Эмоциональный искусственный интеллект

Тема 9. Данные: сбор и аннотирование

Мария Малыгина

Как разметка встроена в пайплайн машинного обучения

Аннотация (разметка) данных используется:

- данные в задачах обучения по прецендентам

$$X \rightarrow Y$$

Множество объектов (X) и множество возможных ответов (Y), между которыми есть неизвестная зависимость

Обучающая выборка – это совокупность прецедентов, то есть пар «объект, ответ». На ней производится настройка модели зависимости.

Пример прикладной задачи:

- задачи классификации

Этап обучения и этап применения

- Модель построена по обучающей выборке (training sample) → оценка качества этой модели, сделанная по той же выборке оптимистически смещённая (явление переобучения).

Проверка на переобучение

- **Контрольная выборка** (test sample) — выборка, по которой оценивается качество построенной модели.

Если обучающая и тестовая выборки независимы, то оценка, сделанная по тестовой выборке, является несмещенной.

Как задаются объекты и какими могут быть ответы?

X – это сложно устроенные объекты. У нас есть способ получения информации об объектах: признаки (features).

Типы признаков – это способы измерения чего-либо об объектах

- Бинарный признак
- Номинальный признак
- Порядковые признаки
- Количественные признаки

Задачи классификации делятся в зависимости от типов ответов

Предсказывает ответы из конечного множества: делим множество на классы (=label)

- Классификация на 2 класса (лежит или стоит корова)
- Классификация на M классов, которые не пересекаются (распознавание символов)
- Классификация на M классов, которые могут пересекаться (дифференциальная диагностика заболеваний)

Annotation strategy

- Predefined Labels (acted or AU)
- Rater-based assessment only (do not necessarily account for the true experienced emotion
- Self-assessment only (problems emanate from the inter-individual differences in interpretation)

Categorical approach – discrete sets of emotions

Theories of basic emotions R. Plutchik (8 emotions), P. Ekman (6 emotions)

Appraisal theories

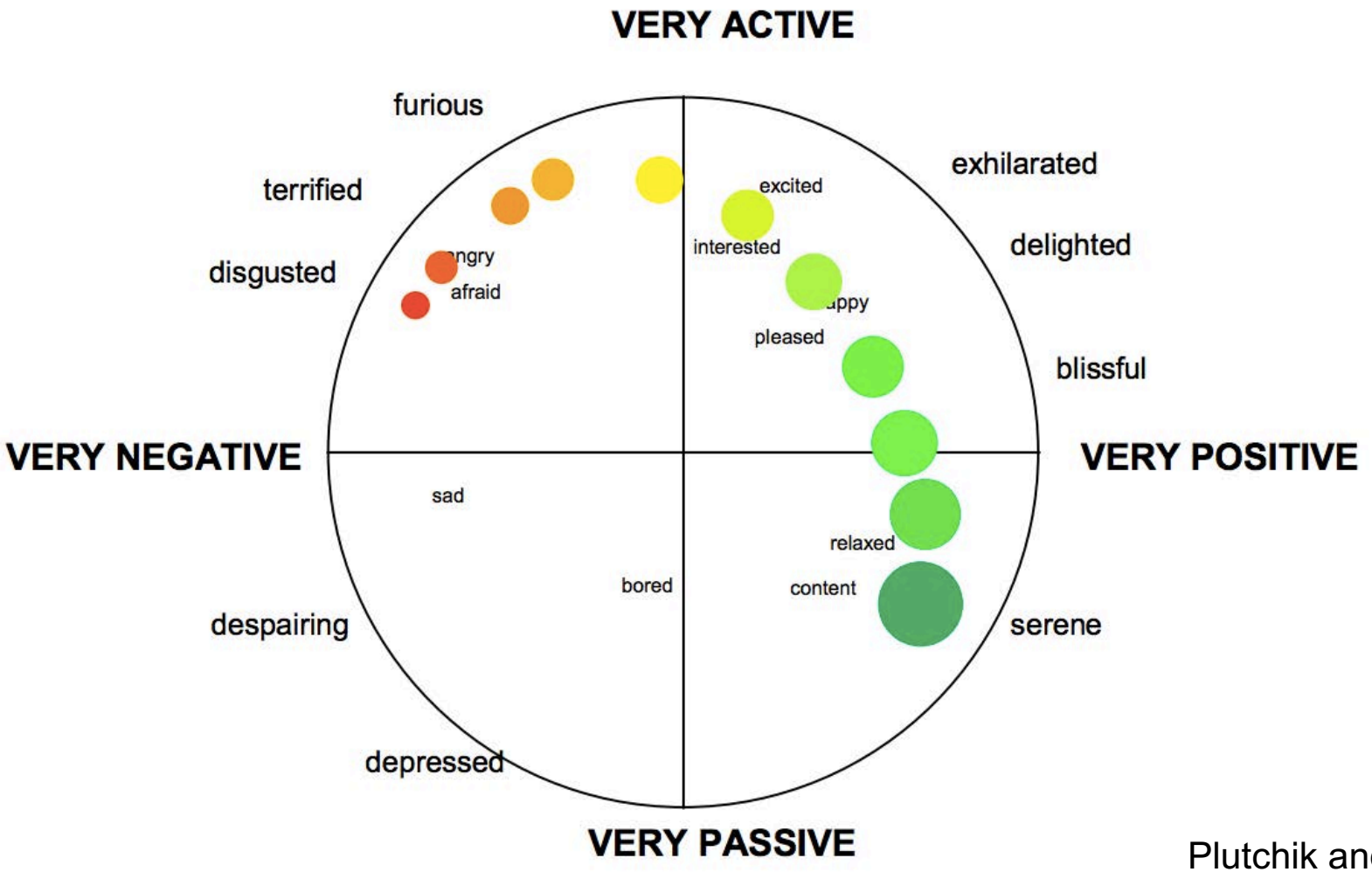
Dimensional approach – combinations of different continuous dimensions

J. Russell Model of Core Affect – a combination of two types of feeling continuums: valence (pleasant to unpleasant) and arousal (low activation to high activation)

