

# Audio-Visual Deep Learning Regression of Apparent Personality

Alejandro Alfonso Hernández

**Directors:** Dr. Sergio Escalera Guerrero  
Cristina Palmero Cantariño  
Dr. Julio Jacques Junior

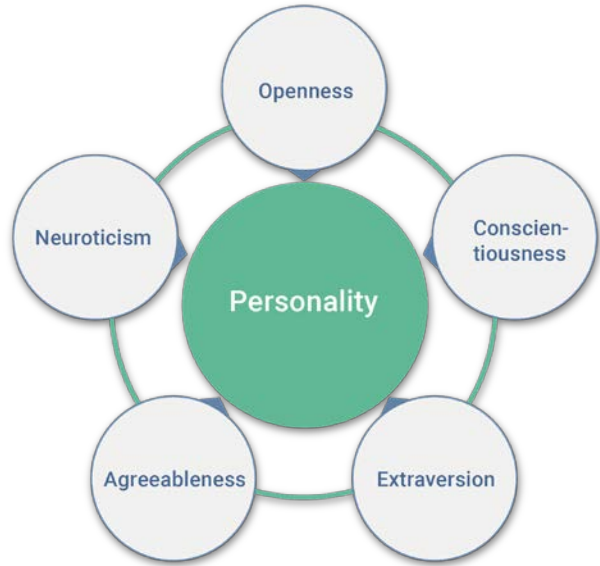
# Personality perception

- According to *Vernon et al.*, people make first impressions about others **from a glimpse** as brief as 100ms or less, and brain activity appears to track social traits even when no explicit evaluation is required.
  - *R. J. Vernon, C. A. Sutherland, A. W. Young, and T. Hartley, “Modeling first impressions from highly variable facial images,” Proc. of the National Academy of Sciences, 2014.*

*“You never get a second chance to make a First Impression”*



# What is the Big-Five model?



# What is Personality Computing?

## Automatic Personality Perception



**Motivation:** Improve the understanding of the variables that can influence the decision making of intelligent systems (specifically deep neural networks) to regress apparent personality, similar to how the human being would.

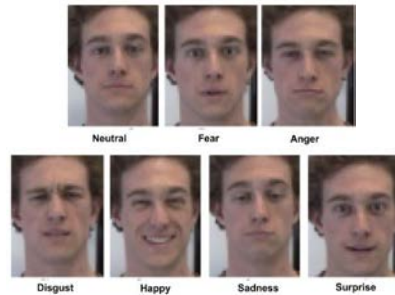
# The goal of our work

- Perform an analysis on the influence of automatic prediction for apparent personality based on the Big-Five model and study its improvement in accuracy related to the handcrafted features, comparing the results by gender, age-ranges and emotions.

