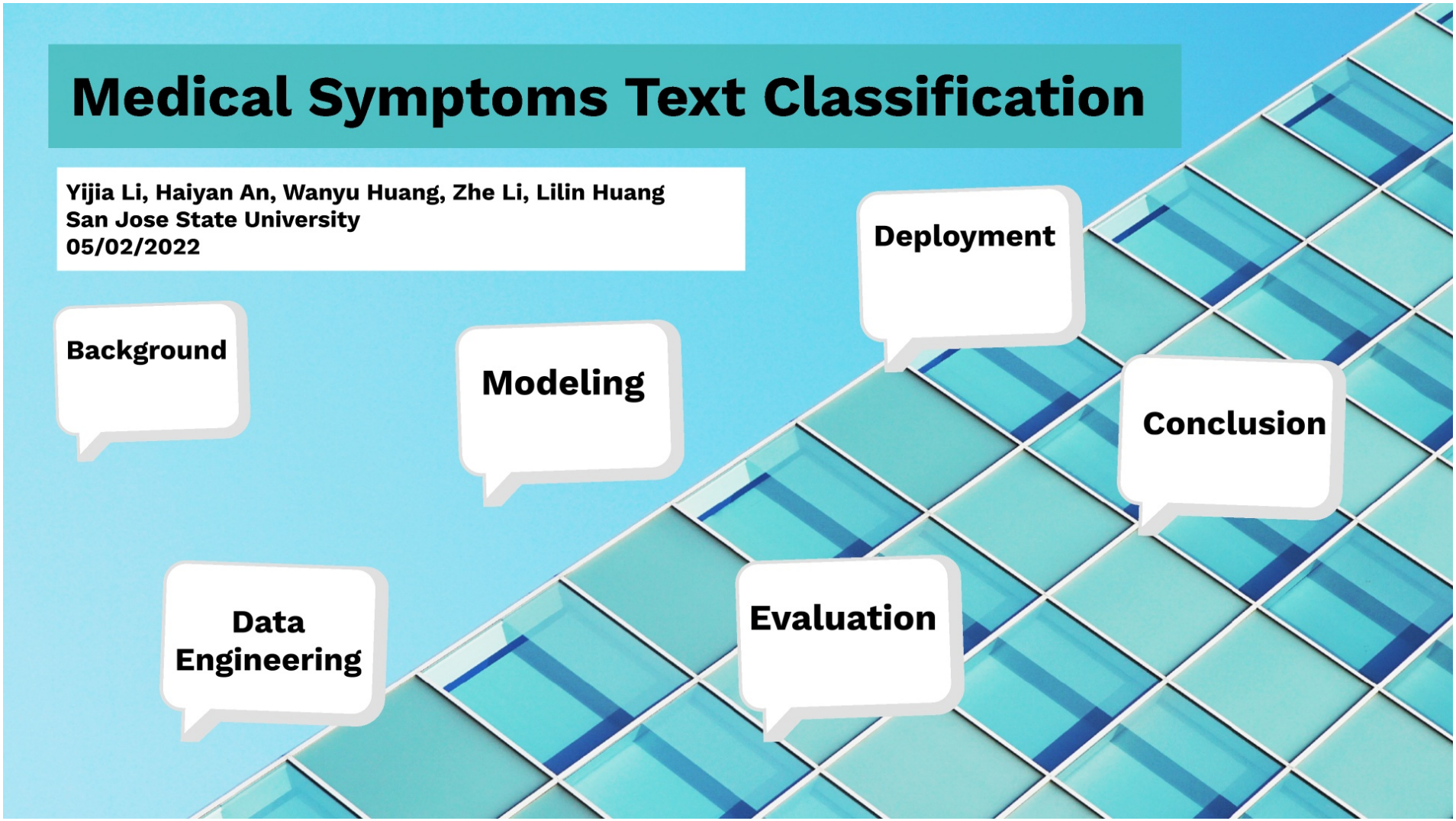

Data Engineering
Text Cleaning, Feature Extraction& Transformation
EDA & Visualization
Data Split

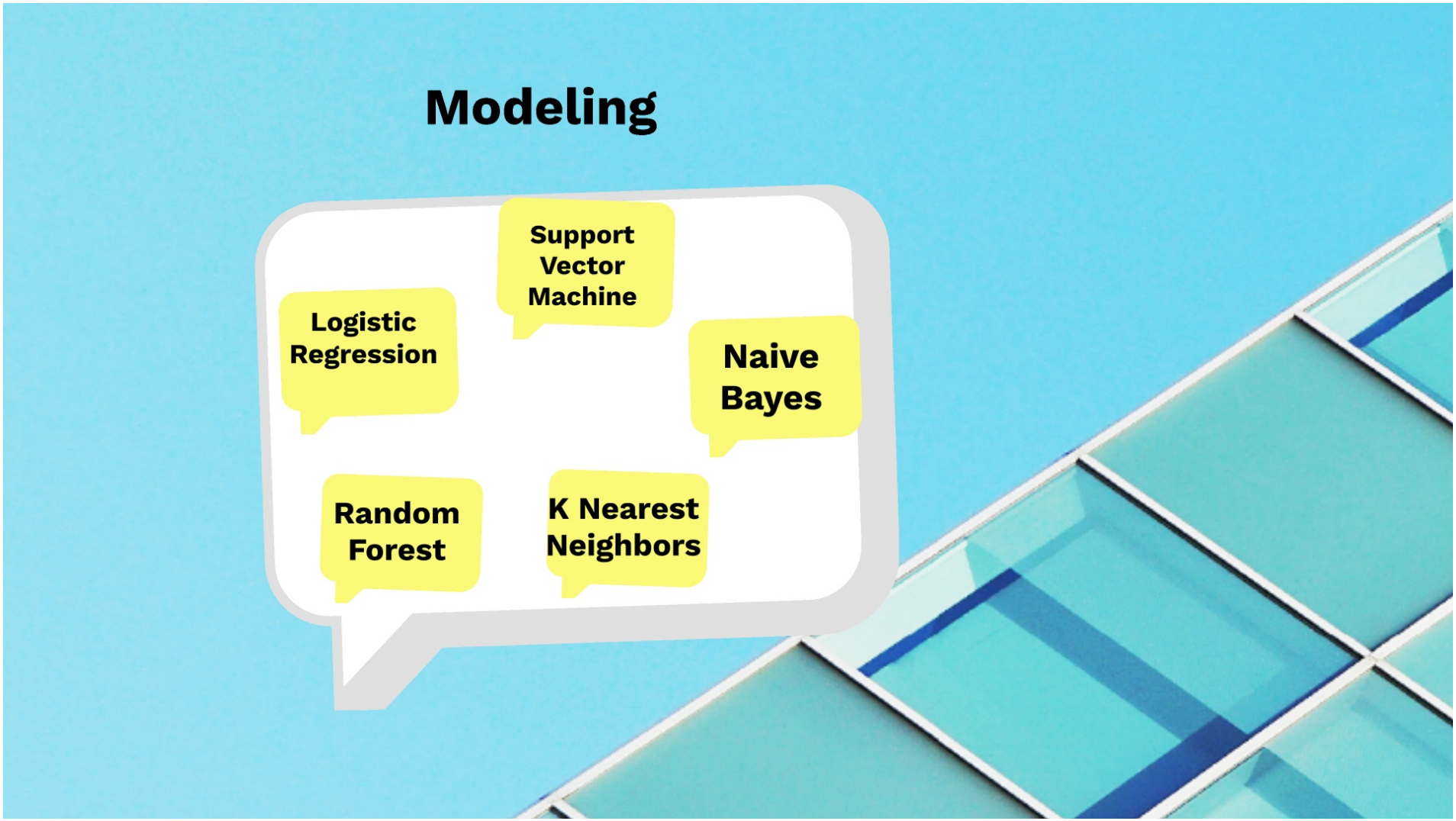
Data Split

For each feature set, Bag of Words, TF-IDF, Word2Vec, and hashing, split into train (80%) and test (20%)



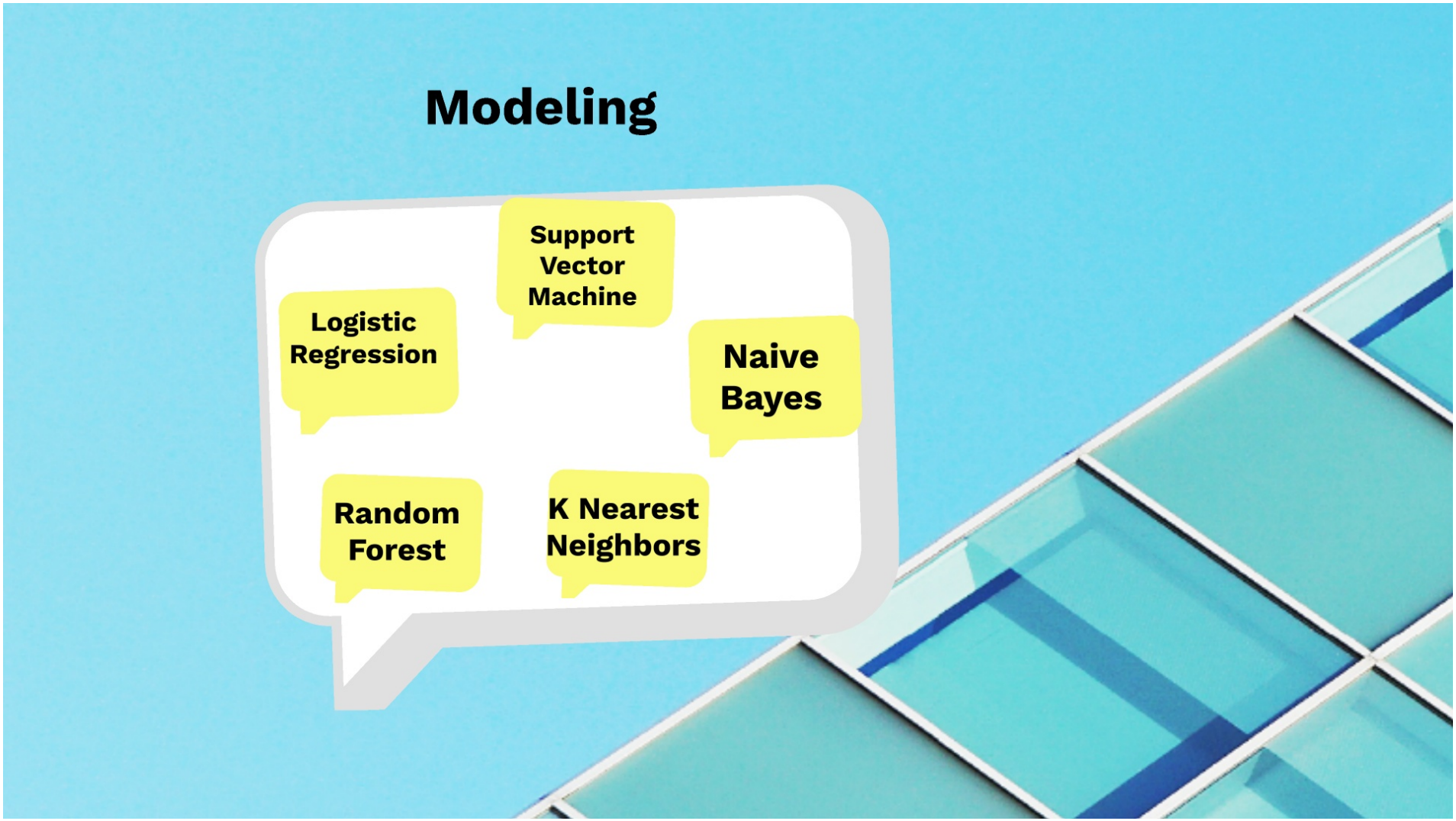

Data Engineering
Text Cleaning, Feature Extraction& Transformation
EDA & Visualization
Data Split


