


Enhancing Baseball Strategy: Classifying Pitch Outcomes on Taken Pitches

DSCI 631 - 001
Caleb Miller
Hashim Afzal
Robert Logovinsky



Motivation

- Goal: Develop a classification model that will predict whether a pitch will be called a strike or ball when the batter does not swing at it
- Why is this important:
 - Allows teams to leverage predictions of calls for various applications:
 - Evaluating catcher framing
 - Making swing decisions
 - Analyzing umpire tendencies



Data Source

- This data was provided by the Philadelphia Phillies
- Dataset sourced from Baseball Savant
 - An MLB-owned platform
- The use of this data is for non-commercial and educational purposes only



Description of Dataset

The dataset contains 351,062 rows of the following columns:

Game and Player Identifiers	Contains unique identifiers for games, plate appearances, pitchers, batters, and catchers
Pitch Details	Pitch type, count of balls/strikes, pitch outcome, and zone
Player Attributes	Handedness of batter and pitcher
Pitch Location	Vertical / horizontal position of the ball and strike zone boundaries

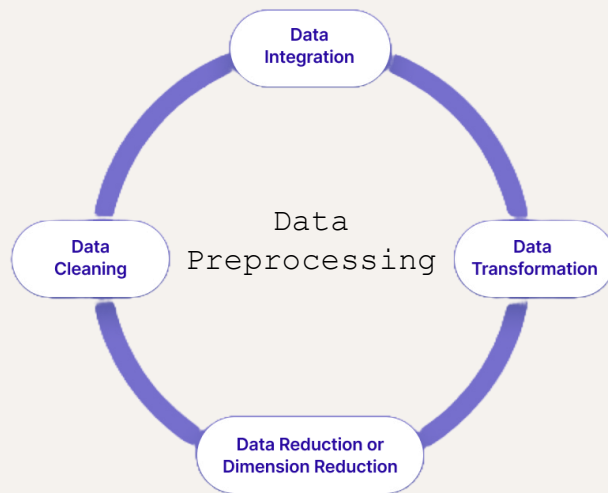
Evaluation Metrics

The strength of our classification models are measured by:

Accuracy	Measures the overall correctness of the model
Precision	Quantifies the model's ability to avoid false positives
Recall	Assesses the model's ability to identify all actual positive instances
F1	A combination of precision and recall that offers a balanced measure

Preprocessing

- 195 rows contained null values
 - This is less than 0.001% of the data, so these rows were dropped
- 'pitcher_name' is dropped
- Changed format of columns:
 - 'description'
 - 'Stand'
 - 'P_throws'
 - 'zone'



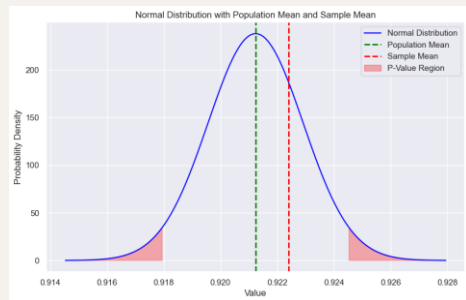
Exploratory Data Analysis Overview

- Statistical Analysis of:
 - Combination of Pitcher and Batter Orientation
 - Pitch Type
- Correct Call Percentage by Pitch Type
- Correct Call Percentage by Pitch Number
- Correct vs Incorrect Call with Standardized Strike Zone
- Percentage of Strikes Called by Pitch Zone

