

# AI Applications to Personnel Selection

Illustrations, best practices and  
future perspectives -

03-2024

Andrei ION (shl.ro)



**SHL.**

People Science. People Answers.

© 2022 SHL and its affiliates. All rights reserved.

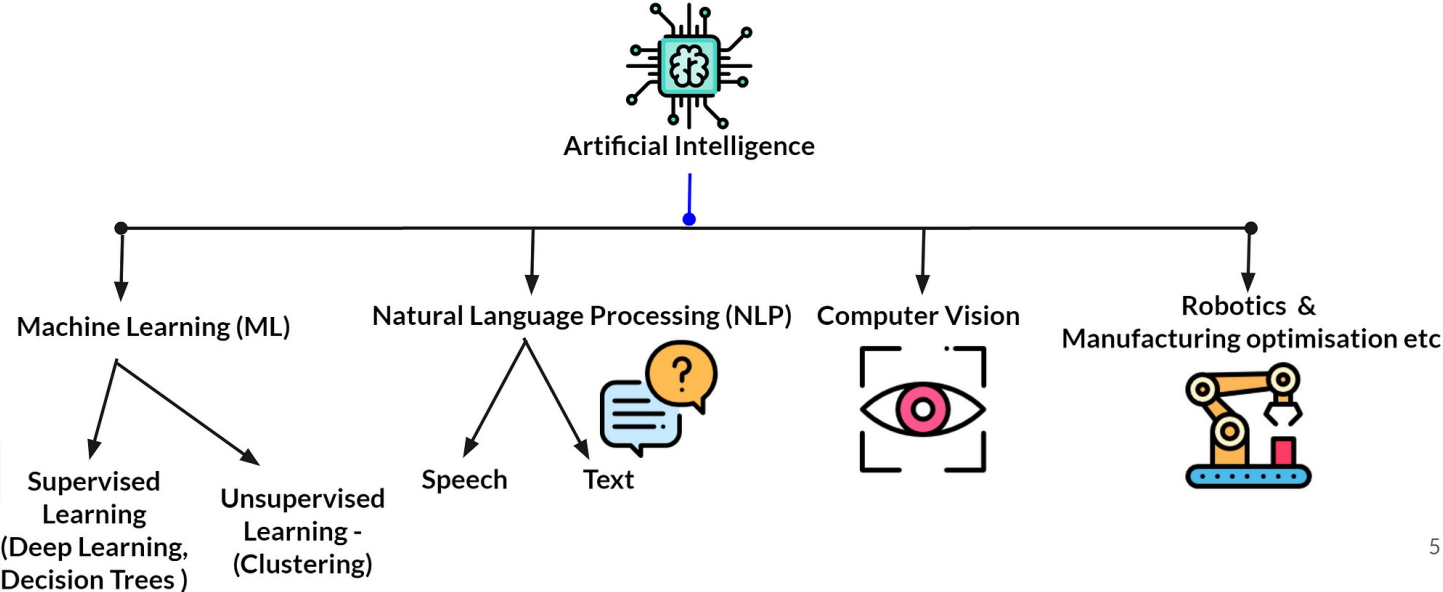
# Topics

At the end of this session you will:

- What is AI applied personnel selection/assessment?!
- Be able to interpret the findings of the Unlocking Potential Report
- Have a consistent approach to developing High Potential employees
- Understand through the Development Action Planner Report how to develop participants not identified as High Potential
- Be capable of organizing personal development planning

# Overview of AI disciplines

- szadasda



# What is 'AI'?

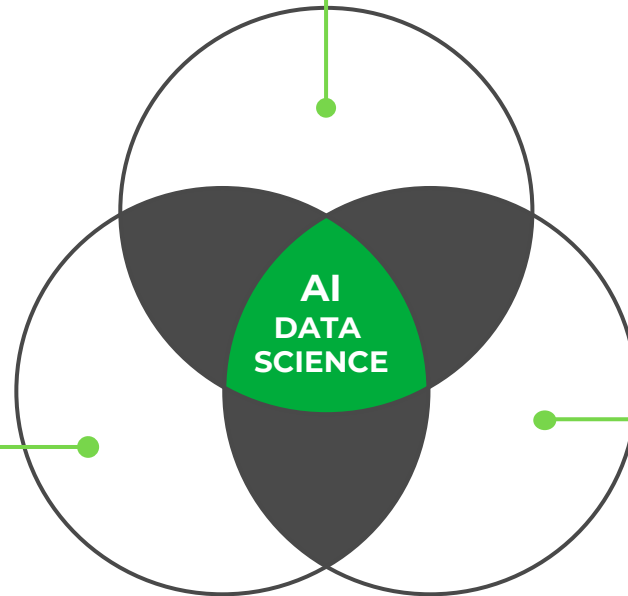
- Artificial Intelligence is a general term for any algorithm or computer program that attempts to simulate human-like intelligence or judgment.
- The definition of AI has evolved over time and tends to vary across fields. However, most definitions include the idea that AI is an effort to replicate tasks and processes, with computers, that are normally thought to require human intelligence. In other words, AI refers to attempts to make machines act intelligently.
- AI applications use machine learning or deep learning, which utilize quantitative models designed to learn patterns from observed data and then apply that information to new scenarios (i.e., new data). For example, Amazon uses machine learning to make future purchase suggestions based on a user's past purchases.
- Many AI applications today are designed to utilize Natural Language Processing (NLP), which enables computer algorithms to parse and extract meaning from natural language (e.g., written or spoken text).

# What is 'AI'?

Interdisciplinary field that applies a wide range of techniques to data for myriad purposes across many domains.

## B. Machine learning (ML)

ML refers to techniques that “learn” patterns in data to make predictions or summarize or score the data



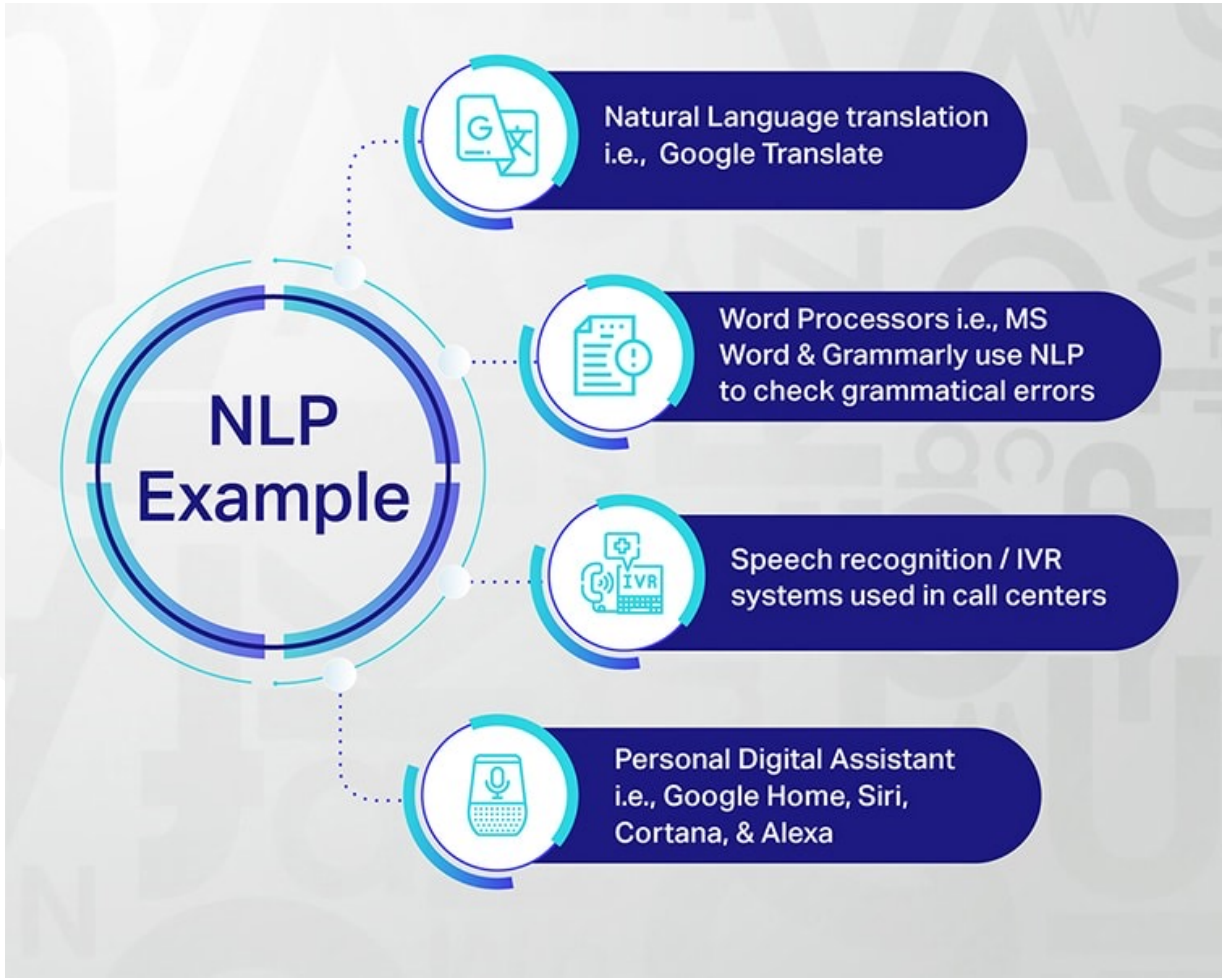
## A. Natural Language Processing (NLP)

specialized techniques for analyzing text data, including both the words used and the relationships among words.

## C. Deep Learning

(DL) is a highly complex type of ML that can be used to analyze any type of data (numeric or text). Its use of neural networks and transformers that have many internal layers of analysis make the mathematical operations not possible to examine directly, thus leading to the “black box” characterization.

# What is 'AI'?



# What is 'AI'?

- NLP examples



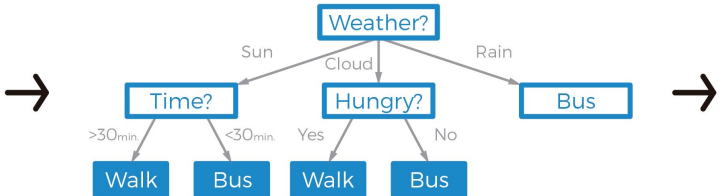
# What is 'AI'?

