INTERNET APPENDIX FOR "NOISE IN EXPECTIONS: EVIDENCE FROM ANALYST FORECASTS"

Tim de Silva[1] David Thesmar[2]

Appendix A. Additional Derivations and Proofs

In this appendix, we provide derivations of the results stated in the main text. Recall that the indices t and h have been suppressed but that all variables implicitly vary across years t and forecasting horizon h.

Proof of Lemma 1. Defining $\varepsilon_i = \pi_i - E(\pi_i \mid X_i, Z_i)$, the first equation holds trivially. The conditional mean independence conditions, $E(\varepsilon_i \mid x_i, z_i) = 0$ and $E(z_i \mid x_i) = 0$, follow from the law of iterated expectations. Starting with the second equality, we have

$$\begin{aligned} E(z_i \mid x_i) &= E\left[E\left(z_i \mid x_i, X_i\right) \mid x_i\right] \\ &= E\left(E(z_i \mid X_i) \mid x_i\right). \end{aligned}$$

From the definition of z_i, we have $E(z_i \mid X_i) = 0$. Combined with the previous expression, this delivers the desired result. The proof for the second equality follows similarly from the law of iterated expectations.

$$\begin{aligned} E(\varepsilon_i \mid x_i, z_i) &= E\left[E\left(\varepsilon_i \mid x_i, X_i, z_i, Z_i\right) \mid x_i, z_i\right] \\ &= E\left[E\left(\varepsilon_i \mid X_i, Z_i\right) \mid x_i, z_i\right], \end{aligned}$$

where the inner expectation equals zero by the definition of ε_i, which delivers the desired result. □

Proof of Lemma 2. Defining $\eta_i = \pi_i - E(F^j \pi_i \mid X_i, Z_i)$, (3) follows trivially. The conditional mean independence condition, $E(\eta_{ij} \mid x_i, z_i) = 0$, follows from the law of iterated expectations.

$$\begin{aligned} E(\eta_{ij} \mid x_i, z_i) &= E\left[E\left(\eta_{ij} \mid x_i, X_i, z_i, Z_i\right) \mid x_i, z_i\right] \\ &= E\left[E\left(\eta_{ij} \mid X_i, Z_i\right) \mid x_i, z_i\right], \end{aligned}$$

where the inner expectation equals zero by the definition of η_{ij}, which delivers the desired result. □

Proof of Lemma 3. Using Lemma 1, we have $MSE^e = E\left[(x_i - \pi_i)^2\right] = E\left(z_i^2\right) + E(\varepsilon_i^2)$. Using Lemma 2 and the definition of consensus forecasts, we have $MSE^a = E\left[(F\pi_i - \pi_i)^2\right] = E\left(b_i^2\right) + E(\eta_i^2) + E(\varepsilon_i^2) + E(\eta_i \varepsilon_i)$. Under the assumption that η_{ij} and ε_i are uncorrelated, subtracting the previous expressions for MSE^e and MSE^a delivers the desired result. □

[1] MIT Sloan. Contact information: www.timdesilva.me, tdesilva@mit.edu.

[2] MIT Sloan, NBER, CEPR.

Proof of Proposition 1. To start, note that the first part of Assumption 1 and Lemma 2 imply

$$F_j \pi_i = g_i + \alpha z_i + \eta_{ij}, \quad g_i \equiv E\left(F\pi_i \mid X_i\right).$$

Taking averages across forecasters to arrive at consensus forecasts, we obtain

$$F\pi_i = g_i + \alpha z_i + \eta_i, \quad \eta_i \equiv \frac{1}{J_i}\sum_j \eta_{ij}.$$

Next, we can derive the noise in consensus forecasts:

$$\begin{aligned}
E\left(\eta_i^2\right) &= E\left[\left(\frac{1}{J_i}\sum_j \eta_{ij}\right)^2\right] \\
&= E\left[\frac{1}{J_i^2}\left[E\left(\sum_j \eta_{ij}\right)^2 \mid J_i\right]\right] \\
&= E\left[\frac{1}{J_i^2}E\left[\sum_j \eta_{ij}^2 + 2\sum_{j<k}\eta_{ij}\eta_{ik} \mid J_i\right]\right] \\
&= E\left[\frac{1}{J_i^2}\sum_j E\left(\eta_{ij}^2 \mid J_i\right)\right] \\
&= E\left[\frac{1}{J_i}\mathrm{var}\left(\eta_{ij}^2\right)\right] = E\left(\frac{1}{J_i}\right)\Sigma.
\end{aligned}$$

The first equality follows by definition, the second from the law of iterated expectations, the fourth by the third part of Assumption 1, and the fifth by the fourth part of Assumption 1. We can similarly derive the bias in consensus forecasts:

$$\begin{aligned}
E\left(b_i^2\right) &= E\left[\left(g_i + \alpha z_i - x_i - z_i\right)^2\right] \\
&= E\left[\left(g_i - x_i\right)^2\right] + (1-\alpha)^2 E\left(z_i^2\right) \\
&= \Delta + (1-\alpha)^2\Theta.
\end{aligned}$$

Combining the previous two results with Lemma 3 delivers the desired result. □

Proof of Proposition 2. The first part of Assumption 1 and Lemma 2 imply

$$F_j \pi_i = g_i + \alpha z_i + \eta_{ij}, \quad g_i \equiv E\left(F\pi_i \mid X_i\right),$$

which gives $F_{ij}^* = \alpha z_i + \eta_{ij}$. Taking variances and applying the orthogonality condition from Lemma 2 gives

$$\mathrm{var}\left(F_{ij}^*\right) = \mathrm{var}(\alpha z_i + \eta_{ij}) = \alpha^2\Theta + \Sigma,$$

delivering the second equation in the proposition. The third equation follows from the third part of Assumption 1:

$$\mathrm{cov}\left(F_{ij}^*, F_{ik}^*\right) = \mathrm{cov}\left(\alpha z_i + \eta_{ij}, \alpha z_i + \eta_{ik}\right) = \alpha^2\Theta.$$

Finally, the first equation follows from applying Lemma 1, Lemma 2, and the second part of Assumption 1:

$$\text{cov}(\pi_i^*, F_{ij}^*) = \text{cov}(z_i + \varepsilon_i, \alpha z_i + \eta_{ij}) = \alpha\Theta.$$

