



CS-401

# Applied Data Analysis

## P1 Milestone

### Authors

Amélie Menoud

András Horkay

Cyrielle Manissadjian

Dominic Stratila

Katia Todorov

## Andras's Ideas

I am going to put my submission here so that when we compare and pick the project we will do, we can compare them easily

### When I Grow Up, I'll Be Anything but a Politician: Professions and Social Attitudes in the New Yorker Caption Contest

This project explores how different professions are portrayed in the New Yorker Caption Contest and how audiences respond to them. While some jobs, such as doctors, are often admired, others, like politicians and lawyers, are frequent targets of humour. Our research will focus on, but might not be limited to: Which professions appear most often in captions? Are certain occupations consistently rated funnier? Are some portrayed in a positive or negative light? To address these questions, we will first identify professions mentioned in captions using a U.S. Census occupation list, accounting also for indirect workplace references (e.g., “courtroom” for lawyers). The professions will then be grouped into broader categories (healthcare, law, education, politics, etc.). Finally, we will analyse their frequency, average funniness scores, and recurring stereotypes. We hope the findings will shed light on societal attitudes towards various professions as reflected in popular culture, and perhaps give ideas which professions to avoid in future careers to avoid being the butt of jokes.

### Laughing at the Forbidden: Taboo Themes in the New Yorker Caption Contest

This project investigates how taboo themes appear in the New Yorker Caption Contest and how audiences respond to them. While taboo jokes may be funny due to shock value, they can also be offensive or harmful. Some key questions we will answer are: Which taboo topics occur most often? How are they rated in terms of funniness? Do their frequencies change over time or with the political landscape in the US? To answer these, we will define taboo categories (e.g., suicide/self-harm, racism, sexism, violence, religion) and identify their presence in captions through keywords and phrases. We will then analyse the frequency at which these topics occur, their associated funniness scores, and any temporal trends, paying attention to events such as elections or economic crises that may influence prevalence. The project will shed light on how societal attitudes towards taboo topics are reflected in humour and how audience perceptions of these themes evolve over time.

### Funny or Stereotypical? Gender Roles in the New Yorker Caption Contest

This project examines how gender is represented in the New Yorker Caption Contest, both in cartoons and in audience-submitted captions. The key questions we will attempt to answer are: Do men appear more frequently than women in the images? When women are depicted, are they more often shown in stereotypical roles (domestic or caregiving) rather than professional settings (as leaders)? In captions, are gendered terms or phrases used that reinforce stereotypes? We will analyse cartoon metadata to measure representation, classify roles to detect stereotypes, and examine captions for gendered language. These steps may involve image as well as text analysis. Finally, we will compare funniness scores to assess whether captions using gendered language or stereotypes are rated differently by audiences. The results will provide insight into the persistence of gender stereotypes in humour and their reception by the public.

## Cyrielle's Ideas

### Why cognitive biases make jokes funnier ?

Cognitive biases are systematic deviations in thinking distorting our perception, judgment, decision-making. Around 250 cognitive biases are referenced, they are classified into 6 categories, we study 3 of them.

- Attentional biases: Do raters become less amused after seeing too many captions (“bored-rater effect”) ? We can track the vote distributions over ranking order to detect potential decrease in average scores. Do certain visual elements (such as anthropogenic animals, uncanny objects/situations) attract more humorous interpretations ? Using the image metadata (objects, locations, entities) to test whether attention-grabbing visuals correlate with higher fun.

- Judgment biases: Is humor works with left-brain/right-brain ? In other word, do logical/emotional joke produces more/less fun ? Use a text-based metric to measure polarity and subjectivity of the caption might help to answer. Calculate statistics per identified groups.

- Cultural biases: ”Do you get it ?” is an often-heard sentence when talking about humorous content. Do culturally referential captions (brands, celebrities, memes) outperform classical ones ? Are winning captions influenced by events ? We can detect proper nouns, cultural entities in captions with text processing, then compare funniness scores via T-tests. Overlay with publication dates (an other dataset need to be merged) to match historical-political context.

### Are we all sheep and do we reproduce the same patterns?

- Diversity across cartoons : Do we keep writing the same jokes forever ? Do trends come back (as in the world of fashion and/or music) ? Do certain humor types reappear over time regardless of image content ? We can extract dominant joke types across years (e.g., existential crisis, sarcasm, animals acting human) using thematic clustering, then correlate with cartoon metadata.

- Diversity within cartoons : Are thousands of captions mainly identical ? Do most people have the same interpretation of the cartoon ? Calculating intra-cartoon diversity using semantic distance metrics, and label clusters with classic humor theory types to detect which “family of humor” dominates and whether originality helps to win the contest.

## Do all caption submissions really have a possibility to win ?

- About fun : Is there a formula for “winning” caption ? Does good writing comes from good ideas or rather good sentences ? Are funny captions sharp, absurd, readable, abstract … ? For each cartoon, we can group captions by humor strategy (statement vs dialogue, clever vs literal), compare pairwise caption clusters telling the “same idea” but with different phrasing. We can also compare cross-cartoon clustering, by plotting a relation between the captions similar to odd elements (uncanny description) or similar to the expected context (description).

- About the System : Is the game fair or algorithmically set up ? Perhaps some captions are highlighted by the competition (reputation bias: Proposals that receive feedback early on accumulate more). We can plot funniness scores vs number of votes to detect power-law effects. We also can play with comparison of low-vote/high-score vs high-vote/moderate-score captions.

## Dominic's Ideas

### Causality or Correlation? Unlikely Trends in the New Yorker Caption Contest

This project explores strange or unexpected correlations between different trends in the Caption Contest that appear unrelated at first sight. For example, Is there a connection between the seasons and jokes about cats? How do jokes about politicians affect the number of people depicted in the cartoon? Does the presence of certain objects (like chairs, doors, or telephones) correlate with particular humor strategies? We will systematically identify such patterns by cross-analysing captions, metadata, and funniness scores across distinct spheres (visual elements, themes, cultural references). The focus is not necessarily to establish causal relationships, but rather to highlight these surprising overlaps and ask whether they reflect deeper societal attitudes, hidden cognitive associations, or simply coincidences. The outcome may provide both serious insights into how people connect ideas in humor and amusing examples of “funny but meaningless” correlations.