# Explainable Machine Learning

*Application Perspectives*

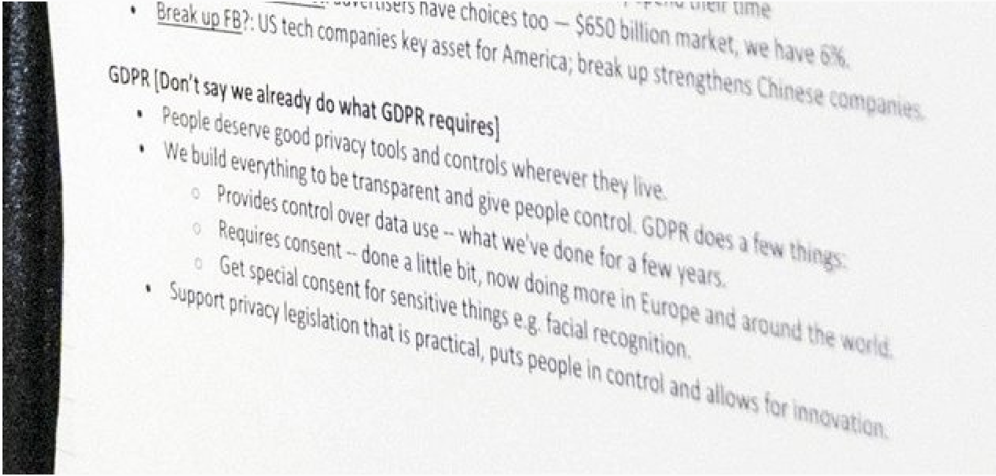
Conrad Sachweh

Heidelberg, April 19, 2018

Heidelberg University

# General Data Protection Regulation

# Motivation



- advertisers have choices too — $650 billion market, we have 6%.
- Break up FB?: US tech companies key asset for America; break up strengthens Chinese companies.

**GDPR [Don't say we already do what GDPR requires]**

- People deserve good privacy tools and controls wherever they live.
- We build everything to be transparent and give people control. GDPR does a few things:
  - Provides control over data use -- what we've done for a few years.
  - Requires consent -- done a little bit, now doing more in Europe and around the world.
  - Get special consent for sensitive things e.g. facial recognition.
- Support privacy legislation that is practical, puts people in control and allows for innovation.



http://www.businessinsider.de/mark-zuckerberg-hearing-notes-details-show-he-was-coached-2018-4

# Background

- Enforcement date: 25 May 2018
- **Regulation** instead of **Directive**
  ⇒ Similar to national laws
- Fines up to €20 million or 4 % of global revenue
- Regardless of company location


4.5.2016 EN Official Journal of the European Union L 119/1

I

*(Legislative acts)*

REGULATIONS

**REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL**

**of 27 April 2016**

**on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)**

**(Text with EEA relevance)**

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Article 16 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee ([1]),

Having regard to the opinion of the Committee of the Regions ([2]),

Acting in accordance with the ordinary legislative procedure ([3]),

Whereas:

(1) The protection of natural persons in relation to the processing of personal data is a fundamental right. Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) provide that everyone has the right to the protection of personal data concerning him or her.

(2) The principles of, and rules on the protection of natural persons with regard to the processing of their personal data should, whatever their nationality or residence, respect their fundamental rights and freedoms, in particular their right to the protection of personal data. This Regulation is intended to contribute to the accomplishment of an area of freedom, security and justice and of an economic union, to economic and social progress, to the strengthening and the convergence of the economies within the internal market, and to the well-being of natural persons.

(3) Directive 95/46/EC of the European Parliament and of the Council ([4]) seeks to harmonise the protection of fundamental rights and freedoms of natural persons in respect of processing activities and to ensure the free flow of personal data between Member States.

([1]) OJ C 229, 31.7.2012, p. 90.
([2]) OJ C 391, 18.12.2012, p. 127.
([3]) Position of the European Parliament of 12 March 2014 (not yet published in the Official Journal) and position of the Council at first reading of 8 April 2016 (not yet published in the Official Journal). Position of the European Parliament of 14 April 2016.
([4]) Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (OJ L 281, 23.11.1995, p. 31).