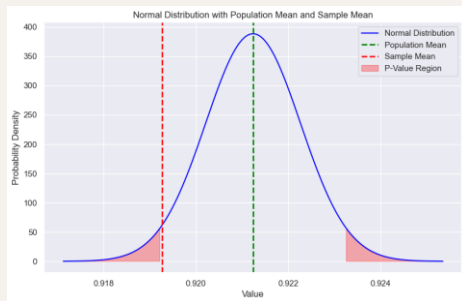


Combinations of Pitcher and Batter Orientation

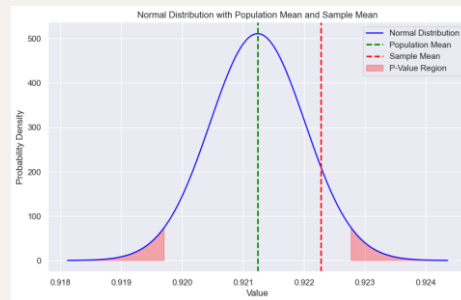
LHP → LHB



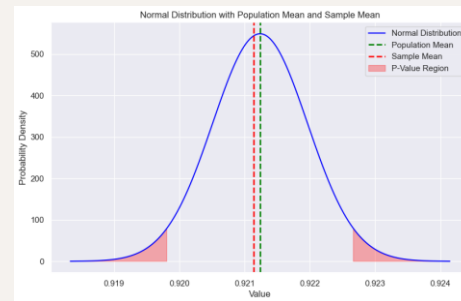
LHP → RHB



RHP → LHB

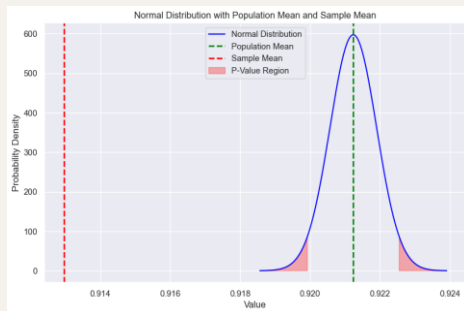


RHP → RHB

