

Creating an Unbiased, Predictive AI Partner for Hiring

Whitepaper

Created by



Introduction

We recently introduced TRACY, the Talent Recommendation Agent Customized to You, our overarching AI hiring system. TRACY is a paradigm shift for hiring, predicting good hires 4x better than interviews and creating new levels of confidence and efficiency never before seen in the hiring process.

TRACY curates a shortlist of recommended hires tailored to the team's needs. Her suggestions take the guesswork out of sifting through hundreds of applicants and shine a spotlight on standout talent. Additionally, she delivers standardized candidate data to hiring teams, which is proven to accurately predict job success. You'll learn the details of this below.

Searchlight Predicted Performance™ is a pillar within our predictive AI for hiring. This is a proprietary artificial intelligence solution created to programmatically reduce bias and guesswork in hiring in order to drive more efficient, equitable, and higher quality hiring decisions.

As we deliver a new and unique AI solution for hiring, it's imperative that we shed light on the internal workings of Searchlight Predicted Performance and answer key questions about how it works, the data that informs it, and how bias is removed from the AI. These are all core issues with using AI in HR today.

In this whitepaper, we will cover:

2	How Our AI Works
5	Model Development
12	Summary of Model Performance
15	Our Commitment to Ethical and Unbiased AI
16	Conclusion

How Our AI Works

The artificial intelligence in **Searchlight Predicted Performance** was designed to overcome two primary challenges that hiring teams face: (1) preventing bias and (2) eliminating guesswork. It's the removal of these two roadblocks that creates confidence in hiring managers and recruiting teams that they're hiring the right person.

Our AI was deliberately designed to prevent bias and eliminate guesswork by ensuring our data sources are simultaneously robust and unbiased, while also establishing consistency that allows us to develop pattern recognition on historical data and accurately extrapolate it on future examples.

Using AI to Prevent Bias

Preventing bias is a critical issue facing hiring teams. With the wider proliferation of artificial intelligence technology, there has been heightened attention on how artificial intelligence could be used to deliver fair and equitable hiring decisions.

The challenge many AI solutions face is that their datasets inherit the biases of their creators and contributors. For example, an AI solution that was trained on data that was overweighted on hiring decisions for a specific class of people might be more inclined to favorably recommend that specific class of people. In an example of AI technology gone wrong, David Heinemeier Hansson, creator of Ruby on Rails, commented on how the initial recommendations made by Apple Card gave him 20x the credit limit of his wife, despite sharing assets.



Our AI delivers hiring recommendations that remove bias from the equation. Unlike more familiar AI systems like ChatGPT, which censors its model from revealing its biases and exact inputs, we directly incorporate unbiased inputs into the decision-making of our AI. We are able to achieve this with:

- **Curriculum (i.e., Data Acquisition, Cleaning, and Selection):** Curating each of our data inputs to focus on unbiased sources of information minimizes the likelihood that our AI trains on biased data. This, in turn, reduces the likelihood of biased recommendations.

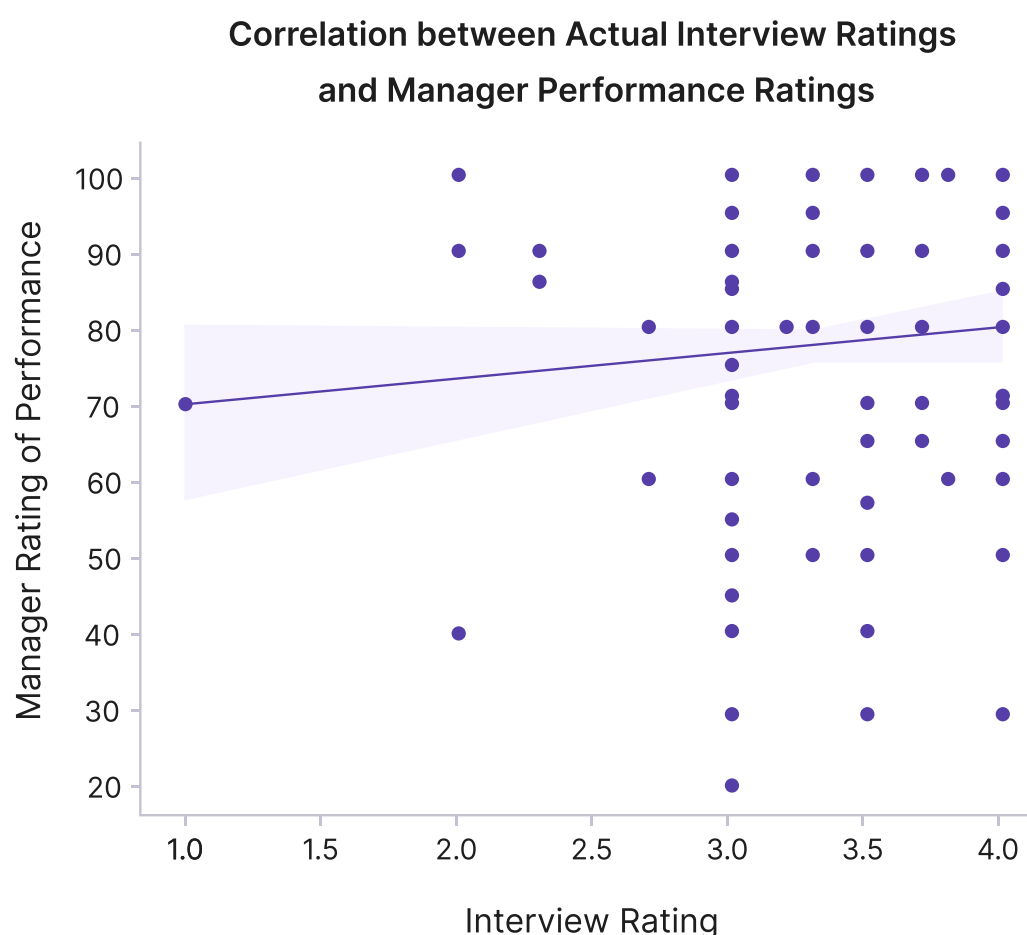
- **Reinforcement:** Penalizing unintended biases our AI may have gained through the training process. This acts as a corrective mechanism for our AI to adjust its recommendations in order to create fair outcomes.
- **Monitoring:** Measuring the decision-making and addressing bias proactively. In addition to Curriculum and Reinforcement, we continuously monitor the decisions and recommendations of our AI, so we uphold the highest of standards for the quality and fairness of our data and recommendations.

