# Summary:

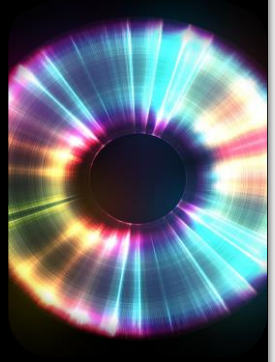
## *Phrases Most Associated with Chest Complaint Discharges*


Chart reviews are a commonplace procedure to maintain quality, compliance, ethics, and consistency of patient medical records. While lots of information can be extracted from structured fields, free-text notes, written by the clinician, can offer invaluable new, unique, or additional insights.

In this case, NLP Discriminative Feature Selection is applied to identify top 100 phrases associated with discharges involving chest complaints. The model can be used for

1. Understanding complaints, conditions, etc. most likely associated with admissions involving chest complaints.
2. Differentiating between cardiac and non-cardiac related chest complaints.
3. Applying these results to the next step of topic modeling and visualizing results in a dashboard for discussion and action planning around this important medical condition.


Repository on GitHub: https://github.com/olga12kz-DS/Discriminative-Feature-Selection-Discharge-Note



# Data Preparation & Feature Engineering

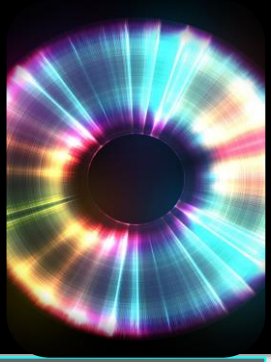
Extract “Discharge Notes” field from the full dataset using **list comprehension and regex pattern in a user defined function.**

Clean up the text by removing unnecessary characters and extra spaces using **regex**.

**Create a label column** to apply supervised learning **(Chi-square model).**

Create **pipelines** to process the text and **display results** in structured and sorted format.

# Optimizing Chi-Square Test Model for Discriminative Feature Selection



## Chi-Square Test

1. Measures how strongly each feature (word or an n-gram) is associated with either class (binary classification), producing a ranked list of most discriminative features.

2. In NLP, relies on CountVectorizer which provides frequency of features by converting text to numbers.

3. Requires class labels as a supervised learning model.

4. Computes Chi-square scores to separate high-signal features from high-noise ones.

5. Allows to fine-tune results to single words or phrases.

## Best Performance | Example of Output

Results filtered to top 100* and at min 2-word phrases

| Feature (phrase) | Chi-Squared Score |
|---|---|
| Substernal chest pain | 8162 |
| Chest pressure | 7845 |
| Stress test | 6117 |
| Chest discomfort | 5647 |
| Cardiac catheterization | 4897 |

**Only top 5 shown for presentation purposes*