Medical Symptoms Text Classification
Yijia Li, Haiyan An, Wanyu Huang, Zhe Li, Lilin Huang
San Jose State University
05/02/2022
Deployment
Background
Modeling
Conclusion
Data Engineering
Evaluation


Modeling
Support Vector Machine
Logistic Regression
Naive Bayes
Random Forest
K Nearest Neighbors

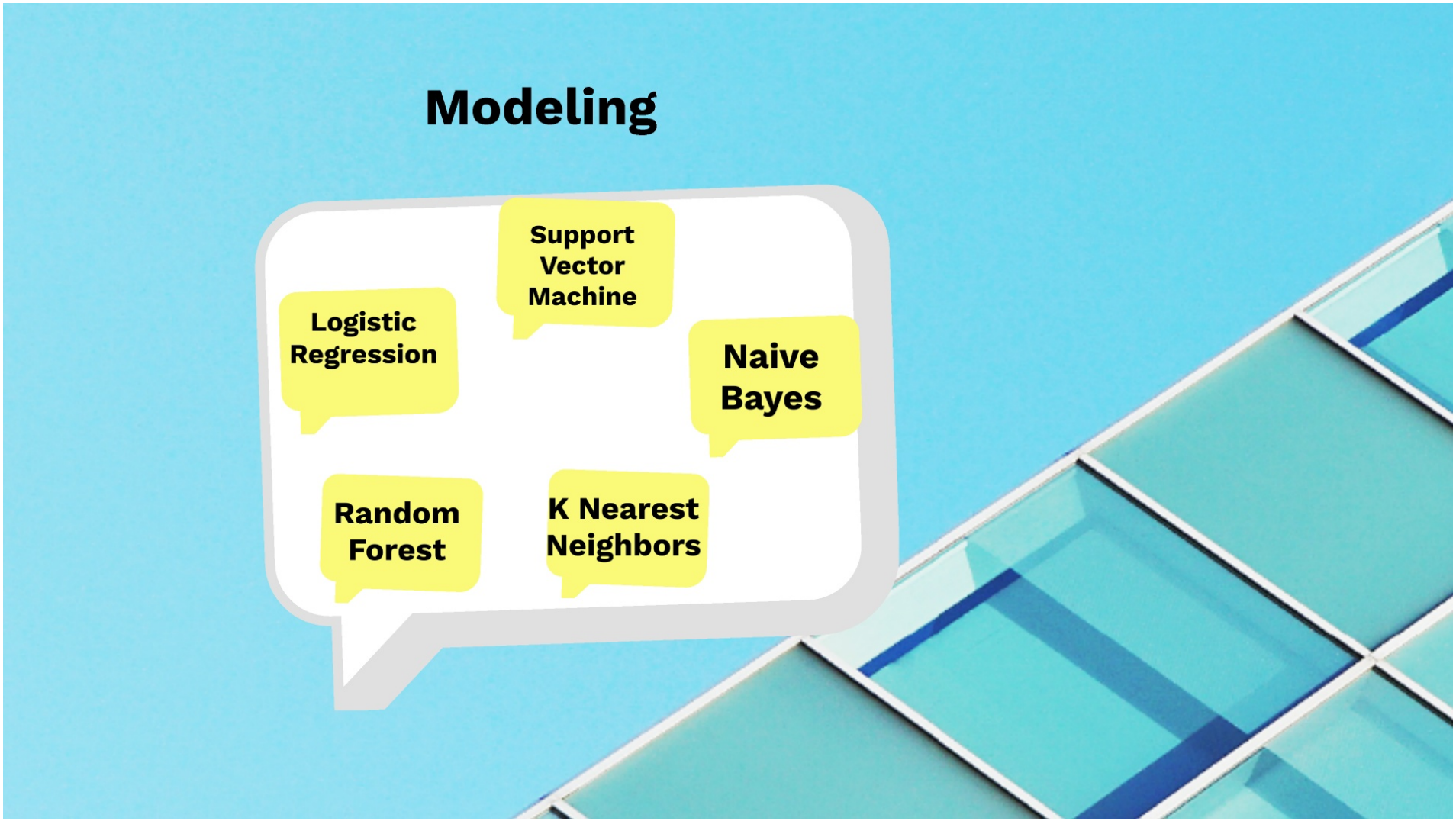
Logistic Regression

- A discriminative classifier
  - the assumption of a linear relationship
- A probabilistic model
  - returns a probability after applying the logistic transformation
- Logistic function is nonlinear
  - the model is more smooth and robust than linear regression
- An inherently binary-class classifier
  - One Vs Rest methods to build 25 classifiers to predict the class
  - scikit-learn library simplify the work


Modeling
Support Vector Machine
Logistic Regression
Naive Bayes
Random Forest
K Nearest Neighbors

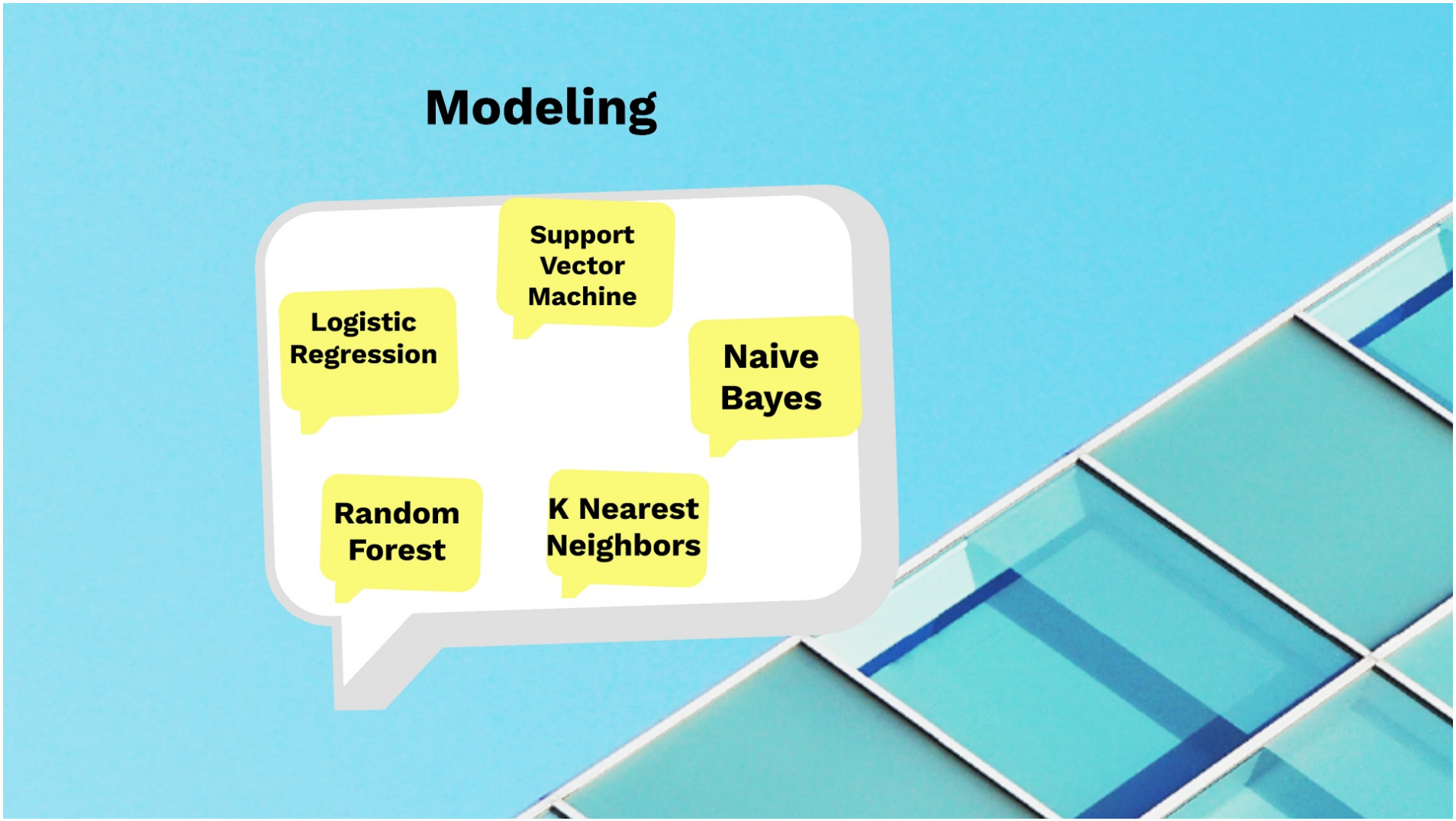
Support Vector Machine

- A margin maximization classifier
  - search for an optimal hyperplane that maximizes the margins
- Considering only support vectors point
  - the points closest to the hyperplane as known as support vectors
  - efficient for high-dimensional and large data sets
- An inherently binary-class classifier
  - One Vs Rest methods to build 25 classifiers to predict the class
  - scikit-learn library simplify the work


Modeling
Support Vector Machine
Logistic Regression
Naive Bayes
Random Forest
K Nearest Neighbors