EU REGULATION 2016/679

## Basis for processing

- Consent must be explicit for purpose
  - e.g: calls recorded for training
- Records of processing activities
  - purpose
  - operator



https://twitter.com/northropgrumman/status/651019923846729728

# Goals

**Responsibility and accountability**

- Explanation of algorithmic decision
  - recommendation systems
  - credit/insurance
- Human intervention (at least safeguards)
- Respect data subjects rights/freedom
- Non-discriminating



https://twitter.com/northropgrumman/status/651019923846729728

# Goals

- Right of access
  - e.g. Facebook export tool
- Right to erasure
- Data breaches
  - 72 h disclosure time
- Pseudoanonymisation
  - encryption keys stored on other system

https://twitter.com/northropgrumman/status/651019923846729728

- Data Protection Officer for every organization
- Responsibility and accountability
- Lawful basis for processing
- Pseudoanonymisation
- Handling of data breaches
- Right of access
- Right to erasure

$\Rightarrow$ Data protection and privacy by design

$\Rightarrow$ Overhaul of standard algorithmic techniques



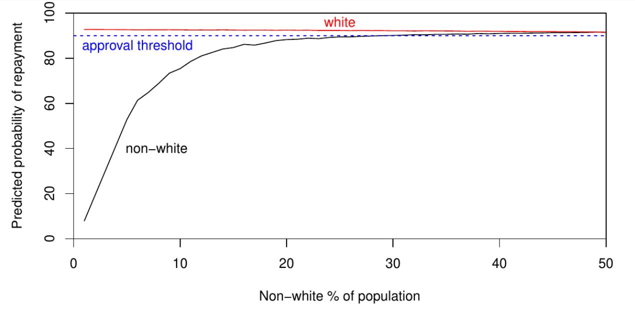

https://martechtoday.com/wp-content/uploads/2017/12/GDPR-with-stars.png

## Problems

- Fines only enforceable with international treaties
- Blockchain vs. right to erasure
- No official ”checklist”
- Big Data is not neutral
    - side-channels even if features removed
    - biases from training set

# Example: Unintended Discrimination

- Favor groups
- Data size 500
- Default 95% probability
- Representation of non-white
- Less uncertainty




Discriminating underrepresented groups in training set with a risk avers logistic regression classifier.



Image credit: European Union regulations on algorithmic decision-making and a ”right to explanation”; Goodman and Flaxman

# Examples

- Postal code
  - revealing racial information
  - info on loan defaulting
- Consumer buying history $\rightarrow$ comparable to medical exam for life insurance

$\Rightarrow$ Meaningful solutions require understanding how result was inferred.

# Outlook

- Companies are planning since years
- Chance to rectify current algorithms
- Better than human counterparts after this?
- Have to start planning algorithms with GDRP in mind

$$\text{collect less data} \Longleftrightarrow \text{better predictions}$$

# Explainable Learning Challenge

# Challenge Overview

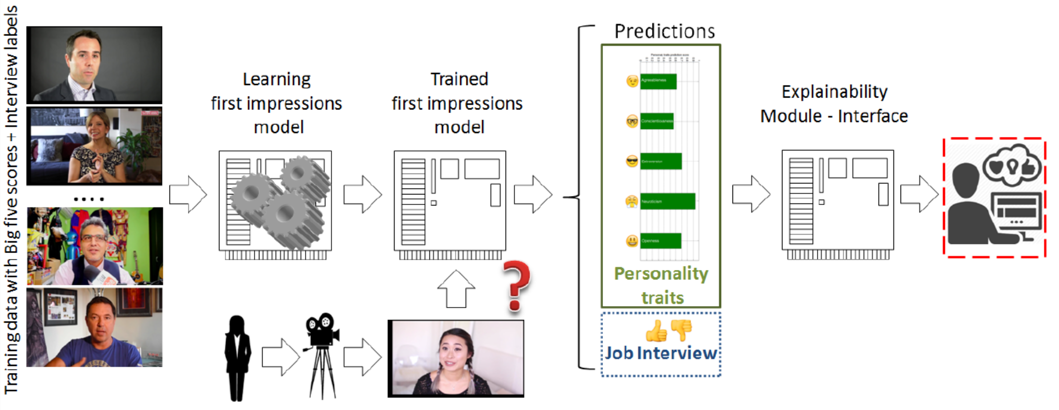





Image credit: Design of an Explainable Machine Learning Challenge for Video Interviews; Hugo et. al.

# Design of an Explainable Learning Challenge for Video Interviews – Objectives

Explainability/interpretability:

- Why is decision preferred over others
- How confident is the algorithm
- How were parameters selected
- Provide text description of reasoning

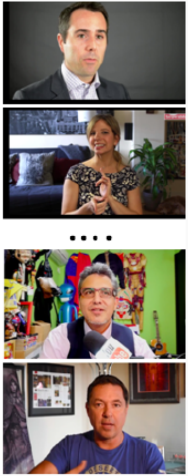


Image credit: Design of an Explainable Machine Learning Challenge for Video Interviews; Hugo et. al.

