
# Psychology Meets Machine Learning

## Interdisciplinary Perspectives on Algorithmic Job Candidate Screening

Cynthia C. S. Liem * Markus Langer † Andrew Demetriou ‡
Annemarie M. F. Hiemstra § Achmadnoer Sukma Wicaksana ¶
Marise Ph. Born ‖ Cornelius J. König **

May 28, 2019

This is the author's manuscript of

> Liem C.C.S. et al. (2018) Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening. In: Escalante H. et al. (eds) Explainable and Interpretable Models in Computer Vision and Machine Learning. The Springer Series on Challenges in Machine Learning. Springer, Cham

For the officially published version of this chapter, please refer to `https://doi.org/10.1007/978-3-319-98131-4_9`, and use the according bibliographic metadata for citations.

The creation of these resources has been (partially) funded by the ERASMUS+ grant program of the European Union under grant no.2017-1-DE01-KA203-003569. Neither the European Commission nor the project's national funding agency DAAD are responsible for the content or liable for any losses or damage resulting of the use of these resources.

*Multimedia Computing Group, Delft University of Technology, Delft The Netherlands, c.c.s.liem@tudelft.nl
†Universität des Saarlandes, Saarbrücken, Germany, markus.langer@uni-saarland.de
‡Multimedia Computing Group, Delft University of Technology, Delft The Netherlands, andrew.m.demetriou@gmail.com
§Erasmus School of Social and Behavioral Sciences, Erasmus University, Rotterdam, The Netherlands hiemstra@essb.eur.nl
¶Datasintesa Teknologi Nusantara (formerly at Delft University of Technology), sukmawicaksana@gmail.com
‖Erasmus School of Social and Behavioral Sciences, Erasmus University, Rotterdam, The Netherlands, m.ph.born@essb.eur.nl
**Universität des Saarlandes, Saarbrücken, Germany, markus.langer@uni-saarland.de, ckoenig@mx.uni-saarland.de

**Abstract**

In a rapidly digitizing world, machine learning algorithms are increasingly employed in scenarios that directly impact humans. This also is seen in job candidate screening. Data-driven candidate assessment is gaining interest, due tohigh scalability and more systematic assessment mechanisms. However, it will only be truly accepted and trusted if explainability and transparency can be guaranteed. The current chapter emerged from ongoing discussions between psychologists and computer scientists with machine learning interests, and discusses the job candidate screening problem from an interdisciplinary viewpoint. After introducing the general problem, we present a tutorial on common important methodological focus points in psychological and machine learning research. Following this, we both contrast and combine psychological and machine learning approaches, and present a use case example of a data-driven job candidate assessment system, intended to be explainable towards non-technical hiring specialists. In connection to this, we also give an overview of more traditional job candidate assessment approaches, and discuss considerations for optimizing the acceptability of technology-supported hiring solutions by relevant stakeholders. Finally, we present several recommendations on how interdisciplinary collaboration on the topic may be fostered.

Keywords: psychology, machine learning, job candidate screening, methodology, explainability, multimodal analysis, interdisciplinarity

# 1 Introduction: algorithmic opportunities for job candidate screening

In a rapidly digitizing world, machine learning algorithms are increasingly employed to infer relevant patterns from data surrounding us as human beings. As a consequence, in many domains, information organization, process optimizations and predictions that formerly required human labor can now be systematically performed at higher efficiency and scalability.

The promise of computer-assisted decision-making has also entered the area of *personnel selection*: one of the oldest research areas within applied psychology. As early as in 1917, the problem of assessing whether candidates would be suitable for a job was recognized as:

> "*the Supreme Problem of diagnosing each individual, and steering him towards his fittest place, which is really the culminating problem of efficiency, because human capacities are after all the chief national resources.*" (Hall, 1917)

This *job candidate screening* problem has been of interest to researchers and practitioners ever since (Ployhart, Schmitt, & Tippins, 2017). 100 years later, richer, multimodal and digital means of candidate presentation have become available, such as video resumes. Such presentation forms may offer more nuanced insight into a candidate; in comparison to paper resumes, ethnic minority applicants perceived digital video resumes as a fairer way of presentation (Hiemstra, Derous, Serlie, & Born, 2012).

Digitization has not only influenced job candidate presentation forms, but also analysis techniques of candidate pools, through the inclusion of algorithmic methods in screening and selection procedures. This especially becomes necessary in case of large applicant pools, but is an actively debated practice. Proponents of automated digital selection methods argue that using algorithmic methods could lead to more diversity and empathetic workplaces, because they help to sidestep pitfalls typically associated with human decision-making. At the same time, caution is warranted because algorithms may be susceptible to bias in data and data labeling. Paradoxically, this bias may especially be harmful to applicants whose attributes are underrepresented in historical data (e.g., ethnic minorities).

## 1.1 The need for explainability

In technologically-assisted personnel selection, technological components replace parts of the selection procedure that formerly were conducted by humans. In alignment with emerging discussions on both fairness, accountability, transparency and ethics in machine learning and artificial intelligence, as well as human interpretability of sophisticated state-of-the-art machine learning models, research into explainability and transparency in algorithmic candidate screening is currently gaining interest (Escalante et al., 2017, 2018; Langer, König, & Fitili, 2018).

Considering technologically-assisted personnel selection, there are several reasons why explainability and transparency can be considered as particularly important:

- *Moral considerations.* Algorithmic decisions on personnel selection consider humans. It should be ensured that these decisions will not be unfair towards, or even harmful to certain population subgroups.
- *Knowledge-related considerations.* Hiring managers, the ultimate adopters of technologically-assisted selection tools, are not computer scientists. Therefore, they might not be able

to develop algorithm-based solutions on their own, nor understand the development process towards an algorithm-based solution.

Within machine learning, particularly through the advances of deep neural networks, very sophisticated and successful statistical models have emerged for performing predictions and classifications, but understanding and interpreting the internal workings of these networks is far from trivial.

- *Concerns about methodological soundness.* Increasingly, commercial ready-to-use solutions are being offered, and their inner workings may be a business secret. Still, regulatory frameworks such as the European General Data Protection Regulation (Council of the European Union, 2016) may grant the explicit right to end users to demand transparency on how their information is processed.

  Furthermore, in practice, a *research-practitioner gap* is frequently observed in personnel selection: several methodologically sound personnel selection procedures and good-practice recommendations that are developed through research never get adopted by hiring managers (N. Anderson, Herriot, & Hodgkinson, 2001). For instance, there are psychometrically sound measures of personality (e.g., Big Five measures (McCrae & Costa, 1999)). However, in practice, a large variety of unvalidated measures are used, that are more appealing to practitioners (Diekmann & König, 2015). Some reasons might simply be that the unvalidated measure is easier to use, or that it appears more efficient and allows more control for hiring managers (Klehe, 2004; König, Klehe, Berchtold, & Kleinmann, 2010). We will discuss main reasons for acceptance and adoption in more detail in Section 5.

For all these reasons, calls for explainability and transparency connect to the concept of *trust*: we want to ensure that a potential technological solution 'does the right thing', without causing harm. At the same time, where to focus on when aiming to 'do the right thing' or 'tackling the most challenging aspect' is differently understood by different people. This is a common issue for domains in which multiple disciplines and stakeholders come together, as for example also noticed in the domain of music information retrieval (Liem et al., 2012). Deeper insight into different disciplinary viewpoints on the problem and the relationships between them–from shared interests to fundamental methodological differences–will have great impact on understanding what would be needed for technological solutions to become truly acceptable to everyone.

## 1.2 Purpose and outline of the chapter

The current chapter emerged from discussions between computer scientists and psychologists in the context of an ongoing collaboration on identifying future-proof skill sets and training resources on Big Data in Psychological Assessment.

Our discussions were inspired by the emerging societal and scientific interest in technological solutions for the personnel selection problem, but also by ongoing concrete data challenges on inferring first-impression personality and interviewability assessments from online video (Escalante et al., 2017, 2018; Ponce-López et al., 2016). These challenges relate to an overall mission "*to help both recruiters and job candidates by using automatic recommendations based on multi-media CVs.*" (Escalante et al., 2017). As a consequence, computer vision and machine

learning researchers are challenged to not only quantitatively, but also qualitatively optimize their algorithmic prediction solutions.

In discussing potential data-driven solutions to these types of challenges, it became clear that the authors of this chapter indeed departed from different methodological focus points, interests, and optimization criteria. We therefore felt the need to more explicitly collect observations of how our various disciplinary viewpoints meet and differ. As a consequence, we contribute this chapter, which is meant as a tutorial which is accessible to computer scientists, psychologists and practitioners alike. Herein, we reflect on similarities and dissimilarities in disciplinary interests, potential common connection points, and practical considerations towards fostering acceptability of technologically-supported personnel selection solutions for various stakeholders, with special interest in questions of explainability. With the current discussion, we aim to move from *multidisciplinary* (Choi & Pak, 2006) debates about technologically-assisted selection mechanisms towards *inter-* and potentially *transdisciplinary* solutions, that can be implemented in responsible ways.

With regard to job candidate screening in personnel selection, we will focus primarily on the *early selection stage* of the process, in which we assume that there are suitable candidates in a large applicant pool but no selection decisions have yet been made. As a consequence, all candidates should be evaluated, and based on the evaluation outcomes a subset of them should be selected for the next selection stage, which may e.g. be an in-person interview.

The remainder of the chapter is outlined as follows:

- In Section 2, we will explain major methodological interests in psychology and computer science (considering machine learning in particular) in a way that should be accessible to practitioners in either discipline. We will also discuss their major similarities and differences.
- Subsequently, in Section 3, we move towards the domain of personnel selection, introducing the domain, its major research questions and challenges, and several important focus areas with key references.
- As a use case, Section 4 discusses a data-driven explainable solution that was developed in the context of the ChaLearn Job Candidate Screening Coopetition, with explicit consideration of potential connection points for psychologists and practitioners.
- Then, Section 5 focuses on research on acceptability of technology-supported personnel selection solutions, as perceived by two categories of user stakeholders in the personnel selection problem: job applicants and hiring managers.
- Finally, in Section 6, considering the various viewpoints provided in this chapter, we will give several recommendations towards interdisciplinary personnel selection solutions.

## 2 Common methodological focus areas

In this section, we will give broad and brief descriptions about how the psychological and computer sciences are conducted. These descriptions are intended to neither be exhaustive nor highly detailed. Rather, they are meant as an introduction to the uninitiated in each field, in vocabulary that should be understandable to all. Our aim is to inspire discussion on the intersections where the two may meet, and the separate paths where they do not. As such,

many of the points are presented with sparse references only where necessary; for readers seeking more thorough explanations and more domain-technical definitions, we will include references to several classical textbooks.

## 2.1 Psychology

### 2.1.1 Psychometrics

Psychology uses procedures, like questionnaires, interview protocols, and role-play exercises as tools to assess and quantify differences between individuals. Unlike direct forms of measurement such as height or weight, psychology investigates *constructs*, which are unseen aspects of individuals such as intelligence and personality. The assumption is that these constructs exist unseen in some quantity, and that individual differences in relation to these constructs are observable using reliable and valid procedures. By examining the relationship between measured constructs and observable behaviors, psychology seeks to increase our understanding of people.

While questionnaires are a commonly used, any systematic procedure used to gather and quantify psychological phenomena can be considered as a psychological instrument. Investigating how well a psychological instrument is measuring what it is supposed to measure is called *psychometrics*. Given that psychological phenomena are both complex and challenging to observe, and that the data collected must be interpreted, a study of the instruments themselves is crucial. Psychometrics can be thought of as the analytical procedures that examine the type of data collected, and estimate how well the variables collected using psychological instruments are reliable and valid. A useful textbook on the subject matter is the book by Furr and Bacharach (Furr & Bacharach, 2014).

### 2.1.2 Reliability

*Reliability* refers to the degree to which the variables produced by a procedure can be shown to be consistent, replicable, and free from measurement error. Similar to instruments in other fields, psychological questionnaires produce measurements that contain random 'noise'. Psychometric methods that assess reliability attempt to quantify the amount of 'signal' to 'noise', and how researchers might increase the amount of signal relative to the noise. By extension, reliability is a matter of degree; although two separate instruments may attempt to measure the same construct, one may have less measurement error than the other.

With regards to questionnaires, reliability is often concerned with *internal consistency*; specifically, how well the individual questions on the survey relate to each other, and to the overall survey scores. As we would expect multiple items on an instrument to measure the same construct, and as we would expect that construct to exist in individuals with some quantity, we would then expect responses to be consistent with each other. Measures of internal consistency, such as the alpha coefficient (Cronbach, 1951), examine the degree to which responses to the items on the test correlate with each other, and with the overall test score. Over the course of the development of an instrument, items that do not correlate well with the rest of the questions may be reworded, removed, or replaced with questions that produce more consistent responses. Thus, an instrument is developed and made sufficiently reliable for use.

Another common form of reliability regards test scores over time; *test-retest reliability* is the degree to which scores administered by one test will correlate with scores from the same test at a different time. Whether test-retest reliability is relevant is related to the construct being

examined. Because we would not expect mood to be perfectly stable–mood is regarded as a 'state' and not a 'trait'—expecting consistent responses on a questionnaire designed to assess mood over time is not sensible. However, because we expect personality to be stable, we would expect a participant's responses on one testing occasion to correlate with their responses on a second testing occasion, and therefore being replicable across occasions.

In situations where individuals are asked give subjective ratings, two forms of reliability are relevant: how reliable the ratings are among a group of raters (inter-rater reliability), and how reliable the multiple ratings are from the same rater (intra-rater reliability). With regards to judgments of relevant constructs, such as personality, *inter-rater* reliability refers to the replicability of ratings across multiple raters who judge a target person. In other words, to what degree do the ratings gathered from multiple people correlate? Conversely, *intra-rater* reliability refers to the degree to which a single person's ratings are consistent. With regards to personality, for example, will the rater judge the same person consistently over time?

The more reliable the instrument, the less random uncorrelated 'noise' compared to an interpretable 'signal' is present. Further, the more reliable the instrument, the more the observed magnitude of construct will approach the true magnitude of the construct. As such, the reliability of instruments is paramount.

### 2.1.3 Validity

However, whether or not a procedure is measuring the underlying construct it is attempting to measure goes beyond whether or not it is consistent. Reliability concerns the more mechanical elements of the instrument, namely the degree to which there is consistency vs. error in the measurements. However, determining how to interpret the measurements gathered by psychological instruments is a matter of *validity*. More specifically, validity refers to the degree to which interpretations of the variables are supported by prior research and theory. In this sense, the measurements produced by a procedure are neither valid nor invalid. Rather, validity is determined by the degree to which the variables produced by the instrument are interpretable as reflecting some psychological phenomenon. Discourse on how best to demonstrate validity has produced a number of validity 'types'. While a complete discussion on validity is beyond the scope of this chapter, a brief summary follows.

*Construct validity* refers to demonstrating and explaining the existence of unseen constructs, also known as '*signs*', beyond their reliable measurement. For example, personality questionnaires are common *instruments* for collecting quantifiable observable behavior about a person. The Big Five (McCrae & Costa, 1999) personality questionnaire is designed to allow researchers to assess personality along 5 dimensions. Specifically, it asks individuals to indicate how strongly they agree with a set of statements from 1 (strongly disagree) to 7 (strongly agree), thus producing a score for each *item*. If scores for the items vary between people, the variance can be quantified and examined, and underlying dimensions can be identified. By demonstrating the emergence of similar numbers of factors in procedures like the Big Five (or other personality questionnaires, such as the NEO-PIR, FFM, or HEXACO) in samples across cultures, and by demonstrating correlations to other meaningful variables, researchers have demonstrated construct validity for personality.

*Criterion validity* refers to the degree to which test scores (the predictor) correlate with specific criterion variables, such as job performance measures. It is often discussed in terms of two types: *concurrent validity*, which refers to the correlation of the predictor and criterion data that are collected at the same time, and *predictive validity*, which refers to the correlation of predictor

data collected during a selection procedure and criterion data collected at a later time.

Predictive validity is often considered the most important form of validity when during a selection procedure, rather than testing for specific and explicit signs that are considered relevant to future job performance measures, the test would rather consist of taking holistic *samples* of intended future behavior. This means of assessment is based on the theory of behavioral consistency, stating that past behavior is the best predictor of future behavior. In this sense, the predictor data may be collected during the selection process, and later correlated with data collected when selected applicants have become employees. For example, a prospective aircraft pilot may be asked to perform an assessment using a flight simulator. If the variables extracted during the flight simulator correlate with later assessments of performance when the candidate has become an employee, the test allows for predictions of future performance. Therefore, we might conclude that the simulator test has demonstrated predictive validity.

In sample-based approaches, decomposition of the observed behavior into constructs is not sought, and as such, construct validity is less relevant. On the other hand, it is relevant whether or not the test produces scores that correlate to certain key criteria, like future ratings of job performance for example.

*Content validity* refers to the degree to which each item, question, or task in a procedure is relevant to what should be tested, and the degree to which all aspects of what should be tested are included. For example, personality research has shown evidence for multiple psychological dimensions, sometimes called personality *facets*. In other words, when we refer to the various aspects of one's personality, such as whether they are extraverted, agreeable, conscientious etc., these are various psychological dimensions that collectively comprise the construct we call 'personality'. Individual psychological dimensions may or may not be shown to correlate with each other, but are shown to be distinct e.g. via the results of an Exploratory Factor Analysis. If we were to develop a new method for assessing personality, the full spectrum of the various personality dimensions must be included in the assessment in order for us to demonstrate content validity. In addition, each element of the procedure must be shown to measure what it is designed to measure. In the case of questionnaires, the actual words in the questions should reflect what it is that they are designed to assess.

*Face validity* is the degree to which the items or tasks look plausible to, and can be understood by participants, and not just to experts. For example, when the test items concern questions on submissive behavior and the test is called the Submissive Behavior Test, participants may be persuaded that it is measuring submissiveness. Another example regards whether participants understand specifically what the questions are asking. If the questions are poorly translated or contain words that are ambiguous or unknown to participants, such as technical jargon or terms that have very specific meanings in one domain but various meanings in other domains, this may affect participant responses. Should the instructions or wording of a questionnaire be confusing to the participants taking it we might also say it lacks face validity.

