FULL METHODOLOGY:

Modeling Corporate Wages for JUST Capital's Rankings of America's Most JUST Companies

September 2022

Kavya Vaghul, Senior Director of Research, JUST Capital
Lisa Simon, Senior Economist, Revelio Labs
Daniel Firester, Lead Data Scientist, Revelio Labs
Rob Marsh, Chief Information Office, JUST Capital

INTRODUCTION

Year after year, JUST Capital's polling has revealed that paying workers a fair, living wage is the top priority for the American public when it comes to just business behavior. But actually measuring whether each company we rank pays its workers fairly can be challenging due to limitations in publicly available data – which mostly focuses on pay equity and minimum wage – and non-standardized reporting on compensation. That's why our annual Rankings rely on models to estimate the state of wages among Russell 1000 companies.

We've partnered with Revelio Labs, a labor market data provider that is working to create the first universal HR database, to leverage their unique employer-level data sets and modeling capabilities to create estimates for corporate wage data used in our Rankings and other non-Rankings research.

This document details the methodology for modeling these estimates, with a particular focus on how we generate values for three of the data points used in our annual Rankings within the "Pays a fair, living wage" Issue:

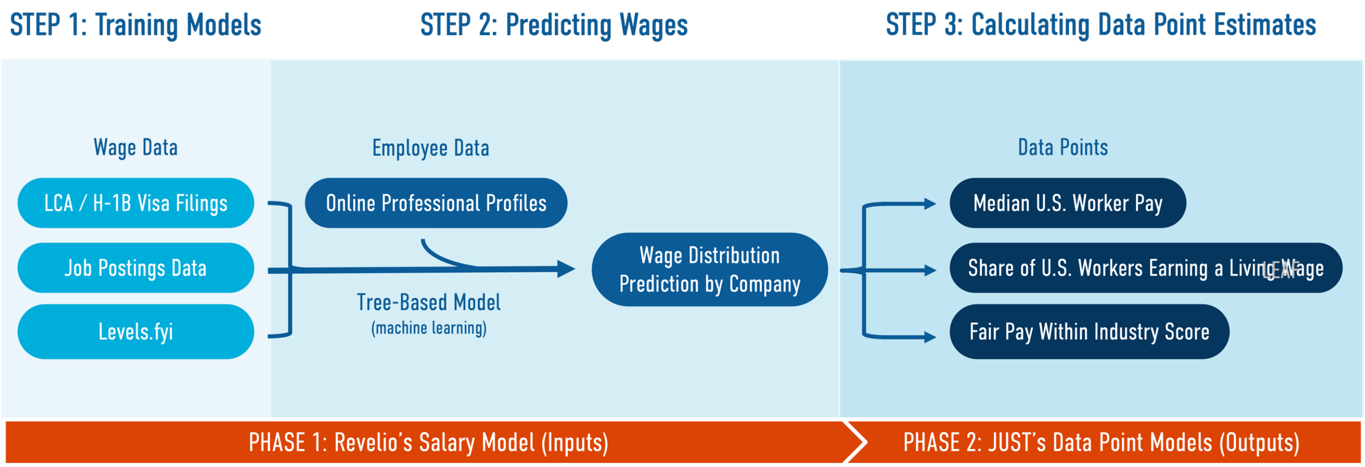
❶ Median U.S. Worker Pay (to compare to CEO Compensation)

❷ Share of U.S. Workers Earning a Living Wage

❸ Fair Pay Within Industry Score

METHODOLOGY SUMMARY

There are two broad modeling phases required to generate the three modeled wage data points used in JUST Capital's annual Rankings: (1) **Revelio's Salary Model** and (2) **JUST's Data Point Models**.