B. Machine Learning vs. C. Deep Learning

## Machine Learning



Input



Decision tree

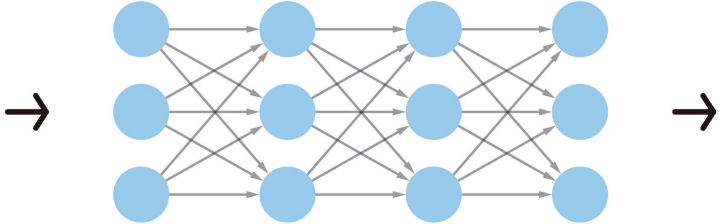


Output

## Deep Learning



Input



Feature extraction + Classification



Output

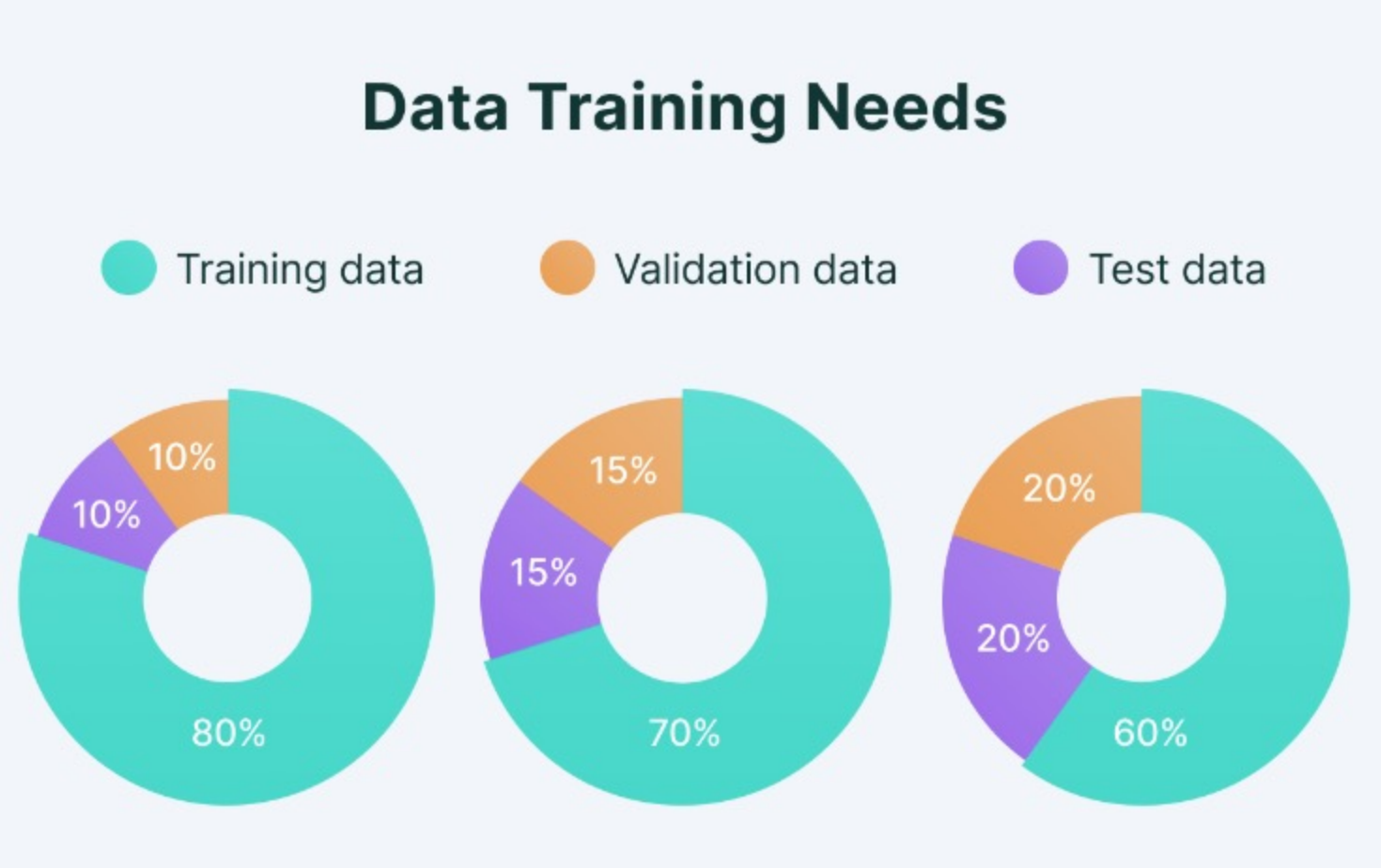


# What is 'AI'?

- Basic operating logic
- To learn and apply patterns in data, ML (either supervised or unsupervised) makes use of what are known as **features**.
- Features are quantifiable properties of a phenomenon being observed that are present in the data. In statistical models, features are sometimes referred to as “independent variables” or simply “**predictors**”.
- AI algorithms find **patterns** in the relationships between **predictors** and an **outcome variable** (referred to as a criterion in I/O Psychology).
- The dataset from which the ML algorithm first identifies patterns amongst features and outcome variables is referred to as the training dataset, as this is the dataset which “trains” the AI.

# What is 'AI'?

- How to train an ML algorithm?! Three stages...



# What is 'AI'?

- Training phase: are collections of examples or samples that are used to 'teach' or 'train the machine learning model. The model uses a training data set to understand the **patterns** and **relationships** within the data, thereby **learning to make predictions** or decisions without being explicitly programmed to perform a specific task.
- **Validation phase:** employing different samples drawn from the dataset to evaluate trained ML models (obtained during the training phase). It is still possible to tune and control the model at this stage. Working on validation data is used to assess the model performance and fine-tune the parameters of the model. This becomes an iterative process.
- **Testing phase:** test data set is a separate sample, an **unseen dataset**, to provide an unbiased final evaluation of a model fit. The inputs in the test data are similar to the previous stages but not the same data.

# What is 'AI'?

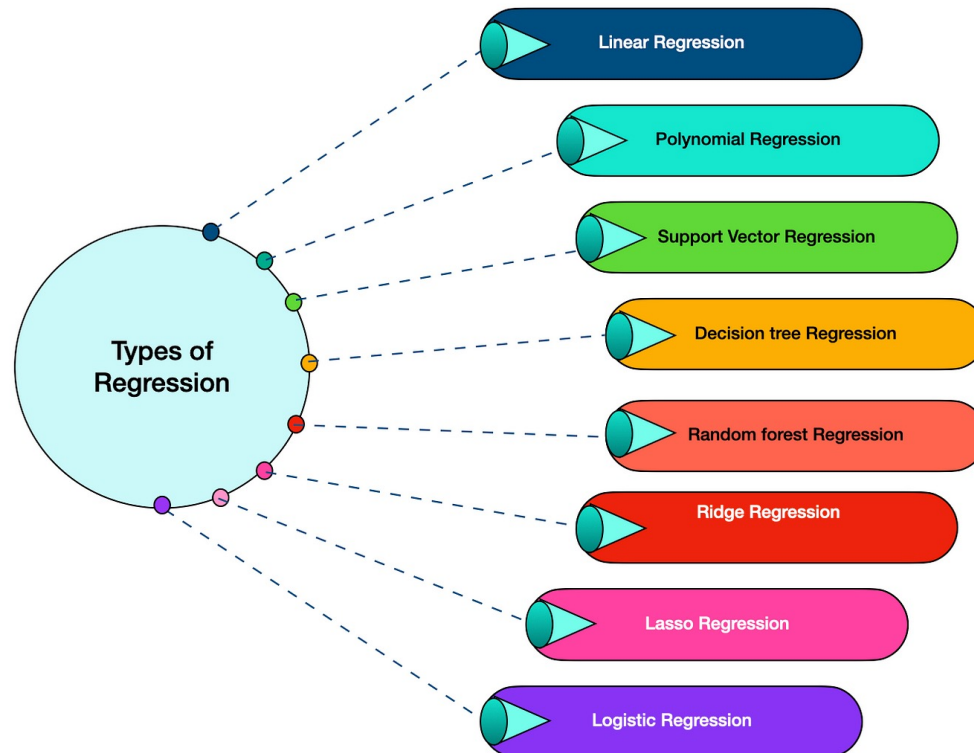
- Training phase: are collections of examples or samples that are used to 'teach' or 'train the machine learning model. The model uses a training data set to understand the **patterns** and **relationships** within the data, thereby **learning to make predictions** or decisions without being explicitly programmed to perform a specific task.
- **Validation phase:** employing different samples drawn from the dataset to evaluate trained ML models (obtained during the training phase). It is still possible to tune and control the model at this stage. Working on validation data is used to assess the model performance and fine-tune the parameters of the model. This becomes an iterative process.
- **Testing phase:** test data set is a separate sample, an **unseen dataset**, to provide an unbiased final evaluation of a model fit. The inputs in the test data are similar to the previous stages but not the same data.

# What is 'AI'?

- For example, in developing an AI-based video interview assessment, certain **features** from the recorded interview – such as words spoken or facial expressions – might be used to predict **subsequent performance** on the job for those who were hired.
- In this scenario the AI application would identify **patterns** between the **video features** and **job performance** in one dataset (the training data), and then attempt to apply those same patterns to another dataset (test or holdout sample).
- If the same patterns do not apply to the test sample, then the AI may have identified the wrong patterns in the training data, and therefore another attempt at learning from the training data could be required.
- This process is known as cross validation, and it is a very important concept for the development of any assessment, but it is particularly important for developing and validating an AI-based assessment.

# What is 'AI'?

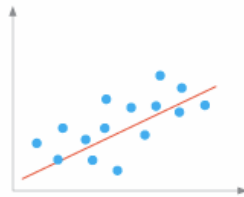
- Prediction is 'optimized' by employing different approaches that maximize prediction – typically, no constraints are imposed on the input data
- Any piece of data can become a potential **predictor**



@arunp77

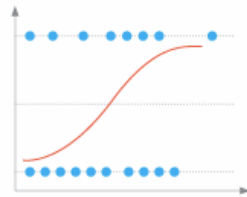
# What is 'AI'?

