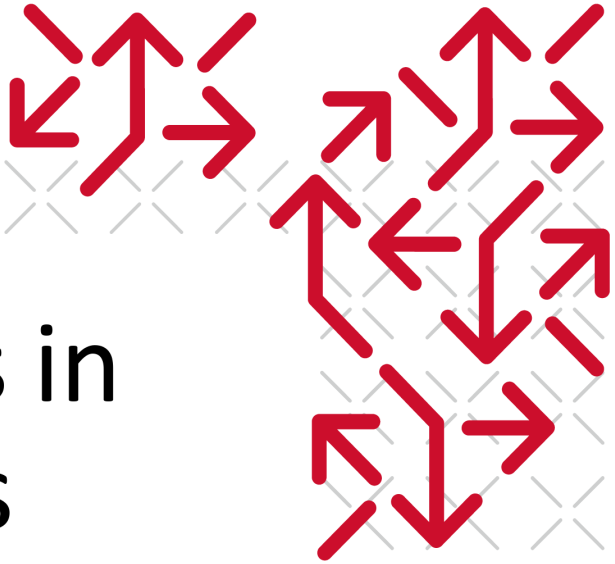
NIM Nuremberg Institute for Market Decisions

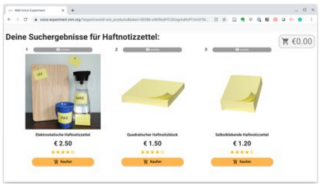


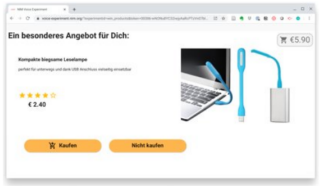
Artificial Voices in Human Choices

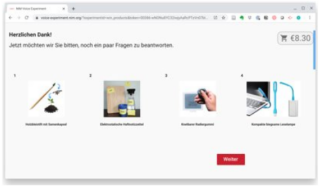
SIXTH MILESTONE

DR. CAROLIN KAISER, RENE SCHALLNER

NIM Nuremberg Institute for Market Decisions







QuestionPro

2

Web Experiment & QuestionPro

TECHNICAL CHALLENGES

3000 participants in 3 parallel experiments on 3 continents

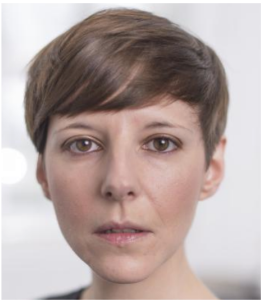
NIM Nuremberg Institute for Market Decisions

- **Backend scaling**
  - Handling 3000+ participants in 3 parallel experiments
  - *(scaling work already done at previous milestone)*
- **English speech synthesis**
  - recording training samples
  - fine-tuning 3 emotion models
  - synthesizing speech in 3 emotions simultaneously
- **Translation of experiment to English**
  - voice prompts & questionnaire
  - 3 different currencies, 3 different pricing schemes
- **Additional Features**
  - QuestionPro- and panel-integration
  - Dynamic stimulus assignment
  - Support for all major browsers

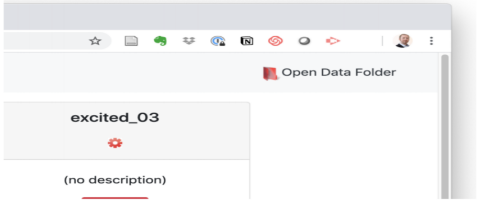
- **Greatest challenge #1**
  - **Only 1 month to go!**
  - 3 months planned
  - luckily, we had worked ahead during our last milestone period
  - we still put ca. 3 months worth into 1
- **Greatest challenge #2**
  - **Unexpected complications with QuestionPro**
  - Technical expertise was hard to find
  - We had to take care of things the QuestionPro platform cannot handle
    - e.g. atomic counting
  - *"All browsers must be supported"*

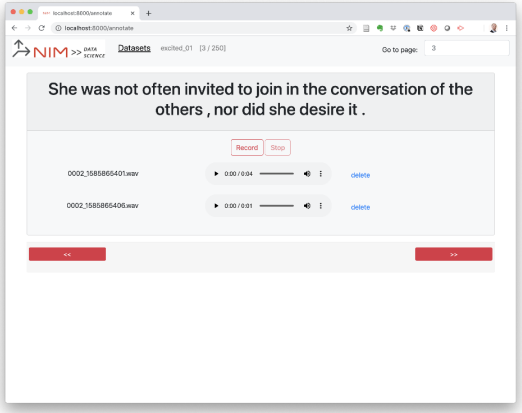
English Speech Synthesis: Recording the Training Samples