*Convergent validity* refers to the degree to which two different instruments, which aim to measure the same construct, produce measures that correlate. For example, we would expect scores from multiple questionnaires that measure Extraversion, a dimension of personality, to correlate. We would further expect a person's loved ones would rate their degree of Extraversion similarly, and that these ratings would correlate with each other and the individual's test scores. Furthermore, we would expect that measures of Extraversion would correlate with related constructs and observable behaviors. On the other hand, *divergent validity* refers to the expectation that the construct an instrument is measuring will not correlate with unrelated constructs. If a measure of Extraversion consistently highly correlates with another personality

dimension, such as Conscientiousness, the measures may not be clearly distinct. In other words, both forms of validity are concerned with the degree to which test scores exhibit relationships to other variables, as would be expected by existing theory.

### 2.1.4 Experimentation and the nomological network

Psychology aims to explain constructs that are not directly observable, by examining the relationships between them, along with their relationships to observable behaviors. This involves demonstrating whether a construct exists in the first place, whether and how it can be reliably measured, and whether and how it relates to other constructs. The complete collection of evidenced and theoretical relationships (or lack thereof) between constructs, along with the magnitudes of their relationships, is called the *nomological network*. The nomological network surrounding a specific construct encapsulates all its relationships to other constructs, some of which will be strong and others of which will be weak.

Psychology develops knowledge by testing hypotheses that expand this network, testing competing theories in the network, or clarifying the magnitudes of the relationships in this network. The researcher derives hypotheses from what one might expect the relationships between variables to be, based on existing research and theory. Procedures are designed to collect data with as little 'noise' as possible, by creating controlled and repeatable conditions, and using reliable and valid instruments. The relationships between the measures from the various instruments are then subjected to statistical tests, usually in the family of general linear modeling (i.e. regression, F-tests, t-tests, correlations etc.), although Bayesian and algorithmic techniques have recently started to appear. In this way, psychology seeks to develop our understanding of the relationship between independent and dependent variables, and by extension, the nomological network surrounding a specific topic.

Although the variables are often described as independent/predictor variables or dependent/outcome/criterion variables, tests are often conducted on concurrent data, where all data points are collected at approximately the same time. As such, the placement of a variable as the independent or dependent may be a matter of statistical modeling, and not whether it is actually making a prediction.

Reliability and validity play an important role in this process. Reliability concerns itself with random error in measurements, which are expected to be uncorrelated with any of the variables being measured. As such, the lower the reliability, the more error in the data, the more attenuated the relationship between constructs will appear to be. The magnitude of the observed effect, in turn, affects the results of statistical significance tests which are often used to determine whether results are interpretable. On the other hand, part of the validation process is demonstrating the effect size of relationships. Specifically, it is necessary to determine how strong relationships between variables are, beyond whether their relationship is statistically significant. Based on prior theory, we often can estimate at least whether a relationship between two constructs ought to be statistically significant, and whether it ought to be strong or weak. When data show the predicted pattern of correlations between constructs, instruments demonstrate validity.

In areas of the nomological network where relationships have yet to be studied, exploratory studies may first be conducted to set the foundation for developing theory. Such studies may include qualitative techniques such as interviews, or questionnaires that allow participants to type their responses freely. Exploratory studies may also include quantitative techniques, such as Exploratory Factor Analysis (EFA): EFA is often used in the development of questionnaires

with Likert-scale items, as it allows the researcher to examine whether or not multiple dimensions are present in the questionnaire, and by extension, the dimensionality of the construct it seeks to measure. By showing how individual items on a questionnaire correlate to one or more latent variables, the researcher can develop the theoretical structure of a construct. For example, personality researchers used such methods to develop theory on the various personality facets. Procedures like EFA may show that certain items on an instrument correlate with a hypothetical axis, much more so than with other hypothetical axes. Based on the wording and content of the questions that cluster together, these hypothetical constructs can be named (e.g., Extraversion vs. Conscientiousness). With an initial estimate of the structure of a construct, researchers can then use a more restricted analytical technique, such as Confirmatory Factor Analysis, to examine whether and how well the exploratory model fits newly collected data.

Psychology researchers are faced with certain limitations, however. The data collection process is often labor-intensive, time is necessary to stay current on theory and research in order to develop hypotheses, and samples are often drawn by convenience leading to a preponderance of student WEIRD samples (Western, Educated, Industrialized, Rich, Democratic) (Henrich, Heine, & Norenzayan, 2010). Nevertheless, by conducting exploratory and confirmatory studies, psychology researchers contribute knowledge about how individual constructs relate to each other and observable behaviors.

## 2.2 Computer science and machine learning

The domain of computer science studies the design and construction of both computers, as well as the automated processes that should be conducted by them. *Generalization* and *abstraction* are important values of the domain. As for generalization, a solution to a problem should not only work in a specific case, but for a broader spectrum of cases–ideally, in any possible case that can be thought of for the given problem. For this reason, it may be needed to not always describe and treat the problem in full contextual detail, but rather in a more abstracted form, that can be used for multiple variants of the problem at once. Here, mathematics and logic contribute the language and governing principles necessary to express and treat generalization and abstraction in formalized, principled ways. Furthermore, *efficiency* and *scalability* are of importance too: through the use of computers, processes should be conducted faster and at larger scale than if their equivalent would be conducted in the physical world only.

Computer processes are defined in the form of *algorithms*, which are sets of explicit instructions to be conducted. Algorithms can be formally and theoretically studied as a scientific domain in itself: in that case, the focus is on formally quantifying and proving their properties, such as lower and upper bounds to the time and memory space they will require to solve a given problem (*computational complexity*). In many other cases, algorithms will rather be used as a tool within a broader computational context.

Within computer science, a domain receiving increasing attention is that of *artificial intelligence* (AI). In popular present-day discourse, 'AI' is often used to indicate specific types of *machine learning*. However, artificial intelligence is actually a much broader domain. While no single domain definition exists, it can be roughly characterized as the field focusing on studying and building intelligent entities. The classical AI textbook by Russell and Norvig (Russell & Norvig, 2010) sketches four common understandings of AI, including 'thinking humanly', 'thinking rationally', 'acting humanly', and 'acting rationally'. Furthermore, a philosophical distinction can be made between 'weak AI' and 'strong AI': in the case of weak AI, machines act as if they are intelligent, and only simulate thinking; in the case of strong AI, machines

would be considered to actually think themselves. While popular discourse tends to focus on strong AI, in practice, many present-day AI advances focus on weak AI in limited, well-scoped domains. Within AI, many subdomains and focus areas exist, including studies of knowledge representation, reasoning and planning, dealing with uncertainty, learning processes, and applying AI in scenarios that require communication, perception, or action.

Machine learning can be considered as the AI subdomain that deals with *automatically detecting patterns from data*. The 'learning' in 'machine learning' denotes the capacity to automatically perform such pattern detections. In the context of the job candidate screening problem, machine learning is the type of AI that most commonly is applied, and therefore, the most relevant subdomain to further introduce in this section. First, we will focus on discussing the main focus points in fundamental machine learning, in particular, *supervised machine learning*. Then, we will focus on discussing how machine learning is typically used in applied domain settings. Following this, the next section will discuss how common methodological focus areas in psychology and machine learning are overlapping, contrasting, and complementing one another.

### 2.2.1 The abstract machine learning perspective

In machine learning, algorithms are employed to learn relevant patterns from data. Different categories of machine learning exist, most notably:

- *Unsupervised machine learning*, in which a dataset is available, but relevant patterns or groupings in the data are initially unknown. Statistical data analysis should be employed to reveal these.
- *Supervised machine learning*, in which in connection to data, known *targets* or labels are provided. The goal will then be to relate the data to these targets as accurately as possible.
- *Reinforcement learning* (Sutton & Barto, 1998), in which the focus is on learning to act towards a desired outcome: an agent should learn those actions in an environment (e.g., game playing actions), that will lead to an optimal reward (e.g., a high score).

In this chapter, we focus on supervised machine learning. With a focus on generalization and optimal exploitation of statistical patterns encountered in data, supervised machine learning algorithms are not pre-configured to specialize in any particular application domain. Therefore, more formally and more abstractly, it can be stated that the goal of a supervised machine learning algorithm is to learn some function $f(\vec{x})$ that relates certain input observations $\vec{x}$ to certain output targets $\vec{y}$, in a way that is maximally generalizable and effective. If $\vec{y}$ expresses categorical class memberships, a *classification* problem is considered. If $\vec{y}$ rather expresses one or more continuous dependent variables, a *regression* problem is considered.

For simplicity, the remainder of this discussion focuses on cases in which $f(\vec{x})$ has the form $f : \mathbb{R}^d \rightarrow \mathbb{R}^1$. In other words, input observations are represented by $\vec{x}$, a $d$-dimensional vector, of which the values are in the set of all real numbers $\mathbb{R}$—in other words, $\vec{x}$ contains $d$ real numbers. $\vec{x}$ should be mapped to a single real number value $y$, expressing the target output.

To learn the appropriate mapping, a *training* stage takes place first, based on a large corpus with various examples of potential inputs $\vec{x}_{train}$, together with their corresponding target outputs $y_{train}$. For this data, the human machine learning practitioner specifies the model that should be used for $f(\vec{x})$. Examples of models can e.g. be a linear model, a decision tree, a

support vector machine, a neural network, or a deep neural network (Bishop, 2006; I. Goodfellow, Bengio, & Courville, 2016). Initially, the parameters that the chosen model should have to optimally fit the data are unknown. For example, for a linear model, these would be the slope and intercept. During the training phase, considering statistical properties of $\vec{x}_{train}$ and $y_{train}$, a model-specific machine learning algorithm will therefore iteratively optimize the necessary model parameters, by minimizing an expert-defined error measure between estimated outputs $\hat{y}$ and true outputs $y$. For example, for a linear model, this may be the sum of squared errors between each $\hat{y}$ and $y$ in the training set.

To assess whether the learning procedure has been successful in a generalizable way, the final reported performance of the learned $f(\vec{x})$ will be computed by running $f(\vec{x})$ on a *test* set, which contains input data that was not used during the training phase. As the final learned $f(\vec{x})$ specifies the necessary mathematical transformation steps that should be performed on $\vec{x}$ in order to predict $y$, it can be used as an optimized *algorithm* for predicting $y$ from $\vec{x}$.

It should be re-emphasized that from a pure machine learning perspective, the only requirement on the nature of $\vec{x}$ and $y$ is that they can be specified in numerical form. The only 'meaning' that $\vec{x}$ and $y$ will have to the model learning procedure, is that they contain certain numeric values, which reflect certain statistical properties. With the focus on finding an optimal prediction function $f(\vec{x})$, the tacit assumption is that finding a mapping between $\vec{x}$ and $y$ makes sense. However, the procedure for learning an optimal $f(\vec{x})$ only employs statistical analysis, and no human-like sense-making. It will not 'know', nor 'care', whether $\vec{x}$ and/or $y$ consider synthetically generated data or real-world data, nor make any distinction between flower petal lengths, census data, survey responses, credit scores, or pathology predictions, beyond their values, dimensionality, and statistical properties. When considering real-world data, it thus is up to the human practitioner to propose correct and reasonable data for $\vec{x}$ and $y$.

While various machine learning models have various model-specific ways to deal with noise and variance, further tacit assumptions are that $\vec{x}$ realistically follows the distribution of future data that should be predicted for, and that $y$ is 'objectively correct', even if it may contain some natural noise. In applied settings, in case the target outputs $y$ consider labels that are obtained through an acquisition procedure (through empirical measurement, or by soliciting human annotations), $y$ also is frequently referred to as 'ground truth', which again implies that $y$ is truthful and trustable.

Being oblivious to human data interpretation, machine learning algorithms will not 'understand' any potential 'consequences' of correct or incorrect predictions by themselves. If such considerations should be taken into account, it is up to the human expert to encode them properly in the defined error measure. For example, in case of binary classification, in which $y$ can only have the values 'true' or 'false', *false negative* classification errors (making a 'false' assessment where a 'true' assessment was correct) and *false positive* classification errors (making a 'true' assessment where a 'false' assessment was correct) may need to be weighted differently. For example, if a binary classification procedure would consider assessing the occurrence of a certain disease in a patient, false negatives (i.e., incorrectly labeling a diseased patient as healthy) may be deemed much graver mistakes than false positives (i.e., incorrectly labeling a healthy patient as diseased), as false negative assessments will cause diseased patients to not be treated. If so, for the error measure employed during learning, the penalty on making a false negative classification should be defined to be much larger than the penalty on making a false positive classification.

### 2.2.2 Machine learning in applied domains

As discussed in the previous section, the focus in fundamental machine learning is on learning $f(\vec{x})$ in an optimal and mathematically well-founded way, considering the given statistical properties of $\vec{x}$ and $y$, as well as the specified error measure. While from a fundamental perspective, it does not matter whether $\vec{x}$ and $y$ are synthetically generated or real-life data, interpretation of $\vec{x}$ and $y$ does matter when machine learning techniques are considered in applied domains, such as computer vision and bioinformatics.

In such applied cases, typically, $y$ represents a dependent variable considering a natural sciences observation, that can objectively be verified in the physical world. For example, it may denote the value depicted by a hand-written number, the occurrence of a disease, the boundaries of a physical object, or the identity of a person. The input data $\vec{x}$ often is the 'raw', high-dimensional result of a noisy sensory measurement procedure: for example, it may express color intensity values of different pixels in an image, an audio waveform, or microarray gene expression data. A human being will not be capable of relating such noisy measurements to their target outputs reliably; in contrast, a machine learning procedure has the power to systematically find relevant properties, rules and correlations between $\vec{x}$ and $y$.

Historically, before initiating the learning procedure, a pre-processing step would be performed on $\vec{x}$. In such a step, raw data measurements would first be turned into semantically higher-level, humanly hand-crafted *features*. For example, the color intensity values of individual pixels in a full image may first be summarized in the form of a histogram; an audio waveform may first be summarized in the form of dominant frequencies over short-time analysis frames. This type of modeling is meant to narrow the *semantic gap* (Smeulders, Worring, Santini, Gupta, & Jain, 2000) between observations that are very obvious to humans, and the noisy low-level measurements from which this observation may be inferable. For example, when provided with pictures of cats and cartoon characters, a human will very easily be able to tell the two apart. However, it is hard to define what color a certain pixel at a certain location should have, in order to belong to a cat or a cartoon character. Generally, objects of focus may also be located at different parts in the image, implying that the exact pixel location may not even be relevant information. When choosing to use a histogram as feature, the picture color values are summarized. The pixel location information is then lost, but we obtain a color and color intensity distribution over the whole image instead. This is therefore a representation of lower dimensionality than when all pixels of the input image are considered in their raw form, but it may give more interpretable information for the statistical model to tell cats apart from cartoon characters.

In recent years, it has increasingly been debated whether going through a feature extraction step is necessary. As an alternative, provided that sufficient training data and powerful deep learning architectures are available, machine learning procedures can be employed for *representation learning* (Bengio, Courville, & Vincent, 2013), directly learning relevant feature representations from $\vec{x}$, without a human expert indicating what information in $\vec{x}$ should be filtered or focused on. Going even further, *end-to-end learning* has also been proposed, in which case the relation between $\vec{x}$ and $y$ is directly learned without the need for an intermediate representation. In many cases, this yields better performance than strategies including intermediate and human-crafted representations (e.g. (Graves & Jaitly, 2014; Long, Shelhamer, & Darrell, 2015)). At the same time, the ultimately learned function from $\vec{x}$ to $y$ becomes harder to interpret for human beings this way.

Since the advent of machine learning, it has been applied to domains which consider phenom-

ena that have natural, physical and objective evidence in the world, although this evidence may not encompass the full breadth of the phenomenon under study. Examples of such domains include speech and natural language (commonly manifesting as spoken audio and text) and music (commonly manifesting as audio). Beyond the physical representation and description of these phenomena, contextual layers of associated modalities, as well as social, human and subjective interpretation, play an important role in the way they are perceived and understood by humans (Davis & Scharenborg, 2017; Liem, Müller, Eck, Tzanetakis, & Hanjalic, 2011).

While machine learning algorithms has proven effective in learning patterns regarding the more descriptive aspects of such phenomena (e.g. (Collobert & Weston, 2008; Hamel & Eck, 2010)), it is still problematic for them to capture notions of true human-like 'understanding' (Hofstadter, 2018; Sturm, 2014). This does not only occur in domains in which 'meaning' may be a shared natural and social phenomenon, with observable and unobservable aspects. Even when the domain considers a pure natural sciences problem with fully objective ground truth, it is not guaranteed that an optimized machine learning procedure mimics human understanding of the problem. This especially can be seen when studying errors made by a seemingly optimized system. In the context of deep neural networks, the notion of *adversarial examples* has emerged: small, humanly unnoticeable perturbations of data on which correct model predictions were originally made, may provoke incorrect model answers with high model confidence (I. J. Goodfellow, Shlens, & Szegedy, 2015).

## 2.3 Contrasting focus areas in psychology and machine learning

Considering the focus areas discussed above, several commonalities and contrasts can be found between interests in psychology and machine learning. Table 1 summarizes several conceptual approximate analogies, as well as their main differences.

In both domains, a prediction task may be studied, involving an $\vec{x}$, $f(\vec{x})$ and $y$. However, the parts of the prediction procedure considered to be of main interest, and the typical types of conclusions being drawn, differ, as also illustrated in Figure 1.

The machine learning concept of training vs. testing has analogues to the difference between exploratory vs. confirmatory factor analysis in psychology. However, in psychology, the focus would be on understanding data, while in machine learning, it is used to verify that a robust model has been trained.