□

Proof of Proposition 3. The first part of Assumption 1 and Lemma 2 imply

$$F_t^j \pi_{it+h} = g_{it}^h + \alpha_h z_{it}^h + \eta_{ijt}^h, \quad g_{it}^h \equiv E\left(F_t \pi_{it+h} \mid X_{it}\right).$$

Then, by definition, $\overline{F}_t^j \pi_{it+h} = g_{it}^h + \alpha_h z_{it}^h$. Forecast revisions are then

$$F_t^j \pi_{it+h} - F_{t-1}^j \pi_{it+h} = \overline{F}_t^j \pi_{it+h} - \overline{F}_{t-1}^j \pi_{it+h} + \eta_{ijt}^h - \eta_{ijt-1}^h.$$

Taking variances, applying the definition of σ_{rev}^2 and $\overline{\sigma}_{rev}^2$, and using the assumption that noise is uncorrelated over time delivers the second equation in the proposition. Forecast errors are equal to

$$\pi_{it+h} - F_t^j \pi_{it+h} = x_{it}^h + z_{it}^h + \varepsilon_{it}^h - \overline{F}_t^j \pi_{it+h} - \eta_{ijt}^h.$$

The CG coefficient is then

$$\begin{aligned}
\beta_{CG} &= \frac{\text{cov}(\pi_{it+h} - F_t^j \pi_{it+h}, F_t^j \pi_{it+h} - F_{t-1}^j \pi_{it+h})}{\text{var}(F_t^j \pi_{it+h} - F_{t-1}^j \pi_{it+h})} \\
&= \frac{\text{cov}(x_{it}^h + z_{it}^h + \varepsilon_{it}^h - \overline{F}_t^j \pi_{it+h} - \eta_{ijt}^h, \overline{F}_t^j \pi_{it+h} - \overline{F}_{t-1}^j \pi_{it+h} + \eta_{ijt}^h - \eta_{ijt-1}^h)}{\sigma_{rev}^2} \\
&= \frac{\text{cov}(x_{it}^h + z_{it}^h + \varepsilon_{it}^h - \overline{F}_t^j \pi_{it+h}, \overline{F}_t^j \pi_{it+h} - \overline{F}_{t-1}^j \pi_{it+h})}{\sigma_{rev}^2} - \frac{\Sigma_h}{\sigma_{rev}^2} \\
&= \overline{\beta_{CG}} \left(\frac{\overline{\sigma}_{rev}^2 - \Sigma_h}{\sigma_{rev}^2} \right),
\end{aligned}$$

where

$$\overline{\beta_{CG}} \equiv \frac{\text{cov}(x_{it}^h + z_{it}^h + \varepsilon_{it}^h - \overline{F}_t^j \pi_{it+h}, \overline{F}_t^j \pi_{it+h} - \overline{F}_{t-1}^j \pi_{it+h})}{\overline{\sigma}_{rev}^2}.$$

Applying the equation derived for σ_{rev}^2 delivers the result. □

Proof of Proposition 4. From Proposition 1, we have

$$MSE^a = \Delta + (1-\alpha)^2 \Theta + \frac{1}{J}\Sigma + E(\varepsilon_i^2).$$

Under the assumption of joint normality,

$$\begin{aligned}
F^{e+a} \pi_i &= x_i + E\left(z_i \mid x_i, F\pi_i\right) \\
&= x_i + \frac{\text{cov}(z_i, \alpha z_i + \eta_i)}{\text{var}(\alpha z_i + \eta_i)} \left(F\pi_i - x_i\right)
\end{aligned}$$

$$= x_i + \beta\left[\alpha z + \eta_i\right], \quad \beta = \frac{\alpha\Theta}{\alpha^2\Theta + \frac{1}{J}\Sigma}.$$

This implies that $MSE^{e+a} = E(\varepsilon_i^2) + (1-\beta\alpha)^2\Theta + \beta^2\frac{1}{J}\Sigma$. Subtracting this from MSE^a delivers the result. □

Proof of Proposition 5. First note that combining Assumption 2 with the law of iterated expectations implies

$$\begin{aligned} E(\pi_i \mid X_i) \equiv x_i &= \left(1-\rho^{h-1}\right)\mu + \rho^{h-1}x_{it}, \\ &\equiv \left(1-\rho^{h-1}\right)\mu + \rho^{h-1}E(EPS_{it+1} \mid X_i). \end{aligned}$$

Therefore, at horizon h, the bias is

$$\begin{aligned} \Delta &= E\left[\left(E(EPS_{t+h}|X_i) - E(F_tEPS_{t+h}|X_i)\right)^2\right], \\ &= E\left[\left(\rho^{h-1}x_{it} - \rho^{h-1}E(F_tx_{it}|X_i)\right)^2\right] = \rho^{2(h-1)}E\left[\left(x_{it} - E(F_tx_{it}|X_i)\right)^2\right] = \rho^{2(h-1)}\Delta^1. \end{aligned}$$

The noise is

$$\begin{aligned} \Sigma &= var(\eta_{t,h}) = var\left(F_tEPS_{t+h} - E\left(F_tEPS_{t+h}|X_i, Z_i\right)\right), \\ &= var\left(F_tx_{t+h} - E(F_tx_{t+h}|X_i, Z_i)\right) = \rho^{2(h-1)}var(\eta_{t,1}) = \rho^{2(h-1)}\Sigma_\eta^1. \end{aligned}$$

The result follows because $\rho < 1$ by assumption. □

Appendix B. Additional Tables and Figures

Table A1. Variables Included in X_{it}

This table lists the set of variables that we include in X_{it}, which we use to form our econometric forecast. As described in Section 1.2, we include two lags of each variable. See Appendix C for a detailed discussion of how we use these variables.

