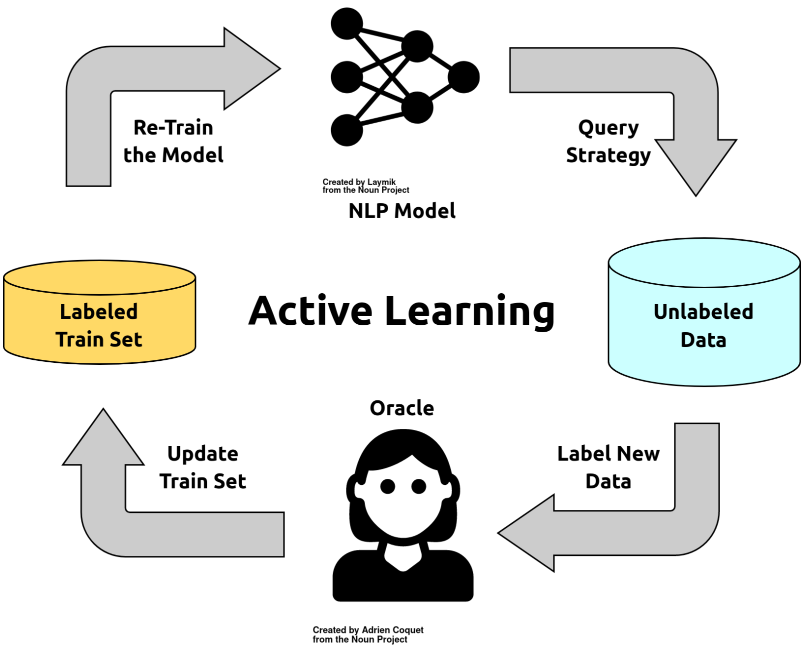
Active Learning NLP Value Classification

Values are abstract motivations that justify opinions and actions (Schwartz 2012). To be societally beneficial, AI should not be morally neutral, but actively strive to align with humans' social goals and interests (Russell et al. 2015). Understanding values is an essential milestone in achieving beneficial AI with applications in fields such as autonomous driving, healthcare, and AI-assisted policy making.

Existing methods for value elicitation typically rely on user surveys (Schwartz, 2012). However, in practical applications (e.g., to conduct meaningful conversations or to identify online trends), artificial agents should be able to identify values on the fly. The growing capabilities of **natural language processing** (NLP) enable the estimation of values from discourse (Mooijman et al. 2018; Hoover et al. 2020). **Value classifiers** can be used to identify the values underlying a piece of text.



NLP value classifiers are typically trained with the supervised paradigm, learning from manually annotated examples. However, due to the subjective and abstract nature of values, annotations are expensive to acquire. Furthermore, in time-pressured situations (e.g., when estimating online trends), manually labeling thousands of examples is unfeasible. **Active learning** (AL) is often used to address the scarcity of labels (Schroeder, 2020). In an AL paradigm, an intelligent strategy is used to iteratively select the most informative data to be annotated next by the (typically human) oracle, from a pool of unlabeled data points. This approach reduces the number of labels needed to achieve the desired performances.

The goal of this project is to implement an AL strategy to train a value classifier. A fully annotated dataset composed of 35k tweets is available (Hoover, 2020), and can be used to test the strategy. The student is invited to compare different choices of *model*, *query strategy* (which data points should be labeled next?), and *stopping criterion* (when should the training end?).

Desired Skills:

- Basic Python experience
- Basic NLP/ML knowledge

For more information, please send an email to Enrico Liscio (e.liscio@tudelft.nl) and Pradeep Murukannaiah (p.k.murukannaiah@tudelft.nl).

References:

Schwartz, S. H. 2012. “An Overview of the Schwartz Theory of Basic Values.” *Online readings in Psychology and Culture* 2(1):1–20.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. “Research Priorities for robust and beneficial artificial intelligence.” *AI Magazine* 36(4):105–114.

Mooijman, Marlon, et al. "Moralization in social networks and the emergence of violence during protests." *Nature human behaviour* 2.6 (2018): 389-396.

Hoover, Joe, et al. "Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment." *Social Psychological and Personality Science* 11.8 (2020): 1057-1071.

Danilevsky, Marina, et al. "A survey of the state of explainable AI for natural language processing." *arXiv preprint arXiv:2010.00711* (2020).