Facial features



Emotions



Raw audio



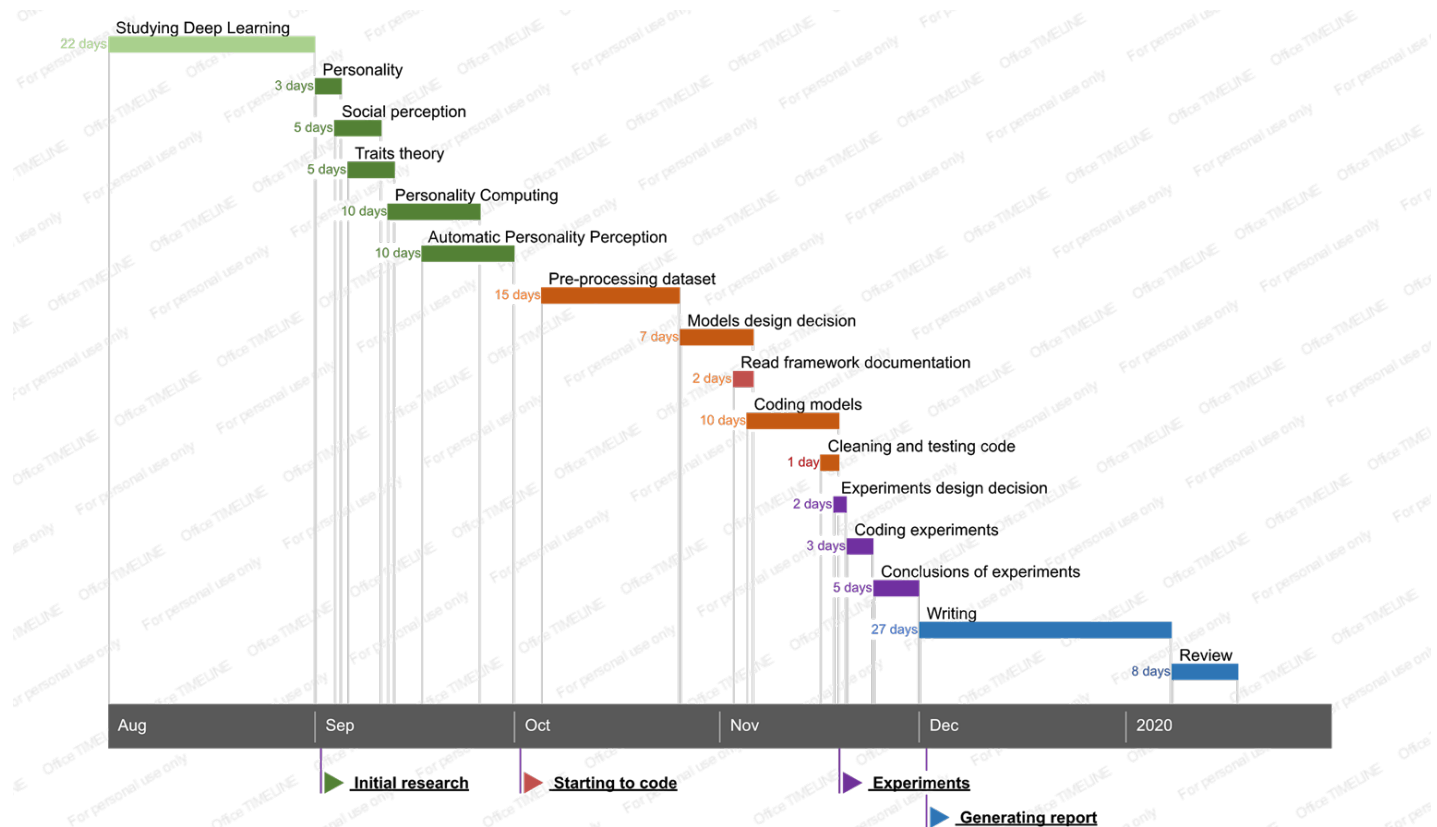
Gender



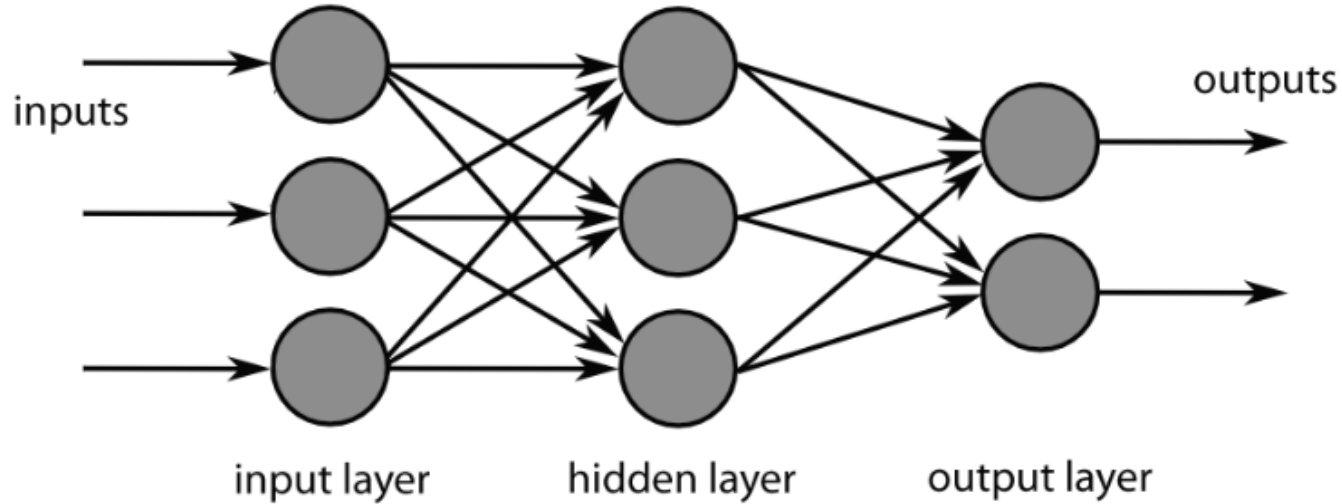
Age



# Planification

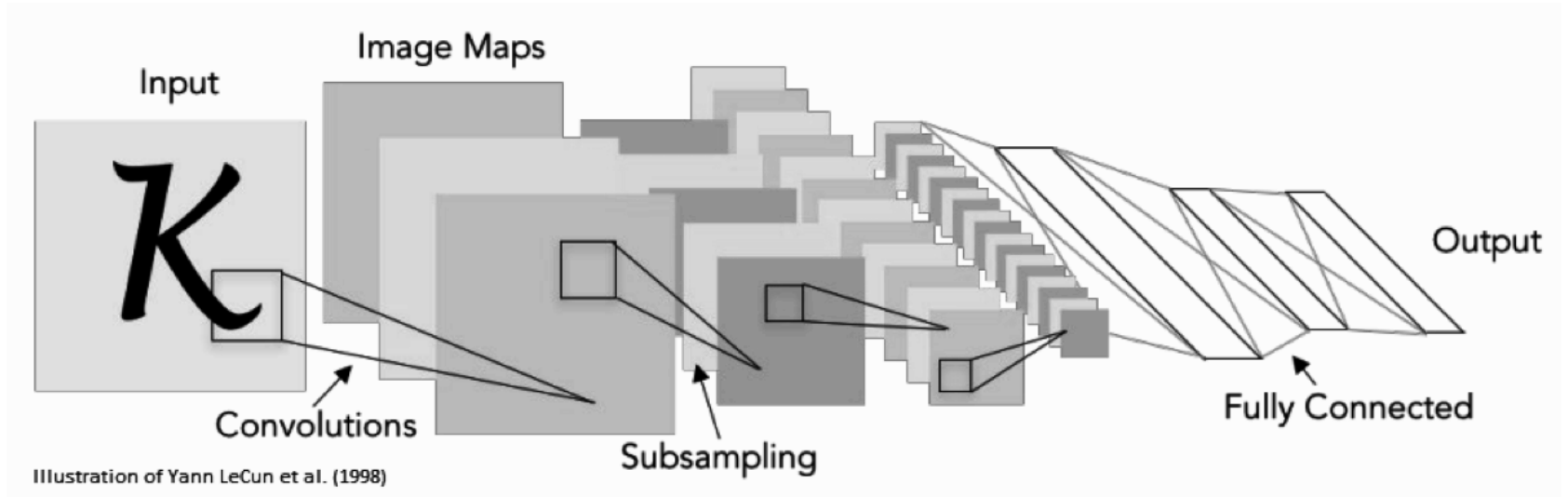


# Multi-Layer Perceptron



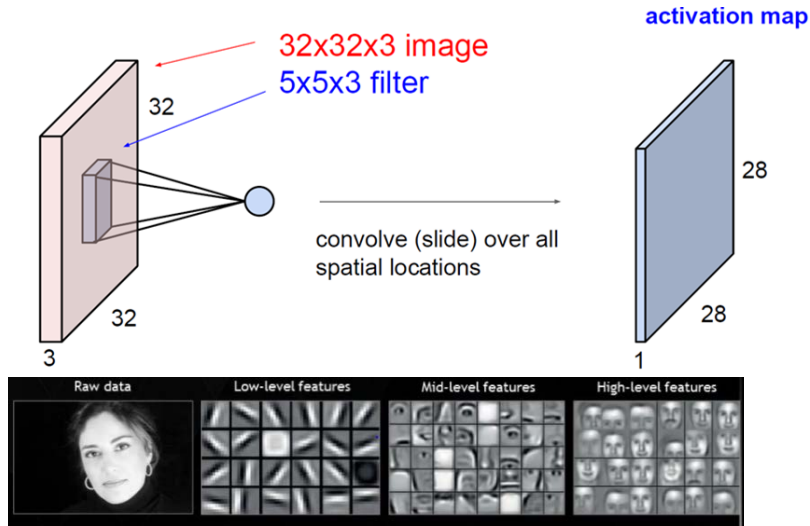
$$f(\sum_i w_i \cdot x_i + b)$$

# Convolutional Neural Networks (CNNs)

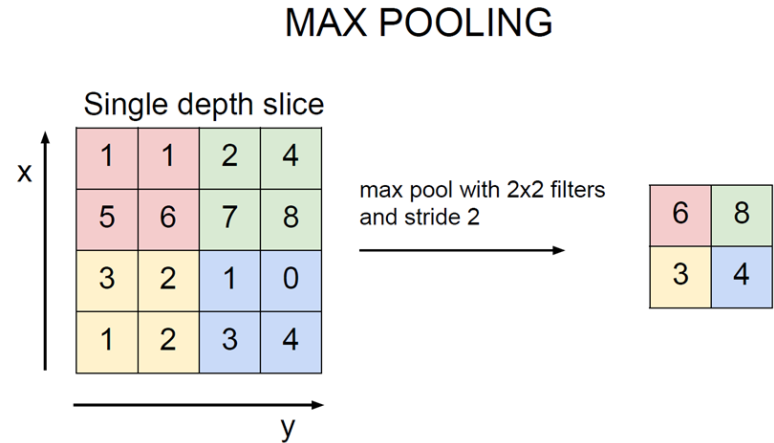


# How CNNs work

- Convolutional layers: convolve the input, the result is a feature map that is passed to the next layer.



