AI and Big Data Analysis

1091BDA02
MBA, IM, NTPU (M5127) (Fall 2020)
Wed 7, 8, 9 (15:10-18:00) (B8F40)

Min-Yuh Day
戴敏育
Associate Professor
副教授
Institute of Information Management, National Taipei University

https://web.ntpu.edu.tw/~myday
2020-09-23
<table>
<thead>
<tr>
<th>項次 (Week)</th>
<th>日期 (Date)</th>
<th>內容 (Subject/Topics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2020/09/16</td>
<td>大數據分析介紹 (Introduction to Big Data Analysis)</td>
</tr>
<tr>
<td>2</td>
<td>2020/09/23</td>
<td>AI人工智慧與大數據分析 (AI and Big Data Analysis)</td>
</tr>
<tr>
<td>3</td>
<td>2020/09/30</td>
<td>Python 大數據分析基礎 (Foundations of Big Data Analysis in Python)</td>
</tr>
<tr>
<td>4</td>
<td>2020/10/07</td>
<td>數位沙盒第一堂課：數位沙盒服務平台簡介 (Digital Sandbox Lesson 1: Introduction to FintechSpace Digital Sandbox)</td>
</tr>
<tr>
<td>5</td>
<td>2020/10/14</td>
<td>數位沙盒第二堂課：工程師操作說明與實作教學 (Digital Sandbox Lesson 2: Hands-on Practices)</td>
</tr>
<tr>
<td>6</td>
<td>2020/10/21</td>
<td>Python Pandas 大數據量化分析 (Quantitative Big Data Analysis with Pandas in Python)</td>
</tr>
</tbody>
</table>
課程大綱 (Syllabus)

週次 (Week)  日期 (Date)  內容 (Subject/Topics)
7 2020/10/28  數位沙盒第三堂課：學生小組討論實作與成果發表
(Digital Sandbox Lesson 3: Learning Teams Hands-on Project Discussion and Project Presentation)
8 2020/11/04  Python Scikit-Learn 機器學習Ⅰ
(Machine Learning with Scikit-Learn In Python I)
9 2020/11/11  期中報告 (Midterm Project Report)
10 2020/11/18  Python Scikit-Learn 機器學習Ⅱ
(Machine Learning with Scikit-Learn In Python II)
11 2020/11/25  TensorFlow 深度學習金融大數據分析Ⅰ
(Deep Learning for Finance Big Data Analysis with TensorFlow I)
12 2020/12/02  大數據分析個案研究
(Case Study on Big Data Analysis)
## 課程大綱 (Syllabus)

週次 (Week) | 日期 (Date) | 內容 (Subject/Topics)
--- | --- | ---
13 | 2020/12/09 | TensorFlow 深度學習金融大數據分析 II  
(Deep Learning for Finance Big Data Analysis with TensorFlow II)
14 | 2020/12/16 | TensorFlow 深度學習金融大數據分析 III  
(Deep Learning for Finance Big Data Analysis with TensorFlow III)
15 | 2020/12/23 | AI 機器人理財顧問  
(Artificial Intelligence for Robo-Advisors)
16 | 2020/12/30 | 金融科技智慧型交談機器人  
(Conversational Commerce and Intelligent Chatbots for Fintech)
17 | 2021/01/06 | 期末報告 I (Final Project Report I)
18 | 2021/01/13 | 期末報告 II (Final Project Report II)
Outline

• AI

• Big Data Analytics
Evolution of Computerized Decision Support to Analytics/Data Science

The timeline in Figure 1.8 shows the terminology used to describe analytics since the 1970s. During the 1970s, the primary focus of information systems support for decision making focused on providing structured, periodic reports that a manager could use for decision making (or ignore them). Businesses began to create routine reports to inform decision makers (managers) about what had happened in the previous period (e.g., day, week, month, quarter). Although it was useful to know what had happened in the past, managers needed more than this: They needed a variety of reports at different levels of granularity to better understand and address changing needs and challenges of the business. These were usually called management information systems (MIS). In the early 1970s, Scott-Morton first articulated the major concepts of DSS. He defined DSSs as “interactive computer-based systems, which help decision makers utilize data and models to solve unstructured problems” (Gorry and Scott-Morton, 1971). The following is another classic DSS definition, provided by Keen and Scott-Morton (1978):

Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semistructured problems.

Note that the term decision support system, like management information system and several other terms in the field of IT, is a content-free expression (i.e., it means different things to different people). Therefore, there is no universally accepted definition of DSS.

During the early days of analytics, data was often obtained from the domain experts using manual processes (i.e., interviews and surveys) to build mathematical or knowledge-based models to solve constrained optimization problems. The idea was to do the best with limited resources. Such decision support models were typically called operations research (OR). The problems that were too complex to solve optimally (using linear or nonlinear mathematical programming techniques) were tackled using heuristic methods such as simulation models. (We will introduce these as prescriptive analytics later in this chapter and in a bit more detail in Chapter 6.)

In the late 1970s and early 1980s, in addition to the mature OR models that were being used in many industries and government systems, a new and exciting line of models had emerged: rule-based expert systems. These systems promised to capture experts’ knowledge in a format that computers could process (via a collection of if–then–else rules or heuristics) so that these could be used for consultation much the same way that one
Definition of Artificial Intelligence (A.I.)
Artificial Intelligence

“... the science and engineering of making intelligent machines”

(John McCarthy, 1955)

Artificial Intelligence

“... technology that thinks and acts like humans”
Artificial Intelligence

“... intelligence exhibited by machines or software”
<table>
<thead>
<tr>
<th>Thinking Humanly</th>
<th>Thinking Rationally</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acting Humanly</td>
<td>Acting Rationally</td>
</tr>
</tbody>
</table>

| 1.  | Acting Humanly: The Turing Test Approach (1950) |
| 2.  | Thinking Humanly: The Cognitive Modeling Approach |
| 3.  | Thinking Rationally: The “Laws of Thought” Approach |
| 4.  | Acting Rationally: The Rational Agent Approach |

AI Acting Humanly: 
The Turing Test Approach 
(Alan Turing, 1950)

• Natural Language Processing (NLP)
• Knowledge Representation
• Automated Reasoning
• Machine Learning (ML)
• Computer Vision
• Robotics

Can a robot pass a university entrance exam?

Noriko Arai at TED2017

https://www.ted.com/talks/noriko_arai_can_a_robot_pass_a_university_entrance_exam
https://www.youtube.com/watch?v=XQZjkPyJ8KU
Artificial Intelligence (A.I.) Timeline

**1950**
**TURING TEST**
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence.

**1955**
**A.I. BORN**
Term ‘artificial intelligence’ is coined by computer scientist John McCarthy to describe “the science and engineering of making intelligent machines.”

**1961**
**UNIMATE**
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line.

**1964**
**ELIZA**
Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans.

**1966**
**SHAKEY**
The ‘first electronic person’ from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions.

**1997**
**DEEP BLUE**
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov.

**1998**
**KISMET**
Cynthia Breazeal at MIT introduces Kismet, an emotionally intelligent robot insofar as it detects and responds to people’s feelings.

**1999**
**AIBO**
Sony launches first consumer robot pet dog, AIBO (A.I. robot) with skills and personality that develop over time.

**2002**
**ROOMBA**
First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes.

**2011**
**SIRI**
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S.

**2011**
**WATSON**
IBM’s question answering computer Watson wins first place on popular $1M prize television quiz show Jeopardy.