Annotation strategy

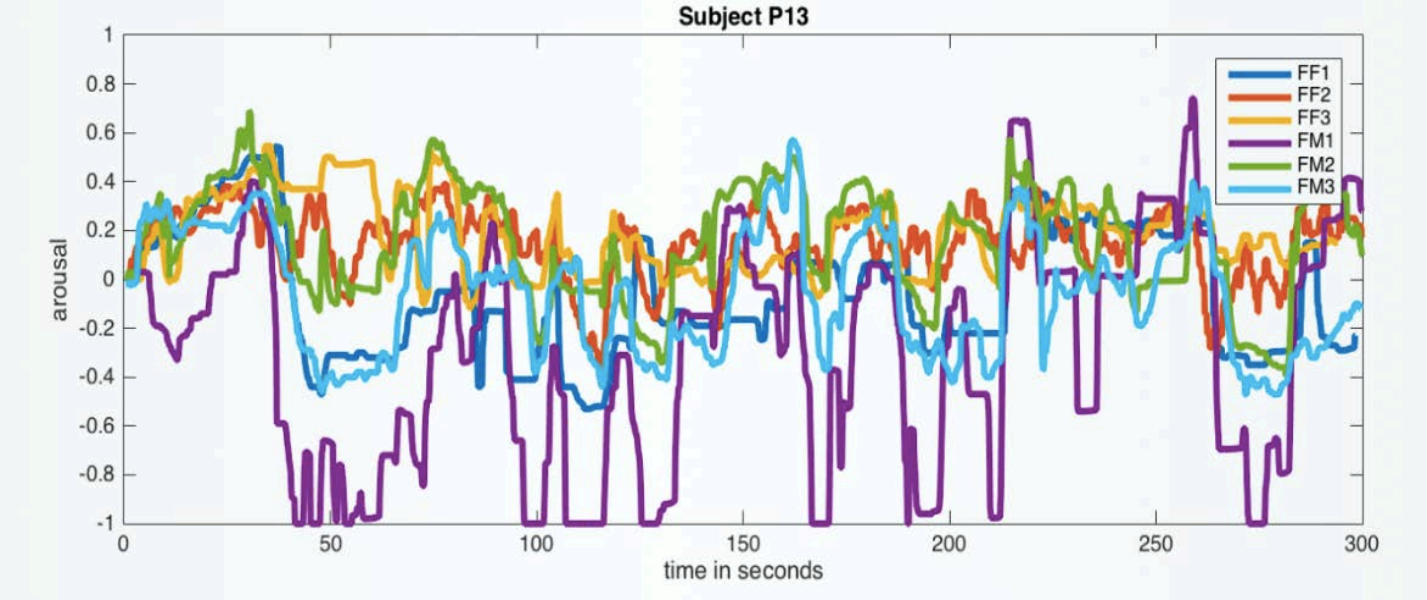
1. Temporal resolution the frame level (e.g., individual frames in a video stream), word level (e.g., individual spoken words), segment level (e.g., sentences or short actions), or session level (e.g., an entire session of interactions)
2. Level of abstraction behavioral expressions (e.g., facial expressions or gestures) or higher level constructs basic emotions (e.g., fear, anger), or nonbasic complex blends of affect and cognition, such as confusion and frustration.

Feeltrace annotation tool (2000)



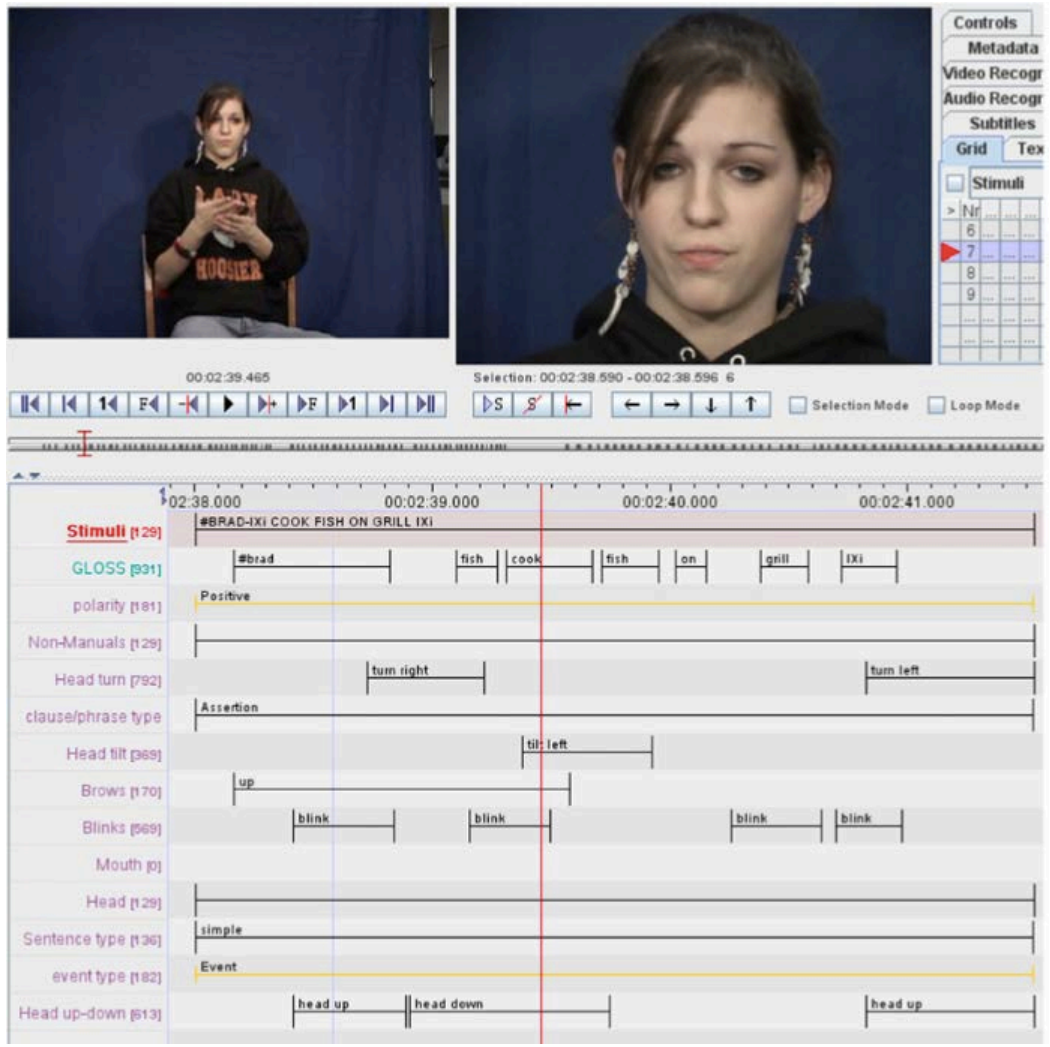
Plutchik and Russell representations

ANNEMO (ANNotating EMOtions) an open-source web-based annotation tool



Dimensional approach for RECOLA dataset
Time-continuous annotation for each affective dimension separately

The Elan tool (Max Planck Institute for Psycholinguistics)

