Idea 1: **Toxic Comment Classification Challenge**

Goal: help online discussion become more productive and respectful.

(1) What is the business problem?

Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

The Conversation AI team, a research initiative founded by Jigsaw and Google (both a part of Alphabet) is working on tools to help improve online conversation. One area of focus is the study of negative online behaviours, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). So far they've built a range of publicly available models served through the Perspective API, including toxicity. But the current models still make errors, and they don't allow users to select which types of toxicity they're interested in finding (e.g. some platforms may be fine with profanity, but not with other types of toxic content).

(2) Who are the intended stakeholders, and why is this problem relevant to them?

- Online and social media platforms
- Filter toxic comments

(3) Where are the datasets available from?

Kaggle
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

(4) What data science approaches do you anticipate you will use to model the business problem as a data science problem? - Supervised--Regression? Supervised--Classification? Unsupervised? Hybrid?

- It should be supervised-Classification

(5) How do you anticipate that the intended clients will use the results of your CP to address the original business problem?

- The results of this study will hopefully help online discussions become more productive and respectful.

Idea 2: **Score Clinical Patient Notes**

Goal: identify specific clinical concepts in patients' notes.

(1) What is the business problem?

When you visit a doctor, how they interpret your symptoms can determine whether your diagnosis is accurate. By the time they're licensed, physicians have had a lot of practice writing patient notes that document the history of the patient's complaint, physical exam findings, possible diagnoses, and follow-up care. Learning and assessing the skill of writing patient notes requires feedback from other doctors, a time-intensive process that could be improved with the addition of machine learning.

Until recently, the Step 2 Clinical Skills examination was one component of the United States Medical Licensing Examination® (USMLE®). The exam required test-takers to interact with Standardized Patients (people trained to portray specific clinical cases) and write a patient note. Trained physician raters later scored patient notes with rubrics that outlined each case's important concepts (referred to as features). The more such features found in a patient note, the higher the score (among other factors that contribute to the final score for the exam).

The goal of this study is to identify specific clinical concepts in patients' notes. Specifically, an automated method is developed to map clinical concepts from an exam rubric (e.g., "diminished appetite") to various ways in which these concepts are expressed in clinical patient notes written by medical students (e.g., "eating less," "clothes fit looser"). Great solutions will be both accurate and reliable.

(2) Who are the intended stakeholders, and why is this problem relevant to them?

- National Board of Medical Examiners
- Having physicians score patient note exams requires significant time, along with human and financial resources. Approaches using natural language processing have been created to address this problem, but patient notes can still be challenging to score computationally because features may be expressed in many ways.

(3) Where are the datasets available from?

Kaggle
https://www.kaggle.com/c/nbme-score-clinical-patient-notes/data

(4) What data science approaches do you anticipate you will use to model the business problem as a data science problem? - Supervised--Regression? Supervised--Classification? Unsupervised? Hybrid?

- It is a supervised problem.

# Idea 3: **Google QUEST Q&A Labeling**

Goal: build predictive algorithms for different subjective aspects of question-answering.

(1) What is the business problem?

Computers are good at answering questions with single, verifiable answers. But, humans are often still better at answering questions about opinions, recommendations, or personal experiences.

Humans are better at addressing subjective questions that require a deeper, multidimensional understanding of context - something computers aren't trained to do well. Questions can take many forms - some have multi-sentence elaborations, others may be simple curiosity or a fully developed problem. They can have multiple intents, or seek advice and opinions. Some may be helpful and others interesting. Some are simply right or wrong. It's hard to build better subjective question-answering algorithms because of a lack of data and predictive models. That's why the CrowdSource team at Google Research, a group dedicated to advancing NLP and other types of ML science via crowdsourcing, has collected data on a number of these quality scoring aspects to build predictive algorithms for different subjective aspects of question-answering.

(2) Who are the intended stakeholders, and why is this problem relevant to them?

- Intelligent Q&A systems
- Results from this study will inform the way future intelligent Q&A systems will get built, hopefully contributing to them becoming more human-like.

(3) Where are the datasets available from?

Kaggle
https://www.kaggle.com/c/google-quest-challenge/data

-

(5) How do you anticipate that the intended clients will use the results of your CP to address the original business problem?

- The results of this study can help the stakeholders such as Google improve the Q&A systems.