The combination of these three factors incorporated into how our AI is developed enables us to prevent bias within our hiring recommendations, which is covered in greater detail [here](#). A third-party certification that Searchlight meets AI regulations' bias audit standards is available [here](#).

Using AI to Eliminate Guesswork

At Searchlight, we aim to reduce guesswork and improve fairness by identifying the shared patterns across candidates and improve consistency of the hiring process. To improve consistency in the hiring process today, companies often attempt to address some of these challenges through interviewer training, structured interview questions, and other standardization strategies to get accurate measurements of a candidate's suitability for a job. However, despite the best efforts of hiring managers and interview teams, we found that most interviews have no correlation with post-hire performance.

We analyzed the correlation of interview performance to post-hire performance outcomes from a subset of customers' interview scorecard data. We found that most interviews had no correlation with post-hire performance and varied considerably from interview to interview. Among the companies within this study, the highest r-value within this study was only 0.14, meaning despite the hours upon hours that companies spent interviewing candidates, their interview processes still invited significant guesswork and randomness.



As an example, see the snapshot above showing the final score for a skills interview correlated to the manager's rating of new high performance post-hire. The correlation is $r=0.09$, which signifies weak or no correlation.

THE PITFALLS OF TRADITIONAL
HIRING PROCESSES CAN BE
ATTRIBUTED TO A FEW FACTORS:

- **Limits of Human Memory:** Research shows that the human brain can only hold up to seven pieces of information at once.
- **Subconscious Biases:** Humans have subconscious biases that subliminally impact our decision-making. These biases can result in interviewers overvaluing candidates who may share common education and experiences, rather than objectively assessing candidates for their skills.
- **Incomplete, Inconsistent Data:** While some interviewers take the time to write a nice paragraph or schedule a debrief meeting, some also only give a thumbs up or thumbs down. This feedback is also inconsistent, with some digging into the technical aspects, while some others might focus on years of experience, with no guarantee that you'll get a full picture. There simply isn't enough time in a traditional hiring process to gather complete information about a candidate.
- **Untrained Interviewing Skillset:** Talent hiring managers and interviewers will rarely be able to spend the core part of their time and energy on improving how they hire.

SEARCHLIGHT'S AI DOES NOT
SHARE THESE CHALLENGES.

- **Augments Human Memory:** AI is able to process magnitudes more information than the human brain, without being overwhelmed.
- **Add De-Biases:** Because of our de-biasing and the control that we exert on the data that our AI is trained on, Searchlight's AI does not suffer from conscious bias. The results of our most recent third-party audit showed: "All impact assessment results showed no bias at the 95% confidence level. Further analysis of the data and the model showed that both the data and model appear well-behaved, and additional test results showed no evidence of bias against minority groups.
- **Automated, Standardized Data:** Searchlight's custom data pipelines use an automated, standardized, written survey format that is exceptional at gathering high-signal, structured feedback on skills and behaviors that predict post-hire performance. Our standardized measurements reduce interviewer variation across our data set.
- **Specialized for Hiring Use Cases:** Our AI is grounded on a principle of consistency and is tailored for hiring, which allows us to develop pattern recognition on historical data and accurately extrapolate it on future examples.

We establish consistency in our AI through a five-step process:

01 Training Data



Studying a substantial number of examples.

02 Feature Generation



Openly considering all the possible interpretations of the data.

03 Feature Selection



Narrowing focus to the most cogent interpretations of the data while simultaneously ignoring the red herrings.