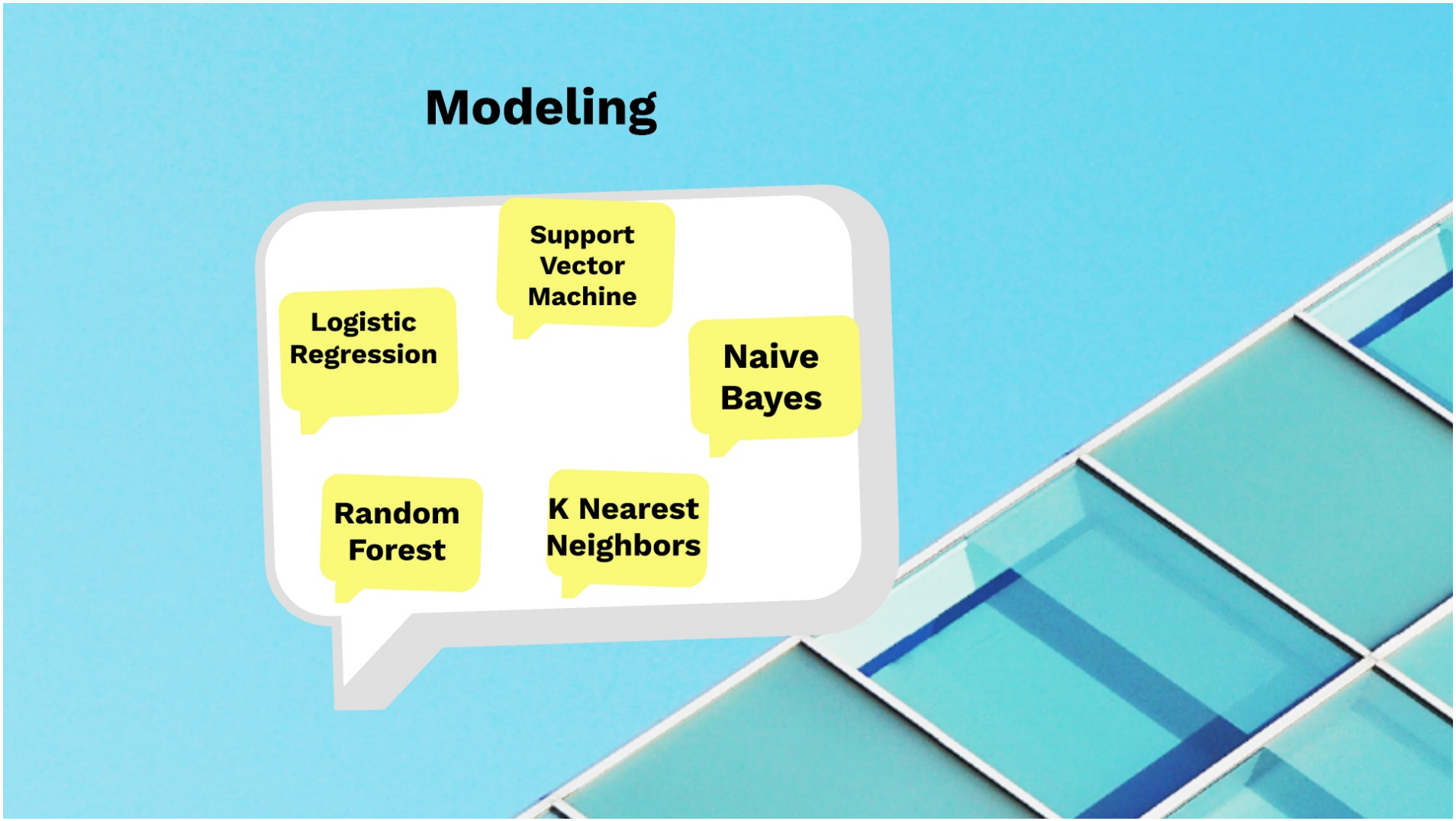
Naive Bayes

- A generative model
  - derive the actual probability of class given data features
  - needs a significant amount of data to get accurate likelihood probabilities
- Relied on strong assumptions
  - Our text features are not conditional independent
  - Hard to assume data distribution for transformed features: Gaussion, multinomial?
- An inherently multi-class classifier
  - find likelihood probabilities for each class
  - implement using scikit-learn library


Modeling
Support Vector Machine
Logistic Regression
Naive Bayes
Random Forest
K Nearest Neighbors

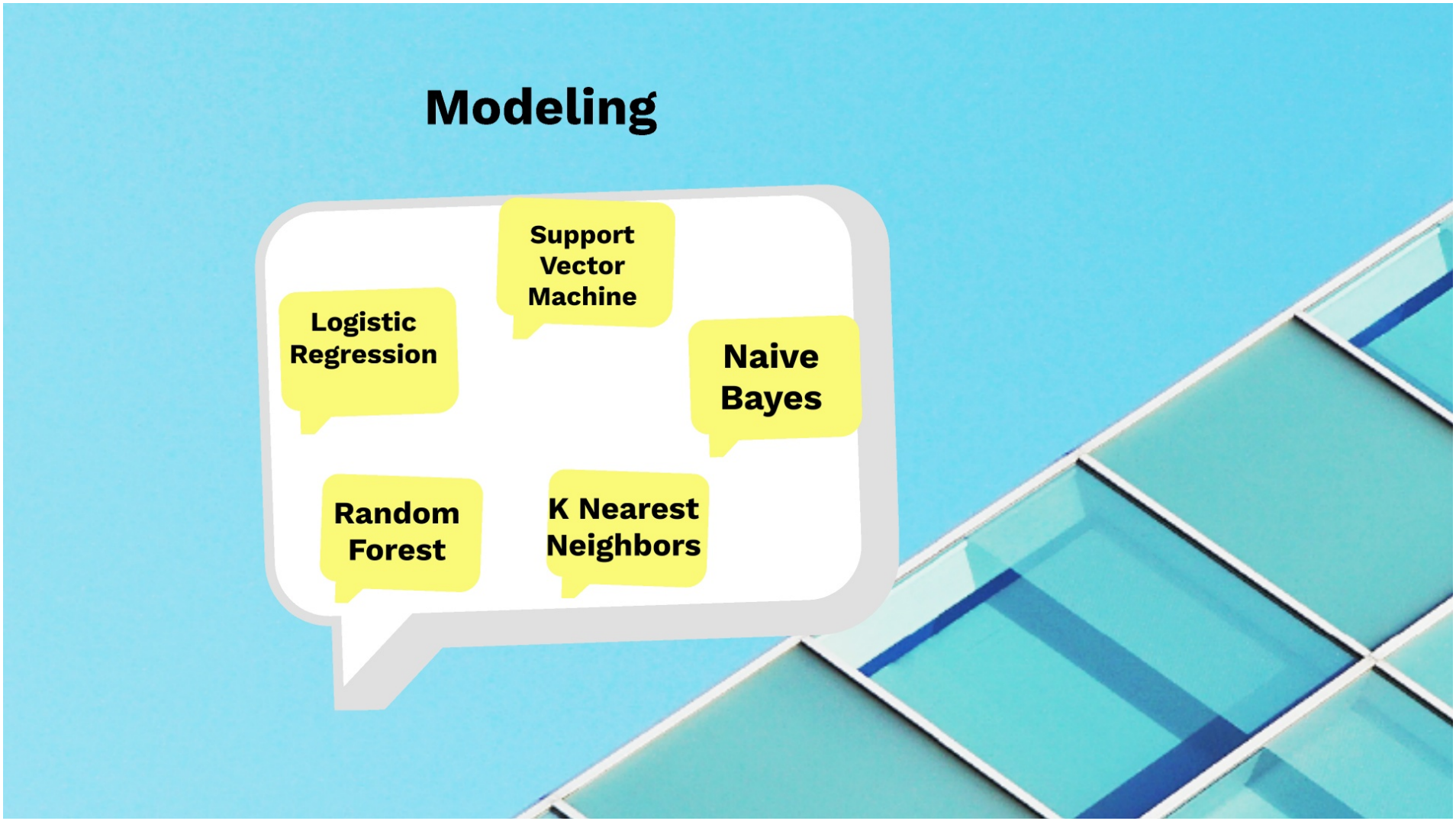
K Nearest Neighbors

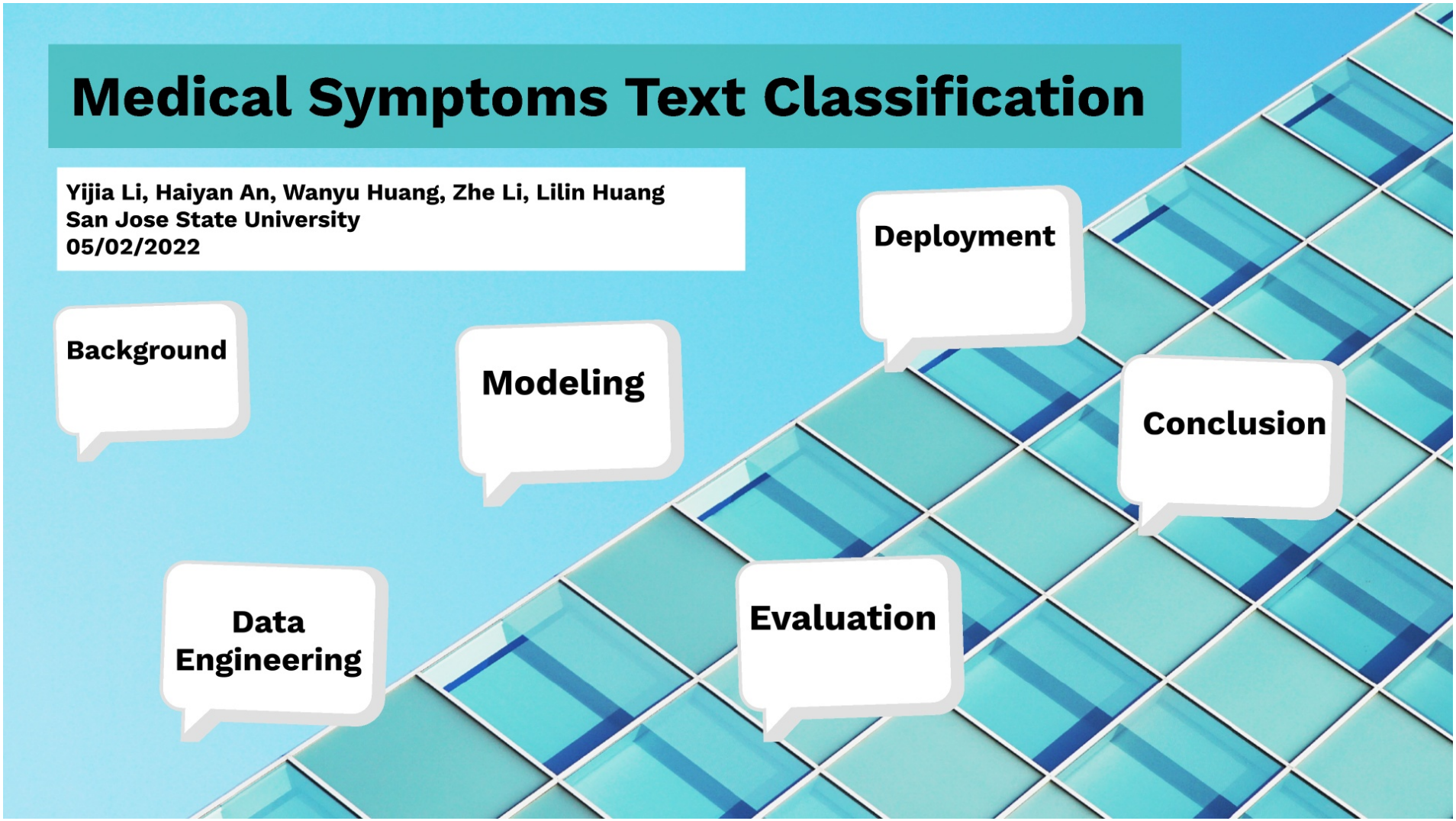
- A memory-based learning model
  - stores feature vectors of training data points in the memory
  - modeling after receive test data points
- A distance-based method
  - data should be normalized
  - handle highly nonlinear decision boundaries and doesn't assume the decision boundary's shape
  - sensitive to hyperparameters choice
- An inherently multi-class classifier
  - predicts the class of unseen data based on the class of neighbors that appears the most time
  - implement using scikit-learn library


Modeling
Support Vector Machine
Logistic Regression
Naive Bayes
Random Forest
K Nearest Neighbors