STEP 1: Training Models
STEP 2: Predicting Wages
STEP 3: Calculating Data Point Estimates
Wage Data
LCA / H-1B Visa Filings
Job Postings Data
Levels.fyi
Employee Data
Online Professional Profiles
Tree-Based Model
(machine learning)
Wage Distribution Prediction by Company
Data Points
Median U.S. Worker Pay
Share of U.S. Workers Earning a Living Wage
Fair Pay Within Industry Score
PHASE 1: Revelio's Salary Model (Inputs)
PHASE 2: JUST's Data Point Models (Outputs)

Revelio Lab's Salary Model leverages several raw data sources to build **training models**, which are then used to **predict** an employee-level salary distribution for each company we rank in the Russell

1000. These distributions are then input into Data Point Models to **calculate estimates** (or outputs) by company for three data points – median U.S. worker pay, share of U.S. workers earning a living wage, and fair pay within industry score – used in both our annual Rankings and other non-Rankings analyses.

PHASE 1: REVELIO'S SALARY MODEL (INPUTS)

Revelio Labs' Salary Model is fueled by **machine learning**, a technique used by data scientists that relies on computer-based algorithms to extract insights from data. These algorithms learn from data and recognize patterns – a stage known as **training** – and can then be applied to **predict** a value or outcome for a new or hypothetical input observation.

STEP 1: Training Models

A robust training model ("supervised machine-learning problem") requires the input of historical or existing raw data that includes a "ground truth" of the outcome (salary, in this case) for the algorithm to learn. The model then identifies different patterns in the raw data to build the context necessary to inform predictions.

In the case of the Salary Model, Revelio Labs uses **three raw salary or wage data sources** for training, all of which include specific employer-level data:

- Labor Conditions Applications: This data comes from the Department of Labor, which maintains company-level data on each H-1B visa filing. These applications include information on company names, individual titles, standard occupational classifications, location, and salaries, among other worker characteristics.
- Online Jobs Postings: This data comes from a labor market intelligence data provider that aggregates job postings data from sites like Indeed. It includes information on company names, job titles, descriptions of role responsibilities, location, and salary ranges, among other worker characteristics.
- Levels.fyi: This data is crowdsourced by workers at mostly technology companies and includes information on company names, job titles, tenure, location, and highly disaggregated compensation, among other worker characteristics.

Because data on work hours is not available in any of these data sets, the Salary Model assumes that all workers are full-time – that is, they work for 40 hours per week over 52 weeks per year.

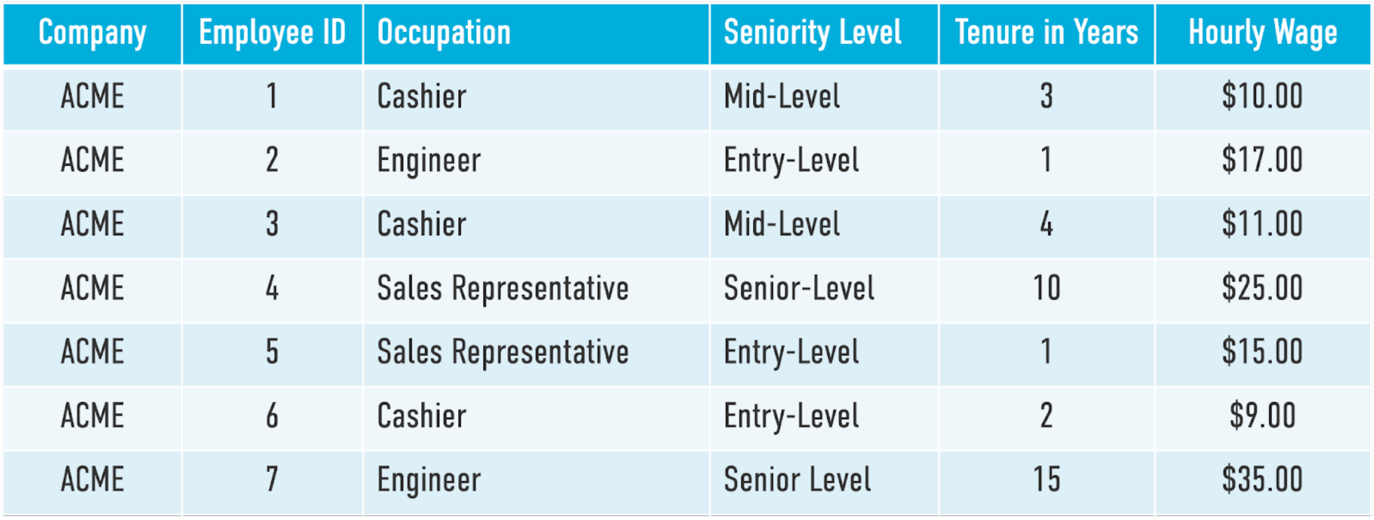
Each raw salary or wage data set has a different nomenclature for companies and workers characteristics, so Revelio Labs undertakes a significant **cleaning and standardization** process prior to building the training models. In this step, every possible reference to a given company or its subsidiary (such as ticker or stock symbol, name, short name, or acronym) are mapped to a unique company identifier, and occupations and seniority levels are categorized into Revelio Labs' custom taxonomy for consistency across datasets.