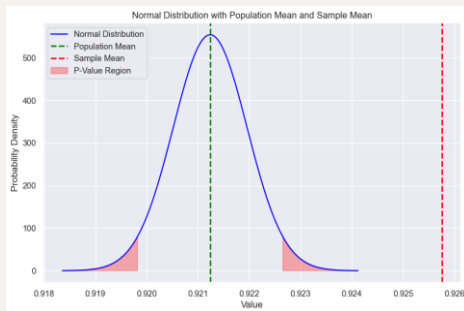


Pitch Type

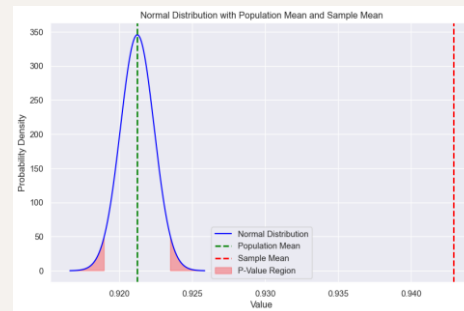
Fastball



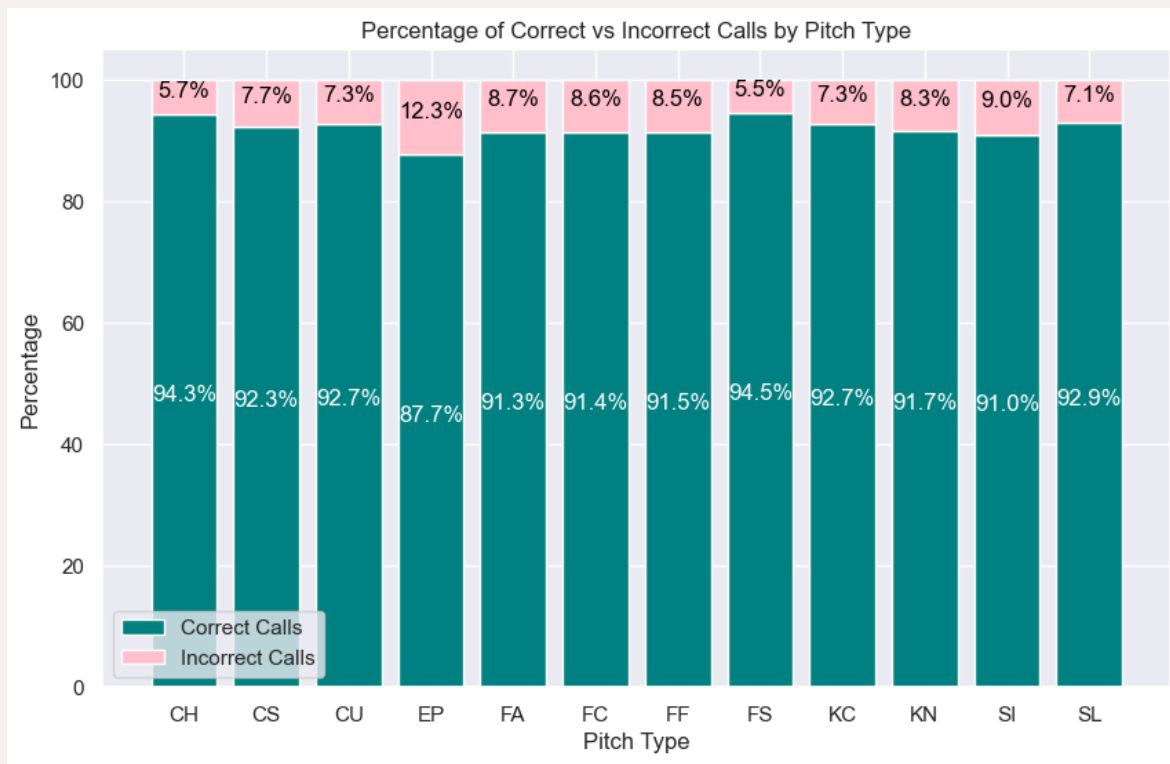
Breaking Ball



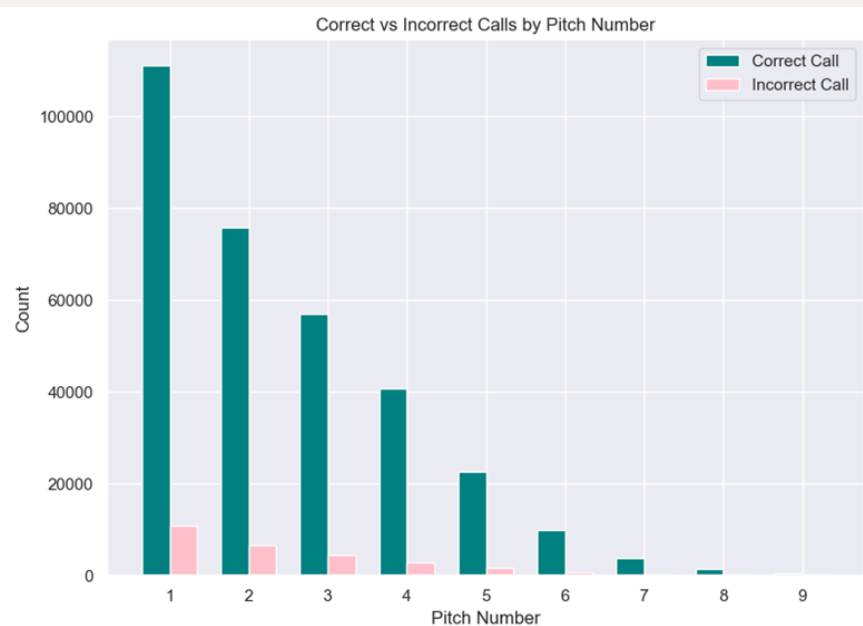
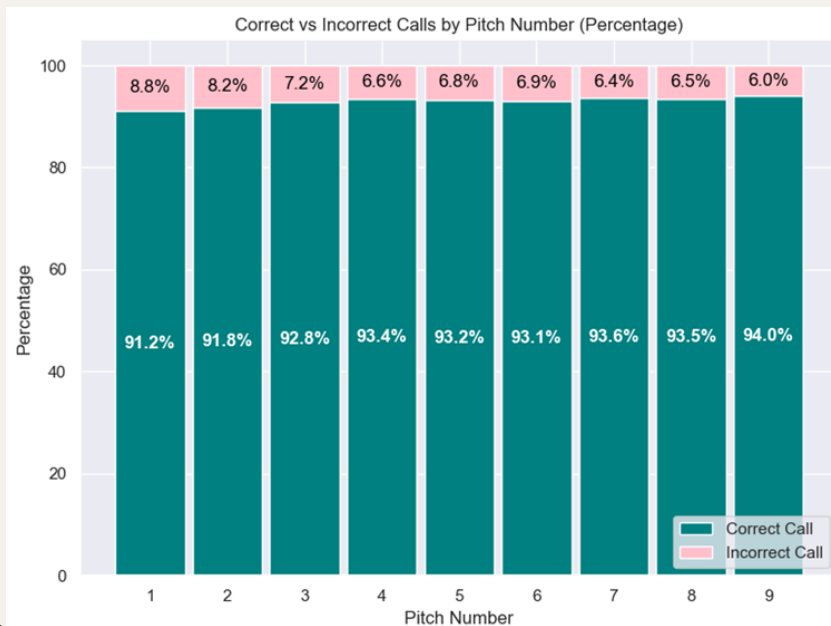
Offspeed



