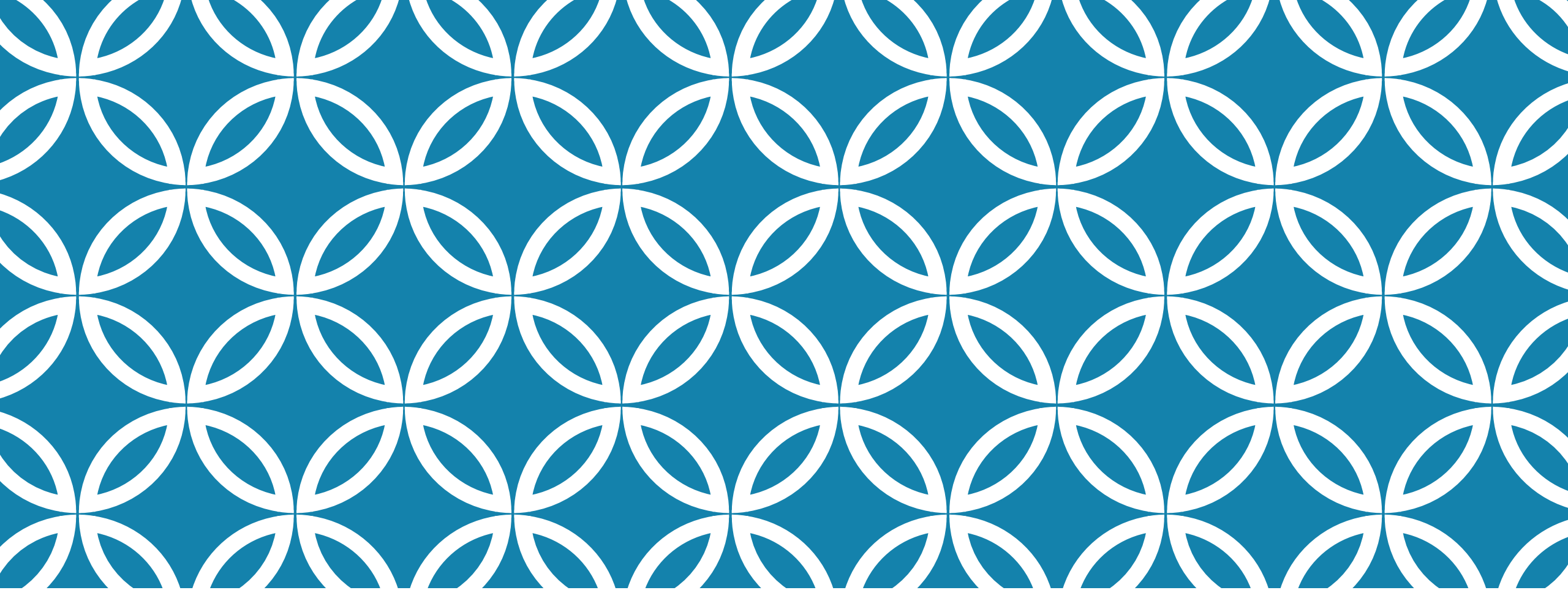


DATA MINING PERSPECTIVES EX: REDDIT

Graph Data Analytics Platform : Damie Brooks

OUTLINE

https://www.tigergraph.com/graph-for-all/

Judging Criteria

Challenge

Solution Proposal

- Reddit Data Extraction
- Graph Data Model
- Machine Learning Algorithms
- End Interaction Project

Project Schedule

Ownership

JUDGING CRITERIA

Most Impactful: Global impact across ages, topics, cultures

Most Innovative Graph: Leverages NLP and numerous graph algorithms

Most Ambitious and Complex Graph: could scale to ~5T nodes

Most Applicable Graph: text-based models for any UGC, ex: social media, blogs, reviews, papers…

CHALLENGE

Humanity is facing major challenges

- Democracies are being questioned throughout the world
- Technologies can represent either a threat or an opportunity

Break confirmation bias

- Foster critical thinking
- Engage with other points of view / those who think differently.