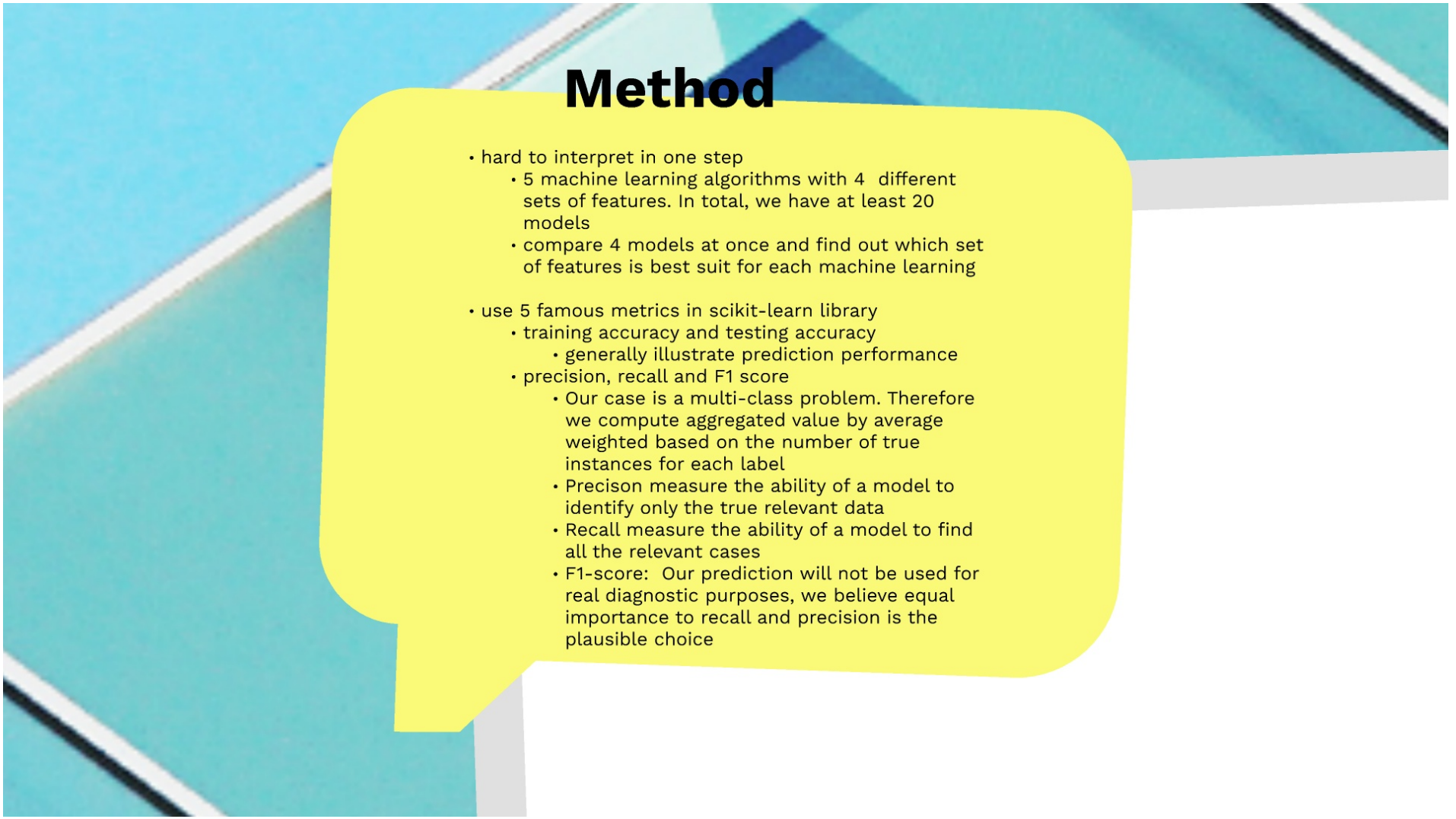
Random Forest

- An ensemble model
  - consists of a large number of individual decision trees
  - individual trees are trained on random subsets of features and samples
- Learn from individual trees' errors
  - data need to be preprocessed so that their errors can be least correlated\
  - sensitive to hyperparameter choices
- An inherently multi-class classifier
  - predicts the class with the most votes from those trees
  - implement using scikit-learn library


Modeling
Support Vector Machine
Logistic Regression
Naive Bayes
Random Forest
K Nearest Neighbors


Medical Symptoms Text Classification
Yijia Li, Haiyan An, Wanyu Huang, Zhe Li, Lilin Huang
San Jose State University
05/02/2022
Background
Modeling
Deployment
Conclusion
Data Engineering
Evaluation


Evaluation
Method
Logistic Regression
Support Vector Machine
K Nearest Neighbors
Naive Bayes
Best
Random Forest



Method

- hard to interpret in one step
  - 5 machine learning algorithms with 4 different sets of features. In total, we have at least 20 models
  - compare 4 models at once and find out which set of features is best suit for each machine learning

- use 5 famous metrics in scikit-learn library
  - training accuracy and testing accuracy
    - generally illustrate prediction performance
  - precision, recall and F1 score
    - Our case is a multi-class problem. Therefore we compute aggregated value by average weighted based on the number of true instances for each label
    - Precison measure the ability of a model to identify only the true relevant data
    - Recall measure the ability of a model to find all the relevant cases
    - F1-score: Our prediction will not be used for real diagnostic purposes, we believe equal importance to recall and precision is the plausible choice


Evaluation
Method
Logistic Regression
Support Vector Machine
K Nearest Neighbors
Naive Bayes
Best
Random Forest

Logistic Regression

- Logistic Regression with bag-of-words extracted data is the best among all LR Classifers
- Logistic Regression with word2vec extracted data is better than 4 % from random guessing since we have 25 classes but still drastically downgraded the performance

2. Metric Table for Logistic Regression

Model	Training Accuracy %	Testing Accuracy %	Testing Precision %	Testing Recall %	Testing F1 Score %
LR_bow	99.568318	99.549887	99.549887	99.574027	99.549153
LR_tf-idf	99.399399	99.474869	99.474869	99.500348	99.474060
LR_hash	85.904655	84.696174	84.696174	85.447843	84.617119
LR_w2v	27.496246	27.906977	27.906977	22.557786	20.560358


Evaluation
Method
Logistic Regression
Support Vector Machine
K Nearest Neighbors
Naive Bayes
Best
Random Forest

Support Vector Machine

Support Vectors Classifier with tf_idf extracted data are the best among all SVCs

Support Vectors Classifier with word2vec extracted data has the same performance issue like LR

3. METRIC TABLE FOR SUPPORT VECTOR CLASSIFIER

Model	Training Accuracy %	Testing Accuracy %	Testing Precision %	Testing Recall %	Testing F1 Score %
SVC_bow	99.587087	99.549887	99.549887	99.574027	99.549153
SVC_tf-idf	99.605856	99.549887	99.549887	99.574027	99.549153
SVC_hash	99.324324	99.174794	99.174794	99.249930	99.176335
SVC_w2v	46.677928	43.735934	43.735934	45.756499	42.670226


Evaluation
Method
Logistic Regression
Support Vector Machine
K Nearest Neighbors
Naive Bayes
Best
Random Forest

Naive Bayes

Guassian Naive Bayes Classifier with tf_idf extracted data are the best among all Naive Bayes Classifiers

Gaussian Naive Bayes Classifier with word2vec extracted data has the same performance issue like LR and SVM

Why

4. METRIC TABLE FOR GAUISSIAN NAIVE BAYES

Model	Training Accuracy %	Testing Accuracy %	Testing Precision %	Testing Recall %	Testing F1 Score %
GNB_bow	90.653153	87.546887	87.546887	89.318237	87.667300
GNB_tf-idf	90.878378	87.921980	87.921980	90.570234	88.415025
GNB_hash	79.298048	75.843961	75.843961	77.614749	76.140452
GNB_w2v	47.184685	48.012003	48.012003	50.995666	48.054367

All dataset has been preprocessed with same step and been extracted PCA components based on 99% variance. The only difference is that word2vec dataset contains only 8 features. We assume the reason behind is those models needs a significant amount of data in high dimensions to get either plausible decision boundaries or accurate likelihood probabilities.

Naive Bayes

Guassian Naive Bayes Classifier with tf_idf extracted data are the best among all Naive Bayes Classifiers

Gaussian Naive Bayes Classifier with word2vec extracted data has the same performance issue like LR and SVM

Why

4. METRIC TABLE FOR GAUISSIAN NAIVE BAYES

Model	Training Accuracy %	Testing Accuracy %	Testing Precision %	Testing Recall %	Testing F1 Score %
GNB_bow	90.653153	87.546887	87.546887	89.318237	87.667300
GNB_tf-idf	90.878378	87.921980	87.921980	90.570234	88.415025
GNB_hash	79.298048	75.843961	75.843961	77.614749	76.140452
GNB_w2v	47.184685	48.012003	48.012003	50.995666	48.054367


Evaluation
Method
Logistic Regression
Support Vector Machine
K Nearest Neighbors
Naive Bayes
Best
Random Forest

K Nearest Neighbors

- K Nearest Neighours with bag-of-words extracted data is the best among all KNNs.
- K Nearest Neighours have high scores in every metric and are consistent with all kind of feature dataset
- Tune hyperparameters
  - choice of distance metrics including ‘euclidean’ and ‘manhattan’
  - the number of K neighbors
  - the choice of weight function for K neighbors
    - ‘uniform’ in which All points in each neighborhood are weighted equally
    - ‘distance’ in which weight points by the inverse of their distance.

7. Metric Table for K Nearest Neighbors

Model	Training Accuracy %	Testing Accuracy %	Testing Precision %	Testing Recall %	Testing F1 Score %
KNN_bow	99.699700	99.699925	99.699925	99.708187	99.699770
KNN_tf-idf	99.831081	99.699925	99.699925	99.708048	99.699734
KNN_hash	99.418168	99.324831	99.324831	99.370567	99.326532
KNN_w2v	99.643393	99.549887	99.549887	99.574682	99.549582


Evaluation
Method
Logistic Regression
Support Vector Machine
K Nearest Neighbors
Naive Bayes
Best
Random Forest

Random Forest

- Surprisingly Random Forest with word2vec extracted data is the best among all RF classifers.
- It also have high scores in every metric and are consistent with all kind of feature dataset
- Tune hyperparameters
  - impurity criteria choice of ‘Gini’ and ‘Entropy’
  - maximum depth of individual trees
  - minimum samples allowed in one node of the tree
  - number of individual trees in the Random Forest

5. METRIC TABLE FOR RANDOM FOREST

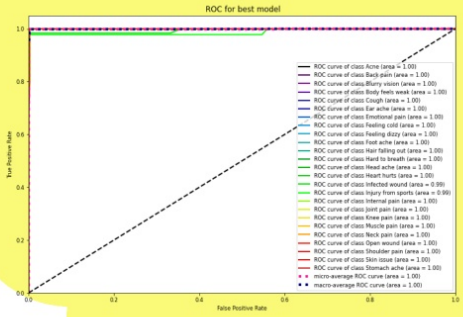
Model	Training Accuracy %	Testing Accuracy %	Testing Precision %	Testing Recall %	Testing F1 Score %
RF_bow	99.774775	99.549887	99.549887	99.574821	99.549617
RF_tf-idf	99.774775	99.549887	99.549887	99.574821	99.549617
RF_hash	99.493243	99.174794	99.174794	99.194755	99.176006
RF_w2v	99.774775	99.549887	99.549887	99.574932	99.549645

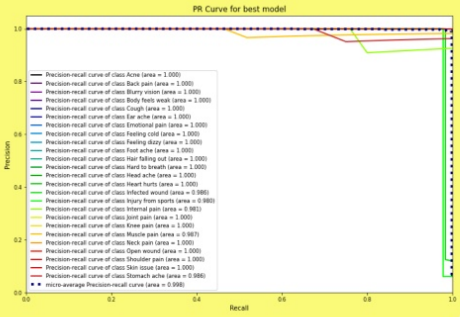

Evaluation
Method
Logistic Regression
Support Vector Machine
K Nearest Neighbors
Naive Bayes
Best
Random Forest