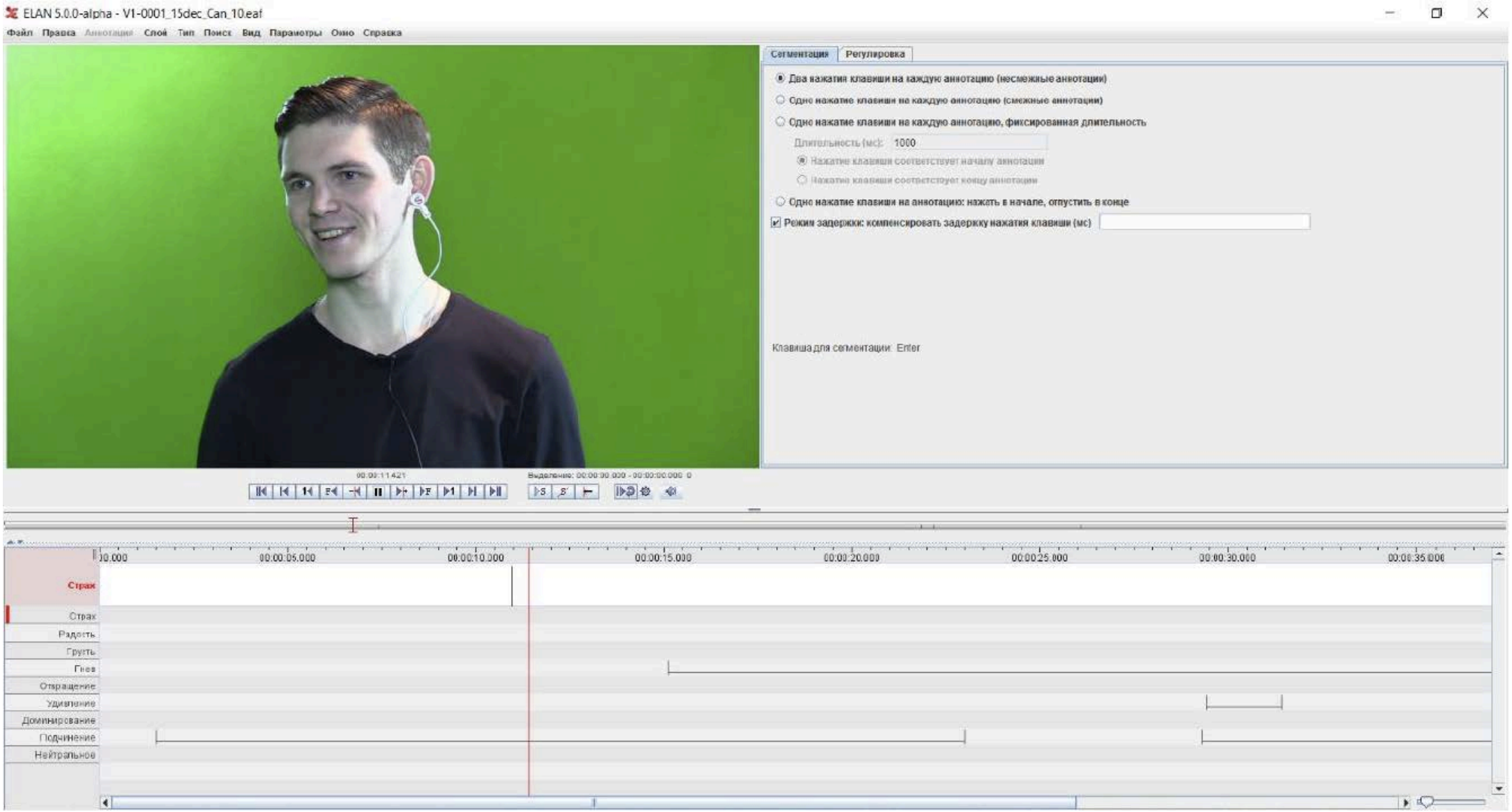


Categorical approach

Ability to set variative length for fragment annotation (i.e. the subject determines the starting and ending point of emotions)

Was used to annotate RAMAS (5 sec videos)

The Elan tool for Russian Multimodal Corpus of Dyadic Interaction



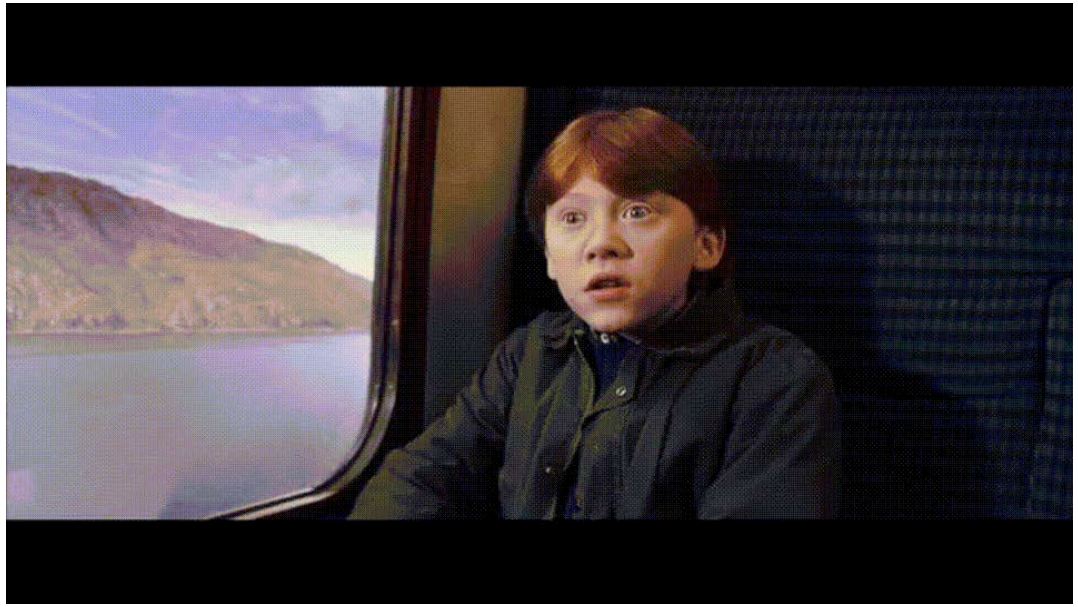
Особенности разметки эмоций. Получение эмоциональных данных

- Acted data – actors portray target affective states (the acted state is the ground truth).
- Induced – эмоциональные состояния, вызванные тем или иным способом. To induce specific affective states using a variety of affective elicitation methods (the induced state is the ground truth).

