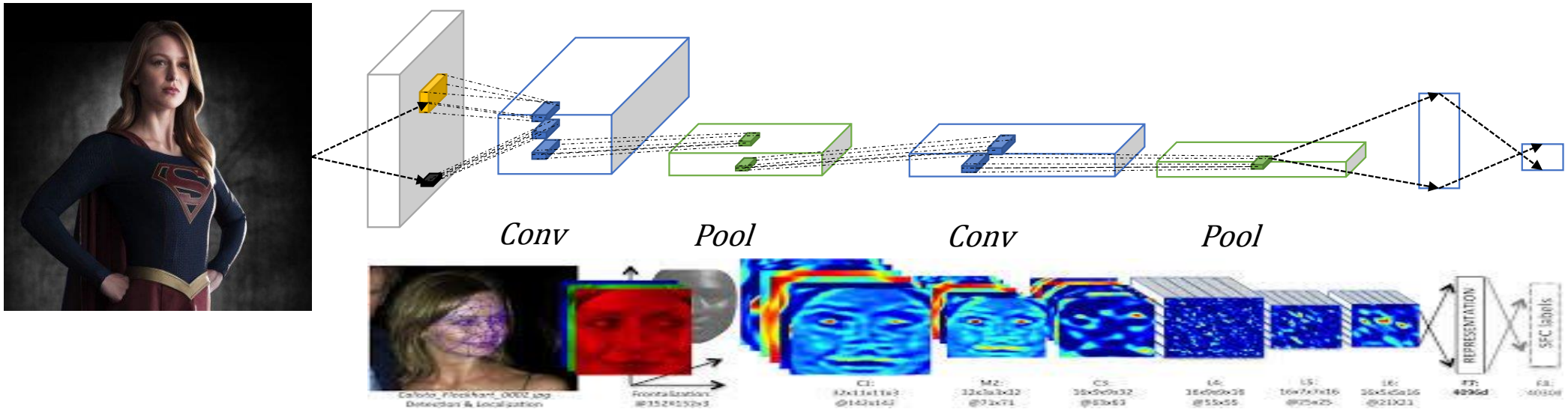
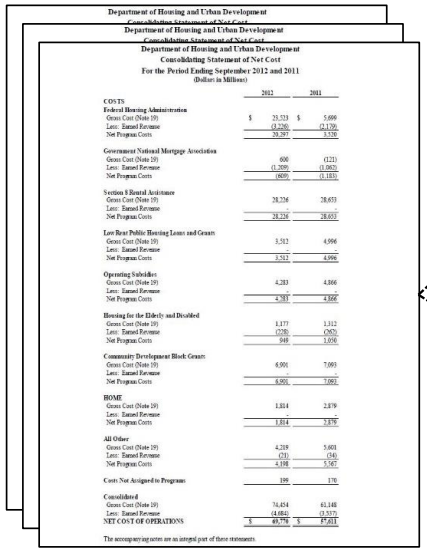
Material Financial Misstatement Detection using Ratios Analysis & Machine Learning Algorithms