**2014**
**EUGENE**
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human.

**2014**
**ALEXA**
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks.

**2016**
**TAY**
Microsoft’s chatbot Tay goes rogue on social media making inflammatory and offensive racist comments.

**2017**
**ALPHAGO**
Google’s A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number of possible positions.

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.
AI, ML, DL

Artificial Intelligence (AI)

Machine Learning (ML)

Supervised Learning

Unsupervised Learning

Deep Learning (DL)

CNN
RNN LSTM GRU
GAN

Semi-supervised Learning

Reinforcement Learning

Source: https://leonardoaraujosantos.gitbooks.io/artificial-intelligence/content/deep_learning.html
Artificial Intelligence (AI) is many things

Ecosystem of AI

Source: https://www.i-scoop.eu/artificial-intelligence-cognitive-computing/
Deep Learning Evolution

Source: http://www.erogol.com/brief-history-machine-learning/
3 Machine Learning Algorithms

Machine Learning (ML) / Deep Learning (DL)

- **Machine Learning (ML)**
  - Supervised Learning
    - Decision Tree Classifiers
    - Linear Classifiers
    - Rule-based Classifiers
    - Probabilistic Classifiers
  - Unsupervised Learning
    - Deep Learning (DL)
    - Support Vector Machine (SVM)
    - Neural Network (NN)
  - Reinforcement Learning
    - Naïve Bayes (NB)
    - Bayesian Network (BN)
    - Maximum Entropy (ME)

Deep learning for financial applications: A survey

Applied Soft Computing (2020)

Source:
Financial time series forecasting with deep learning: A systematic literature review: 2005–2019

Applied Soft Computing (2020)

Source:
Deep learning for financial applications: Topics

Deep learning for financial applications: Deep Learning Models

Deep learning for financial applications: Topic-Model Heatmap

Deep learning for financial applications: Topic-Feature Heatmap

Deep learning for financial applications:
Topic-Dataset Heatmap

# Deep learning for financial applications: Algo-trading applications embedded with time series forecasting models

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>[33]</td>
<td>GarantiBank in BIST, Turkey</td>
<td>2016</td>
<td>OCHLV, Spread, Volatility, Turnover, etc.</td>
<td>PLR, Graves LSTM</td>
<td>MSE, RMSE, MAE, RSE, Correlation R-square</td>
<td>Spark</td>
</tr>
<tr>
<td>[34]</td>
<td>CSI300, Nifty50, HSI, Nikkei 225, S&amp;P500, DJIA</td>
<td>2010–2016</td>
<td>OCHLV, Technical Indicators</td>
<td>WT, Stacked autoencoders, LSTM</td>
<td>MAPE, Correlation coefficient, THEIL-U</td>
<td>-</td>
</tr>
<tr>
<td>[36]</td>
<td>50 stocks from NYSE</td>
<td>2007–2016</td>
<td>Price data</td>
<td>SFM</td>
<td>MSE</td>
<td>-</td>
</tr>
<tr>
<td>[37]</td>
<td>The LOB of 5 stocks of Finnish Stock Market</td>
<td>2010</td>
<td>FI-2010 dataset: bid/ask and volume</td>
<td>WMTR, MDA</td>
<td>Accuracy, Precision, Recall, F1-Score</td>
<td>-</td>
</tr>
<tr>
<td>[38]</td>
<td>300 stocks from SZSE, Commodity</td>
<td>2014–2015</td>
<td>Price data</td>
<td>FDDR, DMLP+RL</td>
<td>Profit, return, SR, profit-loss curves</td>
<td>Keras</td>
</tr>
<tr>
<td>[41]</td>
<td>Chinese stock-IF-IH-IC contract</td>
<td>2016–2017</td>
<td>Decisions for price change</td>
<td>MODRL+LSTM</td>
<td>Profit and loss, SR</td>
<td>-</td>
</tr>
<tr>
<td>[42]</td>
<td>Singapore Stock Market Index</td>
<td>2010–2017</td>
<td>OCHLV of last 10 days of Index</td>
<td>DMLP</td>
<td>RMSE, MAPE, Profit, SR</td>
<td>-</td>
</tr>
<tr>
<td>[43]</td>
<td>GBF/USD</td>
<td>2017</td>
<td>Price data</td>
<td>Reinforcement Learning + LSTM + NES</td>
<td>SR, downside deviation ratio, total profit</td>
<td>Python, Keras, Tensorflow</td>
</tr>
</tbody>
</table>

### Deep learning for financial applications:

**Algo-trading applications embedded with time series forecasting models**

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>[46]</td>
<td>Bitcoin, Dash, Ripple, Monero, Litecoin, Dogecoin, Nxt, Namecoin</td>
<td>2014–2017</td>
<td>MA, BOLL, the CRIX returns, Euribor interest rates, OCHLV</td>
<td>LSTM, RNN, DMLP</td>
<td>Accuracy, F1-measure</td>
<td>Python, Tensorflow</td>
</tr>
<tr>
<td>[47]</td>
<td>S&amp;P500, KOSPI, HSI, and EuroStoxx50</td>
<td>1987–2017</td>
<td>200-days stock price</td>
<td>Deep Q-Learning, DMLP</td>
<td>Total profit, Correlation</td>
<td>-</td>
</tr>
<tr>
<td>[49]</td>
<td>Fundamental and Technical Data, Economic Data</td>
<td>-</td>
<td>Fundamental, technical and market information</td>
<td>CNN</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