Once the raw data is cleaned and standardized, Revelio Labs selects specific algorithms to train, which are picked to best match the attributes of the raw data. Writ large, probabilistic **tree-based** models are the best fit for salary and wage data since they set up a series of if-then rules (or "decisions") to make either numerical (regression) or categorical (classification) predictions.

Here's how tree-based modeling works in the context of the Revelio Salary Model:

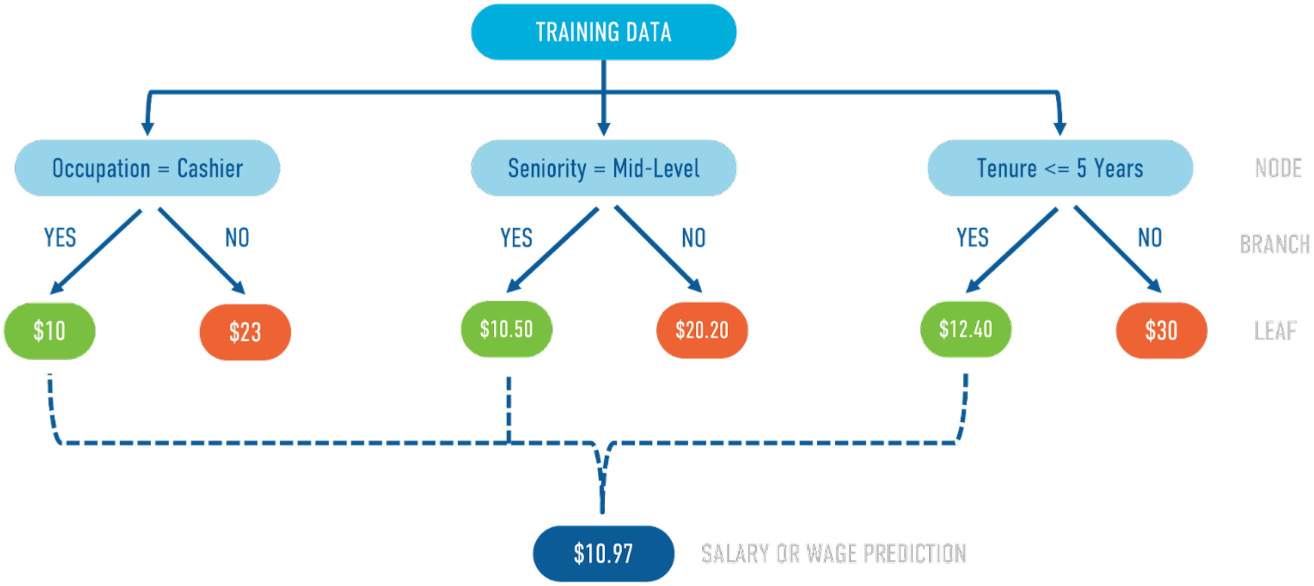
1. The raw data from the Labor Condition Applications, online job postings, and Levels.fyi are fed into their respective training models. Both the Labor Conditions Applications and Level.fyi data are used to train Bayesian Additive Regression Trees (BART model), while the online job postings data trains an Extreme Gradient Boosted Tree (XGBoost model).
2. In a method known as ensembling, the decision tree algorithm goes through covariates (worker characteristics like title, seniority, and tenure, among others) and splits them into leaves based on which cutoffs maximize the difference in outcomes – in this case, different salaries.
3. Each decision tree yields a salary prediction, which is generated by averaging all the values within a given leaf. Independently, however, these predictions are weak because they may focus on a singular worker characteristic.
4. To strengthen the salary prediction, the ensemble method combines the predictions across all decision trees either by taking a weighted or unweighted mean, depending on what type of model is being trained.
5. At this point, the decision tree can then be used to predict a salary (or salary distribution including medians, means, and quantiles) for a "new" or hypothetical observation (worker) at a given company.