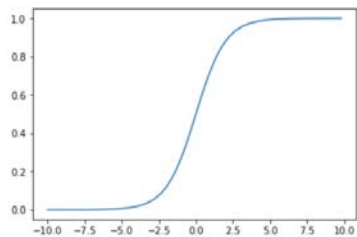
- Pooling layers: provide an approach to down sampling feature maps by summarizing the presence of features in patches of the feature map.





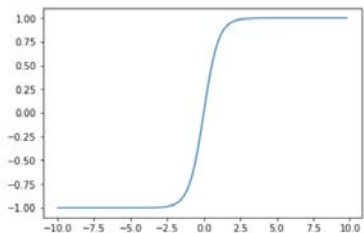
# Activation functions

- Sigmoid



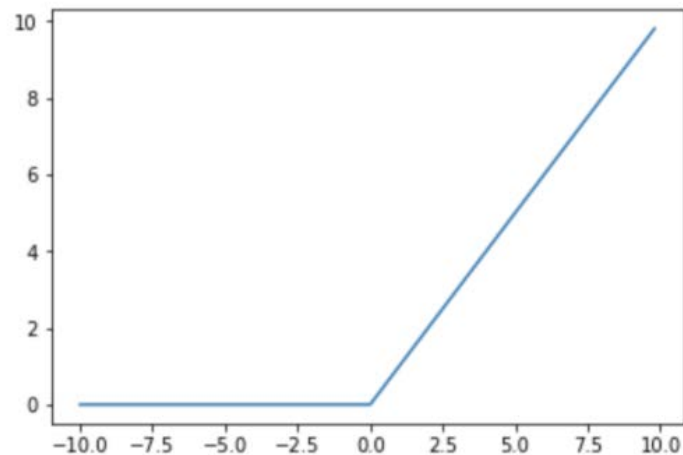
$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$$

- Tangent Hyperbolic (Tanh)