04 Training and Optimization



Going through historical examples, retroactively attempting to predict hiring outcomes, and iteratively learning to minimize mistakes.

05 Testing



Validating abilities on data sets never seen before to ensure the predictive abilities hold up under blind testing.

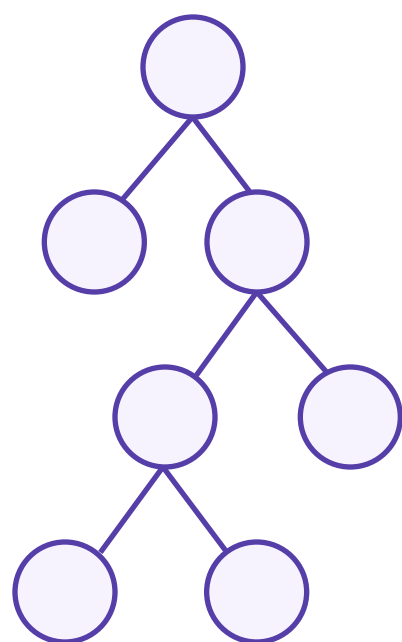
This five-step process eliminates guesswork by efficiently and accurately finding the most meaningful patterns across an endless number of data permutations.

Model Development

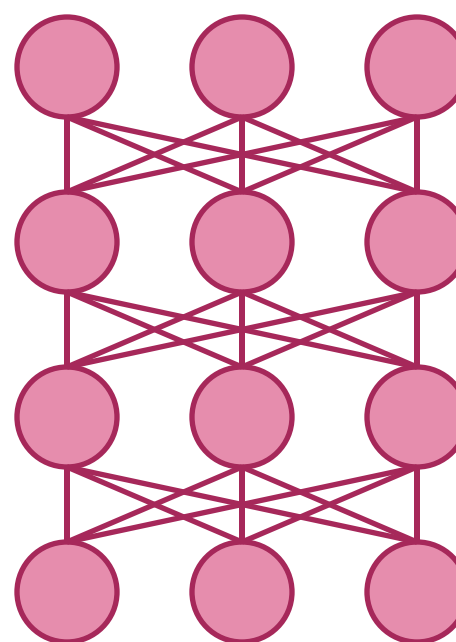
After testing hundreds of versions of machine learning algorithms, ranging from linear regressions to neural networks, our top performing models for the **Searchlight Predicted Performance** were all variations that used groups of decision tree models.

When choosing a model, there is rarely a one size fits all approach. In the case of the **Searchlight Predicted Performance**, we predict a hiring outcome based on a medium-sized data set primarily consisting of tabular data. You can think of tabular data as the kind that you might reasonably organize in a spreadsheet. Our conclusions from testing are corroborated by [recent research](#), which found that for structured, tabular data, decision tree models often outperformed deep learning neural networks. Given both the theoretical and empirical support for using tree models, we selected [LightGBM by Microsoft](#) as our model of choice because of its consistently superior performance.