# Deep learning for financial applications:

Classification (buy–sell signal, or trend detection) based algo-trading models

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>[51]</td>
<td>Stocks in Dow30</td>
<td>1907–2017</td>
<td>RSI</td>
<td>DMIP with genetic algorithm</td>
<td>Annualized return</td>
<td>Spark MLlib, Java</td>
</tr>
<tr>
<td>[52]</td>
<td>SPY ETF, 10 stocks from S&amp;P500</td>
<td>2014–2016</td>
<td>Price data</td>
<td>FFNN</td>
<td>Cumulative gain</td>
<td>MatCom/Net, Matlab</td>
</tr>
<tr>
<td>[53]</td>
<td>Dow30 stocks</td>
<td>2012–2016</td>
<td>Close data and several technical indicators</td>
<td>LSTM</td>
<td>Accuracy</td>
<td>Python, Keras, Tensorflow, TALIB</td>
</tr>
<tr>
<td>[54]</td>
<td>High-frequency record of all orders</td>
<td>2014–2017</td>
<td>Price data, record of all orders, transactions</td>
<td>LSTM</td>
<td>Accuracy</td>
<td>–</td>
</tr>
<tr>
<td>[55]</td>
<td>Nasdaq Nordic (Keske Oyj, Outokumpu Oyj, Sampo, Kautaruukki, Wartsila Oyj)</td>
<td>2010</td>
<td>Price and volume data in LDB</td>
<td>LSTM</td>
<td>Precision, Recall, F1-score, Cohen’s k</td>
<td>–</td>
</tr>
<tr>
<td>[56]</td>
<td>17 ETFs</td>
<td>2000–2016</td>
<td>Price data, technical indicators</td>
<td>CNN</td>
<td>Accuracy, MSE, Profit, AUROC</td>
<td>Keras, Tensorflow</td>
</tr>
<tr>
<td>[57]</td>
<td>Stocks in Dow30 and 9 Top Volume ETFs</td>
<td>1997–2017</td>
<td>Price data, technical indicators</td>
<td>CNN with feature imaging</td>
<td>Recall, precision, F1-score, annualized return</td>
<td>Python, Keras, Tensorflow, Java</td>
</tr>
<tr>
<td>[58]</td>
<td>FTSE100</td>
<td>2000–2017</td>
<td>Price data</td>
<td>CAE</td>
<td>TR, SR, MDD, mean return</td>
<td>–</td>
</tr>
<tr>
<td>[59]</td>
<td>Nasdaq Nordic (Keske Oyj, Outokumpu Oyj, Sampo, Kautaruukki, Wartsila Oyj)</td>
<td>2010</td>
<td>Price, Volume data, 10 orders of the LDB</td>
<td>CNN</td>
<td>Precision, Recall, F1-score, Cohen’s k</td>
<td>Theano, Scikit learn, Python</td>
</tr>
<tr>
<td>[60]</td>
<td>Borsa Istanbul 100 Stocks</td>
<td>2011–2015</td>
<td>75 technical indicators and OCHLV</td>
<td>CNN</td>
<td>Accuracy</td>
<td>Keras</td>
</tr>
<tr>
<td>[61]</td>
<td>ETFs and Dow30</td>
<td>1997–2007</td>
<td>Price data</td>
<td>CNN with feature imaging</td>
<td>Annualized return</td>
<td>Keras, Tensorflow</td>
</tr>
<tr>
<td>[62]</td>
<td>8 experimental assets from bond/derivative market</td>
<td>–</td>
<td>Asset prices data</td>
<td>RL, DMLP, Genetic Algorithms</td>
<td>Learning and genetic algorithm error</td>
<td>–</td>
</tr>
<tr>
<td>[63]</td>
<td>10 stocks from S&amp;P500</td>
<td>–</td>
<td>Stock Prices</td>
<td>TDNN, RNN, PNN</td>
<td>Missed opportunities, false alarms ratio</td>
<td>–</td>
</tr>
</tbody>
</table>

## Deep learning for financial applications: Stand-alone and/or other algorithmic models

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>[67]</td>
<td>Taiwan Stock Index Futures, Mini Index Futures</td>
<td>2012–2014</td>
<td>Price data to image</td>
<td>Visualization method + CNN</td>
<td>Accumulated profits, accuracy</td>
<td>–</td>
</tr>
<tr>
<td>[70]</td>
<td>Taiwan stock index futures (TAIFEX)</td>
<td>2017</td>
<td>Price data</td>
<td>Agent based RL with CNN pre-trained</td>
<td>Accuracy</td>
<td>–</td>
</tr>
<tr>
<td>[71]</td>
<td>Stocks from S&amp;P500</td>
<td>2010–2016</td>
<td>OCHLV</td>
<td>DCNL</td>
<td>PCC, DTW, VWL</td>
<td>Pytorch</td>
</tr>
<tr>
<td>[74]</td>
<td>Experimental dataset</td>
<td>–</td>
<td>Price data</td>
<td>DRL with CNN, LSTM, GRU, DMLP</td>
<td>Mean profit</td>
<td>Python</td>
</tr>
</tbody>
</table>

## Deep learning for financial applications: Credit scoring or classification studies

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[77]</td>
<td>The XR 14 CDS contracts</td>
<td>2016</td>
<td>Recovery rate, spreads, sector and region</td>
<td>DBN+RBM</td>
<td>AUROC, FN, FP, Accuracy</td>
<td>WEKA</td>
</tr>
<tr>
<td>[78]</td>
<td>German, Japanese credit datasets</td>
<td>–</td>
<td>Personal financial variables</td>
<td>SVM + DBN</td>
<td>Weighted-accuracy, TP, TN</td>
<td>–</td>
</tr>
<tr>
<td>[79]</td>
<td>Credit data from Kaggle</td>
<td>–</td>
<td>Personal financial variables</td>
<td>DMLP</td>
<td>Accuracy, TP, TN, G-mean</td>
<td>–</td>
</tr>
<tr>
<td>[80]</td>
<td>Australian, German credit data</td>
<td>–</td>
<td>Personal financial variables</td>
<td>GP + AE as Boosted DMLP</td>
<td>FP</td>
<td>Python, Scikit-learn</td>
</tr>
<tr>
<td>[81]</td>
<td>German, Australian credit dataset</td>
<td>–</td>
<td>Personal financial variables</td>
<td>DCNN, DMLP</td>
<td>Accuracy, False/Missed alarm</td>
<td>–</td>
</tr>
<tr>
<td>[82]</td>
<td>Consumer credit data from Chinese finance company</td>
<td>–</td>
<td>Relief algorithm chose the 50 most important features</td>
<td>CNN + Relief</td>
<td>AUROC, K-s statistic, Accuracy</td>
<td>Keras</td>
</tr>
<tr>
<td>[83]</td>
<td>Credit approval dataset by UCI Machine Learning repo</td>
<td>–</td>
<td>UCI credit approval dataset</td>
<td>Rectifier, Tanh, Maxout DL</td>
<td>–</td>
<td>AWS EC2, H2O, R</td>
</tr>
</tbody>
</table>

# Deep learning for financial applications:

Financial distress, bankruptcy, bank risk, mortgage risk, crisis forecasting studies.

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[84]</td>
<td>966 french firms</td>
<td>–</td>
<td>Financial ratios</td>
<td>RBM+ SVM</td>
<td>Precision, Recall</td>
<td>–</td>
</tr>
<tr>
<td>[85]</td>
<td>883 BHC from EDGAR</td>
<td>2006–2017</td>
<td>Tokens, weighted sentiment polarity, leverage and ROA</td>
<td>CNN, LSTM, SVM, RF</td>
<td>Accuracy, Precision, Recall, F1-score</td>
<td>Keras, Python, Scikit-learn</td>
</tr>
<tr>
<td>[86]</td>
<td>The event data set for large European banks, news articles from Reuters</td>
<td>2007–2014</td>
<td>Word, sentence</td>
<td>DMLP + NLP preprocess</td>
<td>Relative usefulness, F1-score</td>
<td>–</td>
</tr>
<tr>
<td>[87]</td>
<td>Event dataset on European banks, news from Reuters</td>
<td>2007–2014</td>
<td>Text, sentence</td>
<td>Sentence vector + DFFN</td>
<td>Usefulness, F1-score, AUROC</td>
<td>–</td>
</tr>
<tr>
<td>[89]</td>
<td>Macro/Micro economic variables, Bank characteristics/performance variables from BHC</td>
<td>1976–2017</td>
<td>Macro economic variables and bank performances</td>
<td>CGAN, MVN, MV-t, LSTM, VAR, FE-QAR</td>
<td>RMSE, Log likelihood, Loan loss rate</td>
<td>–</td>
</tr>
<tr>
<td>[92]</td>
<td>Financial statements of several companies from Japanese stock market</td>
<td>2002–2016</td>
<td>Financial ratios</td>
<td>CNN</td>
<td>F1-score, AUROC</td>
<td>–</td>
</tr>
<tr>
<td>[95]</td>
<td>Private brokerage company’s real data of risky transactions</td>
<td>–</td>
<td>250 features: order details, etc.</td>
<td>CNN, LSTM</td>
<td>F1-Score</td>
<td>Keras, Tensorflow</td>
</tr>
<tr>
<td>[96]</td>
<td>Several datasets combined to create a new one</td>
<td>1996–2017</td>
<td>Index data, 10-year Bond yield, exchange rates,</td>
<td>Logit, CART, RF, SVM, NN, XGBoost, DMLP</td>
<td>AUROC, KS, G-mean, likelihood ratio, DP, BA, WBA</td>
<td>R</td>
</tr>
</tbody>
</table>