NIM Nuremberg Institute for Market Decisions



- We hired a professional voice artist
- She recorded **500** sentences from public domain audio books in the 3 target emotions:
  - excited, neutral/happy, uninvolved
- Using our cross-platform recording software ensured an efficient workflow:
  - [SPACE] to start and stop recording
  - [->] and [<-] cursor keys to jump to next / previous sentence
  - Quick-jump to recording:
    - enter recording number and press [ENTER]
    - when resuming recording in mid-dataset after a break
  - Sending the recordings means just zipping a folder that is opened by clicking a link at the top right
- Quality of the recordings was very high --> no post-production steps needed!





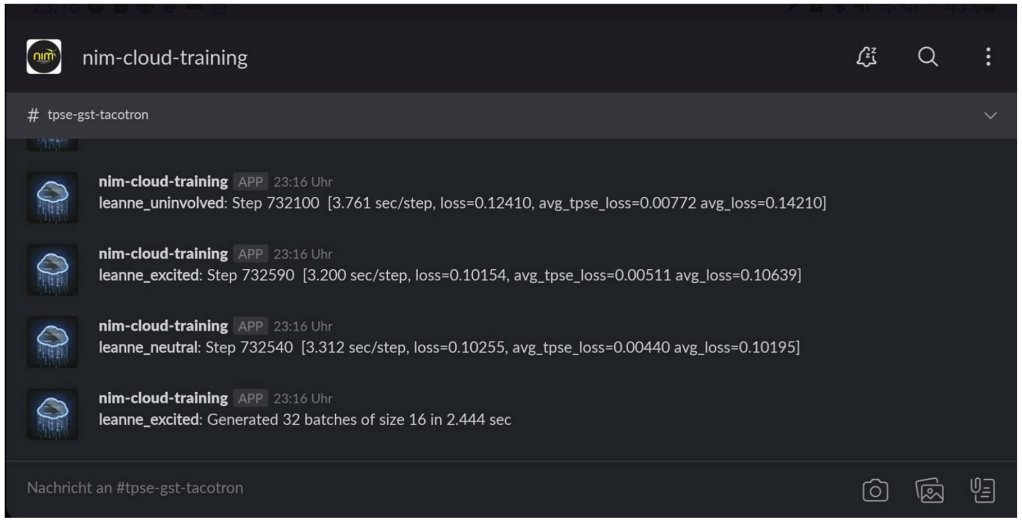
English Speech Synthesis: Fine-tuning the Speech Models

NIM Nuremberg Institute for Market Decisions

> We set up a 4xP100 GPU VM in the Azure cloud

> Pre-processing yielded the following datasets:
  > excited: 500 sentences, 1.28 hours
  > neutral: 500 sentences, 1.38 hours
  > uninvolved: 500 sentences, 1.43 hours

> We fine-tuned all 3 emotional speech models in parallel for ca **48 hours**
  > for uninvolved, batch-size had to be decreased from 16 to 12 because of GPU RAM limits



> Monitoring and automatic backups:
  > To protect us from losing valuable training progress in case anything goes wrong in the cloud
  > Training progress was live-streamed to our private Slack channel #tpse-gst-tacotron
    > --> we were able to monitor the training on-the-go via smartphone app
  > Model snapshots were uploaded to a private Dropbox folder at every 1000 training iterations
    > --> we would always have the latest snapshot for use or for resuming training on another machine

