



Advancing science for life™



# AI.PACA: Patient Assisted Clinical Assessment

Patient States Free-Text LLM

Ignacio Villasmil

Data Science R&D Discovery Intern – Neuromodulation

August 15th, 2023

**GOAL:**

Determine the patient's current state based off their free-text responses

**HOW:**

- 1) Feature engineering of patient responses via text-classification model
- 2) Clustering model on the engineered feature set for determining patient states  
(independent of IBM)

# Feature Engineering Pipeline



# Feature Engineering Pipeline

Boston  
Scientific

- How was your day today? Can you share how the recommendations from the app may have made a difference?
- How was your day (e.g. activity, sleep, pain, medications, mood, pain intensity)?

## Free-Text Passages from Patients

"i had a very nice day went for a walk up to a friends house to visit it was very nice visit not too much pain had some but it wasnt bad"

"tense! go biden!"

"didnt feel very well. hurt a lot today."

## Relevance Filter



## Useful Passages

"i had a very nice day went for a walk up to a friends house to visit it was very nice visit not too much pain had some but it wasnt bad"

"didnt feel very well. hurt a lot today."

## Text Classification



**Pain Score**

**Pain Complexity Score**

**Medication Score**

**Sleep Score**

**Mood Score**

**Anxiety Score**

**Activities Score**

**Socialization Score**

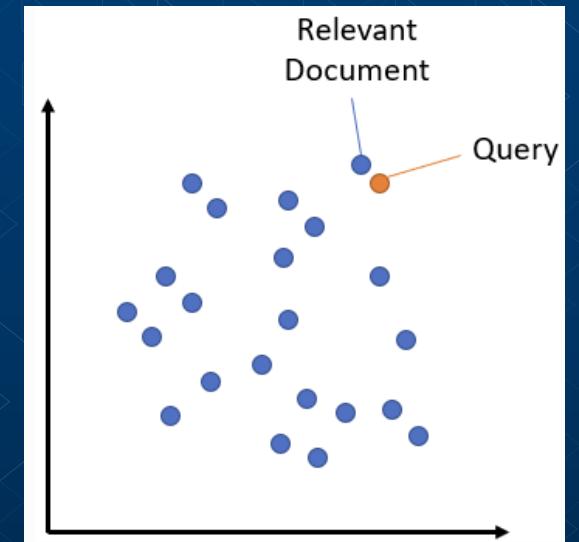
# Relevance Filtering: Asymmetric Semantic Search Sentence Transformer



# Relevance Filtering: Asymmetric Semantic Search Sentence Transformer

Boston  
Scientific

- Used for the filtering of relevant concepts for each patient response by calculating relevancy scores
- Semantic Search (how it works):
  - Create embeddings for all sentences in the passages
  - Create embedding for question being asked in the same vector space (e.g. "How well has the patient been sleeping")
  - Compute cosine similarity to determine semantic overlap
  - A high score represents a good answer for the query (a relevant block of text)

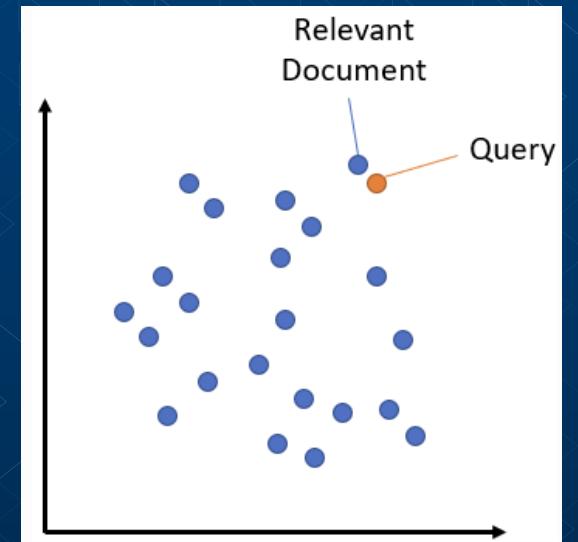




# Relevance Filtering: Asymmetric Semantic Search Sentence Transformer

Boston  
Scientific

- Trained on 215M question-answer pairs from StackExchange, Yahoo, Answers, Google search queries, and Bing search queries
- Filtering via Semantic Search:
  - Semantic Search was performed for each concept, first, on manually created and labeled sentences (these manually created sentences can be thought of as a training set)
  - The resulting scores were used to determine a reasonable score threshold for a concept to be considered relevant to a sentence

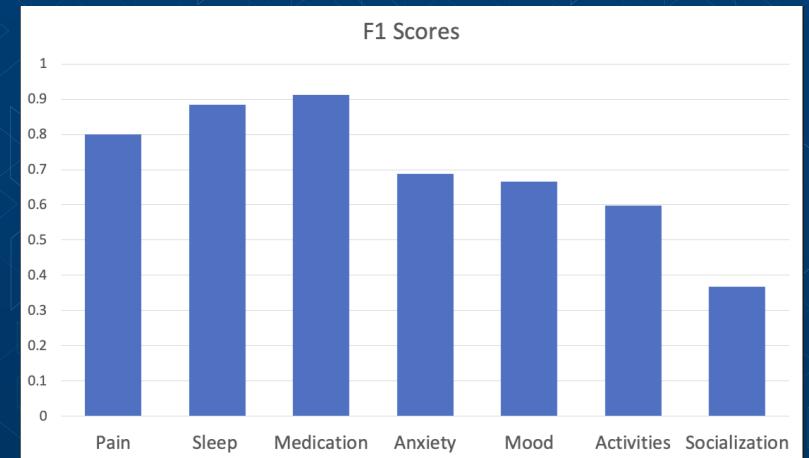




# Semantic Relevance Thresholds

Boston  
Scientific

- For each concept, the thresholds were optimized to result in the largest F1 Score possible
  - F1 Score = harmonic mean of precision & recall
    - Precision = proportion of positive-assigned results that were correct
    - Recall = proportion of actual positives correctly identified
- Socialization feature was not filtered for relevancy



NOTE: Mood & Activities both had multiple queries, and therefore, multiple thresholds (one for each query)

→ If at least one query's threshold was met, then the sentence was marked as relevant



# Semantic Relevance Thresholds

Boston  
Scientific

- F1 Score of Thresholds:
  - Pain Relevance Threshold F1 Score = 0.8
  - Sleep Relevance Threshold F1 Score ≈ 0.8837
  - Medication Relevance Threshold F1 Score ≈ 0.9130
  - Anxiety Relevance Threshold F1 Score ≈ 0.6885
  - Mood Relevance Threshold F1 Score ≈ 0.6667
  - Activities Relevance Threshold F1 Score ≈ 0.5986
  - Socialization Relevance Threshold F1 Score ≈ 0.3684
    - This is why socialization was not filtered for relevancy score

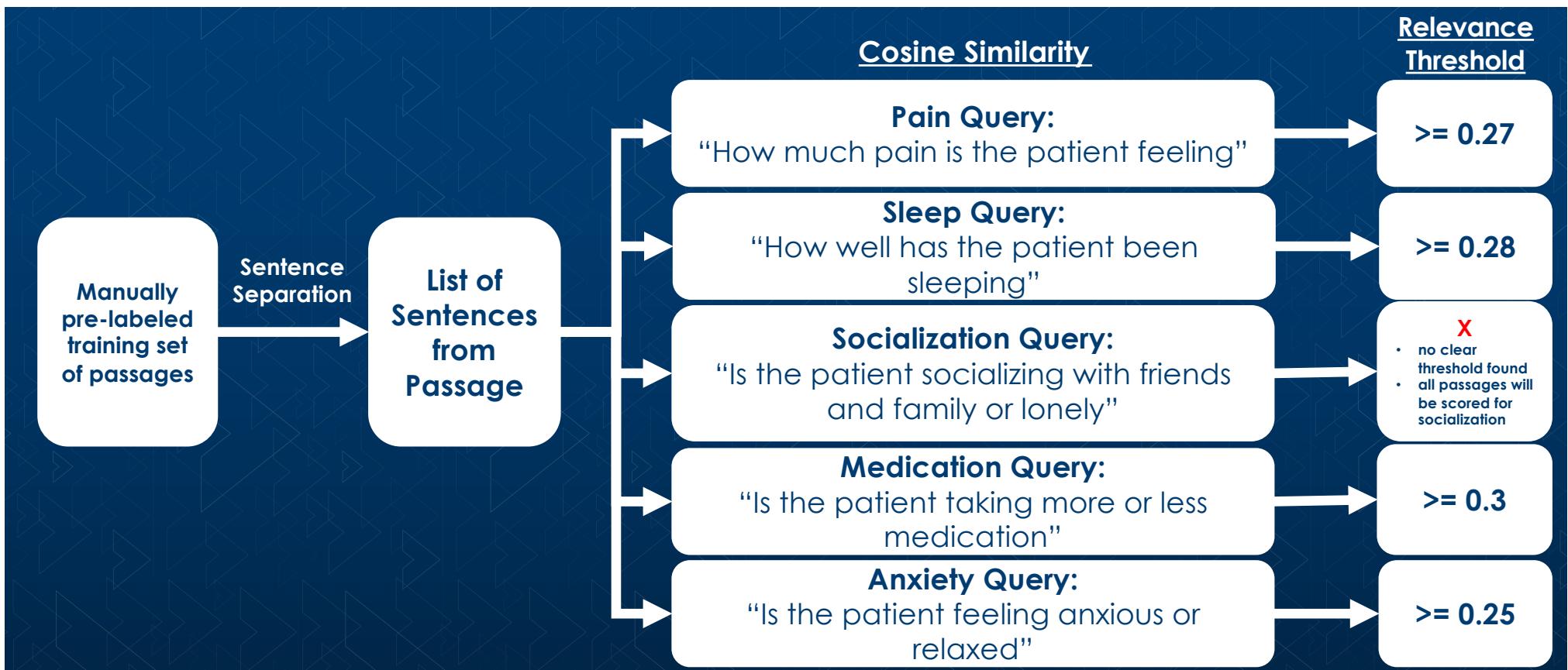
NOTE: Mood & Activities both had multiple queries, and therefore, multiple thresholds (one for each query)

→ If at least one query's threshold was met, then the sentence was marked as relevant