## 5 types of regression



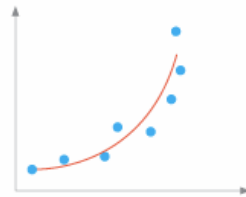
### Linear regression

Predicts a continuous output by modeling a straight-line relationship between input features and target variables, such as estimating the impact of price changes on demand.



### Logistic regression

Models the probability of binary outcomes, such as predicting customer churn; commonly used in classification tasks.



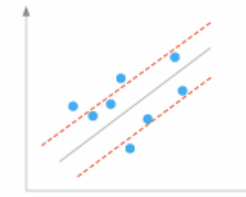
### Polynomial regression

Captures nonlinear relationships, such as estimating the impact of ad spending on sales, by fitting a polynomial curve to data points.



### Time series regression

Predicts future values in a time-dependent data set; often employed to forecast future values based on past observations, as seen in stock market analysis.



### Support vector regression

Approximates a continuous function by identifying a hyperplane that best represents the data's structure; valuable in various applications, including financial market prediction.

# Employing 'AI' in Talent Acquisition



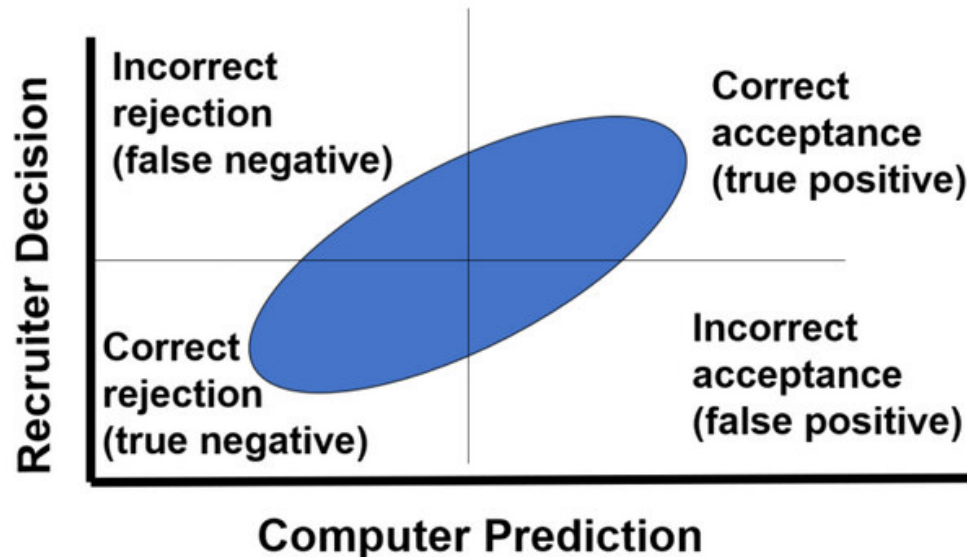
# Employing 'AI' in TA – Supervised ML

- Supervised machine learning refers to when there is a criterion (e.g., job performance, application scores, or selection decisions by the organization) used to create the model (i.e., select variables to include).
- In Data Science, such data are called “labeled data” because each case has a score or decision associated with it that we can use to train the model to predict and from which we can create weights to apply to score future data that are unlabeled.
- The criterion (label) can be either a continuous score or decision (e.g., passed, rejected, withdrew).

# Employing 'AI' in TA – Supervised ML

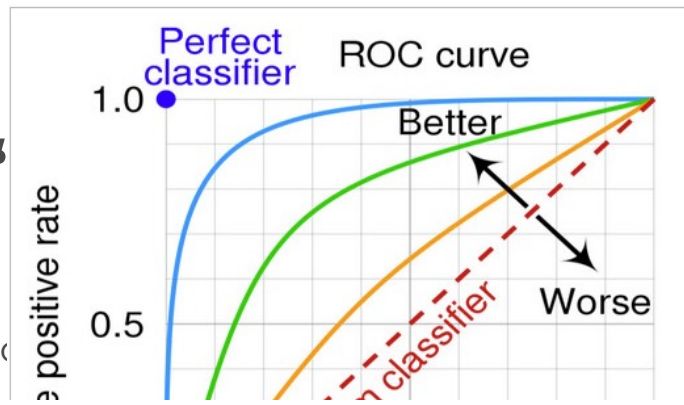
- Models built to predict a decision are often called classification models.
- Researchers using classification will usually evaluate validity, which they might describe as accuracy, in terms of the agreement of the decisions between the computer model prediction and the actual decision (“label”).

## Correct and Incorrect Selection Decisions



# Employing

- These consequences to evaluate accuracy
- Different estimations derived
- One su



# sed ML

ed into a specific metric

**AUC** stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

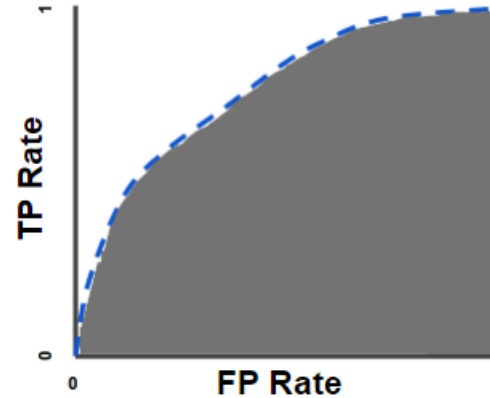
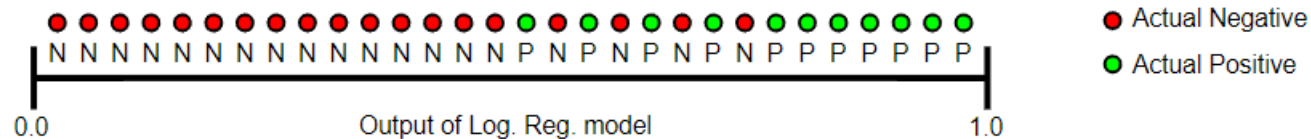


Figure 5. AUC (Area under the ROC Curve).

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. For example, given the following examples, which are arranged from left to right in ascending order of logistic regression predictions:



# Employing 'AI' in TA – Unsupervised ML

- Not all data are labeled (e.g., there is no criterion), however. In these instances, we can use unsupervised machine learning to uncover patterns within the data to summarize it (e.g., identify the topics) or to create measures (e.g., scores) that may be explored for their predictiveness of outcomes in the future.
- Unsupervised machine learning helps us solve problems through dimensionality reduction to cluster data in meaningful ways.
- The term “topic modeling” may be familiar and refers to using unsupervised machine learning on text data where the researcher determines an optimal number of topics based on how the model fits the unstructured text data using metrics such as coherence scores, as well as interpretability of the topics (Valtonen et al., 2022).

# Employing 'AI' in TA (Campion & Campion, 2023)

- Applying AI to talent acquisition processes

Application in TA	Data types	Potentially applicable ML techniques to create model	Pros	Cons
<b>1. Scoring resumes and employment applications (or any candidate-produced string/text)</b>	Numeric and text	Natural language processing (NLP) to analyze the text data and create numeric scores, and then combine with numeric data in applications using a wide range of supervised ML to predict selection scores and/or decisions.	Cost and time saving High volumes	Limited by quality of criterion data  Algorithm interpretation (avoid "black box").
<b>2. Scoring constructed responses to assessments (e.g., interviews, write-in test answers – in-trays etc.)</b>	Primarily text	NLP to analyze the text to identify the content, sometimes using unsupervised ML, to create numeric scores to assess candidates, or to predict other outcomes (e.g., interviewer ratings) using supervised ML (e.g., Koenig et al., 2023, Studies 2–4).	Scoring unstructured responses  Equal consideration to all candidates (unlike typical human review).	Requires criterion data to train models to predict  Algorithm interpretation (avoid "black box").
<b>3. Combining/aggregating scores to increase prediction</b>	Numeric and text	Wide range of supervised ML to optimally predict outcomes (e.g., job performance) and may include many variables and curvilinear relationships (e.g., Koenig et al., 2023, Study 5; Landers et al., 2023).	May increase prediction compared to regression, especially when sample/parameter (n/k) ratio is low.	Increase may be small in large samples given the added complexity.
<b>4. Combining scores to reduce subgroup differences (avoiding bias and adverse impact)</b>	Numeric and text	Pareto Optimal or similar ML, or make adjustments to analysis or data, to increase multiple outcomes simultaneously such as diversity and performance.	May reduce adverse impact without much loss to validity.	May require smaller weights on cognitive predictors, capitalize on chance with small samples, make little improvement, reduce validity, and create prediction bias.

# Employing 'AI' in TA (Campion & Campion, 2023)

- Applying AI to talent acquisition processes (continued...)

Application in TA	Data types	Potentially applicable ML techniques to create model	Pros	Cons
<b>5. Creating test contents (items, response options etc.)</b>	Text from existing questions to create model or use existing pre-trained language models	NLP (transformers using neural networks) to learn word patterns in past questions to create models that can produce similar questions or be used across similar questions or to use available models pre-trained on large public language data sets.	Saves time and effort to write new questions  Can be tuned to produce variety of items.	- requires large training samples;  - untuned models may produce highly similar items; requires researcher judgment
<b>6. Analyzing jobs to determine requirements</b>	Aggregating text from job descriptions or job analysis data.	NLP to extract content to create scores used to predict job analysis ratings with supervised ML.	Saves time in job analysis, especially for low volume applications.	- dependent on having accurate job description  - may be simpler methods if jobs can be linked to external frameworks (e.g., O*NET, ESCO).
<b>7. Inferring skills and personality from narrative application information</b>	Primarily text	Theoretically created closed-word dictionaries or NLP-created open-word dictionaries to score skill or personality in text data usually collected for other purposes to create scores (e.g., letters, statements of interest, responses to interviews or questions on applications, etc.).	Technically simple to use, many dictionaries available, longest history in I-O research compared to other ML, and can be used in many contexts.	- may not be measures of constructs of interest available, evidence of construct validity required; -  - may predict less well than more sophisticated ML.