# Challenge

Video:

- Gestures
- Facial expressions

Spoken word:

- Intonation
- Pitch
- Transcript of video

$\Rightarrow$ "job-interview" prediction
social characteristics – Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness

# Applications

- Recruiters
  - $\Rightarrow$ need explanation of decision
- Negotiators
- Security gates
- Surveillance
- Military

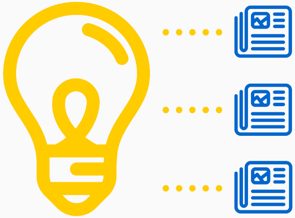




Image credit: https://www.betterbuys.com/cmms/definitive-guide-cmms/cmms-use-cases/

Would you accept algorithms denying your job applications?

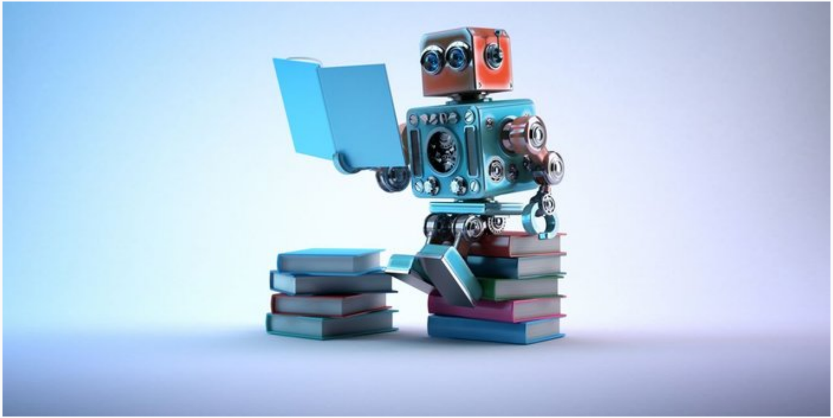






Image credit: Kirillm | Getty Images

# Pro & Con Compared to Human Decision

Advantages:

- objective assessment
- replicable solution

Disadvantages:

- algorithm must be explainable
- built to mimic human decisions

# Employed Data Set

- 10 000 of 15 s clips from 3000 Youtube Videos
- Labeled by humans (Amazon Mechanical Turk)
- Voice transcription (Human transcription service)

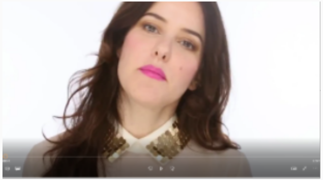




Image credit: Design of an Explainable Machine Learning Challenge for Video Interviews; Hugo et. al.

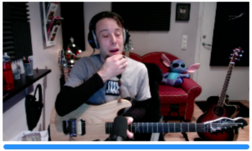
# Evaluation Process

1. Interview recommendation
2. Explanatory mechanism
   - Textual description why decision was made
   - Understand-ability
   - Rational and scientifically common?
3. Code sharing




Image credit: `http://www.evaluationforms.org/wp-content/uploads/Performance-Appraisal-Planning-Process.jpg`

# Proposed Attempts for Explanation of Employed Algorithms



This is to evaluate the quality of participants submission (below in yellow). Please answer all questions on the scale 0-5, **5 is best**.

- **Clarity**: Is the text **understandable** / written in proper English?
- **Explainability**: Does the text provide **relevant explanations** to the hiring decision made?
- **Soundness**: Are the explanations **rational** and, in particular, do they seem scientific and/or related to behavioral cues commonly used in psychology?

**Submission ID: *submission2***

The candidate is talking about himself. He looks quite friendly and comfortable. He could be organized but the background room is really not. Finally it is not sure that he is imaginative due to the monotony of its speech. I would not recommend him for an interview.


| | Target values | Predicted values |
|---|---|---|
| **extraversion** | 0.42 | 0.00 |
| **neuroticism** | 0.41 | 0.00 |
| **agreeableness** | 0.60 | 0.00 |
| **conscientiousness** | 0.39 | 0.00 |
| **interview** | 0.48 | 0.00 |
| **openness** | 0.56 | 0.00 |

**Clarity**
0 1 2 3 4 5 Don't know

**Explainability**
0 1 2 3 4 5 Don't know

**Soundness**
0 1 2 3 4 5 Don't know

Submit



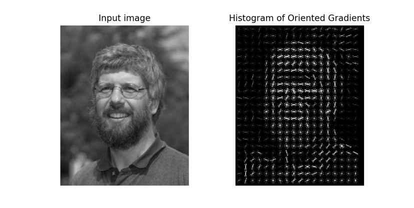
Image credit: Design of an Explainable Machine Learning Challenge for Video Interviews; Hugo et. al.