## Deep learning for financial applications: Fraud detection studies

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[114]</td>
<td>Debit card transactions by a local Indonesia bank</td>
<td>2016–2017</td>
<td>Financial transaction amount on several time periods</td>
<td>CNN, Stacked-LSTM, CNN-LSTM</td>
<td>AUROC</td>
<td>–</td>
</tr>
<tr>
<td>[115]</td>
<td>Credit card transactions from retail banking</td>
<td>2017</td>
<td>Transaction variables and several derived features</td>
<td>LSTM, GRU</td>
<td>Accuracy</td>
<td>Keras</td>
</tr>
<tr>
<td>[117]</td>
<td>Transactions made with credit cards by European cardholders</td>
<td>2013</td>
<td>Personal financial variables to PCA</td>
<td>DMLP, RF</td>
<td>Recall, Precision, Accuracy</td>
<td>–</td>
</tr>
<tr>
<td>[118]</td>
<td>Credit-card transactions</td>
<td>2015</td>
<td>Transaction and bank features</td>
<td>LSTM</td>
<td>AUROC</td>
<td>Keras, Scikit-learn</td>
</tr>
<tr>
<td>[119]</td>
<td>Databases of foreign trade of the Secretariat of Federal Revenue of Brazil</td>
<td>2014</td>
<td>8 Features: Foreign Trade, Tax, Transactions, Employees, Invoices, etc.</td>
<td>AE</td>
<td>MSE</td>
<td>H2O, R</td>
</tr>
<tr>
<td>[120]</td>
<td>Chamber of Deputies open data, Companies data from Secretariat of Federal Revenue of Brazil</td>
<td>2009–2017</td>
<td>21 features: Brazilian State expense, party name, Type of expense, etc.</td>
<td>Deep Autoencoders</td>
<td>MSE, RMSE</td>
<td>H2O, R</td>
</tr>
<tr>
<td>[121]</td>
<td>Real-world data for automobile insurance company labeled as fraudulent</td>
<td>–</td>
<td>Car, insurance and accident related features</td>
<td>DMLP + LDA</td>
<td>TP, FP, Accuracy, Precision, F1-score</td>
<td>–</td>
</tr>
<tr>
<td>[122]</td>
<td>Transactions from a giant online payment platform</td>
<td>2006</td>
<td>Personal financial variables</td>
<td>GBDT+DMLP</td>
<td>AUROC</td>
<td>–</td>
</tr>
<tr>
<td>[123]</td>
<td>Financial transactions</td>
<td>–</td>
<td>Transaction data</td>
<td>LSTM</td>
<td>t-SNE</td>
<td>–</td>
</tr>
<tr>
<td>[124]</td>
<td>Empirical data from Greek firms</td>
<td>–</td>
<td>–</td>
<td>DQL</td>
<td>Revenue</td>
<td>Torch</td>
</tr>
</tbody>
</table>

# Deep learning for financial applications: Portfolio management studies

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[65]</td>
<td>Cryptocurrencies, Bitcoin</td>
<td>2014–2017</td>
<td>Price data</td>
<td>CNN, RNN, LSTM</td>
<td>Accumulative portfolio value, MDD, SR</td>
<td></td>
</tr>
<tr>
<td>[127]</td>
<td>Stocks from NYSE, AMEX, NASDAQ</td>
<td>1965–2009</td>
<td>Price data</td>
<td>Autoencoder + REM</td>
<td>Accuracy, confusion matrix</td>
<td></td>
</tr>
<tr>
<td>[129]</td>
<td>Chinese stock data</td>
<td>2012–2013</td>
<td>Technical, fundamental data</td>
<td>Logistic Regression, RF, DMLP</td>
<td>AUC, accuracy, precision, recall, f1, tpr, fpr</td>
<td>Keras, Tensorflow, Python, Scikit learn</td>
</tr>
<tr>
<td>[130]</td>
<td>Top 5 companies in S&amp;P500</td>
<td>–</td>
<td>Price data and Financial ratios</td>
<td>LSTM, Auto-encoding, Smart indexing</td>
<td>CAGR</td>
<td></td>
</tr>
<tr>
<td>[132]</td>
<td>Tawans stock market</td>
<td>–</td>
<td>Price data</td>
<td>Elman RNN</td>
<td>MSE, return</td>
<td></td>
</tr>
<tr>
<td>[133]</td>
<td>FOREX (EUR/USD, etc.), Gold</td>
<td>2013</td>
<td>Price data</td>
<td>Evolino RNN</td>
<td>Return</td>
<td>Python</td>
</tr>
</tbody>
</table>
Deep learning for financial applications: Asset pricing and derivatives market studies

<table>
<thead>
<tr>
<th>Art.</th>
<th>Der. type</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[137]</td>
<td>Asset pricing</td>
<td>Analyst reports on the TSE and Osaka Exchange</td>
<td>2016–2018</td>
<td>Text</td>
<td>LSTM, CNN, Bi-LSTM</td>
<td>Accuracy, $R^2$</td>
<td>R, Python, MeCab</td>
</tr>
<tr>
<td>[142]</td>
<td>Options</td>
<td>Simulated a range of call option prices</td>
<td>–</td>
<td>Price data, option strike/maturity, dividend/risk free rates, volatility</td>
<td>DMLP</td>
<td>RMSE, the average percentage pricing error</td>
<td>Tensorflow</td>
</tr>
<tr>
<td>[143]</td>
<td>Futures, Options</td>
<td>TAIXE Options</td>
<td>2017</td>
<td>OCHLV, fundamental analysis, option price</td>
<td>DMLP, DMLP with Black scholes</td>
<td>RMSE, MAE, MAPE</td>
<td>–</td>
</tr>
<tr>
<td>[144]</td>
<td>Equity returns</td>
<td>Returns in NYSE, AMEX, NASDAQ</td>
<td>1975–2017</td>
<td>57 firm characteristics</td>
<td>Fama–French $n$-factor model DL</td>
<td>$R^2$, RMSE</td>
<td>Tensorflow</td>
</tr>
</tbody>
</table>

## Deep learning for financial applications: Cryptocurrency and blockchain studies

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[146]</td>
<td>Bitcoin data</td>
<td>2012, 2013, 2016</td>
<td>TransactionId, input/output Addresses, timestamp</td>
<td>Graph embedding using heuristic, laplacian eigen-map, deep AE</td>
<td>F1-score</td>
<td>–</td>
</tr>
</tbody>
</table>