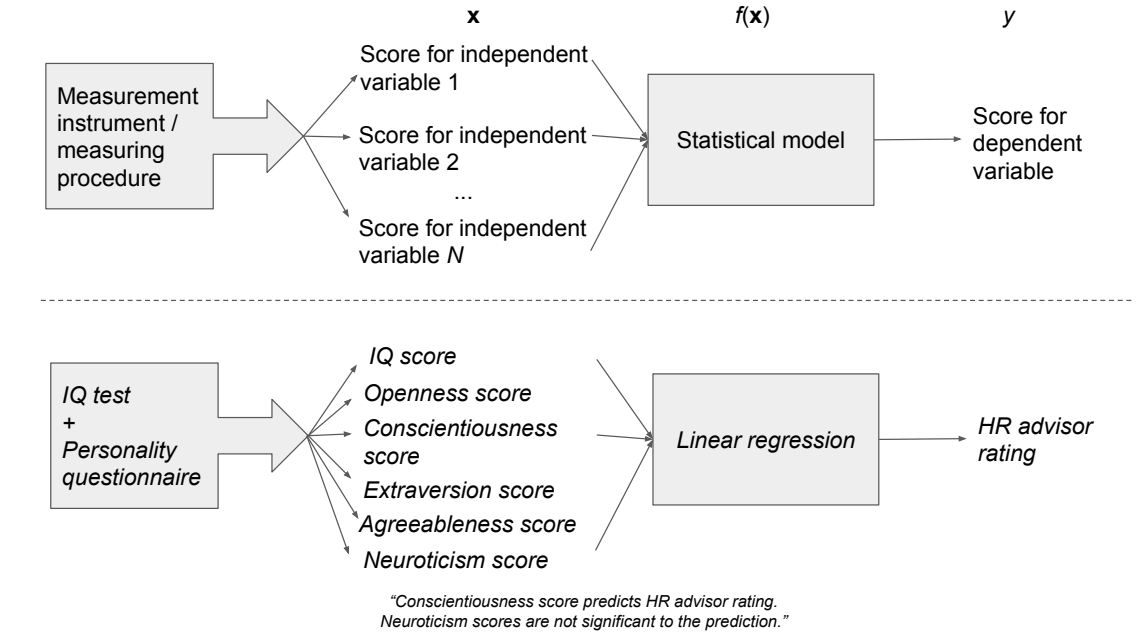
In psychology, human-interpretable meaning of $\vec{x}$ and $y$ is essential: ensuring that $\vec{x}$ will only contain psychometrically validated measurable components that are understandable to a human being, selecting a set of such reasonable components to go into $\vec{x}$, understanding which aspects of $\vec{x}$ then turn out important regarding $y$, and understanding how $y$ human end-users perceive and accept $y$ and $f(\vec{x})$. It is critical that choices of $\vec{x}$ are driven by theory, and corresponding explicit hypotheses about significant relations between the components within $\vec{x}$ and $y$.

The above focus points are out of scope in machine learning. A machine learning expert typically is interested in understanding and improving the *learning procedure*: understanding why $f(\vec{x})$ gets learned in the way it is, where sensitivities lie in the transformation from $\vec{x}$ to $y$, and how prediction errors made by $f(\vec{x})$ can be avoided.

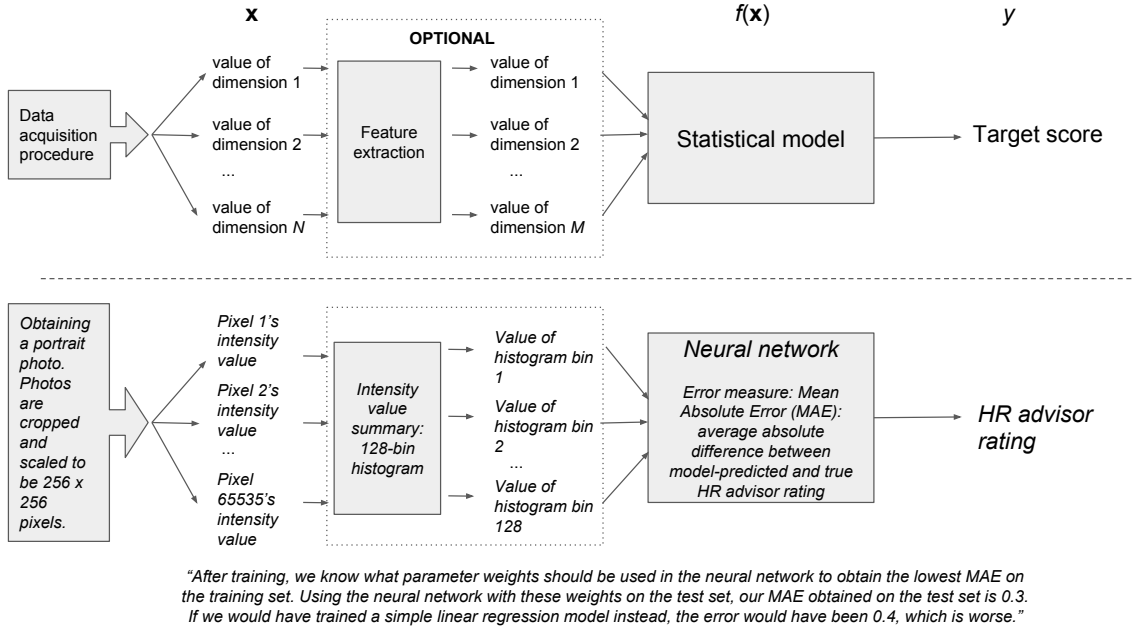
In fundamental machine learning, the focus will exclusively be on this $f(\vec{x})$, and the origins of $\vec{x}$ and $y$ (as well as the reasonableness of any human-interpretable relationship existing between them) will be irrelevant, as long as their statistical properties are well-defined. In applied settings, $\vec{x}$ and $y$ will have further meaning to a human, although in many cases, they

Table 1: Psychology vs. machine learning: conceptual approximate analogies.

| **Psychology** | **Machine Learning** | **Major conceptual differences** |
|---|---|---|
| Exploratory factor analysis | Unsupervised learning | In both domains, if data is available but relationships within the data are unknown, these relationships can be revealed through data analysis. Exploratory factor analysis can be considered as one out of many unsupervised learning techniques, with special focus on explainability of relations in terms of the original input dimensions. |
| Independent/predictor variables | Input data | Each psychological independent variable, as well as its individual dimensions, is human-selected and human-interpretable. In a machine learning setup, input data is usually not manually picked at the individual dimension level. The semantic interpretation of individual dimensions in the data usually also is at a much lower level than that of independent variables in psychology. |
| Variable dimension | Feature | Features express interpretable subinformation in data, where psychological variable dimensions describe interpretable subinformation of an overall variable. Where psychological variable dimensions are explicitly human-selected and human-interpretable, features may be extracted by hand or through an automated procedure. They are still at a semantically lower level than psychological variables dimensions, and not restricted to be psychologically meaningful. |
| Dependent/outcome/ criterion variables | Output/targets/labels/ ground truth (if obtained through acquisition) | These concepts can be considered as equivalents. |
| Statistical model | Statistical model | In psychology, a linear regression model is commonly assumed, and considering other models is typically not the focus. In machine learning, identifying the model that obtains the most accurate predictions (which usually is not a linear regression model) would be the main focus. |
| Model fitting | Training | In psychology, the squared error between predicted and true values will commonly form the error measure to be minimized. In machine learning, more flexible error or cost functions may be used. |




(a) Psychology (in an organizational psychology application).




(b) Machine learning (in a computer vision application).

Figure 1: Prediction pipelines in psychology and machine learning. Abstracted pipelines are given on top, simplified examples of how they may be implemented at the bottom, together with a typical conclusion as would be drawn in the domain.

consider objectively measurable observations in the physical world, with $\vec{x}$ containing raw data

with low-level noisy sensory information.

The flexibility in choosing $f(\vec{x})$ in machine learning is unusual in psychology, where linear regression models are commonly chosen for $f(\vec{x})$, and not typically contrasted with alternative models. The other way around, criterion validity, considering the alignment of $y$ with that what is supposed to be measured, is hardly ever questioned in machine learning settings. In psychology, even though certain types of measures (e.g. supervisor rating as indicator of job performance in the personnel selection problem) tend to dominate, criterion validity is an explicitly acknowledged topic.

When machine learning is to be applied to psychological use cases, $y$ will consider human-related latent concepts, for which no direct and objective measuring mechanisms exist yet in the physical world. When seeking to predict these concepts, it can be debated whether $\vec{x}$ should be expressed at the latent human concept level (constructs/meaningful independent variables) as well. This would be natural for a psychologist, but controversial for a machine learning expert.

Alternatively, an empiricist approach can be taken, purely considering sensory observations, and trying to relate these directly to $y$. This would be natural for a machine learning expert, but controversial for a psychologist. As a possible compromise, if $\vec{x}$ consists of raw data observations, the use of hand-crafted features forms a data-driven analogue to the use of variable dimensions relating to constructs in psychology, even though extracted features will be at a semantically much lower level.

Following these considerations, when applied machine learning methodology is to be integrated in a psychological predictive pipeline, various ways of integration can be imagined. Beyond illustrations of several examples in Figure 1, further example diagrams are illustrated in Figure 2.

1. Keep a traditional psychological pipeline, with traditional input and output data, but consider alternative statistical models to the commonly used linear regression model. This would boil down to varying the choice of statistical model in a traditional psychological pipeline as shown in Figure 1a, top.

2. Keep a traditional machine learning pipeline, (as shown in Figure 1b, top), but ensure that features extracted from raw signals are psychologically informed.

3. Explicitly replace a traditional measurement instrument by a data-driven equivalent. In that case, $\vec{x}$ consists of high-dimensional raw data (e.g., video data), but we wish to turn it into associated traditional instrument scores (e.g., personality trait assessments), so our $y$ can be seen as a transformed version of $\vec{x}$—say, $\vec{x'}$—, at a commonly understood semantic level in psychology, which then can be (re)used in more comprehensive pipelines.

   For going from $\vec{x}$ to $\vec{x'}$, hand-crafted features can also be extracted. Subsequently, a statistical machine learning model is employed to learn correspondences between these feature values, and the traditional instrument scores (Figure 2a).

   Alternatively, instead of performing a hand-crafted feature extraction step, a sophisticated machine learning model can be employed to directly learn a mapping from raw data observations to $\vec{x'}$ (Figure 2b). This would be a way to apply automatic *representation learning* in psychological use cases.

   In feature engineering, a human should explicitly define how an input signal should be transformed, while in representation learning, this would be the task of the chosen statistical model. Especially if it is not very clear how a target instrument score may

concretely relate to information in sensory input data, automated representation learning may therefore yield more optimized mappings than a human can indicate.

In other words, if the predicted target labels are scores of traditional instruments, and the practitioner is sure that criterion and content validity are indeed maintained in the automated learning procedure, representation learning may be an interesting data-driven way to make use of known psychological vocabularies, while bypassing explicit treatment of the semantic gap. However, at the same time, the explicit treatment of the semantic gap through feature engineering can be likened to theory-forming, while in representation learning, a human will have much less control of what the learning algorithm will focus on.

4. Directly seek to learn a meaningful mapping from raw sensory data in $\vec{x}$ to a dependent variable $y$, omitting any intermediate feature or representation extraction steps. This would be an *end-to-end learning* scenario. Conceptually, this approach is close to the representation learning approach mentioned in the previous item. As major difference, in representation learning, the predicted variables are intended to become an alternative to outcomes of a traditional measurement instrument. Therefore, they usually form an intermediate step in a prediction pipeline, replacing the feature extraction block. In case of end-to-end learning, $y$ is the direct output to predict, without including any intermediate explicit representation steps (Figure 2c).

### 2.4 Conclusion

With the main methodological interests of psychology and machine learning being mapped, we now identified relevant contrasts and correspondences between these interests. With this in mind, in the next section, we will proceed by giving an introduction to common personnel selection criteria. Then, Section 4 will illustrate how varying methodological insights into the personnel selection problem can come together in a data-driven solution.

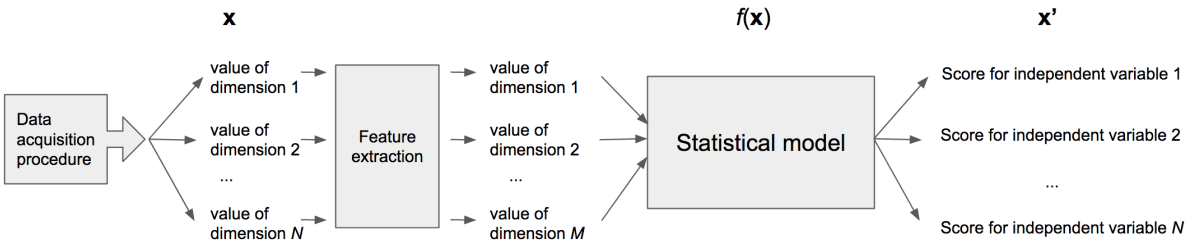
## 3 The personnel selection problem

Historically, personnel selection has been approached as a problem in which *future job performance* should be predicted from job candidate evidence, as provided during the personnel selection stages.

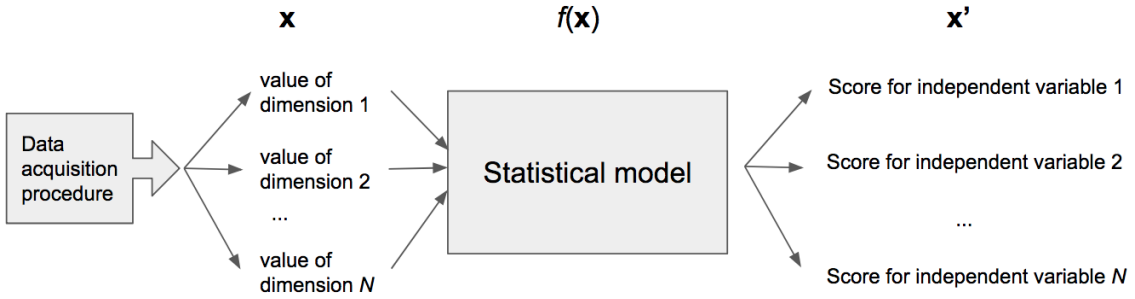
First of all, it is necessary to assume that suitable job candidates exist and that they are willing to apply for the job. Finding these suitable candidates is the focus of recruitment processes. Because it is necessary to have suitable candidates within the applicant pool to be able to select effectively, recruitment and selection are closely intertwined and decisions about selection procedures can influence both processes (Ployhart et al., 2017).

During the early selection stage, the interaction between the applicant and the hiring organization is still low. More precisely, organizations have to rely on limited information (e.g., applicant resumes) in order to decide who to reject and who to keep in the applicant pool. The next stage usually consists of more time-consuming selection procedures, such as face-to-face interviews and/or tests run by assessment centers.

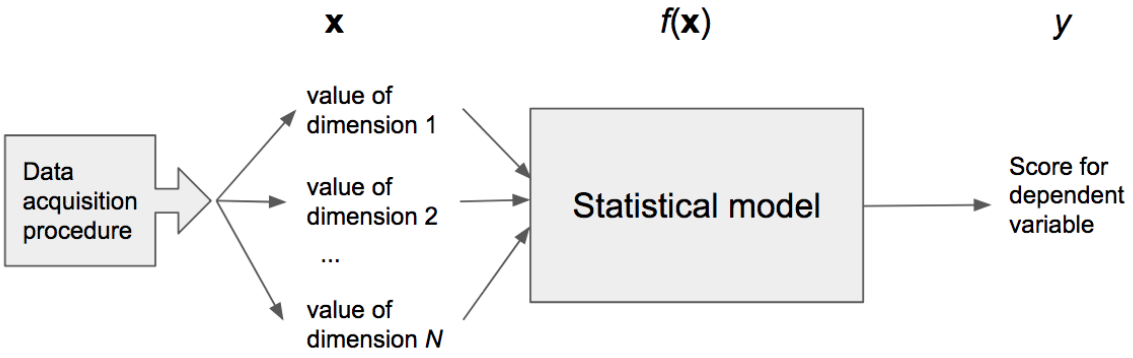
It is based on the hypothesis that individual characteristics such as Knowledge, Skills, Abilities and Other characteristics (*KSAOs*) are predictive of individual outcomes, such as job performance (Guion, 2011). Thus, candidates whose KSAOs fit the job demands are the ones




(a) A machine learning approach replaces a traditional measurement instrument. Hand-crafted features extract information from raw data. These are subsequently used in a prediction pipeline, in which correspondences are learned between obtained feature scores, and psychologically meaningful variable scores that were obtained in correspondence with the raw input data.




(b) A machine learning approach replaces a traditional measurement instrument. Representation learning is applied: a sophisticated statistical model should directly learn the correspondences between raw data input, and corresponding psychologically meaningful variable scores.




(c) A machine learning approach replaces the full psychological pipeline. End-to-end learning is applied: a sophisticated statistical model should directly learn the correspondences between raw data input, and corresponding psychologically meaningful constructs.

Figure 2: Various ways in which psychological and machine learning prediction pipelines can be combined.

that should be hired. This leads to several central classical questions of interest to personnel selection research, in which technological opportunities increasingly play a role, as discussed below.

### 3.1 How to identify which KSAOs are needed?

When an organization needs to select applicants, the first question to be posed is what the organization is looking for. This will be expressed in the form of KSAOs. The logical process to determine KSAOs is to derive these from the job description, and a description of how the job contributes to the organizational goals. For example, if the goal of a hospital is to cure patients, a surgeon in the hospital will be expected to e.g. successfully operate upon patients, correctly analyze the patient's history, coordinate assistants' activities and follow recognized practices during the operation. The needed KSAOs will then, among others, include knowledge and skills regarding techniques for diagnosing and treating injuries and diseases, the ability to tell when something is wrong, and deductive reasoning. Attention to detail, stress tolerance, concern for others and dependability will be further important characteristics.

The KSAOs ideally are derived from a thorough job analysis. A well-known systematic taxonomy of job descriptions, resulting from decades of analyzing jobs, is the occupational net O*NET[1], which forms the largest digital job taxonomy, containing experience and worker requirements and worker characteristics. In practice, however, job descriptions and person specifications sometimes are drawn up in only a few hours by an organization (Cook, 2016).