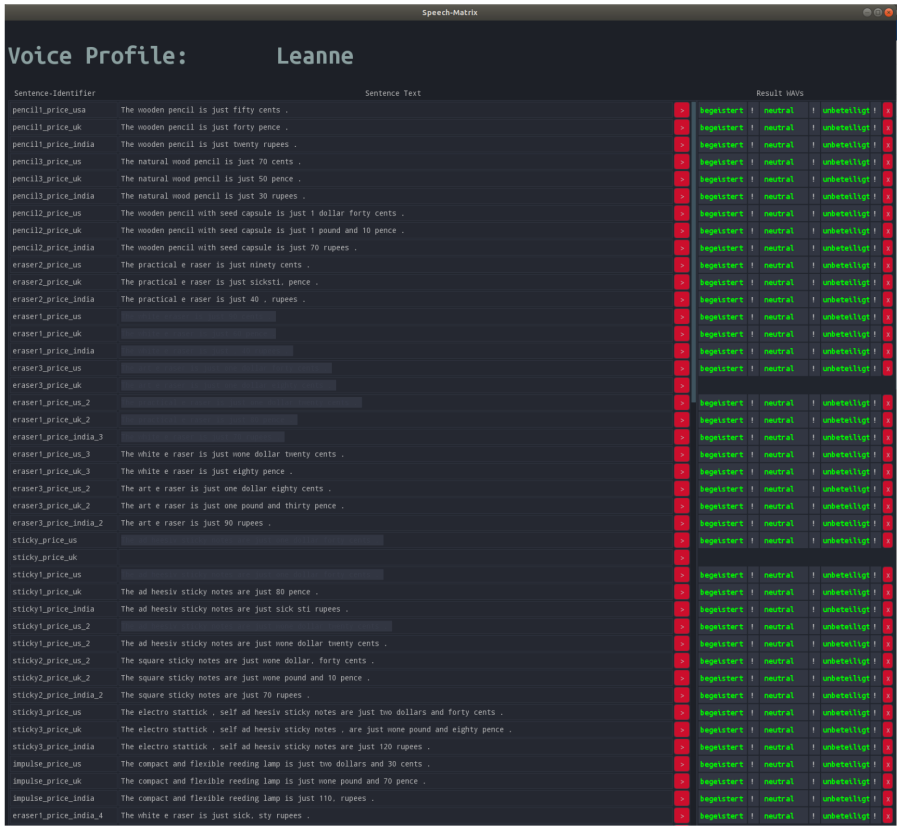
English Speech Synthesis: Synthesizing Speech in 3 Emotions

NIM Nuremberg Institute for Market Decisions

- Using our "speech matrix" tool, we synthesized all sentences in 3 emotions simultaneously on 3 server GPUs:
  - 171 sentences -> audio WAVs:
    - 27 voice prompts x 3 emotions = 81 WAVs
    - 10 prices x 3 emotions x 3 currencies = 90 WAVs

- Azure problems: our 4xP100 training VM "lost" 3 GPUs
  - as with the web server, we moved to Amazon Web Services
  - --> we set up a new VM in AWS with smaller M60 GPUS (like in our workstation) just for synthesis

- Due to high recording quality of training data, the synthesized speech needed ZERO post-production steps. Voice samples could be used as generated!



Integrating our platform with QuestionPro

NIM Nuremberg Institute for Market Decisions

- Instead by clicking on an invitation link, participants get re-directed to our platform from QuestionPro - after a short survery.

- In the survey, their age groups and genders are captured.

- These 2 variables, plus the participant's country are sent to us via URL parameters:
  - *https://voice-experiment.nim.org… &custom1=ID&custom2=AGE&custom3=GENDER&custom4=COUNTRY*

- Instead of tokens from a pre-generated token-pool, we use the passed-in ID as tokens.
  - Since we track token usage, a dummy token pool for QuestionPro was set up.
  - This ID is also passed back to QuestionPro when sending the completion notification.

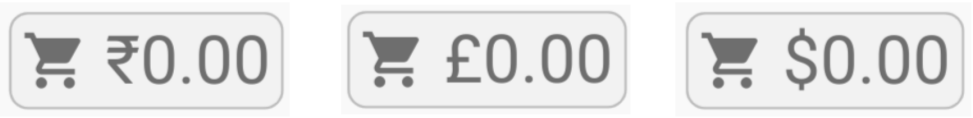
- Since we have 3 experiments, one for each country, the experiment to start is now determined by both the *experiment ID* parameter and the *custom4* parameter from the URL - if the experiment ID starts with *'qpro_'*.