$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

- Rectified Linear Unit (ReLU)



$$f(x) = \max(0, x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

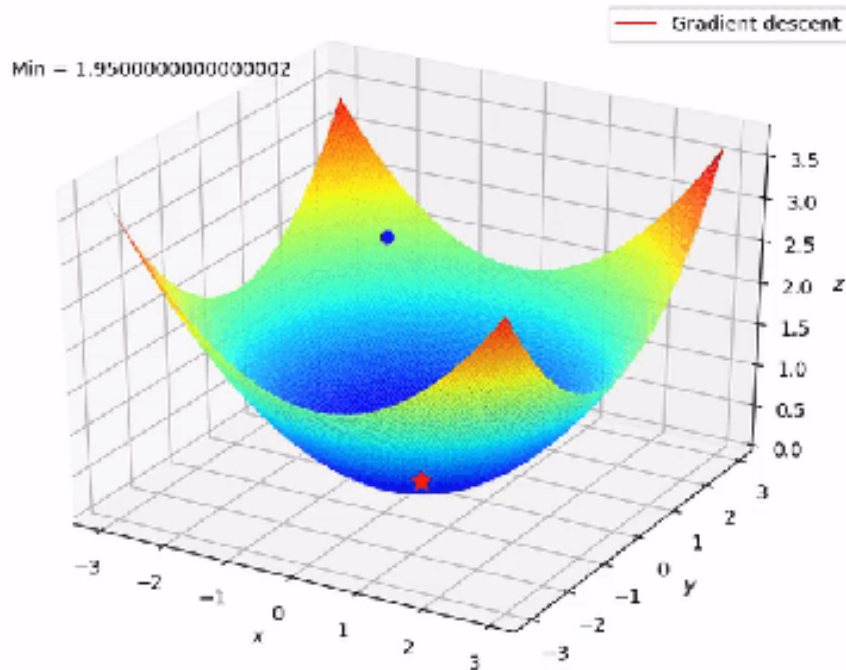
# Optimization algorithms

Optimization algorithms are used to optimize a cost function  $J$  in order to train the neural network.

$$J(W, b) = \frac{1}{m} \sum_{i=1}^m L(y'^i, y^i)$$

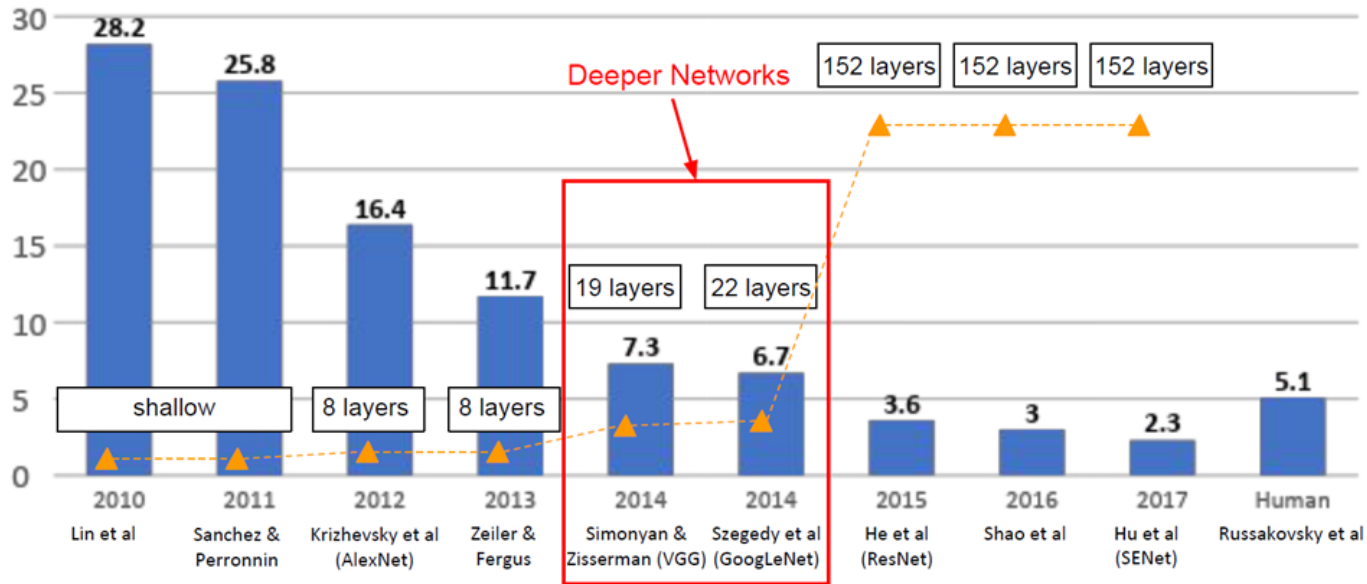
- Stochastic gradient descent (SGD)
- SGD+Momentum
- Adam

## Backpropagation



# Most remarkable CNN architectures

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



# VGG-16



# First Impressions dataset








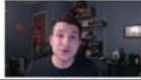












Please assign the following attributes to one of the videos:

- |                                 |             |                   |              |
|---------------------------------|-------------|-------------------|--------------|
| Friendly (vs. reserved)         | <b>Left</b> | <b>Don't know</b> | <b>Right</b> |
| Authentic (vs. self-interested) | <b>Left</b> | <b>Don't know</b> | <b>Right</b> |
| Organized (vs. sloppy)          | <b>Left</b> | <b>Don't know</b> | <b>Right</b> |
| Comfortable (vs. uneasy)        | <b>Left</b> | <b>Don't know</b> | <b>Right</b> |
| Imaginative (vs. practical)     | <b>Left</b> | <b>Don't know</b> | <b>Right</b> |

Who would you rather invite for a job interview?

**Left**      **Don't know**      **Right**

Agreeableness			
Authentic		Self-interested	
			
0.9230	0.9340	0.1098	0.0879
Conscientiousness			
Organized		Sloppy	
			
0.9706	0.9514	0.0873	0.1068
Extraversion			
Friendly		Reserved	
			
0.9156	0.9252	0.0521	0.0933
Neuroticism			
Comfortable		Uneasy	
			
0.9585	0.9791	0.1005	0.0872
Openness			
Imaginative		Practical	
			
0.9777	0.9682	0.0549	0.1113

# Data pre-processing: extracting face images

- Histogram of Oriented Gradients (HOG) feature descriptor combined with linear classifier, an image pyramid and sliding windows detection scheme.
- The detected faces are trimmed and aligned.
- The final face images have a resolution of 224x224x3 pixels.