The characteristics which will be measured during a selection procedure should logically follow from the required KSAOs. In the example of applicants for the occupation of a surgeon, it therefore is important to not only collect information about an applicants' education and experience, but also to measure abilities and traits such as deductive reasoning capacities, attention to detail, concern for others and stress tolerance. A large array of measurement procedures exist to assess applicants' capacities and traits, varying from self-reported personality questionnaires to cognitive tests, work sample tests, structured interviews and role play exercises. As discussed earlier in Section 2, the measures that are explicitly intended to assess constructs (traits, abilities) are often labeled 'signs', whereas measures which aim to assess a sample of relevant performance or behavior (e.g., simulating an operation on a mock patient) are often labeled 'samples'. In practice, most often sign-based measures such as interviews are used (because they are efficient and easy to conduct), although samples often show a good predictive validity (Schmidt & Hunter, 1998).

Smith (M. Smith, 1994) distinguishes between three domains of job characteristics: *universals*, which are characteristics required by all work, *occupationals*, which refer to characteristics required by certain jobs but not others, and *relationals*, referring to characteristics needed to relate to others in specific organizational settings. According to Smith, cognitive ability, vitality, and work importance form the category of universals. The personality factor Conscientiousness (i.e, being organized and structured and able to work on a problem untill the end) may arguably also be seen as a universal. While the aforementioned characteristics have been shown to be relevant for good job performance across most professions, specialized knowledge and certain aspects of personality are examples of occupationals. For a career as a musician, for instance, emotional sensitivity, which is an aspect of emotional intelligence, may be more important than for a job as accountant. Relationals are important to specific settings, and imply a focus

[1] `https://www.onetonline.org`

on values and norms, and the fit ('chemistry') with the people working in those settings such as co-workers, supervisors and management. Relationals mostly are referred to as aspects of *person-organization fit*. More precisely, relationals play an important role when comparing occupations in different organizational settings. For instance, a lawyer in a large commercial bank might require other relationals than a lawyer in a non-profit governmental organization that assists people in poor neighborhoods.

## 3.2 How to measure KSAOs?

After defining which KSAOs are needed, it is necessary to develop or decide for the personnel selection procedures in order to find out which applicants fits the job best. Usually, personnel selection is a multi-hurdle approach, meaning that applicants have to pass different stages before they actually receive a job offer. In a first step, applicants might provide a written resume, afterwards they could be asked to answer to a personality and cognitive ability test. Finally, they might be invited to show their abilities within a face-to-face job interview. Desirably, every single step of the selection process should be psychometrically sound and useful to reveal applicants' KSAOs. As described in Section 2, this means that the selection procedures have to prove to be reliable and valid. For instance, if hiring managers develop a job interview to measure applicants' KSAOs, they have to decide about at least three aspects that may influence psychometric properties of the interview:

- They need to decide for an administration medium. Face-to-face interviews, videoconference interviews and digital interviews all have an impact on applicants' performance ratings (Blacksmith, Willford, & Behrend, 2016; Langer, König, & Krause, 2017) which consequently may affect validity of the interview.
- The degree of standardization of the interview must be decided. This can affect its reliability (Schmidt & Hunter, 1998). In the case of an unstructured interview (i.e., interviewers are allowed to engage in unstructured conversation with the applicant and they have no standardized evaluation criteria), reliability of the interview is at risk because interviewer *A* may evaluate an applicant based on different evaluation standards than interviewer *B*. In other words, if these two interviewers interview the same applicant, the interview scores will likely differ, the interviewers will come to different conclusions about hirability of the applicant, and one interviewer might want to hire while the other might want to reject. In contrast, questions and evaluation of answers in a structured interview are highly standardized. This makes interviews and therefore interview scores more comparable, leading to less noise in the data.
- Lastly, hiring managers need to decide about potential interview questions to capture required KSAOs (Pulakos & Schmitt, 1995). If a job requires programming skills and the interviewer asks questions about applicants' behavior in conflict situations, the interview will neither appear face valid (i.e., applicants would not understand why this is a job related question), nor content valid (i.e., its content will not reflect programming skills as the construct it aims to measure), nor will it be construct valid (i.e., the score on this question will not correlate with other measures capturing programming skills), nor will it demonstrate concurrent (i.e., if the applicant had good grades in a programming course) or predictive (i.e., predict if the applicant will be a good programmer) validity.

To conclude, assessing a selection procedure's reliability means to assess if applicants hirability ratings will be similar for each time that the applicant undergoes (parts of) the selection procedure. In order to evaluate validity of a selection procedure, it is necessary to estimate if a selection procedure appears job related, if it correlates to related constructs and if it predicts important outcomes.

Spreading the attention to other selection procedures, tests focusing on general mental ability (GMA), such as intelligence tests, were shown to have high validity at low application cost (Cook, 2016; Schmidt & Hunter, 1998). Considerable attention has also been paid to personality measures (Morgeson et al., 2007). The five factor model of personality (known as the Big Five: Agreeableness, Conscientiousness, Extraversion, Openness to experience, Neuroticism) (McCrae & Costa, 1999) is widely accepted and used in and outside the field of psychology. In the case of personnel selection, Conscientiousness has especially shown to be a valid predictor for job performance in various organizational contexts (Barrick & Mount, 1991).

However, caution is warranted when assessing personality in the early selection stage, in which resumes are the most frequently used selection instrument. Recruiters may infer impressions from resume data that go beyond the reported factual content. For example, they may attempt to assess an applicant's personality from the resume, which in turn is used to evaluate the applicant's employability. Disconcertingly, there is no research showing that resume-based impressions of applicants' personality are correct. Still these impressions may influence applicants' hirability ratings. In other words, hiring managers have very limited insight into applicants' actual behavior and individual characteristics as they may only have seen applicants' resumes, yet they may still infer much more from the resumes than is appropriate.

This might be a reason why organizations and researchers search for new, efficient sources of information in order to gain additional insights into applicants in early stages of the selection process. However, evidence on the validity of recruiter impressions of the applicants' characteristics based on new, possibly richer sources of applicant information than classical resumes (e.g., from video resumes) is still scarce.

An exception is an experimental study by Waung et al. (Waung, Hymes, & Beatty, 2014) on the effect of resume format on candidate evaluation and screening outcomes among a group of MBA students. When mock applicants were evaluated based on their video resumes, they were rated as less open, extraverted, physically attractive, socially skilled, and mentally capable, and more neurotic than when the same applicants were evaluated based on their paper resumes. Those who were rated as more socially skilled and more conscientious had a higher probability of positive ratings. In another study, Apers and Derous (Apers & Derous, 2017) examined the equivalence of video versus paper resumes on applicants' personality and job suitability ratings. They concluded that resume type did not clearly affect applicant ratings. For instance, personality inferences from video resumes appeared as (in)valid as those from paper resumes. Furthermore, Nguyen & Gatica-Perez (Nguyen, Frauendorfer, Mast, & Gatica-Perez, 2014) developed a computational framework to predict personality based on nonverbal cue extraction. However, with exception to the prediction of Extraversion, results did not support the claim that it is possible to accurately predict various applicant characteristics through automatic extraction of nonverbal cues.

Recent technological developments have opened the door to measuring personality in innovative and possibly more valid ways, such as via Facebook behavior or serious games (Chamorro-Premuzic, Winsborough, Sherman, & Hogan, 2018). These technological developments have sparked interest in both psychologists and computer scientists. For instance, there is evidence that computer-based personality judgments based on digital cues are more accurate than those

made by humans (Youyou, Kosinski, & Stillwell, 2015). The data-driven challenges discussed in this chapter, focusing on predicting personality from online video resumes and YouTube clips (Escalante et al., 2018; Ponce-López et al., 2016) can be considered as further examples of interest in these new algorithm-based methods, even if it has explicitly been presented and disseminated in the technical world.

### 3.3 Dealing with judgment

Selection procedures rely severely on assessors who judge applicants' characteristics. Assessors include interviewers but also assessment center observers and managers assessing work-sample performances. As particularly interviews are among the most frequently used selection methods (e.g. (Ryan, McFarland, Shl, & Page, 1999)), it is important to focus on judgment accuracy and the characteristics of good judges. Furthermore, a focus on ratings by others seems warranted, as it has been proposed that one of the reasons for the relatively low predictive validity of personality measures is the heavy reliance on self-reports, which may contain several biases such as individual differences in faking (Morgeson et al., 2007).

Oh et al. (Oh, Wang, & Mount, 2011) indeed provided evidence for this idea by showing that other-ratings of personality improve the predictive validity of personality for job performance. Similarly, among a sample of sales people, Sitser (Sitser, 2014) was able to demonstrate that other-rated personality traits were able to better predict manager-rated job performance than self-rated traits. In particular, the other-rated personality trait Proactivity, proved to be a strong predictor of job performance. Generally, it can be stated that observer ratings contribute to explaining job performance over and above solely self-report ratings of personality, while this is not the case the other way around (i.e., self-report ratings do not add to explaining variance in job performance over and above the variance explained via observer ratings). However, it has to be noted that observer ratings are also not free from problems, as these ratings might also be faked (König, Steiner Thommen, Wittwer, & Kleinmann, 2017).

As can be seen, studies such as the above have mainly focused on the difference between self- and other-ratings in terms of predictive validity. In the domain of person perception research, the focus has been somewhat different, namely focusing on the search for 'the good judge': "*the oldest concern in the history of research on accuracy is the search for the good judge … the kind of individual who truly understands his or her fellow humans*" (Funder, 1999). In this tradition, Ambady, Bernieri, and Richeson (Ambady, Bernieri, & Richeson, 2000) have demonstrated that merely 'thin slices' of expressive behavior related to Extraversion already result in remarkably accurate judgments of unacquainted judges.

To approach the issue of judgment accuracy, Funder (Funder, 1999) has developed the well-known Realistic Accuracy Model (RAM). RAM states that the degree to which judgments are accurate is moderated by the following factors: good targets, good traits, good information, and finally, good judges (Funder, 2012).

Good targets are very judgeable individuals who may be more transparent than poor targets. Good traits (e.g., extraversion) are more visible than others (e.g., neuroticism) and therefore can be more easily judged. Good information implies good quantity (e.g., a one-hour assessment provides more trait information than a speed-dating exchange) and good quality (e.g., when a person is comfortable and responds to good interview questions, higher-quality information will result). Finally, good judges are better able to detect and use behavior cues to form an accurate personality trait inference.

Yet, HR practices seem to disregard the possibility that individual differences exist in judg-

ment accuracy. A stream of research has focused on potential judge characteristics which may explain individual differences in judgment accuracy. Among these researchers are Christiansen et al. (Christiansen, Wolcott-Burnam, Janovics, Burns, & Quirk, 2005), who used the term *dispositional reasoning* to label individual differences between judges in their complex knowledge of how traits relate to each other and to behaviors, and of situations' potential to elicit traits into manifest behaviors. Christiansen et al. were able to show the importance of dispositional reasoning in predicting judgmental accuracy. Taking this thinking further, De Kock et al. (De Kock, Lievens, & Born, 2015, 2017) provided support for the idea that dispositional reasoning showed incremental validity above general intelligence in predicting judgmental accuracy. In sum, such studies show the importance of asking the question who the external observer is, if we seek better predictive validity of other-ratings.

## 3.4 What is job performance?

So far, a discussion on selection procedures has been provided; however, how 'job performance', the criterion that should be predicted, is appropriately measured has not been discussed. Usually, job performance is considered at an individual level. Frequently, organizations use *supervisor ratings* of past and existing employees as criterion, which are usually easy to generate and/or readily available. However, supervisors are humans, and their ratings may be biased. As a consequence, the usefulness of supervisor ratings as indicators of job performance can be challenged. For example, a supervisor may really like employees who chat about football, which then boosts these employees' performance ratings. Similar issues might occur when designing algorithm-based selection procedures. If the algorithm is trained on predicting supervisor ratings, it will likely learn from biases that supervisors inject into the rating. In the end, the algorithm selects applicants who like to watch football instead of focusing on job relevant skills and abilities.

Beyond supervisor ratings, other common performance indicators for individual employees involve scores regarding sales, number of successful actions or interventions, and customer satisfaction. Recently, new criteria have received attention from researchers and practitioners, namely *extra-role performance*, such as *organizational citizenship behavior* (e.g., helping co-workers), *work engagement* and *deviant behavior* (counterproductive work behavior).

These new criteria may all account for the fact that individual performance, which is most commonly the main criterion of most selection procedures, may not actually translate to organizational performance (Ployhart et al., 2017). For instance, employees showing best possible job performance, but at the same time leading to a negative climate in their teams, may consequently be of more harm for the organization than that they benefit the organization.

Furthermore, in selection research distinction is made between *maximal behavior* (how a person could perform) and *typical behavior* (how a person typically performs). Classical selection procedures, such as job interviews and assessment centers, often only assess applicants' maximal performance, as applicants try to create the best possible impression in such selection situations (Peck & Levashina, 2017). This also implies that they may exhibit impression management behavior (e.g., they exaggerate their past achievements or behave unnaturally in the assessment center (Peck & Levashina, 2017)). Therefore, these selection procedures might not really predict applicants' actual everyday job performance.

### 3.5 Conclusion

In this section, we introduced the job selection problem mainly from a psychological point of view. We highlight that it usually is multi-hurdle approach aiming at finding the best suited applicant given a job description which includes the necessary KSAOs for a job. Selection approaches such as interviews should prove to be valid predictors of relevant criteria (e.g., job performance). In the next section, we will describe a use case of a potential new way of selecting applicants.

## 4 Use case: an explainable solution for multimodal job candidate screening

In this section, we will discuss the data-driven 2017 ChaLearn Looking at People Job Candidate Screening Challenge (Escalante et al., 2017). Besides, as a use case, we will focus on a particular submission to this Challenge (Achmadnoer Sukma Wicaksana & Liem, 2017) and its expansion (Achmadnoer Sukma Wicaksana, 2017). In this work, the chosen solution was explicitly designed to be explainable and understandable to personnel selection experts in psychology.

In alignment with the overall themes of this chapter, the current section will particularly focus on discussions with respect to psychological and machine learning viewpoints on data-driven personnel selection and explainability. As a consequence, technical discussions will only be presented in summarized form; for further details, the reader is referred to the original introduction of the Challenge (Escalante et al., 2017), the overview paper of solutions submitted to the Challenge (Escalante et al., 2018), and the paper and thesis originally describing the solution presented as a use case here (Achmadnoer Sukma Wicaksana, 2017; Achmadnoer Sukma Wicaksana & Liem, 2017).

### 4.1 The Chalearn Looking at People Job Candidate Screening Challenge

The 2017 ChaLearn Looking at People Job Candidate Screening Challenge (Escalante et al., 2017)[2] is part of a series of data-driven 'Looking at People' Challenges, focusing on automated visual analysis of human behavior. For each Challenge, an unsolved analysis problem is proposed, and for this problem, data and target labels are acquired at scale by the Challenge organizers. Subsequently, participant teams sign up to the Challenge, upon which they get access to training data (the data on which solutions are to be trained), as well as validation data (data which can be used for evaluation, while participants are refining their solutions), both also including 'ground truth' target labels. Participants will then propose a final system solution, that will be run on an evaluation dataset, for which the target labels were not released to the participants before.

The Challenge is run in *coopetition* format: on one hand, it is a competition in which centralized data sets are used for training, intermediate validation, and final testing. On the other hand, cooperation is possible and encouraged, as participants are required to openly share their solutions to the problem. As all participants had access to exactly the same data, the Challenge offers useful benchmarking insight, allowing different solutions to be compared against each other.

[2] `http://chalearnlap.cvc.uab.es/challenge/23/description/`

Following an earlier Challenge on apparent personality analysis (Ponce-López et al., 2016), the Job Candidate Screening Challenge focused on predicting apparent personality (the Big Five personality dimensions), as well as interviewability, from short video clips. These can be seen as 'thin slices' (Ambady et al., 2000), giving short but informative insight into a participant.

Given the importance of explainability in job candidate screening processes, the coopetition had both a *quantitative* stage and a *qualitative* stage. The quantitative stage was framed as a pure machine learning problem. For this, the Mean Absolute Error (MAE) was chosen as the evaluation metric, comparing the predictions made by proposed systems with the 'true' scores in the ground truth dataset. MAE comparisons between participant submissions were performed separately for each of the Big Five traits, as well as the interviewability score.

MAE is a common evaluation metric to measure accuracy for a continuous variable. It is a negatively-oriented score, meaning that the lower the score is, the better. It can be turned into a positively-oriented *accuracy* score by subtracting it from 1 ('a perfect system').

More precisely, $MAE$ can be formulated as

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |p_i - g_i|$$

with $N$ being the total number of video excerpts in the test set, $p_i$ being the predicted value for the variable of interest, and $g_i$ being the ground truth value. As a consequence, the Accuracy $A$ can be formulated as

$$A = 1 - MAE.$$

In the qualitative stage, participants were instructed to focus on the explainability of interviewability scores. The required output for this stage was a textual description: it should explain both the workings of a chosen quantitative model, as well as the result of the prediction obtained by using this model. As for the choice of the quantitative model, participants could (re)use any of the solutions submitted to the quantitative stage, or propose a solution of their own. For the assessment of the qualitative textual descriptions, experts in psychological behavior analysis, recruitment, machine learning and computer vision were invited as jury members. Solutions were scored on a scale of 0 to 5 on five criteria:

- **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?
- **Model Interpretability**: Are the explanations useful to understand the functioning of the predictive model?
- **Creativity**: How original / creative are the explanations?

For further details on the Challenge setup and various participants' submissions, the interested reader is referred to the overview papers in (Escalante et al., 2017, 2018).

## 4.2 Dataset

The dataset for the Challenge was acquired as a corpus for first-impression and apparent trait analysis. For this, HD 720p YouTube videos of people facing and speaking English to camera were acquired. Care was taken that the dataset encompassed diversity on several properties, such as gender, age, nationality, and ethnicity. Only good-quality videos in which a unique adult person was facing the camera were considered; from these, at most six 15-second clips were generated for each video, which would not have visual or audio cuts in them. In the end, this yielded 10,000 15-second video clips. For the coopetition, 6,000 of these clips were marked as training data, 2,000 as validation data, and 2,000 as test data, on which the final rankings would be obtained.