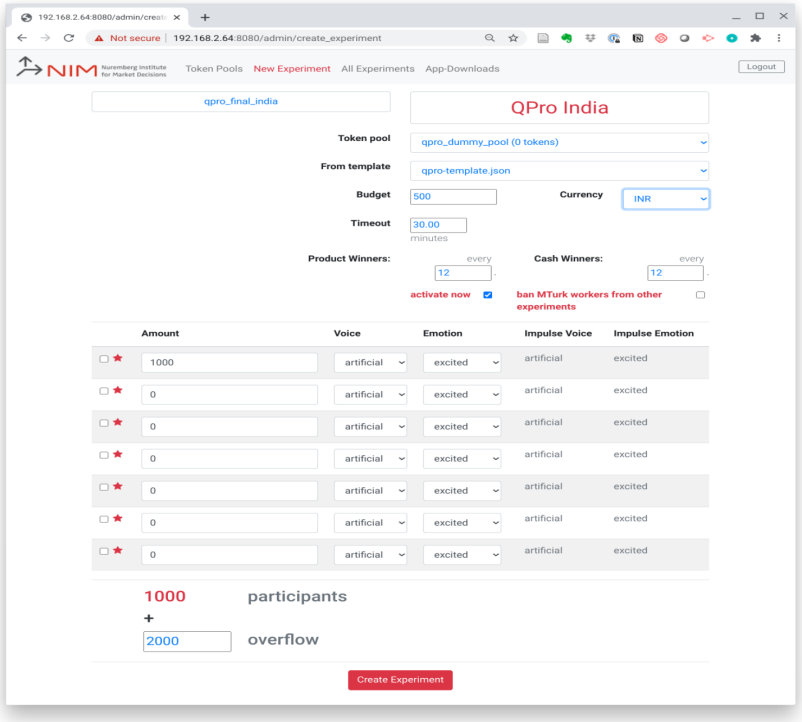
- Our initial questionnaire was shortened, we don't ask for age group or gender anymore.
  - Instead, we save all parameters sent to us by QuestionPro for later use.

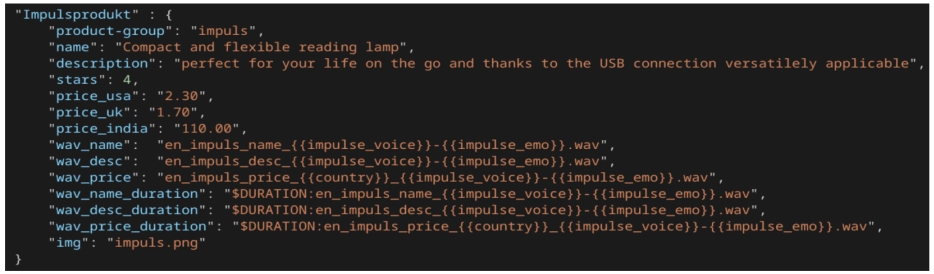
QuestionPro mode and support for currencies

Nuremberg Institute for Market Decisions

- QuestionPro mode is enabled automatically, if the experiment ID starts with qpro_
- For each country, a separate experiment is created
  - each experiment gets its own currency and budget
  - different prices are coded into the JSON template for QuestionPro experiments
- Note that
  - we use an empty dummy token pool
  - we configure 1000 participants with irrelevant voice (will be assigned dynamically)
- The experiment app displays currency symbols based on the configured currency





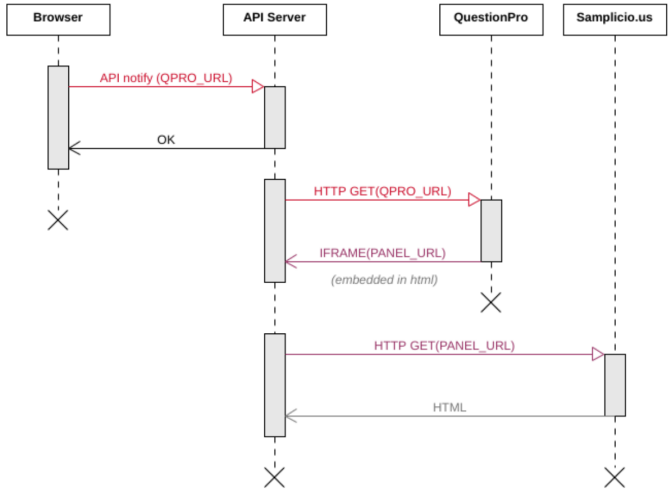


```
"Impulsprodukt" : {
    "product-group": "impuls",
    "name": "Compact and flexible reading lamp",
    "description": "perfect for your life on the go and thanks to the USB connection versatilely applicable",
    "stars": 4,
    "price_usa": "2.30",
    "price_uk": "1.70",
    "price_india": "110.00",
    "wav_name":  "en_impuls_name_{{impulse_voice}}-{{impulse_emo}}.wav",
    "wav_desc":  "en_impuls_desc_{{impulse_voice}}-{{impulse_emo}}.wav",
    "wav_price": "en_impuls_price_{{country}}_{{impulse_voice}}-{{impulse_emo}}.wav",
    "wav_name_duration": "$DURATION:en_impuls_name_{{impulse_voice}}-{{impulse_emo}}.wav",
    "wav_desc_duration": "$DURATION:en_impuls_desc_{{impulse_voice}}-{{impulse_emo}}.wav",
    "wav_price_duration": "$DURATION:en_impuls_price_{{country}}_{{impulse_voice}}-{{impulse_emo}}.wav",
    "img": "impuls.png"
}
```