While the Revelio Salary Model is significantly more complex, the following illustrations provide a hyper-simplified visual explanation of how a tree-based model could be used to predict the wage of a mid-level cashier with less than five years of experience at the fictional company ACME. Suppose the training data for the model includes seven workers at ACME, their characteristics by occupation, seniority level, and tenure, and their hourly wages:



Company	Employee ID	Occupation	Seniority Level	Tenure in Years	Hourly Wage
ACME	1	Cashier	Mid-Level	3	$10.00
ACME	2	Engineer	Entry-Level	1	$17.00
ACME	3	Cashier	Mid-Level	4	$11.00
ACME	4	Sales Representative	Senior-Level	10	$25.00
ACME	5	Sales Representative	Entry-Level	1	$15.00
ACME	6	Cashier	Entry-Level	2	$9.00
ACME	7	Engineer	Senior Level	15	$35.00

The algorithm would split up the data to build the following decision tree, where the value of each leaf is the average hourly wage within the node's grouping. The wage prediction is the average hourly wage across all of the leaves corresponding to the worker characteristics in which we were interested.



STEP 2: Predicting Wages

Once all of the training models have been calibrated, they can be used to predict a salary or wage distribution for the population of employees at any of the companies JUST Capital tracks, data that the models have never seen before.

This requires some information about the expected employee distribution at a given company. The best publicly available estimate of the full employee distribution comes from cleaned and standardized online professional profiles, which include detailed information about companies' current and former employees. Using start and end dates for workers on these online profiles, Revelio Labs produces a point-in-time snapshot of a company's employee distribution by title, seniority level, location, and many other characteristics that are mappable back to the tree-based models. Though this data can track how workers transition within and across companies, the Salary Model only focuses on employees present in the last year. Like with the raw salary and wage data, without data on work hours or employee type (full-time, part-time, temporary, contractor, or others), we assume that all workers in the online profiles are full-time, working for 40 hours per week over 52 weeks per year.

To mitigate the potential biases toward specific types of workers – like corporate or professional staff – that may be overrepresented in an online-only population, Revelio Labs also creates a representative weight for each individual at a given company based on data from other sources like the Bureau of Labor Statistics. These weights become especially important when calculating the data point estimates later on in the process.

With the employee distributions for each company in hand, Revelio Labs feeds each worker's characteristics into the trained models:

1. Each model goes through its respective decision tree for that worker to predict their individual salary distribution.
2. To choose which model's prediction is the best fit for a particular worker, a k-nearest neighbor algorithm is applied to numerically describe how closely that worker is represented in each of the raw data sets that train the model.
3. This process upweights or downweights each model's prediction.
4. Finally, the weights associated with each model help apply a sampling distribution, such that the outputs from the overall Salary Model are random predicted salary samples for each individual at a given company.

In other words, by the end of the Salary Model phase, we have generated a value for the median and mean salary and other salary percentiles for each worker in the U.S. who identifies themselves on online professional profiles as working for a given company, assuming they work full-time.