  Оба подхода опираются на предположение, что внешне наблюдаемые изменения, проявляющиеся в сыгранном действии или возникающие под влиянием индукции, соответствуют действительно возникающей эмоции (адекватно репрезентируют аффект).
- Spontaneous/ in the wild – таких данных на самом деле много, но размеченных практически нет

Acted Facial Expressions in the Wild (AFEW)

Observer-based and assessed with categorical annotation

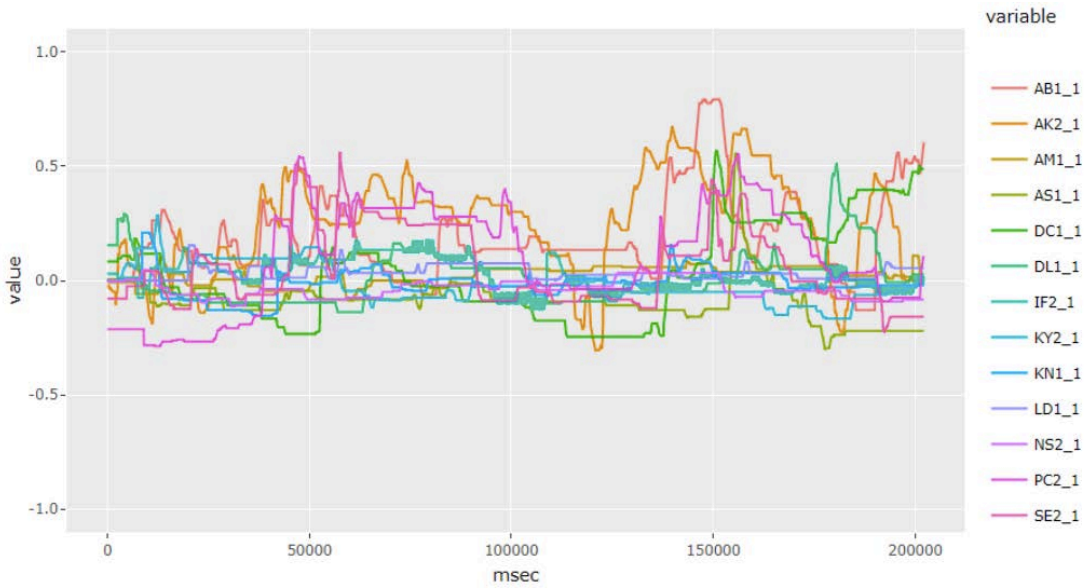


Observers vs self-annotation

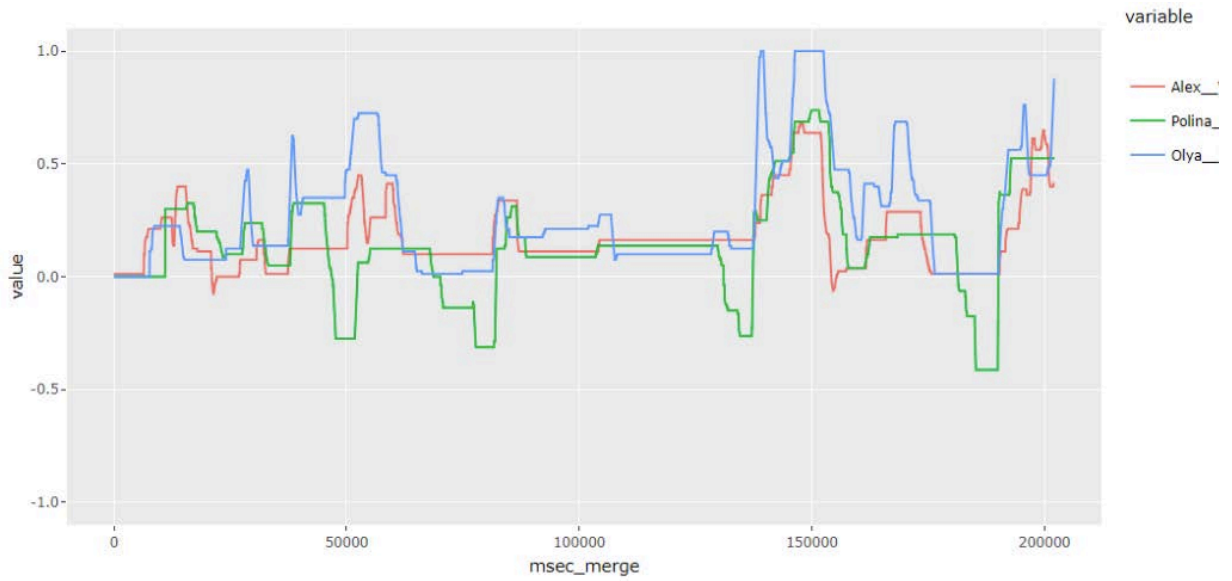
Felt & perceived emotions: one-to-one correspondence between the experience and expression of emotion (?)

- Valence or arousal? Пример: на датасете AMIGOS valence $r=0.44(p < .05)$, and for arousal $r=0.15(p < .05)$ for mimics. дают высокую согласованность по arousal ($r=0.605$), и умеренную по valence ($r=0.264$) для эмоционального и коммуникативного значения кивков и покачиваний
- Positive or negative? ЭМГ-реакции более выражены при просмотре негативных по сравнению с положительными или нейтральными изображениями.

Example: 13 self-annotations



Example: 3 observers annotations



weighted avg	precision	recall	f1-score	support	balanced accuracy
Self-annotation	**0.51**	**0.42**	**0.41**	**172467**	**0.40**
Observers	**0.59**	**0.55**	**0.55**	**188315**	**0.52258**