# Conclusions

- First, there are many situations in which ML may enhance personnel selection. These include prescreening in addition to primary selection procedures; scoring narrative as opposed to numeric data; scoring constructed responses (e.g., write-in comments) as opposed to structured responses (e.g., multiple choice); making tradeoffs and maximizing prediction; reducing subgroup differences; creating test questions; and deriving job requirements.
- Second, there are many potential pros, but also some meaningful potential cons. For example, ML will increase efficiency in large-scale applications, but may not in small scale applications enough to justify the increased complexity, which is a tradeoff in all customized selection systems. Moreover, the prediction improvement of these procedures may be small, especially in large samples compared to traditional or more well-known procedures like regression. The value of ML may depend more on how it can help score data rather than increase prediction, such as text data and constructed responses. The reduction in subgroup differences may or may not be possible, and could actually create prediction bias, but more needs to be known on this front.
- Third, there might also be other uses for ML, such as helping create assessment items or making other tasks easier like job analysis. Fourth, some applications like word dictionaries are actually simple ML that researchers are likely already familiar with, and dictionaries are easily available to novices. There is much yet to be learned and these are just illustrations to stimulate future research

# Deploying 'AI' to Talent Assessment

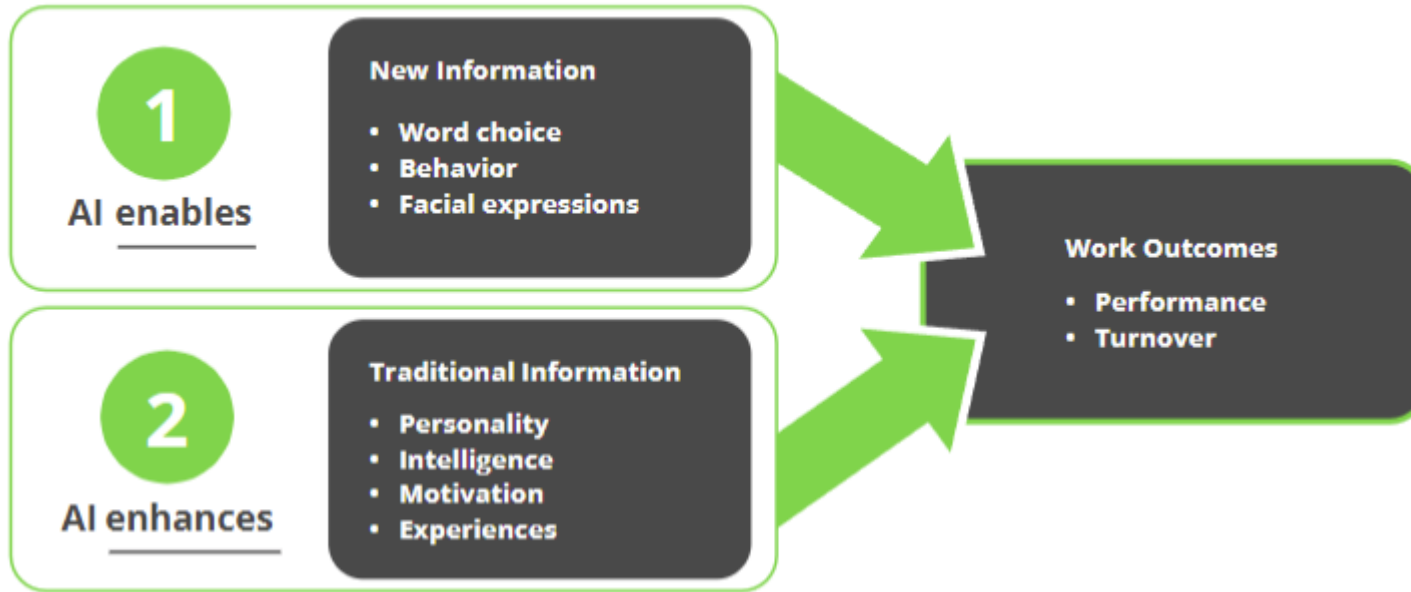


# Employing 'AI' in Talent Assessment

- ML models can be taught to predict:
  - Recruiters' decisions *OR*
  - Job-relevant outcomes (e.g., job-performance, engagement, job satisfaction, turnover intent etc.) *OR*
  - Job-relevant variables (e.g., extraversion, grit, leadership, teamwork etc.).

# Employing 'AI' in Talent Assessment

- HOW?!



# The Smarter Way to Interview Talent Remotely.

Meet thousands of candidates in minutes and wow your shortlist with the smartest interview technology.

SHL Smart Interview Platform

SHL.



# Smart Meet – Automated Scoring – Development Overview – **Verbal Features**

#1: Voice  
to text  
module

#2: NLP  
module

#3: ML  
Analysis

#4: Inter-  
pretability  
phase

# Smart Meet – #1 Voice to text phase

- To generate predictive features used to train AI models, the recording of each interview response must be transcribed to text.
- We use **Microsoft Speech to Text automatic speech recognition (ASR)** service to produce these **text transcripts**. It is also possible to customize acoustic, language, or pronunciation models, which may be useful for dealing with industry-specific vocabulary.
- No speech recognition software has perfect accuracy, as there are always some errors in transcribing audio. The Word Error Rate(WER) is the percentage of incorrect words relative to total words in a sample of transcripts. Incorrect words can be of the following form:
  - Insertion – Words that were not said but are added to the transcript (e.g., “I would ask the manager” becomes “I would ask the man manager”).
  - Deletion – Words that were undetected and do not appear in the transcript (e.g., “I would ask the manager” becomes “I would ask manager”).
  - Substitution – Words that were incorrectly substituted from the audio to the transcript (e.g., “I would ask the manager” becomes “I would task the manager”).
- We conducted a study to estimate the WER for a sample of interview responses. We had humans transcribe the responses and compared the human-generated transcripts to the automated transcript. We found that this service has a WER of less than 15% for responses recorded in American English (**potential issue**). This level of accuracy is usually adequate for developing automatic scoring because most words in the response are not key features driving the score.

# Smart Meet – #2 NLP Phase

- A variety of **text features** (think predictors) are used across our models and each model is fit using the most predictive and interpretable features. The approaches we generally use to generate features include:
  - Bag of Words (BOW): A set of words representing the vocabulary in a corpus of text. We use the BOW feature counts of unigrams, bigrams, trigrams, and tetragrams. All the words are stemmed and stop words are removed. Each response has a large vector, with each element representing an *ngram* (n meaning the number of words in the feature). The value is the number of times the *ngram* appeared in the response. We divide the term frequency counts by Inverse Document Frequency (TF/IDF) to prioritize the most important words/phrases. This way, terms that occur in a high proportion of documents receive less weight. These features help identify typical phrases, which may signify certain meaning, that are signatures of the correct response. Given that a large number of such phrases are selected, it keeps the model generalizable.
  - Word embedding: A pre-trained set of word associations developed from a corpus using deep learning. We use the **Transformer architecture** (used also in Chat GPT), which is the industry standard. We use embeddings to capture the meaning of the words and sentences used, rather than just the word itself. Semantically similar words have similar embeddings. Here, we project the high dimensional word space (with each word as a single dimension) to a low dimension continuous vector space, which represents the meaning space.

# Smart Meet – #3 ML Analysis phase

- We employ multiple methods in creating algorithms to ensure that we mimic human raters as accurately as possible. The general goal is to **use the simplest model that gives acceptable correlation with expert ratings**, because simpler models are easier to understand and more transparent than more complex models (avoiding black box effects).
  - A number of different ML approaches are explored for each model, including off-the-shelf traditional ML approaches like Support Vector Machines (SVM; e.g., Schölkopf, Burges, & Smoia, 1999), random forest (e.g., Biau & Scornet, 2016), and ensemble approaches consisting of more than one model type (e.g., Zhang & Ma, 2012). In instances where these approaches do not meet minimum accuracy thresholds (typically  $r \geq .60$ ) (all of these are supervised ML) – when supervised fails -> deep learning approaches (Vargas, Mosavi, & Ruiz, 2017) are employed. Artificial neural networks to deliver nonlinear processing in multiple layers, which involves the current layer in an algorithm taking the output of the previous layer as an input (Vargas et al., 2017). The primary deep learning technique we employ is BERT (Bidirectional Encoder Representations from Transformers; Devlin, Chang, Lee, & Toutanova, 2018), which allows a deeper sense of language context and flow than do models that are restricted to examining text in a single direction.
  - ML models are derived from a training sample rather than all the data available because a trained model is only considered effective to the extent that the algorithm predicts the criterion (i.e., human ratings) in an independent sample, or test set (Tippins et al., 2021).
  - Then, we apply a k-fold cross-validation procedure using a sample split where 80% of the sample is used for model training and 20% is held out for the test set. In this process, we randomly split the sample into five equally sized test sets, each representing 20% of the cases from the total sample. We develop the model on the data in the 80% of the sample that is not in the first test set, then generate predicted scores in the test set.

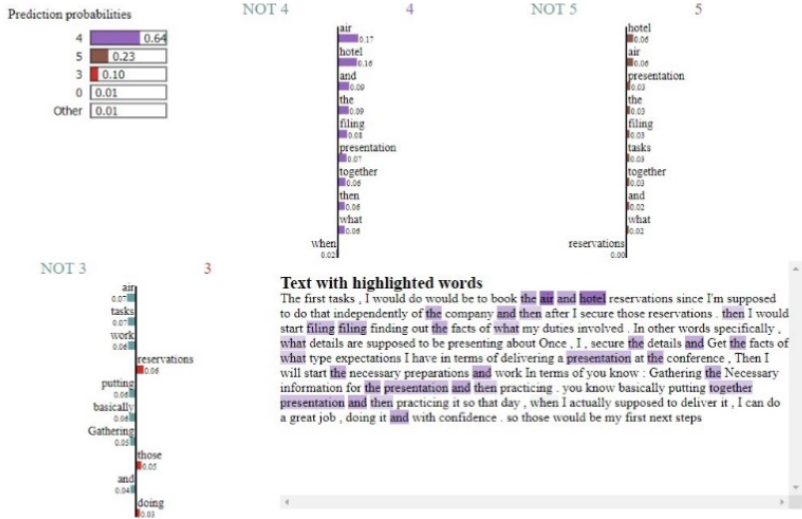
# Smart Meet – #4 Interpretability

- AI models have historically had a black box element to them, in that the process the model uses to make decisions is not transparent to the developer or the user. This is a problem with AI in general, but especially when applied to personnel selection assessments because it is contrary to the principles articulated by the Society for Industrial and Organizational Psychology (SIOP) to guide the use of assessment procedures (Principles; SIOP, 2018).
  - Saliency Maps** - Saliency maps estimate importance by using derivatives of the output with respect to each feature (i.e., how a small change in the feature changes the output) – basically, this assigns a probability of being in one outcome category for each classifier/predictor (in this case – words).
  - Leave one out** - The leave one out method assesses feature importance by removing one feature (e.g., a word) at a time from the response and observing how the level of prediction changes.