Kexing Ding, Yunsen Wang and Xuan Peng

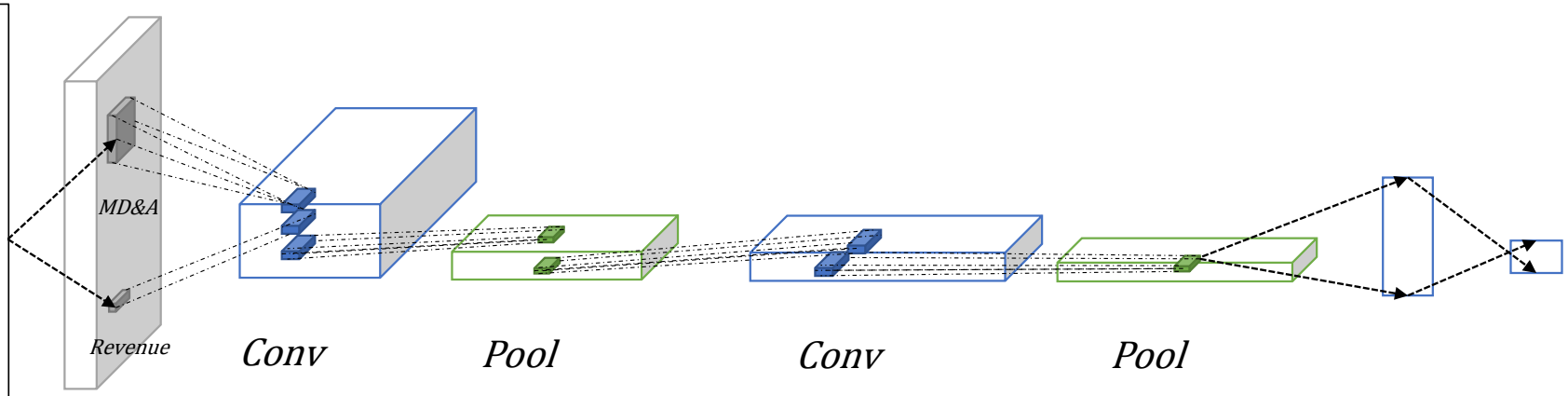


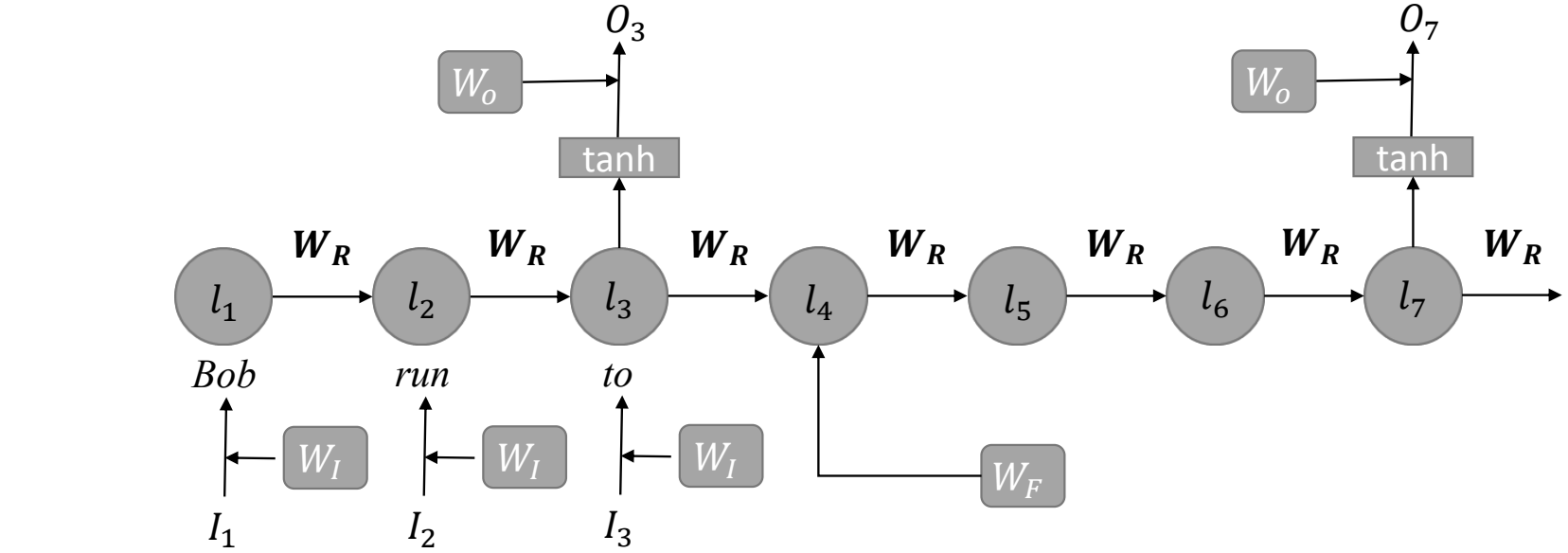


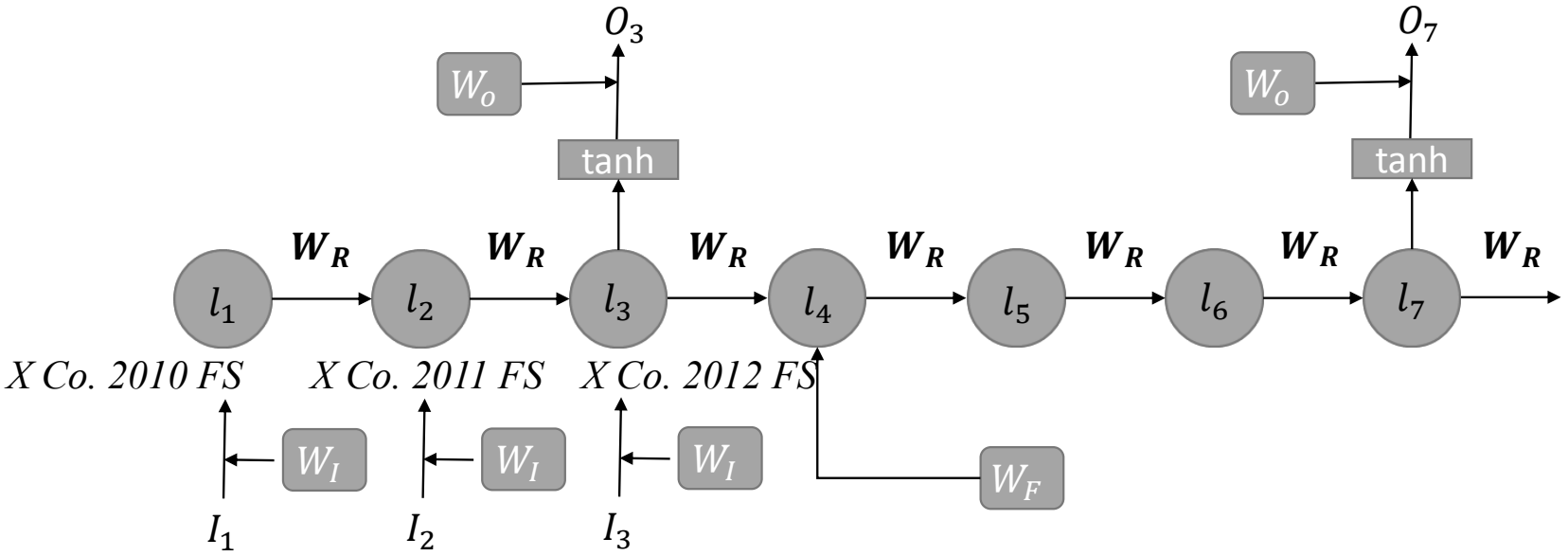
Department of Housing and Urban Development
Consolidating Statement of Net Cost
For the Period Ending September 2012 and 2011
(Dollars in Millions)

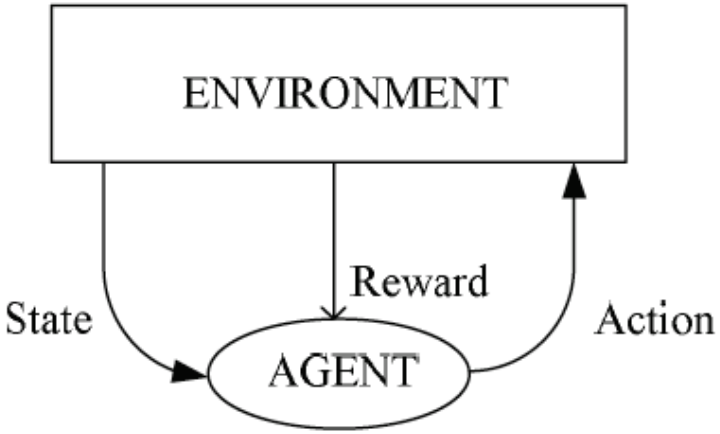
	2012	2011
COSTS		
Federal Housing Administration		
Gross Cost (Note 19)	$ 23,323	$ 5,699
Less: Earned Revenue	(3,226)	(2,179)
Net Program Costs	20,097	3,520
Government National Mortgage Association		
Gross Cost (Note 19)	600	(121)
Less: Earned Revenue	(1,209)	(1,062)
Net Program Costs	(609)	(1,183)
Section 8 Rental Assistance		
Gross Cost (Note 19)	28,226	28,653
Less: Earned Revenue	-	-
Net Program Costs	28,226	28,653
Low Rent Public Housing Loans and Grants		
Gross Cost (Note 19)	3,512	4,996
Less: Earned Revenue	-	-
Net Program Costs	3,512	4,996
Operating Subsidies		
Gross Cost (Note 19)	4,283	4,866
Less: Earned Revenue	-	-
Net Program Costs	4,283	4,866
Housing for the Elderly and Disabled		
Gross Cost (Note 19)	1,177	1,312
Less: Earned Revenue	(229)	(262)
Net Program Costs	948	1,050
Community Development Block Grants		
Gross Cost (Note 19)	6,901	7,093
Less: Earned Revenue	-	-
Net Program Costs	6,901	7,093
HOME		
Gross Cost (Note 19)	1,814	2,879
Less: Earned Revenue	-	-
Net Program Costs	1,814	2,879
All Other		
Gross Cost (Note 19)	4,219	5,601
Less: Earned Revenue	(21)	(34)
Net Program Costs	4,198	5,567
Costs Not Assigned to Programs	199	170
Consolidated		
Gross Cost (Note 19)	74,454	61,148
Less: Earned Revenue	(4,684)	(3,537)
NET COST OF OPERATIONS	$ 69,770	$ 57,611