AMIGOS

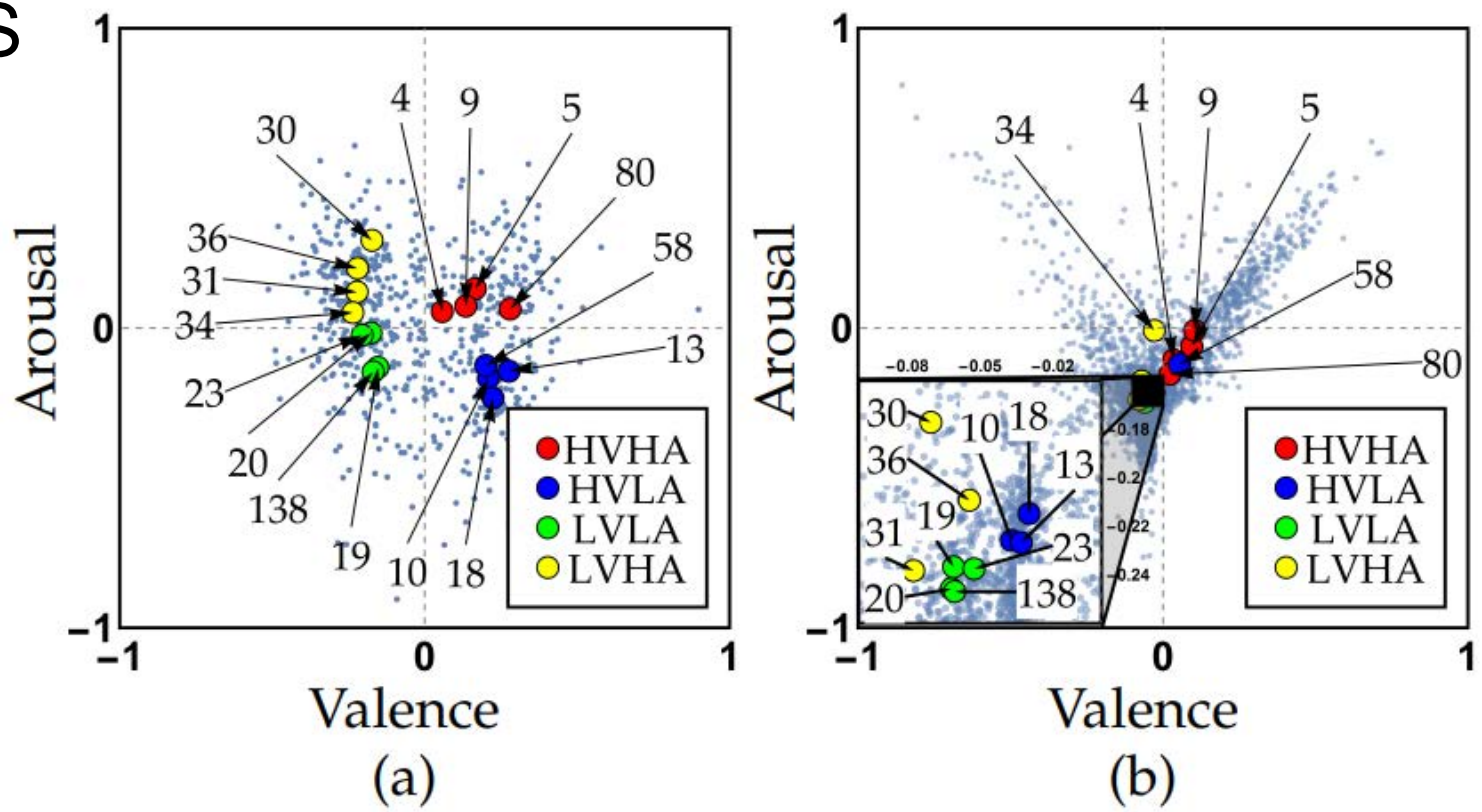


Fig. 4. Distribution of ratings of Valence vs Arousal, for (a) participants' self-assessment of the 16 short videos experiment, and (b) mean external annotations over all annotators for 94 twenty-second segments of the videos of the short videos experiment. Small circles indicate the mean scores over all participants for each of the videos (video ID indicated

Self-annotation vs observer annotation (speech in video games)

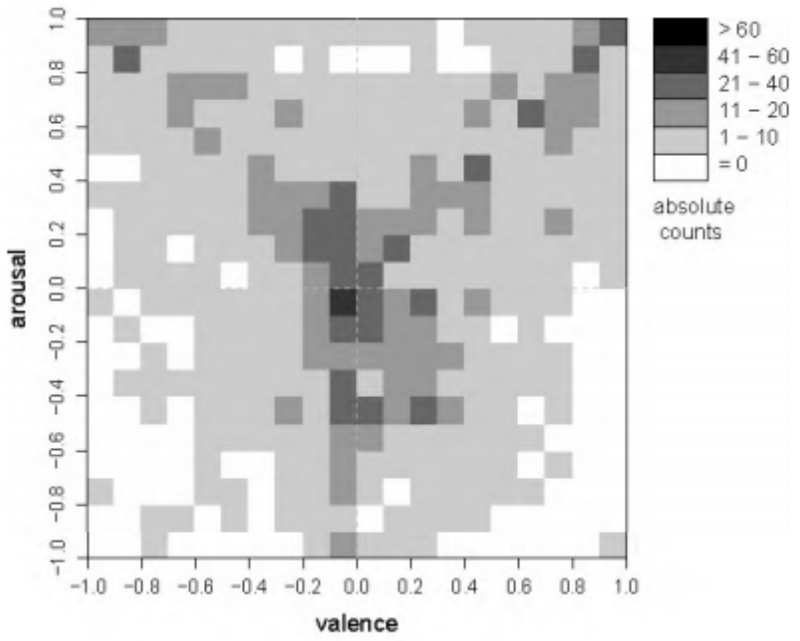


Figure 1: *2D Histogram: the distribution of the 2400 selected speech segments in the Arousal-Valence space, based on the* SELF-*ratings.*

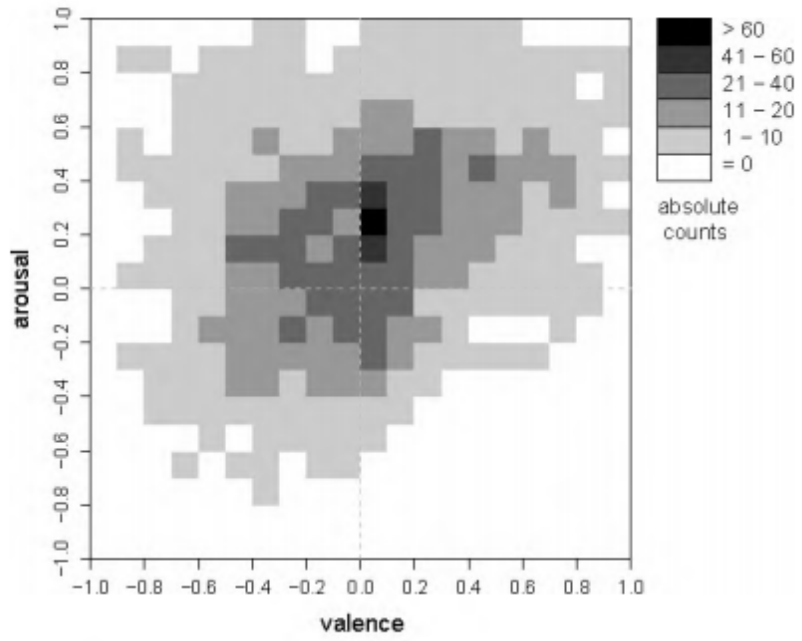


Figure 2: *2D Histogram: the distribution of the 2400 selected speech segments in the Arousal-Valence space, based on the* OTHER.AVG-*ratings.*

Self-assessment relies on

- conscious feelings
- overt actions
- memories of the experience
- meta-cognitive reflections
- unconscious affective components

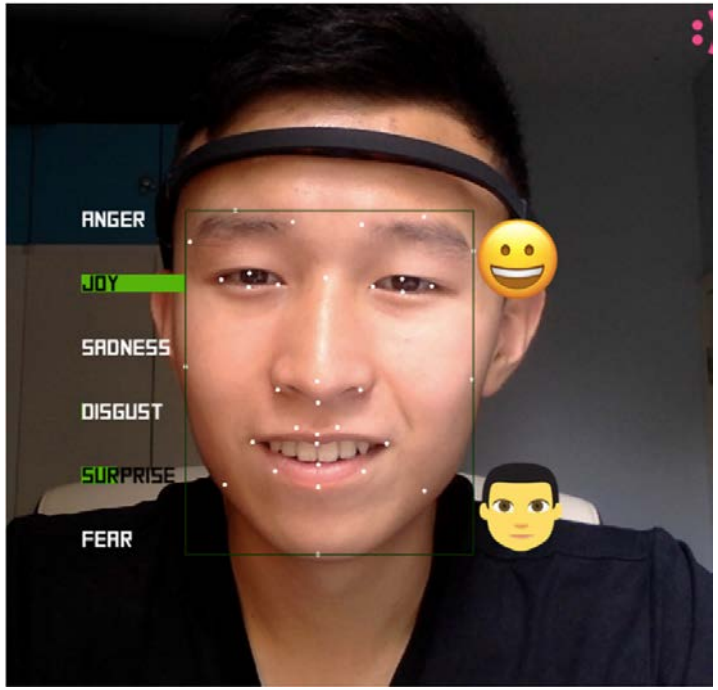
The self must also engage in some form of **reconstructive process** when providing offline annotations of their own affective states. They are also more likely to **distort or misrepresent their affective states** due to biases, such as reference bias or social desirability bias.