Best

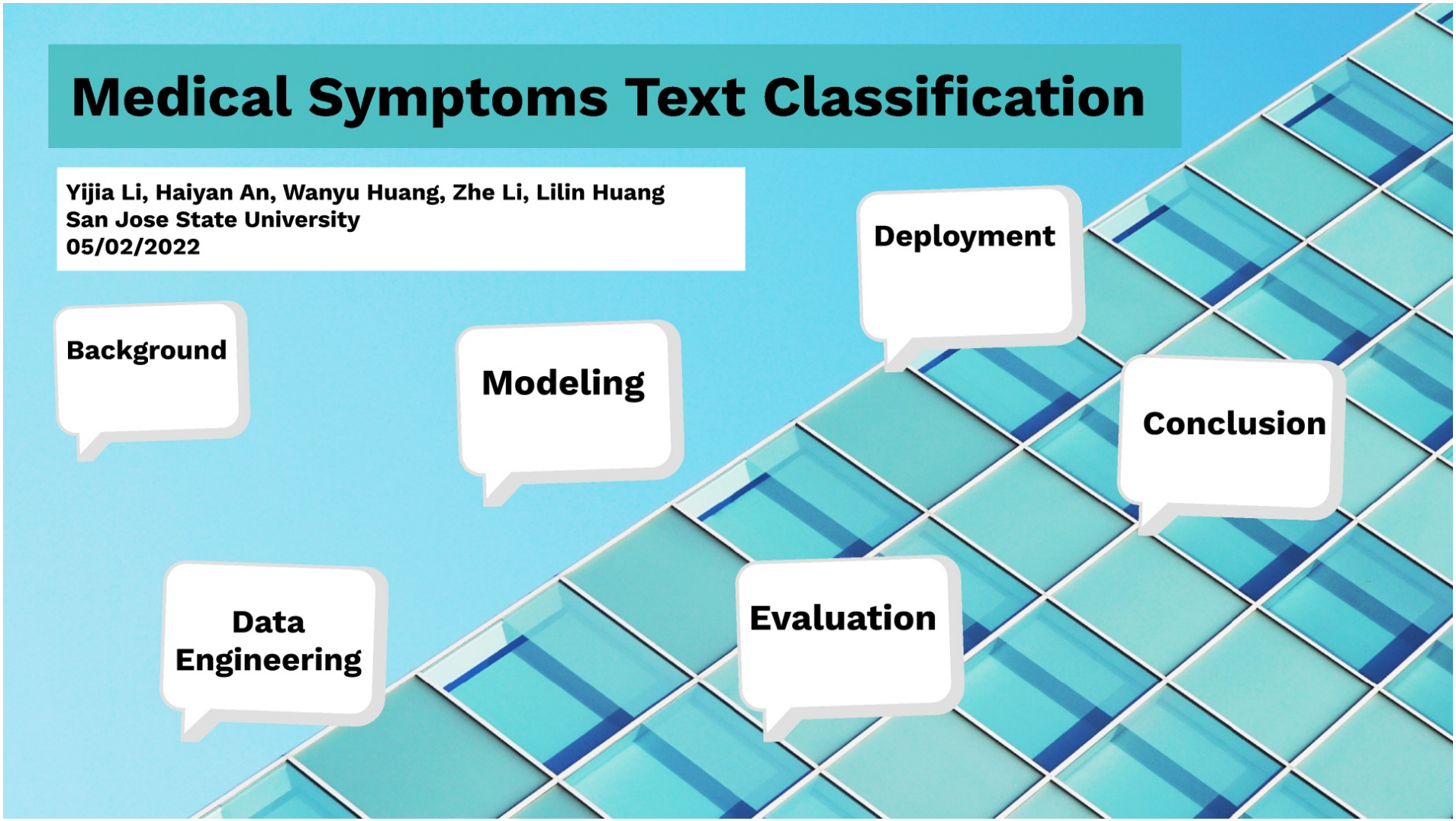
Model	Training Accuracy %	Testing Accuracy %	Testing Precision %	Testing Recall %	Testing F1 Score %
Tuned KNN	99.699700	99.699925	99.699925	99.708048	99.699734
Tuned RF	99.831081	99.699925	99.699925	99.708048	99.699734

After tuning, the Best Model: Random Forest Classifier (impurity criteria = 'Gini', maximum depth = 13, minimum samples allowed = 1, number of individual trees = 50) + Word2Vec Features

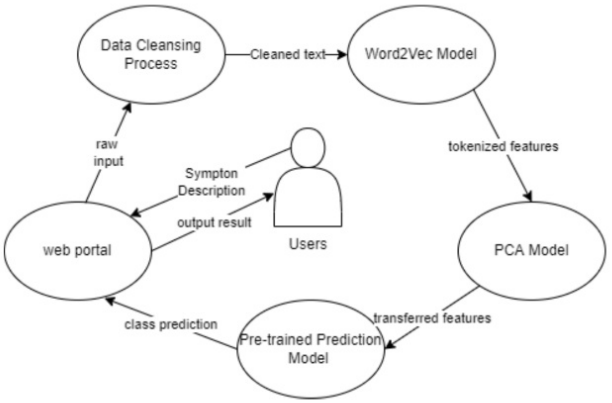




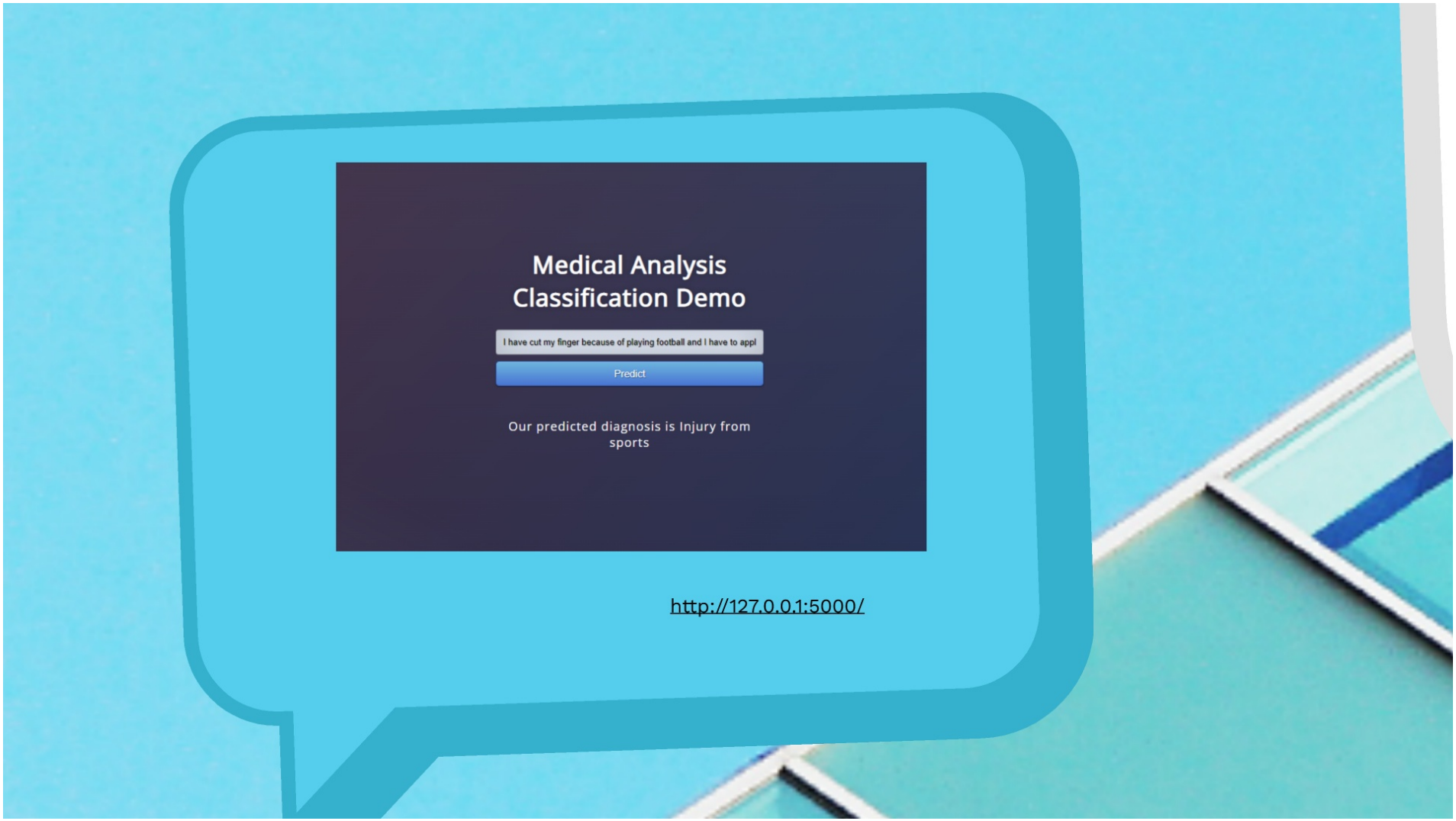

Evaluation
Method
Logistic Regression
Support Vector Machine
K Nearest Neighbors
Naive Bayes
Best
Random Forest


Medical Symptoms Text Classification
Yijia Li, Haiyan An, Wanyu Huang, Zhe Li, Lilin Huang
San Jose State University
05/02/2022
Deployment
Background
Modeling
Conclusion
Data Engineering
Evaluation

