

Affective computing

Аннотирование данных

Малыгина Мария



Данные важны на каждом этапе обучения

- 1 Сбор данных и разметка данных
- 2 Обучение алгоритма на тренировочной выборке
- 3 Проверка качества работы модели на тестовой выборке
- 4 Применение модели к новым данным

Распознавание эмоций в большинстве случаев является задачей обучения по прецедентам (**обучение с учителем**) и относится к типу задач **классификации**.

Как разметка встроена в пайплайн машинного обучения

- Аннотация (разметка) данных используется в задачах обучения по прецедентам.

$$X \rightarrow Y$$

Между объектами (X) и ответами (Y) есть неизвестная зависимость

Обучающая выборка – это совокупность прецедентов, то есть пар «объект, ответ».

Обучающая выборка состоит из размеченных данных. По размеченным данным алгоритм будет подбирать функцию целевой переменной от признаков, наилучшим образом описывающую неизвестную зависимость.

В частном случае

- Во время возникновения или перемены эмоций многие параметры в нашем теле изменяются
 - Дыхание, мимика, потоотделение, ширина зрачка, голос, поза, то, как человек печатает текст и т.д.
- Сигналы улавливаются сенсорами / датчиками
- У нас есть размеченные данные с эмоциями
- Можно находить паттерны в сигналах, которые коррелируют с эмоциями

В частном случае

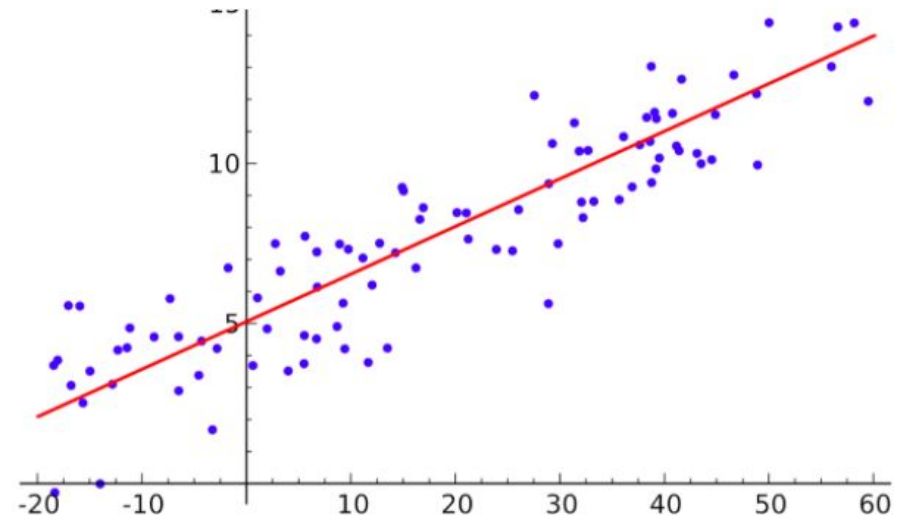
- Во время возникновения или перемены эмоций многие параметры в нашем теле изменяются
 - Дыхание, мимика, потоотделение, ширина зрачка, голос, поза, то, как человек печатает текст и т.д.
- Сигналы улавливаются сенсорами / датчиками
- **У нас есть размеченные данные с эмоциями**
- Можно находить паттерны в сигналах, которые коррелируют с эмоциями

Типичные задачи

Задача классификации – отнесение объекта к одной из категорий на основании его признаков

| | | Predicted class | |
|--------------|-----|----------------------|----------------------|
| | | P | N |
| Actual Class | P | True Positives (TP) | False Negatives (FN) |
| | N | False Positives (FP) | True Negatives (TN) |

Задача регрессии – прогнозирование количественного признака объекта на основании прочих его признаков



1 Подходы к эмоциям на этапе сбора и аннотации данных

Категориальный подход – дискретные наборы эмоций

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

Appraisal theories

Непрерывный подход – комбинации различных измерений

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)