The accompanying notes are an integral part of these statements.



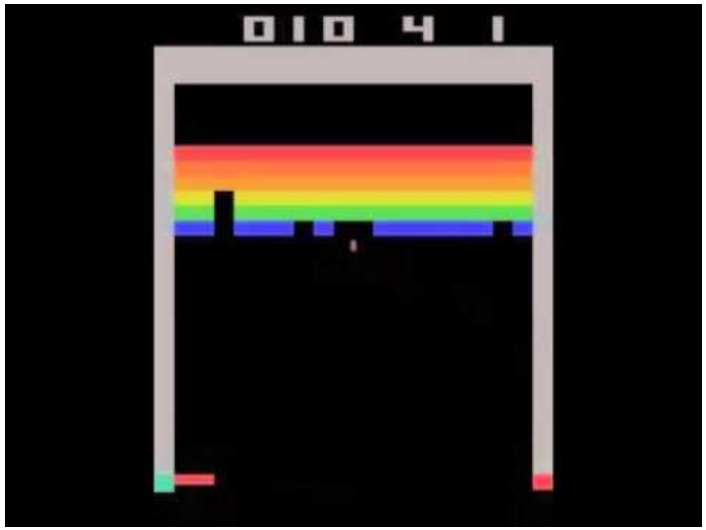

O_3
O_7
W_o
W_o
tanh
tanh
W_R
l_1
l_2
l_3
l_4
l_5
l_6
l_7
Bob
run
to
W_I
W_I
W_I
W_F
I_1
I_2
I_3


O_3
O_7
W_o
W_o
tanh
tanh
W_R
l_1
l_2
l_3
l_4
l_5
l_6
l_7
X Co. 2010 FS
X Co. 2011 FS
X Co. 2012 FS
W_I
W_I
W_I
W_F
I_1
I_2
I_3



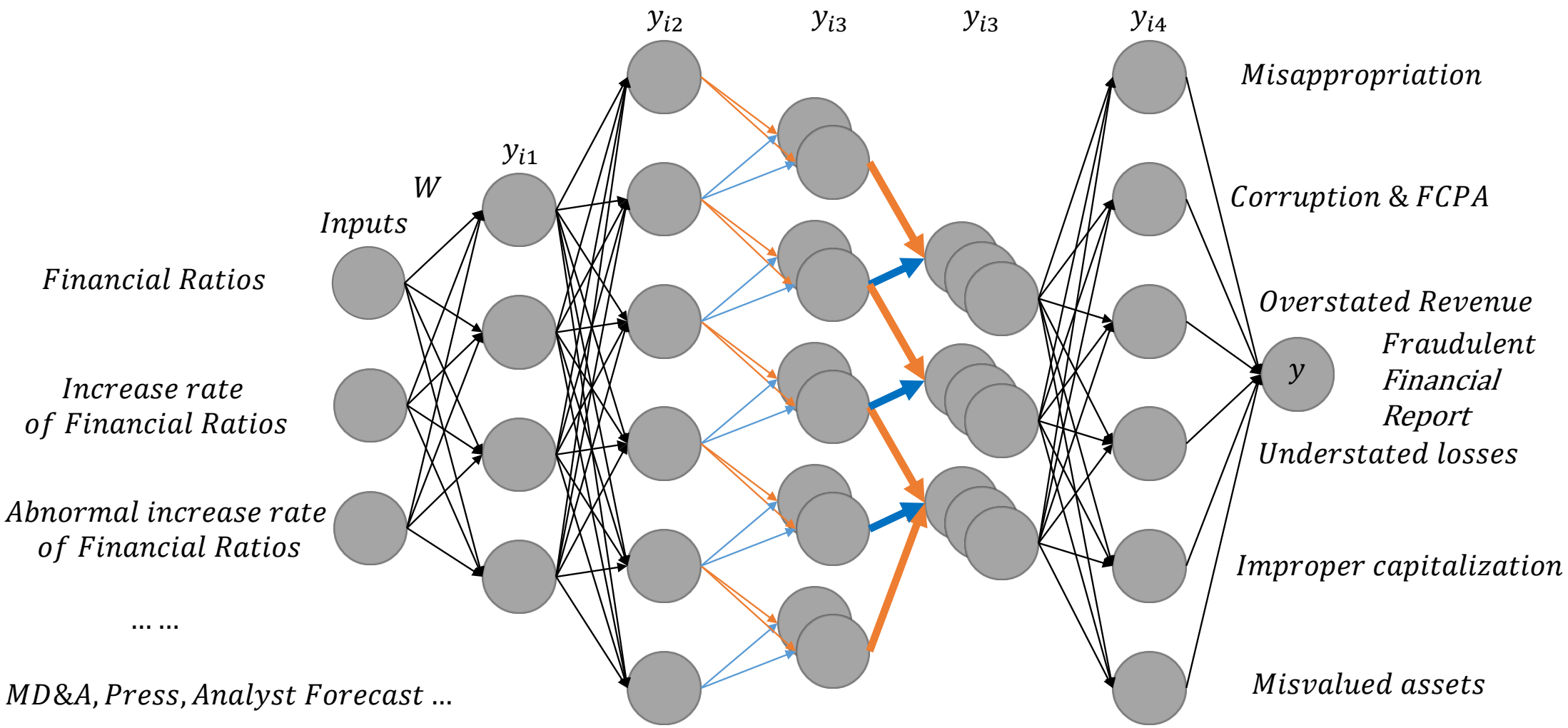
$$V^*(s_t) = \max_{a_t}\left(E[r_{t+1}] + \gamma \sum_{s_t+1} p(s_{t+1}|s_{t,}, a_t, r_t)V^*(s_{t+1})\right)$$

$$Q^*(s_t, a_t) = E\left[r_{t+1} + \gamma \sum_{s_t+1} p(s_{t+1}|s_{t,}, a_t, r_t) \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})\right]$$



NFQ_main() {
input: a set of transition samples D; output: Q-value function Q_N

k=0
init_MLP() $\rightarrow Q_0$;
Do {
generate_pattern_set $P = \{(input^l, target^l), l = 1, \ldots, \#D\}$ where:
$input^l = s^l, u^l,$
$target^l = c(s^l, u^l, s'^l) + \gamma\, min_b Q_k(s'^l, b)$
Rprop_training$(P) \rightarrow Q_{k+1}$
k:= k+1
} WHILE $(k < N)$


y_{i2}
y_{i3}
y_{i3}
y_{i4}
y_{i1}
W
Inputs
Financial Ratios
Increase rate of Financial Ratios
Abnormal increase rate of Financial Ratios
… …
MD&A, Press, Analyst Forecast …
Misappropriation
Corruption & FCPA
Overstated Revenue
Fraudulent Financial Report
Understated losses
Improper capitalization
Misvalued assets
y

Data

- Accounting and Auditing Enforcement Releases (AAER)
  - From 1982 to 2016, SEC has released 3783 AAERs.
  - Respondents involve
    - Officer of the company, auditor and audit firm, officer and company, company, others
  - Structure
    - Respondent, relevant laws and standards, summary, fraud scheme details, sanctions
- Audit Analytics
  - 4.02 Non-reliance Restatement Database since August 2004
  - The reasons for restatement
- Edgar
  - 10K, 10Q, 8K, NT-10K, NT-10Q
  - The detailed fraud schemes, the previous misstated reports and the restated financial reports

Sample Selection

- Pure Player firms
  - Mix of different business segments can be problematic
  - Segment obfuscation

Basic Assumptions

- Financial misstatement is abnormal
- Question: how to decide “normal”
  - Firm fundamentals can be affected by economic factors
  - Firm innovation and development
- We choose peer firms to mimic the normal level

Peer firm Identification

- Traditional SIC or NAICS
  - Hierarchical structure fails to capture firms that are more similar on a variety of dimensions (Clarke 1989).
- Analyst following classification (Ramnath 2002)
  - Analysts are encouraged to cover more than one industry. Supply chain, geographical closeness.

Methodology

- Dynamic industry classification using key ratios
- Clustering Analysis using ratios
- Ratios:
    * profitability ratios:captures performance profitability
    * r1: roa=sale/at
    * r2: profit margin=(sale-cogs)/sale

    * activity ratios: capture features of their operating activities
    * r3: operating cash flow ratio=operating cash flow/sale
    * r4: working capital turnover= sale/working capital

    * expenditure ratios: capture how firms spend money
    * r5: other expense= (xad+xsga)/sale
    * r6: operating expense= xopr/sale

    * balance ratios:
    * r7: liability=total liability/total asset
    * r8: current asset= act/at


0
0
50
100
150
200
250
0
50
100
150
200
250
0
50
100
150
200
250

Validation using restatement

- We calculate the standard deviation of each ratios and mark firm-year observations three standard deviations away as suspicious.
- We next sum the number of suspicious ratios and create **mis_score**.
- High **mis_score** should be associated with high likelihood of restatement. We predict the association to be stronger for dynamic peer groups than traditional industry classification.

Validation results

- Compare our peer and Fama French 49 industry classification:
    - We apply the same method to peer firms identified by SIC classification and compere the significant levels of the ratios in identifying financial restatement.
- As expected, the power of ratios in identifying abnormal ratios is not significant using traditional industry classification (FF 49 industry)

		Logistic Regression Predicted		Support Vector Machine		Artificial Neural Network	
Observed	Fraud Sample 1	Misstated	No-Misstated	Misstated	No-Misstated	Misstated	No-Misstated
Misstated	328	210	118	207	121	219	109
No-Misstated	328	130	198	128	200	130	198
	656	340	316	335	321	349	307
Misstated		64.02%	35.98%	63.11%	36.89%	66.77%	33.23%
No-Misstated		39.63%	60.37%	39.02%	60.98%	39.63%	60.37%
Correct classification			62.20%		62.04%		63.57%
False Negative			35.98%		36.89%		33.23%
False Positive			39.63%		39.02%		39.63%

Observed	Fraud Sample 2	Misstated	No-Misstated	Misstated	No-Misstated	Misstated	No-Misstated
Misstated	1399	921	478	890	509	1044	355
No-Misstated	1399	519	880	609	790	513	886
	2798	1440	1358	340	316	1557	1241
Misstated		65.83%	34.17%	63.62%	36.38%	74.62%	25.38%
No-Misstated		37.10%	62.90%	43.53%	56.47%	36.67%	63.33%
Correct classification			64.37%		60.04%		68.98%
False Negative			34.17%		36.38%		25.38%
False Positive			37.10%		43.53%		36.67%

Observed	Restatement Sample	Misstated	No-Misstated	Misstated	No-Misstated	Misstated	No-Misstated
Misstated	9044	921	478	5620	3424	1044	355
No-Misstated	9044	519	880	3717	5327	513	886
	18088	1440	1358	9337	8751	1557	1241
Misstated		65.83%	34.17%	62.14%	37.86%	74.62%	25.38%
No-Misstated		37.10%	62.90%	41.10%	58.90%	36.67%	63.33%
Correct classification			64.37%		60.52%		68.98%
False Negative			34.17%		37.86%		25.38%
False Positive			37.10%		41.10%		36.67%