Panel A: Collected from WRDS Financial Ratios
The following financial ratios: capei, be, bm, evm, pe_exi, pe_inc, ps, pcf, dpr, npm, opmbd, opmad, gpm, ptpm, cfm, roa, roe, roce, aftret_eq, aftret_invcapx, aftret_equity, preret_noa, pretret_earnat, GProf, equity_invcap, debt_invcap, totdebt_invcap, capital_ratio, int_totdebt, cash_lt, invt_act, rect_act, debt_at, debt_ebitda, short_debt, curr_debt, lt_debt, profit_lct, ocf_lct, cash_debt, fcf_ocf, lt_ppent, dltt_be, debt_assets, debt_capital, de_ratio, intcov, intcov_ratio, cash_ratio, quick_ratio, curr_ratio, cash_conversion, inv_turn, at_turn, rect_turn, pay_turn, sale_invcap, sale_equity, rd_sale, adv_sale, accrual, ptb, divyield

Panel B: Collected from CRSP
Two-digit SIC dummies, return in the month prior to fiscal year-end, cumulative return in the twelve months prior to fiscal year-end excluding the last month, trailing 5-year monthly return volatility (all returns adjusted for delisting), stock price on day of fiscal year-end

Panel C: Collected from Compustat
Natural log of total assets, dummies for year of fiscal report

Panel D: Collected from I/B/E/S
π_{it}, π_{it-1}, π_{it-2}, number of distinct analysts who issue forecasts in the 45 days following the release of the prior FY report

Table A2. Robustness of the Term Structure of Forecasting Accuracy: Contemporaneous Earnings–Yield

This table examines the robustness of the results in Table 2 and Table 4 to including an additional predictor: the earnings–yield at time t on the basis of the stock price at $t_R + 45$, where t_R is defined in Figure 1, calculated as $\frac{EPS_{it}}{P_{it_R+45}}$. Panel A contains the mean squared error of analyst forecasts in the first column, denoted MSE_h^a, and of our econometric forecasts, denoted MSE_h^e, across different forecasting horizons for forecasts of earnings yields, π_{it+h}. Panel B contains the mean squared error of analyst forecasts in the first column, denoted MSE_h^a, and of our econometrician + analyst forecasts, denoted MSE_h^{e+a}, across different forecasting horizons for forecasts of the realization of EPS at $t+h$ divided by price per share at t. The numbers reported in the table are normalized by the mean realization of π_{it+h}^2 at each horizon. In parentheses, we report the Diebold–Marino test statistics for testing the relative accuracy of the two forecasts under a squared loss function, where we calculate the asymptotic variance by performing a bootstrap at the year level with 1,000 iterations. The sample used in this table is slightly smaller than that in Table 1 due to the data restrictions imposed by the inclusion of this additional predictor.

Panel A: Analyst vs. Econometrician

	MSE_h^a	MSE_h^e	
Horizon: h	Analyst	Random Walk	Elastic Net
1 Quarters	4.54%	25.29%	20.5%
		(25.17)	(21.51)
2 Quarters	8.31%	29.25%	19.52%
		(23.07)	(18.83)
3 Quarters	12.94%	33.21%	21.94%
		(19.27)	(18.78)
4 Quarters	18.53%	24.74%	25.22%
		(8.95)	(12.38)

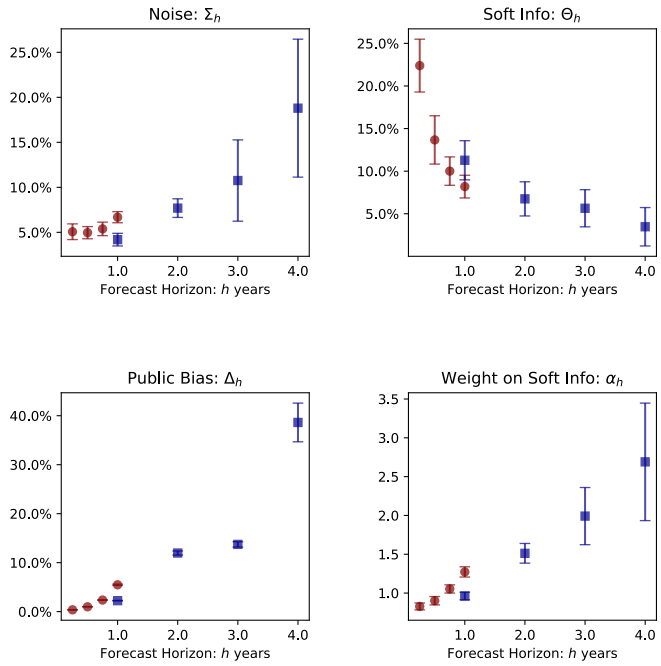
Panel B: Analyst vs. Econometrician + Analyst

	MSE_h^a	MSE_h^{e+a}
Horizon: h	Analyst	Elastic Net
1 Quarters	4.54%	4.5%
		(-2.8)
2 Quarters	8.31%	8.38%
		(1.66)
3 Quarters	12.94%	12.77%
		(-1.2)
4 Quarters	18.53%	17.53%
		(-3.96)

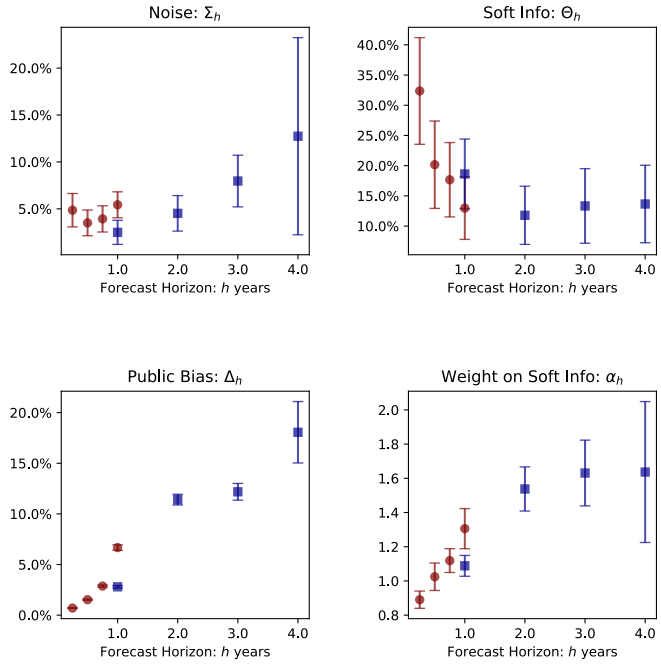
Figure A1. Estimates for Large and Small Firms Separately

We reproduce here Figure 2, separately for large and small firms. We first split the sample into large firms (above median assets) and small ones (below median). Then, we run our estimation procedure on both samples separately. Standard errors are generated through block bootstrap. Because this figure is here for illustrative purposes, we only did 20 bootstraps per panel.

Panel A: Small Firms



Panel B: Large Firms



Appendix C. Additional Details on Supervised Learning Techniques and Forecast Formation

C.1 Supervised Learning Techniques

This section provides a more detailed description of the supervised learning techniques that we explore.