Observers have access

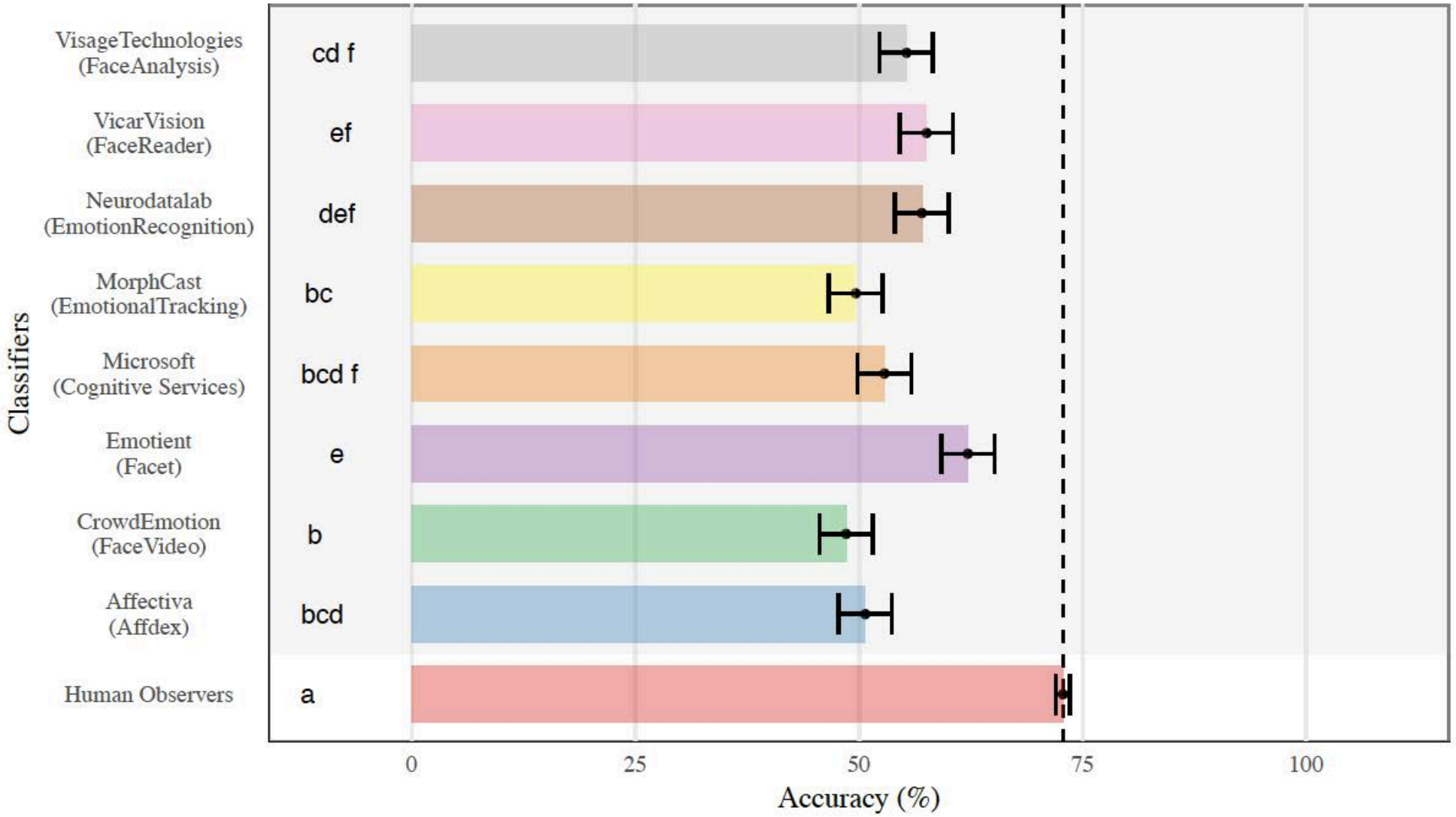
To overt actions and behaviors that can be visibly perceived (e.g., facial features, postures, gestures) and must rely more heavily on inference