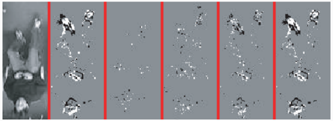
# Approaches

# Example Parts of Pipelines Used

Selection of techniques used:

- Face detection
- Frame differences
- Support vector regression
- Deep convolutional network



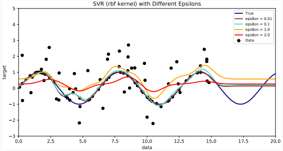
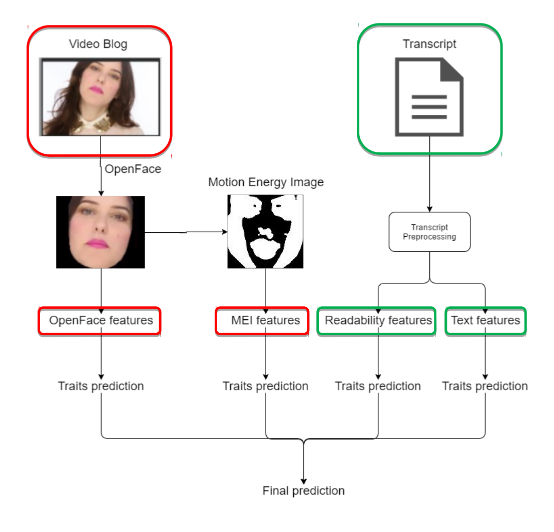





Image credits: `https://www.researchgate.net/publication/200016956_Address-Event_Video_Streaming_over_Wireless_Sensor_Networks`, WikiMedia Shiyu Ji: Svr epsilons demo

# Pipeline of the Winning Paper of the Second Round



# Presented Explanation of Winning Paper

```
****************************************************
* ASSESSMENT REPORT FOR VIDEO 2c42A4Z7qPE.001.mp4: *
****************************************************

On a scale from 0.0 to 1.0, I would rate this person's
interviewability as 0.497947.
Below, I will report on linguistic and visual assessment
of the person.
Percentiles are obtained by comparing the person against
scores of 6000 earlier assessed people.

____________________________________________________
```

# Presented Explanation of Winning Paper

```
*******************
* USE OF LANGUAGE *
*******************
Here is the report on the person's language use:

** FEATURES OBTAINED FROM SIMPLE TEXT ANALYSIS **
Cognitive capability may be important for the job.
I looked at a few very simple text statistics first.

*** Amount of spoken words ***
This feature typically ranges between 0.000000 and 90.000000.
The score for this video is 29.000000 (percentile: 25).
In our model, a higher score on this feature typically leads
to a higher overall assessment score.
```

# Presented Explanation of Winning Paper

```
*******************
* VISUAL FEATURES *
*******************
Here is the report on what I could "see":

*** Action Unit 12: how often was the lip corner pulled? ***
This feature typically ranges between 0.000000 and 1.000000.
The score for this video is 0.148148 (percentile: 82).

*** Action Unit 12: how much was the lip corner pulled on average?
This feature typically ranges between 0.000000 and 2.880709.
The score for this video is 0.333867 (percentile: 81).
```

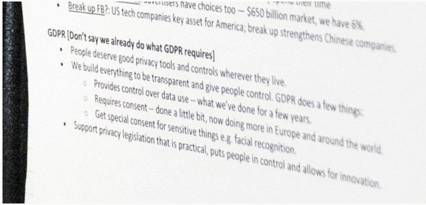
# Conclusion

# Explainability achieved?

- Understandable for an expert
- Unclear if really compliant with GDRP
- Neural networks not explained
- Most of the submitted entries did not use deep learning

# Thanks!

## Thank you for your attention!




References:

European Union regulations on algorithmic decision-making and a "right to explanation"; Goodman and Flaxman

Design of an explainable machine learning challenge for video interviews; Escalante et al.



http://www.businessinsider.de/mark-zuckerberg-hearing-notes-details-show-he-was-coached-2018-4