Elastic net. The first estimator that we explore is elastic net, which is defined as follows for a given set of predictor variables X_{it}:

$$\mathcal{L}(\beta, \alpha_1, \alpha_2) \equiv \sum_i \left[\left(\pi_i - X_{it}' \beta \right)^2 \right] + \alpha_1 ||\beta||_1 + \alpha_2 ||\beta||_2,$$

$$\hat{\beta}^{\text{Lasso}} \equiv \arg\min_{\beta} \mathcal{L}(\beta, \alpha_1, 0), \quad \hat{\beta}^{\text{Ridge}}(\alpha_2) \equiv \arg\min_{\beta} \mathcal{L}(\beta, 0, \alpha_2),$$

$$\hat{\beta}^{\text{Elastic net}} \equiv \arg\min_{\beta} \mathcal{L}(\beta, \alpha_1, \alpha_2).$$

To choose the hyperparameters α_1 and α_2, we use cross-validation on the training set, detailed in Appendix C.2. Intuitively, cross-validation consists of breaking up the training sample into smaller datasets, fitting models on these smaller datasets, and examining which values of the hyperparameters generate the best performance on the other parts of the training set. Importantly, cross-validation is done entirely with the training set to avoid introducing any look-ahead bias.

Tree-based methods. We also consider two tree-based methods: random forest (RF) and gradient-boosted trees (GBT). The building block of tree-based estimators is regression trees, which are nonparametric regression estimators (unlike penalized linear estimators) designed to capture arbitrary nonlinearities among the variables in X_{it}.

We first describe regression trees, which are "grown" in sequential steps to approximate a function. The tree begins with an initial node containing all observations. Next, this initial node is split into two nodes: observations with $x_{it} < c$ and $x_{it} \geq c$. To make this split, the econometrician chooses the variable $x_{it} \in X_{it}$ and c to minimize the MSE. This process of splitting based on a chosen covariate and value continues with the two new subsamples until a terminal criterion is satisfied (e.g., an upper bound on the number of observations in each terminal node or the number of splits is reached). The final regression values are then the averages of the outcome variable across all of the observations remaining in each of the terminal nodes.

The process of growing a regression tree immediately illustrates the potential problem with such trees: they are likely to overfit (i.e., they have high prediction variance), especially if they grow extremely large. Without restrictions on the size of the tree, perfect in-sample fit could be achieved by having one observation in each terminal node, but this will perform terribly out of sample. To address this tendency to overfit, many "ensemble" methods have been developed, which combine several decision trees with a form of regularization to make more accurate out-of-sample predictions. The two tree-based methods that we consider, RF and GBT, are ensemble methods. The core idea behind RF and GBT is to grow many uncorrelated trees and then average their predictions.

RF is constructed based on the intuition of bootstrapping. On each bootstrapped sample, a regression tree is grown with a stopping criterion on the number of splits L with one adjustment: only a random subset of predictor variables is considered at each split.[3] These two steps are then repeated B times, generating B regression trees. Final predictions from the random forest are calculated by averaging predictions across the B regression trees. Averaging across many trees that have different structures due to the randomness in the subset of predictor variables chosen is the regularization in this method that limits overfitting and reduces prediction variance. Similarly to the penalized linear estimators, the two hyperparameters, $\{B, L\}$ can be chosen by means of cross-validation on the training set (see Appendix C.2 for details).

GBT starts by fitting a shallow tree of depth d and calculating the residuals from this regression tree. Then, a second shallow tree of depth d is fit on the residuals calculated from the first tree.[4] This shallow tree is likely to have terrible in-sample fit. To improve its fit, a second shallow tree of depth d is fit on the residuals calculated from the first tree. Predicted values are then formed by adding the predicted values from the two trees, shrinking the predicted values from the latter tree by a factor $\lambda \in (0, 1)$ (regularization). This procedure is repeated B times, after which the predicted value will be a combination of the predicted value from the first tree and the predicted values of the $B-1$ trees scaled by λ. The sequential growing of trees on (pseudo-)residuals from the previous trees makes the trees less correlated, which is why averaging over trees limits overfitting. This method has three hyperparameters, $\{B, d, \lambda\}$, which can be chosen by means of cross-validation on the training set (see Appendix C.2 for details).

C.2 Formation of Forecasts

This appendix describes the formation of our econometric and econometrician + analyst forecasts, including details on the implementation of our machine learning estimators. For expositional simplicity, we present the procedure as pseudo-code. Additional details on the implementation and cross-validation procedures are described at the end of this section.

Pseudo-code. To generate our econometric forecasts with elastic net at time t of π_i, the pseudo-code is as follows. For simplicity, denote as X_{it} the set of variables used in the econometric forecast and as $X_{it}^{e+a} = \{X_{it}, F_t\pi_{it+h}\}$ the set of variables used in the econometrician + analyst forecast. We describe our procedure below for our econometric forecasts, but an analogous procedure is used to form econometrician + analyst forecasts, where we replace X_{it} with X_{it}^{e+a}.

1. Start with the above dataset that contains X_{it} and π_i and $F\pi_i$ for each firm–year
2. Replace all missing values of variables measured at t with industry–time means and then fill all missing values at $t-1$ with values from t and likewise for $t-2$
3. Create year and 2-digit SIC code dummies
4. Initialize $s = 1995$

[3] If all variables are considered at each split, the procedure of forming many trees across bootstrapped samples is called bagging (i.e., bootstrap aggregation).

[4] Thinking of this procedure as operating on residuals from the trees conveys most of the intuition for why boosting works but is a technically incorrect description. Gradient-boosted trees are a particular form of boosting, where trees are successively fit on pseudo-residuals instead of residuals. Pseudo-residuals are defined as the gradient of the objective function, evaluated at each data point.

(a) Create a **training** dataset of observations indexed by i,s in the following set: $\{(i,t) : t \in \{s-5,...,s-1\}\}$

(b) Create a **test** dataset of observations indexed by i,t in the following set: $\{(i,t) : t = s\}$

(c) Trim all independent variables in the **training** dataset based on 5 times the interquartile range

(d) Trim all independent variables in the **test** dataset based on 5 times the interquartile range, with the interquartile range *calculated from the training set*

(e) Standardize all independent variables in the **training** set to have zero mean and unit variance

(f) Standardize all independent variables in the **test** based on means and variances *calculated from the training set*

(g) Fit a machine learning estimator that is one of the following on the training set, using cross-validation as described at the end of this section:

- Elastic net
- Random forest
- Gradient-boosted trees

(h) Generate forecasts on the test set. Calculating the MSE of these forecasts yields the MSEs for our three forecasts for year s.

