

P8: Taking a step back: what is active learning, after all?

Aikaterini-Lida Kalouli, Rebecca Kehlbeck, Rita Sevastjanova,
Miriam Butt, Oliver Deussen, Daniel Keim



P8's Goal

- automatically classify questions into ISQ - NISQ \Rightarrow utilize strengths of active learning
- Support and provide linguistic insights to other projects
- use linguistic insights of other projects

Today's Goal: take a step back

- shed light to machine learning
- introduce active learning as a subfield of machine learning
- present 3 use cases of our RU (and thus present our ongoing work)

Publications (since last retreat)

- **Kalouli, A., R. Crouch and V. de Paiva.** 2020. Hy-NLI: A Hybrid System for State-Of-The-Art NLI. In Proceedings of the 28th International Conference of Computational Linguistics (COLING). Barcelona, Spain.
- **Kalouli, A., R. Sevastjanova, V. de Paiva, R. Crouch and M. El-Assady.** 2020. XplaiNLI: Explainable Natural Language Inference through Visual Analytics. In Proceedings of the 28th International Conference of Computational Linguistics , System Demonstrations (COLING). Barcelona, Spain.
- **Sevastjanova, R., Jentner, W., Sperrle, F., Kehlbeck, R., Bernard, J., El-Assady, M.** QuestionComb: A Gamification Approach for the Visual Explanation of Linguistic Phenomena through Interactive Labeling. TiiS ACM Transactions on Interactive Intelligent Systems, Special Issue, 2020

Machine Learning

'Any sufficiently advanced technology is indistinguishable from magic.'

Arthur C. Clarke

Rather than magic, think of it as tools and technology, that you can use to answer questions with and about your data!

DeepL Übersetzer Lingue

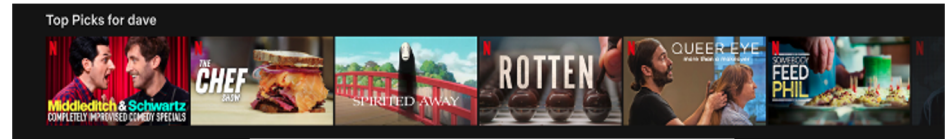
Übersetze Deutsch (erkannt) Übersetze nach Englisch Anpassen

Es kommen immer mehr Sprachen dazu. More and more languages are being added.

BEFORE AFTER

[1]

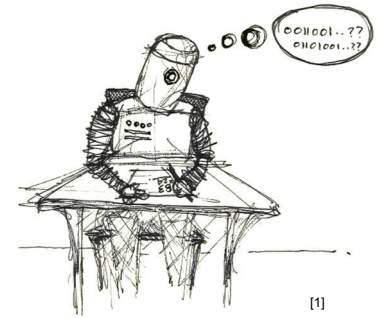
[2]



[1] <https://www.myfluentpodcast.com/e67-deepl-to-learn-english-languages/>
[2] <https://blog.google/products/search/search-language-understanding-ber/>

[3] <https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48>
[4] <http://rejoiner.com/resources/amazon-recommendations-secret-selling-online/>

Machine Learning



→ **Using data to get insights and answer questions**

Using data to

get insights

Clustering
Pattern Mining

Unsupervised,
needs 'just' data

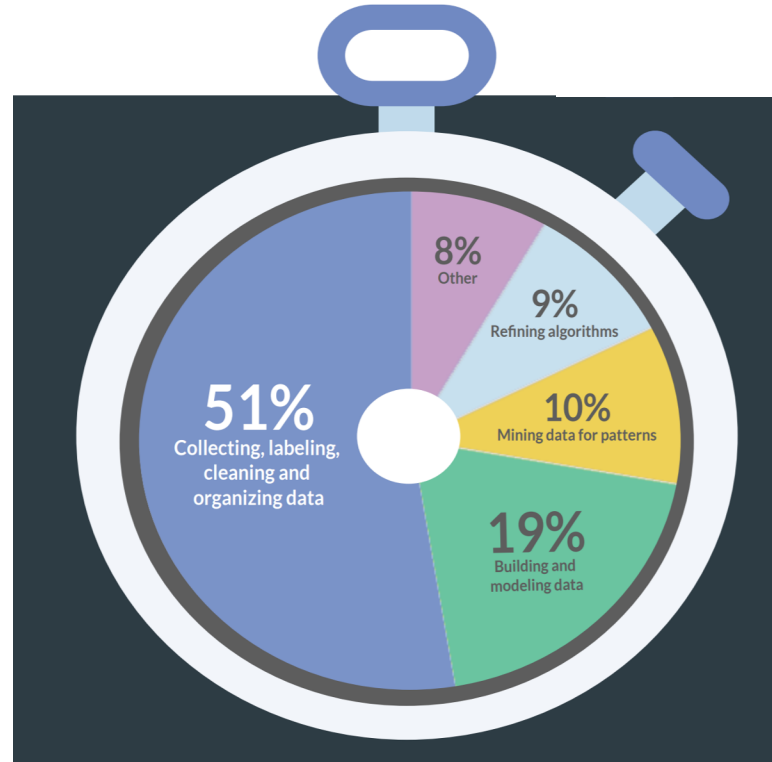
answer questions

Classification
Regression

Supervised,
needs high-quality labeled data

→ This system can create a positive feedback loop, i.e. the output of the system can be used to train the system further.