# Filtering of Relevant Concepts (Asymmetric Semantic Search)

Boston  
Scientific





## Filtering of Relevant Concepts (Asymmetric Semantic Search)

Boston  
Scientific

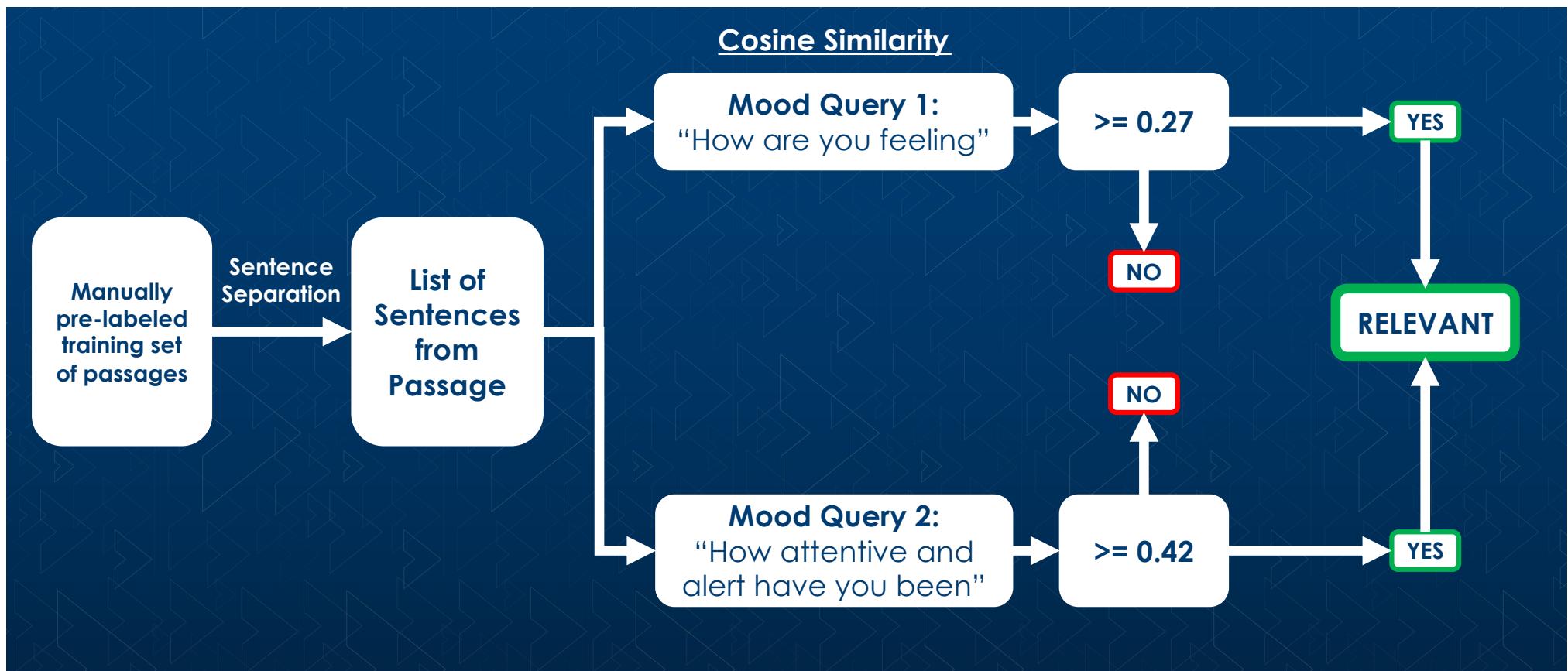
- Similar to the results for socialization, the Mood and Activities concepts also struggled to find a good threshold:
  - Initial Mood Query: “How is the patient’s mood and attentiveness”
  - Initial Activities Query: “Has the patient been active”
- New queries were then attempted for both after looking into the clinical labels to be used for classification of these two concepts:
  - The clinical mood labels to be used for text classification tend to describe either mood or attentiveness in more detail
  - The clinical activities labels to be used for text classification tend to describe a variety of specific activities of interest

→ **For both Mood & Activities, multiple queries were used to find relevancy**



# Filtering of Mood Concept (Asymmetric Semantic Search)

Boston  
Scientific



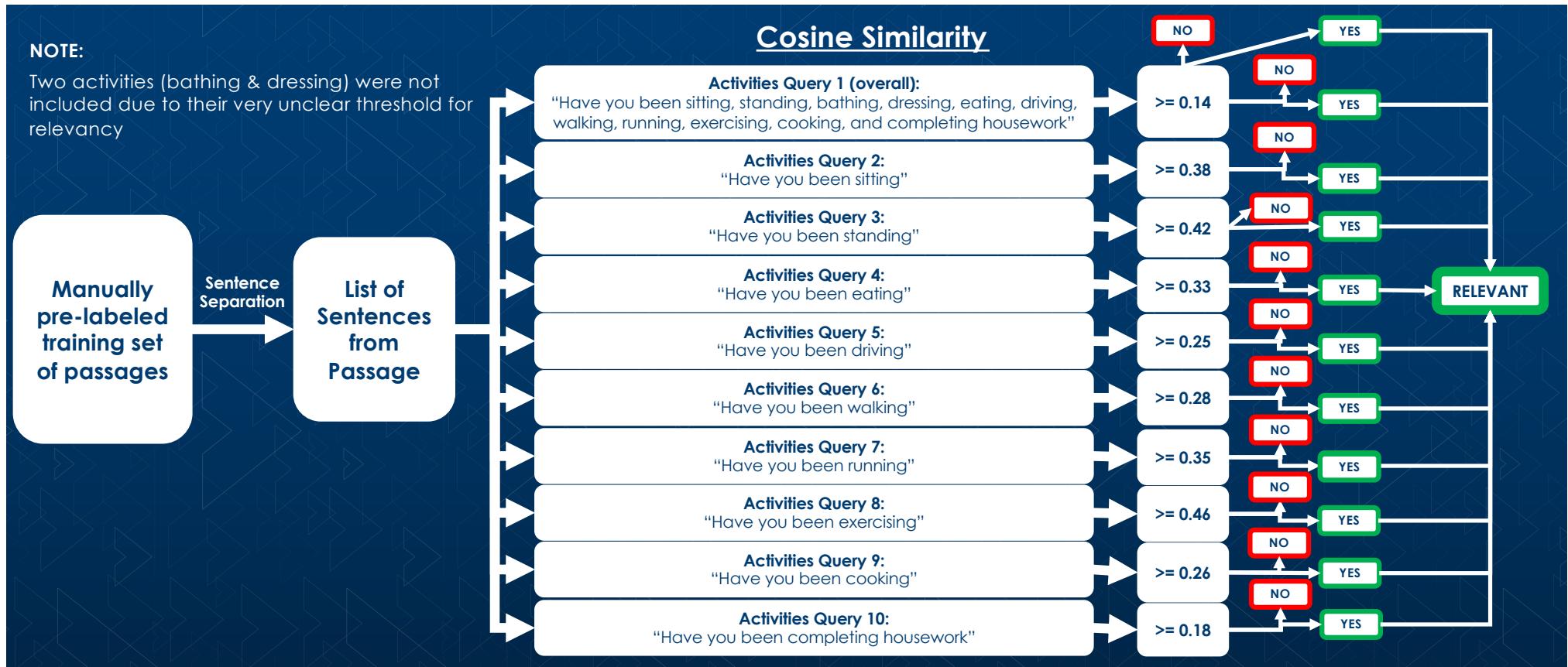


# Filtering of Activities Concept (Asymmetric Semantic Search)



## **NOTE:**

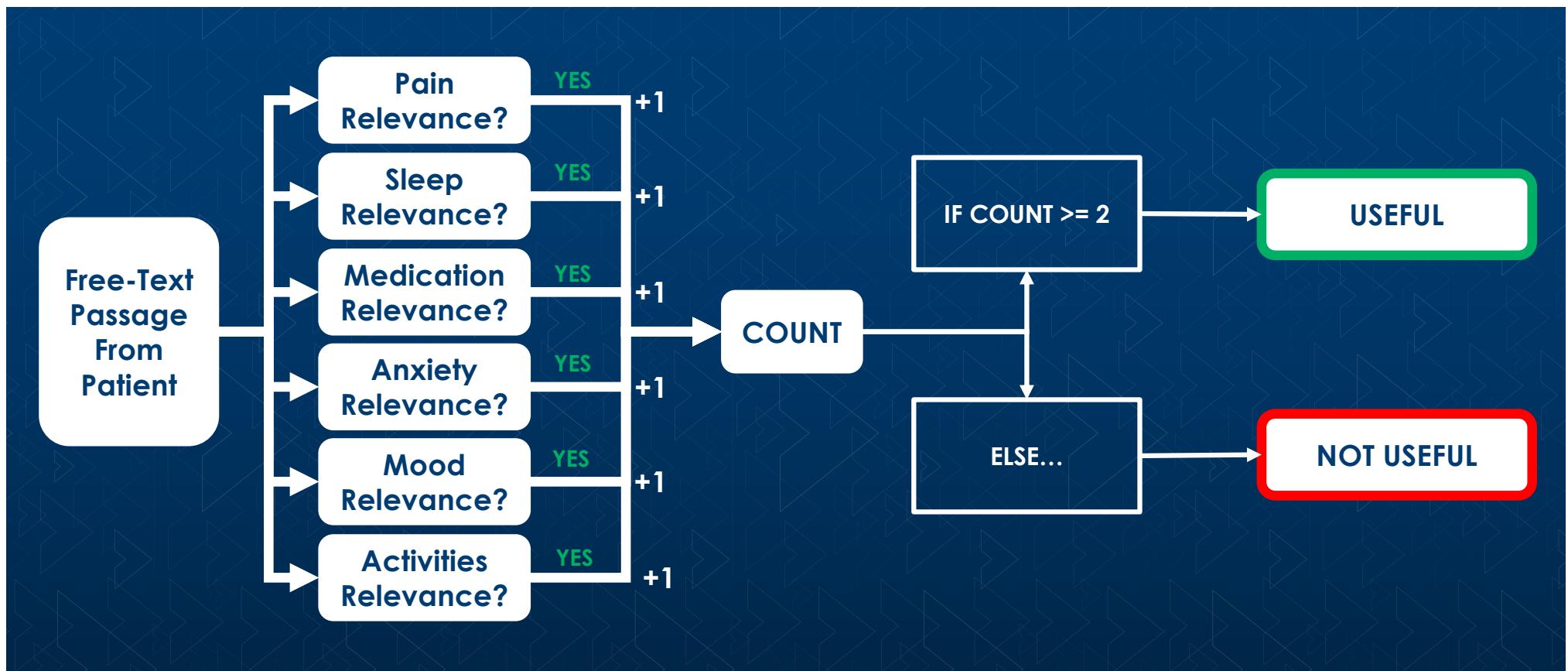
Two activities (bathing & dressing) were not included due to their very unclear threshold for relevancy





# Filtering For Useful Passages

Boston  
Scientific



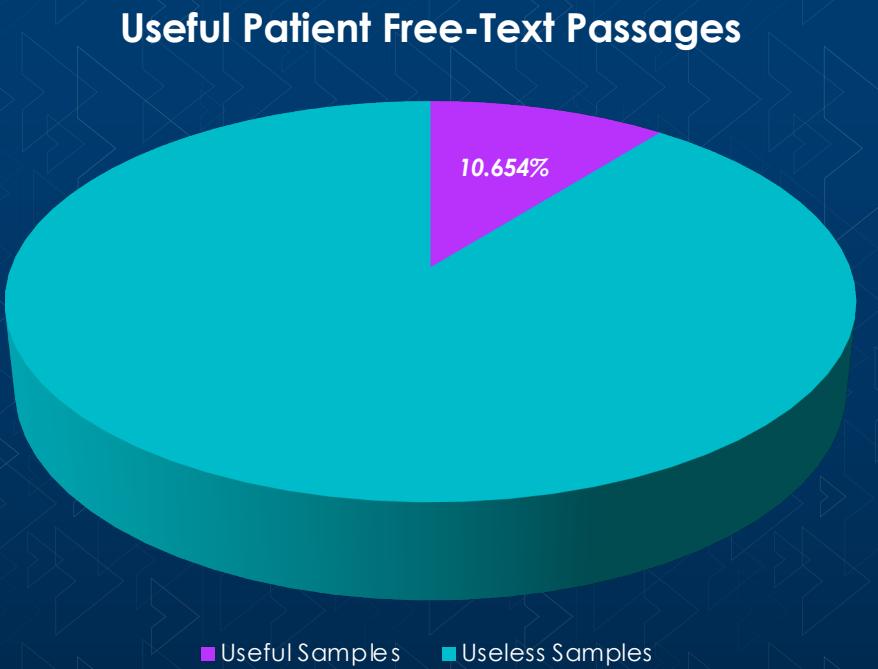
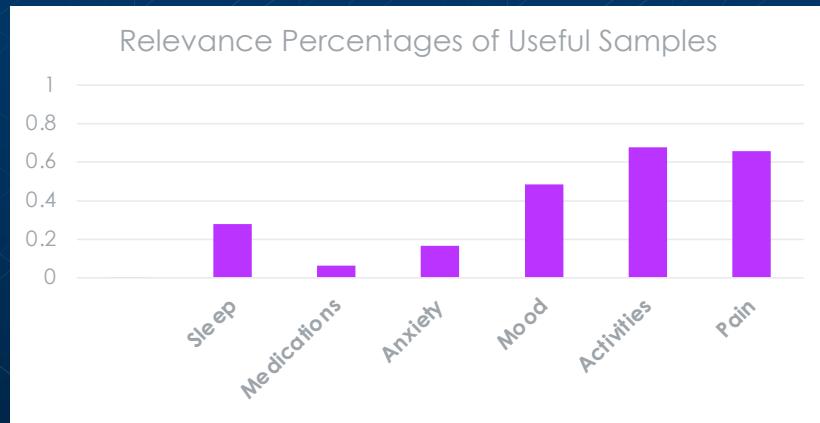


# Filtering For Useful Passages:

Boston  
Scientific

- Results from Filtering Useful Passages:

- About 10.654% of all original 12,080 patient free-text responses were found to be useful (with at least two relevant concepts detected)



# Text Classification To Generate Concept Scores (BART MNLI Model)





# BART Large MNLI (Multi-Genre Natural Language Inference)

Boston  
Scientific

- Natural Language Inference – determining whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) given a "premise"
- Multi-Genre Natural Language Inference (MultiNLI) corpus is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information.

Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction	The man is sleeping.
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	entailment	Some men are playing a sport.

- Key Method: **NLI-based Zero Shot Text Classification**

- To determine whether a sequence belongs to a class, a hypothesis is constructed for the class, and the probabilities of entailment and contradiction are then converted into label probabilities for that class
    - e.g. If we want to determine whether a sequence belongs to the class "politics", we would construct a hypothesis "This text is about politics", and the probabilities of entailment and contradiction are converted into label probabilities



# Convert Clinical Scales to Free-text Features

Boston  
Scientific

## The GAD-7 Scale [18]

### Generalized Anxiety Disorder 7-item (GAD-7)

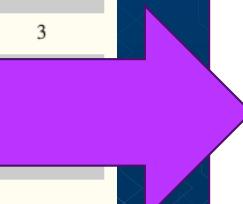
Over the last 2 weeks, how often have you been bothered by the following problems	Not at all	Several days	More than half the days	Nearly every day
1. Feeling nervous, anxious or on edge	0	1	2	3
2. Not being able to stop or control worrying	0	1	2	3
3. Worrying too much about different things	0	1	2	3
4. Trouble relaxing	0	1	2	3
5. Being so restless that it is hard to sit still	0	1	2	
6. Becoming easily annoyed or irritable	0	1	2	
7. Feeling afraid as if something awful might happen	0	1	2	

GAD-7 score obtained by adding score for each question (total points).

A score of 8 points or higher is a reasonable cut-off for needing further evaluation to determine presence and type of anxiety disorder<sup>23, 24</sup>

The following cut-offs correlate with level of anxiety severity:

Score 0-4	: Minimal Anxiety
Score 5-9	: Mild Anxiety
Score 10-14	: Moderate Anxiety
Score 15 or greater	: Severe Anxiety



### Zero-Shot Classification

Examples ▾

very hectic for i have been diagnosed with kidney cancer and it seems i am the only one worried about it. very frustrating

Possible class names (comma-separated)

feeling nervous, not being able to stop or control worrying, worrying too much about

Allow multiple true classes

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.593 s

not being able to stop or control worrying	0.996
worrying too much about different things	0.982
feeling nervous	0.959
trouble relaxing	0.692
•	
•	
•	

[List abbreviated for presentation]



# Text Classification to Generate Concept Scores (BART MNLI Model)

Boston  
Scientific

## Multi-Label Text Classification

Zero-Shot Classification Examples ▾

very hectic for i have been diagnosed with kidney cancer and it seems i am the only one worried about it. very frustrating

Possible class names (comma-separated)  
feeling nervous, not being able to stop or control worrying, worrying too much about

Allow multiple true classes

**Compute**

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.593 s

not being able to stop or control worrying	0.996
worrying too much about different things	0.982
feeling nervous	0.959
trouble relaxing	0.692

Maximize

</> JSON Output

## Single Label Text Classification

Zero-Shot Classification Examples ▾

it seemed to make me tire more easily and the pain was more intense today.

Possible class names (comma-separated)  
feeling little pain, feeling moderate pain, feeling a lot of pain

Allow multiple true classes

**Compute**

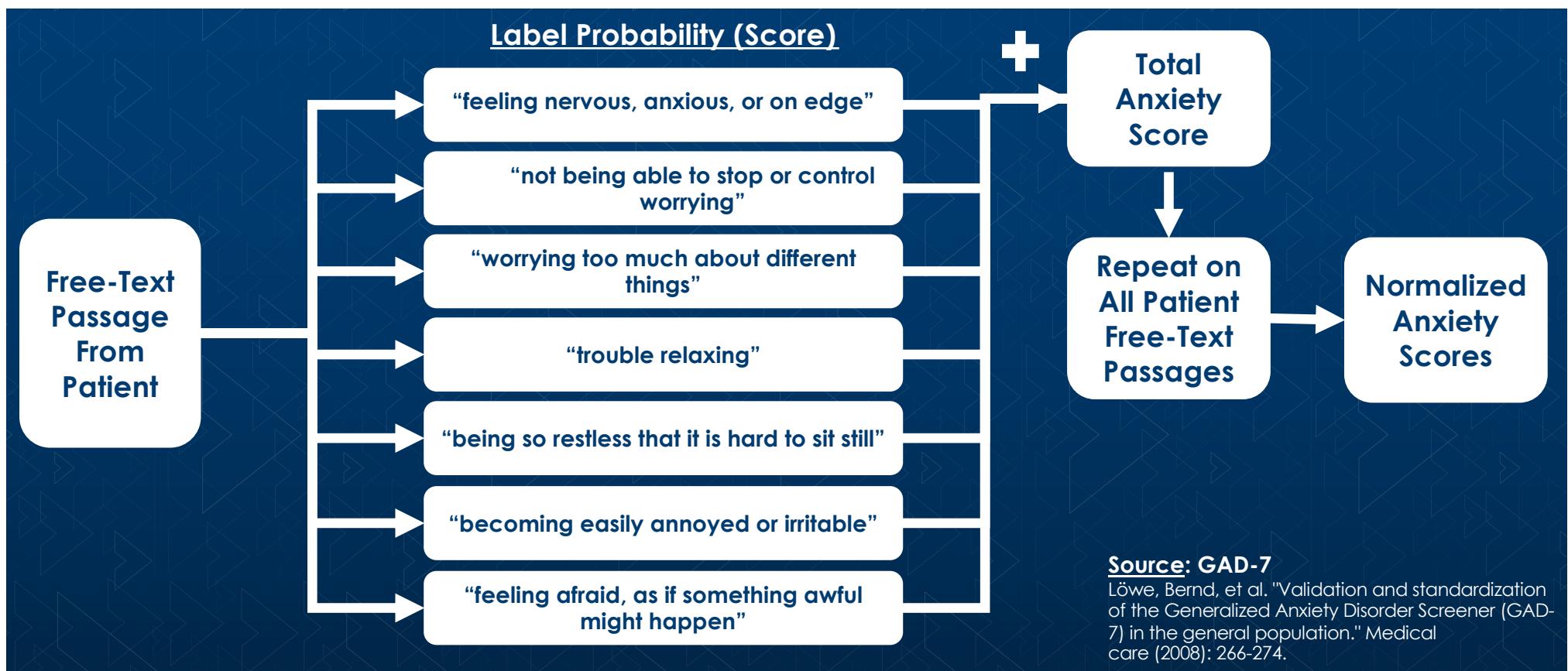
Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.423 s

feeling a lot of pain	0.984
feeling moderate pain	0.011
feeling little pain	0.006



# Multi-label Text Classification: Clinical Labels for Anxiety

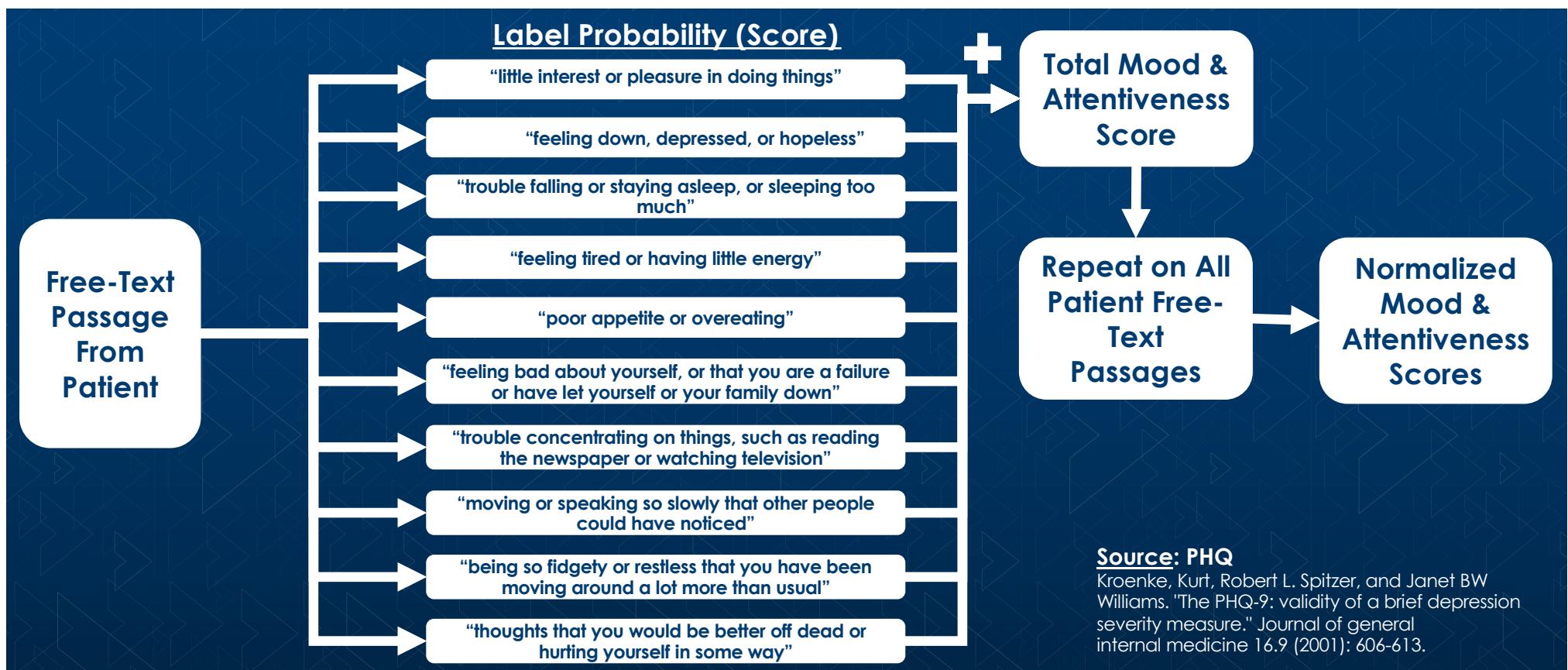
Boston  
Scientific





# Multi-label Text Classification: Clinical Labels for Mood & Attentiveness

Boston  
Scientific



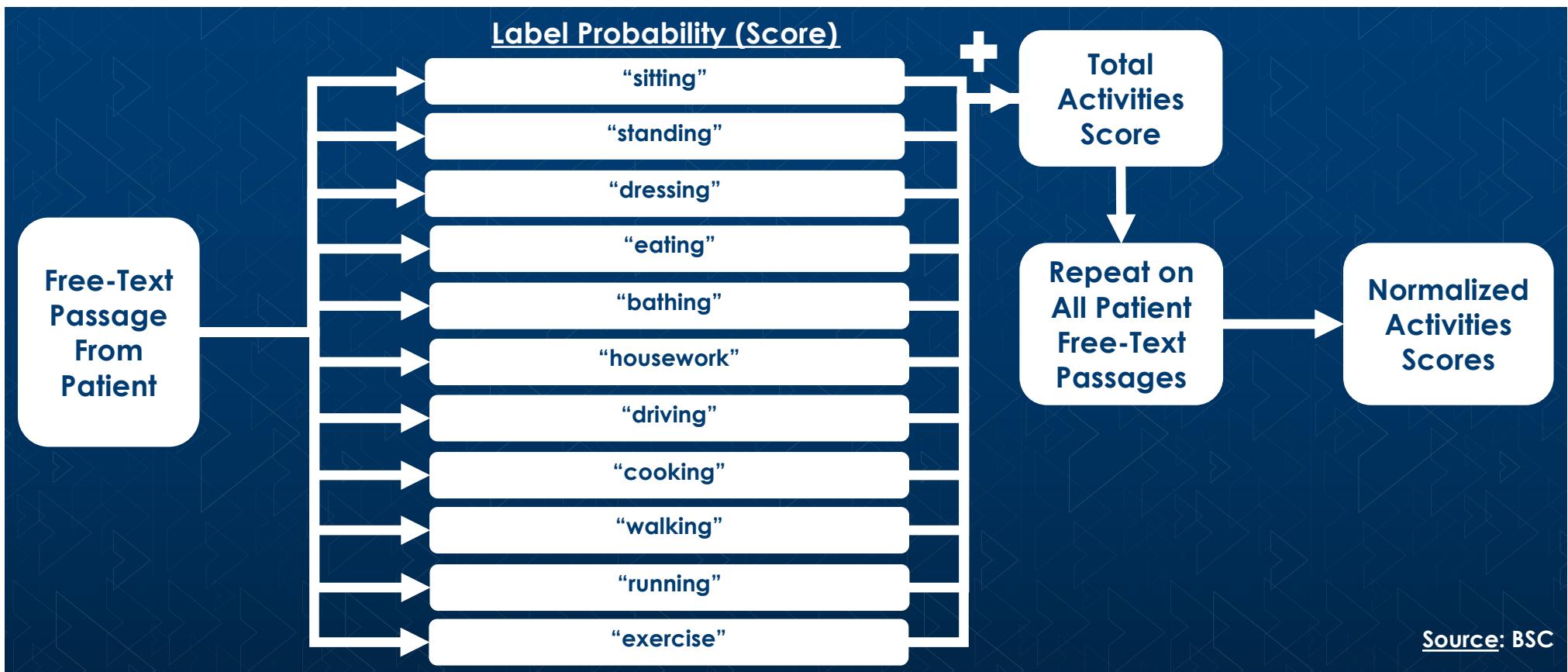
## Source: PHQ

Kroenke, Kurt, Robert L. Spitzer, and Janet BW Williams. "The PHQ-9: validity of a brief depression severity measure." Journal of general internal medicine 16.9 (2001): 606-613.