2 Теории эмоций на этапе обучения

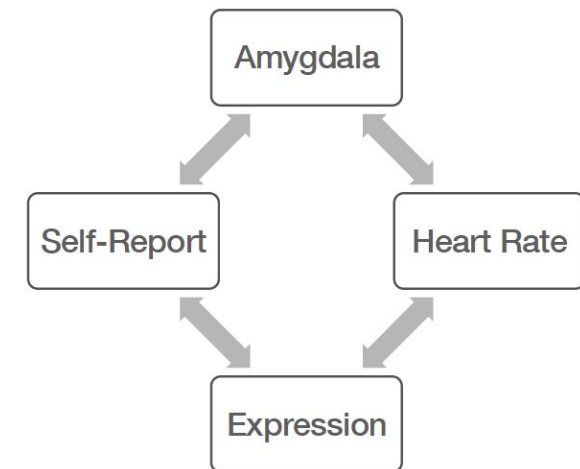
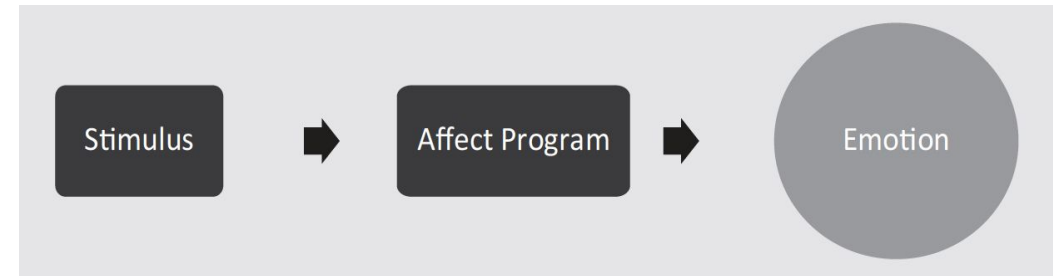
1. Эволюционный подход (S. Tomkins, C. Izard и P. Ekman)

Компоненты эмоций тесно связаны и могут рассматриваться как цепь, связывающая стимулы и реакцию.

- Если все компоненты связаны, то экспрессия лица, физиологическая реакция и ощущаемые эмоции должны четко соответствовать друг другу
- Эмоция может быть измерена любым из компонентов
- Измеренные экспрессии должны предсказывать физиологию и ощущаемые эмоции.

Мультимодальное распознавание должно работать наилучшим образом

=> Если один из компонентов показывает эмоцию X, можем смело утверждать, что она есть



2. Теории когнитивной оценки и теории конструкционизма

Компоненты эмоции жестко не соответствуют друг другу: компоненты влияют друг на друга, но могут быть не синхронизированы.

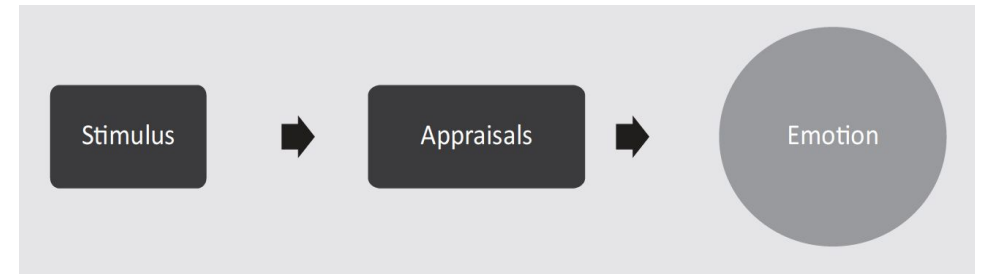
Лицевые экспрессии не должны точно отражать физиологию и ощущаемые эмоции

Теории когнитивной оценки (Ellsworth, Scherer, Frijda): эмоции определяются тем, как человек оценивает текущие обстоятельства.

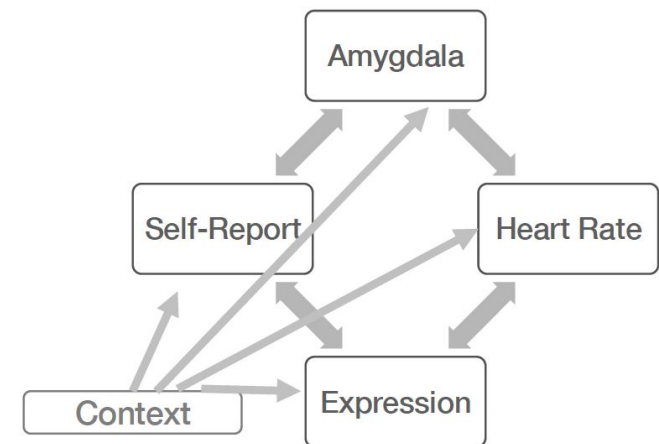
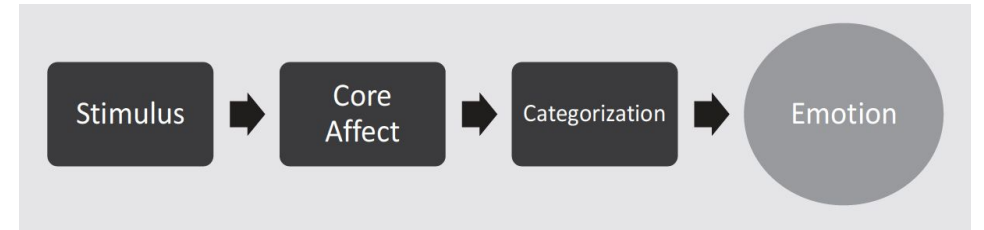
Конструкционистские теории (Barrett): эмоции - это ярлык, который мы присваиваем нашему ощущаемому физиологическому состоянию