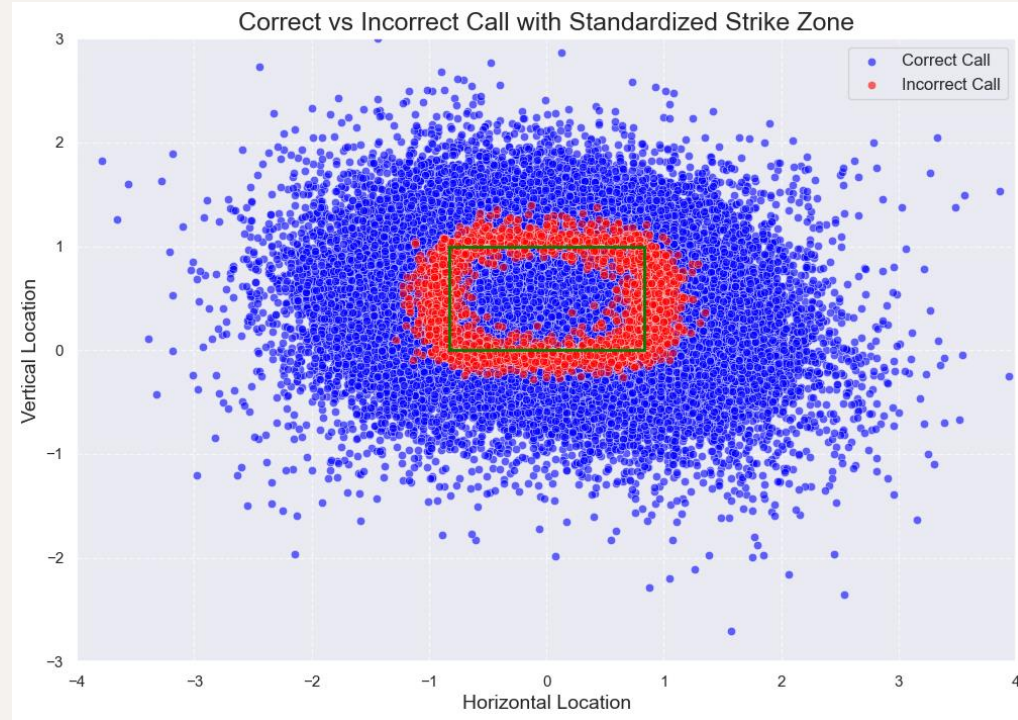
Correct Call Percentage by Pitch Type



Correct Call Percentage by Pitch Number

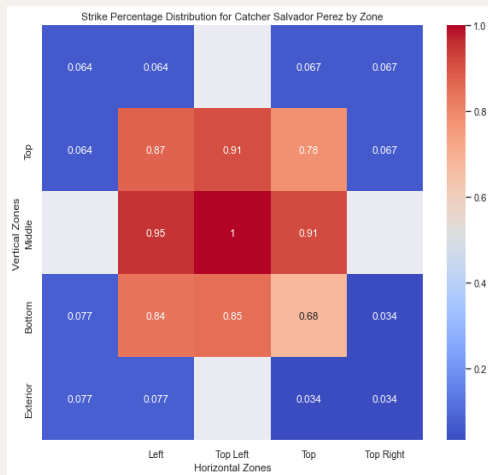


Correct and Incorrect Calls in a Standardized Strike Zone

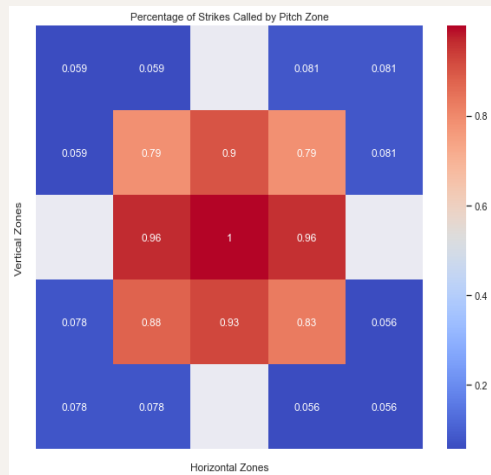


Percentage of Correct Call by Strike Zone

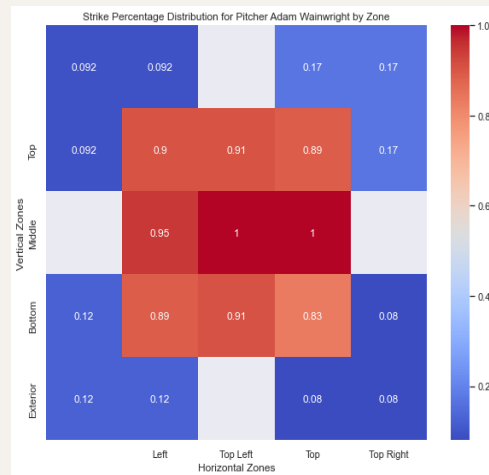
Salvador Perez



Overall



Adam Wainwright



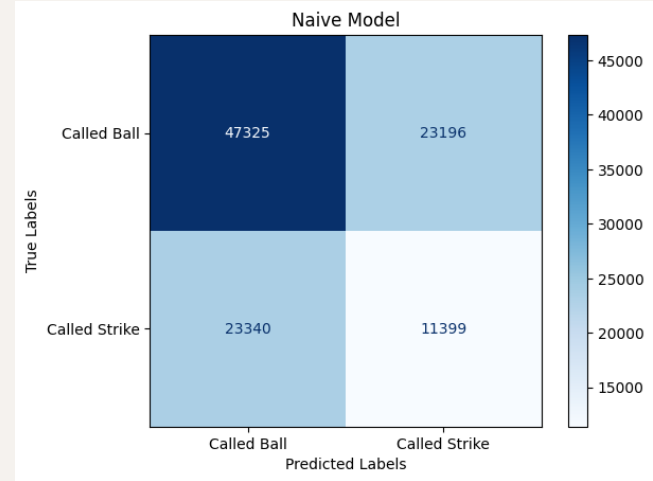
Modeling Overview

Naive Model	True baseline with no machine learning ability
Logistic Regression	Simple, easy to interpret, and strong baseline for classification