(i) Stop if $s = 2021$; otherwise, set $s = s+1$ and continue back to (a)

Cross-validation and implementation details by estimator. We use the following cross-validation and implementation procedures for each machine learning algorithm on our training sets for each model. All procedures are implemented through the `sklearn` package in `Python 3.9`. We use default inputs to all `sklearn` functions mentioned below unless otherwise specified.

- Elastic net: We use 5-fold cross-validation on the training set, implemented with the `ElasticNetCV` function in `sklearn`. We search over a grid of the parameter `l1_ratio` $\in [0.1, 0.99]$, which corresponds to the ratio of the $\mathcal{L}^1$ to $\mathcal{L}^2$ penalty parameters.
- Random forest: We use 5-fold cross-validation on the training set, implemented with the `GridSearchCV` function for `RandomForestRegressor` in `sklearn`. We set `n_estimators` to 1000, corresponding to the number of decision trees in the ensemble, and search over the following grid for each parameter: `max_depth` $\in [4, 8]$, `max_features` $\in [0.3, 1]$, `min_samples_leaf` $\in [1, 5]$, and `min_samples_splits` $\in [2, 10]$. We use bootstrap samples for each decision tree. These parameter choices are similar to those in Gu et al. (2018) and Hansen and Thimsen (2020).
- Gradient-boosted trees: We use 5-fold cross-validation on the training set, implemented using the `GridSearchCV` function for `GradientBoostingRegressor` in `sklearn`. We search over the following grid for each parameter: `n_estimators` $\in [500, 10000]$, `max_depth` $\in [1, 3]$, and `learning_rate` $\in [0.001, 0.1]$. These parameter choices are similar to those in Gu et al. (2018).

Hardware. Rolling estimation with repeated cross-validation is computationally intensive. We parallelize each model estimation across 96 CPUs on the MIT SuperCloud server, with each estimation taking around 100 days of CPU time.

Appendix D. Interpretive Model for MSS Normalization

We write down here a simple model that gives a simple interpretation to the normalized MSE that we use throughout the paper. Take the perspective of a hypothetical agent who seeks to allocate capital across firms. We assume that investing k_i dollars in firm i eventually generates cash flows $\pi_i k_i - \frac{1}{2\gamma}k_i^2$. γ is a measure of returns to scale ($\gamma = \infty$ corresponds to constant returns to scale). This agent is risk neutral and therefore maximizes the sum of all expected cash flows:

$$\Pi_h = \sum_i \left(k_i F\pi_i - \frac{1}{2\gamma}k_i^2 \right)$$

where the expectation is taken using the agent's forecasting rule F. In this simple problem, capital allocation for firm i is $k_i = \gamma F\pi_i$. We compare this allocation to the perfect foresight allocation $k_i^{PF} = \gamma\pi_i$. Trivially, the perfect foresight allocation dominates all forecast-based allocations (including rational ones).

The expected cash flow loss relative to the perfect foresight allocation can then be written as:

$$\frac{\Pi_h^{PF} - \Pi_h^F}{\Pi_h^{PF}} = \frac{MSE_h^F}{MSS_h}, \tag{16}$$

where MSE_h^F is the MSE of the forecasting rule and MSS_h is the realized mean of π_i^2. Thus, (16) shows that normalizing the mean squared errors by the mean squared EPS can be interpreted as the percent allocative loss relative to a perfect foresight optimizer.

Appendix E. Additional Details on GMM Estimation

In this appendix, we discuss the details of our GMM estimation based on the following moment conditions from Proposition 2:

$$\begin{aligned} E\left(\pi_i^* F_{ij}^*\right) &= \alpha\Theta, \\ E\left[\left(F_{ij}^*\right)^2\right] &= \alpha^2\Theta + \Sigma, \\ E\left(F_{ij}^* F_{ik}^* \mid j \neq k\right) &= \alpha^2\Theta. \end{aligned}$$

We have now replaced the covariances with second moments given that all variables are mean zero. The computation of these expectations requires further clarification, given that the third moment varies at a different level than the first two.

We start with an i–t–j panel of individual analyst forecasts for each firm–year discussed in Section 1.1. Denote the size of this dataset as N_0. We then compute all possible interactions between the forecasts of the J_{it} analysts following each firm, resulting in $\binom{J_{it}}{2}$ interactions per i–t. We use this set of interactions for each i–t to construct an i–t–j–k panel that contains the forecasts of analysts j and k and their interaction for all possible j–k pairs. Denote the size of this dataset as N_1. We denote sample expectations taken on this i–t–j–k panel as $\widehat{E}$.

The score vector that we use in our GMM estimation (making t explicit for clarity) is:

$$m(\pi_{it}^*, F_{itj}^*, F_{itk}^*; \alpha, \Theta, \Sigma) = \begin{pmatrix} \frac{N_1}{N_0}\pi_{it}^* F_{itj}^* - \alpha\Theta \\ 1(j=k)\frac{N_1}{N_0}F_{itj}^* F_{itk}^* - \alpha^2\Theta - \Sigma \\ 1(j\neq k)\frac{N_1}{N_1 - N_0}F_{itj}^* F_{itk}^* - \alpha^2\Theta \end{pmatrix}.$$

Note that there is reweighting based on N_1 and N_0. This reweighting ensures that taking the expectation of this score vector on an i–t–j–k panel delivers the same result that we would obtain by calculating these moments on an i–t–j panel. However, we cannot use an i–t–j panel for an estimation because performing GMM requires the sample expectations of each moment condition to be calculated as the sample averages on the same dataset.

Our final step is to generate our parameter estimates by solving:

$$\left(\hat{\alpha}, \hat{\Theta}, \hat{\Sigma}\right) = \underset{\alpha,\Theta,\Sigma}{\arg\min}\, \widehat{E}\left[m(\pi_i^*, F_{itj}^*, F_{itk}^*; \alpha, \Theta, \Sigma)\right]' \widehat{E}\left[m(\pi_i^*, F_{itj}^*, F_{itk}^*; \alpha, \Theta, \Sigma)\right].$$

We solve this optimization problem using a global basinhopping algorithm implemented in `SciPy`. As we did with our MSEs, we normalize by the average squared EPS (calculated on the basis of an unweighted average on the i–t–j panel). We perform this estimation procedure separately for each h.