Machine Learning Process



[1]

Unsupervised Learning

Question: What pattern is in my data?

Get data

Train the model
(= the model learns)

Get insights

Supervised Learning

Question: Does this image show a cat or a dog?

Get data

Label the data

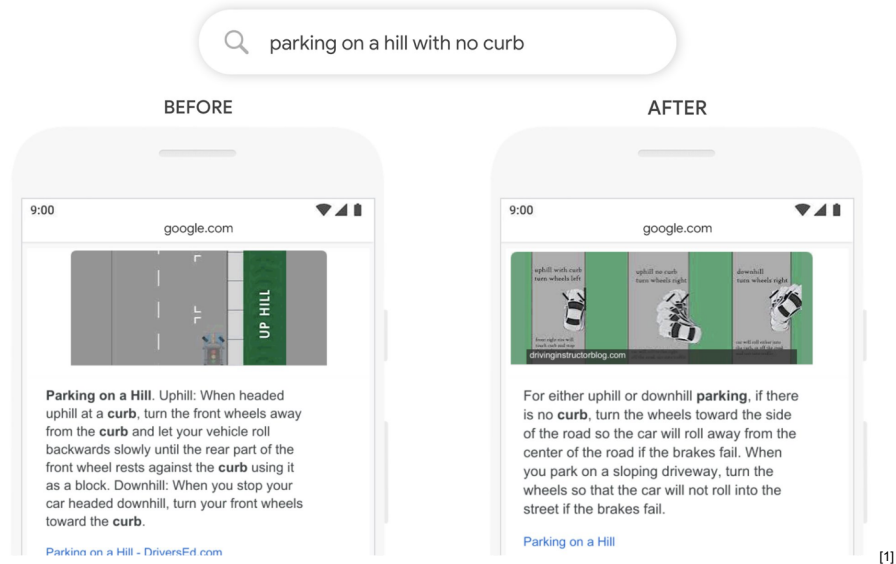
Train the model
(= the model learns)

Answer question

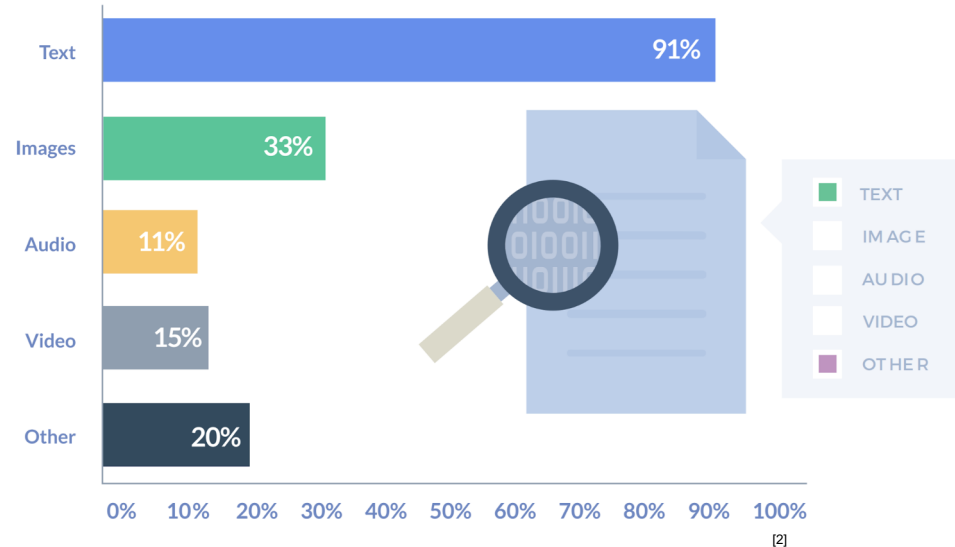
Textual Data

Language has inherent structure while at the same time being very ambiguous.

→ Text data is more difficult to process than other data, even though it is more prevalent.