PHASE 2: JUST'S DATA POINT MODELS (OUTPUTS)

These predicted salary distributions for each company are a basis from which we can calculate many statistics on the state of wages among Russell 1000 companies. For the purposes of JUST Capital's annual Rankings, we leverage the salary distributions to estimate three measures of the fairness or livability of a company's wages:

1. **Median U.S. Worker Pay:** This value is compared to CEO Compensation to measure the vertical fairness of pay within a company.
2. **Share of U.S. Workers Earning a Living Wage:** This value measures the degree to which a company's pay is meeting all workers' basic needs.
3. **Fair Pay Within Industry Score:** This value measures the horizontal fairness of pay between a company and its peers.

STEP 3: Calculating Data Point Estimates

Median U.S. Worker Pay

In 2015, The Securities and Exchange Commission adopted a rule requiring public companies to disclose, in annual proxy statements beginning in 2017, the pay ratio between the Chief Executive Officer's and median employee's compensation. The rule, however, provided companies with significant flexibility in identifying who the median worker is and how to calculate the median, with the ability to change this definition every three years. As a consequence, the publicly reported median worker pay data are incomparable and inconsistent.

JUST Capital's annual Rankings prioritize the inter-company and inter-industry comparability of data, so we choose to model the median worker pay for each company's full-time U.S. workforce. To do so, we take a median across the salaries predicted for each worker who identifies themselves on online professional profiles for each ranked company. This is a weighted median, instead of a raw median, based on the representative weights previously calculated for each individual to account for the fact that worker observations come from an online-only population. Using a similar weighted median approach, we also compute the 5th, 10th, 25th, 75th, 90th, and 95th percentile salaries for each company to paint an aggregate picture of each company's salary distribution.

Share of U.S. Workers Earning a Living Wage

A living wage is the amount of money – prior to taxes – that a worker and their family require to cover the cost of their basic needs where they live, including essentials like housing, food, health insurance, utilities, transportation, and child care, among other necessities like clothing and personal care items. The actual value of a living wage in a given year varies across two dimensions: location and family size. So a family living in a county or Metropolitan Statistical Area (MSA) with a lower cost of living will have a lower living wage value than one living in county or MSA with a higher cost of living, and similarly, a family with no children will have a lower living wage value than one with children who would require child care.

While there are many different sources that calculate living wages or family budgets, JUST Capital uses data from MIT Living Wage Calculator as a benchmark against which we can measure companies' performance. Through a robust feedback process we conducted with external experts, we determined that our data point model would use:

1. A *national* living wage threshold, averaging county-level living wage data weighted by county population in instances where national figures for component costs were not available. As more geographically granular company data becomes available, we will transition to using location-specific living wage thresholds to better customize the model to companies' establishment locations and more accurately capture the geographic variation in living wage values.
2. A family composition of *two full-time working adults and two children*. The living wage is the amount one of the full-time working adults would need to make to contribute to their family's basic needs.

In 2022, the national population-weighted average living wage required for one worker in a family of two full-time working adults and two children to meet their basic needs is $24.16 per hour or roughly $50,250 per year. Each year this value can change depending on inflation and other factors.

To calculate the share of U.S. workers earning a living wage, we leverage the full predicted salary distribution for each individual worker at a company to identify what point in the distribution – or percentile – falls above or below the living wage threshold. For example, if the living wage value of $24.16 per hour fell at the 25th percentile predicted wage for a given worker, we would assume that there is a 75% likelihood that the worker is earning a living wage. We then average across the likelihoods for individual workers within the company, weighting by each of their representative weights previously calculated to account for the fact that worker observations come from an online-only population

In the figure below, we can see a simplified example of how this calculation would play out for that fictional company ACME and its seven workers and how the representative weights help paint a more accurate portrait of the share of living wage earners at a given company:

Company	Employee ID	Occupation	Percentile Matching a Living Wage	Likelihood of Making a Living Wage	Representative Weight	Likelihood x (Representative Weight / Sum of Weights)
ACME	1	Cashier	90	10%	4	0.013
ACME	2	Engineer	10	90%	8	0.232
ACME	3	Cashier	85	15%	10	0.048
ACME	4	Sales Representative	5	95%	3	0.092
ACME	5	Sales Representative	40	60%	3	0.058
ACME	6	Cashier	92	8%	2	0.005
ACME	7	Engineer	2	98%	1	0.032
			Raw Average	54%	Weighted Average	48%

Fair Pay Within Industry Score

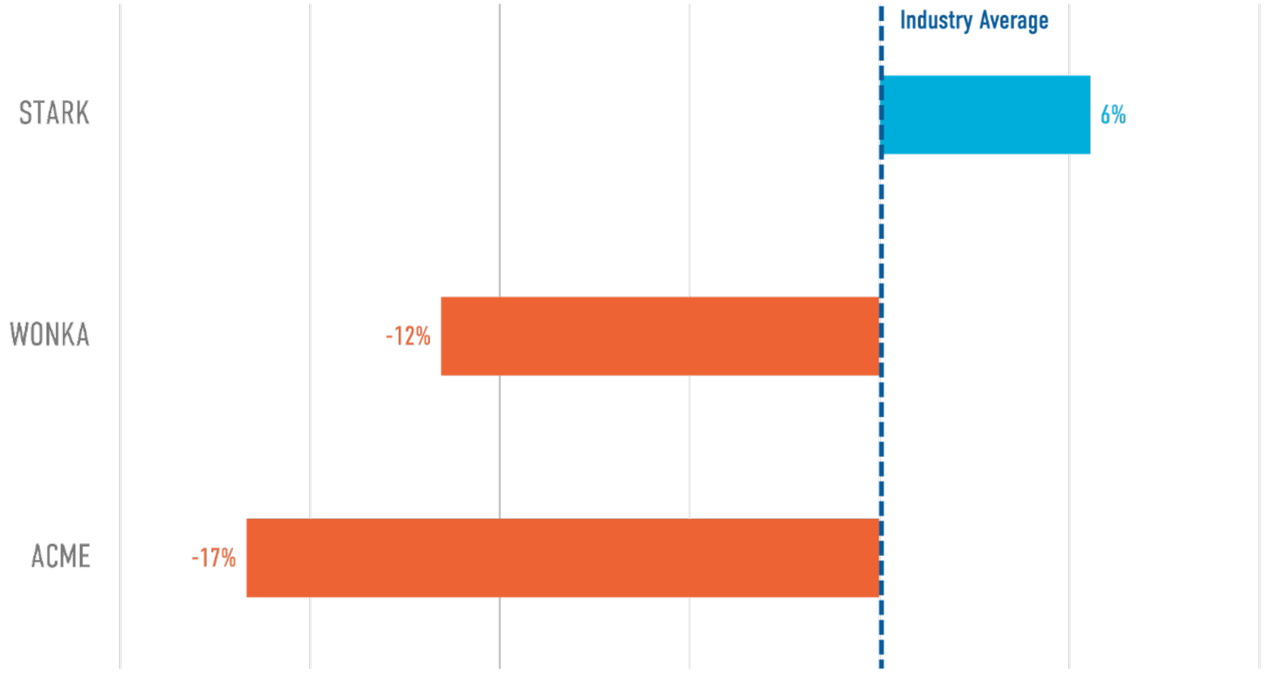
The Fair Pay Within Industry Score is designed to compare a company's pay to its industry peers', across all occupations. In other words, it assesses whether a company is paying more or less fairly on aggregate than its industry peers for similar roles.

To calculate the score, we compare the median salary for a given role at a company to the median salary for that role across the company's industry and compute the percent difference between the company and industry value. To get the final score, we then average all the calculated percent differences by role for the company.