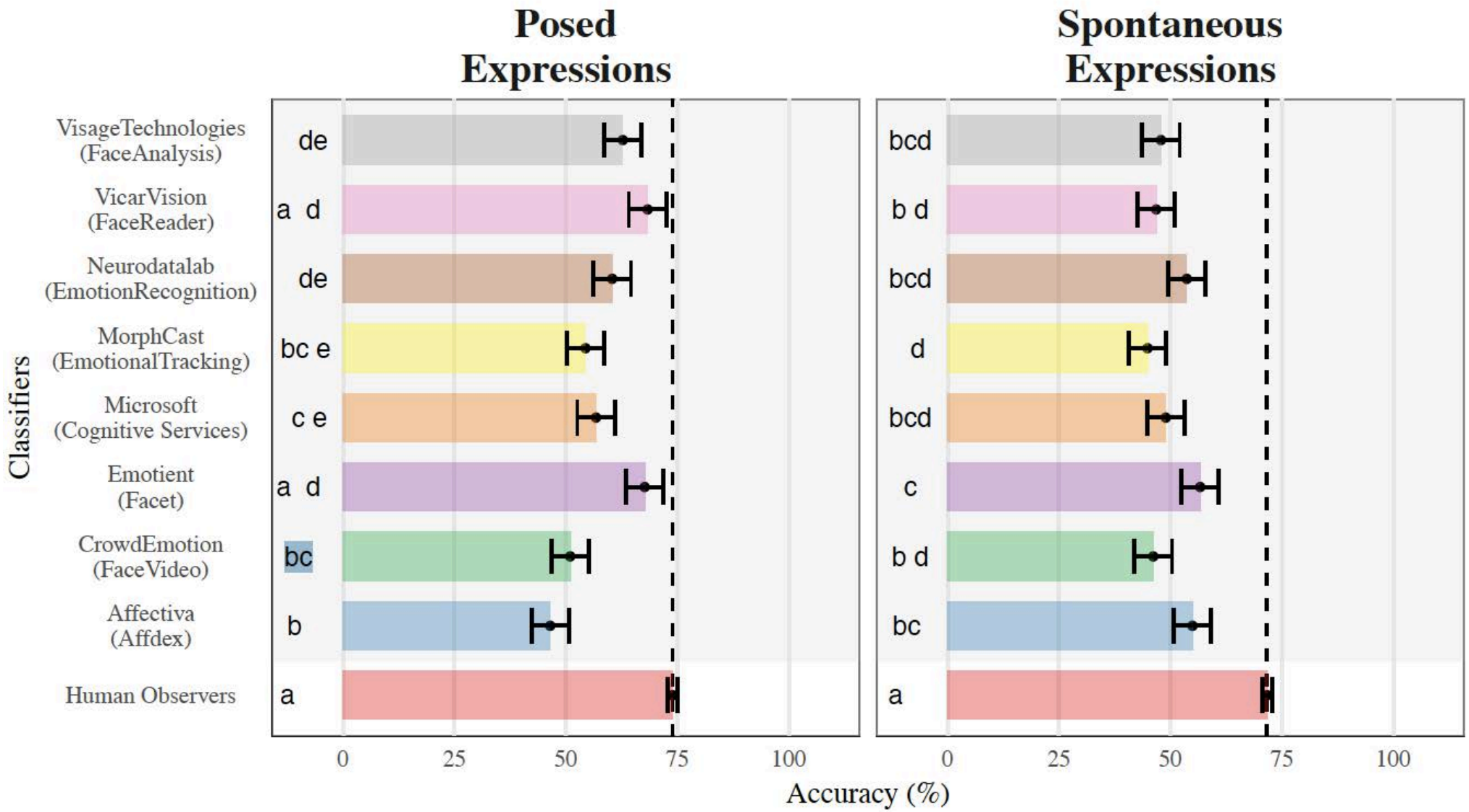
Разметка эмоций по AU

приложение для IOS и Android AffdexMe и AffdexResearch

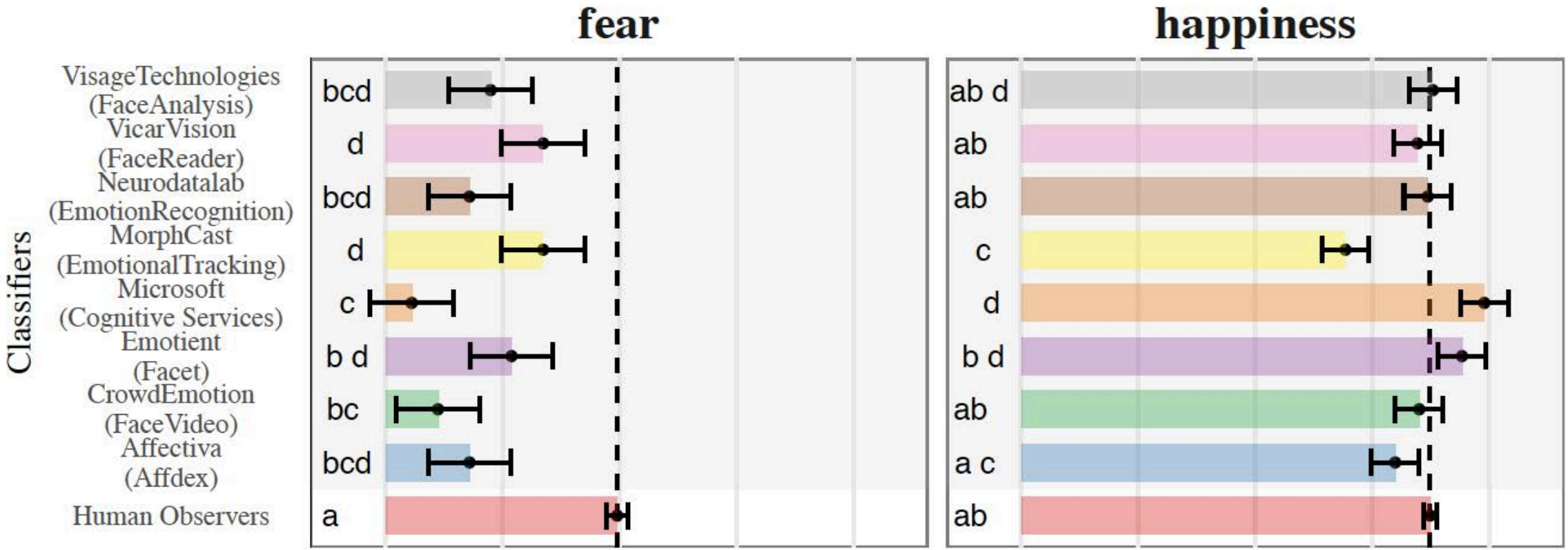


Affective ground truth can never be absolutely obtained




Posed Expressions
Spontaneous Expressions
Classifiers
VisageTechnologies (FaceAnalysis)
VicarVision (FaceReader)
Neurodatalab (EmotionRecognition)
MorphCast (EmotionalTracking)
Microsoft (Cognitive Services)
Emotient (Facet)
CrowdEmotion (FaceVideo)
Affectiva (Affdex)
Human Observers
de
a d
de
bc e
c e
a d
bc
b
a
bcd
b d
bcd
d
bcd
c
b d
bc
a
0
25
50
75
100
Accuracy (%)

Вопрос неоднозначности ground truth



Observer-based annotation

Как измеряется согласованность наблюдателей

The method for calculating inter-rater reliability will depend on the type of data (categorical, ordinal, or continuous) and the number of coders.

Categorical and ordinal data: Kappa

- Kappas in the range of 0 to 0.2 are considered to be slight or poor, 0.2 to 0.4 fair, 0.4 – 0.6 moderate, 0.6 – 0.8 substantial, and 0.8 to 1.0 near perfect
  - Average Obs-Obs kappa of **0.39**
  - Recommended 0.6 kappa for research studies in psychology.