# Multi-label Text Classification: Clinical Labels for **Activities**

Boston  
Scientific

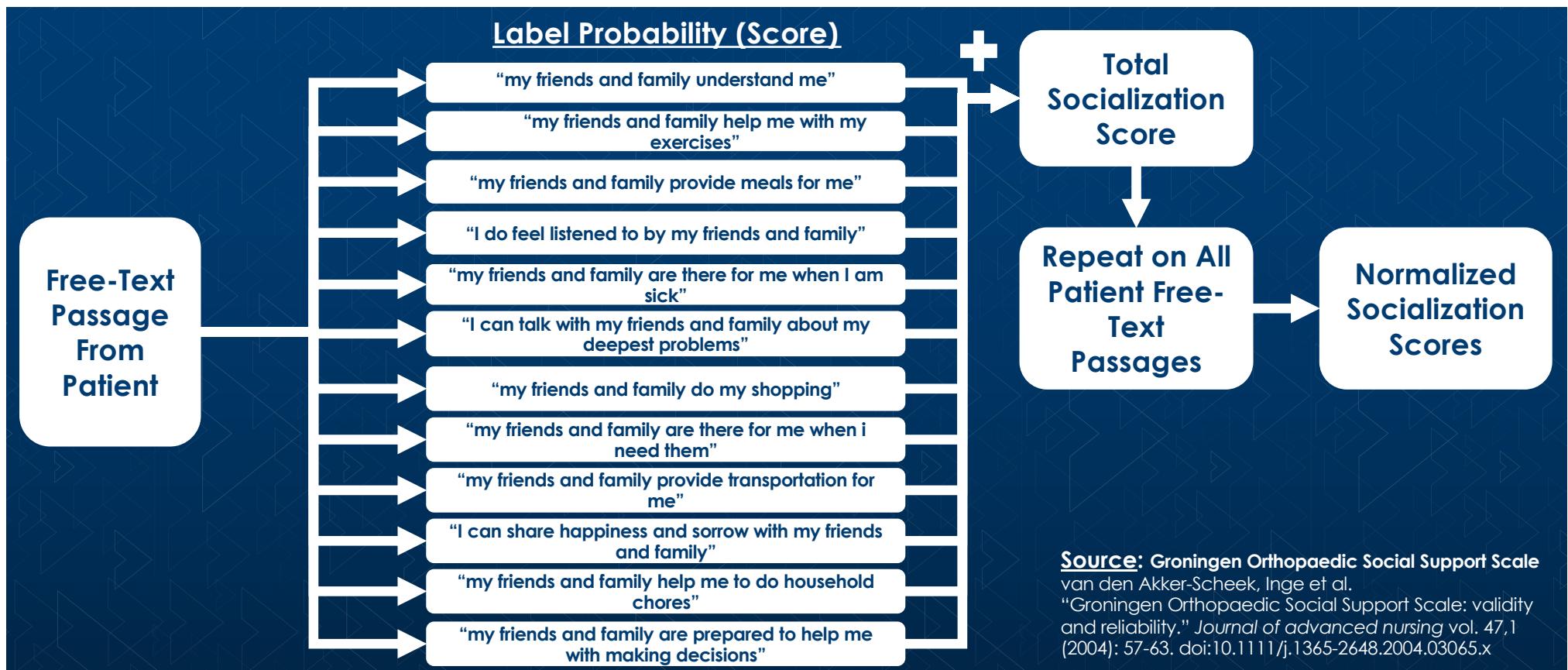


Source: BSC



# Multi-label Text Classification: Clinical Labels for Socialization

Boston  
Scientific

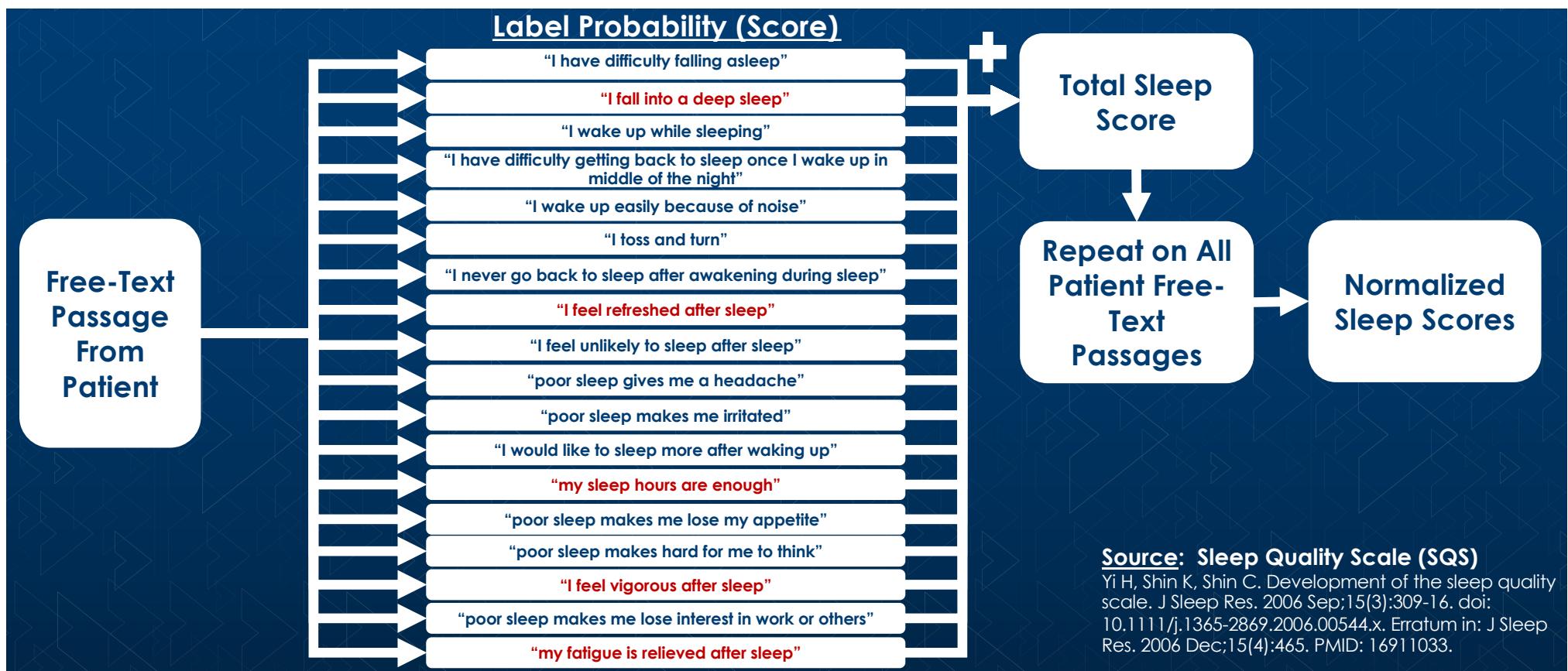


**Source:** Groningen Orthopaedic Social Support Scale  
van den Akker-Scheek, Inge et al.  
"Groningen Orthopaedic Social Support Scale: validity and reliability." *Journal of advanced nursing* vol. 47, 1 (2004): 57-63. doi:10.1111/j.1365-2648.2004.03065.x