Empower users with tools to…

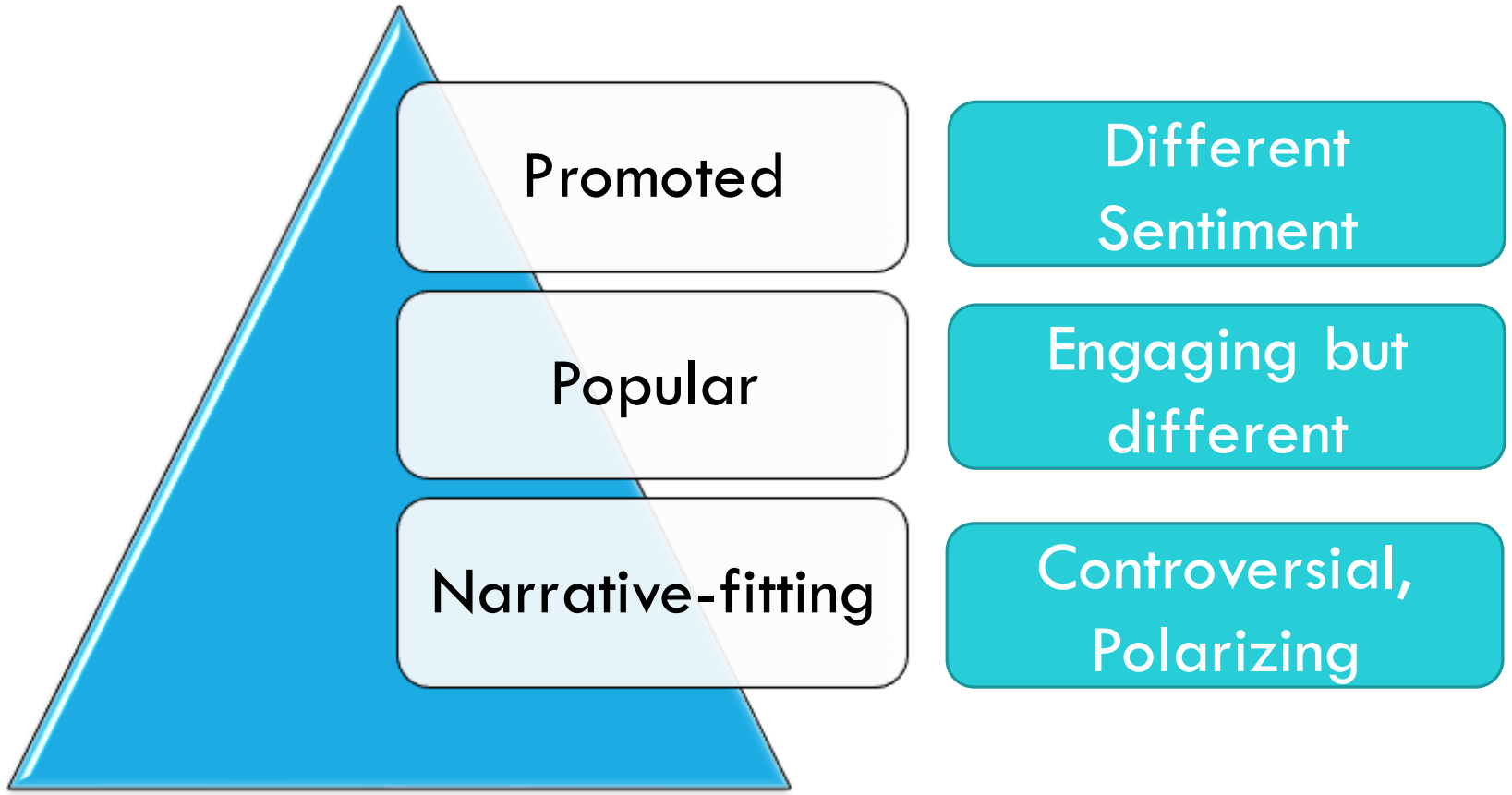
- **Identify bias**
- Seek **diverse perspectives** and opinions on the **same topic**

DEFINING CONFIRMATION BIAS

Confirmation Bias

- Evidence you already believe
  - Similar to your own viewpoints
    - Body text and headline text is similar to your own posts/comments
    - Post is from a user that you frequently comment on with positive sentiment
- Information itself is biased
  - Sentiment is negative, when there are positive sentiment alternatives