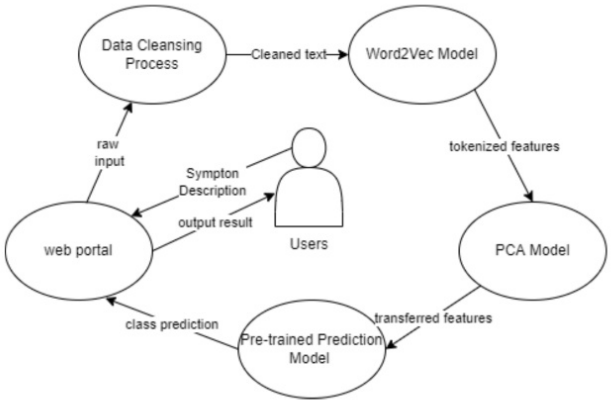
Final Model Deployment



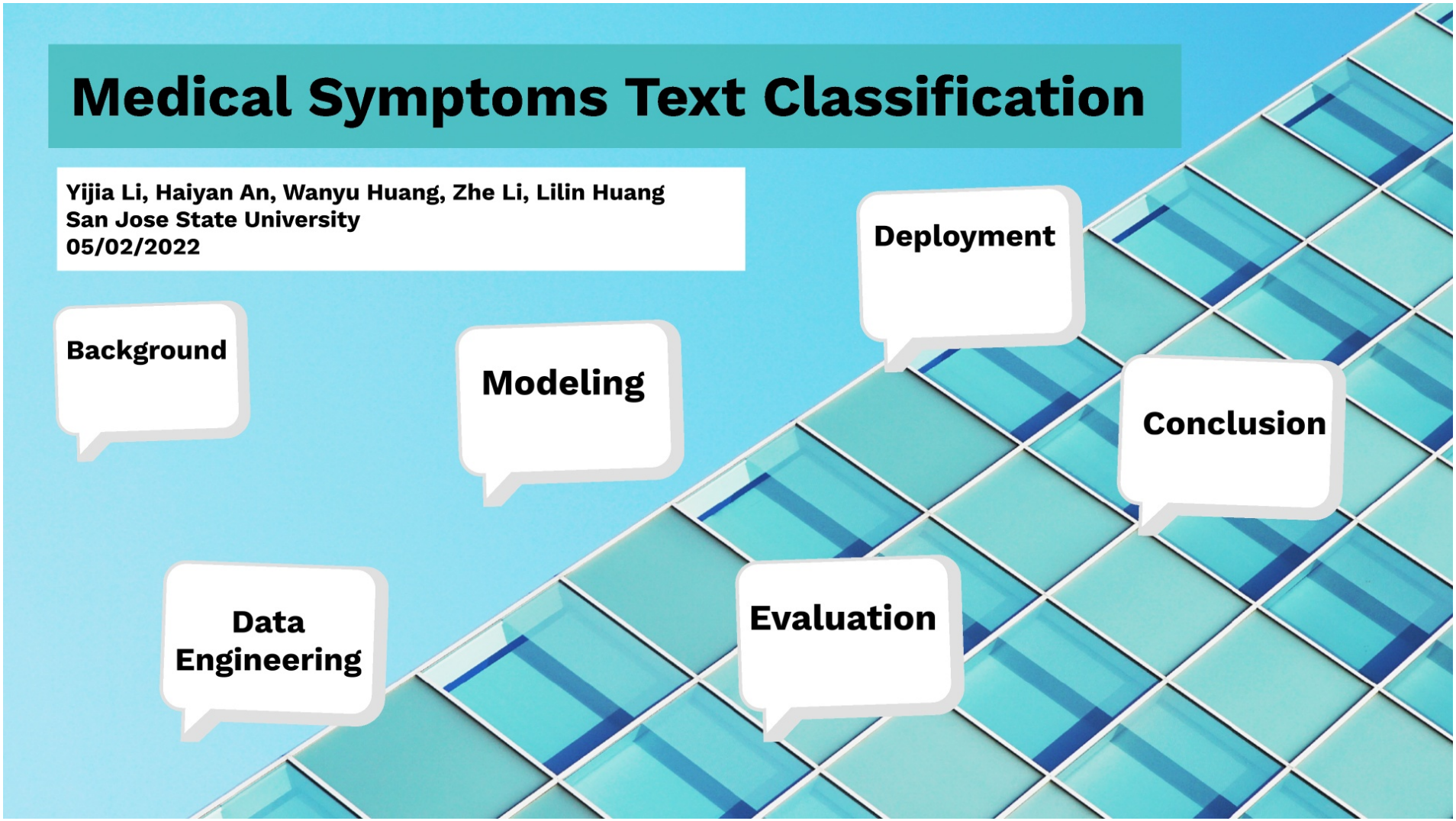
Demo

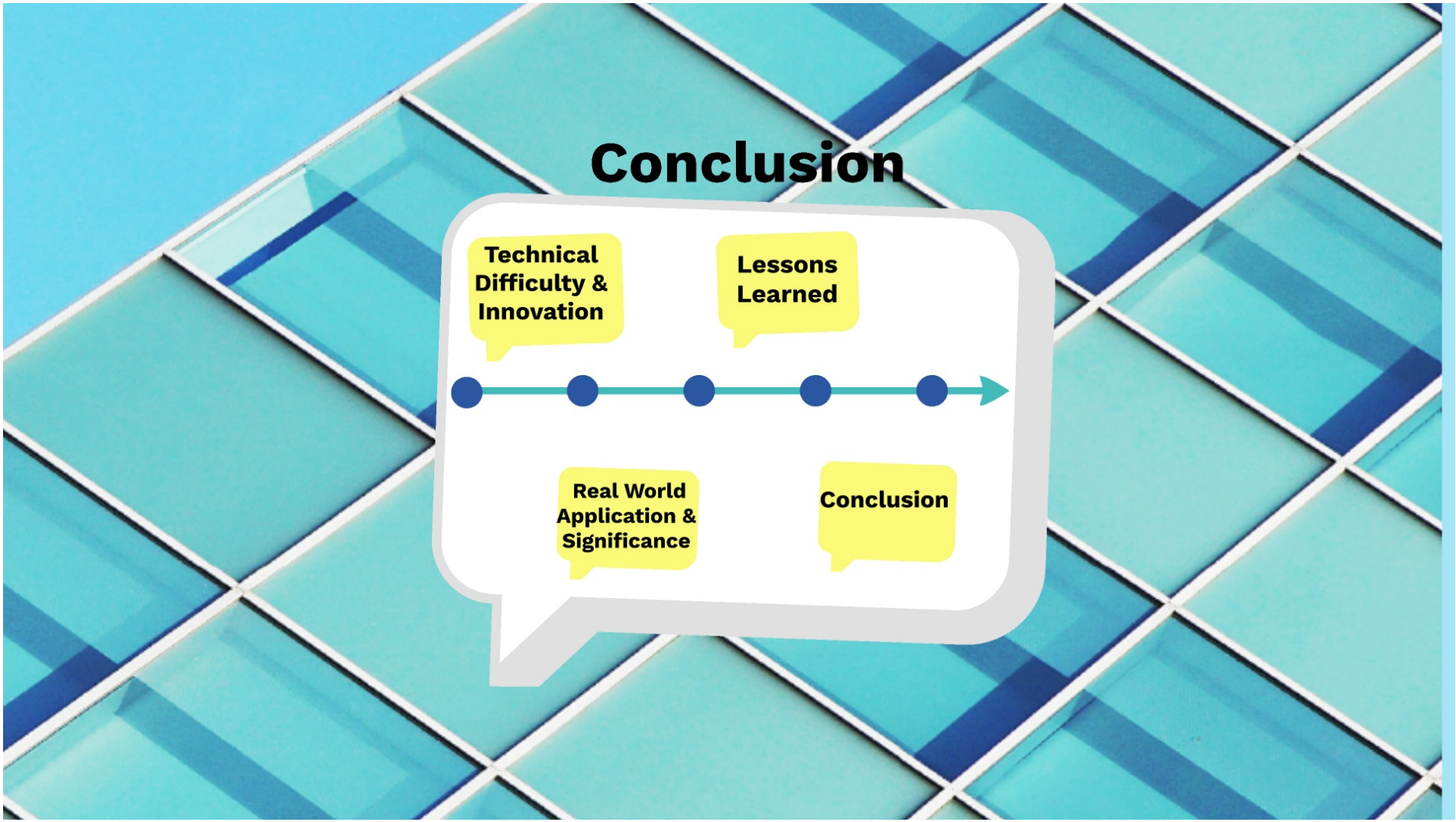

Medical Analysis
Classification Demo
I have cut my finger because of playing football and I have to appl
Predict
Our predicted diagnosis is Injury from sports
http://127.0.0.1:5000/

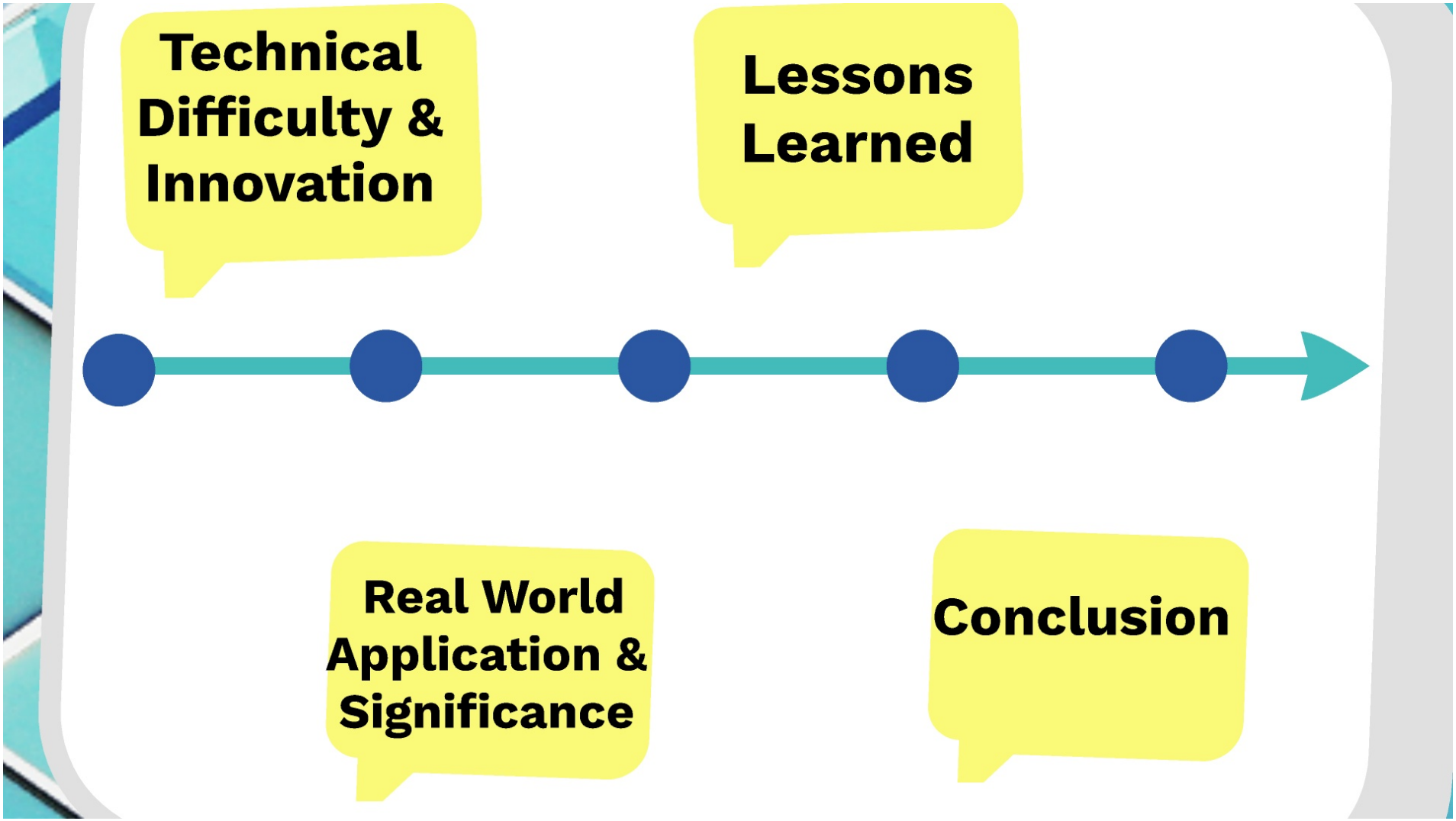
Final Model Deployment



Demo


Medical Symptoms Text Classification
Yijia Li, Haiyan An, Wanyu Huang, Zhe Li, Lilin Huang
San Jose State University
05/02/2022
Background
Modeling
Deployment
Conclusion
Data Engineering
Evaluation


Conclusion
Technical Difficulty & Innovation
Lessons Learned
Real World Application & Significance
Conclusion


Technical Difficulty & Innovation
Lessons Learned
Real World Application & Significance
Conclusion