To calculate standard errors for the three parameters estimated via GMM, we perform a clustered bootstrap at the firm level to account for autocorrelation of earnings and correlation of forecasts across analysts and then re-estimate the GMM on each bootstrap sample.[5] For Δ, we again compute standard errors using a firm-level bootstrap. These

[5] We use the bootstrap procedure from **?**, in which the GMM objective function is not recentered before estimation on each bootstrap sample. This generates standard errors with valid coverage (asymptotically) even under model misspecification.

standard errors are likely to be too tight because they ignore sampling uncertainty in the predictions generated from our machine learning estimators, effectively treating them as raw data that is resampled directly to compute moments. Here, we are constrained by the lack of asymptotic results that characterize the behavior of our statistical learning estimators in large samples. This approach is standard in the literature (e.g., **?**).[6]

Weighting in the Table 3 MSE decomposition. Because of the weighting in our GMM, our resulting parameter estimates equally weight all firm–year–analyst observations (i.e., firms with more analysts will be weighted more). However, the MSEs reported in Table 2 and Table 4 equally weight all firm–year observations because they vary only at the firm–year level. To address this issue, we recompute the expectations required to calculate MSE^a, MSE^a, and $\frac{1}{J}$ on the i–t–j panel used for GMM estimation when reporting their values in Table 3.

[6] We could in principle bootstrap the data and re-estimate our ML models, but this is too computationally intensive. Moreover, it is not clear that every estimator satisfies the regularity conditions required for the bootstrap to be asymptotically valid (**?**).

Appendix F. Afrouzi et al. (2021) Model

In this section, we apply the model proposed by Afrouzi et al. (2021) to our setting. We first briefly describe it, referring the reader to Afrouzi et al. (2021) Section 5 for additional details. We then show that it qualitatively delivers upward-sloping term structures of bias and noise but fails quantitatively in our setting for an intuitive reason.

F.1 Model Description

Consider the same setup as in Section 4.1, including Assumption 2. In addition, assume further that $u_{it}^x \sim \mathcal{N}\left(0, \sigma_u^2\right)$ and allow the mean of x_{it}, denoted μ, to be different from zero. At time t, the analyst needs to form forecasts of EPS_{it+h} for $h \geq 1$. The analyst costlessly observes x_{it-1}, z_{it}^h, and ρ. Importantly, the analyst does not know μ but has the prior $\mu | x_{it-1} \sim \mathcal{N}(x_{it-1}, \underline{\tau})$. The analyst also knows the DGP for EPS_{it}. Because x_{it}^h and z_{it}^h are orthogonal, we assume that the analyst forms forecasts of EPS_{it+h} by first forming optimal forecasts of x_{it+h-1} and z_{it}^h individually and then adding them together. Denote her forecasts by F_t. We assume $\alpha = 1$ for simplicity to focus on forecasts of x_{it}, which we show fail to match the data.

To form $F_t x_{it+h}$, the analyst has access to two sources of information. First, the agent costlessly observes x_{it-1}, which is "at the top of her mind". Secondly, the agent can retrieve additional information from past data to learn about μ, but doing so is costly. She chooses the level of information acquisition to minimize the expected squared error of her forecasts, subject to the cost of information retrieval. Formally, the analyst solves the following problem:

$$\min_{S_{it}} E\left[\min_{F_t x_{it+h}} E\left[\left(F_t x_{it+h} - x_{it+h}\right)^2 | S_{it}\right] + C(S_{it})\right], \tag{17}$$
$$s.t. \quad \{x_{it-1}\} \in S_{it} \subset \mathcal{S}_{it}, \quad \mathcal{S}_{it} = \{s : s \perp \mu | x_{it-1}, x_{it-2}, \ldots\}.$$

The solution to the inner maximization problem is $F_t x_{it+h} = E(x_{it+h} | S_{it})$. As shown in Afrouzi et al. (2021), the assumption that u_{it} is normally distributed implies that the outer maximization problem can be reduced to choosing a belief prediction about μ, denoted τ. Denote the solution to this problem as $\tau^*(h)$.

To further characterize the solution to this problem, Afrouzi et al. (2021) assume that the cost of information retrieval takes the following form:

$$C_t(S_t) \equiv \omega \frac{\exp\left(2\ln(2) \cdot \gamma \cdot \mathbb{I}(S_t, \mu \mid x_{it-1})\right) - 1}{\gamma}.$$

In this expression, $\mathbb{I}(\cdot,\cdot)$ denotes Shannon mutual information, $\omega \geq 0$ governs the overall cost of retrieval, and $\gamma \geq 0$ measures the convexity of the cost function in mutual information. Given this cost function, Afrouzi et al. (2021) show that the choice of belief precision, τ, that solves (17) is:

$$\tau^*(h) = \underline{\tau} \max\left\{1, \left(\frac{\left(1 - \rho^{h+1}\right)^2}{\omega \underline{\tau}}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{18}$$

This choice of $\tau = \tau^*(h)$ implies that the analyst's forecast is the following:

$$F_t x_{it+h} = (1-\rho^{h+1})\left(1-\frac{\underline{\tau}}{\tau^*(h)}\right)\mu + \left(\rho^{h+1} + (1-\rho^{h+1})\frac{\underline{\tau}}{\tau^*(h)}\right)x_{it-1} + \eta_{it}^h,$$
$$\eta_{it}^h \sim \mathcal{N}\left(0, (1-\rho^{h+1})^2 \frac{1}{\tau^*(h)}\left(1-\frac{\underline{\tau}}{\tau^*(h)}\right)\right).$$

Note that $F_t x_{it+h}$ is a random variable because of the expectation noise, η_{it}^h.

F.2 Term Structures of Bias and Noise

Now that we have characterized the analyst's forecasts, we can determine the term structures of bias and noise in this model.

Proposition F1. *In the Afrouzi et al. (2021) model, the term structures of bias and noise are given by:*

$$\Delta_h = \left[(1-\rho^h)\frac{\underline{\tau}}{\tau^*(h-1)}\right]^2 \frac{\sigma_u^2}{1-\rho^2}, \qquad \Sigma_h = \left(1-\rho^h\right)^2 \frac{1}{\tau^*(h-1)}\left(1-\frac{\underline{\tau}}{\tau^*(h-1)}\right).$$

Thus, the Afrouzi et al. (2021) model produces upward-sloping term structures of bias and noise.