Notifying QuestionPro

- QuestionPro needs to be informed about the following 3 conditions:
  - **complete**: when a participant finishes successfully
  - **dropout**: timeout after 30min
  - **over quota**: when the experiment is closed because target number of completions has been reached
- QuestionPro use a panel provider:
  - The panel provider needs to be notified, too
  - QuestionPro embed an invisible <iframe> in their HTML to notify samplicio.us

- Our approach to notify QuestionPro and panel
  - API request to our server containing QPro URL
  - for each URL the server acts as a browser and fetches the QPro web page
    - **now QuestionPro is notified**
  - we parse the response from QuestionPro for embedded iframes
  - we take the iframe target URLs and fetch their web page
    - **now the panel provider is notified, too**
  - we store all received HTTP, including headers and status codes in our DB
  - if anything goes wrong, we re-try
  - **==>** we have detailed response logs in the DB from QPro and panel provider requests



Assigning stimuli dynamically

NIM Nuremberg Institute for Market Decisions

Challenge: On the web, we have to deal with drop-outs

- **Hamburg:** *for comparison*
  - pre-configured 1000 wanted participants with fixed stimuli (emotions)
  - pre-configured 9000 more dynamic participants with no pre-assigned stimuli
  - For the first 1000 the stimulus is fixed
    - after that, the stimulus is assigned in a round robin fashion, based on the missing stimuli

- **QuestionPro:**
  - Only dynamic participants, 3000 per country, number of wanted participants is set to 1000
  - We receive age group and gender => quota groups
    - stimuli must be distributed evenly across quota groups
    - So, for each combination of age- and gender group:
      - we keep a counter
      - we dynamically assign the stimuli in a round robin fashion:
        emo_id = counter[age, gender] % number_of_emotions
    - we maintain separate quota counters for each country
    - **Counter increments are atomic!**

Supporting all major browsers

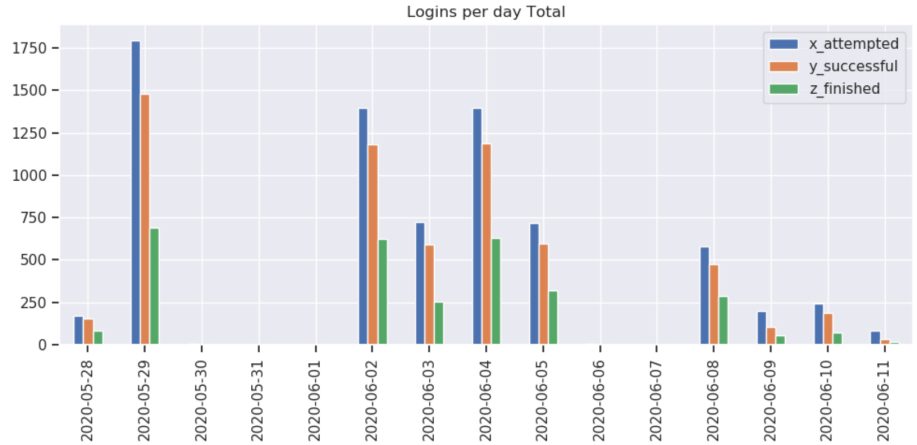
Nuremberg Institute for Market Decisions

- Chrome and (new chromium-based) Edge worked out-of-the-box

- Firefox had issues rendering Markdown with correct line heights and line spacing
  - ***Symptom:*** *Text lines looked cut-off*
  - This is a bug in the Markdown renderer: line hight is not adjusted on lines that contain formatting
  - **Solution:** We patched the open source Flutter Markdown rendering plugin

- Safari had audio problems:
  - It has a very low limit of Audio resources : **only 1 audio resource!** *(compare: other browsers have 8)*
    - ***Symptom:*** *No audio after the audio test screen*
    - In addition, Safari does not follow the web audio standard: shutting down an audio resource and (quickly) opening a new one is not supported, Safari still runs out of audio resources
    - **Solution:** We switched to **re-using a single audio resource**
  - Its JavaScript performance during animations degraded so much that audio timing was affected
    - ***Symptom:*** *speech was cut off at the end*
    - **Solution:** we now **synchronize audio plugin calls with audio timing**
      We start the cut-over timer at the time the audio starts playing, as opposed to after initiating playing