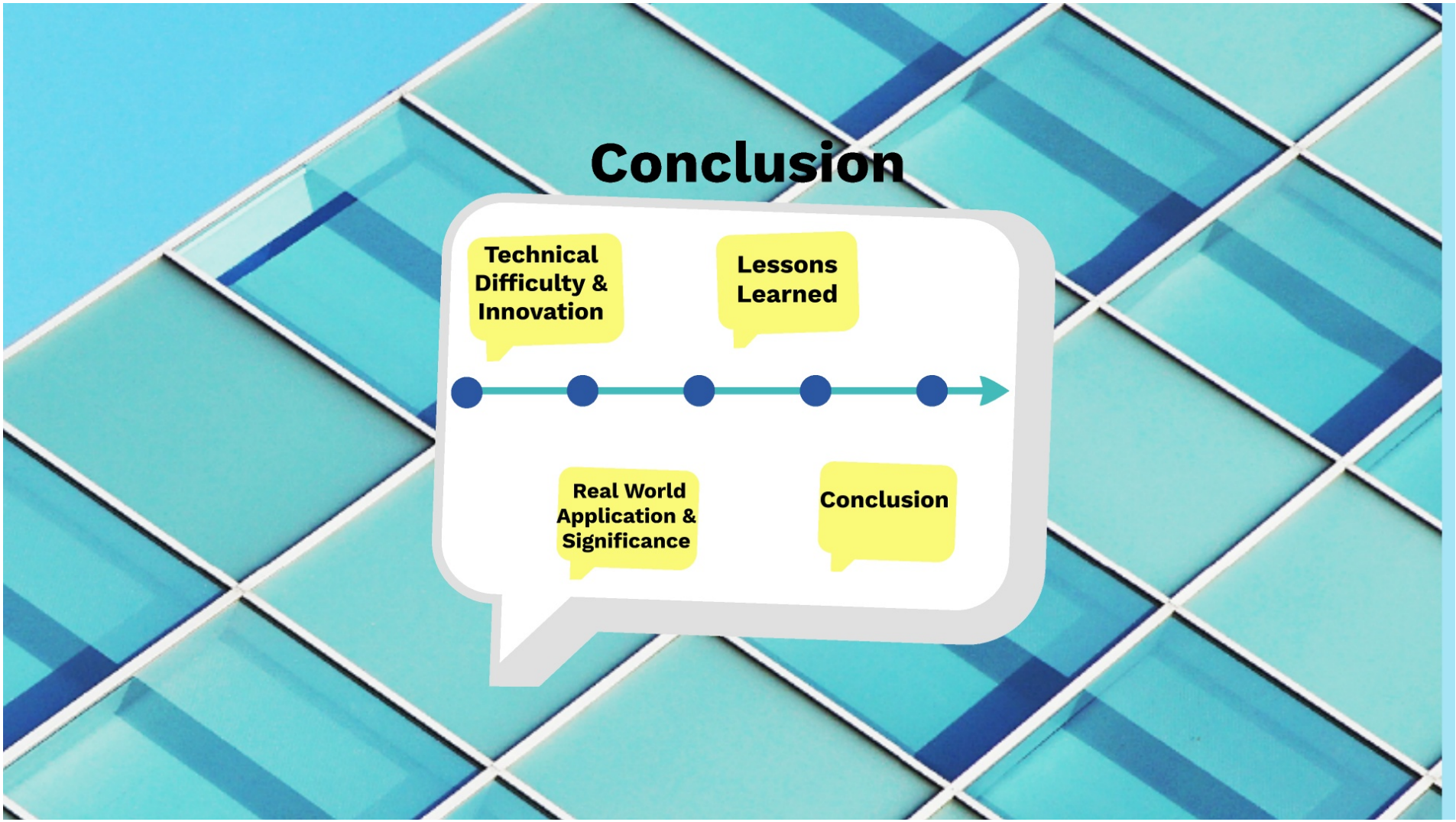
Technical Difficulty & Innovation

Technical Difficulty

- Text Cleaning
- Feature extraction, especially Word2Vec (add function that extends the word vectors to document level)
- Performance evaluation of 20+ models
- Deployment

Innovation

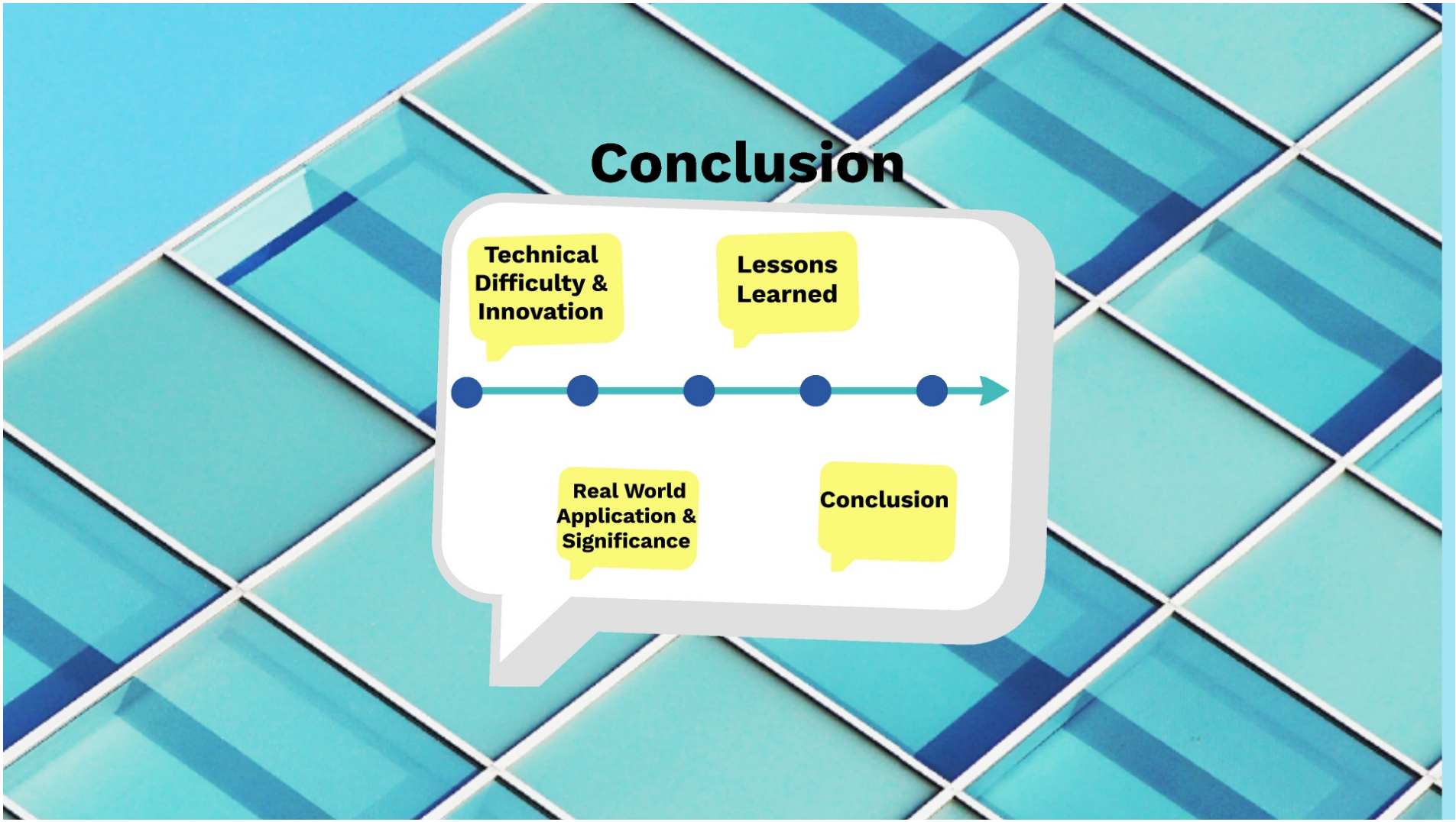
- Find the best combination of machine learning models and feature extraction methods
- Use PCA to reduce feature dimension


Conclusion
Technical Difficulty & Innovation
Lessons Learned
Real World Application & Significance
Conclusion

Real World Application & Significance

As Covid-19 pandemic significantly skyrockets the demand of online medical services, our the medical symptoms text classification system we designed is easily applicable in various medical institutions, adding significant and meaningful life saving and business values to the real world.

- Assisting the medical service institution to identify the patients' symptoms based on text messages, classifying accordingly and transfer to the medical departments with high accuracy and high efficiency.
- Significant reduce direct human contact which reduces the chance of spreading highly infectious diseases like Covid-19.
- Can further extend to be used on the online chat box and auto-reply text messages sent to the phone.


Conclusion
Technical Difficulty & Innovation
Lessons Learned
Real World Application & Significance
Conclusion

Lessons Learned

- Practiced the full cycle of machine learning project including data collection, data processing, modeling, evaluation and deployment
- Features engineering significantly impact the performance of model
    - different sets of algorithms and features perform differently
    - Aggregated word2vec feature set is not suit for LR, SVC, and Gauissan NB
    - Hash vectorized method is apparently weaker than all others
- Accuracy is not an adequate metric especially when models perform nearly perfect
    - There are classes some classifiers have never predicted but still have high accuracy as those classes comprise very small amount in all data
    - During deployment we test some random input and the result seems unexpected even though our accuracy in evaluation is extremely high. Therefore, our model can't really understand the dependency between features and class since we lack of support data - 25 classes but only 6000 support data .