=> Нельзя предполагать, что активация в одном компоненте отражает и лежит в основе эмоции X

Appraisal Theories



Psychological Constructionism



Ground-truth emotion

- Это эмоция, наблюдаемая другими людьми, по поводу которой мнения наблюдателей совпадают
- В то же время это эмоция, которую испытывает и называет сам человек (K.P. Truong, D.A. van Leeuwen, and M.A. Neerinx)

Получение эмоциональных данных

- Сыгранные эмоции – актеры изображают целевые эмоциональные состояния (the acted state is the ground truth).
- Индуцированные эмоции – эмоциональные состояния, вызванные тем или иным способом (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).

- Реалистичные (Spontaneous/ in the wild) – таких данных на самом деле много, но размеченных практически нет

Опирается на предположение, что наблюдатель способен определить эмоциональное состояние другого по внешним проявлениям

Не испытывает
эмоции



Happy

RAVDESS

Доступна оценка
переживаемых эмоций



SEWA

Не может дать
отчет об эмоциях



AFEW

Стратегии разметки данных

- Использование заданных меток (сыгранные эмоции или AU)
- Разметка внешними наблюдателями (не обязательно учитывает реально переживаемую эмоцию, также данные аннотаторов часто не согласованы в оценке объектов)
- Разметка самими респондентами (высокая межиндивидуальная вариативность в ответе на один стимул)

Трудности

Люди чувствуют по-разному

Люди оценивают по-разному

AMIGOS Observers vs self-annotation



Воспринимаемые и переживаемые эмоции: насколько переживаемые эмоции соответствуют их внешней экспрессии?

- **Валентность или активация?** Пример: на датасете AMIGOS по мимике лица valence $r=0.44$ ($p < .05$), и arousal $r=0.15$ ($p < .05$). Эмоциональное и коммуникативное значение **кивков и покачиваний**: дают высокую согласованность по arousal ($r=0.605$), и умеренную по valence ($r=0.264$)
- **Позитивные или негативные эмоции?** ЭМГ-реакции более выражены при просмотре негативных по сравнению с положительными или нейтральными изображениями. (+ наши исследования)