Decision tree



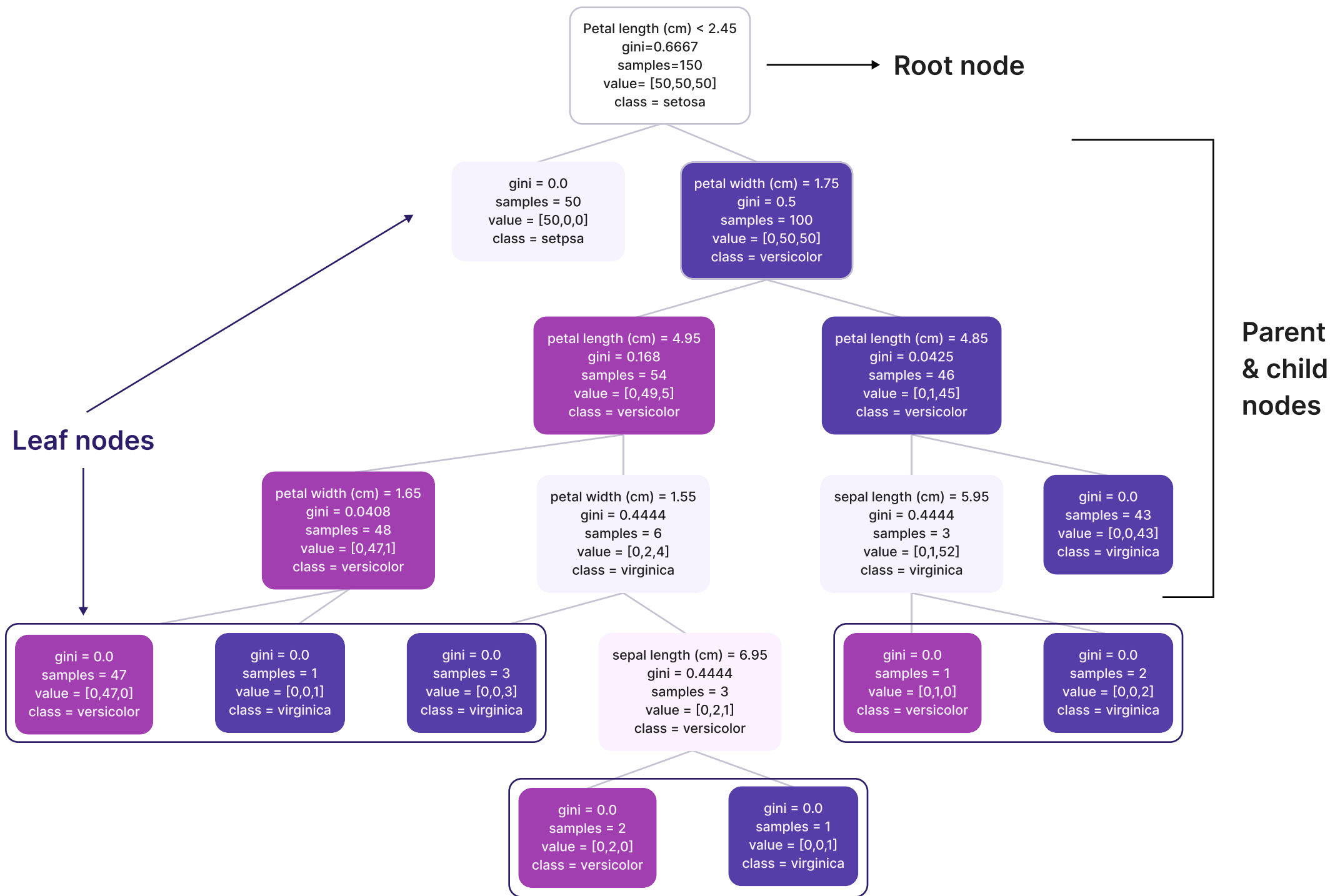
Neural network



Overview of Decision Trees

Decision tree models consist of connected nodes that are linked together by one to two pieces of behavioral data or context. A decision is made about which path to follow based on rules about those pieces of data. At the end of the path, we reach a leaf node, which provides a score. We use a group of these trees, often called a forest, to simplify the decision and minimize assumptions that each tree is responsible for and then combine the results by adding up the scores. Having each tree only consider a subset of the data and then aggregating multiple trees together is analogous to using the wisdom of the crowd to overcome single points of failure.

Example of a Decision Tree



When we train the model, we tune a parameter to decide how many trees to sequentially generate.

Parameters often include things like how many trees to use, how large a tree a can be, how much to weight small errors compared to large errors, learning rates, and many other factors. These fundamentally constrain the training process and allow us to control trade-offs, such as accuracy versus speed, as well as limit risks such as overfitting, which can result in poor performance outside the training data set. Each sequential tree focuses on correcting the errors of the trees that were generated before it. In doing so, we continually reduce the number of errors in the model.

Before we examine how we encode these features into our model, we will explain the data sources we use.

Data Sources

In our decision tree models, our primary data sources are our proprietary assessments, which include peer and manager reviews and Quality of Hire surveys, which encapsulate post-hire performance and retention in new hires' first 90 to 180 days and are strongly correlated to first-year performance reviews. Because of our highly-structured, consistent approach towards collecting this data, we can minimize the amount of bias in our training data without impacting the performance of the AI. From these data sources, we analyze:

1. Relationship Context:

Positive references and recommendations from previous managers or colleagues would increase the score, as they vouch for the candidate's abilities and work ethic. For each reference, we examine the relationship between the candidate and the peer or manager. We take into consideration how many years worked together, the frequency of interactions within a given week, and the qualifications and experiences of each reference provider. Similarly, for our Quality of Hire scores which measure post-hire outcomes, we deliberately collect information from both the manager and the employee at the 90-day and 180-day marks.

2. Peer Ranking and Past Performance:

We also take into consideration how the candidate ranked relative to their peers based on past performance, which provides unique insights into the candidate's professional reputation.

3. Soft Skills:

89% of mishires are due to missing soft skills. Searchlight has a proprietary skills ontology that includes soft skills and working styles that are notoriously difficult to measure, and are empirically linked to post-hire performance. Some example skills include adaptability, collaboration, problem-solving, and influencing others.

4. Cultural Alignment:

This is an assessment of the candidate's values, work styles, and attitudes. The cultural alignment factors that we use are backed by organizational psychology research.

5. Competencies and Job Requirements:

How closely the candidate's profile matches the specific job requirements and responsibilities influences a job match. For each job family, we attach job-specific competencies that further add context to our model. Our scorecard capabilities allow us to evaluate job-specific competencies, so in the case of a sales role, we can incorporate sales-specific competencies like "sales effectiveness," "value selling," etc.

6. Post-Hire Performance Labels as Ground Truth:

Unlike other algorithms that use the "hired" event as the success case, our algorithm trains on performance on the job. This surfaces our commitment to outcomes-based hiring, because we know that there can be large discrepancy between interview ratings and actual employee success.