[1]



[2]

[1] <https://blog.google/products/search/search-language-understanding-bert/>
[2] https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport.pdf

Active Learning

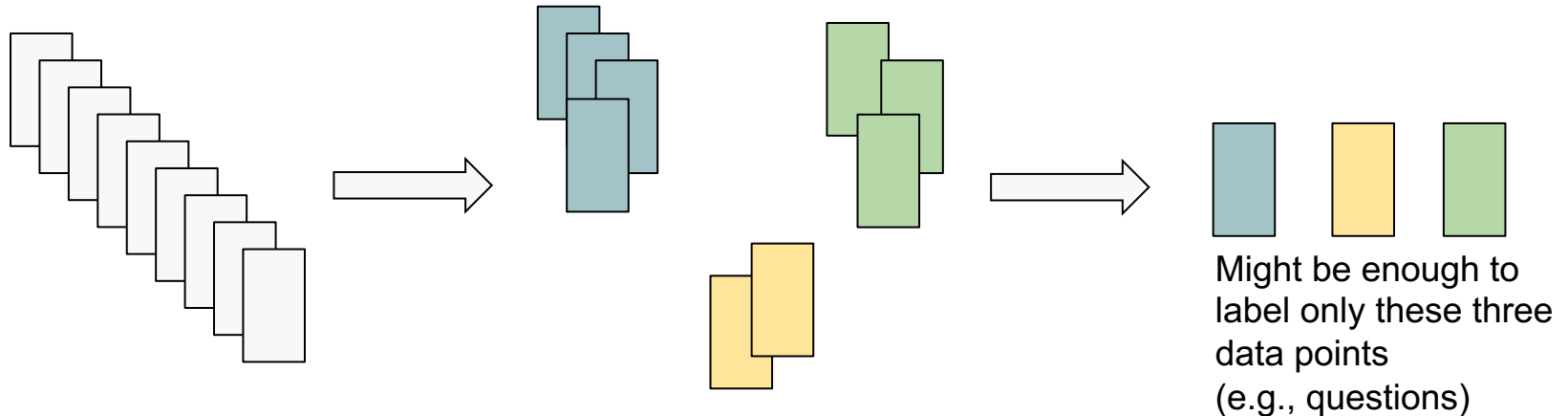
We can use Active Learning
for supervised learning
when we don't have (enough) labeled data.

Active Learning





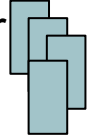
- You have data that could be split into multiple classes (e.g., information-seeking questions vs. rhetorical questions).
- You want to understand whether different classes have interesting/unique patterns.
- But first, you need to determine which data points belong to which class (i.e., you don't have labels yet).
- Here, **active learning might help!**

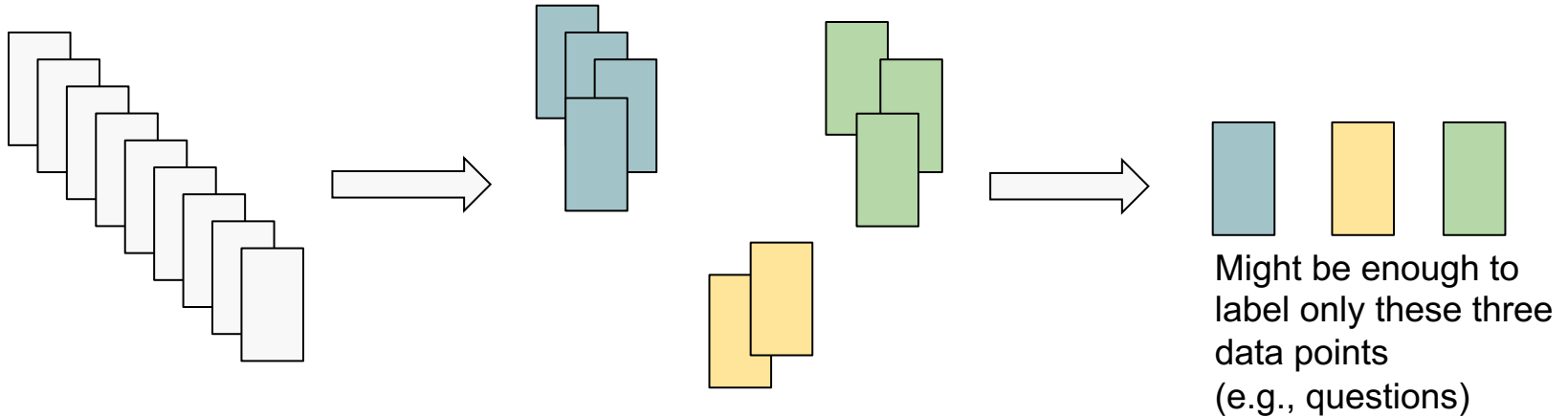
Active Learning

- A special case of Machine Learning for **creating labeled data with a reduced cost.**
- We use machine learning as a tool to **query the annotators wisely.**