Besides the audiovisual video data, speech transcripts were provided for the Job Candidate Screening Challenge, transcribed by a professional transcription service which yielded 435,984 words (out of which 14,535 unique words), with 43 words per clip on average. A full data summary is given in (Ponce-López et al., 2016).

Regarding the annotation of the video clips in terms of personality traits and interviewability, crowdworkers on the Amazon Mechanical Turk platform were provided with an online annotation interface involving pairs of 15-second videos, as shown in Figure 3 (Ponce-López et al., 2016). The following instructions were provided to the crowdworkers:

> "*You have been hired as a Human Resource (HR) specialist in a company, which is rapidly growing. Your job is to help screening potential candidates for interviews. The company is using two criteria: (A) competence, and (B) personality traits. The candidates have already been pre-selected for their competence for diverse positions in the company. Now you need to evaluate their personality traits from video clips found on the Internet and decide to invite them or not for an interview. Your tasks are the following. (1) First, you will compare pairs of people with respect to five traits: Extraversion = Friendly (vs. reserved); Agreeableness = Authentic (vs. self-interested); Conscientiousness = Organized (vs. sloppy); Neuroticism = Comfortable (vs. uneasy); Openness = Imaginative (vs. practical). (2) Then, you will decide who of the 2 people you would rather interview for the job posted.*" (Ponce-López et al., 2016)

Not all possible video pairs were evaluated; instead, the small-world algorithm (Watts & Strogatz, 1998) was used to generate a strategic subset of video pairs with good overall coverage, as it provides high connectivity, avoids disconnected regions in the graph, has well-distributed edges, and a minimum distance between nodes (Humphries, Gurney, & Prescott, 2006). As a result, 321,684 pairs were obtained to label 10,000 videos. In order to convert pairwise scores to cardinal scores, the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952) was fitted using Maximum Likelihood estimation. Detailed explanations on how this can be done are described in (Chen et al., 2016). The final cardinal scores were set to be within the $[0, 1]$ interval. Annotation reliability was verified through reconstruction; the reconstruction accuracy of all annotations was found to be over 0.65, and the apparent trait annotations were found to be highly predictive of invite-for-interview annotations, with a significantly above-chance coefficient of determination of 0.91 (Escalante et al., 2018).

Summarizing the descriptions above, for the quantitative stage of the Challenge, input data consists of 15-second video fragments (video and audio) and their corresponding textual transcripts. The associated target labels consider scores on each of the Big Five personality

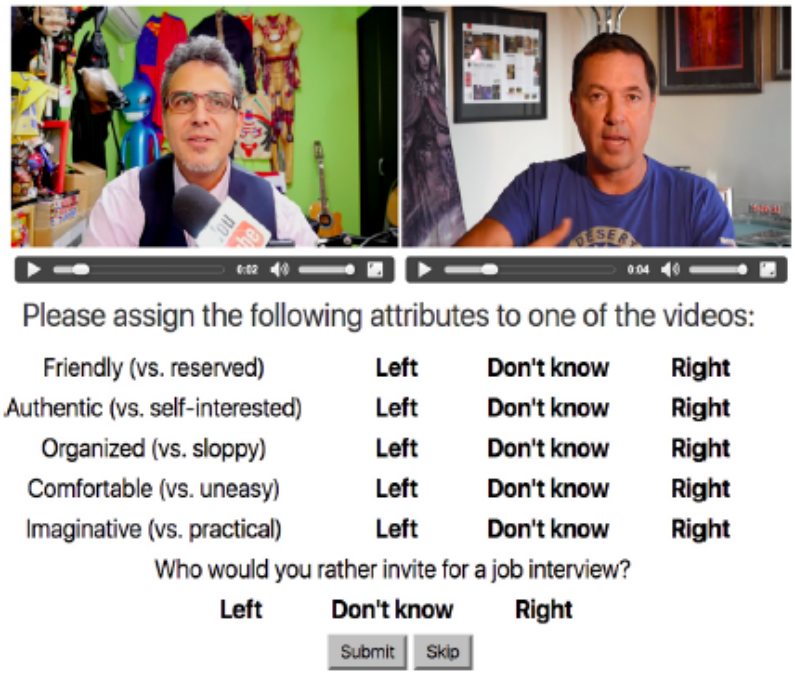



Figure 3: Interface of pairwise comparison to collect labels

traits, as well as interviewability, which all were obtained through crowdsourcing: in all cases, these scores are a numeric value in the $[0, 1]$ range.

### 4.3 General framework of a potential explainable solution

As use case illustration of a potential explainable solution to the Challenge, the work of Achmadnoer Sukma Wicaksana and Liem (Achmadnoer Sukma Wicaksana, 2017; Achmadnoer Sukma Wicaksana & Liem, 2017) is presented here. This work was intended to provide an explainable machine learning solution to the data-driven job candidate screening problem, while explicitly keeping the proposed solution understandable for non-technical researchers and practitioners with expertise in organizational psychology. This was done by designing the system pipeline in consideration of common traditional methodological practice and focus points in job candidate screening (see Section 2). This way, the system was meant as an illustration to trigger discussions and collaborations across disciplines.

The overall system diagram for the proposed system pipeline is given in Figure 4. The general framing closely follows an applied machine learning pipeline (similar to Figure 1b), including an explicit feature extraction step. As such, the setup follows the second suggestion for potential integrations between psychological and machine learning setups, as outlined in Section 2.3.

The input data considers video, audio and text: for each of these, dedicated hand-crafted features are extracted from raw data in various modalities and categories. In other words, the authors proposed several types of information to be extracted from the raw visual, audio and textual data, which all should be understandable with respect to the job candidate screening problem. The details of the chosen categories will be further discussed in Section 4.3.1.

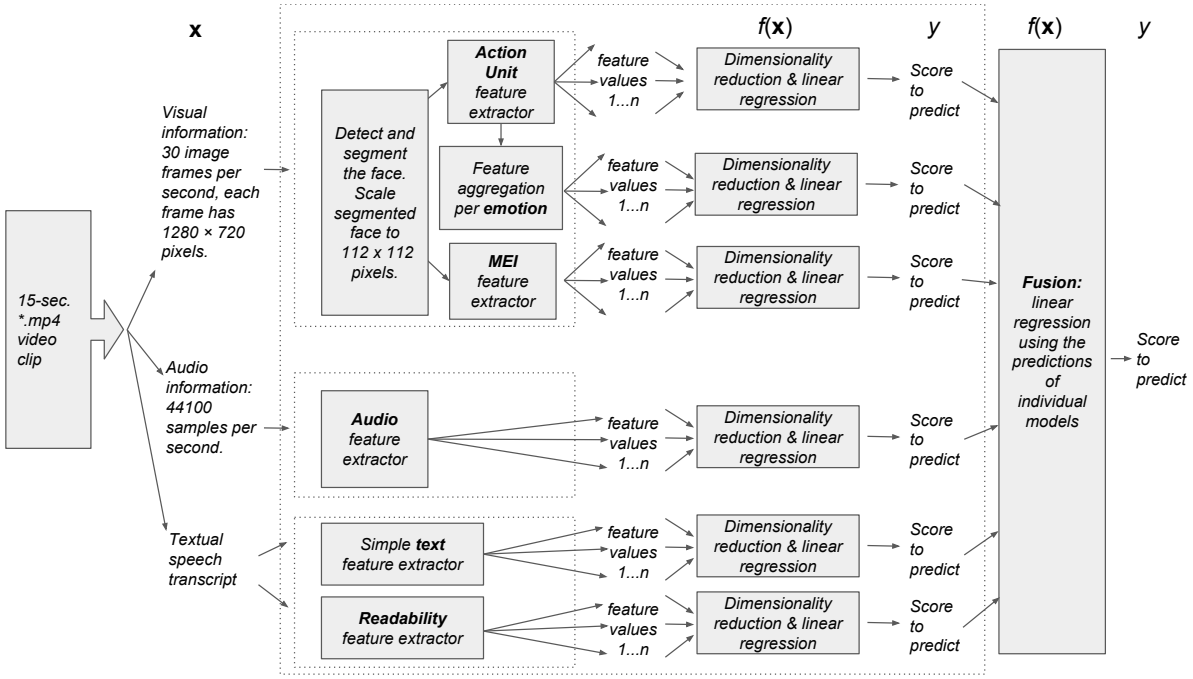



Figure 4: Overall system diagram for the work in (Achmadnoer Sukma Wicaksana, 2017).

The choice to transform the raw data into hand-crafted features, rather than employing an automatically learned representation or an end-to-end learning setup (see Section 2.3), was explicit and deliberate. From an accuracy perspective, machine learning solutions employing an intermediate, hand-crafted feature extraction step typically do not perform as well as solutions which employ heavier automatic learning from raw data. However, as clear benefit, in a hand-crafted feature extraction step, the information extracted from the raw data is controlled and informed by the insight and interpretation of a human practitioner. As such, the explicit definition of features to be extracted in a machine learning pipeline can be seen as an alternative to the explicit choice of theory-driven independent variable dimensions in a traditional psychological setup.

Also regarding the choice of $f(\vec{x})$ (the model that relates the feature values to the dependent variable), it was taken into account that traditional psychological approaches would usually fit a linear regression model. In the current pipeline, this also was done, although in a slightly more elaborate setup than in traditional psychological practice.

First of all, rather than only employing Ordinary Least Squares estimation for the linear model fitting, various regression optimization variants were studied, as further explained in Section 4.3.2. Furthermore, a way had to be found to apply *fusion* of the information from different modalities and feature categories. For this, after training separate linear models per feature category for a dependent variable of interest, the predictions of each of these linear models were used as input to a second regression layer, in which a meta linear model was trained for the dependent variable of interest. This process was separately performed for each of the dependent variables relevant to the Challenge (the scores for each of the Big Five traits, and the interviewability score).

As will further be detailed in the following subsections, within individual feature categories, several dozens of feature dimensions were considered. The final regression step takes six values (one for each feature category) as input. From a traditional psychology perspective, this would be considered a relatively big regression, with many variable dimensions. In contrast, from a machine learning perspective, the approach uses unusually few dimensions: as also will be discussed in Section 4.3.3, it is not uncommon for machine learning pipelines to employ thousands of feature dimensions.

### 4.3.1 Chosen features

The dataset contained information in several modalities: visual information in the video, audio information in the video, and textual information in the form of the speech transcripts.

In the visual modality, information relating to persons' facial movement and expression were considered: in various previous works, these were mentioned as good indicators for personality traits (Borkenau, Brecke, Möttig, & Paelecke, 2009; Naumann, Vazire, Rentfrow, & Gosling, 2009; Waung et al., 2014). More specifically, regarding visual content, the open-source OpenFace library (Baltrušaitis, Mahmoud, & Robinson, 2015) was used to detect and segment the face from frames in each video. Segmented face images were standardized to be 112 x 112 pixels. Beyond segmenting faces, OpenFace also offers a feature extraction library that can extract and characterize facial movements and gaze. Using this feature extraction library, the three visual feature sets were obtained: an Action Unit representation, an Emotion representation, and a Motion Energy Image representation.

Action Units (AU) are subcomponents of facial expressions, which both have been studied in psychology and social and affective signal processing, and which are encoded in the Facial Action Code System (FACS) (Ekman & Friesen, 1978; Ekman & Rosenberg, 2005). OpenFace is able to extract several of these AUs, as listed in Table 2, and indicate AU *presence* (indicating whether a certain AU is detected in a given time frame) and *intensity* (indicating how intense an AU is at a given time frame).

For each AU, three statistical features are derived for usage in our system, aggregating information from the different frames in the particular video. The first feature is the percentage of time frames during which the given AU was visible in a video. The second feature considers the maximum intensity of the given AU in the video. The third feature considers the mean intensity of the AU in the video. As 18 AUs are detected, with three features per AU, 52 features are considered in total for the Action Unit representation.

In affective analysis, combinations of AUs are usually studied. For example, Happiness is evidenced in a face when the cheeks are raised and the lip corners are pulled up. Therefore, AU combinations were hard-coded for the seven basic emotions (Happiness, Sadness, Surprise, Fear, Anger, Disgust and Contempt), as shown in Table 3. Then, the three statistical features as above were considered, but now aggregated over all AUs relevant to the emotion. This yields 21 features in total for the Emotion representation.

Finally, the resulting face segmented video from OpenFace was also used for a Motion Energy Image (MEI) representation. MEI is a grayscale image that shows how much movement happens on each pixel throughout the video, with white indicating a lot of movement and black indicating less movement (Bobick & Davis, 2001). In order to capture the overall movement of a person's face, a Weighted Motion Energy Image (wMEI) is constructed from the resulting face segmented video. wMEI was proposed in the work by Biel et al. (Biel, Aran, & Gatica-Perez, 2011) as a normalized version of MEI, by dividing each pixel value by the maximum pixel

| Action Unit | Description |
|---|---|
| AU1 | Inner Brow Raiser |
| AU2 | Outer Brow Raiser |
| AU4 | Brow Lowerer |
| AU5 | Upper Lid Raiser |
| AU6 | Cheek Raiser |
| AU7 | Lid Tightener |
| AU9 | Nose Wrinkler |
| AU10 | Upper Lip Raiser |
| AU12 | Lip Corner Puller |
| AU14 | Dimpler |
| AU15 | Lip Corner Depressor |
| AU17 | Chin Raiser |
| AU20 | Lip stretcher |
| AU23 | Lip Tightener |
| AU25 | Lips part |
| AU26 | Jaw Drop |
| AU28 | Lip Suck |
| AU45 | Blink |

Table 2: Action Units that are recognized by OpenFace and its description

value.

For the construction of wMEI, it was important to use face-segmented video data, rather than unsegmented full frames. This is because in several cases, videos were recorded in public spaces or while the subject was moving. As a consequence, many pixels in the video corresponding to the background of the scene will also display considerable movement. By only considering face-segmented video data, the focus of analysis will be on the subject's true facial movement. As feature description of the wMEI image of a given video, several statistical features were chosen: the mean, median, and entropy.

For the audio, the focus was on prosodic features, capturing emphasis patterns during speaking. In previous work (Biel et al., 2011), these also were shown to correlate with personality traits. Paralinguistic speech emphasis patterns, which give insight into the tone of voice, have been recognized to be powerful social signals (Nass & Brave, 2005). For this work, speech features were extracted using the MATLAB toolbox developed by the MIT Media Lab (Caneel, 2005; Pentland, 2004). The features that were used are listed in 4; in all cases, the mean and standard deviation over the full video's audio were used. As a consequence, 12 features were used here in total.

Based on findings in organizational psychology, personality traits are not the only (and neither the strongest) predictors for job suitability and hiring decisions. As mentioned in (Schmidt & Hunter, 1998), for example, General Mental Ability (GMA) also is both a valid and strong predictor for job performance.

While formal GMA assessments were not available for subjects in the Challenge dataset, it was considered that language use may indirectly reveal GMA characteristics, such as the use of difficult words. Therefore, for the textual video transcripts, features were chosen that would

| Emotion | Action Units |
|---|---|
| Happiness | 6 + 12 |
| Sadness | 1 + 4 + 15 |
| Surprise | 1 + 2 + 5 + 26 |
| Fear | 1 + 2 + 4 + 5 + 7 + 20 + 26 |
| Anger | 4 + 5 + 7 + 23 |
| Disgust | 9 + 15 |
| Contempt | 12 + 14 |

Table 3: Emotions and its corresponding Action Units that construct them

Table 4: Audio features and its description

| Audio Features | Description |
|---|---|
| F0 | Main frequency of audio |
| F0 conf. | Confidence of F0 |
| Loc. R0 pks | Location of autocorrelation peaks |
| # R0 pks | Number of autocorrelation peaks |
| Energy | Energy of the voice |
| D Energy | Derivative of the energy |

capture the comprehensiveness and sophistication of speech.

Two categories of textual features were considered. First of all, speaking density was approximated by two simple measures: total word count and the number of unique words spoken in the video. Furthermore, linguistic sophistication was approximated by calculating several Readability indexes over the spoken transcripts: ARI (E. A. Smith & Senter, 1967), Flesch Reading Ease (Flesch, 1948), Flesch-Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975), Gunning Fog Index (Gunning, 1952), SMOG Index (McLaughlin, 1969), Coleman Liau Index (Coleman & Liau, 1975), LIX, and RIX (J. Anderson, 1983), as implemented in an open-source contributed library for the Python NLTK toolkit. Each of these Readibility indexes stemmed from existing literature, targeted at quantitative assessment of the reading difficulty level of a given text.

### 4.3.2 Regression model

Prediction of the dependent variable scores was done through regression. Given the large amount of derived related features (for example, multiple alternative Readibility indexes), multicollinearity between input variables is likely to occur. This is undesirable, as the considered feature dimensionality may be higher than the true dimensionality, considering independent components. Furthermore, if a regression model is fitted with highly correlated features as input, it becomes harder to determine the effect per individual feature on the end result.

In order to mitigate the effect of multicollinearity in the model, several techniques were considered. The first one used *Principal Component Regression* (PCR): employing the prominent Principal Component Analysis (PCA) technique before feeding the results to Ordinary Least Squares (OLS) Regression. Next to this, Ridge and Lasso Regression were considered, which

incorporate $l2$ and $l1$ regularization technique on the linear regression model, respectively.

PCA is a linear transformation that converts a set of correlated variables into uncorrelated variables called principal components. This technique also ensures that the highest principal component accounts for the highest variation of data. Thus, by selecting several principal components, data variation over the most important principal component dimensions is maintained, while the amount of dimensions to work with reduces significantly. The transformation from original feature vectors to new principal components can be expressed as a linear matrix multiplication:

$$Y = X * W$$

where $X$ is the original feature matrix, having $N$ rows of $K$-dimensional observations, $W$ is the linear transformation matrix, with $K$ eigenvectors of $M$ dimensions, and $Y$ is the transformed feature matrix, expressing the same $N$ observations as $M$ principal components.