In the future, we intend to incorporate more data, including:

- **Candidate Qualifications and Specialized Skills:** The score takes into account the candidate's educational background, relevant work experience, certifications, and any specialized skills required for the job. Traditionally, these components were captured in a resume and primarily used for screening. We believe incorporating this into our AI will further improve our ability to identify the right candidates.
- **Role-Specific Outcome Metrics:** In the future, role-specific outcome metrics will further enhance the reliability of the **Searchlight Predicted Performance** by training the model to detect role-specific nuances that can better translate specific skills to specific business outcomes. For example, for sales-specific roles, we would be able to train our AI on a sales metric like “quota attainment.” Alternatively, for customer success roles, we could train our AI to be trained on a customer success metric like “net retention.”
- **Experience Relevance:** The relevance of the candidate's past roles to the current job opening plays a role in score calculation. For example, one of the most common challenges hiring teams face is translating experiences from a large company to a startup and discerning which experiences translate and which don't. By doing so, we can better measure and infer transferable skills between organizations.

Feature Encodings

There are many ways that we translate the aforementioned data sources into features that the AI uses for prediction. Here are a number of techniques we use:

- **Feature Interactions:** By examining the interactions between two or more features, we can isolate valuable patterns in our data. An example might be that a manager describing a candidate as “diligent” may be more consistently predictive of high performance. Alternatively, we could look at the ratio between two other traits, such as how “collaborative” someone is relative to how “conflict forward” someone is.
- **Group and Cohort Features:** By examining differences within and between behavior cohorts, we can extract further useful patterns in our data to improve the predictiveness of our AI. For example, there are notable differences in how a much a soft skill trait like “active listening” may be described within the context of an engineering role compared to a sales role.
- **Normalized Features:** A raw value about someone’s tendency to be detail-oriented may not easily distinguish high performers from low performers. However, if we transform the value in the percentile against the whole data set, the model can better understand the significance of this trait.
- **Ordered Features:** In some of our questions in our proprietary assessments, we use a Likert scale, and we can translate this relationship into an ordering. In practice, instead of using text values like “disagree” or “strongly agree,” the machine learning model will encode the value into numbers. Therefore, “disagree” would be encoded as a two, while “strongly agree” would be encoded as a five.
- **One Hot” Features:** Encode whether or not a piece of information is present at a given time.

There are many other techniques, such as hashing or looking at the contrast between different values. More information is [available here](#).

Additional Details on Encoding Soft Skills and Cultural Alignment

Searchlight uses data generated from our own custom data pipelines and our gold standard, structured survey questions. Our collaboration with Industrial and Organizational (IO) psychologists have resulted in a list of six competencies mapped to 30 traits and six universal working styles that empirically relate to job performance and employee satisfaction across different jobs and industries. The competencies measured in our assessment were identified through evaluating our proprietary database of Searchlight hiring assessments and an extensive literature review of decades of I/O Psychology research, such as the [organizational culture profile](#).

Our models looks at structured behavioral data, such as top strengths, strengths, growth areas, and working styles. We look at these behavioral factors both individually and in combination using the techniques described previously. We then translate these into a mathematical interpretation of the data using what is called features. This is how we translate a human-interpretable concept, such as work ethic and integrity, into a mathematical measurement.

For example: we look across all data about a candidate, then use a statistical heuristic, such as skew, to measure how much the data agrees that this candidate is results-driven. We derive over a million such interpretations from our data and select a reasonable number of them in automated fashion such that each incremental feature adds unique insight into our decision-making process. The reason we consider so many interpretations of the data then whittle down to the most significant features is to overcome the pitfalls of dimensionality, in which machine learning models have an exponentially harder time finding consistent patterns when overwhelmed with information.

Why We Decided Against Using Neural Networks

Neural networks are a favored model among companies with large, unstructured data sets. However, for our structured, tabular data, decision tree models are superior for myriad reasons. Very few companies have internal data sets that look like this outside of the Fortune 100, search engines, social media sites, or Wikipedia.

Notably, neural networks are notoriously difficult to explain. They tend to be black boxes around individual decisions, and their results are usually explained in aggregate. In a human-centric process like hiring, every individual decision matters. As we developed the **Searchlight Predicted Performance**, it was critical for us to be able to explain and interpret the recommendations from our AI. The neural network approach would have significantly constrained our ability to deliver cogent explanations to hiring teams on why one candidate was a better fit over another candidate for a given role.

Furthermore, in decision tree models, we maintain the ability to debug and explain individual decisions based on which features impact the final result. This allows us to isolate which pieces of information most impact a decision, which is particularly advantageous as we prioritize fairness and check our data features for bias.