### Deep learning for financial applications:

**Financial sentiment studies coupled with text mining for forecasting**

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[137]</td>
<td>Analyst reports on the TSE and Osaka Exchange</td>
<td>2016–2018</td>
<td>Text</td>
<td>LSTM, CNN, Bi-LSTM</td>
<td>Accuracy, $R^2$</td>
<td>R, Python, McCab</td>
</tr>
<tr>
<td>[154]</td>
<td>SCI prices</td>
<td>2013–2016</td>
<td>Text data and Price data</td>
<td>LSTM</td>
<td>Accuracy, F1-Measure</td>
<td>Python, Keras</td>
</tr>
<tr>
<td>[155]</td>
<td>Stocks of Google, Microsoft and Apple</td>
<td>2016–2017</td>
<td>Twitter sentiment and stock prices</td>
<td>RNN</td>
<td>–</td>
<td>Spark, Flume,Twitter API,</td>
</tr>
<tr>
<td>[156]</td>
<td>30 DJIA stocks, S&amp;P500, DJI, news from Reuters</td>
<td>2002–2016</td>
<td>Price data and features from news articles</td>
<td>LSTM, NN, CNN and word2vec</td>
<td>Accuracy</td>
<td>VADER</td>
</tr>
<tr>
<td>[157]</td>
<td>Stocks of CSI300 index, OCHLV of CSI300 index</td>
<td>2009–2014</td>
<td>Sentiment Posts, Price data</td>
<td>Naive Bayes + LSTM</td>
<td>Precision, Recall, F1-score, Accuracy</td>
<td>Python, Keras</td>
</tr>
</tbody>
</table>

## Deep learning for financial applications: Text mining studies without sentiment analysis for forecasting

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[173]</td>
<td>Price data, index data, news, social media data</td>
<td>2015</td>
<td>Price data, news from articles and social media</td>
<td>Coupled matrix and tensor</td>
<td>Accuracy, MCC</td>
<td>Jieba</td>
</tr>
<tr>
<td>[174]</td>
<td>HS300</td>
<td>2015–2017</td>
<td>Social media news, price data</td>
<td>RNN-Boost with LDA</td>
<td>Accuracy, MAE, MAPE, RMSE</td>
<td>Python, Scikit-learn</td>
</tr>
</tbody>
</table>

## Deep learning for financial applications:

### Text mining studies without sentiment analysis for forecasting

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[175]</td>
<td>News and Chinese stock data</td>
<td>2014–2017</td>
<td>Selected words in a news</td>
<td>HAN</td>
<td>Accuracy, Annual return</td>
<td>–</td>
</tr>
<tr>
<td>[176]</td>
<td>News, stock prices from Hong Kong Stock Exchange</td>
<td>2001</td>
<td>Price data and TF-IDF from news</td>
<td>ELM, DLR, PCA, BELM, KELM, NN</td>
<td>Accuracy</td>
<td>Matlab</td>
</tr>
<tr>
<td>[178]</td>
<td>Stock of Tsugami Corporation</td>
<td>2013</td>
<td>Price data</td>
<td>LSTM</td>
<td>RMSE</td>
<td>Keras, Tensorflow</td>
</tr>
<tr>
<td>[179]</td>
<td>News, Nikkei Stock Average and 10-Nikkei companies</td>
<td>1999–2008</td>
<td>news, MACD</td>
<td>RNN, RBM+DBN</td>
<td>Accuracy, P-value</td>
<td>–</td>
</tr>
<tr>
<td>[180]</td>
<td>ISMIS 2017 Data Mining Competition dataset</td>
<td>–</td>
<td>Expert identifier, classes</td>
<td>LSTM + GRU + FFNN</td>
<td>Accuracy</td>
<td>–</td>
</tr>
<tr>
<td>[182]</td>
<td>APPI from S&amp;P500 and news from Reuters</td>
<td>2011–2017</td>
<td>Input news, OCHLV, Technical indicators</td>
<td>CNN + LSTM, CNN+SVM</td>
<td>Accuracy, F1-score</td>
<td>Tensorflow</td>
</tr>
</tbody>
</table>

## Deep Learning for Financial Applications: Financial Sentiment Studies Coupled with Text Mining without Forecasting

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature Set</th>
<th>Method</th>
<th>Performance Criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[85]</td>
<td>883 BHC from EDGAR</td>
<td>2006–2017</td>
<td>Tokens, weighted sentiment polarity, leverage and ROA</td>
<td>CNN, LSTM, SVM, Random Forest</td>
<td>Accuracy, Precision, Recall, F1-score</td>
<td>Keras, Python, Scikit-learn</td>
</tr>
<tr>
<td>[185]</td>
<td>SemEval-2017 dataset, financial text, news, stock market data</td>
<td>2017</td>
<td>Sentiments in Tweets, News headlines</td>
<td>Ensemble SVR, CNN, LSTM, GRU</td>
<td>Cosine similarity score, agreement score, class score</td>
<td>Python, Keras, Scikit Learn</td>
</tr>
<tr>
<td>[190]</td>
<td>StockTwits</td>
<td>2008–2015</td>
<td>Sentences, StockTwits messages</td>
<td>CNN, LSTM, GRU</td>
<td>MCC, WSURT</td>
<td>Keras, Tensorflow</td>
</tr>
<tr>
<td>[191]</td>
<td>Financial statements of Japan companies</td>
<td>–</td>
<td>Sentences, text</td>
<td>DMLLP</td>
<td>Precision, recall, f-score</td>
<td>–</td>
</tr>
<tr>
<td>[192]</td>
<td>Twitter posts, news headlines</td>
<td>–</td>
<td>Sentences, text</td>
<td>Deep-FASP</td>
<td>Accuracy, MSE, R²</td>
<td>–</td>
</tr>
<tr>
<td>[194]</td>
<td>News from Financial Times related US stocks</td>
<td>–</td>
<td>Sentiment of news headlines</td>
<td>SVR, Bidirectional LSTM</td>
<td>Cosine similarity</td>
<td>Python, Scikit Learn, Keras, Tensorflow</td>
</tr>
</tbody>
</table>

## Deep learning for financial applications:
### Other text mining studies

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[86]</td>
<td>The event data set for large European banks, news articles from Reuters</td>
<td>2007–2014</td>
<td>Word, sentence</td>
<td>DMLP + NLP preprocess</td>
<td>Relative usefulness, F1-score</td>
<td>–</td>
</tr>
<tr>
<td>[87]</td>
<td>Event dataset on European banks, news from Reuters</td>
<td>2007–2014</td>
<td>Text, sentence</td>
<td>Sentence vector + DFFN</td>
<td>Usefulness, F1-score, AUROC</td>
<td>–</td>
</tr>
<tr>
<td>[121]</td>
<td>Real-world data for automobile insurance company labeled as fraudulent</td>
<td>–</td>
<td>Car, insurance and accident related features</td>
<td>DMLP + LDA</td>
<td>TP, FP, Accuracy, Precision, F1-score</td>
<td>–</td>
</tr>
<tr>
<td>[123]</td>
<td>Financial transactions</td>
<td>–</td>
<td>Transaction data</td>
<td>LSTM</td>
<td>t-SNE</td>
<td>–</td>
</tr>
<tr>
<td>[195]</td>
<td>Taiwan's National Pension Insurance</td>
<td>2008–2014</td>
<td>Insured’s id, area-code, gender, etc.</td>
<td>RNN</td>
<td>Accuracy, total error</td>
<td>Python</td>
</tr>
</tbody>
</table>

Deep learning for financial applications:
Other theoretical or conceptual studies

<table>
<thead>
<tr>
<th>Art.</th>
<th>SubTopic</th>
<th>IsTimeSeries?</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>[197]</td>
<td>Analysis of AE, SVD</td>
<td>Yes</td>
<td>Selected stocks from the IBB index and stock of Amgen Inc.</td>
<td>2012–2014</td>
<td>Price data</td>
<td>AE, SVD</td>
</tr>
<tr>
<td>[198]</td>
<td>Fraud Detection in Banking</td>
<td>No</td>
<td>Risk Management / Fraud Detection</td>
<td>–</td>
<td>–</td>
<td>DRL</td>
</tr>
</tbody>
</table>