Proposition F1 shows that noise in increasing with the horizon because of $\tau^*(h-1)$ increasing with the horizon (see eq. (18)). The intuition here is that at longer horizons, it is more useful to know the long-run mean, so the analyst engages in more information retrieval. This greater retrieval provides the analyst with more noisy information, creating large noise in her forecasts.

Additionally, the formula for Δ_h in Proposition F1 shows that bias is increasing with the horizon since $\tau^*(h-1)$ is increasing in h at a slower rate than $1-\rho^h$. This upward term structure of bias results from a combination of two effects. On the one hand, the agent engages in more retrieval, moving her forecast closer to the conditional expectation. On the other hand, the analyst still overweights x_{it-1} in her estimate of the long-run mean, and this overweighting is magnified at longer horizons because the analyst's forecast moves closer to her subjective expectation of the long-run mean. This second effect dominates, generating an upward-sloping term structure of bias.

In sum, the Afrouzi et al. (2021) model generates upward-sloping term structures of bias and noise. We now discuss how this model cannot *quantitatively* fit the data.

F.3 Quantitative Model Fit

The Afrouzi et al. (2021) model has 5 parameters: ρ, σ_u, ω, γ, and $\underline{\tau}$. The key endogenous parameter is $\tau^*(h)$. Rewriting the equation for Δ_h in Proposition F1, we obtain an expression for $\tau^*(h)$ as a function of the bias at horizon

h:

$$\frac{\tau^*(h-1)}{\underline{\tau}} = \frac{\left(1-\rho^h\right)\sigma_u}{\sqrt{\Delta_h\left(1-\rho^2\right)}}. \tag{19}$$

Equation (19) is useful because it express $\frac{\tau^*(h-1)}{\underline{\tau}}$, which is crucial to this model as a function of data moments that we have already estimated (or can estimate).

We focus on annual forecast horizons and estimate $\rho = 0.876$ and $\sigma_u = 0.010$ by regressing $F_t\pi_{it+2}$ on $F_t\pi_{it+1}$. Using our values of Δ_h from Figure 2, we can evaluate (19) at each horizon:

$$\frac{\left(1-\rho^1\right)\sigma_u}{\sqrt{\Delta_1\left(1-\rho^2\right)}} \approx 0.32, \quad \frac{\left(1-\rho^2\right)\sigma_u}{\sqrt{\Delta_2\left(1-\rho^2\right)}} \approx 0.27, \quad \frac{\left(1-\rho^3\right)\sigma_u}{\sqrt{\Delta_3\left(1-\rho^2\right)}} \approx 0.30.$$

By eq. (19), we obtain

$$\frac{\tau^*(1-1)}{\underline{\tau}} \approx 0.32, \quad \frac{\tau^*(2-1)}{\underline{\tau}} \approx 0.27, \frac{\tau^*(3-1)}{\underline{\tau}} \approx 0.30.$$

However, this contradicts eq. (18), which shows that $\forall h$, $\frac{\tau^*(h)}{\underline{\tau}} \geq 1$. Thus, these numerical results illustrate that the Afrouzi et al. (2021) model cannot match our data because it cannot generate public information bias.

The intuition of the quantitative failure of the Afrouzi et al. (2021) model in matching our data is the following. From Proposition F1, the public information bias at horizon h is increasing in σ_u: as $\sigma_u \to 0$, there will be no public information bias because x_{it} will be a deterministic process. In the data, we in fact find that σ_u is quite low. However, despite a low σ_u, we still find substantial forecasting bias—the public information bias is so high that this model would require the analyst to *forget* information to match this level of bias. The heart of this problem is that the model gives the analyst access to x_{it-1}, which is very close to x_{it+h} because σ_u is low. In other words, the fact that the analyst gets to see the machine forecast from the last period, $F^m_{t-1}EPS_{it} = x_{it-1}$, gives her too much knowledge for her forecasts to be as biased (and as noisy) as we find in the data.

References

Afrouzi, Hassan, Spencer Kwon, Augustin Landier, Yueran Ma, and David Thesmar (2021), "New Experimental Evidence on Expectation Formation." *Working Paper*, 1–67.

Angeletos, George-Marios, Zhen Huo, and Karthik A. Sastry (2020), "Imperfect Macroeconomic Expectations: Evidence and Theory." *NBER Macroeconomics Annual*.

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki (2020), "The Impact of Big Data on Firm Performance: An Empirical Investigation." *Working Paper*.

Ball, Ryan T. and Eric Ghysels (2018), "Automated Earnings Forecasts: Beat Analysts or Combine and Conquer?" *Management Science*, 64, 4936–4952.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2011), "Inference for high-dimensional sparse econometric models." *Advances in Economics and Econometrics: Tenth World Congress Volume 3, Econometrics*, 245–295.

Bergman, Peter, Danielle Li, and Lindsey Raymond (2020), "Hiring as Exploration." *Working Paper*.

Bianchi, Francesco, Sydney C. Ludvigson, and Sai Ma (2020), "Belief Distortions and Macroeconomic Fluctuations." *Working Paper*.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer (2016), "Stereotypes." *The Quarterly Journal of Economics*, 1753–1794.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer (2019), "Diagnostic Expectations and Stock Returns." *Journal of Finance*, 74, 2839–2874.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer (2022), "Expectations of Fundamentals and Stock Market Puzzles." *Working Paper*.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer (2020), "Overreaction in Macroeconomic Expectations." *American Economic Review*, 110, 2748–2782.

Bouchaud, Jean Philippe, Philipp Krüger, Augustin Landier, and David Thesmar (2019), "Sticky Expectations and the Profitability Anomaly." *Journal of Finance*, 74, 639–674.

Bradshaw, Mark T., Michael S. Drake, James N. Myers, and Linda A. Myers (2012), "A re-examination of analysts' superiority over time-series forecasts of annual earnings." *Review of Accounting Studies*, 69–76.

Brown, Lawrence D., Andrew C. Call, Michael B. Clement, and Nathan Y. Sharp (2015), "Inside the "Black Box" of sell-side financial analysts." *Journal of Accounting Research*, 53, 1–47.

Brown, Lawrence D. and Michael S Rozeff (1978), "The Superiority of Analyst Forecasts as Measures of Expectations: Evidence from Earnings." *Journal of Finance*, 33, 1–16.

Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard (2020), "Bayesian Solutions for the Factor Zoo: We Just Ran Two Quadrillion Models." *Working Paper*.