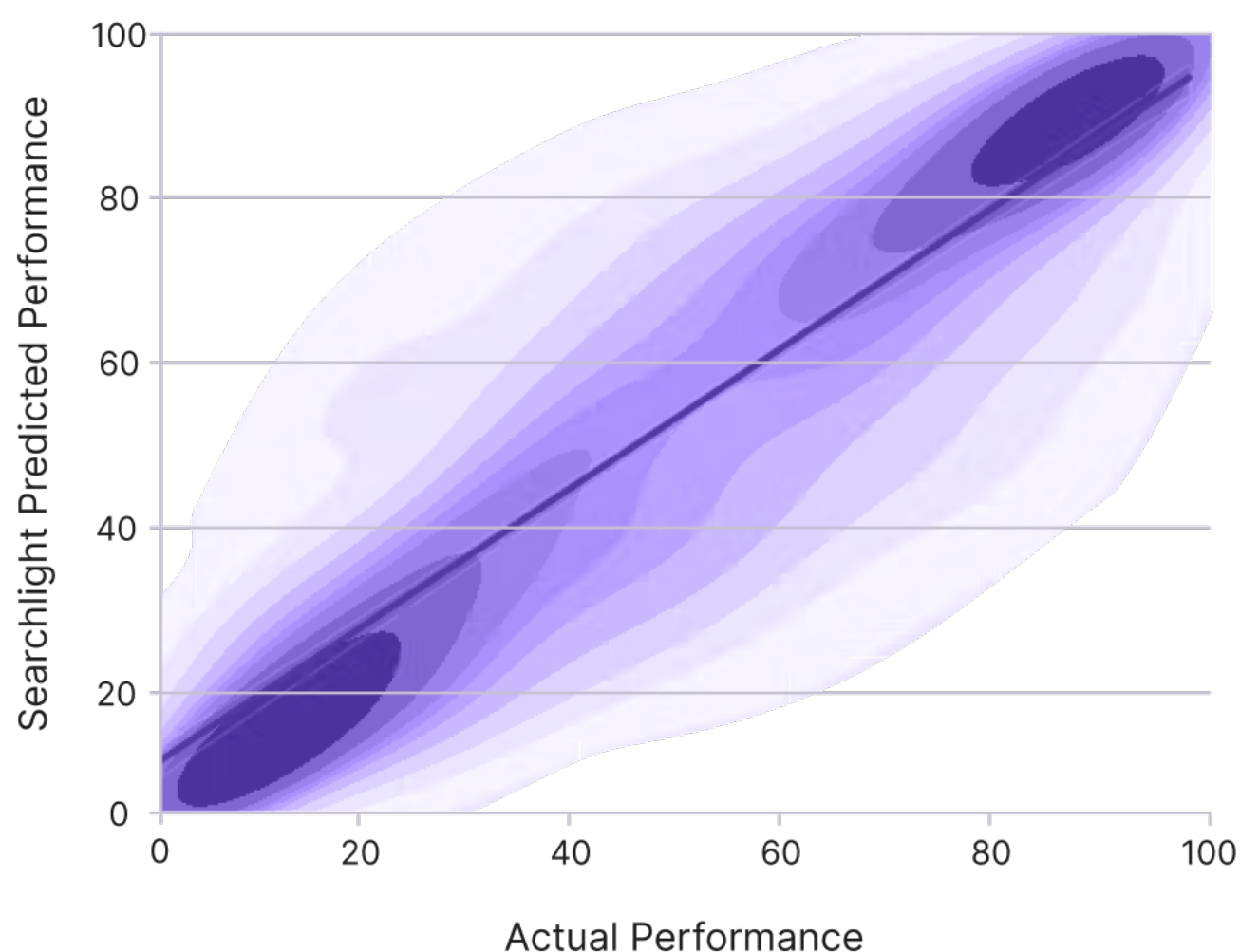
Summary of Model Performance

Across Searchlight's data set, we were able to achieve a correlation of $r=0.55$ between the Searchlight Predicted Performance and actual performance outcomes of a validation data set. This correlation is meaningful for two particular reasons:

1. **These results demonstrate strong correlation.** Typically, correlations above 0.5 are considered a high correlation. Moreover, compared against the 0.14 average correlation that interviews garner in practice according to our internal studies on actual interview and performance data.
2. **These results are corroborated against an independent data set.** This correlation is meaningful because it was measured by an independent data set — separate from our test data set — and corroborated by 5-fold cross validation testing.

Beyond the Pearson r correlation, Searchlight's AI achieves >70% accuracy in predicting high and low performance among candidates.