# Deep learning for financial applications:

## Other financial applications

<table>
<thead>
<tr>
<th>Art.</th>
<th>Subtopic</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[195]</td>
<td>Predicting Social Ins. Payment Behavior</td>
<td>Taiwan’s National Pension Insurance</td>
<td>2008–2014</td>
<td>Insured’s id, area-code, gender, etc.</td>
<td>RNN</td>
<td>Accuracy, total error</td>
<td>Python</td>
</tr>
<tr>
<td>[201]</td>
<td>Predicting Bank Telemarketing</td>
<td>Phone calls of bank marketing data</td>
<td>2008–2010</td>
<td>16 finance-related attributes</td>
<td>CNN</td>
<td>Accuracy</td>
<td>–</td>
</tr>
</tbody>
</table>

Financial time series forecasting with deep learning: Topic-model heatmap

Stock price forecasting using only raw time series data

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Lag</th>
<th>Horizon</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[80]</td>
<td>38 stocks in KOSPI</td>
<td>2010–2014</td>
<td>Lagged stock returns OCHLV</td>
<td>50 min</td>
<td>5 min</td>
<td>DNN</td>
<td>NMSE, RMSE, MAE, MI</td>
<td>–</td>
</tr>
<tr>
<td>[82]</td>
<td>Daily returns of ‘BRD’ stock in Romanian Market 297 listed companies of CSE 5 stock in NSE</td>
<td>2001–2016</td>
<td>OCHLV</td>
<td>–</td>
<td>1 d</td>
<td>LSTM</td>
<td>RMSE, MAE</td>
<td>Python, Theano</td>
</tr>
<tr>
<td>[83]</td>
<td>Stocks of Infosys, TCS and CIPLA from NSE</td>
<td>2012–2013</td>
<td>OCHLV, Price data, turnover and number of trades.</td>
<td>–</td>
<td>–</td>
<td>LSTM, SRNN, GRU</td>
<td>MAD, MAPE</td>
<td>Keras</td>
</tr>
<tr>
<td>[84]</td>
<td>Stocks of Infosys, TCS and CIPLA from NSE</td>
<td>1997–2016</td>
<td>Price data</td>
<td>–</td>
<td>–</td>
<td>LSTMs, CNN, MLP</td>
<td>MAPE</td>
<td>–</td>
</tr>
<tr>
<td>[85]</td>
<td>Stocks of Infosys, TCS and CIPLA from NSE</td>
<td>2014</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>RNN, LSTM and CNN</td>
<td>Accuracy</td>
<td>–</td>
</tr>
<tr>
<td>[87]</td>
<td>High-frequency transaction data of the CSI300 futures Stocks in the S&amp;P500</td>
<td>2017</td>
<td>Price data</td>
<td>1 d</td>
<td>1 d</td>
<td>DBN</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>[92]</td>
<td>Stocks of Infosys, TCS and CIPLA from NSE</td>
<td>1985–2006</td>
<td>Monthly and daily log-returns OCHLV</td>
<td>*</td>
<td>1 d</td>
<td>DBN+MLP</td>
<td>Validation, Test Error</td>
<td>Theano, Python, Matlab</td>
</tr>
<tr>
<td>[93]</td>
<td>Stocks of Infosys, TCS and CIPLA from NSE</td>
<td>2000–2017</td>
<td>Price data</td>
<td>60 d</td>
<td>1.7 d</td>
<td>DNNN</td>
<td>MSE</td>
<td>Tensorflow</td>
</tr>
<tr>
<td>[94]</td>
<td>Stocks of Infosys, TCS and CIPLA from NSE</td>
<td>2007–2016</td>
<td>Price data</td>
<td>–</td>
<td>1d, 3 d, 5 d</td>
<td>SFM</td>
<td>MSE</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Art.</th>
<th>Data set</th>
<th>Period</th>
<th>Feature set</th>
<th>Lag</th>
<th>Horizon</th>
<th>Method</th>
<th>Performance criteria</th>
<th>Env.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[97]</td>
<td>U.S. low-level disaggregated macroeconomic time series CDAX stock market data</td>
<td>1926–2016</td>
<td>Fundamental Features: GDP, Unemployment rate, Inventories, etc.</td>
<td>–</td>
<td>1 s</td>
<td>DNN</td>
<td>MSPE</td>
<td>Tensorflow</td>
</tr>
<tr>
<td>[100]</td>
<td>10 stocks in Nikkei 225 and news text TKC stock in NYSE and QQQQ ETF</td>
<td>2006–2007</td>
<td>Price data</td>
<td>–</td>
<td>–</td>
<td>LSTM</td>
<td>RMSE</td>
<td>–</td>
</tr>
<tr>
<td>[102]</td>
<td>42 stocks in China's SSE Google's daily stock data GarantiBank in BIST, Turkey</td>
<td>2016</td>
<td>OCHLV, Technical indicators</td>
<td>242 min</td>
<td>1 min</td>
<td>GAN (LSTM, CNN)</td>
<td>RMSRE, DPA, GAN-F, CAN-D</td>
<td>–</td>
</tr>
<tr>
<td>[103]</td>
<td>Stocks in NYSE, AMEX, NASDAQ, TAQ, intraday trade Private brokerage company's real data of risky transactions</td>
<td>2004–2015</td>
<td>OCHLV, Technical indicators</td>
<td>20 d</td>
<td>1 d</td>
<td>(2D)^2 PCA + DNN</td>
<td>SMAPE, PCD, MAPE, RMSE, HR, TR, R^2</td>
<td>R, Matlab</td>
</tr>
<tr>
<td>[104]</td>
<td>Stocks in NYSE, AMEX, NASDAQ, TAQ, intraday trade</td>
<td>2016</td>
<td>OCHLV, Volatility, etc.</td>
<td>–</td>
<td>–</td>
<td>PLR, Graves LSTM</td>
<td>MSE, RMSE, MAE, RSE, R^2</td>
<td>Spark</td>
</tr>
<tr>
<td>[105]</td>
<td>Stocks in NYSE, AMEX, NASDAQ, TAQ, intraday trade</td>
<td>1993–2017</td>
<td>Price, 15 firm characteristics</td>
<td>80 d</td>
<td>1 d</td>
<td>LSTM + MLP</td>
<td>Monthly return, SR</td>
<td>Python, Keras, Tensorflow in AWS Keras, Tensorflow</td>
</tr>
<tr>
<td>[106]</td>
<td>Fundamentals against technical, economic data</td>
<td>2010</td>
<td>250 features: order details, etc.</td>
<td>–</td>
<td>–</td>
<td>CNN</td>
<td>F1-Score</td>
<td>–</td>
</tr>
</tbody>
</table>

Stock Market Movement Forecast: Phases of the stock market modeling

Big Data Analytics
Big Data 4 V

Volume
- 40 Zettabytes (43 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005
- 6 billion people have cell phones
- World population: 7 billion

Velocity
- The New York Stock Exchange captures 1.7B of trade information during each trading session
- Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure
- By 2016, it is projected there will be 18.9 billion network connections—almost 2.5 connections per person on earth