XGBoost Classifier	Able to handle nonlinear data, weigh feature importance, and resist overfitting
MLP Classifier	Able to model complex, nonlinear relationships - can provide stronger results with more data

Naive Model

Classification Report:

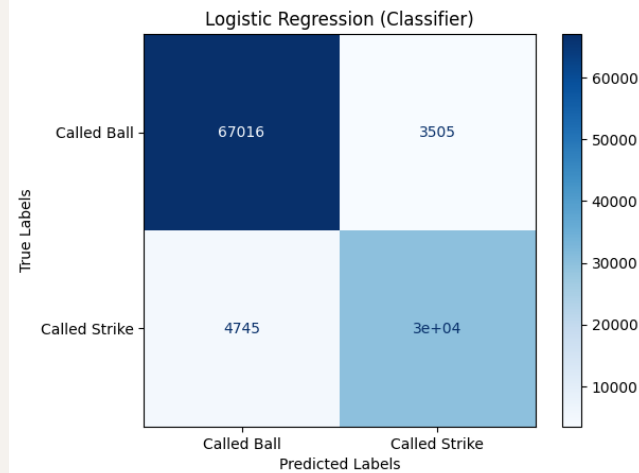
	precision	recall	f1-score	support
Called Ball	0.67	0.67	0.67	70521
Called Strike	0.33	0.33	0.33	34739
accuracy			0.56	105260
macro avg	0.50	0.50	0.50	105260
weighted avg	0.56	0.56	0.56	105260



Logistic Regression

Classification Report:

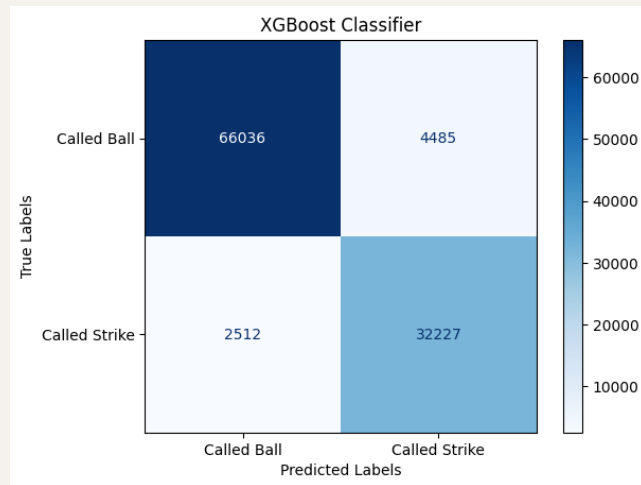
	precision	recall	f1-score	support
Called Ball	0.93	0.95	0.94	70521
Called Strike	0.90	0.86	0.88	34739
accuracy			0.92	105260
macro avg	0.91	0.91	0.91	105260
weighted avg	0.92	0.92	0.92	105260



XGBoost Classifier

Classification Report:

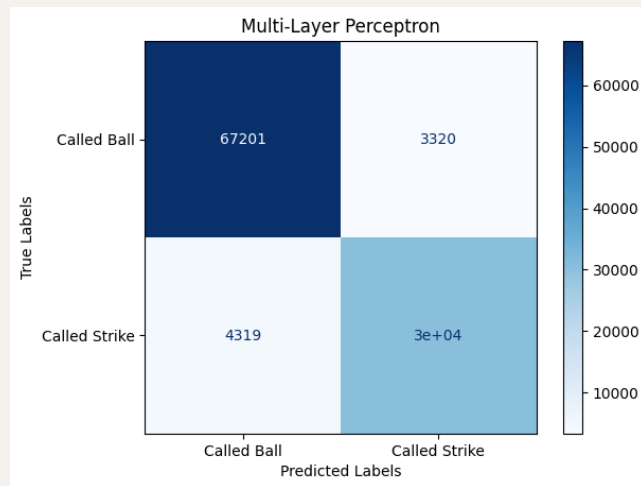
	precision	recall	f1-score	support
Called Ball	0.96	0.94	0.95	70521
Called Strike	0.88	0.93	0.90	34739
accuracy			0.93	105260
macro avg	0.92	0.93	0.93	105260
weighted avg	0.94	0.93	0.93	105260



Multi-Layer Perceptron Classifier

Classification Report:

	precision	recall	f1-score	support
Called Ball	0.94	0.95	0.95	70521
Called Strike	0.90	0.88	0.89	34739
accuracy			0.93	105260
macro avg	0.92	0.91	0.92	105260
weighted avg	0.93	0.93	0.93	105260



Summary of Results

Model	Accuracy	Precision	Recall	F1
Naive	0.56	0.33	0.33	0.33
Logistic Regression	0.92	0.90	0.86	0.88
XGBoost Classifier	0.93	0.88	0.93	0.90
MLP Classifier	0.93	0.90	0.88	0.89

Best Model

- Best Overall Model: XGBoost Classifier
- Best Model for Precision: Logistic Regression and MLP Classifier
- Best Model for Recall: XGBoost Classifier
- The choice of model should be guided by the requirements of the application and the relative importance of precision vs. recall



Challenges and Limitations

- Imbalance of the dataset
 - 66.99% is labeled as a 'ball'
- Model interpretability
 - XGBoost and MLP Classifier can offer high accuracy, but are difficult to interpret
- Models are only as good as the data they are trained on
 - May not be accurate with new data



Recommendations / Future Work

- Test oversampling, undersampling, and SMOTE to see what provides the best results
- Explore additional features that could impact pitch calls
 - Score of the game, weather, umpire, number in attendance, etc.
- Implement cross validation techniques
 - Will generate a more robust model
- Update with new data to help model adapt to any changes in umpiring patterns or rules

Conclusion

- EDA gave us a significantly better sense of which features are important to the models
 - i.e. pitch type, pitch location
- Both XGBoost and MLP Classifiers demonstrated improved predictions compared to umpires
 - This is a step towards automating umpire decision making
- Introducing new data would be a great test of how robust our models are
- The project showed the potential of ML to enhance baseball strategy

Workload Distribution

<u>Caleb</u>	<u>Hashim</u>	<u>Robert</u>
<ul style="list-style-type: none">● Proposal● Visualization● Modeling● ReadME● Report● Presentation	<ul style="list-style-type: none">● Proposal● Preprocessing● Visualization● Presentation	<ul style="list-style-type: none">● Proposal● Preprocessing● Visualization● Presentation