AMIGOS

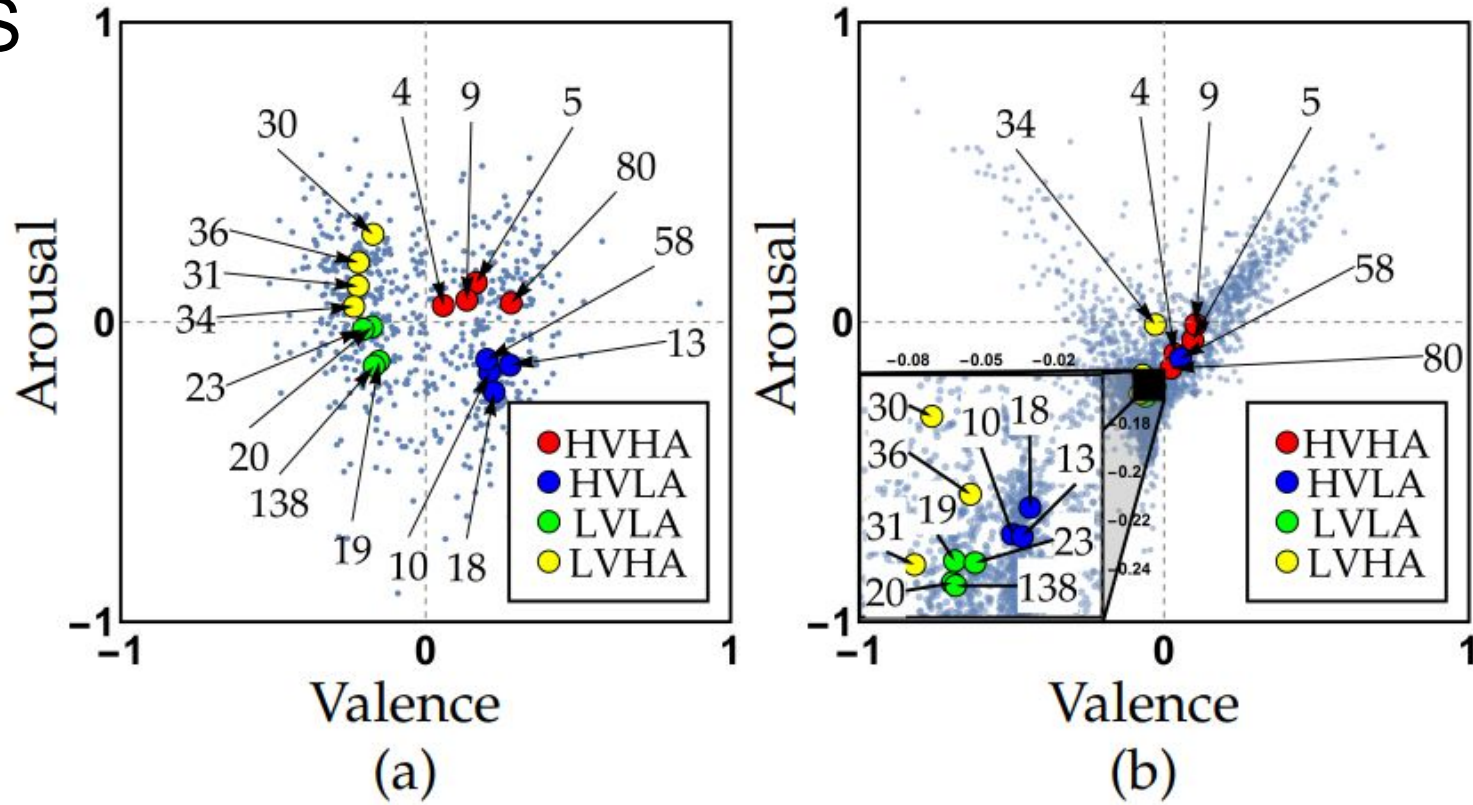


Fig. 4. Distribution of ratings of Valence vs Arousal, for (a) participants' self-assessment of the 16 short videos experiment, and (b) mean external annotations over all annotators for 94 twenty-second segments of the videos of the short videos experiment. Small circles indicate the mean scores over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

TNO-GAMING corpus. Self-annotation vs observer annotation

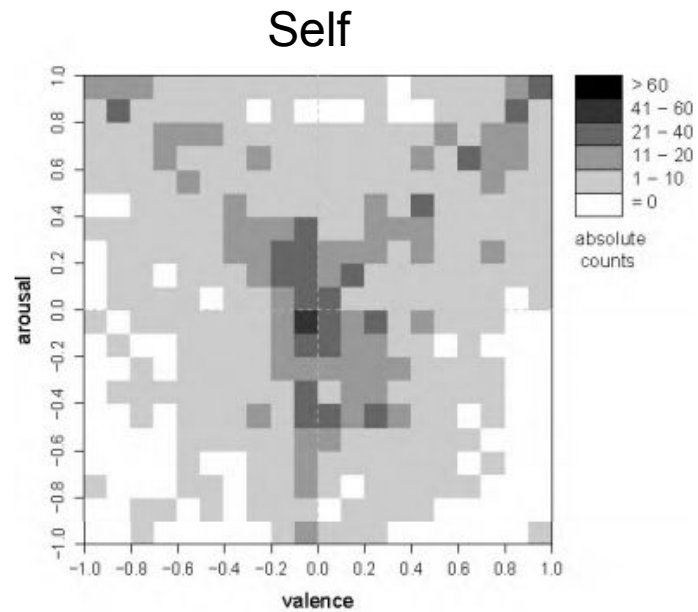


Figure 1: *2D Histogram: the distribution of the 2400 selected speech segments in the Arousal-Valence space, based on the SELF-ratings.*

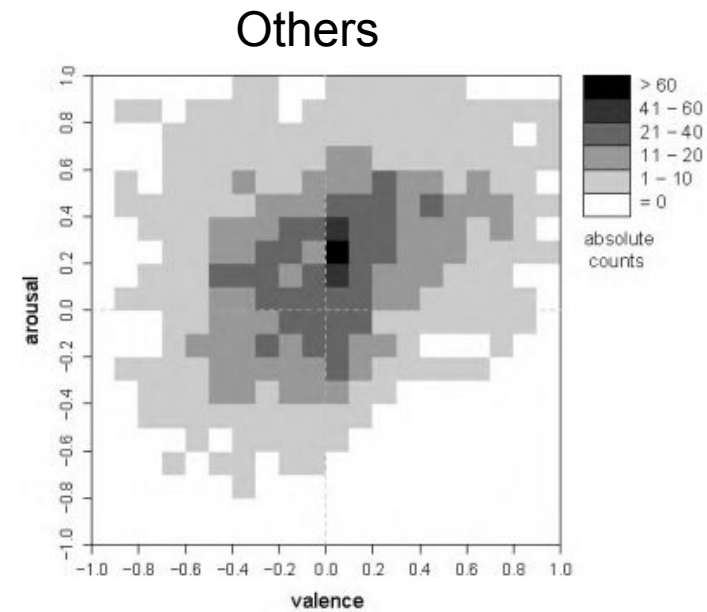


Figure 2: *2D Histogram: the distribution of the 2400 selected speech segments in the Arousal-Valence space, based on the OTHER.AVG-ratings.*