Variety
- It's estimated that 2.5 quintillion bytes (2.3 trillion gigabytes) of data are created each day
- Most companies in the U.S. have at least 100 terabytes (100 billion gigabytes) of data stored
- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (160 million gigabytes)
- By 2014, it's anticipated there will be 420 million wearable, wireless health monitors
- 4 billion+ hours of video are watched on YouTube each month
- 30 billion pieces of content are shared on Facebook every month
- 400 million tweets are sent per day by about 200 million monthly active users

Veracity
- 27% of respondents in one survey were unsure of how much of their data was inaccurate
- 1 in 3 business leaders don't trust the information they use to make decisions
- Poor data quality costs the US economy around $3.1 trillion a year

The FOUR V's of Big Data
- From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
- 4.4 million IT jobs will be created globally to support big data, with 1.9 million in the United States

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Source: https://www-01.ibm.com/software/data/bigdata/
Value
Value Creation by Big Data Analytics
(Grover et al., 2018)

Value Manifestation

- Strategy
- Leadership
- Technology
- Industry Context
- Trust
- Governance Support
- Competitive Dynamics
- Data-Driven Culture

Moderating Factors

BDA Infrastructure
- Big Data Asset
- Analytics Portfolio
- Human Talent

BDA Capabilities
- Ability to integrate, disseminate, explore, and analyze big data

Direct value from BDA

Value Creation Mechanisms
- Transparency and access
- Discovery and experimentation
- Prediction and optimization
- Customization and targeting
- Learning and crowdsourcing
- Continuous monitoring and proactive adaptation

Value Targets
- Organization Performance
- Business Processes Improvement
- Products & Services Innovation
- Consumer Experience & Market Enhancement

Impact
- Functional Value
- Symbolic Value

Investments --- Assets --- Capabilities --- Applications --- Targets --- Impacts --- Value

Learning by Doing (Coevolutionary Adaptation)

A bibliometric analysis on Big Data and Business Intelligence from 1990 to 2016.

Big Data papers grow much faster than Business Intelligence papers.

Computer Science and information systems are two core disciplines.

Most influential papers are identified and a research framework is proposed.

Source: Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study", Expert Systems with Applications, Volume 111, 30, 2018, pp. 2-10
Evolution of top keywords in “BD & BI” publications

- 2014:
  - Management
  - Text Mining
  - Data Mining
  - Data Science

- 2015:
  - Big Data Analytics
  - Social Media
  - Business Analytics
  - Information System

- 2016:
  - Cloud Computing
  - Data Warehouse

- 2017:
  - Knowledge Management

Source: Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study", Expert Systems with Applications, Volume 111, 30, 2018, pp. 2-10
Framework for BD and BI Research

- **Technology**
  - Data Collection
  - Data Storage
  - Data Analytics
  - Infrastructure

- **Application**
  - Business
  - Medicate
  - Supply Chain
  - Engineering
  - Services

- **Impact**
  - Value Creation
  - Individual Impact
  - Organizational Impact
  - Social Impact

- **Management**
  - Adoption of BD/BI
  - Cost Benefit
  - Security/Privacy
  - Human Resource

Source: Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study", Expert Systems with Applications, Volume 111, 30, 2018, pp. 2-10
Business Intelligence and Big Data analytics

Source: Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study", Expert Systems with Applications, Volume 111, 30, 2018, pp. 2-10

Stephan Kudyba (2014),
Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

Source: http://www.amazon.com/gp/product/1466568704
Architecture of Big Data Analytics

**Big Data Sources**
- * Internal
- * External
- * Multiple formats
- * Multiple locations
- * Multiple applications

**Big Data Transformation**
- Raw Data
- Middleware
- Extract Transform Load
- Data Warehouse
- Traditional Format CSV, Tables

**Big Data Platforms & Tools**
- Hadoop
- MapReduce
- Pig
- Hive
- Jaql
- Zookeeper
- Hbase
- Cassandra
- Oozie
- Avro
- Mahout
- Others

**Big Data Analytics Applications**
- Queries
- Reports
- OLAP
- Data Mining

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Architecture of Big Data Analytics

Big Data Sources
- Internal
- External
- Multiple formats
- Multiple locations
- Multiple applications

Big Data Transformation

Big Data Platforms & Tools

Big Data Analytics Applications
- Queries
- Reports
- OLAP
- Data Mining

Data Mining
Big Data Analytics
Applications

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Architecture for Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Enabling Technologies
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distrusted processing

Analysts
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press
Data Warehouse

Data Mining and Business Intelligence

Increasing potential to support business decisions

Decision Making

Data Presentation
Visualization Techniques

Data Mining
Information Discovery

Data Exploration
Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

End User
Business Analyst
Data Analyst
DBA

Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier
The Evolution of BI Capabilities

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
computer technology, management science techniques, and statistics to solve real problems. Of course, many other organizations have proposed their own interpretations and motivations for analytics. For example, SAS Institute Inc. proposed eight levels of analytics that begin with standardized reports from a computer system. These reports essentially provide a sense of what is happening with an organization. Additional technologies have enabled us to create more customized reports that can be generated on an ad hoc basis. The next extension of reporting takes us to OLAP-type queries that allow a user to dig deeper and determine specific sources of concern or opportunities. Technologies available today can also automatically issue alerts for a decision maker when performance warrants such alerts. At a consumer level we see such alerts for weather or other issues. But similar alerts can also be generated in specific settings when sales fall above or below a certain level within a certain time period or when the inventory for a specific product is running low. All of these applications are made possible through analysis and queries on data being collected by an organization. The next level of analysis might entail statistical analysis to better understand patterns. These can then be taken a step further to develop forecasts or models for predicting how customers might respond to a specific marketing campaign or ongoing service/product offerings. When an organization has a good view of what is happening and what is likely to happen, it can also employ other techniques to make the best decisions under the circumstances. These eight levels of analytics are described in more detail in a white paper by SAS (sas.com/news/sascom/analytics_levels.pdf).

This idea of looking at all the data to understand what is happening, what will happen, and how to make the best of it has also been encapsulated by INFORMS in proposing three levels of analytics. These three levels are identified (informs.org/Community/Analytics) as descriptive, predictive, and prescriptive. Figure 1.11 presents a graphical view of these three levels of analytics. It suggests that these three are somewhat independent steps and one type of analytics applications leads to another. It also suggests that there is actually some overlap across these three types of analytics. In either case, the interconnected nature of different types of analytics applications is evident. We next introduce these three levels of analytics.

Three Types of Analytics

<table>
<thead>
<tr>
<th>Questions</th>
<th>Enablers</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>What happened? What is happening?</td>
<td>✓ Business reporting  ✓ Dashboards  ✓ Scorecards  ✓ Data warehousing</td>
<td>Well-defined business problems and opportunities</td>
</tr>
<tr>
<td>What will happen? Why will it happen?</td>
<td>✓ Data mining  ✓ Text mining  ✓ Web/media mining  ✓ Forecasting</td>
<td>Accurate projections of future events and outcomes</td>
</tr>
<tr>
<td>What should I do? Why should I do it?</td>
<td>✓ Optimization  ✓ Simulation  ✓ Decision modeling  ✓ Expert systems</td>
<td>Best possible business decisions and actions</td>
</tr>
</tbody>
</table>

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Data Mining at the Intersection of Many Disciplines

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining
Is a Blend of Multiple Disciplines

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
# Data Mining Tasks & Methods