Future Works

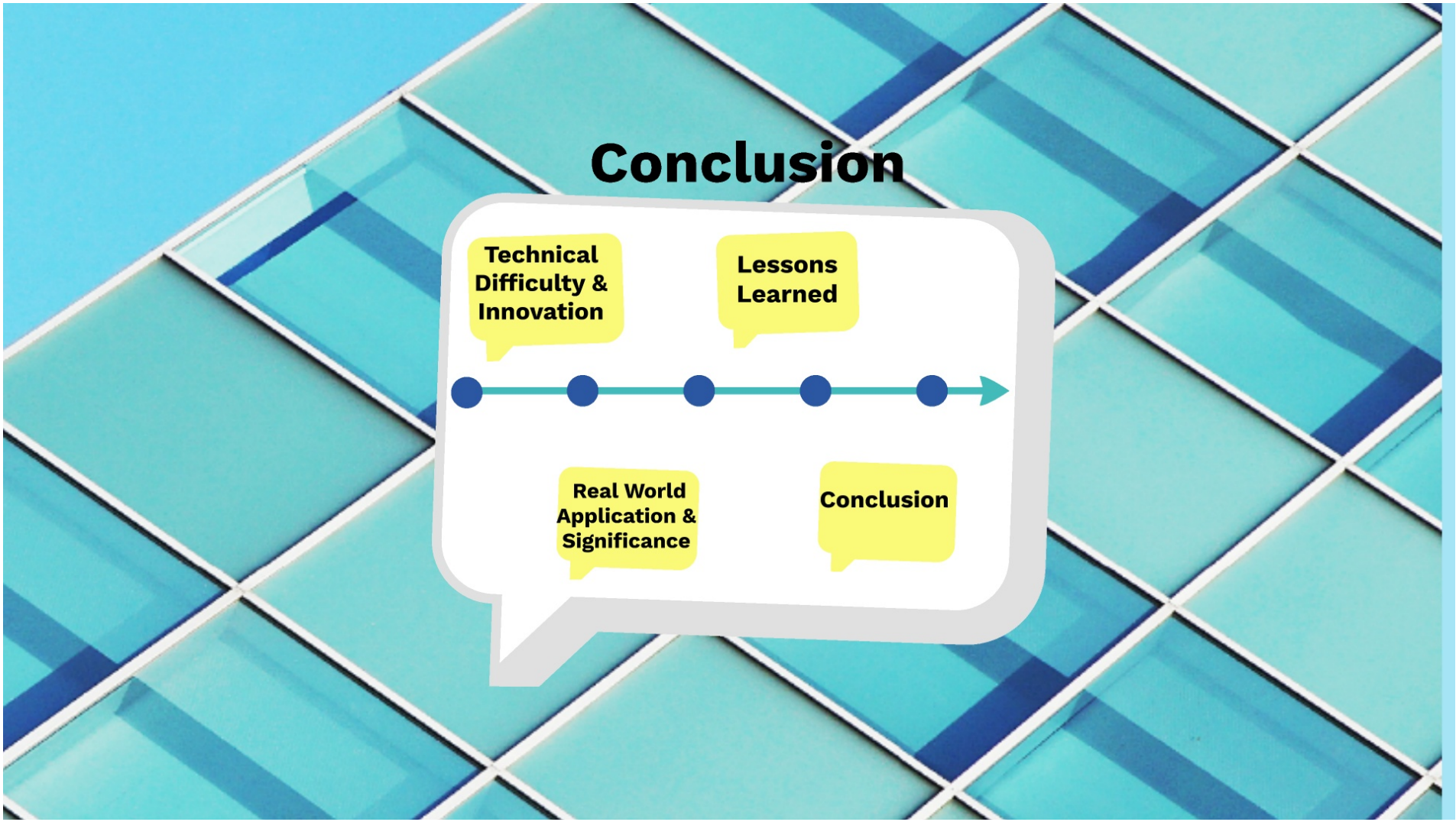
Future Works

- Improving Data Quality to solve the issue of data imbalance and lack of support data
  - collecting more support data
  - binning some classes together to form a more high-level class.
- Deployment with dynamic model
  - model can be continuously improved
  - mitigate our data issue

Lessons Learned

- Practiced the full cycle of machine learning project including data collection, data processing, modeling, evaluation and deployment
- Features engineering significantly impact the performance of model
    - different sets of algorithms and features perform differently
    - Aggregated word2vec feature set is not suit for LR, SVC, and Gauissan NB
    - Hash vectorized method is apparently weaker than all others
- Accuracy is not an adequate metric especially when models perform nearly perfect
    - There are classes some classifiers have never predicted but still have high accuracy as those classes comprise very small amount in all data
    - During deployment we test some random input and the result seems unexpected even though our accuracy in evaluation is extremely high. Therefore, our model can't really understand the dependency between features and class since we lack of support data - 25 classes but only 6000 support data .

Future Works


Conclusion
Technical Difficulty & Innovation
Lessons Learned
Real World Application & Significance
Conclusion

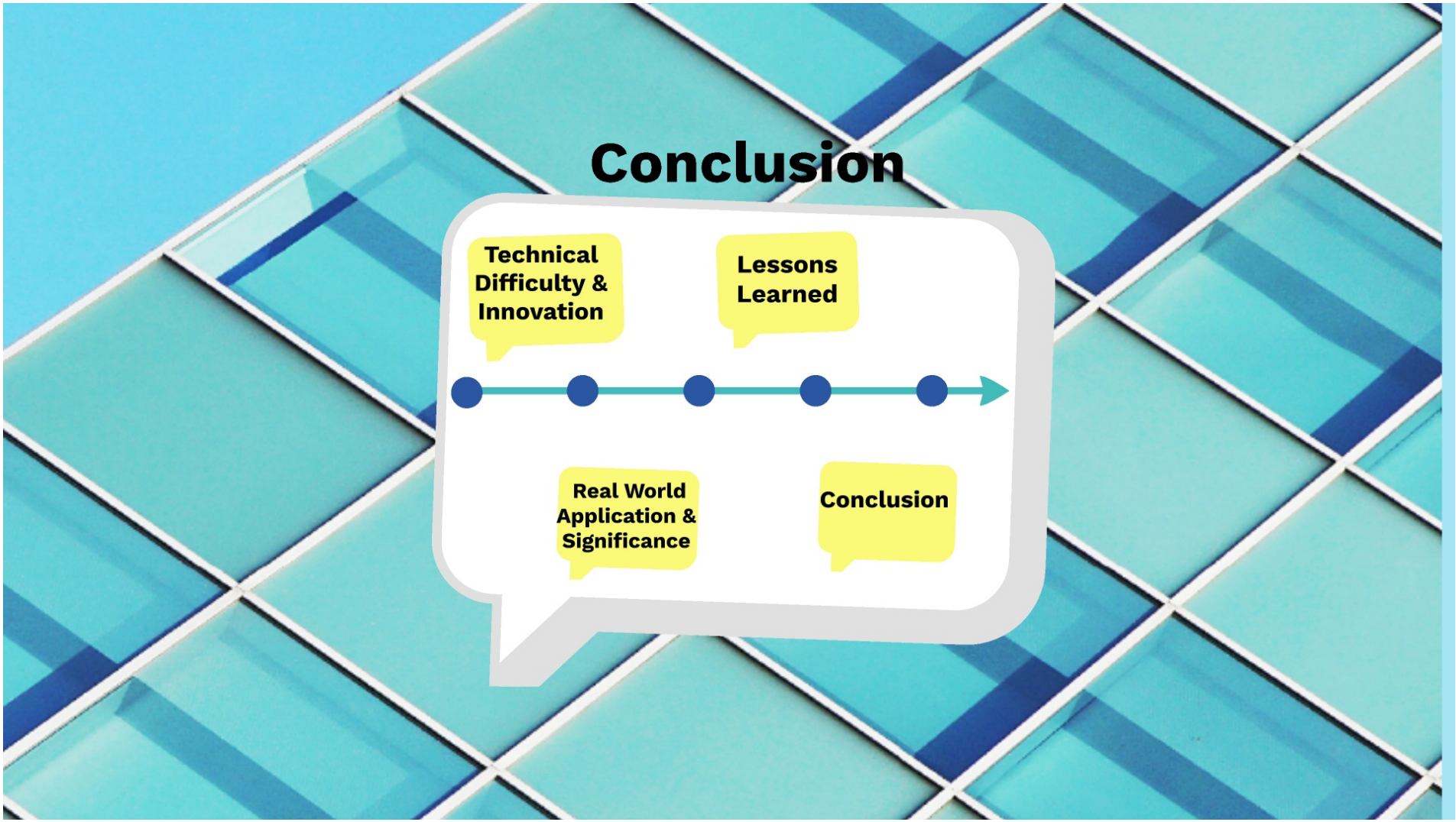
Conclusion

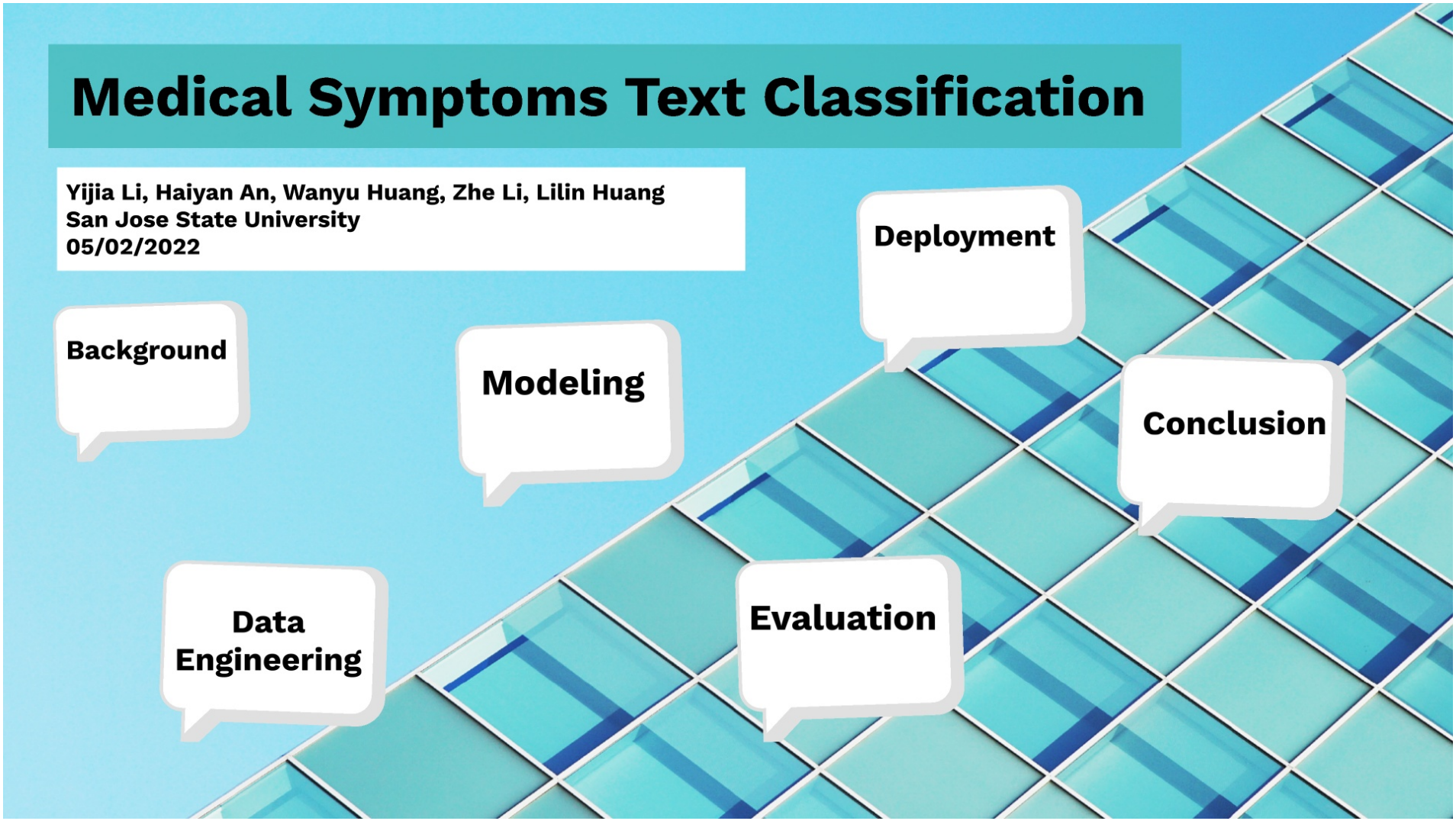
Our model of choice is the tuned random forest classifier using the Word2Vec extracted features. It gives the 99.69% accuracy, 99.69% precision, and 99.71% recall scores in the task of diagnosis classification. It is the model with the best performance compared to the other 21 models.

Our application not only enables patients to early identify the ailment and take preventative actions in advance but also serves as a bridge between patients and real doctors by saving time in matching patients

Our project requiring further work including data quality improving and further digging the relation between features and algorithms to answer why they fit or not. Until that, our model can't really have a realistic and applicable usage

Thank You


Conclusion
Technical Difficulty & Innovation
Lessons Learned
Real World Application & Significance
Conclusion


Medical Symptoms Text Classification
Yijia Li, Haiyan An, Wanyu Huang, Zhe Li, Lilin Huang
San Jose State University
05/02/2022
Background
Modeling
Deployment
Conclusion
Data Engineering
Evaluation