Низкая согласованность felt - perceived.

| | Self – Obs | Obs – Obs |
|---------|------------|-----------|
| Arousal | 0.27 | 0.8 |
| Valence | 0.36 | 0.48 |

TNO-GAMING corpus. Self-annotation vs observer annotation

Модель, обученная на self-разметке, показала результаты хуже, чем модель, обученная на оценках наблюдателей. Воспринимаемые эмоции предсказать легче, чем переживаемые.

Table 3: *Results of Arousal (=A) and Valence (=V) prediction experiments. The baseline results are obtained with a predictor that always predicts Neutrality.*

| | Reference | Test SVR prediction | | | Baseline | |
|---|-----------|---------------------|-------------------------|--------|------------------|-------------------------|
| | | e_{avg} | $\alpha_{\text{ord},5}$ | ρ | e_{avg} | $\alpha_{\text{ord},5}$ |
| A | SELF | 0.41 | 0.22 | 0.25 | 0.45 | -0.07 |
| | OTHER.AVG | 0.21 | 0.42 | 0.55 | 0.31 | -0.18 |
| V | SELF | 0.36 | 0.10 | 0.18 | 0.36 | -0.01 |
| | OTHER.AVG | 0.26 | 0.28 | 0.41 | 0.28 | 0.00 |

Self-assessment relies on

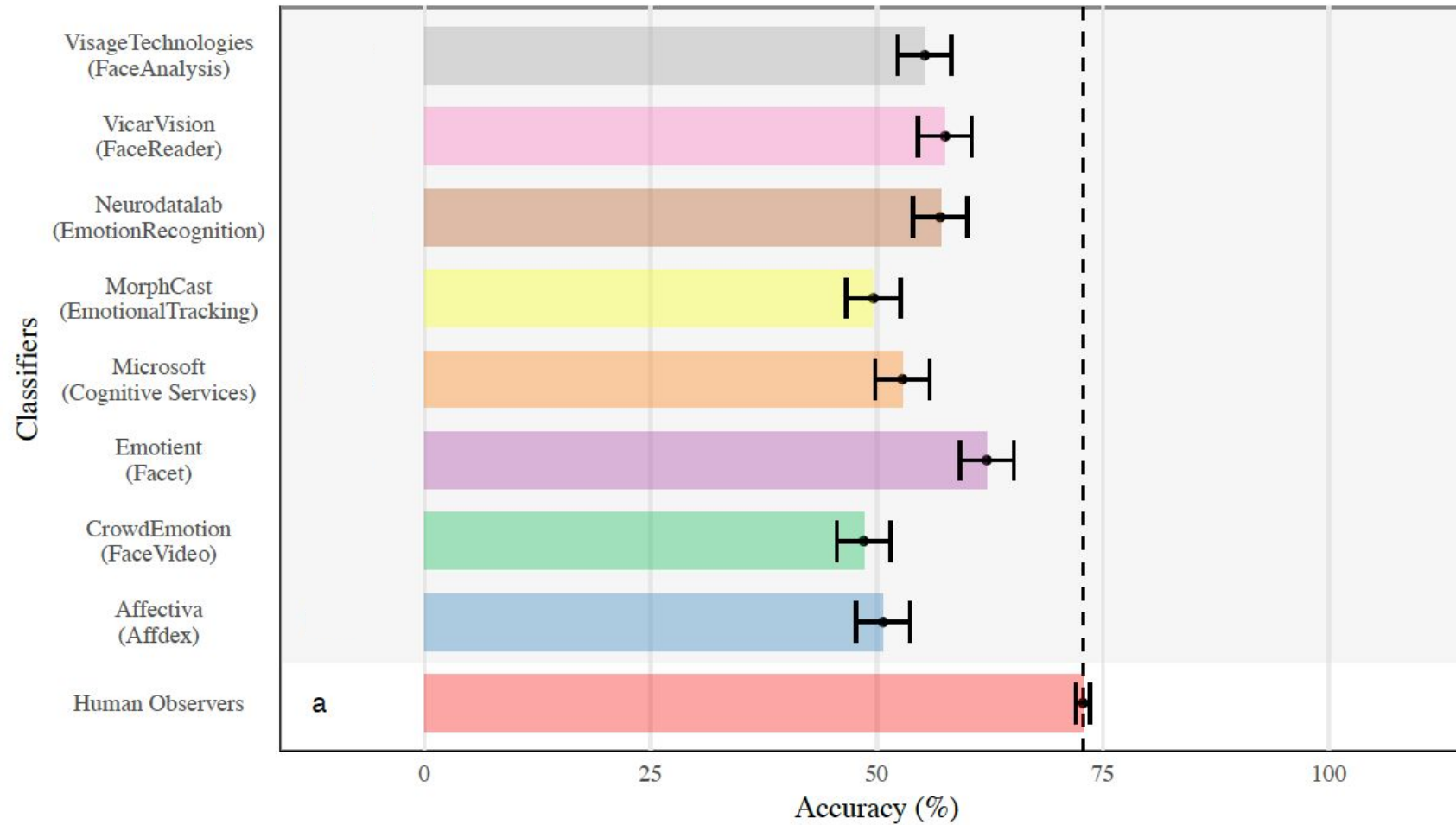
- conscious feelings
- overt actions
- memories of the experience
- meta-cognitive reflections
- unconscious affective components