The figure below provides a stepwise example of how the Fair Pay Within Industry Score would be derived for three fictional companies – ACME, WONKA, and STARK – within the same industry.

	ACME	WONKA	STARK	
Occupation	Median Salary			Industry Median
Cashier	$21,000	$31,500	$42,000	$31,500
Engineer	$75,000	$85,250	$65,750	$75,000
Sales Representative	$66,000	$45,000	$70,000	$66,000
Occupation	Percent Difference from Industry Median Salary			
Cashier	-50%	0%	25%	
Engineer	0%	12%	-14%	
Sales Representative	0%	-47%	6%	
	Average (Across Occupations)			
FAIR PAY WITHIN INDUSTRY SCORE	-17%	-12%	6%	

To interpret these results, we can assume that a value of 0% for the Fair Pay Within Industry Score means that the company's comparative pay across roles is identical to that of the typical company in its industry, so a higher positive value indicates fairer pay. In our example, fictional company STARK, on average, is paying 6% more fairly than the typical company in its industry, while WONKA and ACME are paying 12% and 17% less fairly on average.



HOW IS THIS METHODOLOGY DIFFERENT THAN PREVIOUS YEARS?

For company representatives who have previously engaged in our corporate review period, or for other researchers who have been following our work on wages, you'll notice that this method is a departure from how JUST Capital has previously modeled these three wage data points. We have rigorously vetted this new methodology and output values for each company, and there are four main enhancements we've made to improve data quality to note. This model now:

1. Leverages more company-specific underlying data and much greater transparency about its sources.
2. Utilizes machine learning modeling methods, as opposed to input-output models of prior years.
3. Increases our national living wage threshold, based on a robust feedback process we conducted with external experts that identified a larger family composition (two full-time workers and two children) than the one we used previously (one full-time worker, one part-time worker, and one child).
4. Simplifies the fair pay by industry score, moving away from average percentile ranks to average percent differences.

ABOUT JUST CAPITAL

JUST Capital, an independent, nonprofit organization, makes it easier for people, companies, and markets to do the right thing by tracking the business behaviors Americans care about most. Our research, rankings, indexes, and data-driven tools help people make more informed decisions about where to invest, work, and buy to direct capital towards companies advancing a more just future. America's Most JUST Companies, including the groundbreaking JUST 100, is published annually in Forbes and on JUSTCapital.com.

JUST Capital was co-founded in 2013 by a group of concerned people from the world of business, finance, and civil society – including Paul Tudor Jones II, Deepak Chopra, Rinaldo Brutoco, Arianna Huffington, Paul Scialla, and others. Our mission is to build a more just marketplace that better reflects the true priorities of the American people. We believe that business, and capitalism, can and must be a positive force for change. We believe that if they have the right information, people will buy from, invest in, work for, and otherwise support companies that align with their values. And we believe that business leaders are searching to win back the trust of the public in ways that go beyond money. By shifting the immense resources and ingenuity of the $15 trillion private sector onto a more balanced – and more just – course, we can help build a better future for everyone.

ABOUT REVELIO LABS

Revelio Labs provides workforce intelligence. We absorb and standardize hundreds of millions of public employment records to create the world's first universal HR database, allowing us to see current workforce composition and trends of any company. Our customers include investors, corporate strategists, HR teams, and governments.

For our main data products, we collect unstructured online public profiles, resumes, job postings, employee review data, layoff notices, and other publicly available employment data—and curate and structure this data through the use of proprietary algorithms. Our delivery data files provide data on employee movement, salary, and characteristics, which can be analyzed by positions, skills, geographies, and seniority levels over time.

INTERESTED IN LEARNING MORE?

If you would like to learn more about the methodology behind JUST Capital's and Revelio Labs' wage models or your company's performance in them, please reach out to our corporate engagement team at corpengage@justcapital.com.

If you would like more information or access to company workforce data, including salaries, headcounts, or employee composition to benchmark your company to your peers, please reach out to info@reveliolabs.com.

JUST capital

revelio labs