# Multi-label Text Classification: Clinical Labels for Sleep

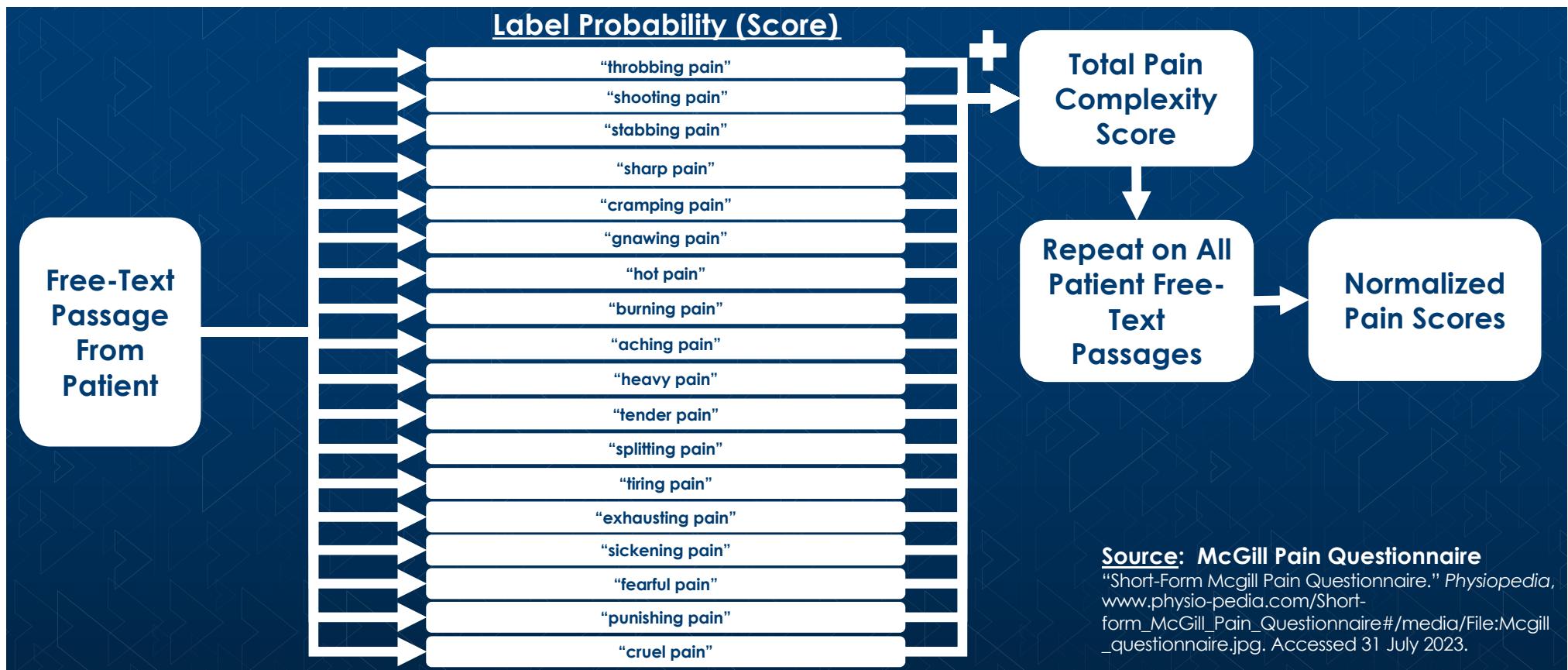
Boston  
Scientific





# Multi-label Text Classification: Clinical Labels for Pain Complexity

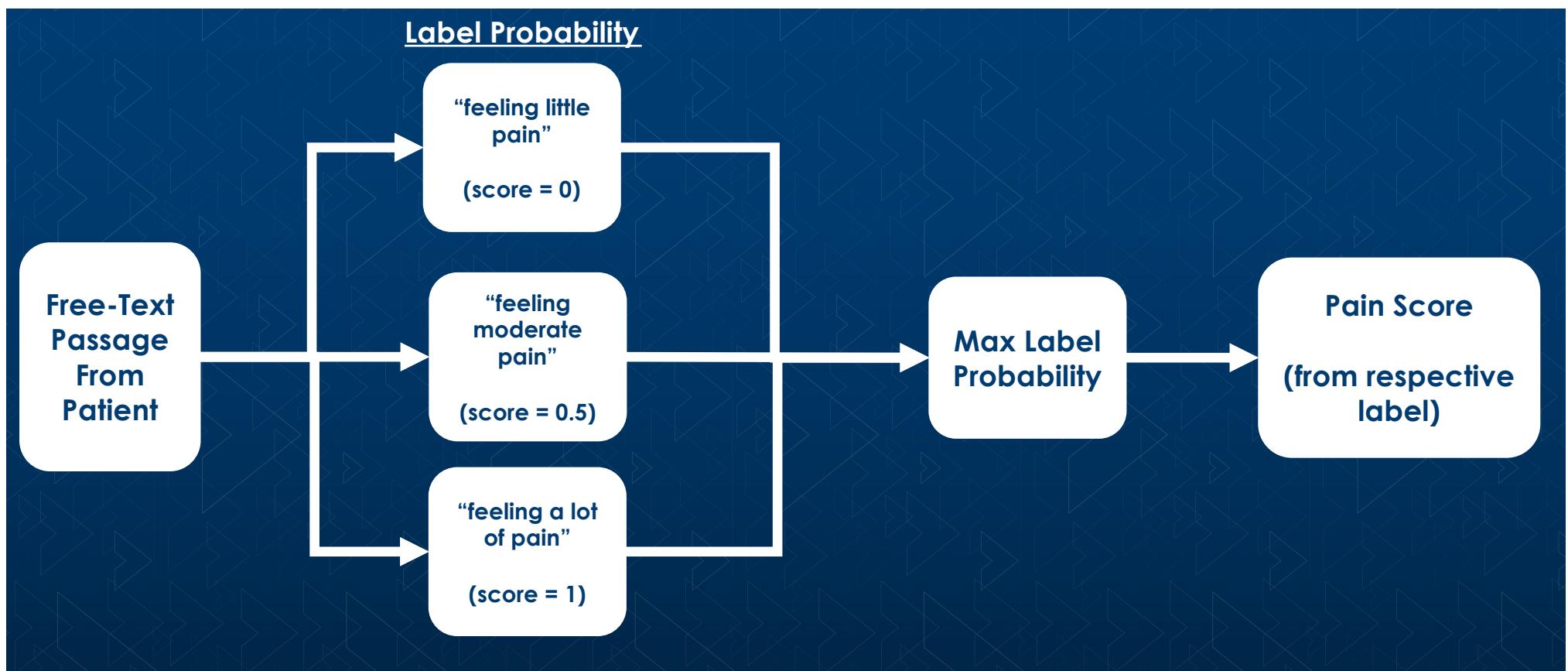
Boston  
Scientific





# Single Label Text Classification: Labels for Pain

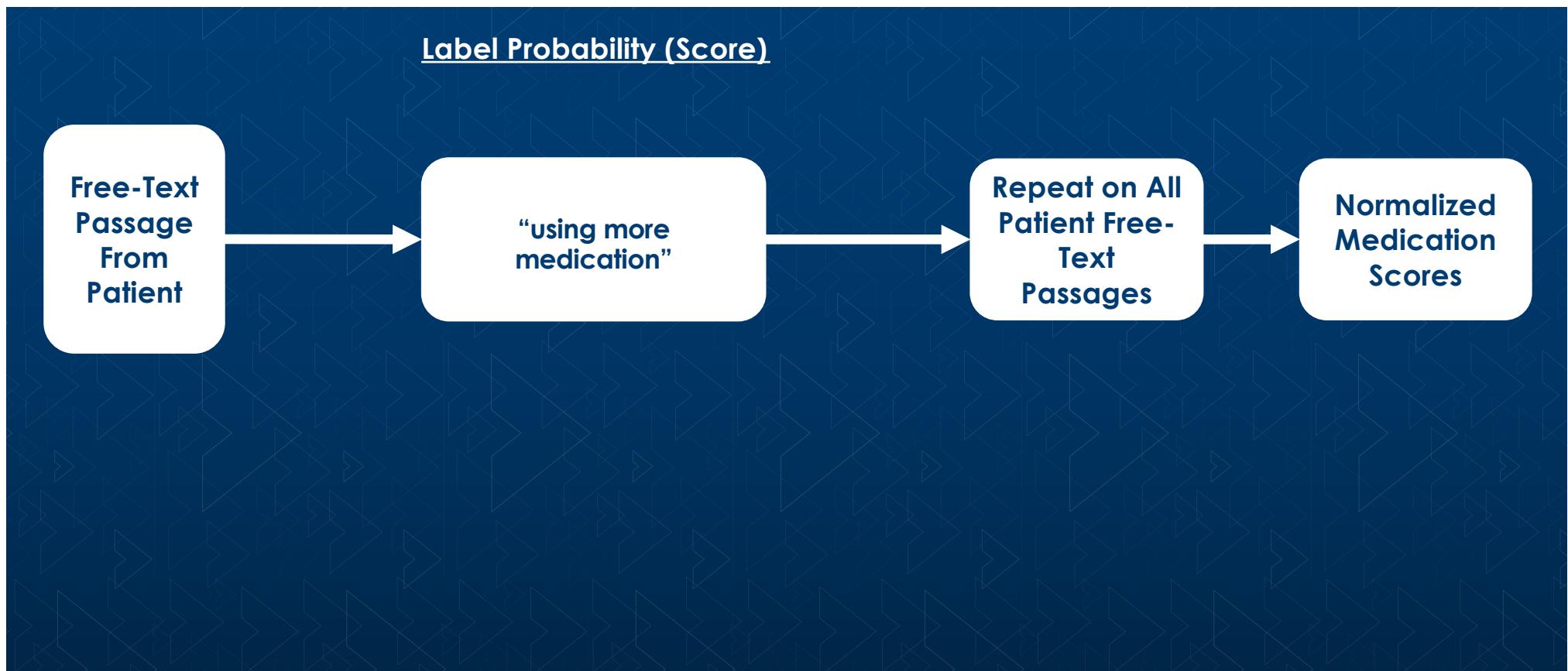
Boston  
Scientific





# Single Label Text Classification: Labels for Medication

Boston  
Scientific



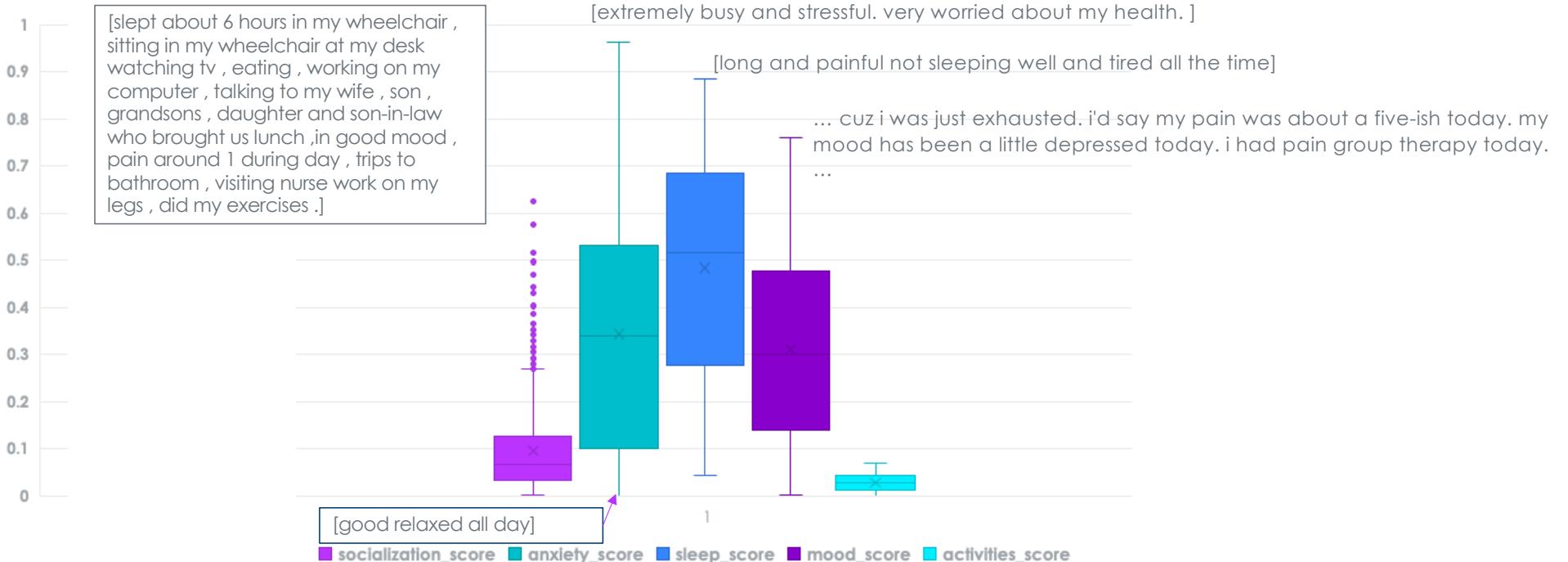
# Results



- Resulting scored data for all patient free-text responses ended up being a fairly sparse data set (a lot of concepts found to be irrelevant in each passage)
- Missing data was imputed
  - Iterative Imputer for imputation of concepts with missing scores (due to low relevancy scores)

# Distributions

Boston  
Scientific

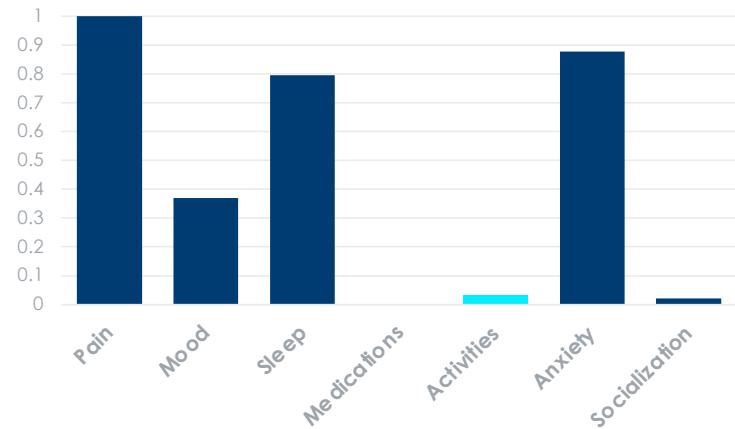




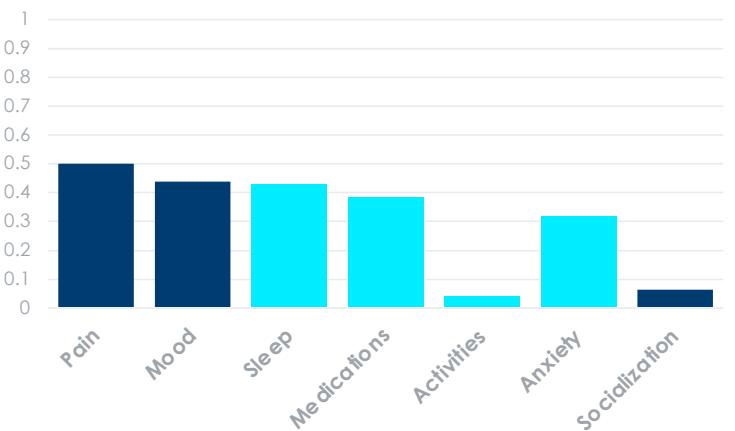
## Sample Results

Boston  
Scientific

i'm not sleeping well my pain has increased i'm still in the same medication my mood goes with my pain if i hurt paranoid