Table 8. Example of Leave One Out method.

Candidate Response	Overall Score	Removed
There is an express 2-day delivery available for an additional \$10.	0.95	-
<b>There</b> is an express 2-day delivery available for an additional \$10.	0.95	There
There <b>is an</b> express 2-day delivery available for an additional \$10.	0.94	is an
There is an <b>express</b> 2-day delivery available for an additional \$10.	0.92	express
There is an express 2- <b>day</b> delivery available for an additional \$10.	0.90	2-day
There is an express 2-day <b>delivery</b> available for an additional \$10.	0.86	delivery
There is an express 2-day delivery <b>available</b> for an additional \$10.	0.91	available
There is an express 2-day delivery available <b>for</b> an additional \$10.	0.94	for
There is an express 2-day delivery available <b>for an</b> additional \$10.	0.95	an
There is an express 2-day delivery available for an <b>additional</b> \$10.	0.91	additional
There is an express 2-day delivery available for an additional <b>\$10</b> .	0.89	\$10





# Smart Meet – #5 Validation

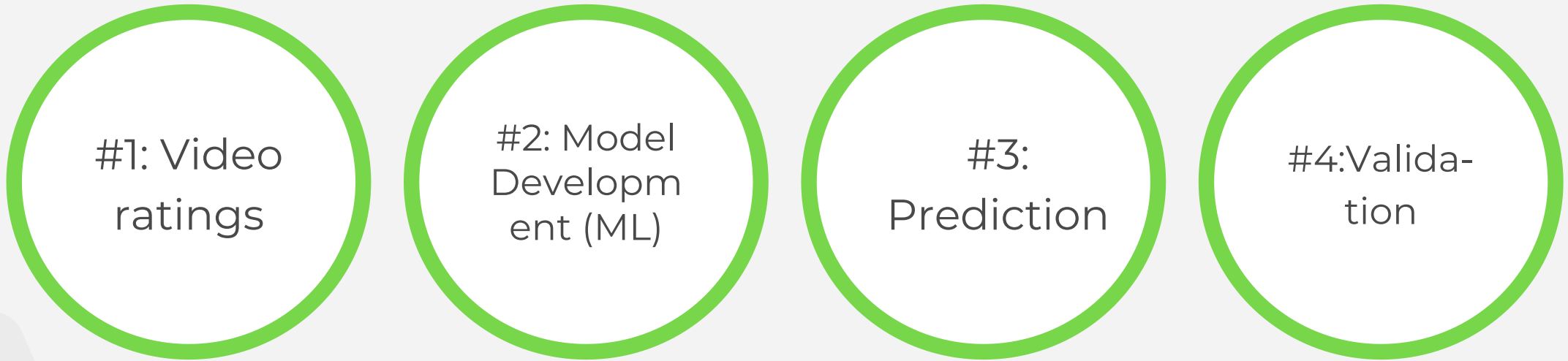
**Table 12.** AI-generated response scores descriptive statistics and standardized mean subgroup differences for gender subgroups.

SID	UCF Component	Female			Male			<i>t</i>	<i>d</i>
		Mean	<i>N</i>	SD	Mean	<i>N</i>	SD		
157	Monitors Performance	2.83	266	1.04	2.74	193	0.96	0.96	-0.09
197	Makes Considered Decisions	3.27	239	0.86	3.37	205	0.82	1.33	0.13
371	Strives to Achieve	2.98	93	1.02	2.74	96	1.07	1.57	N/A
481	Takes Responsibility	2.96	261	0.99	3.03	225	0.92	0.76	0.07
763	Thinks Broadly	2.87	249	0.79	2.53	231	0.99	4.19***	-0.39
163	Performs Repetitive Tasks	2.62	155	1.29	2.23	136	1.19	2.72**	N/A
165	Attends Work Reliably	2.88	176	0.56	2.70	118	0.51	2.73**	N/A
167	Copes with Uncertainty	2.19	177	1.34	2.23	116	1.44	0.22	N/A
171	Persuades Others	2.78	255	1.22	2.58	178	1.27	1.61	N/A
615	Speaks Clearly	2.82	184	1.37	2.63	148	1.45	1.17	N/A

\**p* < .05. \*\**p* < .01. \*\*\**p* < .001.

# Smart Meet – Automated Scoring – Development

## Overview – **Non-verbal Features**



# Smart Meet – #1 Video ratings

- SHL developed a measure of the extent to which an interview respondent is perceived as **confident, calm, engaging**, and has a **positive disposition** in a face-to-face conversation.
- These scores are derived by an **automatic analysis of facial expressions** (non-verbal) and **voice tonality** (para-verbal) of the candidates.
  - First, we rate videos instead of still images. The videos have a duration of at least 30 seconds. This gives the raters ample time to observe the candidates before making a judgment. Moreover, a candidate's final rating is based on a set of videos (at least six) rather than a single video. Also, our rating parameters are expressed behaviors rather than internal states.
  - All the responses were graded by human raters. Our raters have experience working as HR recruiters, soft skills trainers, or possessed a background in industrial and organizational psychology.
  - Finally, each video was rated by at least five independent raters. The mean interrater correlation for the nonverbal scales were .62 for Composure, .62 for Confidence, and .67 for Engagement. The overall score is the average of the individual rating by the raters for each social skill, which had a mean interrater correlation of .69.

# Smart Meet – #2 Model Development – ML

- Nonverbal scores were derived by an automatic analysis of facial expressions and voice tonality of the candidates.
- The video features included various statistical functionals such as mean, standard deviation, distance between extremes, etc., on the Facial Action Unit (AU) intensities over all the frames in a video. We used a variety of Action Units and Pose features.
- The voice features included statistical functionals like mean, standard deviation, moments, peaks, quartiles etc., over speech-related features like pitch, loudness, signal energy, voice quality (jitter, shimmer), MFCC, and spectral shape descriptors.
- We developed machine-learning models that used these features as input and human ratings as the criteria.

# Smart Meet – #3 Prediction – real data

- We conducted a validation study on a sample of 810 candidates to evaluate the extent to which automatic scores predict human ratings. A panel of nine experts graded the candidates on Composure, Confidence, and Engagement. They were provided with the relevant rubrics to grade the responses. The consensus rating among the experts was used as the final human rating. We used the bootstrap method to train the models in multiple stratified samples of 80% of the total sample. Correlations between scores generated by the model and the consensus human scores were computed in the test samples, which were 20% of the total sample. This process was repeated 20 times.

**Table 19.** Correlations between machine-generated nonverbal scale scores and expert ratings.

Region	Composure	Confidence	Engagement
Overall	.65	.68	.68
US/Europe	.65	.67	.67
India	.63	.70	.68

*Note.* All correlations significant at  $p < .001$ .

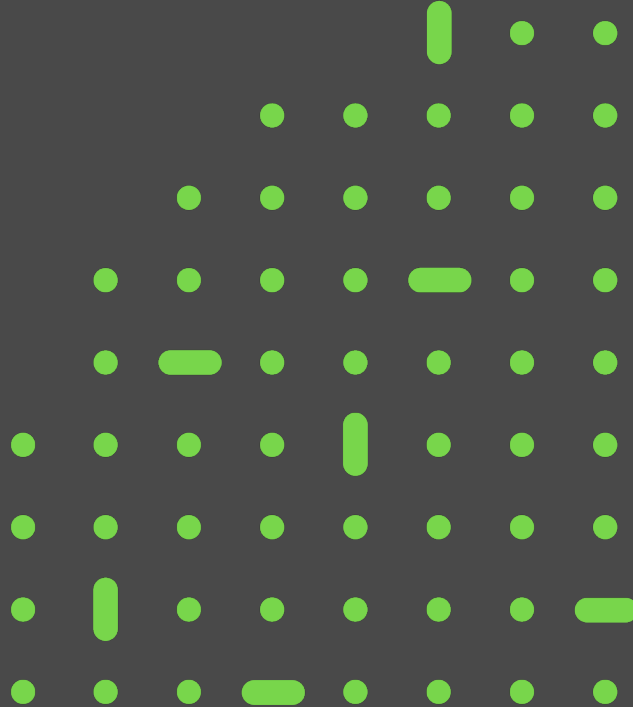
# Smart Meet – #3 Prediction – real data

- Several real-life validation studies were conducted, plus the bias and adverse impact analyses.

**Table 22.** Probability of candidate being selected at varying levels of Workplace Skills and Nonverbal Scores: Study 2

	Workplace Skills	Nonverbal	Log (Select/Reject)	Odds (Select/Reject)	Probability of Selection
Tab	0	0	-9.75	0.00	0.000
	20	0	-5.95	0.00	0.003
	40	0	-2.15	0.12	0.104
Co	0	20	-9.35	0.00	0.000
On	20	20	-5.55	0.00	0.004
Co	40	20	-1.75	0.17	0.148
Co	0	40	-8.95	0.00	0.000
Er	20	40	-5.15	0.01	0.006
*p	40	40	-1.35	0.26	0.206

# MEET 'Smart Meet'



Rep


QUESTION 3

Your manager tells you that you must deliver a presentation at a conference the following week. You have not yet made any preparations for this presentation. The conference is in a different city, and you will have to fly there. You will have to make your own air and hotel reservations and the company will reimburse you later. How would you prepare for this trip? Which tasks would you address first?