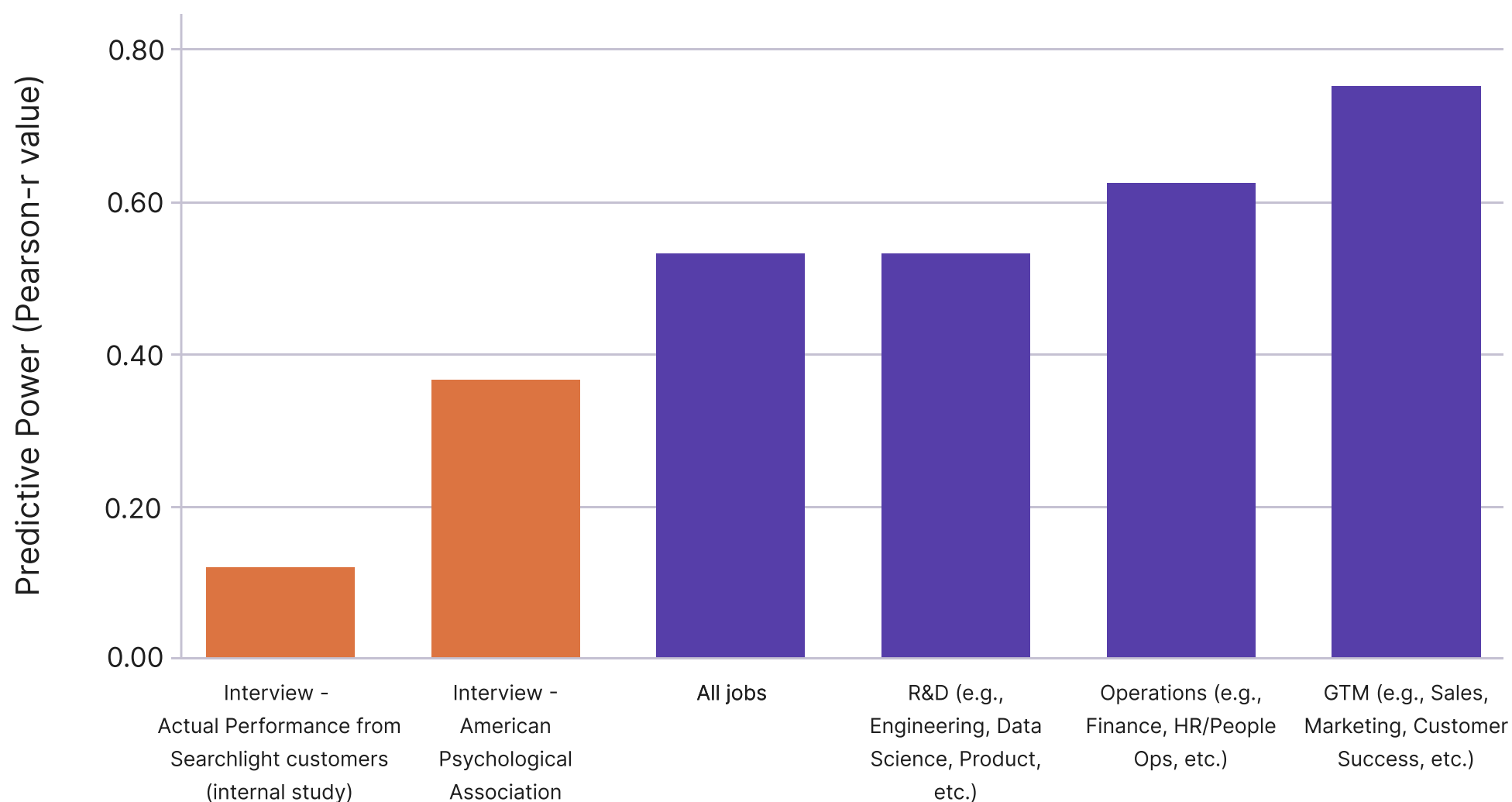
In the graph below, you can see two darker clusters on the bottom left and top right of the chart, demonstrating the superior predictive performance for high and low performance among candidates.



Predictive Accuracy of Searchlight Predicted Performance

Based on prior research which quantitatively reviewed over 100 years of academic research (across industries and occupations in the U.S.), we can conclude the predictive power of the **Searchlight Predicted Performance** (Pearson $r = .55$) is on par, if not better than, the predictive power of most commonly used applicant screening / interviewing tools such as situational judgment tests, years of prior work experience, education, personality traits, and emotional intelligence. Furthermore, when you compare the r -value of the Predicted Performance score against the best performing interviews in practice ($r = 0.14$), we see that Searchlight's Predicted Performance Score is roughly 4x more predictive than a structured interview.

Searchlight Predicted Performance Score compared to other Assessments



Source: [The validity and utility of selection methods in personnel psychology](#). American Psychological Association.

Searchlight Predicted Performance

How the Accuracy of Searchlight Predicted Performance Varies by Job Family


One of the core components our AI measures is soft skills to predict on-the-job performance. As a result, our model tends to have a lower error rate for customer or client-facing roles, such as account executives or customer success, than traditional research and development roles. For roles typically within go-to-market departments, we tend to see correlation closer to 0.75, whereas R&D roles are closer to 0.55. We observe similar patterns with other external-facing roles such as recruiting and select marketing roles.

How Searchlight Predicted Performance Is Visualized in the Searchlight Platform


When you see Searchlight Predicted Performance as Top Candidate, >70% of the time, that candidate becomes a top 20th percentile performer.

Joseph Chang

Product Management (2022). Generated at Apr 25, 2022 1:48 pm



Overall Recommendation



Predicted Performance
Top Candidate
Potential bar raiser

Inputs	
Quantitative Factors Good Neutral • Good • Great	Qualitative Factors Great Neutral • Good • Great

"Quantitative Factors" summarizes the responses from the multiple choice and Likert questions. "Qualitative Factors" summarizes the intangibles (e.g. Behaviors, Culture Continuums) that were detected by our AI, but are harder to see with the naked eye.