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**

Информация об аффективных состояниях (ground truth)
не может быть по-настоящему получена



Процент «ошибки» людей → процент «ошибки» в данных (ground truth)

Как измеряется согласованность

how well multiple annotators can make the same annotation decision for a certain category

Способ подсчета согласованности будет зависеть от типа данных (категориальные, порядковые или интервальные) и количества аннотаторов.

Категориальные и порядковые: Карра

- Карра в диапазоне от 0 до 0.2 считается низкой, от 0.2 до 0.4 приемлемой, 0.4 – 0.6 средней, 0.6 – 0.8 высокой и 0.8 – 1.0 очень высокой.
 - Средняя Obs-Obs карра составляет 0.39
 - Для исследований в области психологии рекомендованное значение карра 0.6
- Интервальные: Intraclass correlation coefficient (ICC), Pearson's r

| Метрика | Условия применения | |
|-------------------------------------|--------------------|--|
| Cohen's Kappa | 2 аннотатора | Все объекты оценены каждым аннотатором. |
| Krippendorff's Alpha, Fleiss's Kapp | N аннотаторов | Каждый объект не обязательно аннотируется каждым аннотатором. Допустимы пропуски в данных. |

Python

- `import sklearn`
`from sklearn.metrics import cohen_kappa_score`
- `import statsmodels`
`from statsmodels.stats.inter_rater import fleiss_kappa`
- <https://pypi.org/project/krippendorff/>
- <https://pypi.org/project/irrcac/0.2.0/>

Разметка внешними аннотаторами: детали

Существуют разные стратегии аннотации данных, исследователи выбирают баланс между количеством аннотаторов (то есть количеством оценок на один объект) и степенью экспертности аннотаторов (то есть качеством, достоверностью их оценки). Для сбора разметки популярны краудсорсинговые платформы, такие как Mechanical Turk.