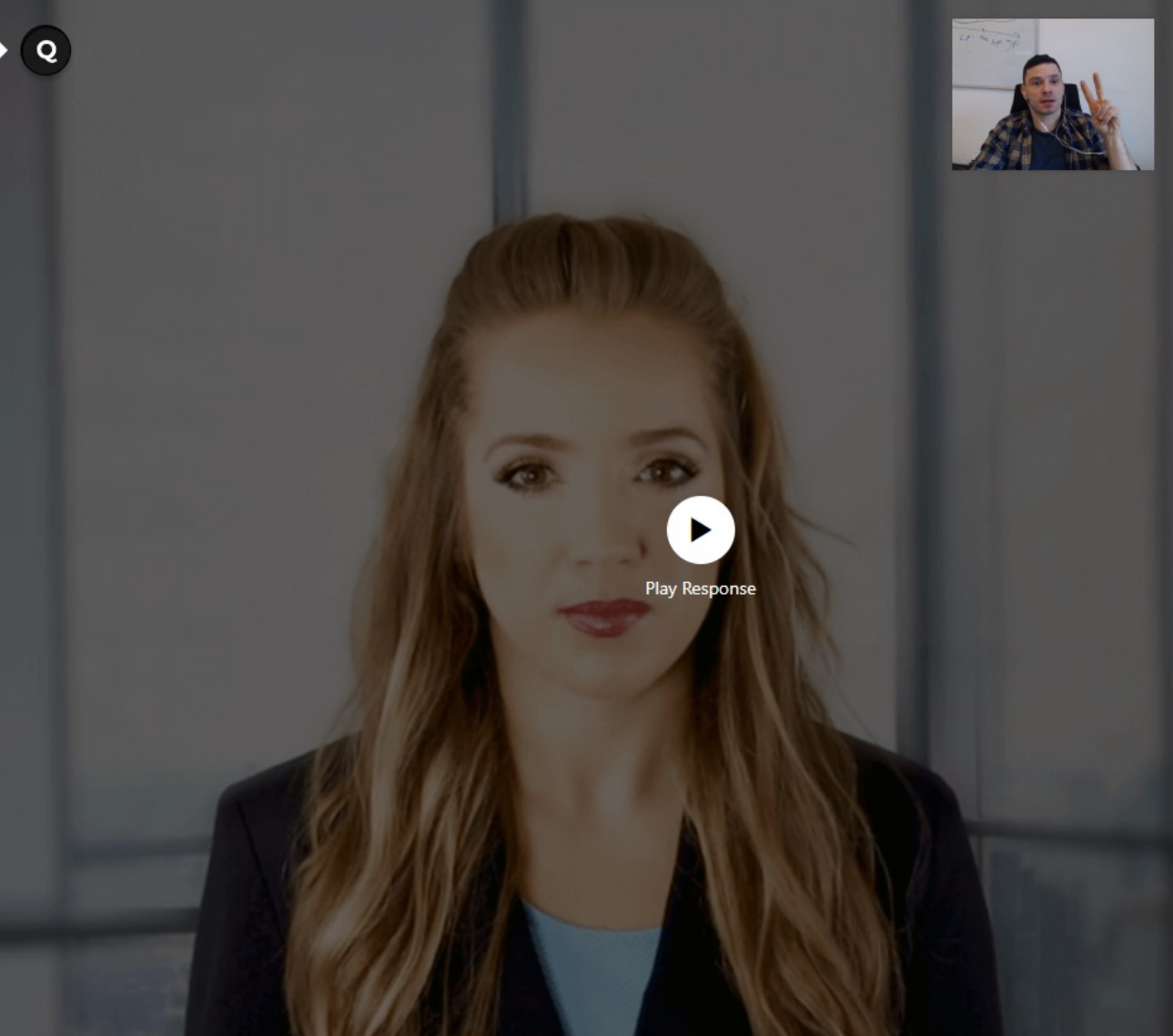





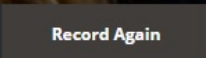
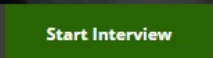
# Report Section 2: Development Activities


**QUESTION 1** < 

This is a sample question to check your device compatibility. It will not be included in your evaluation. Please read the following sentence in a loud and clear voice: "Hi! It's a great day today. I am ready to take the interview. Thanks"



Play Response



## Andrei Ion

Test ID: 270270544141995 | ✉ andrei.ion@fpse.unibuc.ro

Test Date: March 13, 2024

## Test SHL Romania

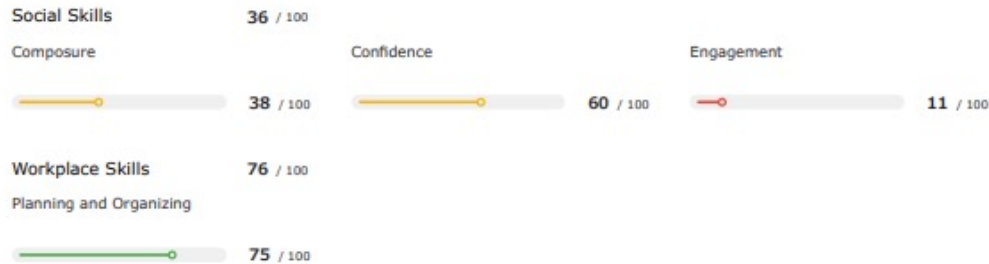
56 /100



Evaluators' score: Pending

## Test SHL Romania

56 / 100



## Score Interpretation

- **Evaluators' score:** Refers to human evaluators rating the candidate on a numeric scale (for example 1-5) from which an overall score is generated.
- **AI-based evaluation (#/100):** Scores generated by artificial intelligence are shown as ratings on a scale of 1-100.
- **Comparison score (percentile):** A score that's been compared against a group of other candidates (also known as a normed score). For example, a candidate in the 60th percentile has scored better than 60% of the people in the comparison group.
- **Absolute score (#/100):** A score based on the number of correct responses. For example, a score of 60/100 means the candidate answered 60% of the questions correctly.

The color coding in this report is as given below:

- Scores between 71 and 100
- Scores between 31 and 70
- Scores between 0 and 30

## 1 | Insights

## Test SHL Romania

56 / 100

Social Skills 36 / 100
 **Composure** 38 / 100

This competency refers to how calm and relaxed the candidate is.

The candidate exhibits a generally relaxed and calm manner. He occasionally appeared anxious during the interview but made efforts to control his reaction.

 **Confidence** 60 / 100

This competency refers to the candidate's confidence level and the extent of surety in her/his responses.

The candidate exhibited a certain level of confidence during the interview. But at times, he seemed uncertain of his responses.

 **Engagement** 11 / 100

This competency refers to candidate's positive emotion and enthusiasm.

The candidate hardly smiled and showed little positive emotion. He lacked enthusiasm during the interview.

Workplace Skills 76 / 100
 **Planning and Organizing** 75 / 100

The ability to prioritize and manage one's workflow while accounting for constraints in time and resources.

- The candidate is adept at planning and scheduling his workflow to ensure smooth execution of projects.
- He is able to accomplish tasks in a timely manner without compromising quality of output.

# Employing 'AI' in Talent Assessment

BEST PRACTICES

# Best practices



## Identify Data Requirements

Consider data minimization, quality, diversity, and security.



## Prioritize Transparency

Develop transparent AI - no "black box" algorithms.



## Design for Fairness

Build fairness into the assessment from the beginning.



## Rigorously Validate

Hold AI assessments to a high standard regarding validity evidence.



## Incorporate Human Oversight

No AI assessment should make decisions without human oversight.



## Disclose Intent

Notify when AI is being used and explain how it works.

# #1 – Data Quality

- Data quality
  - quality data is data that results from a measure that is valid and reliable – a measure that produces predictor data with a low level of error (i.e., low bias and low variance).
  - the AI scoring of asynchronous video interviews (e.g., words a candidate speaks are first transcribed into text) is a complex task. It is imperative that this task is handled with care and maintenance, of this nature, to ensure the reliability of transcription and scoring. For example, in asynchronous video interviews, it is imperative that the words a candidate speaks are accurately transcribed and scored. This is a complex task that requires the development, and ongoing maintenance, of high-quality transcription and scoring algorithms that meet acceptable accuracy and reliability standards.
  - data quality extends beyond the accuracy of the data measured or not related to the task. For example, data quality may not predict actual performance in a criterion that is not well represented by the algorithm. For example, a candidate's response may be produced by the algorithm, but it may not be a criterion that is not well represented by the algorithm.
  - the use of increasing asynchronous video interviews recorded/streaming data for accurate assessment. For example, increasing volume, garbled speech, and other factors may be unsuitable for further assessment.
- Data diversity – little is known about how cultural factors intervene in the roll-out of AI-based data interpretation.
- Data privacy and adaptation ...

**Identify Data Requirements: Best Practice**  
**Maintain a high standard regarding all aspects and considerations of the data that go into developing an AI assessment.**  
**Ensure compliance with existing legal standards (e.g., GDPR, AI Video Interview Act) and guidelines (e.g., SIOP Principles, European Commission’s Guidelines for Trustworthy AI) during the design, development, and operation of an AI assessment.**