Server load and participant numbers



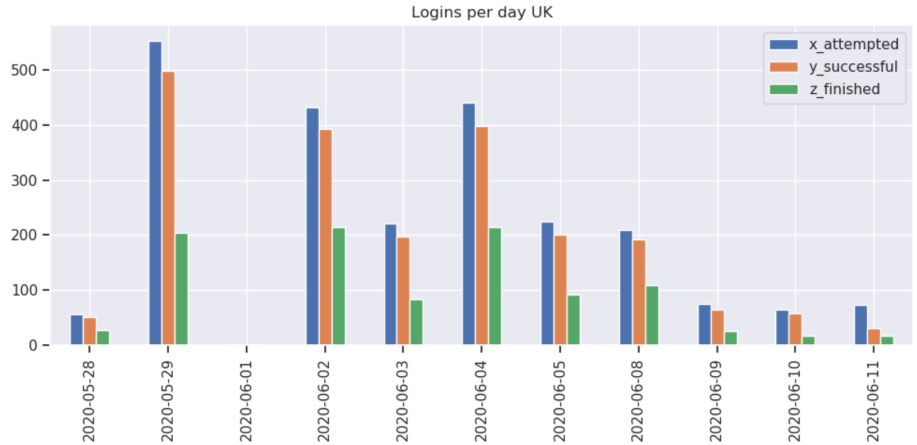
7315
Attempts

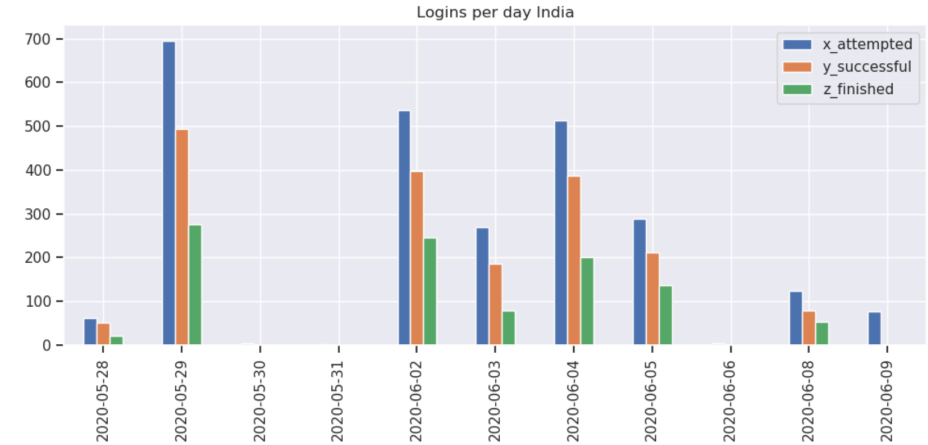
6091
Logins

3161
Partici-
pants

11
days

14 days incl. weekends





Recap of main development efforts for our research

NIM Nuremberg Institute for Market Decisions

- **we researched and implemented** an efficient deep learning method for emotional speech synthesis
  - pre-processing, training, cloud monitoring, …
- **we wrote many tools**, e.g. for:
  - annotation of audio and video datasets
  - recording speech datasets
  - multi-GPU parallel speech synthesis
  - objective evaluation of synthesized speech
- **we developed a pilot study** web app and -server
  - plus an MTurk pilot study
- **we developed the experiment platform**
  - a native Android experiment app for tablets
  - native Android video and heart rate recording apps
  - an entire backend with API server
  - an administration Web interface
- then **we ported it to the web**
- then **we scaled it up** to handle thousands of participants in multiple concurrent experiments
  - and added features like lotteries
- then **we integrated it with QuestionPro** and their panel provider
  - and added support for all major browsers
- not to forget, we **processed and evaluated** all that data
- in the process, **we fixed bugs and submitted patches** to 3 open source projects
- we wrote ca. 23 000 lines of Python, Dart, Go, JS
  - most of which worked flawlessy

NIM Nuremberg Institute for Market Decisions

THANK YOU FOR YOUR ATTENTION

ARTIFICIAL VOICES IN HUMAN CHOICES

SIXTH MILESTONE

DR. CAROLIN KAISER, RENE SCHALLNER

Time spent in experiment