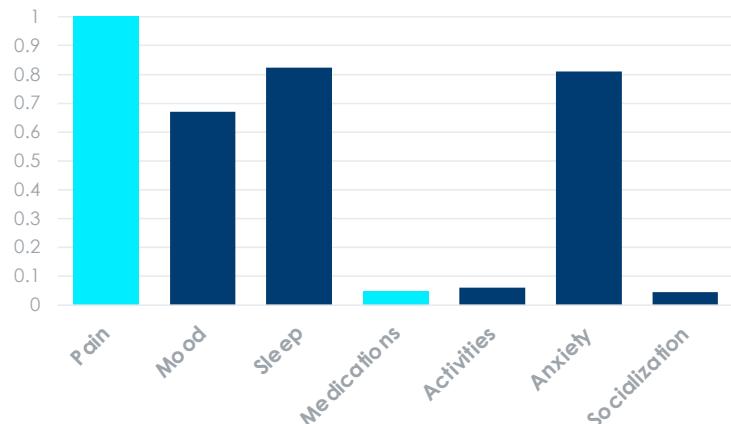


my day was fine my pain is moderate but i'm in a crap mood still taking medication as needed

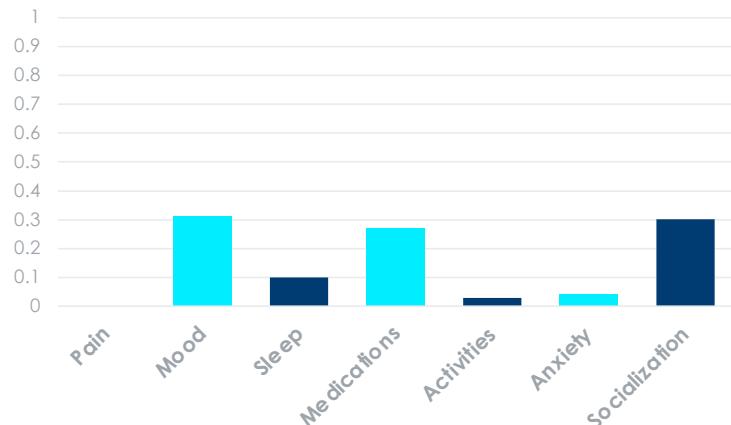


 | Sample ResultsBoston  
Scientific

i've been very tired today. i'm just not sleeping well. i could not even text a friend. i fell a sleep 3 times before sent.



slept well went to breakfast with friends we are visiting and went to see some of the sights

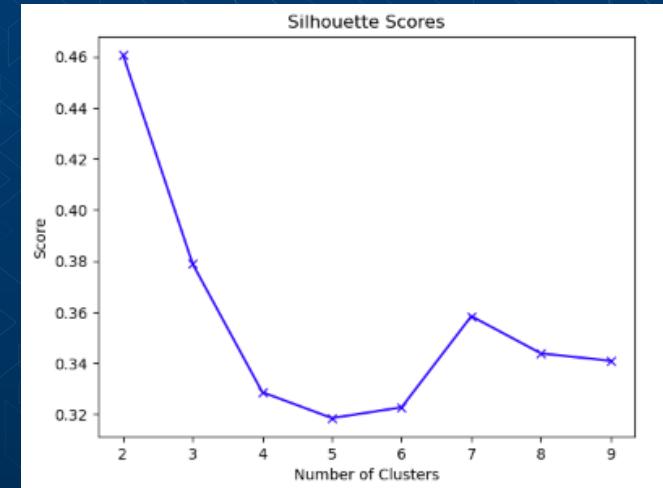
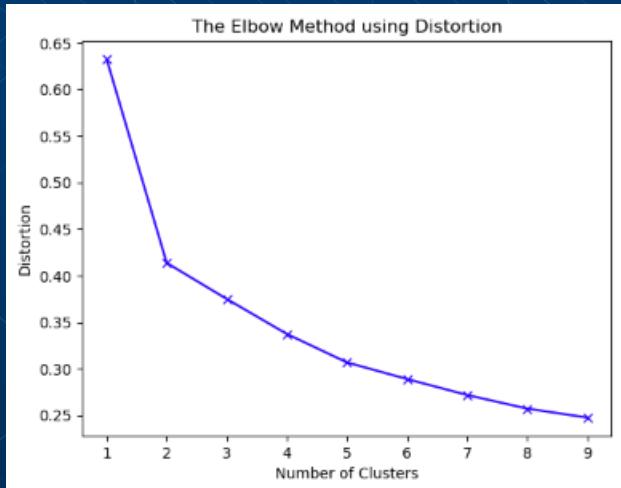
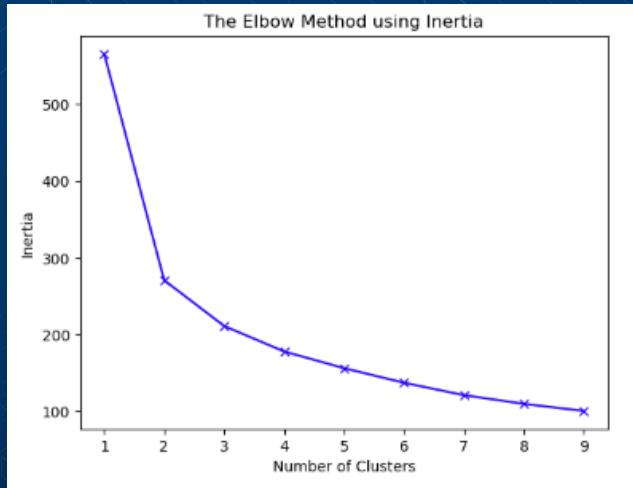


## Clustering Model for Patient States



# Clustering via KMeans

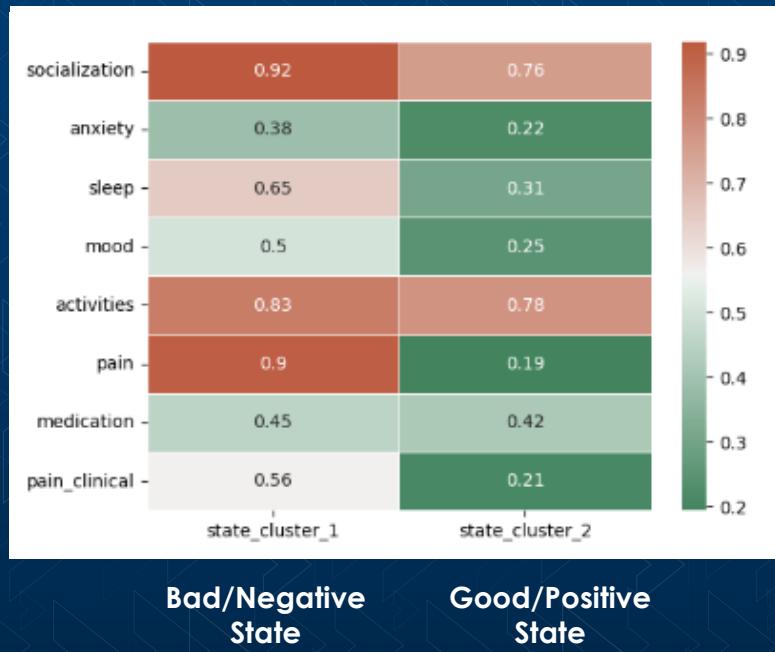
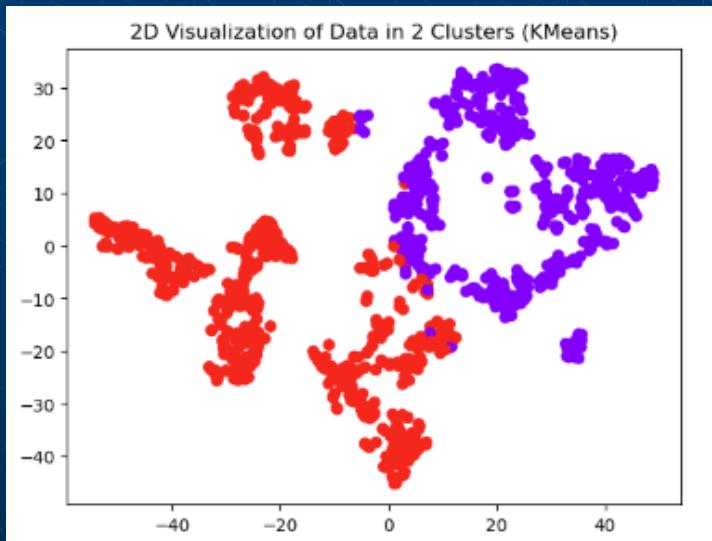
Boston  
Scientific





# Clustering via KMeans: 2 Clusters

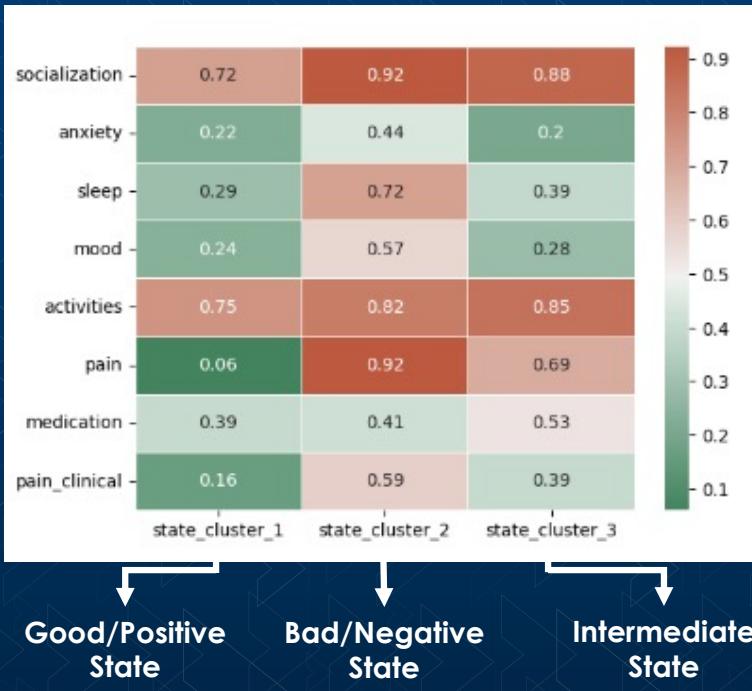
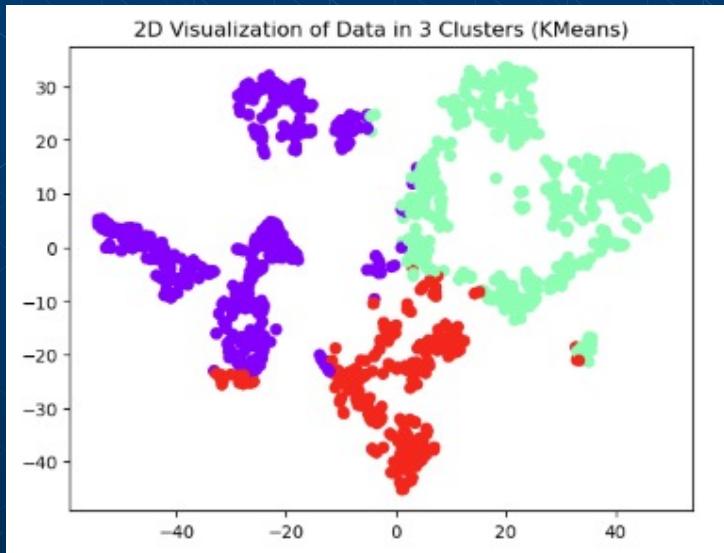
Boston  
Scientific