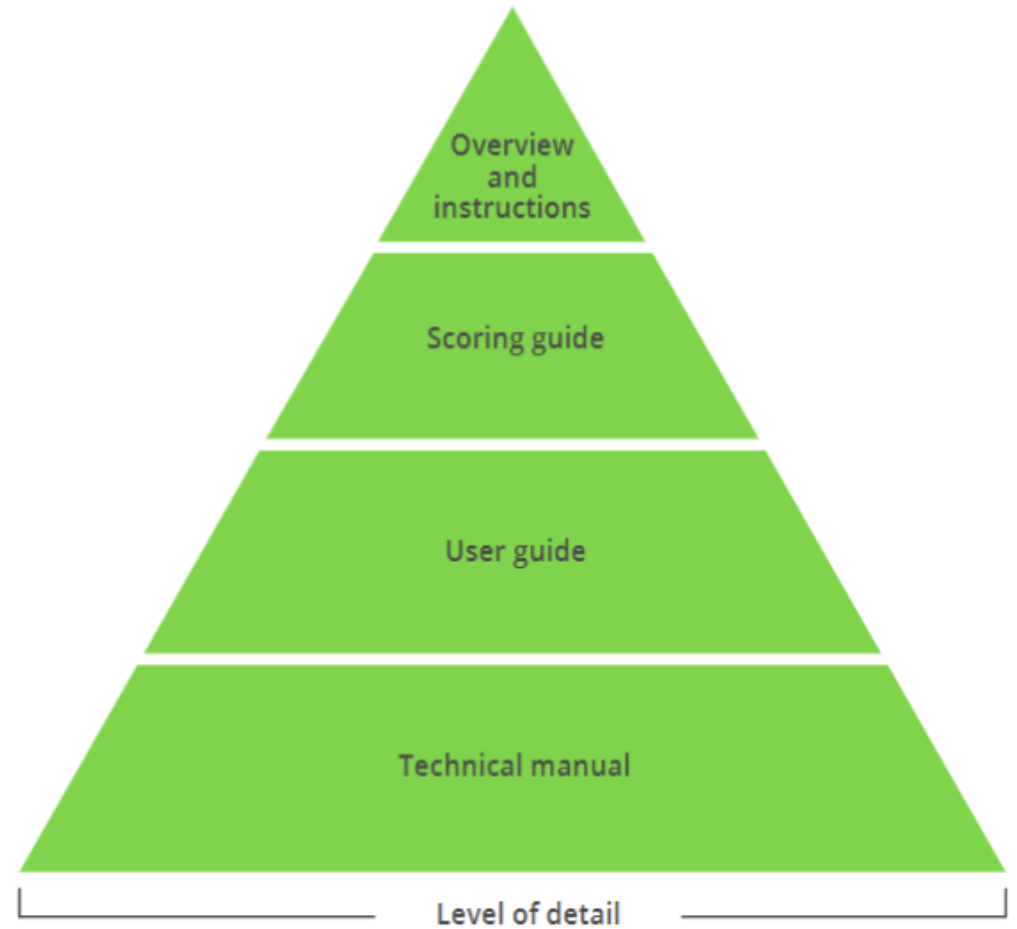
## #2 – Prioritize transparency

- Being able to explain how an AI assessment works, again
- Being able to explain how an AI assessment works, again
- Being able to explain how an AI assessment works, again
- Being able to explain how an AI assessment works, again
- Availability of information regarding how the features are combined to produce a score. The level of information provided should vary by the end user (e.g., hiring manager, candidate).

### Prioritize Transparency: Best Practice

Design, develop, and use AI assessments that are appropriately transparent. Transparency into an AI assessment may include a description of the features scored by the assessment, any constructs and/or characteristics that the features are related to, and some information regarding how the features are combined to produce a score. The level of information provided should vary by the end user (e.g., hiring manager, candidate).

Figure 2. Documenting the Transparency of an AI Assessment





# #3 Design for Fairness

- Equal treatment of all candidates in the selection process.
- Equal access to the constructs being measured by an assessment (i.e., “accessibility”)
- Hiring and selection processes that are non-discriminatory (i.e., without bias)
- Validity is “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA et al.,2014, p. 11).

## Design for Fairness: Best Practice

Begin the development of an AI assessment with fairness in mind. Take steps to proactively remove or reduce bias from the AI assessment during its design and development. Do not rely only on a single test of bias (e.g., the four-fifths rule) after the AI assessment has been developed.

# #4 Validation

- The validation of an AI assessment, as with any other form of assessment, requires the use of methods that are supported by current legal and professional standards. Four such methods of validation are:
  1. Content-related validation - A content-related validation strategy focuses on demonstrating that the content of the assessment (i.e., the features or constructs being assessed) is relevant to the work requirements of the target job. This typically involves input and judgment from SMEs. For example, a word processing assessment can be validated for an administrative assistant role via the consensus from a panel of SMEs that the operation of the word processing software, as measured by the assessment, is an important requirement of t
  2. Construct-related validation - This method focuses on demonstrating that the assessment accurately measures the construct it is intended to measure. This typically involves a SME's judgment of sufficient evidence.
  3. Criterion-related validation - This method focuses on demonstrating a quantitative relationship between assessment scores and a criterion (e.g., performance). This relationship is expressed via a statistical metric (e.g., validity coefficient). The strength of the relationship (expressed via a statistical metric) determines the validity of the assessment.
  4. Generalizing validity evidence - This method focuses on demonstrating that an assessment or AI algorithm is expected to perform similarly in a new context. Generalizing validity evidence typically involves the cross validation of an algorithm, in which the predictive accuracy of the algorithm is assessed in the test or holdout sample.

**Design for Fairness: Best Practice**  
Begin the development of an AI assessment with fairness in mind. Take steps to proactively remove or reduce bias from the AI assessment during its design and development. Do not rely only on a single test of bias (e.g., the four-fifths rule) after the AI assessment has been developed.



# #5 Human Oversight & #6 Disclosure

## **Incorporate Human Oversight: Best Practice**


Design AI assessments to have human oversight throughout its development and during its deployment. No AI system should make a final decision in a high-stakes situation (e.g., determining who to hire) without the possibility of human intervention.

## **Disclose Intent: Best Practice**

- Notify candidates that AI will be used to analyze responses
- Explain how the AI works (e.g., the constructs/ features measured - for example, through a privacy notice)
- Obtain consent to be assessed by AI where and when required
- Provide alternative assessment to those who do not provide consent

# ML vs. Traditional regression in selection decisions


## A simulation of the impacts of machine learning to combine psychometric employee selection system predictors on performance prediction, adverse impact, and number of dropped predictors




Richard N. Landers  Elena M. Auer, Lily Dunk, Markus Langer, Khue N. Tran

First published: 27 March 2023 | <https://doi.org/10.1111/peps.12587> | Citations: 1

We have no known conflicts of interest to disclose.

Author Note. The order of final three authors is alphabetical. EMA is now at BetterUp. ML is now at Fachbereich Psychologie, Philipps-Universität Marburg. The authors also thank Te Chi Cheng for their contributions.

 SECTIONS

 PDF  TOOLS  SHARE

### Abstract

We compare modern machine learning (MML) techniques to ordinary least squares (OLS) regression on out-of-sample (OOS) operational validity, adverse impact, and dropped predictor counts within a common selection scenario: the prediction of job performance from a battery of diverse psychometrically-validated tests. In total, scores from 1.2 billion validation study participants were simulated to describe outcomes across 31,752 combinations selection system design and scoring decisions. The most consistently valuable improvement from adopting MML over traditional regression was from dropping predictors rather than by improving prediction. On average, MML improved prediction of performance from psychometric scale composites only when the ratio of sample size to scale count was less than approximately 3, although algorithm choice, predictor count, and selection ratio affected outcomes as well. We also simulated the effects of design choices when combining item scores, which showed consistent, superior predictive accuracy for several MML algorithms, especially elastic net and random forest, over traditional regression. Given these results, we suggest the potential of machine learning for employee selection is unlikely to be realized in selection systems focusing on the combination of scale composites from previously validated psychometric tests. Instead, it will be realized in unconventional design scenarios, such as the use of individual items to make multiple trait inferences, or with novel data formats like text, image, audio, video, and behavioral traces. We therefore recommend researchers focus on the potential value of MML in future selection contexts rather than continuing to focus on the current value of MML in current selection contexts.

# ML > Human guided processes in generating JDs and in processing qualitative data

PERSONNEL  
PSYCHOLOGY



ORIGINAL ARTICLE | [Open Access](#) |

## Improving measurement and prediction in personnel selection through the application of machine learning

Nick Koenig, Scott Tonidandel , Isaac Thompson, Betsy Albritton, Farshad Koohifar, Georgi Yankov, Andrew Speer, Jay H. Hardy III, Carter Gibson, Chris Frost, Mengqiao Liu ... [See all authors](#)

First published: 14 July 2023 | <https://doi.org/10.1111/peps.12608> | Citations: 1

Papers in this article were submitted and evaluated individually in the traditional peer-review process by five independent reviewers, including regular editorial board members and a special board, judged to make a meaningful contribution after one or more rounds of revisions, accepted independently, and then combined into this thematic article. The co-editors of the special issue decided study ordering, which determined authorship ordering.

SECTIONS



PDF



TOOLS



SHARE

### Abstract

Machine learning (ML) is being widely adopted by organizations to assist in selecting personnel, commonly by scoring narrative information or by eliminating the inefficiencies of human scoring. This combined article presents six such efforts from operational selection systems in actual organizations. The findings show that ML can score narrative information collected from candidates either in writing or orally in response to assessment questions (called constructed response) as accurately and reliably as human judges, but much more efficiently, making such responses more feasible to include in personnel selection and often improving validity with little or no adverse impact. Moreover, algorithms can generalize across assessment questions, and algorithms can be created to predict multiple outcomes simultaneously (e.g., productivity and turnover). **ML has even been demonstrated to make job analysis more efficient by determining knowledge and skill requirements based on job descriptions.** Collectively, the studies in this article illustrate the likely major impact that ML will have on the practice and science of personnel selection from this point forward.

# ML does not safe-guard against adverse impact



ORIGINAL ARTICLE | Open Access |

## Reducing subgroup differences in personnel selection through the application of machine learning

Nan Zhang, Mo Wang, Heng Xu, Nick Koenig, Louis Hickman ✉, Jason Kuruzovich, Vincent Ng, Kofi Arhin, Danielle Wilson, Q. Chelsea Song, Chen Tang, Leo Alexander III, Yesuel Kim

First published: 01 June 2023 | <https://doi.org/10.1111/peps.12593> | Citations: 5