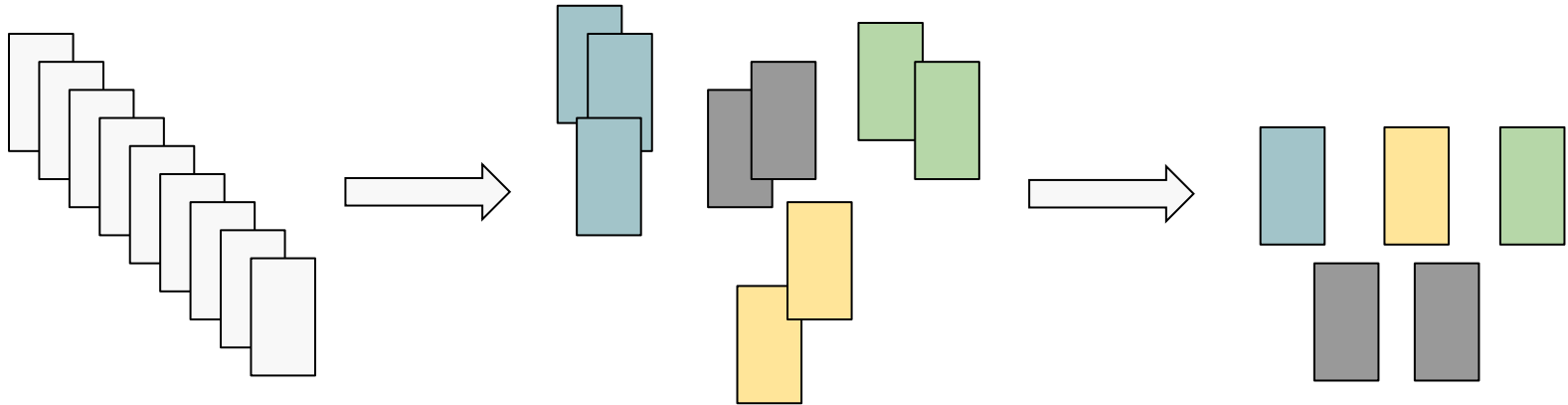
Active Learning

- Why it might be enough? We can use machine learning to learn specific characteristics for  and  and . And then we can assume: if  has a label “x”, then all other  will also have label “x”.



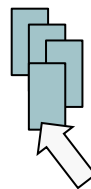
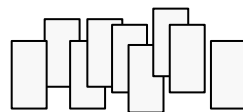
Active Learning

- Usually, data is not so well separable
- Sometimes, we might need more labels
- But still, that would be less work than labeling all data points



Requirements

- Data points (e.g., questions)
- Characteristics that can be automatically or manually extracted from these data points (e.g., pitch values). **Don't worry! There are a plenty of methods/algorithms that we can use!**
- Someone who provides (at least some) labels.



This is a rhetorical question!

Use Case 1: P6 collaboration

- behavioural experiment: 640 conditions, 12 items
- immense labeling effort: time, money, complexity

How can we get “cheaper” labels?



Active Learning

- **Prerequisite:** audio data and its extracted phonetic characteristics
- **Goal:** classify the questions as rhetoric or ISQ
- **Process:** 1. suggest to the annotator the next item to be labeled, depending on whether existing patterns are contradicted or confirmed
2. learn rules about labelled patterns and predict labels for cases that follow the same pattern automatically
⇒ time and effort load are reduced to the minimum

Use Case 2: P2 collaboration (upcoming)

- large corpora, e.g. Asterix, Bible, to be studied for their word order patterns, e.g. position of verb & subject, declarative or interrogative, etc.
- automatic parsing tools can reliably annotate some of these phenomena, but make systematic mistakes in more complex cases

How can we fix these mistakes and apply the fixes on the entire corpus?



Active Learning

- **Prerequisite:** an automatically labeled seed set and a human fixing the mistakes
- **Goal:** create a fully corrected labeled corpus to be used for research questions
- **Process:** 1. observe patterns of common mistakes
2. apply learned fixes to other mistakes or suggest a fix to the human
⇒ time-saving, consistent, avoiding human errors

Use Case 3: P8's own work

- automatic classification of questions into ISQ vs. NISQ requires large labeled data
- labeling is tedious and complex: not even humans agree in some cases

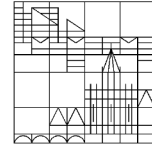
How can we speed up the labeling and deal with disagreements?



Active Learning

- **Prerequisite:** large dialogs labeled for linguistic features, e.g. POS, syntax, semantics
- **Goal:** collect large amounts of reliably labeled data to train final model
- **Support:** 1. automatically label questions with same patterns and forward them to the human for verification
2. suggest the next item to be labeled, based on the created rules
- **Latest Work:** 1. labeled test set of 2000 questions with their contexts
2. integrated distributional semantics into the system
→ experiment with this new dataset and set-up

Universität
Konstanz



Thank you
Questions?
Comments?