# Data pre-processing: raw audio

- We added raw audio as input to our model inspired by the paper of Yağmur Güçlütürk et al. from 2016.
- We only used the first 5 sec of audio per video inspired by the paper of Ricardo Darío Principi et al. from 2019.



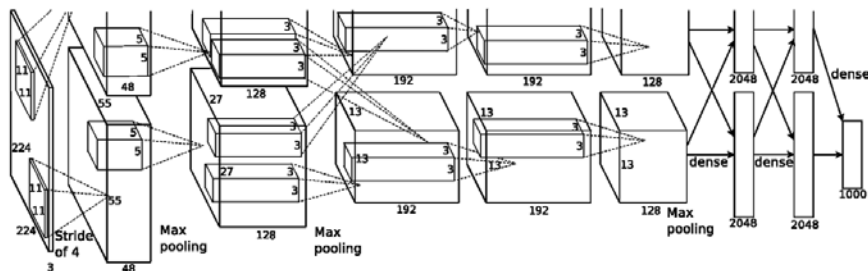
## **Data pre-processing: age and gender**

- Age and gender annotations were used labeled as follows:
  - Age as a positive integer
  - Gender as Male=1, Female=2.



# Data pre-processing: emotions

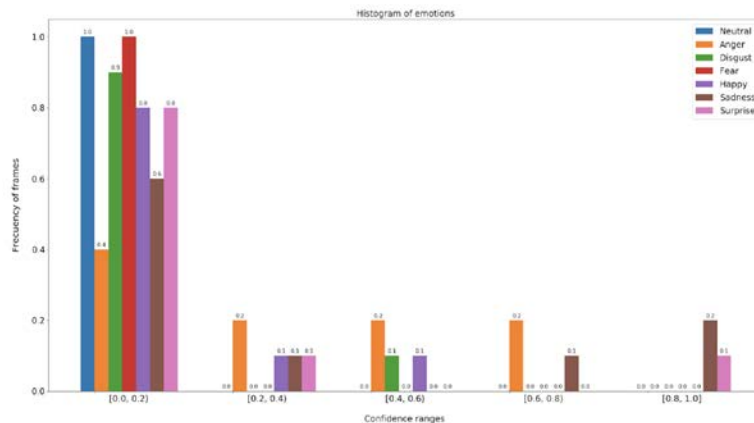
- Pretrained AlexNet architecture for regressing Ekman's basic emotions.



- Ekman's basic emotions.