These principal components then will be fed as input to OLS Regression. This regression technique is a simple linear regression technique that estimates the coefficients by minimizing a loss function with a least squares method:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^P}{\operatorname{argmin}} \|y - X\beta\|_2^2.$$

The other two regression models that are considered for the system incorporate a penalizing function to the least squares regression model. By doing so, they try to shrink coefficients, so that the significance of a subset of input features will be eminent by the value of the coefficients. Coefficient estimation for Ridge and Lasso regression is conducted as follows:

$$\hat{\beta}^{Ridge} = \underset{\beta \in \mathbb{R}^P}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

$$\hat{\beta}^{Lasso} = \underset{\beta \in \mathbb{R}^P}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

with $\lambda$ expressing the tuning parameter. When $\lambda$ is equal to zero, this becomes a least squares regression; when $\lambda$ is infinitely large, the $\hat{\beta}^{Ridge}$ is 0. For other values of $\lambda$, a balance is taken between fitting a linear model and shrinking the coefficients.

These three regression models were considered both for the individual feature category modeling, as well as for the fusion step.

### 4.3.3 Quantitative performance

For understanding the quantitative performance aspects of the system, two experiments were done. First of all, it was assessed which of the three regression techniques would perform best. Secondly, regarding input features, it was assessed whether all features should be used in the system, or only those that through an initial correlation analysis were revealed to be significant ($p < 0.05$) with respect to the dependent variable to be predicted.

From the experimental results, which are reported in detail in (Achmadnoer Sukma Wicaksana, 2017), the best-performing regression technique differed per situation, although the absolute differences in accuracy for the different regression techniques were very small. As

for the choice of feature sets, slightly better results were obtained for using full feature sets, rather than pre-selected feature sets resulting from correlation analysis. Full configurations and detailed performance tables can be found in (Achmadnoer Sukma Wicaksana, 2017).

Quantitative accuracy performance scores of the final, optimized system are reported for all dependent variables in Table 5. For comparison, the table also reports system performance on an earlier published version of the system (Achmadnoer Sukma Wicaksana & Liem, 2017) (which used a smaller feature set and did not yet optimize regression techniques). Furthermore, performance scores are reported for two other proposed solutions: the work in (Gorbova, Lusi, Litvin, & Anbarjafari, 2017), employing similar features to ours, but with a multi-layered perceptron as statistical model; and the work in (Kaya, Gurpinar, & Salah, 2017), which obtained the highest accuracies of all participants in the quantitative Challenge.

This latter work employed several state-of-the-art feature sets, some of which resulting from representations learned using deep neural networks, with considerably higher dimensionality than our features (thousands of feature dimensions). While the system described in this chapter does not outperform the scores of (Kaya et al., 2017), performance differences are small in the absolute sense, at the benefit of an easily understandable model with simple regression architectures, and a much small number of feature dimensions.

Table 5: Comparison of quantitative performance (accuracy) between the system described as use case in this chapter (Achmadnoer Sukma Wicaksana, 2017), an earlier version of the system presented at the ChaLearn workshop (Achmadnoer Sukma Wicaksana & Liem, 2017), and two other proposed solutions for the ChaLearn Job Candidate Screening Challenge.

| Categories | Use case system | Earlier version | (Gorbova et al., 2017) | (Kaya et al., 2017) |
|---|---|---|---|---|
| Interview | 0.8950 | 0.8877 | 0.894 | 0.9198 |
| Agreeableness | 0.9008 | 0.8968 | 0.902 | 0.9161 |
| Conscientiousness | 0.8873 | 0.8800 | 0.884 | 0.9166 |
| Extraversion | 0.9001 | 0.8870 | 0.892 | 0.9206 |
| Neuroticism | 0.8945 | 0.8848 | 0.885 | 0.9149 |
| Openness | 0.8991 | 0.8903 | 0.896 | 0.9169 |

## 4.4 Opportunities for explanation

For the qualitative stage of the Job Candidate Screening Challenge, a textual explanation to accompany a quantitative prediction had to be automatically generated. For the system described in this chapter, the decision was made to generate an extensive report, displaying an explanation and a contextualization of measured values corresponding to each feature used in the system.

The choice was made to describe each feature, and not to make a more optimized textual summary that would pre-filter descriptions of particular variables. This was done, as the authors felt that in a real-life setting, a practitioner with domain knowledge should have the freedom to choose whether to see a full report, or only parts of it. Furthermore, the authors wished to avoid that any information would inadvertently be hidden from an end user, while an end user may actually have interest in it. Indeed, as will be discussed in Section 5.2, perceived controllability of an algorithmic solution is an important requirement for making it acceptable

for end users.

As all features in the system were chosen to be humanly interpretable, a short human explanation was made for each feature, that was printed in the report. Furthermore, as an early screening scenario was adopted, the purpose of the explanation would be to allow for a selection of interviewable candidates to be made from a larger candidate pool. Therefore, for each feature, the score of a candidate for this feature was contextualized against 'what usually would be observed': in this case, the minimum and maximum feature values obtained on the pool of 6,000 earlier rated subjects in the training set. Furthermore, it was indicated at what percentile the current video's score would be with respect to the training set candidates, to further give a sense of how 'usual' this person's observed feature value was.

As all dependent variable score predictions of the system are based on linear data transformations, the weight of each input feature dimension with respect to the final prediction model can easily be traced back. This information was not used for selecting or prioritizing information. However, for those feature values that had the strongest absolute weights with respect to the final prediction, the report would indicate whether this feature value would correlate positively or negatively with the dependent variable.

A sample excerpt from a generated report is given in Figure 5. For possible future work, it will be interesting to develop a more user-friendly presentation of the descriptions, in connection to dedicated user interaction optimizations.

```
********************
* 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 47.000000 (percentile: 62).
In our model, a higher score on this feature typically leads to a higher
overall assessment score.
```

Figure 5: Example description fragment.

## 4.5 Reflection

The ChaLearn Challenge is in many ways interesting to the job candidate screening problem. Generally, the Challenge outcomes suggest that each of the personality characteristics, as well as the interviewability score, can be predicted with high accuracy using algorithmic procedures. At the same time, when aiming to connect these findings to psychological practice, there still are important open questions that will need more explicit attention, in particular regarding validity. For example, a major question to be asked is *what kind of information the ground truth scores truly indicated*.

Regarding validity of the dependent variable scores, the use of crowdsourcing to get non-expert first-impression annotations at scale is interesting. Considering the findings on observer judgment vs. self-reporting in Section 3.3, crowdsourcing could be a useful way to get observer judgments at scale. While crowdsourcing allows for reaching a population with higher diversity than the typical WEIRD (Henrich et al., 2010) samples (Behrend, Sharek, Meade, & Wiebe, 2011), crowdwork usually is offered in a marketplace setting, in which anyone interested in performing a task and meeting the task's qualifications can do so. This means that certain workers may perform many ratings in a batch, but others may only perform a single annotation task and then move on, causing potential annotator biases within the data that are hard to control up front.

Typically, crowdworkers would also perform work for monetary reasons, and only be willing to spend little time on a single task, meaning that the tasks should be compactly presented. This is also evidenced in the way the annotation task was presented (see Figure 3): only a single question is asked per personality trait. Even if this question may have come from a psychometrically validated instrument, there are more underlying facets to a psychological trait than the single question currently being posed. Fully equating the currently posed item questions with the underlying trait (e.g., considering that 'Extraversion == Friendly (vs. reserved)') would not be logical to a psychologist, and this choice should be explicitly defended. While the requirement for crowdsourcing tasks to be compact makes it unrealistic to employ full-length instruments, it still is possible to employ more than one item per trait, and it should be investigated whether doing so will yield higher psychometric reliability and validity.

While it was reported that the personality trait scores were highly predictive for the interviewability score (Escalante et al., 2018), another concern involves potential response bias. Looking at the annotation task, all items were presented with the positive valuation on the left, and the negative valuation on the right. This, together with the pairwise setup, may invite annotators to consistently prefer the person 'they like best'. It is not guaranteed that the commonly advised strategy of reverse wording (varying positively and negatively phrased items) will truly yield better results (van Sonderen, Sanderman, & Coyne, 2013); especially in a crowdsourcing setup, in which workers may be focused on finishing the task fast, high attention to wording variations is not guaranteed. However, this aspect should be researched more deeply.

Looking at the ChaLearn data, especially at what kinds of videos score particularly high and low on each of the traits and the interviewability score (as shown in Table 6), one may wonder whether the first impression ratings may alternatively be interpreted as youthful attractiveness ratings. Again, this may be a consequence of the preference-oriented setup of the annotation task.

Escalante et al. (Escalante et al., 2018) analyzed potential judgment biases in the data regarding ethnicity, race and age, and found low-valued but significant positive biases towards judgment of female subjects on all personality traits except for Agreeableness, and low-valued but significant negative biases towards judgment of African-American subjects. Further analyses on potential age biases indicate that the youngest and oldest people in the dataset (estimated age under 19 or over 60) had below-chance probabilities for interview invitations, and that within the 'common working age' range, younger women and older men had higher prior probabilities of interview invitations. Perfectly performing systems trained on this data will therefore inhibit the same biases, and explicit awareness of this is needed.

Finally, it should be remarked that the data did not consider official job applications, but rather the general impression that candidates would leave in a more spontaneous setting. In a

real application setting, a broader set of KSAOs will be of relevance, and not all personality traits may be equally important to job performance. Therefore, again, the interviewability assessments should at present most strongly be interpreted as preference ratings, rather than true invite-to-interview probabilities.

# 5 Acceptability

So far, explainability in the context of job candidate screening has solely been considered with respect to scientific stakeholders: computer scientists and psychologists interested in data-driven technologically-supported solutions. However, when implementing novel personnel selection approaches, there are two further stakeholders that need to special attention: *applicants* and *hiring managers*.

Applicants are affected by novel personnel selection procedures, as their information and job application will be subject to the novel procedures. Second, hiring managers need to decide which personnel selection procedures are adequate to select applicants for a job.

In this section, research on the acceptance of novel selection technologies by applicants and hiring managers is therefore discussed, as understanding main interests and concerns of these stakeholders will be paramount in successfully implementing novel selection technologies in practice.

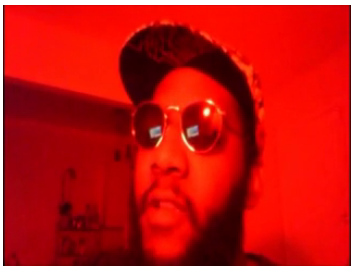
## 5.1 Applicants

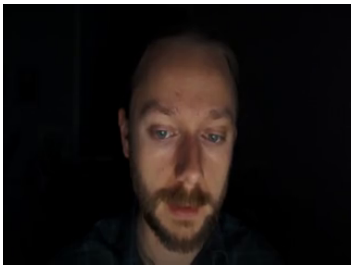
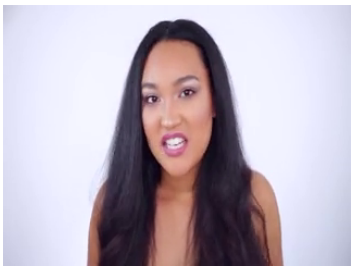
Most research on applicants' acceptance of personnel selection procedures was and still is influenced by Gilliland (Gilliland, 1993), who proposed a model for the justice of selection procedures. In his model, he highlighted the importance of formal (e.g., job relatedness), interpersonal (e.g., interpersonal treatment) and transparency related characteristics (e.g., honest during the selection process) but also distributive justice (e.g., fairness of outcomes) of selection procedures on the overall acceptance of these procedures. Additionally, he pronounced that all of these variables consequently affect applicants' self-perceptions (e.g., self-esteem), reactions to the organization (e.g. organizational attractiveness) and eventually later job performance. Based on his model, scales to measure acceptance of selection procedures were developed (e.g., (Bauer et al., 2001)) and a tremendous amount of research supports the importance of examining acceptance of selection procedures (Chapman, Uggerslev, Carroll, Piasentin, & Jones, 2005).

Unfortunately, research about acceptance of novel technologies for personnel selection lags at least ten years behind current technological possibilities (Ployhart et al., 2017). To be clear, in the last two decades most research focused on the acceptance of technology-mediated job interviews (see (Blacksmith et al., 2016) or web-based testing (Bauer et al., 2006). Just recently, acceptance research has called for studies using more up-to-date technologies (Ployhart et al., 2017) which was answered by Langer and colleagues (Langer, König, & Papathanasiou, 2017) who found that an algorithm-based job interview including automatic analysis of social behavior (e.g., smiling) and a virtual agent as interviewer is less accepted than a videoconference interview with a human interviewer. More specifically, they found that lower transparency and interpersonal warmth of the algorithm-based procedure decreased its acceptance.

In the context of algorithm-based selection procedures, Gilliland's model in combination with findings from the study of Langer and colleagues and research about more classical technology-

Table 6: Snapshots of videos with high and low values for each dependent variable of interest to the quantitative state of the ChaLearn challenge.

| Traits | Extraversion | Agreeableness | Conscientiousness |
|---|---|---|---|
| |  |  |  |
| score | 0.046729 | 0.000000 | 0.048544 |
| |  |  |  |
| score | 0.925234 | 0.912088 | 0.951456 |

| Traits | Neuroticism | Openness | Interview |
|---|---|---|---|
| |  |  |  |
| score | 0.031250 | 0.111111 | 0.149533 |
| |  |  |  |
| score | 0.937500 | 0.977778 | 0.915888 |

enhanced selection approaches can shed light on variables influencing acceptance of algorithm-based selection procedures. More precisely, applicants who are confronted with algorithm-based selection procedures will likely be concerned about *formal characteristics*, *interpersonal characteristics*, and *transparency-related characteristics* of a selection procedure.

First of all, applicants who are screened by any kind of algorithm-based personnel selection approach will be concerned about *formal characteristics* of the procedure. In the terms of Gilliland, these would be perceived job relatedness of the procedure, applicants' opportunity to perform (i.e., applicants' opportunity to show their skills and abilities) and objectivity (i.e., objective treatment during and results of the selection procedure). Regarding job relatedness, if it is obvious for applicants that a selection procedure is relevant to predict job performance, it will be accepted. In the case of algorithm-based selection procedures, there are approaches that appear more job related than others. For instance, using web scraping and machine learning approaches to scan through applicants' social media profiles may appear less job related than a serious game which mimics the aspects of a job and measures actual behavior during the game.

Similar examples are useful to understand that some selection procedures offer more opportunity to perform than others. It may be hard for applicants to put their best foot forward when an organization uses their social media information to evaluate applicants' job fit, whereas algorithm-based job interview solutions could at least appear to provide more opportunity to show one's skills. Compared to classical job interview procedures, however, algorithm-based procedures may provide less perceived opportunity to perform as applicants do not really know how they can influence the algorithm in a way that it will positively evaluate their performance (Langer, König, & Papathanasiou, 2017). In the case of objectivity, algorithm-based solutions could even possess advantages over classical selection procedures, as automatically evaluated resumes or job interviews might be less prone to subjective human influence (e.g., applicants attractiveness; (Gilmore, Beehr, & Love, 1986). However, as discussed in Section 4.5 of this chapter, algorithms themselves might have learned from human biases and consequently not be more consistent than human hiring managers (Caliskan, Bryson, & Narayanan, 2017).

Second, *interpersonal characteristics* of selection procedures influence their acceptance. For instance, the behavior of hiring managers can positively influence applicants' willingness to accept a job offer (Chapman et al., 2005). In the case of algorithm-based personnel selection, applicants might be concerned that human influence is minimized, such that there is no representative of the organization taking his or her time to at least look at their application. Applicants may perceive this as a signal of lower appreciation, thus detrimentally affecting acceptance (Langer, König, & Papathanasiou, 2017). However, positively influencing interpersonal characteristics of algorithm-based selection procedures appears to be challenging. An idea could be to add virtual agents to the algorithm-based selection situation (e.g., in the case of algorithm-based job interviews). However, the results of Langer and colleagues show that this does not seem to entirely solve the problem, and instead introduces new issues, such as negative feelings against the virtual character (which might be caused by the uncanny valley (Mori, MacDorman, & Kageki, 2012)).

Third, *transparency-related issues* seem to relate to applicant reactions. In the sense of Gilliland, a procedure is transparent if applicants are treated honestly, if they receive information about the selection procedure, and if they receive timely and helpful feedback about their performance. It is worth mentioning that the acceptance variables Job relatedness, Opportunity to perform, and Objectivity might all be affected by transparency: for a transparent procedure, it is more obvious if it is job related, if it is possible to show ones skills and abilities, and to evaluate if it treats applicants objectively. More precisely, applicants in a transparent selection procedure

know which decision criteria underlie the selection decision; furthermore, if rejected, they receive information about why they were rejected. In the case of algorithm-based selection, it is not yet commonly made explicit which input variables led to a certain outcome (e.g., a rejection). Therefore, it would be impossible to derive any explanation about which decision criteria were involved. As a consequence, applicants do not know what is expected of them, neither do they know how to improve if they were rejected. In an attempt to increase acceptance of algorithm-based selection tools, incorporating ideas generated in the field of explainable artificial intelligence (Biran & Cotton, 2017) will therefore be useful.

At the same time, Langer and colleagues (Langer et al., 2018) tried to improve transparency of an algorithm-based selection procedure through provision of information about an algorithm-based job interview procedure (e.g., about technical details, and about what an algorithm-based selection procedure is looking for). In the end, participants were positively and negatively affected by this information, indicating that the relation between transparency and acceptance is not just a simple 'the more the better' relation. Instead, it seems that transparency consists of different aspects that need to be addressed in order to understand its influence on acceptance. More precisely, transparency consists of technical details about the selection procedures (e.g., which data are used), justifications of the selection procedure (i.e., why exactly this procedure should be job relevant). Future research should try to reveal other aspects require consideration in order to understand the impact of transparency on acceptance.