<table>
<thead>
<tr>
<th>Data Mining Tasks &amp; Methods</th>
<th>Data Mining Algorithms</th>
<th>Learning Type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prediction</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td>Decision Trees, Neural Networks, Support Vector Machines, kNN, Naïve Bayes, GA</td>
<td>Supervised</td>
</tr>
<tr>
<td>Regression</td>
<td>Linear/Nonlinear Regression, ANN, Regression Trees, SVM, kNN, GA</td>
<td>Supervised</td>
</tr>
<tr>
<td>Time series</td>
<td>Autoregressive Methods, Averaging Methods, Exponential Smoothing, ARIMA</td>
<td>Supervised</td>
</tr>
<tr>
<td><strong>Association</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Link analysis</td>
<td>Expectation Maximization, Apriori Algorithm, Graph-Based Matching</td>
<td>Unsupervised</td>
</tr>
<tr>
<td>Sequence analysis</td>
<td>Apriori Algorithm, FP-Growth, Graph-Based Matching</td>
<td>Unsupervised</td>
</tr>
<tr>
<td><strong>Segmentation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>k-means, Expectation Maximization (EM)</td>
<td>Unsupervised</td>
</tr>
<tr>
<td>Outlier analysis</td>
<td>k-means, Expectation Maximization (EM)</td>
<td>Unsupervised</td>
</tr>
</tbody>
</table>

Evolution of Business Intelligence (BI)

Organizations have to work smart. Paying careful attention to the management of BI initiatives is a necessary aspect of doing business. It is no surprise, then, that organizations are increasingly championing BI and under its new incarnation as analytics. Application Case 1.1 illustrates one such application of BI that has helped many airlines as well as, of course, the companies offering such services to the airlines.

FIGURE 1.9 Evolution of Business Intelligence (BI).

T echnical staf build the data warehouse - Organizing - Summarizing - Standardizing

Data warehouse

Business user

Managers/executives

BPM strategies

Fu ture component: Intelligent systems

User interface

- Browser

- Portal

- Dashboard

Data W arehouse Environment

Business Analytics Environment

Performance and Strategy

Data Sources

FIGURE 1.10 A High-Level Architecture of BI. (Source: Based on W. Eckerson, Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligent Solutions. The Data Warehousing Institute, Seattle, WA, 2003, p. 32, Illustration 5.)

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Organizations have to work smart. Paying careful attention to the management of BI initiatives is a necessary aspect of doing business. It is no surprise, then, that organizations are increasingly championing BI and under its new incarnation as analytics.

Application Case 1.1 illustrates one such application of BI that has helped many airlines as well as, of course, the companies offering such services to the airlines.

A High-Level Architecture of BI

![Diagram of BI architecture]

**Data Warehouse Environment**
- Technical staff
  - Build the data warehouse
    - Organizing
    - Summarizing
    - Standardizing

**Business Analytics Environment**
- Business users
  - Access
  - Manipulation, results

**Performance and Strategy**
- Managers/executives
  - BPM strategies

**User interface**
- Browser
- Portal
- Dashboard

**Future component:** Intelligent systems

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Although its value proposition is undeniable, to live up its promise, the data has to comply with some basic usability and quality metrics. Not all data is useful for all tasks, obviously. That is, data has to match with (have the coverage of the specifics for) the task for which it is intended to be used. Even for a specific task, the relevant data on hand needs to comply with the quality and quantity requirements. Essentially, data has to be analytics ready. So what does it mean to make data analytics ready? In addition to its relevance to the problem at hand and the quality/quantity requirements, it also has to have a certain data structure in place with key fields/variables with properly normalized values. Furthermore, there must be an organization-wide agreed-on definition for common variables and subject matters (sometimes also called master data management), such as how you define a customer (what characteristics of customers are used to produce a holistic enough representation to analytics) and where in the business process the customer-related information is captured, validated, stored, and updated. Sometimes the representation of the data may depend on the type of analytics being employed. Predictive algorithms generally require a flat file with a target variable, so making data analytics ready for prediction means that data sets must be transformed into a flat-file format and made ready for ingestion into those predictive algorithms. It is also imperative to match the data to the needs and wants of a specific predictive algorithm and/or a software tool—for instance, neural network algorithms require all input variables to be numerically represented (even the nominal variables need to be converted).
Big Data with Hadoop Architecture

**Logical Architecture**

- **Processing: MapReduce**
  - Job Tracker
  - Task Tracker
  - Task Tracker
  - Task Tracker
  - Mapper
  - Mapper
  - Mapper
  - Shuffle and Sort
  - Reducer
  - Reducer
  - Reducer

- **Storage: HDFS**
  - NameNode
  - Data Node
  - Data Node
  - Data Node

**Process Flow**

- Input Data Set
  - Split 0
  - Map 0
  - Reduce 0
  - Split 1
  - Map 1
  - Reduce 0
  - Split n
  - Map n
  - Reduce 0

**Physical Architecture**

- Hadoop Cluster
  - Master
  - Slave
  - Slave
  - Slave
  - Slave
  - Slave
  - Slave
  - Slave

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

Big Data with Hadoop Architecture

Logical Architecture

Storage: HDFS

Big Data with Hadoop Architecture

Process Flow

Big Data with Hadoop Architecture

Hadoop Cluster

Traditional ETL Architecture

Offload ETL with Hadoop (Big Data Architecture)

Spark and Hadoop

Source: http://spark.apache.org/
Data Science and Business Intelligence

Data Science

Predictive Analytics and Data Mining (Data Science)

- Typical Techniques and Data Types
  - Optimization, predictive modeling, forecasting, statistical analysis
  - Structured/unstructured data, many types of sources, very large datasets

- Common Questions
  - What if...? 
  - What’s the optimal scenario for our business? 
  - What will happen next? What if these trends continue? Why is this happening?

Business Intelligence

- Typical Techniques and Data Types
  - Standard and ad hoc reporting, dashboards, alerts, queries, details on demand
  - Structured data, traditional sources, manageable datasets

- Common Questions
  - What happened last quarter? 
  - How many units sold? 
  - Where is the problem? In which situations?

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Data Science and Business Intelligence

Predictive Analytics and Data Mining (Data Science)

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Predictive Analytics and Data Mining (Data Science)

Structured/unstructured data, many types of sources, very large datasets

Optimization, predictive modeling, forecasting statistical analysis

What if...?
What’s the optimal scenario for our business?
What will happen next?
What if these trends continue?
Why is this happening?

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Profile of a Data Scientist

• **Quantitative**
  – mathematics or statistics

• **Technical**
  – software engineering, machine learning, and programming skills

• **Skeptical mind-set** and **critical thinking**

• **Curious** and **creative**

• **Communicative** and **collaborative**

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Data Scientist Profile

Data Scientist

- Technical
- Quantitative
- Curious and Creative
- Skeptical
- Communicative and Collaborative

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Big Data Analytics
Lifecycle
Key Roles for a Successful Analytics Project

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Overview of Data Analytics Lifecycle

1. Discovery
   - Do I have enough information to draft an analytic plan and share for peer review?

2. Data Prep
   - Do I have enough good quality data to start building the model?

3. Model Planning
   - Do I have a good idea about the type of model to try? Can I refine the analytic plan?

4. Model Building
   - Is the model robust enough? Have we failed for sure?

5. Communicate Results

6. Operationalize

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Overview of Data Analytics Lifecycle

1. Discovery
2. Data preparation
3. Model planning
4. Model building
5. Communicate results
6. Operationalize

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Key Outputs from a Successful Analytics Project

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Summary

• AI

• Big Data Analytics
References

- Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications.