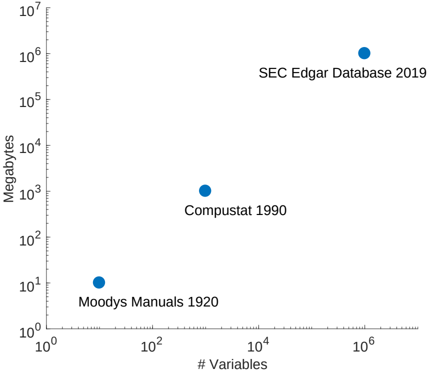
Market Efficiency in the Age of Big Data

Ian Martin   Stefan Nagel

February 2021

Investors' Big Data problem

- Investors forecasting cash flows face huge number of potential predictors
- $\mathbb{E}$ [c.f.] = f(predictors) unknown:
  high-dimensional learning problem
- Consequences for asset pricing? Market efficiency tests? Estimation of risk premia?



High-dimensional learning in asset pricing

- Standard approaches in asset pricing and market efficiency testing assume rational expectations (RE)
  - Assumes away learning problem: investors know $f(.)$ in
    $\mathbb{E}$ [c.f.] $= f$(predictors)
  - Motivates in-sample (IS) tests of "market efficiency":
    IS return predictability = risk premium/mispricing

- We show: when investors learn about $f(.)$ in high-dimensional setting, equilibrium asset prices are such that
  - IS return predictability $\neq$ risk premium/mispricing
  - OOS return predictability = risk premium/mispricing

- Not about econometric problems with predictability tests, but about properties of equilibrium asset prices

Roadmap

Two steps:

1. Investors learn about parameters of cash-flow generating model and price assets accordingly

2. Econometrician analyzes equilibrium prices ex-post using standard return predictability tests
    - Properties of IS tests
    - Properties of OOS tests

1. Asset pricing

Setup

- Economy with N assets, $N \times J$ firm characteristics $\boldsymbol{X}$, $J < N$
- Dividends $\boldsymbol{y}_t$ with growth cross-sectionally predictable based on $\boldsymbol{X}$:
$$\Delta \boldsymbol{y}_t = \boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_t, \quad \boldsymbol{e}_t \sim N(\boldsymbol{0}, \boldsymbol{I}),$$
with normalization $\frac{1}{NJ} \operatorname{tr} \boldsymbol{X}'\boldsymbol{X} = 1$.
- Investors are homogeneous, risk-neutral, and the interest rate is zero.
- Dividend strips: $\boldsymbol{p}_t$ = prices at t of claims to $\boldsymbol{y}_{t+1}$
  - Think: one period $\approx$ one decade

Rational expectations: No return predictability

- RE: investors know value of $\boldsymbol{g}$ in $\Delta \boldsymbol{y}_t = \boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_t$
- Prices

$$\boldsymbol{p}_t = \mathbb{E}_t\, \boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \boldsymbol{X}\boldsymbol{g}.$$

- Dividend strip returns

$$\boldsymbol{r}_{t+1} = \boldsymbol{y}_{t+1} - \boldsymbol{p}_t = \boldsymbol{e}_{t+1}$$

  i.e., unpredictable.

- This is the usual null hypothesis that underlies market efficiency tests, orthogonality conditions, Euler equations.
- But that investors have precise knowledge of $\boldsymbol{g}$ is implausible, especially if J is large $\Rightarrow$ Learning.

Bayesian pricing framework: Prior beliefs

- Before seeing data, investors hold informed prior beliefs

$$\boldsymbol{g} \sim N\left(\mathbf{0}, \frac{\theta}{J}\boldsymbol{I}\right), \qquad \theta > 0$$

  - Proportionality of prior covariance matrix to $\boldsymbol{I}$: w.l.o.g, can always rotate and rescale $\boldsymbol{X}$ to make it hold
  - Variance of the elements of $\boldsymbol{g}$ decline with J: ensures that variance of predictable cash flow growth does not explode when $N, J \to \infty$

- Investors then learn about $\boldsymbol{g}$ by observing $\boldsymbol{X}$ and history $\{\Delta \boldsymbol{y}_s\}_1^t$, summarized by sample average $\overline{\Delta \boldsymbol{y}}_t$.

Bayesian pricing framework: Posterior mean

- Posterior mean is a ridge regression estimator

$$\tilde{\boldsymbol{g}}_t = \boldsymbol{\Gamma}_t (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\overline{\Delta \boldsymbol{y}}_t,$$

  i.e., OLS estimator shrunk towards prior mean of zeros by matrix

$$\boldsymbol{\Gamma}_t = \boldsymbol{Q}\left(\boldsymbol{I} + \frac{J}{N\theta t}\boldsymbol{\Lambda}^{-1}\right)^{-1}\boldsymbol{Q}'$$

  where $\boldsymbol{Q}$, $\boldsymbol{\Lambda}$ from eigendecomposition $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$.

- Shrinkage strong
  - if t small (short time dimension)
  - if θ small (prior tightly concentrated around zero)
  - if J/N is large (large # of predictors)
  - along principal components of $\boldsymbol{X}'\boldsymbol{X}$ with small eigenvalues

Equilibrium realized returns

Proposition

With assets priced based on $\tilde{\boldsymbol{g}}_t$, realized returns are

$$\boldsymbol{r}_{t+1} = \boldsymbol{y}_{t+1} - \boldsymbol{p}_t = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{g} - \boldsymbol{X}\boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\bar{\boldsymbol{e}}_t + \boldsymbol{e}_{t+1}$$

where $\bar{\boldsymbol{e}}_t = \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{e}_s$.

- "underreaction" to $\boldsymbol{X}$ due to shrinkage
- "overreaction" to estimation error in $\tilde{\boldsymbol{g}}_t$, dampened by shrinkage
- unpredictable shock (the only term in RE case)

2. Properties of return predictability tests

In-sample predictability test

- Econometrician cross-sectionally regresses (OLS)

$$\boldsymbol{r}_{t+1} = \underbrace{\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{g}}_{\text{“underreaction”}} - \underbrace{\boldsymbol{X}\boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\bar{\boldsymbol{e}}_t}_{\text{“overreaction”}} + \underbrace{\boldsymbol{e}_{t+1}}_{\text{RE}}$$

  on characteristics matrix $\boldsymbol{X}$ and obtains coefficients

$$\boldsymbol{h}_{t+1} = (\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{g} - \boldsymbol{\Gamma}_t \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\bar{\boldsymbol{e}}_t + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{e}_{t+1}.$$

- Is econometrician likely to find $\boldsymbol{h}_{t+1}$ (jointly) significantly different from zero?
- High-dimensional asymptotics: $N, J \to \infty$, w/ $J/N \to \psi > 0$
  - $\boldsymbol{g}$ still uncertain in investors' minds even with large N
  - Change in N, J alters not only sampling properties of econometric tests, but also investors' learning problem and equilibrium asset prices

In-sample predictability test: RE null

- Consider the return predictability test statistic

$$T_{re} \equiv \frac{\boldsymbol{h}_{t+1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} - J}{\sqrt{2J}}.$$

- Standard approach takes RE as null hypothesis, which implies

$$\boldsymbol{h}_{t+1} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}_{t+1}$$

and

$$T_{re} \xrightarrow{d} N(0,1) \quad \text{as } N, J \to \infty,\ J/N \to \psi > 0.$$

- Properties of test based on T_{re} when econometrician applies it to returns from the Bayesian learning economy where RE null false?
  - Assumption of objectively correct prior: $\boldsymbol{g}$ drawn from the prior distribution

In-sample predictability test: Investor learning

Proposition

The test statistic T_{re} satisfies

$$\frac{T_{re}}{\sqrt{\mu^2+\sigma^2}} - \frac{\mu-1}{\sqrt{2\left(\mu^2+\sigma^2\right)}}\sqrt{J} \xrightarrow{d} N(0,1)$$

where $1 < \mu < 2$ and $1 < \sqrt{\mu^2+\sigma^2} < 2$.

- Therefore,

$$T_{re} \approx \sqrt{\mu^2+\sigma^2}\, N(0,1) + \frac{\mu-1}{\sqrt{2}}\sqrt{J}$$

- Growth of second term with $\sqrt{J}$: prob(rejection) $\to 1$ rapidly as $N, J \to \infty$
  - prob(no rejection) $\to 0$ exponentially fast as $N, J \to \infty$

In-sample trading strategy return

- Consider a characteristics-based trading strategy with weights

$$\boldsymbol{w}_{IS,t} = \frac{1}{N}\boldsymbol{X}\boldsymbol{h}_{t+1}, \qquad r_{IS,t+1} = \boldsymbol{w}'_{IS,t}\boldsymbol{r}_{t+1}$$

  which is IS because $\boldsymbol{h}_{t+1}$ estimated using returns $\boldsymbol{r}_{t+1}$.

- Proposition

$$\lim_{N,J\to\infty, J/N\to\psi} \mathbb{E}\, r_{IS,t+1} = \psi\mu > 0\,,$$

  and the Sharpe ratio $SR_{IS} = \mathbb{E}\, r_{IS,t+1} / \operatorname{var}(r_{IS,t+1})^{1/2}$ *grows at rate* $\sqrt{N}$.

- Reflects look-ahead advantage of econometrician relative to investor in IS regressions in "Big Data" setting $J/N \to \psi > 0$

(Absence of) out-of-sample return predictability

- The econometrician uses $N \times P$ predictors in $\boldsymbol{C}$, with $P \leq J$
  - correlated with those in $\boldsymbol{X}$, but need not fully span $\boldsymbol{X}$

  and runs OLS regression

$$\boldsymbol{b}_{s+1} = (\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}'\boldsymbol{r}_{s+1}$$

- Proposition

$$\mathbb{E}[\boldsymbol{r}_{t+1}(\boldsymbol{C}\boldsymbol{b}_{s+1})'] = 0 \textit{ whenever } t \neq s.$$

- Useful special case: Trading strategy with predicted returns as portfolio weights ($t \neq s$)

$$r_{OOS,t+1} = \boldsymbol{w}'_{OOS,s+1}\boldsymbol{r}_{t+1}, \qquad \boldsymbol{w}_{OOS,s+1} = \boldsymbol{C}\boldsymbol{b}_{s+1}$$

  has $\mathbb{E}[r_{OOS,t+1}] = 0$.

(Absence of) out-of-sample return predictability

- Forward case is natural: Investors are Bayesian so the econometrician cannot "beat" investors in real-time return prediction

- Backwards case $s < t$ is more surprising. Not a tradable strategy, but interesting for academic research:
  - Suggests backwards OOS tests (e.g. Linnainmaa and Roberts 2018) and cross-validation (e.g., Kozak, Nagel and Santosh 2020; Bryzgalova, Pelger, and Zhu 2020) could be appropriate for Bayesian learning setting

- Caution: forward result is likely a general property of Bayesian learning (with objectively correct prior), the backwards result might be somewhat specific to the environment here (e.g., IID cash-flow growth).
  - Generality of this result is an interesting question for future research.

Testing for (absence of) out-of-sample return predictability

- Proposition

  If $s < t$, then in the large N, J, P limit

$$\frac{r_{OOS,t+1}}{\sqrt{\sum_{j=1}^{P} \zeta_{j,s}\zeta_{i,t}}} \xrightarrow{d} N(0,1)\,.$$

- Complication in implementation: Econometrician observes only a single sample conditional on one draw of $\boldsymbol{g}$ and $\mathbb{E}\, r_{OOS,t+1} = 0$ does not imply that $\mathbb{E}\,[r_{OOS,t+1}|\boldsymbol{g}] = 0$.

- Heuristic solution: Estimate variance from uncentered squared returns. Since $\mathbb{E}\, r_{OOS,t+1} = 0$,

$$\text{var}(r_{OOS,t+1}) = \mathbb{E}\, r^2_{OOS,t+1} = \mathbb{E}\,\mathbb{E}[r^2_{OOS,t+1}|\boldsymbol{g}]$$

Out-of-sample moment conditions for risk premia estimation

- Suppose returns from earlier augmented with risk premium/mispricing component $\boldsymbol{X}\boldsymbol{\gamma}_x = \boldsymbol{C}\gamma + \boldsymbol{M}\gamma_m$.

$$\boldsymbol{r}_{t+1} = \boldsymbol{X}\boldsymbol{\gamma}_x + \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{g} - \boldsymbol{X}\boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\bar{\boldsymbol{e}}_t + \boldsymbol{e}_{t+1}$$

- How can we estimate risk premium/mispricing contribution $\gamma'\boldsymbol{C}'\boldsymbol{C}\gamma$ to cross-sectional variation in expected returns associated with $\boldsymbol{C}$?

- Modification of earlier proposition implies

$$\boldsymbol{\gamma}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{\gamma} = \mathbb{E}\left[\boldsymbol{r}_{OOS,t+1}\right]$$

3. Finite-sample analysis: Simulations

Finite-sample analysis: Simulations

- To generate data, we set $\theta = 1$ in

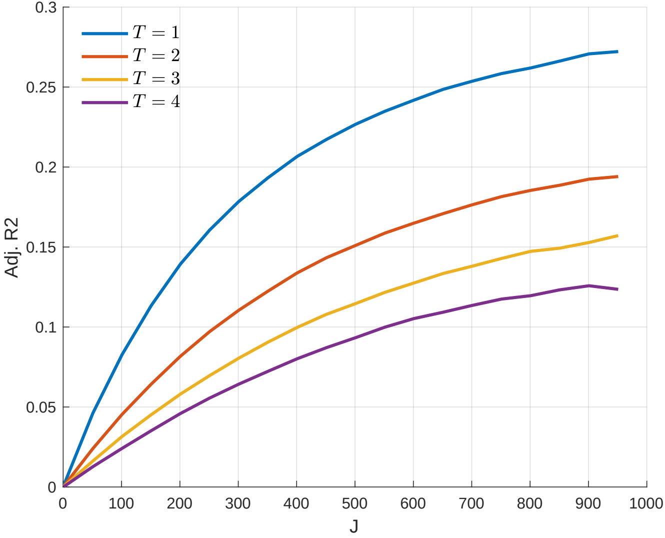
$$\boldsymbol{\Sigma}_g = \frac{\theta}{J}\boldsymbol{I}$$

- θ = ratio of forecastable/residual cash-flow growth variance
  - Based on analyst expectations, Chen, Karceski, and Lakonishok (2003) find forecastable/residual cash-flow growth variance of 0.4 at 10yr horizon

- Simulate cash-flows, prices, returns for $N = 1000$ assets.

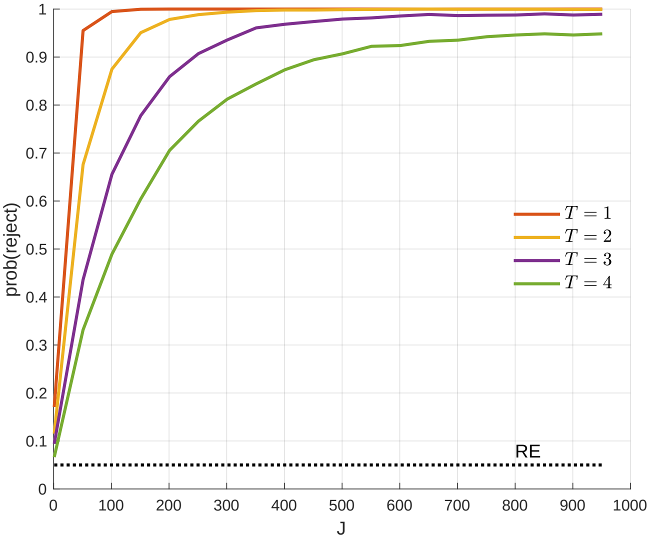
- Econometrician regresses $\boldsymbol{r}_{T+1}$ on $\boldsymbol{X}$ after investors have learned about $\boldsymbol{g}$ for T periods.

Adjusted R^2



0.3
0.25
0.2
0.15
0.1
0.05
0
Adj. R2
0
100
200
300
400
500
600
700
800
900
1000
J
$T = 1$
$T = 2$
$T = 3$
$T = 4$

Rejection probability of no-return-predictability null



Sparsity

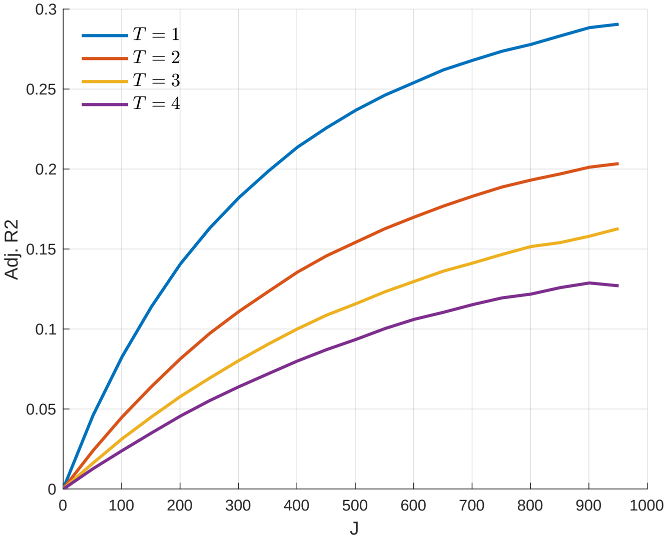
- Suppose prior about $\boldsymbol{g}$ not Normal, but Laplace,

$$f(g_j) = \frac{1}{2b} \exp\left(-\frac{|g_j|}{b}\right), \qquad 2b^2 = \frac{\theta}{J}$$

  and investors use maximum-a-posteriori (MAP) estimator to form $\tilde{\boldsymbol{g}}$

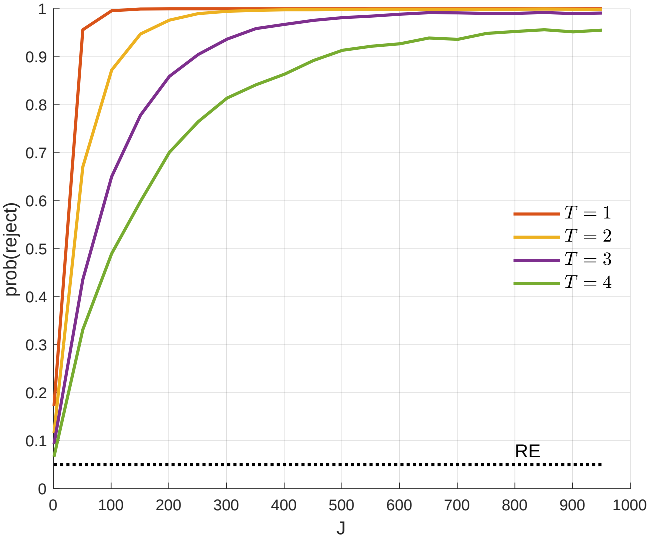
- Then some elements of $\tilde{\boldsymbol{g}}$ can be zero: Sparsity
- $\tilde{\boldsymbol{g}}$ are Lasso regression estimates
- Simulation: Again with objectively correct prior, now $\boldsymbol{g}$ drawn from Laplace distribution

Adjusted R^2



0.3
0.25
0.2
0.15
0.1
0.05
0
Adj. R2
0
100
200
300
400
500
600
700
800
900
1000
J
$T = 1$
$T = 2$
$T = 3$
$T = 4$

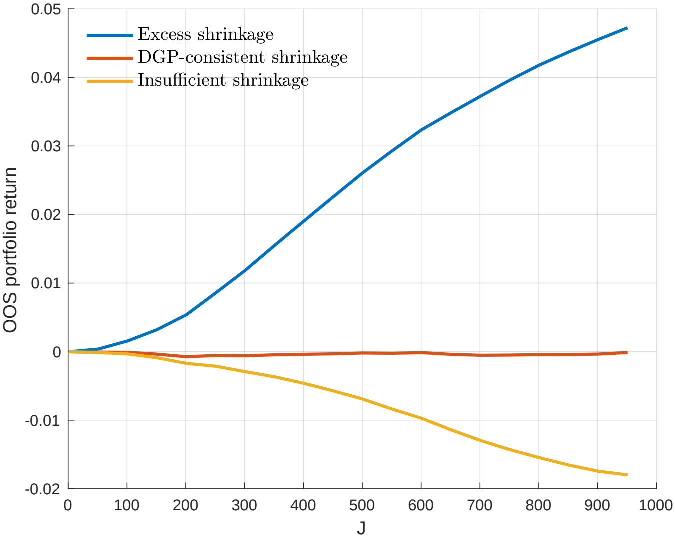
Rejection probability of no-return-predictability null



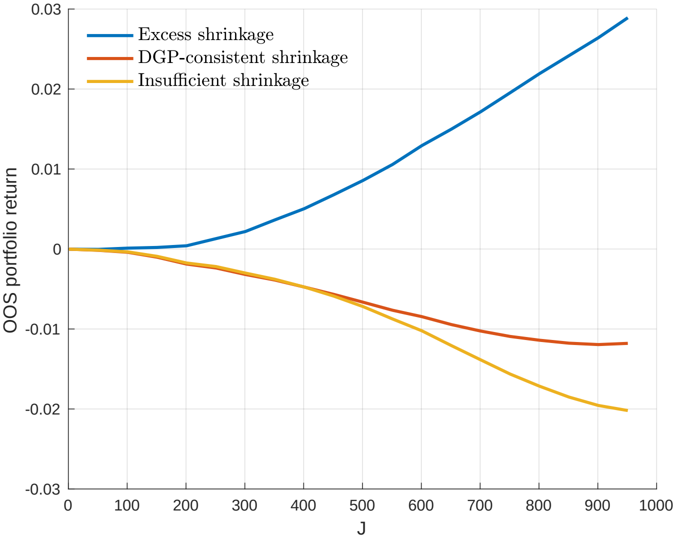
Excess shrinkage or sparsity

- Up to this point, shrinkage or sparsity was purely due to (objectively correct) informative prior beliefs of investors
- Possible reasons for additional sparsity:
  - Predictor variables costly to observe (change over time?)
  - Psychological cost of attention (Sims 2003; Gabaix 2014)
- To incorporate this we now allow prior to be tighter ($\theta = 0.5$) or more dispersed ($\theta = 2$) than distribution that we generate $\boldsymbol{g}$ from ($\theta = 1$)
  - For both normal prior (ridge) and Laplace (lasso)
- Plots: $T = 4$.

OOS portfolio return: Normal prior (ridge)



OOS portfolio return: Laplace prior (lasso)



Implications: Market Efficiency in the Age of Big Data

- In Big Data setting, RE (investors know $\boldsymbol{g}$) implausible and learning (about $\boldsymbol{g}$) has strong effects on asset price properties

- Risk premia & bias theories should focus on explaining OOS, not IS, return predictability
  - Investor learning provides clear motivation for (pseudo-)OOS testing which is lacking in RE framework

- Empirical challenge: characterize cross-section of stock returns OOS
  - Statistical testing with OOS moment conditions
  - Estimation of risk premia with OOS moment conditions

- Broader question for future research: Can puzzling empirical phenomena be explained by agent learning in Big Data settings?
  - Modeling economic decision makers as "machine learners"