- Continuous data: Intraclass correlation coefficient (ICC), Pearson's r

Observer-based annotation

- small amount of skilled (or expert) observers at a relatively high cost per observer
- large number unskilled (or novice) observers at a relatively low cost

Основные причины некачественной разметки

- неоднозначность данных
- плохое руководство для аннотаторов
- недостаток мотивации или знаний у аннотатора

Краудсорсинговые платформы. Их устройство, преимущества, ограничения

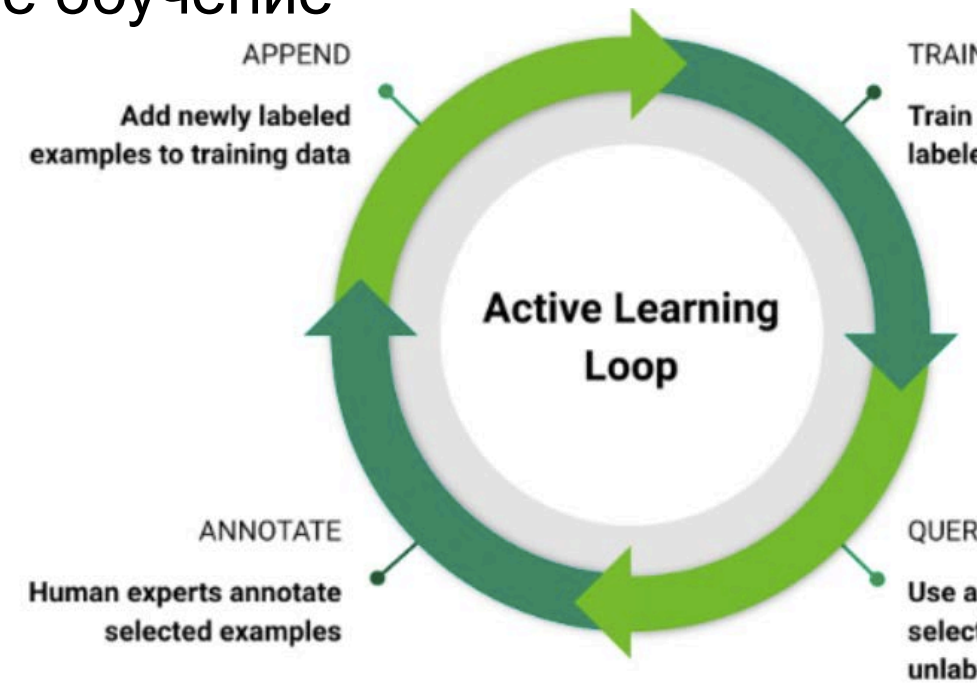
Amazon Mechanical Turk (MTurk), CrowdFlower, Яндекс.Толока

- https://www.mturk.com/
- https://www.crowdflower.com/
- https://toloka.yandex.ru/

- HIT — Human Intelligence Task — задача, которую необходимо решить с помощью человека. Она представляет собой инструкцию для Worker и интерфейс для отправки результата Requester.
- Worker — человек, который решает задачи.
- Requester — человек, который создает задачи.

Оптимизация разметки

- При качественном руководстве для аннотаторов нет смысла в нескольких метках для объекта, если аннотаторами выступают эксперты
- Активное обучение

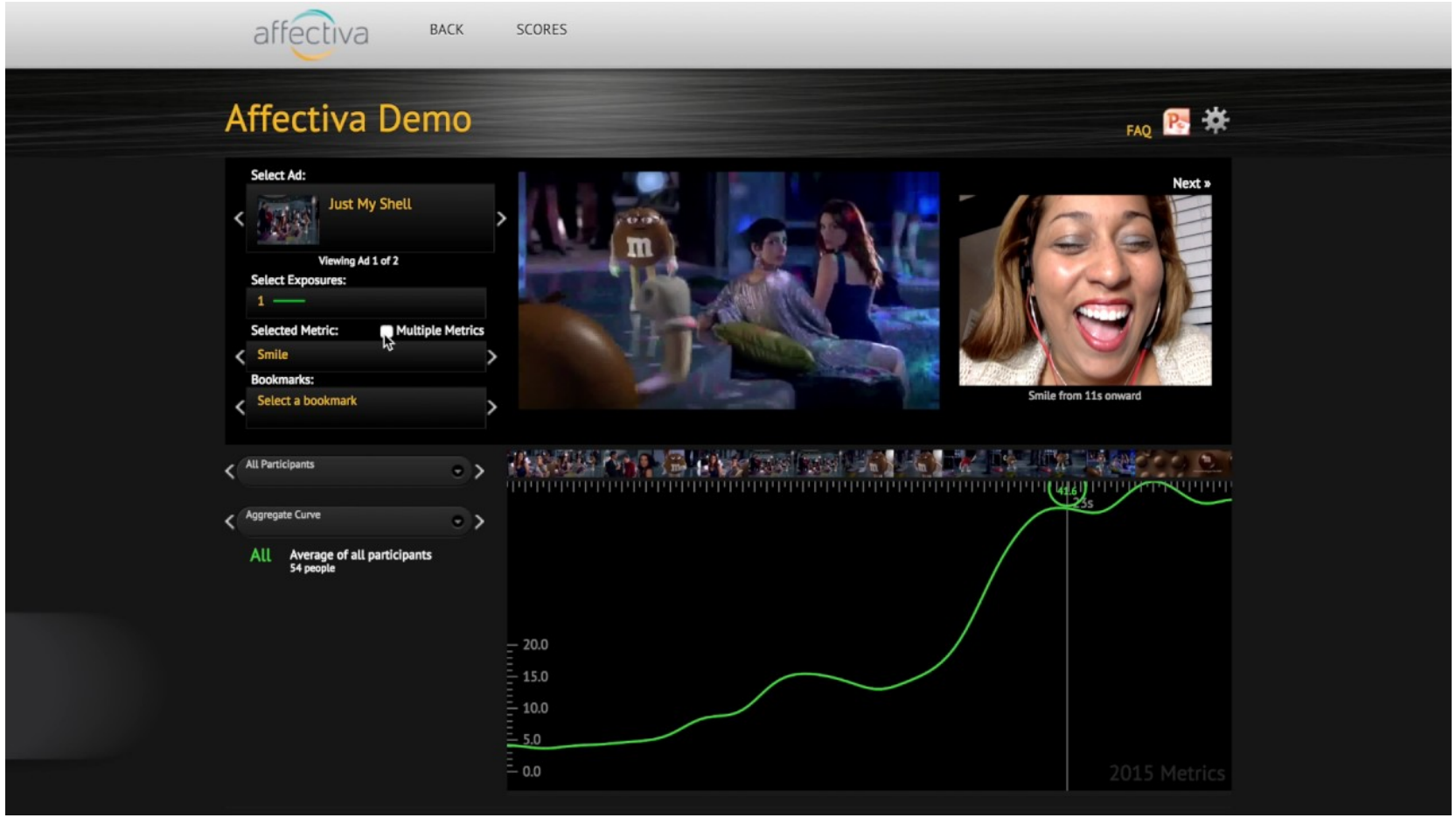


VALUE

- Валентность эмоций операторов кол-центра для автоматической оценки net promoter score (audeering) - лояльность и удовлетворенность обслуживанием;
- Тесты видео для оценки рекламы: engagement - влияет, например, на досматриваемость роликов и виральность рекламы
- Просмотр трейлеров чтобы определить, какой трейлер вызывает наиболее яркую (нужную) реакцию
- Тестирование игр: engagement, насколько человеку интересно

Производные метрики от эмоций (где это нужно и зачем): маркетинг Affectiva и Realeyes

https://www.affectiva.com/product/affdex-for-market-research/



Практическое задание: разметка

https://bit.ly/2HojQ4j