SOCIAL MEDIA ALGORITHMS



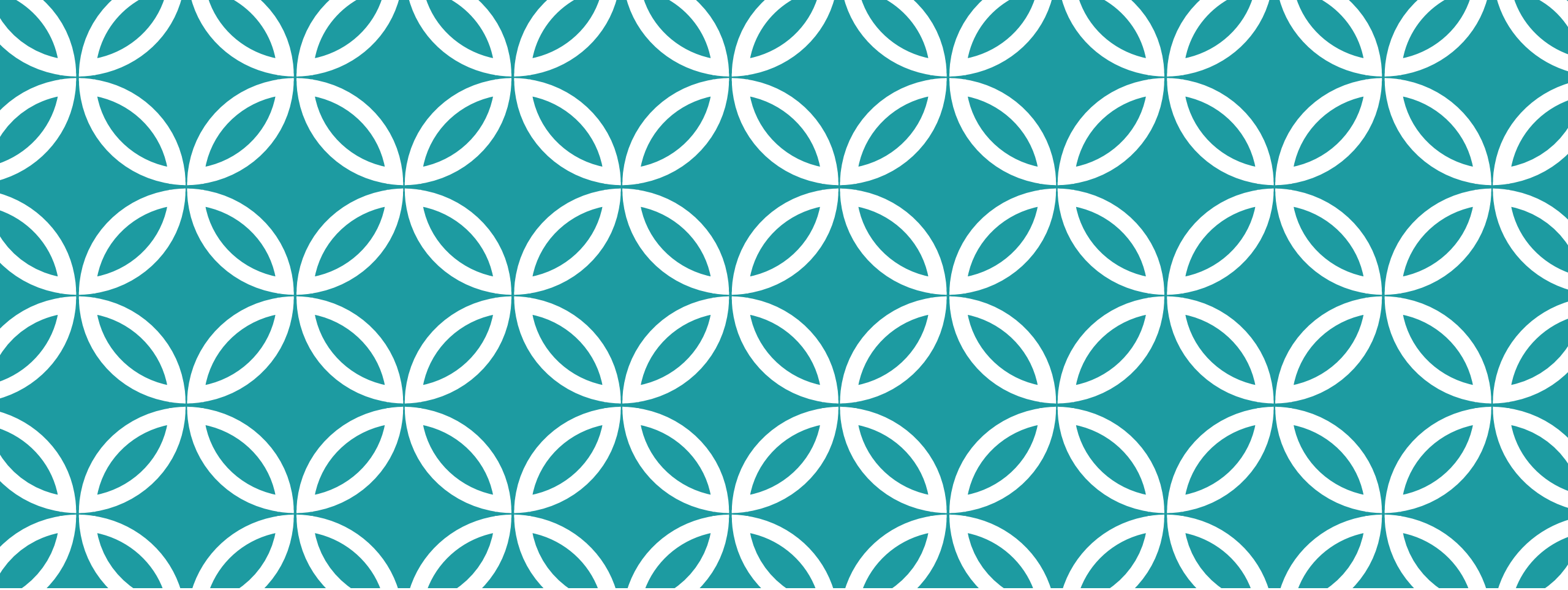
REDDIT UX

Given a post on Reddit, and a user

Compute similarities and offer suggestions

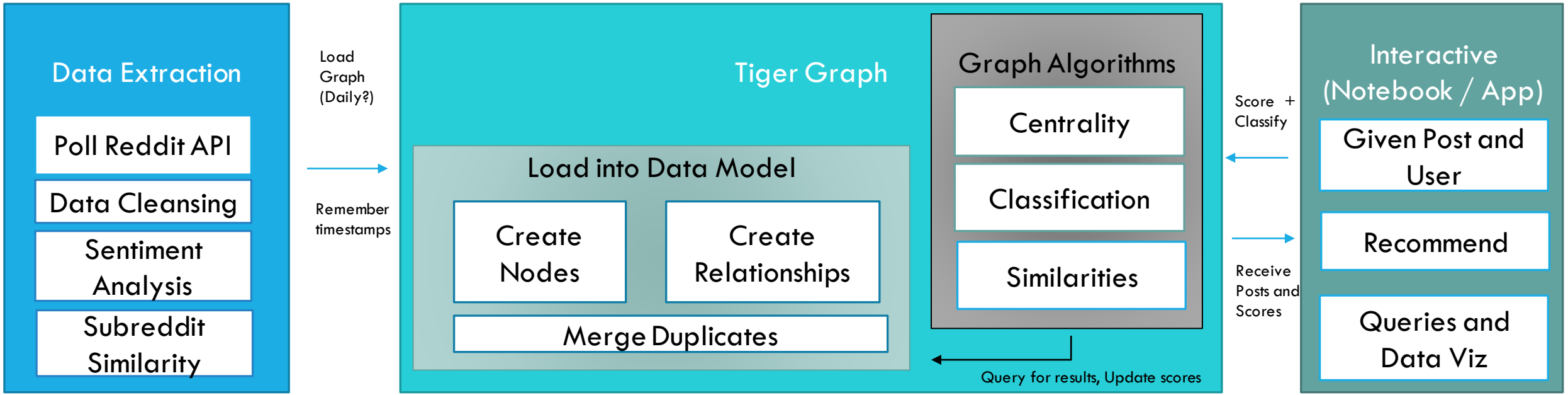
User sees varying perspectives on similar topics: Relevant but different