Основные причины некачественной разметки

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

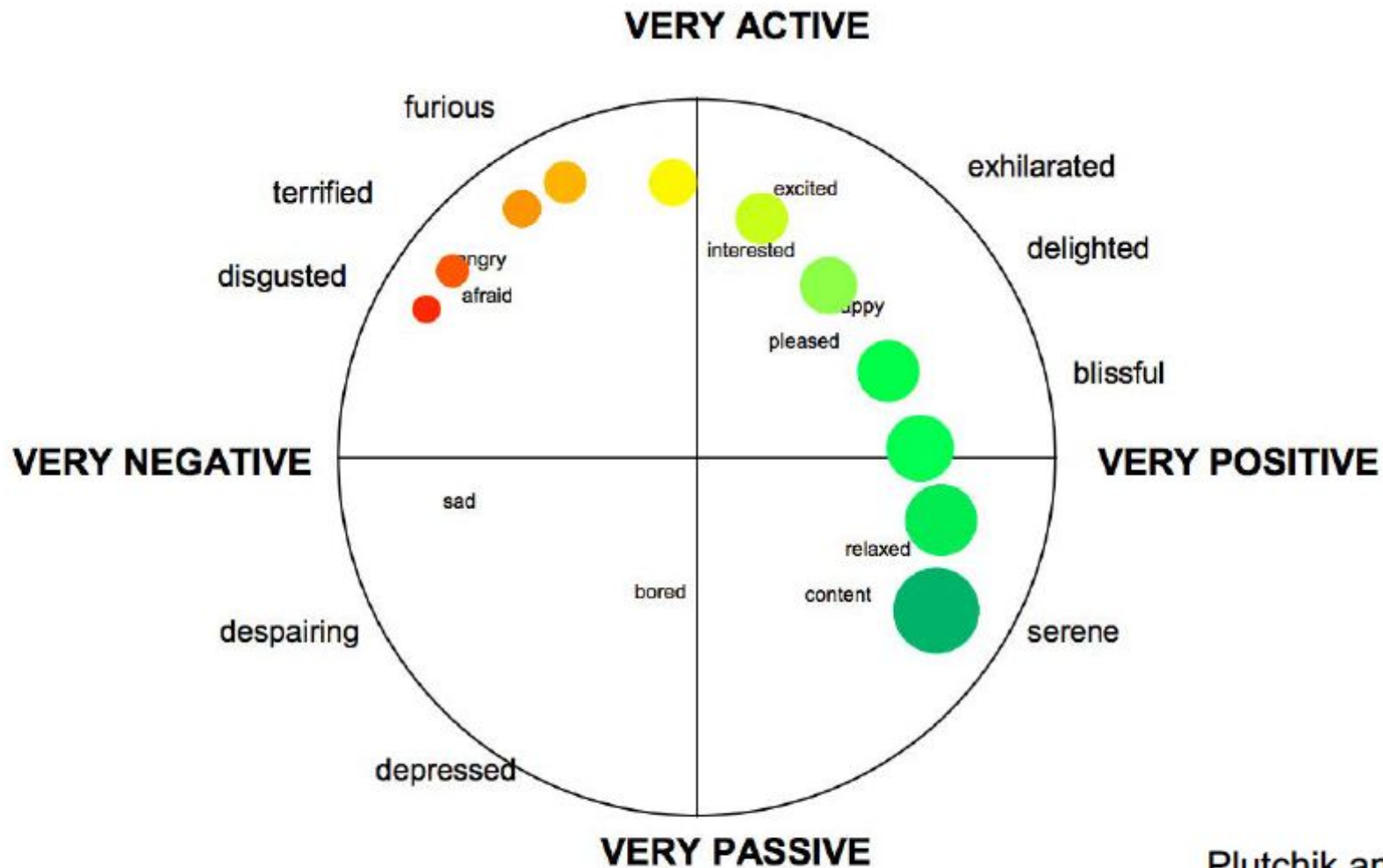
Приложения для аннотации эмоций. Примеры open source веб-приложений

https://github.com/JeanKossaifi/valence_arousal_annotator/wiki

<https://github.com/phuselab/DANTE>

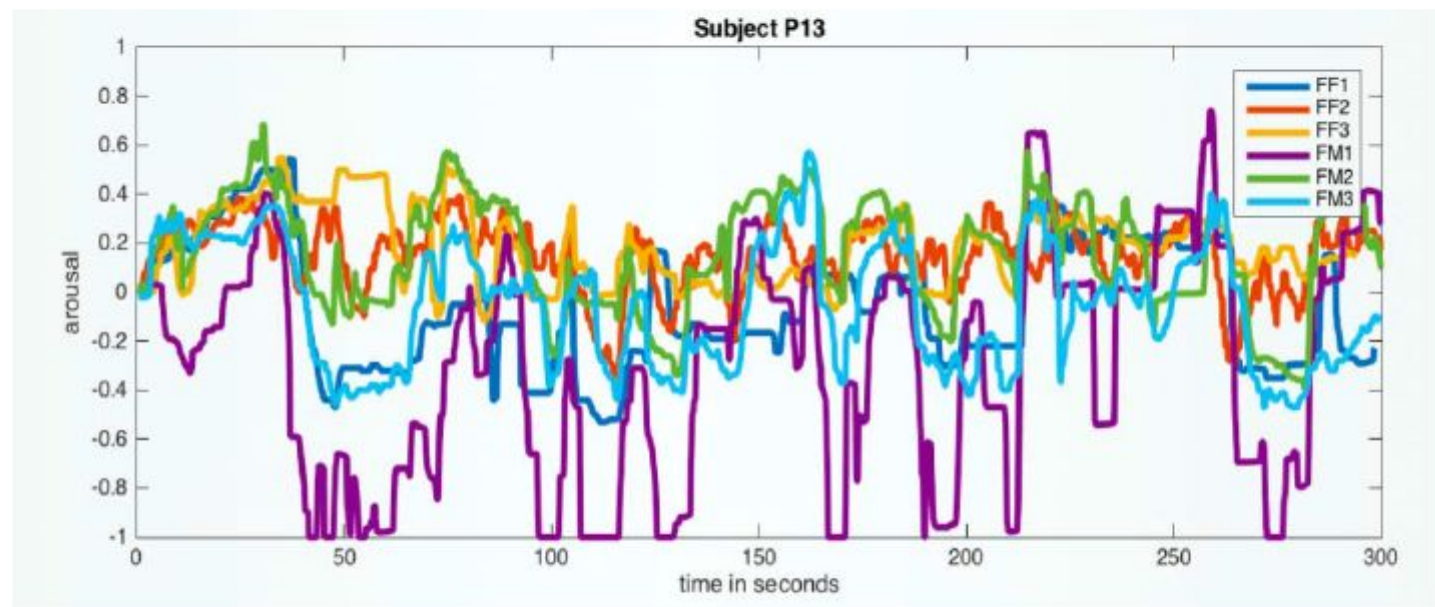
ELAN, ANVIL, Continuous Measurement System, EmuJoy, FeelTRACE, Gtrace, DARMA, CARMA