Why is Joseph rated "Top Candidate"?
Searchlight's AI detected patterns in Joseph's behavioral profile that match those of top-performing hires across Searchlight AI's dataset.

Example of Searchlight Report - see full report sample [here](#)


In contrast, when you see Searchlight Predicted Performance as Risky, >70% of the time, that candidate is a bottom 20th percentile performer.

Mark Scout

Software Engineer (2022). Generated at Apr 25, 2022 1:48 pm [Full report](#)



Overall Recommendation



Predicted Performance
Risky
Negative indicators of job performance

Inputs	
Quantitative Factors Neutral Neutral • Good • Great	Qualitative Factors Neutral Neutral • Good • Great

"Quantitative Factors" summarizes the responses from the multiple choice and Likert questions. "Qualitative Factors" summarizes the intangibles (e.g. Behaviors, Culture Continuums) that were detected by our AI, but are harder to see with the naked eye.

Why is Mark rated "Risky"?
Searchlight's AI detected patterns in Mark's behavioral profile that may negatively impact job performance. Look closely at Delaware's intangibles and the open-text feedback to evaluate the fit with your team.

Our Commitment to Ethical and Unbiased AI

At Searchlight, we're committed to improving hiring through ethical AI. We believe building a diverse team and inclusive culture is win-win for candidates and organizations. The studies back this up: diverse teams financially outperform their peers by 36% (McKinsey reports on [inclusion](#) and [diversity](#)) and inclusive teams outperform their peers by 80% ([Deloitte](#)). But diversity isn't a box to check — it is a commitment to thinking about talent and non-traditional candidates in a new way. Using Ethical AI, we tie diversity initiatives to business outcomes at every part of the hiring process.

Searchlight's AI systems are assessed continually through internal data reporting and periodically through an independent third-party audit from legal, policy, and technical experts in AI. Moreover, we involve human-in-the-loop (HITL) machine learning for our AI systems to prevent unintended bias. With Searchlight, you can be confident that your hiring process will be less biased, more effective, and compliant with current and upcoming laws and regulations.

How Searchlight Proactively Mitigates Bias

Responsible, ethical AI is the foundation of our work at Searchlight. We have implemented many different forms of bias mitigation, which blend AI development with HITL monitoring and intervention.

- We have automated monitoring of the raw behavioral data from our surveys to ensure that there is no racial or gender bias in any of the individual pieces of behavioral data. This crucially ensures that when human decision-makers look at individual factors, they do not make decisions in a biased way. Similarly, machine learning models also have less bias to mitigate when the underlying data used to train and make decisions with is unbiased.
 - For example, when our monitoring system discovered that there was male gender bias when candidates were rated as articulate, we removed this attribute from our training & prediction data set.
- All impact assessment results showed no bias at the 95% confidence level. Further analysis of the data and model showed that both appear well behaved, and additional test results showed no evidence of bias against minority groups.

Compliance to NYC Law 144's Bias Audit Standards

Effective July 2023, the State of New York began enforcing [New York City Local Law 144 \(NYC 144\)](#), which prohibits employers from using an automated employment decision tool (AEDT) to screen candidates or employees for employment decisions unless the tool was subject to a bias audit. Searchlight is proud to be a leader in ethical AI. We are audited by experts in AI to meet NYC Law 144's Bias Audit standards, even though we are not an AEDT.

In accordance with this law, we participated in a third-party audit with INQ Consulting and Armilla, a reputable leader in AI governance, compliance, and regulation. All impact assessment results showed no bias towards any minority group at the 95% confidence level. You can learn [more here](#).

Conclusion

Looking Ahead

As we move forward, we will continue to iterate and improve upon our core AI systems to create new levels of confidence and efficiency in the hiring process. We will do so by:

- Expanding the types of data that are included in our AI training.
- Adding new capabilities to improve recruiter efficiency, reducing manual work so that they can spend more time strategically advising the business.
- Maintaining the highest standards in ethics.

While AI technology is quickly evolving, Searchlight is committed to being a leader in developing predictive and ethical hiring solutions. The Searchlight Predicted Performance is the first step to shifting an existing paradigm within hiring with AI and human collaboration at its core.

Are you curious about what Searchlight can help you accomplish for your business?

Searchlight customers see better new hires, faster onboarding, greater employee lifetime value, and ultimately better business performance.

Successful outcomes include:

- Digital mortgage closing platform provider Snapdocs saved \$3M+ in recruiting costs, and generated \$3M+ in increased revenue because of increased performance.
- B2B online auction platform B-Stock reduced Time to Fill for Customer Success roles by 40%, ramped new hires faster, and reduced 90-day attrition and misalignment rates.
- Online education provider Udemy increased first year new hire retention by 20%.

If you want to learn more, [book a demo](#) with our team.