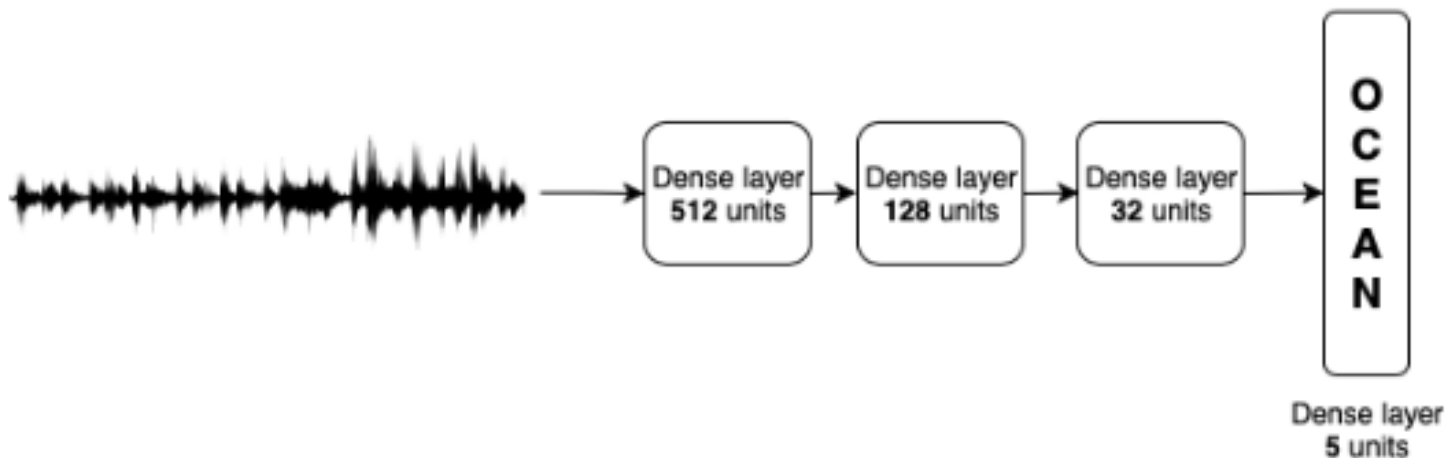


- Histogram of Ekman's universal emotions per confidence range



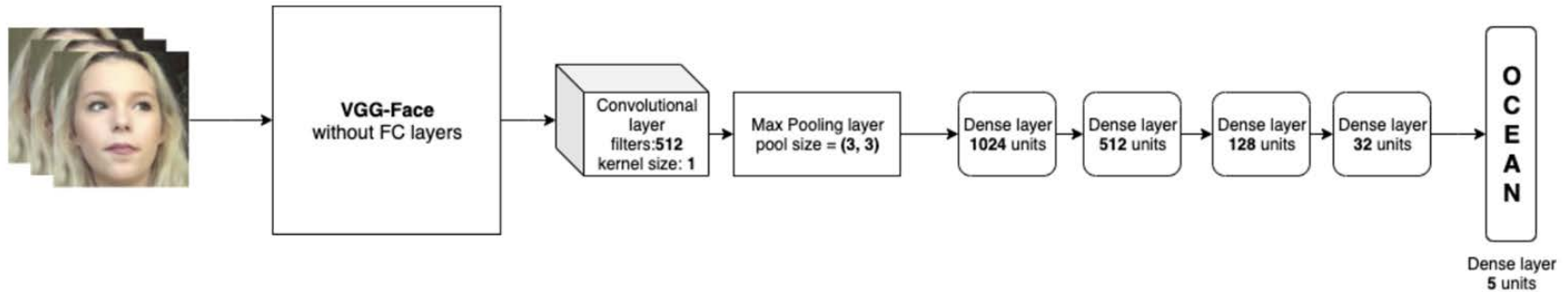
# Proposed models: Audio modality

- **Audio modality:** small deep neural network with raw audio waveform as input and 3 dense layers plus output.



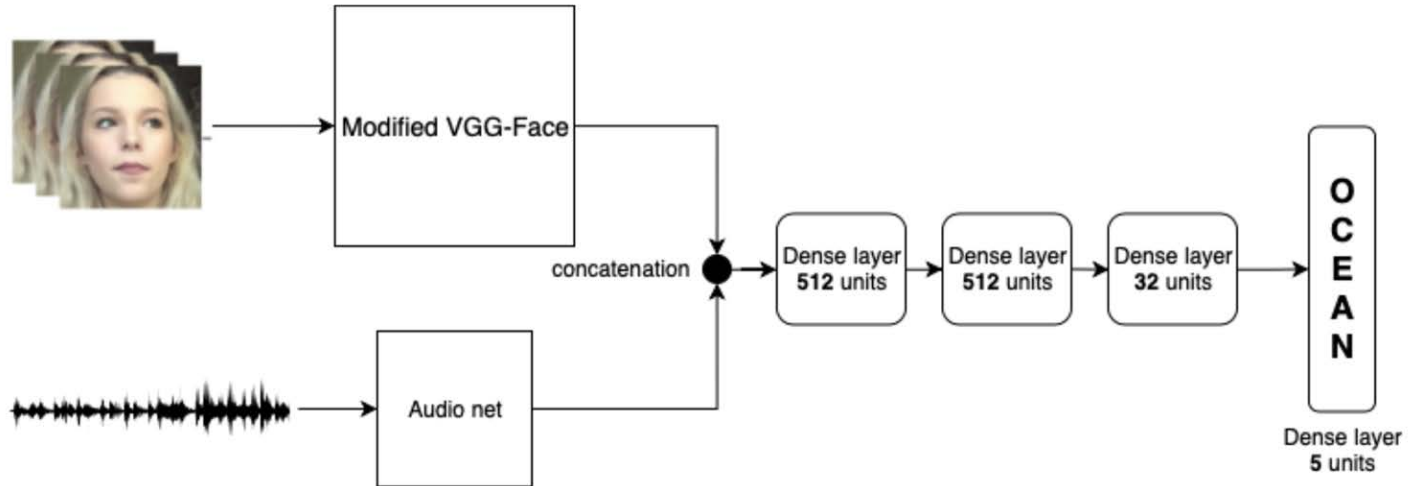
# Proposed models: Visual modality

- **Visual modality:** modified pretrained VGG-Face architecture, with face images as input.



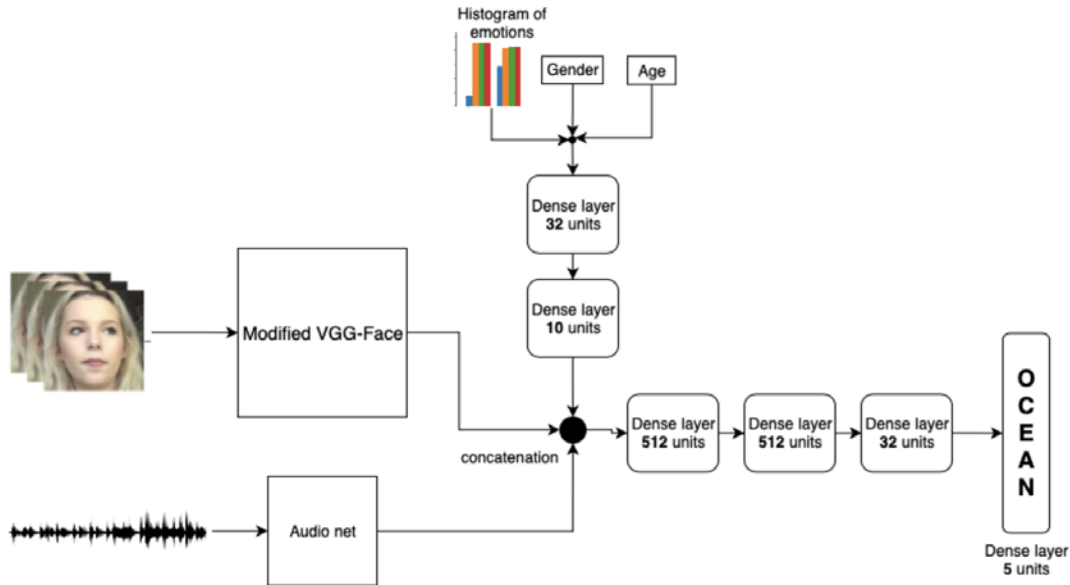
# Proposed models: Audio+Visual modality

- **Audio+Visual modality:** fusion of both previously mentioned models already trained and 3 dense layers plus output layer.