Cao, Kai and Haifeng You (2020), "Fundamental Analysis via Machine Learning." *Working Paper*.

Cassella, Stefano, Benjamin Golez, Huseyin Gulen, and Peter Kelly (2023), "Horizon Bias and the Term Structure of Equity Returns." *Review of Financial Studies*, 36, 1253–1288.

Chen, Qi and Wei Jiang (2006), "Analysts' weighting of private and public information." *Review of Financial Studies*, 19, 319–355.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney K. Newey, and James M. Robins (2016), "Double/Debiased Machine Learning for Treatment and Causal Parameters." *Working Paper*.

Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov (2020), "On cross-validated Lasso in high dimensions." *Annals of Statistics*, 40.

Coibion, Olivier and Yuriy Gorodnichenko (2015), "Information rigidity and the expectations formation process: A simple framework and new facts." *American Economic Review*, 105, 2644–2678.

Daniel, Kent, Avanidhar Subrahmanyam, and David A. Hirshleifer (1998), "Investor Psychology and Security Market Under and Overreactions." *Journal of Finance*, 53, 1839–1885.

D'Arienzo, Daniele (2020), "Maturity Increasing Over-reaction and Bond Market Puzzles." *Working Paper*.

De la O, Ricardo and Sean Myers (2021), "Subjective Cash Flow and Discount Rate Expectations." *Journal of Finance*, 76, 1339–1387.

Dessaint, Olivier, Thierry Foucault, and Laurent Frésard (2020), "Does Big Data Improve Financial Forecasting? The Horizon Effect." *Working Paper*.

Enke, Benjamin and Thomas Graeber (2020), "Cognitive Uncertainty." *Working Paper*.

Eyster, Erik, Matthew Rabin, and Dimitri Vayanos (2019), "Financial Markets Where Traders Neglect the Informational Content of Prices." *Journal of Finance*, 74, 371–399.

Fuster, Andreas, David Laibson, and Brock Mendel (2010), "Natural Expectations and Macroeconomic Fluctuations." *Journal of Economic Perspectives*, 24, 67–84.

Gabaix, Xavier (2014), "A Sparsity-Based Model of Bounded Rationality." *The Quarterly Journal of Economics*, 1661–1710.

Gabaix, Xavier and David Laibson (2017), "Myopia and Discounting." *Working Paper*, 1–43.

Gershman, Samuel J. and Rahul Bhui (2020), "Rationally inattentive intertemporal choice." *Nature Communications*, 11.

Giglio, Stefano and Bryan T. Kelly (2018), "Excess volatility: Beyond discount rates." *The Quarterly Journal of Economics*, 133, 71–127.

Greenwood, Robin and Andrei Shleifer (2014), "Expectations of returns and expected returns." *Review of Financial Studies*, 27, 714–746.

Gu, Shihao, Bryan T. Kelly, and Dacheng Xiu (2018), "Empirical Asset Pricing via Machine Learning."

Hansen, Jorge W and Christoffer Thimsen (2020), "Forecasting Corporate Earnings with Machine Learning." *Working Paper*.

Harford, Jarrad, Feng Jiang, Rong Wang, and Fei Xie (2019), "Analyst career concerns, effort allocation, and firms' information environment." *Review of Financial Studies*, 32, 2179–2224.

Juodis, Artūras and Simas Kucinskas (2019), "Quantifying Noise." *SSRN Electronic Journal*, 2019, 1–61.

Kacperczyk, Marcin, Stijn Van Nieuwerburgh, and Laura Veldkamp (2016), "A Rational Theory of Mutual Funds' Attention Allocation." *Econometrica*, 84, 571–626.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein (2021), *Noise*. Little, Brown Spark, New York, Boston, and London.

Khaw, Mel Win, Ziang Li, and Michael Woodford (2019), "Cognitive Imprecision and Small-Stakes Risk Aversion." *Working Paper*.

Kothari, S. P., Eric C. So, and Rodrigo Verdi (2016), "Analysts' Forecasts and Asset Pricing: A Survey." *Annual Review of Financial Economics*, 1–23.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh (2020), "Shrinking the cross-section." *Journal of Financial Economics*, 135, 271–292.

Kumar, Alok, Ville Rantala, and Ruoxi Xu (2021), "Social Learning and Analyst Behavior." *Journal of Financial Economics*.

Maćkowiak, Bartosz and Mirko Wiederholt (2009), "Optimal sticky prices under rational inattention." *American Economic Review*, 993, 769–803.

Mankiw, N. Gregory and Ricardo Reis (2002), "Sticky information versus sticky prices: A proposal to replace the new Keynesian Phillips curve." *Quarterly Journal of Economics*, 117, 1295–1328.

Manski, Charles F. (2017), "Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise." *NBER Macroeconomics Annual*.

Mullainathan, Sendhil and Jann Spiess (2017), "Machine learning: An applied econometric approach." *Journal of Economic Perspectives*, 31, 87–106.

Nagel, Stefan (2021), *Machine Learning in Asset Pricing*. Princeton University Press, Princeton and Oxford.

Patton, Andrew J. and Allan Timmermann (2010), "Why do forecasters disagree? Lessons from the term structure of cross-sectional dispersion." *Journal of Monetary Economics*, 57, 803–820.

Satopää, Ville, Marat Salikhov, Philip E. Tetlock, and Barb Mellers (2020), "Bias, Information, Noise: The BIN Model of Forecasting." *Working Paper*.

Schmidt-Hieber, Johannes (2020), "Nonparametric regression using deep neural networks with relu activation function." *Annals of Statistics*, 48, 1875–1897.

Sims, Christopher A. (2003), "Implications of rational inattention." *Journal of Monetary Economics*, 50, 665–690.

So, Eric C. (2013), "A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts?" *Journal of Financial Economics*, 108, 615–640.

van Binsbergen, Jules H., Xiao Han, and Alejandro Lopez-Lira (2020), "Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases." *Working Paper*.

Vuolteenaho, Tuomo (2002), "What drives firm-level stock returns?" *Journal of Finance*, 57, 233–264.

Wager, Stefan and Susan C. Athey (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113, 1228–1242.

Woodford, Michael (2003), "Imperfect Common Knowledge and Monetary Policy." In *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, 25–58, Princeton University Press, Princeton, NJ.

Woodford, Michael (2020), "Modeling imprecision in perception, valuation, and choice." *Annual Review of Economics*, 12, 579–601.