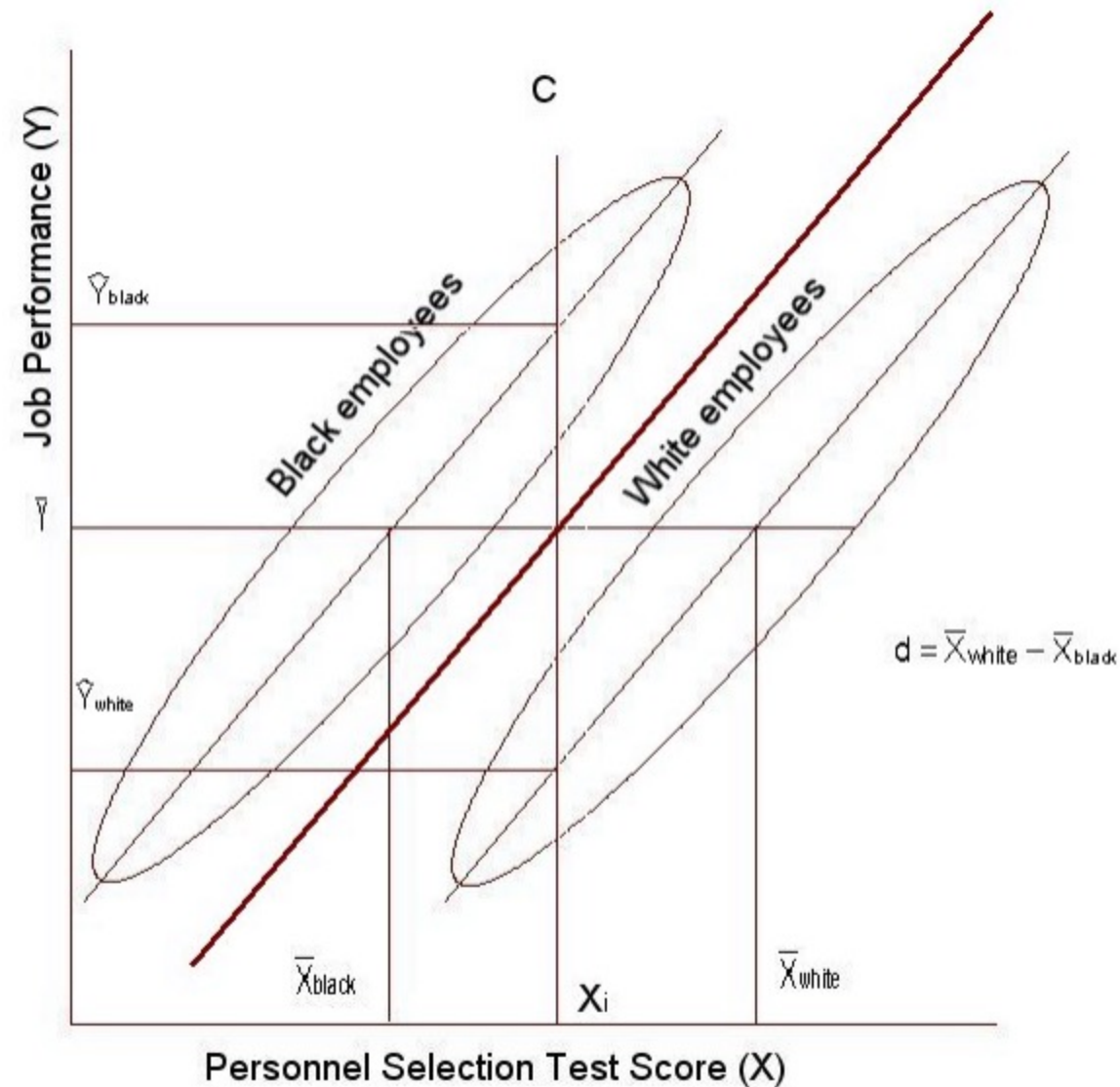
SECTIONS

PDF TOOLS SHARE

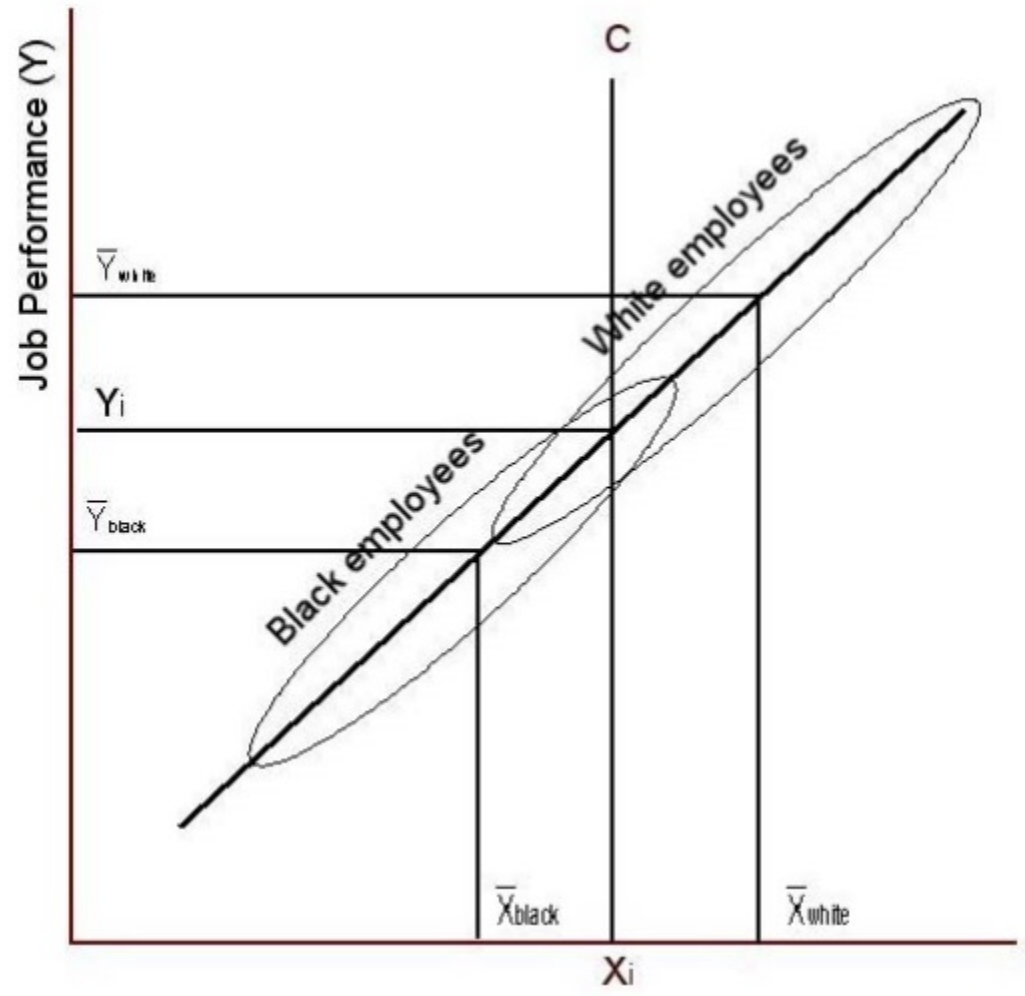
### Abstract

Researchers have investigated whether machine learning (ML) may be able to resolve one of the most fundamental concerns in personnel selection, which is by helping reduce the subgroup differences (and resulting adverse impact) by race and gender in selection procedure scores. This article presents three such investigations. The findings show that the growing practice of making statistical adjustments to (nonlinear) ML algorithms to reduce subgroup differences must create predictive bias (differential prediction) as a mathematical certainty. This may reduce validity and inadvertently penalize high-scoring racial minorities. Similarly, one approach that adjusts the ML input data only slightly reduces the subgroup differences but at the cost of slightly reduced model accuracy. Other emerging tactics involve weighting predictors to balance or find a compromise between the competing goals of reducing subgroup differences while maintaining validity, but they have been limited to two outcomes. The third investigation extends this to three outcomes (e.g., validity, subgroup differences, and cost) and presents an online tool. Collectively, the studies in this article illustrate that ML is unlikely to be able to resolve the issue of adverse impact, but it may assist in finding incremental improvements.

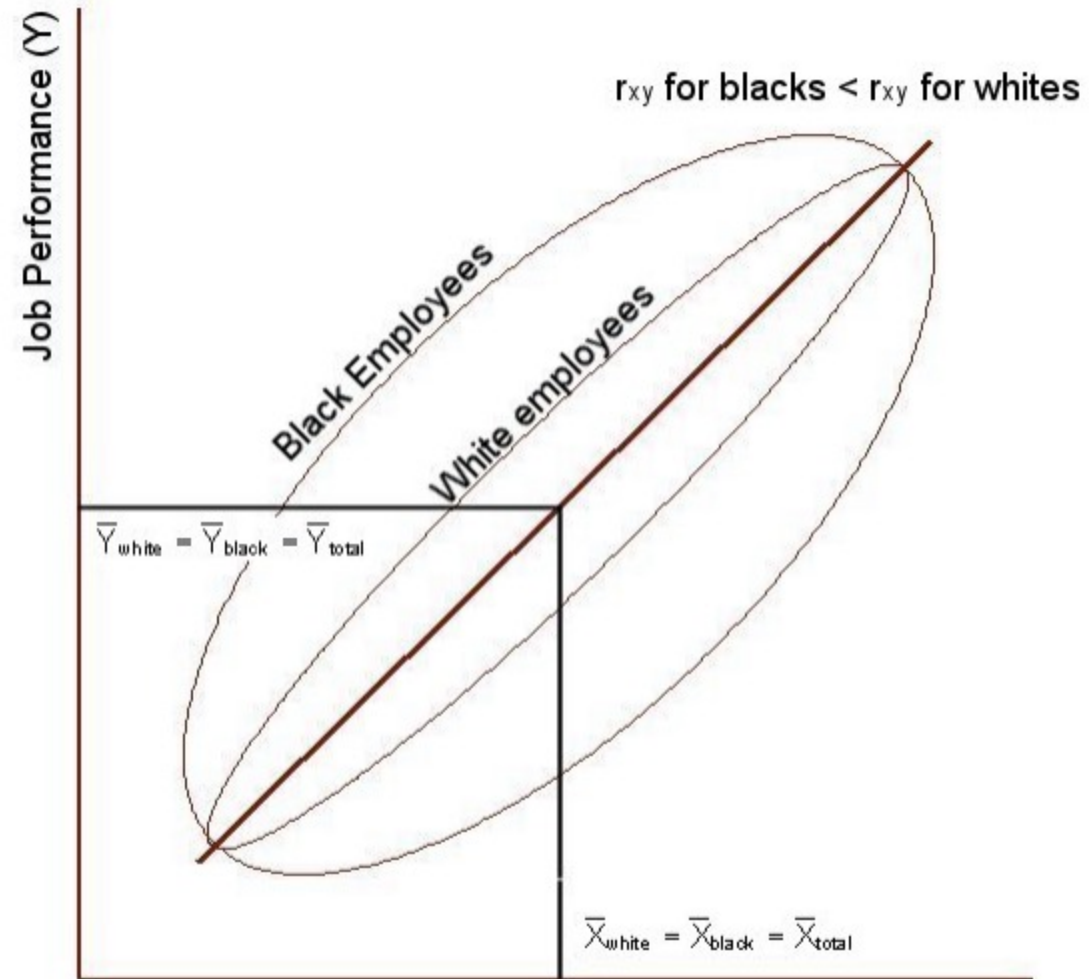
# Adverse impact – test bias



# Adverse impact without test bias



# No adverse impact





# Thank You

For questions, contact us  
[info@shl.ro](mailto:info@shl.ro)



**SHL.**

People Science. People Answers.

© 2022 SHL and its affiliates. All rights reserved.