# Proposed models: Audio+Visual+handcrafted features modality

- **Audio+Visual+handcrafted features modality:** fusion of trained network audio and modified VGG-Face, with a late fusion strategy of the handcrafted features, plus 3 dense layers and the output layer.



# Implementation details: libraries and hardware

- **Programming language**
  - Python 2.7.3.
- **Library for deep neural networks development**
  - Keras 2.1.6. (Tensorflow as backend).
- **Hardware**
  - MacBook Pro (13-inch display, 2018).
  - 4-GPUs NVIDIA GeForce GTX TITAN X with 12GB of memory, property of the University of Barcelona.

## Implementation details: accuracy metric

- We compute the accuracy scores with the following formula:

$$acc_j = 1 - \frac{\sum_{i=1}^N |p_{ij} - gt_{ij}|}{N}$$

## Experiments results: global accuracy scores per traits and modality

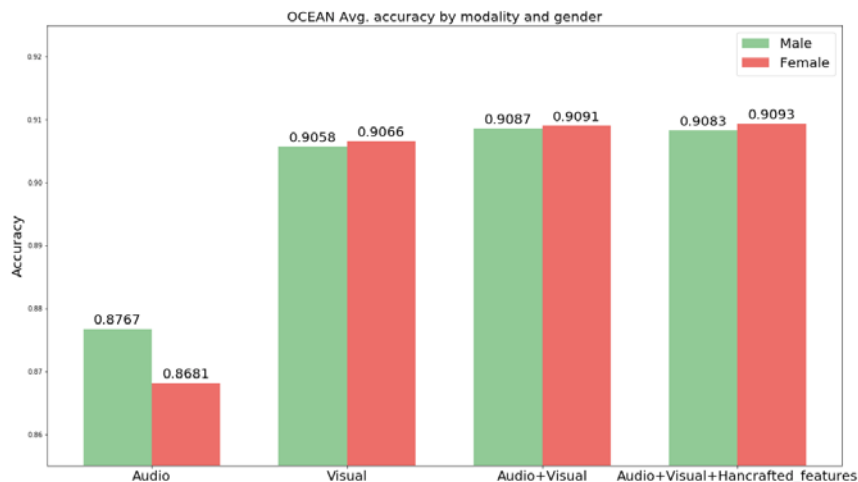
- Global accuracy scores per traits and modalities.

Modalities	O	C	E	A	N	Avg.
Audio	0.87397	0.87333	0.87357	0.87310	0.87362	0.87352
Visual	0.90525	0.91012	0.90783	0.90355	0.90437	0.90623
Audio+Visual	<b>0.90637</b>	<b>0.91447</b>	0.91171	0.90726	0.90465	0.90889
Audio+Visual+Handcrafted features	0.90623	0.91433	<b>0.91174</b>	<b>0.90753</b>	<b>0.90480</b>	<b>0.90893</b>



# Experiments results: comparison per traits and gender

- Average accuracy of modalities per gender.

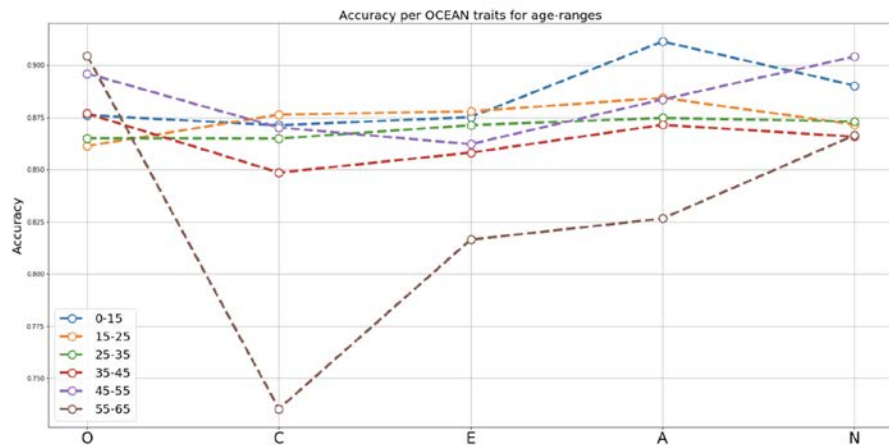


- Accuracy scores per trait, gender and modality.

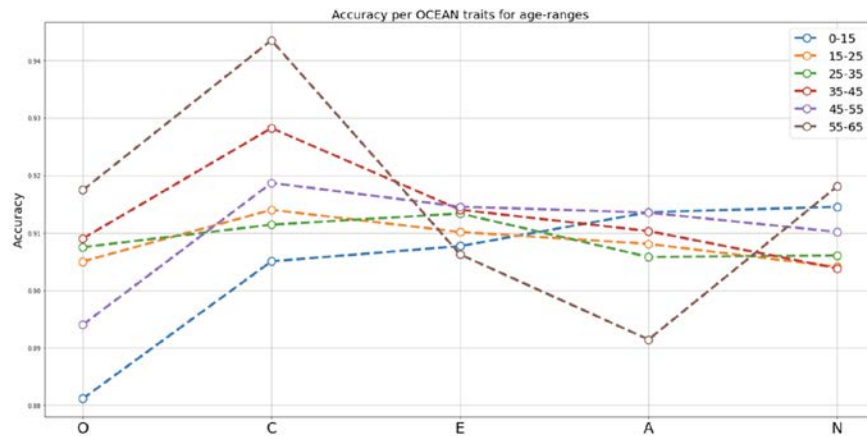
Modalities	O	C	E	A	N	Avg.
Audio	M 0.88042	M 0.86908	M 0.87535	M 0.88030	M 0.87850	M 0.87673
	F 0.85193	F 0.86958	F 0.87213	F 0.87939	F 0.86750	F 0.86811
Visual	M 0.90456	M 0.90900	M 0.90819	M 0.90301	M 0.90401	M 0.90576
	F 0.90582	F 0.91105	F 0.90753	F 0.90402	F <b>0.90467</b>	F 0.90662
Audio+Visual	M <b>0.90650</b>	M <b>0.91362</b>	M <b>0.91097</b>	M <b>0.90700</b>	M 0.90514	M <b>0.90865</b>
	F 0.90627	F 0.91517	F 0.91232	F 0.90747	F 0.90425	F 0.90910
Audio+Visual+handcrafted features	M 0.90563	M 0.91259	M 0.91070	M 0.90686	M <b>0.90563</b>	M 0.90828
	F <b>0.90628</b>	F <b>0.91546</b>	F <b>0.91248</b>	F <b>0.90787</b>	F 0.90461	F <b>0.90934</b>