1. **Controversial:** Similar titles, similar classifications, from the controversial thread (polarizing)
2. **Unpopular Opinion:** Similar titles, similar classifications, less popular or hidden, from new thread
3. **Alternative feeling:** Similar titles, similar subreddit, differing sentiment
4. **Subject Matter Experts:** Most prolific writers in that topic, varying subreddits



SOLUTION DEFINITION

SYSTEM ARCHITECTURE



DATA EXTRACTION

Using HTTP with Python, pull Reddit posts

- Posts can be pulled every second at 100 at a time by subreddit name (or by “all”)
- for 6 week project time
- Pulls include:
  - Subreddit
  - Title
  - Body (selftext)
  - Author
  - Score
  - URL
  - Created Date
  - And more…

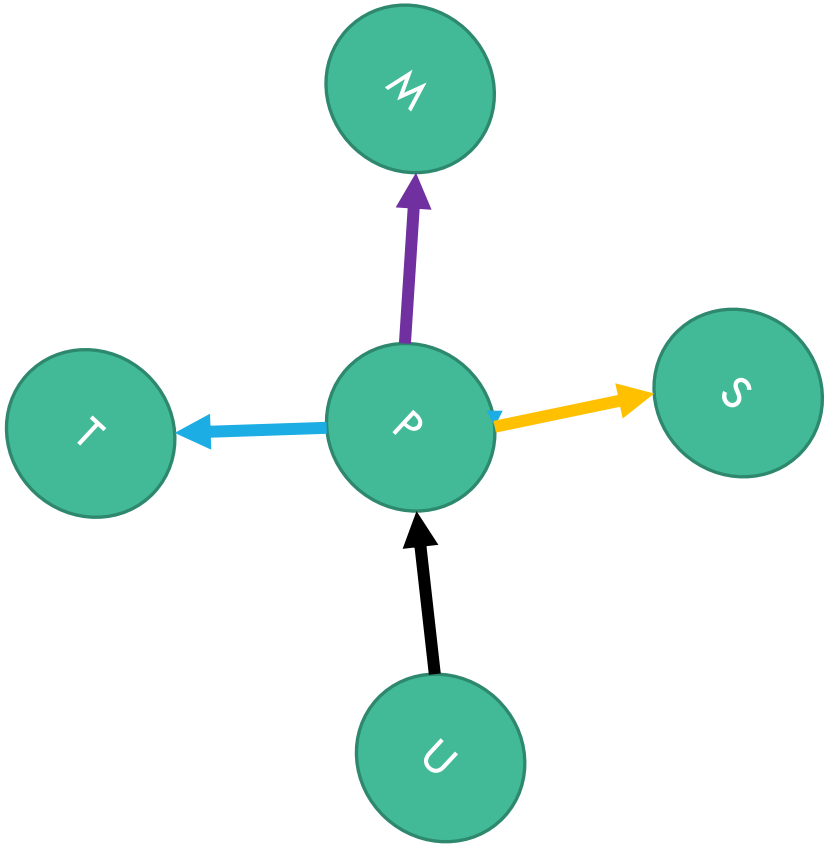
CLASSIFICATION

Stem and Lemminize, remove non alphabetic chars

Classify into categories within Subreddit

GRAPH DATA MODEL

Vertex	Edge
User U	User lists a post
Post P	Post is about subreddit
Subreddit S	
Sentiment M	Sentiment mood of post
PullType T	Post from a pulltype



Can these edges be undirected?

Edge Attributes		Weighted?
NumComments	(p)->(s)	YES
Ups	(p)->(s)	YES
Downs	(p)->(s)	YES
PostedDate	(u)->(p)	NO
SentimentOfTitle?	(u)->(p)	Maybe

ALGORITHMS

Sentiment Analysis on Title

- Positive, Negative, Neutral

Classification on Title

- Semantic meaning on Post
- “What is this post really about?”

Similarity with Factors:

- Title string words
- Sentiment
- Subreddit
- Author

Centrality

- Influencers within a subreddit

END INTERACTION PROJECT

Streamlit app

Graphistry

Jupyter notebooks

SCHEDULE – 6 WEEK PROJECT

Week Ending	Milestone
March 4	Proof of Concept (Data Extraction, Graph Algorithms)
March 11	Tiger Graph, Jupyter, Graphistry Ramp Up
March 18	Scaling the graph size for necessary compute speed
March 25	Test and Refine Bias Detection and Recommendation Engine
April 1	Queries, build Notebook, End Interaction Project
April 8	Project Video Submission