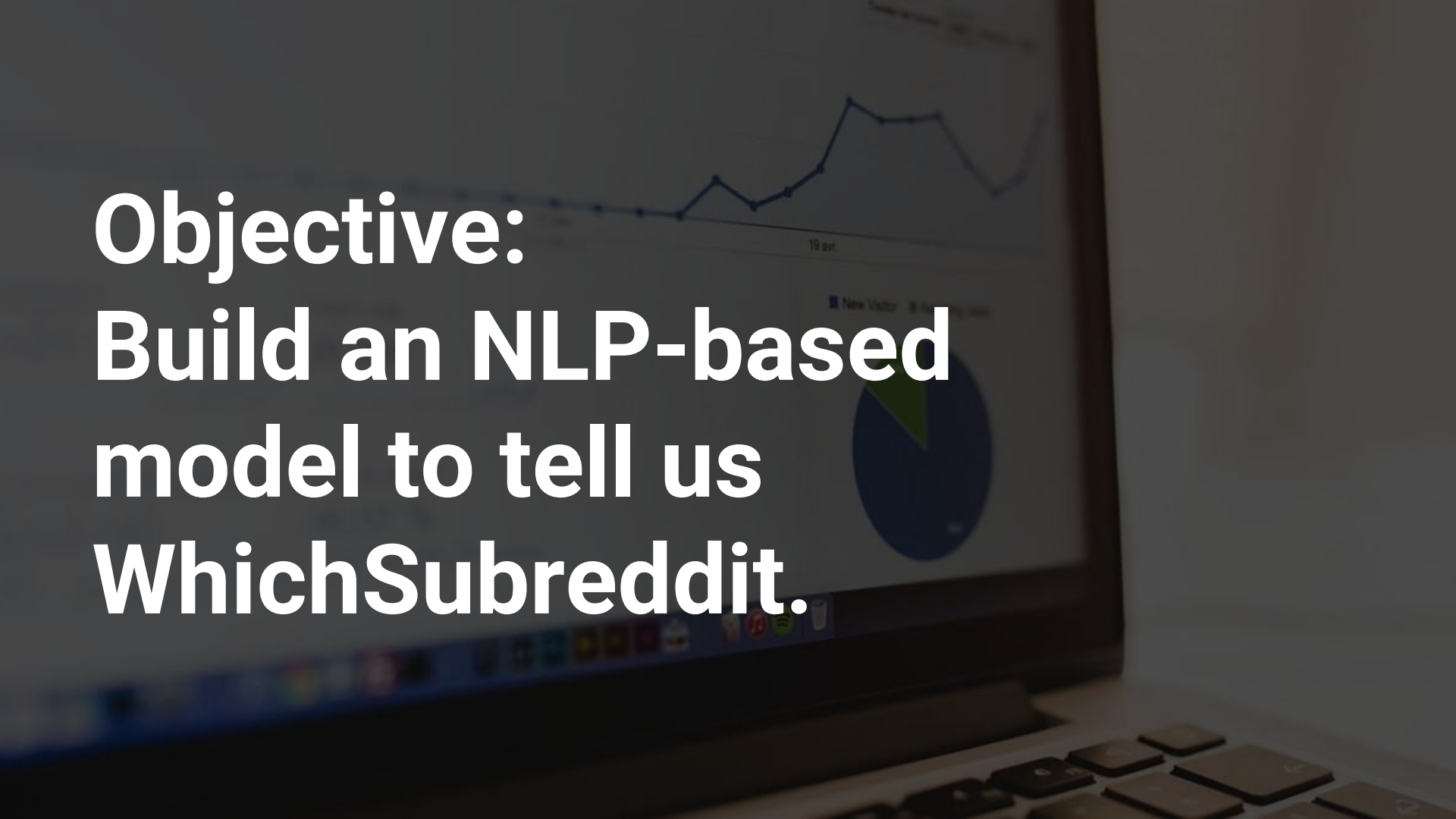


# WhichSubreddit?

NLP Analysis to Unify Specialists



**Objective:**  
**Build an NLP-based**  
**model to tell us**  
**Which Subreddit.**

The background image shows a laptop screen with a dark overlay. On the screen, there is a line graph with a blue line and a pie chart with a blue and green segment. The text is overlaid on the left side of the screen in a bold, white, sans-serif font.

# The problem

Specialists speak different languages. How can we facilitate productive communication across disciplines?

Welcome 欢迎光临 Bienvenue

Bienvenidos  Willkommen

Добро пожаловать Hoş geldiniz

Benvenuti Welkom Dobrodošli

歡迎光臨 Bem-vindo ようこそ

Bonvenon Witamy أهلاً وسهلاً

Aloha Selamat datang ברוך הבא

Được tiếp đãi ân cần 환영합니다

A close-up photograph of a person's hand holding a purple marker, writing on a white surface. The background is blurred, showing some bokeh lights. The text 'The solution' is overlaid in white on the left side of the image.

# The solution

1866 posts collected from two contrasting subreddits (datascience and genetics).

52.7% of posts were from the datascience subreddit.

75% of all posts were used to train the data, reserving 25% to test our model.

93.6% Accuracy obtained using logistic regression. All models were overfit.

## Comparing Models

Model	Train Accuracy	Test Accuracy	Misclassification Rate	Misclassified as Datascience	Misclassified as Genetics
Logistic Regression	99.3%	93.6%	6.4%	25	5
K Nearest Neighbors	96.6%	92.5%	7.5%	19	16
Decision Tree	100.0%	89.1%	10.9%	34	17
Random Forest	99.9%	89.5%	10.5%	29	20
Adaboost	95.6%	90.4%	9.6%	29	16

**Note:** The above models have not undergone gridsearching or optimized in any other manner.

# Room to Improve...

Use non-text features to help differentiate between Subreddits (e.g., number of comments, time since posting, etc.)

Optimize models by tuning hyperparameters

Audience suggestions?