# Experiments results: comparison per traits and age-ranges

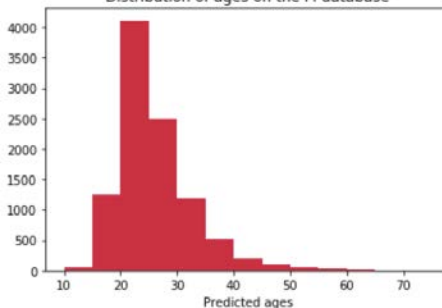
- Audio modality



- Audio+Visual+handcrafted features modality



Distribution of ages on the FI database



# Experiments results: comparison per traits and emotions

Emotion	Works better for	Works worst for
Neutral	<b>Agreeableness</b> (0.9166 with Audio+Visual+handcrafted features modality) <b>Extraversion</b> (0.9145 with Audio+Visual modality)	<b>Openness</b> (0.8183 with Audio modality) <hr/> <i>(If Audio modality is NOT considered)</i> <b>Conscientiousness</b> (0.893 with Visual modality)
Anger	<b>Conscientiousness</b> (0.9139 with Audio+Visual+handcrafted features modality) <b>Extraversion</b> (0.9123 with Audio+Visual+handcrafted features modality)	<b>Openness</b> (0.8707 with Audio modality) <hr/> <i>(If Audio modality is NOT considered)</i> <b>Agreeableness</b> (0.9038 with Visual modality)
Disgust	<b>Conscientiousness</b> (0.9156 with Audio+Visual+handcrafted features modality) <b>Extraversion</b> (0.9142 with Audio+Visual+handcrafted features modality)	<b>Extraversion</b> (0.8614 with Audio modality) <hr/> <i>(If Audio modality is NOT considered)</i> <b>Neuroticism</b> (0.9050 with Visual modality)

Fear	<b>Extraversion</b> (0.9140 with Audio+Visual modality) <b>Neuroticism</b> (0.8999 with Audio+Visual+handcrafted features modality)	<b>Agreeableness</b> (0.8656 with Audio modality) <hr/> <i>(If Audio modality is NOT considered)</i> <b>Agreeableness</b> (0.8851 with Visual modality)
Happy	<b>Conscientiousness</b> (0.9193 with Audio+Visual+handcrafted features modality) <b>Extraversion</b> (0.9068 with Audio+Visual modality)	<b>Openness</b> (0.8548 with Audio modality) <hr/> <i>(If Audio modality is NOT considered)</i> <b>Extraversion</b> (0.8985 with Visual modality)
Sadness	<b>Conscientiousness</b> (0.9129 with Audio+Visual modality) <b>Extraversion</b> (0.9122 with Audio+Visual+handcrafted features modality)	<b>Conscientiousness</b> (0.8642 with Audio modality) <hr/> <i>(If Audio modality is NOT considered)</i> <b>Agreeableness</b> (0.9021 with Visual modality)
Surprise	<b>Conscientiousness</b> (0.9162 with Audio+Visual modality) <b>Extraversion</b> (0.9114 with Audio+Visual+handcrafted features modality)	<b>Openness</b> (0.8554 with Audio modality) <hr/> <i>(If Audio modality is NOT considered)</i> <b>Agreeableness</b> (0.9045 with Visual modality)

# Conclusions

- We evaluated how personality perception from audio-visual cues can be automatically regressed by deep learning strategies.
- Results combining audio and video show better results than isolated modalities, suggesting that both modalities influenced the observers when labeling.
- By including multiple human attributes and features (age, gender, facial expressions of emotion) that can be sources of bias we guided/regularized the network to better learn to regress perception labels.

# Conclusions

- Our results show that:
  - Regression of female values is more accurate than for males,
  - Facial expressions of emotion are highly correlated to personality perception traits,
  - Extraversion and Conscientiousness traits show the higher regression accuracy (thus observers should have higher agreement when labeling those traits).
- To execute this project, it required from a deep knowledge of deep learning. I successfully achieved the knowledge to design customized deep neural network architectures that combine different data modalities, and learnt about associated gradient-based network training strategies.

# Future work

- Future work may include the analysis of other complementary sources of information such as background and clothing understanding, upper-body gestures, heart rate, audio transcription, other camera angles, as well as other attributes such as attractiveness, ethnicity and nationality.
- In addition, we could also regress real personality on the same data and try to establish a link between apparent and real personality.
- We could go even further and extend the study to the analysis of the relationship between two or more people in the same scene.

**Thank you very much for your  
attention!**