Feeltrace annotation tool (2000)



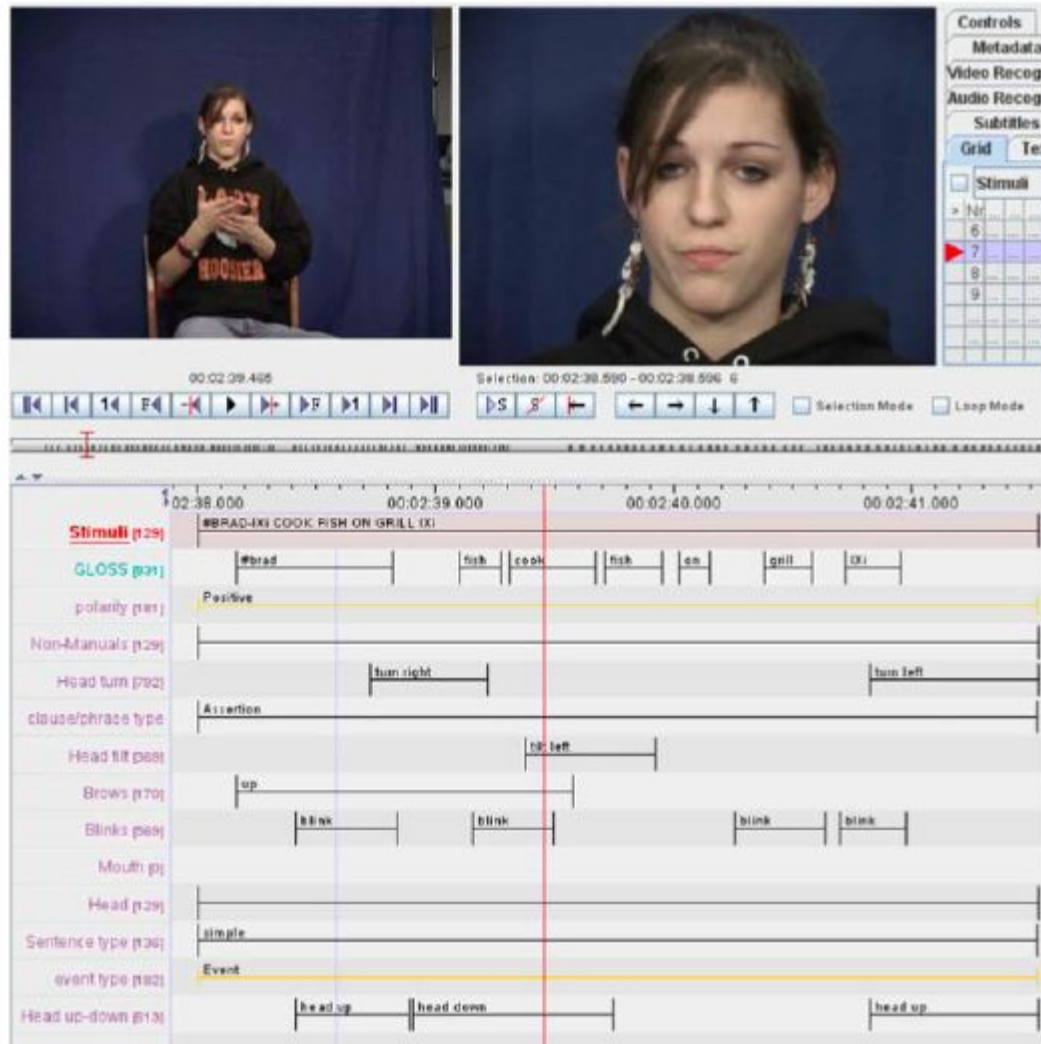
Plutchik and Russell representations

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)

Практическое задание: разметка

<https://bit.ly/2HojQ4j>

Что учесть при разметке: лаг по времени