## 5.2 Hiring managers

In addition to applicants' view on personnel selection situations, the perspective of hiring managers, which is closely related to the perspective of organizations (Klehe, 2004), needs attention, as they are the ones who will be requested to select an applicant based on the information they receive from any type of screening tool. Additionally, they are also the ones who might be afraid of algorithm-based tools, making them superfluous in personnel selection contexts. For the means of raising acceptance of algorithm-based selection tools, it should therefore be an important step to include hiring managers' opinions and ideas about novel selection devices. Based on previous research (Chapman, Uggerslev, & Webster, 2003; Klehe, 2004; König et al., 2010), it is suggested that hiring managers evaluate algorithm-based selection tools considering the tools' perceived *usefulness*, *objectivity*, *anticipated applicant reactions*, *probability of legal actions*, *controllability*, and *transparency*.

Hiring managers expect novel personnel selection methods to be useful to support their everyday work (Chapman & Webster, 2003). In the case of algorithm-based tools, *efficiency* is the first thing that comes to mind, as these tools may have the potential to quickly screen many applicants. Especially as the use of technology has increased the applicant pool for many organizations, algorithm-based screening tools helping to manage the large amount of applications seem to be a logical solution. Additionally, hiring managers seem to be attracted by easy-to-use selection tools (Diekmann & König, 2015) which should be considered when trying to improve acceptance of algorithm-based screening tools. More specifically, easy-to-use seems to imply 'easy-to-apply', to understand, and to interpret (Diekmann & König, 2015).

Secondly, hiring managers hope for *enhancing objectivity* of selection procedures when implementing novel technologies. For instance, Chapman and colleagues (Chapman & Webster, 2003) propose that by reducing human influence on selection situations, adverse impact (i.e., discrimination of minorities (Hough, Oswald, & Ployhart, 2001)) and human biases (e.g., better ratings for more attractive applicants; (Gilmore et al., 1986)) might be reduced. Therefore, if

an algorithm-based tool can actually prove that it is able to increase objectivity of selection situations, hiring managers will appreciate this fact.

Thirdly, hiring managers seem to anticipate *applicant reactions* towards novel selection tools when considering to implement these tools (Klehe, 2004). If hiring managers conclude that applicants may not like a novel selection procedure, it is less likely to be used for future selection procedures. As we have seen in the section on applicant reactions, they actually cover a wide range of different acceptance variables. Currently, it is still unclear which applicant reaction variables hiring managers consider to be most influential. Nevertheless, this makes it clear that algorithm-based tools do not only need to appear adequate to applicants, they also need to appear reasonable in the eyes of hiring managers.

Fourthly, the *probability of legal actions* is closely related to applicant reactions: when applicants react extremely negatively to selection procedures, they might even sue the hiring company (Bauer et al., 2001). In the case of algorithm-based selection tools, legal actions seem possible, especially when an organization cannot prove the algorithms' validity and objectivity in the sense of preventing adverse impact (Klehe, 2004). Generally, following the European General Data Protection Regulation (Council of the European Union, 2016), applicants will also have the right to demand insight into how their data is processed by algorithmic procedures. In the absence of empirical studies relating to these issues, it seems to be hard for organizations and for developers of algorithm-based selection tools to support validity and to provide evidence for unbiased evaluations made by the algorithm.

Regarding validity, there are studies showing that algorithm-based tools correlate with personality (Campion, Campion, Campion, & Reider, 2016) or with job interview performance (Naim, Tanveer, Gildea, & Hoque, 2015) but empirical findings regarding its predictive validity for actual job performance or other important outcomes influencing organizational performance (e.g., organizational citizenship behavior [employees positive behavior at work]) are scarce. Regarding biases in the evaluation of applicants, recent research indicates that this might be a problem, as algorithms can learn from human biases (Caliskan et al., 2017). It is therefore necessary to not only evaluate the predictive validity of the algorithm, but also its development process, in particular its training procedure, in order to realize whether there could be any bias in the training data that may result in biased applicant scoring.

Fifthly, *controllability* (i.e., being able to control a selection situation) could be hard to achieve when using algorithm-based tools. For instance, the scorings and rankings of applicants performed by algorithms may be used for a fully automated pre-screening, but in this case, there is less controllability for hiring managers, which often is unacceptable. Algorithms should therefore offer the possibility to regain control over the decision, when hiring managers want this option. For instance, it might be possible to develop algorithms in which hiring managers can choose to which aspects of applicants they attach more importance (e.g., personality, cognitive ability).

In the context of controllability, it is further important to note that perceived controllability of algorithm-based tools will likely be lower, if hiring managers have the impression that this tool will replace them in any way. Therefore, it should be clear what the algorithm is intended to do in the selection process–generally, a full replacement solution will not meet acceptance, but rather, the algorithm should support and simplify the work of hiring managers.

Sixthly, an antecedent of all the aforementioned conditions for a positive evaluations of algorithm-based selection tools is *transparency* of the procedure. If a tool is transparent to hiring managers, it is easier to evaluate its usefulness, its objectivity, anticipate applicant reactions and the possibility for legal actions, and to assess its controllability (Langer, König, & Papathanasiou,

2017). In this case, transparency would mean that the process in which applicants are evaluated should be *comprehensible* (i.e., it is clear which characteristics and behavior of applicants will be used for their evaluation), *traceable* (i.e., it is possible to have an insight into why one applicant was preferred over another) and *explainable* (i.e., it is possible for hiring managers to formulate feedback to applicants about why they were rejected).

The previous discussion makes it clear that applicants' and hiring managers' acceptance of technology-supported tools can be affected by many different variables; not all of these necessarily relate to the algorithms or technology themselves. In the following and final section, we will discuss where, within the technological realm, acceptability can be fostered and stimulated.

# 6 Recommendations

In previous sections, we have introduced the job candidate screening problem, as well as common methodologies and viewpoints surrounding this problem, perceived by various scientific disciplines and stakeholders. It is undisputed that explainability is important in the context of algorithmic job candidate screening, and technologically-supported hiring in general. It even may be critical for allowing true interdisciplinary collaboration. However, following the discussions throughout this chapter, it becomes clear that 'explainability' in job candidate screening can actually have many different interpretations, and is relevant to many different parties.

As discussed in Sections 2 and 3, for psychologists, explainability will closely relate to understanding what variables are given to the system, whether their inclusion is supported by evidence and theory, and to what extent these variables have been collected using reliable procedures. As discussed in Sections 2 and 4, for computer science researchers with machine learning interests, explainability will mostly lie in understanding why and how an algorithm will learn certain patterns from data. Finally, as discussed in Section 5, for applicants, algorithmic explainability will mostly deal with formal characteristics and transparency-related characteristics (interpersonal characteristics being a matter of presentation), while for hiring managers, explainability will be desired regarding usefulness, objectivity, anticipated applicant reactions, probability of legal actions, controllability, and transparency.

Against these considerations, in this section, we will make several recommendations on how technologically-supported job candidate screening mechanisms can be improved in ways that will raise their acceptance and foster interdisciplinary collaboration, considering all relevant stakeholders.

## 6.1 Better understanding of methodology and evaluation

### 6.1.1 Stronger focus on criterion validity

In early selection procedures, the scoring of candidates will focus on interviewability: the decision of whether or not this candidate should more closely be screened in person by representatives of the entity that is hiring. At the same time, the selection procedure is actually intended as a way to assess future job performance. As such, this aim should be clearly reflected in the procedure and the resulting scores.

At the same time, generally, as discussed in Section 3, there is no single definition of what 'good job performance' exactly means. A more comprehensive set of variables may need to be

assessed here (e.g. not only individual performance, but possibly also organizational citizenship behavior). We expect that exposing these variables transparently to all stakeholders throughout the process will increase trust in the overall system. As another suggestion, it may be useful to more explicitly include validated KSAOs in automated prediction setups.

In machine learning settings, ground truth labeling and further data annotation tasks are commonly done through crowdsourcing. However, most annotation validation methods focus on reliability (high inter-rater agreement, clear majority votes, accurate reconstruction), but not on validity. While this is less of an issue for objectively verifiable phenomena in the natural world, this is problematic in the case of constructs which are not directly observable. To ensure validity, it is advisable to consider psychometric principles when setting up the instruments to solicit the necessary input from humans. The work by Urbano et al. (Urbano, Schedl, & Serra, 2013) on evaluation in music information retrieval gives further useful examples on how comprehensive evaluation, including verification of validity, can be done in computational settings which partially rely on human judgment.

### 6.1.2 Combining methodological focus points

In machine learning, main attention will be given to $f(\vec{x})$, the mapping from input to output, while in psychology, the main attention is given to ensuring the independent variables $\vec{x}$ are explainable given evidence and existing theory. Psychologists also are interested in searching for *mediator* (variables mediating the influence of a predictor on an outcome) and *moderator* variables (variables influencing the relation between other variables), while in machine learning, paying detailed human attention to individual input data dimensions is often irrelevant, also as the input data is usually at semantically lower levels.

As a consequence, while in popular discourse on technologically-supported hiring, the question tends to emerge 'whether human psychologists or algorithms do a better job at candidate assessment', this question does not make much methodological sense. Considering the value of anticipated applicant reactions, probability of legal actions, controllability, and transparency to a hiring manager, as well the desire of applicants for interpersonal relations, the expertise of a human who is knowledgeable about hiring cannot be omitted and replaced by a machine.

There is interest from both the psychology and computer science/machine learning domains to connect their methods to provide better solutions. As mentioned in Section 2.3, data-driven methods can be integrated with the psychological prediction pipeline at several points. They may offer useful and better alternatives to common linear regression models, inform feature engineering, offer data-driven alternatives for traditional measurement instruments, or can be employed in end-to-end learning. It is possible to define explicit feature extraction steps to extract relevant information from raw data; alternatively, relevant–but usually less interpretable–mappings can automatically be learned from the data.

In terms of expected controversy, it will not be controversial, and easily adoptable, to integrate machine learning methods in a traditional psychological pipeline, as an alternative to the common linear regression model. The other way around, a main interesting challenge for machine learning applied in psychological settings is to ensure that information in the prediction pipeline is psychologically informed. One way to do this, as also performed in the system discussed in Section 4, would be to employ the extraction of hand-crafted features from raw signals, even though they will be at a semantically lower level than common psychological instruments and vocabularies.

It will be interesting to consider offering data-driven replacements of traditional measurement instruments. However, in this case, it is important to carefully integrate theory and psychometrically validated findings in the data and target label preparations. While handcrafted feature extraction is considered old-fashioned in machine learning, it is useful to ensure human interpretability of information extracted from raw signals.

If an explicit feature engineering step is omitted, and there rather would be interest in direct representation learning of equivalent outcomes to a traditional measurement instrument, the advantage would be that the first extracted representation will have a well-known form to a psychologist (e.g., a predicted Big Five trait score). At the same time, with the information extraction procedure in representation learning falling fully on the side of a machine learning algorithm, extreme care should be taken: systems do not always learn what they are supposed to learn, but may inadvertently pick up on other patterns in the data (Sturm, 2014). To mitigate this, it is important to consider various concurrent facets of the problem in the representation learning procedure, and perform careful and extensive validation, as e.g. performed in (Kim, Urbano, Liem, & Hanjalic, 2018).

Given the psychological emphasis on understandable data and constructs, the most controversial integration would be to apply end-to-end machine learning approaches in psychological settings. These will not likely allow for meaningful collaborations, as directly learning mappings from raw data to a dependent variable will not be deemed meaningful to a psychologist, due to the lack of clear interpretable variables underlying the prediction procedure.

## 6.2 Philosophical and ethical awareness

Psychology belongs to the social sciences, while computer science belongs to the natural sciences. In combining the two worlds, viewpoints and validation techniques from both these sciences will need to be bridged: the previous subsection has already discussed several ways in which this may be done.

The differences between theory-driven methodology in psychology and data-driven approaches in computer science touch upon philosophical epistemological debates. When formulating theories and hypotheses, do we miss out on important information, or pick what we want to see without solid foundations? At the same time, when blindly and only trusting data, how solid is this data really, and is it justified to fully give up human interpretation?

The difference has also been discussed and debated within the natural sciences, with several authors pointing out that theory will keep playing an essential role, while data at the same time can help in revealing unexpected effects, disproving earlier beliefs, or steering discovery towards theories we did not think of yet (Bar-Yam, 2016; Mazzocchi, 2016).

A major concern regarding machine learning in the context of job candidate screening has been bias. Algorithms do not have an ethical compass by themselves; if training data reveals undesired societal biases, this will be mirrored in any machine learning solution built on top of this data. For example, if a machine learning model intended to assess potential CEOs will be trained on data from the first half of the 20th century, it may infer that being a Caucasian male is a necessary condition in order to be deemed a suitable CEO.

These are undesirable effects, and the machine learning community has started focusing on blind spots and algorithmic improvements to ensure fairness, reduce bias, and avoid that certain population subgroups will be disadvantaged through algorithmic means (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016; Buolamwini & Gebru, 2018; Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012). At the same time, it should be emphasized that societally undesired

effects of algorithmic procedures typically occur because of biased input data, or because of the algorithmic predictions being unjustly held for the absolute truth. The sensibility to deal with this is a shared responsibility between machine learning experts, domain experts regarding the data, and stakeholders involved with practical implementation. In this sense, algorithms may suitably be used as 'mirrors' to reflect on whether predicted outcomes indeed align with their purpose in the broader context of socio-technical social systems (Crawford & Calo, 2016).

## 6.3 Explicit decision support

For many stakeholder parties, having the opportunity for human control in technologically-supported predictions concerning human beings is an important requirement for acceptability. It already was argued that rather than considering technologically-supported solutions as full replacement of a human being, they should rather be considered as complementary assisting tools for decision support. This aligns with the recent proposition by Barabas et al. (Barabas, Dinakar, Ito, Virza, & Zittrain, 2018) to consider algorithmic predictions as indicators of intervention opportunities, rather than binding predictions. We foresee similar opportunities in job candidate screening: as discussed in the previous subsection, algorithms can assist in pinpointing bias and unfairness issues in data, before full decision pipelines are based upon them. Furthermore, they can usefully help in scaling up the early selection stage; however, this mostly would be to provide a selection of potentially interesting job candidates to a human assessor. As such, only a rough first selection may be needed; rather than seeking a full and 'true' ranking and scoring, the only thing that matters may be that a candidate would fall in the upper quadrant of the candidate pool. If so, evaluation metrics should be adjusted accordingly.

## 6.4 The goal of explanation

As discussed throughout this chapter, the need for explanation may lie at different points for researchers in psychology and machine learning, for job applicants, and for hiring managers.

Regarding the academic perspective on 'how good' a prediction model is, a balance between accuracy and explainability needs to be found. Baseline models can be improved in accuracy by increasing model complexity; at the same time, this makes the model's working less understandable for humans. While a model that clearly fails in finding the best applicants will never be accepted, there might be a point at which increasing accuracy does not bring that much benefit, and better comprehensibility will be favored over pushing accuracy another percent.

Throughout the discussions between co-authors in preparing this chapter, we found that literacy regarding each others' methodologies was a first necessary starting point. If the job candidate screening problem should be tackled from an interdisciplinary or transdisciplinary perspective, psychologists will need to gain basic computer science and machine learning literacy, while computer scientists will need to deepen their knowledge on psychometric validation. Preferably, curricula should be developed that do not only train interdisciplinary literacy, but also hands-on basic skills.

In ongoing discussions on explainability in machine learning, common counter-arguments against explainability are that 'humans beings cannot explain their own reasoning processes well themselves' and 'if it works, it just works'. Considering explainability in the context of technologically-supported job candidate screening methods for hiring managers and candidates, an interesting observation is that explainability actually may not be needed so much in positive

cases, but much more so in negative cases: the parties that will demand explainability, will most likely be candidates who do not get hired.

A question here is whether rejected candidates indeed will be helped by explaining why an algorithm did not assess them well; as discussed in (Langer et al., 2018), more transparency about algorithmic procedures and criteria may actually increase user skepticism. Furthermore, pointing the user at candidates who were successful in their place will also not be a pro-active type of feedback. It may be more beneficial to focus on constructive feedback towards future success; it will be a next grand challenge to research whether machine learning can play a role in this.

These observations align to the review on explanation in the social sciences by Miller (Miller, 2017). As a main insight to include in future research, it is mentioned that explanations are often *contrastive*, *selected* and *social*, rather than only being a presentation of causes. However, within AI, also considering the job candidate screening problem, the main focus so far has been almost exclusively on the latter. By more explicitly including contrastive, selected and social elements, it is expected that explanations towards end users will improve in quality and acceptability.

### 6.5 Conclusion

As we discussed throughout this chapter, psychology and machine learning have complementary methodological interests, that may be combined in various novel ways. Careful and explicit treatment of validity, insight into the diversity of explainability opportunities, solid understanding of varying needs and interests of different stakeholders, and literacy across disciplines will be essential in making interdisciplinary collaborations work out in practice. If this can successfully be achieved, we foresee substantial innovation in the field, positively impacting researchers, practitioners and job candidates alike.

## References

Achmadnoer Sukma Wicaksana. (2017). *Layered Regression Analysis on Multimodal Approach for Personality and Job Candidacy Prediction and Explanation.* Retrieved from `http://resolver.tudelft.nl/uuid:a527395d-f42c-426d-b80b-29c3b6478802`

Achmadnoer Sukma Wicaksana, & Liem, C. C. S. (2017). Human-Explainable Features for Job Candidate Screening Prediction. In *Ieee computer society conference on computer vision and pattern recognition workshops* (Vol. 2017-July). doi: 10.1109/CVPRW.2017.212

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000, 1). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology*, *32*, 201–271. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0065260100800064` doi: 10.1016/S0065-2601(00)80006-4

Anderson, J. (1983). Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*, *26*(6), 490–496. Retrieved from `http://www.jstor.org/stable/40031755`

Anderson, N., Herriot, P., & Hodgkinson, G. P. (2001). The practitioner-researcher divide in industrial, work and organizational (IWO) psychology: Where are we now, and where do we go from here? *Journal of Occupational and Organizational Psychology*. doi: 10.1348/096317901167451

Apers, C., & Derous, E. (2017). Are they accurate? Recruiters' personality judgments in paper versus video resumes. *Computers in Human Behavior*. doi: 10.1016/j.chb.2017.02.063

Baltrušaitis, T., Mahmoud, M., & Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *Fg* (Vol. 06, pp. 1–6). Retrieved from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7284869 doi: 10.1109/FG.2015.7284869

Barabas, C., Dinakar, K., Ito, J., Virza, M., & Zittrain, J. (2018). Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In *Proceedings of the conference on fairness, accountability and transparency* (Vol. 81, pp. 1–15). New York: Machine Learning Research. Retrieved from http://proceedings.mlr.press/v81/barabas18a/barabas18a.pdf

Barrick, M. R., & Mount, M. K. (1991). The big fice personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1–26. doi: 10.1111/j.1744-6570.1991.tb00688.x

Bar-Yam, Y. (2016, 9). The limits of phenomenology: From behaviorism to drug testing and engineering design. *Complexity*, *21*(S1), 181–189. Retrieved from http://doi.wiley.com/10.1002/cplx.21730 doi: 10.1002/cplx.21730

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the Selection Procedural Justice Scale. *Personnel Psychology*, *54*, 387–419. doi: 10.1111/j.1744-6570.2001.tb00097.x

Bauer, T. N., Truxillo, D. M., Tucker, J. S., Weathers, V., Bertolino, M., Erdogan, B., & Campion, M. A. (2006). Selection in the information age: The impact of privacy concerns and computer experience on applicant reactions. *Journal of Management*, *32*, 601–621. doi: 10.1177/0149206306289829

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011, 9). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813. Retrieved from http://www.springerlink.com/index/10.3758/s13428-011-0081-0 doi: 10.3758/s13428-011-0081-0

Bengio, Y., Courville, A., & Vincent, P. (2013, Aug). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798-1828. doi: 10.1109/TPAMI.2013.50

Biel, J.-I., Aran, O., & Gatica-Perez, D. (2011). You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube. *Artificial Intelligence*, 446–449. Retrieved from http://www.idiap.ch/{~}jibiel/pubs/BielAranGaticaICWSM11.pdf

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A Survey. In *Proceedings of the 17th international joint conference on artificial intelligence IJCAI* (pp. 8–13). Melbourne, Australia.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. *Springer-Verlag*.

Blacksmith, N., Willford, J. C., & Behrend, T. S. (2016). Technology in the employment interview: A meta-analysis. *Personnel Assessment and Decisions*, *2*, 2. doi: 10.25035/pad.2016.002

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(3), 257–267. doi: 10.1109/34.910878

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th conference on neural information processing systems*. Barcelona. Retrieved from https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, *43*(4), 703–706. doi: 10.1016/j.jrp.2009.03.007

Bradley, R., & Terry, M. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345. Retrieved from http://www.jstor.org/stable/10.2307/2334029 doi: 10.2307/2334029

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *. In *Proceedings of the conference on fairness, accountability and transparency* (Vol. 81, pp. 1–15). Machine Learning Research. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*, 183–186. doi: 10.1126/science.aal4230

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology*, *101*, 958–975. doi: 10.1037/apl0000108

Caneel, R. (2005). *Social Signaling in Decision Making* (Doctoral dissertation, Massachusetts Institute of Technology). Retrieved from http://groupmedia.media.mit.edu/datasets/Social{_}Signaling{_}in{_}Decision{_}Making.pdf

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2018). New Talent Signals: Shiny New Objects or a Brave New World? *Industrial and Organizational Psychology*, *9*(3), 621–640. doi: 10.1017/iop.2016.6

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant attraction to organizations and job choice: A meta-analytic review of the correlates of recruiting outcomes. *Journal of Applied Psychology*, *90*, 928–944. doi: 10.1037/0021-9010.90.5.928

Chapman, D. S., Uggerslev, K. L., & Webster, J. (2003). Applicant reactions to face-to-face and technology-mediated interviews: A field investigation. *Journal of Applied Psychology*, *88*, 944–953. doi: 10.1037/0021-9010.88.5.944

Chapman, D. S., & Webster, J. (2003). The use of technologies in the recruiting, screening, and selection processes for job candidates. *International journal of selection and assessment*, *11*, 113–120. doi: 10.1111/1468-2389.00234

Chen, B., Escalera, S., Guyon, I., Ponce-Lopez, V., Shah, N., & Simon, M. O. (2016). Overcoming calibration problems in pattern labeling with pairwise ratings: Application to personality traits. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 9915 LNCS, pp. 419–432). doi: 10.1007/978-3-319-49409-8{_}33

Choi, B. C., & Pak, A. W. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clinical and Investigative Medicine*. doi: 10.1016/j.jaac.2010.08.010

Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance*, *18*. doi: 10.1207/s15327043hup1802{_}2

Coleman, M., & Liau, T. L. (1975). A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, *60*(2), 283–284. Retrieved from http://content.apa.org/journals/apl/60/2/283 doi: 10.1037/h0076540

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: deep

neural networks with multitask learning deep neural networks with multitask learning. *International Conference on Machine Learning*. doi: 10.1145/1390156.1390177

Cook, M. (2016). *Personnel selection : adding value through people - a changing picture*. Wiley-Blackwell.

Council of the European Union. (2016). *General Data Protection Regulation.* Brussels: European Union. Retrieved from `http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf`

Crawford, K., & Calo, R. (2016, 10). There is a blind spot in AI research. *Nature*, *538*(7625), 311–313. Retrieved from `http://www.nature.com/doifinder/10.1038/538311a` doi: 10.1038/538311a

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*. doi: 10.1007/BF02310555

Davis, M. H., & Scharenborg, O. (2017). Speech perception by humans and machines. In G. Gaskell & J. Mirkovi\'{c} (Eds.), *Speech perception and spoken word recognition.* (pp. 181–204). Routledge Psychology Press.

De Kock, F. S., Lievens, F., & Born, M. P. (2015). An In-Depth Look at Dispositional Reasoning and Interviewer Accuracy. *Human Performance*. doi: 10.1080/08959285.2015.1021046

De Kock, F. S., Lievens, F., & Born, M. P. (2017). A closer look at the measurement of dispositional reasoning: Dimensionality and invariance across assessor groups. *International Journal of Selection and Assessment*. doi: 10.1111/ijsa.12176

Diekmann, J., & König, C. J. (2015). Personality testing in personnel selection: Love it? Leave it? Understand it! In I. Nikolaou & J. Oostrom (Eds.), *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice* (pp. 117–135). Hove, UK: Psychology Press.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference on - itcs '12* (pp. 214–226). New York, New York, USA: ACM Press. Retrieved from `http://dl.acm.org/citation.cfm?doid=2090236.2090255` doi: 10.1145/2090236.2090255

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*.

Ekman, P., & Rosenberg, E. (2005). *What the face reveals*. Oxford University Press.

Escalante, H. J., Guyon, I., Escalera, S., Jacques, J., Madadi, M., Baro, X., … Van Lier, R. (2017). Design of an explainable machine learning challenge for video interviews. In *Proceedings of the international joint conference on neural networks.* doi: 10.1109/IJCNN.2017.7966320

Escalante, H. J., Kaya, H., Salah, A. A., Escalera, S., Gmur Güçlütürk, Y., Güçlü, U., … Salah, A. A. (2018, February). Explaining First Impressions: Modeling, Recognizing, and Explaining Apparent Personality from Videos. *ArXiv e-prints*.

Flesch, R. (1948). A New Readability Yardstick. *The Journal of Applied Psychology*, *32*(3), 221–233. doi: 10.1037/h0057532

Funder, D. C. (1999). *Personality judgment : a realistic approach to person perception*. Academic Press.

Funder, D. C. (2012). Accurate Personality Judgment. *Current Directions in Psychological Science*. doi: 10.1177/0963721412445309

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics : an introduction* (Second Edition ed.). SAGE Publications.

Gilliland, S. W. (1993, October). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review*, *18*, 694–734. doi: 10.2307/258595

Gilmore, D. C., Beehr, T. A., & Love, K. G. (1986). Effects of applicant sex, applicant physical attractiveness, type of rater and type of job on interview decisions*. *Journal of Occupational Psychology*, *59*, 103–109.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. In *Proceedings of the 3rd international conference on learning representations (iclr2015).* San Diego. Retrieved from https://arxiv.org/pdf/1412.6572.pdf

Gorbova, J., Lusi, I., Litvin, A., & Anbarjafari, G. (2017, 7). Automated Screening of Job Candidate Based on Multimodal Video Processing. In *2017 ieee conference on computer vision and pattern recognition workshops (cvprw)* (pp. 1679–1685). IEEE. Retrieved from http://ieeexplore.ieee.org/document/8014948/ doi: 10.1109/CVPRW.2017.214

Graves, A., & Jaitly, N. (2014). Towards End-To-End Speech Recognition with Recurrent Neural Networks. In *Proceedings of the 31st international conference on machine learning (icml-14).*

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions*. Routledge.

Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill. Retrieved from https://books.google.nl/books?id=ofI0AAAAMAAJ

Hall, G. S. (1917). Practical relations between psychology and the war. *Journal of Applied Psychology*. doi: 10.1037/h0070238

Hamel, P., & Eck, D. (2010). Learning Features from Music Audio with Deep Belief Networks. In *International society for music information retrieval conference (ismir).*

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–-83. doi: 10.1017/S0140525X0999152X

Hiemstra, A. M., Derous, E., Serlie, A. W., & Born, M. P. (2012). Fairness Perceptions of Video Resumes among Ethnically Diverse Applicants. *International Journal of Selection and Assessment*. doi: 10.1111/ijsa.12005

Hofstadter, D. (2018, 1). *The Shallowness of Google Translate, The Atlantic.* Retrieved from https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*, 152–194. doi: 10.1111/1468-2389.00171

Humphries, M., Gurney, K., & Prescott, T. (2006). The brainstem reticular formation is a small-world, not scale-free, network. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1585), 503–511. Retrieved from http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2005.3354 doi: 10.1098/rspb.2005.3354

Kaya, H., Gurpinar, F., & Salah, A. A. (2017, 7). Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs. In *2017 ieee conference on computer vision and pattern recognition workshops (cvprw)* (pp. 1651–1659). IEEE. Retrieved from http://ieeexplore.ieee.org/document/8014944/ doi: 10.1109/CVPRW.2017.210

Kim, J., Urbano, J., Liem, C. C. S., & Hanjalic, A. (2018). One Deep Music Representation to Rule Them All? A comparative analysis of different representation learning strategies. *ArXiv e-prints*.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Technical Training, Research B*(February), 49. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf doi:

ERIC:ED108134

Klehe, U.-C. (2004). Choosing how to choose: Institutional pressures affecting the adoption of personnel selection procedures. *International Journal of Selection and Assessment*, *12*, 327–342. doi: 10.1111/j.0965-075x.2004.00288.x

König, C. J., Klehe, U.-C., Berchtold, M., & Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment*, *18*, 17–27. doi: 10.1111/j.1468-2389.2010.00485.x

König, C. J., Steiner Thommen, L. A., Wittwer, A.-M., & Kleinmann, M. (2017). Are observer ratings of applicants' personality also faked? yes, but less than self-reports. *International Journal of Selection and Assessment*, *25*, 183–192. doi: 10.1111/ijsa.12171

Langer, M., König, C. J., & Fitili, A. (2018). Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. *Computers in Human Behavior*, *81*, 19–30. doi: 10.1016/j.chb.2017.11.036

Langer, M., König, C. J., & Krause, K. (2017). Examining digital interviews for personnel selection: Applicant reactions and interviewer ratings. *International Journal of Selection and Assessment*, *25*, 371–382. doi: 10.1111/ijsa.12191

Langer, M., König, C. J., & Papathanasiou, M. (2017). User reactions to novel technologies in selection and training contexts. In *Annual meeting of the society for industrial and organizational psychology (siop).* Orlando, FL.

Liem, C. C. S., Müller, M., Eck, D., Tzanetakis, G., & Hanjalic, A. (2011). The need for music information retrieval with user-centered and multimodal strategies. In *Mm'11 - proceedings of the 2011 acm multimedia conference and co-located workshops - mirum 2011 workshop, mirum'11* (pp. 1–6). doi: 10.1145/2072529.2072531

Liem, C. C. S., Rauber, A., Lidy, T., Lewis, R., Raphael, C., Reiss, J. D., … Hanjalic, A. (2012). Music Information Technology and Professional Stakeholder Audiences: Mind the Adoption Gap. In *Dagstuhl follow-ups* (Vol. 3). Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik. Retrieved from `http://drops.dagstuhl.de/opus/volltexte/2012/3475/` doi: 10.4230/DFU.VOL3.11041.227

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr).* doi: 10.1109/CVPR.2015.7298965

Mazzocchi, F. (2016). Could Big Data be the end of theory in science? *EMBO reports*, *16*(10). Retrieved from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4766450/pdf/EMBR-16-1250.pdf` doi: 10.15252/embr.201541001

McCrae, R. R., & Costa, P. T., Jr. (1999). The five-factor theory of personality. In *Handbook of personality: Theory and research.* Guilford Press. doi: 10.1007/978-1-4615-0763-5{_}11

McLaughlin, G. (1969). SMOG grading: A new readability formula. *Journal of reading*, *12*(8), 639–646. Retrieved from `http://www.jstor.org/stable/40011226` doi: 10.1039/b105878a

Miller, T. (2017, June). Explanation in Artificial Intelligence: Insights from the Social Sciences. *ArXiv e-prints*.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsiderung the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683-729. doi: 10.1111/j.1744-6570.2007.00089.x

Mori, M., MacDorman, K., & Kageki, N. (2012, June). The uncanny valley. *IEEE Robotics & Automation Magazine*, *19*, 98–100. doi: 10.1109/MRA.2012.2192811

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015). Automated analysis and prediction of job interview performance: The role of what you say and how you say it. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* (pp. 1–14). Ljubljana, Slovenia. doi: 10.1109/fg.2015.7163127

Nass, C., & Brave, S. (2005). Wired for Speech: How Voice Activates and Advances the Human Computer Relationship. *Computational Linguistics*, *32*(3), 451–452. Retrieved from `https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=http://search.proquest.com/docview/1037392793?accountid=15115{%}5Cnhttp://vr2pk9sx9w.search.serialssolutions.com/?ctx{%}7B{_}{%}7Dver=Z39.88-2004{%}7B{&}{%}7Dctx{%}7B{_}{%}7Denc=info:ofi/enc:UTF-8{%}7B{&}{%}7Drfr{%}7B{_}{%}7Did=info:sid/ProQ{%}7B{%}25{%}7D3Aeducation{%}7B{&}{%}7Drft{%}7B` doi: 10.1162/coli.2006.32.3.451

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and social psychology bulletin*, *35*(12), 1661–1671. doi: 10.1177/0146167209346309

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*. doi: 10.1109/TMM.2014.2307169

Oh, I. S., Wang, G., & Mount, M. K. (2011). Validity of Observer Ratings of the Five-Factor Model of Personality Traits: A Meta-Analysis. *Journal of Applied Psychology*. doi: 10.1037/a0021832

Peck, J. A., & Levashina, J. (2017). Impression management and interview and job performance ratings: A meta-analysis of research design with tactics in mind. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2017.00201

Pentland, A. (2004). Social Dynamics : Signals and Behavior. *Proceedings of the 3rd International Conference on Developmental Learning, Oct 2004*, *5*, 263–267. Retrieved from `http://vismod.media.mit.edu//tech-reports/TR-579.pdf`

Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the Supreme Problem: 100 Years of selection and recruitment at the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*, 291. doi: 10.1037/apl0000081.supp

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., … Escalera, S. (2016, 10). ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results. In *European conference on computer vision (eccv 2016) workshops* (pp. 400–418). Amsterdam: Springer. Retrieved from `https://link.springer.com/chapter/10.1007/978-3-319-49409-8{_}32` doi: 10.1007/978-3-319-49409-8{_}32

Pulakos, E. D., & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology*, *48*, 289-308. doi: 10.1111/j.1744-6570.1995.tb01758.x

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence - A modern approach (3. internat. ed.)*. Pearson Education.

Ryan, A. M., McFarland, L., Shl, H. B., & Page, R. (1999). An International Look At Selection Practices: Nation and Culture As Explanations for Variability in Practice. *Personnel Psychology*. doi: 10.1111/j.1744-6570.1999.tb00165.x

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin (1998)*.

Sitser, T. (2014). *Predicting sales performance: Strengthening the personality – job performance linkage* (Doctoral dissertation, Erasmus University Rotterdam). Retrieved from https://repub.eur.nl/pub/51139/

Smeulders, A., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/34.895972

Smith, E. A., & Senter, R. J. (1967). Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, 1–14.

Smith, M. (1994). A theory of the validity of predictors in selection. *Journal of Occupational and Organizational Psychology*. doi: 10.1111/j.2044-8325.1994.tb00546.x

Sturm, B. L. (2014). A simple method to determine if a music information retrieval system is a 'horse'. *IEEE Transactions on Multimedia*. doi: 10.1109/TMM.2014.2330697

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning : an introduction*. MIT Press.

Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in music information retrieval. *Journal of Intelligent Information Systems*. doi: 10.1007/s10844-013-0249-4

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: let's learn from cows in the rain. *PloS one*, *8*(7), e68967. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23935915http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3729568 doi: 10.1371/journal.pone.0068967

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of'small-world'networks. *Nature*, *393*(6684), 440–442. Retrieved from http://202.121.182.16/Course/slides2012/NetSci-2012-7.pdf doi: 10.1038/30918

Waung, M., Hymes, R. W., & Beatty, J. E. (2014). The Effects of Video and Paper Resumes on Assessments of Personality, Applied Social Skills, Mental Capability, and Resume Outcomes. *Basic and Applied Social Psychology*, *36*(3), 238–251. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/01973533.2014.894477 doi: 10.1080/01973533.2014.894477

Youyou, W., Kosinski, M., & Stillwell, D. (2015, 1). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(4), 1036–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25583507http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4313801 doi: 10.1073/pnas.1418680112