Leonhard Hennicke Thursday, 1. November 2018

# AYR Research Task: Crowdsourcing

## Why Do We Need Crowdsourcing?

- machine learning works for some tasks, but not for all
- How to evaluate data? How to ensure meaningful categories?

## Why Use Mechanical Turk?

- dozens of crowdsourcing tools
- most focus on surveys or data preparation for big companies
  - small interface, expensive, only big projects
- mTurk has a very open interface, freely create new tasks, AWS integration
- Christian has experience with mTurk

## Mechanical Turk - Basic Concepts

- **Requesters**: create, approve, pay for tasks, set conditions and rewards
- **Workers**: chose tasks, get evaluated
- **HITs**: Human Intelligence Tasks, e.g. What type of diagram is this?
  - created using simple html templates and csv data sheets
- **Assignments**: decide how many workers should execute each task, accept or decline their results
- **Qualifications**: pay extra fees to chose workers that fulfil specific requirements e.g. Master-Qualification or Employment in Software & IT Services

## Three Use Cases for Image Data

1. **Obtaining Training Data**
   - most likely deep learning for image recognition
     ▶ lots of training data needed, too much to create by ourselves
   - needs to be as cheap as possible:
     - no qualification: high failure rate, lots of assignments needed per HIT to have certainty
     - therefore at least Master-Qualification needed (5% extra fee)
2. **Evaluating Data**
   - given the training data we still need to ensure its quality
   - eliminating outliers but avoiding overfitting - hard to do with machine learning
   - requires relatively deep understanding of domain
   - highly subjective if done by mTurk workers
     ▶ reliable results only for easy domains, but might still be helpful
3. **Clustering**
   ◉ finding keywords in images/diagrams
     - mostly very possible with machine learning, but maybe difficult if not much text - abstract concepts hard to recognise

   ◉ giving meaning to clusters
     - machine learning returns clusters but will probably struggle to differentiate meaningful from meaningless ones

   ▶ both work well enough with crowdsourcing

## Mechanical Turk SDK

- lets us automate everything that we can do via Requester UI on the website
- that way we can integrate crowdsourcing into our tool to patch up shortcomings of machine learning