# Clustering via KMeans: 3 Clusters

Boston  
Scientific





# KMeans Clustering: Results & Improvements

Boston  
Scientific

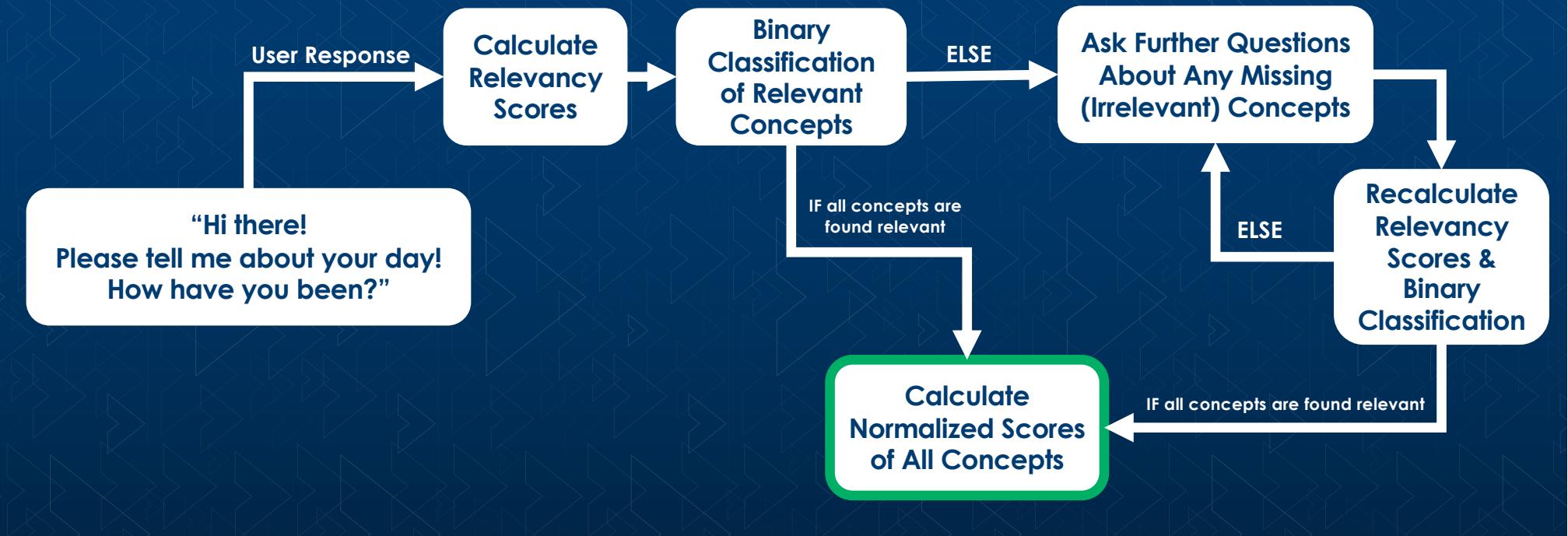
- About 100 samples from each of the 2 KMeans clusters were analyzed to estimate the accuracy of the cluster results:
  - Each of the text samples were manually assigned a state using human judgement
  - Bad/Negative Cluster: 86% correctly clustered samples
  - Good/Positive Cluster: 71% correctly clustered samples
- Next Steps for Improvements:
  - Create ChatBot to further extract information from the patient for missing features/concepts via Q&A
    - Avoids need for imputed scores, which often times may be inaccurate and skewed

# ChatBot



Boston  
Scientific

- All models, functions, labels, queries, and values (e.g. max/min for normalization) were structured into a class
- Using no external LLM models





- For every user response to a question asked about a concept/feature, preface text is appended to the beginning of the response to ensure the MNLI Transformer correctly identifies the concept in the user's response

“it’s gotten worse”

vs.

“In terms of my pain it’s gotten worse”

No clear context from the text

Clearly identified as referring to pain levels



# ChatBot: Initial Example

Boston  
Scientific

- Initially, the canned ChatBot responses were based off relevancy scores, structured through a simple loop to continue asking questions until all concepts are found relevant
  - This would ask the same single text block of questions every time a concept was missing

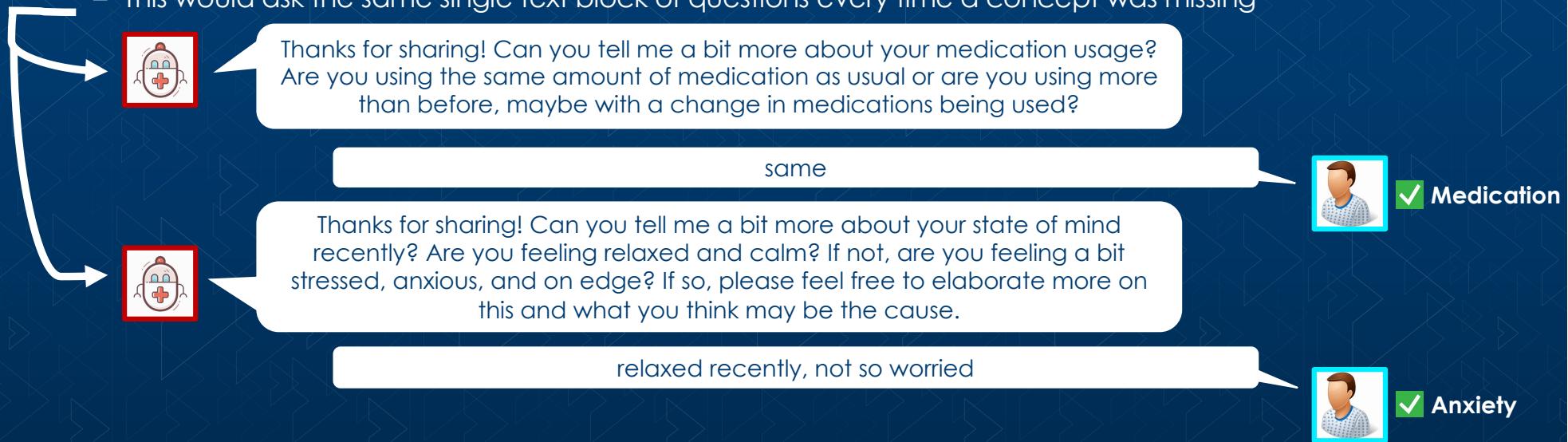




# ChatBot: Initial Example

Boston  
Scientific

- Initially, the canned ChatBot responses were based off relevancy scores, structured through a simple loop to continue asking questions until all concepts are found relevant
  - This would ask the same single text block of questions every time a concept was missing

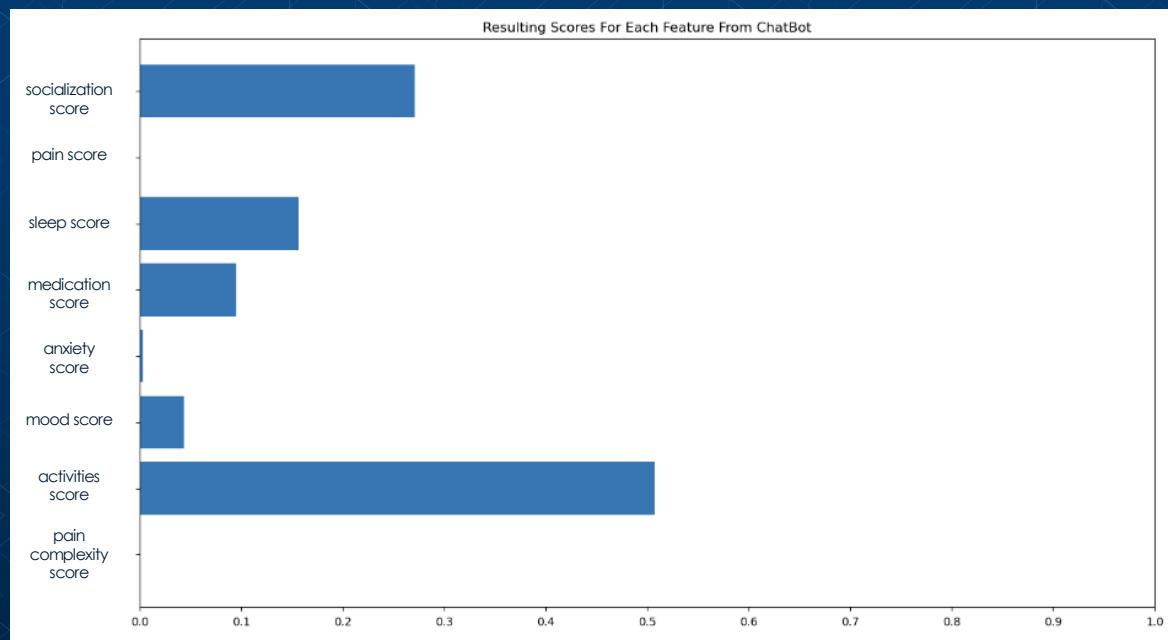




# ChatBot: Initial Example

Boston  
Scientific

- Initially, the ChatBot was created using no external LLM models, and structured through a simple loop to continue asking questions until all concepts are found relevant
  - This would ask the same single text block of questions every time a concept was missing



In this case, for example, a pretty vague (and irrelevant) block of text marked the activities feature as relevant, resulting in a questionable activities score...

This could be avoided & improved by asking deeper diving questions (based off the chat history & goals/interests) via pretrained external LLMs



- Contributions from Llama 2:
  - Keeps track of chat history between LLM and user in order to formulate more detailed follow up questions to obtain more useful information
  - Uses a deep pretrained model to create a variety of improved unique/creative questions
- Same structure of the ChatBot, but this time, Llama 2 provides the opportunity to ask deeper diving and more detailed questions based off the chat history

- **System Prompt** – defining the agent, who they are, what their interests are, and what their task is in this situation
  - “You are a nurse working with chronic pain patients of spinal cord stimulation. You are interested in obtaining enough information from the patient to properly understand their current patient state and their overall well-being. You are very nice, consoling, and welcoming to all patients as a professional nurse. You are interested in 7 key concepts that can be used to determine their overall well-being and patient state. These concepts are pain, sleep, mood, anxiety, medications, activities, and socialization. You will insist on asking as many questions as necessary to know enough information about these 7 concepts from the patient. If a patient does not talk enough about one of these 7 concepts, you will ask them to tell you a bit more about that missing information.”



Boston  
Scientific

- Prompts specific to each concept were appended to the system prompt when asking about a specific concept:
  - **Pain Prompt:** “Generate text asking the user to talk more about their pain recently. Ask if their pain levels are better or worse and if they are feeling more or less pain.”
  - **Sleep Prompt:** “Generate text asking the user to talk more about their sleep recently. Ask if they are having trouble falling or staying asleep and if they are sleeping enough hours. Ask if they are sleeping better or worse and if they are feeling more or less tired.”
  - **Medication Prompt:** “Generate text asking the user to talk more about their medication consumption lately. Ask if they have made any changes to the amount of medications they are consuming and if they have had to start taking new medications. If they did make changes, ask why.”
  - **Activities Prompt:** “Generate text asking the user to talk more about their activities lately. Specifically ask them if they have been sitting, standing, walking, running, exercising, eating, cooking, driving, bathing, dressing, and/or performing housework/chores.”
  - **Socialization Prompt:** “Generate text asking the user to talk more about their socialization lately. Ask if they have been spending time with family and friends lately, and if so, if they are enjoying it.”
  - **Mood Prompt:** “Generate text asking the user to talk more about their mood lately. Ask if they have been mostly in a good or bad mood. Also ask why they have been in that and what they think is causing it.”
  - **Anxiety Prompt:** “Generate text asking the user to talk more about if they have been feeling anxious or relaxed lately. Ask for them to detail why they are feeling that way recently.”



# Enhanced ChatBot with Llama 2: Example

Boston  
Scientific



Hi there! Please tell me about your day! How have you been?



feeling alright, been able to get a lot more work done lately without needing to stop for as many breaks

Activities

Mood

Anxiety



Great to hear that you've been able to increase your productivity without needing to take as many breaks! Can you tell me more about what you've been doing differently lately? Have you made any changes to your workspace or routine?



no changes in my routine besides the scs. the scs has made the job easier



That's great to hear! It's wonderful when you can find ways to make your work easier and more manageable. Can you tell me more about how the spinal cord stimulation (SCS) has helped you? Has it reduced your pain levels or made it more manageable?



Pain

the scs has lowered my pain a lot. the pain is less noticeable now and makes the job easier to perform without constant breaks

} Initial LLM Response About Pain  
(System Prompt + Pain Prompt)

} Follow-up Response From LLM  
About Pain To Ask Deeper  
Questions For Detailed  
Descriptions  
(based off chat history)



# Enhanced ChatBot with Llama 2: Example

Boston  
Scientific



As a nurse, I am interested in obtaining a comprehensive understanding of your overall well-being, including your sleep patterns. Can you tell me more about your sleep recently? Have you been having trouble falling or staying asleep? How many hours of sleep are you getting each night? Have you noticed any changes in your sleep quality or quantity over the past few days or weeks? Are you feeling more or less tired during the day? Any changes in your sleep patterns could provide valuable information in assessing your current patient state. Please feel free to share any details you can about your sleep, and I will do my best to help you.

} Initial LLM Response About Sleep  
(System Prompt + Sleep Prompt)

✓ Sleep

sleep has been more or less the same. the one big improvement has been that i no longer have pain waking me up in the night so i feel better in the morning



Great, it's good to hear that you've noticed an improvement in your sleep quality! Not waking up in the night due to pain is a significant improvement, and it's wonderful that you're feeling better in the morning as a result. Can you tell me more about your sleep patterns? How long does it take you to fall asleep, and how many hours of sleep do you typically get each night? Have you noticed any changes in your dreams or nightmares since starting the spinal cord stimulation therapy? Any additional details you can provide will help me better understand your sleep patterns and overall well-being.

} Follow-up Response From LLM  
About Sleep To Ask Deeper Questions For Detailed Descriptions  
(based off chat history)

i usually get around 7 hours of sleep and it doesn't take me long to fall asleep. not getting nightmares. mostly resting sufficiently to feel good the next morning





# Enhanced ChatBot with Llama 2: Example

Boston  
Scientific



As a nurse, I am interested in obtaining enough information from you to properly understand your current patient state and your overall well-being. I want to know more about your medication usage to determine your overall well-being and patient state. Can you tell me more about your medication consumption lately? Have you made any changes to the amount of medications you are consuming? Have you had to start taking new medications? If you did make changes, can you tell me why?

Initial LLM Response About Medication  
(System Prompt + Medication Prompt)

Medication

i am taking the same amount and type of meds. no changes



Thank you for letting me know. It's important for me to understand your medication routine and how it's been working for you. Can you tell me more about the medications you're taking? What are they for and how often do you take them? This information will help me to better understand your overall health and well-being.

Follow-up Response From LLM  
About Medication To Ask Deeper Questions For Detailed Descriptions  
(based off chat history)

i take acetaminophen and antidepressants just once a day. they help regulate pain





# Enhanced ChatBot with Llama 2: Example

Boston  
Scientific

## FINAL USER TEXT RESPONSE:

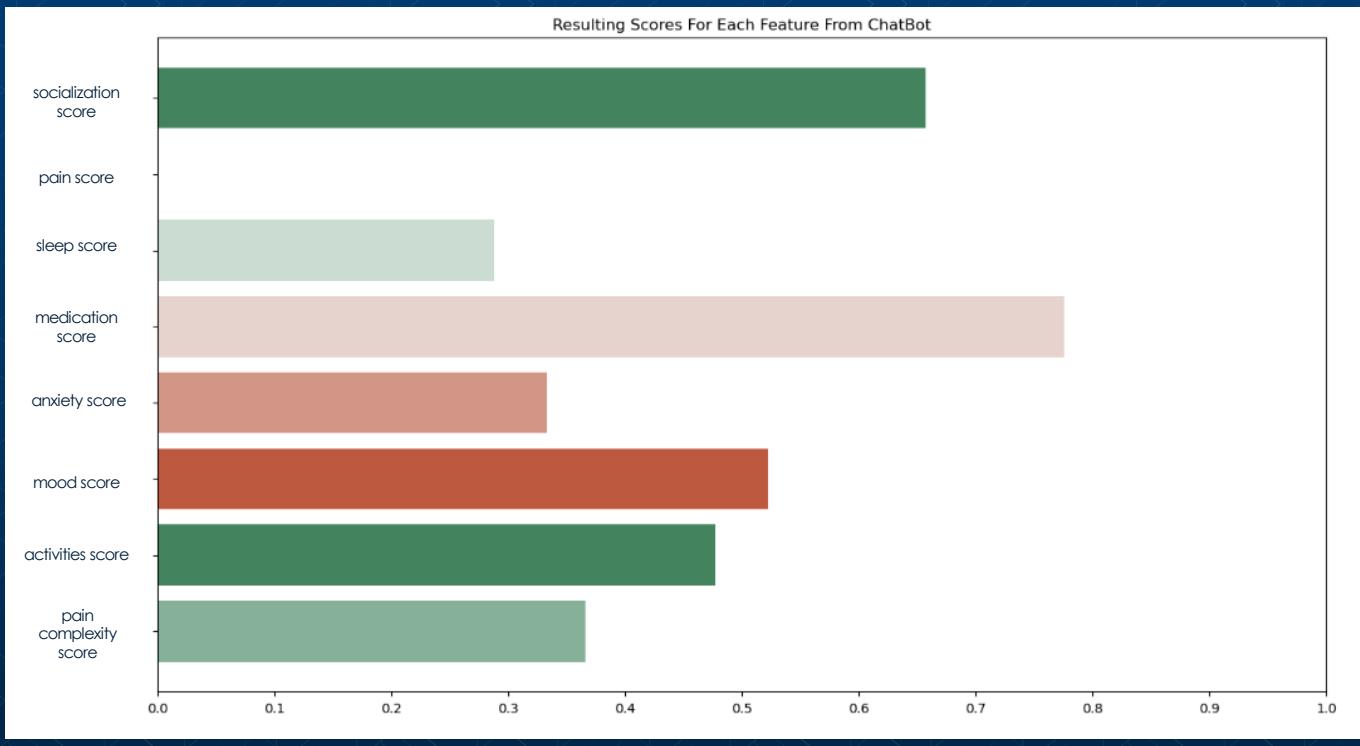
feeling alright, been able to get a lot more work done lately without needing to stop for as many breaks . In terms of my pain no changes in my routine besides the scs. the scs has made the job easier. In terms of my pain the scs has lowered my pain a lot. the pain is less noticeable now and makes the job easier to perform without constant breaks. In terms of my sleep sleep has been more or less the same. the one big improvement has been that i no longer have pain waking me up in the night so i feel better in the morning. In terms of my sleep i usually get around 7 hours of sleep and it doesn't take me long to fall asleep. not getting nightmares. mostly resting sufficiently to feel good the next morning. In terms of my medication i am taking the same amount and type of meds. no changes. In terms of my medication i take acetaminophen and antidepressants just once a day. they help regulate pain



# Enhanced ChatBot with Llama 2: Example

Boston  
Scientific

- ALL user responses are then combined into one text block & used to calculate scores for each of the features



**Predicted cluster from KMeans 2:**

Good/Positive State

**Predicted cluster from KMeans 3:**

Intermediate State

# Future Directions



# Future Directions

Boston  
Scientific

- Obtain scores for larger number of free-text samples to further improve the cluster centroid results
- Correlations with reported scores from patient questionnaires
- One of the biggest obstacles with room for improvement is the inconsistency of the relevancy scores generated by the Semantic Search model
  - Although improved through trying different queries, clearly relevant sentences to a concept are sometimes falling below the threshold while many completely irrelevant sentences are above the threshold
  - Especially now that a ChatBot is designed to ask further deeper questions for more detailed information, the relevancy thresholds can be made more strict (higher thresholds) to ensure that the ChatBot doesn't stop asking questions about a feature prematurely before sufficient information is obtained
- Further improvements to the preface text for user responses to the ChatBot would help obtain better context behind the responses
  - At times, slightly different & more specific preface text is needed to understand the context
- Further reduce the "max\_new\_tokens" parameter for responses from the Llama 2 LLM to avoid each LLM response being too long



# Sources for Clinical Labels

Boston  
Scientific

- Anxiety:
  - [GAD-7 Anxiety-updated\\_0.pdf \(adaa.org\)](#)
  - Löwe, Bernd, et al. "Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population." *Medical care* (2008): 266-274.
- Sleep:
  - Sleep Quality Scale (SQS)
  - Yi H, Shin K, Shin C. Development of the sleep quality scale. *J Sleep Res.* 2006 Sep;15(3):309-16. doi: 10.1111/j.1365-2869.2006.00544.x. Erratum in: *J Sleep Res.* 2006 Dec;15(4):465. PMID: 16911033.
- Mood:
  - [PHQ-9 Microsoft Word - 425C28DF-41A0-084E9F.doc \(montefiore.org\)](#)
  - Kroenke, Kurt, Robert L. Spitzer, and Janet BW Williams. "The PHQ-9: validity of a brief depression severity measure." *Journal of general internal medicine* 16.9 (2001): 606-613.
- Socialization
  - [Table 1 from Groningen Orthopaedic Social Support Scale: validity and reliability. | Semantic Scholar](#)
  - <https://www.semanticscholar.org/paper/Groningen-Orthopaedic-Social-Support-Scale%3A-and-Akker-Scheek-Stevens/b0b15010df73986d76c5c0cdf97e5be656e7f569>
  - Groningen Orthopaedic Social Support Scale
  - van den Akker-Scheek, Inge et al. "Groningen Orthopaedic Social Support Scale: validity and reliability." *Journal of advanced nursing* vol. 47.1 (2004): 57-63. doi:10.1111/j.1365-2648.2004.03065.x
- Activities
  - Developed by BSC

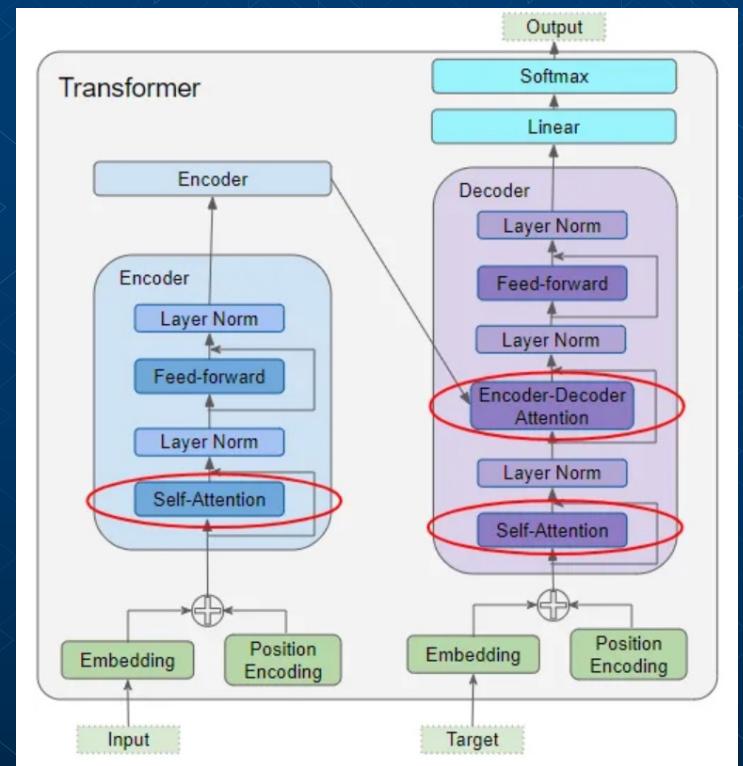
# Transformer Models



# Transformer Models

Boston  
Scientific

- State-of-the-art model for NLP
- Involve the following procedures on text:
  - Text embedding
  - Each word is numerically vectorized
  - Positional encoding
    - Each word is uniquely vectorized to represent its position in a sentence
  - Encoder & decoder with multi-head attention layers
    - Multiple attention heads give the Transformer greater power to encode multiple relationships and nuances for each word





# Transformer Models: Text Classification

Boston  
Scientific

- Most Transformer models are known for things such as translation and text generation
- For text classification in particular, the Transformer model includes one additional head before the output: the classification head
  - Generates prediction of the class label

